Restaurant Segmentation By Crime

Capstone project
By Deepti Chopra

Introduction and Background

- In the Capstone Project, my idea is to show that when driven by venue and location data from FourSquare API, backed up with open source crime data, it is possible to present a new/first-time/solo traveller with a list of attractions to visit supplementd with a graphics showing the crime occurance in the region of the venue.

- A high level approach is as follows:
  1. The travellers decides on a city location
  2. Fetch the top venues from FourSquare
  3. Augment the list of top venues with additional geographical data
  4. Using this additional geographical data the top nearby restaurents are selected
  5. The historical crime within a predetermined distance of all venues are obtained
  6. A map is presented to the to the traveller showing the selected venues and crime statistics of the area.

Who are the end users?

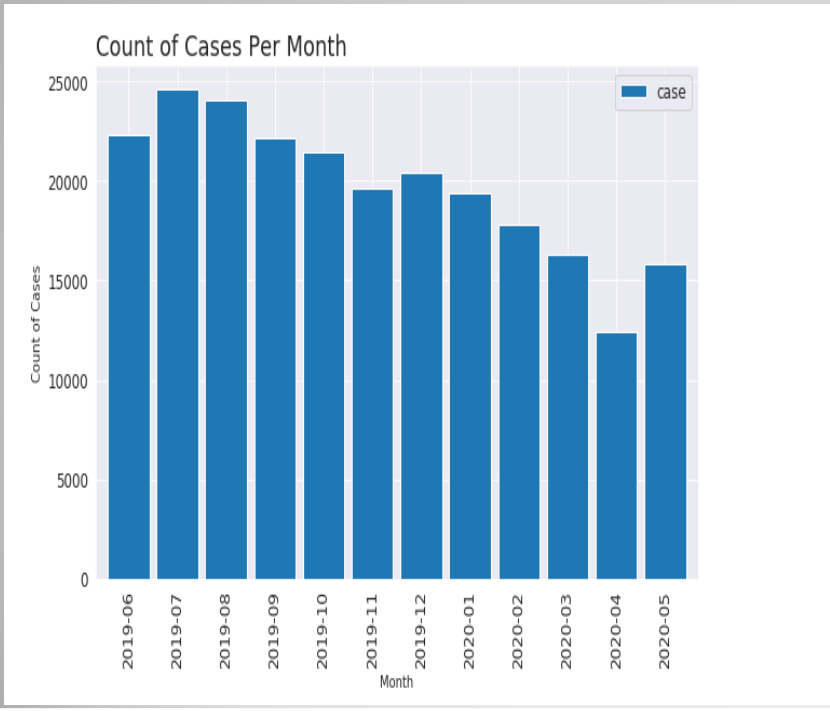
- This solution is targeted at new/first-time/cautious traveller. They want to see all the main sites of a city that they have never visited before but at the same time, for whatever reasons unknown, they want to be able to do all that they can to make sure that they stay clear of trouble i.e. is it safe to visit this venue and this restaurant at 4:00 pm in the afternoon.

Data

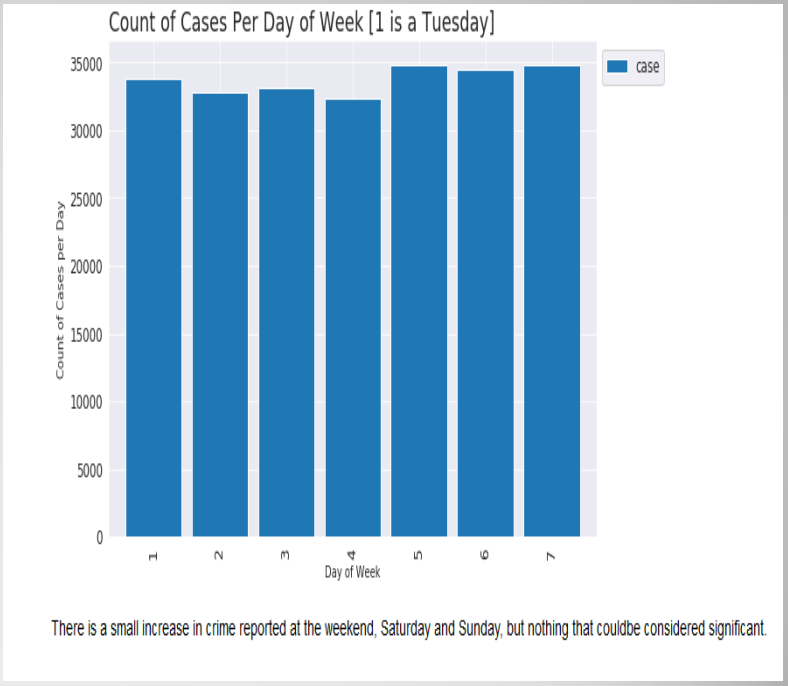
- In this section, I will describe the data used to solve the problem. We discuss the following 4 scenarios:
  1. Query the FourSqaure website for the top sites in Chicago
  2. Use the FourSquare API to get supplemental geographical data about the top sites
  3. Use the FourSquare API to get top restaurent recommendations closest to each of the top site
  4. Use open source Chicago Crime data to provide the user with additional crime data

