Lab 4: Wrangling Flights

The `stat20data` package contains a data set called `flights` that contains *all* of the flights that left from San Francisco International Airport and Oakland International Airports between January 1st and December 31st 2020. You will use this rich data set to learn essential skills of data subsetting and aggregation, here using the `dplyr` package in the `tidyverse`.

To get a sense of what this data frame contains, the documentation can be accessed by pulling up the help file:

```
?flights
```

Create a new .Rmd file to fill in your answers to the following questions. As always, start by loading the necessary packages and data:

```
library(tidyverse)
library(stat20data)
data(flights)
```

1. **select()**: Select from the data set the columns containing the origin and the destination and print the first few rows of the data frame. Do they appear to be ordered alphabetically?
2. **filter()**: Filter the data set to contain only the flights that went to Portland, Oregon and print the first few rows of the data frame. How many were there in 2020?
3. **mutate()**: Mutate a new variable called `avg_speed` that is the average speed of the plane during the flight, measured in miles per hour? (Look through the column names or the help file to find variables that can be used to calculate this.)
4. **arrange()**: Arrange the data set to figure out: which flight holds the record for longest departure delay and what its destination was? What was the destination and delay time for the flight that was least delayed, i.e. that left the most ahead of schedule?
5. **summarize()**: Confirm the records for departure delay from the question above by summarizing that variable by its maximum and its minimum value.

---

6. How many flights left SFO during March 2020?
7. How many flights left SFO during April 2020?
8. Create a bar chart that shows the distribution by month of all of the flights leaving the Bay Area. (Search google, the `ggplot2` cheat sheet, or ask on Ed for help making the axis marks show up correctly.)
9. Create a histogram showing the distribution of departure delays for all flights. Describe in words the shape and modality of the distribution and using numerical summaries (i.e. summary statistics) its center and spread. Be sure to use measures of center and spread that are most appropriate for this type of distribution. Also set the limits of the x-axis to focus on where most of the data lie.
10. Add a new column to your data frame called `before_times` that takes values of `TRUE` and `FALSE` indicating whether the flight took place up through the end of March or after April 1st, respectively.
11. Remake the histograms above, but now separated into two subplots: one with the departure delays from the before times, the other with the flights from afterwards. Can you visually detect any difference in the distribution of departure delays?

