Kenny Dinh

March 1, 2016

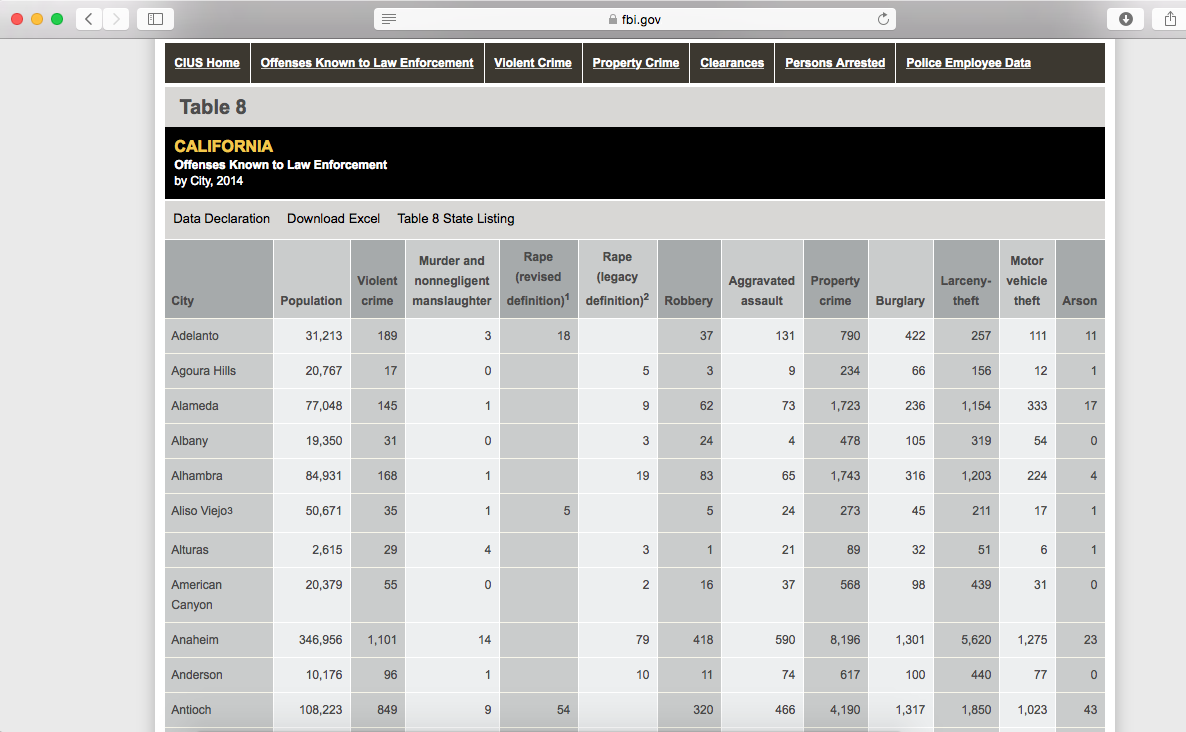
CIS 320

P. Thomas

**CIS 320 Data Analysis Project**

For this data analysis project I decided to do crime analysis on data from the majority cities in California. This data I’ve obtained from the fbi.gov. While navigating the website I’ve found a list of compiled datasets readily available to use and download, each dataset focused on a particular criteria. The Link below is where I found my data set. Data is generated and collected by using a UCR, which is the Uniform Crime Reporting. Furthermore into the data details. The Data is divided up and organized by cities in columns and crime type in rows. Based of this data I would like to try using R to analyze the data and come up with my own percentages for the cities in California ranging which one has the biggest population, highest crime rate, and also which crime occurs most and where.

