Primera evaluación Sistemas Distribuidos y Ubicuos II

Campos Sacramento Yoselin Araceli
Flores Castelán Luis Rafael
Martínez Arteaga Alexis
Reyes Hernández Ana Margarita

ANÁLISIS DE DATOS

TAXIS DE NEW YORK

AGENDA:

1) Adquirir los datos.(trip_data.7z).
2) Descomprimir datos (trip_data.7z).
3) Identificar renglones con errores en los datos.
4) Indicar cuantos son los renglones con errores y a que archivos corresponden.
5) Crear un DataFrame usando la librería pandas.
6) Usar la librería dask.
7) Contar el número de viajes largos en cada archivo
8) De los viajes largos, identificar el número de taxis diferentes.
9) Hacer una gráfica temporal del total de pasajeros agregados por días de la semana y por horas del día.
10) Elegir el vehiculó con más viajes en cada mes y graficar en un mapa.
11) Repetir para las posiciones donde se bajan los pasajeros para el mismo vehículo.

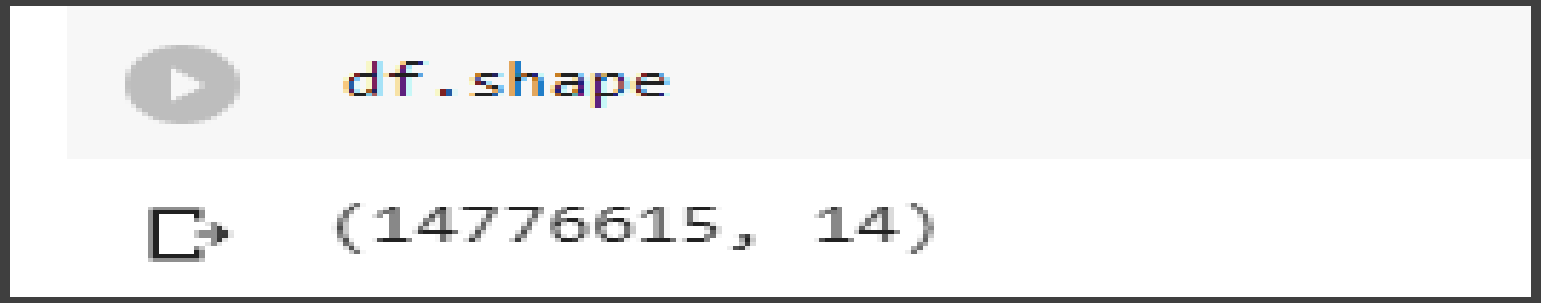
1.-ADQUIRIR LOS DATOS DE LA CARPETA(7Z X TRIP_DATA.7Z).

```
[ ]  1 !wget https://archive.org/download/nycTaxiTripData2013/trip_data.7z
```

2.-DESCOMPRIMIR ARCHIVO(7Z X TRIP_DATA.7Z).

```
1 !7z e /content/trip_data.7z -o/content/ trip_data_1.csv
```

3.- RESPONDER LAS SIGUIENTES PREGUNTAS:



3.1.- Cuantas columnas contiene cada archivo de datos descomprimido.

Respuesta de la pregunta 3.1.- R= Columnas: 14

3.2.- Cuantos renglones tiene cada archivo.

Respuesta de la pregunta 3.2.- R= Filas: 14776615

4.- IDENTIFICAR SI EXISTEN RENGLONES CON ERRORES EN LOS DATOS, POR EJEMPLO, SI HAY COLUMNAS DE MÁS (O DE MENOS), SI HAY CAMPOS VACÍOS, ETC. SI SE DETECTAN RENGLONES CON ERRORES:

4.1.- INDICAR CLARAMENTE CUANTOS SON LOS RENGLONES CON ERRORES Y A QUE ARCHIVOS CORRESPONDEN.

mi_trip_data_1.csv se limpiaron 503635 registros
mi_trip_data_2.csv se limpiaron 477415 registros
mi_trip_data_3.csv se limpiaron 529333 registros
mi_trip_data_4.csv se limpiaron 527348 registros
mi_trip_data_5.csv se limpiaron 1096395 registros
mi_trip_data_6.csv se limpiaron 489113 registros
mi_trip_data_7.csv se limpiaron 440845 registros
mi_trip_data_9.csv se limpiaron 376490 registros
mi_trip_data_10.csv se limpiaron 405720 registros
mi_trip_data_11.csv se limpiaron 495672 registros
mi_trip_data_12.csv se limpiaron 465492 registro

Respuesta de la pregunta 4.1.-

```
df.isna().sum()
```

```
medallion                0
hack_license             0
vendor_id                0
rate_code                0
store_and_fwd_flag     7326207
pickup_datetime          0
dropoff_datetime         0
passenger_count          0
trip_time_in_secs        0
trip_distance            0
pickup_longitude         0
pickup_latitude          0
dropoff_longitude       86
dropoff_latitude        86
dtype: int64
```

4.2.- CREAR UN NUEVO CONJUNTO DE DATOS CON LOS ERRORES ELIMINADOS.

