Capstone-Report

November 29, 2020

1 IBM Data Science Professional Certificate Capstone Project

1.1 Centers of crime and nearby popular venues in Tacoma, Washington

1.1.1 Introduction

This study explores relationships between locations with high rates of reported crime and popular venues nearby in Tacoma, Washington.
Tacoma is an important seaport city on the Puget Sound. In 2012 Tacoma ranked as the 11th busiest container port in North America. The ports of Tacoma and Seattle, only about 30 miles apart, merged in 2015 to create the Northwest Seaport Alliance, which is now the 5th largest port in North America. Tacoma, with an estimated population of 217,827, has earned the nickname "Grit City" due to its heavy industry and tough working class character. Tacoma is also famous in the region for its "aroma".

This study could be of interest to public policy makers, law enforcement, and the concerned citizens of Tacoma. If we find crimes clustered around certain types of venues, then we can seek to understand why. We can even take action to prevent it, or at least avoid being a victim. My initial hypothosis is that we will find increased incidents of certain types of crimes, such as robbery and assult near certain types of venues such as night clubs and strip clubs. Let's find out what the data tells us.

1.1.2 Data

The Tacoma crime data used in this report is available from cityoftacoma.org at https://data.cityoftacoma.org/Public-Safety/Tacoma-Crime/wtqi-kpsn/data. This dataset contains 118k rows and 5 columns (Incident Number, Crime, Occurred On, Approximate Time, and intersection). The intersection column contain the latitude and longitude, which will need to be parsed into two new numeric columns. The incidents span the 6 year period from the beginning of 2014 to the end of 2019.

Data about venues is from Foursquare.com. I will use the "Venue Recomendations" endpoint, which "returns a list of recommended venues near the current location".

I will group the crime data by latitude, longitude, and crime type, and get the count of incidents of crime type at location. Then I will plot the top 100 groups by incident count on the map of Tacoma. Each of these top locations will be passed as the "current location" required by the "Venue Recomendations" endpoint. The recomended venues will be plotted on the map of tacoma along with the crime locations.

To find correlations between crimes and venue categories, I will use one-hot encoding to turn the categorical data into numeric data. Then I will use Pearson correlation to determine if there are any significant relationships between high concentrations of particular types of crime and categories of popular venues in the same location, and also between different crime types at the same location.

1.1.3 Methodology

In this section, I use Pandas to explore and profile the data. Initially, there are 118,182 rows and 5 columns of crime data. Using the min and max functions on the date column reveals the date range of the data is from January 2014 through December 2019. The latitude and longitude data is embedded in the text description of the city block where the incident occurred. I add numeric columns to the dataframe for latitude and longitude, and parse the city block description to populate the new columns. Some rows do not contain latitude and longitude. Those rows are useless, so I cleanse the data by dropping those rows. After data cleansing, there are 110,842 rows and 7 columns. I group the crime data by location and crime and plot the top locations on the map of Tacoma using Folium. I use the Foursquare API to learn about popular venues within a small radius of the crime locations and add these to the Folium map as well.
The goal is to find correlations between crimes and venue categories. These are categorical data, so I use Pandas get_dummies function to create a one-hot encoded, numeric dataframe containing the crimes and venue categories. Then I use the corr function to find Pearson correlation coefficients between crimes and categories, and crimes and other crimes.
In the Results section, I use Seaborn to visualize the discoveries obtained from the data.

A cursory look at the dataset shows that we have 118k rows and 5 columns. The incidents span the 6 year period from the beginning of 2014 to the end of 2019.

The latidude and longitude are embedded in the intersection column. To make grouping and mapping easier we will add two new columns to the dataframe for latitude and longitude.

Some rows are missing latitude and longitude, making them useless for this invenstigation. We will drop those rows.

Now the data is ready to work with. We have 110k rows of data after cleansing.

