Yeshiva University DAV - Math and Statistics Final Project Report

Benjamin Freund, Marlee Goodman, Eli Weiss

**U.S. Aviation: Predicting Delays and Examining the Resulting Economic Costs**

**I. Introduction**

Domestic flight delays cost the U.S. economy approximately $32.9 billion annually. $16.7 billion of that gets borne by passengers, and $3.9 billion of that comes from lost demand. $4 billion is associated with a negative impact of the U.S.’s GDP. The last amount of costs due to delay - $8.3 billion - is with what the airlines are left. Flights, their causes for delay, and the economic consequences will be the subject of this report.

For our project, we made use of a dataset of flights that departed from the New York area in 2017. Using this data, along with a weather database for conditions from every day of 2017, we came up with the following research questions:

1. What is predictive of flights being delayed?
2. What is the estimated cost associated with a delay?

We are looking to determine these answers for a business person in this industry. We hope she will be able to estimate the chances of a flight being delayed via our model, and consequently calculate how much money the delay would cost the airline. She can then suggest changes to reduce the economic cost caused by these delays.

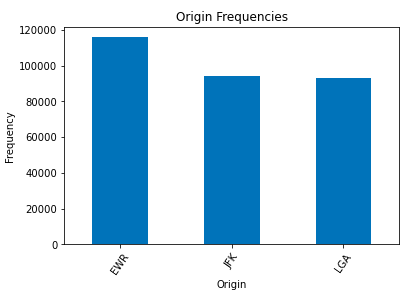
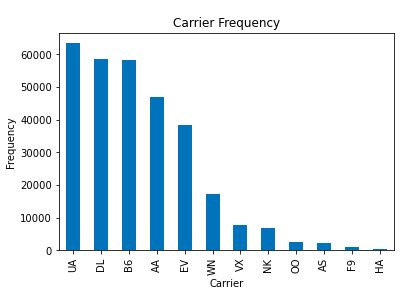
In this essay, we will first introduce our datasets. We will then explain our Exploratory Data Analysis process. We will describe notable findings and how we used them to inform our hypothesis testing direction. The next section will include details about the hypothesis tests on the data. They will explain how we approach the model. Via the data model, we will determine what is predictive of flight delays and the chances of a flight being delayed. Afterward, we’ll share an economic delay model that an investor can use to determine the individual airline flight cost of a delay. We will finish our report with a conclusion, that will include our thoughts about the project process and any future ideas for the subject.

**II. Exploratory Data Analysis**

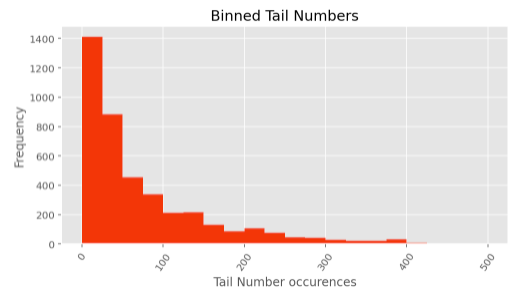
For our project, we initially looked at flight data from the Bureau of Transportation Statistics (BTS) for the US Department of Transportation. Among other flight information in its ‘TranStats’ database, there was data related to 303,748 US flights from 2017 that originated from Newark, Laguardia, and JFK airports (the three major NYC area airports). The flight data was compiled from BTS’s database into a data package by Github user Jay Lee. (jaleetx: [github.com/jayleetx/nycflights](https://github.com/jayleetx/nycflights))

The 18 variables are month, day, actual departure/arrival time, scheduled departure/arrival time, departure/arrival delay, hour & minute, carrier, flight, tail number, origin/destination, airtime, distance, and ‘time\_hour’. A data dictionary explaining these variables can be found in our **Appendix: Exploratory Data Analysis. (**Note: All EDA graphs and explanations can be found in our Github repository: <https://github.com/MarlaGoodman/CMS-Project>)

The most frequently appearing airlines out of the twelve carriers present here are UA (United), DL (Delta), B6 (Jet Blue), AA (American), and EV (Express Jet) - in that order (II.1). The most frequent origin of our flights is Newark (II.2).

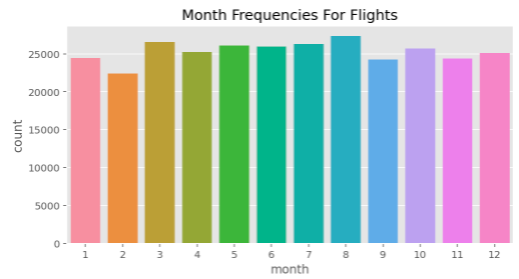
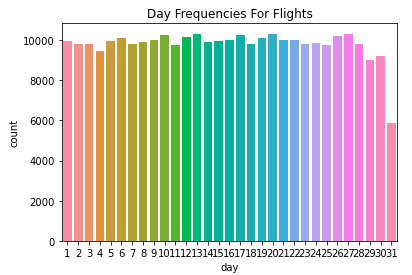
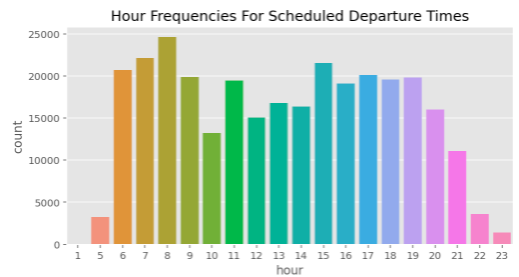


(II.1) (II.2)

Every plane is assigned a registration number, the same way license plates are given to cars. They’re referred to astail numbers. Every U.S. tail number will start with an ‘N’, followed by a number. The ending letters are the carrier abbreviation (e.g. N156EV). Here, there are 4,126 unique tail numbers. We counted all the times each tail number appears and binned the appearance in 25 increment amounts (II.3). The data is skewed right, and the largest category of appearance frequencies is the 0-25 bin. 

