Makenzie Barber

15 July 2022

Working with Data in R

Exploring data analysis (EDA) is key to understanding what your data is trying to tell you. In this report, I will explain and represent in graphs and exact words what the data within “SURVEY10” could tell a user. To begin, I will do a simple count represented by a bar graph of all women and men surveyed, along with the mean differences in both variables. This is done using simple vectors that only call for a specific gender. Next, I will compare data from a subset of all women and men surveyed, along with 3 specific variables (TxtPerDay, GPA, and MinPerDayFaceBook). For simplicity, I will graph GPA to easily spot the differences in GPA by gender. This was done using the “subset()” function, and selecting the aforementioned variables. These comparisons will be analyzed to tell of the specific differences in gender within this population. I have concluded that the differences in gender do not influence GPA. At the end, I will discuss the integrity and issues I discovered within the data, along with a conclusion recapitulating my findings.

The data I have been working with is in the *regclass* package. The data was collected in 2010 from students at the University of Tennessee, who were asked a series of 20 questions, resulting in the SURVEY10 data, with 20 variables and 699 observations. SURVEY10 has been cleaned up to an extent, but it is warned that issues may still exist.

Chart, bar chart

Description automatically generatedTo gain a better understanding of the data, I ventured into an exploratory data analysis by variable, beginning with a simple bar plot (Figure 1) of number of males to females (386 Females to 313 Males). This plot allows the user to better understand the data, as the other variables are answers from the students about themselves.

Graphical user interface, text

Description automatically generated with medium confidenceWith the population being 55% female, you could expect variables like “Height” and “Desired Weight” to have a lower mean, or be skewed slightly. To test this theory, I found the mean of desired weight and height respectively. For comparison, I created a vector of data filtered by “Female” only. I began by comparing desired weight of both genders first, then the desired weight of females.

From here, the user can acknowledge my hypothesis was proven true, as Desired Weight of females was 23.8 units lower than the vector with both genders. The Height vector with female data was 2.9 units lower than Graphical user interface, text

Description automatically generated with medium confidencethe Height vector with both genders. The female vector was created using “%>%” or “piping” method nested in the *dplyr* package, and then used the “filter” function to select values correlating with “Female” within the dataframe, and used the same methodology for values correlating with “Male”. This created a vector with 386 observations across all 20 variables which is flush with the number of women surveyed and the number of variables available.

Text

Description automatically generatedTo gain a deeper understanding of the data across more than one variable, I have created a subset of data for both genders. Using the “subset” function to store the required conditions, I have narrowed it down to 39 women and 33 men. To create these subsets, I utilized the female and male vectors created above. The subsets use the same variables acting as a filter, however, were tweaked based on gender-specific trends. For example, women spent less time on Facebook than men, so I had to adjust the female subset to only include women who spent less than 5 minutes on the website to narrow that one to less than 40 individuals.

Figure 2
 GPA by Gender (Figure 2), was created using the data for GPA using the FSub subset, which contains female data, and the MSub subset, which contains male data. I used the “plot” function to include the GPA for women, and the “lines” function to include the correlating male GPA. The reader will notice in this graph that the “Male” line drops off, and that is because of there is only 33 males. This graph simply highlights the main differences in GPA for subsets named above. I could have used the same methodology to create plots comparing the other variables in my subsets, however, I did not feel it necessary for this report.

The data frame “SURVEY10” is of most completeness and seemingly lacks any integrity issues. That is, until one takes a deeper look. While computing means for heights, I noticed that the overall height mean was around 5’5, which seems normal, but the mean for height in females alone was around 5’3, which seemed a little low, even for women. Upon further inspection, I noticed quite a few data points with heights in the 4’5 and below region. Although it is possible for people to be this height, it seemed irregular for multiple people at a college-age to be that height. This error in data entry compromises the integrity of the height data, and slightly skews Chart, histogram

Description automatically generatedthe data to the left for females and overall, indicating that the skew does come from the female part of the data.

I also ran into issues while creating subsets. I noticed that the easiest way to create subsets may be using the “which” function, however, creating easily graphable data is done through using the subset function. For my specific code, R did not want to connect the “select=” function within “which”, but had no issues when using “subset.” When using “which”, R returned all variables for the female subset I was attempting to create, which made graphing extremely complex. With the “subset” function, I was able to select the columns in which I was interested in and attempts at graphing were less frustrating.

Chart, histogram

Description automatically generatedWith this data being 12 years old, future research could take the data from 2022 (or later), and use it to compare how much preferences, height, weight, and minutes per day on the internet have changed over the last decade. In the future, I would also subset more equally, so that both vectors have exactly the same amount of students.

To conclude, this data offered 20 different variables to isolate, with an identifying factor being gender. From there, one has endless opportunities to do multiple different statistical analyses and discover correlations within gender and other variables. In this report, I discovered that more men were surveyed than men, however, that does not make the averages for women’s data higher. I also discovered that GPA is not reliant on what gender you are. With SURVEY10 being an entire survey of college students, the data under the variables provided must be considered differently to identify easily missed data integrity issues.