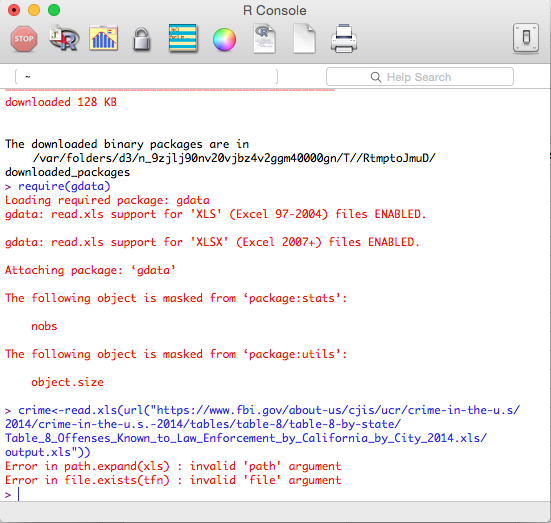
Based on this data analysis if possible to do correctly I would say this information could be useful for people to determine which areas they can avoid. I know this is also flawed due to the fact that the crimes can occur anywhere at anytime, but this not gives people an idea of where it occurs most and I’m sure law enforcement is already aware of this and uses this to its full potential.



<https://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/tables/table-8/table-8-by-state/Table_8_Offenses_Known_to_Law_Enforcement_by_California_by_City_2014.xls>

**R.**

I was not able to load the dataset in by using it’s URL. As I don’t fully understand R my best guess of why it is unable to load is that it is coming from a .xls file rather than a .csv. In order to load the data into R I simply just converted the file into a .csv.



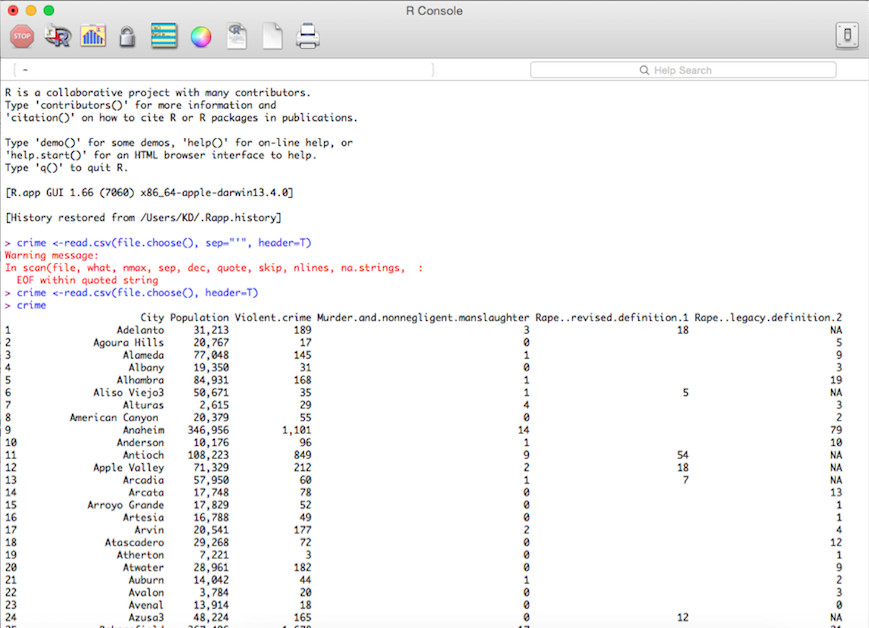
Once after loading the data file into R, which I’ve decided to name “Crime”. The data was inputted into R with only headers and due to my lack of knowledge with R I spent quite sometime figuring out how to format the Data correctly due to the fact that by data set had the first column as headers and as well as rows.

To load the data into R.

This command allowed me choose directly my file with a window screen since I was not able to load the URL into R.

> crime <-read.csv(file.choose(), header=T)

