Lab 1 Summarizing Data

Instructions: Read through and answer or implement the instructions given below. When answering the questions below provide justification for your answer by including the commands and/or the computer output used to answer the question. Submit a written lab report with your answers in Canvas as a Word or PDF document.

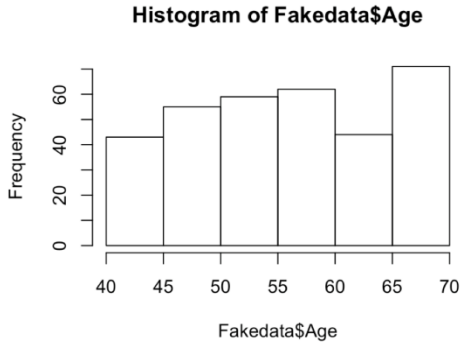
1. Load the Fakedata as you did in the previous lab.
2. **Numerical Summaries of Quantitative Variables**
   In the previous lab you computed means and medians. We would like to extend the tools you have for quantitative variables, in particular for computing the standard deviation. Using what you learned previously about the $ operator, compute the standard deviation of the Pinkylen using the *sd()* command and verify that you get the following:

   Based on this and your previous work you can now compute the mean, median, standard deviation, maximum and minimum of quantitative variables. The other summary that you will use is the IQR, which you will learn later in this lab.

3. **Histogram**
   Creating nice graphics using *R,* follows a similar pattern. Suppose we want a histogram of the Age variable. We can do this by simply entering the command

   hist(Fakedata$Age)

   Which results in something like the following:



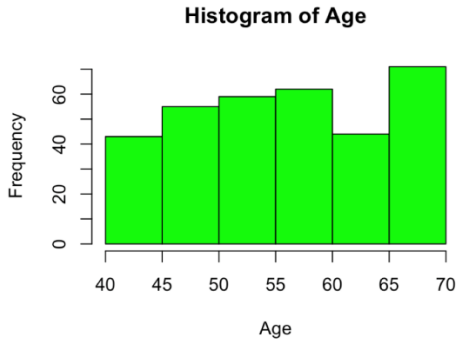
   We can also add other options to color the histogram or to change the number of bins using the breaks command and specifying a number, such as:

   hist(Fakedata$Age, breaks=10, col="green")

   Note the breaks, doesn't always produce exactly the number of boxes as you specify as it does some work to decide how many bins it things is best for your data, but it does provide some flexibility. Further, we can set labels using the additional options main, xlab and ylab, for the title, x-axis label and y-axis label. For example

   hist(Fakedata$Age, breaks=10, col="green", main="Histogram of Age", xlab="Age")

   This would produce the following plot:



Create a histogram of the *Pinkylen* variable, adjust the color to blue, and set the main title to "Histogram of Pinky Length" and the x-axis label to an appropriate label. What is the command to do this? Include a screenshot of the output in your lab report.

4. **Boxplots**

Boxplots are nearly as easy. The command to create a boxplot of the Age variable is

boxplot(Fakedata$Age, main = "My Boxplot", xlab = "Age", col = "green")

Construct a blue boxplot of the Height variable and label the graph appropriately. Provide the command and the output in your lab report.

You may also manually compute the quartiles using the command *quantile()*, which produces all of the quartiles or compute the IQR using the command *IQR().* You need not compute these here.

5. **Summarizing Categorical Data**

Categorical data is only slightly more complicated as it requires that we count the values in each category and then turn them into percents. Once we have this table of percents we can construct the boxplot. We will construct the table using the *table()* command and saving the results as tbl:

tbl <- table(Fakedata$FavColor)

Now enter the command *tbl* to view the counts. You should get:

```
Brown Green  Grey Other
   33   136    64   101
```

Converting these counts to percents requires totaling the number of values and then multiplying by 100. We will do this and store the results in the variable *proptbl.*

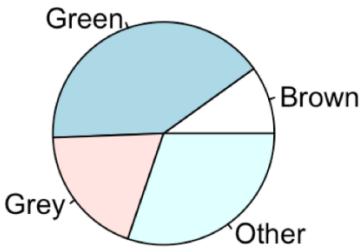
*proptbl <- 100*tbl/sum(tbl)*

You should enter the command *proptbl* to verify that you have

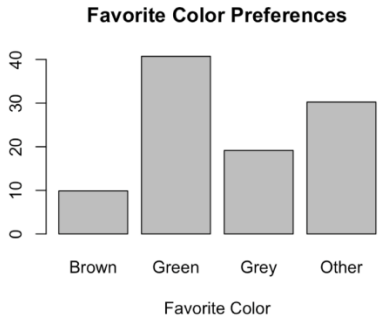
```
   Brown    Green     Grey    Other
9.88024 40.71856 19.16168 30.23952
```

Now creating a pie chart or a bar graph is quite simple. The commands for a pie chart and a bar graph are below along with their output.

pie(proptbl)



barplot(proptbl,xlab = 'Favorite Color', main = 'Favorite Color Preferences')



What are the commands required to compute a bar graph for the Breakfast variable and label it with the appropriate labels?

Application Questions

Read in the data file Italian.csv as you did before. This data are the simulated results of a study attempting to characterize NATO countries as a whole.

> The Civilian American and European Surface Anthropometry Resource (CAESAR) project was a survey of the civilian populations of three countries representing the North Atlantic Treaty Organization (NATO) countries: the United States of America (USA), The Netherlands, and Italy (Robinette et al. 1999, Robinette 2000). One site in Ottawa, Canada was added to the USA sample and it is henceforth referred to as the North American sample. The survey was carried out by the U.S. Air Force, with the help of 1) the contractor, Sytronics Inc., 2) The Netherlands Organization for Applied Scientific Research (TNO), 3) the subcontractor D'Appolonia in Italy, and 4) a consortium of companies under the umbrella of the Society of Automotive Engineers (SAE). (See http://store.sae.org/caesar/ for more information.)

Using this data answer each of the following questions using the appropriate numerical **AND** graphical tools for the particular type of variable. Note that we will be learning more numerical summaries that would also be a typical part of the summary of quantitative variables. Answer each question with a couple of sentence description of your results and include the computer commands and output used to justify your answer. For example, the explanation might look like: "The population mean age is 32.4, indicating a young population. Further, a histogram of the ages shows a skewed right distribution, with much of the population in the younger ages."

6. Summarize the height of the sample included in this survey.
7. Summarize the ages of the sample included in this survey.
8. Summarize the ethnic diversity of those included in this survey.
9. Summarize the gender diversity in those included in this study.