CSE 519: Data Science Fundamentals
Understanding Flight Delays

Project Final Report

Abstract

In this project, our aim is to investigate the underlying factors contributing to flight delays by analyzing various parameters such as wind speed and direction across a wide range of airports. Additionally, we will conduct an in-depth analysis to investigate how arrival and departure patterns across different aircraft carriers might impact the carrier delay times. Moreover, we will explore whether there is any correlation between delays caused by weather conditions and those arising from the cascading effects of arrival and departure delay times.

Note: Our project focuses on data analysis rather than model building.

1 Introduction

The escalating demand for air travel has underscored the necessity of scrutinizing flight delays, a pivotal concern for airlines aiming to streamline operations and enhance profitability. A 2020 study by the FAA projected that flight delays would impose a staggering $32.9 billion cost on the US economy. This encompasses additional operational expenditures for carriers, such as heightened fuel consumption, increased maintenance expenses, and elevated crew costs. Moreover, carriers might face penalty costs based on the extent of delays. For passengers, delays lead to reduced productivity and a loss of confidence in airlines, potentially resulting in increased fares.

Various factors contribute to flight delays, including weather conditions, carrier issues, security concerns, and logistical challenges related to aircraft. This project primarily zooms in on delays induced by weather conditions and their consequent impacts on departure and arrival schedules. By focusing on these aspects, the project aims to conduct a deeper analysis into the multifaceted nature of weather-related delays and their downstream effects on airline operations.

2 Datasets

To gather information on flight delays across various airports, we will utilize the publicly available Airline On-Time Performance Data provided by the Bureau of Transportation Statistics (BTS) [3]. This dataset is renowned for its extensive coverage and reliability in aircraft data.

To investigate the relationship between flight delays and weather conditions, we will analyze data from the Global Surface Summary Of Day dataset, compiled by the National Centers for Environmental Information (NCEI) [5]. This dataset encompasses information from stations nationwide.

Concentrating on the ten major airports depicted in Figure 1, we've compiled a dataset that spans from 2000, comprising approximately 30 airlines per airport, with each airport having its own weather station data. Careful extraction of arrival and departure statistics has been undertaken to gain a comprehensive understanding of scheduling dynamics.

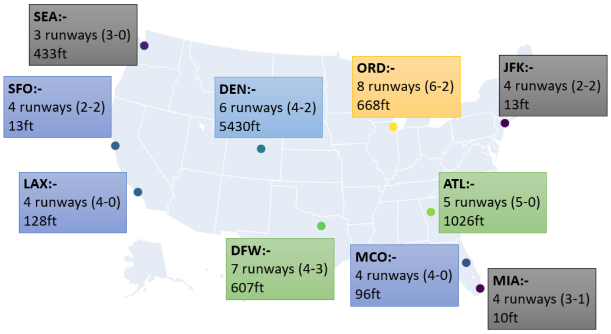


Figure 1: US Map displaying a diverse array of airports used in our analysis, spread across states with varying geographic characteristics. Each text box includes the Airport Code, Number of Runways (parallel - perpendicular), and Altitude. Altitude has been considered a significant factor due to the variability of wind speeds at different altitudes.

Feature	Description (Unit)
TailNumber	Aircraft ID Number
DepDelay	Departure Delay (min)
ArrDelay	Arrival Delay (min)
WeatherDelay	Weather Delay (min)
Origin	Origin Airport Code
Station	Station ID
WSF2	Fastest 2-min wind speed (m/s)
WDF2	Direction of fastest 2-min wind (m/s)
PGTM	Peak gust time (HHMM)

Table 1: The first section of the table includes the primary features utilized in the airline dataset, while the second section encompasses the main features employed in the weather dataset.

3 Data Pre-Processsing and Feature Selection

For our analysis, we utilized an extensive range of features, yet we have outlined the key ones in Table 1. The subsequent subsections will detail the steps taken in data cleaning to prepare it for analysis.

3.1 Airline Data and Features

- Flights that were diverted or canceled have been excluded from our analysis. These exclusions are based on specific flags representing each feature in the dataset.
- Early departures and arrivals are denoted by negative values in the dataset. However, since our analysis focuses on delays, any negative values representing early arrivals or departures have been adjusted to zero.
- To ensure data integrity for our analysis, any rows with NULL values for either the tail number or date have been removed. These fields serve as key identifiers in our analysis.
- Our project focuses on only ten specific airports, and accordingly, data for other airports has been filtered out.

3.2 Weather Data and Features

- Some wind parameters such as WSF2 and WDF2 contain NULL values due to missing data records, primarily from historical data. This issue is not prevalent for dates within this decade.
- For accurate representation of wind direction data, we have grouped directional data into approximately 20 buckets. This categorization facilitates easier visualization through diagrams and enables efficient analysis.

