Deciphering Airline Performance Data

Data Mining B-565

Shraddha Ramprakash Gupta
Radhika Ganesh
Subhadra Mishra

Source: https://www.dreamstime.com/illustration/airport-angry.html

Overview

- Comprehensive dataset crucial for understanding root causes.
- Explore methods for analyzing airline delays and cancellations dataset.
- Uncover insights to refine airline operations and enhance passenger satisfaction.
- Identify primary reasons behind delays/cancellations for operational refinement.
- Prediction of delays on future datasets.

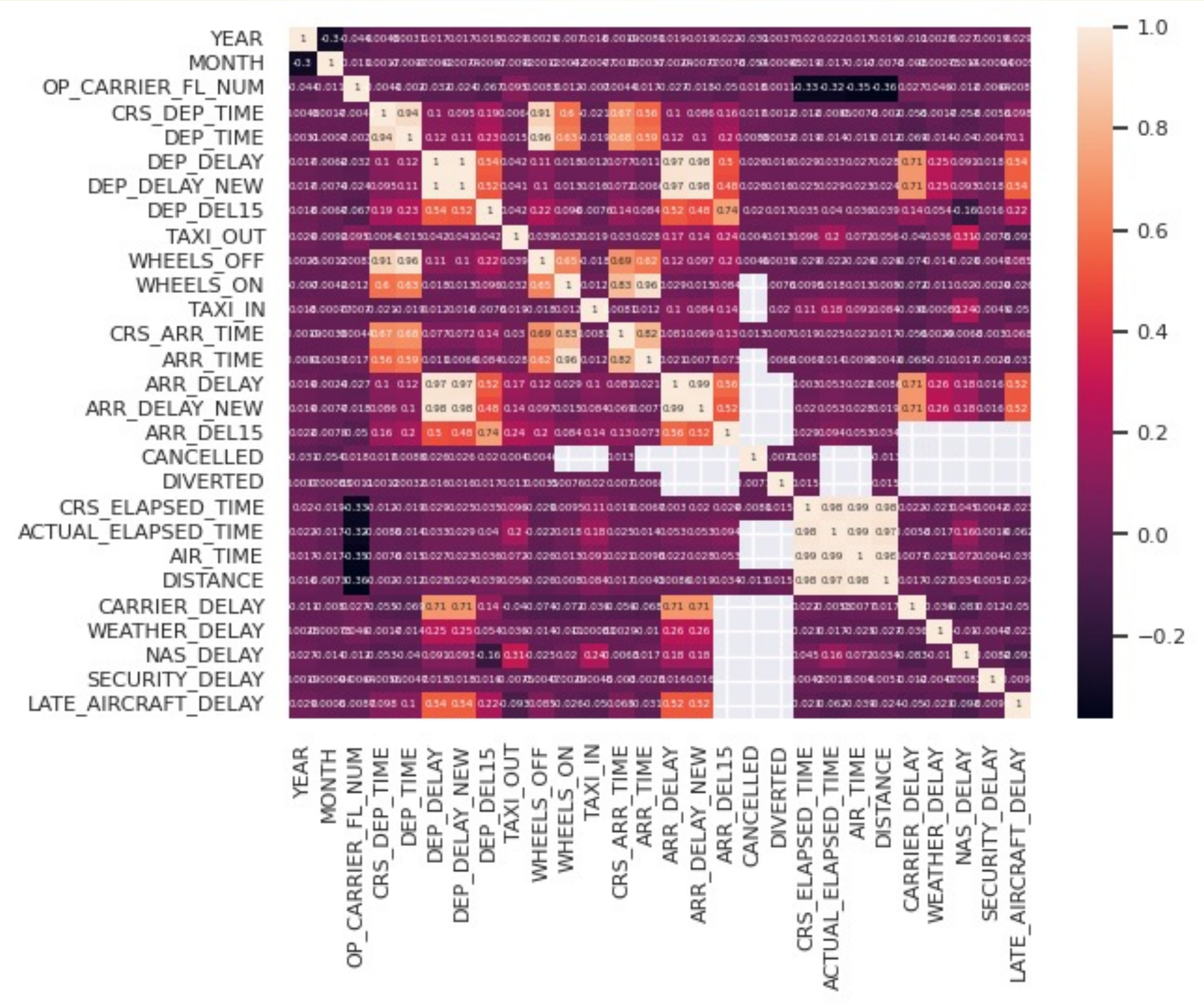
Dataset

- Dataset has been downloaded from the website of Bureau of Transportation Statistics (BTS) which is a part of the US Department of Transportation (DOT).
- Reporting Carrier On-Time Performance data is available from 1987 onwards.
- We have selected the data from January 2022 to August 2023.

