PRÁCTICA BICIMAD SPARK

I. OBJETIVO:

El objetivo de nuestra práctica, consiste en filtrar el uso de las bicicletas a lo largo del 2018 y 2020, en función de los días de la semana y las diferentes franjas de edad, para ver cómo ha afectado la pandemia a su uso.

Para que nos sea más fácil interpretar los resultados, hemos implementado su representación mediante histogramas.

II. INSTRUCCIONES PARA EJECUTAR EL PROGRAMA:

Para poder ejecutar el programa, será necesario instalar la librería Pyspark. Una vez la tengamos instalada, importaremos todos aquellos paquetes que nos puedan ser útiles:

- SparkSession
- IntegerType
- pandas_udf
- PandasUDFType

Además de estos paquetes, también tendremos que importar el paquete date de la librería datetime. Para la visualización de datos empleamos la librería plotly. https://plotly.com/python/

III. DATOS UTILIZADOS:

Para realizar nuestro análisis utilizamos las bases de datos de los meses de enero a diciembre de 2018 y 2020, disponibles en EMT. Hemos cogido los siguientes ficheros:

- Para el año 2018, hemos cogido los siguientes ficheros: 201801_Usage_Bicimad.json, …, 201812_Usage_Bicimad.json
- Para el año 2020 hemos cogido los siguientes ficheros: 202001_ movements.json, …, 202012_movements.json

Todos ellos, los hemos encontrado en la siguiente página web:

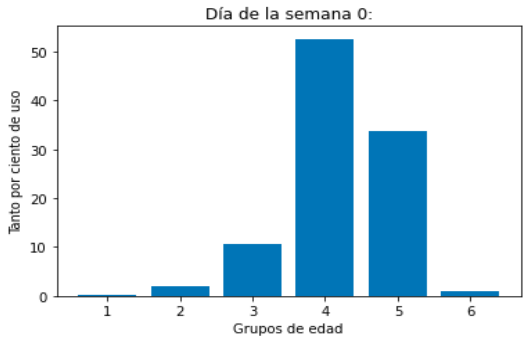
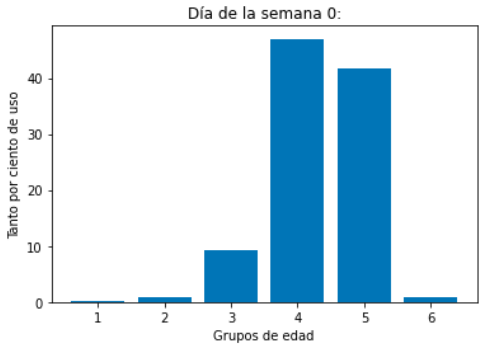
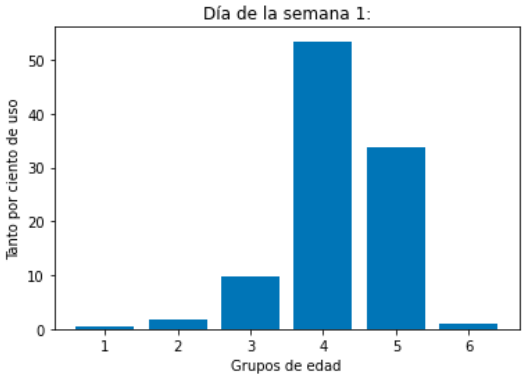
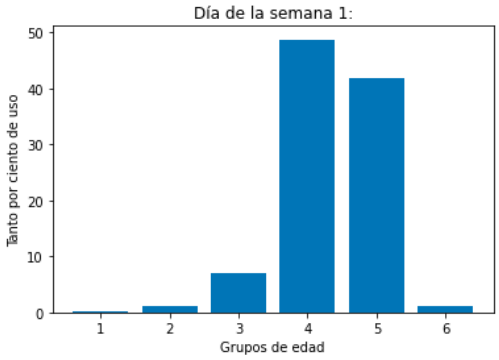
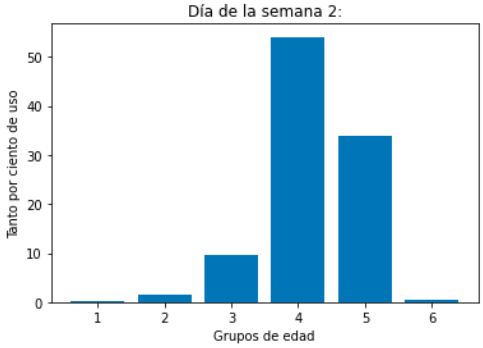
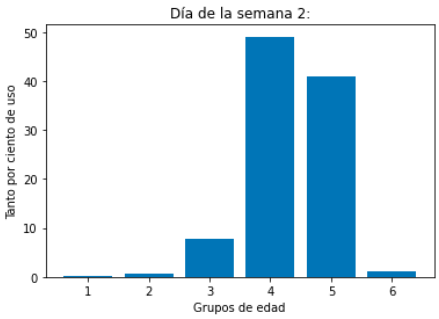
- https://opendata.emtmadrid.es/Datos-estaticos/Datos-generales-(1)

