Sistemas Distribuidos

Primera evaluación

Se evaluarán los siguientes puntos:

Importar datos

1. Adquirir los datos.
    - El conjunto de datos con el que van a trabajar esta disponibles en este enlace y consiste en 12 archivos de texto que tiene informacion del viajes de taxis en la ciudad de NY durante el 2013.
    - En la red interna de LANIA usando cable la descarga del archivo comprimido desde el enlace tomó entre 5 y 10 minutos. Tomen sus provisiones
    - El archivo comprimido tiene un tamaño de aproximadamente 4G, el tamaño de archivo mensual oscila entre 2.0 y 2.5G. La base de datos completa es de aproximadamente 30G.
2. Una vez descargado el archivo pueden descomprimirlo usando la siguiente instruccion

    `!7z x trip_data.7z -o/content/` sin embargo si intentan descomprimir todos los archivos en su sesion se van a terminar el espacio de disco disponible. En esta practica solo van a trabajar con datos correspondientes a los primeros cuatro meses.

Nota: Se inicio trabajando con colab desde un entorne de ejecución remoto después se migro a trabar a un entorno de ejecución local. Lo anterior debido que las sesiones caducan cada 12 horas y los archivos había que descargarlos y depurar los cada expira la sesión, no era muy útil. A demás se conectó con Google Drive y desde ahí se cargaban los archivos, pero a pesar de borrarlos el espacio en disco no se liberaba. Lo cual hacía que la sesión cerrara.

3. Responder las siguientes preguntas:
    1. Cuantas columnas contiene cada archivo de datos descomprimido:

        Todos los archivos tienen 14 columnas

    2. Cuantos renglones tiene cada archivo:

        trip_data_1.csv tiene 14776615 registros
        trip_data_2.csv tiene 13990176 registros
        trip_data_3.csv tiene 15749228 registros
        trip_data_4.csv tiene 15100468 registros
        trip_data_5.csv tiene 15285049 registros
        trip_data_6.csv tiene 14385456 registros
        trip_data_7.csv tiene 13823840 registros
        trip_data_8.csv tiene 12597109 registros
        trip_data_9.csv tiene 14107693 registros
        trip_data_10.csv tiene 15004556 registros

trip_data_11.csv tiene 14388451 registros
trip_data_12.csv tiene 13971118 registros

4. Identificar si existen renglones con errores en los datos, por ejemplo si hay columnas de mas (o de menos), si hay campos vacios, etc. Si se detectan renglones con errores:
    0. 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

    1. crear un nuevo conjunto de datos con los errores eliminados

Explorando los archivos se vio que había registros que tenían distancia con valor Cero, otros que código de región eran muy grandes, que el número de pasajeros era cero o mayor a 9, lo cual sería un número no valido. y otros que sus coordenadas de abordaje están fuera del rango de la ciudad de Nueva York.

5. Crear un `DataFrame` usando la libreria `pandas` y responder lo siguiente:
    0. Indicar el tiempo en segundos, que tarda la libreria `pandas` en leer un archivo a un `DataFrame`

R=1min 47s

    1. Cual es la huella de memoria del proceso usado para generar el objeto `DataFrame`

R= 5104 kb

    2. Cuanta memoria del sistema se usa para crear el objeto anterior

R=4962 kb

    3. Indicar el tiempo que tarda `pandas` en obtener el promedio de la distancia de viaje (`trip_distance`)

R=59.5 ms

6. Todos los pasos siguientes deben **realizarse** usando la libreria `dask`

- Indicar el tiempo que tarda en obtener el promedio de la distancia de viaje (`trip_distance`)*R=**1min 33s***
- Indicar el tiempo que tarda en obtener el promedio de la duracion de viaje (`trip_time_in_secs`)*R=**53.5 s***
- Con las columnas `pickup_datetime` y `dropoff_datetime` como objetos temporales, crear una nueva columna en los dataframes que se llame `duracion` y compararla renglon a renglon con la columna `trip_time_in_secs`
- Definir un `viaje_largo` como aquellos en los que la duracion del viaje (`trip_time_in_secs`) sea mayor a 20 minutos.