(II.3)

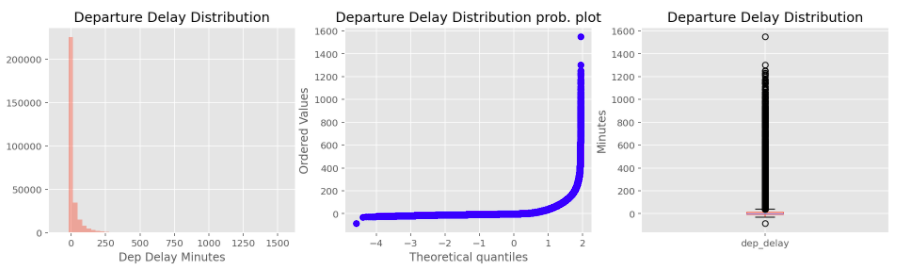
Now that we have looked at physical items related to flights (carrier, tail number, origin), we will examine the time components of the flights (larger versions of these graphs can be found in the Exploratory Data Analysis Appendix - II.4, II.5, II.6 respectively):



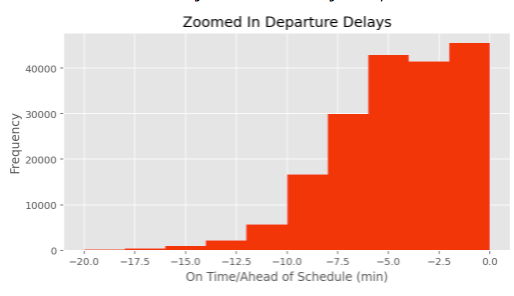
(II.4) (II.5) (II.6)

The hours go from 12:00 AM to 11:59 PM - we can see most flights head out in the 8:00 AM hour. The day variable has a pattern to its graph in that every 3-4 days, the flight frequency per day hits a peak and then drops the day after. February is the month with the least amount of flights, but that is likely because February has the least amount of days in it.

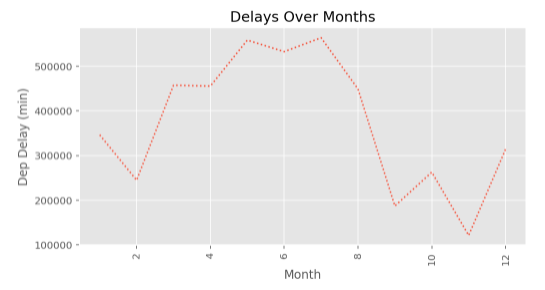
The next time element we want to examine is the departure delays for the flights:

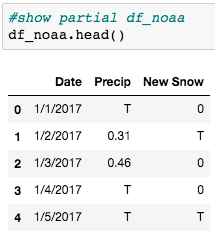


(II.7) (II.8) (II.9)

We can see that the delayed flight variable is heavily skewed to the right. However, let’s look further into the large column in the histogram in (II.7). We will examine the graph that zooms in on the information centered at the 0 minute mark (II.10). From this histogram, we can see that most flights are either on time, or actually ahead of schedule by anywhere from 1 to 20 minutes.

(II.10)

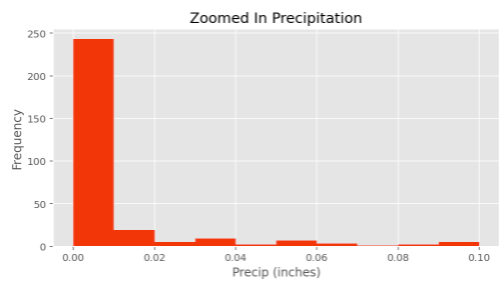
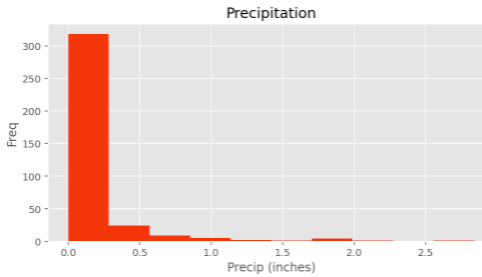
The last time variable we’ll look at is departure delays over months (II.11). In this graph, we see that more departure delay minutes are present between April (4) and July (7). The month with the least amount of departure delay minutes is November (11). 

Our second dataset is from the National Oceanic Atmospheric Administration [Climate](https://w2.weather.gov/climate/xmacis.php?wfo=okx) database. It has weather conditions for each day of 2017 for NYC airports. 

