VISUALIZATION PRACTICAL WORK

Shiny App:

Arrests in the US

Authors (Group 20):

Elisa Mateos Vicente

Serena Alredisi

Yassir Al Bahri

Course:

Big Data & Data Visualization

January 27, 2020

Contents

List of Figures

Chapter 1

Introduction

Each tool was been developed using Rstudio.

1.1 Dependecies & Folder Contents

The following dependencies have been necessary for the execution of the developed apps.

- library(shiny)
- library(tidyverse)
- library(plotly)
- library(ggplot2)
- library(usmap)

The following files are present in the folder and necessary for the full understanding of this report.

- *report.pdf*
- *authors.txt*
- $interactive\_barplot_state.r$
- $interactive\_histogram_crime.r$
- $Heatmap\_and\_Scatterplot.r$
- $Arrest\_Rate\_US$
- $Arrest\_Rate\_Ranking.r$
- $Kmeans.r$

Chapter 2

Problem characterization in the application domain

The selected dataset is "USArrests.csv", Violent Crime Rates by US State. This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states (1973). The percentage of population living in urban areas is also present in the data set. The data frame contains 50 observations and 4 variables.

1. *Murder* numeric Murder arrests (per 100,000).
2. *Assault* numeric Assault arrests (per 100,000).
3. *Rape* numeric Rape arrests (per 100,000).
4. *UrbanPop* numeric Percent of urban population.

The source can be found in the bibliography. *World Almanac and Book of facts* [1] and *Statistical Abstracts of the United States* [2].

Chapter 3

Data and task abstractions

A series of questions are proposed and solved by building a set of *Shiny* Apps in the R environment. The purpose of this questions is to establish a visual analysis and exploration of the data set.

The studied questions are the following:

1. How many murders, assaults and rapes happened in each State? Is this amount different from the mean value of each crime rate?
2. What is the rate crime distribution in the USA?
3. Is there any correlation between the variables?
4. What is the geographical distribution of arrests by state?
5. Which states are the most violent? Which are the less?

Chapter 4

Interaction and visual encoding

4.1 How many murders, assaults and rapes happened in each State? Is this amount different from the mean value of each crime rate?

Figure 4.1 shows an interactive barplot that allows the user to select any State in the US. Each group represents a different crime: *Murder*, *Assault* and *Rape*. Looking at the heights of the bars it is possible to determine which kind of crime is more frequent and how this frequency changes by State. Each crime is defined by a different color and in the legend box the mean values for each crime can be found. The purpose of this plot is to get a quick idea about how the values shown in the plot vary from the mean. For example, in Figure 4.1 the *Assault* rate is higher than the mean value of Assaults in the US.

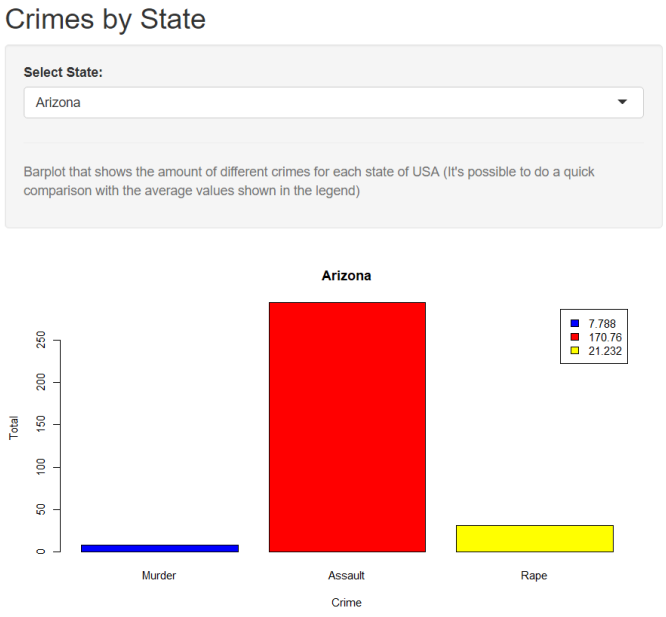

Figure 4.1: One of many output of the app: Crime by state – Run the *interactive_barplot_state.r* file to run the app.

4.2 What is it the distribution of each rate crime in the USA?

Figure 4.2 represents an interactive histogram that allows the user to select the crime variable, the number of bins and the individual observations. The size of the bars in the plot give insight over the shape of the distribution for each crime. In the case of Figure 4.2, the shape of the *Rape* distribution is given by an almost right-skewed histogram. Overall, in this type of histogram the mean is located in the right side of the plot and it is usually greater than either the median or the mode. This shape, thus, indicates that there are a number of data points, perhaps outliers, that are greater than the mode.