Data Preprocessing and Analysis

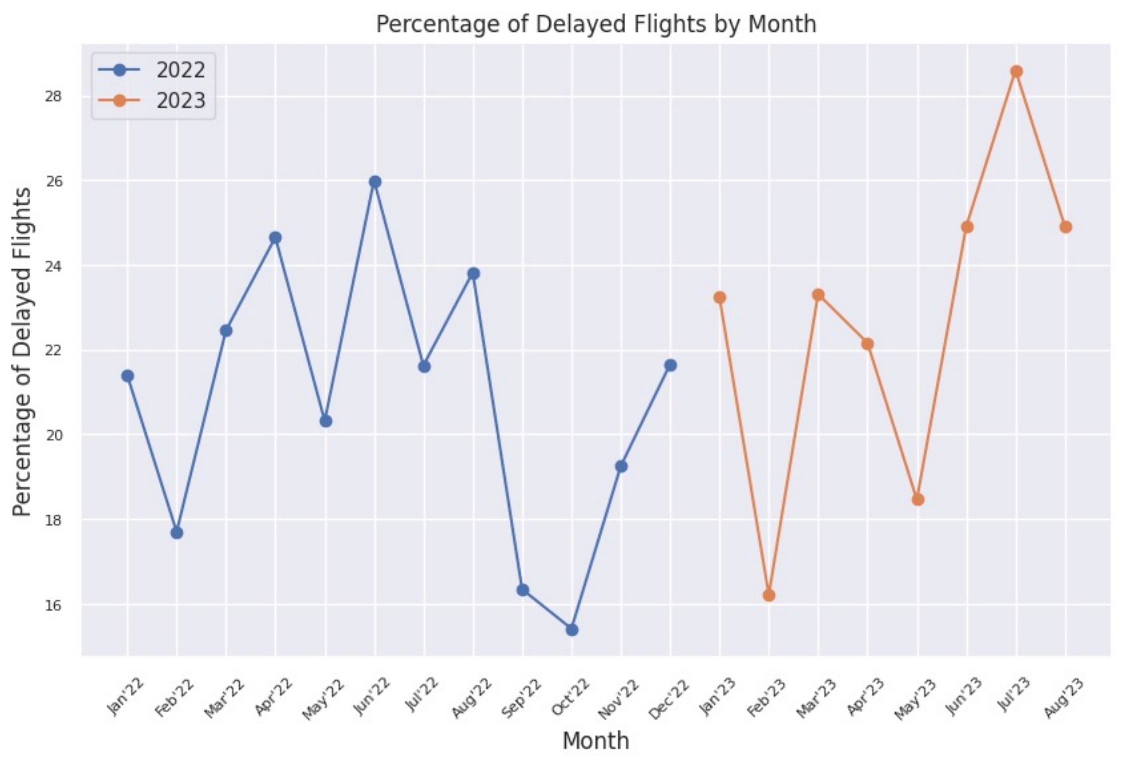
- Data for individual months from January 2022 to August 2023 were downloaded from BTS website and the csv files were merged to prepare the final dataset (70 lakh rows approx. and 34 columns).
- ARR_DEL15 column: Delay = 1, No Delay=0
- Cancelled column: Cancelled: 1, Not Cancelled=0

