Homework 1

Due: 9/4/2024

Classmates/other resources consulted: [type answer here]

Question 1 (8 points)

> **Successfully knitting this file and turning it into a PDF, however you do that, is worth 8 points.**

Question 2 (12 points)

The answers to the following questions can be found in the syllabus.

> a. **What do you need to do before every class?**

Watch the VidGrid video and answer the questions.

> b. **What needs to go at the top of every homework assignment?**

List names of other students worked with and any resources used.

> c. **Should you put your name at the top of every homework assignment?**

No.

> d. **What happens to your lowest homework grade?**

Lowest grade dropped.

> e. **Can you work with other students on the homework assignments?**

Yes, but cannot submit some one elses code as ones own.

> f. **Is letting another student look at your completed homework assignment an example of academic dishonesty?**

Yes, especially if not credited or if code is reused.

g. **Can you use AI, such as ChatGPT, to help you answer a homework question that asks you to explain a concept we learned ``in your own words?“**

No. You must use your own words. ChatGPT can be used for research and explanations, however.

h. **Can you use AI, such as ChatGPT, to help you find bugs in your code?**

Yes.

i. **If you are taking an extension on a homework assignment that is less than 48 hours, and you have not yet reached 144 hours of homework extensions used, do you need to contact Prof. Cannon about it?**

No.

j. **What is the QCL, and what services do they offer?**

The Quantitative and Computing Lab (QCL) is a CMC resource for one on one peer tutoring regarding all courses with a quantitative component.

k. **Per CMC policy, how many hours per week are you expected to be spending on this course?**

12 hours, including time spent in class.

l. **Are you allowed to share any course materials, such as class recordings, preclass videos, homework assignments, and homework solutions, with students who are not enrolled in this class this semester?**

Instructor permission required.

Question 3 (16 points)

The Washington Post maintains a public database of all people in the United States who have been fatally shot by police since 2015. Their public-facing website is https://www.washingtonpost.com/graphics/investigations/police-shootings-database/ (https://www.washingtonpost.com/graphics/investigations/police-shootings-database/), and information about what's in this data and how it's determined is at https://github.com/washingtonpost/data-police-shootings/tree/master/v1 (https://github.com/washingtonpost/data-police-shootings/tree/master/v1). Take a look at what's in this data (what are the columns? how are the values in the columns determined?), and answer all the questions below in your own words.

a. **(4 points) What are some ethical benefits that could result from creating and maintaining this publicly-available database?**

There are several ethical benefits that could come from a database like this. It could lead to increased police accountability and informed public discourse because this information is out in the open. Academic research on the issue of police shootings could also be aided by this.

b. **(4 points) What are some ethical harms that could result from creating and maintaining this publicly-available database?**

The creation of a public database like this one on police shooting deaths also has some potential ethical harms. The largest clear potential harm is regarding victims privacy. Some families of victims may wish to avoid the increased attention that this database could enable. There is also the issue of data misuse or stigmatization, where the data could be misconstrued to reinforce harmful stereotypes.

c. **(4 points) Which ethical challenges (from Section 2 of the data ethics reading) are most relevant to consider when creating, maintaining, or working with this publicly-available data set?**

When working with this dataset several ethical challenges are relevant. Ethical challenges in data collection and use are important because one must ensure that the data accurately represents incidents without bias, respecting the privacy and dignity of individuals involved. Potential harms to fairness and justice should also be considered, as the dataset could influence public perception and policy decisions.

d. **(4 points) In your opinion, do the ethical benefits of creating and maintaining this publicly-available database outweigh the ethical harms? Explain the reasoning behind your opinion. (Because this is asking for your opinion, answers may vary and that's fine)**

I believe the benefits of accountability and informed discussion, outweigh the harms I outline above. With safeguards to protect privacy and data integrity, the database can effectively drive positive change.

Question 4 (4 points)

There are many available resources highlighting data ethics. Read the Center for Democracy and Technology's Statement on AI and Machine Learning at https://cdt.org/ai-machine-learning/ (https://cdt.org/ai-machine-learning/). What principles from our data ethics reading ("Intro to Data Ethics" pages 2-26) are echoed in this statement? What new things did you learn from this reading that were not covered by our previous data reading?

Both resources share key principles like fairness, accountability, and transparency. Both emphasize the need to design automated systems that avoid bias and ensure decisions are fair and transparent.

I think the Center for Democracy and Technology's Statement statement adds new insights about challenges like the difficulty in explaining complex algorithms and the importance of ensuring systems are reliable. It talks about making sure systems work as expected and monitoring them for issues, a practical way to maintain ethical responsibility while still allowing for innovation.

