**Study of the Relationship Among**

**Crime Rate, Educational Attainment, and Household Income in the United States of America**

**Introduction:**

In recent years, public safety became a rising global problem, and the problem is getting worse especially in the U.S. This study is trying to reveal the potential cause(s) of the booming crime rate in the U.S. and the possible way to lower it. This study is highly based on R language with Shiny.

Initially, three questions (assumptions) were raised as the first step of this study and the guideline of the early data acquisition process. The three questions raised are:

* What is the relationship between Crime Rate, Educational Attainment, and Household Income?
* Is the relationship discovered in question one can always be applied (across states/counties)?
* Between Household Income and Educational Attainment, which one of them affects more on the Crime Rate?

**Data Wrangling:**

Data Acquisition

To study this issue and try to answer the questions raised above, below datasets were acquired:

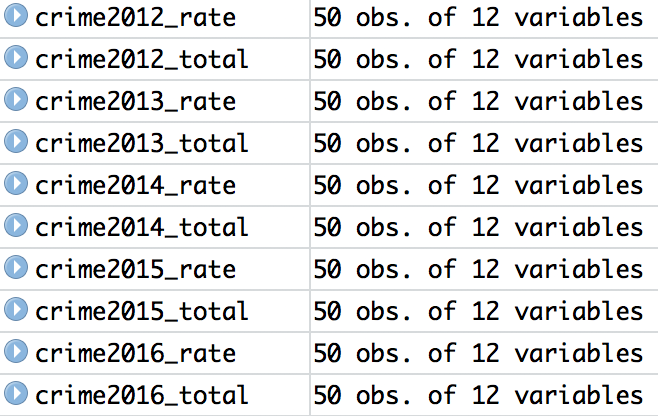
* “Crime in the United States” by FBI, year 2012 - 2016, tabular data, 502 rows \* 13 columns for each year. Contains spatial data of crimes reported by type of crimes in each state of the United States. (<https://ucr.fbi.gov/crime-in-the-u.s/>)
* “EDUCATIONAL ATTAINMENT, American Community Survey 1-Year Estimates” by United States Census Bureau, year 2012 – 2016, tabular data, 54 rows \* 18 columns for the year 2012-2014, 54 rows \* 33 columns for the year 2015 and 2016, contains spatial data for average educational attainment report by age group in each state of the United States. (<https://factfinder.census.gov/>)
* “INCOME IN THE PAST 12 MONTHS, American Community Survey 1-Year Estimates” by United States Census Bureau, year 2012 – 2016, tabular data, 54 rows \* 5 columns for each year, contains spatial data of average household income report by age group in each state of the United States. (<https://factfinder.census.gov/>)

Data Cleaning and Checking

The “Crime in the United States” was .xls files, to convert the files into proper format, they were loaded into Tableau and “Data Interpreter” was applied. Then the files were exported as .csv files.

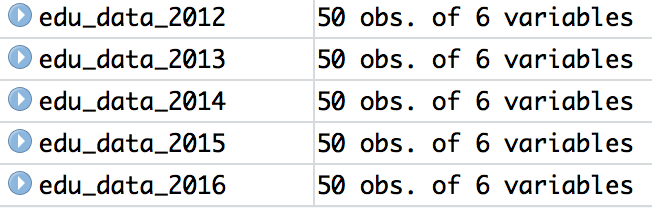
After .csv files were obtained, several additional wrangling was performed:

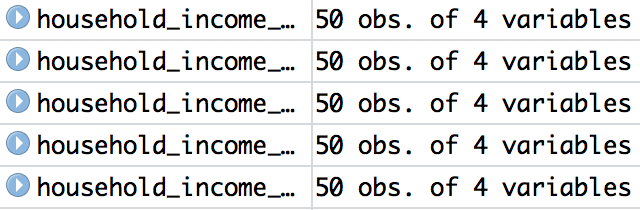
* Identified data inconsistency on column names, fixed by using “colnames” function.
* Numbers that were foot notes in the original data was found in “State” column that is supposed to be string, “gsub” function was used to fix this issue
* Values in “State” columns were converted to lower case by using “sapply” and “tolower” functions.
* Useless columns were removed by applying “subset” function.
* Useless rows were revmoed by applying “df[df$colname == “toberemoved”,]” function to the dataset.
* Each table was split into two different tables separately contains crime rate and crime total data.



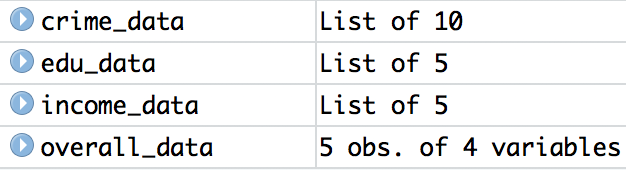
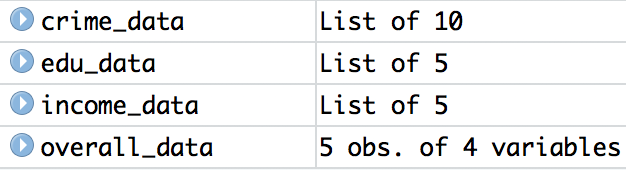
The education and household income datasets were in .csv format, so they were directly loaded into RStudio, and similar wrangling was performed on them.

Additionally, after performed some online about general data used to perform analysis similar to the topic, only few columns of education and household income data was kept, including the ones been used in the final analysis.





After the first data wrangling was completed, and further data checking was performed, the 20 datasets above were transformed to 3 lists, and stored into a RData file with an additional list containing the means of each dataset.

