Classification of Nighttime Accidents based on nightlife spots in the area

Amir Iqbal

June 3, 2021

1. Introduction / Business Problem

The goal of this analysis is to assess, if there is a relationship between the number of late-night accidents and night life activities in the Toronto area. Usually, the bulk of the traffic is in the day when people are travelling to and from their work or they are going out to buy groceries, shopping etc. In general, the traffic is less in the night and less traffic should mean less traffic accidents.

In this project I will analyze, if the number of night life spots in the Toronto area increases the number of accidents in the area at the nighttime or not. The accident data will be divided into daytime accidents and nighttime accidents, and I will focus on the nighttime accidents. The analysis will focus only on the business related to night life activities, like bars, pubs, night clubs etc.

I will use Foursquare API to identify the number of nightlife spots in the Toronto area. The data for nightlife spots will be collected at the postal codes level. The accidents data will also be rolled up at the postal codes level.

* 1. – Problem

The nighttime activities are usually less during the weeknights, but they are comparatively more on the weekends. This project will compare the number of nighttime accidents during weeknights and weekend nights. A higher number of accidents at weekend nights in the areas with higher number of nightlife spots will indicate, nightlife spots may be a contributing factor of higher accidents in the area. This project will determine the impact of nightlife spots on the weekend night accidents and classify the Toronto’s postal districts based on the nighttime accidents and nightlife spots.

* 1. – Potential Users of project

The nightlife business like bars, night clubs, pubs, karaoke bars, etc. can use the analysis done in this project to take additional steps to ensure the safe returns of their customers. The accidents rate seems to be higher during weekend. Business especially over the weekend can offer deals which include vouchers for ride sharing services. The businesses can work with the ride sharing or cab services to ensure easy availability of these services. The area police can use the results to identify in which areas and what days of week the police need to increase the patrol. The health and safety departments of the highly affected neighborhoods can run public awareness campaigns.

2. Description of Data

Data from four different data sources will be retrieved, cleaned, and merged using postal district as the common field among all the data.

**Accidents Data** will be retrieved from the [Toronto open data source portal](https://ckan0.cf.opendata.inter.prod-toronto.ca/download_resource/6719b900-1bf9-4678-b9a0-547e30a24846?format=csv&projection=4326). The data includes accidents data from 2006 till the end of 2019. The accident data is before the start of covid pandemic, so no impact of mandatory closing of businesses due to covid. The data includes the exact time when the accident happened, the location of accident, etc. Accident details are available for more than 16000 accidents over a period of 14 years.

Toronto’s [**Postal Districts Data**](https://cocl.us/Geospatial_data) with the longitude and latitude is available on the internet. The postal districts will be used to define regions in the Toronto city. The analysis will collect and summarize the data at the postal district level.

T**oronto's Neighborhood and Borough's Data** is retrieved from the [WIKIPEDIA page](file:///C:\51%20Projects\Capstone-Project\Doc\WIKIPEDIA%20%20page). This data will allow us to see results at the Borough's level.

**FOURSQUARE** has the data on all the night life venues organized under categories and subcategories. All the night life venues can be found under one main category using the API. The FOURSQUARE API will be used to search the businesses under the category ‘Nightlife Spot’ in each of Toronto’s postal district.

3 - Data Acquisition and Data Cleanup

Data from four different sources will be loaded and cleaned up. The data sources are as follow.

3.1 - Accident Data

Serious Accident data for Toronto area maintained by city government and available from [Toronto open data source portal](https://ckan0.cf.opendata.inter.prod-toronto.ca/download_resource/6719b900-1bf9-4678-b9a0-547e30a24846?format=csv&projection=4326). There are more than 16000 accidents records available in a CSV format. Accident data is available for 14 years starting from 2006 to 2019. Data includes detailed information for each accident, including the date and time, geo coordinates of the accident, district, and neighborhood. Selected eight columns which are relevant for this project. A very small number of records were missing some of the fields, removed the records which are missing key values.

The accident data is missing the postal district of the accident. It includes geo coordinates of the accident, using the geolocator reverse engineered and added the postal district to the accident record.

The project is analyzing the relationship between weekend night accidents and nightlife spots in the area. Added additional columns in the accident dataframe to identify the day of the week and classify the accident as daytime or nighttime.

Accidents which happened from 9 PM to 4 AM were classified as nighttime accidents. All other accidents classified as daytime accidents. Classified accidents as weekend accident, if it happened from Friday to Sunday night, the remaining days classified as weekday nights accidents. Added the column for postal district “FSA” into the dataframe, this column is derived using latitude, longitude and geolocator reverse API. Created a dataframe holding only the nighttime accidents for all days of week.

