Big Data Project

Meina Huang

-> Milestone 1

1. Description

The dataset contains comprehensive flight information, documenting cancellations and delays across different airlines, from 2018 to 2022. This large dataset provides an opportunity for analyzing and predicting flight delays and cancellation probabilities, making it possible to draw insights that can enhance operational efficiency and customer experience within this industry.

2. Data
    a. Location

The dataset can be found on Kaggle:
https://www.kaggle.com/datasets/robikscube/flight-delay-dataset-20182022/data?select=raw

I will be combining the data available for 2018-2022, giving me around 10.2 GB of data to work with.

    b. Attributes

Each year's (5 in total) dataset consists of 60 columns, main ones are:

- FlightDate: The date of the flight.
- Airline: The airline operating the flight.
- Origin: The departure airport.
- Dest: The destination airport.
- Cancelled: Whether the flight was canceled (True/False).
- Diverted: Whether the flight was diverted (True/False).
- CRSDepTime: The scheduled departure time (local time: hhmm).
- DepTime: The actual departure time (local time: hhmm).
- DepDelayMinutes: Difference in minutes between scheduled and actual departure time. Early departures set to 0.
- DepDelay: Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.

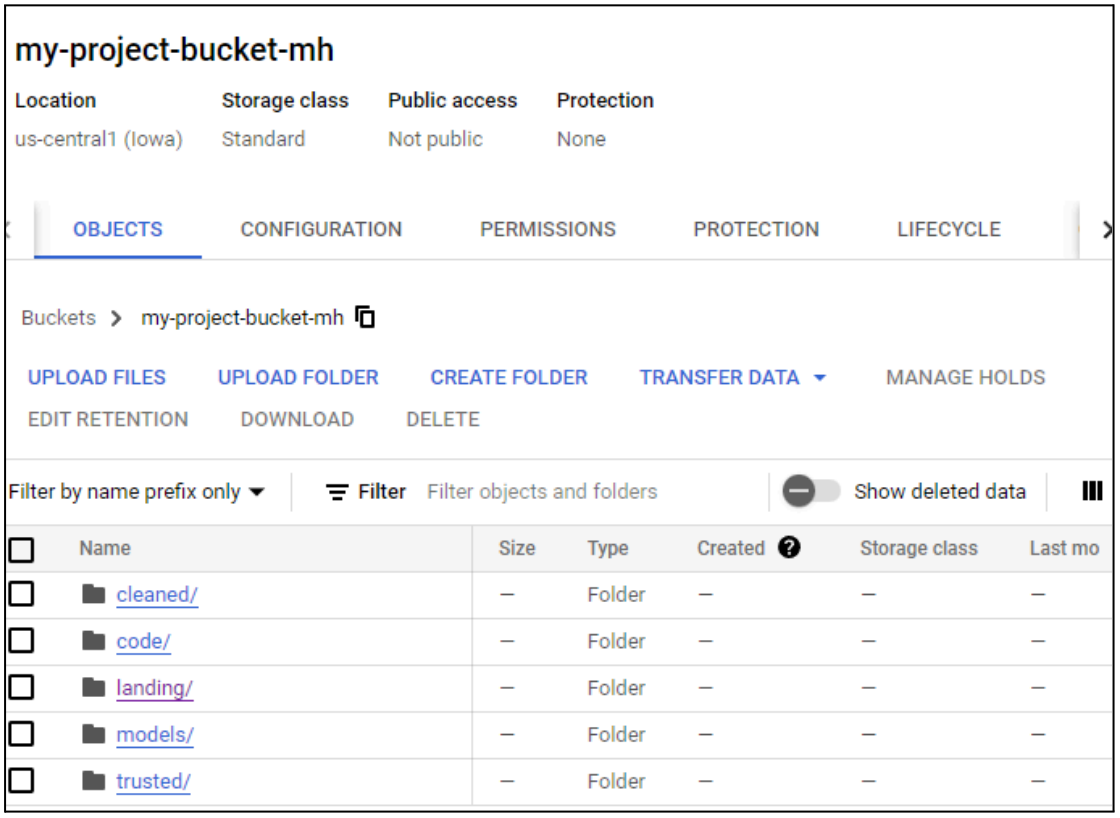
    c. Usage

The amount of people that travel by plane has been increasing and planes have become an important mode of transportation, therefore the prediction of flight delays and cancellations can be useful for customers and airlines. With that, I plan to predict the likelihood of departure delays, utilizing factors such as airline, origin, and flight date.