> crime

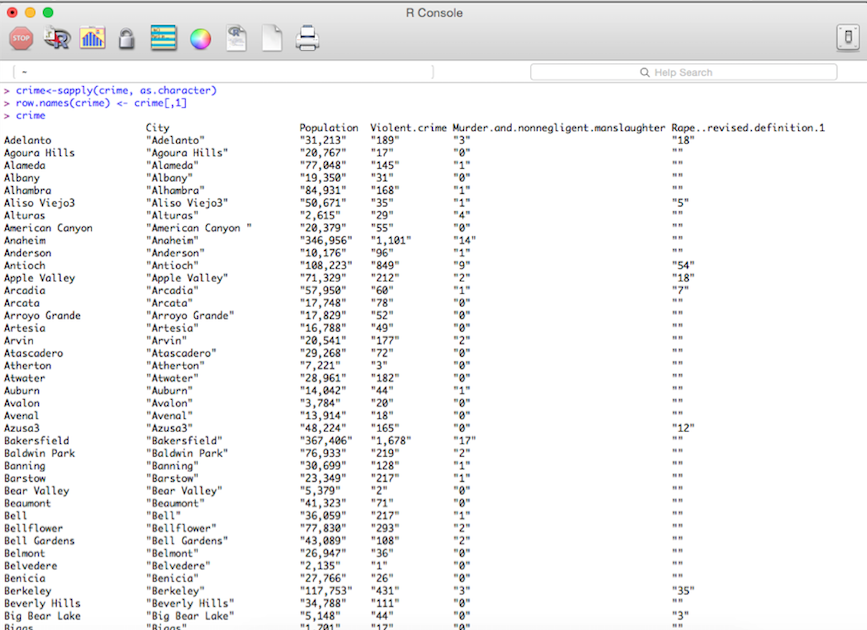


In order to display columns and rows as headers I used the following commands.

> crime<-sapply(crime, as.character)

> row.names(crime) <- crime[,1]

These commands will sort the city as a column on its own but will give me double columns with the same information.



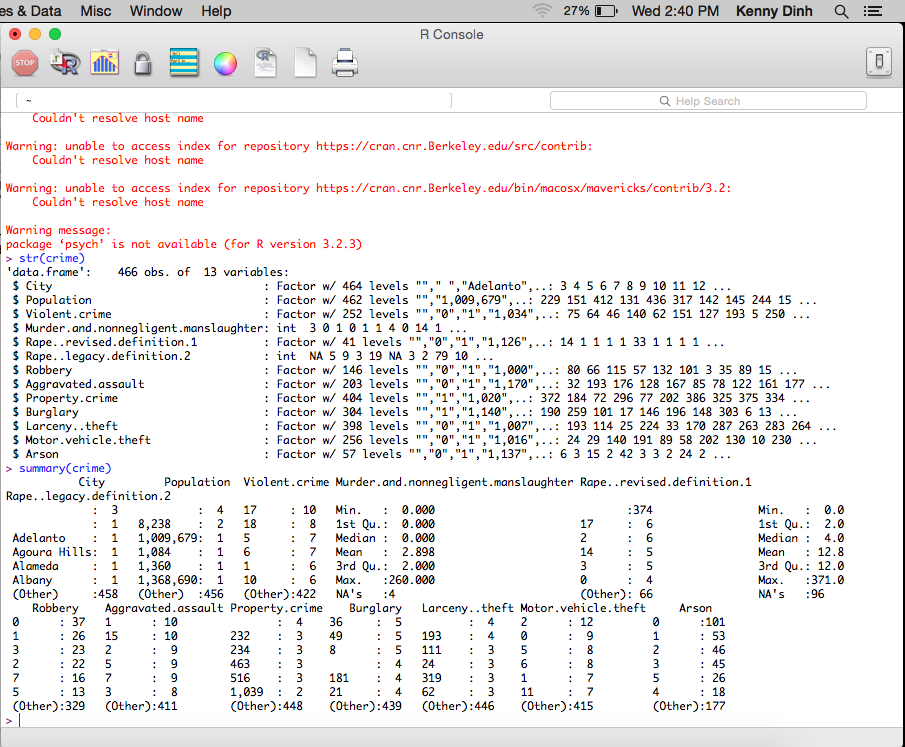
Once data is loaded into R I was able to form a summary and structure command. Reporting what the dataset consists of and what it reports.

Str(crime)

Summary(crime)

Using the str function it reported back the datasets headers and columns. As a new user of R and hardly knowing anything about the language, a such simple step provided me with information in which I found was very useful. Although it only states the obvious I thought it was worthy to post.

Next the summary function which only stated a couple of the cities and a few means and medians. This I couldn’t figure out why it only listed a few and not a complete list. I would imagine if it were listed completely it would be of tremendous use to analyze and evaluate crime rates.



Based on my research and reading upon other Crime Analysis Data report Using R. I am aware of some of the library packages listed, ggplot2, maps, and ggmap. One package that I was very interested in learning and which with research and watching a few videos based off “ggmap” I am aware that this is a very useful tool.

