Airline Flight Performance

Analytics Project - Final Report

ISDS 577 - Group 1

Sairaj Prakash Jadhav

Ray Lien

Keti Lin

Hy Luong

Rucha Nilangekar

Table of Contents

Executive Summary

This report presents a comprehensive analysis of airplane delays within our airline operations, aiming to enhance operational efficiency and customer satisfaction. The research makes use of sophisticated analytics, predictive modeling, and data-driven insights to offer insightful suggestions for strategic decision-making.

We expect to find important insights in our upcoming research into flight delays that will greatly influence how our airline operates. Through careful data analysis, we want to pinpoint the major causes of delays, such as bad weather, backed-up airports, mechanical problems, and flight schedules. We plan to create precise forecasting models that will enable proactive delay management and resource allocation using predictive modeling methods, including decision trees, random forest, and regression analysis. In order to effectively address underlying issues, our goal is to conduct a root cause analysis to identify the source of delays. The results of this investigation are expected to guide operational changes, such as increasing passenger communication, optimizing aircraft schedules during busy times, and gate management. By implementing these data-driven recommendations, we aim to reduce delays, enhance operational efficiency, and ultimately help airlines gain a competitive advantage in the industry.

Data Collection & Preparation

The Carrier On-Time Performance dataset, sourced from the United States Bureau of Transportation Statistics via Kaggle, is publicly available at *kaggle.com*. There are no legal or privacy concerns as the data is collected under the Community Data License Agreement. The dataset is in a common-separated value (CSV) format. The project will utilize two datasets: the original data and a dataset containing airline codes translated to their respective airline names. Our dataset has more than 2 million flights from 1987 to 2020, and it includes 109 different attributes. It will be necessary to preprocess the Carrier On-Time Performance dataset because it's fairly big. We will be utilizing R to clean our dataset.

To start, we imported the dataset into memory and removed the last 48 columns because they were empty.

```
rm( list=ls() ) # remove all existing objects in the environment
gc() # garbage collection

## read in data ##
setwd("D:/CSUF/2023 Fall/ISDS 577/Final Project")

dat = read.csv('airline.csv', head=T, stringsAsFactors=F, na.strings='') # 'read.csv' reads in csv file

dim(dat) # 'dim' gives the dimensions of the data: # of rows, # of columns
str(dat)

# remove the last 48 columns because they're empty
dat1 <- dat[, -c((ncol(dat) - 47):ncol(dat))]
# check
dim(dat1)
```

Figure 0.1: R code for importing the dataset, check dimension and remove columns.

We decided to only use the most recent 10 years (2010-2020) for the most relevant data because a lot has changed since the 1900s. It will also cut our data in half to help with limited computing power.

```
# remove years before 2010
dat2 <- dat1[dat1$Year >= 2010, ]
# check
dim(dat2)
```

Figure 0.2: R code for removing the rows that are before 2010.

After this step, we're left with 663,197 observations and 61 variables. Then, we remove any preliminary multicollinearity variables, such as OriginStateFips and OriginStateName, because it's very clear that these variables will be 100% correlated.

```
# remove preliminary multicollinearity variables
library(dplyr)
dat2 <- dat2 %>% select(-OriginAirportSeqID, -OriginCityMarketID, -OriginCityName,
-OriginStateFips,- OriginStateName, -OriginWac, -DestAirportSeqID, -DestCityMarketID,
-DestCityName, -DestStateFips, -DestStateName, -DestWac, -WheelsOff, -WheelsOn,
-Diverted, -AirTime, -Flights)
```

Figure 0.3: R code for removing the preliminary multicollinearity variables.

There are many people in our group who are using the dataset in different applications. Some applications treat the null value differently. To avoid this issue, we're converting some null values to similar non-null values. For example, in the original dataset, if the flight is canceled, it is classified as 1; However, if it's not canceled, it's empty in this column. Therefore, we're converting that empty value to "0" to prevent incorrect readings from different applications. We are also changing the empty columns

of the delay reasons to “Not Av” for the same reason. This will further ensure that we do not accidentally remove these values in the next step.

```
# replace null cancellation code for non-null value
dat2$CancellationCode <- ifelse(is.na(dat2$CancellationCode), '0', dat2$CancellationCode)

# # replace null reason for delay for non- null value
dat2$CarrierDelay     <- ifelse(is.na(dat2$CarrierDelay), 'Not Av', dat2$CarrierDelay)
dat2$WeatherDelay     <- ifelse(is.na(dat2$WeatherDelay), 'Not Av', dat2$WeatherDelay)
dat2$NASDelay         <- ifelse(is.na(dat2$NASDelay), 'Not Av', dat2$NASDelay)
dat2$SecurityDelay    <- ifelse(is.na(dat2$SecurityDelay), 'Not Av', dat2$SecurityDelay)
dat2$LateAircraftDelay<- ifelse(is.na(dat2$LateAircraftDelay), 'Not Av',
dat2$LateAircraftDelay)
z
```

Figure 0.4: R code for replacing null values with non-null values.

Subsequently, we check for any missing data and plot it out using the *plot()* function. We can see that all of the missing data is less than 2% of their corresponding variables.

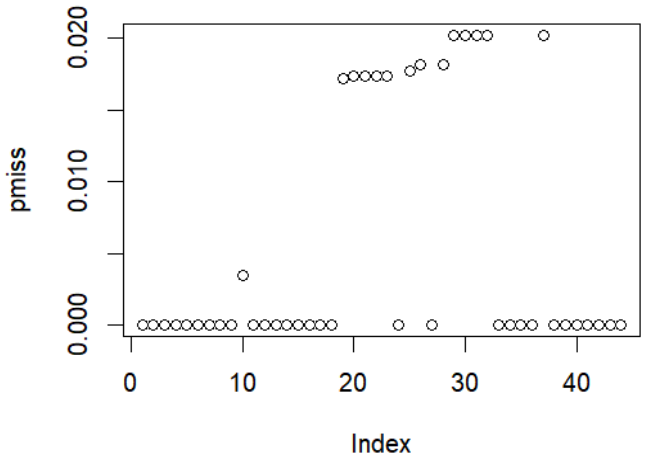


Figure 0.5: Missing values in columns plot chart.

We then print out the names of those columns to further investigate. Those columns are Deptime, Depdelay, DepdelayMinutes, DepDel15, DepartureGroups, TaxiOut, TaxiIn, ArrTime, ArrDelay, ArrDelayMinutes, ArrDel15, ArrivalDelayGroups, Tail_Number, and ActualElapsedTime. Almost all of those are departure and arrival information. They are also very similar in percentage of missing, between around 0.017 and 0.020 percent. Upon further investigation, we found out that those missing values were due to canceled flights. When a flight is canceled, that information is not collected.

```
# check missing ##
matrix.na = is.na(dat2)
pmiss = colMeans(matrix.na) # proportion of missing for each column
nmiss = rowMeans(matrix.na) # proportion of missing for each row
plot(pmiss) # a few columns with high proportion of missing. we want to exclude them.

print(round(pmiss, digits = 2))
```

Figure 0.5: R code for plotting missing values.

Because the ratio of missing values is extremely low, and we might find some interesting things about canceled flights, we decided to leave the canceled flight rows in the dataset. Finally, we organized everything chronologically and exported the *.csv* file to every member for their use. This effectively reduces our data set to 663,197 observations and 44 variables

```
# sort by date
dat2 <- dat2[order(dat2$FlightDate), ]

# exporting the file
write.csv(dat2,'D:/CSUF/2023 Fall/ISDS 577/Final Project/airline_cleaned_csv.csv', row.names
= FALSE)
```

Figure 0.6: R code for sorting and export the dataset.

Data Analysis

Research Question #1: Which day of the week tends to have more delays or cancellations? What are the reasons? How does the on-time performance change across different months or seasons?

Answering this research question will provide airline management with insights into the patterns of delays and cancellations across different days and seasons, which will assist them in optimizing their strategies and operations, ultimately enhancing customer satisfaction. This information can guide resource allocation, such as crew scheduling and maintenance, and inform strategic decisions like flight frequency adjustments and route planning. Additionally, customers will also benefit by understanding which days of the week are more prone to delays or cancellations, as well as the seasonal variations in on-time performance. This information will be crucial for informed travel planning, enabling them to select travel days with historically better punctuality, set realistic expectations for their journeys, and make decisions that will minimize inconvenience.

To answer this research question, we will be using Tableau to generate interesting data visualizations. To statistically confirm the existence of significant differences in delays across various days of the week, we will employ Welch's t-test (Using R – studio). This statistical method is particularly suited for comparing means from two samples that may have unequal variances and sample sizes, making it an ideal choice for rigorously assessing the variations in delay patterns on different days. This method will enable us to ascertain whether the observed variations in data are significant beyond mere chance.

ANALYSIS:

To gain a comprehensive understanding, the analysis will proceed as follows:

1. Monthly delay analysis:
    a) Delayed Flight count
    b) Average Delay durations
    c) Cancellations
2. Daily Analysis:
    a) Delayed flight count
    b) Average Delay durations
    c) Cancellations
3. Delay Reasons
4. Statistical test result

Monthly / Seasonal Analysis (2010 - 2020) :

Throughout the duration of this project, we will categorize airline delays into two primary classifications: delays at departure and delays at arrival. Additionally, we will consider a flight as delayed if it is delayed by 15 minutes or more. Let us begin with an examination of the total number of flights for each month over a period of 2010 to 2020.

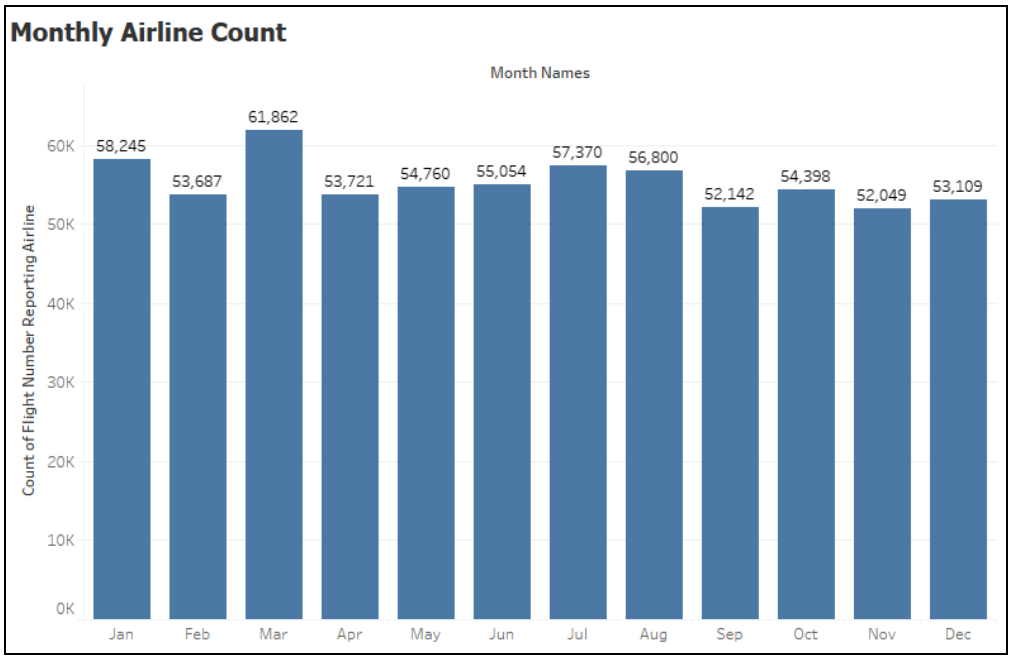


Figure 1.2: Monthly Flight Count (2010 - 2020)

Figure 1.2 indicates that the months of July and August, followed by March and June, experience the highest number of flights, while February has the fewest, with September and November also showing lower figures. The reason behind more frequency of flights during the summer months could be explained by two reasons. Firstly, the summer months typically offer more favorable weather conditions for travel. Secondly, this period coincides with the longest school holidays for children, making it an opportune time for families to schedule vacations.

Moving on, let's look at fig 1.3 and 1.4 illustrating the Monthly count of delayed flights at departure and arrival.

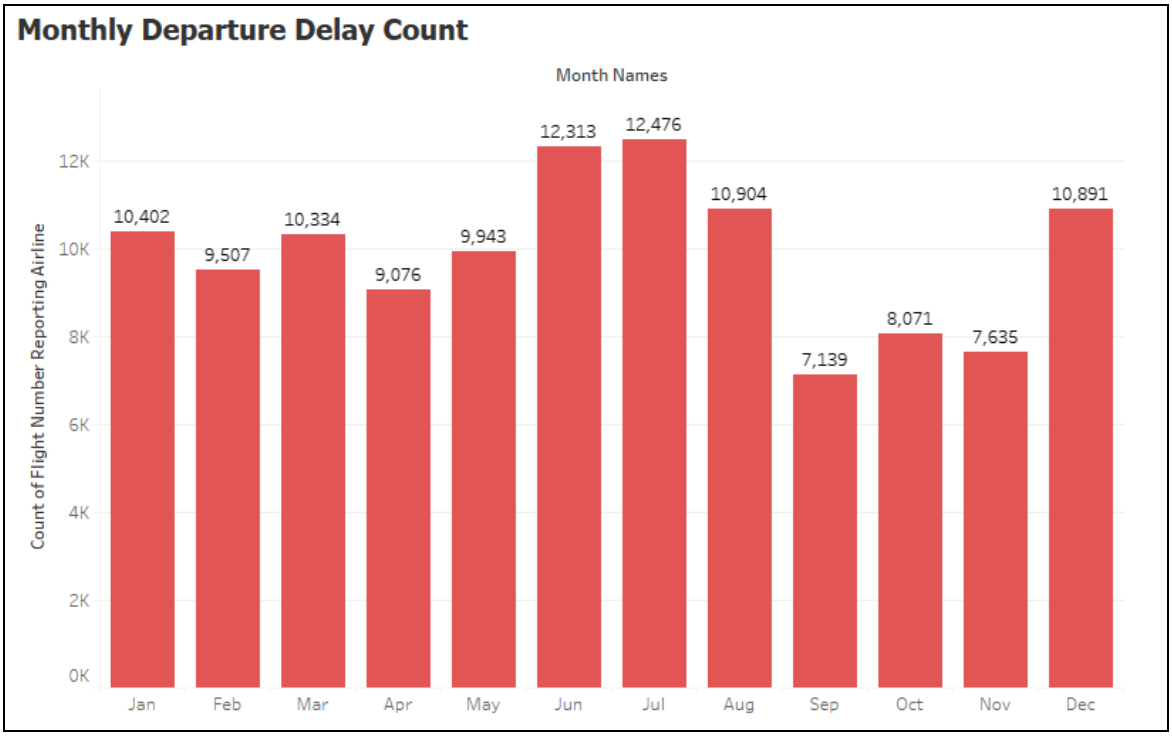


Figure 1.3: Monthly Departure Delay Count (2010 - 2020)

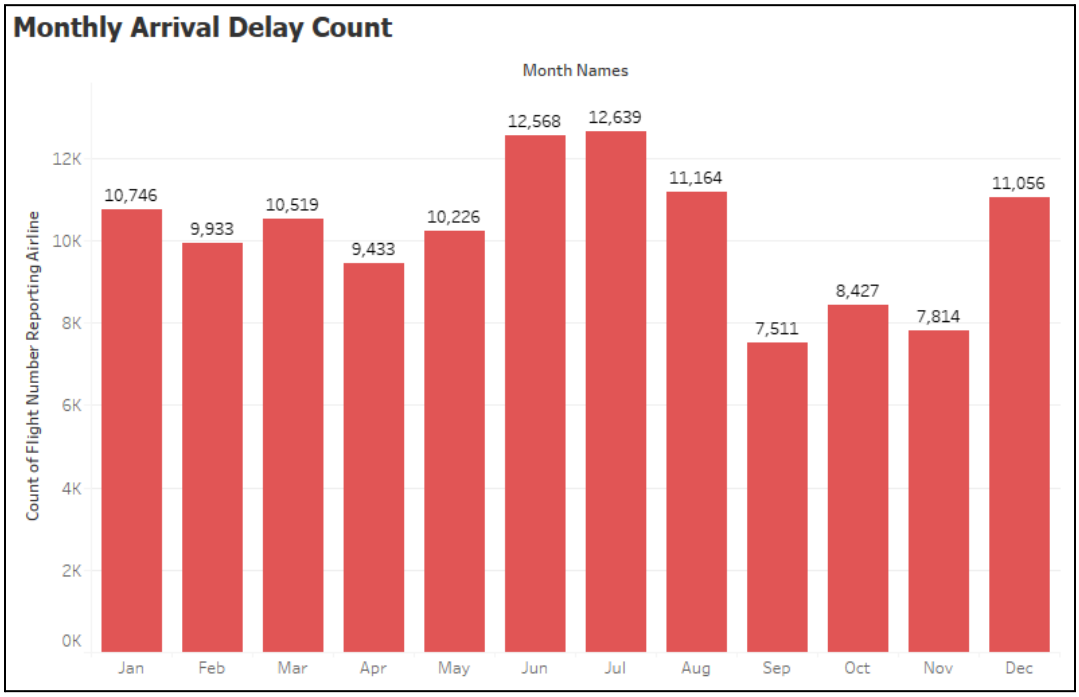


Figure 1.4: Monthly Arrival Delay Flight Count (2010 - 2019)

A detailed analysis of the two bar charts reveals a consistent pattern, with the peak in delayed flights occurring during the months of June, July, August, and December. Conversely, the period encompassing September, October, and November typically registers the fewest delays, characterizing the fall months as a period of lower delay frequencies in flights. Peak travel periods in June, July, August, and December due to school holidays and festive seasons lead to increased air traffic, which can cause congestion-related delays. Additionally, the operational capacity of airlines is often stretched during these high-demand periods, increasing the likelihood of delays. The fall months on the other hand are considered off-season for travel in the U.S. as they coincide with the academic year when families are less likely to vacation, the weather becomes cooler and less predictable, and many travelers are holding off for the winter holiday season. Furthermore, it can be observed that the frequency of delayed flights at arrival is higher than at departure. This phenomenon may stem from various factors. For instance, delays at departure almost invariably lead to subsequent delays at arrival, a subject that will be delved into in future research questions. Additionally, flights are prone to air traffic congestion en route, particularly when headed towards major hub airports. Moreover, delays can compound over time; flights that arrive late often disrupt the schedule of later flights, resulting in a cascading effect of delays throughout the day.