7. Contar el numero de viajes largos en cada archivo

mi_trip_data_1.csv tiene 1584532 viajes tardados
mi_trip_data_2.csv tiene 1641728 viajes tardados
mi_trip_data_3.csv tiene 1964137 viajes tardados
mi_trip_data_4.csv tiene 2100334 viajes tardados
mi_trip_data_5.csv tiene 2297637 viajes tardados
mi_trip_data_6.csv tiene 2255387 viajes tardados
mi_trip_data_7.csv tiene 1951809 viajes tardados
mi_trip_data_8.csv tiene 1797811 viajes tardados
mi_trip_data_9.csv tiene 2289058 viajes tardados
mi_trip_data_10.csv tiene 2408231 viajes tardados
mi_trip_data_11.csv tiene 2226949 viajes tardados
mi_trip_data_12.csv tiene 2276598 viajes tardados

8. De los viajes largos, identificar el número de taxis diferentes (la columna `medallion` contiene un número que identificada a cada uno de los vehículos). ¿Qué vehículos son los que más viajes realizan en cada mes? ¿Son el mismo vehículo?

En el archivo mi_trip_data_1.csv estos son los 10 taxis con más viajes largos:

- DAF57CF25F00457CC6077CD628EC71AC 296
- 8B1E839B6A76E16B17F1A32235E3F7BA 277
- 0076C8327A95E988E721AC33B0FA9D67 253
- 1E7C1EB194CCFD58634305DBE0588B85 249
- 91FD7951320B9228789AF90613D90254 246
- BC41253BA5B3EA6228EC1357F3F1097D 240
- 6BD1B641A1CD55803A21560299B985A7 237
- 832154570CFDD2D7E601518117B187DE 236
- 8211BE04462ADE4621B68E1DFEA54754 232
- BBEFF42C6DF9D215155BBBFF7A0D77FA 226

En el archivo mi_trip_data_2.csv estos son los 10 taxis con más viajes largos:

- 4E834DFB7A8831D0A5B6F9B80092A61F 249
- 89AC2013723DAC319A06BB59B812F546 224
- D1B13C1DB63506CBF3FE37BB6EDF8C16 221

- FCB1BF2054823AB4F0D2A35BB5A26A11 220
- D03A5E091B05BBD0E369D71B211FA618 219
- F9B3A00E6DDCA4F8BF2560DFF36B9E91 214
- D4B5952D54FE462DD585DC1F2865471A 213
- 56EC8E3AA6218867A1341249F26531F3 213
- 4F5BFF21FA397E0B48A18BAAFFEBC2DB 213
- 3F8006D3B159447C6A3FCEB015C791DA 212

En el archivo mi_trip_data_3.csv estos son los 10 taxis con más viajes largos:

- 97E2116EEB09AF20718CC464A13675EB 271
- D2A7720C48ED8BA7FB43E4C6A56D071A 265
- 4E834DFB7A8831D0A5B6F9B80092A61F 264
- 89AC2013723DAC319A06BB59B812F546 258
- D4B5952D54FE462DD585DC1F2865471A 256
- A1F57C8D764D70CD18D2C0F2A98A6D0D 256
- 91FD7951320B9228789AF90613D90254 254
- 2344ACDBE31A22CED6B7782FB8E6960B 252
- FB30B64440B4A7B8DBA9903C7598AD90 252
- 89CBEA6E90D9A967338614180FE2A826 251

En el archivo mi_trip_data_4.csv estos son los 10 taxis con más viajes largos:

- 4E834DFB7A8831D0A5B6F9B80092A61F 278
- 291470C747FCEEEB97C6A2CBB46A46AF 274
- 1E7C1EB194CCFD58634305DBE0588B85 272
- C1A040C016496F3E4CD044EC8074AEA3 268
- E8D4DBC75C6D57BC7D59611125EAD764 267
- D03A5E091B05BBD0E369D71B211FA618 267
- 4313CEF5E971DE658618F468688EE2C2 263
- EC6B0947FCCC49473DE94FCBFEECCE82 261
- 4DC83B013057503607E85BC4BB6581CE 261
- 543340C535A57F9C695443B8CF6DE602 260

En el archivo mi_trip_data_5.csv estos son los 10 taxis con más viajes largos:

- 139D73A09A56D037BE7C56792C7D1FB6 306
- 4E834DFB7A8831D0A5B6F9B80092A61F 295
- E7799CBC76E9F9E7F70880736B918535 294
- 291470C747FCEEEB97C6A2CBB46A46AF 290

- D647D82A8DB7E741EFE115A29CA74322 289
- 4B80F28248C15D3DF27A0B158815D156 287
- A69EF0FF887121C9B29A0169D55653A1 286
- 2344ACDBE31A22CED6B7782FB8E6960B 285
- 7E5A55739A6EFA325F650FD7739135BF 285
- D014BF3B8983EB73F40D4E48BB973932 284

