Sean McLean

ALY 6000 – Project 1

09/29/2023

Introduction

The first project of the course consists of the implementation of vectors and how the usage of functions, brackets, and other tools like data visualizations can be beneficial. The majority of the project's questions delves into the basics of using vectors and serves as a proper introduction to the subject matter. The last few questions provide the opportunity to download a datafile into R Studio which serves as an example where vector building and function usage can be utilized before adding data visualizations to showcase the final product. One of the key aspects to focus on is how to create vectors and then build new ones based off the previously built vectors using functions, operators, and visualizations.

Key Findings

Some of the key findings of the project include how learning that symbols can calculate codes in vectors. The first question provides as a quality introduction on how they can be valuable tools to use in R studio.

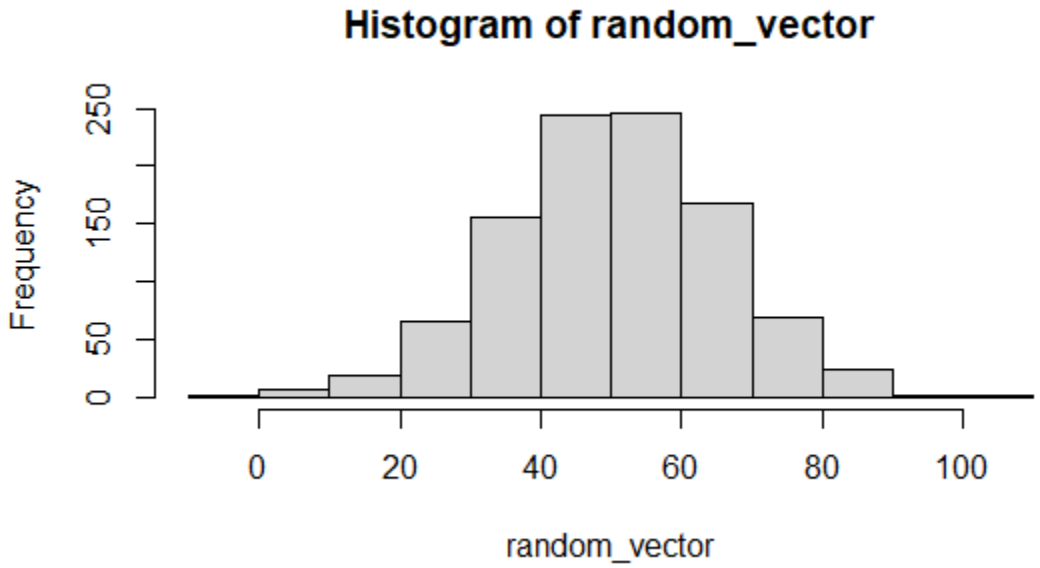
```
> 123 * 453
[1] 55719
> 5^2 * 40
[1] 1000
> TRUE & FALSE
[1] FALSE
> TRUE | FALSE
[1] TRUE
> 75 %% 10
[1] 5
> 75 / 10
[1] 7.5
```

An example of how multiple symbols could be used in a vector was featured on question 14, where the second vector variable has a range of numbers from 10 to 30 that contains only even numbers within that range. Adding 20 to each value changes its range to 30 to 50. Using the same vector variable, if you instead multiply each value by 20 then it changes its range to 200 to 600. Incorporating the function 'greater than or equal to' with a value of 20 to the vector will provide each value with a TRUE or FALSE answer depending on whether the value is greater than 20 (TRUE) or less than 20 (FALSE). And adding the function '!=' with a value of 20 will provide each value with a TRUE or FALSE answer depending on whether the value equals 20 (TRUE) or does not equal 20 (FALSE).