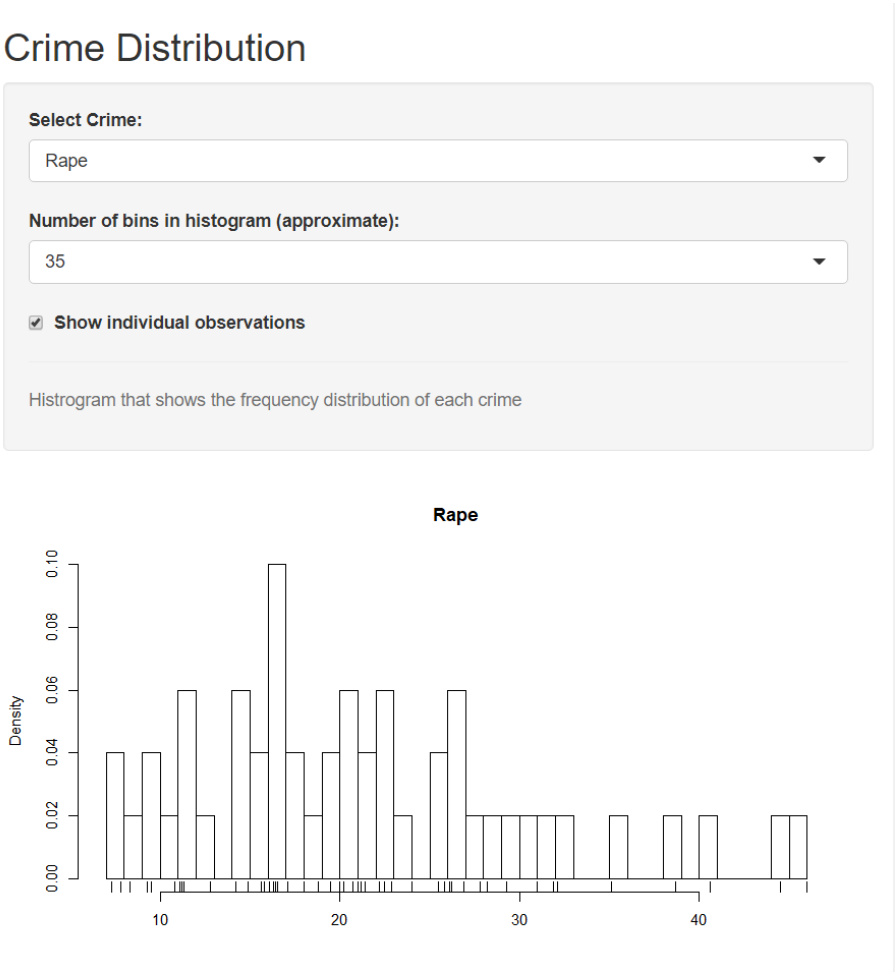

Figure 4.2: One of many outputs of the app: Crime Distribution – Run the *interactive_histogram_crime.r* file to run the app.

4.3 What is the correlation between the variables?

Figure 4.3 is a snapshot of an interactive plot that allows the manual selection of a cell from a correlation heat map to generate a scatter plot of the two corresponding variables. This procedure purpose is to compare the two selected variables. The event data tied to a *plotly_click* event contains the relevant x and y categories. The z value represents the correlation between the variables.

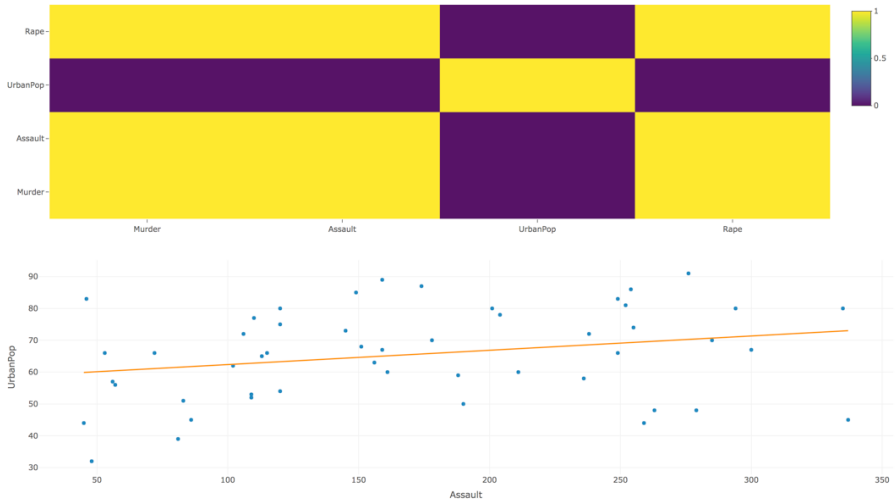

Figure 4.3: One of many outputs of the correlation heat map app – Run the *Heatmap_and_Scatterplot.r* file to run the app.

