IBM – Coursera

Data Science Specialization

Capstone project - Final report

**Analysing and comparing student friendly neighbourhoods in USA**

Kripa Shankar Muthu Kumar– 2020

1. **Introduction:**

The students (all graduate levels) who decides to go to a university strive to know the various characteristics and benefits of the city/location to determine the ease of living.

Both international and citizens search for common parameters in the city/state:

1. Housing

2. Transport facilities (public transport), traffic incidents (crash, weather)

3. Corporate companies and tech startups (internship/full time)

4. Sports and recreation

5. Restaurants, malls, music concerts and game night pubs/bars

6. Libraries, medical centres and more

I have added traffic incidents as the students/parents might like to know the safety of driving through the neighbourhoods.

**Problem**

We must enable the user to determine the best neighbourhood cluster that satisfies most of their parameters

In this project, I try to solve this problem using foursquare location data, New York crash data and machine learning algorithms to compare and find the best places for students to find universities and settle in.

**Stakeholders**

This project will be useful to students, parents and graduate students (job seekers). This will be useful to other third parties who want to see the behaviour and psychology behind students’ requirements that must be around the universities after learning what clusters usually those people fall into.

1. **Data description:**

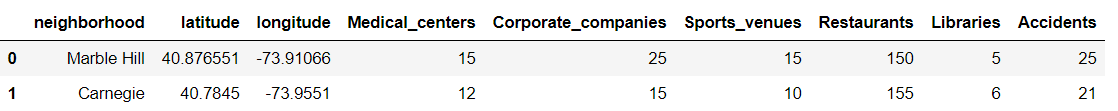
We pull location/venue data that belong to above categories from foursquare and traffic incidents data from the selected state website. In this project, we consider only one state, New York city and its neighborhoods.

I obtain Incident/crash data from [New York Motor Vehicle Collisions – Crashes dataset](https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95). The dataset contains a vast number of features like latitude, longitude, number of persons injured/killed, number of pedestrians/cyclists injured, contributing factor vehicle 1 and more. But we just count the accidents per location or borough and append it to the dataset.

Regarding the venue data, it is same as what we saw in our previous exercise (Week 3). I will extract locations like universities, tech startups/corporate companies, restaurants/pubs, medical centers, events, etc, around Queens, NYC. And, combine with the crash data.

For example, we extract venues using foursquare API: [**https://api.foursquare.com/v2/venues/explore**](https://api.foursquare.com/v2/venues/explore)**.**  We get the data, normalize, find the frequency and append to the dataset.

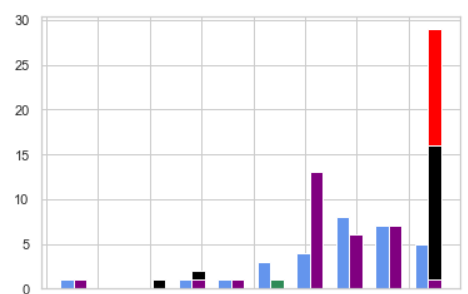
Below is the example of the resultant data.



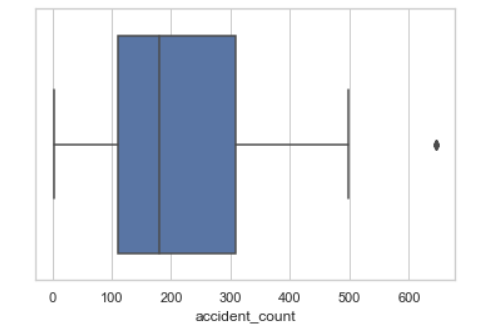
Also, this is a sample dataset for now. Dataset features can get modified as the project gets progressed.

1. **Methodology:**

After integrating the data (including accidents), that is data preparation, we must understand the data.

