Data Mining Project
Chicago Crime

By David Umberger

The purpose of this project is to shed light on the effects of crime and the likelihood of where crimes are committed in the city of Chicago, Illinois. My goal was to better police presence in the Chicago districts by allocating the right amount of resources to the needed areas.

The benefits of this project are numerous. There is certainly the moral reason for lowering what is wrong being the right thing to do. The economic effects of lowering crime in this city are better because this provides a more attractive environment for business so more companies will want to move there. When crime is down, neighborhoods and communities can better themselves and project a more livable image. This will in turn bring more people desiring to live in the city. Along a similar line is aiding in the likelihood of crime prevention will also bring in a higher class of people to the city, professionals for instance, so this can help the local economy recover from the COVID-19 lockdowns. This improved city environment will show a more positive vibe and yield a better outcome where people will want to get out more and enjoy the city. When we consider the effect that crimes of all types have on a community, the results can be sobering so effective crime-fighting is essential for the betterment of society. The backbone of crime-fighting is the police force but what can help is good law-abiding citizens and if these models can help citizens work with police better, that would be a win for the city.

Executive summary:

My objective was to locate and predict the top four most likely crimes in the city of Chicago to help make arrests more effective by associating data rules. I found that assault, battery, criminal damage, and theft are the top four crimes. I was able to break down the crimes into their district areas and appoint dependent variables and used them in modeling algorithms to drill down from there. The results were mixed and with some promise. This data is complicated and has much variety, however, some data attributes are fairly narrowed. My results did show some association between the data so there are some conclusions to be drawn from them. My sample sizes were around ten to thirty thousand to pull good training and testing sets. By breaking up the city's districts into north, central, and south, I was able to look more closely at the data relating to those crimes. My focus in this project was on location and to look at arrests. This shows the number of crimes in those areas. With an average model success rate of around 67%, I found that the models I used do need some work to be more effective.

Dataset:

I used the Crime data from the City of Chicago government page for their crime division. The main dataset contained 1,048,576 records. For my purposes of data that is trainable and testable, the main file was too large since it included all city crime data since 2001. I decided to keep the data relevant and trimmed it down to the years 2019 and 2020. I decided to include 2019 since I noticed that since COVID-19 shut down the city for much of 2020 and the crime rates were significantly lower. 2020 data had only 4,125 records. To put this in perspective, 2019 data comprised 58,736 records, which was more normal.

Case # - The Chicago Police Department RD Number (Records Division Number), which is unique to the incident.	DATE OF OCCURRENCE - Date when the incident occurred. This is sometimes the best estimate.	BLOCK - The partially redacted address where the incident occurred, placing it on the same block as the actual address.	IUCR - The Illinois Uniform Crime Reporting Code. This is directly linked to the Primary Type and Description
PRIMARY DESCRIPTION - The primary description of the IUCR code.	SECONDARY DESCRIPTION - The secondary description of the IUCR code, a subcategory of the primary description.	LOCATION DESCRIPTION - Description of the location where the incident occurred.	ARREST - Indicates whether an arrest was made.
BEAT - Indicates the beat where the incident occurred. A beat is the smallest police geographic area - each beat has a dedicated police beat car.	WARD - The ward (City Council district) where the incident occurred.	FBI CD - Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).	X COORDINATE - The x coordinate of the location where the incident occurred in State Plane Illinois East NAD 1983 projection.
Y COORDINATE - They coordinate the location where the incident occurred in State Plane Illinois East NAD 1983 projection.	LATITUDE - The latitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.	LONGITUDE - The longitude of the location where the incident occurred. This location is shifted from the actual location for partial redaction but falls on the same block.	LOCATION The location where the incident occurred in a format that allows for the creation of maps and other geographic operations on this data portal.

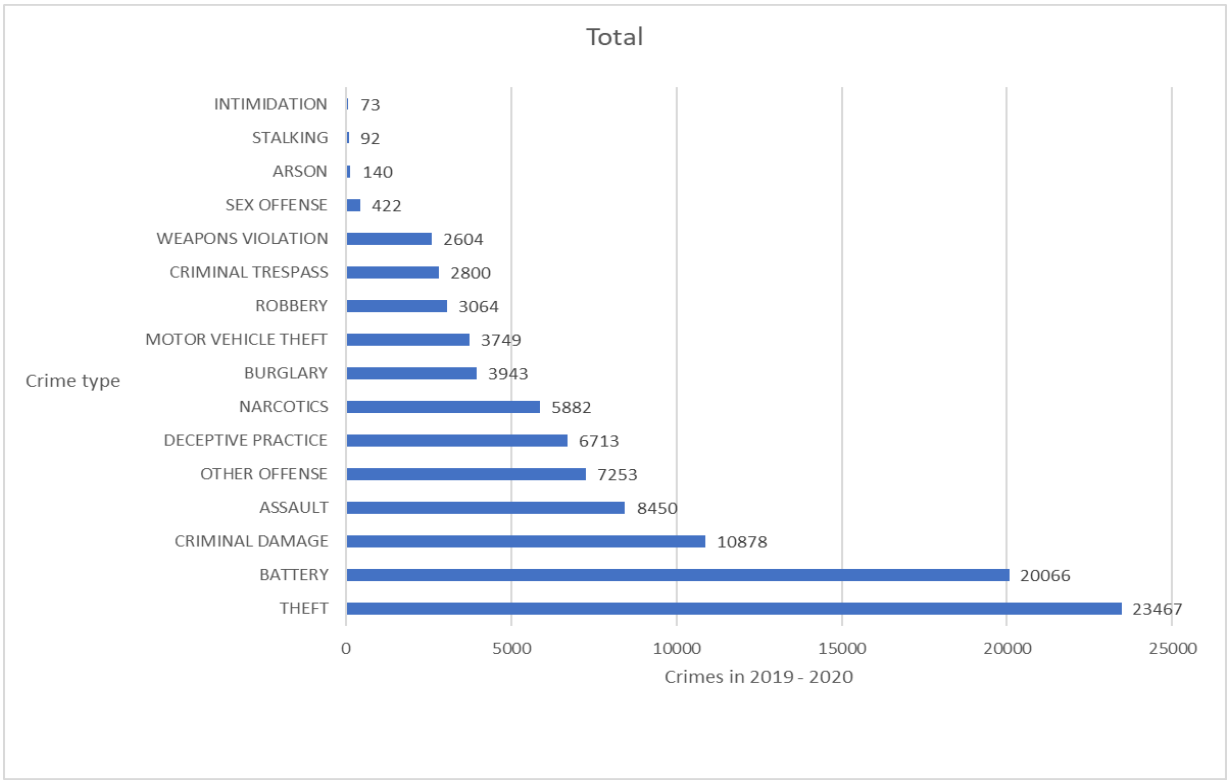
From here, I made the following adjustments to the dataset. I hoped to isolate where certain crimes would occur in the city. To narrow the data, I chose the top 4 most common crimes occurring in the city and decided to work with that data. In the dataset, I added a column to count and pull the descriptive stats by using “If or statements” VBA code in Excel cells. I also added a column at the end entitled “Arrested” and used some more VBA code to change the Arrest column to yes and no. The reason is that the “Arrest” column was a TRUE/FALSE logic attribute when R read the file. This was difficult to work with when I was running several K-means tests so I removed the attribute after I created the yes/no attribute and converted it to a factor. I also created another attribute entitled “District area” by coding the cells to look for the

district numbers I gave had it check and assign to north, central, or south. These actions made the data more workable. The district area datasets resulted in 20,887 records for central, 15,391 records for the south, and 26,583 records for the north. A list of district numbers is found in exhibit 5.

