MEMORIA ENTREGA SPARK (BICIMAD)

Elena Pérez García
Carla Andrea San José Jiménez
Pablo Valenciano Martínez

1. INTRODUCCIÓN

Este trabajo consiste en el análisis y estudio de algunos datos relacionados con la empresa BICIMAD, de bicicletas públicas de Madrid. Nosotros hemos elegido, concretamente, el conjunto de datos “202121_movements.json”, obtenido de la web indicada en el campus virtual. Este conjunto de datos describe todos los trayectos realizados en febrero de 2021 mediante vehículos BICIMAD.

En el paquete de datos se describe un total de 262103 viajes. Para ello, se distinguen los siguientes factores:

- **Tipo de usuario** que ha realizado el viaje (0: Desconocido, 1: Usuario anual, 2: Usuario ocasional, 3: Trabajador de la empresa).
- **Rango de edad** del usuario (0: Desconocido, 1: entre 0 y 16 años, 2: 17 o 18 años, 3: entre 19 y 26 años, 4: entre 27 y 40 años, 5: entre 41 y 65 años, 6: más de 66 años)
- **Tiempo** de viaje
- **Fecha y hora** estimada del viaje
- Estación de **salida**
- Estación de **llegada**

El problema a resolver en este trabajo será encontrar diferencias significativas entre los factores que conciernen a los viajes diferenciando entre los que se han realizado entre semana y los que se han realizado los fines de semana.
Concretamente, hemos realizado una recategorización de la variable **Fecha y Hora** para crear una nueva: **“Viaje realizado durante el fin de semana/Viaje realizado en un día de diario”**.

En cuanto a cuestiones técnicas, todo el trabajo será realizado con la biblioteca pyspark (Spark) de python. Esta librería nos facilitará mucho el trabajo, pues cuenta con diferentes funciones de orden superior como map, reduce, filter, etc. sin las cuales el tratamiento de datos sería mucho más costoso (tanto humana como computacionalmente).

2. DESCRIPCIÓN DE LOS DIFERENTES ESTUDIOS REALIZADOS Y CONCLUSIONES

2.1. TIEMPOS DE USO SEGÚN LA EDAD

Primeramente hemos realizado un estudio de la significación estadística de la interacción entre el factor **"Diario/Fin de semana"** y el **rango de edad.**

Para ello hemos creado la función *obtener_dia_edad_tiempo()*, que toma por argumento un viaje procedente del archivo de datos (y sólo uno, pues la usaremos con un *map*) y devuelve una tupla de la forma **(1: si el viaje se ha realizado en fin de semana o 0: si entre diario, rango de edad, duración en segundos del viaje)**.

También hemos creado la función *estudio_semanal()*, que toma por parámetro el archivo de datos. Lo transformamos de la siguiente forma: le aplicamos, con un *map*, la función anterior, así obtenemos una **lista de tuplas de la forma (1/0,edad,tiempo)**. Ahora filtramos todos estos viajes quedándonos sólo con aquellos que se hayan realizado entre semana (aquellas tuplas cuya primera coordenada sea un 0, usando la función filter). Una vez que nos hemos quedado con estas tuplas, eliminamos de todas su primera coordenada (pues será en todas 0). Así, tenemos ahora un conjunto de viajes que sabemos que se han realizado entre semana, de los cuales sabemos **el rango de edad del viajante y el tiempo que duró el viaje**. Finalmente, agrupamos todos los viajes por rango de edad usando la función *groupByKey()* y hacemos la media de los tiempos para cada rango de edad, también con la función *map*.

Análogamente, hemos creado la función *estudio_findes()* que hace exactamente lo mismo pero filtrando los viajes que hayan sido realizados durante el fin de semana.

