

## Week 2 Lab: Introduction to Data



**10/10** points earned (100%)

Quiz passed!

[Back to Week 2](#)



1 / 1  
points

1.

Create a new data frame that includes flights headed to SFO in February, and save this data frame `assfo_feb_flights`. How many flights meet these criteria?



32735



68



**Correct Response**



1345



3563



2286



1 / 1  
points

2.

Make a histogram and calculate appropriate summary statistics for **arrival** delays of sfo\_feb\_flights. Which of the following is false?

☒ No flight is delayed more than 2 hours.



**Correct Response**

☐ The distribution has several extreme values on the right side.

☐ The distribution is right skewed.

☐ The distribution is unimodal.

☐ More than 50% of flights arrive on time or earlier than scheduled.



1 / 1  
points

3.

Calculate the median and interquartile range for arr\_delays of flights in the sfo\_feb\_flights data frame, grouped by carrier. Which carrier has the highest IQR of arrival delays?

☐ JetBlue Airways

☐ Frontier Airlines

☐ American Airlines

☐ Virgin America

☒ Delta and United Airlines



**Correct Response**



1 / 1  
points

4.

Which month has the highest average departure delay from an NYC airport?

☒ July

**Correct Response**

☐ January

☐ March

☐ October

☐ December



1 / 1  
points

5.

Which month has the highest median departure delay from an NYC airport?

☐ October

☐ January

☐ July

☒ December

**Correct Response**

☐ March



1 / 1  
points

6.

Is the mean or the median a more reliable measure for deciding which month(s) to avoid flying if you really dislike delayed flights, and why?

☐

... ..

Mean would be more reliable as the distribution of delays is symmetric.

- ☐ Median would be more reliable as the distribution of delays is symmetric.
- ☒ Median would be more reliable as the distribution of delays is skewed.



**Correct Response**

- ☐ Mean would be more reliable as it gives us the true average.
- ☐ Both give us useful information.



1 / 1  
points

7.

If you were selecting an airport simply based on on time departure percentage, which NYC airport would you choose to fly out of?

- ☒ LGA



**Correct Response**

- ☐ JFK
- ☐ EWR



1 / 1  
points

8.

Mutate the data frame so that it includes a new variable that contains the average speed, `avg_speed` traveled by the plane for each journey (in mph). What is the tail number of the plane with the fastest `avg_speed`? **Hint:** Average speed can be calculated as distance divided by number of hours of travel, and note that `air_time` is given in minutes. If you just want to show the `avg_speed` and `tailnum` and none of the other variables, use the `select` function at the end of your pipe to select just these two variables with `select(avg_speed, tailnum)`. You can google this tail number to find out more about the aircraft.

- ☐ N779JB
- ☐ N959UW
- ☐ N755US
- ☒ N666DN

**Correct Response**

- ☐ N947UW



1 / 1  
points

9.

Make a scatterplot of `avg_speed` vs. distance. Which of the following is true about the relationship between average speed and distance.

- ☐ The relationship is linear.
- ☐ As distance increases the average speed of flights decreases.
- ☐ The distribution of distances are uniform over 0 to 5000 miles.
- ☒ There is an overall positive association between distance and average speed.

**Correct Response**

- ☐ There are no outliers.
-



1 / 1  
points

10.

Suppose you define a flight to be "on time" if it gets to the destination on time or earlier than expected, regardless of any departure delays. Mutate the data frame to create a new variable called `arr_type` with levels "on time" and "delayed" based on this definition. Then, determine the on time arrival percentage based on whether the flight departed on time or not. What proportion of flights that were "delayed" departing arrive "on time"? (answer should be in the form 0.## where ## is between 2 and 7 decimal places, inclusive)

0.1833639

**Correct Response**

