Licenciatura en Ciencias de DATOS

Trabajo práctico 2

Tópicos de Analítica de Datos con SQL Avanzado

Integrante	LU	Correo electrónico
Dinkel Ayelén	621/15	medina_ayelen@hotmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (+54 011) 4576-3300
http://www.exactas.uba.ar

Índice

1. Introducción

En este trabajo prático se realizarán tareas de exploración y análisis de un dataset real con datos espaciales. Dichas tareas se llevarán a cabo utilizando tecnicas de SQL vistas en las clases y se ejecutaron en el programa de PgAdmin4. Se tiene de datos la red de colectivos del Municipio de General Pueyrredón, que utiliza la solución tecnológica SUBE, dichos datos se encuentran en dos tablas: - "Datos_Eco_Gral_Puey", la cual tiene alrededor de 67 mil registros, que resumen la cantidad de viajes en colectivo realizados en cada línea municipal de General Pueyrredón en un día de 2019, abiertos por hora, tipo de tarifa y coordenadas de referencia. - "Puntos_Interes_Gral_Puey", la cual contiene un conjunto ad-hoc de 12 puntos geográficos de interés para el análisis. Con sus coordenadas de referencia. También se conoce la descripción de los datos, con la cual se va a verificar que los datos que se tienen coincidan con la descripción correspondiente.

2. Validación y perfilado de los datos

Las consultas para este análisis se encuentran en el script "validacionYperfilado.sql". A continuación se explican dichas consultas y las conclusiones que se sacaron de las mismas.

- Para un primer acercamiento con las tablas, se realizó una vista de las primeras diez filas de la tabla "Datos_Eco_Gral_Puey" y de la tabla entera "Puntos_Interes_Gral_Puey". Así observamos la etructura y el formato de los datos. Las figuras 1 y 2 que muestran las salidas de dichas consultas, esta ultima tabla la acompañamos con la imagen geografica de sus datos en la figura 3

	dia date	hora integer	empresa character varying (100)	linea character varying (100)	lat_lon jsonb	tipo_tarifa character varying (10)	cant_trx integer
1	2019-05-15	0	EL LIBERTADOR SRL	BSAS_LINEA_562	{"lat": -38.05, "lon": -57.54}	PLENA	2
2	2019-05-15	0	EL LIBERTADOR SRL	BSAS_LINEA_562	{"lat": -38.046, "lon": -57.542}	BONIFICADA	1
3	2019-05-15	0	EL LIBERTADOR SRL	BSAS_LINEA_562	{"lat": -38.046, "lon": -57.542}	PLENA	1
4	2019-05-15	0	EL LIBERTADOR SRL	BSAS_LINEA_562	{"lat": -38.042, "lon": -57.548}	BONIFICADA	1
5	2019-05-15	0	EL LIBERTADOR SRL	BSAS_LINEA_562	{"lat": -38.042, "lon": -57.548}	PLENA	2
6	2019-05-15	0	EL LIBERTADOR SRL	BSAS_LINEA_562	{"lat": -38.038, "lon": -57.55}	BONIFICADA	1
7	2019-05-15	0	EL LIBERTADOR SRL	BSAS_LINEA_562	{"lat": -38.032, "lon": -57.554}	BONIFICADA	3
8	2019-05-15	0	EL LIBERTADOR SRL	BSAS_LINEA_562	{"lat": -38.032, "lon": -57.554}	PLENA	2
9	2019-05-15	0	EL LIBERTADOR SRL	BSAS_LINEA_562	{"lat": -38.03, "lon": -57.556}	PLENA	1
10	2019-05-15	0	EL LIBERTADOR SRL	BSAS_LINEA_562	{"lat": -38.022, "lon": -57.566}	PLENA	1

Figura 1: Vista de los primeros 10 datos de la tabla "Datos_Eco_Gral_Puey".

