Diego Andrés Gómez Polo
Daniel del Castillo Andrade
Isabela Sarmiento

Proyecto 2, Entrega 1

2a. Perfilamiento de Datos

Se trabajará con un conjunto de datos obtenido del proyecto Infraestructura Visible que contiene información sobre los peajes de Colombia y sus tarifas. Estos datos se encuentran en el archivo *peajes2019.xlsx*, un archivo que contiene 32 columnas y 173 filas, donde cada fila contiene la información detallada de un peaje.

De las 32 columnas, se tienen 20 variables categóricas y 12 numéricas. Hay un total de 789 celdas vacías (14.3% sobre los datos). A continuación, se detalla la información sobre estas variables:

- **cod_via**: representa el código vial en la cual se localiza el peaje. Es una variable categórica con alta cardinalidad, donde 42.3% (63) de los registros son únicos y 13.9% (24) son vacíos.

- **nombre**: representa el nombre del peaje. Es una variable categórica con alta cardinalidad, donde 93.6% (171) de los datos son únicos y no hay registros vacíos.

- **responsable**: representa la entidad territorial o gubernamental encargada de la administración del peaje. Es una variable categórica donde sólo hay 5 registros distintos (INVIAS, Departamento, Distrito, Municipio y Concesión ANI) y no hay registros vacíos. Sólo uno de los datos está bajo control municipal.

- **sector**: representa el sector o ruta donde está ubicado el peaje. Es una variable categórica que contiene, por ejemplo, las poblaciones principales que se encuentran en dicha ruta, donde 79.2% (137) de los registros son distintos y 61.8% (107) son únicos. No hay campos vacíos en esta columna.

- **admon**: representa la entidad encargada actualmente de la administración del peaje. Es una variable categórica con alta cardinalidad, donde 34.4% (55) de los registros son únicos y 7.5% (13) son vacíos, siendo INVIAS la entidad más frecuente.

- **ubicacion**: representa la ubicación del peaje. Es una variable categórica con alta cardinalidad, donde 96.3% (156) de los registros son únicos y 6.4% (11) son vacíos.

- **telefono**: representa el teléfono de contacto del peaje. Es una variable categórica con alta cardinalidad, donde 61.0% (83) de los registros son únicos y 21.4% (37) son vacíos.

- **telefono_g**: Contiene el o los teléfonos del gerente encargado del peaje en cuestión. Hay 50% (68) de valores distintos, presuntamente porque puede haber un gerente para varios peajes del mismo sector. Hay 21.4% (37) celdas vacías.

- **d_cat_i**: representa los vehículos que hacen parte de la categoría 1. Es una variable categórica, donde 1.3% (2) de los registros son únicos y 12.1% (21) son vacíos, siendo “Automóviles y camperos” la categoría más frecuente.

- **d_cat_ii**: representa los vehículos que hacen parte de la categoría 2. Es una variable categórica, donde 16.4% (25) de los registros son únicos y 12.1% (21) son vacíos, siendo “Camiones y Buses 2 ejes pequeños” la categoría más frecuente.

- **d_cat_iii**: representa los vehículos que hacen parte de la categoría 3. Es una variable categórica, donde 15.1% (23) de los registros son únicos y 12.1% (21) son vacíos, siendo “Camiones y Buses 2 ejes grandes” la categoría más frecuente.

- **d_cat_iv**: representa los vehículos que hacen parte de la categoría 4. Es una variable categórica, donde 13.9% (21) de los registros son únicos y 12.7% (22) son vacíos, siendo “3 y 4 ejes” la categoría más frecuente.

- **d_cat_v**: representa los vehículos que hacen parte de la categoría 5. Es una variable categórica, donde 11.3% (17) de los registros son únicos y 12.7% (22) son vacíos, siendo “5 ejes” la categoría más frecuente.

- **d_cat_vi**: representa los vehículos que hacen parte de la categoría 6. Es una variable categórica, donde 21.7% (15) de los registros son únicos y 60.1% (104) son vacíos, siendo “Categoría VI 6 ejes” la categoría más frecuente.

