

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.

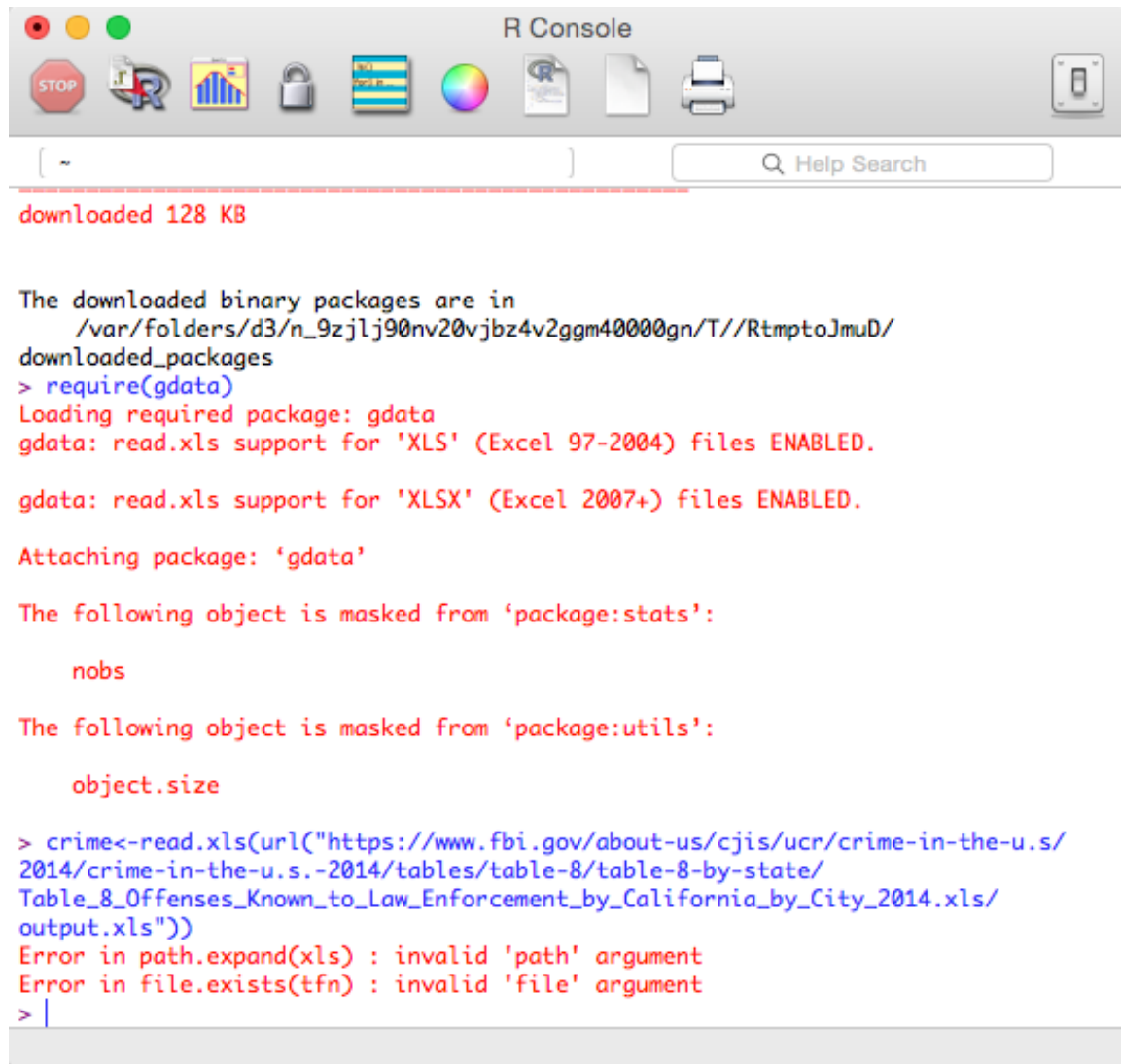
Based on this data analysis if possible to do correctly I would say this information could be useful for people to determine which areas they can avoid. I know this is also flawed due to the fact that the crimes can occur anywhere at anytime, but this not gives people an idea of where it occurs most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



```
R Console

downloaded 128 KB

The downloaded binary packages are in
/var/folders/d3/n_9zjlj90nv20vjbz4v2ggm40000gn/T//RtmptoJmuD/
downloaded_packages
> require(gdata)
Loading required package: gdata
gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.
gdata: read.xls support for 'XLSX' (Excel 2007+) files ENABLED.
Attaching package: 'gdata'

The following object is masked from 'package:stats':

  nobs

The following object is masked from 'package:utils':

  object.size

> crime<-read.xls(url("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/
output.xls"))
Error in path.expand(xls) : invalid 'path' argument
Error in file.exists(tfn) : invalid 'file' argument
> |
```

