Data Mining on Airline on Time Dataset

By Mukund Kalantri, Rohit Kharat, and Reid Glaze

Intro

- In this project, we analyzed the “Airline Reporting Carrier On-Time Performance Dataset” by IBM.
- We considered data from Jan 2022 till April 2022 which is approximately 2 million records, and analyzed the delays and cancellations of these flights.
- We aim to provide useful information to consumers, airline companies, and airports to make important decisions.
- We also made a classification model which predicts the reason for cancellation of a flight and delay prediction model.

Data Preparation

- We had four different files having data for each month.
- We combined all the data and also removed those attributes which were unnecessary for our analysis.
- We also used data from Department of Transportation to get the names of airlines for the airline codes we had in our dataset.

Tools Used

We used python and Google Colab for our project.

We used many libraries in Python like -

- Numpy, and Pandas for handling data
- Scipy and Scikit Learn for statistics
- Matplotlib for evaluation
- Keras and Tensorflow for Machine Learning

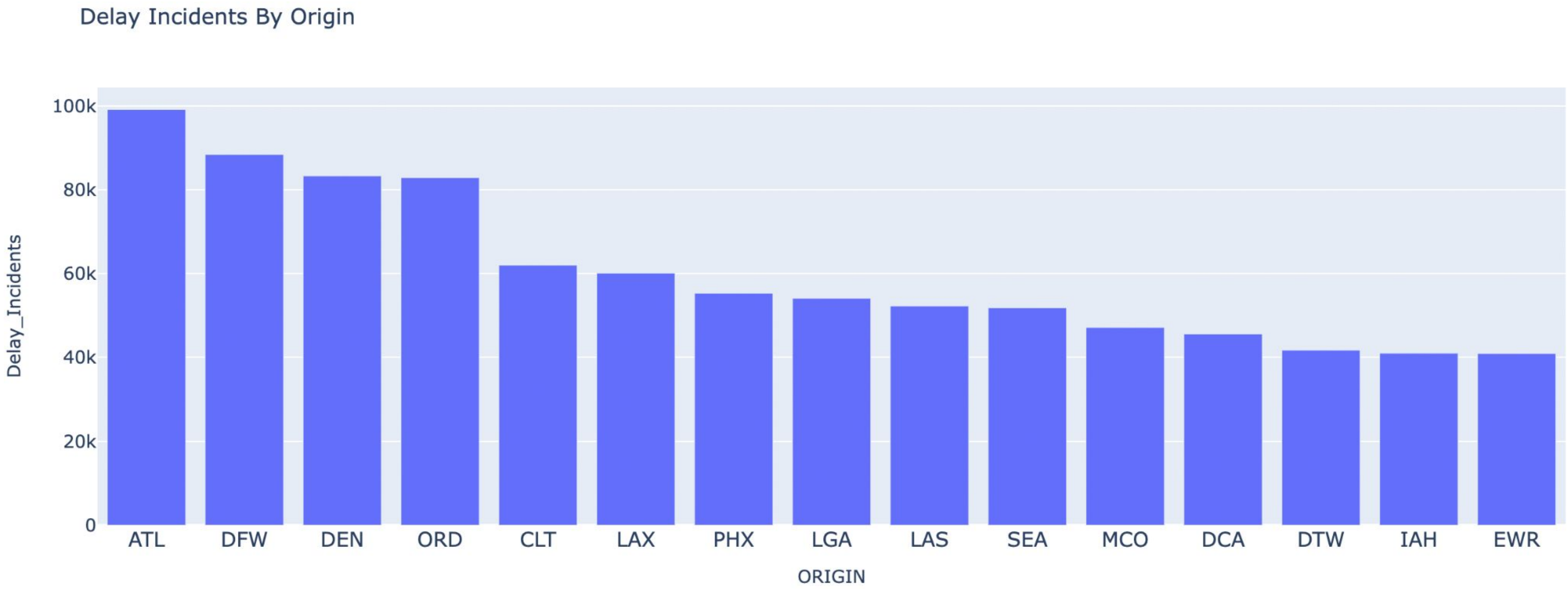
Question 1

Which airports should be focused more on infrastructure development to mitigate the frequency and severity of delays?

