SISTEMAS DISTRIBUIDOS II

EVALUACIÓN DE MEDIO TERMINO

Ismael López Martínez
Ginna Monzerrat Nava Enríquez
Francisco García Ortega
José David Cobos Trujillo

INTRODUCCIÓN

- Como parte de la 1er Evaluación consistió en desarrollar un conjunto de ejercicios con datasets robustos
  - *Bases de datos de traslados en la ciudad de NY*
  - *12 archivos mensual prom. 3gb (total aprox. 30 gb)*

- RETOS
  - *Procesar el conjunto de datos para resolver los 11 ejercicios propuestos.*
  - *Usar técnicas de procesamiento que permitieran optimizar el uso de memoria, almacenamiento y procesamiento.*

RETO: 1, 2 PROCESAMIENTO

- Nuestra solución consistió:
  1. *Leer uno a uno cada archivo del origen “trip_data.7z” (14 columnas)*
  2. *Limpiar los datos*
     1. Usar sólo las columnas necesarias para resolver los problemas (10 columnas).
     2. Eliminar filas que tenían valores nulos (dropna)
     3. Eliminar filas con viajes por arriba de 7 pasajeros.
     4. Delimitar coordenadas que se encuentren en cuadrante de NY
     5. Eliminar distancias mayores a 100 Millas
  3. *Generar nuevos archivos .csv*
  4. *Comprimir cada archivo limpio en formato .gz*
  5. *Resguardarlo en el drive de Google (7.64 GB)*

EJERCICIOS: 3, 4 LIMPIEZA DE DATOS

	Name	Columns	Rows
0	trip_data_12.csv	14	13971118
1	trip_data_11.csv	14	14388451
2	trip_data_10.csv	14	15004556
3	trip_data_9.csv	14	14107693
4	trip_data_8.csv	14	12597109
5	trip_data_7.csv	14	13823840
6	trip_data_6.csv	14	14385456
7	trip_data_5.csv	14	15285049
8	trip_data_4.csv	14	15100468
9	trip_data_3.csv	14	15749228
10	trip_data_2.csv	14	13990176
11	trip_data_1.csv	14	14776615

LIMPIEZA

	Name	Columns	Rows	CantNulos	FueraRangoPasaje	FueraRangoNY	FueraMillas	RowsFinales
0	trip_data_12.csv	10	13971118	112	259	233359	0	13737388
1	trip_data_11.csv	10	14388451	754	136	276720	0	14110841
2	trip_data_10.csv	10	15004556	736	104	170721	0	14832995
3	trip_data_9.csv	10	14107693	78	94	152414	0	13955107
4	trip_data_8.csv	10	12597109	45	320	146214	771	12449759
5	trip_data_7.csv	10	13823840	698	109	206778	0	13616255
6	trip_data_6.csv	10	14385456	338	157	258610	0	14126351
7	trip_data_5.csv	10	15285049	39	86	798965	0	14485959
8	trip_data_4.csv	10	15100468	146	85	257812	0	14842425
9	trip_data_3.csv	10	15749228	293	136	282785	0	15466014
10	trip_data_2.csv	10	13990176	113	95	265047	0	13724921
11	trip_data_1.csv	10	14776615	86	83	272796	0	14503650

5. USO DE PANDAS

- trip_data_1.csv
  - *2.30 gb*
  - *32 ms prom ('trip_distance')*
    - 2.7709756
- No es viable procesar con PANDAS
  - *El uso de memoria por archivo*
    - 3 GB

Memoria antes: 5,106.0625 MB
CPU times: user 47.5 s, sys: 14 s, total: 1min 1s
Wall time: 1min 1s
Memoria después: 8,027.80078125 MB

- Memoria usada:
  - 2,921.73828125 MB (2.85 GB)
- Memoria usada por el objeto dataframe:
  - 1,578 MB (1.54 GB)

6. USO DE DASK

- Procedimiento por cada archivo (limpio).
  - *Se copia del drive de Google al espacio de Colab.*
  - *Se descomprime del formato .gz*
  - *Se carga a dataframe de dask.*
  - *Se borra el archivo .csv del espacio de almacenamiento colab.*
  - *Se realiza las operaciones requeridas.*
- Tiempo estimado
  - *25 mins (recorrer los 12 archivos)*
  - *Memoria 3 gb “elastica”*
  - *Disco 3 gb “elastico”*