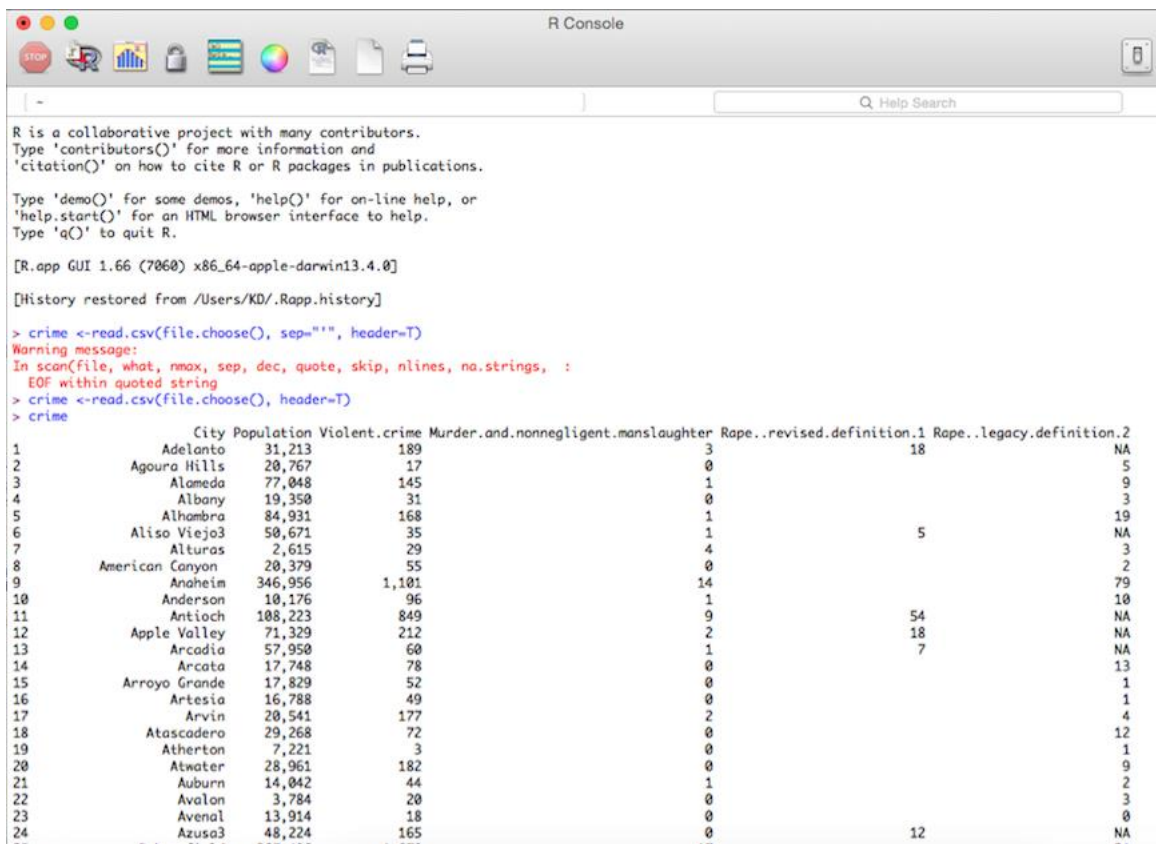
Once after loading the data file into R, which I've decided to name "Crime". The data was inputted into R with only headers and due to my lack of knowledge with R I spent quite sometime figuring out how to format the Data correctly due to the fact that by data set had the first column as headers and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



The screenshot shows an R Console window with the following content:

```
R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.66 (7060) x86_64-apple-darwin13.4.0]
[History restored from /Users/KD/.Rapp.history]

> crime <-read.csv(file.choose(), sep=";", header=T)
Warning message:
In scan(file, what, rmax, sep, dec, quote, skip, nlines, na.strings, :
EOF within quoted string
> crime <-read.csv(file.choose(), header=T)
> crime
```

