Project Proposal Paper

Analysis of Airline Data

Mukund Kalantri
Computer Science
CU Boulder
Boulder, CO USA
muka4041@colorado.edu

Rohit Kharat
Mechanical Engineering
CU Boulder
Boulder, CO USA
rokh4336@colorado.edu

Reid Glaze
Mechanical Engineering
CU Boulder
Boulder, CO USA
regl1257@colorado.edu

PROBLEM STATEMENT / MOTIVATION

In this project, we will analyze airline data taken from approximately 2 million US domestic flights and analyze the delays and cancellations of these flights. We will aim to answer the following questions:

1) Which airports should be focused more on infrastructure development in order to mitigate the frequency and severity of delays?
2) What are the relationships between airline carriers and flight delays/cancellations, so that airlines could improve their services and customers can make informed decisions about which airlines should be preferred?
3) What are the most common reasons for delays in flights based on airports, so that the airport authorities could improve their services?
4) What are the relationships between times in a month or times in a day when there are a lot of cancellations and delays so that customers can make an informed decision while booking flights with respect to their timings?

Additionally, we will also consider building a prediction model that will be used to calculate the arrival delay for a flight based on the other attributes. This could directly be used by the customers and the airport authorities.

LITERATURE SURVEY

This dataset has been used for several different data mining tasks in a Coursera specialization course called “IBM Data Analyst”:

1) Yearly number of flights canceled
2) Average delay time by airline
3) Monthly average delay for each type based on airline
4) Yearly number of flights delayed based on the arrival and departure states

The link to this specialization is: https://www.coursera.org/professional-certificates/ibm-data-analyst

This dataset has also been used in another Coursera course called “Data Analysis with R” where it has been used for:

1) Performing Exploratory Data Analysis
2) Building a prediction model for predicting flight arrival delay
3) Using the R package tidymodels to do an evaluation on the model

The link to this course is: https://www.coursera.org/learn/data-analysis-with-r

PROPOSED WORK

Data Integration

Since this dataset, we are using contains information for over 194 million flights and this amount of data is hard to work with, we are focusing exclusively on flights that have taken place in early 2022. Since it is necessary to download each month separately, we have four different files containing flight data. We plan on combining these files in order to have one workable dataset. We will use the data for the months of January to April of 2022, which includes data for over 2 million US domestic flights.

Data Preprocessing

The initial data preprocessing steps will be importing required libraries and reading the collected data. We will be employing data cleaning techniques such as removing unnecessary features, handling missing data, and filtering outlier data. Data transformation methods will be employed for transforming variables to correct data types for our computations. We will standardize the data using Scikit-learn's libraries for reducing the effect of unscaled data. This data includes 109 attributes. We will not be using all of these attributes in our evaluations, so we will be using dimensionality reduction techniques (e.g. PCA) to keep the most important features for our interesting questions. We would also be using feature engineering techniques to compute new features such as features pertaining to different time periods for our analysis. We will be using data splitting techniques for splitting the feature array and label array into training and testing sets.

Data Analysis

We will perform statistical analysis to describe our data and pattern mining techniques to find hidden patterns in the data. Regression analysis will be done to estimate the relationship between the set of features. Correlation analysis will be done to understand the effect of dependent variables on the independent variable. This will help us to perform predictive analysis to identify future outcomes. Additionally, we will plan to perform a Monte Carlo simulation for analyzing the effect of unpredictable variables on the delay or cancellation of flights. Finally, time series analysis will be used to identify trends, seasonality, and cyclic patterns in our data.

Data Visualization

We will use data visualization techniques to display the findings of our project. This can be used to depict trends, correlations, or patterns in the data. For data reduction or showing important features, chart types such as bar graphs will be useful. We will be using univariate analysis techniques such as distribution plots, box and whisker plots for outliers, and violin plots for kernel density estimation. In the bivariate analysis we will be using line plots, bar plots, and scatter plots for showing important patterns or clusters. We plan on using Tableau and Python data visualization libraries for our visualization software.

Prediction and Classification

We will be focusing on building a delay prediction model. This can be used to predict how long a given flight will be delayed. This can be estimated by examining the significance of attributes in determining the probability and length of a delay, which can be applied to new testing data. Regression analysis will be used for prediction modeling and classification models will be employed for classifying delay or cancellation types based on other features. Monte Carlo simulation methods will be helpful for us to generate models for our analysis as well as to understand the probability distributions for possible outcomes.

DATA SET

We will be using a dataset called "Airline Reporting Carrier On-Time Performance Dataset" that contains information about US domestic flights that have occurred since 1987. This data was compiled from the Bureau of Transportation Statistics.

The link to the data set is:

https://developer.ibm.com/exchanges/data/all/airline/

EVALUATION METHODS

Model evaluation will be done based on our interesting questions and statistical methods will be utilized to infer confidence for our findings.

For statistical analyses, we will be doing Correlation tests, Chi-Square tests, Log Transformation, Normality tests, Student's t-tests, ANOVA, Time series tests, etc.

For evaluating our delay prediction model, we would be doing cross-validation on our model to predict problems like overfitting and underfitting, and using regularization methods to handle these problems.

Once our model is complete, we would be using evaluation metrics, such as accuracy, precision, recall, F1 score, Log Loss, ROC, etc to evaluate our final model.

Confusion matrix will be used to describe the performance of our classification model on a set of test data generated.

TOOLS

Software

- Google Colaboratory will be used so that we can all work on code at the same time.
- Jupyter Labs will be used for data analysis with Python to solve the questions provided.
- RStudio will be used for data cleaning and data pre-processing.
- Tableau will be used for Data Visualization

Libraries

- Numpy and Pandas for data preprocessing
- Matplotlib for data visualization
- Keras and Tensorflow for building a prediction model
- SciKit learn and SciPy for Exploratory Data Analysis and model evaluation

MILESTONES

Task	Deadline
Finish Data Cleaning	July 22
Finish Data Preprocessing	July 25
Brainstorm which techniques to use for Specific Questions	July 27
First Draft to answer questions (other than prediction model)	July 30
Determine which features should be used for the prediction model	Aug 3
Final answer to questions (other than prediction model)	Aug 6
Finalize prediction model	Aug 8
Present Work and submit all remaining assignments	Aug 10

MILESTONES COMPLETED

So far, we have been able to complete all the milestones till date that are listed above.

Question 1

We found out which airports we should focus on infrastructure development with respect to carrier and aircraft delay. This information will help us to mitigate the frequency and severity of delays on passengers.

Question 2

We found out which airlines have the most percentage of their flights delayed during departure and arrival based on different delay times.

Question 3

We found the most common reasons for delays in flights based on origin airports, origin state, and airlines. The analysis was done to find the top 10 candidates which causes the delay in flights, so that information will be helpful for making informed decisions about which airlines and airports should be preferred.

Question 4

We determined which times are optimal for consumers to purchase a flight in order to minimize the probability of delays.

MILESTONES TO DO

Question 4

We wish to add more information related to delays based on the time of the day. The time of the delay is a very important variable and delays cannot be classified in a binary fashion, the same way cancellations can. We hope to find the average times of delays that occur at different times.

Prediction Model

We have not yet started the prediction model, but hope to soon with the majority of the questions being answered. We will analyze the findings of our questions in order to determine what features we will use. Then, we will create a model using cross validation.

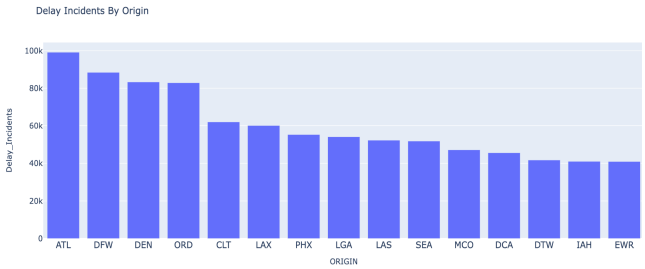
What affects cancellation?

While working on delay analysis, we found that the cancellation information does not relate with the delay time information. So we would work further on finding what causes the cancellation of flights and delay correlates with cancellations.

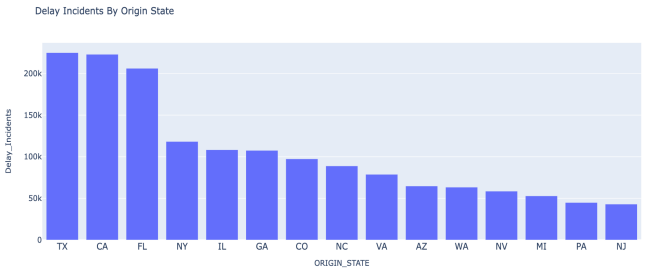
RESULTS SO FAR

Question 1

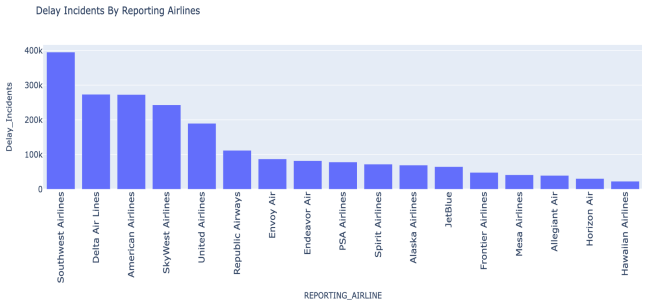
The information we wanted to seek from this interesting question was to understand which airports are affected by delays and how frequent the incidents happened at a particular airport. This information will be helpful in future for infrastructure development considering delays which are related to airport. The analysis performed was helpful for us to understand which airports are frequently affected by delays and whether the type of delay can be used to mitigate that in future.The figure below shows bar graph for top 15 airports having most frequent delays:



The figure below shows bar graph for top 15 states having most frequent delays:



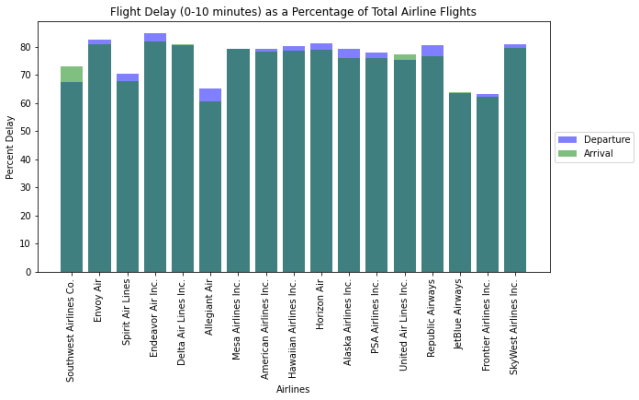
The analysis was further extended to reporting airlines and the figure below shows the most frequent airlines affected by delays:



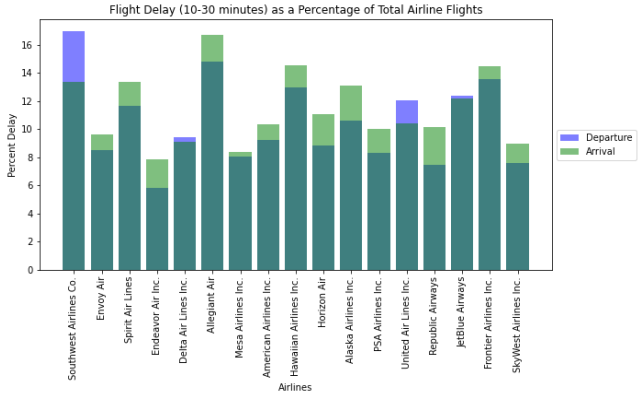
Question 2

For this problem, we wanted to find those airlines which have a lot of delays and cancellations in their flights and which need to improve their services.

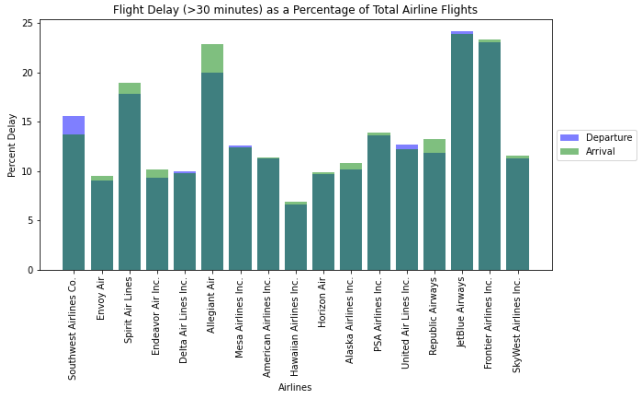
We filtered out our data based on arrival and departure for each airline first. The next step was to find out the unique airlines that we have. Since our original data only has airlines codes, we found the airline names based on these codes from the Bureau of Statistics website. We then looped through all the airlines and filtered out their data from the entire data we have. The next steps were to remove all those flights where there was no delay. Then for both arrival and departure cases, we removed those flights that had missing values in the arrival/departure columns. After this, we categorized all the flights with delays into three categories - Delay less than 10 minutes, Delays between 10 to 30 minutes, and Delays more than 30 minutes. Once we had these counts, we than took out the percentage that these delays were from the total number of flights for that airline. Our results are presented below.



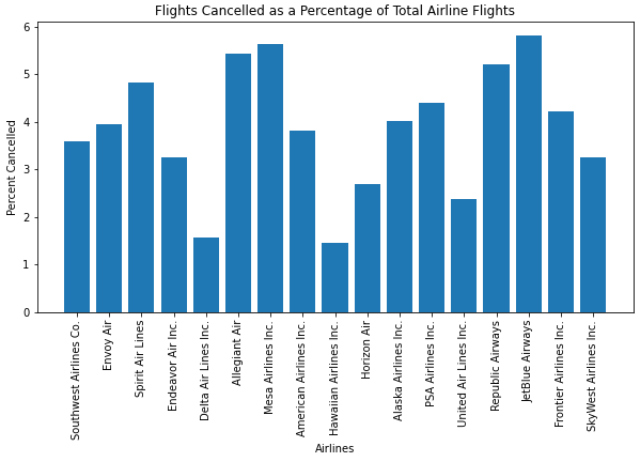
We found that for all of the airlines, at least 60 percent of the flights had short (upto 10 minute) delays. In these, Endavour Air, SkyWest Airlines, and Republic Airways were the airlines with the most delays. Whereas, Delta Airline, JetBlue Airlines, and Frontier Airlines had the best performance. However, the margin here is not much between these airlines and based on the percentage of flight delays for all airlines, we can say that it is pretty normal for any flight to have short delays.



The above bar graph is for Medium length (10-30 minutes) flight delays. We found that in general, the arrival delays were more in comparison to departure delays, and since this happens with most airlines, we can say that the flight times mentioned while booking is not very accurate and is a a little more than what is mentioned. We can also see that Southwest Airlines often get delayed even before leaving airport so they should improve their pre-flight services (ticket checking, boarding luggage, aircraft systems check, etc). The best performing airlines in this category were Endavour Airlines, Mesa Airlines, and SkyWest Airlines. Whereas the worst performers were, Allegiant Air, Frontier airlines, and Hawaiian Airlines. We also saw that at least about 6 percent of all flights irrespective of the airlines gets medium length delays.



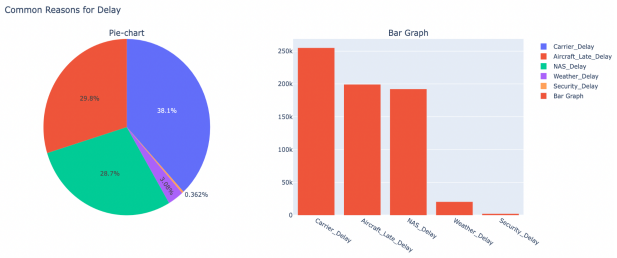
The long (more than 30 minutes) delays category is represented by the bar graph above. This is the category which should be really focused by the airlines to improve their services as such long delays gives really bad experience to the customers and continued bad performance can cause problems to ticket sales for these airlines and affect their business. We observed that JetBlue Airlines, Frontier Airlines, and Allegiant Air have really bad performance here and more than 20 percent of their flights have such long delays. The bar graph shows that their is a pretty big percentage gap in these airlines compared to their competitors, which should concern them and force them into doing detailed analysis to figure out the reasons for delays and improve their services. While the other airlines had comparable performance here, Hawaiian airlines certainly has performed really well which should be a big hope for them in terms of their ticket sales going up.



This last bar graph shows us the percentage of flights getting cancelled by airlines. We can see that these percentages range from 2 to 6 percent of total airline flights in general. While JetBlue again turned out to be the worst performer, the other airlines with bad performance were Mesa Airlines, and Allegiant Air. However, the difference between performance here for the remaining airlines does not vary much, it is worth mentioning that Hawaiian Airlines again had the best performance along with Delta Airlines, which is a big plus point for their business.

