**Analysis of Crime Statistics in American Cities between 1975-2015**

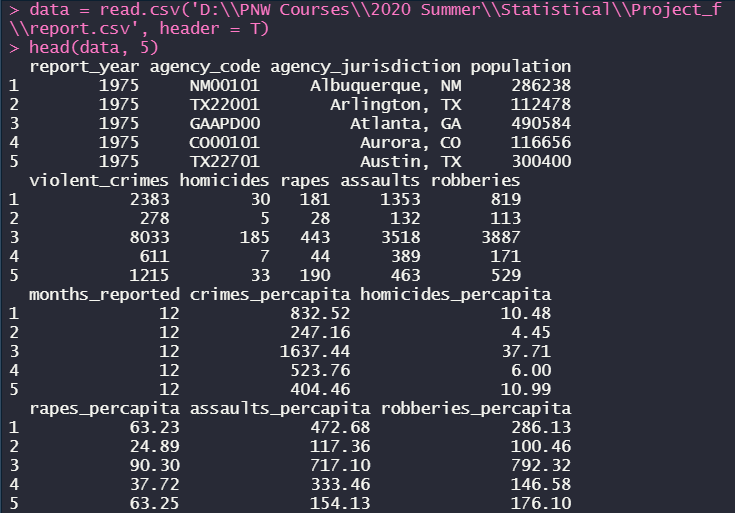
**Statistical Computing – Final Project**

Eric Wei Jordan Nikkel

**Objective:** The purpose of this projectis to determine the crime curve from 1975 to 2015 based on the dataset that FBI collected from the country’s more than 18000 police agencies. National estimates can be inconsistent and out of date, as the FBI took months or years to piece together reports from those agencies that choose to participate in the voluntary program. To try to fill this gap, the Marshall Project collected and analyzed more than 40 years of data on the four major crimes the FMI classifies as violent – homicide, rape, robbery and assault – in 68 police jurisdictions with populations of 250, 000 or greater. We’re gonna use what we’ve learned through this class to extract as much valuable information as we can and do some data analysis.

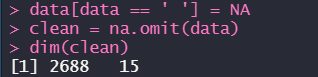
**Data Introduction**: The crime data was acquired from the FBI Uniform Crime Reporting program's "Offenses Known and Clearances by Arrest" database for the year in question, held at the National Archives of Criminal Justice Data. The data was compiled and analyzed by Gabriel Dance, Tom Meagher, and Emily Hopkins of The Marshall Project; the analysis was published as [Crime in Context](https://www.themarshallproject.org/2016/08/18/crime-in-context) on 18 August 2016.

First things first, based on the type of the dataset, we import the data with “**read.csv()**” command, and display the first 5 observations. Then for the convenience of directly accessing the variable, I attach the data and view the dimension of the dataset.

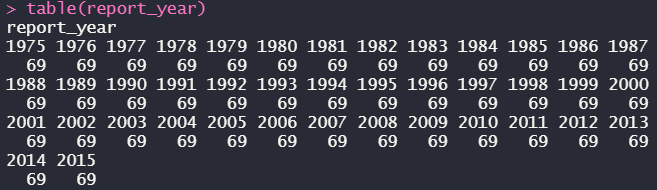




At the very start when I looked through the dataset in ‘excel’, I found there were some missing data denoted by blank space, so I try to clean out the noise by converting them into ‘NA’ first, remove them and generate another new datset called “**clean**”. By viewing the result of the dimension, we can see that there are **2829 – 2688 = 141** variables removed.

