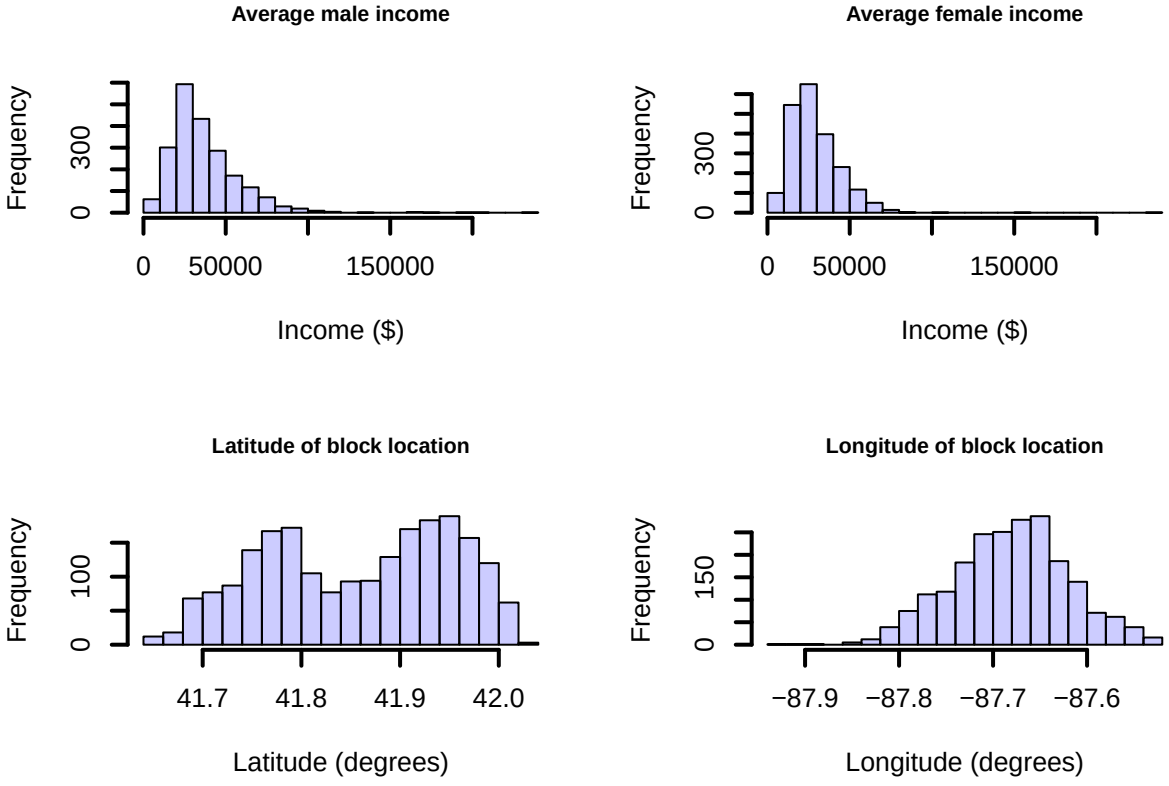
Chicago Crime Analysis

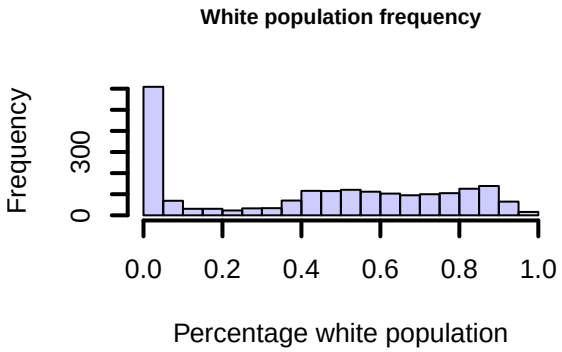
Riccardo Fogliato, Theresa Gebert, Wanshan Li

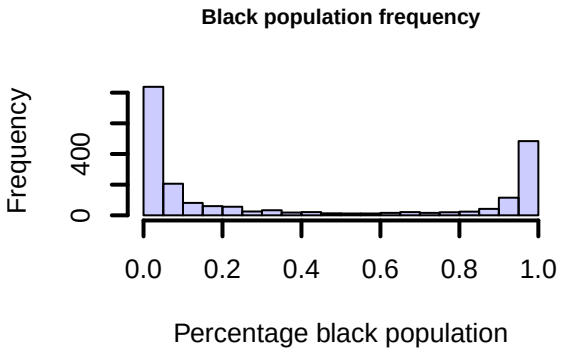
11/7/2017

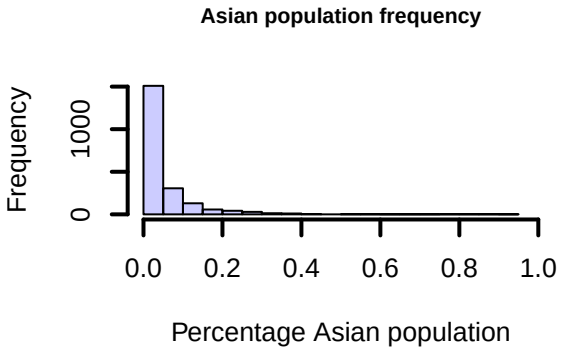
Introduction. A great deal of public policy in recent years has focused on using criminal reports in predictive policing, or forecasting criminal activity in a way that allows for a better deployment of police resources. The City of Chicago maintains a database of all criminal reports, extending from cases in which no arrest was made up to and including homicide. We have collected narcotics crime reports of two types, reports involving possession of marijuana and reports that did not, across a three-year time period and located them by Census block group. Within each block group, we also have information from the 2010 US Census and the 2011 American Community Survey that can provide additional information.

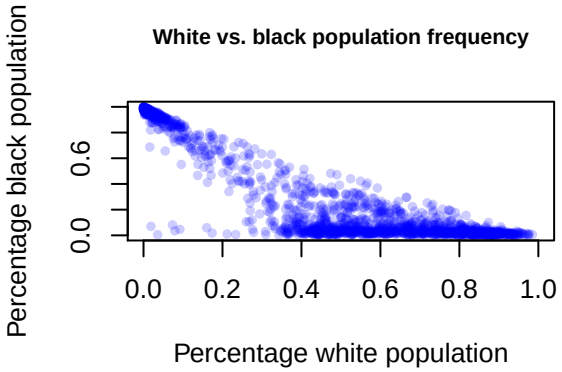
Exploratory Data Analyis. Here we graphically explore our dataset. As usual, income is right-skewed and lower-bounded by zero. Notice that, as expected, certain community areas are either black-dominant or white-dominant. It looks like only about 20-25% of communities are heterogeneous.