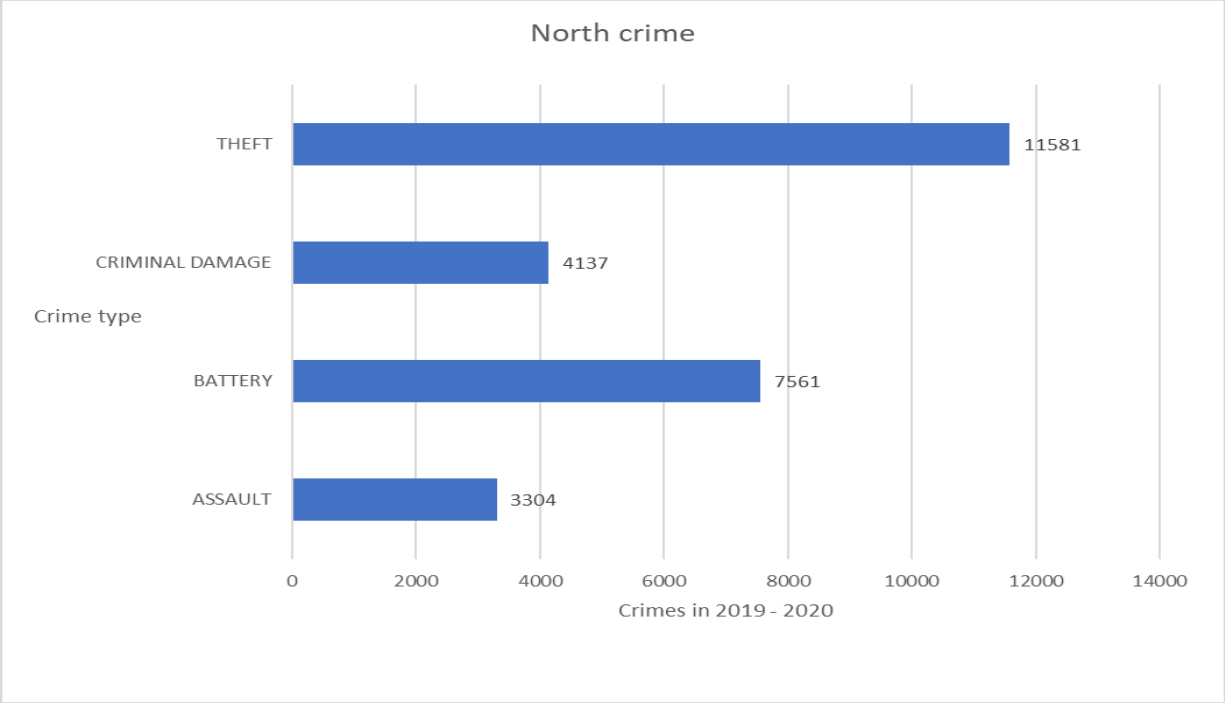
The dependent variables I used were “Arrested”, “Primary type” (of crime), and “District area.” To check for missing data, I ran a few R statements to check for complete rows and then save them to a completed variable name. This was necessary so that my data would have the least resistance in modeling. In analyzing the dataset, I found there were several ways to locate potential crimes by location and this involved some trial and error to find out which attributes worked best to predict the desired results.

Summary data statistics:

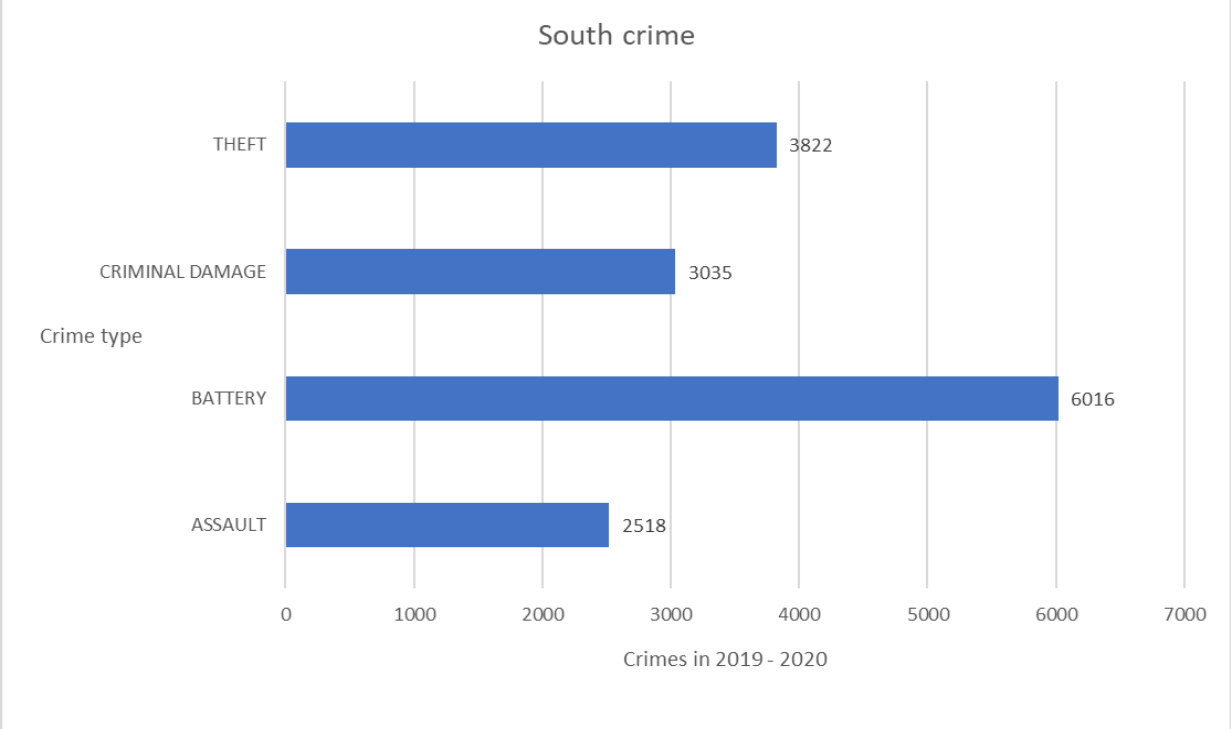
When we look at the data, a good picture involved using several important attributes that included the valuable results that we have. The summary statistics are found in exhibits 1 through 4 but in the context of this project, they were not very helpful since most of the attributes are categorical or discrete. A useful data view was found in pulling the frequencies of crimes from the data set. Looking at the crime type and how often it occurred during the years pulled gives us more of a picture of what the data represents. Below is a summary of total Chicago crimes which included the top 16 crimes.



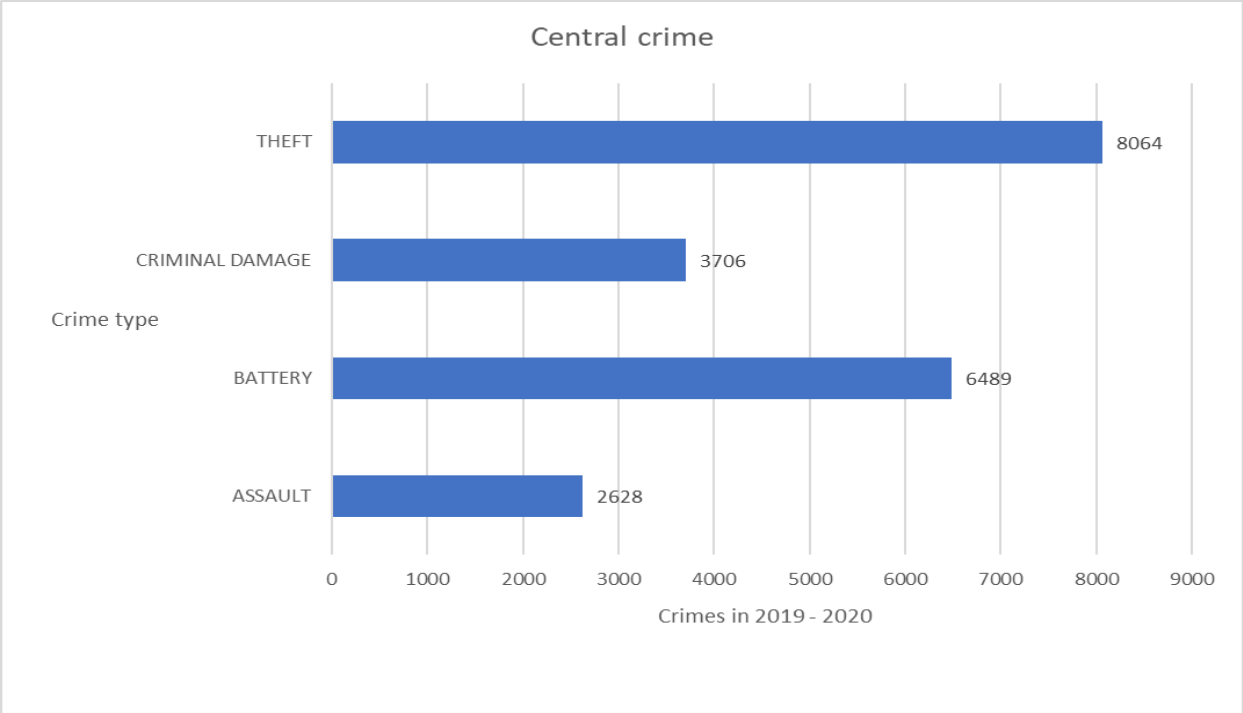
To further drill down, I looked at the top 4 crimes in the city divided into the three district areas. What is interesting to note is the variability of the data in the three areas. The North area shows a large amount of theft with lesser crimes to follow.



The south district displays a larger leaning of battery with a lesser emphasis on theft and the rest.



And with the central district leaning more towards theft but with a lighter count than the north districts. The north and central areas appear to have similar frequencies.



Descriptive modeling:

For the descriptive method, I initially chose to use K-means cluster analysis. I looked through the data to find what relationships I wanted to investigate. After looking through the data attributes to find relevant columns that may fit this model, I decided to test the relationship between “District” and “Beat” since the beat is the smallest police unit of around 8-9 officers in the communities. I ran a few outputs and was not satisfied with the results. There appeared to be a relationship between the two, but the output was not coming out as I had hoped since the data variables that I was using were discrete. I decided to use the Apriori algorithm instead and think this dataset fits better with this model. Some of the source code can be found in exhibit 8. What I found was that the dependent variables that I was attempting to use were better served under a categorical friendly model. Since this algorithm is unsupervised, I wanted to find out if there were relationships between the three district areas, crime types, and arrests in the city.

