FINAL PROYECT

MONTEVIDEO TRANSIT OVERSPEED INQUIRY

Federico Novelli

January, 2022

CASE OF STUDY:

A transit prevention company is keen on realizing a project of transit security improvement as to show its potential to new customers. Due to the high level of exposition this project has, they must ensure to make the highest outcome possible. As the project consists in educating and preventing car accidents and ensuring they respect the speed limits, a study of this matter must be made. The company is settled in Uruguay and wants to approach the capital municipality, the city of Montevideo. The company must choose the spot in Montevideo with highest rate of overspeed's.

USEFUL DATA:

The data needing for this data analysis is the following:

- Montevideo city areas geographic coordinates:
  This is found in the Open Data Catalog from the Data and Statistics window of the Uruguayan government site: https://catalogodatos.gub.uy/dataset?tags=movilidad&organization=intendencia-montevideo
  This also has the information of the postal code of each area needed for registering the validation of the project.
- Montevideo city geographic localization of the car speed measuring devices throughout the city:
  This file also provides the registry of velocity measurements for the hole month of December 2021. This file was extracted through the same site: https://catalogodatos.gub.uy/dataset?tags=movilidad&organization=intendencia-montevideo
- We will also need the localization of each neighborhood:
  The area coordinates are not specific enough and we will need the respective neighborhood to target. This data was found in the Geographic Computer System: https://sig.montevideo.gub.uy/
  The data was downloaded as a DBF file and converted to CSV open UTF8 format file.
- Geographic coordinates of streets of Montevideo city:
  Found in Foursquare site as a developer user API requester: https://es.foursquare.com/developers/projects

METHODOLOGY WALKTHROUGH

Montevideo area coordinates:

After importing the csv file downloaded, we must clean the data. Only the columns Latitude, Longitude, Postal code and City must remain. Geographic coordinates must be shortened and all data without decimals must be transformed to integer. This is to simplify the view. Also change column names for all upper case and easy names. All strings changed also to upper.

Velocity and localization data:

This is the most abundant information. The same changes must be done to data as last file. The important data is: Latitude and Longitude of sensor, street where it is at and the hour of the record. There can be NaN of cero values that must be withdrawn. Categorize the hours into 4 categories depending on the range of the day (night, morning, afternoon, etc). This will be useful for further analysis and sensitivity study of the hours and the speeds. Data must be grouped. The registers of the same spot must be grouped as to shorten the dataframe and the corresponding velocity values must be substituted by the mean and quantity of values used for that mean calculation. This dataframe is the most important. It should be extracted the top critical areas depending on the over speeding and the quantify of cars that do so. Two different analyses will be made. Analyze this data.

Neighborhood geographic information:

We must categorize each measurement of sensors by the neighborhood to which they correspond. When we have the conclusion of target area to assess, the neighborhood must be known. It is also useful to show in map. The dataset has no information of the coordinates of the neighborhoods. Nevertheless, it has the distance to the 0 km spot located in the monument Obelisto located in Plaza Cagancha, Centro. As to cross reference the latitude and longitude of places to their neighborhoods, the distances between measurement points and areas and the 0 km location must be calculated. This way the cross-reference can be done with a certain amount of flexibility and approximation.

Visualization:

Measurement points will be displayed in the map. Also, the top 10 places where top overspeed's are achieved and the top quantity of cars. Clustering will be made to classify critical areas and find an external explanation. algo a linear plot of all the means velocity over speed measurements will be plotted in reference to the distance of the 0km. This is to detect a strong anomaly in certain point. Afterwards, histograms of all de velocities, classified in the range of hours in which the day was categorized, will be plotted. This is to detect in which hour is best to approach the issue of the project. It could be that the time of the day induces another way of driving.

NEEDED LIBRARIES FOR DATA HANDLING:

The libraries that were used for extracting, modifying and visualizing data where the following:

- Numpy
- Pandas
- Json
- Geocoders from Geopy
- Requests
- Pandas.io.json from Json_normalize
- Cm from Matplotlib
- Colors from Matplotlib
- Pyplot from Matplolib
- Folium
- KMeans from Sklearn.cluster
- Sin, Cos, Sqrt, Atan2, Radians from Math

DATA SOURCES:

MONTEVIDEO.CSV DATASET:

This dataset was imported as a csv file. It contains information about the geographical and situational information about houses in Montevideo. Apart from the geographical coordinates, it contains the postal code, the street and the cornering streets of the location, the code of search for the location to which each row refers, the street code and the corresponding city.