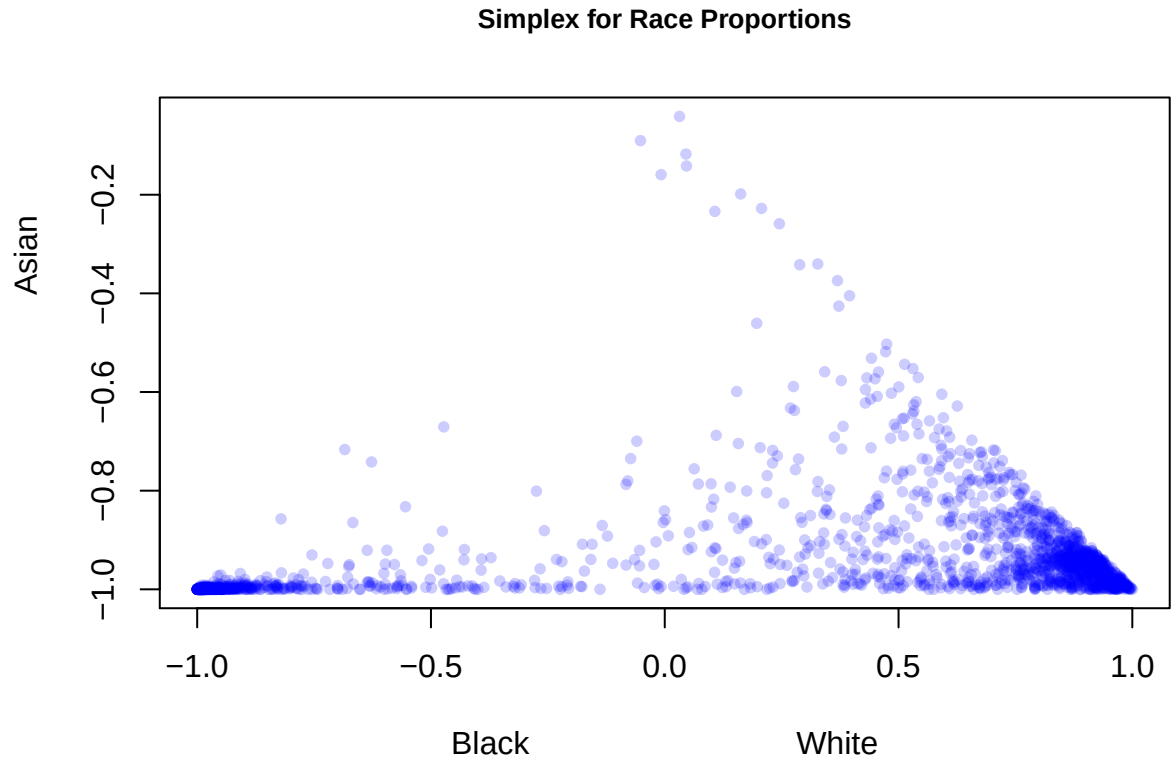






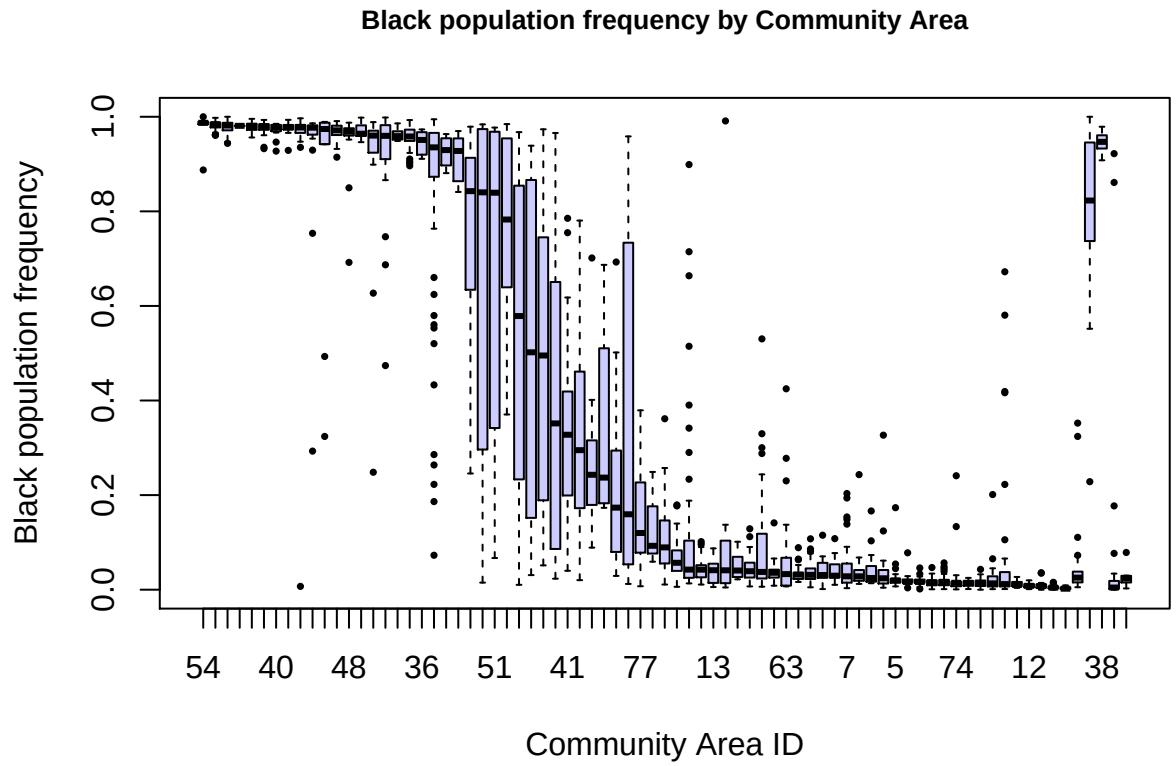


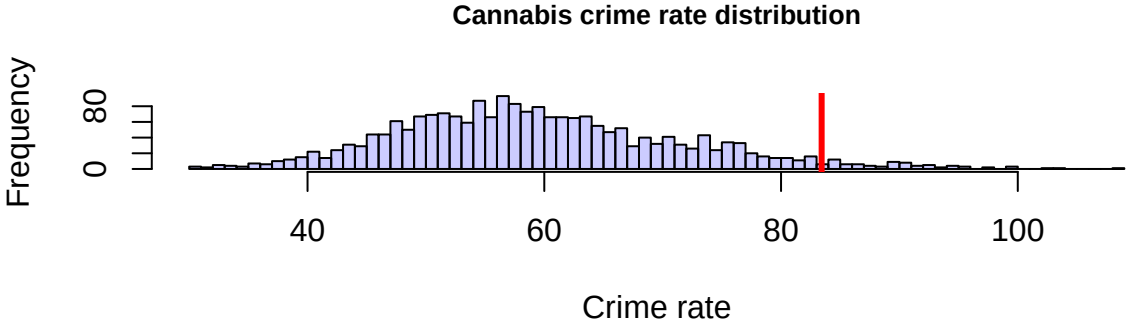
Related to population, we can not perform further analysis. The population is right skewed, but on average 15% of the entire population of the block has not been classified (we do not weight by block's population). However, it might be useful to draw the probability simplex of the three main ethnicities.

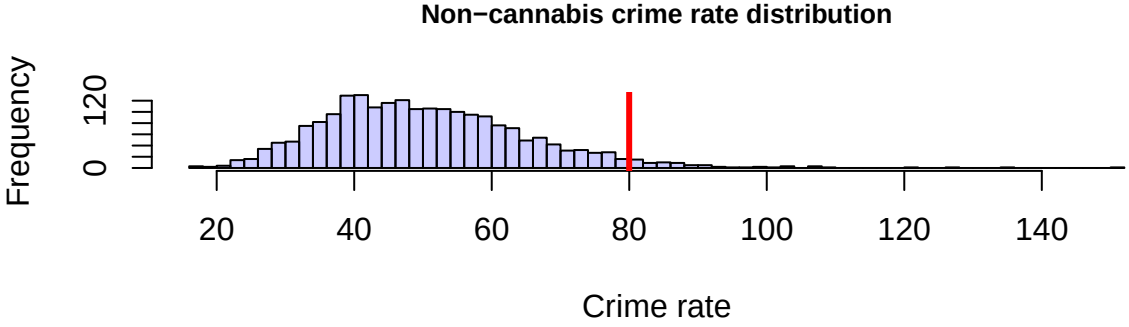


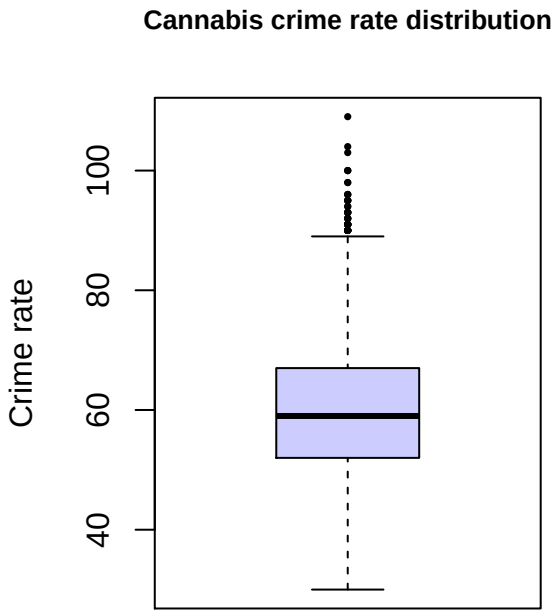
We notice that only a few blocks have predominantly Asian populations. Most blocks have predominantly