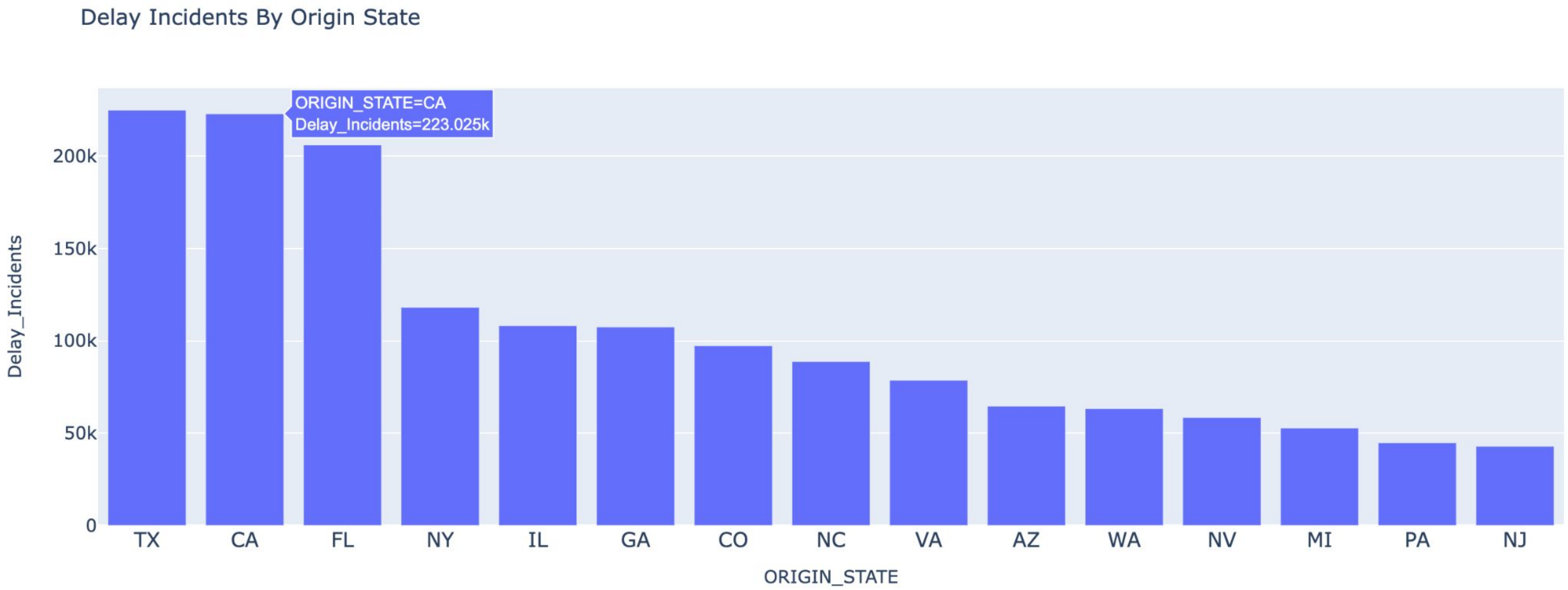
Question 1

1. Filter data based on origin airports, states, and airlines
2. Use bar chart to assess the number of incidents based on airports, states, and reporting airline
3. Airports to be focused more on infrastructure development; mitigate the frequency and severity of delays

Question 1

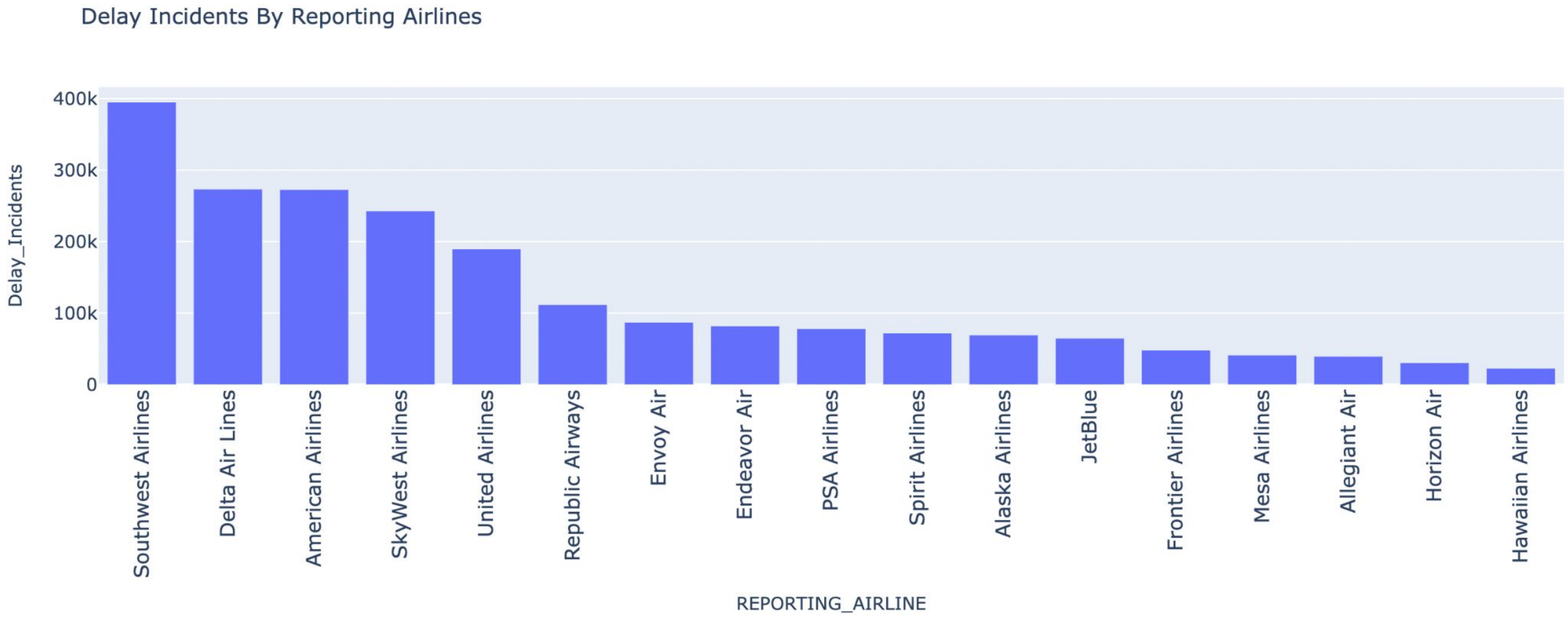

Delay Incidents By Origin
Delay_Incidents
100k
80k
60k
40k
20k
0
ATL
DFW
DEN
ORD
CLT
LAX
PHX
LGA
LAS
SEA
MCO
DCA
DTW
IAH
EWR
ORIGIN

Question 1



Delay Incidents By Origin State

Question 1



Delay Incidents By Reporting Airlines
Delay_Incidents
400k
300k
200k
100k
0
Southwest Airlines
Delta Air Lines
American Airlines
SkyWest Airlines
United Airlines
Republic Airways
Envoy Air
Endeavor Air
PSA Airlines
Spirit Airlines
Alaska Airlines
JetBlue
Frontier Airlines
Mesa Airlines
Allegiant Air
Horizon Air
Hawaiian Airlines
REPORTING_AIRLINE

