TREBALL DE FI DE GRAU

Grau en Estadística

Títol: Estudi Espai-Temporal de l'ús del Bicing a Barcelona mitjançant una plataforma interactiva de Shiny.

Autor: Arnau Nualart Sanz

Director: Josep Anton Sànchez Espigares

Departament: Departament d'Estadística i Investigació Operativa

Convocatòria: Setembre 2023

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Resum

Aquest treball de recerca representa una oportunitat per aprofundir en l'anàlisi espai-temporal de l'ús del Bicing a Barcelona, utilitzant metodologies estadístiques avançades i eines com el Shiny del llenguatge de programació R. Mitjançant l'estudi d'aquestes dades, s'aspira a comprendre millor els patrons d'ús de les bicicletes i identificar àrees d'optimització de les estacions en el sistema del Bicing. Aquest enfocament pot contribuir a la millora del transport urbà sostenible i promoure una mobilitat més saludable i ecològica a la ciutat.

Paraules clau: *Shiny*, *Bicing*, leaflet, Entre-Diari, Components Principals, Clustering, PCA, Profiling

Classificació AMS (MSC 2010)

- 62H25 Factor analysis and principal components; correspondence analysis
- 62H30 Classification and discrimination; cluster analysis
- 65D18 Computer graphics, image analysis, and computational geometry
- 68T10 Pattern recognition, speech recognition

Agraïments

M'agradaria donar les gràcies al director del meu treball Josep Anton Sànchez Espigares, per donar-me suport amb la idea d'aquesta feina, proporcionar-me l'assistència necessària durant tot el procés d'elaboració i ajudar amb qualsevol dubte que pogués sorgir.

TAULA DE CONTINGUT

ÍNDEX DE FIGURES

ÍNDEX DE TAULES

1. INTRODUCCIÓ

La reducció de la contaminació i la sostenibilitat són dues de les coses més buscades en les grans ciutats d'avui en dia. A mesura que les ciutats creixen i es desenvolupen, també ho fan aspectes com ara la contaminació de l'aire, el tràfic congestionat i la falta d'espai verds, entre d'altres. En aquest context, el Bicing va sorgir com una alternativa ecològica i saludable per a la mobilitat urbana que ajudaria principalment als locals a moure's per Barcelona.

Si vius a Barcelona, i et mous per anar a la feina o anar a l'escola/universitat, el servei de Bicing és de les primeres i millors opcions que et venen al cap. Bicing ha estat un exemple de com la innovació en el transport urbà pot tenir un impacte positiu en la qualitat de vida de les persones. A més, a diferència del que passa a moltes ciutats europees els turistes no poden fer servir el Bicing, ja que el lobby d'empreses de lloguer de bicis va pressionar el consistori perquè el sistema vetés els forasters. En qualsevol cas, està pensat per als locals, per a trajectes entre tres i set quilòmetres.

Figura 1.1: Estació de Bicing

El projecte de bicicletes públiques de Barcelona va ser llançat el 2007, gràcies a la inspiració de ciutats europees que ja tenien un sistema similar i la visió del seu creador, Salvador Rueda. Gràcies a això i a les transformacions dels carrers i espais públics, el Bicing ha estat un èxit a Barcelona, ja que ha fet possible que les persones es puguin moure per la ciutat de manera eficient així com millorar la seva salut amb l'exercici físic.

El sistema de bicicletes públiques ha estat un factor clau en la reducció de la contaminació atmosfèrica i sonora a la ciutat, i ha promogut una cultura de transport més sostenible entre els seus usuaris. L'empresa també ofereix esdeveniments als seus abonats i abonades com la

Ruleta del Bicing en la Festa del Bicing, que consisteix a fer girar la ruleta en un estand on se solen repartir més de 100 premis entre els participants.

Barcelona està ple d'estacions per facilitar al ciutadà l'accés de les bicis, ja que en la majoria de les localitzacions sempre hi ha alguna bicicleta per agafar. Tot i això, pot ser que de tant en tant no n'hi hagi cap degut a les hores puntes d'anar a treballar o altres motius, per això hi ha les reposicions de bicis a les estacions efectuades pels treballadors de Bicing.

Respecte a les seves estadístiques, si busquem les dades accessibles del Bicing a Barcelona, veiem que en el *dataset* estudiat hi ha dades per tot el 2018 i la primera part del 2019.

En el nostre cas decidim estudiar la segona setmana de l'any 2019, per tant la que va del 07/01/2019 fins al 13/01/2019, ja que és una setmana laborable normal en la qual no hi ha factors anòmals pel mig, com dies festius, que puguin distorsionar els resultats. El que busquem escollint una setmana com aquesta és saber quin seria el patró d'ús durant les setmanes laborals per així entendre el comportament dels usuaris amb el Bicing per l'àrea de Barcelona.

En termes de les dades que genera el sistema de Bicing són de quantitat massiva. Hi ha un nombre de 456 estacions de bici a Barcelona i els seus voltants, de les quals 411 són per bicis normals i 45 són per bicis elèctriques. A més, cada 5 minuts aproximadament, hi ha una actualització de les dades de totes les estacions. Tot això durant les 24 hores dels set dies de la setmana, el que acaba generant al voltant d'unes 120.000 observacions diferents només per un dia.

Aquesta quantitat de dades ens ajuda molt a estudiar tot el conjunt d'una manera molt més profunda, i tot i que en aquest cas només estudiem set dies podem fer una anàlisi exhaustiva amb el material que tenim.

A més a més, per ajudar la visualització i percepció geogràfica de les estacions i les seves entrades i sortides, es crea una eina a través de R que ajudarà a l'usuari a consultar i entendre millor les dades. Això s'aconsegueix amb la llibreria *Shiny* de R la qual t'ajuda a desenvolupar una aplicació comprensible i fàcil d'utilitzar per a l'usuari que la necessiti per visualitzar els resultats.

Les dades fetes servir s'han extret del Servei de Dades Obertes de l'Ajuntament de Barcelona (Open Data BCN) amb el títol "*Estacions de Bicing, mecàniques i elèctriques, de la ciutat de Barcelona des de l'agost de 2018 al març de 2019*".

Tot i això, el Bicing no ho és tot per als problemes de mobilitat urbana. Com qualsevol sistema, té els seus punts forts i febles, i és necessari analitzar-los per millorar. En aquest treball, es

busca estudiar l'ús del Bicing a Barcelona en termes espai-temporals, amb l'objectiu de proposar millores respecte el sistema i els algoritmes de col·locació de bicicletes.

Amb aquest estudi, s'espera contribuir a la millora del transport urbà sostenible i la reducció de la contaminació a Barcelona, amb la intenció de promoure una millor qualitat de vida per als usuaris i una mobilitat saludable i sostenible.

Com a objectiu primari, a través d'una anàlisi d'una setmana en concret busquem patrons de comportament de l'ús del Bicing a partir de dues dimensions temporals diferents. Això s'aconseguirà a través d'una eina de visualització que intenti representar geogràficament els patrons trobats des d'un punt de vista estadístic de les diferents estacions.

La primera dimensió seria la diària la qual consisteix en analitzar l'ús entre els diferents dies. Amb aquesta es podrà mirar si existeixen diferències entre diferents dies de la setmana, si hi ha diferències entre el cap de setmana i la resta de dies o si els que estan per la meitat com el divendres i dijous són diferents comparats amb el dilluns o dimarts.

La segona dimensió consisteix en l'entre-diària la qual consisteix a analitzar l'ús dintre dels diferents dies. Això s'aconsegueix separant el dia en quatre parts, matí, migdia, tarda i nit. El que es busca amb això és poder mirar si hi ha diferències comparant els diferents horaris en els quals es fan servir les bicicletes. D'aquesta forma mirar si hi ha hores que destaquen per sobre de les altres i per què.

La meva motivació a l'hora de triar aquest tema ve perquè des que vaig arribar a Barcelona el primer any de carrera, tot i que no he arribat a fer servir el Bicing perquè no m'era imprescindible, el veia com un servei còmode, *'eco-friendly*' i saludable que la ciutat oferia per moure'm. Si a això li afegeixes totes les obres viàries que s'han fet per facilitar la vida als ciclistes es transforma en un servei excel·lent. Tenia moltes ganes de treballar en alguna cosa que ajudés al sistema del Bicing per així ajudar a entendre com funciona i exposar problemes o mancances a millorar.

Per dur a terme tot lo comentat, el treball es divideix en diferents parts:

- Descripció del *dataset*
- Metodologia
- Guia aplicació *Shiny*
- Resultats
- Conclusions

2. DESCRIPCIÓ DEL DATASET

En aquest apartat, es proporciona una anàlisi de les dades recopilades i utilitzades en aquest estudi. Veurem en detall cada variable i el que significa per el conjunt de dades. Per fer això comencem amb una breu descripció del *dataset* i posteriorment es mostra una taula on s'aprecien les variables.

La base de dades utilitzada és "2019_01_Gener_BICING_ESTACIONS.csv" extreta d'una base de dades més gran anomenada "*Estacions de Bicing, mecàniques i elèctriques, de la ciutat de Barcelona des de l'agost de 2018 al març de 2019*" d'on es poden extreure altres dades dels diferents mesos del 2018 o començament del 2019.

Les dades contenen 3.979.843 elements amb una suma de 12 variables en total. De totes aquestes observacions només farem servir les que corresponen a la segona setmana de Gener, les altres correspondrien al mes sencer. Seleccionada la setmana a estudiar, queden 868.002 elements els quals s'analitzaran més endavant.

A continuació es mostra un "*Codebook*" de les dades:

Codebook

Variable	Descripció	Tipus	Rang/Categories
Id	Número identificatiu de l'estació de Bicing	Categòrica	0001 – 0496 (No continu)
Type	Tipus d'estació de la Bicicleta	Categòrica	BIKE / BIKE-ELECTRIC
Latitude	Latitud exacta de l'estació	Numèrica	41,36 – 41,45
Longitudes	Longitud exacta de l'estació	Numèrica	2,112 – 2,221
Altitude	Altitud exacta de l'estació	Numèrica	0001 - 0138
StreetName	Nom del carrer de l'estació en qüestió	Categòrica	Gran Via Corts Catalanes, Passeig Marítim, Av. Meridiana, …Altres
StreetNumber	Número del carrer on està situada l'estació	Categòrica	1, 2, 3, 3B, 4, 5, ..., 126, 760 ...Altres

Nom_numero	Variable creada a partir de la unió de *StreetName* i *StreetNumber*	Categòrica	“Gran Via Corts Catalanes, 760”, “Passeig Marítim, 19” ...Altres
Slots	Espais per bicis lliures en l’estació en qüestió	Numèrica	000 – 039
Bikes	Bicis disponibles per utilitzar en l’estació	Numèrica	000 – 039
NearbyStations	Estacions properes a l’estació en qüestió	Categòrica	"71, 79, 209, 406", "76, 109, 220, 350" ...Altres
Status	Estat de l’estació	Categòrica	OPN / CLS
UpdateTime	Hora de l’actualització de les dades de l’estació	Categòrica	“07/01/2019 00:14:13”, “08/01/2019 12:15:12” ...Altres
Dia	Dia en que es recullen les dades de l’estació creada de *UpdateTime*	Categòrica	“07”, “08”, “09”, “10”, “11”, “12”, “13”
Mes	Mes en que es recullen les dades de l’estació creada de *UpdateTime*	Categòrica	“01”
Any	Any en que es recullen les dades de l’estació creada de *UpdateTime*	Categòrica	“2019”
Hora	Hora en que es recullen les dades de l’estació creada de *UpdateTime*	Categòrica	“00” – “23”
Minuts	Minuts en que es recullen les dades de l’estació creada de *UpdateTime*	Categòrica	“00” – “59”
Timestamp	Conversió de una cadena de caràcters *UpdateTime* a un objecte *POSIXct*.	Numèrica	2019-01-07 00:04:00:00 – 2019-01-13 23:55:00:00

Taula 1.1: Descripció variables

Per treballar d'una forma més còmoda es decideix reduir les variables que es faran servir i per començar es treballa amb les variables *id*, *type*, *slots*, *bikes*, *longitude, latitude*, *status* i *updateTime* que son les que ens interessen més i aporten més informació. A més a més, afegirem unes variables que crearem anomenades *Dia*, *Mes*, *Any*, *Hores*, *Minuts, Timestamp* i *nom_numero* les quals estan mencionades anteriorment en el *Codebook*.

A partir de les variables anteriors, crearem un nou *dataframe* el qual ens servirà per estudiar les dades d'una manera més selectiva i precisa. Aquest s'anomenarà "*cleandata_hores*" i constarà de les variables *id*, *dia*, *hores*, *arribades*, *sortides*, *dif*, *part_dia* i nom_*numero*.

2.1. PREPROCESSAT DE LES DADES

En aquest apartat parlaré de quines han sigut les modificacions o els tractaments que s'han dut a terme per l'anàlisi i la visualització de les dades.

Inicialment la base de dades original contenia 3.979.843 observacions, de les quals 868.001 corresponen a la setmana que nosaltres volem estudiar. Un cop reduïdes les observacions, es procedeix a reduir el nombre de variables a analitzar.

Per començar, es passa de les 12 variables originals a unes variables base. Aquestes *són id, type, slots, bikes, status i updateTime*. A partir d'aquestes 6, crearem diverses variables més, les quals s'han esmentat abans que son *Dia, Mes, Any, Hores, Minuts, Timestamp, nom_numero, longitude i latitude*.

Un cop agrupades aquestes quinze variables, decidim ometre els valors *missings*. Això suposa una reducció d'uns 10.226 elements dels quals no hi havia informació en les dades. Això és degut a les dades faltants que hi ha des del moment en que et descarregues el *dataset*. És una incògnita, ja que si estudies els números identificadors de les diferents estacions, hi ha que no arriben a sortir mai, o si apareixen, tenen zero bicis disponibles i un espai (*slot*) lliure, el qual no té sentit.

En el cas de les estacions que no arriben a sortir mai, podem veure un clar exemple en l'estació amb número identificador 10. Si estudiem aquest identificador amb un *subset* al llarg del *dataframe* veiem que no apareix enlloc. Mentre que si estudiem per exemple l'estació amb número identificatiu 43 amb un *subset,* observem que de vegades no es manifesta però a vegades si amb els valors anòmals esmentats abans de 0 i 1 per bicis i *slots*.

Un cop finalitzada la reducció d'observacions segons els que tenen *NA's* o directament no estan a la base de dades, ens quedem amb unes dades depurades de 857.775 observacions. Aquestes són les que formaran el *dataset* anomenat *cleandata* del que més endavant

s'extraurà un més detallat anomenat *cleandata_hores.* Aquest inclourà la creació de variables d'arribades, sortides i la seva diferència de les estacions de tota la setmana classificades segons identificador, dia i hora.

Per ajudar a entendre millor el flux de bicis de les estacions segons les diferents hores, serà aquí quan es crea una nova variable que dona peu més endavant a estudiar els resultats segons les diferents parts del dia. D'aquesta forma es mira si existeixen diferències significatives respecte a les diferents hores que s'agafa o es deixa la bici per així treure conclusions i plantejar un sistema més òptim per al flux de les bicis. Aquestes diferents parts del dia seran Matí, Migdia, Tarda i Nit amb els horaris respectius de 04:00am – 10:00am, 10:00am – 16:00pm, 16:00pm – 22:00pm i 22:00pm – 04:00am.

L'últim pas de creació d'aquest *dataset* serà afegir la latitud i longitud per així després analitzar els fluxos de les estacions geogràficament i extreure conclusions de les diferents àrees que hi ha segons els perfils d'estacions.

3. METODOLOGIA

En aquest apartat es parlarà de les fórmules i els mètodes que s'han fet servir durant aquest estudi.

3.1. CLUSTERING

Aquesta metodologia és una de les més conegudes pels estadístics ja que és de les més treballades gràcies a la utilitat que aporta.

El procediment del *clustering* és una tècnica d'anàlisi de dades que té com a objectiu agrupar observacions similars en conjunts o clústers. Aquesta eina pertany a l'àmbit de l'aprenentatge no supervisat, ja que no requereix etiquetes o categories predefinides per a les dades. Per durho a terme s'utilitzen mètriques de similitud o dissimilitud per mesurar la proximitat entre les observacions, permetent identificar agrupacions naturals sense coneixement previ de la seva estructura.

La metodologia del *clustering* ofereix una perspectiva única per estudiar conjunts de dades complexos, identificant patrons i característiques compartides entre les observacions. A través d'aquest procés, les dades es distribueixen en clústers que després poden ser utilitzats per prendre decisions, segmentar poblacions, detectar anomalies o comprendre millor el comportament de les dades en diferents contextos.

Dins del que és el *clustering* existeixen diferents maneres d'aconseguir la clusterització de les dades, siguin amb variables numèriques o categòriques. Tot i que a continuació es farà un breu resum de les diverses tècniques que hi ha per el *clustering* tant per dades numèriques com per categòriques, cal recalcar que en el nostre cas s'ha utilitzat el *clustering* jeràrquic que pertany a l'apartat de dades numèriques.

3.1.1. CLUSTERING PER DADES NUMÈRIQUES

En aquest tipus de *clustering*, les dades d'entrada són valors numèrics, i l'objectiu és agrupar les observacions similars en clústers basats en les seves característiques quantitatives. En el nostre cas s'ha utilitzat el mètode de *clustering* jeràrquic però la realitat és que hi ha diversos mètodes per dades numèriques, alguns dels quals són:

- **K-means:** És un mètode popular de *clustering* que busca dividir les dades en "k" clústers, on "k" és un nombre predeterminat. Cada clúster té un centroide que representa el centre del clúster. Les observacions s'assignen als clústers en funció de la distància euclidiana al centroide més proper.

- **K-medoids:** Similar al *K-Means*, però en aquest cas, cada clúster té un punt real de les dades com a *medoid*[1]. Això el fa menys sensible a valors atípics o a dades no numèriques.
- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** Aquest mètode es basa en la densitat de les observacions. En aquest cas els clústers són regions de densitat alta, separades per regions amb baixa densitat. És especialment bo per a dades amb diferents densitats o clústers de formes irregulars.

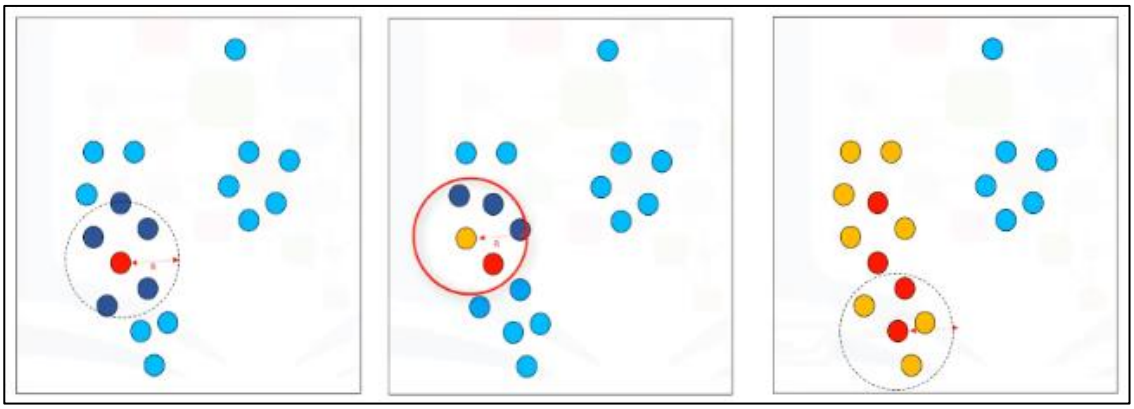

Figura 3.1: Procés de DBSCAN

- **Clustering Jeràrquic:** El *clustering* jeràrquic és una tècnica de *clustering* que organitza les observacions en un arbre jeràrquic de clústers. Aquest arbre té una estructura de branques que permet identificar clústers en diferents nivells de detall. Aquest mètode no requereix fixar prèviament el nombre de clústers, ja que proporciona una representació visual amb un dendrograma que mostra com les observacions es fusionen progressivament en grups més grans.

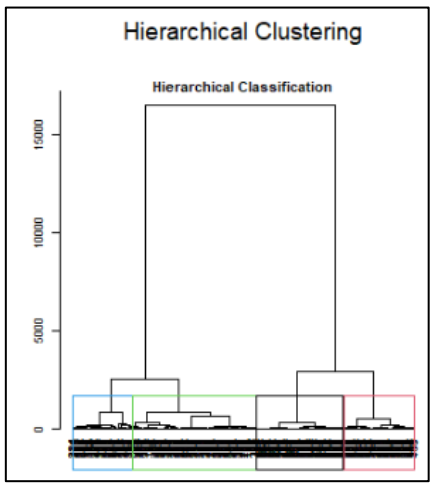

Figura 3.2: Exemple de Dendograma

[1] Medoid: Un medoid pot ser definit com l'objecte d'un grup la dissimilaritat mitjana del qual a tots els objectes en el grup és mínima. És el punt ubicat més cap al centre de tot el grup.

3.1.2. CLUSTERING PER DADES CATEGÒRIQUES

En aquest tipus de *clustering*, les dades d'entrada consisteixen en categories o etiquetes, i l'objectiu és agrupar les observacions que comparteixen les mateixes característiques o propietats. Alguns mètodes de *clustering* per a dades categòriques són:

- **K-Modes:** Aquest mètode és similar al *K-Means*, però s'utilitza per a dades categòriques. Utilitza una mètrica de dissimilitud específica per calcular la distància entre les observacions categòriques.
- **Fuzzy C-Means (FCM):** Aquest és un mètode de *clustering* suau que permet que les observacions pertanyin parcialment a diversos clústers, amb graus d'afinitat en lloc d'assignacions binàries.
- **Clustering Aglomeratiu Jeràrquic (HAC):** Aquest és un mètode de *clustering* que comença tractant cada punt com a clúster individual, seguidament de forma iterativa agrupa els punts més similars fins a formar un dendrograma. La proximitat entre clústers es mesura segons diverses mètriques i al tallar el dendrograma, s'obtenen els clústers finals segons el nivell de semblança desitjada.
- **Latent Class Analysis (LCA):** Aquesta és una tècnica estadística que permet identificar clústers de variables categòriques relacionades, considerant relacions de dependència entre elles.

3.2. PCA (Principal Component Analysis)

El PCA (*Principal Component Analysis*) és una tècnica d'anàlisi de dades que s'utilitza tant en dades numèriques com en dades categòriques per reduir la dimensionalitat i identificar patrons significatius. Aquest mètode transforma les variables originals en noves variables, anomenades components principals, que són combinacions lineals dels atributs originals.

3.2.1. PASSOS PER AL PCA

En aquest apartat, explorarem els passos fonamentals per a dur a terme el PCA, des de la preparació de les dades fins a la selecció dels components principals.

- **Normalització de les dades:** Abans de realitzar el PCA, és important normalitzar les dades numèriques per assegurar que les variables estiguin a la mateixa escala. Aquest

pas és crucial per garantir que les diferents variables contribueixin de manera equitativa al PCA.

- **Càlcul de la matriu de covariància o correlació:** El PCA utilitza la matriu de covariància (o correlació) per avaluar les relacions entre les variables. Aquesta matriu quantifica relació entre totes les parelles de variables i és utilitzada per trobar els vectors propis (*eigenvectors*) i els valors propis (*eigenvalues*) associats.
- **Càlcul dels components principals:** Els components principals són les noves variables que capturen la major variància de les dades originals. Es calculen com una combinació lineal de les variables originals, prioritzant les direccions amb major variabilitat.
- **Selecció dels components principals:** A continuació, s'ha de determinar quants components principals s'han de mantenir. Aquesta decisió pot basar-se en la suma acumulada dels valors propis o en la variància explicada per cada component.
- **Transformació de les dades:** Un cop seleccionats els components principals, les dades són transformades utilitzant aquestes noves variables. Això redueix la dimensió de les dades originals, permetent visualitzar-les i analitzar-les més fàcilment.

Amb l'enfocament en la reducció de dimensions i identificació de components principals, aquest mètode proporciona una manera eficient de representar les dades i identificar patrons importants que condueixen a una millor comprensió i interpretació dels clústers.

4. GUIA APLICACIÓ SHINY

En aquest apartat trobaràs una guia d'utilització de l'aplicació Shiny de Flux Bicing BCN. Aquesta plataforma, com diu el títol, s'ha creat mitjançant RStudio amb les eines interactives que ofereixen les llibreries de 'shiny', 'shinydashboard' i 'shinyjs'.

Aquí trobaràs una descripció de com interactuar amb les seccions, el qual oferirà una visió general sobre les funcions disponibles i com utilitzar-les. Aquesta guia proporciona instruccions pas a pas perquè es pugui visualitzar d'una forma fluida i eficient per a qualsevol tipus de públic.

Com veuràs a continuació, per introduir la plataforma hi ha una secció al principi per explicar el que es veurà i aclarir alguns dubtes que puguin sorgir al llarg de la visualització.

4.1. GUIA APLICACIÓ

Només entrar a la plataforma, et trobaràs amb l'apartat introductori anomenat "Guia Aplicació", el qual correspon a la Figura 4.1. Aquesta secció explica diverses coses importants a tenir en compte durant la visualització. Per començar introdueix les dades, explica breument el que veurà l'espectador i exposa els objectius de la plataforma Shiny.

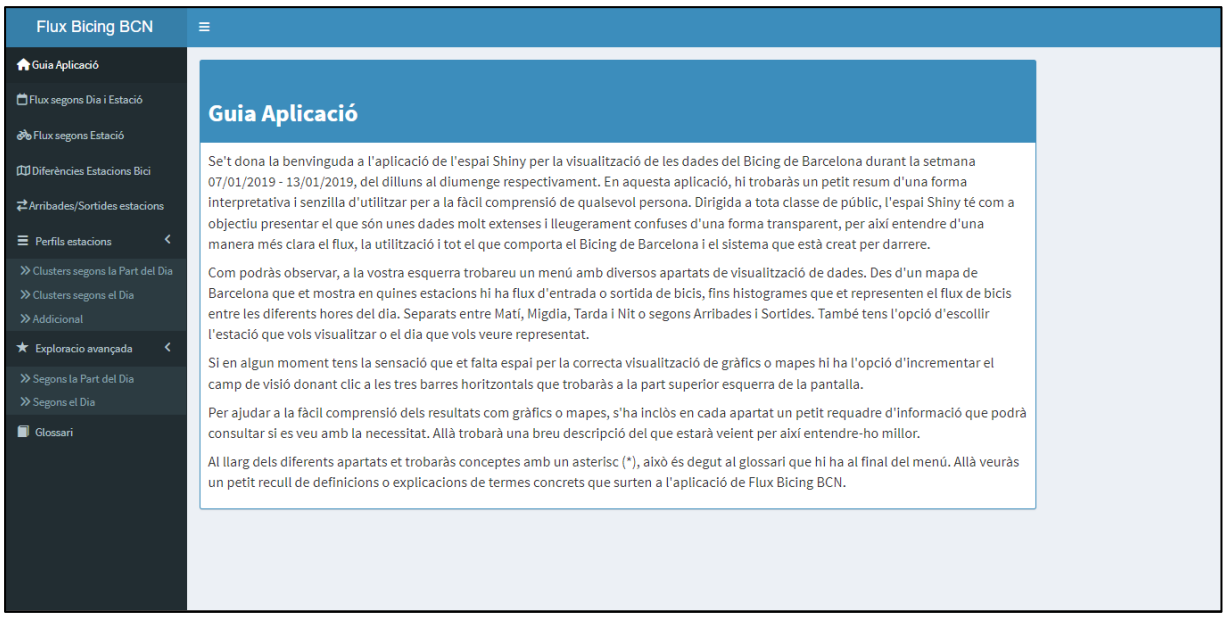

Figura 4.1: Apartat Guia Aplicació

A continuació, dona peu a una explicació del menú que hi ha a la part esquerra i aclareix alguns conceptes tant de les dades com d'algunes seccions interactives. També està inclosa una petita explicació de l'apartat del glossari que es troba al final de l'aplicació i comenta breument els punts d'informació que l'espectador es trobarà a cada apartat. Això facilitarà la comprensió i presentarà l'oportunitat d'explicar i donar llum a conceptes més complexos.