Predictive modeling:

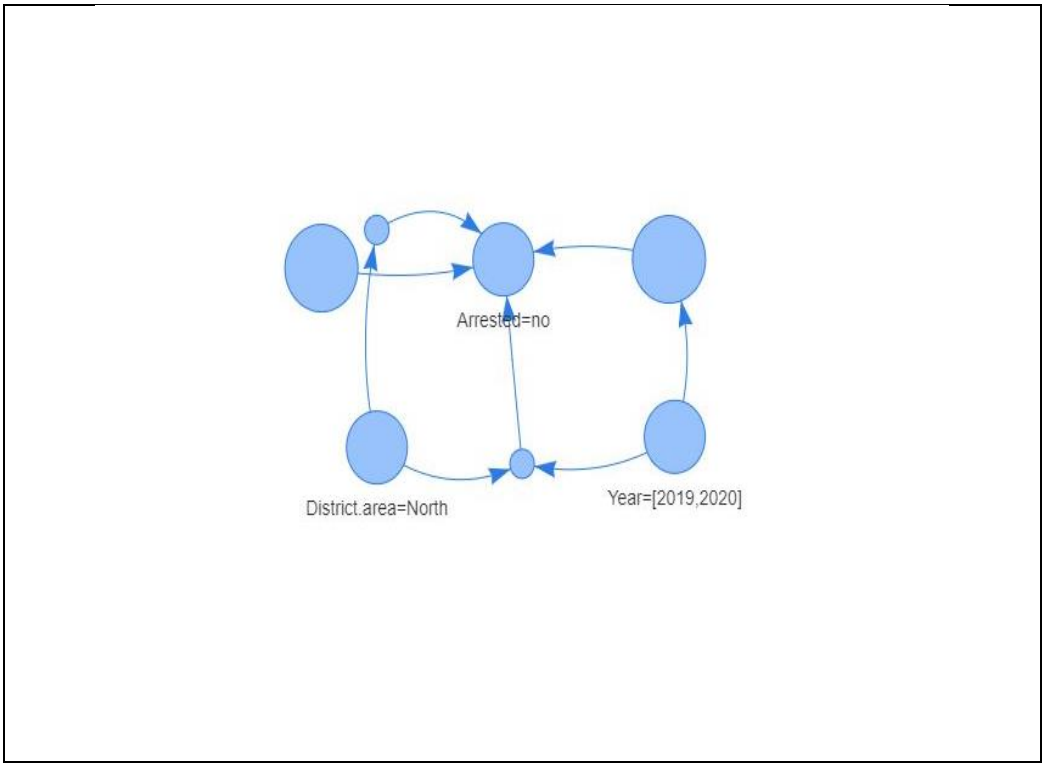
For the predictive methods, I chose to use the Naïve Bayes and decision tree methods. For the decision tree approach, I aimed to look at the number of arrests for the incidents logged to gauge how effective the police coverage in the city and district area is. I was able to make inferences in both of these models using the attribute “Arrested” since it is a yes/no question

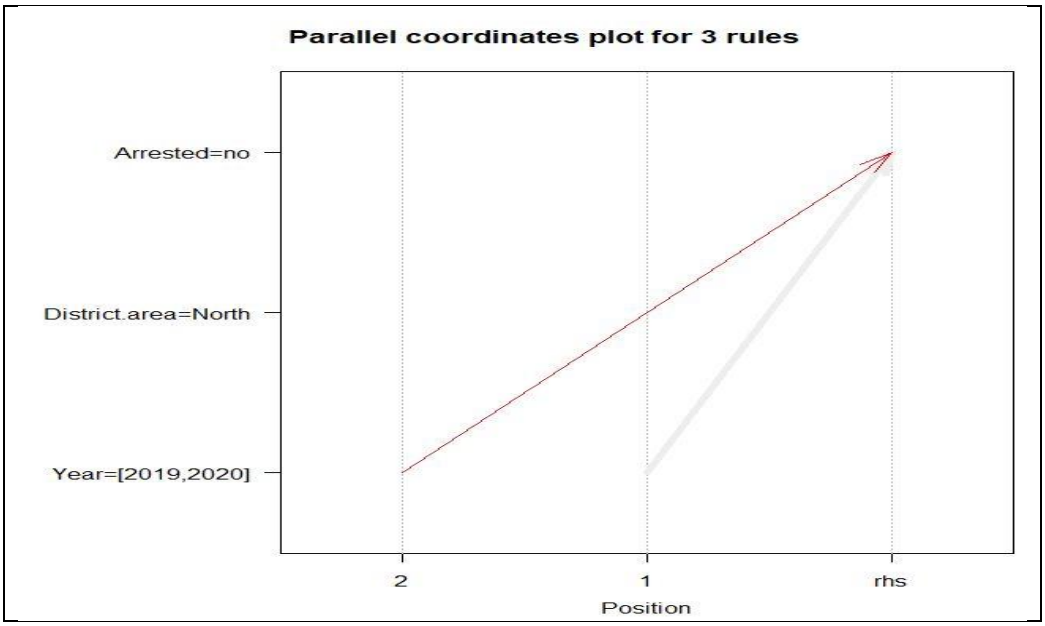
and used several predictors to produce the results. I broke this down by the city and the three-district areas. I think these again will paint a picture of what is happening directly in the communities of Chicago since we want arrests to go up so that crime rates go down. Source code segments can be found in exhibits 9 and 10.

The Naïve Bayes classifier depends more on probability, so this data worked well in our results. The decision tree approach is supervised learning and did take some tweaking with the amount of data used and which predictors to use but, in the end, I was able to pull some solid outcomes to infer results.

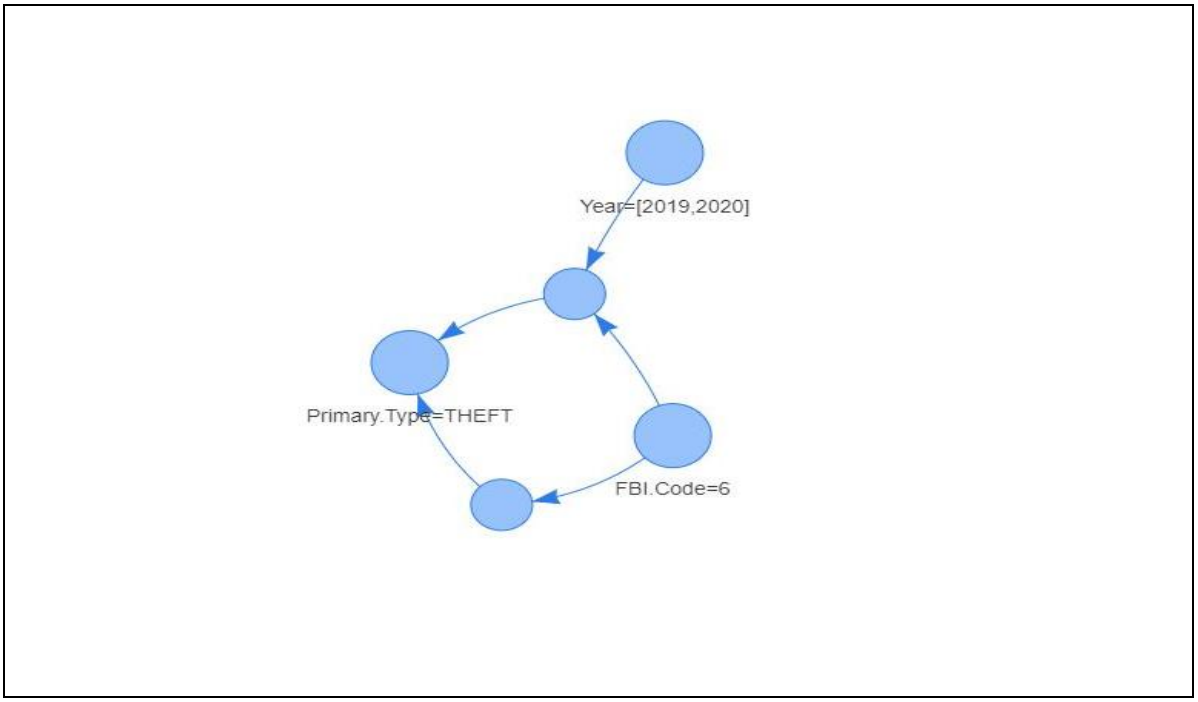
Apriori method results:

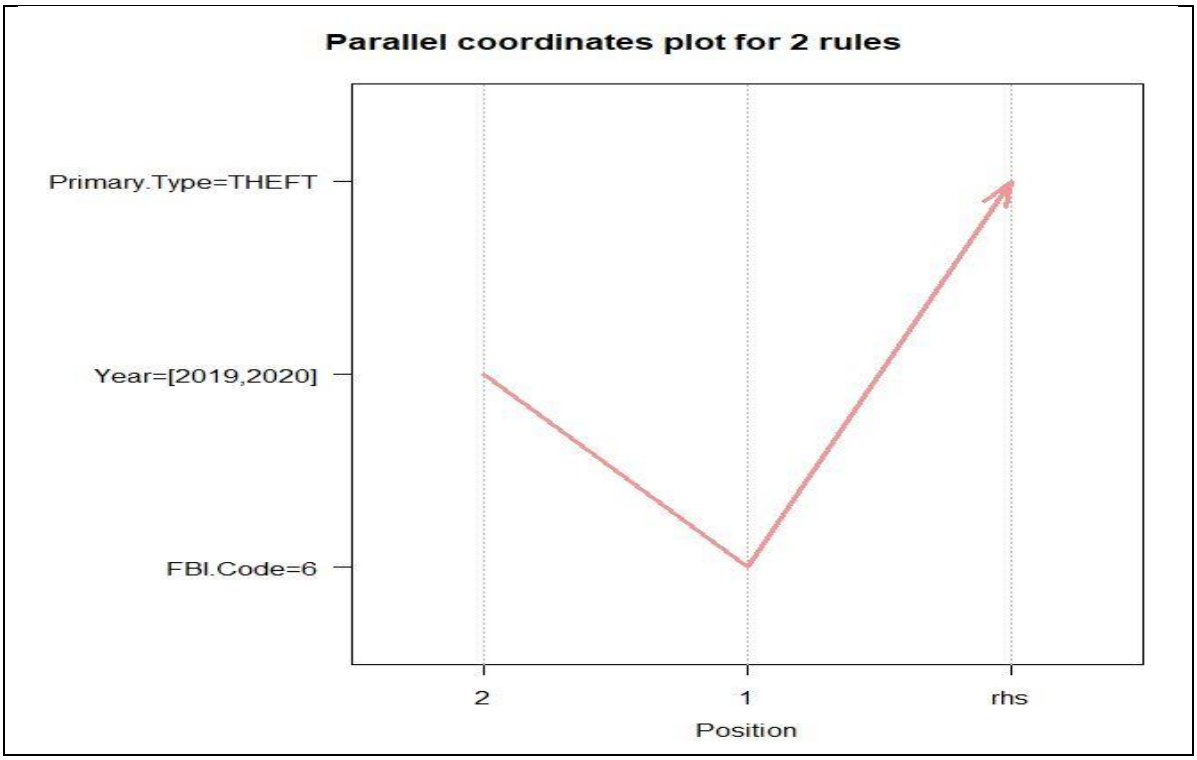
For this method, I pulled two plots – the network plot and the parallel plot. This first plot shows the test for arrested rules to see if there are relationships for arrests made. The network diagram shows the three relationships between the data. This pattern proves to be consistent throughout this data. The parallel plot shows a leaning toward no arrests with most of the data from the north district in the dataset.



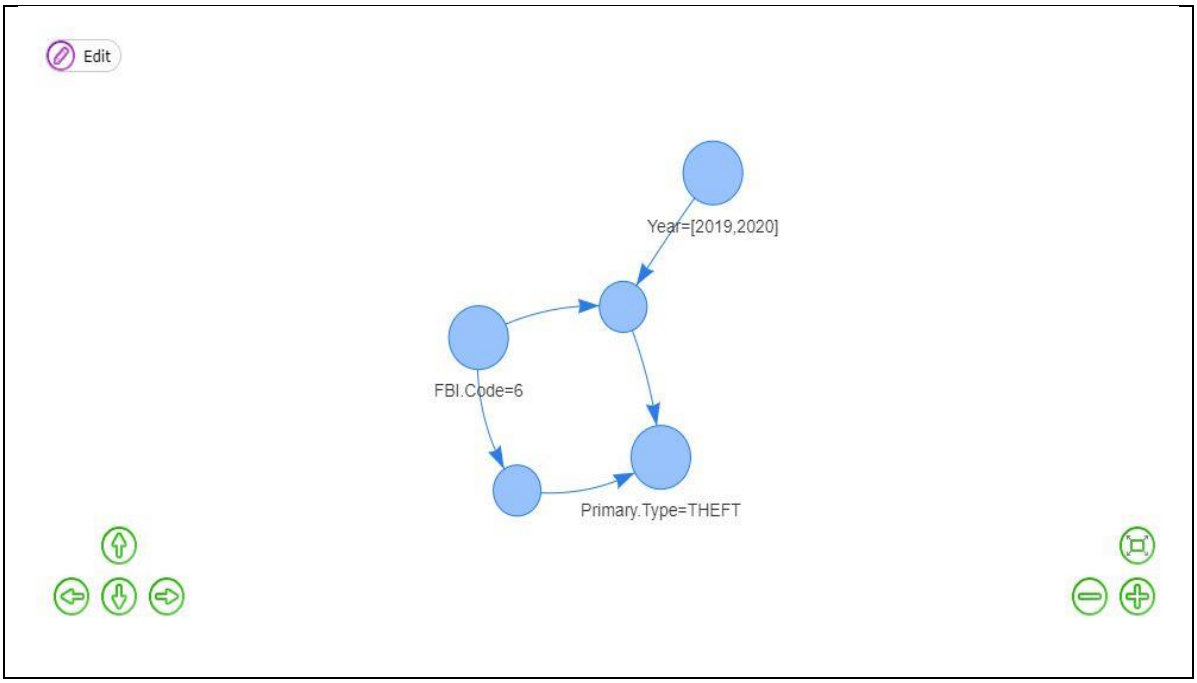


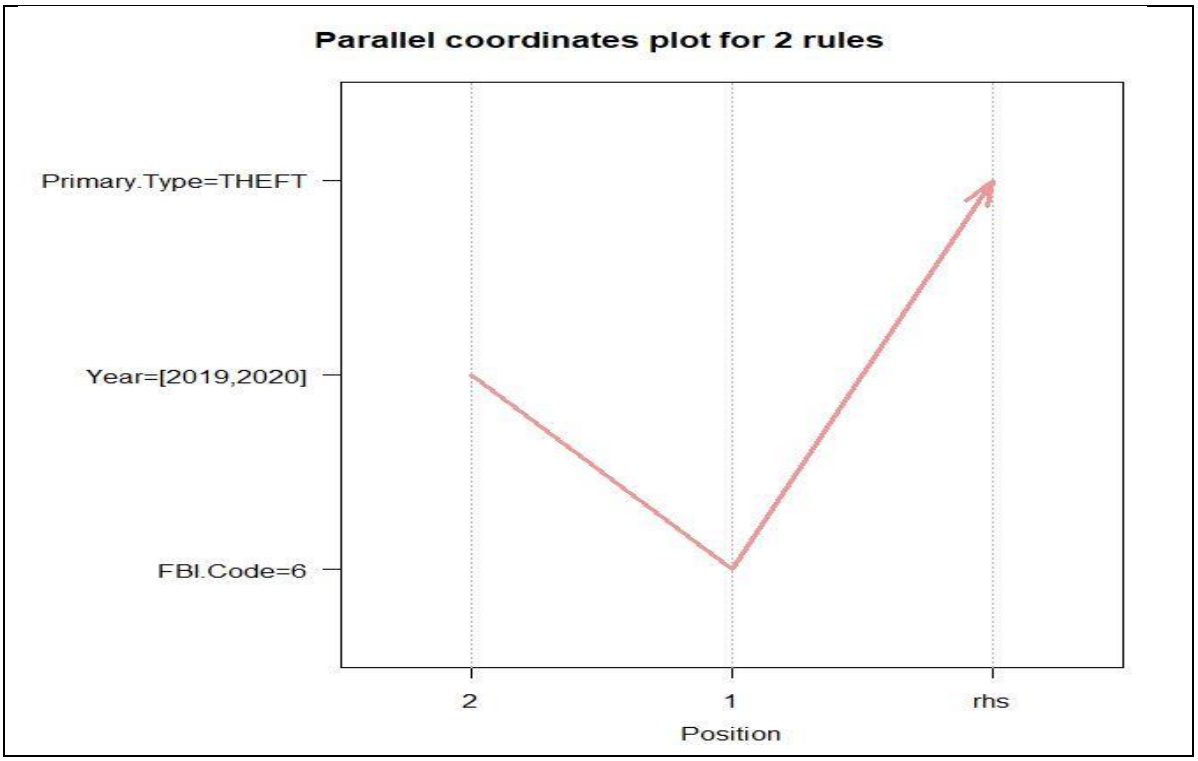
The type of crime was tested next to show a local identifier. These are of particular interest since the incidents are more centered on the specific district area, there will be more focus on the districts of these areas that have the most incidents. The data does favor the north since it has the majority of the records, however, the FBI code showed a particular relationship with theft in the dataset years.





The final test was on the district area. The relationship showed similar results to the type of crime test. There does appear to be a strong relationship between this FBI code and theft.





Naïve-Bayes method results:

A confusion matrix was generated for the total Chicago crime data and each district area for the Arrested cases.