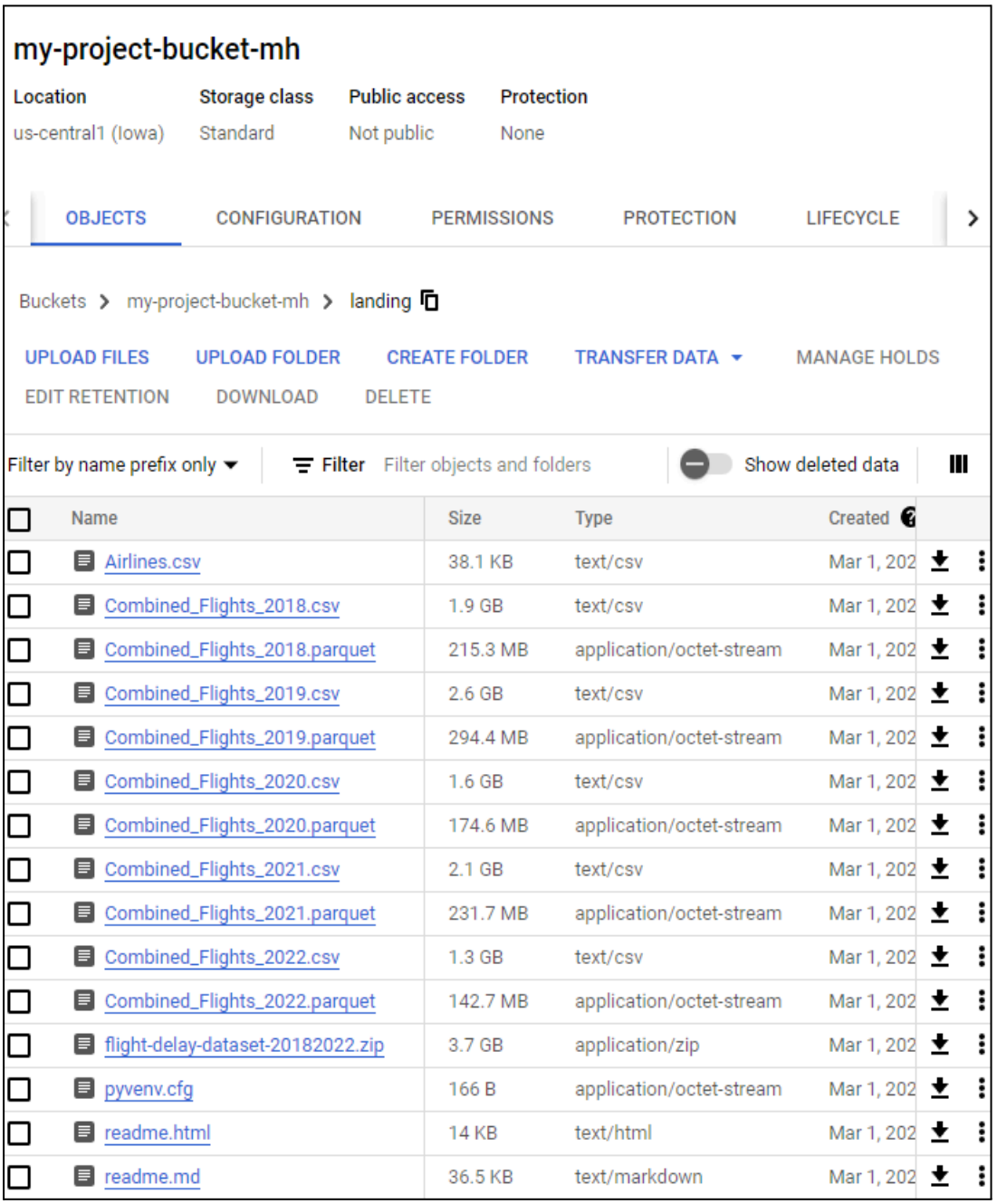
-> Milestone 2

During this milestone, I will gather data from Kaggle and store it in a bucket within GCS. To start, I will get a token from Kaggle and set up an instance in Google VM. I will put the data into a file named "landing" within a bucket named "my project bucket mh". Following this, I will create additional files, such as "cleaned", "code", "models", and "trusted", within the same bucket.

1. GCS Bucket



2. Landing Folder



-> Milestone 3

1. Preparing the Data

A Python script was created to load and clean the dataset from GCS. The cleaned data is stored in the /cleaned folder. Using a Dataproc cluster, a descriptive statistics was produced on the data and generated several graphs to get a general visualization of the data. These include an overview of flights by year, month, and airline. Also, relationships such as canceled vs. diverted flights and the relationship between departure delay and arrival delay were analyzed. Analysis of flights across days of the week, distribution of flight times, and distribution of arrival delays were also made. During the data cleaning process, specific columns were selected for analysis, and the remaining ones were not included. A decision was made to retain missing data in columns such as DepTime, DepDelay, DepDelayMinutes, ArrTime, ArrDelay, and ArrDelayMinutes, as deleting or substituting those for a mean or median could skew the analysis, considering that certain fields might be blank because of canceled flights.

2. Initial Overview

    a. Number of observations

29283782

    b. List of variables

```
FlightDate        datetime64
Airline           object
Origin            object
Cancelled         bool
Diverted          bool
CRSDepTime        int64
DepTime           float64
DepDelay          float64
DepDelayMinutes   float64
ArrTime           float64
ArrDelay          float64
Year              int64
Quarter           int64
Month             int64
DayofMonth        int64
DayOfWeek         int64
CRSArrTime        int64
ArrDelayMinutes   float64
```

    c. Number of missing in the observations

```
FlightDate              0
Airline                 0
Origin                  0
Cancelled               0
Diverted                0
CRSDepTime              0
DepTime            761652
DepDelay           763084
DepDelayMinutes    763084
ArrTime            786177
ArrDelay           846183
Year                    0
Quarter                 0
Month                   0
DayofMonth              0
DayOfWeek               0
CRSArrTime              0
ArrDelayMinutes    846183
```

d. Min/max/avg/stdev for all numeric variables

```
cleaned_flights.parquet Summary statistics for numeric variables:
       CRSDepTime      DepTime     DepDelay  DepDelayMinutes      ArrTime
min      1.000000     1.000000 -1280.000000         0.000000     1.000000
max   2359.000000  2400.000000  7223.000000      7223.000000  2400.000000
mean  1326.261738  1329.295567     9.306866        12.783107  1468.046257
std    482.346300   494.975040    47.280106        46.173367   525.617712

         ArrDelay         Year   Quarter      Month  DayofMonth  DayOfWeek
min  -1290.000000  2018.000000  1.000000   1.000000    1.000000   1.000000
max   7232.000000  2022.000000  4.000000  12.000000   31.000000   7.000000
mean     3.609370  2019.828655  2.448568   6.327840   15.751611   3.974879
std     49.279165     1.342521  1.121206   3.452305    8.778879   2.002314

       CRSArrTime  ArrDelayMinutes
min      1.000000         0.000000
max   2400.000000      7232.000000
mean  1489.003923        12.809917
std    507.288029        45.799592
```

3. Graphs and Charts

a. Overview of Flights

- By Year

Throughout the five years, the percentage of flights that were on time remained relatively stable, varying between 59% and 66%, with an exception in 2020 where the on time percentage was higher at 76%. This could be due to many different factors, including reduced air traffic because of COVID-19 pandemic, resulting in smoother operations and fewer delays. In turn, the higher rate of flight cancellations during 2020 could also be a consequence of the pandemic, with reduced demand for air travel and more travel bans, leading to more cancellations.