	lugar character varying (100)	geom geometry
1	TERMINAL DE OMNIBUS	0101000020E6100000577D9F3F11C84CC0D8D106C48EFE42...
2	CASINO CENTRAL	0101000020E610000021E7E1495FC54CC043FF00667A0043C0
3	FARO PUNTA MOGOTES	0101000020E61000009DAF2B82C7C54CC02B18679BBB0B43...
4	COMPLEJO CHAPADMALAL	0101000020E6100000003959619B01D84CC02D5EBCD20D1A43...
5	PUERTO MAR DEL PLATA	0101000020E6100000006755B85444C54CC0F821844A4B0643C0
6	CENTRO CULTURAL VICTORIA OCAMPO	0101000020E6100000AC56F7F3C4C64CC0BFDCC77D8A0243...
7	LAGUNA DE LOS PADRES	0101000020E610000032BA8DAD30DE4CC0C467FBC576F842...
8	SIERRA DE LOS PADRES	0101000020E6100000507170E483E34CC04C220AE65DF942C0
9	BATAN (CENTRO)	0101000020E61000002A0D251895DA4CC0B3842E54FF0043...
10	FCEYN - UNMDP	0101000020E61000003CB2AE1720C94CC081740E57C20043...
11	MUSEO PROV. ARTE CONTEMPORANEO	0101000020E6100000C307BC3688C54CC01366B5BA93FC42...
12	ESTADIO JOSE MARIA MINELLA	0101000020E6100000A75E8B7382CA4CC0AC6F39E04E0243...

Figura 2: Vista de los datos de la tabla "Puntos_Interes_Gral_Puey".

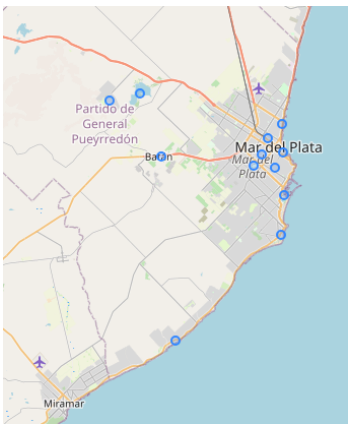

Figura 3: Vista geográfica de los 12 puntos de interés.

Tabla "Puntos_Interes_Gral_Puey":

- Con la informacion de la figura 2 es facil ver que la cantidad de datos son 12, igualmente se genero una consulta con **COUNT** para conocer la cantidad de datos en dicha tabla.
- Tambien observando en la imagen, los nombres de los lugares son todos distintos pero con el punto geometrico se verificó con una consulta contando los valores distintos en la columna geom, se utilizó **COUNT(DISTINCT geom))**. Efectivamente tiene valores distintos.

Tabla "Datos_Eco_Gral_Puey":

- Se realizó una consulta para contar la cantidad de datos, utilizando **COUNT** y el resultado es consistente con la infomracion dada en la descripción de los datos.
- Se verifió que efectivamente los datos provengan de un solo dia utlizando los comandos **COUNT - DISTINCT**.
- Observamos que no hay valores en la columna hora que esten fuera del rango [0,23], se utilizaron restriciones en el comando **WHERE**.
- La columna "tipo_tarifa"posee solamente dos valores posibles como especifica su descripción.
- Aseguramos que la columna çant_trx"tenga unicamente valores mayores a cero como explica la descripcion.
- Otras consultas realizadas sobre esta tabla fueron conocer la cantidad minima y maxima por hora, empresa y linea. Tambien se separó unicamente por hora y empresa sin tener en cuenta cada una de las lineas.

Existencia de nulos o blancos:

Para ambas tablas por separado se consultó sobra la existencia de nulos o vacios con dos consultas utilizando **IS NULL - TRIM() = "**. Para la tabla de los puntos de interes, al ser una tabla chica y por la imagen 2 se puede observar que no posee valores nulos. Y para la tabla de los datos, se realizaron dichas consultas por cada columna. En ninguna de las tablas se observaron datos nulos o faltantes.

3. Consultas

En la siguiente sección se desarrollaran los resultados de las consultas pedidas en los siguientes puntos del trabajo práctico, dichas consultas se encuentran en el script 'consultas.sql'.