En el archivo mi_trip_data_6.csv estos son los 10 taxis con más viajes largos:

- 698A6074D905BC18FE001032FA0B2048 296
- 20801A9439A5BED79DFD9E1C4F833BF3 294
- D4B5952D54FE462DD585DC1F2865471A 290
- 1E7C1EB194CCFD58634305DBE0588B85 288
- DF766411CEF1BFC682CA77FEF6662310 283
- FEBFB5478D15AE3E06E1D0CA674A4C38 280
- 4831383F7FDD5D0DFA715D124BB92BEB 279
- 4018E4BAAF8421F30ADDC502F5BE67EA 278
- 7E3BF2C5869112F3997C2B790E4894FD 277
- E5732B5D88740FBB7D46FB56580818A8 277

En el archivo mi_trip_data_7.csv estos son los 10 taxis con más viajes largos:

- 5466D714601371299033C01FB08BB93B 296
- DAF57CF25F00457CC6077CD628EC71AC 274
- 6BD1B641A1CD55803A21560299B985A7 269
- 0C351740D5081DF4329D6328057F9D44 269
- BC748B498030D391D5ACEA4F0D323D33 266
- 698A6074D905BC18FE001032FA0B2048 261
- 739323AC15DADABE4B54561CB3330C53 256
- 720E3681E2193AD9E52FDD4C6C5B98F9 256
- DF766411CEF1BFC682CA77FEF6662310 255
- 8211BE04462ADE4621B68E1DFEA54754 254

En el archivo mi_trip_data_8.csv estos son los 10 taxis con más viajes largos:

- 6BD1B641A1CD55803A21560299B985A7 321
- 91C232C3BFDC86036057494653C55307 304
- DAF57CF25F00457CC6077CD628EC71AC 303
- 62A9741FAFBF7A57E1E3AB673529B45E 278
- E7765354E342B79E2BBDC5F57777F8C4 272

- BE63343BAD5CD6F99EC435812E332445 270
- 89AC2013723DAC319A06BB59B812F546 264
- 4DC83B013057503607E85BC4BB6581CE 260
- B2A23B78DA7C84A229AF5932A286778C 255
- 570D50E20C4E20D4428EBF94F11DF190 254

En el archivo mi_trip_data_9.csv estos son los 10 taxis con más viajes largos:

- 8C3DB699DA9D5A86780602001DF0892D 317
- 570D50E20C4E20D4428EBF94F11DF190 315
- 62A9741FAFBF7A57E1E3AB673529B45E 301
- 6BD1B641A1CD55803A21560299B985A7 293
- B35B40A7C6563AAC24DF7566CD3434E1 293
- 07CE358C570EBBACBA9E95591C6728C1 278
- 1E7C1EB194CCFD58634305DBE0588B85 278
- 92D01E396D7AC7BD09183A7ED793E274 278
- EC7080E2881E27B08D8F7DEEED9640E8 277
- 7376BAC10BB8455E4AE6A7C3C4552458 276

En el archivo mi_trip_data_10.csv estos son los 10 taxis con más viajes largos:

- D242F08982116B6C6EBEF33FBCC14513 326
- 7E5A55739A6EFA325F650FD7739135BF 325
- DAF57CF25F00457CC6077CD628EC71AC 314
- A6086D46B46031992FE603D412D80F8E 307
- E72CD211B488A8BFADE73DB9B1961F96 307
- 8C3DB699DA9D5A86780602001DF0892D 306
- A6217A107F6C4F243004B51A9FFBC2A8 301
- 240CF09159B8C1890887FF6D61E812AC 299
- 0DC372E1F0AEC7EBBBC97020397DB3E2 296
- 8653D9D8B4AC5647FE3602D28240EE40 296

En el archivo mi_trip_data_11.csv estos son los 10 taxis con más viajes largos:

- 1239300107099B26BF07526F66C30BAF 315
- 8653D9D8B4AC5647FE3602D28240EE40 303
- A18CC3E9191D21F604DFC2423916E6A2 293
- A6086D46B46031992FE603D412D80F8E 292
- 8C3DB699DA9D5A86780602001DF0892D 289
- DAF57CF25F00457CC6077CD628EC71AC 289