> library(ggplot2)

> library(maps)

> plot(crime)

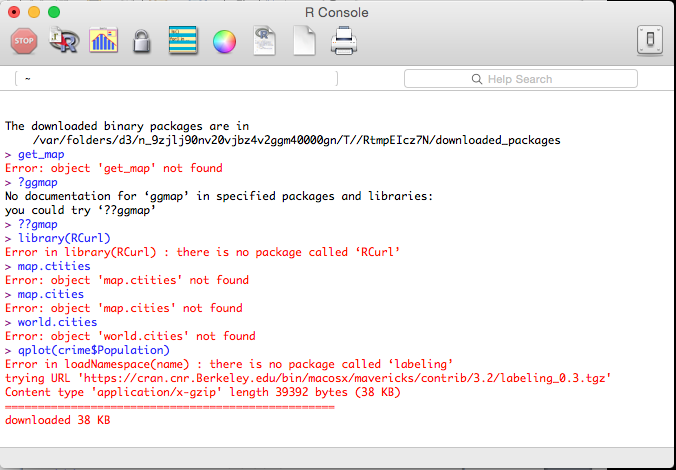
> smoothScatter(crime$Population)

The downloaded binary packages are in

/var/folders/d3/n\_9zjlj90nv20vjbz4v2ggm40000gn/T//RtmpEIcz7N/downloaded\_packages

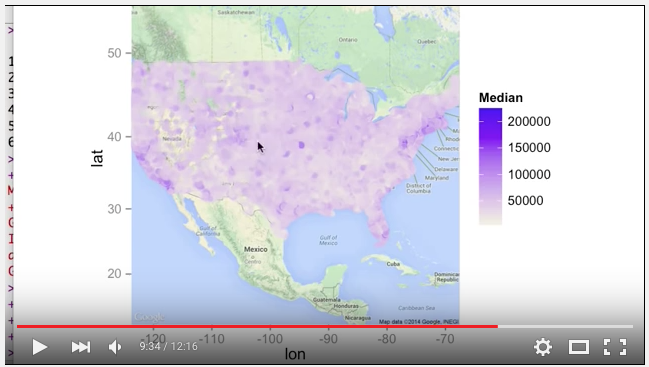
> get\_map

> library(RCurl)



As mentioned earlier the ggmap and ggplot 2 would be awesome if I was fortunate to get it running in R. What it does is that it generate and loads a map into R and plots points, cities, zip codes, and etc. into the map making it visible to see where the crime occurs. Based on a video I watched I tried to generate the same results but received errors as posted in the code above. If successful the code would look similar to this but only concentrated on California as the dataset only reports data based off California Cities.

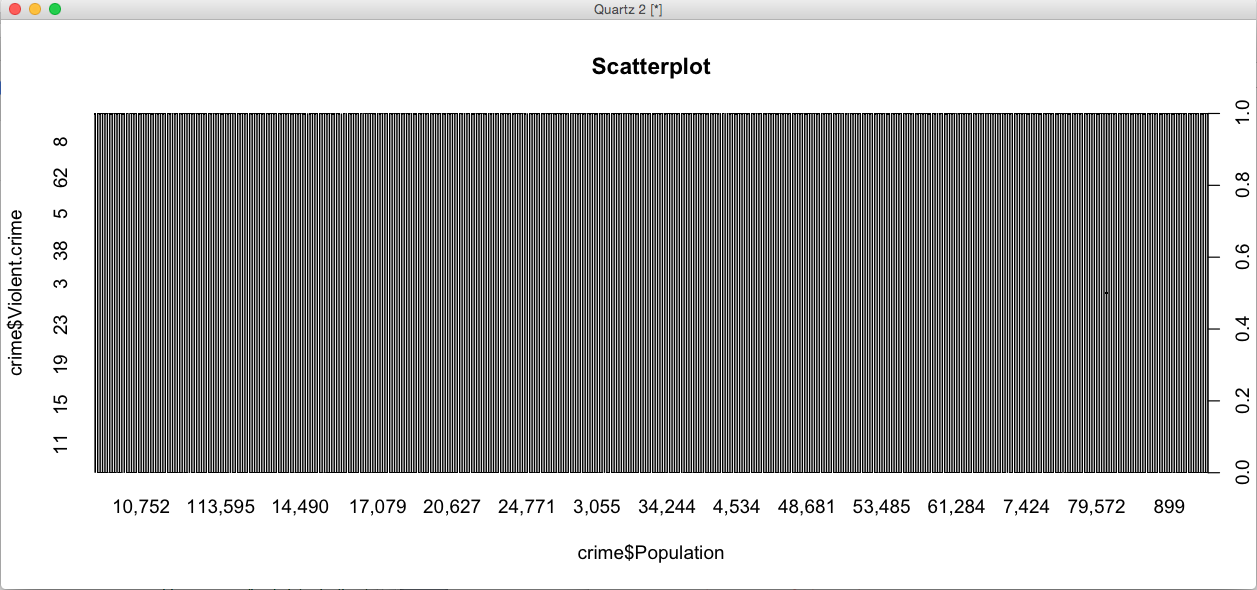
Picture from video: <https://www.youtube.com/watch?v=EtJ-iTZeqTg>



