An Analysis of Crime in Chicago: Can Venues be Used to Predict Crime Rates?

Problem and Background:

For this capstone project, I decided to revisit the Chicago crime dataset that was used in Course 5 of the IBM Data Science Professional specialization, Databases and SQL for Data Science. The city of Chicago is the U.S' third largest city by population behind New York City and Los Angeles. For years, Chicago has struggled with crime, and still is today. However, in recent years, the city has made significant progress in its fight against crime and overall, crime has declined by about 10% since 2016 (Kirkos, 2019). Some categories of crime like shootings, robberies, and car jackings have even seen steeper declines than that (Kirkos, 2019). According to Kirkos (2019), the Chicago Police credit the recent drops in violent crimes in part to their "investments in data-driven policing", which uses data from different sources to identify crime locations and potential crime locations. For this project, I intend to build a model that contributes to these efforts by analyzing the types of crimes occurring in the different community areas and the types of venues that exist in them to see if there is any correlation between the two. Simply put, this study is to determine if the types of venues in a community can be a good predictor of the crimes that exist in that community. This determination will be done by using a combination of clustering and classification algorithms, which will be explained further in the methodology section. When the determination is made, depending on how strong the relationship between crimes and venues is, the findings of this study can be of interest to policy makers and the Chicago Police Department in their efforts to predict crime and develop anti-crime strategies.

Data Requirement:

The data used is gotten from the Chicago data Portal, which holds crime data from 2001 to present. For this project, the working dataset includes only records for the last 5 years, 2014 to 2019. As of April 9, 2019, this dataset contains 1,399,203 records (or rows) with 22 columns (or features). The data can be gotten from https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2 . The top features are Primary Type (which will be relabeled as category) and Community Area (which is a community number that will be matched to a different data set to get the community area name). The 77 community area names and numbers were gotten from another document published by the City of Chicago, which can be gotten at https://www.chicago.gov/content/dam/city/depts/doit/general/GIS/Chicago_Maps/Citywide_Maps/Community_Areas_W_Numbers.pdf . Even though the Chicago Data Portal also offers some data files with that information (including CSV, JSON, and GeoJSON), it's simpler to just copy the names from that document and post in an Excel spreadsheet, since the other source files have a lot of other data that are not relevant to this project. With a little work on the Excel spreadsheet, a clean CSV file of community area numbers and names can be produced. Another piece of data needed for this project is the geographic coordinates of the different community areas in order to visualize them on the map. This data will be gotten through the Geocoder Python package which is documented at https://geocoder.readthedocs.io/index.html . The final dataset that will be needed is the data about the types of venues within each community. This dataset will be gotten from Foursquare through their Places API. This data is returned in JSON format, and the relevant features will be extracted and added into a Pandas data frame.

Methodology:

This section is divided into 2 subsections, exploratory data analysis and machine learning. The first explains the basic analysis that was conducted on the data to get a better understanding of it, while the second explains the machine learning approaches that were used to test for a relationship between the types of venues and the types of crimes.

Exploratory Data Analysis:

Community Areas:

The first check on the data was to visualize the locational data for the community areas that was gotten from the Geocoder package to verify that the locations were accurate. All except one, South Deering (Community Area 51), were accurate. The location finally used for South Deering was gotten from a Google search and corrected into the data frame. Below is a map of the community areas of Chicago prior to clustering.

