Analyzing crime dataset of Detroit city

Anshul Mishra

June 2019

Table of Contents

Table of figures

Introduction

As of 2016, Detroit has the fourth highest murder rate among major cities in the United States after St. Louis, Baltimore and New Orleans and the 42nd highest murder rate in the world. The crime rate has decreased over the years, but the city is overridden with economic downturns and high unemployment. This analysis will help local government agencies as well as tourists to identify geographical areas of interests. Government agencies will be able to make informed and focused decisions to reap out desired outcomes more efficiently. The tourists will be aware of the areas that should be avoided for safe travel and stay. Entrepreneurs can also understand demographics of various areas around city to make better investments for growth and profits.

There are 100's, maybe even 1000's, of travel sites on the Internet, including FourSquare, that will tell you all about places to go, things to see, restaurants to eat at, bars to drink in, nightclubs to part the night away in and then where to go in the morning to get breakfast and a strong coffee. The problems with these sites is that they are one dimensional. If you want to find out all this information about a city you plan to visit next month, you have to do the hard work. Also, just because a venue is the hottest place to go for a night out does not always mean that the unwitting tourist should just ramble in unprepared. The areas surrounding this new venue might be riddled with crime including muggings, car theft and assault, for example. Approach the venue from any direction other than from the north and you could be putting your life in danger. This is when my idea comes in.

Project Idea

My idea for the Capstone Project is to show that when driven by venue and location data from FourSquare, backed up with open source crime data that it is possible to present the cautious and nervous traveler with a list of attractions to visit supplemented with a graphics showing the occurrence of crime in the region of the venue.

A high level approach is as follows:

- The travelers decides on a city location [in this case Detroit]
- The ForeSquare website is scrapped for the venues in the city
- From this list of venues, the list is augmented with additional geographical data
- A map is presented to the traveler showing the selected venues and crime statistics of the area.

Beneficiaries

This solution is targeted for informed decision making. The 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.

Some examples of envisioned users include:

1. Government Agencies

2. Traveler

3. Entrepreneur

Data science aspect of this project

- Data Acquisition
- Data Cleansing
- Data Analysis
- Machine Learning

Data

In this section, I will describe the data used to solve the problem as described previously.It is possible to attempt quite complex and sophisticated scenarios when approaching this problem. However, given the size of the project and for simplicity only the following scenario will be addressed:

1. Query the FourSqaure website for the top sites in Detroit
2. Use the FourSquare API to get supplemental geographical data about the sites
3. Use open source Detroit Crime data to provide the user with additional crime data

Data Acquisition

The first phase of the project is to acquire all of the data that is needed for this project. The initial data required can be broken down into two separate data sets:

1. The FourSquare Venues to Visit in Detroit
2. The Detroit Police Department Crime Data from 1920 to June,2019 (https://data.detroitmi.gov/api/views/6gdg-y3kf/rows.csv?accessType=DOWNLOAD)

Data Cleaning and Analysis

Features to keep from crime database

- Crime ID
- Incident Date & Time
- Offense Category
- Neighborhood
- Latitude
- Longitude

	Crime ID	Offense Category	Incident Date & Time	Neighborhood	Longitude	Latitude
0	3372082	LARCENY	06/17/2019 07:00:00 AM	Outer Drive-Hayes	-82.962068	42.417754
1	3372058	DAMAGE TO PROPERTY	06/17/2019 05:51:00 AM	Midwest	-83.113531	42.355217
2	3372054	LARCENY	06/17/2019 05:34:00 AM	Seven Mile-Rouge	-83.272280	42.429002
3	3372059	LARCENY	06/17/2019 05:02:00 AM	Warrendale	-83.216817	42.343297
4	3372059	AGGRAVATED ASSAULT	06/17/2019 05:02:00 AM	Warrendale	-83.216817	42.343297