- 740B7097FB78DDD012A2CFA4309EB66A 286
- 302EB02F36343A64C452D49463A27C88 281
- 6BD1B641A1CD55803A21560299B985A7 279
- 8321B13849E449225B843F65FC7441AF 279

En el archivo mi_trip_data_12.csv estos son los 10 taxis con más viajes largos:

- BDF61165DAA42F17D35F5875F01B5C7A 311
- A18CC3E9191D21F604DFC2423916E6A2 310
- B80EB4255C8015FA65F7FDDEB24EB8F3 298
- CE8E639180CE24EAAF7007A696B90AE2 293
- 8653D9D8B4AC5647FE3602D28240EE40 293
- E0EC8C572F4B6FC66EE7928BC3FA409A 286
- DAF57CF25F00457CC6077CD628EC71AC 284
- 2DC01415E5E2D7CB1D1E07DE4BA12844 284
- 8DEB70907D00AA1D7FF5E2683240549B 283
- 6BD1B641A1CD55803A21560299B985A7 281

Se los anteriores 120 taxis, solo hay 81 diferentes

El listado de los carros con más de una aparición es:

- 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

El listado de los primeros lugares por cada mes son:

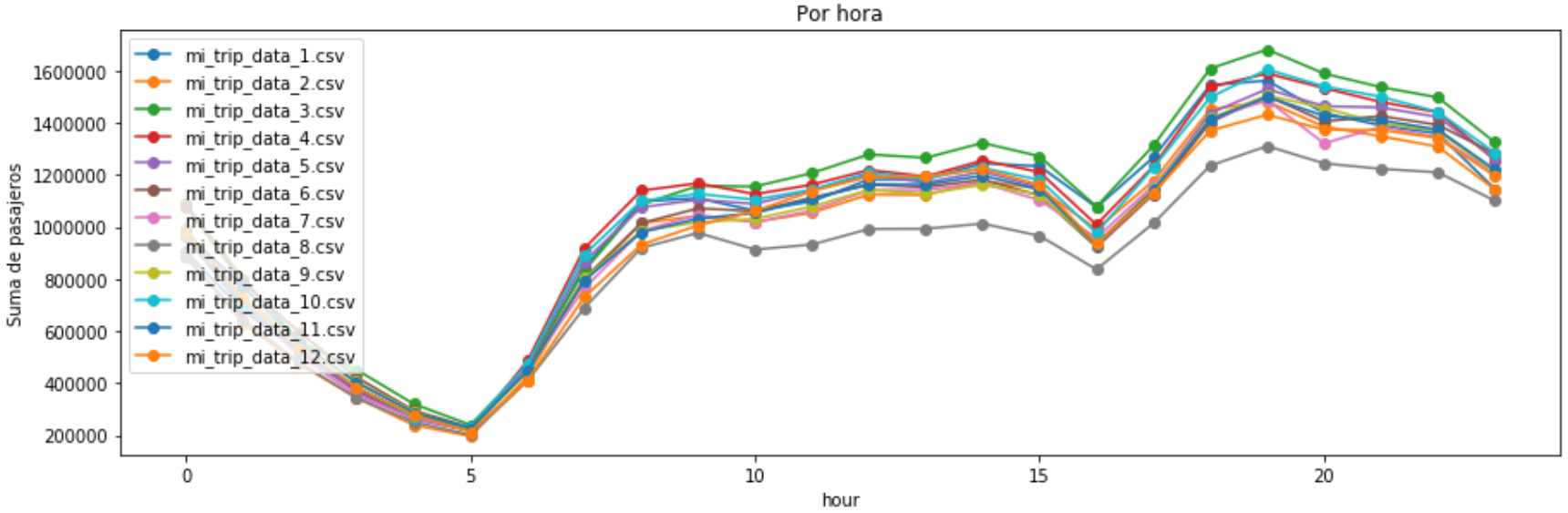
Mes	Medallion
1	DAF57CF25F00457CC6077CD628EC71AC
2	4E834DFB7A8831D0A5B6F9B80092A61F
3	97E2116EEB09AF20718CC464A13675EB
4	4E834DFB7A8831D0A5B6F9B80092A61F
5	139D73A09A56D037BE7C56792C7D1FB6
6	698A6074D905BC18FE001032FA0B2048
7	5466D7146013712990 33C01FB08BB93B
8	6BD1B641A1CD55803A21560299B985A7
9	8C3DB699DA9D5A86780602001DF0892D
10	D242F08982116B6C6EBEF33FBCC14513
11	1239300107099B26BF07526F66C30BAF
12	BDF61165DAA42F17D35F5875F01B5C7A