We've opted to disregard values occurring within a 5-second timeframe as abrupt alterations within such a short interval wouldn't serve as an effective metric for determining delays. Moreover, the variance during this brief period would be considerable. Hence, the decision was made to consider values across a 2-minute time span instead as you can see in Table 1. While the peak gust time might provide a general overview of peak wind speed patterns, it won't be utilized for an in-depth analysis. We've filtered out these particular values and omitted others from our analysis dataset.

4 Experimental Setup

4.1 Wind Speed and Direction Analysis

Before delving into the analysis of flight delays concerning varying wind conditions, it's essential to outline the factors influencing flight delays and understand how wind speed and direction might impact these factors. In aviation, winds are categorized into three types based on the direction they influence: headwind, tailwind, and crosswinds. A headwind occurs when the wind direction opposes the plane's direction; a tailwind occurs when both directions align, and a crosswind transpires when they are perpendicular to each other.

Wind can contribute to flight delays, particularly during high-altitude flight or during takeoff and landing. At higher altitudes, headwinds impede travel time, whereas tailwinds expedite it. This effect is more noticeable during extended-duration flights and may result in delays. However, for our analysis, focusing on domestic flights, we consider this as a relatively minor contributing factor. A more significant factor is the impact of crosswinds on planes during takeoff and landing, which will be our primary focus for analyzing delays.

Crosswinds create challenges for safe airplane takeoffs and landings, leading to waiting times or alterations in runway schedules, consequently causing delays. This becomes particularly challenging in airports with high airplane traffic, resulting in cascading delays, a phenomenon we'll further explore below.

While there isn't a rigid criterion for defining a crosswind as hazardous, and numerous variables lie beyond this project's scope, a general consensus suggests that crosswinds exceeding speeds of 25-30 knots (approximately 13 m/s) pose significant challenges for pilots handling aircraft.

4.2 Runway Analysis

A comprehensive understanding of runway intricacies is pivotal to thoroughly assess how wind impacts flight delays. Airports commonly employ two runway types based on orientation: perpendicular (including intersecting runways without a 90-degree angle) and parallel runways. Perpendicular runways offer flexibility as airports can adapt to crosswind direction, minimizing operational disruptions. However, they cannot accommodate simultaneous takeoffs and landings. In contrast, parallel runways maximize efficiency by allowing simultaneous departures and arrivals but are more susceptible to delays in changing wind conditions, leading to crosswinds.

Before construction of any airport, extensive analysis utilizing historical wind data for the region is conducted. Hence, this paper aims to comprehensively analyze a diverse array of airports, providing detailed deep dive analyses for each. In subsequent sections, we will explore the rationale behind the orientation of runways and their impact on the average flight delay for each airport for different wind conditions.

Initially, our analysis covered the ten airports listed in Figure 1. However, due to space limitations, we've opted to display only five of them in this report. Each runway number indicates the direction in degrees, multiplied by 10. For instance, Runway 13 represents the 130-degree direction.

4.2.1 John F. Kennedy International Airport, New York

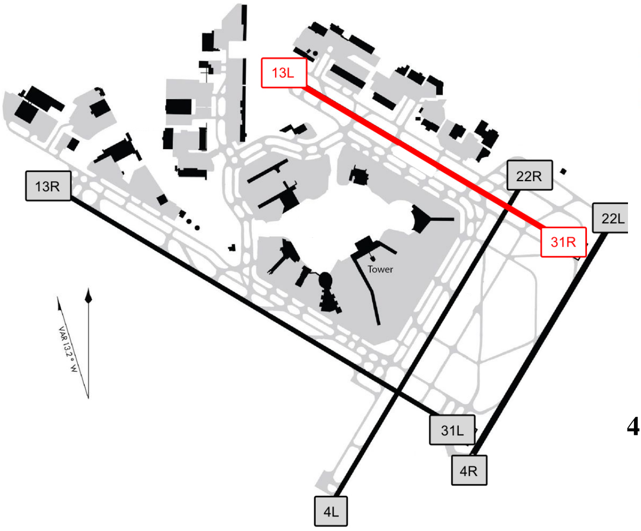


Figure 2: JFK Airport Runway Map

JFK has four runways of which two runways intersect perpendicularly to another runway. They are 13L-31R, 13R-31L, 4L-22R, 4R-22L as shown in the Figure 2.

4.2.2 Los Angeles International Airport, California

LAX has four runways, all of which are parallel: 6L-24R, 6R-24L, 7L-25R, and 7R-25L, as illustrated in Figure 3.

