Analyzing the freeCodeCamp Repository

freeCodeCamp

GitHub

Gerald Soriano

CS 498 – Advanced Computing and Artificial Intelligence

February 17, 2017

RStudio®

Introduction

For the final lab for this class, I decided to conduct a data analysis and visualization of data on GitHub. My original plan was to analyze commit data for the year of 2016. I wanted to observe the total commits made for the year, what region of the world the commit was made, and what time of day the commits were made.

However, due to technical difficulties (names OAuth troubles with GitHub's API), I had to switch my project up a bit.

Design

The final design of my project involves analyzing contributor data of the freeCodeCamp repo and creating 2D & 3D visualizations. I will analyze how many people contributed to the repo, how many repos the contributors themselves created, and how many followers each contributor had.

To change things up a bit, I also decided to do data analysis and visualization through the programming language, R. R is a widely used language in the data analytics world and I think that learning a new language that has popularity in a specific domain would be beneficial for me.

Implementation

R is a simple language and I decided to create two 2D plots and two 3D plots. Packages had to be installed and loaded in order for data visualization to work. The code snippets are as follows:

```
# Create a scatter plot
ggplot(
  data = fccdata,
  aes(
    x = user,
    y = contributions)) +
  geom_point() +
  ggtitle("freeCodeCamp repo Contributions by User") +
  xlab("User") +
  ylab("Contributions") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))

# Visualize the data
plot(density(z), main = "freeCodeCamp User Contributions")

scatter3D(x, y, z, phi = 0, bty="g", main = "freeCodeCamp Contributor
Data")

scatter3D(x, y, z, phi = 0, bty = "g",  type = "h",
          ticktype = "detailed", pch = 19, cex = 0.5, main =
"freeCodeCamp Contributor Data")
```

Results

From the code in the previous section, I was able to create some simple and easy to read visuals.

