Crime in Chicago

Victoria Seliger

April 25, 2018

Section I: Introduction

In my project, I am exploring the relationship between different variables in Chicago and the occurrences of crimes. My working hypothesis is that as temperature increases, so does the number of violent crimes. Additionally, crime is very segregated in Chicago, so a few Wards and community areas will have a large proportion of crime occurrences. These relationships are often referenced in movies and music, and I wanted to see if the relationships had any statistical backing. I am using 2018 crime data imported from the Chicago crime data portal. This data has around 35000 observations and is a record of every crime committed in Chicago, logged by description, ward, community area, type of crime, and other descriptors. This data is especially interesting and relevant to me because as a native Chicagoan, the issues of violence and crime facing our city have always been a part of my life. I am also using temperature data from the Chicago data portal. This is two temperature recordings per day, gathered from beach stations in Chicago. The main difficulty in this project will be manipulating the data so that it is in forms I can work with. For example, averaging two daily temperature reports into one for each day, and turning the crime logs into a crime count for each day.

Section II: Exploratory Data Analysis

First, I read in all the data I will be using. I am using Socrata to read in so that I can use the most up to date data as possible. Both my Crime and Weather data are from the city of Chicago data portal.

My dependent variable is the count of crimes occurring. I will be exploring the relationship between crime and daily temperature, crime and ward, and crime and day of week to see which variable can most accurately explain the occurrence of crime.

Explanatory and Response variables

Crime count by day

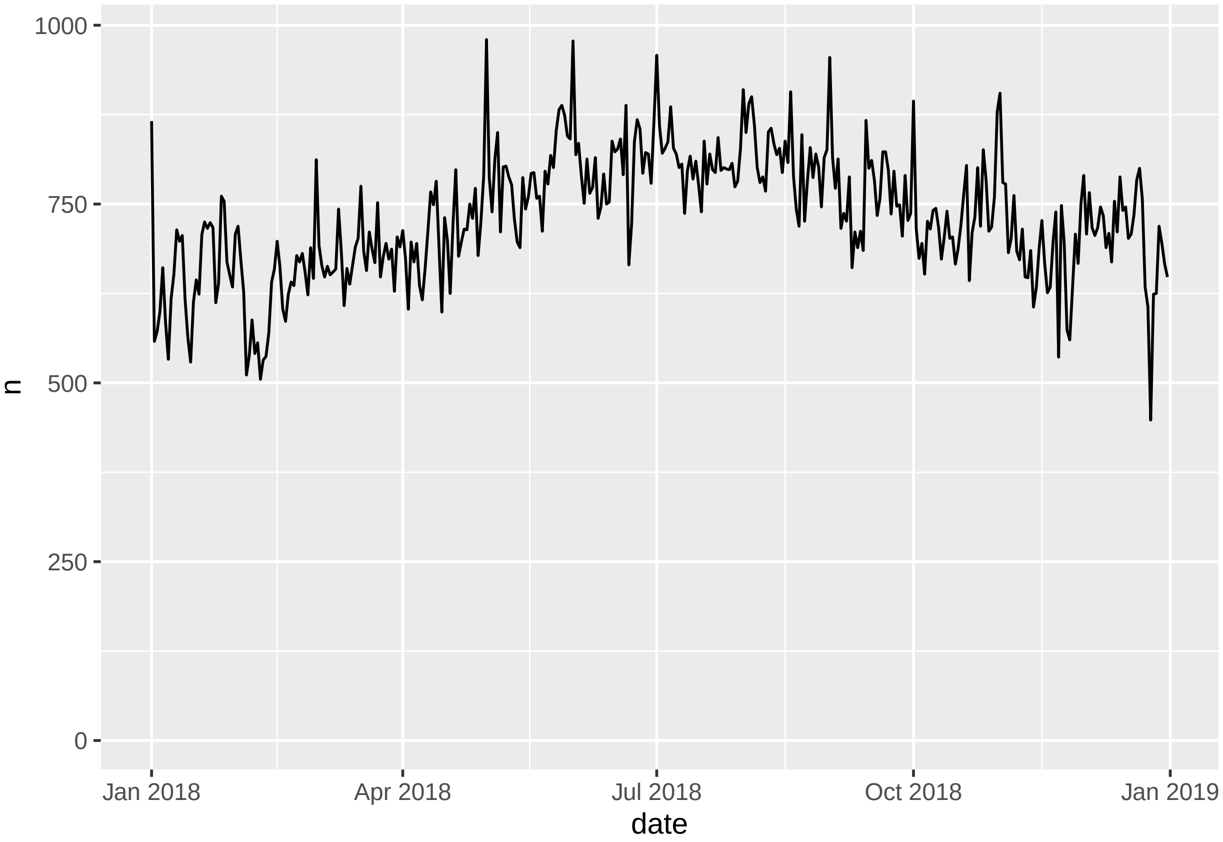
```r
summary(crimeCount)
```

```
##       date                  n
##  Min.   :2018-01-01   Min.   :  8.0
##  1st Qu.:2018-04-02   1st Qu.:673.0
##  Median :2018-07-02   Median :728.0
##  Mean   :2018-07-02   Mean   :730.2
##  3rd Qu.:2018-10-01   3rd Qu.:796.0
##  Max.   :2018-12-31   Max.   :980.0
##  NA's   :1
```

