Predicting flight cancellations in advance with machine learning models

Abstract

Flight cancellations affect travelers, airline companies and the economy. While financial loss and loss of time are the biggest losses to travelers. Airlines additionally experience significant reductions in passenger numbers that have resulted in planes flying empty between airports, which in turn massively reduce revenues for **airlines** and force many **airlines** to lay off employees or declare bankruptcy. In the end the economy suffers from these revenue cuts as well. This project aims at building a model that predicts flight cancellations in advance and will serve as an important tool to save travelers from losses associated with flight cancellations. To achieve this goal, the datasets from 2015 Flight Delays and Cancellations (kaggle.com) downloaded from Kaggle with three files; airlines.csv, airports.csv and flights.csv was used together with the airlinedelaycauses_DelayedFlights.csv also downloaded from Kaggle and a supplementary dataset downloaded from Open data @ OurAirports. Using Python tools like matplotlib, pandas and numpy, these datasets were cleaned, missing values were handled, columns irrelevant to the project were dropped and abbreviations were replaced with full names in the datasets. All three datasets were merged into one dataset after they were cleaned. The resulting dataset had 7755837 rows and 42 columns which were analyzed and visualized using seaborn and matplotlib. Exploratory data analysis revealed that more flights were cancelled in November, January and February. This may be due to harsh winter weather conditions. American Eagle Airlines Inc was responsible for the most flight cancellations at a cancellation rate of 3.5%. Security and Weather delays were most likely to cause flight cancellations and Southwest Airlines Co. has the most patronized flights. All delay features, categorical features except Airlines and IATA CODE, id, ident, Cancellation reason, type, name, tail number, latitude_deg, longitude_deg, elevation_ft, iso_country, iata_code were all dropped. The Airline and IATA CODE were encoded, and the dependent variable 'CANCELLATION' was used to define X and y variables. Train and test sets were split in 80: 20 ratio and three predictive machine learning models were built. These were Linear Regresssion and Decision Tree Classifier which were linear models and a third nonlinear model, Random Forest model. The two linear models were evaluated using R-squared scores, mean absolute error and mean squared error. Of these two the Decision Tree Classifier performed best with an R-squared score of 0.99 compared to R-squared score of 0.64 of the Linear Regression models. The Random Forest Classifier was assessed by calculation accuracy and F1 score. The accuracy and F1 score of of this model were 1.0. Further assessment of this model included the generation of confusion matrix and classification reports. These reports indicated that the model performs better than the initial linear models. Although classification report indicated class imbalance, a precision score of 1.0 was enough to show the model was performing very well. The generation of ROC curve with area under the curve as 1.0 further attests to the very good performance of the Random Forest Classifier Model. The feature importance of the model was calculated using the coefficients and it was shown that the most important features of the Random Forest model were taxi out, elapsed time and arrival time.

INTRODUCTION

Flight cancellations have become unavoidable in-flight travels. The effects of such cancellations lead to travelers losing some money due to missing cruises, prepaid hotel rooms, increased parking fees, cancellation or postponement of business trips and sometimes the additional cost of unplanned air tickets.

Travelers also miss vacation and valuable time with family and friends because of these cancellations.

Building a model that predicts flight cancellations will serve as an important tool that could be used by all air travelers to predict flight cancellations in advance and to prevent loss associated with flight cancellations. It will also give air travelers ample time to replan meetups, reschedule trips, make safer accommodations choices and prevent them from being stranded due to unplanned flight cancellations.

Problem Statement

What predictive model can be used by air travelers to predict flight cancellations in advance?

METHODS

Data collection

The data sources for this project include three csv files namely,

1. The 2015 Flight Delays and Cancellations (kaggle.com) downloaded from Kaggle. It has about three files airlines.csv, airports.csv and flights.csv that have information about flight delays in 2015. However, all the features of the flights.csv was used for this project. This flights.csv file has 5819079 rows and 31 columns. Specific features in the airports.csv and airports.csv files were added to the flights.csv datasets to replace features with abbreviations with full names in the main dataset.
2. The second file that was used for this project is the airlinedelaycauses_DelayedFlights.csv also downloaded from Kaggle. It is made up of similar features to the first file. This data has information about flight delays and cancellations in 2008. It has 1936758 rows and 30 columns.
3. The supplementary dataset was downloaded from Open data @ OurAirports. It is named the airport.csv and has 80559 rows and 18 columns.

Data Wrangling

The initial step of data wrangling involved the cleaning of the flight delays in 2008 and 2015 separately.

Handling missing columns

Flight delays dataset of 2008 was named, Delay_2008. It had 1936758 rows and 30 columns. Missing columns were handled by dropping irrelevant columns like Cancellationcode column. Missing values were replaced with 0. The resulting shape of the dataset was 1936758 rows and 29 columns.

Flight delays dataset of 2015 was named, Delay_2015. It had 5819079 rows and 31 columns. Missing values were replaced with 0. No column was dropped.

Handling columns with abbreviations

Columns with abbreviations like AIRLINE, ORIGIN_AIRPORT and DESTINATION_AIRPORT were replaced with full names from the airlines.csv and airports.csv datasets also downloaded from the Kaggle flight datasets page.

Cleaning supplementary dataset

The supplementary dataset was named Airports. It had 80557 rows and 18 columns. Dropped columns were:

1. gps_code column. This column was dropped because after a careful investigation of this column it was evident that it had a lot of missing values. However, all the contents of this column were replicated in the ident and iata code columns. The ident column also had no missing values.
2. home_link and wikipedia_link columns. These were dropped because information about where data was scrapped is not necessary to this project.
3. local_code column. This is because all of its contents were replicated in ident and iata code columns.
4. iso_country column. This is because this project focused only on airports in the United States of America so there was no need have a column that had only this information.
5. Municipality column. This is because it was irrelevant knowing the names of the airports were enough to satisfy the objective of this project.
6. scheduled_service column. This is because the information in this column was irrelevant to this project.

7. Continent column. This column was also irrelevant to this project.

Handling missing values in supplementary dataset

Missing values in elevation_ft column was replaced with median value of the column.

Data on only US Airports were selected and stored under the name Airports.

Merging all datasets

As initially mentioned, the features of Delays _2008 and Delays_2015 were the similar only that they had been stored under different names. To make it easier to merge the two datasets, columns in Delay_2008 were renamed to match columns in Delay_2015.These two datasets were merged with Delay_2008 on top of Delay_2015. The resulting dataset was named 'Delays'.

The IATA code from the airports.csv dataset was added to Delays dataset.

The cleaned supplementary dataset; 'Airports' was subsequently also merged with Delays using a left join on IATA CODE columns. All missing columns were filled with 0 after the merge.

Columns with float data type were converted to int data type.

Cleaned dataset was saved as fl_delay. It had 7755837 rows and 42 columns.

Exploratory Data Analysis

Understanding features

The EDA stage of the flight cancellation project revealed that distribution of most of the features were right skewed. The initial part of the EDA focused on using visualizations to explore the meaning of features in the dataset. Features like Elapsed Time, Scheduled time and airtime seemed to have similar meaning but the data in these columns were not the same. Scheduled time, though closely related to airtime from the dataset showed some differences which were indicative of delay.

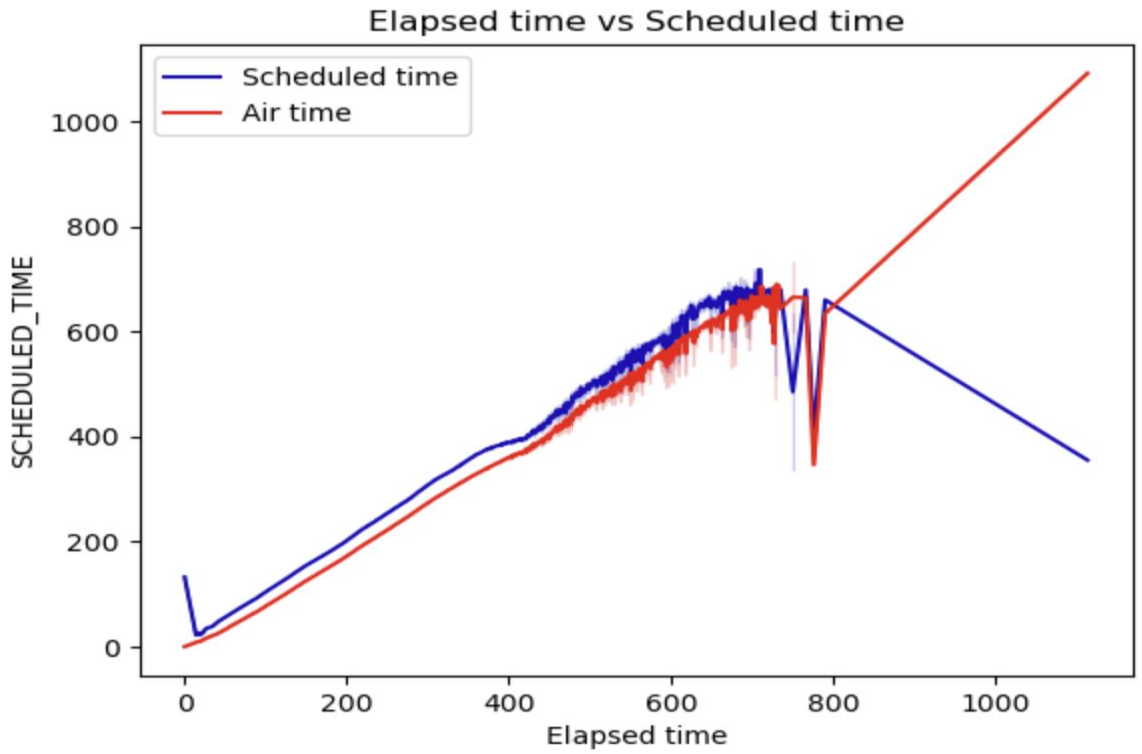


Fig 1 Elapsed time vs Scheduled time

Wheel on and wheel off times were also explored. Visualizing with line plots showed that Wheel on times increased significantly in 2015 compared to 2008. Wheel on time which refers to the time an aircraft touches down on the runway at the airport and Wheel off time which refers to the time the aircrafts wheels leave the runway of an airport. Wheel on and wheel off features were initially thought to be closely related to Arrival time and Departure time features respectively. However, using line plot to analyze these features carefully showed that:

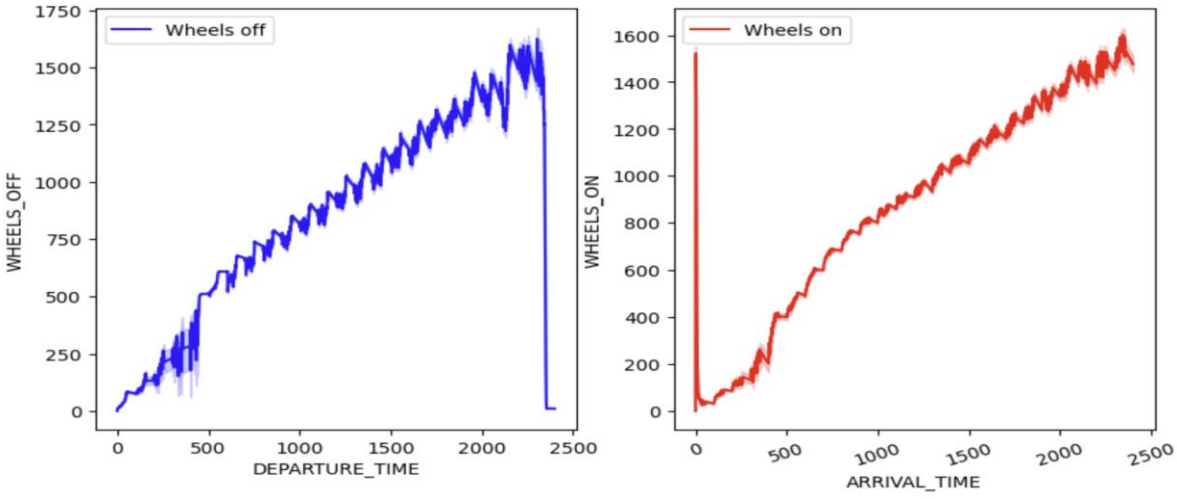


Fig 2 a) Departure time vs Wheels off b) Arrival time vs Wheels on

1. Wheels off time is usually a few minutes before departure time. Which is a realistic finding because the departure time is usually set after the time after the aircraft tyres leave the airport and this time comes later than the time the tyres of the aircraft leaves the airports runway. However, from the line plot, as departure time increases, wheels off time tends to decrease. This may be due to the length of the airport's runways due to the airport's elevation from ground or the size of the airport. Aircraft wheels may leave the runway but still be in the vicinity of the airport.
2. Initial wheels on times are hours ahead of arrival time when the plane is stationary. As arrival time increases, wheels on time lags. This is because the arrival time is calculated by the time the aircraft arrives at the gate. In which case the wheels touch the runway first before the plane moves to the gate. The further the distance from the runway to the gate, the higher the probability that the time difference between the wheels on time and arrival time will differ.

The two graphs indicate that wheels off and wheel on times is not the same as departure and arrival time.

Using EDA, the difference between scheduled arrival and arrival time as well as scheduled departure and departure time was also investigated. It was found that scheduled times do not correspond to the real times in the dataset.

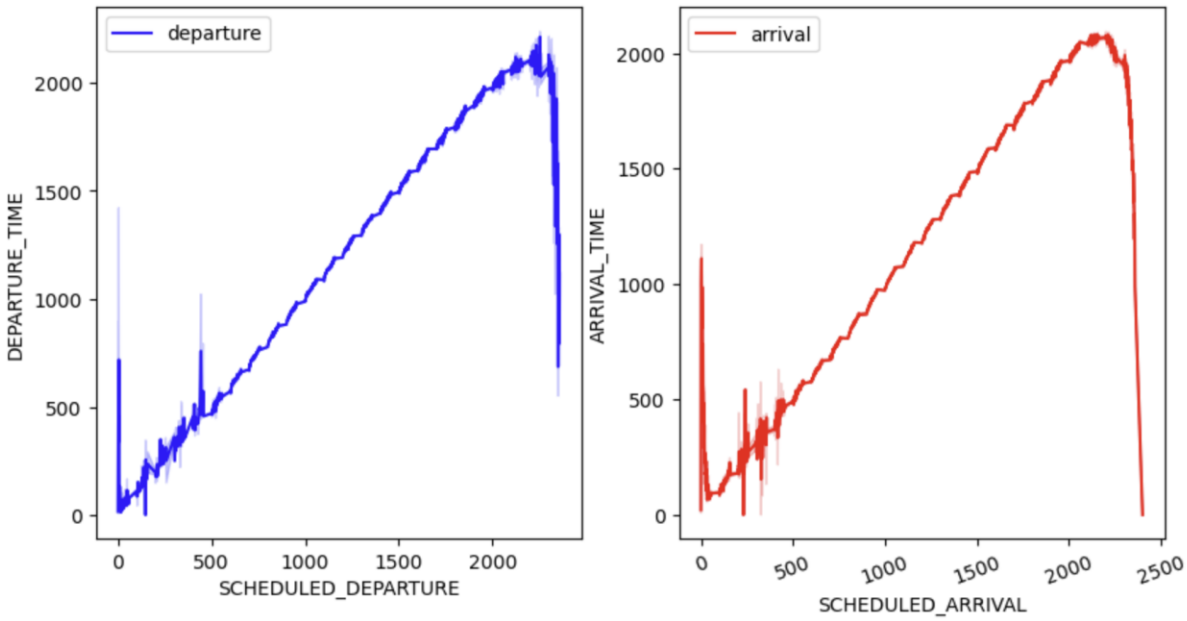


Fig 3 a) Scheduled departure vs Departure time b) Scheduled arrival vs Arrival time

Arrival delay was also plotted against departure delay, and it was found that arrival delay is positively correlated to departure delay. However, arrival delays are not always a consequence of departure delays and vice versa.

Exploring features and observing trends.

Using multiple scatter plots, it was evident that:

a. More departure delays and arrival delays were recorded in 2008 than in 2015.

b. Nas_delays were observed mostly in 2008. Delays that are within the control of the National Airspace System (NAS) may include: non-extreme weather conditions, airport operations, heavy traffic volume and air traffic control.

c. February and April were the months that recorded longest Departure delays and arrival delays.This may be because more passengers were returning from new year celebrations and in April passengers were going on summer holidays.

d. Long weather delays were experienced in December because its winter season

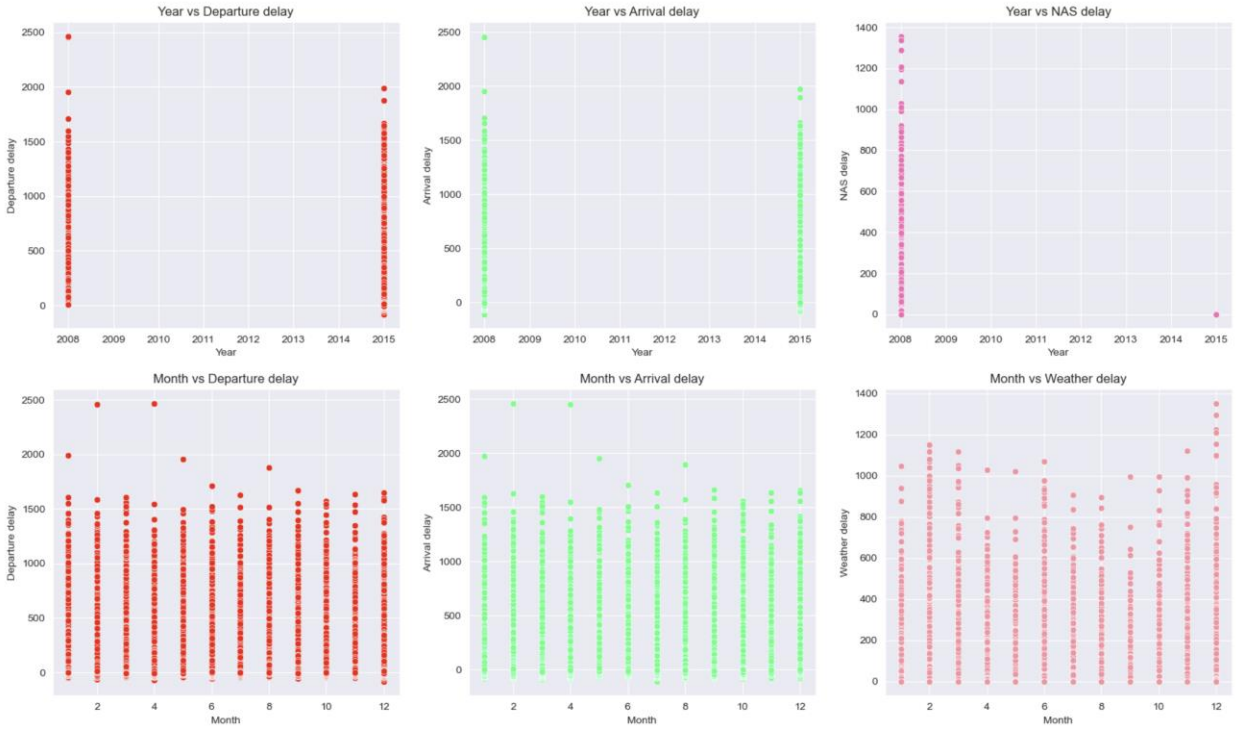
e. Long airline delays were observed on January and August.

f. Long security delays were observed on September, 9th

g. Late aircraft delays were observed on almost all months except January.

h. Security delays were commonly oberved in 2015 than in 2008

i. More departure delays and arrival delays were recorded in 2008 than in 2015



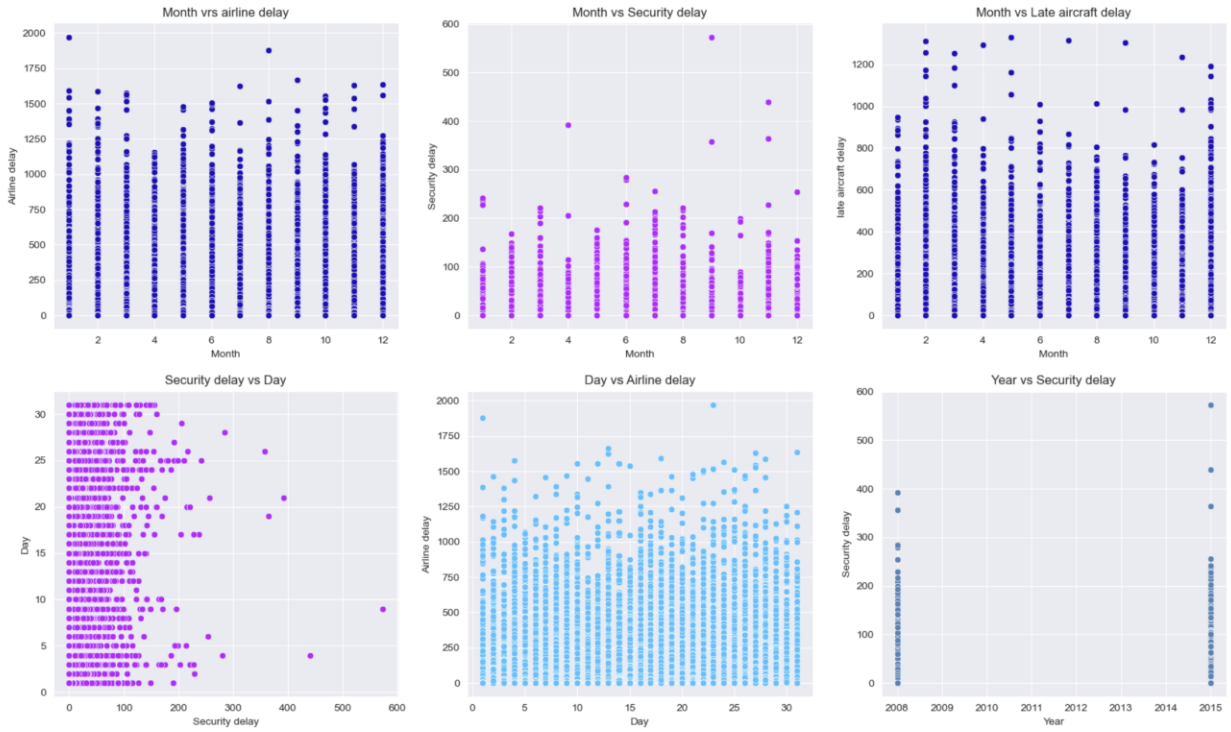


Fig 4 Multiple scatter plots revealing trends in features

Flight Cancellations

The monthly cancelled flights were also investigated using a bar plot. The results showed that more flights were cancelled in November, January and February. This may be due to harsh winter weather conditions.

