Final Exam: Homeworks

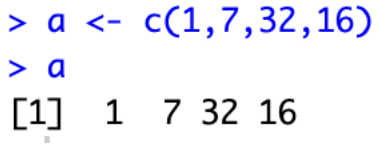
BUSN 231 - Applied Business Statistics
Homework (Exercise)#2
Due Date: September 3, 2023 at 11:59pm

Question#1 (32 points) Use R to compute the following values. After you do so, copy and paste your input and output from R to Word. (Do this for every problem)

Part a)

```
a <- c(1, 7, 32, 16)
a
```

Output:

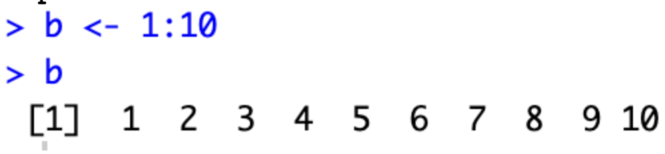


```
> a <- c(1,7,32,16)
> a
[1]  1  7 32 16
```

Part b) Sequences of integers may be created using a colon (:)

```
b <- 1:10
b
```

Output:

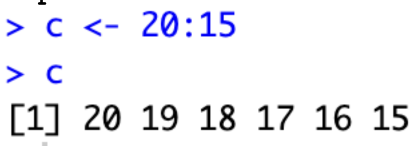


```
> b <- 1:10
> b
 [1]  1  2  3  4  5  6  7  8  9 10
```

Part c)

```
c <- 20:15
c
```

Output:

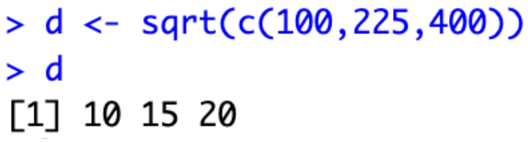


```
> c <- 20:15
> c
[1] 20 19 18 17 16 15
```

Part d) Basic mathematical functions will apply element-by-element

```
d <-sqrt(c(100, 225, 400))
d
```

Output:

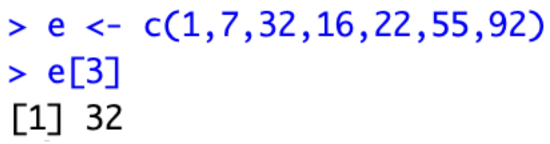


```
> d <- sqrt(c(100,225,400))
> d
[1] 10 15 20
```

Part e) To select subsets of a vector, use square brackets ([ ])

```
e <- c(1, 7, 32, 16, 22, 55, 92)
e[3]
```

Output:

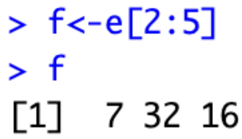


```
> e <- c(1,7,32,16,22,55,92)
> e[3]
[1] 32
```

Part f)

```
f<-e[2:5]
f
```

Output:

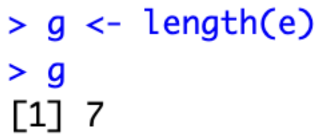


```
> f<-e[2:5]
> f
[1]  7 32 16 22
```

Part g) The number of elements in a vector can be found with the length command.

```
g <- length(e)
g
```

Output:

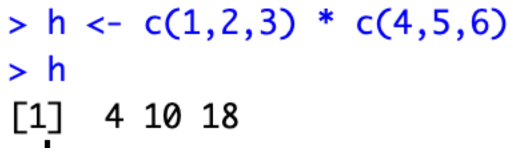


```
> g <- length(e)
> g
[1] 7
```

Part h) If a vector is passed to an arithmetic calculation, it will be computed element-by-element.

```
h <- c(1, 2, 3) * c(4, 5, 6)
h
```

Output:



```
> h <- c(1,2,3) * c(4,5,6)
> h
[1]  4 10 18
```

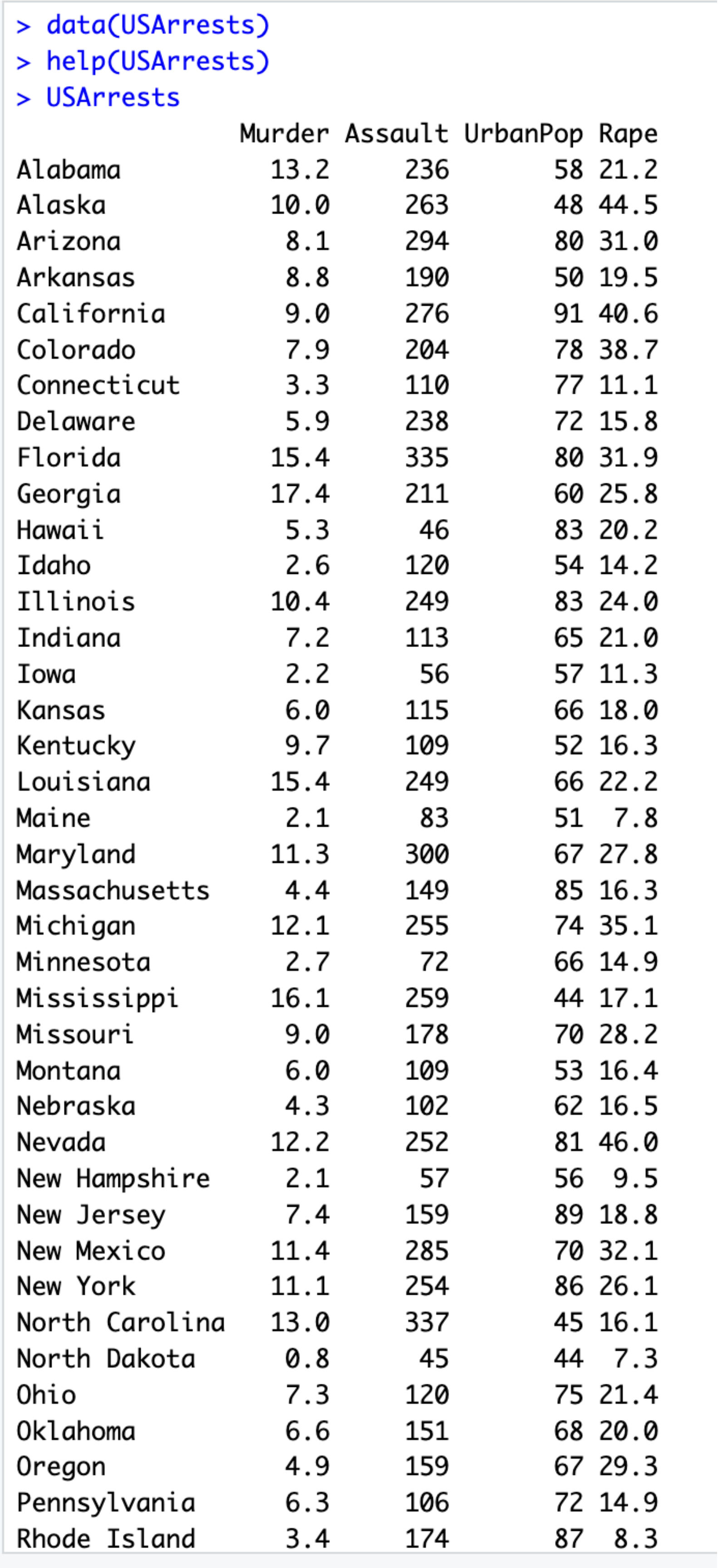
Question#2 (40 points)

This particular dataset, named **USArrests**, contains the number of arrests for murder, assault, and rape for each of the 50 states in 1973. It also contains the percentage of people in the state who live in an urban area. These data are included with R, and you can get the data object and put it in your workspace/environment as an object.

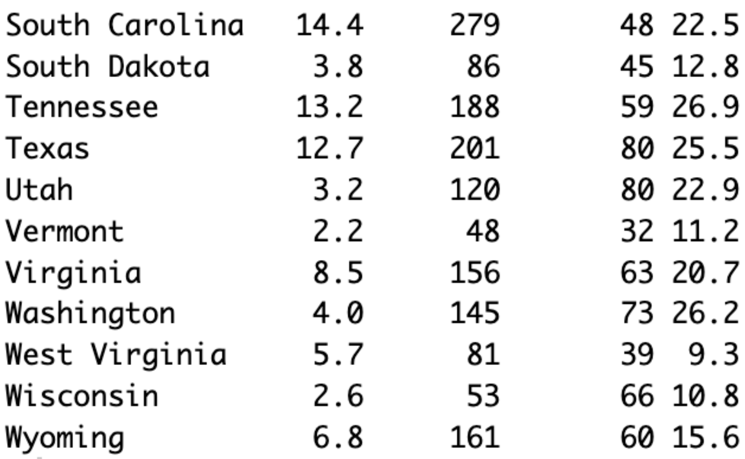
Part a) The data set is pre-loaded with R, so you can load it directly.

```
data(USArrests)
help(USArrests)
USArrests
```

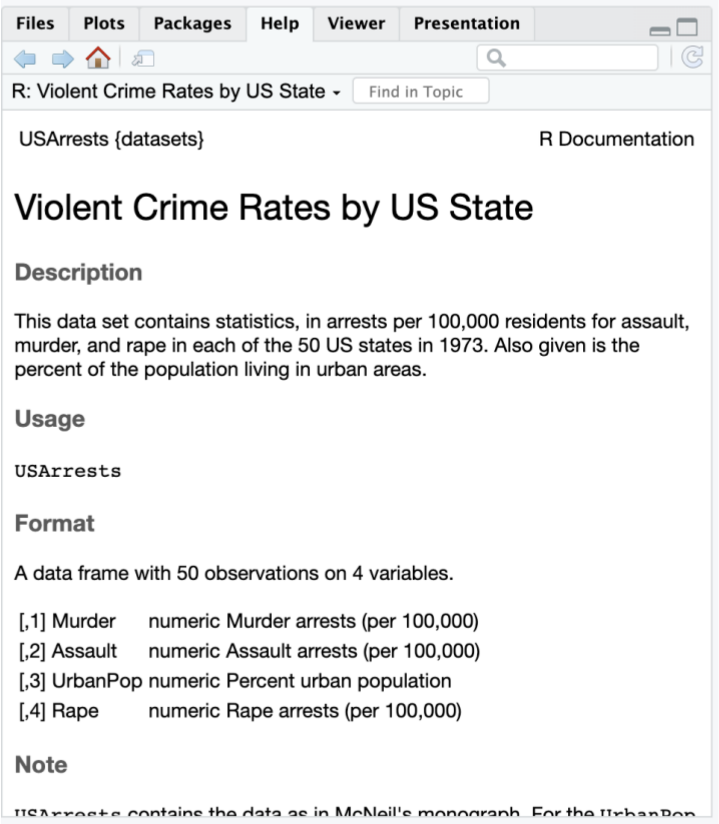
Output:



```
> data(USArrests)
> help(USArrests)
> USArrests
               Murder Assault UrbanPop Rape
Alabama          13.2     236       58 21.2
Alaska           10.0     263       48 44.5
Arizona           8.1     294       80 31.0
Arkansas          8.8     190       50 19.5
California        9.0     276       91 40.6
Colorado          7.9     204       78 38.7
Connecticut       3.3     110       77 11.1
Delaware          5.9     238       72 15.8
Florida          15.4     335       80 31.9
Georgia          17.4     211       60 25.8
Hawaii            5.3      46       83 20.2
Idaho             2.6     120       54 14.2
Illinois         10.4     249       83 24.0
Indiana           7.2     113       65 21.0
Iowa              2.2      56       57 11.3
Kansas            6.0     115       66 18.0
Kentucky          9.7     109       52 16.3
Louisiana        15.4     249       66 22.2
Maine             2.1      83       51  7.8
Maryland         11.3     300       67 27.8
Massachusetts     4.4     149       85 16.3
Michigan         12.1     255       74 35.1
Minnesota         2.7      72       66 14.9
Mississippi      16.1     259       44 17.1
Missouri          9.0     178       70 28.2
Montana           6.0     109       53 16.4
Nebraska          4.3     102       62 16.5
Nevada           12.2     252       81 46.0
New Hampshire     2.1      57       56  9.5
New Jersey        7.4     159       89 18.8
New Mexico       11.4     285       70 32.1
New York         11.1     254       86 26.1
North Carolina   13.0     337       45 16.1
North Dakota      0.8      45       44  7.3
Ohio              7.3     120       75 21.4
Oklahoma          6.6     151       68 20.0
Oregon            4.9     159       67 29.3
Pennsylvania      6.3     106       72 14.9
Rhode Island      3.4     174       87  8.3
```



```
South Carolina   14.4     279        48 22.5
South Dakota      3.8      86        45 12.8
Tennessee        13.2     188        59 26.9
Texas            12.7     201        80 25.5
Utah              3.2     120        80 22.9
Vermont           2.2      48        32 11.2
Virginia          8.5     156        63 20.7
Washington        4.0     145        73 26.2
West Virginia     5.7      81        39  9.3
Wisconsin         2.6      53        66 10.8
Wyoming           6.8     161        60 15.6
```



USArrests {datasets} R Documentation

Violent Crime Rates by US State

Description

This data set contains statistics, in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states in 1973. Also given is the percent of the population living in urban areas.

Usage

```
USArrests
```

Format

A data frame with 50 observations on 4 variables.

[,1] Murder numeric Murder arrests (per 100,000)
[,2] Assault numeric Assault arrests (per 100,000)
[,3] UrbanPop numeric Percent urban population
[,4] Rape numeric Rape arrests (per 100,000)

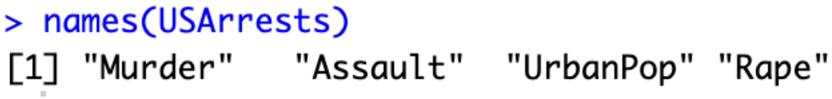
Note

USArrests contains the data as in McNeil's monograph. For the UrbanPop

Part b) You can look at the names contained in this data object (or data frame, as R calls them) by typing:

```
names(USArrests)
```

Output:

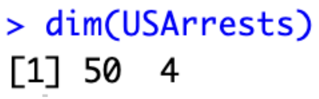


```
> names(USArrests)
[1] "Murder"   "Assault"  "UrbanPop" "Rape"
```

Part c) You can look at how big the data frame is by typing:

```
dim(USArrests)
```

Output:



```
> dim(USArrests)
[1] 50  4
```

Note: please explain the meaning of the numbers

The meaning of the numbers are the size of the dataset, with 50 representing the number of rows or observations (US States), and 4 representing the number of columns or variables (Arrest Charge)

Part d) You can also click on the little spreadsheet icon in the Environment tab to look at the data in spreadsheet format. Alternatively, you can type:

```
View(USArrests)
```

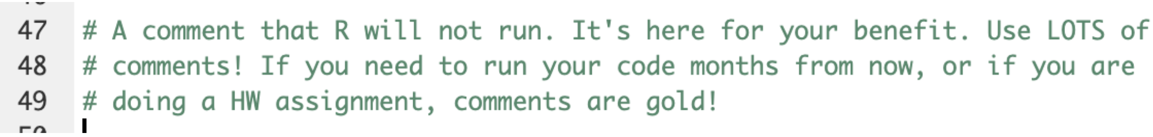
Output:Please explain what happened

View (capital V, the lowercase is not recognized) converts the data set into a sortable table that allows me to view the data

Part e) To include a comment in an R script, type a "#" sign at the beginning of the line. Please type the followings in your R script and take a screenshot and past below.

```
# A comment that R will not run. It's here for your benefit. Use LOTS of
# comments! If you need to run your code months from now, or if you are
# doing a HW assignment, comments are gold!
```

Output:

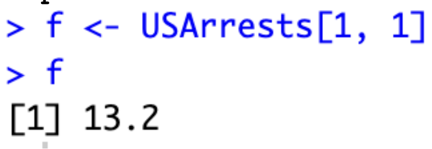


```
# A comment that R will not run. It's here for your benefit. Use LOTS of
# comments! If you need to run your code months from now, or if you are
# doing a HW assignment, comments are gold!
```

Part f) R uses a set of coordinates - first the row number, then the column number. So, in R, the top left entry in the spreadsheet is referenced as [1,1]. In R, you always use square brackets to indicate coordinates. So, for the data frame named `USArrests`, the top left entry in the spreadsheet can be seen by entering:

```
f <- USArrests[1, 1]
f
```

Output:



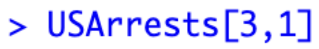
```
> f <- USArrests[1, 1]
> f
[1] 13.2
```

Personal Notes:

The value of 13.2 is the number of Murders in Alabama, which is the first entry in the first column

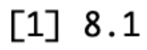
Part g) Enter the command to find the value of the first column, 3rd row.

Command:



```
> USArrests[3,1]
```

Output:

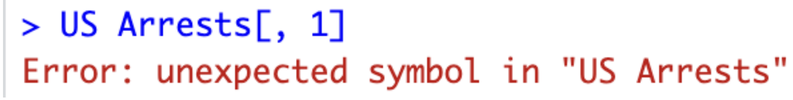


```
[1] 8.1
```

Part h)

```
USArrests[, 1]
```

Output:

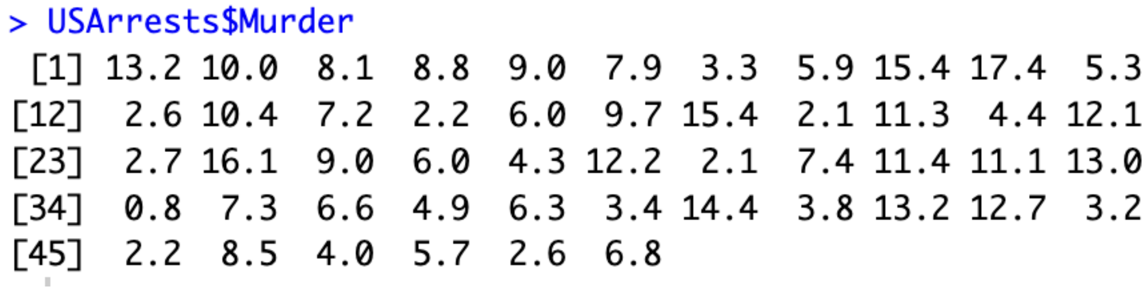


```
> US Arrests[, 1]
Error: unexpected symbol in "US Arrests"
```

Part i) You tell R to give you all the column by excluding the row number, or by putting a **$** sign after the dataframe name and typing the variable name.

```
USArrests$Murder
```

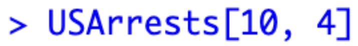
Output:



```
> USArrests$Murder
 [1] 13.2 10.0  8.1  8.8  9.0  7.9  3.3  5.9 15.4 17.4  5.3
[12]  2.6 10.4  7.2  2.2  6.0  9.7 15.4  2.1 11.3  4.4 12.1
[23]  2.7 16.1  9.0  6.0  4.3 12.2  2.1  7.4 11.4 11.1 13.0
[34]  0.8  7.3  6.6  4.9  6.3  3.4 14.4  3.8 13.2 12.7  3.2
[45]  2.2  8.5  4.0  5.7  2.6  6.8
```

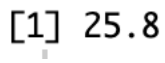
Part j) Get the data element from USArrests in Row 10, Column 4

Command:



```
> USArrests[10, 4]
```

Output:

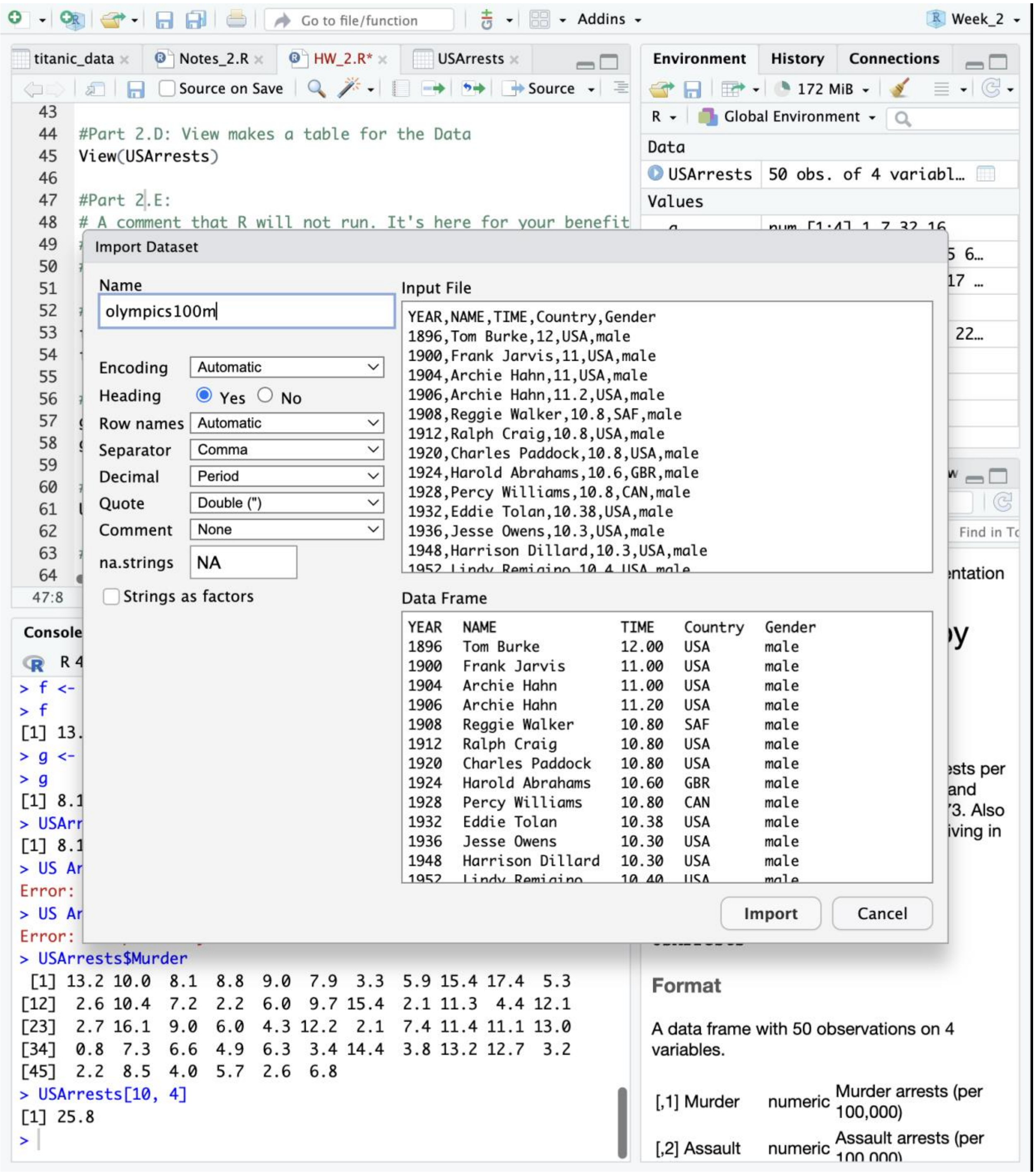


```
[1] 25.8
```

Question#3 (10 points)

You most often save your data in Excel as a CSV file, and therefore need to “transfer” your data from Excel to R. You will read the CSV file into a R dataset.