Al llarg de l'aplicació es troben diferents icones tant en el menú, com el glossari o els punts d'informació que ajudaran a una millor orientació molt més eficaç.

Per començar la visualització dels resultats, cliques sobre l'apartat de "Flux segons Dia i Estació", el qual correspon a la Figura 4.2. Aquí trobaràs un apartat senzill i introductori per donar peu a la resta de dades i resultats. Com es veu a continuació, trobes un gràfic de línies senzill en el qual pots escollir el dia i l'estació que desitges.

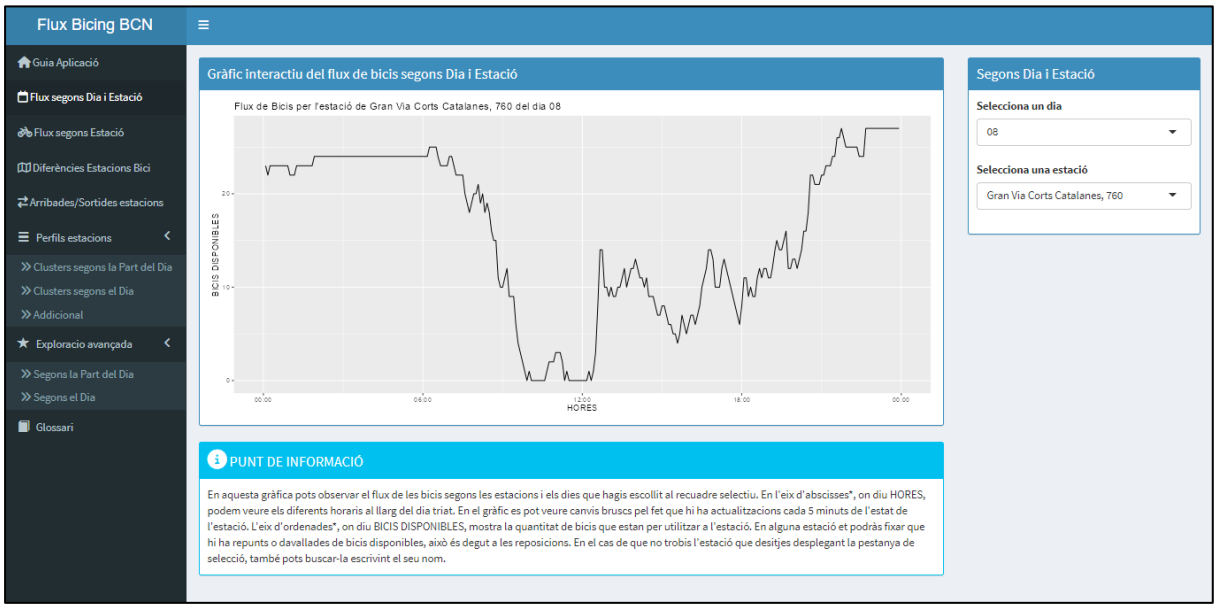

Figura 4.2: Apartat Flux segons Dia i Estació

El requadre petit de la part superior permet escollir el dia i l'estació que vols visualitzar. Com ja diu en el punt d'informació, l'eix d'abscisses representa les hores del dia escollit mentre que l'eix d'ordenades representa les bicis disponibles. Per tant en el cas que hi hagin davallades voldrà dir que en aquell moment estaran agafant les bicis.

El cas de que es vegin molts canvis bruscs en les línies del gràfic, és degut a la constant actualització de l'estat de l'estació. Aproximadament cada cinc minuts hi ha un nou registre en el sistema de qualsevol estació, cosa que fa que el número de registres al llarg del temps sigui immens.

Seguidament a la Figura 4.3 hi ha l'apartat de "Flux segons Estació". Aquest és molt similar a l'anterior degut a que la idea és la mateixa, però en aquest cas l'objectiu és fer la comparació al llarg dels dies. Si et fixes, en la majoria de les estacions hi ha una reducció notable de bicis disponibles durant el cap de setmana (els últims dos dies).

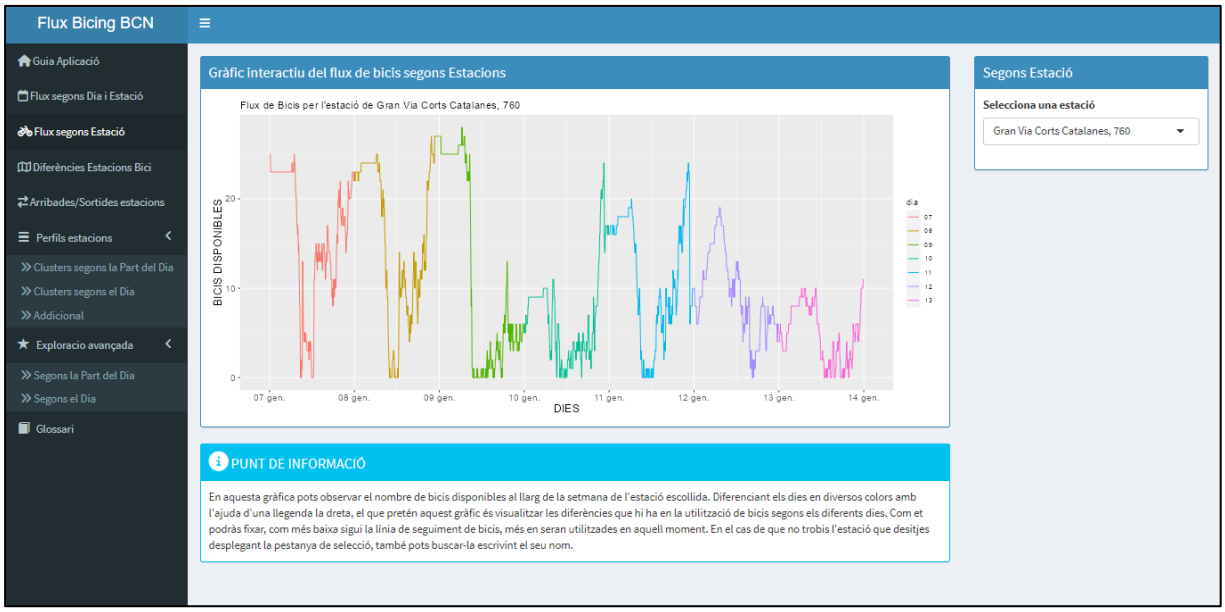

Figura 4.3: Apartat Flux segons Estació

Per assistir a la cerca d'estacions del requadre superior si no troba la que busca, els desplegables de les estacions també contenen la opció d'escriure-la.

A continuació trobes l'apartat de "Diferències Estacions Bici" en la Figura 4.4. Aquí es veu un mapa on mostra la diferència entre arribades i sortides de les estacions que es troben a la ciutat de Barcelona.

Sent cada cercle una estació diferent, si cliques a la que desitges pots veure-hi informació com la seva adreça o la diferència entre les arribades i sortides durant l'hora escollida. Aquests s'escullen al requadre que pots trobar en la part superior a la dreta, on també està la llegenda per donar context als colors del mapa.

Els colors de les estacions ajuden a diferenciar si una estació ha rebut o ha deixat anar més bicis. Una estació verda ha rebut més bicis de les que ha deixat anar mentre que una vermella és el contrari. Una negra ens dona a entendre que no hi ha hagut activitat o que han arribat el mateix nombre de bicis que han marxat. Si et fixes, durant horaris nocturns predominen més els cercle negres, sobretot de 1:00am - 6:00am.

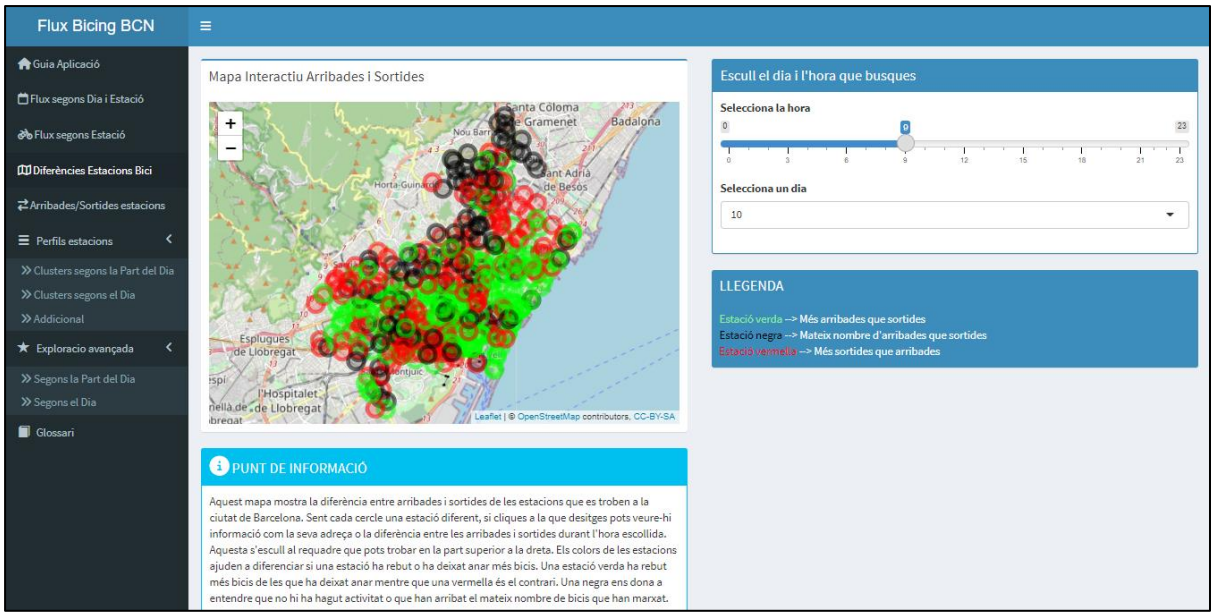

Figura 4.4: Apartat Diferències Estacions Bici

Seguidament, a la Figura 4.5, hi ha l'apartat de "Arribades/Sortides Estacions". Aquí es pot veure un histograma de les arribades o sortides segons la part del dia amb l'opció d'escollir el dia o l'estació que desitgis. Això es fa al requadre de dalt a la dreta, en canvi per canviar entre la visualització d'arribades o sortides has de clicar a la part superior esquerra on es troben els noms amb unes fletxes.

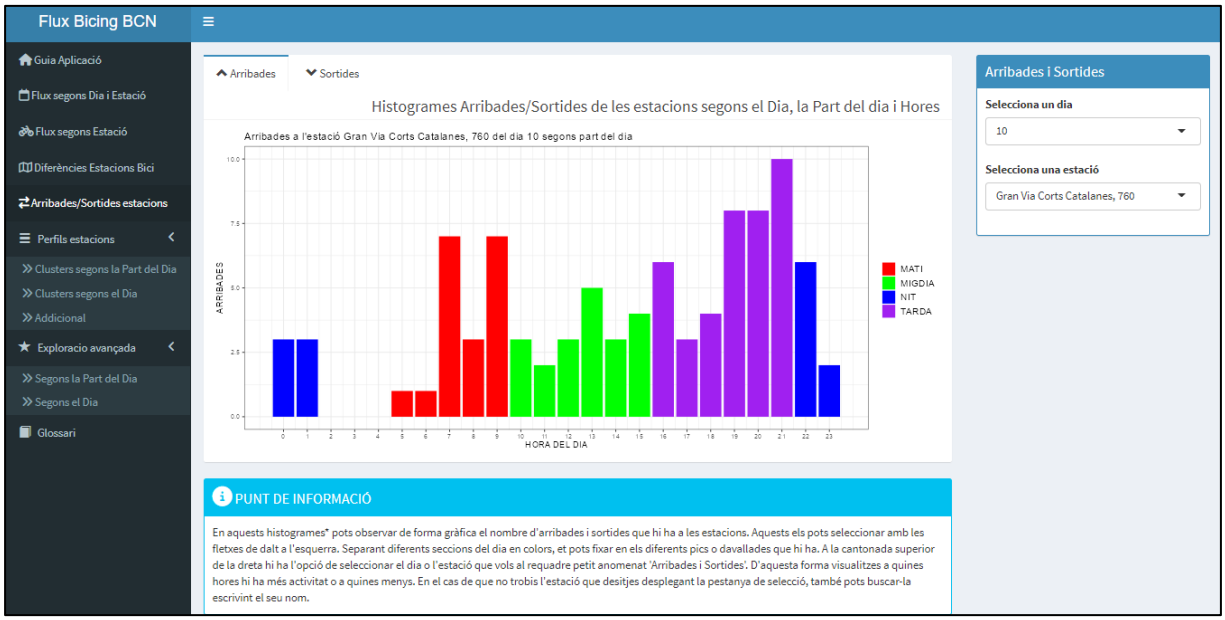

Figura 4.5: Apartat Arribades/Sortides estacions

L'histograma de la Figura 4.5, separa les parts del dia en colors dividint cada grup de forma equitativa. El matí de 4:00am – 9:00am, el migdia de 10:00am – 15:00pm, la tarda de 16:00pm – 21:00pm i la nit de 22:00pm – 3:00am.

Al igual que abans, en el cas de que no trobis l'estació que desitges desplegant la pestanya de selecció, també pots buscar-la escrivint el seu nom.

A continuació entrem a l'apartat del *Clustering* anomenada "Perfils Estacions". Per visualitzar els diferents clústers, s'utilitza les latituds i longituds per localitzar les estacions al mapa de Barcelona. Per començar, com es veu a la Figura 4.6, el subapartat del *Clustering* anomenat "Clústers segons Part del Dia" té en compte les diferents parts del dia. El següent subapartat anomenat "Clústers segons Dia", la Figura 4.7, té en compte l'anàlisi que s'ha fet tenint en compte els dies de la setmana, que van de dilluns a diumenge del 07/01/2019 al 13/01/2019 respectivament.

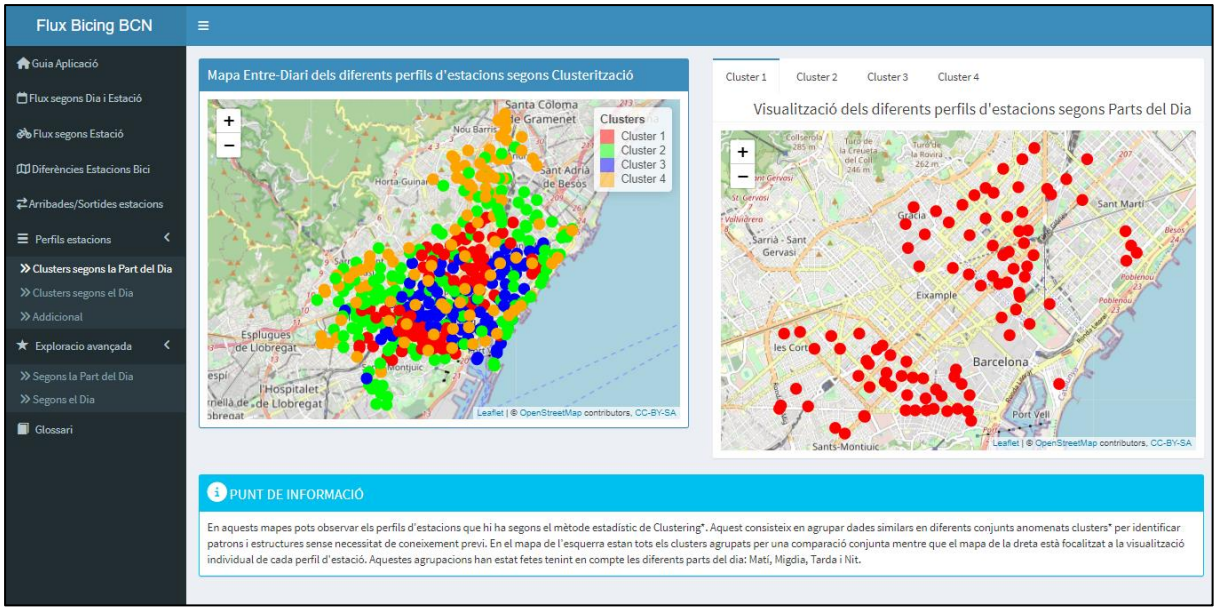

Figura 4.6: Subpartat Clústers segons la Part del Dia

En els dos apartats trobes dos caixes de mapes. En el mapa de l'esquerra estan tots els clústers agrupats per una comparació conjunta mentre que el mapa de la dreta està focalitzat a la visualització individual de cada perfil d'estació. El mapa individual té l'opció a la part superior d'escollir les estacions del clúster que vulguis.

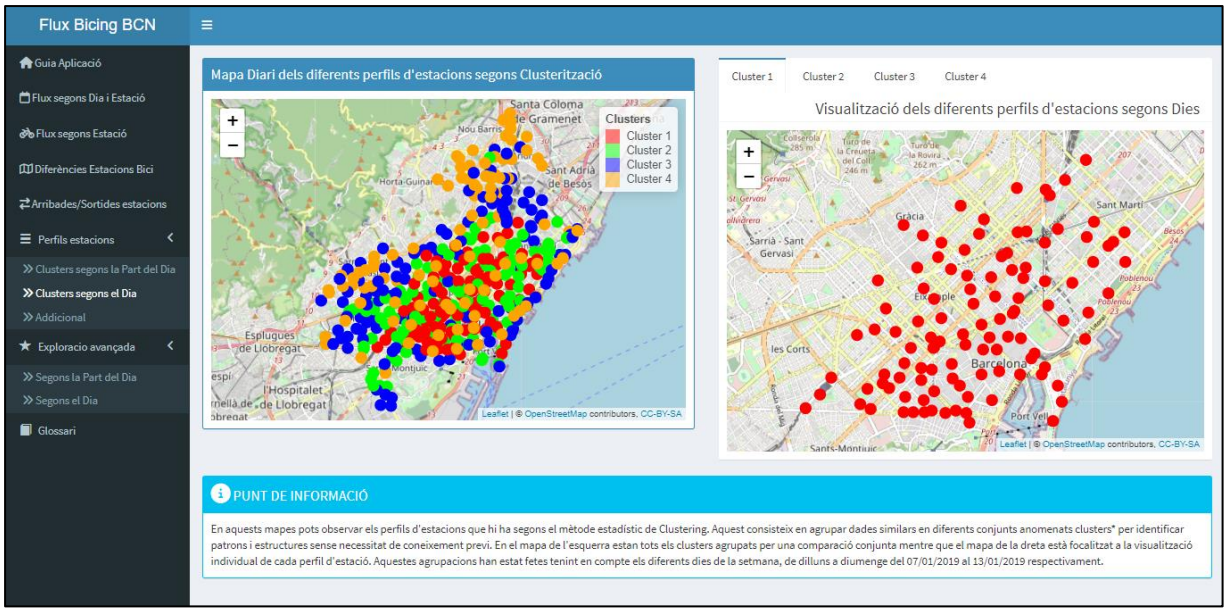

Figura 4.7: Subpartat Clústers segons Dia

Al tercer subapartat de la secció de *Clustering* anomenat "Addicional" que correspon a la Figura 4.8, té com a objectiu mostrar els punts cèntrics de cada clúster de les estacions mitjançant la mitjana.

El mapa de l'esquerra pertany a la tècnica utilitzada per les estacions segons diferents parts del dia mentre que el de la dreta és dels diferents dies de la setmana. Si es clica sobre els punts es veu a quin clúster pertany.

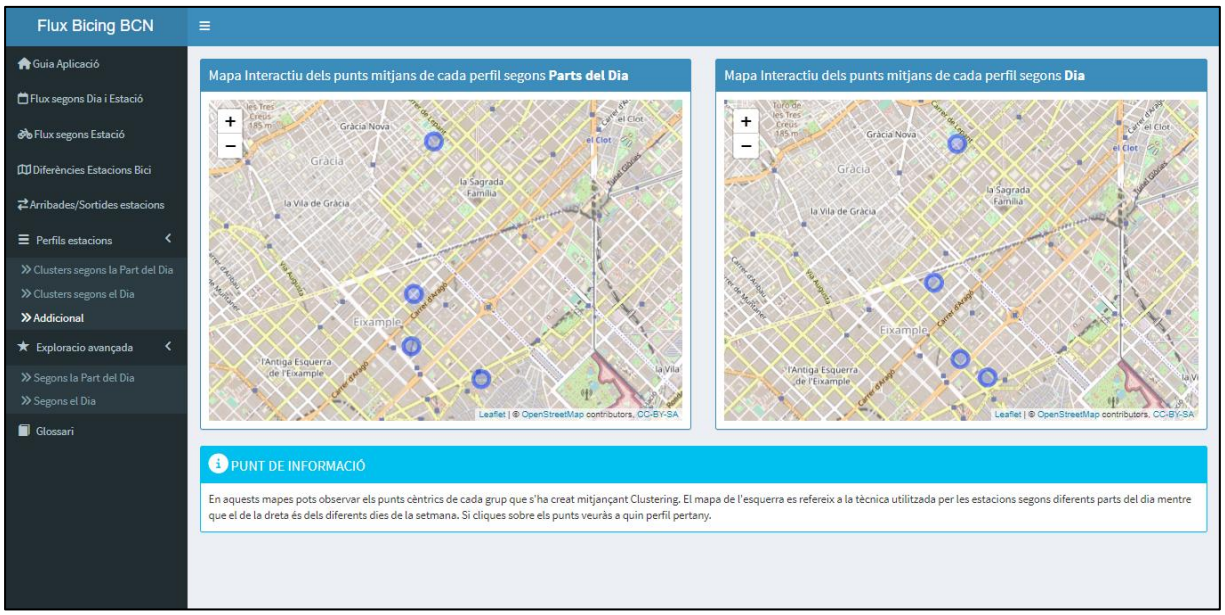

Figura 4.8: Subapartat Addicional

Seguidament, com es veu a la Figura 4.9, hi ha l'apartat anomenat "Exploració avançada". Aquesta secció te l'objectiu de mostrar gràfiques més complexes per una comprensió més profunda sobre els clústers i els diferents perfils que formen les estacions. Al igual que abans,

està separat en dos subapartats. El primer que es veu en la Figura 4.9 anomenada “Segons la Part del Dia” conté un *Scatter Plot* i un *Box Plot* de les dades. Com veurem més endavant, ja que el PCA va concloure que tres components eren suficients per explicar la variabilitat de les dades, hi ha tres *Scatter Plots* graficant les components entre si.

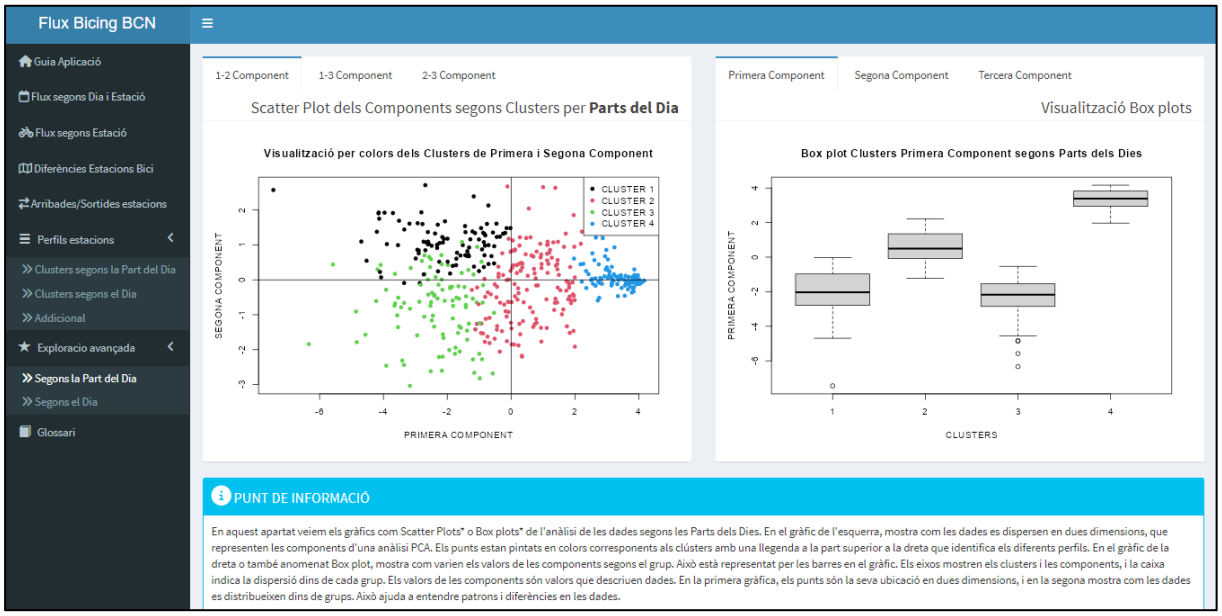

Figura 4.9: Subapartat la Segons Part del Dia

A la part dreta del subapartat es veu una caixa amb un conjunt de *boxplots*. El que es veu en la Figura representa un *boxplot* dels clústers tenint en compte la primera component.

Tot seguit veiem en la Figura 4.10 el següent apartat anomenat “Segons el Dia”. Es veu el mateix que anteriorment, a diferència que en aquest cas només hi ha dues dimensions significatives.

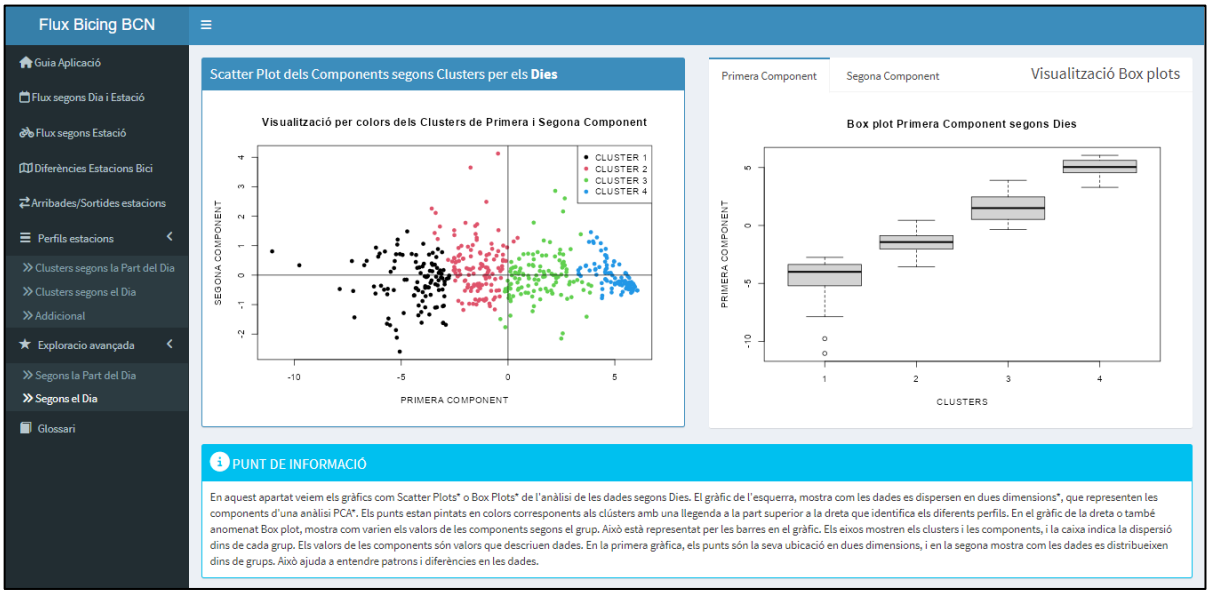

Figura 4.10: Subapartat de Segons el Dia

En termes de diferències clares del gràfic del *Scatter Plot*, es veu que la segona component juga un paper més important en les parts del dia. Mentre que en el gràfic que té en compte els dies de la setmana, la primera component és l'única que es fa notar.

