Licenciatura en Ciencias de DATOS

Trabajo práctico 1

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 que contiene información de viajes del transporte público de pasajeros de Argentina. 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 dato que la tabla tiene alrededor de 1,2 M de registros, lo cuales representan la cantidad de viajes realizados diariamente en cada línea de transporte SUBE en los años 2020, 2021 y 2022. 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.

2.1. Validación inicial de cumplimiento de las especificaciones:

- Para un primer acercamiento con los datos, se realizó una vista de las primeras diez filas. Así observamos la etructura y el formato de los datos. Las figuras 1 y 2 muestran la salida de dicha consulta.

	dia date	nombre_empresa character varying (100)	linea character varying (100)	amba character	tipo_transporte character
1	2020-01-01	EMPRESA BATAN S.A.	BS_AS_LINEA 715M	NO	COLECTIVO
2	2020-01-01	COMPAÃIAÂ DE TRANSPORTE VECINAL S.A.	BS_AS_LINEA_326	SI	COLECTIVO
3	2020-01-01	EMPRESA DE TRANSPORTE PERALTA RAMOS SACI	BS_AS_LINEA_512	NO	COLECTIVO
4	2020-01-01	AUTOBUSES BUENOS AIRES S.R.L. â- TRANSPORTE LARRAZABAL C.I.S.A. â- UNION TRANSITORIA (...	BS_AS_LINEA_514	SI	COLECTIVO
5	2020-01-01	EL URBANO SRL	BS_AS_LINEA_522	SI	COLECTIVO
6	2020-01-01	EL URBANO SRL	BS_AS_LINEA_527	SI	COLECTIVO
7	2020-01-01	TRANSPORTES LINEA 123 S.A.C.I.	BS_ASLINEA_123	SI	COLECTIVO
8	2020-01-01	TRANPORTES VEINTIDOS DE SETIEMBRE S.A.C.	BSAS_LINEA_002	SI	COLECTIVO
9	2020-01-01	GENERAL TOMAS GUIDO S.A.C.I.F.	BSAS_LINEA_009	SI	COLECTIVO
10	2020-01-01	LINEA 10 S.A.	BSAS_LINEA_010	SI	COLECTIVO

Figura 1: Vista de los primeros 10 datos de la tabla (parte1)

tipo_jurisdiccion character	provincia character varying (50)	municipio character varying (50)	cant_viajes integer
MUNICIPAL	BUENOS AIRES	GENERAL PUEYRREDON	2154
PROVINCIAL	BUENOS AIRES	SN	1492
MUNICIPAL	BUENOS AIRES	GENERAL PUEYRREDON	1889
MUNICIPAL	BUENOS AIRES	ALMIRANTE BROWN	4669
MUNICIPAL	BUENOS AIRES	LANUS	187
MUNICIPAL	BUENOS AIRES	LANUS	543
NACIONAL	JN	SD	1927
NACIONAL	JN	SD	6408
NACIONAL	JN	SD	5879
NACIONAL	JN	SD	4531

Figura 2: (parte2)

- Los datos son de los años correctos dichos en la descripción y con un análisis general de la cantidad de datos por año se puede apreciar que se relaciona con los hechos de que en el 2020 con la pandemia se realizaron menos viajes. Estas consultas se ralizaron con un **COUNT**, primero contando todos los datos para ver que se correspondia con la cantidad total informada en la descripción de los mismos, y luego diferenciando por año para tener una noción más especifica, con esta última consulta tambien se observa si los datos corresponden solamente a años 2020, 2021 y 2022.
- Con una segunda consulta se observa que la cantidad total de días por año es la correcta.
- En las columnas **amba,tipo_transporte y tipo_jurisdiccion** que poseen una cantidad finita de valores posibles, se analizó que efectivamente tengan solamente valores válidos. Dichas consultas se llevaron a cabo con el comando de **GROUP BY** agrupando en cada caso por la columna que se queria verificar. Además se agregó en la consulta un **COUNT** para tener información sobre cuantos datos correspondian a cada uno de los valores posibles en cada caso.
- En las columnas **Provincia y Municipio** vemos que tienen las combinaciones validas segun su **Jurisdiccion**. Estas consultas se realizaron agrupando por jurisdiccion y por cada una de las columnas de interés.
- En esas columnas tambien analicé si habia dos formas distintas de escribir al mismo lugar y se puede observar que Neuquen aparece dos veces en **provincia** escrito diferente y en la columna **municipio** las que se repiten son Cañuelas y Presidencia Roque Saenz Peña. Estas observaciones se realizaron linea por linea del resultado obtenido al realizar un **SELECT DISTINCT** sobre la columna de interés.
- Luego vemos que en la columna de cantidad de viajes existen numeros negativos y ceros, lo cual segun la descripcion no podria pasar. Esta conclusion se puede obtener luego de realizar la consulta usando el comando **WHERE** y pidiendo que devuelva solo los campos con la columna cant_viajes ≤ 0.

