21_Project_Week6

October 3, 2020

1 EDA Public Transport System of El ÃĄrea Metropolitana de Valle de AburrÃą

Project Team No 21

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

1.1 Details of the variables

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 is no null values. It is observed that 2% (501.203) of the records contains 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), DD: is the month of the year (1 to 12), YYYY is the year (2019 and 2020) , HH: MM is the hour (00 to 23) and the minutes (00 to 59). The dates of the events are unique per trip, 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 and 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 AburrÃą. 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 detect 19646 records with coordinates (0,0) and also we found 21049 possible outliers

In the LONGITUD column, There are no missing values. However we detect 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 event exists. Also, there are 83900 records above 3 standard deviation from the mean.

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

In the case of BAJANTRASERA variable there are 191317 records above 3 standard deviation 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 the characteristics of these variables, due 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.

The details shown for every variable can be reviewed in the following documents: - Anexo1_SWEETVIZ - Anexo2_EDA

1.2 Analysis of Transport Network 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 explore the files to visualize the detail of the road network of the City of Medellin.

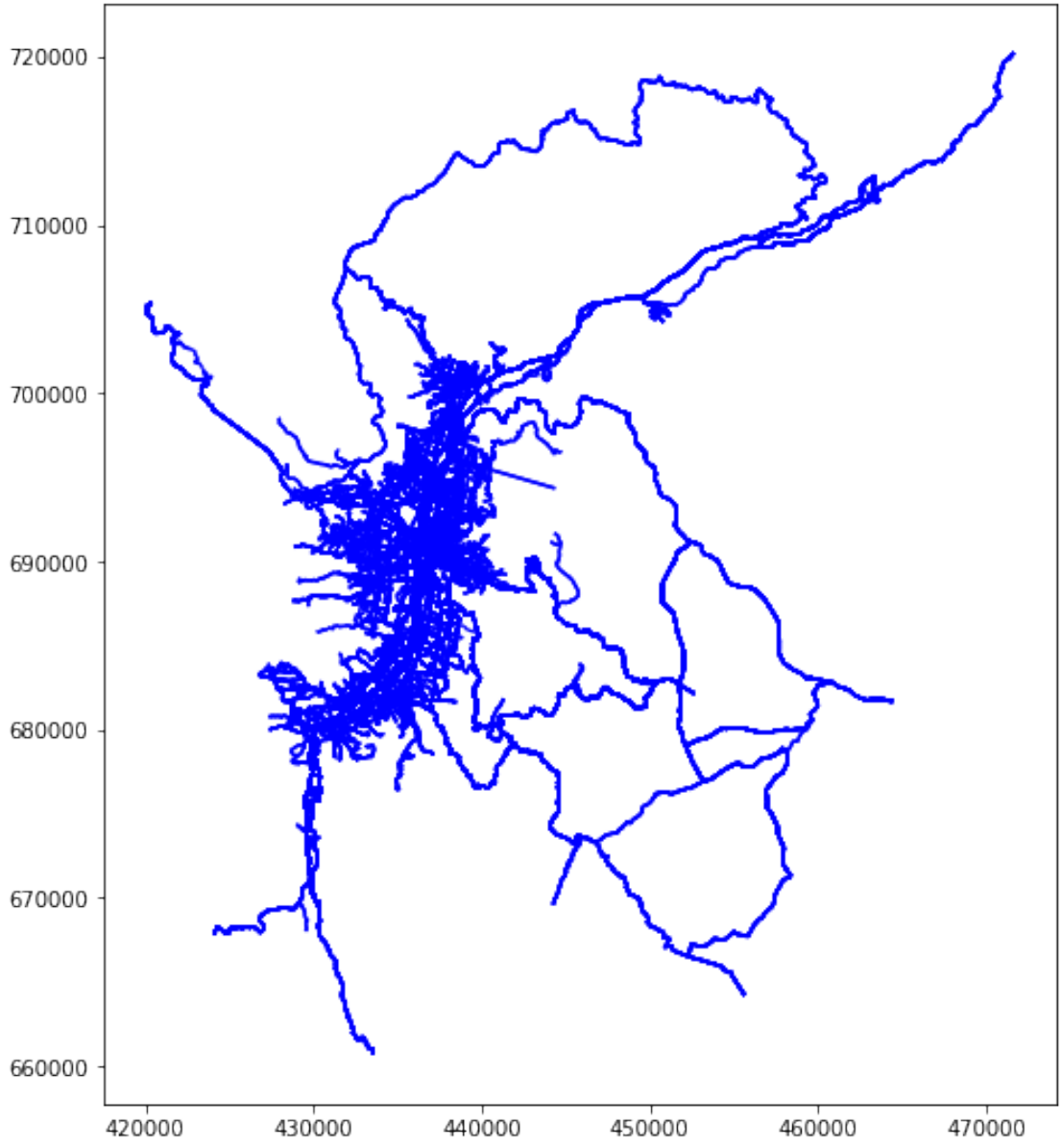
1.2.1 Road network data

Once the data is loaded, we can see that we have `9.470.412` data 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 - 78ÂřW to 72ÂřW - 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.