En el gràfic del Box Plot es veuen grups escalats, això és degut a que la primera component representa la mida la qual ens indica que lo que estigui més a l'esquerra amb un valor molt negatiu són les estacions que tenen més activitat.

Finalment, vists tots els apartats, com es veuen a la Figura 4.11 i Figura 4.12, acabem amb un glossari final per enllestir totes les malenteses que poden haver-hi durant els apartats. Aquí es troba una breu explicació juntament amb un llistat de caixes desplegables que contenen els conceptes. Aquesta secció està creada pensant en la fàcil utilització i la interacció de l'usuari amb la plataforma.

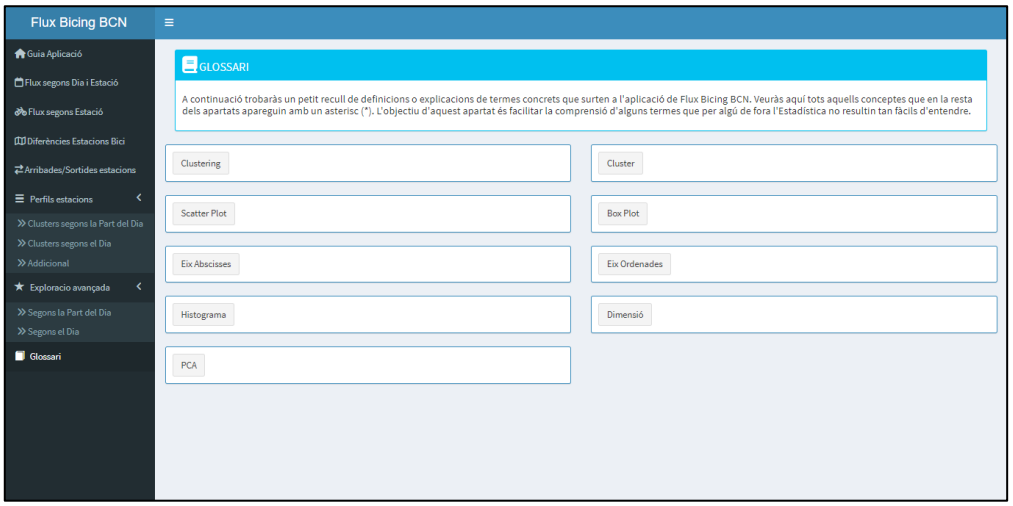

Figura 4.11: Glossari No desplegat

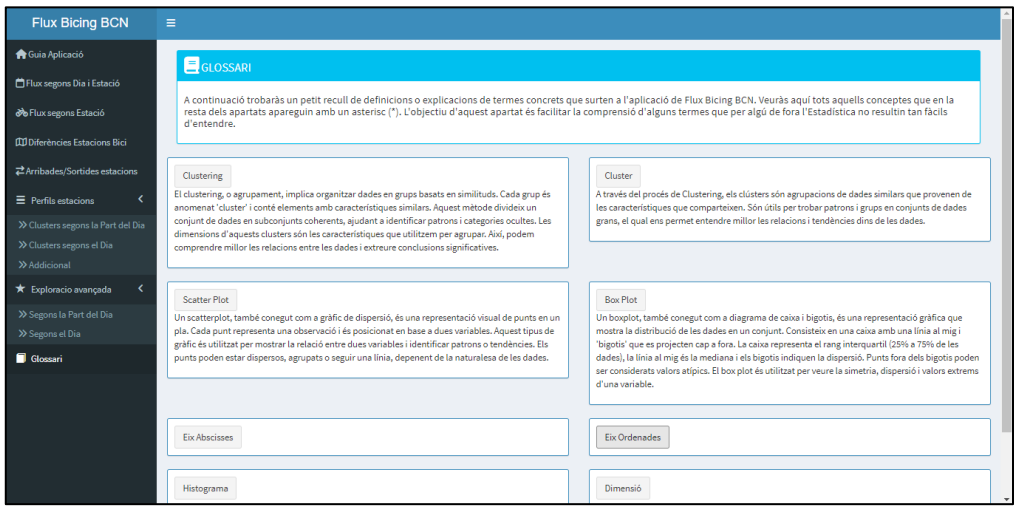

Figura 4.12: Glossari desplegat

En aquest apartat es poden veure termes com *Clustering*, Dimensió, Histograma*, Box Plot*, PCA, *Scatter Plot*, Eix Abscisses, Eix Ordenades o Clúster.

En total la plataforma recull un nombre d'onze seccions informatives o gràfiques d'una manera interactiva perquè l'usuari pugui entendre més enllà com funciona el sistema de Bicing a Barcelona i quin ús fan els ciclistes segons els diferents dies de la setmana o les diferents parts del dia.

5. RESULTATS

En aquest apartat hi ha un recull dels resultats que s'han obtingut, es descompondrà tot el procés que s'ha dut a terme amb les dades, els passos que s'han seguit i a on s'ha arribat després de visualitzar les sortides del nostre codi R.

Com s'ha mencionat anteriorment, en aquest estudi s'ha fet una separació dels resultats per veure si els factors de parts del dia (Entre-Diari) o els dies de la setmana (Diari) tenen un paper clau a jugar en la interpretació de les dades.

Per fer això es fa una exploració del total de dades que tenim. Agafant les entrades i sortides de les estacions reduirem les dimensions fent servir les components principals.

A continuació veurem els diferents passos per cadascun dels diferents factors.

5.1. CREACIÓ DATAFRAMES DEL TREBALL

La creació del principal *dataframe* per treballar les dades ha hagut de seguir diversos passos. Per començar, com s'ha comentat al principi, s'ha fet una reducció de variables de la matriu original.

Un cop es tenen les que es volen, s'han agrupat les dades entre estacions. D'aquesta forma es poden veure les actualitzacions de bicicletes a les estacions d'una forma contínua. En el cas que no s'hagués fet això, es veurien les dades de totes les 456 estacions abans de veure els cinc minuts següents de l'estació desitjada.

Fet això, s'han utilitzat les variables de "bikes" i "slots", les quals es refereixen a les bicis i espais disponibles en les estacions, i s'han calculat les diferències de cada actualització segons els números identificadors. Fetes les diferències de les estacions, s'han agrupat segons si eren positives (arribades) o negatives (sortides) per així unir-les segons les hores que han succeït. A continuació es separen les arribades, les sortides i les diferències segons el dia, l'hora, la part del dia i l'identificador per així crear el *dataframe* definitiu amb el que treballarem.

Per acabar de deixar-lo enllestit, s'inclou el nom i el número de l'estació, els quals serien la variable creada mencionada al principi anomenada "nom_numero". També s'afegeixen les longituds i latituds per després analitzar les dades geogràficament.

En total, després de fer les modificacions necessàries, ens surt un *dataframe* molt més acotat amb només 76.322 observacions. Aquest l'anomenem "cleandata_hores".

Un cop tenim el principal *dataframe* per treballar, s'han de crear dos de nous. Un serà per les dades del nivell Entre-Diari i l'altre per el Diari.

Per fer això simplement creem un nou *dataframe* que contingui una variable identificativa juntament amb les arribades o sortides del període de temps que vulguem.

Pel nivell Entre-Diari, creem dos variables per cada part del dia, una serà les arribades i l'altre les sortides. Les parts del dia són Matí, Migdia, Tarda i Nit. Per tant, sortirà un *dataframe* de 9 variables en total i 456 observacions, una per cada estació.

El mateix farem pel nivell Diari, però en aquest cas seran 15 variables en total, l'identificador i set per les arribades dels dies de la setmana i set per les sortides. Un cop fet això ja tindrem tots els *dataframes* necessaris per treballar i analitzar d'una manera òptima les dades. Els noms d'aquests dos nous *dataframes* són "df_Dia" i "df_partdia".

5.2. NIVELL ENTRE-DIARI

5.2.1. PCA (Principal Component Analysis)

Per començar l'anàlisi del nivell entre-diari s'ha de realitzar un anàlisi de les components principals. Això es fa amb el *dataframe* "df_partdia" el qual conté totes les arribades i sortides de totes les parts del dia. Un cop fem servir la funció *prcomp()* ens retorna el següent:

```
Standard deviations (1, .., p=8):
[1] 2.38587116 1.03257114 0.66298173 0.55321303 0.49394756 0.40570484 0.29057283 0.05303801

Rotation (n x k) = (8 x 8):
                     PC1         PC2         PC3        PC4         PC5        PC6          PC7         PC8
Mati_arribades   -0.2778048 -0.65120433 -0.05922943  0.5407014 -0.08222918 -0.2938065 -0.129199977 -0.3051471
Mati_sortides    -0.3308298  0.41728073 -0.44851888  0.2063284 -0.56532071 -0.2651322 -0.034716513  0.2856572
Migdia_arribades -0.3886399 -0.05895974 -0.24760851 -0.3326879 -0.18169035  0.5703043 -0.402355475 -0.3915907
Migdia_Sortides  -0.3861790 -0.24426335 -0.15755802  0.1165369  0.14952897  0.4526334  0.623534795  0.3698166
Tarda_arribades  -0.3016342  0.53872774  0.29070541  0.5553651  0.36790245  0.1995608 -0.130357213 -0.1835196
Tarda_sortides   -0.3415430 -0.08018301  0.78748689 -0.1610846 -0.44574610 -0.0142442 -0.002605336  0.1787064
Nit_arribades    -0.3903057  0.17349046 -0.02709543 -0.3755581  0.16943208 -0.4388984  0.459885495 -0.4929003
Nit_sortides     -0.3914431 -0.10914691 -0.06171533 -0.2557068  0.50550127 -0.2839003 -0.450468757  0.4760373
```

Figura 5.1: Llista resultats components principals 1

De la funció prcomp() de R obtenim un objecte, força complet, que aglutina els resultats necessaris per a l'estudi de components principals. L'objecte conté una llista *rotation*, que emmagatzema el valor dels *loadings* per a cada component. Analitzar amb detall el vector de *loadings* que forma cada component pot ajudar a interpretar quin tipus d'informació recull cadascuna.

Per exemple, la primera component és completament negativa. Això indica que hi ha una combinació lineal de les variables que resulta en valor negatiu. Si associem aquesta component a la mida, per tant, l'ús que es fa servir de les estacions, ens diu que podria haver sortit positiu o negatiu depenent de l'atzar. Per tant, si ens fixem on hi ha un ús major seria durant la nit i durant el migdia.

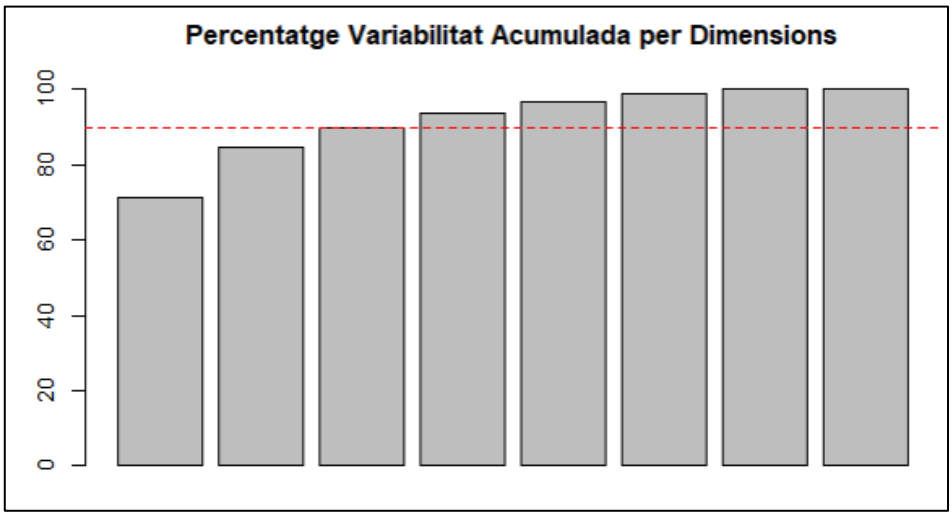

Figura 5.2: Percentatge Variabilitat Acumulada per Dimensions Entre-Diari

Del percentatge de la variabilitat acumulada de la Inèrcia Total, el que busquem nosaltres és una suma del 90%, per tant, tenint en compte aquest valor escollirem tres components per capturar una quantitat prou alta de la variabilitat de les dades.

Seguidament, es calculen i s'emmagatzemen els *eigenvectors*, els *eigenvalues* que formaran posteriorment les components principals. També ho farem amb les projeccions en les 3 dimensions escollides anteriorment.

A continuació hi ha els *Scatter Plots* obtinguts de les components en dues dimensions on es poden distingir lleugerament els números identificatius de les estacions:

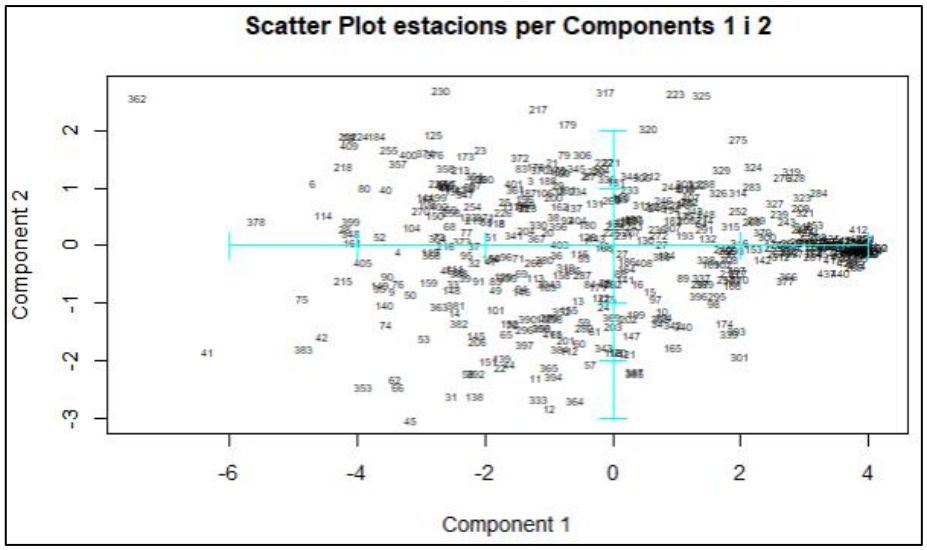

Figura 5.3: Scatter Plot estacions per Components 1 i 2 Entre-Diari

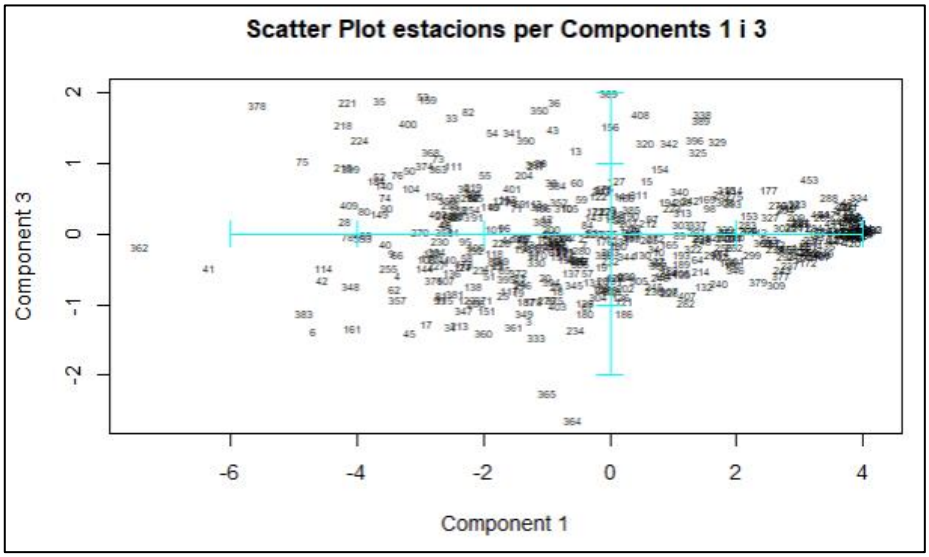

Figura 5.4: Scatter Plot estacions per Components 1 i 3 Entre-Diari

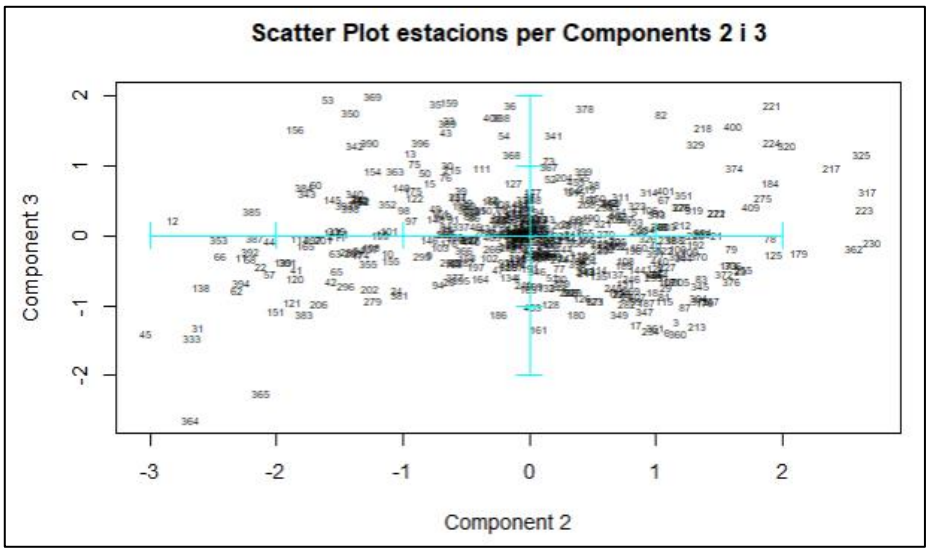

Figura 5.5: Scatter Plot estacions per Components 2 i 3 Entre-Diari

Si en ens fixem en la Figura 5.3 i Figura 5.4 hi ha una acumulació de dades on la component 1 és més positiva i al voltant del 0 de l'eix Y, mentre que a la part negativa es troba molta més dispersió.

Cada punt en els gràfics correspon a una estació, identificada per l'etiqueta. Les coordenades horitzontals i verticals dels punts indiquen els valors de les components respectives. Mitjançant aquests gràfics, és possible observar com les estacions es distribueixen a les noves dimensions. L'ús de diferents parells de components ens ajuda a identificar patrons i agrupacions en funció de les combinacions de components.

Un cop obtingudes les components, s'han procedit a plasmar-les en un pla factorial amb fletxes:

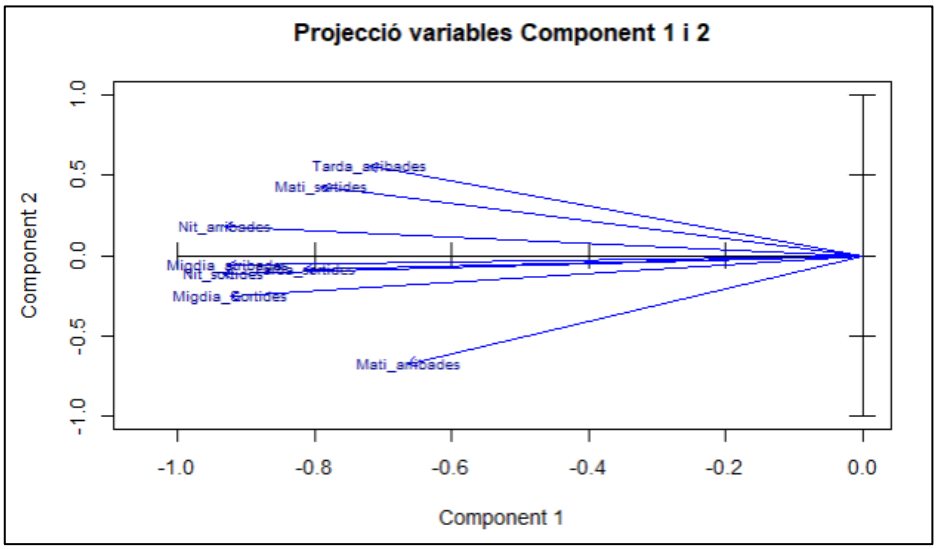

Figura 5.6: Projecció variables Component 1 i 2 Entre-Diari

En la Figura 5.6, totes les variables estan correlacionades negativament amb la primera component, tot i que les que aporten més son "Migdia_arribades", "Migdia_sortides", "Nit_sortides", "Nit_arribades" i "Tarda_sortides". D'aquestes, hi ha que estan relacionades entre elles, "Migdia_arribades", "Migdia_sortides" i "Tarda_sortides" són les més importants per a PC1. Les variables "Tarda_arribades", "Mati_sortides" i "Mati_arribades" són les que estan més correlacionades amb la Segona component, les arribades del matí relacionades negativament i les altres dos de forma positiva.

De la PC2 podem extreure que està relacionada amb els cops que s'utilitzen les bicis per anar a treballar o estudiar ja que les franges horàries encaixen. Si fos així tindria sentit que les sortides del matí i les arribades de la tarda amb una propera correlació positiva estiguin relacionades amb zones residencials mentre que les arribades del matí amb una tendència negativa siguin estacions de zones amb més oficines.

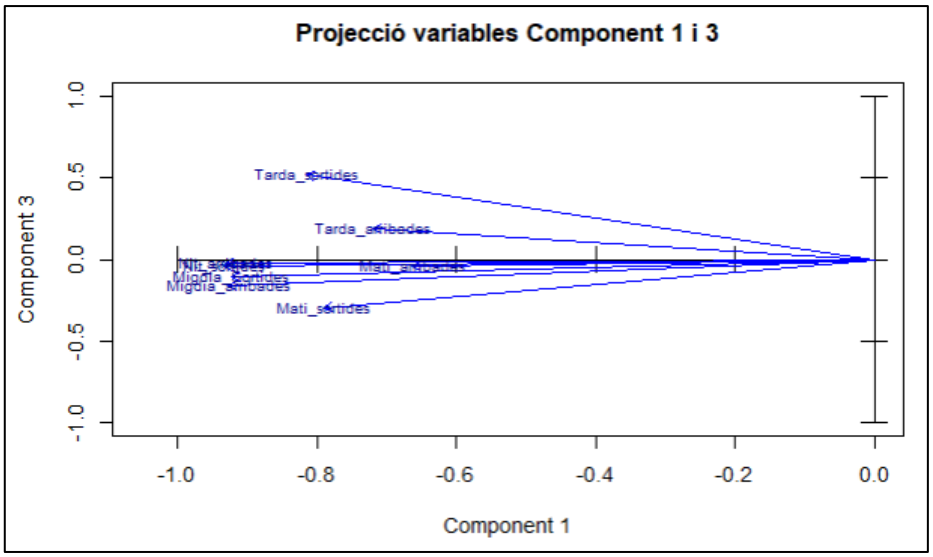

Figura 5.7: Projecció variables Component 1 i 2 Entre-Diari

En la Figura 5.7, totes les variables continuen estant correlacionades negativament amb la primera component, però en aquest cas les que aporten més influencia són les arribades i sortides de la nit i el migdia. Fixant-se en la component 3, "Tarda_arribades" i "Tarda_sortides" estan correlacionades positivament mentre que les sortides del matí tenen una influència negativa. Hi han poques variables que siguin importants tot i que les que aporten més son "Migdia_arribades", "Migdia_sortides", "Nit_sortides", "Nit_arribades" i "Tarda_sortides". D'aquestes, hi ha que estan relacionades entre elles, "Migdia_arribades", "Migdia_sortides" i "Tarda_sortides" són les més importants per a PC1.

Les variables "Tarda_arribades", "Mati_sortides" i "Mati_arribades" son les que estan més correlacionades amb la Segona component, les arribades del matí relacionades negativament i les altres dos de forma positiva.

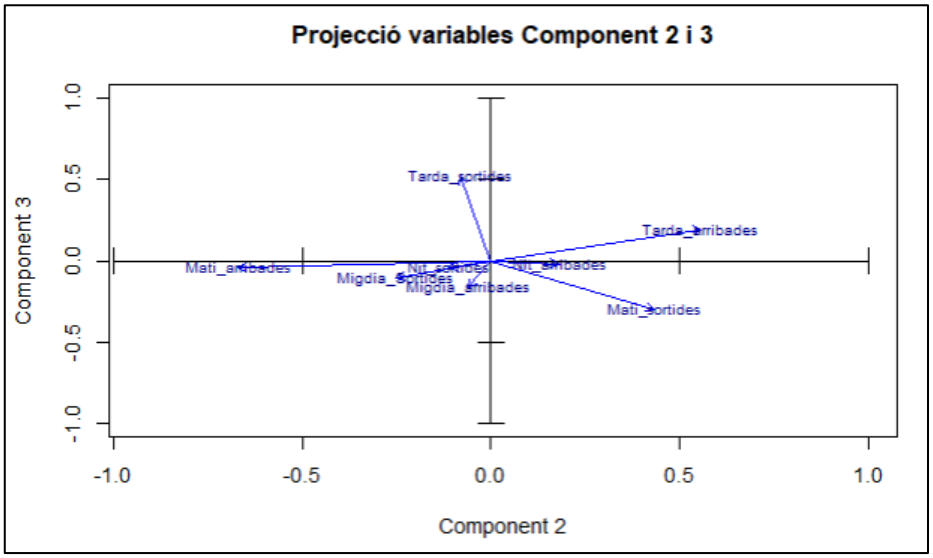

Figura 5.8: Projecció variables Component 2 i 3 Entre-Diari

En la Figura 5.8, com hem vist abans podem distingir que les variables més importants per la component dos són les arribades i sortides del matí juntament amb les arribades de la tarda.

En aquest cas, aquest gràfic ens mostra la importància de les sortides de la tarda de les estacions respecte a la Component 3. Això no es podia diferenciar tan bé en la Figura 5.6 a causa del soroll o la molèstia de les altres variables.

Resumint, es podria dir que la PC1 té a veure amb l'ús que es fa de les estacions, com més negatiu sigui el valor més s'utilitza l'estació. Totes tenen correlació negativa el qual indica que hi ha una combinació lineal de les variables que resulta en valor negatiu. La PC2 està relacionada amb les estacions que s'usen més per anar a la feina o tornar de la feina, i la PC3 amb les estacions que principalment es fan servir per agafar bicis durant la tarda o durant el matí.

A continuació es veu un gràfic més detallat sobre el repartiment dels percentatges segons les dimensions de la separació entre dies de la setmana:

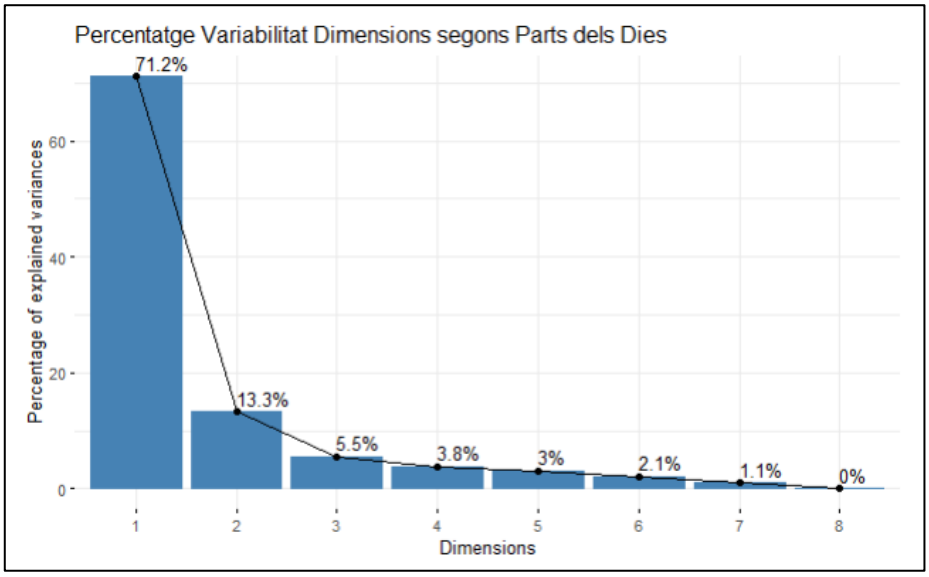

Figura 5.9: Percentatge Variabilitat Dimensions Parts Dies

5.2.2. CLUSTERING

El mètode que he utilitzat per fer els clústers és el mètode de Ward, ja que és un mètode que funciona bé quan la base de dades té poques observacions, en el meu cas tinc només 456 observacions (o estacions), i una de les seves característiques és que minimitza la variància dins de cada clúster.