4.2.3 Hartsfield-Jackson Atlanta International Airport, Georgia

ATL has five runways, all of which are parallel: 8L-26R, 8R-26L, 9L-27R, 9R-27L, 10-28, as shown in Figure 4.

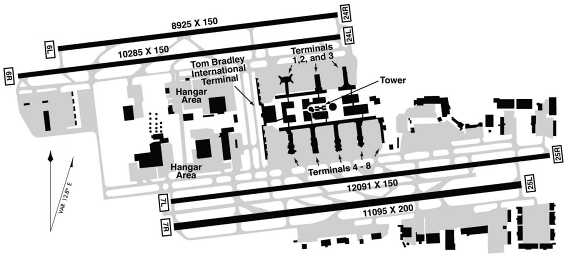


Figure 3: LAX Airport Runway Map

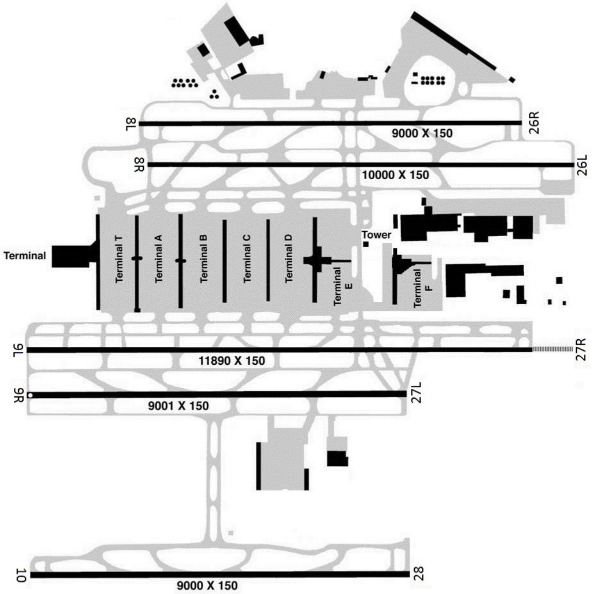


Figure 4: ATL Airport Runway Map

4.2.4 Miami International Airport, Florida

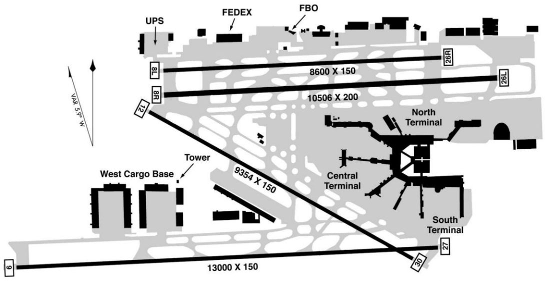


Figure 5: MIA Airport Runway Map

MIA has four runways, with three being parallel and one intersecting another. They are labeled as 8L-26R, 8R-26L, 9-27, and 12-30, as depicted in Figure 5.

4.2.5 Denver International Airport, Colorado

DEN boasts six runways, including two perpendicular runways that do not intersect with any other runway: 7-25,

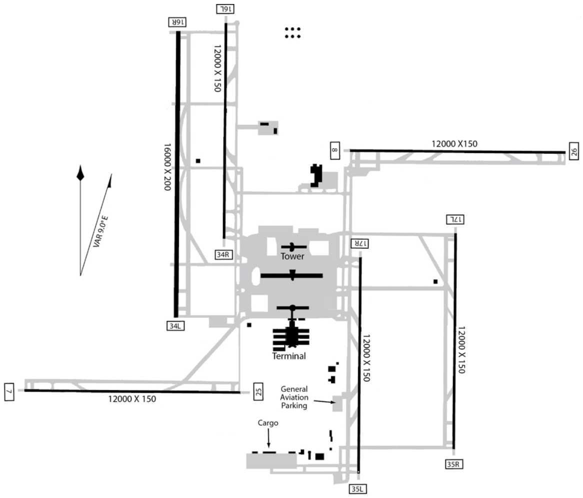


Figure 6: DEN Airport Runway Map

8-26, 16L-34R, 16R-34L, 17L-35R, and 17R-35L, as depicted in Figure 6. A notable reason for selecting this airport is its possession of the longest runways in the US, owing to its higher altitude, where the air density is lower than at sea level. This choice aims to investigate the effect of crosswinds on the arrival and departure of flights using these extended runways.