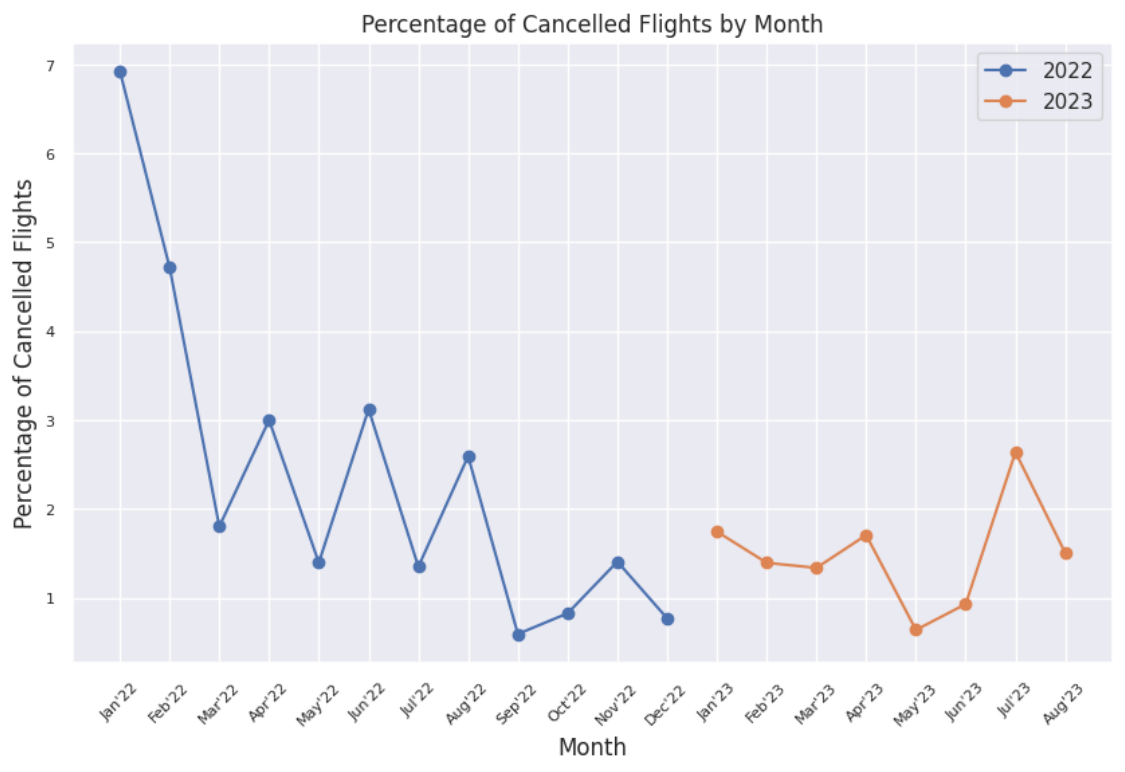
Correlation



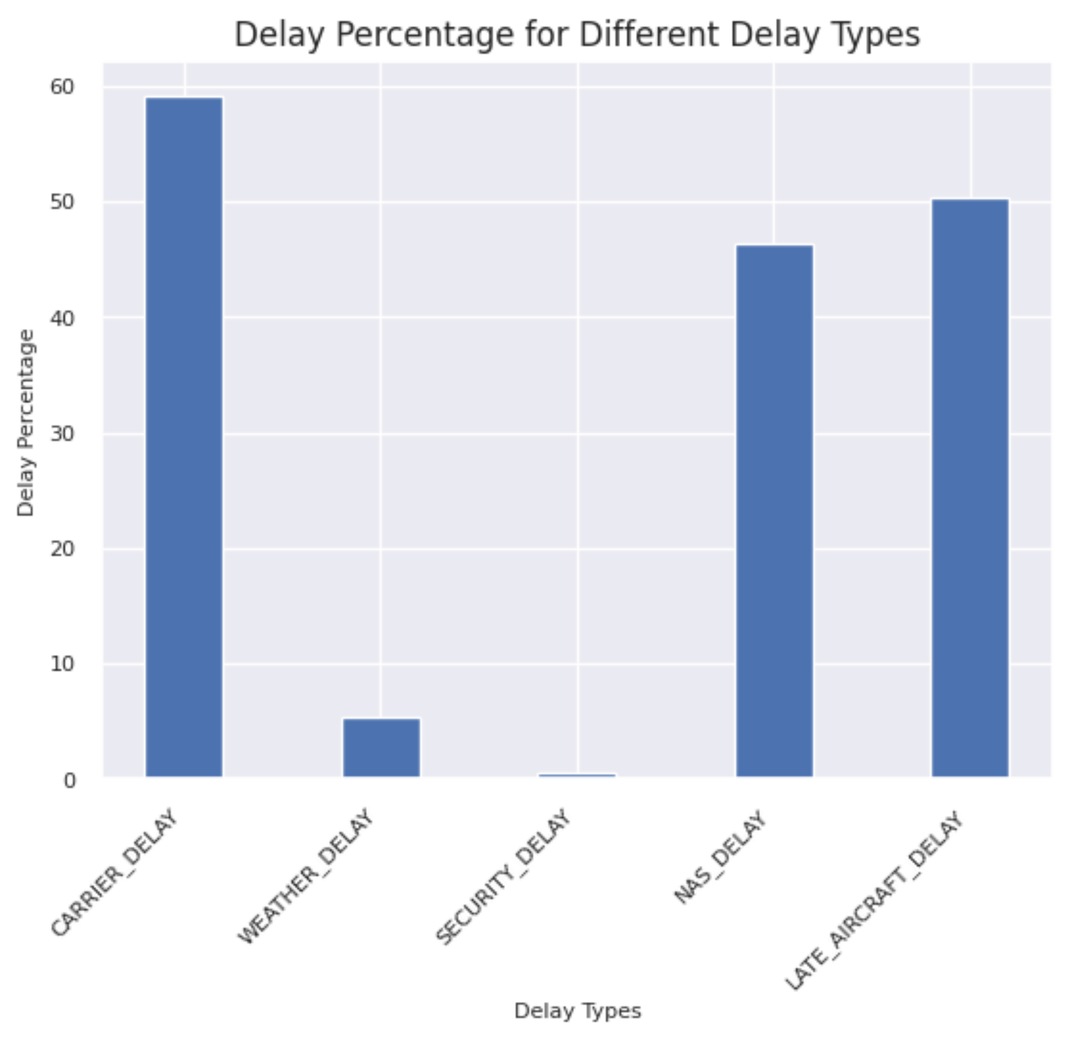
Departure delay and arrival delay are highly correlated with arrival time, carrier delay, and late aircraft delay