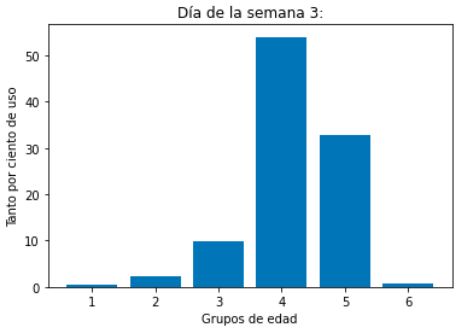
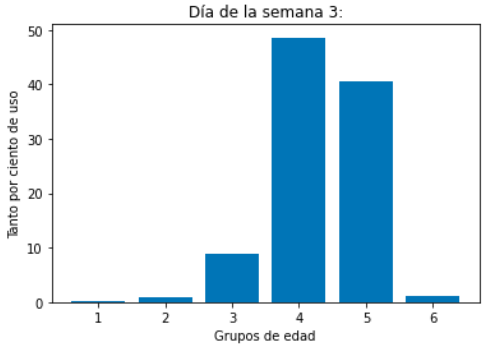
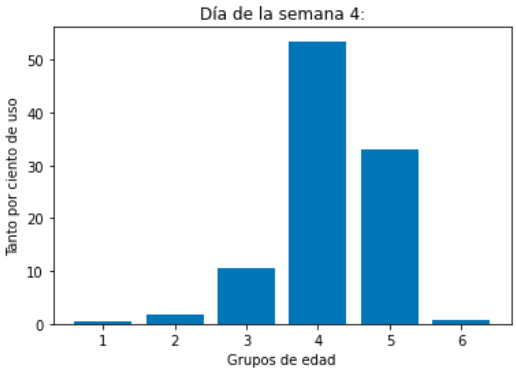
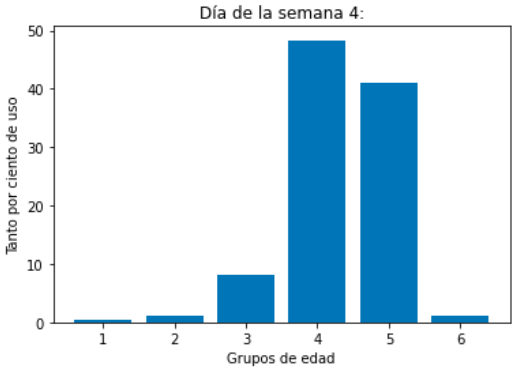
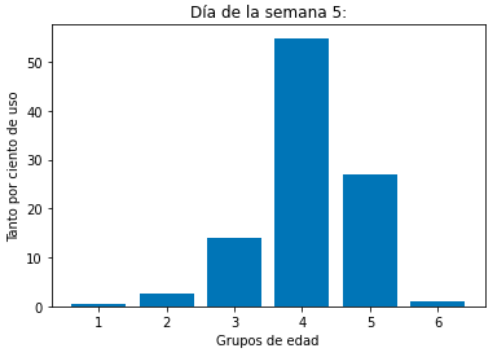
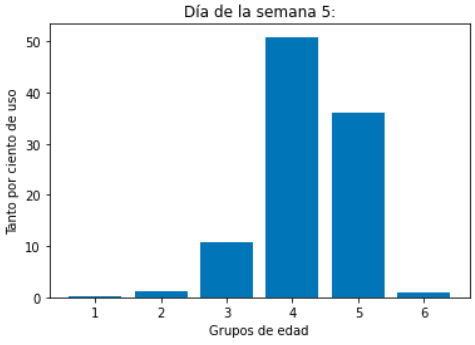
IV. PROCEDIMIENTO:

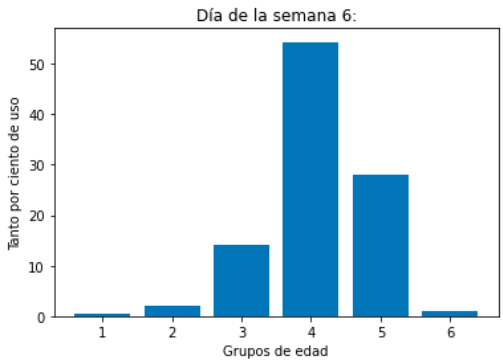
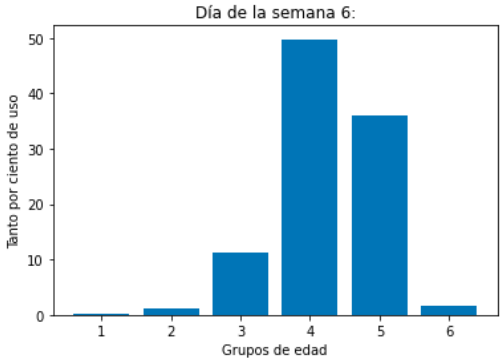
1) Seleccionamos de la base de datos de BiciMAD todos los datos de uso de los años 2018 y 2020.
2) Al tratarse de una base de datos tan extensa, hemos decidido subir los archivos al cluster para agilizar el tratamiento de estos (con el comando scp fiechero usuario@piclusterXX.mat.ucm.es:dest). A continuación, copiamos los ficheros en el sistema de ficheros hdfs (con el comando hdfs dfs –put nombredelfichero.txt nuevonombredelfichero.txt) y, tras aplicar la función take_sample los llevamos de nuevo al sistema de ficheros. Dentro del cluster ejecutamos la función take_sample con la muestra, ejecutamos la función get_sample y creamos el fichero tot_Data.json. Por último, lo descargaremos en nuestro ordenador con el nombre data10000.json
3) A continuación, vamos a trabajar con una muestra aleatoria de 100000 datos.
4) Filtramos y eliminamos todos aquellos datos que no sean de utilidad. Tras esto, nos quedamos con el 49.872% de los datos de esa muestra y creamos una nueva base de datos (df1) añadiendo a los datos anteriores una nueva columna que nos indique el día de la semana (siendo 0 el lunes y 6 el domingo).
5) Repetimos el proceso anterior, pero esta vez la nueva base de datos (df2) tendrá una nueva columna indicando el año.
6) Creamos 2 bases de datos diferenciadas, una para el 2018 (df2018) y otra para 2020 (df2020).
7) Creamos una tabla por año y franja de edad, y lo almacenamos todo en una lista (lista_tablas)
8) A continuación, pasaremos cada elemento de lista_tablas a pandas (lista_pandas), para posteriormente, pasar transformar esta lista_pandas en un diccionario. Gracias a esto, tendremos los datos en una lista de diccionarios, donde cada diccionario se corresponde con una franja de edad.
9) A partir de esta lista de diccionarios, calculamos el porcentaje de uso (en función de la franja de edad y día de la semana. Con toda esta información, construiremos los gráficos de barras, uno por cada día de la semana (franja de edad – porcentaje de uso).
10) Juntamos los datos del lunes al jueves (“labs”) y de viernes a domingo (“fds”).
11) Generamos nuevos gráficos de barras, pero esta vez cada uno de ellos corresponderá a un grupo de edad concreto, y dentro de él relacionaremos el tipo de día, “laborable” (“labs”, de lunes a jueves) o fin de semana (“fds”, de viernes a domingo) con el porcentaje de uso.

