# Homework-3: *Your Name Here*

All of these homework problems are from the *R for Data Science* book. The section numbers (e.g., "3.2.4 Exercises") refer to sections in this book. Although the questions are based on those in the book, some questions ask for additional details or analysis.

When solving these problems, you are allowed to use any method from the book or class, even if that method wasn't yet covered when the exercise was presented in the book.

Write answers that are as complete as possible. If a graph is helpful for formalizing the solution, provide the graph. If a table is helpful, provide a table. In the text part of the answer, outline the progression in your thinking as you perform the analysis.

Note that you should type your answers in RStudio, by typing into the file **Homework-3.sa.Rmd**.

---

## 5.6.7 Exercises

### (1) 5.6.7 Exercise 1 (25 pts; 5 each)

Brainstorm at least 4 different ways to assess the typical delay characteristics of a group of flights. Consider the following scenarios:

- A flight is 15 minutes early 50% of the time, and 15 minutes late 50% of the time.
- A flight is always at least 10 minutes late.
- A flight is 30 minutes early 50% of the time, and 30 minutes late 50% of the time.
- 99% of the time a flight is on time. 1% of the time it's 2 hours late.

For each scenario, using the `flights` dataset, give the analysis, give the answer, and then discuss the findings.

Then discuss this question: *Which is more important: arrival delay or departure delay?* Why? Explain your reasoning.

### (2) 5.6.7 Exercise 3 (5 pts)

Our definition of cancelled flights (`is.na(dep_delay) | is.na(arr_delay)`) is slightly suboptimal. Why? Which is the most important column?

### (3) 5.6.7 Exercise 4 (10 pts)

Look at the number of cancelled flights per 24-hour day. Is there a pattern? Is the proportion of cancelled flights related to the average delay?

### (4) 5.6.7 Challenging Exercise (10 pts)

For each plane, count the number of flights before the first delay of greater than 1 hour.

*Hint*: I found this exercise, which has since been removed from the book, to be challenging but rewarding. A solution involving a `for` loop is easy to conceptualize. But, the Tidyverse way, and also the more efficient way, is to avoid explicit loops. Among various loop-less elegant solutions is to use `row_number()` to number the flights for each plane.

## 5.7.1 Exercises

### (5) 5.7.1 Exercise 2 (10 pts)

Which plane (`tailnum`) has the worst on-time record?

### (6) 5.7.1 Exercise 3 (10 pts)

What time of day should you fly if you want to avoid delays as much as possible?

### (7) 5.7.1 Exercise 5 (10 pts)

Delays are typically temporally correlated: even once the problem that caused the initial delay has been resolved, later flights are delayed to allow earlier flights to leave. Using `lag()`, explore how the delay of a flight is related to the delay of the immediately preceding flight.

## 7.3.4 Exercises

### (8) 7.3.4 Exercise 1 (10 pts)

Explore the distribution of each of the `x`, `y`, and `z` variables in `diamonds`. What do you learn? Think about a diamond and how you might decide which dimension is the length, width, and depth.

### (9) 7.3.4 Exercise 3 (5 pts)

How many diamonds are 0.99 carat? How many are 1 carat? What do you think is the cause of the difference?

### (10) 7.3.4 Exercise 4 (5 pts)

Compare and contrast `coord_cartesian()` vs `xlim()` or `ylim()` when zooming in on a histogram. What happens if you leave binwidth unset? What happens if you try and zoom so only half a bar shows?

## 7.4.1 Exercises

### (11) 7.4.1 Exercise 1 (5 pts)

What happens to missing values in a histogram? What happens to missing values in a bar chart? Why is there a difference?

### (12) 7.4.1 Exercise 2 (5 pts)

What does `na.rm = TRUE` do in `mean()` and `sum()`?

## 7.5.1.1 Exercises

### (13) 7.5.1.1 Exercise 1 (10 pts)

Use what you've learned to improve the visualization of the departure times of canceled vs. non-canceled flights. Explain the ways in which the visualization is an improvement.

### (14) 7.5.1.1 Exercise 4 (10 pts)

One problem with boxplots is that they were developed in an era of much smaller datasets and tend to display a prohibitively large number of outlying values. One approach to remedy this problem is the violin plot. For the diamonds dataset, examine the behavior of the different diamond cuts, using both boxplots and violin plots. Compare the resulting graphs. For this analysis, which is better?

## 7.5.2.1 Exercises

### (15) 7.5.2.1 Exercise 1 (10 pts)

How could you rescale the count dataset above to more clearly show the distribution of cut within color, or color within cut?

### (16) 7.5.2.1 Exercise 2 (10 pts)

Use `geom_tile()` together with dplyr to explore how average flight delays vary by destination and month of year. What makes the plot difficult to read? How could you improve it?

## 7.5.3.1 Exercises

### (17) 7.5.3.1 Exercise 1 (10 pts)

Instead of summarizing the conditional distribution with a boxplot, you could use a frequency polygon. What do you need to consider when using `cut_width()` vs `cut_number()`? How does that impact a visualization of the 2D distribution of `carat` and `price`?

### (18) 7.5.3.1 Exercise 2 (10 pts)

Visualize the distribution of carat, partitioned by price.

### (19) 7.5.3.1 Exercise 3 (5 pts)

How does the price distribution of very large diamonds compare to small diamonds. Is it as you expect, or does it surprise you?