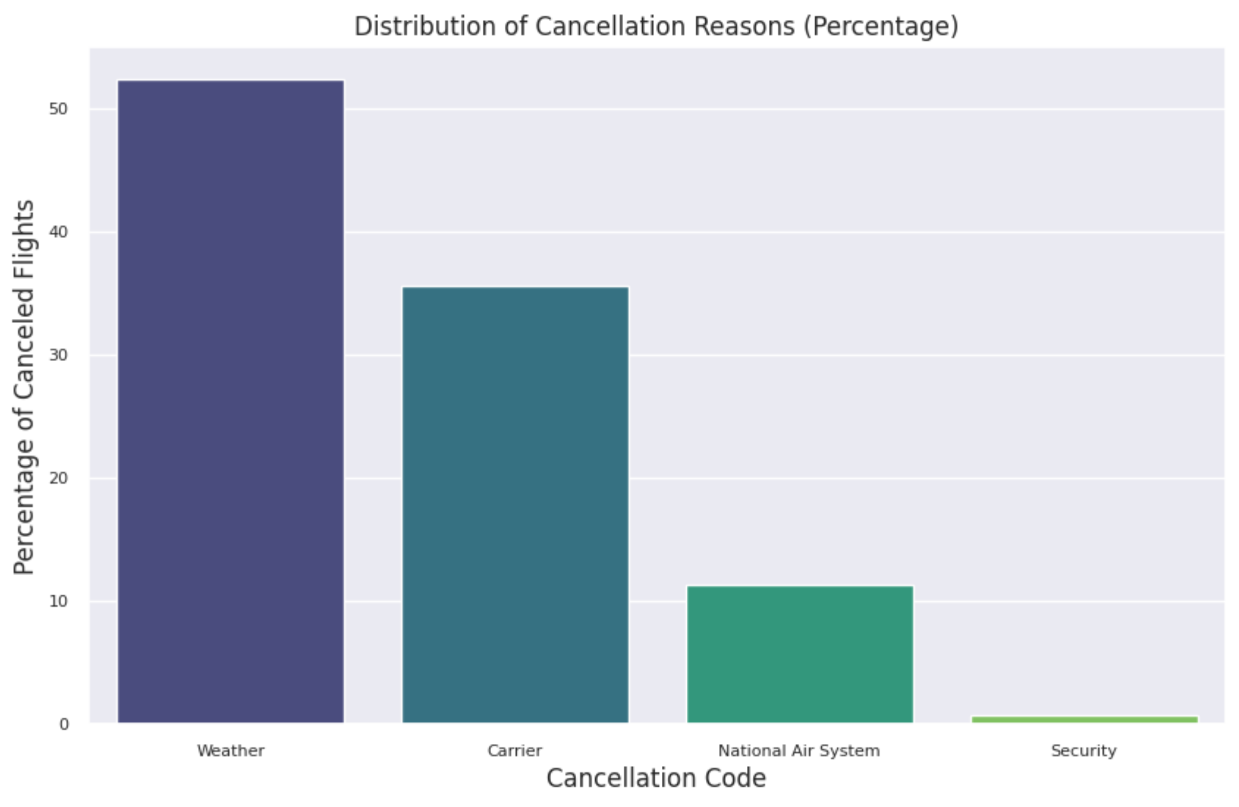
Study of month-wise percentage of delayed and cancelled flights



