Práctica Spark Bicimad

Cristina Sintes, Diana Pérez-Chirinos, Eduardo Garza
Grupo 16

1 Explicación del problema

Para esta práctica hemos desarrollado dos códigos distintos: bicimad1.py, bicimad2.py.
bicimad1.py es la entrega principal que hace un estudio de datos más genérico. Este archivo ha sido probado en cluster sin problemas. El archivo secundario bicimad2.py realiza un estudio de datos más específico. No hemos podido acceder al clúster a fecha 22/05 desde las 17horas aproximadamente, por lo tanto este segundo código ha sido ejecutado con éxito desde la implentación de pyspark en nuestros ordenadores.

En primer lugar haremos una redacción del código bicimad1.py.

Con esta práctica realizaremos distintos estudios:

- Estudiar la edad media de adquisición de las bicicletas por estación tanto diario como en fin de semana.

- Estudiar desde cada estación de donde sale una bicicleta a qué estación suele llegar. También distinguiendo en días de diario y fin de semana.

Se podrían realizar más estudios estadísticos como qué días se cogen más bicicletas o cuál es la estación de la que mas bicicletas se sacan y cuál es la que mas recibe. Hemos seleccionado estos estudios ya que son los que más interesantes nos han parecido.

2 Instrucciones para ejecutar los programas

Explicación del código:

- Inicializamos sparkconf y sparkcontext para cada uno de los datos

- info-date: Calcula la fecha del día que estamos tratando. En realidad, lo que nos va a interesar es saber el día de la semana. Obtenemos el año, semana y día.
- day-traductor: Traduce al día de la semana en el que estamos, lo usamos en info-date
- get-days: Se podría optimizar el código utilizando fromisoformat y isoweekday, pero el server no permite esos atributos de la librería datetime.
- get-edades: Con rdd.map conseguimos en orden los datos de rango de edad, dia, estación de origen y estación de destino. Se omite el rango de edad 0 ya que son datos que no se han podido determinar la edad del usuario.
- is-return: Son los valores que devuelve el código. Según la estación de origen estudiamos, el rango de edad que mas usa bicimad en dicha estación y el destino que tiene mayor frecuencia. Hacemos distinción entre día laboral y fin de semana.
- edad-estacion-final: esta función devuelve el máximo de valores. Una vez tenemos todos los datos del año 2020, los agrupamos y concluimos qué rango de edad usa más las bicicletas y qué destino es el mas frecuente desde cada una de las estaciones a lo largo del año.
- main: tenemos como parámetro los meses, que tenemos que introducir como los archivos json. Añadimos mes a mes y ejecutamos todo el programa hasta que lo convertimos en una única salida.

Se agrupan los datos por estaciones. Están estructurados como una lista de tuplas donde el primer término es el sector de semana en que nos encontremos (laboral o fin de semana) y el segundo es el rango de edad.

Los rangos de edad que se utilizan 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.

Para ejecutarlo debemos poner: python3 bicimad1.py 202001-movements.json 202002-movements.json ... 202012-movements.json.

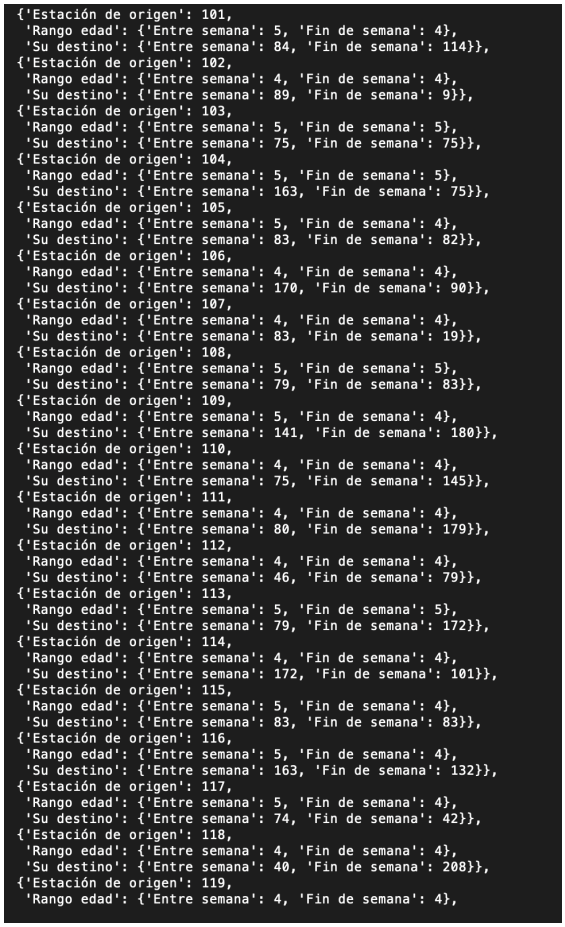
Los archivos 2020i-movements.json i en 1,...,12 son los datos de movimiento sacados de la página oficial de datos estadísticos de EMT Madrid: `https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1)`. De todos los datos que hay recogidos en esta página, utilizaremos los que se llaman Datos de uso de "nombre mes" de 2020.

