Tópicos de Analítica de Datos con SQL Avanzado

4°bimestre 2023

Trabajo Práctico n°1

Alumno: Agustin de Otazua

Introducción

Queremos realizar tareas de exploración y análisis de un dataset real, utilizando como soporte las técnicas SQL vistas en la materia.

Para ello, trabajaremos con la tabla agregada "***Viajes_Transp***", construida en base a datos abiertos de transporte SUBE del Ministerio de Transporte de la Nación.

Esta tabla tiene alrededor de 1,2 M de registros, que presentan la cantidad de viajes en transporte realizados en cada línea de transporte SUBE en los años 2020, 2021 y 2022.

Ejercicio 1

Perfilado

1) Los tipos de datos presentes son:

dia	nombre_empresa	linea	amba	tipo_transporte
date	character varying (100)	character varying (100)	character	character

tipo_jurisdiccion	provincia	municipio	cant_viajes
character	character varying (50)	character varying (50)	integer

2) **Valores distintos**
   - **a.** Hay 9 columnas y 1.206.609 registros (filas)
   - **b.** Hay 338 empresas distintas.
   - **c.** Hay 1435 líneas distintas.
   - **d.** Hay 4 tipos de transporte (colectivo, tren, subte y lancha).
   - **e.** Hay 4 tipos de jurisdicciones.
   - **f.** Hay 22 provincias distintas y 85 municipios.
3) **Porcentajes de valores en algunas variables categóricas**
   - **a.** Un 36% de los registros son de AMBA y el restante 64% del resto del país.
   - **b.** Un 14% de las jurisdicciones son Nacionales, un 47% Provinciales y un 38% Municipales. El restante son valores en blanco.
4) **Valores únicos**
   - **a.** No hay municipios, provincias o empresas con valores únicos.
   - **b.** Hay 3 líneas con valores únicos (520 A, 576 A, 503_SJUAN).
5) Varios de los demás análisis (presencia de nulls o espacios en blanco, estadísticas y demás, están en los siguientes ejercicios)

Ejercicio 2

Me fijo cuántos NULLS o espacios en blanco hay para cada variable.

Realizo la siguiente consulta:

```
select    sum(case when dia is null then 1 else 0 end) dias_null,
                  sum(case when coalesce(nombre_empresa,'') = '' then 1 else 0 end) empresas_null,
                  sum(case when coalesce(linea,'') = '' then 1 else 0 end) lineas_null,
                  sum(case when coalesce(amba,'') = '' then 1 else 0 end) amba_null,
                  sum(case when coalesce(tipo_transporte,'') = '' then 1 else 0 end) transporte_null,
                  sum(case when coalesce(tipo_jurisdiccion,'') = '' then 1 else 0 end) jurisdiccion_null,
                  sum(case when coalesce(provincia,'') = '' then 1 else 0 end) provincia_null,
                  sum(case when coalesce(municipio,'') = '' then 1 else 0 end) municipio_null,
                  sum(case when cant_viajes is null then 1 else 0 end) viajes_null
from viajes_transp;
```

Obtengo los siguientes resultados:

dias_null	empresas_null	lineas_null	amba_null	transporte_null	jurisdiccion_null	provincia_null	municipio_null	viajes_null
0	0	0	0	0	8110	8110	8110	0

Solo en las variables tipo_jurisdiccion, provincia, municio y cant_viajes hay datos faltantes.

Ejercicio 3

A)

```
CREATE VIEW viajes_transp_expand AS
select *, TO_CHAR(dia, 'day') dia_semana, EXTRACT('YEAR' FROM dia) anio
from viajes_transp;
```

B)

```
select anio, tipo_transporte, tipo_jurisdiccion, amba, sum(cant_viajes) cant_viajes_totales,
count(distinct linea) cant_lineas
from viajes_transp_expand
group by cube(anio, tipo_transporte,  tipo_jurisdiccion, amba);
```

Uso cube para poder ver las agregaciones entre las distintas combinaciones de valores.

