Technical Report

Talia Willcott, Stephen Duval, David Perhai, Meghan Roffler

Introduction:

For this project, we examined 2021 flight data in conjunction with weather events to explore patterns in cancellations and delays. The first dataset is from the Bureau of Transportation Statistics and tracks passenger flights across the United States. Our second dataset is from Kaggle and includes weather reports taken from 2,071 airport-based weather stations across the United States. We decided to only look at the data from 2021 for both datasets due to the magnitude of flights and weather events.

This project explores how weather affects the efficiency of flights by examining the timeliness (whether flights are delayed) and consistency (whether flights are cancelled) of the flights based on the current state of the weather. As a stretch goal, we hoped to use machine learning to develop an app-like program that could learn from our 2021 flight and weather data in order to determine whether a flight is cancelled or not. A further stretch goal would be a model that accurately predicts the length of delays. Overall, our results show that weather events are not sufficient to predict the length of flight delays. However, weather seems to be an important feature in determining whether or not a flight is cancelled.

Technologies:

The following is a list of technologies used in this project.

1. Python: using JupyterNotebooks and Azure DataBricks, Python and Pyspark were used to clean and transform the bulk of our data.
2. Azure:
    a. DataLakes: Used for storing the raw files to be used in our ETL
    b. Data Factories: Create a pipeline to return api call results to a SQL database
    c. Data Studio: Used to transform our data into proper CSVs
3. PowerBI: Used for creating reports and a dashboard

Main Questions:

Below is a list of questions we were initially interested in asking. These questions guided the direction of our project:

1. Are there regional patterns for delays and magnitude of delays?
2. Do different states suffer greater delays for the same type of weather events?
3. What weather type causes the most delays? Per state?
4. Which airlines have the fewest and most delays?
5. Do certain airports have more delays on average? Is it correlated with weather or a different delay factor?
6. How much seasonality will we see in the data? How much do holidays affect delays and cancellations?
7. Which factors will have the greatest predictive power in terms of delays?

Data:

1. Bureau of Transportation Statistics (2021 Flight Data)
    a. To download the data, visit the above page and choose the desired fields for monthly flight data. Then download as monthly csv files
    b. Once we merged all 12 months of flight data, our 2021 dataset was 52 columns and over 6 million rows
2. Weather Data from Kaggle
    a. This dataset includes an airport code associated with each weather event (as events are recorded at airports across the country) which was crucial in joining to our flight dataset.
    b. The raw data for 2021 was 15 columns and over 1.2 million rows
3. Current flight data from FlightRadar24
    a. Acquired through an API key.
    b. An account with RapidAPI was necessary
    c. Query1 parameters: List of flights by airlines for given day
    d. Query2 parameter: In-depth flight information using the list of flight ids obtained in Query1

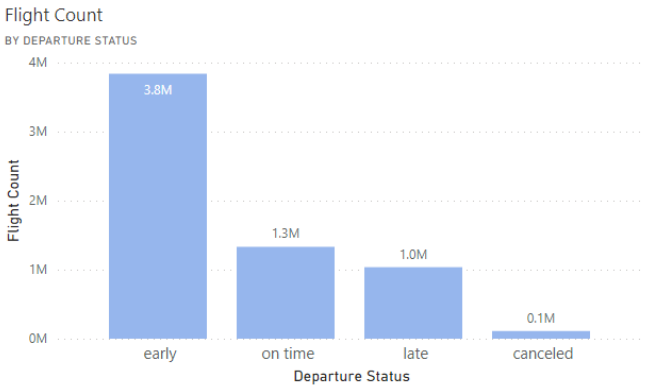
Data Processing:

Below are lists outlining how we processed our live flight data and our weather and flight data.

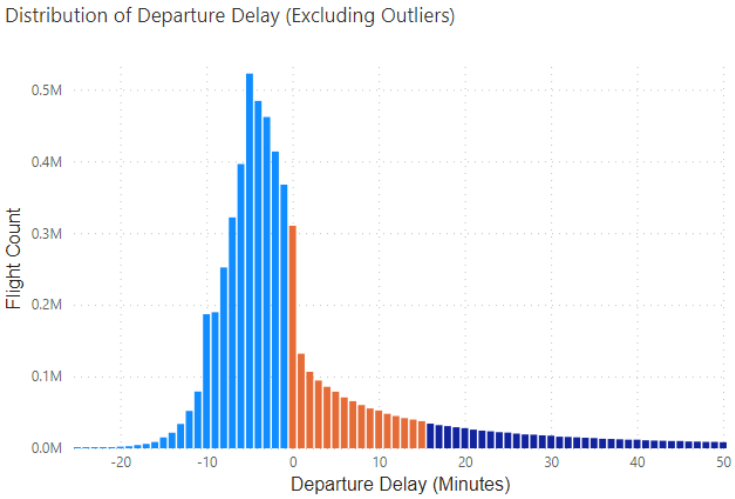
- Weather and Flight CSVs
    - Appended the 12 months of flight data
    - Renamed columns to more logical names (see data dictionary)
    - Cast the columns:['scheduled_departure_time' , 'departure_time' , 'takeoff_time' , 'landing_time' , 'scheduled_arrival_time' , 'arrival_time'] as integer
    - Format the integers as a udf format
    - Concatenate the 'flight_date' columns with the udf format columns, then convert to timestamp
    - Use the timestamps to offset timezones to standard UTC time
    - If a flight is in the air overnight, add a day to the landing date
    - Join weather events to flight data if a weather event occurred during the flight duration
    - Write the complete dataframe to a csv stored in an azure container
- FlightRadar API
    - Call the api with desired parameters
    - Get flights by airline and append to a list
    - Iterate through the list of flights and extract the flightID's
    - Use the flight ID's to get detailed information on each flight
    - Convert the flight information to a DataFrame, and drop unwanted columns
    - Convert unix times to GMT times
    - Calculate the delay times of the day's flights
    - Output the data as a json object to an Azure Storage container

Data Analysis/EDA:

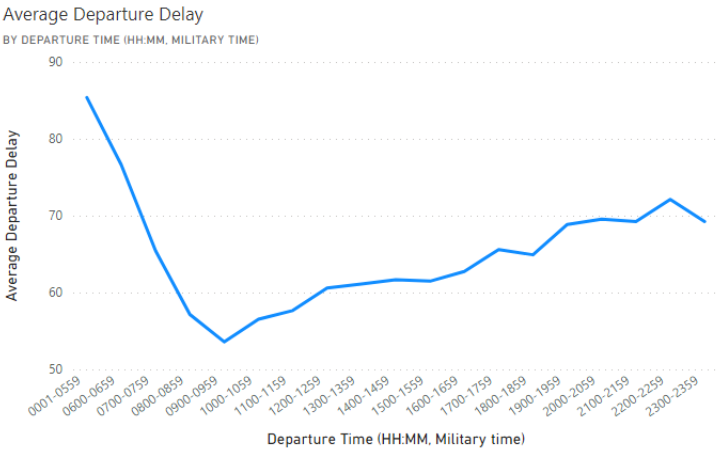
- In total, there were considerably more early flights than delayed/cancelled



