Midterm, STAT 650, Fall 2024

Due: Friday, September 27

Directions:

- This exam should be completed using Quarto and submitted to Canvas as a self-contained HTML or PDF file.
- It is expected that your submission represents your own work and understanding of the material. Do not copy another student's code or written analysis.
- 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(mdsr)
library(nycflights13)
library(NHANES)
```

Exercise 1 (25 points)

For this exercise use the `CIACountries` data set from the `mdsr` package. This package has many of the data sets referenced in our textbook Modern Data Science with R.

The `CIACountries` data set contains seven variables collected for each of 236 countries: population (`pop`), area in sq km (`area`), gross domestic product (`gdp`), percentage of GDP spent on education (`educ`), length of roadways per unit area (`roadways`), internet use as a fraction of the population (`net_users`), and the number of barrels of oil produced per day (`oil_prod`).

```
head(CIACountries)
```

```
         country      pop    area oil_prod   gdp educ   roadways net_users
1    Afghanistan 32564342  652230        0  1900   NA 0.06462444       >5%
2        Albania  3029278   28748    20510 11900  3.3 0.62613051      >35%
3        Algeria 39542166 2381741  1420000 14500  4.3 0.04771929      >15%
4 American Samoa    54343     199        0 13000   NA 1.21105528      <NA>
5        Andorra    85580     468       NA 37200   NA 0.68376068      >60%
6         Angola 19625353 1246700  1742000  7300  3.5 0.04125211      >15%
```

```
dim(CIACountries)
```

```
[1] 236   8
```

Use `ggplot2` to make the following visualizations:

(a) Bar plot of the categorical variable `net_users`.

(b) Histogram of `log10(gpd)`, set the argument `binwidth = 0.2`.

(c) Scatter plot with `log10(pop)` on the x-axis and `log10(gdp)` on the y-axis. Color the points according to `net_users`.

(d) Side-by-side box plots showing the relationship between `log10(gdp)` and `net_users`.

(e) Density plots of `log10(gdp)` for each category of `net_users` (use `facet_wrap()`).

Exercise 2 (25 points)

For this exercise use the `NHANES` data set (see lecture 5).

(a) Make a stacked bar plot with the categorical variables `MaritalStatus` and `HomeOwn`. Map the categories of `HomeOwn` to the fill color of the bars.

(b) Repeat part (a) but display proportions instead of counts. Write 2-3 sentences with your interpretation of this plot.

(c) Make a table with the count and average age for each category of `MaritalStatus`. Arrange the rows of this table according to the average age.

Exercises 3 and 4 use the `flights` data frame from the `nycflights13` package. For a description of the relational database see Chapter 19 of R for Data Science.

Exercise 3 (20 points)

Use `filter()` to find all flights that:

(a) Flew to San Francisco International Airport (SFO).

(b) Departed in summer (July, August, September).

(c) Were operated by United Airlines, and had departure delays that were 10 or more minutes.

(d) Arrived more than two hours late, but did not have a late departure.

Exercise 4 (30 points)

(a) Use `group_by()` and `summarize()` to create a data frame with the following columns:

- Count of the number of flights to each destination.
- Mean arrival delay for each destination.
- Standard deviation of arrival delays for each destination.

Your R code should recreate the following table:

```
# A tibble: 105 x 4
   dest  count arr_delay_mean arr_delay_sd
   <chr> <int>          <dbl>        <dbl>
 1 ABQ     254           4.38         42.0
 2 ACK     265           4.85         30.0
 3 ALB     439          14.4          50.5
 4 ANC       8          -2.5          26.4
 5 ATL   17215          11.3          47.0
 6 AUS    2439           6.02         43.5
 7 AVL     275           8.00         33.6
 8 BDL     443           7.05         42.1
 9 BGR     375           8.03         46.4
10 BHM     297          16.9          56.2
# i 95 more rows
```

(b) Use `left_join()` to combine the data frame of grouped summary statistics from part (a) with the `airports` data frame. The combined data frame should contain additional columns with information about the *destination* airport. Your code should should recreate the following table:

```
# A tibble: 105 x 11
   dest  count arr_delay_mean arr_delay_sd name     lat    lon   alt    tz dst
   <chr> <int>          <dbl>        <dbl> <chr>  <dbl>  <dbl> <dbl> <dbl> <chr>
 1 ABQ     254           4.38         42.0 Albuq~  35.0 -107.   5355    -7 A
 2 ACK     265           4.85         30.0 Nantu~  41.3  -70.1    48    -5 A
 3 ALB     439          14.4          50.5 Alban~  42.7  -73.8   285    -5 A
 4 ANC       8          -2.5          26.4 Ted S~  61.2 -150.    152    -9 A
 5 ATL   17215          11.3          47.0 Harts~  33.6  -84.4  1026    -5 A
 6 AUS    2439           6.02         43.5 Austi~  30.2  -97.7   542    -6 A
 7 AVL     275           8.00         33.6 Ashev~  35.4  -82.5  2165    -5 A
 8 BDL     443           7.05         42.1 Bradl~  41.9  -72.7   173    -5 A
 9 BGR     375           8.03         46.4 Bango~  44.8  -68.8   192    -5 A
10 BHM     297          16.9          56.2 Birmi~  33.6  -86.8   644    -6 A
# i 95 more rows
# i 1 more variable: tzone <chr>
```

(c) Which airports had the longest average arrival delays? Which airports had the greatest variability in their arrival delays? [Hint: use `arrange()`]

(d) In the joined table from part (b), how many rows have `NA` values for the airport `name`? What do these rows represent?

Bonus (5 points)

Use the `flights` data frame to answer the following questions: What month had the highest proportion of canceled flights? What month had the lowest? Interpret any seasonal patterns.