Moving beyond the frequency of airline delays, let us shift our focus on the exploration of the average duration of delays across each month, providing a more comprehensive view of the impact on travel times. For this analysis we will again only consider delay durations of flights that were delayed for 15 minutes or more.

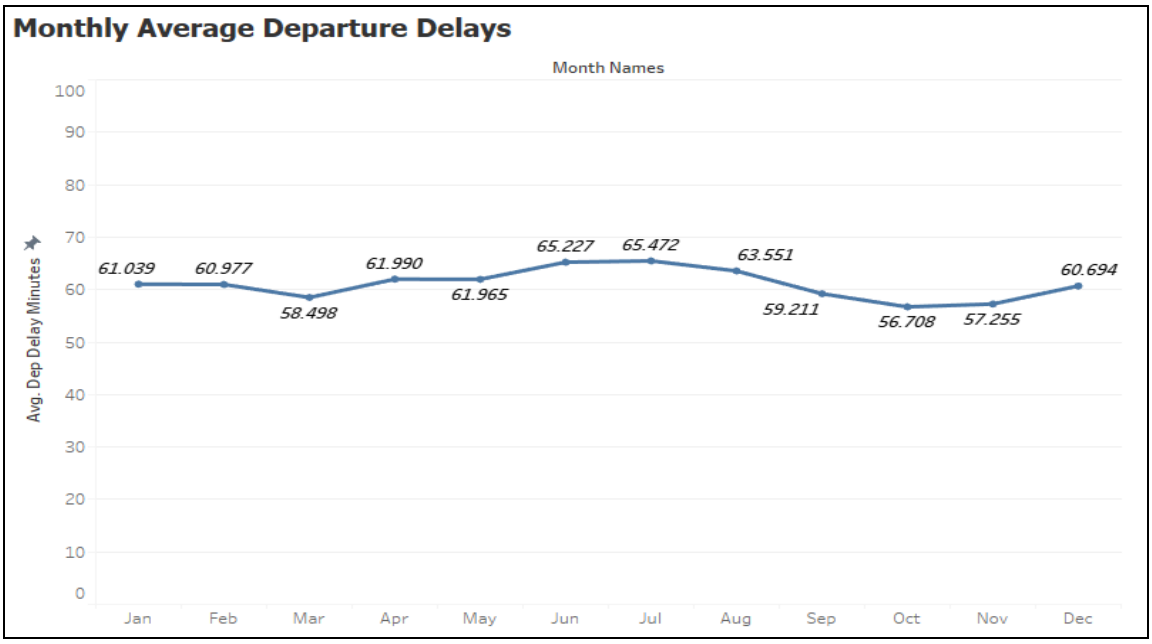


Figure 1.5: Monthly Average Departure Delays (2010 - 2020)

Upon studying figures 1.5 and 1.6 we find that the average delay durations at departure and arrival follow a similar trend, recording the longest delays in the summer months of June, July, and August owing to the high demand and tight airline schedules during those months.

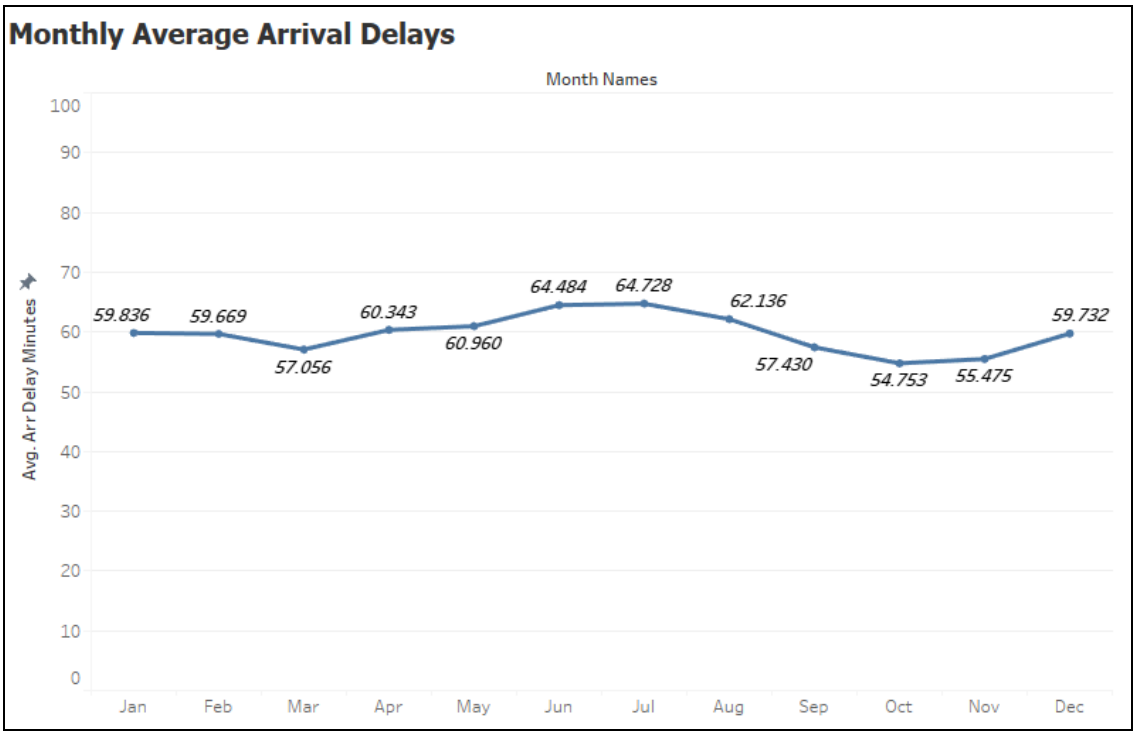


Figure 1.6: Monthly Average Arrival Delays (2010 - 2020)

Additionally, it was intriguing to see that average delay durations decrease during the fall months especially given the often challenging weather conditions characteristic of this period. This trend could be due to various factors, such as less air traffic and hence reduced congestion at airports, which tends to offset the impact of poor weather. Moreover, airlines may improve their operational efficiency and contingency planning after the busy summer season, leading to better management of delays. Additionally, fall typically sees fewer vacation travelers, which can lead to more on-time flights despite the weather. These factors combined may contribute to shorter delay durations during these months.

Next, the count of canceled flights across different months of the year was assessed. Figure 1.7 indicates that the beginning of the year experiences the most cancellations, with January and February being the most affected. This trend is likely due to adverse winter weather conditions during those months. Conversely, November experiences the fewest cancellations. The summer months and December witness a moderate increase in cancellations, potentially due to the high volume of travel and the operational limitations imposed by air traffic control during these peak periods.

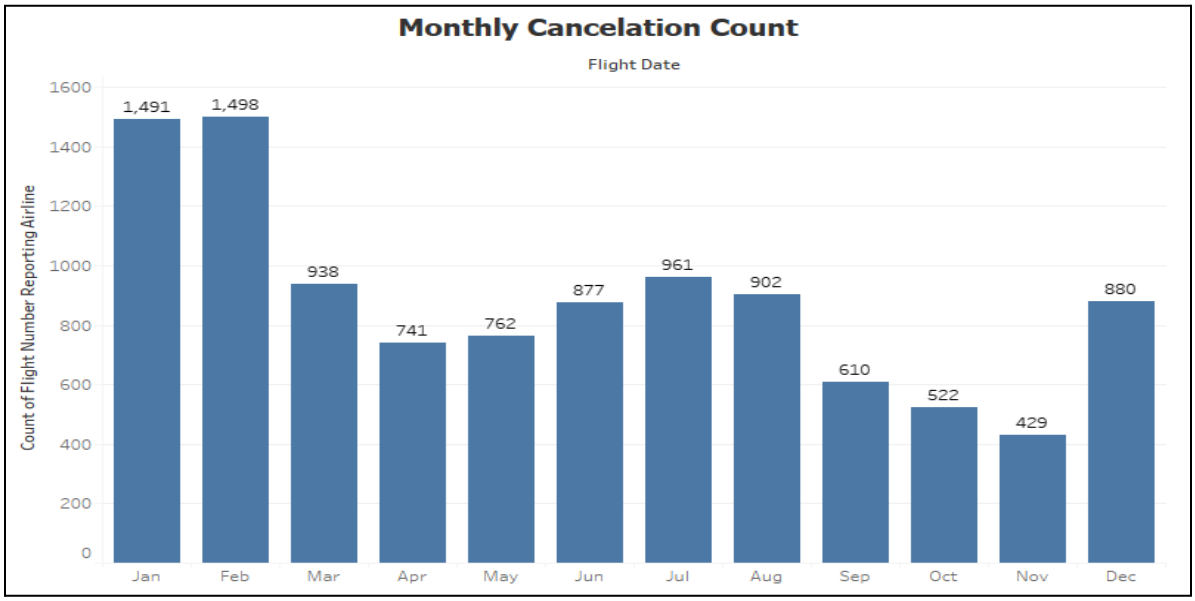


Figure 1.7: Monthly Cancellation Count (2010 - 2019)

Daily Analysis (2010 - 2020):

To begin the analysis, let us look at the bar chart in Figure 1.8 representing the flight count for each day of the week for a 10-year period (2010 - 2020).

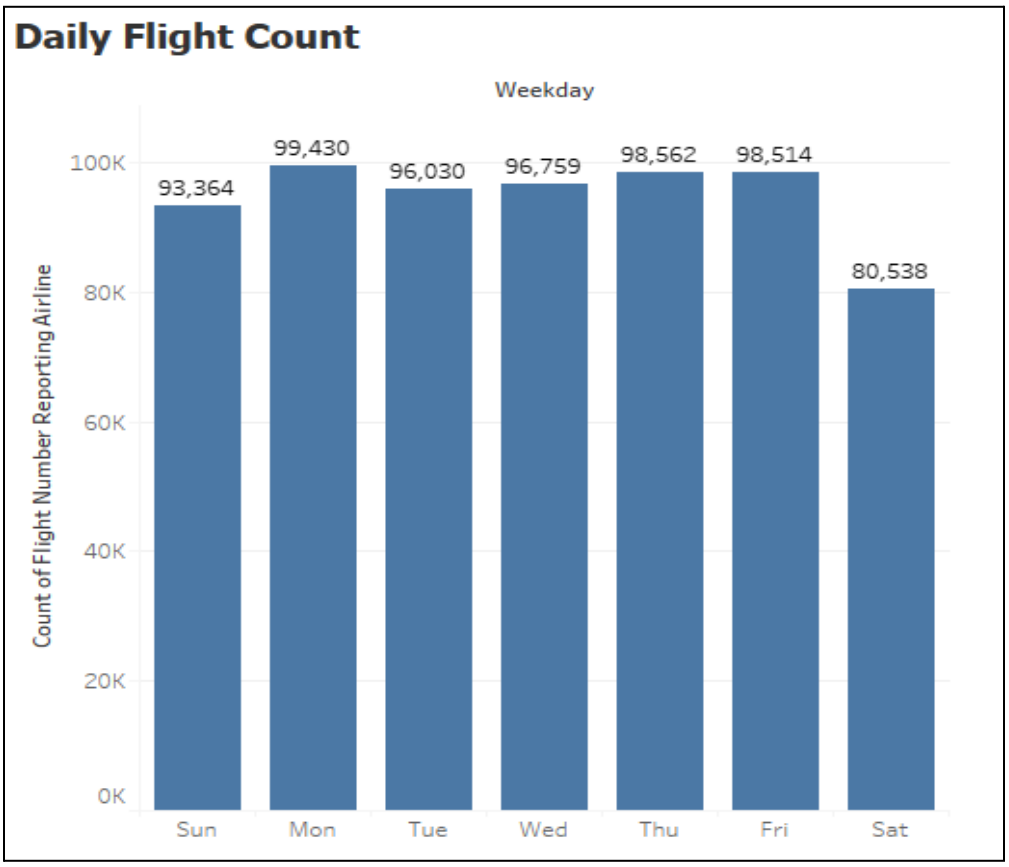


Figure 1.8: Daily Flight Count (2010 - 2020)

The analysis of Figure 1.8 reveals that workdays experience a higher flight count compared to weekends. Mondays see the peak with 99,430 flights, a figure that can be attributed to a surge in business-related travel as professionals commence their workweek. The trend continues with Thursdays and Fridays showing only slightly fewer flights; these days cater to business travelers concluding their trips and leisure travelers embarking on weekend getaways. Tuesdays and Wednesdays experience a relative

decrease in flight numbers, possibly due to a lower demand for midweek travel. Saturdays register the lowest flight activity, reflecting the traditional drop in both business and leisure travel during the weekend.

Now looking at the count of delayed flights at departure and arrival from fig 1.9 and 1.10 we notice that Thursday experiences the highest number of delays, followed by Friday and then Monday.

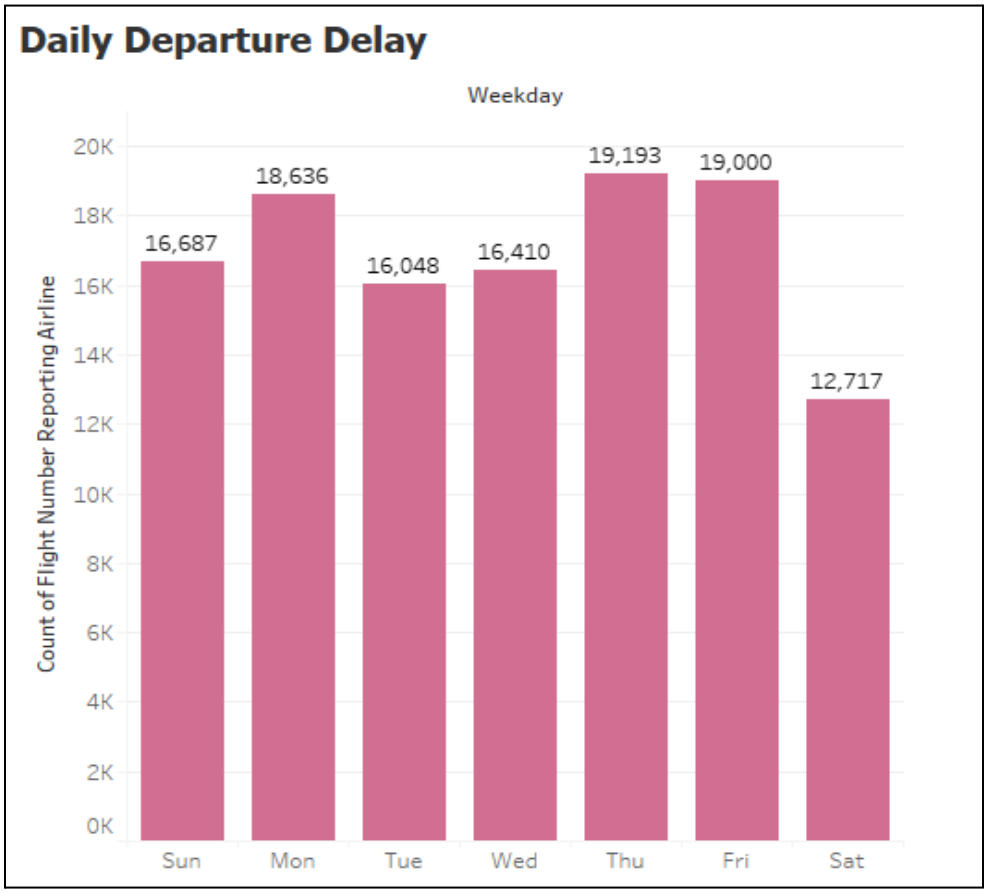


Figure 1.9: Daily Departure Delay Count (2010 - 2020)

Notably, Monday, while having a greater frequency of flights, exhibits a relatively better departure delay performance with only 18.742% of its flights delayed. This compares favorably to Friday's delay rate of 19.286% and Thursday's 19.47%.