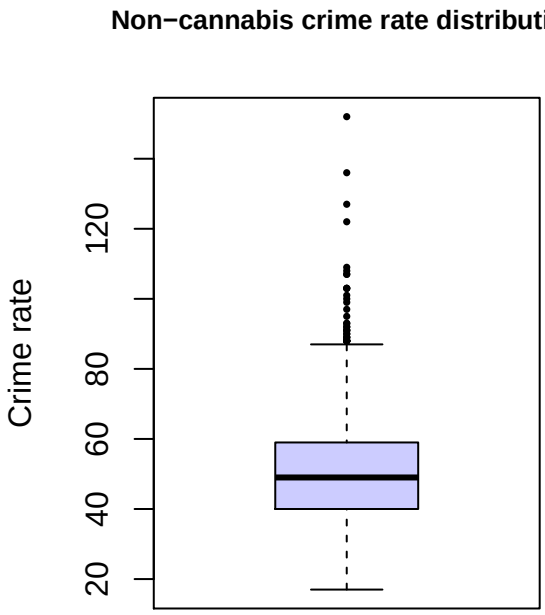
white populations. The blocks including a significant but non-dominant proportion of Asians appear to be primarily the white ones. However, this analysis supports a possible categorization of the three races into the dominant one at the block group level, if we account for the mixed white and black groups.







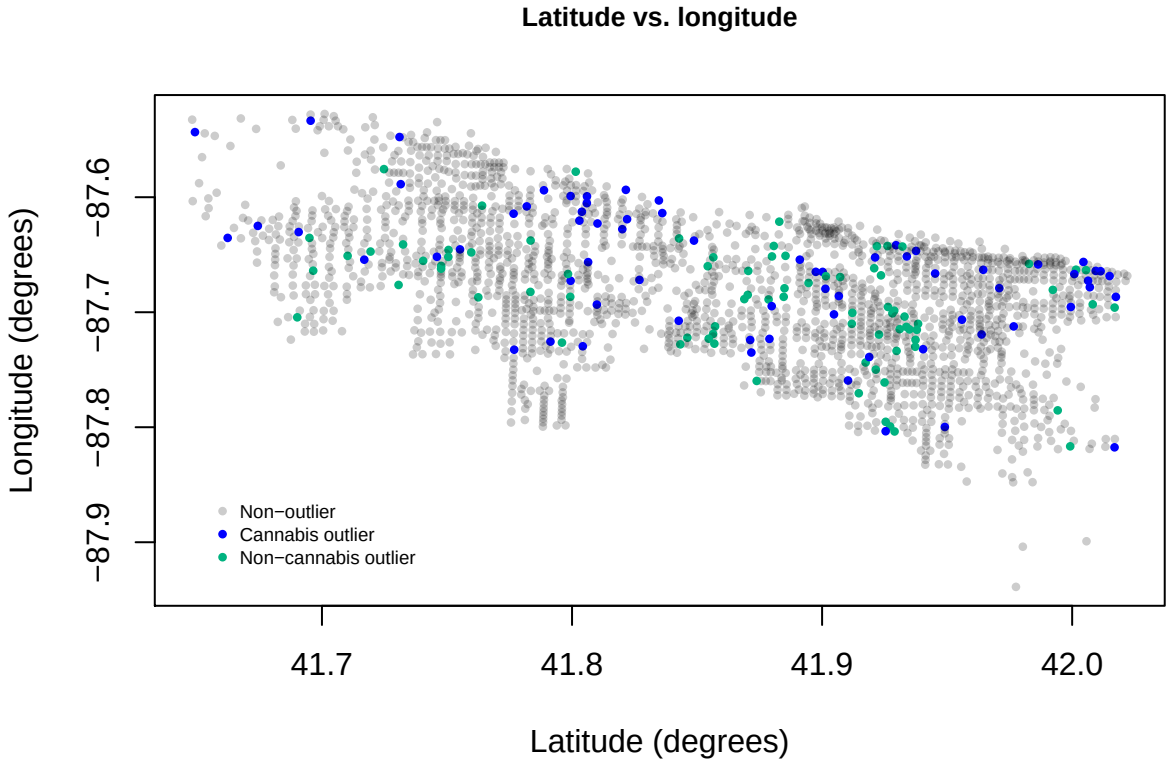




Substantive Questions.

(a) *Identify block groups that have significantly higher or lower crime rates for each of the two crime types than the mean. Are these groups spread evenly throughout Chicago, considering longitude and latitude?*

From the histograms of cannabis and non-cannabis crime rates, it appears reasonable to model crime rates as Normal distributions. In this case, we can use $p < 0.05$ as a heuristic for "significantly" different from the mean. It does not seem like any block groups are significantly below the mean.



This plot shows how the cannabis and non-cannabis outliers are distributed among the rest of the regions. There does not seem to be an obvious relationship between outlier status and geographical location: overall

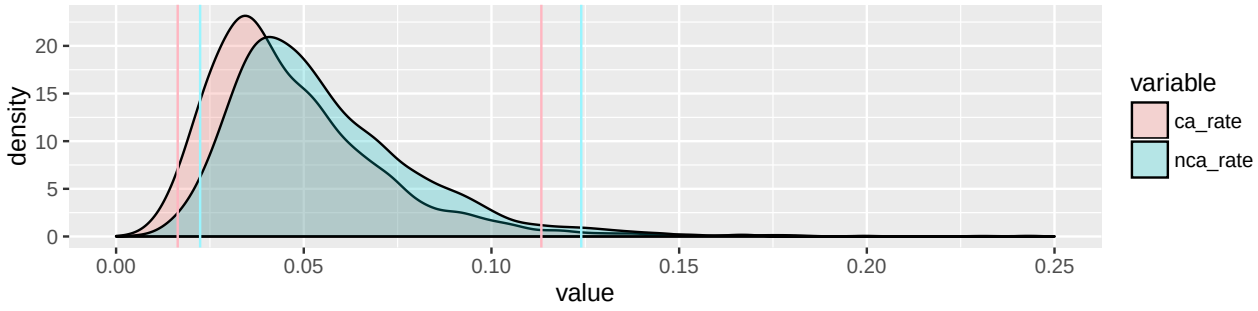
the outliers seem well-distributed. We can test this hypothesis formally by modeling the non-outlier positions as a bivariate Gaussian, and then estimating the probability that our cannabis outliers came from this distribution, and the probability that our non-cannabis outliers came from this distribution.