The total Chicago crime data showed lower results for predictions

```
> table(arrest_results)
      predicted
actual    no   yes
   no  30524 15264
   yes  2242  5252
> nrow(arrest_results[arrest_results$predicted == arrest_results$actual, ]) /
+ nrow(arrest_results)
[1] 0.6714463
```

The north district area showed some slightly lower prediction results

```
> table(arrest_results)
      predicted
actual   no  yes
   no  7176 4233
   yes  441 1394
> nrow(arrest_results[arrest_results$predicted == arrest_results$actual, ]) /
+ nrow(arrest_results)
[1] 0.6470855
```

The central district area also was consistent

```
> table(arrest_results)
      predicted
actual   no  yes
   no  5950 3080
   yes  373 1006
> nrow(arrest_results[arrest_results$predicted == arrest_results$actual, ]) /
+ nrow(arrest_results)
[1] 0.6682678
```

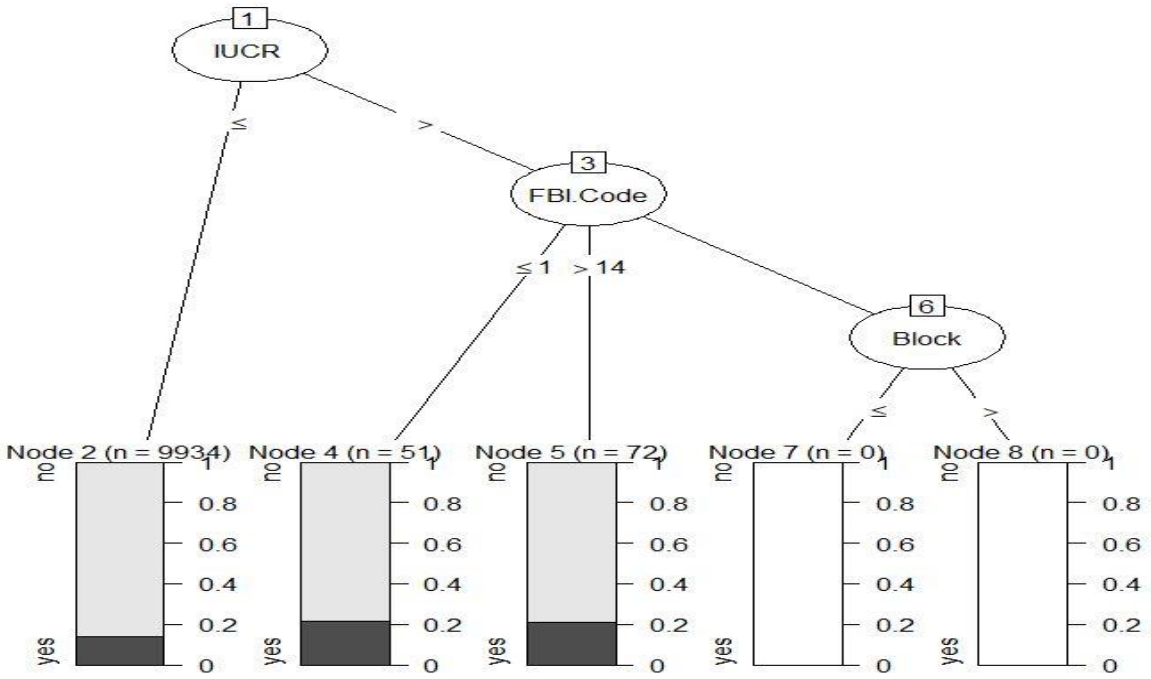
Along with the south district area

```
> table(arrest_results)
      predicted
actual   no  yes
   no  4011 2455
   yes  356  868
> nrow(arrest_results[arrest_results$predicted == arrest_results$actual, ]) /
+ nrow(arrest_results)
[1] 0.6344603
```

Decision tree method results:

A confusion matrix was generated for the total Chicago crime data for “Arrested” cases. The full decision tree output is found in exhibits 6 and 7. The tree plot also displays the results from the analysis. The IUCR and FBI codes again show a stronger prediction association.

```
Evaluation on training data (10057 cases):

           Decision Tree
         ----------------
         Size      Errors

            4 1124(11.2%)   <<


          (a)   (b)    <-classified as
         ----  ----
         8590    67    (a): class no
         1057   343    (b): class yes


        Attribute usage:

        100.00% IUCR
          8.35% FBI.Code
          7.76% Block
```



Model evaluation and recommendations:

The Apriori algorithm showed poor results with support being low around 37% through testing the three district areas to see if there were association rules but did show higher-end confidence, showing a leaning towards only specific data.

```
> inspect(rules[1:2])
    lhs                                rhs                  support   confidence coverage  lift   count
[1] {FBI.Code=6}                    => {Primary.Type=THEFT} 0.3719131 1          0.3719131 2.6888 23313
[2] {FBI.Code=6,Year=[2019,2020]} => {Primary.Type=THEFT} 0.3719131 1          0.3719131 2.6888 23313
> rules <- sort(result, decreasing = TRUE, by = "support")
> inspect(rules[1:2])
    lhs                                rhs                  support   confidence coverage  lift   count
[1] {FBI.Code=6}                    => {Primary.Type=THEFT} 0.3719131 1          0.3719131 2.6888 23313
[2] {FBI.Code=6,Year=[2019,2020]} => {Primary.Type=THEFT} 0.3719131 1          0.3719131 2.6888 23313
```

The arrested rules test is sorted by confidence and then support to see a stronger relationship with leanings towards the heavily populated north district with an arrest.

```
> inspect(rules[1:4])
    lhs                                      rhs             support   confidence coverage lift     count
[1] {District.area=North}                 => {Arrested=no} 0.3636813 0.8606864  0.422548 1.001509 22797
[2] {District.area=North,Year=[2019,2020]} => {Arrested=no} 0.3636813 0.8606864  0.422548 1.001509 22797
[3] {}                                    => {Arrested=no} 0.8593900 0.8593900  1.000000 1.000000 53870
[4] {Year=[2019,2020]}                    => {Arrested=no} 0.8593900 0.8593900  1.000000 1.000000 53870
> rules <- sort(result, decreasing = TRUE, by = "support")
> inspect(rules[1:4])
    lhs                                      rhs             support   confidence coverage lift     count
[1] {}                                    => {Arrested=no} 0.8593900 0.8593900  1.000000 1.000000 53870
[2] {Year=[2019,2020]}                    => {Arrested=no} 0.8593900 0.8593900  1.000000 1.000000 53870
[3] {District.area=North}                 => {Arrested=no} 0.3636813 0.8606864  0.422548 1.001509 22797
[4] {District.area=North,Year=[2019,2020]} => {Arrested=no} 0.3636813 0.8606864  0.422548 1.001509 22797
```

Lastly, looking at the type of crime association rules, they pulled similar results to the district area rules. These two only had two rules to report.

```
> inspect(rules[1:2])
    lhs                                rhs                   support   confidence coverage  lift   count
[1] {FBI.Code=6}                    => {Primary.Type=THEFT} 0.3719131 1          0.3719131 2.6888 23313
[2] {FBI.Code=6,Year=[2019,2020]} => {Primary.Type=THEFT} 0.3719131 1          0.3719131 2.6888 23313
> rules <- sort(result, decreasing = TRUE, by = "confidence")
> inspect(rules[1:2])
    lhs                                rhs                   support   confidence coverage  lift   count
[1] {FBI.Code=6}                    => {Primary.Type=THEFT} 0.3719131 1          0.3719131 2.6888 23313
[2] {FBI.Code=6,Year=[2019,2020]} => {Primary.Type=THEFT} 0.3719131 1          0.3719131 2.6888 23313
```

The Naïve-Bayes tests pulled the following performance results in predicting arrests. They were not as high as I would have hoped but considering the size and variety of the data, this is possible.

Naïve-Bayes performance	Central	South	North	Total
Accuracy	0.67	0.63	0.65	0.67
Error rate	0.33	0.37	0.35	0.33
True positive rate (Sensitivity)	0.73	0.71	0.76	0.70
True negative rate (Specificity)	0.66	0.62	0.63	0.67
False positive rate	0.34	0.38	0.37	0.33
False negative rate	0.27	0.29	0.24	0.30
Precision	0.25	0.26	0.25	0.26
Recall	0.73	0.71	0.76	0.70
F-measure	0.37	0.38	0.37	0.38

The decision tree model pulled better performance than the Naïve-Bayes. I look at the error rate being only 11% compared to the 33 and 34%. The false-positive rate is much lower showing the decision tree almost unwilling to make a positive arrest identifier.

Decision tree performance	Total
Accuracy	0.89
Error rate	0.11
True positive rate (Sensitivity)	0.25
True negative rate (Specificity)	0.99
False-positive rate	0.01
False-negative rate	0.76
Precision	0.84
Recall	0.25
F-measure	0.38

My recommendation for a better model would be to possibly try more combinations and find out why the data is not pulling better rules or showing better results. With that said, some of the challenges could be the data itself in that it may have some structural challenges. This dataset does have much variety with some attributes as unique identifiers and other attributes having only several levels. However, with that said, perhaps with enough code manipulation, we can have a more accurate prediction of crime in Chicago.

