



# Visualization Practical Work

Big Data  
2017/2018

Daniel Silva Reis  
Pedro Afonso Paulino Ferreira de Castro

# Abstract

This report aims to explain the approach our group took when designing a Visual Analytics tool. Our group used the abstraction levels as a reference for its development. The tool is developed using the R programming language and uses Shiny as a framework to produce the web application. The goal of the tool is analyzing the results of several felonies (murder, rape, assault) of each state in the United States of America.

## Table of Contents

<b>Abstract</b>	<b>2</b>
<b>Table of Contents</b>	<b>2</b>
<b>Problem characterization in the application domain</b>	<b>3</b>
<b>Data and Task Abstractions</b>	<b>4</b>
Task Abstraction	4
Why is the visualization being used?	4
What kind of search is done by the user?	4
What kinds of queries are done by the user?	4
Data Abstraction	5
<b>Interaction and visual encoding</b>	<b>5</b>
Views	5
Tabular View	5
Choropleth Map View	6
Region View	7
Scatterplot View	8
Stacked Bar Chart View and Pie Chart View	9
Data Manipulation	10

# Problem characterization in the application domain

The problem consists in designing and implementing a Visual Analytics tool for a chosen dataset. The dataset we picked was the “USA Arrests” one, which contains the arrests due to rape, murder and assault per 100 000 inhabitants, as well as the percentage of population that lives in urban areas, in every state of the United States of America, in 1973.

From the perspective of an analyst, the following questions would probably be the first ones to come up to their mind:

- Which states have the highest/smallest rape/assault/murder rate?
- Between two or more states, which one has the highest rape/assault/murder/urban rate?
- What is the rape/murder/assault rate or the percentage population that lives in urban areas of a specific state?
- Is there any correlation between any of the variables?
- How common is each felony compared to the others?
- What states have a higher/lower urban population/rape/assault/murder rate than X? What states have a higher/lower urban population/rape/assault/murder rate between X and Y?

With that said, we developed our application with answering those 6 questions in mind.

# Data and Task Abstractions

## Task Abstraction

### Why is the visualization being used?

Our application is fundamentally used to consume information. It communicates information well known *a priori* (through the dataset) to analysts, allowing them to make decisions or planning further actions with said information, making it mostly a tool used for communicating information. However, new knowledge can also be inferred from the application. One can also verify hypotheses (like, for instance, if there is any correlation between rape and assault rates). This means that despite its main use of communicating information, our application is also good for discovering knowledge.

### What kind of search is done by the user?

With the questions identified in the “Problem characterization” section in mind, we can infer two kinds of searching done by a user of our application:

- **Lookup:** When the user wants to, for instance, know the rape rate in California, they know exactly what and where to look for.
- **Locate:** If the user wants to know whether there is any correlation between assault and rape rates or not, they know what they are looking for, but don't know where to.

### What kinds of queries are done by the user?

The user may want to query about a single target, in order to find information about that specific target. They may also be interested in comparing two or more states to know, for instance, who has the highest assault rate of them. If they are interesting in knowing a correlation between two rates, or to find a certain trend in data, however, a query to all elements must be performed. So, in the context of our application, all three types of queries can be done: **identify**, **compare** and **summarize**.

# Data Abstraction

As specified in the previous section, we can identify the following tasks:

- Search for trends, specifically dependencies and correlations between attributes.
- Compare two or more values.
- Find a specific value in an attribute.

This means our application will need a representation that targets all data, for the two first tasks, or specifically one attribute, for the latter one. It should also be noted that the USA states are represented by a categorical variable and that the felony and urban population rates are represented by quantitative variables.