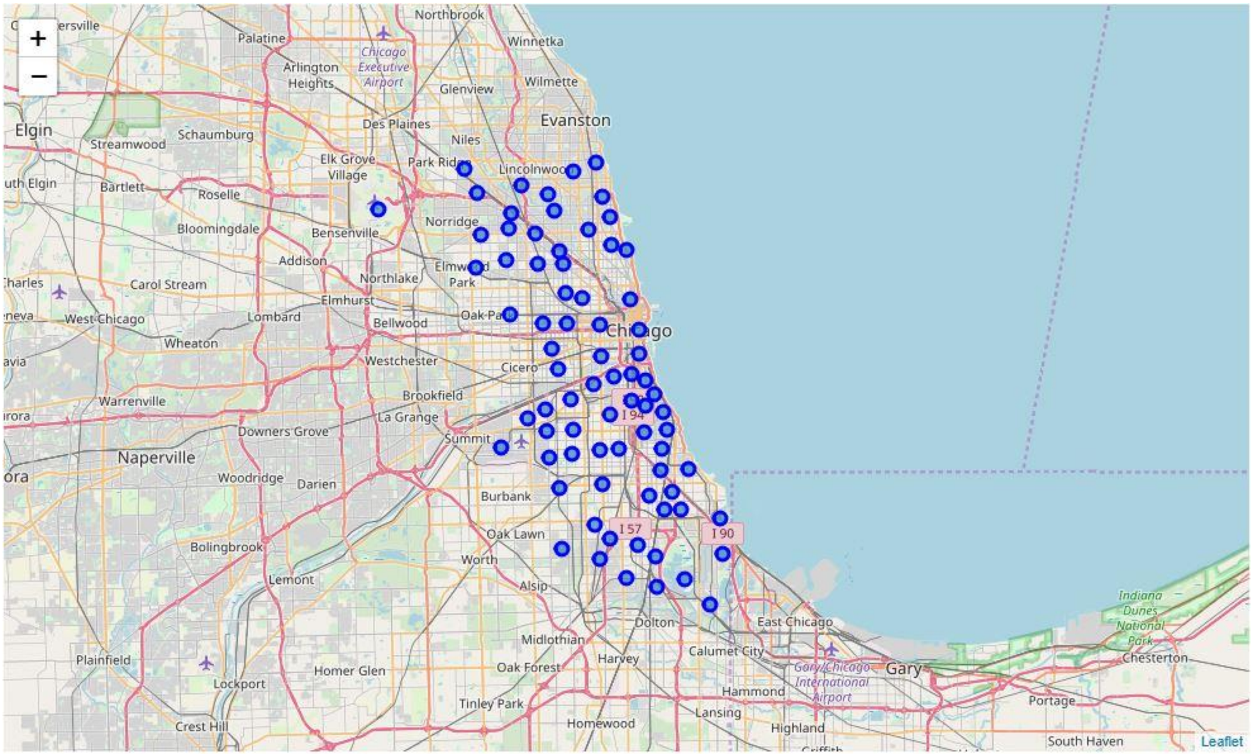

Figure 1: Chicago Community Areas

Crime by Community:

To identify the community areas with the highest crime rates between 2014 to April 2019, the data was grouped by community areas and ranked in descending order. The result showed the top 5 community areas with the most crimes were Austin, Near North Side, Loop, North Lawndale, and South Shore, with Austin significantly distant from the others. The chart below visualizes this observation.