El dendrograma resultant es veu representat en la Figura 5.10 següent:

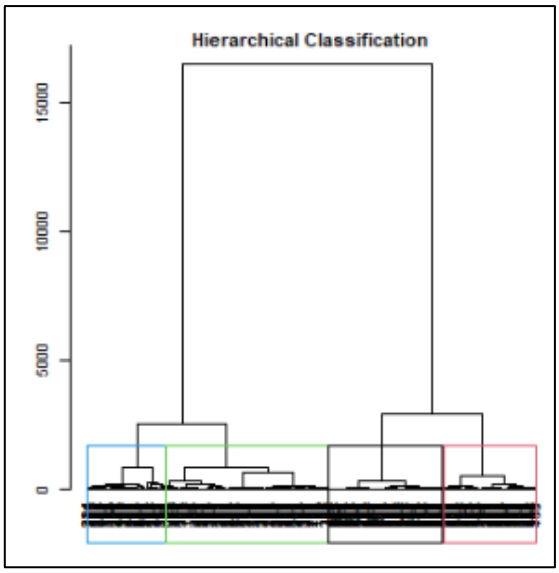

Figura 5.10: Dendograma classificació jeràrquica segons Parts dels dies.

Al dendrograma veiem clarament que el nombre de grups en que es pot dividir la base de dades és 4 i que tots els clústers al contrari del segon son molt similars entre ells. El nombre d'observacions que tindrà cada clúster és:

CLÚSTER	1	2	3	4
Nº observacions	96	157	100	103

Taula 5.1: Observacions segons Clúster Part Dia

5.2.3. PROFILING

A continuació analitzarem en detenció els mateixos *Scatter Plot*s vists anteriorment però aquest cop diferenciant segons els diferents Clústers, primer començarem observant el que plasma en dos dimensions la primera component i la segona component:

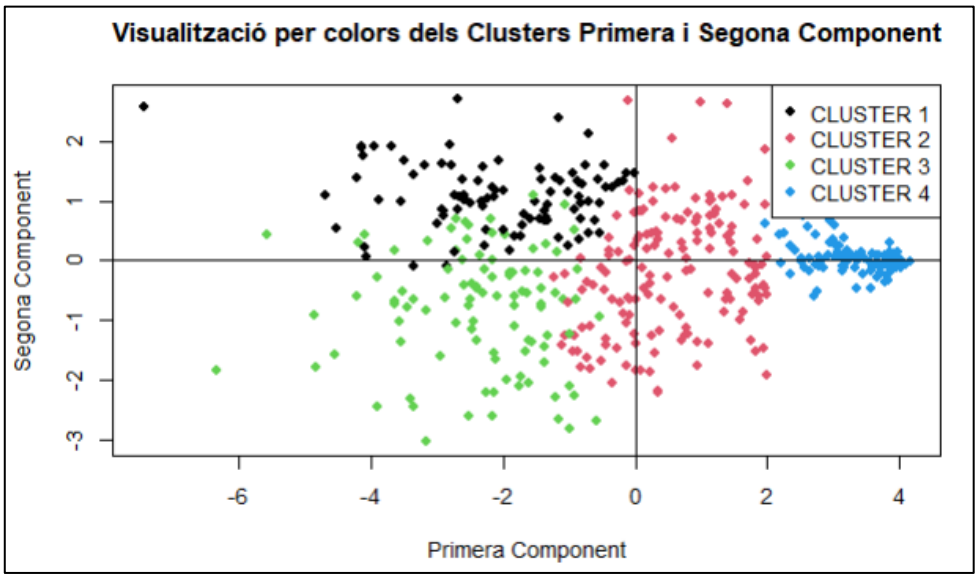

Figura 5.11: Scatter Plot Clústers Component 1-2 Parts Dies

Si mirem la Figura 5.11 i analitzem el gràfic bidimensional de la primera i segona component segons clústers, veiem que el més destacat del gràfic és la diferencia entre el clúster 1 (negre) i el clúster 3 (verd). Aquestes dades són les que estan més disperses si les comparem per exemple amb el clúster 4. Aquestes estan excessivament agrupades i en el sector més positiu de la primera component. Això ens explica que aquestes estacions serien les que s'utilitzen menys. El mateix passa amb la Figura 5.12.

Segons la projecció de variables que hem vist anteriorment, aquesta distribució ens explica que les estacions verdes són les que s'utilitzen més com a destí final d'un viatge durant el matí. Això és sabut ja que si ens fixem en la primera component, les que tenen una tendència més negativa són les estacions que s'utilitzen més. També s'ha de tenir en compte que les estacions que estan per sota de l'eix 0 de la Y son les que estan més relacionades amb les estacions d'arribades durant el matí.

El mateix passa amb les estacions del clúster 1 (negre). Al contrari d'abans però amb la mateix argumentació, les estacions en color negre representen les que s'utilitzen més per a sortides del matí o arribades de la tarda.

Per tant, si fem servir el raonament de la secció final de projecció de variables, les estacions negres de la Figura 5.11 representarien estacions localitzades en zones més residencials mentre que les estacions en color verd estarien en zones amb més oficines.

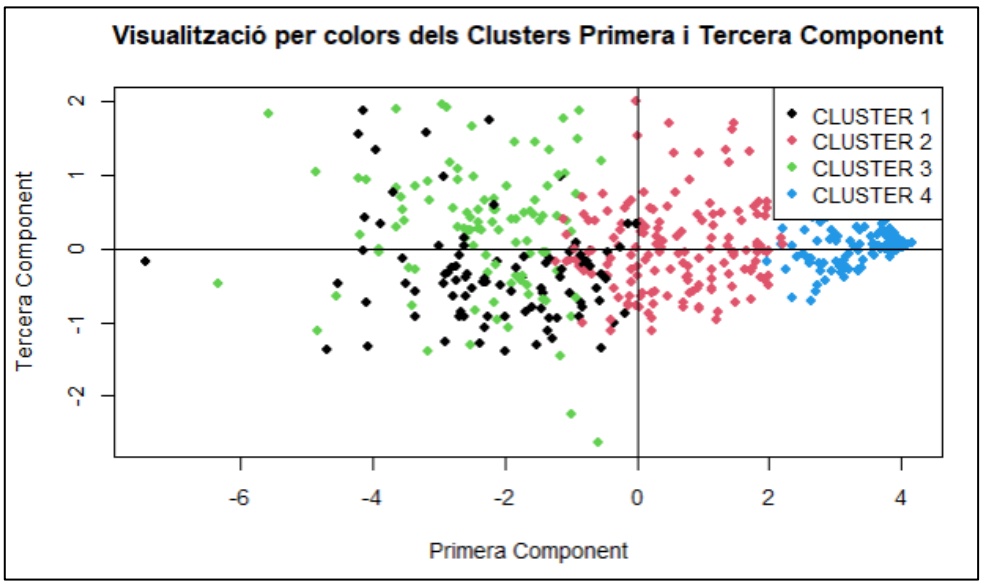

Figura 5.12: Scatter Plot Clústers Components 1-3 Parts Dies

Si ens fixem en la Figura 5.12, la que representa les estacions en un gràfic bidimensional on es diferencien els clústers segons colors, veiem que té una forma similar a la Figura 5.11. Això és degut al pes que té la primera component i la quantitat de variabilitat que aporta.

Al contrari d'abans, ara veiem les estacions agrupades en el clúster 1 i 3 intercanviar-se el paper. Si ens centrem en les estacions del clúster 1 (negres), veiem que està majoritàriament situat al tercer pla cartesià. Com veiem en el Figura 5.8, la projecció de variables ens indica que les estacions amb una tendència negativa de la tercera component són majoritàriament d'on s'agafen més bicis durant el matí. Tenint en compte la primera component, les estacions del clúster 1 segons la Figura 5.12 son les més usades al matí com a punt de sortida.

En canvi, si ens fixem en la part positiva de la Tercera component en la Figura 5.8, podem extreure que les estacions situades en el segon quadrant del pla cartesià en el gràfic bidimensional de la Figura 5.12 són les més utilitzades com a punt de destí durant la tarda. En aquest sector es veuen representades majoritàriament estacions del clúster 3 (verd), tot i que aquest clúster realment està situat al segon i tercer pla cartesià.

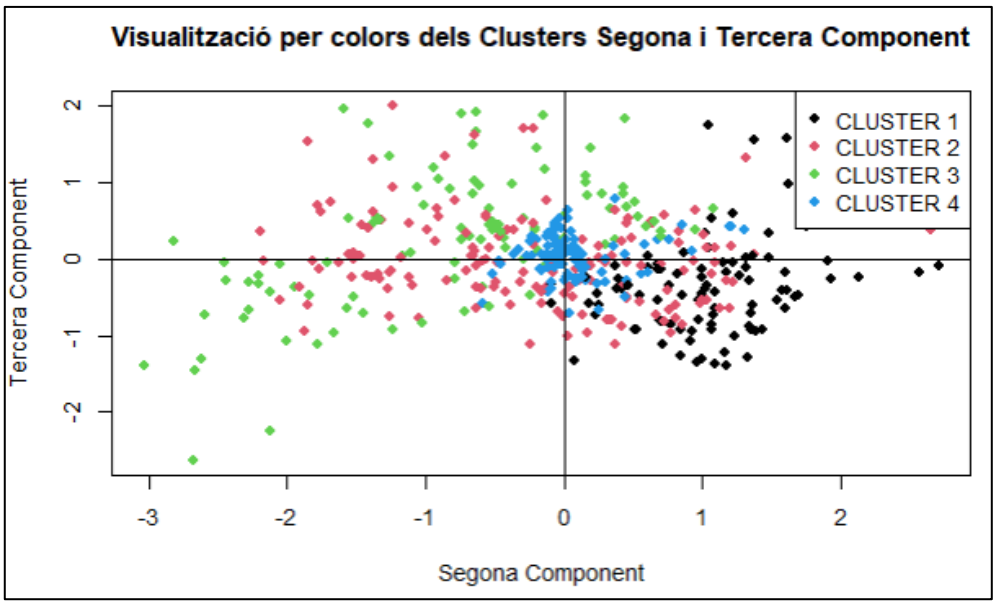

Figura 5.13: Scatter Plot Clústers Component 2-3 Part Dies

De la última Figura 5.13 podem distingir poques coses. El que més ressalta són les estacions en el quart quadrant del pla cartesià que corresponen al clúster 1 i les estacions del clúster 4 agrupades en el centre.

Segons el raonament vist anteriorment amb l'ajuda de la Figura 5.8, el negatiu de la tercera component es refereix a les estacions on hi ha més sortides durant el matí. A més a més, el positiu de la segona component també es refereix a les estacions més utilitzades com a punt de sortida del matí, el qual fa que les estacions negres es situïn en el mateix lloc en la Figura 5.13.

El clúster 4 segueix estan molt agrupat al igual que les Figures anteriors. A part d'això, de la resta de components no podem extreure res més ja que no formen cap altre patró que sigui molt clar.

Per continuar amb la interpretació dels grups creats pel *Clustering*, seguidament veurem uns *boxplots* associats a cada component significativa.

Començant amb la Figura 5.14 que es veu a continuació, veiem dos grups que es diferencien dels altres per tenir una tendència negativa, aquests son el clúster 1 i el clúster 3. Una de les coses que crida l'atenció és lo compactes que estan les dades dins dels diferents grups, sobretot en el clúster 4.

Això dona peu a les observacions anteriors respecte el fet que la primera component es refereix a la mida, o sigui l'ús que es fa de les estacions. Veient aquest Box Plot podem confirmar que les estacions dintre del clúster 1 i el clúster 3 son les més utilitzades mentre que les del clúster 4 son les que menys.

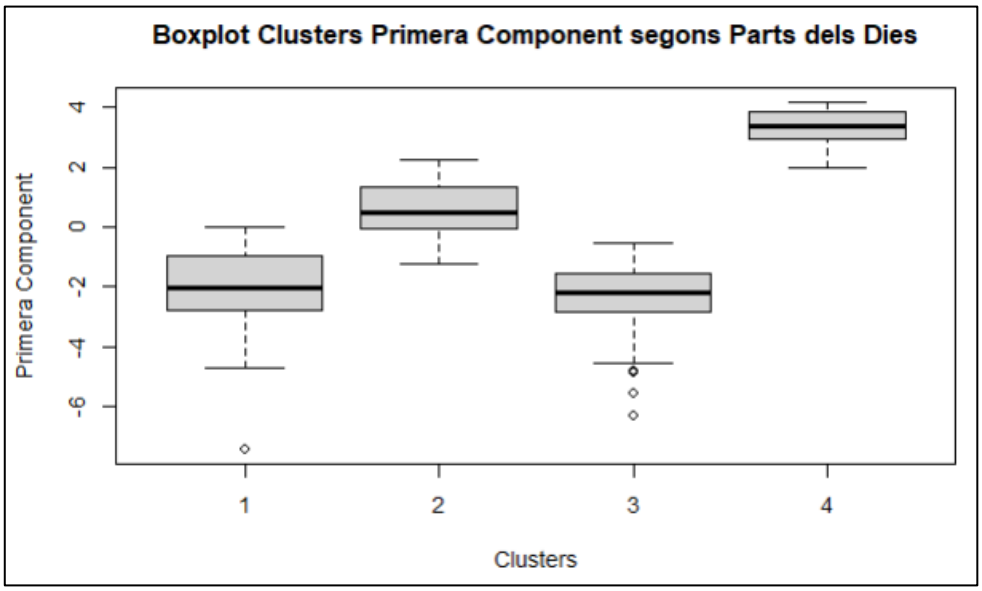

Figura 5.14: Box plot Clúster Parts Dies 1ª Component

Si ens fixem en els clústers restants de la component 2 i component 3 podem extreure poca cosa. Com ens podem fixar en la Figura 5.15, el clúster 1 és el més diferent dels demés amb tendència positiva i el clúster 3 és el que agrupa més estacions amb un valor de la segona component negatiu. On hi ha menys dispersió en la segona i tercera component de forma clara és en el clúster 4. Aquí totes les dades estan molt al voltant del 0.

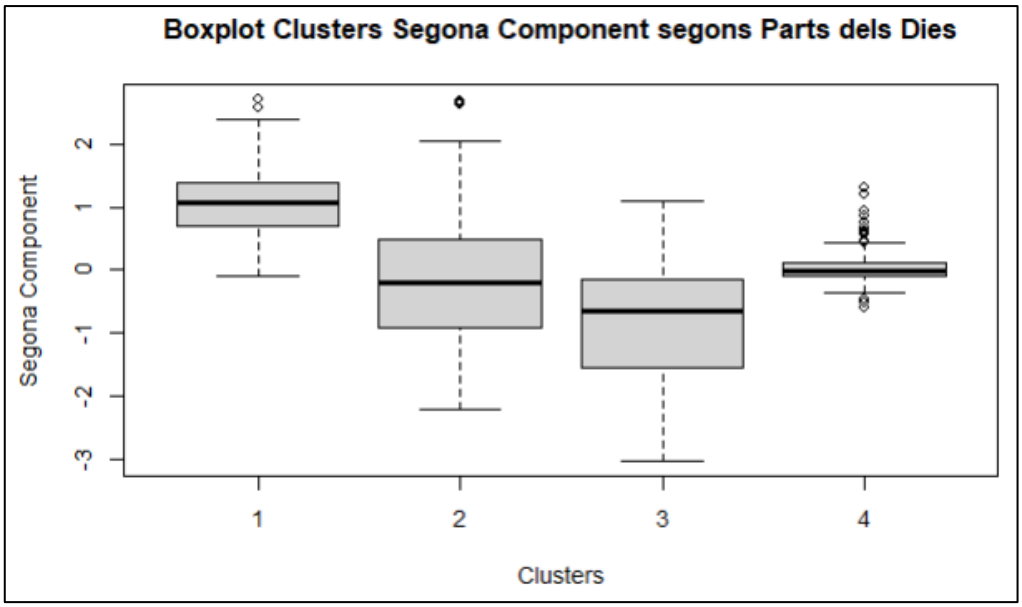

Figura 5.15: Box plot Clúster Parts Dies 2ª Component

Interpretant els resultats del clúster 1 i el clúster 3 de la Figura 5.15, podem dir que el clúster 3 amb una decantació negativa correspon majoritàriament a les estacions amb més arribades al matí mentre que el clúster 1 amb la tendència més positiva correspon a les estacions que s'utilitzen més com a punt de sortida durant el matí o punt d'arribada durant la tarda.

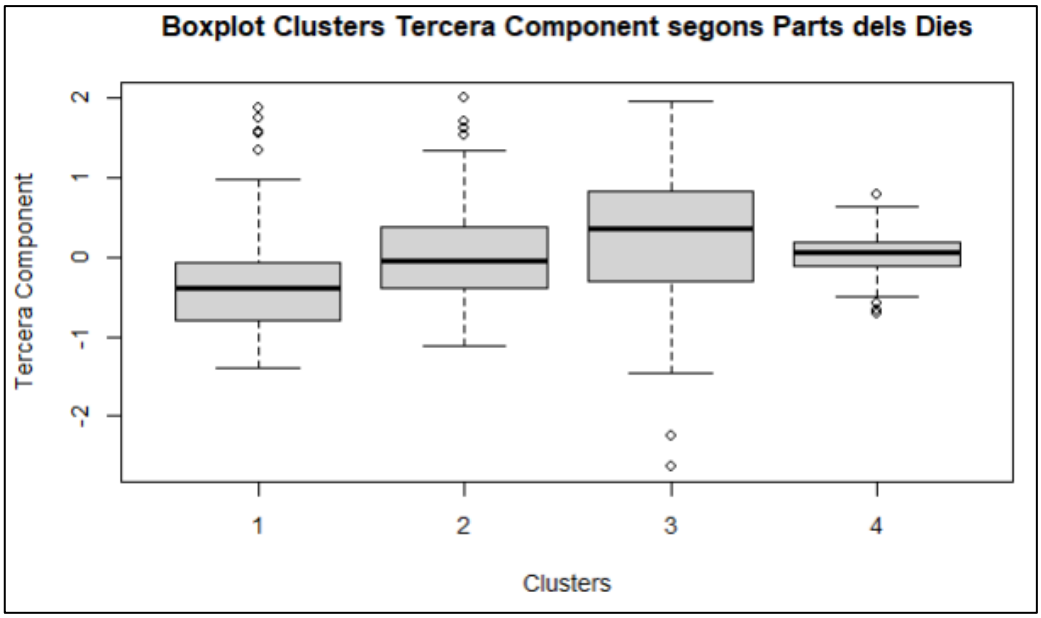

Figura 5.16: Box plot Clústers Parts Dies 3º Component

En la Figura 5.16 poques diferències significatives podem veure. L'únic a destacar és del clúster 3 el qual destaca més de forma positiva, el qual podria ser relacionat amb les estacions que fan de punt de sortida durant la tarda. Els altres clústers són molt propers a 0.

En resum la Figura 5.14 ens representa fàcilment la distribució dels diferents clústers, diferenciant els que tenen estacions amb més ús com el clúster 1 o el clúster 3, mentre que la

Figura 5.15 i Figura 5.16 no aporten gaire informació menys la que dona el clúster 1 i el 3 en la primera i el clúster 3 en la segona, respectivament.

Per finalitzar s'observa una petita taula que mostra les medianes de cada clúster segons les components per tenir present les diferències d'una forma més clara:

	PC1	PC2	PC3
CLUSTER 1	-2.0320090	1.067588918	-0.38978557
CLUSTER 2	0.4943134	-0.197496642	-0.05076775
CLUSTER 3	-2.1729968	-0.655586044	0.36805328
CLUSTER 4	3.3865781	-0.001394995	0.04928122

Taula 5.2: Medianes Clusters segons Components Entre-Diari

Finalment, per acabar amb el *profiling*, veurem la distribució geogràfica dels diferents clústers en un mapa de Barcelona per veure si hi ha diferències significatives. Els mapes que es veuen a continuació s'han creat amb la llibreria "leaflet" de R.

Comencem veient una representació conjunta dels diferents clústers en la Figura 5.17:

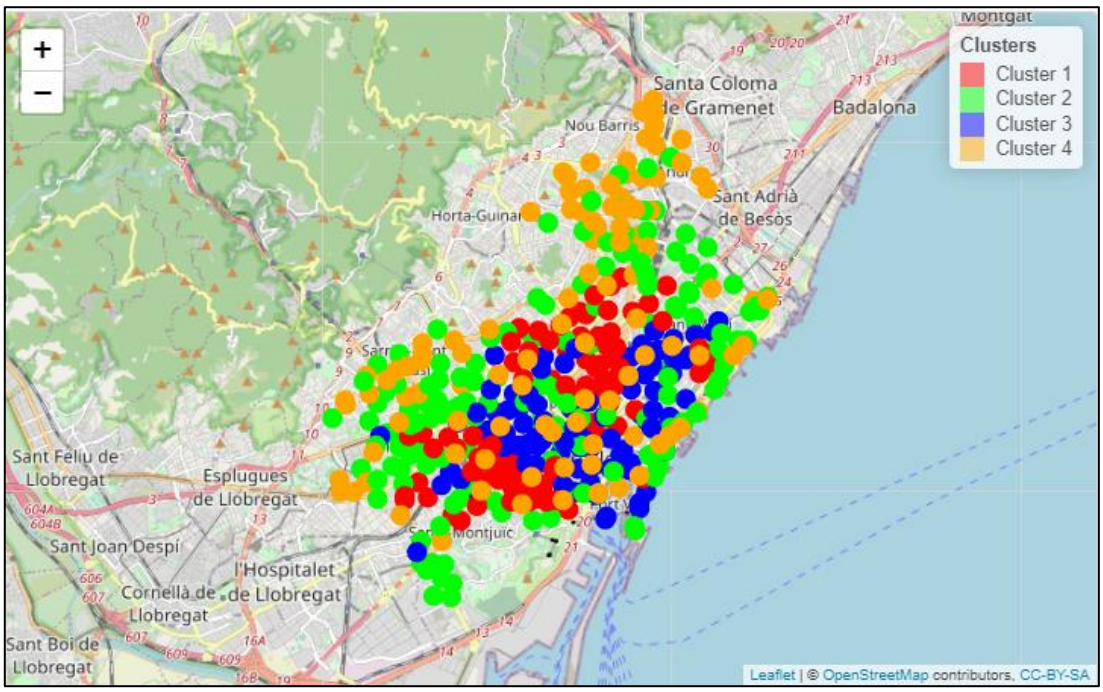

Figura 5.17: Mapa conjunt clusters Entre-Diari

Si ens fixem en els diferents grups de la Figura 5.17, podem veure una diferenciació de la part cèntrica amb la part exterior. El clúster 2 i el 4 (verd i taronja respectivament) envolten els altres d'una manera predominant per la part exterior. Mentre que el clúster 1 i 3 (vermell i blau respectivament) s'agrupen més en la part cèntrica. Si recordem la Figura 5.14, el clúster 1 i 3 representaven les estacions més utilitzades, a més a més amb una diferenciació gràcies a la segona component.

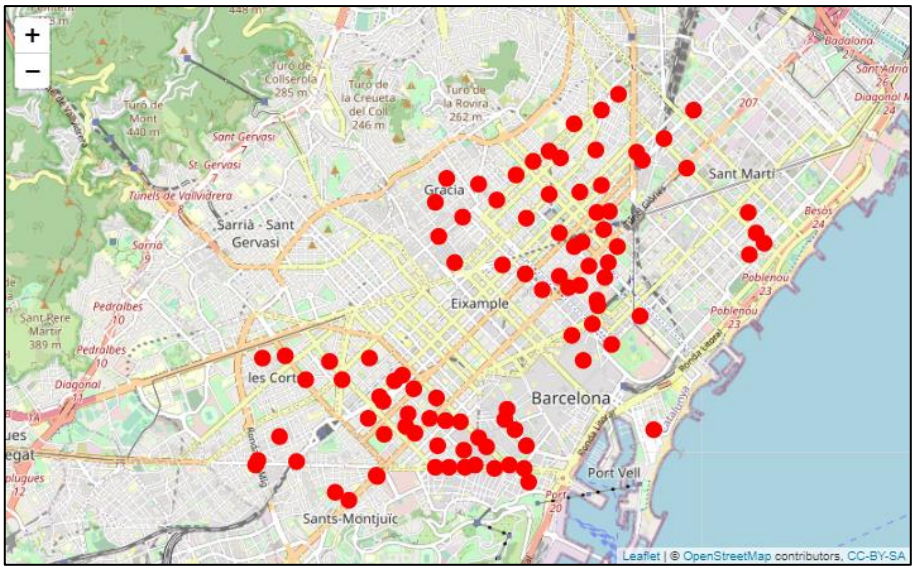

Figura 5.18: Mapa Cluster 1 Entre-Diari

Segons les projeccions de variables vistes anteriorment, el clúster 1 (el vermell en el mapa), es compon majoritàriament d'estacions que es fan servir sobretot com a punt de sortida del matí o punt d'arribada de la tarda. Aquest té una característica curiosa la qual deixa un espai buit d'estacions en la zona de la Ciutat Vella, una zona situada al mig de l'Eixample i la part Sud-est del barri de Sarrià-Sant Gervasi. L'altre clúster més utilitzat, el 3 (el blau en el mapa) es compon majoritàriament d'estacions de punts de destí durant el matí. Aquest és el següent que es veu en la Figura 5.19:

Figura 5.19: Mapa Cluster 3 Entre-Diari

En el diagrama de caixes de la primera component de la Figura 5.14, també destacava el clúster 4 per la seva poca utilització. Com es veu a la Figura 5.20, aquest correspondria a les estacions de taronja que es veuen en el mapa, les quals estan bastant disperses, però es concentren a la part exterior de Barcelona com per exemple l'agrupació que es veu al voltant del barri de Sant Andreu.

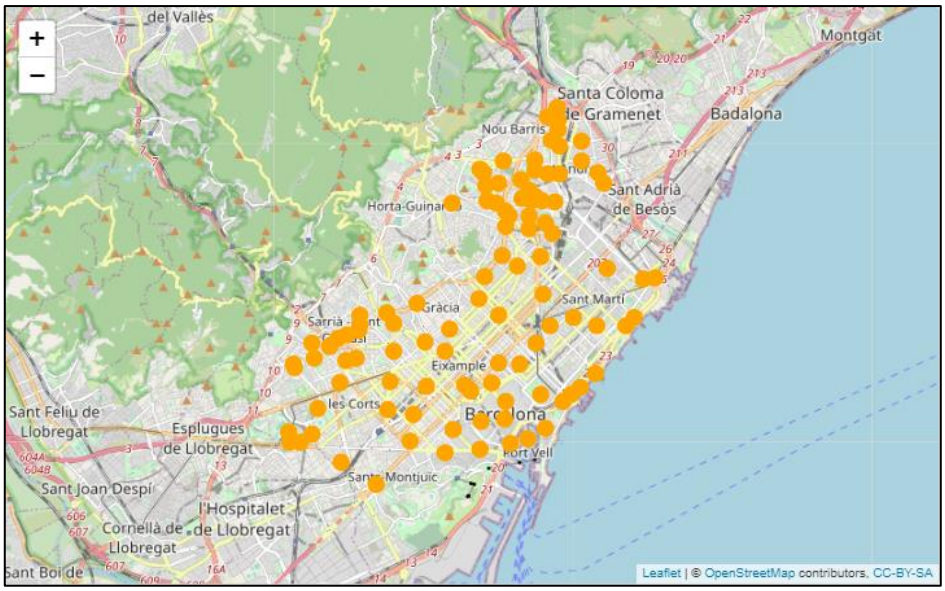

Figura 5.20: Mapa Cluster 4 Entre-Diari

Finalment, veiem la distribució del clúster 2 en la Figura 5.21. Aquest segons els diagrames de caixes o els diagrames de punts és el que aporta menys informació rellevant. En termes geogràfics no veiem cap patró o distribució d'estacions notable. L'únic a destacar és la concentració d'estacions a les zones exteriors: a la platja i a la zona oest al voltant de les Corts i Sarrià-Sant Gervasi. La resta d'estacions estan disperses per la resta de zones de Barcelona.

