BICIMAD ANTES Y DESPUÉS DE LA COVID-19

Práctica Obligatoria Spark:
BICIMAD

GRUPO 15

Introducción

La COVID-19 afectó a muchos sectores, principalmente el turismo, la hostelería y el comercio entre otros. Tras el confinamiento, comprendido entre el 15 de marzo del 2020 y el 21 de junio del 2020, la gente ansiaba moverse, pasear, hacer ejercicio… todo ello con el máximo cuidado y mucha responsabilidad.

En esta práctica estudiaremos los datos de BICIMAD antes, durante y después de dicho confinamiento dado que al tratarse de un servicio de alquiler de bicicletas creímos interesante ver cuánta gente recurrió a él como una forma de hacer ejercicio al aire libre o de transportarse de manera segura dentro de la ciudad sin necesidad de tomar un autobús o meterse en el metro. Es importante mencionar que este estudio está realizado con los datos de la cuidad de Madrid.

Es interesante recordar como fue ese “fin” del confinamiento, dado que hubo diferentes fases de desescalada hasta el 21 de junio que comenzó la nueva normalidad:

- Fase 0: Comienza el 4 de mayo con los paseos en horas determinadas y apertura de algunos locales.
- Fase 1: Comienza el 11 de mayo, se empiezan a abrir pequeños comercios, gimnasios o terrazas de los locales con un 30% de aforo entre otras medidas.
- Fase 2: Comienza el 25 de mayo y ya es posible consumir en el interior de los locales, empiezan las bodas, los espectáculos culturales y se abre el interior de los gimnasios.
- Fase 3: La denominada “fase abierta” que comienza el 8 de junio cuando se flexibilizan
  las movilidades en provincias como medida más destacable.

Metodología y material

Para analizar bien los resultados hemos necesitado una gran cantidad de datos. Los hemos tomado de la web *ww.datos.madrid.es* , de donde hemos seleccionado solo aquellos ficheros que nos interesaban. Dichos ficheros se encuentran en formato JSON. Creímos que era necesario tomar los datos de un año entero: desde agosto del 2019 hasta julio del 2020. Con esto tenemos suficiente información para estudiar los resultados de manera adecuada y completa.

Dentro de dichos ficheros aparece mucha información, a nosotros en particular nos interesa lo siguiente:

- Porcentaje de usuarios de cada tipo por mes:
  - Tipo 1: Usuario anual
  - Tipo 2: Usuario ocasional
  - Tipo 3: Trabajador de empresa
- Tiempo medio de uso en segundos por día de cada mes.
- Número de usos por mes.

Para estudiar los datos comenzamos creando una sesión de Spark con SparkContext() y cargamos los ficheros JSON de los distintos meses. A continuación juntamos todos los ficheros y nos quedamos con una muestra aleatoria de un 1% de los datos mediante la función `.filter(lambda x: random.randint(0, 100) < 1)`. Por último, creamos nuestro RDD con las columnas
'unplug_hourTime','user_type' y *'travel_time´*.

Para el estudio del número de usos por mes la función .countByKey() es suficiente.

Estudiamos el porcentaje de usuarios de cada tipo en cada uno de los meses agrupando los registros por la clave según el tipo de usuario mediante la función `.reduceByKey(lambda x, y: [x[i]+y[i] for i in range(3)])` y dividiendo estos por el número total de registros del mes estudiado. Usamos la función `.map(lambda x: (x[0], [x[1][i]/cuenta_usos[x[0]] for i in range(3)])` para aplicar lo anterior a cada uno de los registros de nuestro RDD.

Para estudiar el tiempo medio de uso de las bicicletas en cada mes aplicamos un procedimiento similar al anterior,
mediante `.reduceByKey(lambda x, y: x+y)` sumaremos todos los tiempos de uso correspondientes al campo *'travel_time'* y estos los dividiremos entre el número de viajes realizados en el mes en cuestión para obtener el tiempo medio de viaje para cada mes usando .map(lambda x: (x[0], x[1]/cuenta_usos[x[0]])). Tras realizar el estudio del tiempo medio observamos que existían registros con un *'travel_time´* mucho más alto de lo normal por lo que mediante la función `.filter(lambda x: x[1] < 10000)` eliminamos aquellos registros con un tiempo de uso mayor a 10000, obteniendo de esta forma el tiempo medio de uso para cada mes de forma más exacta tras la eliminación de estos *outliers.*

