Final Report

Background and Research information:

We have the data of NYPD Motor Vehicle Collisions. Then we are going to use this data to analysis the accidents and traffic happened in the past and find out the most place where and most frequency period time when the accidents and traffic occur so that to predict the future place and time the accidents make be occur to avoid them.

The data we will have is the source from: https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95 where is the NYC open data. The data is collected because of the law, the data will be updated monthly and it can be used for seeing how dangerous/safe intersections are in NYC. We can download it in pdf or excel in order to analysis them for a more depth level.

i. NYPD MVC done this before.

ii. It is important because we can learn from data that when and where is the most accidents happened, and we can send more police officers to the place at that time to control the traffic which can do not waste the police officers time and reduce accidents happening.

iii. Challenges: 1. how to reduce the attributes and dimensional of data.

2. how to deal with miss data.

3. large data set

iv. No ethical considerations.

v. we are helping NYPD to predict the accidents happen and make them using less police officers to reduce more number of accidents.

vi. Too much attributes, and some data are missing, what attributes did not show on websites but we need to consider like weather, how to deal with special case(outlier).

vii. From website data.cityofnewyork.us which the calculation by government.

viii. data is mess and need to be cleaned so that can be used to predict.

ix. I plan to use cross-correlation, PDM and agglomerative clustering to reduce attributes. Then using K-NN deal with missing data or just delete it.

x. When data were cleaned, using K-mean to find the number of clustering by time and place. Therefore, I will use the algorithms of K-means.

xi. How many data or which year of data we using for analysis.

xii. For location, we use L1 norm distance metric to compare the latitude and longitude. For time, we use hour in a day, day in a week, and month in a year to see when is the most number of accidents.

Plot out graphs (at least this graphs should be plot out):

1. Number of accidents vs time. (daily, monthly, yearly)
2. Number of accidents vs place.
3. Number of accidents vs types of accidents

Experiment:

1. I used python matplotlib package to plot different graphs that I analysis.
2. I plot several graphs:

   Accidents frequency against hours in a day

   Accidents frequency against days in a month

   Accidents frequency against months in a year

   Accidents frequency against 2012 to 2018

   SSE against number of K values

   Accidents location with clustering in NYC

3. I pull out the attributes that DATE, TIME, LATITUDE, and LONGITUDE, where DATE and TIME are used to analysis accidents frequency against time period, LATITUDE and LONGITUDE are used to analysis accidents frequency against location.
4. I did not merge others attributes

Discussions

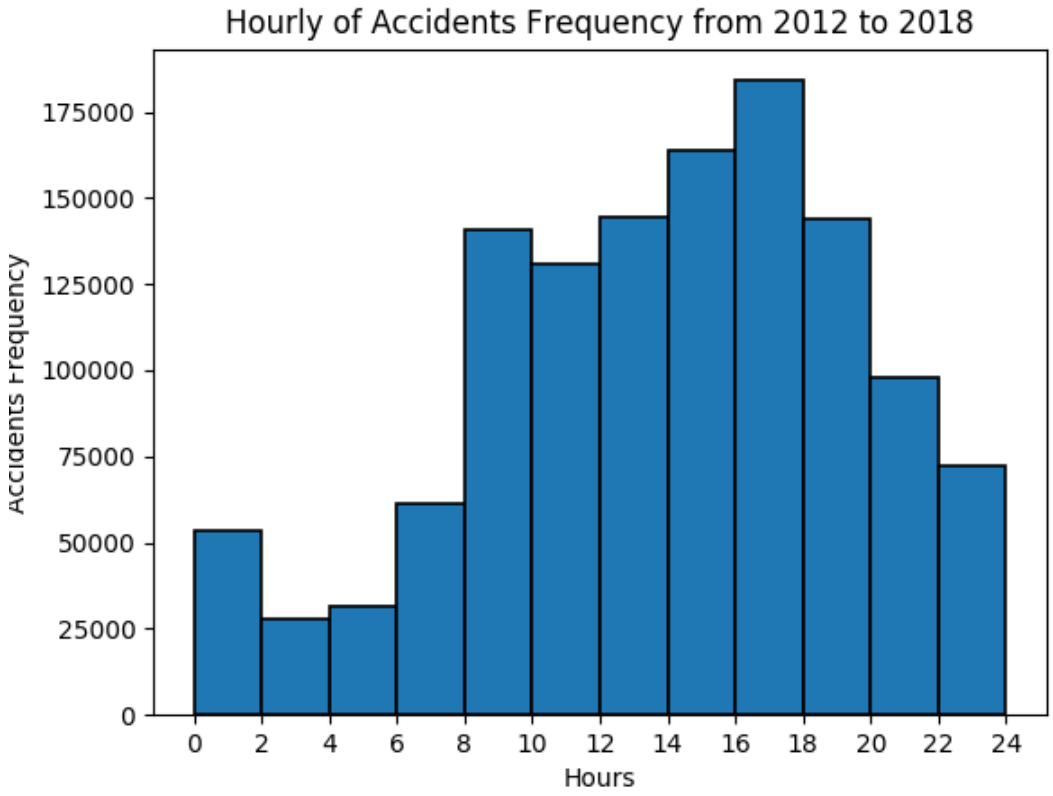
1. I finally used k-means to analysis data.
2. A) First challenge is that there are many missing data in LATITUDE and LONGITUDE. I choice to delete the missing data because using K-NN is not making sense to guess those missing data. The data that I downloaded are sorted by time, the neighbors around missing have not relate to each other so K-NN is not working for them.

