# Lab 2

BSS Stat 20

2022-06-23

## NOTICE

**If you have any collaborators, please write a sentence before Question 1 which acknowledges them. In addition, make sure that the sentence they therefore are also required to write acknowledges you.**

For a template you can follow, see the course syllabus. This notice will be pasted at the beginning of every lab assignment.

**All code MUST be shown. No credit will be given for correct answers without supporting code.**

## Questions

This week's lab will focus on flight data from the San Francisco (SFO) and Oakland (OAK) airports from the year 2020. There are quite a few questions we can answer with this data set! Let's get into it.

You can load the `flights` data-set on the course server by doing the following.

```
## make sure you load the tidyverse library, too
library(tidyverse)

library(stat20data)
data(flights)
```

### Question 1 - Preliminaries and Variable Types

#### part a

What does each row in the data set correspond to / what is the unit of observation in the data set?

#### part b

List three categorical variables in the data set.

#### part c

List three numerical variables in the data set.

**part d**

Name one variable that might be seen as categorical or numerical. Provide your explanation as to why.

## Question 2 - Examining the Data

**part a**

Provide a guess to the format for the numbers in the `dep_time` column. How is the time being recorded? For instance, what would the value of the `dep_time` column in the first row of the data set mean?

**part b**

Provide a guess as to what the units are for the numbers in the `air_time` column.

## Question 3 - Delays

One traveling out of San Francisco might wonder:

> What *flight destination* out of SFO sees the longest departure delays?

**part a**

Some of the values in the `dep_delay` column are positive, but others are negative. What might positive and negative numbers represent in terms of the time of a departure delay, respectively?

**part b**

Use code to find the destination airport out of SFO with the longest median departure delay time.

One might also wonder:

> Which *airline* (carrier) flying out of SFO sees the longest departure delays?

**part c**

Use code to find the airline flying out of SFO with the longest median departure delay time.

**part d**

Use code to find the airline flying out of SFO with the longest *mean* departure delay time.

**part e**

What can you predict about the distribution of departure delay times for the flight carrier you identified in **part c** and **part d**, based on the values you got for the median and mean in those parts, respectively?

**part f**

Confirm your prediction by plotting a histogram of departure delay times out of SFO for this airline carrier. Make sure you label your axes and give the plot a title.

## Question 4 - Before and After COVID-19

On March 11, 2020, the World Health Organization labeled COVID-19 a global pandemic. As we know, the aviation industry took a huge hit following the proliferation of COVID-19. Does that effect show up in this data set?

**part a**

Use code to add a variable to the `flights` data frame which returns `TRUE` for flights taking place before March 11, 2020 and `FALSE` for flights taking place on and after that date. Save your new data frame into an object called `flights4`.

**part b**

Using the `flights4` data set, use code to create another data frame which tabulates the total number of flights taken before and after March 11, 2020.

**part c**

What does your result in **part b** say about how the COVID-19 pandemic affected the number of flights taken on and after March 11, 2020?

**part d**

Now, we will work with `flights` again. Using the original data set, make a bar chart which visualizes the number of flights by month. Make sure to label your axes and give your plot a title. What does this say about the effect of the COVID-19 pandemic on the number of flights taken in the year 2020?

## Question 5 - Creativity!

This question is meant to give you some autonomy when it comes to analyzing this data.

**part a**

If you flew out of SFO or OAK airports during the year 2020, use code to find the tail number of the plane you were on! If you didn't fly out of these airports in 2020, use code to find the tail number of the plane that flew Flight 452 from SFO to Salt Lake City, Utah, U.S.A. (SLC) on August 6.

The code you present should result in a data frame with a single row. The row will correspond to the flight whose tail number you are interested in.

**part b1**

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.) In solving this problem, it will be helpful for you to think about how we answered **part 2b**. Save the new data frame you have to an object called `flights5`.

**part b2**

Using the `flights5` data frame and considering the airport nearest your hometown, which day of the week and which airline seems "best" for flying there from San Francisco (SFO)? Use code to figure this out.

If you're from the Bay Area or from abroad, go ahead and use the Chicago O'Hare International Airport (ORD) as your home airport.

**Note**: It is up to you to define what "best" means! But we will provide you with the code to create a day-of-week column. It uses the `lubridate` package, so you'll need to load that before you do your work on this question. Here's the line (to be used in a pipeline):

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