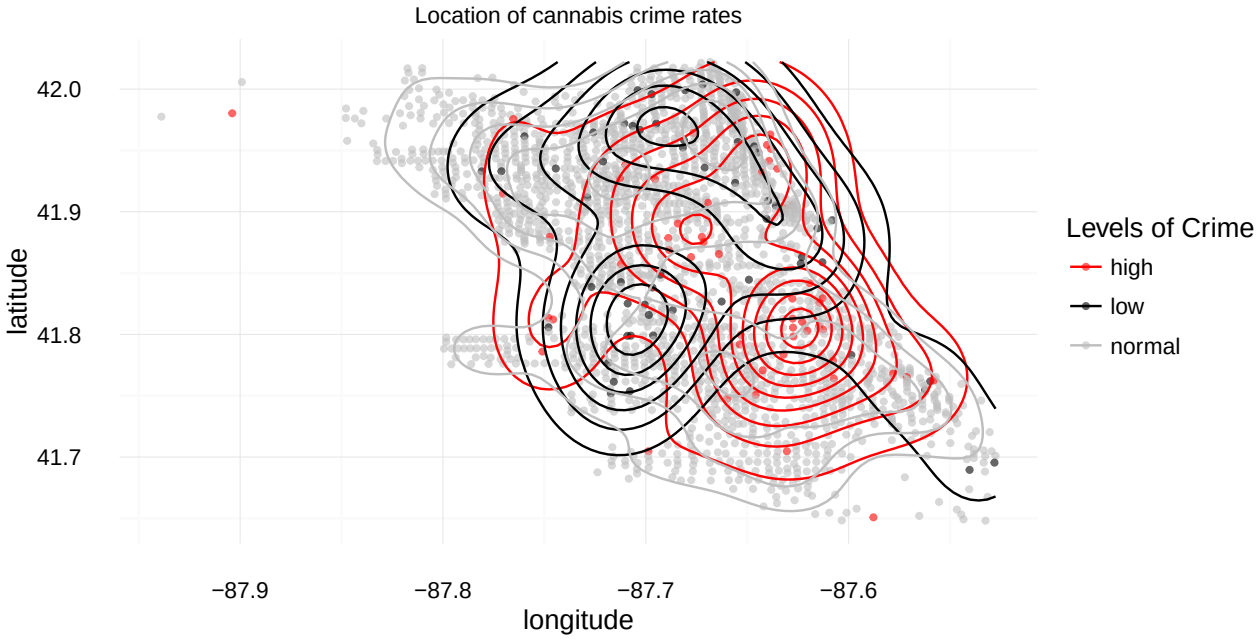
In order to identify block groups with crime rates significantly different from the mean we follow two approaches:

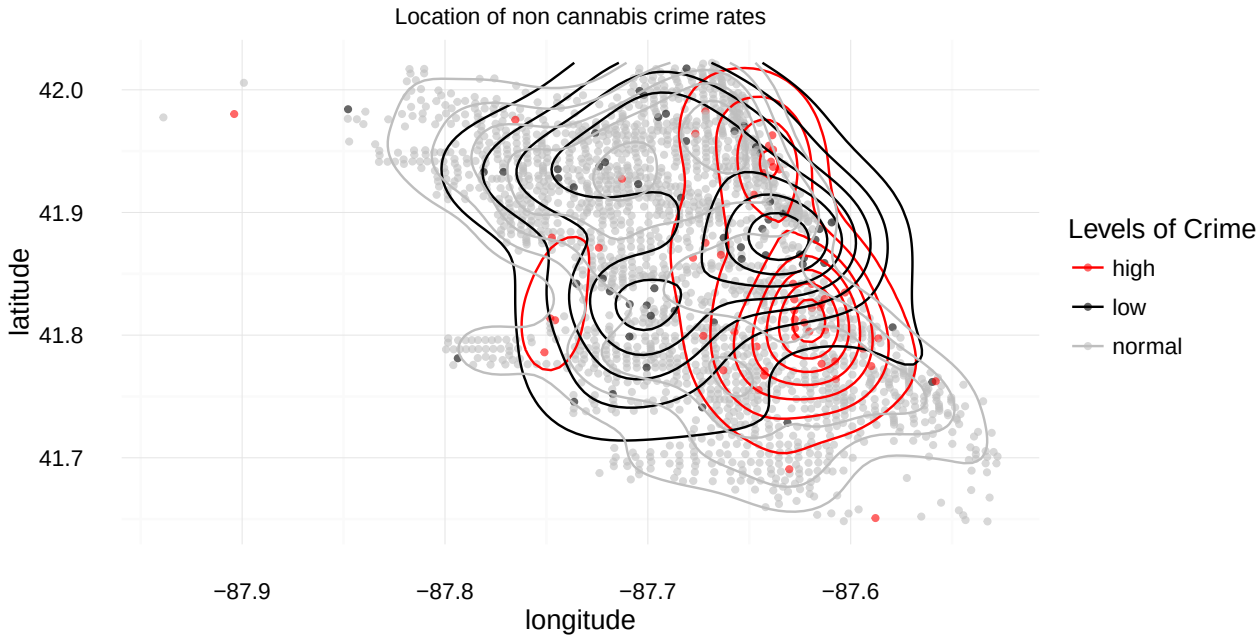
- after transformation of the response variables and of some of the covariates (specifically: log transformation of income for males and females, percentage of race out of total population, log transformation of crime rates), via linear regression we identify observations whose standardised residuals differ more than 2 standard deviations: around 100 groups for every type of crime;
- anaysing the quantiles, we use observations in the α and $1-\alpha$ quantiles, where $\alpha = 0.025$.

We analyze the dataset with both methods, and we reach the same conclusions: there is no clusterisation of low or high crime rates. Hence we decide to report the results related only to the second approach.

The shape of the distribution for the two types of crime is similar, although the values of the quantiles for non-cannabis crimes appear to be slightly shifted to the right. The plot highlights 0.025 and 0.975 quantiles for the two distributions. The second figure shows the spatial distribution of the block groups, highlighting extremely high and low crime rates for both types. Contours of density estimation are reported in the plot.

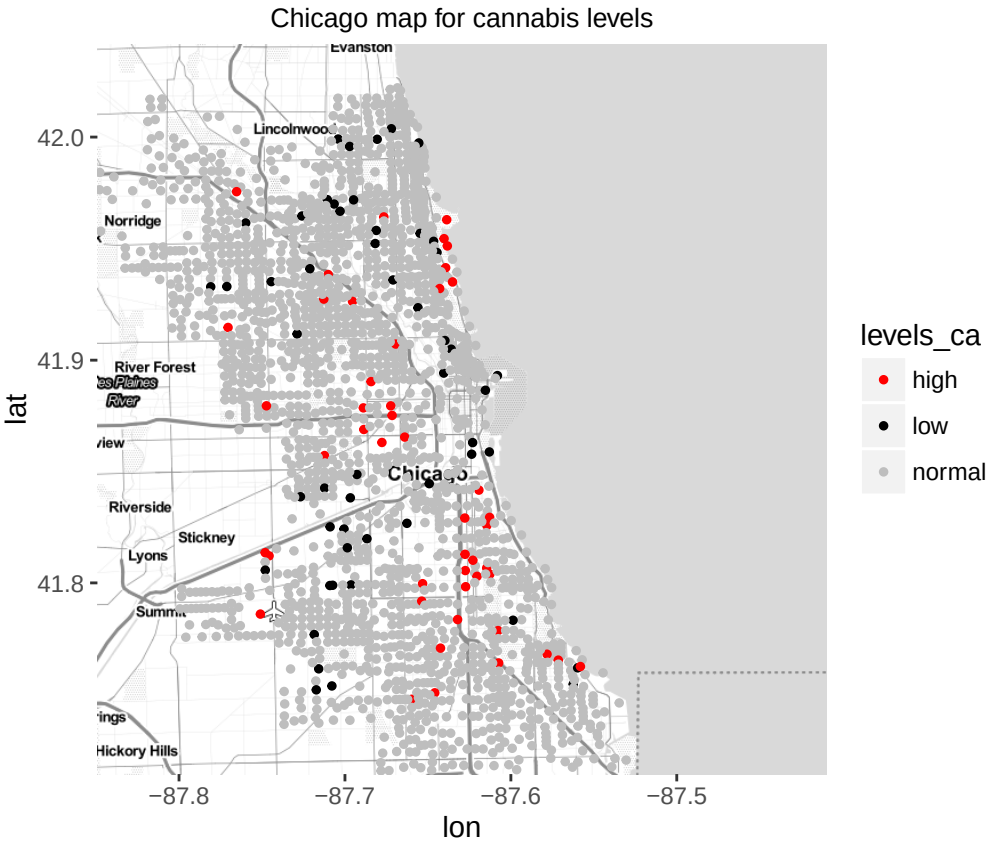






We do not notice any clusterization of points, but, aware of the fact that such a cluster may be hard to notice by eye, we perform Kolmogorov-Smirnoff test to compare the cumulative density functions, otherwise known as Peacock test for multi dimensional distribution. We want to test the null hypothesis under which low, high and normal crime rates are equally distributed on the two dimensional space. For every type of crime, the p-value does not fall below the chosen α-threshold of 5%, and we conclude that we can not reject the null hypothesis. Alternatively, one may perform a Chi-Square test choosing appropriate regions.

