Restaurant Crimes in Chicago

Ricky Donnell Lindsey Jr.

June 8, 2020

1. Introduction

1.1 Background

The City of Chicago is the third-most populous city in the United States of America. Comprised of 77 community areas, with more than 7,300 restaurants.

1.2 Problem

With an estimated population of 2,693,976, within 234 square miles of land, crime is inevitable. Understanding where crimes are most likely to occur as well as what crimes is most likely to occur in each location is essential for any business owner. For this analysis the data will be approached from the perspective of a potential restaurateur, looking for a location to open a restaurant in Chicago.

2. Data Acquisition and Cleaning

2.1 Data Source

The City of Chicago uses the Chicago Open Data Portal to make much of its public data easily accessible. The Chicago Crimes data set being used is open-source and can be found here. The data set reflects reported crimes across Chicago from 2001 to the present day. Data occasionally may be missing column data from a certain category, but it is mostly complete for most entries. The data set tracks twenty-two features, however only ten will be used: ID, IUCR, Primary Type, Location Description, Arrest, Beat, District, Ward, Community Area, and Year.

2.2 Data Cleaning

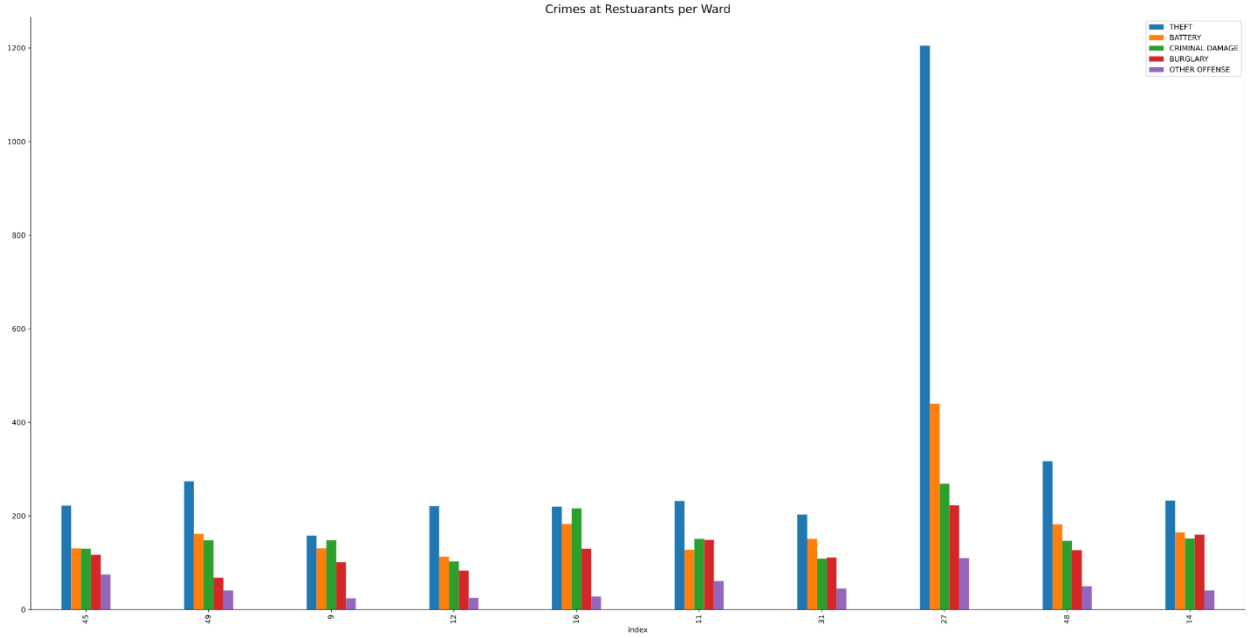
An initial overview of the 7,114,996 rows of data revealed some rather basic information about the data set. The data were filtered by Primary Type, to see the frequency of which crimes were committed as well as to list the unique crimes committed. The most common crime is THEFT. 21.15% of crimes reported are THEFT. The most common location is the STREET. 21.17% of crimes reported occur in the STREET. Behind THEFT, BATTERY CRIMINAL DAMAGE NARCOTICS ASSAULT and OTHER OFFENSE round out top five crimes, with BURGLARY at a close sixth.

