TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Anàlisi, modelització i predicció del preu de compra de l'habitatge amb dades d'un portal web

Autor: David Punset Gálvez

Director: Montserrat Guillén Estany

Departament: Econometria, Estadística i Economia Aplicada

Convocatòria: juny-setembre 2022

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

AGRAÏMENTS

En especial, a la meva tutora Montserrat Guillén, en la part acadèmica per la seva disponibilitat, entrega i implicació tant resolent dubtes com plantejant millores, i sobretot en la part humana, pel seu acompanyament al llarg de tot el projecte, la seva empatia i el seu inestimable tracte.

I a la meva mare, perquè tot li dec a ella.

RESUM

El projecte que procedim a presentar cerca plantejar i desenvolupar un procés complet i autònom, tot garantint l'automatització, fiabilitat i qualitat, que ens permeti extreure dades d'anuncis publicats a la web d'un portal immobiliàri, mitjançant la construcció d'un algoritme que apliqui la tècnica del *web scraping*, transformar-les i projectar-les concebent una base de dades consistent. A partir d'aquesta es procedirà a efectuar una anàlisi descriptiva, una posterior modelització del preu de compra de l'habitatge i la implementació d'un sistema de predicció de preu.

PARAULES CLAU

Preu de compra de l'habitatge, *web scraping*, *data cleaning*, anàlisis descriptiva, distribució, correlació, regressió lineal, regressió robusta, predictiu, aplicació Shiny

ABSTRACT

The project that we are going to present seeks to propose and develop a complete and autonomous process, while guaranteeing automation, reliability and quality, which will allow us to extract data from ads published on the web of a real estate portal, through the construction of an algorithm that applies the web scraping technique, transform and project them by creating a consistent database. Based on this, a descriptive analysis will be carried out, a subsequent modeling of the purchase price of housing and the implementation of a price prediction system.

KEYWORDS

Purchase price of housing, web scraping, data cleaning, descriptive analysis, distribution, correlation, linear regression, robust regression, predictive, Shiny app

CLASSIFICACIÓ AMS (MSC2010):

68T40 Robotics

62J05 Linear regression

62J20 Diagnostics

62M20 Prediction

ÍNDEX

I. INTRODUCCIÓ

El present projecte neix de la motivació de dissenyar i desenvolupar un procés integral, fiable i flexible, que abasti des de l'extracció de la dada bruta fins al lliurament d'una aplicació consumible.

En línia amb això, concretem el nostre procés *end-to-end* en l'objectiu principal d'estudiar els determinants del preu de l'habitatge, prosseguint amb la corresponent modelització i extreure'n un ús predictiu, obtenint les dades directament de la web d'un portal immobiliàri.

Així doncs, el nostre full de ruta es dibuixa sobre quatre blocs: la recol·lecció de les dades i la seva materialització en una base completa, consistent i fiable; l'anàlisi descriptiva de les variables precisades; la modelització del preu de compra de l'habitatge; la implementació d'un sistema de predicció de preu. Els quatre blocs desenvolupats sota el propòsit de garantir tant l'automatització dels processos com la qualitat de les dades i del producte final.

Respecte a l'obtenció de les dades, dissenyarem un algoritme que utilitzant la tècnica del *web scraping*, obtingui les dades que desitgem dels anuncis publicats a la web i les bombegi cap a la nostra màquina, sense ser detectat per mecanismes *anti-scraping* i minimitzant els temps d'execució. El nostre procés actuarà sobre una àrea territorial concreta, escollint per aquesta, la ciutat de València. Un cop l'algoritme hagi finalitzat i disposem dels fitxers amb les dada bruta, entrarà en joc el *data cleaning*. En aquest punt, summament important i d'un grau major per estar consumint d'un origen de dades que no està ideat per això, buscarem garantir la qualitat de la nostra base de dades, tot corregint desajustos per la disposició de les dades a la web, possibles errors de l'algoritme i incongruències entre dades de l'anunci. Seguidament, passarem a conèixer en profunditat les dades mitjançant l'anàlisi exploratòria de les variables recollides, alhora que traçarem el seu tractament, contrastarem les dades amb fonts oficials i seguirem donant feedback, a l'haver excavat més en el comportament de les dades, de la qualitat d'aquestes. Amb la base de dades tancada i l'entrellat de conclusions extretes de l'anàlisi descriptiva, prosseguirem amb la modelització del preu de compra, plantejant dos línies de treball per tal d'obtenir un model aplicable a qualsevol immoble (General) i un restringit a habitatges de València ciutat (No General), així com avaluant les seves corresponents capacitats predictives. Finalment, construirem una aplicació web dissenyada perquè qualsevol consumidor pugui, introduint les dades d'un immoble, obtenir una estimació del preu d'aquest habitatge.

Una eina que permeti al usuari saber si el preu d'un immoble està per sota o per sobre del preu de mercat o quin seria el preu en que es taxaria un immoble amb les característiques que està fixant.

II. METODOLOGIA

Presentat l'esquelet del nostre projecte, és necessari detallar les diferents eines, tècniques i recursos que hem emprat pel seu desenvolupament. Com hem esmentat a la introducció, les decisions preses durant tot el procés han estat supeditades a la idea de no trencar amb l'automatització. En sintonia amb això, també ho ha estat l'elecció de les opcions de treball quant a la metodologia. Així doncs, aquí rau el motiu de la resolució al primer dilema que se'ns plantejava, la font de dades. Optant pel portal immobiliari Idealista, perquè, pel fet d'ésser un dels portals web de referència a nivell nacional, li suposem una interfície web més estable, amb menys canvis que puguin implicar-nos posteriors desenvolupaments en l'algoritme de *web scraping*. Tanmateix, no ens hem decantat, per exemple, pel portal Fotocasa, perquè presenta un sistema *anti-scraping* molt més estricte, suposant això una feblesa i posant en perill la fiabilitat de l'algoritme. Respecte a l'àrea territorial d'estudi seleccionada, hem escollit la ciutat de València. Aquí restem importància a la tria d'una o altra alternativa, seleccionant-la perquè a l'ésser una ciutat, i la tercera del país en nombre d'anuncis publicats a Idealista, ens garanteix un bon nombre d'observacions, alhora que ens és una zona més desconeguda, no com Barcelona, fent més atractiu arribar a certes conclusions tot contrastant-les després. Així mateix, quant al software emprat per construir l'algoritme d'extracció de dades, hem utilitzat UiPath, perquè, a més de ja haver treballat amb ell, les seves funcionalitats de *web scraping* són molt àmplies.

En referència als blocs restants, que s'estenen des del *data cleaning* a l'aplicació d'estimador de preus, hem utilitzat el software estadístic R. Respecte a la tasca de garantir la qualitat de dades, punt crucial del projecte desenvolupat tant a l'apartat de *data cleaning* com a l'anàlisi descriptiva posterior, hem prioritzat en tot moment l'automatització, descartant portar a terme segons quines correccions que comportessin treball manual. Per una altra banda, en l'anàlisi descriptiva hem recorregut a noves fonts de dades, publicades per l'Ajuntament de València, per contrastar el que observàvem sobre la nostra base. Ara bé, no hem utilitzat aquestes dades per modelitzar, ja que això hagués suposat crear-nos una dependència més, amb la consegüent pèrdua d'autonomia. Així mateix, una altra premissa que ens havíem marcat era que l'anàlisi exploratori que efectuéssim fos ric en visualitzacions i així hem procedit, implementat aquestes mitjançant el paquet *ggplot2* de R. Respecte a la modelització, partint d'una divisió aleatòria en un grup d'entrenament i un altre de test (amb proporcions 0.7 i 0.3, respectivament), per evitar el sobreajustament, hem plantejat regressions lineals i robustes, incloent transformacions logarítmiques per garantir el supòsit de normalitat dels residus, estimant els corresponents paràmetres

per mínims quadrats ordinaris (OLS). Finalment, hem construït l'aplicació d'estimador de preus utilitzant la llibreria Shiny.

III. WEB SCRAPING

El plantejament del workflow corresponent al procés de *scraping* ha d'estar subjugat a l'equilibri entre cost i fiabilitat. Es tracta de maximitzar l'eficiència i reduir el temps d'obtenció de la base de dades, sempre sense perdre de vista la probabilitat d'ésser detectats pel portal i capats, fet que no només trencaria òbviament el flux i impossibilitaria l'obtenció de les dades, sinó que posaria en perill futurs intents de *scraping*.

Aquest exercici de balanceig es trasllueix en la forma en que abordem els diferents impediments que sorgeixen i la presa de decisions respecte aquests. El primer dels problemes que se'ns presenten quan ens plantegem fer un *scraping* és el fet que el portal ens impedeix scrapejar més de 60 pàgines d'immobles, extrapolant a número d'anuncis equivaldria a una xifra de 1.800 (30 anuncis per pàgina). Això comporta escollir entre acotar per zona a 1.800 observacions, elecció no desitjable per pèrdua de representativitat, o dividir en N execucions per tal d'abastir-la en la seva totalitat.

S'ha optat per la segona solució és a dir la segmentació, i escollim com a primera divisió la geogràfica, partint del nivell mínim raonable d'agregació. Aquest és el determinat per portal com a zones. En termes de divisió territorial, equivaldria, no en la seva totalitat, a districtes d'una ciutat. Remarco que no hi ha una equivalència total, ja que també el portal recull nivells més petits. El principal problema que es deriva de fer una divisió per territori no és un problema de representativitat, ja que assumim com premissa de partida (més endavant ho avaluarem detingudament de cara a una correcta modelització) que el número d'observacions per agregació ens traçarà una aproximació suficient a la realitat. Ara sí, per contra se'ns afegeix un grau de dificultat més de cara a tractar el problema de duplicitats, que abordarem més endavant, perquè es dona el cas que s'han observat anuncis diferents sobre el mateix immoble però recollits en zones diferents.

Ara bé, la territorial no és l'única divisió que hem de plantejar, ja que una zona pot presentar més de 1.800 immobles. En aquests casos, utilitzaríem la variable preu, ja que ens brinda ordinalitat, aplicant els talls necessaris.

Figura 3.1: Exemple de mapa de localització i buscador per zones d'una ciutat (València) en un portal

Tot seguit, abordem l'extracció de les dades de cada anunci. Aquesta es bifurca en dues opcions d'estructura. La primera d'elles i la més intuïtiva seria enfocar l'algoritme a seleccionar immoble a immoble i anar entrant a les seves adreces d'internet o *urls*. Això suposa que realitzem una *request* al servidor del portal per cada anunci.

Després de realitzar proves, aconseguim, espaiant els temps entre peticions, que el portal no detecti l'algoritme; però, tot i així, sobre els 500 anuncis acaben tallant-nos l'accés a la web per activitat sospitosa. Enfront això, presentem una segona alternativa. Aquesta consisteix en emprar els filtres que ens brinda el portal (estat, planta, característiques...). Així doncs, marcant i desmarcant farem que l'algoritme visualitzi els immobles que compleixen aquella característica. El guany tant en cost com en fiabilitat és inapel·lable, ja que ara canviem d'un escenari, posem per exemple una zona amb 1.000 anuncis, on necessitem 1.000 peticions per saber quins immobles disposen de garatge a un altre on realitzarem tantes peticions com número de pàgines es presentin. Seguint amb l'exemple, si dels 1.000 anuncis, en tenim 300 amb garatge, l'algoritme realitzarà únicament 10 peticions al servidor web (300/30), al ser una variable dicotòmica. En el cas de variables amb més nivells (L), simplement serà cobrir L-1. Aquí agregar, que en termes d'eficiència, sempre serà millor partir dels nivells amb menys freqüència per tal de reduir pàgines. La contra d'aquesta alternativa és fàcilment imaginable sense necessitat de veure la interfície web i és que en la *url* de l'anunci sempre obtindrem més informació. Ara bé, utilitzant la concepció del portal en sí és igualment ràpid pensar que les variables que **Idealista** ha escollit per permetre a

l'usuari acotar la seva recerca són probablement les que el portal creu que esdevenen més rellevant per al client a l'hora d'escollir un habitatge i és plausible pensar que es postularan a la vegada com les més determinants en la determinació del preu d'aquest.

Finalment, també s'han dut a terme les corresponents proves per garantir que aquesta via de la segmentació és factible. Espaiant temps entre peticions, en aquest cas canvis de pàgina, aconseguim que el portal no ens detecti. És essencial el fet que el portal ha de pensar en tot moment, o millor dit, no ha de sospitar que qui està fent les consultes no sigui un usuari i no és esperable que aquest recorri en segons centenars de pàgines. Posem ambdues opcions a sobre de la taula no com alternatives excloents, sinó com dos fluxos independents en el workflow que estem dibuixant. Ara bé, tota variable que puguem aconseguir mitjançant filtres, l'obtindrem per aquesta via òbviament per la diferència de cost i risc. I en relació al risc, per profilaxis, també llançarem sempre primer aquest procés. Amb tot, aquests dos fluxos no esdevindran el primer pas de l'algoritme, sinó que partiran d'un primer procés on recollirem tots els immobles que conformen la zona seleccionada. I en aquesta execució inicial no recollirem únicament les *urls* dels anuncis, sinó que rascarem totes les dades que se'ns brinda en la primera pantalla de visualització de l'anunci. És a dir, la informació que mostra **Idealista** sense necessitat d'entrar en l'enllaç, novament i en sintonia amb el comentat anteriorment, veiem que són dades essencials.

Figura 3.2: Exemple de la primera pantalla d'informació sobre un immoble

Amb tot això, podem procedir a presentar l'esquema del workflow dissenyat. A la figura 2.3 es pot observar que partim amb el procés *Principal*, que hem presentat al paràgraf anterior, que culmina amb l'obtenció de la que anomenem *Database P1*. Aquesta, com hem dit, es constituirà per les *urls* de cada anunci juntament amb el primer conjunt de variables.

Aquesta primera base de dades per una banda nodrirà del enllaços dels immobles al procés *Deep_Dive*, que és el que hem presentat on l'algoritme entra a cada *url* de cada anunci, i a la vegada servirà, mitjançant la intersecció (*Merge*), per completar el nivell o nivells restants de les variables filtre que obtindrem mitjançant el procés *Variable*. L'output d'aquest procés seran un total d'**n** bases (*Database V(i)*), on **n** serà igual al total de nivells de la variable menys 1, com hem explicat abans. Per exemple, per la variable *Estat del immoble*, tenim que pot prendre com a valors les següents categories: *A reformar*, *Obra nova* i *En bon estat*, per tant, obtindrem dos bases amb les *urls* dels immobles que correspongui (*Reform* i *New*), que al fer la intersecció mitjançant la url amb *Database P1*, ens indicarà les *urls* que no son d'immobles a reformar ni obra nova, per tant, els que considerem i considera Idealista com *En bon estat*.

Finalment, obtenim de cada flux una base de dades (*Database 1* I *Database 2*) que unint ens dona la que seria la nostra base final d'una zona; ja que, recordem que la primera divisió que hem realitzat ha estat la d'agrupar per zones i que l'esquema que es presenta esdevindria l'execució d'una zona, tenint per tant N workflows totalment independents. La unió de totes les bases obtingudes per una base final de la ciutat, la durem a terme més endavant, perquè aplicarem el *data cleaning* per cada zona. Tot i que pugui semblar més ineficient, és més fàcil detectar falles en algun procés separant. És interessant fer un darrer aclariment respecte a la seqüencialitat que hem donat al workflow d'executar primer el procés *Variable* que *Deep_Dive*. I és que ho fem no a cada zona, sinó que executem el primer procés per totes les zones i un cop finalitzat, llençaríem el segon de nou per a totes.

Ha sospesat el risc d'aquest segon llançament i a l'ésser processos independents no té un cost afegit procedir així. Per una altra banda, *Deep_Dive* com comentàvem podia ser bloquejant i, avançant-nos és el que ha succeït. Per això, finalment no ha sigut factible implementar el *Deep_Dive* i aquest procés s'ha caigut del projecte.

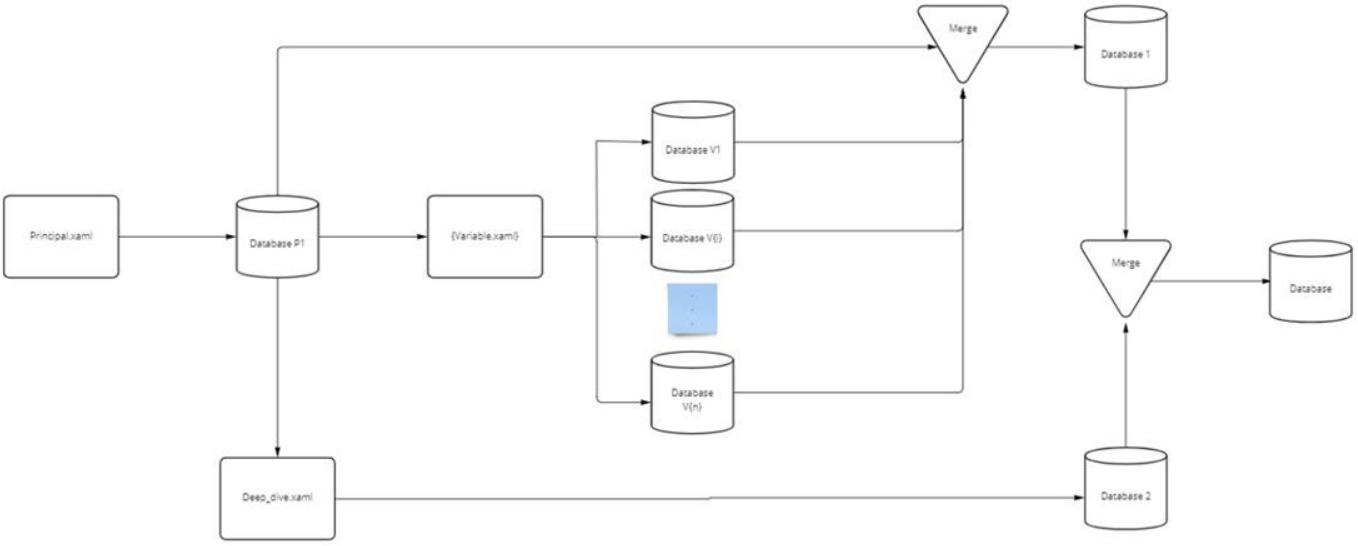
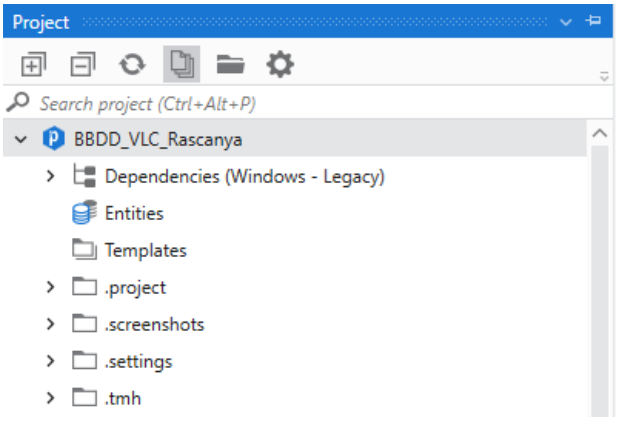


Figura 3.3: Workflow del procés de *scraping* i obtenció de la base de dades

Procedim seguidament a disseccionar els processos *Principal* i *Variable* (el *Deep_Dive* el tenim en un projecte separat) ja des dins del software que hem emprat.
UiPath ens permet visualitzar amb la seva interfície de forma ràpida els processos que composen el nostre projecte de *scraping*.



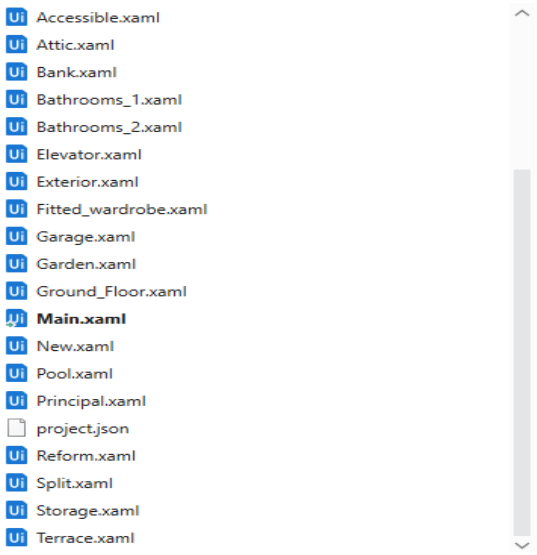


Figura 3.4: Projecte amb els subprocessos del procés *Variable* i el procés *Principal*

Si ens aturem al *Principal*, figura 3.3, podrem observar amb una vista ràpida els passos del procés. Primerament, l'algoritme obra amb el navegador (és preferible Chrome per funcionalitat i és necessari tenir afegida l'extensió corresponent *UiPath Web Automation*) la url que el dirigeix a la pàgina de la zona a scrapejar, amb la totalitat dels immobles, i fa un ancoratge sobre aquesta.

Seguidament i mitjançant selectors (en XML) captura els atributs, juntament amb els seus tags (que hem predefinit prèviament) de l'immoble que està destacant. Una funcionalitat molt necessària és que l'algoritme no falli si troba un error, per exemple si un atribut no existís per a un determinat immoble, cal que ho apunti amb tags diferents als marcats. Tot i així, i per profilaxis, l'algoritme té afegit en aquest punt un *timeout* de 5 segons, a partir dels quals assumim que hi ha hagut alguna falla que no es pot passar per alt i requereix la nostra revisió. Així doncs, l'algoritme itera pels 30 immobles que formen la pàgina i escriu aquests en un fitxer csv. De cara a canviar de pàgina, li hem incorporat a l'algoritme un *delay* de 4 segons; suficient perquè no ens detectin com comportament anòmal.

Hem estimat que l'algoritme tarda en una execució de 1.000 anuncis entre 2.5 i 3 minuts (l'ajust s'ha fet començant amb uns temps molt majors i s'ha anat escurçant a mesura que no saltaven les alarmes). Uns temps que s'allunyen molt de la conducta esperable d'un usuari, però que **Idealista** no arriba a detectar.

Per una altra banda, no deixa de sorprendre que des del portal Idealista tampoc no fan un monitoreig de les *requests* per IP en quant a recerques molt diferents com esdevé que el nostre algoritme sondegi totes les pàgines d'una zona i canviï i repeteixi amb una altra i així successivament. Respecte a això, en sintonia amb l'actitud profilàctica que fem predominar, les execucions s'han llençat en horari nocturn i/o en caps de setmana.

En quant a la resta de subprocessos que configuren el flux *Variable*, els temps no s'allunyen gaire de l'interval del *Principal* (els temps reals són menors perquè únicament llegirà el mateix volum si tots els immobles compleixen aquell filtre), ja que tot i que la part del *scraping* és molt més senzilla (recordem que únicament ataquem a un atribut on tenim l'enllaç tant en *url* com en text, aquest darrer correspondria a la direcció), l'algoritme és molt ràpid i allò que penalitza més és el canvi de pàgina.

Respecte al *Deep_Dive*, tot i que no s'ha mesurat detingudament perquè no ha acabat resultant operatiu, perquè és molt més costós en temps en relació al valor afegit que suposa en quant a informació extra, aproximadament 2-3 segons per anunci.

Així doncs, estimem que per scrapejar una zona al complet l'algoritme tarda sobre uns 25-35 minuts. Després hauríem de sumar-li uns 15-20 min de monitoreig i revisió dels logs i bases obtingudes. Això fa que sigui més costós en temps el número de zones processades que el volum d'aquestes.

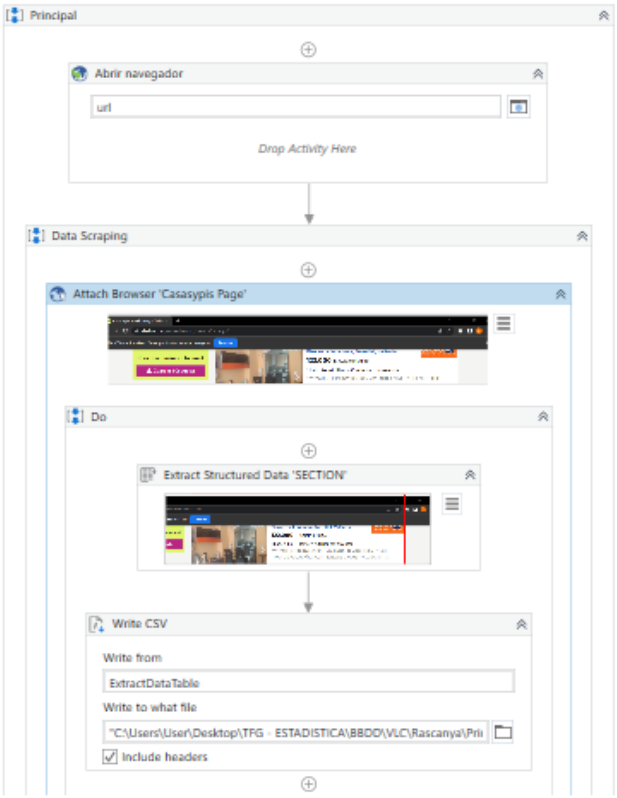


Figura 3.5: Visualització dels passos del procés *Principal*

A continuació, exposem les variables o nivells que obtindríem com a output dels tres processos implementats. Respecte al *Deep_Dive*, el valor afegit i que al no haver-se pogut dut a terme, suposa una pèrdua no menyspreable, està sobretot en la descripció de l'immoble (sobre una descripció completa es podria haver treballat molt millor per detectar anomalies com immobles sense cèdula d'habitabilitat, llogats, ocupats il·legalment...) i en l'any de construcció que trobem en el paquet de text que esmentem com característiques bàsiques.

Obviant això, podem veure a la taula 3.1, l'algoritme finalment emprat recopila no només la informació més significativa, sinó que a més a més, aconsegueix diverses variables més pròpies de l'anunci i molt interessants com el número de fotos o l'agencia immobiliària, si no és un particular, que el publica.

Variable	Origen	Descripció
URL	Principal.xaml	Url anunci immoble
TYPE	Principal.xaml	Tipus d'habitatge
PRICE	Principal.xaml	Preu de venta
PREVIOUS_PRICE	Principal.xaml	Preu anterior de venta (quan es doni)
HAB	Principal.xaml	Número d'habitacions
M2	Principal.xaml	Superfície total
FLOOR	Principal.xaml	Planta
DESC_1	Principal.xaml	Fragment de la descripció publicada
PHOTO	Principal.xaml	Número de fotos de l'anunci
AGENCY	Principal.xaml	Agencia immobiliària que publica l'anunci (quan es doni)
SUBZONE	Principal.xaml	Barri
ZONE	Principal.xaml	Zona

Taula 3.1. Llistat de les variables obtingudes en fer un *scraping* del portal immobiliari

Comentar que respecte a la variable banys no s'ha complert el suggeriment de scrapejar els nivells de la variable menys freqüents, com seria en aquest cas quan presenta més de dos banys. La decisió no ha estat casual, sinó que s'ha preferit optar per les dues categories menys susceptibles d'error i facilitar així la revisió posterior.