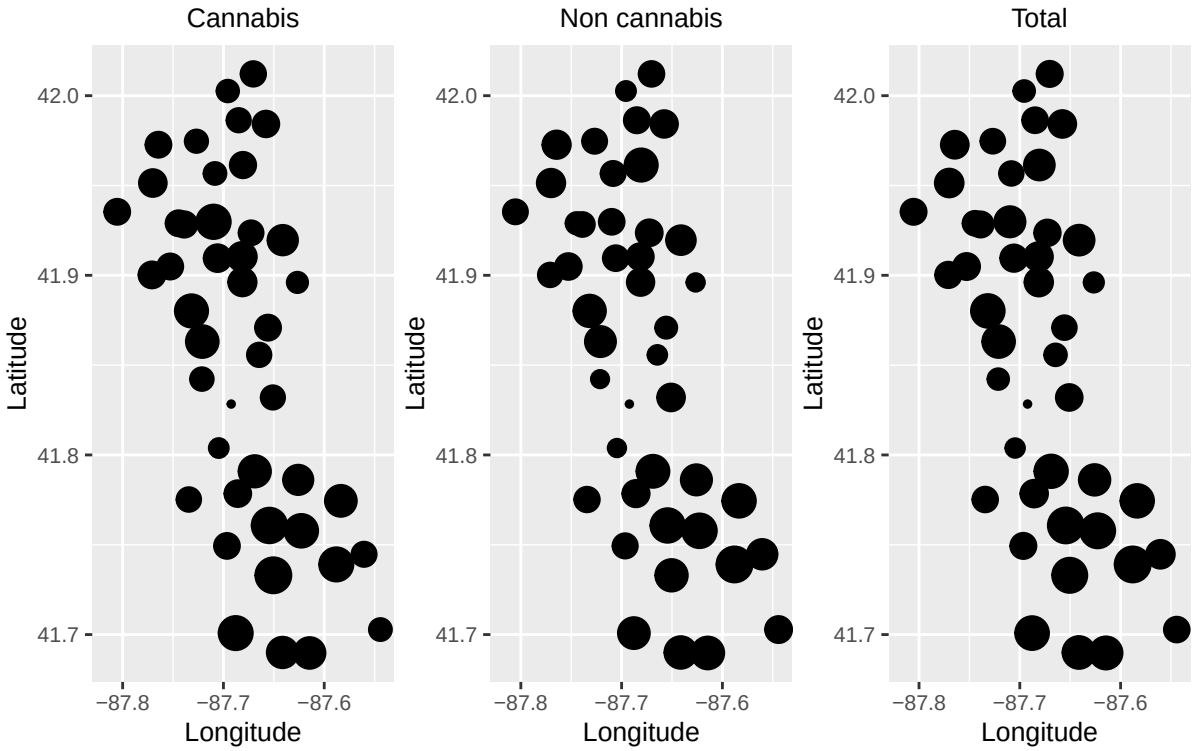
To understand better the pattern, we report the map for non-cannabis crime rates on the chicago map.



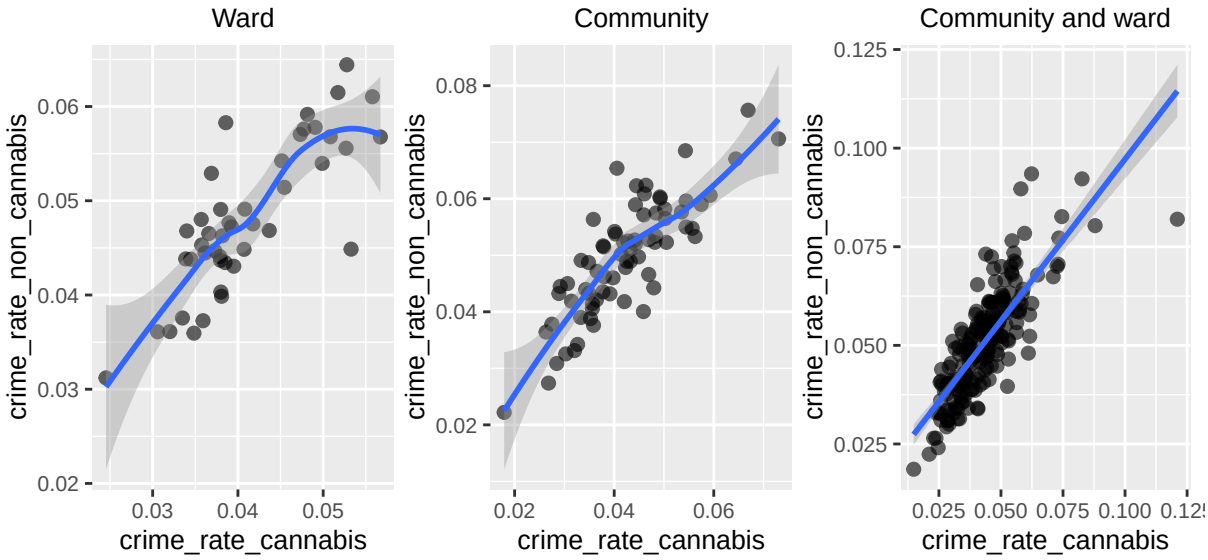
From the map we see that high crime rate seem to be clustered around Hyde Park, but nothing else can be said about the location. Let's conclude noticing that sometime high and low levels of crime occurr in very close locations, which may sound counterintuitive. We attribute this fact either to measurement errors or financial neighborhoods close to living areas.

(b) *How do crime rates vary when we consider Ward and/or Community Area? Is there a relationship between cannabis- and non-cannabis-related reports at these levels?*

We investigate differences if there is any relationship between cannabis and non crime reports at Ward, Community, or jointly Ward and Community levels. Here we have produced plots for crime rates at Ward, Community and both Community and Ward levels. We report only the ones at Ward level, for cannabis, non-cannabis, and total crime rates in sequence. To read the plot, notice that the larger the point, the higher the crime rate is in that specific ward. The number of Wards is 50. We do not notice any spatial clustering of the crime rates, apart from a concentration of higher crime rates in the south (although Wards seem to be larger in size), that we had previously noticed on the map of Chicago.



We turn our analysis to the study of the existence of any kind of relationship between cannabis and non cannabis crimes at the three different levels. We report crime rates at every level, in sequence: ward, community and both. We perform LOESS, whose regression is reported in the plots. We conclude saying that the relationships appears to be pretty linear, and simple linear regression might be a good candidate model.

