Correlation between different clusters of neighborhoods and their traffic accidents in Medellín

Pineda-Jaramillo JD[1],
[1] PhD, Gobernación de Antioquia, corresponding author. Email: jdpineda@unal.edu.co

Abstract:

The traffic accidents is one of the leading cause of death around the world, especially in low- and mid-income countries. In the case of Medellín (Colombia), the traffic accidents have grown in the last years. For this reason, it is important to develop optimal strategies to reduce those accidents. In that scenario, we present an analysis of the trends in the traffic accidents in Medellín and their correlation with different classes of neighborhoods/zones in the city. To achieve this, we used the foursquare API to understand the composition of the different neighborhoods in the city based on venues, then we performed a *K-means clustering* model to define five different clusters of neighborhoods/zones. After performed a correlation between the clusters of neighborhoods/zones in the city with their respective classes of accidents, we obtained that crashes are the most common traffic accident in all clusters, while places with restaurants, bars, soccer fields and concert halls have more run overs than fallen occupants, and places with food trucks, gyms, bars and stores have more fallen occupants and other classes of traffic accidents than run overs. These insights could be used for the Public Administration to conduct strategies, based on different classes of neighborhoods, to reduce these traffic accidents.

Keywords: *K-Means Clustering, Venues, Clusters, Traffic accidents, Foursquare*

1. Introduction

One of the world's leading causes of death and injuries is urban traffic accidents. According to the World Health Organization (WHO), traffic accidents was the leading cause of death for children and young adults aged 5-29 years, and the eight leading cause of death for all age groups surpassing HIV/AIDS, tuberculosis and diarrheal diseases in 2018 [1]. In addition, more than a million people die each year on the world's roads, and 90% of those deaths occur in low- and middle-income countries that represent 82% of the world's population [1], [2].

For this reason, the United Nations declared 2011-2020 as a decade of action for road safety [3], with the aim of reducing the number of deaths of traffic accidents around the world through national and citywide plans, considering five key strategies [4]:

- Road safety management
- Safer roads and mobility
- Safer vehicles
- Safer road users
- Post-crash response

These key strategies are focused on security and prevention, related to strategies in the medium and long term, where the commitment of nations is a key success factor to achieve the goal of the decade, through the design of management plans including investment, transfer and creation of knowledge to face the current problem and to implement sustainable actions in the future.

Colombia is a middle-income country with 13.8 deaths in traffic accidents per 100 000 inhabitants in 2018, where the rate for high-income countries is around 9. Besides, traffic accidents were the second largest cause of violent death in Colombia for the same year, with 26.7% of the total number of deaths [5].

In Medellín, which is one of the largest cities in Colombia, with more than 2.5 million inhabitants (and almost 3.8 million inhabitants in its Metropolitan Area), there has been an increasing trend in the number of traffic accidents. According to 2014 statistics by the Municipality of Medellín [6], from 2008 to 2014, the number of accidents grew a 20.14%. One of the main causes of this problem is the growing number of vehicles, from 2008 to 2014, this figure increased from 767 548 to 1 234 946 (i.e., by 60.9%).

Looking for optimal strategies with the aim of reducing the traffic accidents in Medellín, it is necessary to analyze different characteristics in the traffic accidents in the city. Hence, in this project we want to analyze the trends in the traffic accidents in Medellín, and figure it out if there is a correlation between these traffic accidents and different classes of neighborhoods (neighborhoods with leisure services, or neighborhoods full of parks and plazas, for example).

To perform a clustering task of different neighborhoods in Medellín, we will use the foursquare API to understand how it is composed the different neighborhoods in Medellín, then we will use a *K-means clustering* model to define the clusters.

The main contribution of this project is to perform a detailed analysis of the traffic accidents in Medellín and its correlation with different classes of neighborhoods in Medellín, and to get some useful insights for the Public Administration to conduct strategies, based on different classes of neighborhoods, to reduce these traffic accidents.