6. PROMEDIOS

```
Time_Trip_Distance          0.028929
Mean_Trip_Distance          2.896072
Time_Trip_Time_In_Secs      0.026699
Mean_Trip_Time_In_Secs    820.421583
dtype: float64
```

	Name	Columns	Rows	Time_Trip_Distance	Mean_Trip_Distance	Time_Trip_Time_In_Secs	Mean_Trip_Time_In_Secs
0	trip_data_12.csv	11	13737388	0.027430	2.929154	0.025989	789.093469
1	trip_data_11.csv	11	14110841	0.028155	2.878807	0.026996	775.865342
2	trip_data_10.csv	11	14832995	0.029489	2.954097	0.027711	785.454424
3	trip_data_9.csv	11	13955107	0.027412	2.988236	0.025388	785.902557
4	trip_data_8.csv	11	12449759	0.027522	2.999994	0.023629	1541.909070
5	trip_data_7.csv	11	13616255	0.032929	2.909861	0.026090	750.594135
6	trip_data_6.csv	11	14126351	0.026962	2.945223	0.026089	782.405293
7	trip_data_5.csv	11	14485959	0.028509	2.912618	0.028068	780.132697
8	trip_data_4.csv	11	14842425	0.029111	2.867807	0.028412	747.699503
9	trip_data_3.csv	11	15466014	0.030332	2.846205	0.029428	718.598134
10	trip_data_2.csv	11	13724921	0.027053	2.745731	0.025388	702.955391
11	trip_data_1.csv	11	14503650	0.032245	2.775126	0.027193	684.448976

EJERCICIOS: 6 Y 7

Se convirtieron las columnas "pickup_datetime" y "dropoff_datetime" al formato datetime para trabajar con las diferencias de tiempo que hay entre ellas.

Se generó una columna nueva llamada "duración" con la diferencia entre "dropoff_datetime" y "pickup_datetime" implementando la instrucción ".total_seconds()" para compararla con cada uno de los renglones de la columna "trip_time_in_secs".

- Los resultados se muestran en la columna "Datos Diferentes" en la tabla siguiente

Se definió como "viaje_largo" aquellos cuya duración es mayor a 20 minutos (1200 segundos)

COMPARACIÓN Y CANTIDAD DE VIAJES LARGOS

	Archivo	Rows	DatosDiferentes	ViajesLargos
0	trip_data_12.csv	13737388	13616505	2393939
1	trip_data_11.csv	14110841	13985903	2331471
2	trip_data_10.csv	14832995	14702237	2524007
3	trip_data_9.csv	13955107	13835880	2398956
4	trip_data_8.csv	12449759	12346667	1903110
5	trip_data_7.csv	13616255	13486040	2058195
6	trip_data_6.csv	14126351	14003740	2372239
7	trip_data_5.csv	14485959	14371488	2432819
8	trip_data_4.csv	14842425	14697128	2225154
9	trip_data_3.csv	15466014	15323079	2082677
10	trip_data_2.csv	13724921	13599126	1741161
11	trip_data_1.csv	14503650	14362859	1689386

	Archivo	Taxis_Diferentes
0	trip_data_12.csv	13327
1	trip_data_11.csv	13309
2	trip_data_10.csv	13317
3	trip_data_9.csv	13325
4	trip_data_8.csv	13294
5	trip_data_7.csv	13292
6	trip_data_6.csv	13367
7	trip_data_5.csv	13315
8	trip_data_4.csv	13282
9	trip_data_3.csv	13277
10	trip_data_2.csv	13257
11	trip_data_1.csv	13272

8. De los viajes largos identificar:

- A) El número de taxis diferentes (la columna *medallion* contiene un número que identificada a cada uno de los vehículos).

8. De los viajes largo identificar:

B) ¿Qué vehículos son los que más viajes realizan en cada mes? ¿Son el mismo vehículo? **Tabla A)**

20BA941F62CC07F1FA3EF3E122B1E9B2
(Septiembre, Octubre, Noviembre)

A4FC84D2662D988828DBD26B0948A413
(Junio, Julio)

TABLA A)

