Interactive data visualization using Shiny App in R

Suchit Kumar Mahatman, Adarsha Jung Basnet
Department of Computer Science and Engineering
sm02408918@student.ku.edu.np
basnetaadars123@gmail.com

***Abstract* - It has been frequently observed that a vast amount of data surrounds everyone. Data is treated as a raw material of any research, business, or any other developmental activities. Currently, data has become an essential component of businesses, industries, research organizations, and technological development. We are living in an era of Big data. The volume of data used is rapidly growing every second. Data visualization helps people to the significance of data by summarizing and presenting simply. Data visualization is a process to describe massive data in an accessible, understandable, and straightforward format. A researcher must understand the importance of data visualization and its relationship to research. In the present chapter, we introduce data and visualization of data, and also provide some techniques for visualization of data using statistical software R. R is an open-source programming language that is widely used as a statistical software and data analysis tool. R Studio is an integrated development environment(IDE) for R.**

1.INTRODUCTION

Data:

Data is/are the raw and unorganized facts of the world. When data is/are processed, organized, structured, or in any presented form for some context, useful, it is called information. Knowledge is a personal map/model of information. The data type can be grouped into two major categories, qualitative data, and quantitative data. Qualitative data, which has a numerical value, cannot be assigned, e.g., motivation, confidence. Quantitative data, which has a numerical value, can be assigned, e.g., height, weight, speed. Qualitative data further divided into ordinal and nominal values and quantitative data divided into continuous and discrete values.

A data model is a logical inter-relationships and data flow between different data elements that are involved in building an information system. It is also a documentation way of storing and retrieving. Data models help represent what data is required, and what format is to be used for business or scientific processes.

Data Visualization:

Data visualization is the process of interpreting data and presenting it in a pictorial or graphical format. Data visualization helps the audience to understand the significance of data by some methods, e.g., summarizing and presenting. Another role of Data visualization is a process to describe massive data in an accessible, understandable, and straightforward format. Data visualization is a method to communicate the information clearly and effectively. The enormous growth of data, it has become difficult for research and business organizations to extract crucial information from available data. The research and business organization collects massive data because of analyzing that data capable of making critical business decisions.

Software R:

R is a software environment with a programming language most commonly used for statistical computing and machine learning. It is maintained by the R Foundation. In 1993, R was developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. Now R is available as open-source software. R possesses an extensive catalog of statistical and graphical methods. R is not only entrusted by academics, even some reputed business organizations, including Uber, Google, Airbnb, Facebook, also use R. The scientific community has also recognized computation using R. The enormous help and reference resources of R available on the internet.

Data Visualization Using R:

Matthew N. O. Sadiku et al. (2016) explained some visualization techniques, e.g., line, pie, bar, and scatter with the application, and challenges of data

visualization. The term variate has played an essential role in statistical analysis and data visualization. In contradistinction to a variable, a variate is a quantity which may take any of the values of a specified set with a specified relative frequency or probability. Quantification is the process of assigning numerical values to a variable, and the quantified variable is known as variate. Some statistical analysis is termed with word variate with the prefix, e.g., uni, bi, or multi. The univariate analysis is based upon quantified variables with numerical values one.

2.TECHNICAL BACKGROUND

R programming language and R studio have been used for data visualization for decades. There is a lot of trust among the researchers and scientists as it provides packages (libraries) for easy implementation of different visualization tools. Shiny is an R package that makes it easy to build interactive web apps straight from R. We can host standalone apps on a webpage or embed them in R Markdown documents or build dashboards. We can also extend our Shiny apps with CSS themes, htmlwidgets, and JavaScript actions.

3.RELATED WORKS

Over the decades of R programming there have been many advances in developing various apps using R packages. Many of them are developed to visualize the real world data. There are many apps which are developed using R Shiny in the field of life sciences. Some of them are:

RStudio: Covid-19 Tracker

It's near impossible to list the best dashboards made with R Shiny in life sciences without mentioning Covid-19. The dashboard made by RStudio shows daily updates on the number of cases and deaths among 199 countries.

