PRÁCTICA 3 CICLO DE VIDA ANALÍTICO DEL DATO

PERSISTENCIA Y BUSCADORES

LUCIAN IACOB

1.- Descripción de los datos

En esta práctica vamos a trabajar con famoso dataset "TLC Trip Record Data", el cual recoge los viajes en taxi que tienen lugar en la ciudad de Nueva York.

Debido al gran volumen de datos que representa este dataset se ha decidido trabajar solo con el año 2019, y dentro de este año, solo con los viajes de los icónicos taxis amarillos.

Tras una serie de transformaciones en los datos originales hemos acabado utilizando las siguientes variables:

- ID: Número identificador del registro.
- PU_datetime: Marca de tiempo en formato "YYYY-MM-DD HH:mm:ss" en la que se inicia el viaje.
- Borough: El barrio en el que se desarrolla el viaje (Bronx/Brooklyn/Manhattan/Queens/Staten Island). Nos hemos quedado solo con viajes en los que no se cambie de barrio.
- Trip_duration_min: Duración del viaje en minutos
- Distance_miles: Longitud del recorrido medida en millas.
- Tip: Propina que se le da al taxista en dólares (Solo se registran las propinas dadas con pagos con tarjeta).
- Total: Costa total del viaje aparte de la propia. También en dólares.

Conviene destacar que por mes hay unos 6-7 millones de registros. Para reducir esta cantidad masiva de datos, se ha tomado una muestra aleatoria de unas 120 mil entradas por cada mes.

Por otro lado, hemos utilizado de esa misma página lo datos "Taxi Zone Lookup Table" que contienen información de a qué barrio pertenecen cada una de las ms de 100 zonas de servicio en las que se divide la ciudad.

También hemos utilizado datos geográficos con la delimitación de cada uno de los 5 barrios de la ciudad para poder trabajar con mapas al presentar los datos.

2.- Hipótesis que deseamos comprobar

Es conocido que hay una gran diferencia entre estas 5 zonas aunque pertenezcan a la misma ciudad. A lo largo de miles de películas/reportajes/noticias se ha podido apreciar como pasamos de la gran imagen que tenemos de la gran manzana al pensar en Manhattan a la imagen de barrio marginal y peligroso cuando pensamos en el Bronx.

Este fenómeno ha llevado a la siguiente pregunta:

¿Es tal la diferencia entre los barrios que se nota hasta en los viajes realizados en taxi?

Solo para que nos demos cuenta del diferente nivel adquisitivo del que estamos hablando vamos a adjuntar la siguiente tabla obtenida del Baruch College sobre el ingreso per cápita de cada zona:

	2014	2015	2016	2017
Bronx	$31,556	$32,778	$33,310	$35,564
Brooklyn	$41,399	$43,915	$45,629	$48,758
Manhattan	$152,690	$155,779	$164,056	$175,960
Queens	$40,997	$43,216	$44,031	$46,829
Staten Island	$48,123	$50,894	$51,836	$54,908

A lo largo de la práctica iremos realizando consultas a los datos que nos puedan responder a esta pregunta (o al menos que creamos que nos vayan a aportar información útil respecto al tema).

3.- Descubrimiento de los datos

Mediante el uso del buscador ElasticSearch y su panel de visualización Kibana vamos a ver algunas propiedades de los datos antes de lanzarnos a las queries.

Vamos a ver primero el conteo de viajes agrupados por semana para cada distrito:

- Brooklyn

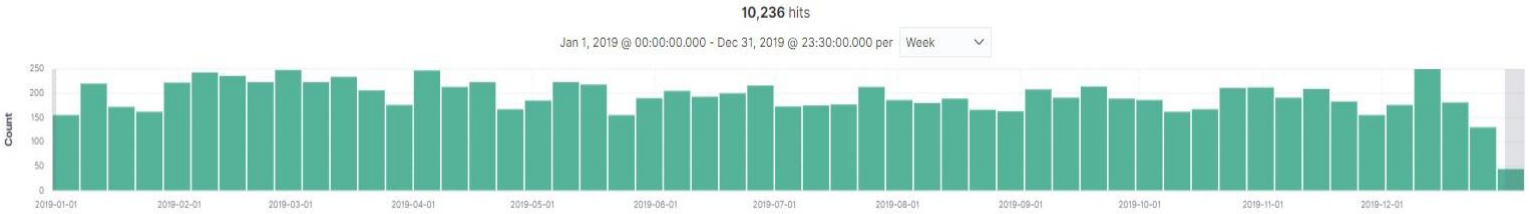

Ilustración 1. Viajes por semana realizados en Brooklyn

- Manhattan

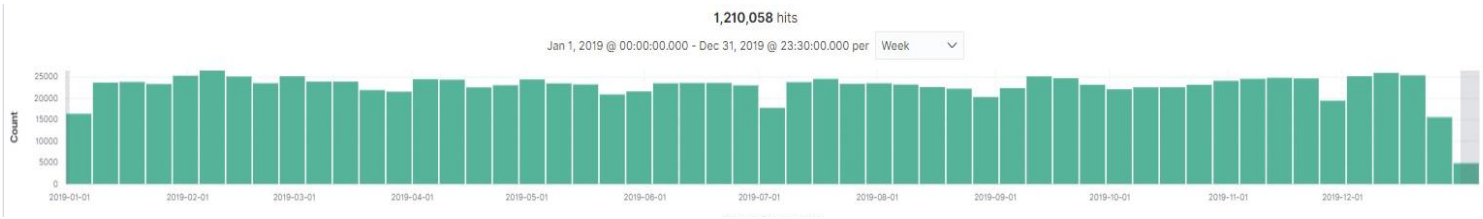

Ilustración 2. . Viajes por semana realizados en Manhattan

- Queens

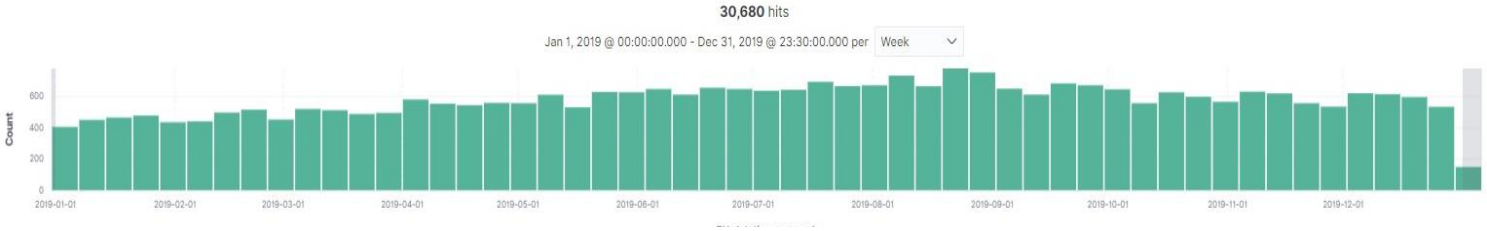

Ilustración 3. Viajes por semana realizados en Queens

- Bronx

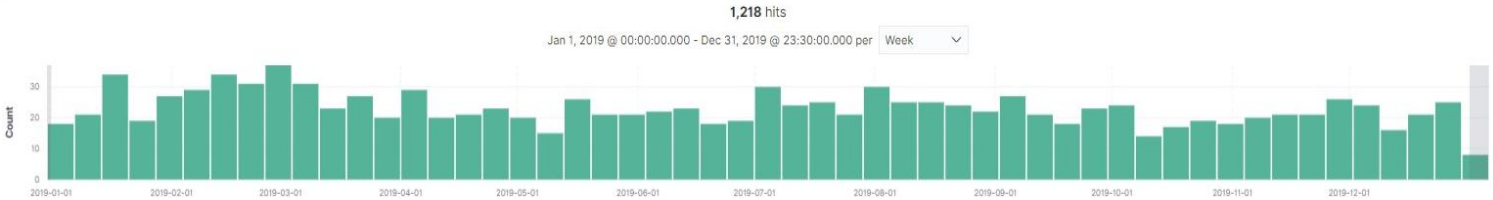

Ilustración 4. Viajes por semana realizados en el Bronx

- Staten Island

Ilustración 5. Viajes por semana realizados en Staten Island

Lo primero que nos llama la atención de estos gráficos es heterogeneidad que hay en la cantidad de viajes realizados según la zona. Aunque solo hemos tomado en torno al 2% de los datos de viajes, la muestra ha sido tomada al azar. Por lo que la diferencia de 1,2 millones de viajes en Manhattan frente a los pocos miles en los demás e incluso solo 30 en Staten Island nos da una pista de que hay una zona claramente dominante.

Esto podría deberse también a razones demográficas, es decir, que la población de Manhattan sea mucho mayor que la de las demás zonas, pero no esto no es así. Atendiendo a los últimos datos, las poblaciones son:

- Bronx: 1,418,207 hab.
- Brooklyn: 2,559,903 hab.
- Manhattan: 1,628,706 hab.
- Queens: 2,253,858 hab.
- Staten Island: 476,143 hab.

Vemos que de hecho se encuentra en tercer lugar respecto a población así podemos ir descartando razones demográficas de este tipo.

También conviene recordar que nos hemos quedado solo con los viajes de taxis amarillos, eliminando los taxis verdes y los vehículos de alquiler, que quizás podrían destacar más en las otras zonas en vez de en Manhattan. Y también es vital señalar que nos hemos quedado con viajes en los que comienzo y final están dentro del mismo barrio.

Por otro lado podemos ver si hay cierta estacionalidad en los viajes según la zona, vayamos una por una:

- Bronx: A principios de año la gente parece viajar más en taxi que en otras épocas. Parece que en el resto del año hay menos movimiento.

- Brooklyn: Sigue una línea similar al Bronx pero sin ser tan acusada la diferencia entre comienzo de año y el resto, habiendo más regularidad.

- Manhattan: Es la zona que presenta más regularidad de todas (quizás por tener más datos). Esto también puede llevarnos a pensar que es un destino turístico muy visitado que hace que en épocas de menos movimiento local (verano por ejemplo) se mantengan los viajes por los turistas.

- Queens: Sigue un patrón distinto a lo anterior, a principio de año es cuando menos viajes venos y luego en épocas que suelen ser más tranquilas aquí vemos más movimiento (posible hipótesis de destino turístico frecuentado otra vez).

- Staten Island: Esta zona presenta tan pocas entradas en nuestra muestra que no podemos deducir estacionalidad alguna por desgracia.

Por último me gustaría mostrar una serie de gráficos aparte:

Si contamos los viajes de todas las zonas y los mostramos agrupando por día en vez de por semana encontramos una regularidad de periodo 7 días aproximadamente, que nos indica una clara regularidad en torno a cuándo se realizan los viajes:

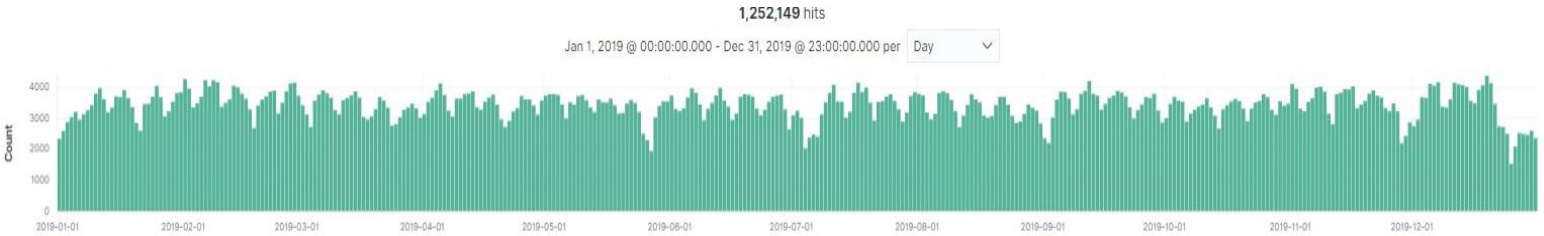

Ilustración 6. Viajes por día realizados en Nueva York

Además de esto me gustaría mostrar el conteo de viajes agrupando por distancia recorrida y por la duración:

- Por distancia recorrida:

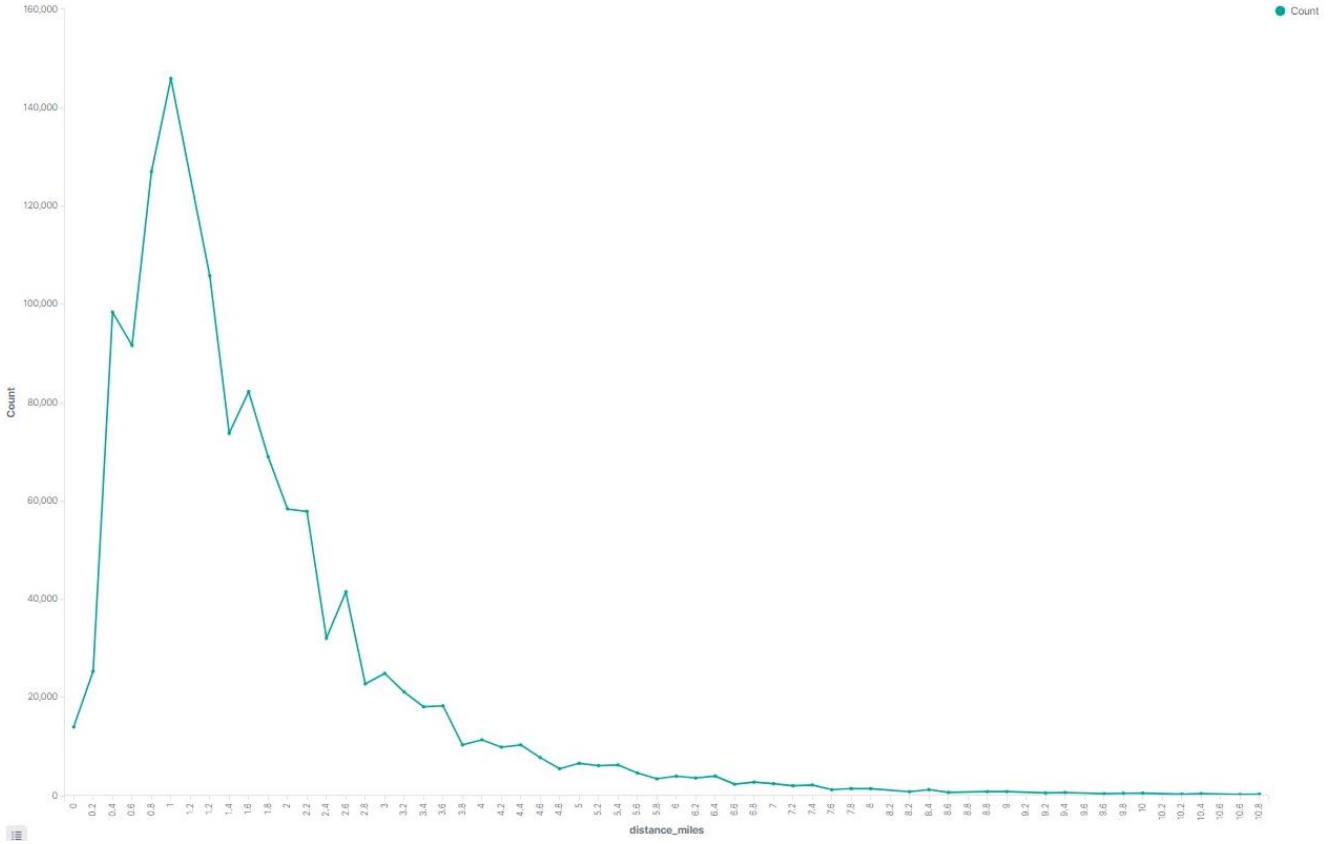

Ilustración 7. Conteo de los viajes en Nueva York según la distancia recorrida

Vemos que el máximo se sitúa en torno a la milla de longitud, viajes más largos resultan extraños.

- Por duración del viaje:

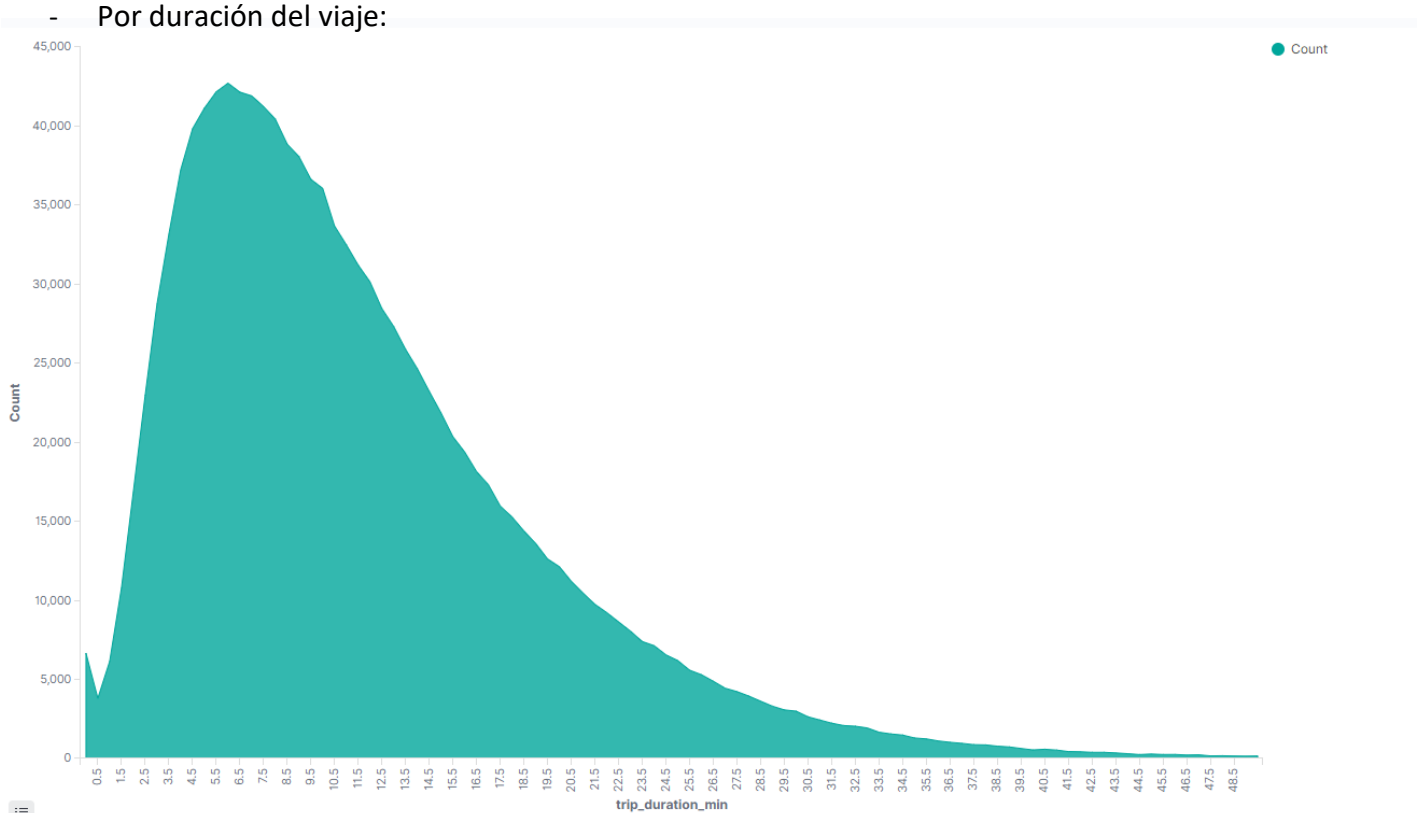

Ilustración 8. Conteo de los viajes en Nueva York según la duración del viaje

Vemos que destacan los viajes cortos en torno a los 6.5 minutos, propio de una ciudad ajetreada en la que uno quiere moverse rápido. Como comentario extra, la función se me asemeja a una distribución chi cuadrado más o menos. Incluso lo que más me inquieta es que me recuerda a la Ley de Planck, pero supongo que eso es debido al bias producido por haber estudiado física porque esa semejanza no es más que mera casualidad.

4.- Analítica realizada

A continuación vamos a realizar una serie de queries a nuestros datos para ver si podemos sacar alguna conclusión. Como lo que queremos es ver diferencias entre los *Borough,* lo que vamos a hacer es ver ciertas medidas agrupadas por zona. Los análisis para cada resultado los dejamos para el final, aquí nos limitamos a mostrarlos:

- Distancia y duración por zona

Hemos calculado la media de la distancia recorrida en los viajes (en millas) y su duración (en minutos). Los presentamos juntos porque son cantidades que suelen ir de la mano (quitando situaciones de mucho atasco). Los resultados son:

```
Total MapReduce CPU Time Spent: 10 seconds 790 msec
OK
Queens  5.53517370621044
Bronx   2.7329556663889187
Brooklyn        2.371383082776211
Staten Island   2.146000004125138
Manhattan       1.8550986583990627
Time taken: 66.906 seconds, Fetched: 5 row(s)
```

Ilustración 9. Media de la distancia recorrida para cada zona

Utilizando funcionalidades de Kibana podemos crear un mapa en el que se nos muestren los 5 barrios de la ciudad y añadir la medida deseada para verlo todo de forma más clara

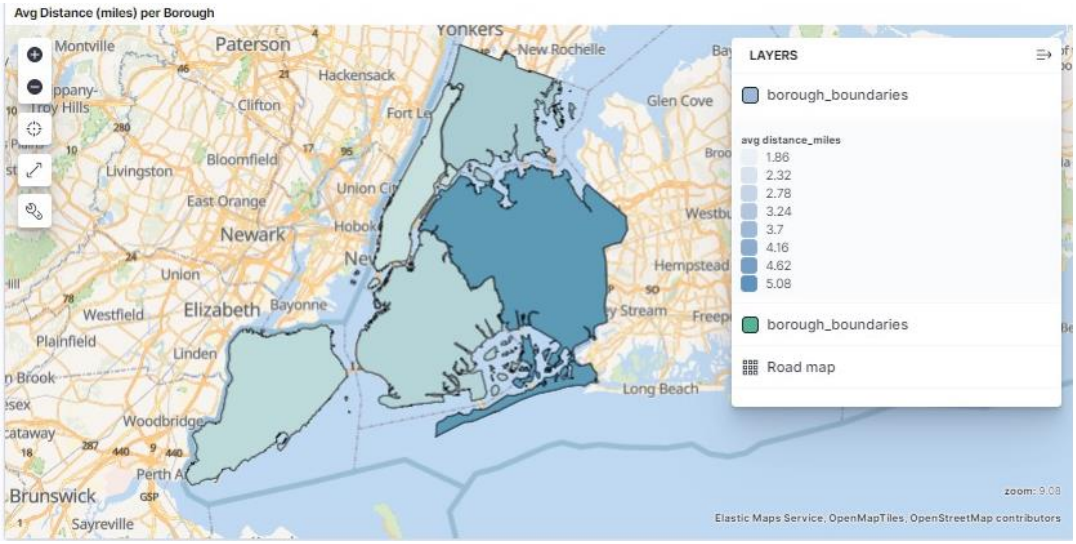

Ilustración 10. Mapa de la media de la distancia recorrida para cada zona.

```
Total MapReduce CPU Time Spent: 11 seconds 120 msec
OK
Staten Island   32.224667293205854
Queens  19.022299885735915
Bronx   15.957717536239041
Brooklyn        15.256462216375803
Manhattan       15.167369638351769
Time taken: 68.626 seconds, Fetched: 5 row(s)
```