   B) Second challenge is the outliers. The first time that I run through k-means and plot the data out that I find out my graph is wired and not make sense. Therefore, I checked that there are 3 points which are far away to each other, so that I knew

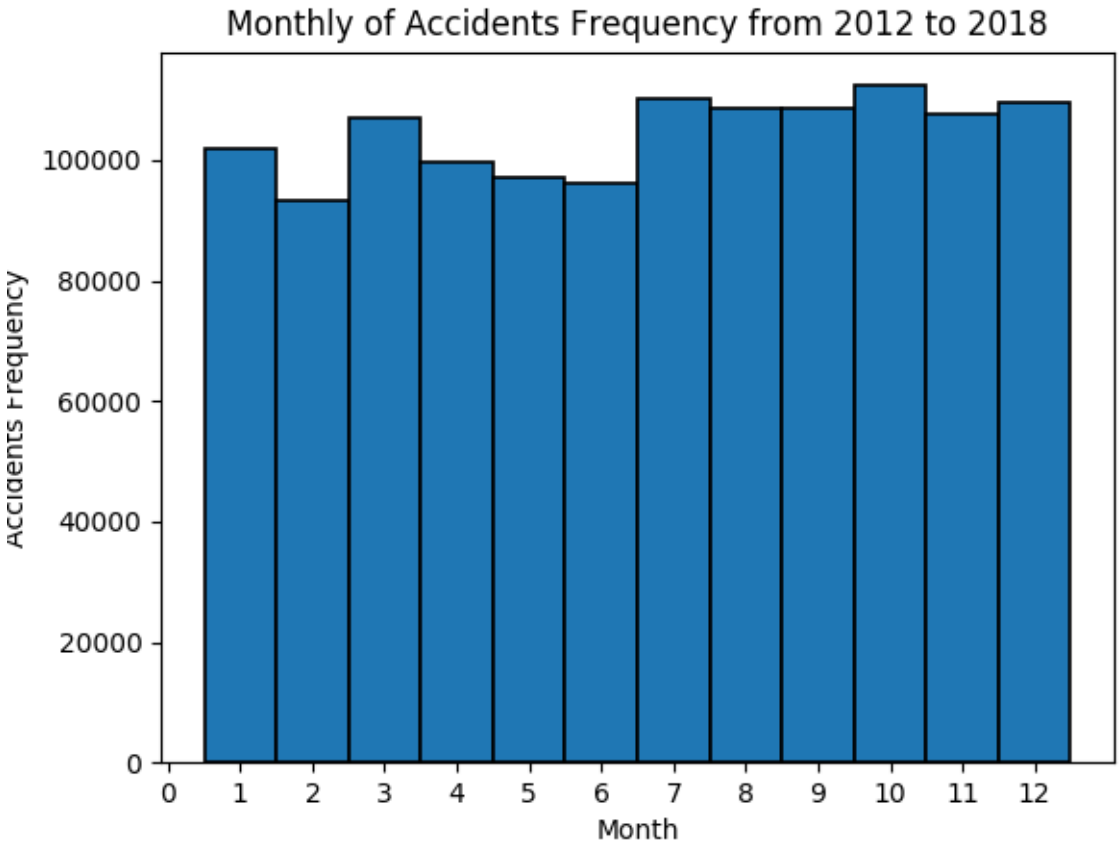
there are outlier in this data and need to be cleaned. Hence, I used mean and standard deviation to find outliers and remove them. Those points that are out of range (mean +- std.dev) seem to be a outliers.