The data of interest is the city of each geo coordinates as well as the postal code. After performing the analysis needed for the case of study, we must select a point in the map of Montevideo and specify the neighborhood, city and postal code. The las two pieces of information are obtained through this dataset.

Size of dataset: 411,682 rows x 15 columns.

AUTOSCOPE_12_2021_VELOCIDAD.CSV DATASET:

This dataset contains the measurements of speeds of cars by several cameras distributed all over Montevideo. The information contains the data registry for the hole month of December 2021. Only this month data was used because of the large amount of data to handle. The information imported includes: geographic location of the measurement, code of the device, street in which it is located and both corner streets, the hour and date of the measurement, the number of the lane of the street in which the car was located and the velocity measured.

The data of interest is the velocity value, the hour in which it was taken and the latitude and longitude. The information regarding the hour will be used to classify the speeds in 4 ranges of hours of the day as to sense the sensitiveness of the velocities due to the portion of the day.

Size of dataset: 4,009,784 rows x 10 columns.

INE_BARRIOS_MVDUTF8.CSV DATASET:

This dataset contains information regarding the names of the neighborhoods. In this case we do not have geographic coordinates. Instead we have the distance in kilometers from the 0km point in Montevideo. This point is located in the Obelisco statue constructed in Plaza Cagancha in the neighborhood called Centro. The information of the dataset is the distance, the name of the neighborhood, its number and code.

The data of interest is the name of the neighborhood and the distance from the 0km point.

Size of dataset: 62 rows x 4 columns.

The geographic coordinates of the 0km point are the following:

Latitude: -34.905869.

Longitude: -56.191386.

DATA CLEANING:

MONTEVIDEO.CSV DATASET:

There are many houses from the same city. We must know the approximated geographic coordinates of each city. In the dataset also are found wrong values regarding other departments out of Montevideo.

Chances made:

- Drop columns that are not needed
- Erase information that does not correspond to the city of interest
- Change case of strigs and columns titles
- Chop decimal numbers of floating points
- Erase NaN values
- Reorder columns due to its utility

The withdrawn of extra information resulted in the following size changes:

Original size: 411,682 rows x 15 columns.

New size: 394554 rows × 4 columns.

Result show:

```
        LAT     LON     PCODE  CIUDAD

0       -34.76  -56.22  12400  ABAYUBA

1117    -34.76  -56.21  12400  ABAYUBA

1116    -34.76  -56.21  12400  ABAYUBA

1115    -34.76  -56.21  12400  ABAYUBA

1114    -34.76  -56.21  12400  ABAYUBA

...     ...     ...     ...    ...

410404 -34.75   -56.11  12400  TOLEDO CHICO

410403 -34.75   -56.11  13000  TOLEDO CHICO

410401 -34.75   -56.11  13000  TOLEDO CHICO

410408 -34.75   -56.11  13000  TOLEDO CHICO

410402 -34.75   -56.11  13000  TOLEDO CHICO
```

<u>AUTOSCOPE_12_2021_VELOCIDAD.CSV DATASET:</u>

There are many data taken in almost the same hour and in the same location. The key is to group the information by geographic location. We must gather all the data of speed measurements taken for each location at a given range of hour and find the mean of each one.

It is better to study the over speeds and under speeds separately as to manipulate each one independently.

Chances made:

- Drop columns that are not needed
- Erase information that does not correspond to the city of interest
- Change case of strigs and columns titles
- Chop decimal numbers of floating points
- Erase NaN and cero values
- Reorder columns due to its utility
- Extract only the hour of the date columna
- Clasify the hour in 4 categories The withdrawn of extra information resulted in the following size changes:

Original size: 4,009,784 rows x 10 columns.

New size: 3674379 rows × 5 columns.

Result show:

```
LAT       LON       DAYTIME             VELOCITY            STREET

0         -34.84    -56.15    noche     45        gral flores

1         -34.89    -56.17    noche     82        arenal grande

2         -34.87    -56.19    noche     30        evaristo ciganda

3         -34.88    -56.16    noche     27        garibaldi

4         -34.87    -56.14    noche     27        bv batlle y ordonez

...       ...       ...       ...       ...       ...

4009777             -34.83    -56.22    madrugada           42        garzon

4009779             -34.92    -56.15    madrugada           35        solano garcia

4009781             -34.81    -56.22    madrugada           82        garzon

4009782             -34.81    -56.22    madrugada           41        garzon

4009783             -34.82    -56.22    madrugada           30        garzo
```

We then spilt the data dependent on weather the velocity is over the limit or under the limit. The general limit speed is 60 km/h.