2. Methodology

Data acquisition and cleaning

The geographical information of Medellín, and information related to traffic accidents in the city, can be found in public administrations website. Specifically, we used two main sources to achieve our objectives:

We gathered the traffic accidents information from 2014 to 2019 from the national government website [7], this information contains details of the generalities related to the location of all traffic accidents occurred in Medellín from January 2014 to June 2019. From this dataset we used specifically information related to the following attributes:

- Accident ID
- Address and location of the accident (with coordinates)
- Type of accident

On the other hand, we gathered information of the geographical information related to all neighborhoods and zones of Medellín from the municipality government website [8], this information contains the following attributes, among others:

- neighborhood
- area of the neighborhood
- geometry of the neighborhood (a set of coordinates which surround the area of the neighborhood)

There were a lot of missing values from the data, especially from the traffic accidents data, for this reason, we performed a data cleaning consisted in filling empty values, deleting rows with much lack of information, treating outliers, and normalizing data to the same format.

Foursquare API

Foursquare is a technology company based in New York City that have built a massive dataset of location data around the world. They crowd-sourced their data and had people use their app to build their dataset and add venues and complete any missing information they had in their dataset.

At the year 2018, its location data was the most comprehensive on the internet and quite accurate that it powers location data for many popular services like Apple Maps, Uber, Snapchat, Twitter and many others, and is currently being used by over 100 000 developers, and this number is growing with time [9].

The Foursquare Places API provides location based experiences with diverse information about venues, users, photos, and check-ins. The API supports real time access to places, snap-to-place that assigns users to specific locations, and geo-tag. Additionally, foursquare allows developers to build audience segments for analysis and measurements [10].

Several authors have used foursquare social media geographic information to perform exploratory and spatial analyses of different aspects, like tourist attraction [11], estimation of building block use [12], temporal activity patterns of different venues [13], and to identify opportunity places for urban regeneration through location based social networks [14].

Clustering venues in Medellín using K-means algorithm

K-Means Clustering is an unsupervised Machine Learning technique used in transportation and other industries and is useful for simplifying large datasets by clustering features with similar values into a smaller number of homogeneous categories [15].

K-Means Clustering is used to divide the data into similar groups with similar features, with the aim of maximizing the heterogeneity between clusters (groups) and the similarities between in-cluster samples [16], [17].

Thus, we will use Foursquare API to explore and analyze different neighborhoods and zones in Medellín using the explore function to get the most common venue categories in each neighborhood/zone, and then use this feature to group the neighborhoods/zones into clusters using the *k*-means clustering algorithm to complete this task.

Finally, we will use the Folium library in Python to visualize the neighborhoods in Medellín and their emerging clusters.

Traffic accidents in Medellín

Once we have the neighborhoods/zones of Medellín and their emerging clusters, we will group the number of traffic accidents in every neighborhood/zone to analyze the most common classes of traffic accidents in each cluster.

After pairing the clusters of neighborhoods with their respective number and classes of accidents, we will obtain important insight about the classes of accidents depending on different kinds of neighborhoods/zones.

3. Results and Discussion

Venues in Medellín

Barrio_Vereda.geojson is a geographical information file related to all neighborhoods and zones of Medellín. We built a dataframe in Python with this file using the *geopandas* library. This file consists in a dataframe of 332 rows and 7 columns, as Table 1 presents. We used specifically the information

regarded to neighborhood and the geometry (pair of coordinates that delimitate the region of the neighborhood/zone).

Table 1. Dataframe created using Barrio_Vereda.geojson.

(332, 7)

	OBJECTID	CODIGO	NOMBRE	SUBTIPO_BARRIOVEREDA	SHAPEAREA	SHAPELEN	geometry
0	661	1422	La Aguacatala	1	622090.156105	3302.658052	POLYGON ((-75.5762310714583 6.194621906760724,...
1	662	0810	El Pinal	1	413416.804617	3271.574553	POLYGON ((-75.54160968023029 6.245319337420042...
2	663	0719	Fuente Clara	1	236441.173403	3022.337658	POLYGON ((-75.60106983601668 6.278324025366627...
3	664	0102	Santo Domingo Savio No.2	1	264750.452451	2943.707654	POLYGON ((-75.54062098774783 6.302374662814517...

