CYO Capstone

Lily Norian

3/3/2022

This is my report for the Capstone CYO Project

###Introduction In this project, I will be using the "Crimes in Chicago" dataset offered by kaggle. Specifically, I will be using the 2012-2017 dataset. This dataset includes data for crimes reported in the city of Chicago during these years. The data includes a lot of important variables and information. However, for the purpose of my analysis, we will focus on District, Primary.Type, Location.Description, and Arrest These are explained below. It is also important to note that each incident is assigned a unique identifier which is in the ID column.

District- the Chicago police district where the crime happened Primary.Type- the primary description of the crime according to the Illinois Uniform Crime Reporting code Location.Description- describes the location of where the crime occurred Arrest- True or False value depending on if an arrest was made

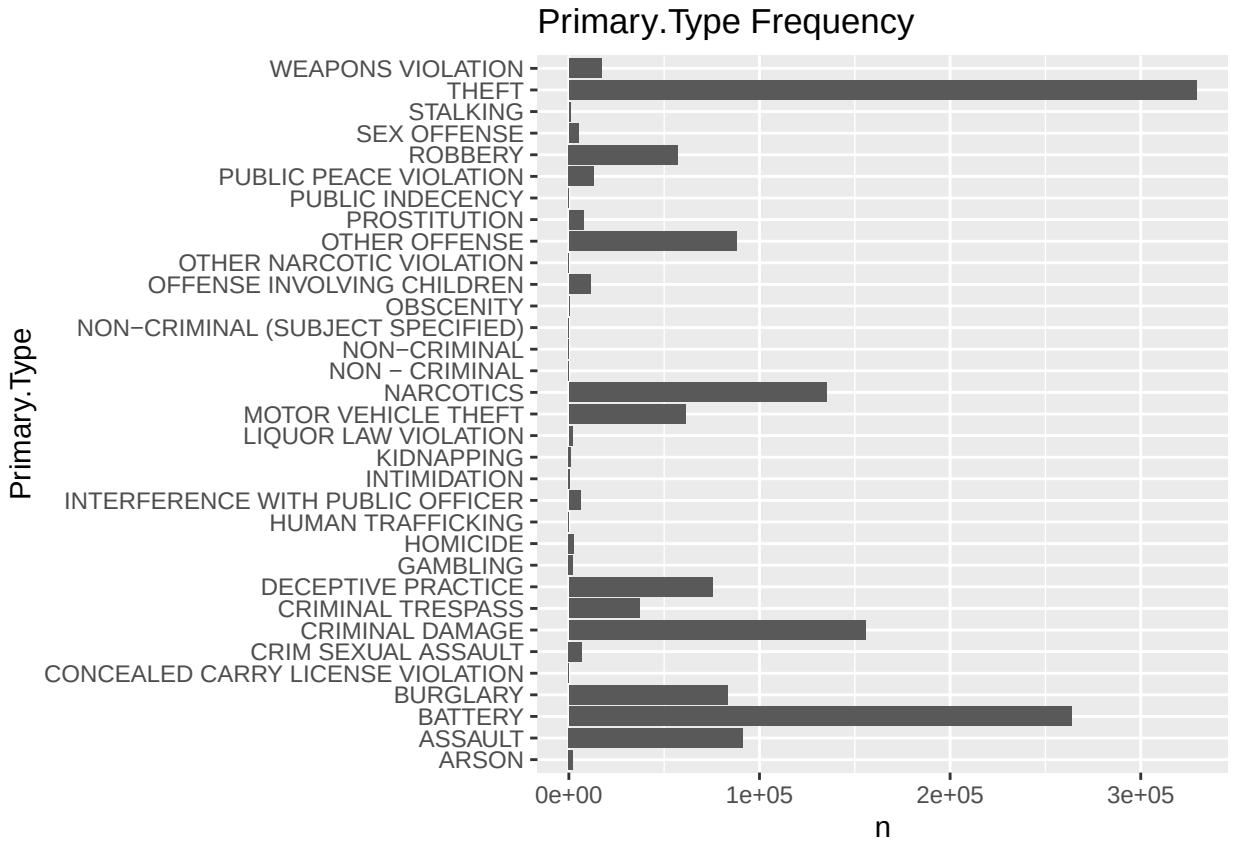
The goal of this project is to understand the relationship between these different variables as they relate to crime in Chicago. This will tell us the safest districts of Chicago, the most common crimes, and predict if an arrest will be made for a crime.

