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

**Fecha:** 20 octubre, 2020

**Proyecto 2 – Perfilamiento de datos proyecto *Infraestructura visible***

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.

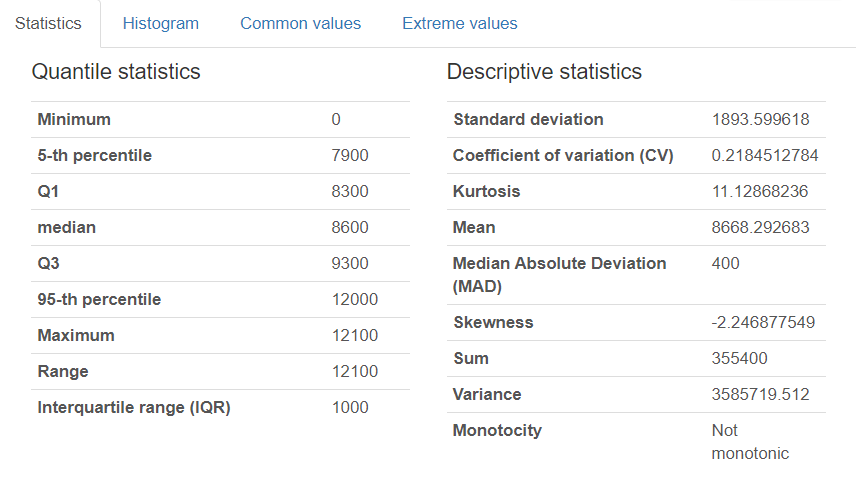
Dentro de esta información se puede encontrar:

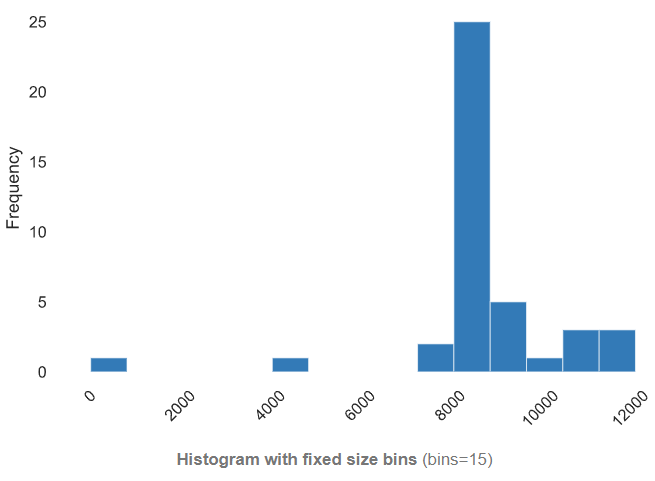
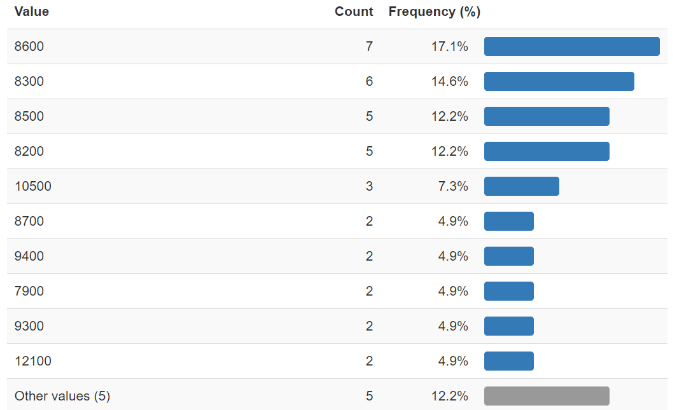
* El nombre de la estación de peaje (nominal).
* Las tarifas dadas por categoría vehicular (numérico/categórico). Existen 14 categorías vehiculares que serán explicadas más en detalle en la tabla que sigue a continuación.
* Los precios que se deben pagar por ejes adicionales de los vehículos pesados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

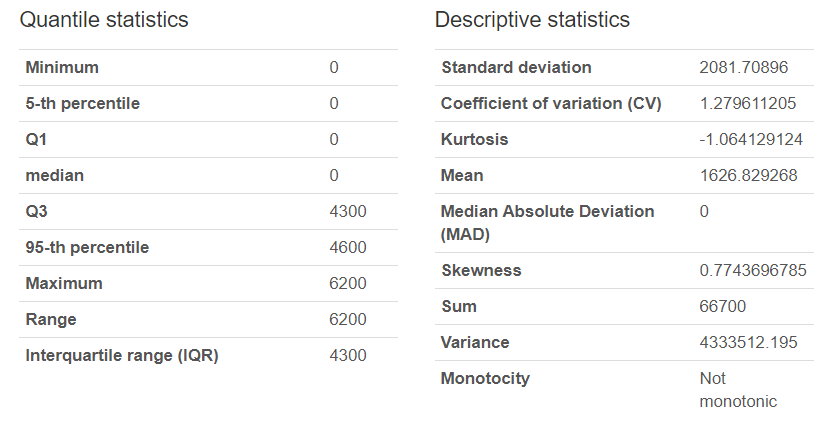
De estos datos se consideran relevantes para el análisis el nombre de la estación de peaje y las tarifas dadas por categoría vehicular, ya que definen de una forma concreta cuales son la tarifas por cada vehículo de cada uno de los peajes del INVIAS en el 2019. Lo anterior, se realiza puesto que uno de los análisis OLAP requeridos a investigar por el proyecto es: ***identificar cuáles son los peajes más costosos por categoría de vehículo.***

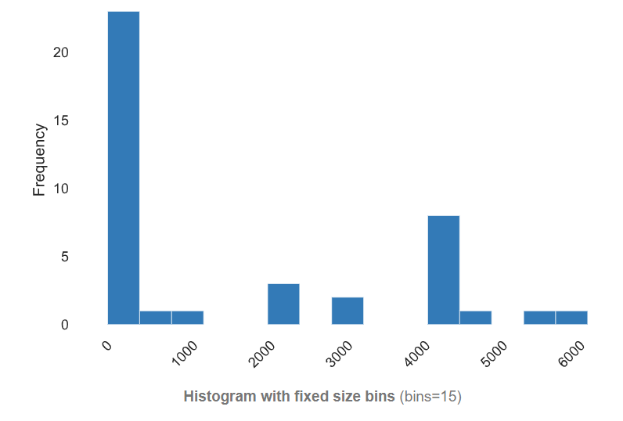
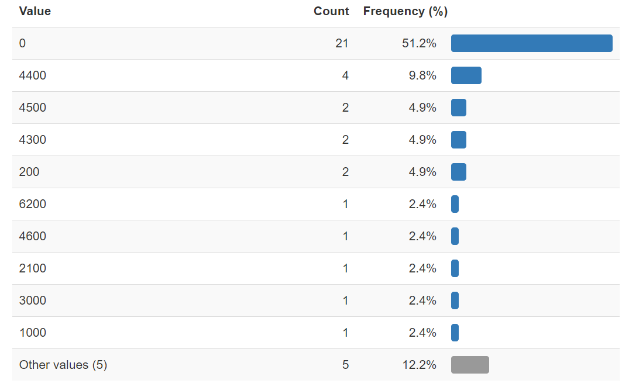
|  |  |  |
| --- | --- | --- |
| **Categoría vehicular** | **Descripción** | **Tipo variable** |
| I | Automóviles, camperos y camionetas | Numérica |
| IE | Vehículos tarifa especial Categoría I | Numérica |
| II | Buses, busetas, microbuses con eje trasero de doble llanta y camiones de dos ejes | Numérica |
| 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 |