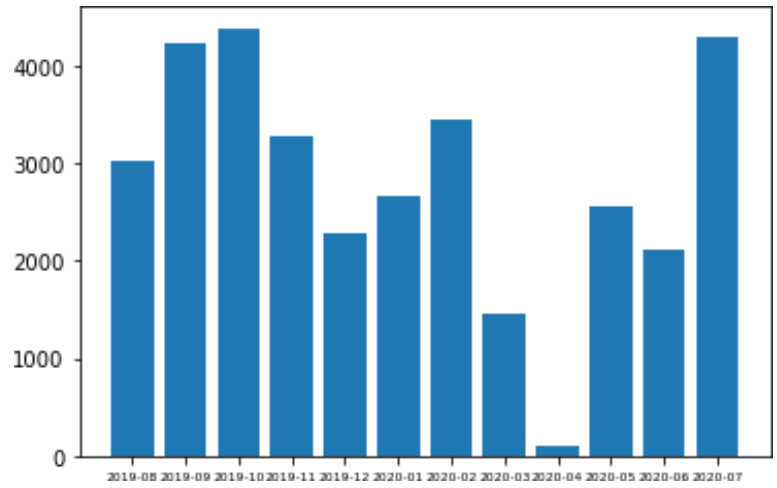
Por último, utilizamos el paquete Matplotlib para crear los histogramas que estudiaremos en las páginas siguientes.

Además incluimos una versión que trabaja con todos los datos pero no muestra los histogramas usando spark.sql. Cabe destacar la gran velocidad de esta frente a SparkContext.

Resultados

Vamos a ir analizando mediante los histogramas generados los datos de los meses mencionados.

Número de usos de BICIMAD por mes:



Agosto 2019:
3017

Septiembre 2019:
4225

Octubre 2019:
4386

Noviembre 2019:
3277

Diciembre 2019:
2285

Enero 2020: 2670

Febrero 2020: 3448

Marzo 2020: 1456

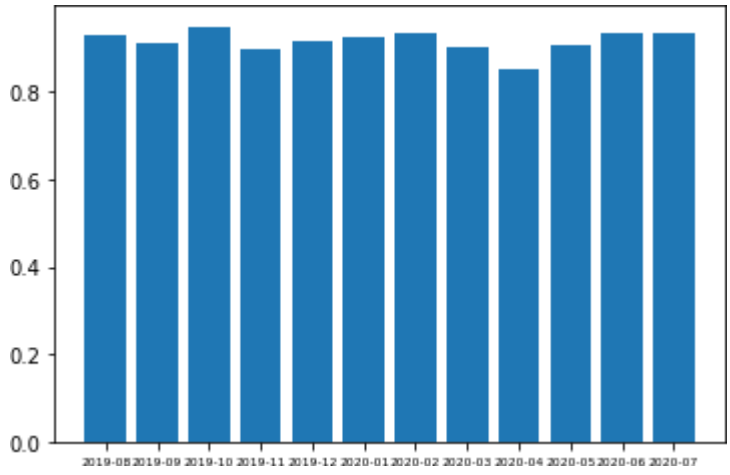
Abril 2020: 103

Mayo 2020: 2567

Junio 2020: 2111

Julio 2020: 4300

Porcentaje de usuarios abonados del tipo 1 (anuales) por mes:



Agosto 2019:
0.9284057010275107

Septiembre 2019:
0.9128994082840237

Octubre 2019:
0.9498404012767898

Noviembre 2019:
0.9002136100091547

Diciembre 2019:
0.9177242888402626

Enero 2020:
0.9262172284644195

Febrero 2020: 0.935614849187935

Marzo 2020: 0.9038461538461539

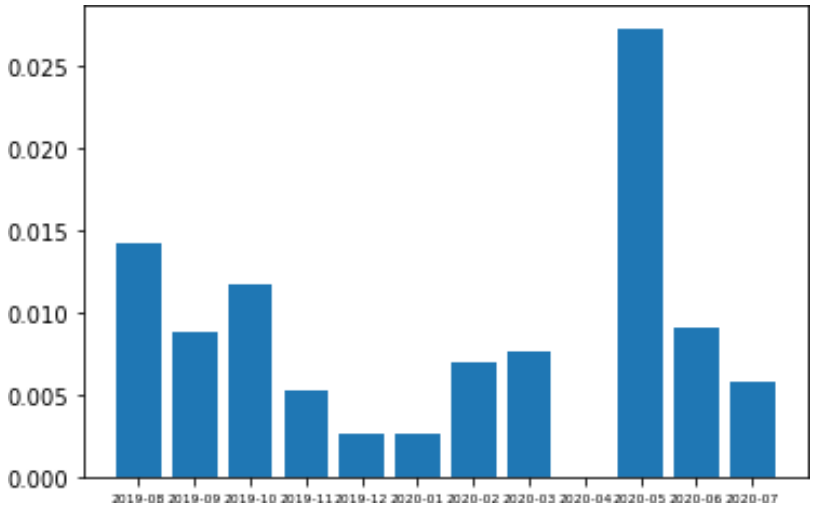
Abril 2020: 0.8543689320388349

Mayo 2020: 0.9076743280093494

Junio 2020: 0.9369966840360019

Julio 2020: 0.9327906976744186

Porcentaje de usuarios abonados del tipo 2 (ocasionales) por mes:



Agosto 2019:
0.014252568776930727

Septiembre 2019:
0.008757396449704143

Octubre 2019:
0.011627906976744186

Noviembre 2019:
0.0051876716509002135

Diciembre 2019:
0.0026258205689279

Enero 2020:
0.0026217228464419477

Febrero 2020:
0.0069605568445475635

Marzo 2020:
0.007554945054945055

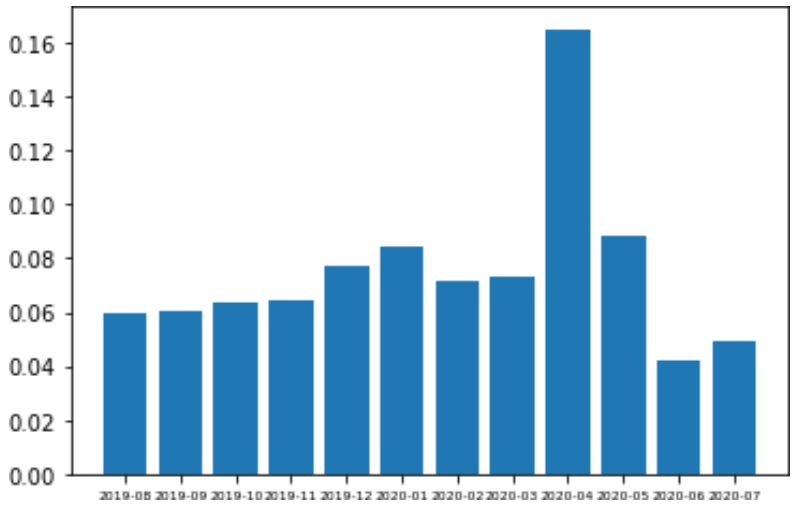
Abril 2020: 0.0

Mayo 2020:
0.027269185820023373

Junio 2020: 0.009000473709142587

Julio 2020: 0.005813953488372093

Porcentaje de usuarios abonados del tipo 3 (trabajadores) por mes:



Agosto 2019:
0.05966191581040769

Septiembre 2019:
0.06059171597633136

Octubre 2019:
0.06338349293205654

Noviembre 2019:
0.06438815990234971

Diciembre 2019:
0.07702407002188184

Enero 2020:
0.08464419475655431

Febrero 2020:
0.07192575406032482

Marzo 2020: 0.07348901098901099

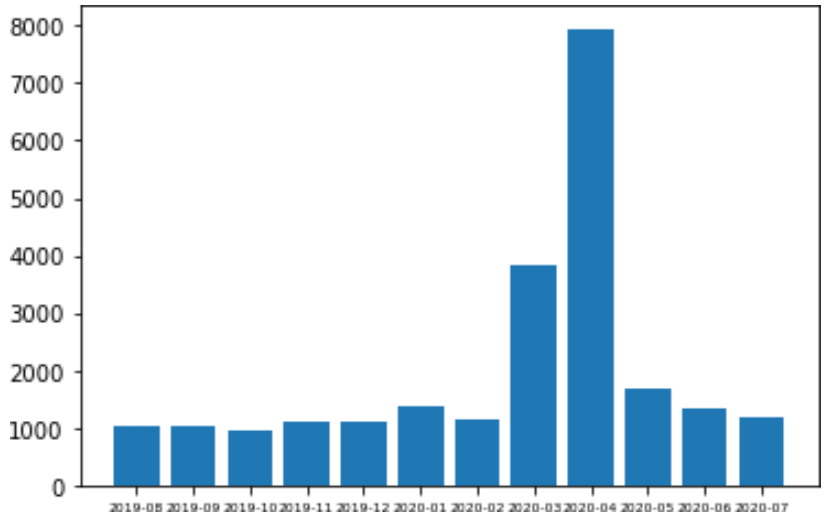
Abril 2020: 0.1650485436893204

Mayo 2020: 0.0884300740163615

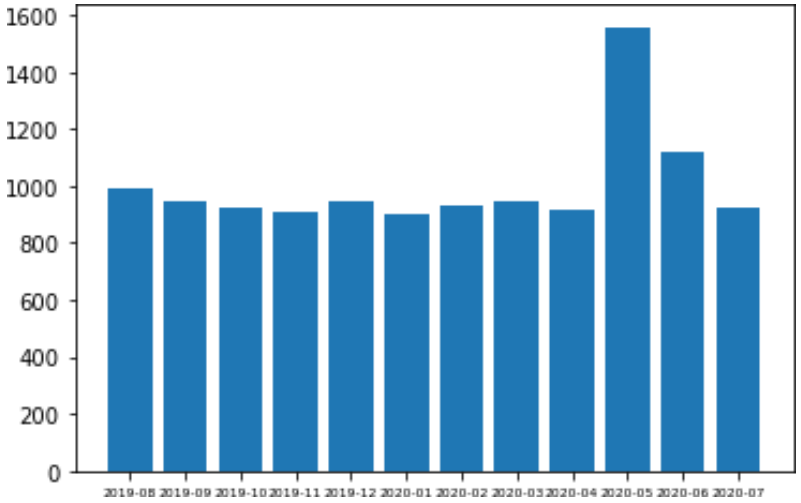
Junio 2020: 0.04168640454760777

Julio 2020: 0.049534883720930234

Por último, vamos a ver los **resultados del tiempo medio de uso en segundos por día de cada mes.** Sin embargo, es necesario explicar porqué vamos a obtener dos tablas. Nos dimos cuenta de que la tabla de tiempo medio en segundos por día y mes tenía un pico muy alto en el mes de mayo. Esto no tenía mucho sentido dado que precisamente en mayo fue cuando fuimos confinados.



Por ello investigamos sobre los datos de mayo del 2019 para ver que había podido suceder. Al ver que se trataba solamente de algunos usuarios cuyo numero de segundos por un solo uso era extremadamente alto concluimos que se trataba de aquellos usuarios que habían alquilado una bicicleta y nunca pudieron devolverla a una estación por culpa del confinamiento. Como
esto “contaminaba” nuestro estudio, hicimos un nuevo histograma quitando esos outliers y
esta vez si que obtuvimos unos resultados que tenían mucho más sentido:



Conclusiones

Vamos a ir analizando cada histograma individualmente y dentro de cada histograma los datos más destacables.

Número de usos de BICIMAD por mes: Como hemos podido observar en el histograma, antes de marzo, que fue cuando comenzó el confinamiento, las cifras de número de usos por día y mes rondaban entre los 2200 y los 4300. Sin embargo, al llegar marzo las cifras caen a 1400 aproximadamente. Intuimos que este dato se refiere a los primeros 14 días de marzo, cuando aún no estábamos confinados. Después, en abril la cifra baja muy significativamente a 103.
Esto tiene sentido, dado que durante el mes entero de abril no se pudo salir a la calle. ¿Por qué entonces hay 103 usos? Se puede deber a varios factores, puede que alguien se saltase la cuarentena, que no pudiese devolver la bicicleta a una estación hasta que nos desconfinaron o lo que es más probable, que esos 103 usos fuesen realizados por usuarios del tipo 3, la gente que si que podía salir de casa y desplazarse porque debían ir a trabajar.

Lo último que nos queda por mencionar de este histograma es que una vez desconfinados y durante las fases de desescalada, las cifras de los usos volvían a la normalidad… como nosotros.

Porcentaje de usuarios abonados de cada tipo por mes: Empezamos por analizar los resultados de los usuarios anuales dado que apenas hay cambios. Su porcentaje sobre el total de los tres tipos de usuarios se mantiene prácticamente igual rondando las cifras 90% - 95%. Sin embargo, nos damos cuenta de que en abril ese porcentaje es menor, un 85% más o menos, por tanto, si este porcentaje baja, significa que algún otro aumenta. Dado que el porcentaje de usuarios ocasionales en abril es 0%, nos damos cuenta de que el porcentaje de usuarios anuales baja porque el de trabajadores aumenta. Recordemos de nuevo que solo podías salir de casa por trabajo, emergencia médica o para ir a un supermercado básicamente.

En cuanto a los usuarios ocasionales, como ya hemos dicho, en abril su porcentaje era de un 0% (no podías bajar a la calle y coger una bici solo para pasear de forma ocasional, pero aquellos que tenían abono anual lo mantuvieron en su mayoría, por eso su porcentaje no varía apenas). Además, en mayo su porcentaje se dispara, creemos que, debido a la desescalada, los paseos por franjas horarias, etc.

Tiempo medio de uso en segundos por día de cada mes: En este apartado nos fijamos en ambos histogramas. En primer lugar, el histograma con outliers. En este observamos que el hecho más destacable es que en mayo y abril su crecimiento es extremadamente alto (de 3800-8000 segundos frente a un uso normal de unos 1000 segundos). Nuestra conclusión al ver que se trata de unos pocos usuarios que aumentan la media es que dichos usuarios no pudieron devolver la bicicleta a una estación BICIMAD. Sin embargo, al observar el histograma sin outliers todo cobra más sentido. El tiempo medio por día y mes es muy similar en casi todos los meses a excepción de mayo, que fue cuando nos desconfinamos. En mayo la cifra es mayor, por lo que intuimos que la gente se volcó en hacer deporte al aire libre o desplazarse con bicicletas para disfrutar plenamente del desconfinamiento, en lugar de usar transporte público o pasear. *Querían hacer algo diferente y fuera de peligro por contagio.*