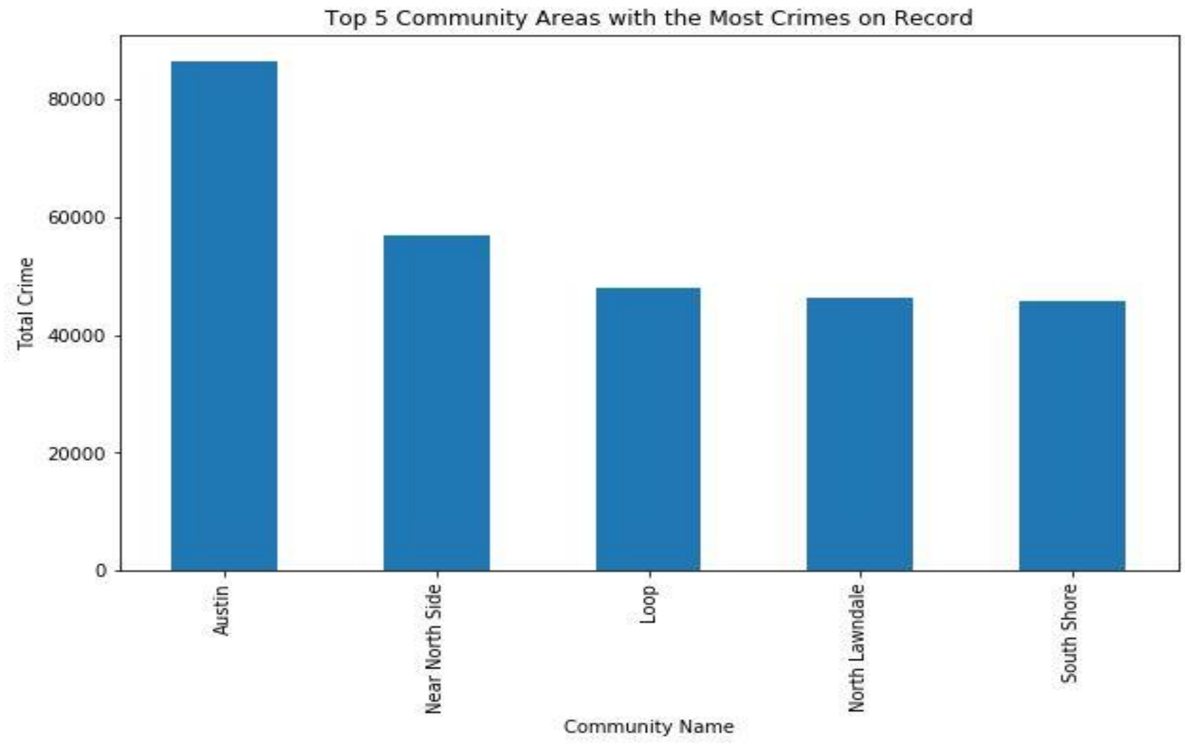

Figure 2: Top 5 Community Areas in Chicago with the Most Crimes (2014 to April 2019)

Total Crime Rates from 2014 to 2018:

To analyze the year over year overall crime trend across Chicago from the data, the dataset was grouped by year. Since at the time of the analysis, 2019 is still in progress, all records for 2019 were excluded. Hence, the chart below only shows the crime trend from 2014 to 2018. The chart shows a steep decline in total crime from 2014 to 2015, an increase in 2016, and a gentle decline from 2016 to 2018. Overall, crime in Chicago has reduced for the observed period (2014 to 2018). These observations are in tandem with Kirkos' (2019) reporting, which credits these drops in crime to "investments in data-driven policing". As for the rise in 2016, this has been blamed on the sharp decrease in street stops that resulted from the "2016 agreement between the Chicago Police Department and the American Civil Liberties Union of Illinois to reform unconstitutional stop-and-frisk practices" (Wheeling, 2018). The result was a sharp increase in homicides, which affected the overall crime rates and the national homicide rates.