Variable/nivell	Origen	Descripció
ACCESSIBLE	Variable.xaml	Habitatge accessible
ATTIC	Variable.xaml	Àtic
GROUND_FLOOR	Variable.xaml	Baix
BATHROOMS_1	Variable.xaml	Un únic bany
BATHROOMS_2	Variable.xaml	Dos banys
EXTERIOR	Variable.xaml	Exterior
NEW	Variable.xaml	Nova construcció
REFORM	Variable.xaml	Per reformar
ELEVATOR	Variable.xaml	Ascensor
GARAGE_1	Variable.xaml	Garatge
GARAGE_2	Variable.xaml	Garatge inclòs en el preu o no
TERRACE	Variable.xaml	Terrassa
STORAGE	Variable.xaml	Traster
GARDEN	Variable.xaml	Jardí
POOL	Variable.xaml	Piscina
SPLIT	Variable.xaml	Aire condicionat
FITTED_WARDROBE	Variable.xaml	Armaris de paret
BANK	Variable.xaml	Habitatge de banc

Taula 3.2. Llistat de les variables addicionals obtingudes en fer un *scraping* del portal immobiliari

Variable	Origen	Descripció
DESCRIPTION	Deep_Dive.xaml	Descripció completa
BASIC_CHARACTERISTICS	Deep_Dive.xaml	Text característiques bàsiques
ENERGY_CERTIFICATE	Deep_Dive.xaml	Certificat energètic

Taula 3.3. Llistat de les variables no obtingudes en fer un *scraping* del portal immobiliari

Per tal de tancar el workflow, partint dels 18 fitxers csv. obtinguts en el procés de *scraping* procedim mitjançant R, ja que en el següent apartat passarem ja a treballar amb ell, a llegir el fitxer *Principal.csv* i anem unint els 17 restants; de tal forma que ens queda una base de dades en brut amb 30 columnes. Aquesta base serà l'input sobre la que treballarem en el *data cleaning*.

En el nostre cas d'estudi, per a València ciutat, s'ha processat la totalitat de les zones que ens presentava **Idealista**. Així doncs han suposat 16 execucions del workflow que en termes absoluts representen vuit hores d'execució, però qu degut a les verificacions i controls manuals, han suposat un esforç que ha durant moltes setmanes. Amb tot, hem obtingut informació d'una xifra de 7.791 anuncis.

Finalment, al llarg de l'apartat hem subratllat com el risc d'ésser detectats com a *bot* ens ha fet prendre mesures com alentir l'algoritme com també hem vist com no hem pogut dur a port el *Deep_Dive*, però no és difícil veure la solució a aquests problemes. I és que ens hem col·locat en un escenari d'una única direcció IP, però podríem haver muntat un sistema amb VPNs que fes sortir l'algoritme des de diferents IPs o, inclús, hi ha serveis web que ens permeten orquestrar des de l'anonimat el que necessitem. No ha estat fortuïta la decisió de no sortir del paradigma d'una única IP, sinó que a la balança, que esmentàvem al inici d'aquest apartat, hi hem sumat al risc i cost, la robustesa, preferint un algoritme alliberat de les dependències respecte a un sistema així.

IV. DATA CLEANING

En aquest apartat, abordem un punt transcendental en el nostre projecte: garantir la qualitat de les dades de la nostra base. Alhora, a l'ésser la construcció d'aquesta mitjançant la tècnica de *data scraping*, es requereix un grau de depuració i transformació major, ja que obtenim les dades d'una font concebuda per ser consumida a nivell d'usuari i no per a un projecte com el present.

Així doncs, aquí entra en joc el següent pas del nostre workflow (tot i que s'ha separat del que hem presentat en l'apartat anterior, no ha estat perquè no sigui un pas més dins d'aquest), el *data cleaning*.

Abans de tot i amb idea de facilitar tant *d*ata cleaning com l'anàlisi descriptiva posterior, procedim amb la construcció d'un identificador, més manejable, del anunci (ID) que extraiem de la url d'aquest mitjançant expressions regulars.

	URL	ID
1	https://www.idealista.com/inmueble/25839422/	25839422

Taula 4.1. Exemple construcció ID

Després d'examinar detingudament cada variable, trobem varis punts problemàtics que anirem esmentant junt amb la resolució proposada i aplicada.

El primer d'ells esdevé la presència de *missings* a les variables HAB, M2 i FLOOR. Ara bé, no ens referim a abordar el tractament de les dades mancants en sí, sinó al fet que quan alguna d'aquestes dades no està informada a la web, l'algoritme ho ignora i escriu la següent variable a la columna de la que falta. De forma que en algunes observacions ens consta la planta on hauria de figurar la superfície i aquesta on hauria de constar el número d'habitacions. Novament fem ús d'expressions regulars per detectar aquests patrons i els corregim, reubicant les dades.

Aquí hem de remarcar la importància de cara al procés de *scraping* de no extreure únicament la xifra de les variables, sinó tota aquella informació, en aquest cas el text "hab." I "m^2", que després necessitem per identificar la dada i depurar-la amb garanties.

	HAB	M2	FLOOR	HAB_v	M2_v	FLOOR_v
172	*NA*	54	Planta 1ª exterior sin ascensor	54 m^2	Planta 1ª exterior sin ascensor	*NA*
181	*NA*	70	Bajo exterior sin ascensor	70 m^2	Bajo exterior sin ascensor	*NA*
361	*NA*	160	Planta 1ª exterior sin ascensor	160 m^2	Planta 1ª exterior sin ascensor	*NA*

Taula 4.2. Exemples d'observacions amb desajust de dades junt amb la seva correcció (per exemple, a la primera fila, 54 m^2 correspon a metres quadrats i no al nombre d'habitacions)

Cal afegir que respecte a la variable FLOOR, en algunes observacions s'ha detectat que també es desajusta quan no ve informada la planta, però en canvi, sí que ve informat si el immoble disposa d'ascensor i/o és exterior. En aquests casos hem procedit a informar la variable com a *missing*. Això no obstant, tot i que disposem de les variables ELEVATOR i EXTERIOR, ens guardem també aquestes referències. Amb l'objectiu, com veurem més endavant, de contrastar una mateixa informació recopilada per diferents vies. El mateix passa amb la mateixa variable FLOOR i les variables GROUND_FLOOR i ATTIC. Per acabar, no ens deixarem de depurar la variable, quan sí tenim la planta indicada, eliminant el text. Quan es tracti d'una planta baixa, informarem la variable FLOOR a zero.

Prosseguim amb la tipologia de l'immoble que ha de venir informada a la variable TYPE. Tanmateix, aquesta la trobem unida a la direcció del immoble. Per tant, utilitzant de nou expressions regulars, extraurem el tipus d'habitatge i, alhora, el barri on s'ubica (SUBZONE). Mentre que la variable ZONE, la creem en el mateix moment de la lectura dels fitxers, ja que els tenim dividits en *paths* per zona. Afegim que en obtenir la tipologia dels immobles, observem el nivell *Finca rústica* per tres observacions. Hem comprovat que es tracta d'habitatges i no de terrenys rústics sense edificació.

Quant al barri, per tal de garantir que no agafem per error el carrer de l'immoble, hem *scrapejat* prèviament tots els barris de les zones estudiades. Tenim així els topònims que hem de trobar dins del text.

	TYPE	TYPE_v
773	Piso	Piso en calle Salvador Giner, 3, El Carme, València
1018	Estudio	Estudio en calle San Vicente Mártir, Sant Francesc, València
1019	Ático	Ático en Sant Francesc, València
1020	Piso	Piso en calle Guillem de Castro, El Carme, València
1033	Dúplex	Dúplex en La Seu, València

Taula 4.3. Exemples de transformació de la variable TYPE

Pel que fa a la variable PREVIOUS_PRICE, hi ha observacions on s'indica únicament el text "*opc.*" i que, per tant, hem de convertir a *missing*, ja que el immoble no té preu anterior informat i l'algoritme està llegint una referència a si existeix a l'oferta una opció a compra o lloguer d'un garatge. Tenim també *missings* a la variable PHOTO quan el anunci no té cap foto publicada. Per tant, transformarem aquests *missings* assignant-los-hi valor zero.

Respecte a la variable AGENCY, haurem d'obtenir el nom de l'agència immobiliària, en cas que no sigui un particular qui publica l'anunci, de la url que hem obtingut de la icona de l'agència que apareixia a l'anunci.

	AGENCY	AGENCY_v
21	lucas-fox-valencia	https://www.idealista.com/pro/lucas-fox-valencia/
23	natalia-navarro	https://www.idealista.com/pro/natalia-navarro/
24	grupo-meri	https://www.idealista.com/pro/grupo-meri/
25	seleccioninmobiliaria	https://www.idealista.com/pro/seleccioninmobiliaria/
26	pilar-fuster	https://www.idealista.com/pro/pilar-fuster/

Taula 4.4. Exemples de transformació de la variable AGENCY

Tot seguit, ens centrem en la variable GARAGE_2, que recull, pels immobles amb garatge, si aquest està inclòs en el preu de venda o si hi ha apart una opció de compra o lloguer juntament amb el preu d'aquest habitatge. A partir d'aquesta variable GARAGE_2, construirem dues variables binàries noves, GARAGE_INCLUDED I GARAGE_NOT_INCLUDED. Pel que fa al increment de preu quan no està inclòs, el recollirem en la variable GARAGE_PRICE, que prendrà valor zero si està inclòs. Mantindrem, per una altra banda, la variable GARAGE, que indicarà únicament si

l'immoble disposa o no de garatge, intacte i de cara a la modelització farem l'avaluació corresponent d'aquest conjunt de quatre variables. Finalitzades aquestes transformacions, la variable GARAGE_2 l'eliminem de la base.

	GARAGE	GARAGE_2	GARAGE_INCLUDED	GARAGE_NOT_INCLUDED	GARAGE_PRICE
102	1	Garaje opc. 60.000 €	0	1	60000
105	1	Garaje opc. 40.000 €	0	1	40000
110	1	Garaje incluido	1	0	0
114	1	Garaje opc. 39.000 €	0	1	39000
115	1	Garaje incluido	1	0	0
122	1	Garaje opc. 70.000 €	0	1	70000

Taula 4.5. Exemples de construcció de variables referents al garatge del immoble

Un altre punt esdevé el crear les columnes corresponents als nivells mancants respecte al número de banys, la planta i l'estat de l'immoble, que tenen més de dos nivells. Recordem que al procés de *scraping*, per tal de minimitzar el cost, deixàvem de capturar un nivell, ja que el podem inferir amb la resta. Així doncs, creem les variables BATHROOMS_3, MIDDLE_FLOORS i GOOD_CONDITION.

Ens restaria, per donar com finalitzades les transformacions de les variables, convertir a numèriques aquelles que correspongui i factoritzar les variables qualitatives.

Així mateix, com hem esmentat abans, amb el compromís sempre present de mesurar i certificar la qualitat de la nostra base, passem a contrastar les variables que hem obtingut d'orígens diferents. És el cas de les referents a si el immoble és exterior o no, disposa d'ascensor i la planta on s'ubica.

Pel que fa a EXTERIOR i ELEVATOR, busquem la incongruència que ens constin com que compleixen la condició, però llegim a EXTERIOR_ELEVATOR el text "interior" i "sin ascensor", respectivament. Fem el mateix exercici a la inversa, detectant el text "exterior" i "con ascensor". Únicament trobem 28 observacions que tenim marcades com que no son immobles exteriors i es contradiu amb el text capturat a EXTERIOR_ELEVATOR. Tot el conjunt pertany a la zona d'Extramurs, fet que ens fa pensar que pugui haver estat fruit d'un error de l'algoritme. En conseqüència, apliquem la correcció marcant com positiu EXTERIOR per a aquestes observacions. Aplicada la correcció, podem eliminar de la base la variable EXTERIOR_ELEVATOR.

Respecte a la planta de l'immoble, utilitzarem GROUND_FLOOR, MIDDLE_FLOORS i ATTIC per trobar contradiccions amb FLOOR, que indica la planta exacta. Tan sols trobem una observació que tenim marcada com a planta intermèdia, però es tracta d'una planta subterrània ("-2"). No aplicarem cap tipus de correcció. En canvi, sí que afegirem el valor zero a aquelles observacions on FLOOR és *missing*, però tenim marcades com GROUND_FLOOR, un total de 49 habitatges. Un altre ajust que durem a terme ens sorgeix quan baixem a la tipologia i trobem que 230 immobles són de construcció horitzontal, però apareixen marcats com MIDDLE_FLOORS. Per tant, els reassignàrem com GROUND_FLOOR.

També hem detectat, pel que fa als nivells REFORM, GOOD_CONDITION i NEW, una incongruència en una observació que apareix marcada com a reformar i obra nova alhora. Al tractar-se únicament d'un immoble, procedim amb una revisió manual consultant la url de l'anunci i confirmem per la descripció completa que es tracta d'un immoble de nova construcció. Per tant, marcarem a 0 la variable REFORM per aquesta observació. Tanmateix, aclarir que no es tracta d'un error del nostre procés, sinó del propi portal permetent que l'habitatge aparegui publicat filtrant tant per "A reformar" com "Obra nova".

Quant a les variables corresponents a la disponibilitat de plaça d'aparcament, hem detectat un únic immoble que es recull dins dels immobles amb garatge (GARAGE), però no es comptabilitza amb garatge inclòs ni no inclòs en el preu (GARAGE_INCLUDED i GARAGE_NOT_INCLUDED). El revisem manualment i veiem que el garatge figura en una opció de compra de una quantia de 0€. Tot i que es podria tractar d'un error amb el preu de l'opció de compra, el categoritzarem com immoble amb el preu del pàrquing inclòs en l'operació (GARAGE_INCLUDED). Alhora, detectem 45 observacions, la totalitat pertanyents al districte de Ciutat Vella, que ens venen categoritzades com GARATGE_NOT_INCLUDED, però el preu associat a l'opció de compra també consta a 0. Ara bé, si revisem les urls dels anuncis, veiem que existeix un preu i, per tant, concloem que es tracta d'un error del nostre algoritme en el procés de *scraping* i procedirem a marcar com *missings* aquestes observacions respecte a GARAGE_PRICE. D'altra banda, entenem la variable GARAGE_PRICE com preu de l'opció de compra, per tant, i en observar preus molt baixos (en un rang de 60 a 150€) referents a lloguers, amb un total de 15 immobles, també els marcarem com *NA*.

Per una altra banda, és interessant remarcar el potencial de la variable DESC_1, que conté la descripció de l'immoble, per detectar i corregir la resta de les variables presents. Ara bé, com que disposem únicament d'un fragment inicial del text de la

descripció, no podent fer una anàlisi completa, no aplicarem cap correcció en aquest punt, sinó que reservarem la variable com a suport en vistes de l'anàlisi descriptiva i la modelització.

Tot i que abordarem el tractament de les dades mancants a l'apartat de modelització, és important subratllar un fet que es pot passar per alt, però que implica una assumpció rellevant, i esdevé que cap de les variables que hem obtingut pel procés *Variable* contindran *missings*. Ara bé, és evident que és possible que en algunes observacions això no es correspongui amb la realitat. Posem, per exemple, un immoble que al portal no consti amb el indicador de que té aire condicionat (SPLIT), però ha estat per omissió o error de l'anunciant. Amb això vull remarcar que mai podrem garantir al complet que les dades són plenament correctes i que no hem d'oblidar la naturalesa pròpia de la nostra font, on hi ha un component humà i una intencionalitat clara. Per tant, admeten un cert error de mesura que ve donat per una transcripció deficient de la realitat de l'habitatge que s'anuncia.

L'admissió dels errors en els anuncis fa que no sigui descabellat pensar que, en alguns casos, la informació mostrada pot estar manipulada amb un biaix positiu o que certes característiques que podrien perjudicar la imatge de l'immoble estiguin invisibilitzades. La darrera depuració que durem a terme esdevé la detecció i eliminació de les observacions duplicades. Posem el focus en el fet que la nostra base de dades està formada per anuncis i es pot donar, i es dóna, el cas que un mateix immoble sigui publicat per usuaris diferents, sobretot si es tracta d'agències immobiliàries.

Primerament, cerquem duplicats per ID, que es poden produir si un habitatge es publica en més d'una zona. Trobem únicament una observació ubicada tant en Benimaclet com Poblats Marítims. Eliminem aquesta segona, ja que el immoble realment es troba a Camí de Vera, Benimaclet.

Com a segon criteri de deduplicació, deixem fora, a part del ID, aquelles variables referents a la ubicació (ZONE i SUBZONE), a l'anunciant (PHONE i AGENCY) i aquelles més manipulables per qui publica (DESC_1, PHONE, PREVIOUS_PRICE). Detectem 504 anuncis duplicats que procedim a eliminar de la nostra base, passant a tenir un total de 7.281 observacions.

Per acabar, aplicat el procés de *data cleaning* i fent balanç final, presentem a la següent taula les variables que conformaran la base de dades del projecte.

Variable	Tipus	Descripció	Valors	Presència NA
ID	Categòrica	Identificador		NO
URL	Categòrica	Url de l'anunci		
ZONE	Categòrica	Zona/districte		NO
SUBZONE	Categòrica	Barri		NO
TYPE	Categòrica	Tipologia immoble	1. *Piso* 2. *Ático* 3. *Dúplex* 4. *Estudio* 5. *Casa o chalet independiente* 6. *Casa de pueblo* 7. *Chalet* 8. *Casa terrera* 9. *Finca rústica*	NO
PRICE	Numèrica	Preu de venta		NO
PREVIOUS_PRICE	Numèrica	Preu anterior de venta		NO
HAB	Numèrica	Número d'habitacions		SÍ
M2	Numèrica	Superfície total		NO
FLOOR	Numèrica	Número de planta		SÍ
DESC_1		Descripció immoble		SÍ
PHOTO	Numèrica	Número imatges publicades		NO
PHONE	Categòrica	Telèfon anunciant		SÍ
AGENCY	Categòrica	Agència		NO

		immobiliària		
ACCESSIBLE	Categòrica		**1** Accessible **0** No accessible	NO
BANK	Categòrica		**1** De banc **0** No propietat bancària	NO
BATHROOMS_1	Categòrica		**1** Un bany **0** Més d'un bany	NO
BATHROOMS_2	Categòrica		**1** Dos banys **0** Un bany o més de dos banys	NO
BATHROOMS_3	Categòrica		**1** Tres o més banys **0** Menys de tres banys	NO
ELEVATOR	Categòrica		**1** Disposa d'ascensor **0** Sense ascensor	NO
EXTERIOR	Categòrica		**1** Exterior **0** Interior	NO
GROUND_FLOOR	Categòrica		**1** Planta baixa **0** Una altra planta	NO
MIDDLE_FLOORS	Categòrica		**1** Planta intermèdia **0** Una altra planta	NO
ATTIC	Categòrica		**1** Última planta **0** Una altra planta	NO
REFORM	Categòrica		**1** Per reformar **0** No necessita	NO

			reforma	
GOOD_CONDITION	Categòrica		**1** En bones condicions **0** Per reformar o obra nova	NO
NEW	Categòrica		**1** Obra nova **0** No nova edificació	NO
FITTED_WARDROBE	Categòrica		**1** Amb armaris de paret **0** Sense armaris de paret	NO
SPLIT	Categòrica		**1** Amb aire condicionat **0** Sense aire condicionat	NO
TERRACE	Categòrica		**1** Amb terrassa **0** Sense terrassa	NO
GARDEN	Categòrica		**1** Amb jardí **0** Sense jardí	NO
POOL	Categòrica		**1** Amb piscina **0** Sense piscina	NO
STORAGE	Categòrica		**1** Amb traster **0** Sense traster	NO
GARAGE	Categòrica		**1** Amb garatge **0** Sense garatge	NO
GARAGE_INCLUDED	Categòrica		**1** Amb garatge inclòs al preu **0** Sense garatge inclòs al preu	NO

GARAGE_NOT_INCLUDED	Categòrica		**1** Amb opció de compra o lloguer de garatge **0** Sense opció de compra o lloguer de garatge	NO
GARAGE_PRICE	Numèrica	Preu de l'opció de compra o lloguer del garatge		NO

Taula 4.6. Variables que formen la base de dades final

V. ANÁLISI DESCRIPTIVA

Disposem d'una base de dades de València ciutat que consta de 7.281 observacions. Aquest conjunt d'immobles està distribuït en 16 zones o districtes, que alhora es divideixen en 71 subzones o barris.

Les dues zones amb més observacions, acumulant un 25.7% dels immobles, esdevenen Ciutat Vella i L'Eixample. Mentre que Benimaclet, amb únicament 71 habitatges, es posiciona com el districte amb menys pes.

Si baixem al nivell d'agregació per subzona, els tres barris amb més presència són Sant Francesc (Ciutat Vella), El Pla del Remei (L'Eixample) i Benicalap (Benicalap); acumulant el 15.1 % del total de la base. En canvi, no superant cap d'ells les 20 observacions, tenim Fonteta de Sant Lluìs (Quatre Carreres), Ciutat Universitària (El Pla del Real) i Beniferri (Campanar).

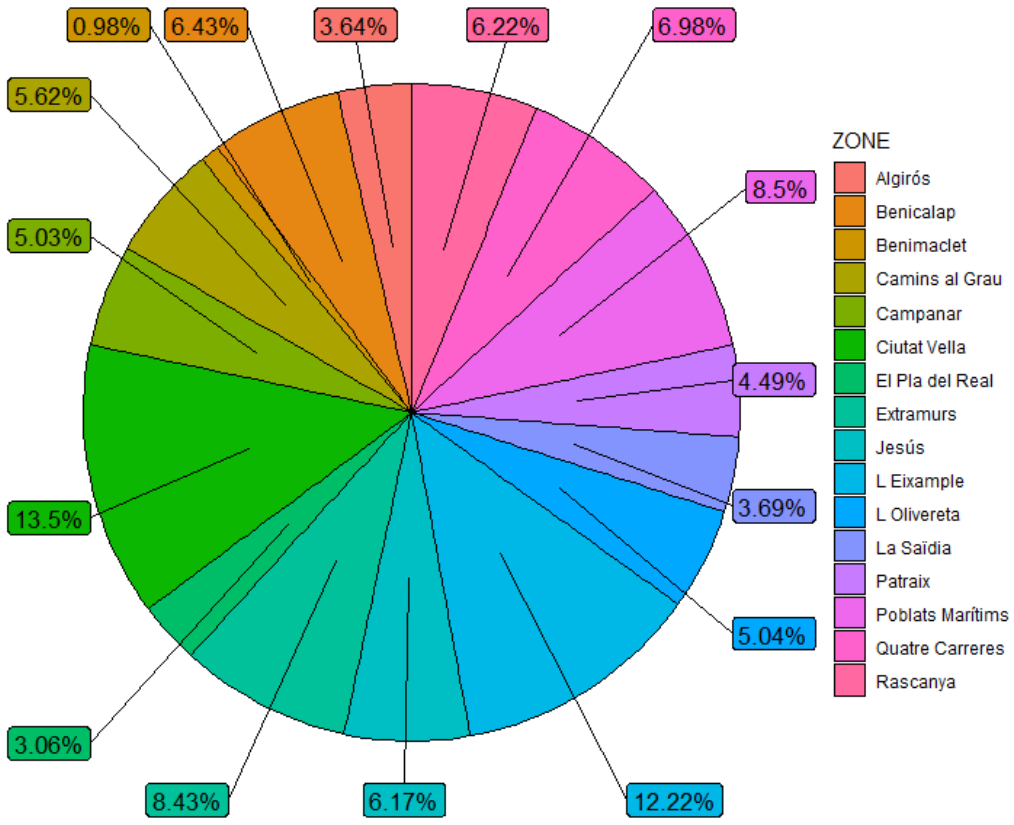


Figura 5.1. *Pie chart* sobre la distribució d'observacions per zona

Per tal de mesurar com de representativa és la distribució d'immobles per districte de la nostra base, ens adrecem a l'Oficina d'Estadística de l'Ajuntament de València. Obtenim dels fitxers que ens faciliten, un per districte amb estadístiques i indicadors de vàries temàtiques actualitzades l'any 2021, les respectives poblacions, superfícies en hectàrees, densitats poblacionals i número d'habitatges censats.

A la següent taula recollim la comparativa del pes de cada districte respecte al total de la nostra base amb la distribució del número d'habitatges per districte a la ciutat de València.

És important subratllar que aquest exercici està supeditat al fet que la nostra base de dades recull immobles en venta i, per tant, que un districte tingui més pes al nostre conjunt de dades que a la realitat, no té perquè suposar una sobrerepresentació de la zona. El número de transaccions immobiliàries, dades que no ens lliuren desagregades per districte, seria una variables més idònia per treballar. No menystenim la comparativa, ja que, tenint clara la limitació, segueix essent una aproximació enriquidora, sobretot pel que fa als valors extrems.

Districte	Població	Superfície (ha.)	Densitat poblacional	Núm. Habitatges	N	PROP (%)	Prop. Habitatges (%)	Dif. Prop.
Ciutat Vella	27.525	169	162,9	19.330	983	13,50	5,02	8,48
L'Eixample	42.853	173,3	247,3	25.195	890	12,22	6,55	5,68
Poblats Marítims	55.760	978,3	57	29.796	619	8,50	7,74	0,76
Extramurs	48.728	197,2	247,1	27.645	614	8,43	7,18	1,25
Quatre Carreres	74.308	1.132,50	65,6	35.854	508	6,98	9,32	-2,34
Benicalap	47.385	221,6	213,8	21.561	468	6,43	5,60	0,82
Rascanya	54.130	262,9	205,9	23.961	453	6,22	6,23	-0,01
Jesús	52.489	298,5	175,8	24.896	449	6,17	6,47	-0,30
Camins al Grau	65.981	236,7	278,8	31.888	409	5,62	8,29	-2,67
L'Olivereta	49.186	198,9	247,3	23.855	367	5,04	6,20	-1,16
Campanar	38.674	523,8	73,8	18.398	366	5,03	4,78	0,25
Patraix	57.833	287,3	201,3	26.895	327	4,49	6,99	-2,50
La Saïdia	47.274	194,4	243,2	24.007	269	3,69	6,24	-2,54
Algirós	36.390	295,9	123	20.446	265	3,64	5,31	-1,67
El Pla del Real	30.667	169,3	181,1	15.666	223	3,06	4,07	-1,01
Benimaclet	28.575	157	182	15.385	71	0,98	4,00	-3,02
València	757.758	5.497	137,9	384.778	7.281			

Taula 5.1. Comparativa amb dades de l'Oficina d'Estadística de l'Ajuntament de València

Podem assumir que tenim una bona mostra respecte als dos districtes amb immobles de la base, Ciutat Vella i L'Eixample, ja que els pesos d'aquestes zones sobre el total són, respectivament, 8.48 i 5.68 punts percentuals superiors als que presenten sobre tot el conjunt de la ciutat. En canvi, tenim cinc districtes amb una diferència negativa superior als dos punts percentuals. Destaquem Benimaclet, com era d'esperar amb únicament 71 observacions, podent assumir que podria suposar un problema en quant a mida mostral si incloem la variable ZONE dins del model i alhora la categoria

corresponent al districte. Aquesta limitació podria mitigar-se agregant Benimaclet a districtes veïns com Rascanya, Algirós o El Pla del Real. Ara bé, aquestes decisions les prendrem quan abordem la modelització, ja que primerament hem d'avaluar si introduïm les variables d'agregació territorial al model i si és així, com les encabirem, perquè afegir-les sense transformació acotaria el nostre model únicament a la ciutat de València.

Per una altra banda, són precisament aquestes hipotètiques transformacions que ens porten a recollir de fonts oficials variables com la població, la superfície i la densitat poblacional de cada divisió territorial, entre altres que veurem més endavant, encarant-les com candidates a substituir la variable ZONE i, si pertoca, SUBZONE. Respecte a aquestes, podem observar com Quatre Carreres esdevé el districte més poblat (9.81% de la població total), el que presenta una àrea més extensa (20.60% del total de la superfície) i alhora el segon menys densament poblat amb 65.6 habitants per hectàrea. Poblats Marítims i Campanar són districtes també perifèrics que tenen unes característiques similars a Quatre Carreres. En contraposició a aquestes zones tenim Camins al Grau, com el districte més densament poblat amb 278.8 ciutadans/ha., juntament amb L'Olivereta que esdevenen dos zones que podríem situar en un segon anell respecte al centre i d'un ús residencial extensiu de la superfície. Veïns a aquests i d'una densitat poblacional similar, però més cèntrics, tenim L'Eixample, Extramurs i La Saïdia. Altrament, separem en aquesta àrea cèntrica Ciutat Vella i El Pla del Real que presenten una densitat i superfície menors, pel que podem deduir que es guanya terreny a l'ús residencial a favor de comerços i altres edificacions de naturalesa més lligada a l'activitat turística.