From the summary of daily crime counts, we can see the range of both the date and count data. We can see that the minimum number of crimes in a day was 8 and the maximum was 980. On average, 728 crimes are committed per day.

```r
ggplot(crimeCount, aes(date,n))+ geom_line()
```

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

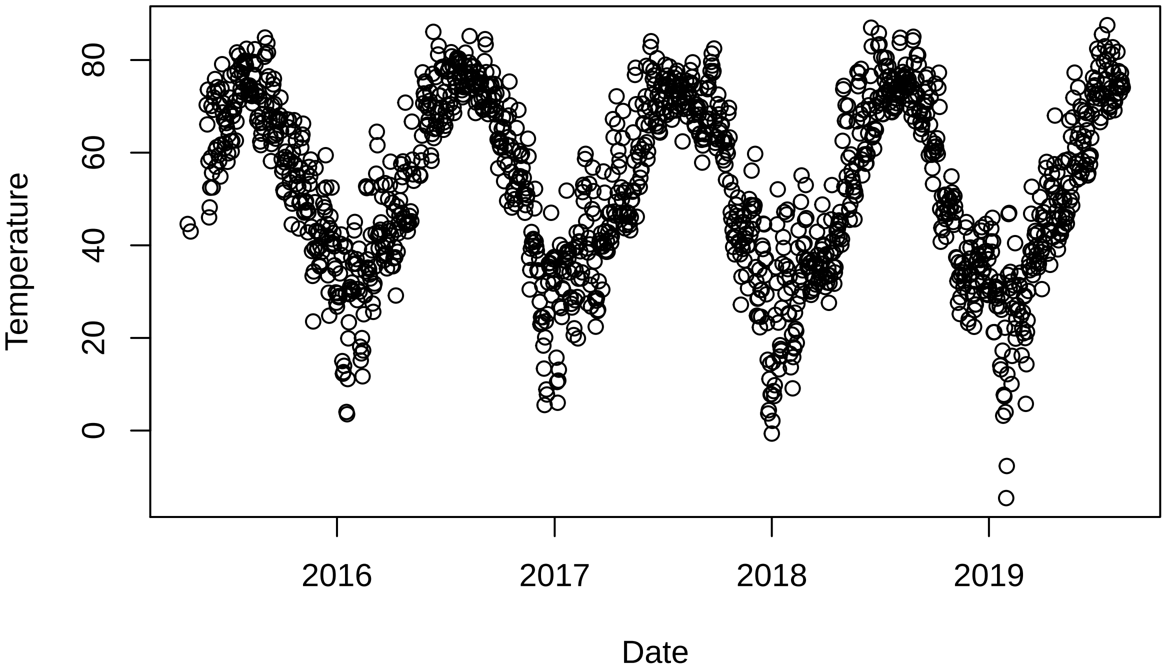


```
summary(tempAdjusted$temp)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##  -14.57   38.45   53.52   52.91   69.82   87.52       5
```

I am looking at this data combined with the temperature data for each day. From the summary of temperature data, we can see the range of temperatures as well as the average daily temperature, 53.5229798.