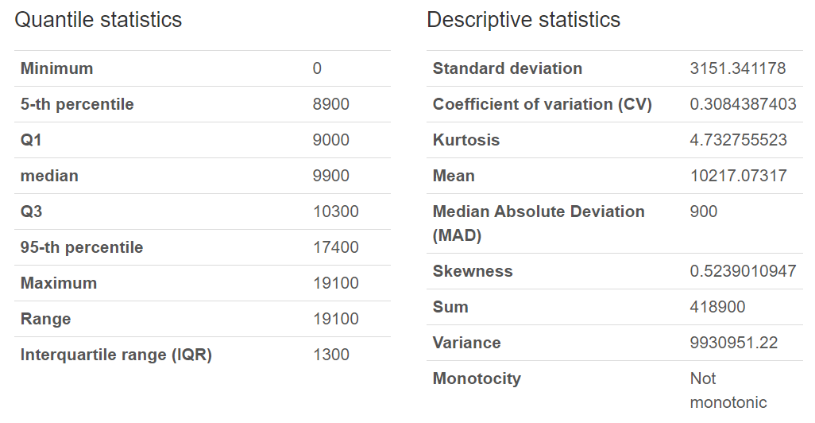
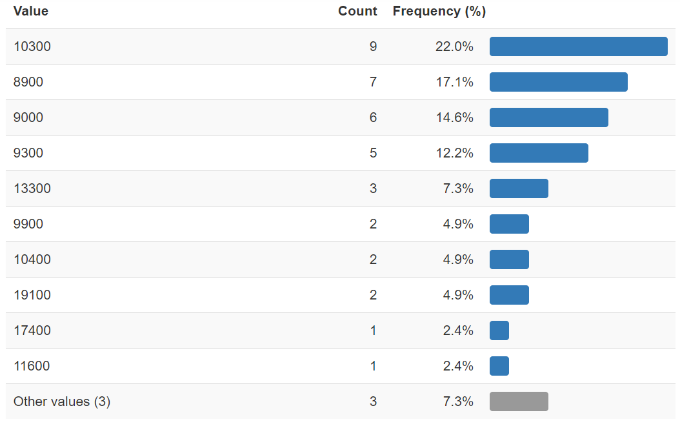
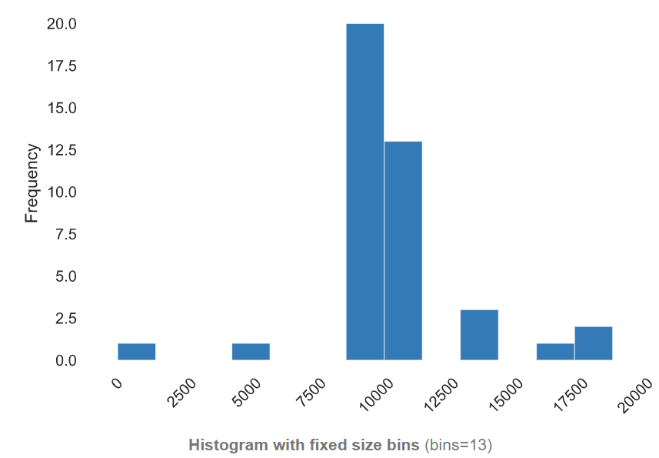
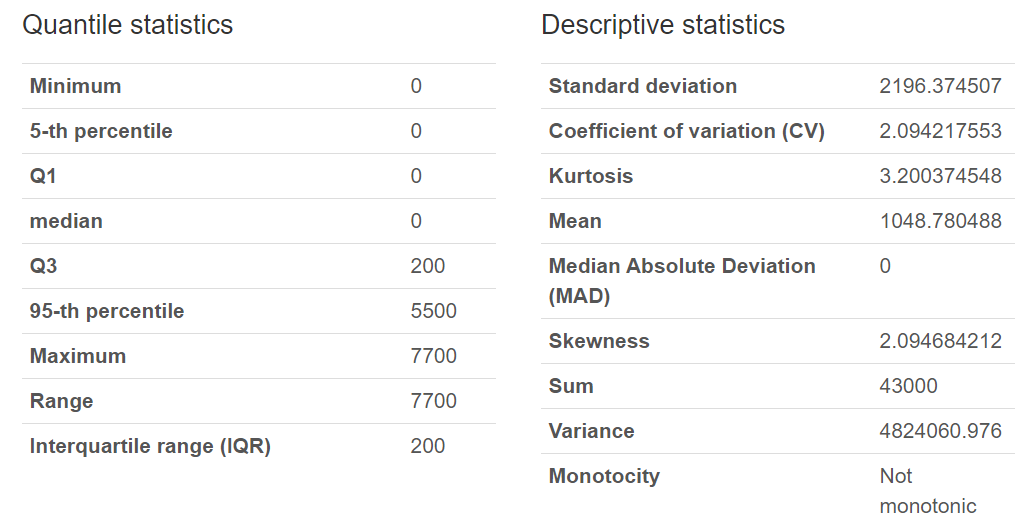
A continuación, vamos a realizar un resumen de las estadísticas descriptivas para las variables numéricas y categóricas de las tarifas por categoría vehicular, teniendo en cuenta que existen múltiples clasificaciones se realizara un análisis por cada categoría con el fin de evitar confusiones. Este análisis es desarrollado con la herramienta *pandas\_profiling* de Python.