Guardar en conjunto los datos nulos

```
df_nulo1 = df[df['store_and_fwd_flag'].isna()]
```

```
df_nulo2 = df[df['dropoff_longitude'].isna()]
```

```
df_nulo3 = df[df['dropoff_latitude'].isna()]
```

```
df_nulo1
```

Eliminar los datos nulos de los conjuntos

```
df = df.dropna(how='all', subset=['dropoff_longitude', 'dropoff_latitude', 'store_and_fwd_flag'])
```

5.- CREAR UN DATAFRAME USANDO LA LIBRERÍA PANDAS Y RESPONDER LO SIGUIENTE:

Función para medir la memoria utilizada para cada proceso.

```
def huella_de_memoria():
  '''Regresa la memoria usada por el proceso de python'''
  mem = psutil.Process(os.getpid()).memory_info().rss
  return (mem / 1024**2)
```

```
huella_de_inicio = huella_de_memoria()
```

Total de la memoria antes de que se ejecuten los procesos, es decir la memoria inicial.

```
huella_de_inicio
```

```
79.62890625
```

5.1.-INDICAR EL TIEMPO EN SEGUNDOS, QUE TARDA LA LIBRERÍA PANDAS EN LEER UN ARCHIVO A UN DATAFRAME.

Respuesta de la pregunta 5.1.- R=

```
t_start = time.time()
df = pd.read_csv('datos_csv/trip_data_1.csv')
t_end = time.time()
print('Tiempo que tarda pandas en leer un archivo DataFrame: {} s'.format(t_end-t_start))
```

```
Tiempo que tarda pandas en leer un archivo DataFrame: 320.99793553352356 s
```

5.2.-CUAL ES LA HUELLA DE MEMORIA DEL PROCESO USADO PARA GENERAR EL OBJETO DATAFRAME.

Respuesta de la pregunta 5.2.- R=

```
huella_de_fin = huella_de_memoria()
```

```
huella_de_fin
```

```
919.9453125
```

5.3.- CUANTA MEMORIA DEL SISTEMA SE USA PARA CREAR EL OBJETO ANTERIOR.

Respuesta de la pregunta 5.3.- R=

```
print(huella_de_fin - huella_de_inicio)
```

```
840.31640625
```

5.4.- INDICAR EL TIEMPO QUE TARDA PANDAS EN OBTENER EL PROMEDIO DE LA DISTANCIA DE VIAJE(TRIP_DISTANCE).

Respuesta de la pregunta 5.4.- R=

```
imp_diferente_cero = (df['trip_distance'] != 0.0)
```

```
tiempo_inicio = time.time()
imp_diferente_cero.mean()
tiempo_final = time.time()
print ('El tiempo que tarda pandas en obtener el promedio de la distancia de viaje (trip_distance): {} s'
       .format(tiempo_final - tiempo_inicio))
```

```
El tiempo que tarda pandas en obtener el promedio de la distancia de viaje (trip_distance): 0.9079926013946533 s
```

6.3 – INDICAR EL TIEMPO QUE TARDA DASK EN OBTENER EL PROMEDIO DE LA DISTANCIA DE VIAJE (TRIP_DISTANCIA).

Archivo	Promedio
1	0.0019996166229248047
2	0.0009975433349609375
3	0.0019948482513427734
4	0.0009973049163818360
5	0.0009973049163818360
6	0.0019948482513427734
7	0.0009982585906982422
8	0.0009903907775878906
9	0.0009944438934326172
10	0.0010056495666503906
11	0.0019965171813964844
12	0.0009965896606445312

6.4 - INDICAR EL TIEMPO QUE TARDA DASK EN OBTENER EL PROMEDIO DE LA DURACIÓN DE VIAJE (TRIP_TIME_IN_SECS).

Archivo	Promedio
1	0.0009925365447998047
2	0.0009975433349609375
3	0.0009970664978027344
4	0.0009970664978027344
5	0.0009973049163818360
6	0.0019948482513427734
7	0.0009973049163818360
8	0.0019948482513427734
9	0.0019946098327636720
10	0.0009944438934326172
11	0.0009951591491699219
12	0.0019946098327636720

Archivo	Diferencias
1	12597109
2	13823840
3	13971118
4	13990176
5	14107693
6	14388451
7	14385456
8	14776615
9	15004556
10	15100468
11	15285049
12	15749228

CON LAS COLUMNAS PICKUP_DATETIME Y DROPOFF_DATETIME COMO OBJETOS TEMPORALES, CREAR UNA NUEVA COLUMNA EN LOS DATAFRAMES QUE SE LLAME DURACIÓN Y COMPARARLA RENGLÓN A RENGLÓN CON LA COLUMNA TRIP_TIME_IN_SECS.

7.- CONTAR EL NUMERO DE VIAJES LARGOS EN CADA ARCHIVO.