Download the file **olympics100m.csv** file from Scholar, and place it on your computer’s desktop (or you may choose any location). Using the “Import Dataset” feature in R Studio, import the data and take a screenshot to answer this question.



```
#Part 2.D: View makes a table for the Data
View(USArrests)

#Part 2.E:
# A comment that R will not run. It's here for your benefit
```

Import Dataset

Name: olympics100m

Encoding: Automatic

Heading: ● Yes ○ No

Row names: Automatic

Separator: Comma

Decimal: Period

Quote: Double (")

Comment: None

na.strings: NA

□ Strings as factors

Input File

```
YEAR,NAME,TIME,Country,Gender
1896,Tom Burke,12,USA,male
1900,Frank Jarvis,11,USA,male
1904,Archie Hahn,11,USA,male
1906,Archie Hahn,11.2,USA,male
1908,Reggie Walker,10.8,SAF,male
1912,Ralph Craig,10.8,USA,male
1920,Charles Paddock,10.8,USA,male
1924,Harold Abrahams,10.6,GBR,male
1928,Percy Williams,10.8,CAN,male
1932,Eddie Tolan,10.38,USA,male
1936,Jesse Owens,10.3,USA,male
1948,Harrison Dillard,10.3,USA,male
1952,Lindy Remigino,10.4,USA,male
```

Data Frame

```
YEAR  NAME              TIME   Country  Gender
1896  Tom Burke         12.00  USA      male
1900  Frank Jarvis      11.00  USA      male
1904  Archie Hahn       11.00  USA      male
1906  Archie Hahn       11.20  USA      male
1908  Reggie Walker     10.80  SAF      male
1912  Ralph Craig       10.80  USA      male
1920  Charles Paddock   10.80  USA      male
1924  Harold Abrahams   10.60  GBR      male
1928  Percy Williams    10.80  CAN      male
1932  Eddie Tolan       10.38  USA      male
1936  Jesse Owens       10.30  USA      male
1948  Harrison Dillard  10.30  USA      male
1952  Lindy Remigino    10.40  USA      male
```

Data

USArrests 50 obs. of 4 variabl…

Values

```
> f <-
> f
[1] 13.
> g <-
> g
[1] 8.1
> USArr
[1] 8.1
> US Ar
Error:
> US Ar
Error:
> USArrests$Murder
 [1] 13.2 10.0  8.1  8.8  9.0  7.9  3.3  5.9 15.4 17.4  5.3
[12]  2.6 10.4  7.2  2.2  6.0  9.7 15.4  2.1 11.3  4.4 12.1
[23]  2.7 16.1  9.0  6.0  4.3 12.2  2.1  7.4 11.4 11.1 13.0
[34]  0.8  7.3  6.6  4.9  6.3  3.4 14.4  3.8 13.2 12.7  3.2
[45]  2.2  8.5  4.0  5.7  2.6  6.8
> USArrests[10, 4]
[1] 25.8
>
```

Format

A data frame with 50 observations on 4 variables.

[,1] Murder numeric Murder arrests (per 100,000)

[,2] Assault numeric Assault arrests (per 100,000)

Note: The data import features can be accessed from the environment window. You can use either "From Text (readr)" or "From Text (base)" to enable importing the CSV file.

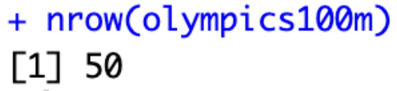
Question#4 (18 points)

After importing data in R you can check and see it with some common functions. Then try the following commands (one at a time) and provide the output

Part a) This function returns the total number of rows in your dataframe.

nrow(olympics100m)

Output:

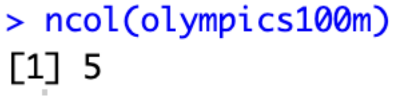


```
+ nrow(olympics100m)
[1] 50
```

Part b) Returns the total number of columns in your dataframe.

ncol(olympics100m)

Output:

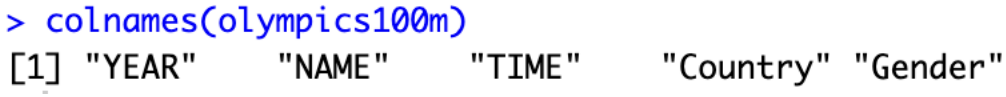


```
> ncol(olympics100m)
[1] 5
```

Part c) This function returns the column headers or column names.

colnames(olympics100m)

Output:

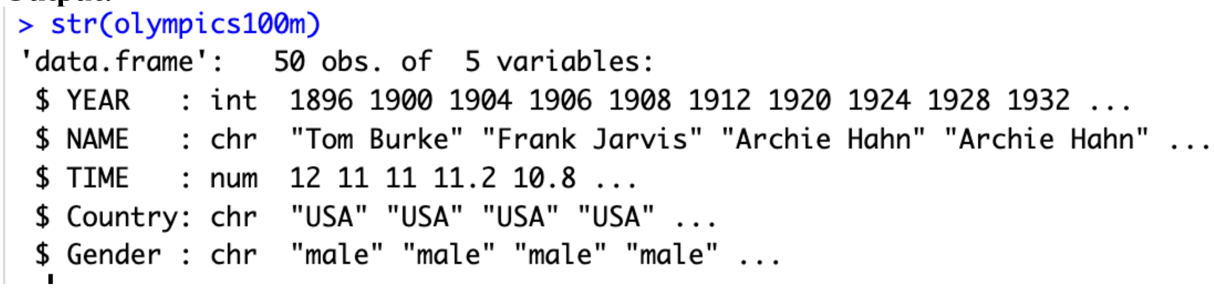


```
> colnames(olympics100m)
[1] "YEAR"    "NAME"    "TIME"    "Country" "Gender"
```

Part d) Returns the structure of your dataframe. Column names with data types and factors.

str(olympics100m)

Output:



```
> str(olympics100m)
'data.frame':	50 obs. of  5 variables:
 $ YEAR   : int  1896 1900 1904 1906 1908 1912 1920 1924 1928 1932 ...
 $ NAME   : chr  "Tom Burke" "Frank Jarvis" "Archie Hahn" "Archie Hahn" ...
 $ TIME   : num  12 11 11 11.2 10.8 ...
 $ Country: chr  "USA" "USA" "USA" "USA" ...
 $ Gender : chr  "male" "male" "male" "male" ...
```

BUSN 231 - Applied Business Statistics

Homework#3

Due Date: September 10 at 11:59pm

The homework submitted should represent your INDIVIDUAL work. Any cases of plagiarism will be treated according to the Honor Code and the CNU procedures. You can ask any questions to me.

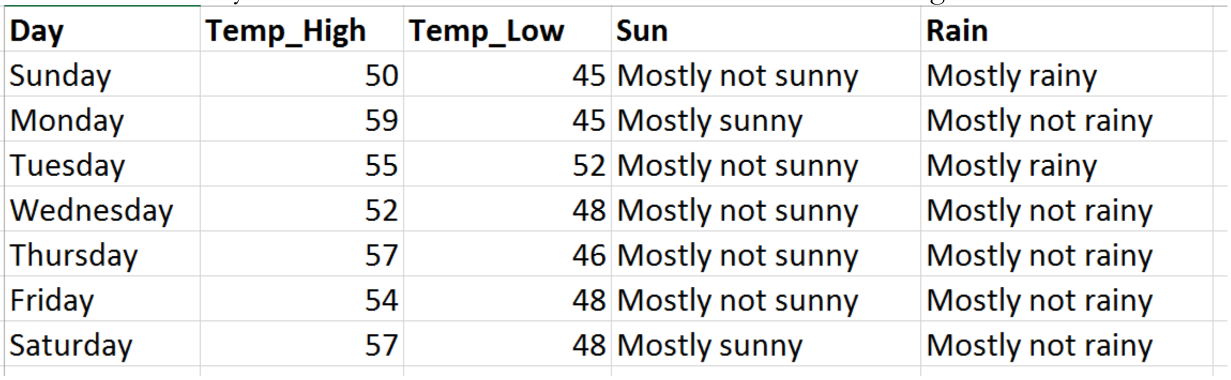
Question#1 (15 points) - CSV File

Suppose you are a meteorologist who wants to track weather condition in the city of New York. Every day you want to record the following data:

- **Day** of the week (0=Sunday, 1=Monday, …, 6=Saturday)
- **Temp_High** – the highest temperature during that day

- **Temp_Low** – the lowest temperature during that day
- **Sun** – how sunny was the weather (0=Mostly not sunny, 1=Mostly sunny)
- **Rain** – how rainy was the weather (0=Mostly not rainy, 1=Mostly rainy)

In the table below you have recorded the weather conditions in NYC during the last week.



Day	Temp_High	Temp_Low	Sun	Rain
Sunday	50	45	Mostly not sunny	Mostly rainy
Monday	59	45	Mostly sunny	Mostly not rainy
Tuesday	55	52	Mostly not sunny	Mostly rainy
Wednesday	52	48	Mostly not sunny	Mostly not rainy
Thursday	57	46	Mostly not sunny	Mostly not rainy
Friday	54	48	Mostly not sunny	Mostly not rainy
Saturday	57	48	Mostly sunny	Mostly not rainy

Your assignment is to **create a csv-file** using the steps below, containing data pictured above.

☑ Step#1: Open any simple text editor (such as Notepad for Windows, or TextEdit for Mac).

☑ Step#2: Create columns, named exactly as shown in the picture above.

☑ Step#3: Enter the data shown above into your data file. Save the data as csv-file and upload to Scholar to answer this question.

Note: You may create a csv-file using MS Excel instead if a text editor.

The dataset has been saved to the assignment in Scholar

Question#2 (35 points) – Frequency Distribution

The Red Lobster restaurant chain conducts regular surveys of its customers to monitor its performance. One of the questions asks customers to rate the overall quality of their last visit. The responses are recorded in the file "**red_lobster.csv**" as 1 = Poor, 2 = Fair, 3 = Good, 4 = Very Good, and 5 = Excellent.

a. ☑ What is the type of data (Discrete, Continuous, Nominal or Ordinal) used in the study?

The following data is Ordinal, as the survey rated/ranked the quality of the visit, meaning that the rankings can be ordered.

b. ☑ Read the data from "**red_lobster.csv**". As we did in lecture, create a tabular summary of the data using *table* and *prop.table* in R. Based on your output, please fill in the table below:

Class	Class Description	Count	Percentage (%)
1	Poor	**17**	**7.69**
2	Fair	**43**	**19.46**

3	Good	**69**	**31.22**
4	Very Good	**68**	**30.77**
5	Excellent	**24**	**10.86**

c. ✅Copy-paste your R code in the space below.

```
str(red_lobster)

lobster_data <- table(red_lobster)

lobster_data

round(prop.table(lobster_data)*100,2)
```

d. ✅Percentage of every response, computed in part (b) above, can be viewed as a probability that a randomly selected customer will provide this response when asked to rate the quality of the last visit to "Red Lobster". What is the probability a randomly selected customer will rate restaurant service as "Good" or higher? Show your computations.

31.22% + 19.46% + 7.69% = 58.37%

e. ✅What is the probability a randomly selected customer will rate restaurant service anything except "Poor" or "Excellent"? Show your computations.

19.46% + 31.22% + 30.77% = 81.45%

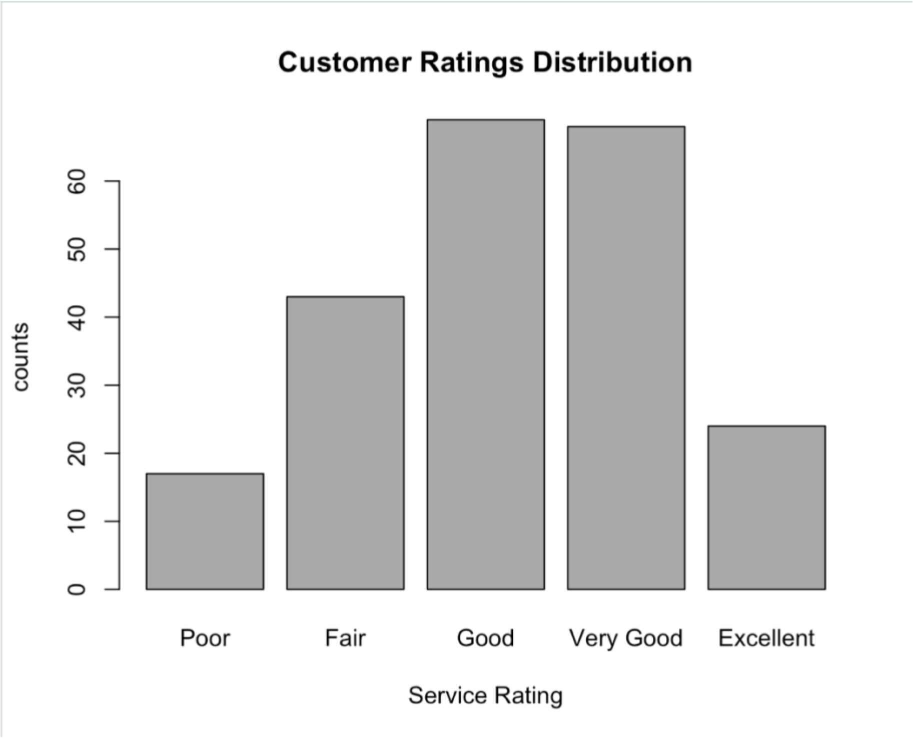
✅**Question#3 (25 points) – Bar Plot**

Construct a bar chart of the Red Lobster data using the *barplot* function in R.

- set main title of the chart to be "Customer Ratings Distribution"
- x-axis label to be "Service Rating"
- y-axis label to be "counts"
- set the color of the bars to be dark gray
- label each column as "Poor", "Fair", "Good", "Very Good", and "Excellent", instead of 1, 2, 3, 4, 5

Copy-paste your R code and bar chart in the space below.

```
barplot(lobster_data,
    main = "Customer Ratings Distribution",
    xlab = "Service Rating",
    ylab = "counts",
    col = "darkgray",
    names.arg = c("Poor","Fair","Good","Very Good","Excellent"))
```



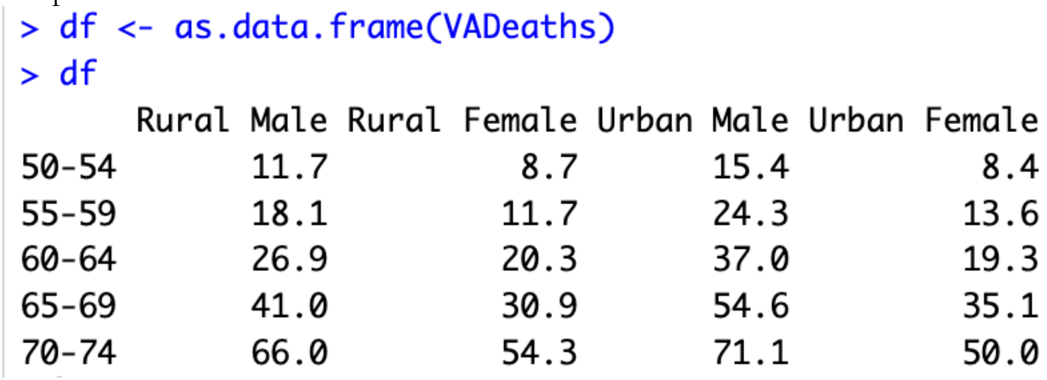
✅**Question#4 (25 points) – Bar Plot** (modified from https://www.theanalysisfactor.com/r-11-bar-charts/)
Here, we'll practice how to create bar plots in R. The function barplot() can be used to create a bar plot with vertical or horizontal bars. Please follow the steps below by typing the given R codes in RStudio. Copy and paste the resulted outputs/figures for each step.

✅Step#1: Import "**VADeaths.csv**" data into R using the read.csv function. Note that you can use the import data feature in RStudio.

✅Step#2: Convert the imported data to a data frame (Note: this is a special step that we need to do for only this data set)

```
df <-as.data.frame(VADeaths)
df
```

Output:

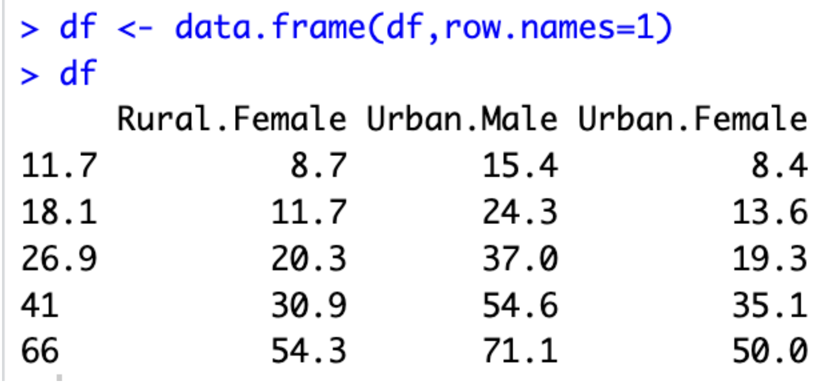


```
> df <- as.data.frame(VADeaths)
> df
      Rural Male Rural Female Urban Male Urban Female
50-54       11.7          8.7       15.4          8.4
55-59       18.1         11.7       24.3         13.6
60-64       26.9         20.3       37.0         19.3
65-69       41.0         30.9       54.6         35.1
70-74       66.0         54.3       71.1         50.0
```

✓ Step#3: Since the first column include the categories (row names), we will convert them from column to row names. (Note: this is a special step that we need to do for only this data set))

df <-data.frame(df, row.names=1)
df

Output:

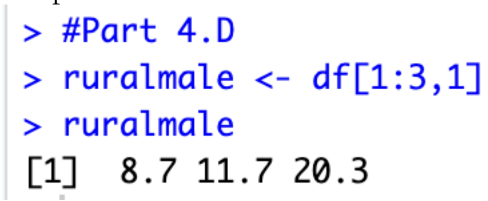


```
> df <- data.frame(df,row.names=1)
> df
     Rural.Female Urban.Male Urban.Female
11.7          8.7       15.4          8.4
18.1         11.7       24.3         13.6
26.9         20.3       37.0         19.3
41           30.9       54.6         35.1
66           54.3       71.1         50.0
```

✓ Step#4: Consider only the subset of the data by selecting "Rural Male" between 50 and 64.

ruralmale <- df[1:3, 1]
ruralmale

Output:

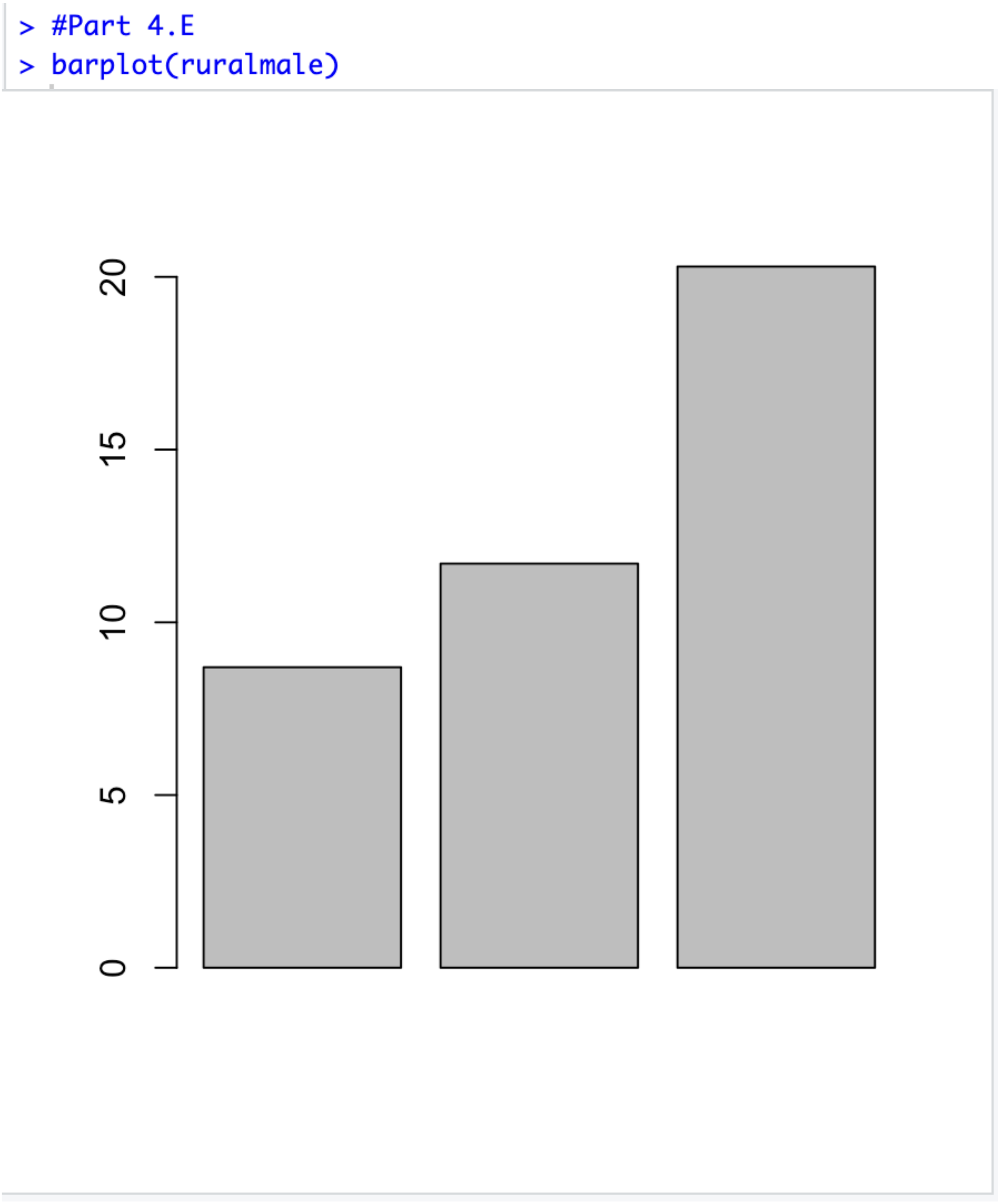


```
> #Part 4.D
> ruralmale <- df[1:3,1]
> ruralmale
[1]  8.7 11.7 20.3
```

✓ Step#5: Bar plot of one variable

barplot(ruralmale)

Output:



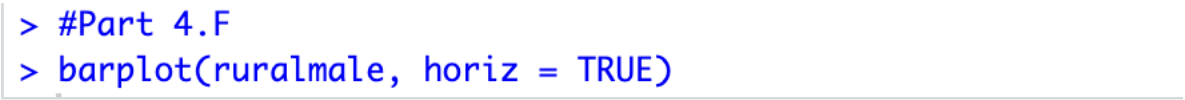
```
> #Part 4.E
> barplot(ruralmale)
```

***This is from the Plots window**

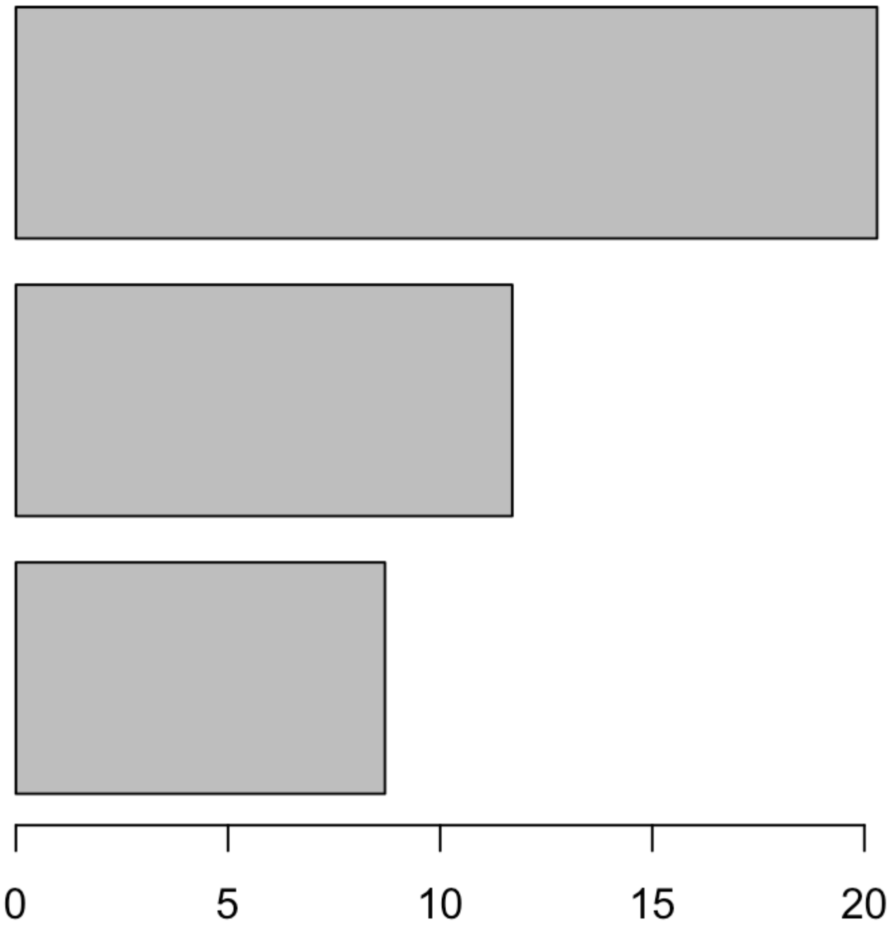
✓ Step#6: Horizontal bar plot

barplot(ruralmale, horiz = TRUE)

Output:



```
> #Part 4.F
> barplot(ruralmale, horiz = TRUE)
```

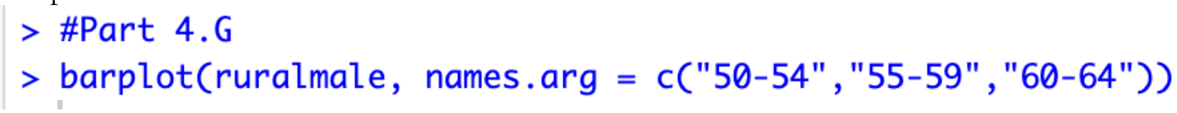


***This is from the Plots window**

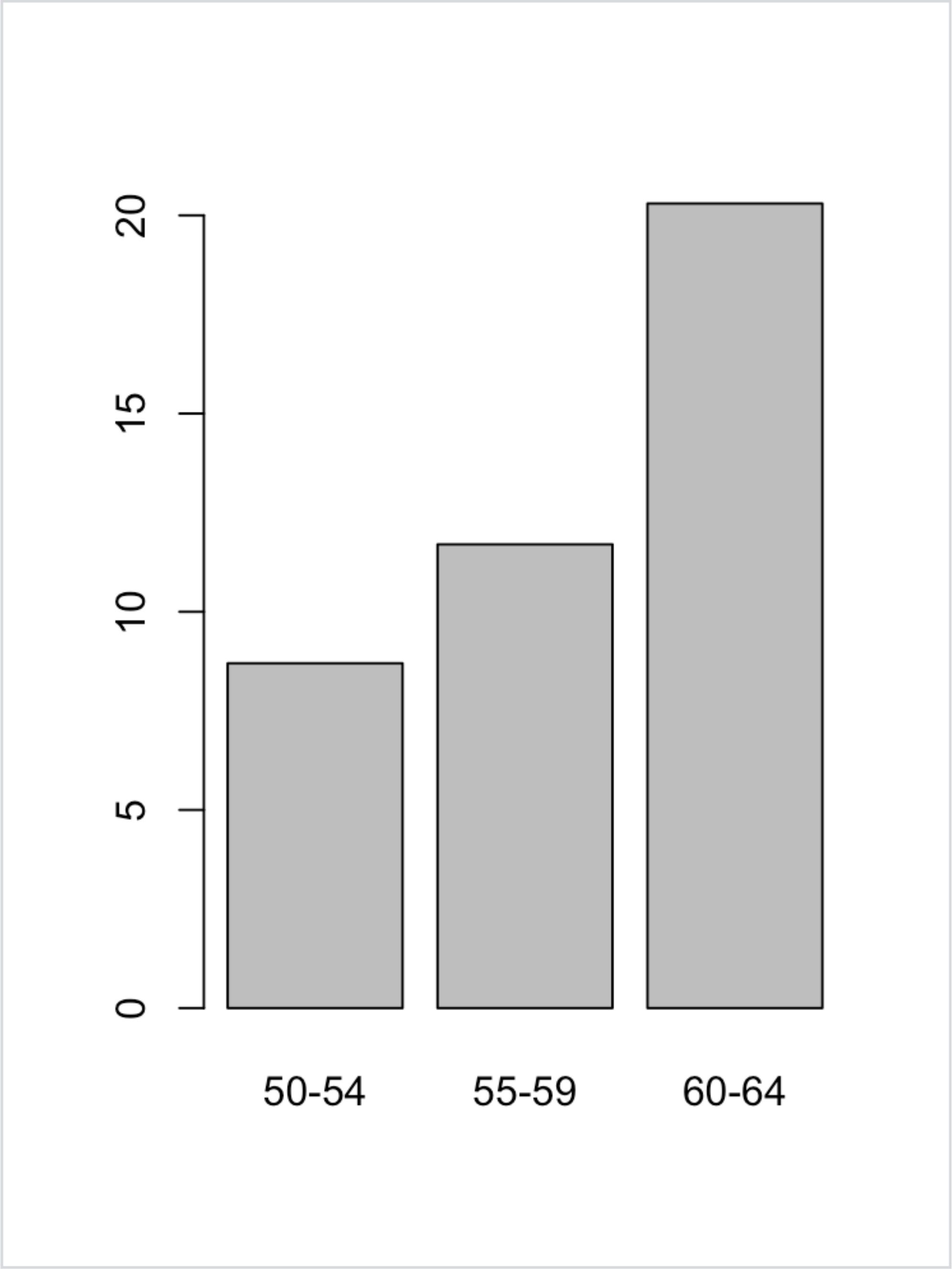
☑ Step#7: Changing group names

barplot(ruralmale, names.arg = c("50-54", "55-59", " 60-64"))

Output:



```
> #Part 4.G
> barplot(ruralmale, names.arg = c("50-54","55-59","60-64"))
```

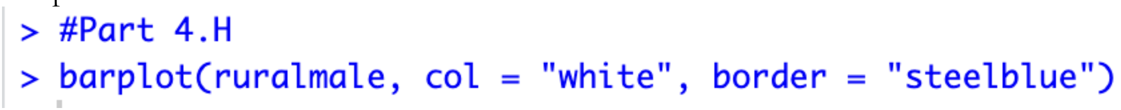


***This is from the Plots window**

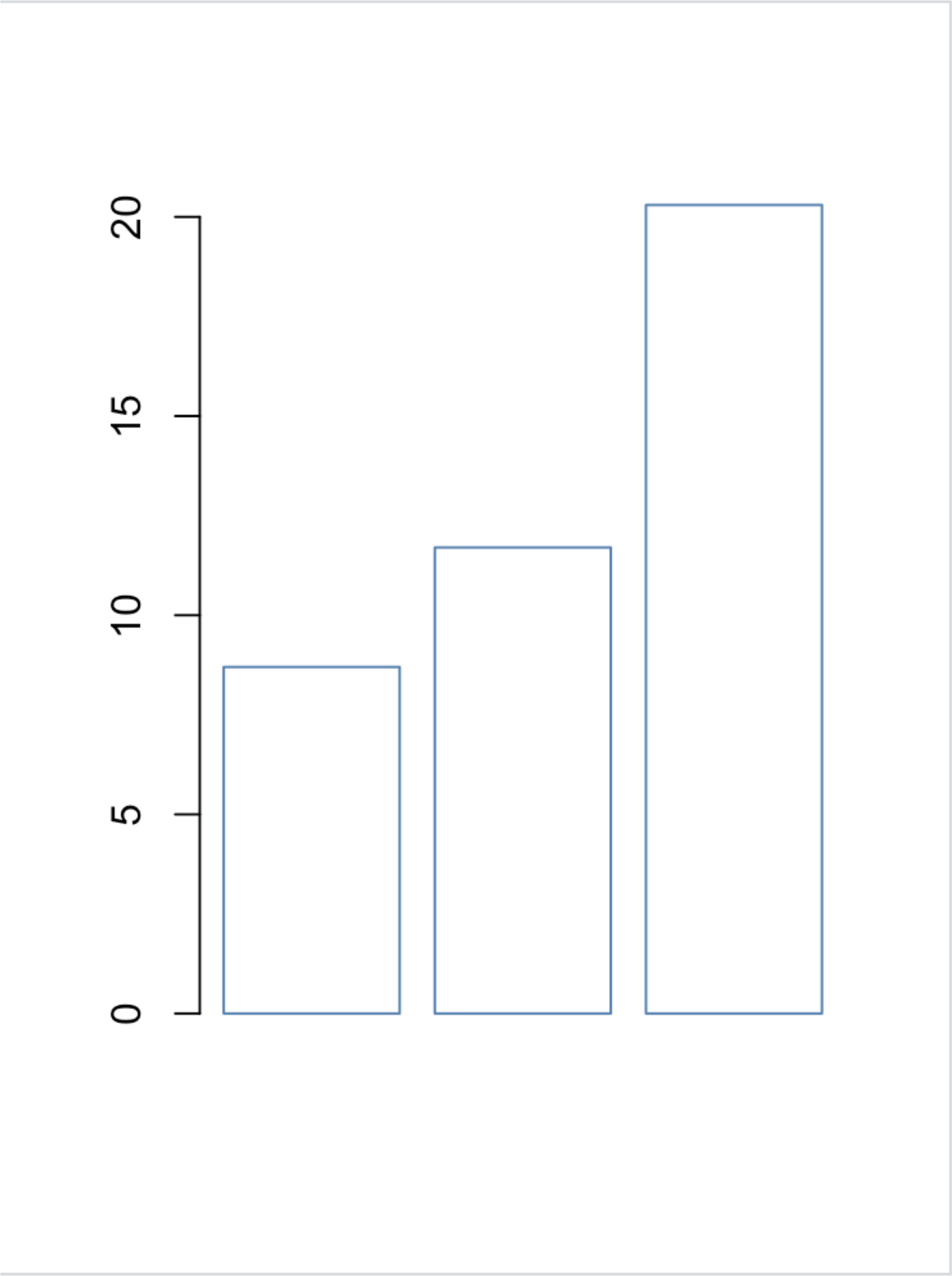
✓ Step#8: Changing border and fill color using one single color

barplot(ruralmale, col = "white", border = "steelblue")

Output:



```
> #Part 4.H
> barplot(ruralmale, col = "white", border = "steelblue")
```

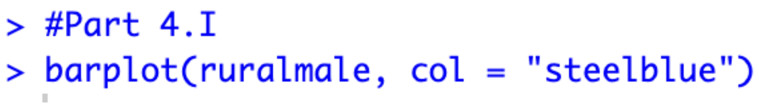


*This is from the Plots window

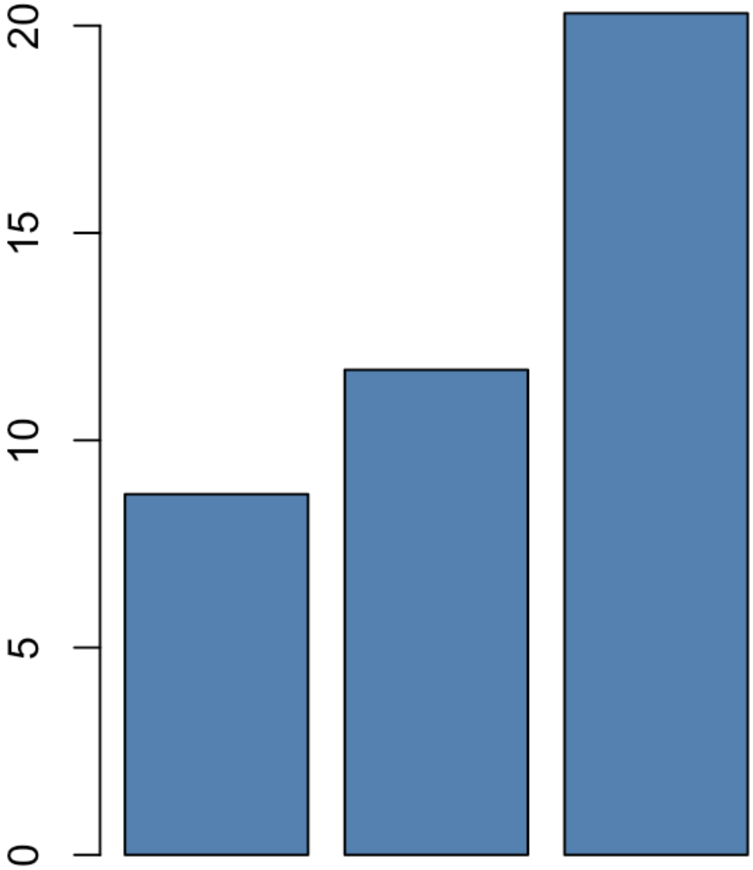
✅ Step#9: Changing fill color using one single color

```
barplot(ruralmale, col = "steelblue")
```

Output:



```
> #Part 4.I
> barplot(ruralmale, col = "steelblue")
```

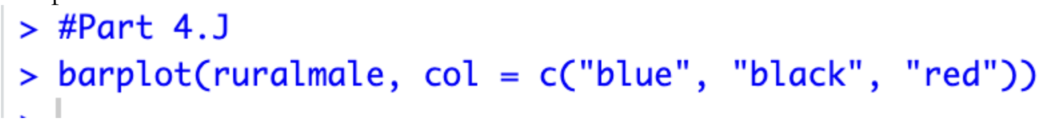


*This is from the Plots window

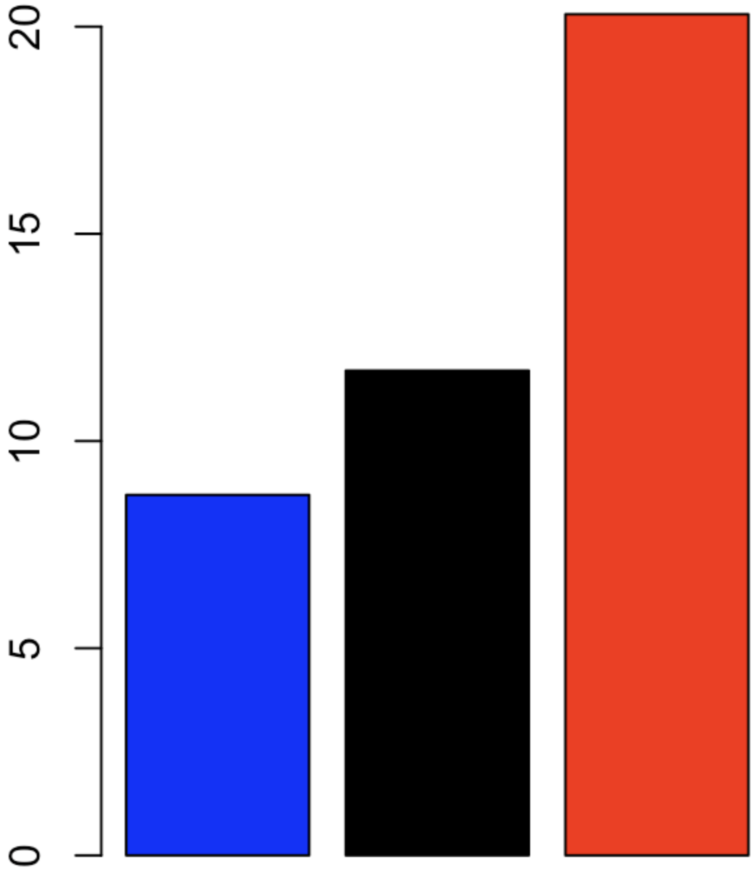
☑ Step#10: Changing fill color using multiple color

barplot(ruralmale, col = c("blue", "black", "red"))

Output:



```
> #Part 4.J
> barplot(ruralmale, col = c("blue", "black", "red"))
```