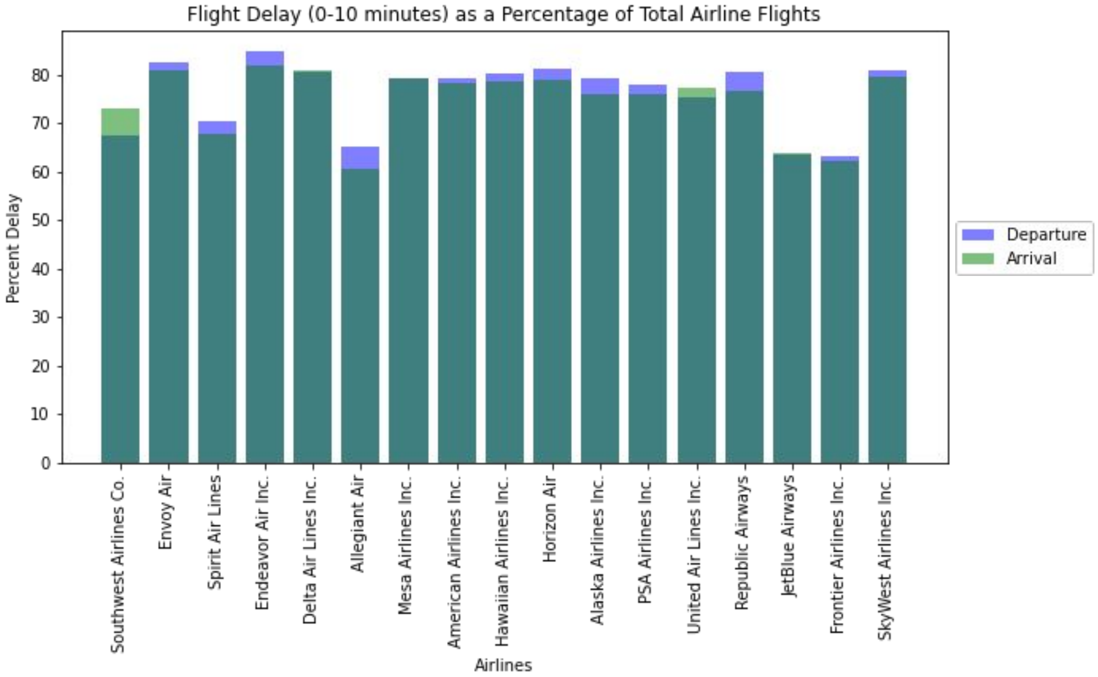
Question 2

What are the relationships between airline carriers and flight delays/cancellations so that airlines could improve their services and customers can make informed decisions about which airlines should be preferred?

Question 2

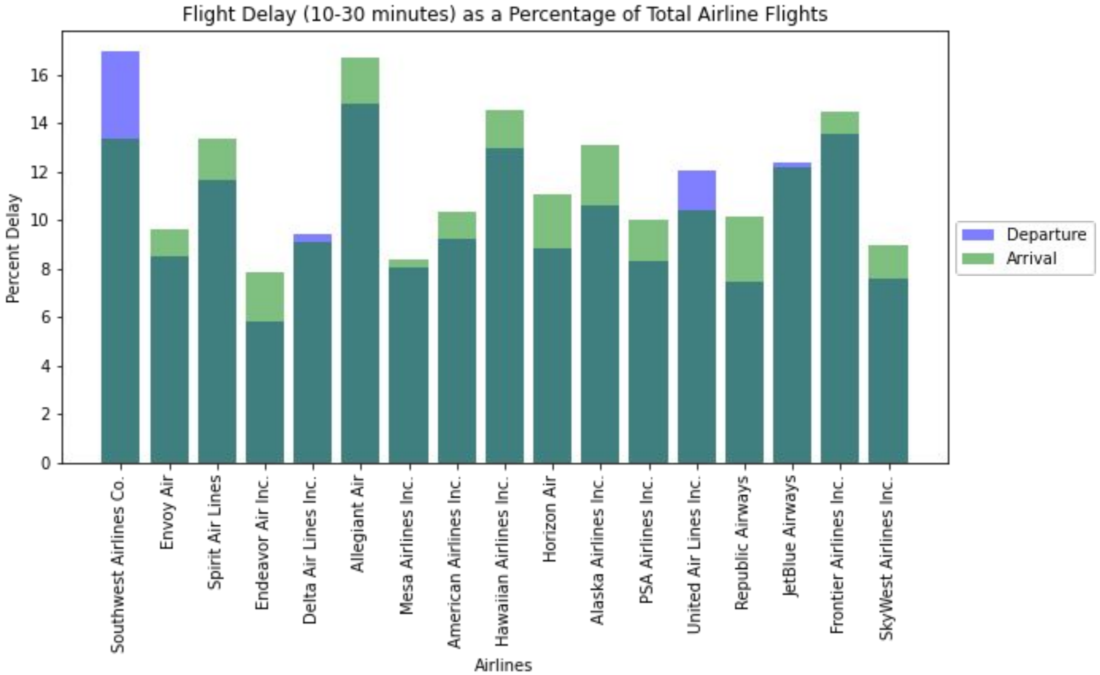
1. Filter out delay specific data
2. For each unique airline:
    a. Slice data for that airline
    b. Count the number of arrival and departure flights delayed for different lengths of time
    c. Calculate the percentage