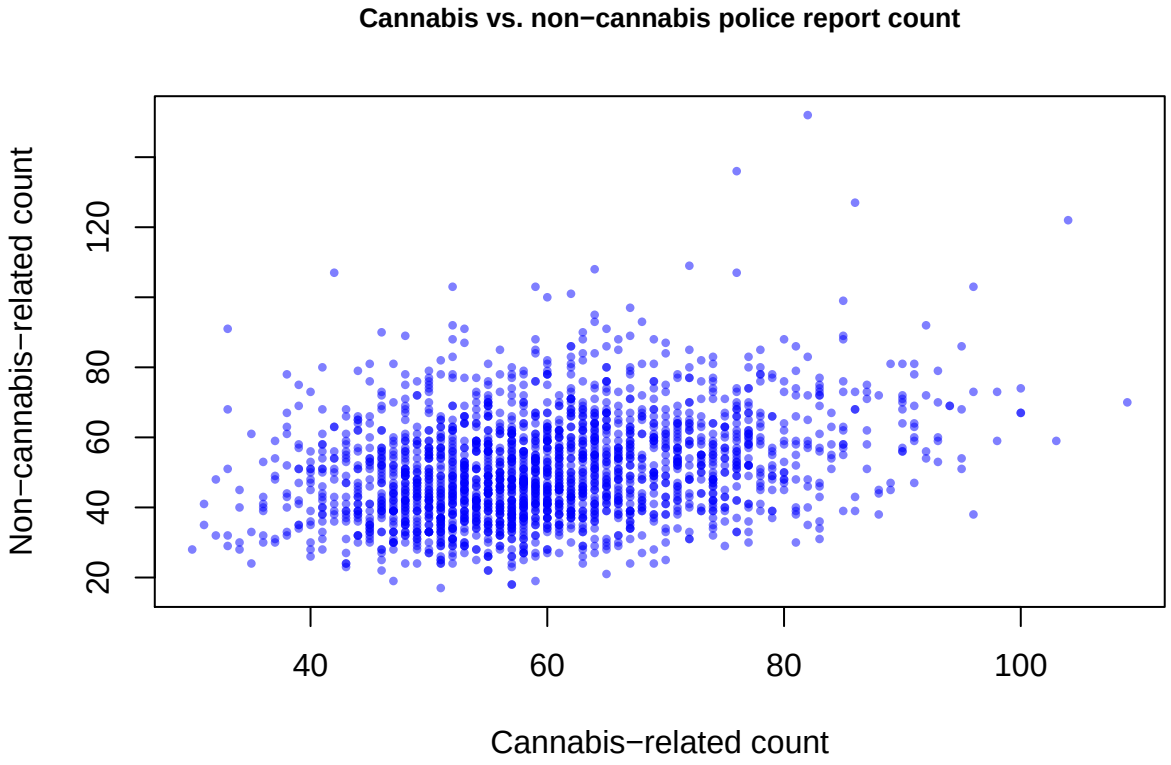


(c) *Is there a correlation between cannabis- and non-cannabis-related police reports in each block group? How does this behave when controlling for race? When controlling for income?*

The cannabis- and non-cannabis-related police reports in each block group are significantly positively correlated. The correlation is 0.3052155. Such a correlation gets significantly stronger for block groups with more black residents. The association between income and the correlation of the two types of crimes is significantly

negative, but quite weak and even ignorable. Therefore, the correlation changes reasonably and significantly when controlling for race, but almost remains the same when controlling for income.

In particular, we utilized a linear regression of cannabis-related crime rates on non-cannabis-related crime rates in block groups to solve this question. The slope coefficient would indicate such correlation. To study the correlation when controlling for income and race, we first introduce factor variables indicating the level of income and the types of raceproportion, and fit linear models containing these two factors. Based on the analysis of main effects, we fit linear models containing original income variable and a variable indicating the proportion of the black in the population, including interaction term between the two. The summary of the regression shows that the cannabis- and non-cannabis-relatedpolice reports in each block group are significantly positively correlated at level $\alpha = 0.05$, with $t(2097) = 57.82$ and $p \approx 0$. The estimated slope coefficient is 0.704. The correlation gets significantly stronger when controlling for race, or the black population proportion, with $t(2095) = 1.863$, and $p = 0.06$, and gets significantly weaker when controlling for income, with $t(2095) = -2.508$ and $p = 0.01$. However, the difference when controlling for income is quite weak, so it would be reasonable to conclude that the correlation gets almost unchanged when controlling for income.



Similarly to the Ward- and Community-Area-level study, we fit a linear model to check the relationship between two types of crime rates on the block group level. Furthermore, we introduce factor variables indicating the level of income and the types of race proportion, and fit a bigger linear model containing interaction terms to studying the correlation when controlling for income and race. The year.ca and year.nc are defined as the number of police reports of two types of crimes in each block groups divided by the total population of that block group. The scatterplots show two obvious outliers. After removing the two outliers, the scatterplot becomes better and we can fit a linear model on that data.

Once we transform the original data by the log, the distribution looks Normal, thus it would be reasonable to fit a linear model on it. The residual plots indicate that the model fits the data relatively well. For race proportion, I group the block groups into three classes: white-dominating class, black-dominating class and asian-dominating class, according to which race has the largest population in that block group.

I fit a regression model containing the interaction term of non-cannabis-related crime reports and race class factor. The summary of the fitted model shows that the correlation between cannabis- and non-cannabis-

related police reports in each block group varies among different race classes. The coefficient of black dominating group is significant.

Table 1: Summary of log.year.ca $\sim$ log.year.nc $*$ race.fac

R^2	$\hat{sigma}$	df	$F_{5,2093}$	p value
0.620	0.262	2093	684	0
Coefficient	Estimate	sde	t statistic	p value
(Intercept)	-1.26150	0.06929	-18.207	**0**
log.year.nc	0.66729	0.01580	42.234	**0**
race.fac2	0.28918	0.11059	2.615	**0.00899**
race.fac3	0.20183	0.53251	0.379	0.705
log.year.nc:race.fac2	0.05339	0.02615	2.041	**0.0413**
log.year.nc:race.fac3	0.04209	0.11798	0.357	0.721

This result inspires us to restrict the race effect to the effect of the proportion of black population. The summary in table shows that the coefficient of the proportion of black population, or generally the effect of race, is significantly positive.

Table 2: Summary of log.year.ca $\sim$ log.year.nc $*$ black.prop

R^2	$\hat{sigma}$	df	$F_{5,2093}$	p value
0.620	0.262	2093	1140	0
Coefficient	Estimate	sde	t statistic	p value
(Intercept)	-1.27140	0.07302	-17.412	**0**
log.year.nc	0.66573	0.01666	39.956	**0**
black.prop	0.31507	0.12948	2.433	**0.0150**
log.year.nc:black.prop	0.05735	0.03078	1.863	**0.0626**

For income, we also firstly introduce factor variables to check the main effects. We divide block groups into two classes, the high income groups whose `income` is larger than the median of population, and low income groups. However, neither the scatterplots nor the summary of the model indicates a significant difference between the two classes, for both original data and the log-transformed data.

Thus we turn to fit a model on origin continuous variable of income. But bifferent from the case of race, for income, if we fit a model based on the log-transformed data, the correlation between the rates of two types of crimes would not have significant difference for different income groups. Thus we fit a model in the original scale, even though the distributions can be skewed, to see the possible effect of income.

The summary of the model shows that the interaction term of `year.ca` and `income` is significant.

From table we can see that the income is significantly associated with the correlation between the crime rates of two types of crimes, but such a association is quite weak, with estimated coefficient smaller than 10^{-5}.

(d) *Similarly, does the relative wealth of a block group contribute to the crime rates? How about the racial distribution? Is there an interaction between these factors?*

Block groups with more black residents have significantly higher crime rates, with a reasonable difference. The income has significant effect on crime rates, but such an effect is quite weak. The interaction between two factors exists and is significant, and the effect of such interaction is also quite weak. Therefore, we have initial analysis which suggests the race distribution contributes more than the relative wealth of a block group.