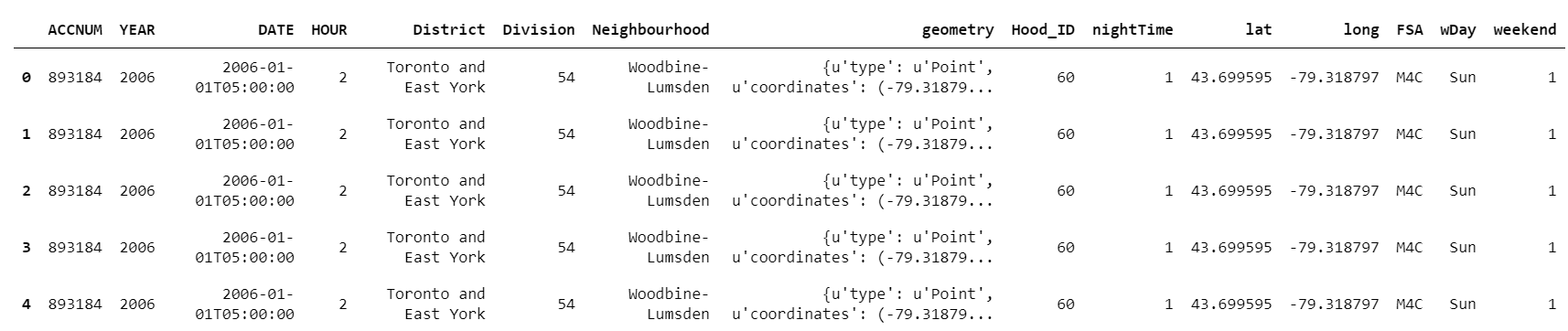


Figure Nighttime accident dataframe

Generated a summary dataframe summarizing all the nighttime accidents by the postal district (FSA). The summarized dataframe includes columns total nighttime accidents, accidents on weekend nights, and accidents on weekday nights by the postal district.

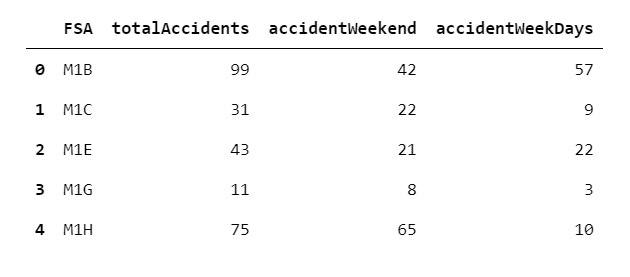


Figure Nighttime accidents summary by Postal District

3.2 - POSTAL Districts

The geo coordinates for the center of the postal districts are required to search the venues using the FOURSQUARE API. Retrieved the postal districts along their longitude and latitude from the [external file](https://cocl.us/Geospatial_data).

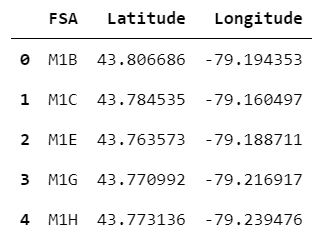


Figure Postal codes with Latitude and Longitude

3.3 - Toronto Boroughs & Neighborhoods

Toronto city has many boroughs and neighborhoods, to identify which postal district falls under which borough and which neighborhoods are covered by the postal districts, retrieved the data from the [WIKIPEDIA page](file:///C:\51%20Projects\Capstone-Project\Doc\WIKIPEDIA%20%20page).

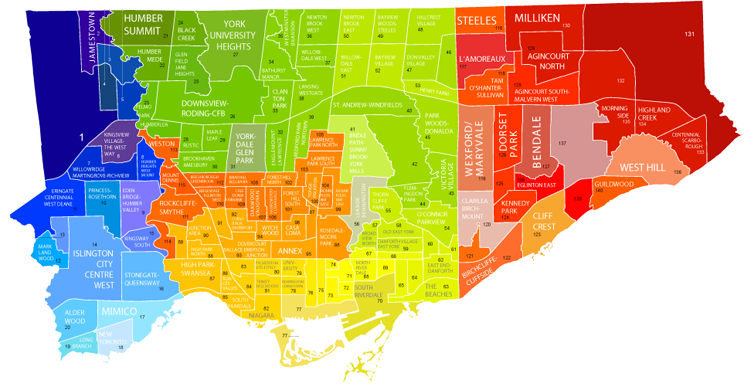


Figure Toronto Boroughs and Neighborhoods

Scraped the data from Wikipedia page using the “BeautifulSoup” package. After scrapping the data from the Wikipedia page, cleaned and filtered the data. Created a dataframe containing postal district(FSA) , Borough and Neighborhood.

