ST 314 Data Analysis 1

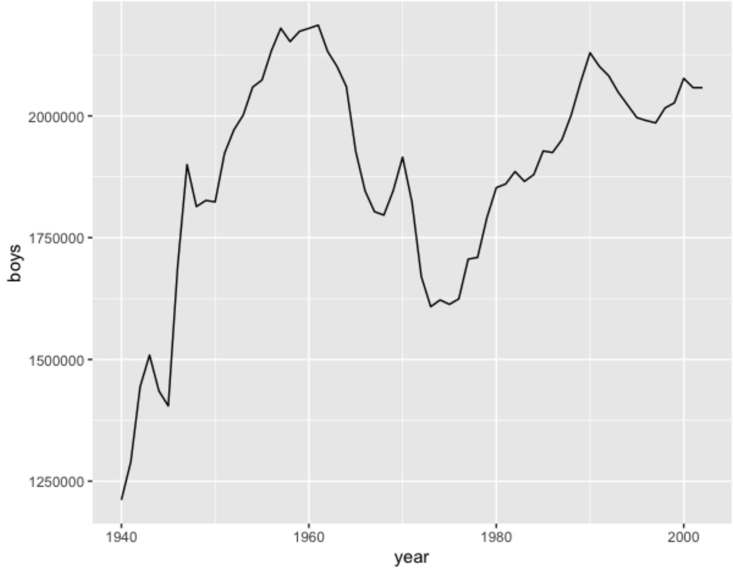
Part 1: Present Day Birth Records in the United States

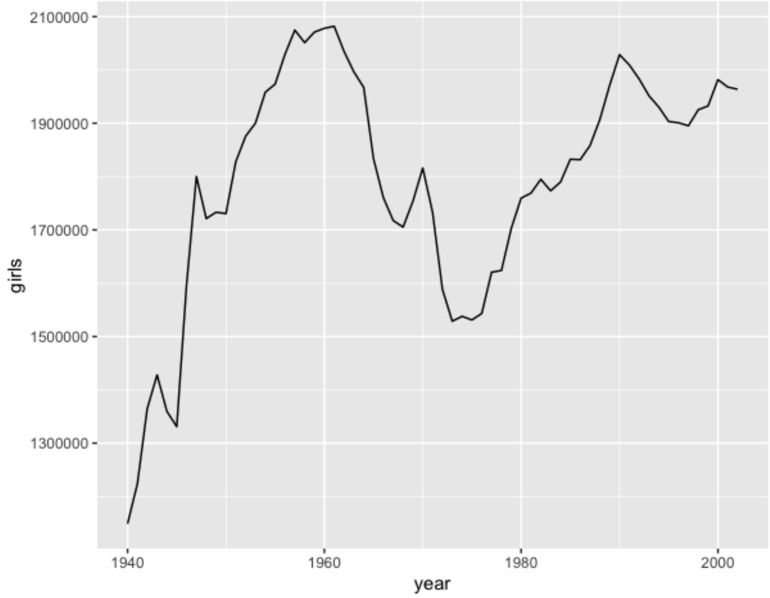
a. (1 point) What years are included in this dataset?
The dataset included the years from 1940 to 2022(e.g., 940, 1941, 1942, 1943, 1944, 1945, 1946, 1947, 1948, 1949, 1950,...,2022)

b. (1 point) What are the dimensions of the data frame? Dimensions refers to the number of rows and variables.
There are 63 rows and 3 variables.

c. (1 point) What are the variable (column) names?
Variable names are year, boys, and girls.

d. (1 point) Make a plot that displays the proportion of boys born over time. What do you see? Does Arbuthnot's observation about boys being born in greater proportion than girls hold up in the U.S.? Include the plot in your response. *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.*





I saw the minimum amount of boys were born in the year 1940. Yes, according to two graphs above that **boys** being born in greater proportion than girls hold up in the U.S. I also used the code `present <- present %>%`
`mutate(more_boys = boys > girls)` to compare birth of girls and boys each year, the result shows all true from 1940 t0 2002, therefore boys being born in greater proportion than girls hold up in the U.S.