Status	OnTime	Delayed	Cancelled
Year			
2018	64.115615	34.330820	1.553565
2019	64.619787	33.481609	1.898604
2020	76.052291	17.953459	5.994249
2021	65.698142	32.542981	1.758876
2022	58.801153	38.178190	3.020657

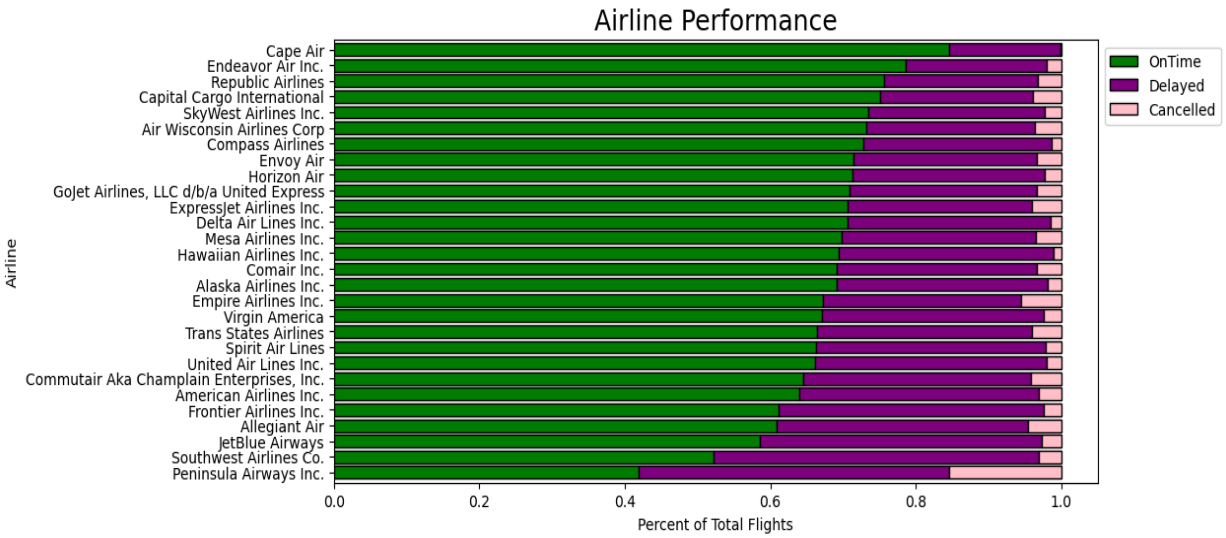
- By Month

September, October, and November tend to have higher rates of on time flights, possibly due to lower overall flight volume during this period. On the other hand, June, July, and August experience a higher rate of delays, which matches with the peak summer travel season when we have an increased flight demand and potential congestion. With that, September shows a lower rate of flight delays, likely due to reduced flight activity after summer. Additionally, the higher cancellation rates during the beginning of the year could be due to operational challenges coming from winter weather conditions.

Status	OnTime	Delayed	Cancelled
Month			
1	68.169907	28.848901	2.981192
2	64.960611	31.916149	3.123240
3	66.363322	28.061425	5.575252
4	64.436275	28.729969	6.833756
5	65.023841	33.064960	1.911198
6	59.488265	38.546106	1.965629
7	61.996702	36.320298	1.683000
8	64.622096	33.242882	2.135022
9	72.924881	25.680296	1.394822
10	69.570264	29.317708	1.112028
11	70.182986	28.976653	0.840361
12	64.138827	34.430717	1.430456

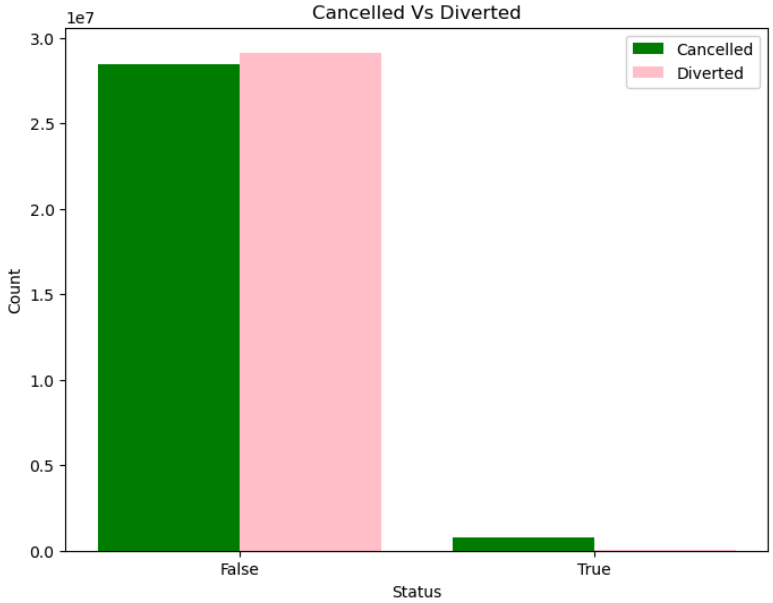
- By Airline

Most airlines show similar rates of on time, delayed, and canceled flights. However, Peninsula Airways, Southwest Airlines, Allegiant Air, and Empire Airlines show higher rates of delays and cancellations compared to other airlines.



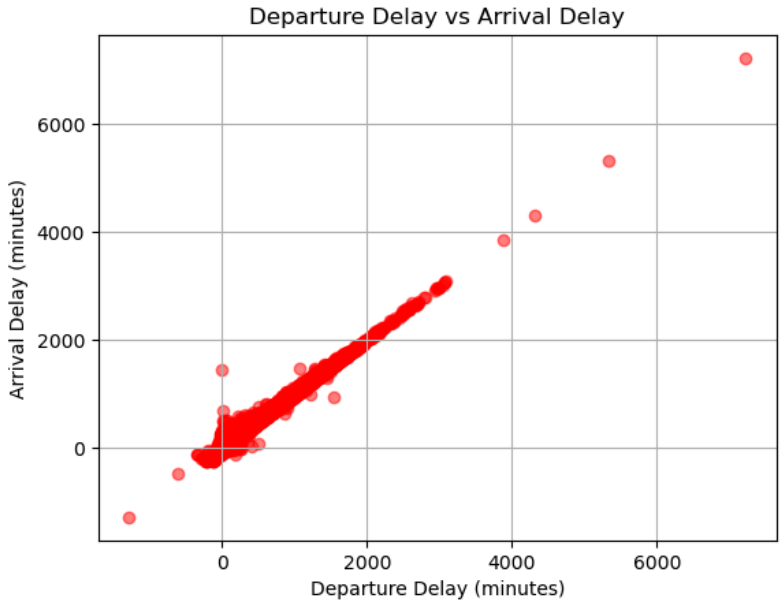
b. Canceled Vs Diverted

The majority of flights typically are not canceled or diverted. However, cancellations are more common than diversions in general, meaning that while both cancellations and diversions are relatively rare, cancellations are more usual than diversions.