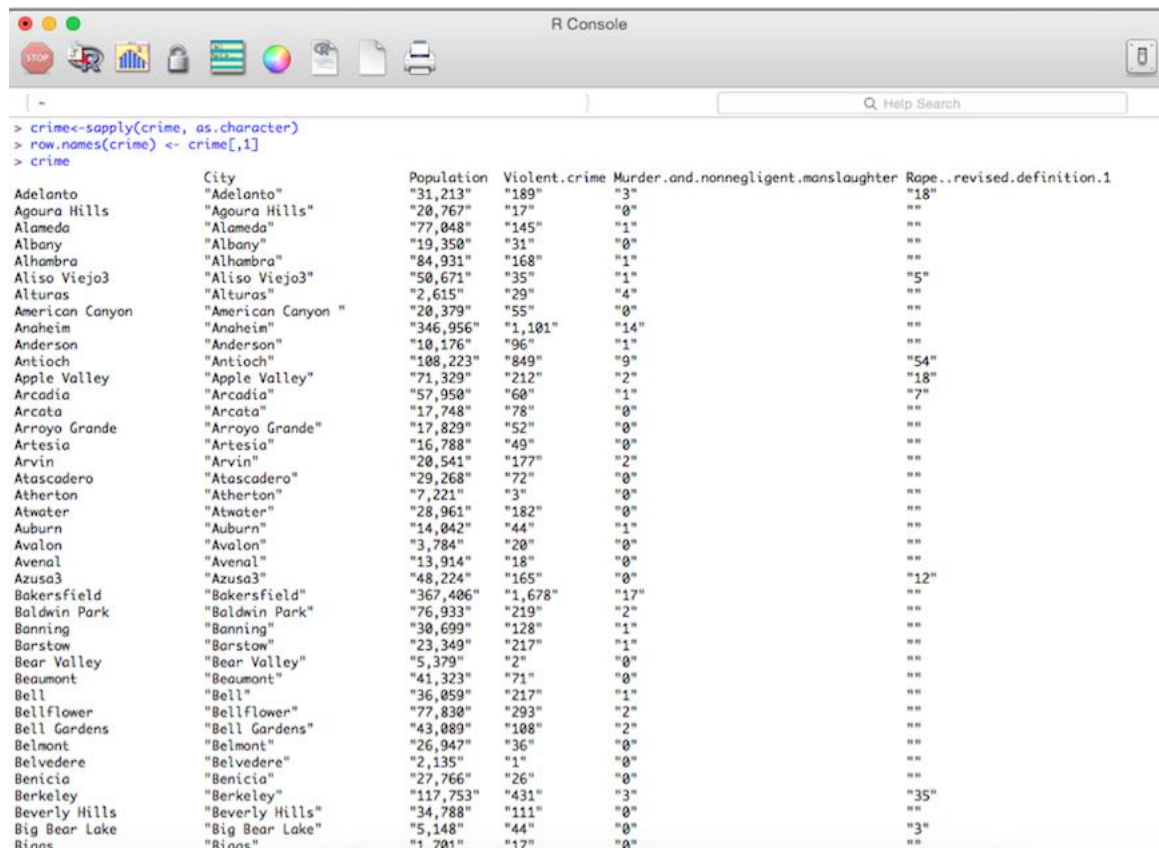
	City	Population	Violent.crime	Murder.and.nonnegligent.manslaughter	Rape..revised.definition.1	Rape..legacy.definition.2
1	Adelanto	31,213	189	3	18	NA
2	Agoura Hills	20,767	17	0		5
3	Alameda	77,048	145	1		9
4	Albany	19,350	31	0		3
5	Alhambra	84,931	168	1		19
6	Aliso Viejo3	50,671	35	1	5	NA
7	Alturas	2,615	29	4		3
8	American Canyon	20,379	55	0		2
9	Anaheim	346,956	1,101	14		79
10	Anderson	10,176	96	1		10
11	Antioch	108,223	849	9	54	NA
12	Apple Valley	71,329	212	2	18	NA
13	Arcadia	57,950	60	1	7	NA
14	Arcata	17,748	78	0		13
15	Arroyo Grande	17,829	52	0		1
16	Artesia	16,788	49	0		1
17	Arvin	20,541	177	2		4
18	Atascadero	29,268	72	0		12
19	Atherton	7,221	3	0		1
20	Atwater	28,961	182	0		9
21	Auburn	14,042	44	1		2
22	Avalon	3,784	20	0		3
23	Avenal	13,914	18	0		0
24	Azusar3	48,224	165	0	12	NA

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.



City	Population	Violent.crime	Murder.and.nonnegligent.manslaughter	Rape..revised.definition.1
Adelanto	"31,213"	"189"	"3"	"18"
Agoura Hills	"20,767"	"17"	"0"	"0"
Alameda	"77,048"	"145"	"1"	"0"
Albany	"19,350"	"31"	"0"	"0"
Alhambra	"84,931"	"168"	"1"	"0"
Aliso Viejo3	"50,671"	"35"	"1"	"5"
Alturas	"2,615"	"29"	"4"	"0"
American Canyon	"20,379"	"55"	"0"	"0"
Anaheim	"346,956"	"1,101"	"14"	"0"
Anderson	"10,176"	"96"	"1"	"0"
Antioch	"108,223"	"849"	"9"	"54"
Apple Valley	"71,329"	"212"	"2"	"18"
Arcadia	"57,950"	"60"	"1"	"7"
Arcata	"17,748"	"78"	"0"	"0"
Arroyo Grande	"17,829"	"52"	"0"	"0"
Artesia	"16,788"	"49"	"0"	"0"
Arvin	"20,541"	"177"	"2"	"0"
Atascadero	"29,268"	"72"	"0"	"0"
Atherton	"7,221"	"3"	"0"	"0"
Atwater	"28,961"	"182"	"0"	"0"
Auburn	"14,042"	"44"	"1"	"0"
Avalon	"3,784"	"20"	"0"	"0"
Avenal	"13,914"	"18"	"0"	"0"
Azusa3	"48,224"	"165"	"0"	"12"
Bakersfield	"367,406"	"1,678"	"17"	"0"
Baldwin Park	"76,933"	"219"	"2"	"0"
Banning	"30,699"	"128"	"1"	"0"
Barstow	"23,349"	"217"	"1"	"0"
Bear Valley	"5,379"	"2"	"0"	"0"
Beaumont	"41,323"	"71"	"0"	"0"
Bell	"36,059"	"217"	"1"	"0"
Bellflower	"77,830"	"293"	"2"	"0"
Bell Gardens	"43,089"	"108"	"2"	"0"
Belmont	"26,947"	"36"	"0"	"0"
Belvedere	"2,135"	"1"	"0"	"0"
Benicia	"27,766"	"26"	"0"	"0"
Berkeley	"117,753"	"431"	"3"	"35"
Beverly Hills	"34,788"	"111"	"0"	"0"
Big Bear Lake	"5,148"	"44"	"0"	"3"
Blaine	"1,701"	"17"	"0"	"0"