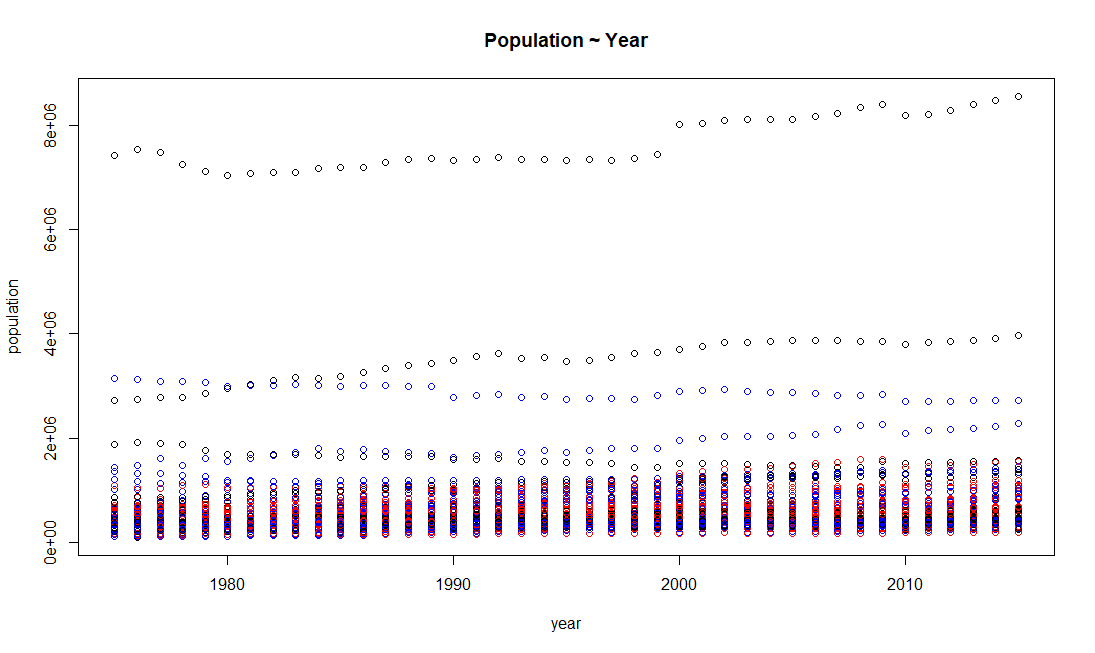


Then I apply the “**table()**” to build a contingency table of the “year” variable and found that each year in this dataset holds the same number, since there are 69 cities counted in all for each year, such that the number is constant.



By using the “**plot()**” function I successfully drew the plot of population within these forty years, and from the plot below we can apparently see that the overall population of each city almost didn’t change a lot, mainly distributed in the range below **4e+06**, a minority of cities held a higher population that close to **8e+06**.



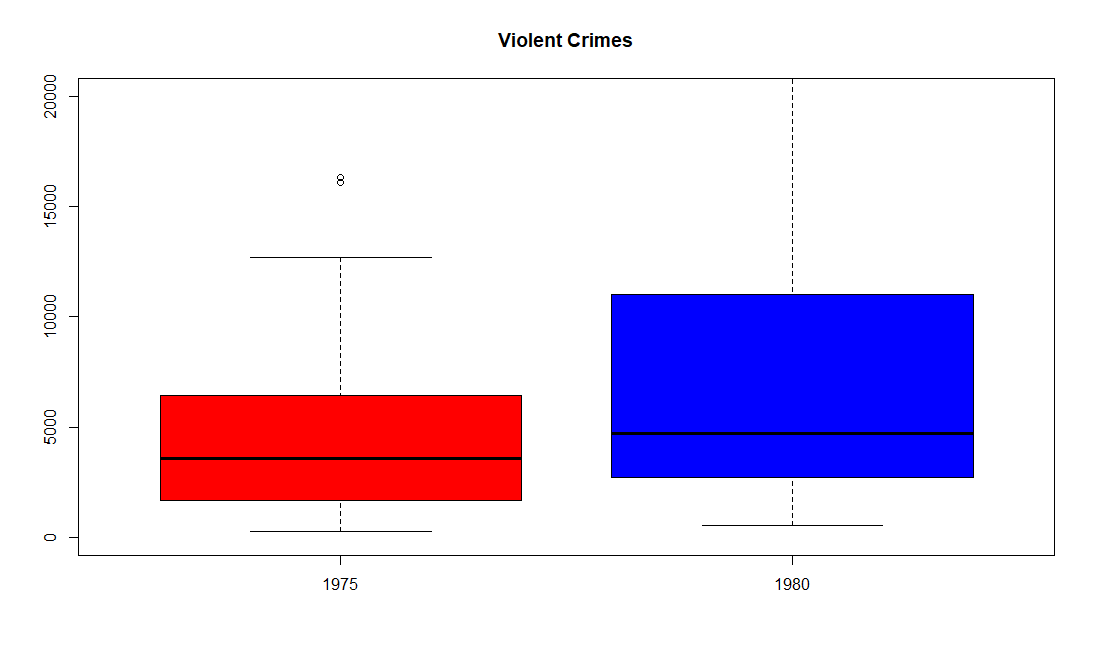


I created two subsets with years of 1975 and 1980, trying to compare the crime condition over these five years and doing some analysis to determine if the “**violent\_crime**” variable was going better or not. And based on the box-plot that I displayed below, very obviously that in the 1980 the violent crimes in America were more rampant than that of 1975, both the median and maximum value are greater and the overall trend is going up.