3 Resultados y conclusiones

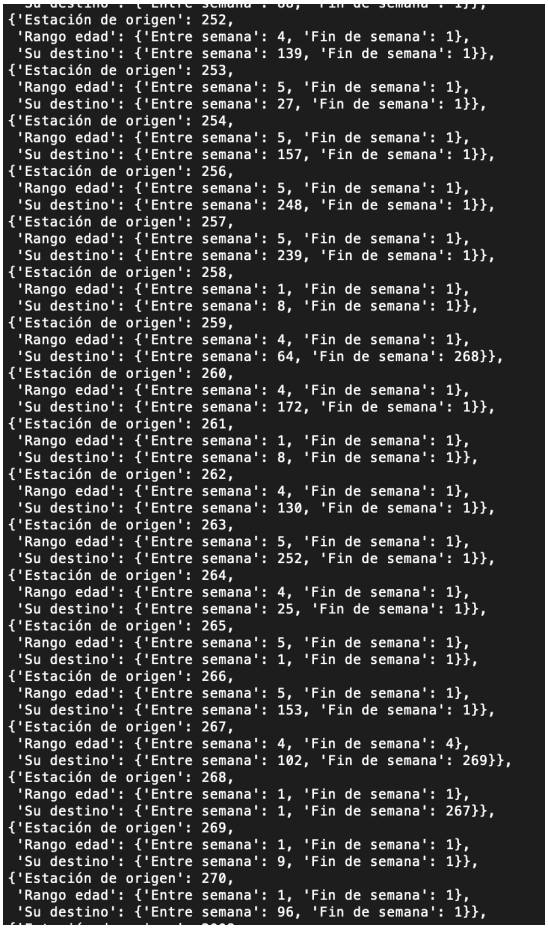
A partir de estos resultados se podrían implementar distintas medidas. Por ejemplo teniendo en cuenta cuál es la estación que más bicicletas recibe, se podría dejar en un estado inicial más vacío. La estación que más bicicletas emite debería tener un estado inicial con más bicicletas disponibles. En los barrios donde haya más tránsito de bicicletas, tanto de salida como de entrada añadir más paradas. Aprovechar los días que menos tránsito hay para sacar de servicio bicicletas averiadas y repararlas, etc. Una vez obtenidos y depurados los datos, las posibilidades de aplicación son enormes.

