Homework 2
of
STAT 3355 Introduction to Data Analysis

Problem 1

Rewrite each code block to comply with the "Homework and Project Code Style Guide"

(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")
```

Answer:

(b)
```
library( ggplot2 )
  head(mpg)
second_version_of_mpg <- mpg[ mpg$cyl == 6,]
second_version_of_mpg$class <- as.character(second_version_of_mpg$class)
```

Answer:

Problem 2

Download the **U.S. Senate 1976–2020** data set on the HARVARD Dataverse. Read the data in its original format (.csv) by using the function `read.csv()` in an appropriate way. In this dataset, there are 3629 observations with 19 variables.

The variables are listed as they appear in the data file.

- year : year in which election was held
- state : state name
- state_po : U.S. postal code state abbreviation
- state_fips : State FIPS code
- state_cen : U.S. Census state code
- state_ic : ICPSR state code
- office : U.S. SENATE (constant)
- district : statewide (constant)
- stage : electoral stage where "gen" means general elections, "runoff" means runoff elections, and "pri" means primary elections.
- special : special election where "TRUE" means special elections and "FALSE" means regular elections
- candidate : name of the candidate in upper case letters
- party_detailed : party of the candidate (always entirely uppercase). Parties are as they appear in the House Clerk report. In states that allow candidates to appear on multiple party lines, separate vote totals are indicated for each party. Therefore, for analysis that involves candidate totals, it will be necessary to aggregate across all party lines within a district. For analysis that focuses on two-party vote totals, it will be necessary to account for major party candidates who receive votes under multiple party labels. Minnesota party labels are given as they appear on the Minnesota ballots. Future versions of this file will include codes for candidates who are endorsed by major parties, regardless of the party label under which they receive votes.
- party_simplified : party of the candidate (always entirely uppercase). The entries will be one of: "DEMOCRAT", "REPUBLICAN", "LIBERTARIAN", "OTHER"
- writein : vote totals associated with write-in canditates where "TRUE" means write-in canditates and "FALSE" means non-write in canditates.
- mode : mode of voting; states with data that doesn't break down returns by mode are marked as "total"
- canditatevotes : votes received by this candidate for this particular party

- totalvotes : total number of votes cast for this election
- unofficial : TRUE/FALSE indicator for unofficial result (to be updated later); this appears only for 2018 data in some cases
- version : date when this dataset was finalized

(a) Turn the variables : year, state, and party_simplified into factor variables.

Answer:

(b) Subset the dataset by extracting the data for the state of Texas. Only keep the columns: year, state, candidatevotes, totalvotes, and party_simplified. **Use this data subset for the rest of the question**

Answer:

(c) Calculate the average and median number of votes received by democratic, republican, libertarian, and other candidates in the state of Texas. Round your numeric answer to the nearest whole number.

Answer:

(d) Identify the years in which the democratic candidate from Texas won.

Answer:

Problem 3

Download the "Teaching Assistant Evaluation Data Set" dataset on the UCI Machine Learning Repository. The link is https://www.kaggle.com/code/johnmantios/teaching-assistant-evaluation/input. You will need to create an account to download it. Read the data in its original format (.data) by using the function `read.table()` or `read.csv()` in an appropriate way and rename each variable according to the web page.

In this dataset, each of the 151 observation corresponds to a unique teaching assistant (TA), so create a variable of TA identification (ID) number and assign an ID number from 1 to 151 to all TAs sequentially. In addition, for simplicity (although it may not be true in this case), assume each class can only have one TA at a time. Therefore, each of the 151 observation corresponds to a unique course at a time. If you see multiple observations share the same instructor ID and course ID, that probably means that the courses occurred in different year or semester.

Answer:

(a) Turn the first variable (whether of not the TA is a native English speaker) into a logical variable, where `TRUE` corresponds to a native English speaker, and `FALSE` otherwise

Answer:

(b) Turn the fourth variable (summer or regular semester) into a logical variable, where `TRUE` corresponds to regular, and `FALSE` corresponds to summer

Answer:

(c) Turn the last variable (class attribute or evaluation score) into an ordered factor variable with levels labeled as 'low', 'medium' and 'high'

Hint: You should get the following response (other than variable names) after applying the `str()` function on the cleaned dataset

```
'data.frame':   151 obs. of  6 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  FALSE FALSE TRUE TRUE TRUE FALSE ...
$ size         : int  19 17 49 33 55 20 19 27 58 20 ...
$ score        : 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 ...
```

Answer:

(d) What is the average and median class size in regular semester? What are those two numbers in summer semester? Round your numeric answer to 2 decimal places.

Answer:

(e) How many native English speaker TAs are there in regular and summer semester, respectively? What are those two numbers for non native English speaker TAs?

Answer:

(f) How many native English speaker TAs in this data? What's the proportion of them received high scores? How many non native English speaker TAs in this data? What's

the proportion of them received high scores ? Round your numeric answer to 2 decimal places.

Answer: