Nombres: Nicolás Andrés Tobo Urrutia (201817465), Vilma Marcela Tirado Gómez (201632317), Juan Felipe Rubio Perdomo (201718384)

Universidad de los Andes
Facultad de Ingeniería

Fecha: 20 octubre, 2020

Proyecto 2

Contenido

Modelos de Data Marts

A continuación, se expresará gráfica y verbalmente el modelo dimensional de data marts definido para este proyecto. En primer lugar, se explicará la granularidad de la tabla de hechos.

Infraestructura visible analiza el comportamiento de los peajes en Colombia. En este negocio existen 2 procesos que deben ser analizados, el recaudo que se hace en los peajes y la cantidad de vehículos que transitan por el mismo. En los datos presentados para realizar este análisis tanto el recaudo como el trafico se miden de forma mensual y se cuentan con los datos desde 2016 hasta 2018. Teniendo en cuenta que los 2 procesos se miden de forma mensual durante 3 años se puede decir que tienen la misma granularidad por lo tanto se pueden analizar usando una sola tabla de hechos. En este caso cada fila en la tabla de hechos representa el trafico y el total de dinero recaudado en dado un peaje y el mes a analizar.

Para responder a las necesidades del negocio y concordar con la granularidad de las tablas de hecho se plantean los siguientes hechos: Fecha_FK, Peaje_FK, Lugar_FK. A continuación, se listarán las diferentes medidas de la tabla de hechos:

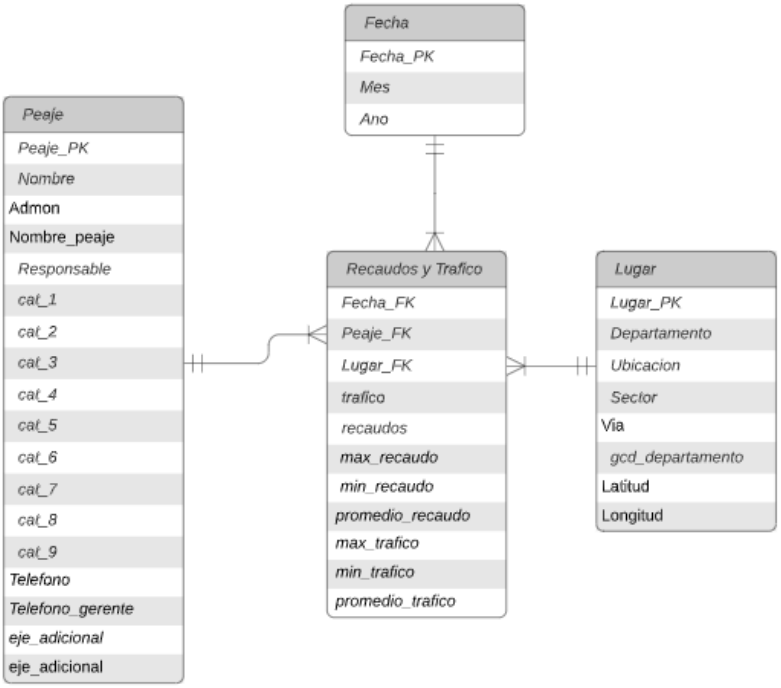
num_trafico	Aditiva
num_recaudos	Aditiva
max_recaudos	Aditiva
min_recaudo	Aditiva
promedio_recaudo	Aditiva
min_trafico	Aditiva
max_trafico	Aditiva
promedio_trafico	Aditiva

Dimensión peaje: Esta dimensión contiene los siguientes atributos Peaje_key, Admon,Nombre,Nombre_proyecto,Responsable,Telefono,Telefono_gerente,cat_1,cat_2,cat_3,cat_4,cat_5,cat_6,cat_7,cat_8,cat_9,eje_adicional,eje_adicional_2. Estos atrubutos nos permiten identificar el peaje por medio de una llave única y caracterizarlo teniendo en cuenta de que proyecto hace parte, que nombre tiene y a cargo de quien esta. De igual forma nos permite conocer las tarifas que tiene ese peaje para cada categoría vehicular.

dimensión fecha: Contiene los atributos Fecha_key, mes, y ano que nos permiten caracterizar la fecha en la cual se genero el proceso acorde a la granularidad explicada anteriormente.

dimensión lugar: Esta dimensión tiene los siguientes atributos Via,Departamento,gcd_departamento,Sector,Ubicacion,Latitud,Longitud,Lugar_key**,** los cuales permiten caracterizar la ubicación geográfica de un peaje.

Teniendo en cuenta lo descrito anteriormente se diseñó el siguiente modelo dimensional.



Perfilamiento de datos y proceso ETL

Perfilamiento

En primer lugar, los datos suministrados para el análisis de este proyecto se encuentran en los siguientes dos archivos: *peajes2019(2).xlsx* y *Tarifas de peaje a cargo del INVIAS - 2019.xlsx.* Con el fin de presentar de mejor forma la descripción y perfilamiento de los datos de cada archivo, desarrollaremos dos secciones.

Descripción y perfilamiento del archivo *Tarifas de peaje a cargo del INVIAS - 2019.xlsx*

Los datos presentados en este documento .xlsx representan de forma general las tarifas de cada peaje en Colombia, en el periodo de 2019. En el documento se encuentran 41 registros con el nombre de cada estación de peaje.

Dentro de esta información se puede encontrar:

- El nombre de la estación de peaje (nominal).
- Las tarifas dadas por categoría vehicular (numérico/categórico). Existen 14 categorías vehiculares que serán explicadas más en detalle en la tabla que sigue a continuación.
- Los precios que se deben pagar por ejes adicionales de los vehículos pesados (numérico). Existen 4 tarifas por cada tipo de vehículo con ejes adicionales.
- La tarifa asociada para el Fondo de Seguridad Vial FOSEVI (numérico).

De estos datos se consideran relevantes para el análisis el nombre de la estación de peaje y las tarifas dadas por categoría vehicular, ya que definen de una forma concreta

cuales son la tarifas por cada vehículo de cada uno de los peajes del INVIAS en el 2019. Lo anterior, se realiza puesto que uno de los análisis OLAP requeridos a investigar por el proyecto es: ***identificar cuáles son los peajes más costosos por categoría de vehículo.*** **(Ver Anexo 1 con tabla descriptiva de categorías por vehículo)*.***

A continuación, vamos a realizar un resumen de las estadísticas descriptivas para las variables numéricas y categóricas de las tarifas por categoría vehicular, teniendo en cuenta que existen múltiples clasificaciones se realizara un análisis por cada categoría con el fin de evitar confusiones. Este análisis es desarrollado con la herramienta *pandas_profiling* de Python. Las graficas arrojadas por la herramienta se encuentran en anexos. (Ver anexo 2 análisis por categoría)

Conclusiones del perfilamiento de datos del archivo *Tarifas de peaje a cargo del INVIAS - 2019.xlsx*

- La clasificación VI es entre las clasificaciones normales, la que menos ocurrencia presenta en los catálogos de tarifas de los peajes.
- Las tarifas de clasificaciones de vehículos especiales pueden ser removidas del análisis, ya que en su mayoría más del 90% de los registros de los peajes, no tienen este tipo de clasificaciones en su catálogo de tarifas.
- La herramienta *panda_profiling* puede convertir datos aparentemente numéricos en categóricos si identifica que no existen muchos valores diferentes entre los registros.
- Las tarifas de categorías de vehículos con clasificación I y II son las que presentan mayor ocurrencia entre todos los peajes.