Exploratory Data Analysis and Statistical Testing

- **Number of Crimes per month**

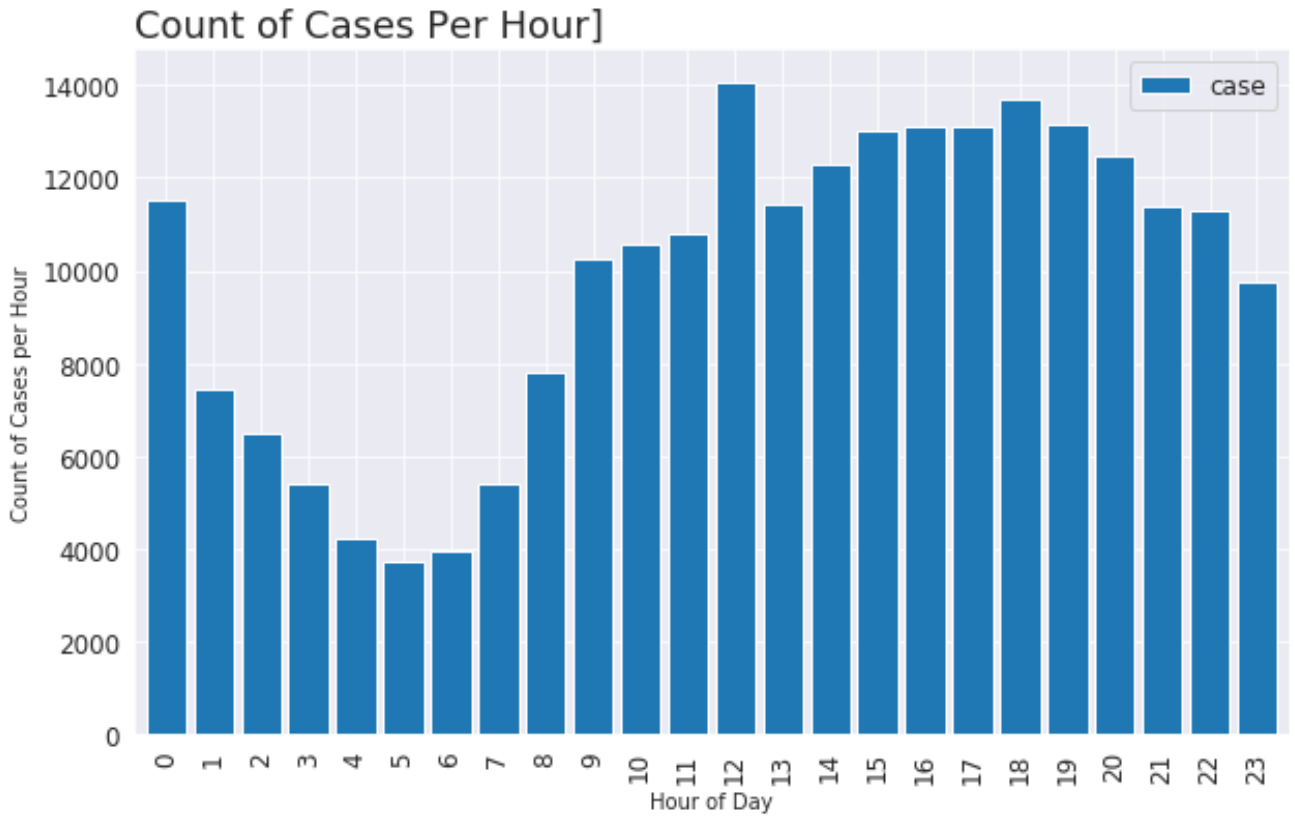


- **Number of crimes occuring on each day**



There is a small increase in crime reported at the weekend, Saturday and Sunday, but nothing that couldbe considered significant.