Figure 1 Filtered crime dataset

Data Processing

1. Clean up the column names:
    - Strip leading & trailing whitespace
    - Replace multiple spaces with a single space
    - Remove # characters
    - Replace spaces with _
    - Convert to lowercase
    - Change the date of occurance field to a date / time object
2. Add new columns for:
    - Hour
    - Day
    - Month
    - Year
3. Split Block into zip_code and street
4. Verify that all rows have valid data

	index	crime_id	offense_category	incident_date_&_time	neighborhood	longitude	latitude	hour	day	month	year	year_month
0	0	3372082	LARCENY	2019-06-17 07:00:00	Outer Drive-Hayes	-82.962068	42.417754	7	1	6	2019	2019-06
1	1	3372058	DAMAGE TO PROPERTY	2019-06-17 05:51:00	Midwest	-83.113531	42.355217	5	1	6	2019	2019-06
2	2	3372054	LARCENY	2019-06-17 05:34:00	Seven Mile-Rouge	-83.272280	42.429002	5	1	6	2019	2019-06
3	3	3372059	LARCENY	2019-06-17 05:02:00	Warrendale	-83.216817	42.343297	5	1	6	2019	2019-06
4	4	3372059	AGGRAVATED ASSAULT	2019-06-17 05:02:00	Warrendale	-83.216817	42.343297	5	1	6	2019	2019-06

Figure 2 Cleaned crime dataset

Predictive modelling

5. Prepare data to include only numerical data and by removing unneeded columns.
    a. Rather than removing columns from df_top10_crimes_2018_june, creating a new df_features DataFrame with just the required columns.
    b. DataFrame df_features will then be processed to remove Categorical Data Types and replace them with One Hot encoding.
    c. Finally the Dependent Variables will be normalized.
    d. There are a couple of further small changes to be made:
    e. Create the X, dependent variables, DataFrame by dropping the Crimes column
    f. Create the y, independent variable
    g. Normalize the X Data

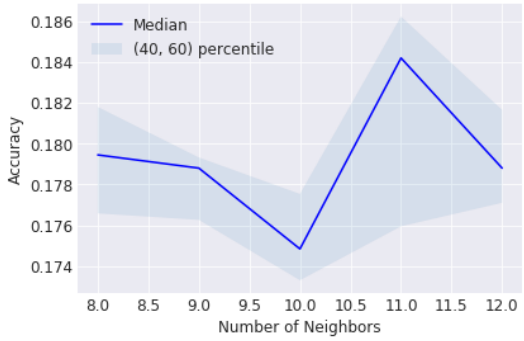
	latitude	longitude	hour_0	hour_1	hour_2	hour_3	hour_4	hour_5	hour_6	hour_7	hour_8	hour_9	hour_10	hour_11	hour_12	hour_13	hour_14	hour_15	hour_16
78698	42.386436	-83.145336	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
78699	42.441512	-83.172150	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
78700	42.348444	-83.041754	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
78701	42.416838	-83.164166	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
78702	42.344425	-83.230505	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3 Dataset for predictive modelling