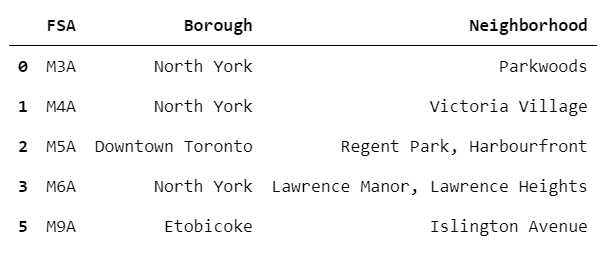


Figure Boroughs and Neighborhoods dataframe

3.4 – Nightlife Spots

FOURSQUARE organize venue data under categories, and one of the main categories is “Nightlife Spots”, venues related to bars, pubs, nightclub, etc. are listed under this main category. For all the postal districts, queried the FOURSQUARE using the postal district longitude, latitude and using category ID. Restricted the results only for the venues which fall under the “Nightlife Spots” category.

Created one dataframe of all venues, found under the “Nightlife Spots” category. Added the postal district and subcategory to the venues dataframe.

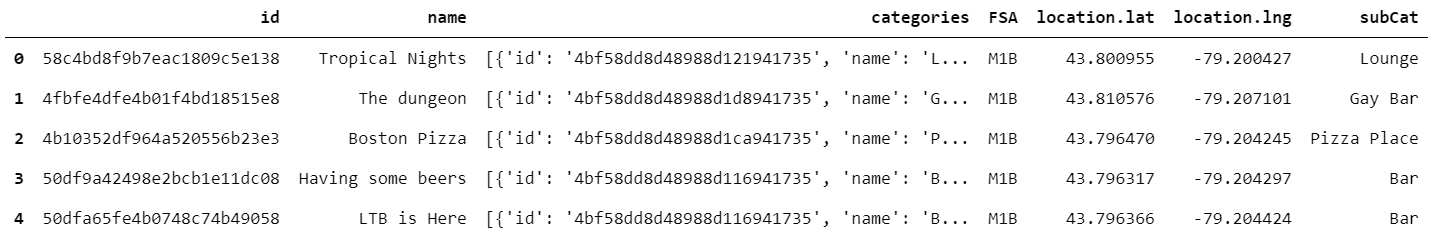


Figure All venues under the “Nightlife Spots” category in the city of Toronto

Next derived a summary dataframe of “Nightlife Spots” venues by the postal districts.

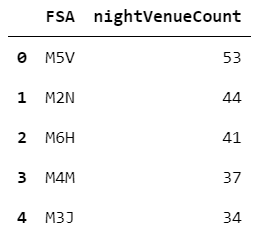


Figure Summary of venues by the postal district

3.5 – Combine data

Data from four data sources have been retrieved, cleaned, and have created detailed and summarized data frames. Postal district is the common key in all the summarized data frame. Next, I will combine all summarized data frames into one combined data frame. This combined dataframe will include all the features required to do the analysis.

Diagram

Description automatically generated

Figure Summarized combine dataframe generated from 4 different sources.

The combined dataframe has venues, accidents data, borough’s, neighborhood all by postal district in one row. This summarized combine data completes the process of acquiring, cleaning, and organizing data.

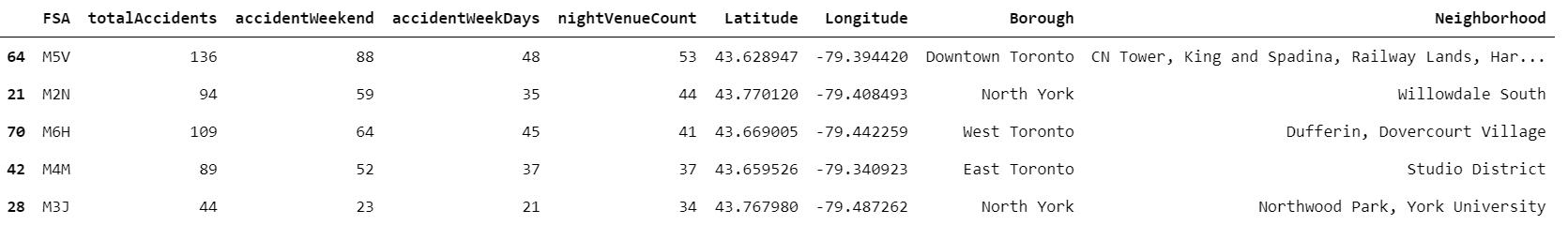


Figure Combined summarized data for all the features.

4. Exploratory Data Analysis

Data from all four sources have been combined into one dataframe. Next, I will evaluate what sort of relationship exists between the key data fields of the combined data.

4.1 Day time Accidents vs Nighttime accidents

The accident data contains the date and exact hour of the accident. Each accident has been classified as nighttime or daytime. Accident will be a nighttime accident, if the accident time is between 9 PM and 4 AM. All other accidents are daytime accidents. Out of 24 hours, 8 hours have been classified as nighttime and remaining 16 hours are classified as daytime. The number of accidents in daytime in general are higher and especially with double the number of hours in daytime, a higher number of accidents are expected in the daytime.

