Base de datos BICIMAD

Silvia Centenera López-Pintor y Álvaro Gaitán Martín

Curso 2022-2023

Para este trabajo se utilizará la base de datos de movimientos del BICIMAD desde abril 2017 hasta junio 2018. La cual contiene información sobre los viajes hechos, entre esta información encontramos el rango de edad y el tiempo de viaje. Por eso decidimos estudiar el tiempo de viaje para cada rango de edad.

1. Explicación de los datos

El tiempo de viaje (por los registros que observamos asumimos que está en minutos) y la edad está clasificada en rangos, van del 0 al 6 y su significado es el siguiente:

ID	Edad
0	No se ha podido determinar la edad.
1	El usuario es menor o igual a 16 años.
2	El usuario tiene 17 o 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 65 años o más.

2. Problema a estudiar

Queremos saber cual es la media de tiempo de viaje de cada rango de edad, para poder saber quiénes hacen los viajes más largos. Esto podría ayudar a intuir el tipo de uso que se le da a la bicicleta. Acompañaremos la media de la desviación típica, el mínimo y el máximo para poder tener una visión más general y entender mejor esa media.

3. Implementación

Como se puede ver en el script, hay una función `main` que accede a hadoop, recoge los nombres de los ficheros, crea una carpeta donde se van a guardar los resultados y aplica otra función `process_file` a cada archivo. Esta última función lee el archivo, filtra filas vacías y las que tengan valores mayor que un máximo (más adelante se explicará por qué se tomó la decisión de agregar una cota superior) y llama a otra función `means_stds` que calcula media, desviación típica, mínimo y máximo. Después se recogen estos resultados, se almacenan en un diccionario y se guarda en la carpeta creada con el nombre "means" + "nombre original".

4. Resultados

Cuando se hizo una primera prueba del script con los datos, el resultado fue una media muy alta y una desviación típica completamente anómala por el contexto de los datos. Viendo el máximo y mínimo apreciamos que había registros con un tiempo de viaje muy grandes, del orden de miles de horas. Desconocemos la causa de estos y la procedencia que pueden tener. Al ser la media y la desviación típica medidas muy susceptibles a los valores extremos se veían muy afectadas por estos datos.

La solución que pensamos es que al llamar al script por la terminal se le pase como argumento la cota superior de tiempo de viaje que se piensa considerar. De esta manera la persona que haga la consulta puede decidir hasta que número tiene sentido, según el estudio que quiera hacer.

Un ejemplo de la ejecución para el archivo `publicbicimad201708_movements.json` con cota superior de 500 minutos da el siguiente resultado:

```
{"0": {"mean": 326.5045398377478, "std_dev": 127.71922255535037, "max": 499, "min": 0},
"1": {"mean": 24.21931818181818, "std_dev": 68.99451107870485, "max": 498, "min": 1},
"2": {"mean": 332.9246901811249, "std_dev": 106.5599007731819, "max": 499, "min": 8},
"3": {"mean": 239.60058577405857, "std_dev": 173.57898284437385, "max": 499, "min": 2},
"4": {"mean": 326.2321154007436, "std_dev": 129.24568557282691, "max": 499, "min": 1},
"5": {"mean": 313.49554498559655, "std_dev": 140.21266495777502, "max": 499, "min": 1},
"6": {"mean": 361.8, "std_dev": 100.53345139951877, "max": 499, "min": 96}}
```


Un ejemplo de la ejecución para el archivo `publicbicimad201709_movements.json` con cota superior de 500 minutos da el siguiente resultado:


```
{"0": {"mean": 348.0134625312736, "std_dev": 103.41951750957985, "max": 499, "min": 2},
"1": {"mean": 34.71516741243635, "std_dev": 88.23437870961705, "max": 499, "min": 0},
"2": {"mean": 331.705078125, "std_dev": 108.9069642059617, "max": 499, "min": 4},
"3": {"mean": 289.8437545335848, "std_dev": 153.8767590586751, "max": 499, "min": 1},
"4": {"mean": 316.24264403403555, "std_dev": 142.11812883891324, "max": 499, "min": 2},
"5": {"mean": 326.23984610889255, "std_dev": 133.3470477415153, "max": 499, "min": 2},
"6": {"mean": 356.44621513944224, "std_dev": 95.68703192940312, "max": 499, "min": 96}}
```


Se puede apreciar que entre estos dos archivos concretos el resultado no varía mucho, se pueden sacar mas o menos las mismas conclusiones. El rango de edad que más minutos de media utiliza este servicio de bicicletas es el 6, aunque bastante seguido del 2 y el 5. Los que más varían en los tiempos son el rango 3 y los que menos el rango 1, este último destacando con una media muy baja comparada con el resto.

5. Utilidad

A parte de observar cual es la media de tiempo por rango de edad, al procesar varios archivos a la vez, este script puede servir para ver cómo evolucionan estos datos en los diferentes meses. Por ejemplo, podríamos pensar que durante el curso escolar los rangos más jóvenes tendrán más actividad, y que los rangos más mayores tendrán una media más estable durante el año. Esto se podría saber aplicando el código a los datos adecuados y observando el resultado.