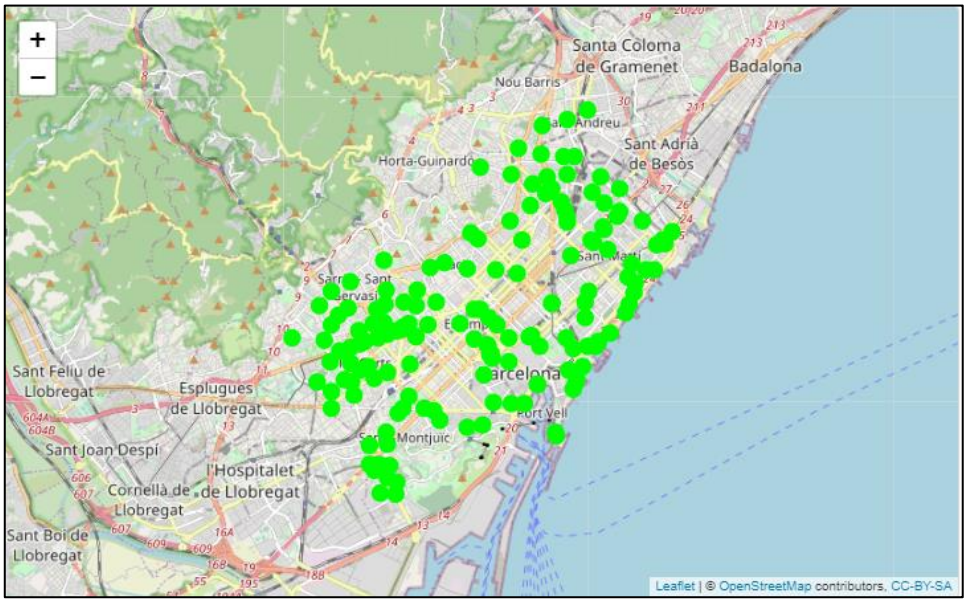

Figura 5.21: Mapa Cluster 2 Entre-Diari

5.3. NIVELL DIARI

5.3.1. PCA (Principal Component Analysis)

Per començar l'anàlisi del nivell diari (els dies de la setmana) s'ha de realitzar un anàlisi de les components principals. Això es fa amb el *dataframe* "df_Dia" el qual conté totes les arribades i sortides de totes les parts del dia. Un cop fem servir la funció *prcomp()* ens retorna el següent:

```
Standard deviations (1, .., p=14):
 [1] 3.50466872 0.78245725 0.49728590 0.43285375 0.41897522 0.37311437 0.34949517 0.26447864 0.23505217
[10] 0.20659984 0.17283691 0.13448366 0.11638142 0.06415331

Rotation (n x k) = (14 x 14):
                          PC1         PC2         PC3         PC4         PC5         PC6         PC7
Dilluns_Arribades   -0.2659739  0.27307173  0.33428886  0.30265555 -0.26583854  0.15238676 -0.022516650
Dilluns_Sortides    -0.2623017  0.30209851  0.36565351  0.28285242 -0.36973013  0.21897519 -0.015622416
Dimarts_Arribades   -0.2701035  0.22628777  0.15521103 -0.20751135  0.28738586 -0.25300179  0.387650057
Dimarts_Sortides    -0.2689222  0.26025434  0.15485226 -0.14918294  0.28679637 -0.20095142  0.450346793
Dimecres_Arribades  -0.2733257  0.18570771 -0.03828829 -0.23212617  0.09988202 -0.14219552 -0.539041160
Dimecres_Sortides   -0.2728784  0.19146240  0.01474180 -0.23434074  0.10232505 -0.20652110 -0.538217330
Dijous_Arribades    -0.2732778  0.07840809 -0.35769507 -0.07654965  0.12498621  0.46677808  0.088834333
Dijous_Sortides     -0.2717656  0.08246175 -0.35382726 -0.08991630  0.12987421  0.55268245  0.045425881
Divendres_Arribades -0.2707394 -0.03272749 -0.40865668  0.16272381 -0.34747378 -0.35229692  0.120640273
Divendres_Sortides  -0.2716163 -0.04452270 -0.42208098  0.16808411 -0.28946399 -0.31057083  0.091160083
Dissabte_Arribades  -0.2644148 -0.33246382  0.09770238  0.38258801  0.35169294  0.02415856 -0.121170002
Dissabte_Sortides   -0.2614060 -0.37214797  0.09098476  0.41983656  0.35647035 -0.08223579 -0.060270673
Diumenge_Arribades  -0.2564491 -0.44922221  0.18615790 -0.35615912 -0.22148672  0.07876348  0.121531950
Diumenge_Sortides   -0.2576511 -0.42101854  0.23243191 -0.36894137 -0.26553378  0.06410262  0.004104825
                          PC8         PC9         PC10        PC11        PC12        PC13        PC14
Dilluns_Arribades    0.40975045 -0.38471298  0.14632348 -0.23168214 -0.36064323 -0.02543876 -0.18773409
Dilluns_Sortides    -0.34029629  0.39946138 -0.10245567  0.19156796  0.30986780 -0.01322061  0.15024144
Dimarts_Arribades   -0.26208050  0.33360148 -0.06939792 -0.23671147 -0.39940850 -0.24965753 -0.22165936
Dimarts_Sortides     0.18718835 -0.35625462  0.03073800  0.28291333  0.37383028  0.22263915  0.23666218
Dimecres_Arribades   0.12257831 -0.06280860 -0.12517117  0.45882840  0.08015629 -0.37713578 -0.34767472
Dimecres_Sortides   -0.03580393  0.11103996  0.14697599 -0.40915879 -0.01287164  0.40802390  0.34881731
Dijous_Arribades    -0.09219100 -0.01215199  0.34169116 -0.30962468  0.41408900 -0.02185634 -0.38827359
Dijous_Sortides      0.09657251  0.02250994 -0.36558930  0.17268533 -0.37849167  0.01774495  0.38052146
Divendres_Arribades  0.06420895  0.03014518 -0.47673958 -0.07616827  0.06301568  0.36852610 -0.30797456
Divendres_Sortides  -0.11705870 -0.05909039  0.48509629  0.12435594 -0.11538354 -0.35753811  0.34664075
Dissabte_Arribades  -0.41837716 -0.16754198  0.16150787  0.28987440 -0.23269761  0.35119065 -0.18432244
Dissabte_Sortides    0.34896058  0.19295845 -0.20454097 -0.25920811  0.25900650 -0.32382340  0.17659888
Diumenge_Arribades   0.38247153  0.40795419  0.28156692  0.22339194 -0.08349307  0.20740328 -0.09495748
Diumenge_Sortides   -0.34281425 -0.44850837 -0.25466147 -0.21804764  0.08853160 -0.20716165  0.09874408
```

Figura 5.22: Llista resultats components principals 2

De la funció prcomp() de R obtenim un objecte, força complet, que aglutina els resultats necessaris per a l'estudi de components principals. L'objecte conté una llista *rotation*, que emmagatzema el valor dels *loadings* per a cada component. Analitzar amb detall el vector de *loadings* que forma cada component pot ajudar a interpretar quin tipus d'informació recull cadascuna.

Per exemple, la primera component igual que anteriorment és completament negativa. Això indica que hi ha una combinació lineal de les variables que resulta en valor negatiu. Si associem aquesta component a la mida, per tant, l'ús que es fa servir de les estacions, ens diu que podria haver sortit positiu o negatiu depenent de l'atzar.

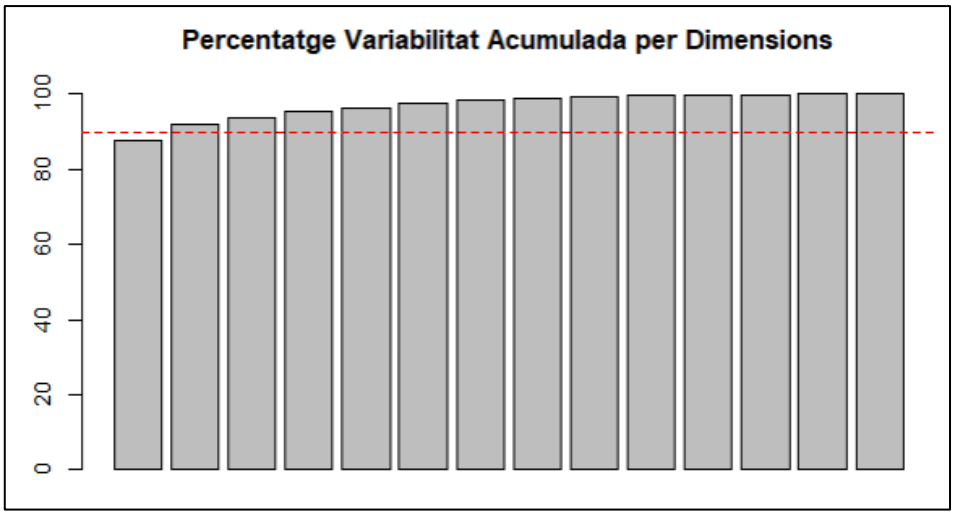

Figura 5.23: Percentatge Variabilitat Acumulada per Dimensions Diari

Del percentatge de la variabilitat acumulada de la Inèrcia Total, el que busquem nosaltres és una suma del 90%, per tant, tenint en compte aquest valor escollirem dos components per capturar una quantitat prou alta de la variabilitat de les dades.

Seguidament, es calculen i s'emmagatzemen els *eigenvectors*, els *eigenvalues* que formaran posteriorment les components principals. També ho farem amb la projecció en les 2 dimensions escollides anteriorment.

A continuació hi ha el *Scatter Plot* obtingut de les components en dues dimensions on es poden distingir lleugerament els números identificatius de les estacions:

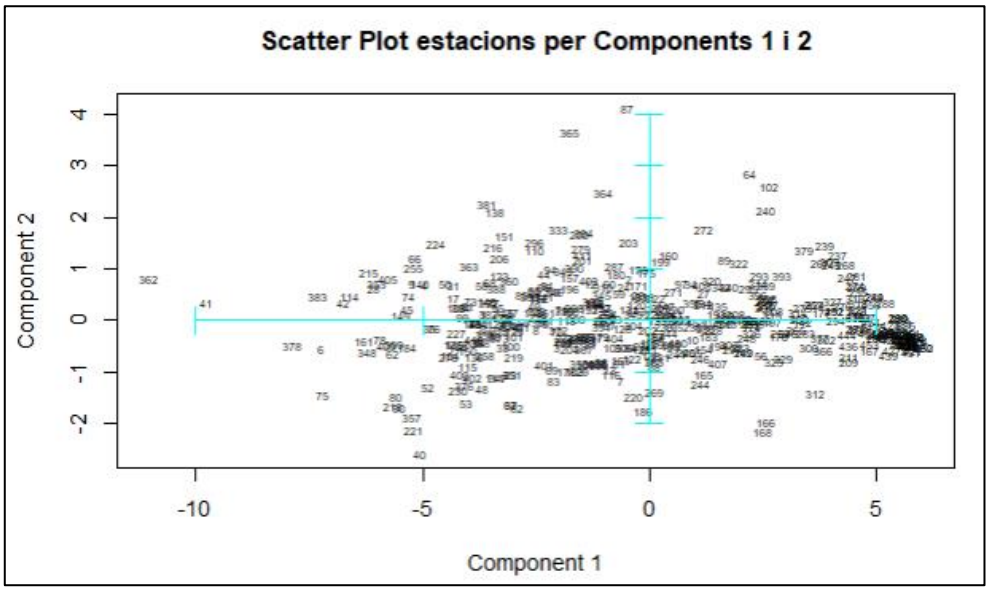

Figura 5.24: Scatter Plot estacions per Components 1 i 2 Diari

Si ens fixem en la Figura 5.19 hi ha una acumulació de dades on la component 1 és més positiva i està per sota del 0 de l'eix Y, mentre que a la part negativa i cèntrica es troba molta més dispersió. Fa l'efecte que d'esquerra a dreta hi ha gradualment més dispersió.

Cada punt en els gràfics correspon a una estació, identificada per l'etiqueta. Les coordenades horitzontals i verticals dels punts indiquen els valors de les components respectives. Mitjançant aquests gràfics, és possible observar com les estacions es distribueixen a les noves dimensions. L'ús de diferents parells de components ens ajuda a identificar patrons i agrupacions en funció de les combinacions de components.

Un cop obtingudes les components, s'han procedit a plasmar-les en un pla factorial amb fletxes:

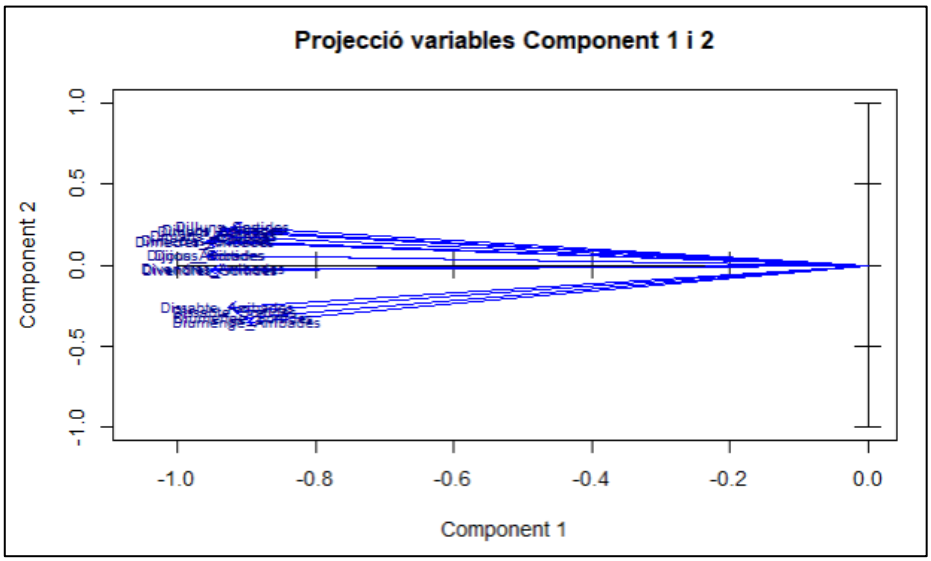

Figura 5.25: Projecció variables Component 1 i 2 Diari

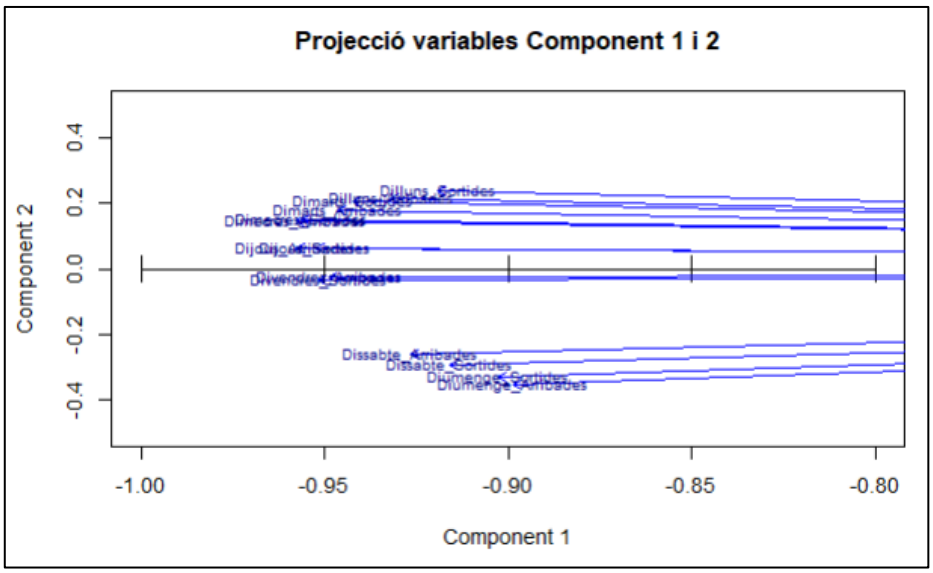

Figura 5.26: Ampliació Figura 5.20

En la Figura 5.20 i la Figura 5.21, totes les variables estan correlacionades negativament amb la primera component. De les variables, hi ha que estan relacionades entre elles, “Dissabte_Arribades”, “Dissabte_Sortides”, “Diumenge_Arribades” i “Diumenge_Sortides”. Totes les variables són les importants per a PC1.

De la PC2 podem extreure que un valor negatiu està relacionat amb una estació que es fa servir més durant el cap de setmana. Si et fixes, veuràs que en la Figura 5.21 hi ha una graduació de positiu a negatiu de dilluns a diumenge respectivament. Per tant, si el 0 és considerat divendres com es veu a la Figura, un valor amb tendència negativa es decantaria més per una estació utilitzada més al cap de setmana mentre que un valor positiu seria una estació que s'utilitza més durant la resta de dies, dilluns a dijous.

Resumint, es podria dir que la PC1 té a veure amb l'ús que es fa de les estacions, com més negatiu sigui el valor més es fa ús de l'estació. Totes tenen correlació negativa el qual indica que hi ha una combinació lineal de les variables que resulta en valor negatiu. La PC2 està relacionada amb els dies de la setmana que facin servir més les estacions, com més negatiu sigui el valor de la segona component més probable és l'estació de ser més utilitzada durant el cap de setmana.

A continuació es veu un gràfic més detallat sobre el repartiment dels percentatges segons les dimensions de la separació entre dies de la setmana:

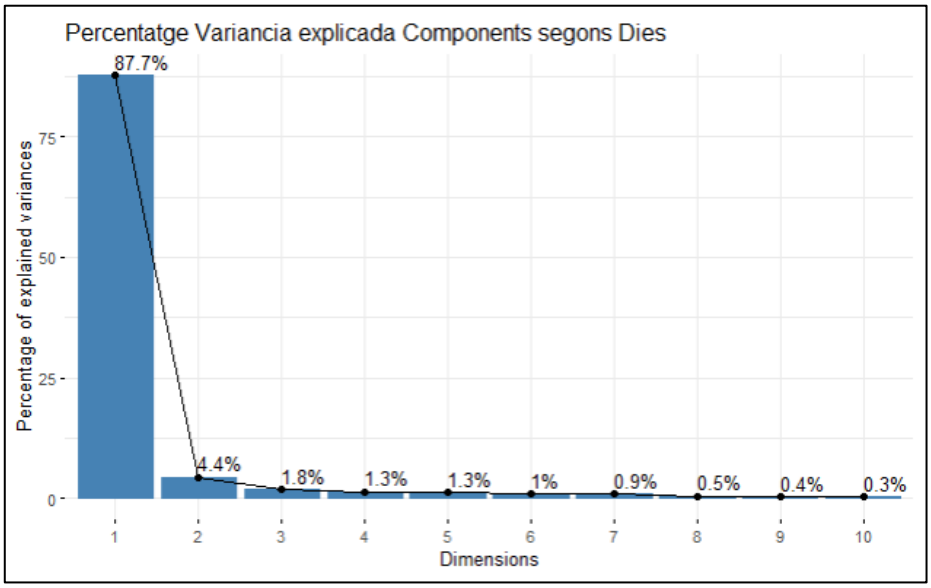

Figura 5.27: Percentatge Variabilitat Dimensions Dies

5.3.2. CLUSTERING

El mètode que he utilitzat per fer els clústers és el mètode de Ward, ja que és un mètode que funciona bé quan la base de dades té poques observacions, en el meu cas tinc només 456 observacions (o estacions), i una de les seves característiques és que minimitza la variància dins de cada clúster.

El dendrograma resultant es veu representat en la Figura 5.9 següent:

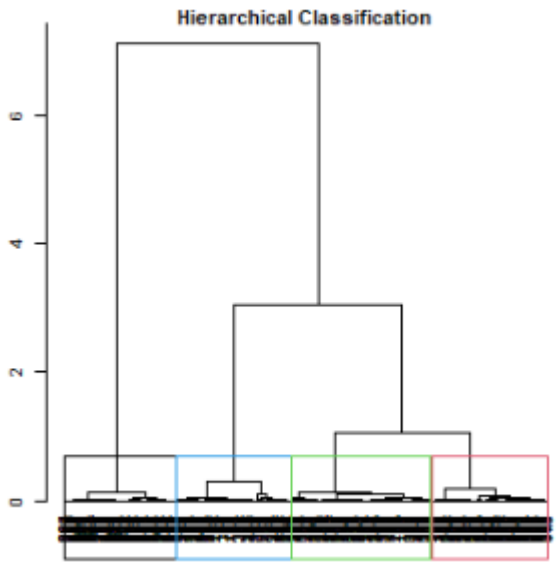

Figura 5.28: Dendograma classificació jeràrquica segons Dies.

Al dendrograma veiem clarament que el nombre de grups en què podrem dividir la base de dades serà 4. Tots els grups son relativament semblants, per tant el nombre d'observacions que tindrà cada clúster és:

CLÚSTER	1	2	3	4
Nº observacions	109	129	120	98

Taula 5.3: Observacions segons Clúster Dies

5.3.3. PROFILING

A continuació analitzarem en detenció el *Scatter Plot* vist anteriorment però aquest cop diferenciant segons els diferents Clústers. Observem en dos dimensions la primera component i la segona component:

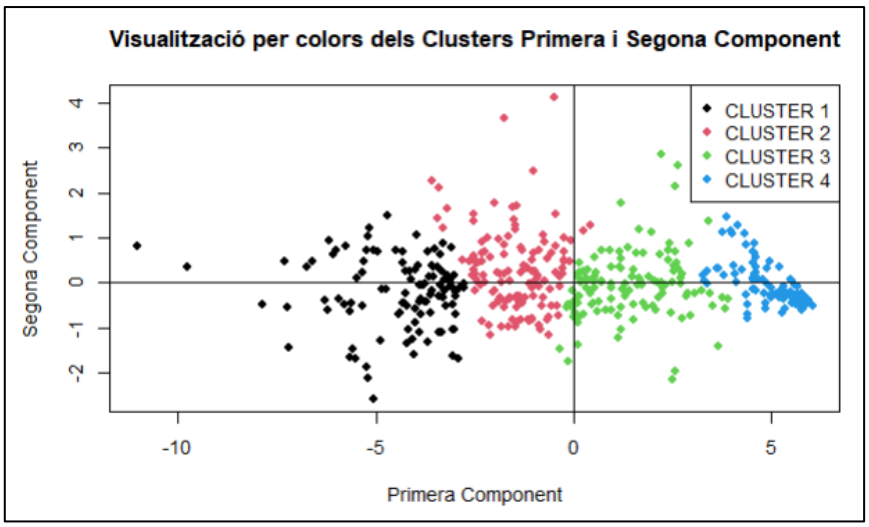

Figura 5.29: Scatter Plot Clústers Component 1-2 Dies

Si mirem la Figura 5.24, lamentablement no podem diferenciar amb els clústers cap diferència clara respecte a la segona component. Ja que la primera component agrupa gairebé la totalitat de la variabilitat no es pot veure res significatiu de la segona component.

El que fa aquest gràfic bidimensional és separar els clústers segons la graduació de la primera component. Com més negativa la primera component més dispersió trobes. Igual que amb l'anàlisi Entre-Diari, aquí trobem una agrupació compacta d'estacions en l'extrem positiu de la PC1, les quals estan associades al clúster 4.

Segons la projecció de variables que hem vist anteriorment, aquesta distribució ens explica que les estacions blaves són les que s'utilitzen al cap de setmana a causa de tenir un cúmul de punts per sota de 0 de l'eix Y. Si això ho connectem amb la poca utilització de l'extrem positiu de la primera component, podem extreure que en la majoria de les estacions del clúster 4 són les que menys es fan servir durant el cap de setmana.

Per continuar amb la interpretació dels grups creats pel *Clustering*, seguidament veurem uns *boxplots* associats a les dues components significatives.

Començant amb la Figura 5.25, veiem un increment constant entre els diferents clústers. Una de les coses que crida l'atenció és lo agrupades que estan les dades dins dels diferents grups. Realment hi ha molt poca dispersió observant les caixes interquartíliques i quan s'acaba un grup respecte a l'eix Y comença un altre.

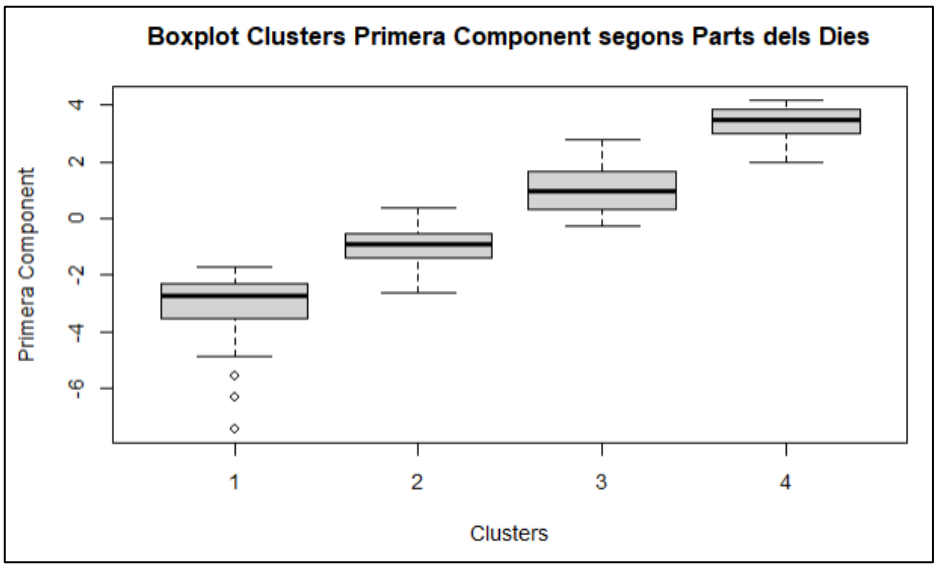

Figura 5.30: Box plot Clúster Dies 1ª Component

Això dona peu a les observacions anteriors respecte al fet que la primera component es refereix a la mida, o sigui l'ús que es fa de les estacions. Veient aquest Box Plot podem confirmar que les estacions dintre del clúster 1 son les més utilitzades mentre que les del clúster 4 són les que menys.

Si ens fixem en el clúster restant de la component 2, podem extreure poca cosa. Com ens podem fixar en la Figura 5.26 tots els clústers tenen valors molt semblants al voltant del 0. L'únic a destacar seria la poca dispersió en el clúster 4 de les seves dades i la lleugera decantació positiva del clúster 2.