Descripción y perfilamiento del archivo *Peajes2019.xlsx*

A continuación, se hará el análisis para los datos contenidos en el archivo *peajes2019(2).xlsx* que consta de una parte muy importante para el desarrollo del proyecto. Esto se debe a que contiene códigos de vías y ubicación geográficas que permitirán hacer un análisis a profundidad de las rutas y no únicamente de los peajes. Un resumen general de lo que se encuentra en este Excel consiste en:

- Códigos de la vía en la que se encuentra el peaje y la ubicación geográfica de los peajes, como el departamento en el que se encuentran y el código departamental.
- El responsable de gestionar el mantenimiento y la operación del peaje (más adelante se expresará específicamente cuáles son los posibles valores).
- Las diferentes categorías que pueden circular por cada tipo de peaje y por último la definición de las diferentes categorías mencionadas anteriormente para cada peaje.

Para analizar la información suministrada y poder tomar decisiones sobre esta es importante conocer qué información es relevante para el negocio. En este caso, es interesante determinar los recaudos por peaje diferenciando las categorías de los vehículos, las rutas, el año y el departamento. Es por esto, que es de vital importancia

mantener toda la información que haga referencia a la ubicación del peaje y también la información básica de los peajes. Por otro lado, es interesante analizar el proyecto por vías y para esto también es necesario guardar la información. Sin embargo, hay información que no aporta al análisis OLAP que espera la empresa, toda la información de los responsables del peaje, la administración y el contacto del peaje puede ser omitida. Además, la definición de las categorías puede ser almacenada en un espacio diferente, aunque por el momento se mantendrá para entender mejor el negocio.

Se hará un breve diccionario con las variables que se mantienen para el estudio:

Nombre	***Definición***	***Tipo***
Cod_via	Código de la vía	ID, Nominal
Nombre	Nombre del peaje	Nominal
Cat_i	Valor de peaje para la categoría i	Numérica
Sector	Secto en el que se encuentra el peaje	Nominal
Ubicacion	Ubicación del peaje	Nominal
D_cat_i	Definición para el peaje de la categoría i	Nominal
Latitud	Latitud	Geográfica/numérica
Longitud	Longitud	Geográfica/numérica
Departamento	Departamento en el que se encuentra ubicado el peaje	Nominal

Ahora se hará un análisis de las variables escogidas a través de la herramienta *pandas_profiling* de *Python*. Esto con el fin de entender mejor las variables que se escogen para el posterior uso en el proyecto. (Ver anexo 3 para ver mas detalles del paso a paso del perfilamiento)

Conclusiones del perfilamiento de datos del archivo *Peajes2019(2).xlsx*

- Hay muchas clasificaciones que no tienen valor por lo que *pandas_profiling* la cuenta como ceros al ser numérica. Esto tiene sentido, ya que, hay muchos peajes que solo aceptan cierto tipo de vehículos, por lo que, no se tomarán como valores faltantes, sino que simplemente se entiende que no presta servicio para ese tipo de categoría.
- La información de contacto y administración puede ser eliminada puesto que no presenta información relevante para el negocio y para el departamento solo se mantendrá uno de los dos diferentes valores.

Las gráficas arrojadas por la herramienta se encuentran en anexos. Adicionalmente se analizó el conjunto con los datos **Recaudo y Trafico en Peajes 2016-2018.xlxs** el cual contiene el tráfico y los recaudos de cada peaje durante estos años, junto con la información de cada peaje, la cual se distribuyó en las dimensiones acorde al modelo dimensional.

Proceso ETL

Extracción

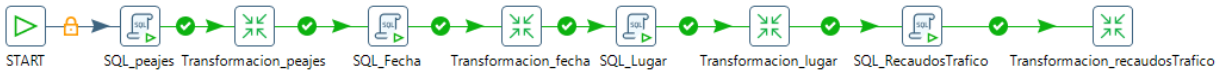
Durante el proceso de extracción de los datos fue posible darnos cuentas de que existían peajes que se encontraban repetidos como ANDES, CHINAUTA y FLANDES entre otros. La repetición de estos peajes en la fuente se debe a que los peajes siguen estando en el mismo lugar y con el mismo nombre, pero hacen parte de diferentes proyectos. Es por esto por lo que muchas veces los primeros meses registrados se encuentran en 0 en la primera fila de nombre repetido y en la segunda los últimos meses se encuentran en 0, se asume que los primeros meses la información fue registrada por un proyecto y posteriormente por otro proyecto. Para manejar con esta repetición se decidió sumar todos los recaudos en una sola fila y tomar como nombre de proyecto el mas reciente ya que se asume que este es el dueño actual del peaje. Se aplico lo mismo en el caso del tráfico.

Transformación

Para la transformación de los datos se utilizó tableau específicamente la herramienta de manejo de bases de datos la cual nos permitió realizar un Full outer Join entre la información de **peajes 2019** y **Recaudo y Trafico en Peajes 2016-2018,** lo anterior nos permitió tener una tabla principal que fue exportada a Excel. A partir de la tabla principal se crearon las tablas de dimensiones con la información descrita en el modelo dimensional. Finalmente, con las llaves de cada tabla dimensional y la información de tráfico y recaudo se creó la tabla de hechos.