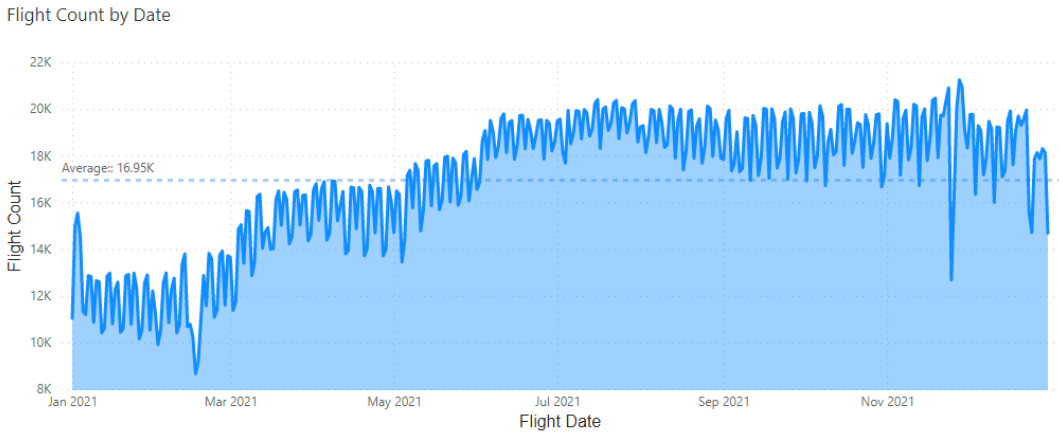
- Mosty flights center around an arrival between 15 minutes early to 15 minutes late



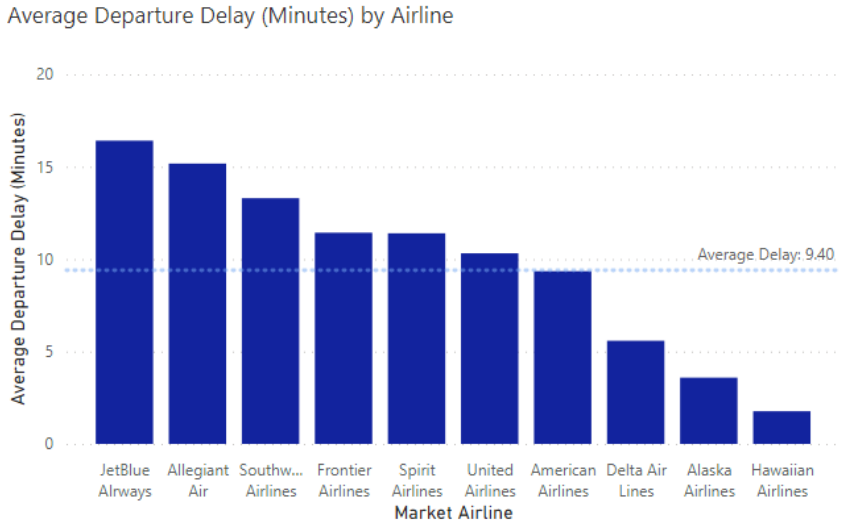
- Average delay time increases throughout the day after 10AM and is the longest from midnight to 6AM



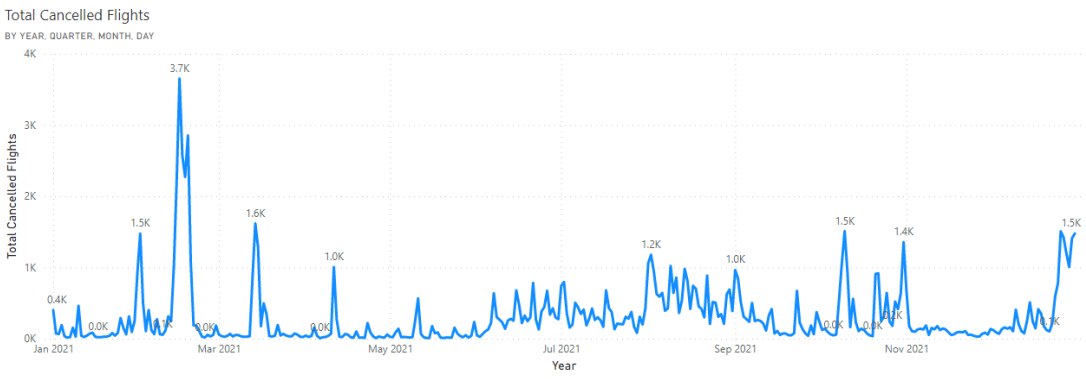
- There are more daily flights in the second half of the year than the first, peaking in Q4. (Likely due to COVID)



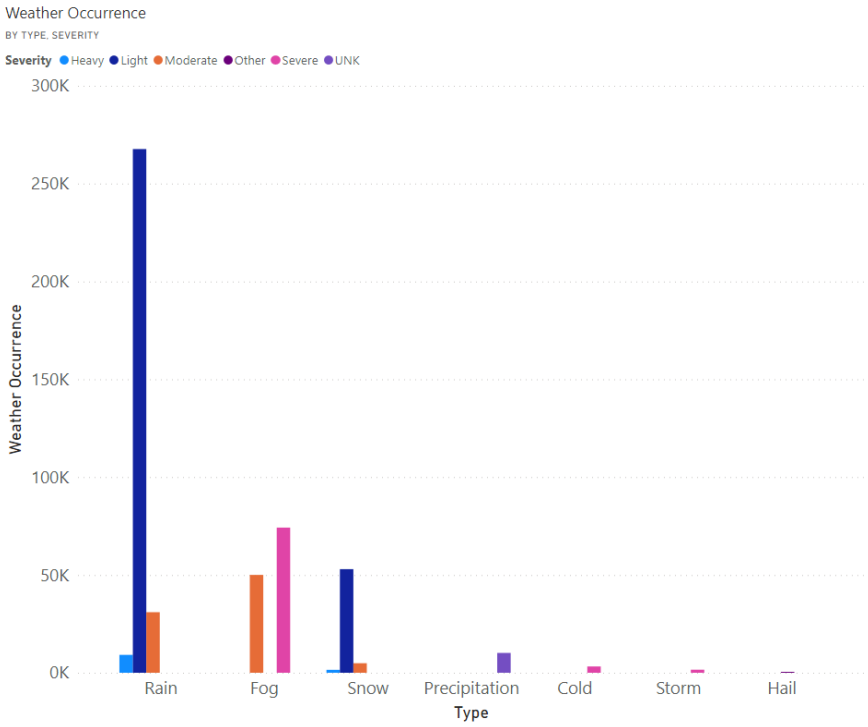
- JetBlue airways had the worst performance in terms of average departure delay, while Delta performed the best of major US airlines