	Archivo	Medallion	Número_De_Viajes
0	trip_data_12.csv	5E3D30644F5CAEA4D1C5A07982D6616E	339
1	trip_data_11.csv	20BA941F62CC07F1FA3EF3E122B1E9B2	327
2	trip_data_10.csv	20BA941F62CC07F1FA3EF3E122B1E9B2	366
3	trip_data_9.csv	20BA941F62CC07F1FA3EF3E122B1E9B2	336
4	trip_data_8.csv	5466D7146013712990 33C01FB08BB93B	288
5	trip_data_7.csv	A4FC84D2662D988828DBD26B0948A413	330
6	trip_data_6.csv	A4FC84D2662D988828DBD26B0948A413	369
7	trip_data_5.csv	20BA941F62CC07F1FA3EF3E122B1E9B2	393
8	trip_data_4.csv	19E063791B0DF5A558B8488180DDAB67	334
9	trip_data_3.csv	DACFA6EF35923081481A22BE96339B6E	308
10	trip_data_2.csv	6FE6DFF9A59C0B64BE0CA64EE2699F08	253
11	trip_data_1.csv	A532B1493C4DD88C450F6796369EAA6F	256

9. GRÁFICAR TOTAL DE PASAJEROS POR DÍA DE LA SEMANA Y POR HORA DE LA SEMANA

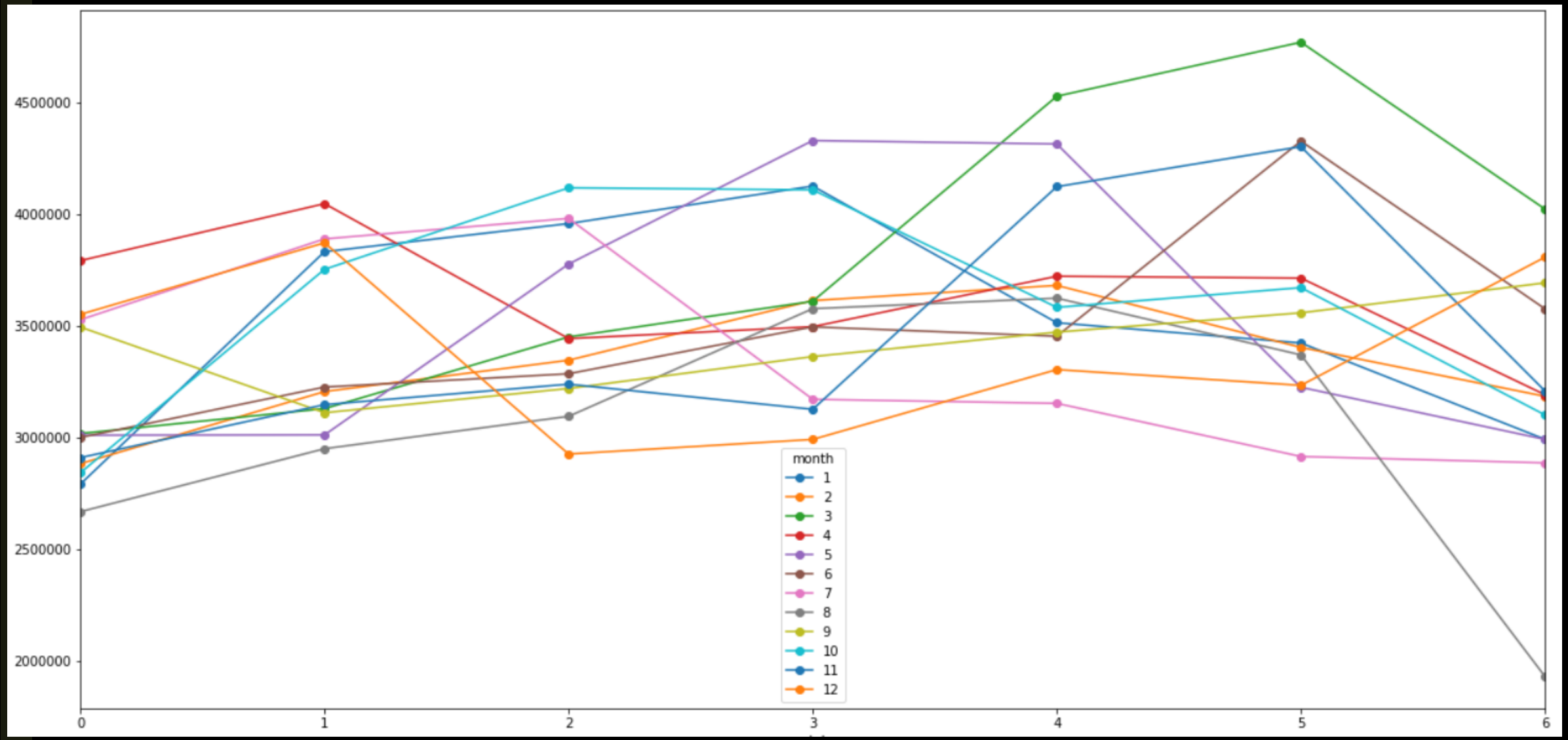
- **Conversión y agrupación de datos por día de la semana**