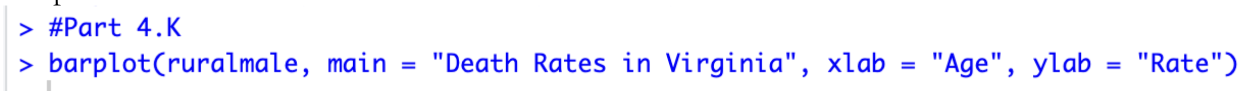


***This is from the Plots window**

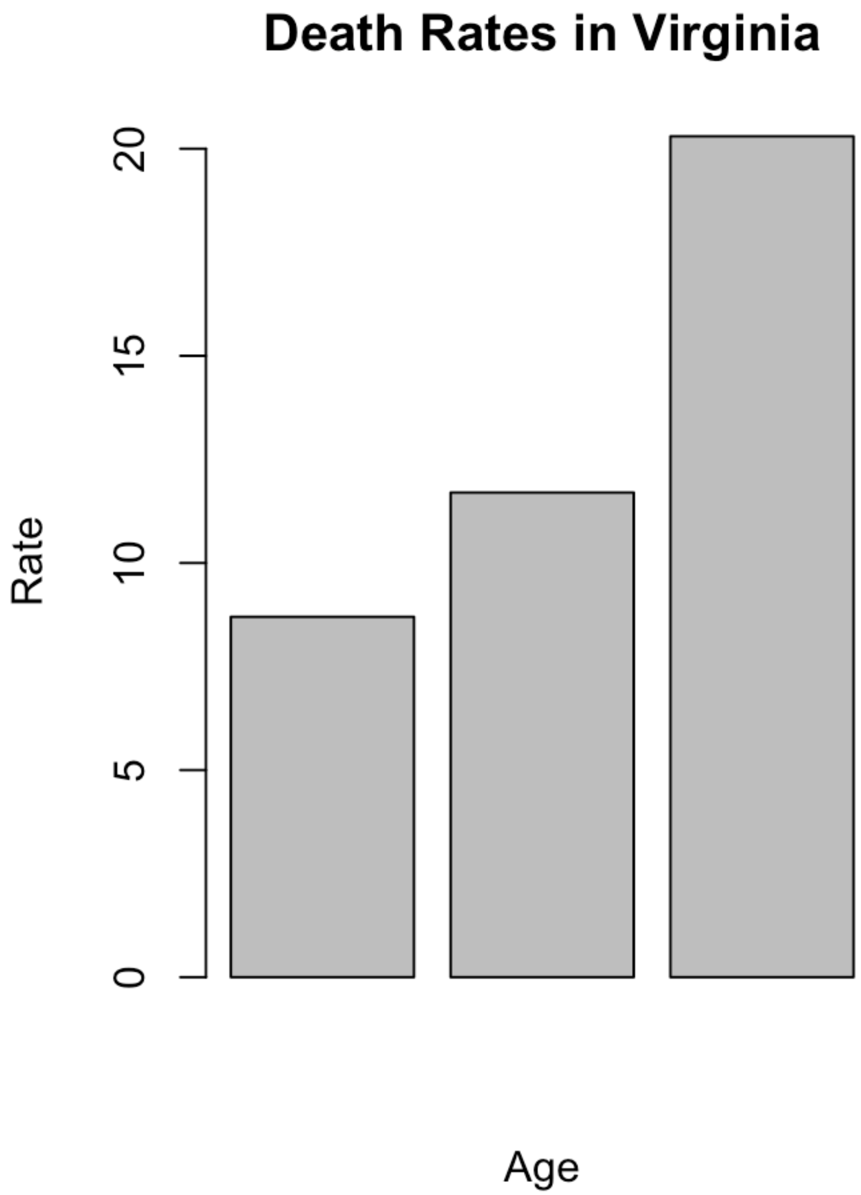
✅ Step#11: Changing main title and axis labels

barplot(ruralmale, main = "Death Rates in Virginia", xlab = "Age", ylab = "Rate")

Output:



```
> #Part 4.K
> barplot(ruralmale, main = "Death Rates in Virginia", xlab = "Age", ylab = "Rate")
```

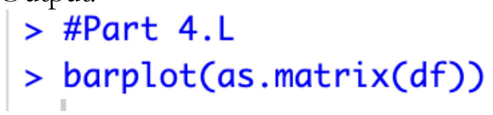


***This is from the Plots window**

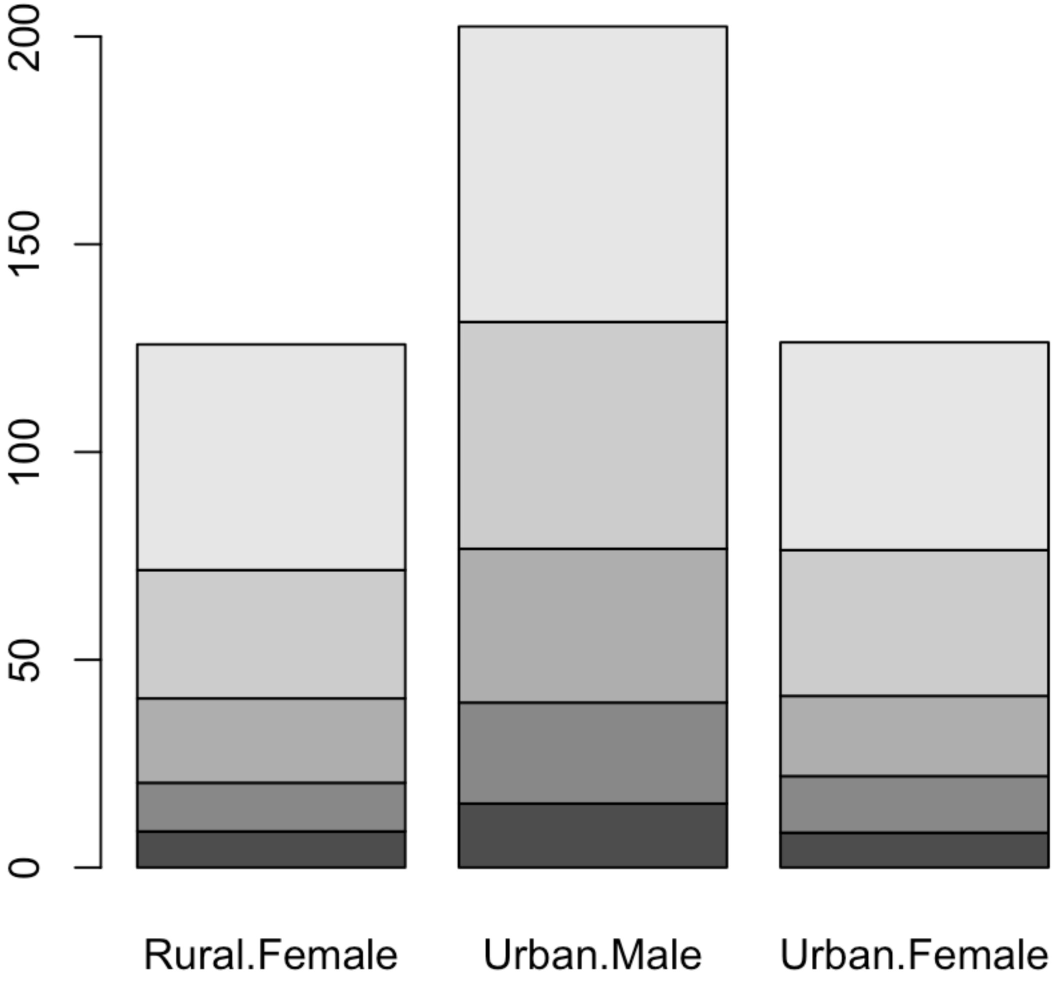
☑ Step#12: Creating stacked barplots in R

barplot(as.matrix(df))

Output:



```
> #Part 4.L
> barplot(as.matrix(df))
```

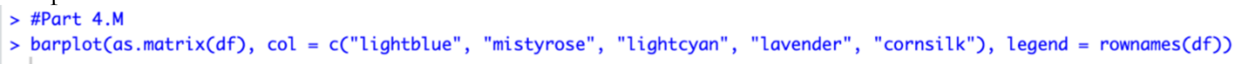


***This is from the Plots window**

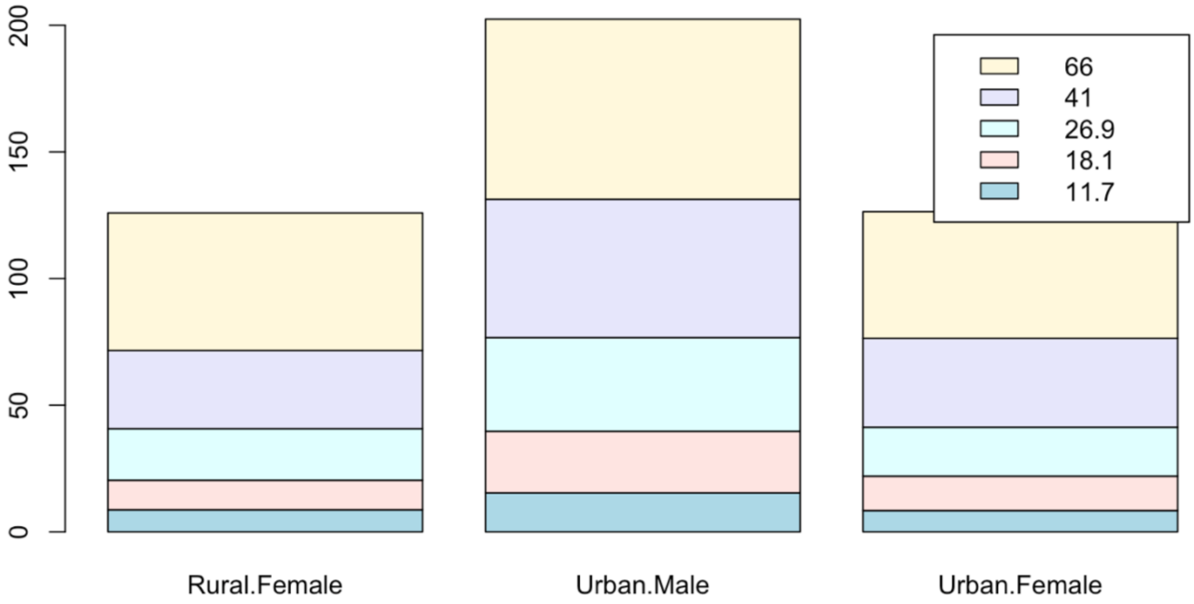
☑ Step#13: All data set using multiple color with legend

barplot(as.matrix(df), col = c("lightblue", "mistyrose", "lightcyan", "lavender", "cornsilk"), legend = rownames(df))

Output:



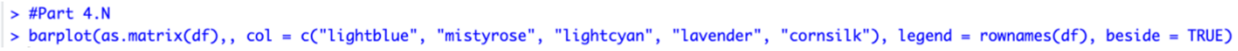
```
> #Part 4.M
> barplot(as.matrix(df), col = c("lightblue", "mistyrose", "lightcyan", "lavender", "cornsilk"), legend = rownames(df))
```



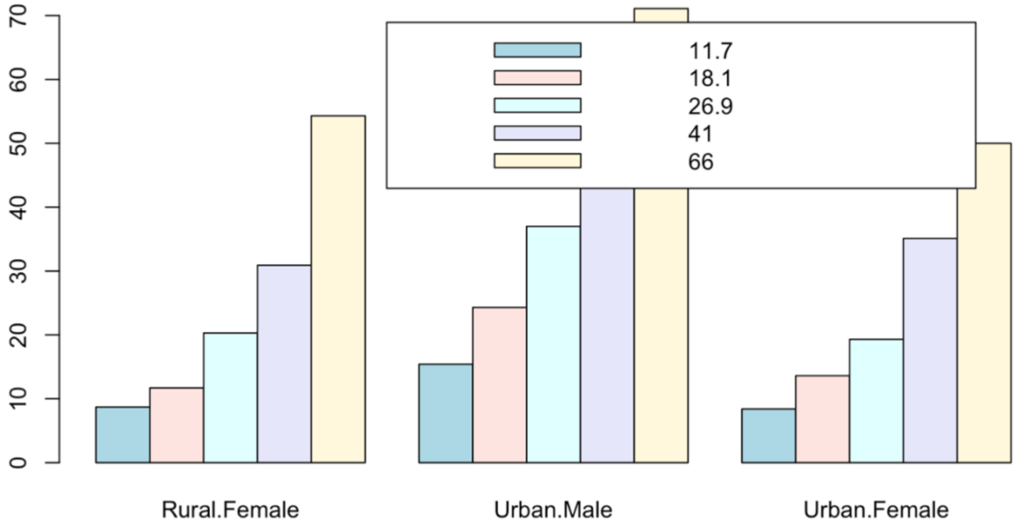
✅ Step#14: Creating grouped bar plots

barplot(as.matrix(df),, col = c("lightblue", "mistyrose", "lightcyan", "lavender", "cornsilk"), legend = rownames(df), beside = TRUE)

Output:



```
> #Part 4.N
> barplot(as.matrix(df),, col = c("lightblue", "mistyrose", "lightcyan", "lavender", "cornsilk"), legend = rownames(df), beside = TRUE)
```



BUSN 231: Applied Business Statistics
Homework#4
Due Date: September 17, 2023 at 11:59pm

The homework submitted should represent your INDIVIDUAL work. Any cases of plagiarism will be treated according to the Honor Code and the CNU procedures. You can ask any questions to me.

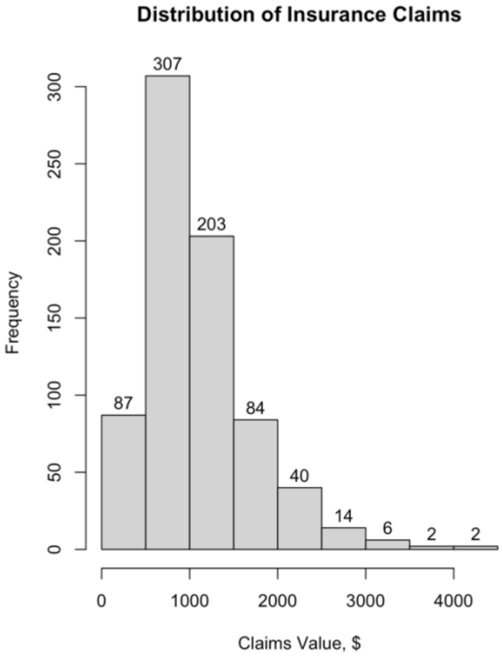
Question#1 (50 points)
An Insurance claims adjuster would like to investigate how the dollar values of auto accidents claims are distributed. He collects a random sample of insurance claims and records the dollar value of the final insurance payment. The results are recorded in the file "**insurance.csv**".

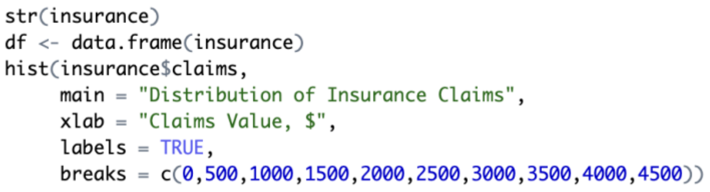
a. What type of data (Discrete, Continuous, Nominal or Ordinal) is collected in the study?

The following data is Discrete, as it is a numerical value that you can count, not measure on a scale.

b. Read the data from the csv-file in R. Construct the histogram of the data using the **hist** function. Please make sure to include the followings in the histogram.
   - set main title to "Distribution of Insurance Claims"
   - label for x-axis to be "Claims Value, $"
   - show frequencies/counts on the top of bars on the histogram
   - use $0 as the lower boundary for the first class, and a class width of $500

   Copy and paste the histogram and R Code below. (Note: you can use histogram-R.R (available in the Week 3 folder) as an example to see how to create an histogram)





```
str(insurance)
df <- data.frame(insurance)
hist(insurance$claims,
     main = "Distribution of Insurance Claims",
     xlab = "Claims Value, $",
     labels = TRUE,
     breaks = c(0,500,1000,1500,2000,2500,3000,3500,4000,4500))
```

c. Fill in the values in the table below according to the histogram.

Between		
>	<=	Frequency
0	500	**87**
500	1,000	**307**
1,000	1,500	**203**
1,500	2,000	**84**
2,000	2,500	**40**
2,500	3,000	**14**
3,000	3,500	**6**
3,500	4,000	**2**
4,000	4,500	**2**
	Total	**745**

d. Characterize the data histogram with respect to (a) **modality**, and (b) **symmetry**. (Note: please take a look at Lecture#8)

With respect to modality, the histogram has one modal class, with insurance claims between $500 - $1,000 having a frequency of 307 claims, so most of the insurance claims that were made were between $500 - $1,000.

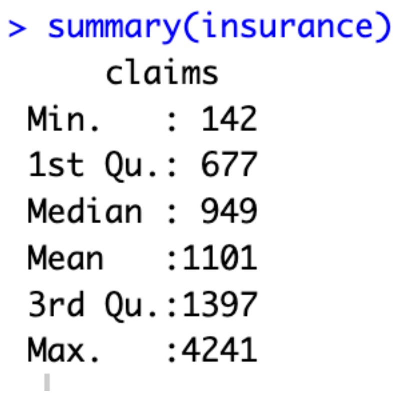
With respect to symmetry, the histogram is asymmetrical. Because the histogram is skewed, the mean, median, and mode are most likely not close to eachother.

e. If the distribution of insurance claims is skewed (left or right) exists, give a brief explanation (in plain English) what do you think is the business reason the data is skewed that way.

The distribution of the graph is right skewed as the data within the histogram visually represents with most of it on the left side. This means that the majority of Insurance claims are less than the mean of the data, and this can be determined by finding the mode of the data. From a business perspective, this checks out. The majority of auto accident claims are not going to be upwards of $4500 (unless you really wreck the car). Theses claims will most likely be in the range of $500 to $1500 depending on the accident, which is enough for the customer to use to fix the car while also not giving out too much money for the sheer amount of claims that have been filed.

Question#2 (40 points)

Analyze numerical data on insurance claims from Question 1 (file "**insurance.csv**"). Specifically, answer the following questions: (Note: You can use Descriptive-R.R (available in the Week#4 folder) as an example)



```
> summary(insurance)
     claims    
 Min.   : 142  
 1st Qu.: 677  
 Median : 949  
 Mean   :1101  
 3rd Qu.:1397  
 Max.   :4241  
```

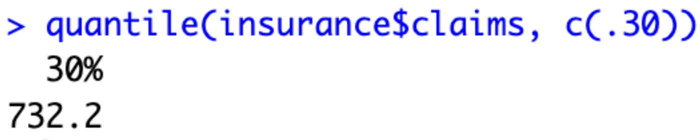
a. What is the mean dollar value of an insurance settlement in the data set sample? (hint: use the *mean* or *summary* function in R)

Mean: 1101

b. What is the range of the middle 50% of the insurance claims in dollar values? Provide the lower and the upper bounds for the said range, as well as the range value. (hint: use the *range* or *summary* function in R)

Range: 3rd Qu – 1st Qu = 1397 – 677 = 720

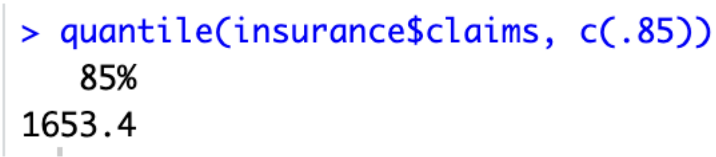
c. **Below** which dollar value are 30% of the least expensive insurance claims? (hint: use the *quantile* function in R)



```
> quantile(insurance$claims, c(.30))
  30% 
732.2 
```

Bottom 30% = $732.2

d. **Above** which dollar value are 15% of the most expensive insurance claims? (hint: use the *quantile* function in R)



```
> quantile(insurance$claims, c(.85))
   85% 
1653.4 
```

Top 15% = $1653.4

Question#3 (10 points)

A coefficient of variation (CV) measures data point dispersion around a mean. Investors use CV to determine risk over return. Their goal is to find that standard deviation shows a lower ratio to mean return, meaning the reward is greater than the risk. The CV formula is calculated by dividing the standard deviation (volatility) of an investment by the expected return (mean).

CV = Standard of deviation / Mean

or
CV = Volatility / Expected return

Investors use this to measure the dispersion of events in order to assess and evaluate risk and volatility of a company or investment. Let's use the scenario of an investor who wants a minimize risk as much as possible. This investor picks among a selection of four investments. He wants to see which offers the best reward relative to risk since he knows that the more risk an investor takes on, the more potential reward. He considers one investment in Amazon, one investment in Apple, one that tracks the S&P 500 index, and a US treasury bond. Let's assume the following:

	Standard Deviation (or Volatility)	**Mean (Expected Return)**	**Coefficient of Variation**
Amazon	20%	10%	2.0
Apple	15%	16%	0.9375
S&P 500 index	10%	10%	1.0
US Treasury bond	1%	5%	0.2

Please fill the table shown in yellow. Which of the four investments should the investor choose? Please explain very briefly why. (No need to do any calculations in R. You can use a basic calculator)
The investor should select the US Treasury Bond as it has the lowest CV. The ratio of Standard Deviation to Mean is only 0.2, which means that the success of the US Treasury bond does not vary a whole lot, thus making it the most secure to invest in.

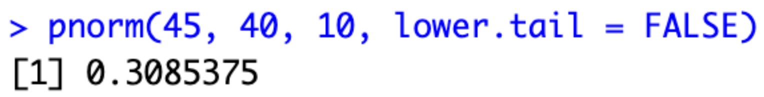
BUSN 231: Applied Business Statistics
Homework#5
Due Date: September 24, 2023 at 11:59pm

The homework submitted should represent your INDIVIDUAL work. Any cases of plagiarism will be treated according to the Honor Code and the CNU procedures. You can ask any questions to me.

Question#1 (45 points)
On average, commuters in Phoenix, Arizona, area require μ= 40.0 minutes to get to work. Assume that for all commuters the times to get to work are normally distributed with a standard deviation of σ= 10 minutes. Joe is an average Phoenix resident he and goes to work every day.

a) What is the probability that on any given day Joe will require over 45 minutes to get to work?