- A polar vortex in mid-February led to an unusual amount of cancelled flights on 2/15/2021

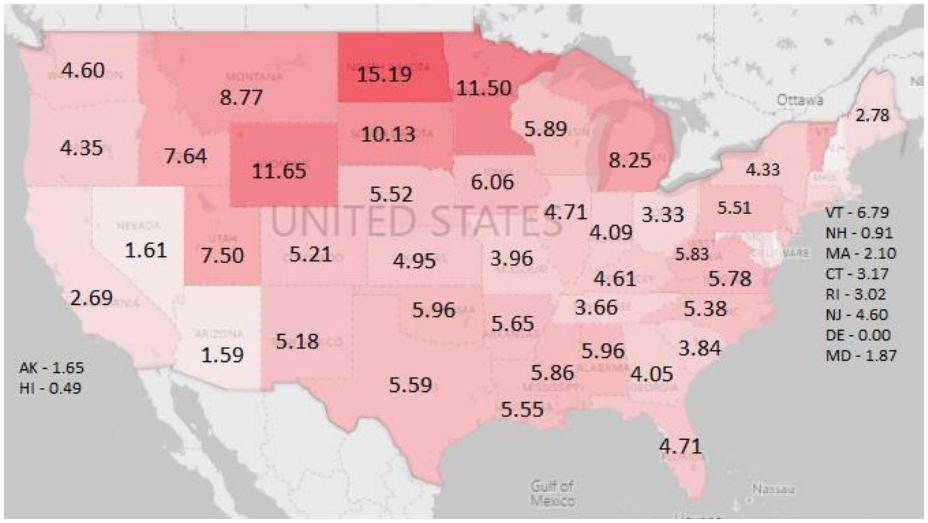


- Rain was the most common weather event. There were more severe weather events than moderate weather events, and more light weather events than both

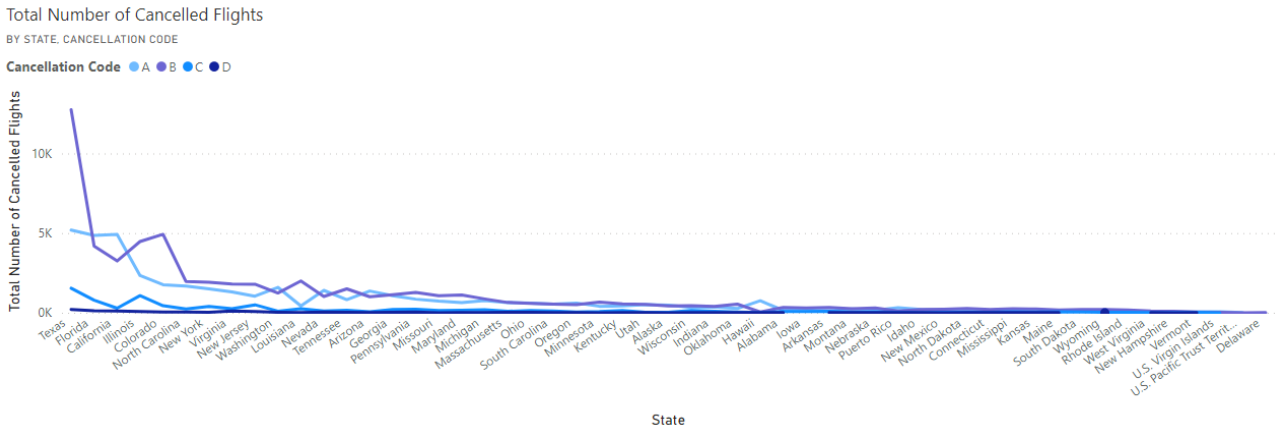


Findings:

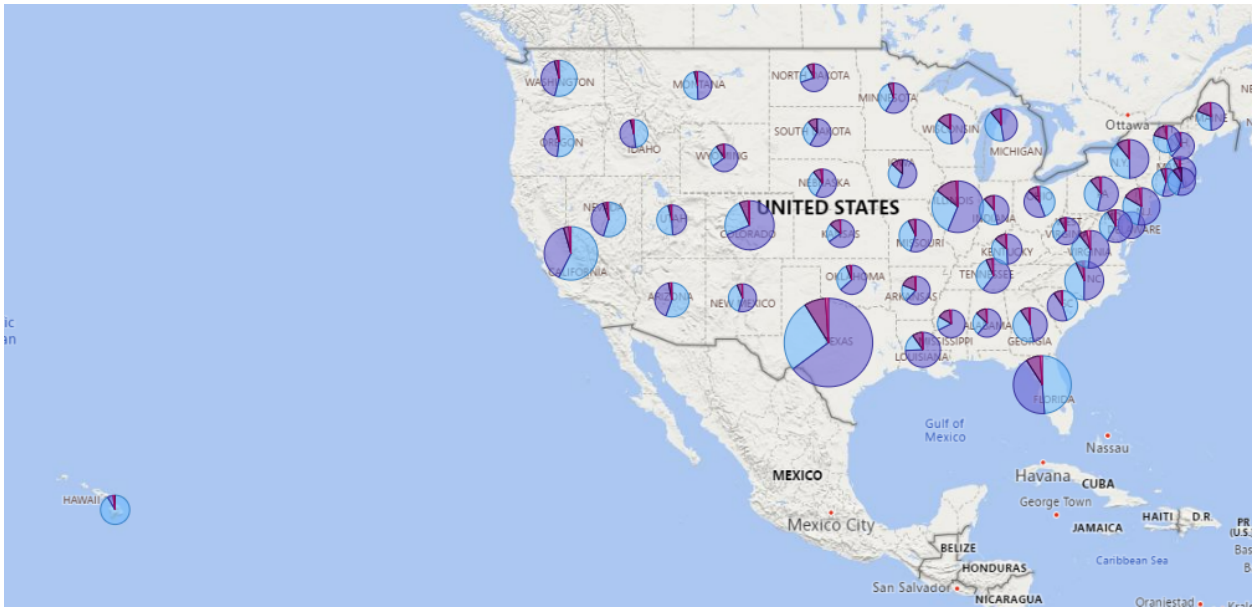
Figure 1. Average weather delay (in minutes) by State