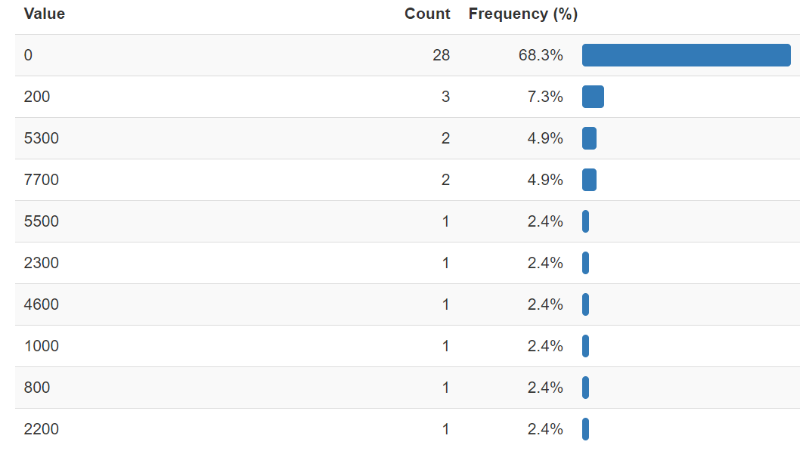
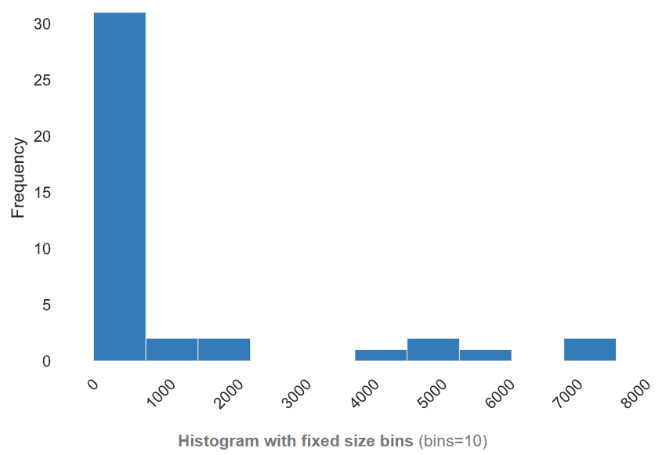
* **Categoría vehicular I:**
  + Estadísticas descriptivas**:**
  + Histogramas de frecuencia de precio de tarifas de categoría I en los peajes:



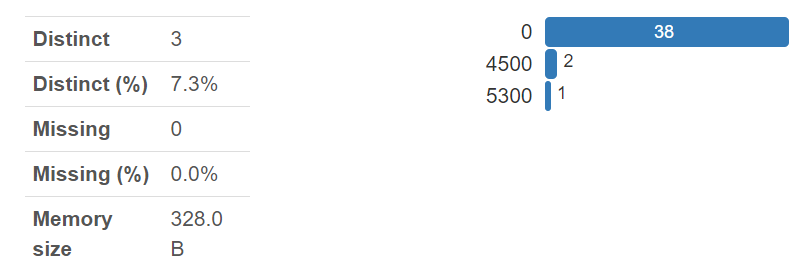
* **Categoría vehicular IE:**
  + Estadísticas descriptivas**:**
  + Histogramas de frecuencia de precio de tarifas de categoría IE en los peajes:

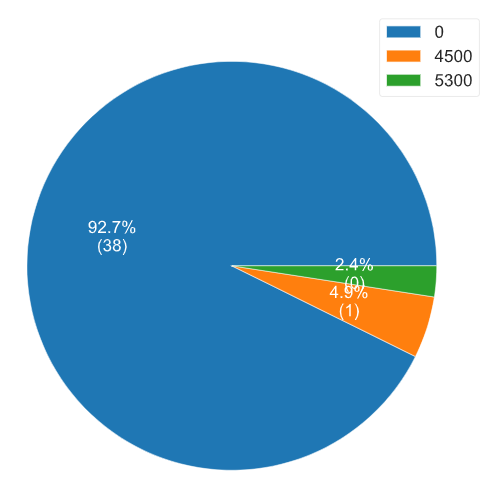


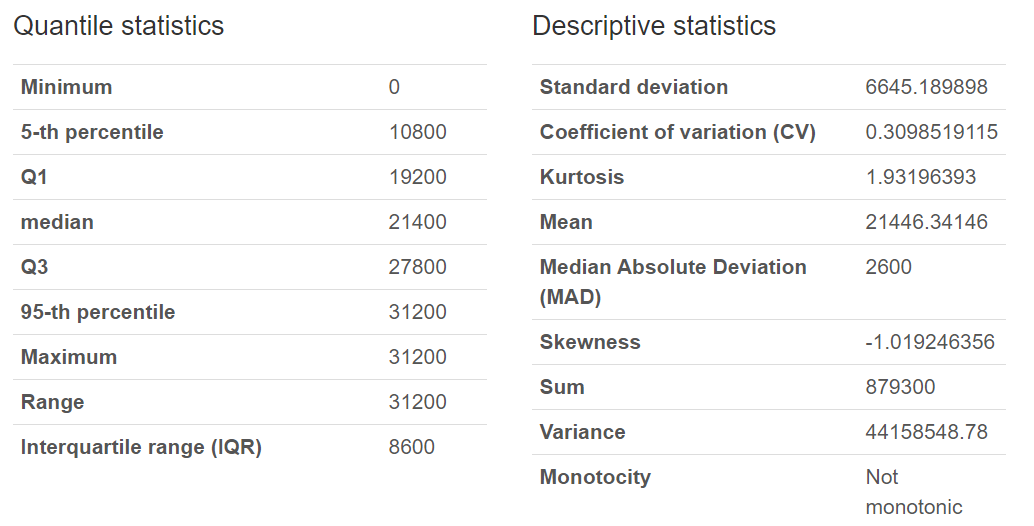
* **Categoría vehicular II:**
  + Estadísticas descriptivas**:**
  + Histogramas de frecuencia de precio de tarifas de categoría II en los peajes:
* **Categoría vehicular IIE:**
  + Estadísticas descriptivas**:**
  + Histogramas de frecuencia de precio de tarifas de categoría IIE en los peajes:



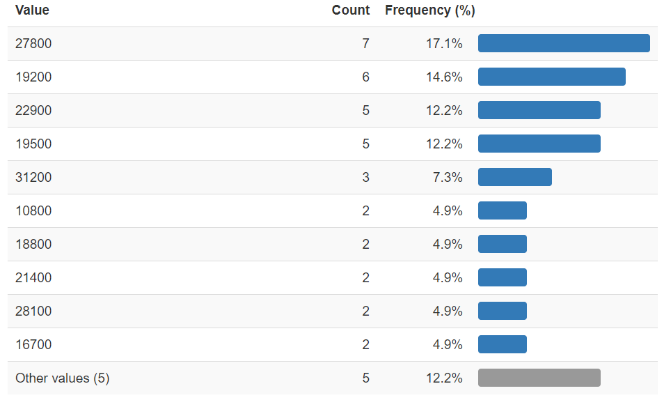
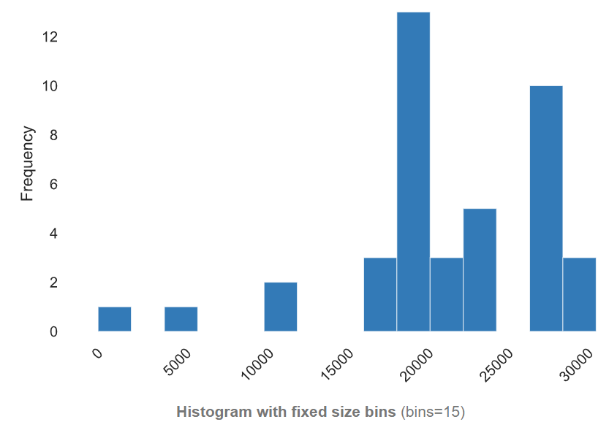
* **Categoría vehicular IIEE:**
  + Número de registros y registros faltantes:

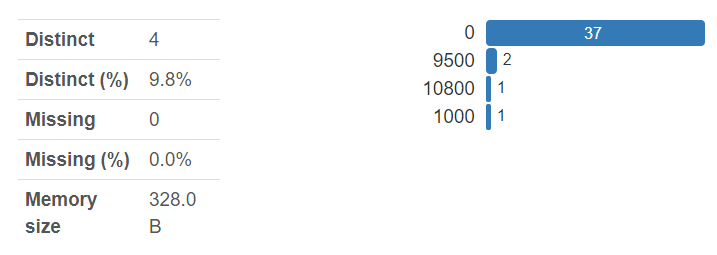
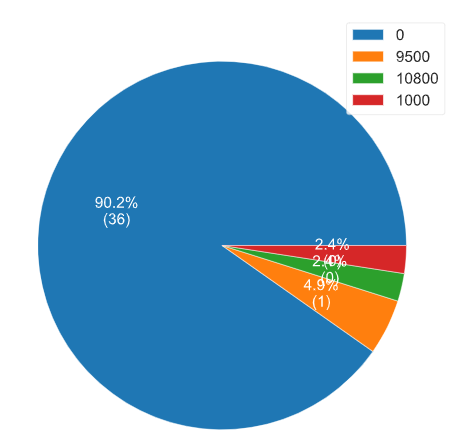


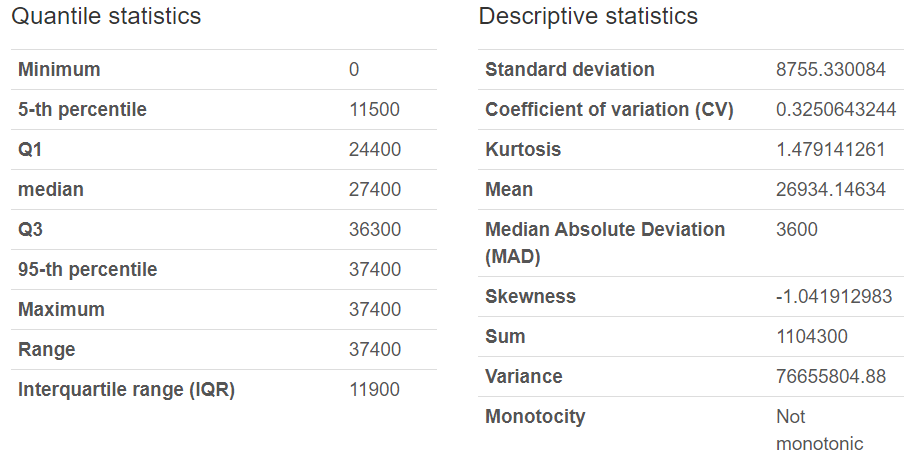
* + Diagrama de pie:
* **Categoría vehicular III:**
  + Estadísticas descriptivas**:**

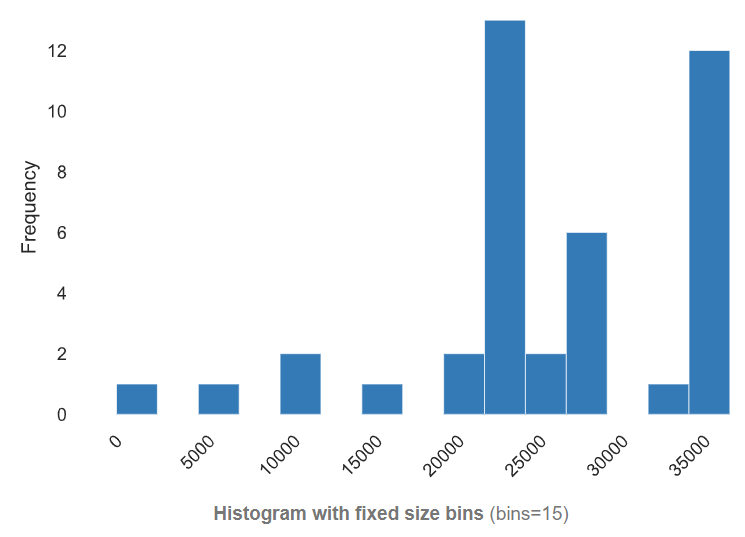
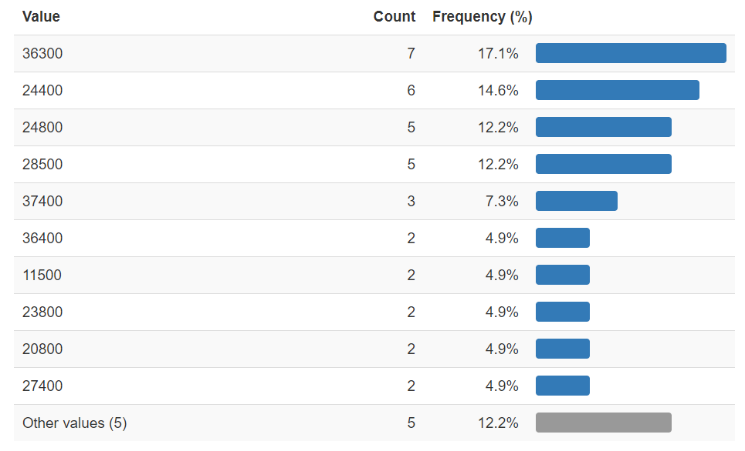
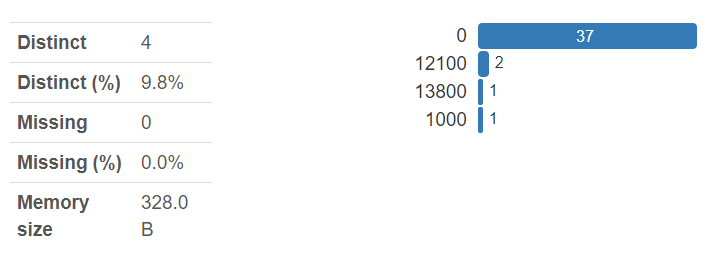
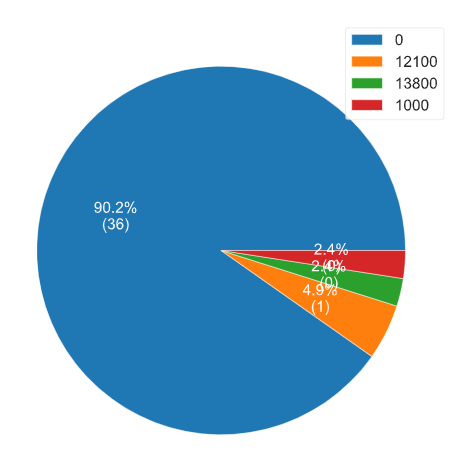
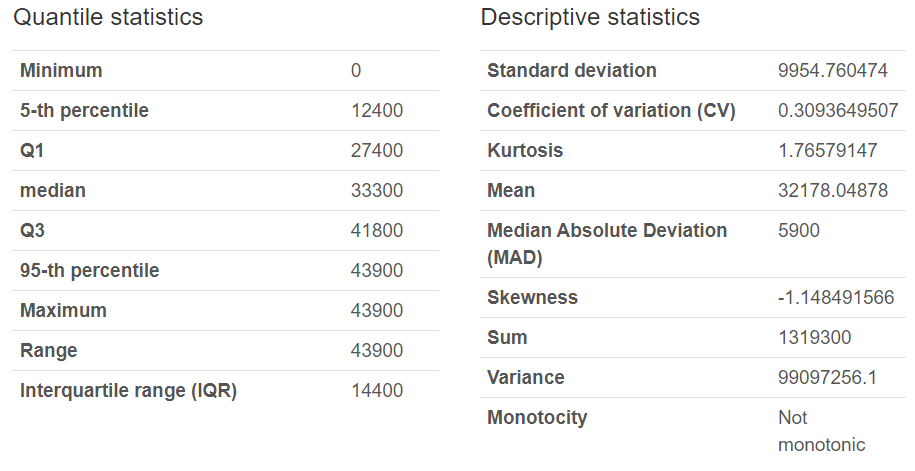


