Capstone project

Traffic Accidents vs Nightlife and the relocation of Traffic Enforcement Cameras according to accident data in Medellín

Camilo Uribe

April 2020

Table of Contents

1. Introduction

In 2019, 6.329 people died in traffic accidents in Colombia, becoming the second cause of violent deaths in the Country, behind homicide [1]. Any measure adopted towards the increase of road safety has the potential of saving hundreds of lives, which should be among the local government top priorities. An insight that can be obtained from existing accidents data in order to propose informed traffic safety policies is the relationship between traffic accidents and nightlife. We will see if exists evidence that supports the hypothesis that there are more serious accidents in the weekends, especially in late night/ early morning hours (from 22:00 to 04:00), and check close venues to serious incidents to see if there are bars or clubs among those venues, which may suggest drunk drivers causing these deaths.

On the second part of the project, we will evaluate one of the existing solutions already implemented: traffic enforcement cameras (also called speed cameras), located in strategical sites, which automatically fine the vehicle owner if that particular road speed limit is violated. The location of these cameras in the city seems arbitrary, causing a lot of citizen complaints, and should be reconsidered according to the behavior of the accidents data, to make sure zones with high risk of accidents are covered.

2. Data

2.1 Data sources

This project will use two datasets. The main dataset consists of georeferenced data for all of the reported traffic accidents in Medellín in 2019. This information is made public online by the local city hall. The original dataset has many columns that contain neighborhood, coordinates, class of accident (collision, running overs, etc), date, hour, severity (if the accident involves injured or dead people), among others. For more information, you can go here.

We will filter our working dataframe to include only the following columns:

- **OBJECTID:** an unique ID of the incident.
- **TIME:** The time of the incident.
- **CLASS:** Type of event (Choque= colision , Volcamiento= car turned over, Atropello= running over, Caida ocupante= Passenger or driver fell out of the vehicle)
- **GRAVITY:** Seriousness of the incident (Solo daños= Just material damages, Herido= Injured people, Muertos= dead people).
- **DAY:** Day of the week when the accident took place.

- **LONGITUDE, LATITUDE:** Coordinates of the accident site.

The second dataset contains the coordinates of all the traffic enforcement cameras (speed cameras) in the city. It can be found here It has the following columns:

- **Address:** Adress of the camera location
- **Max Speed:** Maximum allowed speed at that point
- **Longitude, Latitude:** Coordinates of the camera
- **Coordinates:** Grouped longitude and latitude

2.2 Data cleaning and preparing

The data quality in the datasets was high, however, several cleaning operations were made in order to proceed with the analysis. In the main dataframe, columns were renamed and its indexes traduced, and blank spaces were removed in all columns. Letter casing inconsistencies were removed, and a time format was applied to the time column, in order to allow filtering in the next steps. On the speed cameras dataframe, column names were renamed and traduced, and blank spaces were removed.

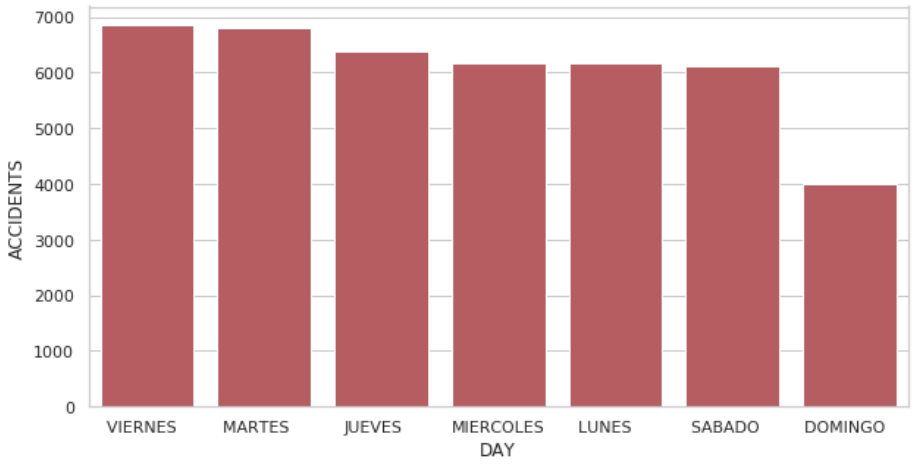
3. Methodology

3.1 Exploratory data analysis

The size of the dataframes was quickly checked. It was determined that there were 42.473 incidents reported in 2019, which represents around 116 accidents per day. We want to see the composition of the accidents, that is, how many involve material damages, injuries and deaths:

	Gravity	Counts
0	HERIDO	23000
1	SOLO DAÑOS	19256
2	MUERTO	217

We can see that more than half of the accidents involve injuries,and fortunately, only a small part of the accidents involve deaths (0.51%). Let's check if the accidents occur more in certain days of the week:



Surprisingly, accidents are well distributed throughout the week, except for Sundays, which have considerably less traffic.