Implications and conclusion:

The Apriori algorithm is a seeker of patterns between attributes showing this data with a large amount of leaning towards the majority since I do have consistent results through two of the tests. The indication here is there are not enough similarities between the attributes to pull more rules. I was only able to capture four rules from the arrested test and two rules from the district area and a crime type test. This could also indicate that some areas do not have many incidents or arrests. Nonetheless, these tests cover all the districts and beats reported in the data. When I ran the arrested test on all three district areas, none of them pulled a single association rule. This again could indicate that the data is not showing enough similarity and when we consider those transaction ID examples from class, there must have been numerous records that did not have enough frequency to survive after pruning the data to the next itemset.

The Naïve-Bayes tests were lower than anticipated. This could be due to several reasons. Due to the size of the data from the total dataset at 62,684 records, 15% of this data was used for training coming out to 9,403 records. For the three district areas, 50% of the data was used for training, still coming in at 13,292, 7,696, and 10,444 records, respectively, there was still enough data to train well. This could be a result of the data in that it may be difficult to predict an arrest. Even if all the police work goes well and all procedures are followed, the resulting arrest may not happen. There could be reasons beyond this data that make arrests more difficult than we can see.

The decision tree tests show that for predicting arrests, the IUCR codes are the best predictor by far, followed by the FBI code and Block. Due to this very large dataset, I used a random sample of 16% of the dataset, amounting to 10,029 records. What is interesting is that the police beats are nowhere to be found. Perhaps since they are the more on the police level and used for police purposes? This could be the result that these codes are assigned to certain people upon arrest and certain FBI codes are assigned upon arrest. Certain blocks in the city may be more prone to arrests?

Upon completing this project, I attest that this data had to be wrestled with much more than I anticipated. Sometimes, it may be more about finding the data to fit the model. Making predictions on arrests and where crimes are located may be much about the outside factors that are involved than the data can lead to. We can only show what the data provides and draw inferences from those results. The predictors in the decision tree show that they may come

from an area that we do not expect. When it comes to predicting the likelihood of where crime may occur, we need to be open to any avenue since this may involve the unexpected.

Appendices:

Exhibit 1 – Chicago total data summary

```
> summary(myfile)
       ID            Case.Number            Date              Block
 Min.   :11552577   Length:62861       Length:62861       Length:62861
 1st Qu.:11596072   Class :character   Class :character   Class :character
 Median :11637368   Mode  :character   Mode  :character   Mode  :character
 Mean   :11672534
 3rd Qu.:11677604
 Max.   :12139735

     IUCR           Primary.Type       Description        Location.Description
 Length:62861       Length:62861       Length:62861       Length:62861
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character


      Beat          District     District.area          Ward
 Min.   : 111   Min.   : 1.00   Length:62861       Min.   : 1.0
 1st Qu.: 532   1st Qu.: 5.00   Class :character   1st Qu.:10.0
 Median :1022   Median :10.00   Mode  :character   Median :24.0
 Mean   :1134   Mean   :11.11                      Mean   :23.4
 3rd Qu.:1724   3rd Qu.:17.00                      3rd Qu.:35.0
 Max.   :2535   Max.   :25.00                      Max.   :50.0
                                                   NA's   :1
 Community.Area    FBI.Code          X.Coordinate      Y.Coordinate
 Min.   : 1.00   Length:62861       Min.   :1098012   Min.   :1813914
 1st Qu.:22.00   Class :character   1st Qu.:1154345   1st Qu.:1859275
 Median :32.00   Mode  :character   Median :1167791   Median :1894178
 Mean   :36.43                      Mean   :1165680   Mean   :1886632
 3rd Qu.:53.00                      3rd Qu.:1176618   3rd Qu.:1909121
 Max.   :77.00                      Max.   :1205116   Max.   :1951507
                                    NA's   :176       NA's   :176
      Year       Updated.On           Latitude       Longitude
 Min.   :2019   Length:62861       Min.   :41.64   Min.   :-87.92
 1st Qu.:2019   Class :character   1st Qu.:41.77   1st Qu.:-87.71
 Median :2019   Mode  :character   Median :41.87   Median :-87.66
 Mean   :2019                      Mean   :41.84   Mean   :-87.67
 3rd Qu.:2019                      3rd Qu.:41.91   3rd Qu.:-87.63
 Max.   :2020                      Max.   :42.02   Max.   :-87.52
                                   NA's   :176     NA's   :176
   Location           Arrested
 Length:62861       Length:62861
 Class :character   Class :character
 Mode  :character   Mode  :character
```

Exhibit 2 – Chicago north data summary

```
> summary(myfile)
       ID            Case.Number            Date              Block              IUCR          
 Min.   :11552596   Length:26583       Length:26583       Length:26583       Length:26583      
 1st Qu.:11594610   Class :character   Class :character   Class :character   Class :character  
 Median :11635440   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :11667835                                                                             
 3rd Qu.:11676158                                                                             
 Max.   :12139735                                                                             

 Primary.Type       Description        Location.Description      Beat           District     
 Length:26583       Length:26583       Length:26583         Min.   : 111.0   Min.   : 1.000  
 Class :character   Class :character   Class :character     1st Qu.: 234.0   1st Qu.: 2.000  
 Mode  :character   Mode  :character   Mode  :character     Median : 834.0   Median : 8.000  
                                                            Mean   : 825.8   Mean   : 8.032  
                                                            3rd Qu.:1214.0   3rd Qu.:12.000  
                                                            Max.   :1834.0   Max.   :18.000  

 District.area          Ward        Community.Area   FBI.Code          X.Coordinate     
 Length:26583       Min.   : 1.00   Min.   : 7     Length:26583       Min.   :1129262  
 Class :character   1st Qu.: 6.00   1st Qu.:28     Class :character   1st Qu.:1160106  
 Mode  :character   Median :20.00   Median :32     Mode  :character   Median :1173068  
                    Mean   :19.95   Mean   :37                        Mean   :1169462  
                    3rd Qu.:27.00   3rd Qu.:56                        3rd Qu.:1177350  
                    Max.   :43.00   Max.   :70                        Max.   :1195809  
                                                                      NA's   :96       
 Y.Coordinate          Year       Updated.On          Latitude       Longitude      
 Min.   :1846427   Min.   :2019   Length:26583       Min.   :41.73   Min.   :-87.80  
 1st Qu.:1868497   1st Qu.:2019   Class :character   1st Qu.:41.79   1st Qu.:-87.69  
 Median :1888349   Median :2019   Mode  :character   Median :41.85   Median :-87.64  
 Mean   :1884769   Mean   :2019                      Mean   :41.84   Mean   :-87.65  
 3rd Qu.:1900927   3rd Qu.:2019                      3rd Qu.:41.88   3rd Qu.:-87.62  
 Max.   :1916265   Max.   :2020                      Max.   :41.93   Max.   :-87.56  
 NA's   :96                                          NA's   :96      NA's   :96      
   Location           Arrested        
 Length:26583       Length:26583       
 Class :character   Class :character   
 Mode  :character   Mode  :character   
```

Exhibit 3 – Chicago central data summary

```
> summary(myfile)
       ID              Case.Number           Date              Block              IUCR          
 Min.   :11552577   Length:20887       Length:20887       Length:20887       Length:20887      
 1st Qu.:11596404   Class :character   Class :character   Class :character   Class :character  
 Median :11637779   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :11674767                                                                              
 3rd Qu.:11677880                                                                              
 Max.   :12138662                                                                              

 Primary.Type       Description        Location.Description      Beat          District     
 Length:20887       Length:20887       Length:20887         Min.   :1111   Min.   :11.00  
 Class :character   Class :character   Class :character     1st Qu.:1433   1st Qu.:14.00  
 Mode  :character   Mode  :character   Mode  :character     Median :1712   Median :17.00  
                                                            Mean   :1788   Mean   :17.64  
                                                            3rd Qu.:2028   3rd Qu.:20.00  
                                                            Max.   :2535   Max.   :25.00  

 District.area          Ward        Community.Area    FBI.Code          X.Coordinate     
 Length:20887       Min.   : 1.00   Min.   : 1.00   Length:20887       Min.   :1098012  
 Class :character   1st Qu.:28.00   1st Qu.: 7.00   Class :character   1st Qu.:1142647  
 Mode  :character   Median :36.00   Median :21.00   Mode  :character   Median :1152030  
                    Mean   :34.47   Mean   :19.81                      Mean   :1151590  
                    3rd Qu.:44.00   3rd Qu.:25.00                      3rd Qu.:1161944  
                    Max.   :50.00   Max.   :77.00                      Max.   :1174588  
                    NA's   :1                                          NA's   :68       
  Y.Coordinate          Year       Updated.On           Latitude       Longitude      
 Min.   :1894064   Min.   :2019   Length:20887       Min.   :41.87   Min.   :-87.92  
 1st Qu.:1905870   1st Qu.:2019   Class :character   1st Qu.:41.90   1st Qu.:-87.75  
 Median :1917765   Median :2019   Mode  :character   Median :41.93   Median :-87.72  
 Mean   :1918405   Mean   :2019                      Mean   :41.93   Mean   :-87.72  
 3rd Qu.:1929861   3rd Qu.:2019                      3rd Qu.:41.96   3rd Qu.:-87.68  
 Max.   :1951507   Max.   :2020                      Max.   :42.02   Max.   :-87.63  
 NA's   :68                                          NA's   :68      NA's   :68      
   Location           Arrested        
 Length:20887       Length:20887       
 Class :character   Class :character   
 Mode  :character   Mode  :character   
```