Quant a la variable SUBZONE, de moment no considerem necessari, accentuat per l'elevat número de barris, baixar a aquest nivell d'agregació com ho hem fet amb els districtes. Abans veurem com es relaciona SUBZONE amb la nostra variable resposta i si detectem barris diferencials, llavors farem el corresponent exercici, sempre amb la idea que no hi ha limitació entre barri i districte en el sentit que, si ho considerem oportú, podem separar un barri del seu districte i elevar-lo a un nivell de ZONE. Precisament per aquest motiu hem anomenat així les variables.

Tot seguit, veurem com es relaciona ZONE i SUBZONE amb la nostra variable resposta, el preu de l'habitatge (PRICE). Presentem la corresponent taula descriptiva sobre el preu del immoble (en milers d'euros, per una millor visualització). Podem observar que L'Eixample, El Pla del Real i Ciutat Vella, districtes cèntrics, es distancien amb claredat en quant a la mitjana i mediana del preu, tot i que presenten una desviació típica

considerablement major que la resta, exceptuant Camins al Grau, accentuada probablement per la presència d'outliers, que podem corroborar ràpidament amb els valors màxims que observem per districte. D'altra banda, per la cua, en termes de mediana, tenim dos districtes veïns: Rascanya i Benicalap. Observem, com era d'esperar, un augment de preu dels districtes en funció a la proximitat al centre de la ciutat.

Respecte als outliers tampoc els abordarem de forma definitiva a l'anàlisi descriptiva, que no significa que no assenyalem la seva presència i/o els excloem puntualment si ho creiem oportú, sinó que ho farem, com comentàvem a l'apartat anterior, quan treballem la modelització.

ZONE	min	max	mean	sd	median	q25	q75
L Eixample	110.00	2952.00	573.66	332.31	500.00	330.00	730.00
El Pla del Real	105.00	2400.00	547.39	352.99	500.00	258.00	672.00
Ciutat Vella	90.00	5029.59	548.67	382.81	450.00	298.75	690.00
Campanar	66.00	2390.00	334.30	259.01	275.00	183.50	400.00
Extramurs	89.30	3500.00	313.42	190.52	275.00	200.00	367.88
Algirós	60.00	927.00	245.39	104.15	229.00	173.00	300.00
Benimaclet	94.50	1150.00	287.43	205.98	225.00	170.00	290.00
Camins al Grau	75.00	2300.00	327.36	338.26	200.00	150.00	350.00
Quatre Carreres	55.60	900.00	222.34	121.78	186.00	147.50	269.62
La Saïdia	59.00	1240.00	205.21	120.51	186.00	140.00	230.00
Poblats Marítims	20.90	2700.00	257.60	270.06	182.00	139.70	267.00
Patraix	68.00	475.00	182.93	74.23	158.00	130.00	226.60
L Olivereta	27.00	1000.00	151.94	88.71	145.00	103.00	174.50
Jesús	49.70	495.00	158.38	72.00	140.00	109.00	185.00
Benicalap	36.80	980.00	181.20	132.07	139.90	92.50	249.81
Rascanya	44.68	1200.00	144.48	115.66	116.21	78.00	162.50

Taula 5.2. Descriptiva de la variable PRICE (en milers d'euros) respecte al factor ZONE

Per acabar de visualitzar el comportament de la nostra variable resposta en funció del districte, representarem ambdues en un boxplot. Ara bé, per tal d'obtenir una imatge més fidedigne, recorrerem a transformar la variable resposta mitjançant la superfície del immoble (M2), amb idea d'aïllar l'efecte d'aquesta darrera variable, passant a

representar el preu del metre quadrat per districte. Afegim que hem tallat l'eix d'ordenades als 6.000 €/m^2, deixant fora 67 immobles.

Novament, podem observar que L'Eixample, Ciutat Vella i El Pla del Real es distancien de la resta presentant un preu del metre quadrat considerablement major. Veiem també que en el cas de L'Eixample, com s'observa també amb claredat a Extramurs, Camins al Grau i, per la cua, a Patraix i Benicalap, s'aprecia una asimetria positiva rellevant, presentant una major dispersió dels preus situats per sobre de la mediana. En contraposició, no apreciem una asimetria esbiaixada a l'esquerra rellevant a cap districte, tot i que sí que s'aprecia una lleugera dispersió major als preus inferiors a la mediana en el cas del Campanar. En general, observem candidats clars a valors atípics superiors a tots els districtes, exceptuant Benimaclet que recordem només aporta 71 observacions a la base, i una dispersió en preus patent. Obviant aquestes limitacions a l'hora d'extreure conclusions, sí que podem considerar que se'ns dibuixa un mapa prou clar que dividiríem en tres agrupacions de districtes de major a menor preu del metre quadrat: el grup del més cèntrics (L'Eixample, Ciutat Vella i El Pla del Real), el conjunt des del Campanar a La Saïdia i cinc districtes més perifèrics que esdevenen Patraix, Jesús i L'Olivereta, veïns entre si i situats al sud-est de la ciutat, i Benicalap i Rascanya, com ja hem comentat abans també veïns, situats al nord de la població.

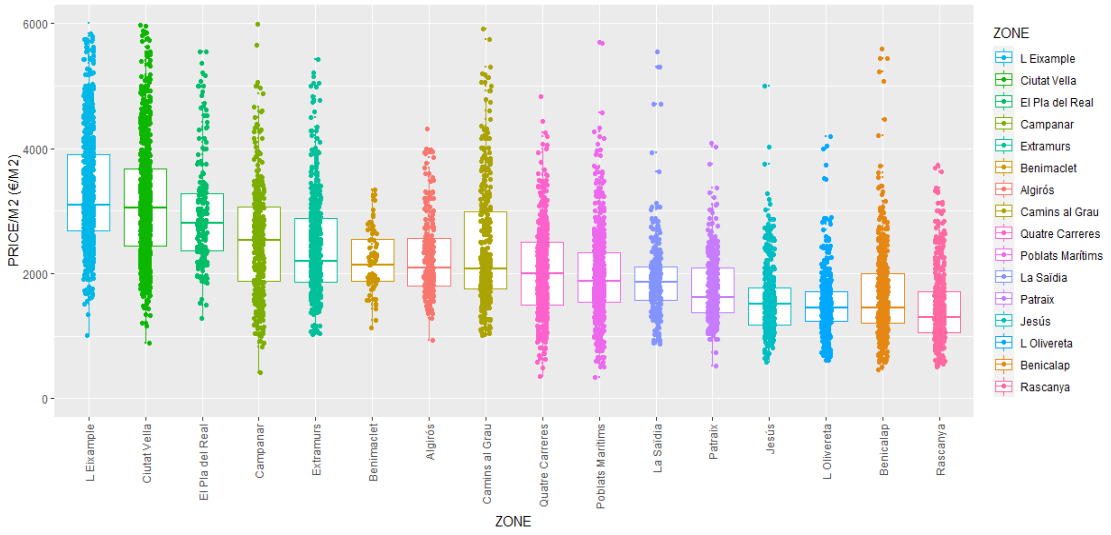


Figura 5.2. Boxplot del preu de l'habitatge (en milers d'euros) per districte

Baixem ara al nivell d'agregació de SUBZONE, mantenint la transformació de la variable resposta a preu del metre quadrat, per trobar que els tres primers barris amb una major mediana i mitjana pertanyen a L'Eixample, Ciutat Vella i Camins al Grau. Sorprèn el cas de Camins al Grau, ja que com a districte es posicionava a una zona mitjana. Això és degut, com ja ens posa en avís tant la dispersió que observem a nivell de districte

com la desviació típica elevada pel que fa al barri, a la presència de valors atípics que posicionen la subzona més amunt d'on hauria d'estar si ometéssim aquests. Destaca també la desviació recollida quant a El Pla del Remei, Sant Francesc i La Xerea; que s'explica i corrobora d'igual forma. És patent, com era d'esperar, que al baixar a nivells de divisió més petits, la influència dels outliers és major i molt més perceptible.

Ara bé, a la sisena posició apareix el barri de La Carrasca, amb 45 immobles que suposen el 16.98% del total d'observacions del districte d'Algirós, que alhora esdevé el setè districte en termes de mediana del preu del immoble. Es tracta d'un cas interessant, ja que posa en relleu l'heterogeneïtat del districte. És a dir, no es tracta de la influència sobre el preu d'unes determinades observacions, sinó que dins de la zona hi ha un barri que es distancia amb claredat dels altres quatre que la conformen, en una mesura suficient com perquè ens puguem qüestionar si ens interessa i és convenient que segueixi formant part d'aquest conjunt. Com ja mencionàvem amb anterioritat, no tenim limitacions respecte a divisions territorials i si en el moment de la modelització, creiem necessari elevar un barri com La Carrasca a nivell de zona o recollir-lo dins d'un altre districte, així procedirem.

SUBZONE	ZONE	n	min	max	mean	sd	median	q25	q75
El Pla del Remei	L Eixample	350	1956.00	2661857.53	12640.62	143510.43	3767.28	3042.45	4636.19
Penya-Roja	Camins al Grau	118	1041.67	8515.28	3562.03	1237.96	3497.92	3000.00	4161.85
Sant Francesc	Ciutat Vella	424	1439.02	3975963.64	12846.85	192925.26	3298.09	2538.00	4106.81
La Xerea	Ciutat Vella	168	1208.33	8273.81	3455.31	1038.39	3274.14	2862.26	3881.58
Gran Via	L Eixample	36	2166.67	4894.74	3363.16	709.90	3229.21	2844.45	3891.30
La Carrasca	Algirós	45	1625.00	3956.04	2830.24	684.31	3203.88	2321.43	3300.00

Taula 5.3. Descriptiva de la variable PRICE/M2 (en €/m^2) respecte al factor SUBZONE, on recollim els 6 primers barris amb una major mediana

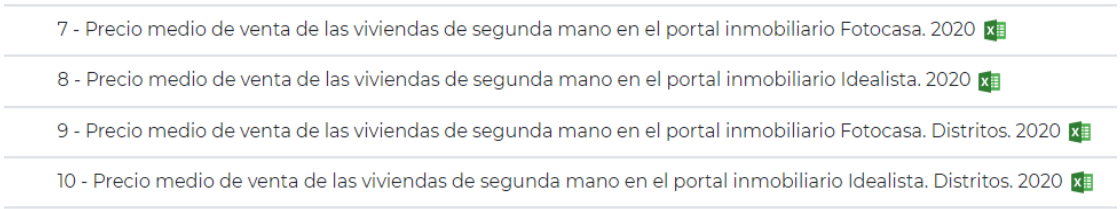
Respecte als barris que presenten un preu del metre quadrat menor en termes de mediana, trobem que es recullen en els districtes de L'Olivereta, Rascanya, Benicalap, Patraix i Poblats Marítims. Aquest últim districte és l'únic dels citats que situàvem en un agregat de districtes d'un grau superior. D'igual forma que detectàvem amb La Carrasca, el barri de Natzaret s'allunya molt dels quatre barris restants que composen Poblats Marítims. El preu mitjà del metre quadrat a Natzaret presenta una diferència de 835.98€ enfront al de El Grau (un 74.56% major), el segon barri amb un preu mitjà més baix del districte.

SUBZONE	ZONE	n	min	max	mean	sd	median	q25	q75
La Fontsanta	L Olivereta	35	614.14	2120.00	1046.75	402.41	863.79	740.26	1249.08
Natzaret	Poblats Marítims	48	597.56	3285.71	1121.08	422.71	1056.90	859.72	1254.11
Els Orriols	Rascanya	167	500.90	10775.86	1166.46	811.71	1077.59	923.00	1253.86
Tres Forques	L Olivereta	72	613.64	3513.51	1251.70	406.05	1216.73	1009.62	1380.85
Ciutat Fallera	Benicalap	57	454.32	3618.75	1258.63	528.35	1242.86	784.72	1642.86
Barrio de Favara	Patraix	27	523.08	2595.24	1349.35	427.01	1314.81	1076.75	1435.29

Taula 5.4. Descriptiva de la variable PRICE/M2 (en €/m^2) respecte al factor SUBZONE, on recollim els 6 últims barris respecte a la mediana

En darrer lloc, tornarem a consultar l'Oficina d'Estadística de l'Ajuntament de València per contrastar amb dades oficials el comportament del preu del immoble envers la zona on es localitza.

Abans d'entrar en detall i abordar les comparatives, voldria remarcar un descobriment tant inesperat com gratificant que és el fet que l'Ajuntament de València posa a disposició del usuari, i per tant com a referència, els preus mitjans de venta dels habitatges de segona mà als portals Fotocasa i Idealista. És més, deixant de banda els valors cadastrals, esdevé l'única font que posa a l'abast del ciutadà; juntament amb una taula dels preus mitjans de venta d'habitatges per trimestres del 2020 elaborada per Grupo Tinsa (una societat de taxació) i el valor de taxació mitjà dels habitatges lliures segons antiguitat, datada al 2021, obtinguda del Ministerio de Transportes, Movilidad y Agenda Urbana.



Imatge 5.1. Taules de preus mitjans dels immobles anunciats a Fotocasa i Idealista (2020) publicats a l'Anuari Estadística de l'Ajuntament de València

Remarco aquest fet pel innegable valor que atorga al nostre projecte, traslluint-se que, com a mínim a ulls d'un òrgan de govern i administració com és l'Ajuntament de

València, els portals immobiliaris són una font de dades de valor pels ciutadans o entitats que ho requereixin. Altrament, no es passa per alt que les dades que publiquen són de fa dos anys i que el nostre projecte té com a meta brindar una metodologia i eines versàtils per analitzar el preu de mercat de l'habitatge amb caràcter més immediat.

Prèviament a contrastar el preu del metre quadrat de la nostra base amb els preus reportats per Grupo Tinsa i els corresponents a Fotocasa i Idealista, delimitarem i exclourem els valors atípics per aconseguir una comparativa més acurada. Per tant, utilitzarem l'amplitud interquartílica (IQR) per obtenir els límits tant inferior, tot i que no aplicarà, com superior que marcaran els outliers. Així doncs, obtenim com a límits, els següents valors: -452.20 i 4965.05 (en €/m^2). Deixem fora 184 observacions, de les quals el 66.30 % pertanyen als districtes de Ciutat Vella (63) i L'Eixample (59). Podríem aplicar també un tall efectiu pels outliers inferiors, però ja havíem vist que la seva presència és mínima. Observem que aplicant l'exclusió dels outliers, obtenim una correcció molt rellevant del preu mitjà del metre quadrat en els districtes de Ciutat Vella (de 7279.68 a 3043.99 €/m^2), L'Eixample (de 6818.07 a 3220.97 €/m^2) i Extramurs (de 6306.25 a 2375.07 €/m^2). En vistes d'això, prosseguirem l'anàlisi descriptiva per la resta de variables deixant fora aquests 184 immobles.

A la següent taula que presentem, posem en contrast el nostre preu del metre quadrat envers els tres preus obtinguts de l'Ajuntament de València (en €/m^2). A més, també ens faciliten i recopilem la renda neta mitjana anual per llar, variable que s'observa a cop d'ull que esdevé clarament explicativa de la nostra variable resposta en funció de la ubicació de l'habitatge.

Districte	PRICE/M2	Preu/m^2 - Grupo Tinsa	Preu/m^2 - Idealista	Preu/m^2 - Fotocasa	Renda neta mitjana anual per llar (en €)
L Eixample	3.221	2.445	2.819	3.109	45.673
Ciutat Vella	3.044	2.283	2.822	3.091	41.976
El Pla del Real	2.850	2.119	2.490	2.800	52.495
Campanar	2.497	1.594	2.108	2.270	37.717
Extramurs	2.375	1.848	2.133	2.196	38.557
Camins al Grau	2.311	1.606	1.836	2.170	33.336
Algirós	2.276	1.569	1.948	2.053	34.742
Benimaclet	2.201	1.506	1.835	1.919	34.135
Quatre Carreres	2.044	1.431	1.625	1.858	29.375
Poblats Marítims	1.960	1.449	1.742	1.896	27.537
La Saïdia	1.886	1.399	1.597	1.605	29.770
Patraix	1.775	1.403	1.445	1.532	31.520
Benicalap	1.614	1.319	1.396	1.775	28.162
Jesús	1.559	1.209	1.338	1.466	28.057
L Olivereta	1.529	1.178	1.273	1.386	26.469
Rascanya	1.462	1.066	1.182	1.370	28.688
València	2.263	1.439	1.824	2.062	32.954

Taula 5.5. Comparativa del preu del metre quadrat per districte amb dades de l'Oficina d'Estadística de l'Ajuntament de València

Podem observar que els preus de la nostra base són majors que els preus obtinguts de les altres tres fonts. Això és degut, per un costat, a que els nostres preus daten del 2022, mentre que els altres estan actualitzats a 2021 i 2020; pel que s'hauria d'aplicar una correcció, que aniria a l'alça, mitjançant l'índex de referència de preus de venda de l'habitatge. Per una altra banda, les diferències de preu són més rellevants respecte a l'estudi de mercat de Grupo Tinsa. Aquestes vindrien explicades, en part, pel fet, que no hem de perdre de vista, que la nostra, com les de Idealista i Fotocasa, estan alimentades per anuncis, aquests com és lògic presentaran un biaix alcista respecte als preus finals resultants de la materialització de la transacció. A l'estudi de mercat de Grupo Tinsa és d'esperar que s'estiguin aplicant correccions en aquest sentit. Per la nostra banda, únicament hem aplicat un ajust eliminant els valors atípics de la nostra base, mentre que desconeixem quins ajustos s'estan aplicant a les altres tres fonts. Per

acabar, les bases de Idealista i Fotocasa són sobre immobles de segona mà, mentre que nosaltres recollim també aquells d'obra nova que, a priori, presentaran un preu de venda major. Tot i així, aquest efecte ha d'ésser molt baix, perquè el pes dels habitatges de nova construcció a la nostra base és només d'un 3.03%.

Ara bé, no ens interessa tant la comparativa a nivell de valors, sinó el contrast entre la ordinalitat dels districte respecte al preu entre bases. Per consegüent, transformarem els valors observats per la posició que ocupa el districte en el conjunt ordenant de major a menor respecte el preu mitjà del metre quadrat. En un escenari ideal, la suma de diferències en valor absolut entre posicions del districte en cada font respecte a la nostra base resultaria nul·la. En canvi, l'extrem oposat donaria quan l'ordre de les posicions fos exactament el invers, que al tenir 16 districtes suposaria una diferència absoluta de 128. Si ho apliquem a les tres fonts, observem que per l'estudi de Grupo Tinsa tenim 10 punts de diferència absolut i 6 punts tant pels preus de Idealista com Fotocasa. Si dividim entre els 128 punts de l'escenari de discordança completa, obtenim una mètrica, que es mou en el rang de 0 a 1, que descriuria la correspondència d'ordinalitat entre bases. Per tant, observem que hi ha poca disparitat, presentant Grupo Tinsa un valor de 0.078 i 0.045 pels portals immobiliaris.

Finalment, i de la mà del que comentàvem amb anterioritat de cercar una variable que de cara a la modelització poguéssim emprar en substitució de ZONE i SUBZONE, tenim la renda neta mitjana anual per llar. Si calculem la correlació entre la renda i el nostre preu mitjà del metre quadrat per districte, obtenim un coeficient de correlació de Pearson igual a 0.889. Així que la renda mitjana d'una zona sembla posicionar-se com una candidata idònia. Observem que el districte amb una renda mitjana més elevada esdevé El Pla del Real amb 52495€, que és el nostre tercer districte en quant a preu mitjà del metre quadrat, i per la cua tenim L'Olivereta amb una renda mitjana que baixa aproximadament a la meitat de la d'una família de El Pla del Real.

Deixem de banda les variables de localització per abordar la categòrica corresponent al tipus d'habitatge (TYPE). Observem que tenim 9 nivells que descriuen la tipologia de l'immoble, però amb una presència a la base molt desigual. Mentre que el 87.68% de les observacions pertanyen al nivell "Piso", tenim "Casa de pueblo", "Casa terrera" i "Finca rústica" que sumen en conjunt únicament 22 immobles (el 0.31% del total). Com era d'esperar en quant pensem en la distribució habitual de tipologies d'immobles d'una ciutat, on imperarà amb força la construcció vertical a l'horitzontal.

TYPE	n	pes (%)
Piso	6223	87.6849373
Ático	361	5.0866563
Dúplex	163	2.2967451
Chalet	152	2.1417500
Casa o chalet independiente	105	1.4794984
Estudio	71	1.0004227
Casa de pueblo	15	0.2113569
Casa terrera	4	0.0563618
Finca rústica	3	0.0422714

Taula 5.6. Número d'observacions i pes dels nivells de la variable TYPE

Davant d'això, i descartant, tot i que seria una opció vàlida, eliminar les observacions d'aquests últims tres nivells, refactorizem agrupant els immobles dins de la categoria de "Casa o chalet independiente".

Tampoc hem de passar per alt que les tipologies descrites com a tal no són independents entre si, en el sentit que un estudi, àtic i dúplex no deixen d'ésser pisos i la resta de nivells, cases. Ara bé, aquesta possibilitat de reduir a dues tipologies, construcció vertical o pis que constituiria el 96.07% de la base i construcció horitzontal o casa amb un pes 3.93%, no l'aplicarem a la descriptiva, sinó que l'avaluarem quan modelitzem. Per una altra banda, Poblats Marítims (121), Benicalap (52) i Rascanya (36) aglutinen en conjunt el 74.91% de les cases de la base. En sintonia amb això, tampoc plantejarem aquí si, en vistes del poc pes i la presència focalitzada de la tipologia casa, l'excloem i ens centrem en modelar el preu de venda d'immobles tipus pis.

Respecte al comportament de la nostra variable resposta en funció de la tipologia de l'habitatge, novament utilitzarem la transformació a preu del metre quadrat, per evitar el component de superfície construïda, per descriure'l. De igual forma que hem fet amb la correcció dels valors atípics, mantindrem el preu del metre quadrat com a variable resposta a l'anàlisi descriptiva de la resta de variables.

TYPE	n	min	max	mean	sd	median	q25	q75
Ático	361	773.20	4959.68	3151.82	865.89	3092.78	2562.19	3764.15
Dúplex	163	660.82	4932.74	2739.71	875.03	2724.14	2250.00	3117.05
Estudio	71	620.92	4058.82	2507.43	863.44	2661.29	1758.04	3072.55
Piso	6223	342.62	4964.29	2212.94	891.24	2050.56	1549.72	2773.72
Chalet	152	359.12	4605.26	1922.81	821.19	1670.68	1295.65	2534.98
Casa o chalet independiente	127	486.73	4576.27	1863.59	883.04	1647.42	1260.87	2338.89

Taula 5.7. Descriptiva de la variable PRICE/M2 (en €/m^2) respecte al factor TYPE

Podem observar que la tipologia "Ático" és la que presenta una mediana més elevada ascendint a 3092.78 €/m^2. En contraposició, tenim els dos nivells corresponents a la construcció horitzontal, "Chalet" i "Casa o chalet independiente", amb les medianes més baixes, 1670.68 i 1647.42 €/m^2, respectivament. Això no resulta tant sorprenent si tornem a la distribució per zona i refresquem que els districtes que apleguen casi tres quarts del total de cases són també els districtes amb un preu mitjà del metre quadrat més baixos (Rascanya, Benicalap i, en una posició més alta, Poblats Marítims). No hem de perdre la vista la limitació que suposa el baix número d'observacions, exceptuant el nivell "Piso", a l'hora d'extreure conclusions robustes.

Per una altra banda, observem que la desviació típica és similar entre categories. Per una millor visibilitat de la dispersió, representarem en un boxplot la variable PRICE/M2 respecte a TYPE.

Es pot observar ràpidament que "Chalet" i "Casa o chalet independiente" presenten asimetria positiva, concentrant-se la major part de les observacions a la part inferior de la distribució, és a dir, preus del metre quadrat més baixos. En canvi, apreciem pel que fa al nivell "Estudio", asimetria esbiaixada a l'esquerra, essent la mediana més alta que la mitjana, i, per tant, amb una concentració major a la part superior de la distribució, preus del metre quadrat més alts.

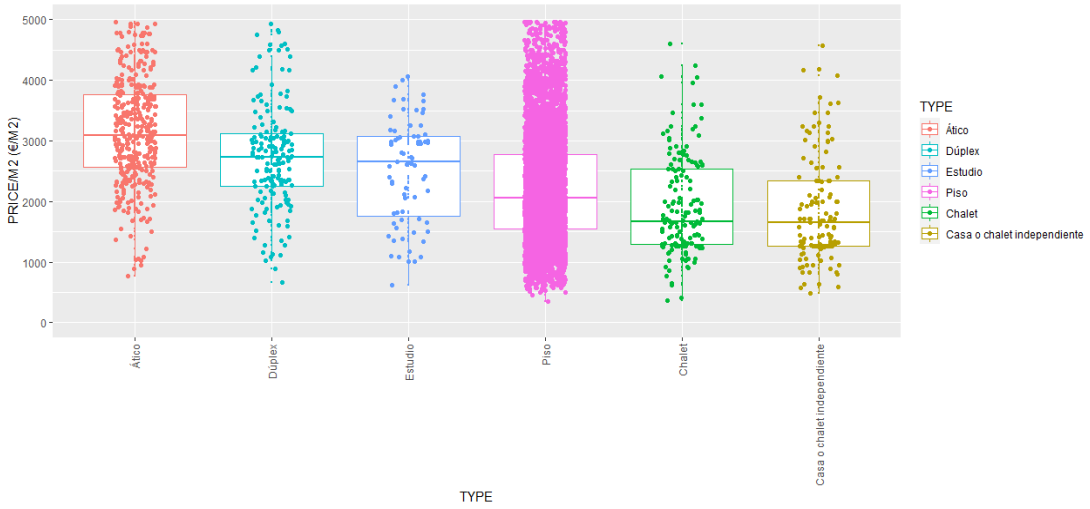


Figura 5.3. Boxplot del preu del metre quadrat (en €/m^2) per tipologia d'habitatge

Abans de prosseguir amb la resta de variables, la majoria d'elles qualitatives, abordarem tant la nostra variable resposta (PRICE) com la que se'ns posiciona, a priori, com més explicativa, la superfície del immoble (M2).

En primer lloc, passem a centrar-nos en la superfície, que recordem no es tracta de la superfície útil, sinó de la total. Afegim una breu taula descriptiva de la variable.

min	max	mean	sd	median	q25	q75
28	895	128.04	70.94	107	83	150

Taula 5.8. Descriptiva de la variable M2 (en m^2)

La superfície mitjana d'un immoble de la nostra base es xifra en 128.04 m^2, amb una desviació típica de 70.94 m^2. Tornem a consultar l'Anuari Estadístic publicat per l'Ajuntament de València, per trobar que, segons dades cadastrals (2021), la superfície construïda mitjana, per immobles amb ús residencial posteriors a 1800, de la ciutat esdevé 109.6 m^2 (el càlcul cadastral està acotat a immobles amb una superfície màxima de 700 m^2, tot i que nosaltres únicament tenim 5 habitatges que superin aquest llindar). La diferència novament pot venir explicada en part, com comentàvem pel que fa als preus, per un biaix alcista per part de l'anunciant a l'hora de publicar el immoble.