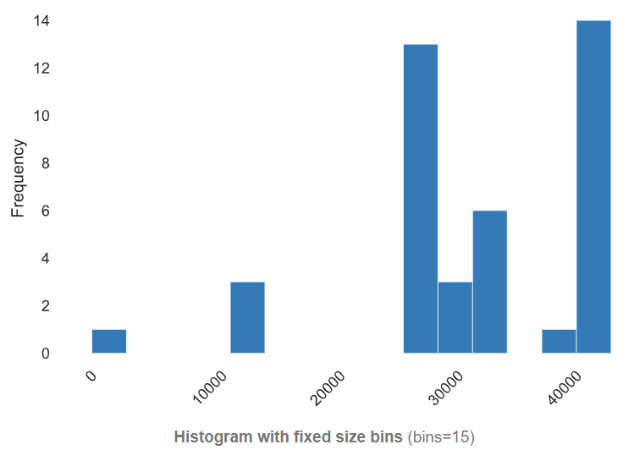
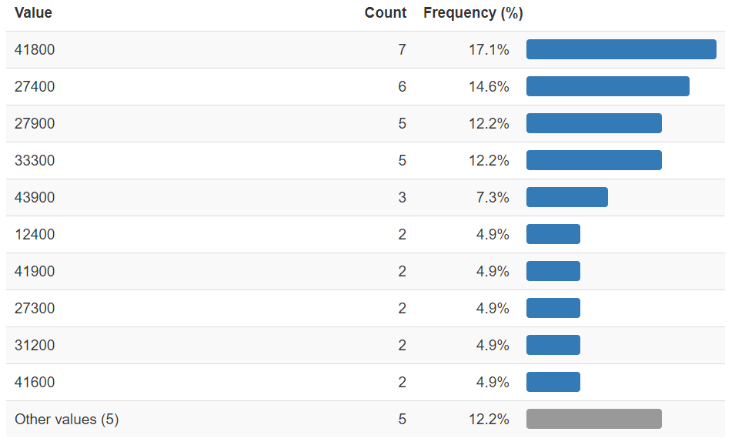
* + Histograma de frecuencia de precio de tarifas de categoría III en los peajes:

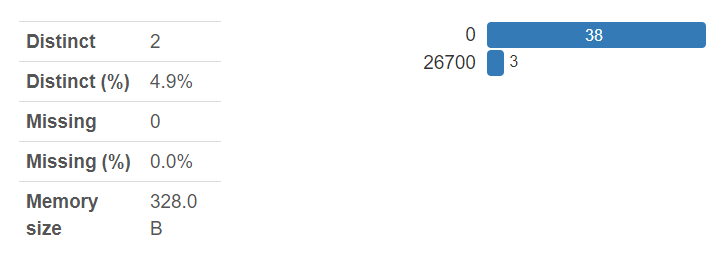
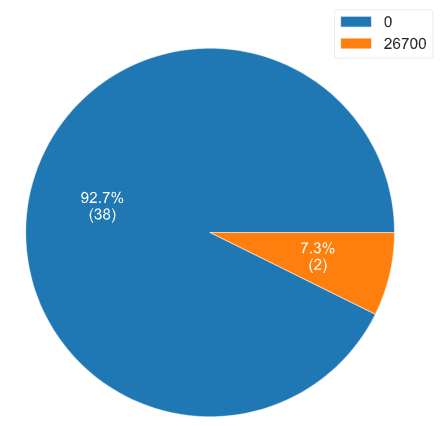
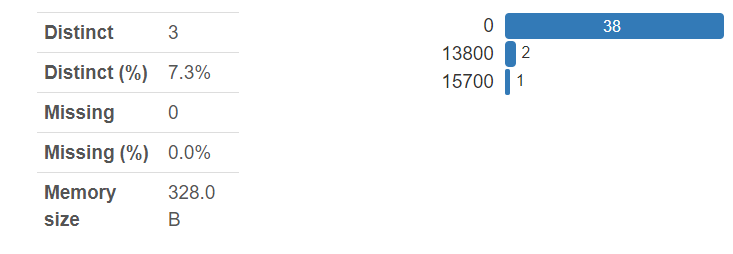


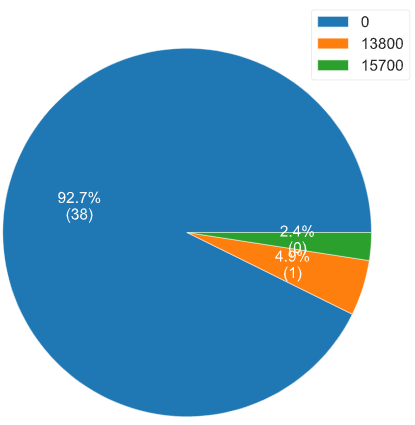
* **Categoría vehicular IIIE:**
  +  Número de registros y registros faltantes:
  + Diagrama de pie:
* **Categoría vehicular IV:**
  + Estadísticas descriptivas**:**

