# Project Notes

## 10/28/2020

* Null analysis:
  + Satisfaction – Impute the missing values with the mean score **DONE**
  + There are 2401 Flights that were cancelled
  + Departure Delay in Minutes – 2345 nulls remain, which are all due to flights being cancelled. 56 flights remain with a delay in minutes, could this be flights that were cancelled after being initially delayed?
  + Departure Delay in Minutes – 2738 NAs, of which 2401 are due to flights being cancelled, and should be kept in the data. 337 remain
  + Flight time in minutes - 2738 NAs, of which 2401 are due to flights being cancelled, and should be kept in the data. 337 remain
  + When filtering on the subset of the data where flights were not cancelled and Arrival Delay in Minutes was negative, I can see that every NA in in Arrival Delay in Minutes corresponds with a NA in Flight time in minutes (337). Since we are unsure of how to interpolate this data, these 337 rows should be dropped form the dataset. **DONE – dropped from dataset**

## 11/14/2020

* Plots:
  + Distribution of Satisfaction scores
    - Distribution is left skewed with a peak at a score of 4 and 3 being the second highest
  + Mean Satisfaction score by Airline
    - Only Northwest Business Airlines Inc. has a mean score below 3.3
    - Cool&Young Airline Inc. is the
  + Distribution of Passenger Ages
    - Slightly right skewed distribution with a peak around 40 years old
  + Count of Passengers by Gender
    - About 20,000 more Female passengers than Males
  + Density Plot of Satisfaction scores by Gender with mean score by Gender vline
    - Men on average give higher Satisfaction score
  + Count of different Travel Types
    - Business travel has the most passengers by far, with Personal second
  + Heatmap of a correlation matrix between all numeric variables
    - Removed NAs
    - Melted the correlation matrix
    - Plotted the melted matrix
    - Added the correlation between each variable
* Data Processing/Cleaning/Munging
  + How to change certain numeric variables to categorical variables
    - Could use the cut function to a column name or use cut within the mutate function of dplyr
    - No\_of\_flights\_p.a
    - %\_of\_Flight\_with\_other\_Airlines
    - No.\_of\_loyalty\_cars