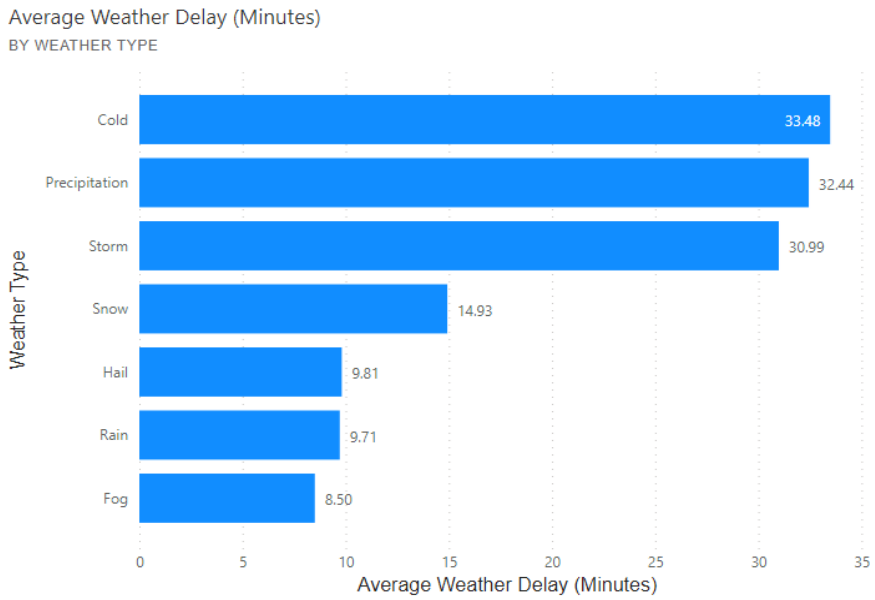
As seen in figure 1, there seems to be regional (state) variations in average departure delay. The range of average departure delay goes from 1.59 minutes at the lowest to 15.19 minutes at the highest, although most states seem to have a value near 5 minutes.



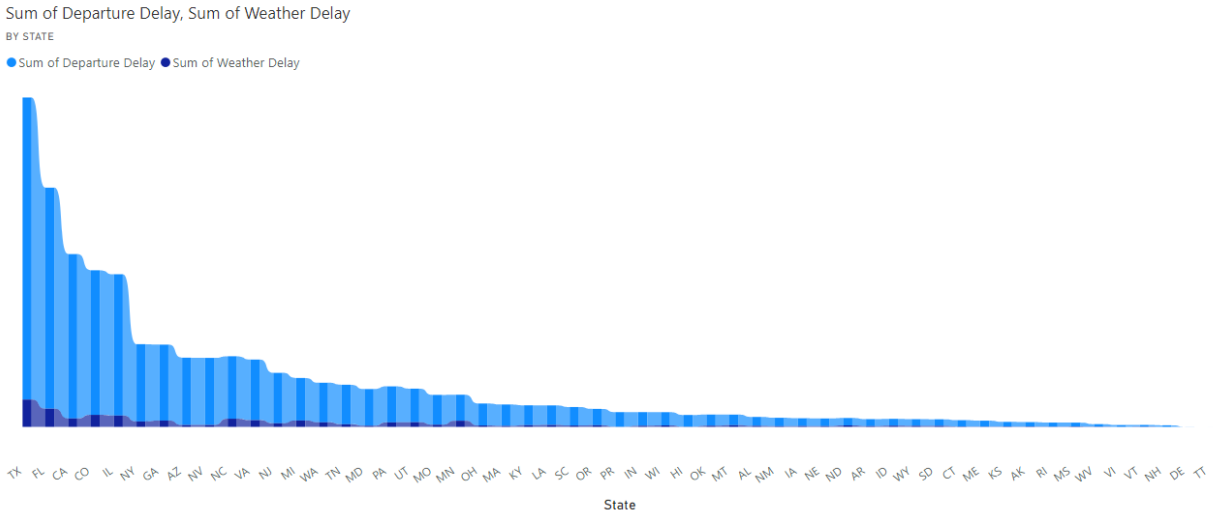
As shown in the visualisation, the highest number of cancelled flights is in Texas and it is due to weather delays (cancellation code B). Due to the winter storm in February 2021, the count of flight cancellations in Texas spiked and this graph illustrates that. The top two and three states with the most cancellations are Florida and California which is due to carrier issues (cancellation code A).

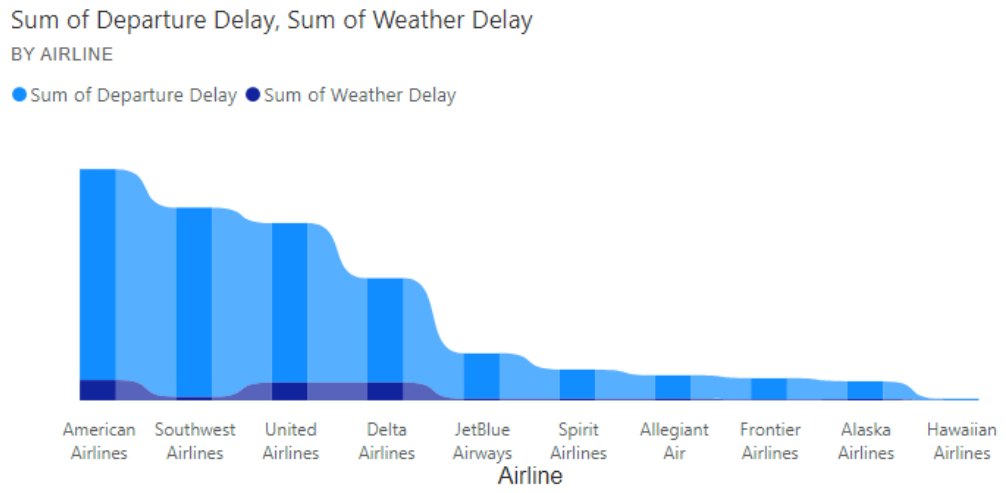


Another way to view the previous statistics is by looking at a map. In this map, the size of the bubble represents the number of cancelled flights and the pie charts in each state show the percentage of cancellation codes. The states with the most cancellations are Texas, Florida and California and the reasons for cancellation are weather, carrier and carrier.

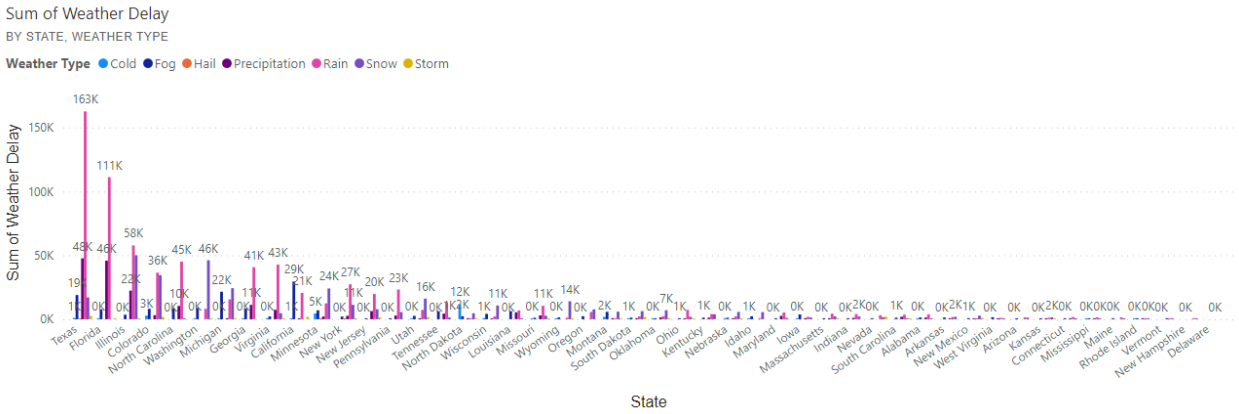


As seen in this illustration, the average weather delay is the highest for cold, precipitation and storm weather events. Fog, rain, and hail have the lowest average weather delay. This is likely due to the average lengths of these weather events (hail and fog usually don't last as long as cold snaps and storms).

