Introduction

The City of Chicago has been focusing on using criminal reports to forecast criminal activity that allows for a better deployment of police resources. Hence, the goal of the following report is to answer the following questions:

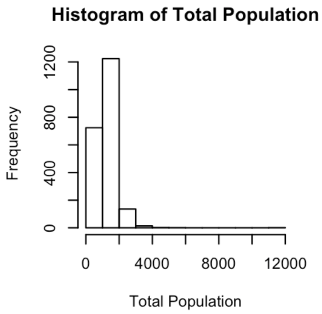
1. Do narcotic-related crimes depend on demographic and geographic factors?
2. Moreover, are cannabis and non-cannabis related crimes correlated?

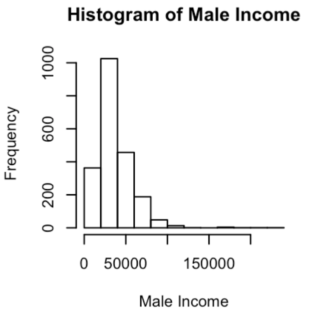
Exploratory Data Analysis

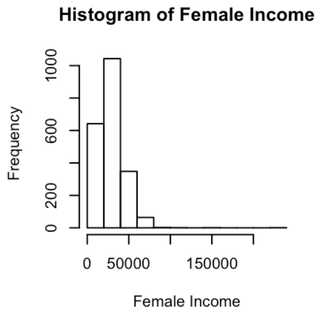
The data set being used consists of a sample of 2,102 observations and 14 variables. A new variable, *CrimeTotal*, which sums the count of narcotic-related crimes per block, has been added to the dataset. The univariate EDA of each variable is summarized below. We will start with the continuous variables:

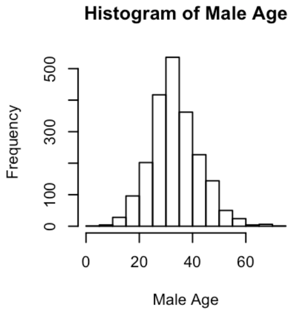
	Minimum	**1st Quartile**	**Median**	**Mean**	**3rd Quartile**	**Maximum**
PopTotal	56.0	892.2	1179.5	1255.1	1494.8	11309.0
IncomeMale	2499	22376	31392	36426	45404	238512
IncomeFemale	2499	18481	25794	28666	35943	235167
AgeMale	4.10	27.70	32.30	33.41	38.90	74.30
AgeFemale	13.80	29.60	34.50	36.17	41.50	82.60
Latitude	41.65	41.78	41.87	41.86	41.94	42.02
Longitude	-87.94	-87.72	-87.68	-87.68	-87.64	-87.53
CrimeC	4.00	15.00	18.00	19.59	23.00	61.00
CrimeNC	1.00	10.00	14.00	16.55	21.00	104.00
pctWhite	0.00	0.02639	0.47150	0.42212	0.71059	0.98491
pctAsian	0.00	0.001302	0.013173	0.047647	0.057291	0.938627
pctBlack	0.00	0.02411	0.10390	0.37846	0.93669	0.99844
CrimeTotal	13.00	27.00	34.00	36.15	43.00	144.00
Ward	1.00	13.00	25.00	25.09	38.00	50.00
Zone	0.00	0.00	1.00	0.5604	1.00	1.00

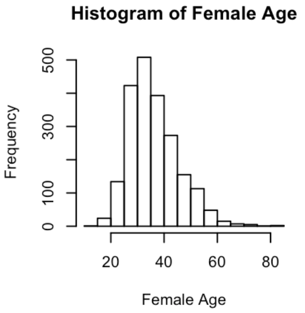
Table 1: Univariate EDA of Continuous Variables