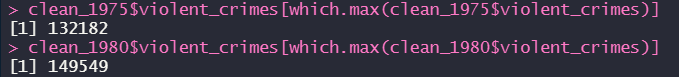


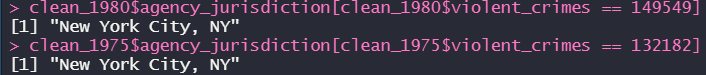




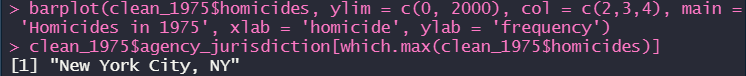


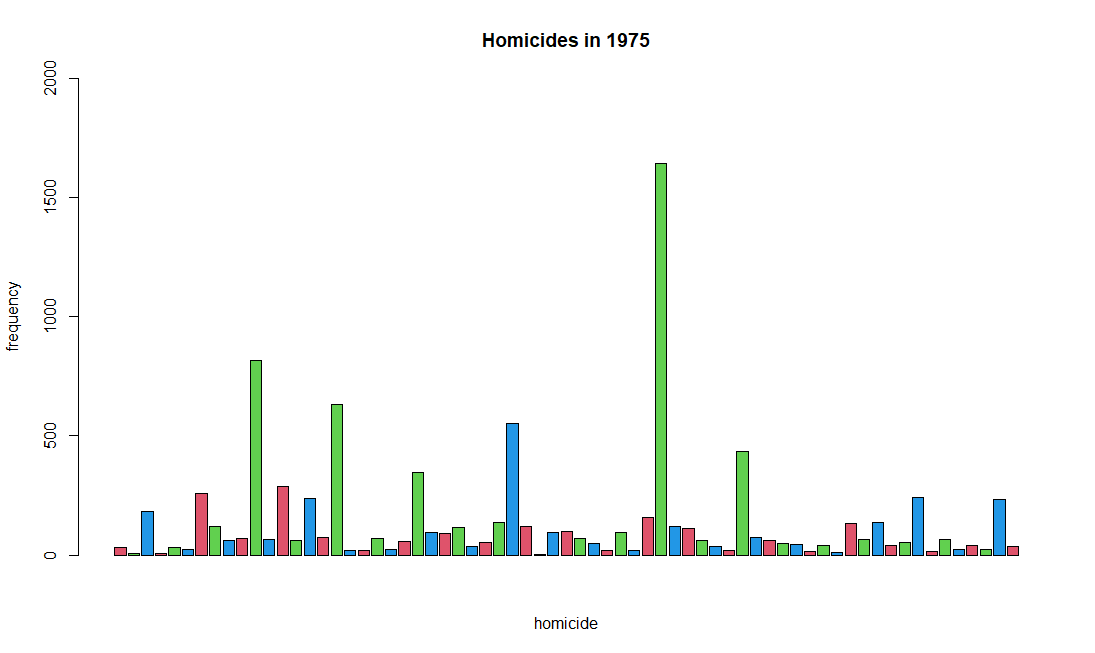
For making a double check, I try to access the maximum value of **violent\_crimes** of both them, and found indeed the index in 1980 is greater, also based on the value I further got the conclusion that **New York was the city where violent crimes were most likely to happen at the time.**





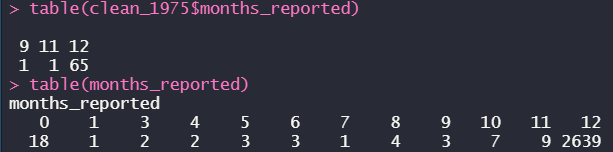
By using the “**barplot()**” function, I check out the distribution of homicides over the America in 1975, and the plot below also indicated that there was a city with an overwhelmingly high homicide rate, and again it ended up showing the city is New York.



