AE-04: NYC Flights

Table of contents

In this activity we explore a random sample of domestic flights that departed from the three major New York City airports in 2013. We will generate simple graphical and numerical summaries of data on these flights and explore delay times. Since this is a large data set, we'll also use some techniques for filtering and grouping our data.

Exercise 1

Look carefully at these three histograms. How do they compare? Are there features revealed in one that are obscured in another? What can you learn about delays in departure time from these histograms?

Exercise 2

In your own words, explain the significance of the three numbers generated by the above R code.

Exercise 3

We just created a new data frame called `pdx_jul_flights` that includes flights headed to PDX in July. How many flights meet these criteria?

Exercise 4

Describe the distribution of the **arrival** delays of these flights using a histogram and appropriate summary statistics.

Exercise 5

Calculate the median and interquartile range for `dep_delay` for flights in the `pdx_flights` data frame, grouped by carrier. Which carrier has the most variable departure delays?

Exercise 6

To answer this question, we can use code like the example above to:

```
-   `group_by` `carrier`, then
-   `summarise` mean departure delays.
```

Which carrier has the highest average delay departing from an NYC airport?

Exercise 7

Rather than choosing the carrier with the lowest mean departure delay, we could instead choose the carrier with the lowest median departure delay. What are the pros and cons of these two choices?

Exercise 8

Based on this boxplot, estimate the median, Q1, and Q3.

Exercise 9

The very first thing we did in this activity was to visualize the distribution of `dep_delay` for the entire `nycflights` data set with a histogram. Now use a boxplot to visualize the same distribution. What do you notice?