

# PRÁCTICA BICIMAD

Antonio Francisco Álvarez Gómez, Marcos Rocha Morales, Marina Bueno García

## 1 Introducción

El objetivo de esta práctica es estudiar qué estaciones son la más populares y la menos populares para dejar la bicicleta en función de 3 franjas horarias distintas (mañana, tarde y noche). También estamos interesados en saber de qué zonas de Madrid (código postal) son los usuarios que más y que menos utilizan el servicio de Bicimad. Por último, estudiaremos dentro de los usuarios que usan más de una vez este servicio al mes aquellos cuyos trayectos son ciclos o caminos. Hemos tomado como referencia el mes de abril de 2017 pero hemos analizado la evolución de los usuarios que cumplen esta última condición a lo largos del año 2017 entre los meses de abril y diciembre.

Para ejecutar el programa es necesario:

1. Tener descargadas las bases de datos utilizadas con los siguientes nombres:  
abril2017.json, mayo2017.json, junio2017.json, julio2017.json, agosto2017.json,  
septiembre2017.json, octubre2017.json, noviembre2017.json, diciembre2017.json.  
Estas bases de datos han sido obtenidas de la siguiente página web:  
[https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-\(1\)](https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1))
2. Ejecutar el programa con el siguiente comando: `python3 practica_bicimad_spark.py`

## 2 Material usado y metodología

Hemos usado como base de datos relativos a la utilización del servicio de bicicletas eléctricas del ayuntamiento de Madrid durante el mes de abril de 2017 (abril2017.json). Las variables que hemos usado en nuestro estudio son:

- user\_day\_code : código del usuario. Todos los movimientos de un mismo usuario para una misma fecha tendrán el mismo código.
- user\_type : número que indica el tipo de usuario que ha realizado el movimiento:
  - 0: no se ha podido determinar el usuario
  - 1: usuario anual
  - 2: usuario ocasional

- 3: trabajador de la empresa
- idunplug\_station: número de la estación de la que se desengancha la bicicleta
- idplug\_station: número de la estación en la que se engancha la bicicleta
- unplug\_hourTime: fecha y franja horaria en la que se realiza el desenganche de la bicicleta. Solo se indica la hora de inicio del movimiento, sin la información de minutos y segundos
- zip\_code: código postal del usuario que ha realizado el movimiento

Por evitar tener datos que produzcan ruido o afecten a nuestras conclusiones hemos decidido quedarnos con los usuarios de tipo 1 y 2. Por las mismas razones, solo nos hemos quedado con los códigos postales válidos.

Con el fin de trabajar solo con las variables indicadas anteriormente, hemos hecho uso de *map*, cuyo parámetro es un función, llamada *mapper*, que desglosa los datos de una línea y la convierte en un diccionario, como se muestra en el siguiente ejemplo:

```
{'usuario': 'df843a74bba22d2b363e4e96f74b99391276e71dfea92e0a11de4fe552f38173',
'user_type': 1,
'start': 41,
'end': 50,
'año': 2017,
'mes': 4,
'dia': 1,
' hora': 1,
'code': '28005'}
```

Por tanto, el resultado final es una lista de diccionarios donde cada índice recoge la información de un solo viaje.

## 2.1 Estación más y menos concurrida

Para estudiar cuál es la estación más concurrida según el momento del día hemos decidido fraccionarlo en 3 intervalos:

- Mañana: [6 : 00 – 14 : 00)
- Tarde: [14 : 00 – 22 : 00)
- Noche: [22 : 00 – 6 : 00)

Una vez filtrado cada viaje en el intervalo correspondiente, creamos una tupla con la siguiente forma (*idunplug\_station*, 1), los agrupamos por el id, los ordenamos de mayor a menor en función del segundo elemento de la tupla, es

decir, en función del número de viajes que tienen esa estación como punto de partida y nos quedamos con el mayor.

Una metodología similar se ha usado para seleccionar la estación de unplug más deficiente, es decir, la estación menos concurrida.

## 2.2 Usuarios que empiezan y acaban en la misma estación

Para estudiar el porcentaje de usuarios cuyos trayectos son un ciclo, es decir, aquellos que empiezan y acaban en la misma estación, sin importar las estaciones en las que se ha parado entre medias seguimos la siguiente metodología. Mapreamos los datos en tuplas de la forma (`user_day_code, [idunplug_station,idplug_station]`) y nos quedamos con aquellos que han realizado más de un trayecto diario. Estos datos los volvemos a filtrar quedándonos con los que empiezan y acaban en la misma estación. Conseguimos así el porcentaje de usuarios que cumplen esa condición.

## 2.3 Usuarios que realizan un camino

Para estudiar el porcentaje de usuarios que realizan un camino, es decir, aquellos que cogen la bicicleta en la misma estación donde la habían dejado anteriormente, esto se debe cumplir en todos los trayectos que realicen. Utilizamos los datos anteriores expresados en forma de tupla y nos quedamos con aquellos que cumplen la condición descrita anteriormente.

