



UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

Valenbisi: Un análisis sobre ruedas

PROYECTO II: Ciencia de Datos

Curso 2022/2023



Javier Luque, Daniel Garijo, Pablo Parrilla, José Valero y Qilu Diana Wu

Presentación

Este proyecto ha sido realizado por alumnos de la ETSINF (Escuela Técnica Superior de Ingeniería Informática) de la Universitat Politècnica de València. Además, se ha realizado en la asignatura de Proyecto II, del segundo cuatrimestre del segundo curso del Grado de “Ciencia de Datos”. A lo largo del cuatrimestre, nuestro equipo ha estado trabajando en este estudio acerca de Valenbisi. Este trabajo ha dado como resultado esta memoria con una serie de análisis elaborados que serán expuestos a lo largo de este documento.

El conjunto de datos utilizado en el proyecto ha sido el resultado de haber observado y modificado varios conjuntos de datos que hemos obtenido, lo cual se explicará más adelante.

Equipo

Este equipo está conformado por cinco integrantes: José Valero, Javier Luque, Daniel Garijo, Pablo Parrilla y Qilu Diana Wu, los cuales, como ya hemos mencionado antes, seremos los encargados de este proyecto acerca de las estaciones de Valenbisi, y de responder a las distintas cuestiones que se nos plantean.

Agradecimientos

Este proyecto no ha tenido mucha colaboración o apoyo por parte de entidades externas. Sin embargo, debemos agradecer a nuestras tutoras, Sara Blanc Clavero y María José Ramírez Quintana, el seguimiento, guía y ayuda que nos han brindado en numerosas ocasiones a lo largo del desarrollo del trabajo.

Además, hay que agradecer sin duda a la Universitat Politècnica de València, que nos ha ofrecido software (Microsoft Office 365), enseñanza y conexión a Internet, así como a los creadores de las páginas web que nos han ofrecido información y a los creadores de las bases de datos, pues han sido el motor principal de este proyecto y, sin ellos, este trabajo habría estado destinado al fracaso. Especial mención a César Ferri, cuyos datos han sido el pilar fundamental del proyecto.

Índice

1. Alcance del proyecto	5
1.1. Introducción al proyecto	5
1.2. Objetivos del proyecto	5
1.3. Utilidad del estudio	5
2. Configuración del proyecto	6
2.1. Fuentes de datos	6
2.2. Integración y transformación de datos	6
3. Resultados obtenidos	7
3.1. Objetivo 1. Análisis introductorio. Centro vs Periferia.	7
3.1.1. Preprocesado	7
3.1.2. Procedimiento	7
3.1.3. Discusión	10
3.2. Objetivo 2. Variables meteorológicas	11
3.2.1. Preprocesado	11
3.2.2. Procedimiento	11
3.2.3. Discusión	12
3.3. Objetivo 3. Agrupación de las estaciones según el día de la semana	14
3.3.1. Preprocesado	14
3.3.2. Procedimiento	14
3.3.3. Discusión	16
3.4. Objetivo 4: Lugares de interés y tramos semanales.	17
3.4.1. Preprocesado	17
3.4.2. Procedimiento	17
3.4.3. Discusión	18
3.5. Objetivo 5. Análisis predictivo	21
3.5.1. Preprocesado	21
3.5.2. Procedimiento	21
3.5.3. Discusión	22
4. Lecciones aprendidas	24
5. Anexos	25

Índice de figuras

1.	Evolución de la media de bicis y bornes en centro y periferia.	8
2.	Evolución de la media de bicis y bornes en centro y periferia (6-2-2022 / 7-2-2022).	8
3.	Cantidad media de bicis disponibles por día de la semana y tramo horario.	9
4.	Frecuencia de medias de variables meteorológicas.	11
5.	Frecuencia de medias de variables meteorológicas con movimiento.	12
6.	Gráfico de barras de bornes libres en prec.	12
7.	Gráfico de barras de variables sin movimiento.	13
8.	Gráfico de barras de variables con movimiento.	13
9.	Mapa de calor del lunes.	14
10.	Número óptimo de clusters con el método de Ward.	15
11.	Agrupación de las estaciones con k-medoides los lunes.	15
12.	Representante del perfil días laborables.	16
13.	Representante del perfil fines de semana.	16
14.	Vista previa de las estaciones con los clusters del lunes.	16
15.	Influencia de las estaciones con respecto de los clusters.	18
16.	Comportamiento de los clusters y su tabla de frecuencias correspondiente. Análisis general.	18
17.	Distribución geográfica de los clusters. Análisis general.	19
18.	Comportamiento en días laborables.	20
19.	Comportamiento en fines de semana.	20
20.	Distribución geográfica en días laborables.	20
21.	Distribución geográfica en fines de semana.	20
22.	Comparación de las bicicletas disponibles y predichas.	23

1 Alcance del proyecto

1.1. Introducción al proyecto

En el presente proyecto, *Valenbisi: un análisis sobre ruedas*, se ha llevado a cabo un exhaustivo análisis sobre el sistema de bicicletas compartidas de la ciudad de València. Este estudio ha buscado comprender los diferentes aspectos relacionados con el uso de este sistema de transporte sostenible.

En primer lugar, se ha realizado un análisis exploratorio de los datos recopilados. Este análisis ha consistido fundamentalmente en comparar cómo se comportan las estaciones centrales respecto a las periféricas.

Además, se ha tenido en cuenta la influencia de variables meteorológicas en el uso de las bicicletas de Valenbisi. Se han recopilado datos climáticos, como la temperatura, el viento y las precipitaciones, y se ha evaluado su relación con los patrones de uso de las bicicletas.

A continuación, se ha aplicado técnicas de clustering para agrupar las estaciones de Valenbisi en función de su similitud en términos de patrones de uso. Esto ha permitido identificar diferentes clusters de estaciones con características y necesidades particulares, lo cual puede ser utilizado para mejorar la planificación y distribución de las bicicletas en la ciudad. Asimismo, se ha considerado la incorporación de información sobre lugares de interés en la ciudad de València, como universidades, hospitales, playas, áreas comerciales...

Por último, se ha desarrollado un modelo predictivo que utiliza los datos recopilados para predecir la demanda futura de Valenbisi. Este modelo puede ser utilizado para planificar la distribución de bicicletas y optimizar la capacidad de las estaciones en función de la demanda esperada en diferentes momentos y estaciones de la ciudad.

1.2. Objetivos del proyecto

Los objetivos del proyecto, que se desarrollarán en la sección 3, son los siguientes:

- Realizar un primer análisis exploratorio que permita estudiar la relación entre las estaciones que se encuentran en barrios periféricos y las estaciones del centro de València.
- Observar el comportamiento de los usuarios que utilizan Valenbisi según variables meteorológicas como la temperatura, el viento y las precipitaciones.
- Analizar si existen agrupaciones de estaciones de Valenbisi en función del día de la semana.
- Analizar la demanda de bicicletas en las estaciones de Valenbisi teniendo en cuenta los lugares de interés cercanos a estas estaciones, valorando, además, si cae entre semana o en fines de semana
- Dada una estación y una hora, predecir el número de bicicletas disponibles.

1.3. Utilidad del estudio

Este proyecto resulta interesante y útil tanto para la empresa encargada de la distribución y recarga de bicicletas, como para los usuarios regulares de Valenbisi, ya que proporciona información sobre: las zonas con un comportamiento similar en relación a la utilización de las bicicletas; el número estimado de bicicletas disponibles en una estación y hora concretas; cómo se ve alterada la demanda de bicicletas según la climatología... . Todo esto puede permitir aplicar una planificación más efectiva y mejorar la experiencia de los usuarios de Valenbisi.

2 Configuración del proyecto

2.1. Fuentes de datos

A lo largo de este proyecto, se ha hecho uso de datos de diversas fuentes:

- Para la obtención de los datos de Valenbisi, se realizó en primer lugar una descarga masiva de ficheros .csv (aproximadamente 8.000 archivos) desde la página de GitHub¹ de César Ferri, profesor de la UPV. Mediante varios programas (archivos .py de Python que pueden ver consultados en el anexo [1]) se realizó la concatenación de todos estos ficheros.
- Para la obtención de datos metereológicos, se diseño un programa en Python (que puede ser consultado en el anexo [2]) que realiza peticiones a la API de AEMET².
- Para la obtención de datos espaciales de los barrios y distritos de la ciudad de València, recurrimos al portal de datos abiertos del ayuntamiento de dicha ciudad³.
- Los datos de variables como el día de la semana o el fichero .xlsx que clasifica las estaciones según su importancia se generaron manualmente.

Para más información sobre la obtención de los datos se puede consultar el anexo [3]

2.2. Integración y transformación de datos

Debido a que los datos se recopilan cada 15 minutos, la base de datos original presentaba una fila por estación, fecha y cuarto de hora. Con el objetivo de que no fuera tan extremadamente costoso computacionalmente, la primera tarea que se realizó fue la de agregar los datos por hora. De esta forma, cada fila representa una estación, fecha y hora concretos.

Posteriormente, a la nueva base de datos se le agregó la información metereológica, el día de la semana... Para esto, se utilizó como variable identificadora (y por tanto, la clave sobre la que se realiza la concatenación) la fecha.