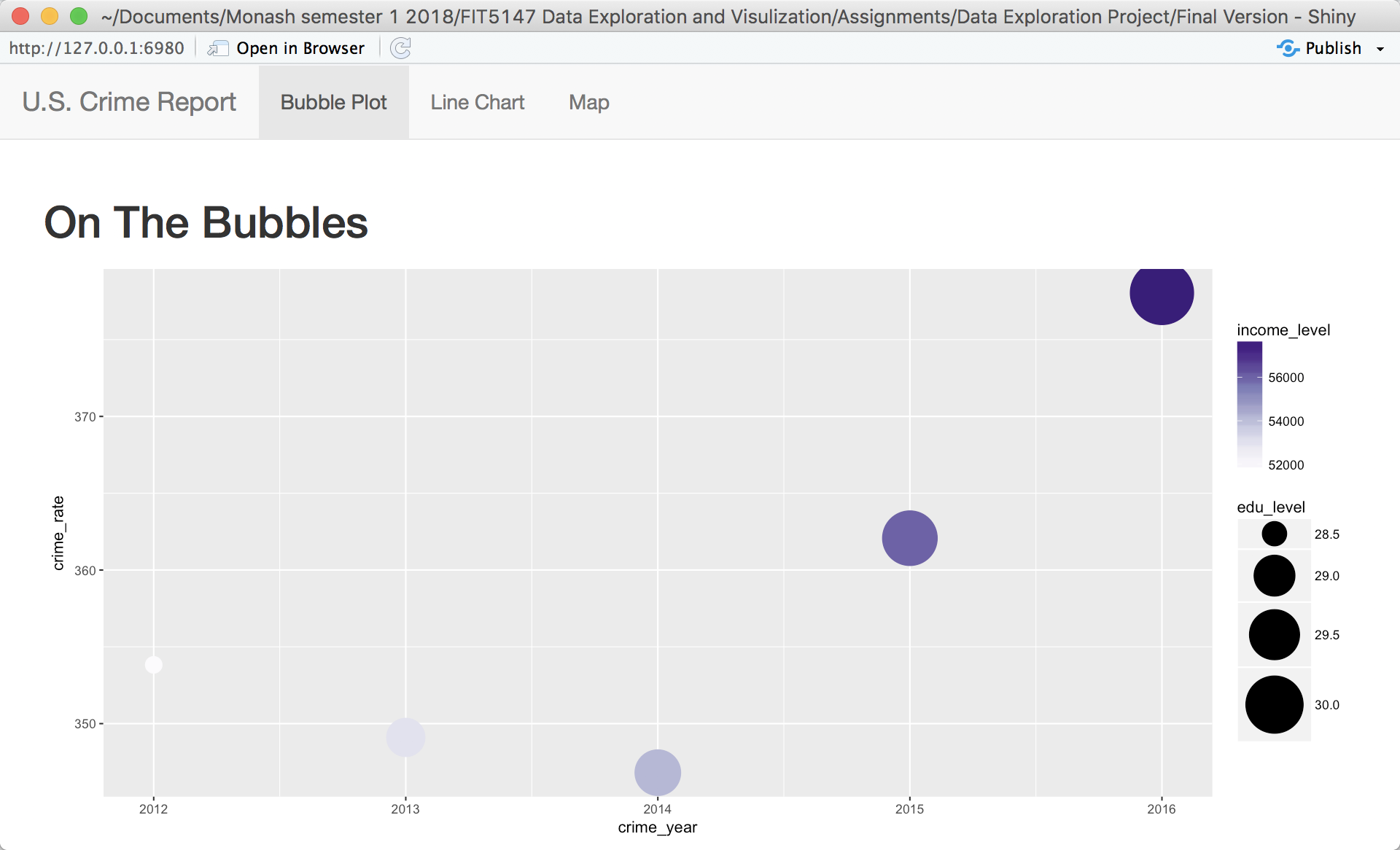


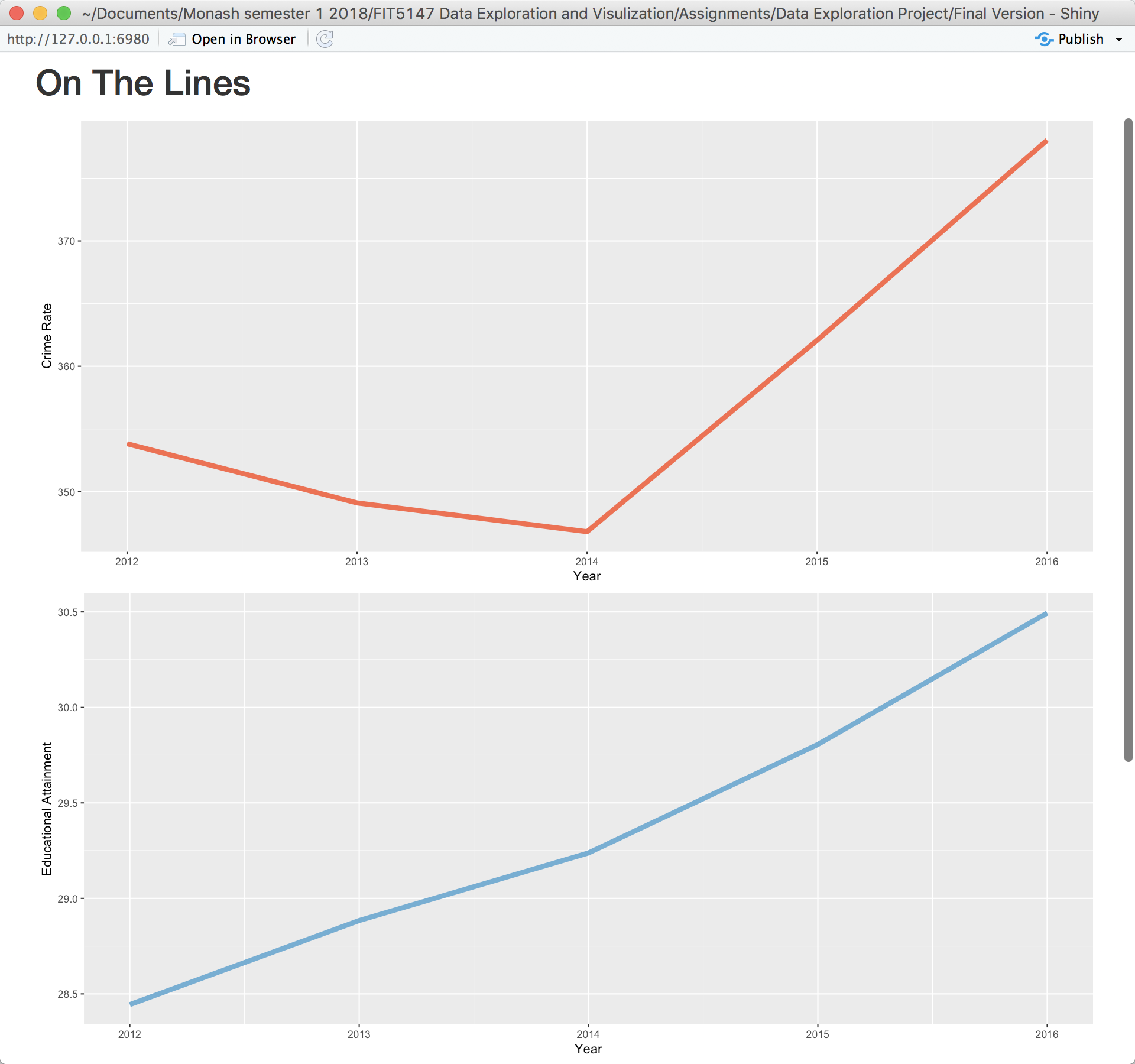
**Data Exploration and Visualization**

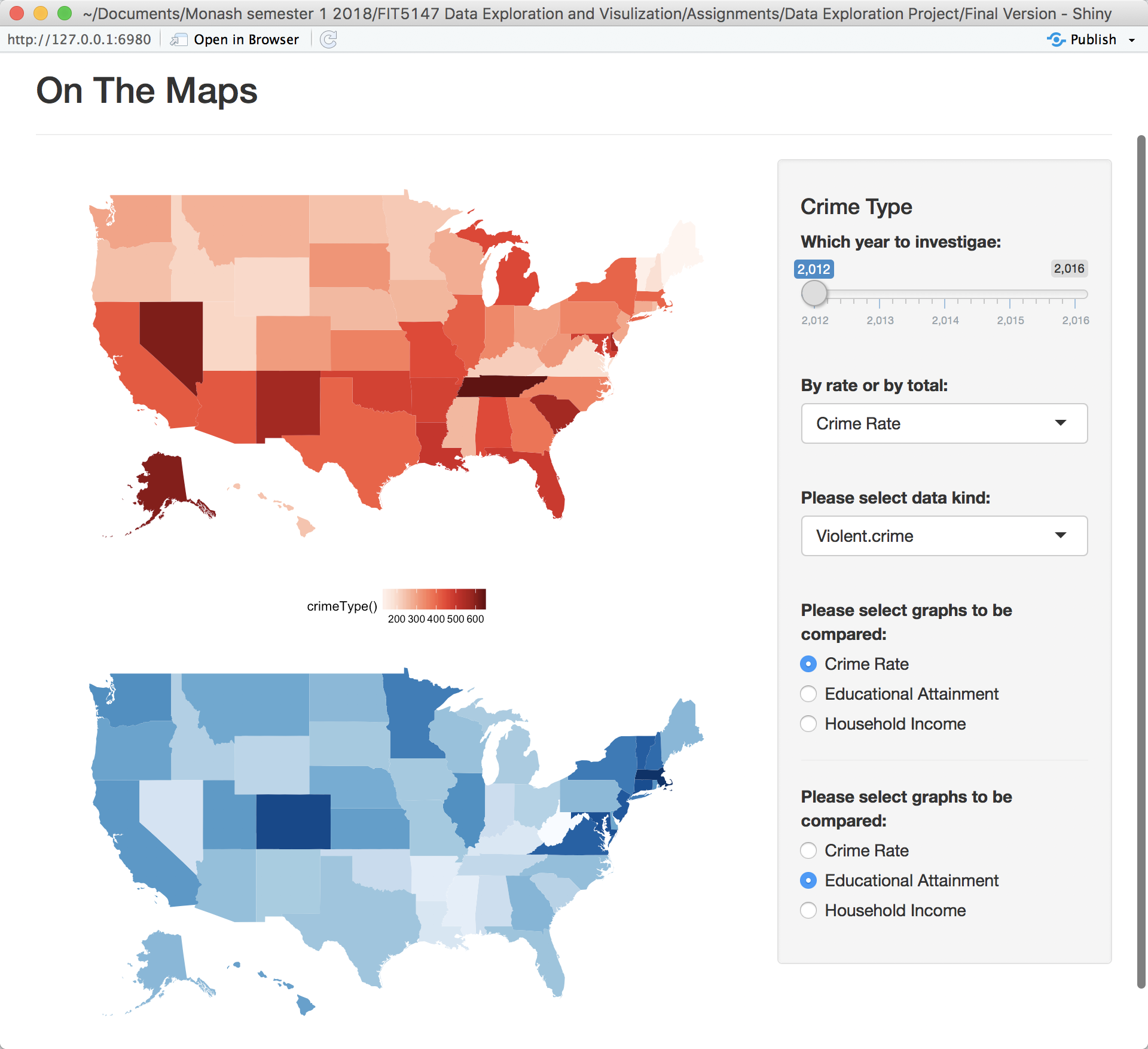
To perform proper data exploration and visualization for this study, Shiny with ggplot2 and fiftystater (a mapping library that can product a better map output comparing to ggplot2) in the environment of RStudio was chosen.

