Analysis on Airline delays and cost predictions: Atlanta airport

Table of Contents:

I. Abstract

"FlightFlow Dynamics" our company specializes in managing and optimizing concessions at American airports aims to leverage machine learning to build a predictive model that forecasts flight delays and early arrivals at U.S. airports. We were hired by ***Atlanta Hartsfield-Jackson International Airport*** (ATL) to analyze their incoming flights and build a model to predict delays.

By analyzing historical data for the year 2021, a period we deem crucial for the sector. This year signifies the peak following the COVID-19 recovery in 2020 and reflects a prosperous phase for the aviation industry. This time period also reflects the disruptions in aviation operation systems as a result of the pandemic. As such, it can provide a robust foundation to start for future changes. This project aims to deliver a transformative solution that optimizes concession performance, and increases airport profitability as a response to the following problem statement:

How can we leverage machine learning to accurately predict flight delays, early arrivals, and in future cancellations in U.S. airports, while optimizing airport operations and creating new revenue streams for airports, without compromising the efficiency and satisfaction of stakeholders?

From our analysis, we can leverage our results to improve ATL under various domains which, all of which ultimately impact costs. Operationally, using our model, ATL can prevent delay propagation by better anticipating disruptions to normal traffic and can then correctly allocate gates, staff, and equipment.

A well established nexus exists between flights and inclement weather conditions. However, this model seeks to investigate the features solely related to the flight itself that can indicate a delay. This base understanding will provide any airport with more tailored, concrete and reliable estimates of costs based on factors less variable than weather. In future iterations of our models, we aim to expand our datasets to include external factors, which would enable airports to adjust their planning in response to fluctuating circumstances.

II. Introduction

Air travel is a critical component of modern transportation, with millions of passengers relying on efficient airport operations daily. Within the last century, the aviation industry has seen remarkable growth which has resulted in substantial economic importance. Whilst aviation creates a high value for customers and other stakeholders, the profit margins are typically low due to high fixed costs and its dependence on external factors. The aviation system is surrounded by different environments: the economic, ecological, social, technological and political environments. Each environment exerts influence on the aviation system and is simultaneously affected by it. Thus, there are numerous factors that can cause an array of disruptions. Disruptions and a low level of uncertainty are an established notion when considering air travel. Thus, it poses a large problem to its primary stakeholders. As such a critical component of our modern transportation system, with millions of passengers relying on efficient operations daily, there is significant data that can be leveraged to mitigate this issue. Thus, the conception of our company: "FlightFlow Dynamics".

In the United States, where air traffic is among the busiest in the world, accurately predicting flight delays, early arrivals, and cancellations is a significant challenge with substantial implications for operational efficiency and customer satisfaction.

Our company specializes in managing and optimizing operations at American airports, a sector that plays a vital role in enhancing passenger experience and generating significant revenue. Efficient airport management depends heavily on the ability to anticipate and adapt to operational disruptions such as flight delays, early arrivals, and cancellations. These propagating disruptions affect not only passengers and airlines but also the concession operations within airports, from retail shops and dining outlets to premium services.

Our mission is to leverage data using machine learning to build predictive models that forecasts flight delays and early arrivals at U.S. airports. This system can be employed by our clients (Airports) to improve resource allocation, enhance passenger experience, reduce disruptions, and increase profits.

By aligning advanced machine learning with our expertise in aviation operations, this project aims to deliver a transformative technical solution. This innovation positions us as a forward-thinking leader in managing and monetizing the dynamic environment of American airports. This predictive system, which is integrated with market data and critical literature will offer our clients new opportunities for monetization. By integrating our insights into the airport's ecosystem, our clients can:

- Reduce delay propagation
- Create new revenue streams
- Reduce energy costs
- Correctly allocate personnel and equipment
- Improve gate profitability
- Increase concessions revenue through alignment with pricing models and stock management
- Improve customer satisfaction and experience

III. Literature Review

Airports are critical to the economic and infrastructural framework of the United States, functioning as pivotal hubs for passenger and freight transportation domestically and internationally. The U.S. airport system, encompassing approximately 519 commercial airports and thousands of regional facilities, is among the most expansive and intricate globally. Despite the severe impact of the COVID-19 pandemic, the industry demonstrated resilience; U.S. domestic air traffic reached approximately 674 million passengers in 2021, marking a recovery from the 2020 but remaining below the pre-pandemic peak of 925 million in 2019 (FAA, 2021).

At the core of this system are airports such as Hartsfield-Jackson Atlanta International Airport (ATL), which has been the world's busiest airport by passenger volume for several years. In 2021, ATL accommodated 75.7 million passengers—a notable increase from 2020 yet substantially below its 2019 peak of 110 million passengers (ACI, 2021; FAA, 2021). Such traffic underscores the operational complexity of U.S. airports, where challenges like flight delays and cancellations persist. Domestic flight cancellations, for instance, improved to 1.76% in 2021 compared to 6% in 2020, yet delays remain a pervasive issue (U.S. Department of Transportation, 2021).

The significance of the American airport network is multifaceted, stemming from its scale, economic influence, and adoption of innovative technologies. With over 5,000 public airports ranging from large international hubs to smaller regional facilities, the Federal Aviation Administration (FAA, 2021) categorizes these airports based on their size and significance. Major hubs such as Atlanta, Los Angeles,

and Chicago serve millions of passengers annually, forming the backbone of both domestic and global travel.

Hartsfield-Jackson Atlanta International Airport serves as a representative case study of the challenges and innovations inherent in managing high-traffic airports. Henriques and Feiteira (2018) identified weather conditions and air traffic congestion as the primary causes of delays at ATL, with frequent summer thunderstorms in Georgia causing delay increases of 5–10%. Their study demonstrated that predictive models could mitigate delays, achieving a 10% annual reduction—a meaningful improvement for an airport with such substantial operational demands.

Further emphasizing ground-level inefficiencies, Neyshabouri and Sherry (2014) analyzed surface operations at ATL. They reported an average ground time of 30 minutes for aircraft, attributing delays to congestion at gates and taxiways. By optimizing ground routing, their model suggested potential reductions of 5–10% in average ground times, offering a feasible solution for improving efficiency.

Yablonsky et al. (2014) provided a broader perspective on ATL's delay dynamics, noting that 60% of flights encounter delays. Despite these figures, the researchers highlighted ATL's effective use of advanced technologies, such as air traffic management and flight surveillance systems, which reduced the average delay to 12 minutes per flight. This technological integration showcases ATL's ability to maintain efficiency despite high operational pressures.

Collectively, these studies illustrate the duality of Hartsfield-Jackson Atlanta International Airport as both a nexus of challenges and a model of innovation. The airport encapsulates the broader realities of the U.S. aviation sector, where infrastructural demands are met with technological advancements and data-driven strategies to optimize operations. These findings underscore the critical importance of ongoing research and innovation in maintaining and enhancing the functionality of major transportation hubs.

IV. Methodology & Results:

We have created a new feature - efficiency ratio, taking into consideration that all flights have different distances and actual flight time. We gained substantial insights into the U.S. aviation industry, carefully identified outliers separately for each airline, and decided to move to the modelling part.

A. Introduction to data

The dataset we utilized consists of 6,311,871 records, providing a comprehensive overview of flight operations across the United States in 2021. This extensive dataset includes 61 columns, capturing a diverse range of flight-related details such as schedules, delays, cancellations, and operational characteristics. The richness of the dataset provides an invaluable resource for analyzing flight performance, identifying factors that influence delays and cancellations, and informing strategies to optimize airport operations. After careful evaluation, we identified 47 features as most relevant to our business objectives, focusing on those with the greatest potential to provide actionable insights.

Description of data:

Index	Feature Name	Data Type	Measurement Units
1	Airline	Nominal	Categorical

2	Origin	Nominal	Categorical
3	Dest	Nominal	Categorical
4	Cancelled	boolean	Boolean
5	Diverted	boolean	Boolean
6	CRSDepTime	Continuous	Coded hours
7	DepTime	Continuous	Hours
8	DepDelayMinutes	Continuous	Minutes
9	DepDelay	Continuous	Minutes
10	ArrTime	Continuous	Hours (24-hour format)
11	ArrDelayMinutes	Continuous	Minutes
12	AirTime	Continuous	Minutes
13	Distance	Continuous	Miles
14	Quarter	Ordinal	Integer
15	Month	Ordinal	Integer
16	DayofMonth	Ordinal	Integer
17	DayOfWeek	Ordinal	Integer
18	Marketing_Airline_Network	Nominal	Categorical
19	Operating_Airline	Nominal	Categorical
20	IATA_Code_Operating_Airline	Nominal	Categorical
21	Tail_Number	Nominal	Categorical
22	OriginAirportID	Nominal	Integer
23	OriginAirportSeqID	Nominal	Integer
24	OriginCityMarketID	Nominal	Integer
25	OriginCityName	Nominal	Categorical
26	OriginStateName	Nominal	Categorical
27	OriginWac	Ordinal	Integer
28	DestAirportID	Nominal	Integer
29	DestAirportSeqID	Nominal	Integer
30	DestCityMarketID	Nominal	Integer

31	DestCityName	Nominal	Categorical
32	DestStateName	Nominal	Categorical
33	DestWac	Ordinal	Integer
34	DepDel15	boolean	Boolean
35	DepartureDelayGroups	Ordinal	Categorized Minutes
36	DepTimeBlk	Nominal	Time Block
37	TaxiOut	Continuous	Minutes
38	WheelsOff	Continuous	Hours
39	WheelsOn	Continuous	Hours
40	TaxiIn	Continuous	Minutes
41	CRSArrTime	Continuous	Coded hours
42	ArrDelay	Continuous	Minutes
43	ArrDel15	boolean	Boolean
44	ArrivalDelayGroups	Ordinal	Categorized Minutes
45	ArrTimeBlk	Nominal	Time Block
46	DistanceGroup	Ordinal	Distance Category
47	DivAirportLandings	Continuous	Integer

Summary statistics:

	CRSDepTime	DepTime	DepDelayMinutes	DepDelay	ArrTime	ArrDelayMinutes	AirTime	Distance	Month	DayofMonth	DayOfWeek
count	6.31E+06	6.20E+06	6.20E+06	6.20E+06	6.20E+06	6.19E+06	6.19E+06	6.31E+06	6.31E+06	6.31E+06	6.31E+06
mean	1.32E+03	1.33E+03	1.28E+01	9.47E+00	1.48E+03	1.25E+01	1.11E+02	7.96E+02	6.97E+00	1.58E+01	4.01E+00
std	4.74E+02	4.87E+02	4.74E+01	4.84E+01	5.14E+02	4.67E+01	6.89E+01	5.83E+02	3.30E+00	8.79E+00	2.01E+00
min	1.00E+00	1.00E+00	0.00E+00	-1.05E+02	1.00E+00	0.00E+00	8.00E+00	3.10E+01	1.00E+00	1.00E+00	1.00E+00
20%	8.33E+02	8.34E+02	0.00E+00	-6.00E+00	1.01E+03	0.00E+00	5.40E+01	3.21E+02	4.00E+00	7.00E+00	2.00E+00
40%	1.14E+03	1.14E+03	0.00E+00	-4.00E+00	1.33E+03	0.00E+00	8.00E+01	5.41E+02	6.00E+00	1.30E+01	3.00E+00
50%	1.32E+03	1.32E+03	0.00E+00	-2.00E+00	1.51E+03	0.00E+00	9.40E+01	6.46E+02	7.00E+00	1.60E+01	4.00E+00
60%	1.45E+03	1.46E+03	0.00E+00	-1.00E+00	1.64E+03	0.00E+00	1.12E+02	8.04E+02	8.00E+00	1.90E+01	5.00E+00
80%	1.81E+03	1.82E+03	1.10E+01	1.10E+01	1.96E+03	1.10E+01	1.53E+02	1.12E+03	1.00E+01	2.50E+01	6.00E+00
max	2.36E+03	2.40E+03	3.10E+03	3.10E+03	2.40E+03	3.09E+03	7.11E+02	5.81E+03	1.20E+01	3.10E+01	7.00E+00

DOT_ID_Operating_Airline	OriginAirportID	OriginAirportSeqID	OriginCityMarketID	OriginWac	DestAirportID	DestAirportSeqID	DestCityMarketID	DestWac	DepDel15	DepartureDelayGroups	TaxiOut
6.31E+06	6.31E+06	6.31E+06	6.31E+06	6.31E+06	6.31E+06	6.31E+06	6.31E+06	6.31E+06	6.20E+06	6.20E+06	6.20E+06
2.00E+04	1.27E+04	1.27E+06	3.17E+04	5.50E+01	1.27E+04	1.27E+06	3.17E+04	5.50E+01	1.73E-01	-1.55E-02	1.62E+01
3.77E+02	1.53E+03	1.53E+05	1.35E+03	2.62E+01	1.53E+03	1.53E+05	1.35E+03	2.62E+01	3.78E-01	2.14E+00	8.58E+00
1.94E+04	1.01E+04	1.01E+06	3.01E+04	1.00E+00	1.01E+04	1.01E+06	3.01E+04	1.00E+00	0.00E+00	-2.00E+00	1.00E+00
1.98E+04	1.11E+04	1.11E+06	3.05E+04	3.30E+01	1.11E+04	1.11E+06	3.05E+04	3.30E+01	0.00E+00	-1.00E+00	1.00E+01
1.98E+04	1.20E+04	1.20E+06	3.10E+04	4.10E+01	1.20E+04	1.20E+06	3.10E+04	4.10E+01	0.00E+00	-1.00E+00	1.30E+01
2.00E+04	1.29E+04	1.29E+06	3.15E+04	4.50E+01	1.29E+04	1.29E+06	3.15E+04	4.50E+01	0.00E+00	-1.00E+00	1.40E+01
2.03E+04	1.32E+04	1.32E+06	3.17E+04	7.20E+01	1.32E+04	1.32E+06	3.17E+04	7.20E+01	0.00E+00	-1.00E+00	1.60E+01
2.04E+04	1.41E+04	1.41E+06	3.30E+04	8.30E+01	1.41E+04	1.41E+06	3.30E+04	8.30E+01	0.00E+00	0.00E+00	2.00E+01
2.05E+04	1.69E+04	1.69E+06	3.61E+04	9.30E+01	1.69E+04	1.69E+06	3.61E+04	9.30E+01	1.00E+00	1.20E+01	2.56E+02