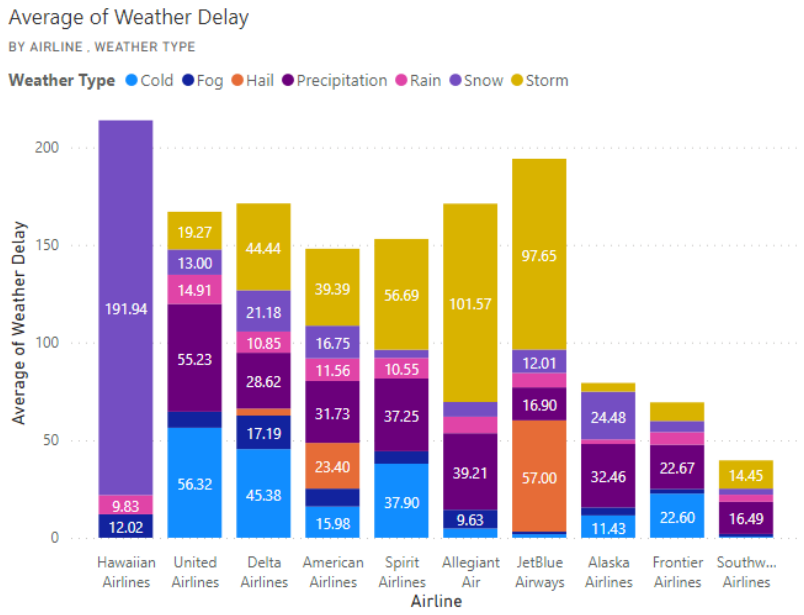




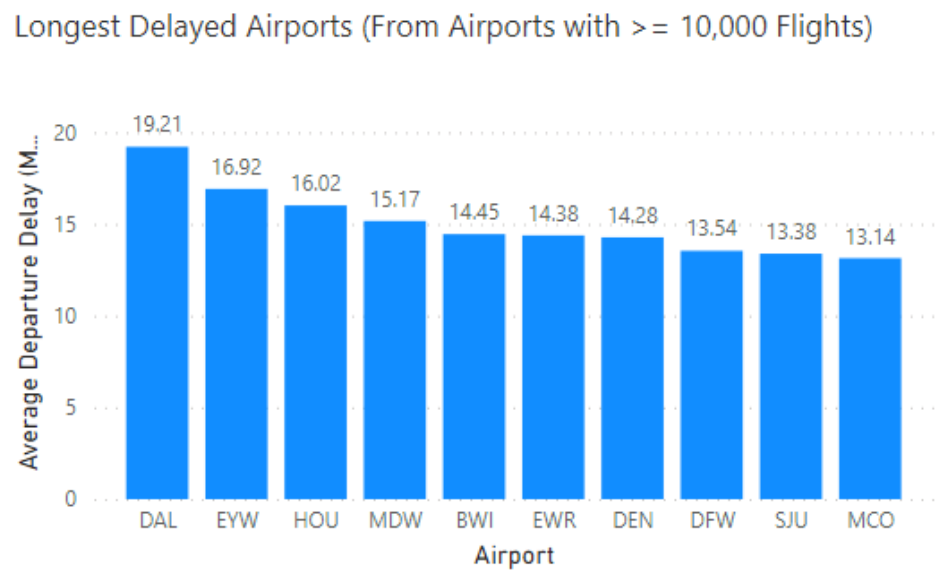
The two visualisations above show total delay in light blue and weather delay in dark blue. The first figure explores how this relationship changes by state while the second figure explores how it changes by airline. A general observable trend is that weather delays increase as delays increase, but are much less frequent.



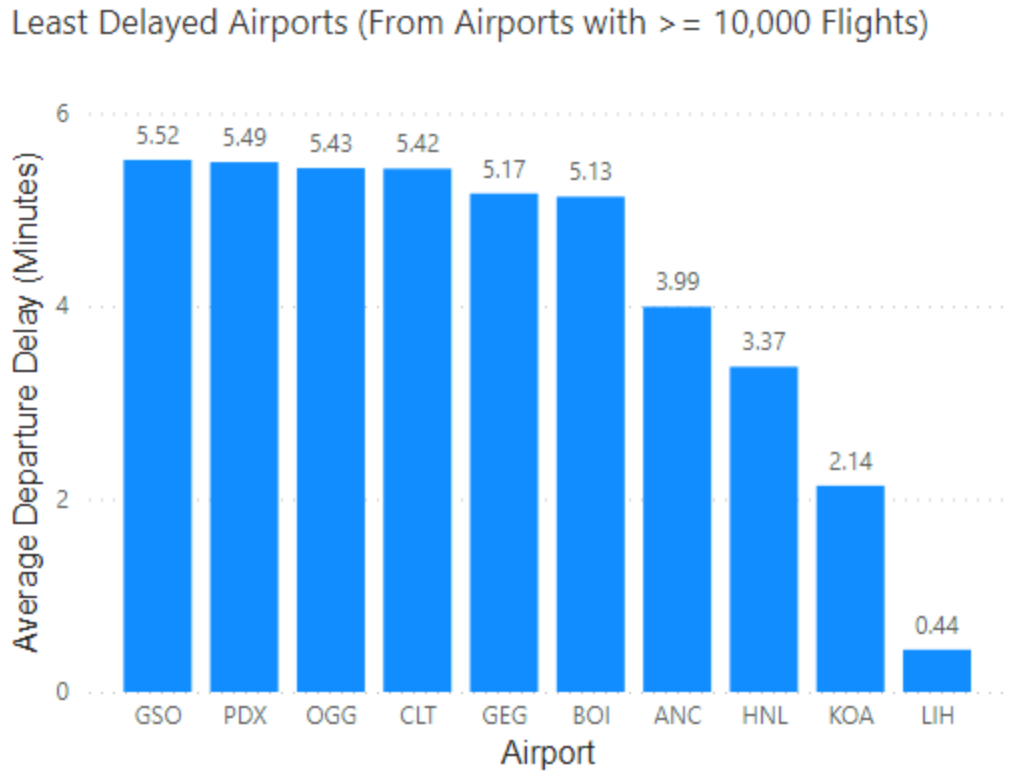
This visualisation above allows us to see which states have the most weather delays and what weather causes these delays. As shown, rain seems to be the leading cause of weather delays for most states. In terms of weather susceptibility, Texas seems most affected by weather events. However, it is important to keep in mind that this graph shows the sum of all delays and so Texas's high values can be due to the large amount of airports in the state.



