STK 353

Practical 2: Scripts

Question sheet

Submission: 30 July 2018, 17:30

This practical is based on a pre-loaded dataset in the R package nycflights13.

Filter

1. Find all flights that
    a. Had an arrival delay of two or more hours
    b. Flew to Houston (IAH or HOU)
    c. Were operated by United, American, or Delta
    d. Departed in summer (July, August, and September)
    e. Arrived more than two hours late, but didn't leave late
    f. Were delayed by at least an hour, but made up over 30 minutes in flight
    g. Departed between midnight and 6am (inclusive)
2. Another useful dplyr filtering helper is between().
    a. What does it do?
    b. Use it to simplify the code needed to answer the previous challenges
3. Missing variables (dataset specific)
    a. How many flights have a missing dep_time?
    b. What other variables are missing?
    c. What might these rows represent?
4. Missing variables:
    a. Why is NA ^ 0 not missing?
    b. Why is NA | TRUE not missing?
    c. Why is FALSE & NA not missing?

Arrange

5. How could you use arrange() to sort all missing values to the start? (Hint: use is.na()).
6. Sort flights to find the most delayed flights. Find the flights that left earliest.
7. Sort flights to find the fastest flights.
8. Which flights travelled the longest? Which travelled the shortest?

Mutate

9. Currently dep_time and sched_dep_time are convenient to look at, but hard to compute with because they're not really continuous numbers. Convert them to a more convenient representation of number of minutes since midnight.

10. Compare air_time with arr_time - dep_time. What do you expect to see? What do you see? What do you need to do to fix it?
11. Compare dep_time, sched_dep_time, and dep_delay. How would you expect those three numbers to be related?
12. Find the 10 most delayed flights using a ranking function. How do you want to handle ties? Carefully read the documentation for min_rank().
13. What does 1:3 + 1:10 return? Why?
14. What trigonometric functions does R provide?

Summarise()

15. Create a graph to explore the relationship between the distance and average delay for each location. The three steps to prepare this data are:

    Group flights by destination.
    Summarise to compute distance, average delay, and number of flights.
    Filter to remove noisy points and Honolulu airport, which is almost twice as far away as the next closest airport.
    Use the pipe function to combine the operations.
16. Define 4 different flight delay characteristics as the one above. Implement each one of these characteristics and list the flights delayed for each one of the characteristics. Choose the characteristic that you think define flight delay the best (including the one above). Look at the number of flight delays per day. Is there a pattern? Plot the average flight delay per day over time.
17. Look at the number of cancelled flights per day. Is there a pattern? Is the proportion of cancelled flights related to the average delay? Plot the number of cancelled flights and average flight delays per day over time on one graph. Use colour, size and shape aesthetics to enhance the interpretation of the graph.