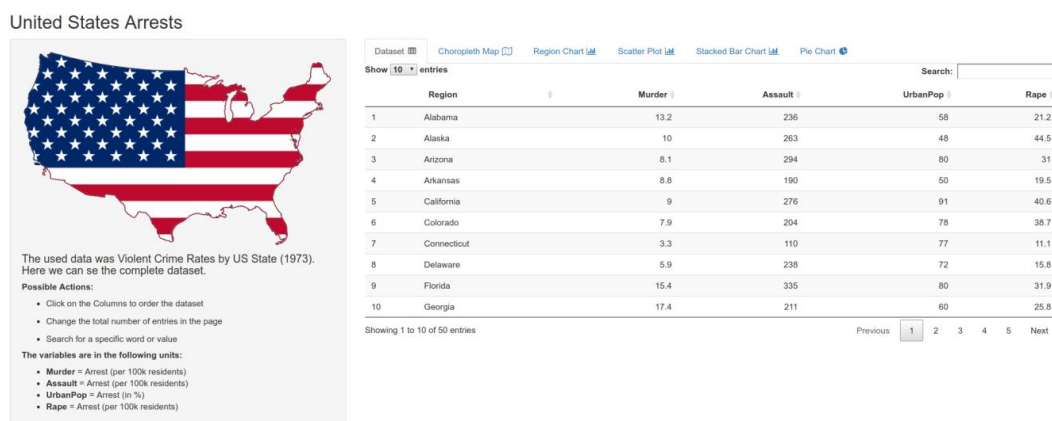
## Interaction and visual encoding

Considering the tasks and data abstractions identified in the previous section, we will need visual representations suitable for focusing on a specific element, several elements and all elements. This means that we will need several views, suitable for each of the user's needs.

## Views

### Tabular View

In this view, the entire dataset is shown in tabular form. The values for a specific state can be quickly retrieved. This view is better suited to answer the question "What is the rape/murder/assault rate or the percentage population that lives in urban areas of a specific state?"



## Choropleth Map View

In this view, a choropleth map can be seen. Choropleth maps take a quantitative attribute and display it in a map, associating the values to the region in the map they represent, through colours. In this view, we can easily see which states have the highest/smallest values for each rate, answering the question “Which states have the higher/smaller rape/assault/murder rate?” and “What states have a higher/lower urban population/rape/assault/murder rate than X? What states have a higher/lower urban population/rape/assault/murder rate between X and Y?”

### United States Arrests

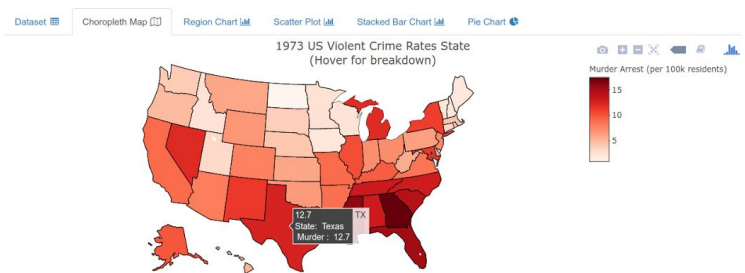
A choropleth map is a thematic map in which areas are shaded or patterned in proportion to the measurement of the statistical variable being displayed on the map, such as population density or per-capita income. Choropleth maps provide an easy way to visualize how a measurement varies across a geographic area or show the level of variability within a region.

In this case, we can choose and vary between the variables "Murder", "Assault", "Rape", "Urban Population" in the dropdown menu and use the slider to select a range of values according to the preference.

Select variable to analyze

Murder

Choose values range:



### United States Arrests

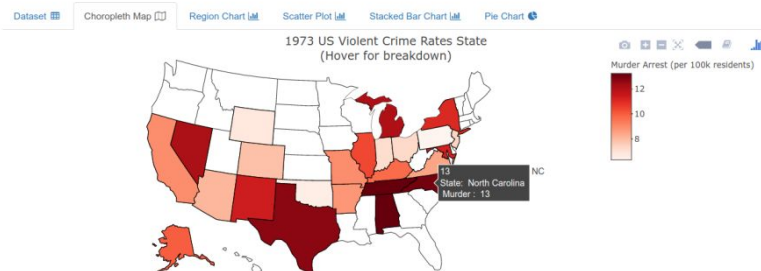
A choropleth map is a thematic map in which areas are shaded or patterned in proportion to the measurement of the statistical variable being displayed on the map, such as population density or per-capita income. Choropleth maps provide an easy way to visualize how a measurement varies across a geographic area or show the level of variability within a region.

In this case, we can choose and vary between the variables "Murder", "Assault", "Rape", "Urban Population" in the dropdown menu and use the slider to select a range of values according to the preference.