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.

```
es & Data Misc Window Help 27% Wed 2:40 PM Kenny Dinh
R Console

Warning: Couldn't resolve host name
Warning: unable to access index for repository https://cran.cnr.Berkeley.edu/src/contrib:
Warning: Couldn't resolve host name
Warning: unable to access index for repository https://cran.cnr.Berkeley.edu/bin/macosx/mavericks/contrib/3.2:
Warning: Couldn't resolve host name
Warning message:
package 'psych' is not available (for R version 3.2.3)
> str(crime)
'data.frame': 466 obs. of 13 variables:
 $ City : Factor w/ 464 levels "", "Adelanto",...: 3 4 5 6 7 8 9 10 11 12 ...
 $ Population : Factor w/ 462 levels "", "1,009,679",...: 229 151 412 131 436 317 142 145 244 15 ...
 $ Violent.crime : Factor w/ 252 levels "", "0", "1", "1,034",...: 75 64 46 140 62 151 127 193 5 250 ...
 $ Murder.and.nonnegligent.manslaughter: int 3 0 1 0 1 1 4 0 14 1 ...
 $ Rape..revised.definition.1 : Factor w/ 41 levels "", "0", "1", "1,126",...: 14 1 1 1 1 33 1 1 1 1 ...
 $ Rape..legacy.definition.2 : int NA 5 9 3 19 NA 3 2 79 10 ...
 $ Robbery : Factor w/ 146 levels "", "0", "1", "1,000",...: 80 66 115 57 132 101 3 35 89 15 ...
 $ Aggravated.assault : Factor w/ 203 levels "", "0", "1", "1,170",...: 32 193 176 128 167 85 78 122 161 177 ...
 $ Property.crime : Factor w/ 404 levels "", "1", "1,020",...: 372 184 72 296 77 202 386 325 375 334 ...
 $ Burglary : Factor w/ 304 levels "", "1", "1,140",...: 190 259 101 17 146 196 148 303 6 13 ...
 $ Larceny..theft : Factor w/ 398 levels "", "0", "1", "1,007",...: 193 114 25 224 33 170 287 263 283 264 ...
 $ Motor.vehicle.theft : Factor w/ 256 levels "", "0", "1", "1,016",...: 24 29 140 191 89 58 202 130 10 230 ...
 $ Arson : Factor w/ 57 levels "", "0", "1", "1,137",...: 6 3 15 2 42 3 3 2 24 2 ...

> summary(crime)
      City      Population Violent.crime Murder.and.nonnegligent.manslaughter Rape..revised.definition.1
Rape..legacy.definition.2
  : 3      : 4      : 17      : 10      Min. : 0.000      :374      Min. : 0.0
  : 1 8,238 : 2      : 18      : 8      1st Qu.: 0.000      : 17      1st Qu.: 2.0
  : 1 1,009,679: 1      : 5      : 7      Median: 0.000      : 2       Median: 4.0
Adelanto : 1 1,084 : 1      : 6      : 7      Mean : 2.898      : 14       Mean : 12.8
Agoura Hills: 1 1,360 : 1      : 1      : 6      3rd Qu.: 2.000      : 3        3rd Qu.: 12.0
Alameda : 1 1,368,690: 1      : 10     : 6      Max. : 260.000      : 0         Max. : 371.0
(Other) : 458 (Other): 456 (Other): 422 NA's : 4      (Other): 66      NA's : 96

  Robbery Aggravated.assault Property.crime Burglary Larceny..theft Motor.vehicle.theft Arson
0 : 37 1 : 10 : 4 36 : 5 : 4 2 : 12 : 0 : 101
1 : 26 15 : 10 : 232 : 3 49 : 5 193 : 4 0 : 9 : 1 : 53
3 : 23 2 : 9 : 234 : 3 8 : 5 111 : 3 5 : 8 : 2 : 46
2 : 22 5 : 9 : 463 : 3 : 4 24 : 3 6 : 8 : 3 : 45
7 : 16 7 : 9 : 516 : 3 181 : 4 319 : 3 1 : 7 : 5 : 26
5 : 13 3 : 8 : 1,039 : 2 21 : 4 62 : 3 11 : 7 : 4 : 18
(Other): 329 (Other): 411 (Other): 448 (Other): 439 (Other): 446 (Other): 415 (Other): 177
> |
```