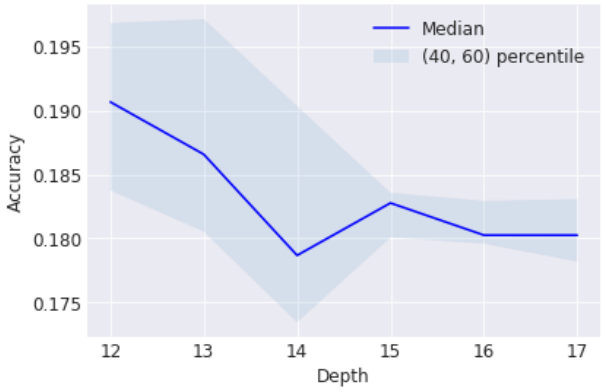
Predictive Models

- K Nearest Neighbor(KNN) –

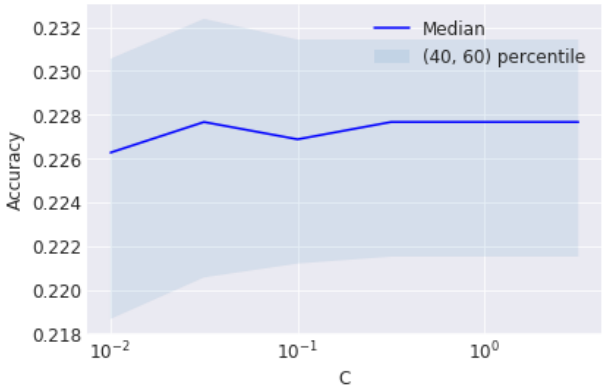
  K = 11



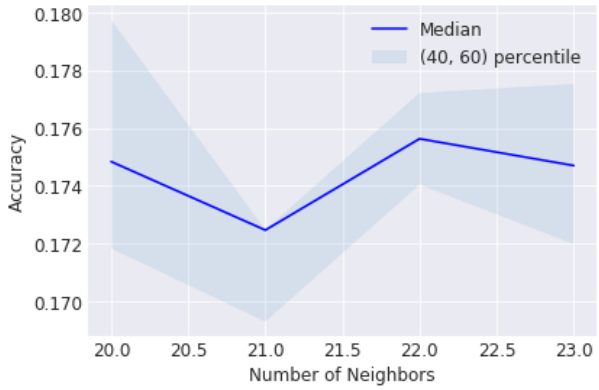
- Decision Tree
  Depth = 12



- Logistic Regression
  C = 0.1



- Decision Forest using a Random Forest
  N_estimator = 21



- **Random Forest** is the best model scoring highest in all measurements, F1-Score, Jaccard and Log Loss. Let's now create a new model. The june,2018(hour = 6) crime data will become the unseen test data for the final model.

Results and Discussion

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.

The following goals were met in this project:

- Identified crime prone areas in Detroit city
- Mapped these crimes onto geographical map of Detroit
- Clustered markers on map for easy identification of crime with an ability to dwell down on more information
- Produced heat map to show crime density in various localities in Detroit
- Understood which month and day of week are more prone to crime based on inferential statistics on historical data
- Did data cleaning and prepared top venues data from four square for mapping crime to venues
- Developed a predictive model to predict which crime category will happen

Results

1. Crime reporting and logging has significantly increased.
2. There is a sharp increase in year 2016 and 2017, Please note - For 2019, it is mid year report.

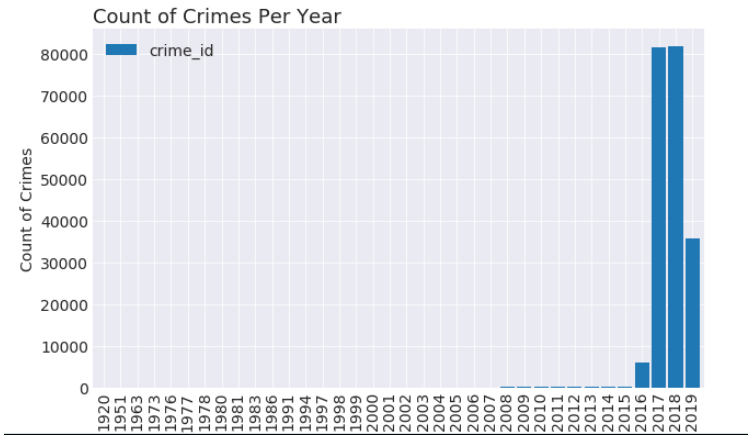


Figure 4 Bar chart showing count of crime per year

3. There is small increase in crimes on Thursday, which is surprising as crimes rate is expected to be higher on weekends. Though the difference is too small to be significant.

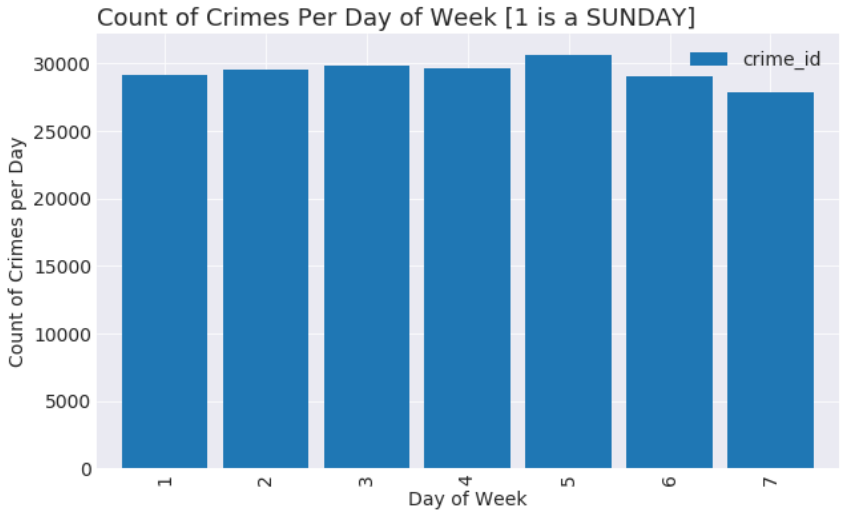


Figure 5 Bar chart showing count of crime per day

4. There is an expected fall-off in reported crime rates after midnight before elevating again after six in the morning. There appears to be a spike around midday.