V. RESULTADOS Y CONCLUSIONES:

A continuación, vamos a comparar las siguientes tablas, en la cuales relacionamos el día de la semana con el porcentaje de uso en función de la franja de edad:

Lunes 2018	Lunes 2020
	
Martes 2018	Martes 2020
	
Miércoles 2018	Miércoles 2020
	

Jueves 2018	Jueves 2020
	
Viernes 2018	Viernes 2020
	
Sábado 2018	Sábado 2020
	

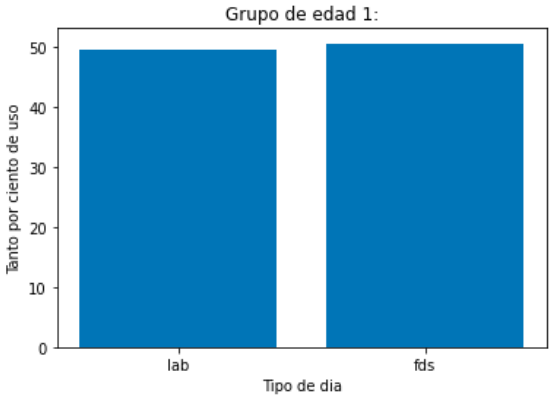
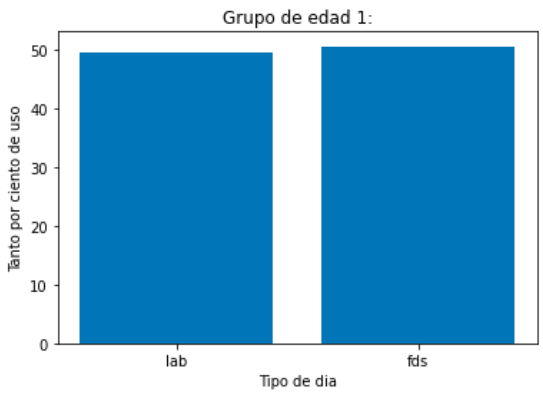
Domingo 2018	Domingo 2020
	