- **d_cat_vii**: representa los vehículos que hacen parte de la categoría 7. Es una variable categórica, donde 18.0% (11) de los registros son únicos y 64.7% (112) son vacíos, siendo “Categoría VII” la categoría más frecuente.

- **d_cat_viii**: representa los vehículos que hacen parte de la categoría 8. Es una variable categórica, donde únicamente hay un registro y el resto son vacíos.

- **d_cat_ix**: representa los vehículos que hacen parte de la categoría 9. Es una variable categórica, donde únicamente hay un registro y el resto son vacíos.

- **departamento**: representa el departamento donde se ubica el peaje. Es una variable categórica, donde 0.6% de los registros son únicos y no hay registros vacíos.

Sobre las variables numéricas se tiene la siguiente información:

- **cat_x**: representa la tarifa que deben pagar los vehículos pertenecientes a la categoría x definidos en la variable correspondiente. Se tienen en total 9 categorías, por lo que se tienen 9 variables correspondientes.

- **eje_adicio**: representa la tarifa en COP que deben pagar los vehículos que tengan un eje adicional.

- **eje_adic_2**: representa la tarifa en COP que deben pagar los vehículos que tengan dos ejes adicionales.

- **latitud**: representa la latitud en la cual se encuentra ubicado peaje y está dada en ° (grados decimales).

- **longitud**: representa la longitud en la cual se encuentra ubicado peaje y está dada en ° (grados decimales).

- **gcd_departamento**: indica el código departamental que le corresponde al peaje al estar ubicado en su respectivo departamento.

En anexos, Figura 6, se encuentra la tabla en la que se detallan los estadísticos respectivos de cada una de las variables numéricas, tales como máximo, mínimo, desviación estándar y el promedio.

A grandes rasgos, como se mencionó anteriormente, las variables **cod_via**, **nombre**, **sector**, **admon**, **ubicacion**, **telefono**, **telefono_g** tienen una alta cardinalidad. Las variables **cat_6** y **cat_7**, **cat_8** y **cat_9** tienen una alta correlación, esto sucede precisamente porque en los casos donde la categoría 6 es 0, la categoría 7, 8 y 9 también. Adicionalmente, es importante mencionar que las variables **d_cat_viii** y **d_cat_vix** tienen un 99.4% de ausencia sobre los registros, por lo que muy pocos peajes llegan a tener 8 y 9 categorías. Las demás categorías también tienen altos porcentajes de ausencia, siendo las más bajas las categorías 1, 2 y 3 con 21.1%. **teléfono** y **teléfono_g** también tienen un porcentaje considerable de ausencia del 21.4%. Por otro lado, las variables de latitud y longitud tienen registros únicos no vacíos y las variables que representan las tarifas de las categorías 6 y 7, **cat_6**, y **cat_7** respectivamente, tienen un alto porcentaje de registros en cero, 57.8% y 59.0%, respectivamente. Asimismo, las variables **eje_adicio** y **eje_adic_2** también tienen altos porcentajes de registros en cero, siendo estos 53.8% y 57.2%, respectivamente.

En mayor detalle, los registros de los peajes para las variables **d_cat_i** a **d_cat_vii** (descripción de las categorías 1 a 7) son cadenas de caracteres similares, pero no iguales. Por ejemplo, en la Categoría II, se observan cadenas tales como “Camiones y buses 2 ejes grandes”, “Categoría III Camiones y buses 2 ejes grandes”, “Camiones y Buses 2 Ejes Grandes” y “Camiones y Buses 2

ejes Grandes”, que representan el mismo concepto con distinta redacción. Esto implica que los valores únicos son por formato y no por la naturaleza de la característica.

Por otro lado, vale la pena contrastar los promedios de las diferentes categorías para observar cuales tienen mayor o menor valor. Esto se observa en la Figura 1. Se puede observar que la categoría con mayor promedio es la Categoría 5, que corresponde por lo general a camiones de 3 y 4 ejes. Esto coincide con la realidad, puesto que este tipo de vehículos tiene que pagar entre 40’000 y 50’000 COP por peaje. Nótese que los valores promedio de las categorías 8 y 9 se omiten. Esto, debido a la cantidad de registros vacíos presentes (únicamente hay un valor por cada categoría, 5’200 y 7’200 COP, respectivamente).

