Homework 1 F21

your name

9/10/2021

Homework Introduction

The goal of this homework is to get started coding in R and in R Markdown.

Add warning = FALSE and message = FALSE to the {} below to make the output look more clean.

In the remaining code chunks, add warning = FALSE and/or message = FALSE only as needed (don't just write both in every chunk.)

```
knitr::opts_chunk$set(echo = TRUE)

#  Before you start, load the tidyverse.
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse
1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Part 1

a. Create the data set Students

On this homework, we create a data set by entering vectors and putting them together to form the tibble, *Students*

In the code chunk below:

First create the following vectors

- ID: A sequence of numbers from 1 to 10 that is unique for each student
- year: Soph, Jr, Sr, Sr, Jr, Soph, Soph, Sr, Jr, Sr
- phonetime: 8, 2, 4, 7, 2, 1, 10, 3, 5, NA
- gpa: 2.75, 3.5, 3.2, 3.5, 3.5, 3, 2.5, 3.3, 2.9, 3.8

- job: FALSE, TRUE, FALSE, TRUE, FALSE, TRUE, FALSE, TRUE, FALSE, TRUE

Join the 5 vectors together in a tibble called Students

Afterwards, remove the 5 vectors from the global environment

Then print out the Student tibble

```
# Write lines of R code to do the following tasks.
# Include comments describing what you are doing.

# First, create the five vectors below:

ID <- 1:10

year <- c("Soph", "Jr", "Sr", "Sr", "Jr", "Soph", "Soph", "Sr", "Jr", "Sr")

phonetime <- c(8, 2, 4, 7, 2, 1, 10, 3, 5, NA)

gpa <- c(2.75, 3.5, 3.2, 3.5, 3.5, 3, 2.5, 3.3, 2.9, 3.8)

job <- c(FALSE, TRUE, FALSE, TRUE, FALSE, TRUE, FALSE, TRUE, FALSE, TRUE)

# Create the Students data frame next:
Students <- tibble(ID,
                   year,
                   phonetime,
                   gpa,
                   job)


# Use an R function to remove the five vectors from the global environment
rm(ID, year, phonetime, gpa, job)

# Print the data frame, by typing Students. You should see a 'tibble' of the
data file.
Students

## # A tibble: 10 x 5
##       ID year  phonetime   gpa job
##    <int> <chr>     <dbl> <dbl> <lgl>
##  1     1 Soph          8  2.75 FALSE
##  2     2 Jr            2  3.5  TRUE
##  3     3 Sr            4  3.2  FALSE
##  4     4 Sr            7  3.5  TRUE
##  5     5 Jr            2  3.5  FALSE
##  6     6 Soph          1  3    TRUE
##  7     7 Soph         10  2.5  FALSE
##  8     8 Sr            3  3.3  TRUE
## 9      9 Jr             5  2.9  FALSE
## 10    10 Sr            NA  3.8  TRUE
```

b. Stats on Students

- Find the mean and median of GPA and phonetime

```
mean(Students$gpa)

## [1] 3.195

median(Students$gpa)

## [1] 3.25

mean(Students$phonetime,
     na.rm = T)

## [1] 4.666667

median(Students$phonetime,
       na.rm = T)

## [1] 4
```

- Create a table showing the frequencies for year.

```
table(Students$year)

##
##   Jr Soph   Sr
##    3    3    4
```

- Calculate the percentage of students that have a job.

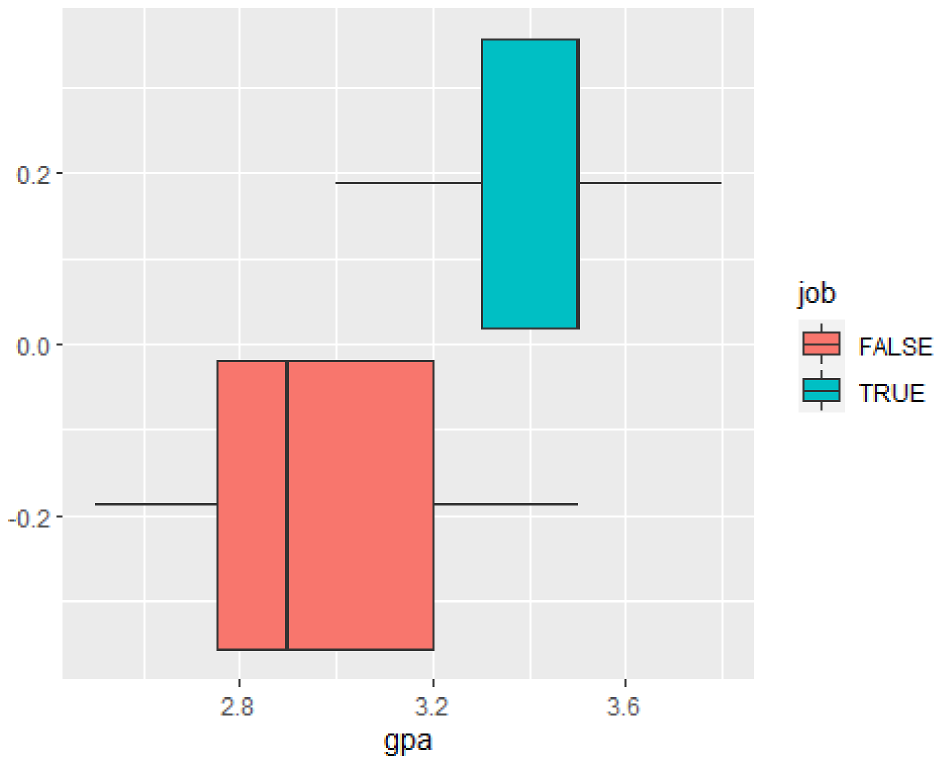
```
mean(Students$job)*100