WheelsOff	WheelsOn	TaxiIn	CRSArrTime	ArrDelay	ArrDel15	ArrivalDelayGroups	DistanceGroup	DivAirportLandings
6.20E+06	6.20E+06	6.20E+06	6.31E+06	6.19E+06	6.19E+06	6.19E+06	6.31E+06	6.31E+06
1.35E+03	1.47E+03	7.68E+00	1.50E+03	3.29E+00	1.73E-01	-3.16E-01	3.65E+00	3.34E-03
4.88E+02	5.10E+02	6.38E+00	4.95E+02	5.01E+01	3.78E-01	2.27E+00	2.28E+00	1.05E-01
1.00E+00	1.00E+00	1.00E+00	1.00E+00	-1.07E+02	0.00E+00	-2.00E+00	1.00E+00	0.00E+00
8.50E+02	1.01E+03	4.00E+00	1.03E+03	-1.80E+01	0.00E+00	-2.00E+00	2.00E+00	0.00E+00
1.16E+03	1.33E+03	5.00E+00	1.34E+03	-1.10E+01	0.00E+00	-1.00E+00	3.00E+00	0.00E+00
1.33E+03	1.51E+03	6.00E+00	1.52E+03	-7.00E+00	0.00E+00	-1.00E+00	3.00E+00	0.00E+00
1.51E+03	1.64E+03	7.00E+00	1.65E+03	-3.00E+00	0.00E+00	-1.00E+00	4.00E+00	0.00E+00
1.83E+03	1.95E+03	1.00E+01	2.00E+03	1.10E+01	0.00E+00	0.00E+00	5.00E+00	0.00E+00
2.40E+03	2.40E+03	2.51E+02	2.40E+03	3.09E+03	1.00E+00	1.20E+01	1.10E+01	9.00E+00

Figure 1. Descriptive statistics with percentiles of all columns.

Based on the descriptive statistics we can see the need to investigate *AirTime* further as it has a very large standard deviation. This highlights that there are very long flights because 80% of the data fall below 1.53 e+02 and the max is 7.11e+02.

B. Columns dropped:

The decision to remove these specific columns is driven by their limited contribution to predicting delays, cancellations, or early arrivals of flights, or by the redundancy of the information they provide. Below, we outline the rationale for excluding each column:

1. **Year and FlightDate:** These columns primarily provide temporal context. Since all the data pertains to flights in 2021, the Year column is redundant. Similarly, while *FlightDate* contains the specific date, it can be represented by more granular temporal columns such as *Month*, *DayOfMonth*, and *DayOfWeek*.
1. **Flight_Number_Marketing_Airline and Flight_Number_Operating_Airline:** These columns represent flight numbers for marketing and operating airlines. While they uniquely identify flights, they do not provide insights into the factors influencing delays or cancellations.
2. **Quarter:** As the dataset includes more granular temporal features like *Month*, *DayOfMonth*, and *DayOfWeek*, the broader grouping by quarter becomes unnecessary for predictive modeling.
3. **Operated_or_Branded_Code_Share_Partners:** This column relates to airline marketing partnerships, which are not expected to influence the timing of departures or arrivals.
4. **DOT_ID_Marketing_Airline and IATA_Code_Marketing_Airline:** These are unique identifiers for airlines, which are already effectively represented by the Airline column, making them redundant.
5. **OriginState, DestState, OriginStateFips, and DestStateFips:** These columns provide location-based state-level information for airports but are less detailed than city or airport-specific columns like *OriginCityName* and *DestCityName*. Including both levels of geographic information could introduce multicollinearity, so the state-level data is excluded.
6. **CRSElapsedTime and ActualElapsedTime:** These columns represent the scheduled and actual flight durations, respectively. Since we are already considering columns like *AirTime* and delay-related metrics, these additional duration measures may be redundant or correlated.

By dropping these columns, we aim to reduce the dimensionality of the dataset, eliminate redundancy, and enhance model interpretability without sacrificing critical predictive information. This step is part of our preprocessing pipeline to focus our analysis on features that are directly related to the problem at hand.

C. Exploratory data analysis

We created a class containing exhaustive univariate and bivariate analysis for better replicability for our future clients. Before analyzing the data we first separated our variables into quantitative and categorical variables. Many of the categorical variables had an exceptionally high number of modalities, which posed challenges for visual representation and interpretability. To address this, we grouped categorical variables into categories of low, medium, and high modality counts to facilitate analysis. Variables with more than 50 modalities were either visualized using specialized graphs (for example: top 20 categories) to ensure interpretability or excluded from graphical representation and analyzed using tabular summaries instead.

Univariate Analysis:

Categorical:

Our categorical univariate analysis was centered around understanding frequencies of each category within a particular variable. Here are a few that are important throughout our analysis:

Based on each flight, we can count the frequencies of each airline. We can see that among the 22 different airlines, SouthWest was almost 17% of the flights in 2021. We will see that later on, the amount of flights per airline will be seen in the delays. These top 10 airlines give us an idea about the market share of the aviation industry. Surprisingly, Delta Airlines and American Airlines, which are giants in the aviation industry, are in the 3rd and fourth position. This may mean that our dataset is not exhaustive given that it is only flights within the United States and for the year 2021.

The departure and arrival deploy groups are pre-made categorical variables with 15 categories.They range from early to extremely delayed. We can see that 60% of the occurrences fall within the first category, which is more than 15 minutes early. In the analysis of Delay groups, we will better understand this distribution.

Airline	Category	Count	Frequency (%)
1	Southwest Airlines Co.	1064640	16.867265
2	SkyWest Airlines Inc.	753343	11.935336
3	Delta Air Lines Inc.	747998	11.850654
4	American Airlines Inc.	736399	11.666889
5	United Air Lines Inc.	446837	7.079311
6	Republic Airlines	332926	5.274601
7	Endeavor Air Inc.	266867	4.228017
8	Envoy Air	255751	4.051905
9	Comair Inc.	222602	3.52672
10	JetBlue Airways	202702	3.211441

DepartureDelayGroups	Category	Count	Frequency (%)
0	-1	3806271	60.303371
1	0	1292454	20.476559
2	1	390307	6.183697
3	2	200675	3.179327
4	3	123663	1.959213
5	4	83020	1.315299
6	12	67207	1.064771
7	5	58583	0.92814
8	6	43627	0.69119
9	7	32995	0.522745
10	-2	30466	0.482678
11	8	25415	0.402654
12	9	19948	0.316039
13	10	16009	0.253633
14	11	12818	0.203078

DistanceGroup	Category	Count	Frequency (%)
1	2	1464741	23.20613
2	3	1260626	19.972303
3	4	1020264	16.164209
4	1	859827	13.622379
5	5	703754	11.149689
6	6	278915	4.418896
7	7	268253	4.249976
8	10	133222	2.110658
9	8	121447	1.924105
10	11	115036	1.822534
11	9	85786	1.359122

Figure 2. Example of summary table of values in categorical variables with count and frequencies of modalities.

These domestic flights cover a wide range of distances, resulting in varying flight times. The lower mileage groups correspond to shorter distances, while the larger categories represent longer flights.

Notably, the greatest frequency is observed within the low-to-mid distance range, highlighting a concentration of flights in this category.

Quantitative:
To better understand the distribution of our quantitative variables, we analyzed them using histograms and box plots.

From the histograms, we observe that most flights have a distance around 315 miles, which corresponds to the midpoint of the first bin. Additionally, there are very few flights in the largest distance bin, with this group representing nearly 0% of the total frequency.

Distance Histogram	Midpoint	Count	Frequency (%)
0	317.16	2966509.00	47.00
1	898.15	2185163.00	34.62
2	1476.25	714926.00	11.33
3	2054.35	235553.00	3.73
4	2632.45	192429.00	3.05
5	3210.55	6783.00	0.11
6	3788.65	5391.00	0.09
7	4366.75	2484.00	0.04
8	4944.85	2613.00	0.04
9	5522.95	20.00	0.00

Figure 3. Example of histogram of values in categorical variables with count and frequencies of modalities.

The variables *TaxiIn* and *TaxiOut* represent the amount of time an aircraft spends moving from landing to the gate and from the gate to takeoff. The boxplots for both variables show similarly left-skewed distributions. The middle 50% of the data is concentrated at the lower end, indicating that most observations have relatively short taxi times. However, the presence of numerous extreme outliers in both box plots suggests that disruptions, poor resource allocation, or other factors may be influencing these times. For example, *TaxiIn* times could be significantly prolonged due to factors such as unavailable gates or airport congestion.

The time variables *CSRDepTime*, *DepTime*, *ArrTime*, and *CSRArrTime* when displayed in a box plot show that they are essentially identical but represented in two different ways. The Common Reporting Standard (CRS), developed by the Organization for Economic Cooperation and Development (OECD), is a global standard for the automatic exchange of financial account information. Since the CRS times and the regularly reported times are the same, we will drop *DepTime* and *ArrTime* from further analysis.

Looking at the *Distance* boxplot, we can confirm our earlier assumption that there are very few long-haul flights. Additionally, since *AirTime* and *Distance* are highly correlated, we expect their distributions to be similar, which is confirmed by the boxplots. However, *Distance* is more left-skewed than *AirTime*, indicating a higher frequency of shorter flights.

From the *ArrTime* boxplot, we can confirm that red-eye flights (flights arriving late at night or early in the morning) make up a much smaller proportion compared to daytime flights. Most departures occur in the morning, which mirrors the distribution seen in the *ArrTime* boxplot for red-eye flights.

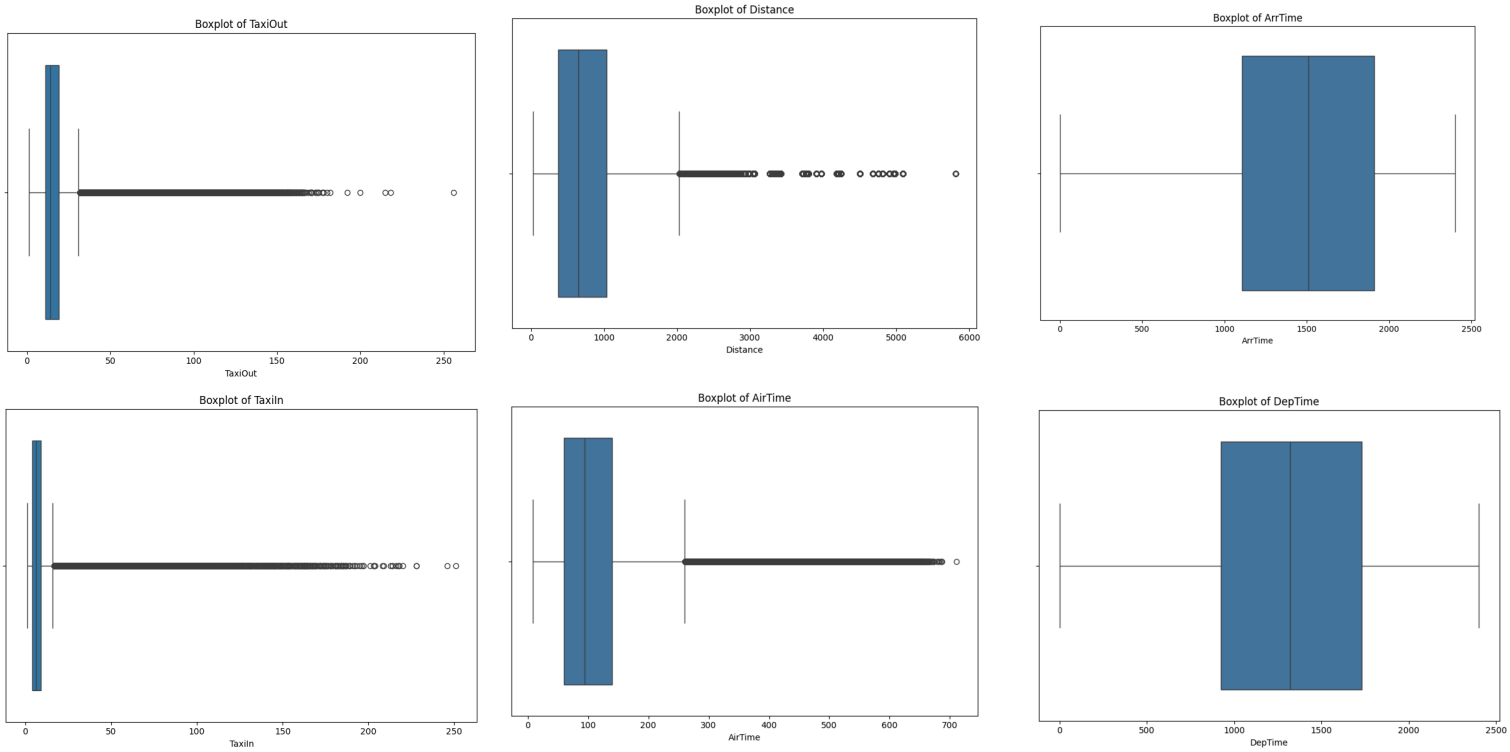


Figure 4. Example of boxplots for quantitative variables.

Bivariate Analysis:

Quantitative Bivariate Analysis:

We used the Pearson correlation coefficient to determine the strength of the relationship between two quantitative variables and plotted the subsequent correlation matrix.

