Assignment 2

Question 1

Find all flights that

1. Had an arrival delay of two or more hours.
2. Flew to Houston (`IAH` or `HOU`).
3. Were operated by United, American, or Delta.
4. Departed in summer (July, August, and September).
5. Arrived more than two hours late, but didn't leave late.
6. Were delayed by at least an hour, but made up over 30 minutes in flight.
7. Departed between midnight and 6am (inclusive).

Question 2

1. Sort `flights` to find the most delayed flights. Find the flights that left earliest.

2. Sort `flights` to find the fastest flights.

Question 3

1. 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?

2. Compare `dep_time`, `sched_dep_time`, and `dep_delay`. How would you expect those three numbers to be related?

3. Find the 10 most delayed flights using a ranking function. How do you want to handle ties? Carefully read the documentation for `min_rank()`.

Question 4

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

2. Which carrier has the worst delays? Challenge: can you disentangle the effects of bad airports vs bad carriers? Why/why not? (Hint: think about `flights %>% group_by(carrier, dest) %>% summarise(n())`)

3. A doctor is studying the effect of smoking on lung cancer for a large number of patients who have records measured at five year intervals. They notice that a large number of patients have missing data points because the patient has died, so they choose to ignore these patients in their analysis. What is wrong with this doctor's approach?

Question 5

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

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

3. For each destination, compute the total minutes of delay. For each flight, compute the proportion of the total delay for its destination.

4. 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.

5. Look at each destination. Can you find flights that are suspiciously fast? (i.e. flights that represent a potential data entry error). Compute the air time a flight relative to the shortest flight to that destination. Which flights were most delayed in the air?

6. Find all destinations that are flown by at least two carriers. Use that information to rank the carriers.

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