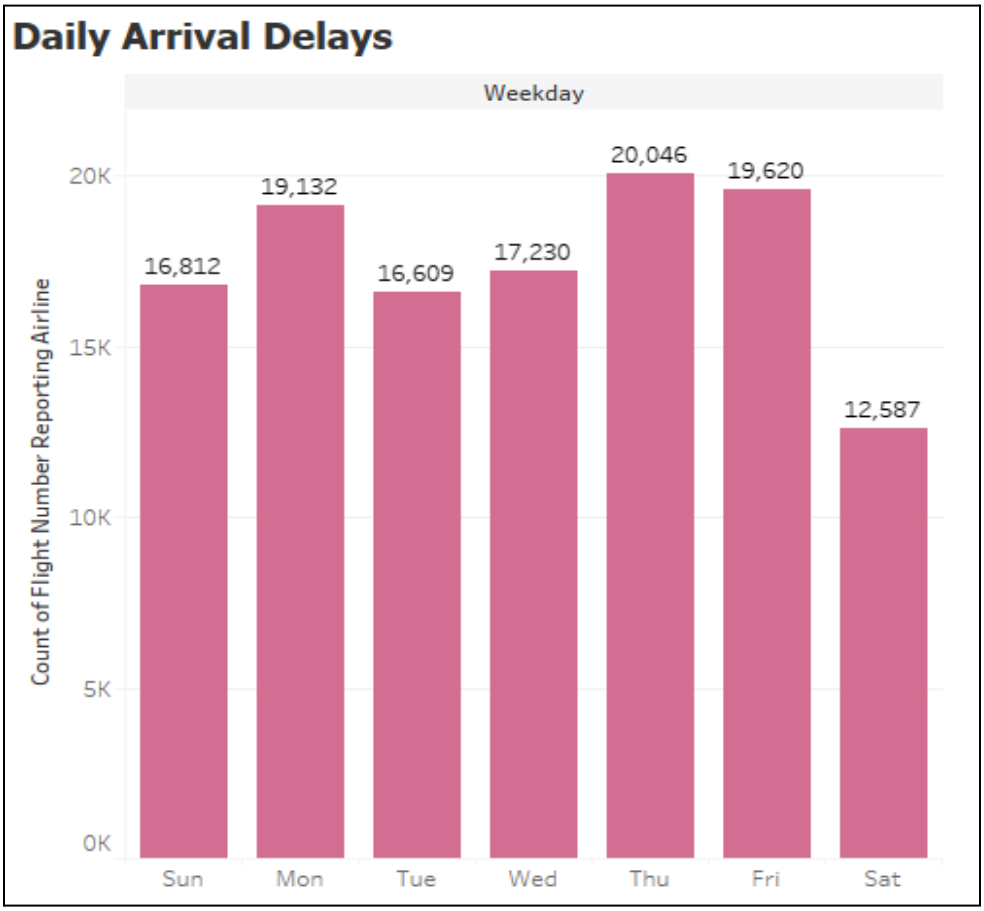


Figure 1.10: Daily Arrival Delay Count (2010 - 2020)

A Similar trend is observed for Arrival delay performance as well, with 19.24% delayed on Mondays, 19.93% on Fridays, and 20.046% on Thursdays. This data suggests that despite the busier schedule, Monday's flight operations are relatively more efficient than those on Thursdays and Fridays. Interestingly, Sunday, despite its lower flight frequency compared to Tuesday and Wednesday, experiences more departure delays. For arrival delays, Wednesday leads, followed by Sunday, then Tuesday. This trend highlights Tuesday as the most efficient travel day during the week,

with fewer delays. Conversely, Saturday, the day with the least overall delays, benefits from a lower frequency of flights.

Next, we will move further to assess the average delay duration for each day of the week. Figure 1.11 shows that Mondays experience the longest average departure delays, approximately 62 minutes. Thursdays and Fridays follow closely, with average delays of around 61.934 and 61.893 minutes respectively.

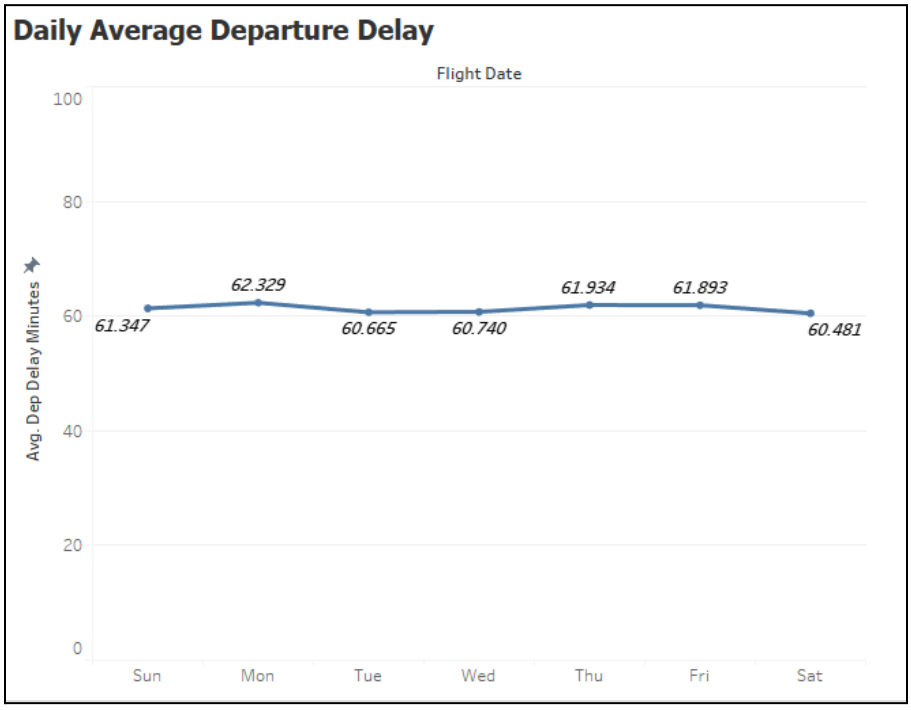


Figure 1.11: Daily Average Departure Delay (2010 - 2020)

Given that these three are the busiest work days, the delay durations are expected due to a combination of increased travel demand and the complexities of managing numerous operational elements. The high demand on these days may lead to congestion both in the air and on the ground. As was observed for the count of

departure delays, Sundays can be seen to have surpassed Tuesdays and Wednesdays in average departure delays. The longer delays on Sundays can be attributed to a significant number of travelers heading home, resulting in a surge of flight activity. This increase in flight volume typically leads to extended delay durations. Lastly, it can be seen that Tuesdays, Wednesdays, and Saturdays, which are the least busy days, exhibit similar patterns in departure delay durations. This consistency indicates a stable trend of delay durations across these days.

Figure 1.12 shows that Mondays experience the longest average arrival delays, approximately 61.223 minutes, similar to the pattern observed in departure delays. However, this time, Sunday surprisingly has surpassed the typically busier days of Thursdays and Fridays by a slight margin, with an average delay duration of 60.677 minutes. Additionally, Saturday's average delay duration surpasses those of the midweek days, Tuesday and Wednesday. This information suggests that weekends experience longer arrival delays in comparison to weekdays.

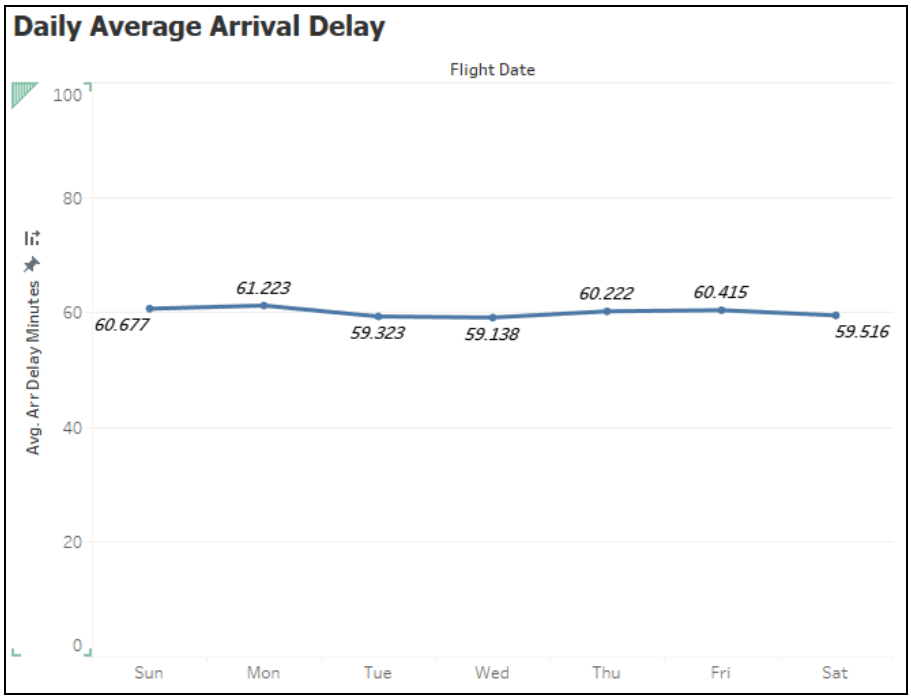


Figure 1.12: Daily Average Arrival Delay (2010 - 2020)

Lastly, the count of canceled flights for each day of the week was assessed.

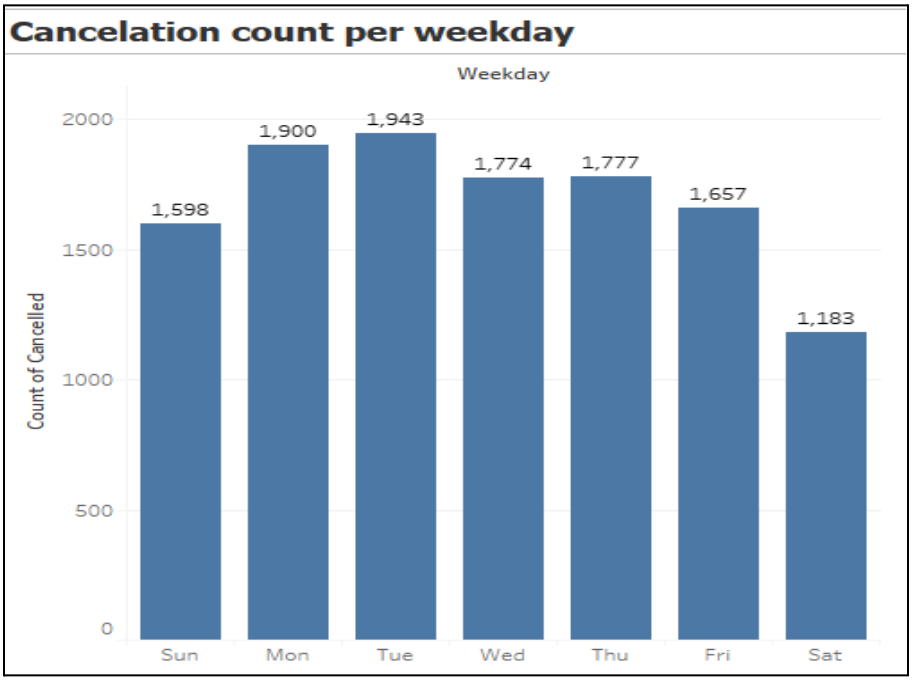


Figure 1.13: Daily Cancellation Count

From Figure 1.13 we observe that weekdays have the most flight cancellations, with the highest observed on a Tuesday followed by Monday. Explaining the reasons behind these cancellations would be difficult as our dataset did not have data related to it. However, these flight cancellations could be due to a mix of controllable and uncontrollable factors. Staffing shortages of pilots and mechanics, along with unaddressed mechanical maintenance issues, significantly contribute to flight cancellations. Additionally, weather-related disruptions, such as storms and snow, frequently cause cancellations, adding to the operational complexity and impacting the reliability of airline services.

Delay Reasons:

Next the most interesting aspect of delay reasons was explored. The reasons were analyzed separately for departure and arrival delays.

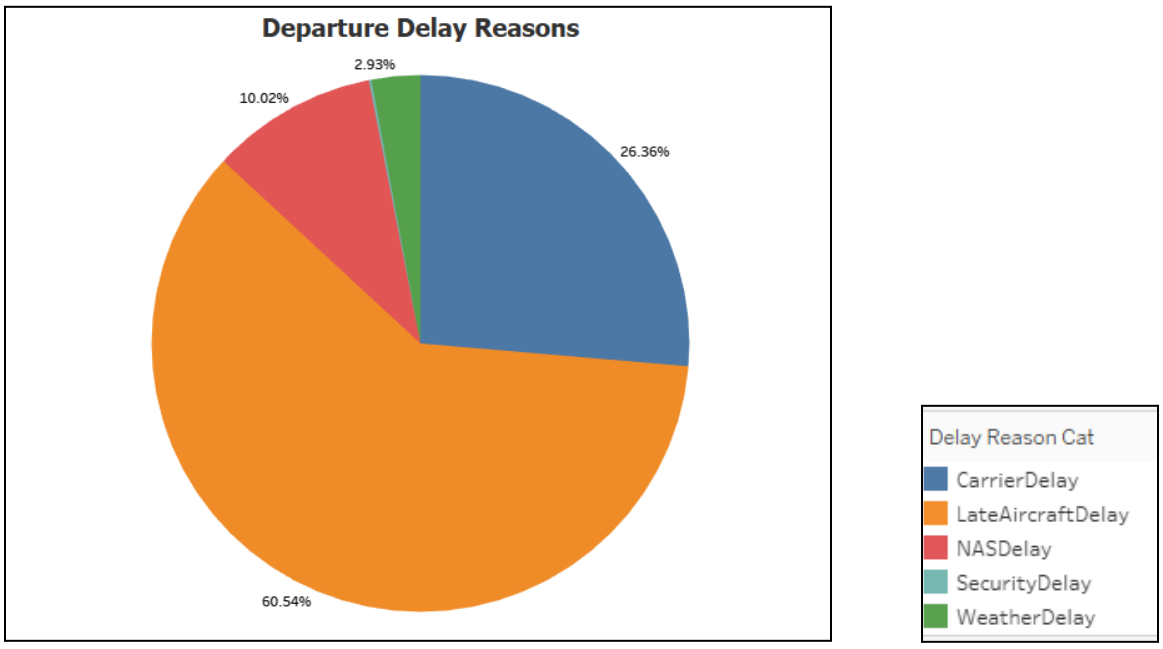


Figure 1.14: Departure Delay Reasons

Figure 1.14 shows a pie chart that categorizes the causes of flight delays into five segments. The largest segment, representing 60.54% of delays, is due to "Late Aircraft Delay," suggesting that previous flights' delays are the primary contributor to subsequent departures being delayed. The next significant cause, at 26.36%, is "Carrier Delay," which could involve factors within the airline's control, such as maintenance or crew problems. "NAS Delay," which stands for National Aviation System and can include weather, heavy traffic, or air traffic control issues, accounts for 10.02% of delays. "Weather Delay," specifically attributed to adverse weather conditions, constitutes 2.93% of the delays. Finally, "Security Delay," likely related to security-related disruptions, makes up the smallest portion at just 0.15% of delays. This distribution highlights that the majority of delays are due to operational and system-wide issues rather than security or weather.

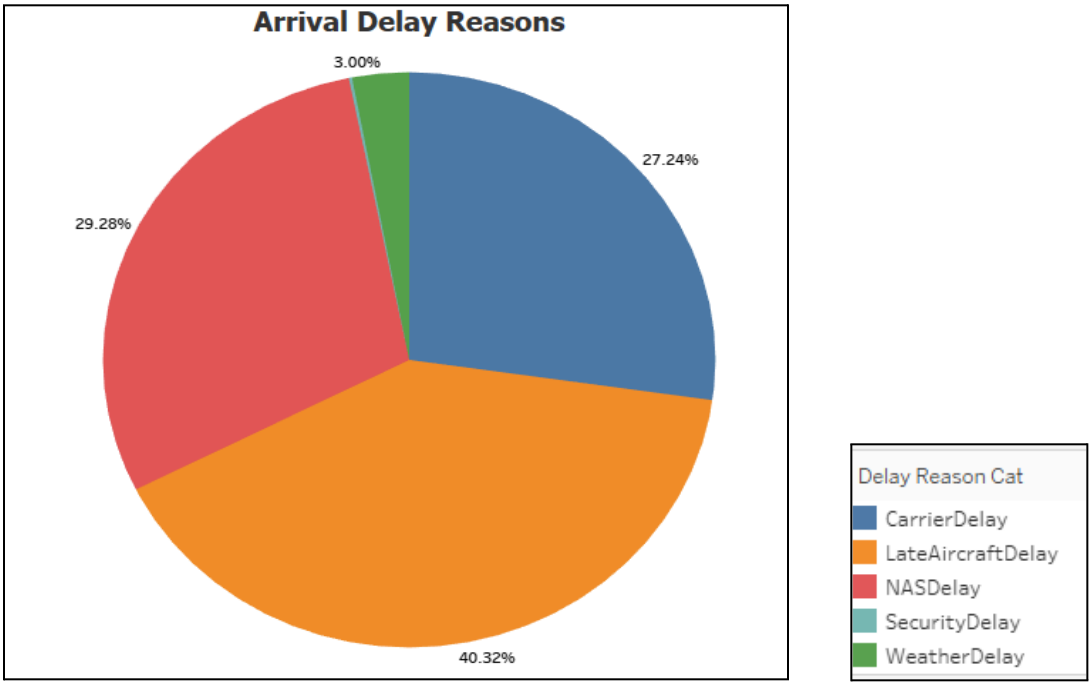


Figure 1.15: Arrival Delay Reasons

Figure 1.15, which presents a pie chart of arrival delay reasons, indicates that 'Late Aircraft Delay' is the most significant factor, accounting for 40.32% of delays. This suggests that issues with prior flights are a major contributor to subsequent arrival delays. The second leading cause is the 'NAS Delay' at 29.28%, which encompasses factors such as air traffic control and widespread system delays. 'Carrier Delay' is the third most common cause at 27.24%, involving delays within the airline's control. Weather-related delays are comparatively less significant at 3%, and security reasons contribute the least at 0.17%.