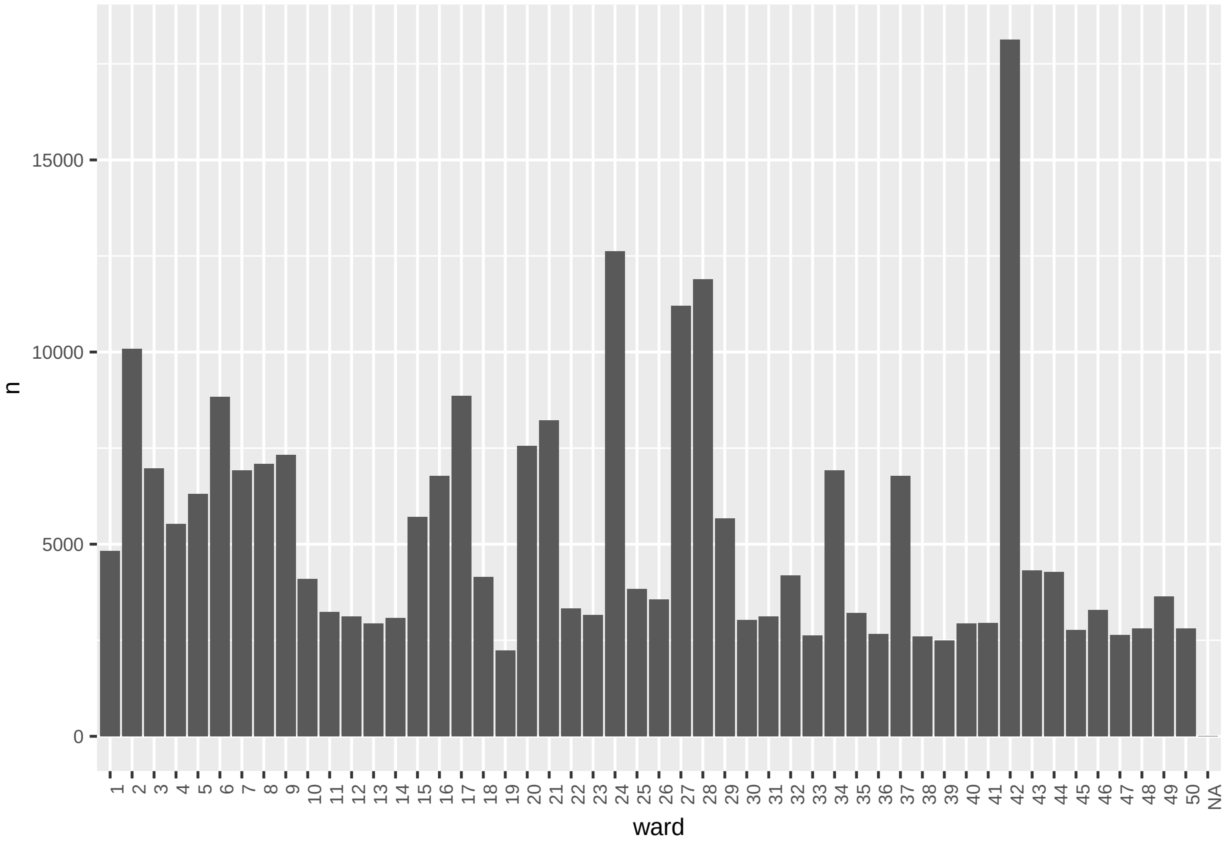
```
plot(tempAdjusted$date, tempAdjusted$temp, xlab = "Date", ylab = "Temperature")
```