To get the coordinates of each neighborhood, we used the *Nominatim* library which permits to achieve with great accuracy the longitude and latitude of each neighborhood/zone (Table 2). However, for specific points, a data wrangling was necessary to get these coordinates. After that, we created the map of the neighborhoods/zones of Medellín using the *Folium* library (Figure 1).

Table 2. Neighborhoods with their respective longitude and latitude.

OBJECTID	CODIGO	NOMBRE	SUBTIPO_BARRIOVEREDA	SHAPEAREA	SHAPELEN	geometry	latitude	longitude
661	1422	La Aguacatala	1	622090.156105	3302.658052	POLYGON ((-75.5762310714583 6.194621906760724,...	6.198763	-75.577201
662	0810	El Pinal	1	413416.804617	3271.574553	POLYGON ((-75.54160968023029 6.245319337420042...	6.244487	-75.545597
663	0719	Fuente Clara	1	236441.173403	3022.337658	POLYGON ((-75.60106983601668 6.278324025366627...	6.277633	-75.605593
664	0102	Santo Domingo Savio No.2	1	264750.452451	2943.707654	POLYGON ((-75.54062098774783 6.302374662814517...	6.299473	-75.540104
665	0302	Las Granjas	1	641349.275023	3964.702306	POLYGON ((-75.54371627443042 6.28265929949213,...	6.279339	-75.549459