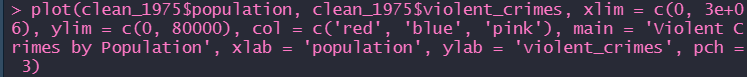


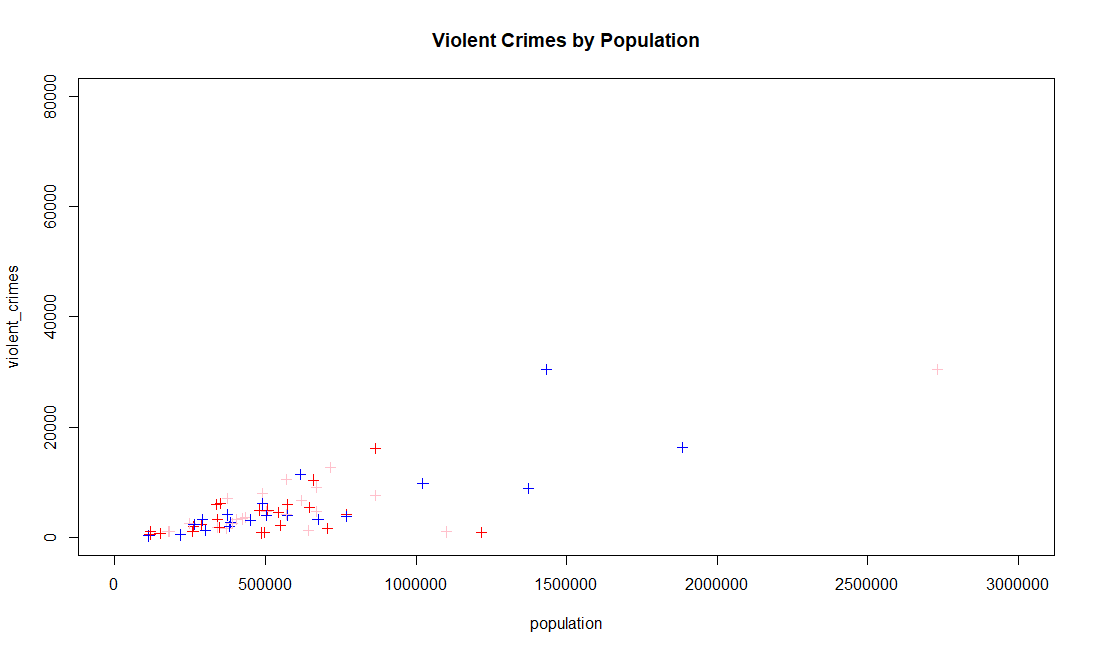


I use the ‘**table()**’ function to view the distribution of the **reports\_months** in 1975, and it indicates that the December is the most frequent one, then I checked it out again based on the overall dataset, and as I thought the **December** is indeed the most reported month and the index is way much greater than other months, than we can conclude that most of the cases happened in the December of each year.

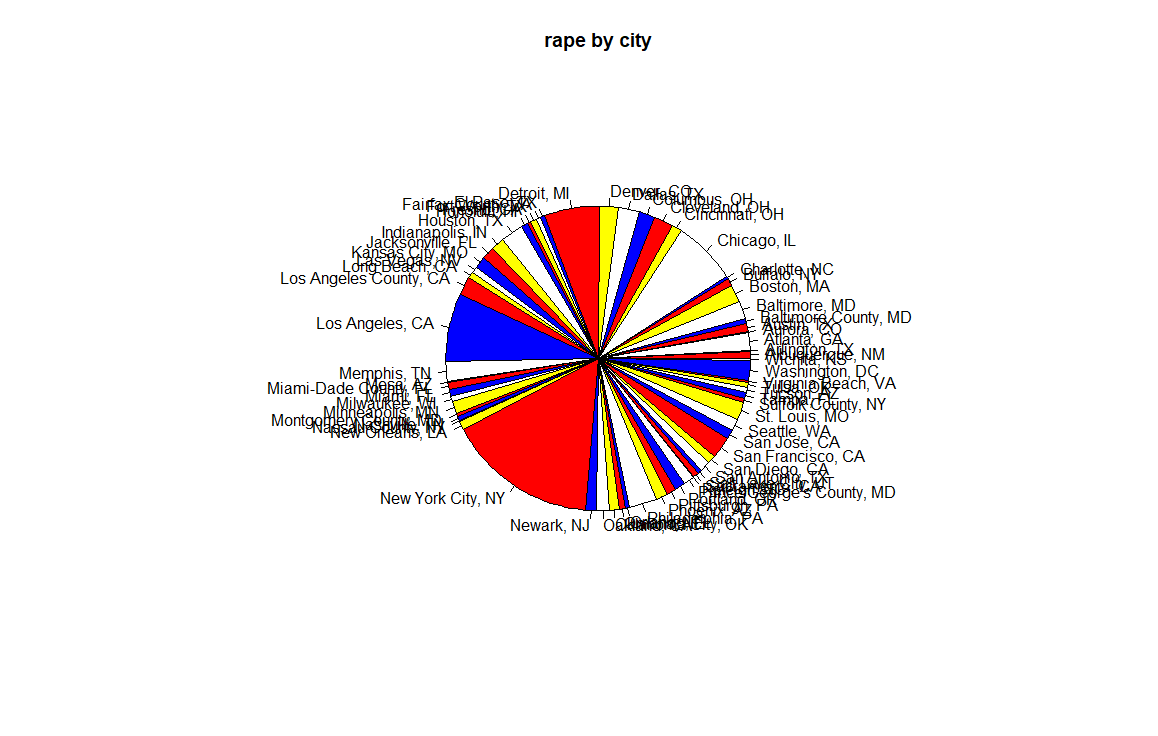


With the variable of ‘**population’** and ‘**violent\_crimes**’, I drafted another scatterplot which displays the relationship between them, and the plot below proves that they are positively relevant.



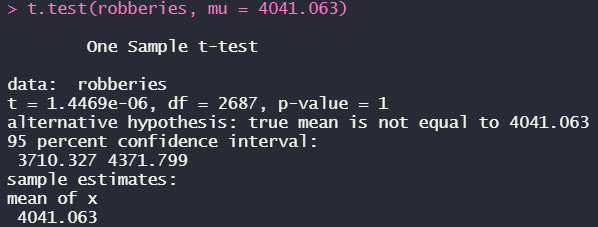


Create a **pie-chart** and observe the city distribution for “**rape**” crime, and as far as I see, the **New York City**, **Los Angeles**, **Chicago** and **Detroit** takes up main percentages of this chart, which means that these four **metropolises** hold a higher probability of “rape” than any other cities.