Archivo	Número de viajes
1	1919857
2	2084036
3	2428514
4	1768887
5	2417638
6	2373698
7	2400513
8	1715237
9	2545155
10	2264701
11	2552684
12	2111851

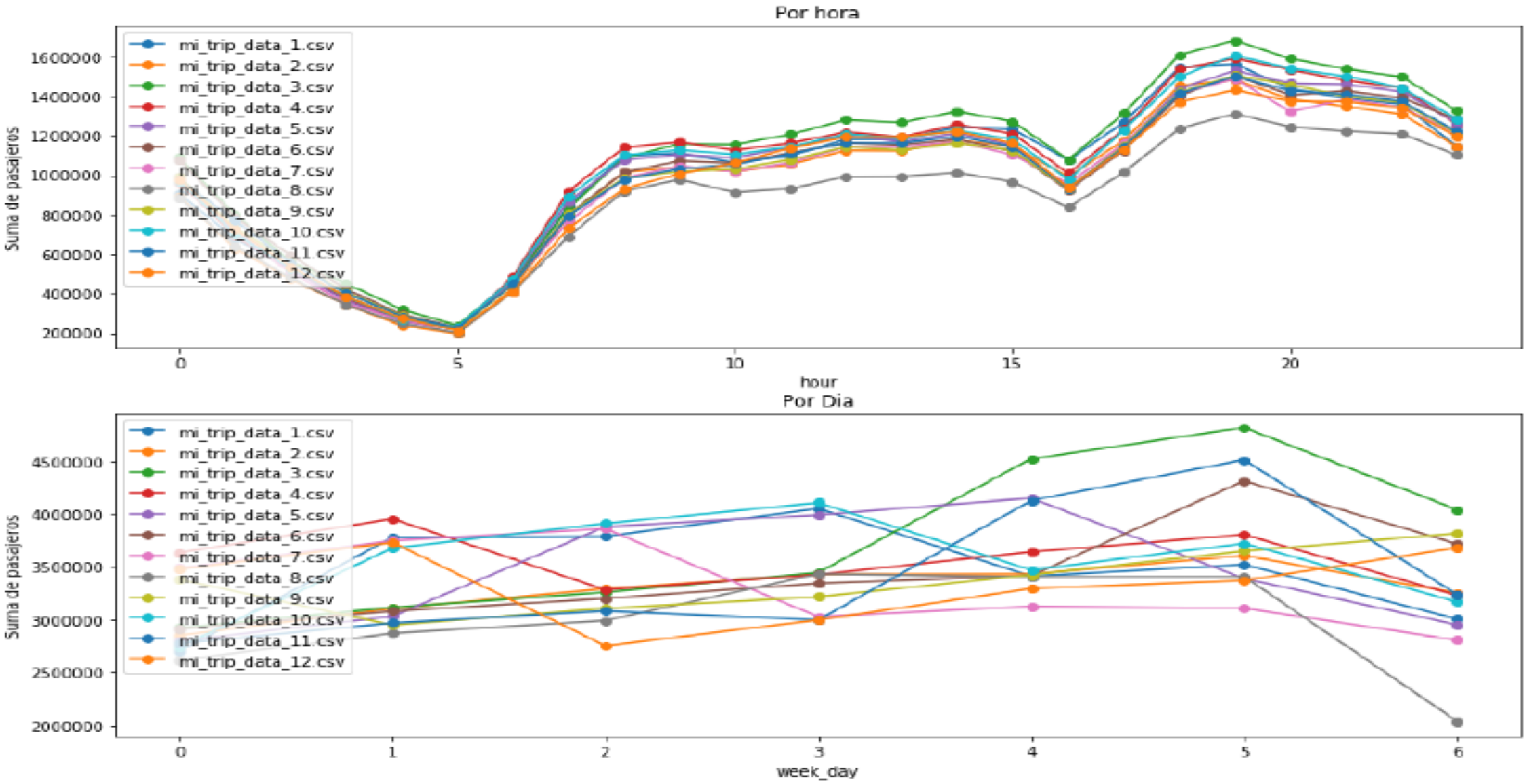
8.- DE LOS VIAJES LARGOS, IDENTIFICAR EL NUMERO DE TAXIS DIFERENTES QUE VEHICULOS SON LOS QUE MAS VIAJES REALIZAN EN CADA MES? SON EL MISMO VEHICULO?