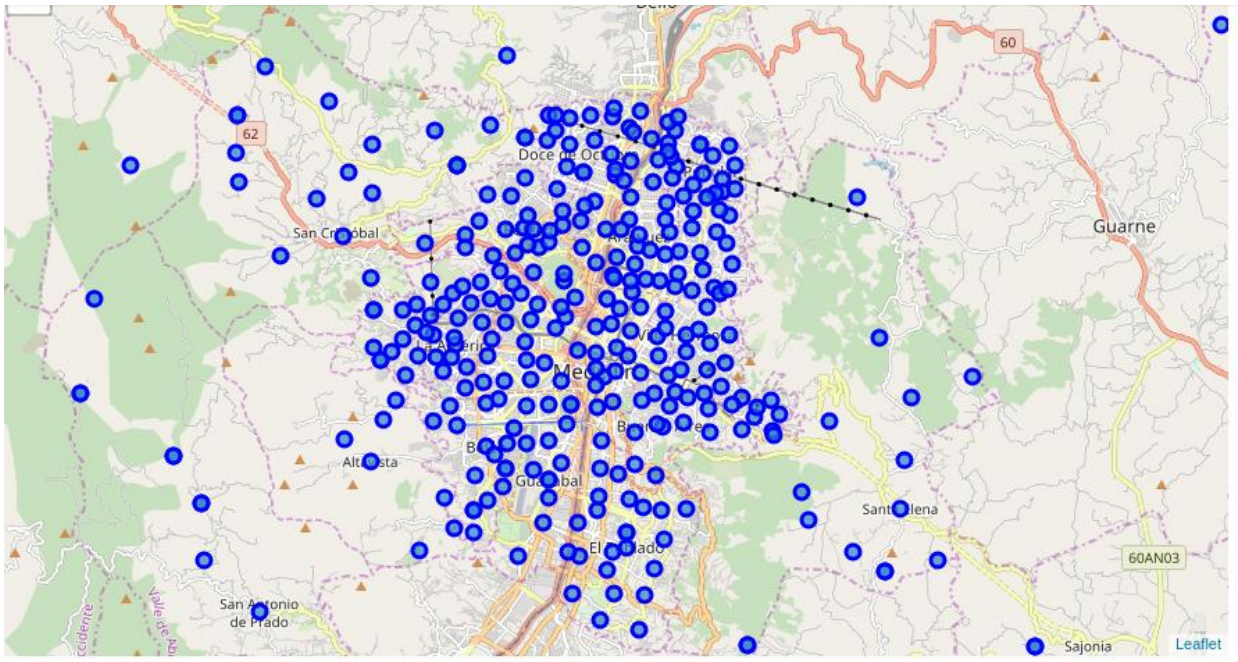

Figure 1. Map of Medellín created using the Folium library with their Neighborhoods/Zones

After defining the location of each neighborhood/zone of Medellín, we explored the Foursquare API to get information related to all the venues of the city of Medellín, achieving a total of 2853 venues with their respective category and location, as we can see in the Table 3 for the first 5 registered venues. There is a total of 237 venues categories.

Table 3. Venues of Medellín and their respective location, category, and neighborhood/zone.

(2853, 7)

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	La Aguacatala	6.198763	-75.577201	Centro Comercial Oviedo	6.199261	-75.575409	Shopping Mall
1	La Aguacatala	6.198763	-75.577201	Mundo Verde	6.199238	-75.575098	Salad Place
2	La Aguacatala	6.198763	-75.577201	Santa Leña	6.199747	-75.575578	Bakery
3	La Aguacatala	6.198763	-75.577201	Mimo's	6.199026	-75.575457	Ice Cream Shop
4	La Aguacatala	6.198763	-75.577201	Medellín Secret	6.198751	-75.575057	Breakfast Spot

Analyzing each Neighborhood

After performing different operations of the data, we obtained the 10 most common venues for each neighborhood /zone, as we can see in Table 4.

Table 4. Ten most common venues for each neighborhood/zone.

	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 Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Aldea Pablo VI	Construction & Landscaping	Restaurant	Doner Restaurant	Food & Drink Shop	Food	Fish & Chips Shop	Fast Food Restaurant	Farmers Market	Farm	Falafel Restaurant
1	Alejandro Echavarría	Shopping Mall	Tram Station	Ice Cream Shop	Gym / Fitness Center	Electronics Store	Event Service	Eye Doctor	Factory	Falafel Restaurant	Zoo
2	Alejandría	Hotel	Shopping Mall	Restaurant	Coffee Shop	Italian Restaurant	Gym	Sushi Restaurant	BBQ Joint	Mediterranean Restaurant	Frozen Yogurt Shop
3	Alfonso López	Burger Joint	Zoo	Donut Shop	Food & Drink Shop	Food	Fish & Chips Shop	Fast Food Restaurant	Farmers Market	Farm	Falafel Restaurant
4	Altamira	Coffee Shop	Breakfast Spot	Art Gallery	BBQ Joint	Zoo	Electronics Store	Food & Drink Shop	Food	Fish & Chips Shop	Fast Food Restaurant

Clustering neighborhoods/zones using their respective common venues

Later, using the K-Means clustering, and after applying the Elbow method for choosing the number of K, we decide to divide our dataset in 5 clusters, as we can see in Figure 2 by different colors.