Crime Count by ward

From the summary of crime count by wards, we can see the range of both the ward and count data. We can see that the minimum number of crimes in a ward is 4 and the maximum was 18131. On average, 4089 crimes are committed in each ward. We can see from the graph that this is not a uniform distribution, with the 42nd ward being an outlier.

```
ggplot(crimeCountWard, aes(ward,n))+geom_col()+theme(text = element_text(size=9),
        axis.text.x = element_text(angle=90, hjust=1))
```



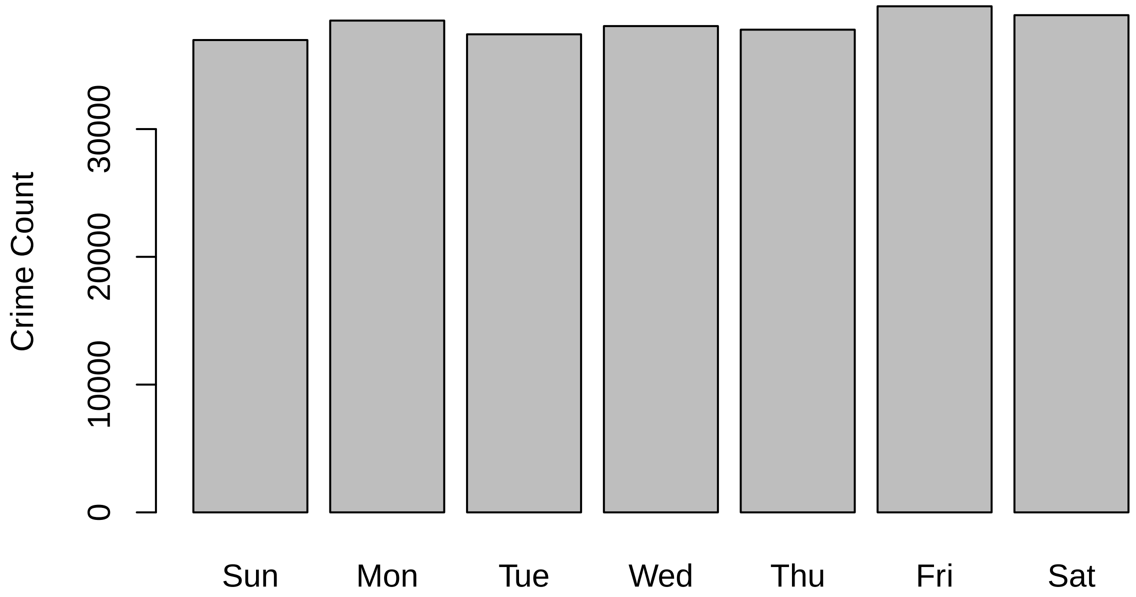
Crime count by day of week

From the summary of crime count by day of week, we can see the count of crime that has occurred on each day of week. We can see that the counts have a pretty small range- there is not much variability between days of the week in terms on number of crimes occurring. From the graph, the data appears skewed left, meaning the median is greater than the mean. More crime tend to happen later on in the week, such as on Friday or Saturday.

```
CrimeData2018socrata$wday <- wday(CrimeData2018socrata$date, label = TRUE)
summary(CrimeData2018socrata$wday)
```

```
##   Sun   Mon   Tue   Wed   Thu   Fri   Sat  NA's
## 36969 38491 37417 38061 37780 39612 38916     8
```

