STAT 408: Midterm Exam
Name:

Please turn in the exam to D2L and include the R Markdown code, a Word or PDF file with output. You are welcome to turn in code and output for each question separately. Please verify that all of the code has compiled and the graphics look like you think they should on your Word or PDF file, you are welcome to upload image files directly if they look distorted in the Word or PDF file.

STAT 408 exams have traditionally had both an in class and a take home portion. This exam will be 100% take home and you are welcome to class notes, code written during class, or other online resources , provided you include an acknowledge of others work (code). **NOTE: you may not discuss the exam with classmates. Direct all questions to the instructor.** The instructor will answer questions related to the data, expectations, and understanding of the exam, but will not fix broken code.

Part 1:

Part 1 are a set of questions that would tend to be given as part of an in class exam. For each question, write a 2 or 3 sentence explanation.

1. (2 points)

Detail at least two principles of good data visualization.

2. (2 points)

Describe the difference between `left_join()`, `inner_join()`, and `full_join()`.

3. (3 points)

Below is a snippet of code from a Shiny app, add comments to each line describing what that line of code accomplishes.

```
ui <- fluidPage(
        mainPanel(
            numericInput(inputId = "test_grade",
                         label = "Estimated Test Score:",
                         min = 1,
                         max = 100,
                         value = 90)
                )
            )
```

Part 2:

This section focuses on data processing and visualization using COVID-19 related tweets.

```
tweets <- read_csv('https://raw.githubusercontent.com/stat408/midterm_f2020/master/Covid_tweets.csv')
```

The tweets dataset contains 4 features:

- Location: User specified location
- TweetAt: Date of tweet, "day - month - year"
- OriginalTweet: text of the tweet
- Sentiment: classified sentiment of the tweet. There are 5 possible values:
  - Extremely Positive
  - Positive
  - Neutral
  - Negative
  - Extremely Negative

1. Data wrangling (20 points)

This question will focus on data manipulation using the covid tweet dataset, which can be accessed using the link above. For full credit, include your code and create neat output by using options like `kable()` for tables and writing results in line with r commands.

a. (2 points)

Print a table that contains the 5 days with the most tweets.

b. (3 points)

Print a table that contains the 5 days with the most tweets, but now sort the rows in chronological order so that the earliest day shows up first and the latest day shows up on the last line.

c. (3 points)

Answer the following questions using inline R tools to summarize your answers. What day of the week has the most tweets? How many tweets occur on that day?

d. (4 points)

How many tweets are geotagged to occur in Montana? Print out a table with the tagged locations. You only need to include those that explicitly state they are in `Montana`.

e. (2 points)

What proportion of the tweets in the dataset are classified to have extremely positive sentiment? Answer in a sentence.

f. (4 points)

For the tweets in the dataset that reference (`#toiletpaper` `#TP` or `#tp`). Create a table that contains the total number that fall in each sentiment category. Sort the table from most extremely positive (first row) to extremely negative (last row).

2. Data Visualization (18 points)

Write a set of captions for each figure. The captions should be 2 - 3 sentences and fully describe the figures.

a. (6 points)

Create a graphic that explores how sentiment changes as a function of day of the week.

b. (6 points)

Create a graphic that explores how sentiment changes over time during the time period covered in the tweets.

c. (6 points)

Make a graphic that explores some other feature of the data. Be creative!