- 6BD1B641A1CD55803A21560299B985A7 6
- DAF57CF25F00457CC6077CD628EC71AC 6
- 1E7C1EB194CCFD58634305DBE0588B85 4
- 4E834DFB7A8831D0A5B6F9B80092A61F 4
- 8C3DB699DA9D5A86780602001DF0892D 3
- 89AC2013723DAC319A06BB59B812F546 3
- D4B5952D54FE462DD585DC1F2865471A 3
- 8653D9D8B4AC5647FE3602D28240EE40 3
- 4DC83B013057503607E85BC4BB6581CE 2
- 570D50E20C4E20D4428EBF94F11DF190 2
- 2344ACDBE31A22CED6B7782FB8E6960B 2
- 698A6074D905BC18FE001032FA0B2048 2
- A6086D46B46031992FE603D412D80F8E 2
- A18CC3E9191D21F604DFC2423916E6A2 2
- D03A5E091B05BBD0E369D71B211FA618 2
- 91FD7951320B9228789AF90613D90254 2
- 7E5A55739A6EFA325F650FD7739135BF 2
- 8211BE04462ADE4621B68E1DFEA54754 2
- DF766411CEF1BFC682CA77FEF6662310 2
- 291470C747FCEEEB97C6A2CBB46A46AF 2
- 62A9741FAFBF7A57E1E3AB673529B45E 2

SE OBTUVIERON DE CADA UNO DE LOS ARCHIVOS SOLO LOS 10 MEJORES IDENTIFICADOS POR EL MEDALLION, EN TOTAL FUERON 120, DE ESOS 81 FUERON DIFERENTES LOS QUE SE MUESTRAN EN LA IMAGEN SON: EL LISTADO DE LOS TAXIS QUE APARECEN MAS DE UNA VEZ.

9.- HACER UNA GRÁFICA TEMPORAL DEL NÚMERO TOTAL DE PASAJEROS AGREGADOS POR DÍAS DE LA SEMANA Y DESPUÉS POR HORAS DEL DÍA. SUGERENCIA: AGREGAR POR EL CORRESPONDIENTE PERIODO TEMPORAL Y GRAFICAR. ¿QUE PUEDEN OBSERVAR DE LAS GRAFICAS?

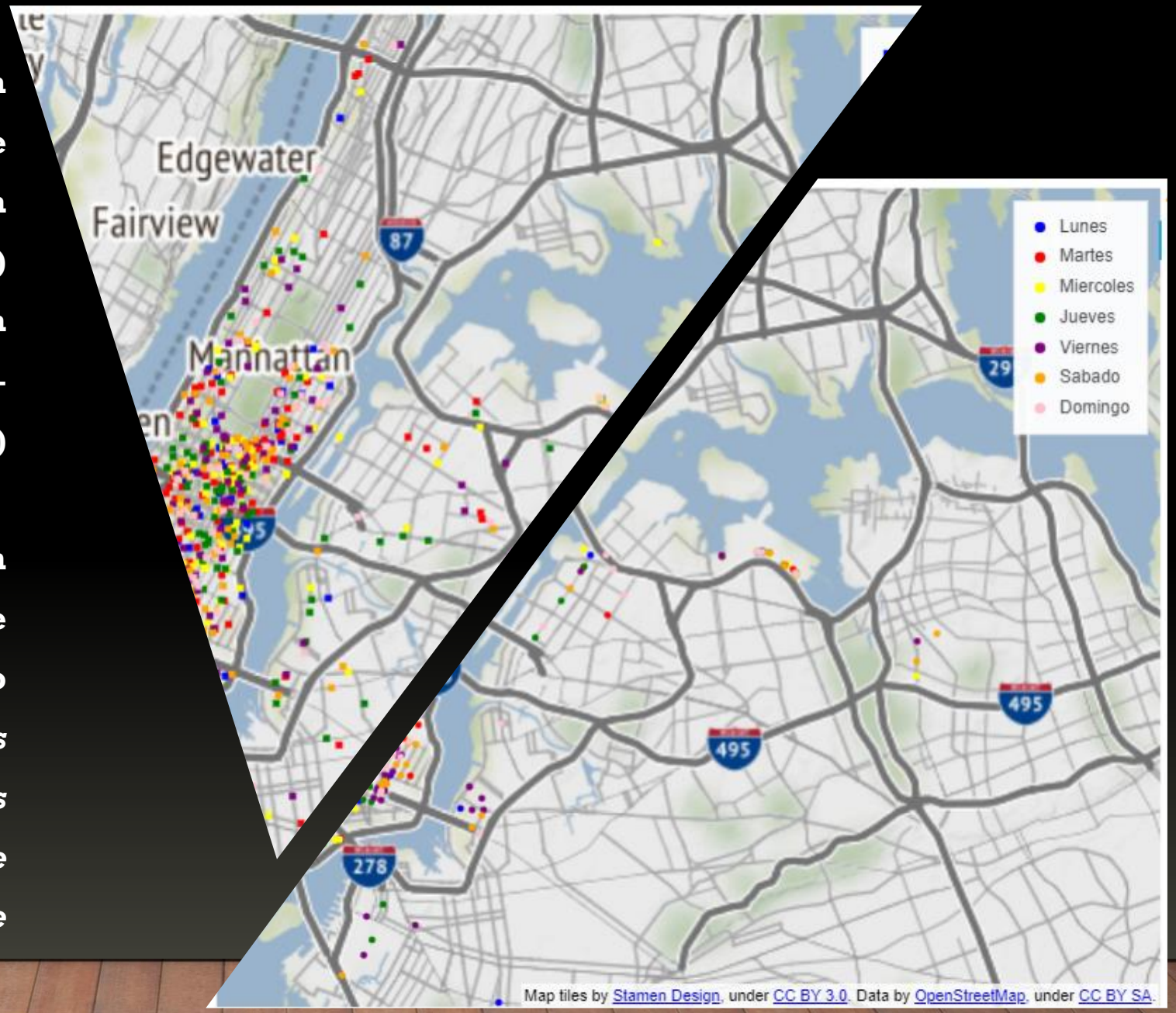
- Se puede observar que por hora tiene una tendencia (patrón) independiente del mes que se esté manejando, está muy marcado el número de pasajes que aborda a cada hora del día. Con lo anterior se podría recomendar a un taxista para maximizar el número de corridas, trabajar de 5 pm a 12 am y podría descansar sin problemas todos los días de 1 am a 7 am.
- En cambio, por día de la semana mes con mes el comportamiento es diferentes, lo cual no marca una tendencia clara, solo se puede apreciar que los viernes, sábados y domingos son los días con más número de pasajes en los taxis, pero no es una regla que en todos los meses se cumpla.
- También se observa que el mes con menos pasaje a toda hora y todos los días de la semana es agosto, puede ser que exista un periodo vacacional. Y marzo es el mes con mayor número de pasajeros, pero corroborar esa tendencia se tendría que comparar varios años.