5 Results

5.1 Flight delay analysis for varying wind conditions

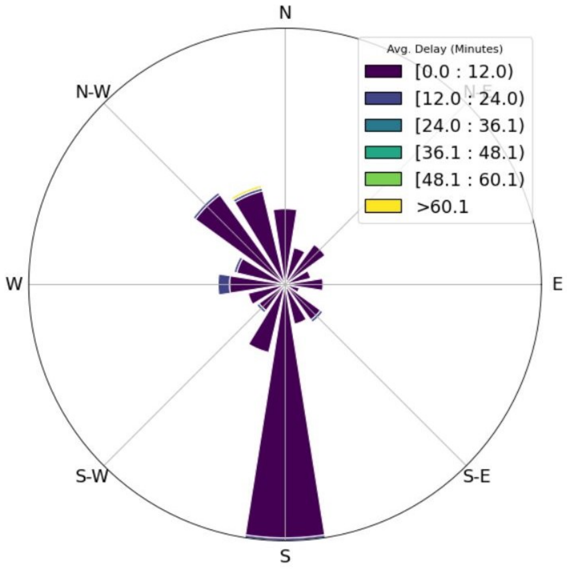


Figure 7: Histogram plot displaying the relationship between wind direction and flight delays for JFK Airport.

In Figure 7, the most frequent wind direction appears to be South. The majority of delays, falling within the 0 to 12 minute range, are associated with this direction. This indicates that delays are more probable when the wind is coming from the South. JFK Airport features two perpendicular runways positioned at 130 and 220 degrees, aligning with the southwest and southeast directions, respectively. Ideally, delays should be minimal when winds are aligned with these angles as they correspond to the runways' orientation. However, challenges arise when the wind direction is 180 degrees, making it difficult for the aircraft control tower to switch runways. This misalignment increases the possibility of crosswinds, leading to airline delays.

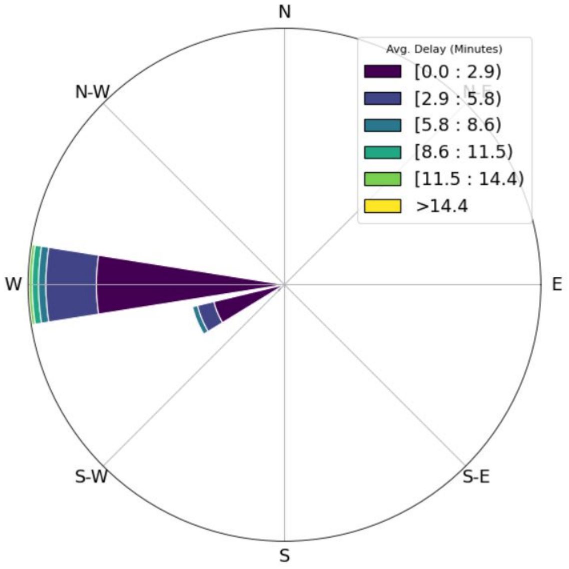


Figure 8: Histogram plot displaying the relationship between wind direction and flight delays for LAX Airport.

In Figure 8, the wind histogram exclusively displays winds from the west direction. This observation elucidates the rationale behind constructing parallel runways at Los Angeles Airport. Historical wind patterns indicate a predominant unidirectional flow, minimizing the probability of crosswinds in this region. Consequently, flight delays at this airport remain low due to the reduced likelihood of crosswind-related issues.

In Figure 9, while the prevailing winds are primarily directed eastward, notable winds are also observed in other directions, such as west and north-west. Atlanta, recognized as the busiest airport in the United States in terms of passenger traffic and aircraft operations, features 5 parallel runways oriented towards the east to maximise efficiency. Although a perpendicular runway design could have mitigated crosswind-related delays, it appears the designers opted for a parallel runway configuration to accommodate the substantial passenger volume efficiently. This choice might elucidate the comparatively higher delay times at Atlanta Airport in contrast to LAX airport due to crosswind related delays.

In Figure 10, the histogram illustrates predominantly eastward winds, with a minor occurrence of south-eastern

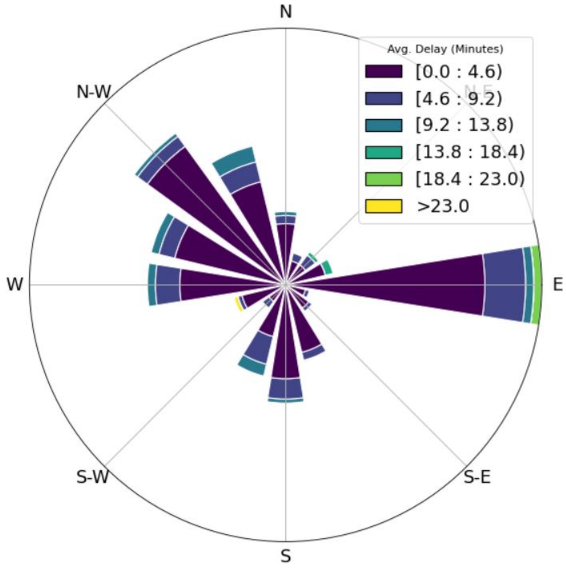


Figure 9: Histogram plot displaying the relationship between wind direction and flight delays for ATL Airport.

