TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Estudi de les diferències entre homes i dones en el nivell salarial d'accés a la jubilació

Autor: Miriam Quero Gramunt

Director: Montserrat Guillen Estany

Departament: Econometria, Estadística i Economia Aplicada

Convocatòria: Juny 2018

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

RESUM

El futur de les pensions és un debat actual en l'àmbit econòmic, social i polític. No només a Espanya, sinó també a la resta d'Europa, els països participen en la reforma dels processos per assegurar la sostenibilitat dels seus sistemes de pensions.

El sistema de pensions a Espanya és públic, obligatori, contributiu i de distribució. Pateix un desequilibri a causa de la reducció dels contribuents per cada beneficiari i pel procés d'envelliment generacional.

La *Muestra Contínua de Vidas Laborables* (MCVL) proporciona informació administrativa de més d'un milió de treballadors i pensionistes a Espanya, per tant, hi ha informació vàlida de tota la seva vida activa.

El propòsit d'aquest treball és analitzar la diferència entre l'últim salari i la primera pensió de jubilació i mirant entre d'altres la diferència que hi ha entre homes i dones. Per dur a terme l'anàlisi, s'utilitzarà el programari estadístic "R" i els paquets relacionats amb la gestió de grans bases de dades.

PARAULES CLAU

Taxa de substitució, pensions, cotització, salari, vida laboral

ABSTRACT

The future of pensions is a current debate in the economic, social and political spheres. Not only in Spain but also in the rest on Europe, the countries are involved in the processes reform in order to ensure the sustainability of their pension systems.

The pension system in Spain is, public, compulsory, contributory and of distribution. It suffers from an imbalance due to the reduction of contributors to each beneficiary and to the process of generational aging.

The MCVL provides administrative information from more than one million workers and pensioners in Spain, therefore, there is enough information of all their active lives.

The purpose of this paper is to analyze the difference between the last salary and the first retirement pension and looking among others for the difference between men and women. To carry out the analysis, the statistical software "R" and the packages related to the management of big databases will be used.

KEY WORDS

Replacement rate, pensions, quote, salary, labour records

Classificació AMS

62P05 *Applications to actuarial sciences and financial mathematics;* **62J05** *Linear regression*; **62D05** *Sampling theory, sample surveys*; **62-07** *Data analysis.*

Índex

INTRODUCCIÓ 3

1.1 LA MOSTRA CONTINUA DE VIDES LABORALS 3

1.2 ESTRUCTURA DE LA MCVL 4

1.3 DIAGRAMA DE LA BASE DE DADES 5

1.4 BASES DE COTITZACIÓ 6

1.5 LA TAXA DE SUBSTITUCIÓ 7

1.6 BRETXA SALARIAL 8

1.7 POBLACIÓ 9

1.8 OBJECTIUS 11

1.9 AGRAÏMENTS 12

DESCRIPCIÓ DELS FITXERS I LA BASE DE DADES 13

2.1 PROGRAMES UTILITZATS 13

2.2 TIPUS DE FITXERS I BASES DE DADES 13

2.3 FITXERS I TAULES 14

2.3.1 DESCRIPCIÓ DE LA TAULA: PERSONAS 14

2.3.2 DESCRIPCIÓ DE LA TAULA: AFILIACIÓN 17

2.3.3 DESCRIPCIÓ DE LA TAULA: BASES DE COTIZACIÓN 19

2.3.4 DESCRIPCIÓ DE LA TAULA: PENSIONES 21

2.3.5 DESCRIPCIÓ DE LA TAULA: CONVIVIENTES 26

2.3.6 DESCRIPCIÓ DE LA TAULA: DATOS FISCALES 26

ANÀLISI DESCRIPTIVA I COMPARACIÓ GRÀFICA 28

3.1 ANÀLISI DESCRIPTIVA DE LA MCVL 28

3.1.1 LECTURA I PROCESOS TAULA *PERSONAL* 28

3.1.2 LECTURA I PROCESOS TAULA *AFILIACIÓN* 30

3.1.3 LECTURA I PROCESOS TAULA *PENSIONES* 36

3.1.4 UNIÓ I PROCESSOS DE PERS_AFI_PREST 41

3.1.5 LECTURA I PROCESOS TAULA COTIZACIÓN 42

3.1.6 UNIÓ FINAL DE LES 4 TAULES 42

3.1.7 DEPURACIÓ DE P_A_P_C 43

3.1.8 CÀLCUL TAXA DE SUBSTITUCIÓ 44

3.1.9 DIAGRAMA DE DISPERSIÓ SIMPLE 45

3.2 ANÀLISI DE REGRESSIÓ 46

3.3 INDIVIDUS AMB MÉS DE 35 ANYS COTITZATS 49

3.4 COMPARATIVA MODELS 53

CONCLUSIONS 54

BIBLIOGRAFIA 56
ANNEXOS...57

INTRODUCCIÓ

La situació econòmica i financera del sistema públic de pensions a Espanya i la necessitat de garantir tant la seva suficiència com la seva sostenibilitat són temes recurrents en els debats sobre política econòmica. En els últims anys, l'augment dels dèficits pressupostaris de la Seguretat Social i la disminució dels recursos disponibles en el seu Fons de Reserva han provocat que l'interès en aquestes qüestions hagi augmentat.

Aquest treball té com a principal objectiu contribuir al debat sobre la taxa de substitució a Espanya, per veure si aquesta és tant elevada com habitualment es diu i si hi ha diferències entre sexes en quan a la relació de l'últim salari i la primera jubilació.

La relació entre l'últim salari i la primera pensió s'anomena taxa de substitució. L'interès d'entendre aquesta relació és el de trobar quins possibles factors la determinen. La dificultat és que cal utilitzar una font estadística molt complexa per a dur a terme aquestes estimacions.

En una noticia recent (Javier G. Jorrín, 2017, p. 1) afirma "*La tasa de sustitución del salario medio respecto a la pensión media en España se sitúa en el entorno del 60%. También es muy elevada la tasa de la primera pensión sobre el último salario, que supera el 80%, esto es, el trabajador sigue cobrando este porcentaje de su salario una vez se jubila.*"

1.1 LA MOSTRA CONTINUA DE VIDES LABORALS

La MCVL, *Muestra Contínua de Vidas Laborales, Ministerio de Trabajo y Seguridad Social, España*, és un conjunt de microdades individuals, però anònimes, extretes dels registres de la Seguretat Social. La informació de la Seguretat Social es completa amb informació fiscal procedent de l'AEAT i amb informació del padró continu facilitada per l'INE.

Per a poder realitzar aquest estudi, s'ha utilitzat la base de dades de la MCVL, la qual, es va haver de sol·licitar prèviament a la *Dirección General de Ordenación de la Seguridad Social* i on es va haver de firmar un document de "condicions per a la recepció i utilització dels fitxers de microdades de la Mostra Contínua de Vides Laborals" on ens mostràvem d'acord a facilitar, a la Seguretat Social, els resultats obtinguts.

Constitueix una mostra representativa de totes les persones que van cotitzar o van cobrar prestacions en un any determinat any en aquest sistema de protecció. És Contínua perquè s'actualitza anualment. Cada mostra, referida a la població en l'any de referència, reprodueix l'historial anterior de les persones seleccionades, remuntant cap enrere fins on es conservin registres informatitzats, i per això es denomina de vides laborals.

En aquest cas, les dades corresponen a una mostra de persones seleccionades a l'atzar entre els que van ser afiliats o pensionistes de la Seguretat Social durant l'any 2015. Per a cada

persona s'inclouen tant dades sobre la seva relació amb la Seguretat Social en l'any esmentat com dades històriques, sempre que aquestes figurin guardades al sistema.

Les dades es presenten sense gairebé cap modificació del seu estat en els registres informàtics i completament desagregades, de manera molt similar a com estan emmagatzemades en les bases de dades de les que procedeixen.

A més, la MCVL serveix per a estudis sobre diferents temes, entre els quals hi ha: la trajectòria laboral al llarg de la vida, mobilitat, probabilitat de trobar feina per als aturats, temporalitat, diferències salarials, estudis d'ocupació sectorials o locals, situació relativa d'homes i dones, immigració, integració laboral de minusvàlids, llicències per cura de fills, anàlisi del temps i la transició de l'ocupació a la jubilació, projecció de la despesa en pensions i la seva relació amb les previsions demogràfiques, efectes redistributius del sistema de pensions i la seva incidència espacial, i finalment, microsimulació de polítiques públiques.

En aquest treball s'analitzaran alguns d'aquests aspectes, fent més èmfasi en les persones de més de 45 anys.

1.2 ESTRUCTURA DE LA MCVL

En aquest estudi ha sigut molt important tenir clar com s'organitza la informació d'aquesta gran base de dades, ja que, la MCVL és formada per diversos arxius i té un pes de 1,19 GB repartit en 22 arxius en format ".txt", dels quals n'hem utilitzat 19. Aquesta informació s'estructura en sis taules:

- *Tabla 1: Personas*
- *Tabla 2: Afiliación (vida laboral)*
- *Tabla 3: Bases de Cotización*
- *Tabla 4: Pensiones*
- *Tabla 5: Convivientes*
- *Tabla 6: Datos Fiscales2 (retenciones IRPF)*

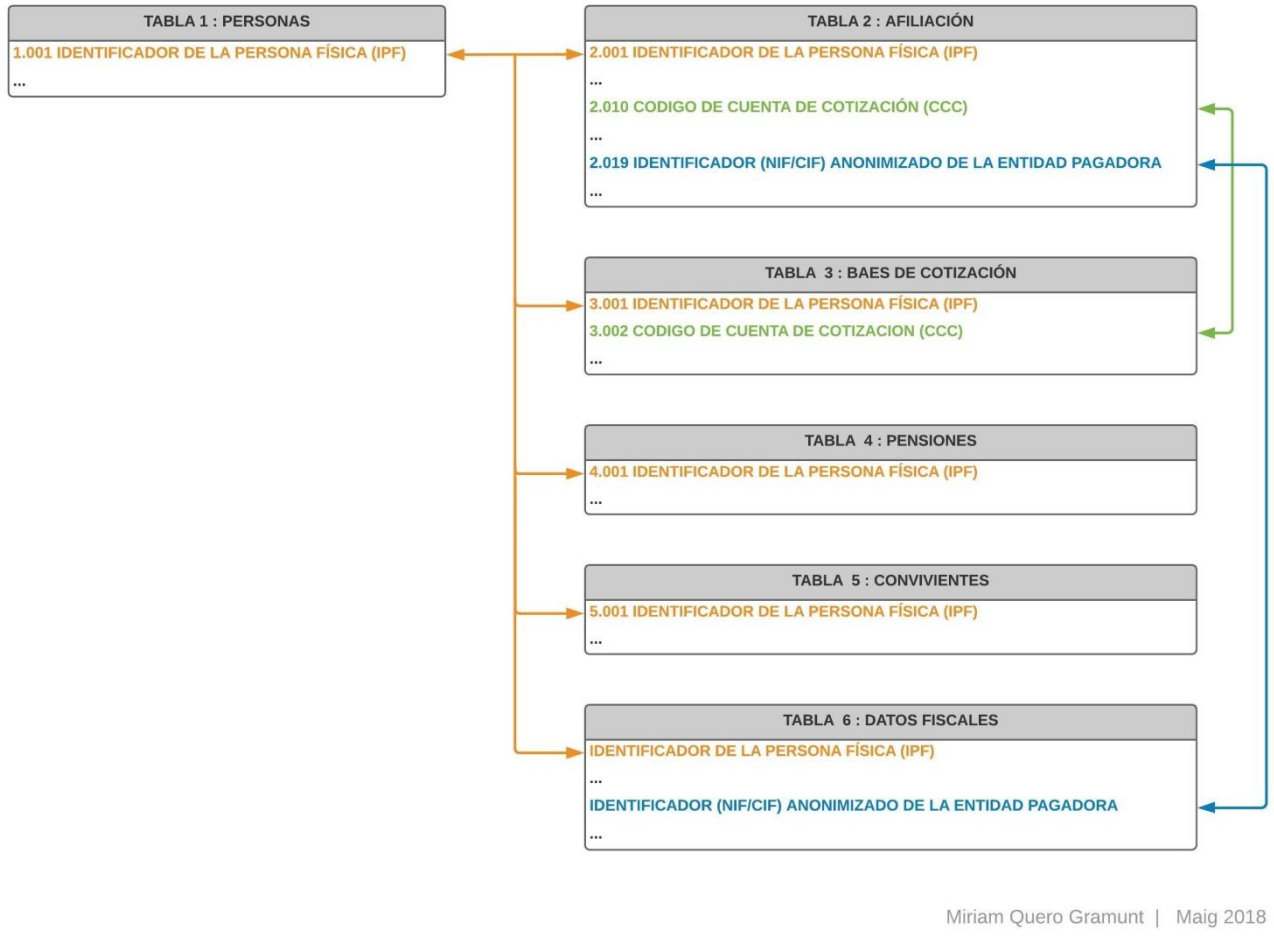
En cadascuna de les taules, la primera columna de cada fila conté una clau: l'identificador, sempre anònim, de la persona física (l'afiliat o pensionista) a la qual es refereixen totes les altres dades (columnes) de cada fila. Aquesta clau identifica a cada persona seleccionada per a la mostra, de manera que permet recuperar tota la informació corresponent a la mateixa persona, que està repartida entre les diferents taules. A part de l'identificador, hi ha altres variables comunes que permeten connectar les taules entre sí.

1.3 DIAGRAMA DE LA BASE DE DADES

S'ha considerat adequat construir aquest diagrama, per a poder representar i apreciar millor la complexitat de la base de dades i les relacions entre les diverses taules.

Es pot observar, que la variable identificadora *1.001 IDENTIFICADOR DE LA PERSONA FÍSICA* de la taula *PERSONAS,* està inclosa en les sis taules de la base de dades, fet possibilita la connexió entre elles.

A més de l'identificador de la persona física, a les taules hi ha altres dades o camps comuns que permeten connectar-les entre si, com és el cas de la variable *2.010 CÓDIGO DE CUENTA DE COTIZACIÓN (CCC)* que permet comunicar la taula *AFILIACIÓN* amb la *BASES DE COTIZACIÓN*. A més a més, la variable *2.019 IDENTIFICADOR (NIF/CIF) ANONIMADO DE LA ENTIDAD PAGADORA* permet enllaçar informació de la taula *AFILIACIÓN* amb la de *DATOS FISCALES*, tot i que aquesta última no s'hagi utilitzat en aquest treball.



Gràfic 1.1 Camps comuns en les taules de la MCVL. Font: Elaboració pròpia.

1.4 BASES DE COTITZACIÓ

La base de cotització és el valor de la remuneració mensual que reben els treballadors en quantitat bruta, en la qual, també s'hi reflecteixen les pagues extra prorratejades. Cada any, la llei estableix uns límits, en el valor als efectes de càlcul del què s'aportarà a la Seguretat Social, que considera oportuns ordenats per categories professionals i atorgant una quantitat mínima i màxima a cadascuna d'elles.

Perquè el treballador quedi protegit per la Seguretat Social s'abona una quota, part per l'empresari, part pel treballador, com a contribució a aquest sistema.

La quota pagada es determina d'acord amb la base de cotització del treballador. Un cop determinada la base de cotització se li apliquen uns percentatges que donen lloc a la quantitat que s'ha d'abonar per tal d'integrar-se al sistema de la Seguretat Social.

La base de cotització és un dels aspectes més importants en revisar les nòmines i la vida laboral.

Com més elevada sigui la base de cotització, més gran serà la quantitat que es percebrà en les prestacions i al revés. Quan aquesta és relativament baixa, també com a conseqüència ho seran l'atur, la prestació per incapacitat laboral, que s'anomena baixa o la jubilació. Aquests termes també són aplicats als règims del camp, mar, empleat de la llar i autònoms.

En concret, la base de cotització es calcula d'acord amb la remuneració total del treballador, amb independència de la seva forma i denominació, que mensualment rebi el treballador per compte d'altri, prorratejant les quantitats que es rebin per períodes superiors al mes. En tot cas, no es computen les dietes i assignacions per despeses de viatge i despeses de locomoció (sempre dins de certs límits).

Com es prorrategen les remuneracions que es reben amb una periodicitat superior al mes, les pagues extres queden integrades en les bases de cotització mensual del treballador. D'aquesta manera, cada treballador té 12 bases de cotitzacions a l'any.

Hi bases de cotització especial per als contractes a temps parcial i per als contractats per a la formació.

En tot cas, les bases de cotització mai podran ser superiors, ni inferiors, a certs límits establerts anualment en la Llei de Pressupostos per a cada exercici econòmic d'acord amb la categoria professional o activitat del treballador.

Encara que hi ha una proporcionalitat entre el salari i la base de cotització, no són el mateix. En aquest treball, però s'ha optat per considerar que són iguals, encara que caldria utilitzar la informació fiscal que ens proporciona la base de dades per a saber el sou exacte.

1.5 LA TAXA DE SUBSTITUCIÓ

La taxa de substitució també coneguda com a taxa de reemplaçament, és un indicador de com un sistema de pensions aconsegueix o no el seu objectiu de proporcionar uns ingressos adequats en el moment de la jubilació respecte als ingressos que el treballador tenia quan estava en actiu. Aquesta taxa es calcula com el percentatge que suposa la pensió de jubilació sobre l'últim sou percebut en l'etapa laboral.

Mikel De La Fuente Lavín (2004) afirma: "*La fórmula de cálculo de las pensiones, en especial las de jubilación, es relativamente generosa en el Estado Español comparada con otros ejemplos de la Unión Europea*" (p. 1).

Una de les preguntes que es planteja molta gent, és quina serà la seva pensió de jubilació. És una pregunta que acostuma a rondar pel cap una vegada s'apropa l'edat per a retirar-se professionalment. Per calcular-la s'ha de tindre en compte molts aspectes, entre els quals destaquen el temps que s'hagi cotitzat a la Seguretat Social i per quant s'ha fet. Però si el que es vol treure és una xifra aproximada, del que es pot ingressar a la pensió de jubilació, es

pot calcular la taxa de substitució, la qual donarà una idea amb una xifra aproximada del que serà la pensió de cada individu.

Aquesta taxa, és un indicador que mostra el que s'acabarà cobrant en la jubilació en relació a l'últim salari. La taxa de reemplaçament és la proporció de prestació de jubilació que s'obté en comparació amb l'últim sou, un càlcul que s'utilitza per saber el poder adquisitiu que es perd amb la jubilació i conèixer si s'haurà de complementar o no, la pensió.

Per exemple, si la taxa de substitució fos del 74% i en el moment de la jubilació d'un individu, l'últim sou és de 2.000 euros, es perdrà el 26%, de manera que quedaria una pensió de jubilació de 1.480 euros. Encara que, cal tenir en compte que aquesta xifra és una mitjana i que la nostra pensió de jubilació depèn de moltes altres coses.

Aquest fet es compliria si la pensió de jubilació es calculés amb la taxa de substitució, però no és du a terme d'aquesta manera. La veritat és que la pensió de jubilació es calcula en funció de la base reguladora que a la vegada es calcula amb les bases de cotització dels anys que corresponguin. Amb la reforma, aquests anys varien depenent de quan es jubili l'individu, abans es prenien els últims 15 anys (180 mesos). Fins que el 2022 serà obligatori calcular la base reguladora comptabilitzant els últims 25 anys que s'hagin cotitzat a la Seguretat Social, el còmput augmenta des de 2013 cada any en un any. D'aquesta manera, jubilar-se al 2015, any que es pren de referència en aquest treball, implica comptabilitzar els últims 18 anys que s'hagi cotitzat. Així, s'hauran de sumar les bases de cotització d'aquests 18 anys i dividir-les entre 252 per esbrinar la base reguladora.

Per tant, aquests fets i que cada persona es troba en una situació amb unes característiques diferents, donen a pensar que calcular la pensió amb l'últim any de salari no és suficient per a fer una bona predicció del que serà la pensió de jubilació de cada individu.

1.6 BRETXA SALARIAL

La igualtat entre dones i homes és un principi constitucional aplicat en els textos normatius de l'Estat espanyol. Per a la seva implementació s'han aprovat lleis i polítiques públiques a l'estat i s'ha habilitat un organisme per a la seva posada en pràctica: l'Institut de la Dona a Espanya. Malgrat això, la igualtat no és objectiu aconseguit i es manifesta en diferents aspectes de la vida social, política i econòmica, com la menor participació de les dones en el mercat laboral i en la vida política, la seva escassa presència en llocs directius i la seva major responsabilitat i dedicació a les tasques domèstiques. La presència femenina inferior a la masculina en la vida pública i privada, l'anomenada divisió sexual del treball, s'inicia al segle XIX i, tot i que des de llavors han tingut lloc grans canvis socials, polítics i econòmics, la desigualtat entre dones i homes segueix persistint en la societat actual.

Sarai Rodríguez (2017) afirma: "*Las interrupciones en las carreras de cotización padecidas por las mujeres suponen un reto para la igualdad en el ámbito de la Seguridad Social y un auténtico desafío para la garantía de la independencia económica de las mujeres,*

especialmente durante la vejez, momento en que éstas afrontan un mayor riesgo de pobreza" (p.1).

L'estudi anterior esmentat, estudia les causes de la bretxa de gènere en el sistema de la Seguretat Social, per després, analitzar els mecanismes correctors d'aquesta situació existent en el sistema d'ordenament jurídic i que intenten compensar la màxima dedicació a la feina no retribuïda, que en molts casos a les dones no els permet desenvolupar la seva carrera professional i aquest fet les aboca a pensions insuficients a la vellesa.

Segons la definició d'Eurostat, la bretxa de gènere, és la diferència entre el salari brut per hora dels homes i el de les dones, expressat com a percentatge del salari brut per hora dels homes. Eurostat el calcula únicament per als assalariats que treballen en unitats de 10 i més treballadors i en el guany per hora inclou els pagaments per hores extraordinàries realitzades, però exclou les gratificacions extraordinàries.

Per a poder analitzar la igualtat entre homes i dones en l'activitat laboral i les retribucions associades a aquesta activitat, s'ha de conèixer el guany anual brut dels treballadors, homes i dones, en funció de les diferents característiques com a ocupació, activitat econòmica, edat, tipus de jornada, etc.

Les diferències entre homes i dones en temes de pensió de jubilació són evidents i es justifiquen perquè les dones solen tenir salaris més baixos i carreres professionals més curtes, de forma que la seva pensió no és tan elevada com la dels homes.

Les diferències entre homes i dones en la taxa de substitució no han estat estudiades en profunditat fins ara.

1.7 POBLACIÓ

Espanya és un dels països en què l'esperança de vida és més elevada. En el cas de les dones, amb gairebé 85 anys d'esperança de vida actual, es situa al capdamunt del rànquing només darrere del Japó. En el cas dels homes, amb gairebé 80 anys d'esperança de vida, es troba també en les primeres posicions (INE, 2018).