* “navbarPage” was selected since there will be multiple plots been displayed.
* “sliderInput” was used to enable year selection.
* “selectInput” was used to enable crime type and crime data type selection.
* “radioButtons” was used to enable selection of graphs to be compared.
* “reactive” was used to the need of dataset change upon the dynamic user inputs.

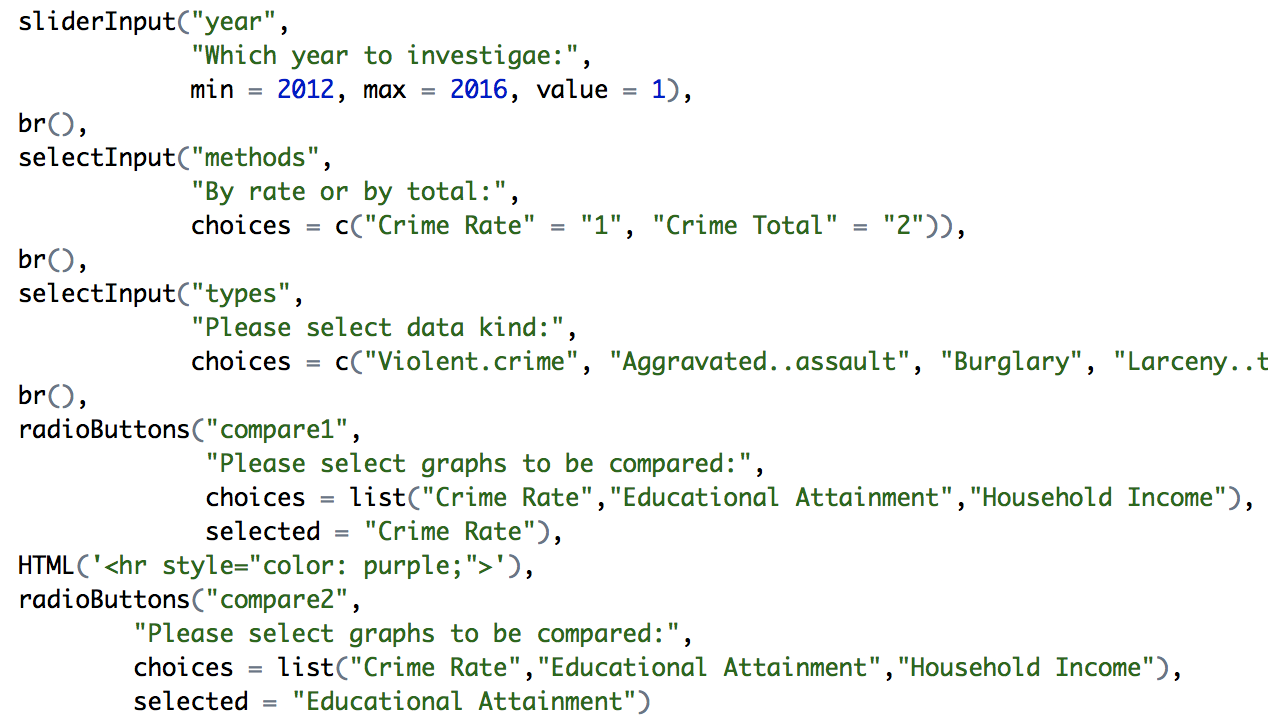
Below visualizations were generated upon execution of the Shiny app.







In the “On The Maps” page, dynamic input was adopted for changing the heatmaps by user’s interest. Users are able to change the data source, which heatmaps to be compared, which type of crime is visualized, and which kind of data to be compared (crime total or crime rate).



(code snippet from ui.R)

