Team 21: Área Metropolitana del Valle de Aburrá Transportation Demand – Exploratory Data Analysis

Features Description

Features Description: Travels Data

The dataset comes from the records made by the devices installed in each of the vehicles in operation of El Valle de Aburrá public transport system and which are part of the Metropolitan Collective Public Transport Management Platform (GTPC).

Therefore the entity gave us a data set with the following columns

FIELD	**DESCRIPTION**
SECUENCIARECORRIDO	Primary key that identifies the track for a vehicle
RECORRIDOFINALIZADO	Complete / incomplete flag (S/N)
IDVEHICULO	unique identifier for a vehicle
CODIGORUTA	unique identifier for a track, each of these identifiers are related to a KML file
FECHAREGISTRO	date and time when the data was recorded
LATITUD	The latitude where passengers board/alight the vehicle*
LONGITUD	The longitude where passengers board/alight the vehicle*
SUBENDELANTERA	Quantity of passengers that board the vehicle through the front door
SUBENTRASERA	Quantity of passengers that board the vehicle through the back door
BAJANDELANTERA	Quantity of passengers that alight the vehicle through the front door
BAJANTRASERA	Quantity of passengers that alight the vehicle through the back door

1. SECUENCIARECORRIDO

Datatype: Integer type number.

Meaning: It is a unique identifier of the trip of a public transport vehicle on a date and on a certain route.

Values taken: The identifier is generated automatically with each new trip.

Utility: By uniquely identifying the trip of a vehicle, it allows obtaining information such as the start and end date of a tour and the number of events where there was passenger movement for any trip recorded in the time horizon of the dataset. Counting the unique values of the variable returns the total recorded trips.

There are no null values. It is observed that 2% (501.203) of the records contain unique id's for "*recorrido*". The most frequent id is 134795181.

2. RECORRIDOFINALIZADO

Datatype: Single character text.

Meaning: a binary variable that indicates if the vehicle's journey or route ended, that is, if it completed the route, or did not.

Values taken: S if the vehicle tour was completed successfully, N otherwise.

Utility: allows a trip to be classified as successful or unsuccessful during the observation time recorded in the dataset. A metric such as the rate of incomplete trips per route, for a certain transport vehicle or per period of time could be constructed.

There are no missing values; 93% of the records correspond to completed routes. Observing unique routes we can say that 88% of those correspond to completed routes and 12% remaining to unfinished routes

3. IDVEHICULO

Datatype: Integer type number.

Meaning: is an unique identifier for each vehicle that operates on the public transportation network whose trips were recorded in the dataset.

Values taken: the number is an internal code of the entity to identify each vehicle in operation.

Utility: It would allow obtaining the characteristics of each vehicle in an eventual crossing of the dataset with another that contains important attributes such as number of seats, number

of foot passengers, maximum capacity, incidents due to mechanical failures, among others. It also allows calculating the number of vehicles in operation on a certain route or section of route or during a certain period of time, that is, the frequency of vehicles.

There are no missing values. The most frequent vehicle is the id 6106. We have 2371 unique vehicles

4. CODIGORUTA

Datatype: Integer type number.

Meaning: Code that identifies the route for transmission to the platform.

Values taken: Unique identifier for each route assigned by the entity.

Utility: It is the analogous to the variable IDVEHICULO. Although each route corresponds to only one vehicle and one route, a route can have different routes as well as different associated vehicles.

There are no missing values. We have 248 routes and the most frequent one is the route 2102

5. FECHAREGISTRO

Datatype: Text containing date and time in YYYY-MM-DD HH:MM:SS 24H format. Meaning: it is the record of the moment in which a passenger movement event occurs, that is, the date and time in which passengers board and alight each time the vehicle stops.

Values taken: It is a date with the format YYYY-MM-DD HH:MM:SS 24H, where,

DD: is the day of the month (0 to 31), MM: is the month of the year (1 to 12), YYYY is the year (2019 and 2020) , HH: MM are the hour (00 to 23) and the minutes (00 to 59) respectively. The dates of the events are unique per travel, vehicle and route.

Utility: It allows filtering any data registered in the dataset, variable or metric calculated for a certain instant or time interval. The minimum interval of interest is the hour and the maximum the day.

There are no missing values. We have values from 2019-11-01 00:00:03 until 2020-05-10 22:21:51, but we only have records from November - 2019 and from march to April - 2020

6. LATITUD & LONGITUD

Datatype: Float type number.

Meaning: It is the exact location in geographical coordinates of the occurrence of a passenger movement event. Latitude is the distance in degrees between any parallel and the line of the equator. Longitude is the measure of the arc between the zero meridian and the meridian of any point.

Values taken: Latitude is between 6.10 and 6.30 degrees to the north (positive values) and longitude between 75.5 and 75.6 degrees to the west (negative values).

Utility: In addition to georeferencing the events, it allows associating them to the nodes that the entity established in the entire public transport network of El Valle de Aburrá. This will be useful when calculating metrics on the number and flow of passengers in certain areas and route segments of the network.

In the **LATITUD** column, There are no missing values. However, we detected 19646 records with coordinates (0,0) and also we found 21049 possible outliers.

In the **LONGITUD** column, There are no missing values. However we detected 19646 records with coordinates (0,0) and also we found 20210 possible outliers

7. SUBENDELANTERA and SUBENTRASERA

Datatype: Integer type number.

Meaning: It counts the total number of passengers who board the bus at each stop on the route (designated by the two previous variables) through the front or rear door respectively.

Values taken: Non-negative integers. Utility: These are the variables required to calculate the passenger load by date, by arc and by route of the public transport network. The sum of the two variables returns the total number of passengers who board a vehicle. This number when filtering by any other variable, range or group of variables, returns the number of boardings per category of variables, the rate of boardings per range, among others. Therefore, it is quite useful to build performance indicators (KPIs) that describe the operation and capabilities of the public bus transport system. By grouping boardings by areas of the transportation network, those that demand more or less transportation services can be identified.

SUBENTRASERA usually occurs when the bus is overloaded, therefore, this variable could be used later, to find cases where such an event exists. Also, there are 83900 records above 3 standard deviations from the mean.

In the case of the **SUBENDELANTERA** variable there are 160010 records above 10 (above 3 standard deviation from the mean).

