**STA 518: Final self-reflection**

Josephat Oyondi

April 15, 2022

Through this course of Statistical computing, I have learned and significantly improved in using R in statistical computing and data analysis. The following are the areas I have gained skills and experience to comfortably use R.

Firstly, I can connect R with Github and use a token in automatically updating my Github repositories, thus making the changes in R studio and automatically updating the repositories after the commit and push. Furthermore, I can format markdown files in Github using different styles such as italics, headers, bold, and embedding media files such as images and links. These will significantly help in building my portfolio and showcasing my projects in R and other languages to other peers and potential employers.

Secondly, I can use R to analyze built-in and external datasets. With only simple commands (such as data()) executed over the console, I could be able to review all types of built-in datasets that I could use in creating visualizations. Additionally, I can import various types of datasets such as XLSX and CSV files into R, store them into a data frame object and conduct various activities such as ‘tidyfying’ and creating different charts. The ability to load and store datasets into an object (e.g. Fig. 1 below) then perform data analysis has removed the fear I had on starting to write long queries to retrieve data from files before even starting to conduct data analysis. In this regard, I can install different packages with a simple command: install.packages("package-name") and load the libraries using the command library(“package-name).

Text

Description automatically generated

Fig. 1: Sample approach of loading data and libraries (Source, Author)

Thirdly, through class activities, such as the ‘ggplot2-intro’, I learned important data visualization skills such as mapping and specifying the x and y-axis information and creating plots using commands such as geom\_boxplot(), geom\_bar(), and embedding one chart over another as shown in Fig. 2 below. As a result, I can determine the most suitable visualization that can be used in presenting different datasets. For instance, it is not good practice to embed a boxplot over a scatter plot. Additionally, I can add more aesthetics to a chart such as color, size, and shape. Hence, these skills could significantly help create more effective visualizations.

Chart, scatter chart

Description automatically generated

Fig. 2: Boxplot and jitter plot for various penguin species (Source, Author)

Fourth, through class activities such as data pipelines and data summarization, I can organize data using commands such as arrange(), combine with c() or c\_bind(), summarize(), and select(). I learned that R is simpler in executing commands such as select() for a specific number of rows from data more than other analysis tools such as SQL. Moreover, I can effectively organize data using commands such as groupby(), pivot\_longer, mutate(), and pivot\_wider(). On the other hand, I can read files using commands such as read\_csv() and write into files using write\_csv(). These skills offer me an opportunity to work with files loaded in R studio or stored in a local machine. Finally, rather than writing long codes for performing simple arithmetic such as finding the mean, I have learned quick ways of using functions such as mean(), and sum() using the values in organizing the data in a summarise() command.

Additionally, I learned that with R, I can combine and display more than one visualization with the help of ‘library("patchwork")’ which enabled me visualize revenue, expenses and profits side by side as shown below (Fig. 3). Furthermore, I can use R to perform basic numeric calculations not only with simple values but also with a long list of data from different data frames.

A picture containing text, indoor, map, screenshot

Description automatically generated

Fig 3. Visualization combination

**Learning objectives achieved**

**Objective 1:**

The ‘https://r4ds.had.co.nz/’ has been resourceful in my preparations and learning. The detailed examples provided have greatly helped in understanding the use of various data manipulation concepts. I feel my overall efforts have been excellent since I can use R programming in analyzing data and extracting insights out of untidy and tidy data. Particularly, our project was one of the challenging task with untidy data. However, we used different functions such as pivot\_wider() and values\_drop\_na() to organize data and delete N/A values, respectively (as shown in the screenshot below). As a result, we were able to clean the data and conduct exploratory analysis in achieving our objectives in analyzing financial information about American Airlines.

Text

Description automatically generated

Fig 4. Data cleaning and management sample R code (Source, Author)

The work in Fig 4. above shows the R code I have used to load an excel file into R studio. The code demonstrates how by having used the read\_excel() function to load data and store it in a data frame (my\_data). The sample code shows how I clean data by removing rows with ‘NA’ values and cleaning the unorganized data into a format that I could use to conduct extracting important insights from data.

For objective 2, Fig 5 (a), (b) and (c) demonstrates the graphical displays, charts and tables i have been able to create through R.