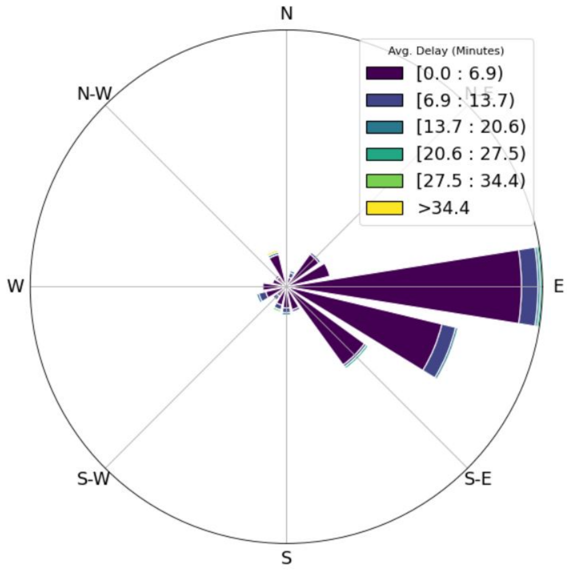


Figure 10: Histogram plot displaying the relationship between wind direction and flight delays for MIA Airport.

winds and negligible winds in other directions. This distribution aligns with more than 90% of the delays lasting under 7 minutes, indicating a correlation between wind directions and runway orientations. The airport features three parallel runways oriented to the east, while one runway is positioned in the south-east direction at 120 degrees. This alignment with runway directions potentially contributes to the shorter delay times observed.

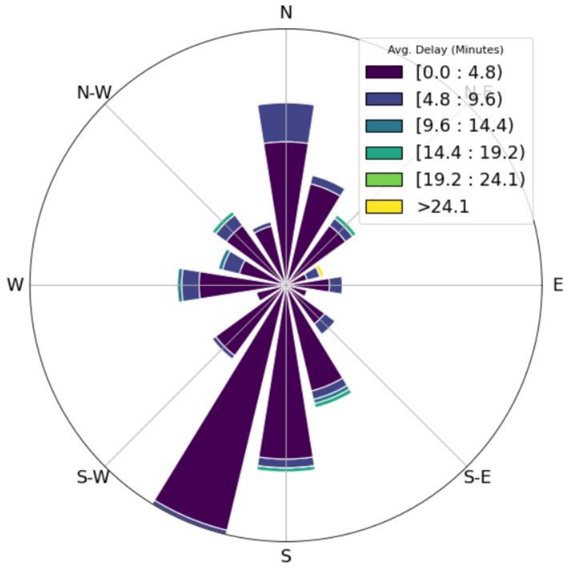


Figure 11: Histogram plot displaying the relationship between wind direction and flight delays for DEN Airport.

In Figure 11, the plot resembles ATL airport, but the prevailing wind is between the south and south-west directions. Denver Airport, in terms of passenger traffic and runway count (it has 6 runways), bears some similarities to Atlanta Airport, explaining this resemblance. However, notable differences include the number of perpendicular runways and Denver's higher altitude. A comparison of the color grading in both plots reveals that over 90% of the delays at DEN are under 5 minutes, whereas the delays at ATL airport are more widely distributed.

5.2 Analysis of arrival delay effect on departure delay

- Departure delay is the difference between the official departure time and the actual departure time of the flight measured in minutes.
- Arrival delay is the difference between the official arrival time and the actual departure time of the flight measured in minutes.
- Positive values for delay indicate that the flight was delayed, negative values indicate the flight departed/arrived early.

From Table 2 and Table 3 we can see that, on average, across all airports, the arrival delay tends to be approximately 10 minutes, while the departure delay averages around 15 minutes. Additionally, the median arrival delay is negative, indicating that, on average, flights tend to arrive early. Similarly, the median departure delay is negative, suggesting that, on average, flights tend to depart earlier than scheduled.