Per tal de fer-nos una idea més clara de la distribució de la nostra variable, passarem a representar-la en un histograma.

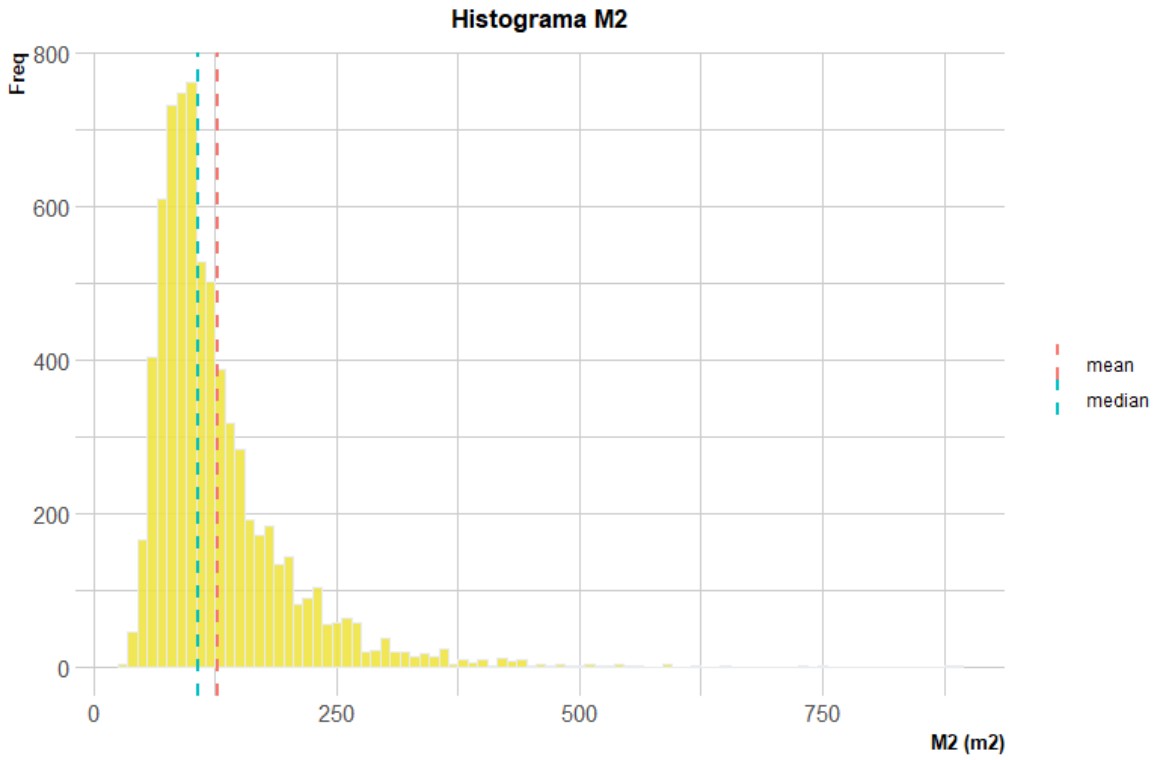


Figura 5.4. Histograma de la superfície total del immoble (en m^2)

Podem observar, com era d'esperar de la naturalesa de la variable, una clara asimetria a la dreta o positivament esbiaixada i curtosis positiva (leptocúrtica), concentrant-se les observacions al voltant de la mitjana. Per una altra banda, veiem observacions que respecte únicament la superfície són clarament candidates a outliers. Si utilitzem novament l'amplitud interquartílica (IQR) per obtenir els límits que marcaran els valors atípics, obtenim que aquests es posicionen en els valors: -17.5 i 250.5 m^2. Per tant, tindríem un total de 430 immobles que excedirien el límit superior, la majoria situats, com era d'esperar en ésser zones que, com hem dibuixat al llarg de l'anàlisi, podríem considerar com més exclusives, en el districte de Ciutat Vella (123) i L'Eixample (112), però destacant també Poblats Marítims amb 63 immobles. Ara bé, ja hem tractat els outliers tenint en compte el preu mitjançant la variable preu del metre quadrat. Per tant, no considerarem cap tractament respecte M2 i no exclourem les 430 observacions subratllades.

Així mateix, tot i que la representació del histograma ens indica que no és així, passarem a contrastar si la variable superfície s'ajusta a una normal. Ho farem representant el corresponent Q-Q plot.

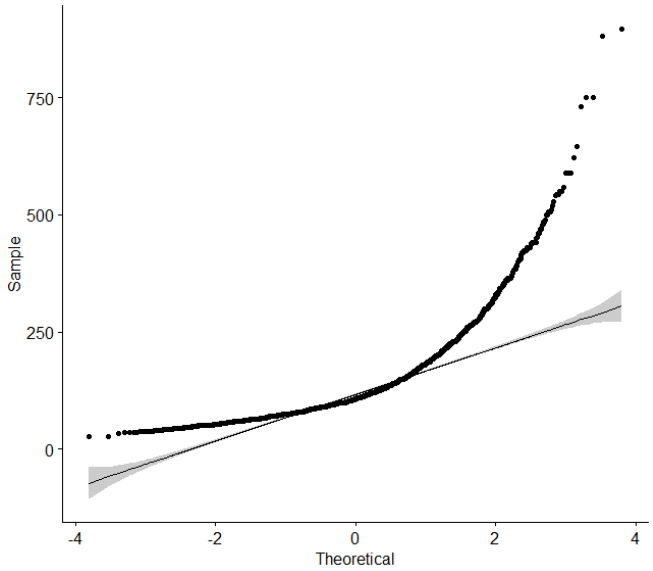


Figura 5.5. Q-Q plot de la variable superfície (en m^2)

Podem observar amb claredat com les nostres dades no presenten normalitat, punt que certificarem aplicant el test de normalitat de Shapiro-Wilk amb una mostra de 5000 observacions. Obtenint el següent output:

```
        Shapiro-Wilk normality test

data:  p2$M2[index]
W = 0.80172, p-value < 2.2e-16
```

Per tant, donant-se un p-valor molt més petit que 0.05, podem refusar que la superfície es distribueixi segons una normal. Per una altra banda, és interessant fer un cop d'ull a la correlació entre preu de l'immoble i la seva superfície construïda. Observem, com era d'esperar, que esdevé alta, ascendint el coeficient de correlació de Pearson a 0.8112.

Tot seguit, agregarem per districte per completar l'anàlisi que hem realitzat del factor localització del immoble i contrastar amb més detall les nostres dades amb fonts

oficials. Per una millor visualització, representarem ambdues variables en un boxplot, tallant l'eix d'ordenades als 600 m^2 i deixant fora 7 immobles.

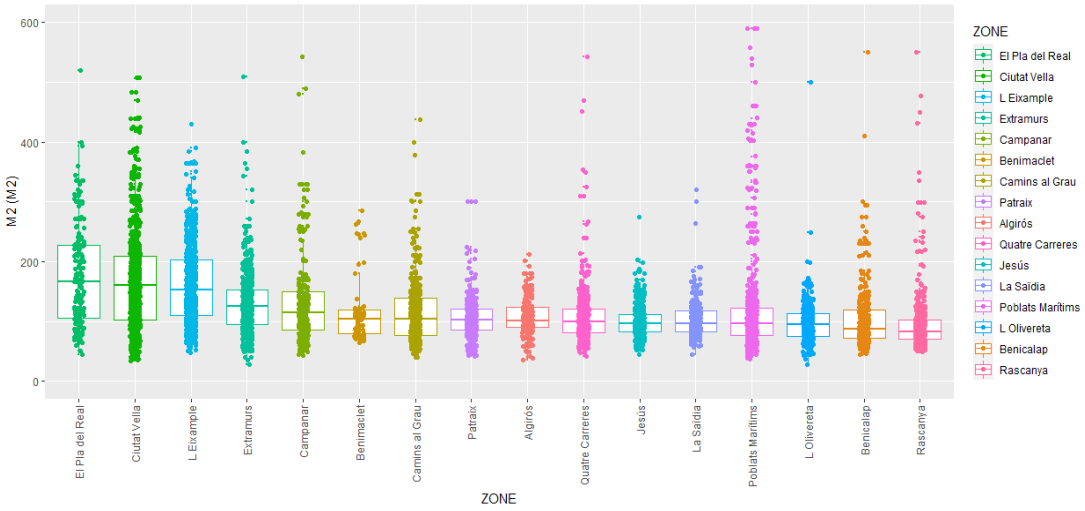


Figura 5.6. Boxplot de la superfície construïda (en m^2) per districte

Podem observar una dispersió, com era esperable després de desgranar els outliers, considerablement més elevada al districtes de Poblats Marítims i, en un segona instància, els tres primers districtes amb una superfície mitjana més alta i destacant també Benimaclet, tot i que recordem que és una zona amb molt poques observacions. En canvi, els districtes de Jesús, Algirós i L'Olivereta són els que presenten una dispersió menor.

Respecte a la mateixa figura 5.2, però en funció del preu del immoble i la taula amb els preus del metre quadrat, veiem que la posició entre districtes balla, però es manté el comportament que ja havíem observat, sobretot en els extrems. Així doncs, tenim El Pla del Real, Ciutat Vella i L'Eixample que es distancien de la resta dibuixant-se com zones amb habitatges més espaiosos i més privatives en quant a preu, alhora esdevenint els districtes amb una renda neta mitjana anual per llar major. En contraposició, tenim zones com Rascanya, Benicalap i L'Olivereta amb immobles més petits i preus més assequibles, acompanyant el fet que les rendes en aquests districtes són les més baixes de la ciutat.

Afegim una taula amb les superfícies mitjanes per districte de la nostra base en contrast amb les cadastrals publicades per l'Ajuntament de València (2021).

Districte	M2	Superfície construïda mitjana (m^2)	Dif. (%)
El Pla del Real	173,61	131,95	31,57
Ciutat Vella	169,21	123,2	37,35
L Eixample	164,52	127,13	29,41
Extramurs	130,91	115,94	12,91
Campanar	127,93	112,04	14,18
Benimaclet	128,39	106,8	20,22
Camins al Grau	116,7	105,3	10,83
Patraix	105,96	109,25	-3,01
Algirós	107,19	109,31	-1,94
Quatre Carreres	111,8	107,06	4,43
La Saïdia	103,8	100,18	3,61
Jesús	101,24	101,79	-0,54
Poblats Marítims	127,11	97,23	30,73
L Olivereta	96,98	96,72	0,27
Benicalap	108,11	101,53	6,48
Rascanya	96,2	99,18	-3,00
València	128,04	109,55	16,88

Taula 5.9. Comparativa de la superfície total (en m^2) per districte amb dades cadastrals publicades per l'Oficina d'Estadística de l'Ajuntament de València

Podem observar que les diferències són majors en els districtes que comentàvem se situen en la part superior en quant a preu del metre quadrat, exceptuant Poblats Marítims, que recordem presenta una dispersió i presència de valors atípics rellevants. En canvi, en els districtes amb una superfície i preu més baix, les dimensions mitjanes dels immobles s'ajusten més amb les reportades per l'Ajuntament de València.

Tot seguit, passem a abordar de forma univariant la nostra variable resposta: el preu del immoble (PRICE). Primerament, per tal de fer-nos una idea de la distribució de la variable, representarem la seva funció de densitat. Per una millor visualització convertim el preu a milers d'euros.

Podem observar que la distribució del preu s'assimila considerablement amb la que dibuixàvem per la superfície, com esperàvem a l'ésser variables d'una naturalesa molt similar. De igual forma, visualitzem per la nostra variable resposta una accentuada asimetria positivament esbiaixada i curtosis positiva.

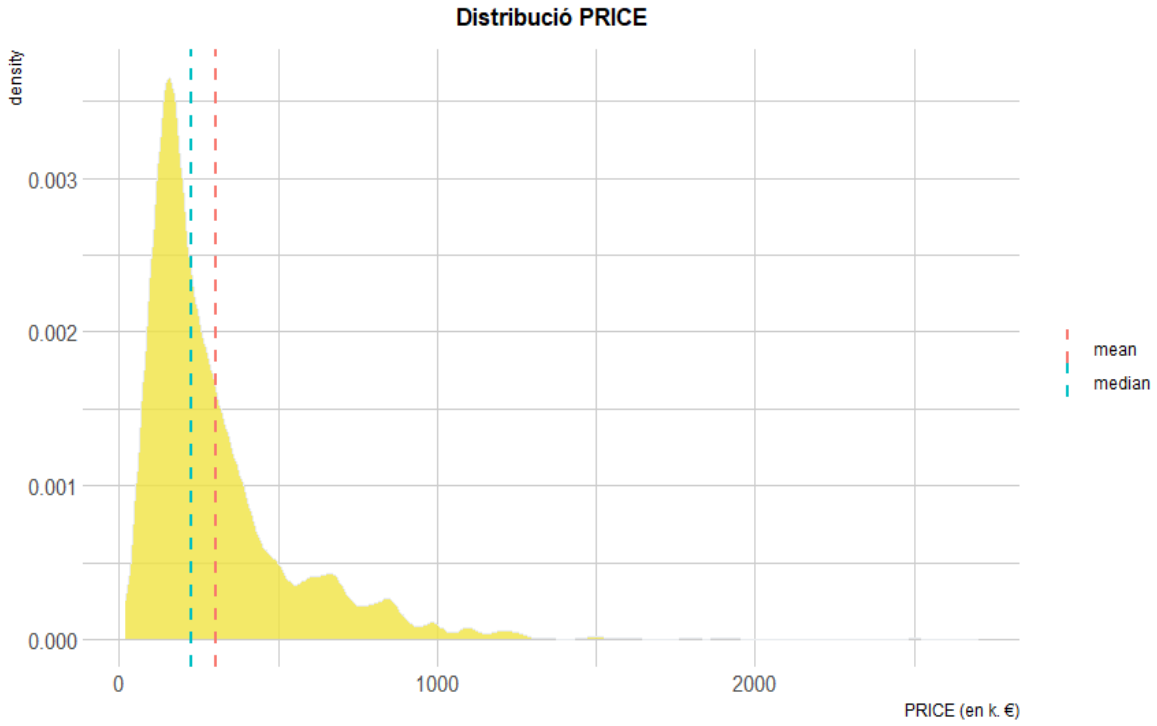


Figura 5.7. Funció de densitat del preu del immoble (en milers d'euros)

Com hem procedit amb la superfície, identificarem, mitjançant l'amplitud interquartílica (IQR), els candidats a valors atípics respecte al preu del immoble. Els llindars obtinguts, en milers d'euros, són: -190 i 714. El límit superior exclouria un total de 510 immobles. El 74.71% d'aquests outliers s'ubiquen als districtes de L'Eixample (200) i Ciutat Vella (181). Tampoc considerarem eliminar aquestes observacions de la base pel mateix motiu, que ja hem tractat els outliers sobre la variable preu del metre quadrat, que per aquestes observacions presenta un preu mitjà de 3586 €/m^2, que esdevé alt, però tampoc descabellat.

Tanmateix, de igual forma que hem procedit amb la superfície i tot i veure clarament, a la representació de la funció de densitat, que els nostres preus no presenten normalitat, representarem el corresponent Q-Q plot.

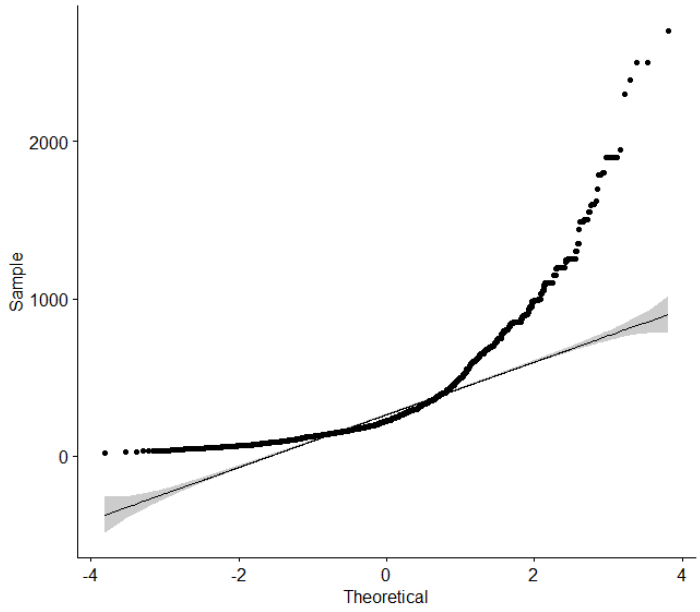


Figura 5.8. Q-Q plot del preu del immoble (en milers d'euros)

Novament, observem amb evidència que la nostra variable resposta no s'ajusta a una normal. Si apliquem el test de normalitat de Shapiro-Wilk amb una mostra de 5000 immobles, obtenim un p-valor molt menor a 0.05, portant-nos a refusar la hipòtesis nul·la que les nostres dades presenten normalitat.

```
        Shapiro-Wilk normality test

data:  p2$PRICE[index]
W = 0.77448, p-value < 2.2e-16
```

Una possible transformació, que ja hem aplicat al llarg de l'anàlisi, és convertir la nostra variable resposta en el preu del metre quadrat. Amb la idea de que la variable resultant podria ajustar-se a una normal, passem a representar la seva funció de densitat. També hi superposem, en verd, la corba de la normal que correspondria per la mitjana i desviació típica de les observacions.

Podem observar com la distribució sembla aproximar-se més a una normal que les de la superfície i preu, malgrat això, tampoc visualitzem que s'ajusti com hauria, fent-nos de guia la corba de la normal representada. Així mateix, advertim asimetria a la dreta i curtosis positiva, tot i que menys accentuades que en el cas de les variables per individual.

Figura 5.9. Funció de densitat del preu del metre quadra (en €/m^2) amb la corba de la normal corresponent (en verd)

No cal que representem el respectiu Q-Q plot, utilitzarem únicament el test de Shapiro-Wilk, amb una mostra de 5000 observacions, per corroborar que tampoc el preu del metre quadrat segueix una distribució normal.

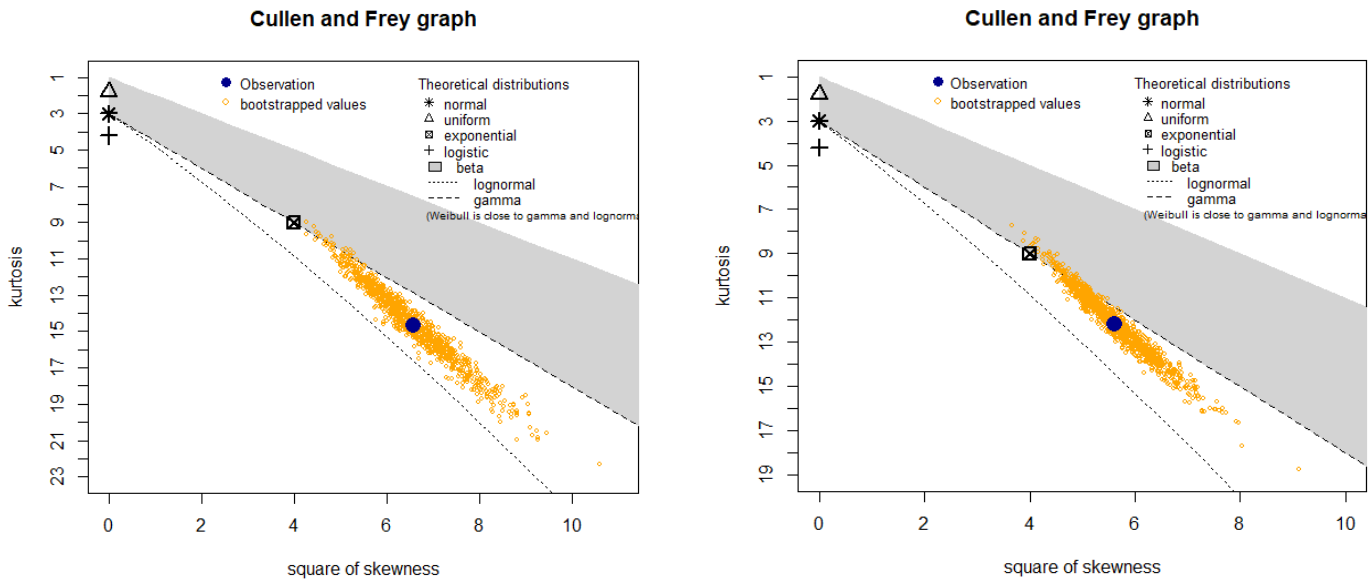
```
        Shapiro-Wilk normality test

data:  (p2$PRICE/p2$M2)[index]
W = 0.96575, p-value < 2.2e-16
```

També en aquesta ocasió, obtenim un p-valor molt inferior a 0.05. Per tant, podem refusar que la variable es distribueixi segons una normal.

Seguidament, en vistes de que no tenim una idea clara sobre les distribucions de les variables superfície, preu i preu del metre quadrat, recorrerem al gràfic de Cullen and Frey, que construirem sobre mostres de 1000 observacions. Com podem percebre, aquest gràfic d'asimetria-curtosis no ens permet dilucidar una aproximació d'ambdues variables a cap distribució concreta. Per una altra banda, podem discernir amb facilitat

l'asimetria i curtosis elevades que caracteritzen ambdues distribucions i que ja havíem apuntat sobre els histogrames d'aquestes.



Figures 5.10. Gràfics Cullen and Frey de les variables M2 i PRICE, d'esquerra a dreta, respectivament

Representem la visualització també per la variable preu del metre quadrat i observem, com era d'esperar amb els anàlisis previs que hem fet, que en aquest cas s'aproxima més a la distribució normal (menor asimetria i curtosis). Podem entreveure que una beta podria posicionar-se com una candidata de garanties.

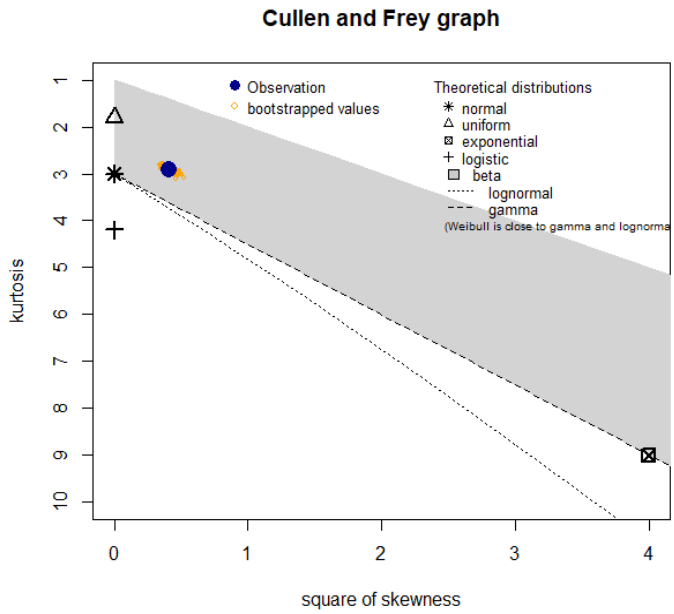
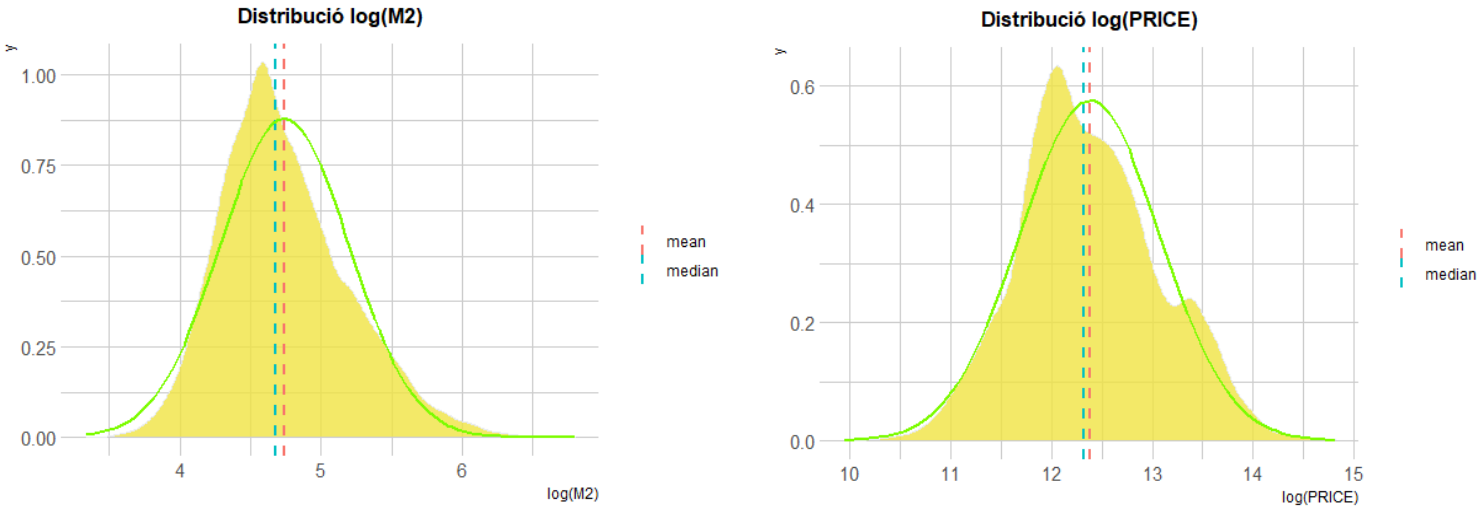


Figura 5.11. Gràfic Cullen and Frey de la variable PRICE/M2

Fonamentant-nos en les visualitzacions que hem anat generant i amb una imatge més clara del comportament de les variables superfície i preu del immoble, se'ns traça com a proposta per garantir la normalitat de les nostres dades, aplicar una transformació logarítmica sobre aquestes.

Així doncs, apliquem la transformació logarítmica sobre ambdues variables i passem a representar les seves funcions de densitat, afegint la corba de la normal (en verd).



Figures 5.12. Funcions de densitat dels logaritmes de les variables M2 i PRICE, d'esquerra a dreta, respectivament

Podem observar com aplicant logaritmes, les nostres variables s'ajusten molt més a una normal. És més que probable que aquesta transformació jugui un paper fonamental per aconseguir un millor ajust quan abordem la modelització.

Després d'explorar la nostra variable resposta junt amb les variables de localització, la tipologia i la superfície, passem ara a prosseguir l'anàlisi de les variables que ens resten. D'una banda, tenim la variable discreta referent al número d'habitacions (HAB). Com ja havíem comentat a l'apartat del *data cleaning*, hi ha una presència molt baixa de missings (80 observacions, que suposen el 1.13% del total) i el rang de valors que pren la variable s'estén des de 1 a 11 habitacions.

Si baixem a la tipologia dels 80 immobles sense informació del número d'habitacions, observarem que 71 d'aquests figuren com "Estudio", que alhora esdevenen la totalitat del estudis de la nostra base. D'aquí i de la definició del concepte d'estudi (tipus d'habitatges on totes les estances, exceptuant el bany, es troben en un mateix espai), veiem que no es tracta de vertaderes dades mancants. Així doncs, optarem per assignar-los-hi un valor i decantant-nos per indicar una habitació i així evitant crear un valor més corresponent a zero habitacions, que alhora seria el valor de la totalitat d'una tipologia d'immoble. Els 9 habitatges restants sí que els considerarem com a missings.

D'altra banda, al tractar-se d'una variable la naturalesa de la qual podríem anomenar de dimensió, com era esperable, observem que la correlació amb la superfície és positiva i ascendeix a 0.6020. Així mateix, també ho és respecte al preu del immoble, tot i que lleugerament menor amb un valor de 0.4728.