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)
```

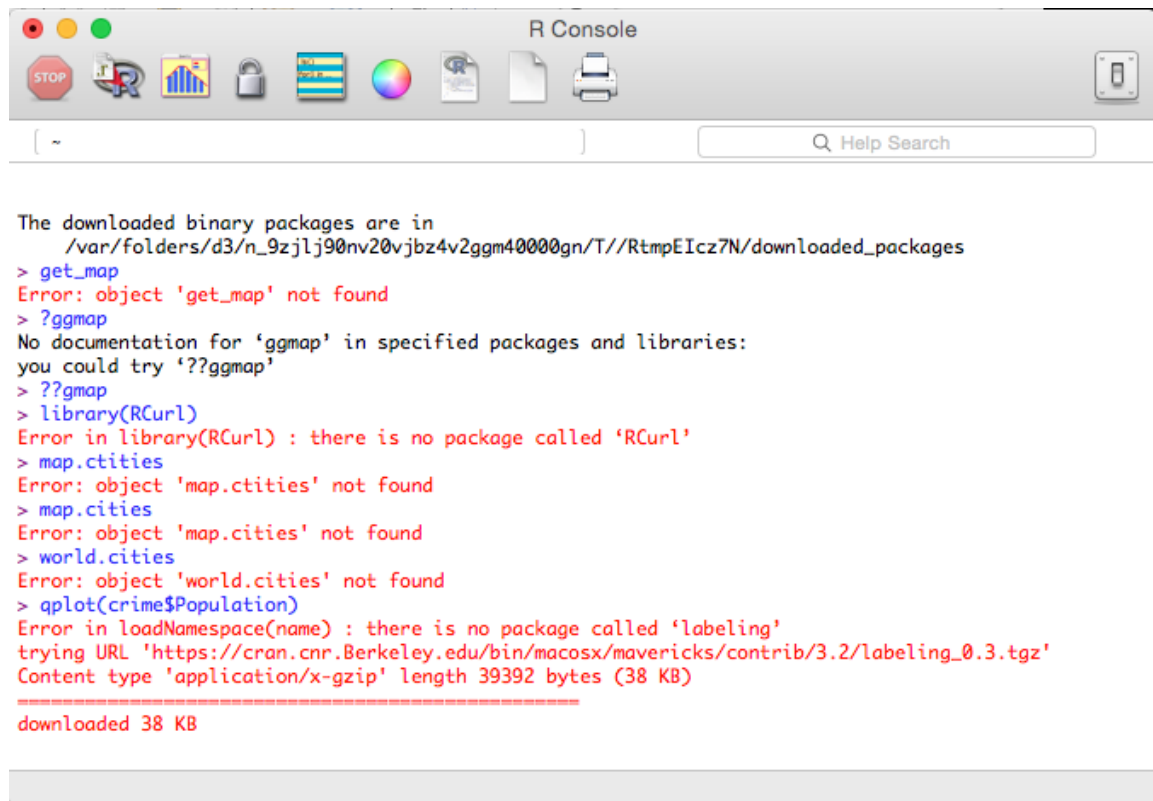
```
> smoothScatter(crime$Population)
```

The downloaded binary packages are in

/var/folders/d3/n_9zjlj90nv20vjbz4v2ggm40000gn/T//RtmpEIcz7N/downloaded_packages