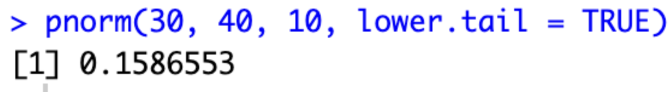
The probability of Joe will require more than 45 minutes to go to work is 30.85%, as we are looking for the area under the normal distribution that is greater than 45.



```
> pnorm(45, 40, 10, lower.tail = FALSE)
[1] 0.3085375
```

b) Joe has just left home from lunch to attend a meeting with the CEO in just 30 minutes. If the CEO routinely fires employees who are tardy for meetings, what is the probability that the Joe will still be employed tomorrow?

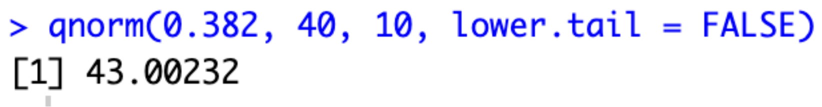
In order for Joe to be employed tomorrow, he must show up on time. Thus, he must arrive in less than 30 minutes, when the mean time to arrive to work is 40 minutes. **Looking at the area under the normal distribution that is less than 30 minutes, the probability of Joe being employed tomorrow is 15.87%.**



```
> pnorm(30, 40, 10, lower.tail = TRUE)
[1] 0.1586553
```

c) Joe is so punctual, that he leaves home for work at exactly same time. His work starts at 8:00 AM sharp. Due to traffic jams during morning rush hours Joe historically was tardy for work on 38.2% of days. What time does Joe leave home for work every morning? Briefly explain your train of thoughts in arriving to the answer.

To find what time joe typically leaves, we need to find when joe is 38,2% late from when he leaves. Considering his average time of travel is 40 minutes with a standard deviation of 10 minutes, we will be looking the value that is in the top 38.2% of the data in the dataset. By running the qnorm function, we can find the value in minutes that it takes for Joe to travel to get to work late 38.2% of the time.



```
> qnorm(0.382, 40, 10, lower.tail = FALSE)
[1] 43.00232
```

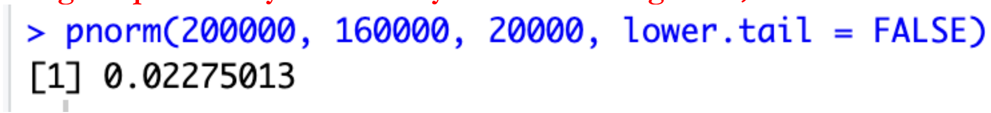
So when Joe is late to work 38.2% of the time he leaves late for work, it takes him approximately 43 minutes to travel to work. Thus, Joe leaves for work at 7:17AM, as although his average time is 40 minutes, he is 38.2% late when he leaves at this time as it would take him 43 minutes to travel to work.

Question#2 (40 points)

Find these probabilities.

a) Weekday sales at a large retailer are normally distributed, with m = +160,000 and s = +20,000. What's the probability that sales tomorrow will exceed $200,000?

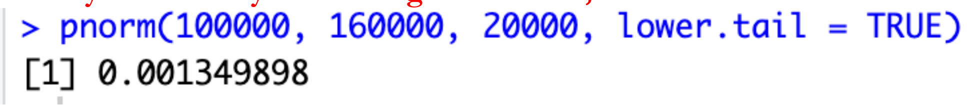
Given we are looking for when sales are above 2 standard deviations away from the mean, this question is not so bad. **With X being +200,000, we are looking for the area above X, meaning the probability of Weekday sales exceeding +200,000 is 2.275%**



```
> pnorm(200000, 160000, 20000, lower.tail = FALSE)
[1] 0.02275013
```

b) What is the probability that sales are less than $100,000 tomorrow?

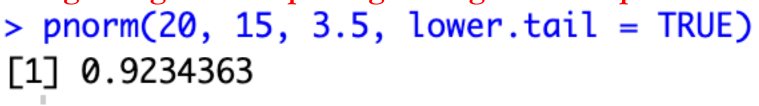
With X being +100,000, we are looking at the area under the curve, meaning the probability of Weekday sales being below +100,000 is .135%



```
> pnorm(100000, 160000, 20000, lower.tail = TRUE)
[1] 0.001349898
```

c) The average weight of packages handled by a delivery service is 15 pounds with a standard deviation of 3.5 pounds. If these weights are normally distributed, what is the probability of finding a package that weighs less than 20 pounds?

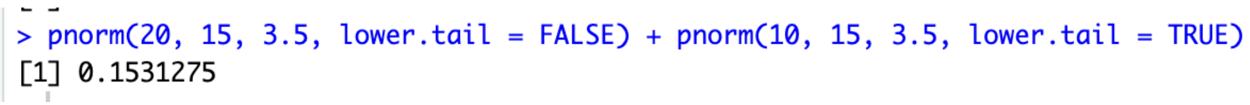
With X being 20, we are looking at the area under the curve, meaning the probability of the average weight of the package being below 20 pounds is 92.344%



```
> pnorm(20, 15, 3.5, lower.tail = TRUE)
[1] 0.9234363
```

d) What is the probability of finding a package that weighs less than 10 or more than 20 pounds?

Because we are looking for the range that X is below 10 pounds and above 20 pounds, we will add the probability of the package being less than 10 pounds from the probability of the package being greater than 20 pounds. **From this, we can determine that the probability of the package being outside 10-20 pounds is 15.313%**



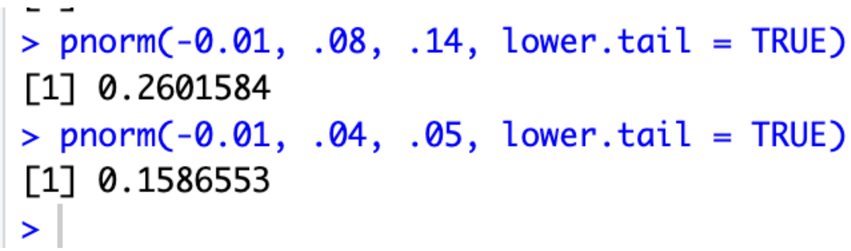
```
> pnorm(20, 15, 3.5, lower.tail = FALSE) + pnorm(10, 15, 3.5, lower.tail = TRUE)
[1] 0.1531275
```

Question#3 (15 points)

You are considering the risk-return profile of two mutual funds for investment. The relatively risky fund promises an expected return of 8% with a standard deviation of 14%. The relatively less risky fund promises an expected return and standard deviation of 4% and 5%, respectively. Assume that the returns are normally distributed.

a) Which mutual fund will you pick if your objective is to minimize the probability of earning a negative return?

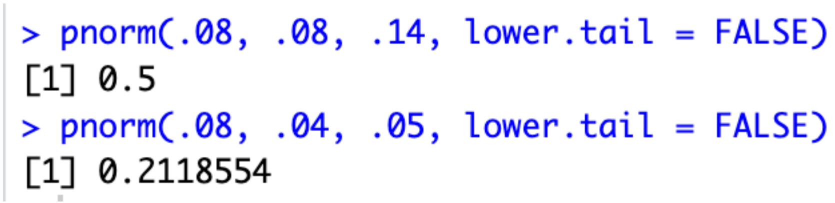
Personally, if I were to invest in a mutual fund, I would select the the find with the expected return of 4% and standard deviation of 5% as the probability for it have a negative return is less than the probability of the 8% with 14% standard deviations



```
> pnorm(-0.01, .08, .14, lower.tail = TRUE)
[1] 0.2601584
> pnorm(-0.01, .04, .05, lower.tail = TRUE)
[1] 0.1586553
> 
```

b) Which mutual fund will you pick if your objective is to maximize the probability of earning a return above 8%?

Given we are looking to maximize probability of earning a return above 8%, I would invest in the 8% expected earnings and 14% standard deviation, as that probability is larger than the probability of the 4%/5% fund.



```
> pnorm(.08, .08, .14, lower.tail = FALSE)
[1] 0.5
> pnorm(.08, .04, .05, lower.tail = FALSE)
[1] 0.2118554
```

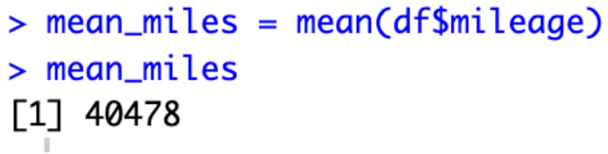
BUSN 231: Applied Business Statistics
Homework#6
Due Date: October 1, 2023 at 11:59pm

The homework submitted should represent your INDIVIDUAL work. Any cases of plagiarism will be treated according to the Honor Code and the CNU procedures. You can ask any questions to me.

Question#1 (50 points)
When you buy new tires for your car, they normally come with a warranty. In order to come up with a warranty, a tire manufacturer must know how long the tires will last on average. In order to do this, the statistician employed by a tire company takes a sample of tires, and measures the mileage when each tire wears off. The results are recorded in the file "**tires.csv**". Answer the following questions:

a. (5 points) What data type is used in the study?
    a. Discrete, as the milage for the tires is counted, not measured per tire. While miles is a unit of measurement, the study is simply recording the amount of miles a tire endures.

b. (5 points) What is the point estimate of the average mileage before a new tire wears off? (Hint: Simply calculate the average mileage)

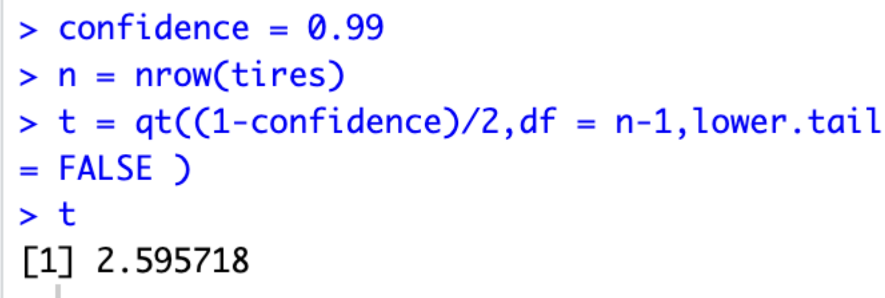


```
> mean_miles = mean(df$mileage)
> mean_miles
[1] 40478
```

The mean time for service is 97.187 seconds

c. (5 point) Based on a point estimate in part (b), does it look like the company can offer customers a warranty of average 40,000 before the tire wears off?
    a. Yes, because the average tire lasts at least 40,478 miles before it cannot be used again under the warranty.

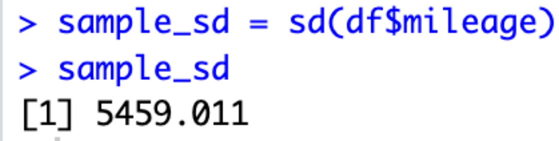
d. (10 points) Suppose we would like to estimate the average wear-off mileage of a tire with a confidence level of 99%. The t-value needed to construct a 99% CI is equal to………. (Hint: calculate the t-value using the **qt** function in R)



```
> confidence = 0.99
> n = nrow(tires)
> t = qt((1-confidence)/2,df = n-1,lower.tail
= FALSE )
> t
[1] 2.595718
```

the T value needed to construct the 99% confidence interval is equal to 2.595718

e. (5 points) What is the value of the sample standard deviation? (Hint: use the **sd** function in R)

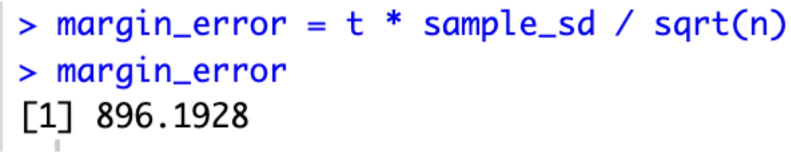


```
> sample_sd = sd(df$mileage)
> sample_sd
[1] 5459.011
```

the sample is equal to 5459.011

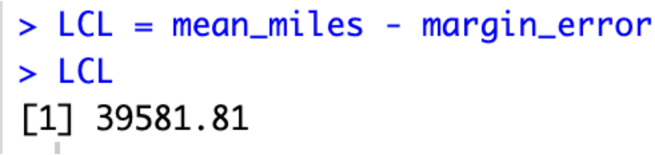
f. (10 points) Write an R code to compute a margin of error for a 99% CI for the average tire wear-off mileage. Calculate and report the LCL and UCL below:

Margin of Error:



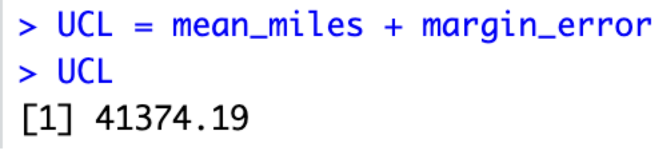
```
> margin_error = t * sample_sd / sqrt(n)
> margin_error
[1] 896.1928
```

LCL =



```
> LCL = mean_miles - margin_error
> LCL
[1] 39581.81
```

UCL =



```
> UCL = mean_miles + margin_error
> UCL
[1] 41374.19
```

g. (10 points) Based on the LCL and UCL in part (f), can the tire manufacturer conclude that the average tire wear-off mileage is 40,000 or more? Briefly explain why or why not.

a. No, because the wear off of the average tire could be greater than 40,000 miles. 99% of the time, the average milage for a tire is between 39,581 miles and 41374 miles, meaning that the tire to wear off below 40,000, which is false advertisement.

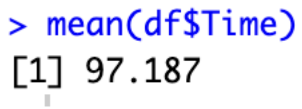
Question#2 (50 points)

The manager of a local fast-food restaurant is interested in improving the service provided to customers who use the restaurant's drive-up window. As a first step in this process, the manager asks an assistant to record the time (in seconds) it takes to serve a large number of customers at the final window in the facility's drive-up system. The file **fast-food.csv** contains a random sample of 200 service times during the busiest hour of the day.

a. (5 points) What data type is used in the study?

   a. This data type is continuous, as the time measured is in a unit of measurement that is measured by the manager of this restaurant. He does not count the time that passes, but rather measures the duration of the interaction with the customers.

b. (5 points) What is the point estimate of the average service time of all customers arriving during the busiest hour of the day at this fast-food operation



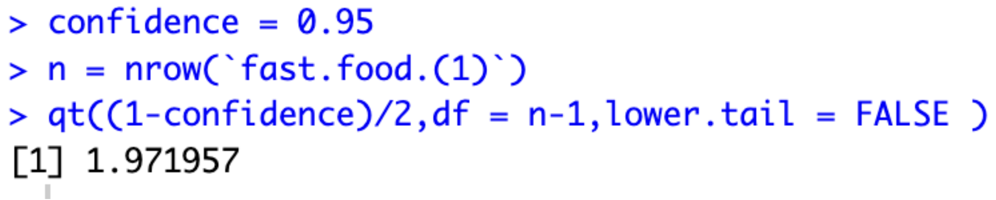
```
> mean(df$Time)
[1] 97.187
```

The mean time for service is 97.187 seconds

c. (5 point) Based on a point estimate in part (b), does it look like the customers wait more than 1 minute on average?

Yes, the customer waits longer than 60 seconds on average, as they typically wait around 1 minute and 37.187 seconds for their order.

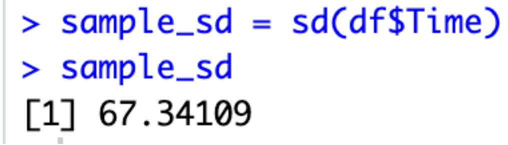
d. (10 points) Suppose we would like to estimate the mean service time with a confidence level of 95%. The t-value needed to construct a 95% CI is equal to………. (Hint: calculate the t-value using the **qt** function in R)



```
> confidence = 0.95
> n = nrow(`fast.food.(1)`)
> qt((1-confidence)/2,df = n-1,lower.tail = FALSE )
[1] 1.971957
```

the T value needed to construct the 95% confidence interval is equal to 1.971957

e. (5 points) What is the value of the sample standard deviation? (Hint: use the **sd** function in R)

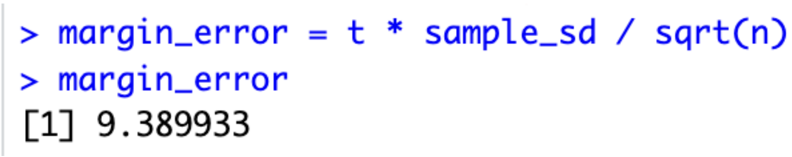


```
> sample_sd = sd(df$Time)
> sample_sd
[1] 67.34109
```

the sample is equal to 1.971957

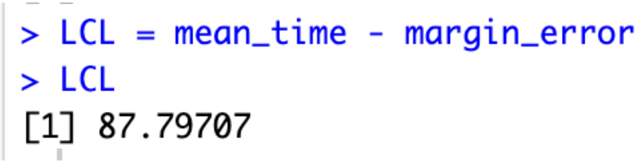
f. (10 points) Write an R code to compute a margin of error for a 95% CI for the mean service time. Calculate and report the LCL and UCL below:

Margin of Error: t = value from 2.D, sample_sd = value from 2.E, n = value from 2.D



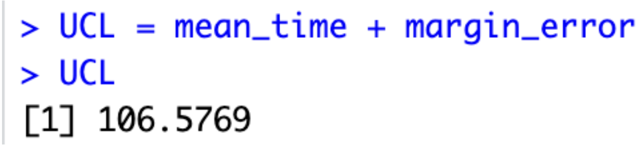
```
> margin_error = t * sample_sd / sqrt(n)
> margin_error
[1] 9.389933
```

LCL = mean_time = value from 2.B



```
> LCL = mean_time - margin_error
> LCL
[1] 87.79707
```

UCL = mean_time = value from 2.B



```
> UCL = mean_time + margin_error
> UCL
[1] 106.5769
```

g. (10 points) If the manager wants to improve service, at least during the busiest time of day, does this confidence interval provide useful information? What useful information does it not provide?

Useful Information:

This information provides a range that allows the manager to estimate that 95% of the time, the average time for his staff to handle high amounts of customers is between 87 and 106 seconds, meaning he is able to accurately account for when productivity is better/worse than usual.

Information Not Provided

Without manually validating the data, we don't know if there are some datapoints that cause values such as the mean or standard deviation to be impacted. There might be outlying data that should be checked. Additionally, this data assumes if there are more customers, the time will take longer. Other tests are able to define if one variable truly impacts another, which is not observed in this study

BUSN 231- Applied Business Statistics
Homework#7
Due Date: October 8, 2023 at 11:59pm

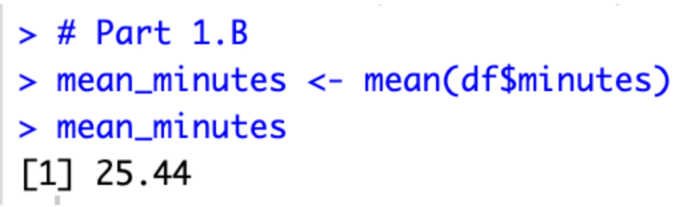
The homework submitted should represent your INDIVIDUAL work. Any cases of plagiarism will be treated according to the Honor Code and the CNU procedures. You can ask any questions to me.

Question#1 (50 points)

The current no-smoking regulations in office buildings require workers who smoke to take breaks and leave buildings in order to satisfy their habits. A study indicates that such workers average 30 minutes per day taking smoking breaks, which negatively impacts the time productivity. To help somewhat reduce the average break time, rooms with power exhausts were installed in the buildings, so that smoking employees do not have to leave building and can return to work faster. To see whether these rooms serve the designed purpose, a random sample of smokers was taken and their total time away from the desk was recorded for one day. The results of the study are recorded in the file "**smoke_break.csv**".

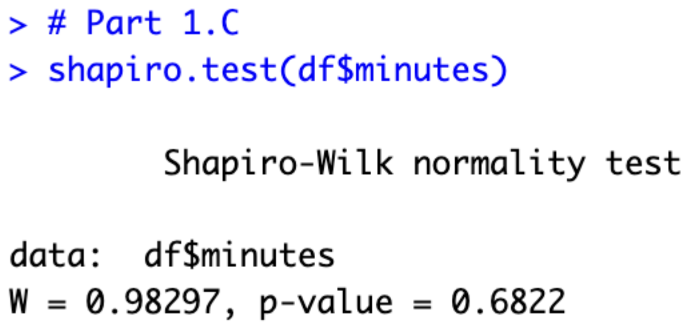
h. (1 point) What data type is used in the study? (numerical or categorical)
   a. Numerical, as it was the length of time to take a smoke break

i. (4 points) What is the point estimate of the average time away from the desk for smokers after the ventilated break rooms were installed?



```
> # Part 1.B
> mean_minutes <- mean(df$minutes)
> mean_minutes
[1] 25.44
```

j. (5 point) Run the Shapiro-Wilk test to see if the data in the sample are normally distributed. In the space below, report the p-value for the test and your conclusion whether the data are normal or not.