Question 5 (10 points)

a. **(2 points) Write R code that creates a variable var1 that has the value 18; a variable var2 that has the value 5; and a variable var3 that has the value 9. Be sure to use the correct assignment operator (not =).**

```
var1 <- 18
var2 <- 5
var3 <- 9
```

b. **(2 points) Write code to combine the values 45, 25, -15, 3.9, 0.46, 401, and -4.8 into one list, and store this list in a variable (with a name of your choosing).**

```
c_result <- c(45, 25, -15, 3.9, 0.46, 401, -4.8)
```

c. **(2 points) For the variable you created in part (b), change the value -15 to be 7 instead. Note you shouldn't create a new list, you should just modify the list you already created.**

```
c_result[3] <- 7
```

d. **(2 points) For the variable you created in part (b) and modified in part (c), find the average of all values and the median of all values. Which do you think does a better job of summarizing the seven values?**

```
c_mean <- mean(c_result)
c_median <- median(c_result)
```

Mean better at summarizing. Median doesn't use information of significance of values.

e. **(2 points) Give two different ways of computing the sum of variables var1, var2, and var3 from part (a). The numbers 18, 5, and 9 should not appear in your code, use var1, var2, and var3 instead.**

```
var_sum <- sum(var1, var2, var3)
```

```
var_list <- c(var1, var2, var3)
var_list_sum <- sum(var_list)
```

Question 6 (5 points)

Suppose you have the variables a and b given as in the following code chunk.

```
a <- c(7, 0, 0, 6)
b <- c(2, 4, 0, 3)
```

a. **(2 points) Write a code chunk that computes the product of a and b. Explain in your own words why you get the answer that you do.**

```
c <- a * b
```

We are finding element wise product of the vectors a and b, so our result c is a vector.

b. **(3 points) What happens when you compute mean(a/b)? Explain in your own words both what happens, and why this happens.**

```
mean(a/b)
```

```
## [1] NaN
```

This attempts to divide each element in a by corresponding element in b. Third element in b is 0, which causes div by zero yielding NaN not a number.

Question 7 (12 points)

Consider the R object mtcars.

a. **(2 points) What data does this object contains, and where does that data come from?**

This object contains an example data set of Motor Trend Car Road Test from 1974. It came from Motor Trend US magazine.

b. **(2 points) Output this data set. Make sure only the first ten rows appear in your knitted file.**

```
head(mtcars, 10)
```

```
##                    mpg cyl  disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
## Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
## Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
## Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
## Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
```

c. **(2 points) Output only the hp column from this data set. Make sure only the first ten rows appear in your knitted file.**

```
head(mtcars$hp, 10)
```

```
##  [1] 110 110  93 110 175 105 245  62  95 123
```

d. **(2 points) What is the smallest hp of all cars in the data set? Use an R command that outputs this value, rather than looking through the data set yourself to find it.**

```
min(mtcars$hp)
```

```
## [1] 52
```

e. **(2 points) Output only the ninth row of this data set.**

```
mtcars[9,]
```

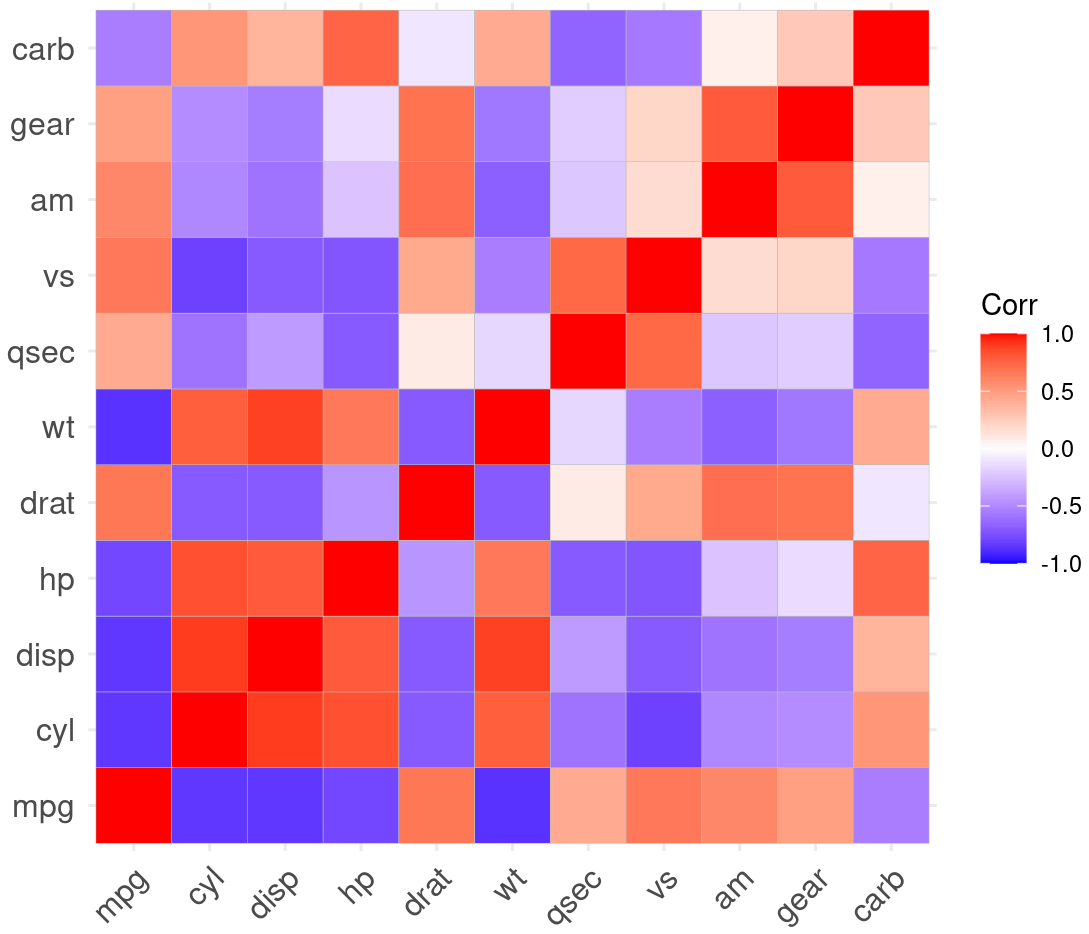
```
##           mpg cyl  disp hp drat   wt qsec vs am gear carb
## Merc 230 22.8   4 140.8 95 3.92 3.15 22.9  1  0    4    2
```