C)The third challenge is the data is much big. There are total 1,034,254 many data in this csv file. Therefore, put all data and run K-means all at once will be very slow and may be forever. Then, I found out that the most cost time is the process to find best K values of k-means, which plot out those many data is fast, so for such huge data set I chose that allowing users to choose from what year to what year they like to analysis. If the data set is smaller that 100,000 then we just go through k-means all at once. If data set is larger than 100,000, we will random choose 100,000 samples from data set and using it to calculate the best K-values for length(data set)/100,000 times and choose the most common one to be the best K-value. If the data set larger than 500,000, it just run it for 5 time at most to find the best K-values.

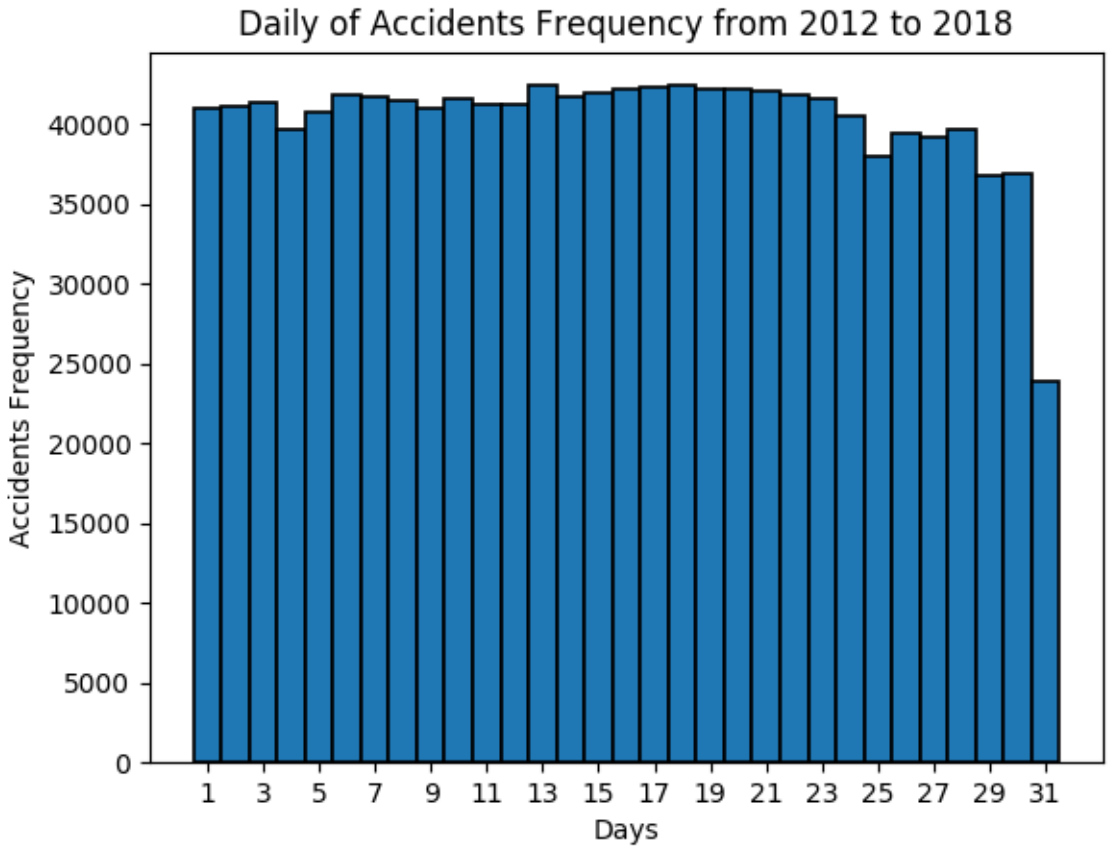
3. The interesting things are that during doing this project, I will find out what algorithm is fit in the data. It just flush in my mind that I know for here I should use what to deal with.
4. The K-means is working the best to clustering the locations against accidents frequency so that I can see where is the place accidents would like to occur.
5. By looking at the result graphs.