Both variables mentioned above need to be analyzed with more detail before taking any action.

8. BAJANDELANTERA and BAJANTRASERA

Datatype: Integer type number.

Meaning: It is the number of passengers who alight through the front / rear door of a vehicle each time it stops.

Values taken: Non-negative integers.

Utility: These are the variables required to calculate the passenger load by date, by arc and by route of the public transport network. The sum of the two variables returns the total number of passengers alighting from a vehicle. This number when filtering by any other variable, range or group of variables, returns the number of alights by category of the variable, the rate of alights by range, among others. Along with the boardings, performance indicators (KPIs) can be built. By grouping the alights in passengers by zones of the transport network, the most or less popular destinations can be identified.

In the case of the **BAJANDELANTERA** variable there are 132858 records above 3 standard deviations from the mean.

In the case of the **BAJANTRASERA** variable there are 191317 records above 3 standard deviations from the mean.

Both variables mentioned above need to be analyzed with more detail before taking any action.

NOTE: The variables **SUBENDELANTERA, SUBENTRASERA, BAJANDELANTERA, BAJANTRASERA** show a normal behavior according to their characteristics, because of that, it is normal one passenger gets on or gets off at the same time. We also discovered that the range of values is between 0 and 99 so eventually we'll discuss how to proceed in this case.

Features Description: Transportation Network and Nodes Data

According to the requirement of the entity, where the analysis of the demand of passengers of the transport network is requested, it is necessary to start from the base of the configuration of the road network on which the vehicles operate.

For this, the entity made available the data of the transport network in gis format, having the following.

Details of the nodes

- emme_nodes.cpg
- emme_nodes.dbf
- emme_nodes.prj
- emme_nodes.shx

Segment Details

- emme_tsegs.cpg
- emme_tsegs.dbf
- emme_tsegs.prj
- emme_tsegs.shx

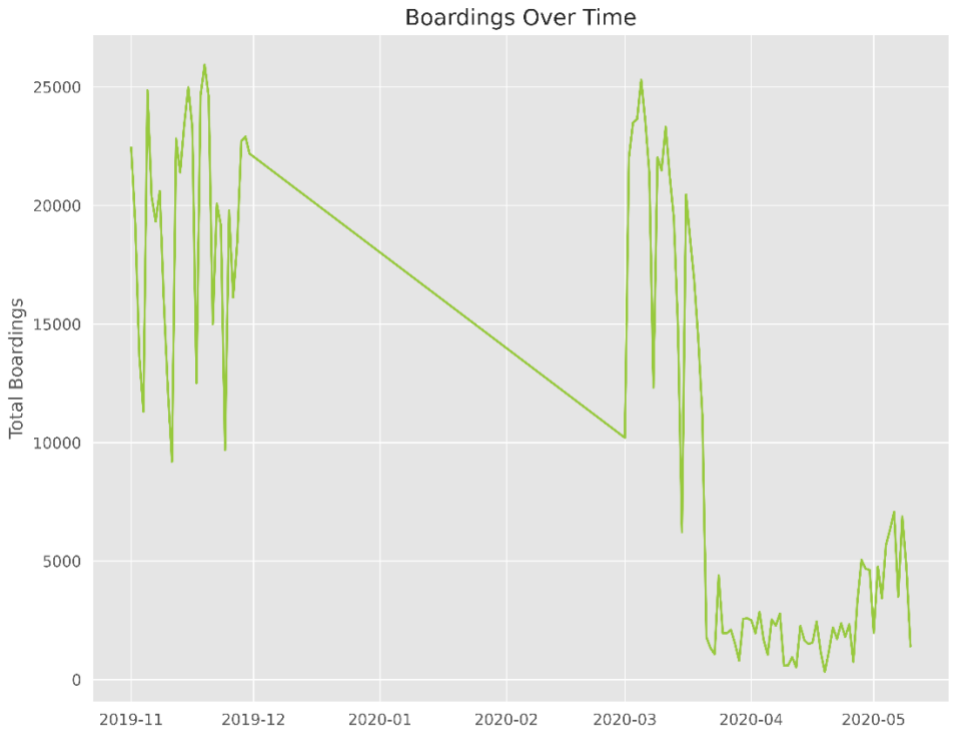
Through the library shapefile we'll explore the files to visualize the detail of the road network of the City of Medellin.

Exploratory Data Analysis

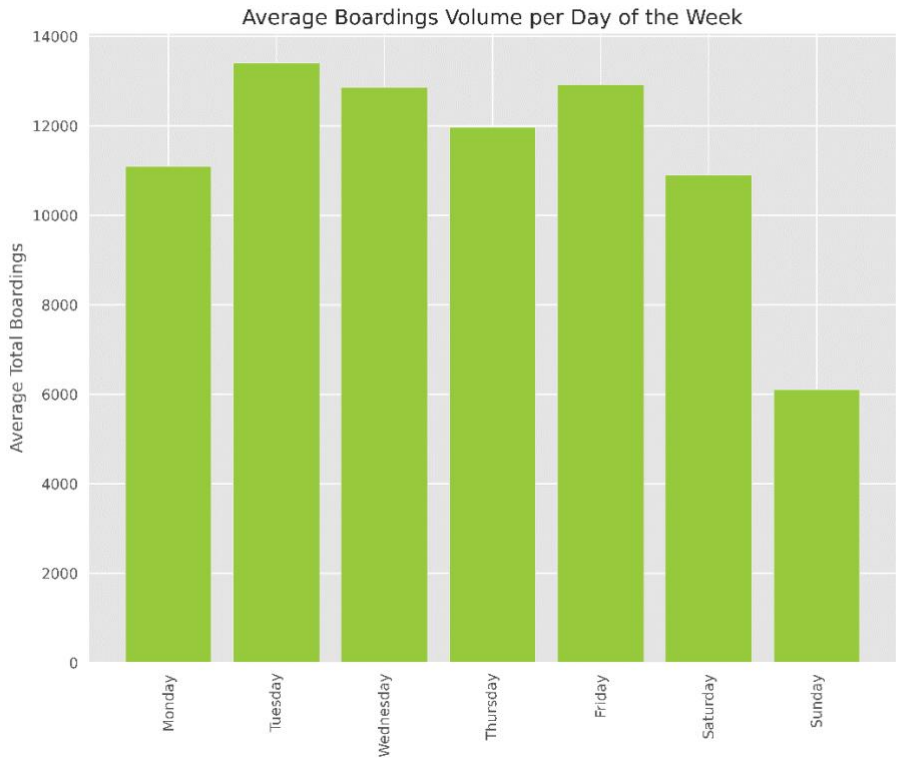
Boardings Data

Before proceeding to handle extreme data, first We will analyze how its behavior has been on the observation window, the data distribution and the most responsible approach.