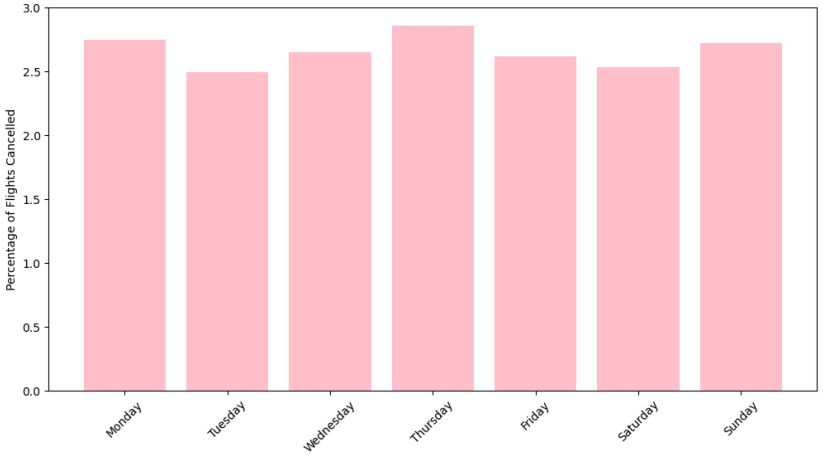
c. Relationship between Departure Delay and Arrival Delay

The graph shows a very strong positive linear relationship, with a few outliers, between departure delay and arrival delay. As departure delay increases, arrival delay tends to increase as well.



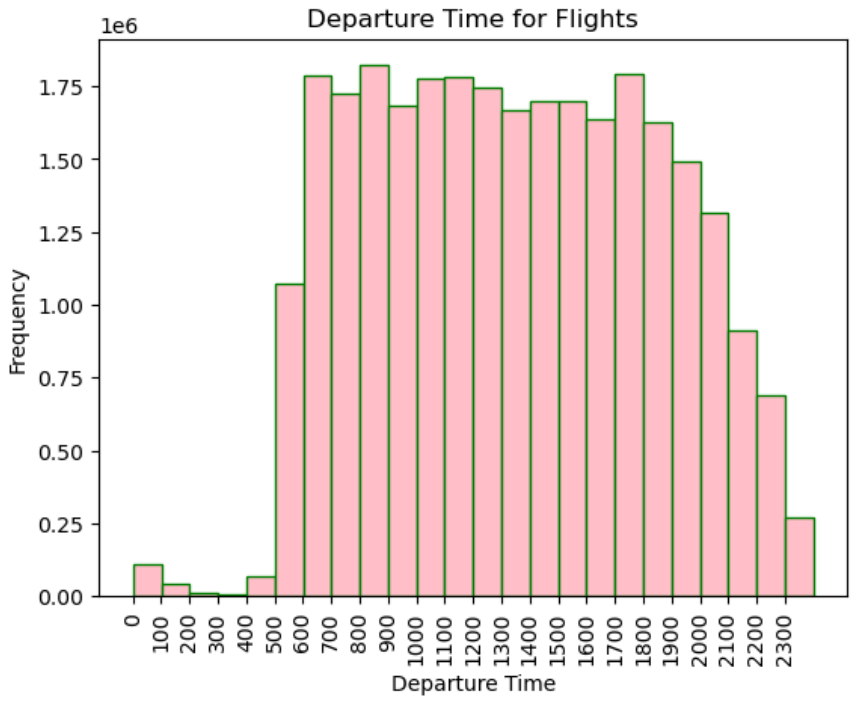
d. Cancellations Across the Week

Monday, Sunday, and Wednesday show higher cancellation rates compared to other days. This means that there is a higher likelihood of flights being canceled during these days, likely because they are the busiest days for air travel. The increased volume of flights on these days might lead to more problems in operations leading to a higher frequency of cancellations.



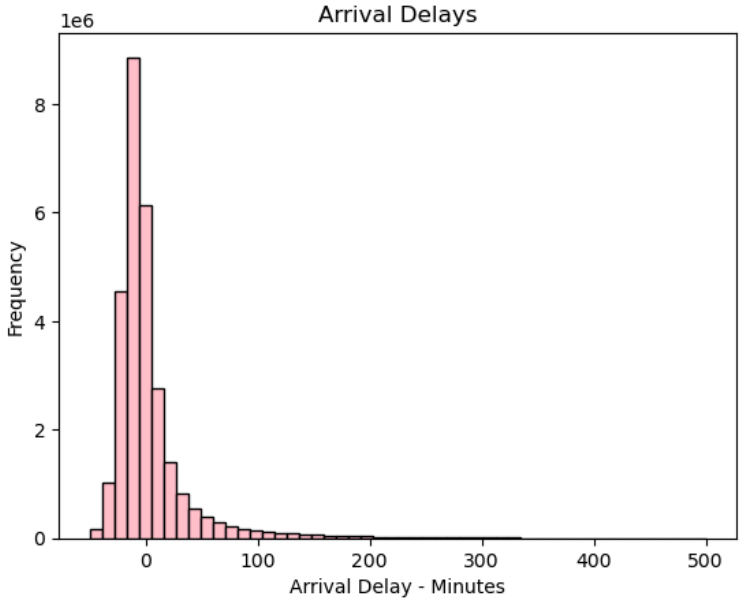
e. Distribution of Flight Times

The average departure time of around 13:30 PM shows an early afternoon peak in flight departures. Additionally, the median departure time also aligns with the average, showing a balanced distribution of flights before and after noon.