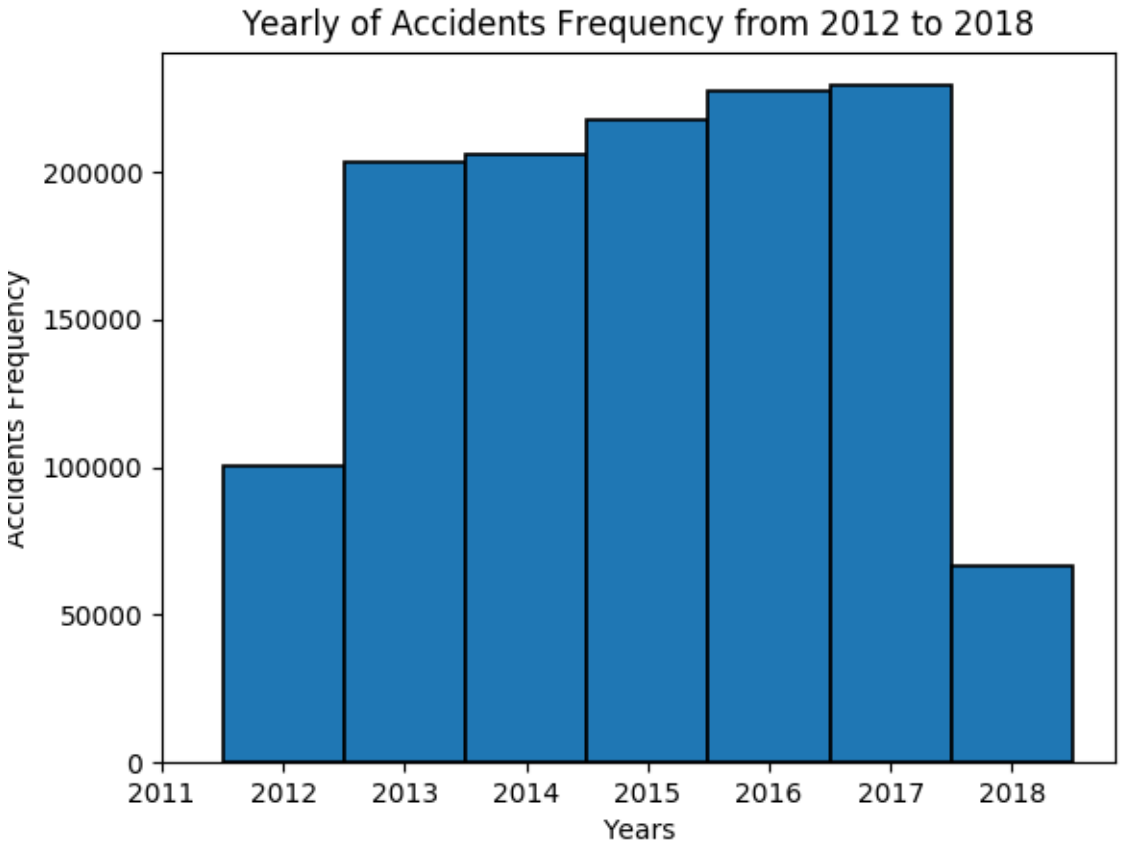
Here is the Hourly graph. From the graph we can know that the frequency accidents happened at the least between the time 0:00 to 8:00 which make sense which most of people were sleeping at this period of time. Then, we saw that there are two peaks where 8:00 to 10:00 and 16:00 to 18:00, which those two periods of time people were going to work, so there were many cars driving on the road the chance of accidents occurs increase.