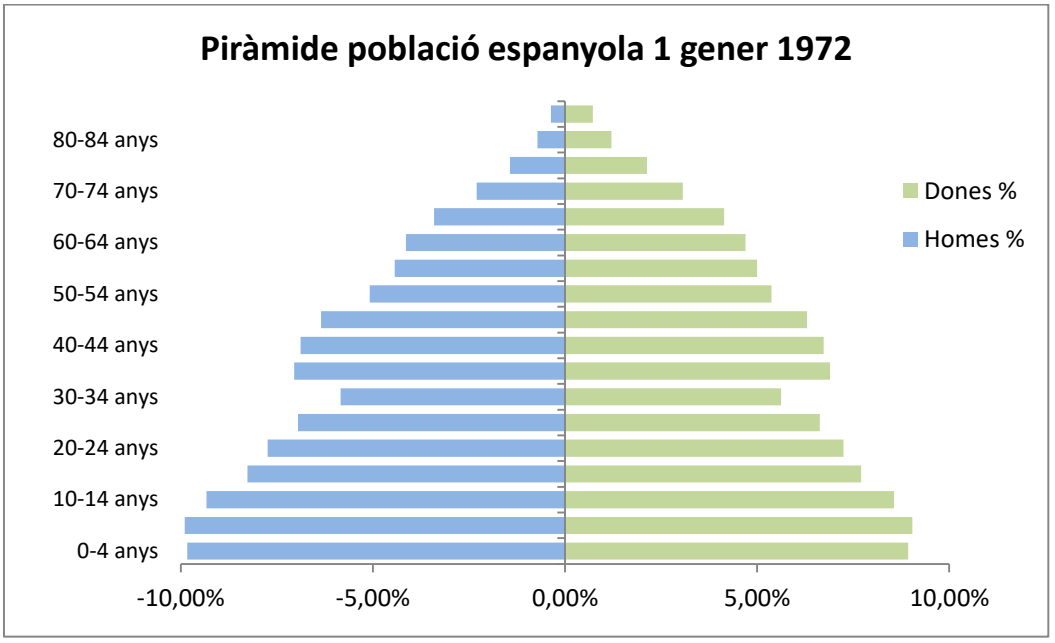
Per tant, tendim a una estructura demogràfica fortament envellida, o el que és el mateix, a una piràmide poblacional fortament invertida, amb una gran massa de població de persones d'edat avançada i un dèficit de població en edat activa.

Al desembre de 2015 hi havia un total de 8.508.482 de pensionistes en tot l'estat espanyol, on s'inclouen les pensions d'invalidesa, les de jubilació, les de viudetat, les d'orfandat i les de favor familiar.

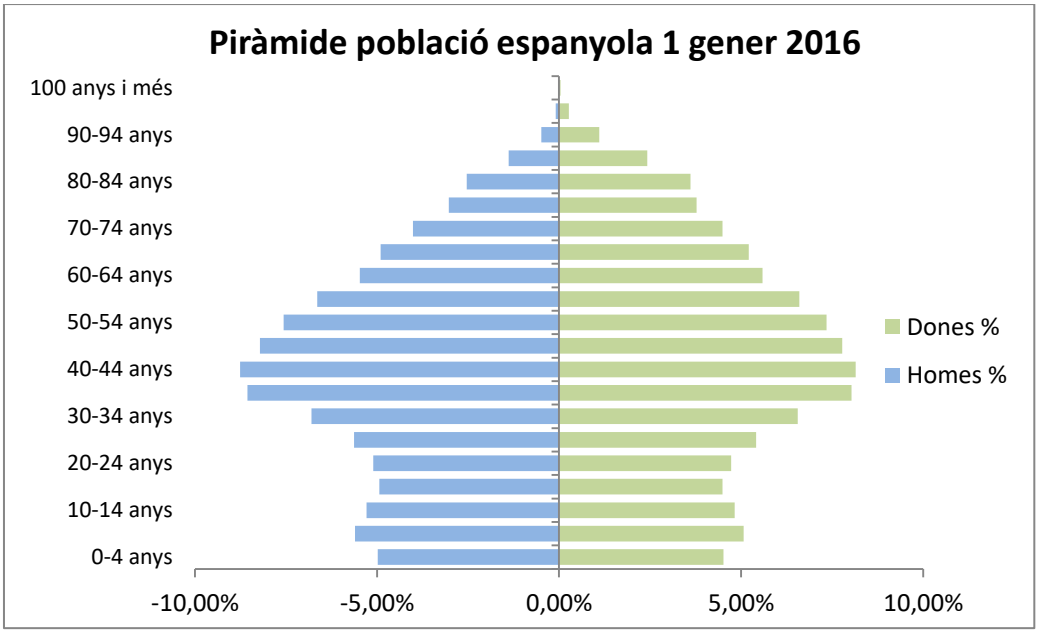
La piràmide de població representa el nombre d'habitants en un país distribuïts per edats, en determinats moments del temps. En aquest cas, s'ha volgut extreure la informació de l'any

1971, ja que, hi ha 44 anys de diferència en comparació al 2015, una xifra bastant elevada per veure els canvis poblacionals.

Per a realitzar les següents piràmides s'han extret les dades de l'INE, a 1 de gener de l'any 1972 i a 1 de gener de l'any 2016. D'aquesta manera es té la població de l'any 1971 i la del 2015 respectivament, que és la que es precisa.



Gràfic 1.2 Piràmide població espanyola 1 gener 1972. Font: *Instituto Nacional de Estadística (INE).* Dades any a 1 de gener de 1972.



Gràfic 1.3. Piràmide població espanyola 1 gener 2016. Font: *Instituto Nacional de Estadística (INE)*. Dades any a 1 de gener de 2016.

Observant les dues piràmides poblacionals, es pot apreciar que la població va envellint segons el comportament d'un conjunt d'indicadors, bàsicament: el nombre de persones grans, la relació de la població d'edat major sobre el total de la població, la taxa de dependència (població jubilada respecte la població activa) i l'edat mitjana de la població.

És molt freqüent separar els individus per gènere, per posar de manifest que l'evolució en la composició poblacional per sexe no té per què ser la mateixa en totes les edats. Al primer gràfic es presenta la piràmide de població a Espanya a 1 de gener del 1972 i al segon la de l'1 de gener de 2016.

Al 1971, la piràmide mostra un decreixement de la població a mesura que augmenta l'edat de les persones, decreixement molt marcat a partir dels 65 anys d'edat, amb uns percentatges de població en edat activa (persones en edat de treballar) notablement superiors als de la població susceptible de rebre prestacions de jubilació.

Quan l'anàlisi es realitza per al 2015 s'observa un canvi en la forma piramidal de 1971, amb un estrenyiment a la base de la piràmide, i un eixamplament a la part central de la mateixa. Dos fenòmens justifiquen bàsicament aquest comportament. D'una banda, es produeix un descens molt accentuat de naixements respecte les dues poblacions. Per un altre, s'observa el desplaçament en la piràmide de la generació del *baby boom*, que ara es concentra a les edats associades a la població activa, o en edat legal de treballar.

A tot això cal sumar, els efectes migratoris que contemplen tant les entrades de persones d'altres països, com les sortides de població espanyola a altres països.

La població major de 65 anys és superior a l'observada en 1971, la qual cosa indica que, a l'any 2015, per a mantenir constant la ràtio de jubilats o beneficis per treballador en actiu, es necessiten més persones en edat legal de treballar que les que es necessitaven al 1971.

1.8 OBJECTIUS

Quan la meva tutora del treball de fi de grau em va proposar aquest projecte, em va explicar que els resultats que obtinguéssim dels objectius que es podien establir, s'utilitzarien per a un projecte major el qual ja estava en marxa. Això va fer que la meva motivació per a dur-lo a terme i complir els objectius que s'enunciaran a continuació s'incrementés.

L'objectiu del present treball és pot dir que és triple. Per una banda, es vol trobar la diferència entre l'últim salari i la primera pensió dels individus separant per sexes. S'enllaçaran diversos arxius per a poder analitzar variables procedents de varis fitxers (tots corresponents a la MCVL).

Per altra banda, també s'ha volgut analitzar si el nivell mitjà de pensió de jubilació en dones és igual al dels homes i així, poder veure si existeixen desigualtats de gènere tant en els ingressos com en la pensió.

A més a més dels objectius anteriors, puc afegir com a objectiu personal i acadèmic, que el fet d'haver-me hagut d'enfrontar a una base de dades de tals dimensions ha sigut tot un repte per a mi, tenint en compte que mai havia treballat amb cap base de dades de similars proporcions i que la informació de la base de dades no havia estat depurada prèviament. Aquest fet ha comportat que s'hagués de partir des de zero, ja que, al no estar la base de dades tractada anteriorment, contenia valors incoherents i inexistents.

Afegir, que el temps dedicat a comprendre l'estructura i la composició d'aquesta base ha sigut relativament alt, ja que, constava de moltes variables i per a poder extreure una informació verídica s'han hagut d'entendre totes les que s'han utilitzat. Per això es pot afirmar que a l'haver analitzat aquestes dades longitudinals amb el software R, he pogut adquirir competències en l'ús del Big data que d'altra manera no hagués obtingut.

Existeixen notables diferències salarials entre homes i dones, però no hi ha gaires estudis que es concentrin en les edats prèvies al moment de la jubilació, i que analitzin les seves causes ni la seva repercussió en el nivell de pensió que s'acaba cobrant.

1.9 AGRAÏMENTS

Agraïments a la Montserrat Guillen Estany, tutora del treball de fi de grau, per la seva exigència, el constant suport, l'orientació i la confiança d'acompanyar-me durant la trajectòria del projecte. A la meva parella, l'Oriol Pausas pel seu interès, recolzament i visió crítica externa i als amics i familiars per estar sempre al meu costat.

DESCRIPCIÓ DELS FITXERS I LA BASE DE DADES

2.1 PROGRAMES UTILITZATS

Per poder realitzar aquest estudi s'ha utilitzat en tot moment, un software estadístic així com un parell de llenguatges de programació.

Aquest software que s'ha esmentat anteriorment, és el *software* estadístic anomenat R. Aquest ha sigut el més utilitzat al Grau d'Estadística, fet que m'ha ajudat molt per a la realització del codi tant per a la lectura, com pel tractament de les dades i les seves pertinents anàlisis realitzades.

La interfície utilitzada del *software* ha sigut l'R-Studio, on s'ha pogut crear el codi del treball amb el propi llenguatge del programa, l'R. A més de l'R, s'ha cregut necessari disposar d'un altre llenguatge per a la implementació del codi, aquest ha sigut l'SQL mitjançant el paquet "sqldf", on bàsicament s'ha necessitat per a poder unir diversos arxius. De l'SQL vaig adquirir certs coneixements gràcies a l'assignatura de Fitxers i Bases de Dades, assignatura que em va motivar a decantar-me per realitzar un treball com aquest.

2.2 TIPUS DE FITXERS I BASES DE DADES

En una primera instància, les microdades no es podien descarregar de la web de la Seguretat Social en qualsevol moment que un vulgui disposar d'elles, sinó que la base de dades MCVL va ser demanada i proporcionada per la Direcció General d'Ordenació de la Seguretat Social, ja que, la seva obtenció era d'accés restringit, per tant, es van haver d'omplir uns documents de confidencialitat i compromís on ens comprometíem a utilitzar-la només per a fins acadèmics i excloent l'ús d'aquesta amb fins administratius.

Un cop la base de dades estava a la nostra disposició es va poder procedir a la lectura dels arxius que la formaven i a la investigació del que era cada fitxer.

Aquesta estava formada per uns arxius en format digital, com ja s'ha dit abans, els quals estaven repartits en 22 fitxers en format ".txt". A la mateixa web de la Seguretat Social, hi havia un document on hi ha explicacions sobre el procés de selecció de la mostra, de la seva estructura i del seu contingut, de manera que així és més fàcil identificar les dades i davant de tot, poder interpretar-les correctament.

Per a que la mostra sigui vàlida i representativa, les persones que s'hi inclouen han de complir aquests criteris de selecció:

- És indispensable que cada individu tingui a la seva disposició un document identificador de la persona física (IPF). Generalment pot ser el Document Nacional d'Identitat (DNI) o el Nº d'Identificació d'Estranger (NIE), on el número d'aquest ha de pertànyer al conjunt de números sel·leccionables (el 4% de tots els números possibles).

- La segona i última condició és que tots els individus que es seleccionin hagin format part de la població espanyola en l'any de referència. La població a tenir en compte la formen, principalment, els individus afiliats en situació d'alta, excloent els qui exclusivament estan registrats a la Seguretat Social a efectes de rebre assistència sanitària, ni els perceptors de pensions no contributives o assistencials, ni qui estan enquadrats en sistemes de previsió social diferents de la Seguretat Social, com podria ser el cas dels funcionaris de Classes Passives, ni tampoc els inscrits com a demandants de feina que no reben ni prestació ni subsidi per l'atur, ni tampoc els perceptors de la Renda Activa d'inserció.

Les dues condicions per a optar a formar part de la MCVL són independents, és a dir, complir una de les dues no influeix en la probabilitat de complir l'altra. El fet d'exigir-les ambdues garanteix que la mostra sempre estarà seleccionant, aleatòriament, un 4% dels individus que compleixen la segona condició (amb igualtat de probabilitats de ser seleccionats). En cada mostra que es realitza es duen a terme els contrastos estadístics adients per a garantir la representativitat de la mostra escollida.

De manera constant es produeixen entrades i sortides d'individus de la població de referència, ja que, és habitual que hi hagi individus que canviïn de feina, es quedin a l'atur, tinguin una pensió, etc. Són fets que depenen de la situació de cada persona per això hi ha canvis amb freqüència.

En canvi, complir la primera condició és un factor gairebé permanent, ja que, el conjunt de números seleccionats no varia, a més de que el número d'identificació de la persona física no és un número que vagi percebent modificacions al llarg de la vida d'una persona, si no que gairebé en tots els casos s'assigna només un identificador, el qual aquest és per sempre. Tanmateix, es podria donar el cas que algú que adquireixi la nacionalitat espanyola canviï el seu NIE per el DNI, en aquest cas sí que canviaria el número i tipus de document, però com ja s'ha dit, no és el més freqüent.

2.3 FITXERS I TAULES

Com s'ha esmentat anteriorment, el conjunt de la base de dades MCVL 2015, estava en format digital, el qual un cop descarregat, estava organitzat en un arxiu comprimit on hi havia 22 fitxers els quals es detallaran a continuació:

2.3.1 DESCRIPCIÓ DE LA TAULA: PERSONAS

En la MCVL de l'any 2015 la taula de *Personas* , aquesta taula té 1.131.430 registres. Aquesta és la taula que inclou tots els individus de la mostra, la qual conté informació bàsica personal de l'individu, sempre mantenint l'anonimat. Aquesta, conté una fila o registre per a cada persona física escollida per a la mostra, amb les columnes pertinents on hi ha les

característiques esmentades procedents de la base de dades de la Seguretat Social i del Padró Municipal. La informació d'aquesta taula es troba en un únic arxiu amb el nom de *MCVL2015PERSONAL_CDF.txt.*

Per diferenciar cada individu dels altres, s'inclou una clau denominada Identificador de Persona Física (IPF), que es concentra a la primera columna de la taula.

També s'inclouen variables com la data de naixement, el sexe, la nacionalitat, la província de naixement, entre d'altres.

En concret, les variables que formen aquesta taula són les següents:

Número i nom de la variable	Etiqueta
1.001 IDENTIFICADOR DE LA PERSONA FÍSICA (IPF)	V1
1.002 FECHA DE NACIMIENTO	V2
1.003 SEXO	V3
1.004 NACIONALIDAD	V4
1.005 PROVINCIA DE NACIMIENTO	V5
1.006 PROVINCIA DE PRIMERA AFILIACIÓN	V6
1.007 DOMICILIO DE RESIDENCIA HABITUAL	V7
1.008 FECHA DE FALLECIMIENTO	V8
1.009 PAÍS DE NACIMIENTO	V9
1.010 NIVEL EDUCATIVO	V10

Taula 2.1 Resum de totes les variables de la taula *PERSONAS*

- *1.001 IDENTIFICADOR DE LA PERSONA FÍSICA (IPF):* aquesta variable és la clau que identifica a cada individu inclòs a la mostra i la distingeix dels demés. Aquesta, també es presenta a les altres taules de la mostra i així facilita la connexió entre elles i així permet relacionar informació de la totalitat de les taules d'un mateix individu i d'aquesta manera s'obté tota la informació que es vol saber sobre un o més individus concret.
  La informació original és formada per una cadena de caràcters de 15 posicions. On les posicions 2 a 11 de la dada original són substituïdes per una cadena aleatòria per així evitar la possible identificació de la persona i mantindre l'anonimat.

- *1.002 FECHA DE NACIMIENTO:* com el seu propi nom indica, aquesta variable correspon a la data de naixement de la persona que està registrada a la Seguretat

Social. La data de naixement es registra completa, amb 8 posicions prenent el format: aaaa/mm/dd, és a dir, l'any, el mes i el dia respectivament.

- *1.003 SEXO:* Indica el gènere masculí o femení, de la persona que està registrada a la Seguretat Social. La dada no canvia a no ser que hi hagués un error o per canvi de sexe de l'individu. És una variable dicotòmica que pren valor "1" per als homes i "2" per a les dones. En altre cas, s'interpreta que no consta la dada.

- *1.004 NACIONALIDAD*: És la relació que uneix a una persona amb un determinat estat, en el cas de la nacionalitat espanyola el vincle serà amb l'estat espanyol. A cada nacionalitat li corresponen 3 posicions alfanumèriques. Per identificar el país del que té la nacionalitat.

- *1.005 PROVINCIA DE NACIMIENTO:* Identifica la província a la qual ha nascut cada individu. Per identificar la província, la variable té assignats uns codis numèrics de dues posicions per a cada província, anant del 01 fins al 52, en el cas d'haver nascut a Espanya. En el cas d'haver nascut a l'estranger se li adhereix la clau 66 al valor.

- *1.006 PROVINCIA DE PRIMERA AFILIACIÓN:* És la província espanyola on es va afiliar l'individu per primera vegada. S'utilitza el mateix mètode que a la variable anterior per identificar el codi en cada registre, però afegint codis específics per a Serveis Centrals (99) i per a determinats territoris en l'àmbit de la gestió de l'Institut Social de la Marina (Règim Especial de Treballadors del Mar, (56)).

- *1.007 DOMICILIO DE RESIDENCIA HABITUAL:* Com el seu nom indica, aquesta variable identifica el municipi on la persona registrada a la Seguretat Social té la seva residència habitual.

- *1.008 FECHA DE FALLECIMIENTO:* Expressa el moment en que la persona que està registrada va morir. S'expressa amb el mateix format que la variable 1.002 FECHA DE NACIMIENTO.

- *1.009 PAÍS DE NACIMIENTO:* És el país on es va donar el naixement de la persona registrada a la Seguretat Social. S'expressa amb un codi alfanumèric de tres posicions.

- *1.010 NIVEL EDUCATIVO:* aquesta variable indica el nivell de coneixements que ha adquirit durant la seva vida la persona, mitjançant la titulació acadèmica o en absència d'aquesta, per les seves habilitats bàsiques. S'expressa amb una clau de dos dígits. Exemple: el codi 10 significa que aquell individu en qüestió no sap ni llegir ni escriure, el 30 és que disposa de graduat escolar, el 40 Batxillerat, Formació Professional de Segon Grau o títols equivalents o Superiors, etc.

2.3.2 DESCRIPCIÓ DE LA TAULA: AFILIACIÓN

En la MCVL de l'any 2015, el total de registres de les taules de *Afiliación* és de 22.381.074. Aquestes taules mostren les dades essencials dels episodis de l'afiliació a la Seguretat Social que ha mantingut la persona en qüestió per a la MCVL al llarg de la seva vida.

El més habitual, és que la taula mostri un únic registre per a cada episodi d'afiliació, el qual s'entén com un període de temps que transcorre entre una data d'alta i una de baixa, ambdues incloses, etc.

Com s'ha dit, mostra casos habituals com l'alta laboral, com de no tant freqüents com l'alta en l'atur, l'alta de conveni especial, etc. Existeixen casos, com els individus que tenen més d'una feina o l'atur parcial, entre d'altres, aquests individus poden tenir registrades varies relacions d'afiliació a la vegada.

A diferència de la taula de *Personas* que hi ha un únic registre per cada individu, la taula de *Afiliación,* al tenir varis registres d'alta i de baixa per a la població escollida, fa que hi hagi més d'una fila en alguns casos de la mateixa persona, ja que, pot haver canviat de feina durant l'any de referència. Això implica que el fitxer sigui molt més gran i hagi d'estar fragmentat en 4 arxius els quals contenen informació sobre els individus que s'han donat d'alta o de baixa algun cop a la seva vida i per tant, queda registrat tots els moviments. Els arxius s'anomenen amb el nom de *MCVL2015AFILIAD1_CDF.txt, MCVL2015AFILIAD2_CDF.txt, MCVL2015AFILIAD3_CDF.txt* i *MCVL2015AFILIAD4_CDF.txt.*

Correspondència entre la taula de *Afiliación* i la de *Personas*

Aquest arxiu es pot relacionar amb la taula anterior a través de la variable identificadora de la persona física i, per tant, es pot tenir informació detallada de caràcter professional a la vegada de saber les característiques personals com l'edat, el sexe, etc.

Correspondència entre la taula de *Afiliación* i la de *Bases de Cotización*

Els registres de la taula *Bases de Cotización* associats amb una relació d'afiliació, representada per un registre a la taula d'afiliació, contenen el mateix identificador de la persona física (IPF) i el mateix codi de compte de cotització secundària (CCC). En general, a cada registre de la taula de *Afiliación* li correspondran tants registres a la de *Bases de Cotización* com anys naturals diferents transcorrin entre la data d'alta i la data de baixa que es trobin en el registre de *Afiliación*.

En aquest fitxer es contemplen nombroses variables, en concret 34 en composen aquest arxiu. Amb aquesta abundant xifra, hi ha informació molt completa per a realitzar diferents estudis, però en aquest treball no ha sigut necessari utilitzar-les totes.

Com ens trobem d'avant d'aquesta situació, s'ha hagut de fer un filtratge previ. Escollint les imprescindibles per a realitzar els càlculs pertinents i unes quantes més per a ampliar la informació. Amb les que no s'ha treballat només es citaran a continuació, deixant pas, a les descripcions de les que si que s'han utilitzat per a l'estudi.

