

# Homework 2

Han Nguyen - TXN200004

09/16/2024

## Problem 1

a)

```
mat <- matrix(c(34, 23, 53, 6, 78, 93, 12, 41, 99), nrow = 3)
df <- as.data.frame(mat)
names(df) <- c("score_given_to_car_on_driving_test",
               "score_given_to_van_on_driving_test",
               "score_given_to_truck_on_driving_test")
```

b)

```
library(ggplot2)
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~
```

```
mpg2 <- mpg[mpg$cyl == 6, ]
mpg2$class <- as.character(mpg2$class)
```

## Problem 2

```
# Read the csv
senate <- read.csv("1976-2020-senate.csv")
head(senate)
```

```
##   year      state state_po state_fips state_cen state_ic  office district
## 1 1976  ARIZONA      AZ         4         86      61 US SENATE statewide
## 2 1976  ARIZONA      AZ         4         86      61 US SENATE statewide
## 3 1976  ARIZONA      AZ         4         86      61 US SENATE statewide
## 4 1976  ARIZONA      AZ         4         86      61 US SENATE statewide
## 5 1976  ARIZONA      AZ         4         86      61 US SENATE statewide
## 6 1976 CALIFORNIA    CA         6         93      71 US SENATE statewide
##   stage special      candidate      party_detailed writein mode
## 1   gen   False      SAM STEIGER      REPUBLICAN   False total
## 2   gen   False WM. MATHEWS FEIGHAN      INDEPENDENT   False total
## 3   gen   False    DENNIS DECONCINI      DEMOCRAT   False total
## 4   gen   False    ALLAN NORWITZ      LIBERTARIAN   False total
## 5   gen   False      BOB FIELD      INDEPENDENT   False total
## 6   gen   False    JACK MCCOY AMERICAN INDEPENDENT   False total
##   candidatevotes totalvotes unofficial version party_simplified
## 1           321236      741210      False 20210114      REPUBLICAN
## 2             1565      741210      False 20210114          OTHER
## 3          400334      741210      False 20210114      DEMOCRAT
## 4             7310      741210      False 20210114    LIBERTARIAN
## 5            10765      741210      False 20210114          OTHER
## 6            82739     7470586      False 20210114          OTHER
```

a)

```
# Convert variables to factor
senate$year <- as.factor(senate$year)
senate$state <- as.factor(senate$state)
senate$party_simplified <- as.factor(senate$party_simplified)
```

b)

```
texas_senates <- subset(senate,
                        state == "TEXAS",
                        select = c("year",
                                   "state",
                                   "candidatevotes",
                                   "totalvotes",
                                   "party_simplified"))
head(texas_senates)
```

```
##   year state candidatevotes totalvotes party_simplified
## 113 1976 TEXAS           20549     3874230          OTHER
## 114 1976 TEXAS           17355     3874230          OTHER
## 115 1976 TEXAS        1636370     3874230     REPUBLICAN
## 116 1976 TEXAS        2199956     3874230     DEMOCRAT
## 259 1978 TEXAS           4018     2312540          OTHER
## 260 1978 TEXAS        1139149     2312540     DEMOCRAT
```

c)

```
# average votes by party
parties <- unique(texas_senates$party_simplified)

avg_votes <- numeric(length(parties))
median_votes <- numeric(length(parties))

# Name for indexing
names(avg_votes) <- parties
names(median_votes) <- parties

# Iterate through each party and calculate avg, median
for (party in parties) {
  avg_votes[party] <- round(mean(
    texas_senates$candidatevotes[texas_senates$party_simplified == party],
    na.rm = TRUE
  ))
  median_votes[party] <- round(median(
    texas_senates$candidatevotes[texas_senates$party_simplified == party],
    na.rm = TRUE
  ))
}

for (p in names(avg_votes)) {
  cat(p, "-", "Average votes:", avg_votes[p], "\n")
}

## OTHER - Average votes: 21533
## REPUBLICAN - Average votes: 3019937
## DEMOCRAT - Average votes: 2416258
## LIBERTARIAN - Average votes: 92815

for (p in names(median_votes)) {
  cat(p, "-", "Median votes:", median_votes[p], "\n")
}
```

```
## OTHER - Median votes: 4564
## REPUBLICAN - Median votes: 2761660
## DEMOCRAT - Median votes: 2112490
## LIBERTARIAN - Median votes: 72657
```

d)

```
# Determine years in which DEMOCRAT candidate from TEXAS won
year_won <- texas_senates$year[
  texas_senates$party_simplified == "DEMOCRAT" &
  texas_senates$candidatevotes == ave(texas_senates$candidatevotes,
    texas_senates$year,
    FUN = max)
```

```
]
year_won <- as.character(year_won)
cat("Years that Democrat won in Texas: ", year_won)
```

```
## Years that Democrat won in Texas: 1976 1982 1988
```