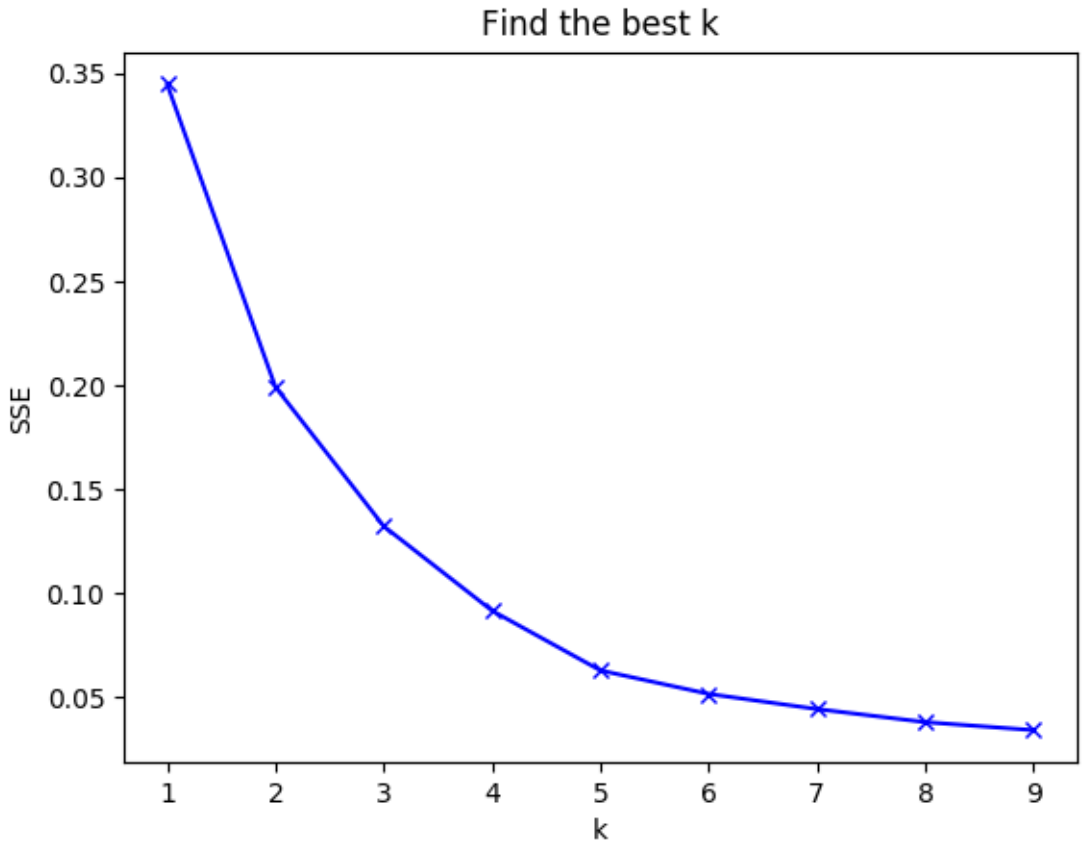
This is the monthly graph. We can see that the right half months where from July to December have more accidents that the left half months where from January to June. The February is a outlier because it has the least day for all the months. Therefore, without looking at February, the most safety month is June, and the most dangerous month is October.