Como resultado al ejecutar el programa obtenemos:

```
crsintes@clus-01:~ $ python3 bicimad1.py 202001_movements.json 202002_movements.json 202003_movements.json 202004_movements.json 202005_movements.json 202006_movements.json 202007_m
ovements.json 202008_movements.json 202009_movements.json 202010_movements.json 202011_movements.json 202012_movements.json
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark/spark-3.0.1-bin-hadoop3.2/jars/spark-unsafe_2.12-3.0.1.jar) to constructor java.nio.DirectByt
eBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
2022-05-21 21:08:48,334 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
2022-05-21 21:08:56,111 WARN spark.SparkContext: Please ensure that the number of slots available on your executors is limited by the number of cores to task cpus and not another cu
stom resource. If cores is not the limiting resource then dynamic allocation will not work properly!
Adding 202001_movements.json
Adding 202002_movements.json
Adding 202003_movements.json
Adding 202004_movements.json
Adding 202005_movements.json
Adding 202006_movements.json
Adding 202007_movements.json
Adding 202008_movements.json
Adding 202009_movements.json
Adding 202010_movements.json
Adding 202011_movements.json
Adding 202012_movements.json
............Starting computations.............
[{'Estación de origen': 1,
  'Rango edad': {'Entre semana': 5, 'Fin de semana': 4},
  'Su destino': {'Entre semana': 65, 'Fin de semana': 175}},
 {'Estación de origen': 2,
  'Rango edad': {'Entre semana': 5, 'Fin de semana': 4},
  'Su destino': {'Entre semana': 175, 'Fin de semana': 175}},
 {'Estación de origen': 3,
  'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
  'Su destino': {'Entre semana': 43, 'Fin de semana': 77}},
 {'Estación de origen': 4,
  'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
  'Su destino': {'Entre semana': 160, 'Fin de semana': 211}},
 {'Estación de origen': 5,
  'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
  'Su destino': {'Entre semana': 19, 'Fin de semana': 19}},
 {'Estación de origen': 6,
  'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
  'Su destino': {'Entre semana': 14, 'Fin de semana': 149}},
 {'Estación de origen': 7,
  'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
  'Su destino': {'Entre semana': 149, 'Fin de semana': 149}},
 {'Estación de origen': 8,
  'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
  'Su destino': {'Entre semana': 75, 'Fin de semana': 157}},
 {'Estación de origen': 9,
  'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
  'Su destino': {'Entre semana': 149, 'Fin de semana': 149}},
 {'Estación de origen': 10,
```



```
{'Estación de origen': 101,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 84, 'Fin de semana': 114}},
{'Estación de origen': 102,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 89, 'Fin de semana': 9}},
{'Estación de origen': 103,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 5},
 'Su destino': {'Entre semana': 75, 'Fin de semana': 75}},
{'Estación de origen': 104,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 5},
 'Su destino': {'Entre semana': 163, 'Fin de semana': 75}},
{'Estación de origen': 105,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 83, 'Fin de semana': 82}},
{'Estación de origen': 106,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 170, 'Fin de semana': 90}},
{'Estación de origen': 107,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 83, 'Fin de semana': 19}},
{'Estación de origen': 108,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 5},
 'Su destino': {'Entre semana': 79, 'Fin de semana': 83}},
{'Estación de origen': 109,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 141, 'Fin de semana': 180}},
{'Estación de origen': 110,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 75, 'Fin de semana': 145}},
{'Estación de origen': 111,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 80, 'Fin de semana': 179}},
{'Estación de origen': 112,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 46, 'Fin de semana': 79}},
{'Estación de origen': 113,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 5},
 'Su destino': {'Entre semana': 79, 'Fin de semana': 172}},
{'Estación de origen': 114,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 172, 'Fin de semana': 101}},
{'Estación de origen': 115,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 83, 'Fin de semana': 83}},
{'Estación de origen': 116,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 163, 'Fin de semana': 132}},
{'Estación de origen': 117,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 74, 'Fin de semana': 42}},
{'Estación de origen': 118,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 40, 'Fin de semana': 208}},
{'Estación de origen': 119,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
```



```
{'Estación de origen': 252,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 139, 'Fin de semana': 1}},
{'Estación de origen': 253,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 27, 'Fin de semana': 1}},
{'Estación de origen': 254,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 157, 'Fin de semana': 1}},
{'Estación de origen': 256,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 248, 'Fin de semana': 1}},
{'Estación de origen': 257,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 239, 'Fin de semana': 1}},
{'Estación de origen': 258,
 'Rango edad': {'Entre semana': 1, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 8, 'Fin de semana': 1}},
{'Estación de origen': 259,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 64, 'Fin de semana': 268}},
{'Estación de origen': 260,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 172, 'Fin de semana': 1}},
{'Estación de origen': 261,
 'Rango edad': {'Entre semana': 1, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 8, 'Fin de semana': 1}},
{'Estación de origen': 262,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 130, 'Fin de semana': 1}},
{'Estación de origen': 263,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 252, 'Fin de semana': 1}},
{'Estación de origen': 264,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 25, 'Fin de semana': 1}},
{'Estación de origen': 265,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 1, 'Fin de semana': 1}},
{'Estación de origen': 266,
 'Rango edad': {'Entre semana': 5, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 153, 'Fin de semana': 1}},
{'Estación de origen': 267,
 'Rango edad': {'Entre semana': 4, 'Fin de semana': 4},
 'Su destino': {'Entre semana': 102, 'Fin de semana': 269}},
{'Estación de origen': 268,
 'Rango edad': {'Entre semana': 1, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 1, 'Fin de semana': 267}},
{'Estación de origen': 269,
 'Rango edad': {'Entre semana': 1, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 9, 'Fin de semana': 1}},
{'Estación de origen': 270,
 'Rango edad': {'Entre semana': 1, 'Fin de semana': 1},
 'Su destino': {'Entre semana': 96, 'Fin de semana': 1}},
```

Ahora vamos a explicar el código bicimad2.py.