3.1. Ejercicio 2

Este ejercicio pedía crear una vista con los datos de la tabla "Datos_Eco_Gral_Puey" sumando la ubicacion en formato geometry, los datos de los puntos de interes de la tabla "Puntos_Interes_Gral_Puey" y los datos del geohash de longitud 6 al que pertenece cada punto. Dicha vista se genero con **CREATE VIEW** y la llamé "datos_geom".

En esta tabla si encontramos nulos,ya que para este item se considera que una ubicacion esta asociada a un punto de interes si se encuentra a menos de 500 metros de la misma, de esta forma no todos los puntos van a ser completados con nombres de lasegunda tabla.

3.2. Ejericio 3

En este ejercicio se pedia mostrar el resultado en el Geometry Viewer de la ubicacion de todos los puntos que corresponden a ubicaciones de las transacciones económicas realizadas en General Pueyrredón. En el resultado de la consulta se excluyeron los puntos inconsistentes, ie. los puntos con coordenadas (0,0). La salida de la consulta se puede observar en la figura 4.

3.3. Ejercicio 4

En este ejercicio se pide incorporar tres puntos a la tabla de los puntos de interes, dichos puntos se pueden elegir libremente mientras cumplan que esten, al menos, a 1200 metros entre sí y del resto de los puntos de interes.

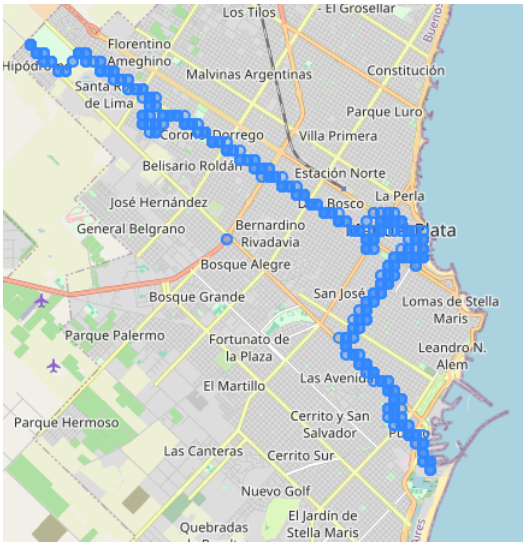

Figura 4: Ubicaciones de las transacciones dentro de Gral. Pueyrredón

Para resolver esta consigna lo que hice fue hacer unas consultas previas observando los puntos que estaban a menos de 1200 metros y comparando con el total de puntos, de esa forma observaba los que no estaban en la interseccion de ambos conjuntos y elegia posibles puntos de interes.

Luego de tener elegidas las coordenadas, compare entre ellas si estaban a mas de 1200 metros de distancia entre si, realizando con el **UNION ALL** y calculando la distancia mediante **ST_DISTANCE**. Las coordenadas elegidas son las que figuran en la tabla 3.3 y sus respectivas distancias las anotadas en la tabla 3.3

Puntos seleccionados de interés			
Coordenada	Punto 1	Punto 2	Punto 3
Latitud	-38.172	-37.926	-37.886
Longitud	-57.64	-57.544	-57.834

Distancias entre los puntos seleccionados	
Puntos	Distancia
1 y 2	28575.951
2 y 3	25887.441
1 y 3	36025.941

Luego verifico que los 15 puntos de interes ninguno esté a menos de 1200 metros entre ellos, utilizando la clausula **(WHERE distancia < 1200)** y obteniendo como resultado que ninguna lo está.

Luego se le agregó el nombre del lugar al que hacen referencia los 3 puntos agregados, para esto se buscó información extra de lo que hay en dichas coordenadas. Los lugares agregados son 'PLAYA PUBLICA', 'ESCUELA MUNICIPAL' y 'ESCUELA DE VUELO' respectivamente en el orden que aparecen sus coordenadas en la tabla 3.3. Por último, en la figura 5 se pueden observar los 15 puntos que se tienen ahora en la tabla "Puntos_interes_gral_puey", donde se marcaron con un cuadrado los puntos agregados **'ESCUELA DE VUELO'** , 'PLAYA PUBLICA' y 'ESCUELA MUNICIPAL' con sus respectivos colores.