En particular, les variables que constitueixen aquesta taula són les següents:

Número i nom de la variable	Etiquetes
2.001 IDENTIFICADOR DE LA PERSONA FÍSICA (IPF)	V1
2.002 RÉGIMEN DE COTIZACIÓN	V2
2.003 GRUPO DE COTIZACIÓN	V3
2.004 TIPO DE CONTRATO DE TRABAJO	V4
2.005 COEFICIENTE DE TIEMPO PARCIAL	V5
2.006 FECHA REAL DEL ALTA EN AFILIACIÓN	V6
2.007 FECHA REAL DE LA BAJA EN AFILIACIÓN	V7
2.008 CAUSA DE BAJA EN AFILIACIÓN	V8
2.009 MINUSVALÍA SEGÚN ALTA EN AFILIACIÓN	V9
2.010 CÓDIGO DE CUENTA DE COTIZACIÓN (CCC)	V10
2.011 DOMICILIO DE ACTIVIDAD DE LA CUENTA DE COTIZACIÓN	V11
2.012 ACTIVIDAD ECONÓMICA DE LA CUENTA DE COTIZACIÓN (CNAE 2009)	V12
2.013 NÚMERO DE TRABAJADORES EN LA CUENTA DE COTIZACIÓN	V13
2.014 FECHA DE ALTA DEL PRIMER TRABAJADOR DE LA CUENTA DE COTIZACIÓN	V14
2.015 TIPO DE RELACIÓN LABORAL (TRL) DE LA CUENTA DE COTIZACIÓN	V15
2.016 COLECTIVO ESPECIAL DE LA CUENTA DE COTIZACIÓN	V16
2.017 EMPLEADOR (TIPO DE IDENTIFICADOR)	V17
2.018 EMPLEADOR (FORMA JURÍDICA) – LETRA NIF DE LA ENTIDAD PAGADORA	V18
2.019 IDENTIFICADOR (NIF/CIF) ANONIMADO DE LA ENTIDAD PAGADORA	V19
2.020 CÓDIGO DE CUENTA DE COTIZACIÓN PRINCIPAL (CCCP)	V20
2.021 PROVINCIA DEL DOMICILIO DE LA CUENTA DE COTIZACIÓN PRINCIPAL	V21
2.022 FECHA MODIFIC.DEL TIPO CONTRATO INICIAL O COEF.TIEMPO PARCIAL INICIAL	V22
2.023 TIPO DE CONTRATO INICIAL	V23
2.024 COEFICIENTE DE TIEMPO PARCIAL INICIAL	V24
2.025 FECHA MODIFIC.TIPO CONTRATO SEGUNDO O COEF.TIEMPO PARCIAL SEGUNDO	V25
2.026 TIPO DE CONTRATO SEGUNDO	V26
2.027 COEFICIENTE DE TIEMPO PARCIAL SEGUNDO	V27
2.028 FECHA DE MODIFICACION DEL GRUPO DE COTIZACIÓN INICIAL	V28
2.029 GRUPO DE COTIZACIÓN INICIAL	V29

2.030 ACTIVIDAD ECONÓMICA DE LA CUENTA DE COTIZACIÓN (CNAE 93)	V30
2.031 SISTEMA ESPECIAL DE TRABAJADORES AGRARIOS (SETA)	V31
2.032 TIPO DE RELACIÓN CON OTRAS ENTIDADES O AUTÓNOMOS	V32
2.033 FECHA DE EFECTO DEL ALTA EN AFILIACIÓN	V33
2.034 FECHA DE EFECTO DE LA BAJA EN AFILIACIÓN	V34

Taula 2.2 Resum de totes les variables de la taula *AFILIACIÓN*

Ara es procedirà a la descripció de les variables utilitzades per a l'estudi present:

- *2.001 IDENTIFICADOR DE LA PERSONA FÍSICA (IPF):* aquesta variable és la que figura en totes les taules de la MCVL i que facilita la relació dels registres d'una mateixa persona en les diverses taules de la mostra. La descripció d'aquesta variable ja s'ha efectuat a la taula de *Personas* i està disponible a l'apartat 2.3.1.

- *2.002 RÉGIMEN DE COTIZACIÓN:* aquesta té la funció d'identificar el règim d'enquadrament del treballador durant l'episodi d'afiliació. Aquesta és una clau numèrica de quatre posicions, on cada règim inclou un tipus de treballadors diferent en funció de l'activitat que du a terme i de la manera que ho exerceix. En termes generals, es pot afirmar que els règims de la Seguretat Social es divideixen en 4 grups: els de tipus General, els Autònoms, de Mar i de Carbó (amb els seus corresponents subgrups identificats amb situacions més especifiques en cotització).

2.3.3 DESCRIPCIÓ DE LA TAULA: BASES DE COTIZACIÓN

La taula de *Bases de Cotización* conté 28.010.404 registres. Aquesta és la tercera taula de la MCVL, la qual reflexa l'import mensual, expressat en cèntims d'euro, de les bases de cotització dels individus seleccionats per a la mostra. Aquesta, es divideix en un conjunt de 13 fitxers de text, els quals els 12 primers corresponen a cotitzacions per compte aliè en general i l'últim, el tretzè correspon als que cotitzen per compte propi i altres situacions especials de cotització.

Aquests prenen el nom de MCVL2015COTIZA1_CDF.txt, MCVL2015COTIZA2_CDF.txt, MCVL2015COTIZA3_CDF.txt, MCVL2015COTIZA4_CDF.txt, MCVL2015COTIZA5_CDF.txt, MCVL2015COTIZA6_CDF.txt, MCVL2015COTIZA7_CDF.txt, MCVL2015COTIZA8_CDF.txt, MCVL2015COTIZA9_CDF.txt, MCVL2015COTIZA10_CDF.txt, MCVL2015COTIZA11_CDF.txt, MCVL2015COTIZA12_CDF.txt i MCVL2015COTIZA13_CDF.txt.

La taula mostra una única fila o registre per cada any natural, persona (IPF) i si existeix, ocupador (CCC) en els que hagi estat vigent alguna relació d'afiliació d'alta.

Quan la Seguretat Social registra més d'una base de cotització durant el mateix mes per a un mateix treballador (IPF) i un mateix codi de compte de cotització (CCC), la quantitat mensual

designada en la fila corresponent de la taula es sempre la suma de totes les bases registrades en aquell mes.

A un registre de la taula de *Bases de Cotización* li correspondrà com a mínim un registre en la de *Afiliación*, podent correspondre-li més d'un quan, al llarg del mateix any natural, el treballador (IPF) ha tingut més d'un episodi diferent d'afiliació en el mateix codi de compte de cotització (CCC).

Es pot donar el cas de que faltin registres de la taula Bases de Cotización quan no s'han localitzat o no s'han pogut determinar les bases de cotització corresponents i també per a determinades relacions d'afiliació en les que no existeix cotització per contingències comunes, com passa generalment quan es perceb subsidi de l'atur.

Aquesta taula consta de 16 columnes en la seva totalitat. La primera columna correspon a la variable Identificador de persona física (variable que s'ha esmentat en les dues taules anteriors) de la persona seleccionada per a la mostra. La segona columna pertany a el codi de compte de cotització i la tercera, a l'any que es referiran les bases de cotització mensuals. Les 13 columnes següents corresponent a les bases de cotització de cada mes de l'any i la última és la suma de tots aquests mesos.

Número i nom de la variable	Etiquetes
3.001 IDENTIFICADOR DE LA PERSONA FÍSICA (IPF)	V1
3.002 CODIGO DE CUENTA DE COTIZACION SECUNDARIA (CCC)	V2
3.003 AÑO DE COTIZACIÓN	V3
3.004 BASE DE COTIZACIÓN MENSUAL CONT.COMUNES	V4
3.005 TOTAL ANUAL BASES DE COTIZACIÓN	V5

Taula 2.3 Resum de totes les variables de la taula *COTIZACIÓN*

- *3.001 IDENTIFICADOR DE LA PERSONA FÍSICA (IPF):* aquesta variable és la que figura en totes les taules de la MCVL i que facilita la relació dels registres d'una mateixa persona en les diverses taules de la mostra. La descripció d'aquesta variable ja s'ha efectuat a la taula de *PERSONAS* i a més apareix a la de *AFILIACIÓN*.

- *3.003 AÑO DE COTIZACIÓN:* aquesta variable indica l'any natural al qual es refereixen les bases de cotització que apareixen a les columnes del registre. Aquesta és una clau numèrica de quatre posicions. A la taula actual apareixen tants registres com anys diferents, sempre que hi hagi una relació d'afiliació vigent en diversos anys naturals.

- *3.004 BASE DE COTIZACIÓN MENSUAL POR CONTINGENCIAS COMUNES:* la dada correspon a la suma de totes les bases de cotització registrades per a un mateix IPF i CCC en el mes corresponent. La base de cotització és la xifra sobre la qual s'aplica el percentatge o tipus de cotització per calcular la quota a ingressar a la Seguretat Social per contingències comunes. Aquesta està formada per 12 columnes on cada una d'elles correspon a un mes de l'any diferent. Ocupa 8 posicions i les dades estan expressades en cèntims d'euro.

- *3.005 TOTAL ANUAL BASES DE COTIZACIÓN:* L'última variable correspon a una única columna, la qual està formada per la suma de les dotze bases mensuals. És a dir, és la suma de totes les columnes de la variable *3.004 BASE DE COTIZACIÓN MENSUAL POR CONTINGENCIAS COMUNES.* Com la variable anterior, aquesta ocupa 8 posicions i les dades estan expressades en cèntims d'euro.

2.3.4 DESCRIPCIÓ DE LA TAULA: PENSIONES

El nombre total de registres de la taula de *PENSIONES* és de 4.671.890. Aquesta presenta les propietats fonamentals de les pensions que perceben, o han percebut en el passat, els individus inclosos en la mostra, com ara el tipus de pensió (incapacitat, jubilació, viduïtat, etc.), la data en que es va reconèixer i l'import dels diferents conceptes de pagament que s'integren en ella.

La informació de la taula *PENSIONES* està en un únic fitxer anomenat MCVL2015PRESTAC_CDF.TXT.

En la taula present, hi ha una majoria d'individus que en cap cas han percebut una pensió procedent de la Seguretat Social. Únicament apareixen registres a la taula de pensions per a aquelles persones incloses en la MCVL que les perceben en l'any de referència o que les van percebre alguna vegada.

En termes generals, per a cada pensió s'inclouen en la taula tantes files com anys naturals hagi estat vigent des del seu reconeixement, de tal manera que cada fila reflecteix les quanties o imports i la situació de a pensió al final d'aquest any.

En ocasions molt poc freqüents, algunes pensions poden tenir diversos registres referits a un mateix any, quan al llarg d'aquest s'ha produït més d'un canvi de situació com, per exemple, si es passa d'estar en alta a estar en suspensió i després, abans que finalitzi l'any, es torna a situació d'alta. En aquests casos el registre que reflecteix la situació al final d'any serà el que mostri la data de situació més recent. S'haurà de decidir el tractament adequat d'aquests casos per evitar computar més d'una vegada la mateixa pensió.

Una persona pot ser titular de diverses pensions al mateix temps si són compatibles. És freqüent la simultaneïtat de pensions de jubilació amb viduïtat, però també es donen casos

de compatibilitat entre pensions de la mateixa classe, però de diferents règims, o fins i tot sent del mateix règim, quan alguna procedeix d'un sistema especial de cotització o d'un antic règim especial que es va integrar en un altre.

Aquesta taula, és una de les que conté més variables de la MCVL, la qual està formada per 42 variables en la seva totalitat. Com a les altres taules, la primera variable és l'identificador de la persona física de l'individu titular de la pensió.

A continuació figura l'any a què es refereixen les dades de la fila o registre. En tercer lloc es dona pas a l'identificador de la prestació, codi alfanumèric que identifica de manera unívoca a cada pensió i que es repeteix en totes les files corresponents a la mateixa pensió. L'identificador de la prestació permetrà distingir les diferents pensions que corresponen a una mateixa persona física (IPF). Després de l'identificador de la prestació s'afegeixen altres 39 columnes amb diferents variables o atributs de la pensió que es relacionen a la taula següent:

Número i nom de la variable	Etiquetes
4.001 IDENTIFICADOR DE LA PERSONA FÍSICA	V1
4.002 AÑO DEL DATO	V2
4.003 IDENTIFICADOR DE LA PRESTACIÓN	V3
4.004 CLASE DE LA PRESTACIÓN	V4
4.005 SITUACIÓN DEL SUJETO CAUSANTE	V5
4.006 GRADO DE INCAPACIDAD	V6
4.007 FECHA DE MINUSVALÍA	V7
4.008 NORMA SOVI	V8
4.009 CLASE DE MÍNIMO	V9
4.010 RÉGIMEN DE LA PENSIÓN	V10
4.011 FECHA DE EFECTOS ECONÓMICOS DE LA PENSIÓN	V11
4.012 BASE REGULADORA	V12
4.013 PORCENTAJE APLICADO A LA BASE REGULADORA	V13
4.014 AÑOS BONIFICADOS	V14
4.015 AÑOS CONSIDERADOS COTIZADOS PARA LA JUBILACIÓN	V15
4.016 IMPORTE MENSUAL DE LA PENSIÓN EFECTIVA	V16
4.017 IMPORTE MENSUAL DE REVALORIZACIÓN	V17
4.018 IMPORTE MENSUAL DE COMPLEMENTOS GARANTÍA MÍNIMOS	V18
4.019 IMPORTE MENSUAL DE OTROS COMPLEMENTOS	V19

4.020 IMPORTE MENSUAL TOTAL DE LA PRESTACIÓN	V20
4.021 SITUACIÓN DE LA PRESTACIÓN (CAUSA DE BAJA)	V21
4.022 FECHA DE SITUACIÓN DE LA PRESTACIÓN	V22
4.023 PROVINCIA DE GESTIÓN DE LA PRESTACIÓN	V23
4.024 NÚMERO DE TITULARES DE UN MISMO SUJETO CAUSANTE	V24
4.025 PRORRATA DE CONVENIO INTERNACIONAL	V25
4.026 PRORRATA DE DIVORCIO	V26
4.027 COEFICIENTE REDUCTOR TOTAL	V27
4.028 TIPO DE SITUACIÓN DE JUBILACIÓN	V28
4.029 COEFICIENTE DE PARCIALIDAD (JUBILACIÓN)	V29
4.030 PRESTACIÓN VITALICIA (ORFANDAD Y VIUDEDAD)	V30
4.031 CONCURRENCIA CON PRESTACIÓN AJENA	V31
4.032 IMPORTE ANUAL PAGAS EXTRA	V32
4.033 IMPORTE ANUAL PAGA DESVIACIÓN IPC	V33
4.034 IMPORTE ANUAL TOTAL DE LA PRESTACIÓN	V34
4.035 AÑO DE NACIMIENTO DEL CAUSANTE DE PENSIÓN DE SUPERVIVENCIA	V35
4.036 PENSIÓN LIMITADA	V36
4.037 COEFICIENTE REDUCTOR DEL LÍMITE MÁXIMO	V37
4.038 COMPATIBILIDAD DE JUBILACIÓN Y TRABAJO	V38
4.039 FECHA ORDINARIA DE JUBILACIÓN	V39
4.040 PERIODO COTIZADO EN EDAD ORDINARIA DE JUBILACIÓN	V40
4.041 PERIODO DE COTIZACIÓN	V41
4.042 PORCENTAJE POR AÑOS COTIZADOS	V42

Taula 2.4 Resum de totes les variables de la taula *PENSIONES*

Com a les altres taules, s'ha fet un filtratge i s'ha decidit utilitzar un nombre de variables en concret. A continuació s'inclourà la descripció de les variables que s'han utilitzat finalment:

- *4.001 IDENTIFICADOR DE LA PERSONA FÍSICA: aquesta variable és la que figura en totes les taules de la MCVL i que facilita la relació dels registres d'una mateixa persona en les diverses taules de la mostra. La descripció d'aquesta variable ja s'ha efectuat a la taula de PERSONAS, i a més apareix a les taules AFILIACIÓN i BASES DE COTIZACIÓN.*

- *4.002 AÑO DEL DATO:* Expressa l'any natural al que es refereixen les dades històriques reflectides en altres camps del registre, tals com l'import de la pensió, el qual el seu estat pot variar amb el temps. En termes generals, a cada pensió se li atribueix tants registres o files com anys naturals s'hagi percebut des de 1996. (Les dates i la situació de prestació permeten esbrinar si s'ha rebut tot l'any o només alguns mesos.

- *4.004 CLASE DE PRESTACIÓN:* Fonamentalment, permet identificar la circumstància que origina la prestació: la incapacitat del treballador o la jubilació, o bé la mort quan sobreviu el seu cònjuge, els seus fills o altres familiars que viuen amb ell i a les seves despeses. A més, també permet identificar alguns factors o situacions que afecten el benefici: grau de discapacitat, tipus d'orfe, la jubilació parcial, etc. Els possibles valors de la variables ocupen dos valors numèrics o alfanumèrics.

- *4.010 RÉGIMEN DE LA PENSIÓN:* Aquesta variable és una clau d'identificació de la totalitat de les normes aplicades en el reconeixement del dret a beneficiar-se, a causa de la inclusió del beneficiari en un determinat grup de treballadors (règim d'afiliació i cotització) durant la seva vida laboral, o en l'atenció al caràcter professional de la contingència (accident laboral o malaltia professional) des del qual es deriva la prestació. Les dades ocupen dues posicions numèriques.

- *4.011 FECHA DE EFECTOS ECONÓMICOS DE LA PENSIÓN:* Data en què neix el dret a l'abonament de la prestació. Els valors possibles en la informació original, El format de la data és l'any (quatre posicions), mes (dos posicions) i dia (dues posicions), és a dir, AAAAMMDD, però en la mostra es sintetitza la informació i només es reflexa l'any i el mes (AAAAMM), donant lloc a 6 posicions.

- *4.012 BASE REGULADORA:* Import calculat a partir de les bases de cotització del subjecte responsable d'un període determinat de temps, normalment com una mitjana corregida, que multiplicada per un percentatge variable determina la quantitat d'una bonificació mensual de la pensió reconeguda prèviament, que, excepte algun cas puntual (pensions limitades o prorratejada, entre d'altres) normalment coincideix amb la pensió eficaç.

- *4.015 AÑOS CONSIDERADOS COTIZADOS PARA LA JUBILACIÓN:* És el temps que es considera cotitzat, expressat en nombre d'anys sencers, a efectes del càlcul de la pensió de jubilació contributiva del sistema de Seguretat Social. S'observa el valor del camp de la base de dades que recull l'any en el període de cotització.
  En general, en pensions de jubilació ordinària (sense jubilació prèvia), reconegudes en l'aplicació de la legislació anterior a la Llei 27/2011, el camp que recull els anys sencers del període de cotització tota reflectirà el nombre d'anys cotitzats arrodonint cap a l'alça, és a dir, comptant com any sencer qualsevol fracció d'any, i tindrà valor zero els camps que recullen la resta de mesos i els dies restants.

Quan es parla de treball a temps parcial, hi ha normes específiques per al càlcul del temps cotitzat a través del qual, des de les hores treballades, es calculen els dies teòrics de cotització, que es multiplica per un coeficient elevador d'1,5. El resultat serà el nombre de dies per a les finalitats dels períodes de cotització mínima i la determinació del percentatge aplicable a la base reguladora.

- *4.016 IMPORTE MENSUAL DE LA PENSIÓN EFECTIVA:* Producte de la base reguladora per el percentatge aplicable a la mateixa, sent l'import resultant minorat per aplicació del límit màxim de les pensions públiques o per aplicació de prorrates (conveni, divorci o altres), o incrementat, si escau, amb l'antiga protecció familiar, o amb les revalorització accidents de treball anteriors a 1974. Aquesta quantitat no inclou les revaloracions ni complements de garantia de mínims. Però si que inclou, quan sigui procedent, el complement per gran invalidesa en la part corresponent a pensió efectiva. Els imports s'expressen en cèntims d'euro, incloent les dades anteriors al canvi de moneda de pesseta a euro, amb un format de set posicions.

- *4.020 IMPORTE MENSUAL TOTAL DE LA PRESTACIÓN:* Aquesta variable és la suma dels imports mensuals de pensió efectiva, de la revalorització, del complement de garantia de mínims i d'altres complements. Inclou, quan és necessari, els complements de gran invalidesa i les seves revalorització. De la mateixa manera que la variable 4.016, els imports s'expressen en cèntims d'euro, incloent les dades anteriors al canvi de moneda de pesseta a euro, amb un format de set posicions. A més a més, és la suma de diversos conceptes que, habitualment, es paguen en 14 mensualitats a l'any (12 ordinàries i 2 extra). No obstant això, en el cas de les pensions d'accidents de treball i de malalties professionals, alguns conceptes s'abonen en només les 12 mensualitats ordinàries, i no en les pagues extra, pel que aquestes seran d'un import una mica inferior. Aquesta dada es refereix a l'import total de cada pensió i no al de la suma de les diferents pensions que pugui percebre simultàniament una mateixa persona.

- *4.021 SITUACIÓN DE LA PRESTACIÓN (CAUSA DE BAJA):* Clau que indica si la pensió estava en situació d'alta, baixa o suspensió al final de l'any de referència del registre (variable "any de la dada"), així com els possibles motius de la baixa (veure comentaris). A cada tipus de situació de la prestació li correspon una clau numèrica de tres posicions la qual es classifica partint d'uns criteris establerts. Com per exemple: 000, 002 o 003 *PRESTACIÓN EN ALTA, 101 EN BAJA POR FALLECIMIENTO, 102 EN BAJA POR SANCION,* etc.
  La major part dels casos de la mostra corresponen a pensions que encara estaven en vigor en el moment de l'extracció de la dada. Per a moltes de les situacions possibles, especialment de suspensió, no es troben casos a la mostra, pel fet que són poc habituals. El més freqüent és que les pensions esdevinguin baixa per defunció o per compliment de l'edat límit (orfandat i familiars). No obstant això, hi ha molts casos de baixa per altres causes.

- *4.022 FECHA DE SITUACIÓN DE LA PRESTACIÓN:* Aquesta identifica l'any i el mes en què es va produir el pas a la situació indicada per la variable "situació de la prestació". El format de la data es presenta en la informació de la mostra de la següent manera: és l'any (quatre posicions) i el mes (dues posicions, és a dir AAAAMM.

- *4.036 PENSIÓN LIMITADA:* Clau que indica que la pensió, o algun dels seus components, està afectat per un límit màxim. En la majoria dels casos es refereix a la quantia màxima fixada, amb caràcter general, per a les pensions públiques, però en altres, amb menor freqüència, es refereix al límit màxim que s'aplica a algunes pensions de viduïtat reconegudes a persones amb càrregues familiars, o al límit màxim que poden arribar als complements de garantia de mínims de la pensió.

A continuació, es detallaran les taules *CONVIVIENTES* i *DATOS FISCALES,* les quals, no s'han cregut necessàries per a l'estudi, però com forma part de la MCVL es fa tot seguit, una breu descripció:

2.3.5 DESCRIPCIÓ DE LA TAULA: CONVIVIENTES

El nombre total de registres de la taula de Convivents és de 1.150.743. Aquesta inclou una fila o registre per cada persona física seleccionada per a la MCVL que figuri registrada en el Padró Municipal Continu, de manera que es pugui identificar el full padronal en què està inscrita. El 96,5% de les persones seleccionades per a la mostra figuren també en el Padró.

A cada fila o registre de la taula apareix l'IPF de la persona seleccionada per a la MCVL, seguit de les dades "data de naixement" i "sexe" d'ella mateixa i de, com a màxim, nou persones més, si n'hi ha, que figurin inscrites en el seu mateix full padronal.
Si han estat seleccionades per MCVL dues o més persones inscrites en el mateix full padronal, les dades dels que viuen en aquestes llars apareixeran repetides en altres tants registres de la taula de Convivents, encara que en diferents columnes.
De cada full padronal s'incorporen a la MCVL, com a màxim, dades de 10 persones, encara que existeixin més inscrites. No s'inclou informació dels fulls padronals on figuren més de 20 persones, per considerar-se que es tracta, probablement, d'institucions.

2.3.6 DESCRIPCIÓ DE LA TAULA: DATOS FISCALES

El nombre total de registres de la taula de *DATOS FISCALES* és de 1.872.704. Essencialment, la taula de *Datos Fiscales* conté informació individualitzada sobre les retribucions satisfetes i les retencions practicades per l'IRPF a les persones incloses en la MCVL durant l'any de

referència, així com algunes dades relatives a la seva situació familiar quan són necessaris per a l'aplicació de reduccions o beneficis fiscals.

ANÀLISI DESCRIPTIVA I COMPARACIÓ GRÀFICA

3.1 ANÀLISI DESCRIPTIVA DE LA MCVL

Després d'haver vist l'estructura, la composició i la descripció de les taules i variables corresponents, en aquest apartat s'analitzarà el que s'ha dut a terme amb algunes de les variables esmentades anteriorment, com la creació de nous dataframes amb la unió de diverses taules de la base de dades MCVL. Finalment s'observarà la relació entre l'últim salari dels individus amb la seva primera pensió, és a dir, la taxa de substitució.