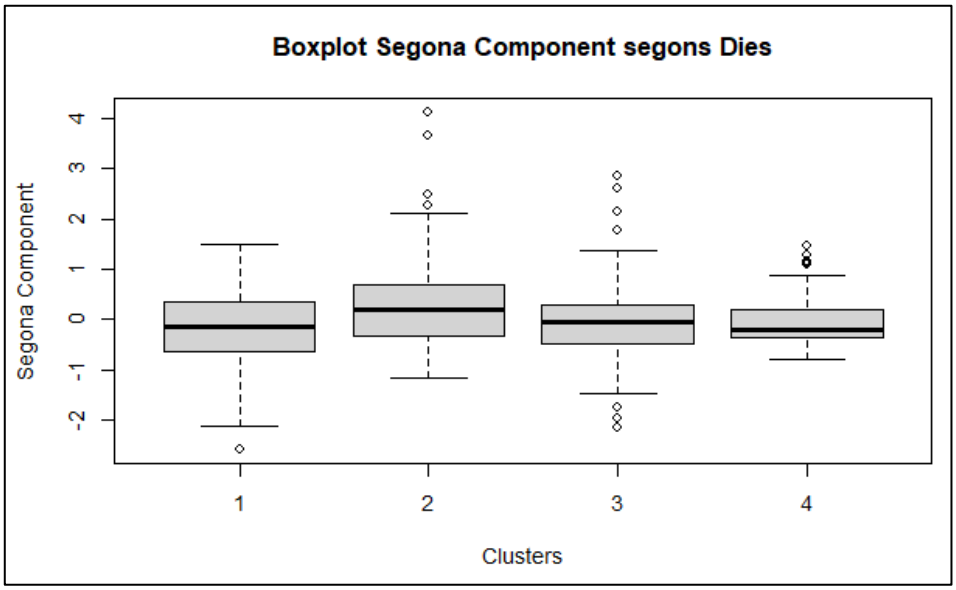

Figura 5.31: Box plot Clúster Dies 2ª Component

En resum la Figura 5.25 ens representa fàcilment la distribució dels diferents clústers, la qual fa pensar en els diferents graus d'ús de les estacions, mentre que la Figura 5.26 ens ofereix una informació nul·la la qual no permet detectar patrons ni distribucions dels clústers d'una forma clara.

Per finalitzar s'observa una petita taula que mostra les medianes de cada clúster segons les components per tenir present les diferències d'una forma més clara:

	PC1	PC2
CLUSTER 1	-4.414090	-0.23195486
CLUSTER 2	-1.455794	0.28159838
CLUSTER 3	1.520170	-0.05742438
CLUSTER 4	4.964416	-0.04236924

Taula 5.4: Medianes Clusters segons Components Dies

Per acabar amb el *profiling*, veurem la distribució geogràfica dels diferents clústers en un mapa de Barcelona per veure si hi ha diferències significatives. Els mapes que es veuen a continuació s'han creat amb la llibreria "leaflet" de R.

Comencem veient una representació conjunta dels diferents clústers en la Figura 5.32:

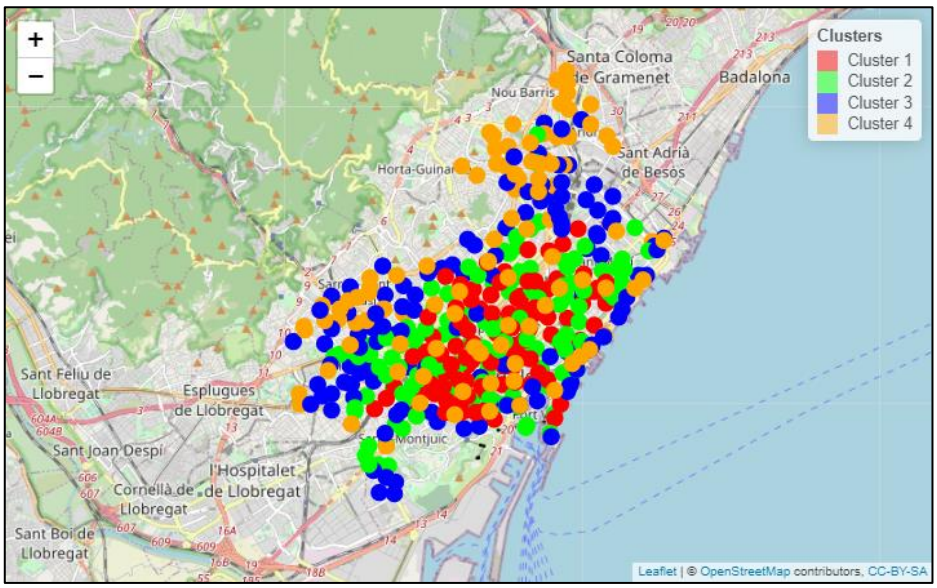

Figura 5.32: Mapa conjunt clusters Diari

Si ens fixem en els diferents grups de la Figura 5.32, podem veure una diferenciació de la part cèntrica amb la part exterior. El clúster 3 i el 4 (blau i taronja respectivament) envolten els altres d'una manera predominant per la part exterior. Mentre que el clúster 1 i 2 (vermell i verd respectivament) s'agrupen més en la part cèntrica, sobretot el clúster 1. Si recordem la Figura 5.30, el clúster 1 representava les estacions més utilitzades.

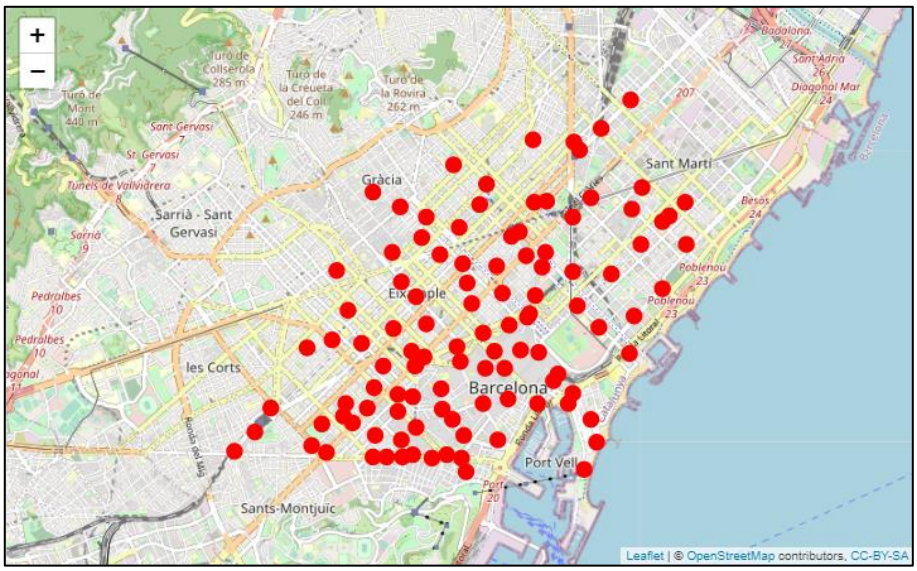

Figura 5.33: Mapa Cluster 1 Diari

Segons els diagrames de caixes vists anteriorment, el clúster 1 (el vermell en el mapa), es compon majoritàriament de les estacions que es fan servir més comparat amb els altres clústers. Aquestes estacions estan principalment situades en zones més cèntriques de Barcelona. El següent clúster més utilitzat, el 2 (el verd en el mapa) es compon d'estacions lleugerament menys cèntriques que el clúster

1 amb més dispersió pel mapa de Barcelona. Aquestes estacions es poden veure a continuació en la Figura 5.34:

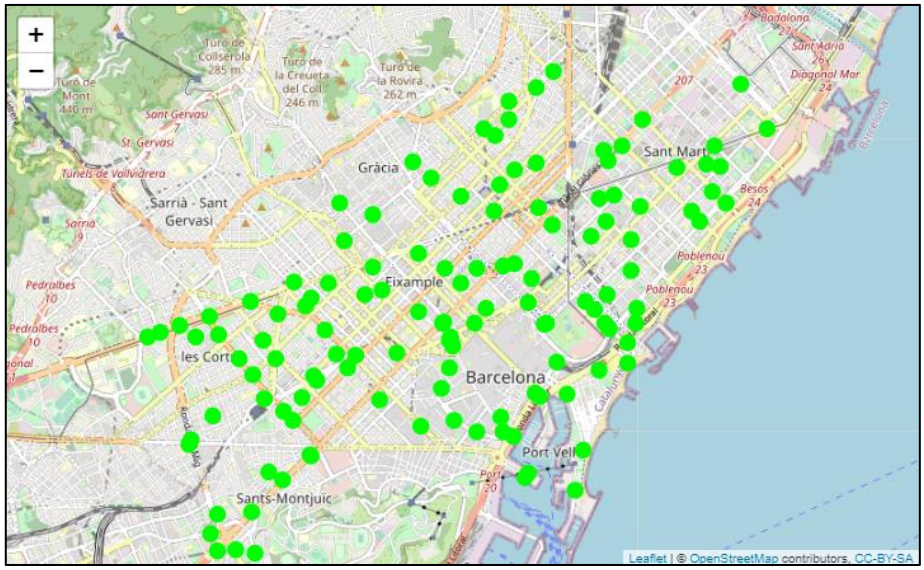

Figura 5.34: Mapa Cluster 2 Diari

En el diagrama de caixes de la primera component de la Figura 5.30, també destacava el clúster 4 per la seva poca utilització. Com es veu a la Figura 5.35, aquest correspondria a les estacions de taronja que es veuen en el mapa, les quals estan bastant disperses, però es concentren a la part exterior de Barcelona com per exemple l'agrupació que es veu al voltant del barri de Sant Andreu. El curiós d'aquest clúster és que les estacions són gairebé les mateixes que en l'apartat d'Entre-Dies que es veu a la Figura 5.20. Això significa que s'estudiï segons els dies de la setmana o les parts dels dies, les estacions que menys es fan servir són les mateixes.

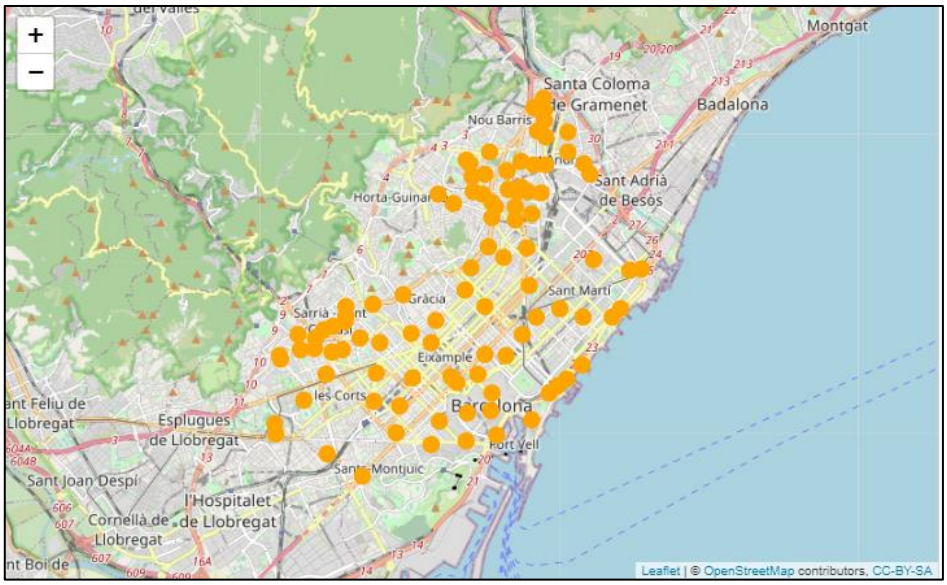

Figura 5.35: Mapa Cluster 4 Diari

Finalment, veiem la distribució del clúster 3 en la Figura 5.36. Aquest grup segons els diagrames de caixes o els diagrames de punts és el que aporta menys informació rellevant. En termes geogràfics

podem veure un patró curiós que agrupa totes les estacions que són menys cèntriques d'una forma clara. Aquestes deixen un buit el mig on se situa l'Eixample, i es localitzen en zones exteriors. Les zones on més es concentren són: la zona marítima, el sector al voltant de la sagrera i la zona de les Corts juntament amb la zona de Sarrià.

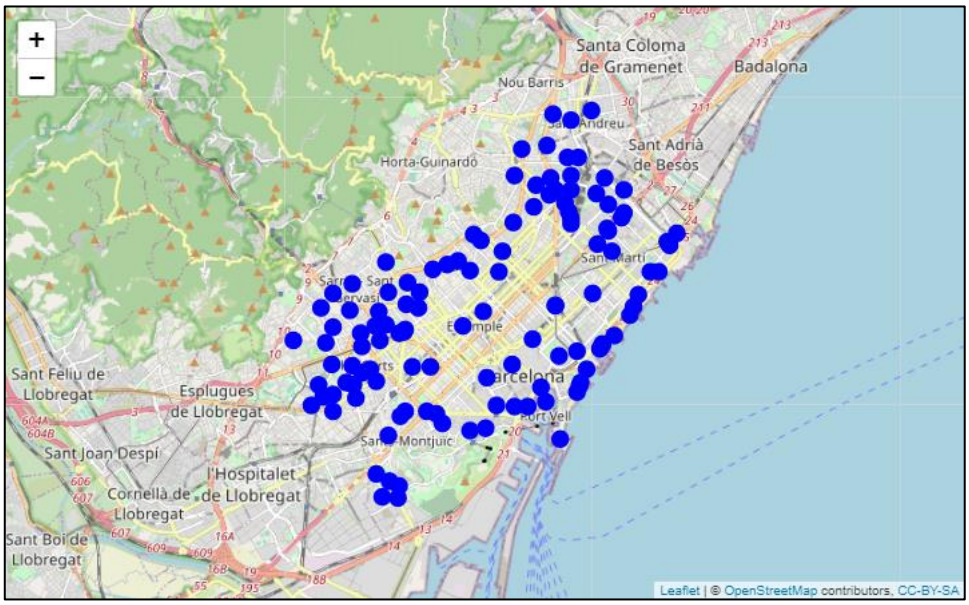

Figura 5.36: Mapa Cluster 3 Diari

6. CONCLUSIONS

Finalitzant aquest treball veiem que els objectius plantejats a l'inici s'han complert. A partir de l'anàlisi exhaustiva realitzada en aquest treball sobre l'ús del Bicing a Barcelona des d'una perspectiva espai-temporal, s'han obtingut conclusions que poden contribuir de manera significativa a la millora del sistema de transport urbà sostenible i a la reducció de la contaminació a la ciutat. Les troballes d'aquest estudi ofereixen una visió detallada dels patrons de comportament dels usuaris del Bicing i permeten identificar millores operatives en el sistema.

En relació amb els objectius principals d'aquesta investigació, s'ha assolit l'objectiu de comprendre els patrons d'ús diari i entre-diari del Bicing a través de l'anàlisi detallada de diferents dimensions temporals. En la dimensió diària, s'ha observat que hi ha variacions significatives en l'ús de les bicicletes entre els diferents dies de la setmana, separant els dies laborables dels festius. Aquesta diferenciació és útil per dissenyar estratègies específiques de promoció i gestió de l'ús del Bicing en funció dels diferents dies. Això es podria aconseguir a través d'una major reposició de bicis segons el període de la setmana en les zones més freqüentades.

Pel que fa a la dimensió entre-diària, l'anàlisi per segments de temps (matí, migdia, tarda i nit) ha proporcionat informació valuosa sobre les preferències d'ús de les bicicletes en diferents moments del dia. Aquesta comprensió pot servir de base per a la millora de la gestió de l'oferta de bicicletes i la distribució de les mateixes segons la demanda específica de cada hora. S'han identificat hores punta d'ús, les quals podrien estar relacionades amb els desplaçaments laborals o altres activitats específiques.

En termes de la plataforma creada, l'elaboració del projecte Shiny per a la visualització i anàlisi dels resultats de l'estudi espai-temporal de l'ús del Bicing a Barcelona ha estat una experiència gratificant, i amb resultats de tota mena. La creació de l'eina a través de R amb la llibreria 'Shiny' i 'Shinydashboard' ha superat les expectatives i s'ha desenvolupat d'una manera fluida i entretinguda. En cap moment aquesta tasca s'ha convertit en una càrrega, sinó que ha sigut una part enriquidora del procés de recerca.

La implementació de la aplicació Shiny ha estat més efectiva del que m'imaginava. Tots els apartats considerats al començament han pres forma juntament amb altres funcionalitats prèviament pensades d'una manera estructurada i intuïtiva. L'eina resulta útil per als usuaris que volen explorar i entendre millor les dades del Bicing, i la seva interfície, juntament amb les característiques de visualització avançada, ofereixen una experiència satisfactòria per als

usuaris que volen comprendre els patrons d'ús de les estacions en diferents dimensions temporals.

La seva utilitat permet als usuaris, amb facilitat, explorar i analitzar la gran quantitat de dades generades pel sistema de Bicing a Barcelona. Aquesta eina no només contribueix al mateix treball, sinó que també té el potencial de ser una eina valuosa per als professionals i investigadors interessats en la mobilitat urbana sostenible i la seva millora continua.

La plataforma està completament adaptada amb caixes d'informació de forma contínua o amb un glossari a la part final per si l'usuari necessita ajuda amb alguns termes o qüestions. Està tot creat perquè l'experiència de l'usuari sigui intuïtiva i hi hagi les mínimes confusions. De principi a final està dotat de gràfics o mapes interactius que permeten canviar els valors de les variables desitjades o visualitzar els resultats d'una altra manera. L'informe conté una explicació detallada amb cada pas a seguir per entendre la interfície i amb imatges que representen cada apartat.

En termes de l'anàlisi geogràfica dels resultats segons els diferents factors de Dies o Entre-Dies hem trobat conclusions diverses. Es destaca la clusterització en quatre grups de les estacions de l'àrea de Barcelona. Aquests clústers revelen diferències significatives en els patrons d'ús. El clúster 1 i 3 es caracteritzen per ser àrees cèntriques de concentració d'estacions més utilitzades, especialment com a punts de sortida del matí o punts d'arribada de la tarda. A més, es destaca la particularitat del clúster 1, amb una zona buida a la part del mig de Barcelona que va de mar a muntanya.

D'altra banda, el clúster 2 i 4 abasten principalment les zones exteriors. El clúster 2 no presenta un patró geogràfic discernible, mentre que el clúster 4 mostra estacions disperses, concentrades a l'exterior de Barcelona i amb una relativa poca utilització.

Respecte a l'estudi del factor diari, també s'ha clusteritzat les dades en quatre grups, alguns amb característiques semblants a les del component anterior de parts del dia (Entre-Diari). El clúster 1 representa estacions molt usades i està situat principalment en zones cèntriques de la ciutat mentre que el clúster 2 mostra estacions menys cèntriques que el clúster 1, amb una distribució més dispersa pel mapa de Barcelona.

D'altra banda, el clúster 3 i 4 són grups de poca utilització que envolten les altres zones, predominantment per la part exterior de la ciutat. El clúster 4, amb estacions escasses i amb una utilització baixa, es concentra a l'exterior, especialment al voltant del barri de Sant Andreu. Curiosament, es destaca que les estacions menys utilitzades són les mateixes en els clústers tant del factor diari com de l'entre-diari. Això suggereix una constant manca d'ús d'aquestes estacions.

Observant les diferències i les característiques de cada clúster format per diferents estacions, es poden organitzar reposicions de bicis d'una forma més eficaç geogràficament perquè sempre hi hagi bicis disponibles o la creació d'estacions en zones més transitades. D'aquesta forma es milloraria el sistema de Bicing, promocionant una millora en el transport d'una forma saludable i sostenible pels usuaris.

En definitiva, mitjançant la clusterització i l'anàlisi geogràfic de les dades, es posa de manifest la diversitat de comportaments d'ús de les estacions de Bicing. Cada clúster ofereix una perspectiva única de la utilització de les estacions en funció de la seva ubicació i les bicis disponibles, el qual permet una comprensió més profunda dels diferents patrons d'ús a la ciutat de Barcelona.

Com a possibles extensions futures, aquest estudi pot estendre el seu impacte més enllà de Barcelona i aplicar-se a altres ciutats europees o mundials amb sistemes de bicicletes similars. La metodologia emprada en aquest treball ofereix una estructura adaptable per explorar patrons d'ús específics en diverses ciutats. A més, seria interessant considerar canvis en la selecció dels dies de la setmana, com per exemple explorar una altra setmana aleatòria o exclusiva de dies festius, per d'aquesta forma captar patrons d'ús especials.

També es podria considerar factors addicionals com l'altitud i els ingressos per zones geogràfiques, com ara barris. Aquesta ampliació podria conduir a la identificació de relacions entre l'ús de les bicicletes i les característiques socioeconòmiques de diferents àrees de la ciutat, contribuint a una millor comprensió dels patrons d'ús en funció de múltiples factors.

Per acabar, el procés d'aquest treball ha estat molt enriquidor, ja que m'ha permès reforçar els meus coneixements estadístics i aprofundir en el món de les dades geogràfiques. L'ús de l'eina Shiny va ser una novetat gratificant, i a través d'aquest projecte, he descobert diverses llibreries de R amb potencial per a futures exploracions. Aquesta experiència no només ha sigut útil per l'estudi sobre el Bicing a Barcelona, sinó que també ha ampliat les meves habilitats en l'anàlisi de dades i programació, preparant-me per reptes futurs.

BIBLIOGRAFIA

CHENG, J; KARAMBELKAR, B; XIE, Y; WICKHAM, H. (2021). leaflet: Create Interactive Web Maps with the JavaScript 'Leaflet' Library. R package version 2.0.4. URL: https://cran.rproject.org/web/packages/leaflet/index.html

CHANG, Winston; BORGES, Barbara. (2021). shinydashboard: Create Dashboards with 'Shiny'. R package version 0.7.2. URL: https://rstudio.github.io/shinydashboard/

RStudio Team (2021). Shiny: Web Application Framework for R. R package version 1.7.4. URL: https://shiny.rstudio.com/

WICKHAM, Hadley, (2009). ggplot2: elegant graphics for data analysis. Springer New York. ISBN 978-0-387-98140-6. URL: http://had.co.nz/ggplot2/book.

WICKHAM, Hadley; FRANÇOIS, Romain; HENRY, Lionel; Müller, Kirill. (2021). dplyr: A Grammar of Data Manipulation. R package version 1.1.2. URL: https://cran.r-project.org/web/packages/dplyr/index.html

ATTALI, Dean; (2021). shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds. R package version 2.1.0. URL: https://cran.r-project.org/web/packages/shinyjs/index.html

MAECHLER, M; ROUSSEEUW, P; STRUYF, A; HUBERT, M; HORNIK, K. (2021). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.4. URL: https://cran.rproject.org/web/packages/cluster/index.html.

KASSAMBARA, A; MUNDT, F. (2020). factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. URL: https://cran.r-roject.org/web/packages/factoextra/index.html

SIEVERT, C. (2021). plotly for R. URL: https://plotly.com/r/

GROLEMUND, G; WICKHAM, H. (2021). lubridate: Make Dealing with Dates a Little Easier. R package version 1.9.2. URL: https://cran.r-project.org/web/packages/lubridate/index.html.

PEBESMA, E. (2021). sp: Classes and Methods for Spatial Data. R package version 1.6.0. URL: https://cran.r-project.org/web/packages/sp/index.html

FOX, J; WEISBERG, S. (2019). car: Companion to Applied Regression. R package version 3.1-1. URL: https://cran.r-project.org/web/packages/car/index.html

GIBERT, Karina, (2022) Apunts de l'assignatura Mineria de Dades impartida per la Karina Gibert, Sergí Ramírez i Dante Conti. URL: https://www-eio.upc.edu/teaching/DocenciaMultivariant/DMGEST/

