Homework 1: Upload by Sunday 29th January by 11:59 pm

Introduction to RMarkdown, Examining Distributions & Examining Relationships

In this assignment you will apply skills in ***examining distributions*** and ***examining relationships*** while creating a document in the RMarkdown format. All statistics homework assignments will be submitted as RMarkdown files, that is, files with the .RMD extension.

Preparation Tasks

1. Before beginning Homework 1 you should have completed the following OLI modules:

   Module 4: Examining Distributions
   Module 5: Examining Relationships

2. From the course site, download the data file "studentData.txt" to the folder in which you store data files for use in R Studio.
3. You may want to have the link http://rmarkdown.rstudio.com/authoring_basics.html open as a reference.

RMarkdown Setup Instructions

1. In R Studio select File→New File→RMarkdown. Select Yes when prompted to install the necessary packages. You must be connected to the internet. This will take a few minutes.
2. Enter Homework 1 as the Title, your name as the Author, and select MS Word as the Default Output Format, then select Okay. A new file will open with a header similar to the one below:

```
---
title:  "Homework Assignment 1"
author: "Eric Ocran"
date:   "14/01/2023"
output: word document
---
```

   An RMarkdown file has three components: metadata, R code, and text. These components are combined when the file is "knitted" to generate an output document, which in our case will be in MS Word format. The specifications for the output document are called metadata and are contained between a pair of triple dashes at the top of the file. For our purposes, we will stick to title, author, date and output, however there are many other specifications that could be included. Metadata uses YAML syntax which is common to many markup languages.

   R code can be entered into an RMarkdown file in two ways, either in a code chunk, which begins with ```` ```{r} ```` and ends with ```` ``` ```` and is shaded grey, or as inline code, which begins with `` `r `` and ends with `` ` ``. Note that the ticks are backticks (the backtick key is usually to the left of the 1-key on the upper left of the keyboard). We will learn to use code chunks first and inline code later in the semester.

   Regular text can be typed anywhere in the document except inside the metadata header and within a code chunk. There are special characters for formatting text and entering math symbols which we will learn as the semester progresses.

   Now, **delete everything below the metadata header** (it's not needed) and then save your RMarkdown file. Be sure to save it to a folder you have created for statistics homework.

3. All homework assignments will begin with a code chunk labeled setup just below the metadata header. Begin by typing:

````
```{r setup, message = FALSE}


```
````

The setup code chunk is where you will install any packages that are needed for the assignment (packages are collections of functions or datasets) as well as import or load any data files that will be needed. For Homework 1 you will need to load the readr package, which contains functions for importing data into R, and you will need to import the studentData.txt data file. (The `message = FALSE` option suppresses R messages from showing up in the knitted MS Word file.)

Type the code `library(readr)` into the code chunk and run the line of code (you can run a line of code by keying Ctrl+Enter or you can use the Run menu). The code chunk should now look like this:

````
```{r setup, message = FALSE}
library(readr)

```
````

You can verify that the readr package has been loaded by clicking on the Packages tab in the lower right pane in R Studio, scrolling down and noting that the box next to readr has been checked.

Now we will import a studentData.txt file which will be needed to complete the homework assignment. Select the Environment tab in the upper right pane, click on the Import Dataset menu and select From Text(readr). This will bring up a dialogue box that will allow you to browse to and select the studentData.txt file. Under Import Options change Delimiter to Tab and then choose Import.

The R code for importing the data also needs to go into the RMarkdown file. When you selected the file to import, code was generated by R Studio and run in the console. Copy and paste the `read.delim()` code from the console into your setup code chunk (remove the console prompt > and any + symbols from the pasted code). Even though the code has already been run and the data file is loaded in your current environment (you should see studentData in the Environment pane), the data file will need to be loaded each time the RMarkdown file is knitted into MS Word. Your setup code chunk should now look similar to the one below (but a different file path!).