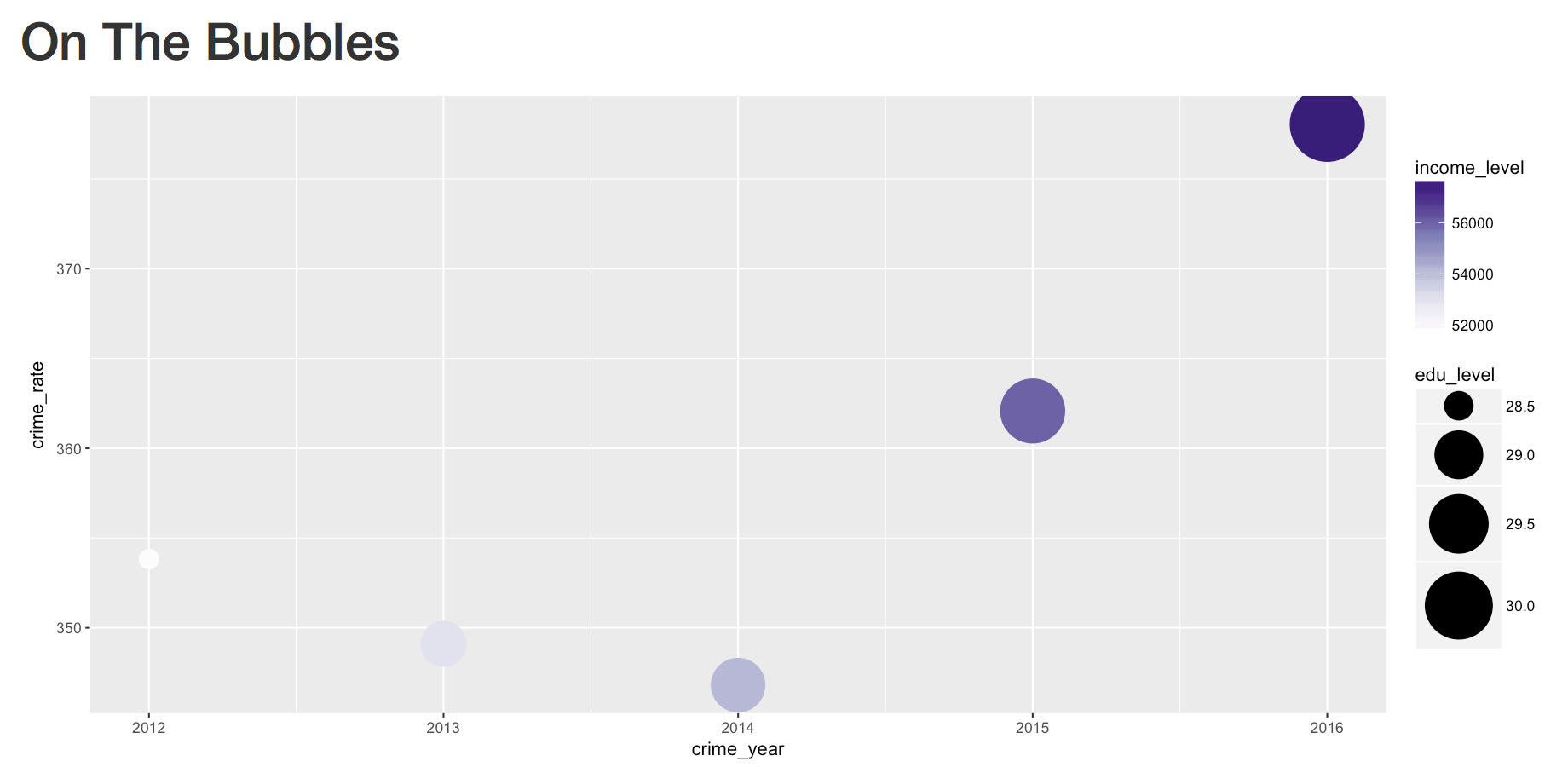


(code snippet from server.R)

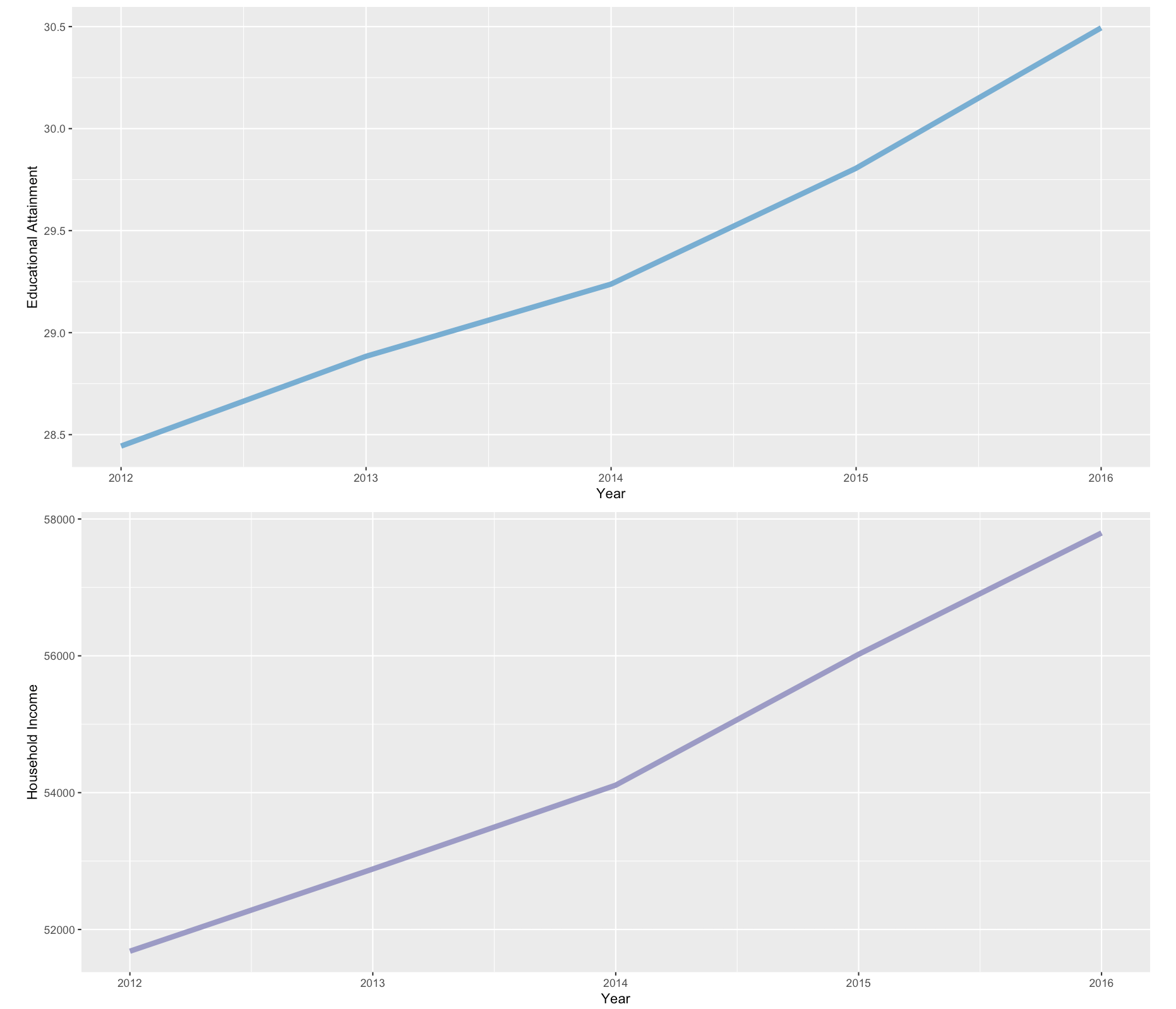
Dynamic input from “selectInput”, “sliderInput”, and “radioButtons” are enabled by using the “reactive” method in the server.R file.