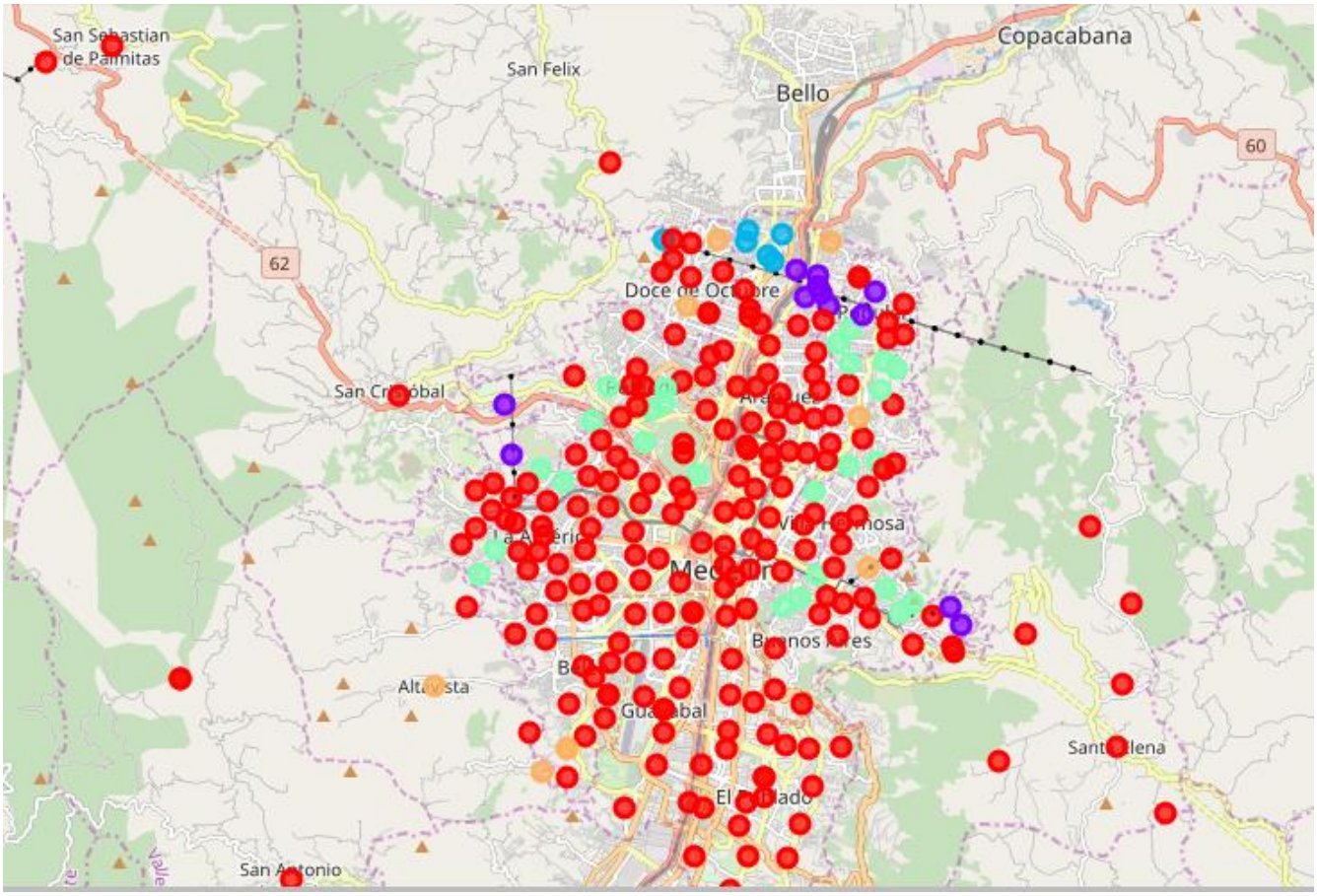

Figure 2. Map of Medellín showing different clusters of neighborhoods/zones by different colors

The 5 clusters found were the following:

- Cluster 0 is the one with neighborhoods full of restaurants, coffee shops and other leisure services.
- Cluster 1 is the one with neighborhoods surrounded by cable car and metro stations. These neighborhoods have restaurants as well.
- Cluster 2 is the one with neighborhoods which have food trucks and bars. Also, these neighborhoods have other leisure places like gyms and stores.
- We can define Cluster 3 as places with parks and plazas.
- Cluster 4 is the one with neighborhoods offering other kind of services, as zoos, soccer fields and concert halls.