First of all, for the period the data was given by the entity, we see that around the first months, November and December the data oscillated without a clear pattern; nevertheless, around the last days of November there was an increase in the passenger boardings. Around December until the first days of March, there was a decrease possibly originated in how the data was previously manipulated by the entity. Then, beginning March there was a substantial increase and an apparent seasonal pattern and a subsequent decay associated with the COVID-19 lockdown.

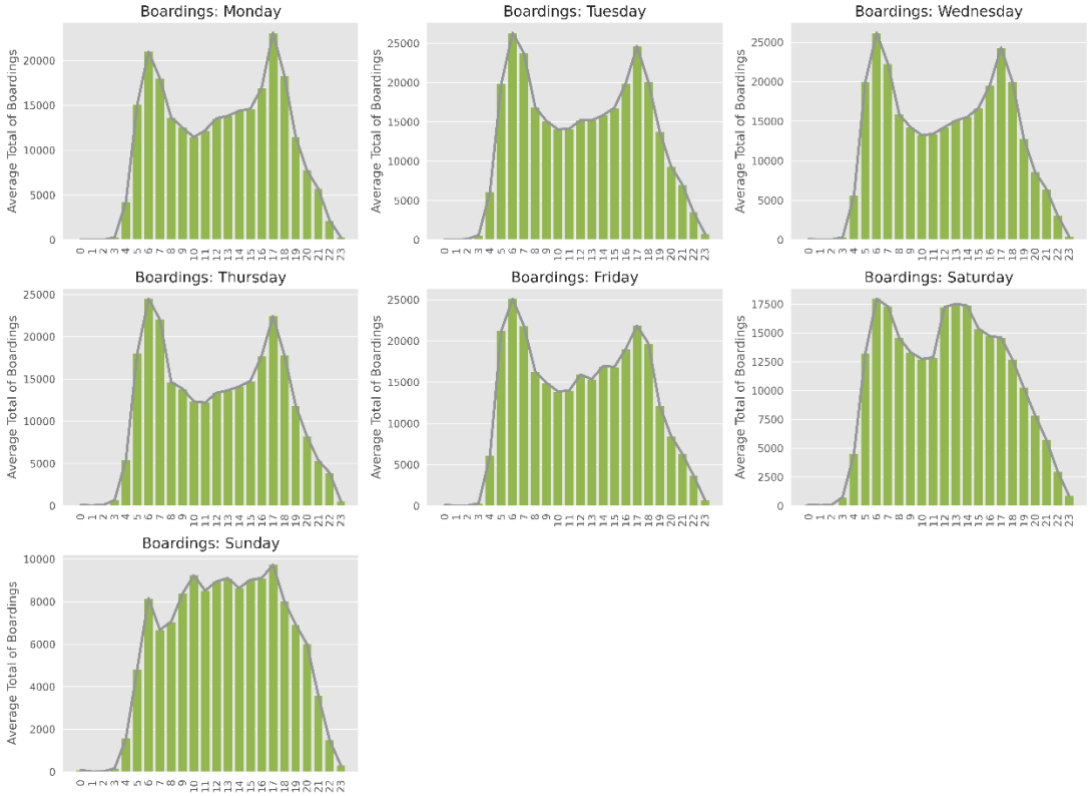


Around April, there was a modest increase and the seasonal behavior previously mentioned can be seen again reaching a peak around the second week of May.



Such mentioned seasonal behavior could be of weekly origin; as we can see at the start of the week the boardings are quite low but as we advance in time, these increase modestly reaching peaks at Tuesday and Friday. There's a valley midweek, i.e., Wednesday and Thursday and a gradual decrease from Friday till Sunday.

We already saw how the boardings behave within the week but what about the behavior throughout the day?

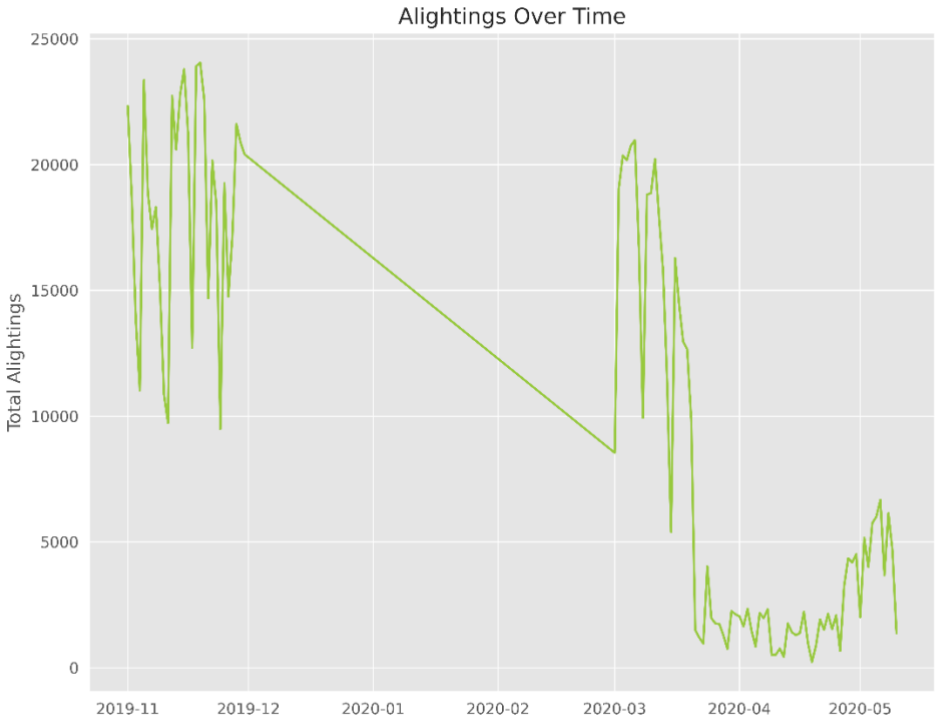


The hourly behavior is almost constant for every day of the week except that the volume of passenger boardings vary depending on each one; for weekdays the repeating pattern consists of an increasing behavior from 12 am or the 0 hours till 6 - 7 a.m reaching its first maximum and then it begins to exhibit a gradual decay which reaches its minimum at 9 - 10 a.m. Next, this tends upward till reaching its second maximum around 5 - 6 p.m. These patterns emerge as a product of the laboral schedules where people begin theirs at the first mentioned peak and end it at the second one, returning home by then.