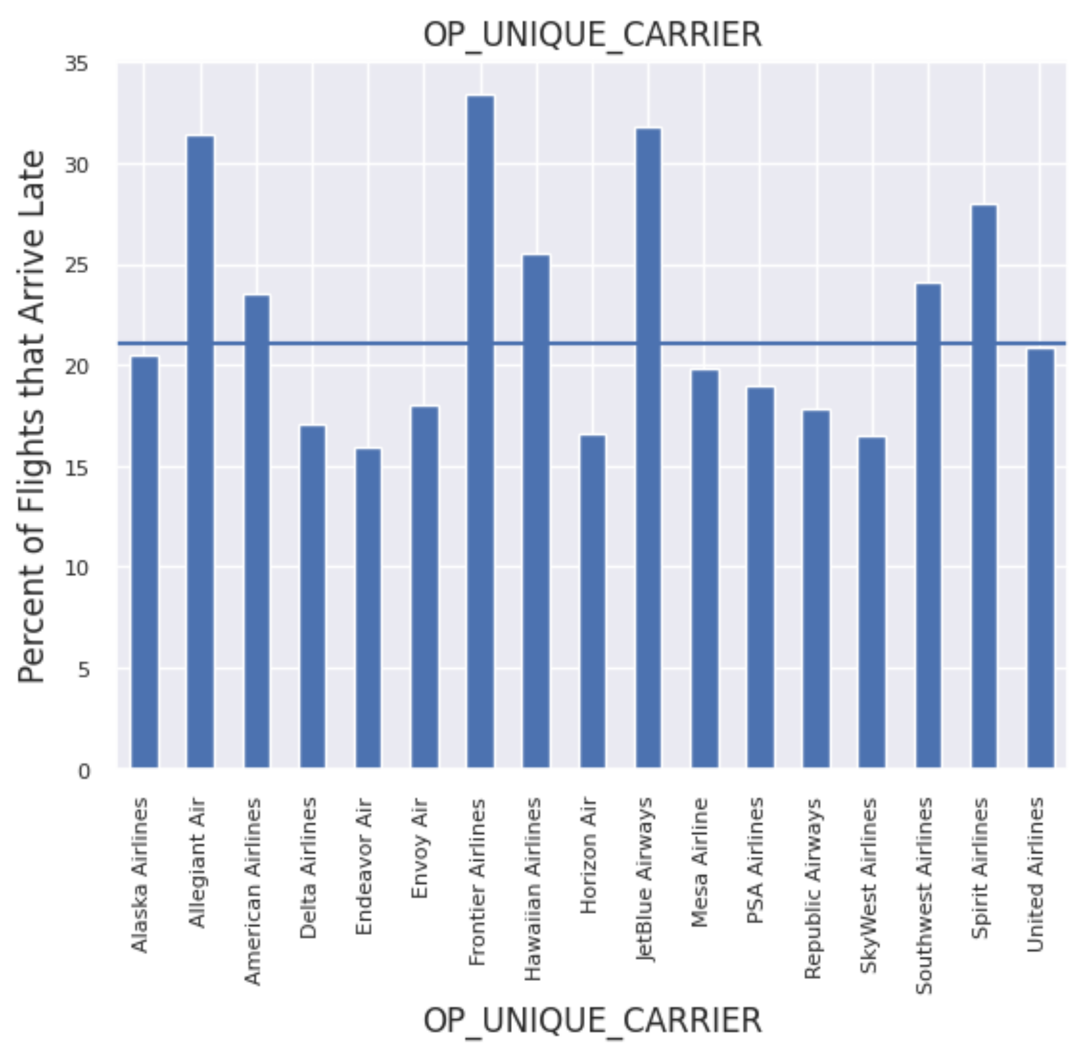


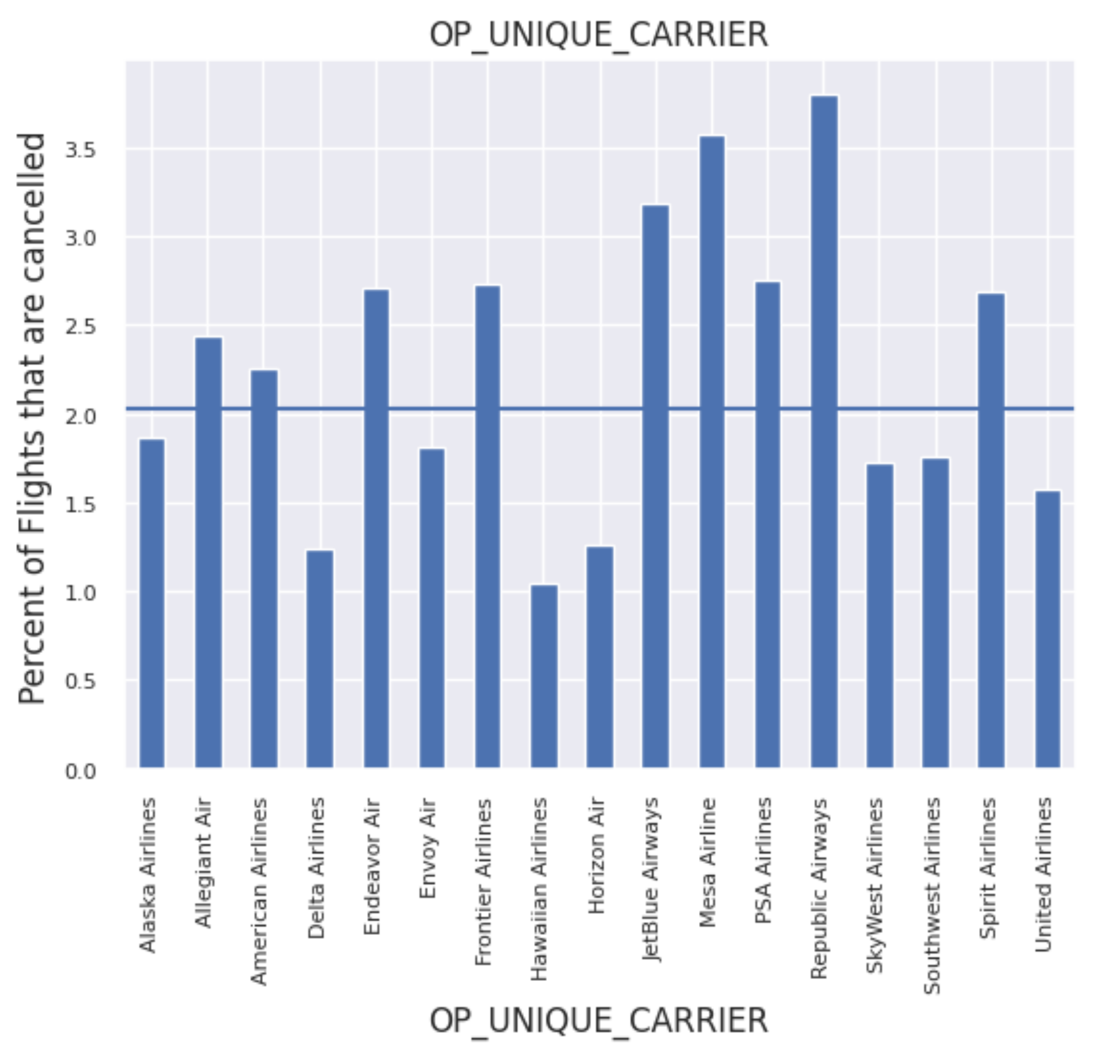