The graph below shows the daytime accidents vs nighttime accidents over the seven days of week.

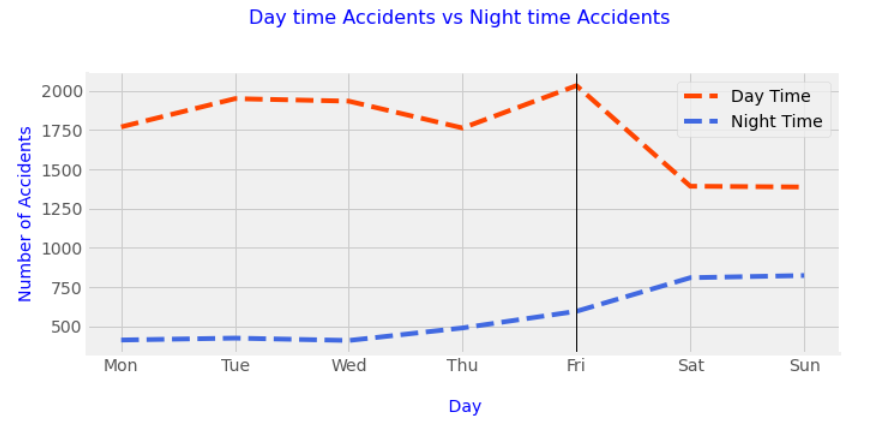


Figure Daytime accidents vs Nighttime accident by day of the week

The graph shows number of accidents in the daytime are higher than nighttime accidents as expected. Data also shows that starting from Friday daytime accidents start dropping from the peak. Saturday and Sunday lowest number of accidents, it does make sense as number of people travelling for work drops significantly over the weekend.

The nighttime accidents increase starting from Friday and maximum number of nighttime accidents are over the weekend. Weekend travel is usually not work related, it is more for personal errands and leisure. Especially the nighttime traffic is usually for leisure and fun activities.

Accident data is from 2006 to 2019, before covid-19 lock downs. The 14 years of data clearly shows the trend, number of nighttime accidents goes up during weekend nights.

4.2 - Accidents in Postal Districts with highest number of Nightlife Spots

Next, I will evaluate the trend of nighttime accidents in the 10 postal districts with the highest number of nightlife spots like bars, pubs, nightclub. This nightlife spots data has been retrieved using the FOURSQUARE API for each of the postal district.

Accidents on three nights (Fri, Sat, Sun) are counted as weekend night accidents and the remaining 4 nights are counted as weekday nights. Although weekday nights are more than weekend nights, eight out of the top ten postal districts have more nighttime accidents on weekend nights compare to weekday nights. Only two postal districts (M4J, M4E) have slightly more weekday nighttime accidents compared to weekend night accidents. Number of accidents in these two postal districts (M4J, M4E) in general are lowest compared to the other districts in the top 10 group.

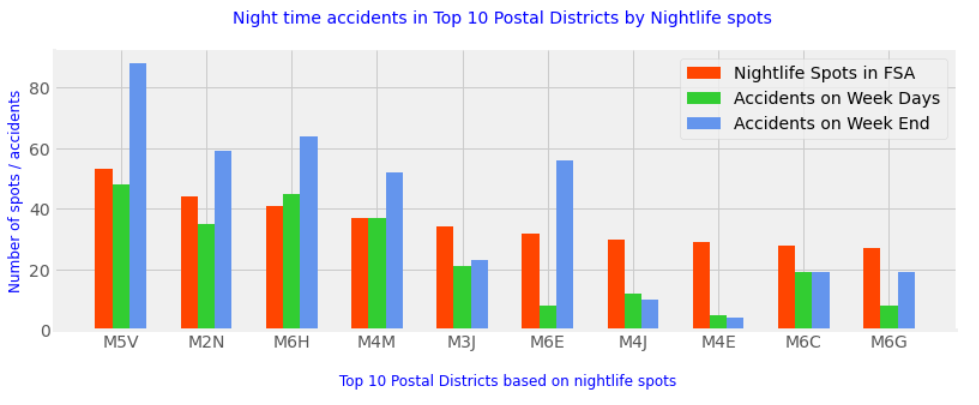


Figure Nighttime accidents in 10 postal districts with highest number of nightlife spots

The data in the ten districts with the highest number of nightlife spots, does show a trend that nighttime accidents on the weekend nights are more than weekday nights. The higher number of nightlife spots in these top 10 postal districts seems to be a contributing factor of higher number of nighttime accidents.