Tot seguit, transformem la variable numèrica factoritzant-la, per tal de poder fer-nos una idea dels preus del habitatge en funció del número d'habitacions de que disposa. En línia amb això, presentem la següent taula resum:

HAB	n	min	max	mean	sd	median	q25	q75
9	6	570.0	1800	1085.83	484.61	997.5	711.25	1400.00
10	14	140.0	1100	731.21	214.30	786.5	650.00	815.00
8	10	230.0	1500	825.30	343.79	774.5	674.25	898.75
6	159	157.0	1700	733.65	277.82	700.0	525.00	900.00
7	25	299.5	1600	663.50	344.99	580.0	395.00	850.00
5	562	54.4	2500	586.30	349.05	550.0	325.00	737.50
11	1	410.0	410	410.00	NA	410.0	410.00	410.00
4	1635	48.0	2700	366.16	251.03	295.0	185.00	475.00
NA	9	130.0	980	280.56	266.05	195.0	165.00	210.00
3	3028	30.6	1600	242.92	176.97	189.0	132.00	299.00
2	1127	27.0	1590	209.96	127.36	180.0	130.00	258.90
1	521	20.9	1490	202.17	114.30	171.0	135.00	235.00

Taula 5.10. Descriptiva de la variable PRICE (en milers d'euros) respecte al factor HAB

Podem observar amb claredat un augment de la mediana i mitjana del preu del immoble a mesura que s'incrementa el nombre d'habitacions que el conformen. Ara bé, allò realment interessant que es dedueix de la taula esdevé el poc pes en observacions que presenten els nivells superiors de la variable. Tenim que els habitatges amb més de 6 habitacions (que concentren 5 nivells del factor) representen únicament el 0.79% (56 observacions) dels immobles de la base. Això ens obre la possibilitat de refactorizar la variable agregant en nivells més amples. Previ a aplicar contrast algun, recorrem a Idealista i veiem que la divisió que escullen pren els següents valors: 1, 2, 3 i 4 o més habitacions (també agreguen 0 habitacions pels estudis, a diferència de nosaltres que els hem fet constar com 1 habitació). Com hem fet amb anterioritat, no és una opció sense fonament utilitzar les agregacions que marca el portal, ja que aquest serveix al consumidor, per tant, al nostre parer seria un error obviar-les i restaria consistència al nostre compromís amb la idea de treballar el mercat amb una base obtinguda d'un portal immobiliari. Tanmateix, postergarem l'avaluació i corresponent decisió quan abordem la modelització.

Cal afegir que no és casual que haguem escollit la nostra variable resposta, en comptes del preu del metre quadrat com veníem fent, per resumir-la en funció del número

d'habitacions. Amb el preu del metre quadrat no veuríem una tendència tant clara de l'augment del preu amb el increment del nombre d'estances. Això no deixa de posar-nos en clar la correlació i interacció existent entre superfície i habitacions, ambdues, com dèiem, variables que consideraríem de dimensió.

Continuem l'anàlisi amb una altra variable, també discreta com l'anterior, que esdevé la planta on s'ubica el immoble (FLOOR). Pel que fa a aquesta, la presència de dades mancants, amb 797 observacions, s'eleva a un 11.23% del total de la base. D'altra banda, el rang de valors que pren abasta des d'una planta -2 a una 29. Si, novament, baixem a la tipologia d'aquests 797 missings, podrem observar que 230 es categoritzen com "Casa o chalet independiente" o "Chalet". Conseqüentment, els hi assignarem com a planta el valor 0 (ja havíem procedit en aquest sentit al procés de *data cleaning* al transformar els valors "Bajo" a 0). Així doncs, reduïm a 567 missings (7.99% del total). En aquest cas, la correlació amb el preu de l'habitatge ja no és tant evident i, tot i que positiva, és baixa amb un valor del coeficient de correlació lineal de Pearson de 0.190.

Passem ara a repetir l'exercici efectuat amb el número d'habitacions, factoritzant la variable i agregant una taula resum on en aquesta ocasió tindrem el preu del metre quadrat del immoble en funció de la planta on s'ubica.

FLOOR	n	min	max	mean	sd	median	q25	q75
29	1	4887.53	4887.53	4887.53	NA	4887.53	4887.53	4887.53
27	1	4595.38	4595.38	4595.38	NA	4595.38	4595.38	4595.38
21	1	4386.79	4386.79	4386.79	NA	4386.79	4386.79	4386.79
25	1	3941.18	3941.18	3941.18	NA	3941.18	3941.18	3941.18
23	1	3788.46	3788.46	3788.46	NA	3788.46	3788.46	3788.46
24	1	3651.85	3651.85	3651.85	NA	3651.85	3651.85	3651.85
13	23	2028.69	4782.61	3288.36	773.52	3421.05	2531.65	3797.01
18	30	2437.50	3956.04	3263.63	281.33	3203.88	3203.88	3300.00
12	28	2314.81	4932.74	3358.49	763.80	3184.90	2788.00	4003.65
19	1	3177.97	3177.97	3177.97	NA	3177.97	3177.97	3177.97
9	88	1064.37	4313.10	2906.42	878.74	3000.00	2009.15	3705.47
15	4	2460.00	3921.57	3081.69	706.35	2972.59	2517.06	3537.21
10	73	790.57	4722.22	2829.71	753.79	2857.14	2287.23	3291.67
8	148	1059.11	4819.28	2950.32	875.08	2856.11	2380.95	3640.88
14	12	1413.04	4302.33	2839.44	869.28	2852.29	2267.58	3349.98
11	42	1491.94	4366.20	2908.07	710.44	2766.04	2452.96	3395.49

20	2	1569.23	3840.00	2704.62	1605.68	2704.62	2136.92	3272.31
17	16	2285.71	4322.37	2857.97	471.03	2690.09	2644.63	2942.58
16	1	2428.57	2428.57	2428.57	NA	2428.57	2428.57	2428.57
7	314	833.33	4959.68	2510.54	900.42	2335.33	1799.00	3183.99
6	420	755.21	4964.29	2410.91	873.85	2239.24	1714.29	2916.64
NA	567	636.36	4932.74	2271.64	977.61	2212.96	1493.90	2825.70
-2	1	2077.92	2077.92	2077.92	NA	2077.92	2077.92	2077.92
2	962	511.32	4960.32	2239.03	860.96	2057.19	1642.06	2755.17
3	1023	555.56	4960.32	2219.10	933.65	2050.00	1475.60	2908.25
5	754	454.32	4957.98	2159.26	917.37	2048.81	1450.40	2724.14
4	784	500.90	4945.05	2273.49	992.53	2047.18	1528.16	2931.03
1	1214	342.62	4960.32	2156.81	792.18	2021.93	1566.20	2650.60
0	584	359.12	4605.26	1886.07	815.19	1704.55	1282.61	2342.63

Taula 5.11. Descriptiva de la variable PRICE/M2 (en €/m^2) respecte al factor FLOOR

Per un costat, observem com la planta baixa esdevé el nivell que presenta un preu mitjà del metre quadrat més baix. Aquest fet vindria explicat en gran part per la localització dels corresponents immobles amb un pes del 49.31% dels districtes de Rascanya, Benicalap i Poblats Marítims, recordant també que la majoria d'habitatges de construcció horitzontal i, per tant, planta baixa, s'ubiquen en aquestes zones. Altrament, podem observar una distribució molt desigual entre els diferents nivells en termes d'observacions. Això ens portaria a una refactorització en agregacions majors. Ara bé, a diferència del número d'habitacions, hem obtingut amb el procés de *scraping* una variable alternativa a FLOOR referent a agrupacions superiors d'aquesta i que tenim dividida en tres variables dicotòmiques: GROUND_FLOOR, MIDDLE_FLOORS i ATTIC.

FLOOR2	n	min	max	mean	sd	median	q25	q75
ATTIC	484	773.20	4959.68	3001.89	930.59	2949.38	2346.95	3672.73
MIDDLE_FLOORS	6029	342.62	4964.29	2240.53	893.50	2090.23	1563.22	2823.53
GROUND_FLOOR	584	359.12	4605.26	1886.07	815.19	1704.55	1282.61	2342.63

Taula 5.12. Descriptiva de la variable PRICE/M2 (en €/m^2) respecte als factors ATTIC, MIDDLE_FLOORS i GROUND_FLOOR

És necessari subratllar el fet que amb aquestes agregacions podem identificar els àtics, mentre que amb el número de planta no es poden determinar. Per una altra banda, podem observar com no hi ha presència de missings, recordant que quan es tracta de variables que anomenàvem variables filtre, no es donaven dades mancants. Així que, no podent assumir que la informació publicada no és correcta i entenent que es tracta d'anuncis on únicament no s'especifica la planta exacta, utilitzar les variables GROUND_FLOOR, MIDDLE_FLOORS i ATTIC respecte a FLOOR pel que fa a la completesa de dades. De igual forma que amb la variable HAB, l'avaluació i decisió de quines variables emprar i com refactorizar, si pertoca, la variable FLOOR, les durem a terme quan modelitzem.

Una altra variable, que, com el nombre d'habitacions i la superfície, podem considerar de dimensió, és el número de banys del immoble, que tenim dividida entre variables dicotòmiques: BATHROOMS_1, BATHROOMS_2, BATHROOMS_3 (un bany, dos banys i tres o més banys; respectivament).
Si prenem la variable com numèrica podem observar, com era d'esperar, una correlació alta amb la superfície, ascendint a 0.663, com també respecte al preu, amb un valor del coeficient de correlació lineal de Pearson igual a 0.666.

BATHROOMS	n	min	max	mean	sd	median	q25	q75
BATHROOMS_3	847	639.66	4960.32	2962.50	903.13	2965.35	2362.80	3512.40
BATHROOMS_2	3298	359.12	4960.32	2407.10	864.31	2250.00	1767.48	2950.61
BATHROOMS_1	2952	342.62	4964.29	1901.98	812.40	1724.14	1295.85	2354.23

Taula 5.13. Descriptiva de la variable PRICE/M2 (en €/m^2) respecte als factors BATHROOMS_1, BATHROOMS_2 i BATHROOMS_3

A diferència de les anteriors variables, pel número de banys únicament tenim les dades agrupades en aquests tres nivells. Podem observar clares diferències respecte a la mitjana i la mediana del preu entre nivells, alhora que la dispersió en aquests esdevé similar entre ells. De cara a la modelització haurem de contrastar si les diferències entre els preus mitjans, recordant que la nostra variable resposta esdevé el preu de l'immoble i no el preu del metre quadrat com aquí estem traçant, són significatives, tenint en compte la interacció amb la resta de variables del model.

Una altra variable, que a priori ha de tenir un efecte significatiu de cara a la determinació del preu d'un habitatge, és l'estat de l'immoble que tenim separat, novament, per nivells en tres variables dicotòmiques: REFORM, GOOD_CONDITION i NEW. Si ens fixem en el pes de cada nivell sobre la base, veurem que els immobles d'obra nova únicament representen el 3.03% del total dels immobles (215 observacions), mentre que els habitatges a reformar i aquells que es considera en bones condicions esdevenen el 14.43% i 82.54% (1024 i 5858 observacions), respectivament. Així doncs, tenim una distribució molt desigual entre els tres nivells.

Per a una millor visualització del comportament del preu del metre quadrat en funció de l'estat de l'habitatge, representarem ambdues variables en el seu corresponent boxplot.

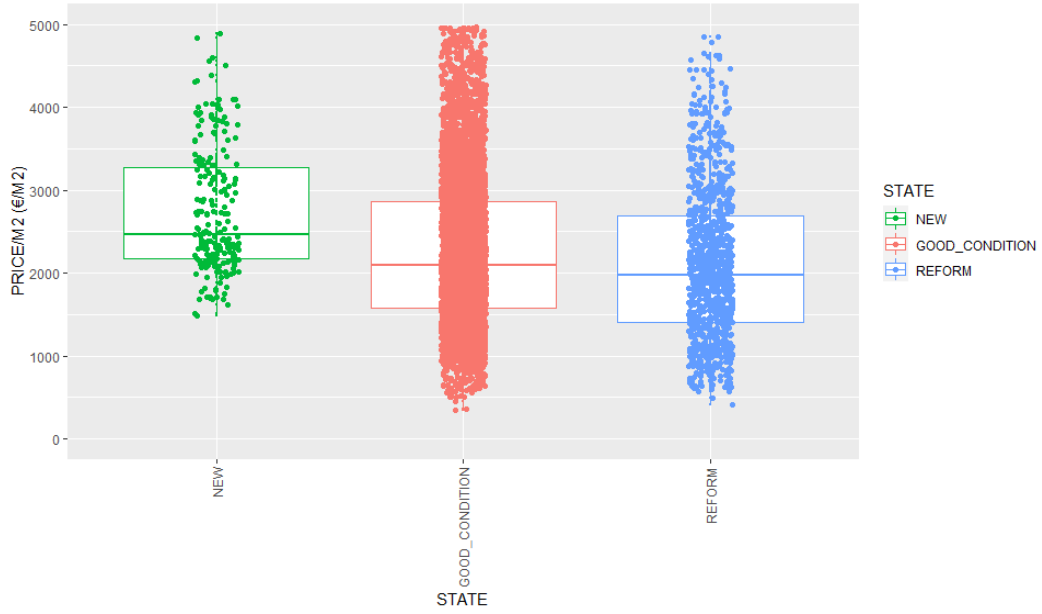


Figura 5.13. Boxplot del preu del metre quadrat (en €/m^2) per estat del immoble

Per un costat observem una diferència rellevant (468.32€ més que l'observat al nivell GOOD_CONDITION) entre el preu mitjà dels habitatges de nova construcció respecte als altres dos nivells, com era esperable. Ara bé, sorprèn que la diferència entre els pisos a reformar i els que estan en bones condicions sigui tant baixa. Aquest fet ve explicat en gran part per la localització dels immobles a reformar, donant-se que el 49.80% dels habitatges es troben ubicats en els districtes L'Eixample, Ciutat Vella i Extramurs (201, 196 i 113 observacions, respectivament), zones que, amb un preu mitjà del metre quadrat considerablement major a la mitjana de la ciutat,

categoritzàvem com més “exclusives”. Podem apreciar l’efecte contrari dins del nivell d’obra nova, amb un pes major de les zones amb un preu mitjà inferior com Rascanya, Benicalap i Patraix; en concordança amb l’asimetria positiva i la dispersió que observem en el boxplot. Així doncs, haguéssim esperat una diferència de preus en els extrems majors a l’observada sense la interacció de la variable ZONE.

Continuem amb les variables ACCESSIBLE i EXTERIOR, que recordem que indiquen, respectivament, si l’habitatge està adaptat per persones amb mobilitat reduïda i si les habitacions principals disposen de finestres orientades a la via pública o, pel contrari, a un pati interior. Respecte a l’accessibilitat del immoble, és necessari fer un incís per assenyalar que aquesta esdevé la variable, la informació de la qual, presenta menys garanties de correspondre’s amb la realitat. Malauradament, ens trobem en un context social que té un camí molt llarg per recórrer en termes, no només de sensibilització, sinó de legislació i compliment en referència als drets de les persones amb mobilitat reduïda. Si l’Estat ja no garanteix aquests drets pel que fa la via pública, és descabellat pensar que l’anunciant d’un portal immobiliari tindrà clara la definició d’habitatge accessible i actuarà d’acord amb aquesta amb responsabilitat a l’hora de donar aquesta dada. Encara és més poc creïble que s’apliqui algun tipus de control sobre la informació. Per exemple, sense anar més lluny, consultant els immobles accessibles que no disposen d’ascensor trobem 5 habitatges que no són planta baixa. Per tant, no compleixen ni el mínim d’accessibilitat i els marcarem com a no accessibles. Trobem també un sisè immoble, però descobrim que a la descripció sí que figura que l’edifici té ascensor, per tant, modificarem la variable ELEVATOR.

A banda d’aquesta darrera revisió, no entrarem a discutir més enllà la veracitat de la informació donant per vàlida la variable ACCESSIBLE com qualsevol altra de les presents.

Passem ara a resumir ambdues variables en funció del preu del metre quadrat.

ACCESSIBLE	n	min	max	mean	sd	median	q25	q75
1	558	414.55	4932.74	2534.86	923.63	2397.97	1813.65	3164.10
0	6539	342.62	4964.29	2240.11	912.88	2083.33	1550.00	2835.05

EXTERIOR	n	min	max	mean	sd	median	q25	q75
1	6119	359.12	4964.29	2308.08	917.95	2144.14	1602.99	2905.51
0	978	342.62	4960.32	1982.98	860.82	1833.33	1315.17	2521.07

Taules 5.14. Descriptives de la variable PRICE/M2 (en €/m^2) en funció dels factors ACCESSIBLE i EXTERIOR

Veiem que tant els immobles accessibles com exteriors presenten un preu mitjà del metre quadrat major que els seus nivells oposats. Pel que fa al pes sobre el total en quant a observacions, la distribució entre nivells és molt desigual. En el cas de la variable ACCESSIBLE, el menor pes amb un 7.86% el presenta el nivell positiu, mentre que per EXTERIOR és el nivell negatiu (interior) el que presenta únicament un 13.78% del total d'observacions.
Fent l'exercici d'agregar factors que podrien interaccionar amb els dos presents, com per exemple, la zona en ambdós casos i l'estat del immoble en el cas d'ACCESSIBLE (ja que recull el nivell corresponent a obra nova i podríem associar a priori nova construcció amb una major adaptació del immoble, però cap de les 558 observacions correspon a immobles d'obra nova), no trobem indicis de possibles interaccions interessants.

En canvi, on sí que visualitzem a priori una interacció candidata a ésser significativa és en la següent variable que explorarem: ELEVATOR (amb la variable FLOOR o els factors GROUND_FLOOR, MIDDLE_FLOORS i ATTIC). Ara bé, prèviament a analitzar la interacció, resumirem el preu del metre quadrat únicament respecte a la disponibilitat o no d'ascensor.

ELEVATOR	n	min	max	mean	sd	median	q25	q75
1	5473	523.08	4964.29	2427.72	879.75	2272.73	1741.94	2990.38
0	1624	342.62	4893.62	1709.11	818.18	1452.53	1094.33	2167.17

Taula 5.15. Descriptiva de la variable PRICE/M2 (en €/m^2) en funció del factor ELEVATOR

Com esperàvem, el preu mitjà del metre quadrat és considerablement major en els immobles que disposen d'ascensor que en aquells que no. La distribució en número d'observacions és desigual, acaparant els habitatges amb ascensor el 77.12% del total de la base.

Seguidament, representarem en un boxplot el preu del metre quadrat respecte a la interacció, que esmentàvem prèviament, entre el factor ELEVATOR i GROUND_FLOOR, MIDDLE_FLOORS i ATTIC. En canvi, no farem el mateix exercici entre ELEVATOR i TYPE, perquè directament no aplica als habitatges de construcció horitzontal, ja que la totalitat d'aquests figuren sense ascensor (tot i que no tenim una definició específica per part del portal, interpretarem la variable ELEVATOR en un sentit únic d'habitatge plurifamiliar).

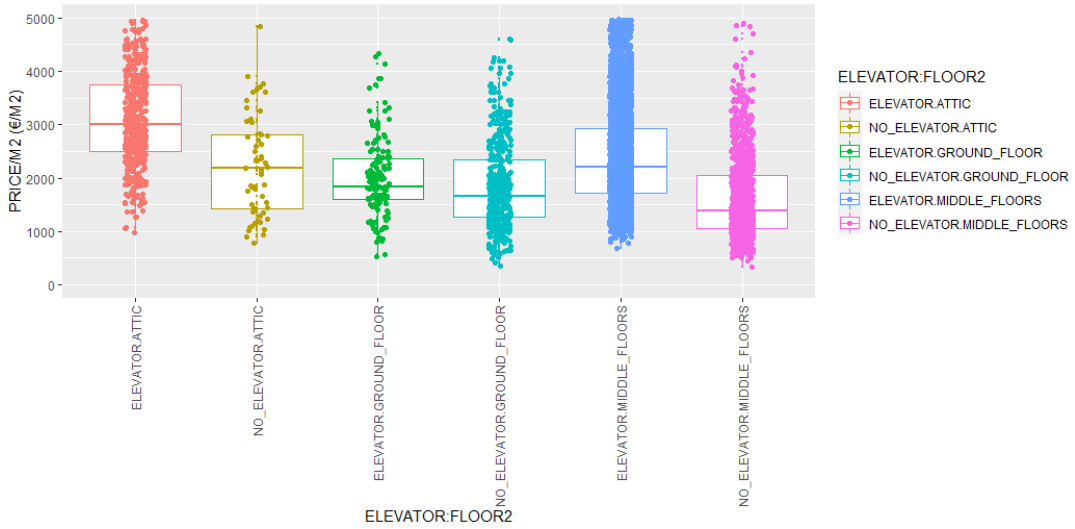


Figura 5.14. Boxplot del preu del metre quadrat (en €/m^2) en funció de la interacció de d'ELEVATOR amb GROUND_FLOOR, MIDDLE_FLOORS i ATTIC (FLOOR2)

És interessant contrastar amb la visualització, com esperàvem, que el fet de disposar d'ascensor no marca un diferencial en el preu mitjà del metre quadrat en immobles ubicats en una planta baixa. Mentre que la diferència entre el preu mitjà del metre quadrat amb ascensor i sense ascensor en els baixos puja tant sols a 166.21€, en el cas dels àtics i dels pisos de planta intermèdia ascendeix a 879.7€ i 751.83€. Per una altra banda, observem que la falta d'ascensor en el nivell àtic fa que el preu mitjà d'aquest i d'un immoble de planta intermèdia que sí disposa d'ascensor, s'equiparin (2225.79 €/m^2 i 2380.62 €/m^2, respectivament). D'igual forma també ho podem visualitzar en un

nivell inferior, entre els immobles ubicats en plantes intermèdies sense ascensor i els habitatges de planta baixa.

Prosseguim amb les variables referents a la disponibilitat de plaça d'aparcament amb l'habitatge: GARAGE, GARAGE_INCLUDED, GARAGE_NOT_INCLUDED, GARAGE_PRICE. Quant a la variable GARAGE, veiem que únicament un 22.29% dels immobles mostren alhora anunciat que disposen de garatge. Podem observar com els immobles amb pàrquing presenten un preu mitjà del metre quadrat 448.75€ major que aquells que no en disposen.

GARAGE	n	min	max	mean	sd	median	q25	q75
1	1582	359.12	4960.32	2612.00	863.47	2499.52	1979.38	3179.72
0	5515	342.62	4964.29	2163.25	907.59	1978.49	1474.14	2734.88

Taula 5.16. Descriptiva de la variable PRICE/M2 (en €/m^2) en funció del factor GARAGE

Si baixem ara a la factorització dels immobles amb garatge respecte a si aquest està o no inclòs en el preu de l'operació de venta de l'habitatge, observem que un 20.67% dels immobles amb garatge l'ofereixen en una opció de compra independent. Podem observar, com era d'esperar, que els immobles amb pàrquing inclòs presenten el major preu mitjà del metre quadrat (2640.98 €/m^2), seguits d'aquells en que el garatge no està inclòs (2500.79 €/m^2). Ambdós nivells amb un preu mitjà considerablement superior als immobles sense pàrquing anunciat (2163.25 €/m^2).

GARAGE2	n	min	max	mean	sd	median	q25	q75
GARAGE_INCLUDED	1255	359.12	4960.32	2640.98	890.83	2531.65	1973.42	3225.74
GARAGE_NOT_INCLUDED	327	1029.41	4960.32	2500.79	739.97	2360.00	2000.00	2976.97
NOT_GARAGE	5515	342.62	4964.29	2163.25	907.59	1978.49	1474.14	2734.88

Taula 5.17. Descriptiva de la variable PRICE/M2 (en €/m^2) en funció del factor GARAGE2, construït mitjançant GARAGE_INCLUDED i GARAGE_NOT_INCLUDED

Respecte al preu de les opcions de compra de garatge (GARAGE_PRICE), tenim 267 immobles que tenen un import informat, perquè recordem que al procés de *data cleaning* havíem transformat a *missings* els preus que venien a 0 o, amb un valor molt baix, corresponien a lloguers. Podem observar un preu mitjà de l'opció de compra de 27588€. D'altra banda, al comptar tant poques observacions no avaluarem una interacció que se'ns presenta com força interessant i que esdevindria entre el preu d'aquestes opcions de compra i el districte on s'ubica el immoble, responent a avaluar si observem diferències considerables entre zones més aglomerades i altres més dilatades (com podríem associar a districtes més cèntrics i perifèrics, respectivament).

Tot seguit, encarem sis variables dicotòmiques corresponents a característiques que a priori podríem considerar en general més secundàries en la recerca d'un immoble: FITTED_WARDROBE, SPLIT, TERRACE, GARDEN, POOL, STORAGE. Això no es tradueix necessàriament en un menor pes en la determinació del preu de l'habitatge.

Procedim a resumir el preu del metre quadrat respecte a aquest conjunt de factors.

FITTED_WARDROBE	n	min	max	mean	sd	median	q25	q75
1	3800	500.90	4964.29	2472.64	888.67	2355.37	1779.14	3033.21
0	3297	342.62	4960.32	2021.98	889.80	1834.95	1350.00	2530.62

Taula 5.18. Descriptiva de la variable PRICE/M2 (en €/m^2) en funció del factor FITTED_WARDROBE

SPLIT	n	min	max	mean	sd	median	q25	q75
1	3804	523.08	4964.29	2522.68	911.19	2428.57	1798.44	3111.64
0	3293	342.62	4950.00	1963.63	828.41	1815.22	1336.63	2421.57

Taula 5.19. Descriptiva de la variable PRICE/M2 (en €/m^2) en funció del factor SPLIT

TERRACE	n	min	max	mean	sd	median	q25	q75
1	2422	359.12	4960.32	2500.47	954.32	2404.45	1748.08	3157.58
0	4675	342.62	4964.29	2140.40	872.32	1977.53	1495.49	2685.71

Taula 5.20. Descriptiva de la variable PRICE/M2 (en €/m^2) en funció del factor TERRACE

GARDEN	n	min	max	mean	sd	median	q25	q75
1	483	575.54	4932.74	2569.18	890.47	2564.10	1835.27	3207.14
0	6614	342.62	4964.29	2240.94	915.09	2077.85	1555.56	2834.13

Taula 5.21. Descriptiva de la variable PRICE/M2 (en €/m^2) en funció del factor GARDEN

POOL	n	min	max	mean	sd	median	q25	q75
1	459	1043.48	4932.74	2881.22	698.65	2825.62	2446.31	3323.52
0	6638	342.62	4964.29	2220.55	915.06	2042.34	1541.67	2794.60

Taula 5.22. Descriptiva de la variable PRICE/M2 (en €/m^2) en funció del factor POOL

STORAGE	n	min	max	mean	sd	median	q25	q75
1	1193	359.12	4960.32	2557.81	905.57	2445.78	1891.89	3134.33
0	5904	342.62	4964.29	2203.77	907.96	2027.03	1521.35	2777.78

Taula 5.23. Descriptiva de la variable PRICE/M2 (en €/m^2) en funció del factor STORAGE

Com és raonable al tractar-se de característiques positives, la presència de l'atribut comporta un preu mitjà del metre quadrat major en totes les variables. També era d'esperar un nombre baix d'immobles amb piscina o jardí (al voltant del 6% de pes sobre el total d'observacions en ambdós factors). Així mateix, hem comprovat mitjançant la interacció entre POOL i GARDEN, ja que es presentava com la més

susceptible d'una major correspondència entre nivells, que a priori no esdevindran focus de problemes de multicol·linealitat.