**Conclusion:**

By studying the bubble chart of the treated “across the U.S.” data, it can be discovered that household income and educational attainment are in a positive relation, when the relationship between crime rate and household income or the relationship between crime rate and educational attainment are hard to be identified.

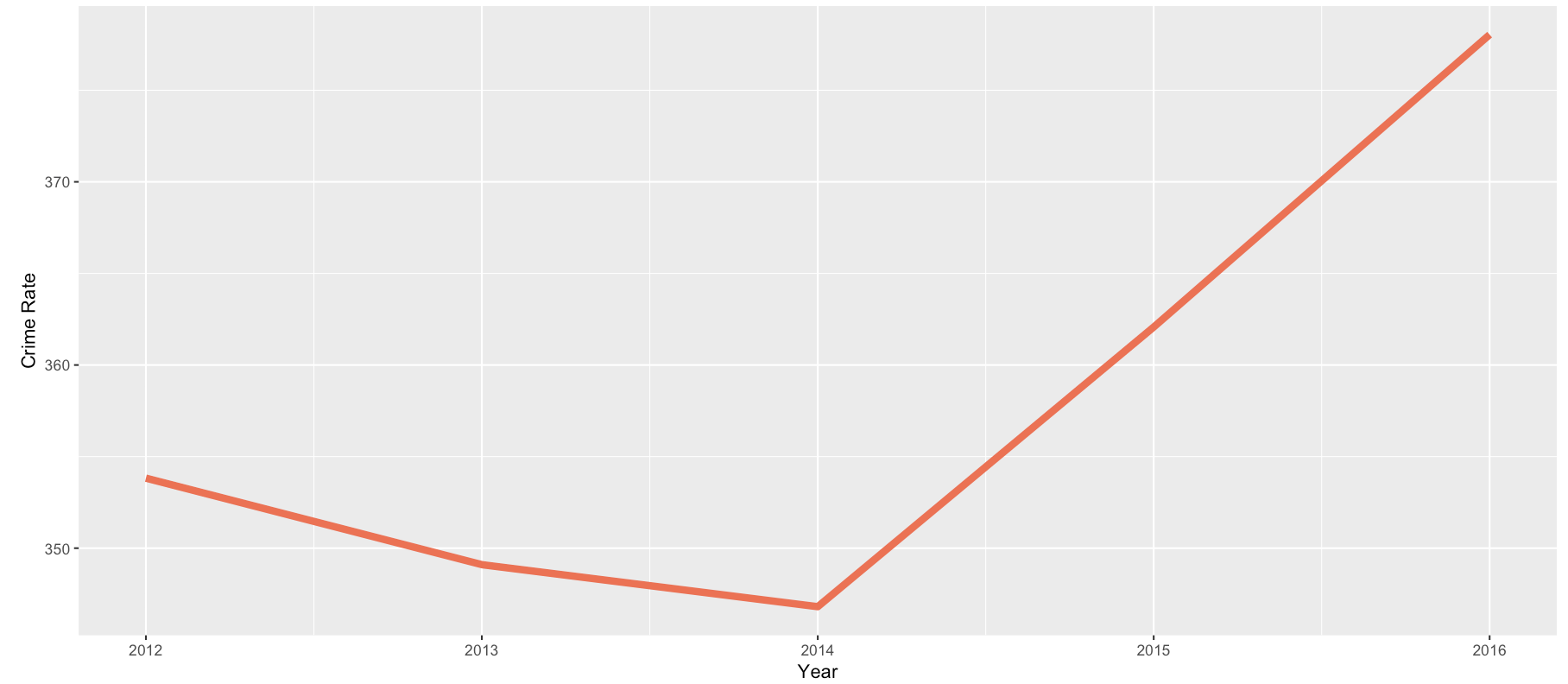


A clearer relationship could be identified by looking at the line charts. The line charts of household income and educational attainment across are almost identical, even a minor fluctuation on the year of 2014 are the same.