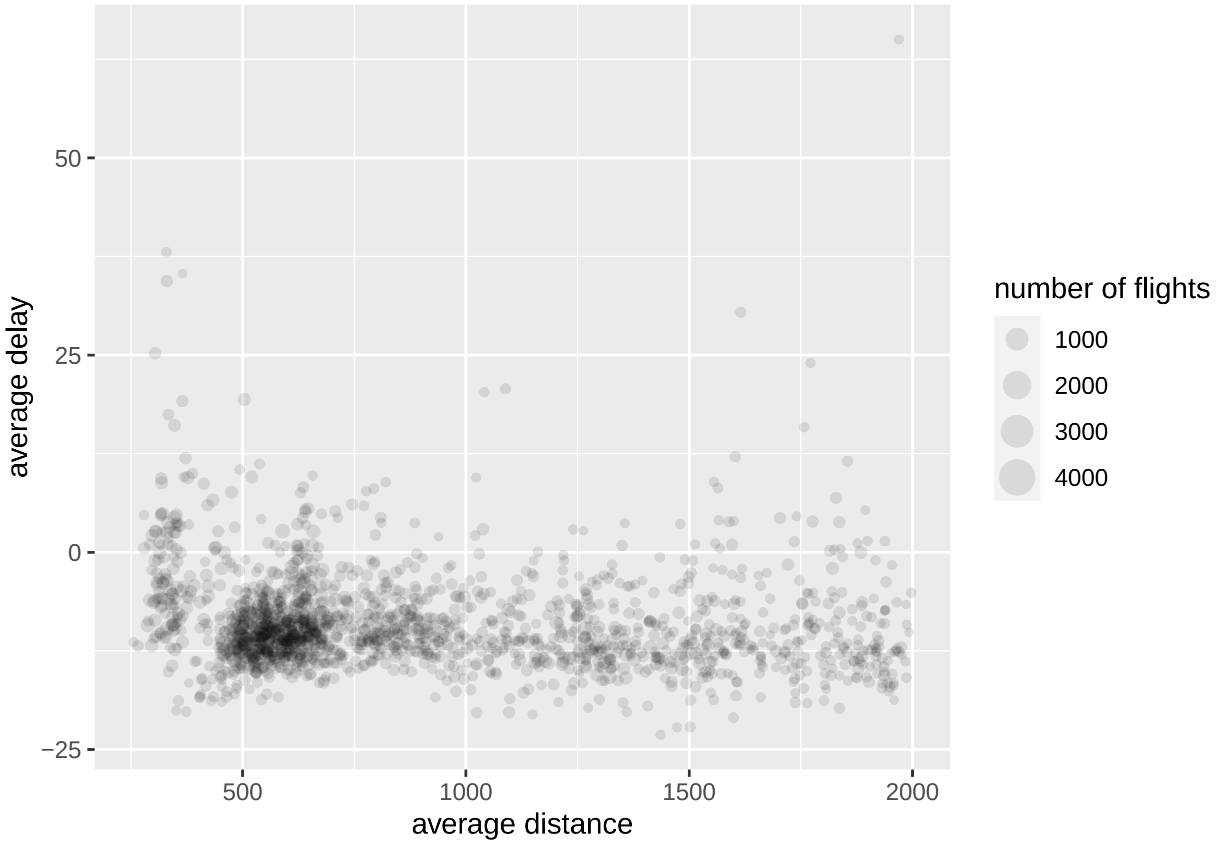
12. If you flew out of OAK or SFO during this time period, what is the tail number of the plane that you were on? If you did not fly in this period, find the tail number of the plane that flew JetBlue flight 40 to New York's JFK Airport from SFO on May 1st.
13. What proportion of the flights left on or ahead of schedule?
14. Create a plot that captures the relationship of average speed vs. distance and describe the shape that you see. What phenomena related to taking flights from the Bay Area might explain this structure? Make any modifications needed to help you see all of the data.

---

15. What is the most common destination of the flights from the Bay Area? The most distant destination?
16. For OAK and SFO separately, what proportion of the flights left on or ahead of schedule?
17. Create a data frame that contains the median and interquartile range for departure delays, grouped by carrier. Which carrier has the lowest typical departure delay? Which one has the least variable departure delays?
18. For flights leaving SFO, which month has the highest average departure delay? What about the highest median departure delay? Which of these measures is more reliable for deciding which month(s) to avoid flying if you really dislike delayed flights?
19. Each individual airplane can be uniquely identified by its tailnumber in the same way that people can be by their social security numbers. Which airplane flew the farthest during this year for which we have data? How many times around the planet does that translate to?
20. What is the tailnumber of the fastest plane that in the dataset? What type of plane is it (google it!)? Be sure to be clear how you're defining fastest.
21. Using the airport nearest your hometown, which day of the week and which airline seems best for flying there from San Francisco (if you're from near SFO or OAK or from abroad, use Chicago as your hometown)? Be clear on how you're defining *best*. (note that there is no explicit weekday column in this data set, but there is sufficient information to piece it together. The following line of code can be added to your pipeline to create that new column. It uses functions in the `lubridate` package, so be sure to load it in at the start of this exercise).

```
mutate(day_of_week = wday(ymd(paste(year, month, day, set = "-")), label = T))
```

22. The plot below displays the relationship between the mean arrival delay and the mean distance traveled by every plane in the data set. It also shows the total number of flights made by each plane by the size of the plotted circle. Please form a single chain that will create this plot, starting with the raw data set. You will also want to exclude the edge cases from your analysis, so focus on the planes that have loged more than 20 flights and flown an average distance of less than 2000 miles.


50
25
0
−25
average delay
500
1000
1500
2000
average distance
number of flights
1000
2000
3000
4000