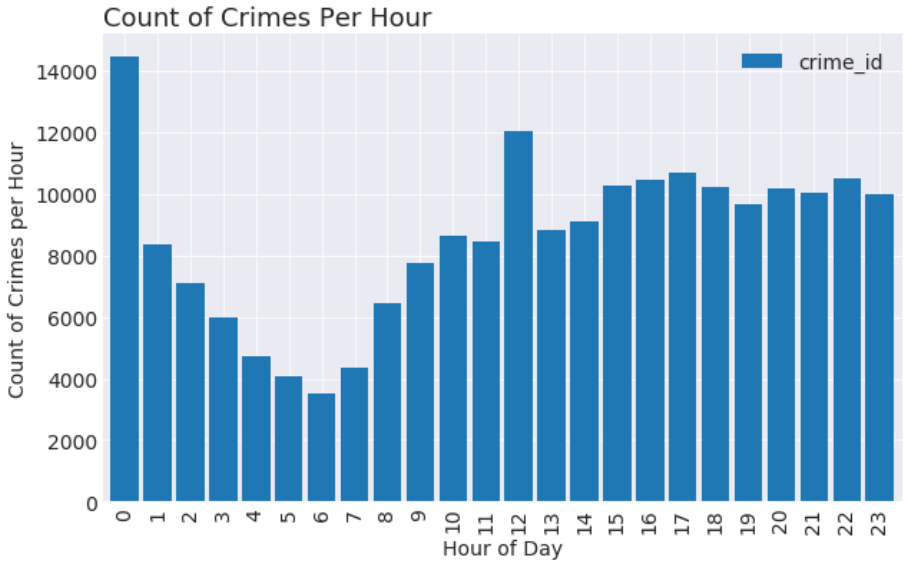


Figure 6 Bar chart showing count of crime per hour

5. Top 3 crimes
    - Assault
    - Damage to Property
    - Larceny

6. For top 3 crimes
    - Crime rates peak in May and June

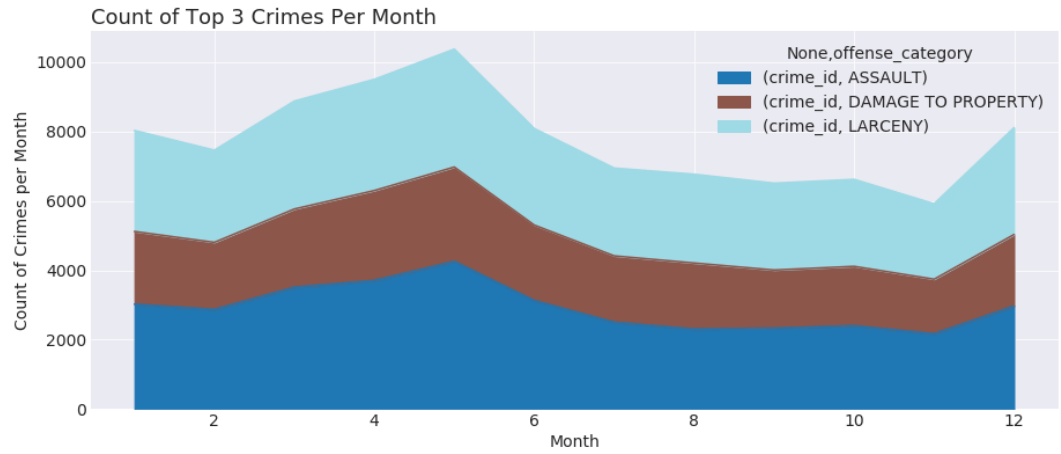


Figure 7 Area plot showing crime count per month

- Second half of weak have higher crime rates than first half

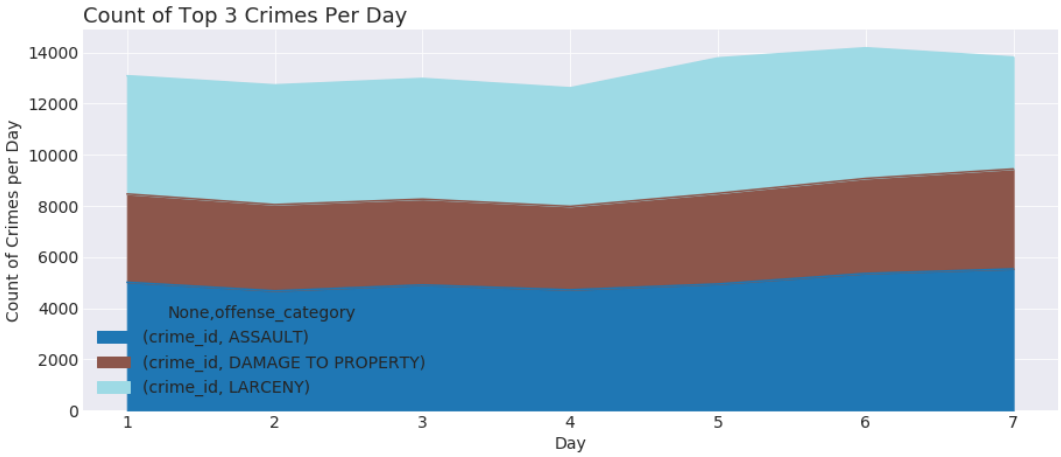


Figure 8 Area plot showing crime count per day

- Morning 5am is the safest time in Detroit, midnight and 12 noon being the most unsafe time