Figure 12(a) illustrates the correlation between arrival delay and departure delay for four major airports: Hartsfield-Jackson Atlanta International, Miami International, John F. Kennedy International, and Los Angeles International. The analysis unveiled a correlation coef-

Airport	Max	Min	Mean	Median
Hartsfield-Jackson Atlanta International	1484	-155	6.97	-5
Los Angeles International	1761	-90	10.63	-2
John F. Kennedy International New York	1547	-62	15.64	-2
Miami International	1678	-41	12.32	-2
Orlando International	1285	-94.0	8.43	-3
Seattle Tacoma International	1642	-95.0	4.0	-4
Denver International	1501	-117	5.85	-4
Dallas Fort Worth International	2536	-153	5.16	-5
San Francisco International	1339	-117	12.10	-3
Chicago O'Hare International	2050	-134	9.04	-5

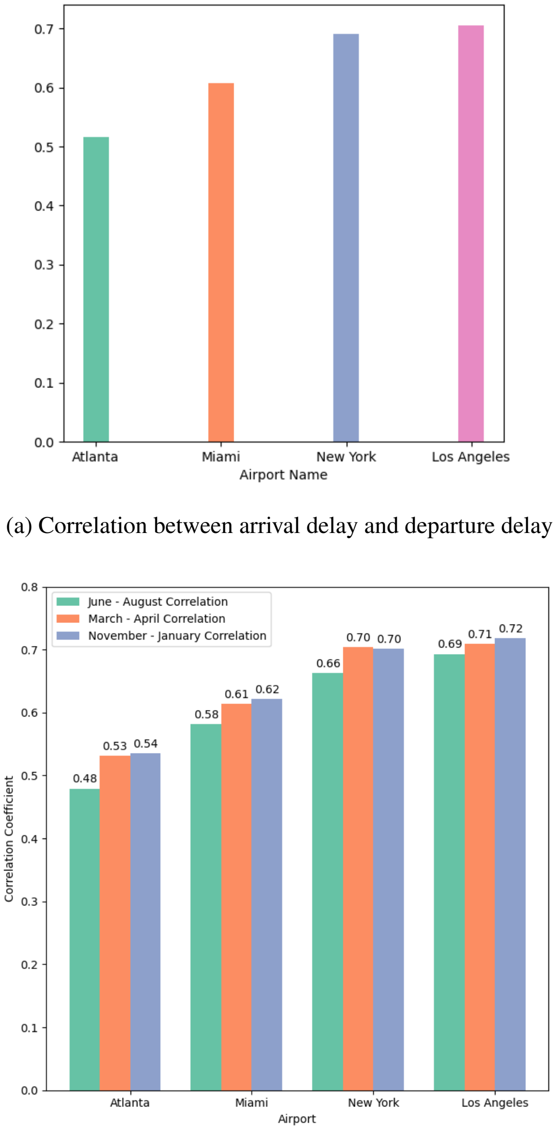
Table 2: Arrival Delay Statistics(Minutes)

Airport	Max	Min	Mean	Median
Hartsfield-Jackson Atlanta International	1484	-155	6.97	-5
Los Angeles International	1761	-90	10.63	-2
John F. Kennedy International New York	1547	-62	15.64	-2
Miami International	1678	-41	12.32	-2
Orlando International	1670	-68	14.68	-1
Seattle Tacoma International	1544	-191	9.44	-2
Denver International	1560	-162	11.50	-1
Dallas Fort Worth International	4225	-65	10.80	-2
San Francisco International	1476	-62	16.345	-1
Chicago O'Hare International	2130	-65	14.75	-2

Table 3: Departure Delay Statistics(Minutes)

ficient exceeding 0.5, indicating a robust positive linear relationship between arrival and departure delays. This consistent pattern suggests that flights experiencing delays upon arrival are also more likely to encounter delays upon departure. Further exploration into operational factors contributing to these delays could offer valuable insights for enhancing airline scheduling and operational efficiency.

Notably, the correlation coefficient for Hartsfield-Jackson Atlanta International Airport is comparatively lower than that of the other three airports, despite its status as the busiest airport. This discrepancy may be attributed to superior scheduling strategies employed by Hartsfield-Jackson Atlanta International Airport, differentiating it from its counterparts. Investigating these strategies further could unveil valuable practices that contribute to the airport's efficiency in managing delays.



(a) Correlation between arrival delay and departure delay

(b) Seasonal Correlation of Arrival and Departure Delays

Figure 12

Figure 12(b) illustrates the seasonal correlation between arrival delay and departure delay for the four airports under consideration. The seasonal data was extracted by filtering airport data based on three distinct categories: June - August (Summer), March - April (Spring), and November - January (Winter). Notably, the correlation coefficients are consistently higher during the Winter season for all airports. This observation may be attributed to significant travel events during this period, such as Thanksgiving, Christmas, and New Year. The increased number of travelers during these holidays likely contributes to heightened airport traffic, resulting in a