Por otro lado, se agregó la información geométrica tanto de los barrios como de los distritos de la ciudad de València. Esta información (los polígonos que forman los barrios y los distritos) permitió determinar en qué barrio y en qué distrito se situa cada estación.

Para más información sobre la integración y transformación de datos se puede consultar los anexos [4] y [5].

¹<https://github.com/ceferra/valenbici>

²<https://opendata.aemet.es/centrodedescargas/inicio>

³<https://valencia.opendatasoft.com/pages/home/>

3 Resultados obtenidos

3.1. Objetivo 1. Análisis introductorio. Centro vs Periferia.

Para comenzar a buscar respuestas a los objetivos, primero hay que familiarizarse con el comportamiento de las estaciones de valenbisi. En este apartado en concreto, se estudiará la relación entre las estaciones que se encuentran en barrios periféricos y las estaciones del centro de València.

3.1.1. Preprocesado

Partiendo del dataframe “df_merge” obtenido en el tratamiento de los datos, en primer lugar se eliminan las estaciones 11, 17, 47, 113, 155, 166 y 232. Esto se debe a que durante los análisis, se ha encontrado que estas estaciones tienen varias observaciones en las que el número de bicicletas disponibles es 0 y el número de bornes disponibles también es 0; lo cual hace pensar que se trata de un error y podría afectar a los resultados. Si bien, habrá usos futuros del dataframe “df_merge” original del tratamiento en los que se incluyan estas estaciones.

A continuación, se seleccionan las observaciones de aquellas estaciones consideradas “a ojo” periféricas y céntricas. Así, se han seleccionado los barrios del centro de València y de la periferia y se han guardado los datos de estas estaciones en el dataframe “data2”. En el RMarkdown, que puede consultarse en el anexo [6], se puede visualizar cuales son todas estas estaciones seleccionadas.

Una vez se tienen todas las observaciones deseadas, se han creado dos nuevas variables “datetime_char” y “datetime”, que combinan los valores de “fecha” y “hora_hora”. La única diferencia entre estas dos nuevas variables es que la primera es tipo texto y la segunda contiene objetos POSIXct.

3.1.2. Procedimiento

A continuación, se puede comenzar a hacer las transformaciones necesarias para contestar a la pregunta de si hay relación o se diferencian las estaciones céntricas de las periféricas. Para empezar, se obtiene un dataframe en el que cada observación corresponde a un período de una hora; y se guardan los valores de la media de bornes libres y de bicis disponibles de todas las estaciones céntricas, y la media de bornes libres y de bicis disponibles de todas las estaciones periféricas. Dicho dataframe tiene la estructura de la tabla 1.

datetime	bicis_centro	bicis_periferia	bornes_centro	bornes_periferia
2022-12-02 00:00:00	2.602941	8.989362	17.29706	9.975177
2022-12-02 01:00:00	2.338235	9.138298	17.66176	9.826241
2022-12-02 02:00:00	1.941176	9.276596	18.05882	9.684397
...

Cuadro 1: Dataframe **datetimes**

A partir de este dataframe, en la Figura 1 se puede visualizar la evolución de los valores del centro y de la periferia a lo largo del tiempo. De especial interés es la evolución que tienen a lo largo de un día.

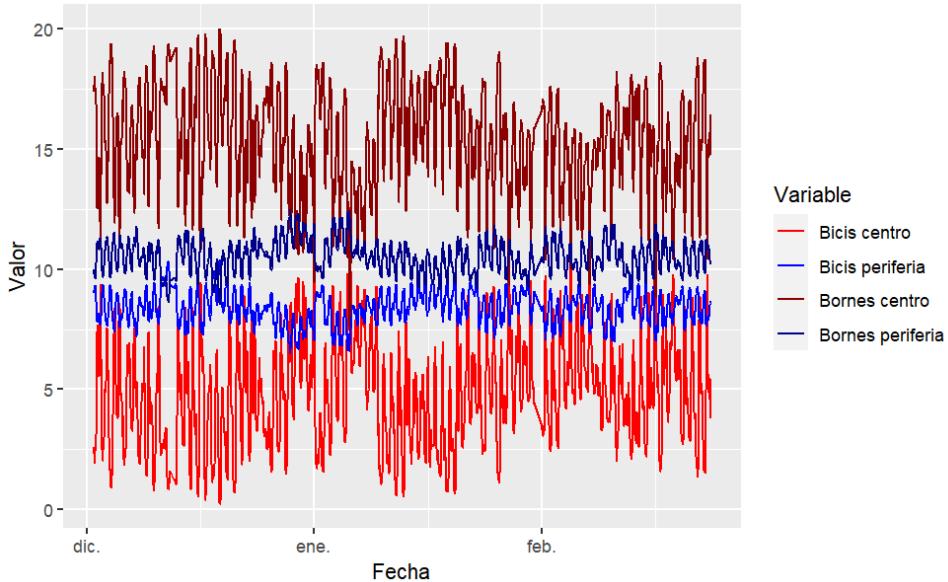


Figura 1: Evolución de la media de bicis y bornes en centro y periferia.

En el gráfico 1 se puede observar cómo en el centro hay muchos más bornes libres que bicicletas disponibles. En cambio, en la periferia los valores son muchos más parecidos y lo normal es que haya tanto bicicletas de sobra como espacios de sobra para aparcar.

También se ha visualizado el mismo gráfico con muestras de 10 estaciones de centro y 10 estaciones de periferia y se ha llegado a las mismas conclusiones.

Para poder visualizar mejor la evolución diaria, se han filtrado los datos seleccionando solo las observaciones de 2 días. En el gráfico de la Figura 2 se puede observar la evolución durante el día 6 y 7 de febrero (lunes y martes).

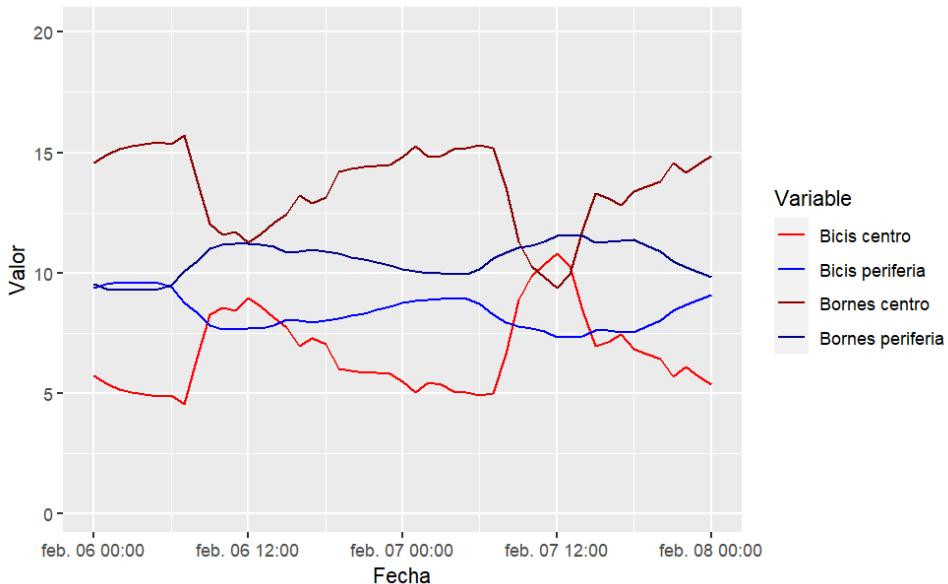


Figura 2: Evolución de la media de bicis y bornes en centro y periferia (6-2-2022 / 7-2-2022).

Se puede observar que la cantidad de bicis disponibles en el centro sube claramente en el mediodía y se mantiene baja por la mañana y por la noche. En cambio, las estaciones de la periferia tienen una evolución contraria y no tan acentuada.

Sin embargo, cabe destacar que la tendencia en el fin de semana es muy diferente a los días laborales y la cantidad de bicis disponibles aumenta en el centro durante la noche. En la periferia, sin embargo, se mantienen bastante estables el número de bicis disponibles y de bornes libres durante todo el día.

Estos análisis se han realizado sobre otros pares de días laborales y fines semanas, llegando a las mismas conclusiones.

A partir de aquí, se plantea la cuestión de si la cantidad de bicis en centro y la periferia está relacionada con el tramo horario. Por tanto se, considerarán los siguientes tramos horarios:

1. Madrugada: Hasta las 06:00h.
2. Mañana: De 06:00h a 09:00h.
3. Media mañana: De 09:00h a 11:00h.
4. Mediodía: De 11:00h a 16:00h.
5. Tarde: De 16:00h a 20:00h.
6. Noche: A partir de las 20:00h.

Con esto, se crea un nuevo dataframe llamado “tramos” que guardará la media de cantidad de bicis disponibles de todos los días de cada estación, por día de la semana y tramo horario. Además, se incluirá el tamaño de la estación, la zona (centro o periferia) y las coordenadas. Estas tres últimas variables tendrán, obviamente, el mismo valor para todas las observaciones de la misma estación.