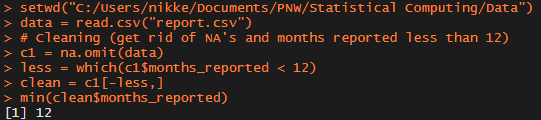
For calculating the mean value of the “**robbery**” variable in the root dataset, I use the “**mean()**” function to determine it, but in order to ensure the accuracy, I also apply the “**t.test()**” to get the validation, the result is as below. The **Null Hypothesis** in this test is the size of the difference relative to the variation is 0, yet the **Alternative Hypothesis** is not, from the p-value that we get, we have strong evidence to support Null Hypothesis to say that they don’t have any differences of mean value for it’s way greater than significance level(0.05), so the conclusion that we get is **the mean value of “robbery” is exactly 4041.063**.



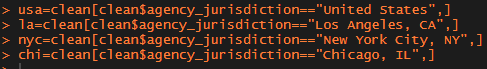


\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

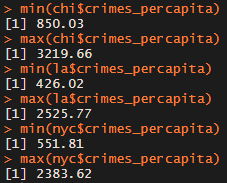
**Most violent cities:** Violent crime rates often influence models that determine quality of living for cities and are also generally of interest to members of the population. I thought it would be interesting to compare crime rates among America’s 3 largest metropolises, New York City, Los Angeles and Chicago. In order to do this, I cleaned the original data set by omitting all NA values using **“na.omit”** and additionally filtered out instances where less than 12 months were reported in order to increase accuracy. The **“min”** function was used to determine whether or not the subset creation was successful.



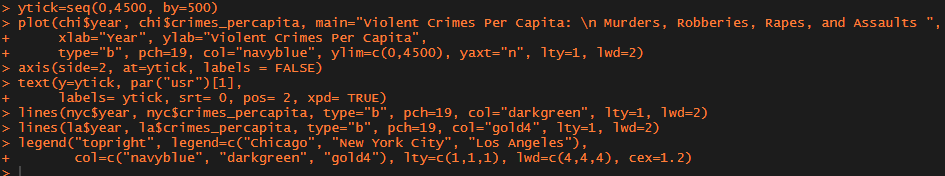
Subsets were created for each city in question for ease of use and visualization.



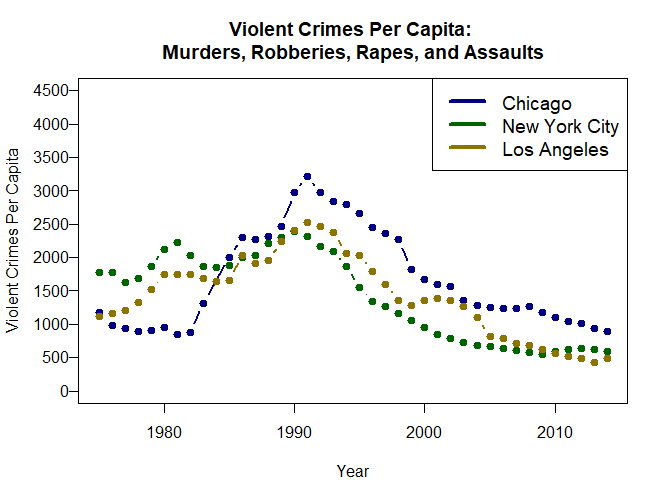
Per capita rates were used in these comparisons in order to control for population differences between the three cities. The **min** and **max** values of overall crimes per capita for each city were determined in order to help develop the dimensions for visualizations plots.



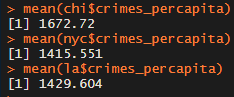
The minimum value found was **426.02** and the max value was **3219.66**. These helped influence the dimensions of the y-coordinate in the subsequent plot. The following code was used to develop the plot for visualization.



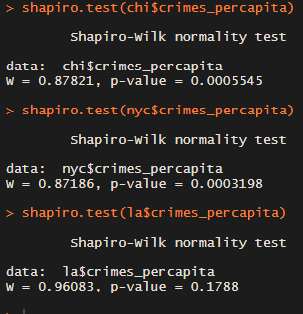
The following plot was developed:

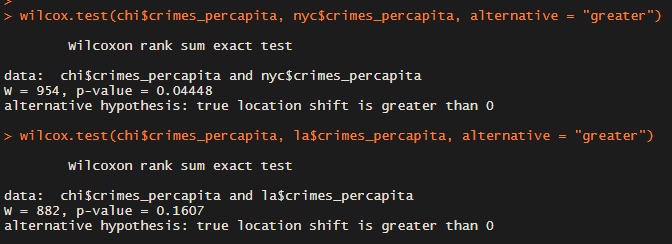


Though there are some slight variations, I inferred from the plot that, in general, Chicago appears to be the most violent city, followed by Los Angeles and then New York City in that order. We can test the hypothesis that Chicago is more violent than Los Angeles and New York City by running some t-tests. These inferences are initially supported by taking the mean value for per capita crimes for each city. The mean per capita crime rate in this data set is **1672.72** for Chicago, **1429.604** for Los Angeles, and **1415.551** for New York City.