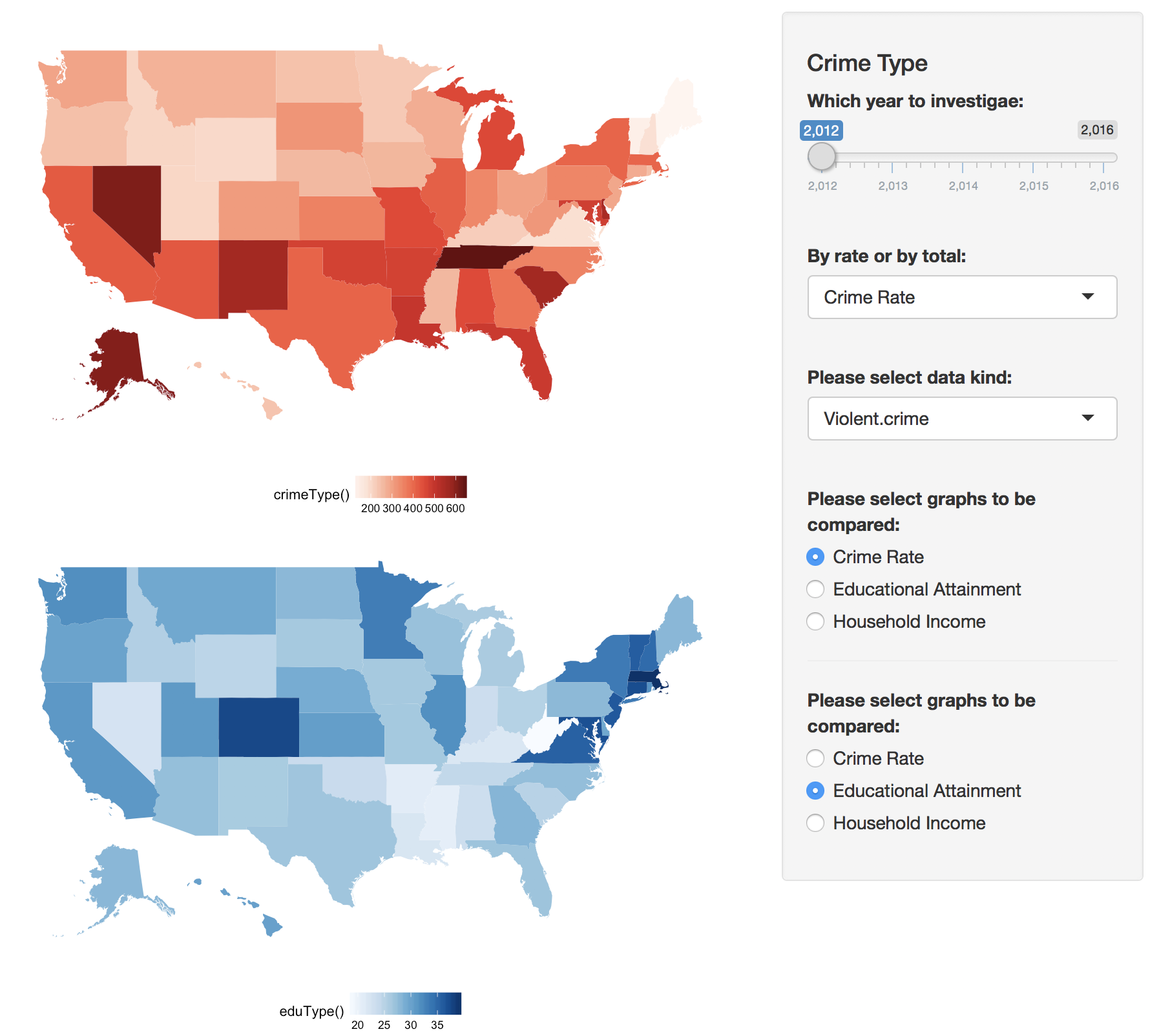


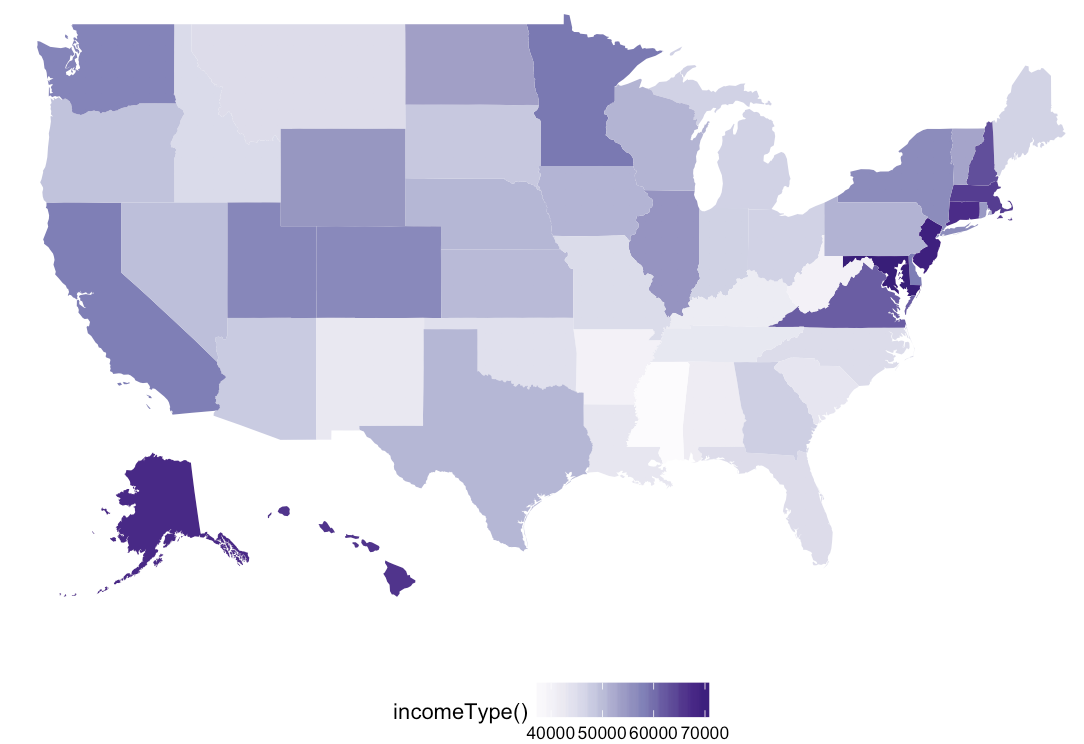
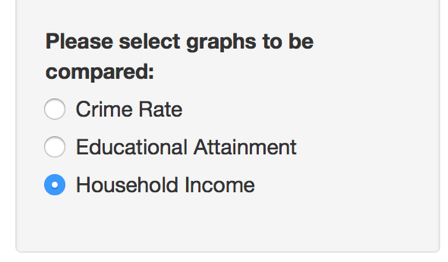
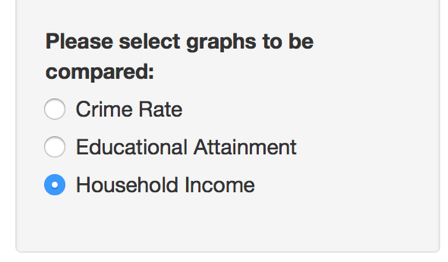
Probably due to the small size of the dataset, on the aspect of trending, no obvious relationship between crime rate and other two datasets are found.