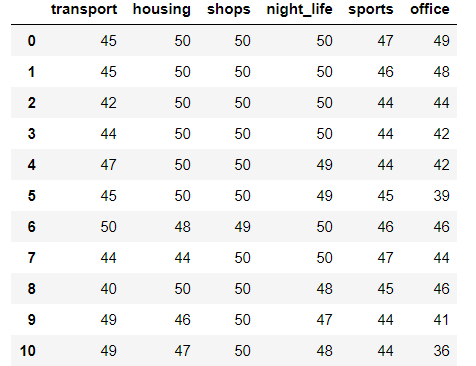


Venue distribution

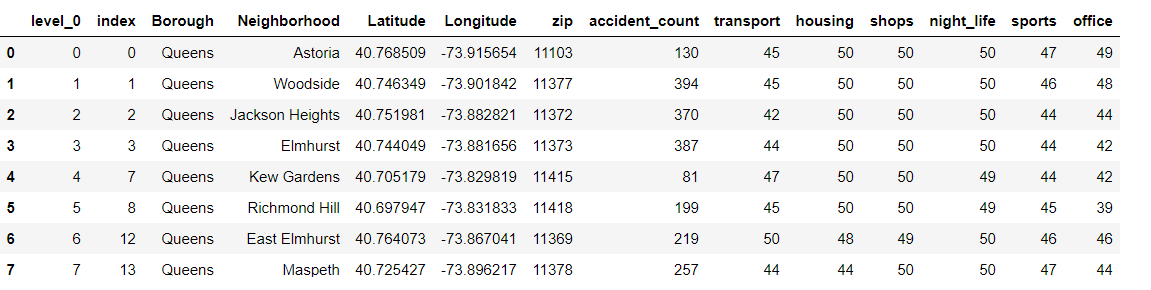


Accident count

Venue distribution plots the following venue facilities data



The final data are formed by combining accident count, neighbourhood data and the facilities data. Zip is obtained using google maps



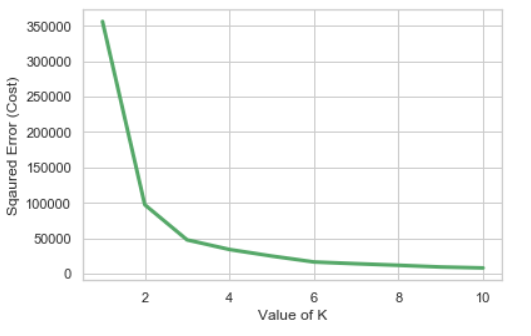
**K means clustering**

Using K means clustering, we cluster the following queen’s data to find areas with both facilities and accident-less prone areas.

We perform clustering because we don’t have labels for the process that we do. And, this is a segmentation procedure. K means calculates all the possibilities and positions the centroids accordingly.

For clustering, we index data frame from accident\_count to office column and perform.

To find an optimal value of k, we use elbow method

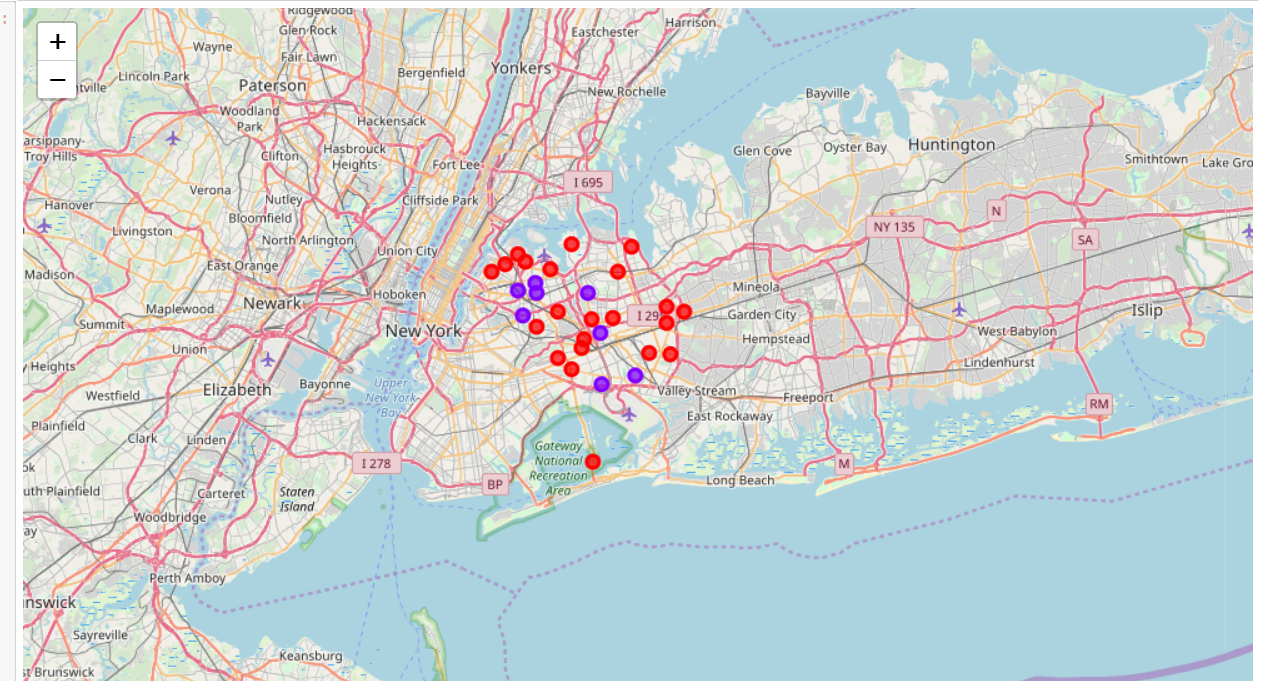


From this graph, we see that optimal number of clusters could be 2.