Before we do t-tests, it is important to check that the data falls under the normal distribution, considering that it is one of the assumptions required for t-testing. We can do this by running the Shapiro-Wilk normality test using the function **“shapiro.test”**. In this test, the null hypothesis is that the distribution is normal and the alternative hypothesis is that the distribution is not normal. The results are as follows.



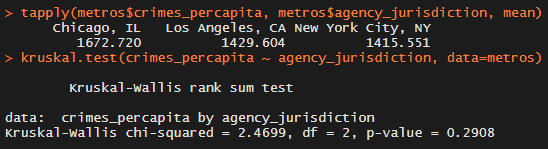
The p-values obtained for the tests on both the Chicago and New York City data are less than 0.05, implying that the data is not normally distributed. The p-value for Los Angeles’s data is, however, greater than 0.05, indicating that this data is normally distributed. Considering most of the data is not normally distributed, we will utilize the Mann-Whitney U-test instead of the t-test which we will run using the **“wilcox.test()”** function. 

The results of the Mann-Whitney U-tests are shown above. We used a one-tailed test ( **alternative = “greater”** ) because we are testing to see whether or not the rates are higher for Chicago. The p-value of less than 0.05 for the Chicago vs New York City indicates that we can reject the null hypothesis that there rate is not higher in Chicago. This indicates that the per capita violent crime rates across the years given by this data are in fact higher for Chicago when compared to New York City. The p-value for the Chicago vs Los Angeles test, however, is greater than 0.05, which indicates that Chicago’s per capita crime rate is not significantly higher than Los Angeles’s.

We can look for differences between the violent crime rates of these cities all simultaneously using the non-parametric Kruskal-Wallis test to analyze differences in the means. First, we will have to combine the data for the three cities into one data frame using the **“rbind()”** function.

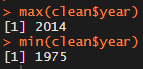


Now, we can utilize the tests provided by the package **PASWR**, which we call using the **“library()”** function.

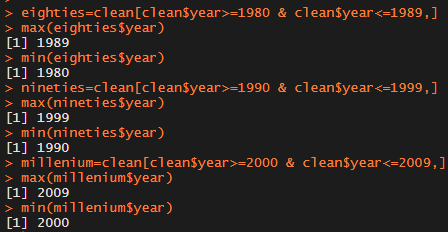


Even though we determined that Chicago was more violent than New York using the Mann-Whitney U-test, the results of our Kruskal-Wallis test gave us a p-value higher than 0.05, indicating that there are no significant differences between the violent crime rates of these cities when compared simultaneously.

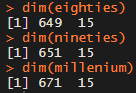
**Comparing time periods:** Another frequent topic of debate by historians and scholars is which time periods throughout US history have been the most violent. We can determine that to some extent using the data given. Typically, when analyzing crime rates, it is common to compare decades to one another. Given that the years provided in this data start at 1975 and end at 2014, we will leave out the 1970s and the 2010s and will only compare the 1980s, the 1990s, and the 2000s.



We will begin by creating subsets for each decade to help with ease of handling the data and for creating plots for visualization.

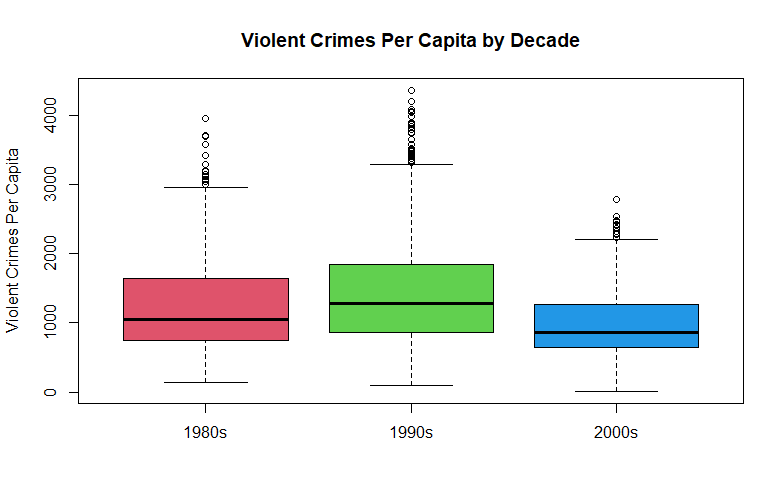


The number of observations for each subset are very similar but slightly different due to removal of some observations during cleaning.

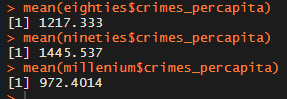


We will begin comparisons by using a boxplot to visualize the data. Again, we will use crimes per capita to control for population differences that may have occurred between the decades.