###Methods/Analysis After loading the data, I began my exploration. To start, I wanted to familiarize myself with the data. To accomplish this, I found the frequency of the different districts, crime types, and unique idenitifiers to ensure that there was no repetition. These values are shown below:

```
##      n_ID n_Primary.Type n_District
## 1 1456714             33         25
```

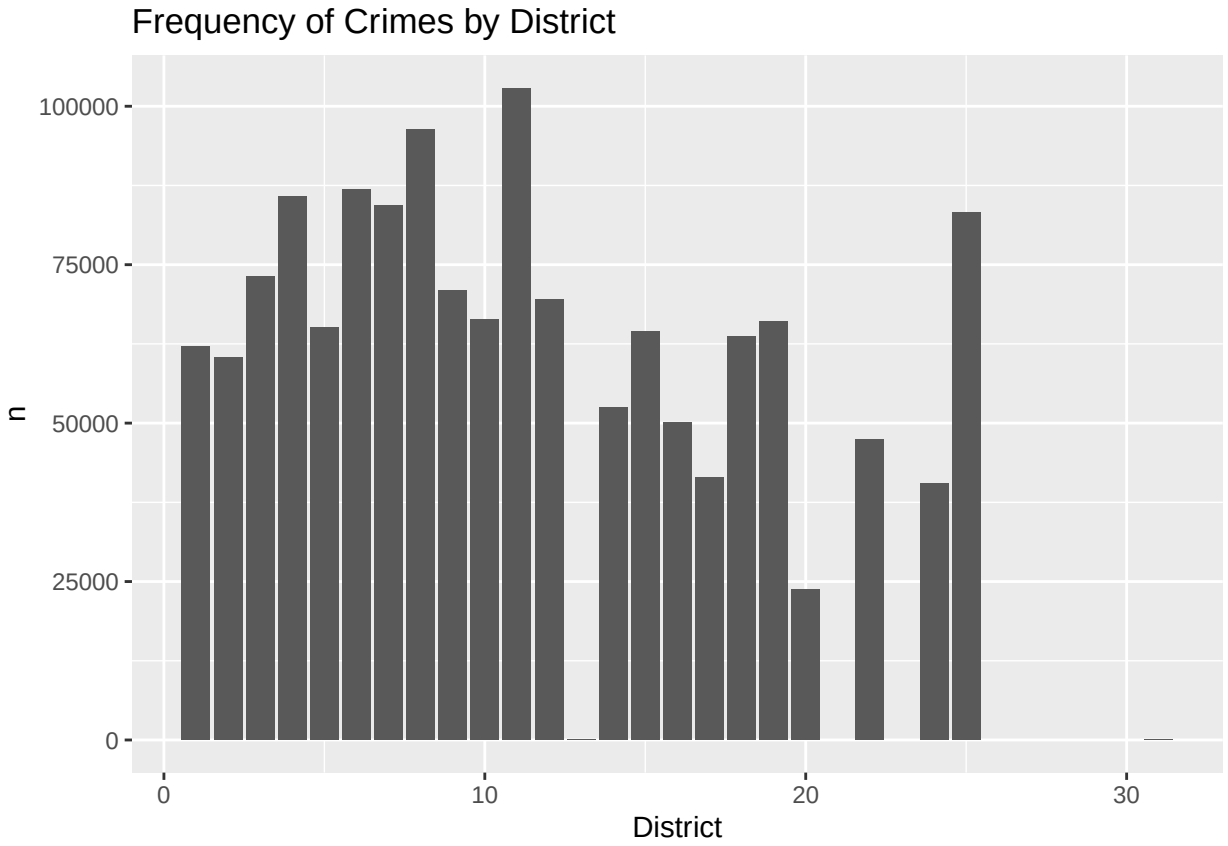
There were a couple columns that I deemed unnecessary for my analysis like the longitude and latitude and others that I removed from the data for future analysis. I removed any NAs and duplicates to clean the data. Then, I made a series of plots to show frequencies and relationships between variables.