The data was then filtered by Ward, Beat, and Community Area. Most communities tended to follow the same trend for crimes by primary type. No feature was best at identifying the optimal location however, sub-optimal locations with high crime rates were easily identifiable.

2.3 Feature Selection

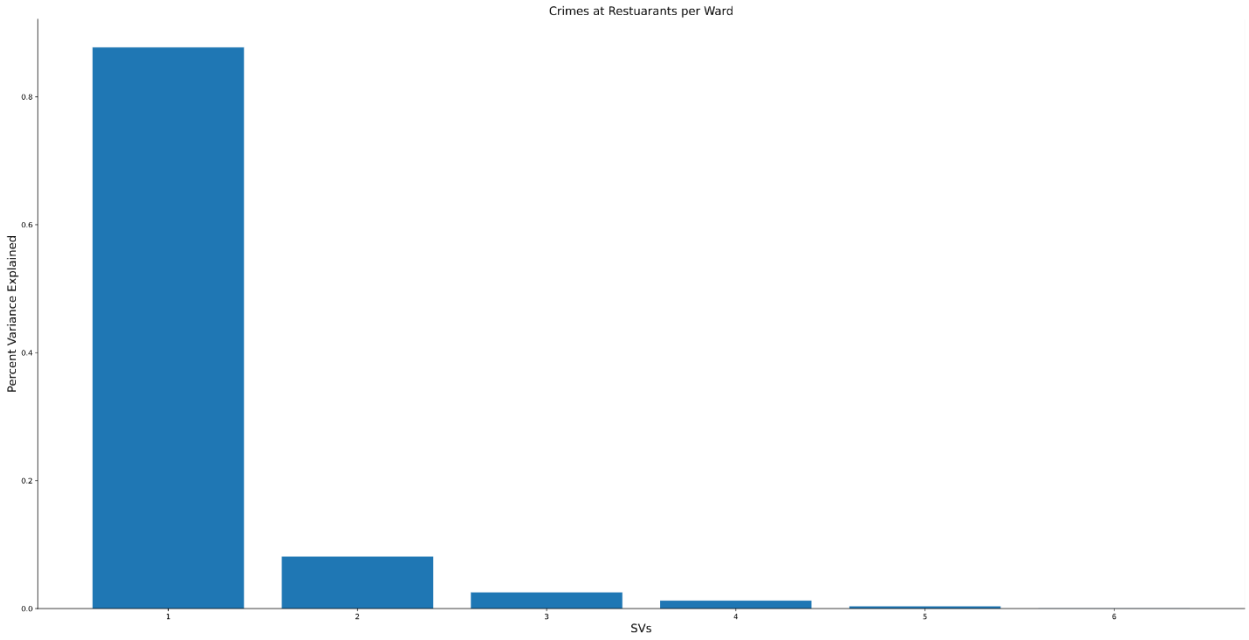
The data set was filtered to remove any rows that we're missing any column data. Upon which Location Description was now considered as the primary feature for filtering the data set.

Only criminal offenses that occurred at a restaurant were kept. Illustrating that on average the most common crime was theft at ~200 reported thefts at restaurants per Ward.