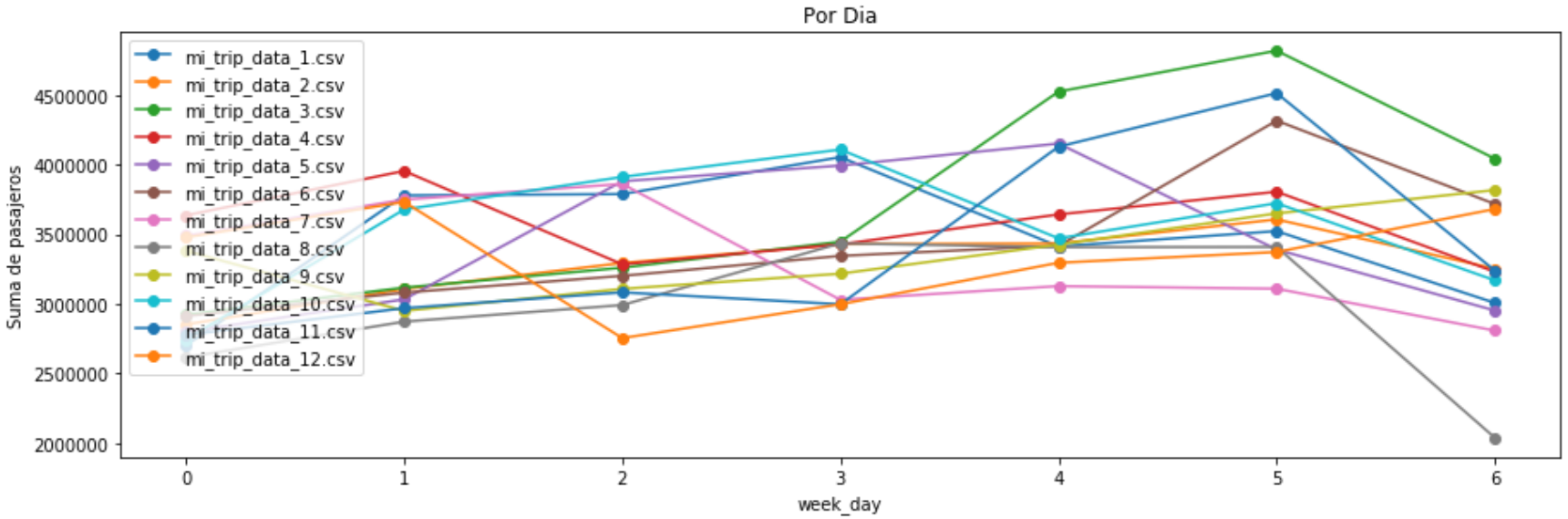
De lo anterior podemos concluir que hay taxis que tiene clientes fijos o muy bien estudiado el flujo de pasajeros, ya que 9 de lo 12 primeros lugares aparecen mas de una ves la lista de los 120 taxis con más viajes largos por año.

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 (patron) 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.