The first plots show the frequencies of crime types and crimes in districts. These graphs as seen below show that theft is the most common crime and battery is the second most common. We can also see that district 11 had the most crimes reported. This gives us a better understanding of the data and crime environment in



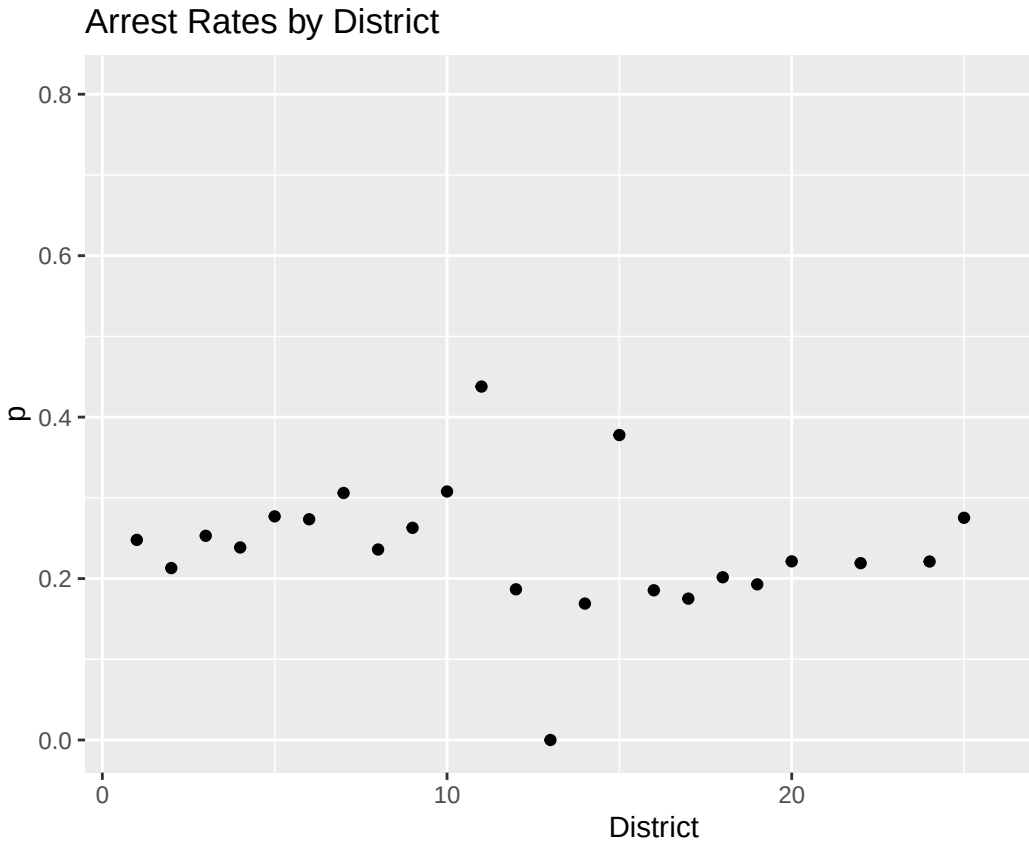
Chicago.

```
## Warning: Removed 1 rows containing missing values (position_stack).
```