Los resultados que obtenemos son lo siguientes:

Medias de los tiempos de viaje realizados **entre semana** según la edad:

[(0, 1176.312300057936), (1, 1620.454980842912), (2, 981.2971698113207), (3, 1053.3499308437067), (4, 945.1009192255036), (5, 1032.1195643532872), (6, 1222.1041095890412)]

Medias de los tiempos de viaje realizados durante el **fin de semana** según la edad:

[(0, 1110.2198850730913), (1, 297.92246894994355), (2, 826.1958041958042), (3, 1049.7946603298853), (4, 944.174389574075), (5, 1075.1365607750365), (6, 1129.2574552683895)]

Observamos que el grupo de edad que pasa más tiempo viajando en BICIMAD entre semana es el grupo 1 **(entre 0 y 16 años)** y el que más durante los fines de semana es el grupo 6 **(más de 66 años)**.

Una vez obtenidos estos datos, hemos creado la función *resta_findes()* que calcula la resta de las medias de tiempos según la edad, para ver así en qué rangos hay diferencias significativas. El resultado obtenido es el siguiente:

[(0, 66.09241498484471), (1, 1322.5325118929684), (2, 155.10136561551656), (3, 3.5552705138213696), (4, 0.9265296514286092), (5, 43.016996421749354), (6, 92.84665432065162)]

Como el grupo 0 de rango de edad es “Rango desconocido” no lo tendremos en cuenta para el estudio. Observamos que la diferencia más grande se encuentra en el rango de edad 1 **(entre 0 y 16 años).** La media de duración de viajes de este grupo durante el fin de semana es de 297.92 segundos, mientras que entre semana es de 1620.4 segundos.

2.2. FRECUENCIA DE VIAJES SEGÚN LA EDAD

Otro aspecto a estudiar de este conjunto de datos es la cantidad de viajes realizados según el tramo semanal y la edad. Hemos creado la función *contador()* que toma por parámetro el archivo de datos el cual convertiremos en un rdd para trabajar con él quedándonos solo con los factores **Diario/Fin de semana** y la **edad** de cada viaje, a través de la función *obtener_dia_edad().* Una vez hecho esto, dividimos el conjunto en dos: “rdd_contador_diario” y “rdd_contador_findes”, en cada uno de ellos recogeremos los viajes realizados en cada tramo semanal correspondiente siendo la única información de cada viaje **la edad del usuario.** Así, usando la función *Counter()* obtenemos cuántos viajes se han realizado en cada tramo semanal (diario o fin de semana) según la edad.

Obtenemos los siguientes resultados:

Recuento de viajes realizados **entre semana** según la edad:

{0: 39699, 4: 15339, 5: 12579, 3: 2892, 1: 1044, 6: 365, 2: 212}

Recuento de viajes realizados los **fines de semana** según la edad:

{0: 103022, 4: 40054, 5: 35508, 3: 7154, 1: 2657, 6: 1006, 2: 572}

Como el grupo 0 corresponde a “Rango de edad desconocido” no podemos afirmar que, aunque haya diferencia significativa, sea relevante para el estudio.
Observamos que el rango de edad que más viajes realiza entre semana es el 4 **(entre 27 y 40 años),** y también durante los fines de semana. Lo mismo pasa para el rango de edad que menos viajes realiza, que es el 2 **(17 o 18 años),** tanto para fines de semana como para días de diario. Este dato es coherente, pues el rango 2 solo aborda 2 años de edad. Concluimos con que los rangos de edad están igualmente ordenados en cuanto a frecuencia de viajes tanto los fines de semana como los días de diario.

Diferencias en valor absoluto:

{0:63323, 4: 20169, 5: 27475, 3: 4262, 1: 38, 6: 360, 2: 2292}

Observamos que en los rangos de edad 4 y 5 se observa una gran diferencia de número de viajes entre el fin de semana y los días de diario. Recordamos que estos rangos de edad se corresponden a 4: entre 27 y 40 años y 5: entre 41 y 65. Estos grupos usan el servicio BICIMAD de una forma bastante más considerada los fines de semana que entre semana.

