Chicago Crimes: Predicting Crime in the Third Largest U.S. City

Christopher Craig

2024-03-15

Executive Summary

Chicago, IL is the third largest city in the United States in terms of population, only trailing New York and Los Angeles. When many Americans think of Chicago, they also think of crime as this is the topic most heavily dominated in the news cycles when covering Chicago. The amount of crime is a talking point with politicians, an often called upon theme when comparing to other cities of similar size.

For this analysis, I wanted to look at a hypothetical situation of a newly elected mayor who wants to predict the amount of crime the city will experience in order to properly allocate resources and set expectations for his/her term. The mayor has worked with criminologists and knows that cities generally experience more crime in the summer months when the weather is warmer and less crime when it's cold [1]. Knowing this, the mayor hypothesizes they can use historical Chicago weather data and other factors such as where the crime was committed as well as what kind of crime was committed in order to predict the number of crimes the city will experience.

This analysis will call upon two distinct data sets, both pulled from the Kaggle repository. The first data set has historical reported crime in Chicago, with observations such as what kind of crime was reported, when it was reported, where it was reported and if an arrest was made. This data set excludes murders and was initially extracted from the Chicago Police Department's CLEAR (citizen Law Enforcement Analysis and Reporting) system[2].

The second data set has historical weather information for Chicago, such as recorded temperature, humidity, wind speed, atmospheric pressure and date and time of the observations. Since date is the unique identifier in both data sets, I was able to merge the two data sets in order to formulate a prediction using two methods: 1) The root mean squared error (RMSE) and 2) Random Forests . I prefer the RMSE model as we can closely interpret this to the standard deviation in our prediction and we have a clear goal while running the model: to decrease the RMSE as much as possible. The model starts with the average amount of crime, then I will add the effects of the month, type of crime, district the crime occurred, as well as the average temperature of the month the crime occurred.

$$Y_{u,i} = \mu + b_i + \varepsilon_{u,i}$$

I also wanted to see if the regression improves with Random Forests as this model averages multiple decision trees, each tree being randomly different which should reduce instability and improve our model. This will give our hypothetical mayor options and show how well the models perform against one another.

First, I will explore both data sets prior to making the prediction using the RMSE and Random Forest algorithms.

[1] https://crimesciencejournal.biomedcentral.com/articles/10.1186/s40163-022-00179-8

[2] https://www.kaggle.com/datasets/adelanseur/crimes-2001-to-present-chicago

Exploratory Analysis of the Chicago Crime Dataset

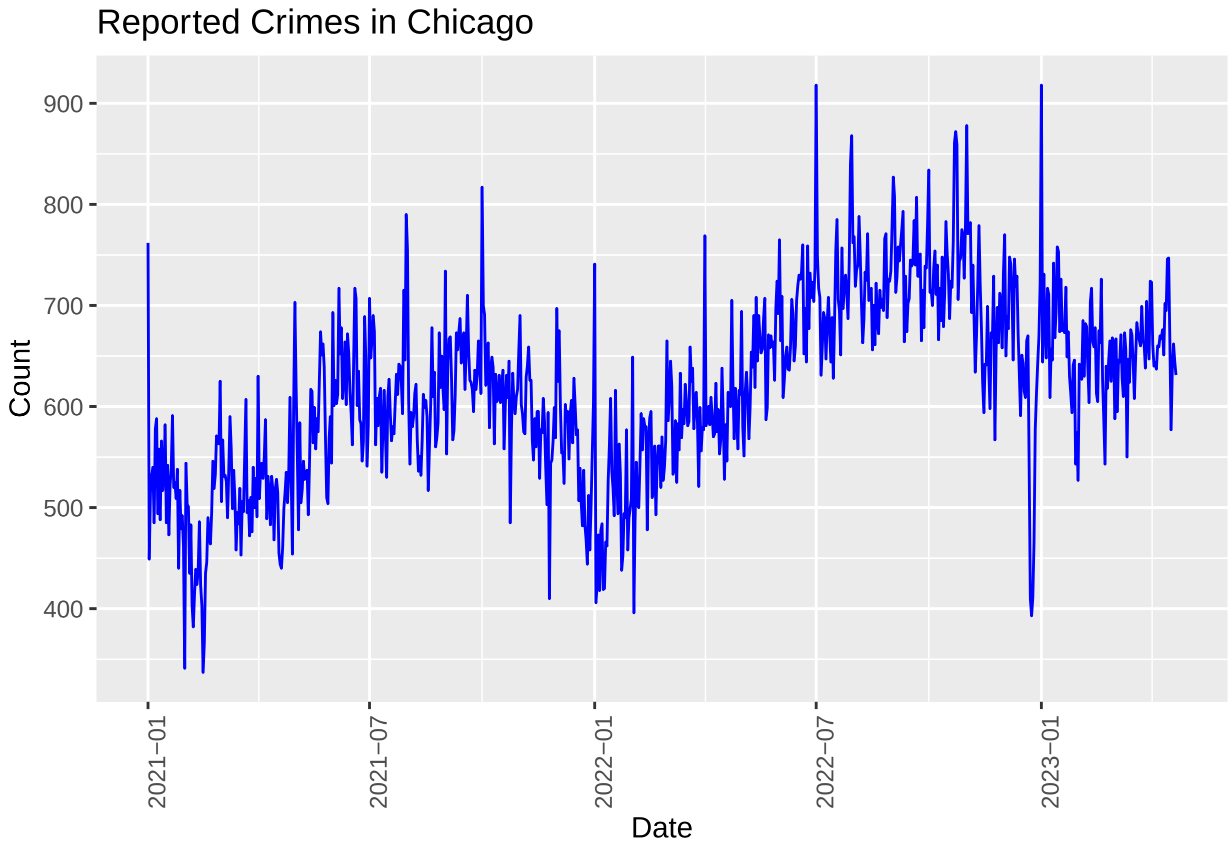
Let's first analyze the Chicago Crimes dataset. An important reminder: This data set reflects reported incidents of crime that occurred in the city of Chicago but excludes murders.

First, let's look at the structure of the Chicago Crimes data set:

```
## 'data.frame':    519926 obs. of  14 variables:
##  $ X                   : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ ID                  : int  12571973 12343475 12602803 12540388 12541139 12445976 12519170 1254049
##  $ Date                : chr  "12/19/2021 07:23:00 AM" "04/16/2021 08:45:00 PM" "10/21/2021 11:00:00
##  $ Block               : chr  "042XX S MOZART ST" "056XX N RIDGE AVE" "083XX S STONY ISLAND AVE" "08
##  $ Primary.Type        : chr  "BATTERY" "THEFT" "OTHER OFFENSE" "THEFT" ...
##  $ Location.Description: chr  "SIDEWALK" "OTHER (SPECIFY)" "OTHER (SPECIFY)" "CONVENIENCE STORE" ...
##  $ Arrest              : chr  "true" "false" "false" "false" ...
##  $ Domestic            : chr  "true" "false" "false" "false" ...
##  $ Beat                : int  921 2013 412 632 911 533 914 322 1431 833 ...
##  $ District            : int  9 20 4 6 9 5 9 3 14 8 ...
##  $ Ward                : int  15 48 8 6 12 9 11 6 1 13 ...
##  $ Year                : int  2021 2021 2021 2021 2021 2021 2021 2021 2021 2021 ...
##  $ time                : chr  "1 07:23:00 AM" "1 08:45:00 PM" "1 11:00:00 AM" "1 06:00:00 AM" ...
##  $ new_date            : chr  "2021-12-19" "2021-04-16" "2021-10-21" "2021-11-14" ...
```