Con las muestras de 20 estaciones en el formato del dataframe “tramos”, se han creado animaciones de diagramas de barras (ver anexo [6]) que distinguen el tramo horario y evolucionan día a día. En estas animaciones se da algún indicio de las evoluciones vistas en los diagramas de líneas, pero no deja claro que esta evolución sea así. Por otra parte, se ve claramente que las estaciones de la periferia suelen estar más llenas que las del centro, con más bicis disponibles y menos bornes libres.

Finalmente, con la intención de corroborar las conclusiones obtenidas a partir de los diagramas anteriores, se desarrolla un mapa, que puede ser observado en la Figura 3, de las estaciones a partir del dataframe “tramos”. En este mapa cada estación tendrá 42 anillos alrededor suyo del radio de las bicis disponibles en cada momento, uno por cada tramo horario de cada día de la semana. Estos anillos tendrán un color según el tramo horario al que pertenezcan. Además, se añadirá un anillo negro con el tamaño de la estación (estos anillos negros se pueden visualizar en el mapa interactivo del RMarkdown que puede consultarse en el anexo [6] y no en este documento, para ayudar a la visualización).

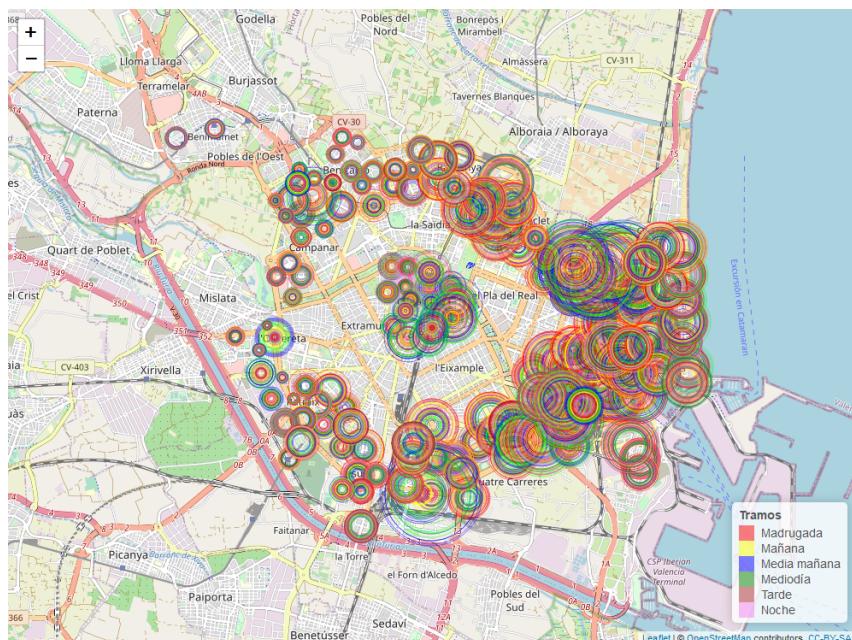


Figura 3: Cantidad media de bicis disponibles por día de la semana y tramo horario.

Se puede observar que en las estaciones centrales los anillos exteriores son de color azul y verde, perteneciente a la media mañana y mediodía; mientras que en las estaciones periféricas los anillos exteriores son de los colores pertenecientes al principio y final del día.

Desde el mapa interactivo se puede visualizar mucho mejor, y gracias a los anillos negros se puede ver también que las estaciones periféricas tienen en su mayoría mayor porcentaje de ocupación de bicis disponibles que las estaciones centrales. Cabe destacar que la zona universitaria, que pertenece a la periferia, tiene un comportamiento más semejante a las estaciones centrales que al resto de la periferia.

3.1.3. Discusión

A partir de los análisis realizados, se puede llegar a la conclusión de que hay más bicis disponibles en la periferia que en centro, donde suele haber muy pocas, al inicio del día. Según va avanzando la mañana, la cantidad de bicis disponibles empieza a disminuir en la periferia y sube bruscamente en el centro (y también en la zona excepcional de las universidades). Por último, una vez pasa el mediodía, la cantidad de bicis sigue una evolución opuesta a la de la mañana, pero más larga (durante más horas); hasta que finalmente, se estabiliza en una situación similar a la de la mañana.

La principal causa de esta evolución a lo largo del día es la gente que vive en la periferia de la ciudad y utiliza Valenbisi para acudir al trabajo o a la universidad. Por ello, a primeras horas de la mañana, en las estaciones periféricas empieza a disminuir el número de bicicletas con destino a las estaciones centrales y universitarias. Una vez los ciudadanos acaban sus jornadas laborales, regresan igualmente a sus casas, dejando las bicicletas como las encontraron por la mañana.

Este suceso del desplazamiento al trabajo explica porque en los fines de semana no se produce la misma evolución que en los días laborales; ya que mucha gente no va a trabajar (o a la universidad) los fines de semana.

3.2. Objetivo 2. Variables meteorológicas

Para este objetivo, se busca relacionar los datos de las estaciones obtenidas con la base de datos de la AEMET, de forma que se pueda observar si influyen las precipitaciones, la velocidad media del viento o la temperatura media a la demanda de bicicletas de Valenbisi.

3.2.1. Preprocesado

En base al conjunto de datos “df_merge”, se ha creado un nuevo dataset para filtrar aquellas variables que interesan del resto, creando de esta forma “df_merge2”, que esta compuesto por las variables características de los bornes y bicicletas, junto a todas aquellas variables que hacen referencia a los sucesos climáticos y metereológicos como son “tmed” (temperatura media), “prec” (precipitaciones) y “velmedia” (velocidad media del viento).

Así pues, se obtiene un dataset de 9 variables con las que poder empezar a realizar los análisis deseados. Centrándose en las variables climáticas, se las ha tenido que transformar a numéricas (puesto que estaban como categóricas) y sustituir las comas por puntos para que se puedan aplicar las medias y obtener sus consecuentes frecuencias.

Además de estas variables, también se ha tratado la variación de las estaciones para obtener cuáles son las que tienen un mayor nivel de movimiento. Para conseguir esta variable (denominada “df_variacion”, se ha tenido que partir del dataset “df_merge”, ordenarlo por las variables “number_”, “fecha” y “hora_hora” y obtener la variación mediante una función. De esta forma, se tienen ya todos los datos disponibles para proceder a los análisis.

Para más información sobre el preprocesado y para ver todos los gráficos no incluidos en el presente documento, ver el anexo [7]

3.2.2. Procedimiento

Teniendo ya todos los datos que se necesitan transformados, se puede actuar ya de acuerdo a los análisis que se quieren realizar, sabiendo que se desea observar si existe alguna diferencia de cambio de demanda de las bicicletas de Valenbisi cuando se dan sucesos climáticos/meteorológicos respecto cuando no las hay. Para ello, se han creado diferentes datasets filtrados según los valores de las variables que se buscan (por ejemplo, en el caso de las precipitaciones, se ha creado un dataset donde los valores en “prec” son mayores a 1, lo que se puede considerar ya precipitaciones moderadas, y otro dataset donde los valores en “prec” son iguales a 0, lo que implica que no llueve).

Desde estos datasets con un valor como punto de inflexión para diferenciar cuando se da un suceso climático respecto cuando no, se han realizado las medias de las frecuencias de bornes libres y de las bicicletas disponibles para diferenciar entre los dos datasets de la misma variable si realmente influye el suceso meteorológico o no. A partir de la tabla de la Figura 4, se puede obtener ya alguna conclusión.

```
##   Media_prec Media_vel Media_tmed
## 1  0.5851133 0.6044911  0.5955746
## 2  0.5892091 0.5847167  0.5856007
## 3  0.3986329 0.3837375  0.3905497
## 4  0.3969539 0.4010250  0.3999425
```

Figura 4: Frecuencia de medias de variables meteorológicas.

Partiendo de las frecuencias de los datasets de cada variable meteorológica, se ha profundizado aún más en los análisis buscando como afectan dichas variables a las estaciones que tengan un mayor nivel de movimiento (demanda de bicicletas), para lo cual se ha creado el dataset “df_variacion” mencionado anteriormente.

Con este nuevo dataset, se han vuelto a realizar los análisis sobre las variables metereológicas, pero esta vez, filtrando por aquellas estaciones que tengan un valor de valoración mayor a la media de todas las estaciones.

```

##   Media_prec_mov Media_vel_mov Media_tmed_mov
## 1      0.5504299      0.5684721      0.5654061
## 2      0.5795999      0.5858315      0.5825677
## 3      0.4293568      0.4201789      0.4133801
## 4      0.4036326      0.3985609      0.4018804

```

Figura 5: Frecuencia de medias de variables meteorológicas con movimiento.

Cabe destacar que se han realizado otro tipo de análisis, como una matriz de correlaciones o un mapa de calor que nos corroboran todos los análisis al explicarnos las relaciones existentes entre cada variable empleada en este objetivo.