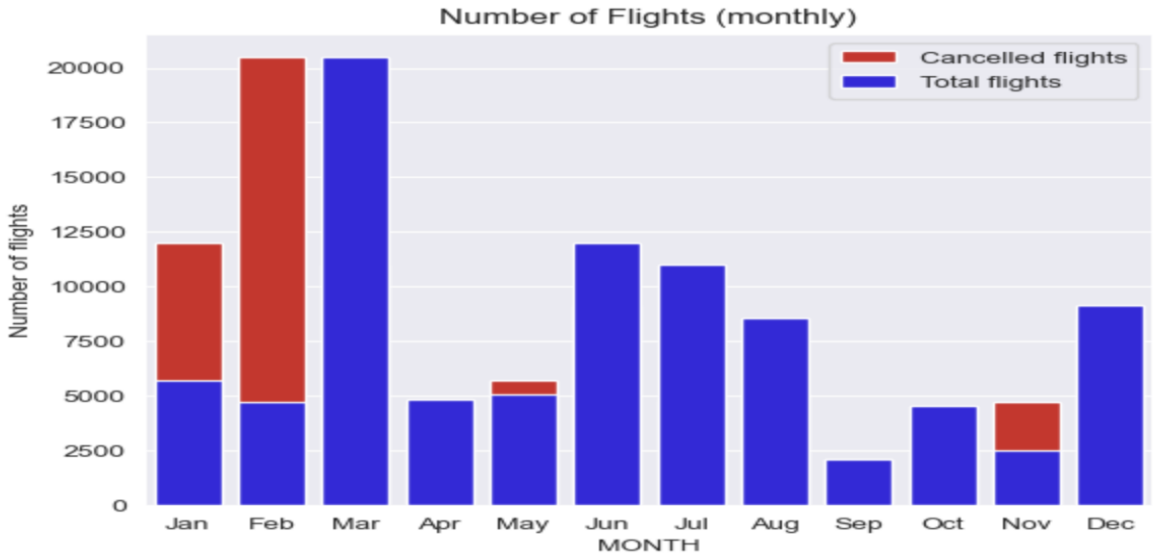


Fig 5 Stacked bar chart of cancelled flights and total flights

Airlines responsible for the most flight cancellations had a cancellation rate of 3.5% and was American Eagle Airlines Inc.

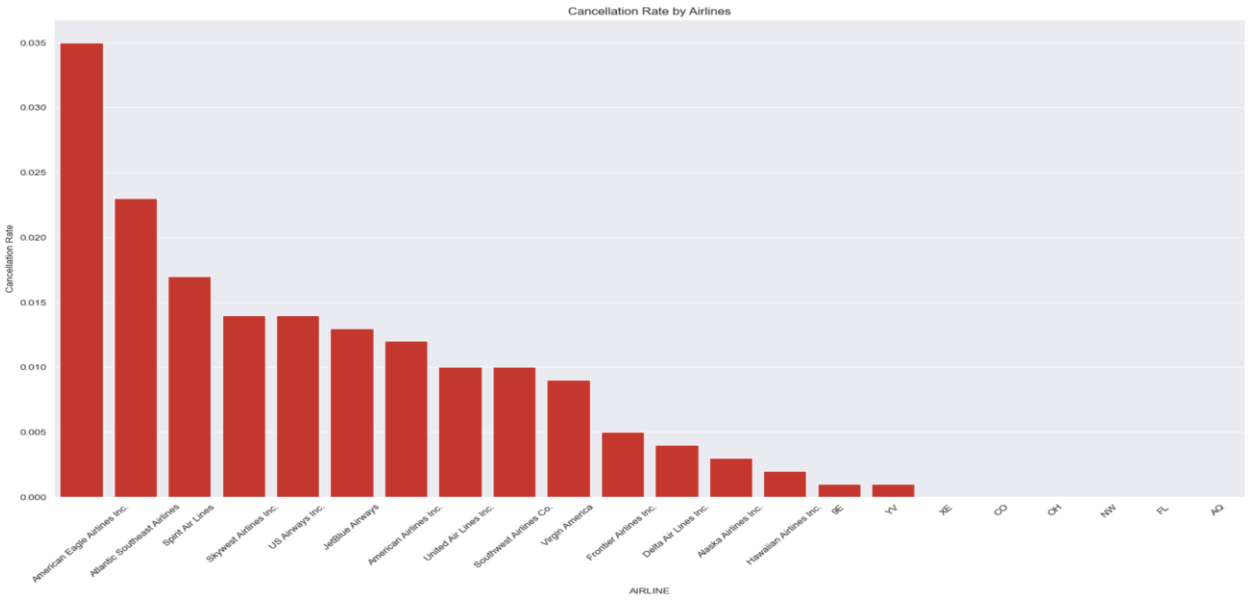


Fig 6 Airline Cancellation rate

Exploring the different types of delays

There were eight types of delays in the dataset and each of these delays were analyzed. It was showed using multiple line plots that Security delays which is caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines more than 29 minutes at screening areas was more likely to cause flight cancellation than any other delay.

Weather delays were also more likely to cause flight cancellation.

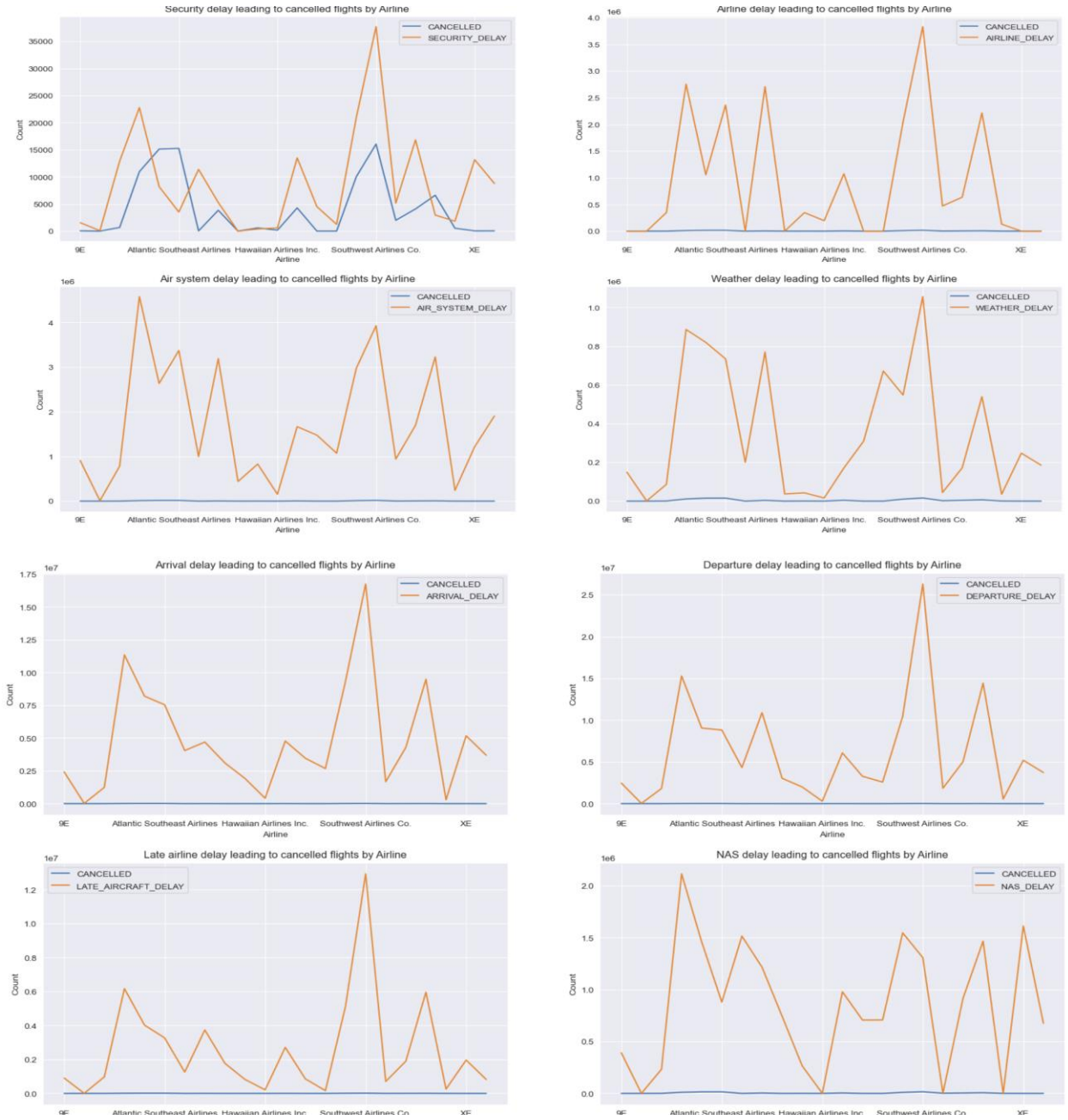


Fig 7 Multiple line plots showing the delays that led to flight cancellations

Exploring Airport Characteristics that lead to flight delays and cancellation

With respect to the most visited airports, Hartsfield Jackson Atlanta International Airport was the most visited origin and destination airport.