Per a poder aconseguir veure aquests resultats, s'ha hagut de fer de dur a terme una sèrie de passos que s'explicaran en els següents apartats.

3.1.1 LECTURA I PROCESOS TAULA *PERSONAL*

Primer de tot, es va procedir a la lectura de l'arxiu "MCVL2015PERSONAL_CDF.txt" amb gairebé 1,2 millions d'observacions al programa *RStudio*, arxiu el qual conté tota la informació de la taula *PERSONAL*, mitjançant la comanda *read.table, d*onant nom al nou dataframe *personal.*

Una vegada llegida la taula *PERSONAL*, es vol obtenir l'edat dels individus d'aquesta. Això es du a terme aplicant la comanda *floor* (permet agafar un sol argument numèric "x" i retorna un vector numèric que conté els nombres enters) a la variable *V2* (*1.002 FECHA DE NACIMIENTO),* la qual està previament dividida entre 100 per a que ens quedi només l'any de naixement i d'aquesta manera, es pot restar a l'any 2015. Aquesta nova variable es crea amb el nom d'"Edat" i s'afegeix a la taula *PERSONAL*.

Una vegada es té l'edat, es crea una taula de freqüències entre la variable nova creada Edat amb la *V3 (1.003 SEXO),* per poder observar el nombre d'individus que hi ha en cada edat exacta diferenciant pel gènere femení o masculí.

Es du a terme un etiquetatge de la variable *V3 (1.003 SEXO),* on els valors 0 s'etiqueten amb "NA", els 1 amb "Home" i els 2 amb "Dones".

Per a veure la freqüència absoluta i relativa d'homes i de dones en la taula *PERSONAL*, es creen uns dataframes amb la funció "table" per a la freqüencia absoluta i "prop.table" per a la freqüència relativa, on s'obtenen els següents resultats:

```
> taulafreq
  Sexe Frequència absoluta Frequència relativa
1 Dona               560810                46.94
2 Home               634024                53.06
3   NA                    3                 0.00
```

Taula 3.1 Freqüència absoluta i relativa de la variable Sexe

Es pot apreciar, que a la mostra hi ha aproximadament un 6% més d'homes que de dones, això, amb valors absoluts es transforma en 634.024 homes, 560.810 de dones i 3 NA's.

Com en la taula de freqüències anterior s'observa que hi ha NA's, es decideix procedir a eliminar-los i a repetir la taula.

```
> taulafreq
  Sexe Freqüència absoluta Freqüència relativa
1 Dona               560810                46.94
2 Home               634024                53.06
```

Taula 3.2 Freqüència absoluta i relativa de la variable Sexe sense NA's

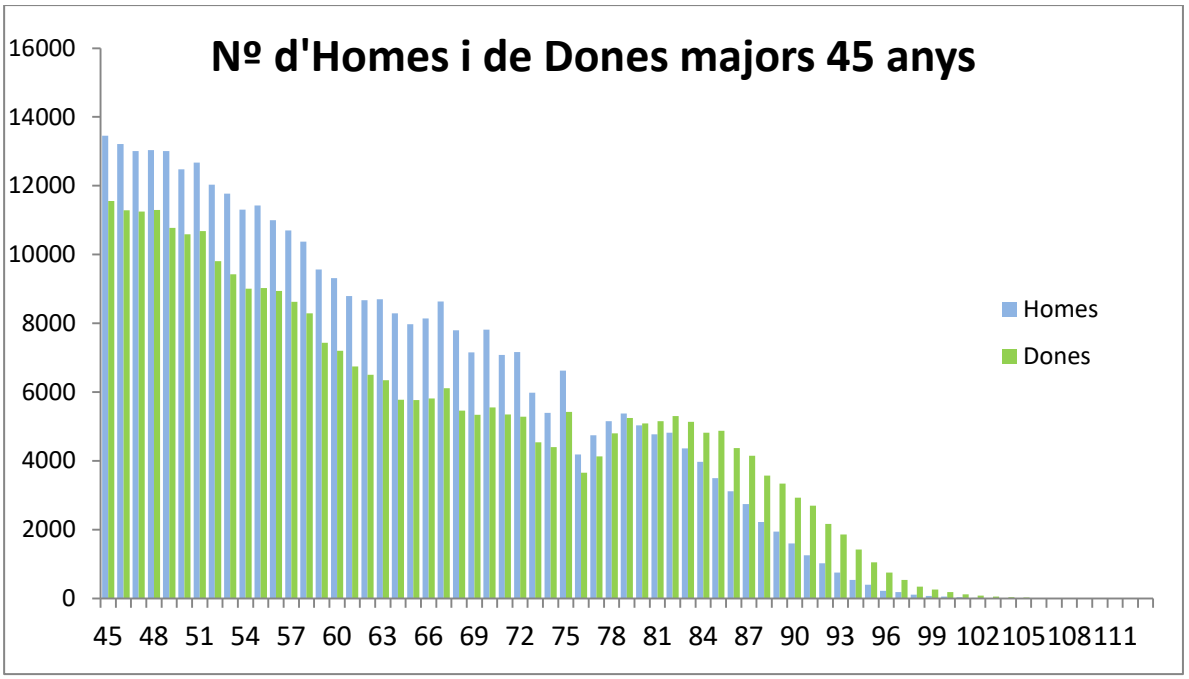
S'han recodificat els valors de la variable V3(*1.003 SEXO*), que conté el sexe de l'individu, passant a valdre 0 els homes i 1 les dones.

A més a més, l'etiquetatge previ de la variable V3, es modifica etiquetant els valors 0 amb "Home" i els 1 amb "Dones", ja que, ja no hi ha *NA's*.

A continuació, s'observarà una taula de freqüències i un gràfic d'homes i de dones però només dels adults de més de 44 anys:

```
 Sexe
Edat  Home  Dona        Edat  Home  Dona
  45 13454 11559         80  5029  5087
  46 13217 11285         81  4773  5153
  47 13012 11247         82  4819  5306
  48 13038 11298         83  4362  5133
  49 13010 10778         84  3974  4821
  50 12480 10586         85  3497  4877
  51 12676 10682         86  3113  4371
  52 12026  9807         87  2741  4147
  53 11774  9426         88  2219  3569
  54 11309  9010         89  1939  3340
  55 11430  9023         90  1599  2924
  56 10995  8942         91  1250  2694
  57 10698  8624         92  1018  2169
  58 10371  8293         93   749  1861
  59  9565  7436         94   537  1422
  60  9313  7200         95   401  1052
  61  8796  6748         96   220   751
  62  8667  6505         97   181   535
  63  8700  6342         98   110   340
  64  8285  5780         99    70   256
  65  7975  5768        100    54   180
  66  8144  5815        101    35   121
  67  8632  6112        102    18    84
  68  7793  5462        103    10    55
  69  7150  5335        104     8    30
  70  7816  5557        105     4    23
  71  7081  5345        106     3    10
  72  7166  5284        107     1     3
  73  5978  4536        108     0     3
  74  5396  4398        109     0     2
  75  6624  5418        110     2     2
  76  4184  3654        111     0     1
  77  4746  4125        113     0     1
  78  5157  4798        114     2     3
  79  5379  5247
```

Taula 3.3 Freqüència absoluta de la variable Sexe en adults de 45 anys o més



Gràfic 3.1 Freqüència absoluta de la variable Sexe en adults de 45 anys o més

S'ha realitzat un breu descriptiu per observar el contingut de la taula *PERSONAS.*

```
> summary(personal)
       V1                V2                V3                V4                V5                V6                V7
 Min.   :      31   Min.   :190101   Min.   :0.0000   N00    :1095791   Min.   : 0.00   Min.   : 0.00   Min.   :    0
 1st Qu.: 507642   1st Qu.:195205   1st Qu.:0.0000   N17    :  18840   1st Qu.:11.00   1st Qu.:10.00   1st Qu.:10000
 Median :1001646   Median :196707   Median :0.0000   N12    :  13315   Median :28.00   Median :28.00   Median :28007
 Mean   :1324618   Mean   :196465   Mean   :0.4694   N23    :   5692   Mean   :28.72   Mean   :24.58   Mean   :24711
 3rd Qu.:1615708   3rd Qu.:197902   3rd Qu.:1.0000   N22    :   5444   3rd Qu.:41.00   3rd Qu.:36.00   3rd Qu.:36000
 Max.   :3766547   Max.   :201510   Max.   :1.0000   N05    :   4550   Max.   :66.00   Max.   :52.00   Max.   :52001
                   NA's   :187                       (Other):  51202
       V8                V9                V10               edat
 Min.   :     0   N00    :1018213   31     :209879   Min.   : 0.00
 1st Qu.:     0            : 44300   20     :146000   1st Qu.: 36.00
 Median :     0   N12    : 15555   30     :128722   Median : 48.00
 Mean   :  3175   N17    : 15227   42     : 98949   Mean   : 50.41
 3rd Qu.:     0   N22    : 11743   40     : 98552   3rd Qu.: 63.00
 Max.   :201604   N09    : 10801   22     : 93179   Max.   :114.00
                  (Other): 78995   (Other):419553   NA's   :187
```

Taula 3.4 Descriptiu de la taula *PERSONAS*

Es coneix, que cada persona té uns canvis de cotització al llarg de la seva vida i per a la realització d'aquesta anàlisi interessa l'últim canvi que ha fet cada individu en concret, ja que, un individu pot realitzar diversos treballs al llarg de la seva vida. Per exemple, una persona pot treballar durant molts anys al sector del règim agrari, després al de mines, al d'autònoms, etc. Per aquest motiu, només es té en compte l'últim regim al que ha cotitzat.

3.1.2 LECTURA I PROCESOS TAULA *AFILIACIÓN*

Després d'haver realitzat els descriptius de la taula *PERSONAS*, es procedeix a realitzar la lectura de tots els fitxers de la taula *AFILIACIÓN* com es detalla a continuació:

Un cop s'ha procedit a la lectura del primer fitxer de la taula "MCVL2015AFILIAD1_CDF.TXT" amb 6,2 millions d'observacions, creant un nou dataframe amb el nom d'*afilia1b*, s'instal·la

el paquet “plyr”. “Plyr” és un conjunt d'eines que resolen un una serie de problemes comuns: cal dividir un gran problema en peces manejables, operar en cada peça i tornar a col·locar totes les peces. Amb aquest paquet i la funció “arrange”, es pot prosseguir a ordenar la variable *V1 (2.001 IDENTIFICADOR DE LA PERSONA FÍSICA (IPF)*), i tot seguit per *V7* (2.007 *FECHA REAL DE LA BAJA EN AFILIACIÓN*). D’aquesta manera s’obtindrà la data més recent de cada persona (així si un individu s’ha donat de baixa a la Seguretat Social més d’una vegada, només es tindrà en compte l’ultima de les vegades).

A més del paquet “plyr”, s’instal·la el paquet “dplyr”, el qual, entre d’altres usos, permet que les operacions en els marcs de dades es puguin expressar de manera concreta perquè no calgui repetir el nom del marc de dades.

Aquest paquet, permet agrupar per cada persona i de cada una s’està agafant la data més recent que ha canviat de cotització, creant la nova variable *novaV7.* També s’ha creat *count*, que és el número de vegades que ha canviat el seu estat a l’afiliació i *novaV2* que és el tipus de regim de cotització de cada individu. D’aquesta manera, s’eliminen els registres sobrants dels individus obtenint una fila per a cada individu amb la data de la seva última baixa a l’afiliació.

Es realitzen els mateixos passos, per als altres tres fitxers "MCVL2015AFILIAD2_CDF.TXT" amb 6 millions d’observacions, "MCVL2015AFILIAD3_CDF.TXT" amb 6,6 millions d’observacions i "MCVL2015AFILIAD4_CDF.TXT" amb 3,3 millions d’observacions de la taula *AFILIACIÓN*.

Un cop es tenen els 4 fitxers de la taula *AFILIACIÓN* units, amb les *variables V1, novaV7*, *count* i *novaV2*, es procedeix a unir el dataframe *personal* amb el dataframe *unio3*, el qual, s’ha creat amb la unió dels 4 fitxers d’*AFILIACIÓN*.

Amb la unió d’algunes variables de les taules *PERSONAL* i *AFILIACIÓN,* es pot extreure per sexes, el nombre d’individus de 45 anys en endavant que hi ha per a cada tipus de règim de cotització, com es pot veure a continuació:

	111	112	114	115	121	131	132	134	135	136	137	138	140	161	163	521	531	540	611
45	9870	22	0	0	0	0	21	0	0	0	0	33	6	161	461	2712	0	3	8
46	9667	18	0	0	0	0	16	2	0	0	0	29	11	184	413	2685	1	2	9
47	9514	14	1	0	0	0	26	2	0	0	0	21	18	173	443	2611	1	7	11
48	9523	17	0	0	0	0	13	2	0	0	1	26	15	186	374	2658	0	5	19
49	9486	9	0	0	3	0	11	6	0	0	2	29	13	188	349	2687	0	3	20
50	9078	21	0	0	0	0	18	2	0	0	1	15	6	175	337	2607	1	7	21
51	9119	10	0	0	0	0	14	0	0	0	0	24	27	203	359	2669	2	6	23
52	8544	19	0	0	0	0	9	1	0	0	0	35	17	203	353	2615	1	15	20
53	8437	11	0	0	0	0	17	2	0	0	0	25	27	167	329	2488	1	7	29
54	7992	10	0	0	3	0	8	1	0	0	0	21	28	154	344	2490	1	7	36
55	8107	7	0	0	0	0	9	0	0	0	0	34	103	147	289	2475	1	13	20
56	7765	10	0	0	0	0	6	0	0	0	0	26	152	123	267	2345	2	27	30
57	7517	9	0	0	2	0	8	0	0	1	0	18	185	146	238	2273	0	15	40
58	7340	9	0	1	0	0	10	0	0	0	0	14	261	125	221	2101	2	24	29
59	6676	7	0	0	2	0	3	2	0	0	0	23	335	113	176	1919	3	41	34
60	6321	4	0	0	1	0	4	1	0	0	0	13	361	86	166	2003	0	64	35
61	5954	7	0	0	1	0	4	1	0	0	0	9	431	95	162	1818	1	53	30
62	5968	8	0	0	1	0	7	1	0	0	0	10	433	109	132	1667	2	65	50
63	5848	10	0	1	0	0	8	0	0	0	0	5	457	120	135	1720	2	95	52
64	5540	9	1	0	1	0	4	1	0	0	0	3	437	94	135	1682	3	104	44
65	5233	7	1	0	3	0	4	0	0	0	0	8	388	63	139	1694	2	120	43
66	5382	9	1	0	2	0	4	0	1	0	0	6	404	78	137	1686	3	131	50
67	5735	17	0	1	6	0	8	0	0	0	0	7	361	106	137	1732	2	152	80
68	5158	12	0	0	2	0	3	0	0	1	0	7	321	74	127	1579	1	158	79
69	4651	17	0	1	1	0	1	0	0	0	0	1	301	3	106	1525	0	111	109
70	5087	8	1	2	2	0	3	0	0	0	0	2	345	7	123	1545	0	135	142

Taula 3.5 Nombre d'homes majors o iguals a 45 anys per a cada tipus de règim de cotització (Part 1)

	613	640	721	740	811	812	813	814	821	822	823	825	840	899	911	940	1211	1221	1240
45	3	0	0	0	51	6	4	8	0	0	1	9	3	12	15	4	0	0	0
46	1	0	3	0	44	5	4	6	0	0	3	11	5	19	11	10	0	1	0
47	1	0	3	0	34	4	11	4	1	0	4	15	3	18	9	15	0	0	0
48	1	0	0	0	42	10	11	7	4	2	2	11	3	19	10	28	1	2	0
49	3	0	3	0	35	5	9	7	1	2	3	11	5	20	9	33	0	0	1
50	2	0	4	0	35	9	6	3	1	5	4	6	2	13	7	26	0	0	0
51	4	0	5	0	46	8	8	5	1	4	0	8	5	19	11	43	0	1	0
52	0	0	7	0	34	6	8	2	1	4	0	9	1	15	5	37	0	0	0
53	3	0	6	0	48	9	5	9	1	4	1	12	2	16	7	40	1	0	0
54	3	0	12	0	39	13	8	7	0	0	1	11	5	9	7	42	0	1	0
55	3	0	11	0	50	8	7	4	1	0	3	11	2	17	36	19	0	0	0
56	6	0	14	0	33	5	3	4	0	0	1	9	7	17	75	4	0	1	0
57	5	0	18	0	40	3	5	3	3	2	1	5	5	31	66	8	0	0	0
58	2	0	7	0	38	11	6	1	2	0	1	9	7	22	65	4	0	1	0
59	4	0	13	0	33	8	4	5	2	1	1	1	4	25	55	8	1	2	0
60	6	0	22	0	45	12	5	5	0	2	5	9	5	28	55	5	0	0	0
61	4	0	17	0	44	8	4	6	1	1	2	8	8	20	41	9	1	0	0
62	5	0	22	0	29	6	3	5	0	0	1	9	4	23	52	11	0	0	0
63	6	0	27	0	39	11	6	0	1	1	1	6	3	22	40	23	4	4	0
64	9	0	31	0	26	7	8	4	0	0	5	9	12	17	31	13	1	2	0
65	10	0	39	0	40	8	7	3	0	5	4	9	6	23	25	20	1	1	0
66	10	0	34	0	44	5	8	6	1	3	0	9	10	28	24	13	0	1	0
67	14	0	50	0	39	5	1	3	0	0	4	10	6	22	36	11	5	0	0
68	10	0	45	2	26	6	9	8	1	2	1	7	9	23	25	12	3	2	0
69	33	1	52	1	31	7	3	5	1	0	2	10	10	31	28	5	0	1	0
70	38	2	78	0	33	4	7	3	1	3	1	6	7	36	40	14	3	6	1

Taula 3.6 Nombre d'homes majors o iguals a 45 anys per a cada tipus de règim de cotització (Part 2)

	111	112	115	121	132	134	135	136	137	138	140	161	163	521	531	540	611
45	8937	10	0	0	65	7	0	0	0	570	23	240	197	1420	2	4	12
46	8689	7	0	0	66	9	0	0	1	572	27	247	220	1356	1	3	16
47	8613	8	0	0	64	6	0	0	3	574	25	270	208	1363	0	6	26
48	8616	8	0	0	69	5	0	1	0	576	33	282	175	1423	0	5	20
49	8120	7	0	0	67	11	0	0	0	539	41	282	139	1432	5	3	20
50	8008	8	0	1	51	5	0	0	0	499	26	278	153	1418	3	8	16
51	8007	6	0	0	70	9	0	0	2	500	41	279	151	1440	2	12	23
52	7223	12	0	0	68	7	0	0	1	545	42	223	175	1366	0	8	23
53	6872	6	0	0	74	11	0	0	0	493	74	182	194	1335	1	4	26
54	6636	6	0	0	67	3	0	0	0	464	73	187	177	1231	2	12	28
55	6533	2	0	0	57	9	0	0	0	492	98	201	200	1211	3	17	37
56	6446	6	0	0	60	6	1	0	0	508	117	194	153	1258	1	13	35
57	6110	3	0	0	60	5	0	0	1	467	143	188	169	1228	1	29	39
58	5867	6	0	0	51	6	0	0	1	399	211	173	167	1147	1	35	45
59	5163	2	0	0	38	2	0	0	1	402	216	171	127	1034	0	26	46
60	4809	6	2	3	40	4	0	1	0	357	250	163	139	1109	0	49	51
61	4470	6	0	0	49	8	0	0	0	337	262	159	108	957	3	77	57
62	4190	2	0	1	40	5	0	0	0	312	258	148	136	992	1	69	76
63	3954	8	0	0	47	2	0	0	0	291	317	169	126	973	1	83	69
64	3514	10	0	0	34	2	0	0	1	249	304	134	118	900	0	88	67
65	3311	8	0	1	22	3	0	0	0	241	317	111	123	967	0	125	61
66	3126	4	0	1	37	3	0	1	0	202	317	145	132	1038	1	138	96
67	3123	9	0	1	29	1	0	0	0	193	351	196	85	1091	0	139	85
68	2686	10	0	4	32	8	0	0	0	101	274	126	136	940	0	149	95
69	2533	6	0	0	24	6	0	0	0	38	224	18	141	913	0	124	175
70	2401	6	0	1	30	3	0	2	0	27	238	16	132	992	0	113	190
71	2174	7	0	1	19	3	0	0	0	6	189	7	125	907	0	124	157
72	2024	9	0	1	18	3	0	0	0	11	129	5	128	844	0	134	164
73	1627	7	1	1	21	2	0	0	0	8	112	2	86	646	0	82	158
74	1458	8	0	0	29	5	0	0	0	5	103	1	104	610	0	83	135

Taula 3.7 Nombre de dones majors o iguals a 45 anys per a cada tipus de règim de cotització (Part 1)

	613	640	721	740	811	812	813	814	825	840	899	911	940	1211	1221	1240
45	0	0	0	0	4	0	0	0	6	0	0	0	0	0	4	0
46	1	0	1	0	6	0	0	1	5	0	1	2	0	1	1	0
47	2	0	1	0	4	0	0	0	4	0	3	1	0	2	1	0
48	2	0	1	0	10	0	0	0	10	0	0	0	0	2	2	0
49	0	0	3	0	5	0	0	0	6	0	1	1	0	7	8	0
50	2	0	2	0	6	0	1	1	5	1	2	0	0	4	4	0
51	5	0	0	0	5	0	0	1	15	0	2	2	1	4	9	1
52	5	0	2	0	4	0	0	0	6	0	0	0	0	6	10	0
53	5	0	4	0	3	0	0	0	10	0	2	1	0	12	14	1
54	3	0	4	0	5	0	0	0	7	0	1	0	0	13	5	0
55	3	1	7	0	3	0	0	0	11	0	1	0	0	14	16	0
56	3	0	4	0	2	0	0	0	6	0	0	1	1	10	18	0
57	1	0	9	0	3	0	0	0	13	0	2	0	1	17	15	0
58	3	0	12	1	4	0	0	0	5	1	1	1	0	17	13	0
59	3	0	8	0	1	0	0	0	12	0	2	0	0	21	21	0
60	3	0	15	2	2	0	0	0	5	1	2	1	0	18	26	0
61	6	0	15	0	1	0	0	0	12	0	6	0	1	24	25	1
62	5	0	17	0	0	0	0	1	9	0	4	0	0	34	32	0
63	2	0	21	0	2	0	0	0	8	1	1	0	0	34	37	0
64	3	0	25	2	0	0	0	0	7	0	0	0	0	45	44	0
65	6	1	37	0	2	0	0	0	9	0	3	0	0	49	72	4
66	4	0	28	1	0	0	0	0	17	0	1	0	0	58	94	3
67	9	3	37	0	3	0	0	1	16	0	0	1	0	69	128	2
68	2	0	43	0	1	0	0	0	8	0	0	0	0	94	153	2
69	17	2	53	1	2	0	0	0	9	2	2	0	0	106	159	9
70	17	4	53	2	0	0	0	0	7	7	3	1	0	121	166	6
71	13	4	84	4	0	0	0	1	9	3	0	1	0	135	158	7
72	20	2	80	6	1	0	1	0	7	1	3	1	0	132	186	6
73	11	3	196	6	1	0	0	0	4	0	3	0	0	112	147	5
74	9	2	251	4	1	0	1	0	8	2	0	1	0	117	135	8

Taula 3.8 Nombre de dones majors o iguals a 45 anys per a cada tipus de règim de cotització (Part 2)

S'adjunta a continuació, una taula amb la descripció dels codis de la variable novaV2.

