Facultad de Ciencias Matemáticas

Programación Paralela

curso 2021-2022

Práctica 4: Análisis de datos de BiciMAD con Spark

Esther Arribas García, Lucía Bragado Pérez, Álvaro Seco Silva

Mayo, 2022

Índice

1. Introducción

En esta práctica se trabaja en el planteamiento, diseño e implementación de una solución a un problema de análisis de datos reales utilizando Spark. El conjunto de datos de estudio es el proporcionado por el ayuntamiento de Madrid sobre el uso del sistema de bicicletas de préstamo BiciMAD.

Realizaremos una comparativa del uso de este servicio pre y post pandemia, principalemente el tiempo de uso según el grupo de edad, la cantidad de salidas y llegadas según el barrio, y el número de usuarios únicos totales al mes.

Para visualizar los resultados obtenidos, mostraremos histogramas y tablas.

2. Motivación y definición del problema a resolver

Con vistas a realizar un análisis de datos que pueda ofrecer resultados interesantes desde el punto de vista social, consideramos que una buena idea sería hacer una comparativa del uso de BiciMAD antes y después de la pandemia por COVID-19.

De esta manera, seleccionamos los datos correspondientes a los meses de Mayo a Octubre (ambos inclusive) de los años 2019 y 2020.

Analizamos los siguientes puntos de dicho conjunto de datos:

- Tiempo medio de uso de la bicicleta en función del grupo de edad (6 meses pre-confinamiento VS 6 meses post-confinamiento).
- Cantidad de salidas y llegadas por barrio (pre-confinamiento y post-confinamiento).
- Cantidad de usuarios (sin repeticiones) de BiciMAD cada mes (pre-confinamiento VS post-confinamiento).

3. Datos empleados

Los datos que hemos empleado en el estudio están disponibles el siguiente enlace Datos BiciMAD proporcionados por el Ayuntamiento de Madrid.

De este conjunto de datos extraemos los correspondientes a los meses que nos interesan. Los datos de cada mes vienen dados por separado en un archivo con extensión *.json*.

Cada línea de estos ficheros es un objeto de JavaScript, con los siguientes atributos:

- **_id**: identificador del movimiento
- **user_day_code**: código del usuario (para una misma fecha todos los movimientos de un mismo usuario tienen el mismo código).
- **idplug_base**: número de la base en la que se engancha la bicicleta.
- **user_type**: número que indica el tipo de usuario que ha realizado el movimiento. Como los posibles valores cambian del año 2019 al 2020, no tendremos en cuenta este atributo en el análisis comparativo pre-post pandemia.
- **idunplug_base**: número de la base de la que se desengancha la bicicleta.
- **travel_time**: tiempo total en segundos, entre el desenganche y el enganche de la bicicleta.
- **idunplug_station**: número de la estación de la que se desengancha la bicicleta.

- **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

- **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)
- **zip_code**: número de la estación de la que se desengancha la bicicleta.

4. Diseño e implementación de la solución

Para trabajar de forma paralela con los datos, usaremos *rdd*.

En primer lugar, para el análisis del tiempo medio en función de la edad y las salidas y llegadas por barrio creamos dos rdd, uno para cada año (*rdd19* y *rdd20*), uniendo los archivos descargados de cada mes.

En segundo lugar, para el apartado de cantidad de usuarios por mes pre y post confinamiento, consideramos dos listas, una para cada año, *lrdd19* y *lrdd20* cuyos elementos son los rdd independientes correspondientes a cada mes(sin unir, a diferencia de *rdd19* y *rdd20*).

Para el análisis del tiempo medio en función del grupo de edad en cada año, de cada línea del rdd (*rdd19* o *rdd20*) obtenemos la tupla (grupoEdad, tiempoViaje) a través de un map (*mapper_edad*). Después agrupamos estos datos por grupo de edad con un groupByKey obteniendo tuplas (grupoEdad, listaTiemposViaje). Finalmente, para cada grupo de edad hacemos la media de los elementos de la listaTiemposViaje para obtener la media por grupos de edad. Representamos el resultado en un histograma doble.

