PRÁCTICA FINAL: Análisis del tráfico de Madrid durante la pandemia del COVID

CARLOS VEGA GONZÁLEZ

Introducción

En este proyecto se requiere los siguiente:

La pandemia de COVID-19 es una enfermedad infecciosa causada por el coronavirus SARS-CoV-2. Se descubrió por primera vez en Wuhan, China en diciembre de 2019 y se propagó rápidamente a nivel mundial. El primer caso confirmado de COVID-19 en España se informó el 31 de enero de 2020, y desde ese momento la enfermedad se extendió por todo el país, afectando a millones de personas. La pandemia tuvo un gran impacto en la vida de los ciudadanos españoles: para controlar la propagación del virus, se implementaron medidas restrictivas como un confinamiento domiciliario, el cierre de fronteras, el cierre de negocios y escuelas o restricciones a la movilidad. Esto llevó a una disminución en la actividad comercial y una reducción en el tráfico de vehículos en las carreteras. Además muchas actividades pasaron a realizarse desde los domicilios gracias al teletrabajo o las clases online, y se cancelaron viajes no esenciales. Todo esto impactó de forma significativa a los desplazamientos por carretera.

1: Atendiendo a los datos de la introducción, investigar y definir los períodos relevantes desde el comienzo de la pandemia hasta la actualidad que puedan haber afectado a la movilidad de las personas (p.e antes y después de la pandemia, confinamiento estricto, restricciones de movilidad entre provincias, etc...) y seleccionar los datos de intensidad de tráfico que se correspondan con esos periodos que hayas definido.

2: Realizar un estudio usando SPARK, de cómo ha afectado la pandemia del covid-19 al tráfico en la ciudad de Madrid. Puedes complementar los datos del tráfico con otros datasets que consideres relevantes para enriquecer el estudio.

1. Análisis previo

Primero haremos un breve análisis para entender mejor los datos que tenemos. Realizando la operación .printSchema() y .show() sobre el dataset podemos observar que tenemos un total de 9 columnas:

```
root
 |-- id: integer (nullable = true)
 |-- fecha: timestamp (nullable = true)
 |-- tipo_elem: string (nullable = true)
 |-- intensidad: integer (nullable = true)
 |-- ocupacion: double (nullable = true)
 |-- carga: integer (nullable = true)
 |-- vmed: double (nullable = true)
 |-- error: string (nullable = true)
 |-- periodo_integracion: integer (nullable = true)
```

```
+----+-------------------+---------+----------+---------+-----+----+-----+------------------+
|  id|              fecha|tipo_elem|intensidad|ocupacion|carga|vmed|error|periodo_integracion|
+----+-------------------+---------+----------+---------+-----+----+-----+------------------+
|1001|2019-12-01 00:00:00|      M30|      1008|      3.0|    0|60.0|    N|                 5|
|1001|2019-12-01 00:15:00|      M30|       924|      3.0|    0|59.0|    N|                 5|
|1001|2019-12-01 00:30:00|      M30|       984|      2.0|    0|60.0|    N|                 5|
```

Nos interesa trabajar con la intensidad, ya que es el valor que mide la cantidad de tráfico que hay en un tramo concreto a una hora concreta.

Por otro lado, tenemos el tenemos el tipo_elem que nos indica la zona donde se miden los datos y tenemos M_30 y URB. Se puede utilizar para acotar más todavía nuestro rango de intensidad.

2. Estudio de la intensidad de tráfico durante el covid-19

Primero estableceré los rangos que se van a analizar y su porqué:

Pre-covid: se muestran los datos de antes del covid-19 para poder comparar y va desde 01/12/19 al 13/03/20.

Primer estado de alarma: Declaración del primer estado y del confinamiento total en España. Su duración fue del 14/03/20 al 20/06/20,

Nueva normalidad: Tras el primer estado de alarma, llegó la nueva normalidad que va desde el 21/06/20 al 20/10/20.

Segundo estado de alarma: esta declaración va desde 21/10/20 al 09/5/21.

Después del segundo estado de alarma: abarca la franja del 10/5/21 al 1/11/21.

Pre-covid

Primero cargamos los csv con los que vamos a trabajar:

```
diciembre19 = spark.read.csv('12-2019.csv', header=True, inferSchema=True, sep=";")
enero20 = spark.read.csv('01-2020.csv', header=True, inferSchema=True, sep=";")
febrero20 = spark.read.csv('02-2020.csv', header=True, inferSchema=True, sep=";")
marzo20 = spark.read.csv('03-2020.csv', header=True, inferSchema=True, sep=";")
```

Una vez tengamos los datos cargados, filtramos por la fechas que nos interesan:

```
diciembre19 = diciembre19.filter(dayofmonth(to_date(col("fecha"),"yyyy-mm-dd"))>=1)
marzo20 = marzo20.filter(dayofmonth(to_date(col("fecha"),"yyyy-mm-dd"))<=13)
```