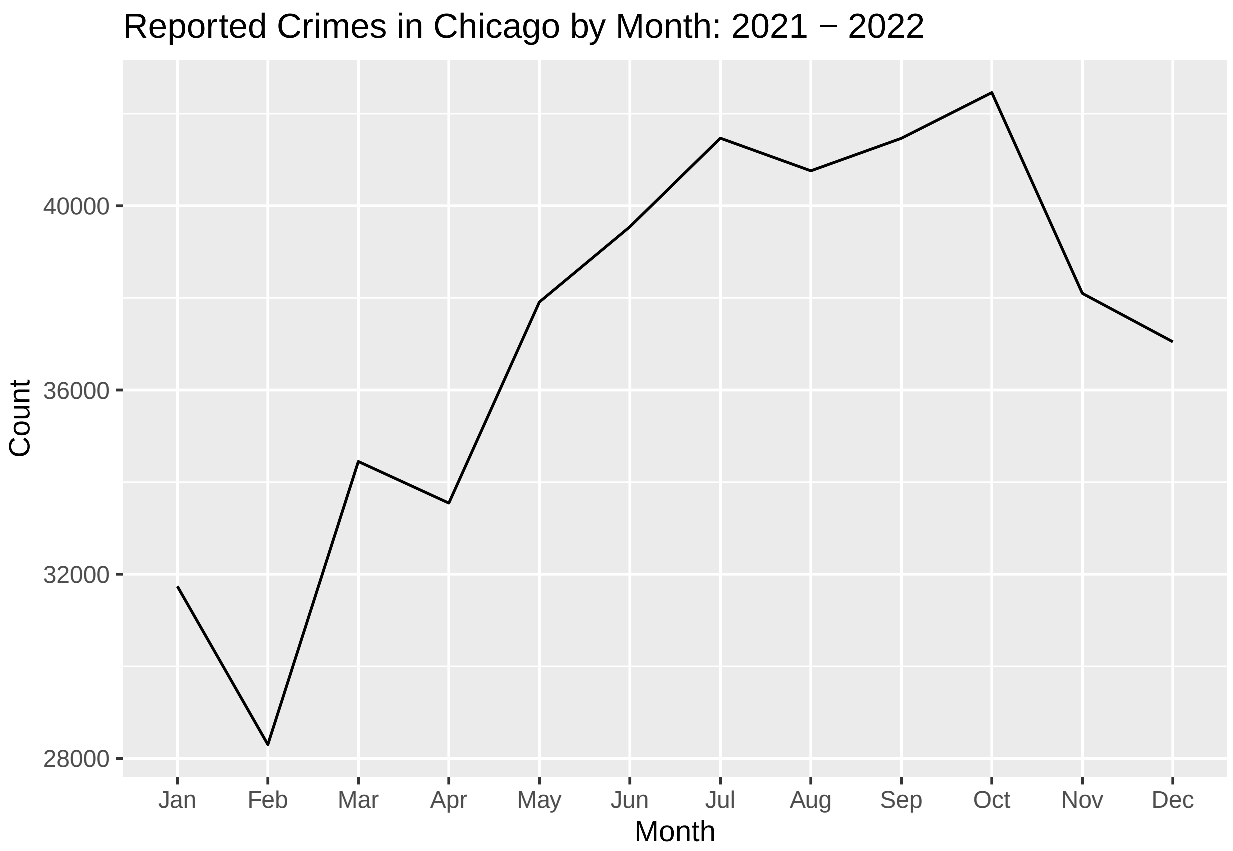
We see we have some key fields to work with, including the date, type of crime observed (Primary.Type), and District. The date will be important because further on in our analysis, I will delineate the data by month as our hypothetical mayor would like to predict the amount of crime on a monthly basis.

Next, let's take a high level look at the trend of crimes reported over time.



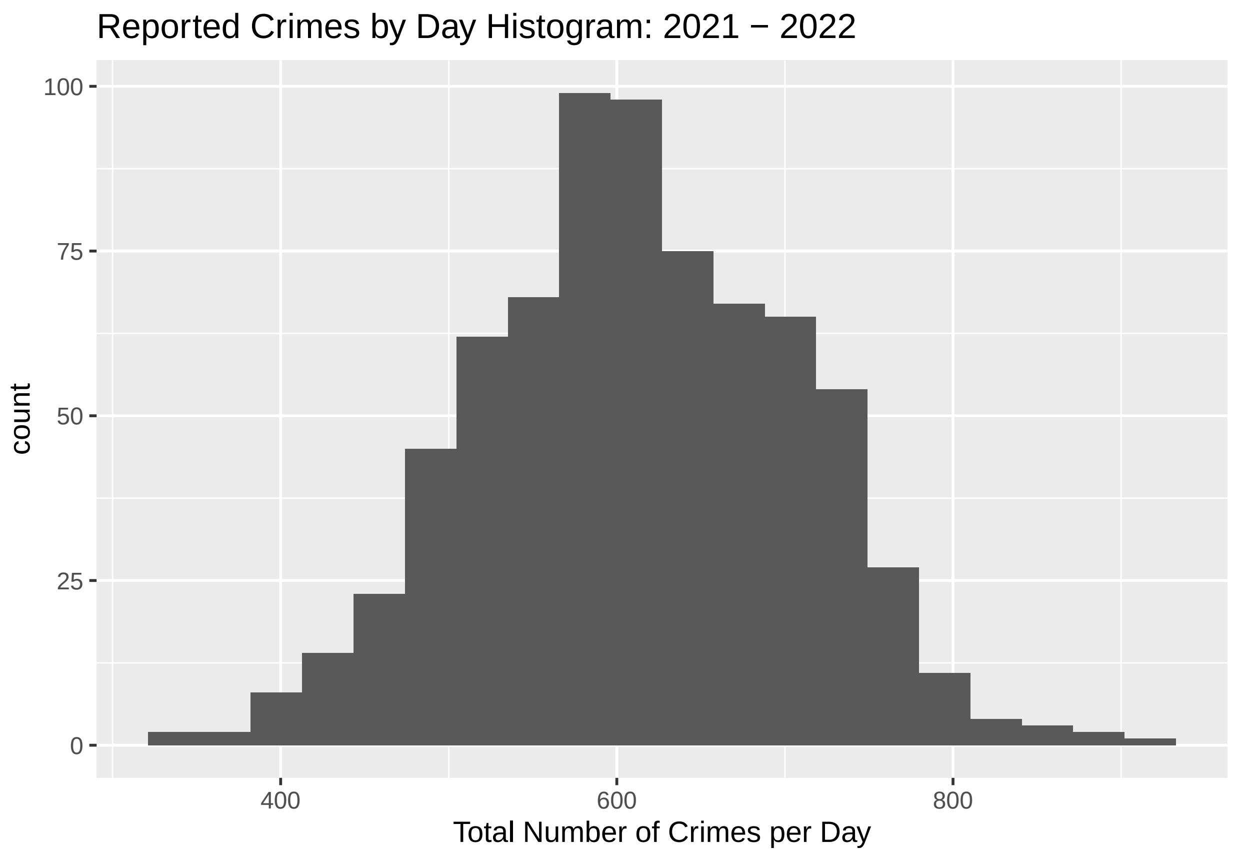
We notice a few things here. First, our hypothesis appears to be directionally correct: we see the number of reported crimes trend upwards in the summer months and trend downward around January of each year. We also see our data starts in 2021 and goes into 2023 so we have a good amount of observations to work with.