Figura 5: Vista de los 15 puntos de interés.

3.4. Ejercicio 5

En este ejercicio se queria conocer las horas diurnas de mayor demanda y las 3 horas diurnas de menor demanda, considerando el periodo entre las 6 y las 20hs en todo el municipio y luego para las zonas dentro de un radio de 500m de distancia de unos puntos seleccionados.

Para dichas consultas se sumó (**SUM**) la cantida de demanda en el periodo de tiempo seleccionado utilizando **WHERE - BETWEEN** agrupando por hora y ordenando por la cantidad de demanda, **GROUP BY - ORDER BY**. Para el item con las zonas en especifico se utlizo otro **WHERE IN** seleccinando que la columna 'lugar' se encuentre entre la lista que era de interés. Con dichos datos se realizo un ranking (**RANK**) y seleccioné las dos horas con mayor demanda y luego las 3 con menor como se pedia.

En la figura 6 se observan los resultados en un gráfico de barras de la consulta sobre todo el municipio y en la figura 7 los resultados de las zonas seleccionadas. Los gráficos se realizaron con la opción de **Graph Visualiser** en el PgAdmin.

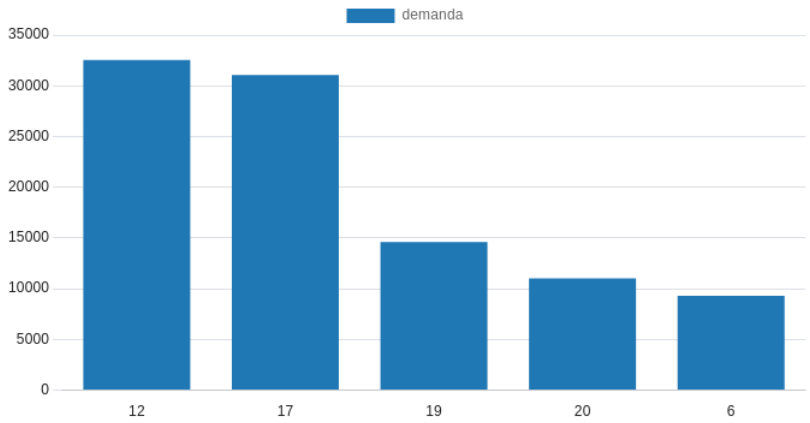

Figura 6: Horas de mayor y menor demanda teniendo en cuenta todo el municipio.

Con estos gráficos se puede observar que la hora de mayor demanda en las zonas seleccionadas es la segunda hora de mayor demanda en todo el municipio y luego la hora que comparten es la de menor demanda. Para poder realizar un mejor análisis y comparación entre las distintas consultas generé el gráfico de curvas de la evolución por horas. En la figura 8 se encuentra la evolución teniendo en cuenta todo el municipio y en la figura 9 en las zonas específicas.

En estas curvas se puede apreciar que tienen un comportamiento similar, a mayor escala cuando se trata de

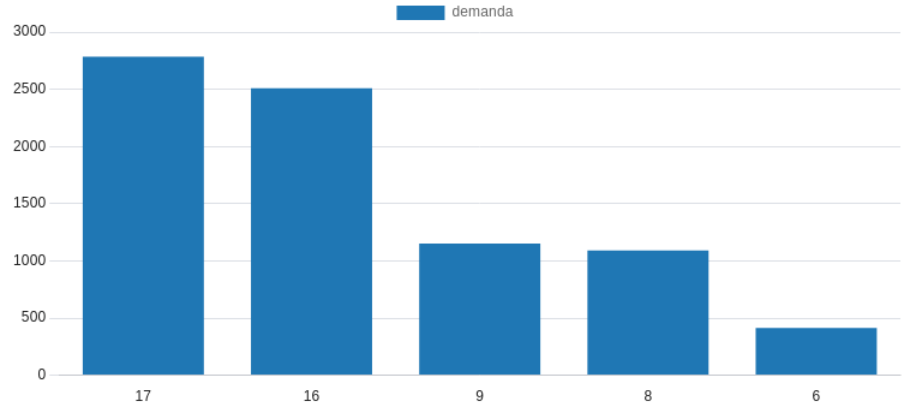

Figura 7: Horas de mayor y menor demanda teniendo en cuenta zonas espécificas.

