# Homework 2

Adyan Rahman

## Problem 1

Part a:

```
# Matrix has unnecessary spaces and extra lines. Those were removed and altered.
mat <- matrix(c(34, 23, 53, 6, 78, 93, 12, 41, 99), nrow = 3)

# Nothing needed to be changed here!
df <- as.data.frame(mat)

# The names were too long, and was written with periods in between instead
# of underscores
names(df) <- c("car_drivertst_scr", "van_drivertst_scr", "truck_drivertst_scr")
```

Part b:

```
# Spaces in front and after "ggplot2" were removed. ggplot and 2 were combined
# to get rid of the space, as that would of caused an error
library(ggplot2)

# Got rid of unnecessary space for the next line and in front/behind "mpg"
head(mpg)
```

```
## # A tibble: 6 x 11
##   manufacturer model displ  year   cyl trans      drv     cty   hwy fl    class
##   <chr>        <chr> <dbl> <int> <int> <chr>      <chr> <int> <int> <chr> <chr>
## 1 audi         a4      1.8  1999     4 auto(l5)   f        18    29 p     compa~
## 2 audi         a4      1.8  1999     4 manual(m5) f        21    29 p     compa~
## 3 audi         a4      2    2008     4 manual(m6) f        20    31 p     compa~
## 4 audi         a4      2    2008     4 auto(av)   f        21    30 p     compa~
## 5 audi         a4      2.8  1999     6 auto(l5)   f        16    26 p     compa~
## 6 audi         a4      2.8  1999     6 manual(m5) f        18    26 p     compa~
```

```
# changed the variable name to "mpg2" to make it more concise
mpg2 <- mpg [mpg $ cyl == 6, ]

# Combined "mpg" and "2" in order to make sure an error isn't thrown.
mpg2$class <- as.character (mpg2$class)
```

## Problem 2

Accessing the 1976 - 2020 senate data set

```r
# Setting the working directory
setwd("C:\\Users\\ktzr1\\OneDrive\\Desktop\\STAT3355 Datasets")

# Opening the senate data set
senate <- read.csv("1976-2020-senate.csv")
```

Part a:

```r
# Putting the columns needed to be factors into a vector
names <- c('year' ,'state', 'party_simplified')

# Converting the columns into factor variable
senate$party_simplified <- factor (senate$party_simplified)
senate$year <- factor(senate$year)
senate$state <- factor(senate$state)
```

Part b:

```r
# Creating a subset where every observation is in Texas
texas_senate <- senate [senate$state == "TEXAS" , c("year", "state", "candidatevotes",
                                                    "totalvotes", "party_simplified")]
```

Part c:

```r
# Calculating the total of all the party votes, average votes, and median votes
# by calling the candidate votes and the party from the Texas Senate subset
party_totals <- tapply(texas_senate$candidatevotes, texas_senate$party_simplified, sum)
party_average <- round(tapply(texas_senate$candidatevotes, texas_senate$party_simplified, mean))
party_median <- round(tapply(texas_senate$candidatevotes, texas_senate$party_simplified, median))

# Putting the results into a 4x4 data frame in order to make it organized
result_df <- data.frame(
  Party = names (party_totals),
  Total_Votes = as.integer (party_totals),
  Average_Votes = as.integer (party_average),
  Median_Votes = as.integer (party_median)
)

# Printing data frame to view the calculated data
print (result_df)
```

```
##         Party Total_Votes Average_Votes Median_Votes
## 1    DEMOCRAT    38660127       2416258      2112490
## 2 LIBERTARIAN     1206600         92815        72657
## 3       OTHER      409126         21533         4564
## 4  REPUBLICAN    48318995       3019937      2761660
```

Part d:

```r
# Find the maximum number of candidate votes in Texas
max_votes <- max(texas_senate$candidatevotes)
```

```r
# Filter for Democratic candidates who received the maximum number of votes
democratic_winners <- subset(texas_senate, party_simplified == "Democratic" & candidatevotes == max_vot

# Get the year(s) when the Democratic candidate(s) won with the highest number of votes
winning_years <- democratic_winners$year

# Print the year(s)
print(winning_years)
```

```
## factor()
## 24 Levels: 1976 1978 1980 1982 1984 1986 1988 1990 1992 1994 1996 1998 ... 2021
```