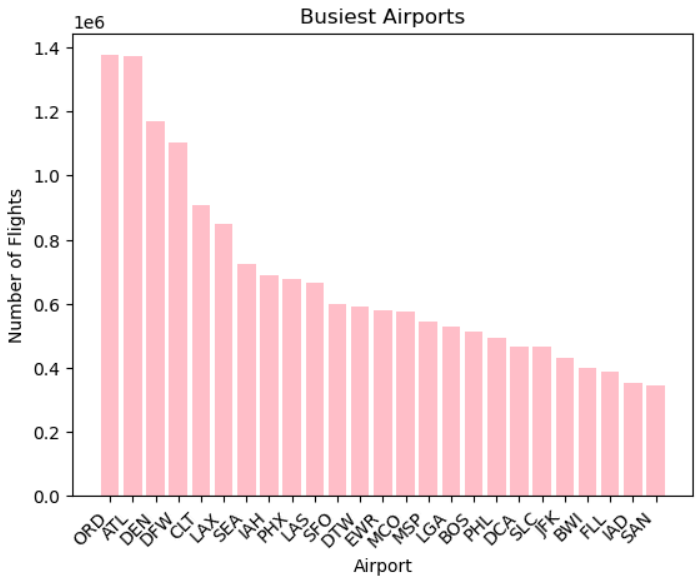
f. Distribution of Arrival Delays

The distribution of arrival delays shows that the majority of flights arrive on time, or with delays typically less than 20 minutes. This pattern indicates that while delays occur, they are generally small and do not significantly impact flights.



g. Busiest Airports

ORD and ATL are shown as the busiest origin airports. As major hubs, they tend to have a higher volume of flight. The high volume of flights from these hubs can lead to a higher likelihood of delays, because of congestion and operational issues that usually happens to busy airports.



4. Summary

The analysis presented an overview of the data for flights during 2018-2022, giving insights into many aspects of flight operations and performance. To start, it was identified a stability in on time performance across years, with an uncommon increase in the year of 2020, reflecting the impact of external factors such as the COVID-19 pandemic on air travel. Also, the seasonal variations in flight delays and cancellations highlight different patterns throughout the year, with some months showing higher rates of disruptions, likely due to factors such as weather and peak travel seasons. Additionally, the busiest airports and airlines were identified.

Challenges in feature engineering might come from handling missing data and special cases. Handling missing data needs to be done carefully, especially in columns essential for analysis such as arrival times, delays times, and cancellations. Because the pandemic was an event and it might have impacted a lot of air travelling, numbers in delay or cancellation data may need a more careful approach because they might reflect special and extraordinary circumstances. In addition, making sure that a proper encoding of categorical variables like airlines will be important.

-> Milestone 4

1. Overview

For this milestone, PySpark was used to read and process the flight delay data. The process consisted of various stages including reading the source data, normalizing it, performing feature engineering, splitting it into training and testing sets, modeling, validation, and evaluation of predictive models. Data with engineered features was saved to the /trusted folder and the trained models to the /models folder in GCS. Table with column name, data type, and feature engineering treatment is included here.

Features:

Column	Data Type	Variable Type	Indexer	Encoder	Scaler
FlightDate	Timestamp	Date			
Airline	String	Categorical	StringIndexer	OneHotEncoder	
Origin	String	Categorical	StringIndexer	OneHotEncoder	
CRSDepTime	Long	Continuous			MinMax
DepTime	Double	Continuous			MinMax
DepDelay	Double	Continuous			MinMax
ArrTime	Double	Continuous			MinMax
ArrDelay	Double	Continuous			MinMax
Year	Long	Continuous			MinMax
Quarter	Long	Continuous			MinMax
Month	Long	Continuous			MinMax
DayofMonth	Long	Continuous			MinMax
DayOfWeek	Long	Continuous			MinMax
CRSArrTime	Long	Continuous			MinMax
ArrDelayMinutes	Double	Continuous			MinMax

Label:

DepDelayMinutes - Double - Continuous - 1.0 if delay is > 0 minute, otherwise 0.0

After the feature engineering, random columns were selected and split the data into training and test sets. The feature engineering step involved creating a VectorAssembler to combine features, such as departure delay minutes, then, two models were created and a pipeline was constructed to go through feature transformation and model training. Challenges in feature engineering might come with model training where it might be difficult to handle large datasets efficiently to avoid performance limitations.

2. Results

a. Delay Status

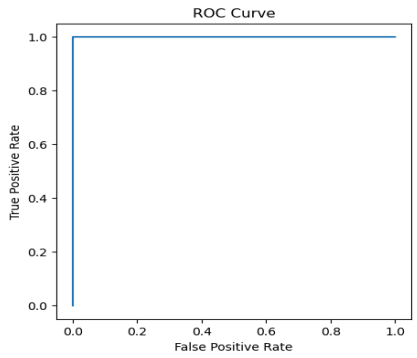
We have a binary classification model where the goal is to predict whether a flight will experience a departure delay (label=1.0) or not (label=0.0). The dataset contains various features related to flights, such as airline, origin airport, scheduled departure and arrival times, and actual departure and arrival times.

The model first predicts on test data and calculates performance metrics, then later, multiple models are trained with various hyperparameters. After that, the best performing model is selected, and its performance metrics are evaluated based on test data.

In the first part, the confusion matrix indicates that the model performed well, with a high number of true positives (109,632) and true negatives (231,119), and relatively low false positives (24) and false negatives (6). Accuracy, precision, and recall are 98% or higher, and F1 is really close to 1.