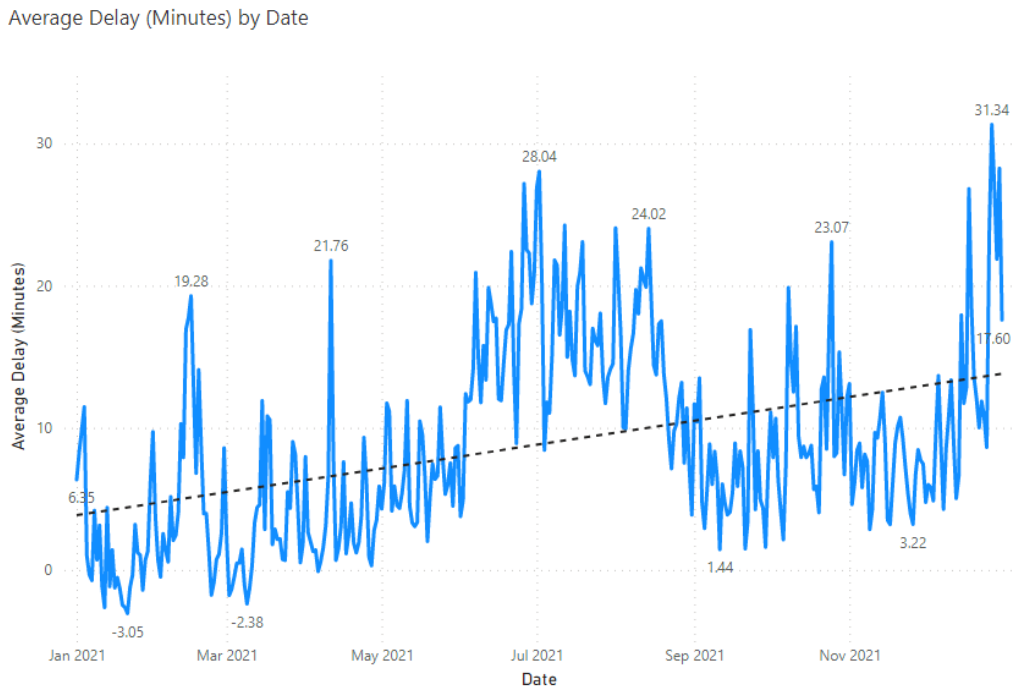
This illustration shows the average weather delay by airline and type of weather. This allows us to see which kinds of weather cause the most delays for each airline. It is interesting to note that Hawaiian airlines seem to experience the longest delays from fog where other airlines do not. JetBlue and Allegiant experiences the longest delays from storms.



Of the 10 airports listed above, only SJU (San Jose, Puerto Rico) and EYW(Key West) would be classified as major US airports. The other two lie in tropical areas with unpredictable weather and high precipitation and severe rainstorms. Weather is a strong predictor of delays, so it is unsurprising that tropical storms would affect the average delays of these locations. While it is not true that all major airports suffer longer delays on average than their smaller counterparts, another factor that affects flight delays is airport traffic.



Of the 10 airports listed above, only Charlotte services more than 100,000 flights per year. Interestingly, Charlotte is the only airport not located on the Pacific side of the United States. It stands to reason that smaller airports would have smaller average delays, due to less air traffic and fewer opportunities for delays to arise from events other than weather.



Lastly, this graph shows that average delays seem to fluctuate throughout the year. Overall though, average delays are getting longer as the year goes on, peaking around New Years.

Machine Learning

This section outlines the models used to predict flight cancellations and delays. This project has two predictive goals:

1) whether a flight will be cancelled or not and
2) whether a flight will be delayed or not.

For both predictions, we chose to use a Random Forest Classifier. Due to the nature of our combined flight and weather data, it was necessary to create four separate data sets and four corresponding models. The data was divided based on whether there was a flight event at the origin or destination at the time of scheduled departure. They are outlined in more detail below:

- Set 1: Flights with weather events occurring at both origin and destination at scheduled takeoff.
- Set 2: Flights with a weather event occurring at origin at scheduled takeoff.
- Set 3: Flights with a weather event occurring at destination at scheduled takeoff.
- Set 4: Flights that had no weather event occurring at either location at scheduled takeoff.

For both our models, we used the following predictor variables:

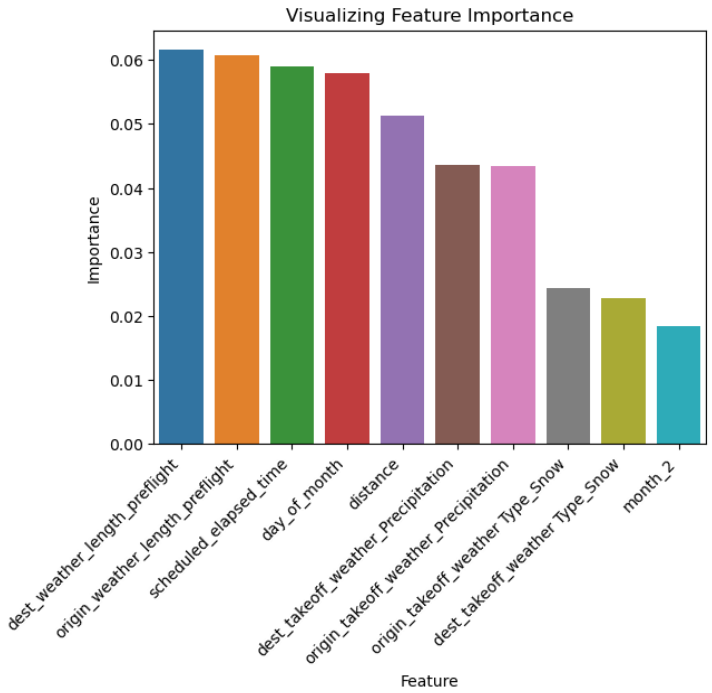
- Month of flight (dummy)
- Day of week (dummy)
- Marketing Airline Code (dummy)
- Operating Airline Code (dummy)
- Origin Airport Code (dummy)
- Destination Airport Code (dummy)
- Departure Time Block (dummy)
- Arrival Time Block (dummy)
- Day of month (int)
- Scheduled Elapsed Time in mins (int)
- Distance in miles (int)
- Origin Weather Severity at Scheduled Takeoff (0-4)
- Destination Weather Severity at Scheduled Takeoff (0-4)
- Origin Precipitation at Scheduled Takeoff in inches (float)
- Destination Precipitation at Scheduled Takeoff in inches (float)
- Minutes Weather Event has been going at Origin before Scheduled Takeoff (int)
- Minutes Weather Event has been going at Destination before Scheduled Takeoff (int)