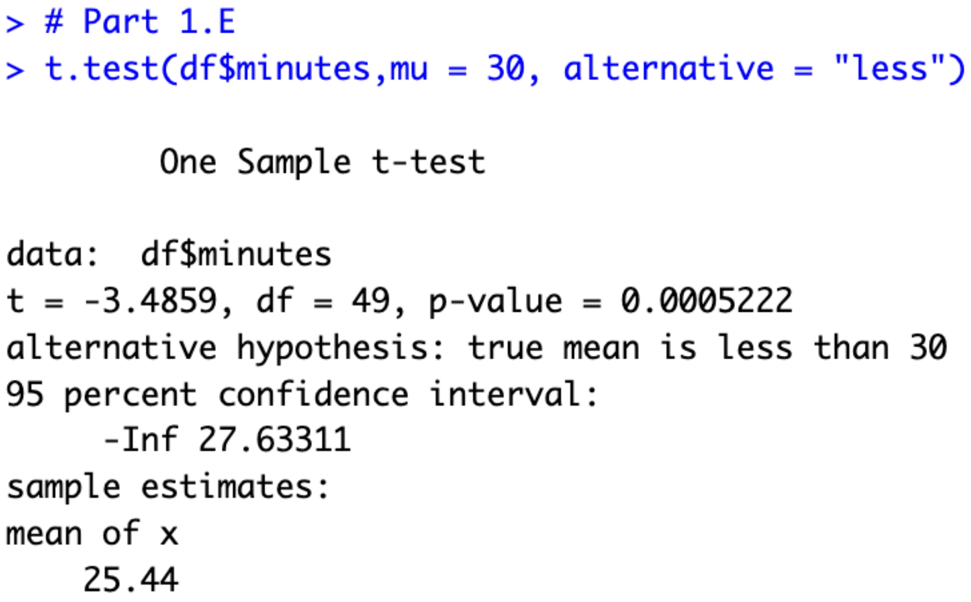


```
> # Part 1.C
> shapiro.test(df$minutes)

        Shapiro-Wilk normality test

data:  df$minutes
W = 0.98297, p-value = 0.6822
```

The data would be considered normal as the returned p-value is greater than the significance level of 0.05

k. (10 points) Formulate the null and alternative hypotheses that need to be tested in order to find out if the break rooms serve the desired purpose.
   a. Null Hypothesis: Mean >= 30
   b. Alternative Hypothesis: Mean < 30

l. (15 points) Compute the p-value for the hypotheses formulated in part (d)



```
> # Part 1.E
> t.test(df$minutes,mu = 30, alternative = "less")

        One Sample t-test

data:  df$minutes
t = -3.4859, df = 49, p-value = 0.0005222
alternative hypothesis: true mean is less than 30
95 percent confidence interval:
     -Inf 27.63311
sample estimates:
mean of x
    25.44
```

m. (15 points) Conclude the hypothesis test and formulate your findings in plain English. At 5% significance level, is there enough evidence in the sample collected to conclude that the dedicated break rooms serve their purpose to reduce the average break time? In order to receive full credit, you must provide written explanation for your answer.

   a. No, because the p-value of 0.0005222 is less than the 0.05 significance level to determine that there is enough evidence to conclude that the mean time is less than 30 minute after the break room renovations. Thus, we will reject the alternative hypothesis of the mean smoke break being less than 30 minutes.

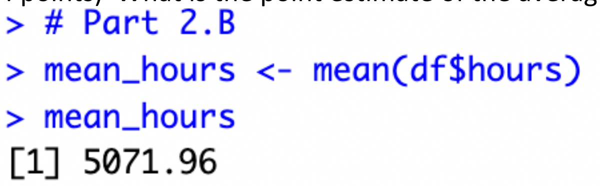
Question#2 (50 points)

Lithonia Lighting, manufacturer of light bulbs, recently came up with a new design of LED bulb. The company would like to be able to say to the customers that the new bulb will last over 5,000 hours. One hundred light bulbs were randomly selected, and their lifetime was measured (hours until bulb "burns out"). The data are recorded in the file "**light_bulbs.csv**".

a. (1 point) What data type is used in the study? (numerical or categorical)

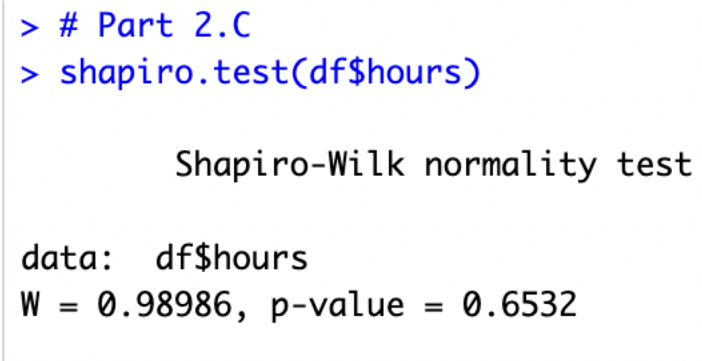
   a. Numerical, because this the time that it will take for a light to go out

b. (4 points) What is the point estimate of the average lifetime of a light bulb?



```
> # Part 2.B
> mean_hours <- mean(df$hours)
> mean_hours
[1] 5071.96
```

c. (5 points) Run the Shapiro-Wilk test to see if the data in the sample are normally distributed. In the space below, report the p-value for the test and your conclusion whether the data are normal or not.



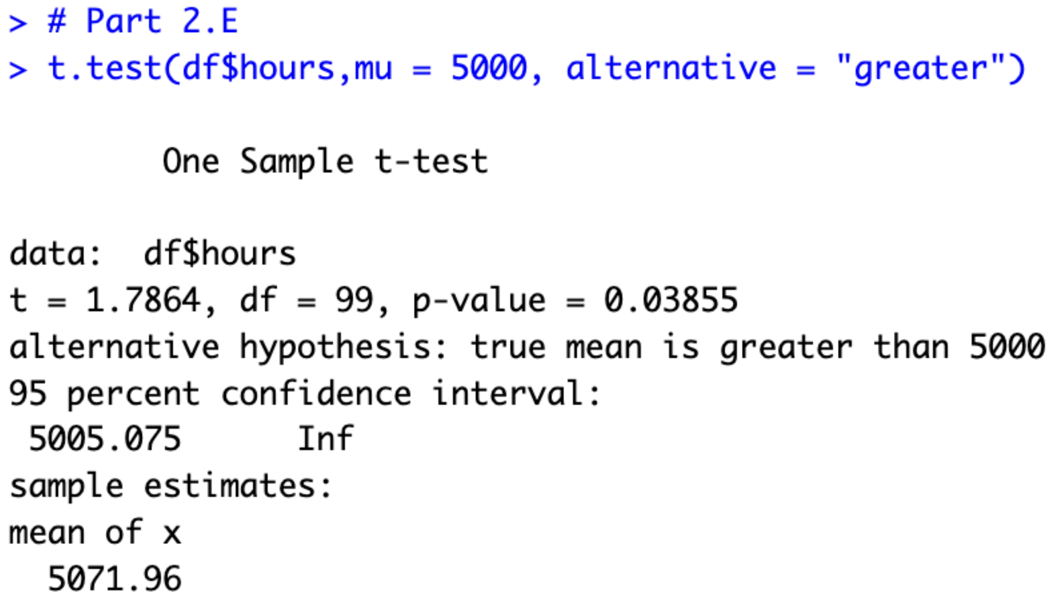
```
> # Part 2.C
> shapiro.test(df$hours)

        Shapiro-Wilk normality test

data:  df$hours
W = 0.98986, p-value = 0.6532
```

The data would be considered normal as the returned p-value is greater than the significance level of 0.05

d. (10 points) Formulate the null and alternative hypotheses that need to be tested in order to find out if the light bulb lasts on average longer than 5,000 hours.

   a. Null: Mean <= 5000
   b. Alternative: Mean > 5000

e. (15 points) Compute the p-value for the hypotheses formulated in part (d) above.



```
> # Part 2.E
> t.test(df$hours,mu = 5000, alternative = "greater")

        One Sample t-test

data:  df$hours
t = 1.7864, df = 99, p-value = 0.03855
alternative hypothesis: true mean is greater than 5000
95 percent confidence interval:
 5005.075      Inf
sample estimates:
mean of x
  5071.96
```

f. (15 points) Conclude the hypothesis test and formulate your findings in plain English. At 5% significance level, is there enough evidence in the sample collected to conclude that the new light bulb lasts longer than 5,000 hours? In order to receive full credit, you must provide written explanation for your answer.

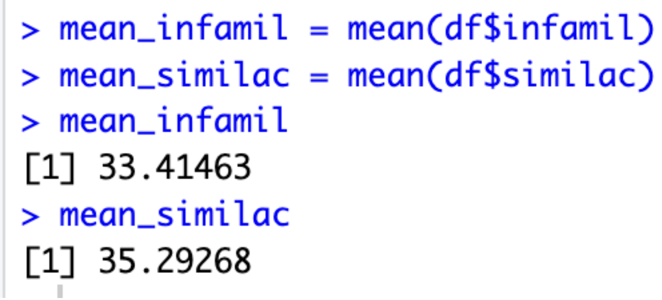
a. No, because the p-value of 0.03855 is less than the 0.05 significance level to determine that there is enough evidence to conclude that the mean time is greater than 5000 hours for a light bulb to go out. Thus, we will reject the alternative hypothesis of the mean time is greater than 5000 hours.

BUSN 231: Applied Business Statistics
Homework#9
Due Date: October 22, 2023 at 11:59pm

The homework submitted should represent your INDIVIDUAL work. Any cases of plagiarism will be treated according to the Honor Code and the CNU procedures. You can ask any questions to me.

Question#1 (50 points)
Newborn babies normally lose between 5% to 7% of the weight in the first week of life. For proper development, it is important to ensure that babies quickly gain weight after the initial weight loss. Two leading manufacturers of baby-food are Infamil and Similac. Similac recently came up with the new baby formula, and claims that it allows babies to gain weight faster than the Infamil's formula. To test this claim, a pediatrician at Riverside Health Care facility decides to track weight gain between week 1 and week 4 of life for a sample of babies fed with Similac and for a different sample of babies fed with Infamil formulas. The weight gains (in ounces) are recorded in the file "**baby_food.csv**".

n. (5 point) What are the point estimates of the average weight gains for babies fed with Infamil and for babies fed with Similac?

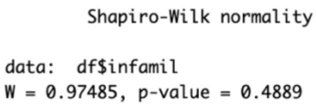


```
> mean_infamil = mean(df$infamil)
> mean_similac = mean(df$similac)
> mean_infamil
[1] 33.41463
> mean_similac
[1] 35.29268
```

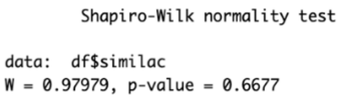
o. (5 points) Does it appear that feeding babies with Similac's formula leads to a higher weight gain? Briefly explain why or why not.
Yes, because the point of estimate for feeding babies with Similac is greater than the point of estimate for feeding babies with Infamil

p. (10 point) Formulate null and alternative hypotheses that the pediatrician needs to test in order to statistically confirm or disprove that the Similac's new formula leads to a bigger weight gain than Infamil.
H0: (Mean of Similac - Mean of Infamil) <= 0
HA: (Mean of Similac - Mean of Infamil) > 0

q. (10 points) Test both samples for normality. Report p-values for Shapiro-Wilk tests for both samples. State your conclusion of the samples are normally distributed or not.



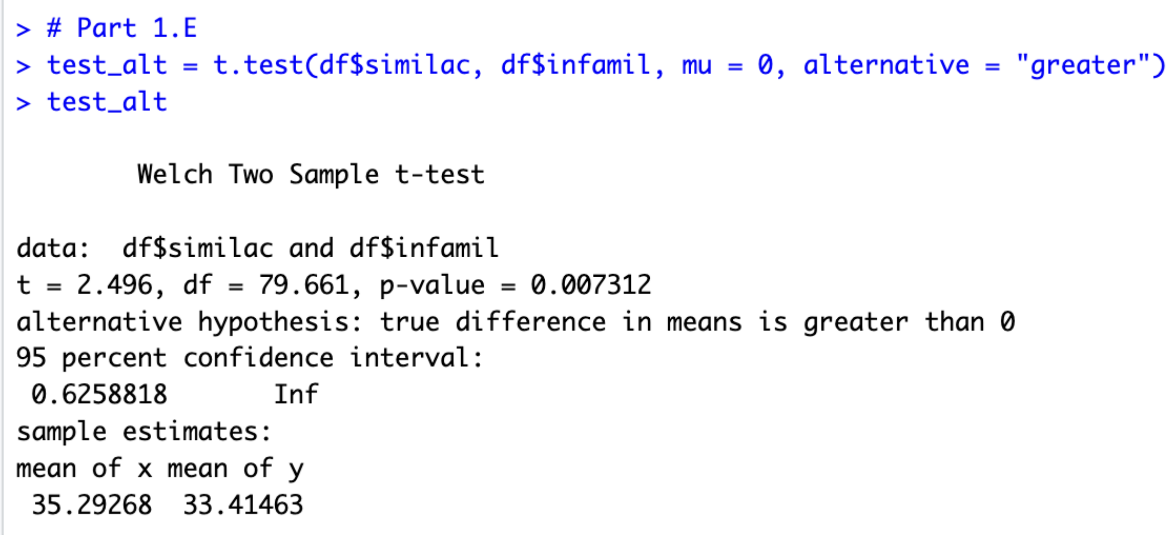
```
        Shapiro-Wilk normality test

data:  df$infamil
W = 0.97485, p-value = 0.4889
```



```
        Shapiro-Wilk normality test

data:  df$similac
W = 0.97979, p-value = 0.6677
```

Both samples are normally distributed, as both the p-values are above the assumed significance level of a = 0.05

r. (10 points) For the hypotheses formulated above in part (c), find the p-value.



```
> # Part 1.E
> test_alt = t.test(df$similac, df$infamil, mu = 0, alternative = "greater")
> test_alt

        Welch Two Sample t-test

data:  df$similac and df$infamil
t = 2.496, df = 79.661, p-value = 0.007312
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 0.6258818       Inf
sample estimates:
mean of x mean of y 
 35.29268  33.41463 
```

s. (10 points) At 1% significance level, can the pediatrician conclude that babies gain more weight when fed with Similac's formula? Explain why or why not.

No, because the p-value of this test is 0.0073, which is less than the significance level of a = 0.01. Thus, there is not enough evidence to conclude that Similac causes greater weight gain in babies compared to Infamil, so we will be rejecting the null hypothesis.

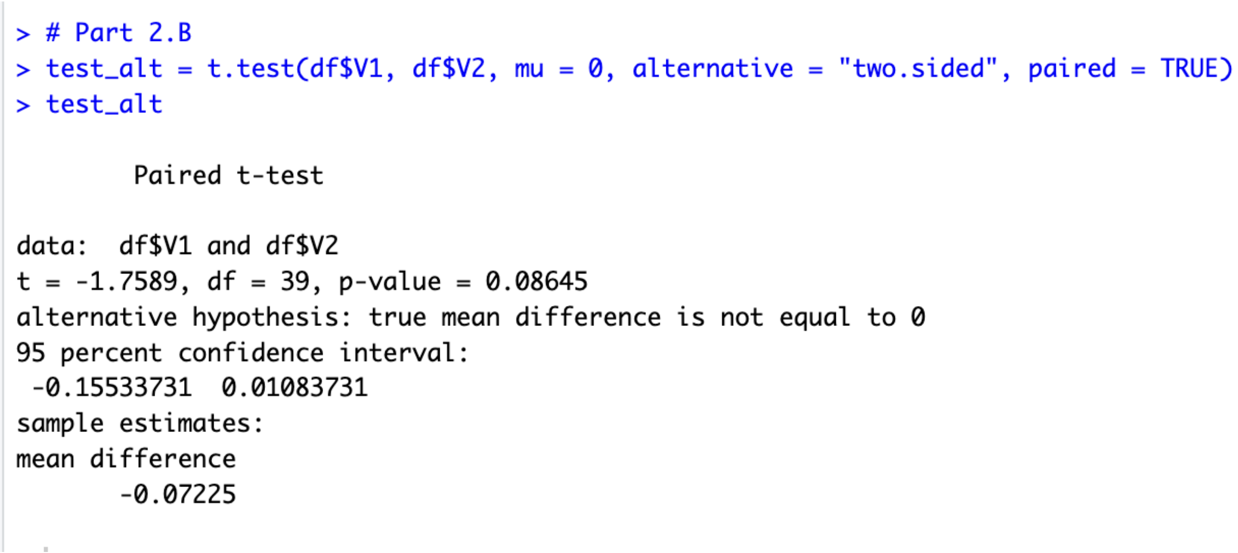
Question#2 (50 points)
A pupillometer is a device used to measure the changes in pupil dilations as the eye is exposed to different visual stimuli. Marketing researchers believe that there is a direct relationship between the amount of pupil dilation change and the person's interest in the visual stimulus. Therefore, marketing researchers sometimes use pupillometer to help evaluate the potential customers' interest in new products (Optical Engineering, March 1995). The Design and Market Research Laboratories has created two different image patterns for the promotion of a new laptop, and would like to see if the two

patterns generate different levels of interest from potential buyers. The researchers randomly selected 40 volunteers to participate in the marketing study. First, each person was shown pattern 1, and after 20 minutes of rest, each person was shown pattern 2. Each time the pattern was shown, the change in dilation of the volunteer's pupils was measured in millimeters, and recorded in the file "**pupillometer.csv**".

g. (20 point) Formulate the appropriate hypotheses to test in order to determine if the two patterns differ in level of interest towards them from potential consumers.

H0: (Mean of Pattern1- Mean of Pattern2) = 0
HA: (Mean of Pattern1- Mean of Pattern2) != 0

h. (15 points) What is the p-value for the hypotheses formulated above?



```
> # Part 2.B
> test_alt = t.test(df$V1, df$V2, mu = 0, alternative = "two.sided", paired = TRUE)
> test_alt

        Paired t-test

data:  df$V1 and df$V2
t = -1.7589, df = 39, p-value = 0.08645
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
 -0.15533731  0.01083731
sample estimates:
mean difference
       -0.07225
```

i. (15 points) Formulate the conclusion for your investigation. At 5% significance level, do the volunteers' reactions indicate that both patterns are equally stimulating for potential customers? Provide explanation to your answer.

At the 5% significance level, it can be concluded the volunteers' reactions indicate that both patterns are equally stimulating for potential customers. Because the p-value of the t test (0.08645) is greater than the significance level (0.05), there is sufficient evidence to back this claim. Thus, we fail to the reject the null hypothesis.

BUSN 231: Applied Business Statistics
Homework#10
Due Date: October 29, 2023 at 11:59pm

The homework submitted should represent your INDIVIDUAL work. Any cases of plagiarism will be treated according to the Honor Code and the CNU procedures. You can ask any questions to me.

Question#1 (35 points) - Chi-squared Goodness-of-Fit

Internet search engines are used by people daily. In 2008 Google processed 42%, Yahoo – 21%, and Bing –19% of all searches. The rest of the Internet searches market (18%) was shared among smaller companies, many of which are now defunct (Ask.com, Quora.com, etc.). Since then online search engines expanded aggressively. To see if the market distribution of among search engines changed, a survey was conducted. One thousand people were asked to identify their search engine of choice and responses were recorded as “Bing”, “Google” “Yahoo” and “other”. The data are recorded in the file “**internet_searches.csv**”.

j. (5 point) What type of data is used in this study?

   a. The Data is Nominal, for these are qualitative values that cannot be ranked, only categorized

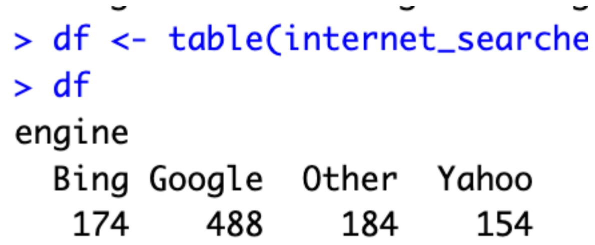
k. (10 points) Formulate null and alternative hypotheses to test if the search engines market redistributed since 2008.

   H0: P_Google = 0.42,P_Yahoo = 0.21, P_Bing = 0.19, P_Other = 0.18

   H1: That distribution of the searches by engine have changed since 2008

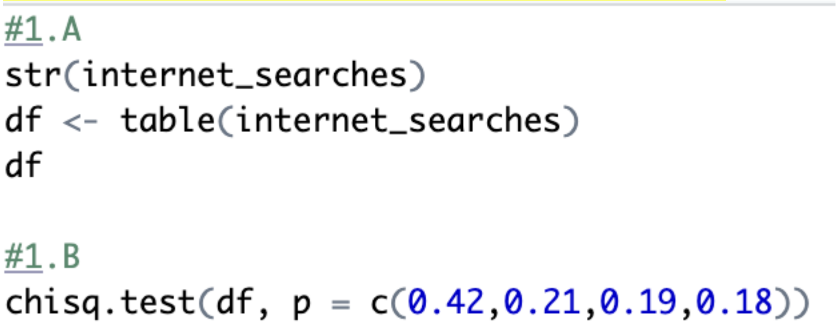
l. (10 points) What are actual observed frequencies of responses for Bing, Google, Yahoo, and other. If null hypothesis was true:

   Bing = 174
   Google = 448
   Other = 184
   Yahoo = 154

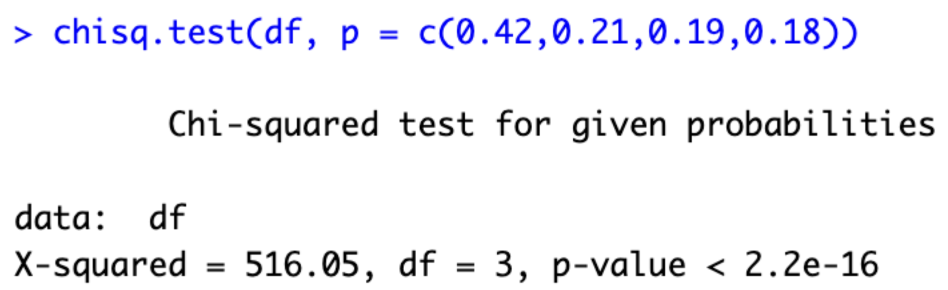


```
> df <- table(internet_searche
> df
engine
  Bing Google  Other  Yahoo
   174    488    184    154
```

m. (5 points) Using the **chisq.test()** function, find the p-value for the hypotheses formulated above. Copy-paste your R code below and report the p-value found.



```
#1.A
str(internet_searches)
df <- table(internet_searches)
df

#1.B
chisq.test(df, p = c(0.42,0.21,0.19,0.18))
```



```
> chisq.test(df, p = c(0.42,0.21,0.19,0.18))

        Chi-squared test for given probabilities

data:  df
X-squared = 516.05, df = 3, p-value < 2.2e-16
```

n. (5 points) In plain English, what conclusion can we make from the test? Can we claim that the search market re-distributed between companies as compared to 2008?