Esta práctica lo que realiza es:

- Calcula el número total de bicicletas que se cogen cada día de la semana distinguiendo los rangos de edad, y el rango de edad que más bicicletas saca cada día.

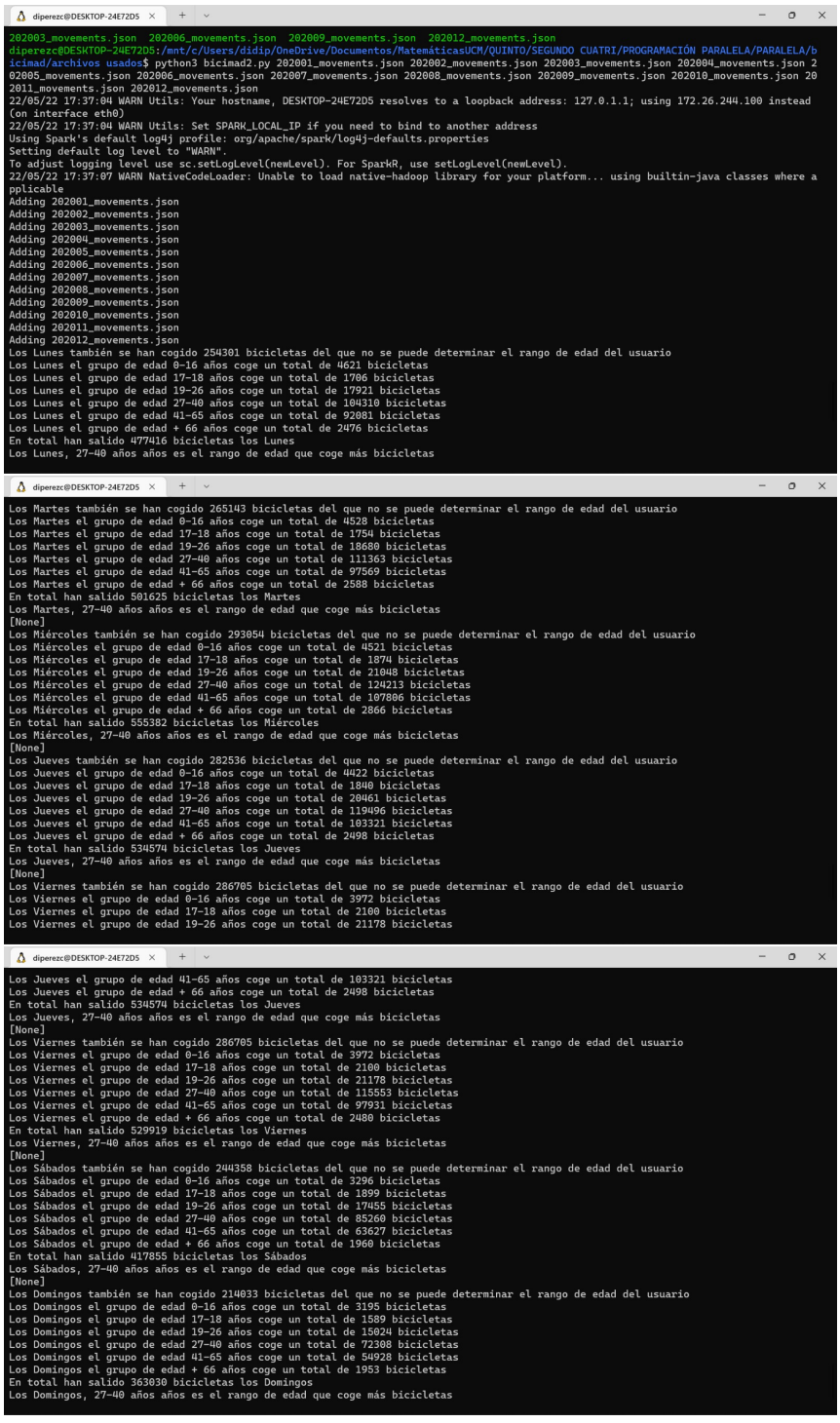
Como funciones nuevas tenemos:

- get-info: Filtra los datos según los días de la semana.
- cuantos-salen-por-dia-edad: Recoge en una lista los datos diarios de adquisiciones de bicletas separando por grupo de edad. Y obtenemos el índice del máximo grupo de edad, es decir, el grupo de edad que más bicicleteas ha obtenido ese día. Obviando el grupo de edad 0 que no nos aporta información. Devuelve una 3-tupla que tiene en la posicion 0 el día de la semana, en la posición 1 la lista con los datos recogidos y en la posición 2 el índice del máximo.
- is-return: Devuelve la información obtenida anteriormente de manera ordenada.