```
> get_map
```

```
> library(RCurl)
```

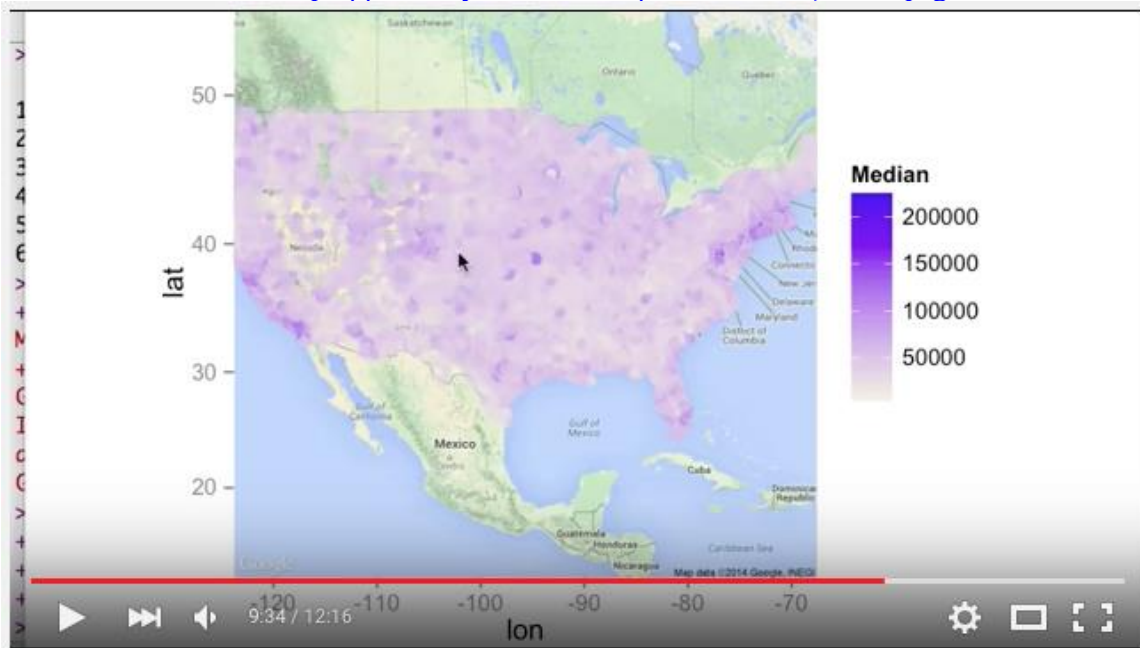


```
R Console

The downloaded binary packages are in
/var/folders/d3/n_9zjlj90nv20vjbz4v2ggm40000gn/T//RtmpEIcz7N/downloaded_packages
> get_map
Error: object 'get_map' not found
> ?ggmap
No documentation for 'ggmap' in specified packages and libraries:
you could try '??ggmap'
> ??ggmap
> library(RCurl)
Error in library(RCurl) : there is no package called 'RCurl'
> map.ctities
Error: object 'map.ctities' not found
> map.cities
Error: object 'map.cities' not found
> world.cities
Error: object 'world.cities' not found
> qplot(crime$Population)
Error in loadNamespace(name) : there is no package called 'labeling'
trying URL 'https://cran.cnr.Berkeley.edu/bin/macosx/mavericks/contrib/3.2/labeling_0.3.tgz'
Content type 'application/x-gzip' length 39392 bytes (38 KB)
downloaded 38 KB
```

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.

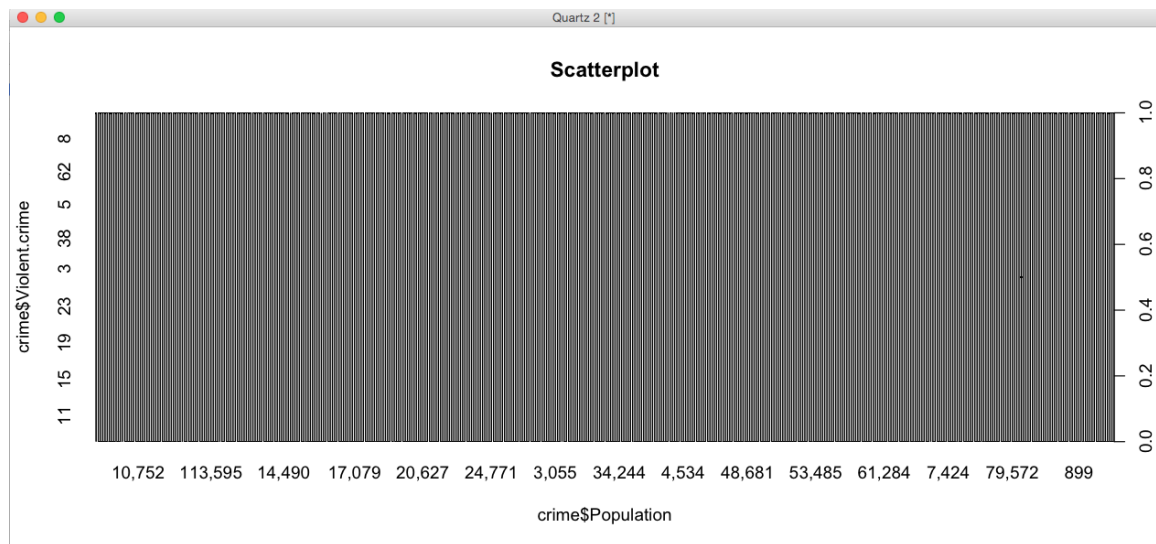
Picture from video: <https://www.youtube.com/watch?v=Etj-iTZeqTg>



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 below gave me this graph. It is safe to say 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")
```



