UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS MATEMÁTICAS

ESTUDIO SOBRE BICIMAD

Pablo Alonso Ciria, María Correas Crespo y Siria Catherine Íñiguez Brito

Programación Paralela

Curso académico 2021-2022

1. Introducción y Objetivos

A lo largo de este trabajo realizaremos un análisis sobre el uso de las bicicletas públicas BICIMAD en Madrid. En concreto, trataremos de examinar la afluencia en el uso de bicicletas en torno a Moncloa. Específicamente, centraremos nuestro estudio en el año 2019, durante el primer semestre, pues, en principio, lo consideramos un año adecuado, ya que se trata del anterior a la pandemia, por lo que entendemos que reflejará de una forma más realista el alquiler de estas bicicletas públicas.

En concreto, la muestra de nuestro estudio se basará en las estaciones cercanas a Moncloa: De

ESTACIÓN	CÓDIGO
Moncloa	116
Arcipreste de Hita I	117
Arcipreste de Hita II	118
Fernando el Católico 19	168
Cea Bermúdez, 59	160
Paseo de Moret	119

ellas solo tomaremos los usuarios que llegan a las mismas entre el mes de Enero y Junio de 2019. Nuestros objetivos son los siguientes:

- **Rango de edad:** buscamos agrupar el conjunto de viajes Bicimad con destino alguna de nuestras estaciones según el rango de edad de los usuario para conocer que tipo de público tienen.
- **Franja horaria:** tratamos de hacer un estudio de la afluencia de bicicletas en cada tramo horario para averiguar cuál es el periodo temporal en que se utiliza de forma más significativa estas bicicletas y que tienen como destino llegar a Moncloa.
- **Estación de salida:** queremos conocer de que distritos provienen los usuarios que acaban en Moncloa.
- **Duración del recorrido:** Por último, estamos interesados en conocer la duración media de todos los viajes realizados que tienen como destino Moncloa.

2. Material y metodología

Para llevar a cabo este estudio, nos apoyaremos en el conjunto de datos facilitados por BICIMAD: un sistema de alquiler público de bicicletas eléctricas de la ciudad de Madrid gestionado por la EMT. Todos los datos han sido extraídos de su página oficial de datos abiertos; https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1). Primero de todo, debemos conocer la descripción de los datos ofrecidos, así como los campos que la componen y su significado.

Los datos acerca del servicio BiciMAD, se ofrecen en formato JSON (formato de texto sencillo para el intercambio de datos). La información presentada es relativa a los movimientos de las bicicletas, es decir, el desplazamiento de una bici desde una estación de origen hasta una estación de destino donde se incluye información relativa a la estación de partida, estación de destino, datos del usuario de forma anonimizada... Describimos a continuación los datos concretos que utilizaremos para llevar a cabo nuestro propósito:

- **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:** franja horaria en la que se realiza el desenganche de la bicicleta. Se facilita la hora de inicio del movimiento, sin la información de minutos y segundos, por lo tanto, todos los movimientos iniciados durante la misma hora, tendrán el mismo dato de inicio.

- **travel_time:** tiempo total en segundos, entre el desenganche y el enganche de la bicicleta.
- **user_type:** número que indica el tipo de usuario que ha realizado el movimiento. Sus posibles valores son:

  0: No se ha podido determinar el tipo de usuario

  1: Usuario anual (poseedor de un pase anual)

  2: Usuario ocasional

  3: Trabajador de la empresa
- **ageRange:** número que indica el rango de edad del usuario que ha realizado el movimiento. Sus posibles valores son: 0: No se ha podido determinar el rango de edad del usuario

  1: El usuario tiene entre 0 y 16 años

  2: El usuario tiene entre 17 y 18 años

  3: El usuario tiene entre 19 y 26 años

  4: El usuario tiene entre 27 y 40 años

  5: El usuario tiene entre 41 y 65 años

  6: El usuario tiene 66 años o más

En nuestro caso particular, asociamos cada distrito con las identificaciones de las estaciones que le corresponden para su posterior utilización dentro de la implementación.