Moving forward there was a lot of code and functions I was not able to perform in R. Although I wish it wasn’t the case and would be nice to see R work for me. As I obviously need to learn more about R and its functions and utilities it would make this project a lot more understanding. Although I wasn’t able to Run R successfully I was able to run certain plots. The code for the plots and graphs I was able to run are below.

Using the code bloew gave me this graph. It is safe to saw that I Received the wrong results most likely due to my coding. But to my understanding it is suppose to show the comparison of the graphs. In this case showing Population in regards to Violent Crimes that has occurred. I believe that based of that I would be able to evaluate the percentage of the Violent crimes occurring within the given population.

> plot(crime$Population, crime$Violent.crime, main="Scatterplot", xlab="crime$Population", ylab="crime$Violent.crime")



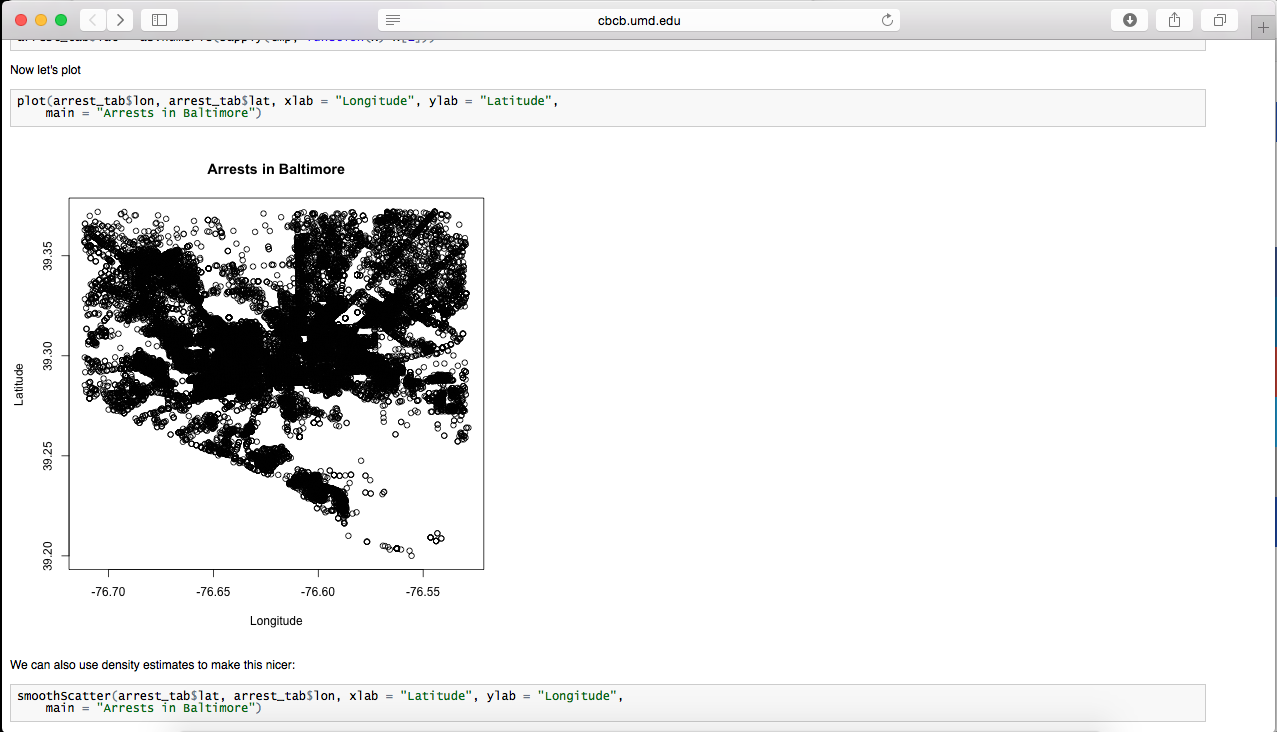
If I were to change the variable after “$” it would generate a different graph to whatever variable I choose to type in from the dataset. Another thing I learned from this project is that the only comparison graphs I received was using the graphs would list informative information. If I was able to load the maps in correctly I believe the maps would be very clear and informative.