Para hacer más visible las frecuencias para cualquier lector, se han creado varios gráficos de barras horizontales de forma que se pueda apreciar cuando varía la frecuencia de bornes libres/bicicletas disponibles para cada variable, incluyendo también la variación de las estaciones en algunos gráficos. Concretamente, se han creado gráficos para cada variable climática (“prec”, “tmedz” “velmedia”) para sus frecuencias sin filtrar por mayor movimiento y filtrando por este. Y por último, se han creado también dos gráficos de las tres variables juntas, una con todas las frecuencias sin filtrar por mayor movimiento, y el otro filtrando por este.

3.2.3. Discusión

Después de mencionar tanto las frecuencias que se han obtenido en todas las variables en general, como filtrando por aquellas con mayor movimiento, como los gráficos mencionados anteriormente creados a partir de las propias frecuencias para hacer las conclusiones de los análisis más ilustrativos, se van a interpretar los gráficos para discutir los resultados:

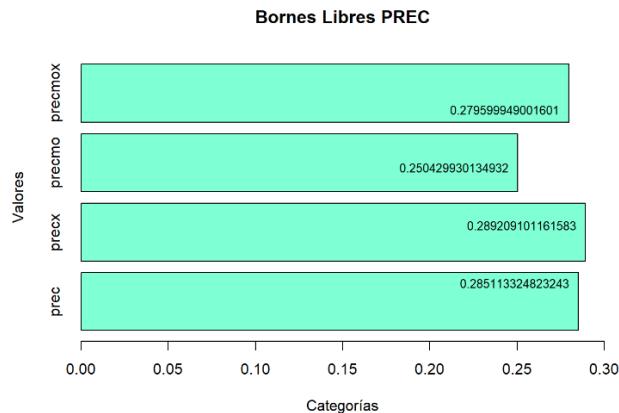


Figura 6: Gráfico de barras de bornes libres en prec.

En el gráfico de la figura 6 se muestran las variables de precipitaciones tanto con y sin filtro de movimiento centradas en los bornes libres. Como se puede observar, la frecuencia de bornes libres es menor cuando llueve (“prec”) respecto a cuando no llueve (“precx”). Este hecho se ve mucho más claro cuando nos fijamos en las variables filtradas por movimiento, donde se da el mismo caso pero con una distancia mucho mayor entre frecuencias. Por otra parte, se sabe que el gráfico de bicicletas disponibles es el complementario de este gráfico.

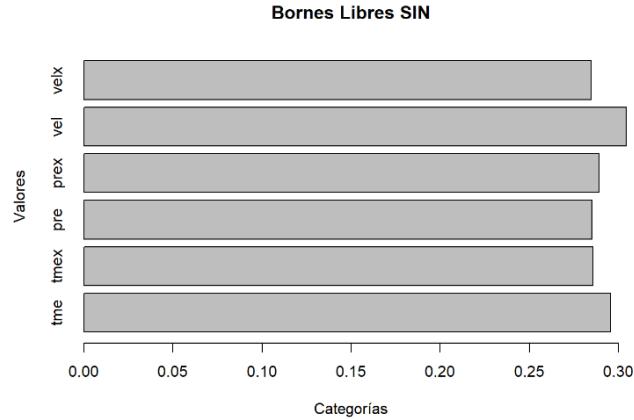


Figura 7: Gráfico de barras de variables sin movimiento.

Centrándose en las seis variables del análisis inicial, en el gráfico de la figura 7 se observa la frecuencia de las variables a nivel general (sin filtrar por movimiento) en relación a los bornes libres (puesto que bicicletas disponibles es su complementario, se centra solo en este caso), donde se puede ver que hay una mayor frecuencia de bornes libres en vel (“velmedia”), prex (“prec_exc”) y tme (“tmed”). Esto indica que sin filtrar por movimiento, hay una mayor proporción de bornes libres, y por ende una cantidad mayor de bicicletas circulando, cuando el nivel de velocidad media del viento es significante, cuando la temperatura media es elevada y cuando no llueve.

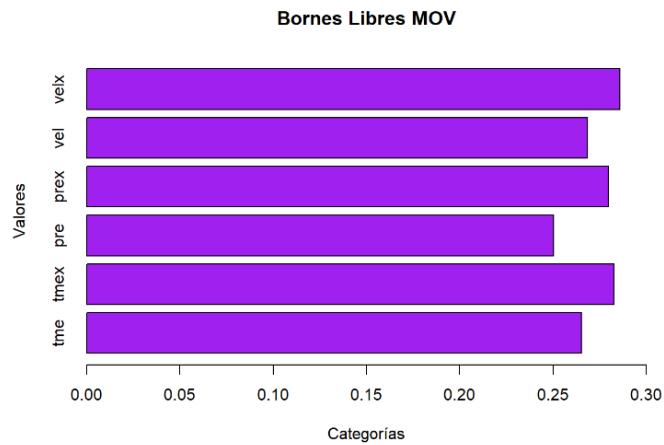


Figura 8: Gráfico de barras de variables con movimiento.

Para concluir, en la figura 8 se puede observar el mismo gráfico que antes pero esta vez filtrado por aquellas estaciones con una mayor variación en sus datos, es decir un mayor nivel de movimiento. Así pues, se puede observar como en este caso las mayores frecuencias de bornes libres son las de las variables velx (“velmedia_exc”), prex (“prec_exc”) y tmex (“tmed_exc”), lo que implica que cuando hay más bicicletas en circulación (y más bornes libres) es cuando no llueve, no hay viento (o no se registra un nivel significativo de este) y la temperatura media es menor a la media. Cabe recalcar que cuando se filtra por mayor movimiento en las estaciones, las frecuencias difieren en mayor distancia respecto a su dataset contrario al punto de inflexión que se toma para dividirlos (respecto a cuando no se filtra por variación).

3.3. Objetivo 3. Agrupación de las estaciones según el día de la semana

Este tercer objetivo trata de encontrar patrones en las estaciones en función del día de la semana y sus porcentajes de ocupación horaria.

3.3.1. Preprocesado

Primeramente, se cargó el *data frame* agrupado del que se obtuvo un subconjunto de variables a utilizar. Como se ha comentado ya en el objetivo 1, existen algunas estaciones que a cierta hora no hay ni bicicletas disponibles ni bornes libres. En este caso, en vez de eliminar estaciones enteras, se consideró quitar únicamente las filas incoherentes (1371 filas), ya que se haría el promedio de los porcentajes de ocupación por hora y suprimir estaciones enteras impediría analizarlas y ver su comportamiento.

A continuación, se incluyó una nueva variable que permitiría saber el porcentaje de ocupación de bicicletas en cada estación a cada hora del día, para que el análisis no se basase en el promedio de bicicletas (que podría considerarse alto o bajo según el total de bornes), sino en su porcentaje. Esto se hizo multiplicando la columna de bicis disponibles, *avg_av*, por 100, y dividiéndolo por su total de bornes, *avg_total*.

Como la idea era buscar los patrones de los 7 días de la semana, se creó una función, *generar_tabla_dia*, para generar las 7 tablas, que se utilizarían durante todo el proceso de análisis. Se nombraron como *df_diasem*, siendo *diasem* el día de la semana (sin tildes).

Para más información sobre las variables y la función creada, así como sobre los gráficos y el mapa interactivo (figura 14) que se mostrarán posteriormente, consultar el anexo [8].

3.3.2. Procedimiento

Tras la obtención de los *data frames* se procedió con el análisis de agrupamiento.

En primer lugar, se cogió un subconjunto de atributos que serían necesarios, más adelante, para la adición de los clusters de cada día de la semana y su representación en un mapa. Concretamente, se incluyeron los identificadores (*number_*) y nombres (*name*) de las estaciones, así como sus coordenadas (*latitud* y *longitud*). Se le asignó el nombre de *df_cluster* al nuevo *data frame*.

Se siguieron los mismos pasos para cada día. Se empezó quitando la primera variable de los *df_diasem*, ya que no sirve para decidir si una estación va a un cluster u otro, además de que sacaría conclusiones erróneas.

Luego, se realizaron mapas de calor para ver si existían tendencias de agrupación entre las estaciones. En la Figura 9, por ejemplo, se muestra el mapa de calor del lunes.

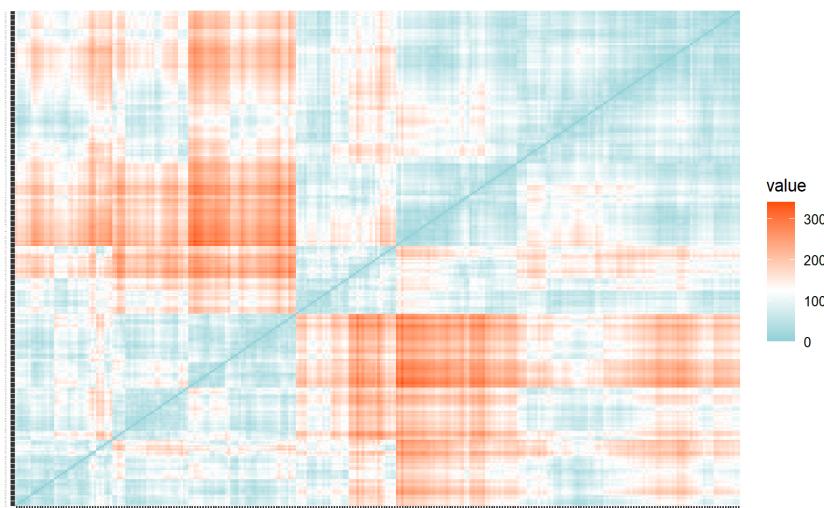


Figura 9: Mapa de calor del lunes.