Por hora
mi_trip_data_1.csv
mi_trip_data_2.csv
mi_trip_data_3.csv
mi_trip_data_4.csv
mi_trip_data_5.csv
mi_trip_data_6.csv
mi_trip_data_7.csv
mi_trip_data_8.csv
mi_trip_data_9.csv
mi_trip_data_10.csv
mi_trip_data_11.csv
mi_trip_data_12.csv
Suma de pasajeros
1600000
1400000
1200000
1000000
800000
600000
400000
200000
0
5
10
15
20
hour
Por Dia
mi_trip_data_1.csv
mi_trip_data_2.csv
mi_trip_data_3.csv
mi_trip_data_4.csv
mi_trip_data_5.csv
mi_trip_data_6.csv
mi_trip_data_7.csv
mi_trip_data_8.csv
mi_trip_data_9.csv
mi_trip_data_10.csv
mi_trip_data_11.csv
mi_trip_data_12.csv
Suma de pasajeros
4500000
4000000
3500000
3000000
2500000
2000000
0
1
2
3
4
5
6
week_day

10. Elegir el vehículo con más viajes en cada mes y graficar en un mapa los sitios donde se suben pasajeros agrupados por o día de la semana (un color distinto para cada día) o 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.)

11. Repetir para las posiciones donde se bajan los pasajeros para el mismo vehículo donde se bajan los pasajeros para el mismo vehículo

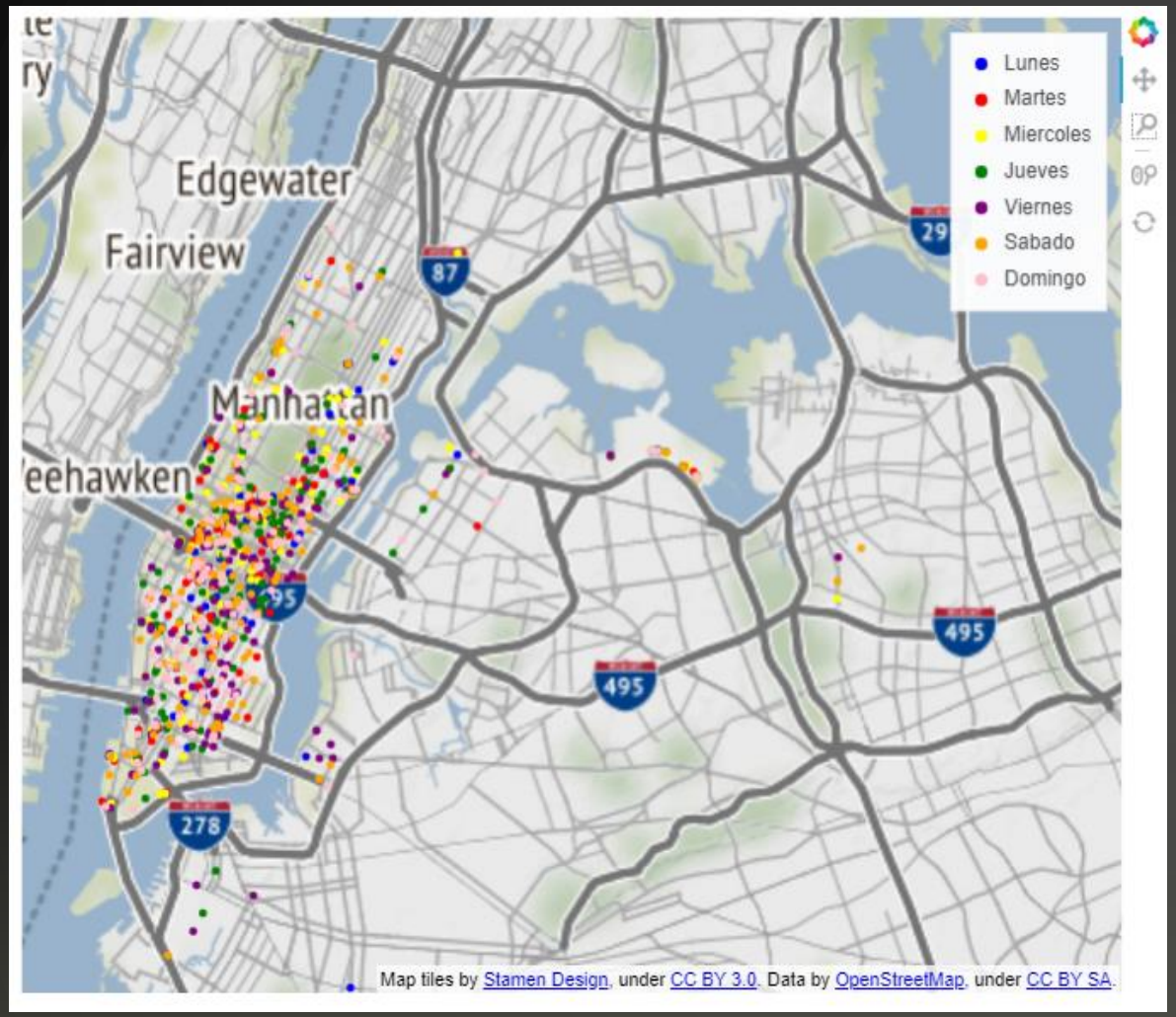
En el archivo de colab se pueden ver todos los gráficos aquí solo esta una muestra de los mapas generados correspondientes al mes de Enero del taxi con más viajes largos en ese mes.