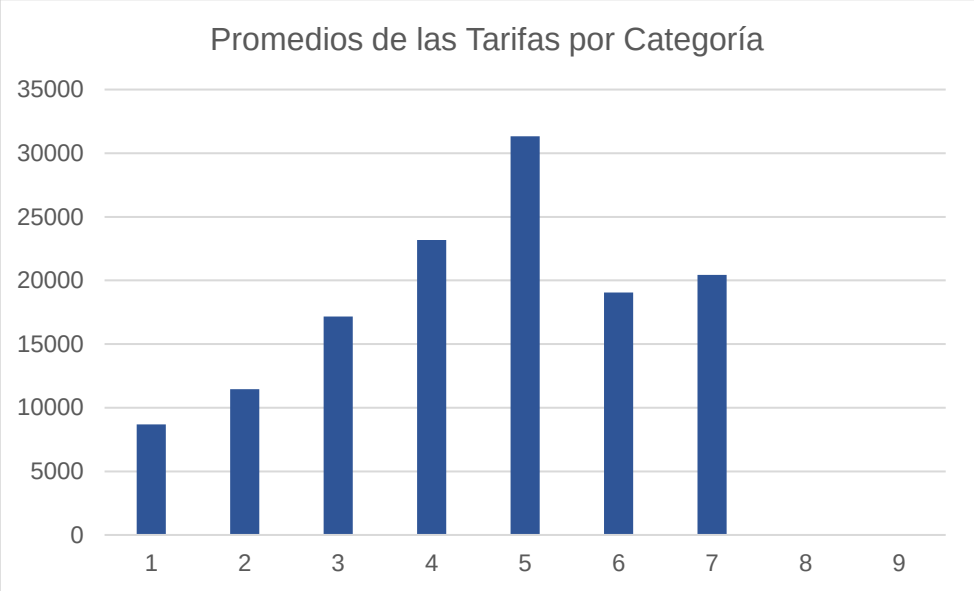


Figura 1: promedio de las tarifas en los peajes por categoría.

Finalmente, en la Figura 7 y Figura 8 se puede contrastar las fracciones de registros no vacíos por cada categoría y el *heatmap* de correlación de Pearson entre variables numéricas, respectivamente. Nótese que las variables **d_cat_viii** y **d_cat_ix**, que únicamente tienen un registro cada una, son las columnas más pequeñas.

2b. ETL

A grandes rasgos, el flujo de trabajo para la extracción, transformación y carga de los datos se evidencia resumido en la Figura 2. En mayor detalle, primero se usó Excel para extraer los datos en un formato válido para su posterior carga en los servicios de Google, e.g., la remoción del símbolo $ para características de carácter monetario. Seguidamente, los datos fueron cargador al sevidor Cloud de Google Drive para tenerlos disponibles al uso de los servicios de Google. Finalmente, se cargaron a BigQuery, donde se obtuvo el resultado final de las dimensiones deseadas (más detalle en la siguiente sección).

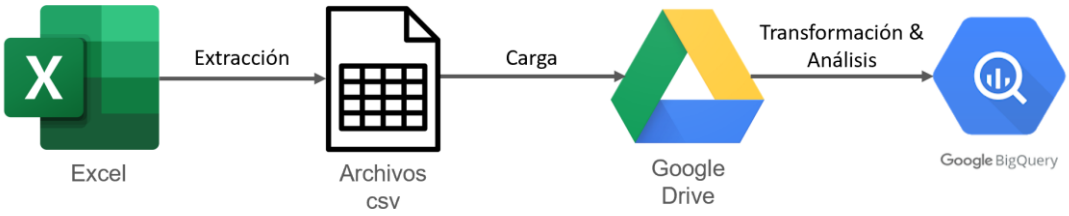


Figura 2: proceso ETL para el proyecto.