Exhibit 4 – Chicago south data summary

```
> summary(myfile)
       ID              Case.Number           Date              Block              IUCR          
 Min.   :11552587   Length:15391       Length:15391       Length:15391       Length:15391      
 1st Qu.:11598230   Class :character   Class :character   Class :character   Class :character  
 Median :11640109   Mode  :character   Mode  :character   Mode  :character   Mode  :character  
 Mean   :11677618                                                                              
 3rd Qu.:11679917                                                                              
 Max.   :12138151                                                                              

 Primary.Type       Description        Location.Description      Beat           District     
 Length:15391       Length:15391       Length:15391         Min.   : 411.0   Min.   : 4.000  
 Class :character   Class :character   Class :character     1st Qu.: 511.0   1st Qu.: 5.000  
 Mode  :character   Mode  :character   Mode  :character     Median : 622.0   Median : 6.000  
                                                            Mean   : 779.7   Mean   : 7.574  
                                                            3rd Qu.: 723.0   3rd Qu.: 7.000  
                                                            Max.   :2234.0   Max.   :22.000  

 District.area          Ward       Community.Area    FBI.Code          X.Coordinate     
 Length:15391       Min.   : 3.00   Min.   :43.00   Length:15391       Min.   :1148940  
 Class :character   1st Qu.: 8.00   1st Qu.:46.00   Class :character   1st Qu.:1170139  
 Mode  :character   Median :10.00   Median :53.00   Mode  :character   Median :1176431  
                    Mean   :14.36   Mean   :58.01                      Mean   :1178243  
                    3rd Qu.:20.00   3rd Qu.:69.00                      3rd Qu.:1184006  
                    Max.   :34.00   Max.   :75.00                      Max.   :1205116  
                                                                       NA's   :12       
  Y.Coordinate          Year       Updated.On          Latitude       Longitude      
 Min.   :1813914   Min.   :2019   Length:15391       Min.   :41.64   Min.   :-87.73  
 1st Qu.:1838582   1st Qu.:2019   Class :character   1st Qu.:41.71   1st Qu.:-87.65  
 Median :1849610   Median :2019   Mode  :character   Median :41.74   Median :-87.63  
 Mean   :1846830   Mean   :2019                      Mean   :41.73   Mean   :-87.62  
 3rd Qu.:1854767   3rd Qu.:2019                      3rd Qu.:41.76   3rd Qu.:-87.60  
 Max.   :1868373   Max.   :2020                      Max.   :41.79   Max.   :-87.52  
 NA's   :12                                          NA's   :12      NA's   :12      
   Location           Arrested        
 Length:15391       Length:15391       
 Class :character   Class :character   
 Mode  :character   Mode  :character   
```

Exhibit 5 – Chicago police districts and beats

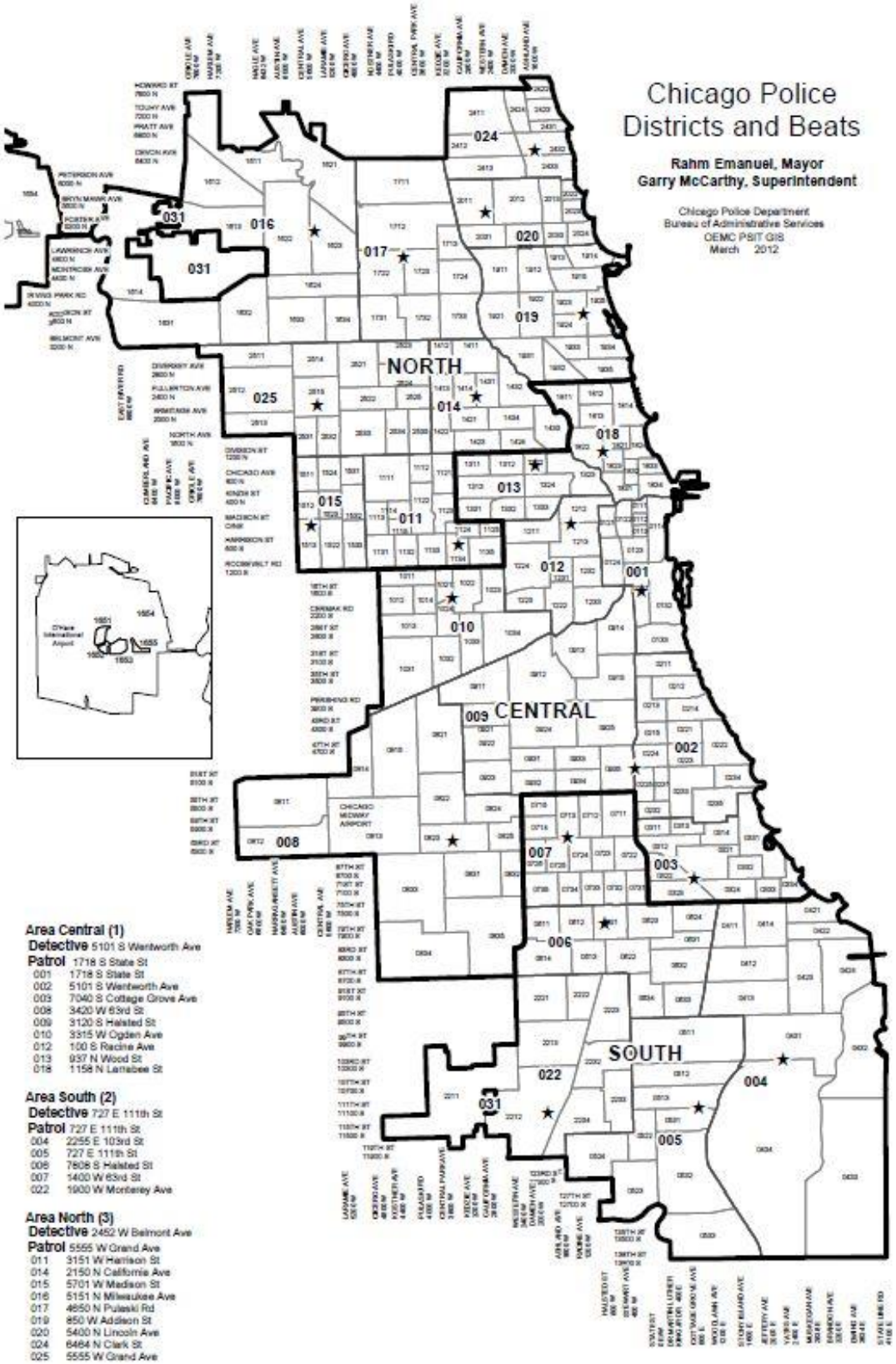


Exhibit 6 – Decision tree output 1

```
Call:
C5.0.default(x = train[, predictors], y = train$Arrested)


C5.0 [Release 2.07 GPL Edition]         Wed Mar 03 02:39:59 2021
-------------------------------

Class specified by attribute `outcome'

Read 10057 cases (12 attributes) from undefined.data

Decision tree:

IUCR in {820,890,545,560,486,1310,496,810,460,430,1320,497,870,051A,484,495,
:        1345,520,530,041A,552,880,420,1305,850,440,485,555,1340,479,483,865,
:        482,895,553,487,475}: no (9217/994)
IUCR in {860,454,498,558,051B,453,557,550,554,461,462}:
:...FBI.Code = 14: no (0)
    FBI.Code in {08A,08B}: yes (60/7)
    FBI.Code in {6,04B,04A}:
    :...Block in {100XX S CALUMET AVE,002XX W WASHINGTON ST,078XX S UNION AVE,
        :         076XX S CORNELL AVE,053XX S WESTERN AVE,034XX W 55TH ST,
        :         022XX S DR MARTIN LUTHER KING JR DR,017XX N OAK PARK AVE,
        :         062XX W GRACE ST,051XX S AUSTIN AVE,002XX N HOYNE AVE,
        :         026XX W MONROE ST,082XX S MARYLAND AVE,048XX S MICHIGAN AVE,
        :         111XX S VERNON AVE,019XX W 21ST PL,061XX S EBERHART AVE,
        :         013XX W ROSCOE ST,031XX N SHEFFIELD AVE,083XX S SANGAMON ST,
        :         004XX S WESTERN AVE,059XX S KOMENSKY AVE,032XX W WARREN BLVD,
        :         017XX N SHEFFIELD AVE,0000X W RANDOLPH ST,010XX E 100TH PL,
        :         010XX S MONITOR AVE,039XX S ELLIS AVE,037XX S DEARBORN ST,
        :         044XX S ST LOUIS AVE,039XX N CUMBERLAND AVE,
        :         006XX W HARRISON ST,060XX N CLAREMONT AVE,
        :         008XX N WASHTENAW AVE,020XX W 71ST ST,068XX S MARSHFIELD AVE,
        :         035XX W NORTH AVE,065XX S ST LAWRENCE AVE,
        :         016XX N LOCKWOOD AVE,074XX N OAKLEY AVE,006XX W JACKSON BLVD,
        :         002XX S STATE ST,012XX W 115TH ST,100XX S COMMERCIAL AVE,
        :         131XX S LANGLEY AVE,079XX S INGLESIDE AVE,052XX S PULASKI RD,
        :         083XX S JEFFERY BLVD,014XX N HUDSON AVE,013XX W 15TH ST,
        :         001XX W Madison St,022XX N LINCOLN AVE,030XX N GREENVIEW AVE,
        :         059XX S JUSTINE ST,056XX W LE MOYNE ST,013XX N ASHLAND AVE,
        :         120XX S LA SALLE ST,122XX S PEORIA ST,016XX N LINDER AVE,
        :         055XX W MONROE ST,071XX S HALSTED ST,055XX W WASHINGTON BLVD,
```

