# Project Notes

## Business Questions

* Which airlines and origin cities have the highest rates of departure delays and cancellations?
  + Variables: Airline Code, Airline Name, Origin City, Flight Cancelled, Departure Delay in Minutes
* Do vacation destinations have a higher satisfaction?
  + Variables: Satisfaction, Type of Travel, Destination City
  + After identifying destinations where customers will go for a weekend getaway or vacation, filtering out business travel, and grouping by vacation spot or not, there is no increased satisfaction of the flight if the customer is headed toward a vacation spot.
  + This goes against the initial hypothesis that satisfaction would be increased when heading towards a vacation destination because customers would be excited and/or less stressed
* # A tibble: 2 x 4
* Vacation avg\_satisfaction count std
* *<int>* *<dbl>* *<int>* *<dbl>*
* 1 0 2.75 22039 0.856
* 2 1 2.75 28088 0.859
* Which demographics are most price sensitive?
  + Variables: Age, Gender, Price Sensitivity
* Which airlines have the highest satisfaction scores? Distribution of the scores should be factored in as well.
  + Variables: Satisfaction, Airline Code, Airline Name

Chart

Description automatically generated

* Which variables affect Satisfaction most?
  + Variables: Type of Travel, Flight Cancelled, Departure Delay in Minutes, Arrival Delay in Minutes, Flight Time in Minutes, Flight Distance, Arrival Delay Greater 5 Mins, Airline Code, Shopping Amount at Airport, Eating and Drinking at Airport, Class, Airline Status, Age, Gender, Price Sensitivity, No of Flights P.A., % of Flight With Other Airlines, No. of Flight With Other Airlines
  + Use Linear regression and

## 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?
  + Arrival Delay in Minutes – 2738 NAs, of which 2401 are due to flights being cancelled, and should be kept in the data. 337 remain, which should be dropped **[DONE]**
  + Flight time in minutes - 2738 NAs, of which 2401 are due to flights being cancelled, and should be kept in the data. 337 remain, which should be dropped **[DONE]**
  + 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 **[DONE]**
      * Created cat\_No\_of\_Flights and split into Low Medium and high by cutting into thirds
    - %\_of\_Flight\_with\_other\_Airlines **[DONE]**
      * Created `cat\_%\_of\_FLight\_with\_other` by cutting the column into thirds
    - No.\_of\_loyalty\_cars **[DONE]**
      * Created a new column, cat\_loyalty\_cards, for the None, Low, Medium, High categories based on the distribution of cards a passenger has
    - Age **[DONE]**
      * Created 8 intervals in the cat\_Age column
  + Use dummy variables:
    - Airline Status **[DONE]**
    - Type of Travel **[DONE]**
    - Class **[DONE]**
    - Cat\_Age **[DONE]**
    - Cat\_loyalty\_cards **[DONE]**
    - cat\_No\_of\_Flights **[DONE]**
    - cat\_%\_of\_FLight\_with\_other **[DONE]**
* Summarization and Exploration
  + Grouped by Destination City to see the how many people have arrived there **[DONE]**

## 11/18/2020

* Find out if Vacation spots had higher satisfaction scores
  + Read the cleansed data set into the Vacay R file
  + Created vector of unique destinations, and wrote the vector to a csv
  + Mapped based on user input, what destinations were vacation destinations
  + Merge the mapping with the dataset, filter out “Business Travel” to make sure we are only looking a leisure or travel on one’s own time
  + Group the data by whether the destination is a vacation destination or not
  + Take the mean of the satisfaction score, count, and standard deviation
  + No difference!

> avg\_score\_vac\_dest <- df\_merged %>%

+ filter(Type\_of\_Travel != "Business travel") %>%

+ group\_by(Vacation) %>%

+ summarize(avg\_satisfaction = mean(Satisfaction),

+ count=n(),

+ std=sd(Satisfaction))

> avg\_score\_vac\_dest

# A tibble: 2 x 4

Vacation avg\_satisfaction count std

*<int>* *<dbl>* *<int>* *<dbl>*

1 0 2.75 22039 0.856

2 1 2.75 28088 0.859

## 11/19/2020

* Wrote Data cleaning and processing functions in DataCleaning.R **[DONE]**
  + Process\_flight\_survey function cleans the entire dataset and does the feature engineering
    - This can be imported into any script and run so the data does not have to be stored on the hard drive
* Begin narrowing scope of variables for regression
  + Should I make dummy variables out of the Airline Name?
* Figure out null values that are due to flight cancellations
  + Filter for flights cancelled that have null delays, fill with 0
    - There will still be nulls remaining because there are some columns that are na when the delay is not