4.3 - Nighttime Accidents by each Borough

The top 10 postal districts indicate a relationship between the number of nightlife spots and number of weekend night accidents. Let’s now analyze this relationship at the Toronto city level. The city is divided into nine boroughs, and the graph below shows the number of nightlife spots, number of weekday night accidents and number of weekend night accidents for each brough.

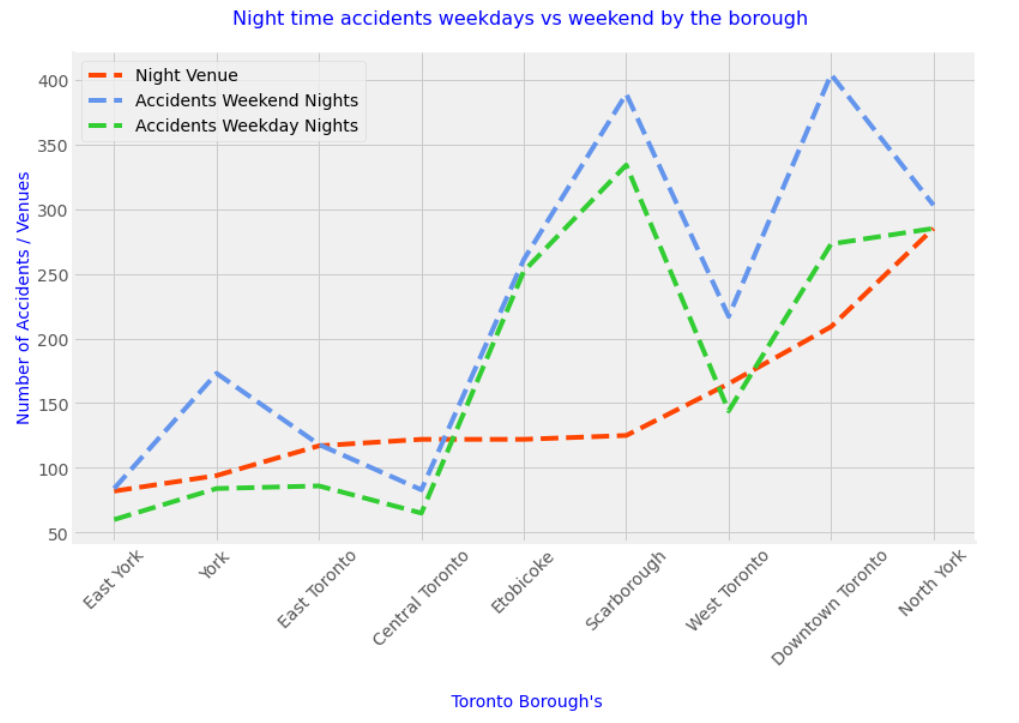


Figure Nighttime accidents by the boroughs

The graph shows that, number of accidents increases as the number of nightlife spots increases. The weekend night accidents are more than weekday night accidents in all the nine boroughs. The data confirms a higher number of nightlife spots in a borough contributes to higher number of accidents and the weekend night accidents are more than weekday night accidents.

5. Modeling

Exploratory data analysis indicates number of nightlife spots in a postal district increases accidents on weekend nights. Let’s run an unsupervised algorithm K-means to classify the postal districts.

5.1 Selected Features

I will be using following three features to run the **k-means** classification algorithm. The numbers for these features have been rolled up at the postal district level.

1. Accidents Weekday Nights
2. Accidents Weekend Nights
3. Nightlife Spots

5.1 - Standardized the Data using the Standard Scaler

Before running the K-means algorithm, the numerical data needs to be standardized. It will ensure the feature(s) with bigger values do not dominate the results. I used the StandardScaler to standardized the data for the three features selected.

5.2 Determine Optimal Value for K

Best results from k-means algorithm can be achieved by using the optimal value for the input parameter number of clusters (k). I will use the Elbow method to find the best value for the input parameter number of clusters (k).

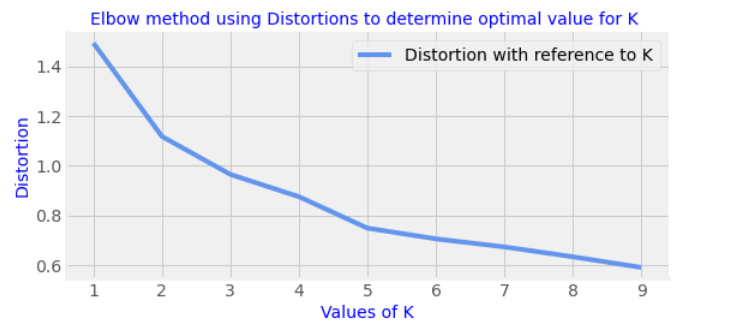


Figure Distortion with reference to K

Ran the k-means algorithm on the standardized data. Ran the k means algorithm for different values of k from 1 to 10 and collected the distortion value for each of the K. The graph above shows the value of k and corresponding distortion value. The value of k at 4 is the optimal value based on the shape of the graph. I will use 4 as input parameter to classify the data into clusters.