Select variable to analyze

Murder

Choose values range:



In this view, we can see a bar chart. A bar chart takes a categorical and a quantitative value and is usually used to compare values or lookup specific views. This view is mostly useful for answering the questions “Between two or more states, which one has the highest rape/assault/murder/urban rate?” and “What is the rape/murder/assault rate or the percentage population that lives in urban areas of a specific state?”.


A bar chart or bar graph is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a line graph.

In this case, we can choose and vary between the variables "Murder", "Assault", "Rape", "Urban Population" in the dropdown menu and use the slider to select a range of values according to the preference.

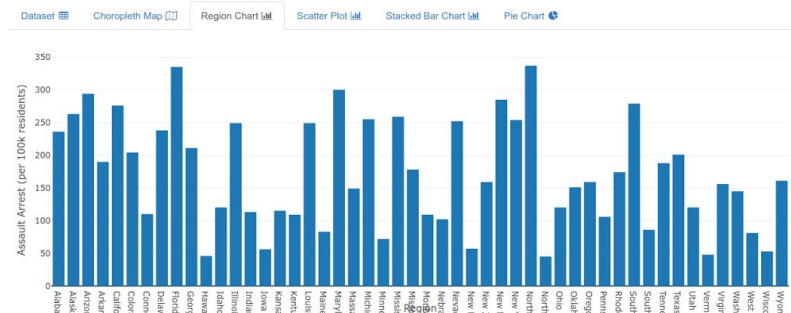
## Select variable to analyze

Assault

Choose values range:



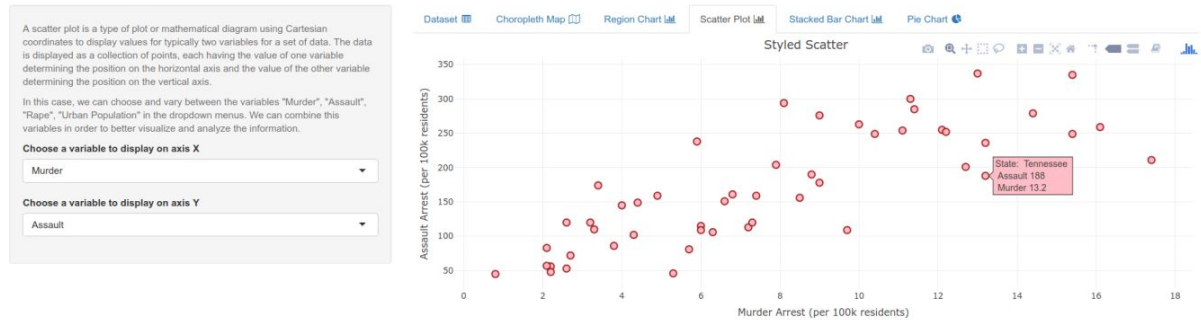
45 74.2 103.4 132.6 161.8 191 220.2 249.4 278.6 307.8 337



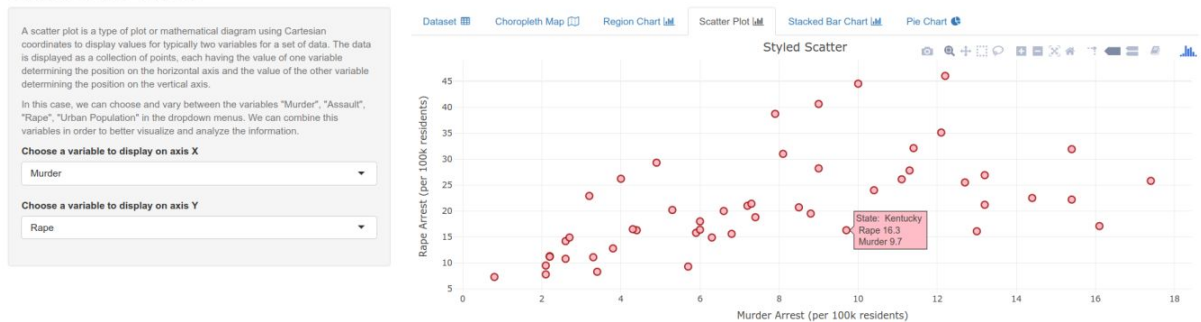
# Scatterplot View

In this view, we can see a scatterplot. Scatterplots take two quantitative attributes, each taking horizontal and vertical spatial position channels. They are mostly used to find extreme values. We can also easily check for correlations between attributes, which allows us to answer the question “Is there any correlation between any of the variables?”

