STA305 Take-home Midterm #1 - Fall 2023

Rules

1. Your solutions must be written up based on the R Markdown (Rmd) file called midterm-<your-name>.rmd (**use your name to replace**) using the attached template midterm-your-name.Rmd (downloaded from Canvas). This file must include your code and write up for each task. **Knit the Rmd and submit both Rmd and the knitted pdf by the submission deadline**.

2. Your reference of the exam is limited to my lecture slides, any files that are available on our Canvas site, your homework files and the help documentation of R. You are not allowed to use Internet search or AI tools to help you. You are also not allowed to communicate with anyone (other than the instructor) regarding any questions about the exam. If you have questions, please use the chat group to ask them and I will try my best to respond on time.

3. Late submission will not be accepted. Please do not wait until the last minute to knit/submit. Save your intermediate work timely to avoid unnecessary loss of your work. Knit your document for each question rather than knit the whole document in the end to avoid failure in debugging before the deadline.

4. You may only use `tidyverse`, `openintro` and `nycflights13` (and its dependencies) for this assignment. Your solutions may not use any other R packages other than those packaged loaded by defaulty in Rstudio.

5. You must **answer each question in words** along with R codes in the markdown notebook. Even if the answer seems obvious from the R output, make sure to state it in your narrative as well. For example, if the question is asking what is 2 + 2, and you have the following in your document, you should additionally have a sentence that states "2 + 2 is 4."

```
2 + 2
```

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

6. For each question, create summary table or graph (or both) following the instructions to provide evidence of your answer. You can create multiple tables or graph if necessary.

7. For all graphs, they need to be polished in details including but not limited to

- have proper axis labels and legends
- labels should have units when applicable
- have proper titles
- have reasonable font sizes for all texts in the graph
- the plotting range should be reasonably good to illustrate particular data trends
- use color when necessary to add contrast between different elements of the graph

Grading Criterion (100% in total + 10% bonus)

Each question (including bonus) accounts for 10% of the total score. For each question, completeness of the approach and coding work accounts for 30%; correctness of figure/table accounts for 30%; correctness of your answer accounts for 20%; graph details account for 20%.

Exam Questions

You must both use codes (including graphs or results printed out by the codes) and words to answer each question.

1. The `nycflights13` data set

For the following tasks, use data sets of the `nycflights13` package:

- `flights`: data of all flights that departed NYC in 2013
- `weather`: hourly meteorological data for each airport
- `planes`: construction information about each plane
- `airports`: airport names and locations
- `airlines`: translation between two letter carrier codes and names

Task list

a) Create histograms of departure time for each month in one graph (there should be 12 histograms). Each bar should represent counts of flights between 12am-1am, 1am-2am, ..., 11pm-12am etc. Choose proper setting for the bar plot and polish your graph to make it look reasonably nice (for this question, graph details account for 50% of the score).

b) What are the top ten destinations that had most flights from NYC airports in 2013? Print a table that lists those airports in descending order of frequency and shows the number of flights heading to each airport.

c) For flights that have a positive departure delay (`dep_delay` is greater than zero), create a bar plot where the x variable is month and the y variable is the mean depature delay for each month. According to your graph, which month has the worst delay?

d) Which day had the highest mean temperature in each of the origin airport based on the `weather` data set? Create table or graph to answer this question.

e) Find a way to select all **overnight flights** (also called "Red Eye Flights" that depart at late night and arrive in the morning) from the data set. Compare the average travel distance of all overnight flights with that of non-overnight flights. What do you find?

f) (Bonus) Create a visualization that effectively shows whether there is a relationship between the average daily departure delay and the average daily temperature. Summarize your findings from the graph.

2. The `babies` data set from `openintro` package

After loading `openintro` library, a data set named `babies` should be ready to explore. The following questions are based on this data set.

a) The variable `parity` is a binary indicator of first pregnancy (0 for first pregnancy and 1 otherwise). Create a new column named "First_pregnancy_flag" which has the value "First Pregnancy" if `parity` is 0 and "Not First Pregnancy" if `parity` is 1. Then create a boxplot to compare the distribution of gestation between these two groups using the newly defined variable.

b) The variable `smoke` is an indicator of smoking mother or not (0 for non-smoker, 1 for smoker and `NA` for unknown). Create a new column named "Smoker_flag" which has the value "Smoker" if `smoke` is 1, or "Non-smoker" if `smoke` is 0, or "Unknown" if `smoke` is 'NA'. Then create a boxplot to compare the distribution of gestation vs different smoking groups using the newly defined variable.

c) Create a graph to study the relationship between `gestation` and `age`. You should have a scatterplot and a smooth line plot on the same graph. Is there any correlation between the two variables based on your graph?

d) Create a proper graph to visualize the effect of gestation length and first-pregnancy on the birthweight of babies. Summarize your findings from the graph.

e) BMI, the Body mass index, is a value derived from the mass (weight) and height of a person. Create a graph to visualize the effect of mother's smoker status and BMI on the birthweight of babies. For your reference, BMI follows the formula

$$\text{BMI} = \frac{\text{mass in pounds}}{\text{height in inches}^2} \times 703$$