Introduction to RStudio

This first data analysis assignment is intended to help you familiarize yourself with the RStudio interface and get comfortable running small chunks of R code. If you have not done so already, please work through the R Tutorial in the Start Here module on Canvas.

For this assignment, you will need the Intro\_to\_RStudio.R file and the loan50.csv dataset. Both can be found on the Data Analysis 1 assignment page.

You can find a description of the variables recorded in the loan50.csv dataset on the [OpenIntro Statistics loan50 info page](https://www.openintro.org/data/index.php?data=loan50).

The Intro\_to\_RStudio.R script walks you through using some of the basic, built-in functions in R. Read through and run each line of code to ensure you understand what the functions are doing and what types of output each produces. The script uses two variables: loan\_amount and homeownership. For the assignment you’ll submit, you will practice using two **different variables**. Please make sure the assignment you submit uses the correct variables (specified in the questions below).

## Part 1 - Exploring a Single Quantitative Variable

For this portion of the assignment, you’ll practice using R to explore the annual\_income variable in the loan50.csv data set.

a. (2 points) Construct a histogram of the annual income data. Include informative labels and a title. Include your histogram below. *To copy or save a graph from RStudio, click the Export button just above the preview of the graph. From there you can choose to Save Image or Copy to Clipboard.*

Chart

Description automatically generated

b. (2 points) Construct a boxplot of the annual income data. Include informative labels and a title. Include your boxplot below.

Chart, box and whisker chart

Description automatically generated

c. (2 points) Using the histogram you constructed in part a and the boxplot from part b, describe the shape of the distribution of the annual income variable and comment on the presence of any outliers.

Considering the histogram constructed in part a and the boxplot from part b, the shape of the distribution of the annual income variable is right-skewed. There are 4 outliers based on the boxplot from part B.

d. (1.25 points) Calculate the mean of the annual income data.

The mean of the annual income data is $86,170 (formula: **mean(loan50$annual\_income)**).

e. (1.25 points) Calculate the median of the annual income data.

The median of the annual income data is $74,000 (formula: **median(loan50$annual\_income)**).

f. (2 points) Which measure of center (mean or median) is more appropriate for these data? Why? Consider the shape of the distribution discussed in part c.

The median would be the most appropriate measure of center for this data because looking at the income for the households, there exists a couple of outliers due to the fact of a few billionaires but their percentage is relatively small. Therefore, the mean is not resistant to outliers, which pulls the mean value up. The median on the other hand is resistant to outliers. So, the median would be a better measure of central tendency.

g. (1.25 points) Calculate the standard deviation of the annual income data.

The standard deviation is 57566.5 (formula: **sd(loan50$annual\_income)**).

h. (1.25 points) Calculate the interquartile range of the annual income data.

The interquartile range is 43,750 (formula: **IQR(loan50$annual\_income)**).

## Part 2 – Visualizing Two Variables

Let’s continue to explore the annual income data, but now consider how annual income data may vary between loan status (current or fully paid).

a. (2 points) Construct a side-by-side boxplot for annual income broken up by loan status. Include informative labels and a title.

Chart

Description automatically generated

b. (2 points) How do the distributions of annual income compare for loan status? Comment on the shape, center, spread, and presence of outliers for the two groups.

The shape of the “Distribution of Annual Income based on loan status” is more rightly skewed than the shape of the general “Distribution of Annual Income”, which is less rightly skewed. The mean of the annual income data is $86,170 and the median is $74,000. Therefore, both boxplots are almost evenly spread out with the black line in the second boxplot close to the center. In both boxplots, there are 4 outliers.

## Part 3 – Exploring a Single Categorical Variable

Finally, we’ll focus our attention only on the loan status variable.

a. (2 points) Construct a table of counts for the loan status variable. Report the number of observations in each category below.

Constructing a table for the loan status column reports 6 fully paid cases and 44 who has yet to pay the loan amount among the 50 loan cases.

A picture containing graphical user interface

Description automatically generated

b. (2 points) Construct a table of proportions for the loan status variable. Report the proportions for each category below.

Constructing a table for the loan status column reports 12% of fully paid cases (0.12) and 88% who has yet to pay the loan amount (0.88) among the 50 loan cases.

Table

Description automatically generated with low confidence

c. (2 points) Construct a barplot that displays the distribution of loan status types. Include informative labels and a title. Include your barplot below.

Chart

Description automatically generated with medium confidence

**Gradescope Page Matching (2 points)**

When you upload your PDF file to Gradescope, you will need to match each question on this assignment to the correct pages. Video instructions for doing this are available in the Start Here module on Canvas on the page “Submitting Assignments in Gradescope”. Failure to follow these instructions will result in a 2-point deduction on your assignment grade. Match this page to outline item “Gradescope Page Matching”.