3.2 Traffic accidents vs Nightlife

We want to see the behavior of the data on weekends, in late night/ early morning hours. We will limit our search to fatal accidents, which are the most severe, and look for the size of the resulting dataset:

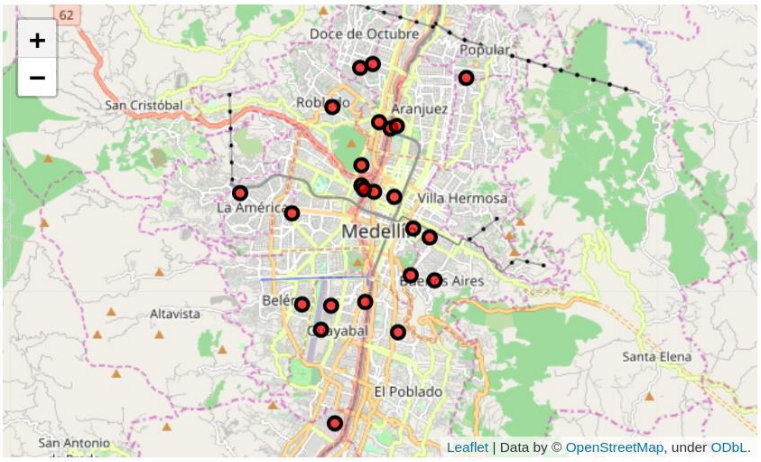
In [14]:
```
df_dead= df2[df2.GRAVITY.str.contains("MUERTO")]

df_dead= df_dead.reset_index().drop(columns=['index'])
df_dead['TIME'] = pd.to_datetime((df_dead['TIME']))
df_dead = (df_dead.set_index('TIME')
          .between_time('22:00:00', '04:00:00')
          .reset_index()
          .reindex(columns=df_dead.columns))

df_dead['TIME']=df_dead['TIME'].dt.time
```

In [15]:
```
search_values=["VIERNES","SABADO","DOMINGO"]
df_wknd= df_dead[df_dead.DAY.str.contains('|'.join(search_values ))]
df_wknd.shape
```

Out[15]: (24, 7)

As we can see, there are only 24 accidents late night/ early morning (from 22:00 to 04:00) on weekends which involve fatal victims. That's around 0.06% of the total accidents. We will visualize the location of these fatal accidents in a map:

Out[18]:



We can see that fatal accidents occurring on weekend nights are almost randomly distributed along the city, with a slight agglomeration downtown.

In the next step, we want to check if there are coincidences between the location of the accidents and the presence of clubs, bars, pubs or other sites where alcohol is consumed, in order to suggest the hypothesis that some of these accidents are caused by drunk drivers. For this task, the Foursquare API was used in order to retrieve the most common close venues to each accident and draw conclusions from the results. This will continue in the results section.

3.3 Relocation of traffic enforcement cameras

Addressing the potential relocation of speed cameras in the city, we refer to the second dataset, containing their coordinates an address, to see the total amount of cameras. There are 77 speed cameras currently working.

We will work under the assumption that zones with high concentration of accidents are more risky and need to be intervened with speed cameras installation. My approach to this problem is to create n clusters of accidents, where n is the total number of cameras currently operating (77). The chosen model for the creation of the clusters is the K-means model, which is one of the simplest available. The K-means model works with centroids that represent the “average point” of each cluster, which in this case refers to a location in the city.

Under these aforementioned assumptions, the centroids of the K-means will correspond to points in the map with many reported accidents nearby, suggesting these are high risk zones that need to be intervened with the installation of cameras. For the accidents clusters, only accidents with injuries or deaths will be considered. In the results sections the findings of the model will be presented.

4. Results

4.1 Traffic accidents vs Nightlife

Out of the 24 fatal accident locations on weekends late night/early morning, there are only three where the closest venue are bars, and there is only one point where the closest venue is a Nightclub. In the following figure a part of the resulting dataframe is shown.

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Com V
0	715213	Gym	Construction & Landscaping	Bus Station	Park	Wine Shop	Eye Doctor	Co Resta
1	717337	Food Truck	Recreation Center	Construction & Landscaping	Dessert Shop	Wine Shop	Eye Doctor	Colon Resta
2	718353	Restaurant	Pizza Place	Bakery	Bookstore	Latin American Restaurant	Burger Joint	Far M
3	718780	Burger Joint	Pizza Place	Café	Housing Development	Sandwich Place	Bookstore	Resta
4	719435	Bakery	Italian Restaurant	Nightclub	Café	Colombian Restaurant	BBQ Joint	Sanc

In the discussion section we will refer back to these results.

4.2 Relocation of traffic enforcement cameras

The sklearn library was used to preprocess the data and set up the K-means model on the accident data, with 77 clusters. A dataframe containing each centroid cluster and coordinates was created.

```
In [41]:
1 centroids= df3.groupby('Cluster').mean()
2 centroids['Cluster'] = centroids.index
3 centroids.head()
```

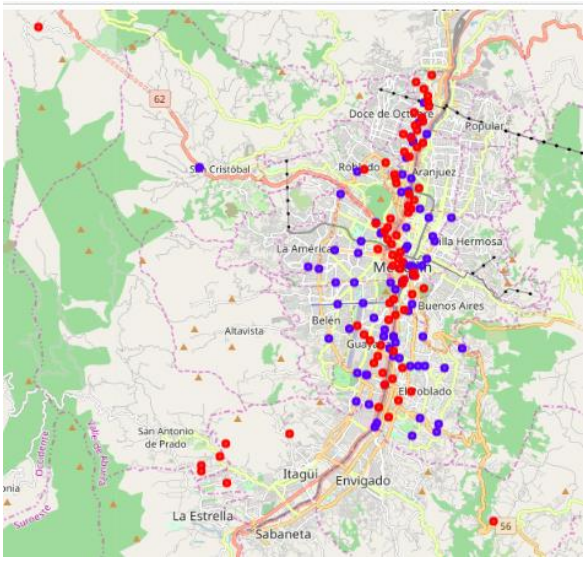
Out[41]:

Cluster	Longitude	Latitude	Cluster
0	-75.574825	6.247128	0
1	-75.568275	6.286070	1
2	-75.579784	6.203342	2
3	-75.578296	6.225906	3
4	-75.571441	6.265524	4

```
In [45]:
1 centroids.shape
```

Out[45]: (77, 3)

The results were plotted as points in the map. Blue dots are the current location of the cameras, while the Red dots represent the suggested location according to the implemented model:



5. Discussion

5.1 Relationship between traffic accidents and nightlife

As It can be seen in the results section, out of the 24 fatal accident locations, there are only three where the closest venue are bars, and there is only one point where the closest venue is a Nightclub.

This doesn't bring any strong evidence that these incidents might be caused by drunk drivers. Moreover, even if the driver was in fact drunk, the accident location might not be close to the place where the person got drunk, or the driver could have drunk in a different place, not a bar or nightclub. In order to have a deeper understanding of the possible causes of each accident, and the proposal of better road safety policies, more data needs to be retrieved, like alcohol breath test results on accident sites.

5.2 Relocation of traffic enforcement cameras

We can see that the suggested new camera locations agglomerate along the most concurred roads across the city, e.g. the zones with higher risk. There are also some new

locations in peripheral neighborhoods, especially in the San Antonio de Prado borough where there have been numerous accidents and currently there are no speed cameras.

We have to make clear that many of the new locations don't correspond exactly to roads, due to the nature of the model that doesn't take into account their location, but give a good approximation of potential locations and has to be interpreted carefully. Also, there are outliers that can be ignored given their extreme location (far away from any road at all).

There are also new locations very close to existing speed cameras, which raise a concern about the true usefulness of speed cameras to make the roads safer and save lives (their presence hasn't stopped accidents). A more robust study should be conducted on the real benefit of this measure.

5. Conclusions

- There is no strong evidence that accidents caused between 22:00 and 04:00 on the weekends are caused by drunk drivers. The location of the accidents by itself it's not enough to draw conclusions about the possible causes of these events.
- In the city of Medellín, there are several zones with high risk of accidents that aren't covered by traffic enforcement cameras, and should be reconsidered in future interventions. In any case, the presence of many of these cameras in the zones with high accident reports raises questions about its usefulness in preventing said accidents. This should be addressed by the competent authorities and policy makers in future efforts for improving road safety.

6. Bibliography

[1] https://www.datos.gov.co/Justicia-y-Derecho/Muertes-violentas-seg-n-sexo-Colombia-comparativo-/h6ai-y2zm