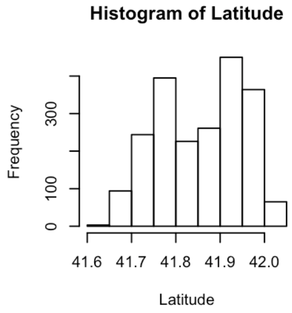

Histogram of Total Population

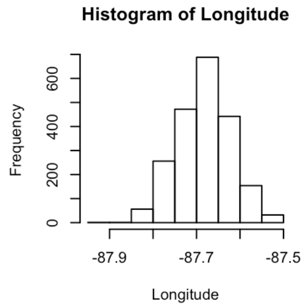

Histogram of Male Income

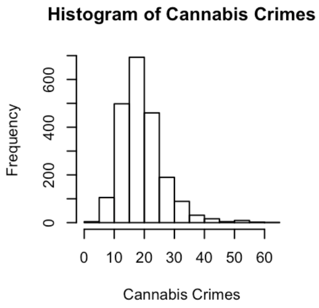

Histogram of Female Income

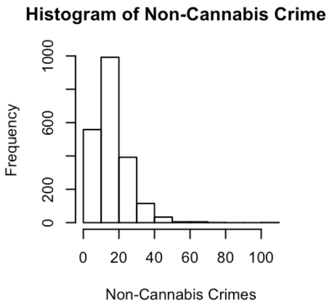

Histogram of Male Age

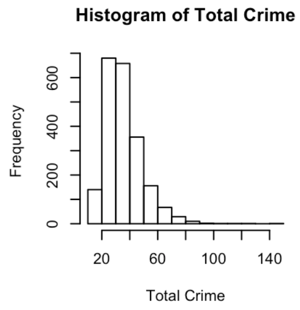

Histogram of Female Age

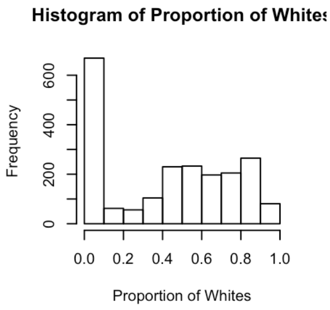

Histogram of Latitude

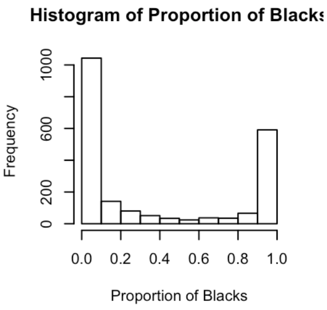

Histogram of Longitude


Histogram of Cannabis Crimes


Histogram of Non-Cannabis Crime


Histogram of Total Crime


Histogram of Proportion of Whites


Histogram of Proportion of Blacks

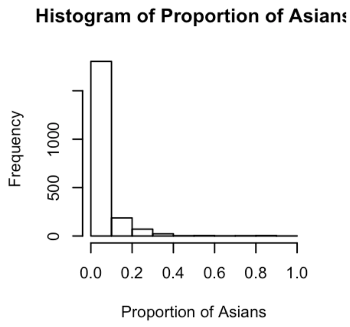

Histogram of Proportion of Asians

Figure 1: Histogram of Continuous Variables

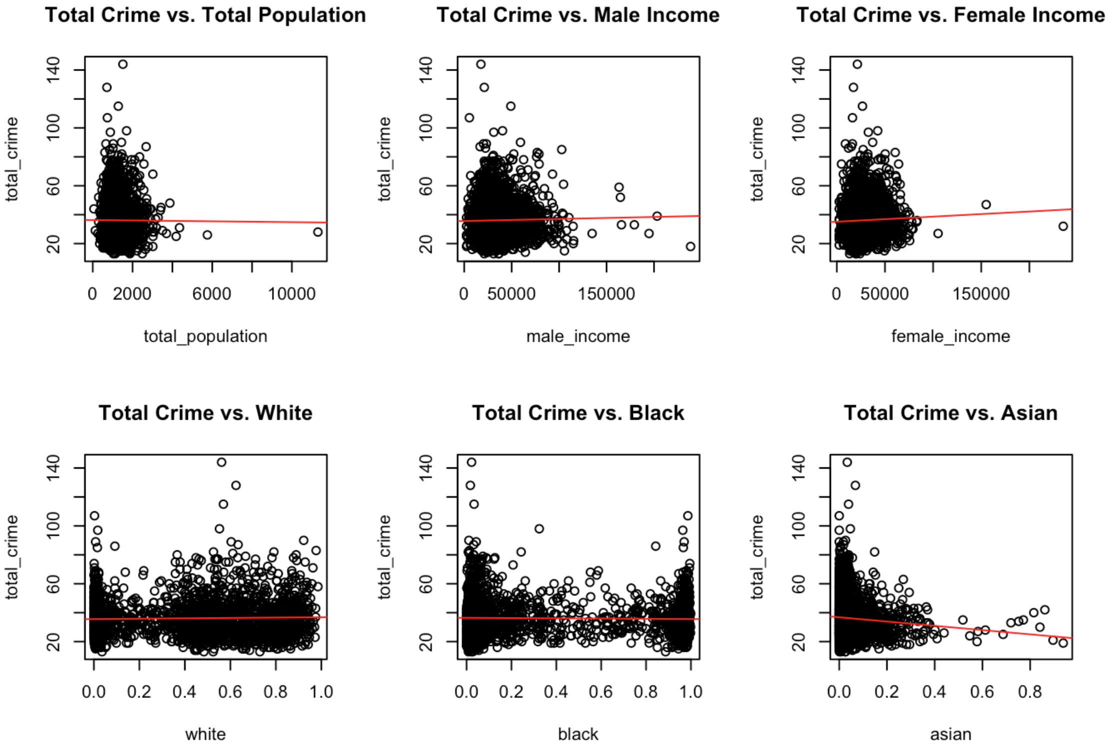
Some of the variables are right-skewed, so it is worthwhile to consider transformations. This point will be touched upon again in the future.

Next, we will move on to the categorical variable:

Zone	**Count**	**Percentage**
0	1,178	56.0%
1	924	44.0%

Table 2: Univariate EDA of Zone Variable

The multivariate EDA of each variable paired to the CrimeTotal variable is summarized below:



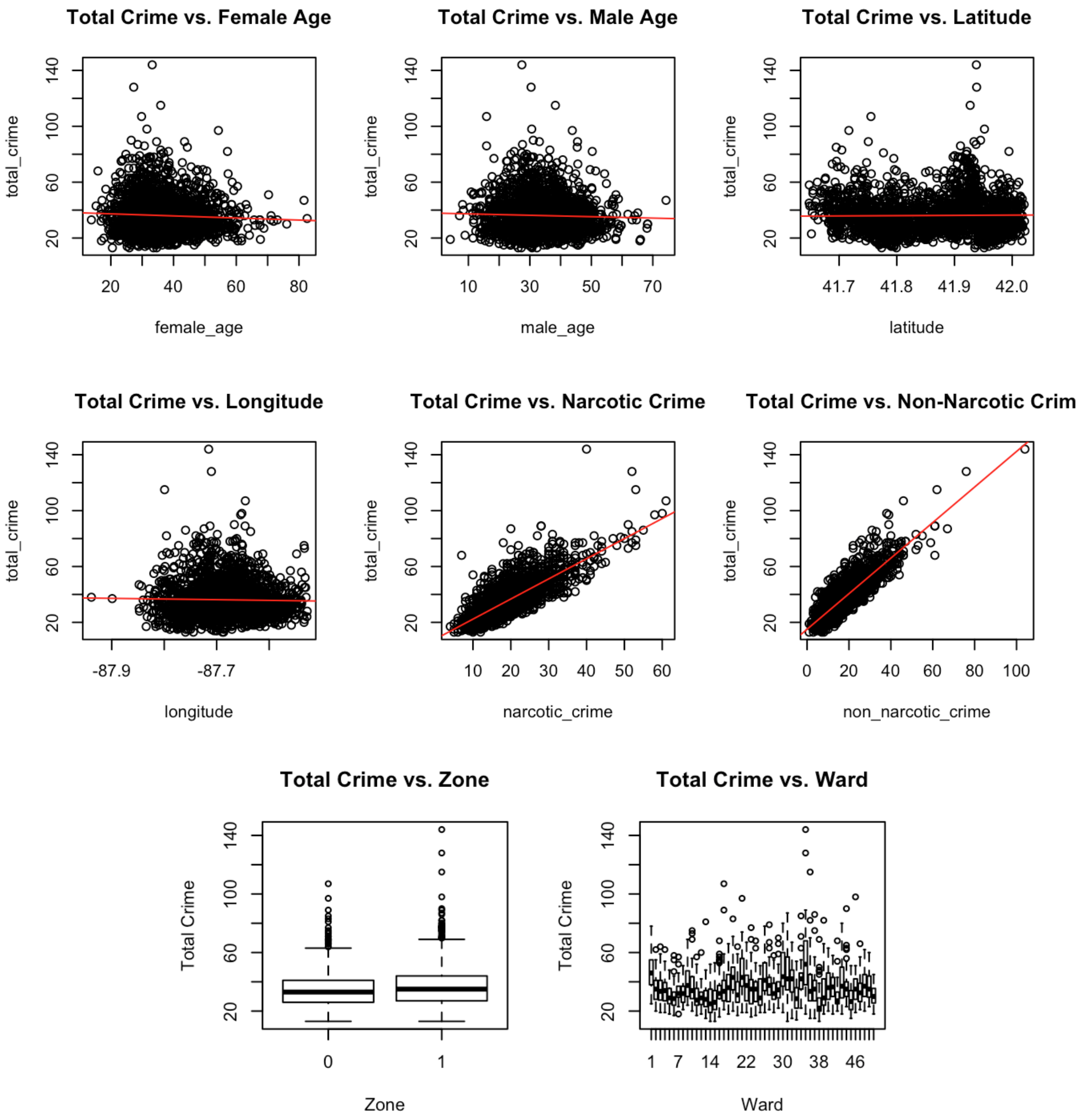


Figure 2: Multivariate EDA of All Variables and Response Variable

There is generally no correlation between *CrimeTotal* and the variables, besides with the age, income, and *pctAsian* variables. Even so, the correlations are quite weak.

The multivariate EDA of each variable paired to the CrimeC and CrimeNC variables are not shown in this report, since they are similar to the CrimeTotal variable.

For each type of crime, the top 5% highest and the bottom 5% lowest crime rate blocks are shown below.

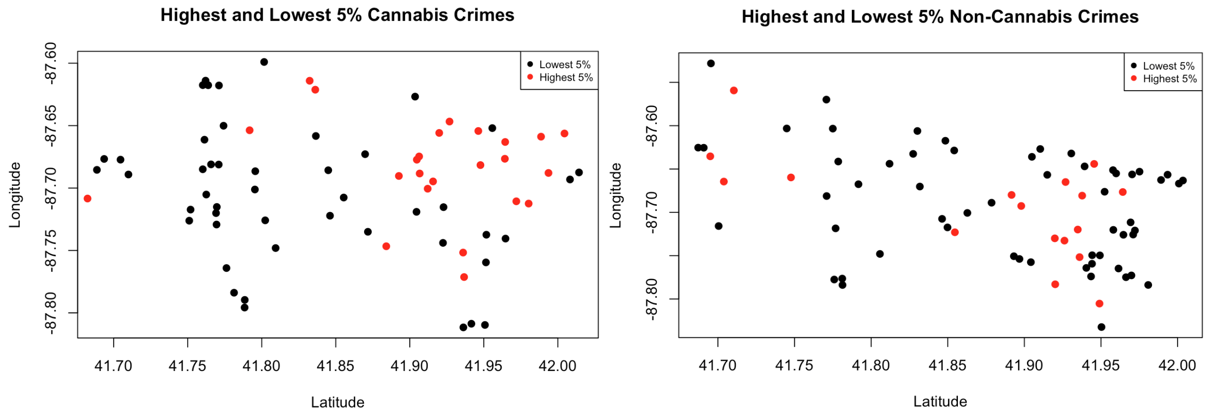


Figure 3: Highest and Lowest 5% Narcotic Crimes

The geographical pattern for high crime rates for both types of crime roughly lie in similar latitudes and longitudes of 41.90 and -87.75, respectively. On the other hand, the low crime rates for both types of crime are more scattered and appears to be random.

Initial Modeling & Diagnostics

Multivariate EDA and the LASSO regression were performed to determine the most suitable variables to include in the two candidate multiple linear regression models. The motivation behind using the LASSO as opposed to other feature selection methods, such as stepwise regression, is to avoid applying methods intended for one test to many tests.

Further assumptions to note throughout this report are that:
(1) *Zone* and *Ward* are discrete variables of categorical nature, so they are converted into factors.
(2) All other variables remain continuous.
(3) *CrimeC* and *CrimeNC* were eliminated from the model to avoid multicollinearity.