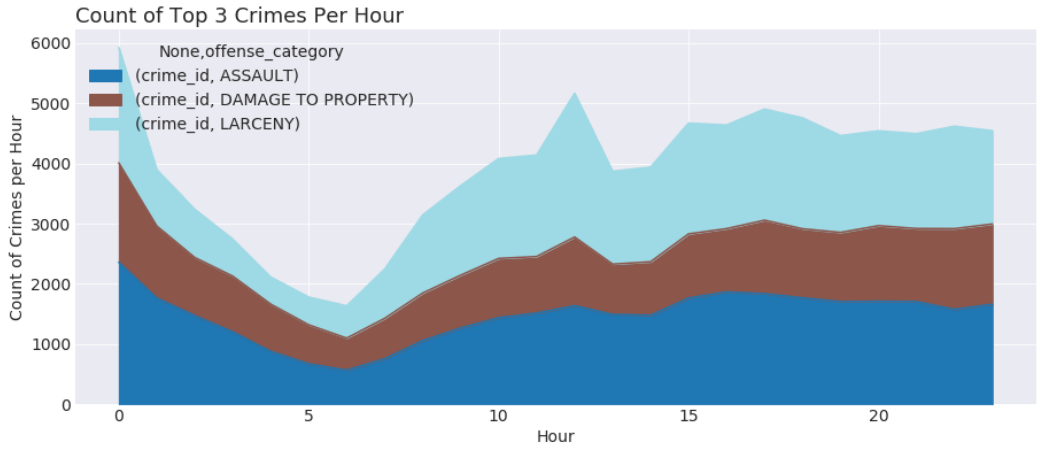


Figure 9 Areaplot showing crime count per hour

7. Due to computational limitation focusses analysis on crimes that happened at noon in June of 2018 only

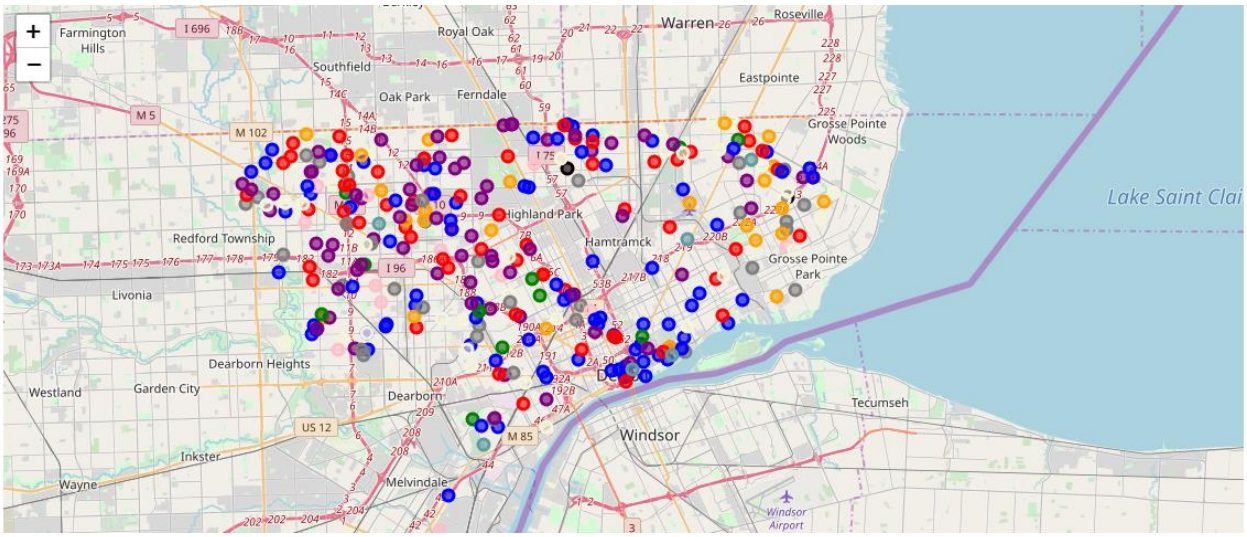


Figure 10 Marker map of Detroit showing crimes that happened on noon in June,2018