`e.` (1 point) In what year did we see the most total number of births in the U.S.? *Hint:* First calculate the totals and save it as a new variable. Then, sort your dataset in descending order based on the `total` column. You can do this interactively in the data viewer by clicking on the arrows next to the variable names. To include the sorted result in your report you will need to use two new functions. First we use `arrange()` to sorting the variable. Then we can arrange the data in a descending order with another function, `desc()`, for descending order. The sample code is provided below.

```
present %>%
arrange(desc(total))
```

`The most total number of births in the U.S in 1961.`

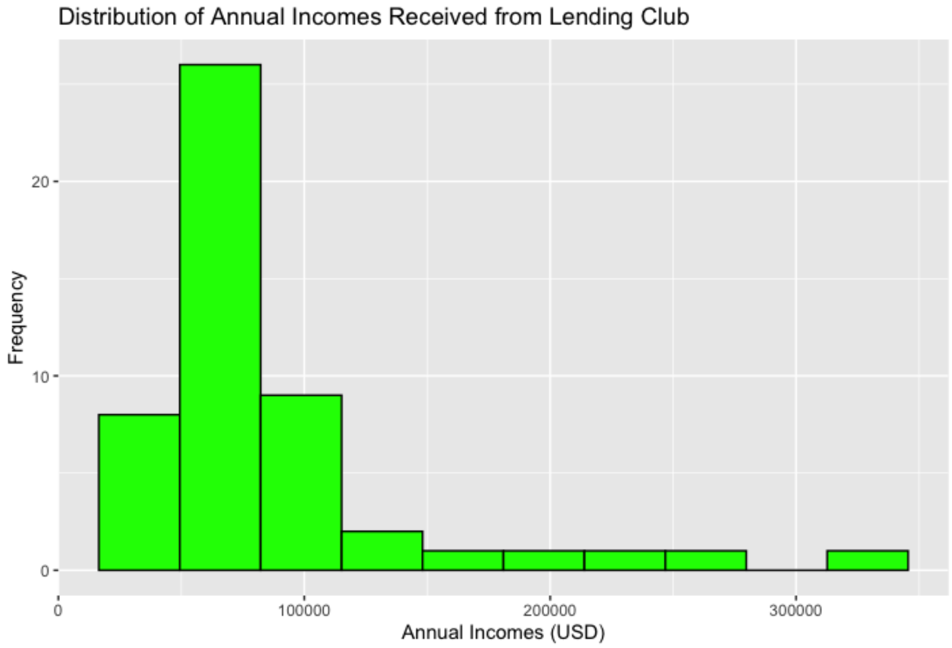
Part 2: Loan Data from Lending Club

Part 2 of the Introduction to RStudio & Tutorial uses two variables from the Lending Club data: 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).

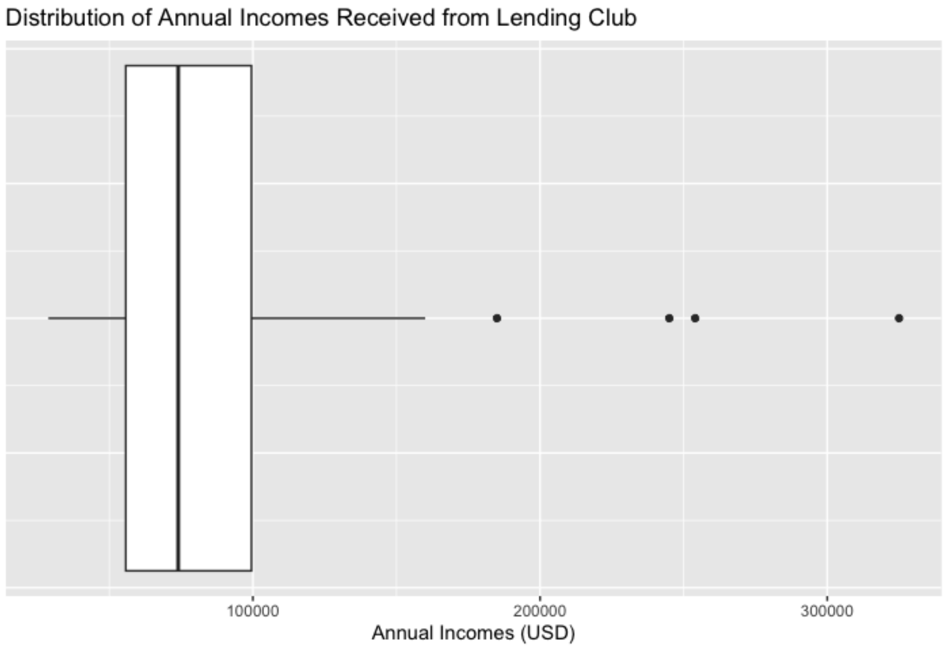
Exploring a Single Quantitative Variable

For this portion of the assignment, you'll practice using R to explore the **annual_income** variable in the loan50.csv data set.

a. (1 point) Construct a histogram of the annual income data. Include informative labels and a title. Include your histogram below.



b. (1 point) Construct a boxplot of the annual income data. Include informative labels and a title. Include your boxplot below.



c. (1 point) 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.

The shape of the histogram is right-skewed and unimodal. Unimodal means there is only one maximum in the histogram. There are 4 outliers in the boxplots: the first is 185000, the second is 245000, the third is 254000, and the fourth is 325000. These outliers are extreme values that fall far from the center.

d. (1.5 points) Calculate the mean of the annual income data.