f. **(2 points) Suppose you want to make a correlation plot, and you know it is part of the ggcorrplot package. Install and load this package (you do not need to include the installation command in your knitted file). You will know it works if the following command produces a plot (remove the # in front of this command before running it):**

```
# install.packages("ggcorrplot")
library(ggcorrplot)
```

```
## Loading required package: ggplot2
```

```
ggcorrplot(cor(mtcars))
```



Question 8 (10 points)

Call this function on the following input: difference(3,7), difference(b = 5), difference(4), difference(b = 8, a = 5, neg = TRUE), and difference(7, neg = TRUE). For each, explain in your own words in 1-2 sentences why you get the answer that you do.

```
difference <- function(a = 0, b = 1, neg = FALSE){
  s <- a - b
  if(neg == FALSE)
      return(s)
    else
      return(-s)
}
```

a. **difference(3,7)**

```
difference(3,7)
```

```
## [1] -4
```

You get -4 because inputs are passed in as a, b and a - b = -4. Not inverted because neg == FALSE.

b. **difference(b = 5)**

```
difference(b = 5)
```

```
## [1] -5
```

You get -5 because a is default 0 and b is specified as 5 in input and 0 - 5 = -5.

c. **difference(4)**

```
difference(4)
```

```
## [1] 3
```

You get 3 because a is set to 4 by input and b default is 1 and 4 - 1 = 3.

d. **difference(b = 8, a = 5 neg = TRUE)**

```
difference(b = 8, a = 5, neg = TRUE)
```

```
## [1] 3
```

You get 3 because 5 - 8 = -3 and neg = TRUE, which inverts the result of -3 to 3.

e. **difference(7, neg = TRUE)**

```
difference(7, neg = TRUE)
```

```
## [1] -6
```

You get -6 because b default is 1 and 7 - 1 = 6. NEG = TRUE gets negative of 6, yielding -6.

Question 9 (7 points)

a. (3 points) **The median of 4, 11, 7, 9, and 5 is 7 (the middle number, when they are put in order from smallest to largest). Explain why the following command doesn't correctly compute the median, and give a command that uses the median()function to correctly compute the median of these five numbers.**

```
median(4, 11, 7, 9, 5 )
```

```
## [1] 4
```

You need to input the object as a vector / list, otherwise it just evaluates the first object. Like so:

```
median(c(4, 11, 7, 9, 5))
```

```
## [1] 7
```

b. (4 points) **Consider the following two commands. Explain in your own words why you get the answer you do for each of them.**

```
max(c(5,9,2), c(3, 11, 1))
```

```
## [1] 11
```

```
pmax(c(5,9,2), c(3, 11, 1))
```

```
## [1]  5 11  2
```

The max command finds max value across both lists. The pmax or piece wise max command findings the maximum for each corresponding pair in the list.

Question 10 (13 points)

> a. **(2 points) Display your favorite math equation below. Be sure to use math mode so that it is displayed nicely.**

$e^{i\pi} + 1 = 0$

> b. **(2 points) How do you make a greater than or equal sign show in your knitted file? Search online to find the answer; a useful link is https://detexify.kirelabs.org/classify.html (https://detexify.kirelabs.org/classify.html)**

$\geq$

> c. **(9 points) Find the html file HW_example.html attached to this assignment in Canvas. Write RMarkdown code to replicate this html document (minus the title, author, and date) below.**

(Note: If you are knitting to PDF, you can also find HW_example.pdf attached to this assignment in Canvas)

(Note: On certain systems, the list is displayed with white circles rather than black circles; that's perfectly fine!)

Section: LaTeX

In an obtuse triangle, $a^2 + b^2 \geq c^2$. The formula for the *area of a trapezoid* is

$$A = \frac{b_1 + b_2}{2}h.$$

Subsection: lists

The five **Claremont Colleges** are:

- CMC
- HMC
- Scripps
- Pitzer
- Pomona

Question 11 (3 points)

When you were stuck on a question on this assignment, what did you do? Were the things you tried useful/productive? Did you follow any of the troubleshooting/debugging hints given in class on 8/28? What would you do the same or different if you get stuck on a homework question in the future?

I was stuck with the writing portion. Just kept pushing through it. Also referenced 8-28 notes.