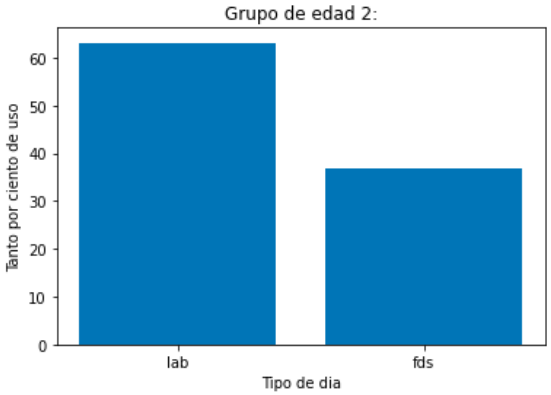
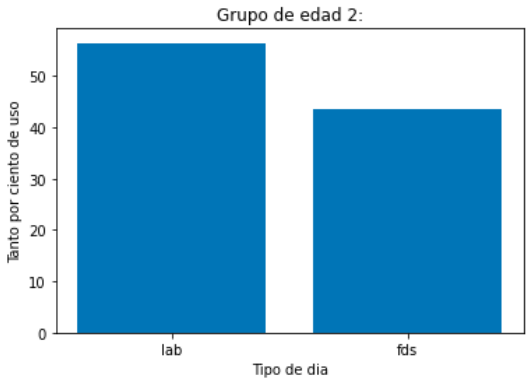
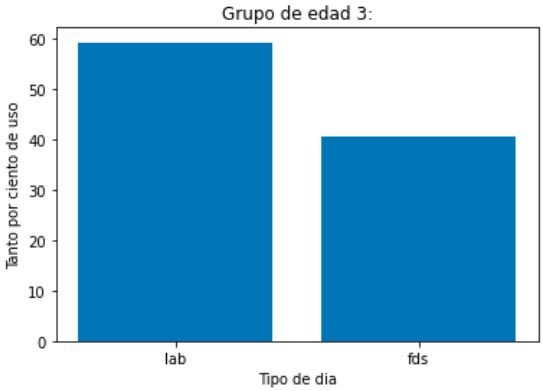
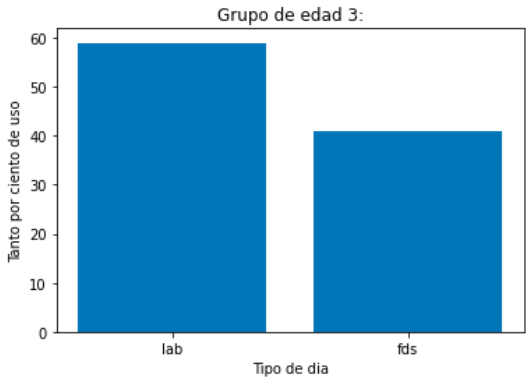
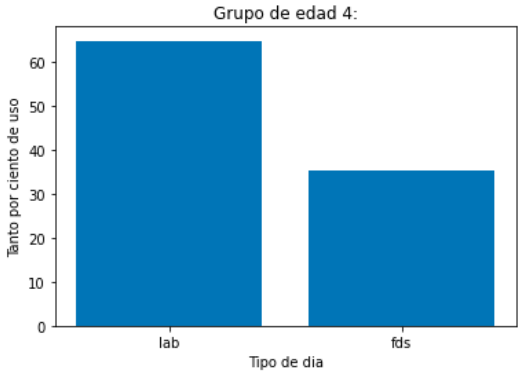
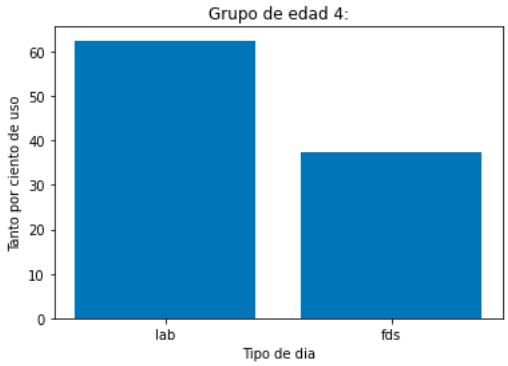
Viendo estos gráficos, podemos decir que, tanto en el 2018 como en el 2020, los grupos de edad que más utilizan las bicicletas son (de mayor a menor porcentaje): la 4 (personas entre 27 y 40 años), la 5 (de 41 a 65) y la 3 (de 19 a 26 años).