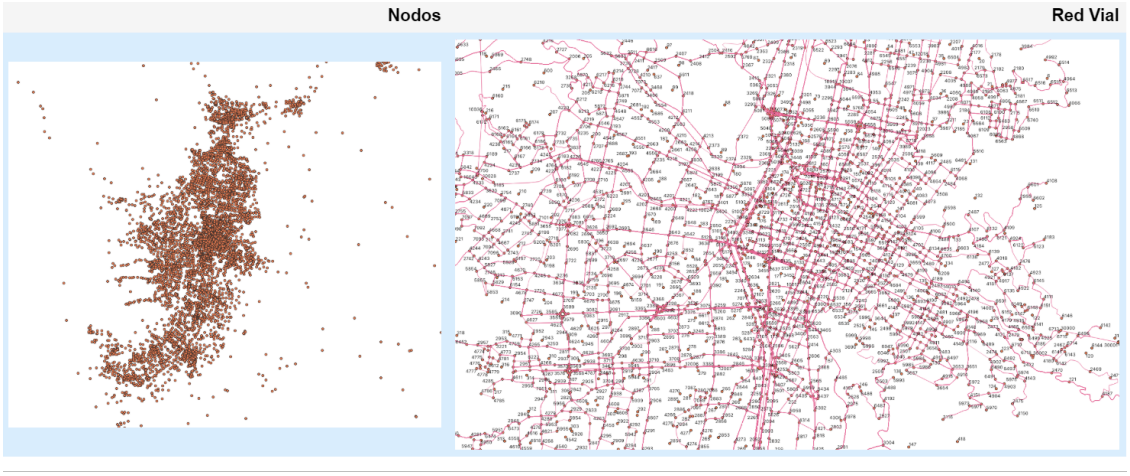


1.2.2 Node data

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

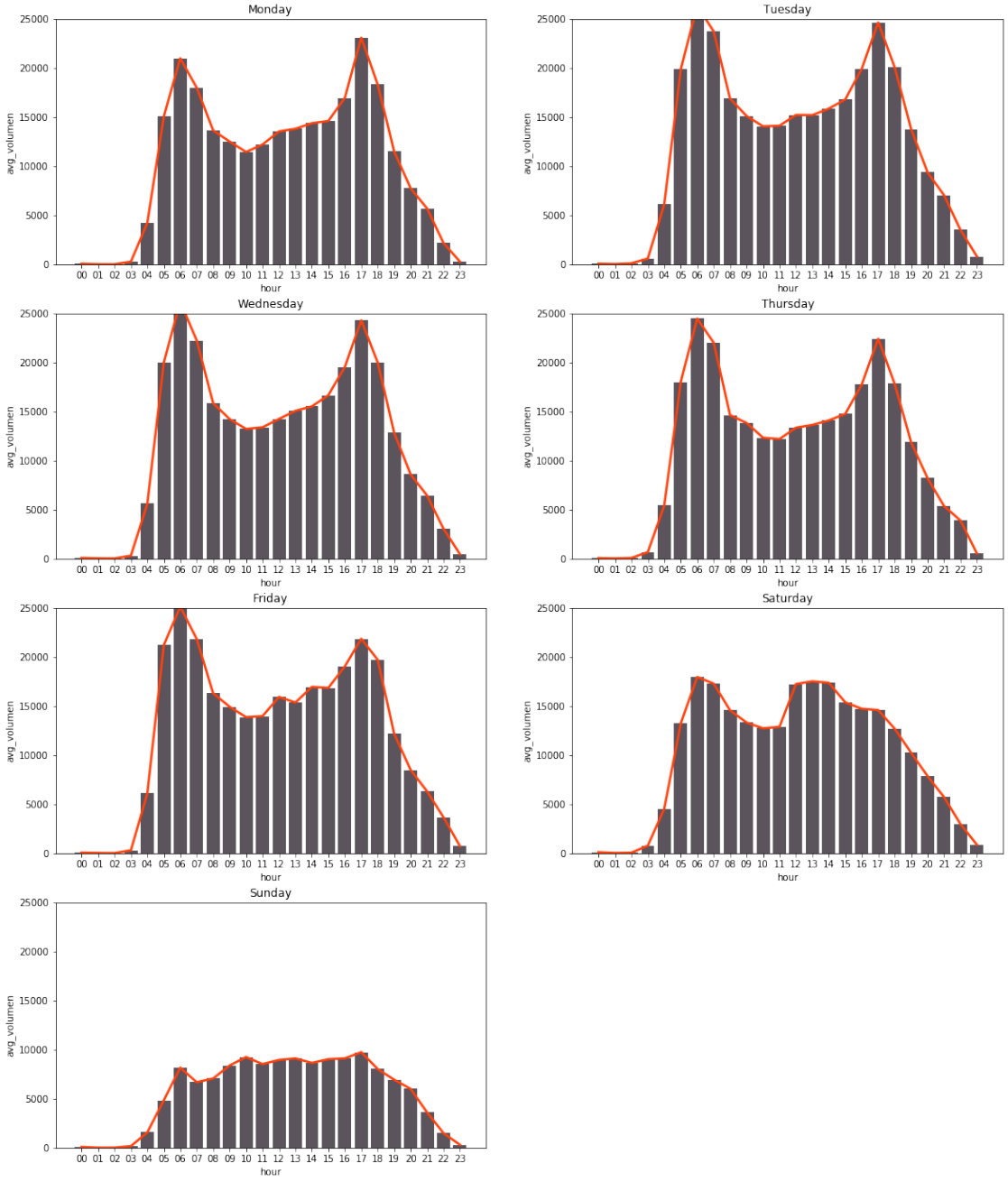
Configuration parameters:

Name: WGS 84 / UTM zone 18N Axis Info [cartesian]: - E[east]: Easting (metre) - N[north]: Northing (metre) Area of Use: - name: World - N hemisphere - 78ÂřW to 72ÂřW - 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