Los cuadros azules refejan la cercanía de las estaciones. Se observan unos cuantos grupos (entre 2 y 5) definidos, lo que afirma la relación entre las estaciones y se puede proceder con el análisis.

Se eligió un método de partición para el análisis, pues se obtuvieron mejores resultados que con los métodos jerárquicos. Específicamente, se trabajó con el k-medoides (o PAM), ya que este da una estación real como centro (medoide) en vez de uno ficticio como en k-means.

Para la elección del número de clusters se empleó el método de Ward sobre dos criterios: el método del codo con el coeficiente de Silhouette y la variabilidad intra-cluster. La intención es maximizar el primero y minimizar el segundo. Siguiendo con el ejemplo del lunes, se presentan los gráficos resultantes en la Figura 10

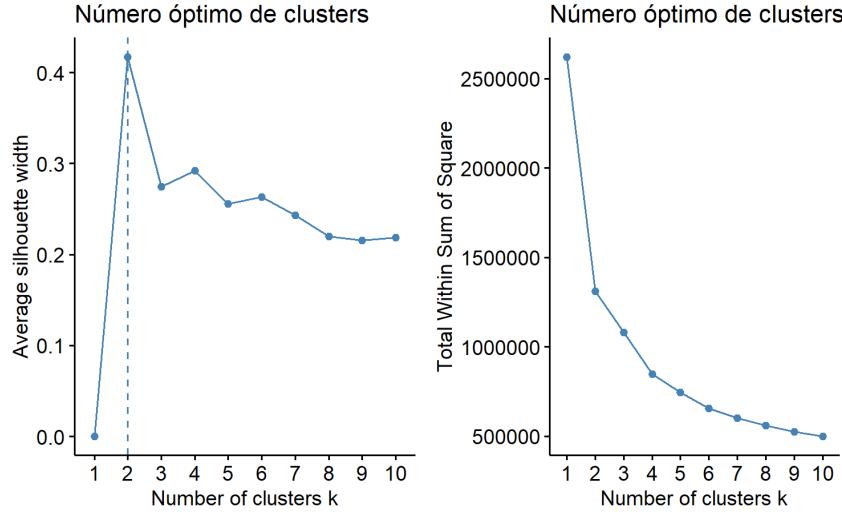


Figura 10: Número óptimo de clusters con el método de Ward.

En particular, se eligieron 4 clusters para la agrupación porque, a pesar de que 2 sea el número óptimo con Silhouette, la suma intra-cluster es lo suficientemente pequeña. Con 4 clusters, se obtuvo el gráfico de la Figura 11.

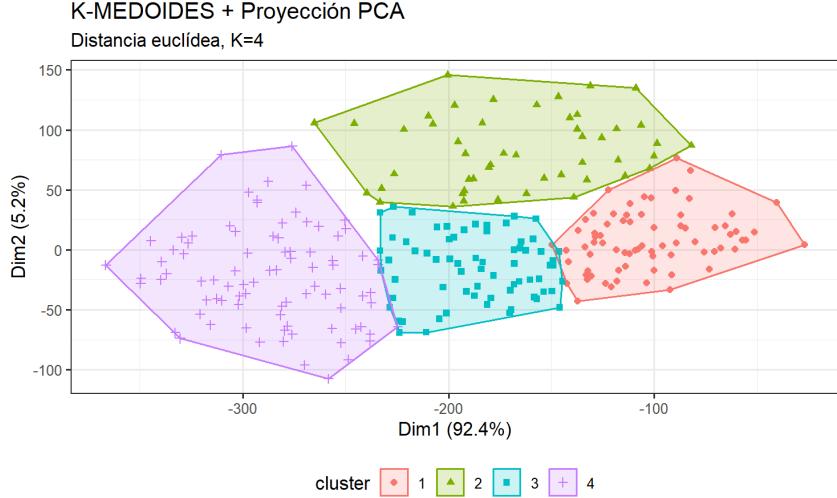


Figura 11: Agrupación de las estaciones con k-medoides los lunes.

Durante el análisis, se observaron las diferencias entre los patrones de agrupación de los clusters los días laborables y los fines de semana. Se muestra un representante para los días laborables (Figura 12) y otro para los fines de semana (Figura 13).

En la Figura 12, se aprecia el movimiento de los clusters 2, 3 y 4. Las pautas que siguen los dos últimos son similares, tendiendo a una falta de disponibilidad de bicis entre las 7 de la mañana y las 15 del mediodía. El cluster 2, en cambio, hace todo lo contrario; durante esas horas tiene muchas bicis disponibles. Esto

puede ser debido a que las estaciones de los clusters 3 y 4 suelen ser las que se utilizan para desplazarse a algún lugar de interés, como la universidad u hospital (se estudiará en el Objetivo 4), mientras que las que se incluyen en el segundo cluster suelen ser estaciones poco frecuentadas y de las que se trasladan las bicicletas por la madrugada a estaciones más transitadas, explicando la baja disponibilidad de bicis en la madrugada. Respecto al cluster 1, su movimiento es prácticamente nulo.

A diferencia de los días laborables, los fines de semana presentan poco movimiento en las estaciones. Los clusters 1 y 2 no presentan casi movimiento, pero se advierte de la bajada de disponibilidad a partir de las 8 de la mañana en el cluster 3. Lo más probable es que sea debido a las personas que solo trabajan de mañanas los sábados, lo que explica también el movimiento por las tardes.

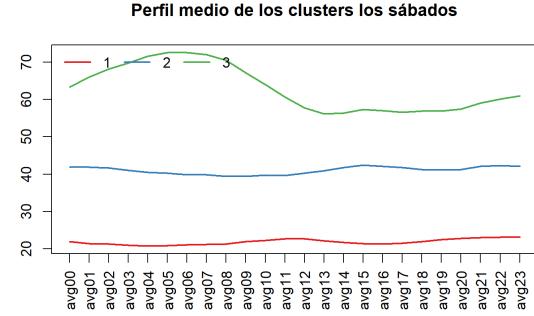
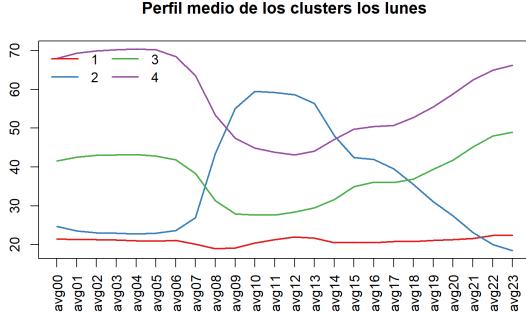


Figura 12: Representante del perfil días laborables. Figura 13: Representante del perfil fines de semana.

Finalmente, para poder crear mapas donde los puntos se coloreasen por cluster, se añadieron a *df_cluster* las columnas con los clusters a los que pertenecían cada estación según el día de la semana (7 columnas adicionales).

3.3.3. Discusión

En definitiva, se han visto notorias diferencias entre cómo actúa una estación un día laborable frente a un día de fin de semana. Básicamente, se debe a que en los fines de semana la gente no suele salir (porque no trabaja ni tiene clase) o, por lo menos, no usan las bicis como medio de transporte, y sí lo hacen en días laborables.

A modo de conclusión, se hizo un mapa interactivo que indica los clusters a los que pertenece cada estación en función del día de la semana. En la Figura 14 se muestra una vista previa de cómo quedaría.

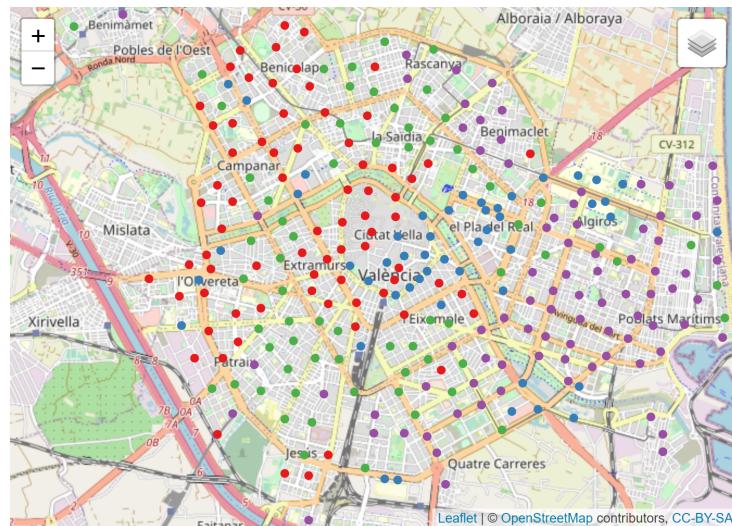


Figura 14: Vista previa de las estaciones con los clusters del lunes.

3.4. Objetivo 4: Lugares de interés y tramos semanales.

Este apartado tiene como meta encontrar diferencias de comportamiento entre las estaciones con lugares de interés, estudiar sus patrones y obtener el contraste entre el periodo laborable y el fin de semana.