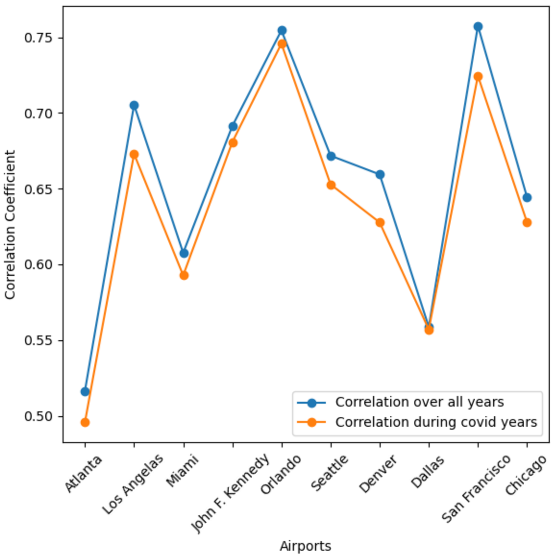


Figure 13: Comparing Correlations: During the Covid-19 Period vs. Overall Trends"

notable increase in the correlation coefficient compared to other seasons. Further investigation into the specific factors influencing these patterns during the Winter season could provide valuable insights into the dynamics of airport operations during peak travel periods.

In defining the COVID years as spanning from 2019 to 2021, an examination of the correlation between arrival delay and departure delay revealed a consistent trend across ten different airports which is depicted in figure 13. Notably, the correlation during the designated COVID period was found to be lower compared to the overall correlation. This divergence in correlation could be attributed to multifaceted adjustments within the aviation industry prompted by the pandemic. Airlines strategically modified flight schedules, often reducing frequencies and altering routes. The introduction of health and safety protocols, involving enhanced screening and distancing measures, potentially impacted the efficiency of boarding and departure processes. Furthermore, the surge in flight cancellations and rescheduling during the pandemic emerged as a factor influencing the observed correlation between arrival and departure delays. These findings underscore the intricate interplay of operational changes and external influences on the temporal relationship between arrival and departure delays during the unprecedented challenges posed by the COVID-19 pandemic.

6 Statistical Analysis

6.1 Impact Assessment of Arrival Delay on Departure Delay

In this analysis, I performed a comprehensive examination to uncover the intricate relationship between departure delay and arrival delay, aiming to grasp the temporal intricacies influencing the timely commencement of flights. The variable of interest, ’Arrival Delay,’ represents the temporal deviation between the actual and scheduled arrival times of flights. Employing a meticulous regression analysis, I sought to quantify the extent to which changes in arrival delay are associated with changes in departure delay. The statistical significance of the regression coefficients, gauged through metrics such as the F-statistic and R-squared, played a pivotal role in deciphering the impact of arrival delay on departure delay. Additionally, a t-test was executed to scrutinize the individual significance of the ’Arrival Delay’ coefficient. Here are the key hypotheses:

- **Null Hypothesis:** The coefficient for ’Arrival Delay’ is zero, indicating no discernible impact on departure delay.
- **Alternative Hypothesis:** The coefficient for ’Arrival Delay’ is non-zero, suggesting a statistically significant influence on departure delay.

Assumptions intrinsic to the t-test, including the normality and independence of residuals, were rigorously examined to underpin the reliability of my findings. Through these analyses, I aim to determine whether ’Arrival Delay’ serves as a statistically significant predictor of ’Departure Delay.’

Dependent Variable	Departure Delay
R-squared	0.267
F-Statistic	2.022e+06
Method	Least Squares
No. Observations	5560002
Df Residuals	5560000
Df Model	1
Covariance Type	non robust

Table 4: Unveiling the Relationship: Regression Analysis of Departure Delay vs. Arrival Delay

Variable	**constant**	**Arrival Delay**
Coefficient	4.9301	0.5471
Standard Error	0.014	0.0
t	346.75	1422.06
P > |t|	0.0	0.0

Table 5: T Test Results for the regression model in Table 4

Table 4 displays the outcomes of the regression analysis performed for Hartsfield-Jackson Atlanta International Airport, providing insights into the relationships between variables. Additionally, Table 5 presents the results of the T-Test conducted on the previously analyzed regression model, further evaluating the statistical significance of the model’s coefficients. Key insights drawn from the results include:

- The positive coefficient for Arrival Delay in the regression analysis (0.5471) implies that, on average, an increase in Arrival Delay is associated with a corresponding increase in Departure Delay.

- The model, as indicated by the high R-squared value (0.267) and a significant F-statistic (2.022e+06), is effective in explaining the variability in Departure Delay.
- The T-Test reinforces the significance of the Arrival Delay coefficient, with a T-Statistic of 1422.057 and a p-value of 0.0.

The results yield compelling evidence affirming the hypothesis that Arrival Delay serves as a substantial predictor of Departure Delay at Hartsfield-Jackson Atlanta International Airport. Intriguingly, this consistent trend persisted across diverse airports subjected to the same analysis, underscoring the universal significance of Arrival Delay as a predictive factor for Departure Delay. These collective findings underscore the broader implications and suggest the potential applicability of operational insights across various airport contexts.