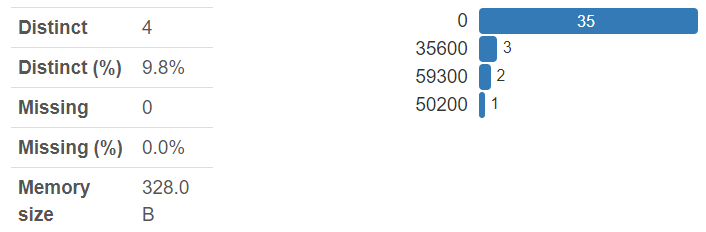
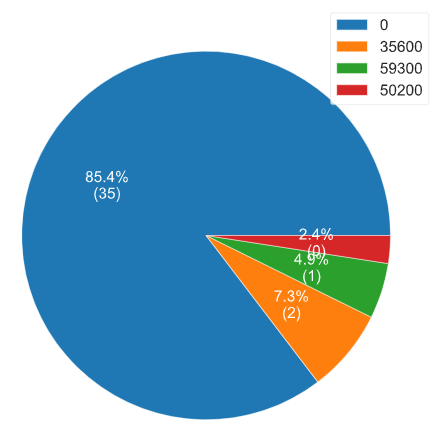
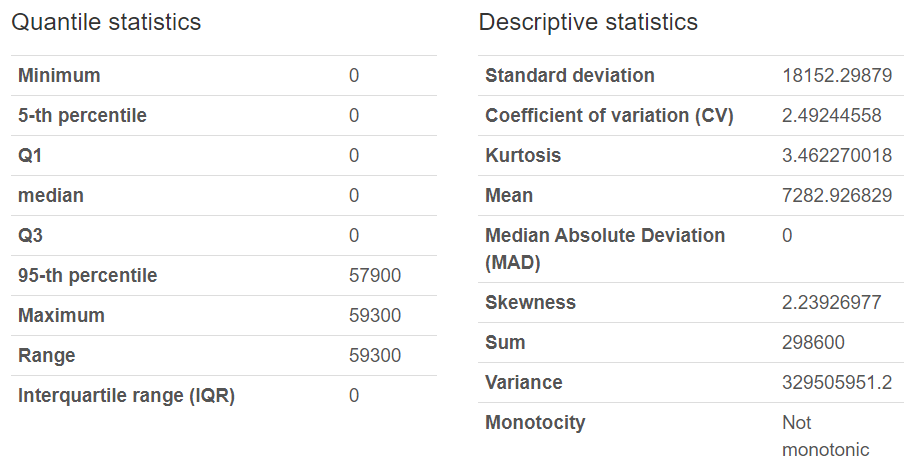


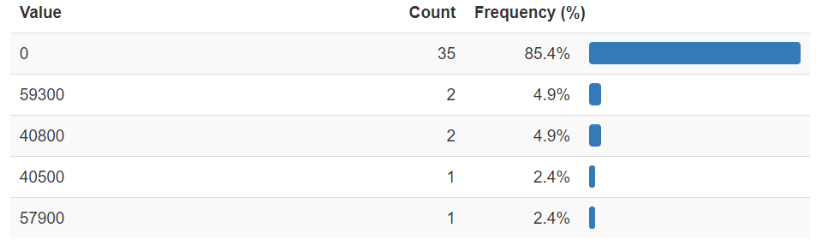
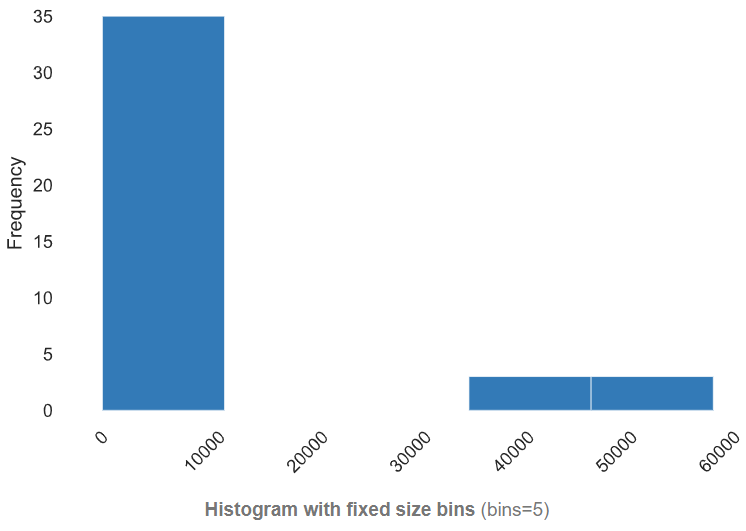
* + Histograma de frecuencia de precio de tarifas de categoría IV en los peajes:
* **Categoría vehicular IVE:**
  + Número de registros y registros faltantes:
  + Diagrama de pie:
* **Categoría vehicular V:**
  + Estadísticas descriptivas**:**
  + Histograma de frecuencia de precio de tarifas de categoría V en los peajes:



* **Categoría vehicular VB:**
  + Número de registros y registros faltantes:
  + Diagrama de pie:
* **Categoría vehicular VE:**
  + Número de registros y registros faltantes:
  + Diagrama de pie:



* **Categoría vehicular VI:**
  + Número de registros y registros faltantes:
  + Diagrama de pie:
* **Categoría vehicular VII:** 
  + Estadísticas descriptivas**:**
  + Histograma de frecuencia de precio de tarifas de categoría I en los peajes:



**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.

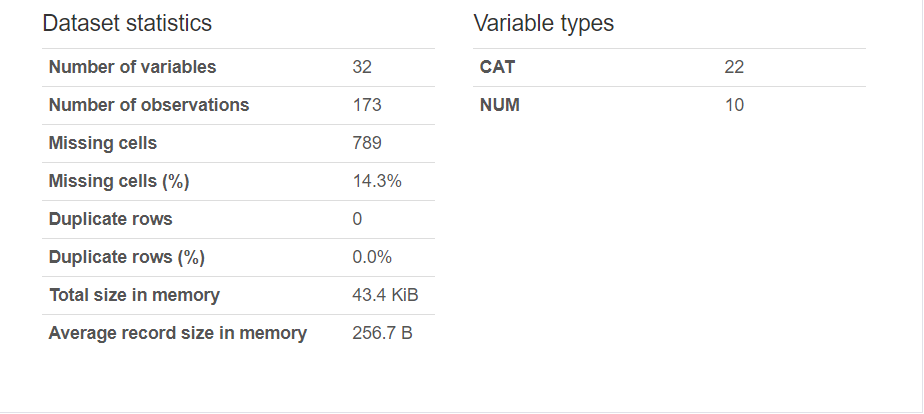
Para analizar la información suministrada y poder tomar decisiones sobre esta es importante conocer qué información es relevante para el negocio. En este caso, es interesante determinar los recaudos por peaje diferenciando las categorías de los vehículos, las rutas, el año y el departamento. Es por esto,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

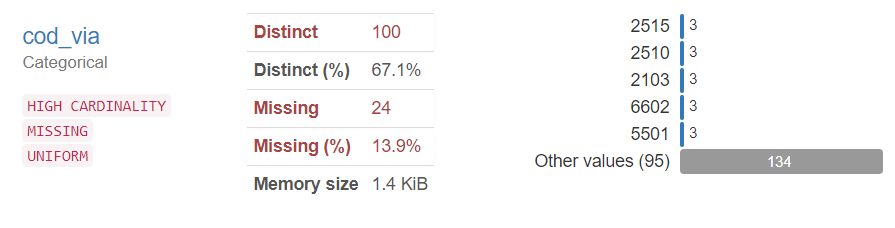
|  |  |  |
| --- | --- | --- |
| **Nombre** | **Definición** | **Tipo** |
| Cod\_via | Código de la vía | ID, Nominal |
| Nombre | Nombre del peaje | Nominal |
| Cat\_i | Valor de peaje para la categoría i | Numérica |
| Sector | Secto en el que se encuentra el peaje | Nominal |
| Ubicacion | Ubicación 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.