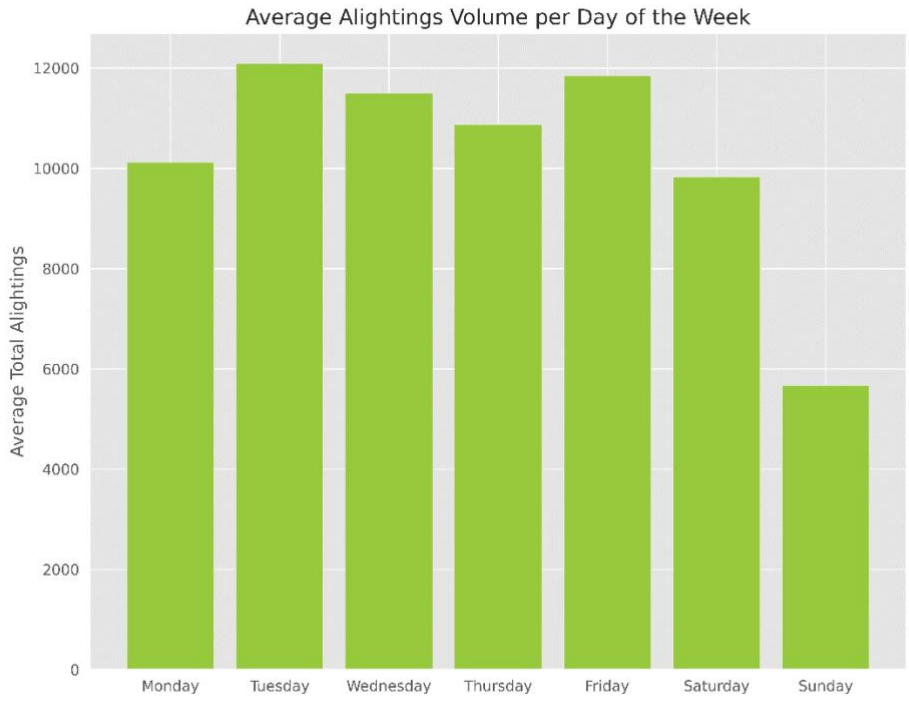
For Saturday the pattern is quite similar but the valley is more pronounced and the gap is smaller such that minimum is reached faster than for the weekdays. Again, the maximum is reached around 6 - 7 a.m. and the minimum, around 10 a.m. The increase is faster and the second maximum is reached around 12 p.m. This again, could be explained as a product of the laboral schedule which goes from 8 a.m. until 12 p.m. on Saturdays. After reaching its second maximum, the number of boardings descends till midnight where it reaches a minimum.

On Sundays, there are constant peaks from 6 a.m. every two hours approximately till 6 p.m. where there's a maximum of passenger boardings. These peaks could be associated with outdoors activities such as going to the church, lunching outside among others.