Due to the fact that 2023 only includes partial data for the year, we will remove this year to get a clean 2021 - 2022 data set. After doing so, let's see the total number of crimes per month



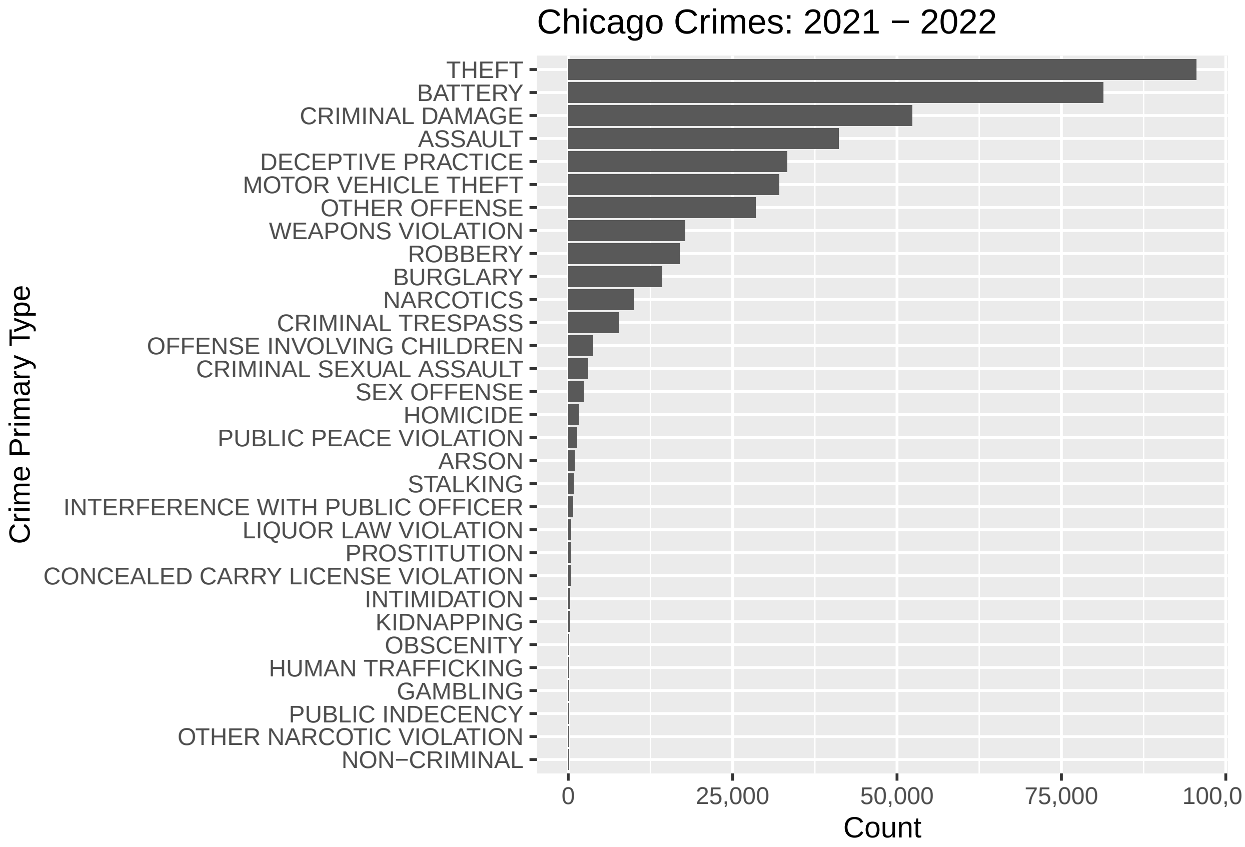
We see something interesting here. The number of of crimes reported over this two year period is higher in May through Sep but we also see that October is the month with the highest total number of crimes reported in this data set. However, after October we see the anticipated decline into November. February is the month in this data set with the lowest total number of reported crimes.

Next, let's look at a histogram of the total number of reported crimes.



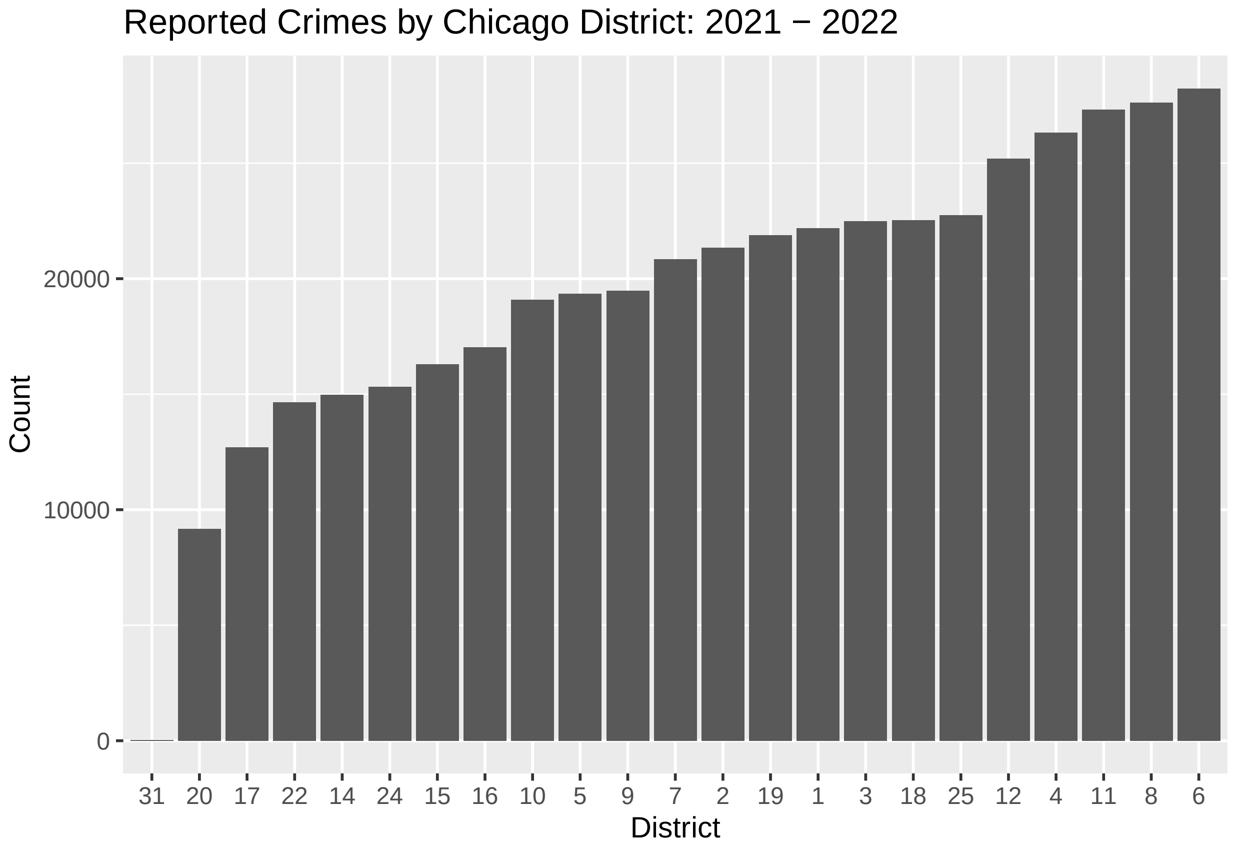
We see a normal distribution with approximately 600 crimes reported per day most often seen in the data set. That is certainly a lot of reported crime!