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

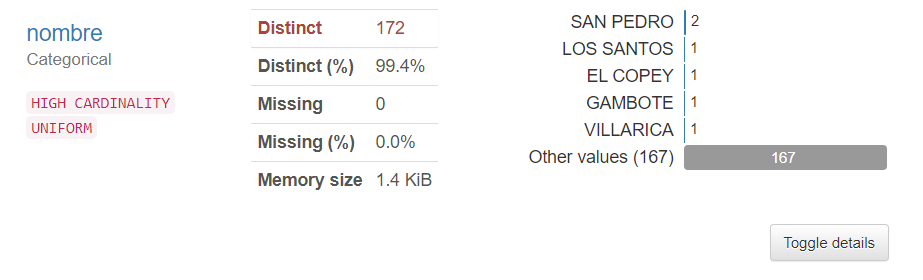


**Código de la vía:**

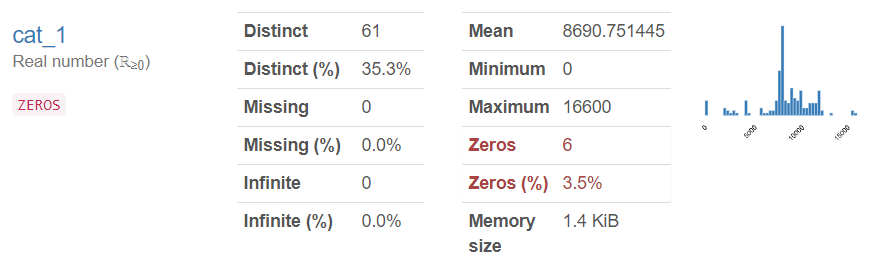


* Se resalta que con esto se hallarán las rutas para el análisis OLAP.

**Nombre:**

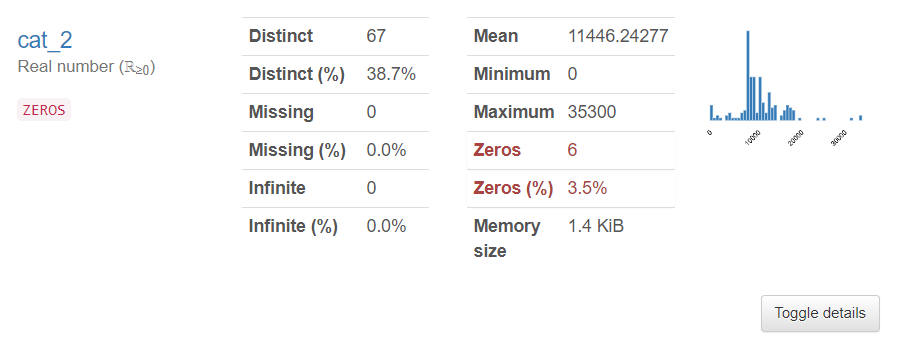


**Cat\_1:**

****

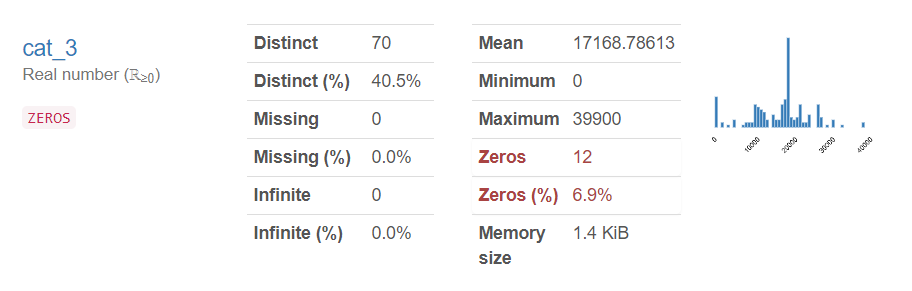
Acá se puede observar como la categoría 1 tiene una distribución con un centro observable en el que circundan los precios por lo que facilita el análisis de los valores.

**Cat\_2:**

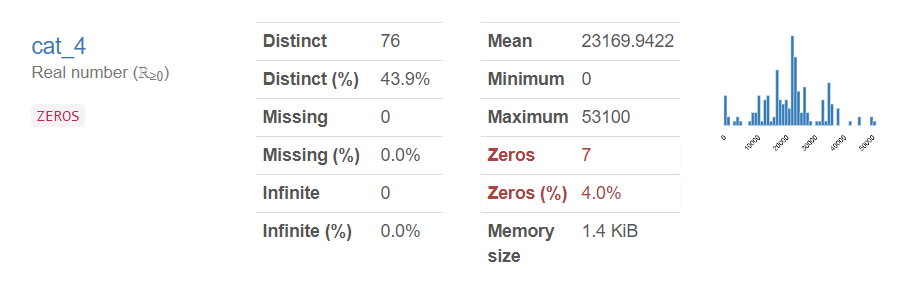
****

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.