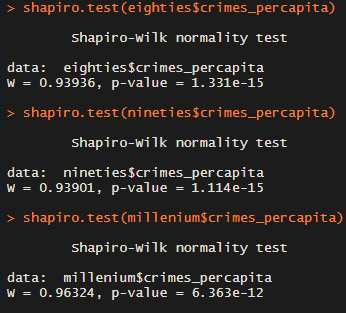




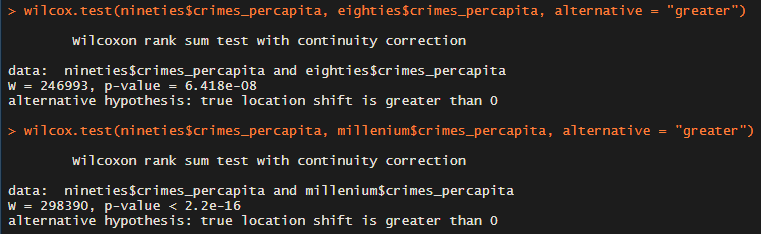
Evidently, the 1990s appear to be the most violent decade, followed by the 1980s and the 2000s, in that order. This is supported by viewing the mean crime per capita for each decade directly.



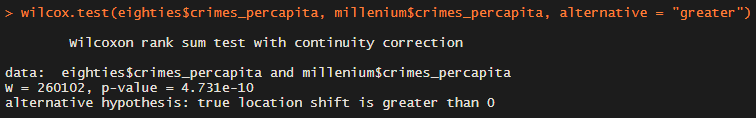
Before testing whether or not the 1990s does have a higher rate of violent crimes per capita, we will assess the normality of the data sets to determine which test we will use. We will do this once again by using the Shapiro-Wilk normality test through the **“shapiro.test()”** function.



All p-values obtained were lower than 0.05, indicating we can reject the null hypothesis and assume that none of the data is normally distributed. Once again, we will use the Mann-Whitney U-test through the **“wilcox.test()”** function and will assume that the 1990s had higher rates ( **alternative = “greater”** ).



The p-values for each test were lower than 0.05, allowing us to reject the null hypothesis. This indicates that violent crime rates per capita were in fact higher in the 1990s than in the 1980s and the 2000s. For additional insight, we will compare the 2000s and the 1980s as well, under the assumption that rates were higher in the 1980s.

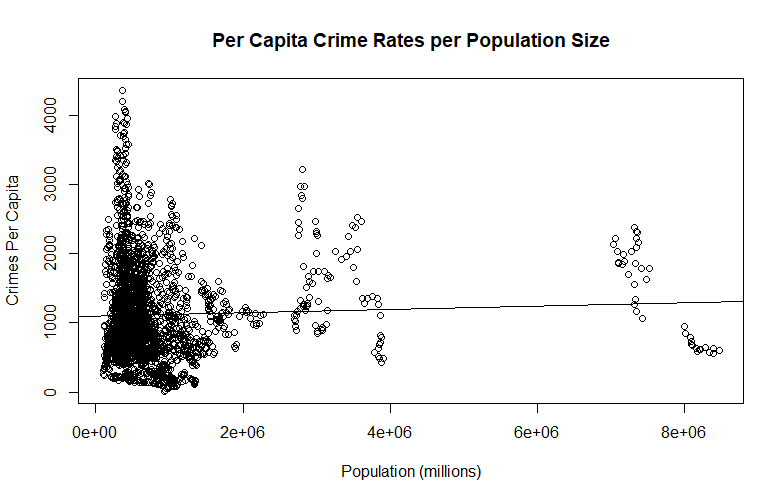


Similarly, the p-value here is less than 0.05. This allows us to conclude that the 1980s experienced higher per capita rates of violent crime compared to the 2000s.

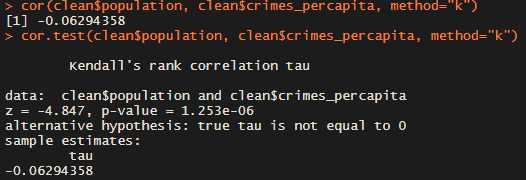
**Modeling crime and population:** Another interesting aspect that could be of note is whether or not per capita crime rates are correlated with population to determine whether or not relative crime rates increase in larger cities. We will begin by simply plotting the data to see if there may be correlations.



Initially there does not seem to be many promising trends based on the plot below. There does appear, however, to be a slight positive correlation. We can investigate this further by taking a more detailed look at the model and running a correlation test.