Carga

Una vez se obtuvieron las tablas del modelo dimensional en Excel se exportaron a formato csv para poder realizar el proceso de carga con la herramienta Spoon de Pentajo, el flujo utilizado para la carga de datos es el siguiente:



Como se puede ver se creo la respectiva tabla para cada una de las dimensiones en la base de datos y posteriormente fueron pobladas utilizando los csv generados anteriormente usando el nodo de transformación

A continuación, se muestra el resultado de los datos ya cargados

	lugar_key [PK] character varying (30)	departamento character varying (30)	ubicacion character varying (150)	sector character varying (100)	via character varying (20)	gcd_departamento integer	latitud double precision	longitud double precision
1	5.749465359-75.62728596	Antioquia	KM 03 RUTA 25B01 Via Pinta...	La Pintada - Penalisa	25B01	5	6	-76
2	5.968901178-75.59407236	Antioquia	KM 38+800 RUTA 2509; Via ...	La Pintada - Primavera	2509	5	6	-76
3	6.046953761-75.6598771	Antioquia	Km. 89+396 Ruta Nacional 60...	La Mansa - Primavera	6003	5	6	-76

Ilustración 1 Primeras filas de la dimensión lugar

	fecha_key [PK] character varying (20)	mes character varying (10)	ano integer
1	2016Ene	Ene	2016
2	2016Feb	Feb	2016
3	2016Mar	Mar	2016
4	2016Abr	Abr	2016

Ilustración 2 Primeras filas de la dimensión fecha

	peaje_key [PK] character varying (150)	nombre character varying (200)	admon character varying (200)	nombre_proyecto character varying (200)	responsable character varying (200)	cat_1 integer	cat_2 integer	cat_3 integer	cat_4 integer	cat_5 integer
1	EL PLACER-Concesionaria VÃal Un...	EL PLACER	Concesionaria VÃal UnÃon de...	RUMICHACA - PASTO	ConcesiÃ³n ANI	9900	10400	22100	28800	33200
2	DAZA-INVIAS	DAZA	INVIAS	RUMICHACA - PASTO - CHAC...	INVIAS	0	0	22200	29100	33600
3	CANO-INVIAS	CANO	INVIAS	RUMICHACA - PASTO - CHAC...	INVIAS	10000	10700	0	0	0

cat_6 integer	cat_7 integer	cat_8 integer	cat_9 integer	telefono character varying (200)	telefono_gerente character varying (200)	eje_adicional bigint	eje_adicional_2 bigint
0	0	0	0	3173682092 â€“ 3173310921	MÃ³vil1: 3174275839, atiende...	0	0
0	0	0	0	-1	-1	9100	9400
0	0	0	0	-1	-1	9100	9400

Ilustración 3 Primeras filas de la dimensión peaje

	trafico bigint	recaudo bigint	fk_fecha [PK] character varying (20)	fk_peaje [PK] character varying (150)	fk_lugar [PK] character varying (30)
81	226546	2305675800	2016Mar	TUNIA-INVIAS	2.701690113-76.53681132
82	201084	2111622100	2016May	TUNIA-INVIAS	2.701690113-76.53681132
83	214318	2231469450	2016Nov	TUNIA-INVIAS	2.701690113-76.53681132
84	212793	2209695700	2016Oct	TUNIA-INVIAS	2.701690113-76.53681132
85	199745	2113931150	2016Sep	TUNIA-INVIAS	2.701690113-76.53681132
86	7554	2377006000	2017Abr	TUNIA-INVIAS	2.701690113-76.53681132

Ilustración 4 Ejemplo de la tabla de hechos

Las imágenes mostradas anteriormente se tomaron haciendo de la base de datos en Postgres. Estos datos constan de 4717 filas y 31 columnas, donde se encontraron los siguientes estadísticos.

- Máximo para recaudo: 12,552,733,500
- Mínimo de recaudo: 0
- Máximo para tráfico: 2,185,983,800
- Mínimo de tráfico: 0

Análisis OLAP

Con el fin de probar todo el modelo dimensional generado a lo largo del documento, se crearon 4 análisis OLAP 3 propuestos por el cliente y 1 propuesto por el grupo.

Análisis OLAP 1: Información de recaudo y tráfico por mes en cada departamento

a) El resultado de este análisis OLAP está dirigido a solventar el objetivo de negocio planteado por Infraestructura visible: *Análisis de los recaudos hechos en peajes y sus variaciones en el tiempo por peaje especifico, ruta a la que pertenece el peaje y departamentos en Colombia. Mostrar Máximos, Mínimos, Promedios y distribución por ubicación geográfica.*

b) A partir de la información dada por el análisis, un analista de negocio puede realizar un promedio de recaudos anuales por departamento. Esto le permitirá identificar cuales departamentos generan el mayor recaudo en un historial de tres años, y con base a ello poder definir en cuales lugares seria ideal construir nuevos peajes, para generar más recaudo.

c)

Row Labels	Recaudo	Trafico	MaximoTrafico	MaximoRecaudo
2016				
Abr				
Antioquia	15751864050	1316805	536188	7765660000
Atlantico	5860873600	696872	286384	2047208700
Bolivar	7517496200	995677	455233	1939218100
Boyaca	3833319450	411777	335927	3072710950
Caldas	5594194250	453103	116819	1418827800
Casanare	761539300	65591	65591	761539300
Cauca	4605458900	463544	277106	2617425900
Cesar	10748166200	16474752	15697200	2690094300
Cordoba	4871623650	730684	406289	1870379500
Cundinamarca	58186753650	5417698	1025509	9415170700
Huila	4999028000	237695	148280	1641735000
La Guajira	2577783500	299462	135545	840841000
Magdalena	4478884100	290996	158484	2052236300
Meta	14289456100	1077391	261428	6196049300
Narino	1757557000	181944	181944	1757557000
Norte de Santander	1615072700	593625	366972	1115880600
Quindio	2725188800	196234	196234	2725188800
Risaralda	6828190850	501469	290987	4119695600
Santander	13579139700	1010897	139862	2546234700
Sucre	2577627900	395076	297445	1534592800
Tolima	13158981900	866528	377511	5067527150
Valle del Cauca	21031445900	1736907	412043	4079244350
Ago				
Antioquia	22214322050	1801111	634257	9066145900
Atlantico	6287052400	744601	311820	2088896100
Bolivar	8051026400	1079987	481812	2022578800
Boyaca	4433642200	483870	385503	3504303700
Caldas	7647599450	613130	156531	2310237250

Análisis OLAP 2: Información de recaudo y tráfico por código de vía y peajes pertenecientes a ellas.

a) El resultado de este análisis OLAP está dirigido a solventar el objetivo de negocio planteado por Infraestructura visible: *Análisis del tráfico en Colombia y sus variaciones en el tiempo por peaje especifico, ruta a la que pertenece el peaje y departamentos en Colombia. Mostrar Máximos, Mínimos, Promedios y distribución por ubicación geográfica.*

b) A partir de la información dada por el análisis, un analista de negocio puede visualizar cuales son las vías, los peajes que la conforman y los tráficos que existen en los peajes por cada mes en un año. Con ello se puede analizar cuales vías son las mas transitadas y con ello cuales requieren reparaciones por el desgaste generado por el continuo uso de las vías.

c)