After having each dataframe, we group it by its location. The velocity value is substituted by the mean value and also added the number of numbers taken into account. Locations coordinates numbers are chopped to 2 decimas for a better grouping and less exact.

New size of each dataframe: 62 rows × 4 columns.

Result show for over speed values dataframe:

```
LAT     LON     MEAN_OVER       COUNT_OVER
0       -34.92  -56.16  69.84   247
1       -34.92  -56.15  79.55   1102
2       -34.91  -56.17  69.84   1346
3       -34.91  -56.16  76.49   18194
4       -34.91  -56.15  73.04   9574
...     ...     ...     ...     ...
57      -34.82  -56.22  69.82   144
58      -34.82  -56.21  65.91   66
59      -34.81  -56.22  74.16   732
60      -34.80  -56.22  71.72   18
61      -34.79  -56.22  72.17   816
```

Result show for under speed values dataframe:

```
LAT     LON     MEAN_UNDER      COUNT_UNDER
0       -34.92  -56.16  42.35   22626
1       -34.92  -56.15  29.72   33358
2       -34.91  -56.17  41.42   17330
3       -34.91  -56.16  32.25   107931
4       -34.91  -56.15  31.52   83928
...     ...     ...     ...     ...
57      -34.82  -56.22  32.08   33383
58      -34.82  -56.21  38.63   17633
59      -34.81  -56.22  39.14   39968
60      -34.80  -56.22  33.74   8505
61      -34.79  -56.22  43.78   14327
```

We also found the top 10 over speeding locations and the top 10 locations where most people exceeded the speed limit. Statistical values where generated for each top 10 list.

As to use there values with the next dataset, the distance from the 0km point for each measurement location is needed. After preforming calculations, appending the dataframes of over and under speeds and appending the distance values, we ended up with the following information:

	LAT	LON	DISTANCE	MEAN_OVER	COUNT_OVER	MEAN_UNDER	COUNT_UNDER
0	-34.92	-56.16	3.30	69.84	247	42.35	22626
1	-34.92	-56.15	4.10	79.55	1102	29.72	33358
2	-34.91	-56.17	2.00	69.84	1346	41.42	17330
3	-34.91	-56.16	2.90	76.49	18194	32.25	107931
4	-34.91	-56.15	3.80	73.04	9574	31.52	83928
...	...	...	...	...	...	...	...
57	-34.82	-56.22	9.90	69.82	144	32.08	33383
58	-34.82	-56.21	9.70	65.91	66	38.63	17633
59	-34.81	-56.22	11.00	74.16	732	39.14	39968
60	-34.80	-56.22	12.10	71.72	18	33.74	8505
61	-34.79	-56.22	13.20	72.17	816	43.78	14327

62 rows × 7 columns

INE_BARRIOS_MVDUTF8.CSV DATASET:

For this dataset we only need the information of the distance and the name of each neighborhood. A good indicator that the previous data was well manipulated and grouped is that we ended up will the same number of rows as number of neighborhoods.

The manipulation of data consisted it:

- Keeping only 2 necessary columns
- Changing columns titles and orders
- Cleaning names with strange symbols
- Changing all names with upper case
- Chopping the distance values to 2 decimals as the previous calculations.

Result show:

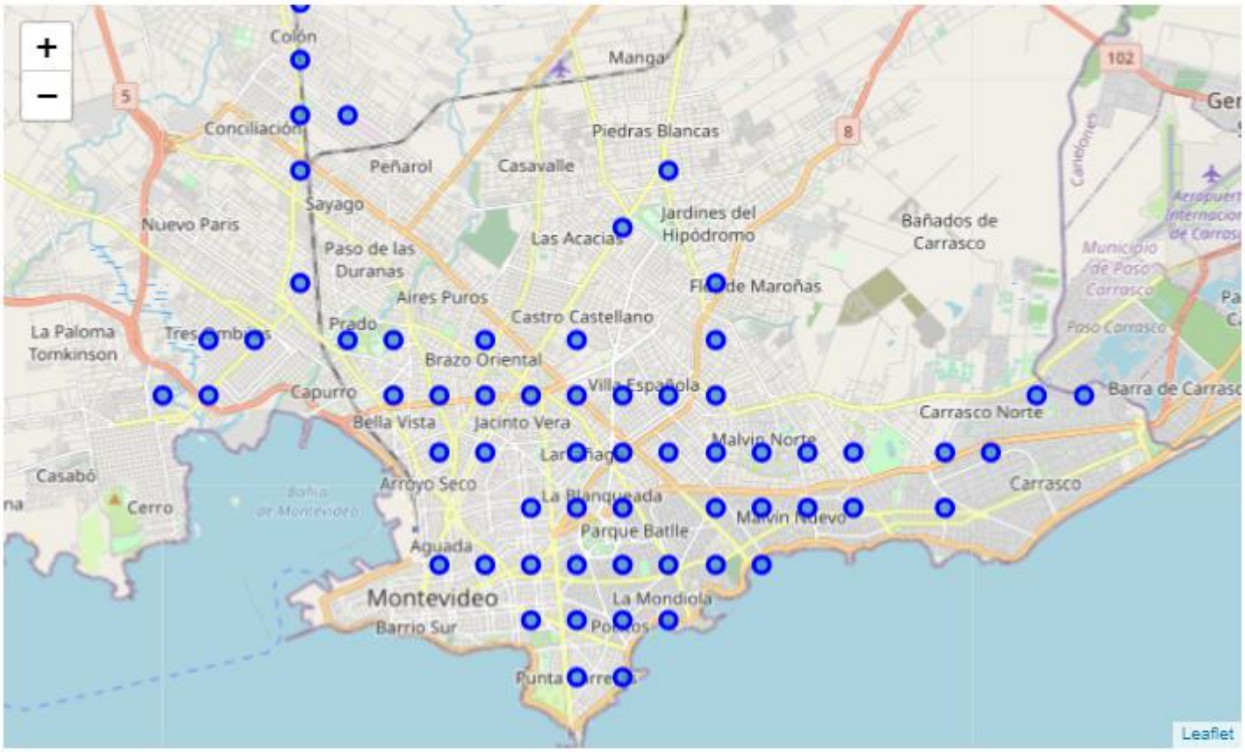
```
DISTANCE        BARRIO
0       2.10    CIUDAD VIEJA
1       1.30    CENTRO
2       0.70    BARRIO SUR
3       2.30    CORDON
4       0.80    PALERMO
...     ...     ...
57      29.70   VILLA GARCIA, MANGA RUR.
58      5.40    MANGA
59      3.10    POCITOS
60      3.20    BELVEDERE
61      3.30    LA TEJA
62 rows × 2 columns
```

We then cross-references this dataframe with the one that includes the velocity measurements summary as to find to which neighborhood each measurement corresponded.

VISUALIZING DATA SOURCE LOCATION:

Location of each grouped speed measuring point:

We show in the map the different point of measurement chosen for each neighborhood:



TOP 10 CRITICAL PLACES:

Using the data of the top 10 places where the highest speeds where achieved and the top 10 places where more quantity of cars exceeded the limit, we found the following statistics:

Top 10 places where the highest speeds were achieved:

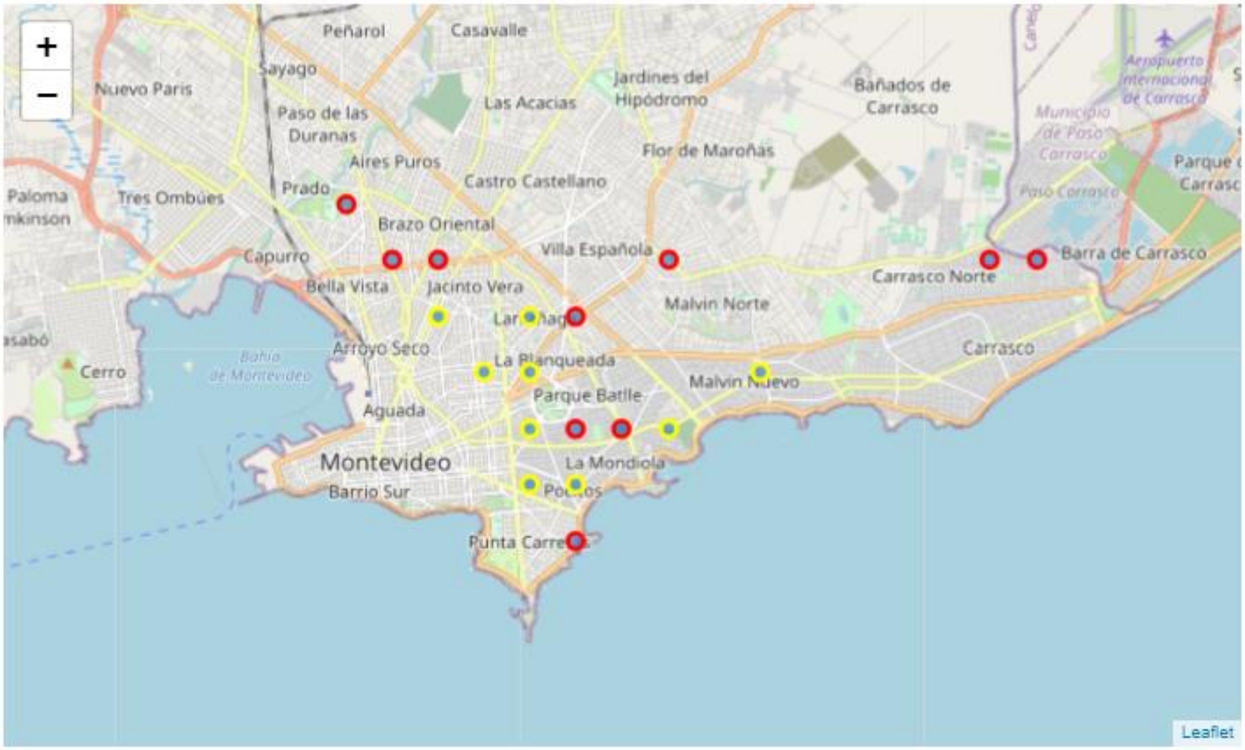
LAT	LON	MEAN_OVER		COUNT_OVER
count	10	10	10	10
mean	-34	-56	80	2279
std	0	0	2	5046
min	-34	-56	77	13
25%	-34	-56	78	30
50%	-34	-56	80	173
75%	-34	-56	82	983
max	-34	-56	84	16099

Top 10 places where more cars exceeded speed limit:

LAT	LON	MEAN_OVER		COUNT_OVER
count	10	10	10	10
mean	-34	-56	73	10223
std	0	0	4	4146
min	-34	-56	68	6090
25%	-34	-56	71	6961
50%	-34	-56	72	9228
75%	-34	-56	72	11910
max	-34	-56	83	18194

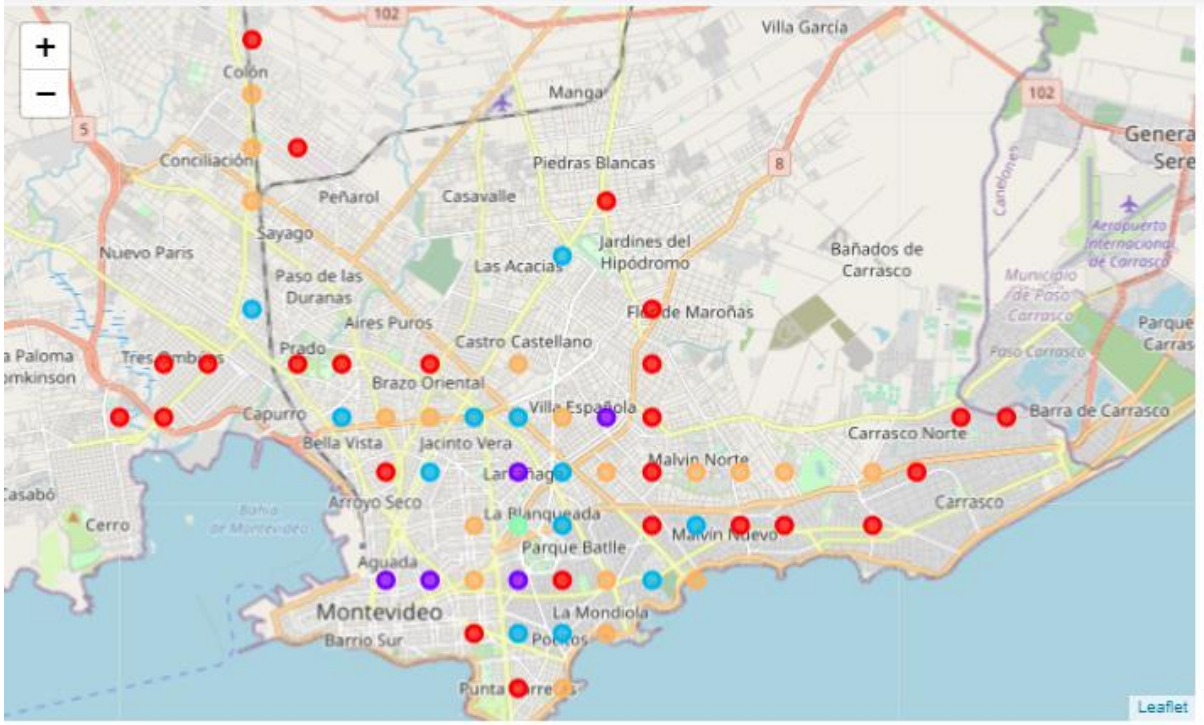
VISUALIZING LOCATIONS OF TOP CRITICAL PLACES:

In red dots it the show the places where highest speeds were achieves. In yellow dots the places where more cars exceeded the speed limit.

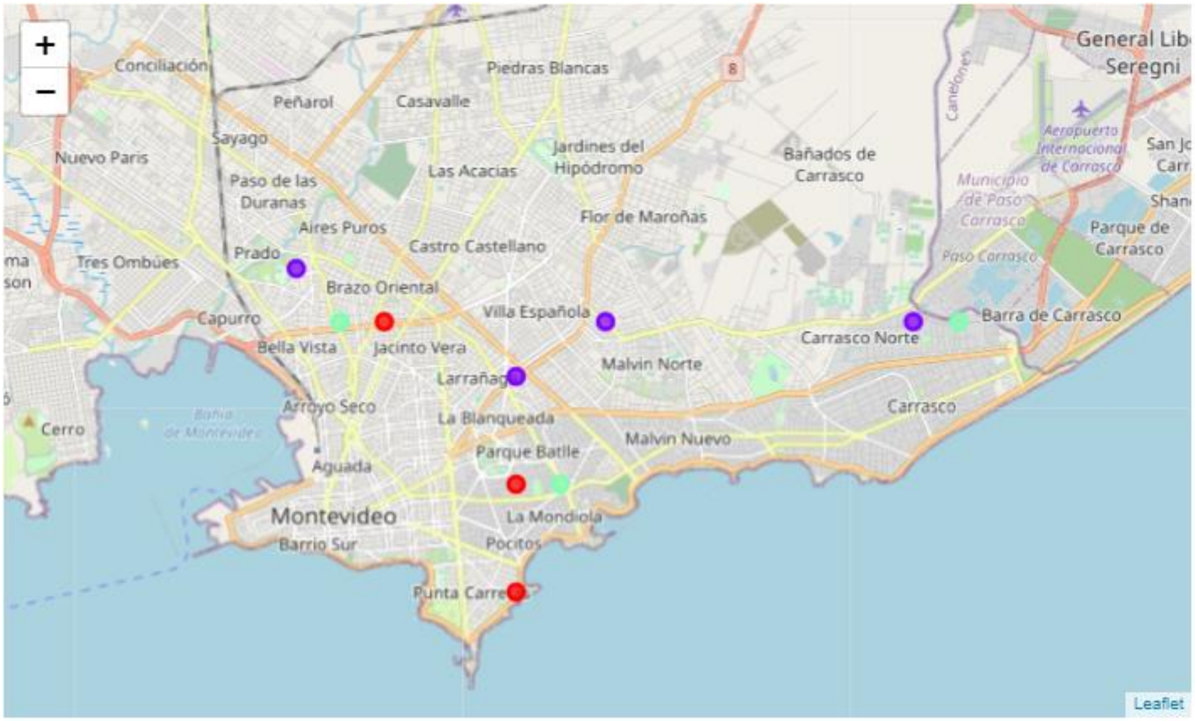


CLUSTERING:

We grouped measuring places into 5 clusters as to see if some clusters where critical areas and areas of interest.