2c.

Con base en la evidencia estadística resultado del análisis de la sección 2a, se descartaron ciertos atributos para la dimensión de peajes, reteniendo únicamente: **nombre**, **responsable**, **cod_via**, **latitud**, **longitud**, **departamento**, **gcd_departamento**, **sector** y **admon**, como se observa en la Figura 3. Esto, debido a que los campos de teléfono, descripciones de categoría y tarifa por ejes adicionales no proveen información estadística útil al negocio (estos campos cumplen una función descriptiva), obteniendo así la mayor granularidad estadística útil posible sobre el hecho. Un ejemplo de las filas de esta nueva dimensión se puede observar en la Figura 4. Las características principales de la dimensión se observan en la Figura 5 (los resultados estadísticos correspondientes a los atributos se encuentran en la sección 2a).

Query editor

```
1 SELECT * FROM `bi202020.prueba2.Peaje` LIMIT 1000
```

Run | Save query | Save view | Schedule query | More

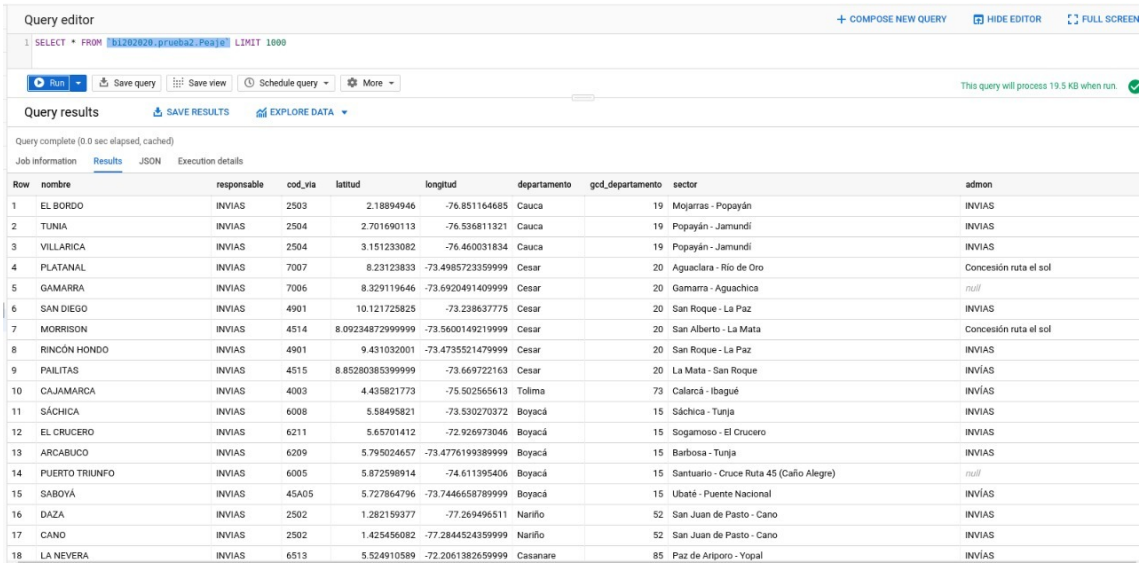
Peaje — QUERY TABLE

Schema | Details | Preview

Field name	Type	Mode	Policy tags	Description
nombre	STRING	NULLABLE		
responsable	STRING	NULLABLE		
cod_via	STRING	NULLABLE		
latitud	FLOAT	NULLABLE		
longitud	FLOAT	NULLABLE		
departamento	STRING	NULLABLE		
gcd_departamento	INTEGER	NULLABLE		
sector	STRING	NULLABLE		
admon	STRING	NULLABLE		

Edit schema

Figura 3: características de los atributos para la nueva dimensión de peaje.



Query editor

```
SELECT * FROM `bi202020.prueba2.Peaje` LIMIT 1000
```

Run | Save query | Save view | Schedule query | More — This query will process 19.5 KB when run.