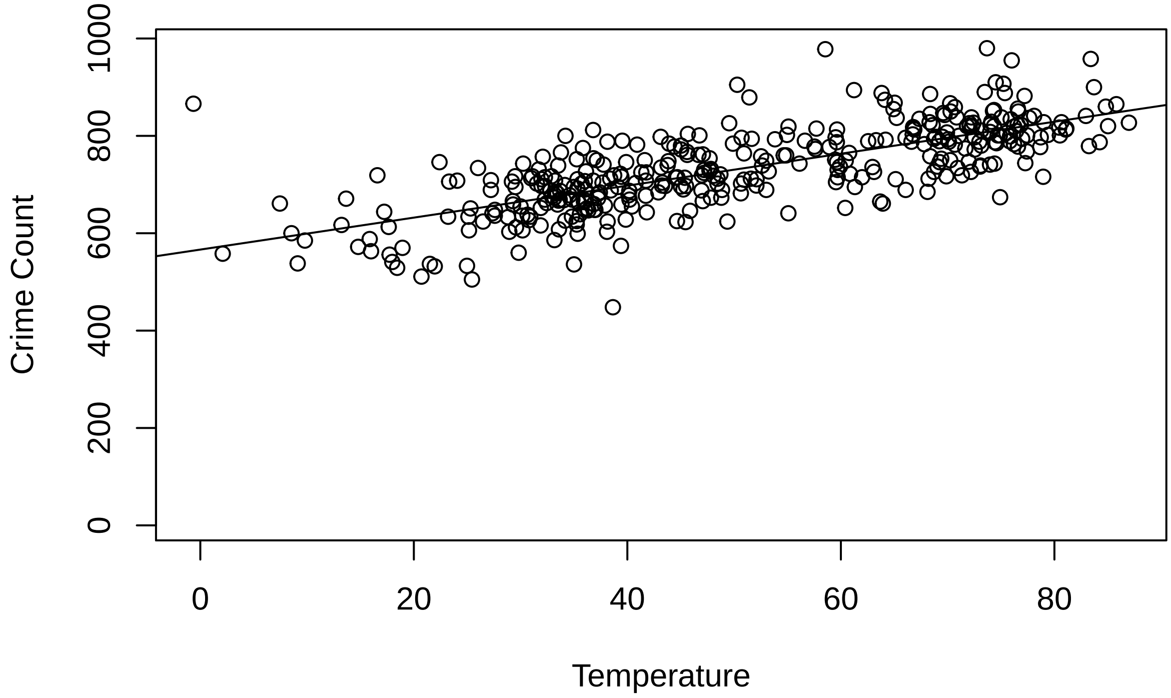
```
plot(CrimeData2018socrata$wday, ylab ="Crime Count" )
```



Section III: Simple Linear Regression

For my regression, I will explore the relationship between temperature and number of crimes committed on a given day. I expect that as temperature increases, the number of crimes committed increases.

```
fit1<-lm(combinedSet$n~combinedSet$temp)
plot(combinedSet$temp, combinedSet$n, xlab = "Temperature", ylab = "Crime Count")+abline(fit1)
```



```
## integer(0)
```

As seen in my regression model, there is a positive relationship between temperature and crimes committed, confirming my hypothesis. The coefficient is 3.2819222, which means that according to the model, for every 1 degree higher the temperature is, 3.2819222 more crimes are committed that day, holding all other variables constant. The R^2 value of my regression line is 0.5354451, meaning that temperature explains 0.5354451 of variability in crimes committed, suggesting that this model is not very good. This is surprising to me because I thought temperature would have been a good indicator, but this value may be low because I am only using data from 2018, and we are only 4 months into the year.

Section IV: Multiple Linear Regression

The model I wanted to explore was crime count per day as explained by weekday, temperature, and month.

I started by looking at a regression of crime count per day with predictors weekday and daily temperature.

```
summary(fit4)
```

```
##
## Call:
## lm(formula = combinedSet$n ~ factor(combinedSet$weekday) + combinedSet$temp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -231.092  -32.952   -0.911   32.065  307.923
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    560.2081    11.0100  50.882  < 2e-16 ***
## factor(combinedSet$weekday)Mon  -8.6823    11.1020  -0.782 0.434704
## factor(combinedSet$weekday)Tue   3.3068    11.1021   0.298 0.765990
## factor(combinedSet$weekday)Wed  -2.4199    11.1023  -0.218 0.827582
## factor(combinedSet$weekday)Thu  37.2264    11.1019   3.353 0.000885 ***
## factor(combinedSet$weekday)Fri  23.9908    11.1019   2.161 0.031363 *
## factor(combinedSet$weekday)Sat -16.3261    11.1017  -1.471 0.142280
## combinedSet$temp                 3.3010     0.1542  21.402  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.88 on 357 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.5771, Adjusted R-squared:  0.5689
## F-statistic: 69.61 on 7 and 357 DF,  p-value: < 2.2e-16
```

As seen in my regression model, there is some relationship between weekday and temperature and crimes committed per day. The p values for the variables at the 10% significance level indicate that all the variables are significant predictors.