Study of delay and cancellation reasons



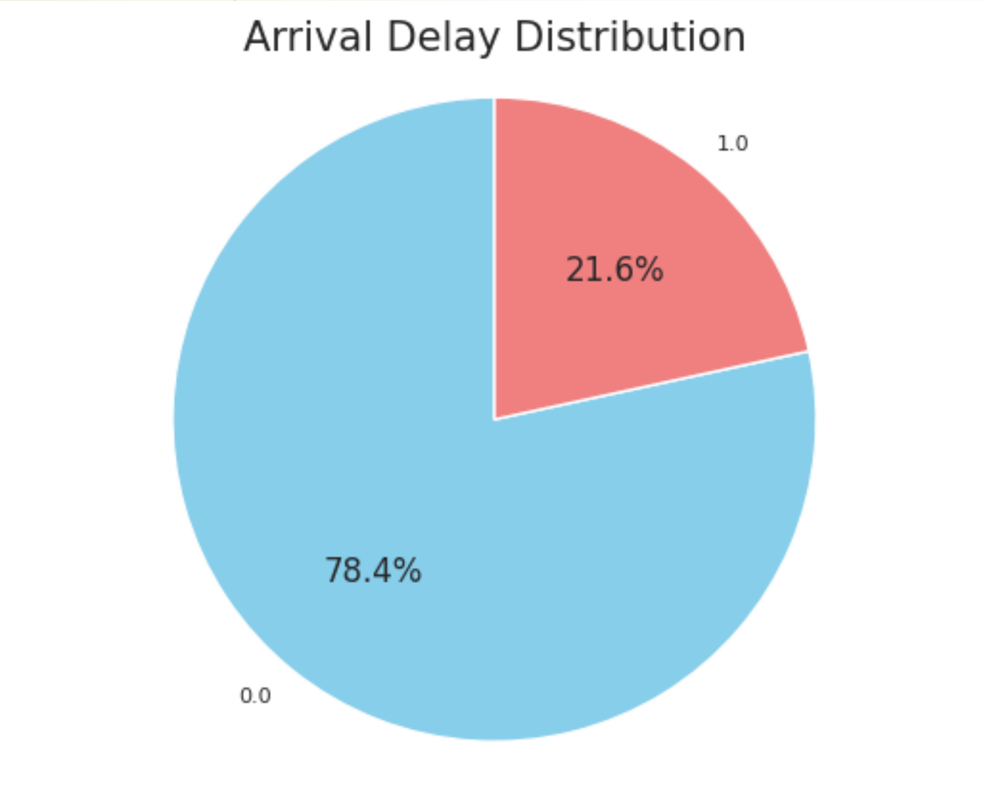