Because the p-value is less than the compared significance level of 0.05, we are able to conclude that we must fail to reject the alternative hypothesis, meaing the distribution of search engine use changed.

Question#2 (30 points) - Chi-squared Independent

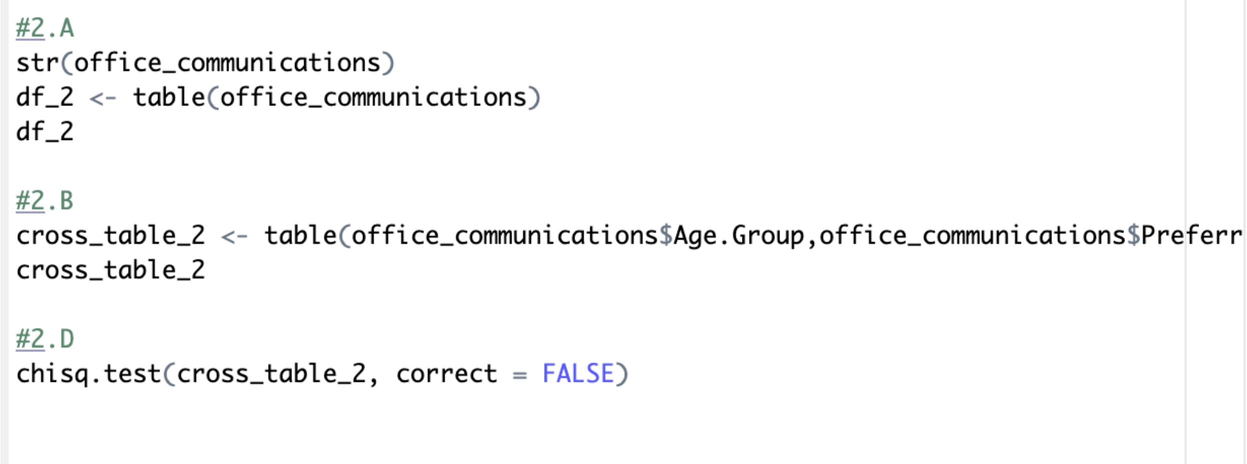
In 2003 USA Today reported on preferred types of office communications for different age groups ("Talking Face-to-Face vs. Group Meetings", USA Today, October 13, 2003). The results were based on a survey of a group of people in each of the then mature age groups (Generation Y, Generation X, Boomer, Mature). Possible responses were: Group Meetings, Meetings with Individuals, Email, Other (such as Skype). The data collected are recorded in the file "**office_communications.csv**".

a. (10 points) Formulate which hypotheses you need to test in order to investigate if a relationship exists between age group and preferred way of office communications (in other words, are each "generation" of workers' preferences of communications ways different).

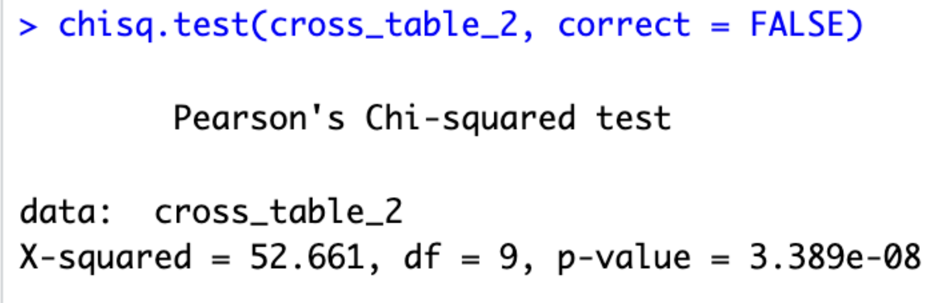
H0 : Variables "Age Group" and "Preferred Communication" are independent from each other

H1 : Variables "Age Group" and "Preferred Communication" are dependent

d. (5 points) Using the **chisq.test()** function, find the p-value for the hypotheses formulated above. Copy-paste your R code below and report the p-value found.



```
#2.A
str(office_communications)
df_2 <- table(office_communications)
df_2

#2.B
cross_table_2 <- table(office_communications$Age.Group,office_communications$Preferr
cross_table_2

#2.D
chisq.test(cross_table_2, correct = FALSE)
```



```
> chisq.test(cross_table_2, correct = FALSE)

        Pearson's Chi-squared test

data:  cross_table_2
X-squared = 52.661, df = 9, p-value = 3.389e-08
```

e. (5 points) At 5% significance level, does the sample collected provide enough evidence that different age groups prefer different ways of office communications? Briefly explain the reasoning behind your answer.

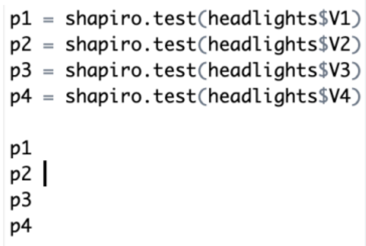
No, the sample collected does not show enough evidence to prove that different age groups prefer different communications. Thus, we will reject the null hypothesis and fail to reject the alternative hypothesis as this study's p-value is less than the 5% significance level.

Question#3 (35 points) - ANOVA

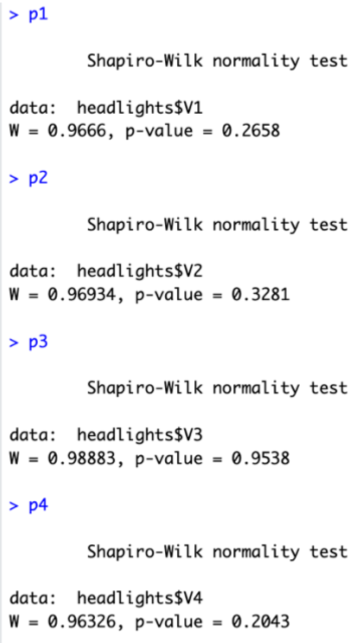
A car manufacturer designed four different head lights and would like to compare them. One of the tests performed on the headlights is how far away (in feet) a driver can read a road sign in night conditions with the headlights on. The company equips four identical cars each with one of the four lamp designs. Next four different groups with 41 drivers each are selected. Each driver in the first group drives the first car, equipped with design 1 lamp. When a driver can read a road sign in the dark, illuminated by the headlight, distance is recorded. The test is replicated with each driver in the second group driving car 2, equipped with design 2 lamp, and results are recorded. The process is continued with groups 3 and 4, driving cars 3 and 4, equipped with design 3 and 4 lamps. The data are recorded in the file "**headlights.csv**". The car company would like to determine if the four designs are any different in terms of average distance at which a road sign can be recognized at night.

a. (5 points) Explain why ANOVA is an appropriate test for this problem.
    a. ANOVA tests allow us to compare 3 or more population means to find the differences between three of more independent groups

b. (5 points) Test the normality requirements for the data. Report p-values, and conclusion form the tests



```
p1 = shapiro.test(headlights$V1)
p2 = shapiro.test(headlights$V2)
p3 = shapiro.test(headlights$V3)
p4 = shapiro.test(headlights$V4)

p1
p2
p3
p4
```



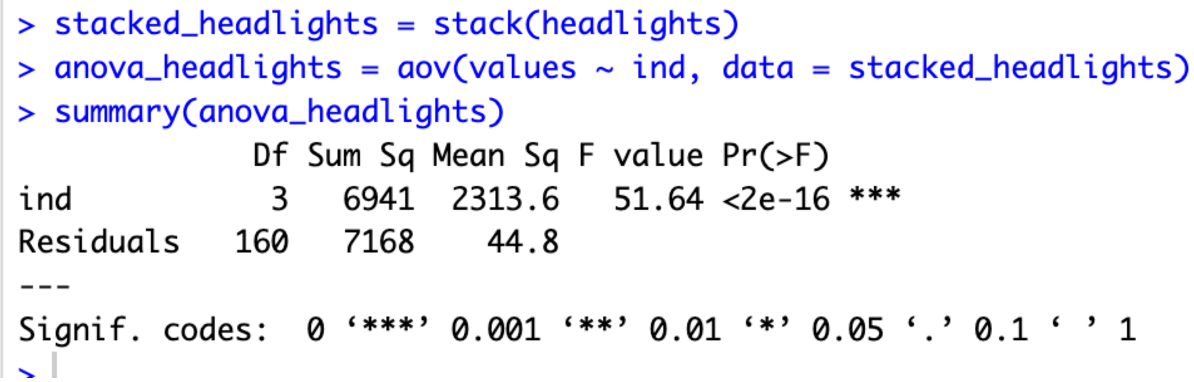
```
> p1

        Shapiro-Wilk normality test

data:  headlights$V1
W = 0.9666, p-value = 0.2658

> p2

        Shapiro-Wilk normality test

data:  headlights$V2
W = 0.96934, p-value = 0.3281

> p3

        Shapiro-Wilk normality test

data:  headlights$V3
W = 0.98883, p-value = 0.9538

> p4

        Shapiro-Wilk normality test

data:  headlights$V4
W = 0.96326, p-value = 0.2043
```

After running a Shapiro test on all four sets of data, it can be determined that all groups are normally distributed for their p-value is greater than the assumed 0.05 significant level. Therefore, we can use the 1-way ANOVA analysis.

c. (10 points) Formulate null and alternative hypotheses to be tested with this data.

H0: The means of the groups are equal
H1: Not all the means are equal

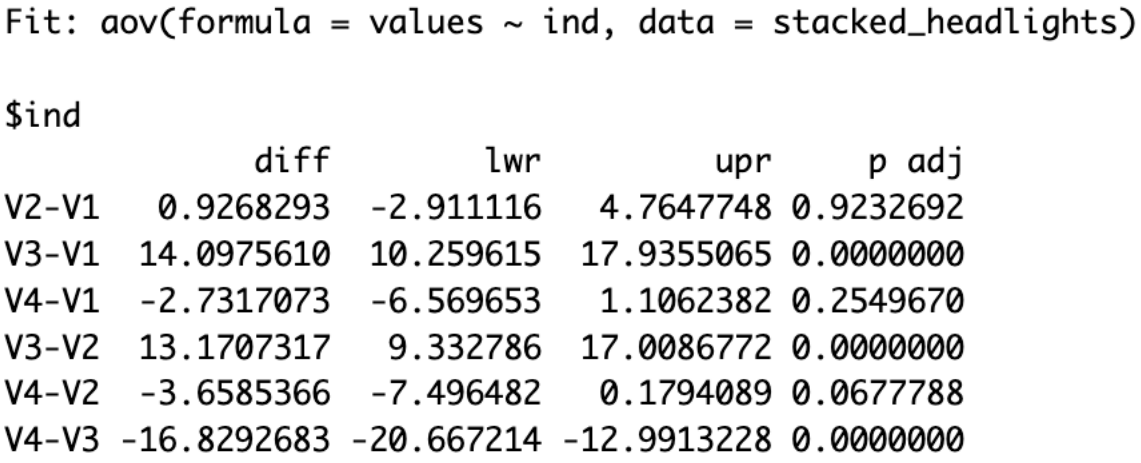
d. (5 points) Perform the appropriate test using R, report the p-value, and your conclusion, in plain English, from this test.



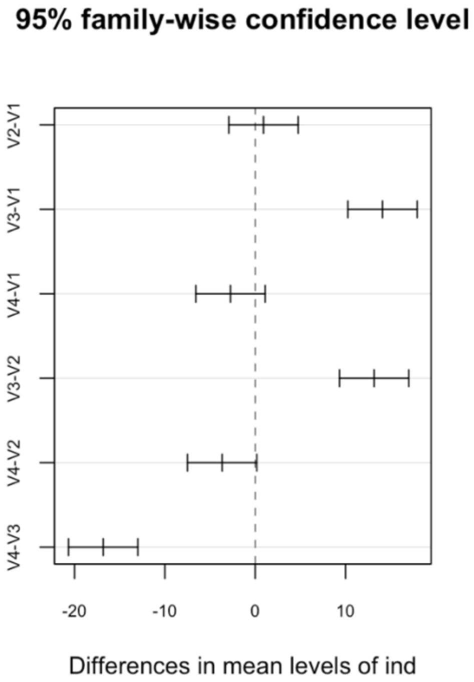
```
> stacked_headlights = stack(headlights)
> anova_headlights = aov(values ~ ind, data = stacked_headlights)
> summary(anova_headlights)
             Df Sum Sq Mean Sq F value Pr(>F)    
ind           3   6941  2313.6   51.64 <2e-16 ***
Residuals   160   7168    44.8                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

From this test, the data proved to not be significant, so we will be rejecting the null hypothesis, meaning there are differences in the mean between groups of data

e. (10 points) Conduct Tukey Test and plot the cart shows the mean difference for each pair of lamp designs, and comment on which design (if any) could be better/worse than others. Copy-paste the chart below. (Hint: use TukeyHSD() r code)



```
Fit: aov(formula = values ~ ind, data = stacked_headlights)

$ind
             diff        lwr         upr     p adj
V2-V1   0.9268293  -2.911116   4.7647748 0.9232692
V3-V1  14.0975610  10.259615  17.9355065 0.0000000
V4-V1  -2.7317073  -6.569653   1.1062382 0.2549670
V3-V2  13.1707317   9.332786  17.0086772 0.0000000
V4-V2  -3.6585366  -7.496482   0.1794089 0.0677788
V4-V3 -16.8292683 -20.667214 -12.9913228 0.0000000
```



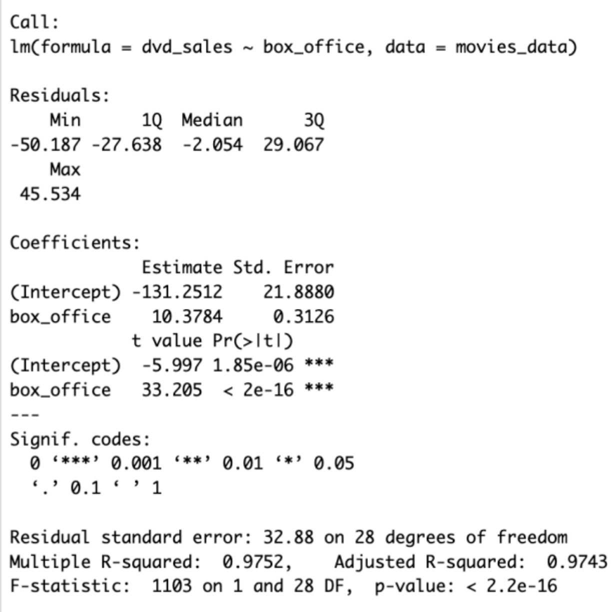
In terms of better designs, V1 and V2 are great as they are either consistently around the mean viewing distance or are above it, which is good in this scenario. For V3 and V4, they are typically behind it, and there for are not are great of designs.

BUSN 231: Applied Business Statistics
Homework#12 - Regression
Due Date: November 12, 2023 at 11:59pm

The homework submitted should represent your INDIVIDUAL work. Any cases of plagiarism will be treated according to the Honor Code and the CNU procedures. You can ask any questions to me.

Question#1 (50 points) - Simple Linear Regression
A company that holds DVD distribution rights to movies previously released only in theaters wants to estimate the sales of DVDs based on box office success of a movie. It will help this company to plan how many DVDs to produce and how to organize transportation. A company collected records on recently released movies. For each movie it recorded the box office gross (in $ millions), and number of DVDs sold (in thousands). The data are recorded in the file "**movie.csv**".

t. (5 point) Calculate the coefficient of correlation between DVD sales and box Office performance. Report the coefficient of correlation and characterize the strength of correlation.



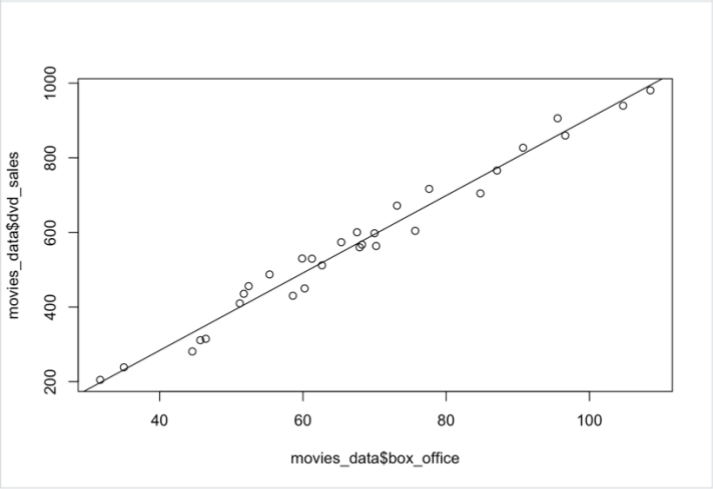
```
Call:
lm(formula = dvd_sales ~ box_office, data = movies_data)

Residuals:
    Min      1Q  Median      3Q 
-50.187 -27.638  -2.054  29.067 
    Max 
 45.534 

Coefficients:
             Estimate Std. Error
(Intercept) -131.2512    21.8880
box_office    10.3784     0.3126
            t value Pr(>|t|)    
(Intercept)  -5.997 1.85e-06 ***
box_office   33.205  < 2e-16 ***
---
Signif. codes:  
  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 
  ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.88 on 28 degrees of freedom
Multiple R-squared:  0.9752,	Adjusted R-squared:  0.9743 
F-statistic:  1103 on 1 and 28 DF,  p-value: < 2.2e-16
```

R^2, which is the Coefficient of Determination is found in the Multiple R-Squared value, equaling 0.9752. Taking the square root of this value gives us the coefficient of correlation, which is 0.9752

u. (10 points) Construct a model which allows predicting the number of DVDs sold (in thousands) based on box office gross (in $ millions). Report your regression model equation below.

dvds_sold = -131.2512 + 10.3784 * box_office
*Based on the results from #1.A

v. (5 point) Create a scatter plot for the data (using DVD sales as your target variable Y). Plot the regression line on the same plot. Copy-paste the plot below.



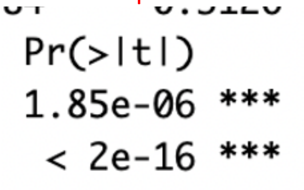
w. (5 points) Is this a statistically significant model? Test the regression slope to see if movie box office is statistically related to movie DVD sales.

To determine this, we need our null and alternative hypothesis:

H0: m != 0

H1: m = 0

Per the simple linear model from #1.A, here is the p-value



```
Pr(>|t|)
1.85e-06 ***
 < 2e-16 ***
```

Since the p-value is less than 0.0, we can reject H0 in favor of H1. Thus, we can determine that the statistically related, as the regression equation is different from 0.

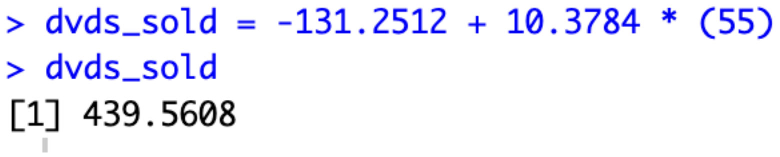
x. (5 points) Interpret the slope of the regression model. Provide a clear, plain English interpretation.

If the movie in the box office generates $0 from their movie, than -131.2512 thousand copies of their movie will sell on DVD. From this point, the amount of DVDs sold comes down to 10.3784 * the amount of money($) made in the box office

y. (5 points) Characterize how good your model fits the data. Explain the basis for your assessment. HINT: use model's R-squared.

The Coefficient of Determination, per #1.A, is R^2 = 0.9752. Given that the closer R^2 = 1 makes a perfect model, it can be deducted that the model fits the data almost perfectly, only allowing for 2.48% of the variation between box office numbers and dvd sales to go unexplained

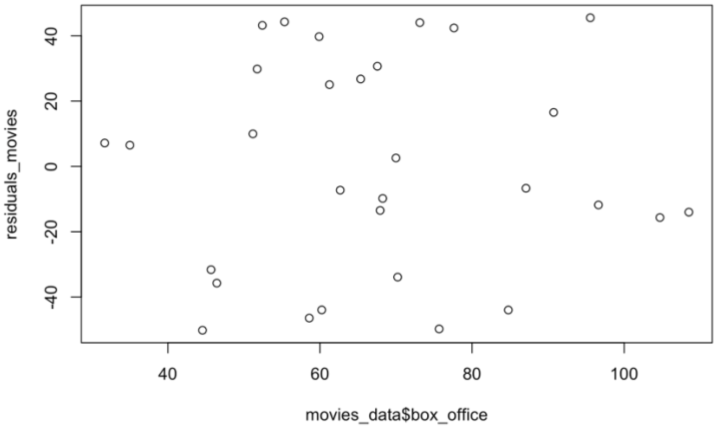
z. (5 points) For a movie with a box office gross of $55,000,000 ($55 Million), what are the projected DVDs sales? Provide details of your computations.