Alightings Data

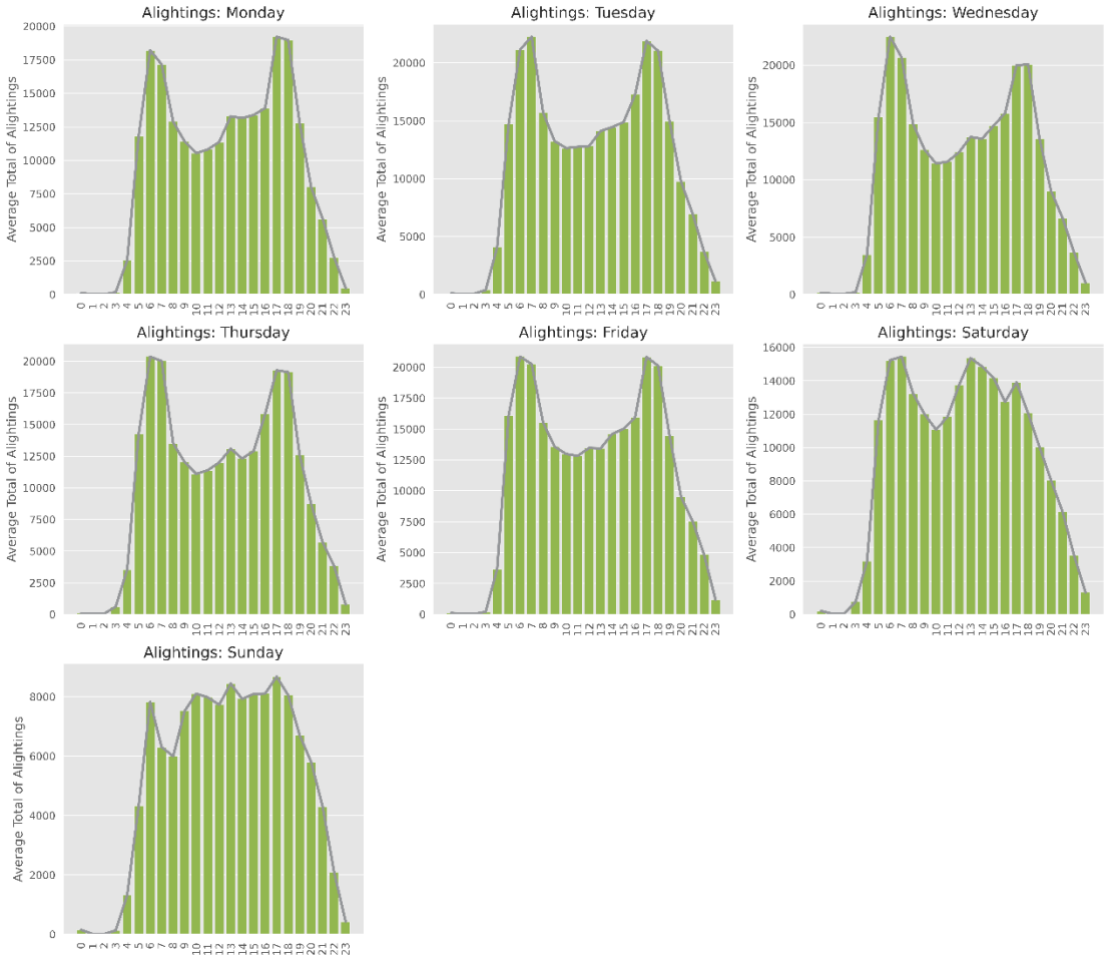


The temporal pattern for the alightings is quite similar to that of the boardings with maximums at the beginning of the observation window (November approximately) and a decay till the start of March. Then, there is an increase reaching a maximum around half March and an eventual decay till the beginning of April, again, possibly because of the COVID-19 lockdown. There exists a modest increase from mid April till May where again, a Maximum is observed. Also it can be observed a seasonal pattern which could be of weekly origin as we will inspect next.



Weekly, the pattern for the alightings is almost the same as the one of the boardings with two modes for Tuesday and Friday each and a small valley around midweek, i.e., Wednesday and Thursday. After Friday the alighting number descends till Sunday and beginning the week these are low in contrast with the rest of the week.

Now, we proceed as with the Boardings to examine how the alightings vary throughout the day,

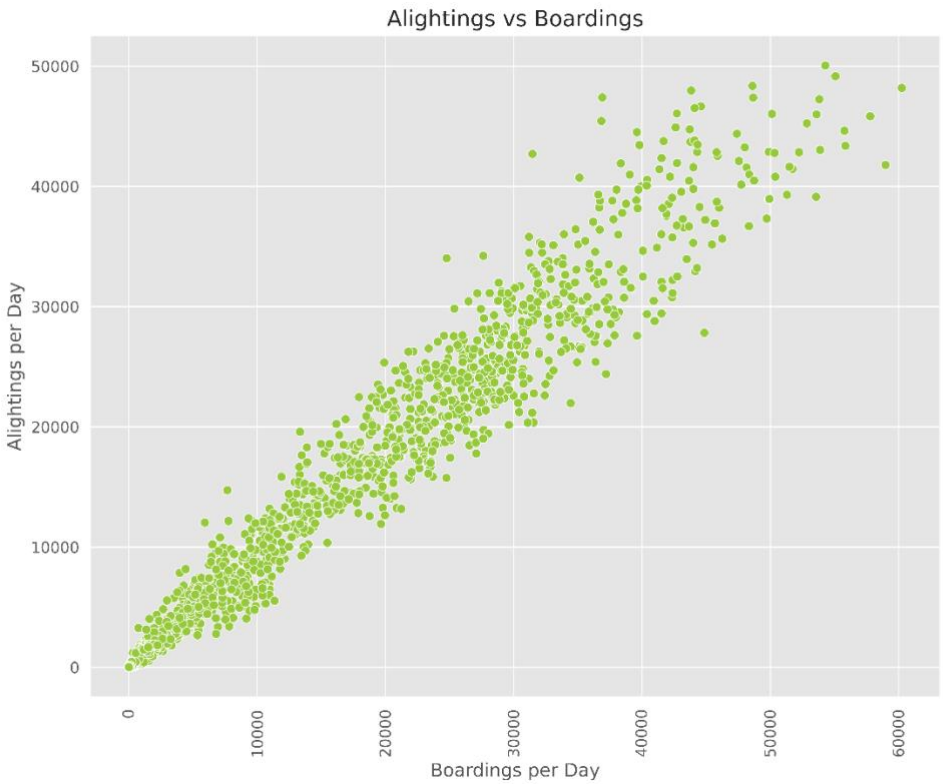


The pattern for the alightings is almost similar to the boardings with a bimodal distribution throughout the day where there's a maximum at 6 - 7 a.m. and another at 5 - 6 p.m for weekdays. As for the case of the boardings these two peaks are associated with the laboral schedules such that around them the passenger flux is greater than the rest of the day, this includes both taking the bus to get home or work and the corresponding alighting. Again, on Saturday the gap between maximus is tighter as the laboral schedule is shorter than for the weekdays going from 8 a.m. to 12 p.m. most of the cases where these two hours are the ones where the maximums are reached. Again there's a descense in the alightings values as the day goes by till reaching a minimum at 11 p.m.

On Sundays, the multimodal pattern is repeated reaching a maximum almost every two hours probably again, associated with the corresponding boardings.

Now that we have seen how the boardings and alightings behave each one separately, it could be of interest to see if they correspond to each other, that is, we would expect that the same quantity of passengers that board a vehicle to be the same which alights from it.

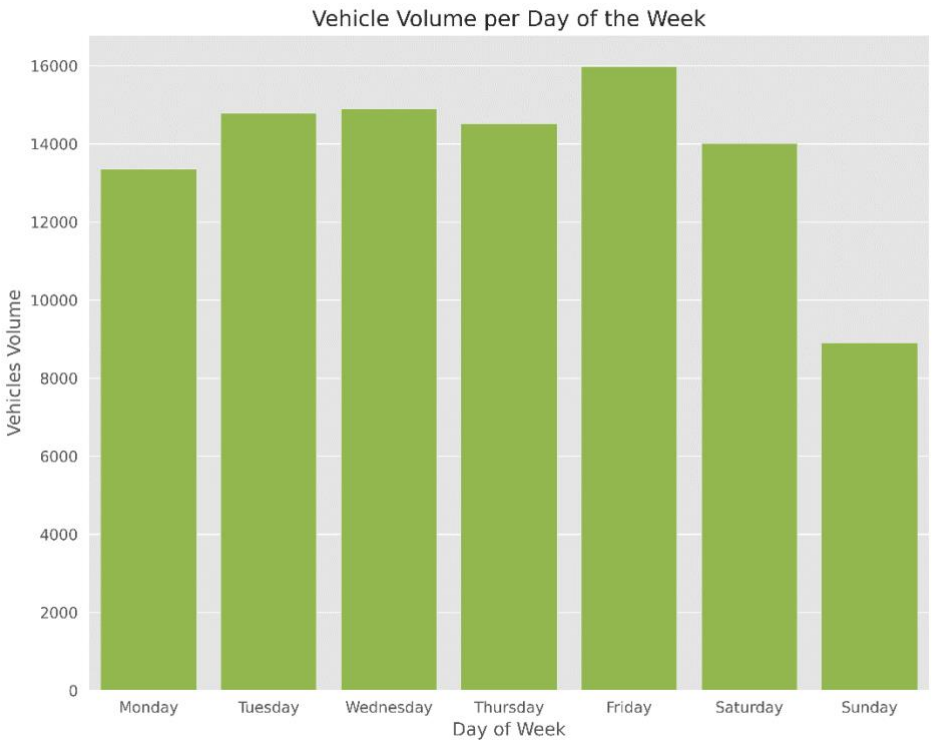
Through a scatterplot it can be easily corroborated