Statistical Results:

Statistical tests are crucial for making sure research findings are solid before they're shared with the public and those who make policies. They work as a detailed check to confirm that the evidence isn't just due to chance. This way, we can be confident about the research outcomes and know they're not just random guesses. As part of this research question, we conducted a similar statistical test called Welch's t-test to confirm the results for the average delay durations at departure and arrival for different days of the week. This test is used to compare two groups and see if they are different from each other in some way – typically in their average values. For this research question, the way we divided the groups was one day of the week in comparison to the rest of the days, and essentially, we made seven such comparisons for each day of the week. For example, Monday with the rest of the days of the week, Tuesday with the rest of the days of the week, etc.

Departure Delay Duration Results:

```
Welch Two Sample t-test (Monday with rest of the days)
data:  monday_group and rest_of_days_group
t = 2.3355, df = 39292, p-value = 0.01952
mean of x mean of y
35.51873  34.65097


Welch Two Sample t-test (Tuesday with rest of the days)
data:  tuesday_group and rest_of_days_group
t = -1.6472, df = 33126, p-value = 0.09952
mean of x mean of y
34.22714  34.87359


Welch Two Sample t-test (Wednesday with rest of the days)
data:  wednesday_group and rest_of_days_group
t = -0.70342, df = 34620, p-value = 0.4818
mean of x mean of y
34.55593  34.82247


Welch Two Sample t-test (Thursday with rest of the days)
data:  thursday_group and rest_of_days_group
t = 2.1971, df = 41033, p-value = 0.02802
mean of x mean of y
35.45547  34.65942


Welch Two Sample t-test (Friday with rest of the days)
data:  friday_group and rest_of_days_group
t = 1.4564, df = 40386, p-value = 0.1453
mean of x mean of y
35.23875  34.69993


Welch Two Sample t-test (Saturday with rest of the days)
data:  saturday_group and rest_of_days_group
t = -3.913, df = 26049, p-value = 9.139e-05
mean of x mean of y
33.31398  34.97196
```

```
Welch Two Sample t-test (Sunday with the rest of the days)
data:  sunday_group and rest_of_days_group
t = -0.39038, df = 34351, p-value = 0.6963
mean of x mean of y
34.65507  34.80678
```

The results from the Welch Two Sample t-tests comparing each day of the week with the rest for departure delays can be interpreted as follows:

Monday: With a t-value of 2.3355 and a p-value of 0.01952, there is a statistically significant difference in departure delays for Mondays compared to the rest of the days. The average delay on Monday (35.51873) is higher than the average for the rest of the days (34.65097).

Tuesday: The t-value is -1.6472 and the p-value is 0.09952, which is not statistically significant under the typical 0.05 threshold. The average delay on Tuesday (34.22714) is slightly lower than the average for the rest of the days (34.87359).

Wednesday: With a t-value of -0.70342 and a p-value of 0.4818, there's no significant difference in delays. The average delay on Wednesday (34.55593) is slightly lower than the rest of the days (34.82247).

Thursday: The t-value is 2.1971 and the p-value is 0.02802, indicating a significant difference. The average delay on Thursday (35.45547) is higher than the average for the rest of the days (34.65942).

Friday: The t-value is 1.4564 and the p-value is 0.1453, which does not indicate a significant difference. The average delay on Friday (35.23875) is slightly higher than the rest of the days (34.69993).

Saturday: With a t-value of -3.913 and a very low p-value (9.139e-05), there is a significant difference. The average delay on Saturday (33.31398) is lower than the average for the rest of the days (34.97196).

Sunday: The t-value is -0.39038 and the p-value is 0.6963, indicating no significant difference. The average delay on Sunday (34.65507) is close to the average for the rest of the days (34.80678).

Based on the mean delay times, the order of days in descending departure delays is: Monday >> Thursday >> Friday >> Sunday >> Wednesday >> Tuesday >> Saturday. These results are identical to the results that we observe from the Tableau visualizations.

Arrival Delay Duration Results:

```
Welch Two Sample t-test ( Monday with the rest of the days)
data:  monday_group_1 and rest_of_days_group_1
t = 4.022, df = 39409, p-value = 5.782e-05
mean of x mean of y
35.63862  34.16698


Welch Two Sample t-test (Tuesday with the rest of the days)
data:  tuesday_group_1 and rest_of_days_group_1
t = -3.2306, df = 34684, p-value = 0.001237
mean of x mean of y
33.34055  34.56199


Welch Two Sample t-test (Wednesday with the rest of the days)
data:  wednesday_group_1 and rest_of_days_group_1
t = -2.3134, df = 36812, p-value = 0.02071
mean of x mean of y
33.72501  34.56199
```

```
Welch Two Sample t-test (Thursday with the rest of the days)
data:  thursday_group_1 and rest_of_days_group_1
t = 1.9578, df = 42763, p-value = 0.05027
mean of x mean of y
34.96776  34.28332


Welch Two Sample t-test (Friday with the rest of the days)
data:  friday_group_1 and rest_of_days_group_1
t = 1.4162, df = 41481, p-value = 0.1567
mean of x mean of y
34.82108  34.31173


Welch Two Sample t-test (Saturday with the rest of the days)
data:  saturday_group_1 and rest_of_days_group_1
t = -3.2472, df = 25192, p-value = 0.001167
mean of x mean of y
33.16407  34.54164


Welch Two Sample t-test (Sunday with the rest of the days)
data:  sunday_group and rest_of_days_group
t = -0.39038, df = 34351, p-value = 0.6963
mean of x mean of y
34.65507  34.80678
```

The results from the Welch Two Sample t-tests for arrival delays comparing different days of the week to the rest are as follows:

Monday: The t-value is 4.022, with a very low p-value (5.782e-05), indicating a statistically significant higher mean arrival delay for Mondays (35.63862) compared to the rest of the days (34.16698).

Tuesday: The t-value is -3.2306, with a p-value of 0.001237, showing a significantly lower mean delay for Tuesdays (33.34055) compared to the rest (34.56199).

Wednesday: The t-value is -2.3134, with a p-value of 0.02071, suggesting a significantly lower mean delay for Wednesdays (33.72501) than the rest (34.56199).

Thursday: The t-value is 1.9578, with a p-value of 0.05027, indicating a borderline significantly higher mean delay for Thursdays (34.96776) compared to the rest (34.28332).

Friday: The t-value is 1.4162, with a p-value of 0.1567, which is not statistically significant. The mean delay for Fridays (34.82108) is slightly higher than the rest (34.31173), but the difference is not significant.

Saturday: The t-value is -3.2472, with a p-value of 0.001167, showing a significantly lower mean delay for Saturdays (33.16407) compared to the rest (34.54164).

Sunday: The t-value is -0.39038, with a p-value of 0.6963, indicating no significant difference in mean delay for Sundays (34.65507) compared to the rest (34.80678).

Based on these mean delay times, the order of days in descending arrival delays is:

Monday >> Thursday >> Friday >> Sunday >> Wednesday >> Tuesday >> Saturday. However, these results do not match the results that we observed through the Tableau visualizations. Hence, we can say that this might have happened because the actual difference between groups might be too small to detect, indicating that while there may be a difference, it's not strong enough to be statistically significant.

Conclusion and Recommendations:

The research clearly indicates that airline performance fluctuates with the seasons. Increased travel demand during the summer months and December holidays leads to more flights and, consequently higher delays due to escalated air traffic and operational challenges. Furthermore, the weekly delay analysis reveals interesting patterns: Mondays are the busiest with the most delays, while Saturdays, despite fewer overall delays, tend to have longer average arrival delays than midweek days, underscoring a unique weekend operational dynamic. The causes of delays also vary; late aircraft and carrier delays are predominant in departures, while arrivals are affected by these as well as NAS delays. This analysis highlights the complexities in managing airline operations, influenced significantly by seasonal travel trends and varying weekly schedules.

To mitigate airline delays predominantly caused by Late Aircraft Delay, NAS Delay, and Carrier Delays, a multi-faceted approach is recommended. This includes optimizing aircraft turnaround processes, enhancing coordination with the National Aviation System for better traffic management, and investing in advanced technology for improved communication and infrastructure. Proactive maintenance, efficient crew management, and robust operational planning are key to addressing carrier-related delays. Additionally, improving weather forecasting and response strategies, coupled with effective customer communication and contingency planning, can significantly reduce delays. Training staff to handle peak times and allocating resources wisely further supports these efforts, aiming at overall operational efficiency and reduced delay occurrences.

Research Question #2: Are there certain types of flights (e.g., red-eye flights, early morning departures) that are more likely to experience delays?

After generating data visualizations in Tableau, it is evident there are specific types of flights that are more prone to delays during the day. To gain a comprehensive understanding, the analysis will proceed as follows:

1. Airline-Specific Delays:
    a. Identify the airline with the highest incidence of delays.
    b. Research the causes of the delays with airlines that have the highest incidence through customer reviews.
2. Average Departure and Arrival Delays:
    a. Examine the average departure and arrival delays in minutes. This analysis will provide insights into the temporal aspects of delays.

First, the count of airline departure delays was assessed using three key variables in Tableau: the Reporting Code, Departure Time Block, and Departure Time Delay for flights experiencing delays exceeding 15 minutes. The Reporting Code, also known as the International Air Transport Association (IATA) airline code, is a unique two-digit identifier assigned to each airline. For example, on the Count of Airline Departure Delays heatmap, airline WN displays the deepest shade of blue. Airline WN represents Southwest which indicates that it has the highest occurrence of delays exceeding 15 minutes. A closer examination of the time blocks showed that Southwest (WN) typically experienced 9,455 delays during the military time range of 6:00 to 6:59, leading to delayed arrivals. In addition, Southwest recorded the highest count of delays

at 8,405 during the time range of 18:00 – 18:59. A majority of the pronounced shade of blue on the Count of Airline Arrival Delays heatmap are in between later times of the day such as noon and afternoon hours. This shows that most of the airline departure delays and arrival delays are during the afternoon where the counts are the highest. Based on this correlation, delayed departures cause delayed arrivals, which is a critical issue that needs further research.

Count of Airline Depature Delays

	Reporting Airline																				
Dep Time Blk	9E	AA	AS	B6	CO	DL	EV	F9	FL	G4	HA	MQ	NK	OH	OO	UA	US	VX	WN	XE	YV	YX
0001-0559	174	1,784	431	710	110	1,646	882	295	63		268	405	221	276	1,538	1,203	506	6	2,208	98	199	186
0600-0659	811	4,808	1,366	1,845	368	5,909	3,263	867	620	159	296	2,405	627	395	5,418	4,176	1,305	83	9,455	652	590	646
0700-0759	625	5,997	1,555	1,845	384	6,448	2,271	619	671	215	496	2,040	734	568	3,735	3,808	1,962	455	8,305	397	635	389
0800-0859	690	5,133	1,279	1,672	328	6,526	2,776	599	719	124	652	1,946	538	295	4,185	4,086	1,600	258	8,997	491	802	480
0900-0959	729	4,358	1,125	1,606	306	5,381	2,826	420	353	140	611	1,861	488	610	4,010	3,312	1,508	339	7,711	449	727	434
1000-1059	847	4,531	1,144	1,564	259	4,921	3,496	786	872	149	611	2,185	400	361	4,229	3,110	1,208	221	8,673	599	674	385
1100-1159	712	4,518	1,115	1,663	295	6,119	2,901	491	655	104	463	2,059	487	663	4,890	3,154	1,733	241	7,615	478	643	503
1200-1259	843	4,585	987	1,311	331	5,674	3,504	418	681	143	492	2,297	413	450	4,147	3,307	1,070	210	7,676	548	1,015	455
1300-1359	767	4,495	1,089	1,484	300	5,458	2,981	513	539	152	612	1,824	372	578	4,964	3,165	1,457	252	7,908	568	508	413
1400-1459	691	4,702	824	1,563	327	4,864	3,219	631	625	184	619	2,096	357	519	3,583	2,873	1,368	237	7,505	621	872	427
1500-1559	798	4,267	1,084	1,424	321	5,812	2,902	614	588	165	600	2,023	490	456	4,902	2,918	1,119	175	7,535	445	454	437
1600-1659	722	4,162	819	1,414	229	5,203	3,337	567	636	123	454	2,067	476	497	4,083	2,848	1,380	215	7,885	604	903	417
1700-1759	905	4,455	1,351	1,565	379	6,400	3,148	506	635	176	468	1,861	409	582	4,927	3,552	1,607	386	8,429	556	672	526
1800-1859	452	4,742	1,358	1,798	284	3,697	2,288	718	705	136	381	2,110	483	403	3,763	2,800	1,452	275	8,189	452	750	445
1900-1959	700	3,236	1,037	1,601	276	5,573	2,594	646	468	99	325	1,627	498	453	3,461	3,130	1,062	204	8,047	503	702	415
2000-2059	398	3,562	752	1,331	178	3,107	1,637	481	440	93	303	1,541	617	355	3,238	1,803	1,192	148	6,610	241	372	357
2100-2159	301	1,849	681	1,125	176	2,836	1,203	408	370	91	207	762	451	89	1,921	1,607	332	92	4,447	202	224	320
2200-2259	140	1,827	283	665	26	2,429	467	208	312	3	147	392	214	283	1,326	1,280	995	60	1,690	58	309	119
2300-2359	9	551	405	639	41	758	59	136	87	5	47	2	191		201	835	136	104	76	21	14	2

Figure 2.1: Count of Airline Departure Delays

Count of Airline Arrival Delays

	Reporting Airline																				
Arr Time Blk	9E	AA	AS	B6	CO	DL	EV	F9	FL	G4	HA	MQ	NK	OH	OO	UA	US	VX	WN	XE	YV	YX
0001-0559	17	1,961	829	1,644	141	1,666	119	423	132		186	238	486	18	574	2,362	363	72	2,401	13	56	71
0600-0659	104	1,041	320	468	81	1,252	597	166	30	5	325	446	137	194	1,096	926	556	45	802	115	160	73
0700-0759	502	1,645	466	815	100	2,353	1,652	348	284	19	330	1,338	309	183	3,487	1,703	467	110	5,237	305	312	339
0800-0859	515	3,207	922	1,391	168	4,315	2,520	297	311	108	378	1,625	420	518	3,173	2,282	1,331	158	6,302	430	588	327
0900-0959	743	4,251	1,088	1,402	194	4,900	2,594	771	742	157	441	2,018	522	367	4,046	2,848	1,078	171	8,187	401	628	415
1000-1059	588	4,466	1,070	1,753	315	5,776	2,954	488	660	107	509	1,840	514	581	4,126	3,018	1,777	220	7,734	502	701	478
1100-1159	870	4,570	858	1,322	329	5,483	3,138	368	660	133	602	2,367	458	353	3,907	3,137	1,102	164	7,759	539	882	500
1200-1259	793	4,297	1,099	1,629	226	5,264	2,962	468	575	174	603	1,988	395	546	4,792	2,828	1,382	224	7,732	475	594	398
1300-1359	790	4,569	850	1,540	298	5,320	3,267	603	636	182	544	1,973	384	594	4,206	2,679	1,507	204	7,691	561	800	409
1400-1459	785	4,125	1,065	1,475	314	5,767	3,122	644	587	169	468	2,050	450	451	4,825	3,242	1,097	196	7,484	569	558	493
1500-1559	706	4,750	822	1,246	227	4,944	2,839	582	570	130	458	1,991	442	590	3,870	2,799	1,385	260	7,725	424	904	391
1600-1659	877	4,379	1,330	1,415	404	6,418	3,657	535	598	148	484	1,896	371	420	5,083	3,658	1,596	344	8,147	626	588	460
1700-1759	602	4,786	1,107	1,498	321	4,612	2,747	532	663	148	426	2,460	471	577	4,201	2,961	1,218	229	8,197	538	776	440
1800-1859	871	4,242	1,055	1,512	312	6,019	3,421	773	629	127	407	1,851	444	439	4,129	3,527	1,357	243	8,405	584	741	410
1900-1959	510	4,903	937	1,347	414	5,391	2,324	550	574	153	425	1,975	536	586	4,377	2,805	1,715	255	8,164	575	730	457
2000-2059	692	4,531	1,294	1,705	302	5,503	2,827	608	724	169	411	1,635	414	315	3,749	3,926	1,467	269	7,641	427	545	494
2100-2159	627	4,857	1,252	1,741	214	5,646	2,057	449	576	116	485	1,786	444	595	3,688	3,358	1,729	306	7,585	314	700	414
2200-2259	392	3,477	1,302	1,576	325	4,120	1,779	744	630	136	317	1,104	528	190	2,597	2,342	833	363	6,407	368	282	364
2300-2359	315	3,250	970	1,269	220	3,810	1,013	557	432	71	244	786	726	291	2,315	2,405	975	121	5,062	190	480	399

Figure 2.2: Count of Airline Arrival Delays

According to the Bureau of Transportation Statistics from January 2010 to December 2020, the primary causes of delays for Southwest Airlines can be attributed to specific factors. These include air carrier delay (5.57%), aircraft arriving late (8.85%), and national aviation system delay (3.39%). Among these three main causes, an air carrier delay encompasses a range of issues, including aircraft maintenance concerns, crew availability, ground operations, aircraft cleaning and loading, and logistical challenges.

(For clarification, the Bureau of Transportation Statistics classifies a flight as delayed when the aircraft arrives 15 or more minutes behind its scheduled time.)