Let's look at the number of incidents for each type of crime.

```
In [293]: tacoma_crime.Crime.value_counts()
```

```
Out[293]: Theft From Motor Vehicle                        20888
          All Other Larceny                               18004
          Burglary/Breaking & Entering                    14077
          Motor Vehicle Theft                             11493
          Destruction/Damage/Vandalism of Property        11179
          Shoplifting                                      6678
          Simple Assault                                   5800
          Aggravated Assault                               3386
          Credit Card/Automatic Teller Fraud               2944
          Robbery                                          2718
          Impersonation                                    2570
          Theft of Motor Vehicle Parts/Accessories         2504
          False Pretenses/Swindle/Confidence Game          2038
          Counterfeiting/Forgery                           1411
          Identity Theft                                    706
          Drug/narcotic Violations                          665
```

```
Weapon Law Violations                           504
Violation of No Contact/Protection Order        501
Arson                                           479
Stolen Property Offenses                        453
Theft From Building                             353
Pornography/Obscene Material                    277
Extortion/Blackmail                             193
Wire Fraud                                      184
Kidnaping/Abduction                             168
Intimidation                                    162
Purse-Snatching                                 136
Prostitution                                     80
Pocket-Picking                                   74
Murder and Nonnegligent Manslaughter             55
Drug Equipment Violations                        51
Theft From Coin Operated Machine or Device       41
Embezzlement                                     26
Welfare Fraud                                    24
Assisting or Promoting Prostitution              12
Human Trafficking/Involuntary Servitude           2
Human Trafficking/Commercial Sex Acts             2
Negligent Manslaughter                            2
Justifiable Homicide                              1
Hacking/Computer Invasion                         1
Name: Crime, dtype: int64
```

Latitude and longitude values are not the exact location of the crime, but the coordinates for the 100 address of the block. That means we can group crimes by location as well as the type of crime. To start with, we will just look at the top incident count by crime and location overall.

```
In [7]: groups = tacoma_crime.groupby(['Latitude','Longitude','Crime'])['Incident N
        # let's just look at the top 100 crime at location groupings
        tacoma_crime_groups = pd.DataFrame(groups, columns=['Latitude','Longitude',
        print(tacoma_crime_groups)
```

```
     Latitude   Longitude                           Crime  Count
0   47.217857 -122.467608                     Shoplifting    550
1   47.217857 -122.467608               All Other Larceny    526
2   47.242213 -122.482564                     Shoplifting    392
3   47.217857 -122.467608        Theft From Motor Vehicle    341
4   47.242290 -122.482583                     Shoplifting    263
..        ...         ...                             ...    ...
95  47.242290 -122.482583        Theft From Motor Vehicle     46
96  47.242213 -122.482564    Burglary/Breaking & Entering     46
97  47.231273 -122.495200             Motor Vehicle Theft     46
98  47.238167 -122.479216               All Other Larceny     46
99  47.257194 -122.443916                  Simple Assault     45
```

```
[100 rows x 4 columns]
```

```
In [296]: tacoma_crime_groups.Crime.value_counts()
```

```
Out[296]: Theft From Motor Vehicle                          35
          Shoplifting                                       28
          All Other Larceny                                 19
          Simple Assault                                     6
          Burglary/Breaking & Entering                       5
          Robbery                                            2
          Motor Vehicle Theft                                2
          Aggravated Assault                                 1
          Pornography/Obscene Material                       1
          Destruction/Damage/Vandalism of Property           1
          Name: Crime, dtype: int64
```

The largest groupings crime at location, by far, are “Theft From Motor Vehicle”, “Shoplifting”, and “All Other Larceny”. We also see groupings of other crimes, including violent crime.

Now we will see how these points are distributed on the map.

We see the crimes are mostly clustered around downtown, along the freeways, and at the Tacoma mall. These are primarily the business districts and not residential areas.

Now we will explore the popular businesses nearest to these crime clusters using the Foursquare API.

1.2 Results

In this section, I will summarize the discoveries obtained from the data.

We see strong correlations between different crimes at the same location, but with only two exceptions, correlations between crimes and venue categories are moderate to weak. When using a 100 meter radius, we see strong correlation (0.704907) between the category “Seafood Restaurant” and the crime “Burglary/Breaking & Entering”, and 0.573753 between the category “Music Store” and the crime “Pornography/Obscene Material”. When considering a 200 meter radius, the strongest correlation coeficient is 0.311282 between the category “Seafood Restaurant” and the crime “Robbery”.