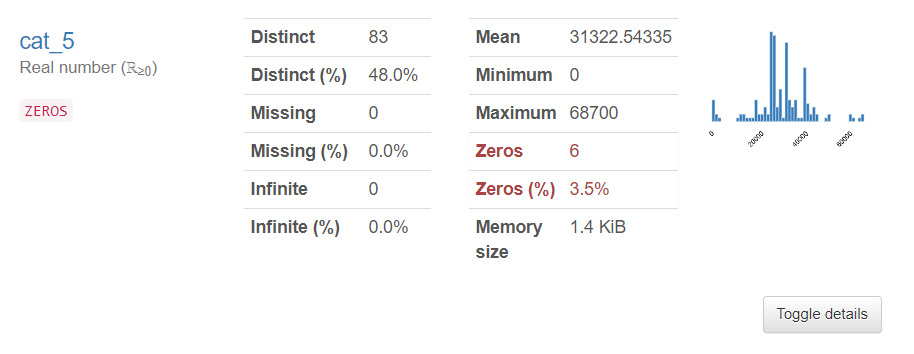
**Cat\_3:**

****

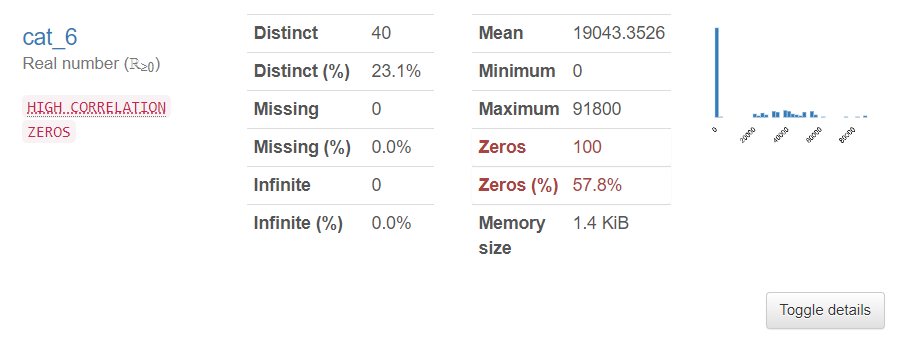
**Cat\_4:**

****

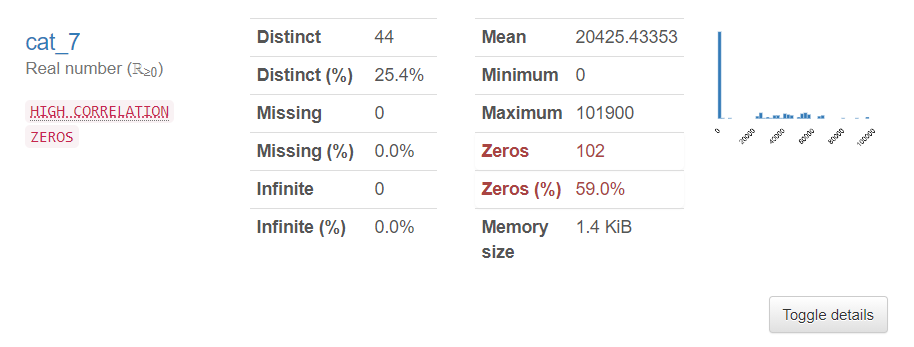
**Cat\_5:**

****

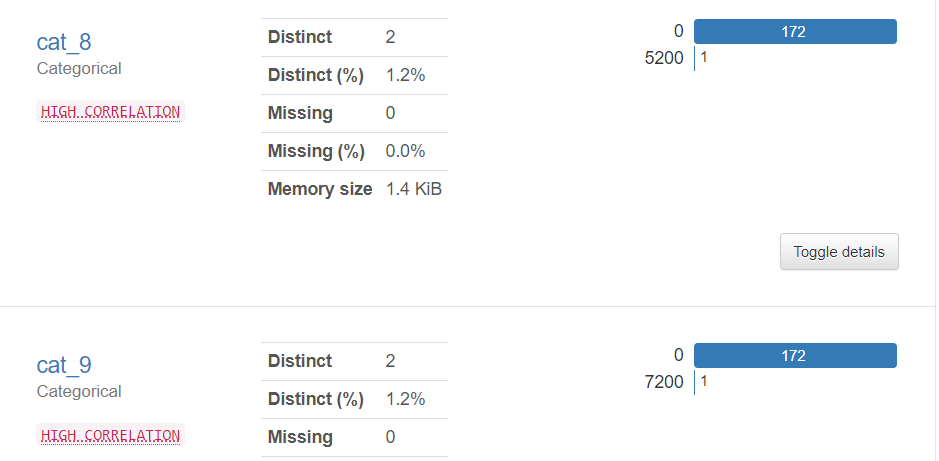
**Cat\_6:**

****

**Cat\_7:**

****

**Cat\_8 y Cat\_9:**

****

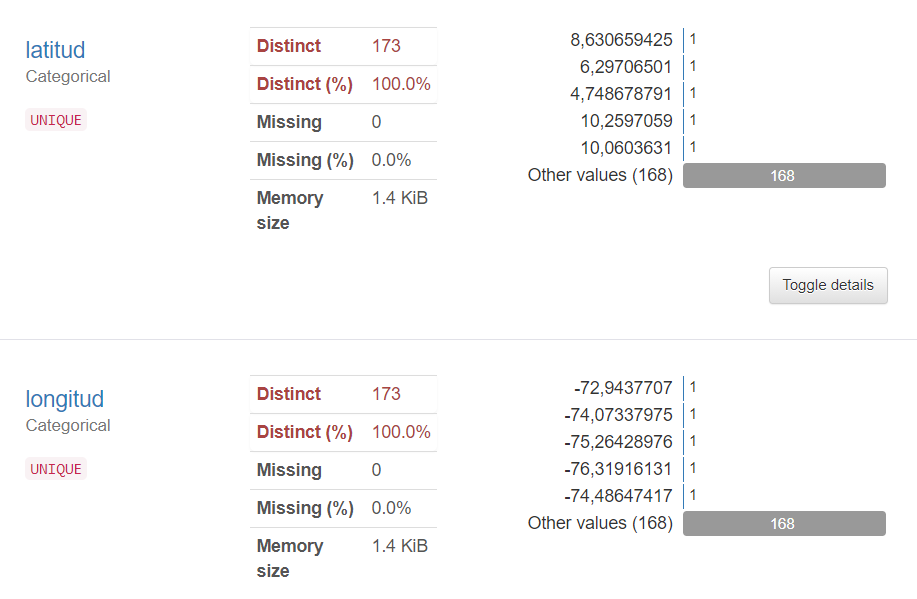
**Def\_i:**

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.

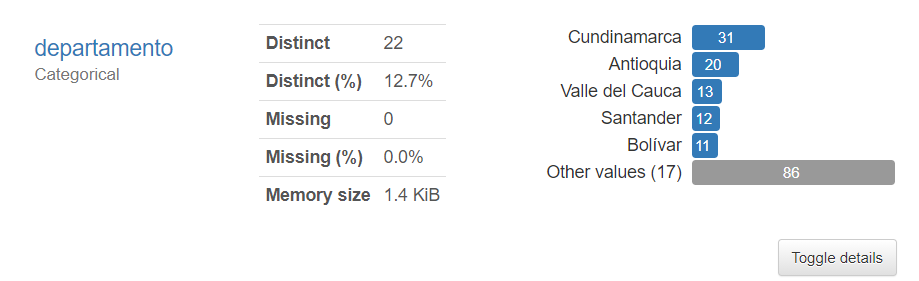
**Latitud y Longitud:**

****



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

**Departamento:**

****

**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.
* La herramienta *panda\_profiling* puede convertir datos aparentemente numéricos en categóricos si identifica que no existen muchos valores diferentes entre los registros.