**Graphs in R.**

Some of the graphs that I was able to generate were the plots and a smooth scatter plots. Off my some research in similar analysis reports I was able to understand and comprehend the logic behind these graphs. A useful analysis report was this the Baltimore Data Analysis with R. The link is below.

Here in this link it provided graphs and how they would’ve turned out if the maps were loaded into R correctly. Due to many errors I was not able to generate a map. But I was able to have graphs.

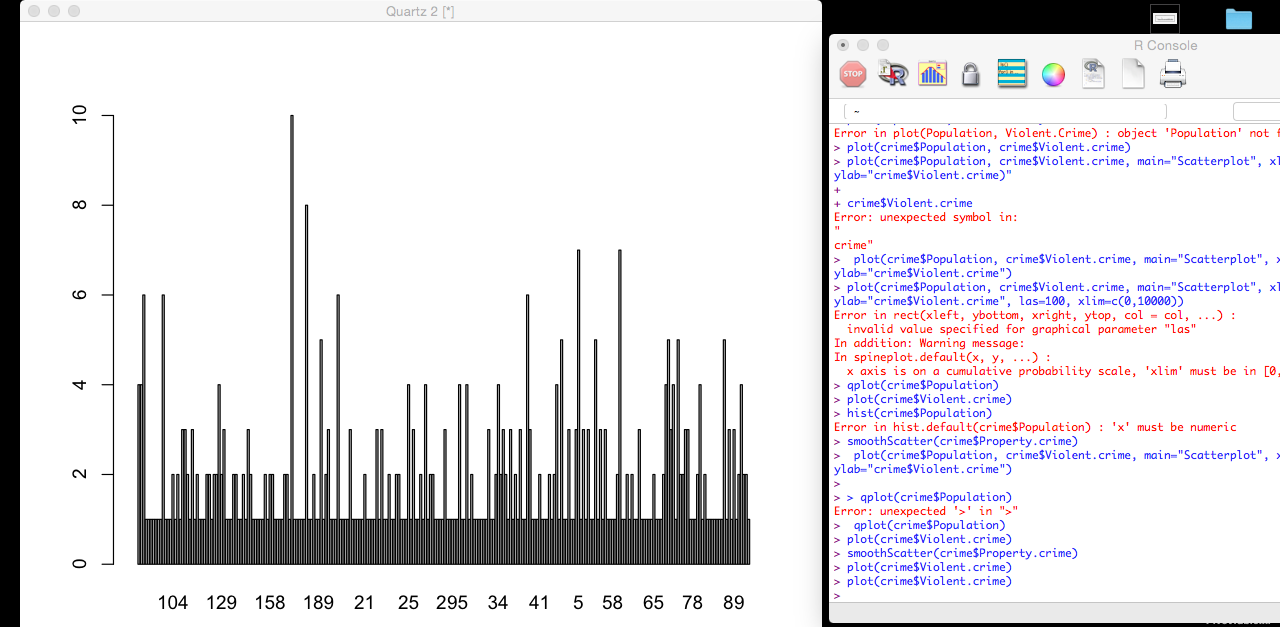
<http://cbcb.umd.edu/~hcorrada/CFG/lectures/lect20_Rintro/baltimore.html>