DISTRITO	identificación = correspondencia
Centro	1 <= iden <= 63
Moncloa	116 <= iden<= 126 ó iden = 132
Tetuán	139 <= iden <= 143
Chamberí	156 <= iden <= 161
Chamartín	144<= iden <= 155 ó 168 <= iden <= 173 ó iden = 138
Salamanca	92 <= iden <= 115 ó 162 <= iden <= 167
Retiro	64 <= iden <= 91
Arganzuela	En otro caso

Como cabía esperar, el conjunto de datos es demasiado voluminoso, por lo que será necesario hacer uso de la programación paralela, en concreto, utilizaremos como herramienta de trabajo Python, donde nos apoyaremos de librerías como pyspark (esencial para la Ciencia de Datos), json (debido al formato del conjunto de datos) y matplotlib (creación de gráficas).

Una vez recopilada la información e implementado las funciones necesarias, realizaremos una pequeña conclusión y valoración de los resultados, por lo que en este punto será de gran utilidad la librería matplotlib, para según el caso, utilizar gráficos de barras, de sectores, de líneas y así comprender mejor los resultados, además de complementarlos con imágenes visuales e intuitivas.

3. Implementación

La implementación de este trabajo es relativamente sencilla. Lo tenemos preparado para que podamos decidir cuantos meses queremos examinar, en caso de que no se explicite, toma solo el mes de mayo de 2019. Para determinar los meses que queremos utilizar basta con incluirlos como argumento al llamar al documento .py. Por ejemplo, si queremos leer los meses de marzo abril y junio tendriamos que escribir en la terminal: *python3 bicimad_grupo3.py 3,4,6* los meses separados por comas. Nuestro estudio se basa en los meses de enero a junio de 2019. El primer proceso que se realiza es la extracción de los datos de cada mes y lo filtramos para quedarnos solo con los viajes

con destino alguna de nuestras estaciones y acabamos uniendo los rdd de cada mes en un único rdd.

Luego el procedimiento es muy similar, para cada caso elegimos los campos que necesitamos con un map, agrupamos por el campo esencial y calculamos el número de viajes para cada campo esencial. Así, por ejemplo, para las franjas horarias, nos quedamos solo con los valores de *'unplug_hourTime'* y *'idplug_station'*, los agrupamos por la franja horaria y hacemos el conteo del número de viajes en cada franja. Un poco diferente es el cálculo del tiempo medio de trayecto en el que en vez de hacer el conteo calculamos la media de *'travel_time'*.

Por último, creamos las gráficas más útiles en cada caso gracias a la librería matplotlib. Veamos ahora los resultados obtenidos.