# Problem 3

Part a:

```r
# Setting working directory
setwd("C:\\Users\\ktzr1\\OneDrive\\Desktop\\STAT3355 Datasets")

# Reading the data table through read.csv to get the different variables
ta_data <- read.csv("tae.data", header = FALSE)

ta_data$TA_ID <- 1:nrow (ta_data)

# Rename the columns
names(ta_data) <- c("English Speaking", "Course Instructor", "Course",
                    "Semester Type", "Class Size", "Class Attribute", "TA_ID")

# Convert values in the column to logical by setting 1 to TRUE
ta_data$`English Speaking` <- ta_data$`English Speaking` == 1
```

Part b:

```r
# Convert values in the column to logical by setting 2 to TRUE for
# 2 = regular
ta_data$`Semester Type` <- ta_data$`Semester Type` == 2
```

Part c:

```r
# Convert evaluation score to an ordered factor variable with labels
# 'low', 'medium', and 'high'
ta_data$`Class Attribute` <- factor(ta_data$`Class Attribute`,
                                    levels = c(1, 2, 3),
                                    labels = c("low", "medium", "high"),
                                    ordered = TRUE)
```

Part d:

```r
# Calculate average and median class sizes for regular semester
regular_class_data <- subset(ta_data, `Semester Type` == TRUE)
regular_mean <- mean(regular_class_data$`Class Size`)
regular_median <- median(regular_class_data$`Class Size`)

# Calculate average and median class sizes for summer semester
summer_class_data <- subset(ta_data, `Semester Type` == FALSE)
summer_mean <- mean(summer_class_data$`Class Size`)
summer_median <- median(summer_class_data$`Class Size`)

# Create a data frame to store the results
summary_df <- data.frame (
  Semester = c("Regular", "Summer"),
  Average_Class_Size = c(round(regular_mean, digits = 2), round(summer_mean, digits = 2)),
  Median_Class_Size = c(round(regular_median, digits = 2), round(summer_median, digits = 2))
)

# Printing the data frame
print (summary_df)
```

```
##   Semester Average_Class_Size Median_Class_Size
## 1  Regular              29.34                29
## 2   Summer              19.70                20
```

Part e:

```r
# Convert semester to factor with appropriate labels
ta_data$`Semester Type` <- factor(ta_data$`Semester Type`, labels = c("Summer", "Regular"))

# Convert English Speaking to factor
ta_data$`English Speaking` <- factor(ta_data$`English Speaking`, labels = c("English", "Non English"))

# Count the number of native and non-native English speaker TAs in regular semester
regular_eng <- sum(ta_data$`Semester Type` == "Regular" & ta_data$`English Speaking` == "English")
regular_noneng <- sum(ta_data$`Semester Type` == "Regular" & ta_data$`English Speaking` == "Non English

# Count the number of native and non-native English speaker TAs in summer semester
summer_eng <- sum(ta_data$`Semester Type` == "Summer" & ta_data$`English Speaking` == "English")
summer_noneng <- sum(ta_data$`Semester Type` == "Summer" & ta_data$`English Speaking` == "Non English")

# Create a data frame to store the counts
ta_counts <- data.frame (
  Semester = c("Regular", "Summer"),
  English = c(regular_eng, summer_eng),
  Non_English = c(regular_noneng, summer_noneng)
)

# Print the data frame
print(ta_counts)
```

```
##   Semester English Non_English
## 1  Regular     108          20
## 2   Summer      14           9
```

## Problem 4

The article relates to my past experiences with group work in the sense that people in my past have hitchhiked and taken advantage of me doing the work. They just sit there and give very minimal effort thinking that they will get credit for it. One way to overcome this is to always communicate with one another. Always be clear with each other and make sure everyone is doing their part.