Table 3: Summary of year.ca $\sim$ year.nc $*$ income

R^2	$\hat{\sigma}$	df	$F_{3,2095}$	p value
0.616	0.005	2095	1122	0
Coefficient	Estimate	sde	t statistic	p value
(Intercept)	4.272e-03	5.340e-04	7.999	**0**
year.nc	8.874e-01	2.908e-02	30.517	**0**
income	3.868e-08	1.445e-08	2.678	**0.00747**
year.nc:income	-1.915e-06	7.636e-07	-2.508	**0.01221**

In particular, we analysed the effects of two factors by firstly studying the main effects and interaction of two factor variables in (c) for income and race proportion, and then fit a linear model containing original continuous variables of income and the proportion of the population that is black.

Based on the feature of the data, we also fit a Poisson regression model with the same predictors. Both linear model and Poisson model for (d) cannot be fit quite well (with quite large RSS and residual deviance). But since we are focusing on the significance of coefficients, that does not matter much. The summaries of two linear models for one factor each indicates that block groups with higher income have significantly larger crime rates, though the difference is very small. Furthermore, block groups with a higher proportion of black residents would have significantly larger crime rates, with a reasonable difference. The interaction between two factors exists and is significant, but the effect of such an interaction is very weak.

The summaries of two fitted models for one factor indicates that the crime rate for high income class is significantly larger, with $t(2097) = 2.70$ and $p = 0.007$. And the crime rate for black-dominating block groups is significantly larger, with $t(2097) = 13.14$ and $p \approx 0$. The summaries for one-factor models are attached below (race.fac2 indicates black, race.fac3 indicates asian, income.fac2 indicates high income).

Table 4: Summary of log.year.crime $\sim$ income.fac

R^2	$\hat{sigma}$	df	$F_{1,2097}$	p value
0.42	0.003	2097	7.26	0.007
Coefficient	Estimate	sde	t statistic	p value
(Intercept)	-3.49196	0.01297	-269.295	**0**
income.fac2	0.04944	0.01834	2.695	**0.00709**

Table 5: Summary of log.year.crime $\sim$ race.fac

R^2	$\hat{sigma}$	df	$F_{2,2096}$	p value
0.40	0.077	2096	88	0
Coefficient	Estimate	sde	t statistic	p value
(Intercept)	-3.55700	0.01126	-315.934	**0**
race.fac2	0.23981	0.01825	13.137	**0**
race.fac3	-0.09089	0.09872	-0.921	0.357

The summary for the model containing interaction term indicates that the interaction between `income` and `race` is significant at level $high.income * black$. That is, there is significant interaction between income and race factors. (Again, race.fac2 indicates black, race.fac3 indicates asian, income.fac2 indicates high income.)

Based on these facts, we can also fit a model on two continuous covariates: income and proportion of black

Table 6: Summary of log.year.crime $\sim$ income.fac $*$ race.fac

R^2	$\hat{\sigma}$	df	$F_{5,2093}$	p value
0.40	0.093	2093	43	0
Coefficient	Estimate	sde	t statistic	p value
(Intercept)	-3.63229	0.01717	-211.563	**0**
income.fac2	0.13053	0.02261	5.774	**0**
race.fac2	0.30297	0.02496	12.136	**0**
race.fac3	-0.07946	0.10859	-0.732	0.46436
income.fac2:race.fac2	-0.09874	0.03705	-2.665	**0.00775**
income.fac2:race.fac3	0.23133	0.25623	0.903	0.36673

population. The summary of the fitted model is:

Table 7: Summary of log.year.crime $\sim$ income $*$ black.prop

R^2	$\hat{sigma}$	df	$F_{3,2095}$	p value
0.40	0.097	2095	75	0
Coefficient	Estimate	sde	t statistic	p value
(Intercept)	-3.715e+00	2.734e-02	-135.911	**0**
income	3.852e-06	6.933e-07	5.556	**0**
black.prop	4.106e-01	5.209e-02	7.882	**0**
income:black.prop	-3.062e-06	1.631e-06	-1.877	**0.0606**

Based on the discrete feature of the data, it may be natural to fit a Poisson regression model

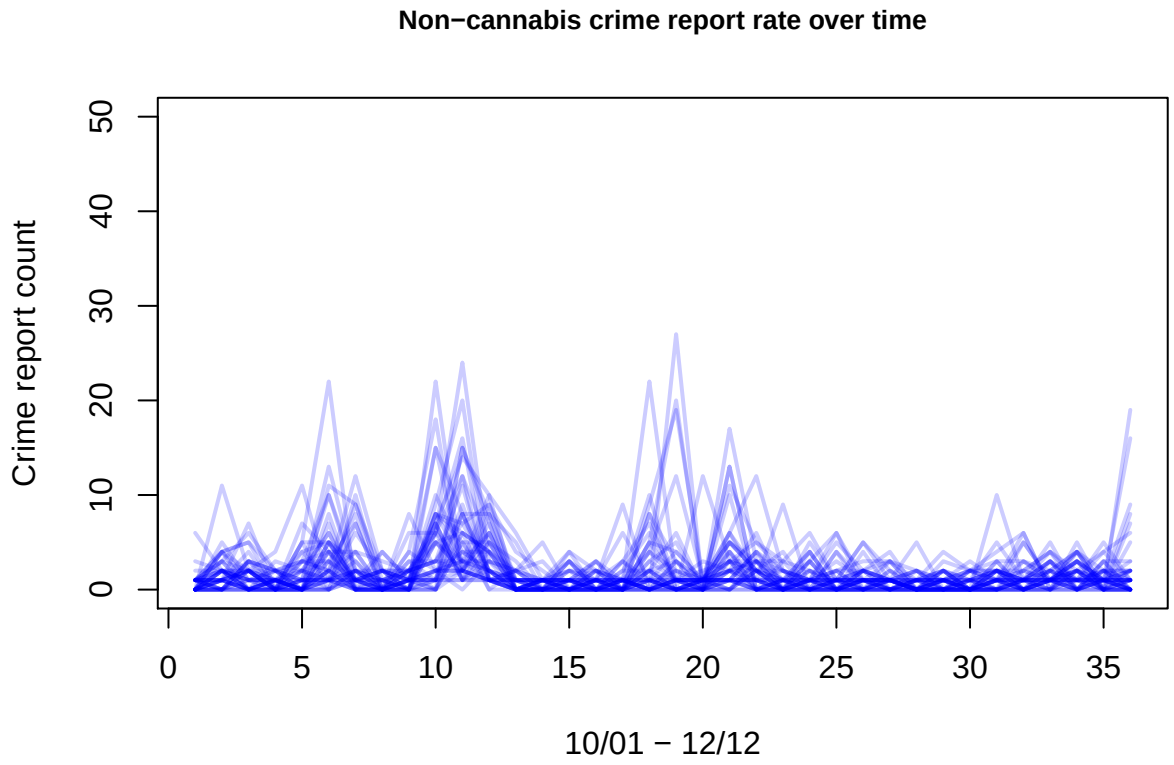
$$\log(E[\texttt{year.crime.num}|\texttt{income}, \texttt{black.prop}]) \sim \beta_0 + \beta_1 \texttt{income} + \beta_2 \texttt{black.prop} + \beta_3 \texttt{income} : \texttt{black.prop} + \log(poptotal) \quad (1)$$

where `year.crime.num` is the total number of (cannabis- and non-cannabis-related) crime reports in each block group per year. $\log(poptotal)$ is the offset item. However, such a model cannot be fitted and an error "cannot find valid coefficients ..." is raised. If we set the offset term as $\log(poptotal/100)$ or $\log(poptotal/1000)$, the model can be fitted but quite bad. For example, the residual deviance is too large (3402650 on 2095 degrees of freedom). The summary (which is omitted here) shows that the coefficients for income, the proportion of the black, and the interaction between the two, are all significant.