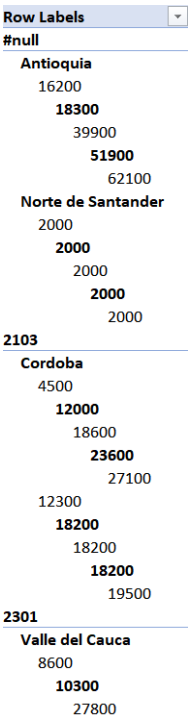
Row Labels	MaximoTrafico	PromedioTrafico	MinimoTrafico	Trafico
2016				
Antioquia				
#null				
PALMITAS				
Abr	0	0	0	0
Ago	102756	102756	102756	102756
Dic	111372	111372	111372	111372
Ene	0	0	0	0
Feb	0	0	0	0
Jul	113653	113653	113653	113653
Jun	0	0	0	0
Mar	0	0	0	0
May	0	0	0	0
Nov	103254	103254	103254	103254
Oct	114494	114494	114494	114494
Sep	94356	94356	94356	94356
SAN CRISTOBAL				
Abr	0	0	0	0
Ago	102953	102953	102953	102953
Dic	136925	136925	136925	136925
Ene	0	0	0	0
Feb	0	0	0	0
Jul	111671	111671	111671	111671
Jun	0	0	0	0
Mar	0	0	0	0
May	0	0	0	0
Nov	103944	103944	103944	103944
Oct	113155	113155	113155	113155
Sep	95336	95336	95336	95336
2509				

Análisis OLAP 3: Información de tarifas de peajes por categoría, departamento y código de vía.

a) El resultado de este análisis OLAP está dirigido a solventar el objetivo de negocio planteado por Infraestructura visible: *Análisis del costo de viajar a través de la vía primaria de Colombia y sus variaciones en el tiempo por categoría de vehículo y Ruta. Mostrar Máximos, Mínimos, Promedios y distribución por ubicación geográfica, ruta y puntos deseados de viaje.*

b) A partir de la información dada por el análisis, un analista de negocio puede calcular cual es el costo promedio de pago de peajes de un vehículo para una vía promedio. Esto es útil al momento de calcular rutas que tomarían las personas para evitar peajes y así reducir costos en los viajes. Por lo tanto, el negocio puede plantearse el poner un peaje en esos lugares.

c)



Row Labels
#null
Antioquia
16200
18300
39900
51900
62100
Norte de Santander
2000
2000
2000
2000
2000
2103
Cordoba
4500
12000
18600
23600
27100
12300
18200
18200
18200
19500
2301
Valle del Cauca
8600
10300
27800

Análisis OLAP 4: Recaudos por vía y peaje.

a) El resultado de este análisis OLAP está dirigido a solventar el objetivo de negocio planteado por Infraestructura visible: *Análisis de los recaudos hechos en peajes y sus variaciones en el tiempo por peaje especifico, ruta a la que pertenece el peaje y departamentos en Colombia. Mostrar Máximos, Mínimos, Promedios y distribución por ubicación geográfica.*

b) En la siguiente tabla se puede hacer un análisis para cada peaje de sus recaudos y las estadísticas para sus recaudos en un año específico. Además, se puede analizar para cada vía, es decir se puede analizar rutas. Con el fin de calcular el recaudo total de un trayecto.

Row Labels	Recaudo	PromedioRecaudo	MinimoRecaudo	MaximoRecaudo
#null				
EL ESCOBAL				
2016				
Norte de Santander	470196750	39183062.5	31516000	47241000
2017				
Norte de Santander	543778250	45314854.166666666667	40087500	51503200
2018				
Norte de Santander	602063100	50171925	42268700	58381000
PALMITAS				
2016				
Antioquia	10217393872	851449489.33333333	0	1801628300
2017				
Antioquia	22631573050	1885964420.83333333	1595590300	2292398450
2018				
Antioquia	24724505800	2060375483.33333333	1665738300	2574916900
SAN CRISTOBAL				
2016				
Antioquia	10609842878	884153573.16666667	0	2119734528
2017				
Antioquia	22756278650	1896356554.16666667	1609127450	2471531500
2018				
Antioquia	24638960200	2053246683.33333333	1686556800	2661981900
2103				

Trabajo en equipo

Durante el desarrollo del proyecto, se decidió realizar reuniones entre los miembros del equipo con el fin de socializar el desarrollo del proyecto, revisar las actividades realizadas, identificar tareas faltantes y administrar las responsabilidades establecidas para cada una de las micro entregas y para la entrega final del proyecto.

A continuación, se presenta la tabla de resumen que establece la organización de los integrantes del equipo frente a las actividades requeridas para el proyecto. Los puntos obtenidos por cada integrante se asignaron de manera equitativa teniendo en cuenta el cumplimiento de cada estudiante con la tarea asignada, la disposición frente a la asignación de dichas tareas y el tiempo requerido para finalizarlas.

Integrante	Tareas Realizadas	Tiempo Utilizado	Puntos
Juan Rubio	• Comprensión y preparación de los datos. • Modelado y Evaluación (ETL). • Análisis de resultados.	10 horas.	33.3
Vilma Tirado	• Comprensión y preparación de los datos. • Modelado y Evaluación (ETL). • Análisis de resultados	10 horas.	33.3
Nicolás Tobo	• Comprensión y preparación de los datos. • Modelado y Evaluación (ETL). • Análisis de resultados	10 horas.	33.3

Anexos

1. Tabla categorías vehículos

Categoría vehicular	Descripción	Tipo variable
I	Automóviles, camperos y camionetas	Numérica
IE	Vehículos tarifa especial Categoría I	Numérica
II	Buses, busetas, microbuses con eje trasero de doble llanta y camiones de dos ejes	Numérica
IIE	Vehículos tarifa especial Categoría II	Numérica
IIEE	Vehículos tarifa especial Categoría IIEE	Categórica
III	Camiones de tres ejes	Numérica
IIIE	Vehículos tarifa especial Categoría III	Categórica
IV	Camiones de cuatro ejes	Numérica
IVE	Vehículos tarifa especial Categoría IV	Categórica
V	Camiones de cinco ejes	Numérica
VB	Vehículos tarifa especial Categoría VB	Categórica
VE	Vehículos tarifa especial Categoría V	Categórica
VI	Camiones de seis ejes	Categórica
VII	Camiones de siete ejes	Numérica

2. Análisis por categoría archivo Tarifas con Pandas Profiling

- **Categoría vehicular I:**
  - Estadísticas descriptivas**:**

Statistics | Histogram | Common values | Extreme values

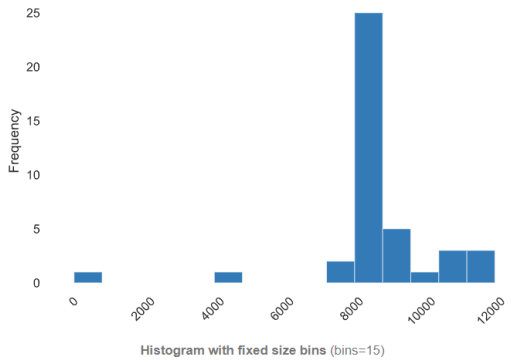
Quantile statistics

Minimum	0
5-th percentile	7900
Q1	8300
median	8600
Q3	9300
95-th percentile	12000
Maximum	12100
Range	12100
Interquartile range (IQR)	1000