Cuando los datos estén filtrados juntamos todos los datasets en uno y realizamos un groupBy sobre el id junto con el calculo de la media de las intensidades. Creamos la columna Fases. Y por último eliminamos la columna de id, y volvemos a realizar otro groupBy pre esta vez sobre Fases y realizamos de nuevo el cálculo de la media de las intensidades.

```
unionFase1 = diciembre19.union(enero20).union(febrero20).union(marzo20)
intFase1 = unionFase1.groupBy("id").mean("intensidad")
intFase1 = intFase1.withColumn("Fases", lit('Pre Covid-19'))
intFase1 = intFase1.withColumnRenamed("avg(intensidad)", "intensidad")
intFase1 = intFase1.drop("id").groupBy("Fases").mean("intensidad")
```

De esta manera calculamos la media de la intensidad de cada fase que he establecido. Obteniendo el siguiente resultado:

```
+-----------+-----------------+
|      Fases|  avg(intensidad)|
+-----------+-----------------+
|Pre Covid-19|419.7335557378334|
+-----------+-----------------+
```

Primer estado de alarma

Para realizar el calculo de la intensidad durante el confinamiento realizamos las mismas operaciones que se han mencionado anteriormente, cambiando simplemente las fechas y los datasets.

El rango establecido sería el siguiente:

```
marzo20 = marzo20.filter(dayofmonth(to_date(col("fecha"),"yyyy-mm-dd"))>=14)
junio20 = junio20.filter(dayofmonth(to_date(col("fecha"),"yyyy-mm-dd"))<=20)
```

Obtenemos el siguiente resultado:

```
+------------+-----------------+
|       Fases|  avg(intensidad)|
+------------+-----------------+
|Primer estado|136.3331458375607|
+------------+-----------------+
```

Nueva normalidad

El filtro es el siguiente:

```
mayo20 = mayo20.filter(dayofmonth(to_date(col("fecha"),"yyyy-mm-dd"))>=21)
octubre20 = octubre20.filter(dayofmonth(to_date(col("fecha"),"yyyy-mm-dd"))<=20)
```

Y los resultados obtenidos so los siguientes:

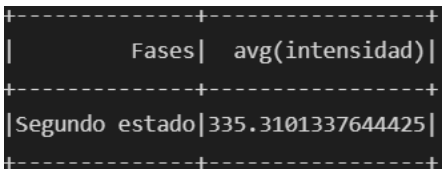
```
+--------------+-----------------+
|          Fases|  avg(intensidad)|
+--------------+-----------------+
|Nueva normalidad|288.1112356542706|
+--------------+-----------------+
```

Segundo estado de alarma

El filtro utilizado es el siguiente:

```
octubre20 = octubre20.filter(dayofmonth(to_date(col("fecha"),"yyyy-mm-dd"))>=21)
mayo21 = mayo21.filter(dayofmonth(to_date(col("fecha"),"yyyy-mm-dd"))<=9)
```

Y los resultados obtenidos:



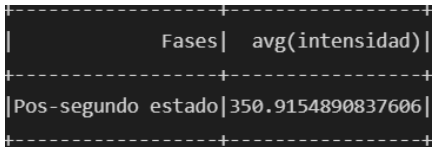
Fases	avg(intensidad)
Segundo estado	335.3101337644425

Después del segundo estado de alarma

Por último, muestro el filtro:

```
mayo21 = mayo21.filter(dayofmonth(to_date(col("fecha"),"yyyy-mm-dd"))>=9)
noviembre21 = noviembre21.filter(dayofmonth(to_date(col("fecha"),"yyyy-mm-dd"))<=1)
```

Los resultados serian los siguientes:



Fases	avg(intensidad)
Pos-segundo estado	350.9154890837606

Unión de los datos

Se muestra la unión de todos los resultados anteriores, ordenados de primera fase a la última. Se puede observar que los datos disminuyen claramente durante el confinamiento y van aumentando conforme se va dando más libertad de movilidad.

Fases	Intensidad media de tráfico
Pre Covid-19	419.733556
Primer estado	136.333146
Nueva normalidad	288.111236
Segundo estado	335.310134
Pos-segundo estado	350.915489

Como ha afectado la pandemia del covid-19 al tráfico de Madrid

Para realizar este estudio he utilizado dos dataset de datos de la comunidad, entre ellos de ruido y multas a vehículos. Se ha estudiado en las mismas

Los datos obtenidos son los siguientes:

- Ruido: Los dataset de ruido se corresponden con datos recogidos en diferentes partes de Madrid.

Fases	Contaminacion acustica media
Pre Covid-19	60.593719
Primer estado	56.954675
Nueva normalidad	58.949540
Segundo estado	59.893164
Pos-segundo estado	59.067206

- Multas: Los dataset usados son de la comunidad de Madrid. Aquí se muestra la cantidad de multas de todo tipo que se han puesto en las diferentes fases del covid-19.

Fases	Cantidad de multas
Pre Covid-19	804063
Primer estado	487713
Nueva normalidad	553510
Segundo estado	1310436
Pos-segundo estado	1093701

Podemos observar, que en ambos estudios los datos disminuyen durante el confinamiento, y a partir de ahí los datos aumentan. En el dataset de multas, podemos observar que hay un incremento bastante alto en el segundo estado y después, debido a que se utiliza la función count y los rangos de meses que utilizo son más grandes que los otros, al tener más datos hay más multas.