The LASSO regression gave the following model:

$$\textbf{\textit{CrimeTotal}} = \beta_0 + \beta_1 \textbf{\textit{IncomeFemale}} + \beta_2 \textbf{\textit{AgeMale}} + \beta_3 \textbf{\textit{AgeFemale}} + \beta_4 \textbf{\textit{Ward(level=2}} + \ldots + \beta_{53} \textbf{\textit{Ward(level=49)}} + \beta_{54} \textbf{\textit{pctBlack}} + \beta_{55} \textbf{\textit{pctAsian}} + \beta_{56} \textbf{\textit{Zone(level=1)}} + \varepsilon_i$$

Based on the multivariate EDA, the variables that exhibited correlation values of above 0.5 were chosen to be included in the multiple linear regression model as produced:

$$\textbf{\textit{CrimeTotal}} = \beta_0 + \beta_1 \textbf{\textit{IncomeMale}} + \beta_2 \textbf{\textit{IncomeFemale}} + \beta_3 \textbf{\textit{AgeMale}} + \beta_4 \textbf{\textit{AgeFemale}} + \beta_5 \textbf{\textit{Ward(level=2)}} + \ldots + \beta_{54} \textbf{\textit{Ward(level=49)}} + \beta_{55} \textbf{\textit{Latitude}} + \beta_{56} \textbf{\textit{pctWhite}} +$$

$$\beta_{57}\boldsymbol{pctBlack} + \beta_{58}\boldsymbol{Zone(level{=}1)} + \varepsilon_i$$

To determine which of the two models outperforms the other, we look at the prediction error and bootstrap to measure the uncertainty in the mean-squared-errors. First, we perform 5-fold cross-validation and compute the average of the error values to determine the optimal model. The prediction errors are shown below:

Model	Prediction Error
EDA	141.70
LASSO	141.86

Table 3: Prediction Error of Models

Although the prediction error for the EDA model appears to be lower, the two values are very close. We turn to bootstrapping to measure the uncertainty and find that the LASSO model performs better than the EDA model, since every difference is positive at approximately 0.38.

The diagnostics for the LASSO model are shown below:

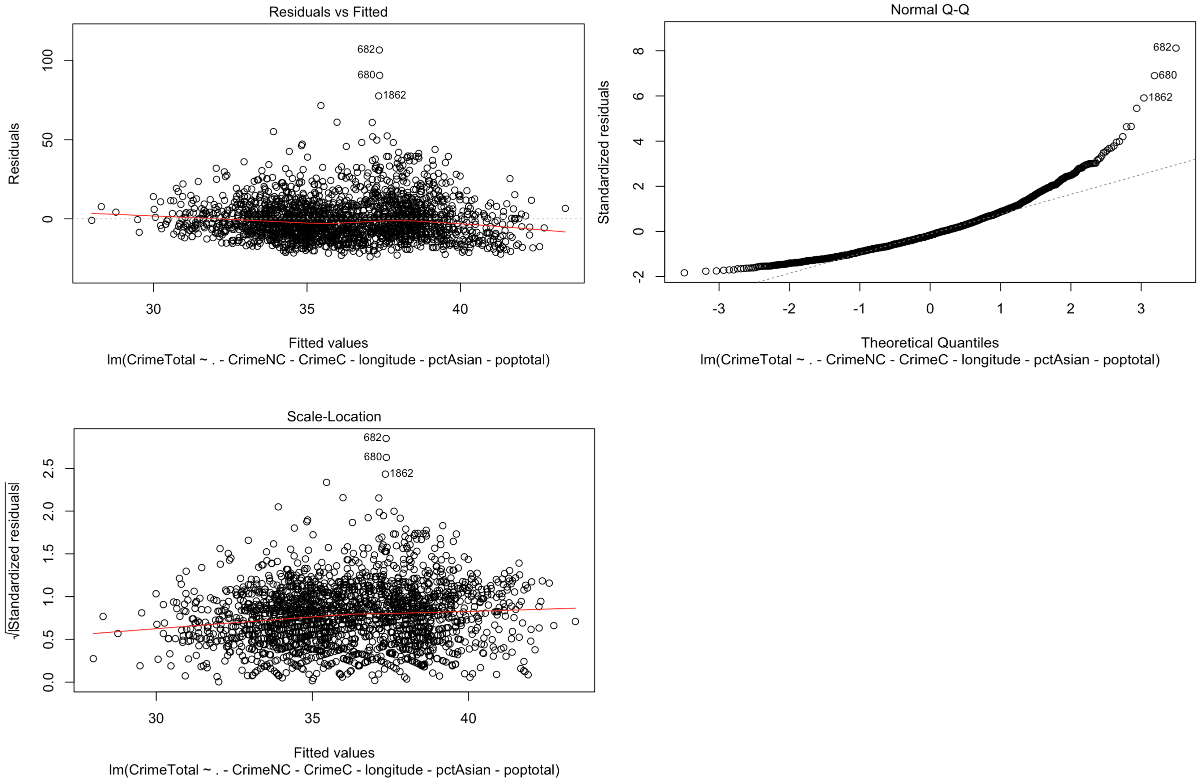


Figure 4: Diagnostic Plots of LASSO Model

The residuals vs. fitted values plot shows a generally equally spread of residuals around the horizontal line without distinct patterns, suggesting that there are no non-linear relationships. The assumption of linearity is upheld.

The Q-Q plot shows the residuals are lined well on the bottom left end, but gradually go out of place as it increases. In general, the residuals do not deviate severely and hence suggests the residuals are normally distributed.

The Scale-Location plot shows the residuals are spread equally and randomly along the ranges of the predictors. The assumption of homoscedasticity is upheld.

Basic transformation of variables, i.e. log, sqrt, sq, were attempted to determine possible improvements in the model, but no improvements were apparently. Nevertheless, all model assumptions were upheld.