4. Discusión

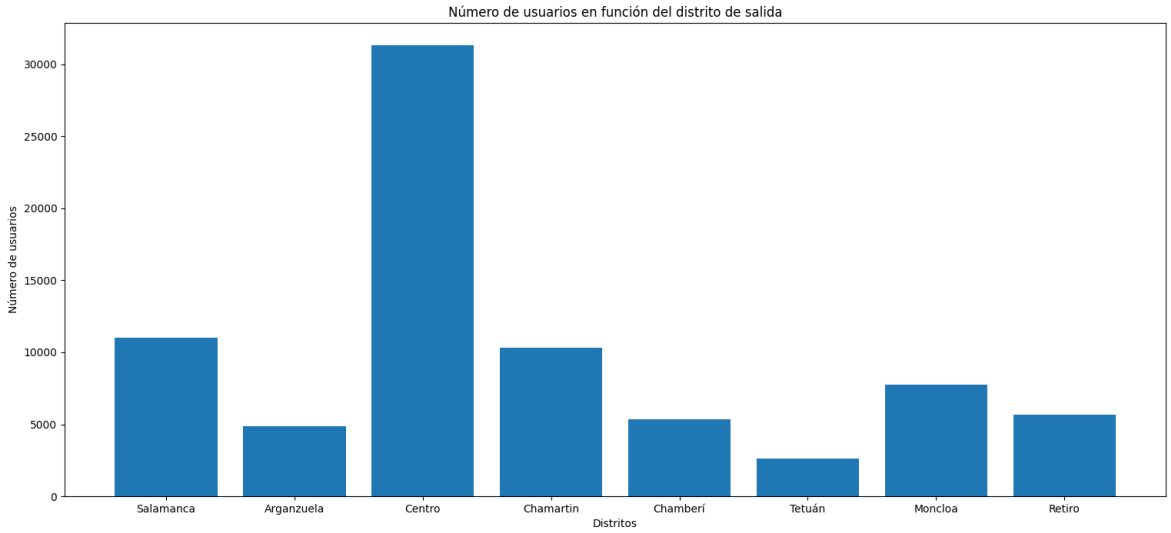
4.1. Estación de Salida

Los datos obtenidos sobre los distritos de origen son los siguientes:

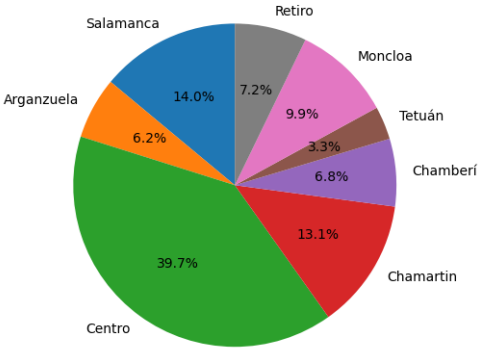
ESTACIÓN	Salamanca	Arganzuela	Centro	Chamartín
Nº USUARIOS	11005	4875	31303	10295

ESTACIÓN	Chamberí	Tetuán	Moncloa	Retiro
Nº USUARIOS	5337	2608	7771	5673

y obtenemos las gráficas: Como se puede observar, la gran mayoría de usuarios provienen del



distrito Centro, más concurrido y es el que tiene mayores límites de movilidad a motor. También se puede ver que una cantidad importante de viajes proceden del mismo distrito de Moncloa o del vecino Chamberí, luego son desplazamientos relativamente cortos. También podemos contemplar estos datos con un formato de gráfico de sectores, con el porcentaje de usuarios que le corresponden a cada distrito:



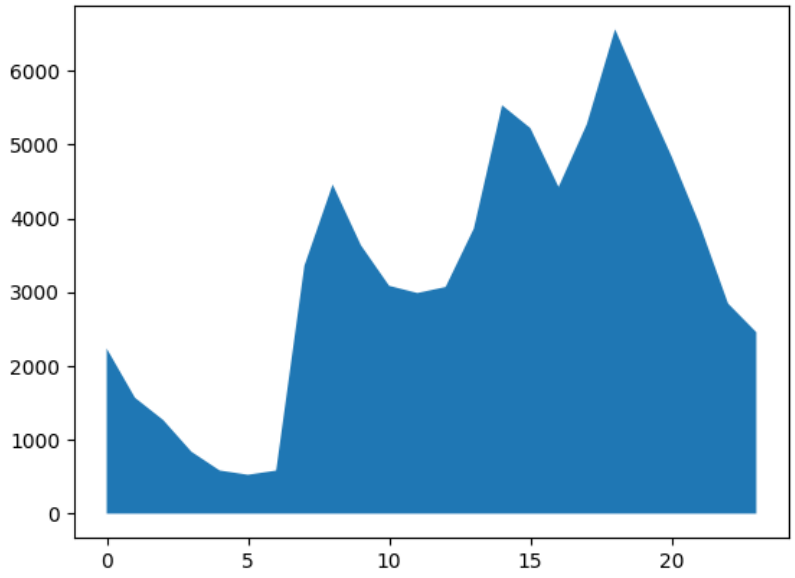
4.2. Franja Horaria

Los datos obtenidos para cada hora del día son los siguientes:

HORA	00	01	02	03	04	05	06	07	08	09	10	11
Nº USUARIOS	2238	1571	1269	839	585	527	584	3360	4461	3638	3088	2990

HORA	12	13	14	15	16	17	18	19	20	21	22	23
Nº USUARIOS	3072	3865	5535	5225	4427	5278	6565	5687	4842	3914	2848	2459

Y obtenemos el gráfico lineal de la evolución de la frecuentación de usuarios por hora. Los resultados



son coherentes, durante las horas nocturnas se observa un valle con baja afluencia de bicicletas mientras que la hora punta de usuarios se corresponde con las 18h de la tarde.