- Clusters of crime locations were visible, particularly around the periphery of Detroit.

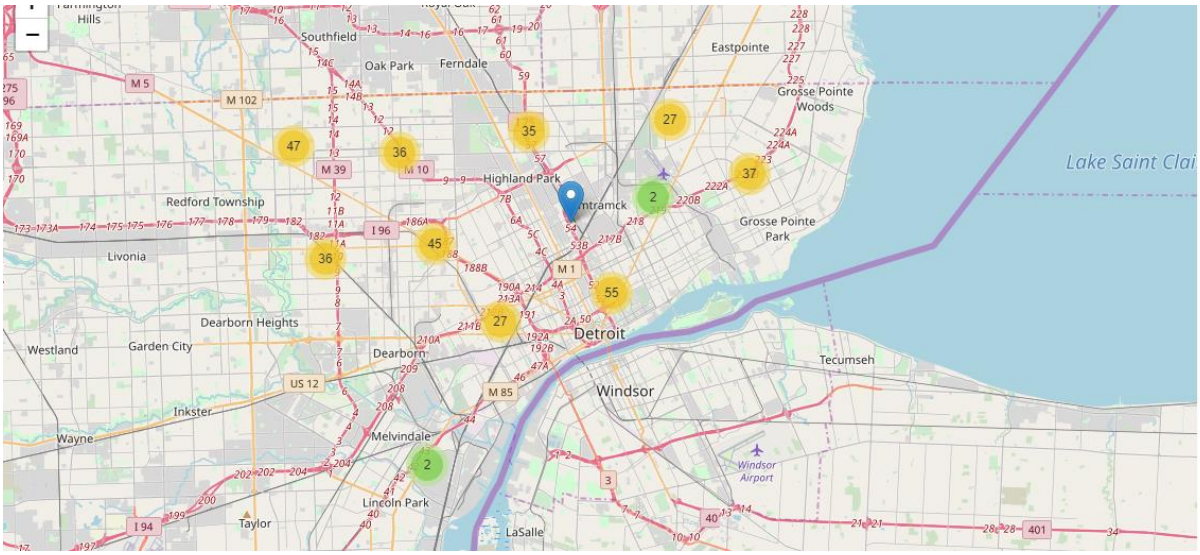


Figure 11 Marker map of Detroit showing cluster of crime of noon, June,2018

- Heat map shows **Greektown and Macomb street have a high crime rate occurrence**