Question 2

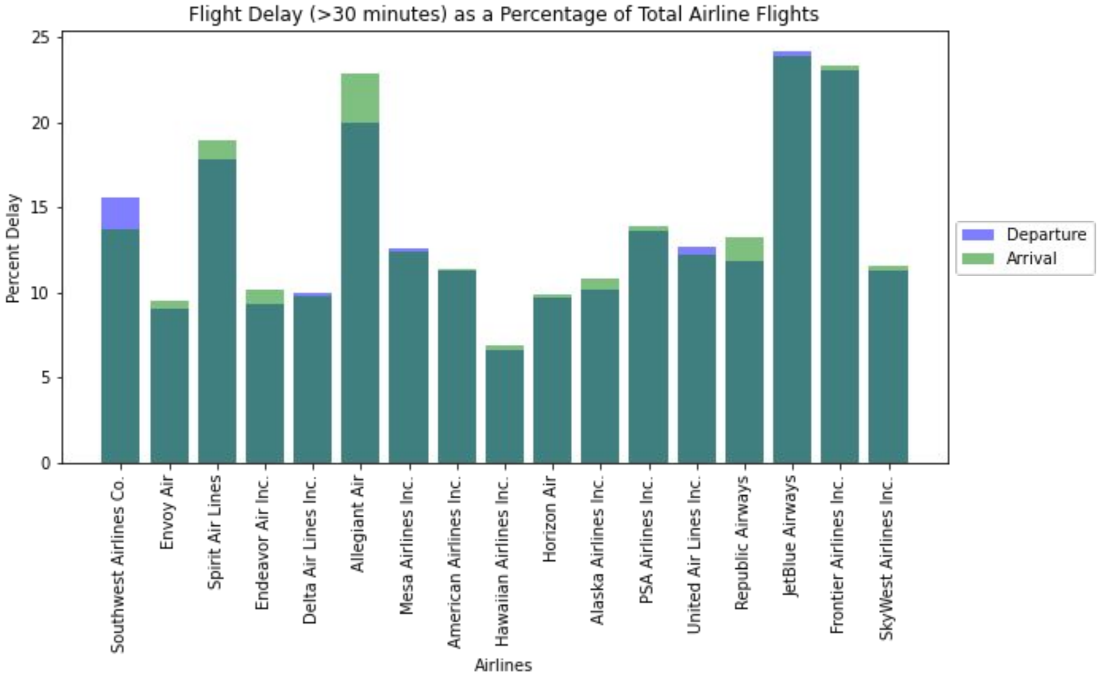


Flight Delay (0-10 minutes) as a Percentage of Total Airline Flights
Percent Delay
0
10
20
30
40
50
60
70
80
Airlines
Southwest Airlines Co.
Envoy Air
Spirit Air Lines
Endeavor Air Inc.
Delta Air Lines Inc.
Allegiant Air
Mesa Airlines Inc.
American Airlines Inc.
Hawaiian Airlines Inc.
Horizon Air
Alaska Airlines Inc.
PSA Airlines Inc.
United Air Lines Inc.
Republic Airways
JetBlue Airways
Frontier Airlines Inc.
SkyWest Airlines Inc.
Departure
Arrival

Question 2


Flight Delay (10-30 minutes) as a Percentage of Total Airline Flights
Percent Delay
0
2
4
6
8
10
12
14
16
Airlines
Southwest Airlines Co.
Envoy Air
Spirit Air Lines
Endeavor Air Inc.
Delta Air Lines Inc.
Allegiant Air
Mesa Airlines Inc.
American Airlines Inc.
Hawaiian Airlines Inc.
Horizon Air
Alaska Airlines Inc.
PSA Airlines Inc.
United Air Lines Inc.
Republic Airways
JetBlue Airways
Frontier Airlines Inc.
SkyWest Airlines Inc.
Departure
Arrival

Question 2


Flight Delay (>30 minutes) as a Percentage of Total Airline Flights
Percent Delay
0
5
10
15
20
25
Airlines
Southwest Airlines Co.
Envoy Air
Spirit Air Lines
Endeavor Air Inc.
Delta Air Lines Inc.
Allegiant Air
Mesa Airlines Inc.
American Airlines Inc.
Hawaiian Airlines Inc.
Horizon Air
Alaska Airlines Inc.
PSA Airlines Inc.
United Air Lines Inc.
Republic Airways
JetBlue Airways
Frontier Airlines Inc.
SkyWest Airlines Inc.
Departure
Arrival

Question 3

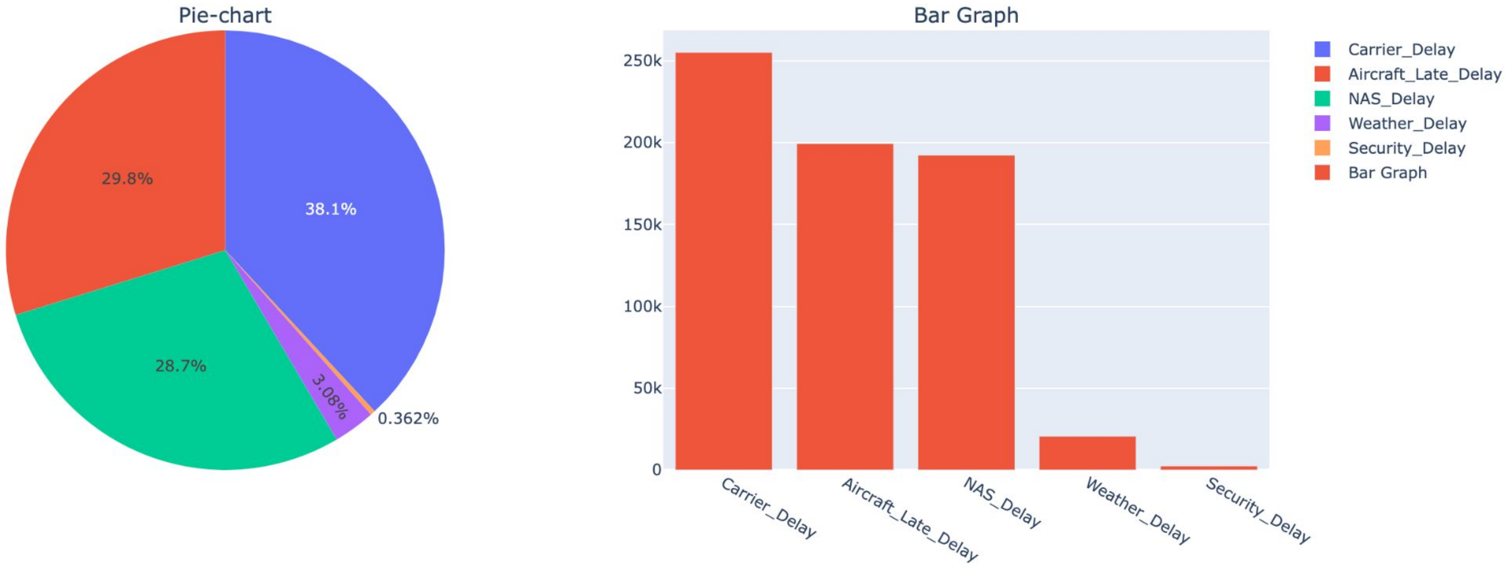
What are the most common reasons for flight delays based on airports, so that the airport authorities could improve their services?