A continuació, aprofundirem en una variable d'una naturalesa diferent a totes les tractades prèviament i que esdevé el número d'imatges del immoble publicades a l'anunci (PHOTO). Aquesta discreta, que ens ha generat molt interès, l'hem de separar de la resta precisament perquè trenca amb l'associació d'immoble amb anunci, ja que es tracta d'una característica pròpia d'aquest darrer. Així doncs, avançant-nos, si incloem la variable en els nostres models, forçarem a que el input d'aquests siguin habitatges de portals immobiliaris. Fet que no s'ha de prendre com un inconvenient, però que s'ha de tenir present de cara a la utilitat dels models que presentem.

Si passem a fixar-nos en la distribució de la variable i construïm la corresponent taula de freqüències, observarem que la moda se situa en el valor 0, tenint 1523 immobles sense cap imatge (21.46% del total), mentre que la mediana i la mitjana es posicionen en 18 i 18.96 imatges, respectivament. Per una altra banda, sorprèn que el tercer quartil se situí en tan sols 28 imatges, ja que haguéssim esperat que aquest valor fos menor, és a dir, menys immobles amb tantes imatges. Alhora trobem valors extrems com és el cas d'un pis amb 171 imatges publicades, que correspon al màxim observat. Enfront això, revisem manualment alguns d'aquests anuncis amb la sospita que pugui donar-se un problema d'imatges repetides, però no en detectem, tot i que si que hi ha forces fotografies que són quasi idèntiques (existeix un control per part del portal en aquest sentit).

Representem en un histograma la nostra variable per tal de tenir una visió més clara de la distribució descrita.

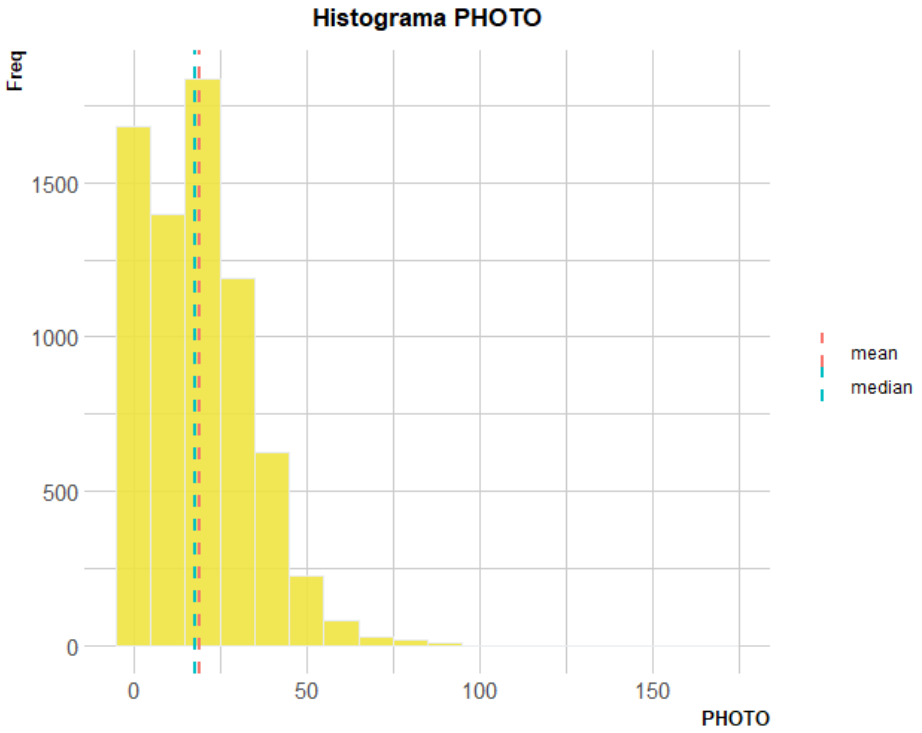


Figura 5.15. Histograma del número d'imatges del immoble publicades a l'anunci

La idea darrera de la ingesta i l'anàlisi present neix de la intuïció que el número d'imatges anunciades puguin explicar en alguna mesura el preu del immoble. Així doncs, passem a calcular la correlació entre PRICE i PHOTO, obtenint un coeficient de correlació de Pearson de 0.2817. Ara bé, mentre que, com esperàvem, la correlació és positiva, aquesta, al ser menor a 0.30, la consideraríem dèbil. Per tant, a priori sembla que el número d'imatges no ens donarà el joc que esperàvem. Tanmateix, això no resta valor al potencial de les imatges, sinó que, avançant-nos a les propostes de millora, potser el focus hauria de centrar-se en l'anàlisi de les pròpies imatges.

En darrer lloc, i no fortuïtament, ja que tot i manifestar potencial, són més complexes d'abordar i manejar posteriorment a la modelització, analitzarem les variables BANK, PHONE i AGENCY conjuntament i DESC_1.
Comencem resumint el preu del metre quadrat respecte a si es tracta d'un immoble pertanyent a una entitat financera o no.

BANK	n	min	max	mean	sd	median	q25	q75
0	6975	359.12	4964.29	2276.58	913.64	2117.65	1579.27	2866.67
1	122	342.62	4179.49	1502.97	787.73	1267.46	934.97	1912.89

Taula 5.24. Descriptiva de la variable PRICE/M2 (en €/m^2) en funció del factor BANK

Podem observar que tenim únicament 122 immobles que siguin propietat d'entitats financeres (1.72% del total), mentre que el seu preu mitjà del metre quadrat, amb 1502.97 €/m^2, és un 33.98% inferior al dels habitatges que no són propietat de bancs o fons d'inversió. No obstant això, si afegim a l'equació el factor districte, veurem que hi ha una fort desequilibri entre zones. Sense anar més lluny, els quatre districtes amb un preu mitjà del metre quadrat més baix (Rascanya, L'Olivereta, Jesús i Benicalap) acabalant el 45.90% d'aquests immobles. Aquest fet no tant sols explica que la diferència de preus respecte al factor BANK sigui tant gran, sinó que posa en relleu una realitat social, ja que aquests quatre districtes presenten les menors rendes netes mitjanes anuals per llar de la ciutat.

D'altra banda, esperàvem un nombre tant baix d'immobles de propietat financera? Veient el pes de fons d'inversió com Blackstone i KKR al mercat immobiliari espanyol sembla difícil de creure i aquí rau el motiu d'analitzar conjuntament la variable BANK amb les referents al telèfon de contacte (PHONE) i l'agència immobiliària que publica l'anunci (AGENCY).

Cap dels habitatges que tenim categoritzats com immobles de banc té informada l'agència immobiliària. Ara bé, sí tenim el telèfon de contacte i a través d'aquest podem veure que tres immobiliàries bancàries acaparen 110 dels 122 immobles (Solvia, 68; Haya, 27; Aliseda, 15). Així mateix, si ens fixem en la distribució de la variable AGENCY detectem almenys dos nivells "holapisos1" i "solviafranquiciavalencia" (119 i 8 habitatges, respectivament) que corresponen a agències que treballen en exclusiva amb immobles d'entitats financeres (Hola Pisos és una web pertanyent a la immobiliària bancària Anticipa Real Estate). Així doncs, tenim com a mínim 127 immobles que haurien d'estar categoritzats com immobles de banc, però no figuren com a tal. Això, juntament amb el fet que tenim 3101 missings a la variable AGENCY on podríem trobar més casos, posa en relleu les poques garanties que presenta la variable BANK o la política del portal per fer valer la qualitat dels seus anuncis envers aquesta condició. Tanmateix, no aplicarem transformació alguna sobre els 127 immobles que hem identificat, ja que el procés ha sigut en part manual. Per

una altra banda, no ens plantegem introduir ni PHONE ni AGENCY en els nostres models, fet pel qual no aprofundirem en el seu anàlisi. PHONE i AGENCY, en sintonia amb el número d'imatges, són variables que trenquen amb l'associació entre immoble i anunci, ja que són pròpies d'aquest darrer, i tot i que òbviament existeixen agències, per exemple, que es cataloguen com "de luxe" i, per tant, sí podríem observar un comportament de preu diferenciat respecte a la variable AGENCY i PHONE, no creiem que aporti utilitat alguna i sí el inconvenient que comentàvem d'ésser pròpies de l'anunci i no del immoble, així com esdevenir candidates a interpretacions errònies.

Finalment, plantejarem el tractament de la variable DESC_1, on, recordem, hem recollit un fragment, no la totalitat, de la descripció del immoble publicada a l'anunci. El disposar únicament d'una fracció del text és una circumstància limitant, ja que no podem dur a terme un anàlisi complet i, per tant, una de les seves utilitats immediates que esdevindria fer-ne ús per contrastar la resta de variables i així sumar un grau més a l'exercici de garantir la qualitat de les dades, no se'ns presenta viable. No obstant això, sí que podem treure partida a la variable per detectar paraules clau que emmascarin factors amb influència sobre la determinació del preu d'un immoble i que al no disposar d'ells a la base resten com soroll. Seria el cas, per exemple, d'immobles que es venen arrendats a tercers, ocupats o amb càrregues com sense cèdula d'habitabilitat o en un estat de ruïna (hi ha una gran diferència entre que el immoble necessiti una reforma, que recollim a la variable REFORM, i que es trobi en un estat inhabitable).

Així doncs, efectuarem l'exercici de la següent forma: primerament, analitzem les descripcions dels immobles que presentin un preu del metre quadrat (PRICE/M2) que podríem considerar com valor atípic, amb la idea de trobar paraules clau i patrons, com els comentats al paràgraf anterior, que expliquin aquest comportament. Aquí hem de fer un breu incís per deixar clar que per una banda plantegem l'aplicació en un sentit negatiu (és a dir, factors que expliquin un preu més baix del immoble) i que la detecció i exclusió que vam fer dels outliers respecte PRICE/M2 va resultar únicament sobre immobles amb preus elevats, per tant, haurem de tensar el límit inferior per recollir observacions (ho farem disminuint de 1.5 a 0.5 el multiplicador del IQR). Un cop tenim els patrons (els completarem amb aquells que ja considerem de partida) aplicarem una recerca d'aquests sobre tota la base detectant els immobles amb aquests factors negatius ocults.

De l'anàlisi de les descripcions dels immobles amb el preu del metre quadrat més baix (207 observacions), destaquem els següents elements: "alquilado", "con inquilinos",

“ocupado”, “ocupada”, “subasta”. Com comentàvem, completarem el llistat amb possibles variacions d’aquests termes i altres que ja consideràvem d’inici. D’altra banda, les referències més freqüents eren entorn a la necessitat d’una reforma total o una rehabilitació, fins i tot hem trobat un parell de casos de locals comercials que es poden habilitar com habitatge (és a dir, que no tenen cèdula d’habitabilitat). A més, hem descobert esments a immobles de banc, corroborant la poca fiabilitat de la variable BANK al no estar aquests categoritzats com a tal.

Respecte al llistat de paraules clau que finalment hem definit, recollit a la taula 5.25, hem d’aclarir que esdevé vital que la paraula per si sola reculli un únic sentit possible. És a dir, descartem, per exemple, utilitzar el terme “reforma”, ja que pot trobar-se en un context positiu com podria ser que el immoble es transmet reformat.

Elements	Context
“alquilado”, “alquilada”, “con inquilinos”, “arrendado”, “arrendada”	Immoble llogat a tercers
“ocupado”, “ocupada”, “ocupas”, “okupado”, “okupada”, “okupas"	Immoble ocupat
“subasta”	No es tracta d’una transacció directa, sinó d’una subhasta
“sin cédula”,”sin cedula”, “ruina”	Immoble en un estat inhabitable

Taula 5.25. Llistat de paraules clau sobre la variable DESC_1

Així mateix, prosseguim fent una recerca dels elements concretats sobre tota la base i categoritzant els immobles amb coincidència textual en una nova variable (FLAG). Aquesta l’hem plantejat com a dicotòmica, és a dir, no classifiquem les observacions pel context de la paraula clau.

Detectem un total de 59 observacions, que, com esperàvem, presenten un preu mitjà del metre quadrat menor que els immobles lliures de coincidència. Serà a la modelització on, més concretament a l’anàlisi dels residus, on avaluarem la utilitat de la variable FLAG.

FLAG	n	min	max	mean	sd	median	q25	q75
0	7038	342.62	4964.29	2264.76	916.87	2105.26	1566.37	2857.14
1	59	637.35	4046.05	2086.60	936.04	1891.89	1410.84	2767.86

Taula 5.26. Descriptiva de la variable PRICE/M2 (en €/m^2) en funció del factor FLAG

VI. MODELITZACIÓ

6.1 Model General i Model No General

Un cop hem finalitzat l'anàlisi descriptiva de les variables que configuren la nostre base, abordem en aquest apartat la materialització de tot el procés traçat i executat en propostes de models, que ens permetin estimar el preu d'un immoble en funció de les variables significatives en la seva determinació.

El primer punt a qüestionar-nos ens porta a bifurcar el nostre plantejament en dos línies pivotant en torn a la inclusió i no inclusió de les variables de localització del immoble (ZONE i SUBZONE). En consonància, treballarem en pos de presentar dos models finals: un que no inclourà les variables espacials i, per tant, general i sense la limitació d'ésser aplicat a immobles d'una àrea concreta (en el nostre cas, València ciutat) i un segon afegint les variables de localització i que tindrà l'exclusivitat d'ús amb immobles de la ciutat, però que esperem que presenti un major ajust. Ara bé, avançant-nos, buscant salvar els punts forts d'ambdós plantejaments, presentarem una darrera proposta on substituirem les variables de localització per la renda neta mitjana anual per llar observada en la ubicació, gràcies a l'anàlisi previ on hem corroborat la correlació entre ambdues variables.

D'altra banda, en les dues línies de plantejament partirem d'un model de regressió lineal múltiple, on optarem, en el cas de les variables categòriques, per definir com a basals els nivells que presentin el menor preu mitjà (fent una excepció en el cas de la tipologia, l'estat i la planta del immoble on, per facilitar la interpretació, escollirem FLAT, GOOD_CONDITION i MIDDLE_FLOORS com a nivells basals). Alhora, per evitar el sobreajustament, procedirem a fraccionar aleatòriament la nostra base en un subconjunt d'entrenament i un de test (amb proporció d'observacions de 0.7 i 0.3, respectivament). Així mateix, a mesura que avancem amb la modelització, anirem sospesant les diferents alternatives de tractament d'algunes de les nostres variables que hem presentat al llarg de l'anàlisi descriptiva. Ara bé, com també hem comentat a l'anàlisi, no considerem la inserció de les variables PHOTO, PHONE i AGENCY en cap dels nostres models.

Així doncs, iniciarem la modelització amb el primer plantejament: sense incloure ZONE ni SUBZONE. Per una altra banda, quant a la variable HAB, d'entrada la refactoritzarem en els 4 nivells presentats a l'anàlisi (1, 2, 3 i 4 o més habitacions). Respecte a la planta del immoble, utilitzarem els factors GROUND_FLOOR, MIDDLE_FLOORS i ATTIC, i per la

tipologia, agruparem els nivells “Casa o chalet independiente” i “Chalet” en un sol nivell, que d’aquesta forma englobarà els habitatges de construcció horitzontal (HOUSE). Finalment, del conjunt de variables referents al garatge, incorporarem d’inici únicament GARGE_INCLUDED.

$$PRICE = \beta_0 + \beta_1 HOUSE + \beta_2 STUDIO + \beta_3 DUPLEX + \beta_4 PENTHOUSE$$
$$+\beta_5 M2 + \beta_6 HAB2 + \beta_7 HAB3 + \beta_8 HAB4 + \beta_9 GROUND\_FLOOR + \beta_{10} ATTIC$$
$$+\beta_{11} ACCESSIBLE + \beta_{12} NO\_BANK + \beta_{13} BATHROOMS\_2 + \beta_{14} BATHROOMS\_3$$
$$+\beta_{15} ELEVATOR + \beta_{16} EXTERIOR + \beta_{17} REFORM + \beta_{18} NEW$$
$$+\beta_{19} FITTED\_WARDROBE + \beta_{20} SPLIT + \beta_{21} TERRACE + \beta_{22} GARDEN$$
$$+\beta_{23} POOL + \beta_{24} STORAGE + \beta_{25} GARAGE\_INCLUDED + \epsilon$$

Fórmula 6.1. Equació del Model G1, on Ɛ correspon al terme de pertorbació

En referència als factors amb més d’un nivell, el basal del model estarà conformat per pisos ubicats en una planta intermèdia, en bones condicions i que únicament disposen d’una habitació i un bany.

Tot seguit, procedim a efectuar l’estimació per Mínims Quadrats Ordinaris (OLS).

```
Call:
lm(formula = PRICE ~ ., data = base1)

Residuals:
     Min       1Q   Median       3Q      Max
-1219181   -62031   -12190    43155  1331670

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -50671.56   15607.49  -3.247 0.001176 **
M2                  2415.51      39.39  61.318  < 2e-16 ***
ACCESSIBLE1         4637.20    6705.03   0.692 0.489221
ELEVATOR1          32877.99    5373.96   6.118 1.02e-09 ***
EXTERIOR1            -16.36    5714.60  -0.003 0.997716
GROUND_FLOOR1      -46144.95    9470.14  -4.873 1.14e-06 ***
ATTIC1              10274.44   15358.04   0.669 0.503531
REFORM1             -8186.08    5423.58  -1.509 0.131274
NEW1                25989.88   11542.81   2.252 0.024391 *
FITTED_WARDROBE1     8568.77    4146.33   2.067 0.038824 *
SPLIT1              31466.44    4238.10   7.425 1.32e-13 ***
TERRACE1            12441.82    4326.96   2.875 0.004052 **
GARDEN1            -10538.53    8444.60  -1.248 0.212104
POOL1               30516.99    8893.11   3.432 0.000605 ***
STORAGE1            10720.03    5206.99   2.059 0.039568 *
GARAGE_INCLUDED1    -1734.98    5348.51  -0.324 0.745660
PENTHOUSE1          69390.50   17576.00   3.948 7.99e-05 ***
DUPLEX1               481.86   14351.78   0.034 0.973218
HOUSE1            -105082.77   13458.95  -7.808 7.07e-15 ***
STUDIO1            -29580.68   20083.60  -1.473 0.140848
HAB21              -27977.56    8657.36  -3.232 0.001239 **
HAB31              -55106.01    8132.12  -6.776 1.38e-11 ***
HAB41              -49548.16    8934.38  -5.546 3.08e-08 ***
NO_BANK1             6690.92   14754.01   0.453 0.650210
BATHROOMS21         31630.57    4703.36   6.725 1.95e-11 ***
BATHROOMS31        174708.57    8516.95  20.513  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 126700 on 4933 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.7441,    Adjusted R-squared:  0.7428
F-statistic: 573.9 on 25 and 4933 DF,  p-value: < 2.2e-16
```

Output 6.1. Resum Model G1

Podem observar com el conjunt de variables ACCESSIBLE, EXTERIOR, GARDEN, GARAGE_INCLUDED i BANK no es posicionen com factors amb efectes significatius en la determinació del preu d'un habitatge en presentar els contrastos individuals un p-valor major a 0.05. Pel que fa a les qualitatives amb més d'un nivell com la tipologia, l'estat i la planta del immoble, la interpretació dels contrastos individuals de les seves corresponents variables fictícies ens pot portar a extreure conclusions errònies. Per tant hem de recórrer a crear una ANOVA, corroborant que els tres predictors sí tenen un efecte significatiu sobre la variable resposta. Altrament, sorprenen els coeficients associats a les dummies referents al número d'habitacions, ja que apareixen com a negatius, al contrari del que esperàvem en ésser el nivell basal una única habitació. Això se'ns presenta com un indici de que pot existir multicol·linealitat entre els nostres regressors.

Així mateix, no optarem per utilitzar directament els p-valors associats als predictors com a base de selecció, sinó que ens inclinarem per aplicar el mètode *stepwise* mixt.

Utilitzarem com a criteri l'Akaike (AIC), que tendeix a ésser més restrictiu i penalitzar en major mesura la inclusió de predictors que el R^2 ajustat.

Obtenim com a millor model del procés de selecció el següent:

$$PRICE = \beta_0 + \beta_1 HOUSE + \beta_2 STUDIO + \beta_3 DUPLEX + \beta_4 PENTHOUSE + \beta_5 M2 + \beta_6 HAB2 + \beta_7 HAB3 + \beta_8 HAB4 + \beta_9 GROUND\_FLOOR + \beta_{10} ATTIC + \beta_{11} BATHROOMS\_2 + \beta_{12} BATHROOMS\_3 + \beta_{13} ELEVATOR + \beta_{14} REFORM + \beta_{15} NEW + \beta_{16} FITTED\_WARDROBE + \beta_{17} SPLIT + \beta_{18} TERRACE + \beta_{19} POOL + \beta_{20} STORAGE + \epsilon$$

Fórmula 6.2. Equació del Model G2, on Ɛ correspon al terme de pertorbació

```
Call:
lm(formula = PRICE ~ ., data = base1)

Residuals:
     Min       1Q   Median       3Q      Max
-1217185   -62055   -12260    43086  1332544

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       -44337.72    8847.21  -5.011 5.59e-07 ***
M2                 2413.69      39.22  61.549  < 2e-16 ***
ELEVATOR1         33200.25    5284.49   6.283 3.62e-10 ***
GROUND_FLOOR1     -46941.78    9398.34  -4.995 6.10e-07 ***
ATTIC1             9922.01   15347.68   0.646  0.51800
REFORM1           -7934.10    5386.10  -1.473  0.14080
NEW1              26852.90   11485.27   2.338  0.01943 *
FITTED_WARDROBE1   8439.12    4114.13   2.051  0.04029 *
SPLIT1            31585.40    4204.82   7.512 6.89e-14 ***
TERRACE1          12346.91    4314.34   2.862  0.00423 **
POOL1             25009.70    7768.67   3.219  0.00129 **
STORAGE1          10167.27    5120.79   1.985  0.04715 *
PENTHOUSE1        69864.67   17565.87   3.977 7.07e-05 ***
DUPLEX1             365.19   14343.80   0.025  0.97969
HOUSE1           -105417.74   13406.72  -7.863 4.57e-15 ***
STUDIO1           -29804.65   20067.86  -1.485  0.13756
HAB21             -27737.93    8643.73  -3.209  0.00134 **
HAB31             -54960.95    8115.90  -6.772 1.42e-11 ***
HAB41             -49357.17    8910.54  -5.539 3.20e-08 ***
BATHROOMS21        31440.46    4677.10   6.722 1.99e-11 ***
BATHROOMS31       174855.39    8468.20  20.648  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 126700 on 4938 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.744,     Adjusted R-squared:  0.743
F-statistic: 717.6 on 20 and 4938 DF,  p-value: < 2.2e-16
```

Output 6.2. Resum Model G2

Així doncs, observem que han quedat fora les 5 variables que assenyalàvem com que no tenien un paràmetre associat significatiu al nostre primer model. El nou model presenta un mateix R^2 ajustat que el primer, esdevenint força elevat i ascendit el valor d'aquest a 0.743. És a dir, el model és capaç d'explicar el 74.3% de la variabilitat observada en la determinació del preu de l'habitatge. Per últim, novament l'estadístic

F presenta un p-valor associat molt inferior a 0.05, per tant, podem rebutjar que no hi hagi cap regressor amb efecte significatiu.

Prosseguim amb la validació pel nostre model G2 dels supòsits de normalitat, linealitat, independència i homoscedasticitat.

Per tal de diagnosticar si els residus es distribueixen segons una normal, primerament utilitzarem el corresponent Q-Q plot per comparar els residus estandarditzats amb una distribució normal teòrica.

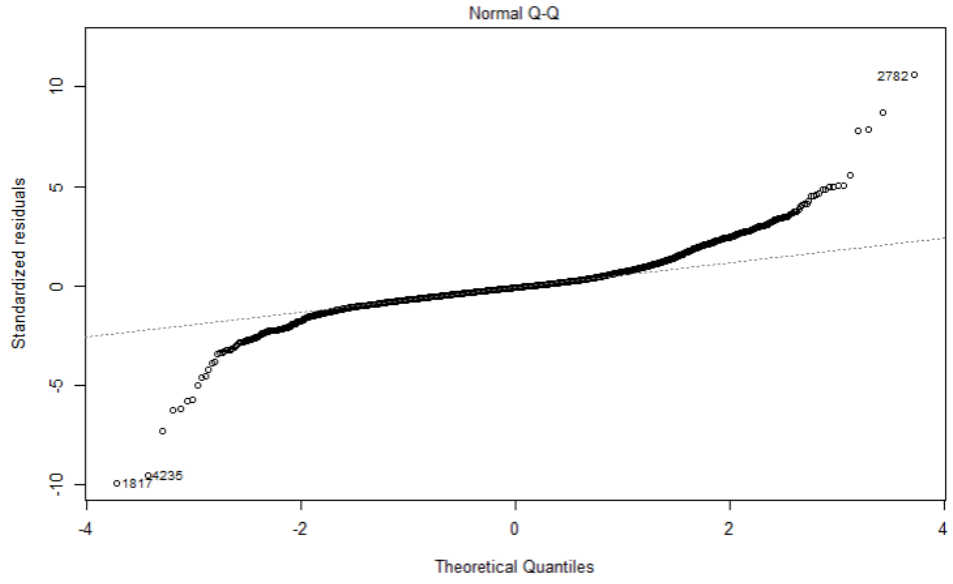


Figura 6.1. Model G2. Q-Q plot dels residus estandarditzats

Com avançàvem a l'anàlisi descriptiva, observant que la nostra variable resposta no es distribuïa segons una normal i suggeríem en enfront això una transformació logarítmica, podem visualitzar amb claredat que no es compleix el supòsit de normalitat dels residus.

Ho corroborem amb el test de normalitat de Shapiro-Wilk, obtenint un p-valor molt inferior a 0.05 i, per tant, rebutjant la hipòtesi nul·la de normalitat.

```
        Shapiro-Wilk normality test

data:  model2$residuals
W = 0.87499, p-value < 2.2e-16
```

Pel que fa a la linealitat, és a dir, que la variable resposta estigui linealment relacionada amb les variables independents, visualitzarem els residus enfront els valors ajustats.

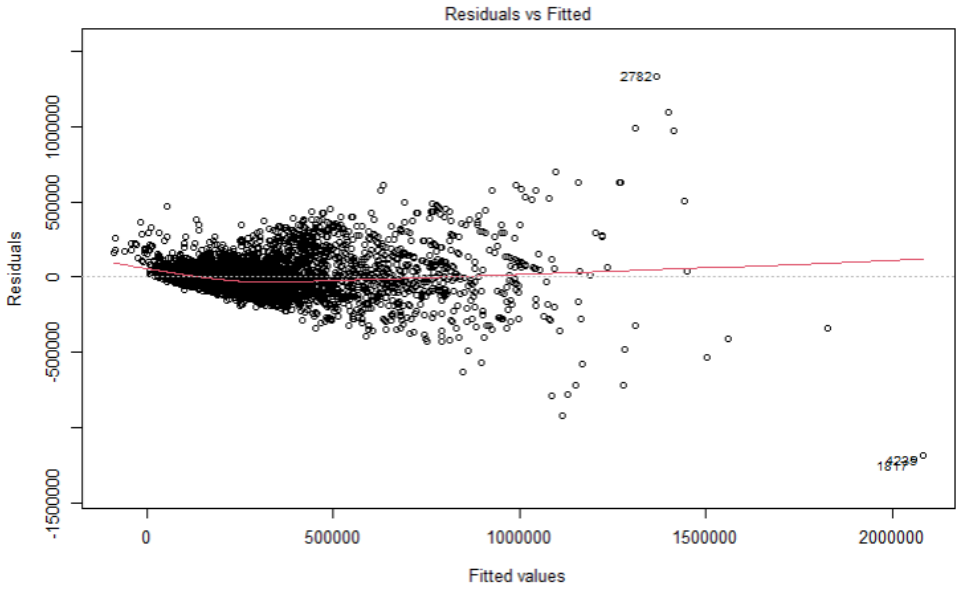


Figura 6.2. Model G2. Gràfic de dispersió residus vs valors ajustats

Podem observar que la corba d'ajust (en vermell) no acaba d'ésser recta i horitzontal, de forma que no visualitzem el comportament desitjat que suposaria que els residus i els valors observats es distribueixen entre si aleatòriament. Això ens suggereix que una alternativa a una aproximació lineal podria esdevenir més desitjable. Alhora també ens posa en alerta respecte al compliment del supòsit d'homoscedasticitat. D'altra banda, es reporten 3 valors atípics (observacions 1817, 2782 i 4235), que ja identificàvem al Q-Q plot.