(II.11)

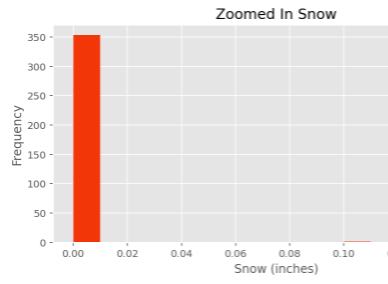
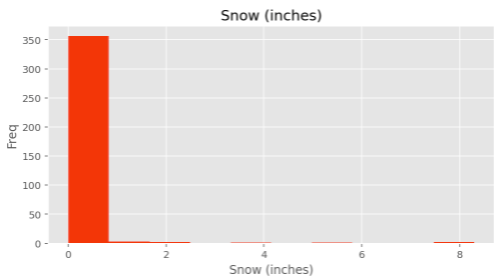
In the dataset (II.12), we have the date for each day of 2017, the precipitation variable, and the snow variable. The ‘T’, or ‘trace’ in the columns is used to describe a very small amount of precipitation that results in no measurable accumulation. If ‘trace’ was converted to a numerical amount, it would equal 0.00. No matter how many traces you add together, the sum will never be more than a trace. Therefore, we converted all the T’s in our data set to 0. (II.12)

Precipitation histogram and a zoomed in graph of the large column:



(II.13) (II.14)

Snow histogram and a zoomed in graph of the large column:



(II.15) (II.16)

1/100 (0.01) of an inch of rain/snow is the first measurable amount of precipitation reported by The National Weather Service. This would not leave puddles on the ground and would slightly wet the surface.This might occur during a light shower for 2-5 minutes or rain drizzle for 2 hours

**III. HYPOTHESIS TESTING**

After performing some initial data cleaning and EDA, we did two hypothesis tests to determine which variables should be used in our model. The first was a t-test comparing average weekend delays to average weekday delays - to compare the two sample means. The null hypothesis was that there was no statistically significant difference between the two delays, while the alternative hypothesis stated that there was such a difference.

In order to perform this hypothesis test, we did some data manipulation. As pictured above, the flights dataset has both month and day variables. Furthermore, all flights in the dataset are from 2017, so we have the year too. We needed to create a day of the week variable which could properly tell us which day of the week corresponded to which day, month, and year combination. We built out these new variables via Python. Additionally, we needed another variable that would give us the total delay. This was achieved by adding together the arrival and departure delays already found in our dataset, and assigning the mean of the total delays to any NAs within the variable.

Next, we created two samples, one containing 190 weekend flights and the other containing 190 weekday flights. 190 is less than 10% of the 303,748 flights in our dataset, so the samples are independent. Additionally, both samples were generated using Python, so they are simple random samples. We run into a problem when we look at the samples’ distributions. Both distributions are definitively right skewed. However, with such a large sample size, we can accept the skewness and proceed with caution.

To perform the t-test, we used Python code from the SciPy library. The alpha level we used was 0.05, which is the equivalent of a 95% confidence interval. Our t-test returned a t-statistic of 1.469 and a p-value of 0.143. Since our p-value is larger than our alpha level, we fail to reject the null hypothesis. We say that there is no statistically significant difference between average weekend delays and average weekday delays. Therefore, day of the week is not a variable we will be including in our model.

The second hypothesis test was a test comparing average snow induced delays to average rain induced delays. Again, we deemed a t-test to be an appropriate statistical test because we wanted to compare two sample means. The null hypothesis was that there is no statistically significant difference between average snow induced delays and average rain induced delays. The alternative hypothesis was that such a difference existed.

Again, before the test itself, we had to perform some data manipulation. Our main challenge was to join the flights dataset with the weather dataset. This was difficult, as they did not have an immediately apparent similarity. The closest similar variable was the date variable, where the weather dataset had a date column in the format of, “YYYY-MM-DD” and the flights dataset had three separate columns for day, month, and year. To join the two datasets, we broke the date column in the weather dataset into three different columns representing day, month, and year, respectively. We then performed an inner join on the datasets, joining on day, month, and year. This would ensure all flights with weather data would be included, even if there were multiple flights recorded on the same day.

Using Python ,we then divided the dataset into two simple random samples. The first sample contained 190 flights where it only snowed and did not rain. The second sample consisted of 190 flights where it only rained and did not snow. It is important to note that flights where it rained and snowed were not included in either sample. Both samples are independent, as they both contain 190 observations, which is less than 10% of the total observations in the dataset. The issue, as before, comes with the distribution of the two samples. Both samples are right skewed, which makes sense for the reasons outlined above. However, we can still proceed with caution due to our large sample sizes. That being said, it is interesting to note that the delays for flights with snow are slightly more centered around 0 (the distribution is slightly narrower), while the delays for flights with rain are slightly more spread out and have a few more extreme values. This could imply that flights with snow are slightly more likely to be delayed for a shorter amount of time.

Using the SciPy library, we performed this t-test with an alpha level of 0.05, or a 95% confidence interval. We got a t-statistic of 2.414 and a p-value of 0.016. Since our p-value is smaller than our alpha level, we reject the null hypothesis. We say that there is a statistically significant difference between the average snow delays and the average rain delays. We suggested before that snow induced delays may be shorter, but further research would need to be done to determine if this is actually the case. Due to the statistically significant difference, we will include rain and snow in our models.

**IV. MODELLING**

After performing statistical tests to see if our factors have significance, we built several machine learning models to see if we could predict if a particular flight was going to be delayed or not.