Para analizar el número de salidas y llegadas según el barrio, en primer lugar identificamos cada código de estación con un barrio gracias a la información proporcionada por el Ayuntamiento de Madrid en el siguiente enlace. Creamos un diccionario de claves los nombres de los barrios y valores los códigos de estación correspondientes a cada barrio. Para cada línea del rdd correspondiente obtenemos la tupla (codEstaciónSalida, 1) a través de un map (*mapper_unplug_station*). Después, usamos el diccionario que asocia códigos de estación con barrios para transformar cada codEstaciónSalida en un nombre de barrio, a través de un map (*map_asociar_barrio*). A continuación, usamos un un groupByKey para tagrupar por barrio obteniendo tuplas (barrio, listaDeUnos), donde hay tantos unos en la listaDeUnos como salidas haya habido en dicho barrio. Por tanto, aplicamos un map que transforme cada listaDeUnos en la suma de sus elementos. Así ya tenemos la tabla de tuplas (barrio, númeroDeSalidas). Filtramos las tuplas donde el barrio es distinto de "nada", asignado al obtener un código de estación no registrado en la página del Ayuntamiento. Por último, hacemos un sortBy con el opuesto de númeroDeSalidas para ordenar los barrios de mayor a menor número de salidas. El número de llegadas por barrio se analiza de manera análoga.

Para el análisis de la cantidad de usuarios por mes pre y post confinamiento, para cada lista *lrdd19* y *lrdd20* y cada elemento de dichas listas (que corresponde a un mes) transformamos cada línea del fichero rdd a una única variable códigoUsuario a través de un map (*mapper_usuario_unico*). Despúes, eliminamos códigos repetidos (pues queremos contar usuarios diferentes) con un groupByKey. Una vez eliminadas las repeticiones, obtenemos la cantidad de usuarios como la longitud de la lista de códigos después de la agrupación.

5. Instrucciones para ejecutar el código

Para poder obtener los resultados se necesita ejecutar el código *bicimad.py* desde la misma carpeta donde estén guardados los archivos *json* correspondientes a los meses de Mayo a Octubre de 2019 y 2020, con el nombre que reciben al ser descargados desde Datos BiciMAD, esto es, los siguientes 12 archivos:

201905_Usage_Bicimad.json,

201906_Usage_Bicimad.json,

201907_movements.json,

201908_movements.json,

201908_movements.json,

201909_movements.json,

201910_movements.json,

202005_movements.json,

202006_movements.json,

202007_movements.json,

202008_movements.json,

202009_movements.json,

202010_movements.json.

6. Evaluación de resultados

La comparativa pre-post confinamiento del tiempo medio de uso de la bicicleta en función del grupo de edad se muestra en el siguiente diagrama de barras:

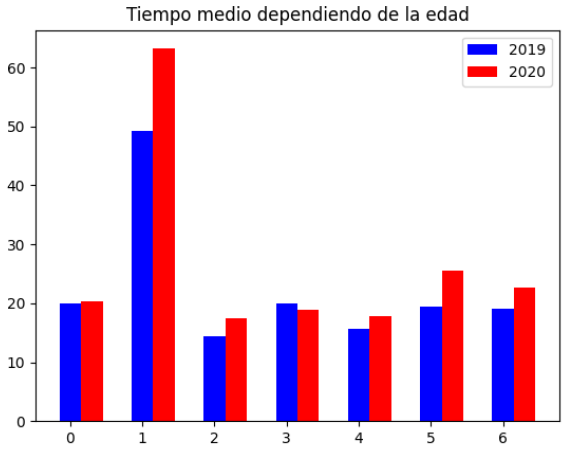

Figura 1: Tiempo medio (en minutos) de uso según grupo de edad pre-post pandemia

Notamos que hay el grupo de edad 1, correspondiente a niños y adolescentes (de 0 a 16 años) es el grupo que usa el servicio de BiciMAD por mayor tiempo tanto en pre-confinamiento (unos 50 min de media) como en post-confinamiento (una hora de media). El resto de grupos de edad obtienen tiempos medios muy similares entre sí en ambos años (en el rango de los 15 min a los 25min).

En cuanto a la comparativa pre-post pandemia, el tiempo medio de uso después de la pandemia suele ser ligeramente superior al de antes de la pandemia, salvo para el grupo de edad 3 (de 19 a 26 años) en la que el tiempo medio es mayor antes de la pandemia. Estos resultados resultan coherentes, pues los ciudadanos que decidían usar el servicio de Bicimad posiblemente repetían los mismos trayectos que hacían antes de la pandemia, obteniendo tiempos medios similares.