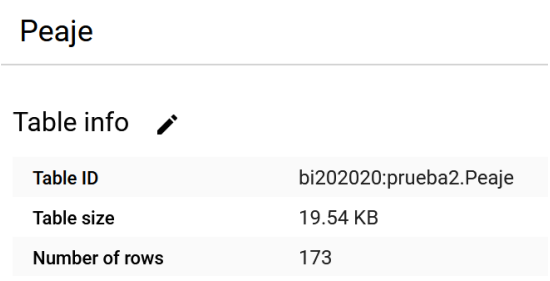
Query results — SAVE RESULTS | EXPLORE DATA

Query complete (0.0 sec elapsed, cached)

Job information | Results | JSON | Execution details

Row	nombre	responsable	cod_via	latitud	longitud	departamento	gcd_departamento	sector	admon
1	EL BORDO	INVIAS	2503	2.18894946	-76.851164685	Cauca	19	Mojarras - Popayán	INVIAS
2	TUNIA	INVIAS	2504	2.701690113	-76.536811321	Cauca	19	Popayán - Jamundí	INVIAS
3	VILLARICA	INVIAS	2504	3.151233082	-76.460031834	Cauca	19	Popayán - Jamundí	INVIAS
4	PLATANAL	INVIAS	7007	8.23123833	-73.4985723359999	Cesar	20	Aguaclara - Río de Oro	Concesión ruta el sol
5	GAMARRA	INVIAS	7006	8.329119646	-73.6920491409999	Cesar	20	Gamarra - Aguachica	null
6	SAN DIEGO	INVIAS	4901	10.121725825	-73.238637775	Cesar	20	San Roque - La Paz	INVIAS
7	MORRISON	INVIAS	4514	8.09234872999999	-73.5600149219999	Cesar	20	San Alberto - La Mata	Concesión ruta el sol
8	RINCÓN HONDO	INVIAS	4901	9.431032001	-73.4735521479999	Cesar	20	San Roque - La Paz	INVIAS
9	PAILITAS	INVIAS	4515	8.85280385399999	-73.669722163	Cesar	20	La Mata - San Roque	INVÍAS
10	CAJAMARCA	INVIAS	4003	4.435821773	-75.502565613	Tolima	73	Calarcá - Ibagué	INVÍAS
11	SÁCHICA	INVIAS	6008	5.58495821	-73.530270372	Boyacá	15	Sáchica - Tunja	INVIAS
12	EL CRUCERO	INVIAS	6211	5.65701412	-72.926973046	Boyacá	15	Sogamoso - El Crucero	INVIAS
13	ARCABUCO	INVIAS	6209	5.795024657	-73.4776199389999	Boyacá	15	Barbosa - Tunja	INVIAS
14	PUERTO TRIUNFO	INVIAS	6005	5.872598914	-74.611395406	Boyacá	15	Santuario - Cruce Ruta 45 (Caño Alegre)	null
15	SABOYÁ	INVIAS	45A05	5.727864796	-73.7446658789999	Boyacá	15	Ubaté - Puente Nacional	INVÍAS
16	DAZA	INVIAS	2502	1.282159377	-77.269496511	Nariño	52	San Juan de Pasto - Cano	INVIAS
17	CANO	INVIAS	2502	1.425456082	-77.2844524359999	Nariño	52	San Juan de Pasto - Cano	INVIAS
18	LA NEVERA	INVIAS	6513	5.524910589	-72.2061382659999	Casanare	85	Paz de Ariporo - Yopal	INVÍAS

Figura 4: ejemplo de las filas de la nueva dimensión peaje y estadísticas de la consulta.



Peaje

Table info

Table ID	bi202020:prueba2.Peaje
Table size	19.54 KB
Number of rows	173

Figura 5: estadísticas de la nueva dimensión.