1. **Results:**

K means assigns 0 cluster label to less accident-prone areas and 1 to more accident-prone areas.

Here is the accident map plotted for queens, New York city, New York.

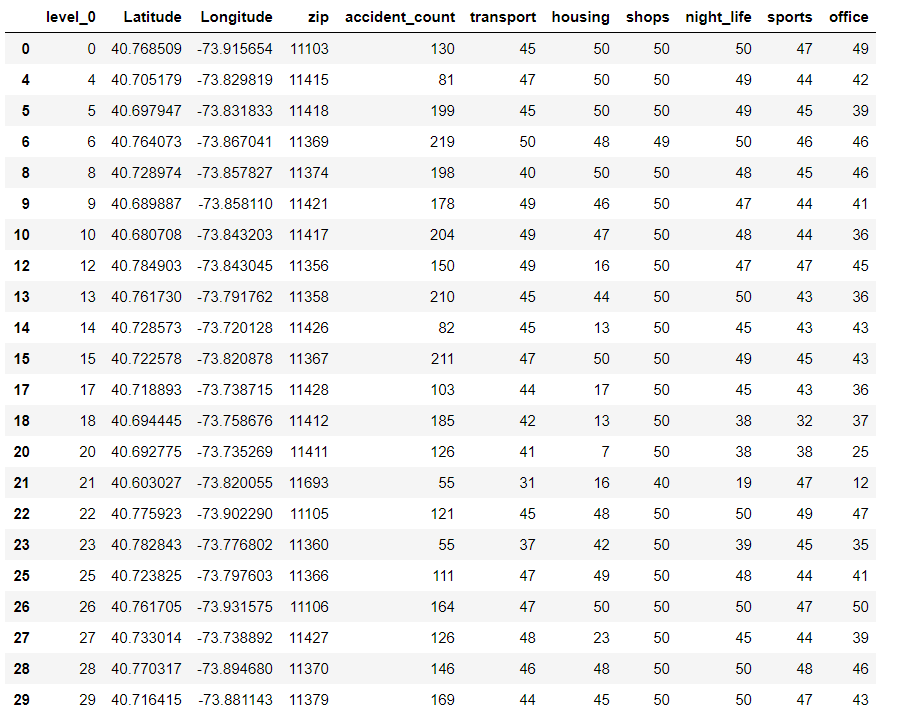


Purple belongs to cluster label 1. Read belongs to cluster label 0.