ANNEX

- SHINY UI

```
library(shinydashboard)
library(shiny)
library(ggplot2)
library(dplyr)
library(leaflet)
library(sp)
library(lubridate)
library(car)
library(cluster)
library(factoextra)
library(plotly)
library(shinyjs)

ui <- dashboardPage(
  dashboardHeader(title="Flux Bicing BCN"),
  dashboardSidebar(
    sidebarMenu(
      menuItem("Guia Aplicació", tabName = "tab0", icon =icon("house")),
      menuItem("Flux segons Dia i Estació", tabName = "tab1", icon =
icon("calendar")),
      menuItem("Flux segons Estació", tabName = "tab2", icon =
icon("bicycle")),
      menuItem("Diferències Estacions Bici", tabName = "tab3", icon =
icon("map")),
      menuItem("Arribades/Sortides estacions", tabName = "tab4", icon =
icon("exchange")),
      menuItem("Perfils estacions", tabName="tab5",icon= icon("menu-
hamburger", lib= "glyphicon"),
                    startExpanded = TRUE,
                    menuSubItem("Clusters segons la Part del Dia", tabName
= "submenu1"),
                    menuSubItem("Clusters segons el Dia", tabName =
"submenu2"),
                    menuSubItem("Addicional", tabName="submenu3")),
      menuItem("Exploracio avançada", tabName = "tab6", icon= icon("star",
lib="glyphicon"),
                    startExpanded = TRUE,
                    menuSubItem("Segons la Part del Dia",
tabName="submenu5"),
                    menuSubItem("Segons el Dia", tabName = "submenu4")),
      menuItem("Glossari", tabName = "tab7", icon= icon("book",
lib="glyphicon"))
  ),

    tags$head(tags$style(HTML('.shiny-server-account { display: none;
}'))),

    uiOutput("userpanel")
  ),

  dashboardBody(
      tabItems(
      tabItem("tab0",
              fluidRow(
                box(
                  title=tags$h3("Guia Aplicació", style = "font-size: 30px;
font-weight: bold;"), solidHeader=TRUE, status= "primary",
                  width= 10,
                  p(style="font-size: 18px",
"
Se't dona la benvinguda a l'aplicació de l'espai Shiny per la visualització
de les dades del Bicing
```

```
de Barcelona durant la setmana 07/01/2019 - 13/01/2019, del dilluns al
diumenge respectivament. En aquesta aplicació, hi trobaràs un petit resum
d'una forma interpretativa i senzilla d'utilitzar per a la fàcil comprensió
de qualsevol persona.
Dirigida a tota classe de públic, l'espai Shiny té com a objectiu presentar
el que són unes dades
molt extenses i lleugerament confuses d'una forma transparent, per així
entendre d'una
manera més clara el flux, la utilització i tot el que comporta el Bicing de
Barcelona i el sistema que
està creat per darrere."),
                  p(style="font-size: 18px",
"Com podràs observar, a la vostra esquerra trobareu un menú amb diversos
apartats de visualització
de dades. Des d'un mapa de Barcelona que et mostra en quines estacions hi
ha flux d'entrada o sortida
de bicis, fins histogrames que et representen el flux de bicis entre les
diferents hores del dia.
Separats entre Matí, Migdia, Tarda i Nit o segons Arribades i Sortides.
També tens l'opció d'escollir
l'estació que vols visualitzar o el dia que vols veure representat."),
                  p(style="font-size: 18px",
"Si en algun moment tens la sensació que et falta espai per la correcta
visualització de gràfics o mapes
hi ha l'opció d'incrementar el camp de visió donant clic a les tres barres
horitzontals que trobaràs a la
part superior esquerra de la pantalla.",
                  p(style="font-size: 18px",
"Per ajudar a la fàcil comprensió dels resultats com gràfics o mapes, s'ha
inclòs en cada
apartat un petit requadre d'informació que podrà consultar si es veu amb la
necessitat. Allà trobarà una breu descripció del que estarà veient per així
entendre-ho millor."),
                  p(style="font-size: 18px",
"Al llarg dels diferents apartats et trobaràs conceptes amb un asterisc
(*), això és degut al glossari que hi ha al final del menú. Allà veuràs un
petit recull de definicions o explicacions de termes concrets que surten a
l'aplicació de Flux Bicing BCN."))
                )
              )),
      tabItem("tab1",
              fluidRow(
                box(title="Gràfic interactiu del flux de bicis segons Dia i
Estació",solidHeader=TRUE,width= 9, plotOutput("bike_plot"),status=
"primary"),
                box(title="Segons Dia i Estació", solidHeader=TRUE,width=3,
status= "primary",
                    selectInput("dia", label = "Selecciona un dia",
                  choices = unique(fitxer_char$dia)),
                    selectInput("estacio1", label = "Selecciona una
estació",
                    choices = unique(fitxer_char$nom_numero)))),
              fluidRow(
                box(title=HTML('<div><i class="fa fa-info-circle"
style="font-size: 24px;"></i> PUNT DE
INFORMACIÓ</div>'),width=9,status="info",solidHeader=TRUE,
                    "En aquesta gràfica pots observar el flux de les bicis
segons les estacions i els dies que hagis escollit al recuadre selectiu. En
l'eix d'abscisses*, on diu HORES, podem veure els diferents horaris al
llarg del dia triat. En el gràfic es pot veure canvis bruscs pel fet que hi
ha actualitzacions cada 5 minuts de l'estat de l'estació. L'eix
d'ordenades*, on diu BICIS DISPONIBLES, mostra la quantitat de bicis que
estan per utilitzar a l'estació. En alguna estació et podràs fixar que hi
ha repunts o davallades de bicis disponibles, això és degut a les
reposicions. En el cas de que no trobis l'estació que desitjes desplegant
la pestanya de selecció, també pots buscar-la escrivint el seu nom.")
              )
              ),
      tabItem("tab2",
              fluidRow(
```

```
                box(title="Gràfic interactiu del flux de bicis segons
Estacions",solidHeader=TRUE,width=9, plotOutput("bike_plot2"),status=
"primary"),
                box(title= "Segons Estació",solidHeader=TRUE, width=3,
status= "primary",
       # Selecciona estacio
       selectInput(inputId = "estacio2", label = "Selecciona una estació",
                   choices = unique(fitxer_char$nom_numero)))
              ),
              fluidRow(
                box(title=HTML('<div><i class="fa fa-info-circle"
style="font-size: 24px;"></i> PUNT DE INFORMACIÓ</div>'), width= 9,
status="info",solidHeader=TRUE,
                    "En aquesta gràfica pots observar el nombre de bicis
disponibles al llarg de la setmana de l'estació escollida. Diferenciant els
dies en diversos colors amb l'ajuda d'una llegenda la dreta, el que pretén
aquest gràfic és visualitzar les diferències que hi ha en la utilització de
bicis segons els diferents dies. Com et podràs fixar, com més baixa sigui
la línia de seguiment de bicis, més en seran utilitzades en aquell moment.
En el cas de que no trobis l'estació que desitjes desplegant la pestanya de
selecció, també pots buscar-la escrivint el seu nom.")
              )
      ),
      tabItem("tab3",
              fluidRow(
                box(title= "Mapa Interactiu Arribades i Sortides", width=
6, leafletOutput("mapadif"),status= "primary"),
                box(title="Escull el dia i l'hora que
busques",solidHeader=TRUE, status= "primary",
        sliderInput(inputId="HORA", label= "Selecciona la hora",
                    min=0, max=23, value=0),
        selectInput(inputId="DIA", label= "Selecciona un dia",
                    choices=unique(cleandata_hores$dia))),
                box(title="LLEGENDA", background= "light-blue",
                   HTML( " <span style='color: lightgreen;'>Estació
verda</span> --> Més arribades que sortides <br/>"," <span style='color:
black;'>Estació negra</span> --> Mateix nombre d'arribades que sortides
<br/>","<span style='color: red;'>Estació vermella</span> --> Més sortides
que arribades"))
              ),
              fluidRow(
                box(title=HTML('<div><i class="fa fa-info-circle"
style="font-size: 24px;"></i> PUNT DE INFORMACIÓ</div>'), width=
6,status="info",solidHeader=TRUE,
                    paste0("Aquest mapa mostra la diferència entre
arribades i sortides de les estacions que es troben a la ciutat de
Barcelona. Sent cada cercle una estació diferent, si cliques a la que
desitges pots veure-hi informació com la seva adreça o la diferència entre
les arribades i sortides durant l'hora escollida. Aquesta s'escull al
requadre que pots trobar en la part superior a la dreta. ",
"Els colors de les estacions ajuden a diferenciar si una estació ha rebut o
ha deixat anar més bicis. Una estació verda ha rebut més bicis de les que
ha deixat anar mentre que una vermella és el contrari. Una negra ens dona a
entendre que no hi ha hagut activitat o que han arribat el mateix nombre de
bicis que han marxat. Si et fixes durant horaris nocturns predominen més
els cercle negres, sobretot de 1:00am - 6:00am."))
              )
      ),
      tabItem("tab4",
               fluidRow(
                tabBox(title= "Histogrames Arribades/Sortides de les
estacions segons el Dia, la Part del dia i Hores", width=9, id= "tabset1",
                       tabPanel(HTML(paste0(icon("chevron-up", lib=
"glyphicon"), " Arribades")),  plotOutput("bike_plot3"),status= "primary"),
                       tabPanel(HTML(paste0(icon("chevron-down", lib=
"glyphicon"), " Sortides")), plotOutput("bike_plot4"),status= "primary")),
       box(title= "Arribades i Sortides",solidHeader=TRUE, width=3, status=
"primary",
       # Selecciona el dia
       selectInput(inputId= "day", label ="Selecciona un dia",
```

```
                    choices= unique(cleandata_hores$dia)),
       # Selecciona l'estació
       selectInput(inputId = "estacio3", label = "Selecciona una estació",
                    choices = unique(cleandata_hores$nom_numero)))
              ),
               fluidRow(
                 box(title=HTML('<div><i class="fa fa-info-circle"
style="font-size: 24px;"></i> PUNT DE INFORMACIÓ</div>'), width= 9,
status="info",solidHeader=TRUE,
                     "En aquests histogrames* pots observar de forma gràfica
el nombre d'arribades i sortides que hi ha a les estacions.
Aquests els pots seleccionar amb les fletxes de dalt a l'esquerra. Separant
diferents seccions del dia en colors, et pots fixar en els diferents pics o
davallades que hi ha. A la cantonada superior de la dreta hi ha l'opció de
seleccionar el dia o l'estació que vols al requadre petit anomenat
'Arribades i Sortides'. D'aquesta forma visualitzes a quines hores hi ha
més activitat o a quines menys. En el cas de que no trobis l'estació que
desitjes desplegant la pestanya de selecció, també pots buscar-la escrivint
el seu nom.")
               )
      ),
      tabItem("submenu1",
              fluidRow(
                box(title= "Mapa Entre-Diari dels diferents perfils
d'estacions segons Clusterització ", leafletOutput("mapaperf1"),
status="primary", solidHeader = TRUE),
                tabBox(title="Visualització dels diferents perfils
d'estacions segons Parts del Dia", id="tabset2",
                       tabPanel("Cluster 1", leafletOutput("Perf1")),
                       tabPanel("Cluster 2", leafletOutput("Perf2")),
                       tabPanel("Cluster 3", leafletOutput("Perf3")),
                       tabPanel("Cluster 4", leafletOutput("Perf4"))),
              ),
              fluidRow(
                box(title=HTML('<div><i class="fa fa-info-circle"
style="font-size: 24px;"></i> PUNT DE INFORMACIÓ</div>'), width=12,
status="info",solidHeader=TRUE,
                     "En aquests mapes pots observar els perfils d'estacions
que hi ha segons el mètode estadístic de Clustering*. Aquest consisteix en
agrupar dades similars en diferents conjunts anomenats clusters* per
identificar patrons i estructures sense necessitat de coneixement previ. En
el mapa de l'esquerra estan tots els clusters agrupats per una comparació
conjunta mentre que el mapa de la dreta està focalitzat a la visualització
individual de cada perfil d'estació. Aquestes agrupacions han estat fetes
tenint en compte les diferents parts del dia: Matí, Migdia, Tarda i Nit.")
              )),
      tabItem("submenu2",
              fluidRow(
                box(title= "Mapa Diari dels diferents perfils d'estacions
segons Clusterització", leafletOutput("mapaperf2"), status="primary",
solidHeader = TRUE),
                tabBox(title="Visualització dels diferents perfils
d'estacions segons Dies", id="tabset2",
                       tabPanel("Cluster 1", leafletOutput("Perf5")),
                       tabPanel("Cluster 2", leafletOutput("Perf6")),
                       tabPanel("Cluster 3", leafletOutput("Perf7")),
                       tabPanel("Cluster 4", leafletOutput("Perf8"))),
              ),
              fluidRow(
                box(title=HTML('<div><i class="fa fa-info-circle"
style="font-size: 24px;"></i> PUNT DE INFORMACIÓ</div>'), width=
12,status="info",solidHeader=TRUE,
                    "En aquests mapes pots observar els perfils d'estacions
que hi ha segons el mètode estadístic de Clustering. Aquest consisteix en
agrupar dades similars en diferents conjunts anomenats clusters* per
identificar patrons i estructures sense necessitat de coneixement previ. En
el mapa de l'esquerra estan tots els clusters agrupats per una comparació
conjunta mentre que el mapa de la dreta està focalitzat a la visualització
individual de cada perfil d'estació. Aquestes agrupacions han estat fetes
```

```
tenint en compte els diferents dies de la setmana, de dilluns a diumenge
del 07/01/2019 al 13/01/2019 respectivament.")
                )),
      tabItem("submenu3",
              fluidRow(
                box(title=HTML("Mapa Interactiu dels punts mitjans de cada
perfil segons <b>Parts del Dia</b>"), leafletOutput("MitjPart"),
solidHeader = TRUE, status ="primary"),
                box(title=HTML("Mapa Interactiu dels punts mitjans de cada
perfil segons <b>Dia</b>"), leafletOutput("MitjDia"), solidHeader = TRUE,
status ="primary")
              ),
              fluidRow(
                box(title=HTML('<div><i class="fa fa-info-circle"
style="font-size: 24px;"></i> PUNT DE INFORMACIÓ</div>'), width=
12,status="info",solidHeader=TRUE,
                    "En aquests mapes pots observar els punts cèntrics de
cada grup que s'ha creat mitjançant Clustering. El mapa de l'esquerra es
refereix a la tècnica utilitzada per les estacions segons diferents parts
del dia mentre que el de la dreta és dels diferents dies de la setmana. Si
cliques sobre els punts veuràs a quin perfil pertany. ")
              )),
      tabItem("submenu4",
              fluidRow(
                box(title=HTML("Scatter Plot dels Components segons
Clusters per els <b>Dies</b>"),plotOutput("Clust1"), status="primary",
solidHeader=TRUE),
                tabBox(title="Visualització Box plots", id="tabset5",
                       tabPanel("Primera Component", plotOutput("Clust2")),
                       tabPanel("Segona Component",
plotOutput("Clust3")))),
              fluidRow(
                box(title=HTML('<div><i class="fa fa-info-circle"
style="font-size: 24px;"></i> PUNT DE INFORMACIÓ</div>'), width=
12,status="info",solidHeader=TRUE,
                    "En aquest apartat veiem els gràfics com Scatter Plots*
o Box Plots* de l'anàlisi de les dades segons Dies. El gràfic de
l'esquerra, mostra com les dades es dispersen en dues dimensions*, que
representen les components d'una anàlisi PCA*. Els punts estan pintats en
colors corresponents als clústers amb una llegenda a la part superior a la
dreta que identifica els diferents perfils. En el gràfic de la dreta o
també anomenat Box plot, mostra com varien els valors de les components
segons el grup. Això està representat per les barres en el gràfic. Els
eixos mostren els clusters i les components, i la caixa indica la dispersió
dins de cada grup. Els valors de les components són valors que descriuen
dades. En la primera gràfica, els punts són la seva ubicació en dues
dimensions, i en la segona mostra com les dades es distribueixen dins de
grups. Això ajuda a entendre patrons i diferències en les dades.")
              )
              ),
      tabItem("submenu5",
              fluidRow(
                tabBox(title=HTML("Scatter Plot dels Components segons
Clusters per <b>Parts del Dia</b>"), id="tabset6",
                       tabPanel("1-2 Component", plotOutput("Clust4")),
                       tabPanel("1-3 Component", plotOutput("Clust5")),
                       tabPanel("2-3 Component", plotOutput("Clust6"))),
                tabBox(title="Visualització Box plots", id="tabset7",
                       tabPanel("Primera Component", plotOutput("Clust7")),
                       tabPanel("Segona Component", plotOutput("Clust8")),
                       tabPanel("Tercera Component",
plotOutput("Clust9")))),
              fluidRow(
                box(title=HTML('<div><i class="fa fa-info-circle"
style="font-size: 24px;"></i> PUNT DE INFORMACIÓ</div>'), width=
12,status="info",solidHeader=TRUE,
                    "En aquest apartat veiem els gràfics com Scatter Plots*
o Box plots* de l'anàlisi de les dades segons les Parts dels Dies. En el
gràfic de l'esquerra, mostra com les dades es dispersen en dues dimensions,
que representen les components d'una anàlisi PCA. Els punts estan pintats
```

```
en colors corresponents als clústers amb una llegenda a la part superior a
la dreta que identifica els diferents perfils. En el gràfic de la dreta o
també anomenat Box plot, mostra com varien els valors de les components
segons el grup. Això està representat per les barres en el gràfic. Els
eixos mostren els clusters i les components, i la caixa indica la dispersió
dins de cada grup. Els valors de les components són valors que descriuen
dades. En la primera gràfica, els punts són la seva ubicació en dues
dimensions, i en la segona mostra com les dades es distribueixen dins de
grups. Això ajuda a entendre patrons i diferències en les dades.")
                )
                ),
        tabItem("tab7",
                box(title=HTML('<div><i class="fa fa-book" style="font-
size: 25px;"></i> GLOSSARI</div>'), status= "info",solidHeader = TRUE,
width=12,HTML("<h4 style='font-size: 16px;'>A continuació trobaràs un petit
recull de definicions o explicacions de termes concrets que surten a
l'aplicació de Flux Bicing BCN. Veuràs aquí tots aquells conceptes que en
la resta dels apartats apareguin amb un asterisc (*). L'objectiu d'aquest
apartat és facilitar la comprensió d'alguns termes que per algú de fora
l'Estadística no resultin tan fàcils d'entendre.</h4>")),
                useShinyjs(),
                fluidRow(box(
      status = "primary",
      solidHeader = TRUE,
      actionButton("ClusterING", "Clustering"),
      div(id = "InfoClustering", style = "display: none;",
          p("El clustering, o agrupament, implica organitzar dades en grups
basats en similituds. Cada grup és anomenat 'cluster' i conté elements amb
característiques similars. Aquest mètode divideix un conjunt de dades en
subconjunts coherents, ajudant a identificar patrons i categories ocultes.
Les dimensions d'aquests clusters són les característiques que utilitzem
per agrupar. Així, podem comprendre millor les relacions entre les dades i
extreure conclusions significatives.")
      )
    ),
                box(
      status = "primary",
      solidHeader = TRUE,
      actionButton("Cluster", "Cluster"),
      div(id = "InfoCluster", style = "display: none;",
          p("A través del procés de Clustering, els clústers són
agrupacions de dades similars que provenen de les característiques que
comparteixen. Són útils per trobar patrons i grups en conjunts de dades
grans, el qual ens permet entendre millor les relacions i tendències dins
de les dades. ")
      )
    )),fluidRow(
                box(
      status = "primary",
      solidHeader = TRUE,
      actionButton("ScatterPlot", "Scatter Plot"),
      div(id = "InfoScatter", style = "display: none;",
          p("Un scatterplot, també conegut com a gràfic de dispersió, és
una representació visual de punts en un pla. Cada punt representa una
observació i és posicionat en base a dues variables. Aquest tipus de gràfic
és utilitzat per mostrar la relació entre dues variables i identificar
patrons o tendències. Els punts poden estar dispersos, agrupats o seguir
una línia, depenent de la naturalesa de les dades.")
      )
    ),
                box(
      status = "primary",
      solidHeader = TRUE,
      actionButton("BoxPlot", "Box Plot"),
      div(id = "InfoBox", style = "display: none;",
          p("Un boxplot, també conegut com a diagrama de caixa i bigotis,
és una representació gràfica que mostra la distribució de les dades en un
conjunt. Consisteix en una caixa amb una línia al mig i 'bigotis' que es
projecten cap a fora. La caixa representa el rang interquartil (25% a 75%
de les dades), la línia al mig és la mediana i els bigotis indiquen la
```

```
dispersió. Punts fora dels bigotis poden ser considerats valors atípics. El
box plot és utilitzat per veure la simetria, dispersió i valors extrems
d'una variable. ")
      )
    )),fluidRow(
    #Eix Abscisses, Eix Ordenades, Histograma, Dimensió
    box(
      status = "primary",
      solidHeader = TRUE,
      actionButton("Abscisses", "Eix Abscisses"),
      div(id = "InfoAbs", style = "display: none;",
          p("L'eix d'abscisses, també conegut com a eix horitzontal, és la
línia horitzontal en un gràfic que representa una variable. En un
scatterplot, l'eix d'abscisses mostra una de les dues variables que es
comparen. En un boxplot, l'eix d'abscisses no és tan rellevant com en
altres tipus de gràfics, ja que principalment es concentra en mostrar les
distribucions i estadístiques de les dades verticalment.")
      )
    ),
    box(
      status = "primary",
      solidHeader = TRUE,
      actionButton("Ordenades", "Eix Ordenades"),
      div(id = "InfoOrd", style = "display: none;",
          p("L'eix d'ordenades, també conegut com a eix vertical, és la
línia vertical en un gràfic que representa una variable. En un scatterplot,
l'eix d'ordenades mostra l'altra variable que es compara. En un box plot,
l'eix d'ordenades és essencial per mostrar les distribucions de les dades i
les estadístiques com la mediana i el rang interquartil. És clau per
entendre la variabilitat i el comportament vertical de les dades")
      )
    )),fluidRow(
    box(
      status = "primary",
      solidHeader = TRUE,
      actionButton("Histograma", "Histograma"),
      div(id = "InfoHist", style = "display: none;",
          p("L'histograma és una representació gràfica de la distribució de
les dades en un conjunt. L'eix d'abscisses, o eix horitzontal, mostra els
intervals o categories de valors, mentre que l'eix d'ordenades, o eix
vertical, indica la freqüència o proporció d'observacions dins de cada
interval. L'histograma permet veure com les dades estan distribuïdes i
identificar patrons comuns o inusuals. És útil per explorar la forma i la
tendència de les dades i és ampliament utilitzat en estadística i l'anàlisi
de dades.")
      )
    ),
    box(
      status = "primary",
      solidHeader = TRUE,
      actionButton("Dimensio", "Dimensió"),
      div(id = "InfoDim", style = "display: none;",
          p("Les dimensions que sorgeixen en mètodes com PCA són noves
perspectives creades a partir de les variables originals. En PCA, aquestes
dimensions s'anomenen 'components principals'. Les dimensions creades
capturen diferents nivells de variabilitat o patrons en les dades
originals. Són utilitzades per reduir complexitat, destacar
característiques clau i facilitar la visualització i l'anàlisi de dades en
una forma més comprensible.")
      )
    )), fluidRow(box(
      status = "primary",
      solidHeader = TRUE,
      actionButton("PCA", "PCA"),
      div(id = "InfoPCA", style = "display: none;",
          p("L'Anàlisi de Components Principals (PCA) és una eina que pren
moltes dades complexes i les simplifica en noves formes per tal que puguem
veure els patrons importants i entendre millor les relacions entre elles.
Identifica patrons i varietat en un conjunt de dades, convertint-les en
noves dimensions anomenades 'Components Principals'.  Aquestes noves
```

```
dimensions són combinacions lineals de les variables originals i capturen
la major part de la variabilitat. Les components principals són útils per
visualitzar dades en menys dimensions, reduir el soroll (eliminar
distraccions) i destacar patrons ocults.")
      )
    )))))
      )
```

- SHINY SERVER

```
server <- function(input, output) {

  selected_data1 <- reactive({
    fitxer_char %>%
      filter(dia == input$dia, nom_numero == input$estacio1) %>%
      select(timestamp, bikes)
  })

  output$bike_plot <- renderPlot({
    ggplot(selected_data1(), aes(x = timestamp, y = bikes)) +
      geom_line() +
      labs(x = "HORES", y = "BICIS DISPONIBLES") +
      ggtitle(paste0("Flux de Bicis per l'estació de ", input$estacio1, "
del dia ", input$dia)) +
      scale_x_datetime(date_labels = "%H:%M")
  })


  #ARRIBADES

       dades_estacio1 <- reactive({
     cleandata_hores %>%
    filter(dia == input$day, nom_numero == input$estacio3) %>%
    select(hores, dia, arribades, part_dia)
   })

      dades_estacio2 <- reactive({
     cleandata_hores %>%
    filter(dia == input$day, nom_numero == input$estacio3) %>%
    select(hores, dia, sortides, part_dia)
   })

   maxim <- reactive({
  max(c(dades_estacio1()$arribades,dades_estacio2()$sortides))
})

  output$bike_plot3<- renderPlot({
     ggplot(dades_estacio1(), aes(x = hores, y = arribades, fill =
part_dia)) +
  geom_bar(stat = "identity", position = "dodge") +

  labs(title =paste("Arribades a l'estació",input$estacio3,"del
dia",input$day, "segons part del dia "),x = "HORA DEL DIA", y =
"ARRIBADES") +
  scale_fill_manual(values = c("red", "green", "blue", "purple"),
                    labels = c("MATI", "MIGDIA", "NIT", "TARDA")) +
  scale_y_continuous(limits=c(0,maxim())) +
  scale_x_continuous(breaks = 0:23) +
  theme_bw()+
  theme(legend.text = element_text(size = 12))+
  guides(fill = guide_legend(title = NULL))
  })

  #SORTIDES

  output$bike_plot4<- renderPlot({
```

```
     ggplot(dades_estacio2(), aes(x = hores, y = -(sortides), fill =
part_dia)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title =paste("Sortides de l'estació",input$estacio3,"del
dia",input$day, "segons part del dia "),x = "HORA DEL DIA", y = "SORTIDES")
+
  scale_fill_manual(values = c("red", "green", "blue", "purple"),
                    labels = c("MATI", "MIGDIA", "NIT", "TARDA")) +
  scale_y_continuous(limits=c(0,maxim())) +
  scale_x_continuous(breaks = 0:23) +
  theme_bw() +
  theme(legend.text = element_text(size = 12))+
  guides(fill = guide_legend(title = NULL))
  })

  selected_data2 <- reactive({
    fitxer_char %>%
      filter(nom_numero == input$estacio2) %>%
      select(dia, timestamp, bikes) %>%
      arrange(dia, timestamp)
  })

  output$bike_plot2 <- renderPlot({
    ggplot(selected_data2(), aes(x = timestamp, y = bikes, color = dia)) +
      geom_line() +
      labs(x = "DIES", y = "BICIS DISPONIBLES") +
      ggtitle(paste0("Flux de Bicis per l'estació de ", input$estacio2))+
      scale_x_datetime(date_labels = "%d %b", date_breaks = "1 day")+
      theme(
      axis.title = element_text(size = 14),
      axis.text = element_text(size = 12)
    )
  })

  color_palette <- colorFactor(c("red", "black", "green"), levels =
c("red", "black", "green"))

  #MAPA DIFERENCIES ARRIBADES I SORTIDES
  output$mapadif<- renderLeaflet({
  leaflet() %>%
  addTiles()%>%
  addCircleMarkers(data=cleandata_hores[cleandata_hores$dia==input$DIA &
cleandata_hores$hores==input$HORA,],
             lng= ~longitude, lat= ~latitude,
             color= ~color_palette(ifelse(dif > 0, "green", ifelse(dif ==
0, "black", "red"))),
             popup= ~paste("<h3> Estació Bicing </h3>","<b> Adreça: </b>",
nom_numero, "<br>", "<b> Diferència: </b>", dif))
})
#  MAPA DIFERENTS PERFILS D'ESTACIO <PART DIA>

  cluster_colors <- c("#ff0000", "#00ff00", "#0000ff", "#ffa500")
  output$mapaperf1 <- renderLeaflet({
    leaflet(partdia) %>%
  addTiles() %>%
  addCircleMarkers(
    lng = ~longitude,
    lat = ~latitude,
    color = ~color,
    radius = 5,
    opacity = 1,
    fillOpacity = 1,
    label = ~id
  ) %>%
  addLegend(
    position = "topright",
    colors = cluster_colors,
    labels = paste("Cluster", 1:length(cluster_colors)),
    title = "Clusters"
  )
```

```
  })

  #PERFILS DIFERENCIATS <PART DIA>
  output$Perf1<- renderLeaflet({
    leaflet(partdia[partdia$Clust==1,]) %>%
  addTiles() %>%
  addCircleMarkers(
    lng = ~longitude,
    lat = ~latitude,
    color = ~color,
    radius = 5,
    opacity = 1,
    fillOpacity = 1
  )
  })
   output$Perf2<- renderLeaflet({
    leaflet(data=partdia[partdia$Clust==2,]) %>%
  addTiles() %>%
  addCircleMarkers(
    lng = ~longitude,
    lat = ~latitude,
    color = ~color,
    radius = 5,
    opacity = 1,
    fillOpacity = 1
  )
  })
    output$Perf3<- renderLeaflet({
    leaflet(data=partdia[partdia$Clust==3,]) %>%
  addTiles() %>%
  addCircleMarkers(
    lng = ~longitude,
    lat = ~latitude,
    color = ~color,
    radius = 5,
    opacity = 1,
    fillOpacity = 1
  )
  })
     output$Perf4<- renderLeaflet({
    leaflet(data=partdia[partdia$Clust==4,]) %>%
  addTiles() %>%
  addCircleMarkers(
    lng = ~longitude,
    lat = ~latitude,
    color = ~color,
    radius = 5,
    opacity = 1,
    fillOpacity = 1
  )
  })

#  MAPA DIFERENTS PERFILS D'ESTACIO <DIARI>
  output$mapaperf2 <- renderLeaflet({
    leaflet(df_Dia) %>%
  addTiles() %>%
  addCircleMarkers(
    lng = ~longitude,
    lat = ~latitude,
    color = ~color,
    radius = 5,
    opacity = 1,
    fillOpacity = 1,
    label = ~id
  ) %>%
  addLegend(
    position = "topright",
    colors = cluster_colors,
    labels = paste("Cluster", 1:length(cluster_colors)),
    title = "Clusters"
```

```
)
})

#PERFILS DIFERENCIATS <DIARI>
output$Perf5<- renderLeaflet({
  leaflet(df_Dia[df_Dia$Clust==1,]) %>%
addTiles() %>%
addCircleMarkers(
  lng = ~longitude,
  lat = ~latitude,
  color = ~color,
  radius = 5,
  opacity = 1,
  fillOpacity = 1
)
})
 output$Perf6<- renderLeaflet({
  leaflet(data=df_Dia[df_Dia$Clust==2,]) %>%
addTiles() %>%
addCircleMarkers(
  lng = ~longitude,
  lat = ~latitude,
  color = ~color,
  radius = 5,
  opacity = 1,
  fillOpacity = 1
)
})
  output$Perf7<- renderLeaflet({
  leaflet(data=df_Dia[df_Dia$Clust==3,]) %>%
addTiles() %>%
addCircleMarkers(
  lng = ~longitude,
  lat = ~latitude,
  color = ~color,
  radius = 5,
  opacity = 1,
  fillOpacity = 1
)
})
   output$Perf8<- renderLeaflet({
  leaflet(data=df_Dia[df_Dia$Clust==4,]) %>%
addTiles() %>%
addCircleMarkers(
  lng = ~longitude,
  lat = ~latitude,
  color = ~color,
  radius = 5,
  opacity = 1,
  fillOpacity = 1
)
})

   #MAPAS PUNTS MITJS DIA
   dta<- df_Dia %>%
group_by(Clust) %>%
summarize(n_obs = n(),
          mean_long= mean(longitude),
          mean_lat= mean(latitude))

   output$MitjDia<- renderLeaflet({
   leaflet() %>%
addTiles() %>%
addCircleMarkers(data=dta,
           lng= ~mean_long, lat= ~mean_lat,
           popup= ~paste("<b> Cluster: </b>", Clust))
   })

   #MAPAS PUNTS MITJS PARTDIA
   dtapart<- partdia %>%
```

```
  group_by(Clust) %>%
  summarize(n_obs = n(),
            mean_long= mean(longitude),
            mean_lat= mean(latitude))

     output$MitjPart<- renderLeaflet({
       leaflet() %>%
  addTiles() %>%
  addCircleMarkers(data=dtapart,
             lng= ~mean_long, lat= ~mean_lat,
             popup= ~paste("<b> Cluster: </b>", Clust))
     })

     #CLUSTERS AVANÇATS DIA
 output$Clust1<- renderPlot({
   plot(PSI_Dia[,1],PSI_Dia[,2],col=cdia$x,main="Visualització per colors
dels Clusters de Primera i Segona Component",pch =16,xlab="PRIMERA
COMPONENT",ylab="SEGONA COMPONENT")
abline(h = 0, col = "black")
abline(v = 0, col = "black")
legend("topright",c("CLUSTER 1", "CLUSTER 2", "CLUSTER 3", "CLUSTER
4"),pch=16,col=c(1:4))
 })
 output$Clust2<- renderPlot({
   #Primera Component
boxplot(PSI_Dia[,1]~cdia$x,main="Box plot Primera Component segons Dies",
ylab="PRIMERA COMPONENT", xlab="CLUSTERS")
 })
output$Clust3<- renderPlot({
   #Segona Component
boxplot(PSI_Dia[,2]~cdia$x,main="Box plot Segona Component segons
Dies",ylab="SEGONA COMPONENT", xlab="CLUSTERS")
 })

#CLUSTERS AVANÇATS PART DIA
output$Clust4<- renderPlot({
  plot(PSI_Part[,1],PSI_Part[,2],col=cpart$x,main="Visualització per colors
dels Clusters de Primera i Segona Component",pch =16, xlab="PRIMERA
COMPONENT",ylab="SEGONA COMPONENT")
abline(h = 0, col = "black")
abline(v = 0, col = "black")
legend("topright",c("CLUSTER 1", "CLUSTER 2", "CLUSTER 3", "CLUSTER
4"),pch=16,col=c(1:4))
})
output$Clust5<- renderPlot({
  plot(PSI_Part[,1],PSI_Part[,3],col=cpart$x,main="Visualització per colors
dels Clusters de Primera i Tercera Component",pch =16,xlab="PRIMERA
COMPONENT",ylab="TERCERA COMPONENT")
abline(h = 0, col = "black")
abline(v = 0, col = "black")
legend("topright",c("CLUSTER 1", "CLUSTER 2", "CLUSTER 3", "CLUSTER
4"),pch=16,col=c(1:4))
})
output$Clust6<- renderPlot({
  plot(PSI_Part[,2],PSI_Part[,3],col=cpart$x,main="Visualització per colors
dels Clusters de Segona i Tercera Component",pch =16,xlab="SEGONA
COMPONENT",ylab="TERCERA COMPONENT")
abline(h = 0, col = "black")
abline(v = 0, col = "black")
legend("topright",c("CLUSTER 1", "CLUSTER 2", "CLUSTER 3", "CLUSTER
4"),pch=16,col=c(1:4))
})
output$Clust7<- renderPlot({
  #Primera Component
boxplot(PSI_Part[,1]~cpart$x,main="Box plot Clusters Primera Component
segons Parts dels Dies", ylab= "PRIMERA COMPONENT", xlab="CLUSTERS")
})
output$Clust8<- renderPlot({
  #Segona Component
```

```
boxplot(PSI_Part[,2]~cpart$x,main="Box plot Clusters Segona Component
segons Parts dels Dies", ylab= "SEGONA COMPONENT", xlab="CLUSTERS")
})
output$Clust9<- renderPlot({
  #Tercera Component
boxplot(PSI_Part[,3]~cpart$x,main="Box plot Clusters Tercera Component
segons Parts dels Dies", ylab= "TERCERA COMPONENT", xlab="CLUSTERS")
})

observeEvent(input$PCA, {
    shinyjs::toggle("InfoPCA")
  })
observeEvent(input$Cluster, {
    shinyjs::toggle("InfoCluster")
  })
observeEvent(input$ScatterPlot, {
    shinyjs::toggle("InfoScatter")
  })
observeEvent(input$BoxPlot, {
    shinyjs::toggle("InfoBox")
  })
observeEvent(input$Dimensio, {
    shinyjs::toggle("InfoDim")
  })
observeEvent(input$Ordenades, {
    shinyjs::toggle("InfoOrd")
  })
observeEvent(input$Abscisses, {
    shinyjs::toggle("InfoAbs")
  })
observeEvent(input$Histograma, {
    shinyjs::toggle("InfoHist")
  })
observeEvent(input$ClusterING, {
    shinyjs::toggle("InfoClustering")
  })
}

runApp(list(ui=ui, server=server))
```