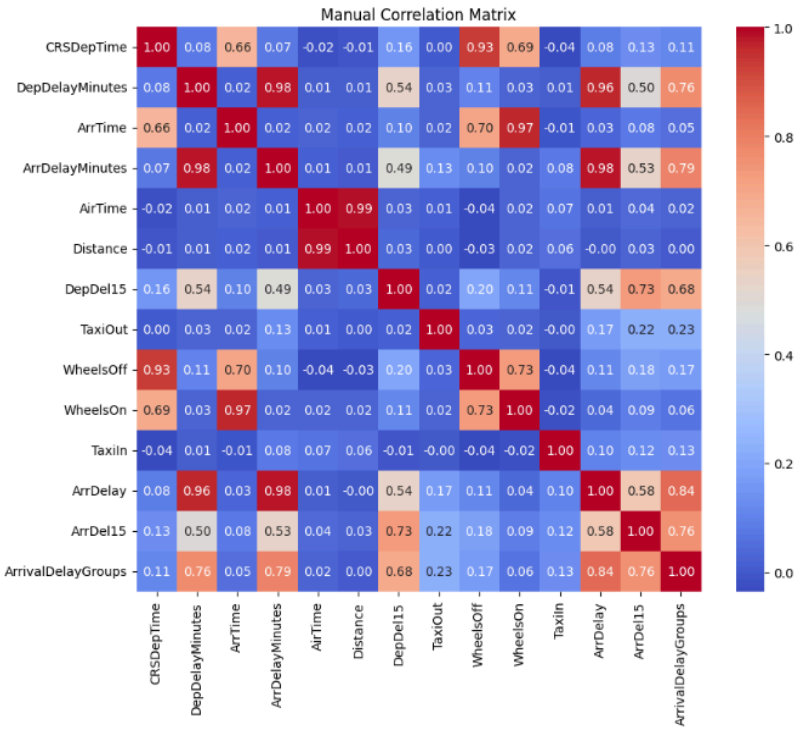


Figure 5.Correlation matrix for quantitative variables.

Based on the correlation matrix, we can see a high correlation between ArrDelayMinutes and DepDelayMinutes. DepDel15 and ArrDel15 are "flags" which means a delay was more or less than 15 minutes. This flagging effect can be seen in the correlation between DepDel15 and DepDelayMinutes. Distance and AirTime are very correlated. For analysis of efficiency we will keep it and remove it before modeling. The other highly correlated variables are removed at the end of the bivariate analysis section.

Categorical Bivariate Analysis:

To examine the relationship between two categorical variables, we created frequency tables and plotted stacked bar charts to visualize the distribution of categories within each variable.

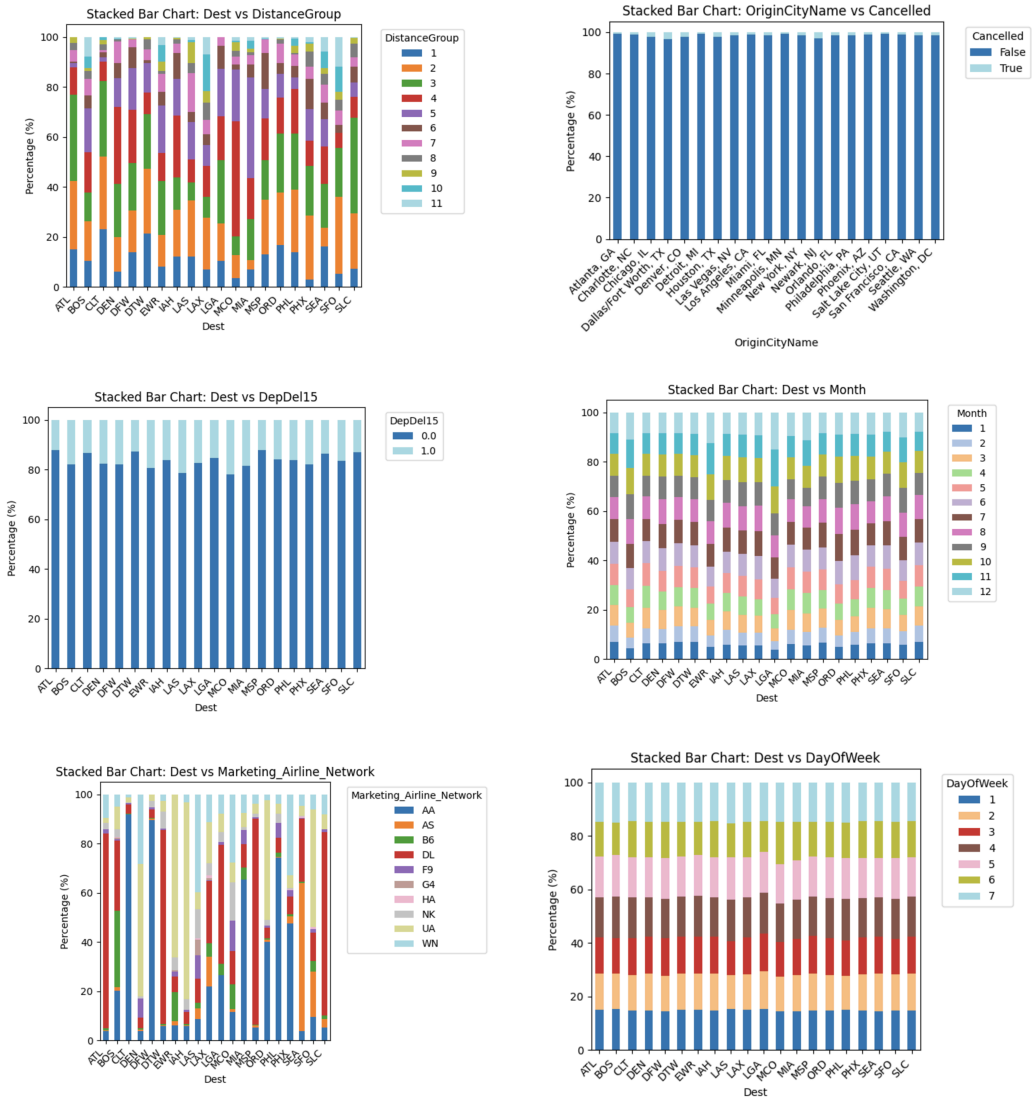


Figure 6. Examples of stacked bar charts for categorical variables.

Focusing on the airport in Atlanta Georgia, we observe that the distribution of flight numbers remains consistent throughout the year, with no significant seasonal variation. A similar proportional distribution is seen across different days of the week. Regarding flight delays, more than 60% of flights arrive slightly earlier than expected, while around 15% experience delays. In the case of *DepDel15* (departure delay of more than 15 minutes), approximately 85% of flights arriving at ATL are on time, with no delays exceeding 15 minutes.

We also observed that the percentage of canceled flights across all airports, including Atlanta, is very small. This is an indication to further investigate the characteristics of these flights to understand the factors contributing to their occurrence.

The stacked bar chart of Dest and Marketing_Airline_Network shows that the majority of flights departing from Atlanta are from the company Delta. We will keep this proportion in mind throughout our further analysis. Likewise, we will retain the proportion of mid ranged distance flights for further analysis.

Mixed Variable Analysis:

To examine relationships between variables, we conducted Analysis of Variance (ANOVA) tests and measured their strength using Cramer's V test. Several variables had values of 1.0 or very close to it, indicating a strong relationship. Variables with high Cramer's V values, which exceeded the upper threshold, were excluded from modeling to reduce redundancy. However, some variables were retained temporarily to facilitate further exploratory analysis before being removed later in the process.

Variables: Airline vs Marketing_Airline_Network
Chi-squared: 52307846.7394, p-value: 0.0000000, Cramer's V: 0.9596
Association strength: Very large
Variables: Airline vs Operating_Airline
Chi-squared: 132549291.0000, p-value: 0.0000000, Cramer's V: 1.0000
Association strength: Very large
Variables: Dest vs DestCityName
Chi-squared: 2354327883.0000, p-value: 0.0000000, Cramer's V: 1.0000
Association strength: Very large
Variables: Dest vs DestStateName
Chi-squared: 328217292.0000, p-value: 0.0000000, Cramer's V: 1.0000
Association strength: Very large
Variables: Origin vs OriginCityName
Chi-squared: 2354327883.0000, p-value: 0.0000000, Cramer's V: 1.0000
Association strength: Very large
Variables: Marketing_Airline_Network vs IATA_Code_Operating_Airline
Chi-squared: 52307846.7394, p-value: 0.0000000, Cramer's V: 0.9596
Association strength: Very large
Variables: Operating_Airline vs DOT_ID_Operating_Airline
Chi-squared: 132549291.0000, p-value: 0.0000000, Cramer's V: 1.0000
Association strength: Very large
Variables: Origin vs OriginStateName
Chi-squared: 328217292.0000, p-value: 0.0000000, Cramer's V: 1.0000

Association strength: Very large

We created side-by-side bar charts to visually identify patterns among the mixed variables. One notable finding is that the majority of flights fall into the departure delay group -1. This will be further explored in the *Delay Analysis* section. An interesting trend we observed is that February exhibits the highest likelihood of flight cancellations, suggesting the influence of seasonality and potentially other external factors. We found that most variables related to canceled and diverted flights have missing values. This will be examined in more detail in the *Analysis of Canceled Flights* section.

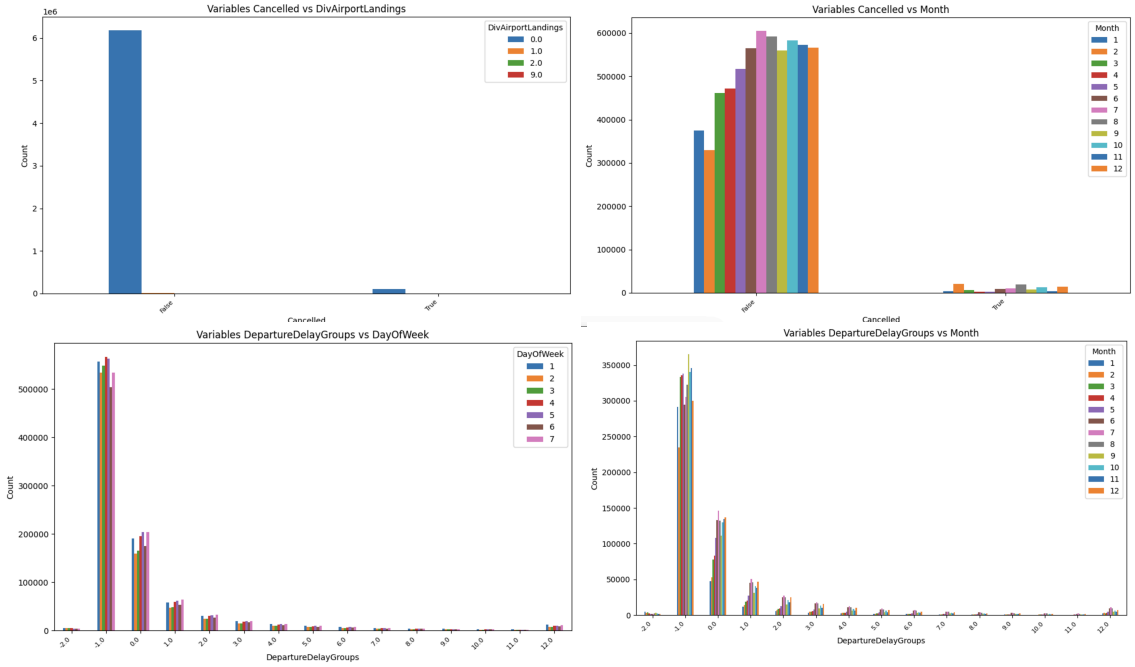


Figure 7. Stacked bar charts to analyze categorical and quantitative data together.

D. Analysis of Canceled and Diverted flights:

Investigation of the validity of canceled flights was crucial in determining our way of imputing missing values. We needed to determine if the flights that contained null values were actual flights and not canceled ones. We ensured the validity of our null values by separating our dataset into canceled and not canceled. Using the not canceled dataset we were able to properly handle any valid null values. Within this non canceled dataframe we can see 3 null values. Because there were only a few null values, we decided to drop them. We were then able to determine that the other null values that were present in the original DataFrame were the result of a canceled flight or diverted, so logically those values would be missing.

E. Analysis of Delays

When slicing the data looking at the different characteristics of each variable, we can determine that the time variables are different representations of each other. Based on the table below we can make observations about the different time variables.