Results

To test whether or not there seems to be a relationship between total crime rate and geographic and demographic variables, we fit the following full model with all variables:

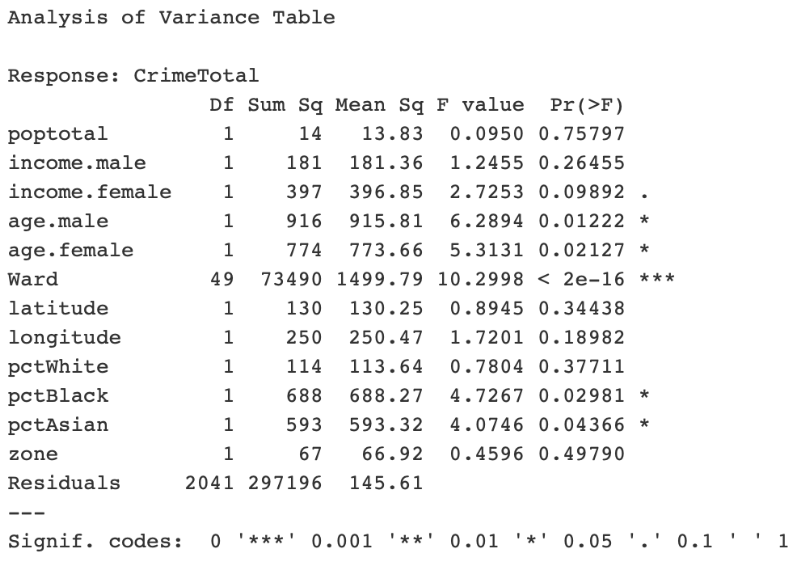
$$\boldsymbol{CrimeTotal} = \beta_0 + \beta_1\boldsymbol{PopTotal} + \beta_2\boldsymbol{AgeMale} + \beta_3\boldsymbol{AgeFemale} + \beta_4\boldsymbol{IncomeMale} + \beta_5\boldsymbol{IncomeFemale} + \beta_6\boldsymbol{Ward(level=2)} + \ldots + \beta_{55}\boldsymbol{Ward(level=49)} + \beta_{56}\boldsymbol{pctBlack} + \beta_{57}\boldsymbol{pctAsian} + \beta_{58}\boldsymbol{pctWhite} + \beta_{59}\boldsymbol{Latitude} + \beta_{60}\boldsymbol{Longitude} + \beta_{61}\boldsymbol{Zone(level=1)} + \varepsilon_i$$

We state the null as there being no relationship between total crime rate and geographic and demographic variables. In other words:

$$H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

$$H_1: \beta_j \neq 0 \text{ for at least one } j, j=1...p$$

The F-test for overall significance in the ANOVA table is shown below for the full model:



```
Analysis of Variance Table

Response: CrimeTotal
                Df Sum Sq Mean Sq F value  Pr(>F)    
poptotal         1     14   13.83  0.0950 0.75797    
income.male      1    181  181.36  1.2455 0.26455    
income.female    1    397  396.85  2.7253 0.09892 .  
age.male         1    916  915.81  6.2894 0.01222 *  
age.female       1    774  773.66  5.3131 0.02127 *  
Ward            49  73490 1499.79 10.2998 < 2e-16 ***
latitude         1    130  130.25  0.8945 0.34438    
longitude        1    250  250.47  1.7201 0.18982    
pctWhite         1    114  113.64  0.7804 0.37711    
pctBlack         1    688  688.27  4.7267 0.02981 *  
pctAsian         1    593  593.32  4.0746 0.04366 *  
zone             1     67   66.92  0.4596 0.49790    
Residuals     2041 297196  145.61                    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

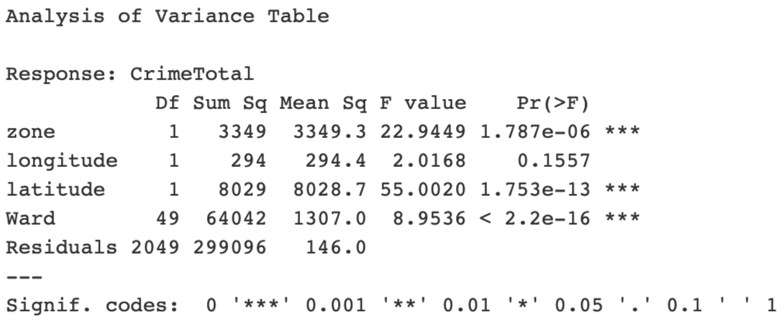
Table 4: ANOVA Table of Full Model

The predictors *AgeFemale*, *AgeMale*, *Ward*, *pctBlack*, *pctAsian* have p-values lower than the significance value of α = 0.05, implying that they are statistically significant in this model.

An F-test was conducted on the following reduced model with **only** geographic variables:

$$\boldsymbol{CrimeTotal} = \beta_0 + \beta_6\boldsymbol{Ward(level{=}2)} + \ldots + \beta_{55}\boldsymbol{Ward(level{=}49)} + \beta_{59}\boldsymbol{Latitude} + \beta_{60}\boldsymbol{Longitude} + \beta_{61}\boldsymbol{Zone(level{=}1)} + \varepsilon_i$$

The corresponding ANOVA table is displayed below:



```
Analysis of Variance Table

Response: CrimeTotal
            Df Sum Sq Mean Sq F value    Pr(>F)    
zone         1   3349  3349.3 22.9449 1.787e-06 ***
longitude    1    294   294.4  2.0168    0.1557    
latitude     1   8029  8028.7 55.0020 1.753e-13 ***
Ward        49  64042  1307.0  8.9536 < 2.2e-16 ***
Residuals 2049 299096   146.0                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