  Se convierten los datos de la columna "pickup_time" a un formato datetime y se agrupan por mes y por día de la semana y se suma la cantidad de pasajeros en el rango de agrupación y así presentar la gráfica del total de pasajeros por día de la semana,

- **Conversión y agrupación de datos por hora del día**

  Se convierten los datos de la columna "pickup_time" a un formato datetime y se agrupan por mes y hora para poder sumar la cantidad de pasajeros y graficarlos por el total de pasajeros por hora del día,

GRÁFICA DE NÚMERO DE PASAJEROS POR DÍA DE LA SEMANA



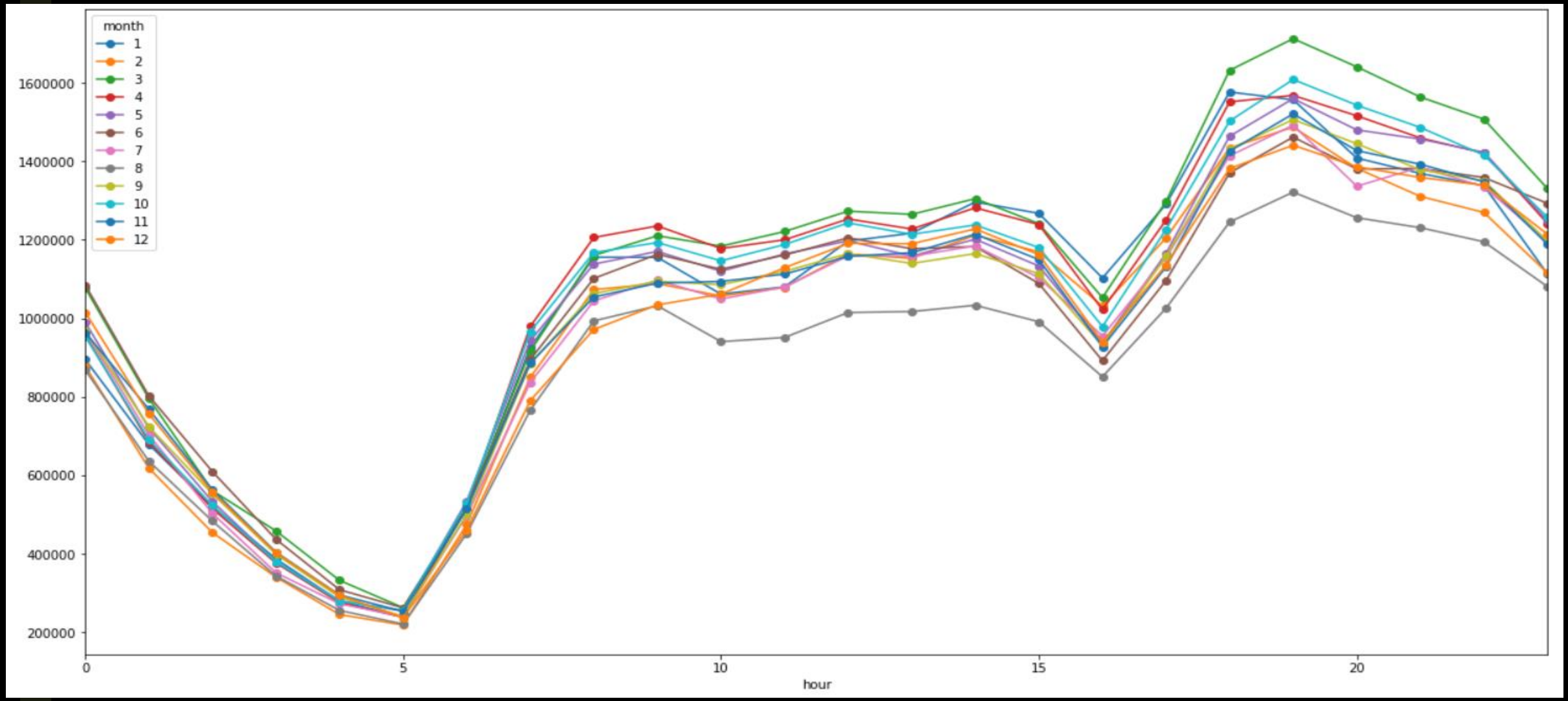
9.1 CONCLUSIONES DE GRÁFICA POR DÍA DE LA SEMANA