Analyzing traffic accidents

Accidentalidad_georreferenciada_2014-2019.csv is a dataset which contains details of the generalities related to the location of more than 226 000 traffic accidents occurred in Medellín from January 2014 to June 2019. Below we present in the Table 5 the first rows of the features we used for the analysis and in the Figure 3 the location of 2% of random traffic accidents from the dataset (we just present 2% of the dataset due to limitations in memory ram, but the analysis was carried out using 100% of the dataset).

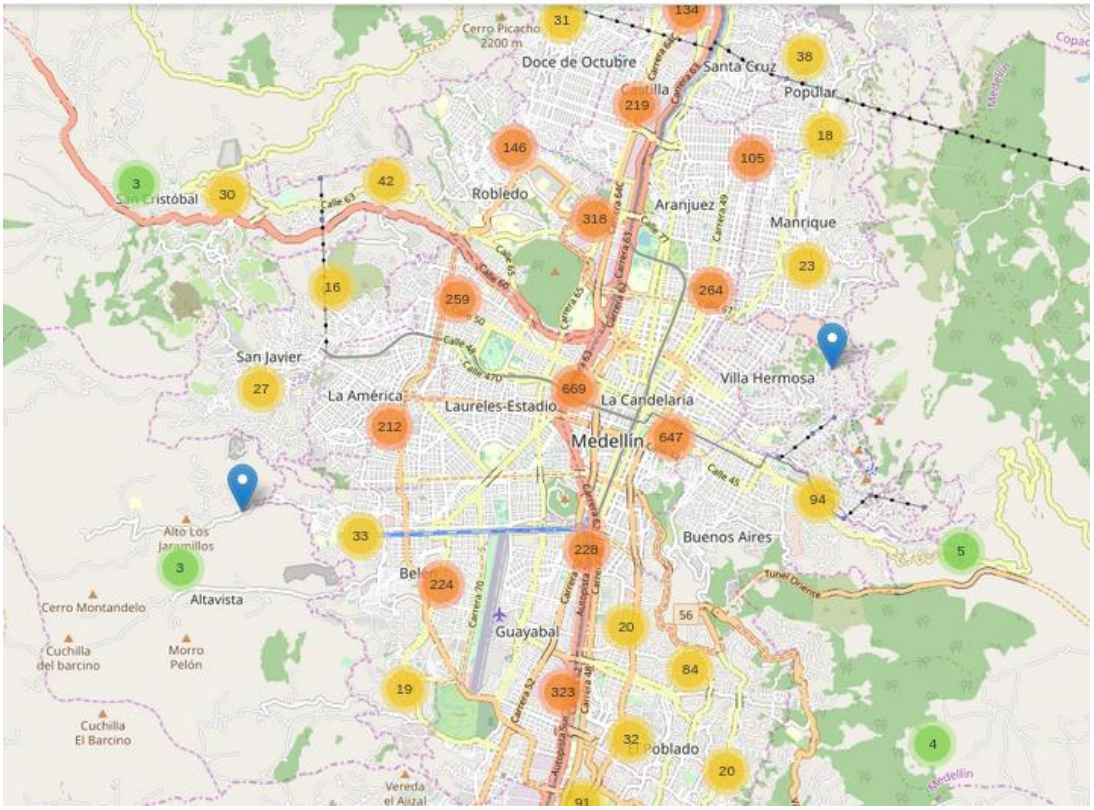

Figure 3. Location of 2% random traffic accidents in Medellín in the period of 2014-2019.

Table 5. Features used for the analysis of traffic accidents (first five rows)

	class	longitude	latitude
0	crash	-75.575177	6.256572
1	crash	-75.566388	6.246808
2	other	-75.574233	6.206502
3	other	-75.619871	6.249727
4	other	-75.592195	6.282603

Later, we used the coordinates of each traffic accident to assign them to each neighborhood/zone of Medellín using column "geometry" from Table 2, then awe grouped how many traffic accidents were inside each neighborhood/zone of each class, as we show in Table 6 for the first 5 neighborhoods.