As well as the usage of functions for vector creation and processing, the implementation of brackets can help extract values that can be stored into new vectors. Adding Boolean statements to a vector will show whether a value is going to be either a true or false

statement (Phillips, 2018). In question 23, the vector is indexed with logical vectors which means that they have only TRUE of FALSE values. So, the way the c vector is indexed the only TRUE values are the second and fourth values so 12 and 5 which are the only values that will be returne d. The vector contains 10 values that range from 10 to 30 and only contains the even numbers in the range. On question 25 the 'seq' function is used to also show a range from 10 to 30 and only contains the even numbers in the range, followed by question 26 that uses the greater than or equal to operator that returns all values between 20 and 30.

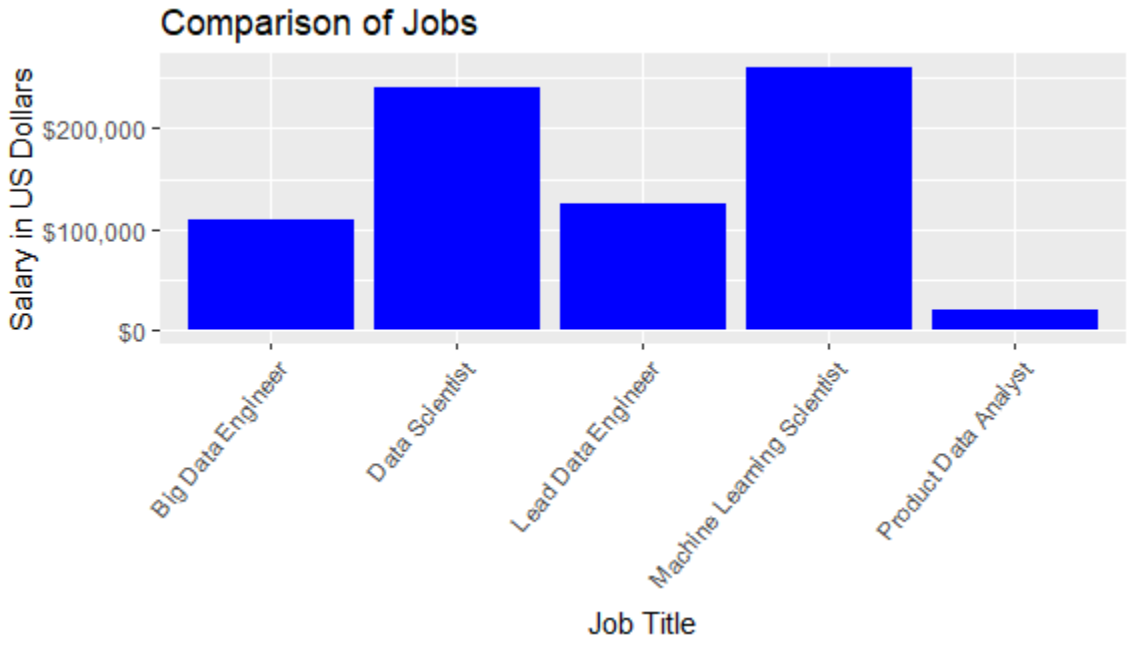
One interesting part of the project I think from my research was the set.seed function that will return a certain amount of random samples from a vector (Zach, 2022). So, in this case on question 30 it will provide five samples from the vector. The vector will run under the circumstances it has provided which is a total of 10 samples that have to be between 0 and 1000 in value. I had trouble understanding this code but if I had to guess it would be that the set.seed function will return five samples from the vector under the conditions it has set. Another example of the set.seed function is on question 37, where the histogram has a mean of 50 and where the top of the bell shaped curve would lie. One standard deviation has a value of 15 so on both sides of the mean the range for one standard deviation would be between 35 and 65 (Geyer, 2023). This would indicate that 68 percent of the values will fall within that range. The histogram is provided below that was a part of question 38.



The histogram has a mean of 50 where the top of the bell-shaped curve would lie in regards to how the data is dispersed. One standard deviation has a value of 15 so on both sides of the mean the range for one standard deviation would be between 35 and 65. This would indicate that 68 percent of the data will fall within that range. Two standard deviations would total 30 on one side and 60 on both sides of the mean so 95 percent of the data would fall within the range of 20 and 80.