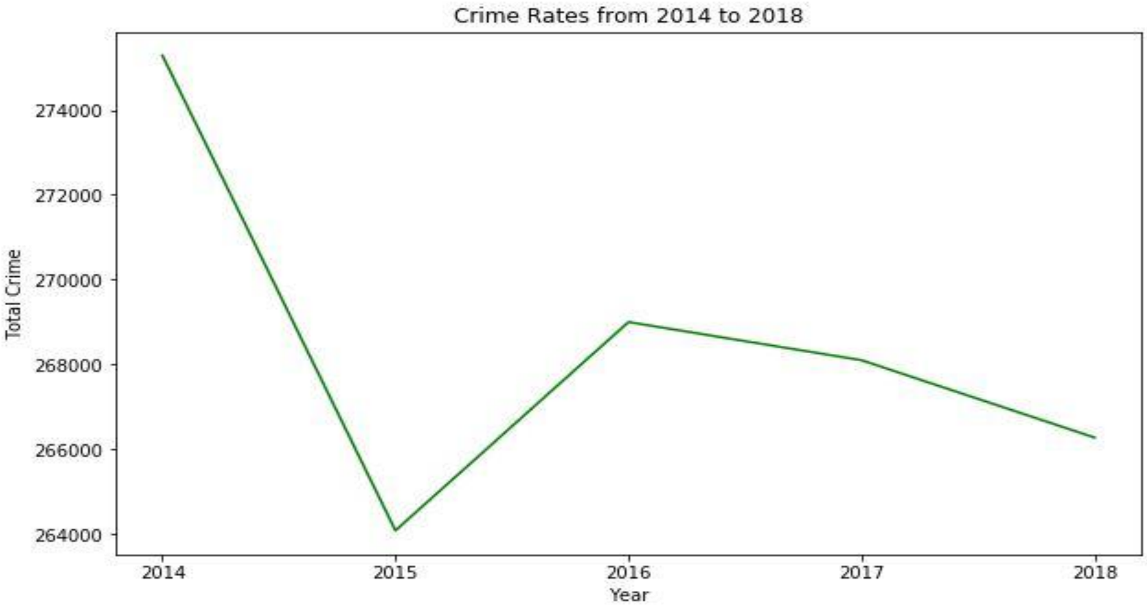

Figure 3: Crime rate in Chicago from 2014 to 2018

Crime Trends for Top 5 Crime Categories:

In addition to analyzing the overall crime trend through the years, the same analysis was done for the different crime categories, specifically the top 5 crime categories. This was done by charting the frequency for the top 5 crime categories, theft, battery, criminal damage, assault, and narcotics. Of these 5 categories, theft, battery, criminal damages, and assault have seen fluctuations through the observation period, there've been years with slight rise followed by years with declines of similar proportion, over and over again. The noteworthy observation was the steep decline in narcotics related crimes during that period. This can be attributed to the relaxation of penalties for nonviolent drug offenders and the expansion of programs that help nonviolent drug offenders receive treatment for substance abuse (Ramos, 2018). The chart below shows this trend.

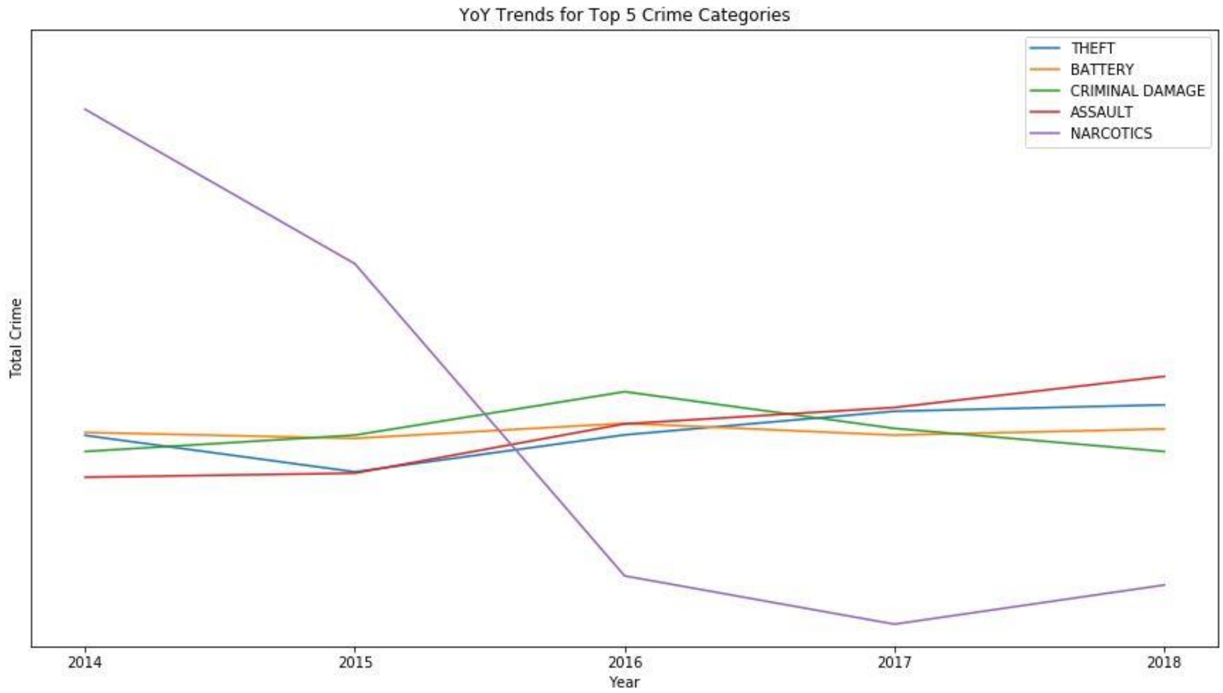

Figure 4: Year over Year Trend for Top 5 Crime Categories in Chicago.