Mean of the annual income is: 86170

e. (1.5 points) Calculate the median of the annual income data.

Median of the annual income is: 74000

f. (1 point) Which measure of center (mean or median) is more appropriate for these data? Why? Consider the shape of the distribution discussed in part c.

According to the right-skewed and Unimodal shape, the median is more appropriate to measure the center for these data. Because there are some extreme data in the dataset using the median to measure the center of the dataset will reduce the influence of those extreme data.

g. (1.5 points) Calculate the standard deviation of the annual income data.

The standard deviation of the annual income data is :57566.5

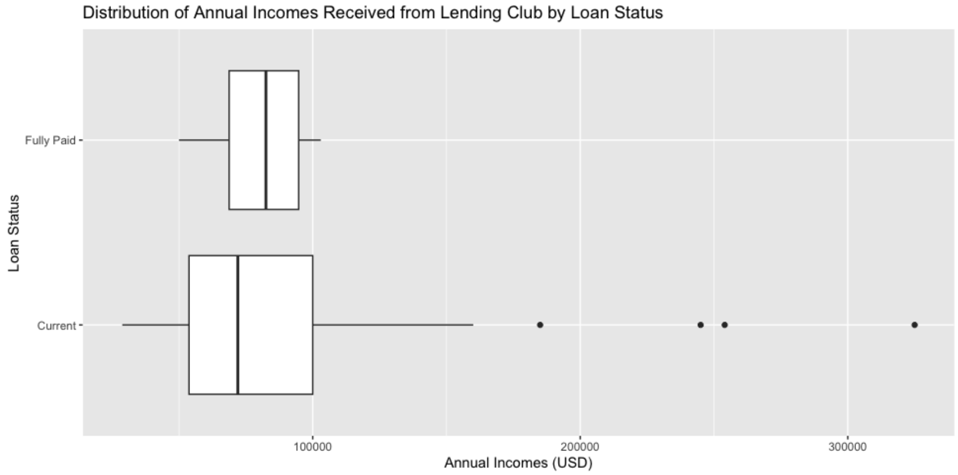
h. (1.5 points) Calculate the interquartile range of the annual income data.

The interquartile range of the annual income data is: 43750

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).

i. (1 point) Construct a side-by-side boxplot for annual income broken up by loan status. Include informative labels and a title.



j. (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.

According to two boxplots:
Shape: The shape for the fully paid is symmetric but the shape for the Current looks like right skewed.
Center: For the Current’s median is smaller than the median of Fully Paid.
Spread: The interquartile range of Fully Paid is narrower than the interquartile of Current.
Presence of outliers: for the Fully Paid there are no outliers but for the Current there are four outliers in the boxplot.

Exploring a Single Categorical Variable

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

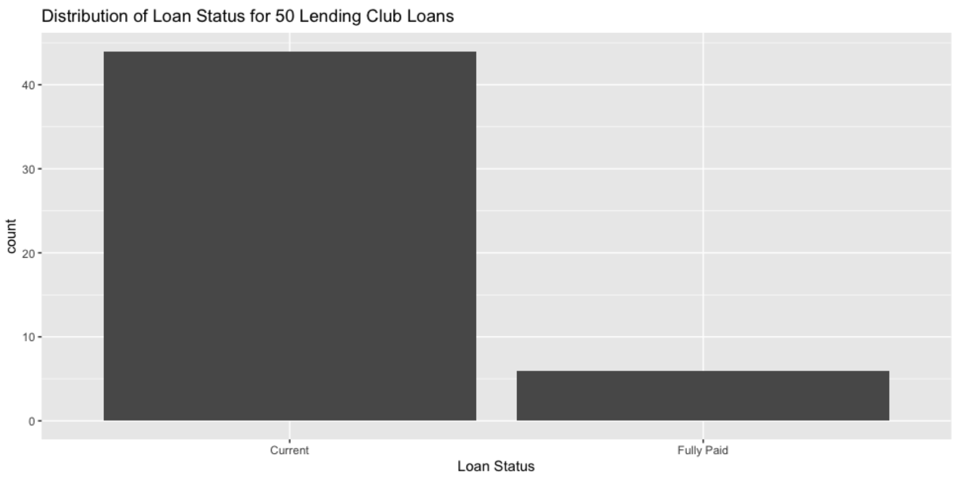
k. (2 points) Construct a table of counts for the loan status variable. Report the number of observations in each category below.

Current: 44 Fully Paid: 6

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

Current: 0.88 Fully Paid: 0.12

m. (1 point) Construct a barplot that displays the distribution of loan status types. Include informative labels and a title. Include your barplot below.



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".