## Problem 3

```
tae <- read.table("tae.data", sep = ",", header = FALSE)
names(tae) <- c("english_speaker", # 1 = English speaker, 2 = non-English
               "instructor",      # categorical (25 categories)
               "course",          # categorical (26 categories)
               "regular",         # 1 = Summer, 2 = Regular
               "class_size",      # numeric
               "score")          # 1 = Low, 2 = Medium, 3 = High
# Add TA ID based on row number
tae$TA_ID <- 1:nrow(tae)

head(tae)
```

```
##   english_speaker instructor course regular class_size score TA_ID
## 1                1         23      3        1         19     3     1
## 2                2         15      3        1         17     3     2
## 3                1         23      3        2         49     3     3
## 4                1          5      2        2         33     3     4
## 5                2          7     11        2         55     3     5
## 6                2         23      3        1         20     3     6
```

a)

```
# Turn first variable into logical variable
tae[, 1] <- tae[, 1] == 1
head(tae)
```

```
##   english_speaker instructor course regular class_size score TA_ID
## 1              TRUE         23      3        1         19     3     1
## 2             FALSE         15      3        1         17     3     2
## 3              TRUE         23      3        2         49     3     3
## 4              TRUE          5      2        2         33     3     4
## 5             FALSE          7     11        2         55     3     5
## 6             FALSE         23      3        1         20     3     6
```

b)

```
# Turn 4th variable into logical variable
tae[,4] <- tae[,4] == 2
head(tae)
```

```
##   english_speaker instructor course regular class_size score TA_ID
## 1             TRUE         23      3   FALSE         19      3      1
## 2             FALSE        15      3   FALSE         17      3      2
## 3             TRUE         23      3    TRUE         49      3      3
## 4             TRUE          5      2    TRUE         33      3      4
## 5             FALSE          7     11    TRUE         55      3      5
## 6             FALSE        23      3   FALSE         20      3      6
```

c)

```
# Turn the last variable (class attribute or evaluation score) into an ordered factor variable with levels
tae$score <- factor(tae$score,
                    levels = c(1,2,3),
                    labels = c("low", "medium", "high"),
                    ordered = TRUE)
head(tae)
```

```
##   english_speaker instructor course regular class_size score TA_ID
## 1             TRUE         23      3   FALSE         19 high      1
## 2             FALSE        15      3   FALSE         17 high      2
## 3             TRUE         23      3    TRUE         49 high      3
## 4             TRUE          5      2    TRUE         33 high      4
## 5             FALSE          7     11    TRUE         55 high      5
## 6             FALSE        23      3   FALSE         20 high      6
```

d)

```
# Average
avg_regular <- round(mean(tae$class_size[tae$regular == TRUE], na.rm = TRUE), 2)
avg_summer  <- round(mean(tae$class_size[tae$regular == FALSE], na.rm = TRUE), 2)

# Median
median_regular <- round(median(tae$class_size[tae$regular == TRUE], na.rm = TRUE), 2)
median_summer  <- round(median(tae$class_size[tae$regular == FALSE], na.rm = TRUE), 2)

# Print results with labels
cat("Regular semester - Average:", avg_regular, "Median:", median_regular, "\n")
```

```
## Regular semester - Average: 29.34 Median: 29
```

```
cat("Summer semester - Average:", avg_summer, "Median:", median_summer, "\n")
```

```
## Summer semester - Average: 19.7 Median: 20
```

e)

```

# English speakers
native_regular <- sum(tae$english_speaker & tae$regular)
native_summer  <- sum(tae$english_speaker & !tae$regular)

# Not English speakers
non_native_regular <- sum(!tae$english_speaker & tae$regular)
non_native_summer  <- sum(!tae$english_speaker & !tae$regular)

cat("Native English TAs - Regular:", native_regular, "Summer:", native_summer, "\n")

## Native English TAs - Regular: 20 Summer: 9

cat("Non-native English TAs - Regular:", non_native_regular, "Summer:", non_native_summer, "\n")

## Non-native English TAs - Regular: 108 Summer: 14

```

f)

```

total_native <- sum(tae$english_speaker)
high_native  <- sum(tae$english_speaker & tae$score == "high")
prop_native  <- round(high_native / total_native, 2)

# Non-native English TAs
total_non_native <- sum(!tae$english_speaker)
high_non_native  <- sum(!tae$english_speaker & tae$score == "high")
prop_non_native  <- round(high_non_native / total_non_native, 2)

cat("Native English TAs:", total_native, ". High score proportion:", prop_native, "\n")

## Native English TAs: 29 . High score proportion: 0.62

cat("Non-native English TAs:", total_non_native, ". High score proportion:", prop_non_native, "\n")

## Non-native English TAs: 122 . High score proportion: 0.28

```