2.3. ESTACIONES MÁS TRANSITADAS

Finalmente, hemos estudiado cuáles son las estaciones más transitadas en el caso de los días de diario y de los fines de semana. Para ello hemos creado la función *estaciones()* que dado un viaje del archivo de datos, nos devolverá una tupla de la forma **(Diario/Finde, estación de salida, estación de llegada)**. Una vez hecho esto, llamaremos a esta función desde *transitadas_semanal()*, que hará el recuento de las estaciones más transitadas durante los días de diario, y desde la función *transitadas_finde()* que hará lo mismo pero para el fin de semana. Aclaramos que, para el tránsito de una estación hemos tenido en cuenta tanto que haya sido de salida como de llegada. Los resultados obtenidos son los siguientes:

Las 10 **estaciones** más transitadas **entre semana**

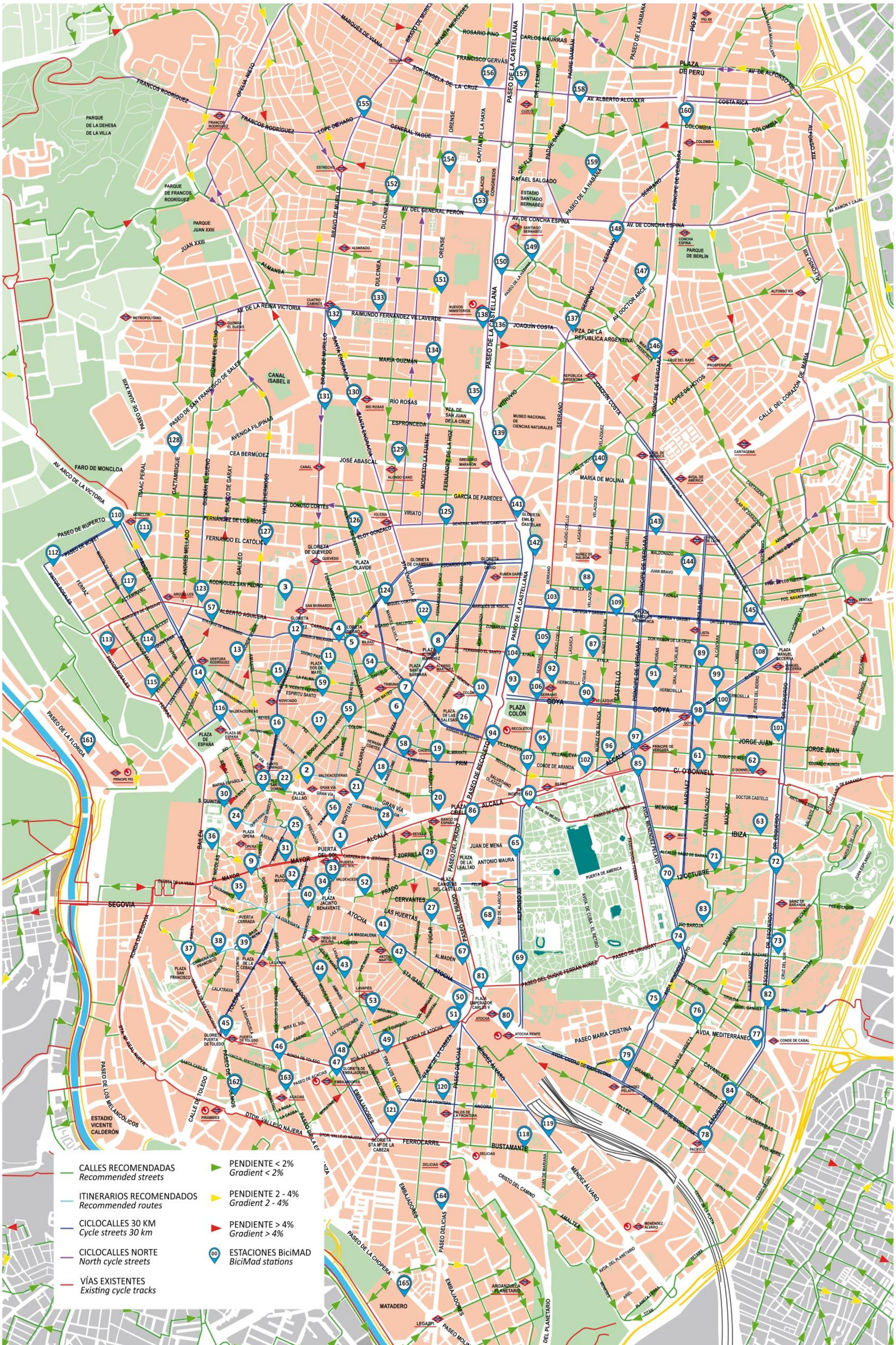
{**43**: 1911, **57**: 1725, **132**: 1504, **175**: 1446, **208**: 1223, **220**: 1188, **90**: 1174, **129**: 1172, **135**: 1166, **49**: 1133}