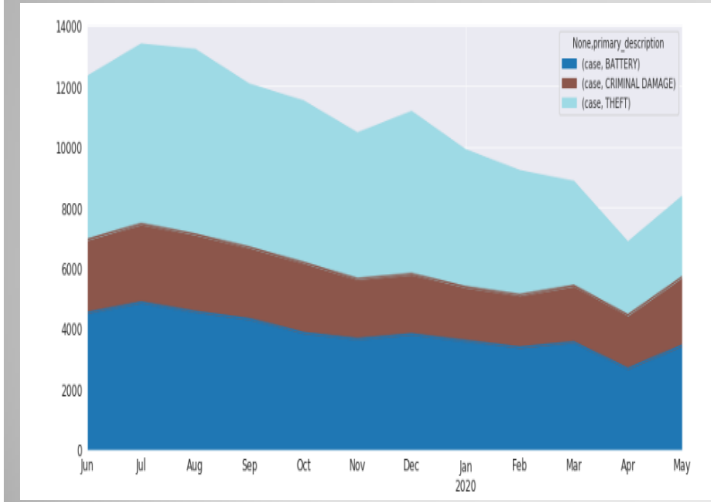
Number of crimes occuring in each hour

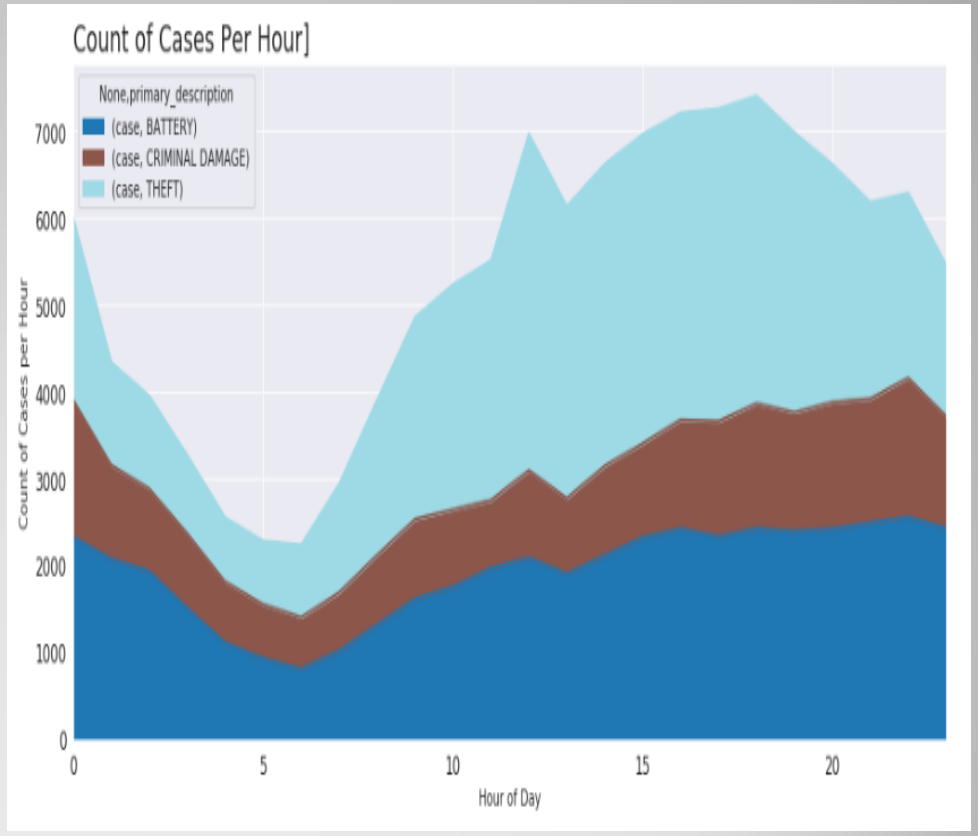


There is an expected fall-off in reported crime rates after midnight before elevating again after eight in the morning.

3 most commonly occurring crimes:

	primary_description	case
30	THEFT	55293
2	BATTERY	46223
6	CRIMINAL DAMAGE	25611





Crime Categories

Modelling and Machine Learning

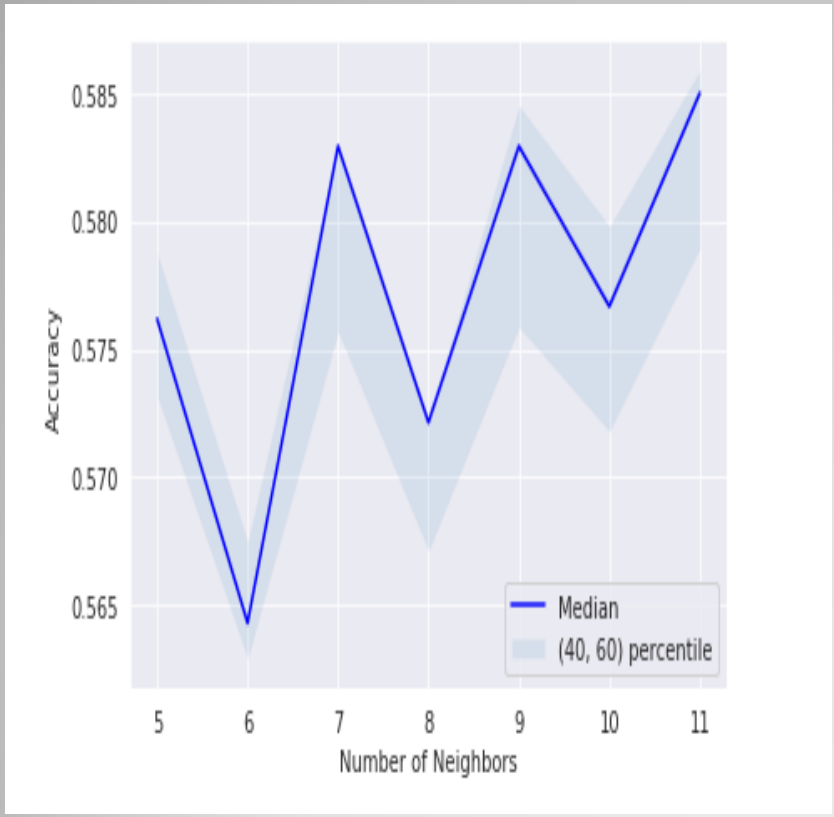
- Before we start modelling we need to prepare the data frame to include only numerical data and by removing unneeded columns.
- Rather than removing colums from df_crimes a new df_features DataFrame was created with just the required columns. This df_features DataFrame was then processed to remove Categorical Data Types and replace them with One Hot encoding. Finally the Dependant Variables were Normalised.
- The Features DataFrame looked like this:

	latitude	longitude	hour_0	hour_1	hour_2	hour_3	hour_4	hour_5	hour_6	hour_7	hour_8	hour_9	hour_10	h
1	41.841609	-87.658034	0	0	0	0	0	0	0	0	0	0	0	0
8	41.969365	-87.728061	0	0	0	0	0	0	0	0	0	0	0	0
9	41.803121	-87.609460	0	0	0	0	0	0	0	0	0	0	0	0
10	41.733015	-87.552709	0	0	0	0	0	0	0	0	0	0	0	0
11	41.708243	-87.639226	0	0	0	0	0	0	0	0	0	0	0	0

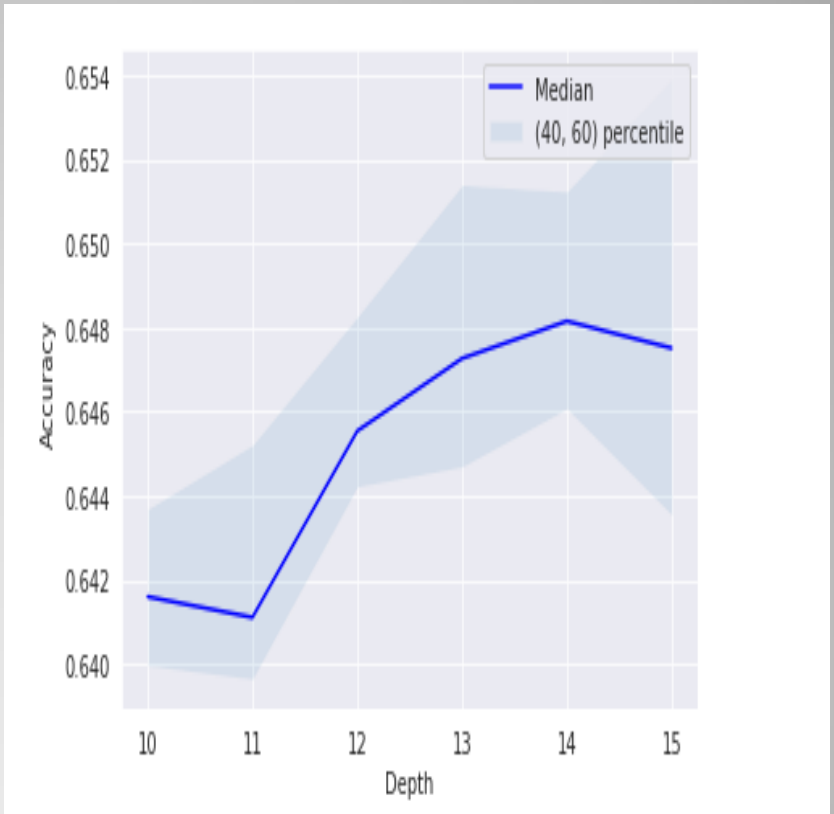
Machine Learning Models

- Five model type were then chosen to be evaluated:
  1. K Nearest Neighbours
  2. Decision Trees
  3. Logistic Regression
  4. Naive Bayes
  5. Decision Forest using a Random Forest

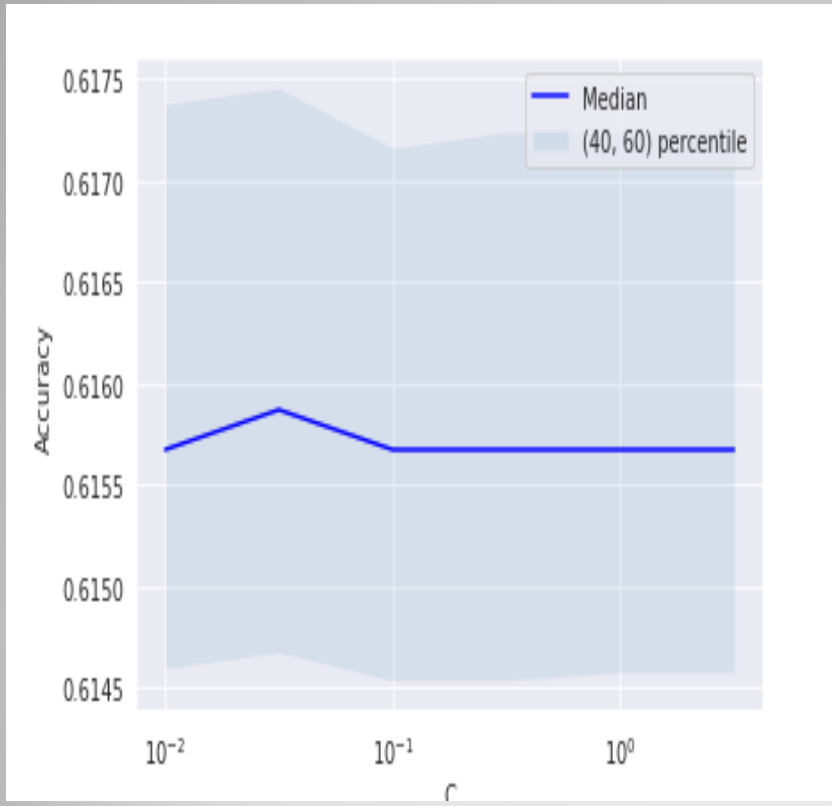
K Nearest Neighbor(KNN)



