title: “STAT3355(HW-2)” author: “Tulasi Janjanam” date: “2024-09-14” output: pdf_document: default html_document: default —

Problem 1

(a)

```
car_test_scores <- matrix(c(34, 23, 53, 6, 78, 93, 12, 41, 99), nrow = 3)

df <- as.data.frame(car_test_scores)

names(df) <- c("car_score",
               "van_score",
               "truck_score")
print(df)
```

```
##   car_score van_score truck_score
## 1        34         6          12
## 2        23        78          41
## 3        53        93          99
```

(b)

```
library(ggplot2)
head(mpg)
```

```
## # A tibble: 6 × 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…
```

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

Problem 2

(a)

```
senate_data <- read.csv("/Users/tulasijanjanam/Downloads/dataverse_files/1976-2020-senat
e.csv")

senate_data$year <- as.factor(senate_data$year)
senate_data$state <- as.factor(senate_data$state)
senate_data$party_simplified <- as.factor(senate_data$party_simplified)
#Checking
str(senate_data)
```

```
## 'data.frame':    3629 obs. of  19 variables:
##  $ year            : Factor w/ 24 levels "1976","1978",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ state           : Factor w/ 50 levels "ALABAMA","ALASKA",..: 3 3 3 3 3 5 5 5 5 5
...
##  $ state_po        : chr  "AZ" "AZ" "AZ" "AZ" ...
##  $ state_fips      : int  4 4 4 4 4 6 6 6 6 6 ...
##  $ state_cen       : int  86 86 86 86 86 93 93 93 93 93 ...
##  $ state_ic        : int  61 61 61 61 61 71 71 71 71 71 ...
##  $ office          : chr  "US SENATE" "US SENATE" "US SENATE" "US SENATE" ...
##  $ district        : chr  "statewide" "statewide" "statewide" "statewide" ...
##  $ stage           : chr  "gen" "gen" "gen" "gen" ...
##  $ special         : chr  "False" "False" "False" "False" ...
##  $ candidate       : chr  "SAM STEIGER" "WM. MATHEWS FEIGHAN" "DENNIS DECONCINI" "ALL
AN NORWITZ" ...
##  $ party_detailed  : chr  "REPUBLICAN" "INDEPENDENT" "DEMOCRAT" "LIBERTARIAN" ...
##  $ writein         : chr  "False" "False" "False" "False" ...
##  $ mode            : chr  "total" "total" "total" "total" ...
##  $ candidatevotes  : int  321236 1565 400334 7310 10765 82739 3748973 3502862 31629 1
04383 ...
##  $ totalvotes      : int  741210 741210 741210 741210 741210 7470586 7470586 7470586
7470586 7470586 ...
##  $ unofficial      : chr  "False" "False" "False" "False" ...
##  $ version         : int  20210114 20210114 20210114 20210114 20210114 20210114 20210
114 20210114 20210114 20210114 ...
##  $ party_simplified: Factor w/ 4 levels "DEMOCRAT","LIBERTARIAN",..: 4 3 1 2 3 3 4 1
3 3 ...
```

(b)

```
texas_data <- subset(senate_data, state == "TEXAS",
                     select = c("year", "state", "candidatevotes", "totalvotes", "party_
simplified"))
dim(texas_data)
```

```
## [1] 64  5
```

```
head(texas_data)
```

```
##     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)

```
dim(texas_data)
```

```
## [1] 64  5
```

```
head(texas_data)
```

```
##     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
```

```
party_votes_tx <- aggregate(candidatevotes ~ party_simplified, data = texas_data,
                            FUN = function(x) c(mean = mean(x), median = median(x)))

mean_votes <- round(party_votes_tx$candidatevotes[, "mean"])
median_votes <- round(party_votes_tx$candidatevotes[, "median"])


party_votes_tx_summary <- data.frame(
  party_simplified = party_votes_tx$party_simplified,
  avg_votes = mean_votes,
  median_votes = median_votes
)

print(party_votes_tx_summary)
```

```
##   party_simplified avg_votes median_votes
## 1         DEMOCRAT   2416258      2112490
## 2      LIBERTARIAN     92815        72657
## 3            OTHER     21533         4564
## 4       REPUBLICAN   3019937      2761660
```

(d)

```
texas_data <- subset(senate_data, state == "TEXAS" & stage == "gen",
                     select = c(year, state, candidatevotes, totalvotes, party_simplifie
d))
texas_winners <- texas_data[texas_data$party_simplified == "DEMOCRAT" &
                            texas_data$candidatevotes == max(texas_data$candidatevotes),
"year"]
print(texas_winners)
```

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

Problem 3

(Info)