C)

Calculo algunos estadísticos de la variable cant_viajes.

i) Considerando la totalidad de los datos:

```
select sum(cant_viajes),
        avg(cant_viajes),
        STDDEV(cant_viajes),
        min(cant_viajes),
        max(cant_viajes),
        PERCENTILE_CONT(0.5) within group (order by cant_viajes) mediana,
        PERCENTILE_CONT(0.25) within group (order by cant_viajes) Q1,
        PERCENTILE_CONT(0.75) within group (order by cant_viajes) Q3,
        (4 * PERCENTILE_CONT(0.75) within group (order by cant_viajes)) -
            ( 3 * PERCENTILE_CONT(0.25) within group (order by cant_viajes)) Q3_mas_3IQR,
        (4 * PERCENTILE_CONT(0.25) within group (order by cant_viajes)) -
            ( 3 * PERCENTILE_CONT(0.75) within group (order by cant_viajes)) Q1_menos_3IQR
from viajes_transp v;
```

sum	avg	stddev	min	max	mediana	q1	q3	q3_mas_3iqr	q1_menos_3iqr
8650725536	7169	16945	-43	603766	1913	445	6900	26265	-18920

Aparece una cantidad de viajes negativa en el mínimo. No estoy seguro si esto tiene alguna interpretación o si fue un error. En las descripción de los datos, esta variable debería ser no negativa.

ii) Abierto conjuntamente por año y AMBA:

```
select anio,
        amba,
        sum(cant_viajes),
        avg(cant_viajes),
        STDDEV(cant_viajes),
        min(cant_viajes),
        max(cant_viajes),
        PERCENTILE_CONT(0.5) within group (order by cant_viajes) mediana,
        PERCENTILE_CONT(0.25) within group (order by cant_viajes) Q1,
        PERCENTILE_CONT(0.75) within group (order by cant_viajes) Q3,
        (4 * PERCENTILE_CONT(0.75) within group (order by cant_viajes)) -
                ( 3 * PERCENTILE_CONT(0.25) within group (order by cant_viajes)) Q3_mas_3IQR,
        (4 * PERCENTILE_CONT(0.25) within group (order by cant_viajes)) -
                ( 3 * PERCENTILE_CONT(0.75) within group (order by cant_viajes)) Q1_menos_3IQR
from viajes_transp_expand v
group by anio, amba;
```

anio	amba	sum	avg	stddev	min	max	mediana	q1	q3	q3_mas_3iqr	q1_menos_3iqr
2020	NO	266842253	1261	2102	0	40067	495	146	1518	5634	-3970
2020	SI	1525623170	10510	19548	-43	603766	5502	2013	12264	43017	-28740
2021	NO	488534194	1855	2471	0	28439	906	248	2559	9492	-6685
2021	SI	2344464335	15916	21770	-15	477857	10392	3985	20613	70497	-45899
2022	NO	703794890	2409	3185	0	31842	1204	339	3272	12071	-8460
2022	SI	3321466694	22552	31539	0	535299	14769	5769	28268	95765	-61728

En CABA durante 2020 y 2021 hubo registros con una cantidad negativa de viajes. En esos mismos casos también se observa una diferencia notable entre el promedio y la mediana, sugiriendo una fuerte asimetría en la distribución.

iii) Abierto conjuntamente por año, AMBA, tipo jurisdicción, tipo transporte:

```
select anio,
        amba,
        tipo_jurisdiccion,
        tipo_transporte,
        sum(cant_viajes),
        avg(cant_viajes),
        STDDEV(cant_viajes),
        min(cant_viajes),
        max(cant_viajes),
        PERCENTILE_CONT(0.5) within group (order by cant_viajes) mediana,
        PERCENTILE_CONT(0.25) within group (order by cant_viajes) Q1,
        PERCENTILE_CONT(0.75) within group (order by cant_viajes) Q3,
        (4 * PERCENTILE_CONT(0.75) within group (order by cant_viajes)) -
          ( 3 * PERCENTILE_CONT(0.25) within group (order by cant_viajes)) Q3_mas_3IQR,
        (4 * PERCENTILE_CONT(0.25) within group (order by cant_viajes)) -
          ( 3 * PERCENTILE_CONT(0.75) within group (order by cant_viajes)) Q1_menos_3IQR
from viajes_transp_expand
group by anio, amba, tipo_jurisdiccion, tipo_transporte;
```

No dejo los resultados porque son muchos.