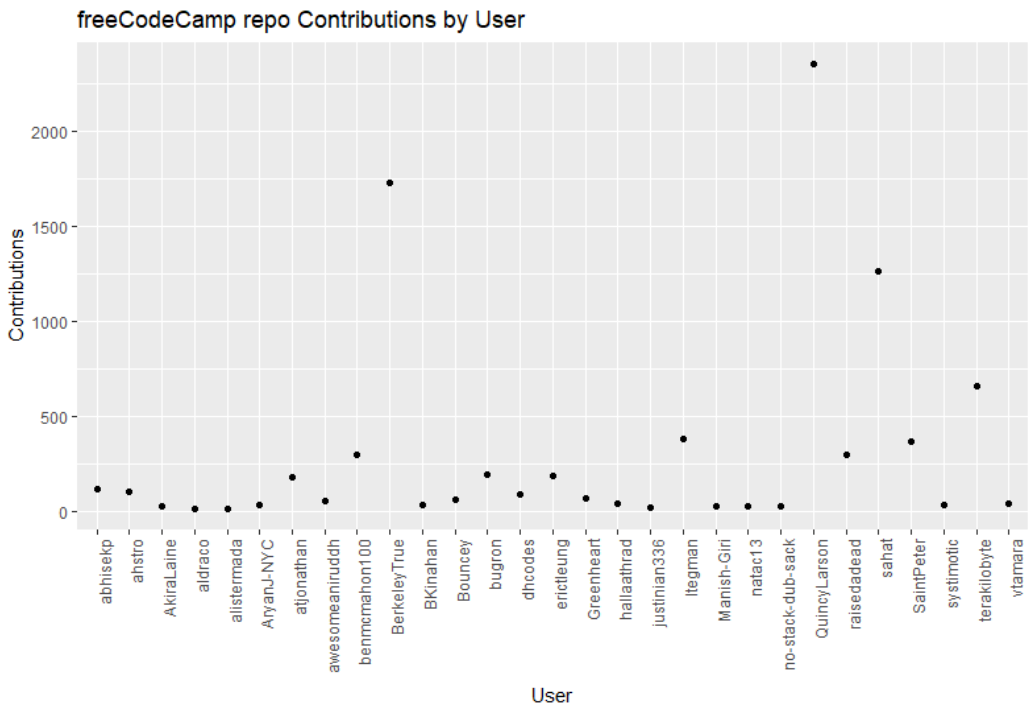

Figure 1. *A scatter plot of the contributor data for the freeCodeCamp repo*

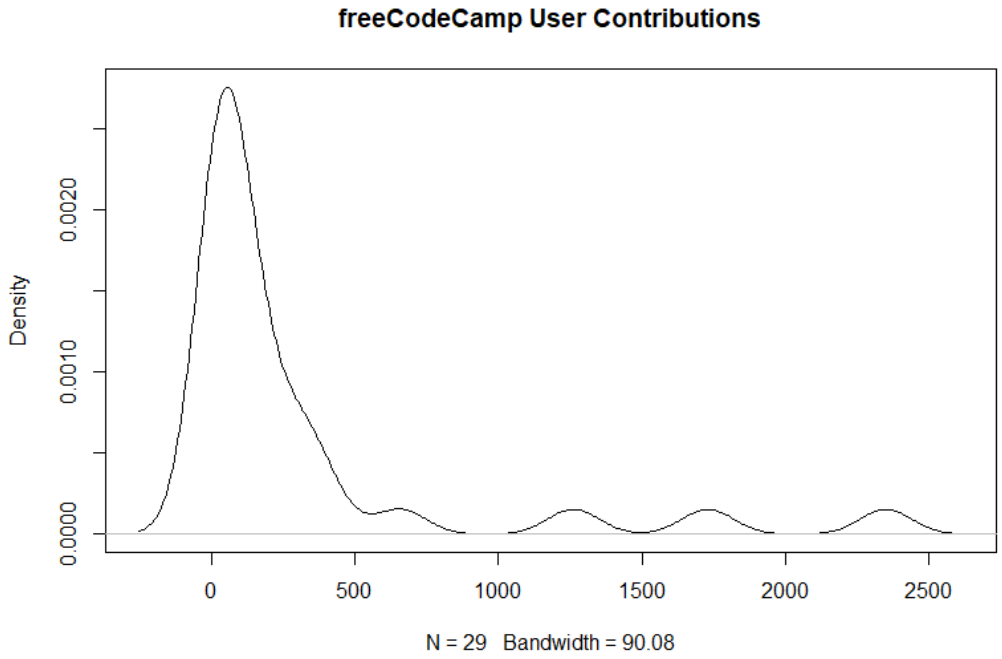

Figure 2. *A density plot of the contributor data for the freeCodeCamp repo*

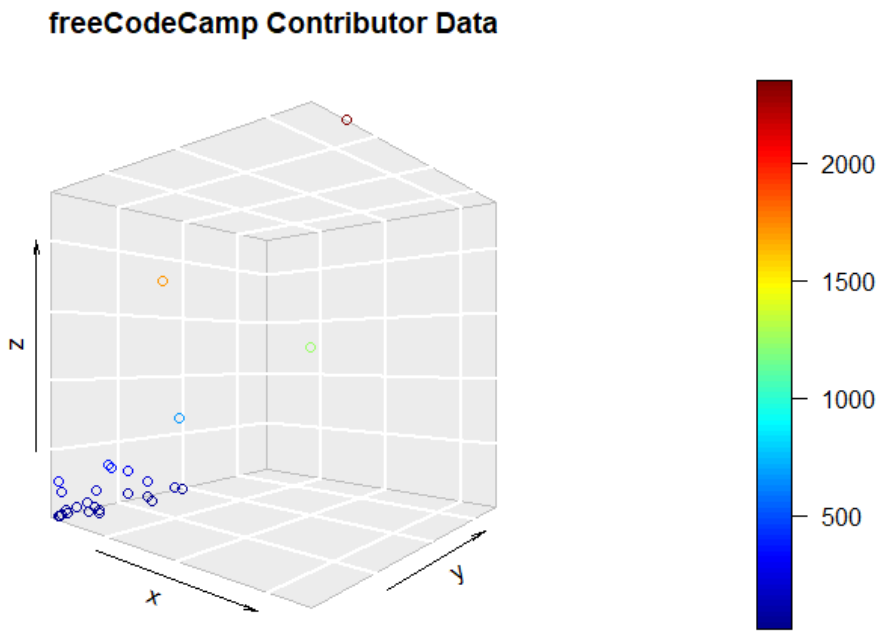

Figure 3. *A 3D plot of the contributor data for the freeCodeCamp repo*

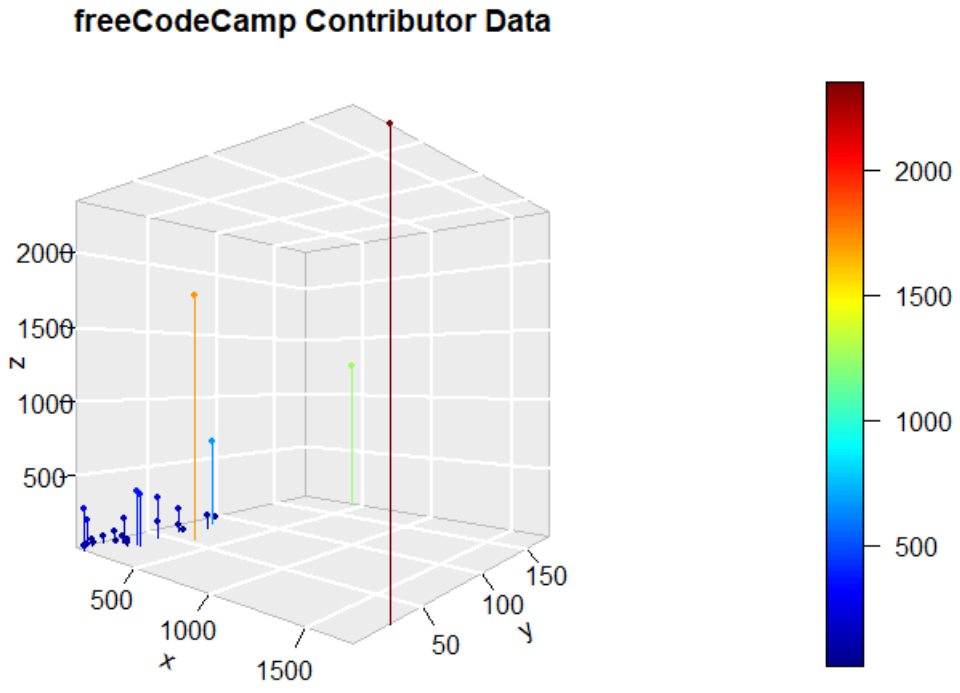

Figure 4. *Another 3D plot of the contributor data for the freeCodeCamp repo*

Conclusion

Looking at the scatter plot in Figure 1, most users had contributions below the 500 mark. There were only four contributors outside of that mark who were above and beyond in their contributions. Figure 2 illustrates that same observation in another setting through a density plot. The 3D plots in Figures 3 and 4 do a better job with highlighting the outliers in the freeCodeCamp repository contributor data. By analyzing this data, it is easy to see how there has been a considerable amount of contributions set force for this software project.

The freedom of this lab report allowed me to have even more fun with testing out and applying the new found knowledge I received from this course. Though I still wish that I could have analyzed real "big data", learning a new language as a result from this final project is still a big plus. A recommendation for this lab is to have a showcase of works that previous students have completed for the final project. This will give my fellow students and I even more ideas for implementations. Overall, this lab was good and fun!

Important Links and Resources

My repo for the project: https://github.com/sorianog

freeCodeCampe repo: https://github.com/freeCodeCamp/freeCodeCamp

RStudio Download: https://www.rstudio.com/products/rstudio/download/

Learning R through Pluralsight: https://www.pluralsight.com/courses/r-data-science