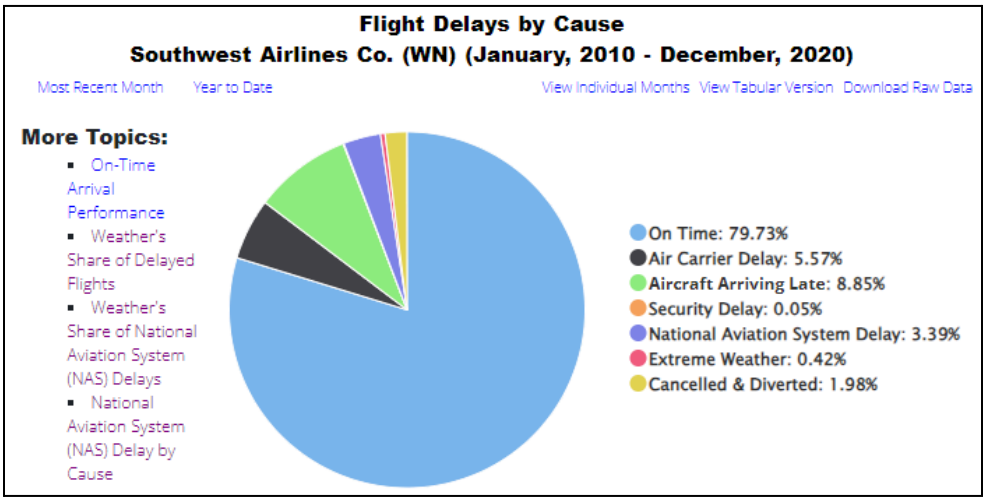


Figure 2.3: Flight Delay Causes for Southwest (2010 - 2020)

Additionally, the National Aviation System (NAS) delay pertains to disruptions within the broader national air transportation network which is not directly caused by a specific airline. This NAS would include weather, volume, equipment, closed runway,

and others. The main factors are weather (51.26%), volume (33.07%), and closed runway (10.61%).

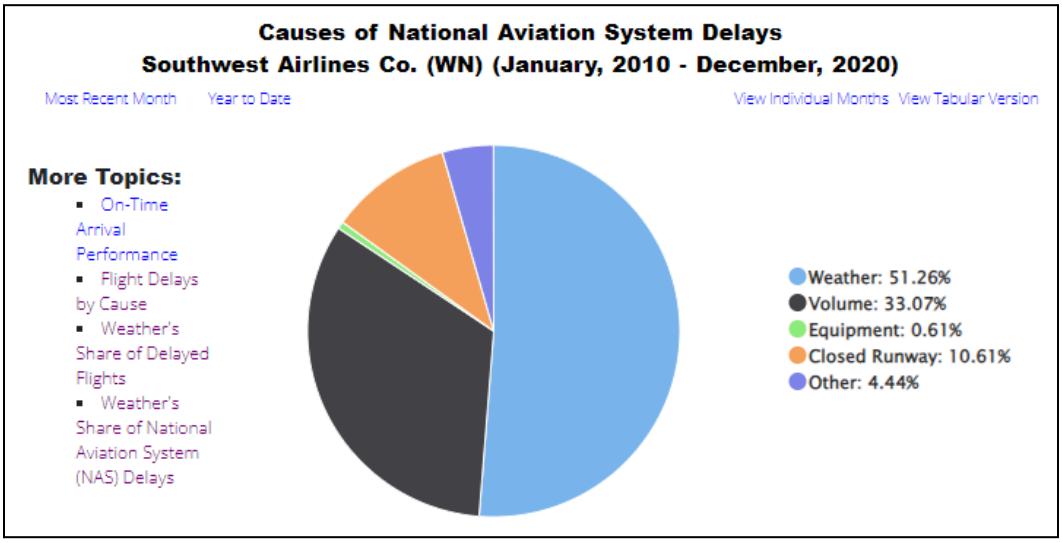


Figure 2.4: Causes of National Aviation System Delays for Southwest (2010 - 2020)

Next, we looked at the average departure and arrival delay in minutes. In the "Average Departure Delay in Minutes" bar chart, a peak emerged during the 17:00 – 17:59 time frame, with an average delay of 16.99 minutes.

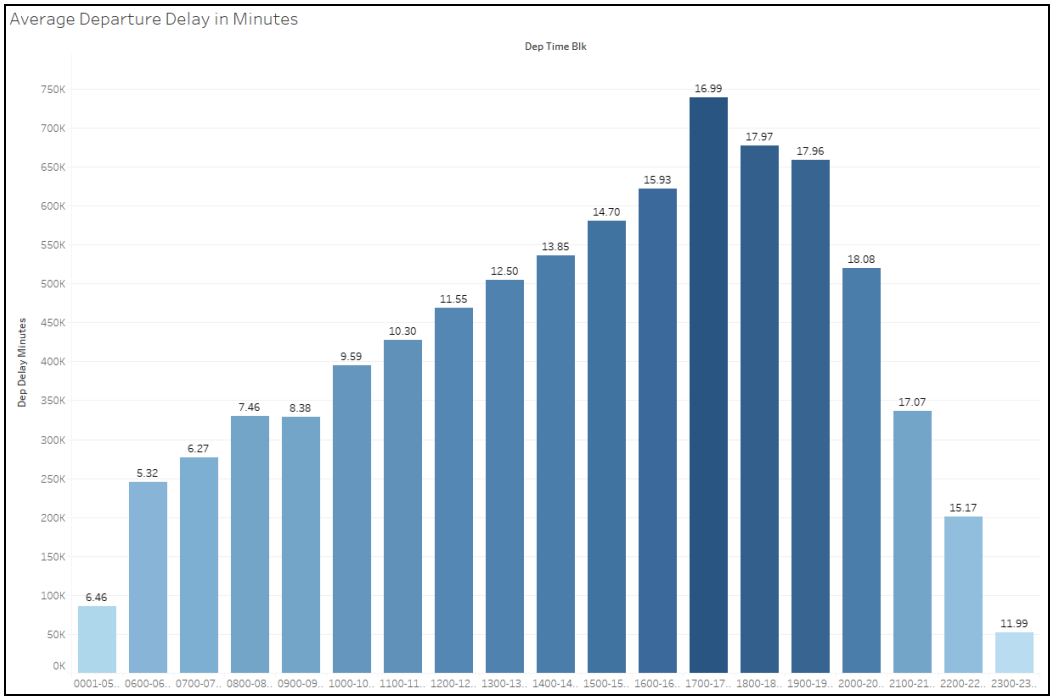


Figure 2.5: Average Departure Delay in Minutes

In addition, examining the "Average Arrival Delay in Minutes," we identified another peak delay period from 20:00 to 20:59 that averages to 17.35 minutes. Conversely, the early morning and late-night hours displayed the lowest delay occurrences. This pattern suggests a higher volume of flights and subsequent delays in the afternoon through evening hours. This observation aligns with our earlier findings, emphasizing the impact of departure delays on subsequent arrival times. Our analysis provides valuable insights into the temporal patterns of delays, highlighting specific timeframes where operational adjustments may yield significant improvements. These findings further the importance of effective scheduling and proactive delay management in optimizing air travel operations.

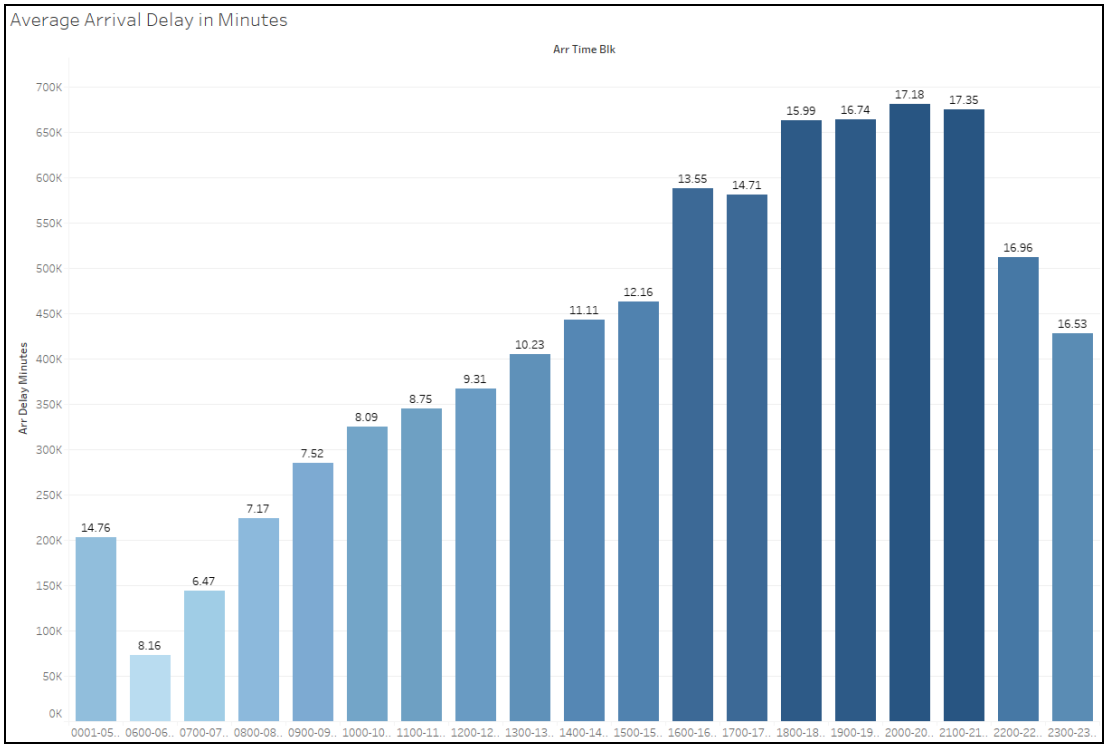


Figure 2.6: Average Arrival Delay in Minutes

Research Question #3: What is the relationship between flight distance and the likelihood of delay? Are there specific routes or flight numbers that are consistently delayed?

To understand the relationship between flight distance and the likelihood of delay, we can perform the following steps:

- Load and inspect the data to understand its structure and content.
- Filter out flights that were not delayed.
- Analyze the relationship between flight distance and delay by calculating the likelihood of delay for different distance groups.
- Visualize the results to better understand the relationship.

Here's the bar chart visualizing the relationship between flight distance groups and the likelihood of delay:

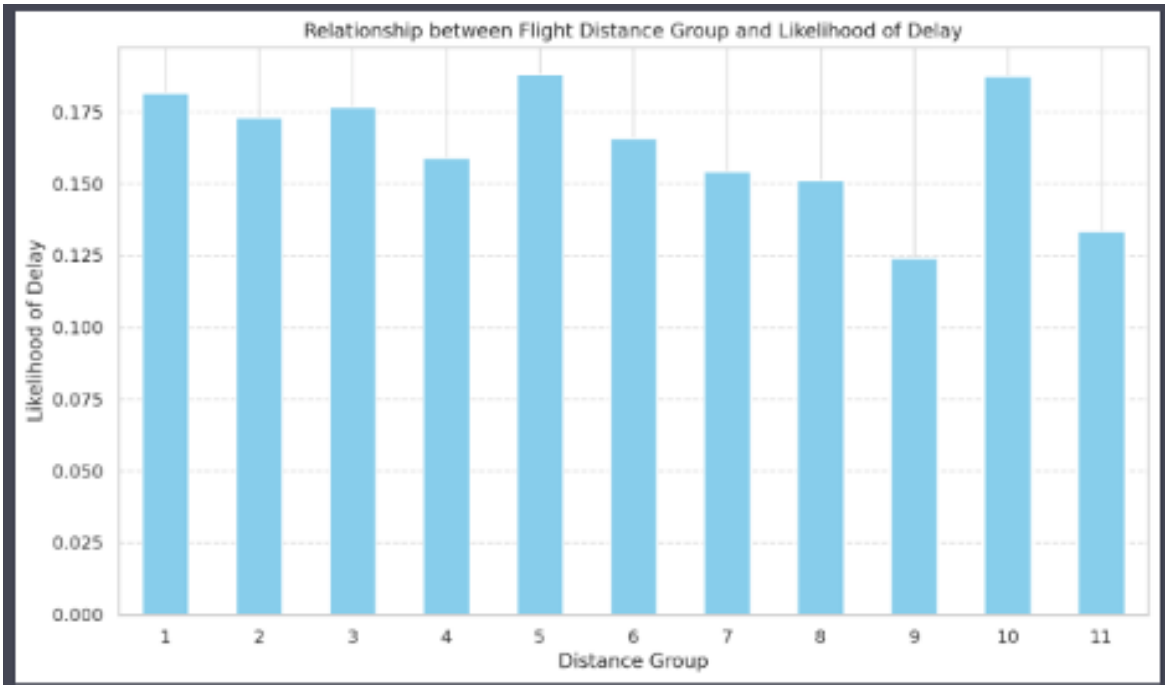


*Figure 3.1: R*elationship between flight distance groups and the likelihood of delay

From the chart, we can observe the following:

1. Flights in distance group 1 (shortest distances) have a relatively high likelihood of delay.
2. The likelihood of delay tends to decrease for the subsequent distance groups, with a noticeable drop in groups 2 and 3.
3. The likelihood of delay then remains relatively steady for the middle distance groups.
4. There's a slight increase in the likelihood of delay for the longest flights (group 9 onwards).

These observations suggest that flight distance does play a role in the likelihood of delay, but other factors may also be at play. For example, shorter flights might be more susceptible to delays due to the higher frequency of takeoffs and landings, while longer flights might have more buffer time to compensate for any initial delays.

The heatmap displays the correlation between different variables. A correlation of 1 indicates a perfect positive relationship, while a correlation of -1 indicates a perfect negative relationship. A correlation close to 0 suggests no linear relationship between the variables.

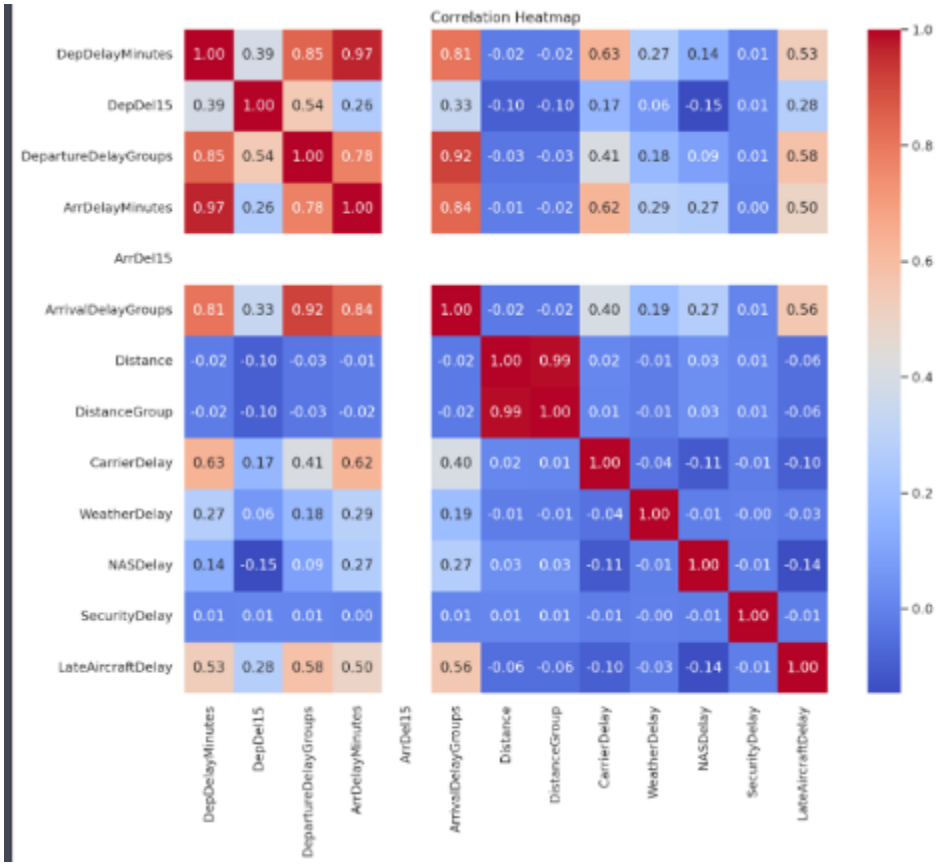


Figure 3.2: Correlation Heatmap

From the heatmap, we can observe the following:

- Flight distance (Distance and DistanceGroup) does not have a strong correlation with departure or arrival delay minutes (DepDelayMinutes, ArrDelayMinutes), or with departure or arrival delay indicators (DepDel15, ArrDel15).
- There is a strong correlation between departure and arrival delay minutes, as well as between their respective indicators.

- The different types of delays (CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay) have weak to moderate correlations with each other and with departure or arrival delays.

This indicates that flight distance does not have a significant impact on the likelihood of a flight being delayed. Other factors, such as weather or carrier-related issues, might have a more significant impact on delays.