Para ejecutarlo será de manera análoga a bicimad1.py, es decir, mismo procedimiento y mismos archivos.

Como resultados y conclusiones:

- Se observa que el grupo de edad que más uso hace de las bicicletas es 27-40 años indistintamente del día de la semana. También se observa que el que menos uso hace es el de 17-18 años, quizás debido a la franja tan reducida de años que tiene ese rango. Seguido de el grupo +66, cosa que cabría esperar.
- Los días que más uso se hace del servicio bicimad es el miércoles, con 555382, jueves y viernes. En ese orden decreciente de uso. Los días que menos uso se hace son el Domingo, con 363030, y el Sabado. Por lo tanto vemos que se dá mayor uso entre semana más que los fines de semana.
- Vemos que el día y grupo de edad que más uso hace es: Miercoles el grupo 27-40 (124213 desplazamientos), lo cual concuerda perfectamente con lo estudiado antes.


diperezc@DESKTOP-24E72D5
202003_movements.json 202006_movements.json 202009_movements.json 202012_movements.json
diperezc@DESKTOP-24E72D5:/mnt/c/Users/didip/OneDrive/Documentos/MatemáticasUCM/QUINTO/SEGUNDO CUATRI/PROGRAMACIÓN PARALELA/PARALELA/bicimad/archivos usados$ python3 bicimad2.py 202001_movements.json 202002_movements.json 202003_movements.json 202004_movements.json 202005_movements.json 202006_movements.json 202007_movements.json 202008_movements.json 202009_movements.json 202010_movements.json 202011_movements.json 202012_movements.json
22/05/22 17:37:04 WARN Utils: Your hostname, DESKTOP-24E72D5 resolves to a loopback address: 127.0.1.1; using 172.26.244.100 instead (on interface eth0)
22/05/22 17:37:04 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to another address
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
22/05/22 17:37:07 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Adding 202001_movements.json
Adding 202002_movements.json
Adding 202003_movements.json
Adding 202004_movements.json
Adding 202005_movements.json
Adding 202006_movements.json
Adding 202007_movements.json
Adding 202008_movements.json
Adding 202009_movements.json
Adding 202010_movements.json
Adding 202011_movements.json
Adding 202012_movements.json
Los Lunes también se han cogido 254301 bicicletas del que no se puede determinar el rango de edad del usuario
Los Lunes el grupo de edad 0-16 años coge un total de 4621 bicicletas
Los Lunes el grupo de edad 17-18 años coge un total de 1706 bicicletas
Los Lunes el grupo de edad 19-26 años coge un total de 17921 bicicletas
Los Lunes el grupo de edad 27-40 años coge un total de 104310 bicicletas
Los Lunes el grupo de edad 41-65 años coge un total de 92081 bicicletas
Los Lunes el grupo de edad + 66 años coge un total de 2476 bicicletas
En total han salido 477416 bicicletas los Lunes
Los Lunes, 27-40 años años es el rango de edad que coge más bicicletas
diperezc@DESKTOP-24E72D5
Los Martes también se han cogido 265143 bicicletas del que no se puede determinar el rango de edad del usuario
Los Martes el grupo de edad 0-16 años coge un total de 4528 bicicletas
Los Martes el grupo de edad 17-18 años coge un total de 1754 bicicletas
Los Martes el grupo de edad 19-26 años coge un total de 18680 bicicletas
Los Martes el grupo de edad 27-40 años coge un total de 111363 bicicletas
Los Martes el grupo de edad 41-65 años coge un total de 97569 bicicletas
Los Martes el grupo de edad + 66 años coge un total de 2588 bicicletas
En total han salido 501625 bicicletas los Martes
Los Martes, 27-40 años años es el rango de edad que coge más bicicletas
[None]
Los Miércoles también se han cogido 293054 bicicletas del que no se puede determinar el rango de edad del usuario
Los Miércoles el grupo de edad 0-16 años coge un total de 4521 bicicletas
Los Miércoles el grupo de edad 17-18 años coge un total de 1874 bicicletas