Question 3

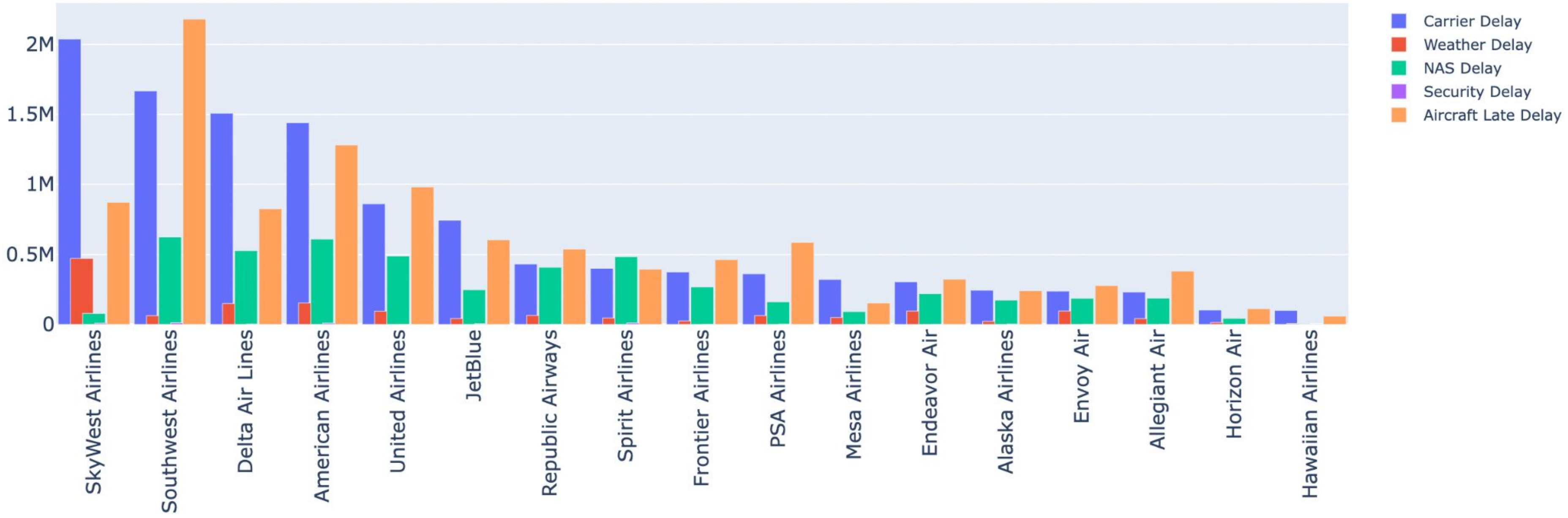
1. Filter data to find common reasons for delays
2. Use bar chart and pie chart to understand the common reasons of delays for a particular airline or airport
3. Helpful to understand which airports are affected the most by a particular type of delay and to provide relevant information to airport authorities to work on improving their services

Question 3

Common Reasons for Delay



Question 3



2M
1.5M
1M
0.5M
0
SkyWest Airlines
Southwest Airlines
Delta Air Lines
American Airlines
United Airlines
JetBlue
Republic Airways
Spirit Airlines
Frontier Airlines
PSA Airlines
Mesa Airlines
Endeavor Air
Alaska Airlines
Envoy Air
Allegiant Air
Horizon Air
Hawaiian Airlines
Carrier Delay
Weather Delay
NAS Delay
Security Delay
Aircraft Late Delay