3.4.1. Preprocesado

Para este objetivo, se han hecho varias transformaciones y adiciones de los datos, siendo una de las transformaciones similar a la realizada en el anterior objetivo.

En primer lugar, cargamos los datos del dataframe agrupados por horas y concatenados, y sobre este dataframe se harán las transformaciones. Para evitar algunas estaciones con error, el primer paso es eliminar las estaciones que tienen valor 0 en avg_av y avg_free (lo cual es incoherente porque una es la contraria de la otra).

Para identificar las estaciones que se encuentran cerca de lugares de interés en València, se han examinado los mapas convencionales de la ciudad y el mapa de la red de Valenbisi y mapas físicos. Los datos se han ido recogiendo en un archivo excel, que contiene columnas binarias de las diferentes categorías de lugares de interés.

En el excel, mediante un anidamiento de condicionales, se obtuvo una columna de categoría. Para obtener las etiquetas más significativas, se estableció una jerarquía, aunque no era del todo realista ya que convenía preservar algunas estaciones, como por ejemplo las de playa al ser muy pocas. De mayor a menor, los condicionales eran: Playa, Centro Comercial, Hospital, Tren, Monumento, Universidad, Instituto, Deporte y nada.

Finalmente, para tratar los datos y, al igual que se ha hecho anteriormente, se hace la conversión de las bicis ocupadas en cada estación dependiendo de los bornes totales en un porcentaje de ocupación, teniendo ya escalados los datos. De manera similar al objetivo anterior, se han agrupado todas las estaciones y se ha hecho la media de todos los datos por horas y por estación. De esta manera, cada fila era una estación y cada una de las 24 columnas añadidas es una hora concreta del día.

Finalmente, se hace un filtro de las estaciones que en la variable "Categoría" tienen un valor diferente de "Nada" (es decir, un lugar de interés cercano), y se han obtenido un total de 99 estaciones, que son los datos que necesitamos para el análisis general, así como las variables que realmente nos serán útiles, como el id, el nombre, las coordenadas en latitud y longitud y su categoría.

3.4.2. Procedimiento

El planteamiento del objetivo es tratar de ver si influyen los variados puntos de interés en el estudio. Se debe encontrar si existe realmente relación entre su comportamiento, agrupar las estaciones por comportamiento similar y visualizar el mismo y su localización. Para obtener el comportamiento similar de las estaciones, hemos planteado un clustering que agrupará las estaciones. Antes de realizarlo, se hacen las comprobaciones para ver si hay agrupamiento y observamos tanto analíticamente como gráficamente que si existen tendencias a comportarse de manera similar.

Tras varios ensayos y a raíz de los resultados, obtenemos cuatro clusters. Para comprobar que existe relación entre los clusters y los lugares de interés asignados a cada estación, se decide hacer un test de independencia y un Análisis Factorial de Correspondencias, con el objetivo único de observar la relación entre las variables solamente, confirmando las sospechas de una fuerte correlación. De cada cluster, hemos sacado su medoide (la observación que más se asemeja al centro del cluster) y se han visualizado los valores de sus variables en una gráfica.

De esta manera, observamos el comportamiento de los clusters, pero su interpretación está intrínsecamente ligada a los sitios con algún interés cerca, de manera que se añade al lado de la gráfica una tabla de frecuencias cruzada de los clusters y los lugares de interés. Tras esto, se visualiza en un mapa las localizaciones de las estaciones coloreadas por el cluster al que pertenecen.

Para finalizar con el objetivo, se debe visualizar si existe, además, diferencia entre días laborables y fines de semana. Para ello, se repite el procedimiento de clustering y visualizado (no igual con las comprobaciones para saber si hay agrupamiento o los test de independencia, pues ya la hemos confirmado anteriormente), pero esta vez se hará una vez con las filas cuyo día de la semana difiera de sábado y domingo y otra al revés.

Para más información sobre el procedimiento, consultar el anexo [9].

3.4.3. Discusión

3.4.3.1 Análisis de los clusters

En primer lugar, se muestra la relación que tienen las estaciones con lugares concretos cerca con respecto del cluster que se les ha asignado.

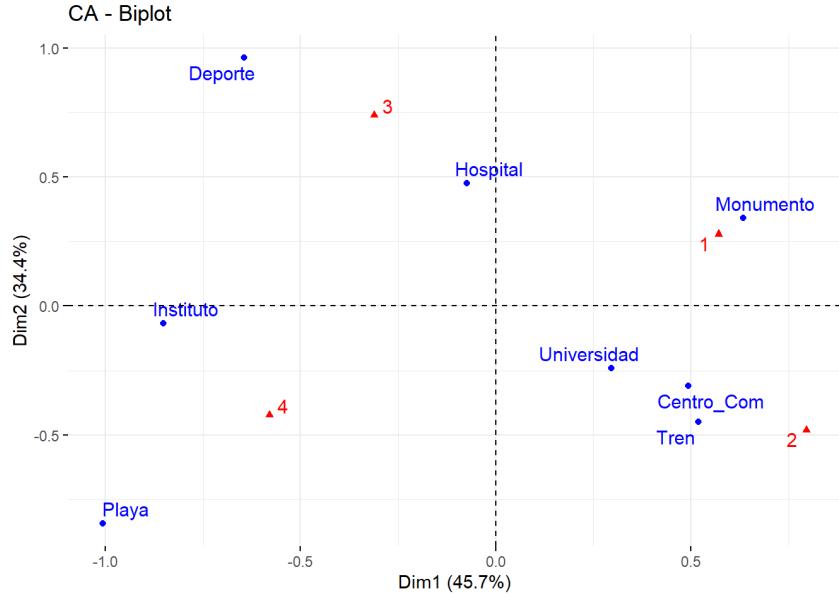


Figura 15: Influencia de las estaciones con respecto de los clusters.

En la Figura 15 se pueden vislumbrar ciertos resultados antes de ver el comportamiento de los clusters. El cluster uno se puede ver que está muy relacionado con monumento, siendo muy contrario al cluster 4 y teniendo correlación negativa con la playa. Ahora bien, en la primera componente se observa relación positiva con universidad, centro comercial y tren, siendo que estos dos últimos han influido mucho más en el segundo cluster. Con respecto al tercer cluster, el hospital y el deporte parecen estar relacionados con el tercer cluster. Por último, en el cuarto cluster se aprecia la relación con el instituto y con la playa, aunque también muchísima relación con el deporte en la primera componente.

	1	2	3	4
Centro_Com	5	8	2	5
Deporte	1	0	6	2
Hospital	7	0	6	5
Instituto	0	0	3	6
Monumento	8	2	1	1
Playa	0	0	0	10
Tren	0	3	1	1
Universidad	4	5	2	5

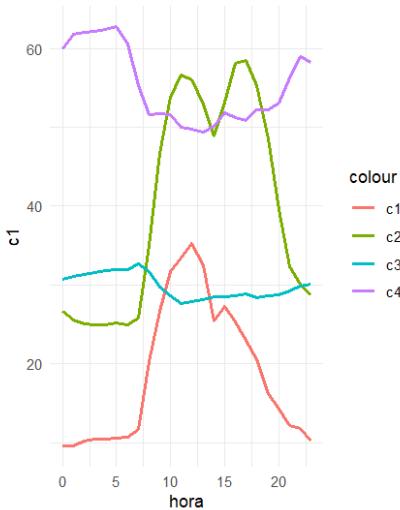


Figura 16: Comportamiento de los clusters y su tabla de frecuencias correspondiente. Análisis general.

En la Figura 16 se puede observar cierta tendencia por parte de los lugares de interés por situarse en clusters concretos. Todas las estaciones de la playa caen en el cuarto cluster, la mayoría de los monumentos

en el primero y las de deporte en el tercer cluster. El primer cluster se puede observar muy vacío a primera horas de la mañana, y se va llenando a lo largo del día, aunque sufre un bajón según se acercan las dos de la tarde y ya va disminuyendo. Se deduce de los gráficos que este cluster se caracteriza por las consultas externas de los hospitales, por los monumentos que seguramente solo abren por las mañanas y por los centros comerciales, además de las estaciones de la universidad de la facultad de medicina que han sido clasificados como hospital.

El segundo cluster, se caracteriza por muy alta ocupación a lo largo del día. Principalmente se ocupa alrededor de las 11 de la mañana, teniendo en cuenta la afluencia a las estaciones de tren. Sin embargo, sobre las dos de la tarde, da un bajón coincidiendo con el primer cluster, coincidiendo con la hora de salir de trabajar y de la universidad, así como la hora de abandonar los centros comerciales. Sin embargo, vuelve a subir a las cuatro y a las seis de la tarde, coincidiendo con la asistencia a clases de universidad y a los centros comerciales.

El tercer cluster se caracteriza, principalmente, por los lugares con instalaciones deportivas y con hospitales. No se observa mucho movimiento en este cluster, más allá de una ligera desocupación a media tarde. En el último cluster está totalmente caracterizado por las estaciones de la playa (todas caen aquí), por dos tercios de las estaciones de los institutos y por ser el que más estaciones agrupa. Principalmente, tiene una alta ocupación por la mañana, que se reduce a lo largo de la tarde (la gente abandona la playa coincidiendo en un periodo de invierno, y los alumnos de los institutos).