```
> dvds_sold = -131.2512 + 10.3784 * (55)
> dvds_sold
[1] 439.5608
```

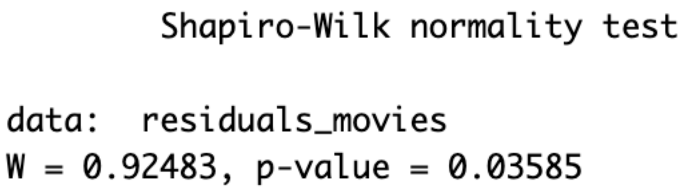
439.5608 Thousand DVDs Sold

aa. (5 points) Calculate model's residuals. Create a residuals plot vs movies' box office. Copy-paste a residual plot, and report if residual plot indicates any problems with the regression model.



The residuals vs box office plot demonstrates homoscedasticity, meaning that there are the same number of residuals above and below 0 with no pattern and an equal distribution. This makes no threat to the regression model

bb. (5 points) Test residuals for normality as well. Comment on whether or not the residuals are normally distributed.



```
        Shapiro-Wilk normality test

data:  residuals_movies
W = 0.92483, p-value = 0.03585
```

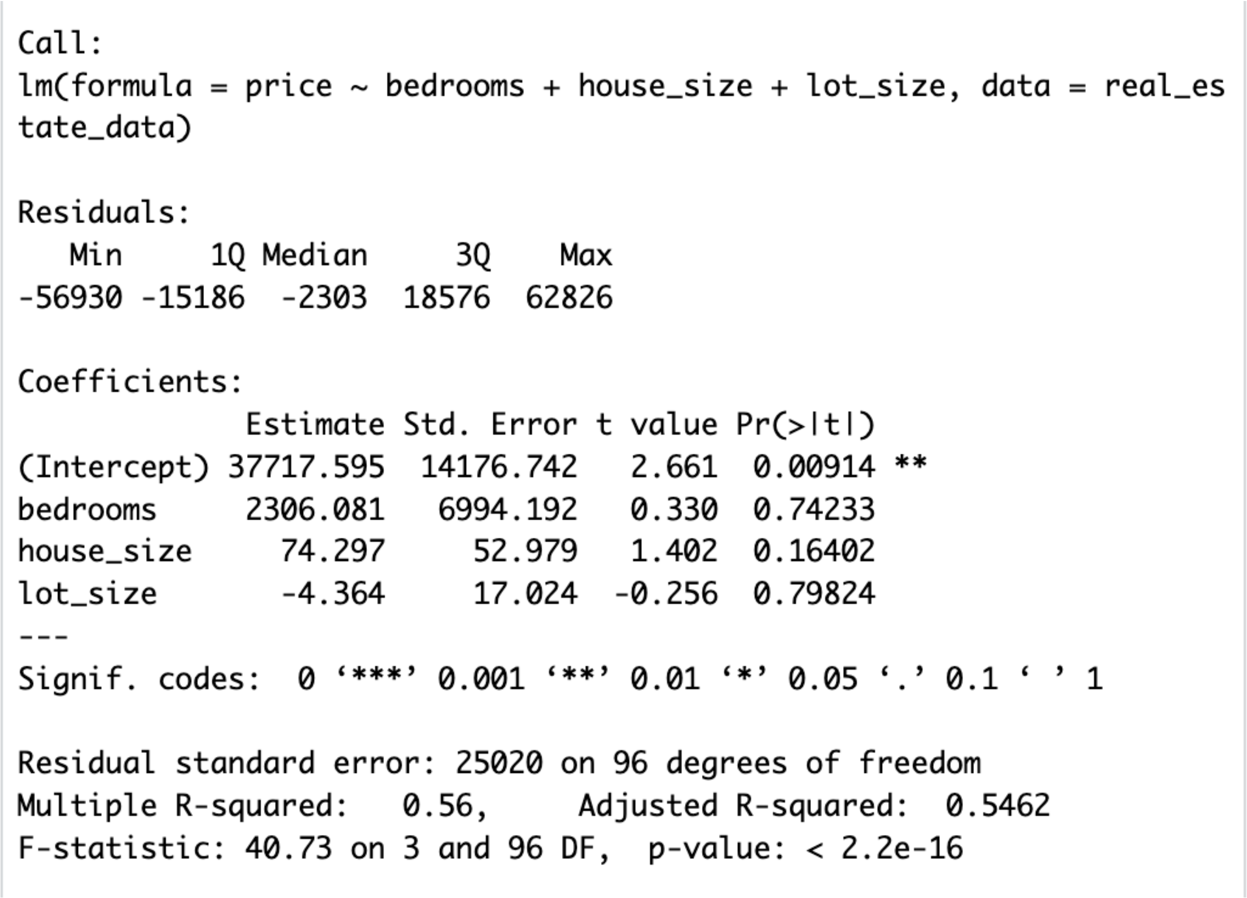
After testing the residuals for normality, it can be concluded that the residuals are normally distributed as the p-value is greater than the significance level of 0.01, so the residuals are normally distributed

Question#2 (50 points) - Multiple Linear Regression

A real estate agent wants to develop a model to predict the selling price of a home. The agent believes that the most important variables in determining the price of a house are its:

- Size of the house (square footage)
- Number of bedrooms
- Lot size where the house is built (also in square feet).

He collects from the market the data on a number of houses, which includes for every house the information listed above as well as the house price. The data are recorded in the file "**real_estate.csv**".



```
Call:
lm(formula = price ~ bedrooms + house_size + lot_size, data = real_es
tate_data)

Residuals:
   Min     1Q Median     3Q    Max 
-56930 -15186  -2303  18576  62826 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept) 37717.595  14176.742   2.661  0.00914 **
bedrooms     2306.081   6994.192   0.330  0.74233   
house_size     74.297     52.979   1.402  0.16402   
lot_size       -4.364     17.024  -0.256  0.79824   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25020 on 96 degrees of freedom
Multiple R-squared:  0.56,     Adjusted R-squared:  0.5462 
F-statistic: 40.73 on 3 and 96 DF,  p-value: < 2.2e-16
```

o. (10 point) Construct and report a regression model that allows predicting the selling price of the house if one knows house size, number of bedrooms, and a lot size.

Price = 37717.595 + 2306.081 * bedrooms + 74.297 * house_size + -4.364 * lot_size

p. (5 points) Characterize how good your model fits the data. Explain the basis for your assessment. HINT: use model’s R-squared.

Per #2.A, the adjusted R^2 = 0.5462. Given that the closer $R^2 = 1$ makes a perfect model, it can be deducted that the model fits about 50/50, only allowing for ~45% of the variation between the variables predicting price of the sold house to go unexplained. Needs improvement.

q. (5 points) Provide a plain English interpretation for the coefficients (the slopes) of the model built in (a) above. For each predictor variable, write your interpretation clearly and fully.

- Intercept(b0): When all other predictors are 0, then the base price for the real estate being sold is $37,717.595.

- Bedrooms(b1): while keeping all other variables unchanged, this is the increased rate of change experienced as more bedrooms are found in the house being sold. If all variables remain unchanged, and there is 1 bedroom in the house, the price automatically increase by $2,306.081

- House_size(b2): while keeping all other variables unchanged, this is the increased rate of change experienced as the size of the house increases. If all variables remain unchanged, and there is 1 sq ft found in the house, the price automatically increase by $74.297

- Lot_size(b3): while keeping all other variables unchanged, this is the decreased rate of change experienced as the property of the house increases. If all variables remain unchanged, and there is 1 sq ft found at the house, the price automatically descrease by -$4.364

r. (10 points) Are there any "useless" (not significant) predictor variables in the regression model? For each predictor, state the hypotheses that you are testing, p-value, and your conclusion whether or not this predictor is statistically significant. Use 5% significance level.

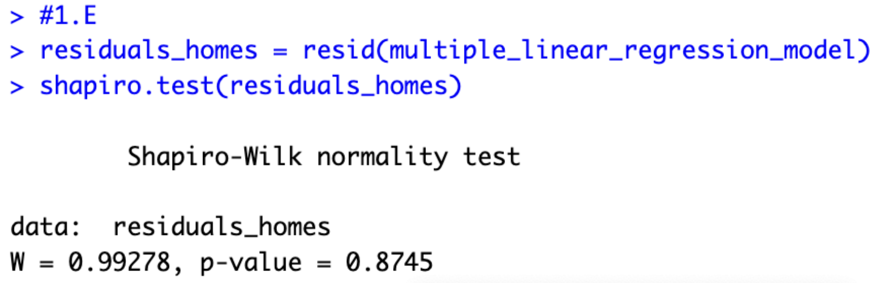
- Bedrooms
  - Hypothesis
    - H0: bi = 0
    - H1: bi != 0
  - P-value
    - P = 0.74233
  - Statistically Significant?
    - Given the p-value is larger than the 0.05 significance level, we fail to reject the null hypothesis, which means the variable is equal to 0. Thus, this predictor is considered insignificant, or useless, because it has no impact in predicting the price of the house sold.

- House Size
  - Hypothesis
    - H0: bi = 0
    - H1: bi != 0
  - P-value
    - P = 0.16402
  - Statistically Significant?
    - Given the p-value is larger than the 0.05 significance level, we fail to reject the null hypothesis, which means the variable is equal to 0. Thus, this predictor is considered insignificant, or useless, because it has no impact in predicting the price of the house sold.

- Lot Size
  - Hypothesis
    - H0: bi = 0
    - H1: bi != 0
  - P-value
    - P = 0.79824
  - Statistically Significant?
    - Given the p-value is larger than the 0.05 significance level, we fail to reject the null hypothesis, which means the variable is equal to 0. Thus, this

predictor is considered insignificant, or useless, because it has no impact in predicting the price of the house sold.

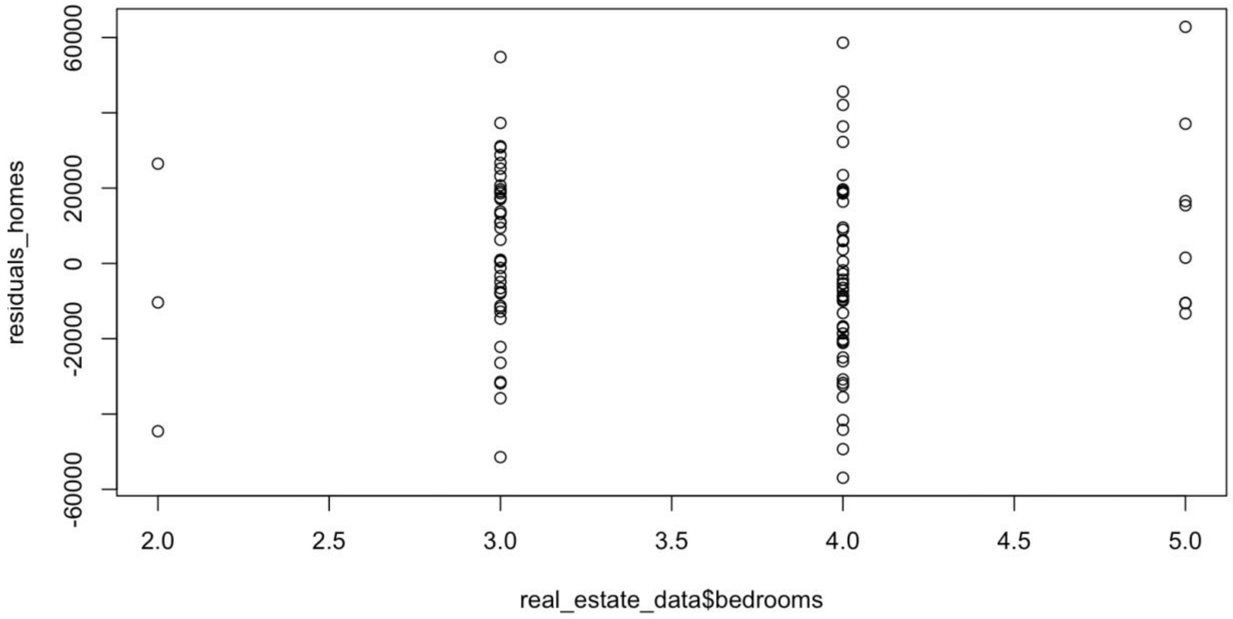
s. (10 points) Calculate model's residuals and test them for normality. Report your findings.

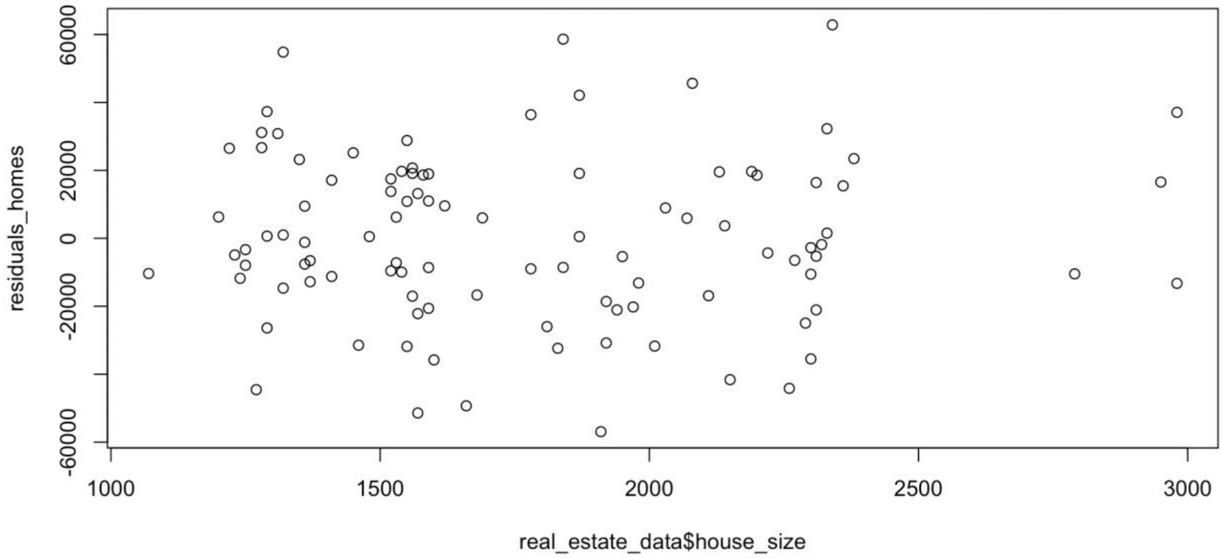


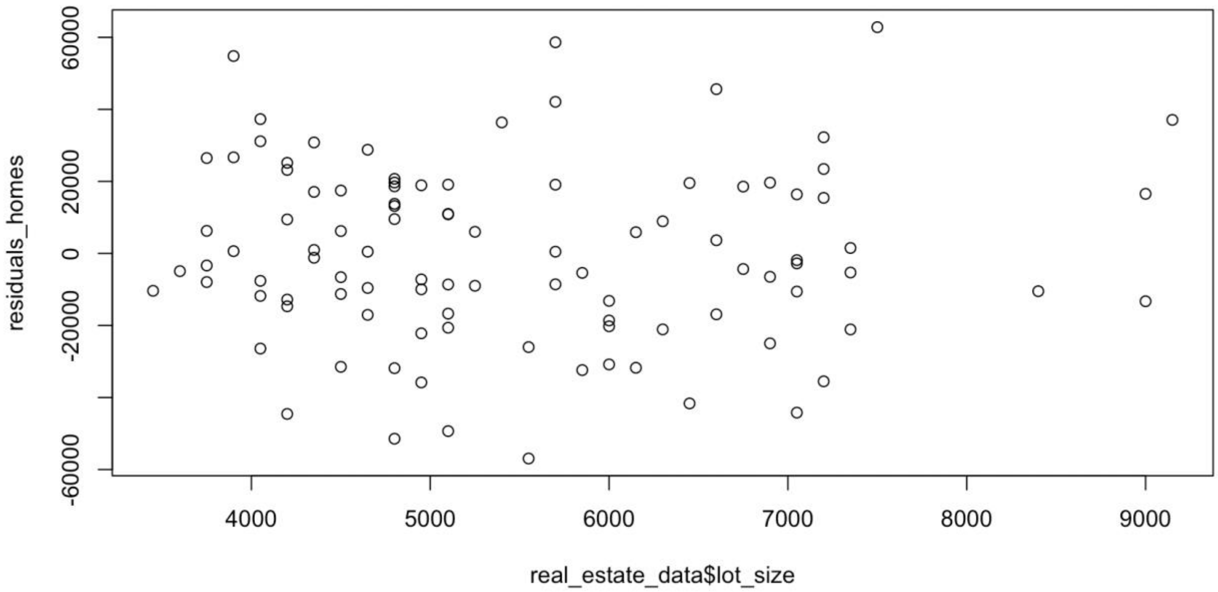
```
> #1.E
> residuals_homes = resid(multiple_linear_regression_model)
> shapiro.test(residuals_homes)

        Shapiro-Wilk normality test

data:  residuals_homes
W = 0.99278, p-value = 0.8745
```

- The data is normally distributed

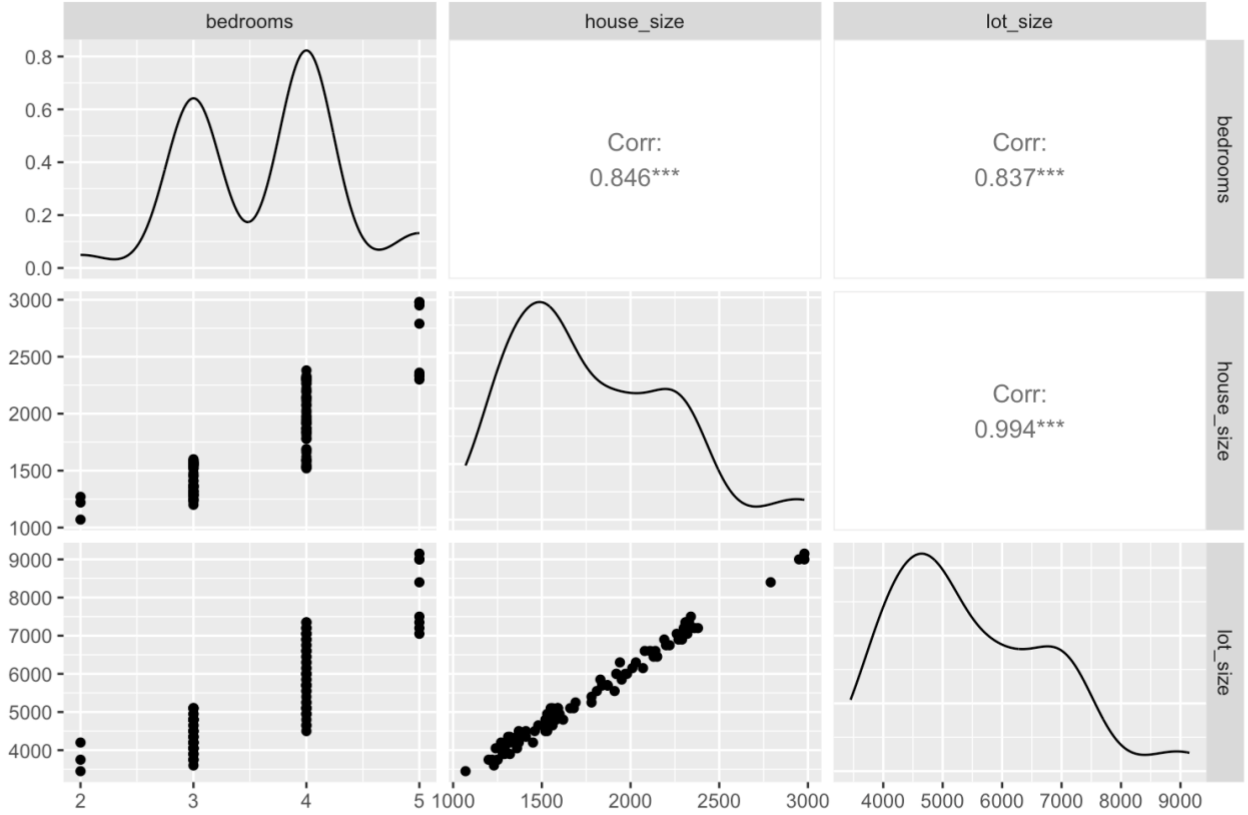
t. (5 points) Create residuals plots (vs house size, vs number of bedrooms, and vs lot size). Copy-paste the residuals plots in the space below. Comment if any of the residuals plots indicate any potential problems.




residuals_homes
-60000
-20000
0
20000
60000
1000
1500
2000
2500
3000
real_estate_data$house_size


residuals_homes
-60000
-20000
0
20000
60000
4000
5000
6000
7000
8000
9000
real_estate_data$lot_size

u. (5 points) Check the model for multicollinearity. Copy-paste the correlations matrix in the space below. Comment on whether or not the model suffers from multicollinearity.



- This model suffers from multicollinearity as all three correlations are between 0.7 and 1