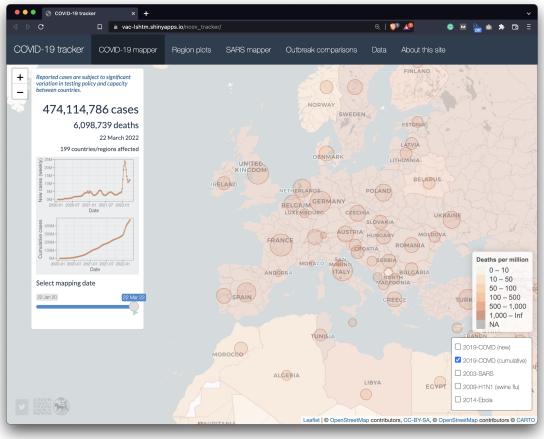

Fig:Covid-19 Tracker dashboard by RStudio

The dashboard allows its end-users (citizens, researchers, government officials) to keep track of the pandemic stats from the very beginning (Jan 22, 2022). It has a section including an interactive world map, region plots (cases in regions of interest), and much more.

There are thousands or even tens of thousands of Covid-19 dashboards out there, so what makes this one unique? Well, it's built with R Shiny, it's free to access, and has the source code available on GitHub.

RStudio: Genome Browser

This dashboard shows a visualization based on Circos, which is a way of visualizing whole genomes. It uses pancreatic adenocarcinoma tumor samples data from various donors, provided by ICGN.

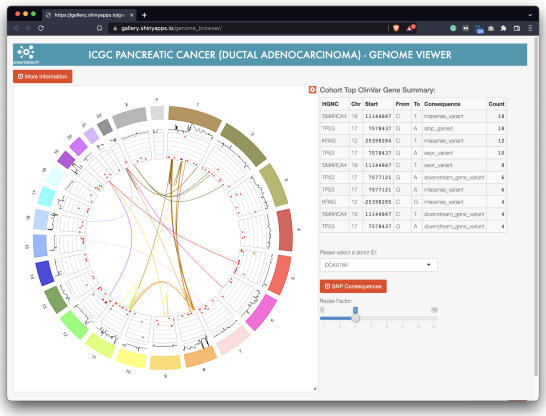

Fig: Genome Browser dashboard by RStudio

If we are not familiar with the Circos-style plot, here's a crash course. It shows several layers, starting from the outer one representing color and size distinguished chromosomes, moving clockwise from chromosome 1 to 22, then X and Y. Inside the ring, we can see a line representation of copy number mutations.

The dashboard serves researchers in the area of genomics and bioinformatics to make sense of complex data. It's not the most useful dashboard for people without advanced domain knowledge.

RStudio: ShinyMRI

The ShinyMRI dashboard visualizes 3D MRI images and is made entirely with R Shiny. It was also recognized as an honorable mention in the 2019 Shiny Contest.

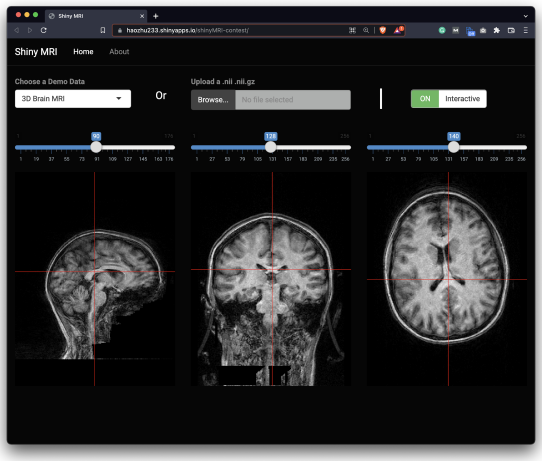

Fig – ShinyMRI dashboard by RStudio

What makes this dashboard unique is the ability to upload our own files, either in .nii or .nii.gz format. That functionality makes the dashboard extremely useful for medical professionals. Sure, it's in a PoC state now, but a little tweaking and added functionality could take it a long way.

Appsilon: Bee Colonies

The Bee Colonies dashboard was made by R/Shiny Developer, Ryszard Szymański, in one day.

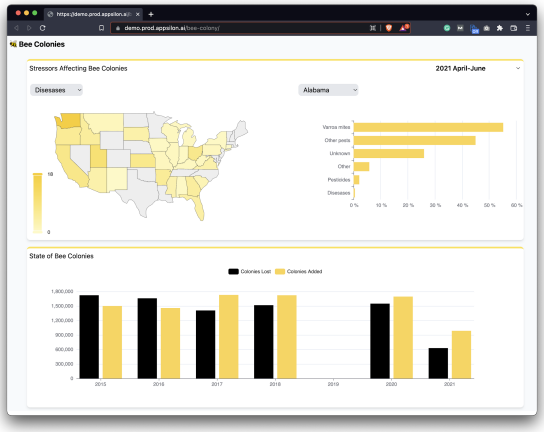

Fig: Appsilon's Bee Colonies dashboard

The dashboard shows stressors affecting bee colonies (diseases, pesticides, and others) in different US states for a given period. It also shows you how many bee colonies were lost and added in a given year, ranging from 2015 to 2021.