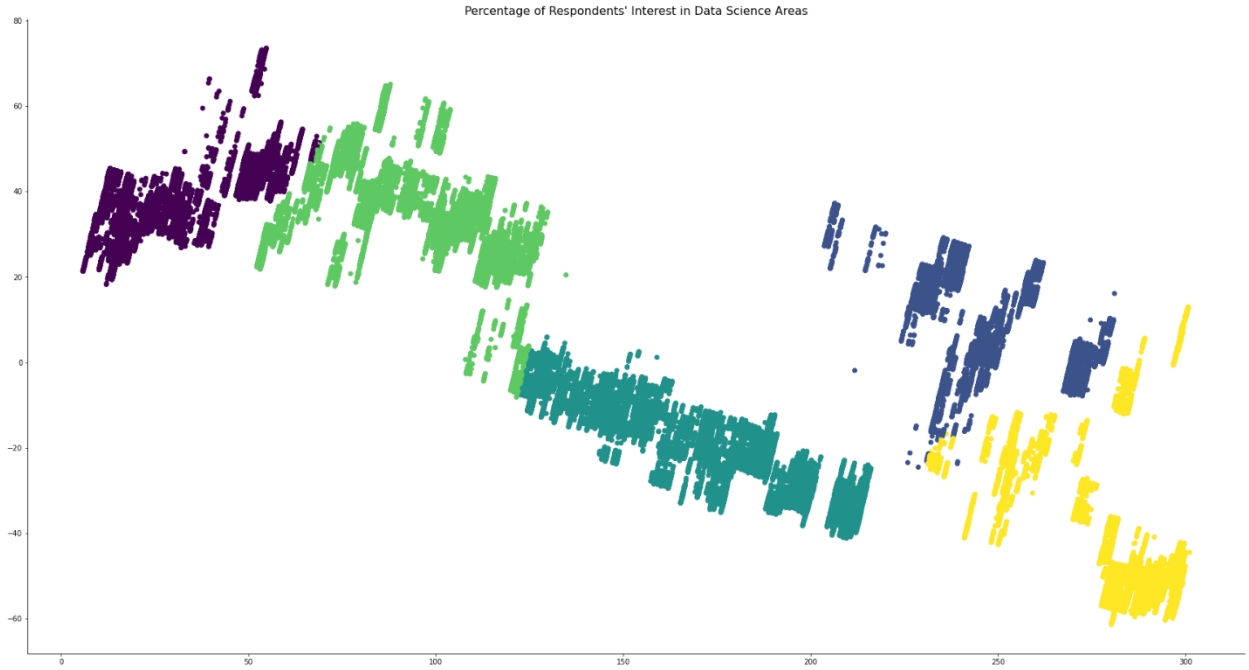


3. Exploratory Data Analysis

To better explore the data set containing only restaurants, all features were label encoded. Representing the values with numerical dummy values would allow for much easier analysis. To better understand the impact of each feature dimensionality reduction was now required. Truncated Singular Value Decomposition was used. Truncated SVD was used over PCA since the labeled values are discrete. The truncated SVD illustrated that that eighty-four percent of the variance of the data could be explained by the first singular value.



After plotting the data set for the axis of the first and second singular values, the clustering of the data points became more evident. K Means clustering algorithm was used to find an optimal number of clusters for fitting the data points. Five clusters seemed to best provide the best result for the data points.



4. Short Comings of Analysis

Further Analysis of the data set could not be completed for a couple of key reasons that will now be discussed. Firstly, the scope of the project from the beginning. The project in its inception was far too broad. Even within narrowing the scope, a precise question was never asked. Secondly, the project's data set was simply too large for an entry-level project. Although, having a great grasp on the theoretical knowledge, the application of said knowledge is not as easy. Ultimately like all real projects, time restraints have limited any further sinking of resources into this project.

5. Conclusion

From the dataset, I was able to illustrate that the is a sense of regularity to the crimes and where they occur throughout Chicago. However, limited ability to properly analyze the data set ultimately limited my question of 'Where would be the best place to establish a restaurant in The City of Chicago.

6. Future Improvements

The reason behind many of the issues was inexperience when handling the dataset. Dimensionality reduction on discrete data must be handled differently that of continuous data. This compounded with the large data set causes a plethora of issues beyond my current capabilities. Upon further reflection, the use of the relative frequency of each value count may have been a better methodology to use. This potentially could have negated the need for label encoding. The use of relative frequencies may allow for the use of Principal Component Analysis (PCA) for dimensionality reduction. Principle component analysis could then be used to better illustrate the explained variance of a given feature of the dataset.