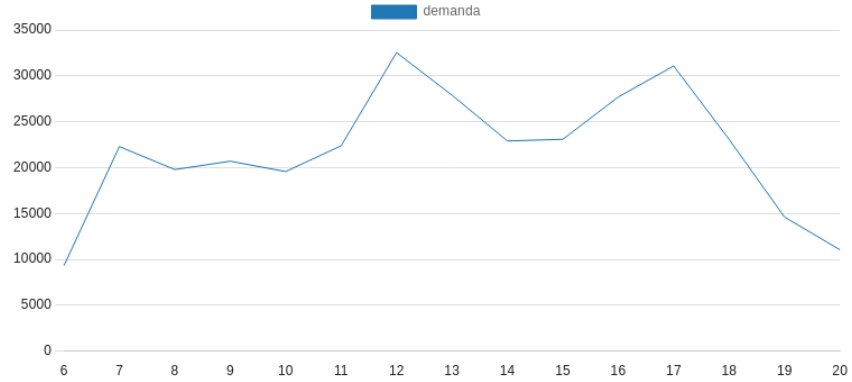

Figura 8: Curva de demanda por hora en todo el municipio.

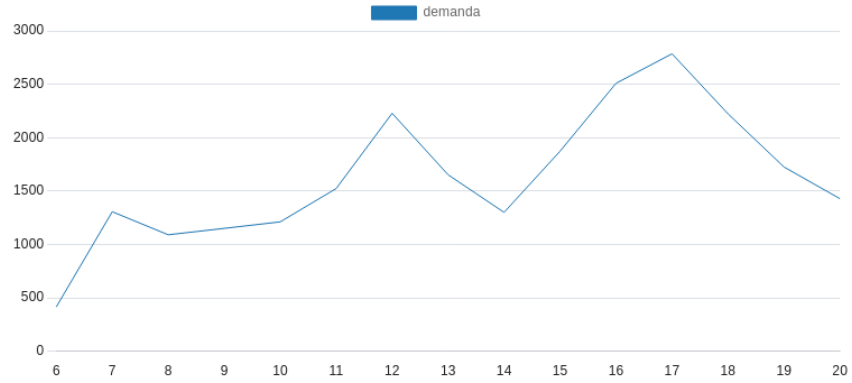

Figura 9: Curva de demanda por hora en las zonas espécificas.

todo el municipio y observando que el pico de mayor demanda a las 12hs, se ve representado igualmente en la curva de las zonas aunque no sea de las dos horas de mayor demanda. Se puede concluir que la demanda esta altamente realacionada con la franja horaria, las zonas tienen su peso generando algunos swap entre horas cercanas en el ranking.

3.4.1. Otras consultas

En esta sección dentro del ejercicio 5, se pide realizar una serie de consultas considerando las zonas determinadas por geohashes de longitud 6.

- Se quiere conocer las 5 zonas de mayor cantidad de transacciones económicas registradas en el día, para ello fue necesario sumar la cantidad de transacciones (columna 'cant_trx') agrupando por zona y luego ordenar por esas cantidades. Nos quedamos con las primeras 5 utilizando **FETCH FIRST 5 ROWS ONLY** y el Geometry Viewer que se obtiene es el que se encuentra en la figura 10, donde se puede observar que las zonas con más transacciones se encuentran en el centro de Mar del Plata que es una de las ciudades más grandes y pobladas del partido que estamos analizando.

Figura 10: 5 zonas con mayor cant de transacciones.

- Ahora se quiere conocer las 5 zonas de mayor cantidad de transacciones pero discriminadas según el tipo de tarifa. La consulta se realizó similar a la anterior, solo que en este caso se necesitó realizar un ranking haciendo una particion sobre el tipo de tarifa y ordenando según las cantidades de cada tipo, luego se eligieron las 5 mejores de cada una. La salida de esta consulta está representada en la tabla 3.4.1,el tipo de tarifa y sus cantidades de transacciones, donde se observa que la cantidad de tarifas plenas es siempre mayor que las tarifas bonificadas. En la figura 11 se observa en el mapa las zonas que poseen esas cantidades.