As stated before, the relationship between boardings and alights is perfectly linear such that the same quantity that boards the vehicle is the same which alights from it. Furthermore, this can be corroborated using the correlation coefficient, which in this case was of 98.30%.

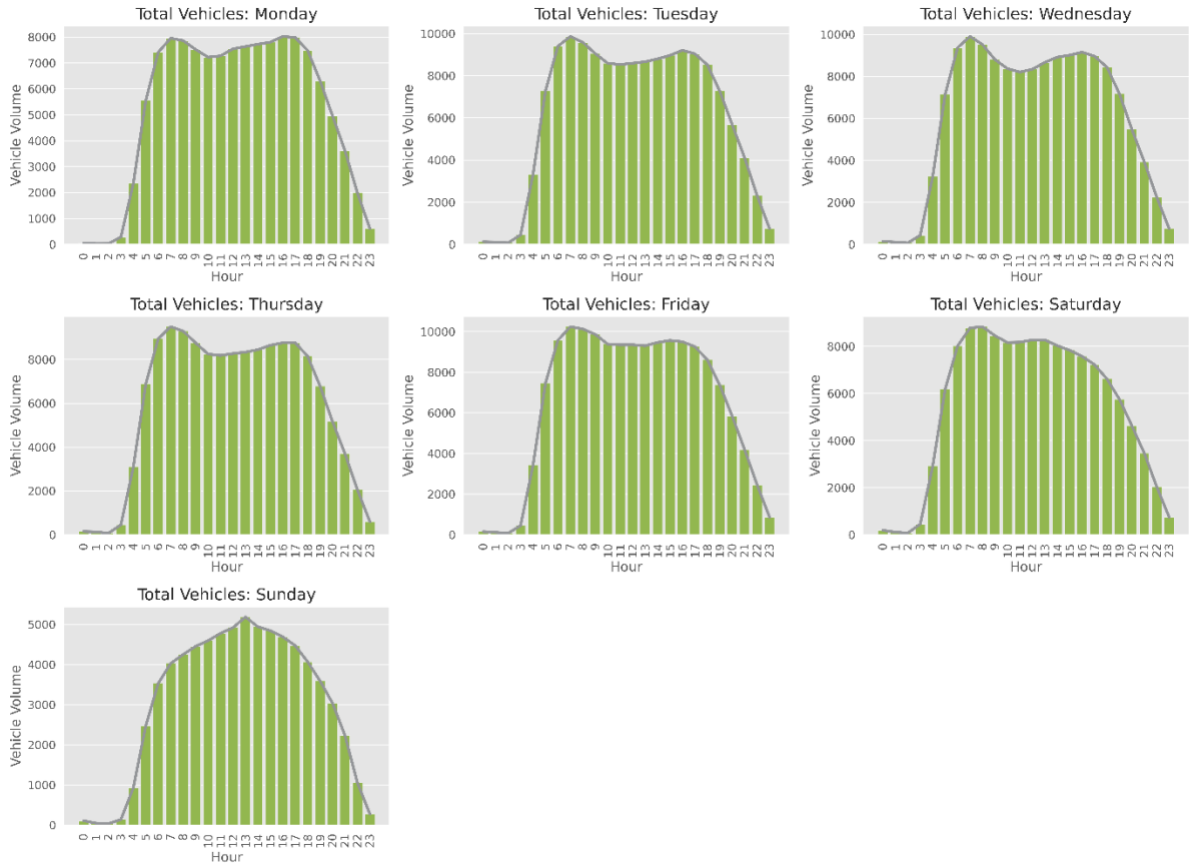
Vehicles Data

Overall, in the transportation network there were 2371vehicles. Let's proceed to see how these behave regarding time



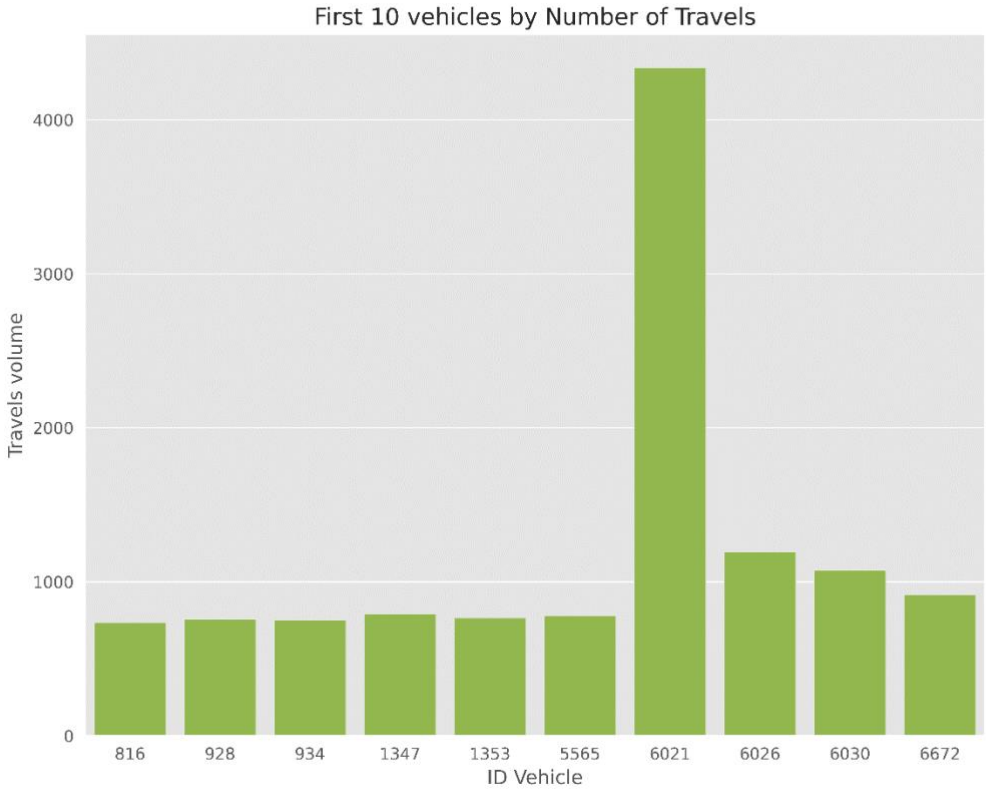
The vehicles volume behavior per day of the week is quite similar to those of the boardings and alightings reaching a maximum of vehicles in circulation on Friday and two other maximums on Tuesday and Wednesday. After the maximum is reached there's a gradual decrease till sunday corresponding to the lowest passenger volume along the week.