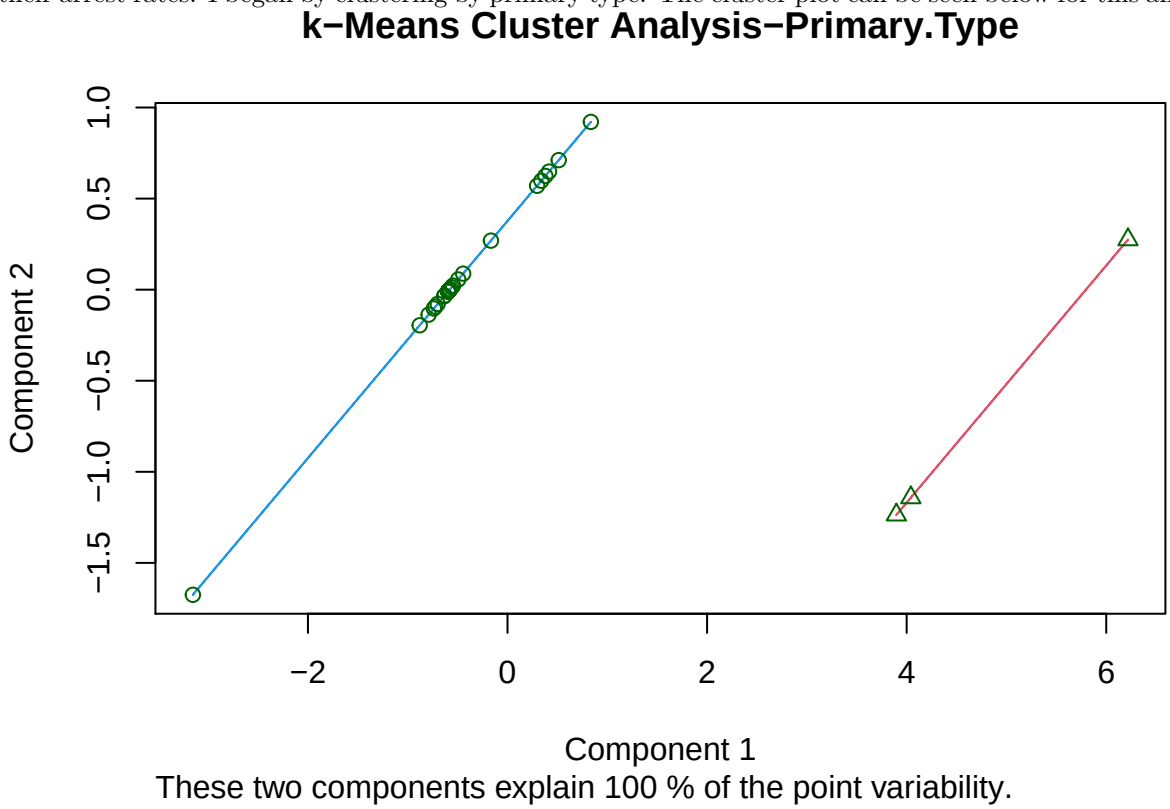
I then made plots that show the frequency of one variable given another variable. I found these plots difficult to read and not particularly helpful in my analysis so I am omitting them from this report.

Since I decided to focus on the arrest rates from the dataset, I plotted the relationship between district and primary tyoes and the arrest rates. The most useful relationship I found was that between the district and ar-



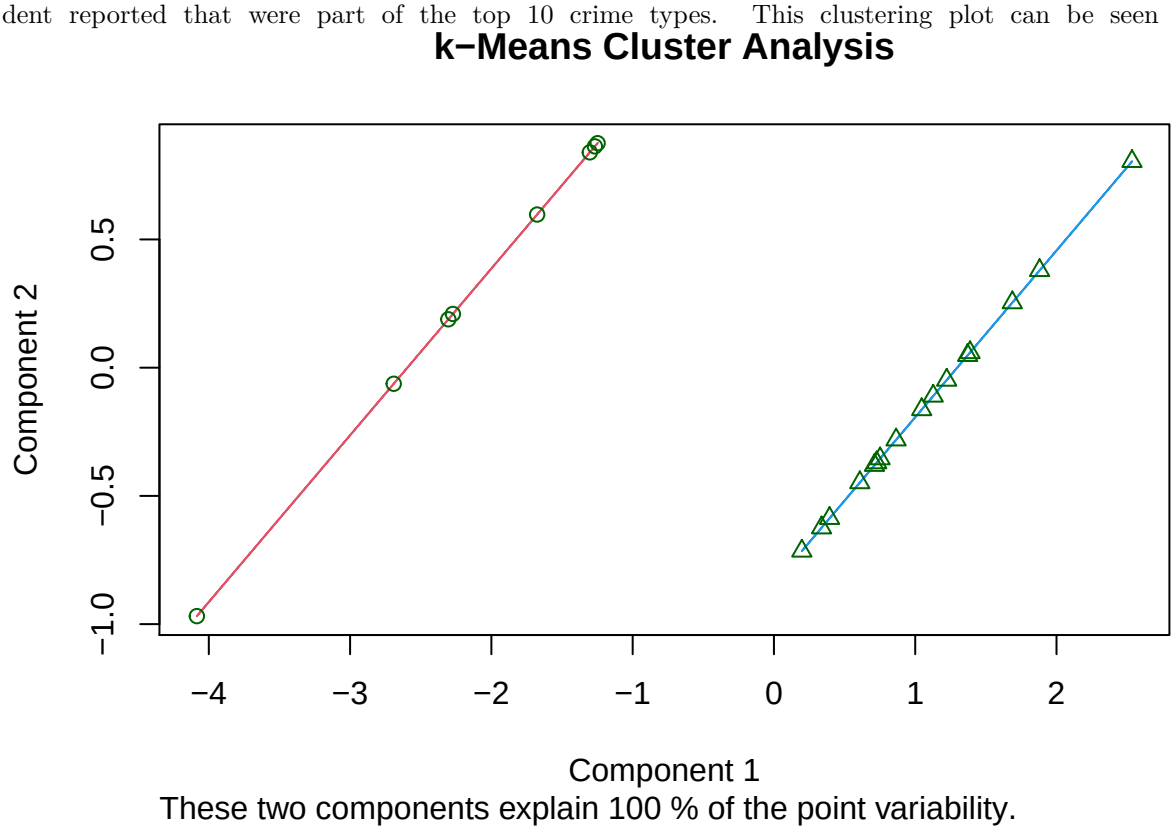
rest rates. This plot is shown below.