Orden de cantidad de transacciones según tipo de tarifa					
Tipo de tarifa | Orden	1°	2°	3°	4°	5°
Bonificada	11905	5256	5201	4659	4638
Plena	12614	6079	5834	5682	5417

- En este caso se quiere conocer las 5 zonas de mayor cantidad de transacciones considerando por separado las horas 7, 12 y 17. La consulta se realizó similar a la anterior, quedandonós con las horas especificas para el analisis y generando el ranking con una particion sobre la hora y ordenando según las cantidades, luego se eligieron las 5 mejores de cada hora. La salida de esta consulta está representada en la tabla 3.4.1 y en la figura 12. En la tabla se muestran los resultados numericos de las horas y suss respectivas cantidades de transacciones, donde se observa que la hora con más transacciones son las 17hs, luego las 12hs y por ultimo las 7hs.

Orden de cantidad de transacciones según las horas seleccionadas					
Hora | Orden	1°	2°	3°	4°	5°
7hs	754	635	498	475	473
12hs	1453	1113	898	886	860
17hs	2847	1310	1223	970	926

- Se quiere conocer las 5 zonas en las que pasan la mayor cantidad de lineas diferentes y saber la cantidad de lineas que pasan por dicho lugar. La consulta se realizo con un **COUNT DISTINCT** en la cantidad de lineas y se ordenó de forma decreciente en esas cantidad de lineas. En la figura 13 se encuentran señaladas en el mapa estas 5 zonas con mayor cantidad de lineas, teniendo 25 lineas la que se encuentra más al norte, las del centro

Figura 11: Zonas con mayor cantidad de transacciones segun tipo de tarifas

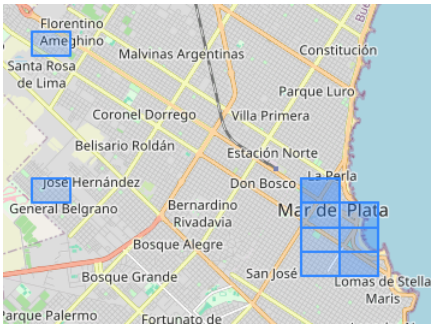

Figura 12: Zonas con mayor cantidad de transacciones diferenciado por horas especificas.

24 lineas y las de abajo 23 lineas cada una. No hay mucha diferencia de cantidad de lineas, lo que tiene sentido al estar pegadas las zonas y seguramente comparten la mayoria de esas lineas. Tambien se observa que son las zonas representadas en la figura 10, las cuales son las que tenian mayor cantidad de transacciones.

- En esta consulta se queria conocer las zonas por las que pasa una sola linea, para conocer estas zonas se utilizó **WHERE** para quedarnos solamente con las que tenian el resultado de cantidad de lineas en 1. El resultado geografico se puede observar en la figura 14, donde vemos que las zonas estan a las afueras del partido.