Table 6. Number of traffic accidents by class for each neighborhood/zone in Medellín

neighborhood	crash	fallen_occupant	other	overturning	run_over
Aguas Frías	9	3	6	1	2
Aldea Pablo VI	25	12	6	2	19
Alejandro Echavarría	130	23	38	11	38
Alejandría	408	18	23	9	14
Alfonso López	714	198	201	55	148

Correlation between traffic accidents and venues in Medellín

After having the different clusters of neighborhoods/zones with their respective venues categories, and their respective number of accidents by class, we merged both datasets to obtain the Table 7.

Table 7. Merged dataframe showing the number of traffic accidents by class, and the cluster label for each neighborhood/zone in Medellín

neighborhood	crash	fallen_occupant	other	overturning	run_over	Cluster Labels
Aldea Pablo VI	25	12	6	2	19	0
Alejandro Echavarría	130	23	38	11	38	0
Alejandría	408	18	23	9	14	0
Alfonso López	714	198	201	55	148	0
Altamira	207	34	40	12	23	0

Later, we calculated the mean number of traffic accidents by class by cluster labels as we show in Table 8.

Table 8. Mean number of traffic accidents by class by cluster labels

Cluster Labels	crash	fallen_occupant	other	overturning	run_over
0	1156.239766	136.953216	165.000000	49.526316	158.076023
1	189.636364	46.181818	50.454545	14.636364	74.181818
2	267.166667	61.666667	56.166667	17.833333	46.166667
3	238.375000	42.750000	51.250000	15.083333	47.458333
4	189.625000	44.750000	53.625000	16.000000	68.375000

Finally, we analyzed the percentage of each class of traffic accident per each cluster of neighborhoods/zones in Medellín, obtaining the Figure 4.

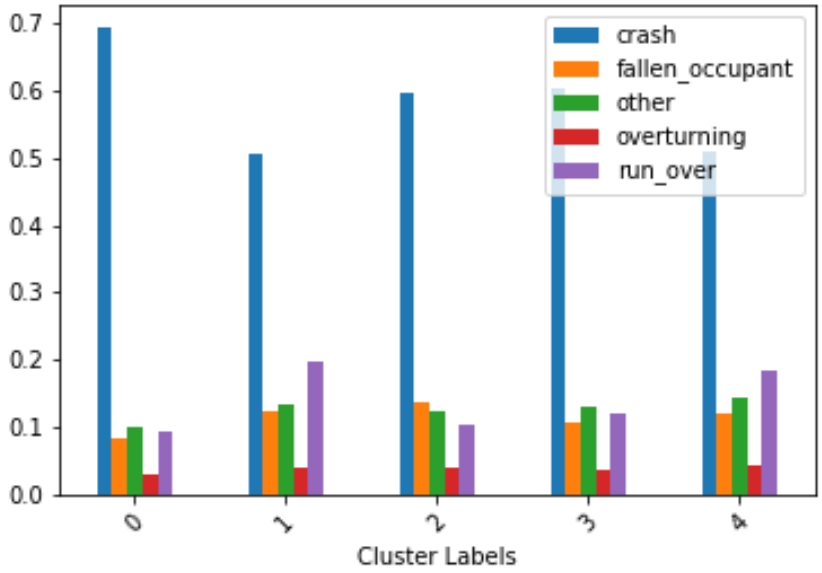

Figure 4. Percentage of each class of traffic accident per each cluster of neighborhoods/zones in Medellín.

From the Figure 4, we can conclude the following:

- Crashes are the most common traffic accidents in every cluster
- Clusters 1 and 4 have more run_overs than fallen_occupants and other kind of traffic accidents (neighborhoods with restaurants, bars, soccer fields and concert halls).
- Cluster 2 has more fallen_occupants and other kind of traffic accidents than run_overs (places with food trucks, gyms, bars and stores).
- Overturning is the least frequent traffic accident in all neighborhoods

The code and data can be found in the following repository: https://github.com/jdpinedaj/Coursera_Capstone

4. Conclusions

From the correlation between the different clusters analyzed and their respective classes of accidents, it is important to note that crashes is the most common traffic accident by far in each cluster.