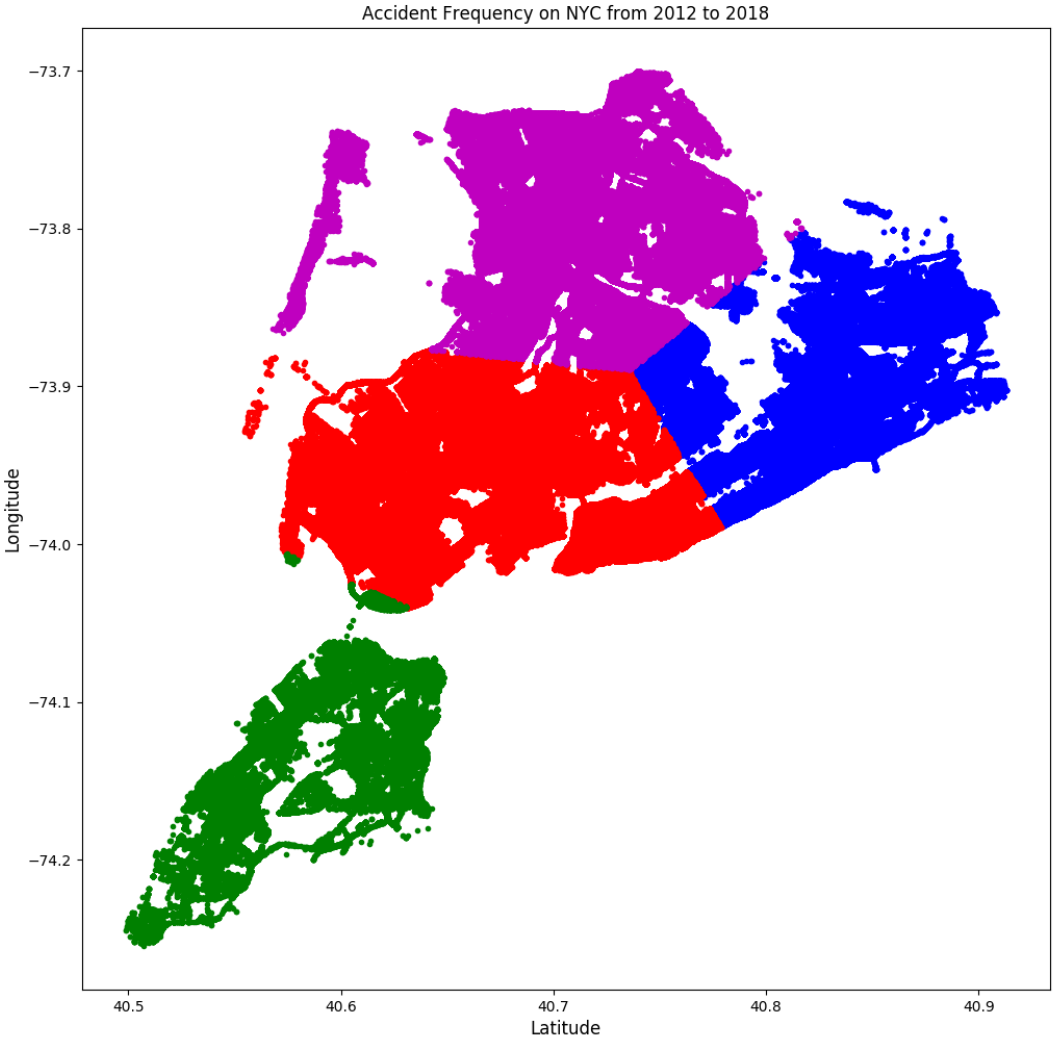
Here is the daily graph. We can see that the couple days at the end of month have less accidents, the reason is that some months do not have 31st and Feb do have 30st and 31st and sometime no 29th .



Here is the Yearly graph. The graph shows that the accidents happened is the least at 2018 because we are current at 2018 and it is not finish yet. Other than that, 2012 is the year that less accidents that all others. Then, we can see that the frequency of accidents is increasing year by year, so the next frequency of accidents will be also increased.



Here is the SSE against to the value of k. It is easily to see that the SSE decreasing by increasing the value of k. Then, by calculating distances of each point to the origin, I found out the most closet point is when k=4. Therefore, the k=4 is the best k value for this data set to implement k-means.



Here is the location of NYC after k-means clustering. We can see that there are four main clusters, and those four areas (Manhattan, Queens, Brooklyn, Staten Island), are the highly frequency of accidents locations in NYC. Compare four areas, the Staten Island is the most accident happened location where **number of accidents/areas** the density of accidents.

6. Using of K-means:

   First, read csv file to get data.

   Plot out the histogram graph where frequency against period of time.

   Data about time do not need to clean where there is no missing data.

   Second, selecting subset of data by period of time. Cleaning missing data. Cleaning outlier data.

   For 1 to 10, try all value of k and find out the cost of them. If the data set is too large, separate to smaller data set and use sample to calculate k values. Using the value where the point is the closet to the origin. Run K-means, then plot out the clusters.

Conclusion:

After this semester, I did learn many methods to deal with data. The purpose we clean and analysis data is to predict what will happen in the future. Then, not only for this project, there are homework which I can truly using what I learn to analysis data, and I did get the information that the data tells us. I am so glad that I choose this class.