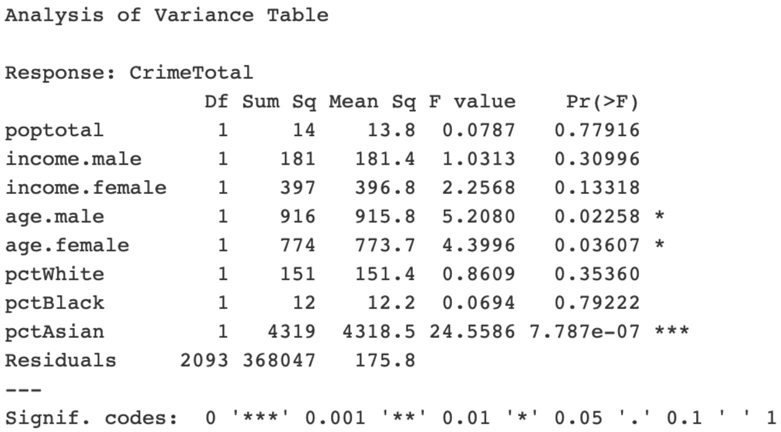
Table 5: ANOVA Table of Reduced Model with Geographic Variables

The predictors *Zone*, *Latitude*, *Ward* have p-values lower than the significance value of α = 0.05, implying that they are statistically significant in this model.

An F-test was conducted on the following reduced model with **only** demographic variables:

$$\boldsymbol{CrimeTotal} = \beta_0 + \beta_1\boldsymbol{PopTotal} + \beta_2\boldsymbol{AgeMale} + \beta_3\boldsymbol{AgeFemale} + \beta_4\boldsymbol{IncomeMale} + \beta_5\boldsymbol{IncomeFemale} + \beta_{56}\boldsymbol{pctBlack} + \beta_{57}\boldsymbol{pctAsian} + \beta_{58}\boldsymbol{pctWhite} + \varepsilon_i$$