RÉGIMEN DE COTIZACIÓN

(Las claves seguidas de "B" corresponden a regímenes extinguidos)

Clave		Denominación del régimen o sistema especial
0111		REGIMEN GENERAL
0112		REGIMEN GENERAL (ARTISTAS)
0113		REGIMEN GENERAL (DE UN REGIMEN EXTINGUIDO)
0114		REGIMEN GENERAL (DE UN REGIMEN EXTINGUIDO)
0115	B	REGIMEN GENERAL (FERROVIARIOS)
0121		REGIMEN GENERAL (REPRESENTANTES COMERCIO)
0131		REGIMEN GENERAL (SISTEMA ESPECIAL COTIZACION)
0132		REGIMEN GENERAL (SISTEMA ESPECIAL COTIZACION)
0133	B	REGIMEN GENERAL (SISTEMA ESPECIAL COTIZACION)
0134		REGIMEN GENERAL (SISTEMA ESPECIAL COTIZACION)
0135		REGIMEN GENERAL (SISTEMA ESPECIAL COTIZACION)
0136		REGIMEN GENERAL (SISTEMA ESPECIAL COTIZACION)
0137		REGIMEN GENERAL (SISTEMA ESPECIAL COTIZACION)
0138		REGIMEN GENERAL (S.E.C. EMPLEADOS HOGAR)
0140		REGIMEN GENERAL (CONVENIO ESPECIAL)
0150		REGIMEN GENERAL (SITUACION ESPECIAL)
0151		CUENTA CONVENCIONAL
0152		CUENTA CONVENCIONAL
0160		CUENTA CONVENCIONAL
0161		REGIMEN GENERAL (S.E.C. AGRARIO INACTIVIDAD)
0163		REGIMEN GENERAL (S.E.C. AGRARIO CCC ACTIVIDAD)
0170		CUENTA CONVENCIONAL
0180		CUENTA CONVENCIONAL
0521		R.E. AUTONOMOS
0522	B	R.E. AUTONOMOS (DE UN REGIMEN EXTINGUIDO)
0531		R.E. AUTONOMOS (PREST. CESE DE ACTIVIDAD)
0540		R.E. AUTONOMOS (CONVENIO ESPECIAL)
0611		R.E. AGRARIO CUENTA AJENA
0612	B	R.E. AGRARIO CUENTA AJENA (J.REALES)
0613	B	R.E. AGRARIO CUENTA AJENA (COTIZ.DE EMPRESAS)
0640	B	R.E. AGRARIO CUENTA AJENA (CONVENIO ESPECIAL)
0650	B	R.E. AGRARIO CUENTA AJENA (SITUACIÓN ESPECIAL)
0721	B	R.E. AGRARIO CUENTA PROPIA
0722	B	R.E. AGRARIO CUENTA PROPIA (JORNADAS TEORICAS)
0723	B	R.E. AGRARIO CUENTA PROPIA (COBERTURA A.T.S.S.)
0740	B	R.E. AGRARIO CUENTA PROPIA (CONVENIO ESPECIAL)
0800	B	R.E. MAR (CTA.CONVENCIONAL COTIZ.EN DESEMPLEO)
0811		R.E. MAR (C.AJENA GRUPO 1)
0812		R.E. MAR (C.AJENA GRUPO 2A)
0813		R.E. MAR (C.AJENA GRUPO 2B)
0814		R.E. MAR (C.AJENA GRUPO 3)
0821		R.E. MAR (ASIMILADOS C. AJENA GRUPO 1)
0822		R.E. MAR (ASIMILADOS C.AJENA GRUPO 2A)
0823		R.E. MAR (ASIMILADOS C.AJENA GRUPO 2B)

Taula 3.9 Claus i denominacions del règim o sistema especial (Part 1)

RÉGIMEN DE COTIZACIÓN

(Las claves seguidas de "B" corresponden a regímenes extinguidos)

Clave		Denominación del régimen o sistema especial
0825		R.E. MAR (AUTONOMOS)
0831		R.E. MAR (AUTONOMOS PREST. CESE DE ACTIVIDAD)
0840		R.E. MAR (CONVENIO ESPECIAL)
0850		R.E. MAR (C.AJENA SITUACIÓN ESPECIAL)
0899		R.E. MAR (CTA.CONVENCIONAL COTIZ.EN DESEMPLEO)
0911		R.E. MINERIA CARBON
0940		R.E. MINERIA CARBON (CONVENIO ESPECIAL)
0950		R.E. MINERIA CARBON (SITUACIÓN ESPECIAL)
1200	B	R.E. EMPLEADOS HOGAR
1211	B	R.E. EMPLEADOS HOGAR (FIJOS)
1221	B	R.E. EMPLEADOS HOGAR (DISCONTINUOS)
1240	B	R.E. EMPLEADOS HOGAR (CONVENIO ESPECIAL)
1250	B	R.E. EMPLEADOS HOGAR (FIJOS SIT.ESPECIAL)

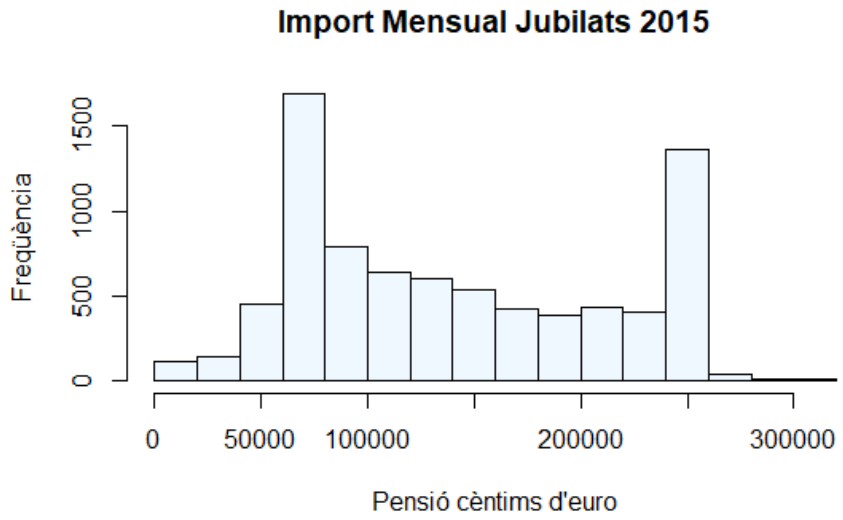
Taula 3.10 Claus i denominacions del règim o sistema especial (Part 2)

3.1.3 LECTURA I PROCESOS TAULA *PENSIONES*

Una vegada llegit l'únic fitxer de la taula *PENSIONES,* "MCVL2015PRESTAC_CDF.TXT" amb 4,6 millions d'oservacions, es vol obtenir l'any de jubilació. Per obtindre'l s'agafen els primers quatre caracters de la variable V11 (*4011: FECHA EFECTOS ECONOMICOS DE LA PENSIÓN*) amb la funció *substr,* la qual, conté l'any i el mes de jubilació i només es vol extreure l'any.

Una vegada tenim l'any de jubilació net (sense el mes), es crea un dataframe amb el nom de *jubilats2015*, el qual està format per: l'any de jubilació del 2015 i que l'any natural al que es refereixen les dades històriques de la mostra sigui del 2015, així s'obtindran els jubilats del 2015, també es té en compte la condició d'agafar els registres quan la V4 (*4.004 CLASE DE LA PRESTACIÓN*) prengui el valor 21, ja que, són les pensions de jubilació, és a dir, són els que tenen l'alta de pensió de jubilació i per últim, es vol la V10 (*4.010 RÉGIMEN DE LA PENSIÓN*) amb els valors 1 o 75, ja que, aquests pertanyen a la condició de que siguin de règim general on s'excloent els miners, els policies, etc.

Tots aquest són jubilats de l'any 2015 (tenen 65 anys, són recent jubilats), que es donen d'alta al 2015 i tenen una pensió de jubilació de règim general.



Gràfic 3.2 Histograma pensions jubilats any 2015

Es du a terme un histograma per veure amb claretat l'import mensual de les pensions dels jubilats del 2015. Es realitza amb el dataframe *jubilats2015* i mostrant la variable V20 *(4.020 IMPORTE MENSUAL TOTAL DE LA PRESTACIÓN),* que és la suma dels imports mensuals de pensió efectiva, de la revalorització, del complement de garantia de mínims i d'altres complements. Inclou, quan és necessari, els complements de gran invalidesa i les seves revalorizacions, per tant, es pot dir que és l'import mensual total de la pensió de cada individu i està expressat en cèntims d'euro.

```
summary(jubilats2015$V20)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0   73570  125600  141000  212800  312200
```

Taula 3.11 Descriptiu del nombre d'individus jubilats l'any 2015

```
 dim(jubilats2015)
[1] 8028   43
```

Taula 3.12 Dimensions del dataframes jubilats2015

Es pot veure en aquests resums, que hi ha 8.028 individus jubilats l'any 2015 d'aquesta base i amb les condicions citades després de l'histograma. També es pot apreciar, que rebre un pensió va des dels 0€ fins als 3.122€ mensuals, amb una mitjana de 1256€ mensuals. On un 25% d'aquests individus, és a dir, 2.007 individus reben una pensió de menys de 735,70€ al mes.

Cal recalcar, que la pensió prengui valor 0, vol dir que encara no s'ha tramitat i pagat la pensió a alguns individus.

Ara partint del dataset *jubilats2015,* s'agafen només els jubilats que estiguin donats d'alta variable V21 *(4.021 SITUACIÓN DE LA PRESTACIÓN (CAUSA DE BAJA)*) amb els valors: 001, 002 o 003, ja que, els altres implica que estan donats de baixa per motius com baixa per mort, baixa per sanció, etc.

```
dim(jubilats2015alta)
[1] 7991   43
```

Taula 3.13 Dimensions del dataframes jubilats2015alta

```
 (dim(jubilats2015)-dim(jubilats2015alta))
[1] 37  0
```

Taula 3.14 Diferència entre total de jubilats l'any 2015 i els que estan donats d'alta

Es pot veure com ha disminuit el nombre d'individus passant de 8.028 a 7.991. Això vol dir que s'han donat de baixa 37 persones l'any 2015 (el mateix any en que s'havien jubilat).

Al tenir tantes dades es possile que hi hagi algun duplicat, Per veure si hi ha duplicats s'instal·la el paquet *sqldf*:

Es fa un *select distinct* de *jubilats2015* amb V1 que és on hi ha l'identificador per veure si hi ha algú repetit. Se'n troba un per tant, possiblement és algú que té dues altes o treballava a dos llocs.

Per veure qui és l'individu que està dues vegades, es fa aquesta subquery (on es compta quantes vegades surt cadascú i es selecciona el que surt més d'una vegada):

```
jubilats2015V1<-sqldf("select v1 from (select V1, count(*) as n from jubilats2015alta group by V1) where n>1")
jubilats2015V1
      v1
1 538544
```

Taula 3.15 Identificadors dels individus que es repeteixen

Ara s'ha de prendre la decisió de que es fa amb aquest individu. Es podria escollir el primer dels dos duplicats, però al tenir tantes dades es preferible eliminar les dues files que contenen aquest identificador, ja que, no es té cap criteri per decidir quin és el bo.

No serà necessari eliminar aquests casos dels altres fitxers, ja que, com s'utilitzarà un "Left Join" ja no les agafarà.

Una vegada s'ha netejat el duplicat, es procedeix a fer una reducció d'aquest últim dataframe, escollint només les variables que es creuen necessàries per a l'estudi. Les variables incloses en el nou dataset són les següents: V1 (*1.001 IDENTIFICADOR DE LA PERSONA FÍSICA (IPF)*), V4 (*4.004 CLASE DE LA PRESTACIÓN)*, V10 *(4.010 RÉGIMEN DE LA PENSIÓN)*, V11 *(4.011 FECHA DE EFECTOS ECONÓMICOS DE LA PENSIÓN)*, V12 *(4.012 BASE REGULADORA)*, V15 *(4.015 AÑOS CONSIDERADOS COTIZADOS PARA LA JUBILACIÓN)*, V16

(4.016 IMPORTE MENSUAL DE LA PENSIÓN EFECTIVA), V20 *(4.020 IMPORTE MENSUAL TOTAL DE LA PRESTACIÓN)*, V21 *(4.021 SITUACIÓN DE LA PRESTACIÓN (CAUSA DE BAJA))*, V22 *(4.022 FECHA DE SITUACIÓN DE LA PRESTACIÓN)*, V36 *(4.036 PENSIÓN LIMITADA)* i anyjub (variable creada anteriorment, la qual, conté l'any de jubilació).

A continuació, es presenta una mostra del dataframe *jubi2015alta* on s'observa les el contingut de les variables de 6 individus:

	V1	V4	V10	V11	V12	V15	V16	V20	V21	V22	V36	anyjub
194		21	1	201504	223760	41	211174	211174	0	201504	0	2015
369		21	1	201502	73328	34	51330	59350	0	201502	0	2015
370		21	1	201502	197016	43	161553	161553	0	201502	0	2015
525		21	1	201505	119862	43	91095	91095	0	201505	0	2015
552		21	1	201504	285781	47	256088	256088	0	201504	1	2015
559		21	1	201501	206049	35	148355	148355	0	201501	0	2015

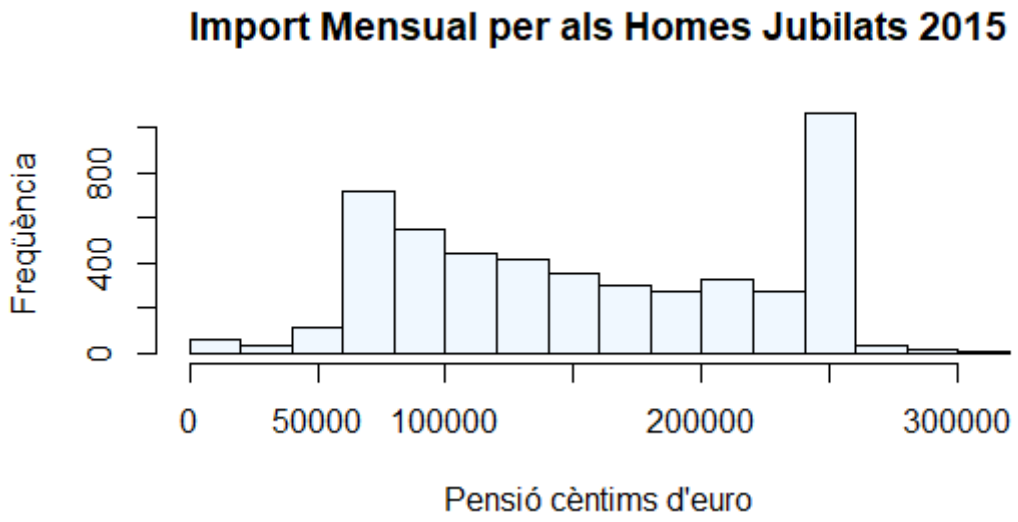
Taula 3.16 Mostra del dataframes *jubi2015alta*

Ara partint d'aquest nou dataset *jubi2015alta*, es vol procedir a unir-lo amb el dataframe *personal.* Abans de procedir a unir els dos dataframes, es crearà un de nou, *personalmini*, on serà una reducció del dataset *personal,* ja que, només es necessiten dues variables per a poder reconèixer si l'identificador correspon al sexe femení o masculí.

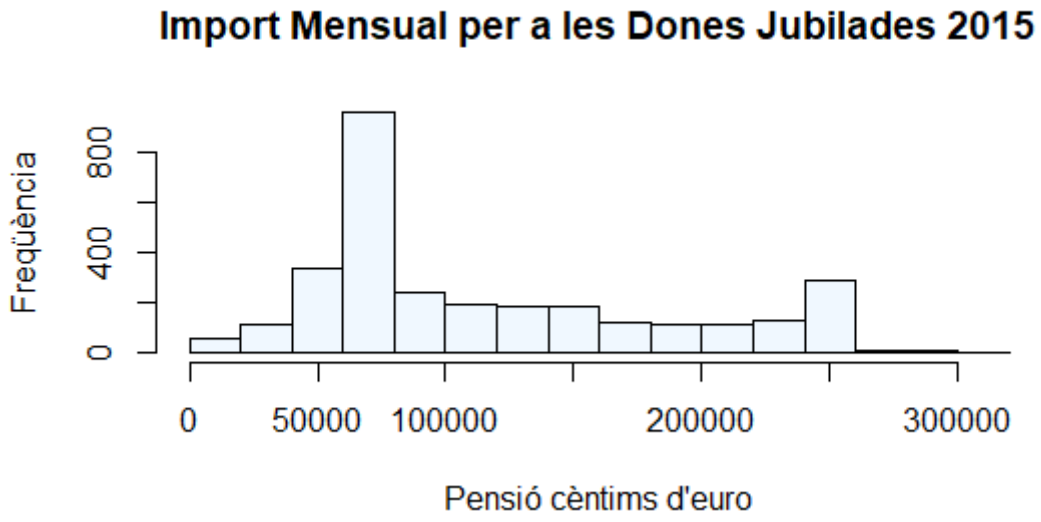
S'utilitza la comanda "left join" del paquet "sqldf" per a unir *jubilats2015alta* amb *personalmini,* partint de la V1 (*1.001 IDENTIFICADOR DE LA PERSONA FÍSICA (IPF))* que està en els dos donant un nom nou *F1*.

Després de tots aquests procediments, es pot donar pas a fer dos histogrames, un per als homes i un per a les dones, així amb la variable V20 i la V3 es podrà veure per separat els jubilats donats d'alta per cada sexe i l'import mensual del total de la prestació en cèntims d'euro. Com es pot observar a continuació:

HISTOGRAMES I RESUMS PENSIONS PER GÈNERE DE JUBILATS L'ANY 2015



Taula 3.17 Import mensual per als homes jubilats de l'any 2015



Taula 3.18 Import mensual per a les dones jubilades de l'any 2015

Pensions dels homes

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	91580	148600	157200	231200	312200

Taula 3.19 Descriptiu de les pensions dels homes jubilats de l'any 2015

Pensions de les dones

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	60190	84000	114500	161100	309100

Taula 3.20 Descriptiu de les pensions dels dones jubilats de l'any 2015

En referència als histogrames anteriors, es poden extreure resultats a primera vista amb molta claredat.

Es pot observar, que el salari on hi ha un nombre més elevat d'homes és al voltant dels 2500€ respecte al de les dones que el més repetit és el que està al voltant dels 750€. A més a més, com es pot veure als resums, el 75% dels homes cobren una pensió mensual d'almenys 915,8€ en front al 75% de les dones que cobren almenys 601,9€. Per tant, hi ha una clara diferència entre els dos gràfics i els dos resums.

Es pot concloure, només disposant d'aquesta informació, es pot afirmar que els homes tenen unes pensions més favorables que les dones.

3.1.4 UNIÓ I PROCESSOS DE PERS_AFI_PREST

Es segueix amb la mateixa dinàmica d'anar unint els dataframes per aconseguir el dataframe final.

Es recorda que prèviament s'havia agafat l'última afiliació (entre d'altres variables), unida també al dataset *personal, m1.* D'aquest s'escullen les següents variables amb el nom de *PERS_AFI,* el qual conté: V1, edat, V2, V3, V4, V5, V7, novaV7, count i novaV2. Les quals les 4 primeres procedeixen al dataframe *personal* i les restants al d'*unió3* que provenen o estan calculades del fitxer d'*AFILIACIÓN*.

La variable novaV7 es recorda que és l'última afiliació i es requereixen les bases de cotització (es veurà més endavant), per comparar aquestes amb el que està cobrant de pensió cada individu.

Es pren com a base *F1*, el qual, s'uneix amb un "left join" a PERS_AFI, així es tindrà la informació escollida de les taules de personal, afiliació i prestació junt amb el que interessa per a l'estudi. Donant nom al nou dataframe, PERS_AFI_PREST on s'utilitza la variable V1 per a la connexió entre datasets.

Ara es fa un petit descriptiu de la variable V15 *(4.015 AÑOS CONSIDERADOS COTIZADOS PARA LA JUBILACIÓN)* amb "summary" per observar el nou dataframe.

```
summary(PERS_AFI_PREST$V15)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   32.00   39.00   36.53   43.00   61.00
```

Taula 3.21 Descriptiu de la variable V15 del dataframe PERS_AFI_PREST

S'aprecia que com a mínim hi ha individus amb 0 anys cotitzats i com a màxim 61. Surt un valor tant elevat com 61 anys, ja que, una persona no està obligada a jubilar-se amb 65 anys, sinó que, si ho vol pot fer amb 80 anys, per exemple. També s'observa que la meitat de les persones de la base de dades han cotitzat 39 anys i de mitjana n'han cotitzat 36,5.

3.1.5 LECTURA I PROCESOS TAULA COTIZACIÓN

Després d'haver realitzat els descriptius anteriors, encara falta per llegir les dades de l'última taula que s'utilitzarà en aquest estudi: la taula *COTIZACIÓN*, una taula que distribueix les seves dades en 13 arxius, de la qual es procedeix a realitzar la lectura de tots els fitxers d'aquesta com es detalla a continuació:

Un cop s'ha procedit a la lectura del primer fitxer de la taula "MCVL2015COTIZA1_CDF.TXT" amb 2,3 milions d'observacions, creant un nou dataframes, el qual, es reduirà per obtenir només les variables d'interès que són: V1 (*3.001 IDENTIFICADOR DE LA PERSONA FÍSICA (IPF))*, V3 *(3.003 AÑO DE COTIZACIÓN)* i tot seguit per V16 *(3.005 TOTAL ANUAL BASES DE COTIZACIÓN*), la qual, aquesta última és la suma de totes les columnes de la variable *3.004 BASE DE COTIZACIÓN MENSUAL CONT.COMUNES*.

Una vegada es té el primer fitxer llegit i reduït, amb les variables que es requereixen, es porta a terme un altre pas: es fa un "subset" a la variable any de cotització (V3) on es vol només els anys iguals a 2014, d'aquesta manera es podrà observar el que van cotitzar els individus l'any previ al 2015.

Com hi ha 13 fitxers que contenen tota la informació de *COTIZACIÓN,* es realitzen els mateixos passos, per als altres 11 fitxers:

"MCVL2015COTIZA2_CDF.TXT" amb 2,4 milions d'observacions,
"MCVL2015COTIZA3_CDF.TXT" amb 6,6 milions d'observacions,
"MCVL2015COTIZA4_CDF.TXT" amb gairebé 2,2 milions d'observacions,
"MCVL2015COTIZA5_CDF.TXT" amb gairebé 2,5 milions d'observacions,
"MCVL2015COTIZA6_CDF.TXT" amb 2,2 milions d'observacions,
"MCVL2015COTIZA7_CDF.TXT" amb 2,3 milions d'observacions,
"MCVL2015COTIZA8_CDF.TXT" amb 2,3 milions d'observacions,
"MCVL2015COTIZA9_CDF.TXT" amb 1,9 milions d'observacions,
"MCVL2015COTIZA10_CDF.TXT" amb gairebé 1,2 milions d'observacions,
"MCVL2015COTIZA11_CDF.TXT" amb 950 mil observacions i
"MCVL2015COTIZA12_CDF.TXT" amb 403 mil observacions.

L'últim fitxer de la taula *COTIZACIÓN,* correspon als individus per compte propi*,* el qual no es llegeix, ja que, per aquest anàlisi es decideix treballar només amb individus que tinguin una cotització per compte aliè i/o altres situacions especials, però no es vol tenir en compte els autònoms.