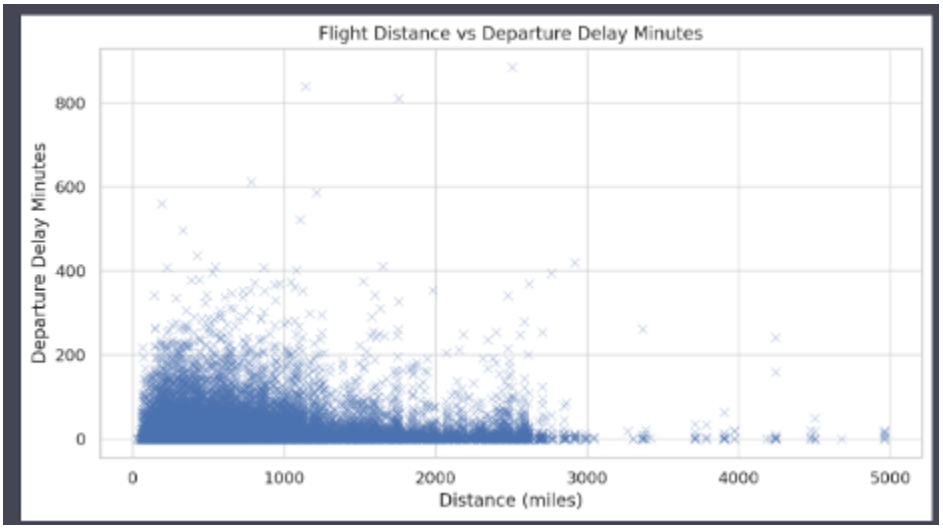


Figure 3.3: Flight Distance vs Departure Delay Minutes

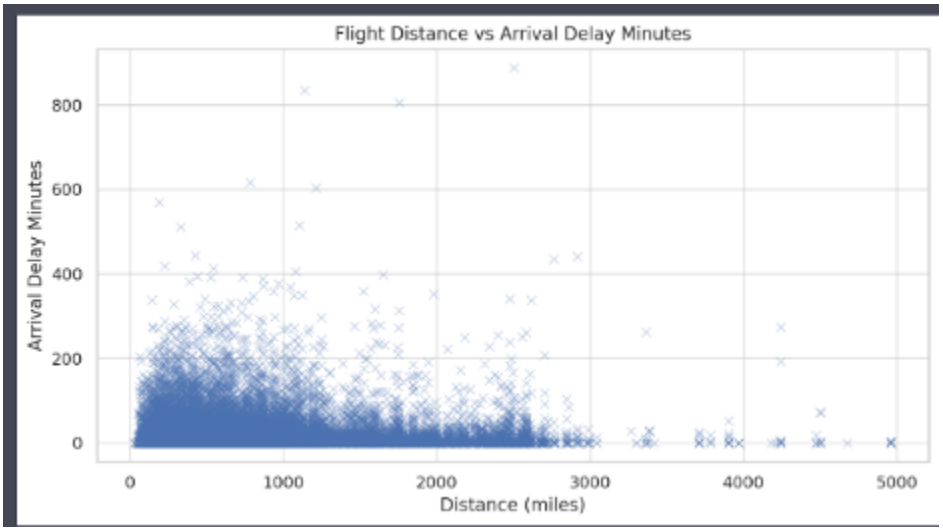


Figure 3.4: Flight Distance vs Arrival Delay Minutes

The scatter plots show the relationship between flight distance and departure/arrival delay minutes. From the visualizations, we can see that there is no clear trend between flight distance and delays.

The correlation coefficient between flight distance and departure delay minutes is 0.0069, and between flight distance and arrival delay minutes is −0.0007. Both correlation coefficients are close to zero, indicating that there is no significant linear relationship between flight distance and delays.

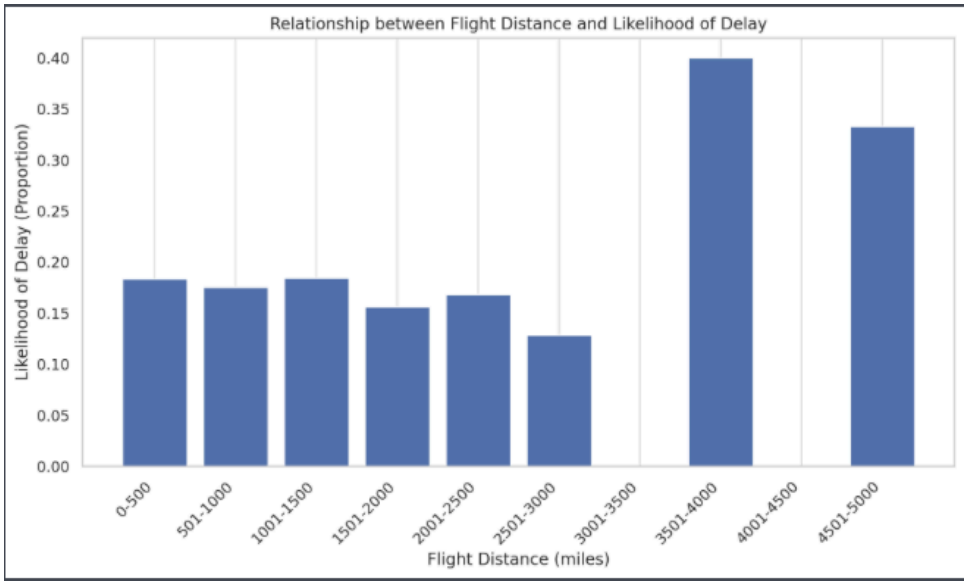


Figure 3.5: Relationship between Flight Distance and Likelihood of Delay

Based on the plot, we can see that the likelihood of delay does not have a clear linear relationship with flight distance. The proportion of delayed flights seems to fluctuate across different distance bins, with no apparent trend. This suggests that flight distance is not a strong predictor of the likelihood of delay. Other factors, such as weather conditions, air traffic, and airline operations, may play a more significant role in determining flight delays.

In conclusion, based on this dataset, we can say that flight distance does not significantly affect the likelihood of departure or arrival delays.

We have identified flight numbers and routes with consistently high average delays:

- There are 3006 flight numbers with consistently high average departure and arrival delays.
- There are 2434 routes with consistently high average departure and arrival delays.

For the identified flight numbers with consistently high delays:

- The average carrier delay is 4.96 minutes.
- The average weather delay is 0.82 minutes.
- The average NAS delay is 3.42 minutes.
- The average security delay is 0.02 minutes.
- The average late aircraft delay is 6.36 minutes.

For the identified routes with consistently high delays:

- The average carrier delay is 4.98 minutes.
- The average weather delay is 0.80 minutes.
- The average NAS delay is 4.13 minutes.
- The average security delay is 0.02 minutes.
- The average late aircraft delay is 6.30 minutes.

Overall Average Departure Delay: 13.32minutes

Overall Average Arrival Delay: 12.75minutes

Top 5 Airlines with Highest Average Departure Delay:

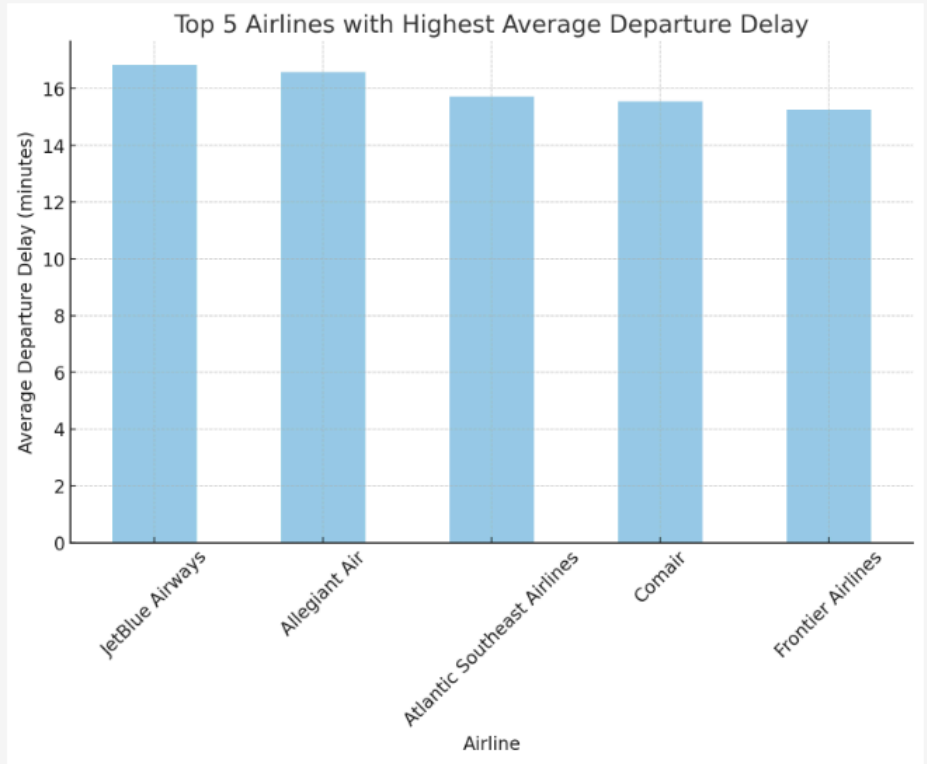


Figure 3.6: Top 5 Airlines with Highest Average Departure Delay

A spokesperson for JetBlue pointed to the carrier's many flights in the "congested weather-prone northeast corridor," which they said affects delays.

A spokesperson for Allegiant pointed to high demand and staff shortages in addition to outside factors like weather as a reason for delays.

The bar chart above shows the top 5 airlines with the highest average departure delay in minutes. The airlines with the highest average departure delays are:

JetBlue Airways: 16.83minutes

Allegiant Air: 6.59minutes

Atlantic Southeast Airlines: 15.73minutes

Comair: 15.55minutes

Frontier Airlines: 15.27minutes

Top 5 Airlines with Highest Average Arrival Delay:

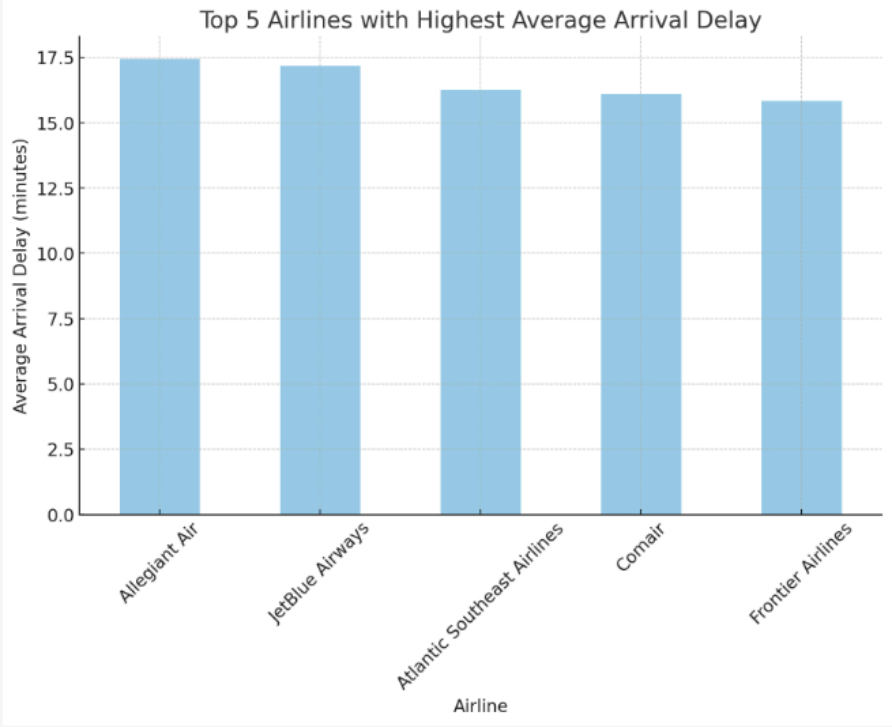


Figure 3.7: Top 5 Airlines with Highest Average Arrival Delay

The bar chart above shows the top 5 airlines with the highest average arrival delay in minutes. The airlines with the highest average arrival delays are:

Allegiant Air: 17.43minutes

JetBlue Airways: 17.17minutes

Atlantic Southeast Airlines: 16.24minutes

Comair: 16.09minutes

Frontier Airlines: 15.84minutes

Top 5 Flight Numbers with Highest Average Departure Delay:

6889: 131.00 minutes

7177: 121.00 minutes

7158: 110.50 minutes

6061: 94.54 minutes

7098: 93.50 minutes

Top 5 Flight Numbers with Highest Average Arrival Delay:

7177: 140.00 minutes

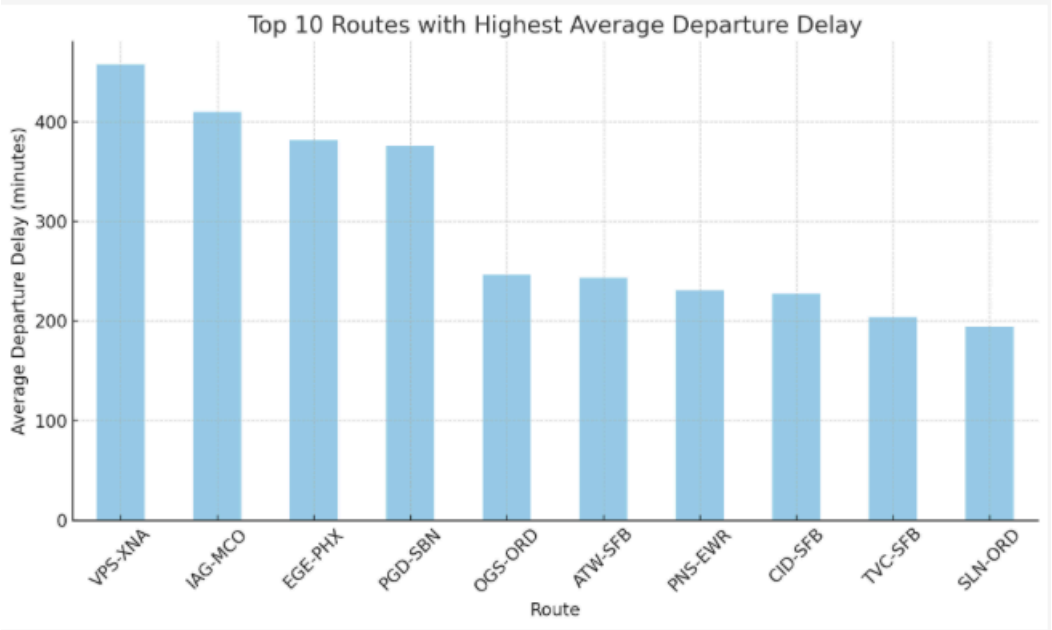
6889: 136.00 minutes

7171: 101.50 minutes

7020: 99.00 minutes

6061: 96.40 minutes

Routes with Highest Average Departure Delay:



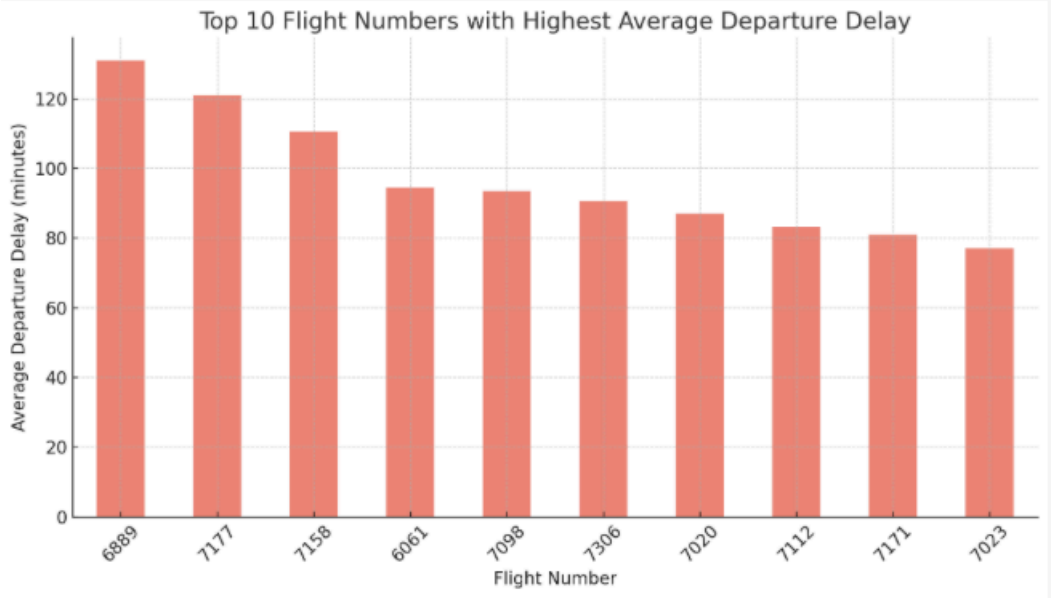


Figure 3.8, 3.9: Top 10 Flight Numbers and Top 10 Routes with Departure Delay

The bar charts above show the top 10 routes and flight numbers with the highest average departure delay in minutes. From the charts, we can see that certain routes and flight numbers consistently experience higher departure delays than others.