```
ta_data <- read.table("/Users/tulasijanjanam/Downloads/archive/tae.data", sep = ",", hea
der = FALSE)
colnames(ta_data) <- c("eng_speaker", "instructor_id", "course_id", "regular", "size",
"score")
ta_data$ta_id <- 1:nrow(ta_data)

str(ta_data)
```

```
## 'data.frame':    151 obs. of  7 variables:
##  $ eng_speaker  : int  1 2 1 1 2 2 2 2 1 2 ...
##  $ instructor_id: int  23 15 23 5 7 23 9 10 22 15 ...
##  $ course_id    : int  3 3 3 2 11 3 5 3 3 3 ...
##  $ regular      : int  1 1 2 2 2 1 2 2 1 1 ...
##  $ size         : int  19 17 49 33 55 20 19 27 58 20 ...
##  $ score        : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ ta_id        : int  1 2 3 4 5 6 7 8 9 10 ...
```

(a)

```
ta_data$eng_speaker <- ta_data$eng_speaker == 1
```

(b)

```
ta_data$regular <- ta_data$regular == 1
```

(c)

```
ta_data$score <- factor(ta_data$score, levels = c(1, 2, 3), labels = c("low", "medium",
"high"), ordered = TRUE)
str(ta_data)
```

```
## 'data.frame':    151 obs. of  7 variables:
##  $ eng_speaker  : logi  TRUE FALSE TRUE TRUE FALSE FALSE ...
##  $ instructor_id: int  23 15 23 5 7 23 9 10 22 15 ...
##  $ course_id    : int  3 3 3 2 11 3 5 3 3 3 ...
##  $ regular      : logi  TRUE TRUE FALSE FALSE FALSE TRUE ...
##  $ size         : int  19 17 49 33 55 20 19 27 58 20 ...
##  $ score        : Ord.factor w/ 3 levels "low"<"medium"<..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ ta_id        : int  1 2 3 4 5 6 7 8 9 10 ...
```

(d)

```
# regular semesters
regular_size_avg <- mean(ta_data$size[ta_data$regular], na.rm = TRUE)
regular_size_median <- median(ta_data$size[ta_data$regular], na.rm = TRUE)

# summer semesters
summer_size_avg <- mean(ta_data$size[!ta_data$regular], na.rm = TRUE)
summer_size_median <- median(ta_data$size[!ta_data$regular], na.rm = TRUE)

# results
regular_size_avg <- round(regular_size_avg, 2)
regular_size_median <- round(regular_size_median, 2)
summer_size_avg <- round(summer_size_avg, 2)
summer_size_median <- round(summer_size_median, 2)

cat("Regular Semester - Average:", regular_size_avg, "Median:", regular_size_median,
"\n")
```

```
## Regular Semester - Average: 19.7 Median: 20
```

```
cat("Summer Semester - Average:", summer_size_avg, "Median:", summer_size_median, "\n")
```

```
## Summer Semester - Average: 29.34 Median: 29
```

(e)

```
# Native
native_regular <- sum(ta_data$eng_speaker & ta_data$regular)
native_summer <- sum(ta_data$eng_speaker & !ta_data$regular)

# Non-native
non_native_regular <- sum(!ta_data$eng_speaker & ta_data$regular)
non_native_summer <- sum(!ta_data$eng_speaker & !ta_data$regular)

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

```
## Native English Speakers - Regular: 9 Summer: 20
```

```
cat("Non-Native English Speakers - Regular:", non_native_regular, "Summer:", non_native_
summer, "\n")
```

```
## Non-Native English Speakers - Regular: 14 Summer: 108
```

(f)

```
# Native
native_total <- sum(ta_data$eng_speaker)
native_high <- sum(ta_data$eng_speaker & ta_data$class_attribute == "high")
native_high_prop <- round(native_high / native_total, 2)

# Non-native English speaker TAs and proportion who received high scores
non_native_total <- sum(!ta_data$eng_speaker)
non_native_high <- sum(!ta_data$eng_speaker & ta_data$class_attribute == "high")
non_native_high_prop <- round(non_native_high / non_native_total, 2)

cat("Total Native English Speaker TAs:", native_total, "\n")
```

```
## Total Native English Speaker TAs: 29
```

```
cat("Proportion of Native English Speaker TAs with High Scores:", native_high_prop,
"\n")
```

```
## Proportion of Native English Speaker TAs with High Scores: 0
```

```
cat("Total Non-native English Speaker TAs:", non_native_total, "\n")
```

```
## Total Non-native English Speaker TAs: 122
```

```
cat("Proportion of Non-native English Speaker TAs with High Scores:", non_native_high_pr
op, "\n")
```

```
## Proportion of Non-native English Speaker TAs with High Scores: 0
```