Ilustración 11.Media de la duración del viaje para cada zona

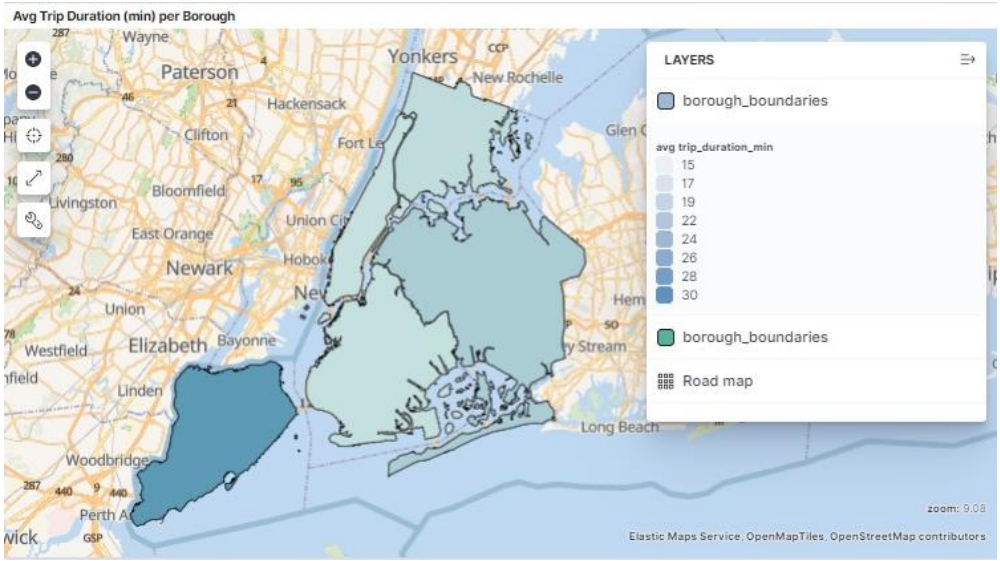

Ilustración 12. Mapa de la media de la duración del viaje para cada zona

- Coste total del viaje y propina por zona (en dólares)

```
Total MapReduce CPU Time Spent: 10 seconds 990
OK
Staten Island   31.161666782697043
Queens  23.65741469704849
Bronx   17.06577167135154
Manhattan       15.039107415594858
Brooklyn        14.367261164075208
Time taken: 69.112 seconds. Fetched: 5 row(s)
```

Ilustración 13. Media del coste total del viaje para cada zona

```
Total MapReduce CPU Time Spent: 12 seconds 150 msec
OK
Manhattan       1.7632642572124126
Queens  1.703742992004088
Staten Island   1.356333335240682
Brooklyn        1.0278902116705924
Bronx   0.2683743829250164
Time taken: 78.647 seconds, Fetched: 5 row(s)
```

Ilustración 14. Media de la propina entregada al taxista para cada zona.

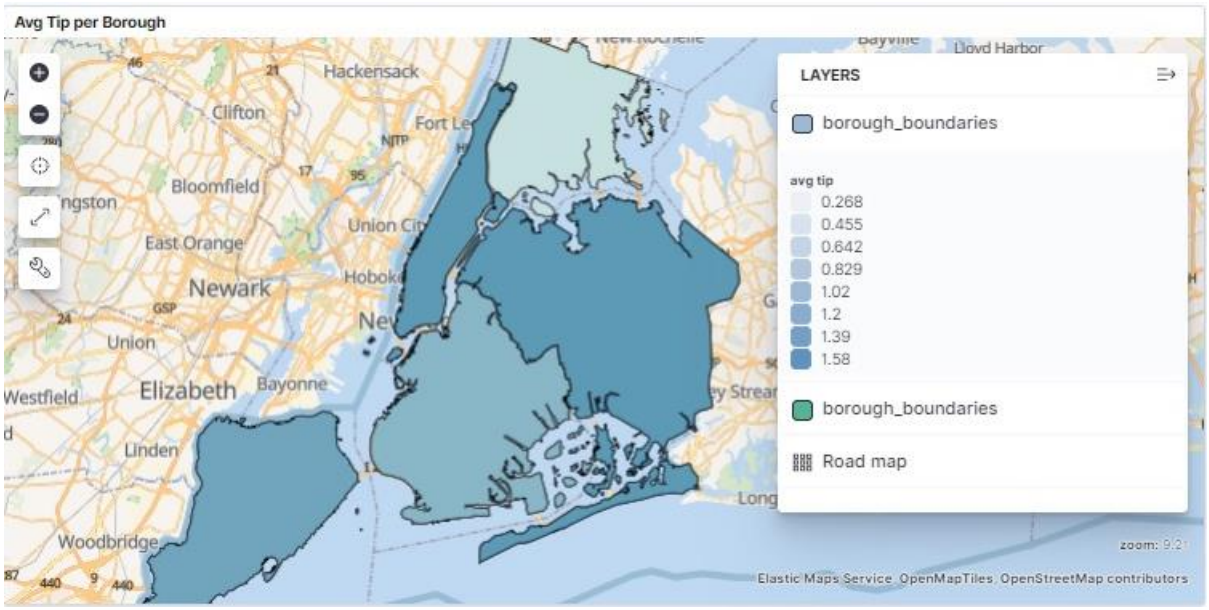

Ilustración 15. Mapa de la media de la propina entregada al taxista para cada zona

Desgraciadamente no sé cómo poner nombre a cada zona en el mapa para que se vea de forma clara cuál es. Pero aquí se puede ver una imagen que indica cuál es cuál.