Index	DepDelay	DepDelayMinutes	ArrDelay	ArrDelayMinutes	CRSDepTime	DepTime	DepTimeBlk	ArrTime	CRSArrTime	ArrTimeBlk
0	-10	0	-25	0	724	714	0700-0759	818	843	0800-0859
1	-5	0	-9	0	922	917	0900-0959	1031	1040	1000-1059
2	-9	0	-29	0	1330	1321	1300-1359	1501	1530	1500-1559
3	-9	0	-8	0	1645	1636	1600-1659	2002	2010	2000-2059
4	-6	0	-22	0	1844	1838	1800-1859	1903	1925	1900-1959
5	-2	0	-26	0	1650	1648	1600-1659	1808	1834	1800-1859
6	-1	0	27	27	1652	1651	1600-1659	1929	1902	1900-1959
7	-3	0	-4	0	1245	1242	1200-1259	1452	1456	1400-1459
8	-9	0	-15	0	726	717	0700-0759	821	836	0800-0859
9	-5	0	-31	0	2045	2040	2000-2059	2144	2215	2200-2259
10	2	2	-6	0	1030	1032	1000-1059	1435	1441	1400-1459
11	-10	0	-29	0	1506	1456	1500-1559	1447	1516	1500-1559
12	8	8	8	8	1032	1040	1000-1059	1204	1156	1100-1159
13	-9	0	-10	0	732	723	0700-0759	939	949	0900-0959
14	-9	0	-32	0	902	853	0900-0959	1058	1130	1100-1159
15	-11	0	-35	0	700	649	0700-0759	814	849	0800-0859
16	-7	0	-7	0	2035	2028	2000-2059	2148	2155	2100-2159
17	-8	0	-15	0	538	530	0001-0559	645	700	0700-0759
18	-10	0	-26	0	600	550	0600-0659	718	744	0700-0759
19	-10	0	-21	0	2030	2020	2000-2059	2145	2206	2200-2259

Figure 8. Analysis of variables related to arrival and departure delays

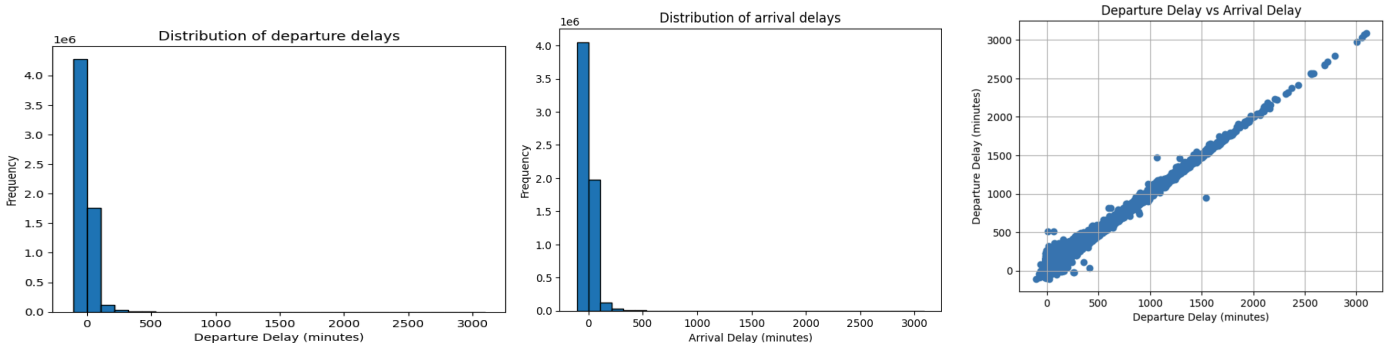


Figure 9. Investigation of arrival and departure delays

The variables *Arrival Time* and *CSRDepTime* are recorded in local time format (hhmm). By dividing these values by 100, we created a new variable representing the hour to better analyze the distribution. Additionally, we observed that *DepTime* aligns with *DepTimeBlk* (hourly intervals), which asserts similar information. Therefore, we can further reduce the data by retaining only one of these variables.

Examining the correlation between *ArrDelay* and *DepDelay*, we found they are highly correlated, so we will retain only one of them for modeling. It is also important to note that *DepDelay* and *ArrDelay* can take negative values, indicating flights arriving or departing earlier than scheduled. On the other hand, *DepDelayMinutes* and *ArrDelayMinutes* record only positive delays, with zero values for flights that were early. This distinction highlights delays while separating them from early arrivals or departures. Since *DepTime* and *ArrTime* are simply timestamps, we will use them to analyze time-of-day patterns

ArrDelay	ArrDelayMinutes																
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max	
ArrivalDelayGroups																	
-2	1583783	-22.921501	6.509963	-107	-26	-21	-18	-16	1583783	0	0	0	0	0	0	0	
-1	2466704	-8.452073	4.179725	-15	-12	-9	-5	-1	2466704	0	0	0	0	0	0	0	
0	1067324	5.65057	4.211112	0	2	5	9	14	1067324	5.65057	4.211112	0	2	5	9	14	
1	401559	21.008963	4.273577	15	17	21	25	29	401559	21.008963	4.273577	15	17	21	25	29	
2	199467	36.278327	4.287548	30	32	36	40	44	199467	36.278327	4.287548	30	32	36	40	44	
3	119535	51.434149	4.299149	45	48	51	55	59	119535	51.434149	4.299149	45	48	51	55	59	
4	79751	66.535304	4.307373	60	63	66	70	74	79751	66.535304	4.307373	60	63	66	70	74	
5	56743	81.576758	4.316791	75	78	81	85	89	56743	81.576758	4.316791	75	78	81	85	89	
6	42103	96.657815	4.303714	90	93	96	100	104	42103	96.657815	4.303714	90	93	96	100	104	
7	31781	111.672226	4.302145	105	108	112	115	119	31781	111.672226	4.302145	105	108	112	115	119	
8	24765	126.73834	4.347631	120	123	127	130	134	24765	126.73834	4.347631	120	123	127	130	134	
9	19838	141.717159	4.322076	135	138	142	145	149	19838	141.717159	4.322076	135	138	142	145	149	
10	15457	156.703047	4.315422	150	153	157	160	164	15457	156.703047	4.315422	150	153	157	160	164	
11	12258	171.776636	4.302565	165	168	172	175	179	12258	171.776636	4.302565	165	168	172	175	179	
12	64802	333.646631	228.732561	180	206	247	343	3089	64802	333.646631	228.732561	180	206	247	343	3089	

Figure 10. Understanding boundaries of Arrival Delay Groups

Examining the distributions of *ArrDelay* and *DepDelay*, we observe that aside from the two negative delay groups and the last delay group, the standard deviations across all other groups are consistently around 4 minutes. Analyzing the distribution of delay minutes within each group reveals that the vast majority of flights experience minimal to no arrival delays, with a significant concentration of data points near 0 minutes. Key insights from the data include the observation that a large proportion of flights arrive early or on time. For example, the *ArrDelayGroups* "-2.0" group, which represents early arrivals, contains over 15.8 million flights with an average delay of approximately -22.9 minutes. Similarly, the "0.0" group, representing on-time arrivals, includes over 10.6 million flights with a mean delay close to zero. These trends indicate that most flights adhere closely to their scheduled arrival times or even arrive ahead of schedule.

After understanding the distributions of the groups, we labeled each group to facilitate better understanding throughout our analysis, shown in the two plots below.

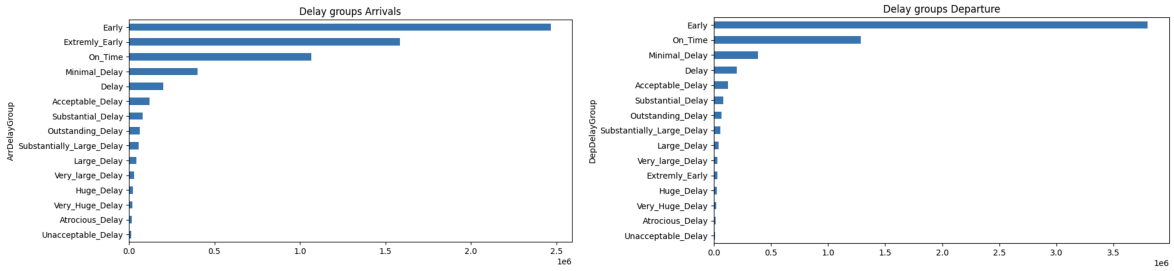


Figure 11. Distribution of Delay Groups for arrival and departure

F. Analysis of delays with time

In order to further explore delays, we aggregated them by different time variables: DepDelay and ArrDelay by hour of the day, DepDelay by day of the month, DepDelay by day of the week

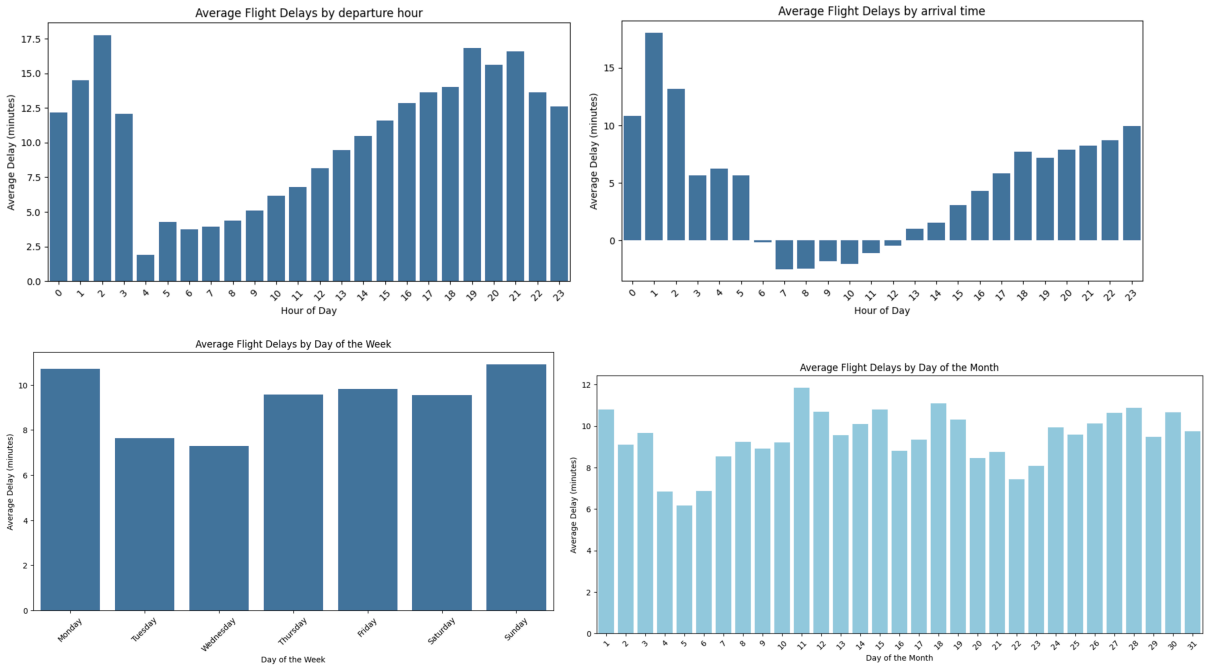


Figure 12. Relationship between delays and time related variables

A rough cyclical pattern in delays emerges across the days of the month, with noticeable dips in delays around the first and third weeks.

Examining delays by the hour of the day reveals a distinct trend for both arrivals and departures. From 7h-12h, average delays are negative, indicating that many flights arrive earlier than scheduled. In contrast, the highest delays occur between 6:00 PM and 2:00 AM, supporting the concept of delay propagation. This aligns with the commonly held belief that morning flights are more reliable than those in the afternoon or evening.

This observation underscores the importance of *DepHour* in capturing patterns that resonate with real-world experiences. Our client can leverage this discovery to collaborate with airlines experiencing frequent delays and optimize gate and terminal allocations to mitigate congestion and enhance operational efficiency.

The table below displays the percentages of flights across *DepDelayGroup* categories by month. The majority of flights fall into the *Extremely Early*, *Early*, or *On Time* groups. However we notice higher delay group percentages in the summer and winter months, which aligns with peak travel seasons.

DepDelayGroup	Extremly_Early	Early	On_Time	Minimal_Delay	Delay	Acceptable_Delay	Substantial_Delay	Substantially_Large_Delay	Large_Delay	Very_large_Delay	Huge_Delay	Very_Huge_Delay	Atrocious_Delay	Unacceptable_Delay	Outstanding_Delay
Month															
1	1.301357	77.610255	12.576046	3.128328	1.528794	0.924431	0.644672	0.438057	0.348363	0.251195	0.202611	0.159366	0.131070	0.113185	0.642270
2	1.027111	71.088124	16.042615	4.335308	2.148977	1.352070	0.886195	0.612258	0.446438	0.358973	0.272115	0.222308	0.187990	0.132413	0.887106
3	0.787333	72.334741	16.861744	4.140178	1.844423	1.077723	0.682862	0.485432	0.352725	0.273232	0.213286	0.158553	0.130534	0.102299	0.554934
4	0.533139	71.178409	17.725118	4.289548	1.922828	1.122375	0.727355	0.513590	0.368034	0.297062	0.221628	0.166168	0.141519	0.110283	0.682945
5	0.389047	65.268499	20.815275	5.257267	2.471194	1.449873	0.978911	0.698309	0.529251	0.384399	0.303840	0.244583	0.195008	0.156084	0.858460
6	0.256100	52.248743	23.471662	7.927204	4.354062	2.832388	1.989443	1.404020	1.096024	0.842057	0.653314	0.535127	0.432580	0.340875	1.616400
7	0.258643	50.455858	24.167494	8.349844	4.527571	2.932774	2.073783	1.519525	1.144825	0.847883	0.664181	0.534694	0.414160	0.345686	1.763080
8	0.363537	54.475047	22.247226	7.707768	4.263562	2.757092	1.842915	1.328680	1.000193	0.750949	0.590769	0.461406	0.361505	0.302073	1.547277
9	0.541352	65.229381	19.848665	5.529584	2.715879	1.612609	1.058202	0.758465	0.544034	0.427966	0.319052	0.237501	0.195473	0.153982	0.827855
10	0.374807	58.373407	22.325246	6.945101	3.678780	2.256578	1.473438	1.030375	0.742908	0.552754	0.425698	0.333716	0.258238	0.202361	1.026592
11	0.337894	60.324022	23.398542	6.620729	3.140457	1.746950	1.115015	0.753185	0.544404	0.396423	0.300156	0.226951	0.186593	0.139770	0.768910
12	0.302117	52.969317	24.106568	8.250621	4.432118	2.708261	1.786511	1.205107	0.888831	0.678039	0.496450	0.376983	0.305480	0.248136	1.245460

Figure 13. Patterns of different types of delays for each month

G. Efficiency Analysis:

To analyze efficiency we create an EfficiencyRatio defined by AirTime divided by Distance. We used this Efficiency Ratio to understand if flights are delayed due to lack of efficiency. This can be interpreted as different altitudes which denote different speeds, circling the airport before landing, etc. When we sort the highest Efficiency Ratio values by ascending = False, we can see many airlines are not efficient.