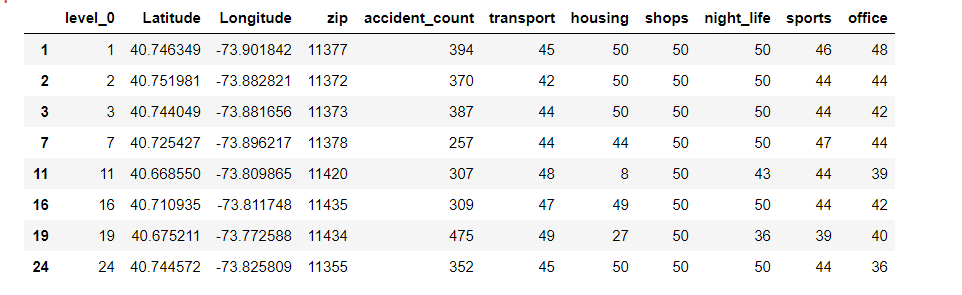
1. **Discussion**

From examining clusters, we obtain the following data frame

This belongs to cluster 0 – less accident-prone areas



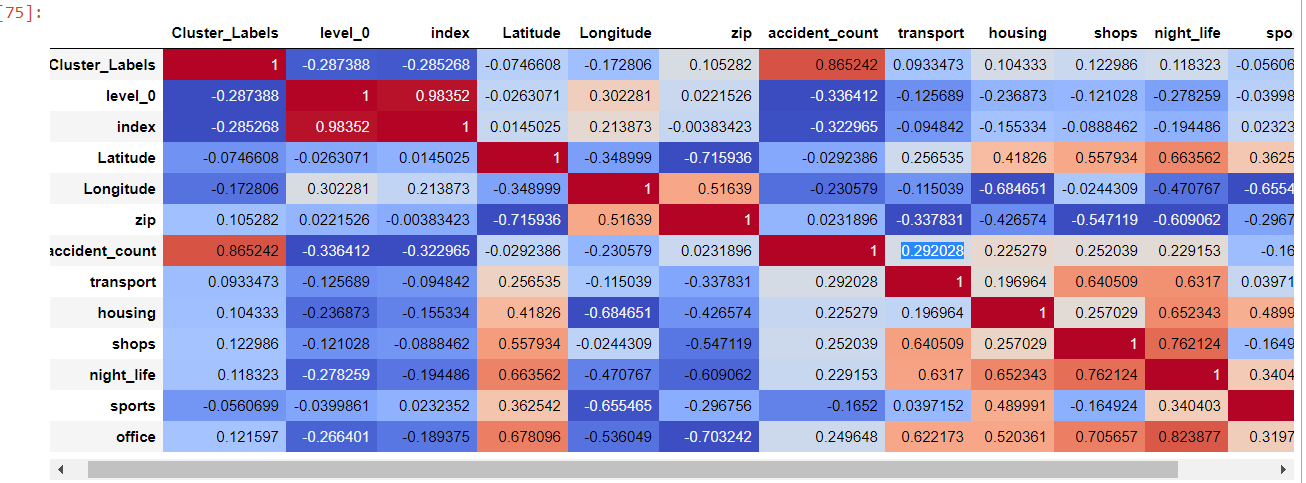
This belongs to cluster 1 – more accident-prone areas

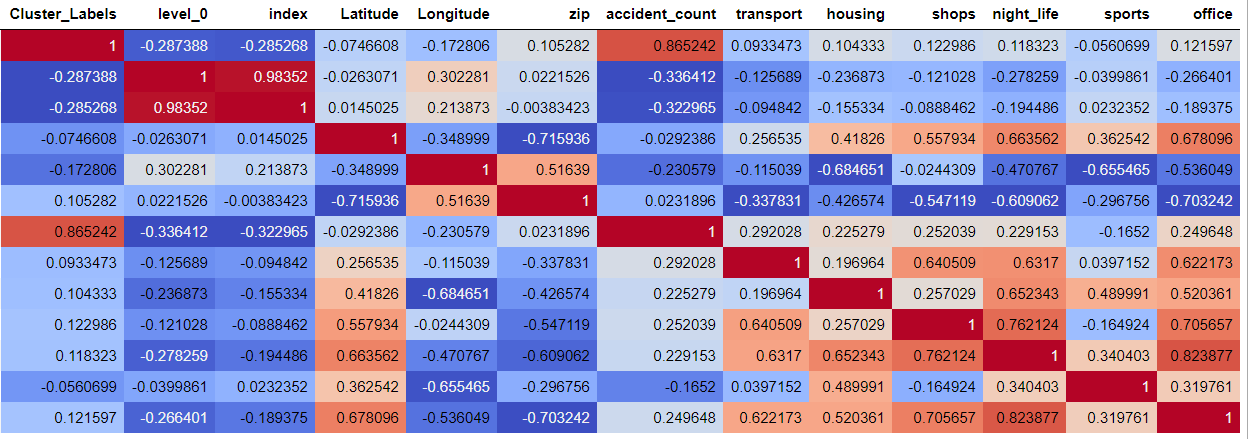


The K means clustering correctly partitioned the high volume of accidents area to low volume of accidents area. Firstly, I thought of finding the correlations between the venues/facilities to the accident\_prone areas. However, just by a glance, it looks like there is not a significant impact of venues, but, still, notably, in hight volume of accidents area, there are a lot of transport options and office areas. On the map, we could see that there are a lot of accidents around Airport area.

The lack of significant impacts of venues could be because of my data collection. Even though, I tried to get the accident count for zip code, I couldn’t get the venue spots zip code wise. It is due to some API and data model issues. However, I will try to diagnose this issue more and try to fix it.

Correlations matrix





**VI. Conclusion**

Thus, this project will be useful for students to select the areas that they would like to get housing, jobs or go to places to enjoy. Accidents clustering also helps students and parents discern places that they have to be cautious about. Whole data collection and preparation were done using foursquare API and Google Maps (reverse geo coding)