The coefficients are:

Intercept (Sunday)- The model says that at zero degrees of temperature, 560.2081488 crimes are committed on Sundays, holding all other variables constant.

Monday- -8.6822762, which means that according to the model, on Mondays -8.6822762 more crimes are committed on Mondays than on Sundays, holding all other variables constant.

Tuesday- 3.3067895, on Tuesdays, 3.3067895 more crimes are committed than on Sundays, holding all other variables constant.

Wednesday- -2.4198899, on Wednesdays, -2.4198899 more crimes are committed than on Sundays, holding all other variables constant.

Thursday- 37.2264117, on Thursdays, 37.2264117 more crimes are committed than on Sundays, holding all other variables constant.

Friday- 23.990779, on Fridays, 23.990779 more crimes are committed than on Sundays, holding all other variables constant.

Saturday- -16.3261451, on Saturdays, -16.3261451 more crimes are committed than on Sundays, holding all other variables constant.

The temperature coefficient of temperature 3.300977 indicates that for every 1 degree higher the temperature goes, 3.300977 more crimes are committed per day, holding all other variables constant.

The adjusted R^2 value of my regression line is 0.5688537, meaning that the predictors weekday and temperature explains 0.5688537 of variability in crimes committed, suggesting that this model is somewhat good.

Next, I decided to look at a regression of crime count per day with predictors month and daily temperature.

```
summary(fit5)
```

```
##
## Call:
## lm(formula = combinedSet$n ~ factor(combinedSet$month) + combinedSet$temp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -272.417  -31.602   -2.395   30.562  245.324
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  582.2922    13.6434  42.679  < 2e-16 ***
## factor(combinedSet$month)Feb -41.9852    14.5675  -2.882  0.00419 **
## factor(combinedSet$month)Mar   6.5333    14.3936   0.454  0.65018
## factor(combinedSet$month)Apr  25.4600    14.9334   1.705  0.08909 .
## factor(combinedSet$month)May  58.5672    18.0040   3.253  0.00125 **
## factor(combinedSet$month)Jun  49.3174    19.6841   2.505  0.01268 *
## factor(combinedSet$month)Jul  40.9687    20.9657   1.954  0.05148 .
## factor(combinedSet$month)Aug  44.2868    21.4189   2.068  0.03940 *
## factor(combinedSet$month)Sep   8.3326    19.7101   0.423  0.67273
## factor(combinedSet$month)Oct  15.5075    16.5831   0.935  0.35036
## factor(combinedSet$month)Nov   7.3590    14.5273   0.507  0.61278
## factor(combinedSet$month)Dec  40.0229    14.2627   2.806  0.00529 **
## combinedSet$temp               2.5385     0.3366   7.541    4e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 55.63 on 352 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.6012, Adjusted R-squared:  0.5876
## F-statistic: 44.21 on 12 and 352 DF,  p-value: < 2.2e-16
```

This model explains slightly more of the crime than the last one, at 0.5875592.

Lastly, I created a model using weekday, month, and temperature as predictors for crime committed per day.

```
summary(fit6)
```