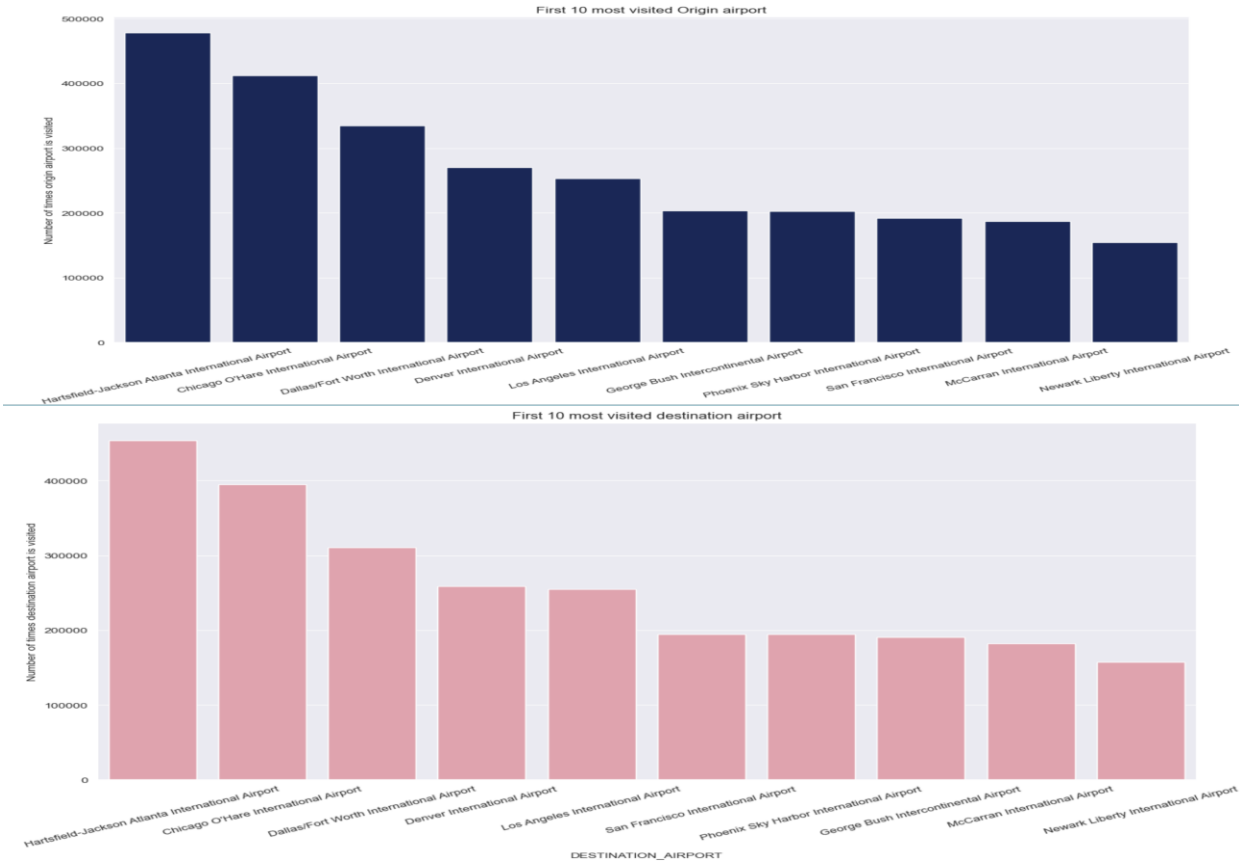


Fig 8 a) First 10 most visited origin airports b) first 10 most visited destination airports

The location and structure of airports was essential in preventing delays and subsequent flight cancellations. Airports with high elevation affect airline performance in many ways. The increased elevation means these airports have increased runway distances which also means increased landing distances due to reduced air density. This property of airports makes it difficult for landing and takeoff and may lead to departure delays and arrival delays. Pilots must calculate takeoff and landing distances as well as adjust their approach and departure procedures to ensure safe operations at high elevation airports.

The most elevated airport was Gunnison Crested Butte Regional Airport and Aspen-Pitkin County Airport (Sardy Field).

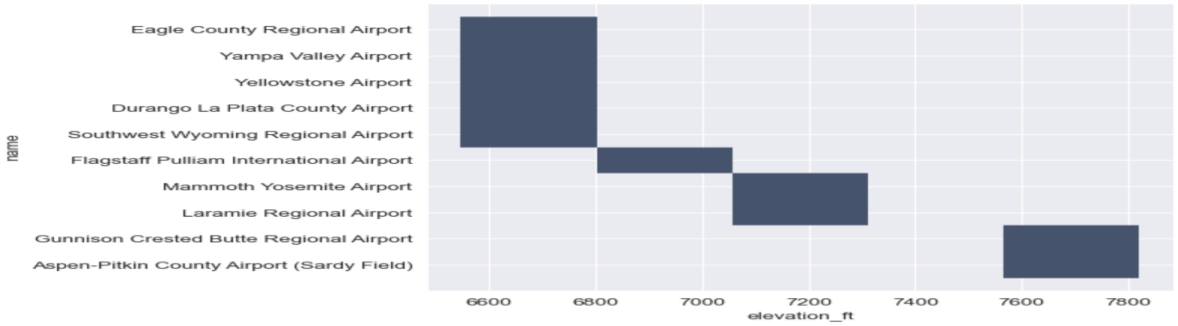
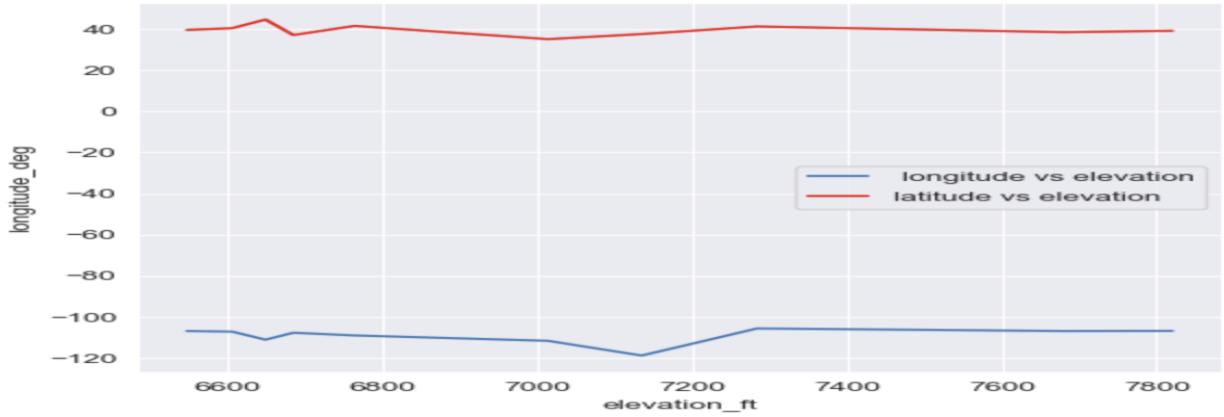


Fig 9 first 10 most airports with high elevation

Longitude and latitude give information about the position or location of an airport in US. Latitude measures how far north or south a location is from the Equator while longitude measures how far east or west a location is from the prime meridian at Greenwich. At higher elevations, latitude and longitude appear more stable but that is not the case. Elevation does not influence longitude or latitude.



Fi9 10 Elevation vs longitude and latitude degrees

Airports situated at high elevations were, however, mostly associated with more delays as shown in the chart below.

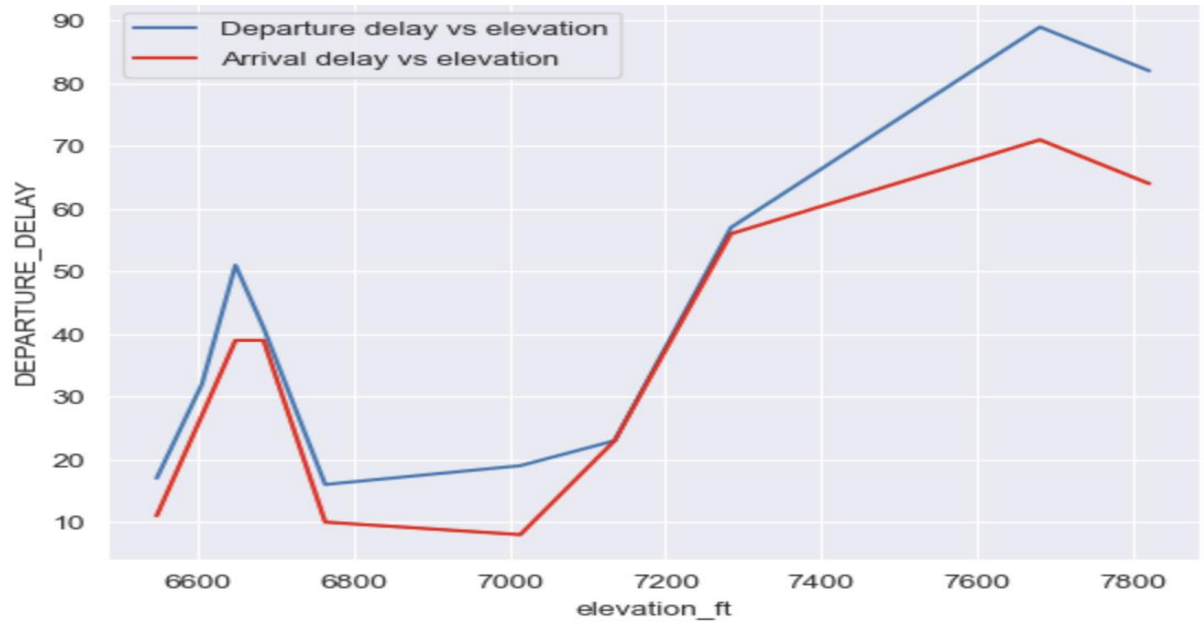


Fig 11 Elevation vs departure delays

Airline performance and contribution to flight delays and cancellations

Analyzing delays caused by airlines like the air system delays, airline delays, late aircraft delays, NAS_ delay, departure delay, diverted and arrival delays revealed that:

i. Virgin America Airlines had more airline delays whereas Alaska airline has less airline delays