La cantidad de usuario por semana en cada mes no parece guardar una relación, siendo agosto el mes con menor número de pasajeros a excepción de jueves y viernes, días que registra mayor cantidad de pasajeros a comparación de diciembre y julio. Justo en esos dos días, marzo es el mes con más pasajeros.

En algunos meses como Enero o Junio, el periodo de la semana laboral (Lunes a Viernes) se mantiene estable, mientras que en fin de semana aumenta considerablemente, lo que podría parecer que en época vacacional existe menor cantidad de pasajeros.

- Los días más productivos son los viernes y sábados
- El mes de marzo es uno de los más productivos los fines de semana

GRÁFICA DE NÚMERO DE PASAJEROS POR HORA DEL DÍA



9.2 CONCLUSIONES DE GRÁFICA POR HORA DEL DÍA

A diferencia de la agrupación por días de la semana, este tiene un comportamiento muy parecido todos los meses, el mes con menor número de pasajeros es Agosto, confirmando lo que se observaba en la gráfica por días.

También podemos ver que a partir de las 19:00 horas el número de pasajeros comienza a decaer hasta las 05:00 de la mañana, dónde de nuevo comienza a crecer el número de viajes, esto puede ser derivado de los horarios de trabajo, ya que se mantiene estable en un horario de 9:00 a 18:00 horas, derivado de los horarios de salida, comienza a crecer de nuevo hasta decaer.

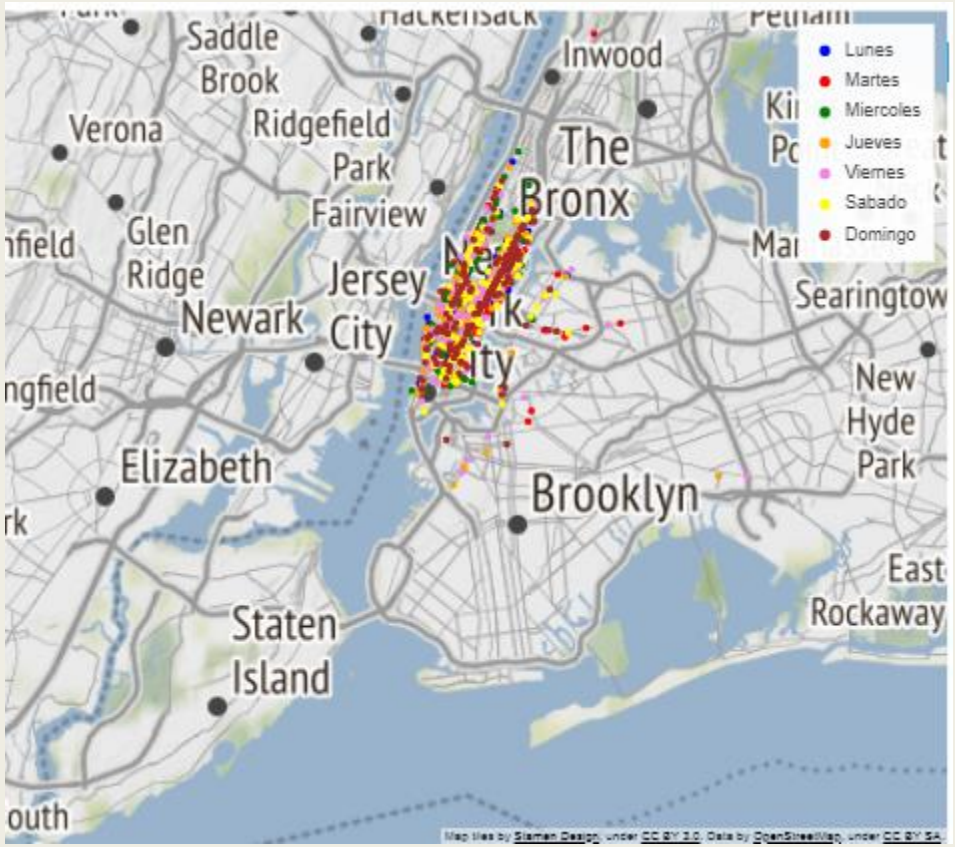
- En el rango de 00:00 a 05:00 horas los taxis bajan su cantidad de pasajeros, se empieza a recuperar inmediatamente después de pasar las 05:00.
- En el mes de Agosto, a toda hora del día hay pocos pasajeros.
- El mes de Marzo es el mes con más pasajeros.
- En el rango de las 18:00 a 23:00 horas es cuando los taxistas registran más pasajeros.
- Dato: En Estados Unidos las jornadas laborales son de las 9:00 a las 18:00 horas.