Las siguientes tablas muestran la cantidad de salidas y llegadas por barrio en los meses estudiados de 2019 y 2020.

	Llegadas 19
Embajadores	165902
Universidad	143427
Justicia	135693
Rios_Rosas	122009
Recoletos	119327
Acacias	117393
Palacio	92810
El_Viso	91615
Sol	90613
Goya	83936
Jerónimos	76412
Cuatro_Caminos	61470
Cortes	61028
Palos_de_Moguer	60231
Arguelles	58317
Niño_Jesus	55982
Castellana	39922
Casa_de_Campo	39048
Gaztambide	39010
Almagro	38611
Chopera	36654
Pacifico	35142
Ibiza	34021
Estrella	33277
Delicias	33097
Adelfas	33016
Hispanoamerica	32785
Arapiles	28071
Fuente_del_Ebro	25400
Lista	25047
Guindalera	20934
Trafalgar	19711
Nueva_españa	19161
Atocha	14973
Castillejos	12564
Imperial	10298
Concepcion	9546
Moscardo	6803
Castilla	5968
Valdeacederas	5705
San_Diego	5474
San_Isidro	4016
Prosperidad	4002
Marroquina	3697
Vallehermoso	3580
Media_Legua	3291
Pueblo_Nuevo	2840
Bellas_vistas	2447
Ciudad_Jardin	1878
La_Paz	1536
San_Pascual	1488
Puerta_del_Angel	1416
Ventas	1376
Berruguete	999

	Llegadas 20
Embajadores	142143
Universidad	122384
Justicia	106303
Rios_Rosas	102324
Recoletos	96892
Acacias	92217
El_Viso	88250
Palacio	77044
Sol	74248
Goya	65745
Jeronimos	55209
Palos_de_Moguer	51997
Cortes	50995
Arguelles	48102
Fuente_del_Ebro	46023
Cuatro_Caminos	43450
Almagro	42596
Delicias	41395
Niño_Jesus	38870
Gaztambide	35427
Ibiza	34775
Guindalera	33343
Pacifico	32446
Lista	32225
Castellana	32071
Chopera	32063
Casa_de_Campo	30632
Estrella	30202
Hispanoamerica	28546
Arapiles	24791
Atocha	22513
Castilla	20831
Adelfas	20467
Nueva_españa	20325
Vallehermoso	19551
Media_Legua	17072
Trafalgar	14652
Concepcion	14584
San_Isidro	11158
Ciudad_Jardin	10201
Moscardo	9812
La_Paz	9691
Valdeacederas	9565
Puerta_del_Angel	9225
San_Diego	8948
Berruguete	8536
San_Pascual	8422
Bellas_vistas	7925
Pueblo_Nuevo	6968
Castillejos	6790
Marroquina	6773
Imperial	6580
Prosperidad	6477
Ventas	5066
Pavones	3356
Carmenes	154
Numancia	27

	Salidas 19
Embajadores	164792
Universidad	142889
Justicia	135280
Rios_Rosas	120511
Recoletos	119661
Acacias	117070
El_Viso	93353
Palacio	92461
Sol	90267
Goya	84915
Jeronimos	76225
Cuatro_Caminos	62104
Cortes	61200
Palos_de_Moguer	59960
Arguelles	58455
Nino_Jesus	55282
Castellana	40479
Casa_de_Campo	39331
Gaztambide	39210
Almagro	38811
Chopera	36833
Pacifico	35095
Ibiza	33857
Delicias	33132
Hispanoamerica	33094
Adelfas	32947
Estrella	32682
Arapiles	28280
Lista	25380
Fuente_del_Ebro	25285
Guindalera	20849
Trafalgar	19883
Nueva_espana	19097
Atocha	14941
Castillejos	12749
Imperial	10442
Concepcion	9519
Moscardo	6831
Castilla	5920
Valdeacederas	5609
San_Diego	5437
San_Isidro	4077
Prosperidad	4028
Marroquina	3653
Vallehermoso	3619
Media_Legua	3268
Pueblo_Nuevo	2850
Bellas_vistas	2492
Ciudad_Jardin	1867
La_Paz	1573
San_Pascual	1535
Puerta_del_Angel	1496
Ventas	1385
Berruguete	996