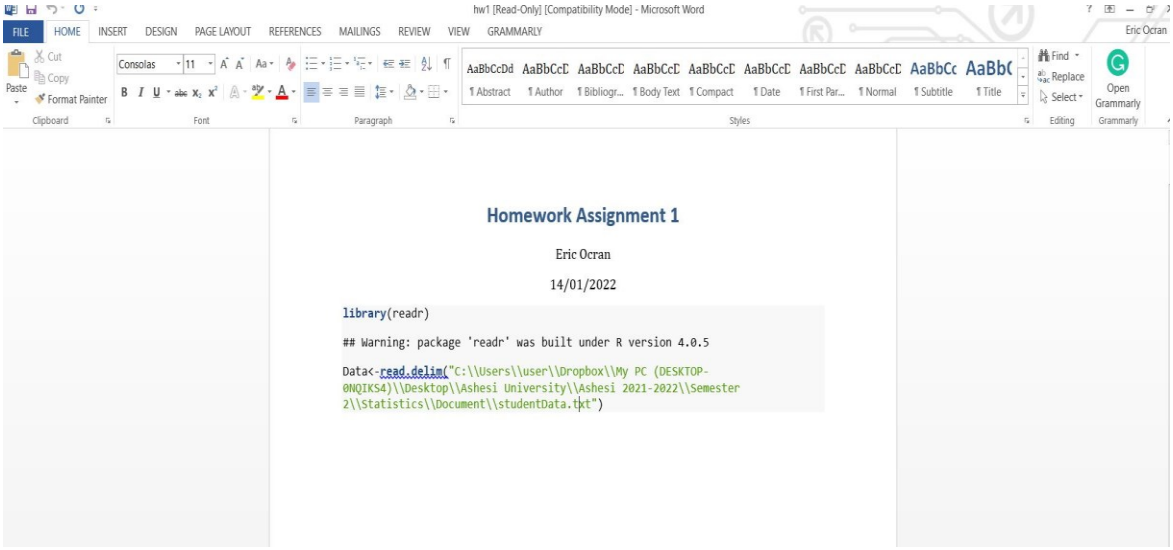
````
```{r setup, message = TRUE}

library(readr)

Data<-read.delim("C:\\Users\\user\\Dropbox\\My PC (DESKTOP-
0NQIKS4)\\Desktop\\Ashesi University\\Ashesi 2021-2022\\Semester
2\\Statistics\\Document\\studentData.txt")

```
````

Finally, save your RMarkdown file and then click the Knit button to render what you have so far into MS Word. If you receive any errors get help to resolve them before moving on. The most likely error is that you forgot to remove the console prompt and any + symbols from the code you copied from the console into your code chunk. The R language is case sensitive which is another common source of errors. The knitted MS Word document should look something like the image below:



Now, close the MS Word file and continue with the assignment. It is good practice to render (Knit) after completing each homework question to check your work and troubleshoot any errors as you go along.

Homework Questions

1. Recall from OLI Module 4 the visual displays that are appropriate for a single categorical variable and a single quantitative variable. Also, you may want to refer to OLI sections that give the R code for creating pie charts, tables, and tables of proportions, histograms, and boxplots.

   Type `### Question 1` into RMarkdown a couple of lines below the setup code chunk. In RMarkdown, the hash symbol indicates that the text is a header when the file is rendered.

   Next, create a new code chunk below the Question 1 header by typing ```` ```{r} ```` followed by a couple of enters and then ```` ``` ````.

   A good strategy is to test code first in the console and then copy and paste it into the code chunk (don't forget to remove the console prompt > and any + signs).

   Below the code chunk, give a written response to the homework question if one is required.

   a. Create a pie chart to show visually the percentage of the program majors of the statistics students. The pie chart should have a relevant main title and each section of the pie chart should be appropriately labeled, including the % value (see code example in OLI Module 4 Learn By Doing activity for Pie Charts on page 17). Give a written summary of the results below the code chunk that generates the pie charts. Remember, in statistics a summary includes *numbers*, *units,* and *context*.

   b. Create a bar chart to visualize the distribution of a second categorical variable in the statistics students survey data that has more than two categories (i.e. not gender or yes/no questions). The height of the bars should show percent values and be in descending order. The bar chart should have a main title, category labels, and an appropriate y-axis label.

      Below is a sample code for a hypothetical question about the favorite chocolate bar of statistics students.

```
#tabulate and view counts for each factor
t1 <- table(studentData$chocolate);t1

#reorder the factor levels from highest to lowest based on observed counts in t1
choc <- factor(studentData$chocolate, levels = c("Kit
Kat","Cadbury","Goldentree","Bounty","Other"))

#tabulate and view counts for each factor which are now in descending order
t2 <- table(choc);t2

#generate the bar chart with a title and y-axis as percentages
barplot(100*t2/length(choc), main = "Chocolate Bar Preferences of Ashesi
Students", ylab = "Percentage")
```

2. Type `### Question 2` into RMarkdown a couple of lines below the code chunk and text for Question 1. Next, create a new code chunk by typing ```{r} followed by a couple of enters and then ```.

    a. Create two histograms, one showing the distribution in the weekly expenditure of statistics students on food and a second histogram showing the age distribution of the Statistics students. Be sure to include a descriptive main title and appropriate x and y-axis labels. Below is some sample code from the histogram example from class.

    ```
    hist(data$score, breaks = 20, col = "lightblue", ylim = c(0, 5), xlim = c(0, 25), main="Flappy Bird
        Scores", xlab = "Highest Score in 2 Minutes", ylab = "Counts")
    ```

    b. If you think the distributions are roughly symmetrical, calculate the mean and standard deviation as measures of center and spread. If you think the distributions are skew, calculate the median and IQR as measures of center and spread. **Calculations should be done using R code in the code chunk for question 2**.

    c. Write three to four sentences comparing the two distributions. Say something about the shape, center, spread, and whether there appear to be outliers in each distribution.