However, a booming crime rate since 2015 was identified from this line chart, and this finding is in line with the news reports in 2015, “The report shows that there was an overall increase in violent crime last year, making clear what each of us already knows: that we still have so much work to do” (Berman, 2016).



Looking at the map data comparison, it is pretty clear that crime rate is on the contrary of educational attainment and household income. The darker areas on the crime rate map generally has a lighter color on education and household income map.

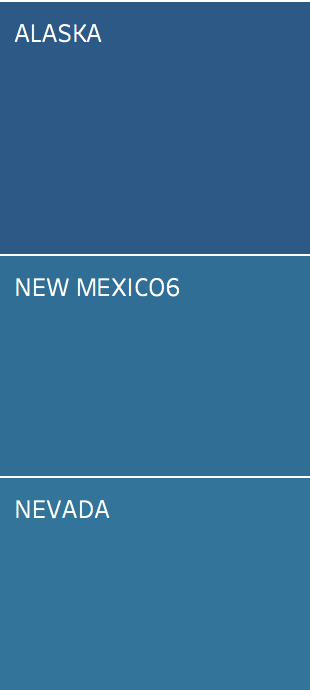
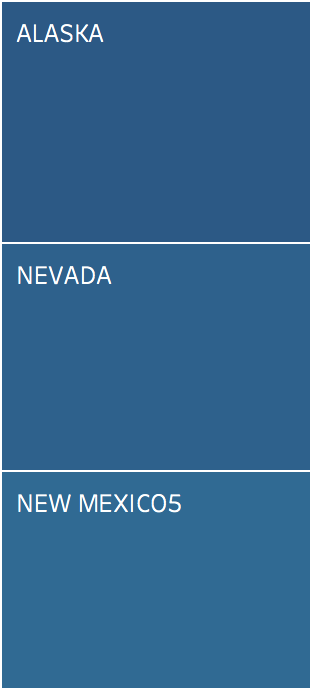
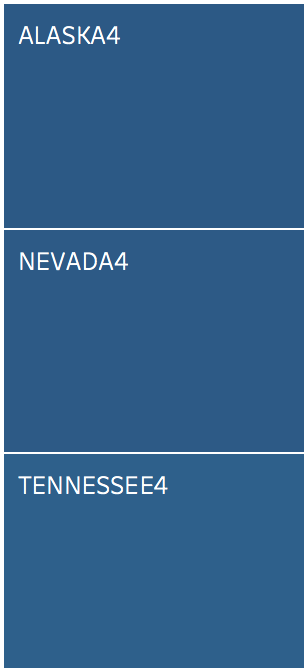
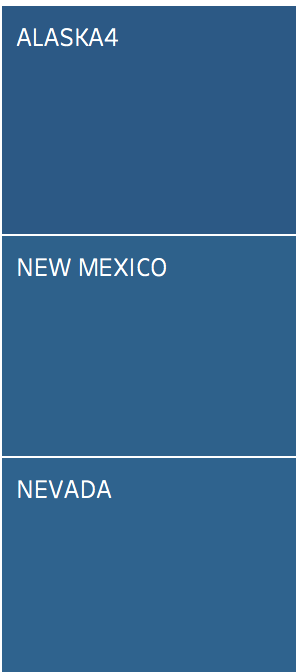




With supporting of the data exploration and visulizations above, the three questions now can be answered:

* In general, negative relationships were identified between both crime rate vs. educational attainment and crime rate vs. household income, and a positive relationship was found between educational attainment and household income. However, possibly due to the lack of data, no obvious trending relationship was identified between crime rate and other two datasets.
* Not always. An interesting finding was spotted by looking at the three map visualizations, Alaska has a medium educational attainment, high household income, with an extreme high crime rate.
* Since no obvious relationship was found between crime rate and other two datasets, some additional datasets and explorations will be needed to answer this question.

An interesting finding by building treemaps of total crime rates, the states with highest crime rates were totally out of my guess of California, New York, and Texas.

2015

2014

2013

2012

2016

**Reference:**

Berman, M. (2016, September 26). *Violent crime and murders both went up in 2015, FBI says.* Retrieved April 26, 2018, from https://www.washingtonpost.com/news/post-nation/wp/2016/09/26/violent-crime-and-murders-both-went-up-in-2015-fbi-says/?utm\_term=.b0c7d1534bf7