Homework 2

Data Science I

1. Each of these tasks can be performed using a single data verb. For each task, say which verb it is:
   a. Find the average of one of the variables.
   b. Add a new column that is the ratio between two variables
   c. Sort the cases in descending order of a variable
   d. Create a new data table that includes only those cases that meet a criterion.
   e. From a data table with three catgeorical vairables A, B, and C, and a quantiative variable X, produce a data frame that has the same cases but only the variables A and X.

2. Use the nycflights13 package and the flights data frame to answer the following questions:
   a. What month had the highest proportion of cancelled flights?
   b. What month had the lowest?
   c. Interpret any seasonal patterns with a graphic.
   d. What plane (specified by tailnum variable) traveled the most times from New York City airports in 2013?
   e. Plot the number of trips per week over the year.

3.Now we will use the nycflights13 package and the flights and planes tables to answer the following questions
   a. What is the oldest plane (specified by the tailnum variable) that flew from New York City airports in 2013?
   b. How many airplanes that flew from New York City are included in the planes table?
   c. How many planes in the planes table have a missing manufacture date?
   d. What are the five most common manufacturers?

4. We will use the Master and Batting table from the Lahman package to investigate the Relative Age Effect. The Relative Age Effect is an attempt to explain anomalies in the distribution of birth month among athletes. Breifly, the idea is that children born just after the age cut-off for participation will be as much as 11 months older than their fellow athletese, which is enough of a disparity to give them an advantage. That advantage will then be compounded over the years, resulting in notably more professional athletes born in these months. Display the distribution of birth months of baseball players who batted during the decade of the 2000s. How are they distributed over the calendar year? Does this support the notion of a relative age affect?

5. The Violations data set in the mdsr package contains information reguarding the outcome of health inspections of restaurants in New York City. Use these data to calculate the median violation score by zip code for zip codes in Manhattan with 50 or more inspections. Note that if no violation was recorded, we would like to consider this a score of a 0. What pattern do you see between the number of inspections and the median score?