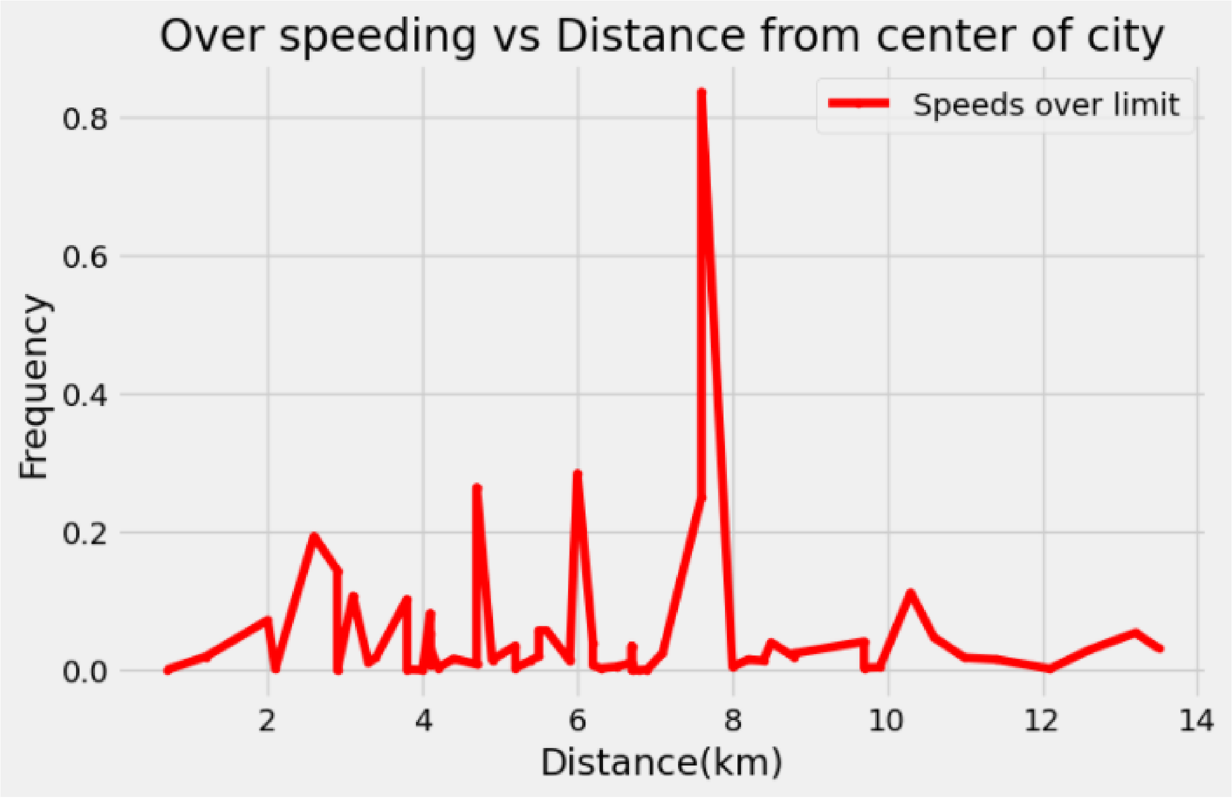


The top 10 speed measurements where clustered into 3 groups to detect further critical areas.



DATA ANOMALLY DETECTION:

Being able to visualize all the over speeding frequency of occurrences depending on the distances will help us detect which critical places are the keenest on having speed problems. In other words, in what place more people tend to commit an illegal driving.



We see that around the 7km distance there is a place where more cars tend to overspeed. This is the target place to conclude the project survey.

SENSITIVITY STUDY:

Another factor to study is in what range of hours the illegal driving occurs. Is the project pretends to deal with this problem, it is also useful to know at what hour is more convenient to educate people that transit this place.

For this, the gather the hole velocity measurements without grouping it but after filtering bias data. We then classify it by 4 ranges of hours:

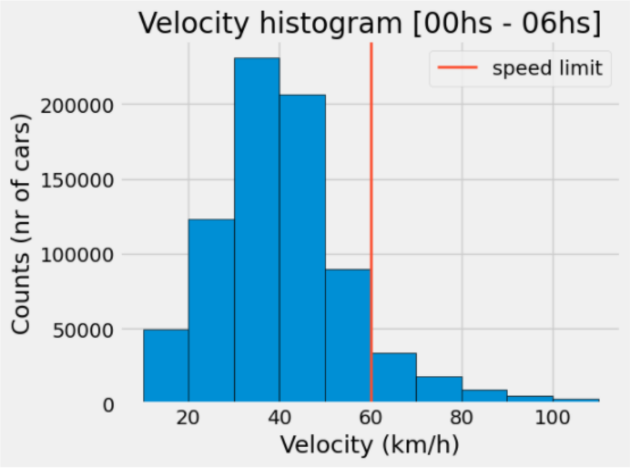
00hs to 6hs = before morning

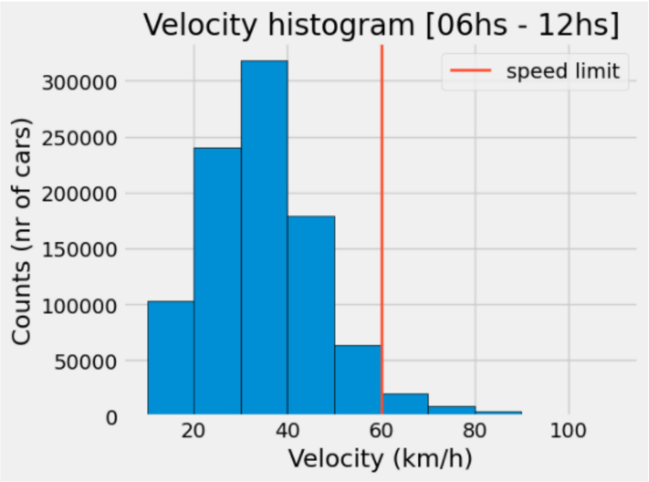
6hs to 12hs = morning

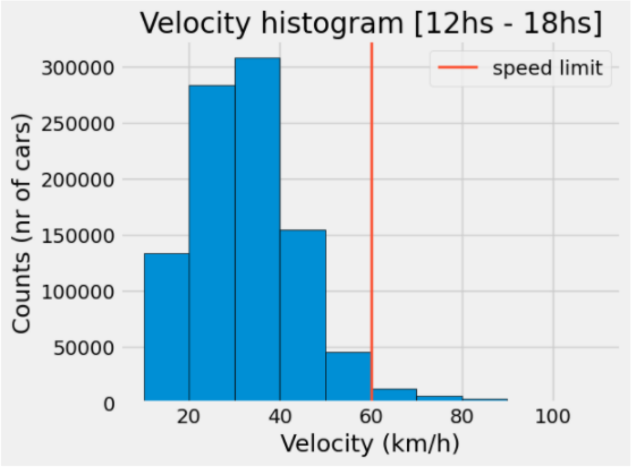
12hs to 18hs = afternoon

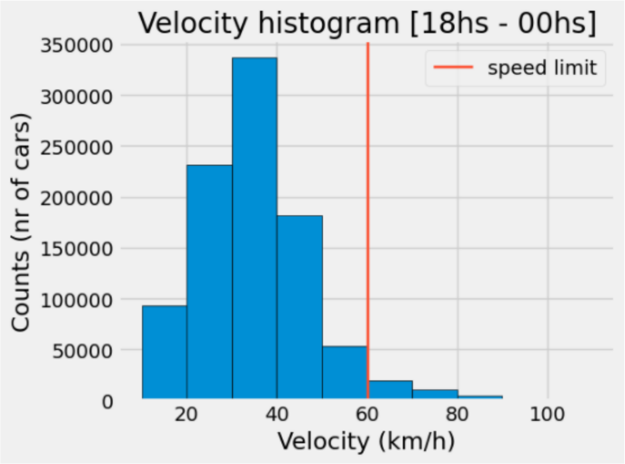
18hs to 00hs = night

The we plotted a histogram for each hour range and see if there where difference in the driving speeds.









We see that the is no noticeable change in the speed patron due to the part of the day. As conclusion, the hour does not affect driving.

CONCLUTION:

Based on the data analyzed we can conclude that there is a tendency to over speed in the surrounding areas of the center of the city that is the center of business. This could be due to the stress of leaving and reaching work places. The critical point are not found near the business area, near 0km point, because the transit is too dense to overspeed.

Transit control devices with capability of giving a ticket, as well as police man, must be strengthen in the surrounding areas of the business center.

The hour of the day is not a factor that affects the driving conduct of cars. This reflects that the project of the company could be preformed at any hour of convenience.

The critical spot found where the highest rate of cars speeds over the legal limit are achieves is has the following location information:

FREQUENCY OF ILLIGAL SPEED DRIVING: 88%

NEIGHBORHOOD: CARRASCO

CITY: MONTEVIDEO

LATITUDE: -34.87

LONGITUD: -56.25

This is the chosen place for preforming the project of the company to which the survey was realized.