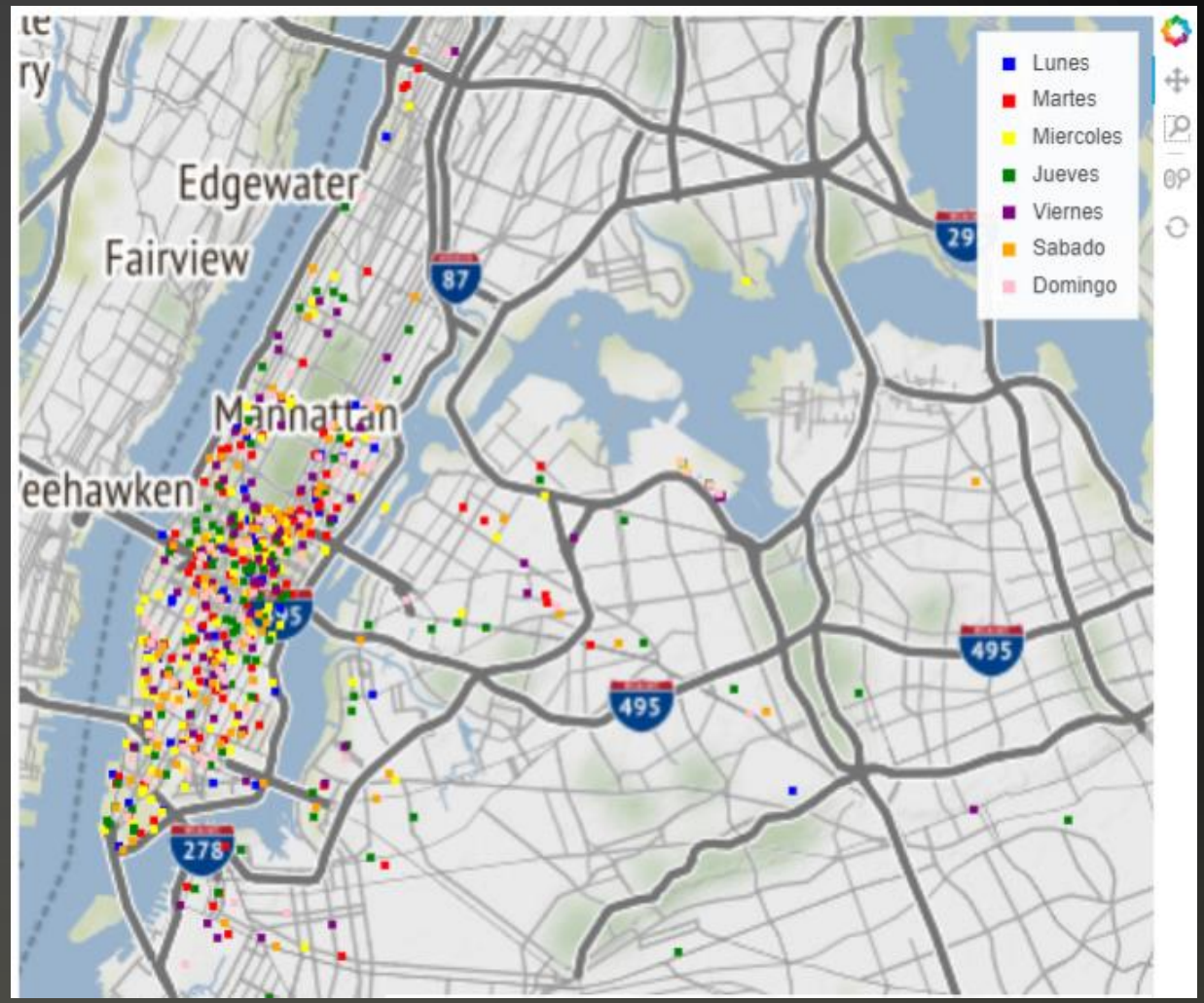


TAXIS CON MAS VIAJES LARGOS POR MES

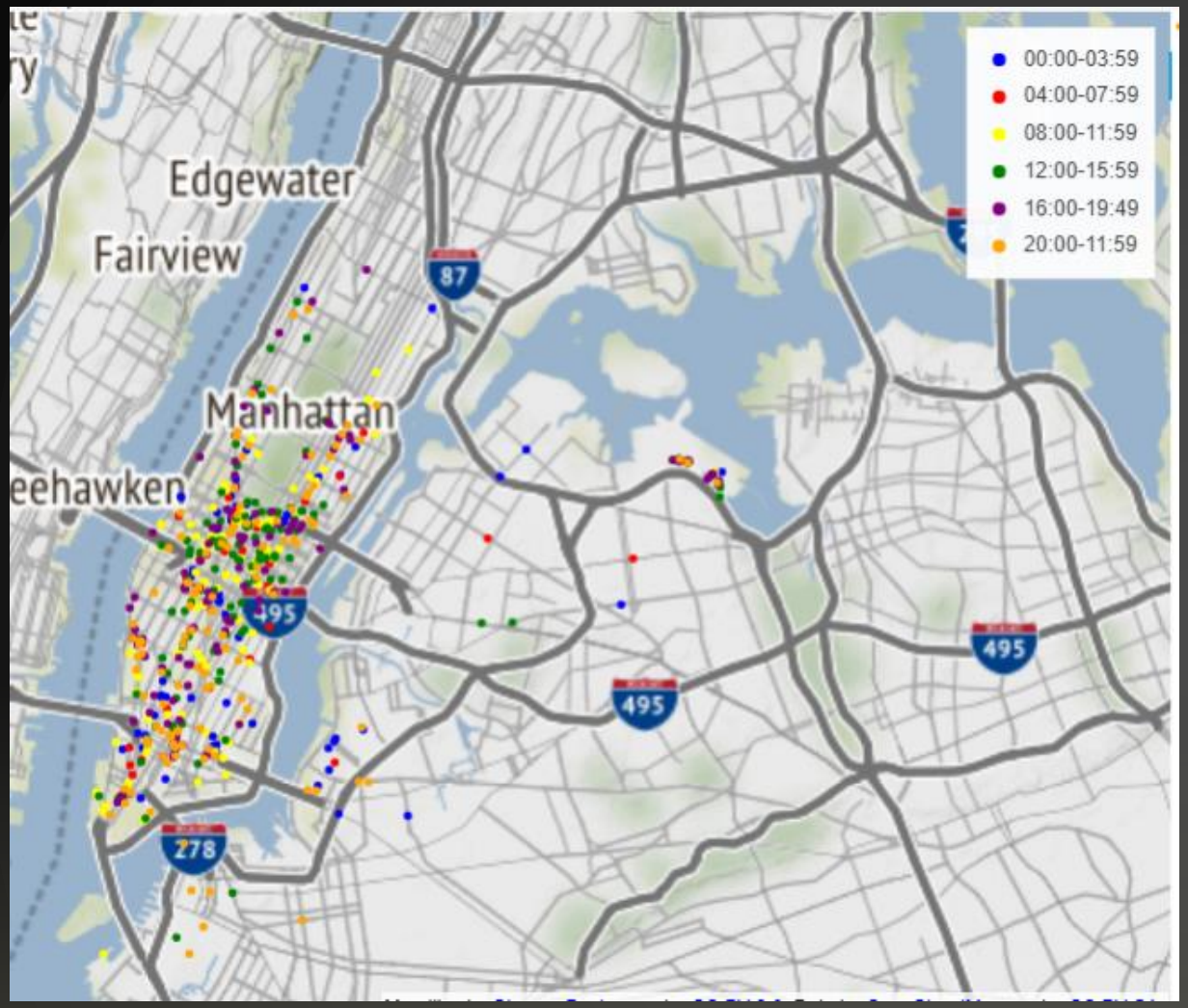
- Enero:DAF57CF25F00457CC6077CD628EC71AC -- 296
- Febrero: 4E834DFB7A8831D0A5B6F9B80092A61F -- 249
- Marzo: 97E2116EEB09AF20718CC464A13675EB -- 271
- Abril: 4E834DFB7A8831D0A5B6F9B80092A61F -- 278
- Mayo: 139D73A09A56D037BE7C56792C7D1FB6 -- 306
- Junio: 698A6074D905BC18FE001032FA0B2048 -- 296
- Julio: 5466D71460137129903C01FB08BB93B -- 296
- Agosto:6BD1B641A1CD55803A21560299B985A7 -- 321
- Septiembre: 8C3DB699DA9D5A8678060200 1DF0892D -- 317
- Octubre: D242F08982116B6C6EBEF33FBCC14513 -- 326
- Noviembre: 1239300107099B26BF07526F66C30BAF – 315
- Diciembre: BDF61165DAA42F17D35F5875F01B5C7A -- 311