- Hora y mes punta por zona:

Se ha buscado la hora del día y el mes del año más concurrido de cara a viajes en taxi. Otros periodos de tiempo como el día o la semana no se han considerado al no creer que aporten información más escarificadora que los otros dos.

```
Total MapReduce CPU Time Spent: 3 seconds 80 msec
OK
Bronx   8       84
Brooklyn        23      663
Manhattan       18      82257
Queens  14      1783
Staten Island   12      6
Time taken: 50.962 seconds, Fetched: 5 row(s)
```

Ilustración 16. Hora del día más concurrida y conteo del número de viajes realizados por zona.

```
Total MapReduce CPU Time Spent: 3 seconds 260 msec
OK
Bronx   3       123
Brooklyn        2       956
Manhattan       11      101988
Queens  8       3237
Staten Island   3       7
Time taken: 50.873 seconds, Fetched: 5 row(s)
```

Ilustración 17. Mes del año más concurrido y conteo del número de viajes realizados por zona.

5.- Evaluación de los resultados

Vayamos en el mismo orden que hemos utilizado en el apartado anterior. Como nota aclaratoria previa creo conveniente dejar un poco de lado los resultados obtenidos respecto a Staten Island porque 30 registros frente a la cantidad del resto resultan insignificantes y pueden llevarnos a error.

- Distancia y duración por zona

Vemos que Manhattan se sitúa al final de ambos rankings (aunque no por mucha diferencia) mientras que las dos zonas con menor PIB (Queens y el Bronx) per cápita se sitúan en la cima. La posible forma de ver esto es que cuanto mayor es el nivel adquisitivo de la persona, menos le importa gastarse dinero en un taxi para un viaje corto. En cambio, con poder adquisitivo bajo, si toca un viaje corto uno se lo piensa dos veces y se plantea ir andando o utilizar medios de transporte más baratos, quedando relegado el taxi para viajes más largos donde no quede otra.

- Coste total y propina entregada por zona

El coste total, como era de esperar, sigue la tendencia de los rankings de duración y distancia recorrida en el viaje por lo que poco podemos añadir.

Lo que sí me gustaría destacar es la propina entregada al taxista. Aunque parezca un detalle insignificante e incluso algo muy aleatorio, la propina es algo que está implantado en la sociedad (como un valor general que comparte la gente por decirlo de alguna manera).

Y, como esperaba, se confirma que a mayor PIB per cápita, mayor cantidad de propina se deja en el viaje (quitando a la población de Queens que parecen más generosos que los demás). Siendo la propina media de Manhattan unas 7 veces mayor que la del Bronx.

- Hora punta

Tanto en esta media como en la de mes sí que considero un problema la gran diferencia de registros que hay entre las zonas, porque hasta ahora hemos calculado medias mientras que aquí hemos contado valores.

Vemos que en Manhattan la hora más saturada son las 6 de la tarde, que la podría situar con la hora en la que se termina la jornada laboral de trabajos de tipo oficina (el conocido nine to five que dicen) mientras que en el Bronx tenemos que la hora más transitada son las 8 de la mañana, cuando la gente suele ir a trabajar. En las demás tenemos horas distribuidas de una manera que no logro encontrar un patrón claro que nos indique que hay diferencia entre las zonas debido al diferente nivel de vida. Quizás con más datos se podría ver algo mejor.

- Mes punta

Respecto al mes, nos encontramos en la misma situación de antes, no hay un patrón claro. Lo que es más, si cogemos el número total de registros para cada zona y dividimos por 12 vemos ese valor no es muy distinto al valor máximo que nos devuelve la query. Esto nos indica que todos los meses presentan más o menos la misma de viajes (de nuevo, obviemos los 30 registros de Staten Island), provocando que el mes más concurrido lo sea por algún fenómeno de tipo aleatorio más que por algún patrón de fondo.

Estas dos últimas queries no han sido muy esclarecedoras pero nos han dado alguna pista respecto a la regularidad de los viajes a lo largo del año en las distintas zonas.

Después de todo el análisis realizado podemos afirmar que sí, *la diferencia en el nivel de vida en cada uno de los 5 barrios que tiene la ciudad de Nueva York sí que se refleja en los viajes realizados por los ciudadanos*. Quizás si trabajásemos con el dataset completo en vez de esta reducida parte que hemos tratado obtendríamos otros resultados, pero eso requeriría de una infraestructura algo mayor y bastante más paciencia.