On the other hand, two important insights were noted. First, the places with restaurants, bars, soccer fields and concert halls have more run overs than fallen occupants, and places with food trucks, gyms, bars and stores have more fallen occupants and other classes of traffic accidents than run overs.

It is important to mention that run overs and fallen occupants generally have a major severity than crashes, and these traffic accidents are common in places with leisure services. For this reason, it is necessary that Public Administration to conduct strategies, based on different classes of neighborhoods, to reduce these traffic accidents and their severity.

As future works, it is possible to analyze if it exists some implications in the infrastructure elements and the severity of traffic accidents, for working in improvements in these elements to reduce the severity of traffic accidents.

5. References

[1] World Health Organization, "Global Status Report on Road Safety," 2018.

[2] F. Wegman, "The future of road safety: A worldwide perspective," *IATSS Research,* vol. 40, no. 2, pp. 66-71, 2017.

[3] United Nations, "Resolution 64/255. Improving Global Road Saftey," NY, 2010.

[4] United Nations, "Global Plan for the Decade of Action for Road Safety 2011-2020," NY, 2011.

[5] Instituto Nacional de Medicina Legal y Ciencias Forenses, "Forensis 2018 - Datos para la Vida," Bogotá, 2018.

[6] Alcaldía de Medellín, "Informe anual de accidentalidad 2014," Medellín, 2014.

[7] Ministerio TIC - Colombia, "www.datos.gov.co," [Online]. Available: https://www.datos.gov.co/browse?Informaci%C3%B3n-de-la-Entidad_Departamento=Antioquia&Informaci%C3%B3n-de-la-Entidad_Municipio=Medell%C3%ADn&q=transporte&sortBy=relevance. [Accessed 12 February 2020].

[8] Alcaldía de Medellín, "Alcaldía de Medellín OpenData," [Online]. Available: https://geomedellin-m-medellin.opendata.arcgis.com/datasets/c844f0fd764f41b2a808d8747457de8a_4. [Accessed 12 February 2020].

[9] IBM, "IBM Coursera," [Online]. [Accessed 12 Febrary 2020].

[10] BaluNaik - FindNearMe, "github," [Online]. Available: https://github.com/BaluNaik/FindNearMe. [Accessed 12 February 2020].

[11] P. Fernández-Costa, I. da Silva, R. Ribeiro and J. Mercedes, "Strategy for extraction of Foursquare's social media geographic information through data mining," *Bulletin of Geodetic Sciences,* vol. 25, no. 1, 2019.

[12] S. Spyratos, D. Stathakis, M. Lutz and C. Tsinaraki, "Using Foursquare place data for estimating building block use," *Environment and Planning B: Urban Analytics and City Science,* vol. 44, no. 4, pp. 693-717, 2017.

[13] K. D'Silva, A. Noulas, M. Musolesi, C. Mascolo and M. Sklar, "Predicting the temporal activity patterns of new venues," *EPJ Data Science,* vol. 7, no. 13, 2018.

[14] P. Martí, C. García-Mayor and L. Serrano-Estrada, "Identifying opportunity places for urban regeneration through LBSNs," *Cities,* vol. 90, pp. 191-206, 2019.

[15] J. Pineda-Jaramillo, "A review of Machine Learning (ML) algorithms used for modeling travel mode choice," *DYNA,* vol. 86, no. 211, pp. 32-41, 2019.

[16] J. Hair Jr, W. Black, B. Babin and R. Anderson, Multivariate Data Analysis, Harlow, UK: Seventh Ed., Pearson, 2014.

[17] C. Fraley and A. Raftery, "How many clusters?. Which Clustering method?. Answers via model-based cluster analysis," *The Computer Journal,* vol. 41, no. 8, pp. 578-588, 1998.