Enero: DAF57CF25F00457CC6077CD628EC71AC--296

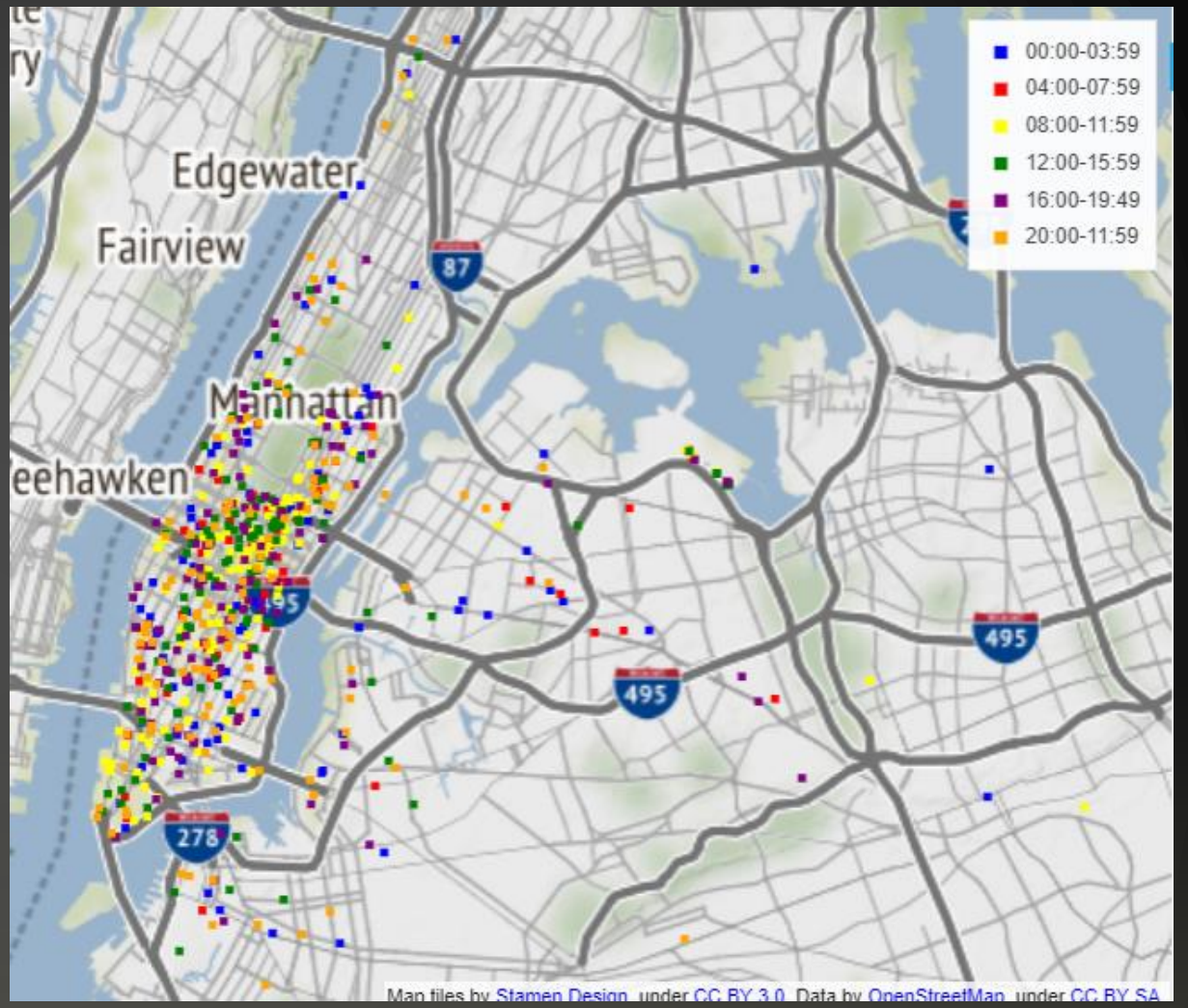




- Pick up
- Dropoff



- Pick up



- Dropoff

COMPAÑEROS Y DOCENTE.

Gracias ...

!

Por la atención prestada.