Los Miércoles el grupo de edad 19-26 años coge un total de 21048 bicicletas
Los Miércoles el grupo de edad 27-40 años coge un total de 124213 bicicletas
Los Miércoles el grupo de edad 41-65 años coge un total de 107806 bicicletas
Los Miércoles el grupo de edad + 66 años coge un total de 2866 bicicletas
En total han salido 555382 bicicletas los Miércoles
Los Miércoles, 27-40 años años es el rango de edad que coge más bicicletas
[None]
Los Jueves también se han cogido 282536 bicicletas del que no se puede determinar el rango de edad del usuario
Los Jueves el grupo de edad 0-16 años coge un total de 4422 bicicletas
Los Jueves el grupo de edad 17-18 años coge un total de 1840 bicicletas
Los Jueves el grupo de edad 19-26 años coge un total de 20461 bicicletas
Los Jueves el grupo de edad 27-40 años coge un total de 119496 bicicletas
Los Jueves el grupo de edad 41-65 años coge un total de 103321 bicicletas
Los Jueves el grupo de edad + 66 años coge un total de 2498 bicicletas
En total han salido 534574 bicicletas los Jueves
Los Jueves, 27-40 años años es el rango de edad que coge más bicicletas
[None]
Los Viernes también se han cogido 286705 bicicletas del que no se puede determinar el rango de edad del usuario
Los Viernes el grupo de edad 0-16 años coge un total de 3972 bicicletas
Los Viernes el grupo de edad 17-18 años coge un total de 2100 bicicletas
Los Viernes el grupo de edad 19-26 años coge un total de 21178 bicicletas
diperezc@DESKTOP-24E72D5
Los Jueves el grupo de edad 41-65 años coge un total de 103321 bicicletas
Los Jueves el grupo de edad + 66 años coge un total de 2498 bicicletas
En total han salido 534574 bicicletas los Jueves
Los Jueves, 27-40 años años es el rango de edad que coge más bicicletas
[None]
Los Viernes también se han cogido 286705 bicicletas del que no se puede determinar el rango de edad del usuario
Los Viernes el grupo de edad 0-16 años coge un total de 3972 bicicletas
Los Viernes el grupo de edad 17-18 años coge un total de 2100 bicicletas
Los Viernes el grupo de edad 19-26 años coge un total de 21178 bicicletas
Los Viernes el grupo de edad 27-40 años coge un total de 115553 bicicletas
Los Viernes el grupo de edad 41-65 años coge un total de 97931 bicicletas
Los Viernes el grupo de edad + 66 años coge un total de 2480 bicicletas
En total han salido 529919 bicicletas los Viernes
Los Viernes, 27-40 años años es el rango de edad que coge más bicicletas
[None]
Los Sábados también se han cogido 244358 bicicletas del que no se puede determinar el rango de edad del usuario
Los Sábados el grupo de edad 0-16 años coge un total de 3296 bicicletas
Los Sábados el grupo de edad 17-18 años coge un total de 1899 bicicletas
Los Sábados el grupo de edad 19-26 años coge un total de 17455 bicicletas
Los Sábados el grupo de edad 27-40 años coge un total de 85260 bicicletas
Los Sábados el grupo de edad 41-65 años coge un total de 63627 bicicletas
Los Sábados el grupo de edad + 66 años coge un total de 1960 bicicletas
En total han salido 417855 bicicletas los Sábados
Los Sábados, 27-40 años años es el rango de edad que coge más bicicletas
[None]
Los Domingos también se han cogido 214033 bicicletas del que no se puede determinar el rango de edad del usuario
Los Domingos el grupo de edad 0-16 años coge un total de 3195 bicicletas
Los Domingos el grupo de edad 17-18 años coge un total de 1589 bicicletas
Los Domingos el grupo de edad 19-26 años coge un total de 15024 bicicletas
Los Domingos el grupo de edad 27-40 años coge un total de 72308 bicicletas
Los Domingos el grupo de edad 41-65 años coge un total de 54928 bicicletas
Los Domingos el grupo de edad + 66 años coge un total de 1953 bicicletas
En total han salido 363030 bicicletas los Domingos
Los Domingos, 27-40 años años es el rango de edad que coge más bicicletas