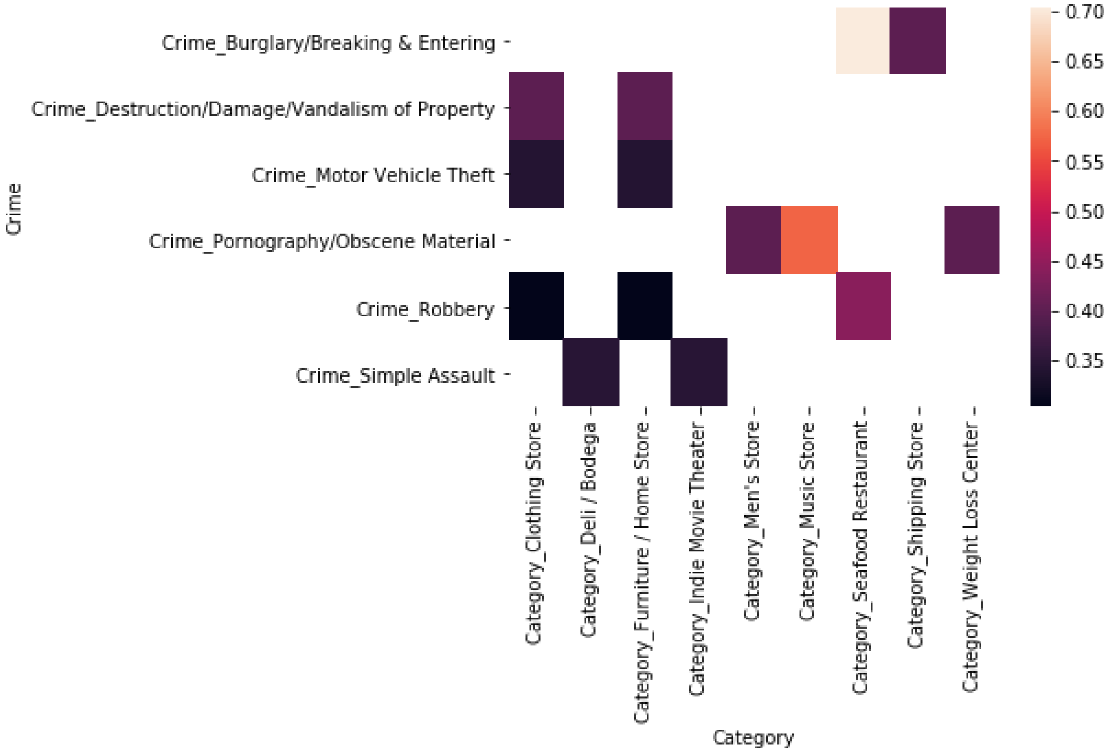
```
In [190]: # remove correlations between one crime and another crime. I only want to
          crime_to_crime = [item for item in corr_100.index.values if item[1].start
          corr_100.drop(index=crime_to_crime, inplace=True)
          corr_100
```

```
Out[190]:
          Crime_Robbery                                      Category_Clothing Store
                                                             Category_Furniture / Home
          Crime_Motor Vehicle Theft                          Category_Clothing Store
                                                             Category_Furniture / Home
          Crime_Simple Assault                               Category_Indie Movie Theat
                                                             Category_Deli / Bodega
          Crime_Burglary/Breaking & Entering                 Category_Shipping Store
```

```
Crime_Destruction/Damage/Vandalism of Property Category_Clothing Store
                                               Category_Furniture / Home
Crime_Pornography/Obscene Material             Category_Weight Loss Cente
                                               Category_Men's Store
Crime_Robbery                                  Category_Seafood Restaurar
Crime_Pornography/Obscene Material             Category_Music Store
Crime_Burglary/Breaking & Entering             Category_Seafood Restaurar
```