Next let's take a look at what kinds of crime are being reported in the data set and if there's any crime that gets reported on more than others.



This graph clearly shows certain types of crime are reported more often than others with Theft, Battery, Criminal Damage, and Assault reported the most.

Now, let's see if there are districts within Chicago where crime is reported more than others.



This graph clearly shows us that certain districts have more reported crime than others. Districts 4, 6, 8, 11, 12 each reported over 25,000 crimes in this two year period.

We now know some key pieces from the Chicago Crimes data set that will help us in our prediction: we know the amount of crimes reported varies by district, the types of crime varies, and the amount of crime reported will vary by month (typically colder months report less crime).

Next, let's look at the Chicago weather data set and merge with the Chicago Crimes data set.

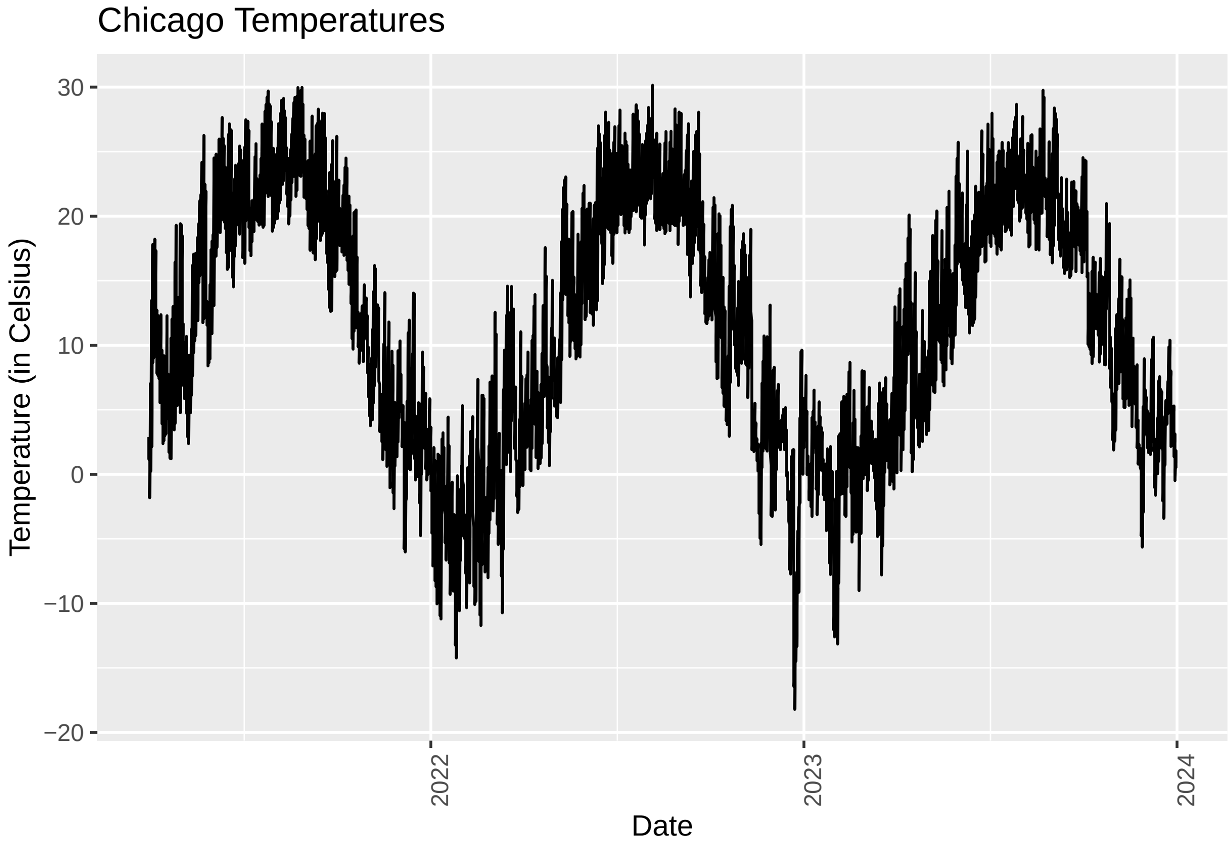
Exploratory Analysis of the Chicago Weather Dataset & Merging with the Crimes Dataset

Let's first look at the structure of the weather data set:

```
## 'data.frame':    24108 obs. of  11 variables:
##  $ date     : Date, format: "2022-08-06" "2022-08-06" ...
##  $ YEAR     : int  2022 2022 2022 2021 2021 2021 2021 2022 2021 2023 ...
##  $ MO       : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ DY       : int  6 6 6 28 24 28 28 6 27 23 ...
##  $ HR       : int  14 13 15 13 14 14 12 12 14 14 ...
##  $ TEMP     : num  30.1 30.1 30 30 30 ...
##  $ PRCP     : num  0.07 0.06 0.08 0.01 0.07 0.02 0.02 0.02 0 0.01 ...
##  $ HMDT     : num  59.6 59 61.4 67.1 61.6 ...
##  $ WND_SPD  : num  3.34 3.03 3.84 3.89 4.82 4.25 3.55 2.87 3.26 6.59 ...
##  $ ATM_PRESS: num  99.4 99.5 99.4 99.6 99.2 ...
##  $ REF      : int  202208 202208 202208 202108 202108 202108 202108 202208 202108 202308 ...
```