Ejercicio 5

Algunas observaciones de la tabla obtenida en el ejercicio 3.B:

1. La cantidad de viajes totales aumenta bastante entre 2020 y 2022. Esto se debe a las restricciones de movilidad que aparecieron con la Pandemia que empezó en 2020. Para 2022 la situación epidemiológica ya estaba mucho más normalizada.
2. Los colectivos son el medio de transporte más usado, mientras que
3. Las lanchas son el medio de transporte menos usado.
4. Los viajes de colectivos municipales por AMBA duplican a los que no circulan por AMBA.
5. Los colectivos tienen muchas más líneas que los otros medios de transporte.
6. Solo hay trenes nacionales. Esto se debe a que cruzan más de una provincia.
7. No hay transporte en lancha en CABA.
8. Entre 2020 y 2021 el número total de líneas aumentó en 161, pero entre 2021 y 2022 solo se redujo en 2.

Ejercicio 6

A)

```
select tipo_jurisdiccion, provincia, municipio, tipo_transporte,
                    linea, cant_viajes_acum, cantidad_dias_actividad
from
        (select t.tipo_jurisdiccion, t.provincia, t.municipio, t.tipo_transporte, t.linea,
                        sum(t.cant_viajes) cant_viajes_acum,
                        count(distinct t.dia) as cantidad_dias_actividad,
                        rank() over (partition by t.tipo_jurisdiccion, t.provincia, t.municipio,
t.tipo_transporte
                                                order by sum(t.cant_viajes) desc) as rank
        from viajes_transp_expand t
        where anio = 2022
        group by tipo_jurisdiccion, provincia, municipio, tipo_transporte, linea
        )
where rank = 1
order by cant_viajes_acum desc
```

La subconsulta se arma simplemente porque no puedo llamar “**where rank = 1”** en ella.

B)

```
with viajes_por_mes as (
	select t.tipo_jurisdiccion, t.provincia, t.municipio, t.linea,
			EXTRACT('MONTH' FROM dia) mes_actual, sum(t.cant_viajes) viajes_mes_actual,
			lag(EXTRACT('MONTH' FROM dia)) over (partition by tipo_jurisdiccion, provincia, municipio, linea) mes_previo,
			lag(sum(t.cant_viajes)) over (partition by tipo_jurisdiccion, provincia, municipio, linea) viajes_mes_previo
	from viajes_transp_expand t
	where AMBA = 'SI' and tipo_transporte = 'COLECTIVO' and anio = 2022 --anio > 2020
	group by tipo_jurisdiccion, provincia, municipio, linea, EXTRACT('MONTH' FROM dia)
	)
select  linea, tipo_jurisdiccion, provincia, municipio,
		mes_previo, viajes_mes_previo,
		mes_actual, viajes_mes_actual,
		variacion_mensual_viajes
from (
	select linea, tipo_jurisdiccion, provincia, municipio,
			mes_previo, viajes_mes_previo,
			mes_actual, viajes_mes_actual,
			(viajes_mes_actual::numeric / coalesce(viajes_mes_previo, viajes_mes_actual)::numeric - 1) variacion_mensual_viajes,
			rank() over (partition by tipo_jurisdiccion, provincia, municipio, linea
		order by (viajes_mes_actual::numeric / coalesce(viajes_mes_previo, viajes_mes_actual)::numeric - 1) desc) as rank
	from viajes_por_mes
	)
where rank = 1;
```

El mes que más aparece como de mayor variación en el número de viajes es **marzo**. Probablemente sea porque coincide con el fin de las vacaciones de verano y el inicio del ciclo lectivo.

Conclusiones

Realizamos un análisis exploratorio del dataset: número de valores distintos por variable, estadísticos (percentiles, max, min, etc.) y algunas observaciones de la realidad presentada en los datos.

También pudimos aplicar técnicas analíticas de SQL como agregaciones por ventanas y una introducción a modelos multidimensionales.