It helps decision-makers to early detect potential problems of adding bee colonies to one location instead of the other.

4.PROPOSED SYSTEM/METHODOLOGY

Installing R and R studio

R is maintained by an international team of developers who make the language available through the web page of The Comprehensive R Archive Network. The top of the web page provides three links for downloading R. We follow the link that describes our operating system: Windows, Mac, or Linux.

We installed R and R studio by simply downloading them from their respective webpages. After the installation was complete we were able to write our code in the R studio using R programming language.

Importing libraries

Shiny is the main library we used in both our apps. With the help of Shiny package we were able to create a ui and server in our app to display histogram and correlation plot.

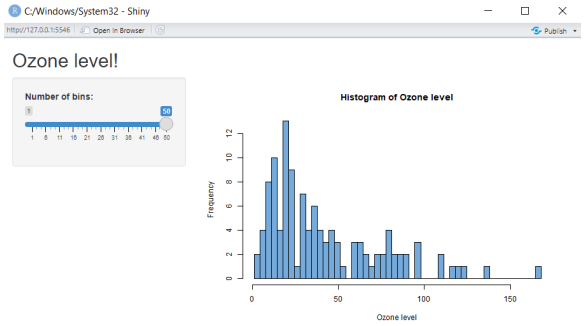

Fig: Histogram visualization using Shiny package

We used other packages in our correlation app. Bslib package is used to style our app by giving it a "minty" theme.

Modeldata package provides datasets for visualization. We used three datasets from this package to show the correlation between its attributes. The Data Explorer package is used to list data for evaluation in the correlation plot. Plotly package is used to show the correlation plot. Tidyverse package is used to create the correlation plot using ggplot.

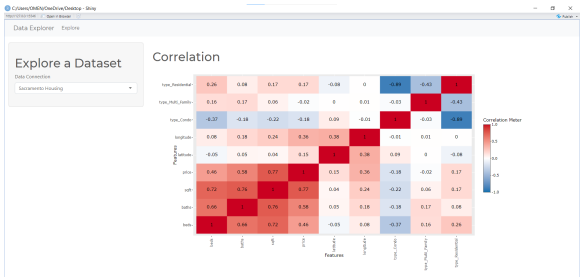

Fig: Correlation plot of Sacramento Housing Dataset

5.DATASETS

A dataset is a structured collection of data generally associated with a unique body of work.

The simplest and most common format for datasets you'll find online is a spreadsheet or CSV format — a single file organized as a table of rows and columns. But some datasets will be stored in other formats, and they don't have to be just one file. Sometimes a dataset may be a zip file or folder containing multiple data tables with related data.

The dataset we used for this project is a CSV file. We have chosen 4 different datasets for visualization provided by the Model data packages in Rstudio.

Air Quality, Stack Overflow, Car Prices, Sacramento Housing are the datasets we use in our applications.

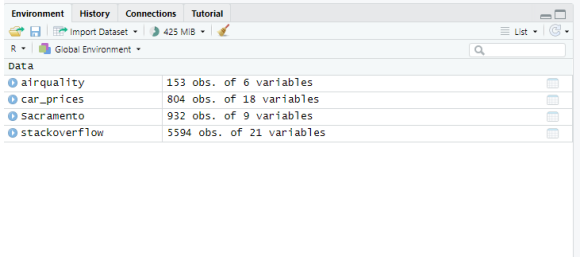

Fig: Datasets used in our applications

6.EXPERIMENTAL ANALYSIS

In our histogram project, we took an air quality dataset for observation of ozone level from May to June. The frequency information in the histogram obtained shows that the level of Ozone remained the same for a time period.

As we can see from our visualization, most of the day ozone level is from 0 to 50. But there are also some days when the ozone level spikes over 150.
Using the slider bar method we have controlled the number of bars in the histogram to be shown. By increasing or decreasing the slide bar we can see the data are mostly clustered between 0-50.

