**Week 2 Lab: Introduction to Data**

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? 68
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
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? Delat and united Airlines
4. Considering the data from all the NYC airports, which month has the highest average departure delay? July
5. Which month has the highest median departure delay from an NYC airport? December
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? Median would be more reliable as the distribution of delays is symmetric.
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
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. N666DN
9. Make a scatterplot of avg\_speed vs. distance. Which of the following is true about the relationship between average speed and distance. There is an overall postive association between distance and average speed.
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 percent of flights that were "delayed" departing arrive "on time"? [NUMERIC INPUT] 0.18