Descriptive statistics

Standard deviation	1893.599618
Coefficient of variation (CV)	0.2184512784
Kurtosis	11.12868236
Mean	8668.292683
Median Absolute Deviation (MAD)	400
Skewness	-2.246877549
Sum	355400
Variance	3585719.512
Monotocity	Not monotonic

  - Histogramas de frecuencia de precio de tarifas de categoría I en los peajes:



Value	Count	Frequency (%)
8600	7	17.1%
8300	6	14.6%
8500	5	12.2%
8200	5	12.2%
10500	3	7.3%
8700	2	4.9%
9400	2	4.9%
7900	2	4.9%
9300	2	4.9%
12100	2	4.9%
Other values (5)	5	12.2%

- **Categoría vehicular IE:**
  - Estadísticas descriptivas**:**

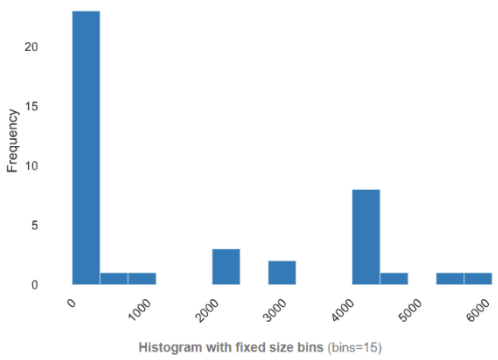
Quantile statistics

Minimum	0
5-th percentile	0
Q1	0
median	0
Q3	4300
95-th percentile	4600
Maximum	6200
Range	6200
Interquartile range (IQR)	4300

Descriptive statistics

Standard deviation	2081.70896
Coefficient of variation (CV)	1.279611205
Kurtosis	-1.064129124
Mean	1626.829268
Median Absolute Deviation (MAD)	0
Skewness	0.7743696785
Sum	66700
Variance	4333512.195
Monotocity	Not monotonic

  - Histogramas de frecuencia de precio de tarifas de categoría IE en los peajes:



Value	Count	Frequency (%)
0	21	51.2%
4400	4	9.8%
4500	2	4.9%
4300	2	4.9%
200	2	4.9%
6200	1	2.4%
4600	1	2.4%
2100	1	2.4%
3000	1	2.4%
1000	1	2.4%
Other values (5)	5	12.2%

- **Categoría vehicular II:**
  - Estadísticas descriptivas**:**

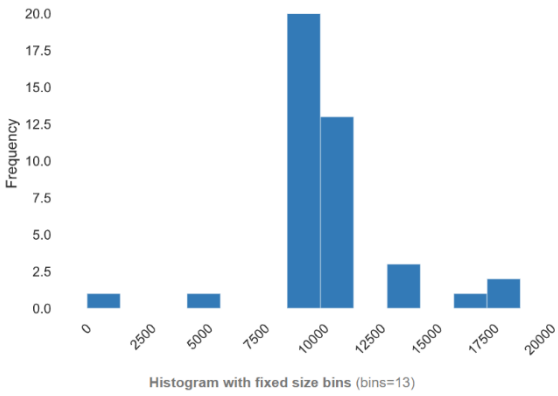
Quantile statistics

Minimum	0
5-th percentile	8900
Q1	9000
median	9900
Q3	10300
95-th percentile	17400
Maximum	19100
Range	19100
Interquartile range (IQR)	1300

Descriptive statistics

Standard deviation	3151.341178
Coefficient of variation (CV)	0.3084387403
Kurtosis	4.732755523
Mean	10217.07317
Median Absolute Deviation (MAD)	900
Skewness	0.5239010947
Sum	418900
Variance	9930951.22
Monotocity	Not monotonic

  - Histogramas de frecuencia de precio de tarifas de categoría II en los peajes:



Value	Count	Frequency (%)
10300	9	22.0%
8900	7	17.1%
9000	6	14.6%
9300	5	12.2%
13300	3	7.3%
9900	2	4.9%
10400	2	4.9%
19100	2	4.9%
17400	1	2.4%
11600	1	2.4%
Other values (3)	3	7.3%

- **Categoría vehicular IIE:**

- Estadísticas descriptivas**:**

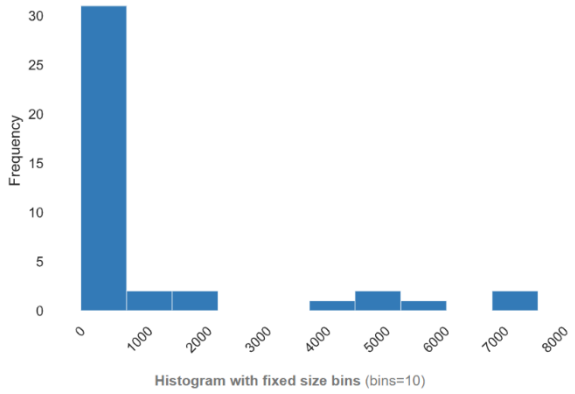
Quantile statistics

Minimum	0
5-th percentile	0
Q1	0
median	0
Q3	200
95-th percentile	5500
Maximum	7700
Range	7700
Interquartile range (IQR)	200

Descriptive statistics

Standard deviation	2196.374507
Coefficient of variation (CV)	2.094217553
Kurtosis	3.200374548
Mean	1048.780488
Median Absolute Deviation (MAD)	0
Skewness	2.094684212
Sum	43000
Variance	4824060.976
Monotocity	Not monotonic

- Histogramas de frecuencia de precio de tarifas de categoría IIE en los peajes:

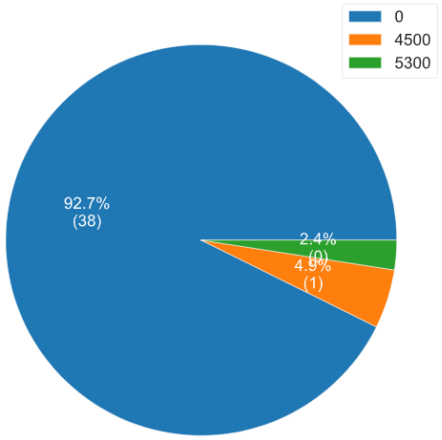


Value	Count	Frequency (%)
0	28	68.3%
200	3	7.3%
5300	2	4.9%
7700	2	4.9%
5500	1	2.4%
2300	1	2.4%
4600	1	2.4%
1000	1	2.4%
800	1	2.4%
2200	1	2.4%