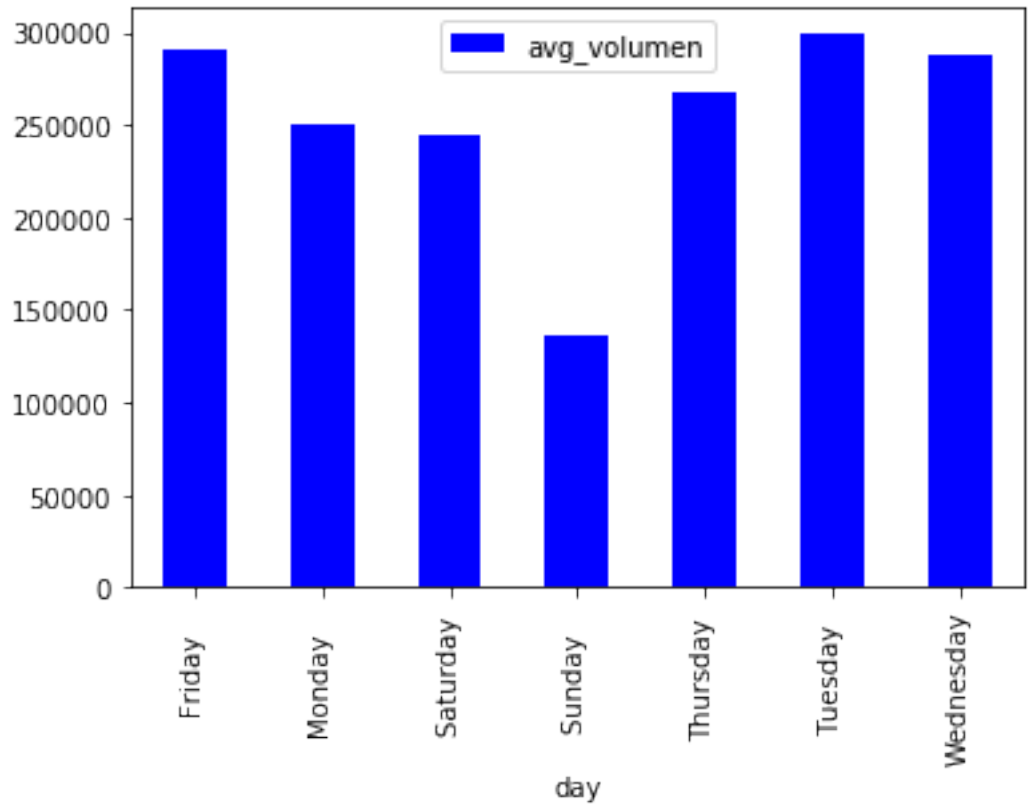


1.2.3 Quantity of Passengers on boarding

We can see that in days during the week, the peaks of the boarding passenger are around 5:00 am and 7:00 am, according to the working hours, and also, we can see another peek around 4:00 pm and 6:00 pm when the people return to their home from work. This peaks during the week is around the 24000 with a maximum 26343 in Tuesdays. During the weekend, the can see a diferent behavior on average of passenger that get on the bus due to the mayority of the population do not trip to the work on weekends, especially on sundays, where the boarding passenger decrease drastically with a maximum average per hour of 9736 passenger and there are no peaks in any hour of the day.