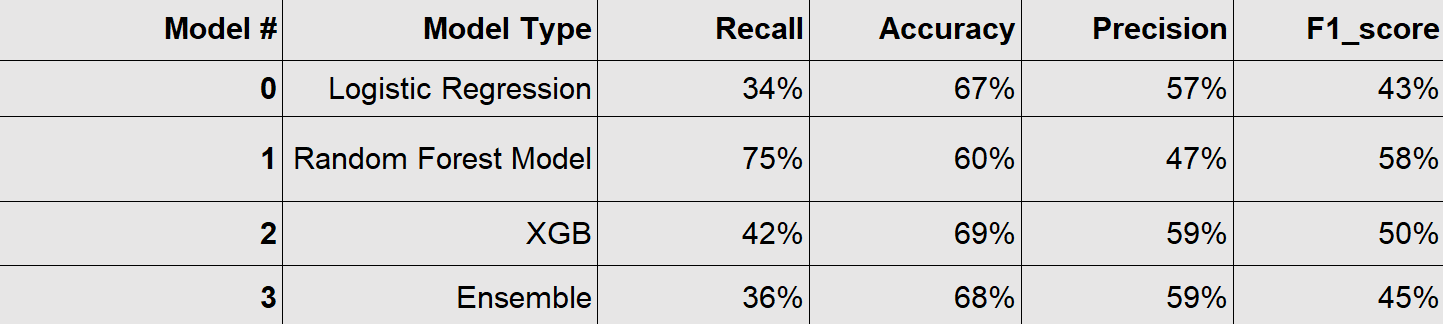
We chose the following factors as predictors for our models:

* Month of the year (one hot encoded)
* Hour of the day (one hot encoded)
* Rain (boolean)
* Snow (boolean)
* Count of flights on the day for a carrier at the airport (scaled between 0 and 1)
* Running Count of delays for this plane (scaled between 0 and 1)

We used these predictors to build 4 machine learning models (Logistic Regression, Random Forest, XGBoost, and an Ensemble of the Logistic Regression and the Random Forest).

Making Sense of the model outputs and comparing their performance**:**

After building the models and seeing their outputs, it’s important that we build KPIs and metrics from these confusion matrices to compare them and see which model performs the best.



IV.1

See **Appendix: Model Confusion Matrices** to see a detailed look at the confusion matrix for each model.

We see that the Logistic Regression Model has high accuracy (% of Trues/Falses that are correctly classified) at 67%. However, it has a low recall (% of Trues that are correctly classified) at 34%. With such a low recall, this model is not preferred.

We see that the Random Forest Model has the highest recall at 75%. However, it has a relatively low precision (% of predicted Trues that are actually true) at 47%.

We see that the XGBoost Model has relatively low recall at 42%, a relatively high precision at 59%, and a relatively high accuracy at 69%.

We see that the Ensemble Model has high accuracy at 68% and a high precision of 59%. However, it has a low recall at 36%. With such a low recall this model is not preferred.

After analyzing all of the models, we must choose which metric is the most important to our study, since one model does not clearly outperform across the board. It follows that, in our study, a False Positive is preferable to a False Negative. This is because if an airline is expecting losses due to delays, they will be able to prepare for the delay. In the event that a delay doesn’t happen, there is no harm. However, if there is no prediction for a delay and then a flight is delayed, the airline could suffer losses they are unprepared for causing issues for the company.

XGBoost performs the best in precision and second best in recall. Random Forest performs the best in recall and the lowest in precision. We prefer the Random Forest model because of its high recall.

Understanding our preferred model: Random Forest utilizes Decision Trees.

It generates many trees and each tree provides a prediction for the class of the observation. Then the model takes the average prediction of all the Decision Trees and uses this average probability to determine if the observation belongs to a particular class. More information on Random Forests can be found here: <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>

**V. ECONOMIC MODEL**

To determine the monetary loss associated with flight delays, we created an economic loss model. The main cost components for a flight are fuel, crew, maintenance, and aircraft ownership. We researched each component and determined how much more a delayed flight would cost per item.

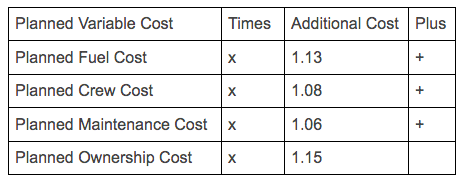
Fuel: In 2017, airlines used around 15B gallons of fuel, at $1.5 per gallon. Fifteen billion gallons times a price of $1.5 per gallon equals $22.5B for U.S. aviation fuel. We also discovered that delays cost airlines $8.3B. The organization Airlines For America, or A4A, has determined that fuel makes up almost 35% of costs when it comes to delays. Thirty-five percent of $8.3B delay costs equals $2.9B. Therefore, of $22.5B spent on fuel, $2.9B of it was due to delay. With 2.9B/22.5B, we find nearly 13% of fuel costs came from delays. If a flight is delayed, the airline can expect to increase its fuel cost for that flight by 13% on average.

Crew: On average, it is estimated that crew expenses account for about one-third of total airline costs, or about $45B (BTS). The FAA oversees 16.4M flights annually. This means airlines will spend $2,750 on average for crew costs per flight. A4A determines that about thirty-three percent of all delay costs come from crew overtime -- 33% of total airline delay costs ($8.3B) equals $2.739B. $2.739 billion divided by 16.4M flights gives a crew cost delay per flight of $167, on average. To get the ratio of crew cost delayed to total delayed, we will divide $167 by $2,750 (i.e. 6%). If a flight is delayed, the airline can expect to increase its crew cost for that flight by 6% on average.

Maintenance: A4A estimates that the maintenance cost takes up sixteen percent of total delayed costs. 16% of $8.3B comes out to $1.328B. That, divided by our 16.4 million flights, equals $80. On average, a flight will have a maintenance cost of $80 due to delay. According to the FAA, maintenance costs airlines $1,000 per flight. This means we divide $80 by $1,000 and get 8%. If a flight is delayed, the airline can expect to increase its maintenance costs for that flight by 8% on average.