The corresponding ANOVA table is displayed below:



```
Analysis of Variance Table

Response: CrimeTotal
                Df Sum Sq Mean Sq F value    Pr(>F)    
poptotal         1     14    13.8  0.0787   0.77916    
income.male      1    181   181.4  1.0313   0.30996    
income.female    1    397   396.8  2.2568   0.13318    
age.male         1    916   915.8  5.2080   0.02258 *  
age.female       1    774   773.7  4.3996   0.03607 *  
pctWhite         1    151   151.4  0.8609   0.35360    
pctBlack         1     12    12.2  0.0694   0.79222    
pctAsian         1   4319  4318.5 24.5586 7.787e-07 ***
Residuals     2093 368047   175.8                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 6: ANOVA Table of Reduced Model with Demographic Variables

The predictors *AgeMale* and *AgeFemale* have p-values lower than the significance value of α = 0.05, implying that they are statistically significant in this model. From the analysis, we reject the null hypothesis, implying that at least one of the regressors contributes significantly to the model.

However, solely relying on the p-value is not valid to determine the absolute statistical significance of the variables. Some variables may be correlated and their coefficients and *t* statistics can change depending on which other variables are included in the regression. Further analysis would be needed to fully establish which specific variables affect the total crime rate.

Next, we analyze the effect of being above or below the river on cannabis-related crime counts versus non-cannabis related crime counts. We construct the model with the Zone variable as well as the AgeFemale, AgeMale, Ward, pctBlack, pctAsian variables deemed significant from part (9):

$$\boldsymbol{CrimeC} = \beta_0 + \beta_2\boldsymbol{AgeMale} + \beta_3\boldsymbol{AgeFemale} + \beta_6\boldsymbol{Ward(level{=}2)} + \ldots + \beta_{55}\boldsymbol{Ward(level{=}49)} + \beta_{56}\boldsymbol{pctBlack} + \beta_{57}\boldsymbol{pctAsian} + \beta_{61}\boldsymbol{Zone(level{=}1)} + \varepsilon_i$$

$$\boldsymbol{CrimeNC} = \beta_0 + \beta_2\boldsymbol{AgeMale} + \beta_3\boldsymbol{AgeFemale} + \beta_6\boldsymbol{Ward(level{=}2)} + \ldots + \beta_{55}\boldsymbol{Ward(level{=}49)} + \beta_{56}\boldsymbol{pctBlack} + \beta_{57}\boldsymbol{pctAsian} + \beta_{61}\boldsymbol{Zone(level{=}1)} + \varepsilon_i$$

We state the null as the location above or below the river having no effect on the cannabis and non-cannabis related crime rates. In other words:

$$H_0: \beta_2 = \beta_3 = \beta_6 = \ldots = \beta_{55} = \beta_{56} = \beta_{57} = \beta_{61} = 0$$
$$H_1: \beta_j \neq 0 \text{ for at least one } j,\ j = 2, 3, 6, \ldots, 55, 56, 57, 61$$

The regression coefficient and p-values for the *Zone(level=1)* variable for both models are shown below:

Type of Crime	Regression Coefficient	p-value
Cannabis (CrimeC)	1.8267	5.97e-05
Non-Cannabis (CrimeNC)	6.3636	7.14e-14

Table 7: Regression Coefficients of CrimeC and CrimeNC Regressed on Zone, Latitude, and Longitude

The estimated intercept is 1.8267 higher among zone 1 compared to zone 0 for cannabis crimes, and the estimated intercept is 6.3636 higher for zone 1 than zone 0 for non-cannabis related crimes. In other words, holding all other variables constant, the cannabis related crimes are 1.8267 higher in zone 1 compared to zone 0, and the non-cannabis related crimes are 6.3636 higher in zone 1 compared to zone 0. In the context of this problem, cannabis and non-cannabis related crimes are higher on the north of the river.

A p-value based on the z-score was extracted from the coefficients, giving a value of 0.76033. This number is below the significance value of α = 0.05, suggesting that the location above or below the river affects the cannabis related versus non-cannabis related crime counts in each block.

A new data set with total crime per ward and total population per ward as variables was created. The top 5 wards with the highest total count of crime is shown below:

Ward	Total Crime
35	2,361
1	2,203
21	2,173
34	2,060
19	2,031

Table 8: Top 5 Wards With Highest Total Crime

Fitting a model with the count of crime per ward as the response and total population as a covariate reranks the wards. However, the ranks by total crime and the ranks by residuals are quite similar, shown in Table 9 with the top 5 wards with the highest residuals:

Ward	Residual
35	839.726
21	671.963
1	644.106
34	536.217
19	525.016

Table 9: Top 5 Wards With Highest Total Crime Residuals

This observation suggests that controlling for the total population does not have an influence on the wards with highest crime rates.