## United States Arrests



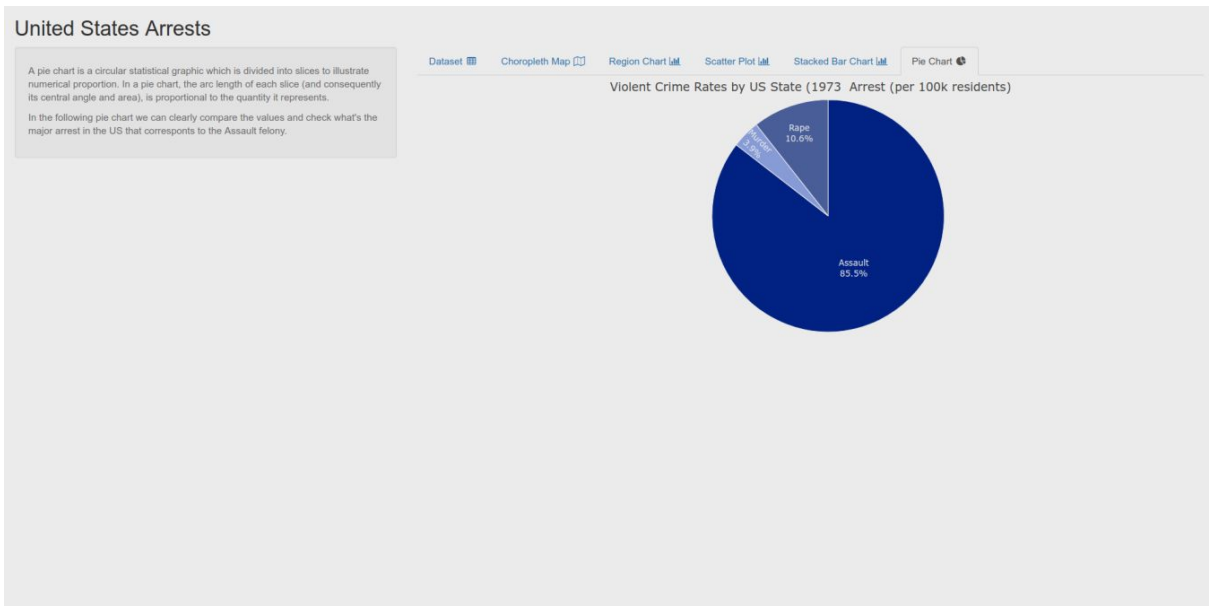
## United States Arrests





# Stacked Bar Chart View and Pie Chart View

Stacked bar charts take two categorical key-values and a quantitative value and display, using colour and length as visual channels. Both the length of each part and the whole encode a value. A pie chart takes a categorical value and a quantitative value and allow us to understand the relative value of a part compared to the whole. In these views, we can compare the values of each felony added up together and understand which crimes are more and less common. This answers the question “How common is each felony compared to the others?”. The Stacked Bar Chart View allows us to get the answer to that question in regards to each specific state, whereas the Pie Chart View allows us to get the answer for the entire United States.



## Data Manipulation

In order to enhance the experience of the user in finding answers to those questions, we must let the user do some manipulation of the data:

- The user can choose which felony (murder, assault, rape, urban population) they want to show in the views that allow for that flexibility (Choropleth Map View and Region View) . There is synchronism between these views, which means that if you change the attribute in one of them it also changes in the other one.
- By letting the user filter the states by their name, searching for a specific state becomes much easier.
- By letting the user filter a specific attribute's values according to a range (higher than X, lower than X, between X and Y) we make it easier to understand which states have the highest/lowest rates in a specific attribute and which states have a higher/lower urban population/rape/assault/murder rate than X? What states have a higher/lower urban population/rape/assault/murder rate between X and Y?are closer to each other in a specific attribute.

The user should also be able to easily switch between views according to its needs.