The plot shows that the relationship between $x = Assault$ and $y = Rape$ has a perfect positive correlation. This means that an increase in one of the variables translates into the same increase in the other. On the other hand, $x = UrbanPop$ and $y = Rape$ show no relationship. The linear regression tool represents different behaviors for each pair of variables. The overall conclusion is that the regions with the most urban population tend to show the highest rates of *Assault* and *Rape*, but the $Murder$ rate is more or less equal in all the States.

4.4 What is the geographical distribution of arrests by state?

This data set has a heavy geographical value. It would not be complete without a geographical analysis. Figure 4.4 is one of the outcomes for the Arrest Rates by State. A choropleth of the US States map in which each state has a different lightness encoding based on the variable studied. The darker the shade, the higher the rate.

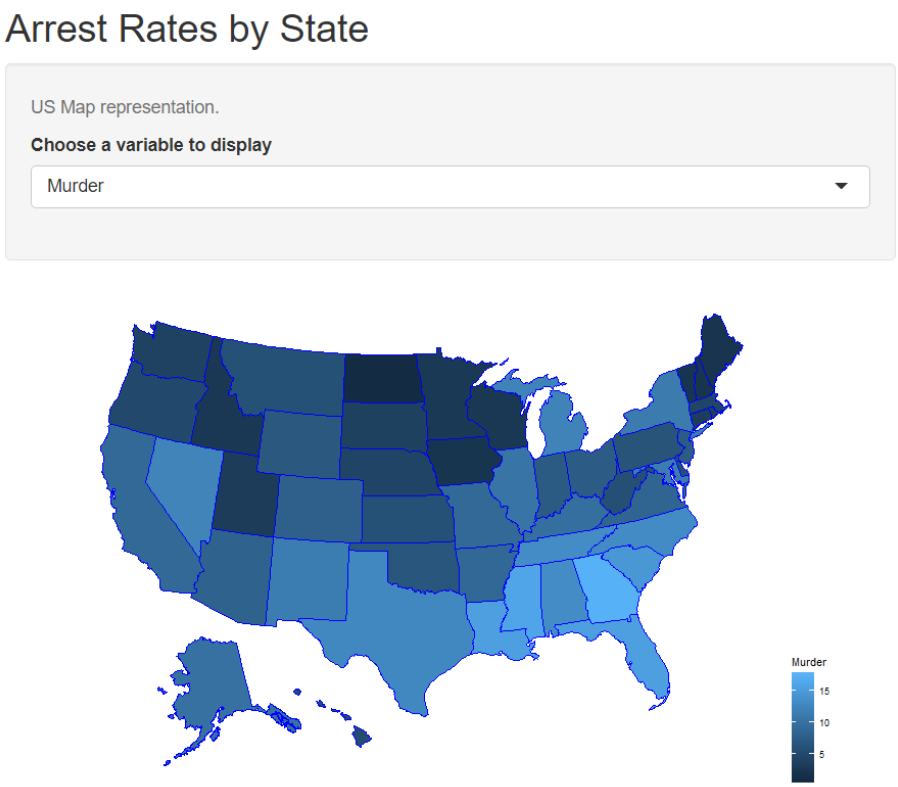

Figure 4.4: One of three outputs of the Arrests by State app – Run the *Arrest_Rate_US.r* file to run the app.

4.5 Which states are the most violent? Which are the less?

Similarly to the previous question, Figure 4.5 is a ranking of states based on the crime variable selected and represented by an horizontal bar plot. Figure 4.4 allows the evaluation of crime rate geographically, whilst Figure 4.5 uses that same data and ranks it from highest to lowest. It is very interesting to compare both of these questions to guarantee an ideal understanding of how each crime behaves in each state.