EJERCICIO 10 Y 11

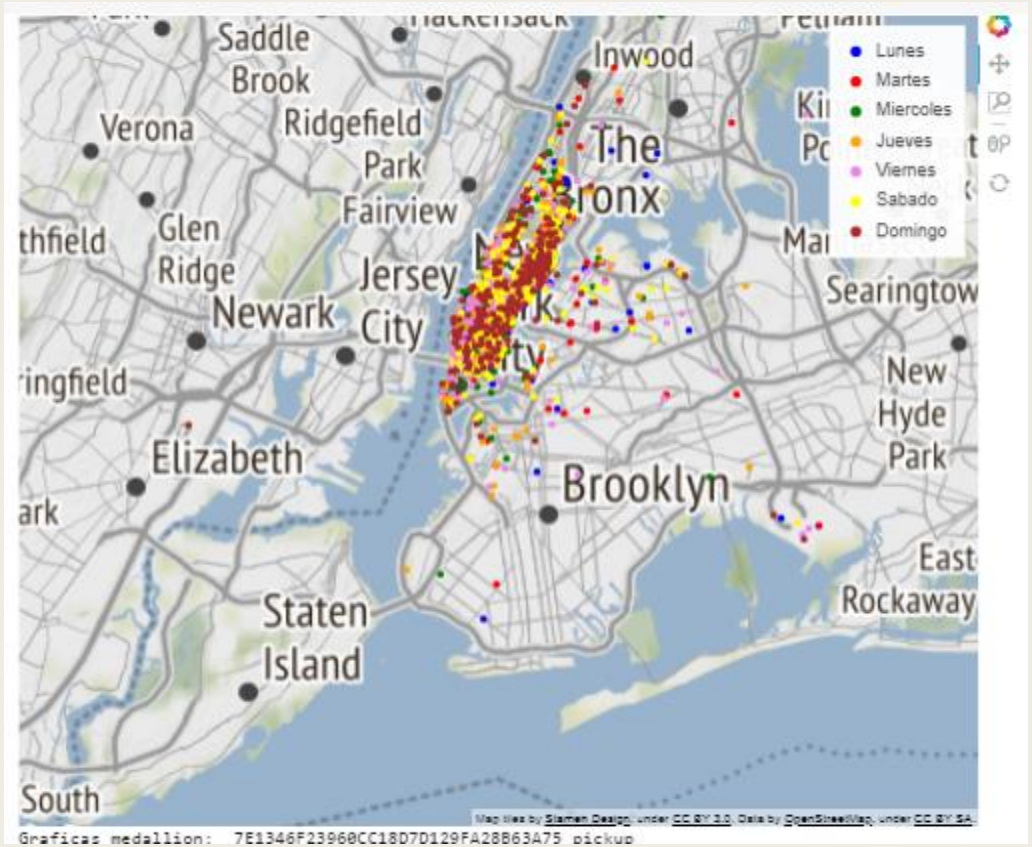
Elegir el vehículo con más viajes en cada mes y graficar en un mapa los sitios donde se suben pasajeros agrupados por día de la semana (un color distinto para cada día) hora del día (un color distinto para cada intervalo de cuatro horas,00:00 - 03:59, 04:00--07:59, 08:00-11:59, etc.)