5.3 - Run K Means Algorithm on the Data

Data is standardized, and optimal value for K has been determined using the meth. Let’s now run the k-means unsupervised classification algorithm.

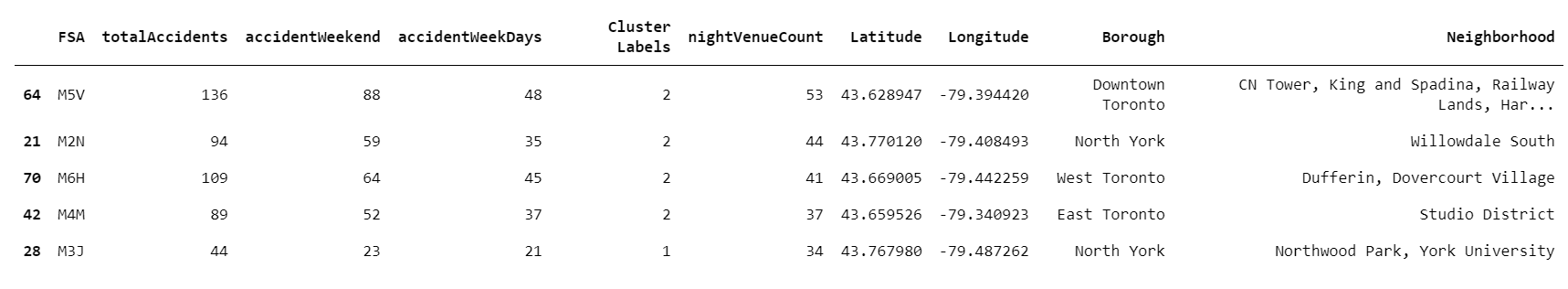


Figure Combine dataframe with cluster labels added.

Ran the k-means algorithm, model classified the postal districts based on the three selected features into four clusters. Added the cluster label to each row in the combined dataframe.

6 - Results & Discussion

I will now review and discuss the results of classification algorithm. The model has divided data into four clusters. The table below summarize the results by the cluster.

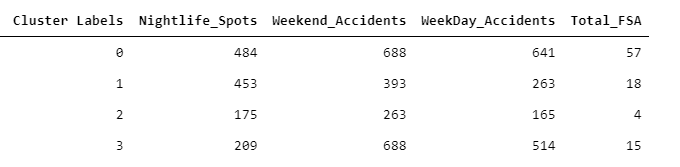


Figure Summarized results by cluster

The table above shows rolled up numbers by each cluster for all the three features. The table also shows how many postal districts were selected in each cluster.

The data from the above table is shown as a graph for easy interpretation of data. The graph is showing the three selected features by each cluster.

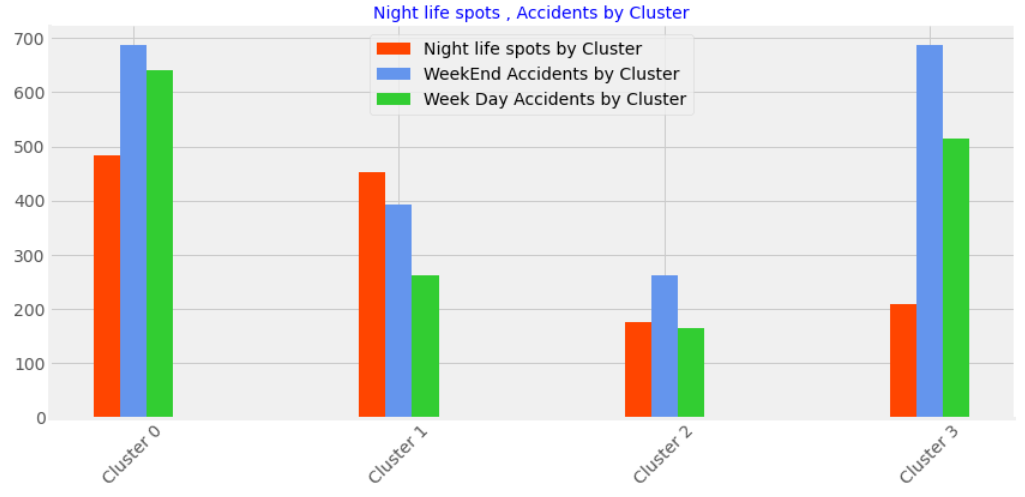


Figure Accident, venue counts by cluster.