Las 10 **estaciones** más transitadas durante los **fines de semana**

{**57**: 4559, **43**: 4508, **208**: 3615, **175**: 3406, **132**: 3120, **49**: 3068, **220**: 3053, **9**: 2992, **163**: 2965, **13**: 2942}

La estación número 43 se encuentra entre **Tirso de Molina y Lavapiés**. La estación número 57 en **Argüelles**. Con este estudio podríamos decidir si en alguna de estas estaciones convendría aumentar el número de bases de la estación, para así evitar un posible colapso. Sin embargo, en nuestro conjunto de datos no aparece ninguna información de la disponibilidad de bicis de cada estación, por lo que esta cuestión no nos es alcanzable.

Podemos observar la ubicación del resto de las estaciones a través del mapa que hemos importado en la siguiente página:


CALLES RECOMENDADAS
Recommended streets
ITINERARIOS RECOMENDADOS
Recommended routes
CICLOCALLES 30 KM
Cycle streets 30 km
CICLOCALLES NORTE
North cycle streets
VÍAS EXISTENTES
Existing cycle tracks
PENDIENTE < 2%
Gradient < 2%
PENDIENTE 2 - 4%
Gradient 2 - 4%
PENDIENTE > 4%
Gradient > 4%
ESTACIONES BiciMAD
BiciMad stations
PASEO DE LA CASTELLANA
PLAZA DE PERÚ
PARQUE DE LA DEHESA DE LA VILLA
PARQUE DE FRANCOS RODRIGUEZ
PARQUE JUAN XXIII
ESTADIO SANTIAGO BERNABEU
CANAL ISABEL II
MUSEO NACIONAL DE CIENCIAS NATURALES
FARO DE MONCLOA
PLAZA DE ESPAÑA
PUERTA DEL SOL
PLAZA COLÓN
GOYA
ALCALÁ
GRAN VÍA
MAYOR
SEGOVIA
ATOCHA
PASEO DEL PRADO
PASEO DE RECOLETOS
PUERTA DE AMÉRICA
PASEO MARÍA CRISTINA
ESTADIO VICENTE CALDERÓN
PASEO DE LOS MELANCÓLICOS
MATADERO
PASEO DE LA CHOPERA
BUSTAMANTE
FERROCARRIL
EMBAJADORES
PASEO DELICIAS
AVDA. MEDITERRÁNEO
DOCTOR CASTELO
IBIZA
MENORCA
JORGE JUAN
C/. O'DONNELL
MARÍA DE MOLINA
JOAQUÍN COSTA
AV. DE CONCHA ESPINA
AV. DEL GENERAL PERÓN
RAIMUNDO FERNANDEZ VILLAVERDE
AV. DE LA REINA VICTORIA
FRANCOS RODRÍGUEZ
AV. ARCO DE LA VICTORIA
PASEO DE LA FLORIDA
CALLE DEL CORAZÓN DE MARÍA