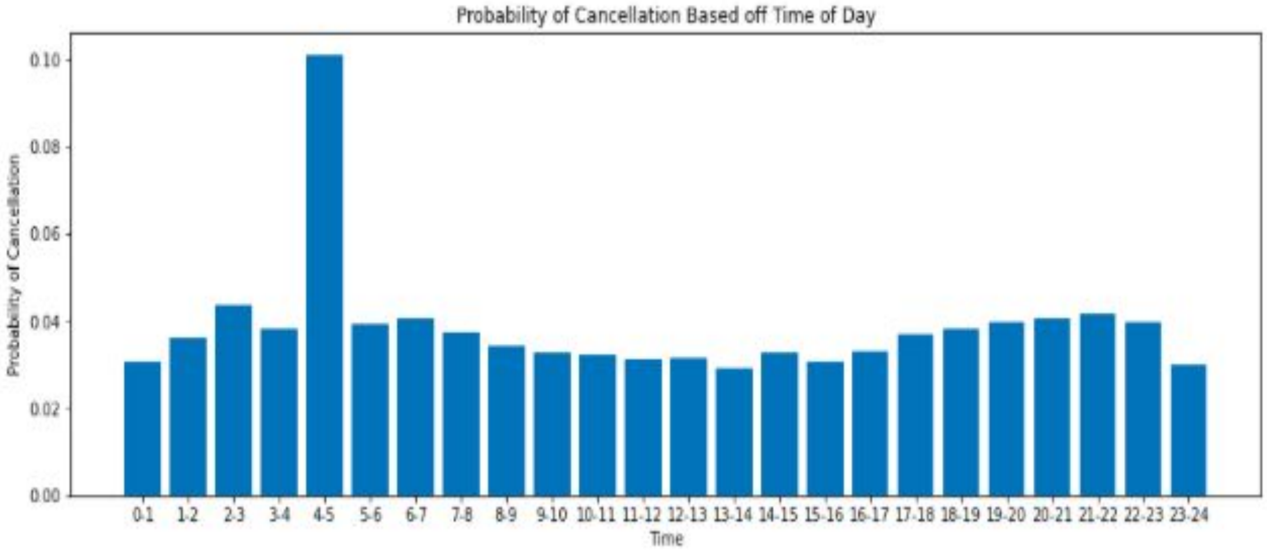
Question 4

What are the relationships between times in a day and days of the week when there are a lot of cancellations and delays so that customers can make an informed decision while booking flights concerning their timings?

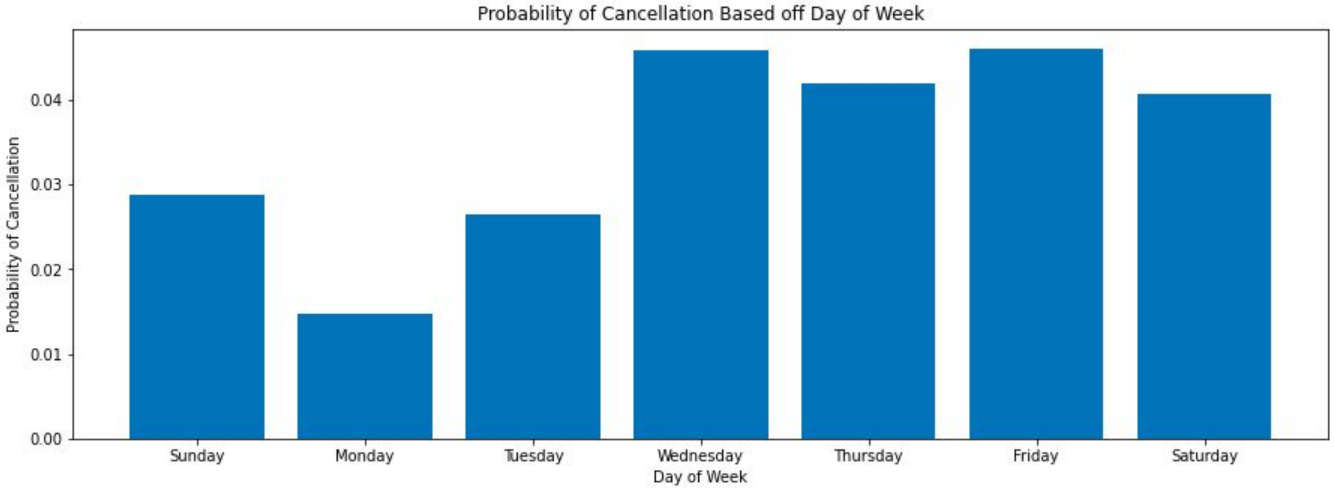
Question 4

1) Categorize by day of the week or hour of the day. Count the number of objects in each category.
2) Filter by cancellation or type of delay (short, medium, long).
3) Categorize the filtered data by day of the week or hour of the day and count the number of objects in each category.
4) Divide the categorized filtered data by the categorized unfiltered data to get probability data.

Probability of Cancellation (by the Hour)



Probability of Cancellation (Day of Week)



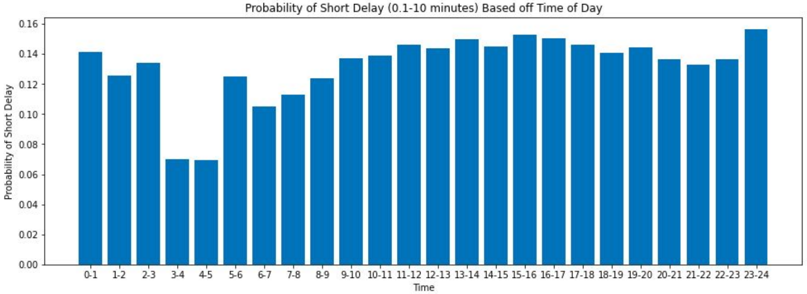
Probability of Delay (By the Hour)

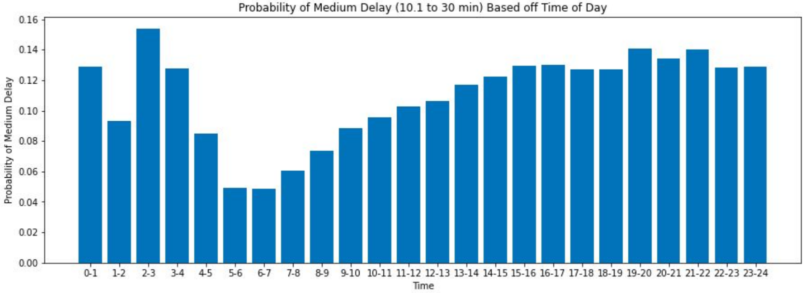
Top to bottom:

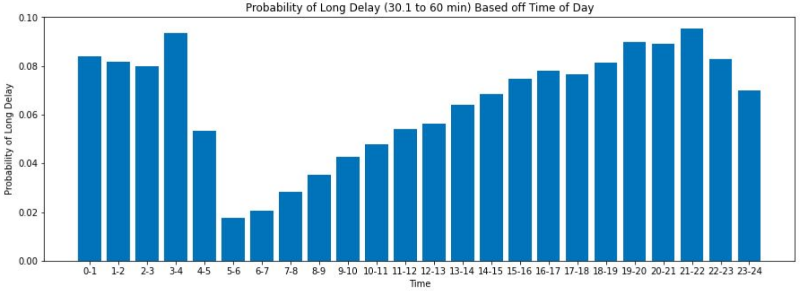
Short delay: 0-10 mins

Medium delay: 10-30 mins

Long delay: 30-60 mins







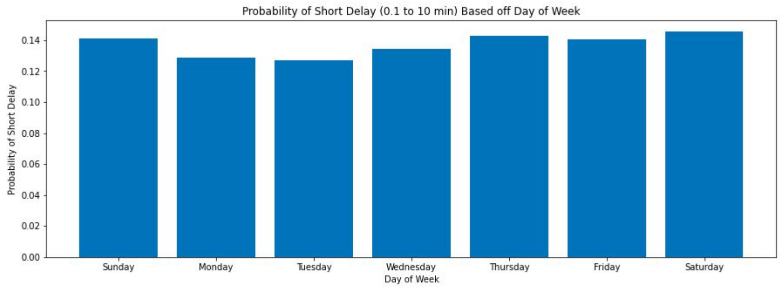
Probability of Delay (Day of Week)

Top to bottom:

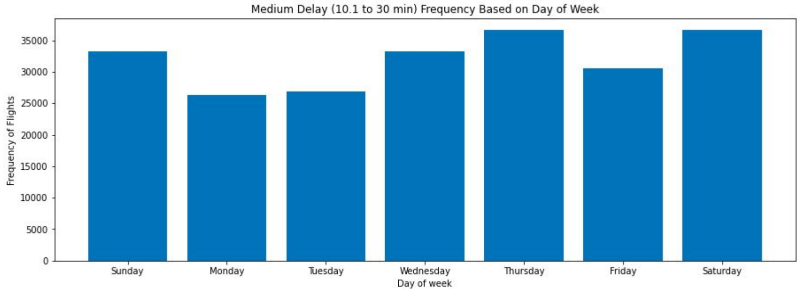
Short delay: 0-10 mins

Medium delay: 10-30 mins

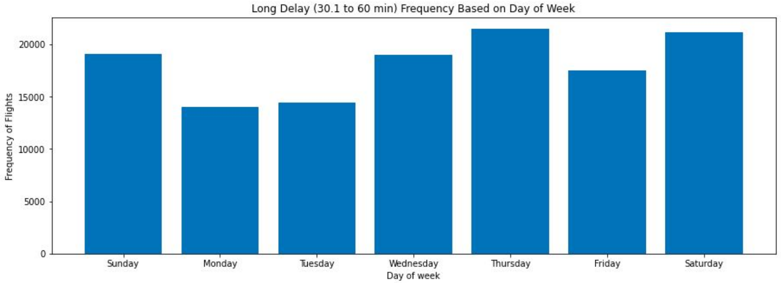
Long delay: 30-60 mins



Probability of Short Delay (0.1 to 10 min) Based off Day of Week



Medium Delay (10.1 to 30 min) Frequency Based on Day of Week



Long Delay (30.1 to 60 min) Frequency Based on Day of Week

Classification for Reason of Flight Cancellation

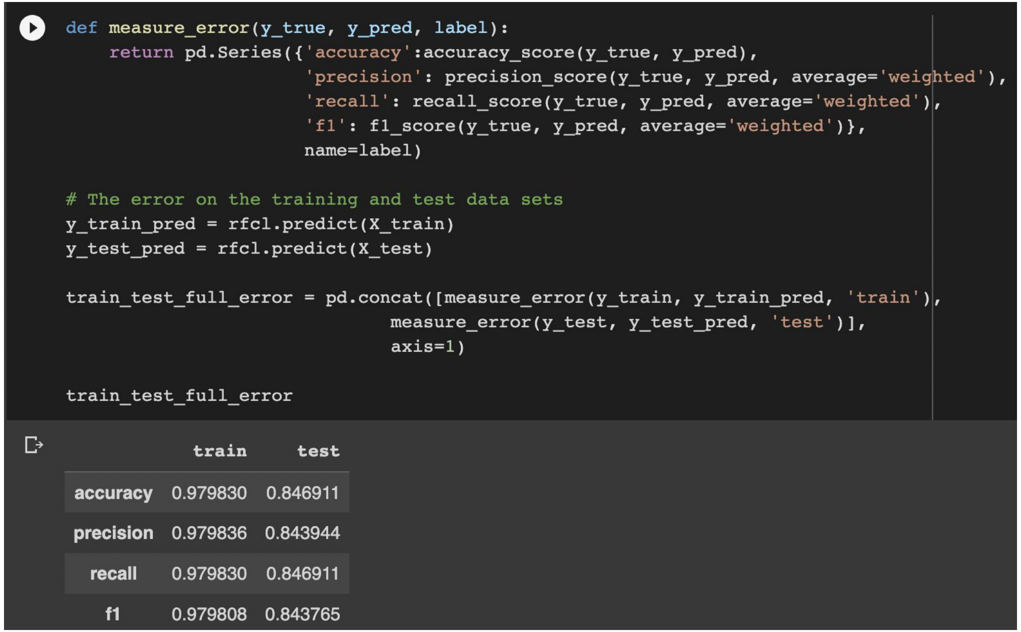
- Several models were built for multi-class classification task of classifying cancellation code.
- This task helped us in identifying which attributes leads to cancellation of flight.
- Also, this helped us to understand the reason of cancellation based on the model and infer the pattern behind cancellations

Classification for Reason of Flight Cancellation

The models which we built:

- Decision Tree
- Random Forest Classifier
- Extra Trees Classifier
- Gradient Boosting Classifier
- XGBoost Classifier
- AdaBoost Classifier
- KNN
- Naive Bayes
- Logistic Regression