- Se identificó cual era el taxi que tenía mayor cantidad de viajes en cada mes por medio de su medallion y se agruparon los registros relacionados con él.
- Se crearon funciones para dibujar los puntos de carga y descarga de pasajeros dentro del mapa, dentro de las funciones se convirtieron las columnas 'pickup_datetime' y 'dropoff_datetime' a datatime para poder aplicar condiciónes de selección y crear dataframe.
- Dentro de la función encargada de dibujar puntos por día, se agregó al dataframe la columna 'weekday' que contenía el día de la semana para posteriormente crear subconjuntos de dataframe correspondientes a cada uno de los días de la semana.
- Dentro de la función encargada de dibujar por rango de horas, se agregó al dataframe la columna 'hour' que contenía un strftime de hora y minuto para posteriormente crear subconjuntos de dataframe correspondientes a cada uno de los rangos de horas solicitados.
- Se implemento una función para cambiar las Coordenadas latitud y longitud de carga y descarga de pasajeros al formato x y de Web Mercator ESRI para poder dibujar los puntos en los mapas.

EJERCICIO 10 Y 11

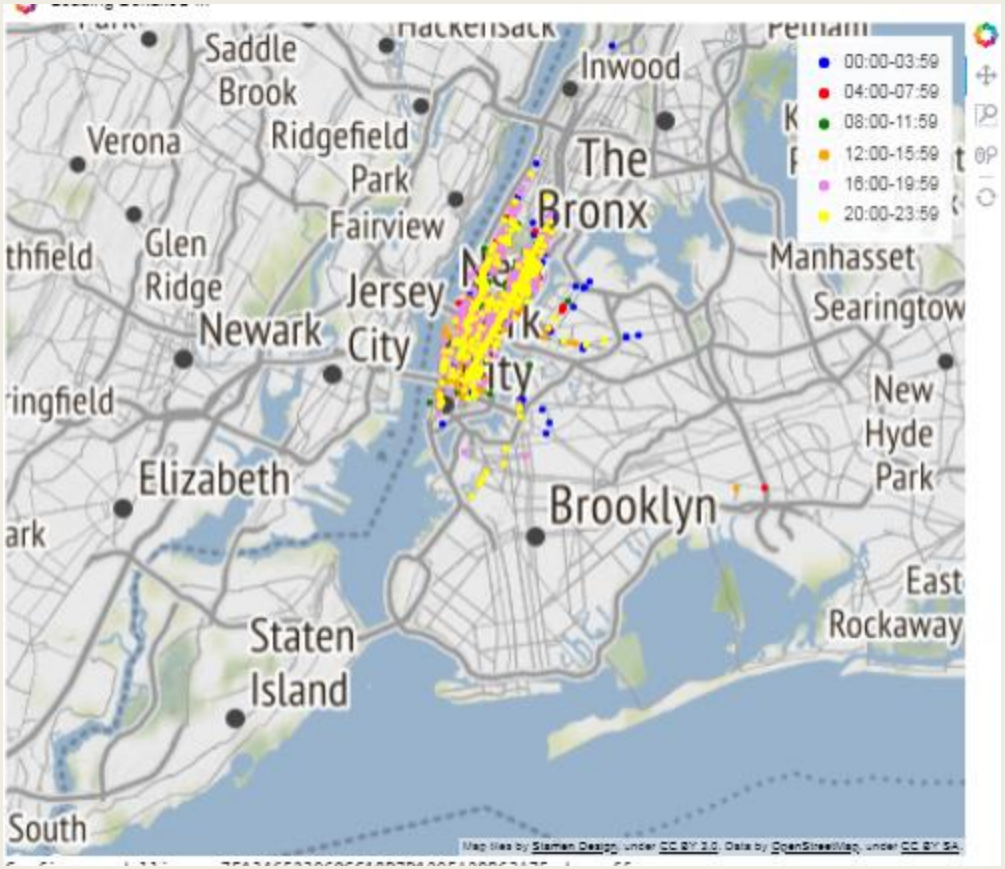


Carga de pasajeros

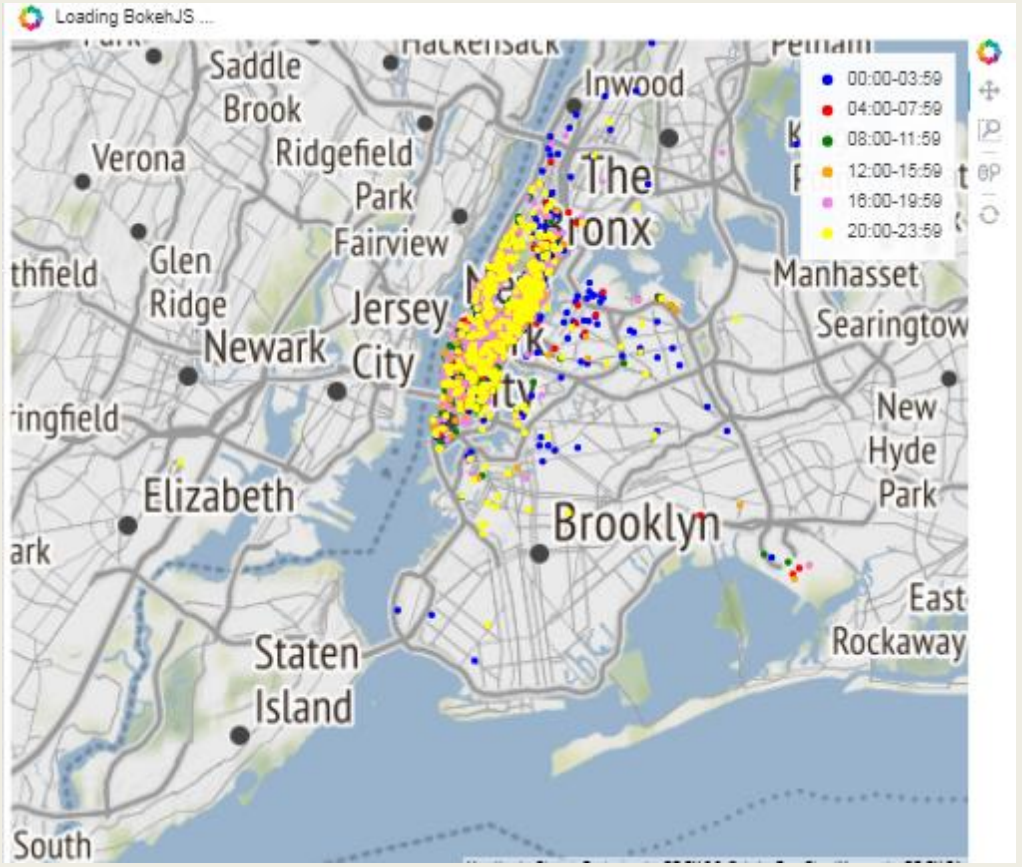


Descarga de pasajeros

EJERCICIO 10 Y 11



Carga de pasajeros



Descarga de pasajeros

EJERCICIO 10 Y 11