Lligant amb el diagnòstic d'homoscedasticitat, representar les arrels dels residus estandarditzats enfront els valors ajustats ens donarà una visió més acurada per avaluar si efectivament la variància dels errors no es manté constant.

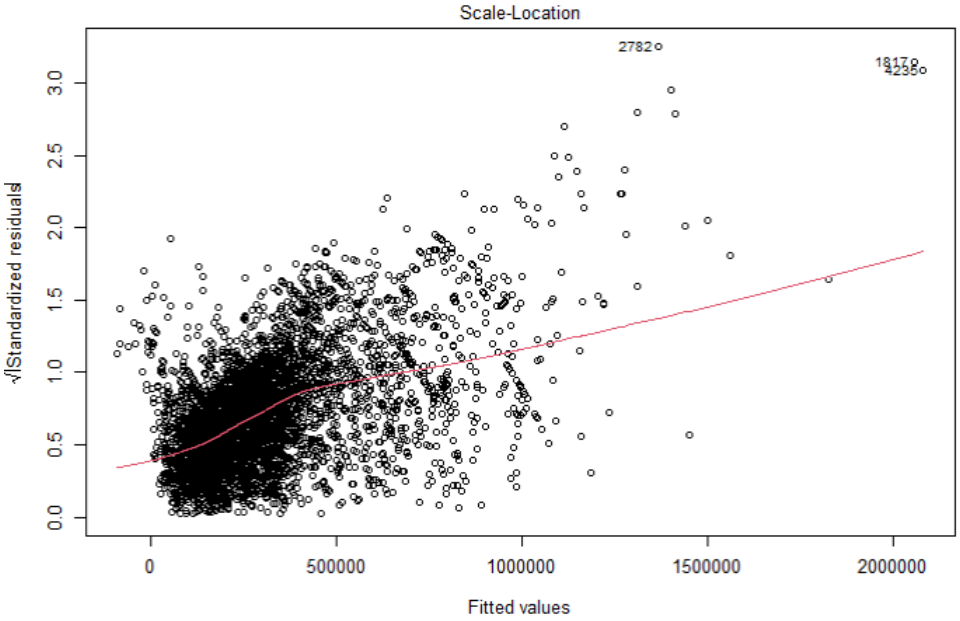


Figura 6.3. Model G2. Gràfic de dispersió de les arrels dels residus estandarditzats vs valors ajustats

Podem apreciar amb claredat un possible problema latent d'heteroscedasticitat, que corroborem amb el test de Breusch-Pagan, obtenint un p-valor molt inferior a 0.05 i, per tant, rebutjant la hipòtesi nul·la de variància constant dels residus.

```
        studentized Breusch-Pagan test

data:  model2
BP = 1528.1, df = 20, p-value < 2.2e-16
```

Com a proposta per evitar que la presència d'heteroscedasticitat tingui un impacte important sobre l'estimació del nostre model, plantejarem l'alternativa d'ajustar una regressió robusta, menys sensible a l'incompliment del supòsit de variància constant dels residus com també enfront la presència d'outliers.

Aquests valors atípics influents els identificarem mitjançant el corresponent *Influence plot* i haurem de procedir a remoure'ls de la nostra base.

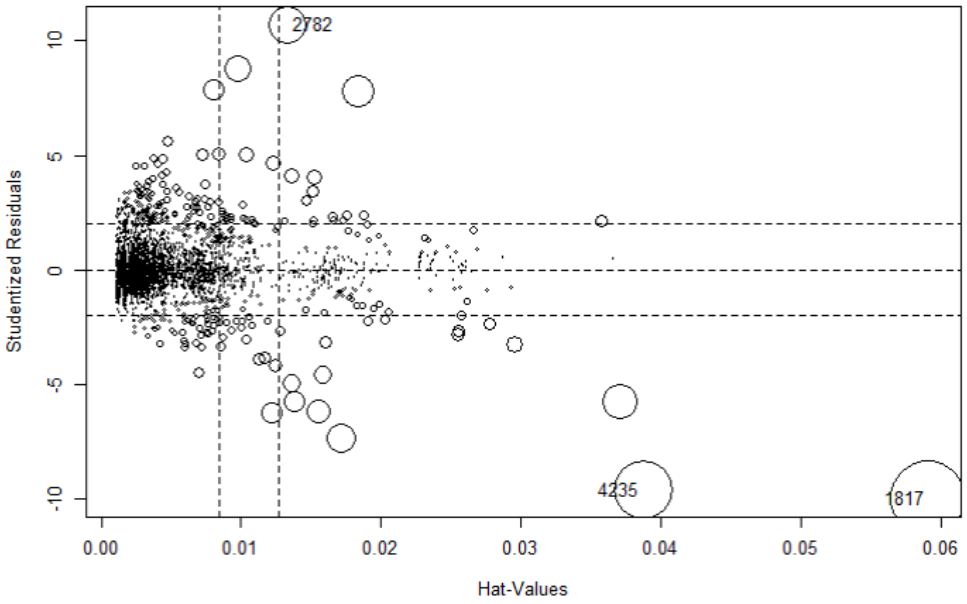


Figura 6.4. Model G2. *Influence plot*

Així doncs, detectem com outliers influents les mateixes 3 observacions que ja havíem assenyalat en visualitzacions anteriors. Les observacions 1817 i 4235 presenten un preu molt menor a l'esperat. En ambdós casos la superfície del immoble anunciada és clarament superior a la real. A més, el immoble 1817 té la tipologia mal informada, ja que a la descripció s'indica una casa, no un pis. Quant a l'observació 2782, amb un preu molt superior a l'esperat que ascendeix a 2.7 milions d'euros, l'explicació la trobem ràpidament a la descripció al llegir que es tracta d'un "Histórico palacete con 100 años de antigüedad".

El darrer diagnòstic que realitzarem buscarà contrastar si existeix multicol·linealitat entre els nostres regressors, fet que implica un increment en la variància dels coeficients de regressió estimats comportant una afectació de cara a determinar la seva significança. Amb aquesta idea, calcularem l'índex VIF (Factor de Inflació de la Variància).

M2	ELEVATOR	GROUND_FLOOR	ATTIC	REFORM
2.460330	1.517459	2.088593	4.508996	1.124174
NEW	FITTED_WARDROBE	SPLIT	TERRACE	POOL
1.126122	1.299713	1.357034	1.286287	1.138952
STORAGE	PENTHOUSE	DUPLEX	HOUSE	STUDIO
1.115127	4.460959	1.268600	2.221822	1.168358
HAB2	HAB3	HAB4	BATHROOMS2	BATHROOMS3
3.114445	4.958709	5.564637	1.682700	2.326323

Output 6.3. Model G2. VIF dels regressors

Podem observar VIF elevats pel que fa a les dummies corresponents a la tipologia, planta i, com ja havíem esmentat en observar coeficients negatius, número d'habitacions. Si el VIF és més gran que 5 o 10, pot ésser indici de que els coeficients associats als regressors no són del tot confiables. En el nostre cas, només la dummy HAB4 presenta un VIF superior a 5, però la resta de dummies referents a les habitacions, ATTIC i PENTHOUSE voregen el llindar. L'explicació a la presència de multicol·linealitat respecte al número d'habitacions es troba en la seva relació de dependència amb la superfície al ser ambdues dues variables de dimensionalitat. En canvi, encara és més clara la relació entre ATTIC i PENTHOUSE, ja que la primera recull els pisos que s'ubiquen a l'última planta del immoble i la segona a la tipologia àtic, que engloba immobles que s'ubiquen a la planta més elevada. Valorarem més endavant com procedir enfront això.

Un cop finalitzat el diagnòstic del nostre model G2, resumirem en una taula les conclusions extretes i les propostes a implementar a modus de correcció i pal·liatiu.

Supòsit	Validació	Proposta
Normalitat	Els residus no es distribueixen segons una normal	Transformació logarítmica
Linealitat	Possibles indicis de no linealitat	Regressió robusta + Valorar altres aproximacions
Homoscedasticitat	No es compleix la variància constant dels residus	Regressió robusta
Outliers influents	Detectem 3 valors atípics candidats a ésser influents	Regressió robusta + Eliminació dels valors atípics influents
Multicol·linealitat	Indicis de multicol·linealitat en presentar 4 regressors un VIF superior a 5	Valorar el replantejament dels nivells de la variable HAB i de les dummies ATTIC i PENTHOUSE

Taula 6.1. Conclusions del diagnòstic de compliment de supòsits fonamentals per part del nostre model G2

Seguidament, procedirem seqüencialment, a abordar els supòsits incomplerts i aplicar i avaluar les propostes suggerides.

Primerament, respecte a la normalitat dels residus, durem a terme la transformació logarítmica de la nostra variable resposta i la superfície (M2). Utilitzarem les mateixes variables que en el nostre primer model i novament procedirem a efectuar l'estimació per Mínims Quadrat Ordinaris (OLS).

```
Call:
lm(formula = LOG_PRICE ~ ., data = base2)

Residuals:
     Min       1Q   Median       3Q      Max
-1.50646 -0.22304 -0.00809  0.21997  1.37431

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.115837   0.085135  83.583  < 2e-16 ***
ACCESSIBLE1       0.001165   0.017657   0.066 0.947388
ELEVATOR1         0.243800   0.014204  17.164  < 2e-16 ***
EXTERIOR1        -0.012315   0.015052  -0.818 0.413316
GROUND_FLOOR1    -0.119252   0.024942  -4.781 1.79e-06 ***
ATTIC1            0.023754   0.040435   0.587 0.556917
REFORM1          -0.023194   0.014290  -1.623 0.104648
NEW1              0.132041   0.030400   4.343 1.43e-05 ***
FITTED_WARDROBE1  0.044184   0.010917   4.047 5.26e-05 ***
SPLIT1            0.121541   0.011166  10.885  < 2e-16 ***
TERRACE1          0.027762   0.011424   2.430 0.015123 *
GARDEN1          -0.027447   0.022222  -1.235 0.216831
POOL1             0.123417   0.023409   5.272 1.41e-07 ***
STORAGE1          0.034937   0.013708   2.549 0.010845 *
GARAGE_INCLUDED1 -0.014292   0.014067  -1.016 0.309672
PENTHOUSE1        0.164583   0.046273   3.557 0.000379 ***
DUPLEX1           0.030903   0.037826   0.817 0.413986
HOUSE1            0.067647   0.035141   1.925 0.054283 .
STUDIO1          -0.048064   0.052879  -0.909 0.363420
HAB21            -0.217156   0.022832  -9.511  < 2e-16 ***
HAB31            -0.368110   0.021787 -16.896  < 2e-16 ***
HAB41            -0.346889   0.024403 -14.215  < 2e-16 ***
NO_BANK1          0.213371   0.038849   5.492 4.17e-08 ***
BATHROOMS21       0.169905   0.013175  12.896  < 2e-16 ***
BATHROOMS31       0.384419   0.023051  16.677  < 2e-16 ***
LOG_M2            1.043610   0.018469  56.507  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3336 on 4933 degrees of freedom
  (8 observations deleted due to missingness)
Multiple R-squared:  0.7738,    Adjusted R-squared:  0.7726
F-statistic: 674.9 on 25 and 4933 DF,  p-value: < 2.2e-16
```

Output 6.4. Resum Model G3-1

Podem observar que el nostre R^2 ajustat millora lleugerament i que la variable BANK ha passat a presentar efectes significatius en la determinació del preu de l'habitatge. Ara bé, no ens avançarem en extreure conclusions, ja que ens resta abordar els altres supòsits, sinó que ens centrarem en avaluar la normalitat dels residus sota el nou escenari. Per tant, passem a representar el Q-Q plot corresponent:

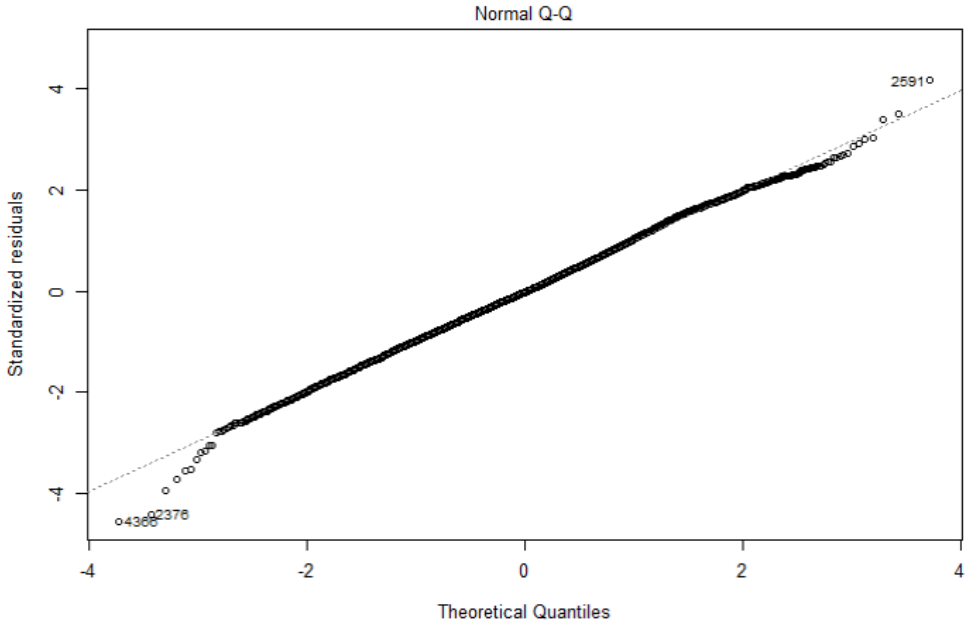


Figura 6.5. Model G3-1. Q-Q plot dels residus estandarditzats

Podem observar com amb la transformació logarítmica els residus s'ajusten molt millor a una normal, exceptuant en els extrems i intensificat per la presència de valors atípics. Corroborem amb el test de normalitat de Shapiro-Wilk, obtenint un p-valor molt major que sense la transformació, però encara inferior a 0.05 i, per tant, no podem acceptar la hipòtesi nul·la de normalitat. Tanmateix, i a falta d'excloure els outliers influents, no sembla que tinguem un problema preocupant quant a l'assumpció de normalitat.

```
        Shapiro-Wilk normality test

data:  model31$residuals
W = 0.99861, p-value = 0.0002743
```

Tot seguit, passarem a aplicar la regressió robusta, incloent la transformació logarítmica.

```
Call:
lmrob(formula = LOG_PRICE ~ ., data = base2, fast.s.large.n = Inf)
 \--> method = "MM"
Residuals:
      Min        1Q    Median        3Q       Max
-1.559779 -0.221503 -0.008802  0.221057  1.362591

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.016230   0.106828  65.678  < 2e-16 ***
ACCESSIBLE1       0.002191   0.017873   0.123 0.902438
ELEVATOR1         0.235950   0.016972  13.902  < 2e-16 ***
EXTERIOR1        -0.017655   0.014916  -1.184 0.236625
GROUND_FLOOR1    -0.111942   0.029372  -3.811 0.000140 ***
ATTIC1            0.015759   0.034921   0.451 0.651810
REFORM1          -0.013783   0.015849  -0.870 0.384547
NEW1              0.127371   0.025269   5.041 4.81e-07 ***
FITTED_WARDROBE1  0.041436   0.011125   3.724 0.000198 ***
SPLIT1            0.123719   0.011180  11.066  < 2e-16 ***
TERRACE1          0.025586   0.012069   2.120 0.034058 *
GARDEN1          -0.011889   0.022939  -0.518 0.604275
POOL1             0.118028   0.021173   5.574 2.62e-08 ***
STORAGE1          0.033737   0.013897   2.428 0.015235 *
GARAGE_INCLUDED1 -0.009677   0.013650  -0.709 0.478415
PENTHOUSE1        0.169354   0.041627   4.068 4.81e-05 ***
DUPLEX1           0.031884   0.031312   1.018 0.308599
HOUSE1            0.055957   0.047267   1.184 0.236536
STUDIO1          -0.060150   0.069928  -0.860 0.389740
HAB21            -0.254799   0.025639  -9.938  < 2e-16 ***
HAB31            -0.415192   0.024652 -16.842  < 2e-16 ***
HAB41            -0.405647   0.029293 -13.848  < 2e-16 ***
NO_BANK1          0.206841   0.048542   4.261 2.07e-05 ***
BATHROOMS21       0.167791   0.013774  12.182  < 2e-16 ***
BATHROOMS31       0.382715   0.024855  15.398  < 2e-16 ***
LOG_M2            1.078091   0.023858  45.189  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Robust residual standard error: 0.3236
  (8 observations deleted due to missingness)
Multiple R-squared:  0.7798,    Adjusted R-squared:  0.7786
Convergence in 13 IRWLS iterations
```

Output 6.5. Resum Model G3-2

Passem a avaluar novament l’homoscedasticitat amb un gràfic de dispersió entre les arrels dels residus i els valors ajustats.

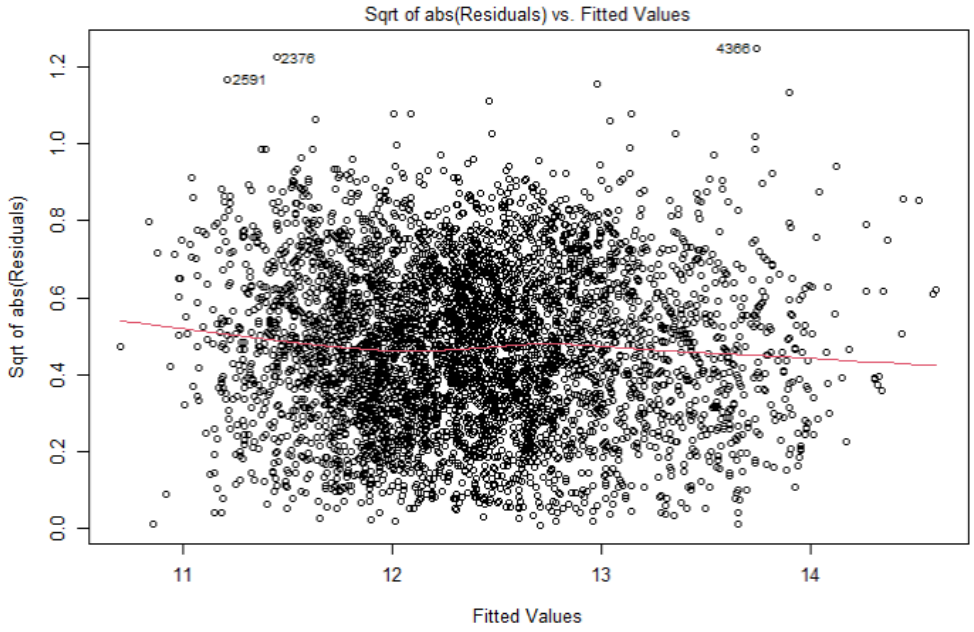


Figura 6.6. Model G3-2. Gràfic de dispersió de les arrels dels residus vs valors ajustats

Podem observar una millora considerable en termes de variància constant dels residus i de linealitat respecte al model G2. Observem alhora 3 valors atípics (observacions 2376, 2591 i 4366), també identificats en el Q-Q plot del model G3-1. Passem a eliminar-los de la nostra base juntament amb els 3 outliers que havíem detectat en el model G2 i que ara no apareixen com influents. Tanmateix, la regressió robusta en si ja minimitza la distorsió resultant de la presència d'outliers.

Respecte a la multicol·linealitat, havíem vist indicis de dependència entre les dummies ATTIC i PENTHOUSE (tot i que, els VIF associats a aquestes no superaven el llindar de 5), així com, entre la superfície i el número d'habitacions del immoble. Descartem eliminar variables per trencar aquestes relacions, ja que les considerem imprescindibles pel seu efecte significatiu en la determinació del preu d'un habitatge. Com alternativa tenim el replantejament del tractament de les variables HAB i FLOOR, ambdues incloses com a qualitatives de 4 i 3 nivells, respectivament. Hem tornat a estimar el nostre model, però definint ambdues variables com discretes i no hem observat evidències de multicol·linealitat. Ara bé, no ens sembla alarmant el problema de multicol·linealitat, exceptuant de cara a la interpretació dels coeficients de les dummies d'habitacions, i prioritzant la usabilitat del nostre model final, no apliquem canvis sobre aquest, optant per assumir el grau de col·linealitat present.

Així doncs, passem a presentar el nostre model final, eliminant els 4 regressors amb paràmetres no significatius (ACCESSIBLE, EXTERIOR, GARDEN, GARAGE_INCLUDED).

$$LOG\_PRICE = \beta_0 + \beta_1 HOUSE + \beta_2 STUDIO + \beta_3 DUPLEX + \beta_4 PENTHOUSE + \beta_5 LOG\_M2 + \beta_6 HAB2 + \beta_7 HAB3 + \beta_8 HAB4 + \beta_9 GROUND\_FLOOR + \beta_{10} ATTIC + \beta_{11} BATHROOMS\_2 + \beta_{12} BATHROOMS\_3 + \beta_{13} ELEVATOR + \beta_{14} REFORM + \beta_{15} NEW + \beta_{16} FITTED\_WARDROBE + \beta_{17} SPLIT + \beta_{18} TERRACE + \beta_{19} POOL + \beta_{20} STORAGE + \beta_{21} NO\_BANK + \epsilon$$

Fórmula 6.3. Equació del Model G-Final, on Ɛ correspon al terme de pertorbació

```
Call:
lmrob(formula = LOG_PRICE ~ ., data = base5, fast.s.large.n = Inf)
 \--> method = "MM"
Residuals:
      Min        1Q    Median        3Q       Max
-1.338977 -0.221555 -0.008465  0.222346  1.164601

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.00201    0.10497  66.707  < 2e-16 ***
ELEVATOR1         0.23234    0.01676  13.864  < 2e-16 ***
GROUND_FLOOR1    -0.11416    0.02895  -3.944 8.13e-05 ***
ATTIC1            0.01468    0.03489   0.421 0.673893
REFORM1          -0.01484    0.01583  -0.938 0.348532
NEW1              0.13069    0.02494   5.241 1.66e-07 ***
FITTED_WARDROBE1  0.03977    0.01106   3.595 0.000328 ***
SPLIT1            0.12243    0.01114  10.989  < 2e-16 ***
TERRACE1          0.02375    0.01200   1.980 0.047795 *
POOL1             0.10911    0.01762   6.192 6.41e-10 ***
STORAGE1          0.02991    0.01360   2.200 0.027873 *
PENTHOUSE1        0.17037    0.04161   4.095 4.29e-05 ***
DUPLEX1           0.03222    0.03114   1.035 0.300855
HOUSE1            0.05004    0.04718   1.061 0.288939
STUDIO1          -0.06011    0.06975  -0.862 0.388822
HAB21            -0.25502    0.02559  -9.964  < 2e-16 ***
HAB31            -0.41521    0.02448 -16.960  < 2e-16 ***
HAB41            -0.40659    0.02903 -14.004  < 2e-16 ***
NO_BANK1          0.19405    0.04669   4.156 3.29e-05 ***
BATHROOMS21       0.16573    0.01365  12.139  < 2e-16 ***
BATHROOMS31       0.37814    0.02465  15.337  < 2e-16 ***
LOG_M2            1.08190    0.02343  46.185  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Robust residual standard error: 0.3237
  (8 observations deleted due to missingness)
Multiple R-squared:  0.7797,    Adjusted R-squared:  0.7788
Convergence in 13 IRWLS iterations
```

Output 6.6. Resum Model G-Final

El nostre model general final presenta un R^2 alt que s'eleva a 0.779. Per tant, el model és capaç d'explicar el 77.9% de la variabilitat observada en la configuració del preu d'un habitatge.

Tot seguit, és moment d'interpretar els coeficients estimats i extreure conclusions. En primer lloc, quant als signes dels paràmetres, deixant de banda els ja comentats de les

dummies que conformen el número d'habitacions, no trobem incongruències amb el que esperàvem. Els únics que presenten signe negatiu són GROUND_FLOOR, REFORM i STUDIO. És a dir, que un habitatge estigui ubicat en una planta baixa comporta una disminució del preu enfront al mateix immoble ubicat en una planta superior, com també, que un immoble es trobi per reformar implica una caiguda del preu respecte a que estigui en bones condicions o sigui de nova construcció i que es tracti d'un estudi comporta també un preu menor que si fos un pis. D'altra banda, podem observar que un augment de la superfície en un 10.82% es trasllada en un increment d'un 10% en el preu de compra del immoble. A més, advertim, per exemple, que la disponibilitat d'ascensor dilata el preu de l'habitatge en un 26.15%, mentre que, el fet de que el immoble ocupi una planta baixa suposa una caiguda del preu en un 10.79 respecte a que se situï en una planta intermèdia.

Havent obtingut el model final de la primera línia de plantejament que posàvem sobre la taula i que recordem buscava no incloure variables de localització per evitar així restringir la seva utilitat a una àrea donada (en el nostre cas, València Ciutat); passem a incloure el factor districte (ZONE). Pel que fa al barri (SUBZONE), l'excloem en presentar un gran nombre de nivells amb una major desigualtat en la distribució d'observacions, fet que ens portaria a agrupar barris i complicaria, sense garanties d'aportar un valor afegit, en excés el nostre model.

Per una altra banda, com comentàvem al inici de l'apartat, tindríem una alternativa a introduir els districte de la ciutat. Com hem comprovat a l'anàlisi exploratori, existeix una clara correlació entre el preu mitjà de l'habitatge i la renda neta mitjana anual per llar per zones. Aprofitant això podríem agrupar els districtes per rangs de rendes en conjunts i desenvolupar el corresponent model que, molt probablement millorant l'ajust del primer, no presentaria la limitació de cenyir-se a una zona concreta. Ara bé, no desenvoluparem aquest plantejament, perquè implicaria una dependència del nostre projecte amb una variable externa, amb una segona font de dades. Una altra opció podria girar entorn a dibuixar anells respecte al centre, basant-nos en la teoria d'un major preu en zones cèntriques respecte a la perifèria, però podria esdevenir un problema si tan sols una zona no ho compleix i presenta un preu mitjà molt major o menor a l'esperat.

Tot seguit, passem al segon enfocament, on partirem de l'anàlisi del compliment dels supòsits fonamentals que hem dut a terme amb el model general i, per tant, utilitzarem tant la transformació logarítmica del preu i la superfície com la regressió robusta.