6.1 Postal Districts by Clusters Map

Below figure shows the map of Toronto with the distribution of postal districts by cluster. Each circle in the map represents a postal district. Each cluster has its own unique color, and the number inside the circle represent the cluster number.

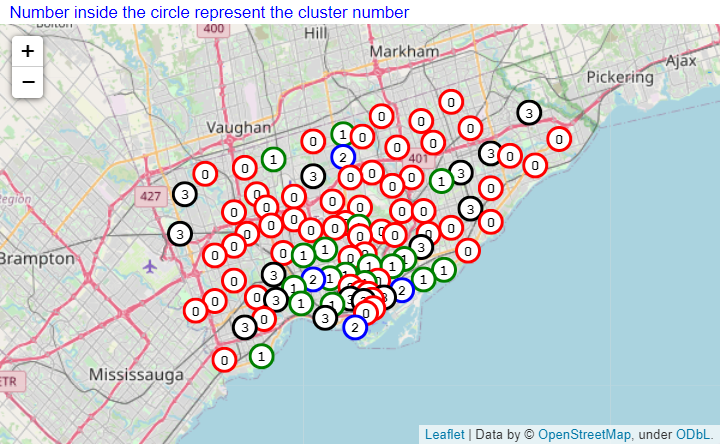


Figure Postal districts in the Toronto city with their cluster label

6.2 Results by Cluster

Modeling algorithm has grouped all postal districts into four clusters. Next, I will discuss how data is organized in each cluster with respect to average numbers. Calculated the average number of accidents, venues, and postal districts from the data for all the four clusters. It will allow us to see how each specific cluster with respect to average numbers did.

6.3 Cluster 0 Results

All numbers in Cluster 0 are more than average per cluster. Cluster 0 has the highest number of total accidents across all clusters. The Weekend Accidents are more than Weekday Accidents and almost half of the postal districts are in cluster zero.

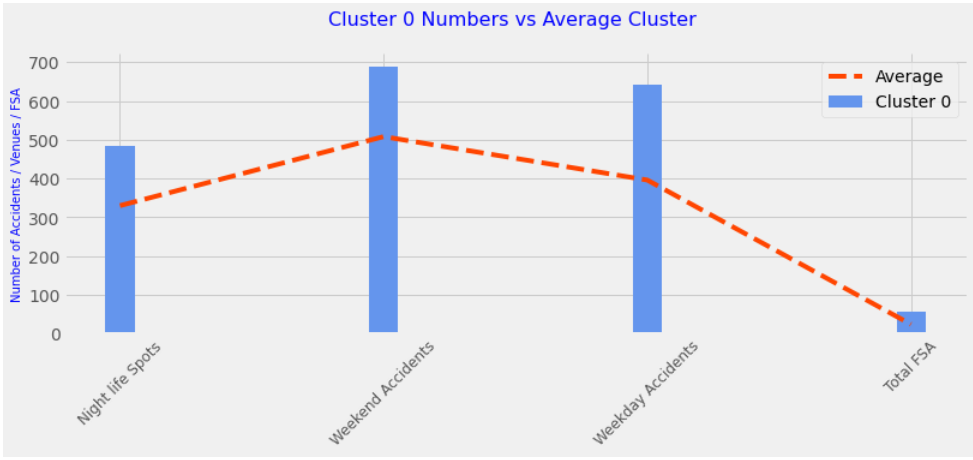


Figure Cluster 0 numbers with respect to average cluster numbers.

6.4 Cluster 1 Results

Cluster 1 has relatively higher number of Night life spots. The weekend accidents or weekday accidents both are less than average per cluster. The weekend accidents are more than weekday Accidents.

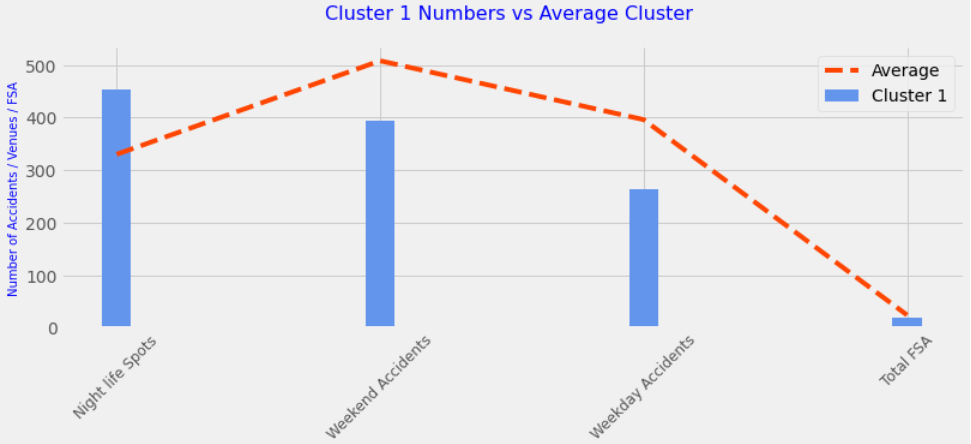


Figure Cluster 1 numbers with respect to average cluster numbers.