Arrests Versus No Arrests:

With the data records, some of the incidents resulted in arrests while others did not. The chart below shows the observation that majority of the incidents on record during this period resulted in no arrests. The data set doesn’t give any indication as to why arrests may be made or not.

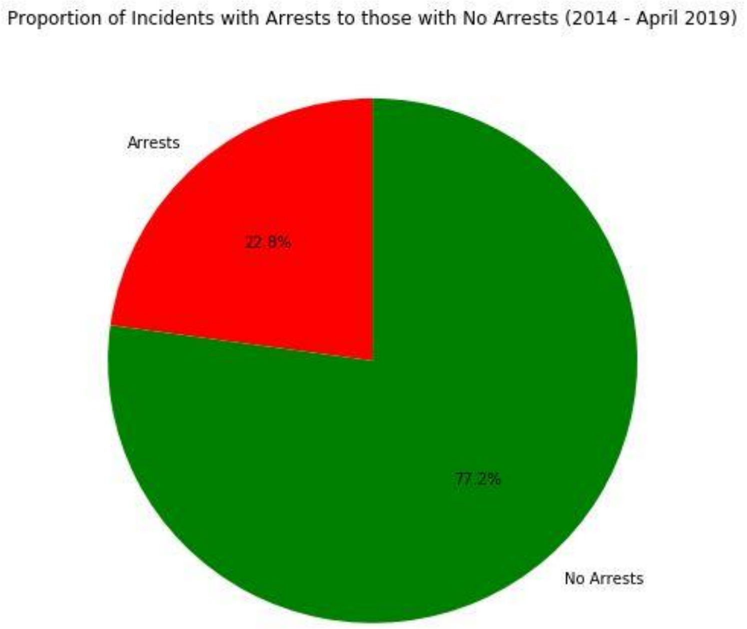

Figure 5: Proportion of Incident with Arrests to those Without Arrests (2014 to April 2019)

Locations with the Most Crime:

This final data exploration exercise was to identify the locations with the most crime. The observation was that the streets and private residences are the locations with the highest number of crimes for this period. Although the dataset separates streets from sidewalks and residences from apartments, those pairs can be grouped. The first pair are transit locations while the second pair are housing locations. Essentially, most of the crimes occur where people live and on the routes through which they travel to conduct their day-to-day activities.