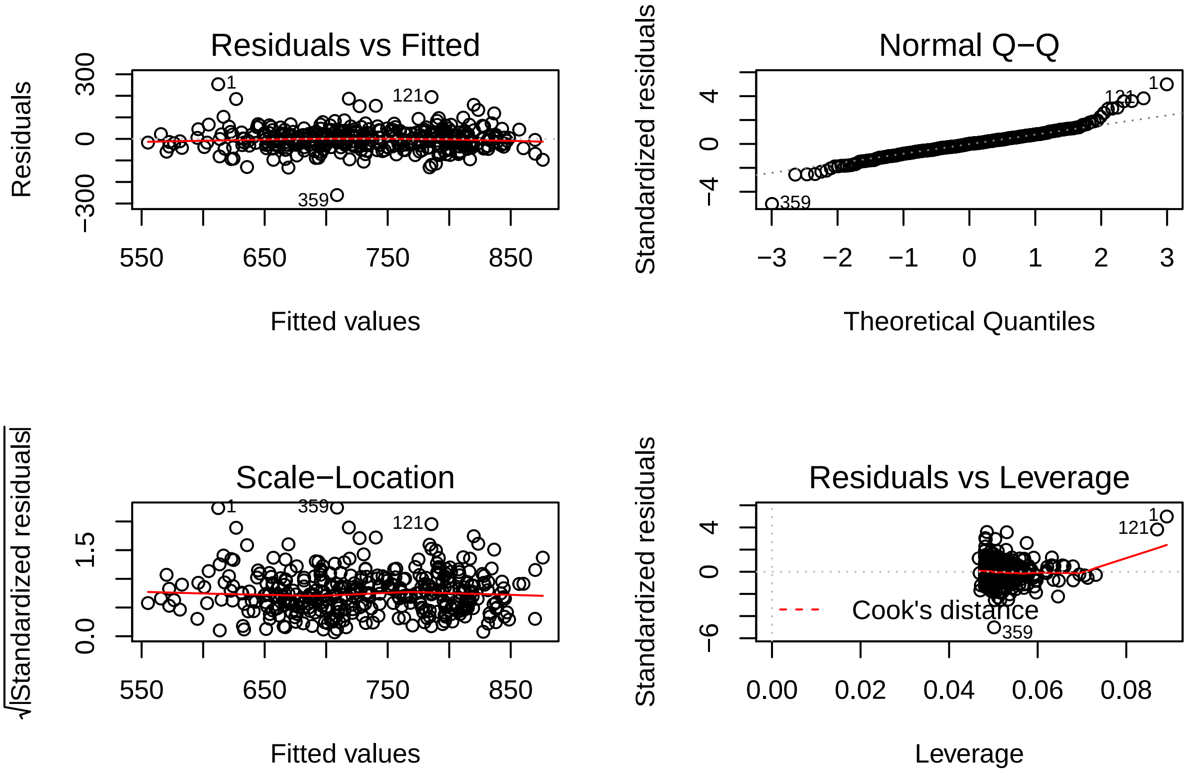
```
##
## Call:
## lm(formula = combinedSet$n ~ factor(combinedSet$weekday) + factor(combinedSet$month) +
##     combinedSet$temp)
```

```
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -260.810  -28.842    0.311   27.198  253.748
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     574.1854    14.7861  38.833  < 2e-16 ***
## factor(combinedSet$weekday)Mon   -7.2864    10.4175  -0.699 0.484749
## factor(combinedSet$weekday)Tue    4.2108    10.4193   0.404 0.686362
## factor(combinedSet$weekday)Wed   -1.4724    10.4228  -0.141 0.887740
## factor(combinedSet$weekday)Thu   37.3836    10.4242   3.586 0.000384 ***
## factor(combinedSet$weekday)Fri   24.2948    10.4257   2.330 0.020365 *
## factor(combinedSet$weekday)Sat -15.5024    10.4123  -1.489 0.137434
## factor(combinedSet$month)Feb   -43.1315    13.9637  -3.089 0.002172 **
## factor(combinedSet$month)Mar     4.0458    13.8112   0.293 0.769744
## factor(combinedSet$month)Apr    23.9887    14.3202   1.675 0.094805 .
## factor(combinedSet$month)May    53.6648    17.2769   3.106 0.002052 **
## factor(combinedSet$month)Jun    44.4849    18.9030   2.353 0.019165 *
## factor(combinedSet$month)Jul    36.1161    20.1179   1.795 0.073490 .
## factor(combinedSet$month)Aug    37.2042    20.5732   1.808 0.071415 .
## factor(combinedSet$month)Sep     4.2982    18.9215   0.227 0.820433
## factor(combinedSet$month)Oct    12.8146    15.8986   0.806 0.420785
## factor(combinedSet$month)Nov     4.0961    13.9378   0.294 0.769020
## factor(combinedSet$month)Dec    39.7734    13.6854   2.906 0.003893 **
## combinedSet$temp                 2.6430     0.3232   8.177 5.57e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.3 on 346 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:   0.64,  Adjusted R-squared:  0.6213
## F-statistic: 34.18 on 18 and 346 DF,  p-value: < 2.2e-16
```