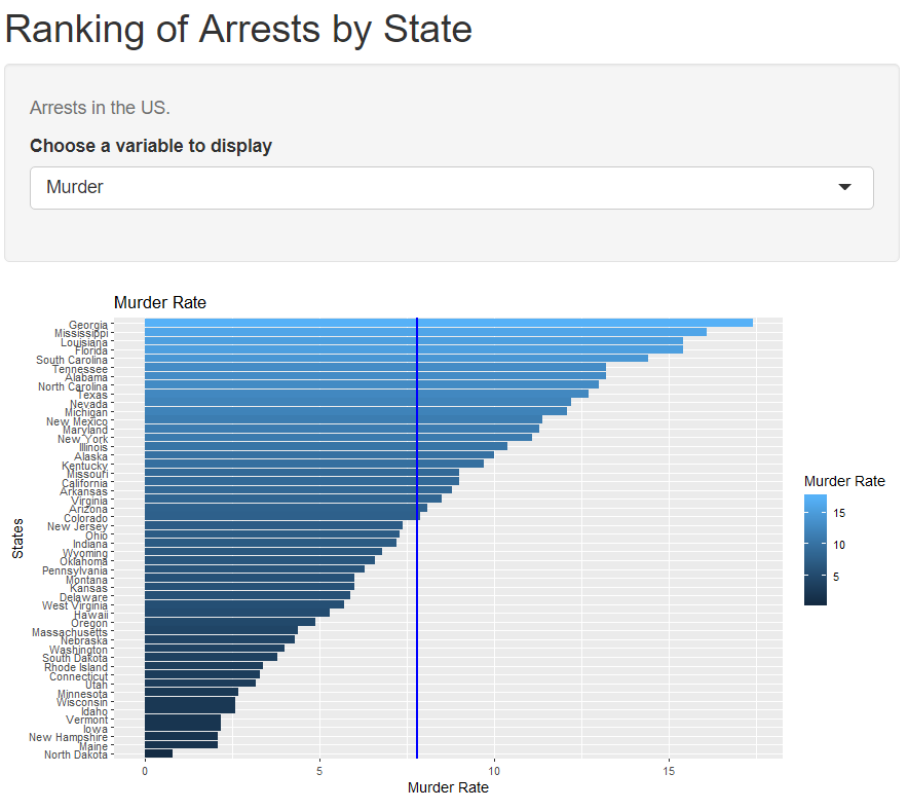

Figure 4.5: One of three outputs of the Arrest Rate Ranking of States app – Run the *Arrest_Rate_Ranking.r* file to run the app.

Chapter 5

Algorithm implementation

In this section a first glance of the clustering analysis performed on the data set is shown (Figure 5.1). The chosen algorithm is K-means. The reason it was selected instead of the EM algorithm is because it performs hard assignments of data points to clusters whilst the EM performs soft assignments. K-means is a particular limit of EM for Gaussian mixtures.

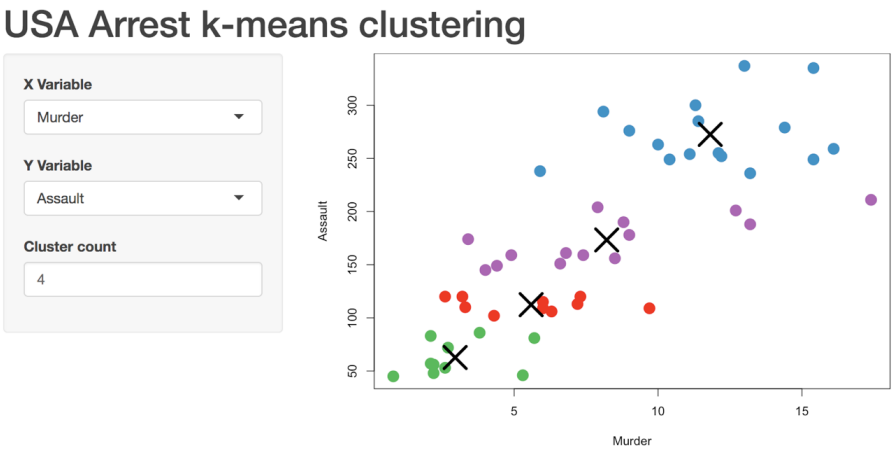

Figure 5.1: Output of the K-means algorithm – Run the *Kmeans.r* file to run the app.

The distances from each point to it centroid can be seen in Figure 5.1. The algorithm classifies the data into four clusters (optimal number). This plot is a simple visualization of clusters using a pair of variables from the data set.

Chapter 6

Conclusion

Visualization is a very vast field. Different tools for all kind of studies with static representations dominating the landscape of data analysis. This work takes a different approach, expanding the capabilities of visualization by providing an interactive platform thanks to the *Shiny* tool.

Several feats have been achieved. Shrinking the amount of individual representations by combining different visualizations or providing the ability to interact with each piece of knowledge in anyway that is found fit. This last perk allows to any user to explore and interact with the apps and extract it own conclusions with may even differ from those of the creators of the app. This is found to be, in the opinion of the authors, the greatest advantage of this methodology.

Bibliography

[1] World Almanac Books, *World Almanac and Book of facts*, 1975.

[2] US Census Bureau, *Statistical Abstracts of the United States*, 1975.

[3] R Studio, *Shiny Tutorial*, `https://shiny.rstudio.com/`.