Aircraft Ownership: A4A says that aircraft ownership constitutes thirteen percent of total delay costs. 13% times $8.3B equals $1.079B. We also know that ownership makes up 17% of total U.S. airline direct operating costs of $40B. $40B times 17% equals $6.8B. $1.079B ownership delay costs are a part of that total ownership cost of $6.8B, which comes out to 15%. If a flight is delayed, the airline can expect to increase its ownership cost for that flight by 15% on average.

On average, a flight delayed can have an estimated cost of:



**VI. CONCLUSION**

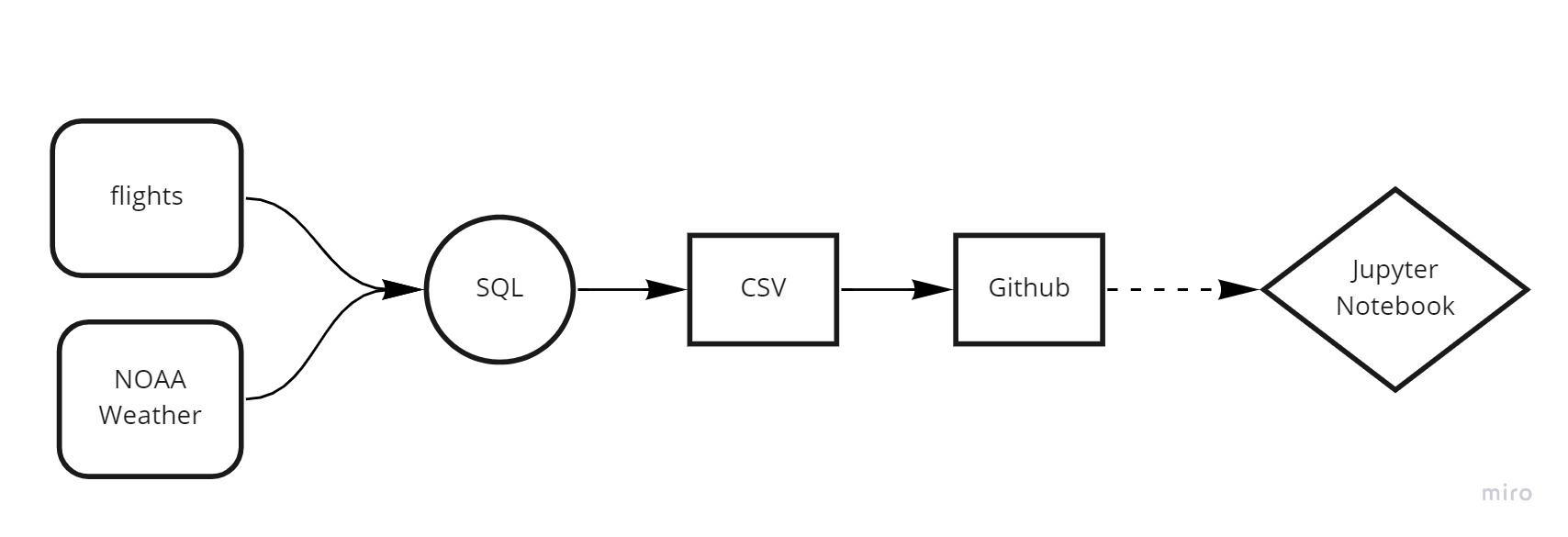
From the variables inputted into our model, we find that 36% of flights are delayed, on average. The cost of a delayed flight, on average, will depend on the individual flight’s planned variable costs. However, each airline can use the economic model pictured above to determine the delay cost. This is the major takeaway for an airline business person.

The major takeaways for us were learning how to do hypothesis testing and modelling in Python, and understanding the tradeoff between recall and precision (and what best fit our project). Moving forward, we would like to test more variables to see what else can contribute to our model. We would also like to develop a model that increases both recall and precision. To really continue with this topic, we would also look into flights that are ahead of schedule. We would determine the likelihood of a flight being ahead of schedule and the positive resulting economic impact.

**\*Every part of this project (data, scripts, essay, slideshow) can be found on the project’s Github:**

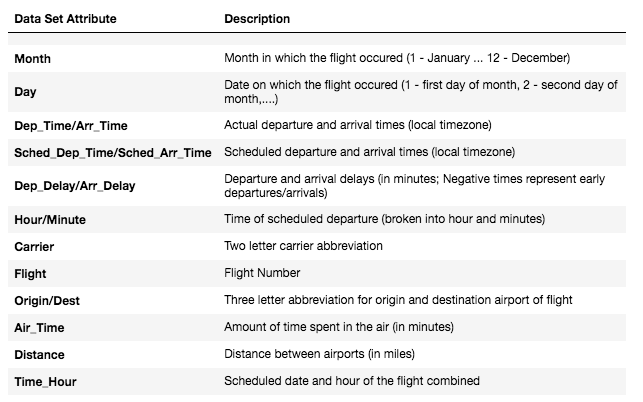
https://github.com/MarlaGoodman/CMS-Project