Now, regarding the hour of the day, the vehicle volume behaves as follows



Where there's a bimodal pattern from Monday to Friday similar to those of the boardings and the alightings around the hours where there's a maximum flow of passengers in the area; that is, around 6 - 7 a.m. and 5 - 6 p.m. On saturday, there's a maximum around 6 - 7 a.m. as for the weekdays but the behavior for the rest of the day doesn't vary too much reaching a modest second maximum around noon as for the case of passenger volume (boardings and alightings).

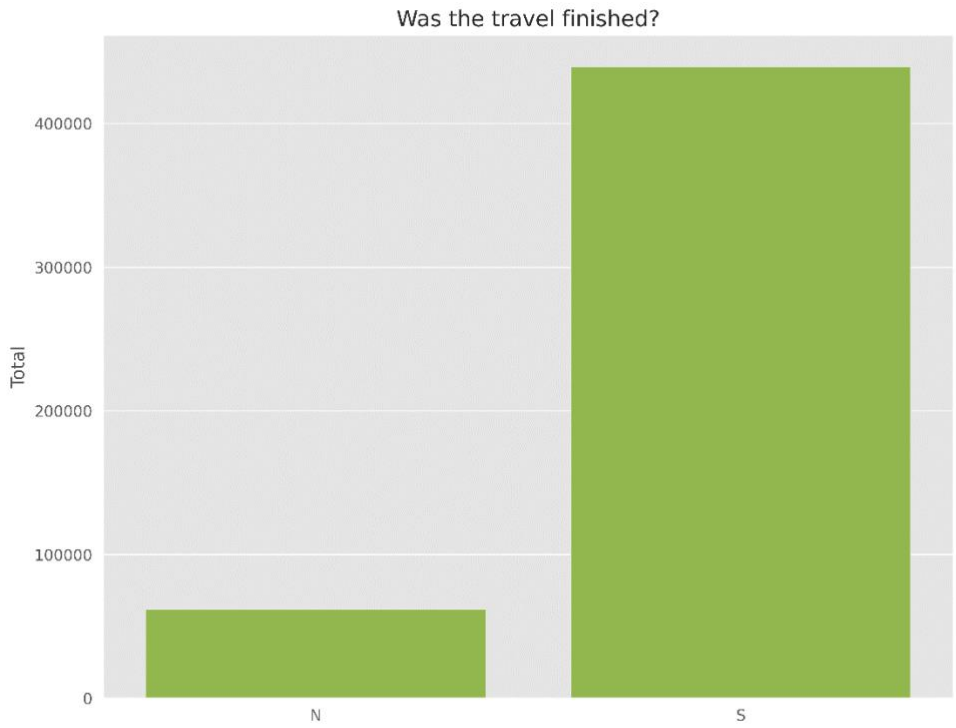
During Sunday there's a unimodal distribution reaching its maximum at 1 p.m. confirming the suspicion that around this hour most the people in the city go outside, probably tolunch.



The travels are mostly made by the vehicle with ID 6021 making around 4000 travels throughout the window observation; the next one in order of realised travels is the 6026 with over 1000 travels and then, the 6030 with around 1000 travels. From this we could infer that those with an ID around 6000 are the ones that make the most travels.