Index	AirTime	Distance	EfficiencyRatio	Origin	Dest
715827	51	31	1.645	WRG	PSG
4596994	108	73	1.479	DEN	COS
4425483	119	83	1.434	ATL	CSG
4425230	144	106	1.358	CHA	ATL
2736577	39	31	1.258	WRG	PSG
2022635	96	77	1.247	SFO	MRY
5997678	126	113	1.115	SPS	DFW
428966	34	31	1.097	PSG	WRG
2731595	32	31	1.032	WRG	PSG
4425199	106	106	1	ATL	CHA

Figure 14. Analysis of Efficiency of flights in terms of air time and distance.

This analysis highlights the variability in aircraft operations and their potential impact on delays. To explore this further, we developed a function that calculates the average airtime for each distance group across airlines. By examining these metrics both graphically and numerically, we can observe how differences in airline efficiency influence operational performance. Additionally, we observed a high correlation of 0.98 between *Distance* and *AirTime*. To simplify our model and enhance its predictive power, we decided to drop *AirTime* and retain *Distance* as it offers equivalent insights while being more directly interpretable for predicting delays. Additionally, *Distance* is not based on a time value and will always be objective, which led to our decision to drop *AirTime*.

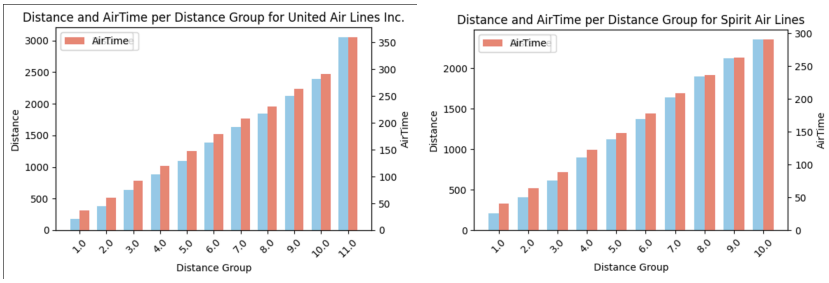


Figure 15. Figures of Distance and AirTime per Airline.

H. Analysis of Airlines:

To understand delays across airlines, we aggregated the data and visualized it graphically. The chart below illustrates the distribution of flights within various departure delay categories for each airline.

Southwest Airlines Co. emerges as the airline with the highest number of flights, showing a broad distribution ranging from minimal to significant delays, along with a notable proportion of on-time flights. Similarly, Delta Air Lines Inc. and American Airlines Inc. demonstrate a mix of on-time performance and varying levels of delays.

This analysis highlights the frequency and severity of delays, which have a direct impact on passenger flow and dwell time at the airport. Such insights can be instrumental in designing strategies to mitigate delays and enhance airport operations.

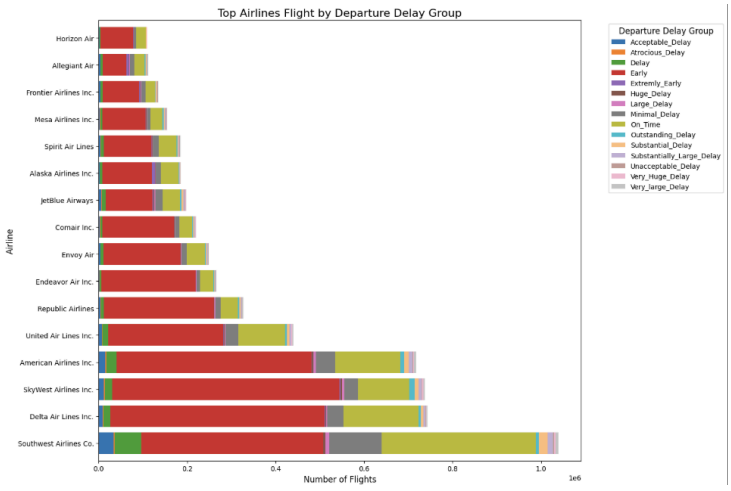


Figure 16. Analysis of distribution of delay types for each airline.

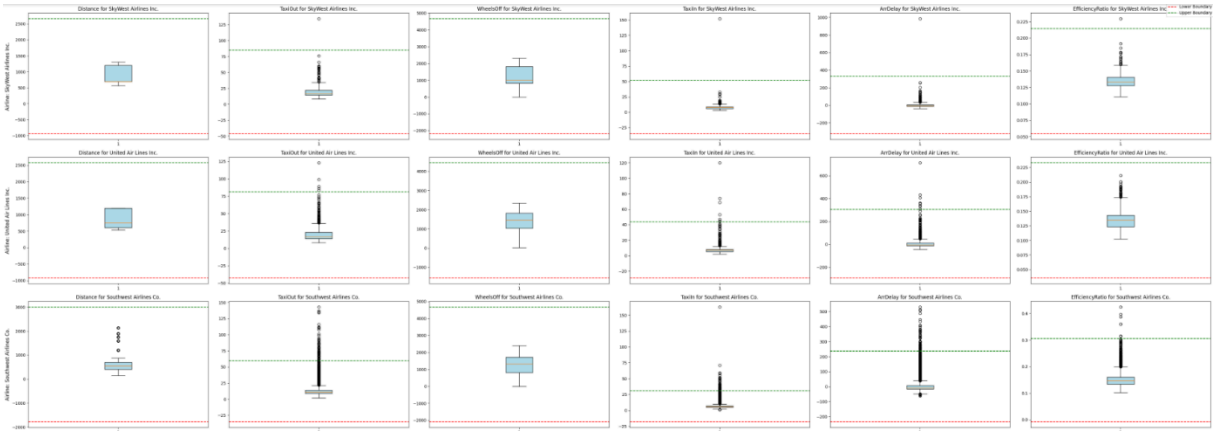


Figure 17. Boxplots of grouped data by airline and origin

I. Preprocessing and data cleaning

After conducting an initial exploration of the data, we narrowed our focus to flights arriving at ATL, our client. To improve the analysis, we removed outliers separately for each combination of airline and origin. This approach was necessary because analyzing the data as a whole could misclassify certain flights as outliers based solely on overall trends. For instance, flights covering shorter distances would dominate the dataset, potentially labeling longer flights as outliers. Below you can find plot for set of airlines with outliers for our variables.

By isolating each origin-airline combination, we accounted for the unique characteristics of each route. This allowed us to identify outliers more accurately, focusing on cases where flight times or operational efficiency deviated significantly from expectations for that specific route. Below we can observe the different distribution and where the outlier boundaries lie. The substantial differences in these boundaries validate our decision to remove outliers on a route-specific basis, ensuring the integrity and relevance of the data for further analysis.

J. Unsupervised analysis

To uncover underlying patterns and structures within our data, we employed unsupervised learning techniques. Unlike supervised learning, which relies on labeled data for predictive modeling, unsupervised learning operates without predefined outcomes. This makes it particularly useful for identifying natural groupings, reducing dimensionality, and gaining deeper insights into the intrinsic relationships and distributions within the dataset.

This approach is especially important for our dataset, which is large and contains a variety of entities, such as airports, airlines, origins, destinations, delay characteristics, and technical flight details. Understanding how these elements relate to one another will be crucial. These unsupervised learning methods will guide our efforts to identify key areas for delay mitigation and operational improvements.

In this section, we will focus on two key unsupervised learning techniques: Principal Component Analysis (PCA), clustering, and clustering using PCA dimensions.

To facilitate unsupervised exploration, we identified a unique unit of analysis. We aggregated the data by *Airline*, *Origin*, *Month*, and *Day*, using the mean to represent each day as a unique data point. This approach allowed us to focus on meaningful patterns at the daily level.

To prepare for unsupervised learning, it was crucial to numerically encode categorical variables. We decided to remove the *TailNumber* variable due to its high cardinality, containing over 3,000 unique values, which can vary by day. Using dummy encoding would introduce excessive noise, while label encoding would result in nonsensical mean calculations and make the results harder to interpret.

K. Principal components Analysis

Principal Component Analysis (PCA) reduces the complexity of high-dimensional data while preserving its essential trends and patterns. It achieves this by transforming the data into a set of uncorrelated dimensions that effectively summarize the original dataset. PCA is especially valuable for visualizing and interpreting data, as projecting rows and columns onto these new dimensions helps elucidate the underlying meaning of each component.

The first initial step of conducting PCA is ensuring there is sufficient correlation among variables. After sufficient correlation was confirmed by examining the correlation matrix, we performed the KMO and Bartlett's test of sphericity which tells us if our data is fit for PCA. Our aloha level is 0.05.

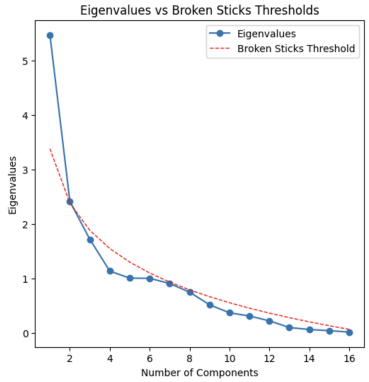
Barlett's test statistic: 1309932.022
p-value: 0.0000

KMO: 0.77

The P-value of Bartlett's test statistic is within our five percent confidence interval and therefore we can reject the null hypothesis and we should conduct PCA. The Kaiser-Meyer-Olkin (KMO) test is 0.777, which means our data is adequately suited (medium bin in range of indexes) and confirms our data is fit for PCA.

We determined the number of components by analyzing explained variance, Broken Sticks threshold, and Karlis-Saporta-Spinaki test.

Karlis-Saporta-Spinaki threshold: 1.026



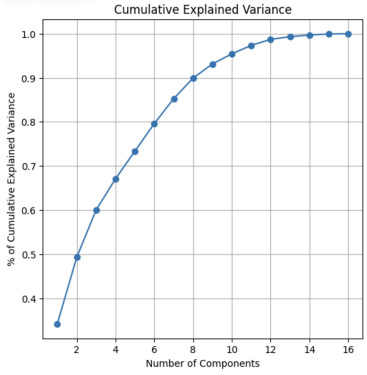


Figure 18. Scree plot with broken sticks rule and cumulated explained variance for PCA

Based on the Broken Sticks threshold and explained variance, it is clear we should retain two factors. We deemed that the explained variance represented by two components (49%) to be sufficient, especially given this type of data. The third principal component represents only 10 % of explained variance and therefore is not worthwhile to retain at the cost of overall interpretability. The Karlis-Saporta-Spinaki test concluded that we could retain four factors. We tried to plot them pair by pair, but it was not interpretable and 4-dimensional space is hard for a human's brain to interpret. Because we want to put precedence on interpretability, we retained two factors. We then plotted the correlation circle and computed the correlation between the variables and factors, cosine squared of each of the variables, and variable contributions.

Contributions of each variable to the first two factors (in %):	F1	F2
ArrDelayMinutes	15.906	32.816
Distance	0.197	32.261
Quarter	0.277	28.564
DayOfWeek	0.009	1.237
DepDelayMinutes	15.938	1.138
DepDel15	12.184	1.030
DepartureDelayGroups	16.344	0.793
TaxiOut	1.024	0.629
WheelsOff	1.787	0.530
TaxiIn	0.375	0.357
ArrDelay	16.284	0.324
ArrDel15	12.441	0.170
DepHour	1.396	0.139
ArrHour	1.006	0.007
EfficiencyRatio	0.088	0.003
ArrDelayGroup_encoded	4.743	0.002

Correlation between variables and factors:	F1	F2
Distance	0.104	0.093
Quarter	0.123	0.007
DayOfWeek	0.022	-0.009
DepDel15	0.816	-0.058
DepartureDelayGroups	0.946	-0.138
TaxiOut	0.237	-0.013
WheelsOff	0.313	0.882
TaxiIn	0.143	0.113
ArrDelay	0.944	-0.166
ArrDel15	0.825	-0.088
DepHour	0.276	0.890
ArrHour	0.235	0.830
EfficiencyRatio	-0.193	-0.123
ArrDelayGroup_encoded	0.509	-0.064

Cos² of the variables on the first two factors:	F1	F2
DepDelayMinutes	0.972	0.0277
ArrDelayMinutes	0.967	0.0332
Distance	0.556	0.4444
Quarter	0.997	0.0030
DayOfWeek	0.856	0.144181
DepDel15	0.995	0.0050
DepartureDelayGroups	0.979	0.020957
TaxiOut	0.997	0.0030
WheelsOff	0.112	0.8884
TaxiIn	0.616	0.384047
ArrDelay	0.970	0.029914
ArrDel15	0.989	0.011358
DepHour	0.088	0.9121
ArrHour	0.074	0.926041
EfficiencyRatio	0.241	0.7592
ArrDelayGroup_encoded	0.984	0.0155

Figure 19. Contributions of variables for PCs, their correlations and cos2 .

Looking at the variable contributions we see that *DepDelayMinutes, ArrDelayMinutes, DepartureDelayGroups, ArrDelay, DepDel15, ArrDel15* are dominant variables in Factor 1. These variables contribute most to the variance explained by PC1. Their loadings range from 0.81 to 0.95, showing that Factor 1 is highly associated with overall delays. Very important insight in F2 - 32% contribution of WheelsOff, it is almost twice the highest contribution in Factor 1.

Based on the cosine squared values, we can conclude that for Factor 1, the variables *WheelsOff*, *DepHour*, and *ArrHour* are not well represented. Additionally, Factor 2 appears to be less effectively represented by the variables compared to Factor 1, as Factor 2 represents less explained variance.