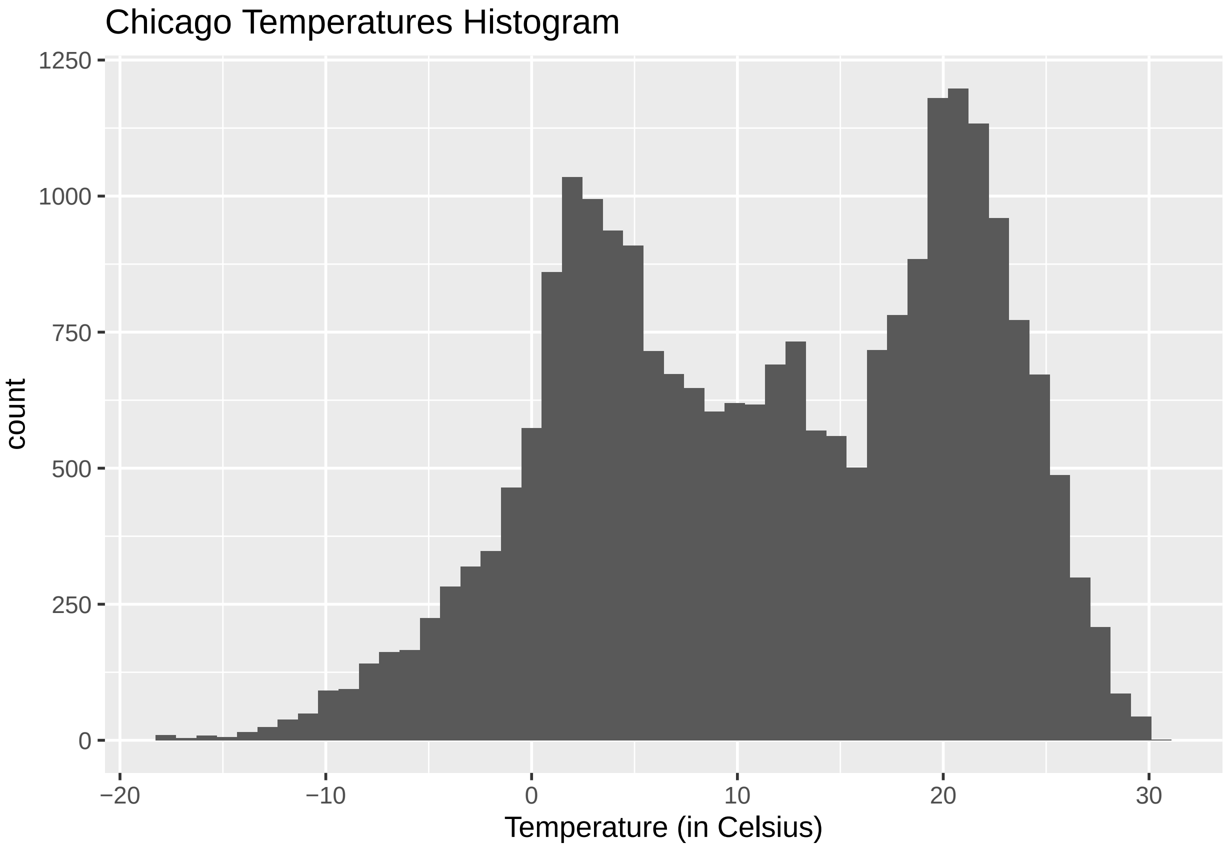
There are some key pieces of data we can utilize: from the Year, Mo, and Dy attributes we formulated a date that we can then use to join with our crimes data set. While we will solely be using the TEMP (temperature) variable, we see we also have observations on the amount of precipitation (PRCP), humidity (HMDT), wind speed (WND_SPD) and atmospheric pressure (ATM_PRESS).

Let's look at temperatures by day in Chicago



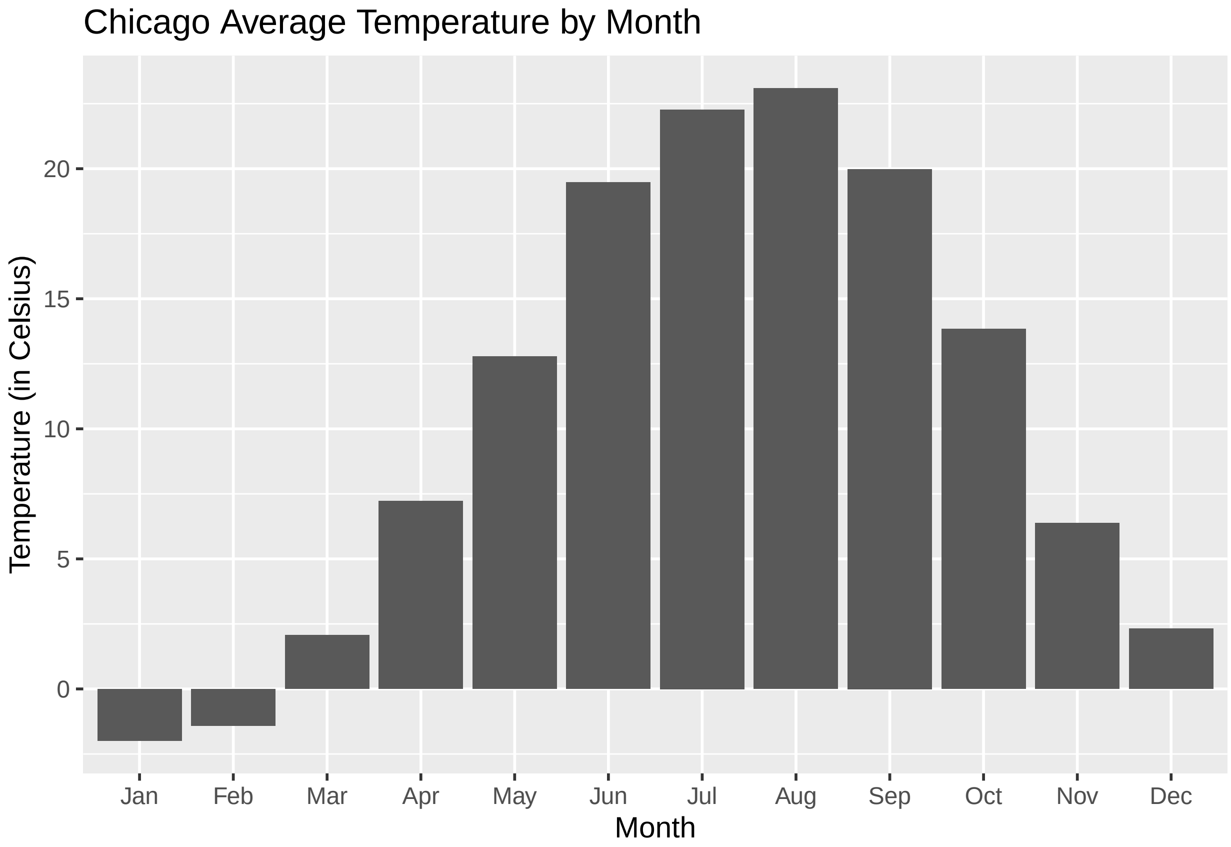
This chart isn't overly insightful, we see temperatures get higher in summer months and lower in winter months. The range is 30 to -20 Celsius and we see this data goes from 2021 to 2024 which is good because there is overlap between this and the observation dates in the crimes data set.

Using this data set we can look a a histogram of temperatures:



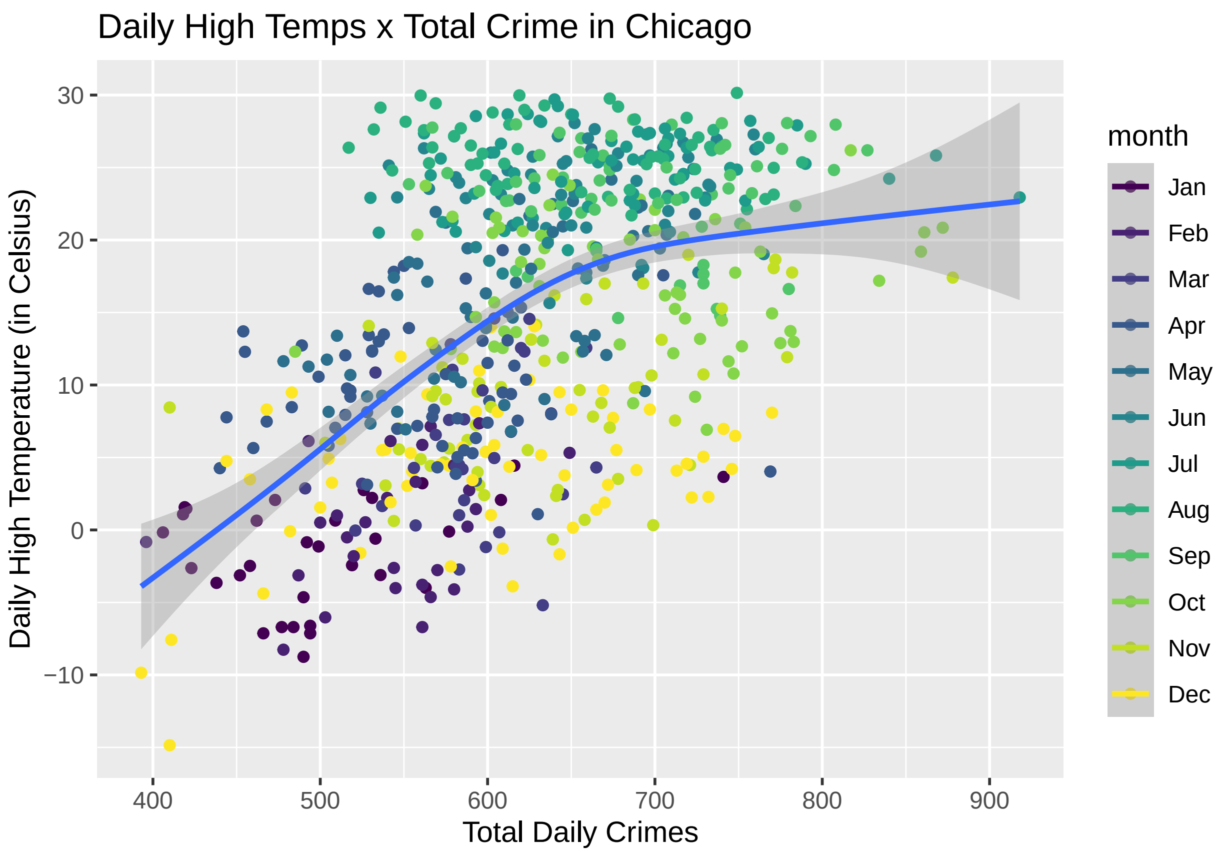
This histogram shows we have a bi-modal distribution, with peaks around 2 and 24 degrees celcius.

We can also view the average temperatures by month in the weather data set:



Again, not overly insightful but good to see how the temperature widely varies in Chicago.

Let's now do an initial join, by date, with the Crimes data set to see if there is a correlation between temperatures and reported crime. For this I am going to use the maximum temperature reported for a given date



It appears from this table there is a positive correlation between daily high temps and daily total reported crimes. We can also look at the correlation coefficient between high temps and total crimes.

```
## # A tibble: 1 x 1
##   Correlation_Coefficient
##                     <dbl>
## 1                   0.521
```

We see from the correlation coefficient that there is moderate correlation between high temps and total crime. I suspect the correlation would be higher but we see in the exploratory analysis on the crimes data set that October was the month with the most crime and we know from the exploratory analysis that there are warmer months than October.

Now let's join the key elements from the weather data set to the crimes data set. Since our hypothetical mayor wants to predict the number of crimes by month, I am going to group each data set by month and use that as the join.

```
## # A tibble: 6 x 9
## # Groups:   month, Primary.Type [1]
##   month Primary.Type District total  TEMP   PRCP  HMDT WND_SPD ATM_PRESS
##   <ord> <chr>           <int> <int> <dbl>  <dbl> <dbl>   <dbl>     <dbl>
## 1 Jan   ARSON               1     3 -1.99 0.0580  86.3    6.59      99.6
## 2 Jan   ARSON               2     6 -1.99 0.0580  86.3    6.59      99.6
## 3 Jan   ARSON               4     6 -1.99 0.0580  86.3    6.59      99.6
## 4 Jan   ARSON               5     5 -1.99 0.0580  86.3    6.59      99.6
## 5 Jan   ARSON               6     1 -1.99 0.0580  86.3    6.59      99.6
## 6 Jan   ARSON               7     7 -1.99 0.0580  86.3    6.59      99.6
```

Modeling using RMSE to predict Number of Crimes Per Month

In order to predict the number of monthly crimes, I first partitioned our compiled table into a training and test set with a 80/20 split, respectively. I chose 80/20 to prevent over fitting the on this data set. I believe this produces an optimal analysis for that crime data to be predicted. The code to do so is listed below.

```
 ## This creates the RMSE function we will use for our prediction

RMSE <- function(true_crime, predicted_crime) {
  sqrt(mean((true_crime - predicted_crime)^2))
}

## This partitions the data into Test and Train Sets

y <- compiled$total
index <- createDataPartition(y, times = 1, p = 0.2, list = FALSE)
test_set <- compiled[index, ]
train_set <- compiled[-index, ]
```

The first model will assume the average. Inserting this into our algorithm, the RMSE on this simplistic approach is 117.76.

```
# Calculates average
mu <- mean(train_set$total)

# First Model simply takes average mu and applies to RMSE
model1_rmse <- RMSE(test_set$total, mu)

# Create a tibble with results
model1_results <- tibble(Method = "Avg", RMSE = model1_rmse)
model1_results
```

```
## # A tibble: 1 x 2
##   Method  RMSE
##   <chr>  <dbl>
## 1 Avg     118.
```

The second model will build upon this by adding the effect the month has to the prediction. As we saw in the exploratory analysis, the number of crimes varies by month. When adding this to our model, the RMSE improves slightly to 117.99.

```
 # Creates averages by month
month_averages <- train_set %>% group_by(month) %>% summarize(b_month = mean(total - mu))

# Adds monthly averages to the prediction
prediction <- test_set %>% left_join(month_averages, by = 'month') %>% mutate(predicted = mu+ b_month) %

# Calculates RMSE
month_rmse <- RMSE(test_set$total, prediction)

# Create a tibble with results

model2_results <- tibble(Method = "Month", RMSE = month_rmse)
model2_results
```

```
## # A tibble: 1 x 2
##   Method  RMSE
##   <chr>  <dbl>
## 1 Month   118.
```

The third model will add the averages by crime type. Per the exploratory analysis, some crimes are reported more often than others. When adding this to our model, the RMSE greatly improves to 61.13.

```
# Creates averages by crime type
type_averages <- train_set %>% left_join(month_averages, by = 'month') %>%
  group_by(Primary.Type) %>% summarize(b_type = mean(total - mu - b_month))

# Add crime type averages to prediction
prediction <- test_set %>% left_join(month_averages, by = 'month') %>%
  left_join(type_averages, by = 'Primary.Type') %>% mutate(predicted = mu + b_month+b_type) %>%
  pull(predicted)

# Calculates RMSE
type_rmse <- RMSE(test_set$total, prediction)

# Create a tibble with results
model3_results <- tibble(Method = "Type", RMSE = type_rmse)
model3_results
```

```
## # A tibble: 1 x 2
##   Method  RMSE
##   <chr>  <dbl>
## 1 Type    61.1
```

The fourth model will add the averages by district as we saw some districts had much more crime compared to other districts around Chicago. This improves the RMSE to 56.79 after adding to our model.

```
# Create averages by district
district_averages <- train_set %>% left_join(month_averages, by = 'month') %>%
  left_join(type_averages, by = 'Primary.Type') %>% group_by(District) %>%
  summarize(b_district = mean(total - mu - b_month - b_type))

# Add district averages to prediction
prediction <- test_set %>% left_join(month_averages, by = 'month') %>%
  left_join(type_averages, by = 'Primary.Type') %>% left_join(district_averages, by = 'District') %>%
  mutate(predicted = mu + b_month + b_type + b_district)  %>% pull(predicted)

# Calculates RMSE
district_rmse <- RMSE(test_set$total, prediction)

# Create a tibble with results
model4_results <- tibble(Method = "District", RMSE = district_rmse)
model4_results
```

```
## # A tibble: 1 x 2
##   Method     RMSE
##   <chr>     <dbl>
## 1 District   56.8
```

The fifth and final model will add the averages by temperature to the prediction. While we know there is correlation between the month and the temperatures, this should account for some unpredictability when there are months that have relatively high crime and lower temps (e.g. October). Also, I will set the minimum predicted value to be 0 since there can't be negative reported crime. This improves the RMSE to 55.98, exceeding our target of 60.

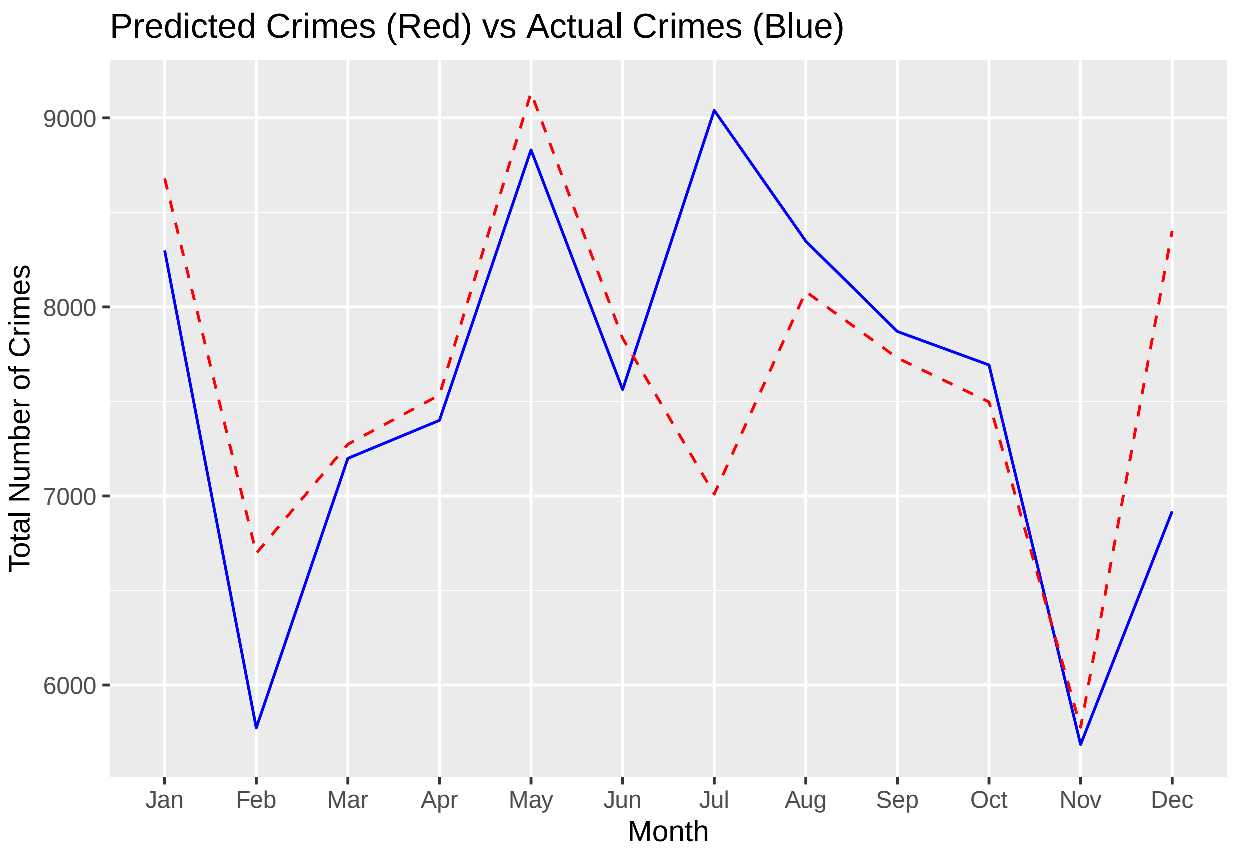
```
# Creates averages by temperature
temp_averages <- train_set %>% left_join(month_averages, by = 'month') %>% left_join(type_averages, by =
  left_join(district_averages, by = 'District') %>% group_by(round = round(TEMP)) %>%
  summarize(b_temp = mean(total - mu - b_month - b_type - b_district))

# Adds temperature averages to prediction
# Also set the minimum predicted value to be 0 (there can't be negative crimes)
prediction <- test_set %>% left_join(month_averages, by = 'month') %>%
  left_join(type_averages, by = 'Primary.Type') %>% left_join(district_averages, by = 'District') %>%
  mutate(round = round(TEMP)) %>% left_join(temp_averages, by = 'round') %>%
  mutate(predicted = mu + b_month + b_type + b_district + b_temp) %>%
  mutate(predicted = ifelse(predicted < 0,0,predicted)) %>% pull(predicted)

#Calculates RMSE
temp_rmse <- RMSE(test_set$total, prediction)

# Create a tibble with results
model5_results <- tibble(Method = "Temperature", RMSE = temp_rmse)
model5_results
```

```
## # A tibble: 1 x 2
##   Method       RMSE
##   <chr>       <dbl>
## 1 Temperature  56.0
```

To summarize how the RMSE progressed through the five models:

```
##            Models      RMSE
## 1          1. Avg 117.76248
## 2      2. + Month 117.98562
## 3       3. + Type  61.13370
## 4 4. + District  56.79114
## 5       5. + Temp  55.98494
```