**Data Architecture:**

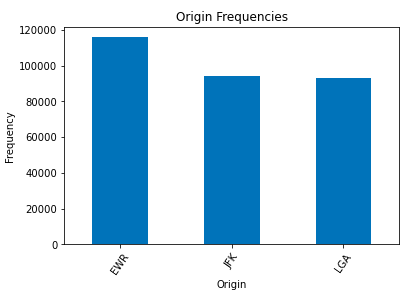
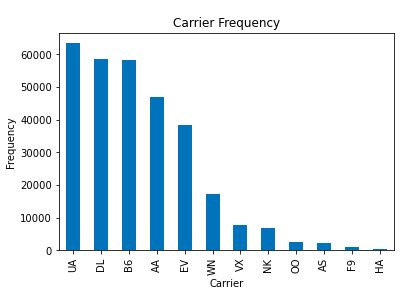


**Appendix: Exploratory Data Analysis**

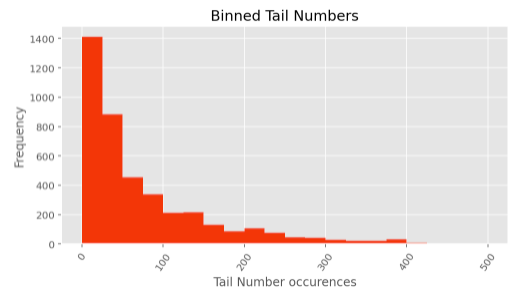
1. Data Dictionary



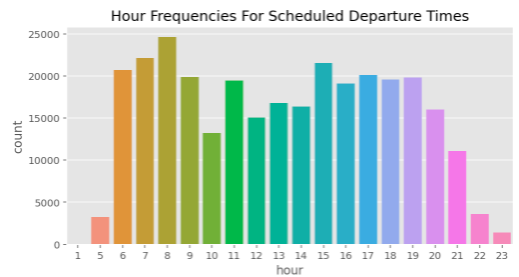
II.1 II.2



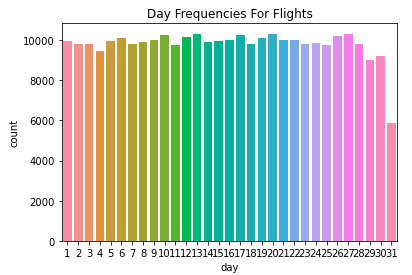
II.3



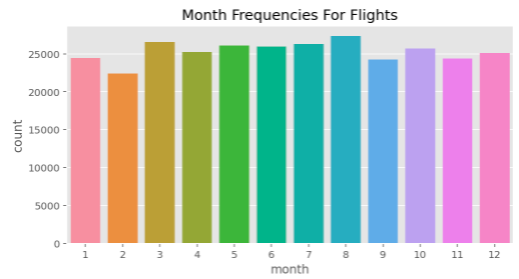
II.4



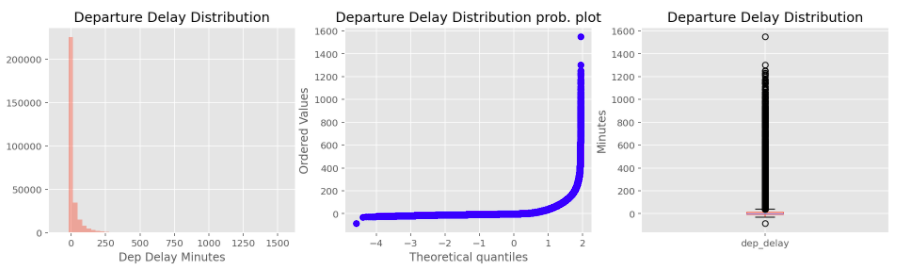
II.5



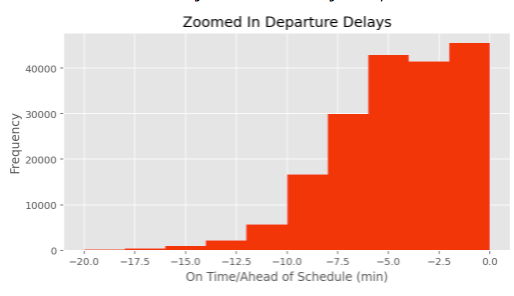
II.6



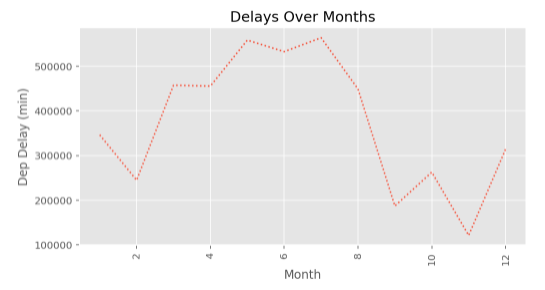
II.7,II.8,II.9



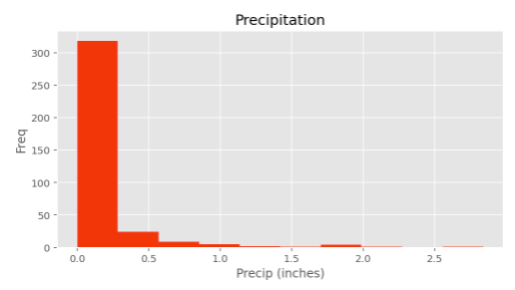
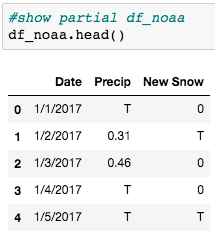
II.10