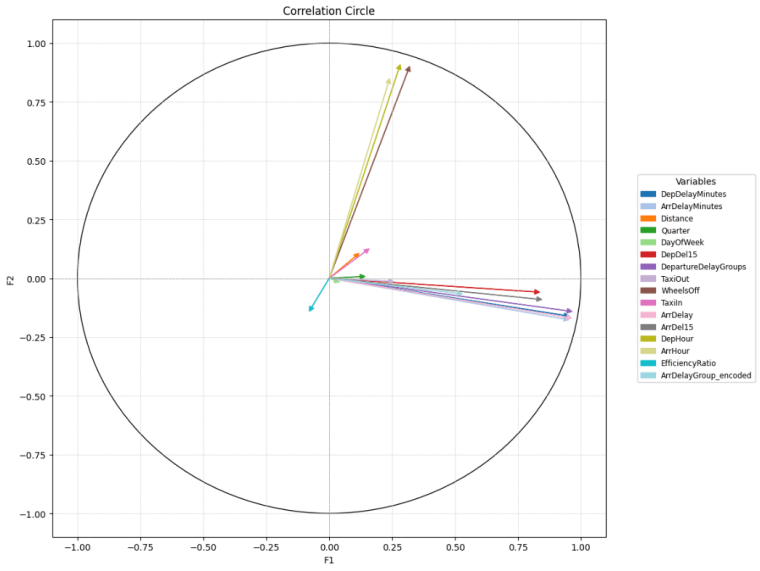


Figure 20. Correlation circle for 2 PCs.

From the correlation circle we can see that only *EfficiencyRatio* is uncorrelated to all the other variables and is very far from the boundary of the circle. We can interpret *EfficiencyRatio* as not having strong correlation and therefore weak explainability for Factor 1.

We then projected the individuals (days in combination with airlines) onto the two dimensions with highlighted extreme variables to better interpret what each dimension represents.

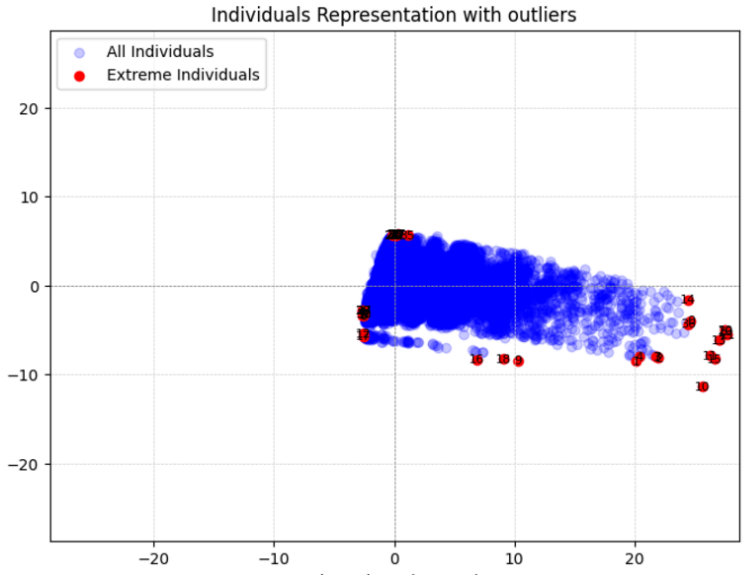


Figure 21. Individuals' representation with highlighted extreme points.

Given the concentration of data points, we specifically extracted extreme points along Dimension 1 and 2.

As we previously asserted, it can determine that F1 is likely related to delays and Factor 2 is more likely related to efficiency in terms of time or operational performance. We can see this through the variable contribution of *Distance* and the correlation of Factor 2 with variables such as *DepHour* and *ArrHour*.

From the extreme values on the individuals representations, we can identify about 14 Airlines with 5 as the most extreme.

We can take Frontier Airlines for example. We see that their points along Dimension 1 are in the range of 10-21 (high positive values) with low values for Dimension 2. We observe that this Airline can also depict patterns closer to zero on dimension 1 and 2. For example, observation 12 (F1 = -2.56, F2 = -5.38) and 17 (F1 = -2.56, F2 = -5.72). These observations assert a lack of efficiency. Delta Air Lines Inc also has high values for F1. We can conclude that SouthWest and Delta are two candidates to look out for since they have frequent delays. American Airlines has one extreme Factor 1 value on October 13th.

Closely **examining extreme** observations in Dimension 1:

1) Row 10: Frontier Airlines Inc, from DEN on 19th December.
   - F1 = 25.67 which means high delay score. This flight had a big delay because we can confirm the row-wise behaviour of Factor 1 and Factor 2. We can then say that the airline suffers from substantial delays, both at departure and arrival. The Factor 1 score indicates a high likelihood of delayed flights.
   - F2 = -11.37 means extreme negative efficiency for this flight. Because we have a set of rows with the same behavior, we can suggest poor operational efficiency, possibly due to late departures, long taxi times, which further exacerbate the delays.

2) Row 13: Frontier Airlines Inc, from LGA on 9th November
   - F1 = 27.06 means extremely high delay score. This flight is significantly delayed, likely suffering from both departure and arrival delays.
   - F2 = -6.16 means negative efficiency. Flight is not operating efficiently, meaning there could be inefficient taxi times, late takeoffs, or operational disruptions. We can then say that this company as a whole is inefficient and suffers from delays since it appears in outliers we have similar behaviour several times.

Closely **examining extreme** observations in Dimension 2:

Extreme values assert operational efficiency (low or high) based on high *WheelsOff* and *DepHour* and *ArrHour* contributions. From the plot we can see extreme values in the range -9 to -16 and 22- 37.

1) Row 23: JetBlue Airways from FLL on 15th September
   - F1 = -0.18 means a very low delay score. This flight is typically on-time or only experiences minimal delays.
   - F2 = 5.65 means high efficiency. That flight was operated efficiently, likely due to quick taxi times, timely departures, and smooth operations overall.

2) Row 37: Spirit Airlines from MCO on 31st October
   - F1 = 0.36 means very low delay score. It means the flight had a minimal delay and was on time. Because of the row-wise behavior for this airline with a low Factor 1 we can conclude that delays don't occur for this company frequently.

- F2 = 5.79 asserts high efficiency. We have a set of rows with high Factor 2, it shows that Spirit Airlines operates with good efficiency, suggesting timely departures and quick taxiing.

L. Clustering Analysis

Cluster analysis identifies similarities and groups within data by leveraging row-wise similarities, as opposed to PCA, which focuses on column-wise variance. We employed clustering to uncover inherent groupings in the data and to analyze the characteristics and behaviors of the observations within these clusters.

To achieve this, we utilized the K-means algorithm, which partitions the dataset into a predefined number of clusters (k). The algorithm assigns each observation to the cluster with the nearest mean, which iteratively updates the cluster centroids to minimize intra-cluster variance.

To determine the optimal number of clusters, we generated an elbow plot using a range of cluster values and their corresponding inertia. The elbow plot identifies the point where inertia - representing the within-cluster sum of squared distances - experiences a significant drop before plateauing. Using this plot, we were able to select k = 2 and proceed with our analysis.

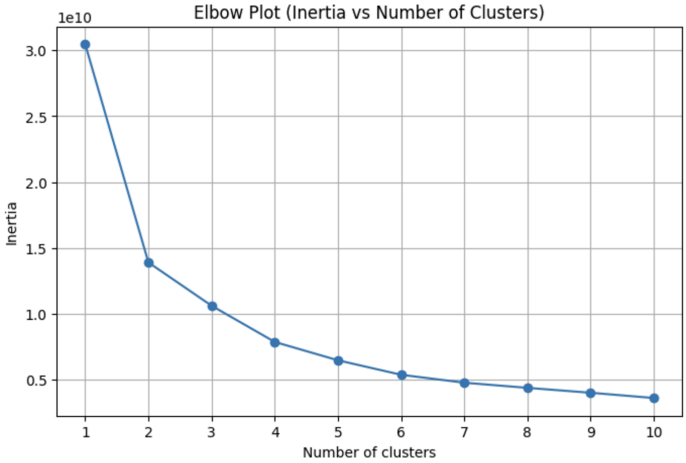


Figure 22.Elbow plot for Kmeans clustering

Using our selected k = 2, we generated silhouette scores and plotted them respectively for each cluster.

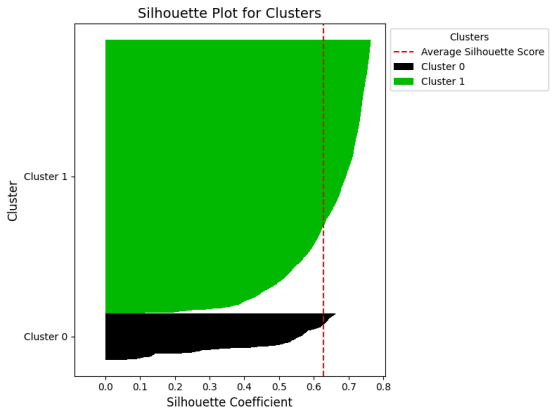


Figure 23. Silhouette plot for 2 Kmeans clusters

The silhouette plot is sharp enough and both clusters are over the threshold. Both clusters are above 0.6 where a value close to 1 shows that a point is far from the neighboring clusters, and a value of 0 indicates that such a point is remarkably close to the decision boundary between the two neighboring clusters. Neither value is below the average score which asserts cohesion within black cluster is good. We can see cohenison in the green cluster is not as good in comparison.

To get a better understanding of the clusters, we conducted descriptive statistics. We then grouped the data by each variable and evaluated the summary statistics. We can see that for many of the variables, the clusters have a lot of overlap in terms of distributions.

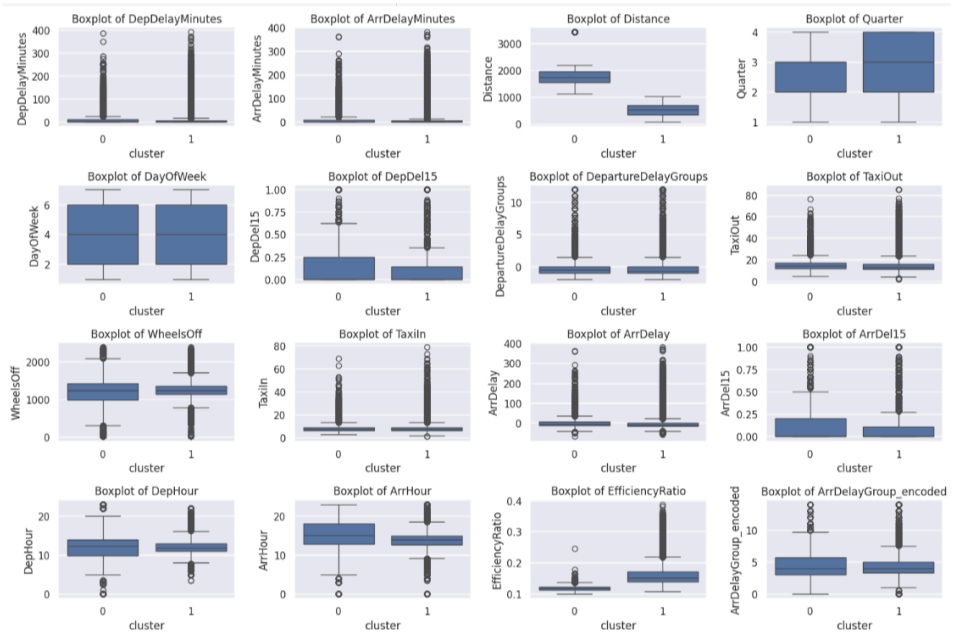


Figure 24. Boxplots of all variables for clusters to analyze difference

We can say that cluster 0 likely indicates longer haul flights based on the distance variables mean and standard deviation which is twice that of cluster 1.

Cluster 0 contains 13,288 days of flights for each airline and origin, while Cluster 1 includes nearly six times as many flights, with a total of 77,962. Cluster 1 exhibits lower mean values for *ArrDelay* and *DepDelayGroups*, along with shorter average distances, suggesting it primarily consists of shorter routes with better time performance.

The variables *ArrHour*, *DepHour*, *Quarter*, and *DayOfWeek* show similar values across both clusters, indicating that these factors do not significantly differentiate the two clusters. This suggests that, despite the presence of long-haul flights, these variables do not have a substantial impact on the cluster differentiation.

ArrDelay and DepartureDelayGrups show lower mean times for cluster 1 (-3.56 and -.26) compared to cluster 0 (.43 and -.06). ArrHour and DepHour are approximately the same for both clusters, therefore we cannot interpret those variables. Taking a look at DepDel15 and ArrDel15: cluster 0 16.8% of the flights had departure delays exceeding 15 minutes and 15.9% of the flights had arrival delays of over 15 minutes. Meanwhile, cluster 1 has a lower value. 12.6% of cluster 1's departures were more than 15 minutes delayed and 11.6% of their arrivals exceeded 15 minutes. When we look at the mean ArrDelay, because it is negative, we can conclude that the flights arrived earlier than expected.

TaxiIn and TaxiOut indicate that Cluster 0 is associated with longer average taxi-out times with a mean of 15.2 minutes and an average taxi-in time of 8.1 minutes. Cluster 1 has an average taxi-out time of 14.4 minutes and average taxi-in time of 7.8 minutes. We can associate the longer taxi times for the long haul flights although there is no literature that states longer flights have longer taxi times.

Unsupervised learning conclusions

The overall understanding of our PC analysis implies based on the unsupervised models we were able to extract companies who are deemed extreme disruptions to the flight environment and our clients can utilize this information through a variety of our suggested solutions. For example, maintaining a separate and upcharge gate for frequently disruptive airlines. We were also able to group some of the characteristics that long haul and short distance flights have in common. These long haul flights are characterized by moderate to severe delays, while shorter distance flights are characterized with better time performance.

M. Feature Engineering

After analyzing the PCA, which brought us insights that some airlines are performing worse, we realized that we need to create separate variables for each airline and route
We have created 4 additional variables: average air time, average departure delay, average efficiency, average arrival delay group (implying previous flights).