## [1] 50
```

c. Plots of Students

Create a boxplot of gpa by job.

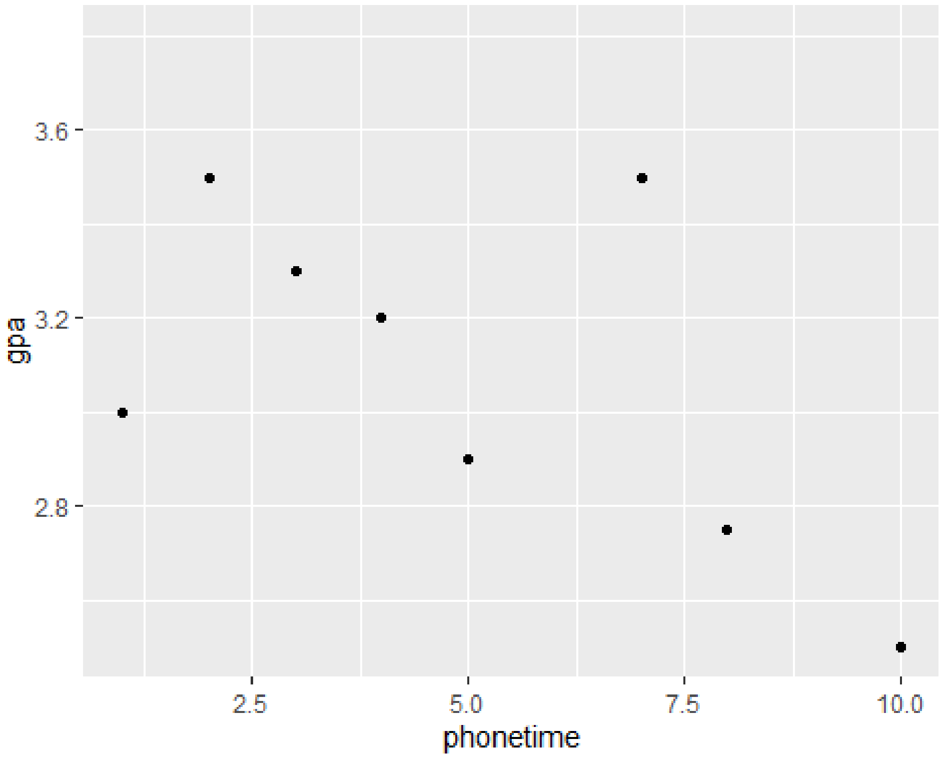
Run the code described below.

```
#  Use ggplot to make a boxplot of gpa by job.
#  Include fill=job inside geom_boxplot() so the boxplots are different
colors.

ggplot(data = Students,
       mapping = aes(x = gpa, fill = job)) +
  geom_boxplot()
```



Next, create a scatterplot of gpa by phonetime

```
# Use ggplot to make a scatterplot of gpa by phonetime.

ggplot(data = Students,
       mapping = aes(x = phonetime,
                     y = gpa)) +
  geom_point()

## Warning: Removed 1 rows containing missing values (geom_point).
```



Describe here the two relationships you observe.

What seems to be the effect of having a job for these students?

What seems to be the effect of time spent on one's phone?

Do you think these (made-up) students are typical, or do you think the actual trend among all students could be different?

Part 2

a. Read "Lebron James.csv"

- Download the data file 'Lebron James.csv' from Blackboard and put it in the same folder as this markdown file.
- Create data frame LBJ using **read.csv()** then save it as a tibble using **tibble()**.
- Print the first 10 rows of the data set.