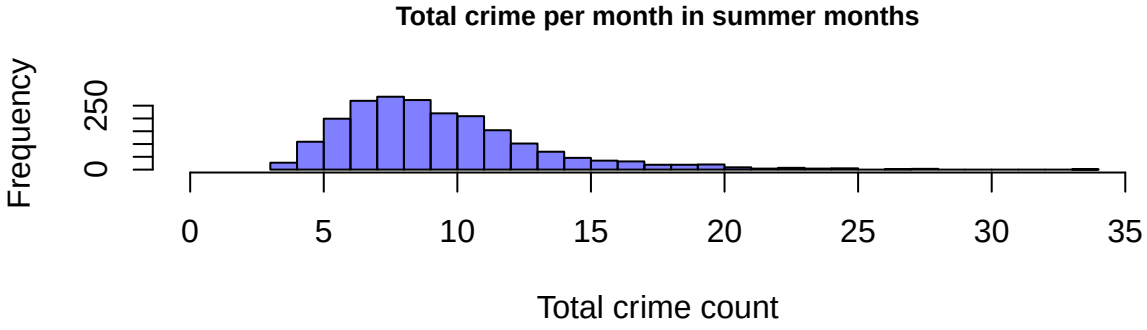
The summary of the Poisson model indicates almost the same result, except that by the Poisson model, block groups with higher income have significantly lower crime rates, but the difference is also very small like that in a linear model. Such a contradiction and small magnitude of coefficient shows that the income is just weakly associated with crime rates.

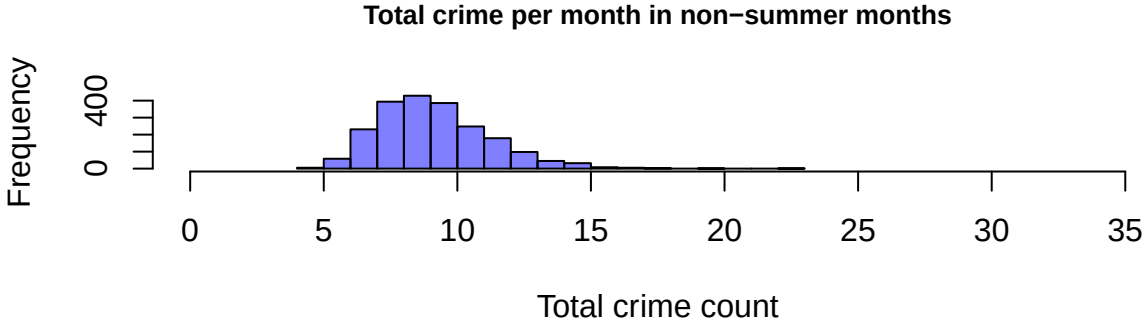
(e) *Some crime rates are known to rise during the summer months and cycle throughout the year. Do these reports have similar trends?*

Exploratory data analysis of the evolution of crime rates over time shows that there is probably a pattern.



We can formally test this by creating an indicator "summer month" and testing whether the crime rates in summer months are different from non-summer months. The distributions do look like the crime rate in summer months has a longer right tail; that is, more high values.





We can use a simple t-test to see whether these two distributions come from the same population or are different.

Though the Normality assumptions are dubious since the summer data is right-skewed, this test rejects our hypothesis that the summer and non-summer monthly crime rate distributions by block are different ($p < 0.001$). As an initial pass, this suggests further investigation into the difference between monthly crime rates in the summer vs. non-summer.

(f) *Using your insights from the previous sections, construct a predictor for the total number of crimes in each block group from October to December 2012 using crimes from January 2011 through September 2012, plus all demographic information available. Consider training the model on the same period of time in 2011 using all crimes that preceded it. Which factors – seasonal, local temporal, demographic, geographic – play the strongest roles?*

Given our insights from the previous sections, we know there is seasonality present in crime rates and that crime rates vary significantly by block group. A simple, but likely effective, predictor for the total number of crimes in each block group from October to December 2012, is the total number of crimes in each block group from October to December 2011. Initially, we do not even need to incorporate demographic information, since we already know many of these are contained within the block groups themselves.

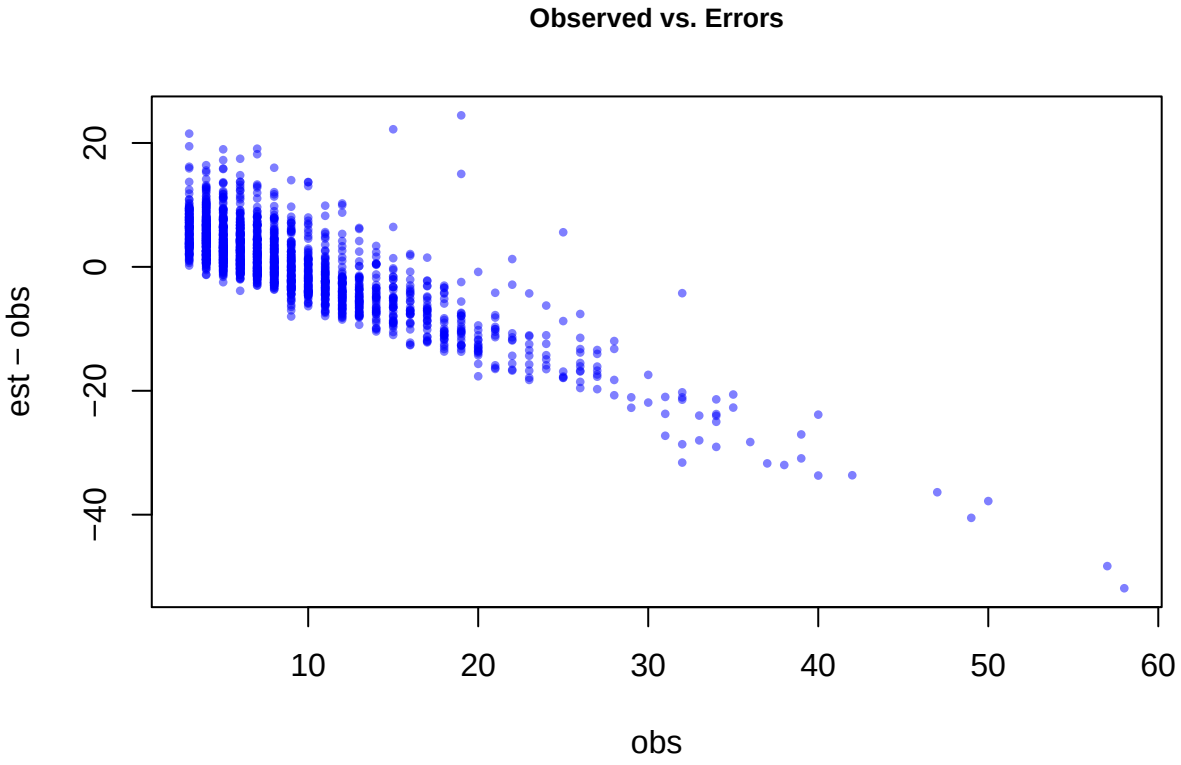
Using this model, we can calculate our MSE.

```
## [1] 74.44392
```

We can now ask the question, is this estimate improved by incorporating demographic information? When we model the total number of crimes in October thru December 2011 per block as a Poisson model fitted on black population percentage, male age, income, and Ward, the fitted values of this model actually lower the MSE.

```
## [1] 50.64395
```

Furthermore, when we use even more data, not only the months October thru December of 2011, but also all the data up to September 2012, and then take the average monthly crime rate, we lower our MSE even further.



```
## [1] 49.55379
```

Since we are simply using the GLM to predict future (theoretically unobserved) values vs. inference, it is not as necessary to check the model assumptions, which are pretty unreasonable (as discussed in the previous section).