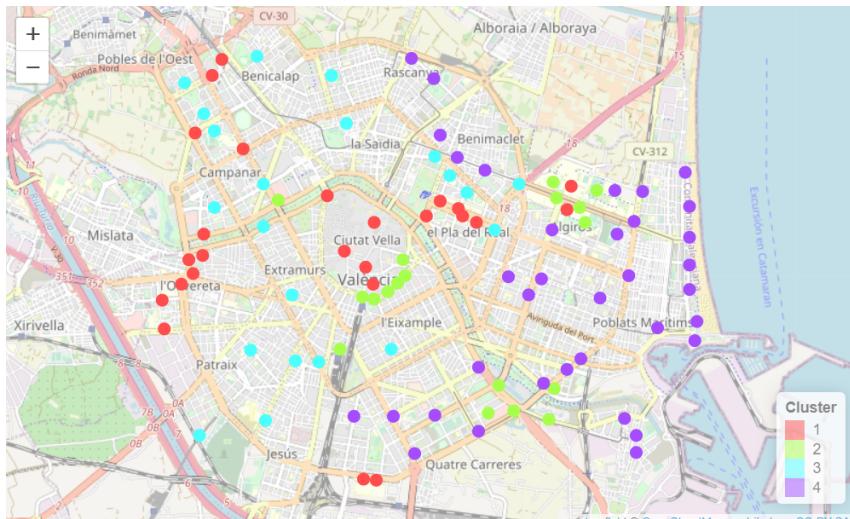


Figura 17: Distribución geográfica de los clusters. Análisis general.

En el mapa de la Figura 17, se puede ver que los clusters tienen cierta coherencia geográfica pues, en general, las estaciones del mismo cluster están agrupadas. El cuarto cluster ocupa todo el este de València (como ya imaginábamos al ver la playa) la zona universitaria y está presente en el sur y en el norte. Se puede ver la presencia del primer cluster en el oeste de Valencia, en el Hospital La Fe y en la parte de la universidad de Blasco Ibáñez. El segundo cluster ocupa parte de la universidad y la calle de Colón en el centro y el tercero se encuentra algo más repartido.

3.4.3.2 Diferencias entre el fin de semana y los días laborables

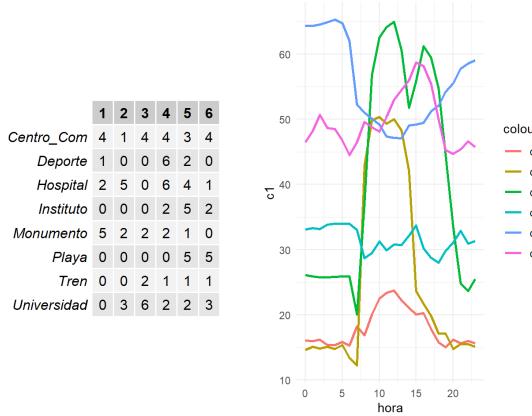


Figura 18: Comportamiento en días laborables.

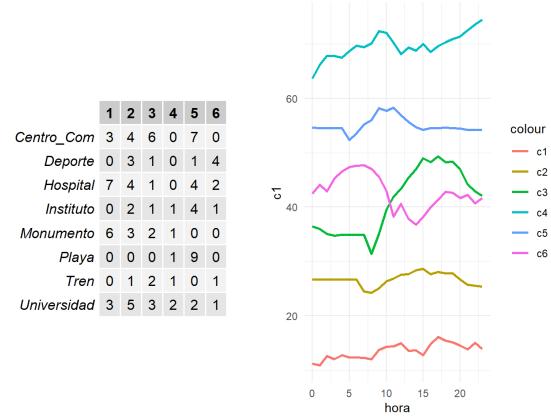


Figura 19: Comportamiento en fines de semana.

La característica más llamativa que se puede observar a simple vista, analizando los gráficos de las figuras 18 y 19, es la diferencia del número de clusters, pues en ambos subconjuntos se hace necesario escoger seis clusters un análisis óptimo. Otra característica que salta a la vista es la notable diferencia de variaciones en ambas gráficas. En el fin de semana, los clusters demuestran una variación muy inferior a la de los días laborables, siendo las variaciones mucho más livianas y en menor número.

La otra característica más destacable es la diferencia en el reparto de las estaciones con respecto a los clusters y cómo implica un cambio de comportamiento con respecto a uno y a otro periodo de la semana, indicando que los nuevos clusters, como era de esperar, han sufrido cambios.

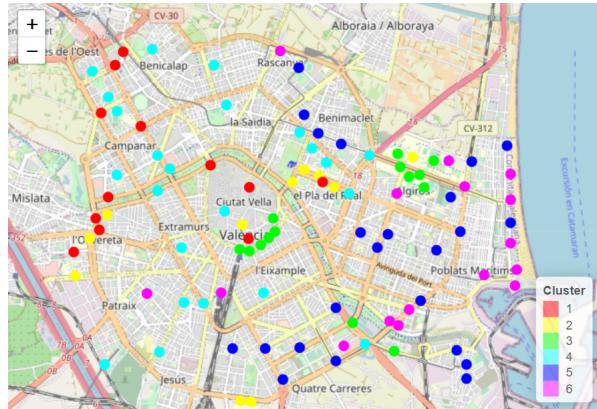


Figura 20: Distribución geográfica en días laborables.

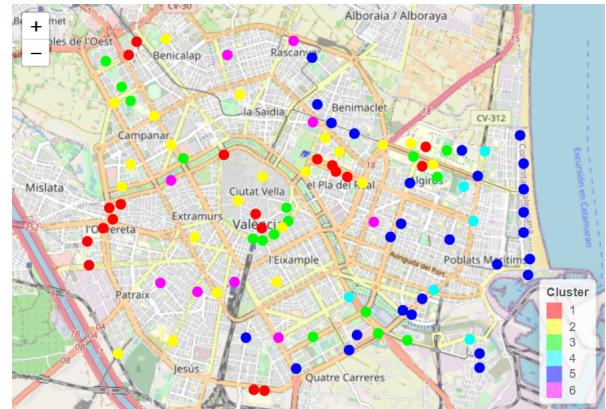


Figura 21: Distribución geográfica en fines de semana.

Finalmente, en las figuras 20 y 21, se observan grandes cambios en la distribución de las estaciones. En el día laborable, el este de Valencia (dejando de lado las universidades) se observa que predominan los puntos azules oscuros y magentas, y azules claros en la mitad oeste de Valencia, así como poca presencia de los puntos verdes y amarillos. En el mapa de fines de semana, desaparecen radicalmente los puntos azules claros y magentas, viéndose mermados y dispersos por el mapa. Los puntos rojos cobran algo más de protagonismo el fin de semana, al igual que los amarillos.

Como conclusión, diremos que los puntos de interés si afectan en gran medida al comportamiento de las estaciones de bicicletas, así como existe diferencia entre el comportamiento en fin de semana y en días laborables de estas estaciones.

3.5. Objetivo 5. Análisis predictivo

Este último apartado pretende describir el proceso seguido así como los resultados obtenidos por el modelo que intenta predecir el número de bicicletas disponibles.

3.5.1. Preprocesado

A continuación, se analizarán todas las cuestiones relativas al quinto y último objetivo. Si se desea ver el código utilizado para su desarrollo así como interactuar con el gráfico de la figura 22, se recomienda consultar el anexo [10]

Una vez hecho este inciso, para la realización de este objetivo, en primer lugar, se seleccionaron únicamente las variables que eran necesarias: el identificador de la estación, el número de bicis disponibles, el día de la semana, la hora y la fecha.

Como se analizará posteriormente, el orden de las filas es clave para que el programa cree una buena matriz de datos a la cual aplicar el modelo de regresión. En consecuencia, posterior conversión de la variable “fecha” a objetos fecha mediante la función “as.Date()”, se ordenó la base de datos por fecha y hora.

Posteriormente, la variable “hora” se convirtió en una variable de tipo hora mediante la función “as.POSIXlt()” y, una vez hecho esto, se convirtió a factor. Por su parte, las variables “día_semana” e “id_estación” también se convirtieron a factores. Finalmente, la variable “bicis_disponibles” se convirtió a variable numérica y, como consecuencia de la comentada agrupación por horas, se truncó.

Adicionalmente, se creó el conjunto de datos test con datos desde el 9 de febrero de 2023 (inclusive) hasta el 23 de febrero de 2023 (inclusive). Este conjunto no se utilizará para realizar la predicción (para ello se construirá una matriz) pero sí que será donde se almacenen las predicciones.

3.5.2. Procedimiento

Para intentar predecir el número de bicis, se partía de una estructura de datos como la de la tabla 2.

id_estación	bicis_disponibles	día_semana	hora	fecha
1	0	V	0	2022-12-02
2	6	V	0	2022-12-02
3	4	V	0	2022-12-02
...