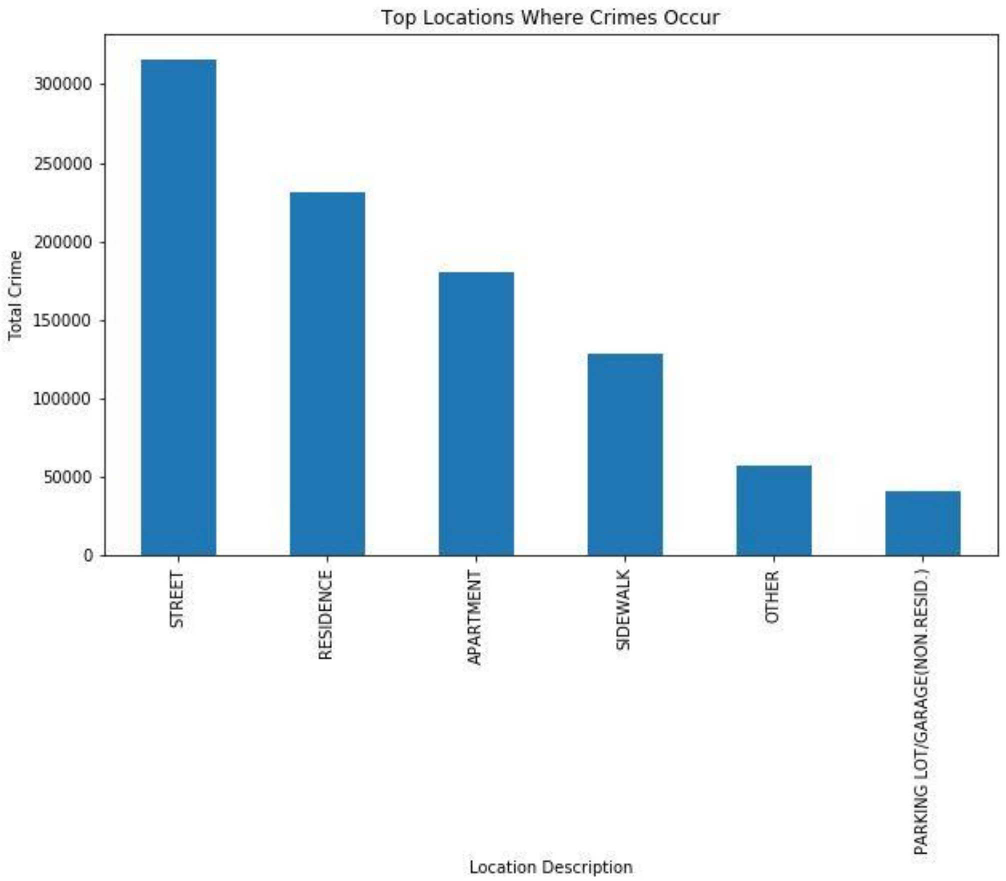

Figure 6: Top Locations in Chicago Where Crimes Occur.

Machine Learning:

Having analyzed the crime dataset from the different dimensions of community areas, time, crime category, arrest or no arrests, and locations, I set out to determine whether the types of venues in the different community have a correlation with the crime rates in those communities. This section used a combination of clustering, classification, and multiple regression to make this determination. The results of the different findings are discussed in their respective sections.

Clustering:

The first part was to use clustering to group the community areas into 2 clusters based on the frequency of the different crime categories in those communities. This was done using Python's built-in K-Means clustering algorithm. The number of clusters was set to 2 so that the target or dependent variable of the upcoming classification should be binary and suitable for logistic regression. Below is a visualization of the clustered map.