    Be sure to knit the document after every homework question in order to test that everything renders properly. Waiting till the end to knit makes it more difficult to troubleshoot any problems.

    Use `### Question X` to begin each new question

3. The Governor of the Bank of Ghana collected data on the bounced check fees, in cedis, for a sample of 25 banks for direct-deposit customers who maintain GH₵1,000 balance. Below is a stem-and -leaf display of the data generated in R.

    9 | 147

    10 | 02238

    11 | 125566777

    12 | 223489

    13 | 02

Note: 11 | 7 = 117

    a. List the first 5 values in the data set.
    b. What is the median amount of fees paid?
    c. Does the dataset have a mode? Explain.
    d. Complete the following sentence “The middle 50% of customers paid between ____ and ____ cedis as fees.

e. What is the IQR for the dataset?
f. Determine if there are any outliers using the IQR rule. (Show calculations for the IQR rule in a code chunk.)

4. A box and whisker plot is a useful visual display for exploring relationships between a categorical explanatory variable and a quantitative response variable, that is C → Q relationship. Look through the statistics students' data frame. Identify a quantitative variable that you think might differ in part by one of the categorical variables. For example, is the number of hours spent on studies different for CS, BA, and MIS students? In this case, Program major is the categorical explanatory variable, and hours spent on studies is the quantitative response variable. Here's another example: is weekly expenditure related to gender?

    a. Use the sample code below as a guide to generate side-by-side boxplots for the two variables you choose (do not choose age and gender).

    ```
    boxplot(studentData$age ~ studentData$gender,main="Age of Statistics Students by
            Gender",ylab="Age (years)")
    ```

    The ~ symbol reads "modeled as" or "a function of" so the command above instructs R to display the quantitative response variable age "as a function of" the categorical explanatory variable gender.

    b. Calculate the 5-number summary of the quantitative response variable for each group of the categorical explanatory variable. This is done with the *tapply()* function. Use the code below as a guide.

    ```
    tapply(studentData$age,studentData$gender,fivenum)
    ```

    c. Write three to five sentences summarizing the relationship between the explanatory and response variables. Don't forget to include numbers, units and context in your summary.

5. The visual display for exploring the relationship between two quantitative variables, Q → Q, is a scatterplot. Refer to the code in the two OLI Module 5 Learn by Doing activities on Linear Relationships.

    Choose two quantitative variables from the statistics students' data set. Identify one as the explanatory variable (x-axis variable) and one to be the response variable (y-axis variable). If there is no clear dependency, this may be an arbitrary decision.

    a. Generate a scatterplot of the data. The plot should have a descriptive main title and appropriate x and y-axis labels **with units**. (Note that in the plot() function the x-axis variable comes first and the y-axis variable second)
    b. Generate a linear model and add the regression line to the scatterplot. Note that in the lm() function, the format is lm(y-axis variable ~ x-axis variable).
    c. Calculate the correlation coefficient. (Note that in the cor() function, the x-axis variable comes first and the y-axis variable comes second; if cor() returns NA, then there is likely missing data, add the parameter use = "complete.obs" as the third argument in the function to tell R to remove missing data).
    d. Based on the pattern you observe in the scatter plot and the value for the correlation coefficient r, does there appear to be a linear relationship between the two variables? Comment on the direction and strength of a linear relationship and any possible outliers.

    Below is some sample code.

    ```
    plot(Data$SocialMedia,Data$Sleep, main="Hours of Sleep as a function of Hours on Social
        Media",xlab="Time on social media (hours)",ylab="Hours of sleep")
    model = lm(Data$Sleep ~ Data$SocialMedia)
    abline(model)
    cor(Data$Sleep,Data$SocialMedia,use="complete.obs")
    ```

6. The operations manager of Hankook tires wants to compare the actual inner diameters of two grades of tires, each of which is expected to be 575 millimeters. A sample of five tires of each grade was selected, and the results representing the inner diameters of the tires, ranked from smallest to largest, are as follows:

   Grade A : 588, 579, 585, 580, 583
   Grade B : 570, 574, 572, 571, 578

   a. For each of the two grades of tires, compute the mean, median, and standard deviation.
   b. Which grade of tire is providing better quality? Explain.
   c. What would be the effect on your answers in (a) and (b) if the last value for grade B were 588 instead of 578? Explain.

ASSIGNMENT SUBMISSION: Knit the final version of your completed assignment in order to make sure it renders properly in MS Word. Troubleshoot any problems. Important: Upload the .**RMD** file to Canvas (not the MS Word file!)**.** The first step in the grading process for homework is knitting the submitted .RMD file. If it doesn't render, your homework will be reduced to a maximum C grade.

Upload by 11:59 pm on Sunday.

The only time a pie chart is appropriate is at a baker's convention.

Never show a bar chart at an AA meeting.