	Salidas 20
Embajadores	141588
Universidad	122298
Justicia	106178
Rios_Rosas	101586
Recoletos	97636
Acacias	92081
El_Viso	88589
Palacio	76970
Sol	74040
Goya	66201
Jeronimos	55253
Palos_de_Moguer	51629
Cortes	51346
Arguelles	48454
Fuente_del_Ebro	45624
Cuatro_Caminos	44721
Almagro	43071
Delicias	41138
Niño_Jesus	38803
Gaztambide	35753
Ibiza	34661
Guindalera	33450
Lista	32654
Pacifico	32330
Castellana	32100
Chopera	31945
Casa_de_Campo	30757
Estrella	29829
Hispanoamerica	28824
Arapiles	25034
Atocha	22523
Castilla	20634
Adelfas	20529
Nueva_españa	20372
Vallehermoso	19516
Media_Legua	16803
Trafalgar	14688
Concepcion	14511
San_Isidro	10920
Ciudad_Jardin	10160
La_Paz	9511
Valdeacederas	9456
Puerta_del_Angel	9384
Moscardo	9152
San_Diego	8879
Berruguete	8466
San_Pascual	8396
Bellas_vistas	7896
Castillejos	7264
Marroquina	6711
Imperial	6620
Pueblo_Nuevo	6479
Prosperidad	6460
Ventas	5037
Pavones	3274
Carmenes	162
Numancia	38

Observamos que los barrios más concurridos antes de la pandemia (Embajadores, Universidad, Justicia...) también lo son después de la misma, aunque con menos salidas y llegadas. Lo mismo ocurre para los barrios menos concurridos.

El siguiente histograma doble muestra la comparativa pre y post pandemia de la cantidad de usuarios únicos del servicio de BiciMAD:

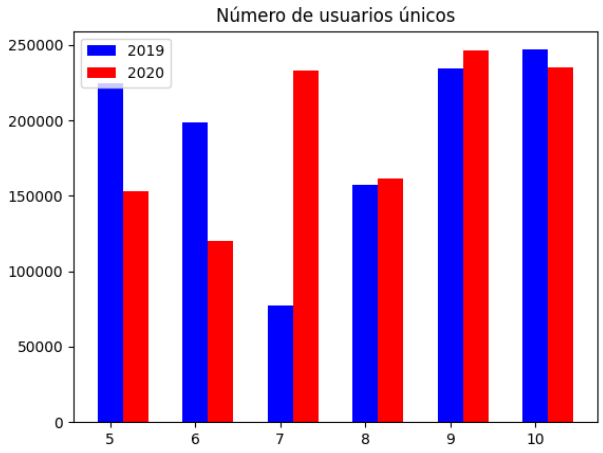

Figura 2: Cantidad de usuarios únicos de Bicimad en cada mes pre y post pandemia

Esta gráfica nos proporciona mucha información desde el punto de vista social. En primer lugar, en el mes de Mayo, correspondiente al momento en que se volvió a las calles después del confinamiento, había cierta reticencia a usar Bicimad, siendo la diferencia respecto al año anterior mayor que 50000 usuarios. Esta diferencia se mantuvo durante el mes de Junio.

Sin embargo, en el mes de Julio observamos un gran cambio: el uso en 2019 cae drásticamente (posiblemente debido a la gente abandonando la capital por vacaciones) mientras que en 2020 aumentan los usos en unos 100000 usuarios, probablemente por el miedo de la ciudadanía a salir de vacaciones y exponerse a un mayor contagio, usar la bicicleta fue un gran refugio para muchas personas. Además, el transporte con bicicleta, al ser al aire libre, minimiza el riesgo de contagio con respecto a otros medios de transporte.

Tras el mes vacacional de Julio, en los sucesivos meses de Agosto, Septiembre y Octubre tras la pandemia se consiguió equilibrar la cantidad de usuarios que había el año anterior, ratificando la vuelta a la normalidad.

7. Conclusión

En definitiva, en esta práctica se ha trabajado con la librería *pyspark* de Python para hacer un análisis comparativo del servicio Bicimad antes y después de la pandemia.

Para ello hemos trabajado con RDDs (Resilient Distributed Datasets), que sirven para operar de forma paralela con un conjunto inmutable y particionado de datos.

Los resultados obtenidos nos permiten obtener conclusiones significativos de un servicio real en nuestra ciudad.