6.5 Cluster 2 Results

All numbers in Cluster 2 are significantly less than average per cluster. Cluster 2 has smallest number of night life spots. The weekend accidents are more than weekdays accidents.

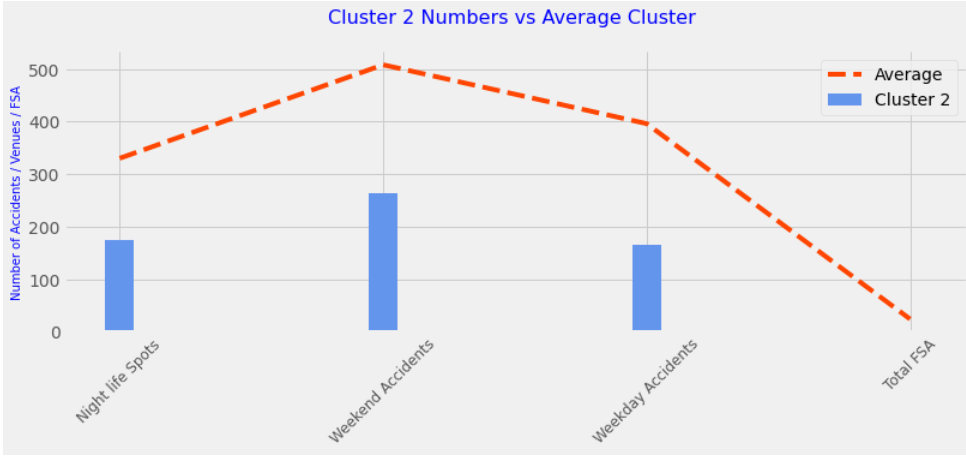


Figure Cluster 2 numbers with respect to average cluster numbers.

6.6 Cluster 3 Results

Cluster 3, relatively small number of night life spots. The weekend accidents, and weekday accidents are more than average. The weekend accidents are more than weekday accidents.

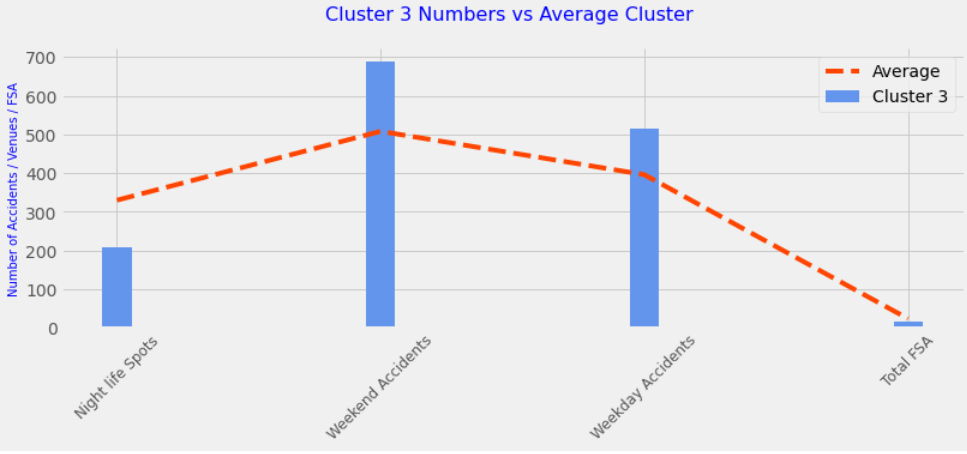


Figure Cluster 3 numbers with respect to average cluster numbers.

7. Conclusion

In this project, I analyzed the 14 years of nighttime accident data for the city of Toronto, and impact of nightlife spots in the city on the number of nighttime accidents. The data shows daytime accidents which are significantly higher during the weekdays dropped over the weekend and the nighttime accidents went up over the weekend.

Nighttime accidents which were low during the weekday went up significantly on the weekend. The data confirms number of accidents on weekend nights are higher than weekday nights. Analysis further shows that accidents in the postal districts with higher number of nightlife spots have significantly more nighttime accidents compare to postal districts with low number of nightlife spots. The analysis further shows that nighttime accidents on the weekend nights are higher compared to weekday nights.

I used the classification modeling to group the postal directs based on number of accidents and nightlife spots. The nightlife business in the postal districts with higher number of weekend night accidents can take additional steps to ensure the safety of their customers. The area police can use the analysis to identify in which postal districts and what days of week the police need to increase the patrol. The health and safety departments of the highly affected postal districts can run public awareness campaigns.