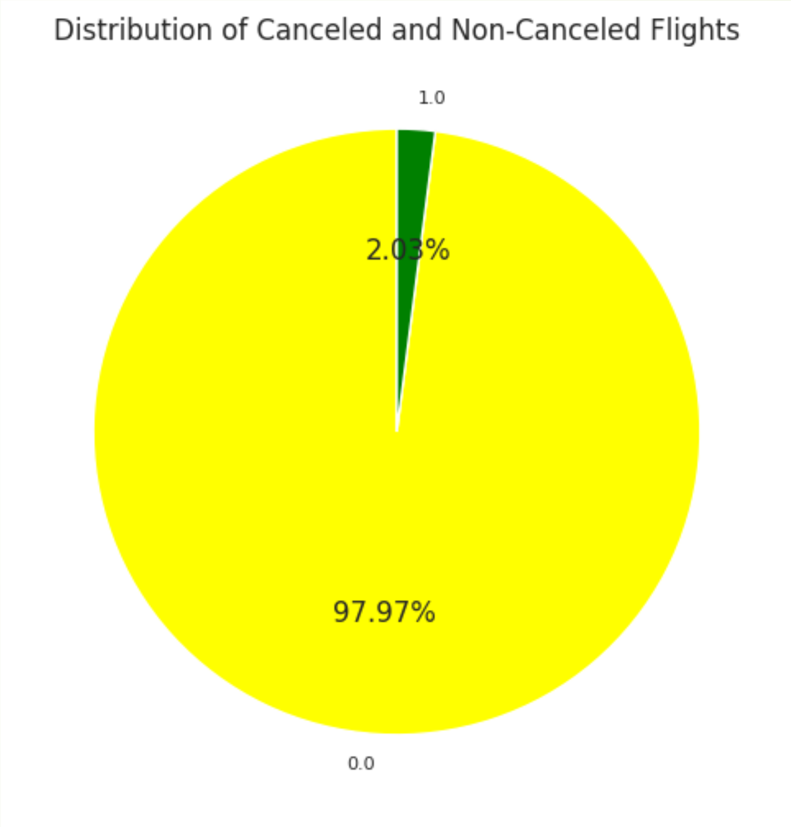
Carrier wise distribution of delayed and cancelled flights