Before supervised learning one important moment should be taken into account and done. Supervised learning in our case will predict something for the future, so for future flights we cannot use some flight characteristics, as they relate to the very fact of the flight - the delay time of departure, the number of hours in the air. However, parameters such as taxiing time are planned in advance, the distance is known, so we keep them.

N. Supervised analysis

Our supervised analysis focused on multiclass classification, with the target variable representing delay groups across 15 classes of varying severity. Given the inherent imbalance in our target variable, we employed stratified sampling during the train-test split to preserve the class distribution. We opted not to apply oversampling or undersampling techniques, as we believed that generating synthetic data or removing observations could compromise the applicability of our results. Imbalanced data is a frequent challenge in the real-world and especially in the aviation industry. We aim to ensure that our future models remain robust under such conditions. Additionally, we could not use many variables because they are direct indicators of the present flight. However, features such as taxiing or distance are planned in advance so we can use them in our analysis.

To address the classification task, we experimented with a range of tree-based models, including Gradient Boosting Classifier, XGB Classifier, Random Forest Classifier, and LightGBM Classifier. While all these models rely on decision trees as their foundation, they employ distinct algorithms and optimization strategies, offering varied performance characteristics and trade-offs.

Random forest results:

- Test accuracy 0.6495, Kappa 0.4413, F1 0.6120
- Early has a large diagonal value (23434), meaning most predictions for this class were correct (true positive)
- Extremly_Early also has a significant diagonal value (13709), which is good.
- For the Minimal_Delay class, you see some misclassifications into Early and On time delay classes (1160 and 917).
- This suggests overlap or confusion in the features for these classes. It might be due to similar patterns in the data for these labels.
- On time flights were misclassified as early as 5140 cases .
- Extremely early cases were misclassified as early as 5125.
- And early were misclassified as extremely early in 3469

The performance of other models can be found in the notebook, they showed worse results. We will not discuss them and tell you which model was the best.

Among these baseline models, XGB Classifier performed the best. The algorithm is a combination of gradient boosting with advanced regularization, parallelism, and efficient computation.The XGB classifier is well-equipped for tabular data, which is our case. Within the baseline model, we set the parameter objective='multi:softprob'. This parameter specifies that the model uses the softmax function to output class probabilities for multiclass classification.

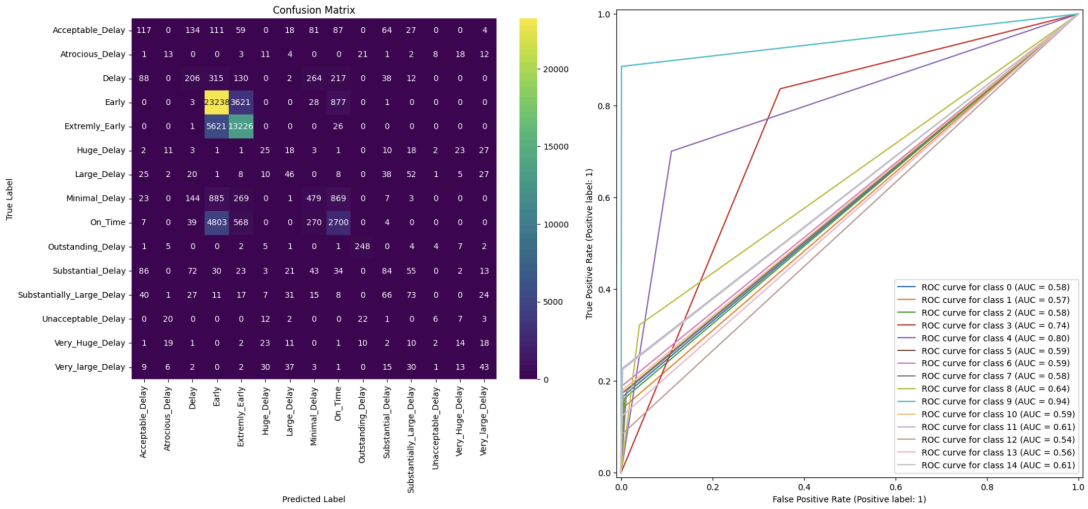


Figure 25. Confusion matrix and ROC AUC curve for Random Forest

Best Performing Model:

The XGBoost (XGB) classifier outperformed all other models, with most models performing in the 0.5-0.6 range, while XGB showed a clear advantage.

Across all models, predicting significantly large delays proved challenging, as many were incorrectly predicted as on-time. This may be due to the class imbalance in the data and the lack of training on features that contribute to such extreme delays. A similar issue occurred with predicting minimal delays, which were often classified as on-time. This pattern was consistent across all models.

Examining the ROC curve, we found that the most accurately predicted class was Class 9 (Outstanding Delay and Very Huge Delay) with an AUC of 0.56, while Atrocious Delay (AUC 0.57) was the least accurately predicted, as confirmed by the confusion matrix. Conversely, Early and Extremely Early classes were predicted relatively well, likely due to the abundance of such observations compared to extreme delays.

Overall, given the nature of the data and the distribution of classes, the model performed reasonably well. We will proceed with this model for the final sections of the analysis.

Next we will evaluate variable importance for this model in the next section, then we will take that into account when tuning our final model.

O. Hyperparameter Tuning

After determining our optimal base model, we tuned hyper parameters to ensure robustness. We employed three methods of hyperparameter tuning to the most finely tuned final model.

We decided to tune both learning and tree based parameters for optimal results which is an advantage of this model as it is an ensemble algorithm.

tree_method: The algorithm used for constructing the tree. (exact, approx, hist, gpu_hist)

- Subsample: If subsample is set too low, there may not be enough data for each tree to capture meaningful patterns. We observed that the differences in model metrics for subsample values (e.g., 0.6 to 1.0) are small, suggesting that the model is relatively stable to variations in subsampling. When analyzing the confusion matrix, we found that misclassifications for on-time flights were 4,598 (compared to 4,803 for the initial model), and the number of correct predictions increased to 2,819 (compared to 2,700 in the initial model). This indicates that

subsample has a positive impact on the model's performance. A subsample range of 0.7 to 0.8 introduces enough randomness to help reduce overfitting, while still providing enough data for each tree to learn effectively. This range strikes a good balance between reducing model complexity and ensuring adequate learning.

- Max_depth: Represents the maximum depth of a tree (complexity).
- Min_child_weight: Represents the minimum sum of instance weights/ number of samples required in a child node. Higher values prevent the model from learning overly specific patterns (overfitting) by making the trees more conservative
- eta: this is the learning rate. The default is .3. The learning rate is crucial to efficiently and correctly convergeing the model on a global minimum.
- Colsample_bytree: Represents the fraction of features to sample for building each tree. Values between 0 and 1 help to reduce overfitting by preventing the model from relying too heavily on any single feature.
- Alpha: Controls L1 regularization in XGBoost. L1 adds a penalty to the absolute magnitude of the feature weights, encouraging sparsity in the model. For us it can mean that our DataFrame likely benefits from minimal regularization, suggesting that most features are relevant.

1. **Manual/ Sequential Tuning**

Manual tuning can be very powerful because it allows individual parameters to be maximized independent of other parameters. We found the ideal set of parameters to be:
alpha = 0.0001
eta = 0.5
colsample_bytree = 1
min_child_weight = 7
subsample=0.7
tree_method = 'exact'
max_depth = 12
produced an accuracy of:

Training and test metrics for XGBClassifier:
TRAIN ACCURACY: 0.9578, TEST ACCURACY: 0.6934
TRAIN KAPPA: 0.9377, TEST KAPPA: 0.5359
TRAIN F1: 0.9577, TEST F1: 0.6821

Max Depth Tuning: As we increased the *max_depth*, the accuracy improved, which is logically consistent. In XGBoost, increasing the tree depth enhances the model's ability to capture more complex patterns in the data. Deeper trees allow the model to split the data into smaller, more specific regions, helping it identify intricate patterns and relationships. This is particularly advantageous for large datasets, where more complexity is often required.
Overfitting Consideration: Due to system limitations, we were unable to further increase the depth. However, from the trend observed (with the difference between *max_depth* = 11 and *max_depth* = 12 being minimal), we can infer that increasing the depth further could lead to overfitting. The model may start memorizing the training data rather than generalizing, potentially harming performance on unseen data.

Results for max_depth = 12 on Test set:

Accuracy: 0.7026
Precision: 0.6846
Recall: 0.7026
F1-score: 0.6886

The exact tree construction method was computationally the most expensive. However, it is beneficial when features like *EfficiencyRatio* require precise thresholds to maximize predictive power, as the *exact* method excels at identifying these thresholds. Improved Performance: After switching to the exact method, we observed an improvement in the model's performance. For on-time flights, the misclassifications decreased to 4,663, compared to 4,803 in the initial model. Additionally, the number of correct classifications increased to 2,754, up from 2,700 in the initial model. This indicates that the exact method has enhanced the model's ability to predict on-time flights accurately.

The eta parameter provided us with good results. As the learning rate controls the step size at each boosting iteration, it helps determine how quickly the model converges. We observed significant improvements in the prediction of on-time flights. The number of correct predictions increased to 3,085, and misclassifications decreased to 4,288, a noticeable improvement from the initial model. This demonstrates that adjusting the learning rate has positively impacted model performance.

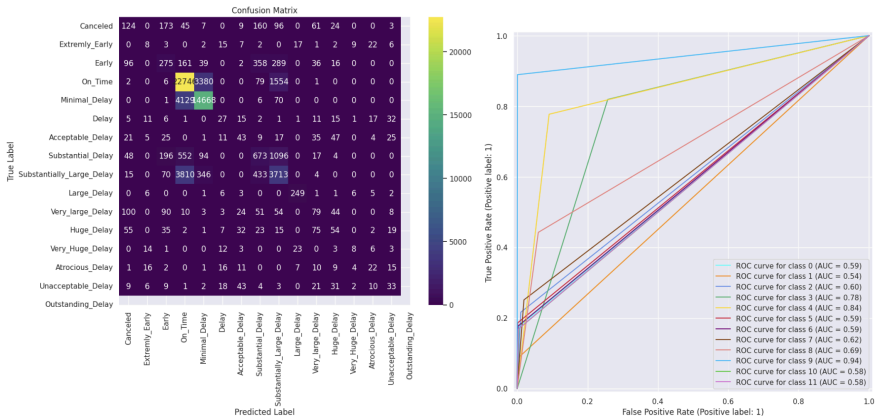


Figure 26. Confusion matrix and ROC AUC curve for tuned XGBoost with all best tuned parameters

2. Simultaneous Tuning via RandomizedSearchCV

Through RandomizedSearchCV, the best set of parameters are:

Tree_method = hist
subsample= 0.9
Min_child_weight = 9
Max_depth = 10
Eta = 0.3
Colsample_bytree = 1.0
Alpha = 0.01
Best score: 0.687

We can see that in comparison to manual tuning the optimal tree-method found here is ‘hist’. Additionally the alpha is larger but the eta is smaller. This means the steps taken by this tuned model are smaller than the manually tuned model, which in theory would provide a better score as it could find a better minimum. The search also found this to be optimal, likely because it is computationally faster.

3. **Stratified K Folds Cross Validation**

Because our target variable is imbalanced, we opted for stratified K Fold cross validation to maintain class proportions of our target variable in each of the K folds. We set the number of splits to 10.
Using the parameters form the manually tuned optimal model, our results were:
Maximum Accuracy That can be obtained from this model is: 0.7017560943941312 (Fold: 9
Minimum Accuracy: 0.6960
Overall Accuracy: 0.7000
Standard Deviation is: 0.00159

P. Feature importance:

To ensure the best interpretability, we decided to employ three different methods of feature importance to gain maximum insights.

1. Built in model importance function

In this built- in function, we tried both the gain and weight parameters, which gave us varying results. The gain parameter asserted that the variables representing each airline were less important than the other included variables while the weight method showed nearly the opposite results. The weight parameter tells us how frequently a feature is used in splits across all trees while the gain parameter tells us how much a feature improves the model's performance when it is used for splitting.
In our case, we will interpet using the gain parameter. WheelsOff and EfficiencyRatio are the most important predictors of any delay class, which is corroborated by our exploratory analysis.

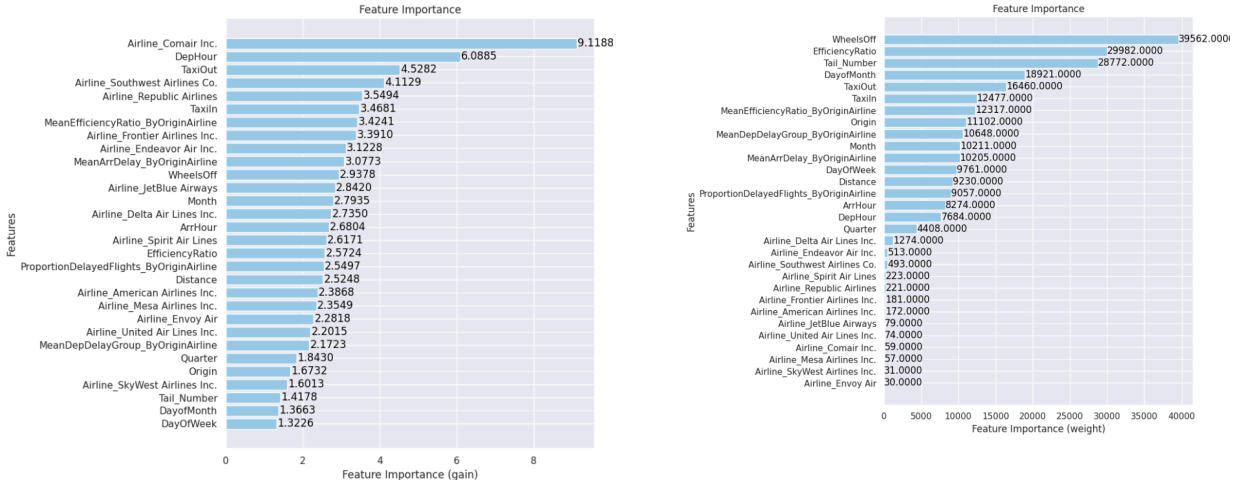


Figure 27. Feature importance of XGBoost with gain and weight methods.