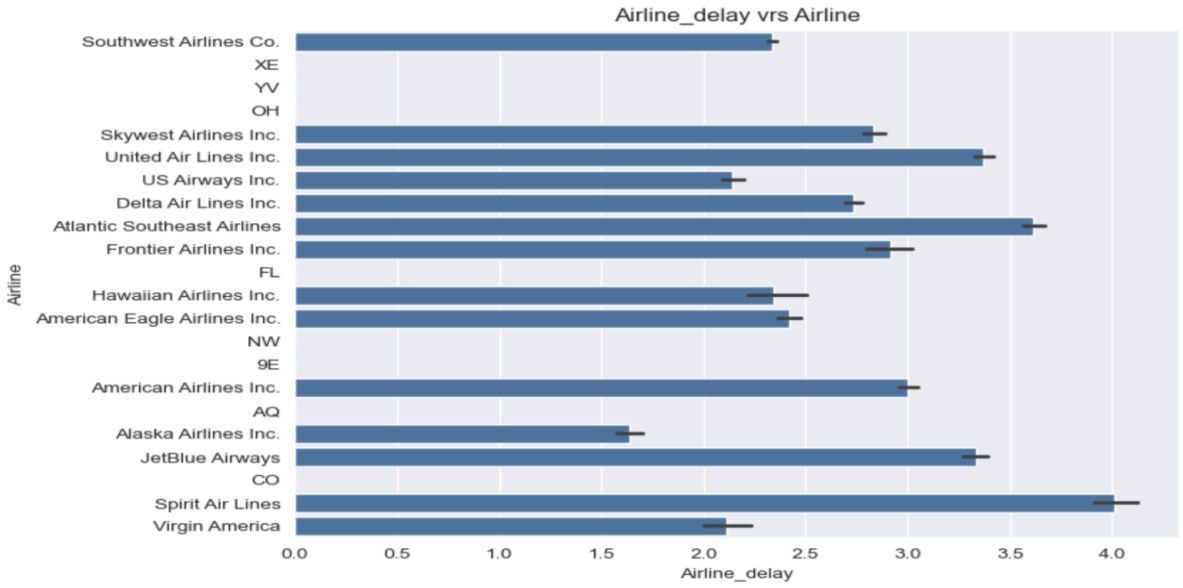


Fig 12 Airline delay vs Airlines

ii. Airline departure delays last longer than airline arrival delays. Airline arrival delays were however more common than airline departure delays.

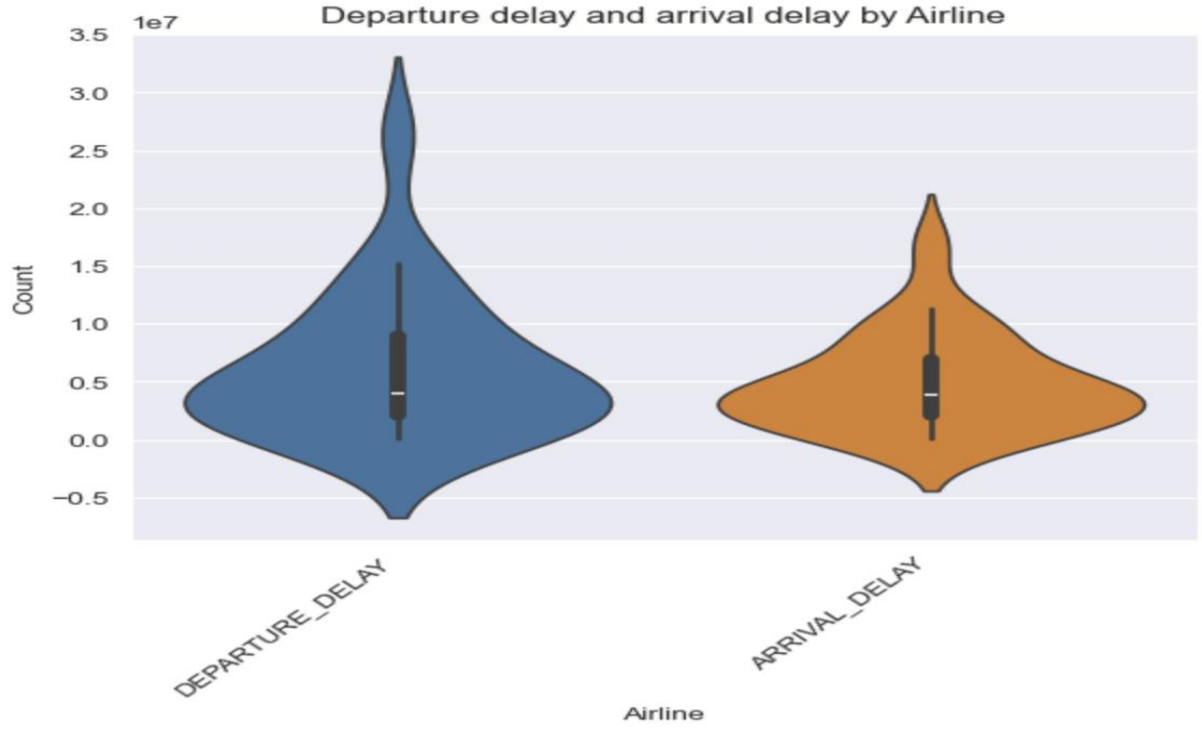


Fig 13 Violin plots of departure delay and arrival delay by airlines

iii. Late Aircraft delay is an arrival delay at an airport due to the late arrival of the same aircraft at a previous airport. The ripple effect of an earlier delay at downstream airports is referred to as delay propagation.

With reference to the chart below, FL airlines have the latest aircraft delay whereas Hawaiian Airlines Inc has the least late aircraft delay.

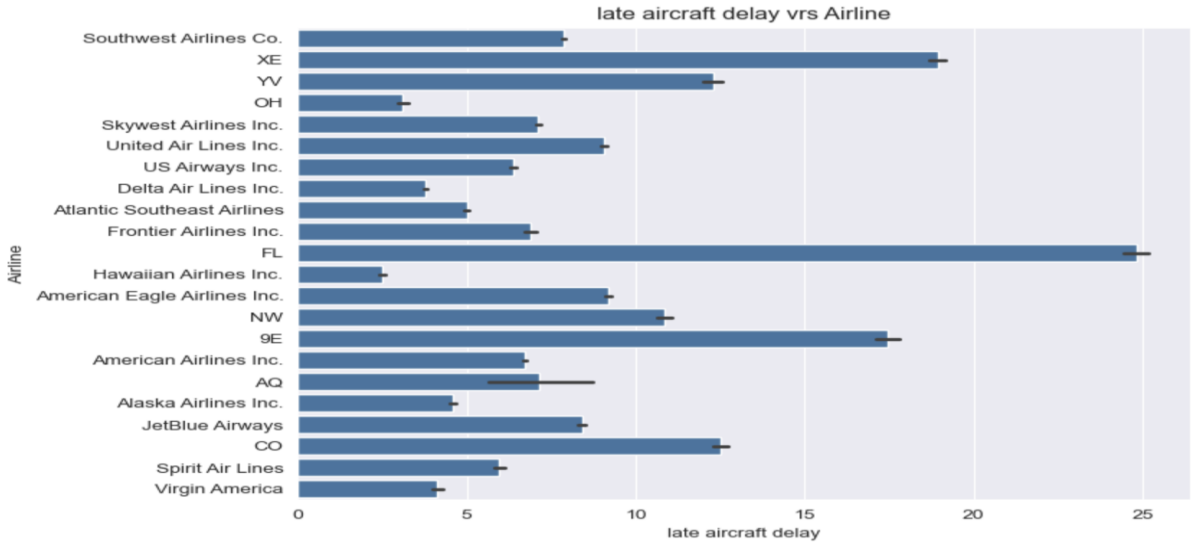


Fig 14 Late aircraft delay vs airline

iii. The airline with the most Air system delays was YV whereas Hawaiian Airlines Inc had the least air system delay.

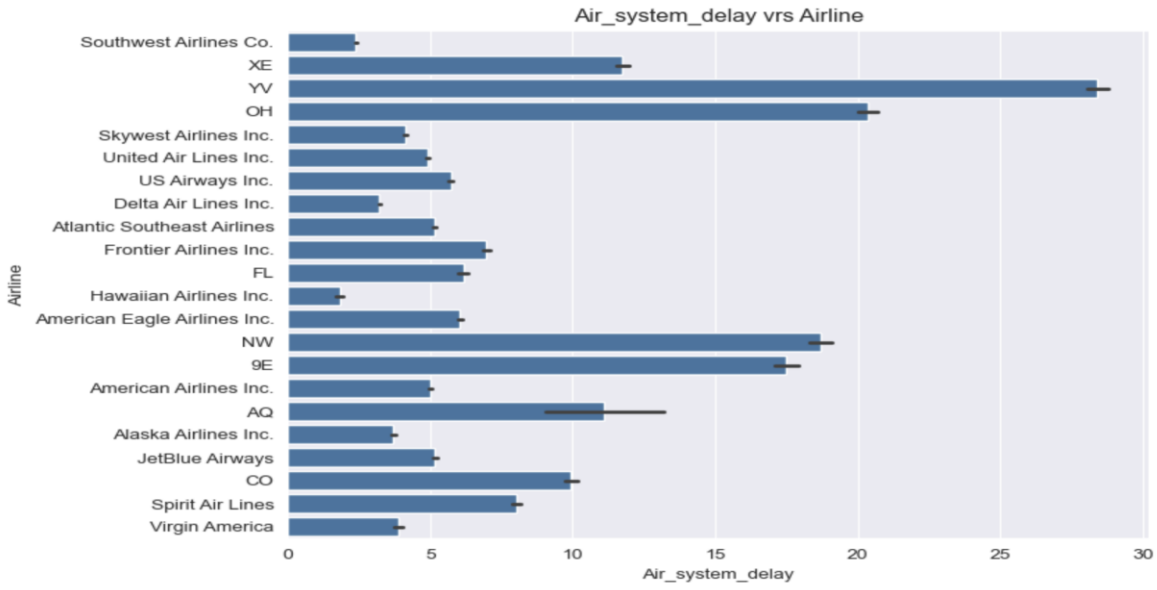


Fig 15 Air system delay vs airline

iv. AQ airlines has the most diverted flights whereas Hawaiian Airlines Inc were the least diverted flights.