```
LBJ <- read.csv("Lebron James.csv")

head(LBJ, n = 10)

##     Season Team Home Opponent Minutes_Played Shot_Attempts Shot_Proportion
## 1  2006/07  CLE Home      WAS       40:38:00            11           0.458
## 2  2006/07  CLE Away      SAS       41:53:00            14           0.538
## 3  2006/07  CLE Away      CHA       38:06:00             3           0.231
## 4  2006/07  CLE Home      ATL       47:17:00            13           0.500
## 5  2006/07  CLE Home      CHI       37:50:00            6           0.462
## 6  2006/07  CLE Home      BOS       43:32:00            9           0.529
## 7  2006/07  CLE Away      NYK       41:20:00           10           0.526
## 8  2006/07  CLE Home      POR       38:54:00           10           0.667
## 9  2006/07  CLE Home      MIN       40:30:00           12           0.500
## 10 2006/07  CLE Away      WAS       33:20:00            8           0.400
##    Rebounds Assists Steals Blocks Turnovers Personal_Fouls Points
Game_Result
## 1         10       5      0      2         5              2     26
Win
## 2         10       4      1      1         2              3     35
Win
## 3          9       7      0      1         2              0     16
Loss
## 4          7       6      2      1         2              1     34
Loss
## 5          4      12      3      2         3              0     19
Win
## 6          8       5      3      0         2              4     38
Win
## 7          4       6      2      1         3              2     29
Win
## 8          7       7      2      1         4              2     32
Win
## 9          9       6      1      2         4              1     37
Win
## 10         5       4      2      0         3              2     20
Loss
##    Point_Differential
## 1                   3
## 2                   7
## 3                  -4
## 4                  -9
## 5                  19
## 6                   1
## 7                   6
## 8                  13
## 9                  16
## 10                -12
```

b. Better at Home or Away?

Calculate the 5 number summary for *Shot_Proportion* when Lebron plays at home and away. Use the **aggregate()** function. Then describe the difference, if any, between when he plays home vs away games.

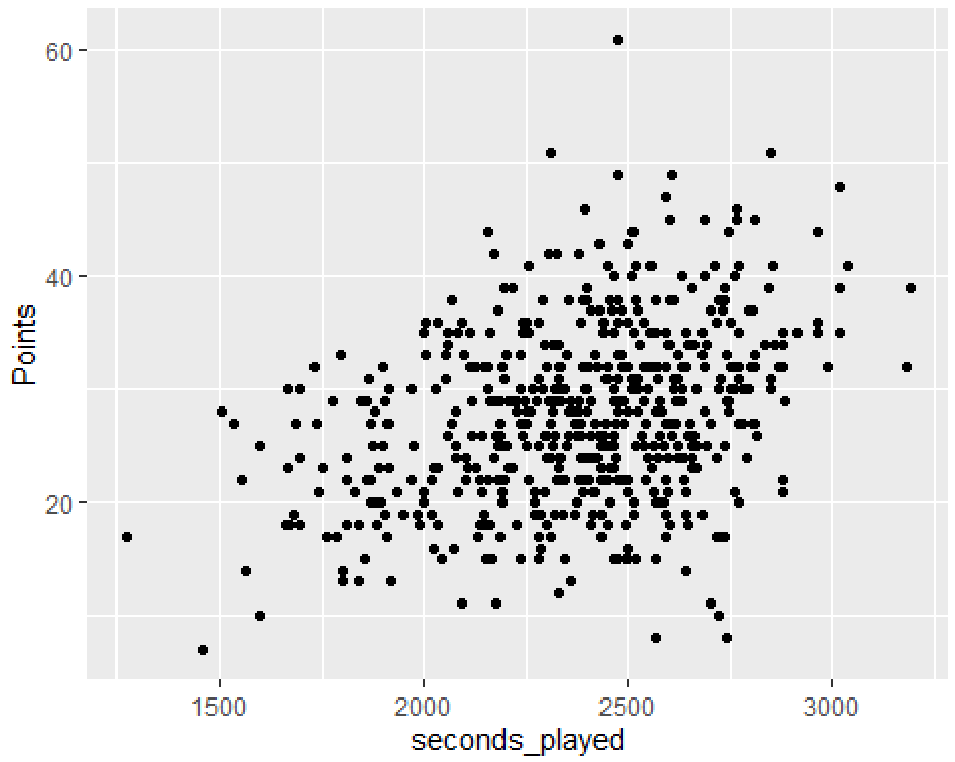
```
tapply(X = LBJ$Shot_Proportion,
       INDEX = LBJ$Home,
       FUN = summary)
```

```
## $Away
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1110  0.4210  0.5000  0.4915  0.5630  0.8330
##
## $Home
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2140  0.4440  0.5000  0.5145  0.5867  0.8460
```

c. Points by Time Played: Plot and Correlation

Run the specified code below. When done, describe the relationship between time played (in seconds) and points scored.

```
# Make a scatterplot of Points by time_played.

# Need to convert minutes played from a string to a number (in seconds).
# Use the code below for the conversion

LBJ <- LBJ %>%
  mutate(time = lubridate::ms(substr(Minutes_Played, 1, 5)),
         seconds_played = lubridate::period_to_seconds(time))



#  Create the scatterplot:

LBJ %>%
  ggplot(aes(x = seconds_played, y = Points)) +
  geom_point()
```



```
#  Use the function cor.test(xvector, yvector) to help assess the
relationship

cor.test(LBJ$seconds_played, LBJ$Points)

## 
##  Pearson's product-moment correlation
## 
## data:  LBJ$seconds_played and LBJ$Points
## t = 8.4784, df = 537, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2668832 0.4159793
## sample estimates:
##       cor 
## 0.3435946
```