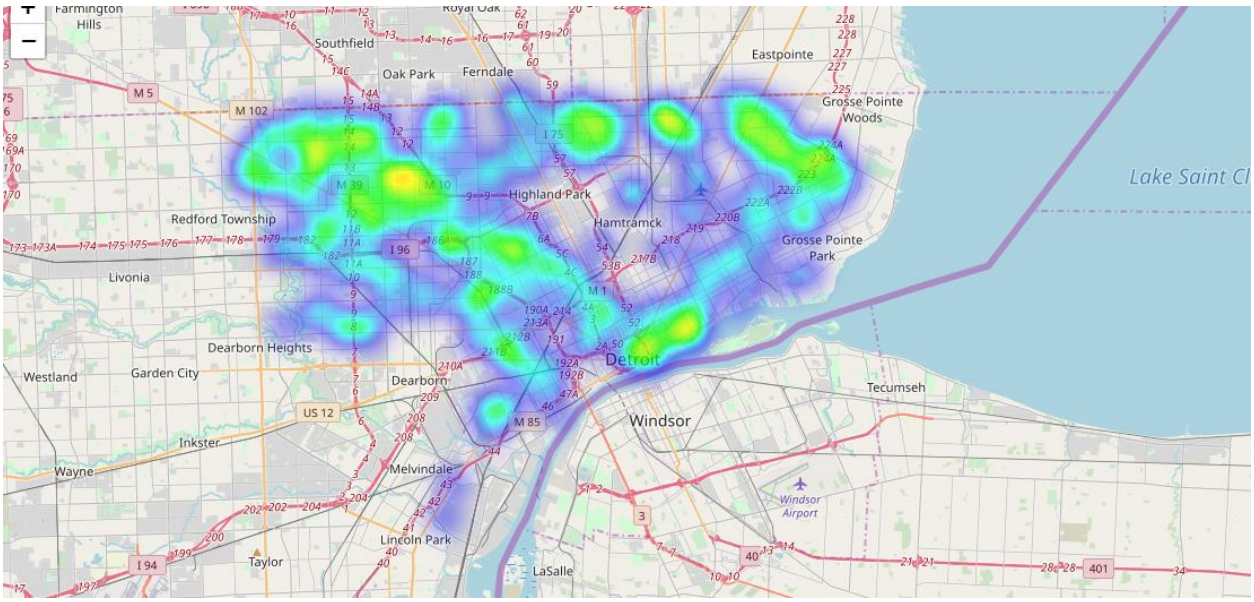
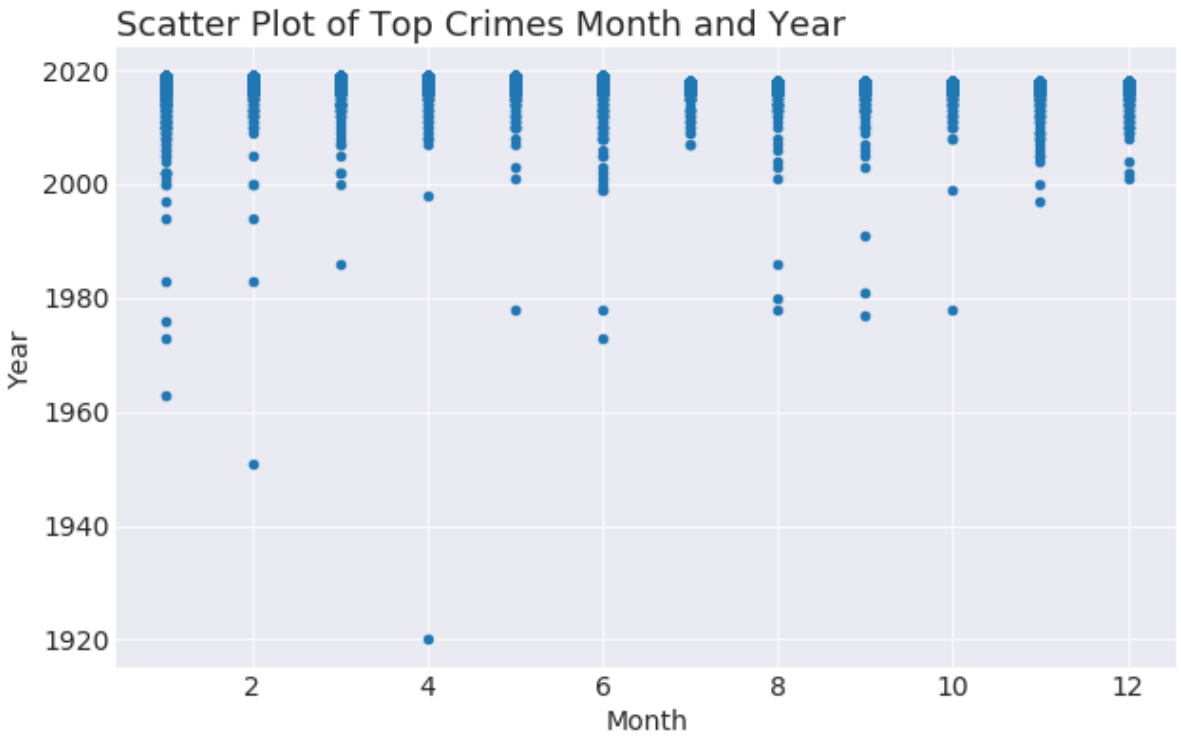
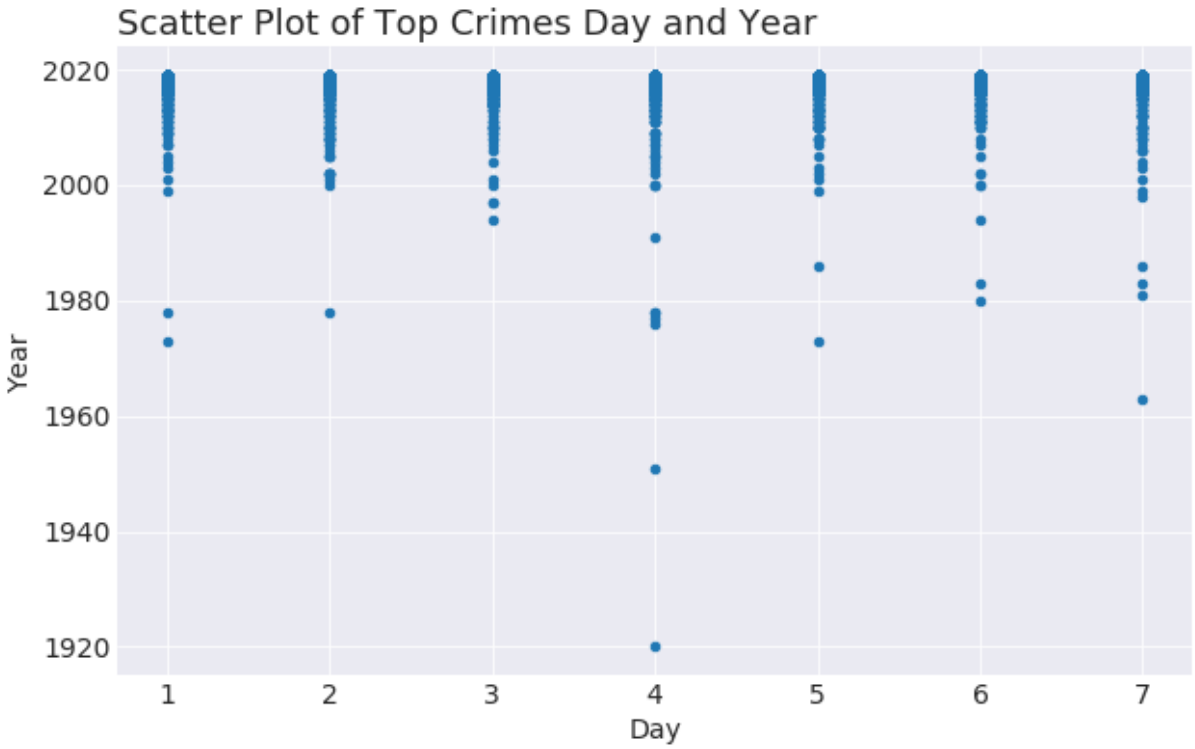


Figure 12 Heat map of Detroit showing crimes that happened on noon in June,2018

8. January has been the most dangerous month



9. Wednesday has been the most dangerous day from 1920 to present in Detroit



10. The model suggests latitude and longitude are the most important feature that influence the model.

Conclusion

Based on the analysis, this project offers a way for travelers to analyses venues to travel from foursqaure api quickly. The crime data analysis suggests the following:

- Avoid travelling to Detroit in January
- Take extra precautions on Wednesday if planning to travel
- Avoid travelling to Detroit city peripheral areas as they are more prone to crime
- Avoid being outdoors during noon

Future Work

- Link venue data to crime data and make a combined visualization for on the spot recommendations based on crime level of locality.
- Include more dependent variables and collect more dataset to create more accurate predictive model.