Exam 2, STAT 450

Due: Friday, November 8

Directions:

- This exam should be completed using Quarto and submitted to Canvas as a self-contained HTML or PDF file.
- Your solutions to this exam must be your own work.
- Make sure your Quarto document is well-formatted: label each exercise with a header, use separate code chunks for your answers to each exercise, and any written analysis should be formatted as plain text outside of the code chunks. Points may be deducted for poor formatting.

First, load the following R packages:

```
library(tidyverse)
library(nycflights13)
```

Question 1

Use `read_csv()` to read the data sets `county.csv` and `votes.csv` into R. Both data sets can be downloaded from Canvas.

The `county` data contains demographic information for each US county. Variable descriptions:

- `fips`: unique identifier for counties
- `county`: name of county
- `state`: state abbreviation
- `pop2014`: population estimate, 2014
- `pct_bachelors`: Bachelor's degree or higher, percent of persons age 25+

The `votes` data contains information on voting outcomes for the 2016 presidential election. Variable descriptions:

- `fips`: unique identifier for counties
- `votes_clinton`: number of votes for Hillary Clinton
- `votes_trump`: number of votes for Donald Trump
- `total_votes`: total number of votes

a

Confirm that `fips` is a unique identifier for the rows in the `county` data frame.

b

Use `inner_join()` to combine the `county` and `votes` data frames, using `fips` as the key. Call the resulting, joined data frame `county_votes`.

c

Use `mutate()` to add a new column to the `county_votes` data frame called `pct_clinton`, which is defined as the number of votes for Hillary Clinton divided by the total number of votes, and then multiplied by 100:

```
pct_clinton = 100 * votes_clinton / total_votes
```

d

Use `ggplot()` to make a scatter plot with `pct_bachelors` on the x-axis and `pct_clinton` on the y-axis. Use `geom_smooth()` to add a smooth trend line. Describe the relationship between the variables in the scatter plot.

e

Use `filter()` to subset the rows of `county_votes` corresponding to counties that are in California (CA). Which CA counties had the highest percentage of votes for Clinton? Which CA counties had the lowest percentage of votes for Clinton?

Question 2

Use `group_by()` and `summarize()` to compute the mean departure delay for each origin airport in `flights`. Then join this table of grouped summaries with the `airports` table. To improve presentation relocate the airport `name` to the first column, and arrange the rows according to the average delay. This is what the resulting, joined table should look like:

```
# A tibble: 3 x 10
  name          origin  count dep_delay_mean   lat   lon   alt    tz dst   tzone
  <chr>         <chr>   <int>          <dbl> <dbl> <dbl> <dbl> <dbl> <chr> <chr>
1 La Guardia    LGA    104662           10.3  40.8 -73.9    22    -5 A     Amer~
2 John F Kenne~ JFK    111279           12.1  40.6 -73.8    13    -5 A     Amer~
3 Newark Liber~ EWR    120835           15.1  40.7 -74.2    18    -5 A     Amer~
```

Question 3

This question uses the `weather` data frame from the `nycflights13` package.

```
glimpse(weather)
```

a

Use a `for` loop to count the number of `NA` values in each column of `weather` (Hint: use the `is.na()` function in your code). The output of your code should look like this:

```
    origin       year      month        day       hour       temp       dewp
         0          0          0          0          0          1          1
     humid   wind_dir wind_speed  wind_gust     precip   pressure      visib
         1        460          4      20778          0       2729          0
 time_hour
         0
```

b

Repeat part **a**, but this time use the `apply()` function instead of a `for` loop.