3.1.6 UNIÓ FINAL DE LES 4 TAULES

Una vegada s'han llegit els 12 fitxers i s'han realitzat els passos explicats prèviament a cada fitxer, es du a terme la unió d'aquests 12 donant lloc a un nou dataframe anomenat COTIZA, el qual, conté tota la informació de tots els individus de la base de dades respecte a la taula *COTIZACIÓN,* això si, exceptuant les variables que no s'han volgut incloure.

Seguint el mateix guió, es procedeix a la unió del dataframes PERS_AFI_PREST amb COTITZA que és el que s'acaba de crear amb els fitxers de cotització. S'uneixen amb l'ajuda del paquet "sqldf" partint de l'identificador, que com s'ha dit anteriorment està a tots els fitxers per facilitar les unions.

Finalment, s'aconsegueix un dataframe que conté tota la informació necessària per a realitzar la comparativa que es vol fer en aquest estudi, la de comparar l'últim salari amb la primera pensió respecte el gènere.

A continuació es presenta una mostra d'aquesta unió de les quatre taules nombrada PERS_AFI_PREST_COTIZA:

v1	v4	v10	v11	v12	v15	v16	v20	v21	v22	v36	anyjub	pv1	sexe	v1.1	edat	v2	v3	v5	v7	novaV7	count	novaV2	cv1	cv3	cv5
194	21	1	201504	223760	41	211174	211174	0	201504	0	2015	194	0	194	64	195101	0	1	1059	20150427	7	111	194	2014	3127257
369	21	1	201502	73328	34	51330	59350	0	201502	0	2015	369	0	369	61	195402	0	10	1059	20150226	56	111	369	2014	903600
370	21	1	201502	197016	43	161553	161553	0	201502	0	2015	370	0	370	62	195302	0	1	1059	20150223	10	111	370	2014	2602440
525	21	1	201505	119862	43	91095	91095	0	201505	0	2015	525	0	525	61	195405	0	1	1000	20150523	17	111	525	2014	903600
552	21	1	201504	285781	47	256088	256088	0	201504	1	2015	552	0	552	65	195004	0	29	1059	20150410	17	111	552	2014	667960
559	21	1	201501	206049	35	148355	148355	0	201501	0	2015	559	0	559	61	195401	0	34	1059	20150118	41	111	559	2014	903600

Taula 3.22 Mostra del dataframe PERS_AFI_PREST_COTIZA

3.1.7 DEPURACIÓ DE P_A_P_C

Ara el que es vol veure és quines persones han cotitzat més d'una vegada l'any 2015, sigui per canvi de feina instantània, eventual, etc. Per aconseguir això, utilitzant el paquet "dplyr" s'obtenen aquests resultats que s'uneixen al dataframes PERS_AFI_PREST_COTIZA donant un nou nom PERS_AFI_PREST_COTIZA_TOTAL. Passarà a dir-se P_A_P_C posteriorment, que serà aquest últim el que es depurarà.

Una vegada es sap que hi ha individus que surten dues vegades o més al dataframe, es procedeix a localitzar-los. Amb la funció "duplicated" es pot apreciar totes les files del data set on l'individu ha cotitzat més d'una vegada l'any 2015, això s'observa a la variable numccc, creada al dataset PERS_AFI_PREST_COTIZA_TOTAL que és la que compta quantes vegades surt l'identificador, per així saber is està duplicat o no.

Es decideix que amb els duplicats s'utilitzarà la funció "distinct", la qual reté només les files úniques i elimina totes les que tenien algun duplicat. Es decideix realitzar aquesta operació, amb motiu de ser coherents amb les decisions preses, ja que, si per exemple es decidís guardar el primer registre de tres que n'hi ha, no seria equiparable als altres registres ja que aquest igual té uns valors molt diferents a algú que ha cotitzat tot un mateix any seguit però perquè un potser ha cotitzat la meitat de mesos que l'altre i no per algun altre factor. Per tant, només es tindran en compte els individus amb les mateixes condicions respecte al nombre de vegades que han cotitzat durant l'any de referència, és a dir, només el que tinguin un registre (no hauran canviat de feina).

S'aprofita per eliminar tot tipus de duplicat, tant columnes, files (com s'acaba d'explicar), files que continguin valors negatius i files que continguin NA's. Un cop s'ha netejat el data set P_A_P_C, es denomina amb un nom nou PAPCNET.

Després de fer la neteja de P_A_P_C es passa de tenir 7.708 a 6.411 registres al nou dataframes PAPCNET.

3.1.8 CÀLCUL TAXA DE SUBSTITUCIÓ

Com s'ha esmentat al principi d'aquest document, la taxa de substitució també és coneguda com a taxa de reemplaçament, aquesta és un indicador de com un sistema de pensions aconsegueix o no proporcionar uns ingressos adequats en el moment de la jubilació respecte als ingressos que el treballador tenia quan estava en actiu. Aquesta taxa es calcula com el percentatge que suposa la pensió de jubilació sobre l'últim sou percebut en l'etapa laboral.

Un dels objectius d'aquest estudi, era arribar a conèixer la taxa de substitució. Gràcies a tots els passos que s'han dut a terme fins aquest punt, garanteixen que es pugui fer el càlcul d'aquesta.

Primer de tot, es necessita obtindre la pensió per als 12 mesos de l'any de cada individu. S'afegirà una nova variable, nombrada pensio12, a PAPCNET, la qual contindrà la suma de les pensions mensuals, ja que, s'haurà multiplicat per dotze la variable V20 (*4.020 IMPORTE MENSUAL TOTAL DE LA PRESTACIÓN)* i d'aquesta manera es tindrà el total anual.

Una vegada es té la pensió anual per individu, es pot dur a terme els càlculs per obtindré la taxa de substitució. Es torna a repetir el procés d'afegir una nova variable, taxa, amb el procediment següent: es divideix la variable pensio12 entre la variable cV5, (antiga V16 (*3.005 TOTAL ANUAL BASES DE COTIZACIÓN))*, amb aquest càlcul s'obtindrà la taxa de substitució de cadascun dels individus del dataframe PAPCNET.

Es realitzen els següents descriptius de les variables que s'acaben de crear:

```
summary(PAPCNET$taxa)
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.0577   0.7309   0.9550   4.1900   2.9450 877.9000
```

Taula 3.23 Descriptiu de la variable taxa del dataframe PAPCNET

Es pot observar en l'anterior descriptiu, que hi ha uns valors fora del que seria habitual obtenir en la taxa de substitució. Això és donat perquè la taxa de substitució només té en compte l'últim salari que ha cobrat aquell individu i la pensió és calculada amb els últims 15 anys que ha cotitzat la persona.

```
summary(PAPCNET$pensio12)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
26570  951700 1579000 1739000 2508000 3746000
```

Taula 3.24 Descriptiu de la variable pensio12 del dataframe PAPCNET

Es pot observar que la pensió mínima es de 265,7€ anuals i la màxima de 37.460€. La pensió mitjana anual que rep un individu és de 17.390€. A més, s'aprecia que un 25% dels individus reben com a màxim 9.517€ i un 25% dels individus reben com a mínim 25.080€.

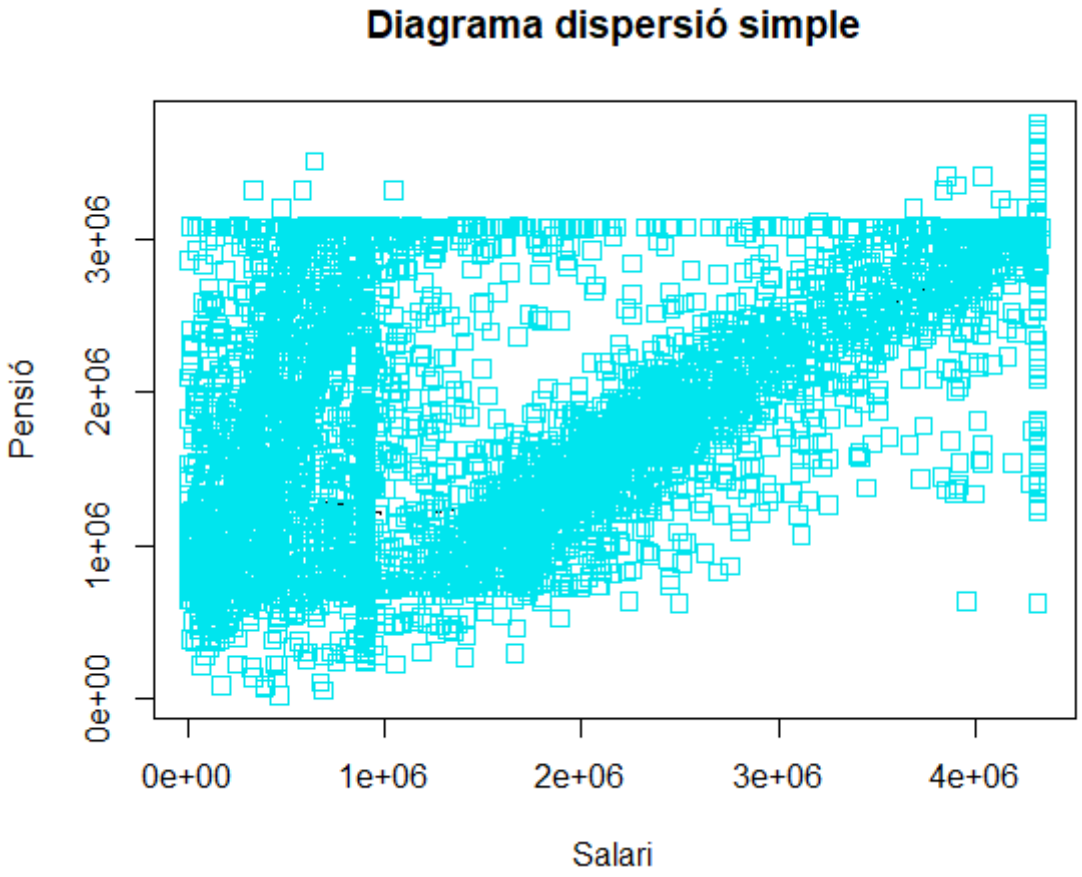
Per altra banda, en el següent descriptiu es poden observar les xifres del total anual de les bases de cotització cV5 (*3.005 TOTAL ANUAL BASES DE COTIZACIÓN*). On el mínim és de 2,5€ i el màxim de 43.400€ anuals, amb una mediana de 9.036€ anuals.

```
summary(PAPCNET$cV5)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 2557  665400  903600 1574000 2343000 4340000 
```

Taula 3.25 Descriptiu de la variable cV5 del dataframe PAPCNET

3.1.9 DIAGRAMA DE DISPERSIÓ SIMPLE

Ara per veure amb claredat les variables d'interès, es du a terme un diagrama de dispersió simple. Es realitzarà a partir de dos vectors que provenen del dataframe PAPCNET i així obtindré la relació entre el salari i la pensió.



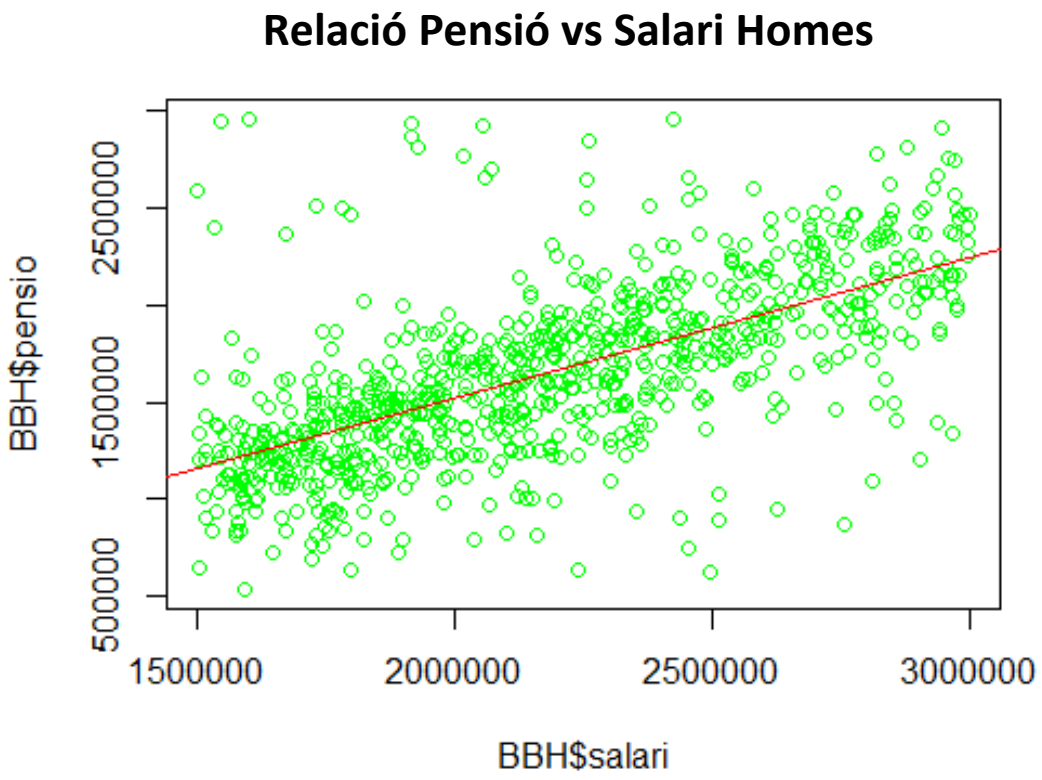
Gràfic 3.3 Diagrama de dispersió simple del Salari i la Pensió del dataframes PAPCNET

Es pot veure com aquest gràfic de dispersió segueix una tendència positiva a partir dels 10.000€ de l'eix Salari i de més dels 5.000€ de l'eix Pensió. Per tant, es pot creure que els individus amb valors anteriors a aquests, deuen ser casos on l'últim any de feina cotitzaven molt menys que durant els anys previs o viceversa. Fet el qual, normalment no és el més habitual, per aquest motiu s'haurà de prendre la següent mesura:

Per a tenir resultats més concisos, es decideix passar un filtre a aquest dataframe perquè tots aquests casos no es tinguin en compte, ja que, es pot afirmar que es poden considerar com a casos extrems i no interessa tindre'ls a la mostra, ja que, possiblement distorsionarien els resultats.

3.2 ANÀLISI DE REGRESSIÓ

Es decideix fer dos diagrames per separat, un diagrama per homes i un per dones. A més, es decideix filtrar individus que hagin tingut un salari anual superior a 15.000€ i inferior a 30.000€ l'any de referència, també és limita la pensió anual, escollint només els individus amb una pensió inferior a 30.000€ anuals i per últim, tenir els individus que no hagin canviat de feina.
D'aquesta manera es pretén disposar d'uns individus amb unes condicions mínimament similars i així poder centrar-se de manera més precisa en la diferència entre sexes respecte la taxa de substitució i no es distorsionen els resultats.



Gràfic 3.4 Diagrama de dispersió simple entre la Pensió i el salari dels homes del dataframe BBH

En aquest gràfic on es relaciona l'últim salari i la primera pensió dels homes, després d'aplicar els filtres prèviament esmentats, es pot veure amb claredat la tendència positiva que comença al voltant dels 11.000€ de pensió anuals per uns 15.000€ de salari cotitzats i va incrementant progressivament fins als 22.000€ de pensió per 30.000€ de salari aproximadament.

En el següent descriptiu, es pot observar la taxa de substitució per als homes. Aquesta oscil·la entre el 0,25 i l'1,90. Tot i que la mitjana és del 0,76. A més a més, el 50% dels homes tenen una taxa de substitució entre el 0,67 i el 0,83.

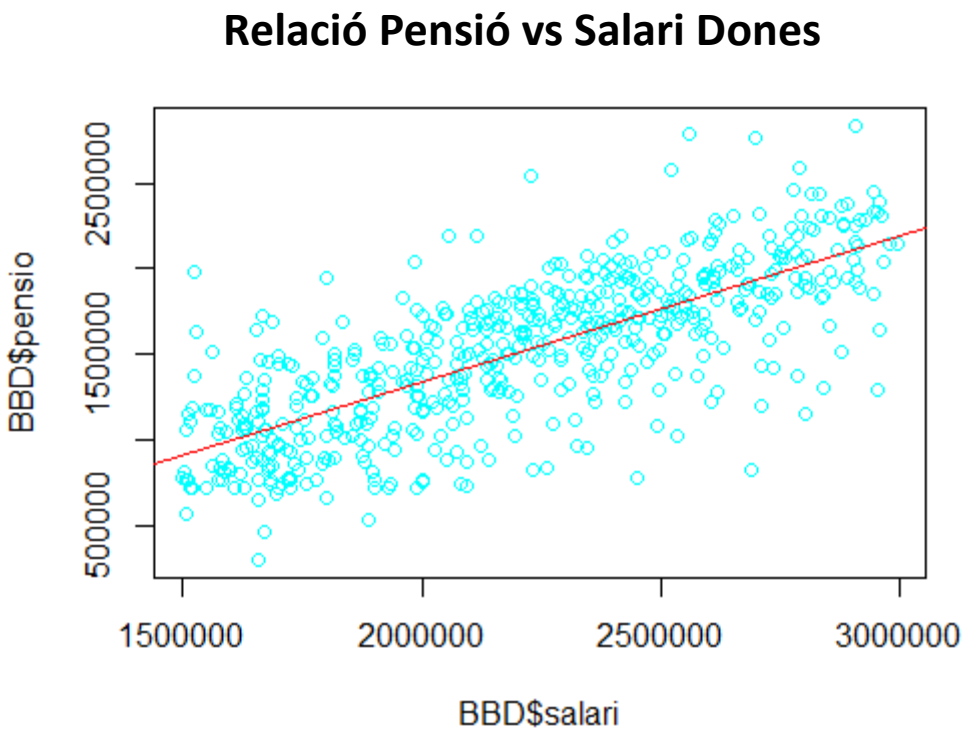
```
> summary(BBH$taxa)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2509  0.6708  0.7557  0.7584  0.8343  1.9010
```

Taula 3.26 Descriptiu de la variable taxa del dataframe BBH (Homes)

```
> summary(BBD$taxa)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1795  0.5860  0.6980  0.6810  0.7781  1.2990
```

Taula 3.27 Descriptiu de la variable taxa del dataframe BBD (Dones)

En l'anterior descriptiu, es pot observar la taxa de substitució per a les dones. La qual, oscil·la entre el 0,18 i l'1,30. Tot i que la mitjana és del 0,70. A més, el 50% de les dones tenen una taxa de substitució entre el 0,59 i el 0,78.



Gràfic 3.5 Diagrama de dispersió simple entre la Pensió i el salari de les dones del dataframe BBD

En el gràfic anterior, és relaciona l'últim salari i la primera pensió de les dones, d'igual manera que en el dels homes, una vegada s'han aplicat els filtres prèviament esmentats, es pot veure que segueix una tendència positiva com en l'altre sexe en qüestió, però en aquest cas en comptes de començar al voltant dels 11.000€ de pensió anuals comença al voltant dels 9.000€ per uns 15.000€ de salari cotitzats i va incrementant progressivament fins als 22.000€ de pensió per 30.000€ de salari aproximadament.

Si s'observa amb detall, s'aprecia que el gràfic de les dones té la pendent de la recta més pronunciada verticalment que la dels homes, ja que, la tendència comença al punt on hi ha una pensió inferior per a les dones respecte el mateix salari cotitzat tant per homes com per dones.

Si es comparen les xifres del descriptiu de la taxa de substitució dels homes amb els de les dones, es pot apreciar que amb les mateixes condicions imposades per als dos gèneres es tenen uns resultats molt dispars. On primer de tot, la mitjana de la taxa de substitució per als homes és del 0,76 i la de les dones és del 0,70. Un 25% de la població masculina disposa d'una taxa que va des de 0,25 fins a 0,67, mentrestant la de les dones va de 0,17 fins a 0,58. De la mateixa manera passa amb l'interval superior, el 25% dels homes tenen unes taxes de substitució que fluctuen entre el 0,83 i l'1,90 i en canvi, per a les dones les xifres són de 0,77 i 1,30.

En el següent model es pot observar la relació entre la pensió i el salari. S'han tingut en compte les següents condicions que corresponen a les utilitzades en els gràfics anteriors: individus que hagin tingut un salari anual superior a 15.000€ i inferior a 30.000€ l'any de referència i individus amb una pensió inferior a 30.000€ anuals.
Es pot afirmar, que la variable salari és significativa respecte a la pensió. A més, per cada unitat que augmenta el salari, la pensió augmenta en 0,729 unitats.

```
Call:
lm(formula = BBB$pensio ~ BBB$salari)

Residuals:
     Min       1Q   Median       3Q      Max
-1208412  -204305    -6507   179574  1795987

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.432e+04  5.476e+04   0.261    0.794
BBB$salari  7.295e-01  2.466e-02  29.578   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 371700 on 1394 degrees of freedom
Multiple R-squared:  0.3856,    Adjusted R-squared:  0.3852
F-statistic: 874.9 on 1 and 1394 DF,  p-value: < 2.2e-16
```

Model 1 Relació entre la pensió i el salari del dataframe BBB

En el model que s'observa a continuació es vol observar si aquest conjunt de variables afecten a la taxa de substitució: el sexe, l'edat de l'individu i el nombre de vegades que es canvia de feina. A més de les condicions esmentades en el model anterior.
Es pot afirmar que les tres variables esmentades, si es tenen en compte les tres a la vegada, afecten a la variable resposta, ja que, són significatives.
S'observa que el sexe és negatiu i molt significatiu, aquest fet vol dir que ser dona fa disminuir la taxa de substitució respecte ser home, per tant, elles tenen una taxa significativa més baixa, ja que, si sexe pren valor 1, més baixa serà la taxa de substitució (on home pren

valor 0 i dona 1). El mateix passa amb l'edat, com més elevada sigui més baixa serà la taxa de substitució. Per últim, la variable numccc, mostra que com més vegades s'hagi canviat l'individu de feina més elevada serà la taxa de substitució.

```
Call:
lm(formula = taxa ~ sexe + edat + numccc, data = BBB)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98418 -0.11404 -0.01102  0.07867  1.57480

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.512684   0.189052   2.712  0.00676 **
sexe        -0.101734   0.011080  -9.181  < 2e-16 ***
edat        -0.008346   0.002850  -2.928  0.00345 **
numccc       0.803773   0.034196  23.505  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2172 on 1631 degrees of freedom
Multiple R-squared:  0.3037,    Adjusted R-squared:  0.3024
F-statistic: 237.1 on 3 and 1631 DF,  p-value: < 2.2e-16
```

Model 2 Relació del conjunt de variables Sexe, Edat i numccc a la variable taxa del dataframe BBB

3.3 INDIVIDUS AMB MÉS DE 35 ANYS COTITZATS

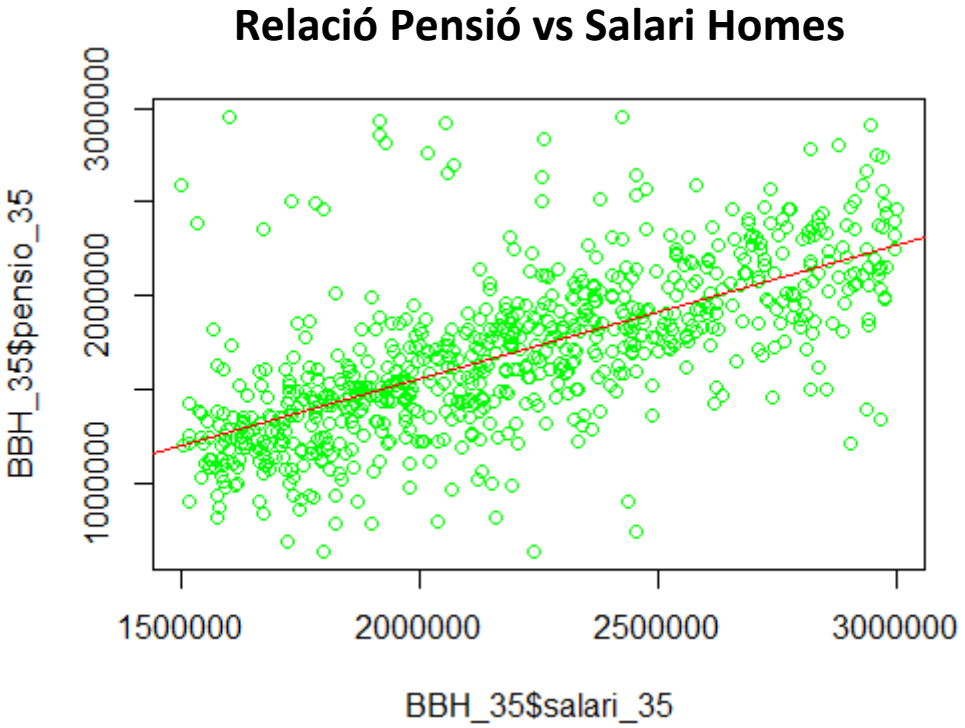
A continuació, es decideix tornar enrere i repetir el procés d'unir totes les taules amb les mateixes condicions exceptuant la del nombre d'anys cotitzats de cada individu. Es decideix que els individus que es volen analitzar hagin cotitzant un mínim de 35 anys. D'aquest amanera es podrà observar si el fet d'haver estat cotitzant aquest període de temps afecta a la taxa de substitució o no.

```
summary(PAPCNET_35$taxa_35)
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
0.0577   0.7393   0.9923   4.2160   3.2090 877.9000
summary(PAPCNET_35$pensio12_35)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
26570 1326000 1945000 2008000 2799000 3746000
summary(PAPCNET_35$cV5)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 3260  699200 1130000 1747000 2709000 4340000
```

Taula 3.28 Descriptius de les variables taxa_35, de pensió12_35 i de cV5 del dataframe PAPCNET

En els anteriors descriptius, es pot observar que no hi ha una gran diferència si es compara amb els seus respectius abans de que s'apliqués la condició de que els individus havien de tenir 35 anys mínim cotitzats.