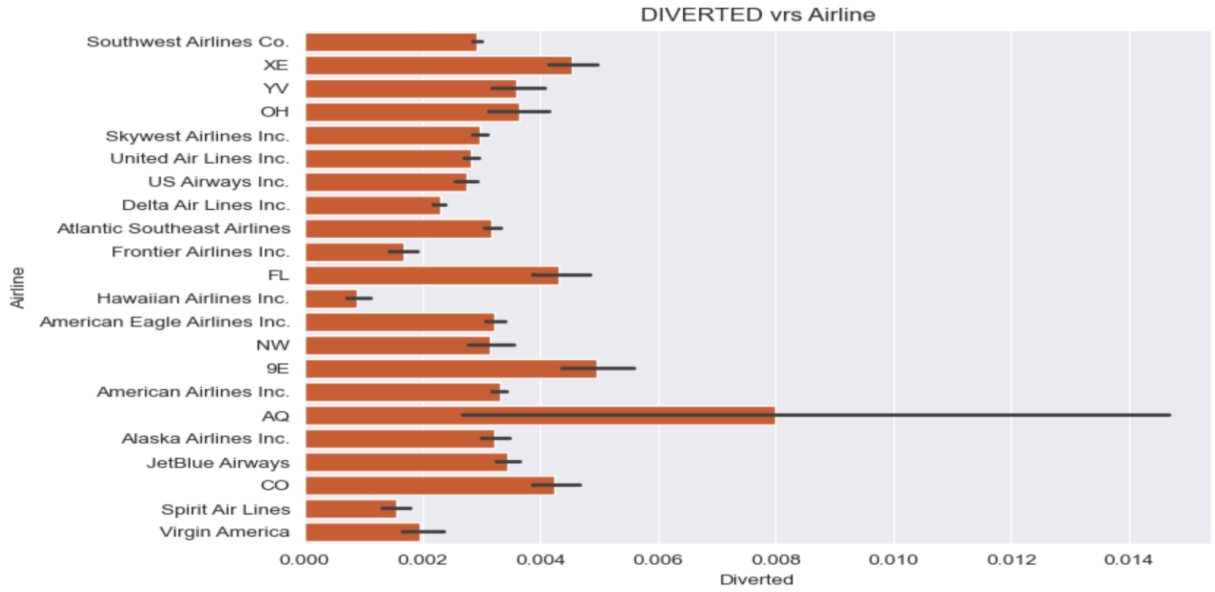


Fig 16 Diverted vs airline

v. Diverted flights mostly result in departure delays.

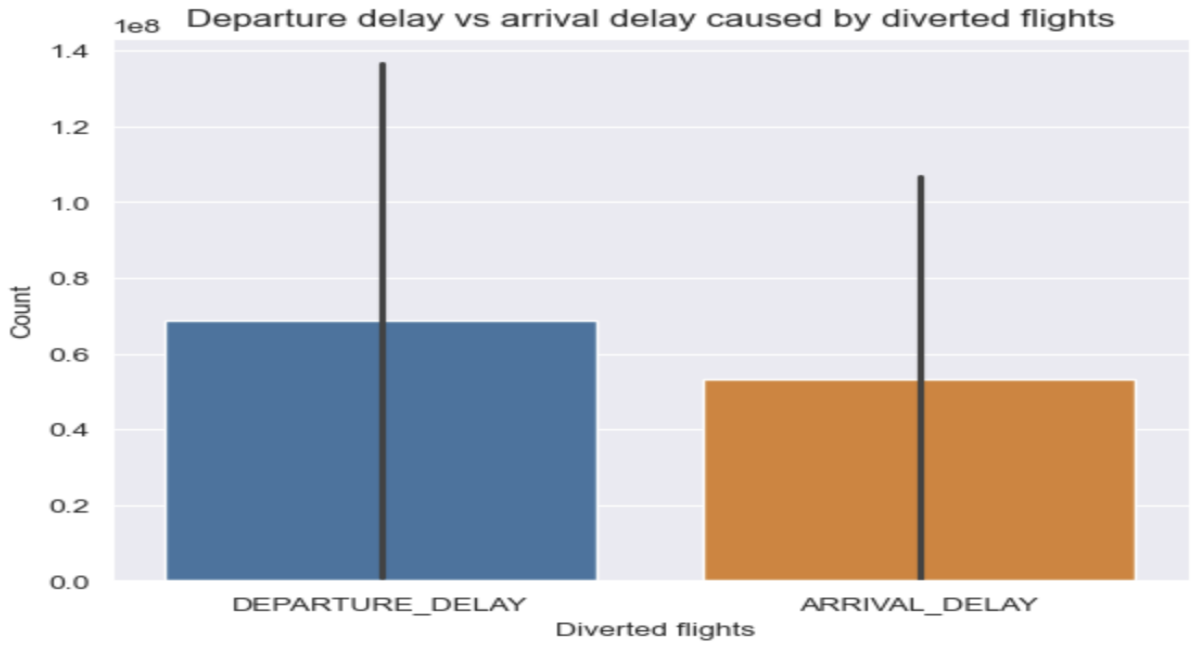


Fig 17 Distribution of departure delays and arrival delays caused by diverted flights

vi. Southwest Airlines Co. has the most patronized flights.

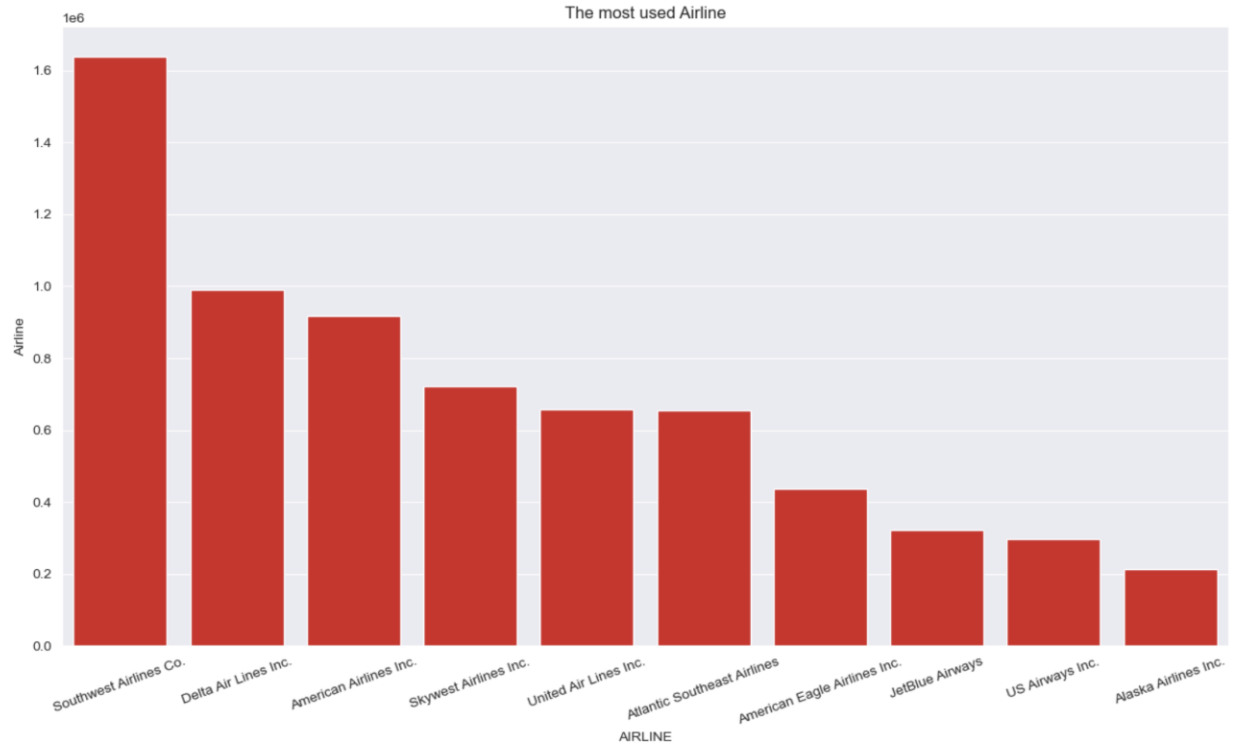


Fig 17 Most patronized flights

vii. Airline delays are mostly because of airline departure delays, arrival delays and late aircraft delays. Airline security delays contribute the least to airline delays.

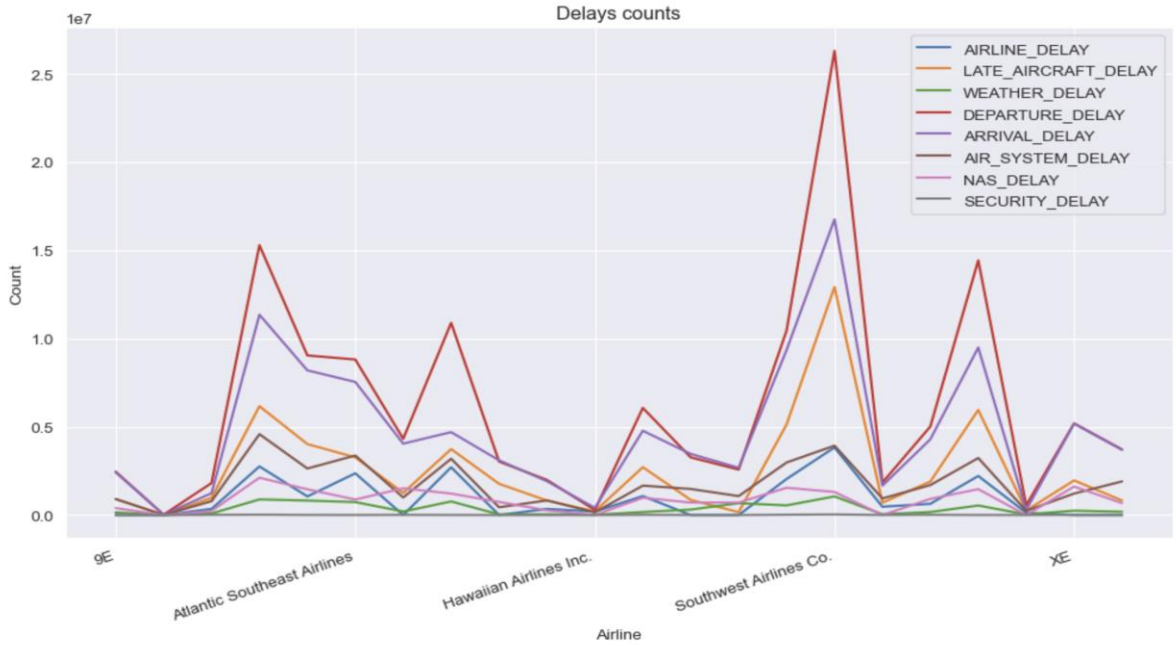


Fig 18 Caused of airline delays

Weather delay is caused by extreme or hazardous weather conditions that are forecasted at point of departure, enroute, or on arrival. From the chart, long distances flights tend to have longer hours of weather delays.