Text

Description automatically generated

Fig 5 (a) R code showing how to summarize data in R. (Source, Author)

Chart, scatter chart

Description automatically generated

Fig. 5 (b): Visualization showing jitter plot (Source, Author)

Text

Description automatically generated

Fig. 5(c) Code showing summarization technique (Source, Author)

**Objective 2:**

Throughout the semester, I have used R in creating multiple presentations and conducting exploratory analysis by using different functions such as ggplot and user-defined functions. I successfully achieved this objective by implementing various graphical presentations of line charts, boxplots, scatter plots and area charts. In addition, I was able to create a shinny app using ‘shiny’ package (Fig. 6a,b,c). Shiny app was significantly beneficial in learning how to create more interactive reports on a browser layout, other than creating visualizations within the Rmd file. Consequently, I preferred presenting some of my visualizations (particularly for revenue, expenses and profits of American Airlines) as shown the following three screenshots.

Chart

Description automatically generated

Fig. 6a. Shiny app revenue

Chart, line chart

Description automatically generated

Fig 6b. Shiny expense

Chart

Description automatically generated

Fig 6.c. Shiny revenue

From the shiny app above, I learned that Shiny presented plots an easy-to-read grid that enhances the accuracy of insights obtained from the data. Additionally, the ability to add more aesthetics to the data (such as fill, color and group) helped in generating visualizations based on multiple metrics and datasets. Finally, a notable lesson I learned from the design of visualizations in Shiny is that R programs can use data created outside the ‘ui’ and ‘server’ arguments, thus allowing the user to visualize data in a Shiny app efficiently.

***Objective 3:***

I achieved this objective in activity 11 of the course. In this activity I learned how to use four random sampling techniques for exponential, normal, uniform and chi-square distribution. For instance, in Fig. below, I created a function for generating random samples for all the four distributions. In this example, the generate\_n\_samples function takes only one argument (n – sample size) then generates random values for each distribution and store the records as a tibble (data frame) for further analysis. In this case, one of the important lessons I learned here, is that in R programming, a function can be used in storing data to a single variable, and the variable can be reused in different analytics tasks (Fig. 7).

**Text

Description automatically generated**

Fig. 7. User defined function example

Furthermore, using the in-built functions (mean() and sd()), in was able to obtain the mean and standard deviations all the distributions. Besides, I can use the randomization function in calculating the means using custom functions as shown below (Fig. 8).

**Text

Description automatically generated**

Fig 8. Random sampling sample

On the other hand, I generated multiple samples using the `purr::map\_\*` and `BASE::replicate` functions. These functions helped in exploring the Central Limit Theorem theory. However, the main challenge I encountered in generating multiple samples (e.g. of size 1000) took long to generate and led to slow performance of R-studio.

Finally, performed a distribution exploration of four probability models (using a helper function) including binomial, exponential, normal and uniform distribution. For instance, the Fig. below shows the graphic representation of means of each probability model.

Timeline

Description automatically generated

Fig. 9: Random sampling simulations

***Objective 4:***

The help function in R to learn more about various functions and data by executing commands in the console as shown in the example below.

A screenshot of a computer

Description automatically generated with medium confidence

Fig. 5: Source documentation use example (Source, Author)

**Objective 5:**

Can write, clear, efficient, and well-documented codes using comments and structured format using pipelines. Moreover, building of pipelines were helpful in writing a structured code to achieving tasks such as summarization and numeric calculations.

Through the user defined functions, I have learned how to create my functions on manipulating data, searching through rows in data frames and creating visualizations. The main benefit I gained with functions is that they significantly reduce the work of redoing some tasks that perform almost the same activities. For instance, the figure below, I created a function that conducts a search in a data frame and return all the rows that matches the string. In this case, I was able to reuse the code multiple times in extracting accurate data for visualization various financial metrics about the American Airlines (main organization chosen for our group project).

External resources such as StackOverflow have helped troubleshoot difficult problems in R. I mostly referred to StackOverflow at the start of classes when am not prepared well to understand how some functions are used in R.

**Grade:** I could give myself an A for this course. As noted by Kohn (2011) “Grades don’t prepare children for the real world.” However, through the activities and tasks have been able to complete using R, I believe I can get A grade, since I feel am prepared to do all that pertains to data analysis using R programming. Although my grade might not be important in the technical real-world, I believe the grade could offer me an opportunity to show my R programming skills in to people who still believe in the power of grades.

**References**

Kohn, A. (2011). The Case Against Grades. *Educational Leadership.*