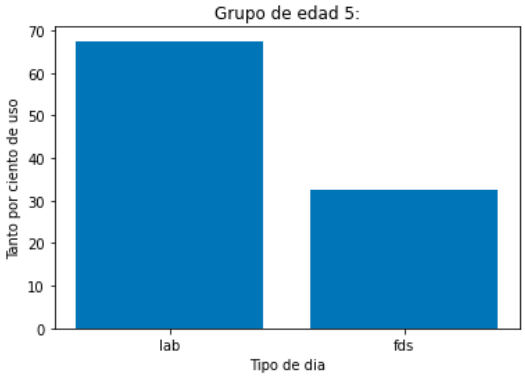
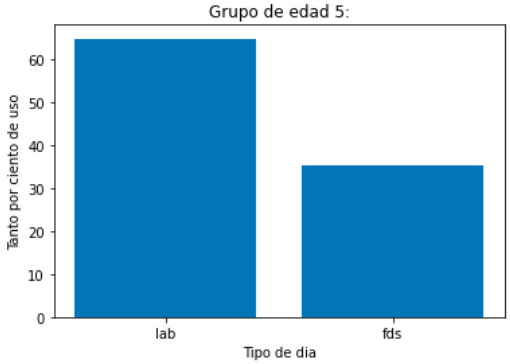
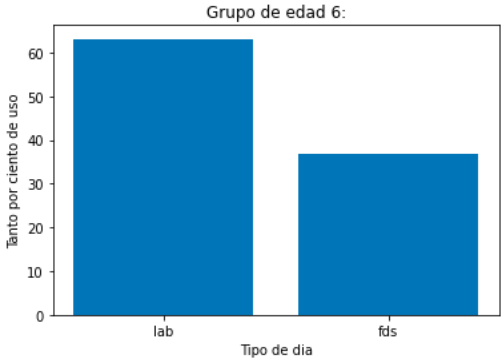
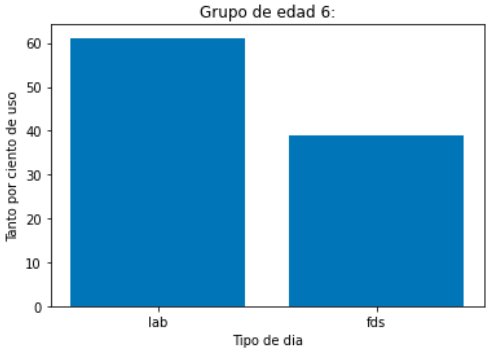
Si ahora, echamos un vistazo a estas columnas en los años proporcionados, podemos observar que, tras la pandemia, ya en el 2020, se produjo un aumento de casi el 10% de usuarios del grupo de edad 5, mientras el uso de las bicicletas en personas del grupo 4, cayó casi un 5%.

Por otro lado, podemos observar que, en los grupos de edad restantes, el porcentaje de usuarios tanto antes (en 2018), como después de la pandemia (en 2020) el porcentaje de usuarios apenas ha sufrido cambios.

Ahora, mostraremos los gráficos de cada grupo de edad que relacionan el tipo de día (laborable o fin de semana) con el porcentaje de uso.

Grupo de edad 1 en 2018	Grupo de edad 1 en 2020
	

Grupo de edad 2 en 2018	Grupo de edad 2 en 2020
	
Grupo de edad 3 en 2018	Grupo de edad 3 en 2020
	
Grupo de edad 4 en 2018	Grupo de edad 4 en 2020
	

Grupo de edad 5 en 2018	Grupo de edad 5 en 2020
	
Grupo de edad 6 en 2018	Grupo de edad 6 en 2020
	

En este caso, globalmente, podemos decir que, el mayor porcentaje de uso en todos los grupos de edades tanto antes como después de la pandemia se da en el grupo de días laborables (lab). Sí que es cierto que al tratarse de un grupo de 4 días frente al grupo fin de semana (fds) que son 3, es normal que este porcentaje sea mayor.

Si ahora comparamos la situación pre y post pandemia podemos apreciar ligeros cambios, muy similares. Vemos, que el porcentaje de uso de las bicicletas en los fines de semana a lo largo del 2020 aumentó ligeramente frente a los porcentajes iniciales del 2018.

Estas variaciones, podrían deberse (entre otras cosas) a la disminución del tráfico en Madrid. Al haber menos coches en las calles, la bicicleta ha podido ser considerada como un medio de transporte más seguro que en el 2018 en el que la circulación en Madrid era mayor (no había restricciones de movilidad) y por ello más gente se ha desplazado en bicicleta con este servicio.