VPS-XNA: 458.00 minutes

IAG-MCO: 410.00 minutes

EGE-PHX: 381.67 minutes

PGD-SBN: 376.33 minutes

OGS-ORD: 246.57 minutes

ATW-SFB: 243.50 minutes

PNS-EWR: 231.00 minutes

CID-SFB: 227.50 minutes

TVC-SFB: 204.00 minutes

SLN-ORD: 194.00 minutes

VPS: Destin-Fort Walton Beach Airport

XNA: Northwest Arkansas Regional Airport

IAG: Niagara Falls International Airport

MCO: Orlando International Airport

EGE: Eagle County Regional Airport

PHX: Phoenix Sky Harbor International Airport

PGD: Punta Gorda Airport

SBN: South Bend International Airport

OGS: Ogdensburg International Airport

ORD: O'Hare International Airport

ATW: Appleton International Airport

SFB: Orlando Sanford International Airport

PNS: Pensacola International Airport

EWR: Newark Liberty International Airport

CID: The Eastern Iowa Airport

TVC: Cherry Capital Airport

SLN: Salina Regional Airport

Flight Numbers with Highest Average Departure Delay:

Flight 6889: 131.00 minutes

Flight 7177: 121.00 minutes

Flight 7158: 110.50 minutes

Flight 6061: 94.54 minutes

Flight 7098: 93.50 minutes

Flight 7306: 90.67 minutes

Flight 7020: 87.00 minutes

Flight 7112: 83.20 minutes

Flight 7171: 81.00 minutes

Flight 7023: 77.00 minutes

Next, we will perform a similar analysis for arrival delays to identify consistently delayed routes and flight numbers based on arrival times.

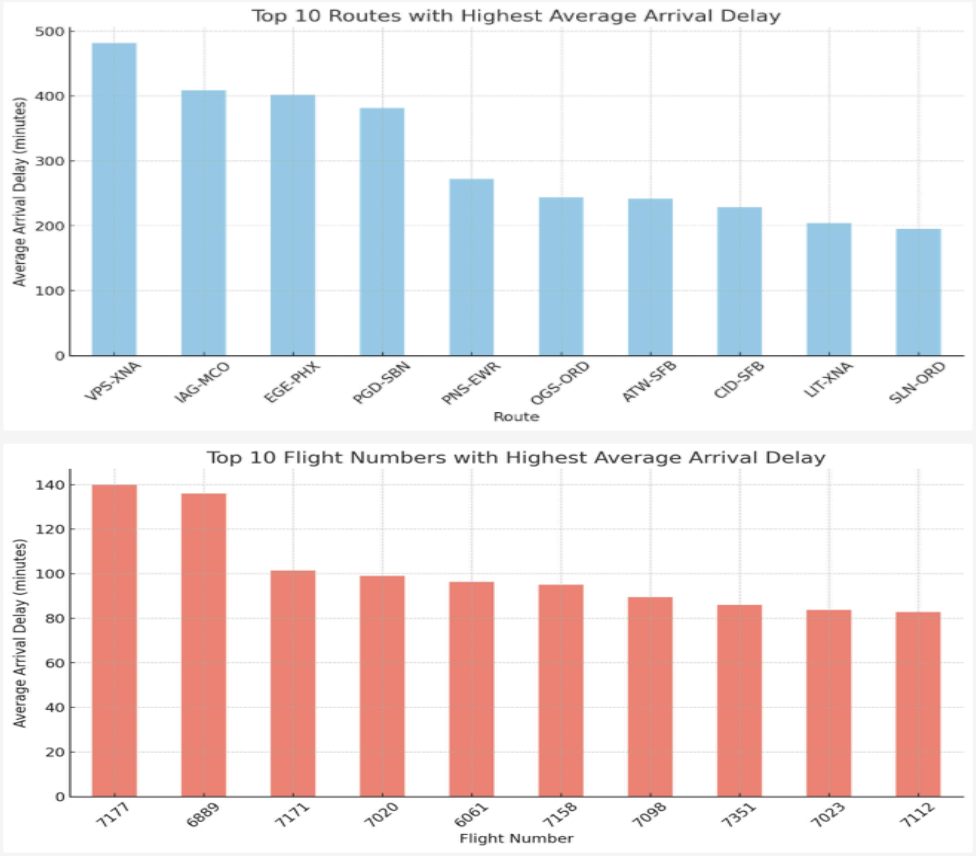


Figure 3.10, 3.11: Top 10 Flight Numbers and Top 10 Routes with Arrival Delay

The bar charts above show the top 10 routes and flight numbers with the highest average arrival delay in minutes. Similar to the departure delay analysis, we can see that certain routes and flight numbers consistently experience higher arrival delays than others.

Routes with Highest Average Arrival Delay:

VPS-XNA: 482.00 minutes

IAG-MCO: 409.00 minutes

EGE-PHX: 401.67 minutes

PGD-SBN: 381.33 minutes

PNS-EWR: 272.00 minutes

OGS-ORD: 243.57 minutes

ATW-SFB: 242.00 minutes

CID-SFB: 228.50 minutes

LIT-XNA: 204.00 minutes

SLN-ORD: 195.40 minutes

Flight Numbers with Highest Average Arrival Delay:

Flight 7177: 140.00 minutes

Flight 6889: 136.00 minutes

Flight 7171: 101.50 minutes

Flight 7020: 99.00 minutes

Flight 6061: 96.40 minutes

Flight 7158: 95.00 minutes

Flight 7098: 89.50 minutes

Flight 7351: 86.00 minutes

Flight 7023: 83.75 minutes

Flight 7112: 82.80 minutes

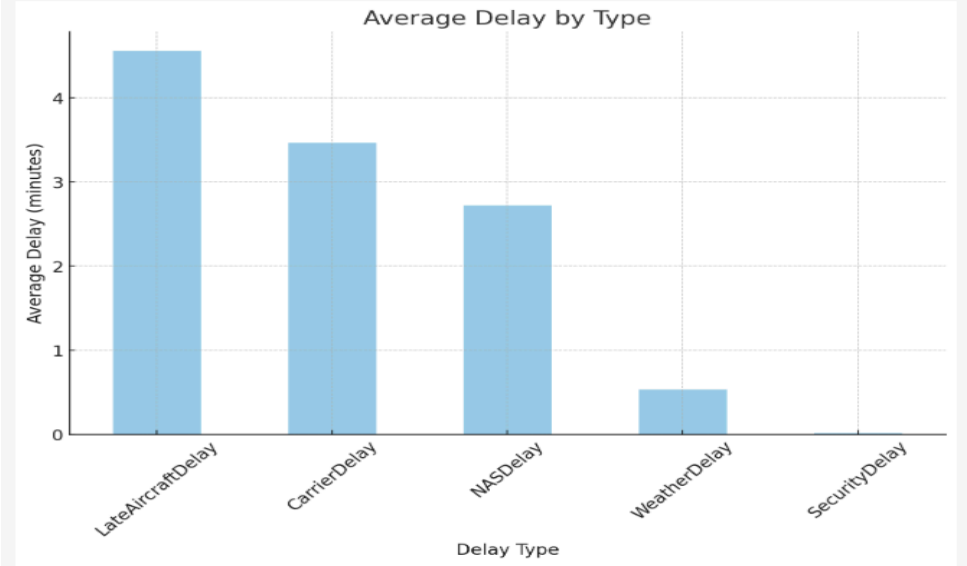


Figure 3.12: Average Delay by Type

The bar chart above shows the average delay in minutes for each delay type. From the chart, we can see that LateAircraftDelay and CarrierDelay are the most common reasons for flight delays, followed by NASDelay, WeatherDelay, and SecurityDelay.

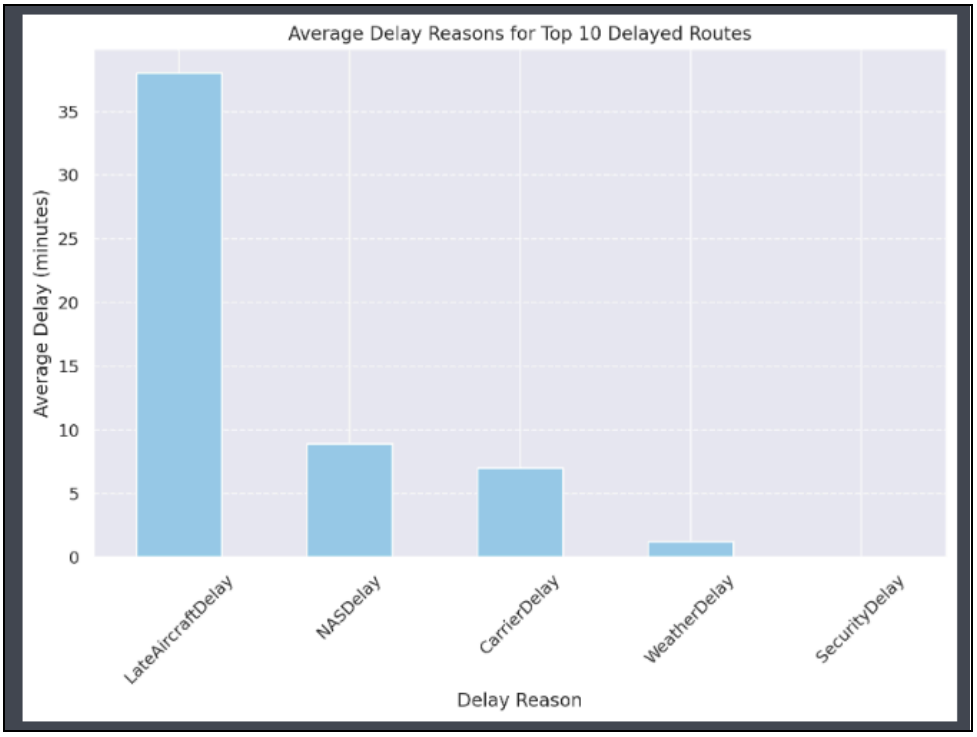


Figure 3.13: Average Reasons for Top 10 Delayed Routes

The bar plot above shows the average delay (in minutes) for each delay reason (CarrierDelay, WeatherDelay, NASDelay, SecurityDelay, LateAircraftDelay) on the top 10 most delayed routes.

Recommendations for Improvement:

Weather Delays: Airlines can improve weather-related delays by closely monitoring weather conditions and adjusting flight schedules accordingly. They can also invest in more advanced technology to better handle adverse weather conditions.

Carrier Delays: Airlines should regularly maintain and inspect aircraft to prevent mechanical issues and delays. They should also ensure that flight crews are properly trained and that there is sufficient staffing to handle flights.

Late Aircraft Arrival: Airlines can improve late aircraft arrival delays by optimizing flight schedules and turnaround times. They should also have contingency plans in place for when flights are delayed due to late aircraft arrival.

Coordinate with NAS: Coordinate with the National Aviation System (NAS) to minimize delays caused by air traffic control, airport operations, and other NAS-related issues.

Enhance Security Procedures: Although Security Delay is the least common reason for delays, it is still important to review and enhance security procedures to ensure the safety of passengers while minimizing delays.

Short Distances (0-500 miles): Flights in this category have average arrival delays of approximately 5.44 minutes and departure delays of 8.72 minutes. Carrier, weather, and National Air System (NAS) delays are notable, with carrier delays averaging around 17.92 minutes.

Medium Distances (500-1500 miles): As the distance increases, the arrival and departure delays show a slight variation, with arrival delays around 4.75 to 4.52 minutes and departure delays ranging from 9.43 to 10.40 minutes. Carrier and NAS delays remain significant contributors.

Longer Distances (1500-3000 miles): Interestingly, arrival delays slightly decrease for flights covering 1500 to 2000 miles, averaging around 2.53 minutes, but slightly increase for flights covering 2500 to 3000 miles. Departure delays hover around 9.93 to 10.40 minutes. NAS and carrier delays continue to be prominent.

Very Long Distances (3000-5000 miles): For the longest flight distances, arrival delays vary, with some distance groups showing slightly higher delays (like 5.07 minutes for 3000-3500 miles) and others showing lower or even negative delays. Departure delays are higher in this category, especially for flights covering 4500-5000 miles (average 14.51 minutes). Carrier delays are notably high, particularly for the 4500-5000 miles category, averaging 73.68 minutes.

Overall Trend: There isn't a straightforward linear relationship between distance and delay. While longer flights (especially those over 3000 miles) tend to have higher departure delays, arrival delays do not consistently increase with distance. In some distance groups, arrival delays even decrease or become negative, indicating early arrivals.

In conclusion, while flight distance does influence the likelihood of delay, it is not the sole factor. The data suggests that other factors such as carrier-specific issues, weather, and NAS delays play significant roles, irrespective of the flight distance. Therefore, while longer flights might be prone to higher departure delays, they do not necessarily experience proportionally higher arrival delays.

Research Question #4: Is there a correlation between the departure delay and arrival delay for flights? Which airline(s) consistently have the best and worst on-time performance?

Understanding the dynamics between departure and arrival delays is pivotal in the aviation industry, not only for optimizing operations but also for improving the overall passenger experience. Delays can have a ripple effect, impacting airline costs, passenger plans, and even the broader economy. Furthermore, identifying airlines with the best and worst on-time performance can be invaluable for consumers making travel choices and for carriers striving for operational excellence.

To dissect these questions, our analysis will encompass several investigative steps:

- **Data Loading and Inspection**: We begin by loading the comprehensive dataset from Kaggle, which includes various data points such as flight numbers, scheduled and actual departure/arrival times, and more. An initial inspection will help us understand the data's structure and integrity.
- **Correlation Analysis**: We will compute the Pearson correlation coefficient to quantify the relationship between departure and arrival delays. This will reveal if there is a statistically significant linear relationship between the two variables.
- **Performance Benchmarking**: By grouping the data by airline, we will calculate average departure and arrival delays to benchmark on-time performance. This approach will highlight the airlines that consistently perform above or below industry standards.

- **Visualization**: To aid in the interpretation of our findings, we will present a series of visualizations:

1. A scatter plot to visually assess the correlation between departure and arrival delays.
2. A histogram to examine the distribution of departure and arrival delays.
3. A box plot to depict the spread and identify outliers in delay times.
4. A bar chart to rank airlines based on their average total delays, identifying the best and worst performers.

By analyzing and visualizing the data, we aim to provide a nuanced understanding of how departure delays may affect arrival times and which airlines lead or lag in punctuality. These insights could be instrumental for stakeholders in making informed decisions and implementing strategies to mitigate delays.

Analysis of Departure Delays

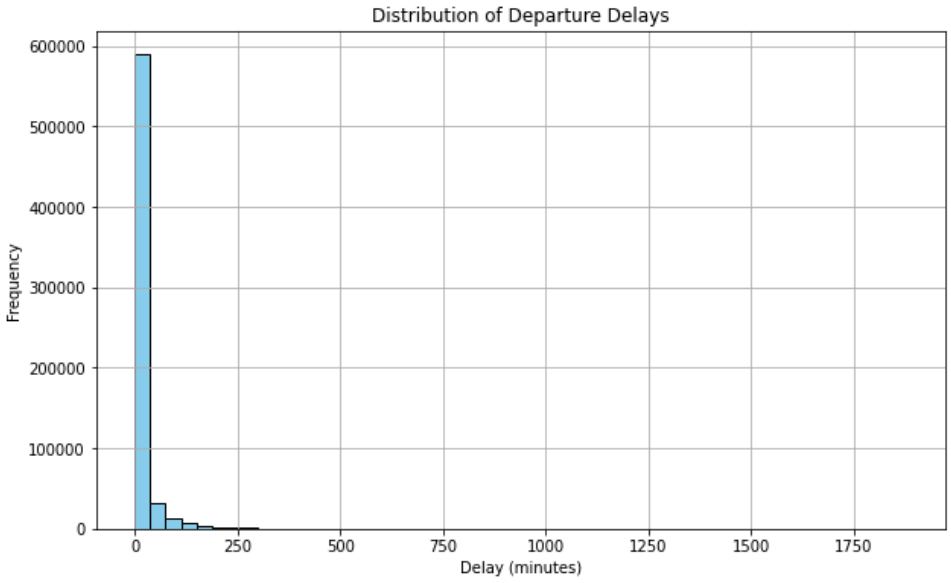


Figure 4.1: Histogram representation of Departure Delays

The accompanying histogram provides a visual representation of the distribution of departure delays across all flights within the dataset obtained from Kaggle.
From the histogram, we can infer the following points:

- **Majority On-Time**: A predominant number of flights depart on time or with minimal delays, as indicated by the first bar's height.
- **Right-Skewed Distribution**: There's a swift decline in frequency as delay time increases, with longer delays being less common.
- **Occasional Extended Delays**: The presence of bars past the 250-minute mark, though low in height, identifies the occasional substantial delays.
- **Operational Efficiency**: The data suggests that most flights are managed efficiently, with only a minority experiencing significant delays.

The data portrayed in the histogram is a crucial indicator of airline efficiency regarding departure times. While the bulk of flights manage to keep delays to a minimum, the spread of data to the right highlights the challenges airlines face in mitigating longer delays. This information could be foundational for airlines looking to improve their departure punctuality, as it emphasizes the need to address the causes of both frequent short delays and rarer extended delays.