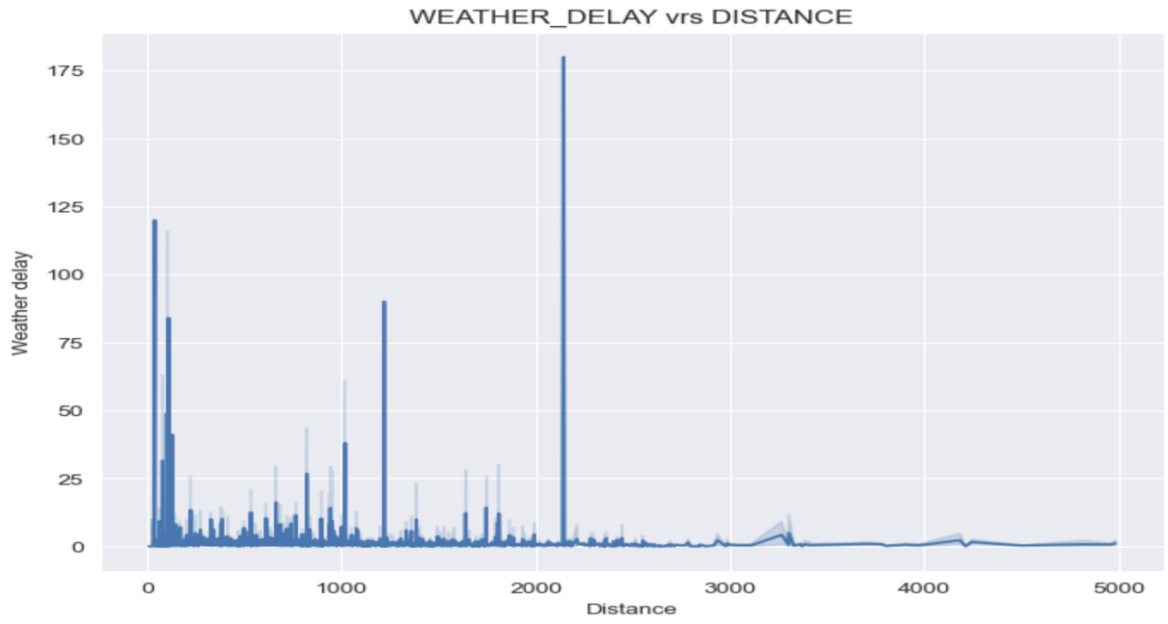


Fig 19 Distance vs Weather delays

The distance feature in the dataset may have been used to calculate the flight time. This is a possible explanation of the strong correlation between Elapsed time and distance.

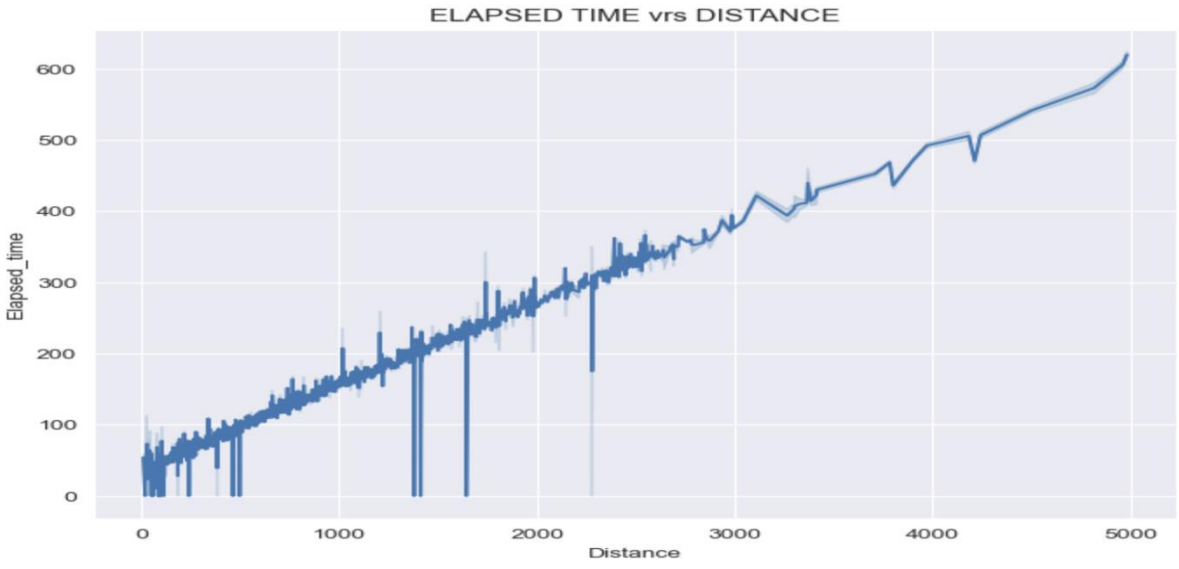


Fig 20 Distance vs Elapsed time

Heatmap

The final stage of EDA constituted construction of a heatmap, to further reveal relationships between features.

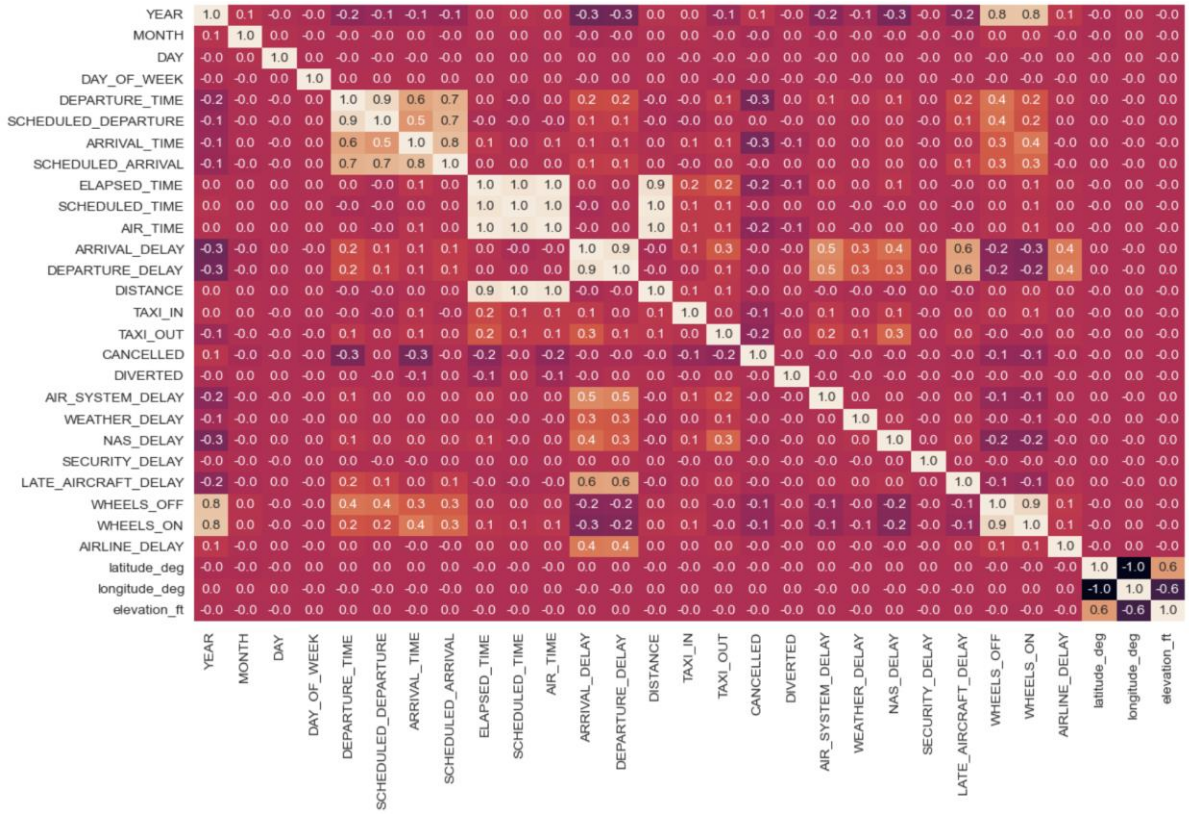


Fig 21. Heatmap showing correlations between features

Strong correlations were noted when correlations are equal or greater than 0.5. These relationships were found between:

From the heatmap, strong correlations were noted when correlations are equal or greater than 0.5. These relationships were found between:

1. Departure time, scheduled departure, arrival time and scheduled arrival.

2. Elapsed time, airtime and scheduled time.

3. Arrival delays and departure delays.

4. Air system delay, arrival delay and departure delay.

5. Late aircraft delay, arrival and departure delay.

6. Distance and Elapsed time.

7. Wheels off and Wheels on.

8. Year, wheels off and wheels on.

9. Latitude, longitude and elevation ft.

PREPROCESSING

All delay features were dropped from the dataset prior to modelling. Columns like the id, ident, Cancellation reason, type, name, tail number, latitude_deg, longitude_deg, elevation_ft, iso_country, iata_code were all dropped. Categorical variables were also dropped except Airline and IATA code. These two categorical columns were encoded, and the first column of the dummy variable was deleted.

X as all variables except the cancelled feature. y variables were also defined as only the Cancelled feature. Data was split into training and test sets in the ratio 80:20 respectively.

Training and test sets were saved using the pickle module.

MODELLING

Three machine learning models were built to predict flight cancellation.

The first model created was a Linear Regression Model.

The second model created was Decision Tree Regressor with the max depth parameter set to 2.

The third model created was a Random Forest Classifier Model. The n_estimators parameter was set at 8.

These models were assessed, and the best performing model was selected.

RESULTS AND DISCUSSION

The first model created to predict flight cancellation was Linear Regression model.

The intercept of this model is -22.988276.

The second model was a decision tree regressor with maximum depth parameter set to 2.

These models were linear models and were assessed by calculating R-squared score, the mean absolute error and the mean squared error. The result of the assessment is found in table 1.

Table 1 Linear Regression and Decision Tree Regressor Models. Assessment Results.

Model Assessment	Linear Regression	Decision Tree Regressor
R-squared	0.64	0.99
Mean absolute error	0.032	0.002
Mean squared error	0.064	0.0104

The R-squared score of the Decision Tree Regressor model is 0.99 which indicates that accuracy of the prediction of this model is better than the Linear model which was 0.64. The mean absolute error was 0.0002 and the mean squared error was also 0.0104 which are smaller than the errors calculated of the Linear Regression model.

The third model was a Random Forest Classifier model. The n_estimators parameter was set at 8. The accuracy of this model was 1 and the f1-score was 1. These scores indicate the best performance of the Random Forest Classifier model.

```
y_pred = RF_MODEL.predict(X_test_loaded)
y_pred_prob = RF_MODEL.predict_proba(X_test_loaded)


# Assessing RF model

ac = accuracy_score(y_test_loaded, y_pred)

f1 = f1_score(y_test_loaded, y_pred, average='weighted')
cm = confusion_matrix(y_test_loaded, y_pred)

print('Random Forest: Accuracy=%.3f' % (ac))

print('Random Forest: f1-score=%.3f' % (f1))
```

```
Random Forest: Accuracy=1.000
Random Forest: f1-score=1.000
```

Fig 22 Accuracy and F1 score of Random Forest Classifier Model

Confusion matrix and classification report was generated.

```
# Printing confusion matrix

print(cm)
```

```
[[1533029          0]
 [     19      18120]]
```

Fig 23 Confusion matrix of Random Forest Classifier Model

The confusion matrix indicated that the model was able to identify these;

18125 True negatives

1533029 True positives

0 False positive

19 False negatives

The classification reports showed that there is class imbalance. One class was 1533029 and the other class was 18139. For this reason, the accuracy of the model was not as relevant as the precision and recall of the model. The final performance of the model showed both precision and recall scores to be 1.0. This was the best score any model could have. It can therefore be concluded that the model was performing well.

Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1533029
1	1.00	1.00	1.00	18139
accuracy			1.00	1551168
macro avg	1.00	1.00	1.00	1551168
weighted avg	1.00	1.00	1.00	1551168

Fig 24 Classification report of Random Forest Classifier Model

A Receiver Operating Characteristics curve was also generated.

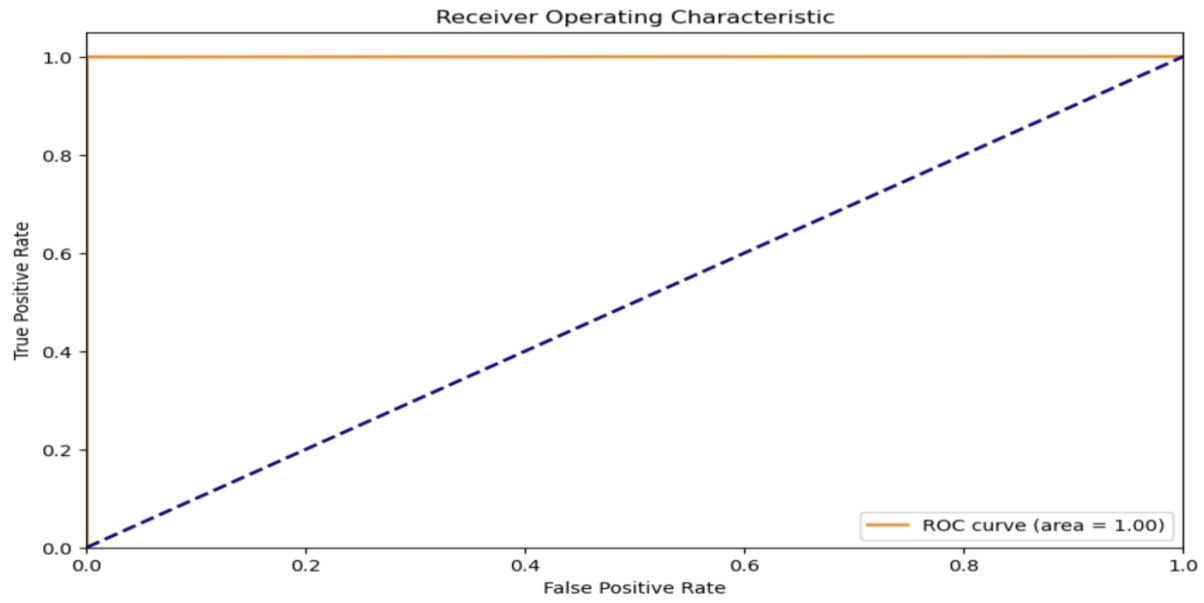


Fig 25 Receiver Operating Characteristics curve

The graph proves that the diagnostic ability of the model is good. The area under the curve value is 1.0 and it shows that the model can rank a randomly positive instance higher than a randomly negative instance.

The feature importance of this Random Forest model was also generated and visualized.

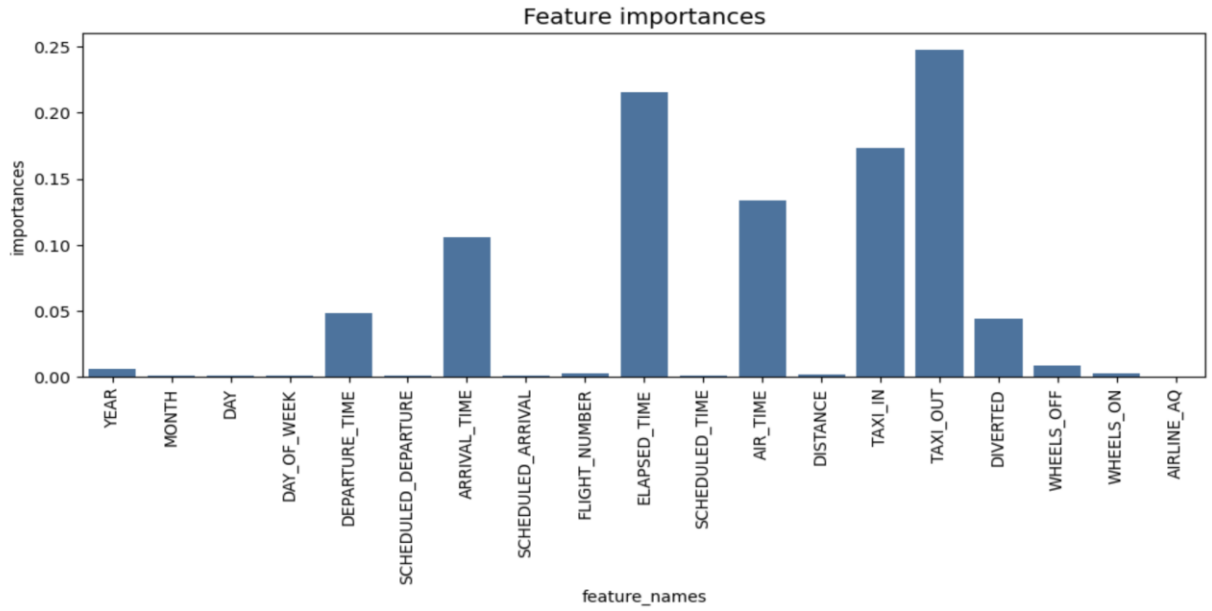


Fig 26 Feature importance chart

The most important features of this model are the taxi out, elapsed time and taxi in features.

Comparing these three models, the Random Forest model performed the best.

MODEL LIMITATIONS

Although the Random Forest Model can differentiate true positives from true negatives quite well, the model still had trouble categorizing 19 of the training data. Though negligible, this resulted in 19 false negatives as shown in the confusion matrix generated earlier.

False negative in this case means that the model predicts that flights are not cancelled when in fact they are.

Comparatively, having false negatives results in worse scenarios than false positives do although there are no false positives.

False negative means that travelers will lose their precious time and money to book flights, only to find out at the last minute that these flights have been in fact been cancelled whereas false positive means that 1 traveler will have to reschedule their trips because the prediction that flights will be cancelled was false.

MODEL IMPROVEMENT

False negatives are limitations of the Random Forest model because this predictive model was built to solve the issue of false negatives. This is an indication that the current state of the model does not solve objective of the project.

To improve the model, the threshold line must be adjusted so that the model predicts more false positives and less false negatives since the latter will cost us more than the former.

The threshold line that resulted in the earlier confusion matrix was 0.6.

In this specific case the threshold line must be reduced.

```
# new confusion matrix after reducing threshold to 0.3

precision_recall_threshold(p, r, thresholds, 0.3)
```

	pred_neg	pred_pos
neg	1533018	11
pos	0	18139

Fig 27. Confusion matrix after reducing threshold to 0.3

After several iterations, the threshold line was reduced from 0.6 to 0.3 and that reduced false negative to zero and increased false positive to 11.

CONCLUSION

This project used three datasets: 2015 Flight Delays and Cancellations (kaggle.com) downloaded from Kaggle. It has about three files airlines.csv, airports.csv and flights.csv that have information about flight delays in 2015, the airlinedelaycauses_DelayedFlights.csv also downloaded from Kaggle and a supplementary dataset downloaded from Open data @ OurAirports.

These datasets were cleaned and merged into one dataset with 7755837 rows and 42 columns. These columns were YEAR, MONTH, DAY , DAY_OF_WEEK ,DEPARTURE_TIME, SCHEDULED_DEPARTURE, ARRIVAL_TIME ,SCHEDULED_ARRIVAL ,AIRLINE , FLIGHT_NUMBER ,TAIL_NUMBER, ELAPSED_TIME, SCHEDULED_TIME, AIR_TIME ARRIVAL_DELAY ,DEPARTURE_DELAY, ORIGIN_AIRPORT ,DESTINATION_AIRPORT , DISTANCE , TAXI_IN, TAXI_OUT , CANCELLED ,DIVERTED , AIR_SYSTEM_DELAY , WEATHER_DELAY ,NAS_DELAY, SECURITY_DELAY ,LATE_AIRCRAFT_DELAY, WHEELS_OFF , WHEELS_ON ,CANCELLATION_REASON , AIRLINE_DELAY, IATA_CODE , id , ident , type ,name ,latitude_deg , longitude_deg ,elevation_ft ,iso_country, iso_region and iata_code.

Exploratory data analysis reveal that more flights were cancelled in November, January and February. This may be due to harsh winter weather conditions. American Eagle Airlines Inc was responsible for the most flight cancellations at a cancellation rate of 3.5%.

Security and Weather delays were most likely to cause flight cancellations.

Southwest Airlines Co. has the most patronized flights.

All delay features were dropped from the dataset prior to modelling. Columns like the id, ident, Cancellation reason, type, name, tail number, latitude_deg, longitude_deg, elevation_ft, iso_country, iata_code were all dropped. Categorical variables like the origin and destination airports were also dropped. The remaining categorical variable used for modelling were Airline and IATA code. These two categorical columns were encoded, and the first column of the dummy variable was deleted.

The dependent variable was the 'Cancelled' feature.

Three predictive machine learning models were built using the Linear Regression, Decision Tree Classifier, Random Forest Classifier algorithms.

The two linear models were evaluated using R-squared scores, mean absolute error and mean squared error. Of these two the Decision Tree Classifier performed best with an R-squared score of 0.99 compared to R-squared score of 0.64 of the Linear Regression models.

The third model, Random Forest Classifier was nonlinear. The accuracy and F1 score of this model were 1.0. The confusion matrix and classification reports indicate that the model performs better than the initial linear models. It had a precision score of 1.0.

The ROC curve further revealed that the model was performing very well because the area under the curve was 1.0.

The feature importance of the model was calculated using the coefficients and it was shown that the most important features of the Random Forest model were taxi out, elapsed time and arrival time.

A confusion matrix generated shows that the model was unable to accurately categorize training data leading to 19 false negatives. The model threshold line was reduced to 0.3 to reduce false negatives to 0 to the detriment of increasing false positive to 11.

This was done because it will cost more to falsely predict that flights are not cancelled when they are truly cancelled than to predict that flights are cancelled when they are truly not cancelled.

This did not only improve the performance of the model but improved the suitability of the Random Forest models' predictions in addressing the objective of the project.