- **Conclusiones mapeo de puntos por mes**

Se logro observar en los mapas que la mayoría de los viajes se concentraron en la zona de MANHATTAN siendo los días domingo los de mayor cantidad de viajes en un horario de 20:00 a 23:59 y el taxi con mayor cantidad de viajes fue el correspondiente al medallion 19E063791B0DF5A558B8488180DDAB67.

EJERCICIO 10 Y 11

Medallion	Mes	Día	Hora
7E1346F23960CC18D7D129FA28B63A75	Enero	Lunes, Sábado y Domingo	00:00-03:59, 16:00-19:59, 20:00-23:59
0C9C589C0AD57ECCB633CB90A33DC37A	Febrero	Sábado y Domingo	16:00-19:59 ,20:00-23:59,
19E063791B0DF5A558B8488180DDAB67	Marzo	Sábado y Domingo	20:00-23:59
19E063791B0DF5A558B8488180DDAB67	Abril	Sábado y Domingo	00:00-03:59, 20:00-23:59
20BA941F62CC07F1FA3EF3E122B1E9B2	Mayo	Sábado y Domingo	00:00-03:59, 20:00-23:59
A4FC84D2662D988828DBD26B0948A413	Junio	Sábado y Domingo	00:00-03:59, 20:00-23:59
A4FC84D2662D988828DBD26B0948A413	Julio	Sábado y Domingo	16:00-19:59, 00:00-03:59, 20:00-23:59
5466D7146013712990 33C01FB08BB93B	Agosto	Sábado y Domingo	16:00-19:59, 00:00-03:59, 20:00-23:59
2905ABD21DF99EA5B9741EA063266632	Septiembre	Domingo	00:00-03:59, 20:00-23:59
20BA941F62CC07F1FA3EF3E122B1E9B2	Octubre	Domingo	20:00-23:59
19E063791B0DF5A558B8488180DDAB67	Noviembre	Domingo	20:00-23:59
F3E844649503D2A5A44DD729348E7336	Diciembre	Sábado y Domingo	20:00-23:59