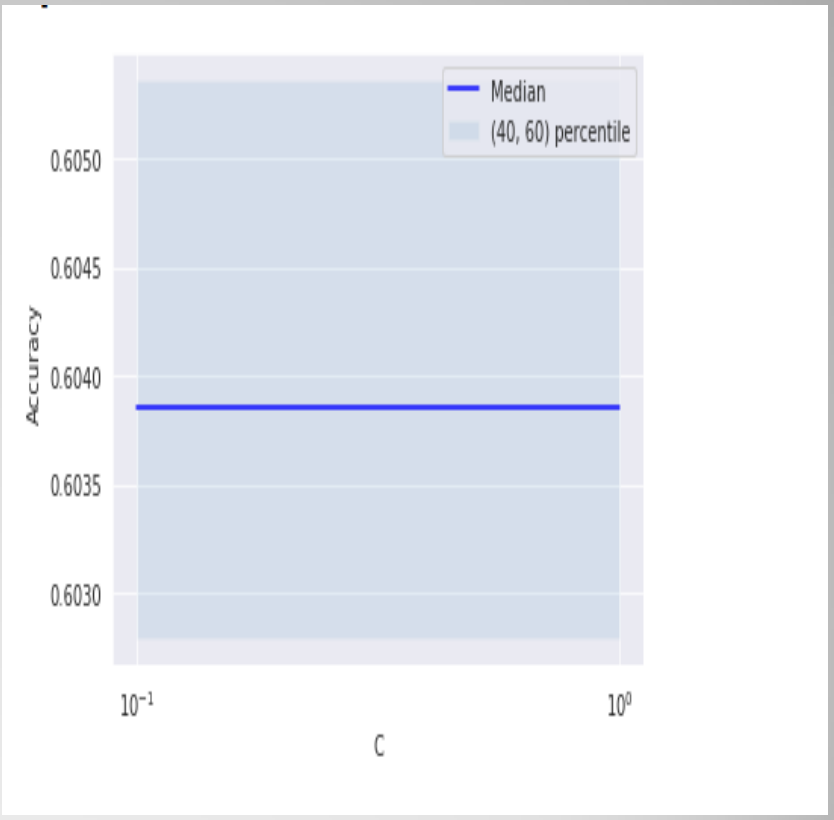
Decision Tree



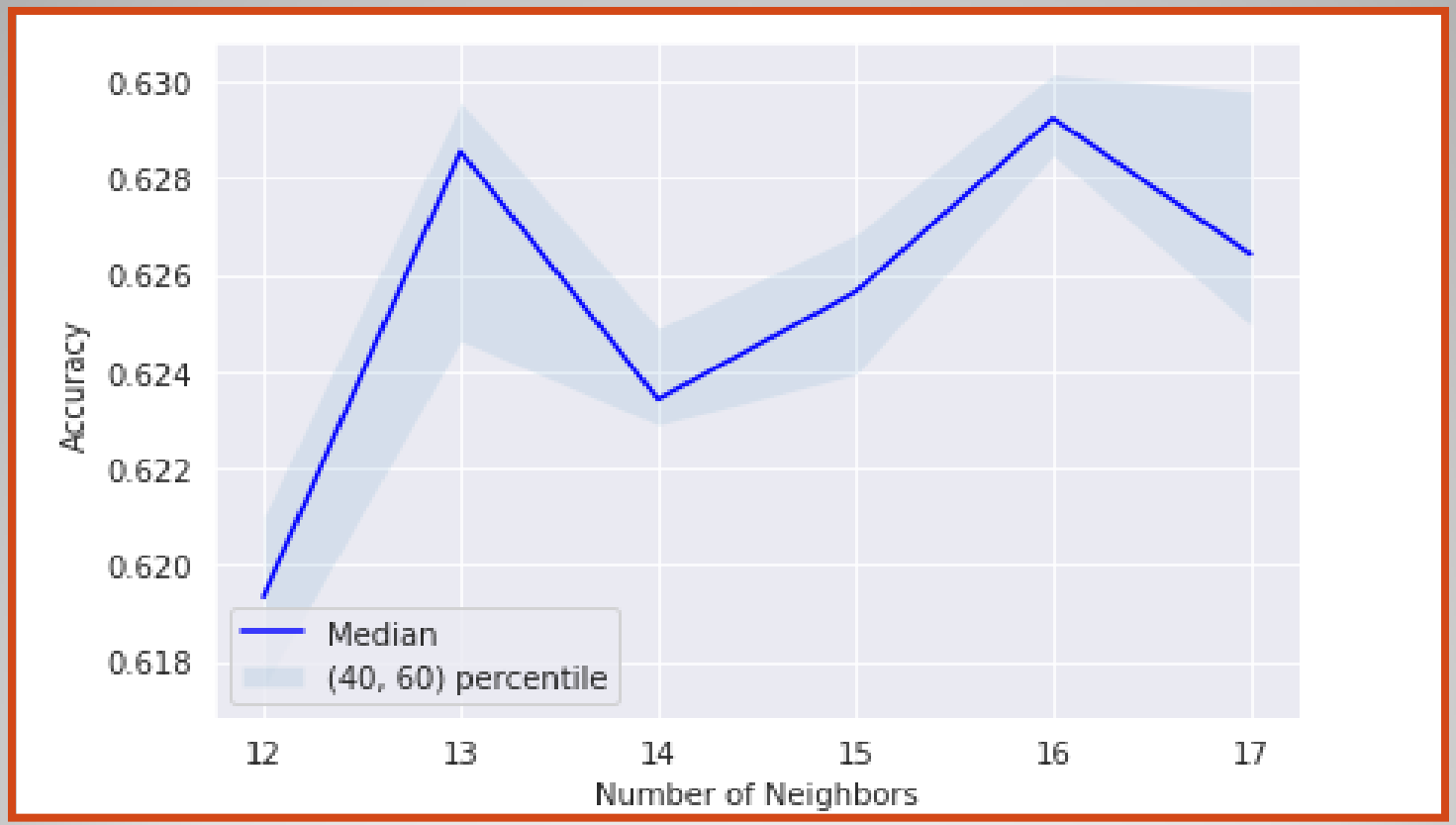
Logistic Regression



Naive Bayes

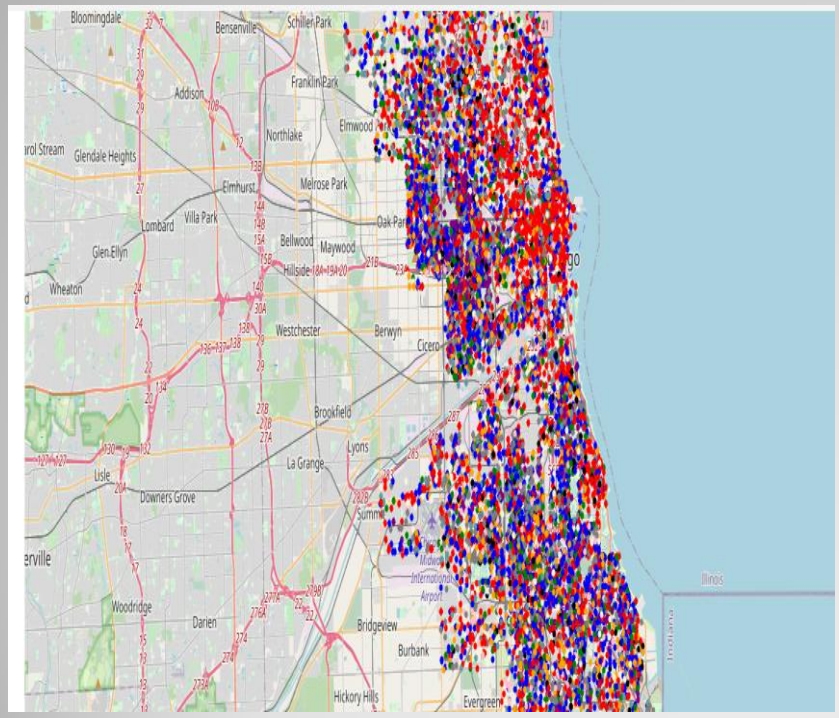


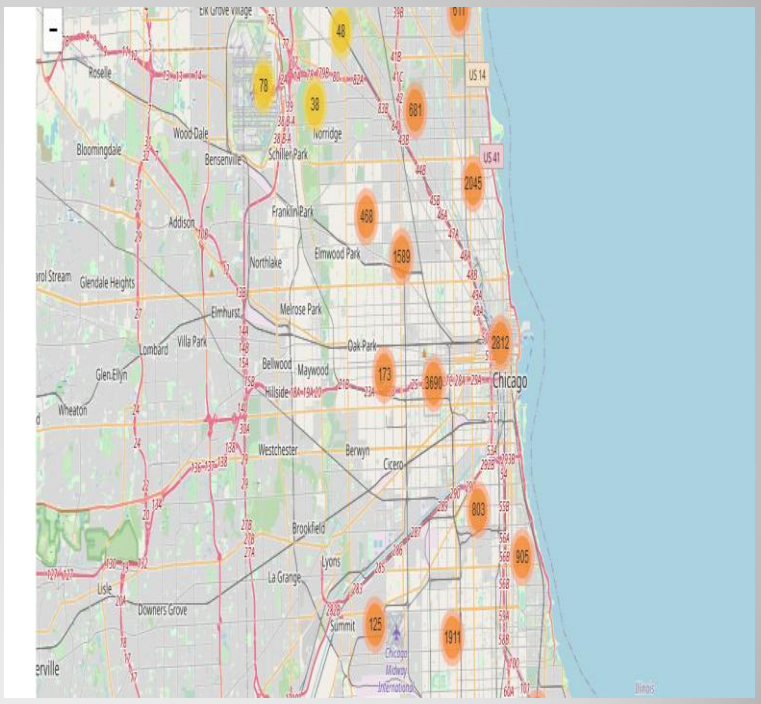
Random Forest

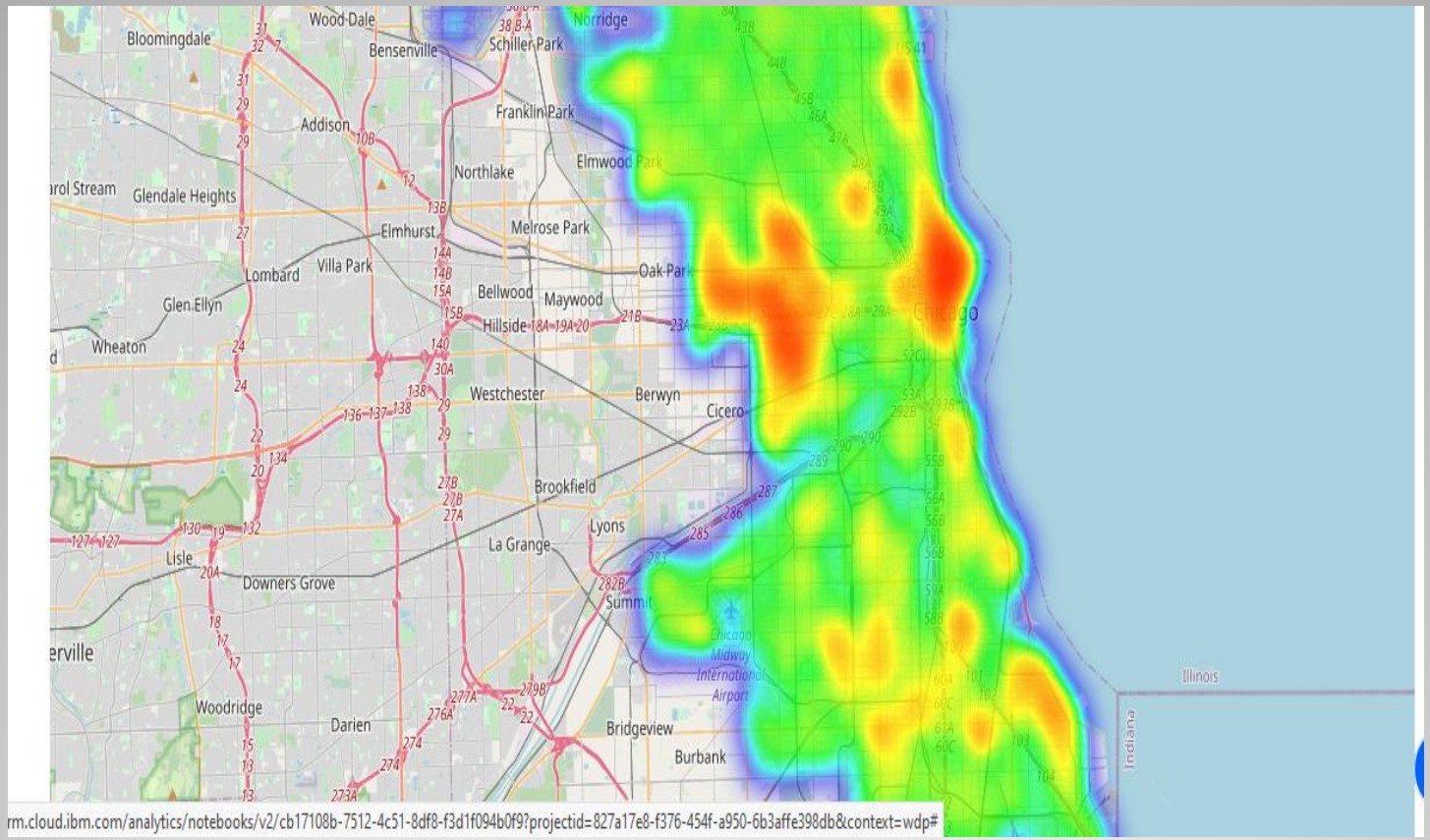


RESULTS

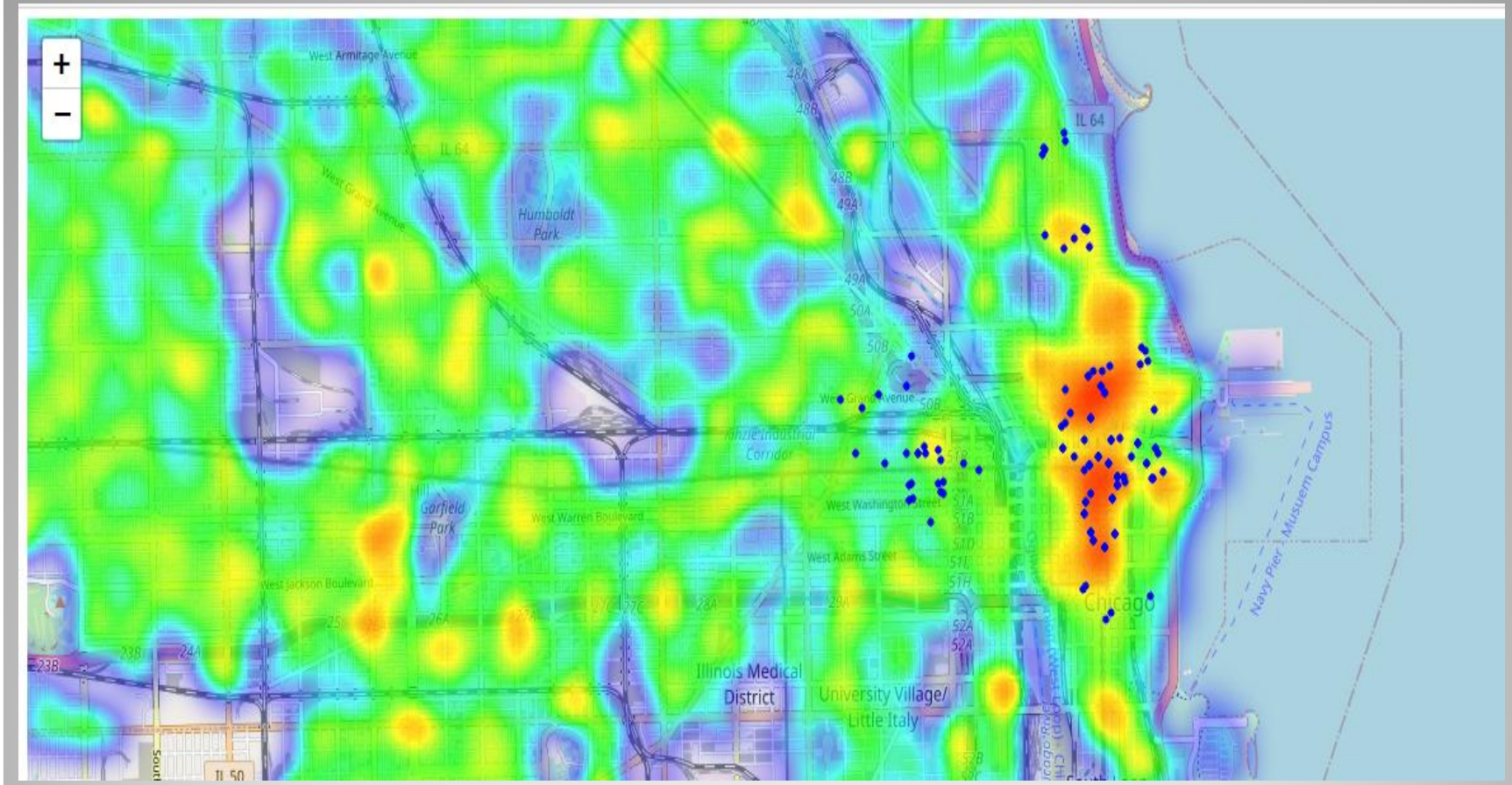
- **CRIME MAP-**







Crime heatmap



Popular restaurants visualized on heatmap

DISCUSSION

- In all we were able to build visualizations that would help the tourists decide on locations to visit.
- The decision is going to be of the traveller. Whether he/she wants to totally avoid an area that has high crime rate or want to visit a popular attraction despite that.
- We can clearly make more such maps for other categories of attractions as we have done for restaurants.

Conclusion

- Although all of the goals of this project were met there is definitely room for further improvement and development as noted below. However, the goals of the project were met and, with some more work, could easily be devleoped into a fully phledged application that could support the cautious traveller in an unknown location.
- Of the contributing data the Chicago Crime data is the one where more data would be good to have. Also not every city in the world makes this data freely available so that is a drawback.
- FourSquare proved to be a good source of data but frustrating at times. Despite having a Developer account I regularly exceeded my hourly limit locking me out for the day. This is why Pickle was used to store the captured data.


THANK
YOU