- PCA, CLUSTERING I PROFILING

```
library(shinydashboard)
library(shiny)
library(ggplot2)
library(dplyr)
library(leaflet)
library(sp)
library(lubridate)
library(car)
library(cluster)
library(factoextra)
library(plotly)

fitxer <- read.csv("D:/Universitat/Quart de
carrera/TFG/Important/fitxer.csv")

data <- read.csv("D:/Universitat/Quart de
carrera/TFG/Important/prova2.txt")
data <- data[complete.cases(data),]

data$latitude<- as.numeric(data$latitude)
data$longitude<- as.numeric(data$longitude)

# Combinacio dades bicing amb dades estació per obtenir noms i números de
#carrers de estació
```

```
fitxer <- fitxer %>%
  left_join(data %>% select(id, streetName, streetNumber),
            by = "id")
nomnumero<- paste(fitxer$streetName, fitxer$streetNumber, sep=", ")
fitxer$nom_numero<- nomnumero
fitxer<- fitxer[,c(-13,-14)]

#Combinacio dades de bicing amb les dades de l'estació per obtenir latitud
#i longitud
fitxer <- fitxer %>%
  left_join(data %>% select(id, latitude, longitude),
            by = "id")

#Busquem si tenim NA's i on estan

# Cerquem quines files tenen NA a la columna "x".
rows_with_nas <- is.na(fitxer$longitude)

# Fem un subset al marc de dades per incloure només les files amb NA a la
#columna "x".
print(fitxer[rows_with_nas, ])

#Treiem NA's

cleandata<- na.omit(fitxer)

## Extraccio i Anàlisis Arribades/Sortides

#CREACIO DATAFRAME ORDENAT SEGONS ID

cleandata_order <- cleandata %>%
  arrange(id)

library(dplyr)
library(lubridate)

cleandata_order <- cleandata_order %>%
  group_by(id) %>%
  mutate(diff_col = c(0, diff(bikes))) %>%
  ungroup()

#Extraccio nombre arribades i sortides cada hora:

#Agrupem les dades per estació i hora, i sumem les arribades (diferències
#positives) i les sortides (diferències negatives) per a cada hora.

cleandata_hores <- cleandata_order %>%
  group_by(id, dia, hores) %>%
  summarise(arribades = sum(diff_col[diff_col > 0]),
            sortides = sum(diff_col[diff_col < 0]), .groups= "drop")

cleandata_hores$dif<- cleandata_hores$arribades + cleandata_hores$sortides

#Afegim longitud i latitud a Cleandata_hores

cleandata_hores <- cleandata_hores %>%
  left_join(select(data, id, latitude, longitude), by = "id")

# Creacio Divisió horaria

# Divisió horària en intervals de 6 hores
intervals <- cut(cleandata$hores, breaks = c(0, 4, 10, 16, 22, 24), right =
FALSE,
                 labels = c("Nit", "Mati", "Migdia", "Tarda", "Nit"))
intervals2<- cut(cleandata_hores$hores, breaks = c(0, 4, 10, 16, 22, 24),
right = FALSE,
                 labels = c( "Nit","Mati", "Migdia", "Tarda", "Nit"))
intervals3<- cut(cleandata_order$hores, breaks = c(0, 4, 10, 16, 22, 24),
right = FALSE,
```

```
                labels = c( "Nit","Mati", "Migdia", "Tarda", "Nit"))

# Afegim nova columna amb la divisió horària
cleandata$periode_dia <- as.character(intervals)
cleandata_hores$part_dia <- as.character(intervals2)
cleandata_order$periode_dia<- as.character(intervals3)

table(cleandata$periode_dia)

## Analisis Periodes Diaris

# Creem el gràfic de barres per arribades i sortides segons el dia

Graf1<- function(estacio, data){
  dades_estacio <- cleandata_hores %>%
    filter(id == estacio) %>%
    select(hores, dia, arribades, part_dia)

(ggplot(dades_estacio[dades_estacio$dia==data,], aes(x = hores, y =
arribades, fill = part_dia)) +
  geom_bar(stat = "identity", position = "dodge") +
    annotate(geom = "text", x = 23, y = max(dades_estacio$arribades),
             label = paste0("Max: ", (max(dades_estacio$arribades))-1), vjust
= 1) +
  labs(title =paste("Arribades a l'estacio",estacio,"del dia",data, "segons
part del dia "),x = "Hora del dia", y = "Diferència mitjana d'arribades i
sortides") +
  scale_fill_manual(values = c("red", "green", "blue", "purple"),
                    labels = c("Matí", "Migdia", "Nit", "Tarda")) +
  scale_x_continuous(breaks = 0:23) +
  theme_bw())

}

Graf2<- function(estacio, data){
 dades_estacio <- cleandata_hores %>%
    filter(id == estacio) %>%
    select(hores, dia, sortides, part_dia)

  (ggplot(dades_estacio[dades_estacio$dia==data,], aes(x = hores, y = -
(sortides), fill = part_dia)) +
  geom_bar(stat = "identity", position = "dodge") +
      annotate(geom = "text", x = 23, y = -(min(dades_estacio$sortides)),
             label = paste0("Max: ", (-(min(dades_estacio$sortides)))-1 ,),
vjust = 1) +
  labs(title =paste("Sortides de l'estacio",estacio,"del dia",data, "segons
part del dia "), x = "Hora del dia", y = "Diferència mitjana d'arribades i
sortides") +
  scale_fill_manual(values = c("red", "green", "blue", "purple"),
                    labels = c("Matí", "Migdia", "Nit", "Tarda")) +
  scale_x_continuous(breaks = 0:23) +
  theme_bw())

}
Graf1(1,7)
Graf2(1,7)

cleandata_hores<- read.csv("D:/Universitat/Quart de
carrera/TFG/Cleandata_hores1.csv", sep=";")

#PART DEL DIA
sumes <- aggregate(cbind(arribades, -(sortides)) ~ id + part_dia, data =
cleandata_hores, sum)

df_partdia <- reshape(sumes, idvar = "id", timevar = "part_dia", direction
= "wide")

names(df_partdia) <- sub("^.*\\.", "", names(df_partdia))
```

```
colnames(df_partdia)= c("id", "Mati_arribades", "Mati_sortides",
"Migdia_arribades", "Migdia_Sortides", "Tarda_arribades", "Tarda_sortides",
"Nit_arribades", "Nit_sortides")

#DIA
sumes <- aggregate(cbind(arribades, -(sortides)) ~ id + dia, data =
cleandata_hores, sum)

df_Dia <- reshape(sumes, idvar = "id", timevar = "dia", direction = "wide")

names(df_Dia) <- sub("^.*\\.", "", names(df_Dia))

colnames(df_Dia)= c("id", "Dilluns_Arribades", "Dilluns_Sortides",
"Dimarts_Arribades", "Dimarts_Sortides", "Dimecres_Arribades",
"Dimecres_Sortides", "Dijous_Arribades", "Dijous_Sortides",
"Divendres_Arribades", "Divendres_Sortides", "Dissabte_Arribades",
"Dissabte_Sortides", "Diumenge_Arribades", "Diumenge_Sortides")

# Analisis segons Entre-Dia

dP<- df_partdia[,-c(1)]

#objects()
#attributes(dP)

## Visualització de les dades

attach(dP)

#R està entenent correctament les meves variables

sapply(dP,class)

#Fem una llista de les variables numèriques

numeriques<-which(sapply(dP,is.numeric))
numeriques

dP<-dP[,numeriques]

## ANALISIS COMPONENTS PRINCIPALS DE DP

pc1 <- prcomp(dP, scale=TRUE) # matriu de correlacions
print(pc1)

#QUIN PERCENTATGE DE LA INERCIA TOTAL ESTÀ REPRESENTAT EN SUBESPAIS?

inerProj<- pc1$sdev^2
totalIner<- sum(inerProj)
pinerEix<- 100*inerProj/totalIner
barplot(pinerEix, main="Percentatges de la Inèrcia total de les
Components")
abline(h = 90, col = "red", lty = 2)

### Inercia Acumulada en Subespais
#De la primera compnent fins la vuitena dimensió

barplot(100*cumsum(pc1$sdev[1:dim(dP)[2]]^2)/dim(dP)[2], main="Percentatge
Variabilitat Acumulada per Dimensions")
abline(h = 90, col = "red", lty = 2)

percInerAccum<-100*cumsum(pc1$sdev[1:dim(dP)[2]]^2)/dim(dP)[2]

#SELECCIÓ DIMENSIONS SIGNIFICATIVES

nd = 3

print(pc1)
attributes(pc1)
```

```
#EMMAGATZEMATGE EIGENVALUES, EIGENVECTORS I PROJECCIONS EN LES nd
DIMENSIONS

View(pc1$x)
dim(pc1$x)
dim(dP)
dP[400,]
pc1$x[400,]

Psi = pc1$x[,1:nd]
dim(Psi)
Psi[400,]

#CREACIÓ ETIQUETES PER INDIVIDUS I VARIABLES

iden = row.names(dP)
etiq = names(dP)
ze = rep(0,length(etiq))

### PLOT INDIVIDUS

eje1<-1
eje2<-2
eje3<-3

plot(Psi[,eje1],Psi[,eje2], type="n", main="Scatter Plot estacions per
Components 1 i 2", xlab="Component 1", ylab="Component 2")
text(Psi[,eje1],Psi[,eje2],labels=iden, cex=0.5)
axis(side=1, pos= 0, labels = F, col="cyan")
axis(side=3, pos= 0, labels = F, col="cyan")
axis(side=2, pos= 0, labels = F, col="cyan")
axis(side=4, pos= 0, labels = F, col="cyan")

plot(Psi[,eje1],Psi[,eje3], type="n", main="Scatter Plot estacions per
Components 1 i 3", xlab="Component 1", ylab="Component 3")
text(Psi[,eje1],Psi[,eje3],labels=iden, cex=0.5)
axis(side=1, pos= 0, labels = F, col="cyan")
axis(side=3, pos= 0, labels = F, col="cyan")
axis(side=2, pos= 0, labels = F, col="cyan")
axis(side=4, pos= 0, labels = F, col="cyan")

plot(Psi[,eje2],Psi[,eje3], type="n", main="Scatter Plot estacions per
Components 2 i 3", xlab="Component 2", ylab="Component 3")
text(Psi[,eje2],Psi[,eje3],labels=iden, cex=0.5)
axis(side=1, pos= 0, labels = F, col="cyan")
axis(side=3, pos= 0, labels = F, col="cyan")
axis(side=2, pos= 0, labels = F, col="cyan")
axis(side=4, pos= 0, labels = F, col="cyan")

### Projecció variables
Phi = cor(dP,Psi)
View(Phi)

X<-Phi[,eje1]
Y<-Phi[,eje2]
Z<-Phi[,eje3]

plot(Psi[,eje1],Psi[,eje2],type="n",xlim=c(-1.05,max(X,0)), ylim=c(-1,1),
main="Projecció variables Component 1 i 2", xlab="Component 1",
ylab="Component 2")
axis(side=1, pos= 0, labels = F)
axis(side=3, pos= 0, labels = F)
axis(side=2, pos= 0, labels = F)
axis(side=4, pos= 0, labels = F)
arrows(ze, ze, X, Y, length = 0.07,col="blue")
text(X,Y,labels=etiq,col="darkblue", cex=0.7)

plot(Psi[,eje1],Psi[,eje3],type="n",xlim=c(-1.05,max(X,0)), ylim=c(-1,1),
main="Projecció variables Component 1 i 3", xlab="Component 1",
ylab="Component 3")
```

```
axis(side=1, pos= 0, labels = F)
axis(side=3, pos= 0, labels = F)
axis(side=2, pos= 0, labels = F)
axis(side=4, pos= 0, labels = F)
arrows(ze, ze, X, Z, length = 0.07,col="blue")
text(X,Z,labels=etiq,col="darkblue", cex=0.7)

plot(Psi[,eje2],Psi[,eje3],type="n",xlim=c(min(X,0),1), ylim=c(-1,1),
main="Projecció variables Component 2 i 3", xlab="Component 2",
ylab="Component 3")
axis(side=1, pos= 0, labels = F)
axis(side=3, pos= 0, labels = F)
axis(side=2, pos= 0, labels = F)
axis(side=4, pos= 0, labels = F)
arrows(ze, ze, Y, Z, length = 0.07,col="blue")
text(Y,Z,labels=etiq,col="darkblue", cex=0.7)

## CLUSTERING ENTRE-DIA

dd<- Psi[,1:3]
#CLUSTERING JERARQUIC
library(cluster)

d  <- dist(dd)
h1 <- hclust(d,method="ward.D")
plot(h1)

#El coeficient de correlació cophenètic és una mesura que varia entre 0 i
1. Un valor més proper a 1 #indica una millor conservació de les relacions
de distància originals en la solució jeràrquica. En altres #paraules, un
valor més alt indica una millor qualitat del clustering jeràrquic en termes
de la seva #capacitat per representar les similituds o distàncies entre les
observacions originals.

coph_mat <- cophenetic(h1)

coph_corr <- cor(as.vector(d), as.vector(coph_mat))
cat("Coeficient de correlació cophenètic:", coph_corr)

k<-4
c2 <- cutree(h1,k)
#Mides grupals
table(c2)

#Recordatori

pc1$rotation

## Scatter Plot Clusters Entre-Dia
#PSI 1/2

c1<-c2
plot(Psi[,1],Psi[,2],col=c2,main="Visualització per colors dels Clusters
Primera i Segona Component",pch =16, xlab="Primera Component",ylab="Segona
Component")
abline(h = 0, col = "black")
abline(v = 0, col = "black")
legend("topright",c("CLUSTER 1", "CLUSTER 2", "CLUSTER 3", "CLUSTER
4"),pch=16,col=c(1:k))


#PSI 1/3

c1<-c2
plot(Psi[,1],Psi[,3],col=c1,main="Visualització per colors dels Clusters
Primera i Tercera Component",pch =16,xlab="Primera Component",ylab="Tercera
Component")
abline(h = 0, col = "black")
abline(v = 0, col = "black")
```

```
legend("topright",c("CLUSTER 1", "CLUSTER 2", "CLUSTER 3", "CLUSTER
4"),pch=16,col=c(1:k))



#PSI 2/3

c1<-c2
plot(Psi[,2],Psi[,3],col=c2,main="Visualització per colors dels Clusters
Segona i Tercera Component",pch =16,xlab="Segona Component",ylab="Tercera
Component")
abline(h = 0, col = "black")
abline(v = 0, col = "black")
legend("topright",c("CLUSTER 1", "CLUSTER 2", "CLUSTER 3", "CLUSTER
4"),pch=16,col=c(1:k))

## Box Plots Cluster Entre-Dies

#Primera Component
boxplot(PSI_Part[,1]~c2,main="Boxplot Clusters Primera Component segons
Parts dels Dies", ylab= "Primera Component", xlab="Clusters")

#Segona Component
boxplot(PSI_Part[,2]~c2,main="Boxplot Clusters Segona Component segons
Parts dels Dies", ylab= "Segona Component", xlab="Clusters")

#Tercera Component
boxplot(PSI_Part[,3]~c2,main="Boxplot Clusters Tercera Component segons
Parts dels Dies", ylab= "Tercera Component", xlab="Clusters")

cdg <- aggregate(as.data.frame(dd),list(c2),median)
cdg

#Unió Clusters amb dataframe

df_partdia$Clust<- c2

data <- read.csv("D:/Universitat/Quart de
carrera/TFG/Important/prova2.txt")
data <- data[complete.cases(data),]

data$latitude<- as.numeric(data$latitude)
data$longitude<- as.numeric(data$longitude)

library(dplyr)
df_partdia$longitude<- data$longitude
df_partdia$latitude<- data$latitude

dta<- df_partdia %>%
  group_by(Clust) %>%
  summarize(n_obs = n(),
            mean_long= mean(longitude),
            mean_lat= mean(latitude))

## Mapes Entre-Dia

#Mapa mitjana diferents clusters

mrap<- leaflet() %>%
  addTiles() %>%
  addCircleMarkers(data=dta,
             lng= ~mean_long, lat= ~mean_lat,
             popup= ~paste("<b> Cluster: </b>", Clust))

mrap

library(leaflet)
library(dplyr)

cluster_colors <- c("#ff0000", "#00ff00", "#0000ff", "#ffa500")
```

```
df_partdia <- df_partdia %>%
  mutate(color = cluster_colors[Clust])

map <- leaflet(df_partdia) %>%
  addTiles() %>%
  addCircleMarkers(
    lng = ~longitude,
    lat = ~latitude,
    color = ~color,
    radius = 5,
    opacity = 1,
    fillOpacity = 1,
    label = ~id
  ) %>%
  addLegend(
    position = "topright",
    colors = cluster_colors,
    labels = paste("Cluster", 1:length(cluster_colors)),
    title = "Clusters" )

map

mapet<- function(a){

  filtered_df <- df_partdia %>%
  filter(Clust == a)

map <- leaflet(filtered_df) %>%
  addTiles() %>%
  addCircleMarkers(
    lng = ~longitude,
    lat = ~latitude,
    color = ~color,
    radius = 5,
    opacity = 1,
    fillOpacity = 1
  )

map
}
mapet(1)
mapet(2)
mapet(3)
mapet(4)

------------------------------------------------------------------------
--------------------

# Analisis segons Dia

dD<- df_Dia[,-c(1)]

## Visualització Dades

#CREATION OF THE DATA FRAME OF CONTINUOUS VARIABLES

attach(dD)

#Està R entenent les meves variables?
sapply(dD,class)

#Creem una llista de les variables numèriques

numeriques<-which(sapply(dD,is.numeric))
numeriques

dD<-dD[,numeriques]
sapply(dD,class)
```

```
## Analisis Components Principals de dD

pc1 <- prcomp(dD, scale=TRUE)
print(pc1)

#Quin percentatge total de la inèrcia està representat en subespais

inerProj<- pc1$sdev^2
totalIner<- sum(inerProj)
pinerEix<- 100*inerProj/totalIner
barplot(pinerEix)

### Inèrcia Acumulada en subespais

#De la primera component a la catorzena dimensió del subespai

barplot(100*cumsum(pc1$sdev[1:dim(dD)[2]]^2)/dim(dD)[2], main="Percentatge
Variabilitat Acumulada per Dimensions")
abline(h = 90, col = "red", lty = 2)

### Selecció Dimensions Significatives

nd = 2

print(pc1)


### EMMAGATZEMATGE EIGENVALUES, EIGENVECTORS I PROJECCIONS EN LES nd
DIMENSIONS

dim(pc1$x)
dim(dD)
dD[400,]
pc1$x[400,]
Psi = pc1$x[,1:nd]

dim(Psi)
Psi[400,]

#Etiquetes per individus i variables

iden = row.names(dD)
etiq = names(dD)
ze = rep(0,length(etiq))

### Plots dels individus

eje1<-1
eje2<-2

plot(Psi[,eje1],Psi[,eje2], type="n", main="Scatter Plot estacions per
Components 1 i 2", xlab="Component 1", ylab="Component 2")
text(Psi[,eje1],Psi[,eje2],labels=iden, cex=0.5)
axis(side=1, pos= 0, labels = F, col="cyan")
axis(side=3, pos= 0, labels = F, col="cyan")
axis(side=2, pos= 0, labels = F, col="cyan")
axis(side=4, pos= 0, labels = F, col="cyan")

### Projecció variables

Phi = cor(dD,Psi)

X<-Phi[,eje1]
Y<-Phi[,eje2]


plot(Psi[,eje1],Psi[,eje2],type="n",xlim=c(-1,-0.8), ylim=c(-0.5, 0.5),
main="Projecció variables Component 1 i 2", xlab="Component 1",
ylab="Component 2")
axis(side=1, pos= 0, labels = F)
```

```
axis(side=3, pos= 0, labels = F)
axis(side=2, pos= 0, labels = F)
axis(side=4, pos= 0, labels = F)
arrows(ze, ze, X, Y, length = 0.07,col="blue")
text(X,Y,labels=etiq,col="darkblue", cex=0.7)

## CLUSTERING DIA

dd<- Psi[,1:2]

library(cluster)

# CLUSTERING JERÀRQUIC

d  <- dist(dd)
h1 <- hclust(d,method="ward.D")
plot(h1)

#El coeficient de correlació cophenètic és una mesura que varia entre 0 i
1. Un valor més proper a 1 #indica una millor conservació de les relacions
de distància originals en la solució jeràrquica. En altres #paraules, un
valor més alt indica una millor qualitat del clustering jeràrquic en termes
de la seva #capacitat per representar les similituds o distàncies entre les
observacions originals.

coph_mat <- cophenetic(h1)

coph_corr <- cor(as.vector(d), as.vector(coph_mat))
cat("Coeficient de correlació cophenètic:", coph_corr)

k<-4
c2 <- cutree(h1,k)
#Mides grupals
table(c2)

### Scatter Plots segons Cluster

c1<-c2
plot(Psi[,1],Psi[,2],col=cdia$x,main="Visualització per colors dels
Clusters Primera i Segona Component",pch =16,xlab="Primera
Component",ylab="Segona Component")
abline(h = 0, col = "black")
abline(v = 0, col = "black")
legend("topright",c("CLUSTER 1", "CLUSTER 2", "CLUSTER 3", "CLUSTER
4"),pch=16,col=c(1:k))

### Box Plots segons clusters

#Primera Component
boxplot(PSI_Dia[,1]~c2,main="Boxplot Primera Component segons Dies",
ylab="Primera Component", xlab="Clusters")

#Segona Component
boxplot(PSI_Dia[,2]~c2,main="Boxplot Segona Component segons
Dies",ylab="Segona Component", xlab="Clusters")

cdg <- aggregate(as.data.frame(dd),list(c2),mean)
cdg

#Unió Clusters amb dataframe

df_Dia$Clust<- c2

data <- read.csv("D:/Universitat/Quart de
carrera/TFG/Important/prova2.txt")
data <- data[complete.cases(data),]

data$latitude<- as.numeric(data$latitude)
data$longitude<- as.numeric(data$longitude)
```

```
library(dplyr)
df_Dia$longitude<- data$longitude
df_Dia$latitude<- data$latitude

dta<- df_Dia %>%
  group_by(Clust) %>%
  summarize(n_obs = n(),
            mean_long= mean(longitude),
            mean_lat= mean(latitude))



## MAPAS DIA

#Mapa mitjana diferents clusters

mrap<- leaflet() %>%
  addTiles() %>%
  addCircleMarkers(data=dta,
             lng= ~mean_long, lat= ~mean_lat,
             popup= ~paste("<b> Cluster: </b>", Clust))

mrap

library(leaflet)
library(dplyr)

cluster_colors <- c("#ff0000", "#00ff00", "#0000ff", "#ffa500")

df_Dia <- df_Dia %>%
  mutate(color = cluster_colors[Clust])

map <- leaflet(df_Dia) %>%
  addTiles() %>%
  addCircleMarkers(
    lng = ~longitude,
    lat = ~latitude,
    color = ~color,
    radius = 5,
    opacity = 1,
    fillOpacity = 1,
    label = ~id
  ) %>%
  addLegend(
    position = "topright",
    colors = cluster_colors,
    labels = paste("Cluster", 1:length(cluster_colors)),
    title = "Clusters"
  )

map

mapet<- function(a){
filtered_df <- df_Dia %>%
  filter(Clust == a)

map <- leaflet(filtered_df) %>%
  addTiles() %>%
  addCircleMarkers(
    lng = ~longitude,
    lat = ~latitude,
    color = ~color,
    radius = 5,
    opacity = 1,
    fillOpacity = 1
  )

map
}
mapet(1)
```

```
mapet(2)
mapet(3)
mapet(4)
```