Cuadro 2: Base de datos original

Para pasar de una estructura de datos como la indicada en la tabla 2 a la matriz de datos necesaria para realizar la predicción se utilizó una función, que recibe como parámetros la base de datos, un id de estación y una hora.

Para realizar la predicción, se decidió, después de probar otras variables con errores mayores y coeficientes de determinación menores, escoger como variables predictoras las bicis disponibles en las 3 horas anteriores y en la misma hora que se intenta predecir pero en el mismo día de la semana anterior (por ejemplo, si se quiere predecir el número de estaciones en un estación a las 14 horas, las variables predictoras son el número de bicis disponibles a las 13 horas, a las 12 horas, a las 11 horas y el número de bicis a las 14 horas del mismo día de la semana pero en la semana anterior).

Teniendo esto en cuenta, la matriz que devuelve dicha función, aunque eliminando columnas que se eliminan en la siguiente función, tiene la forma de la tabla 3.

semana_anterior	hora_21	hora_22	hora_23	hora_0
0	2	4	3	2
0	4	2	2	0
1	3	0	0	1
...

Cuadro 3: Matriz con las 3 horas anteriores

Una vez creada la matriz, el siguiente paso es realizar las predicciones, para lo cual se creó otra función, que recibe como parámetros el id de una estación y una hora. Esta función realiza una regresión lineal múltiple, donde la variable respuesta son los datos de la hora que se pasa como parámetro y las variables predictoras son el resto de las variables (las bicis disponibles en las 3 horas anteriores y en la misma hora del mismo día pero de la semana anterior).

Una vez definidas las dos funciones, se procede con la obtención de las predicciones de los datos de testeo. Para cada fila, se llama a la función predictora (pasándole como parámetros el id de la estación y la hora correspondiente) que devuelve un vector de valores y otro de índices. Con estos dos objetos, se asignan a la variable *predicho* de las filas en cuestión los valores correspondientes.

Después de este proceso, el conjunto de test tiene la forma de la tabla 4.

id_estación	bicis_disponibles	día_semana	hora	fecha	predicho
1	14	J	0	2023-02-09	14.6455
2	2	J	0	2023-02-09	1.7246
3	1	J	0	2023-02-09	2.0524
...

Cuadro 4: Datos de test con la predicción

3.5.3. Discusión

Para comprobar la calidad de la predicción llevada a cabo (y por tanto de los 276 modelos generados), se han utilizado medidas como el Root Mean Square Error (RMSE), el Mean Absolute Error (MAE) y el coeficiente de determinación (R^2). Sobre el conjunto de test, dichos indicadores toman los siguientes valores:

- RMSE = 1.766471
- MAE = 1.133816
- R^2 = 0.9170628

En general, estos valores son considerablemente buenos. Sin embargo, para medir si realmente la predicción puede considerarse aceptable, decidimos crear un modelo que predijera mediante medias (la media de una estación, un día de la semana a una hora concreta). Los mismos indicadores para este modelo toman los siguientes valores:

- RMSE = 5.072797
- MAE = 3.982729
- R^2 = 0.3382811

Concluimos, por tanto, que el modelo es aceptable (cómo mínimo mejor que el modelo basado en medias) en cuanto a las predicciones que realiza.

Por su parte, para analizar un caso concreto, en la figura 22 se puede observar un gráfico que muestra la evolución del número de bicicletas disponibles (reales en azul y predichos en naranja) el día 2023-02-14 en la estación 267.

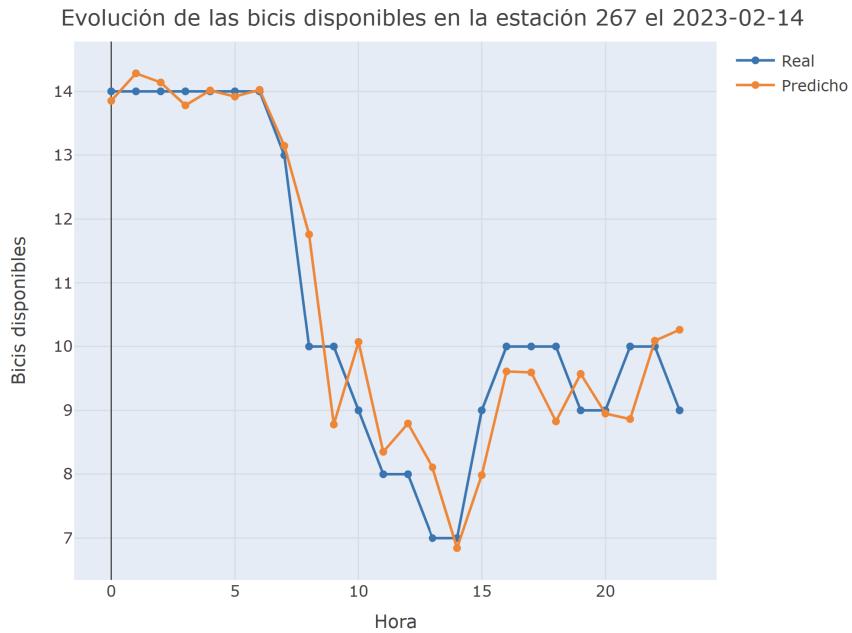


Figura 22: Comparación de las bicicletas disponibles y predichas.

Analizando la Figura 22, se puede concluir que, para este caso concreto, las predicciones realizadas son considerablemente buenas ya que, para un gran porcentaje de las horas, la diferencia entre el valor real y el predicho es menor a 1.

Adicionalmente, se ha creado una página web ⁴ en la cual los usuarios pueden consultar cuántas bicis habrá disponibles en una estación, un día de la semana, a una hora concreta. Debido a que para poder utilizar el modelo de regresión creado sería necesario disponer de datos en tiempo real, el modelo que subyace a la web es el mencionado modelo basado en medias. Dejamos como propuesta de futuro la consideración de datos en tiempo real, lo que permitirá utilizar el modelo de regresión y que, por tanto, proporcionará predicciones con un error menor.

⁴<https://valenbisi.shinyapps.io/prediccion/>

4 Lecciones aprendidas

Algunas de las lecciones que hemos aprendido como consecuencia de la realización de este proyecto y que, muy probablemente, nos serán útiles para nuestros futuros proyectos en Ciencia de Datos son:

- Interdependencia de algunos objetivos y trabajo en equipo. Durante el proyecto, nos dimos cuenta de que algunos de nuestros objetivos estaban interrelacionados, lo que nos enseñó la importancia de trabajar en equipo y brindarnos apoyo mutuo para alcanzar nuestros objetivos de manera eficiente.
- Estructura lógica y coherente en la presentación de resultados. Dado que, como se ha mencionado, algunos de nuestros objetivos estaban interconectados, aprendimos la importancia de presentar los resultados de manera lógica y coherente. Comprendimos que los hallazgos de un análisis podían ser relevantes para otros y que la presentación ordenada de los resultados facilitaba su comprensión.
- Potencia de R en análisis estadísticos. Durante el proyecto, pudimos apreciar que R es considerablemente más potente que Python cuando se trata de realizar análisis estadísticos, ya que este lenguaje ofrece una amplia gama de paquetes y funciones especializadas, lo que facilita la implementación de técnicas analíticas avanzadas como puede ser el análisis clustering.
- Comparación con valores de referencia. Durante la realización del análisis de regresión, descubrimos que es útil comparar los resultados obtenidos con valores de referencia, ya que esta comparación nos permitió evaluar el rendimiento y la relevancia de nuestros resultados.

Además de estas reflexiones, consideramos importante enumerar y valorar las herramientas de software utilizadas por el equipo:

- R y RStudio. Este proyecto nos ha permitido aprender un nuevo lenguaje de programación, tan potente para el análisis de datos como lo es R. Todos y cada uno de los análisis que se han llevado a cabo en nuestro proyecto han sido realizados utilizando R.
- Repositorios para programas en R. Este proyecto nos ha permitido conocer plataformas, como la Shiny Apps o Rpubs, en las cuales poder publicar los trabajos realizados en R. La primera de las plataformas mencionadas nos permitió publicar la página web donde los usuarios pueden obtener la predicción de las bicicletas disponibles y la segunda nos permitió publicar los RMarkdowns.
- Python. En este proyecto, Python fue utilizado en la parte inicial del mismo. Su uso se limitó al procesamiento los datos de las bicicletas disponibles (concatenación de los diferentes archivos csv) y la obtención de los datos climatológicos mediante la realización de consultas a la API de AEMET.
- GitHub. En este proyecto, se ha hecho uso de esta plataforma para poder compartir los programas Python utilizados.

5 Anexos

- [1] https://github.com/valenbisi/obtencion_datos
- [2] https://github.com/valenbisi/main_aemet.py
- [3] https://github.com/valenbisi/hito_1.pdf
- [4] https://github.com/valenbisi/hito_2.pdf
- [5] <https://rpubs.com/tratamiento>
- [6] https://rpubs.com/analisis_exploratorio
- [7] <https://rpubs.com/metereologia>
- [8] <https://rpubs.com/clustering>
- [9] <https://rpubs.com/interes>
- [10] <https://rpubs.com/prediccion>