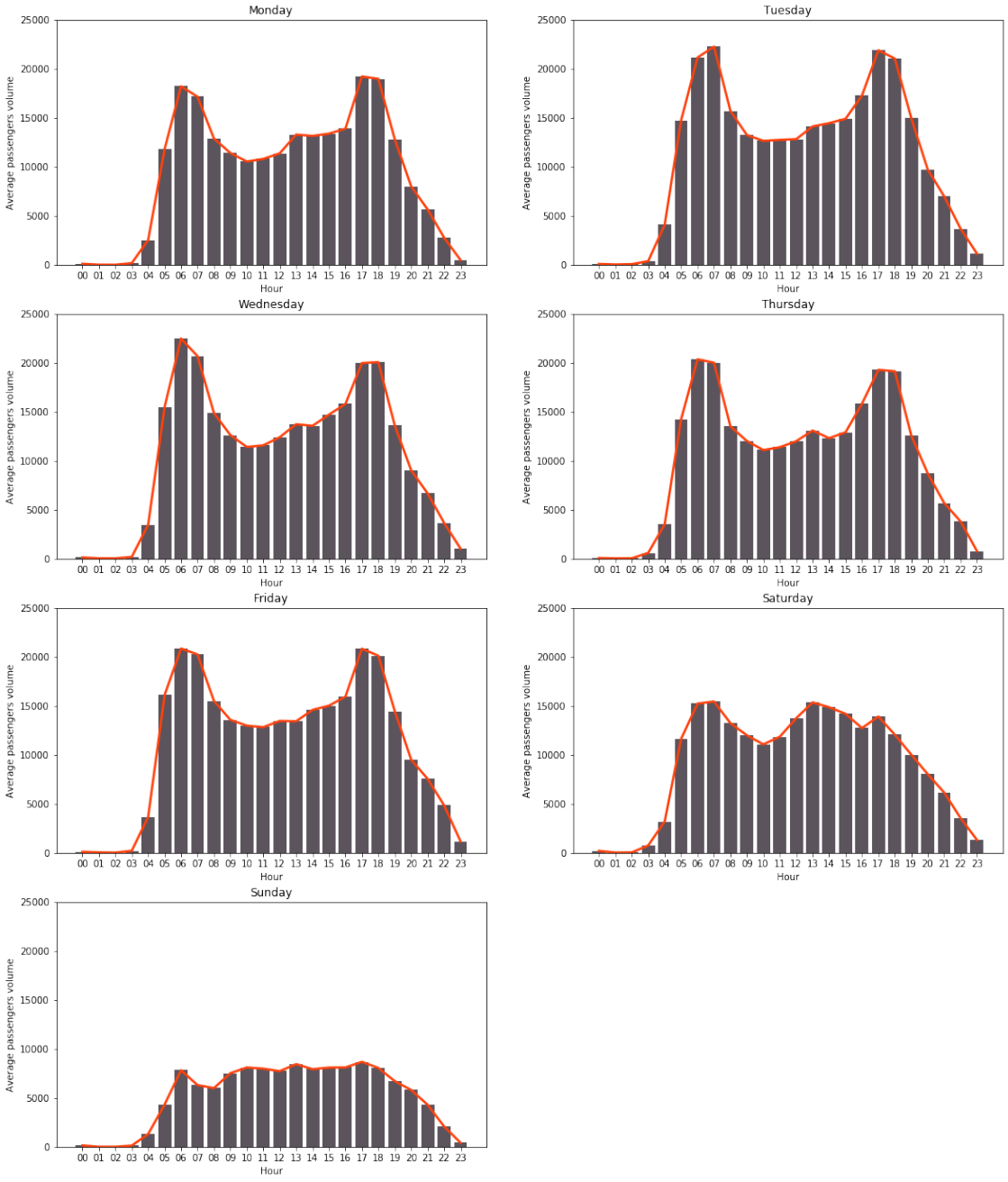


One interesting thing that we can see is that we find that the Tuesdays and Fridays are the days