In the second part where multiple models were trained, the confusion matrix had a moderate to low performance compared to the first part. Despite an AUC of 1.0, the model's performance metrics are not the best, with a high number of false positives (329,319) and false negatives (5,107) compared to true positives (5,107) and true negatives (686,151). The accuracy is around 69%, indicating that there were a significant number of misclassifications. Although precision is 100%, recall is below 2%, meaning that the model failed to effectively identify delayed flights. Overall, the model effectively distinguishes between delayed and non delayed flights, but it struggles with accurately identifying delayed flights. This might be because there is a great difference between the two types of flight, or any error in the feature selection.

AUC: 1.0
|label| 0.0| 1.0|
| 0.0|686151| 0|
| 1.0|329319|5107|
Accuracy, Precision, Recall, F1 Score
(0.6873207705053121, 1.0, 0.015270941852607153, 0.03008249566316087)



As mentioned before, precision is 1.0, meaning that the model is predicting the positives correctly, which also signs with the perfect ROC curve. However, the recall is under 2%, so while the model is very accurate when it predicts delayed flights, it is missing a significant portion of those still.

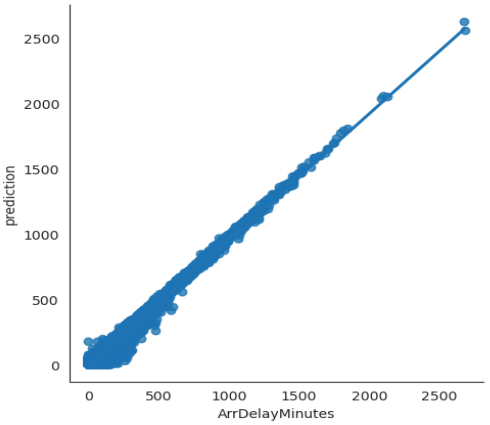
*Feature weights:
(Not full output)
Coefficient 448: ArrDelay 0.0028658999623831674
As the arrival delay minutes increase, the likelihood of the flight being classified as delayed also increases by approximately 0.003.

b. Delay Amount

RMSE: On average, the model's predictions for arrival delays are off by approximately 9.34 minutes. Because flight delays can cause significant operational and passenger impact, a low RMSE is what we want for accurate planning and resource allocation.

R-squared: Approximately 96% of the variability in arrival delays can be explained by departure delays. This value indicates that the linear regression model effectively captures the relationship between departure and arrival delays.

Actual vs. Predicted



The graph shows a positive linear relationship between arrival delay minutes and the prediction values. As the arrival delay minutes increase, the prediction value also increases. The data points are scattered closely around the linear regression line, indicating a good fit of the linear model to the data.

*Feature weights:
bestModel coefficients [0.9748168137449849]
bestModel intercept 0.40876317997420963
For every minute increase in departure delay, the model predicts an increase of approximately 0.98 minutes in the arrival delay, and if the departure delay is 0, there would still be an estimated delay of approximately 0.41 minutes.

3. Summary

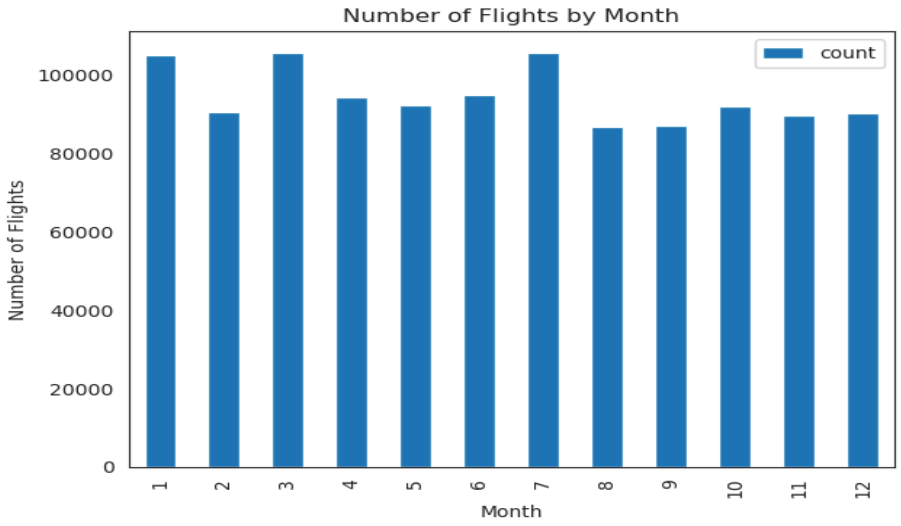
The first model initially performed well in distinguishing departure delays, however when it went through multiple training, there was a high number of false positives and false negatives. The binary classification model focuses on the prediction of departure delays, enabling stakeholders to identify flights at risk of delays. This allows for proactive measures to be taken to minimize the impact of delays on both operational performance and customer experience. With the high false positives and negatives resource allocation for flights would not be accurate, impacting operational efficiency and passenger satisfaction. The second model effectively forecasts arrival delay minutes based on departure delay minutes. This linear regression model aims to forecast arrival delay minutes based on departure delay minutes, providing airlines and airports with insights for proactive planning and resource allocation. By accurately predicting delays, airlines can optimize schedules, and improve passenger satisfaction. While the first model struggled with high false positives and negatives, the second model seemed to perform well in forecasting arrival delay minutes based on departure delay minutes.

-> Milestone 5

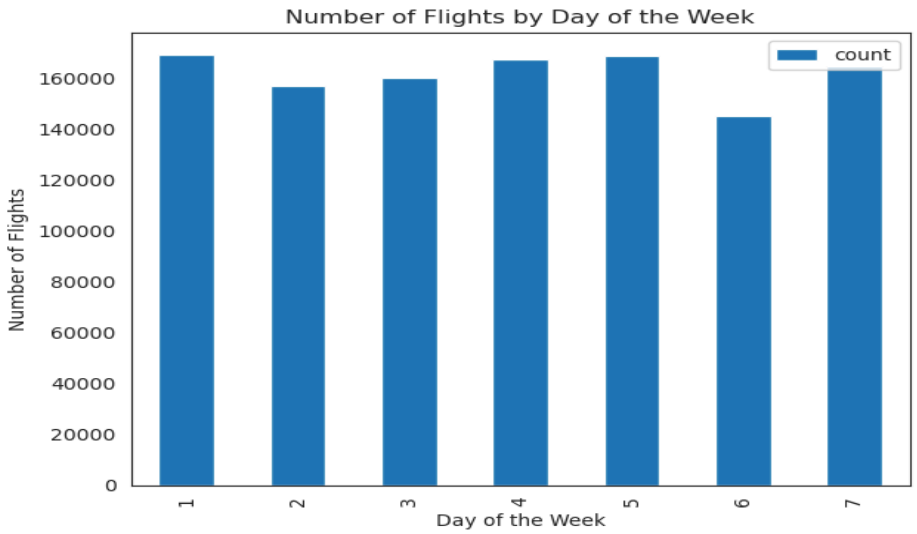
1. Visualizations

For this milestone, some visualizations were created:

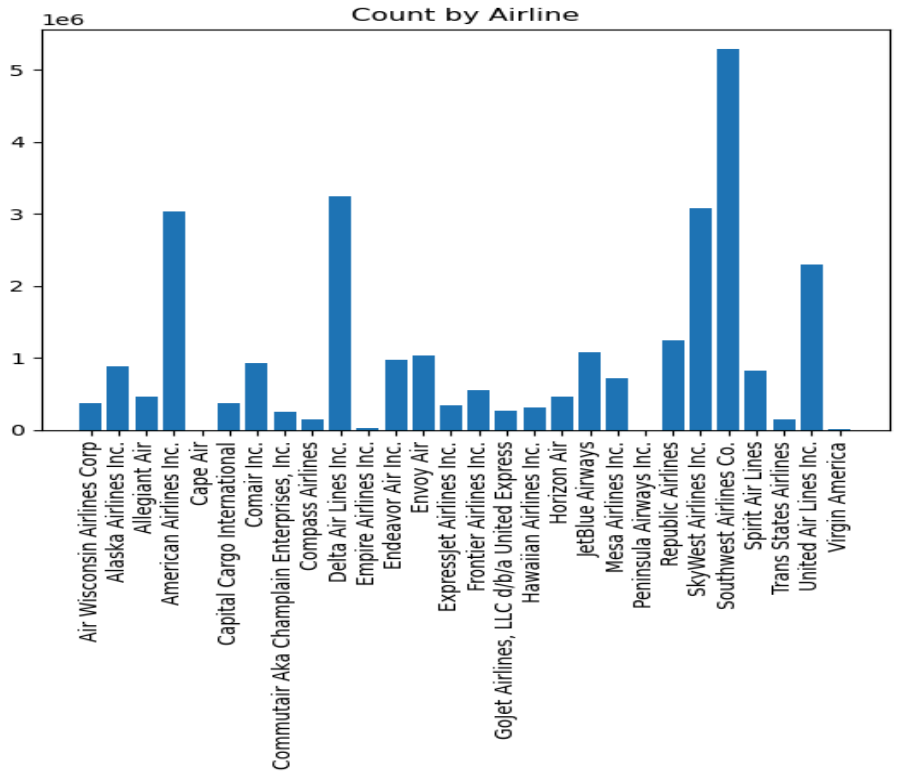
- Flights by Month: January, March, and July often see increased flight activity due to seasonal factors like holidays and breaks. January peak in travel might be due to New Year's vacations, March due to spring break, and July due to summer holidays.



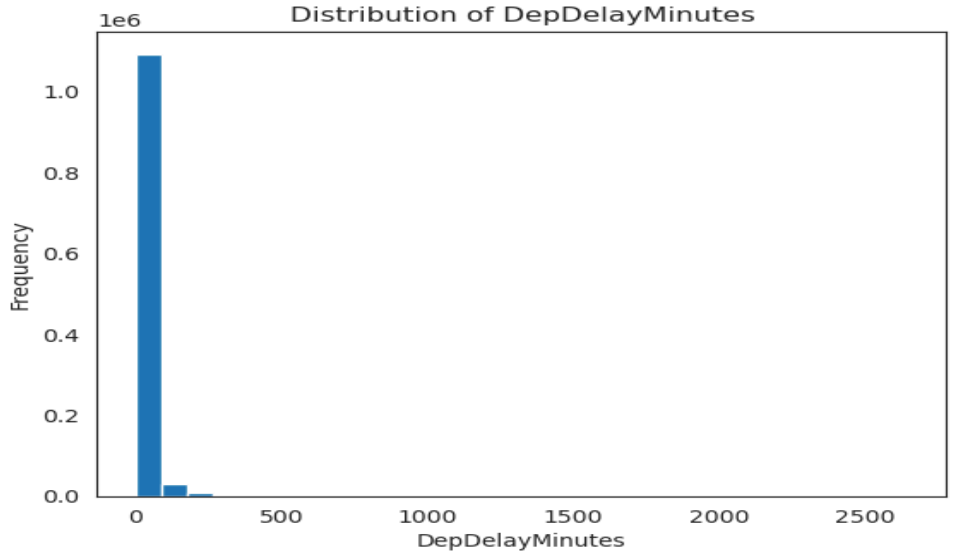
- Flight by Day of the Week: Weekdays show a higher volume of air travel. Specifically, Monday, Thursday, and Friday coming as peak days with the highest number of flights, this could be due to business schedules, where travel is common at the beginning and end of the workweek. Because people would go to work on Mondays, and go back to their home toward the end of the week.