7 Meta Analysis

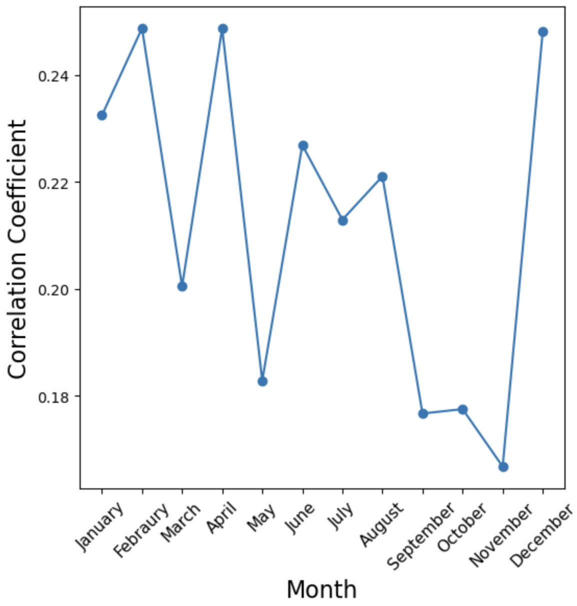


Figure 14: Correlation of Arrival Delay vs Wind Speeds by Month for ATL Airport.

We aimed to analyze the impact of diverse wind conditions on creating a chain of delays. Our hypothesis revolves around the notion that airports experiencing strong crosswinds may encounter delays for both departing and arriving flights. Consequently, this could lead to subsequent flight delays, as the runways remain occupied with the earlier delayed flights, resulting in longer queues.

To test our hypothesis, we plotted a correlation between arrival delays and wind speeds by month. Interestingly, colder winter months exhibit a stronger correlation compared to other months, although it's important to note that correlation does not imply causation. Conclusively attributing arrival delays directly to wind speed is challenging. While higher wind speeds typically impede flight, the wind direction plays a crucial role. When the wind aligns with the aircraft, it doesn't cause delays, whereas crosswinds tend to do so.

Hence, analyzing the cascading effect concerning wind direction is challenging due to its directional vector nature, which demands extensive expertise for an in-depth meta-analysis.

8 Conclusion

This paper presents a thorough investigation into the causes of flight delays, specifically focusing on wind conditions and cascading effects. The analysis involves rigorous statistical examinations of various parameters, and the paper outlines the steps taken for data selection and pre-processing. The central findings highlight a significant and consistent interdependence between arrival and departure delays across multiple airports. Statistical tests, including regression analysis and hypothesis testing, confirm the existence of a discernible cascading effect, wherein delays in arrival impact subsequent departures. The time series analysis uncovers nuanced patterns, with monthly variations suggesting potential seasonal influences. Notably, findings during the Covid-19 period underscore the adaptability of these relationships to external disruptions. Visualizations, such as correlation comparisons, provide intuitive insights, emphasizing the universal nature of these dependencies across diverse airports. This project contributes valuable knowledge for optimizing airline operations and underscores the importance of strategic planning to minimize the ripple effects of delays throughout the flight itinerary.

9 Challenges

- Disentangling the impact of external factors, especially weather conditions, posed a challenge in understanding the true causes of delays.
- Aligning arrival and departure times became a temporal challenge, demanding precision to capture the true temporal dependencies.
- Interpreting statistical results posed challenges, emphasizing the need to transform numerical findings into actionable insights for practical relevance in airline operations.
- The nature of wind directional data, being inherently directional in nature, presents challenges in creating comprehensive visualizations.

References

[1] L. Carvalho, A. Sternberg, L. Maia Goncalves, A. Beatriz Cruz, J.A. Soares, D. Brandao, D. Carvalho, and E. Ogasawara, 2020, "On the relevance of data

science for flight delay research: a systematic review," *Transport Reviews*.

[2] Yuemin Tang, 2021. "Airline Flight Delay Prediction Using Machine Learning Models." In *2021 5th International Conference on E-Business and Internet (ICEBI 2021)*, October 15-17, 2021, Singapore, Singapore. ACM, New York, NY, USA, 7 Pages.

[3] Airline On-Time Data - `https://www.bts.dot.gov/`

[4] Rule OST 2000-8164 to determine delay categories - `https://www.regulations.gov/document/DOT-OST-2000-8164-0059`

[5] NOAA National Centers of Environmental Information. 1999. "Global Surface Summary of the Day - GSOD. 1.0." *NOAA National Centers for Environmental Information*.

[6] A Comparative Analysis of Delay Propagation on Departure and Arrival Flights for a Chinese Case Study by Zhe Zhing (https://www.mdpi.com/2226-4310/8/8/212).