Anexos

	cat_1	cat_2	cat_3	cat_4	cat_5	cat_6	cat_7	cat_8	cat_9	eje_adicio	eje_adic_2	latitud	longitud	gcd_departamento
count	173	173	173	173	173	173	173	173	173	173	173	173	173	173
mean	8690.751	11446.24	17168.79	23169.94	31322.54	19043.35	20425.43	30.0578	41.6185	4199.422	4278.035	6.427413	-74.7104	35.54913295
std	3013.232	5503.619	7896.162	10556.25	12244.77	24498.75	27202.51	395.3487	547.4059	4719.424	5546.179	2.541052	1.164883	24.57934853
min	0	0	0	0	0	0	0	0	0	0	0	1.087284	-77.4087	5
25%	8000	8900	12300	17600	27300	0	0	0	0	0	0	4.544131	-75.5874	17
50%	8600	10300	18700	24400	30500	0	0	0	0	0	0	5.727865	-74.8203	25
75%	10400	13900	21400	28500	37700	42400	44600	0	0	8700	8400	8.604856	-73.7752	54
max	16600	35300	39900	53100	68700	91800	101900	5200	7200	15900	20800	11.60651	-72.1567	85

Figura 6: resumen de estadísticos de las variables numéricas.

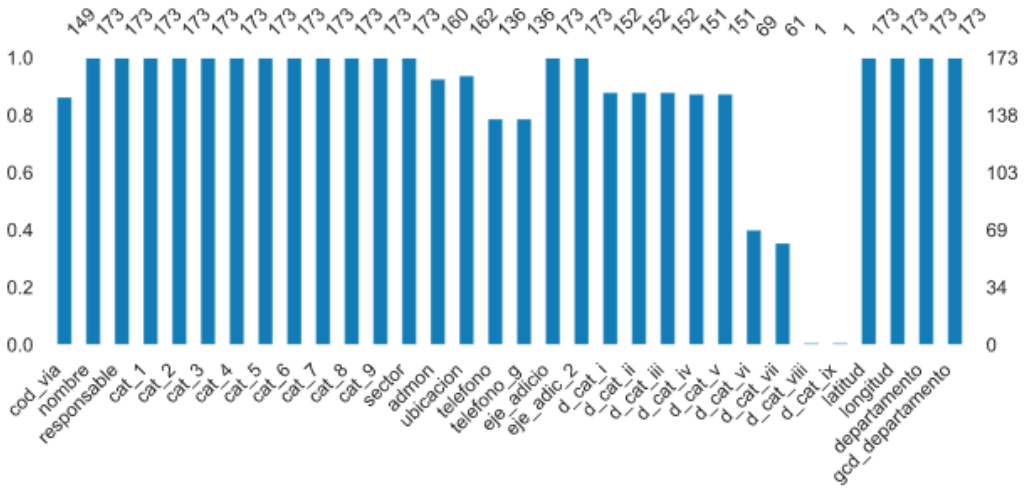


Figura 7: fracción decimal de registros no vacíos por categoría.

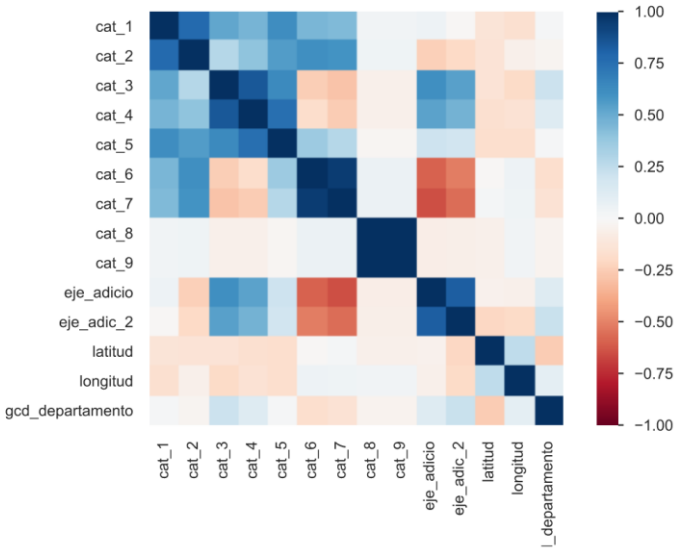


Figura 8: *heatmap* de correlación de Pearson.