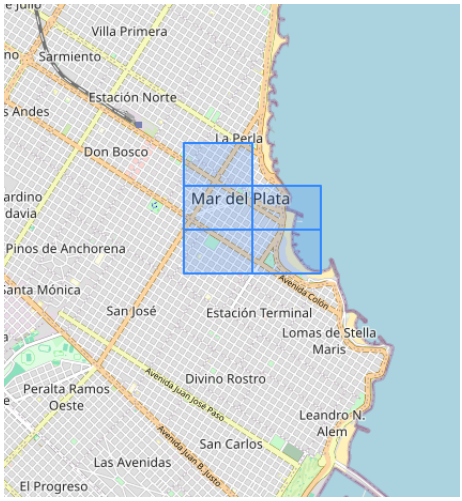

Figura 13: 5 zonas con mayor cantidad de lineas

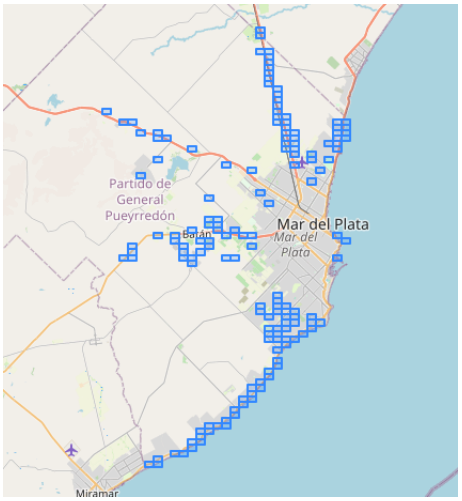

Figura 14: Zonas con una sola linea.

- Analizando en conjunto las respestas obtenidas en las consultas anteriores, se puede observar que la zona centrica del partido donde se encuentra la mayor cantidad de transacciones, como muestra la figura 13, es tambien la zona donde se realizan la mayor cantidad de transacciones diferenciadas por el tipo de tarifa y por el tipo de hora. En la figura 11 se puede ver que hay interseccion entre las zonas que tienen mayor cantidad de tarifas bonificadas y plenas, se puede concluir que no tiene que ver la zona con el tipo de tarifa y si con la cantidad de transacciones que se realizan en dicha zona. Algo similar ocurre con la cantidad de transacciones diferenciadas por hora que se muestra en la figura 12 donde las del centro son las que tienen la mayora cantidad en las horas 17,12 y 7. Las zonas que se encuentran mas alejadas en dicho mapa, hacen referencia a la cantidad de transacciones a las 7hs. Todo este analisis de cantidad de transacciones concuerda con las zonas que poseen la mayor cantidad de lineas, las cuales se encuentran en el centro y las que poseen solo una linea en los alrededores.

Por ultimo se quiere realizar unas consultas referidas a la cantidad de lineas:

- Se quiere conocer los geohashes de longitud 6 en los cuales se observó a la línea de mayor cantidad de pasajeros transportados, y cuáles a la línea de menor cantidad e identificar para cada una de ellas el geohash-hora de mayor cantidad de ascensos.

  Para llevar a cabo dicha consulta, se seleccionó la linea con mayor cantidad de pasajeros (**SUM**) agrupando por linea y ordenando por dicha cantidad de forma decreciente. Del mismo modo ordenando de forma ascendente la cantidad de pasajeros se tiene en la primera fila la linea con menor cantidad de pasajeros y con esa informacion

se agrupa segun las zonas donde pasan esas lineas. El resultado de dicha consulta se puede observar en la figura 15 las zonas por las que pasa la linea con mayor cantidad de pasajeros y la de menor cantidad de pasajeros. Como en el la figura no se puede diferenciar las zonas, se muentra en las figuras 16 , 17 y 18 tres ejemplos de que linea pasa en cada lugar. Observando mas puntos en el Geometry Viewer del pgAdmin, se puede reconocer que la linea con menor cantidad de pasajeros ("BS_AS_LINEA 720M") recorre solamente la ruta de Batán hacia el centro de Mar del Plata y la linea cno mayor cantidad de pasajeros ("LINEA_511") recorre las zonas de alrededor, como se observa en las dos figuras 17 y 18 de una punta a la otra pasando por el centro.

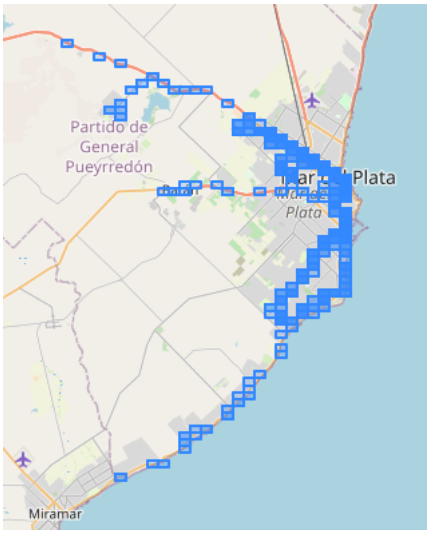

Figura 15: Zonas por la que pasa la linea con mayor y linea con menor cant de pasajeros.

