Predicting Fatal Accidents by Location

By: Edwin Johnson

1. Introduction

1.1 Background

Accidents are one the highest probability causes of death, comparable to fatalities caused by firearms in the United States. It is therefore reasonable and expected that governments and individuals take the necessary precautions and make investments to infrastructure in order to prevent accidental deaths. To optimize for this reduction, it becomes necessary to understand the features that cause fatalities and to explore possible reasons or correlated factors that make any accident a fatal one.

1.2 Problem

This project aims to find whether the location of the fatal accidents and their nearby venues affect whether the accident will become a fatal one.

1.3 Interest

Government officials would be interested as previously stated to decrease or even eliminate fatalities. If an accident is clustered around a certain area it would be in that government officials' best interest to act or improve infrastructure at that location. Individuals as well would be interested in obtaining the knowledge to avoid being another statistic. One assumption would be that most fatal accidents happen on highways or interstates. Another by the airport or near bus stops or where there would be greater pedestrian traffic. This project aims to discover if these factors play a role in fatalities.

2. Data

The datasets this project will use is from Kaggle dataset Killed or Seriously Injured (KSI) Toronto and the Foursquare API for location and venue details.

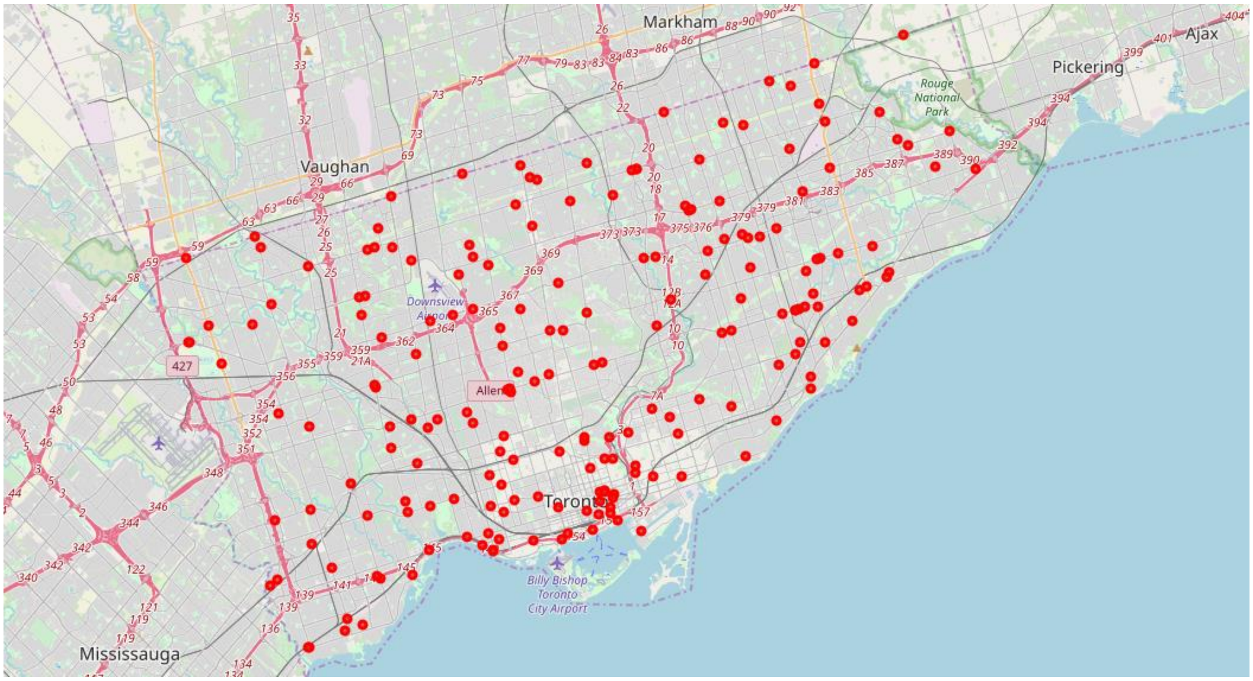
The Kaggle dataset has the location in latitude and longitude, date, time, types of accidents such as pedestrian, vehicle, and if the accident caused a fatality.

These features in addition to the nearest venues in 100 meters of the location of the accident would be used to make predictions as to where fatalities are clustered and the types of the accidents.

3. Methodology

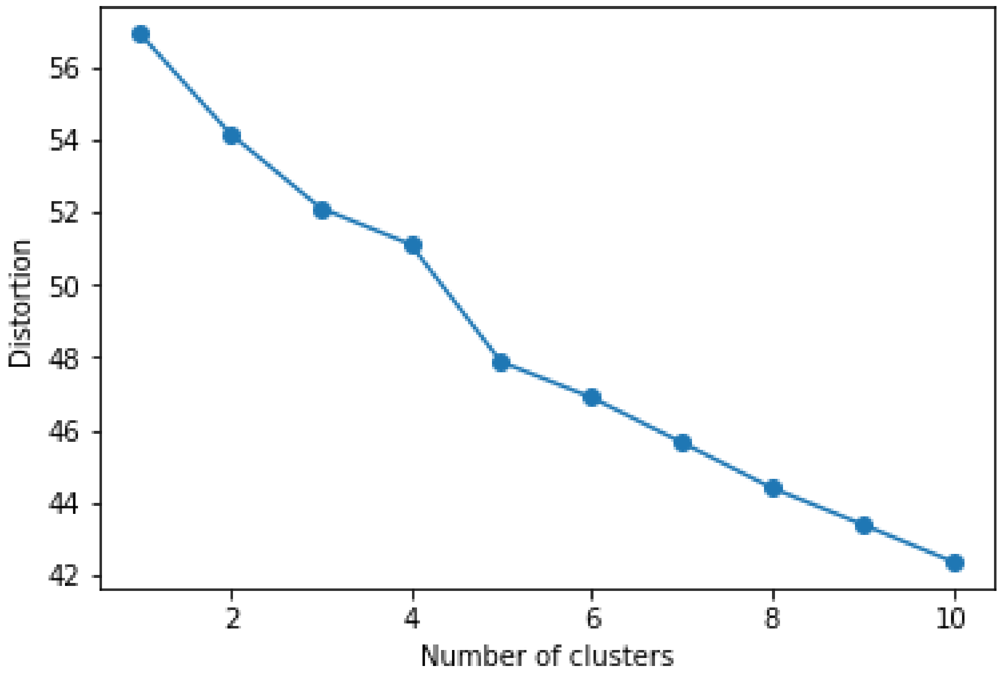
Exploring the KSI dataset shows that it is thoroughly cleaned and has no missing value. The data set is selected for the most recent years of data 2015 and up. This is two part one because too many individual datapoints cannot be plotted using folium and secondly we are making an assumption that as the years go by the most common areas where fatalities occur are improved or circumvented by the local government.