Luis Rafael Flores Castelán Examen Medio Sistemas Distribuidos2
MAC-LANIA



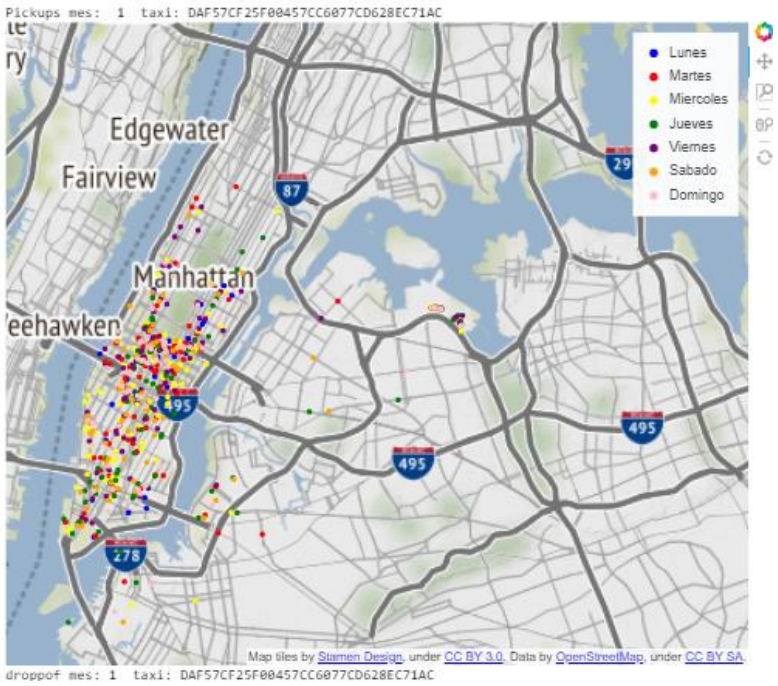


Luis Rafael Flores Castelán Examen Medio Sistemas Distribuidos2
MAC-LANIA

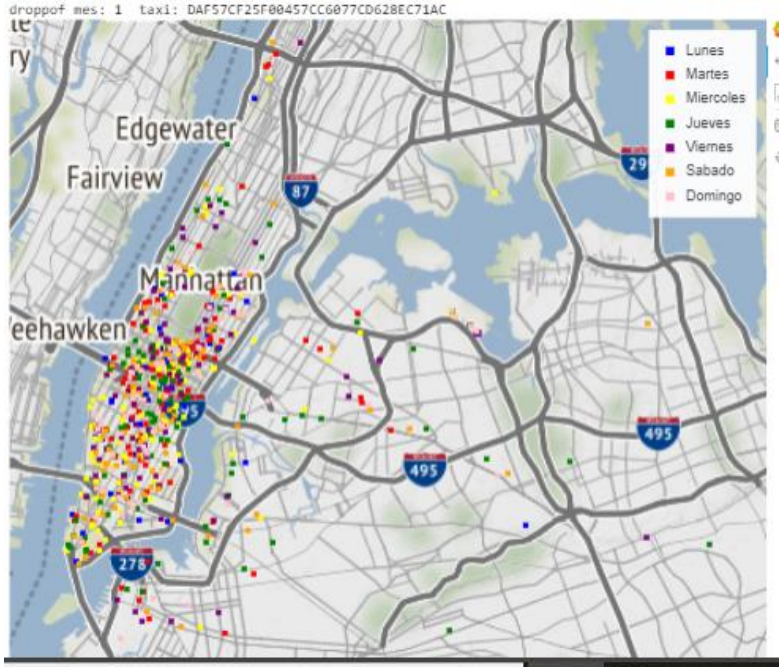
10. Elegir el vehiculo con mas viajes en cada mes y graficar en un mapa los sitios donde se suben pasajeros agrupados por
    - dia de la semana (un color distinto para cada dia)
    - hora del dia (un color distinto para cada intervalo de cuatro horas,00:00 - 03:59, 04:00--07:59, 08:00-11:59, etc.)
11. Repetir para las posiciones donde se bajan los pasajeros para el mismo vehiculodonde se bajan los pasajeros para el mismo vehiculo

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

Subidas mes 1 por dia

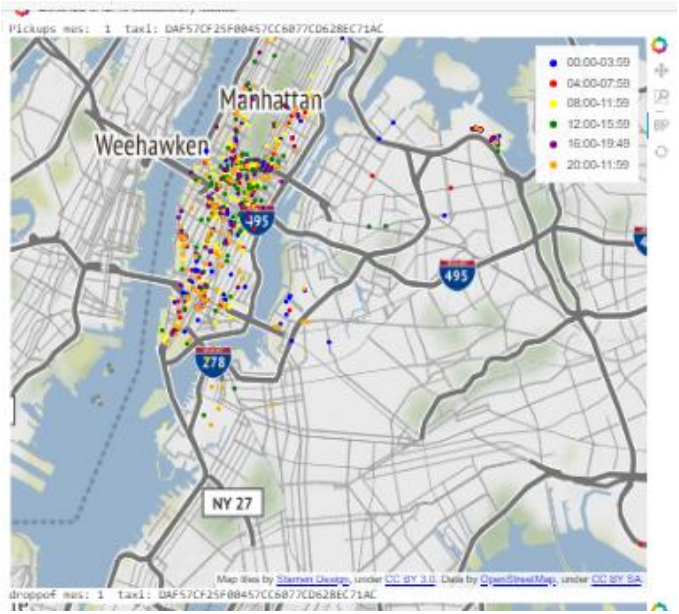


Bajadas mes uno por dia



Luis Rafael Flores Castelán Examen Medio Sistemas Distribuidos2
MAC-LANIA

Subidas por rango de horas mes uno



Bajadas por rango de horas mes uno