Així doncs, ajustem la regressió robusta utilitzant les mateixes variables que en la línia de modelització general agregant el factor ZONE. El nivell basal per la nostra variable de localització serà el districte amb un preu mitjà de l'habitatge menor, Rascanya.

```
Call:
lmrob(formula = LOG_PRICE ~ ., data = base6, fast.s.large.n = Inf)
 \--> method = "MM"
Residuals:
      Min        1Q    Median        3Q       Max
-1.387605 -0.154600 -0.001047  0.155593  1.051588

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.888878   0.080927  97.482  < 2e-16 ***
ACCESSIBLE1        0.007448   0.012543   0.594  0.55270
ELEVATOR1          0.212950   0.011522  18.482  < 2e-16 ***
EXTERIOR1         -0.015290   0.011548  -1.324  0.18554
GROUND_FLOOR1     -0.087493   0.021358  -4.097 4.26e-05 ***
ATTIC1             0.047403   0.029123   1.628  0.10365
REFORM1           -0.107319   0.010591 -10.133  < 2e-16 ***
NEW1               0.217720   0.022744   9.573  < 2e-16 ***
FITTED_WARDROBE1   0.024769   0.007801   3.175  0.00151 **
SPLIT1             0.075230   0.007878   9.549  < 2e-16 ***
TERRACE1           0.036973   0.008547   4.326 1.55e-05 ***
GARDEN1            0.011025   0.018331   0.601  0.54755
POOL1              0.224859   0.019916  11.291  < 2e-16 ***
STORAGE1           0.032890   0.010086   3.261  0.00112 **
GARAGE_INCLUDED1   0.088989   0.010638   8.365  < 2e-16 ***
PENTHOUSE1         0.137584   0.034014   4.045 5.31e-05 ***
DUPLEX1            0.047535   0.028404   1.674  0.09429 .
HOUSE1             0.360475   0.034163  10.552  < 2e-16 ***
STUDIO1           -0.054852   0.043290  -1.267  0.20518
HAB21             -0.130206   0.018312  -7.110 1.32e-12 ***
HAB31             -0.179333   0.018492  -9.698  < 2e-16 ***
HAB41             -0.220691   0.020820 -10.600  < 2e-16 ***
NO_BANK1           0.078809   0.045565   1.730  0.08376 .
BATHROOMS21        0.131614   0.009228  14.263  < 2e-16 ***
BATHROOMS31        0.252010   0.018164  13.874  < 2e-16 ***
ALGIROS1           0.418567   0.021168  19.774  < 2e-16 ***
BENICALAP1         0.056508   0.019655   2.875  0.00406 **
BENIMACLET1        0.382295   0.030572  12.505  < 2e-16 ***
CAMINS_AL_GRAU1    0.361826   0.019601  18.460  < 2e-16 ***
CAMPANAR1          0.294773   0.024923  11.828  < 2e-16 ***
CIUTAT_VELLA1      0.739917   0.019273  38.390  < 2e-16 ***
EL_PLA_DEL_REAL1   0.599684   0.023979  25.009  < 2e-16 ***
EXTRAMURS1         0.457330   0.019423  23.546  < 2e-16 ***
JESUS1             0.053038   0.019473   2.724  0.00648 **
L_EIXAMPLE1        0.805040   0.019135  42.071  < 2e-16 ***
L_OLIVERETA1       0.041548   0.020127   2.064  0.03904 *
LA_SAIDIA1         0.218706   0.022224   9.841  < 2e-16 ***
PATRAIX1           0.139070   0.020638   6.739 1.78e-11 ***
POBLATS_MARITIMS1  0.320201   0.020201  15.851  < 2e-16 ***
QUATRE_CARRERES1   0.232908   0.020476  11.375  < 2e-16 ***
LOG_M2             0.815144   0.016989  47.981  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Robust residual standard error: 0.2279
  (8 observations deleted due to missingness)
Multiple R-squared:  0.8863,    Adjusted R-squared:  0.8853
Convergence in 14 IRWLS iterations
```

Output 6.7. Resum Model NG-1

Podem observar com els factors ACCESSIBLE, EXTERIOR i GARDEN no es presenten com a regressors amb una influència significativa, com també es donava amb la modelització general. Passem a excloure aquestes 3 variables, a les qual sumarem el factor BANK al comprovar que deixa de tenir un coeficient associat significatiu, per obtenir el nostre model final.

$$LOG\_PRICE = \beta_0 + \beta_1 HOUSE + \beta_2 STUDIO + \beta_3 DUPLEX + \beta_4 PENTHOUSE$$
$$+\beta_5 LOG\_M2 + \beta_6 HAB2 + \beta_7 HAB3 + \beta_8 HAB4 + \beta_9 GROUND\_FLOOR + \beta_{10} ATTIC$$
$$+\beta_{11} BATHROOMS\_2 + \beta_{12} BATHROOMS\_3 + \beta_{13} ELEVATOR + \beta_{14} REFORM + \beta_{15} NEW$$
$$+\beta_{16} FITTED\_WARDROBE + \beta_{17} SPLIT + \beta_{18} TERRACE + \beta_{19} POOL + \beta_{20} STORAGE + \beta_{21} GARAGE\_INCLUDED$$
$$+\beta_{22} ALGIROS + \beta_{23} BENICALAP + \beta_{24} BENIMACLET + \beta_{25} CAMINS\_AL\_GRAU + \beta_{26} CAMPANAR$$
$$+\beta_{27} CIUTAT\_VELLA + \beta_{28} EL\_PLA\_DEL\_REAL + \beta_{29} EXTRAMURS + \beta_{30} JESUS + \beta_{31} L\_EIXAMPLE$$
$$+\beta_{32} L\_OLIVERETA + \beta_{33} LA\_SAIDIA + \beta_{34} PATRAIX + \beta_{35} POBLATS\_MARITIMS + \beta_{36} QUATRE\_CARRERES + \epsilon$$

Fórmula 6.4. Equació del Model NG-Final, on Ɛ correspon al terme de pertorbació

```
Call:
lmrob(formula = LOG_PRICE ~ ., data = base7, fast.s.large.n = Inf)
 \--> method = "MM"
Residuals:
      Min        1Q    Median        3Q       Max
-1.381662 -0.154464 -0.002053  0.155598  1.051375

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        7.953310   0.067986 116.985  < 2e-16 ***
ELEVATOR1          0.213487   0.011455  18.636  < 2e-16 ***
GROUND_FLOOR1     -0.084955   0.020918  -4.061 4.95e-05 ***
ATTIC1             0.046982   0.029151   1.612 0.107098
REFORM1           -0.107098   0.010586 -10.117  < 2e-16 ***
NEW1               0.218452   0.022916   9.533  < 2e-16 ***
FITTED_WARDROBE1   0.025568   0.007739   3.304 0.000961 ***
SPLIT1             0.075902   0.007845   9.675  < 2e-16 ***
TERRACE1           0.037642   0.008548   4.404 1.09e-05 ***
POOL1              0.228541   0.018094  12.631  < 2e-16 ***
STORAGE1           0.033471   0.010021   3.340 0.000844 ***
GARAGE_INCLUDED1   0.089631   0.010608   8.449  < 2e-16 ***
PENTHOUSE1         0.137595   0.034054   4.041 5.41e-05 ***
DUPLEX1            0.048167   0.028351   1.699 0.089390 .
HOUSE1             0.358150   0.033954  10.548  < 2e-16 ***
STUDIO1           -0.052488   0.042994  -1.221 0.222219
HAB21             -0.129884   0.018264  -7.112 1.31e-12 ***
HAB31             -0.177884   0.018415  -9.660  < 2e-16 ***
HAB41             -0.219532   0.020748 -10.581  < 2e-16 ***
BATHROOMS21        0.130491   0.009190  14.199  < 2e-16 ***
BATHROOMS31        0.250330   0.018070  13.853  < 2e-16 ***
ALGIROS1           0.419385   0.021077  19.898  < 2e-16 ***
BENICALAP1         0.057075   0.019654   2.904 0.003701 **
BENIMACLET1        0.380143   0.030389  12.509  < 2e-16 ***
CAMINS_AL_GRAU1    0.362535   0.019597  18.499  < 2e-16 ***
CAMPANAR1          0.295208   0.024813  11.897  < 2e-16 ***
CIUTAT_VELLA1      0.740140   0.019125  38.699  < 2e-16 ***
EL_PLA_DEL_REAL1   0.599277   0.023942  25.031  < 2e-16 ***
EXTRAMURS1         0.460537   0.019308  23.852  < 2e-16 ***
JESUS1             0.052150   0.019421   2.685 0.007271 **
L_EIXAMPLE1        0.805877   0.019014  42.383  < 2e-16 ***
L_OLIVERETA1       0.041510   0.019899   2.086 0.037030 *
LA_SAIDIA1         0.218743   0.022279   9.818  < 2e-16 ***
PATRAIX1           0.139743   0.020628   6.775 1.39e-11 ***
POBLATS_MARITIMS1  0.319748   0.020037  15.958  < 2e-16 ***
QUATRE_CARRERES1   0.231694   0.020327  11.398  < 2e-16 ***
LOG_M2             0.814871   0.016916  48.170  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Robust residual standard error: 0.2277
  (8 observations deleted due to missingness)
Multiple R-squared:  0.8862,    Adjusted R-squared:  0.8853
Convergence in 11 IRWLS iterations
```

Output 6.8. Resum Model NG-Final

Esdevé molt destacable el R^2 ajustat que assolim i que s'eleva a 0.885, major que el corresponent al model general, com esperàvem després de comprovar de forma exhaustiva a l'anàlisi descriptiva les diferències de comportament del preu en funció de la ubicació del immoble. En conclusió, amb el nostre model no general aconseguim explicar el 88.5% de la variabilitat en la determinació del preu de l'habitatge a la ciutat de València.

Així doncs, d'igual forma que amb el model general, passem a interpretar els coeficients estimats i obtenir respostes. Respecte als signes dels paràmetres, observem que es repeteix el mateix escenari que al model no general, essent GROUND_FLOOR, REFORM i STUDIO, les úniques variables que presenten un coeficient associat negatiu, obviant novament les fictícies d'habitacions. Altrament, destaquen els coeficients associats al factor districte, essent tots positius a l'haver definit Rascanya, zona amb el menor preu mitjà de l'habitatge, com basal. Tenim, per exemple, el fet que un immoble se situï a l'Eixample comporta un augment en el seu preu d'un 123.87% respecte a si estigués situat a Rascanya, mentre que, ho faria només en un 4.24% si el reubiquéssim de Rascanya a L'Olivereta. Quant al pes d'un augment de la superfície, aquest hauria d'elevar-se a un 8.15% per obtenir un conseqüent increment del 10% en el preu de l'habitatge.

Com a darrer punt, restaria fer un aclariment respecte a les interaccions, com és el cas de la disponibilitat d'ascensor amb la planta on s'ubica el immoble, que hem presentat a l'anàlisi descriptiva. Hem decidit no introduir interaccions en els nostres models al fer balanç del benefici i l'augment de complexitat que se'n derivaria.

6.2 Capacitat predictiva

Arribats a aquest punt en que hem seleccionat i presentat els nostres dos models finals, ens restaria analitzar les seves corresponents capacitats predictives sobre observacions que no haguem utilitzat en la seva estimació, amb la idea de comprovar que no hi ha problemes de sobreajustament i generalitzin de forma adequada davant noves observacions. Aquí radica el motiu de que haguem aplicat l'estratègia de dividir aleatòriament la nostra base en un grup d'entrenament i un de test (recordem que en proporcions 0.7 i 0.3, respectivament).

En primer lloc, generarem prediccions per ambdós models utilitzant les dades de test. Amb aquestes i els respectius valors observats prosseguim a calcular l'error quadràtic mig (RMSE), l'error absolut mig (MAE) i el R^2 (entès com el quadrat de la correlació entre els valors observats i els predits) per a cada model. Altrament, en ambdós plantejaments hem aplicat la transformació logarítmica, per tant, de cara als valors absoluts, aplicarem exponencial sobre els valors predits i la resposta LOG_PRICE.

Model	RMSE	MAE	Rsquared
G-Final	121866	77848	0.7412
NG-Final	99674	59758	0.8228

Taula 6.2. RMSE, MAE i R^2 dels models G-Final i NG-Final

Com esperàvem, el model NG-Final presenta menors RMSE i MAE i un major R^2 que el model G-Final. És a dir, la inclusió del factor localització millora considerablement la capacitat predictiva del nostre model. Tanmateix, calculem l'error percentual absolut mig (MAPE), obtenint un 27.32% pel model general i un 19.65% pel model no general. Així doncs, podríem considerar el model acotat a la ciutat de València un candidat a bon model (MAPE < 20%), mentre que el model general deixa més que desitjar en termes de precisió respecte a les seves estimacions predictives.

Per concloure i fer-nos una imatge més clara del comportament de l'error predictiu en els models, passarem a representar un gràfic de dispersió dels valors predits enfront els observats en el nostre grup de test.

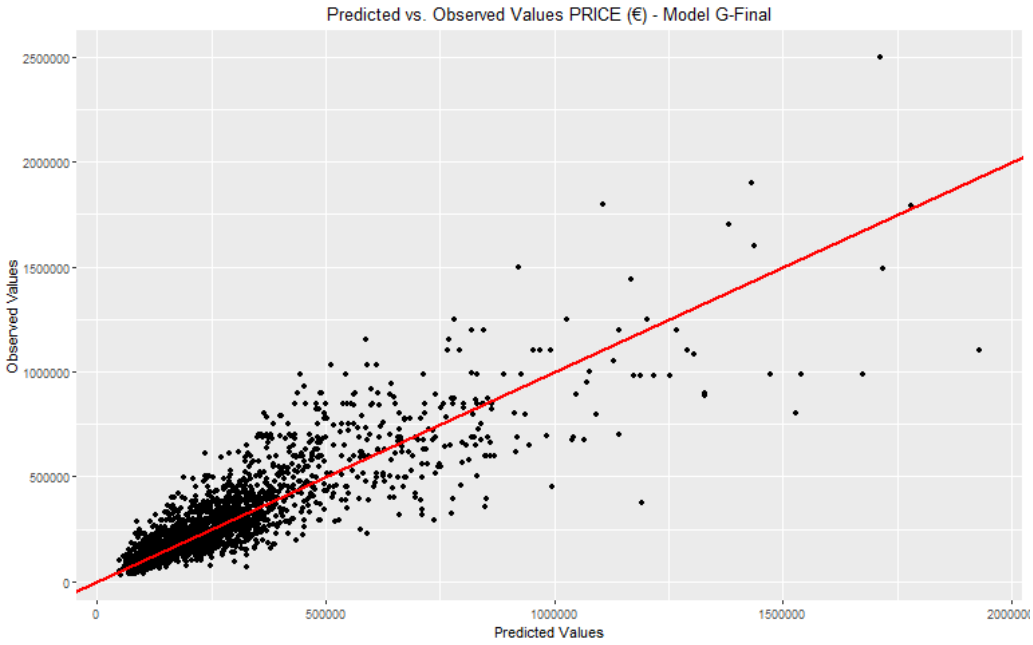


Figura 6.7. Model G-Final. Gràfic de dispersió dels valors predits vs els observats

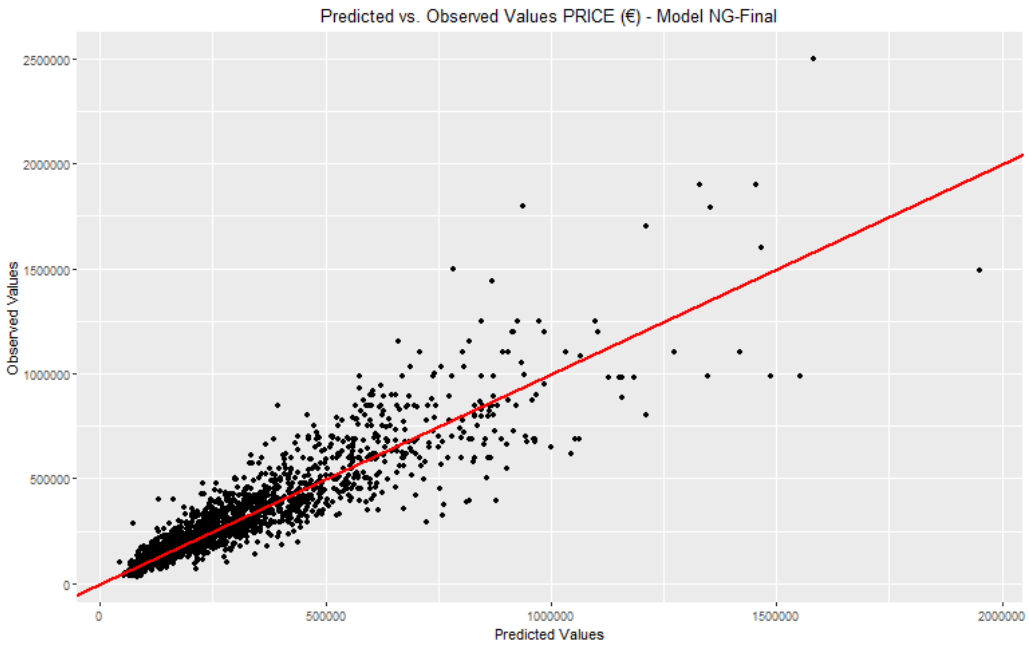


Figura 6.8. Model NG-Final. Gràfic de dispersió dels valors predits vs els observats

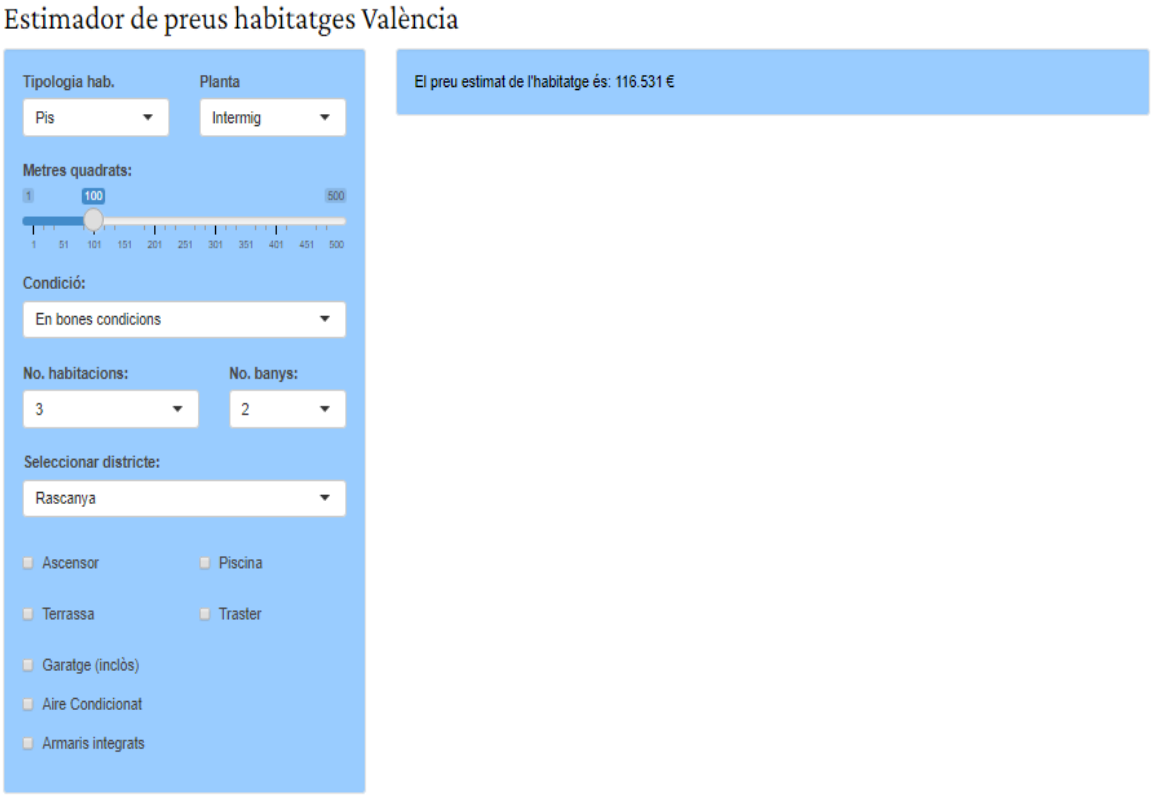
En ambdues visualitzacions, a banda d'una millora amb el model NG-Final, podem apreciar un patró clar, que no esdevé una sorpresa, i és un major error de predicció en immobles amb un preu més elevat.

VII. APLICACIÓ SHINY: ESTIMADOR DE PREUS D'HABITATGE VALÈNCIA CIUTAT

Un cop hem finalitzat amb les dos línies de modelització presentades, correspon productivitzar el treball dut a terme. És per aquesta raó que hem acoblat una aplicació que permet al usuari, tot introduint les dades definides, obtenir una estimació del preu de l'habitatge. El consumidor podrà veure, si ho fa sobre un immoble que li interessi, si aquest està sobrevalorat o, pel contrari, si és una bona oportunitat de mercat; si, en canvi, no té cap immoble concretitzat, però si té en ment quines característiques hauria de tenir, podrà visualitzar el preu que hauria de pagar.

Així doncs, amb aquesta idea en ment, hem utilitzat Shiny per crear l'aplicació. Com a motor per aquesta, hem emprat el nostre model no general, per tant, es tracta d'una eina d'ús restringit a València ciutat.

Que una aplicació tingui vida o no depèn senzillament de que se'n faci ús. Per tal de que això es compleixi, hem dissenyat la interfície perquè sigui fàcil i ràpid de manejar, customitzant i ordenant les variables d'entrada i escollint el botó interactiu més adequat per cada una d'elles.



Imatge 7.1. Interfície de l'aplicació Shiny

VIII. CONCLUSIONS

És moment de refrescar els objectius cabdals del nostre projecte i avaluar en quina mesura hem pogut assolir-los, alhora que passem llistat al cúmul de claus que hem extret al llarg del procés.

Primerament, la principal meta del present projecte és dual: mentre que hem dissenyat i desenvolupat un flux de treball, sota la premissa d'avantposar l'automatització, que nodrint-se de les publicacions d'anuncis d'un portal immobiliari ens permet obtenir una fotografia nítida i fidedigna del comportament del preu de l'habitatge a una àrea territorial concreta, així com la seva modelització i ús predictiu; hem trencat amb la font de dades convencional, dibuixant, si més no, noves possibilitats de treball com també estratègies i mitjans per materialitzar-les, sempre i quan tinguem dades publicades en algun racó de la xarxa.

Així doncs, hem buscat doblegar la limitació que suposa, tot i el paradigma actual d'informació, aconseguir disposar d'una base de dades de garanties. I és precisament aquí on radica el motiu de que haguem tingut sempre actius dos termòmetres: automatització i qualitat de dades; el pes dels quals s'ha traslluït en el plantejament i les decisions preses en el transcurs del projecte. En referència a l'automatització, alhora entenent aquesta en clau d'eficiència i viabilitat, és obvi que on ha pres un major pes, per la dificultat suscitada, ha estat en la construcció de l'algoritme que alimenta la nostra base de dades. Hem hagut d'esprémer la tècnica de *web scraping* per intentar pouar les dades desitjades, minimitzant els temps d'execució alhora que salvaguardant-nos de ser detectats pel portal. Aquí hem de valorar positivament que, exceptuant la descripció completa, el text de característiques bàsiques i el certificat energètic del immoble, hem pogut extreure totes les variables que havíem considerat d'un inici. D'altra banda, hem calculat que els temps operatius per extreure les dades d'una zona oscil·len entre els 40 i 55 minuts (25-35 minuts d'execució de l'algoritme + 15-20 minuts de revisió de logs i fitxers obtinguts), trobant-se el major cost en temps en el número de zones i no en el volum d'aquestes. Sota la nostra experiència, són unes bones xifres, però no podem negar que esdevindrien limitants si volguéssim augmentar la freqüència d'extracció, per exemple, a diària, existint encara un gran potencial de millora. Respecte a l'altre estendard del nostre projecte, la qualitat de dades, s'ha composat per dos blocs: el *data cleaning* i l'anàlisi descriptiva. Mentre que el *data cleaning* suposa el primer tractament de les dades bombejades per l'algoritme, on, detectant patrons, hem corregit tant desajustos resultants de la disposició de les dades a la interfície web com algun error puntual de l'algoritme , alhora que hem

definit i assegurat l'esquema de variables i els nivells que les configuren; a l'anàlisi descriptiva, no partint ja de la dada bruta, hem baixat a analitzar exhaustivament el comportament de les variables, servint-nos de feedback per detectar incongruències i aplicar les correccions necessàries, sempre i quan les incoherències seguissin patrons extrapolables i la solució es pogués automatitzar. Aquí afegir un punt més aplicat i que esdevé el contrastar les nostres observacions amb dades de fonts oficials, tot i que, a la pràctica no ha saltat cap alarma i allò que observàvem s'apropava gratament al que ens reportaven. Remarcar la importància que hem donat a garantir la qualitat de la nostra base dades, un temps invertit que a vegades sol infravalorar-se, però que de no consolidar-se, podria comportar efectes molt negatius sobre el potencial tant explicatiu com predictiu del nostre treball.

Centrem-nos ara en l'entrellat de conclusions que hem extret del propi exercici. L'esquelet d'aquest es traça en tres blocs: l'anàlisi descriptiva, la modelització i l'aplicació; que podríem traduir, respectivament, a exploració, materialització i productivització. En primer lloc, referent a l'anàlisi descriptiva, hem pogut contrastar que tant la nostra variable resposta com la superfície no es distribueixen segons una normal, confirmant a la modelització la superfície com el factor amb un major pes en la determinació del preu d'un immoble. Així mateix, hem vist com la tipologia amb un preu mitjà més elevat corresponia als àtics, com també es donava a nivell de planta on s'ubica el immoble. Un altre factor determinant ha esdevingut l'estat del immoble, amb el nivell REFORM com el de menor preu. D'altra banda, hem pogut comprovar que, com esperàvem, totes les variables dicotòmiques, exceptuant de la referent a si es tracta d'un immoble de banc, presentaven un major preu pel nivell positiu, desmarcant-se la disponibilitat d'ascensor com el factor amb un pes més significatiu sobre el preu i caient, és a dir, presentant a la modelització paràmetres associats no significatius, si el immoble és accessible, exterior i si gaudeix o no de jardí. Altres variables, com l'agència que publica l'anunci, el telèfon de contacte i la descripció, ens han servit per contrastar la coherència d'altres factors i per detectar observacions en situació especial, com seria que el immoble està ocupat o llogat, però no les hem inclòs dins dels models. Ara bé, hi ha un factor que el destaquem per sobre de la resta i esdevé la localització del immoble, no perquè sigui, darrera de la superfície, la variable amb una major influència sobre el preu, sinó perquè ens ha portat a plantejar dos línies de modelització en funció de la seva inclusió o no. Hem vist que dels dos models resultants, ambdós amb un R^2 ajustat molt alt, el general, o sense inclusió del factor districte, ajustava pitjor i presentava una capacitat predictiva menor que el no general. Tanmateix, hem presentat la renda neta mitjana anual per llar, molt correlacionada amb el districte, com un bon substitut d'aquest. Finalment, hem utilitzat el model no

general per crear una aplicació Shiny, dotant així d'una eina perquè qualsevol persona pugui utilitzar-la per saber el preu esperat del immoble que cerca.

IX. BIBLIOGRAFIA

Oficina d'Estadística Ajuntament de València. *Catastro inmobiliario 2021. Municipios del Área Metropolitana de València*, 2021. Disponible a: https://www.valencia.es/estadistica/CatPub/files/CatastroInmobAM/R%C3%BAstic%20i%20urb%C3%A0%202021.pdf.

Oficina d'Estadística Ajuntament de València. *Anuario Estadístico de la Ciudad de València*, 2021. Disponible a: https://www.valencia.es/estadistica/CatPub/files/Anuario2021_Conceptos.pdf.

UiPath Inc. Uipath Forum. Disponible a: https://forum.uipath.com/.

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Disponible a: https://www.R-project.org/.

Benoit, K. (2011). *Linear regression models with logarithmic transformations*. London School of Economics, London, 22(1), 23-36.

Ghosalkar, N. N., & Dhage, S. N. (2019). *Real estate value Prediction using linear regression*. In 2018 fourt international conference on computing communication control and automation (ICCUBEA) (pp. 1-5). IEEE.