## 2.4 Usuarios más y menos asiduos según su código postal

Al igual que antes hemos contado las estaciones más y menos asiduas ahora hacemos lo mismo con el código postal: agrupamos los datos en forma de tupla (`codigo_postal,1`) según el código, los ordenamos y nos quedamos con el mayor y el menor.

# 3 Resultados

Los datos obtenidos en el mes de abril de 2017 son:

## 3.1 Estación más y menos concurrida

Obtenemos que los id de las estaciones más y menos frecuentes para dejar la bicicletas por tramos horarios son:

	Mañana	Tarde	Noche
más frecuente	163	64	57
menos frecuente	2008	119	119

Figure 1: `idunplug_station` más y menos frecuente

Con un porcentaje de uso por franja horario de:

	Mañana	Tarde	Noche
porcentaje	29.77%	54.50%	15.73%

Figure 2: Porcentaje de uso por franjas horarias

### 3.2 Usuarios más y menos asiduos según su código postal

Obtenemos que los usuarios con código postal 28005 son los más asiduos a BICI-MAD mientras que los de código postal 28009 son los menos asiduos. Veamos que zonas de Madrid son:

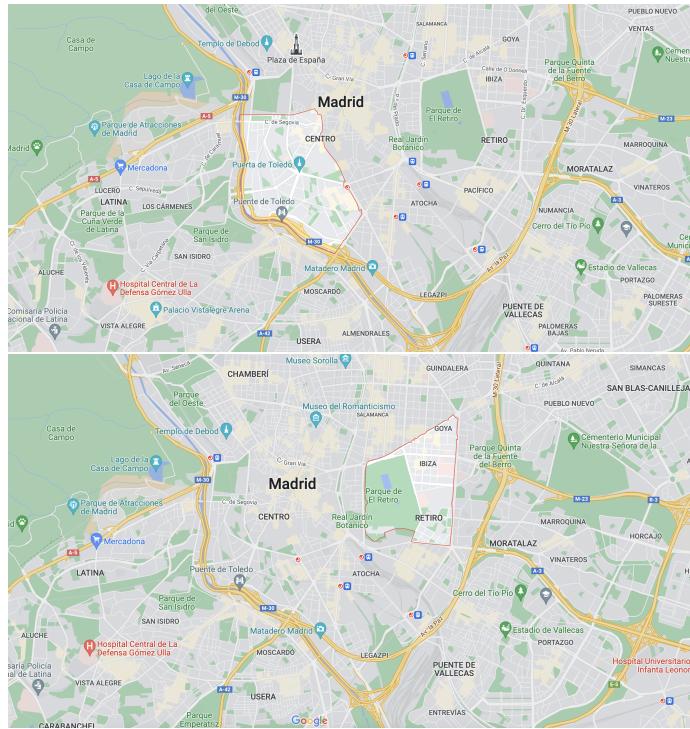


Figure 3: Zona más frecuentes (arriba) y zonas menos frecuentes (abajo)

### 3.3 Usuarios que empiezan y acaban en la misma estación

Obtenemos que más de un 26% de los usuarios que han utilizado la bici en abril de 2017 han repetido, es decir, han cogido la bicicleta más de una vez. Entre ellos, obtenemos que casi un 42% empiezan y terminan sus trayectos en bici en la misma estación.

### 3.4 Usuarios que realizan un camino

Entre los usuarios que repiten sabemos que más de un 47% de los usuarios realizan un camino.

Por último, nos ha parecido interesante estudiar como ha variado el porcentaje de usuarios que realizan ciclos o caminos en sus trayectos en bici desde el mes de abril de 2017 hasta diciembre de 2017.

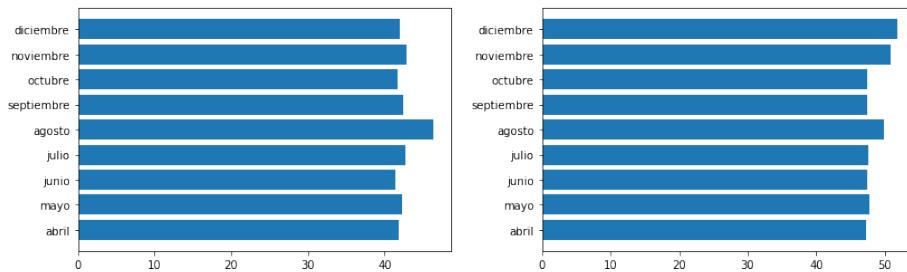


Figure 4: Evolución del porcentaje de usuarios que lleva a cabo ciclos y caminos

## 4 Conclusión

Concluimos que:

- El intervalo del día con más usuarios es la tarde, de 14:00 a 22:00.
- Más de un cuarto de los usuarios utiliza la bici más de una vez en el mes de abril de 2017.
- Agosto es el mes donde más usuarios realizan ciclos
- Diciembre es el mes donde más usuarios realizan caminos
- Los usuarios con código postal 28005 son los más asiduos y los de 28009 los que menos.