2. **Permutation feature importance**

The baseline score of the model is 0.698, meaning it correctly predicts the class for 69.8% of the test samples. This serves as our reference for evaluating the importance of different features.
Similarly to the built-in model importance (gain), *WheelsOff* is the most important, with a significant decrease in accuracy of 0.43 when permuted. This indicates that *WheelsOff* is a critical feature for making accurate predictions making it a vital feature in our model.
DepHour, EfficiencyRatio, and MeanEfficiencyRatio_ByOriginAirline are also highly important. The hour of departure likely plays a significant role, as certain peak hours may influence delays and flight performance. *EfficiencyRatio* measures historical airline performance and is crucial in predicting delays, as airlines with better past efficiency (fewer delays, better operational performance) tend to have more predictable outcomes.

Similarly to our built-in feature importance model (weight), the dummy-encoded airlines appear to have little importance because they are ranked at the bottom of the plot. This suggests that the model relies more on continuous or aggregated features rather than on the specific dummy encoding of categories.

The aggregated mean features related to specific origins and airlines are more impactful. This suggests that certain airlines (as observed in PCA) are less efficient, which affects the predictions. These aggregated mean features highlight the significant role of airline-specific operational patterns in delay prediction.

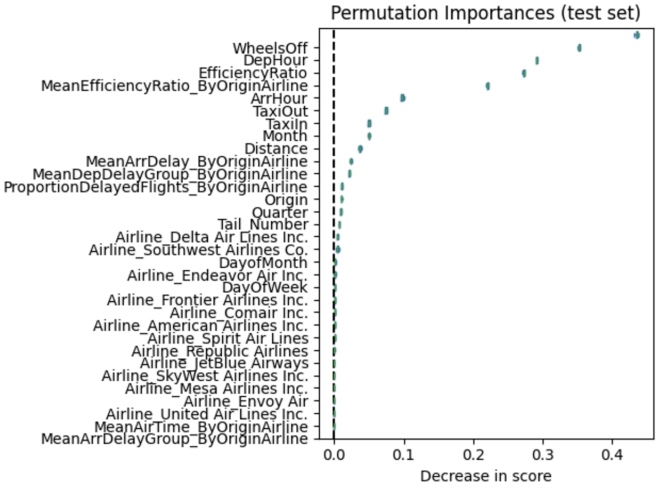


Figure 28. Permutation importance of XGBoost

3. SHAP analysis

The shapley additive explanations (SHAP) algorithm is a powerful model-agnostic method that uses a unique game theory algorithm to fairly allocate the contribution of each feature to the model's predictions. We employed SHAP for feature importance analysis due to its ability to provide clear, interpretable insights into how each feature influences model predictions, regardless of the underlying model type.

After initializing the SHAP explainer and fitting it to our tuned model, we applied it to predict on the test set (X_test) and derived the SHAP values for each feature. To better interpret these values, we applied a

softmax function to convert the SHAP values into probabilities, making it easier to understand their impact on the prediction outcomes.

The beeswarm plot below shows the distribution of the SHAP values across all classes for each variable. This visualization provides a clear picture of how each feature contributes to the model's predictions, with the color representing the feature value (from low to high), and the width of the plot indicating the distribution of SHAP values.

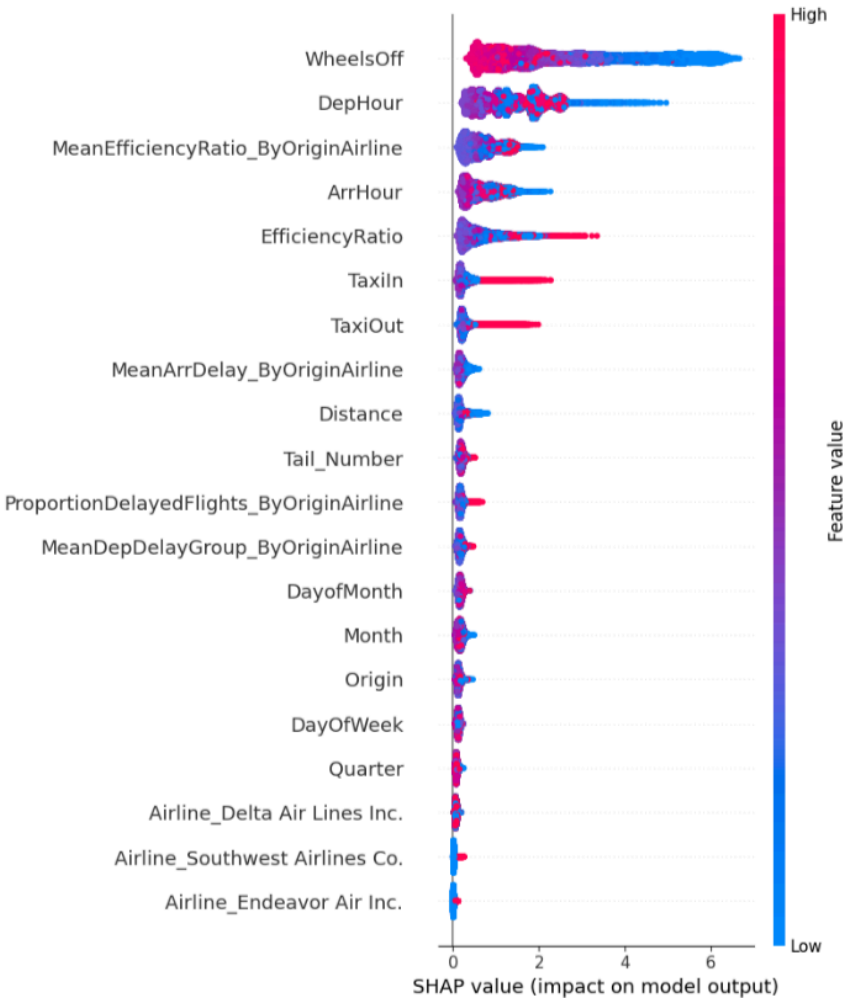


Figure 29. Overall SHAP values representation

A higher mean absolute SHAP value suggests that a feature has a greater influence on the model's prediction. Each dot represents the SHAP value for an instance

As observed with permutation importance and the built-in model package using the weight method, the most important and impactful features in defining the delay group are *WheelsOff, DepHour, and MeanEfficiencyRatio_ByOriginAirline.* The SHAP values for *WheelsOff* show that higher values tend to correspond to higher SHAP values, suggesting a strong positive influence on the model's prediction. In

contrast, higher values of *TaxiIn* seem to have a slightly negative impact on the SHAP value, indicating that longer taxi times may result in a small decrease in the likelihood of certain delay classes.

To further illustrate the contribution of each feature, we took sample rows from each class and analyzed the individual contributions each variable made to the predicted delay class through waterfall plots.

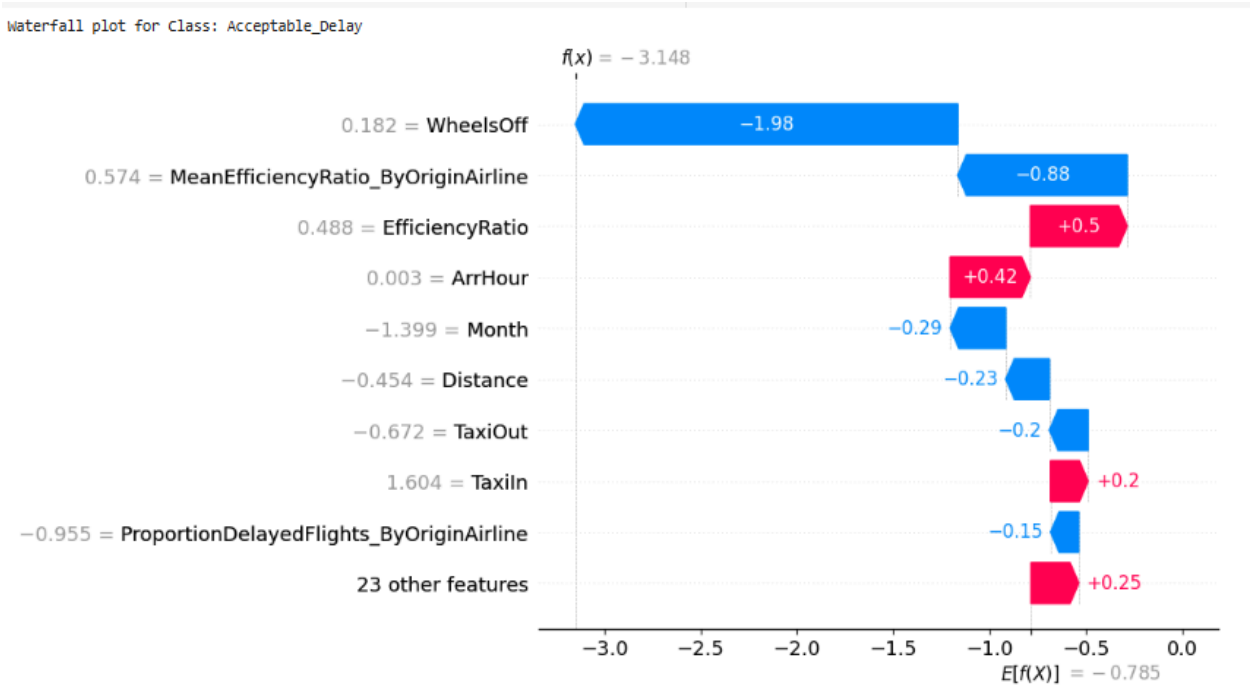


Figure 30. SHAP values for 1 instance assigned to class Acceptable delay

In this waterfall plot for the Acceptable Delay predicted class, the predicted SHAP value for this particular delay group is f(x) = -3.148, while the average prediction is E[f(x)] = -0.785. The *WheelsOff* feature contributed -1.98 to the decrease in the predicted probability of an "Acceptable Delay." This means that higher values of WheelsOff lowered the likelihood of this delay group being predicted.

The *MeanEfficiencyRatio_ByOriginAirline* contributed -0.88, further decreasing the predicted probability of the Acceptable Delay class. This suggests that lower efficiency ratios (indicating worse performance by the airline or origin) were associated with a decrease in the probability of the flight being classified as Acceptable Delay.

On the other hand, *EfficiencyRatio* without considering airline and origin had a positive contribution of 0.5, meaning that higher efficiency ratios generally increased the likelihood of the Acceptable Delay class being predicted. This means that flights operated by more efficient airlines tend to be classified as Acceptable Delay more frequently.

V. Conclusion

Our analysis has provided valuable insights into the aviation landscape surrounding ATL, enabling us to offer actionable recommendations. Through our unsupervised analysis, we were able to identify airlines that exhibit distinct disruptive delay patterns. This information allows ATL to proactively schedule and anticipate potential delays, particularly for airlines like Delta and Southwest, which can be better prepared for these disruptions. We are also able to identify groups of characteristics that indicate operational efficiency which our client can gain valuable insights from.

By leveraging these findings, ATL can optimize gate management and implement more efficient flight scheduling, particularly for connecting flights. Personalized insights into each airline's performance can help improve operational decision-making and reduce delays. We can now successfully implement our initial value propositions to our clients:

- Reduce delay propagation
- Create new revenue streams
- Reduce energy costs
- Correctly allocate personnel and equipment
- Improve gate profitability
- Increase concessions revenue through alignment with pricing models and stock management
- Improve customer satisfaction and experience

Since our analysis is based on 2021 data, the model can be retrained annually to incorporate the most recent data, allowing ATL to continuously refine its predictions and adapt to evolving trends in the aviation industry.

VI. Discussion

Future Improvements:

While our current model provides valuable insights, it does have limitations, particularly when it comes to predicting the emergence of new airlines or capturing the evolving behavior of existing ones. Since our analysis is based on 2021 data, it may not be entirely relevant for the future, given the potential changes in airline operations and industry dynamics over time. The model is more effective for short-term decision-making, where past data can help forecast the near future, but it may not account for broader shifts in the industry.

One promising avenue for improvement is the development of a time-series model using data from multiple years. Using this historical data, we could predict delays for 2025, considering both seasonal patterns and external factors such as holidays or events. Time-series forecasting would require a substantial amount of historical data, but with sufficient data from both public and proprietary sources, we could build a more robust model that incorporates seasonality, external variables, and emerging patterns to improve long-term predictions.

By working closely with clients over multiple years, we would have access to a rich dataset that would enhance our ability to predict delays more accurately, taking into account the continuous evolution of airline operations and external factors.

References

Airports Council International. (2021). *World airport traffic dataset*.

Bureau of Transportation Statistics. (2021). *National Transportation Statistics*.

Federal Aviation Administration. (2021). *2021 Annual Traffic Report*.

Henriques, R., & Feiteira, I. (2018). Predictive modelling: flight delays and associated factors, hartsfield–Jackson Atlanta International airport. *Procedia computer science*, *138*, 638-645.

Neyshabouri, S., & Sherry, L. (2014, January). Analysis of airport surface operations: a case-study of Atlanta Hartsfield airport. In *Proceedings of the Transportation Research Board 93rd Annual Meeting*, Washington, DC, USA (pp. 12-16).

U.S. Department of Transportation. (2021). *Air Travel Consumer Report: December 2021*.

Yablonsky, G., Steckel, R., Constales, D., Farnan, J., Lercel, D., & Patankar, M. (2014). Flight delay performance at hartsfield-jackson atlanta international airport. Journal of Airline and Airport Management, 4(1), 78-95.

.