DS 413/613 MIDTERM EXAM SUMMER 2022

Instructions: Generate an Rmardown file and a Word file that shows all required code and output.

**1) U** <- c(“Maine” , “Texas”, “Delaware”, “Oregon”, “Utah”, “Vermont”, “Ohio”)

a) Is **U** an atomic vector or a list ?

b) Use and show R code that will extract the elements “Maine” and “Vermont”.

c) Use and show R code that will extract all elements except “Texas”.

d) Use and show R code that will produce length of **U.**

**2) V** = list(“Chicago”, k = list( 2, 6, 18, 24), FALSE, 13, 1.3, y = 1:10)

a) Is **V** an atomic vector or a list ?

b) Use and show R code that will extract the 5th element of **V**.

c) If the vector **V** is a list, use and show R code to identify the type of each object in **V.**

**3)** Copy paste and run the tribble given below.

tribble( ~John, ~Raymond, ~Martha, ~Alice, ~Juan,

86, 77, 81, 88, 90,,

79, 78, 85, 81, 78,

76, 75, 88, 94, 81,

84, 90, 71, 84, 89,

100, 80, 93, 85, 84,

90, 73, 70, 88, 93,

) -> TestScores

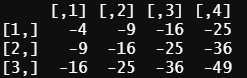
TestScores

a) Use and show R code (a map function) to find the median for each column.

b) Use and show R code (a map function) to find the cube root of each column element.

c) Use and show R code (a map function) to convert each column value to 0.

**4)** Use and show R code, as demonstrated in class to produce the following matrix,



**5)**

a) Show and use a census API key that gives you access to the **ACS** data. Do not use my API key, use and show your own key.

b) Using **ACS** census data from 2015, show and use R code to do the following to produce a tibble that shows the median income estimates and the margin of errors for white males ages 35 - 44 in the counties of California. The required variable code starts with the characters BO1OO1. Use the table to find the other characters.

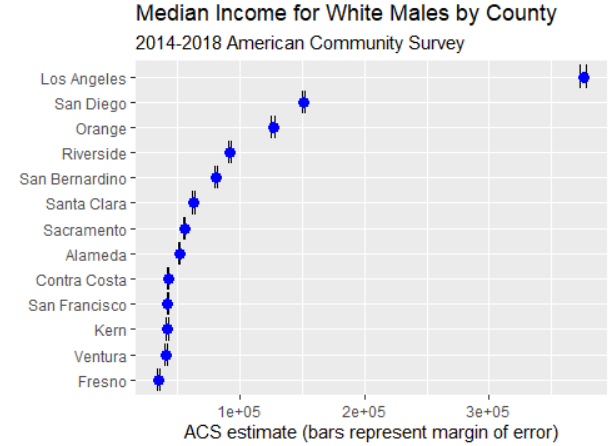
The first five rows of your data table are provided below:



c) Use dplyr functions to change your table of part a so that it reflects estimates that are greater than $30,000 dollars and list the estimates in descending order.

d) Use and show ggplot R coding to produce a scatter plot that features x = natural log of **moe** plotted against y = natural log of **estimate.** Does your plot suggest a linear relationship between the varibles ? If so, what general trend can be inferred? (Use the full data table that you generated for part b)

e) Use and show R code that will produce the following graph for the data generated in part c



**6)** Provided below is the famous poem **“ Stopping by the Wood On a Snowy Evening” by Robert Frost.**

Use the text mining sequence of steps and the R code modeled in class to **a)** create a tibble, **b)** find line locations of words, **c)** produce a word frequency table, and **d)** and create a bar graph data visualization plot that will also display word frequency trends.

Hint( do not forget to process a single spaced body of text ; be careful about commas and double quotation marks. Use the examples demonstrated in class.)

Whose woods these are I think I know.

His house is in the village though;

He will not see me stopping here

To watch his woods fill up with snow.

My little horse must think it queer

To stop without a farmhouse near

Between the woods and frozen lake

The darkest evening of the year.

He gives his harness bells a shake

To ask if there is some mistake.

The only other sound’s the sweep

Of easy wind and downy flake.

The woods are lovely, dark and deep,

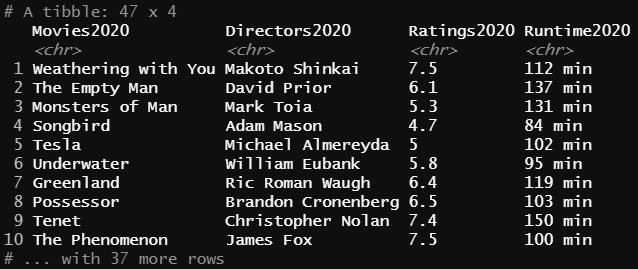
But I have promises to keep,

And miles to go before I sleep,

And miles to go before I sleep.

**7)** Now using the same body of text found in **Problem 6,** use and show R code to create a word cloud. You can use the coding and methods that were illustrated in class or you can use alternate coding of your choice to create the word cloud.

**8**) Go to the link <https://www.imdb.com/list/ls096735829>and use Selector Gadget , as demonstrated in class to collect data specific to the movie names, the directors of the movies, the movie ratings, and the running times of the movies. Your code should produce the final resulting table given below. Hint: (you may have to first create a data frame, and then convert the data frame to the tibble shown below.



The following problem is for Graduate students only (Use the *diamonds* data table)

**9)**

a) Use and show R code that shows both column variables x and y, of the diamonds data table contain the value 4.93. How many times does the number 4.93 appear in each column ?

b) Use and show R code that shows neither column variable x or y, contain the value 3.62.

c) Now show and use R code to find all values that the column variables x and y have in common.