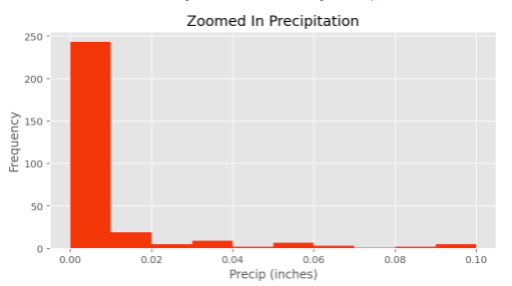
II.11



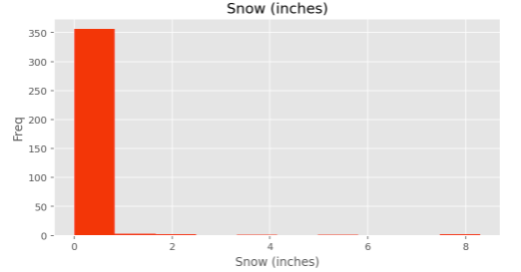
II.12 II.13



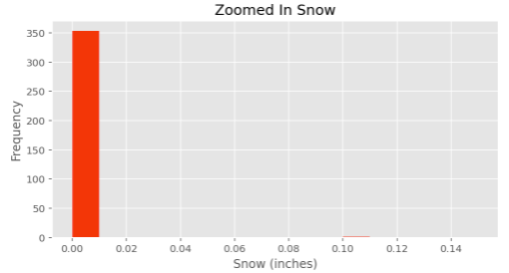
II.14



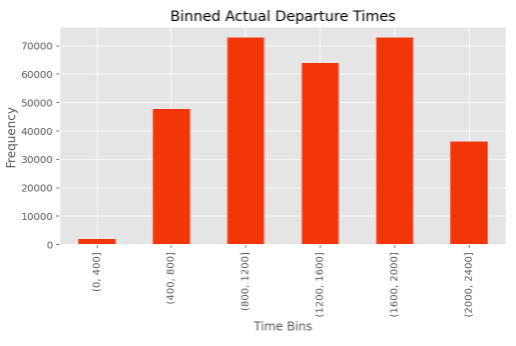
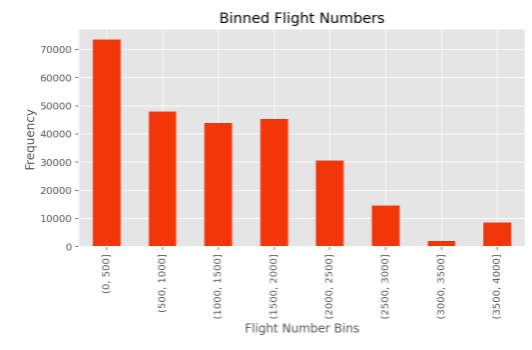
II.15

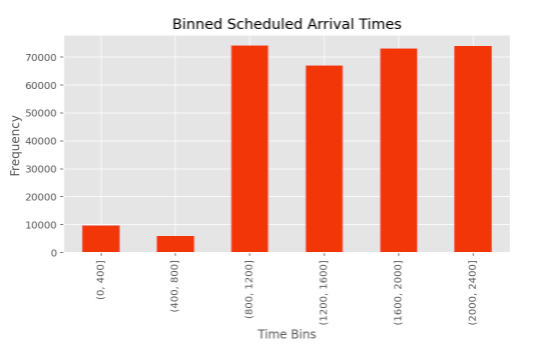
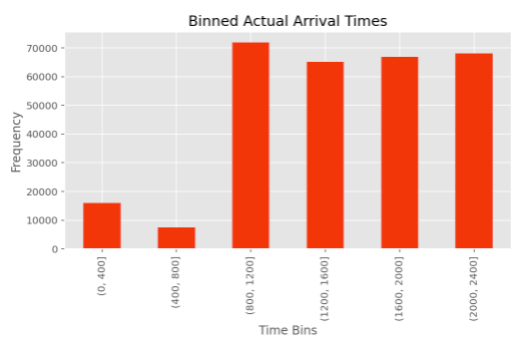


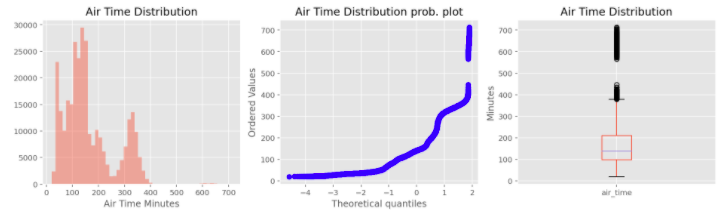
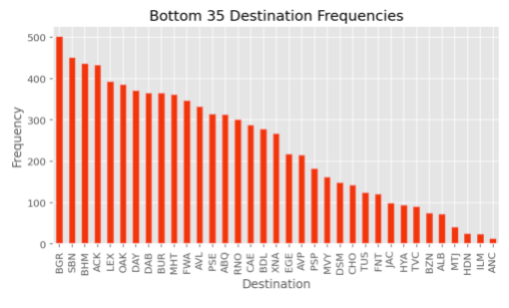
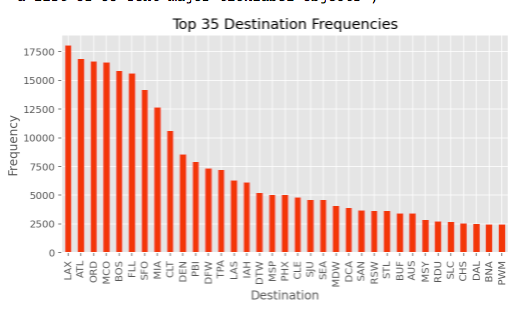
II.16