En el següent gràfic s'observa la relació de l'últim salari i la primera pensió dels homes que han cotitzat 35 anys o més. A més, es tenen en compte les mateixes condicions que prèviament s'han esmentat, on es filtra per individus que hagin tingut un salari anual superior a 15.000€ i inferior a 30.000€ l'any de referència, també és limita la pensió anual, escollint només els individus amb una pensió inferior a 30.000€ anuals i per últim, tenir els individus que no hagin canviat de feina.



Gràfic 3.6 Diagrama de dispersió simple entre la Pensió i el salari dels Homes del dataframe BBH_35

En el primer descriptiu, es pot observar la taxa de substitució per als homes. Aquesta oscil·la entre el 0,29 i l'1,84. Tot i que la mitjana és del 0,78. A més a més, el 50% dels homes tenen una taxa de substitució entre el 0,69 i el 0,84. Es pot apreciar, que respecte la taxa de substitució calculada en l'apartat 3.2 ANÀLISI DE REGRESIÓ, la mitjana passa de valdre 0,76 a 0,78 i el 50% dels homes passa de tenir una taxa de substitució entre 0,67 i 0,83 a tenir 0,69 i 0,84 .

En el segon descriptiu, s'observa la taxa de substitució de les dones. Aquesta oscil·la entre 0,18 i 1,30 amb una mitjana del 0,75. A més el 50% de les dones tenen una taxa entre el 0,69 i el 0,81, aquest interval ha augmentat respecte la taxa de substitució calculada en l'apartat 3.2 ANÀLISI DE REGRESIÓ, on l'interval anava del 0,58 al 0,68. Una altra dada que ha canviat molt és la mitjana, la qual, passa de valdre 0,68 a valdre 0,75.

```
> summary(BBH_35$taxa_35)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.2854  0.6907  0.7718  0.7747  0.8402  1.8440
```

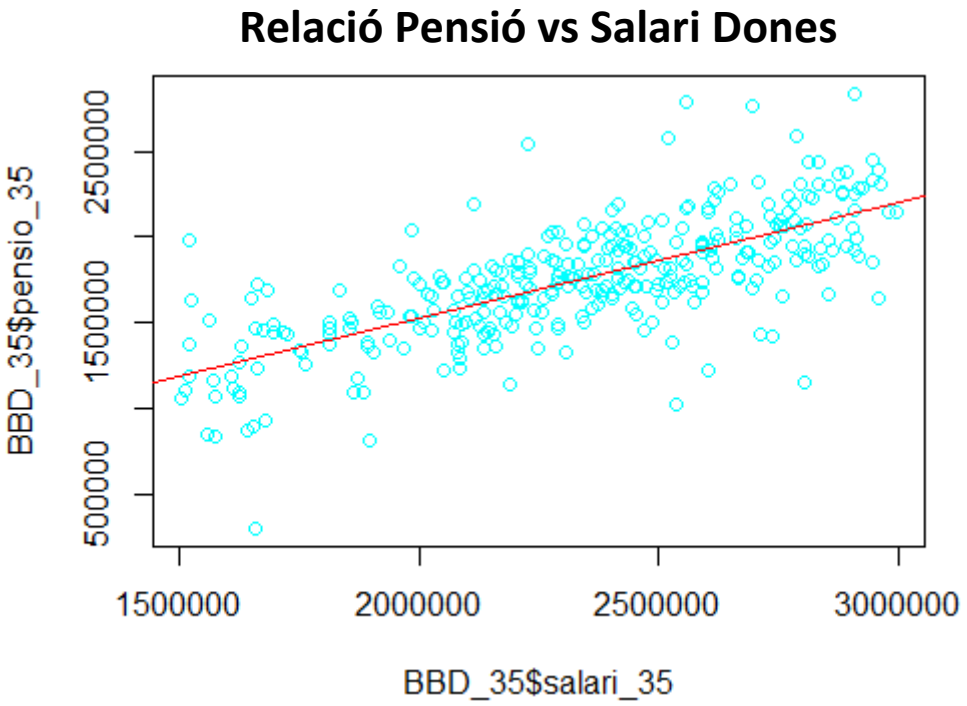
Taula 3.29 Descriptiu de la variable taxa_35 del dataframe BBH_35

```
> summary(BBD_35$taxa_35)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1795  0.6903  0.7523  0.7511  0.8119  1.2990
```

Taula 3.30 Descriptiu de la variable taxa_35 del dataframe BBD_35

Aquest resultats mostren que les dones reben un gran canvi en les seves taxes depenent de si se'ls té en compte els anys cotitzats o no. En canvi, els homes es pot apreciar que els canvis que perceben són mínims.

En el següent diagrama de dispersió, es relaciona la pensió amb el salari de les dones. Es pot veure com la línia de la tendència en comptes de començar per sota dels 10.000 euros de pensió anuals, comença per sobre, això vol dir que es parteix d'un poder adquisitiu més elevat que si no es tenen en compte el nombre d'anys cotitzats.



Gràfic 3.7 Diagrama de dispersió simple entre la Pensió i el salari de les Dones del dataframe BBD_35

En el següent model es pot observar la relació entre la pensió i el salari. S'han tingut en compte les següents condicions que corresponen a les utilitzades en els gràfics anteriors: individus que hagin tingut un salari anual superior a 15.000€ i inferior a 30.000€ l'any de referència, individus amb una pensió inferior a 30.000€ anuals, els quals no hagin canviat mai de feina i individus que hagin cotitzat 35 anys o més.

Es pot afirmar, que la variable salari és significativa respecte a la pensió. A més, per cada unitat que augmenta el salari, la pensió augmenta en 0,642 unitats.

```
Call:
lm(formula = BBB_35$pensio_35 ~ BBB_35$salari_35)

Residuals:
     Min       1Q   Median       3Q      Max
-1123973  -189437   -25450   139640  1647346

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.931e+05  5.897e+04    4.97 7.77e-07 ***
BBB_35$salari_35 6.421e-01  2.604e-02   24.66  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 342600 on 1070 degrees of freedom
Multiple R-squared:  0.3625,    Adjusted R-squared:  0.3619
F-statistic: 608.3 on 1 and 1070 DF,  p-value: < 2.2e-16
```

Model 3 Relació entre la pensió i el salari del dataframe BBB_35

En el model que s’observa a continuació es vol observar si el conjunt de variables que s’esmenten a continuació afecten a la taxa de substitució: el sexe, l’edat de l’individu i el nombre de vegades que es canvia de feina. A més de les condicions esmentades en el model anterior.

Es pot afirmar que les tres variables esmentades, si es tenen en compte les tres a la vegada, afecten a la variable resposta, ja que, són significatives.

S’observa que el sexe és negatiu i no significatiu, aquest fet vol dir que no es rellevant el sexe en quant a tenir una taxa de substitució més alta o més baixa. En canvi, l’edat és significativa, podent afirmar que com més elevada sigui més gran serà la taxa de substitució. Per últim, la variable numccc també significativa, mostra que com més vegades s’hagi canviat l’individu de feina més elevada serà la taxa de substitució.

```
Call:
lm(formula = taxa_35 ~ sexe + edat + numccc, data = BBB_35)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73552 -0.06717 -0.00282  0.05752  1.07984

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.588788   0.176437  -3.337 0.000876 ***
sexe        -0.019178   0.010208  -1.879 0.060552 .
edat         0.011598   0.002661   4.358 1.44e-05 ***
numccc       0.621871   0.035720  17.409  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1539 on 1068 degrees of freedom
Multiple R-squared:  0.2303,    Adjusted R-squared:  0.2282
F-statistic: 106.5 on 3 and 1068 DF,  p-value: < 2.2e-16
```

Model 4 Relació del conjunt de variables Sexe, Edat i numccc a la variable taxa_35 del dataframe BBB_35

3.4 COMPARATIVA MODELS

Si es comparen els Models 1 i 2, on es vol observar la relació del salari amb la pensió, es pot afirmar que quan no es tenen en compte el nombre d'anys cotitzats, la variable pensió augmenta en 0,729 unitats per cada augment d'unitat de la variable salari en comptes d'augmentar en 0,642 unitats quan es tenen en compte individus amb 35 anys cotitzats o més.

Per altra banda, si es comparen els Models 3 i 4, on es vol veure si afecten a la taxa de substitució les següents variables: el sexe, l'edat de l'individu i el nombre de vegades que es canvia de feina, s'observa que el model on no es té en compte el nombre d'anys cotitzats, el sexe és rellevant a l'hora de tenir una taxa de substitució més elevada. En canvi, en el model que es tenen en compte els 35 anys o més cotitzats, el sexe no afecta a la taxa de substitució.

CONCLUSIONS

El volum del conjunt de dades ha implicat més dedicació a l'hora de llegir les dades, ja que aquesta, com ja s'ha esmentat, conté milions d'observacions. Des de l'inici s'ha treballat amb la totalitat de les dades i a mesura que anava avançant el treball s'ha anat escurçant a l'aplicar els filtres i condicions que es desitjaven.

L'anàlisi es pot dividir en dues parts, la primera la més extensa on s'ha anat reduint la base fins a tindre el salari, la pensió, el sexe, la taxa de substitució, entre d'altres; i una segona part on amb les mateixes característiques s'ha exclòs els individus amb un nombre d'anys cotitzats inferiors a 35. Per així poder veure si els anys cotitzats tenen un efecte important a la taxa de substitució.

Una de les principals deduccions és que, el fet de mesurar la taxa de substitució ha estat una tasca molt complicada perquè hi ha molts casos diferents. Per a poder mesurar-la en general, s'ha hagut de prendre la decisió de centrar-se en una població concreta.

A la primera part, s'ha hagut de restringir a un subgrup que són els que estan en aquests intervals (individus amb salari entre 15.000 i 30.000€, individus amb una pensió inferior a 30.000), perquè així s'ha cregut que s'ha eliminat salaris molt baixos, o molt elevats que serien els que distorsionarien l'anàlisi, ja que, aquests són casos que tenen una situació més específica. D'aquesta manera, eliminant aquests casos s'aconsegueix treure un grup de persones amb unes condicions més homogènies. Llavors en aquest nombrós subgrup es veu que la influència en la taxa de substitució és negativa per les dones en contra del que es podria pensar. Per tant, les dones tenen una taxa de substitució significativament inferior als homes.

Després dels càlculs efectuats, es pot afirmar que en determinat tipus de persones si que es pot determinar una relació entre el sexe i la taxa de substitució, aquesta no és gaire gran però si que es pot afirmar que n'hi ha.

En referència a tenir en compte l'últim salari o una mitjana elaborada amb els últims anys de cotització. El més adequat seria tenir en compte la mitjana dels últims anys, ja que, prendre l'últim salari com a referència pot no ser ni una bona estimació ni un bon mètode, ja que aquest, pot haver estat afectat per circumstàncies diferents per a cada individu. S'ha arribat a la conclusió que seria millor tenir en compte més anys, que no només l'últim salari per a fer una millor predicció.

A priori es pot pensar que les dones tenen salaris més baixos i per tant, les seves pensions proporcionalment serien més altes igual que la seva taxa de substitució. Però quan es fa l'anàlisi estadística es veu que és al revés, les taxes de substitució de les dones són més baixes que les dels homes, a continuació s'argumentarà.

Les dones tenen carreres professionals més inestables, això es tradueix en salaris més baixos, però també més irregulars i això té un impacte en la pensió.

La taxa de substitució és un indicador de benestar de la gent gran, perquè si les taxes de substitució fossin del 100%, els individus es jubilarien i seguirien tenint el mateix poder adquisitiu. Ja es coneix que aquest fet és impossible, ja que, seria insostenible. Però si que, s'ha vist que aquesta taxa no és igual per a tothom. Es conclou doncs, que usant la taxa de substitució com a indicador de benestar, s'està creant una desigualtat, ja que, és una mesura que es comporta de forma diferent per a homes i per a dones.

A la segona part, s'agafen les persones que tinguin un mínim de 35 anys cotitzats a més de les condicions de la primera part. Així es pot observar si canvien els resultats al tenir en compte aquesta condició.

A la primera part, es tenen taxes de substitució del 76% per als homes i del 68% per a les dones una diferència molt pronunciada d'un 8% de diferència.

En canvi en la segona part de l'anàlisi, s'obtenen unes taxes del 78% per als homes i 75% per a les dones.

Aquestes dades mostren una clara evidència de diferència de gènere, ja que, quan no es tenen en compte el nombre d'anys cotitzats la desigualtat es abismal, però aquí es podria justificar que hi ha moltes dones amb carreres irregulars, ja que com s'ha dit anteriorment, aquestes tenen situacions laborals més inestables per la maternitat, la cultura d'aquest país que generalment aboca a les dones a deixar la feina si han de cuidar de familiars, etc.

Per altra banda, quan sí que es tenen en compte els anys cotitzats, la taxa de substitució de les dones augmenta fins a arribar a ser gairebé igual a la dels homes. Aquí és quan es pot afirmar que hi ha una clara diferencia, ja que, tot i tenir les mateixes condicions tant els homes com les dones, elles segueixen tenint una taxa de substitució de dos punts per sota de la dels homes.

A més, aquests resultats mostren que els homes no reben un gran canvi en les seves taxes depenent de si se'ls té en compte els anys cotitzats o no. És més, es pot apreciar que les diferències que s'obtenen són mínimes.

Una limitació d'aquest treball és que s'han usat les bases de cotització, que era la informació disponible a la mostra, com a equivalents al salari percebut. Donat que hi ha un màxim de les bases, alguns salaris poden ser superiors a la base que s'està considerant. No hi ha el mateix problema en la part del mínim, ja que el salari no pot ser inferior a la base de cotització mínima. Això portaria a major diferències entre últim salari i primer pensió i per tant, les taxes de substitució reals, en alguns casos en què el salari és superior a la base de cotització màxima permesa, poden ser encara més baixes que les que s'han estimat aquí.

La principal conclusió del treball és l'existència de fortes diferències en les taxes de substitució mitjanes dels homes i les dones. Aquest és un fet que no només no és reconegut quan es parla de les pensions i es generalitza a un dir que la taxa global és del 80%, sinó que a més a més, s'ha vist al llarg de la realització d'aquest treball que té una dificultat elevada.

BIBLIOGRAFIA

Barr, N. y Diamond, P. (2012*) La reforma necesaria. El futuro de las pensiones. Ed. El hombre del tres. Madrid.*

García Jorrín, J. (2017, 17 novembre). *España es ya el segundo país europeo con mayor generosidad en sus pensiones*. *El Confidencial*, p. 1. Recuperat de https://www.elconfidencial.com/economia/2017-11-17/pensiones-generosas-espana-segundo-puesto-tasa-sustitucion_1479153/

Salarios, ingresos, cohesión social. (2018). Recuperat de http://www.ine.es/ss/Satellite?L=es_ES&c=INESeccion_C&cid=1259925408327&p=1254735110672&pagename=ProductosYServicios%2FPYSLayout

De la Fuente Lavín, M. (2004). *LA TASA DE SUSTITUCIÓN DE LAS PENSIONES.* Recuperat de http://www.ehu.eus/ojs/index.php/Lan_Harremanak/article/viewFile/5099/4953

Rodríguez Gonzalez, S. (2017). *Desigualdad por causa de género en la Seguridad Social: carreras de cotización y prestaciones.* Recuperat de http://www.ehu.eus/ojs/index.php/Lan_Harremanak/article/view/18894/17203

INE. (2016, desembre). INE. *Boletín Mensual de Estadística. Diciembre 2016.* Recuperat de http://www.ine.es/daco/daco42/bme/c17.pdf

INE. (s.d.). [*Principales series desde 1971. Población residente por fecha, sexo y grupo de edad*] [*Conjunto de datos*]. Recuperat de http://www.ine.es/jaxiT3/Datos.htm?t=10258

plyr: Tools for Splitting, Applying and Combining Data. (2016, 8 juny). Recuperat de https://cran.r-project.org/web/packages/plyr/index.html

Wickham, H., François, R., Henry, L., & Müller, K. (2018, 19 maig). *A Grammar of Data Manipulation*. Recuperat de https://cran.r-project.org/web/packages/dplyr/dplyr.pdf

ANNEXOS

```
#LECTURA BBDD
setwd("D:/UB/TFG/MCVL/MCVL2015 CDF")
personal<-read.table(file="MCVL2015PERSONAL_CDF.TXT", sep=";")
head(personal)
personal$edat=2015-floor(personal$V2/100)
table(personal$edat,personal$V3)
bb<-personal
head(bb)

attach(bb)

#### TAULA 1 ####
V3_f<-factor(V3)
levels(V3_f)<-c("NA","Home","Dona")
V3_f
taula1<-table(V3_f)
taula1 #frequència absoluta

taula1b<-round(prop.table(table(V3_f)), 4)*100
taula1b #frequència relativa

#marcs de dades
marco1<-data.frame(taula1)
marco2<-data.frame(taula1b)

#agrego noms a les variables
names(marco1)<-c("Sexe","Frequència absoluta")
names(marco2)<-c("Sexe","Frequència relativa")
marco1
marco2

#uneixo els dos marcs
taulafreq<-merge(marco1,marco2,by="Sexe")
taulafreq

######Com s'observa que hi ha NA's es procedeix a eliminar-los i a repetir
#la Taula1

#Eliminar els 0 de la variable sexe
personal=personal[personal$V3!=0, ]
summary(personal$V3)
personal$V3[personal$V3 == 1] <- 0
personal$V3[personal$V3 == 2] <- 1
head(personal)
bb<-personal
head(bb)
attach(bb)

#### TAULA 1 ####
V3_f<-factor(V3)
```

```
levels(V3_f)<-c("Home","Dona")
V3_f
taula1<-table(V3_f)
taula1 #frequència absoluta

taula1b<-round(prop.table(table(V3_f)), 4)*100
taula1b #frequència relativa

#marcs de dades
marco1<-data.frame(taula1)
marco2<-data.frame(taula1b)

#agrego noms a les variables
names(marco1)<-c("Sexe","Frequència absoluta")
names(marco2)<-c("Sexe","Frequència relativa")
marco1
marco2

#uneixo els dos marcs
taulafreq<-merge(marco1,marco2,by="Sexe")
taulafreq

### TAULA 2 ####
attach(bb)
maj44<-bb$edat[bb$edat>44]
sexe<-V3[-c(which(bb$edat<=44))]
taula2<-table(maj44, sexe)
taula2
class(taula2)
table(maj44)
summary(personal)

################################
### ordenar AFILIA1

setwd("D:\\MCVL")
afilia1b<-read.table("MCVL2015AFILIAD1_CDF.TXT", sep=";", colClasses =
c(rep("integer", 7), rep("NULL", 27)), header=FALSE)

library(plyr)
afilia1b_ord=arrange(afilia1b,V1,desc(V7))

library(dplyr)
afilia1b_ord_net <- afilia1b_ord %>% group_by(V1) %>%
  summarise(novaV7=first(V7), count=n_distinct(V7), novaV2=first(V2))
######Proves
head(afilia1b_ord_net)

### ordenar AFILIA2
afilia2b<-read.table("MCVL2015AFILIAD2_CDF.TXT", sep=";", colClasses =
c(rep("integer", 7), rep("NULL", 27)), header=FALSE)

afilia2b_ord=arrange(afilia2b,V1,desc(V7))
```

```
afilia2b_ord_net <- afilia2b_ord %>% group_by(V1) %>%
  summarise(novaV7=first(V7), count=n_distinct(V7), novaV2=first(V2))

head(afilia2b_ord_net)
#################################

### ordenar AFILIA3
afilia3b<-read.table("MCVL2015AFILIAD3_CDF.TXT", sep=";", colClasses =
c(rep("integer", 7), rep("NULL", 27)), header=FALSE)

afilia3b_ord=arrange(afilia3b,V1,desc(V7))

afilia3b_ord_net <- afilia3b_ord %>% group_by(V1) %>%
  summarise(novaV7=first(V7), count=n_distinct(V7), novaV2=first(V2))

head(afilia3b_ord_net)
#################################

### ordenar AFILIA4

afilia4b<-read.table("MCVL2015AFILIAD4_CDF.TXT", sep=";", colClasses =
c(rep("integer", 7), rep("NULL", 27)), header=FALSE)

afilia4b_ord=arrange(afilia4b,V1,desc(V7))

afilia4b_ord_net <- afilia4b_ord %>% group_by(V1) %>%
  summarise(novaV7=first(V7), count=n_distinct(V7), novaV2=first(V2))

head(afilia4b_ord_net)
#################################
dim(afilia1b_ord_net)
dim(afilia2b_ord_net)

unio<-rbind(afilia1b_ord_net, afilia2b_ord_net)
dim(unio)

unio2<-rbind(unio, afilia3b_ord_net)
unio3<-rbind(unio2, afilia4b_ord_net)
head(unio3)
dim(unio3)

##TAULA 3
######################

(m1 <- merge(personal, unio3, by.x = "V1", by.y = "V1"))
head(m1)

table(m1$edat,m1$novaV2,m1$V3)

###Homes
b<-m1[-c(which(m1$edat<=44)),]
c<-b[-c(which(b$V3!="0")),]
```

```
taula3Homes<-table(c$edat, c$novaV2)
taula3Homes
print(taula3Homes)
#Dones

b<-m1[-c(which(m1$edat<=44)),]
c<-b[-c(which(b$V3!="1")),]

taula3Dones<-table(c$edat, c$novaV2)
taula3Dones
print(taula3Dones)


#### FICHERO PRESTACIONES
pensiones<-read.table(file="MCVL2015PRESTAC_CDF.TXT", sep=";")
head(pensiones)
pensiones$anyjub=substr(pensiones$V11,1,4)
jubilats2015=subset(pensiones, V2==2015 & anyjub==2015 & V4==21 &
(V10==1|V10==75))

#Faig un histograma per veure les pensions anuals dels jubilats del 2015

hist(jubilats2015$V20, main = paste("Import Anual Jubilats 2015"), xlab =
"Pensió cèntims d'euro", ylab = "Freqüència", col = "aliceblue")
summary(jubilats2015$V20)
dim(jubilats2015)
#jubilats que estiguin d'alta (variable 4.021 són els 001,002 i 003)
jubilats2015alta=subset(jubilats2015, (V21==000 | V21==002 | V21==003) )
head(jubilats2015alta)
dim(jubilats2015alta)
(dim(jubilats2015)-dim(jubilats2015alta))

#Per veure si hi ha duplicats paquet sql:
#install.packages(sqldf)
library(sqldf)
#comprobacio
#quantes files té el jubilas2105alta
nrow(jubilats2015alta)
#veure si hi ha algú repetit:
jubilats2015V1<-sqldf("select distinct V1 from jubilats2015alta")
nrow(jubilats2015V1)

#Per veure qui és l'individu que està dues vegades
jubilats2015V1<-sqldf("select v1 from (select V1, count(*) as n from
jubilats2015alta group by V1) where n>1")
jubilats2015V1
#l'individu que està dues vegades té l'identificador 538544

jubilats2015altabo=subset(jubilats2015alta, V1!=538544) #Eliminar les dues
files amb aquest identificador ja que no tinc cap criteri per decidir quin
és el bo.
head(jubilats2015altabo)
```

```
###########NOVES VARIABLES AFEGIDES
jubi2015alta=jubilats2015altabo[,c("V1","V4","V10","V11","V12","V15","V16",
"V20","V21","V22","V36","anyjub")]
head(jubi2015alta)

#Ara partint de jubilats2015altabo, anirem ajuntant amb un left join de
l'identificador i el sexe per saer si són homes o dones
personalmini=personal[,c("V1","V3")]
names(personalmini)=c("pV1", "sexe") #canviar noms
head(personalmini)
library(sqldf)
F1=sqldf("select * from jubi2015alta left join personalmini on
jubi2015alta.V1=personalmini.pV1")
head(F1)

#Compararem les pensions dels homes i les dones fent un histograma per
separat
#homes
hist(subset(F1$V20,F1$sexe==0), main = paste("Import Mensual per als Homes
Jubilats 2015"), xlab = "Pensió cèntims d'euro", ylab = "Freqüència", col =
"aliceblue")
summary(subset(F1$V20,F1$sexe==0))

#dones
hist(subset(F1$V20,F1$sexe==1), main = paste("Import Mensual per a les
Dones Jubilades 2015"), xlab = "Pensió cèntims d'euro", ylab =
"Freqüència", col = "aliceblue")
summary(subset(F1$V20,F1$sexe==1))
#summary(round(subset(F1$V20,F1$sexe==1)),4)*0.01  igual pero AMB DECIMALS

#FALTA AFEGIR LA ULTIMA AFILIACIO per saber lo ultim que han cobrat
#tenim la persona que s'ha jubilat al 2015 i el sexe

head(m1)
PERS_AFI=m1[,c("V1","edat","V2","V3","V5","V7","novaV7","count","novaV2")]
head(PERS_AFI)

library(sqldf)
#Prenc com a base Fi, Li junto PERS_AFI (i tindré Personal, afiliació i
prestació JUNT amb el que m'interessa)
PERS_AFI_PREST=sqldf("select * from F1 left join PERS_AFI on
F1.pV1=PERS_AFI.V1") #ara tenim també novaV7 que correspont a la data en
que s'ha jubilat

head(PERS_AFI_PREST)
summary(PERS_AFI_PREST$V15)
summary(PERS_AFI_PREST$edat)

##################################
#Lectura FITXERS COTIZA
##################################
```

```
coti1<-read.table("MCVL2015COTIZA1_CDF.TXT", sep=";", header=FALSE)
head(coti1)
C1=coti1[,c("V1","V3","V16")]
C1_14=subset(C1, (V3==2014))
dim(C1_14)

coti2<-read.table("MCVL2015COTIZA2_CDF.TXT", sep=";", header=FALSE)
head(coti2)
C2=coti2[,c("V1","V3","V16")]
C2_14=subset(C2, (V3==2014))

coti3<-read.table("MCVL2015COTIZA3_CDF.TXT", sep=";", header=FALSE)
head(coti3)
C3=coti3[,c("V1","V3","V16")]
C3_14=subset(C3, (V3==2014))

coti4<-read.table("MCVL2015COTIZA4_CDF.TXT", sep=";", header=FALSE)
head(coti4)
C4=coti4[,c("V1","V3","V16")]
C4_14=subset(C4, (V3==2014))

coti5<-read.table("MCVL2015COTIZA5_CDF.TXT", sep=";", header=FALSE)
head(coti5)
C5=coti5[,c("V1","V3","V16")]
C5_14=subset(C5, (V3==2014))

coti6<-read.table("MCVL2015COTIZA6_CDF.TXT", sep=";", header=FALSE)
head(coti6)
C6=coti6[,c("V1","V3","V16")]
C6_14=subset(C6, (V3==2014))

coti7<-read.table("MCVL2015COTIZA7_CDF.TXT", sep=";", header=FALSE)
head(coti7)
C7=coti7[,c("V1","V3","V16")]
C7_14=subset(C7, (V3==2014))

coti8<-read.table("MCVL2015COTIZA8_CDF.TXT", sep=";", header=FALSE)
head(coti8)
C8=coti8[,c("V1","V3","V16")]
C8_14=subset(C8, (V3==2014))

coti9<-read.table("MCVL2015COTIZA9_CDF.TXT", sep=";", header=FALSE)
head(coti9)
C9=coti9[,c("V1","V3","V16")]
C9_14=subset(C9, (V3==2014))

coti10<-read.table("MCVL2015COTIZA10_CDF.TXT", sep=";", header=FALSE)
head(coti10)
C10=coti10[,c("V1","V3","V16")]
C10_14=subset(C10, (V3==2014))

coti11<-read.table("MCVL2015COTIZA11_CDFF.TXT", sep=";", header=FALSE)
head(coti11)
```

```
C11=coti11[,c("V1","V3","V16")]
C11_14=subset(C11, (V3==2014))

coti12<-read.table("MCVL2015COTIZA12_CDF.TXT", sep=";", header=FALSE)
head(coti12)
C12=coti12[,c("V1","V3","V16")]
C12_14=subset(C12, (V3==2014))
#LA 13 no la llegeixo perque són els autonoms

dim(coti1)
dim(coti2)

uniocoti1<-rbind(C1_14, C2_14)
dim(uniocoti1)

uniocoti2<-rbind(uniocoti1, C3_14)
uniocoti3<-rbind(uniocoti2, C4_14)
uniocoti4<-rbind(uniocoti3, C5_14)
uniocoti5<-rbind(uniocoti4, C6_14)
uniocoti6<-rbind(uniocoti5, C7_14)
uniocoti7<-rbind(uniocoti6, C8_14)
uniocoti8<-rbind(uniocoti7, C9_14)
uniocoti9<-rbind(uniocoti8, C10_14)
uniocoti10<-rbind(uniocoti9, C11_14)
COTIZA<-rbind(uniocoti10, C12_14)

dim(COTIZA)
names(COTIZA)=c("cV1", "cV3","cV5") #canviar noms
head(COTIZA)

PERS_AFI_PREST=data.frame(PERS_AFI_PREST)
COTIZA=data.frame(COTIZA)

############################
#####ARA UNEIXO PERS_AFI_PREST amb l'ultima COTITZACIÓ
############################

PERS_AFI_PREST_COTIZA=sqldf("select * from PERS_AFI_PREST left join COTIZA
on PERS_AFI_PREST.pV1=COTIZA.cV1")
head(PERS_AFI_PREST_COTIZA)
library(sqldf)
library(dplyr)

PERS_AFI_PREST_COTIZA_TOTAL <- PERS_AFI_PREST_COTIZA %>% group_by(cV1) %>%
  summarise(numccc=n(), basesanualtotal=sum(cV5)) %>% ungroup(cV1)

head(PERS_AFI_PREST_COTIZA_TOTAL)
class(PERS_AFI_PREST_COTIZA_TOTAL)

PERS_AFI_PREST_COTIZA_TOTAL=data.frame(PERS_AFI_PREST_COTIZA_TOTAL)
```

```
P_A_P_C=sqldf("select * from PERS_AFI_PREST_COTIZA_TOTAL left join
PERS_AFI_PREST_COTIZA on
PERS_AFI_PREST_COTIZA_TOTAL.cV1=PERS_AFI_PREST_COTIZA.V1")
head(P_A_P_C)
class(P_A_P_C)
dim(P_A_P_C)

#Detectar duplicats de P_A_P_C

anyDuplicated(P_A_P_C) # Número de fila
P_A_P_C[duplicated(P_A_P_C), ]  # Fila

dim(P_A_P_C)

#Primer elimino la columna duplicada cV1 i em quedo amb la primera cV1
P_A_P_C <- P_A_P_C[, !duplicated(colnames(P_A_P_C))]

#Eliminar duplicats de P_A_P_C
PAPC<-P_A_P_C %>% distinct(cV1, .keep_all = TRUE)
dim(PAPC)

#files que continguin un valor negatiu
DF<-PAPC [rowSums (PAPC <0) == 0,]

#eliminar NAs
PAPCNET<-DF[complete.cases(DF), ] #s'elimina 1 fila només, la unica que
conté NAs
head(PAPCNET)

#Elimino columnes duplicades (identificadors)
PAPCNET$V1 <- NULL
PAPCNET$V1.1 <- NULL
PAPCNET$pV1 <- NULL
PAPCNET$V2 <- NULL
#PAPCNET$V3.1 <- NULL (V3.1 -> 1.003 SEXO)
head(PAPCNET)

#Calcular pensió per 12 mesos

PAPCNET$pensio12 <- PAPCNET$V20 * 12

#Calcular Tasa de sustitució= Pensió/Sou -> Tasa=v20/cV5

PAPCNET$taxa <- PAPCNET$pensio12 / PAPCNET$cV5

summary(PAPCNET$taxa)
summary(PAPCNET$pensio12)
summary(PAPCNET$cV5)

#Diagrama de Dispersió Simple
# vectors a partir del dataframe PAPCNET per obtindre la
#relació entre el salari i la pensió
```

```
Salari <- PAPCNET$cV5
Pensio <- PAPCNET$pensio12
#plot(salari,pensio, panel.first = grid(8,8), pch = 0, cex = 1.2, col =
"blue")
plot( Salari, Pensio, panel.first = lines(lowess(Salari, Pensio), lty =
"dashed"),
      pch = 0, cex = 1.2, col = "turquoise2", main = paste("Diagrama
dispersió simple"))

# MODEL

BBB=subset(PAPCNET, pensio12<3000000 & cV5>1500000 & cV5<3000000)
BBB$salari <- BBB$cV5
BBB$pensio <- BBB$pensio12
head(BBB)
plot(BBB$salari, BBB$pensio, main="Gràfica relació Sou vs Salari",  col =
"blue")
abline(lm(BBB$pensio~BBB$salari), col="red")
summary(lm(BBB$pensio~BBB$salari))
summary(BBB$taxa)


Y<- BBB$taxa
reg<-lm(taxa ~ sexe + edat + numccc,  data=BBB)
summary(reg)


#AQUEST HOMES:
summary(PAPCNET$taxa)
BBH=subset(PAPCNET, sexe==0 & numccc==1 & pensio12<3000000 & cV5>1500000 &
cV5<3000000)
BBH$salari <- BBH$cV5
BBH$pensio <- BBH$pensio12
head(BBH)
plot(BBH$salari, BBH$pensio, main="Gràfica relació Pensió vs Salari Homes",
col = "green")
abline(lm(BBH$pensio~BBH$salari), col="red")
summary(lm(BBH$pensio~BBH$salari))
dim(BBH)
#taxa homes salari entre 15.000 i 30.000
summary(BBH$taxa)

BBBh=subset(PAPCNET, sexe==0)
dim(BBBh)
#taxa homes total
summary(BBBh$taxa)

#AQUEST DONES:
summary(PAPCNET$taxa)
BBD=subset(PAPCNET, sexe==1 & numccc==1 & pensio12<3000000 & cV5>1500000 &
cV5<3000000)
BBD$salari <- BBD$cV5
BBD$pensio <- BBD$pensio12
```

```
head(BBD)
plot(BBD$salari, BBD$pensio, main="Gràfica relació Pensió vs Salari Dones",
col = "cyan")
abline(lm(BBD$pensio~BBD$salari), col="red")
summary(lm(BBD$pensio~BBD$salari))
dim(BBD)
#taxa dones salari entre 15.000 i 30.000
summary(BBD$taxa)

BBBd=subset(PAPCNET, sexe==1)
dim(BBBd)
#taxa dones total
summary(BBBd$taxa)

#Model regresió lineal
Y<- PAPCNET$taxa
reg<-lm(taxa ~ sexe + edat + numccc + V5,  data=PAPCNET)
summary(reg)

Y<- BBB$taxa
reg<-lm(taxa ~ edat + V5,  data=BBB)
summary(reg)

Y<- PAPCNET$taxa
reg<-lm(taxa ~ numccc,  data=PAPCNET)
summary(reg)

cor(PAPCNET$sexe, PAPCNET$edat)
cor.test(PAPCNET$sexe, PAPCNET$edat)

###################PART 2############################
################El mateix però ara amb persones amb 35 anys o més cotitzats

#### PRESTACIONS

jubilats2015_35=subset(pensiones, V2==2015 & anyjub==2015 & V4==21 &
(V10==1|V10==75) & V15>=35)
head(jubilats2015_35)

hist(jubilats2015_35$V20, main = paste("Import Anual Jubilats 2015"), xlab
= "Pensió cèntims d'euro", ylab = "Freqüència", col = "aliceblue")
summary(jubilats2015_35$V20)
dim(jubilats2015_35)

#agafo els jubilats que estiguin d'alta (variable 4.021 són els 001,002 i
003)
jubilats2015alta_35=subset(jubilats2015_35, (V21==000 | V21==002 |
V21==003) )
dim(jubilats2015alta_35)
 (dim(jubilats2015_35)-dim(jubilats2015alta_35))

#install.packages(sqldf)
library(sqldf)
```

```
#comprobacio
#Nombre de files jubilas2105alta
nrow(jubilats2015alta_35)

jubilats2015V1_35<-sqldf("select distinct V1 from jubilats2015alta_35")
nrow(jubilats2015V1_35)

#Per veure qui és l'individu que està dues vegades
jubilats2015V1_35<-sqldf("select v1 from (select V1, count(*) as n from
jubilats2015alta_35 group by V1) where n>1")
jubilats2015V1_35

jubilats2015altabo_35=subset(jubilats2015alta_35, V1!=538544)
head(jubilats2015altabo_35)

###########NOVES VARIABLES AFEGIDES
jubi2015alta_35=jubilats2015altabo_35[,c("V1","V4","V10","V11","V12","V15",
"V16","V20","V21","V22","V36","anyjub")]
head(jubi2015alta_35)

##Llegir taula personal i etiquetes sexe
personalmini_35=personal[,c("V1","V3")]
names(personalmini_35)=c("pV1", "sexe") #canviar noms
head(personalmini_35)
library(sqldf)
F1_35=sqldf("select * from jubi2015alta_35 left join personalmini_35 on
jubi2015alta_35.V1=personalmini_35.pV1")
head(F1_35)

#Compararem les pensions dels homes i les dones fent un histograma per
separat
#homes
hist(subset(F1_35$V20,F1_35$sexe==0), main = paste("Import Mensual per als
Homes Jubilats 2015"), xlab = "Pensió cèntims d'euro", ylab = "Freqüència",
col = "aliceblue")
summary(subset(F1_35$V20,F1_35$sexe==0))

#dones
hist(subset(F1_35$V20,F1_35$sexe==1), main = paste("Import Mensual per als
Dones Jubilades 2015"), xlab = "Pensió cèntims d'euro", ylab =
"Freqüència", col = "aliceblue")
summary(subset(F1_35$V20,F1_35$sexe==1))

(m1_35 <- merge(personal, unio3, by.x = "V1", by.y = "V1"))
head(m1_35)

PERS_AFI_35=m1_35[,c("V1","edat","V2","V3","V5","V7","novaV7","count","nova
V2")]
head(PERS_AFI_35)

library(sqldf)
PERS_AFI_PREST_35=sqldf("select * from F1_35 left join PERS_AFI_35 on
F1_35.pV1=PERS_AFI_35.V1")
```

```
head(PERS_AFI_PREST_35)
summary(PERS_AFI_PREST_35$V15)
summary(PERS_AFI_PREST_35$edat)

#################
COTIZA_35<-COTIZA
PERS_AFI_PREST_35=data.frame(PERS_AFI_PREST_35)
COTIZA_35=data.frame(COTIZA_35)

############################
#####ARA UNEIXO PERS_AFI_PREST amb l'ultima COTITZACIÓ
############################

PERS_AFI_PREST_COTIZA_35=sqldf("select * from PERS_AFI_PREST_35 left join
COTIZA_35 on PERS_AFI_PREST_35.pV1=COTIZA_35.cV1")
head(PERS_AFI_PREST_COTIZA_35)
library(sqldf)
library(dplyr)

PERS_AFI_PREST_COTIZA_TOTAL_35 <- PERS_AFI_PREST_COTIZA_35 %>%
group_by(cV1) %>%
  summarise(numccc=n(), basesanualtotal=sum(cV5)) %>% ungroup(cV1)

head(PERS_AFI_PREST_COTIZA_TOTAL_35)
class(PERS_AFI_PREST_COTIZA_TOTAL_35)

PERS_AFI_PREST_COTIZA_TOTAL_35=data.frame(PERS_AFI_PREST_COTIZA_TOTAL_35)

P_A_P_C_35=sqldf("select * from PERS_AFI_PREST_COTIZA_TOTAL_35 left join
PERS_AFI_PREST_COTIZA_35 on
PERS_AFI_PREST_COTIZA_TOTAL_35.cV1=PERS_AFI_PREST_COTIZA_35.V1")
head(P_A_P_C_35)

class(P_A_P_C_35)
dim(P_A_P_C_35)

#Detectar duplicats de P_A_P_C

anyDuplicated(P_A_P_C_35) # Número de fila
P_A_P_C_35[duplicated(P_A_P_C_35), ]  # Fila

dim(P_A_P_C_35)

#Primer elimino la columna duplicada cV1 i em quedo amb la primera cV1
P_A_P_C_35 <- P_A_P_C_35[, !duplicated(colnames(P_A_P_C_35))]

#Eliminar duplicats de P_A_P_C
PAPC_35<-P_A_P_C_35 %>% distinct(cV1, .keep_all = TRUE)
dim(PAPC_35)

#files que continguin un valor negatiu
DF_35<-PAPC_35 [rowSums (PAPC_35 <0) == 0,]
```

```
#eliminar NAs
PAPCNET_35<-DF_35[complete.cases(DF_35), ] #s'elimina 1 fila només, la
unica que conté NAs
head(PAPCNET_35)

#Elimino columnes duplicades (identificadors)
PAPCNET_35$V1 <- NULL
PAPCNET_35$V1.1 <- NULL
PAPCNET_35$pV1 <- NULL
PAPCNET_35$V2 <- NULL
#PAPCNET$V3.1 <- NULL (V3.1 -> 1.003 SEXO)
head(PAPCNET_35)

#Calcular pensió per 12 mesos

PAPCNET_35$pensio12_35 <- PAPCNET_35$V20 * 12

#Calcular Tasa de sustitució= Pensió/Sou -> Tasa=v20/cV5

PAPCNET_35$taxa_35 <- PAPCNET_35$pensio12_35 / PAPCNET_35$cV5

summary(PAPCNET_35$taxa_35)
summary(PAPCNET_35$pensio12_35)
summary(PAPCNET_35$cV5)


#Diagrama de Dispersió Simple

Salari_35 <- PAPCNET_35$cV5
Pensio_35 <- PAPCNET_35$pensio12_35
#plot(salari,pensio, panel.first = grid(8,8), pch = 0, cex = 1.2, col =
"blue")
plot( Salari_35, Pensio_35, panel.first = lines(lowess(Salari_35,
Pensio_35), lty = "dashed"),
      pch = 0, cex = 1.2, col = "turquoise2", main = paste("Diagrama
dispersió simple"))


### MODEL
BBB_35=subset(PAPCNET_35, pensio12_35<3000000 & cV5>1500000 & cV5<3000000)
BBB_35$salari_35 <- BBB_35$cV5
BBB_35$pensio_35 <- BBB_35$pensio12_35
head(BBB_35)
plot(BBB_35$salari_35, BBB_35$pensio_35, main="Gràfica relació Sou vs
Salari",  col = "blue")
abline(lm(BBB_35$pensio_35~BBB_35$salari_35), col="red")
summary(lm(BBB_35$pensio_35~BBB_35$salari_35))

###Model regresió lineal
Y<- BBB_35$taxa_35
reg<-lm(taxa_35 ~ sexe + edat + numccc,  data=BBB_35)
summary(reg)
```

```
#AQUEST HOMES:
summary(PAPCNET_35$taxa_35)
BBH_35=subset(PAPCNET_35, sexe==0 & numccc==1 & pensio12_35<3000000 &
cV5>1500000 & cV5<3000000)
BBH_35$salari_35 <- BBH_35$cV5
BBH_35$pensio_35 <- BBH_35$pensio12_35
head(BBH_35)
plot(BBH_35$salari_35, BBH_35$pensio_35, main="Gràfica relació Pensió vs
Salari Homes",  col = "green")
abline(lm(BBH_35$pensio_35~BBH_35$salari_35), col="red")
summary(lm(BBH_35$pensio_35~BBH_35$salari_35))
dim(BBH_35)
summary(BBH_35$taxa_35)

BBBh_35=subset(PAPCNET_35, sexe==0)
dim(BBBh_35)
summary(BBBh_35$taxa_35)

#AQUEST DONES:
BBD_35=subset(PAPCNET_35, sexe==1 & numccc==1 & pensio12_35<3000000 &
cV5>1500000 & cV5<3000000)
BBD_35$salari_35 <- BBD_35$cV5
BBD_35$pensio_35 <- BBD_35$pensio12_35
head(BBD_35)
plot(BBD_35$salari_35, BBD_35$pensio_35, main="Gràfica relació Pensió vs
Salari Dones",  col = "cyan")
abline(lm(BBD_35$pensio_35~BBD_35$salari_35), col="red")
summary(lm(BBD_35$pensio_35~BBD_35$salari_35))
dim(BBD_35)
summary(BBD_35$taxa_35)

BBBd_35=subset(PAPCNET_35, sexe==1)
dim(BBBd_35)
summary(BBBd_35$taxa_35)
```