Downloading the datafile into the project allowed for the first chance to data shape and manipulate the data using several different methods and functions on question 42. The 'head' function showed just the first six rows of each table in the datafile. By adding the 'n' function, it

will return how many rows in each table that it has been instructed to, so the first seven rows. The ‘names’ function provides all the column names in every table in the datafile. The vector 'smaller_dataframe' was created and uses the ‘select’ function for the 'first_dataframe' datafile and the two columns 'job_title' and 'salary_in_usd' from the datafile. This returns just the two columns from the datafile for the new vector. The vector which contains the 'job_title' and 'salary_in_usd' columns are then ordered in descending order of the 'salary_in_used' column by using the ‘desc’ function. The vector which features the 'job_title' and 'salary_in_usd' columns uses the greater than function with a value of 8000 for the 'salary_in_usd' column which returns only the rows with those values. The vector features three columns (job_title, salary_in_usd, salary_in_euros) from the downloaded datafile. The values in the 'salary_in_euros' column is computed by taking the values in the 'salary_in_usd' column and multiplying each value by .94. The vector then uses the ‘slice’ function that extracts and returns just the selected rows from the datafile. The plotting of the vector includes the job titles and their salaries, the title of the plot, the proper x and y axis labels, the color of the bars, the scale of the x axis in dollars, and the angle of the job title text on the y-axis.



Recommendations

What would make the project more effective is increasing the number of plots and graphs from the datafile. The histogram in question 38 was a great way to see the vector visually and gave me a better idea of what the standard deviation function looks like. This would allow for more data visualizations and to learn more about plotting functions for future project use. With the total of functions and operators that were practiced and researched at the beginning of the project, more of those being used on the datafile instead of just from question 42 would provide

more opportunities to shape the data. These minor revisions would make the project more versatile and effective in its goal.

Conclusion

The project overall sets the foundation for future course work in R Studio and for improved execution in future weekly projects. The importance of vectors and how they are used in coding is a key first step in project building. This was especially vital for a beginning user of R Studio and helps the user understand basic properties and fundamentals in coding. The project is a springboard to how to use vectors in future data sets and using the necessary tools to clean, shape, and manipulate data.

Works Cited

Dragonfly Statistics. (2013, January 5). *Creating Sequences with R.* YouTube. https://www.youtube.com/watch?app=desktop&v=Hw-j-hzAUNM

GeeksForGeeks. (2021, December 21). *Replicate elements of vector in R programming - rep() Method.* GeekForGeeks. https://www.geeksforgeeks.org/replicate-elements-of-vector-in-r-programming-rep-method/

Douglas, A., Roos, D., Mancini, F., Couto, A., & Lusseau, D. (2023, September 11). *Working With Vectors.* An Introduction to R. https://intro2r.com/vectors.html

Phillips, N. (2018, January 22). *Logical Indexing*. YaRrr! The Pirate's Guide to R. https://bookdown.org/ndphillips/YaRrr/logical-indexing.html

Naveen. (2022, August 1). *Remove Specific Value From Vector*. SparkBy{Examples}. https://sparkbyexamples.com/r-programming/r-remove-from-vector-with-examples/#remove-specific-value-from-vector

Zach. (2022, August 16). *How (And When) to Use set.seed in R*. Statology. https://www.statology.org/set-seed-in-r/#:~:text=The%20set.,time%20you%20run%20the%20code

Stat 5101, Geyer. (2023, September 29). *Random Variates*. University of Minnesota – School of Statistics. https://www.stat.umn.edu/geyer/old/5101/rlook.html#:~:text=rnorm%20is%20the%20R%20function,standard%20deviation%20of%20the%20distribution

Kabacoff, R. I. (2022). Creating a Dataset. *R in action: Data analysis and graphics with R and tidyverse* (3rd ed.). Manning Publications. ISBN 978-1-617-29605-5