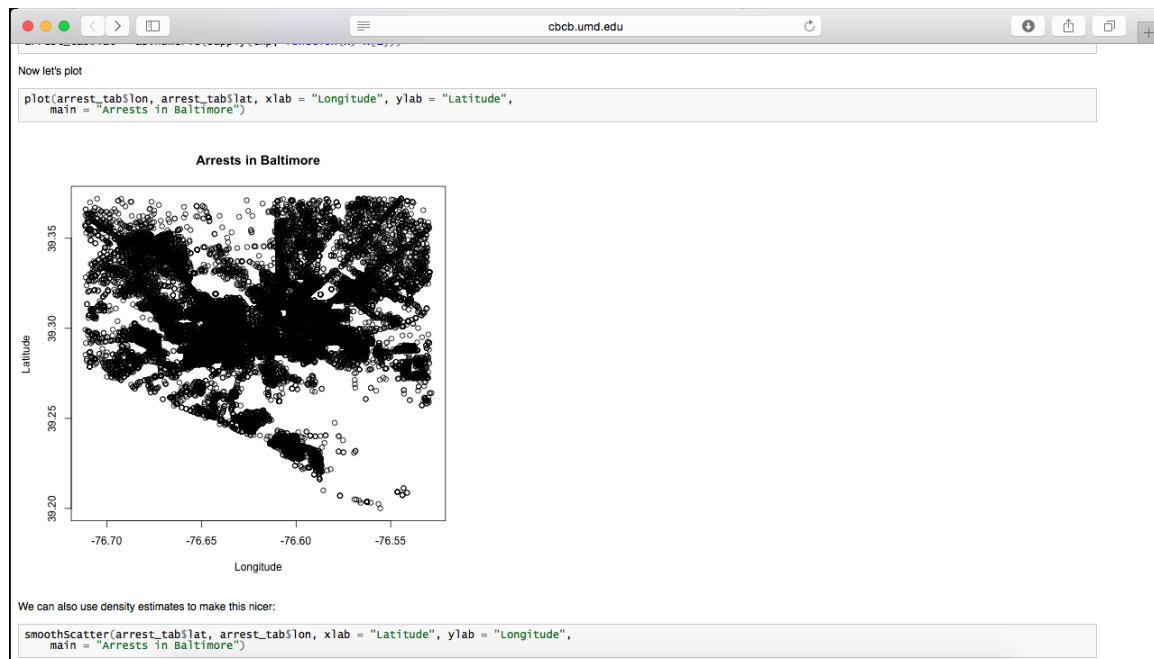
If I were to change the variable after "\$" it would generate a different graph to whatever variable I choose to type in from the dataset. Another thing I learned from this project is that the only comparison graphs I received was using the graphs would list informative information. If I was able to 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.

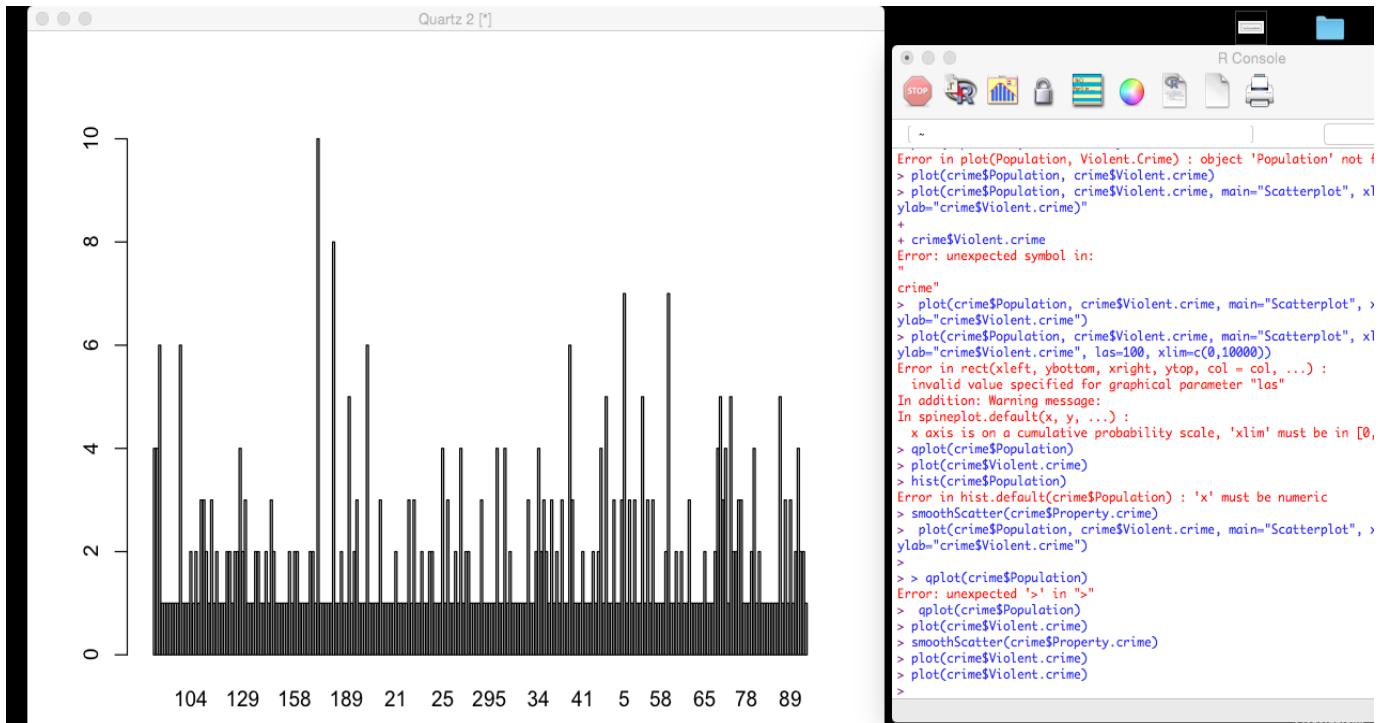
<http://cbcb.umd.edu/~hcorrada/CFG/lectures/lect20 Rintro/baltimore.html>