Using the graph and plots command listed below I was able to generate these.

plot(crime$Violent.crime)

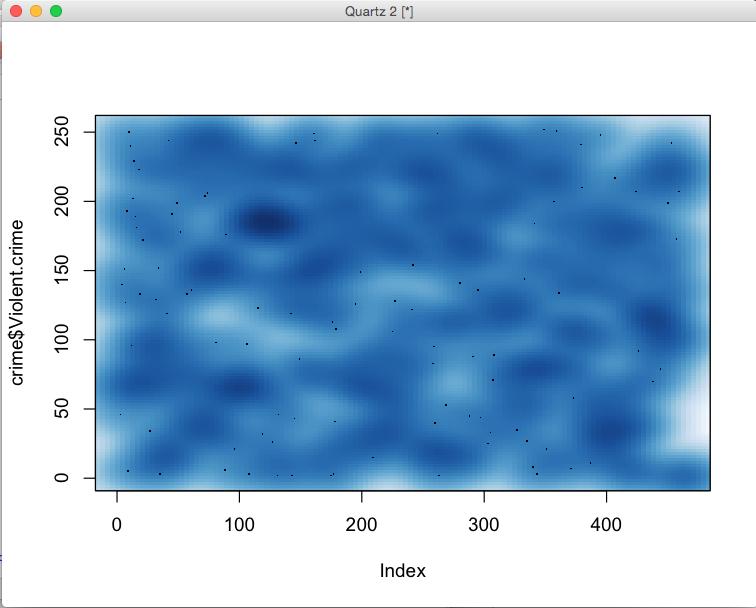
Below is the graph I was able to generate this based off of Violent. Crime. And shows how much it occurs within the graph



Moving on the scatter plots and graphs below.

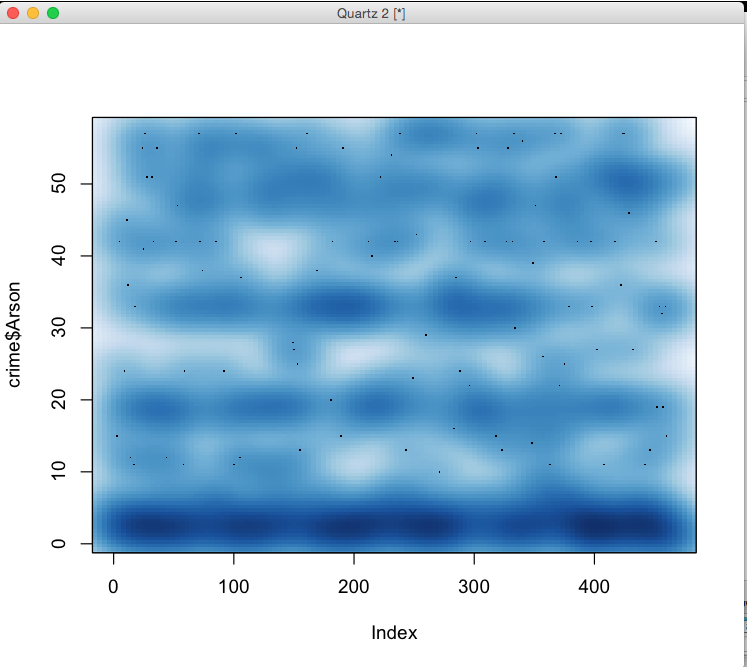
smoothScatter(crime$Violent.crime)

This shows the graphs of Violent.crimes



smoothScatter(crime$Arson)

This Graph shows the scatter graphs of Arson Crimes.



In conclusion of this project if I were able to generate the codes in R correctly and load packages in without failing I believe it would display more useful information and more straightforward graphs. Although the data was categorized into violent crimes, and non violent crimes, robbery, assault, and arson. My goal was to generate calculations and percentages of what occurs more. And out of a given population amount what is the percentage that a “certain” crime has occurred and what are the chances of it occurring again.