2.2. Existencia de nulos o blancos:

Vemos que no hay existencia de NULL en los datos pero si de campos vacios en las columnas de **Provincia, municipio y tipo_jurisdiccion** y las tres tienen los campos vacios solo con el tipo de transporte **SUBTE**, lo cual tiene sentido ya que solo estan en Capital Federal. Dichas consultas se llevaron a cabo usando los comandos de **GROUP BY - WHERE**, con la primera agrupando segun las columnas de interes y el segundo para ver los campos vacios o viendo los nulls con la consulta **is null**. Y para terminar de completar el analisis de los datos faltantes en el tipo de transporte **SUBTE** se verificó tanto si tenia faltantes en todas las columnas mencionadas anteriormente como si todas los valores faltantes correspondian a dicho transporte.

2.3. Perfilado de los datos

Sumado al análisis inicial que se realizó de las cantidades de los datos en conjunto con la validación, se realizó un perfilado de los datos. A continuación se enumeran las consultas realizadas, los comandos utilizados y las conclusiones que se fueron sacando de cada resultado obtenido.

- Se realizó una consulta para conocer de cada linea, diferenciando por año, la cantidad de viajes totales en el año, la cantidad minima y maxima de viajes en un dia que realizó, un promedio de esos viajes, una moda y la cantidad real de dias que realizo viajes esa linea. En esta consulta se utilizaron los comandos de **MIN, MAX, AVG, MODE Y COUNT**.
- En la respuesta de la consulta anterior se aprecia que hay varias lineas donde la cantidad total de dias es mucho menor a la cantidad total de dias del año, con lo cual hay muchos datos faltantes o trabajaron muchos menos dias. Para tener una vista mas especifica, la consulta sigue diferenciando por linea y año pero seleccionando solo las que la cantidad de dias es menor a 183, aproximadamente la mitad de dias del año.
- Luego se analizó la cantidad de empresas y la cantidad de lineas que posee cada empresa, sumando otras dos consultas se buscó cual es la cantidad minima y maxima de lineas que posee una empresa, un numero promedio de lineas por empresa y una moda. En una consulta extra se averiguo el nombre de las empresas que poseen la cantidad minima y maxima de lineas.
- Realizando consultas sobre la cantidad de viajes, se diferencio por tipo de tranporte para obtener el que mas viajes realiza, con el resultado de la cosulta se puede ver que son los colectivos los que mas viajes realizan. Dichas conclusiones se pueden sacar facilmente ya que se utilizó **ORDER BY** ordenando por a cantidad sumada previamente y de forma descendente. Continuando con el analisis de cantidad de viajes y ordenando de la misma forma, se consulto sobre caad jurisdiccion y sobre cada provincia, esta ultima se realizó unicamente con el tipo de transporte **COLECTIVO** dado que es el que mayor viajes tiene y que se enuentra en todas las provincias. En todas estas consultas, al estar ordenadas de forma descendente se puede observar facilmente cual es el maximo y minimo en cantida de viajes en cada consulta.

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 3

En la primera parte de este ejercicio creamos una vista con **CREATE VIEW** que la llamé "VIAJES", donde en dicha tabla se encuentran los mismos datos de la tabla original, sumando dos columnas extras **ANIO - DIA_SEMANA**. La agregacion de estas columnas, aparte de ser un pedido del ejercicio, van a facilitar consultas posteriores en donde se quiere tener facilmente la informacion que nos brindan.

Para la siguiente consulta se pedia calcular la cantidad de viajes y la cantidad de líneas diferentes, para ello se utilizó el comando **CUBE** para realizar las combinaciones posibles entre los atributos (año, tipo de jurisdiccion, amba, tipo de transporte). Dicha consulta se realizo sobre la vista creada anteriormente llamada "Viajes", la cual ya posee la columna de 'anio'.

Por ultimo se pidio generar estadisticos univariados para la columna de cantidad de viajes, para dicho trabajo se utilizaron los comandos **AVG, MAX, MIN,STDDEV Y PERCENTILE CONT**, donde AVG: calcula un promedio, STDDEV el desvio estandar y PERCENTILE_CONT nos devuelve el percentil del porcentaje pedido realizado sobre la variable que se esta estudiando. Ademas se agregaron limites superiores e inferiores que representan el rango intercualtil. Todo este analisis se llevo acabo en todo el conjunto de datos (incluyendo los datos que no son validos, por ej se conoce que hay valores en la columna cantidad de viajes que son menores a cero y eso no podria suceder), y luego en dos subgrupos del mismo, siendo primero un subconjunto por año y en AMBA y por ultimo sumando tambien el tipo de jurisdiccion y tipo de transporte.

Los resultados obtenidos de los estadisticos en toda la base de datos fueron:

Estadisticos sobre el total de los datos									
Total viajes	Promedio	Min	Max	SD	Mediana	Q1	Q3	Lim inf	Lim sup
8650725536	7169.45	-43	603766	16945.23	1913	445	6900	-9237.5	16582.5

3.2. Ejercicio 5

En esta seccion se analizan 10 caracteristicas libres de interes del dataset obtenido en la consulta 3b. En primer lugar se creo una vista de la tabla que se obtiene en dicha consulta, la cual llamamos: **'Viajes_amba'** y sobre ella se realizaran las consultas.

Para un primer acercamiento con la tabla en cuestion, se puede observar la figura 3 que contiene las 10 primeras filas.

	anio numeric	tipo_jurisdiccion character	amba character	tipo_transporte character	viajes bigint	lineas bigint
1	2020	MUNICIPAL	NO	[null]	157637900	315
2	2020	NACIONAL	NO	[null]	5301446	11
3	2020	PROVINCIAL	NO	[null]	103902907	453
4	2020	[null]	NO	[null]	266842253	776
5	2020		SI	[null]	73928371	7
6	2020	MUNICIPAL	SI	[null]	281466898	115
7	2020	NACIONAL	SI	[null]	742471680	156
8	2020	PROVINCIAL	SI	[null]	427756221	129
9	2020	[null]	SI	[null]	1525623170	407
10	2020	[null]	[null]	[null]	1792465423	1167

Figura 3: Vista de las primeras 10 filas de la tabla 'Viajes_amba'

A partir de la tabla **'Viajes_amba'** se realizaron las siguientes consultas:

- Si en todas las jurisdicciones tienen todos los tipos de transporte. Se realizo la consulta con un **GROUP BY** en ambos campos y se obtuvo que el medio de transporte 'COLECTIVO' es el unico que se encuentra en cada tipo de jurisdiccion, el tipo 'LANCHA' solo en 'PROVINCIAL' y 'TREN' solo 'NACIONAL'.
- Conocemos de analisis anteriores que la cantidad de viajes fue subiendo a lo largo de los años pero no se analizaron diferenciando el medio de transporte, entonces realice la consulta agrupando anio y tipo_transporte y se obtuvo que la mayor cantidad de viajes es con el tipo de transporte 'COLECTIVO' y que el año 2020 que posee menos cantidad de viajes en general, igualmente sigue siendo mayor la cantidad de viajes que otro tipo de transporte en el año 2022. Un analisis pasa con el medio de transporte 'TREN' aunque los viajes en 'SUBTE' en el año 2022 si superan los de 'TREN' en 2021 y 2020 pero solo los de ese año.
- Luego se analizó la cantidad de viajes separando lo que es AMBA y lo que no, diferenciando por tipo de transporte. Los resultados se ordenaron por cantidad de viajes totales asi se puede observar que la mayor cantidad de viajes se realizan en colectivo y dentro de AMBA.
- Se observó la diferencia de la cantidad de viajes realizados por año y separando por jurisdiccion. Como era de esperar el año 2022 es el que en su mayoria tiene mas cantidad de viajes, mantiene los analisis por año que se hicieron anteriormente. Por otro lado se observa que en cada año el orden descendente de cantidad de viajes por jurisdiccion respeta ser NACIONAL-PROVINCIAL-MUNICIPAL.
- Se analizó cuantas son las lineas operan en cada jurisdiccion. Se ordenó la salida de forma descendente para poder conocer cual es la jurisdiccion con mayor cantidad de lineas, la cual resultpo ser la provincial con 2003 lineas.
- Continuando con el analisis sobre la cantidad de lineas, se decidio diferenciar entre las que se encuentran fuera y dentro de AMBA. Se obtuvo que en su mayoria se encuentran fuera de AMBA.
- Analizando la cantidad de viajes por jurisdiccion y diferenciando si esta en AMBA o no, se observa que en su mayoria se encuentra dentro de AMBA. El orden de NACIONAL-PROVINCIAL-MUNICIPAL en cantidad de viajes ya se habia observado en una consulta anterior.
- Relacionando las ultimas dos consultas y viendo que la mayor cantidad de lineas se encuentran fuera de AMBA pero la cantidad de viajes es mucho mayor dentro de ellas, realice una consulta donde las relacione a la cantidad de lineas con su total de viajes. Con el resultado se puede observar que realmente tienen esa relacion, aunque haya menos lineas dentro de amba son las que más viajes realizan.
- Se analizó la cantidad de lineas por año y se puede notar un aumento de lineas entre 2020 y 2021, de 1186 a 1351 y luego una disminucion pero mas pequeña en relacion al aumento anterior entre los años 2021 y 2022, de 1351 a 1345. Sumando a esta consulta se añadio otra similiar pero que nos diferencie por tipo de tranporte y el resultado es el esperado, que los cambios sobre la cantidad de lineas son sobre el tipo de transporte 'COLECTIVO'.

3.3. Ejercicio 6

* Para la primera consulta, donde se queria conocer de un año en especifico (2022) las lineas con mayor cantidad de viajes de cada jurisdicion y tipo de transporte se realizó una consulta que contiene dos partes, en la primera se seleccionaron las columnas de los datos pedidos a conocer junto con la suma de los viajes, la cantidad de dias y se genero un ranking de las lineas segun su cantidad de viajes. Para ello se utilizaron los comandos **SUM, COUNT Y RANK con PARTITION BY y ORDER BY**.

De esta forma se obtenia un orden descendente de las lineas segun su cantidad de viajes, por lo tanto en la segunda parte de la consulta se selecciona unicamente las lineas que estan primeras en cada orden que fue particionado por tipo de jurisdiccion, provincia, municipio y tipo de transporte.

En el resultado se puede apreciar que el ferrocarril Roca posee mayor cantidad de viajes, en segundo lugar el subte B y luego un colectivo municipal en el municipio de Moreno.

* Para la segunda consulta, donde se queria averiguar el mes de mayor variacion intermensual porcentual considerando los viajes totales mensuales de las lineas de colectivo en amba realice la busqueda en tres partes:

- En la primera se seleccionaron las variables de las que se necesitaba informacion, se separo los datos por linea, tipo de jurisdiccion con su municipio y provincia correspondiente, con ellos se busco el mes, año, la suma de la cantidad

de viajes y crearon las columnas con la informacion del mes anterior. Para esto ultimo se utilizo la función **LAG con PARTITION BY y ORDER BY**. Esta función devuelve el valor de un campo evaluado en el N-ésimo registro previo dentro de la partición, para obtener la cantidad de viajes del mes anterior.

- En la segunda parte, con toda la información ya obtenida sobre cada mes y su anterior se calculó la variacion intermensual porcentual y se genero un ranking segun esa variacion obtenida sin tener en cuenta los null. De esta forma se tiene en orden cuales fueron los meses con mayor variacion.

- En la ultima parte lo unico que se hizo fue quedarnos con las filas que tenian el primer puesto en el orden del ranking generado anteriormente. En esta llamada, para los datos que quedaron en null no se tuvieron en cuenta en el rank pero se observó que empezaba a nombrar igualmente desde el puesto 2, entonces es el primer valor en la columna orden que aparece para cada maximo.

4. Conclusiones

Luego de varias consultas realizadas a la base de datos de transporte se tiene un conocimiento más amplio sobre la informacion que nos brindan dichos registros. Analizando las conclusiones parciales que se fueron sacando en cada proceso se puede observar las relaciones entre unas y otras, al abordar los datos desde distintos angulos se encuentran relaciones fuertes entre las consultas. Como por ejemplo, desde un principio nontamos que la cantidad de viajes aumenta a lo largo de los tres años analizados, lo que por un lado concuerda con la realidad vivida en dichos años, como tambien se mantiene firme en las conclusiones parciales de otras consultas que involucran las variables de año y cantidad de viajes. Tambien se puede concluir que el tipo de transporte 'COLECTIVO' es el que mayor cantidad de viajes posee y si se discrimina por tipo de jurisdiccion, la que tiene el mayor volumen es el 'MUNICIPAL'.

Se notó una relación tal vez algo antintutiva con respecto a la cantidad de lineas en AMBA y fuera de ella con la cantidad de viajes realizados en la zona, en un principio se podria llegar a relacionar que si hay mas lineas tambien hay mas cantidad de viajes, pero ese no es el caso de la discriminacion en cuanto a estar dentro o fuera de AMBA. En este caso se podria investigar un poco sobre el origen de los datos, si esta relacionada la informacion de que la linea pertenece a AMBA y si realiza viajes dentro de ella solamente o si cruza los limites.

Para las distintas consignas se utilizaron varias herramientas vistas en las clases, las primeras consultas fueron mas simples y a lo largo del trabajo se fueron combinando varias de ellas para realizar consultas mas especificas, obteniendo asi una informacion más detallada de los datos.