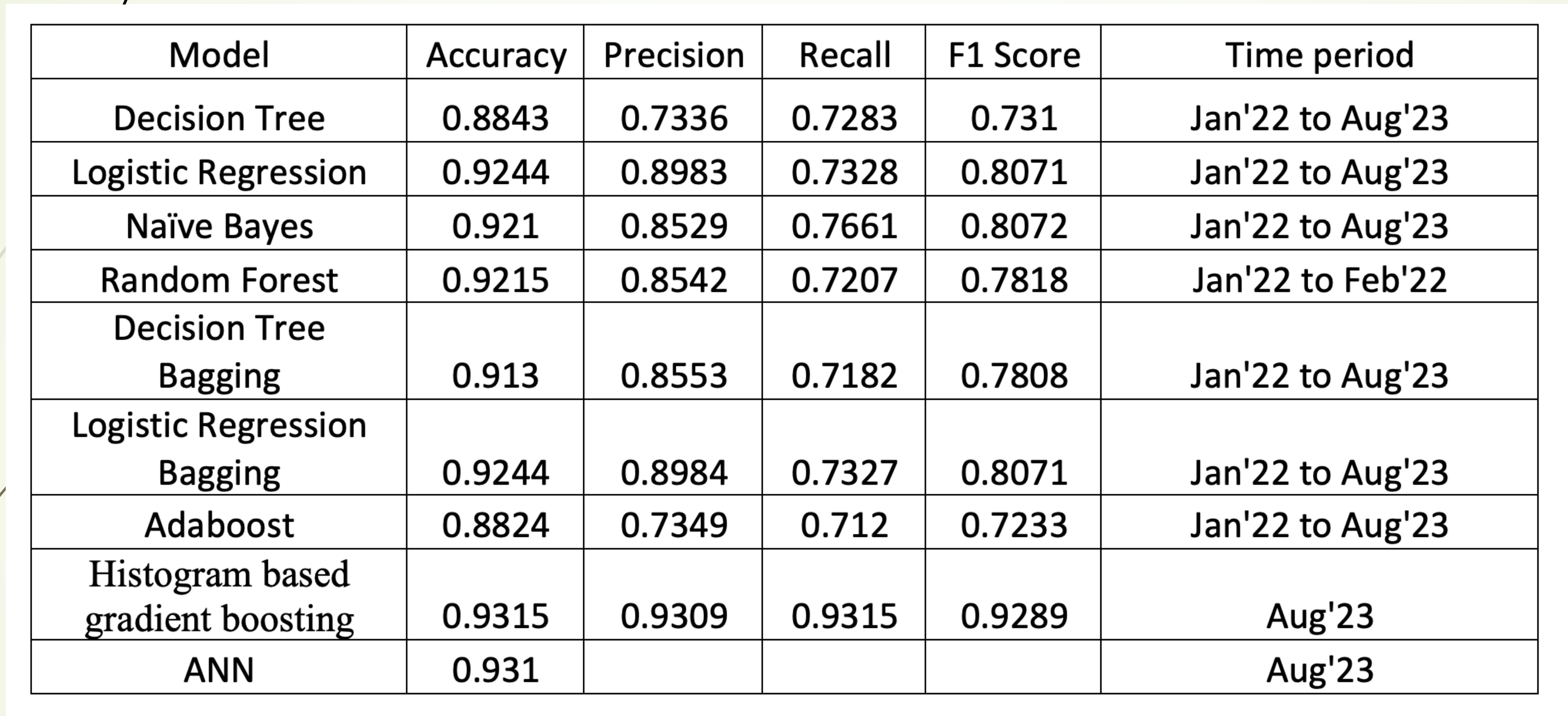
Delayed and cancelled flights distribution





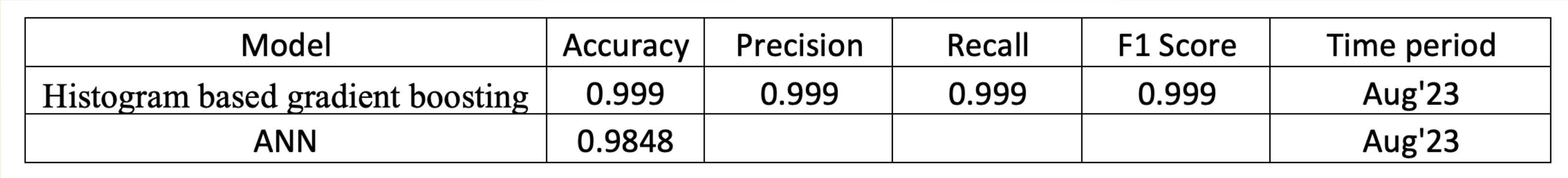
Classification Models

Delayed data results:



Model	Accuracy	Precision	Recall	F1 Score	Time period
Decision Tree	0.8843	0.7336	0.7283	0.731	Jan'22 to Aug'23
Logistic Regression	0.9244	0.8983	0.7328	0.8071	Jan'22 to Aug'23
Naïve Bayes	0.921	0.8529	0.7661	0.8072	Jan'22 to Aug'23
Random Forest	0.9215	0.8542	0.7207	0.7818	Jan'22 to Feb'22
Decision Tree Bagging	0.913	0.8553	0.7182	0.7808	Jan'22 to Aug'23
Logistic Regression Bagging	0.9244	0.8984	0.7327	0.8071	Jan'22 to Aug'23
Adaboost	0.8824	0.7349	0.712	0.7233	Jan'22 to Aug'23
Histogram based gradient boosting	0.9315	0.9309	0.9315	0.9289	Aug'23
ANN	0.931				Aug'23

Cancellation data results:



Model	Accuracy	Precision	Recall	F1 Score	Time period
Histogram based gradient boosting	0.999	0.999	0.999	0.999	Aug'23
ANN	0.9848				Aug'23

Thank You!