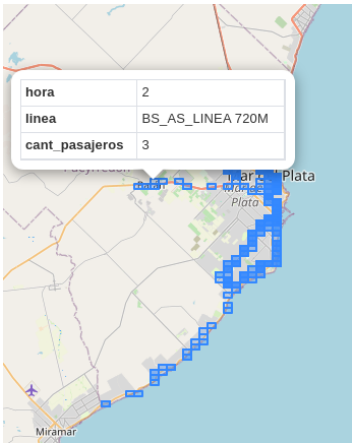

Figura 16: Zona que recorre la linea con menor cant de pasajeros.

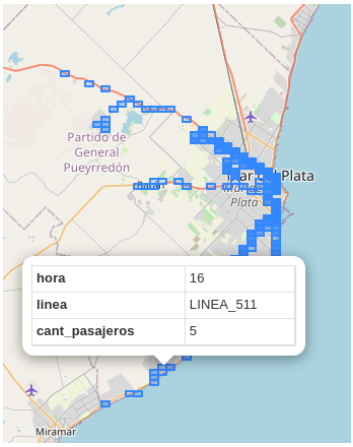

Figura 17: Zona que recorre la linea con mayor cant de pasajeros.

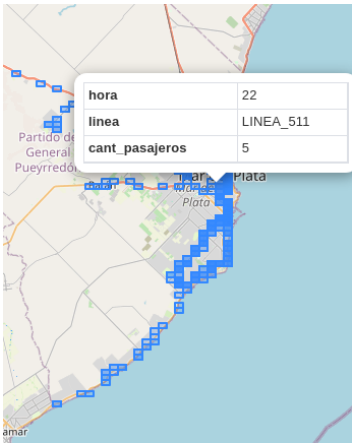

Figura 18: Zona que recorre la linea con mayor cant de pasajeros.

- Se queria conocer la linea que pasa por la mayor cantidad de los puntos de interés, conociendo tambien cuáles son los puntos de interés visitados. Sin tener en cuenta los puntos agregados en el ejercicio 4. Entonces para esta consulta se seleccionaron las lineas y sus respectivos lugares que no sean del conjunto agregrado en el inciso 4 utilizando (**WHERE lugar NOT IN**), agrupando por linea y ordenando por la cantidad de lugares distintos (**COUNT(DISTINCT lugar**). Con dicha subconsulta se obtiene la linea y luego se busca en la base de datos los lugares por los que pasa la linea asi se obtienen los nombres, la salida de la consulta se puede observa en la figura 19

–cantidad de lineas por empresa select distinct(empresa)0.75–cantidad de lineas por empresa select distinct(empresa)

Figura 19: Lugares por los que pasa la linea.

4. Conclusiones

Se hizo un análisis sobre las zonas del partido General Pueyrredón y se observaron que las zonas donde mas transacciones se realizan están en el centro, se observa que en ese centro se encuentra la ciudad de Mar del Plata, lo cual tiene sentido para ese partido de la costa. También en dichas zonas se encuentran las lineas con mayor cantidad de pasajeros y hay mas cantidad de lineas que en los alrededores de la ciudad. Se analizaron también las demandas por hora en el municipio y en algunas zonas determinadas, al compararlas se encuentran fuertes relaciones que tienen que ver mas con las horas que por la zona en las que se encuentra esa demanda.

5. Cambios realizados

- Se realizaron cambios en el informe, se realizo casi por completo agregando las descripciones de las consultas realizadas y mostrando los resultados de las mismas con tablas y gráficos. Agregando tambien introducción y conclusiones.
- Se cambiaron los **COUNT** por **SUM** en los casos que estaba mal utilizado el primer comando.
- Tambien se cambio el comando para realizar los ranking, que no fue algo mencionado en esta devolución pero si en la anterior.
- Se cambio la consulta del ejercicio 3, intentando ahora si resolver lo que se pedía. Al igual la consulta del punto 5b(ii), obteniendo los 5 mejores de cada tarifa.
- Se completaron los últimos dos items del ejercicio 5.