which has the most boarding passengers.



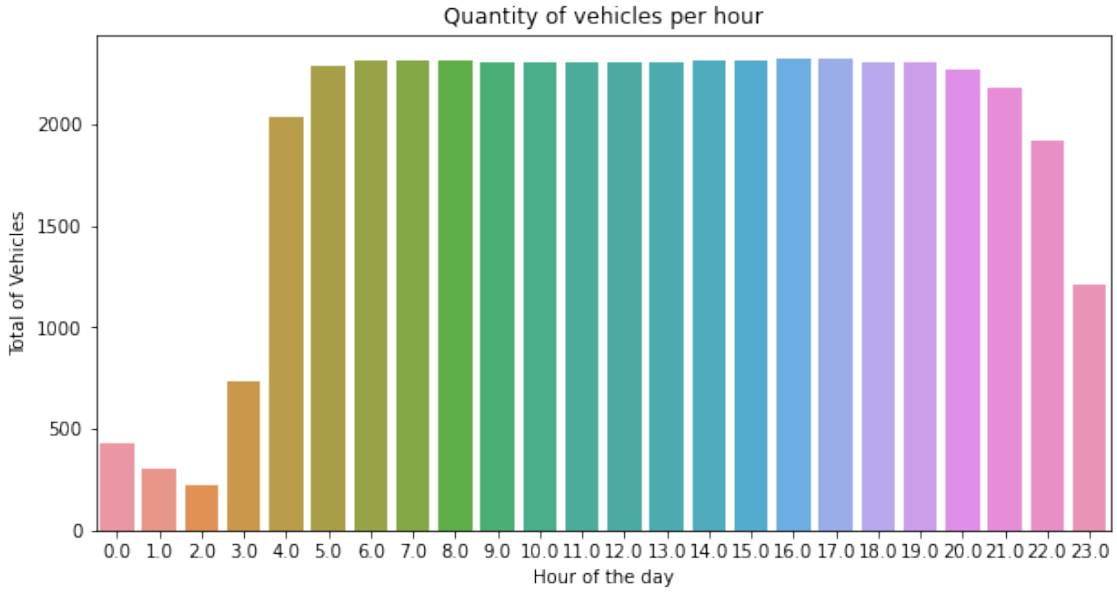
1.2.4 Average Alightings per hour and per day of week