A train-test split of 85/15 was used due to the size of the datasets involved, as well as a random state of seed 50 for consistency. Results were as follows:

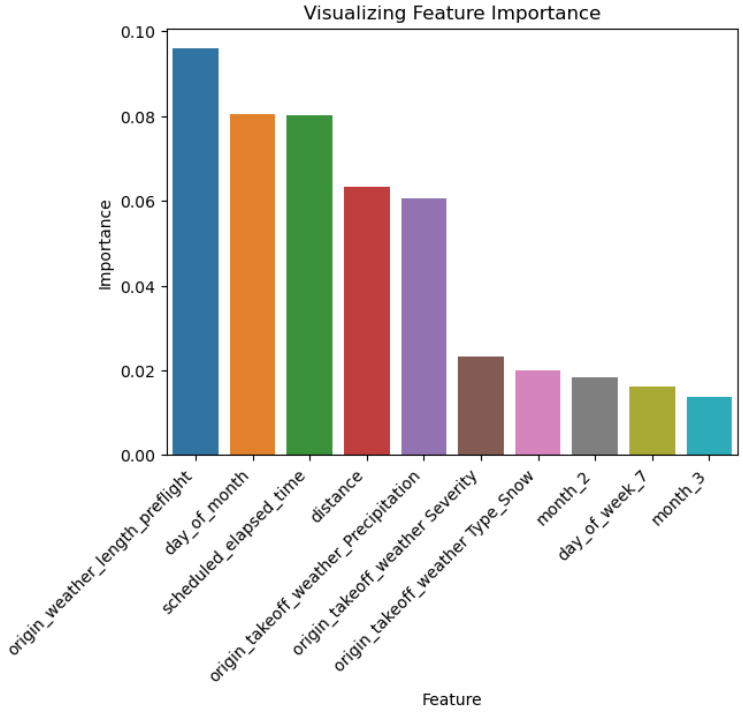
<u>Cancellation Prediction</u>

- Model with weather events at both locations

- AUC: 0.994
- Cancelled/Not_Cancelled Proportion: 7.2%

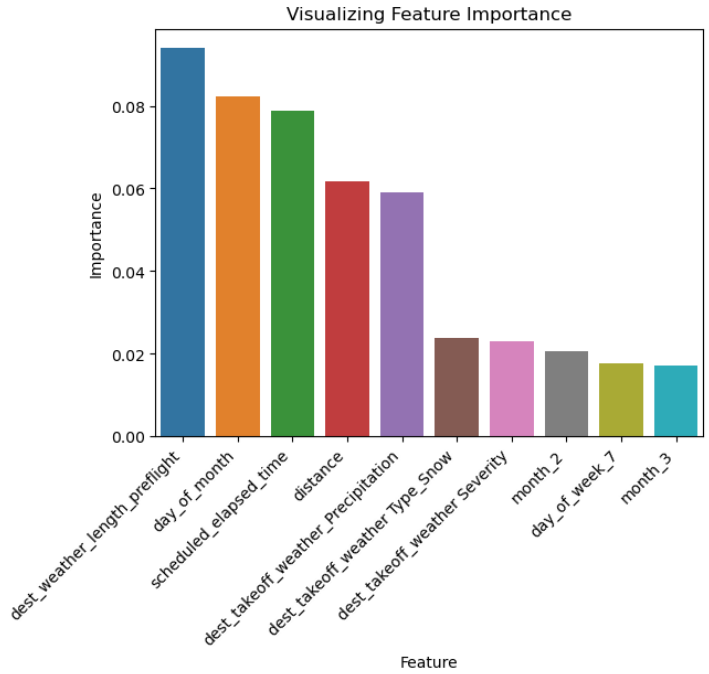


- Model with weather event at origin
  - AUC: 0.956
  - Cancelled/Not Cancelled Proportion: 4.1%

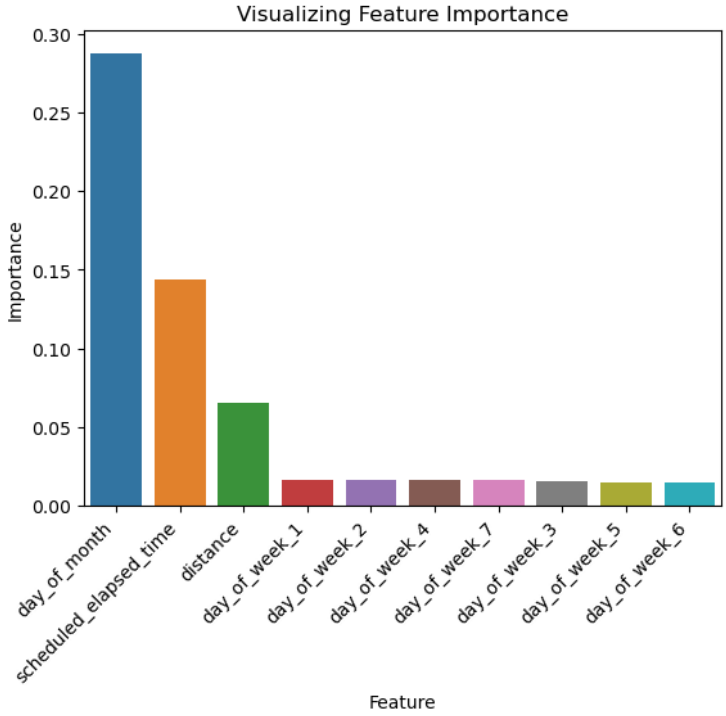


- Model with weather event at destination
  - AUC: 0.707

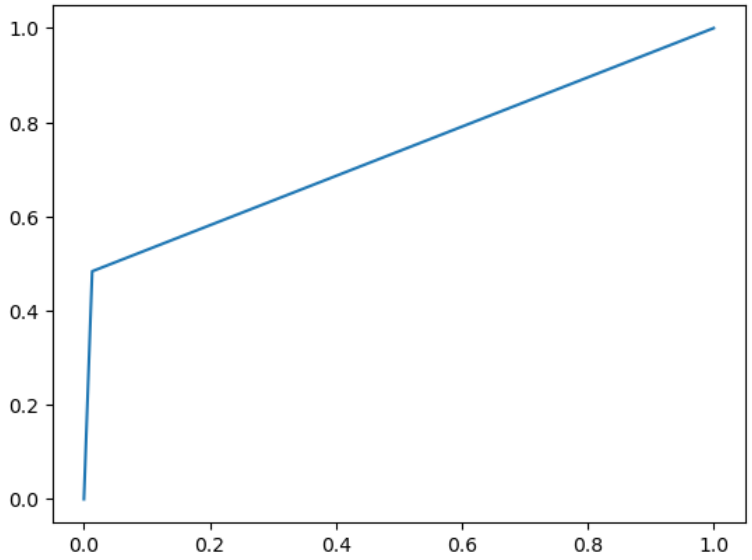
- Cancelled/Not Cancelled Proportion: 3.9%

