DATA 412/612 /FINAL EXAM/FALL 2021

**Instructions**: Do not help or accept help from anyone as you work on and complete the exam. You are at liberty to use your book, past graded exams, or any online resource that you feel will be helpful. Send all r code and associated output as an r markdown file and as a knitted word file document.

library(tidyverse) library(stringr) library(dplyr)

library(ggplot2) library(tidyr) library(reshape2)

library(readr) library(forcats)

**1)**

**Using appropriate R code, read in the emailed excel spread sheet**

***Collegedata. Use read\_csv(“Collegedata.csv”)***

The data in Collegedata contains a subset of the variables found in the

2016 to 2017 [College Scorecard](https://collegescorecard.ed.gov/data/)

database. These data contain information on colleges in the United States.

The variables included are:

- `UNITID` and `OPEID`: Identifiers for the colleges.

- `INSTNM`: Institution name

- `ADM\_RATE`: The Admission Rate.

- `SAT\_AVE`: Average SAT equivalent score of students admitted.

- `UGDS`: Enrollment of undergraduate certificate/degree-seeking students

- `COSTT4\_A`: Average cost of attendance (academic year institutions)

- `AVGFACSAL`: Average faculty salary

- `GRAD\_DEBT\_MDN`: The median debt for students who have completed

- `AGE\_ENTRY`: Average age of entry

- `ICLEVEL`: Level of institution (1 = 4-year, 2 = 2-year, 3 = less than 2-year).

- `MN\_EARN\_WNE\_P6`: Mean earnings of students working and not enrolled 6 years

after entry (so students who graduated in the 2009 to 2010 academic year)

**2)**

Given the level of the institution, does there appear to be an association between the average cost of attendance**(x variable**) and the mean earnings of students six years after graduation**(y variable)**? Make appropriate plots to justify your response. (Hint: Generate two plots to make your decision, first a standard scatter plot involving the two continuous variables mentioned and then a facet plot over the appropriate categorical variable)

**3)**

**Use r code to produce a histogram of the average age of entry. Comment on the distribution of this variable.**

**4)**

**Use r code that will produce output that shows the 10 institutions that have the highest average age of entry?**

**5)**

**There are many universities with "American University" in the name. E.g.**

**"American University of Puerto Rico" and "National American**

**University-Ellsworth AFB Extension". Use R code to create a data frame, called**

**`americandf`, that contains just the data from universities with**

**"American University" in the name.**

**6)**

**Provide r code that will produce the number of colleges from the College Score data frame that have average SAT scores that are above 1000. ( Do not produce the data frame. Your code should only yield the number)**

**7)**

**Provide r code that will show a data frame that lists the 10 highest Average SAT scores in decreasing order. A partial data frame is given below.**

UNITID OPEID MN\_EARN\_WNE\_P6 INSTNM SAT\_AVG ADM\_RATE UGDS COSTT4\_A

*<dbl>* *<dbl>* *<chr>* *<chr>* *<dbl>* *<dbl>* *<dbl>* *<dbl>*

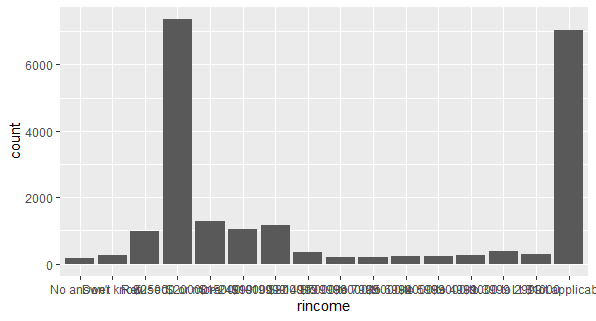
1 110404 113100 59800 Calif~ 1555 0.0807 979 63471

2 166683 217800 107300 Massa~ 1519 0.0794 4489 63250

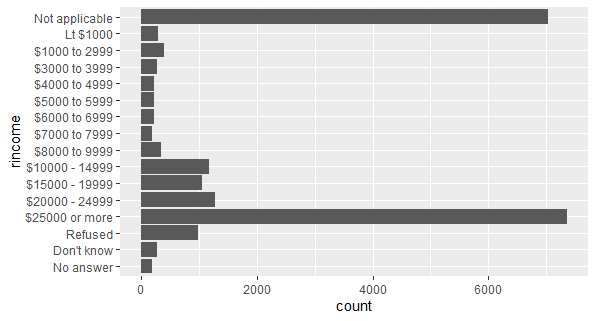
3 144050 177400 73700 Unive~ 1508 0.0794 5978 70100