- **Categoría vehicular IIEE:**
  - Número de registros y registros faltantes:

Distinct	3
Distinct (%)	7.3%
Missing	0
Missing (%)	0.0%
Memory size	328.0 B

- Diagrama de pie:



- **Categoría vehicular III:**
  - Estadísticas descriptivas**:**

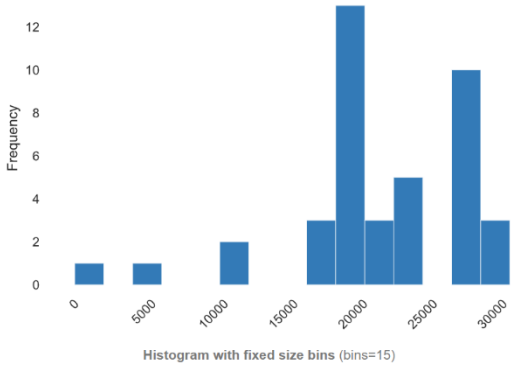
Quantile statistics

Minimum	0
5-th percentile	10800
Q1	19200
median	21400
Q3	27800
95-th percentile	31200
Maximum	31200
Range	31200
Interquartile range (IQR)	8600

Descriptive statistics

Standard deviation	6645.189898
Coefficient of variation (CV)	0.3098519115
Kurtosis	1.93196393
Mean	21446.34146
Median Absolute Deviation (MAD)	2600
Skewness	-1.019246356
Sum	879300
Variance	44158548.78
Monotocity	Not monotonic

  - Histograma de frecuencia de precio de tarifas de categoría III en los peajes:

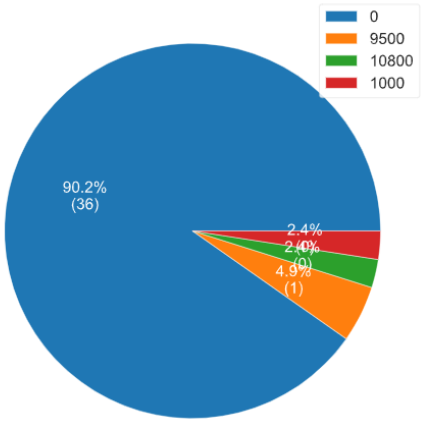


Value	Count	Frequency (%)
27800	7	17.1%
19200	6	14.6%
22900	5	12.2%
19500	5	12.2%
31200	3	7.3%
10800	2	4.9%
18800	2	4.9%
21400	2	4.9%
28100	2	4.9%
16700	2	4.9%
Other values (5)	5	12.2%

- **Categoría vehicular IIIE:**
  - Número de registros y registros faltantes:

Distinct	4
Distinct (%)	9.8%
Missing	0
Missing (%)	0.0%
Memory size	328.0 B

  - Diagrama de pie:



- **Categoría vehicular IV:**
  - Estadísticas descriptivas**:**

Quantile statistics

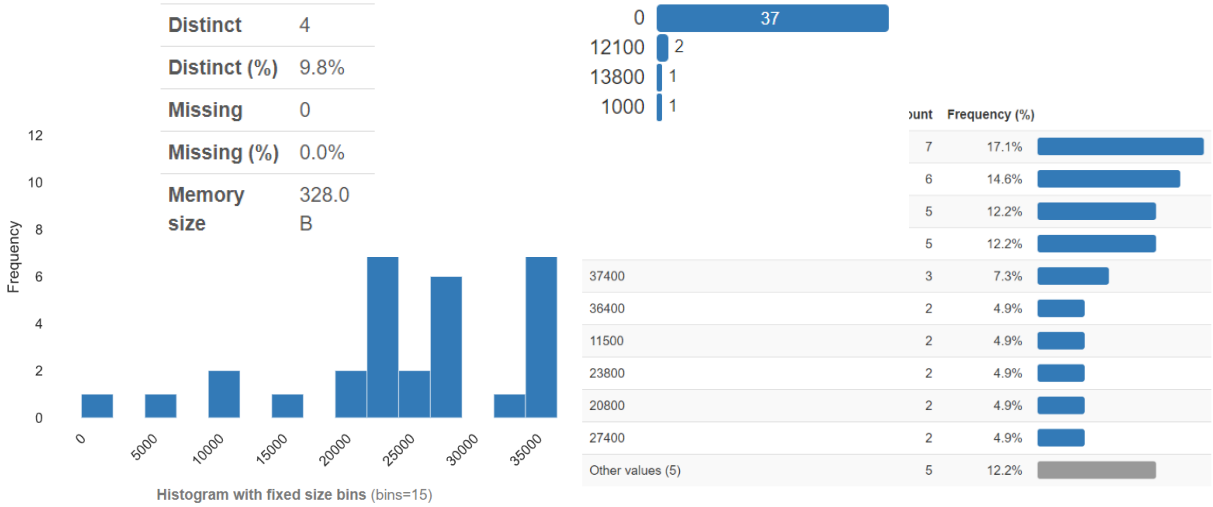
Minimum	0
5-th percentile	11500
Q1	24400
median	27400
Q3	36300
95-th percentile	37400
Maximum	37400
Range	37400
Interquartile range (IQR)	11900

Descriptive statistics

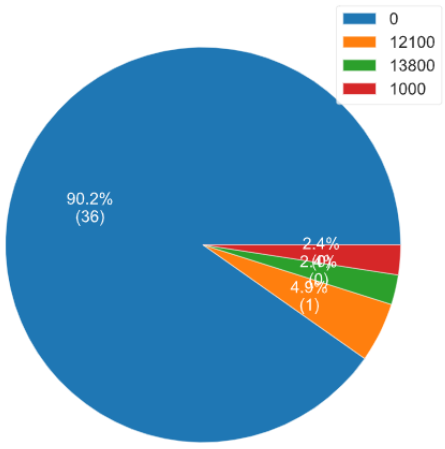
Standard deviation	8755.330084
Coefficient of variation (CV)	0.3250643244
Kurtosis	1.479141261
Mean	26934.14634
Median Absolute Deviation (MAD)	3600
Skewness	-1.041912983
Sum	1104300
Variance	76655804.88
Monotocity	Not monotonic

- Histograma de frecuencia de precio de tarifas de categoría IV en los peajes:

- **Categoría vehicular IVE:**
  - Número de registros y registros faltantes:



  - Diagrama de pie:



- **Categoría vehicular V:**

- Estadísticas descriptivas**:**

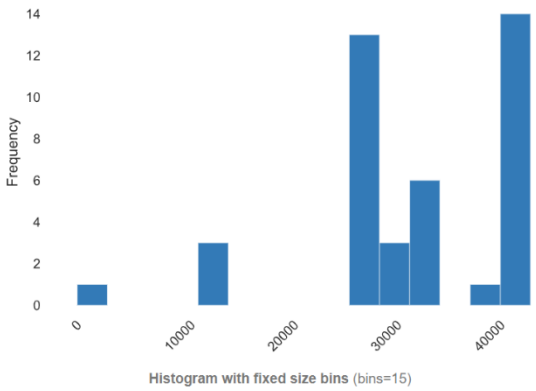
Quantile statistics

Minimum	0
5-th percentile	12400
Q1	27400
median	33300
Q3	41800
95-th percentile	43900
Maximum	43900
Range	43900
Interquartile range (IQR)	14400

Descriptive statistics

Standard deviation	9954.760474
Coefficient of variation (CV)	0.3093649507
Kurtosis	1.76579147
Mean	32178.04878
Median Absolute Deviation (MAD)	5900
Skewness	-1.148491566
Sum	1319300
Variance	99097256.1
Monotocity	Not monotonic

- Histograma de frecuencia de precio de tarifas de categoría V en los peajes:

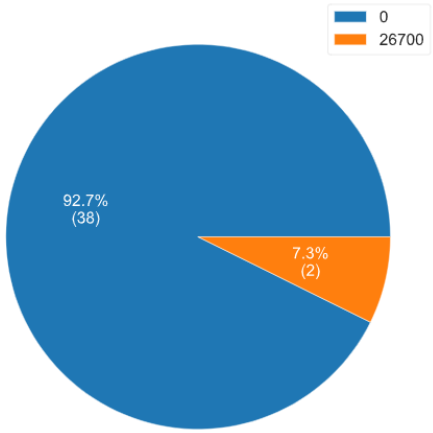


Value	Count	Frequency (%)
41800	7	17.1%
27400	6	14.6%
27900	5	12.2%
33300	5	12.2%
43900	3	7.3%
12400	2	4.9%
41900	2	4.9%
27300	2	4.9%
31200	2	4.9%
41600	2	4.9%
Other values (5)	5	12.2%

- **Categoría vehicular VB:**
  - Número de registros y registros faltantes:

Distinct	2
Distinct (%)	4.9%
Missing	0
Missing (%)	0.0%
Memory size	328.0 B

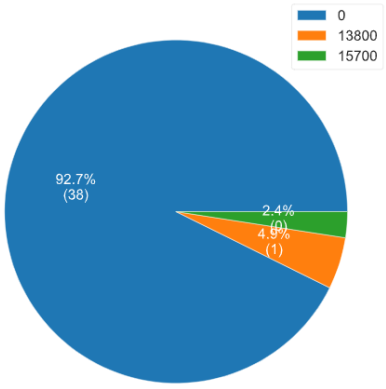
- Diagrama de pie:



- **Categoría vehicular VE:**
  - Número de registros y registros faltantes:

Distinct	3
Distinct (%)	7.3%
Missing	0
Missing (%)	0.0%
Memory size	328.0 B

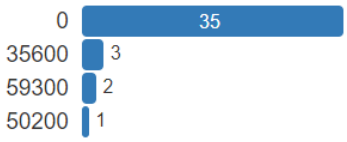
  - Diagrama de pie:



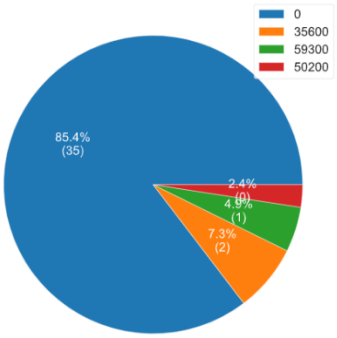
- **Categoría vehicular VI:**

- Número de registros y registros faltantes:

Distinct	4
Distinct (%)	9.8%
Missing	0
Missing (%)	0.0%
Memory size	328.0 B



- Diagrama de pie:



- **Categoría vehicular VII:**
  - Estadísticas descriptivas**:**

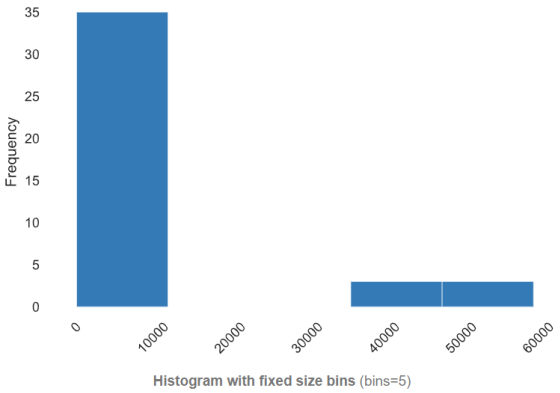
Quantile statistics

Minimum	0
5-th percentile	0
Q1	0
median	0
Q3	0
95-th percentile	57900
Maximum	59300
Range	59300
Interquartile range (IQR)	0

Descriptive statistics

Standard deviation	18152.29879
Coefficient of variation (CV)	2.49244558
Kurtosis	3.462270018
Mean	7282.926829
Median Absolute Deviation (MAD)	0
Skewness	2.23926977
Sum	298600
Variance	329505951.2
Monotocity	Not monotonic

  - Histograma de frecuencia de precio de tarifas de categoría I en los peajes:



Value	Count	Frequency (%)
0	35	85.4%
59300	2	4.9%
40800	2	4.9%
40500	1	2.4%
57900	1	2.4%

3. Análisis categorías perfilamiento archivo Peajes2019.xlsx

A continuación, se encuentran las estadísticas descriptivas de todos los datos:

Dataset statistics

Number of variables	32
Number of observations	173
Missing cells	789
Missing cells (%)	14.3%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	43.4 KiB
Average record size in memory	256.7 B

Variable types

CAT	22
NUM	10

Código de la vía:

cod_via
Categorical

HIGH CARDINALITY
MISSING
UNIFORM

Distinct	100
Distinct (%)	67.1%
Missing	24
Missing (%)	13.9%
Memory size	1.4 KiB

2515	3
2510	3
2103	3
6602	3
5501	3
Other values (95)	134

- Se resalta que con esto se hallarán las rutas para el análisis OLAP.

Nombre:

nombre
Categorical

HIGH CARDINALITY
UNIFORM

Distinct	172
Distinct (%)	99.4%
Missing	0
Missing (%)	0.0%
Memory size	1.4 KiB

SAN PEDRO	2
LOS SANTOS	1
EL COPEY	1
GAMBOTE	1
VILLARICA	1
Other values (167)	167

Toggle details

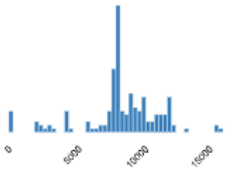
Cat_1:

cat_1
Real number ($\mathbb{R}_{\geq 0}$)

ZEROS

Distinct	61
Distinct (%)	35.3%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	8690.751445
Minimum	0
Maximum	16600
Zeros	6
Zeros (%)	3.5%
Memory size	1.4 KiB



Acá se puede observar como la categoría 1 tiene una distribución con un centro observable en el que circundan los precios por lo que facilita el análisis de los valores.

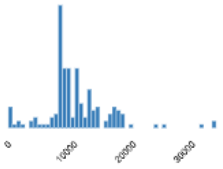
Cat_2:

cat_2
Real number ($\mathbb{R}_{\geq 0}$)

ZEROS

Distinct	67
Distinct (%)	38.7%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	11446.24277
Minimum	0
Maximum	35300
Zeros	6
Zeros (%)	3.5%
Memory size	1.4 KiB



Toggle details

Esta categoría cuenta buses y camiones pequeños con 2 ejes pequeños y también sigue una distribución orientada a un valor aunque hay algunos casos excepcionales por fuera de la desviación.

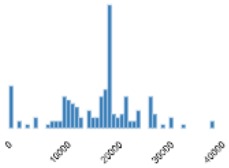
Cat_3:

cat_3
Real number (ℝ≥0)

ZEROS

Distinct	70
Distinct (%)	40.5%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	17168.78613
Minimum	0
Maximum	39900
Zeros	12
Zeros (%)	6.9%
Memory size	1.4 KiB



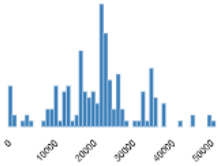
Cat_4:

cat_4
Real number (ℝ≥0)

ZEROS

Distinct	76
Distinct (%)	43.9%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	23169.9422
Minimum	0
Maximum	53100
Zeros	7
Zeros (%)	4.0%
Memory size	1.4 KiB



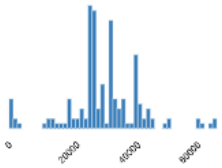
Cat_5:

cat_5
Real number (ℝ≥0)

ZEROS

Distinct	83
Distinct (%)	48.0%
Missing	0
Missing (%)	0.0%
Infinite	0
Infinite (%)	0.0%

Mean	31322.54335
Minimum	0
Maximum	68700
Zeros	6
Zeros (%)	3.5%
Memory size	1.4 KiB



Toggle details

Cat_6:

cat_6

Real number (ℝ≥0)

HIGH_CORRELATION

ZEROS

Distinct	40	Mean	19043.3526
Distinct (%)	23.1%	Minimum	0
Missing	0	Maximum	91800
Missing (%)	0.0%	Zeros	100
Infinite	0	Zeros (%)	57.8%
Infinite (%)	0.0%	Memory size	1.4 KiB

Toggle details

Cat_7:

cat_7

Real number (ℝ≥0)

HIGH_CORRELATION

ZEROS

Distinct	44	Mean	20425.43353
Distinct (%)	25.4%	Minimum	0
Missing	0	Maximum	101900
Missing (%)	0.0%	Zeros	102
Infinite	0	Zeros (%)	59.0%
Infinite (%)	0.0%	Memory size	1.4 KiB

Toggle details

Cat_8 y Cat_9:

cat_8

Categorical

HIGH_CORRELATION

Distinct	2
Distinct (%)	1.2%
Missing	0
Missing (%)	0.0%
Memory size	1.4 KiB

Toggle details

cat_9

Categorical

HIGH_CORRELATION

Distinct	2
Distinct (%)	1.2%
Missing	0

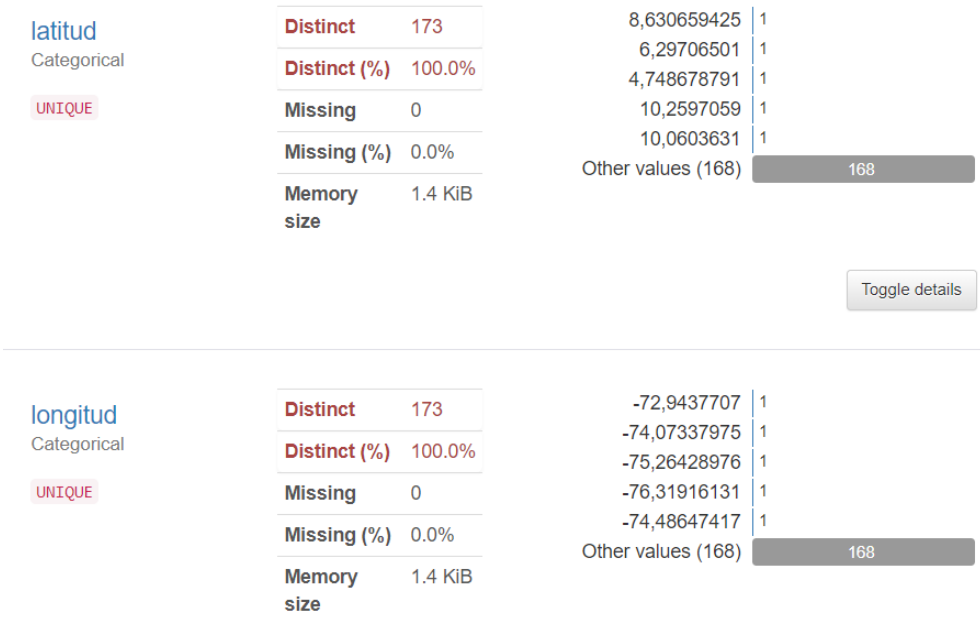
Def_i:

Para esta categoría de variables no se agregarán las estadísticas puesto que son variables nominales que definen el tipo de vehículo que debe pagar la cat_i correspondiente. En este caso es mejor listar los posibles valores de la variable:

- Automóviles y Camperos
- Camiones y buses 2 ejes pequeños
- Categoría III camiones y buses 2 ejes grandes.

Es decir, es la definición de los tipos de vehículos que pagan la categoría con índice i.

Latitud y Longitud:

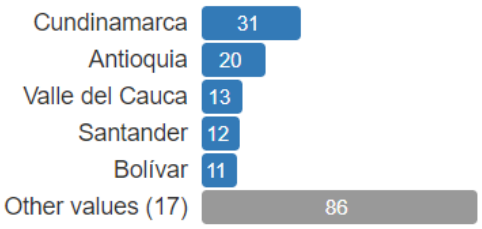


Gráfica de los peajes a través de sus latitudes y longitudes a través de la aplicación: https://maps.co/gis/#. Información dinámica se encuentra en: https://maps.co/map/5f90af7ade9523389543433c99a2

Departamento:

departamento
Categorical

Distinct	22
Distinct (%)	12.7%
Missing	0
Missing (%)	0.0%
Memory size	1.4 KiB



Toggle details