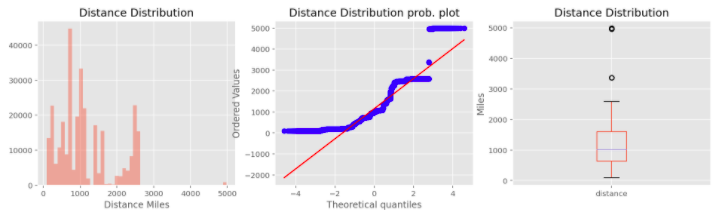


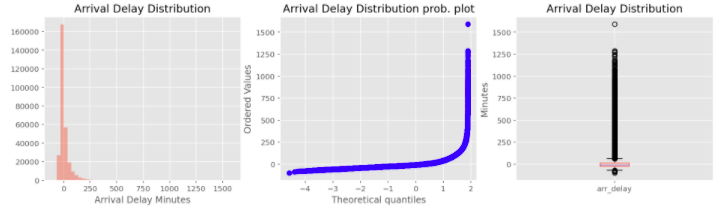
Other:

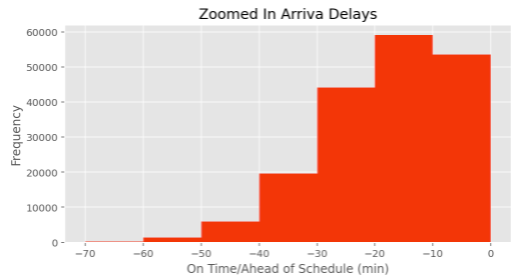








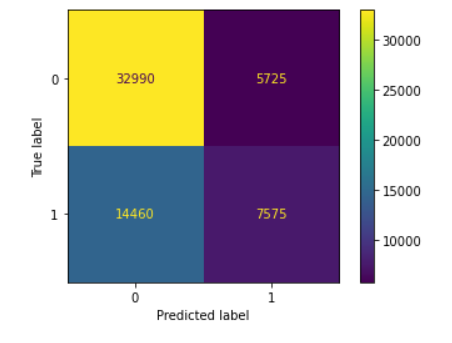




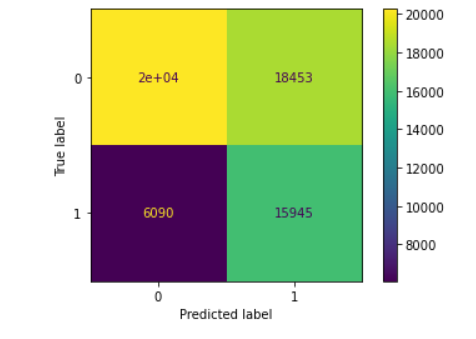
**Appendix: Model Confusion Matrices:**

(0 = on time, 1 = delayed)

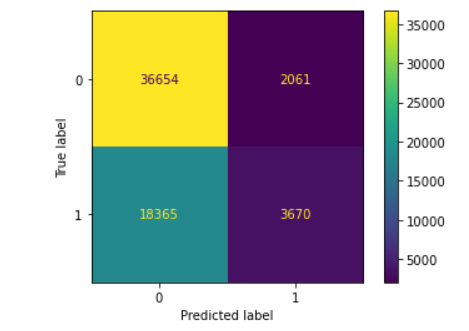
1. Logistic Regression:



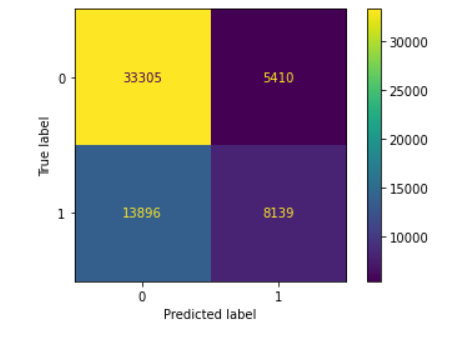
2. Random Forest:



3. XGBoost



4. Ensemble



Sources for Economic Model:

<https://www.statista.com/statistics/978677/operating-expenses-united-airlines/>

<https://www.statista.com/statistics/197690/us-airline-fuel-consumption-since-2004/>

<https://www.airlines.org/dataset/per-minute-cost-of-delays-to-u-s-airlines/#>

<https://www.airlines.org/data/annual-u-s-impact-of-flight-delays-nextor-report/>

<https://www.airlines.org/argus-us-jet-fuel-index/>

<https://www.airlines.org/dataset/a4a-quarterly-passenger-airline-cost-index-u-s-passenger-airlines/#>

<https://www.faa.gov/regulations_policies/policy_guidance/benefit_cost/media/econ-value-section-4-op-costs.pdf>

<https://www.faa.gov/air_traffic/by_the_numbers/>

<https://catsr.vse.gmu.edu/pubs/Delay_Cost_Model_US_Ferguson.pdf>

<https://isr.umd.edu/NEXTOR/pubs/TDI_Report_Final_10_18_10_V3.pdf>

<https://www.icao.int/mid/documents/2017/aviation%20data%20and%20analysis%20seminar/ppt3%20-%20airlines%20operating%20costs%20and%20productivity.pdf>

<https://transportgeography.org/contents/chapter5/air-transport/airline-operating-costs/>

<https://www.transportation.gov/administrations/assistant-secretary-research-and-technology/what-cost-airline-fuel-means-you>