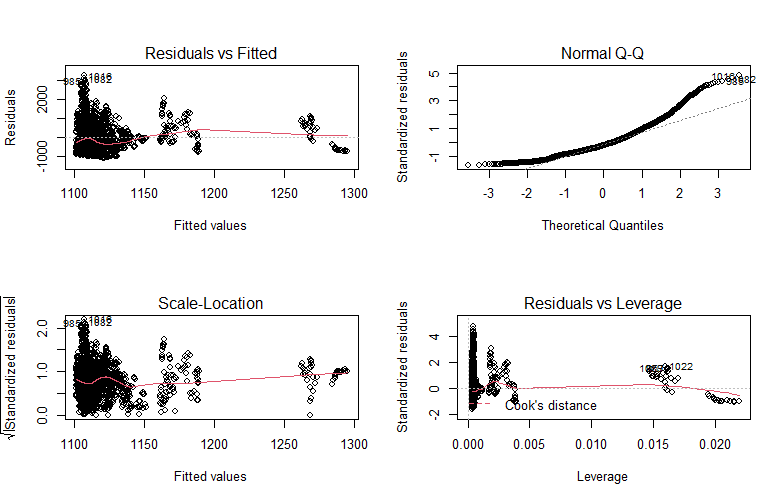


Considering that the data is not normally distributed or rank ordered, we will use Kendall’s tau coefficient to look at correlation through the **“cor()”** and **“cor.test()”** functions.



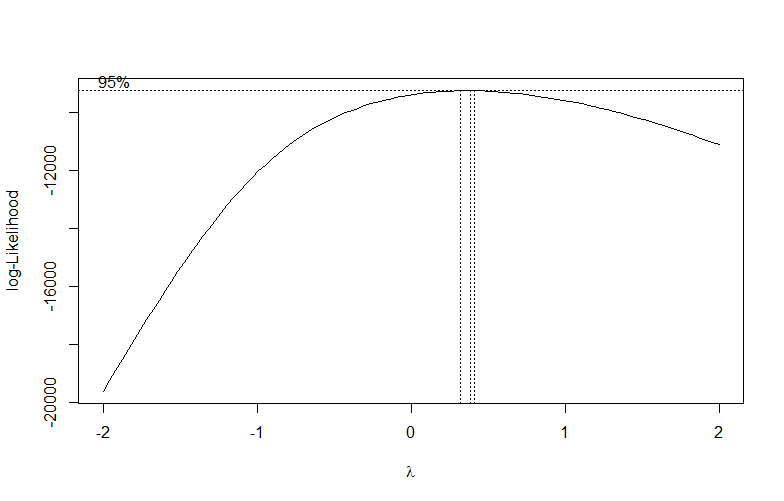
The correlation coefficient shown from theses tests is **-0.063** which actually indicates an ever so slight negative correlation. The p-value for the cor.test is less than 0.05 which indicates this correlation is significant. However, these results indicate that it is difficult to determine a correlation between these two variables. We will still continue though to analyze the model.



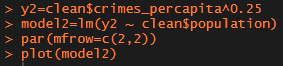


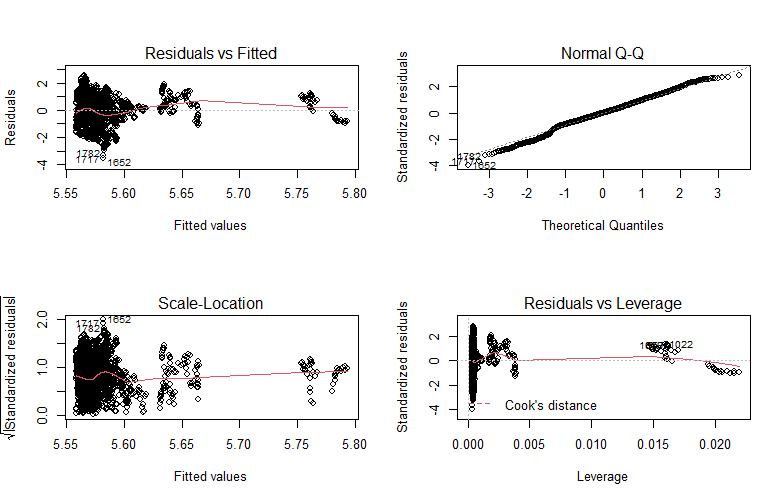
Evidently and unsurprisingly, this model is not terrific. The line for Residuals vs Fitted is pretty skewed , as we would expect the red lines and dotted lines to overlap for a good model. Additionally, the Normal Q-Q plot is poor, as we would expect the values to fall in line with the dotted line, but they veer off pretty far, especially towards the tails. We can still attempt to fix this model using a Box-Cox transformation just for fun, though. I already have installed the “MASS” package, so it is first just a matter of calling it from the library and obtaining the Box-Cox plot.





The maximum likelihood estimation here appears to fall pretty close to 0.25, so we will raise our values for the response variable (crimes per capita) to the power of 0.25 to fit into our new model.





The Box-Cox transformation actually helped fix the Normal Q-Q plot for this model, but the Residuals vs Fitted plot did not improve and may have even gotten worse.

As it turns out, it is not really feasible to model crimes per capita using population alone. Population could potentially serve as an explanatory variable in a multiple linear regression model if other explanatory variables are provided as well. Unfortunately, population is likely the only good independent variable provided in this data set.