4 166027 215500 102200 Harva~ 1506 0.054 7447 64400

**8) Using the gss\_cat data frame, write r code that will produce the bar graph below. And explain in one or two sentences why the bar graph is difficult to interpret.**



**9a)** Now write r code from the same data set that produce the transformed bar graph and comment on why it is an improvement



**9b)**

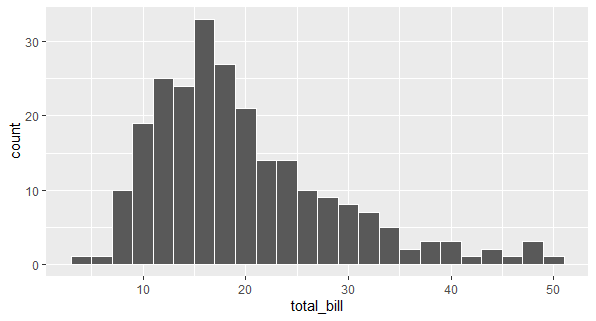
**Use r code to produce the *tips* data frame from the reshape2 package. Name three categorical variables in the data frame.**

**10)**

**Use r code to indicate how many levels exist for the factor *day* in the tips data frame and determine the frequency of each level.**

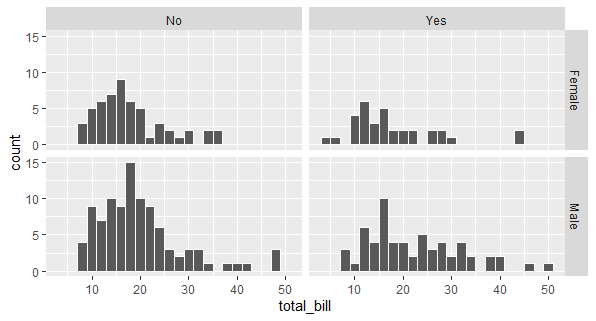
**11)**

**Produce r code that will produce the following histogram from the *tips* data frame**



**12)**

**Write r code that will produce the following histograms from the *tips* data frame**



**13)**

Using the **stringr::words** data set along with **str\_subset** code, produce Rcode that will show a 9 letter word that has the letter **a** in themiddle.

**14)**

Produce a string that will force a match for the regular expression **\\””\**  Use and show the R code command **writelines** to confirm your answer

**15)**

Describe in words (two or three sentences) what the following regular expression will match

^.\*e$

**16)**

Using the methods demonstrated in class regarding Factors and Forcats, use and show R code to create a **factor** that will enable you to sort the string vector **(“eight”, “four”, “ten”, “two”)** according to quantity, not alphabetical order.

**17)**

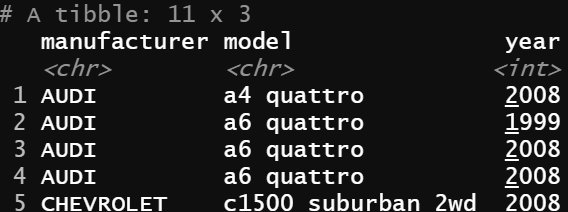
Using the mpg data set, use and show R code that will produce a table that shows the average city mileage (mean for **cty**) for each manufacturer.

**18)**

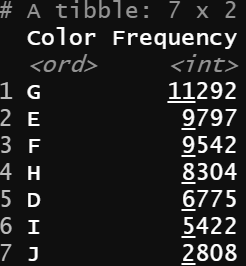
Using the mpg data set and methods and code illustrated in the emailed Factors and Forcats R file: use and show R code to generate a scatter plot that clearly shows how different manufacturers compare with one another with regards to average or mean city mileage (cty)

**Graduate student problems**

**19**) To the mpg data table, apply a stringr function to print all observations of the manufacturer variable in upper case letters. Then use R coding to produce rows 15 to 25. The first five rows of the table are shown below. Note that the row numbers 1 – 5 correspond to rows 15 – 20.



**20)** To the diamonds data table, apply R code to produce the table give below. Note that the variable name color has been changed to Color.



**21)** Now using the table produced in problem 20 and ggplot coding, produce the bar graph shown below.