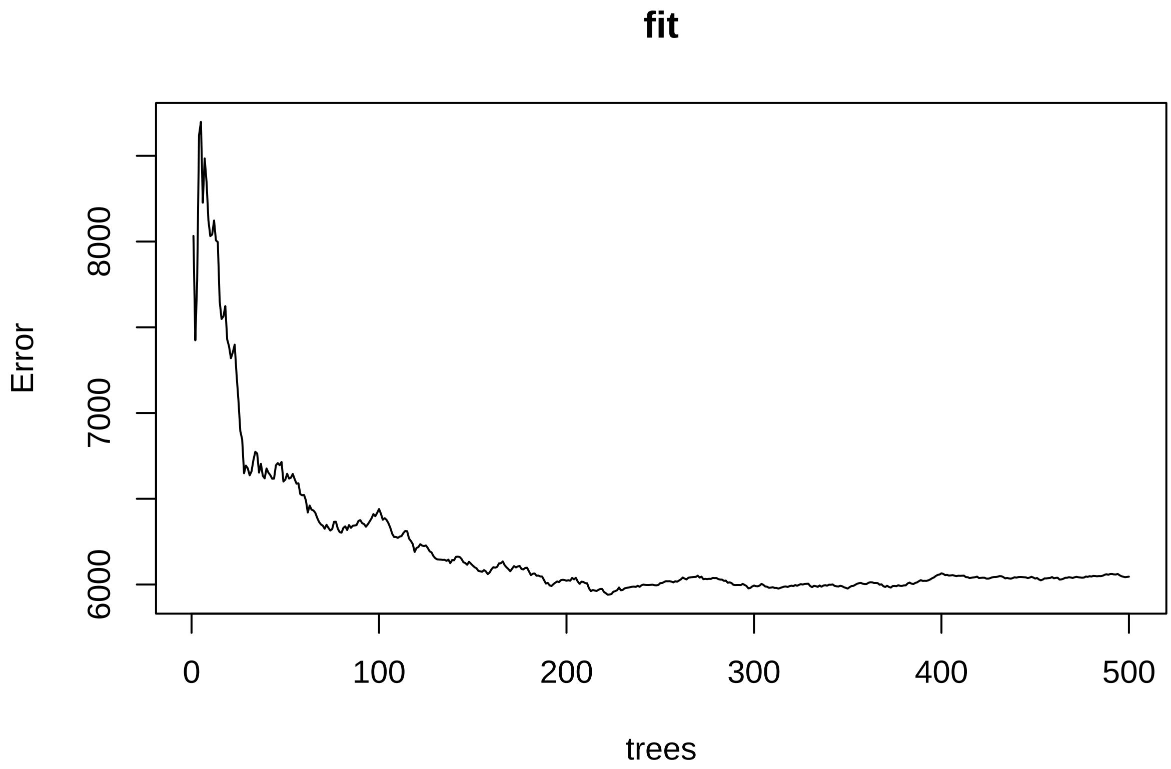
We can visually see how well our model performs compared to actual number of crimes in the test set.



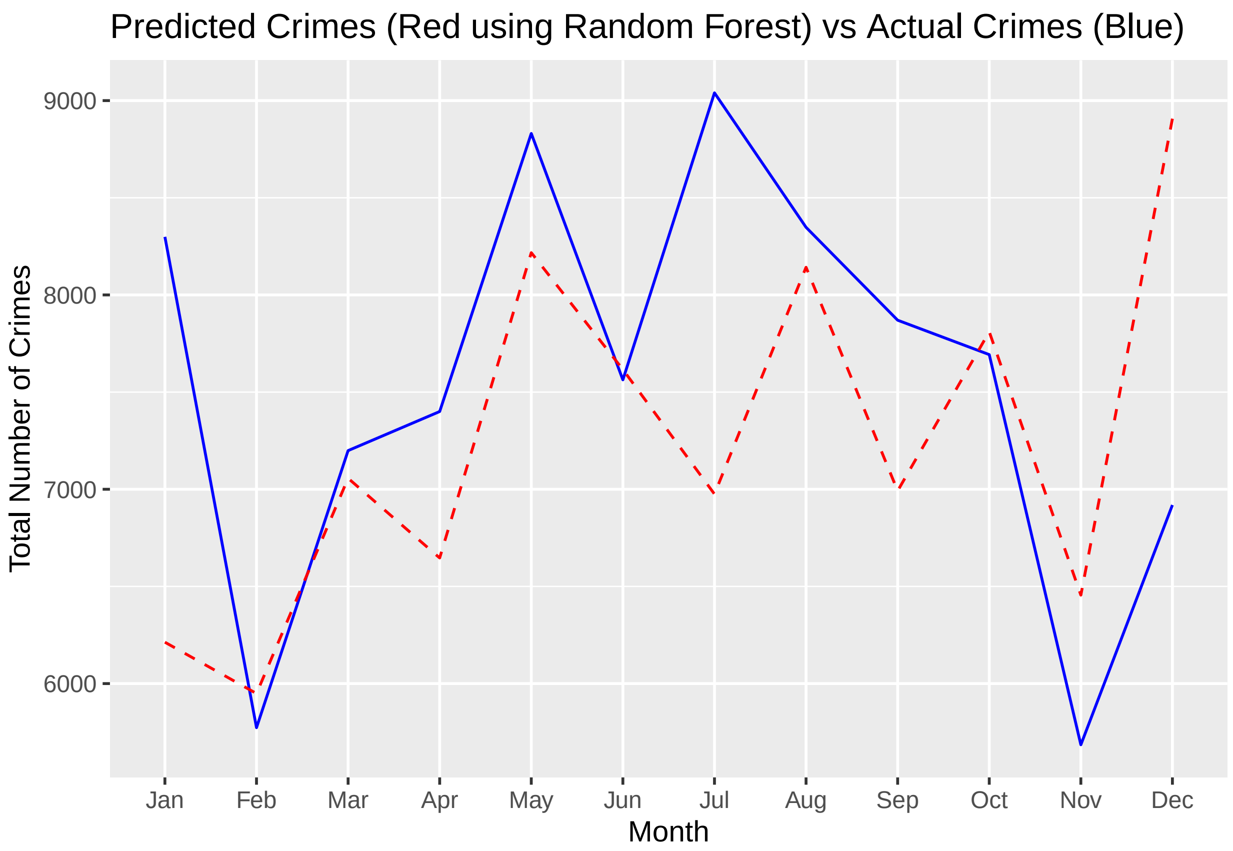
Modeling using Random Forest to predict Number of Crimes Per Month

I believe an RMSE of around 56 to be very good in this case, implying the average error off by 56 crimes per month, per crime type, per district. However, let's see if Random Forests can improve upon this.

First let's fit the Random Forest model on the training set and plot the results.



This shows us the accuracy stabilizes around 200 decision trees. Now let's apply the prediction to the test set and create a new graph to see how it compares to the actual crimes reported by month.



Lastly, we can calculate an RMSE using the predictions captured by the Random Forest model to see if the RMSE is better.

```
## # A tibble: 1 x 2
##   Method         RMSE
##   <chr>         <dbl>
## 1 Random Forest  80.7
```

Surprisingly, we find that using the Random Forest did not improve the RMSE we calculated prior. I suspect this is due to the fact that our compiled data set was built upon time series data.

Conclusion

In conclusion, the final RMSE calculated in our modeling is 55.98. I can take these findings to the hypothetical mayor ane explain that based on the month, district, type of crime, and average temperature, our average standard deviation is around 56 reported crimes compared to actuals. While not perfect, it's clearly better than guessing our simply using the average as a predictor.

With this knowledge, the hypothetical mayor can use this prediction to appropriately deploy resources, for example, more police officers during months in which our model predicts more crimes will be reported. Politically, the mayor may use this to set expectations with his/her constituents and use it as political capital if they are able to report a lower actual than the prediction.

While I touched on temperature as a factor in the model, other meteorological factors could be considered in future iterations, such as considering days with heavy snow or rain and seeing how this impacts the amount of crime in the city. Non meteorological factors could also be in play such as holidays or other large events the city of Chicago hosts.

An interesting aspect for future work would be to apply this model to other major cities to predict their crime levels. New York and Los Angeles would be candidates as well as major international cities such as Paris or Madrid.

References

Chicago Crimes Kaggle Data set: https://www.kaggle.com/datasets/utkarshx27/crimes-2001-to-present

Chicago Weather Kaggle Data set: https://www.kaggle.com/datasets/curiel/chicago-weather-database

https://crimesciencejournal.biomedcentral.com/articles/10.1186/s40163-022-00179-8