Correcting by population size in general is reasonable, because some data points could be skewed. For example, some wards may have significantly less inhabitants but majority of them commit crimes.

Lastly, to investigate whether or not there is a relationship between cannabis and non-cannabis related police reports in each block group, a new model with cannabis-related crime as the response and non-cannabis related crime as the predictor is constructed. The initial reduced model is as follows:

$$\boldsymbol{CrimeC} = \quad \beta_0 + \beta_1 \boldsymbol{CrimeNC} + \varepsilon_i$$

And the null and alternate hypothesis we want to test is as follows:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

The resulting output is shown in Table 10:

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	15.07729	0.30416	49.57	<2e-16

CrimeNC	0.27278	0.01611	16.93	<2e-16

Table 10: Regression Coefficient of CrimeC Regressed on CrimeNC in Initial Reduced Model

The p-value of the output is below < 2.2e-16, suggesting that we reject the null hypothesis and that there is a relationship between cannabis and non-cannabis related crimes.

The model diagnostics are shown below:

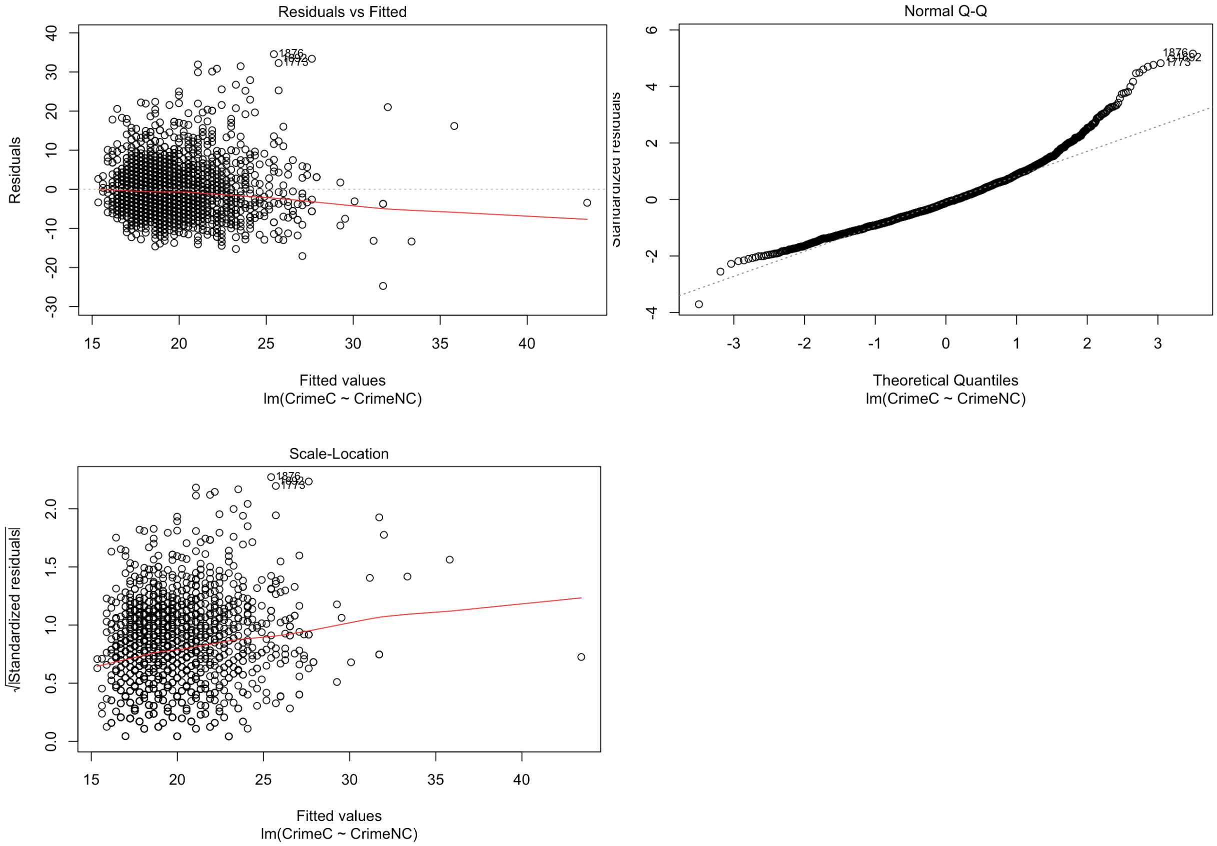


Figure 5: Diagnostic Plots of Initial Reduced Model

However, most notable in the diagnostics is the Q-Q plot that shows a scatter of points towards the top right that diverge from the line. The residuals vs. fitted plot also exhibits a minor negative relationship. The assumptions of linearity and normal distribution are violated.

In an attempt to improve the model and its assumptions, as well as controlling for the other variables, the final full model is constructed:

$$\boldsymbol{CrimeC} = \beta_0 + \beta_1 \boldsymbol{AgeMale} + \beta_2 \boldsymbol{AgeFemale} + \beta_3 \boldsymbol{log(IncomeMale)} +$$

$$\beta_4 \boldsymbol{log(IncomeFemale)} + \beta_5 \boldsymbol{log(PopTotal)} + \beta_6 \boldsymbol{pctAsian} +$$

$$\beta_7 \boldsymbol{pctWhite} + \beta_8 \boldsymbol{pctBlack} + \beta_9 \boldsymbol{Latitude} + \beta_{10} \boldsymbol{Longitude} +$$
$$\beta_{11} \boldsymbol{CrimeNC} + \beta_{12} \boldsymbol{Ward(level=2)} + \ldots + \beta_{61} \boldsymbol{Ward(level=49)} +$$
$$\beta_{62} \boldsymbol{Zone(level=1)} + \varepsilon_i$$

The *IncomeMale*, *IncomeFemale*, *PopTotal* variables are transformed with a *log()* function to improve the model. The summary of the model is displayed in Table 11:

Coefficients	Estimate	Std. Error	t value	Pr(>ltl)
(Intercept)	1.601e+03	9.167e+02	1.747	0.080826
CrimeNC	2.710e-01	1.802e-02	15.040	<2e-16

Table 11: Regression Coefficient of CrimeC Regressed on CrimeNC in Final Full Model

The model diagnostics are shown below:

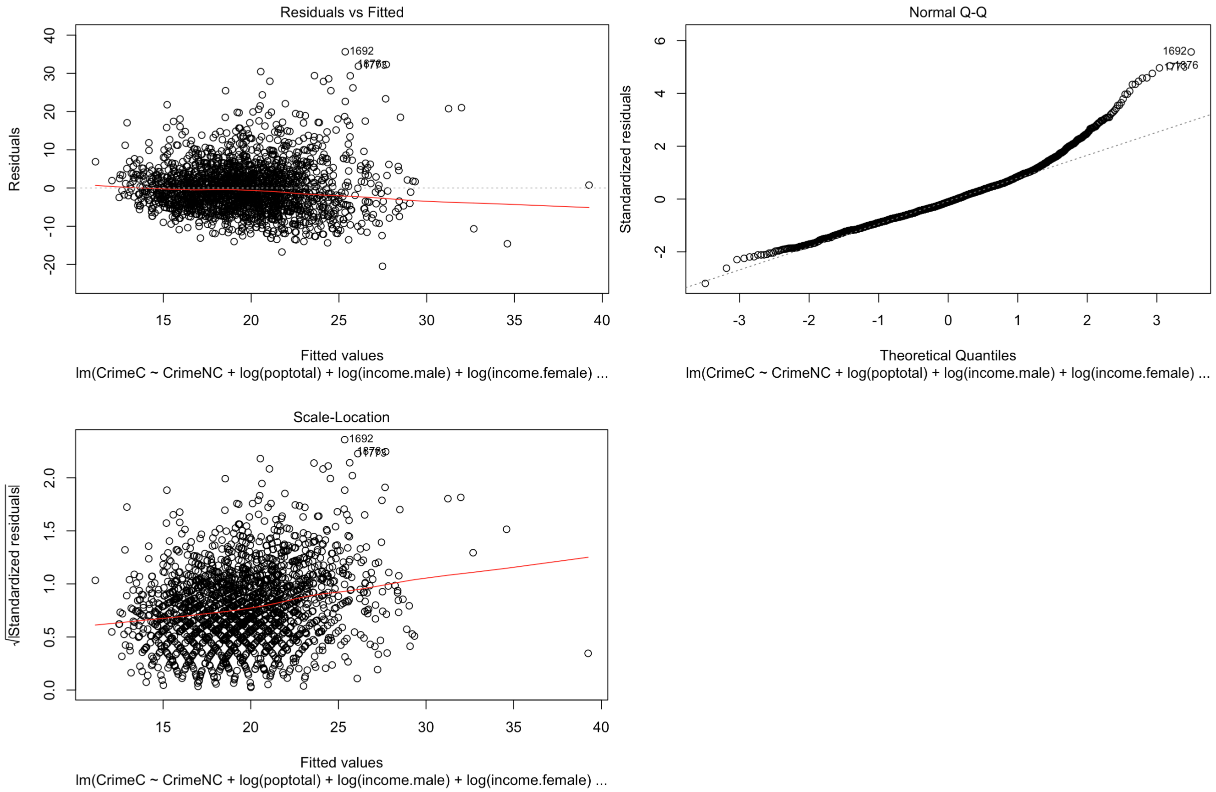


Figure 6: Diagnostic Plots of Final Full Model

The diagnostic plots show a slight improvement, though further enhancements can still be made, which will be discussed in the last section.

The regression coefficient of *CrimeNC* decreased from the initial model to the final model. This observation suggests that there is at least one other covariate that has predictive power affecting the cannabis-related crime rate.

Conclusion

With respect to the first and second research hypotheses, we may conclude there is strong evidence suggesting that there are variables that are statistically significant when determining the dependency of narcotic-related crimes on demographic and geographic factors.

In the context of this problem, there could be a trend in the age of each gender and ethnicity proportion in each block that correlates to the crime rates. Ward and Zone suggest that certain regions of Chicago has higher crime rates than others based on the geographic and political delineations. This trend may be due to the income disparities of each region, which is another variable to consider for further analysis.

With respect to the third research hypothesis, we may conclude there is evidence suggesting that cannabis and non-cannabis related crimes are related. The test concludes that CrimeNC and at least one other covariate has predictive power affecting the cannabis-related crime rate. However, EDA and variable selection were not performed when fitting the above model, suggesting that overfitting may be likely.

There are a vast selection of other variable and model selection methods that could be explored and potentially show a better fit than the ones performed in this report. Interaction terms and potential confounders could also be analyzed to ensure a stronger model.