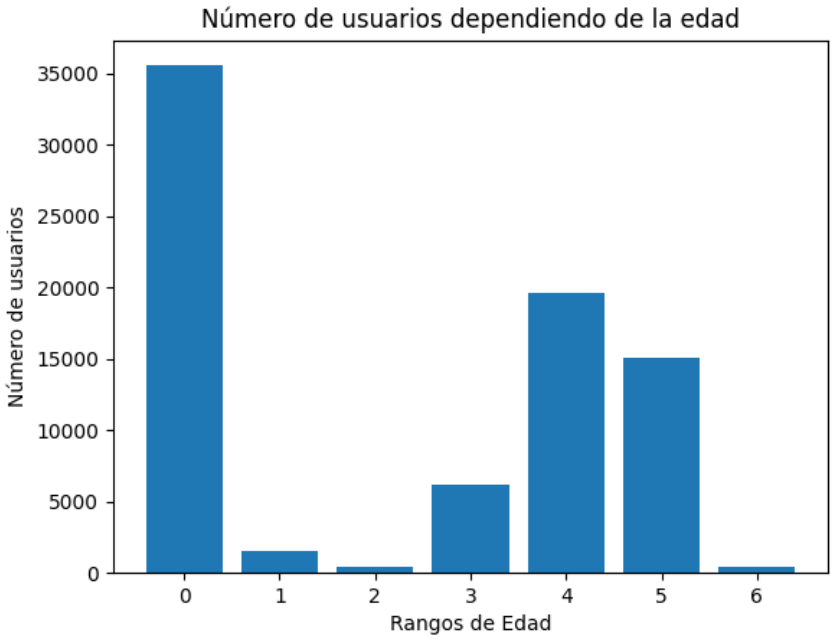
4.3. Rango de Edad

Los datos obtenidos para cada rango de edad es el siguiente:

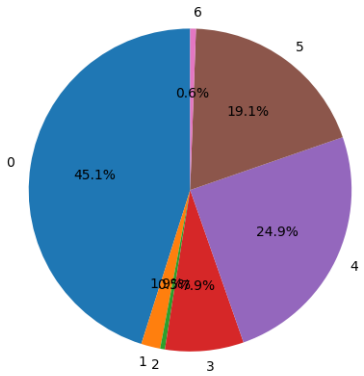
RANGO EDAD	0: desconocida	1: 0-16 años	2: 17-18 años	3: 19-26 años
Nº USUARIOS	35573	1510	375	6207

RANGO EDAD	4: 27-40 años	5: 41-65 años	6: más de 66 años
Nº USUARIOS	19672	15075	455

Y obtenemos las siguientes gráficas de barras y sectores: Como se puede observar, desgraciadamente



los datos obtenidos no tienen suficiente calidad como para poder extraer conclusiones. Más del 45 % de los usuarios no han registrado su edad y por lo tanto no podemos saber exactamente el rango de edad (puede ocurrir que los usuarios jovenes no se hayan molestado en registrar su edad o todo lo contrario). Entre los datos que sí conocemos podemos destacar que el rango de edad predominante es el del adulto joven (entre 27 y 49 años)



4.4. Duración media del recorrido

Finalmente, hemos tomado la duración media del trayecto de los usuarios que tienen como destino estaciones cercanas a Moncloa y nos hemos encontrado con que esa media ronda los 18 minutos. Una duración razonable para ir desde el distrito centro (distrito favorito para los usuarios) hasta Moncloa.