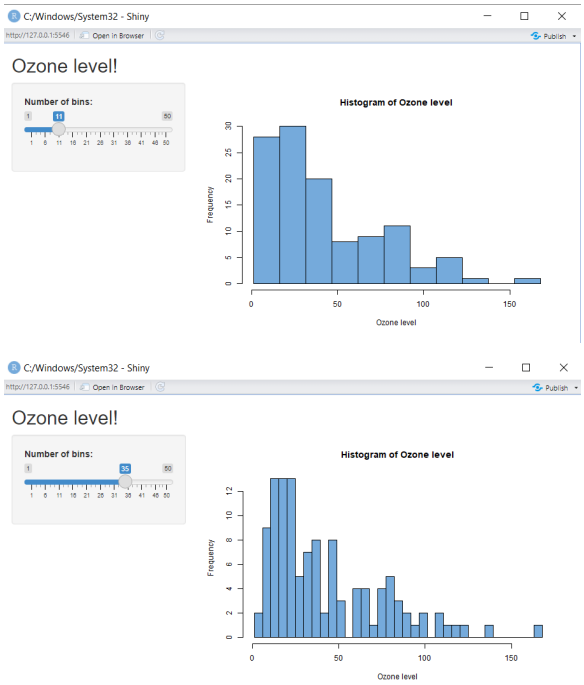

Fig: Two images showing the use of a slide to control the number of bars in the histogram app.

In our correlation project we have taken three datasets namely: Stack Overflow, Car Prices, Sacramento Housing to visualize the data and analyze the correlation between their attributes. The strength of correlation is determined by correlation coefficient which varies between -1 to +1.

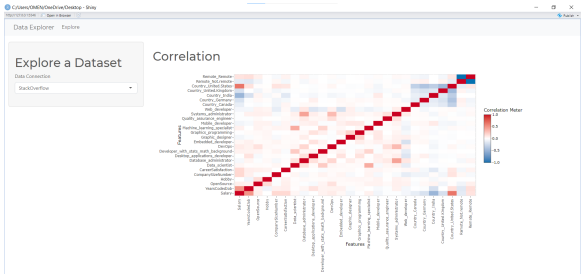

Fig: Correlation between attributes of Stack Overflow dataset

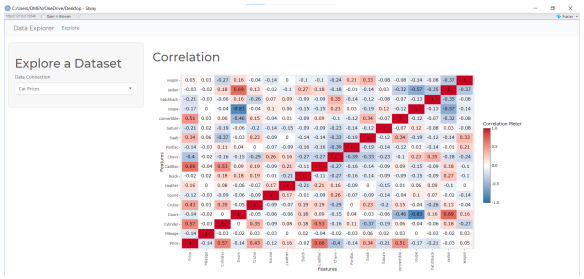

Fig: Correlation between attributes of Car Prices dataset

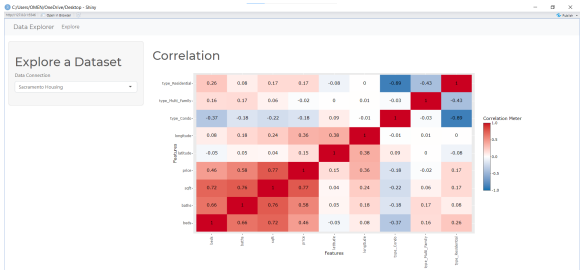

Fig: Correlation between attributes of Sacramento Housing dataset

In the above figure we can see that the correlation between attributes are shown with the help of colors. If the color in the plot is red or moving from white to red then they are positively correlated, else if it is blue or moving from white to blue they are negatively correlated.

7.DISCUSSION AND CONCLUSION

There are lots of things to learn from a project. Using R and R shiny we came to understand the basics of building an app with R studio. It was very easy to use and interact with. Package installation, importing datasets, and even easy to run apps on the emulator as it is provided by Rstudio itself. We also understood the use of different visualization tools such as histogram and correlation plot in different types of data.

8.REFERENCES

1. R studio, Available:www.rstudio.com
2. R Programming language, Available: https://cran.r-project.org/bin/windows/base/
3. Anjali pant, R.S. Rajput, “Introduction To Research Data And Its Visualization Using R”, October 2019, Available: https://www.researchgate.net/publication/336982016_Introduction_To_Research_Data_And_Its_Visualization_Using_R
4. Dario Radečić, ”R Shiny in Life Sciences – Top 7 Dashboard Examples”, 29 March, 2022, Available: https://appsilon.com/r-shiny-in-life-sciences-examples/
5. Saurav Kaushik, “Creating Interactive data visualization using Shiny App in R (with examples)”, 17 October, 2016 Available: https://www.analyticsvidhya.com/blog/2016/10/creating-interactive-data-visualization-using-shiny-app-in-r-with-examples/