Classification for Reason of Flight Cancellation



```python
def measure_error(y_true, y_pred, label):
    return pd.Series({'accuracy':accuracy_score(y_true, y_pred),
                      'precision': precision_score(y_true, y_pred, average='weighted'),
                      'recall': recall_score(y_true, y_pred, average='weighted'),
                      'f1': f1_score(y_true, y_pred, average='weighted')},
                      name=label)

# The error on the training and test data sets
y_train_pred = rfcl.predict(X_train)
y_test_pred = rfcl.predict(X_test)

train_test_full_error = pd.concat([measure_error(y_train, y_train_pred, 'train'),
                              measure_error(y_test, y_test_pred, 'test')],
                              axis=1)

train_test_full_error
```

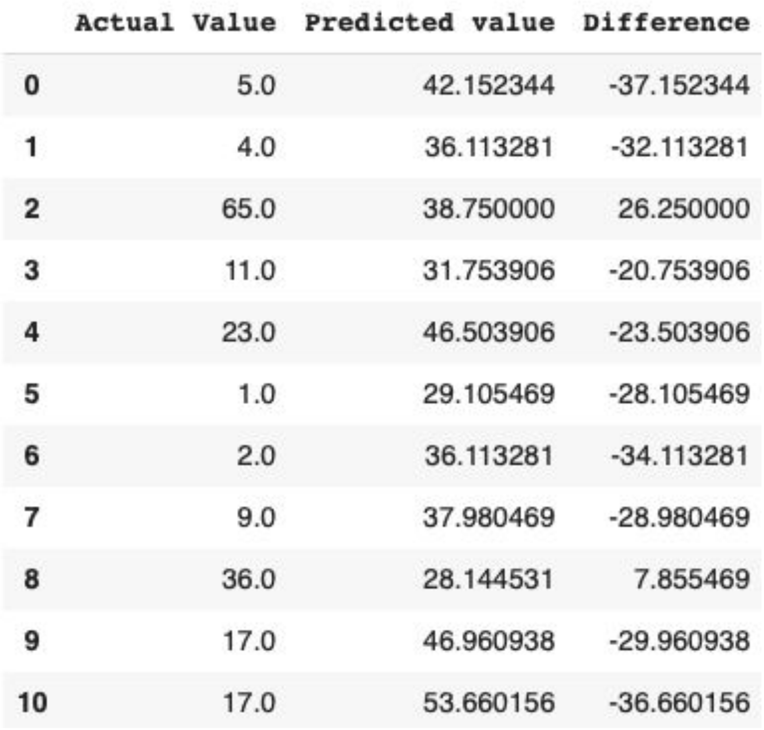
	train	test
accuracy	0.979830	0.846911
precision	0.979836	0.843944
recall	0.979830	0.846911
f1	0.979808	0.843765

Delay Prediction Model

- Multiple Linear Regression
- Neural Networks

Multiple Linear Regression

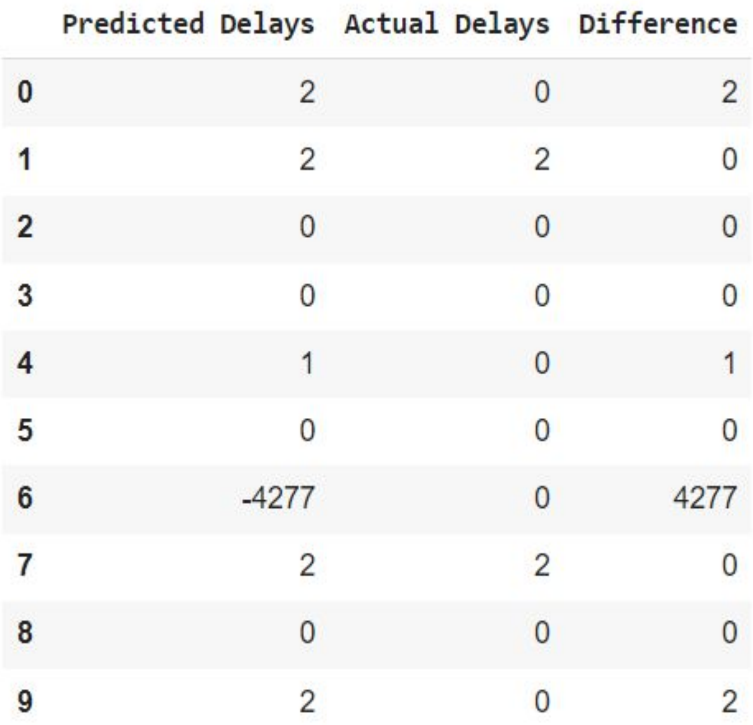
- Weak performance
- One-hot encoding for categorical attributes
- Curse of dimension
- Only two attributes considered



	Actual Value	Predicted value	Difference
0	5.0	42.152344	-37.152344
1	4.0	36.113281	-32.113281
2	65.0	38.750000	26.250000
3	11.0	31.753906	-20.753906
4	23.0	46.503906	-23.503906
5	1.0	29.105469	-28.105469
6	2.0	36.113281	-34.113281
7	9.0	37.980469	-28.980469
8	36.0	28.144531	7.855469
9	17.0	46.960938	-29.960938
10	17.0	53.660156	-36.660156

Neural Networks

- Good performance
- Label encoding for categorical attributes
- No curse of dimension
- Many attributes considered
- Can be improved



	Predicted Delays	Actual Delays	Difference
0	2	0	2
1	2	2	0
2	0	0	0
3	0	0	0
4	1	0	1
5	0	0	0
6	-4277	0	4277
7	2	2	0
8	0	0	0
9	2	0	2

Conclusion

- Our analysis will help the airport as well as the airline authorities to make informed decisions and give efficient updates to passengers

- Our classification and prediction model will be useful for understanding the reasons behind delays/cancellations as well as we will able to predict those for passengers

Thank You