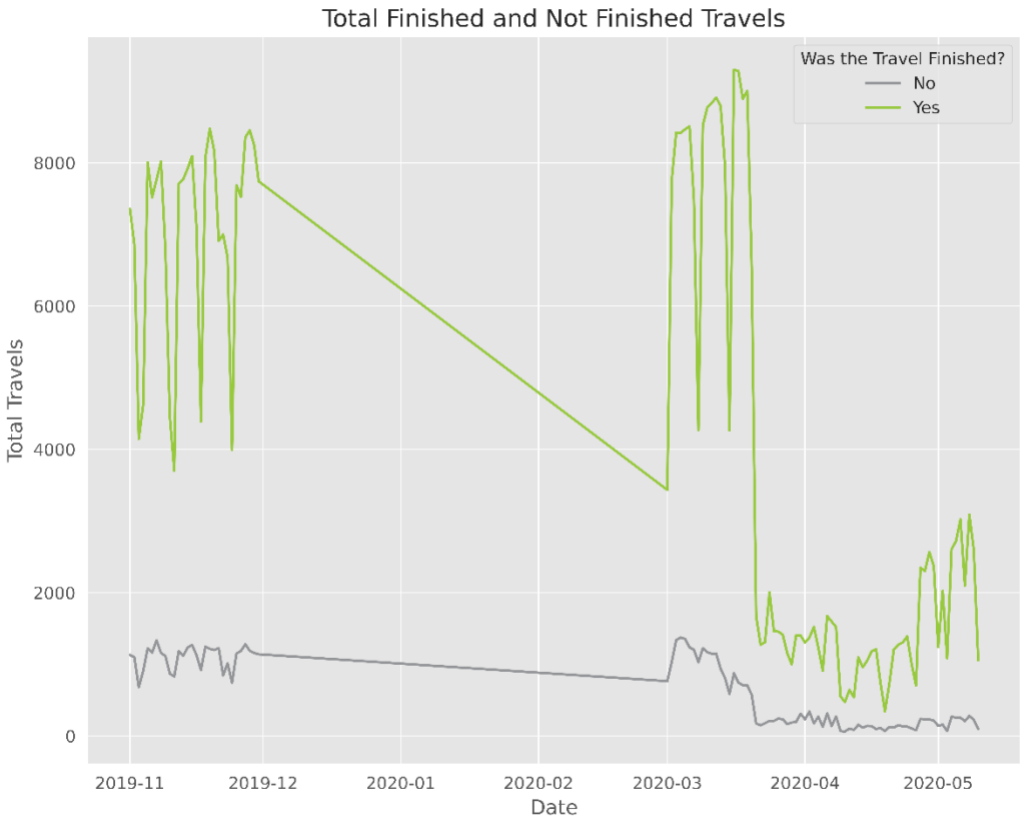
Finished Travels

Breaking down the travels made by whether if they were finished or not, we have,



Where **N** symbolizes that the travel wasn't finished and **S** that it was (**N** and **S** are initials for No and Sí or No and Yes in Spanish). So then, for the totality of travels made along the observation period, 87.67% were finished while 12.33% weren't caused mainly by different factors, such as mechanical or electrical failures, change of itinerary, among others.

Observing the behavior for finished and not finished travels as stated by the passenger volume data, the total of these have has decreased along the last months with a slight increase for may; nevertheless, something positive is that the number of not finished travels has decreased steadily for the last two months, indicating a better control and faster solutions regarding mechanical failures and a better planning of the routes.



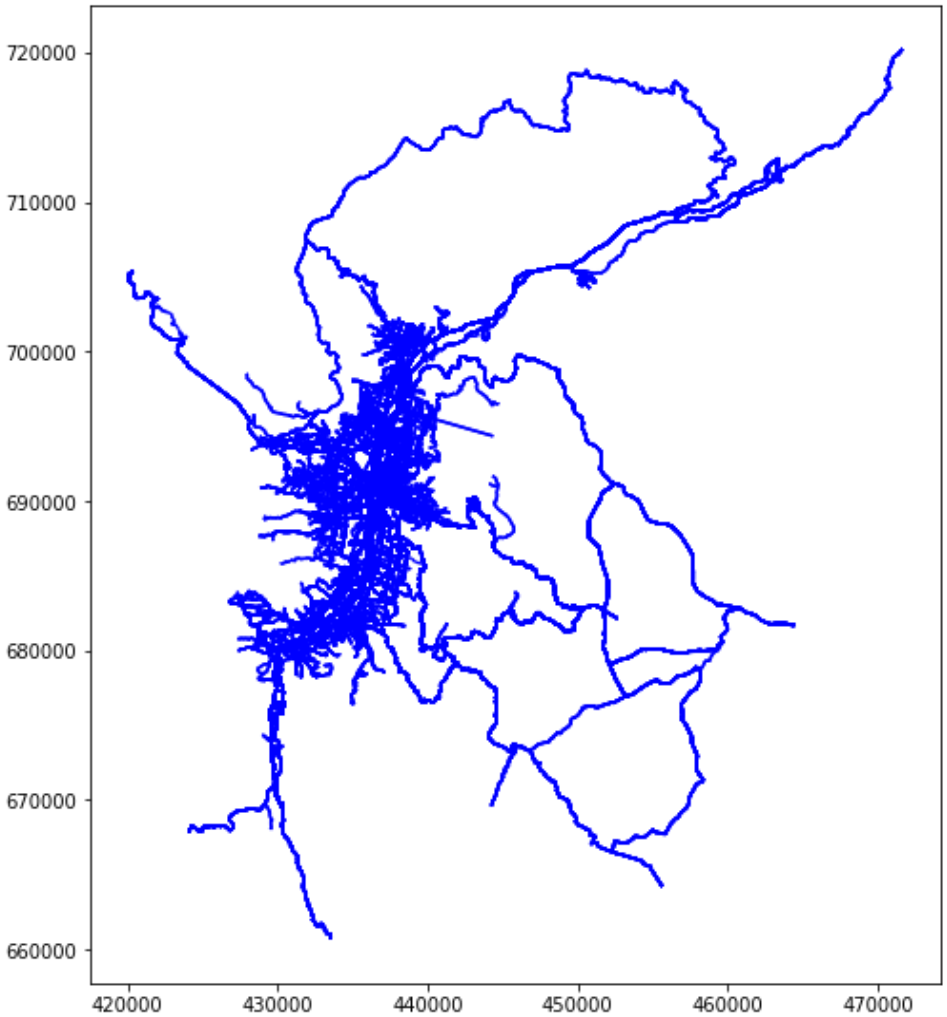
Road Network Data

Once the data is loaded, we can see that we have 9470412 observations from the road network of the City of Medellin.

Configuration parameters of the shape of the road network are:

```
Name: WGS 84 / UTM zone 18N Axis Info [cartesian]: - E[east]: Easting (metre) - N[north]: Northing (metre) Area of Use: - name: World - N hemisphere - 78°W to 72°W - by country - bounds: (-78.0, 0.0, -72.0, 84.0) Coordinate Operation: - name: UTM zone 18N - method: Transverse Mercator Datum: World Geodetic System 1984 - Ellipsoid: WGS 84 - Prime Meridian: Greenwich
```

This graph illustrates the data of the segments that make up the network.



Nodes Data

Once the data is loaded, we can see that we have 186.660 nodes that make up the segment network.

Configuration parameters:

```
Name: WGS 84 / UTM zone 18N Axis Info [cartesian]: - E[east]: Easting (metre)
- N[north]: Northing (metre) Area of Use: - name: World - N hemisphere - 78°W
to 72°W - by country - bounds: (-78.0, 0.0, -72.0, 84.0) Coordinate Operation:
- name: UTM zone 18N - method: Transverse Mercator Datum: World Geodetic System
1984 - Ellipsoid: WGS 84 - Prime Meridian: Greenwich
```