The graphical results are those that would be expected: on average, the highest volume of passengers who get off the transport network vehicles occurs on weekdays, while the volume is lower on weekends. It is clear that the peak hours in the morning are between 5 a.m. and 7 a.m., and peak hours at night are between 5 p.m. and 6 p.m. An in-depth analysis of the data will require segmenting the public transport network by zones to see if it influences the behavior of passenger unloading. It would be expected that this would be the case, since those areas that are residential or that are not centers of business, industry or tourism would show a greater volume of passengers descending from the vehicles in the evening hours (compared to the morning hours) when the residents return to their homes after the day's activities.



1.2.5 Vehicles per hour

It is observed that the flow of vehicles remains relatively the same between 05:00 a.m. and 09:00 p.m. and After 10:00 p.m. decreases the supply of vehicles. This behavior is expected as night and early hours do not have high demand.



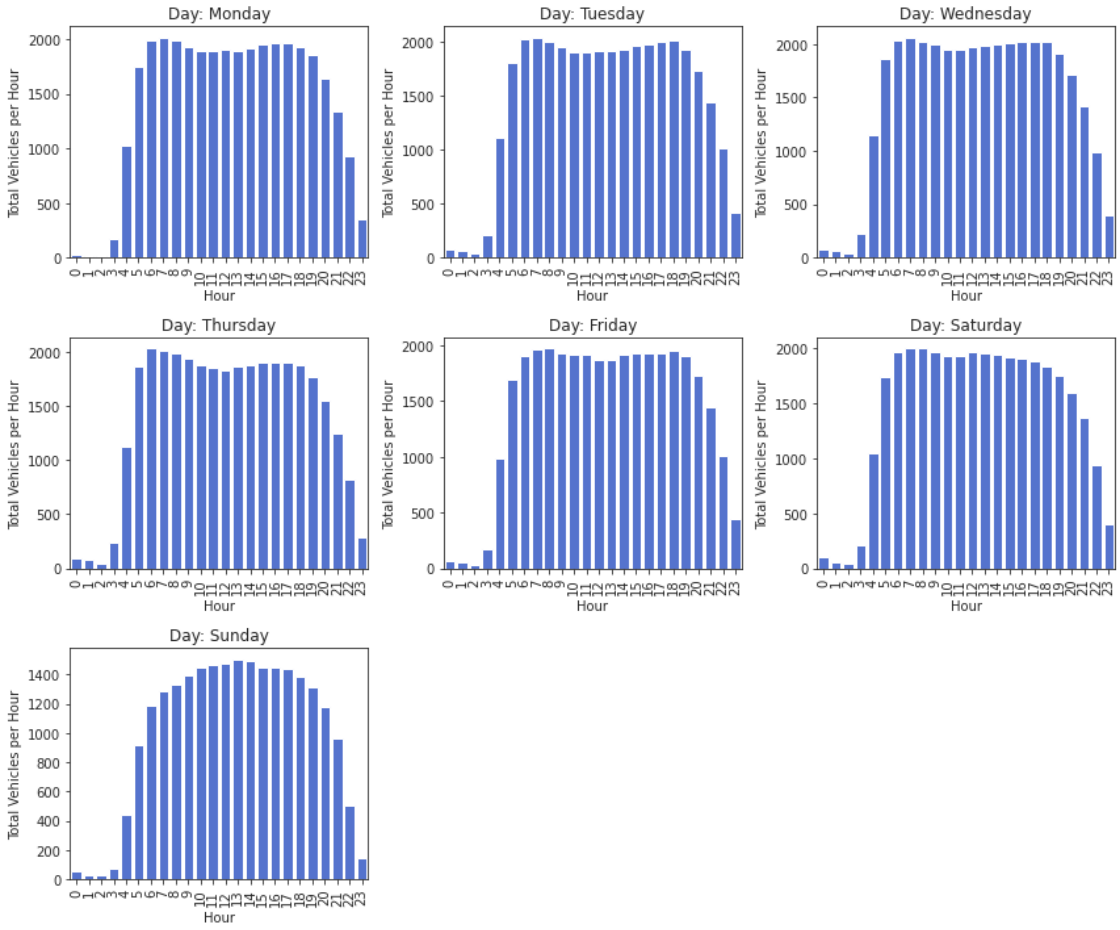
1.3 Vehicles per Day of the Week and Hour