This model has the strongest adjusted r squared of my models, at 0.6213228. meaning that the predictors weekday and temperature explains 0.6213228 of variability in crimes committed, suggesting that this model is good. My adjusted r squared increased as I added the variables of weekday month, and temperature together.

Diagnostics

```
par(mfrow=c(2,2))
plot(fit6)
```



Graph 1: Residuals vs. fitted

My residuals look random and there is no non-linear pattern present. Additionally, the slope of the fitted line of residuals is pretty flat. Both of these indicate independence and a model that fits well.

Graph 2: Normal Q-Q

This plot indicates my residuals are normally distributed as they are mostly all tight on the dashed line. This plot indicates a few potential outliers, such as observations 40, 15, and 1 but overall the fit is good and indicated a strong model.

Graph 3:Scale Location

The fit line has a flat slope and points look randomly scattered, indicating homoskedasticity. This means that variance is random and residuals are spread equally along predictors. This plot again indicates outliers 1,15, and 40.

Graph 4:Residuals vs. Leverage

This graph indicates influential points. None of my points look overly influential, with the exception of observation 1 in the upper right corner. While not very influential, observations 15 and 40 have slightly larger residuals.

Section V: Hypothesis Testing

Days of the week

H_0- mu$_{\text{Monday}}$=mu$_{\text{Friday}}$

The null hypothesis is that the average amount of crimes on Monday and Friday are equal.

H_a- mu$_{\text{Monday}}$???mu$_{\text{Friday}}$

The alternative hypothesis is that the amounts of crimes on Monday and Friday are not equal.

I conducted a t test on whether or not the true mean of counts over the two days of week were equal.

```
countMonday<-crimeCount$n[crimeCount$weekday=="Mon"]
countFriday<-crimeCount$n[crimeCount$weekday=="Fri"]
t.test(countMonday,countFriday,alternative="two.sided", conf.level = 0.95)
```

```
##
##  Welch Two Sample t-test
##
## data:  countMonday and countFriday
## t = -1.6534, df = 102, p-value = 0.1013
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -63.408784   5.754938
## sample estimates:
## mean of x mean of y
##  719.5577  748.3846
```

Based on how small my p-value from the test is, we can reject the null and say with 95% confidence that there is a difference in the true means of crime counts between Mondays and Fridays. This is additionally confirmed based on the 95% confidence interval not including 0 (no difference in means).

Section VI: Conclusions

Overall, this project was a very interesting experience for me. I got to face firsthand some of the difficulties of working with real world data and also got to learn about and understand my city a little bit better. Most of my challenges throughout this project were related to figuring out how to make my data work with what I wanted to do. I had to make many adjustments to get my two data sources to match in format so that I could work with them together.

It may have been interesting to separate crimes into different categories and analyze predictor variables this way. For example, nonviolent crimes such as burglary or domestic incidents may be more routine while violent crimes such as assault and murder could be spur of the moment or caused by other underlying variables. Additionally, it may have been useful to try to find lurking variables such as whether or not school is in session. This could be lurking because school is out in the summer, which could increase crime while making it seem like the relationship between temperature and crime was higher than it was.

I wasn't able to find a perfect model to describe the occurrence of crime in Chicago, but throughout trying different models I was able to see which variables had more and less correlation with crime and build a model that was pretty good. My final model was a multiple regression of crime with predictors weekday, month and temperature. The adjusted r squared of this model was 0.4768, and 9 out of 11 predictors were significant based on the p-values. It was very interesting to explore the relationships of these variables in a context that was so applicable to my life.

Pledge

On my honor, I have neither given nor recieved any unauthorized assistance on this work.

Victoria Seliger