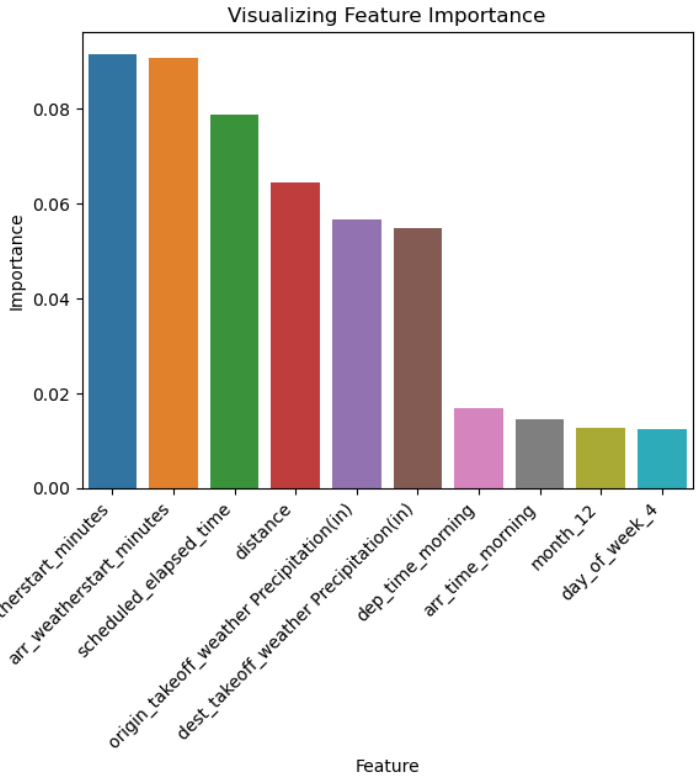


- Model with no weather event at either location
  - AUC: 0.579
  - Cancelled/Not Cancelled Proportion: 1.3%

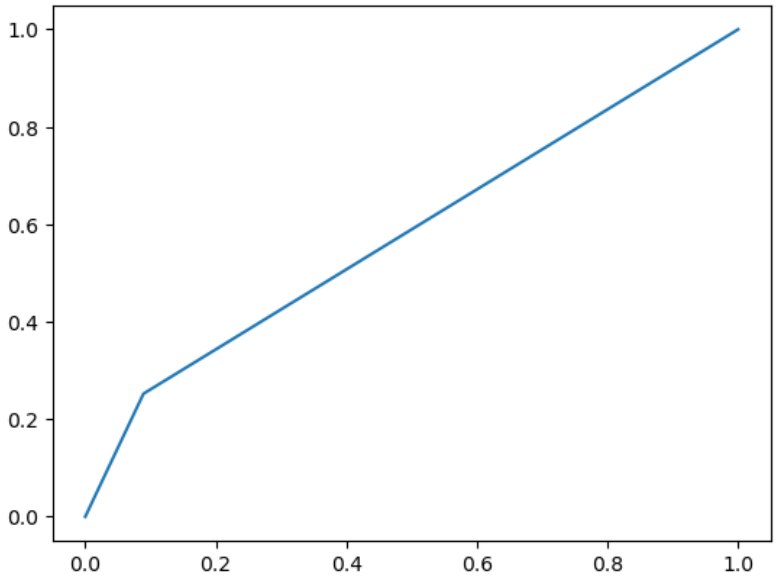


Delay Prediction

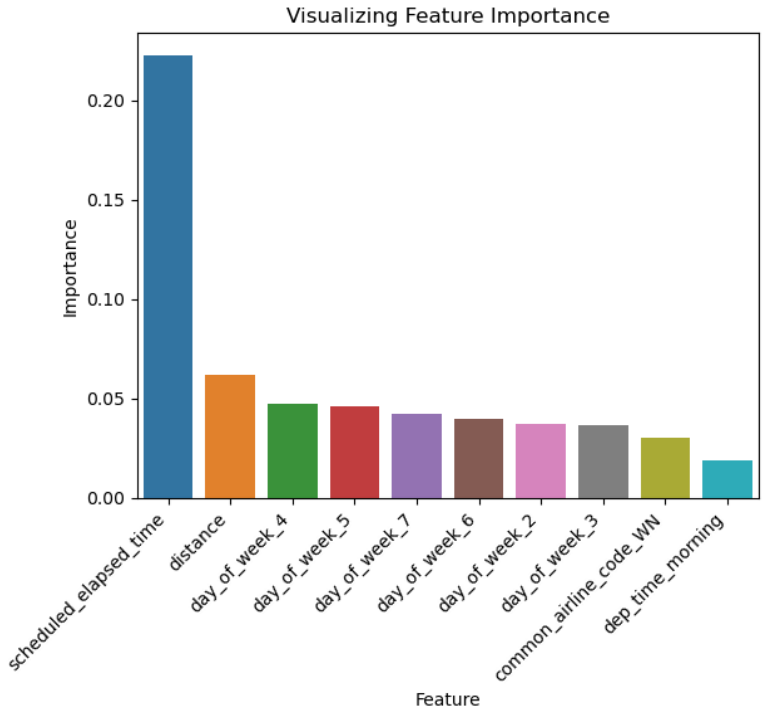
- Model with weather events at both locations
  - AUC: 0.735
  - 

  - Most important features: Departure weather start length, arrival weather start length, and scheduled elapsed time
  - 

- Model with no weather events at either location
  - AUC: 0.582



- Most important features: scheduled elapsed time, distance, day of week



It was surprising that day of the month was of high feature importance! But it is clear by the AUC's that the datasets that had flights with weather events were far more accurate in both cancellation and delay predictions. It's even clearer that weather, and specifically weather at the origin, makes a big difference in if a flight gets cancelled or not. The length of time the weather event had been going on before the flight seems to be a huge factor as well in our predictions.

Conclusion:

We find that weather seems to be an important factor in making flight predictions. Specifically, when there is a weather event in both the origin and destination, the flight is 5.5 times more likely to be cancelled than when there is no weather event. Furthermore, having a weather event in our prediction made our model 3.5 times more accurate in correctly predicting whether a flight is cancelled. The additional features that were important in our prediction models are scheduled elapsed time and distance of flight.

However, it is important to keep in mind that our project only looks at 2021 data. The arctic storm that streaked across the southern states certainly affected some of the conclusions we could make from this dataset. COVID also significantly decreased the number of flights in the earlier half of the year.

The next steps for our project include:

- Analysing several years of flight and weather data
- Fine tuning the delay prediction to be able to predict delay lengths
- Creating a data factory capable of handling and storing live weather and flight data to be used in machine learning models
- Creating an app or widget that houses live flight predictions for consumer and airline use