If the total number of vehicles that transit in a day is counted and disaggregating in turn by day of the week, it can be seen that from Monday to Friday, the pattern is approximately similar for the hours where there are a greater volume of vehicles that is, between 5 and 7 in the morning corresponding to the interval in which people travel to their places of work.

Similarly, there is an increase for the same days, around 5 and 6 in the afternoon or between the 17 and 18 hours which is associated with the return from work places to people's homes.

For Saturday, a similar maximum is experienced between 7 and 8 in the morning, as is the case on weekdays, but later a gradual decrease is experienced until the last hours of the day; that is, the bimodal pattern observed from Monday to Friday is not so marked on Saturday.

On Sunday, there is a unimodal behavior such that a maximum number of vehicles in circulation is reached around 1 and 2 in the afternoon or between the 13 and 14 hours. This behavior can be caused because the people tend to engage in outdoor activities along this day



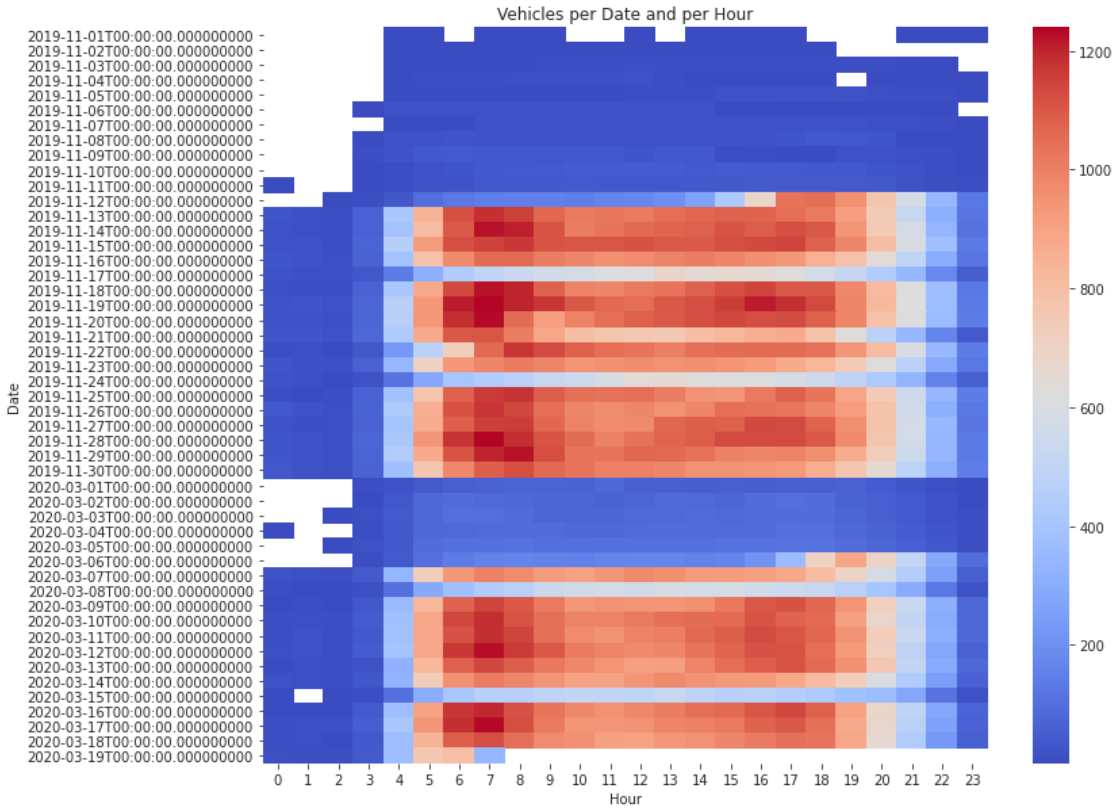
1.3.1 Vehicles per Date and Hour

During the first days of November, the lowest values for vehicle circulation are experienced, starting to have a greater flow in the middle of the month in question, this possibly originates in the information delivered by the entity that was previously organized and only delivered the records for some months where the data was considered reliable.