Exhibit 7 – Decision tree output 2

```
                    012XX S INDEPENDENCE BLVD,006XX N ST CLAIR ST,
                    053XX N MILWAUKEE AVE,032XX W ADDISON ST,061XX S ARCHER AVE,
                    016XX N PULASKI RD,082XX S HALSTED ST,090XX S COMMERCIAL AVE,
                    001XX W DIVISION ST,026XX N ELSTON AVE,123XX S WALLACE ST,
                    032XX N BROADWAY,018XX W FULLERTON AVE,089XX S ASHLAND AVE,
                    052XX N BROADWAY,046XX S HALSTED ST,016XX N AUSTIN AVE,
                    065XX W DIVERSEY AVE,032XX N CLARK ST,030XX W 26TH ST,
                    0000X E GARFIELD BLVD,036XX S INDIANA AVE,015XX E 55TH ST,
                    081XX S GREEN ST,069XX S JEFFERY BLVD,008XX S WESTERN AVE,
                    003XX S CICERO AVE,093XX S ESCANABA AVE,012XX N DEARBORN ST,
                    011XX S MASON AVE,008XX W 87TH ST,103XX S TORRENCE AVE,
                    094XX S JUSTINE ST,009XX W 115TH ST,084XX S STEWART AVE,
                    047XX S HALSTED ST,072XX S EAST END AVE,050XX W MONTROSE AVE,
                    001XX N AUSTIN BLVD,025XX S HAMLIN AVE,029XX N ASHLAND AVE,
                    047XX S WESTERN AVE,016XX W BELMONT AVE,072XX S ROCKWELL ST,
                    019XX W MONTEREY AVE,055XX N CLARK ST,061XX N WESTERN AVE,
                    045XX N BROADWAY,064XX W IRVING PARK RD,035XX N ELSTON AVE,
                    022XX S SACRAMENTO AVE,036XX S WASHTENAW AVE,
                    059XX S MAPLEWOOD AVE,029XX W ADDISON ST,046XX W BELMONT AVE,
                    013XX S CANAL ST,025XX N NARRAGANSETT AVE,
                    050XX S CALIFORNIA AVE}: yes (350/60)


Evaluation on training data (10057 cases):

           Decision Tree
         ----------------
         Size      Errors

            4 1124(11.2%)   <<


          (a)   (b)    <-classified as
         ----  ----
         8590    67    (a): class no
         1057   343    (b): class yes


        Attribute usage:

        100.00% IUCR
          8.35% FBI.Code
          7.76% Block
```

Exhibit 8 – Apriori source code

```
crime <- read.csv("Chicago Working Data.csv")

nrow(crime[complete.cases(crime),])

workdata = crime[complete.cases(crime),]

## Used three different Apriori functions to pull the rules
associated with these intended variables (Arrested, Crime type
and District area)

result <- apriori(workdata, parameter = list(sup = 0.35, conf =
0.8, target = "rules"), appearance = list(default = "lhs", rhs =
c('Arrested=yes','Arrested=no')))

result <- apriori(workdata, parameter = list(sup = 0.35, conf =
0.8, target = "rules"), appearance = list(default = "lhs", rhs =
c('Primary.Type=ASSAULT','Primary.Type=BATTERY',
'Primary.Type=CRIMINAL DAMAGE','Primary.Type=THEFT')))


result <- apriori(workdata, parameter = list(sup = 0.35, conf =
0.8, target = "rules"),appearance = list(default = "lhs", rhs =
c('District.Area=North','District.Area=Central','District.Area=S
outh')))

rules <- sort(result, decreasing = TRUE, by = "support")
rules <- sort(result, decreasing = TRUE, by = "confidence")
inspect(rules[1:4])

Two-key plot -
plot(rules, shading="lift", control=list(main = "Two-key plot"))
plot(rules, method = "paracoord", shading = "confidence")


igrh <- plot(top_rules, method = "graph")
igrh_df <- get.data.frame(igrh, what = "both")
nodes <- data.frame(id = igrh_df$vertices$name,
value = igrh_df$vertices$support,
title = ifelse(igrh_df$vertices$label=="",
igrh_df$vertices$name,
igrh_df$vertices$label), igrh_df$vertices )
edges <- igrh_df$edges
network <- visNetwork(nodes, edges) %>%
visOptions(manipulation = TRUE) %>%
visEdges(arrows = 'to', scaling = list(min = 2, max = 2)) %>%
visInteraction(navigationButtons = TRUE)
network
```

Exhibit 9 – Decision tree source code

```
myfile <- read.csv("Chicago Working Data.csv", na.strings = '?')

nrow(myfile[complete.cases(myfile),])

myfile = myfile[complete.cases(myfile),]

sample_size <- floor(0.16 * nrow(myfile))

training_index <- sample(nrow(myfile), size = sample_size,
replace = FALSE)

train <- myfile[training_index,]

test <- myfile[-training_index,]

predictors <-
c('ID','Case.Number','Block','IUCR','Primary.Type','Beat',

'District','District.area','Ward','Community.Area','FBI.Code')

train$Arrested <- as.factor(train$Arrested)

model <- C5.0(x = train[, predictors], y = train$Arrested)

summary(model)
```

Exhibit 10 – Naïve-Bayes source code

```
myfile <- read.csv("Chicago Working Data.csv", na.strings = '?')

nrow(myfile[complete.cases(myfile),])

myfile = myfile[complete.cases(myfile),]

sample_size <- floor(0.15 * nrow(myfile))

training_index <- sample(nrow(myfile), size = sample_size,
replace = FALSE)

train <- myfile[training_index,]

test <- myfile[-training_index,]

train$Arrested <- as.factor(train$Arrested)

myfile.model <- naiveBayes(Arrested ~ . , data = train)

myfile.predict <- predict((myfile.model), test, type = 'class')

arrest_results <- data.frame(actual = test[,'Arrested'],
predicted = myfile.predict)

table(arrest_results)

nrow(arrest_results[arrest_results$predicted ==
arrest_results$actual, ]) /

nrow(arrest_results)


myfile <- read.csv("Chicago South.csv", na.strings = '?')

myfile <- read.csv("Chicago Central.csv", na.strings = '?')

myfile <- read.csv("Chicago North.csv", na.strings = '?')

nrow(myfile[complete.cases(myfile),])

myfile = myfile[complete.cases(myfile),]

sample_size <- floor(0.5 * nrow(myfile))

training_index <- sample(nrow(myfile), size = sample_size,
replace = FALSE)

train <- myfile[training_index,]

test <- myfile[-training_index,]

train$Arrested <- as.factor(train$Arrested)
```

```
myfile.model <- naiveBayes(Arrested ~ . , data = train)

myfile.predict <- predict((myfile.model), test, type = 'class')

arrest_results <- data.frame(actual = test[,'Arrested'],
predicted = myfile.predict)

table(arrest_results)

nrow(arrest_results[arrest_results$predicted ==
arrest_results$actual, ]) /

nrow(arrest_results)
```

Resources:

Beat Officers. https://home.chicagopolice.org/, 2021. https://home.chicagopolice.org/community-policing-group/how-caps-works/beat-officers/#:~:text=Under%20CAPS%2C%20a%20team%20of,to%20know%20their%20beat%20officers. Accessed March 1, 2021

Crimes. https://www.chicago.gov/, 2021. https://www.chicago.gov/city/en/dataset/crime.html. Accessed February 23, 2021

https://data.cityofchicago.org/, 2021. https://data.cityofchicago.org/Public-Safety/Crimes-One-year-prior-to-present/x2n5-8w5q. Accessed March 2, 2021

https://images.fastcompany.net/, 2021. https://images.fastcompany.net/image/upload/w_596,c_limit,q_auto:best,f_auto/wp-cms/uploads/2020/03/1-90471389-exclusive-chicagoand8217s-new-brand-identity-could-save-the-city-dollar10-million-a-year.jpg. Accessed March 2, 2021

https://www.extendoffice.com/documents/excel/3293-excel-count-frequency-of-values.html, 2021. https://www.extendoffice.com/documents/excel/3293-excel-count-frequency-of-values.html. Accessed March 3, 2021

https://news.wttw.com/, 2012. https://news.wttw.com/sites/default/files/Map%20of%20Chicago%20Police%20Districts%20and%20Beats.pdf. Accessed March 1, 2021