For my actual analysis, I wanted to use k-means clustering to find group districts and crime types together by their arrest rates. I began by clustering by primary type. The cluster plot can be seen below for this analysis:



Then, to simplify analysis I decided to focus on the top 10 crime types as these are the most common crimes that are reported in the city of Chicago and still has the potential to tell us a lot of useful information about crime in Chicago.

I then performed a k-means clustering analysis on the districts and arrest rates only for the incident reported that were part of the top 10 crime types. This clustering plot can be seen below:



For the second part of my analysis, I wanted to predict whether or not an arrest would be made for a crime reported in Chicago. To accomplish this, I performed a regression analysis. I split the dataset into train and test sets with a 80:20 ratio as this is a standard starting point and I assume it will capture most of the variance in my model. My final model used district, crime primary type, and location to predict if an arrest would be made for the associated crime. I decided to use these variables as I could see in the k-means clustering that the different districts, types, and locations could be grouped together in terms of their arrest rates.

###Results The k-means clustering algorithms, allow us to group together districts and crime types by the rate that arrests are made for these crimes. Through this analysis, I learned that battery, theft, and criminal damage are grouped together for low arrest rates. However, these are the three most common crime types so it makes sense that the rate of arrests is lower than the others because of the frequency of these crimes. My second k-means algorithm essentially groups together the districts by arrest rates for the top 10 crime types. This algorithm gives us a sense of the likelihood of an arrest being made for one of the top 10 crimes in the different districts of Chicago. It is more likely for an arrest to be made in districts 10, 11, 15, 17, 20, 21, 22, and 24. My regression model had an RMSE or 0.32 which was considerably lower than the RMSE of my first model using only the mean arrest rates. Each variable that I added into the model, lowered the RMSE which tells me that my model somewhat accurately predicts the probability of an arrest being made. The variables that I analyzed, do in fact impact the probability of an arrest being made for a crime reported in the city of Chicago.

###Conclusion Overall, I would say that I am happy with what I accomplished in this project. I think that the way I chose to analyze the dataset, taught me a lot about the data and I believe that you can see the progression of my analysis throughout my work. However, I think that there is a lot more that could be

done. For example, I was not confident using the k-means clustering algorithm and I think that more could be done with it. I think that more could be learned about the data by potentially looking at the time of the crime. This might also be a useful variable to include in the regression analysis to more accurately predict if an arrest will be made. Also, I would like to point out that even though the RMSE for my regression analysis is rather low, since I am only predicting "True" or "False," I would like this number to be even lower. Adding in an analysis of the time that the crime was committed may help in this.