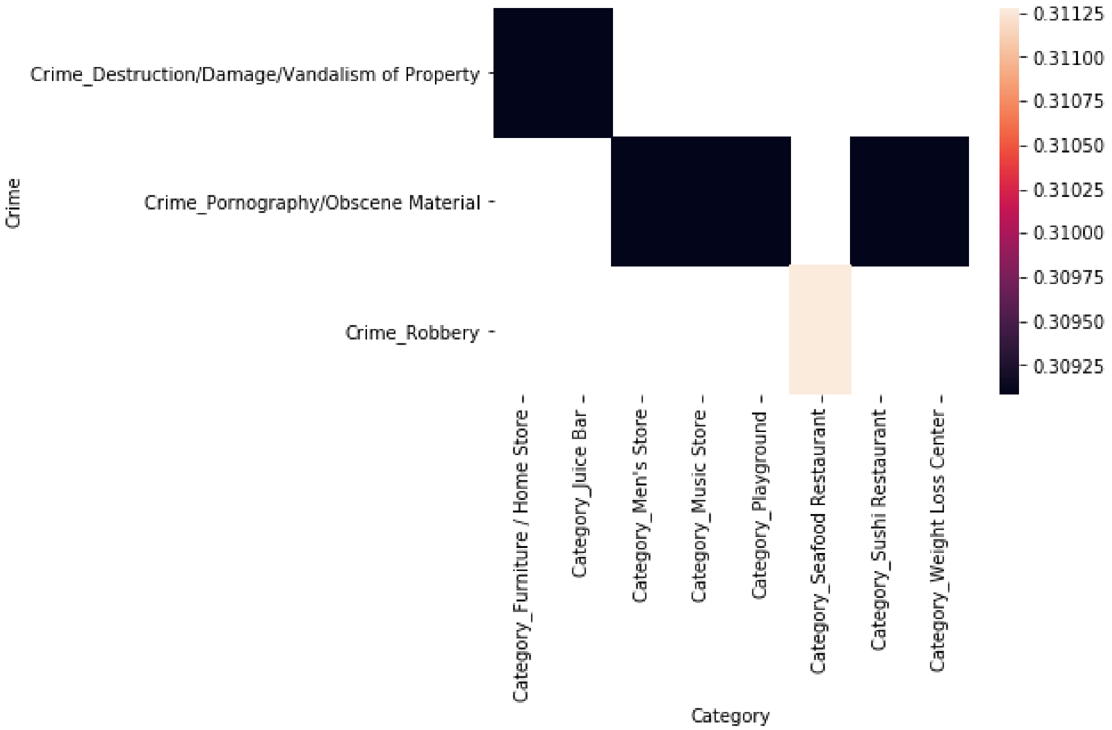
```
In [188]: # Heatmap with a 100 meter radius from crime location
          import seaborn as sns
          import matplotlib.pyplot as plt
          %matplotlib inline

          sns.heatmap(p_100)
          plt.show()
```



```
In [194]: # Heatmap with a 100 meter radius from crime location
          import seaborn as sns
          import matplotlib.pyplot as plt
          %matplotlib inline

          sns.heatmap(p_200)
          plt.show()
```



1.3 Discussion

Disappointingly, with a couple of exceptions, only moderate to weak correlations were found between any crime location and the category of popular venues in the vicinity. The strong correlations that were found would appear to be only coincidental. That is to say, I can't think of any reason to expect seafood restaurants to be highly correlated with Burglary/Breaking & Entering.

1.4 Conclusion

In conclusion, I did not find sufficient evidence of any significant correlation between the places where crime occurs and the types of venues in the vicinity.
I did find that certain crimes are highly correlated with one another, such as Robbery with Destruction/Damage/Vandalism of Property and Motor Vehicle Theft with Destruction/Damage/Vandalism of Property. A likely explanation for this is that a single event often involves both crimes.
It is also worth note that places like parking lots and garages are not venues, per se, so are not returned by Foursquare's venues API. Visually inspecting the map makes it clear that Theft From Motor Vehicle occurs frequently in large parking areas, such as the Park and Ride garage and the Tacoma Mall. Perhaps more interesting results could be achived by repeating the study with a different set of location data.

```
In [ ]:
```