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

```
plot(crime$Violent.crime)
```

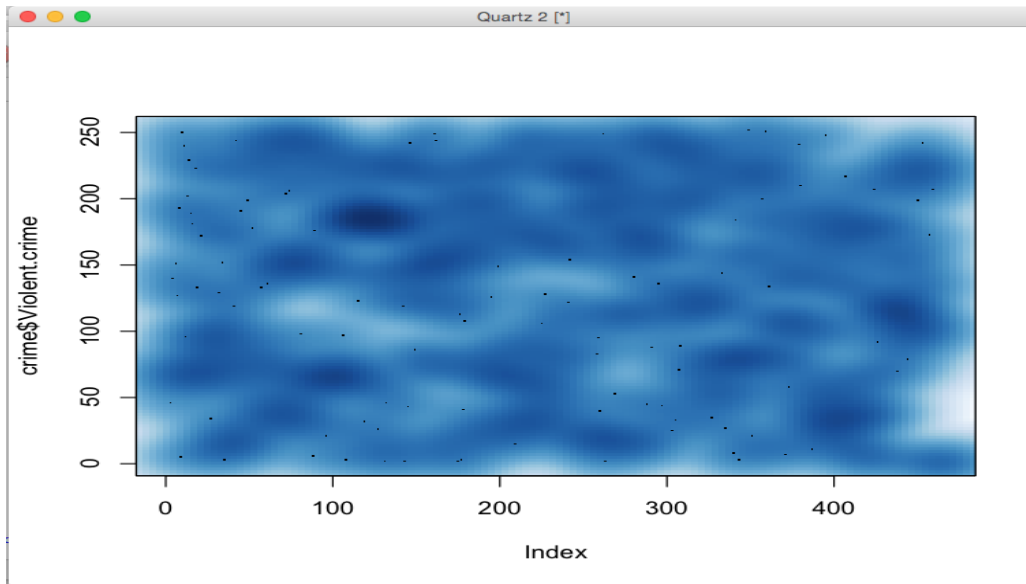
Below is the graph I was able to generate this based off of Violent.
Crime. And shows how much it occurs within the graph



Moving on the scatter plots and graphs below.

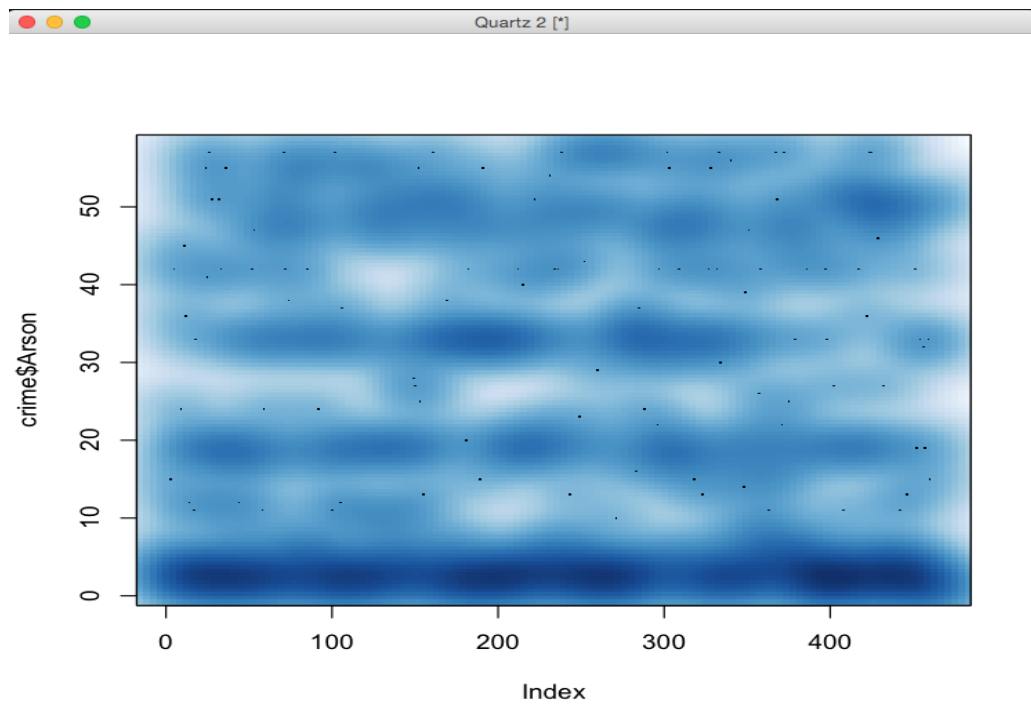
```
smoothScatter(crime$Violent.crime)
```

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.