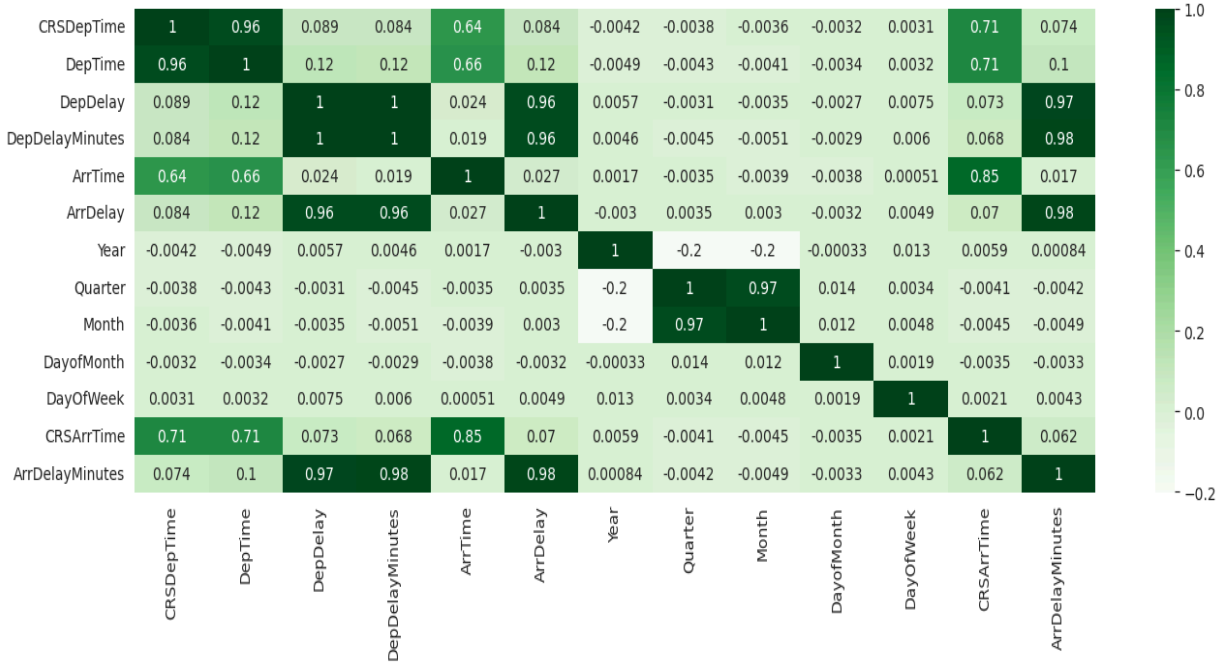
- Flight by Airline: Southwest Airlines operates the highest number of flights, while Delta Air Lines, SkyWest Airlines, and American Airlines operate a similar number of flights. This seems consistent because Southwest has an extensive route and is a well known company, thus contributing to more flight schedules.



- Departure Delays: Frequency for delays is extremely low, showing that flights usually depart on time, with 0 delay minutes. Most observations cluster around low or no delay, with some cases of long delays. Overall, most flights operated fairly close to their scheduled departure times.



- Correlation: There are strong positive correlations between DepDelay and DepDelayMinutes, and ArrDelay and ArrDelayMinutes. This is expected as these variables are representing the same information in different formats. There are strong positive correlations between DepDelay and DepDelayMinutes, and ArrDelay and ArrDelayMinutes, suggesting that departure delays are associated with arrival delays. This is also expected, if a flight departures later, it will probably arrive later. Also, there are moderate negative correlations between Quarter and DepDelay, ArrDelay, suggesting delays might vary by quarter or season.



2. Summary

Air travel is a complex system influenced by various factors, and analyzing flight data can provide many insights. One observation made with the current data is the distinct seasonal patterns in flight activity, with peaks observed during January, March, and July, likely due to holidays and school breaks. The busiest travel days tend to be weekdays, particularly Mondays, Thursdays, and Fridays, reflecting the schedules of business travelers. Seasonal observation is supported by the correlation analysis, showing moderate negative correlations between Quarter and delay. Among airlines, Southwest Airlines stands out as operating the highest number of flights, aligning with its extensive route network and strong market presence. Overall, this analysis of flight data shows seasonal trends, airline operations, delay distributions, and variable interrelationships. These insights not only deepen our understanding of air travel but also inform strategies for airlines and airports to enhance operational efficiency, customer satisfaction, and contingency planning in the aviation industry.

-> Milestone 6 - URL for project on Github: https://github.com/hmeina/big-data-flights

In conclusion, this project analyzed flight data from 2018 to 2022 to gain insights into the operational dynamics of the aviation industry. Throughout the project, I explored big data processing using PySpark and developed skills in creating machine learning pipelines.

The pipeline process started with data preparation, where flight data was loaded and processed using PySpark. Through a comprehensive exploratory data analysis, I was able to get interesting insights into the dataset's characteristics, for feature engineering work later on. Feature engineering was another important step taken, involving normalization tasks and the creation of engineered features such as departure delay minutes. Vector Assembler was used for smooth feature combination, and a binary label to predict departure delays was also created. In the next step, a binary classification model to predict departure delays and a linear regression model to forecast arrival delay minutes based on departure delays were created. Using PySpark's machine learning libraries, comprehensive model training and evaluation was carried out, and I examined each model performance through various metrics including confusion matrices, accuracy, precision, recall, F1 score, RMSE, and R squared.

The visualizations gave valuable insights into flight trends across different aspects such as month, day of the week, airline, and departure delays. These visuals not only highlighted important trends but also helped to better understand how different features correlate and impact model performance.

Main insights from the project emphasizes the complex connection between departure and arrival delays within the aviation industry. While the binary classification model initially appeared effective at differentiating delayed and non delayed flights, there were some issues during the multiple training sections, leading to a decrease in accuracy.

On the other hand, the linear regression model seemed to work for forecasting arrival delay minutes, offering valuable insights for proactive planning and resource allocation. Strong positive correlations between departure and arrival delay emphasizes the crucial role of timely departures in ensuring punctual arrivals.

Air travel is complex and influenced by various factors, as shown by the distinct seasonal patterns seen in flight activity. Peaks during January, March, and July were likely due to holidays and school breaks. Additionally, weekday observations revealed Mondays, Thursdays, and Fridays as peak travel days, aligning with business travelers' schedules. Correlation analyses further supported seasonal observations, highlighting moderate negative correlations between Quarter and delay. Among airlines, Southwest Airlines stood out for operating the highest number of flights, indicative of its extensive route network and market presence. Overall, this analysis of flight data provided insights into seasonal trends, airline operations, delay distributions, and interrelationships. These insights deepens the understanding of air travel, and can be used to create strategies for enhancing operational efficiency, customer satisfaction, and contingency planning within the aviation industry.