As of the date mentioned, a pattern begins to be seen that is repeated almost in a general manner throughout the period of interest. The volume of vehicles begins to rise between 3 and 4 hours to reach the highest point between 7 and 8 hours. Around 9 to 16 hours, the flow remains approximately constant and a peak is experienced at around 17 and 18 hours as seen in the daily breakdown. The flow begins to descend until 22 hours arriving at 23 where it begins to take the minimum until 2 hours the next day.

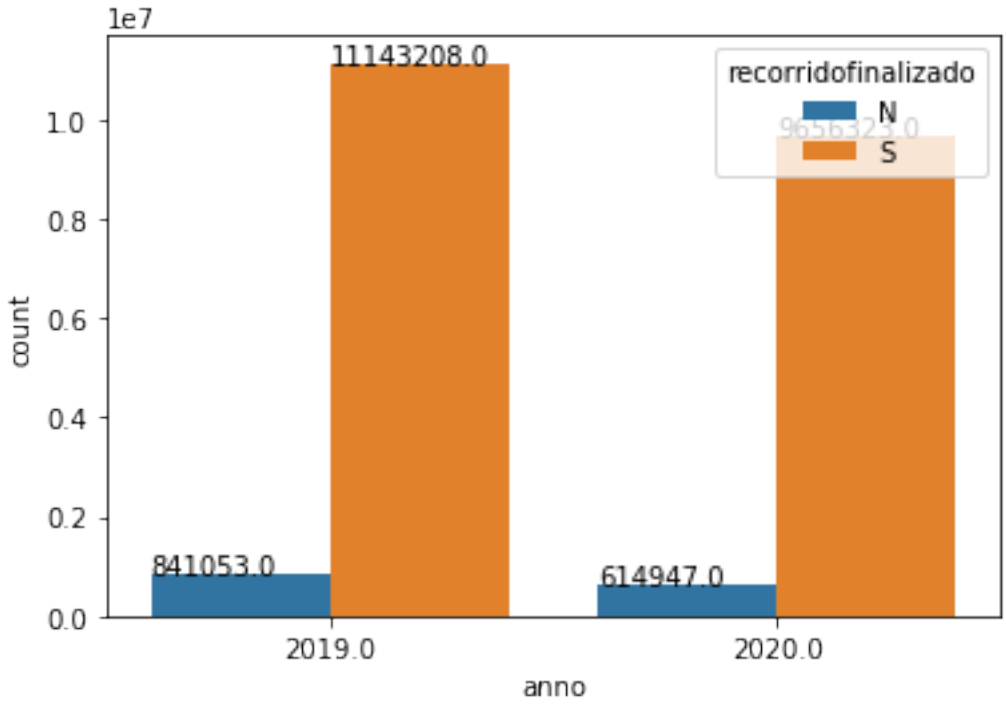
No observations were found from January to the beginning of March, and for those of that month, the flow of vehicles is minimal between days 1 and 6. From day 7 the behavior stabilizes and acquires the behavior that was appreciated for the month of November.

The highest points of vehicle flow occur approximately at the middle and end of each month.



1.3.2 Routes finished

It is observed that there is a number of routes that were not terminated: 2019 had 841053 and as for 2020, we have 614947. There can be different factors, such as mechanical or electrical failures, change of itinerary, among others.


1e7
11143208.0
recorridofinalizado
N
S
1.0
0.8
0.6
0.4
0.2
0.0
count
841053.0
614947.0
2019.0
2020.0
anno