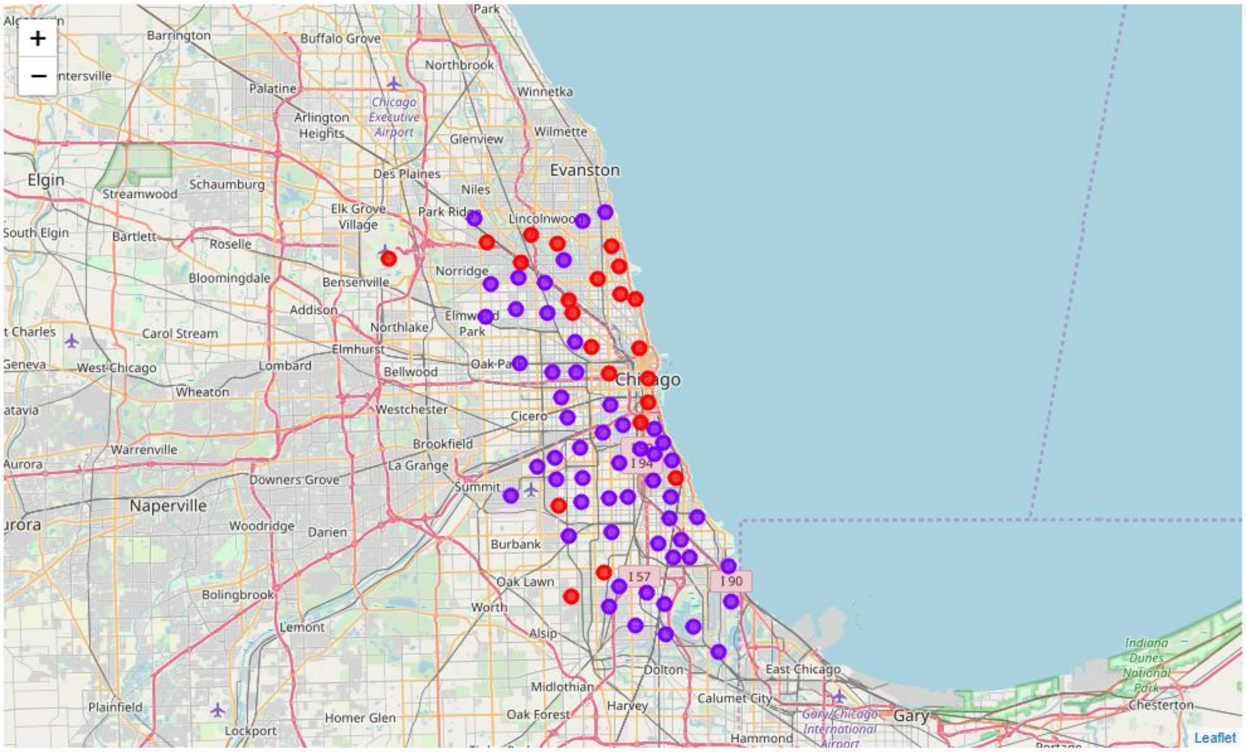

Figure 7: Clusters of Community Areas in Chicago Based on Average Frequency of Crime Categories

Classification:

In a similar way in which the average frequency of crime categories was derived, a data frame was built to contain the average frequency of venue categories for each community. For each community area, the corresponding cluster labels from the clustering section above (0 and 1) was merged into this data frame and used as the target variable or label that needs to be predicted. So, essentially, the goal of this classification model is to use the venue categories in a community area to predict which cluster the community area will belong to. The data was split into 70% to 30% test-to-training ratio and logistic regression was used as the classification algorithm for training the model. Upon using the test sample to evaluate the model accuracy, the Jaccard index showed a 57% similarity score between the test labels and the predicted labels, the F1-score was 0.41 (which is far from a perfect score of 1 on a scale from 0 to 1), and the Log loss was 0.77 (which is also far from a perfect score of 0 on a scale from 0 to1). Although the Jaccard similarity score was almost encouraging, the other evaluation metrices were discouraging.

Multiple Regression:

In a second attempt to determine whether venue categories within a community could be a strong predictor of the overall crime rates within that community, multiple regression was used. The multiple regression model took in the average frequencies of the different venue categories as its predictor variables and the total crime for each community as its target variables. Upon evaluating the model's performance, the residual sum of squares was really high (which is no good), and the coefficient of determination was -0.67 (also terrible). Although the typical range for the coefficient of determination is from 0 to 1, with smaller values worse than larger values, negative values are possible with the python function that was used. According to the scikit-learn.org website, negative coefficients of determination indicate that the model is arbitrarily worse. In general, the coefficient of determination or R^2 explains how much of the variance between the actual values and the predicted values is explained by the model.

Discussion and Conclusion:

Based on the observations from the analysis above, the types of venues in a community are not good indicators of the types of crimes that occur in the community or the community's overall crime rates. However, the above 50% accuracy in the classification model is an interesting observation that can be made a subject for future research. In that alternate research, I'll recommend testing for relationships between specific crime categories and venue categories as opposed to using all the crime categories (33) and venues categories (246) for the entire analysis. This is because using all the categories as features resulted in many sparse predictor variables with values of 0. While as a whole, the venue categories may not be strong predictors of the crime categories or total crime rates, there might be stronger correlations between specific pairs of venue categories and crime categories. That will require iterative modeling and evaluation and is an exercise for another day.

The code for this project can be found at :
https://github.com/joekmfonfu/ibm_ds_capstone_porject/blob/master/capstone_code.ipynb

This report is also available on my blog at: https://datascience-jkm.blogspot.com/

Reference

Kirkos, B. (2019, January 01). Chicago murder rate drops for second year in a row. Retrieved April 9, 2019, from https://www.cnn.com/2018/12/31/us/chicago-murders-drop-2018/index.html

Ramos, E. (2018, February 4). Chicago Drug Arrests Reach Historic Lows, But Those Busted Could Still Fill Stadiums. Retrieved May 14, 2019, from https://www.wglt.org/post/chicago-drug-arrests-reach-historic-lows-those-busted-could-still-fill-stadiums#stream/0

Wheeling, K. (2018, March 30). What Caused Chicago's Spike in Violent Crime? Retrieved May 14, 2019, from https://psmag.com/social-justice/chicago-spike-in-violent-crime