Question 3

The answer we were seeking from this question was what were the most common reasons for delays based on airport information. This information would be helpful for us to understand which airports are affected the most by a particular type of delay and to provide relevant information to airport authorities to work on improving their services. The analysis was further extended considering origin state of flights as well as which airlines have the most frequent cases of delays. The figure below shows the most common reasons of delays:

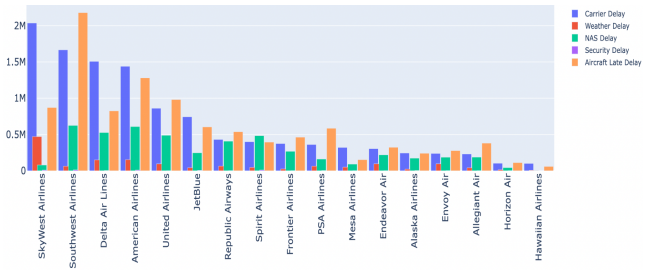


The figure below shows the origin airports most affected by carrier delays:

The figure below shows the origin airports most affected by security delays:

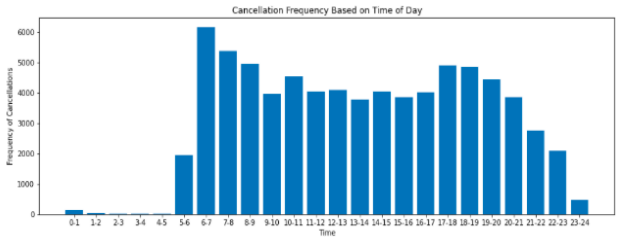
The figure below shows the origin airports most affected by late aircraft delays:

Further the analysis was extended to understand which airlines where affected by a particular type of delay.



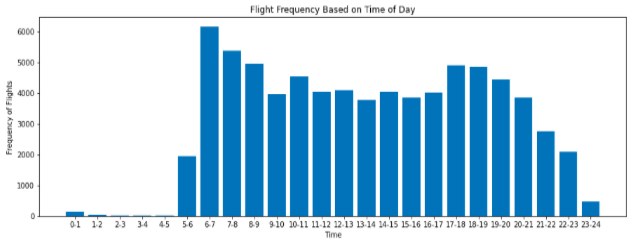
Question 4

For this problem, we wanted to find information that would be useful to the consumer when booking a flight. Since our flight data only includes 4 months of the year, we decided to focus more specifically on times of the day. First, we focused exclusively on cancellations. We filtered out all the canceled flights into a dataframe. Then we separated these flights into

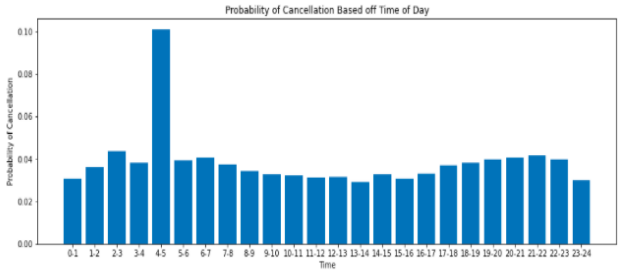


24 different categories based on hourly intervals of departure times. This classification method resembles a decision tree.

This bar graph shows that the highest frequency of cancellations occurs between 6am and 7am. However, this data is not particularly useful for the consumer because it does not give any information on probability. We performed the same type of classification for the data without filtering for cancellations and came up with a similar looking graph.



Next, we divided the frequency of cancellations for each time slot by the frequency of flights that departed in the same time slot. We came up with the following graph.



It appears that there is one significant outlier. Flights that departed between 4am and 5am had nearly a 10 percent chance of being canceled. It is unclear why this was the case. The rest of the distribution appears to be a lot more consistent, with the cancellation rate hovering between 2.9 and 4.4 percent. According to this data, a consumer should book a flight between 1-2pm if they wish to avoid cancellations. The second lowest probability of cancellations occurs between 11pm and 12am, but this probability increases after 1am and before 11pm. Generally speaking, a consumer should aim to depart in the middle of the day and avoid the time slot between 4 am and 5 am if they wish to avoid cancellations.