We used folium to plot the accidents and see if there is anything, we can immediately tell from the accident data. The venue data from the Foursquare API was appended to the data frame by location of each accident. The immediate area is the most important as it makes little sense in searching a mile away from the area of the accident so 100m search radius was the query sent with the API call.

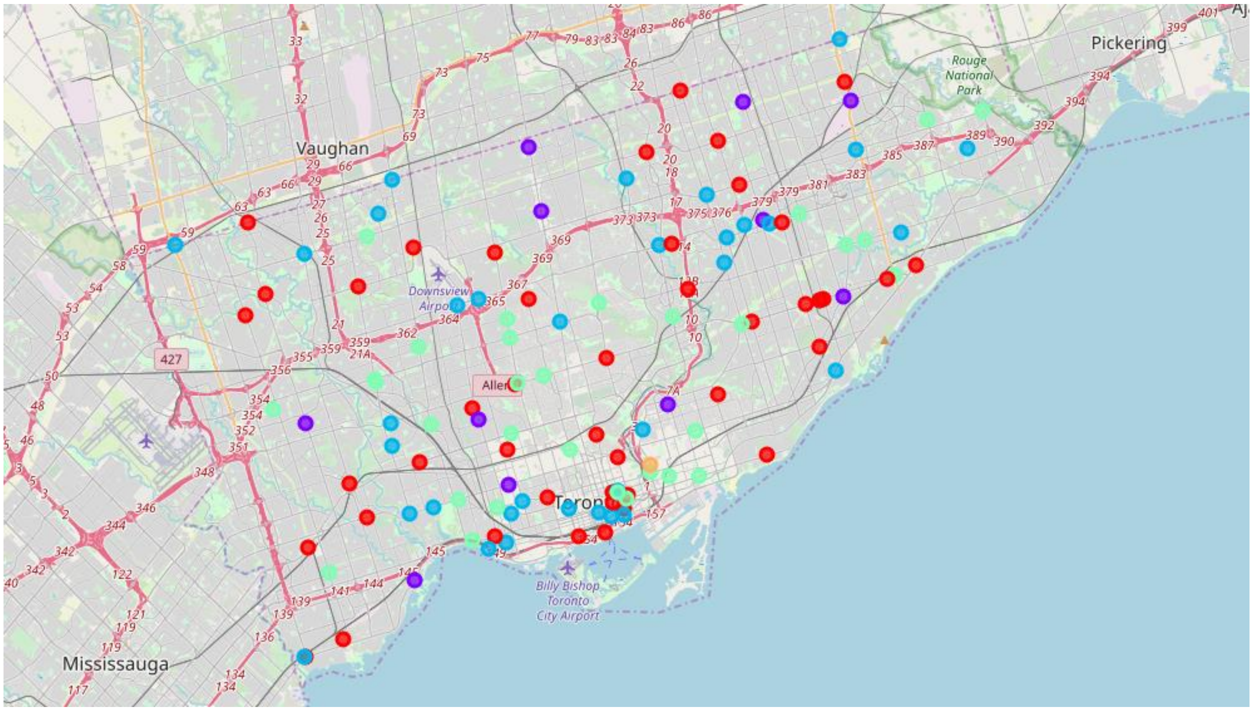


Machine learning was applied using K-means. First one-hot encoding for each of the unique venue locations. Afterwards we applied K-means using the elbow method to determine the best number of clusters to use. Afterwards we plotted the cluster data using folium and obtained data for each cluster individually for analysis.

Elbow Method:



Clustered Accidents:



4. Results

The results were inconclusive. There were a plethora of restaurants and cafes around each location but none that were common. Neither did the clusters indicate any sort of commonality in each one as venue category seemed to be equally prevalent although cluster 3 and 4 had less members in each.

5. Discussion

There were some interesting observations such as that most fatal accidents do not occur at an interstate location like one would assume. Although they are close by with the exception of a few, most fatalities occur on normal streets. More analysis could be done on the time of day or a supervised learning algorithm could be used to make a clearer analysis.

6. Conclusion

In this project we have identified the problem and utilized the Kaggle dataset as well as Foursquare API to see if we can determine if the venues nearby an accident are correlated to its likelihood of fatality. We applied K-means to discover if the clusters would be able to indicate a clear correlation of venue to fatality. However, the unsupervised machine learning model failed to determine any way humans can apply the clusters for future recommendations for government officials or individuals. As such the best recommendation we can make is to be constantly vigilant while driving as death could occur anywhere.