Analysis of Arrival Delays

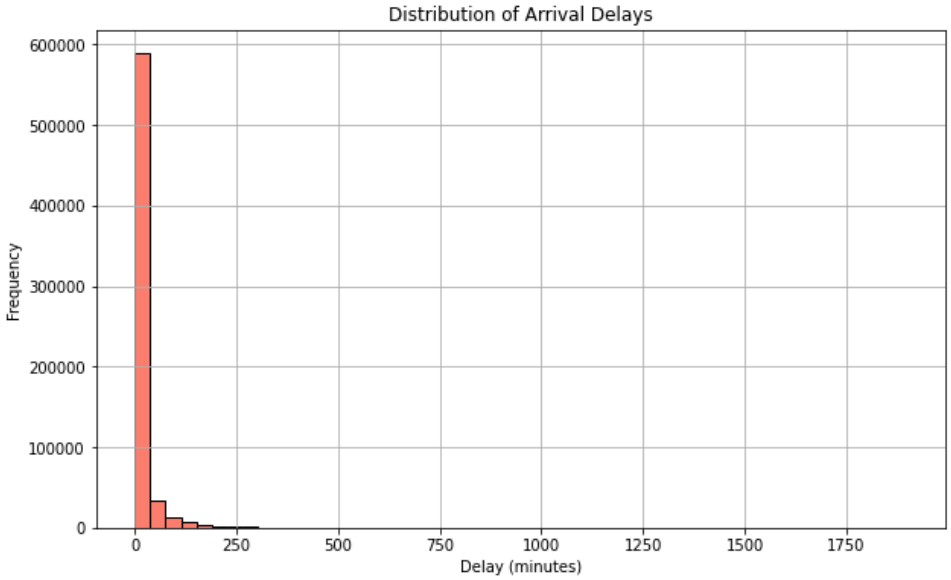


Figure 4.2: Histogram representation of Arrival Delays

This histogram depicts the distribution of arrival delays across flights:

- **Predominance of Timely Arrivals**: The towering initial bar indicates that most flights arrive on schedule or with slight delays.
- **Decrease in Frequency with Longer Delays**: The rapid decline in bar height past the initial one suggests that significant arrival delays are relatively rare.
- **Presence of Extreme Delays**: Although uncommon, the graph identifies some instances of extreme delays, which may warrant further investigation.
- **Operational Insight**: The concentration of flights with minimal arrival delays implies efficient in-flight time management and recovery from any departure delays.

The patterns observed offer a snapshot of arrival punctuality, highlighting an overall trend towards on-time performance with a few exceptions.

Correlation Between Departure and Arrival Delays

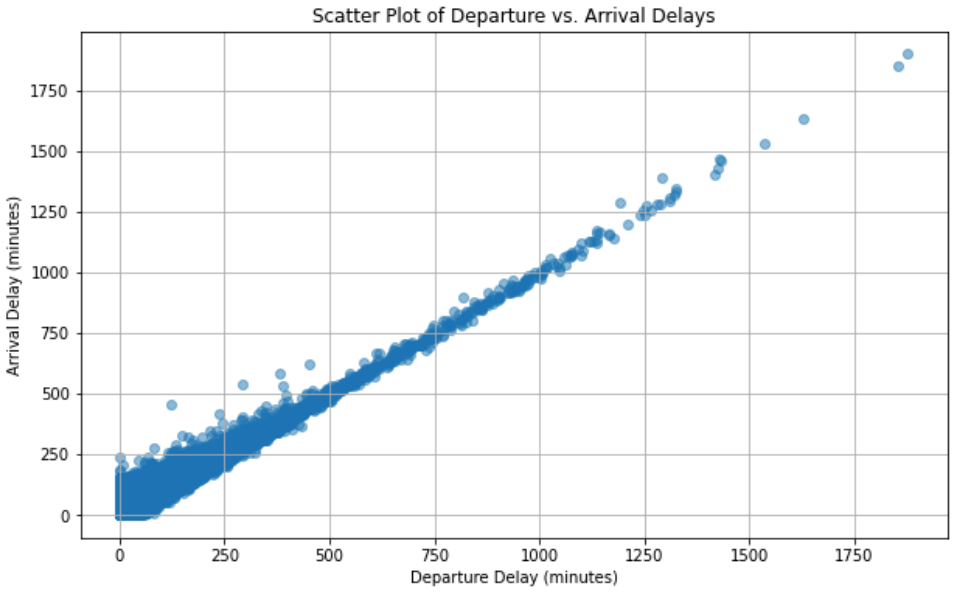


Figure 4.3: Scatter plot representing the relationship between Departure and Arrival Delays

The scatter plot provides a visual correlation between departure and arrival delays:

- **Direct Relationship**: A clear positive trend indicates that flights that depart late tend to also arrive late.
- **Proportionality**: The density of points along a line through the origin suggests a proportional relationship, where the longer the departure delay, the longer the arrival delay tends to be.
- **Outliers**: A few points far from the main cluster highlight exceptional cases with substantial delays, likely due to specific disruptive events.

- **Operational Adjustments**: Despite some variability, the trend shows limited instances of flights making up time in the air, pointing to potential constraints in reducing arrival delays.

This pattern affirms the direct impact of departure timing on arrival accuracy and underscores the importance of punctual departures for on-time arrivals.

Dispersion of Departure and Arrival Delays

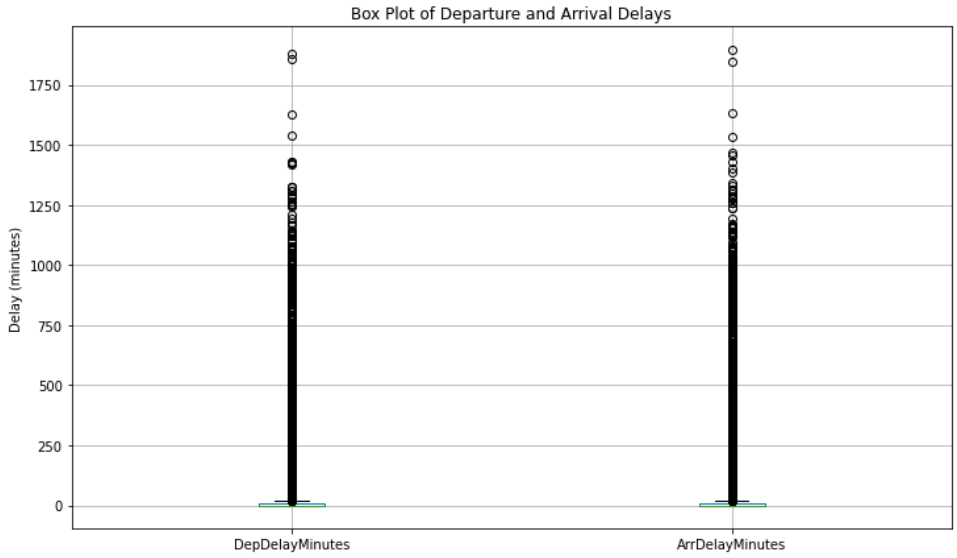


Figure 4.4: Box plot representing the comparison between Departure and Arrival Delays

The box plot comparison between departure and arrival delays illustrates several aspects of flight delay patterns:

- **Central Tendency**: The medians of both departure and arrival delays are low, suggesting that more than half of the flights experience short delays.

- **Variability**: The interquartile ranges (IQRs) indicate a moderate spread of delays around the median, with arrival delays showing a slightly wider IQR than departure delays.
- **Outliers**: Numerous outliers for both departure and arrival delays point to exceptional cases where delays are significantly longer than typical.
- **Symmetry**: The similar distribution patterns between departure and arrival delays underscore their interrelated nature.

This visual analysis highlights that while most flights manage to depart and arrive with minimal delay, there are notable exceptions that could impact overall operational efficiency and passenger satisfaction.

Analysis of WORST On-Time Performance Airlines

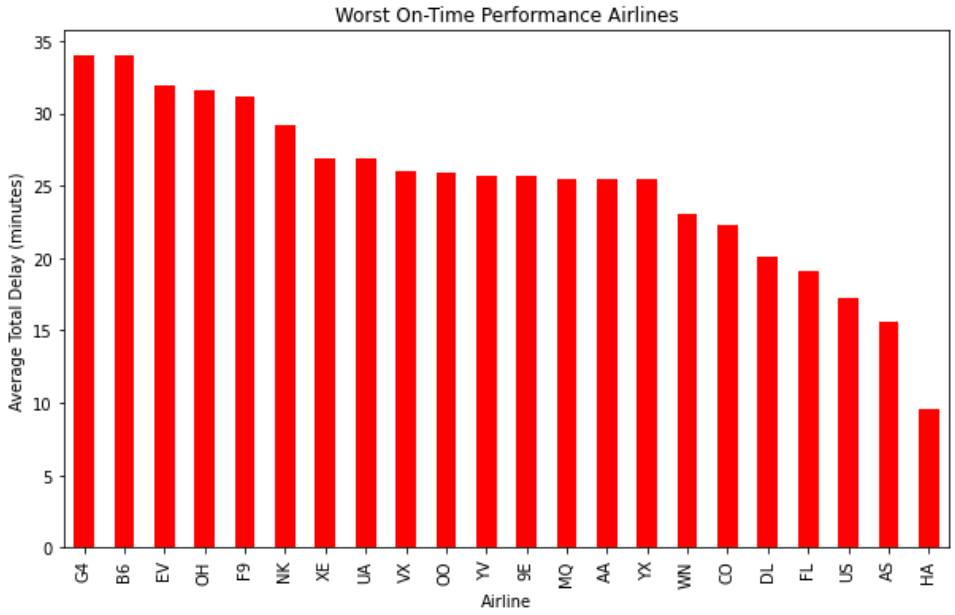


Figure 4.5: Bar Chart representing the WORST on-time performing flights

The bar chart presented here ranks airlines based on their average total delay, providing a clear visualization of those with room for improvement in on-time performance. Allegiant Air LLC, JetBlue Airways, ExpressJet Airlines, PSA Airlines, and Frontier Airlines are identified as the carriers with the highest average delays.

The reasons for these delays are multifaceted and may be influenced by the budget-conscious nature of some of these carriers:

- **Cost-Saving Measures**: Budget airlines often operate with minimal margins for error to keep costs low. This can result in tight scheduling with little room for delay recovery.
- **High Traffic Hubs**: Carriers like JetBlue that serve busy airports are more exposed to operational delays due to air traffic congestion, especially during peak hours.
- **Rapid Turnaround Times**: Regional airlines, such as PSA and ExpressJet, face quick turnaround challenges, which are exacerbated by the high volume of daily flights typical for low-cost models.
- **Weather and Geographical Impacts**: Airlines operating in regions with unpredictable weather patterns may encounter more frequent delays, a situation that budget carriers are less equipped to manage due to their limited resources.
- **Scheduling and Fleet Utilization**: The need to maximize aircraft usage can lead to schedules that have little buffer for delays, impacting overall performance.

In recognizing these challenges, it becomes evident that while budget airlines offer economical options for travelers, the trade-off can be an increased likelihood of delays. This highlights an opportunity for such airlines to explore innovative operational strategies to improve their on-time performance without compromising their cost-effective approach.

Analysis of BEST On-Time Performance Airlines

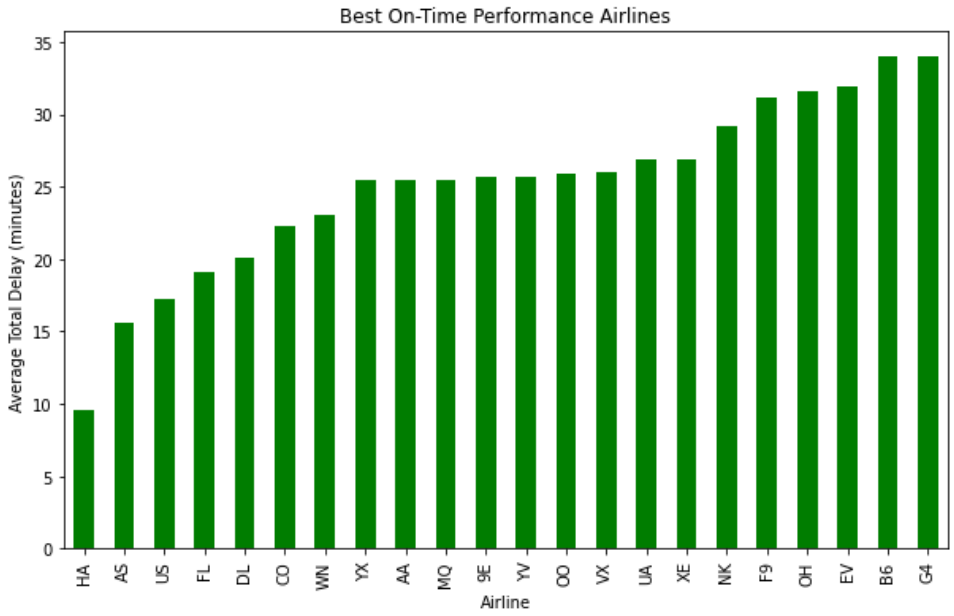


Figure 4.6: Bar Chart representing the BEST on-time performing flights

The bar chart represents the average total delay for each airline, with shorter bars signifying superior performance. Illustrated in this graph are the airlines that have excelled in maintaining schedule integrity, with Hawaiian Airlines, Inc. and Alaska Airlines leading the pack.

Factors contributing to the success of the top performers are likely to include:

- **Optimized Scheduling**: These airlines may have more efficient scheduling systems, allowing ample time between flights to accommodate potential delays.
- **Effective Operations**: Strong operational strategies, including proactive maintenance and quick turnaround practices, help keep delays to a minimum.
- **Geographical Advantages**: Hawaiian Airlines, for example, operates in a region with favorable weather conditions and less congested airspace, which can reduce delay incidents.
- **Customer-Centric Policies**: A focus on customer satisfaction may drive these airlines to prioritize on-time performance, aligning crew, ground staff, and systems to this goal.
- **Resource Allocation**: Efficient allocation of resources such as gates and staffing can also contribute to smoother operations and fewer delays.

This chart not only showcases the airlines that are performing well but also sets a benchmark for others in the industry. It suggests that a combination of strategic planning, resource management, and perhaps even geographical location plays a crucial role in achieving high on-time performance ratings.

CONCLUSION AND RECOMMENDATIONS BASED ON OUR FINDINGS:

Our analysis of the correlation between departure and arrival delays across various airlines has yielded clear insights. The data confirms a significant positive

correlation: generally, flights that depart late also arrive late. This emphasizes the critical nature of timely departures in achieving on-time arrivals, a key determinant of customer satisfaction and operational success.

Our analysis has identified airlines that consistently outperform and underperform in terms of on-time records. Hawaiian Airlines, Inc. and Alaska Airlines emerged as leaders, likely benefiting from efficient scheduling, effective operations, and in some cases, geographical advantages. On the other end of the spectrum, airlines like Allegiant Air LLC and JetBlue Airways faced challenges, often attributed to the constraints of low-cost operational models, including tight scheduling and limited resources that reduce their margin for managing delays.

The visualizations provided have allowed us to delve deeper into the distribution of delays, highlighting the airline's operational efficiencies and areas for improvement. They also serve as a call to action for those airlines lagging to adopt best practices from industry leaders.

In conclusion, this research underscores the importance of strategic operational management in minimizing delays. It offers a roadmap for airlines seeking to improve their on-time performance, thereby enhancing passenger experiences and maintaining a competitive edge in the aviation market.

Research Question #5: Can machine learning models accurately predict flight delays based on historical data from this dataset?

In pursuit of enhancing airline and airport operations to mitigate flight delays and improve the overall customer experience, this report employs machine learning techniques with Python, leveraging the powerful Pandas library for data analytics. The objective is to address the question of predicting flight delays by employing various machine learning methods and subsequently evaluating their accuracy and performance. Here is the methodology that we are going to deploy and analyze the data set:

1. **Data Preparation:**
    a. Handling Missing Data:
        - Identify and remove missing data points to ensure a clean dataset for analysis.
    b. Multicollinearity and Variable Selection:
        - Identify and eliminate columns with multicollinearity to enhance model interpretability.
        - Eliminate unnecessary variables that may not contribute significantly to the prediction.
    c. Dummy Variable Creation:
        - Transform categorical variables into dummy variables to facilitate their inclusion in the machine learning models.
2. **Machine Learning Models:**
    a. Regression Method:

- Implement multiple regression to establish a baseline predictive model.

- Evaluate the model's performance and interpretability.

b. Decision Tree Method:

- Employ decision tree algorithms to capture non-linear relationships in the data.

- Tune hyperparameters for optimal performance.

c. Random Forest Method:

- Utilize the random forest algorithm to improve predictive accuracy and handle complex relationships.

- Assess the impact of ensemble learning on model performance.

3. Analysis:

a. Evaluate Model Results:

- Compare the predictive capabilities of the three methods.

- Assess the strengths and weaknesses of each model.

b. Error Calculation:

- Calculate relevant metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) for each model.

c. Accuracy Comparison:

- Quantitatively compare the accuracy of the multiple regression, decision tree, and random forest methods.

Data Preparation

First, we have to import the dataset into memory. Because the dataset is very large, we've divided it into numerical and categorical columns for ease of use as shown in Figure 5.1.

```
# Class: ISDS 577
# Author: Hy Luong
# CAPSTONE PROJECT

# In[10]: Reading in the data set
import pandas as pd
import numpy as np
df = pd.read_csv('D:/CSUF/2023 Fall/ISDS 577/Final Project/airline_cleaned_csv.csv')
# print(df)
column_names = df.columns.tolist()
```

Figure 5.1: Python code for importing the data set and separating numerical/categorical columns.

Since the cleaned dataset that was distributed to the team is very general. We must refine it further to make it suitable for our machine-learning task. The initial step involves visualizing missing values in the dataset using Seaborn's heatmap. This provides a clear overview of the locations and patterns of missing data, aiding in decision-making for subsequent handling strategies (Figure 5.2 and 5.3).

```
# In[20]: Cleaning the missing values
import seaborn as sns
import matplotlib.pyplot as plt

# Create a boolean mask for missing values (True for missing, False for non-missing)
missing_mask = df.isna()  # You can also use df.isnull()
```

```
# Use seaborn's heatmap to visualize the missing value map
plt.figure(figsize=(8, 6))
sns.heatmap(missing_mask, cmap='viridis', cbar=False)
plt.title('Missing Value Map')
plt.show()
```

Figure 5.2: Python code for visualizing missing data.

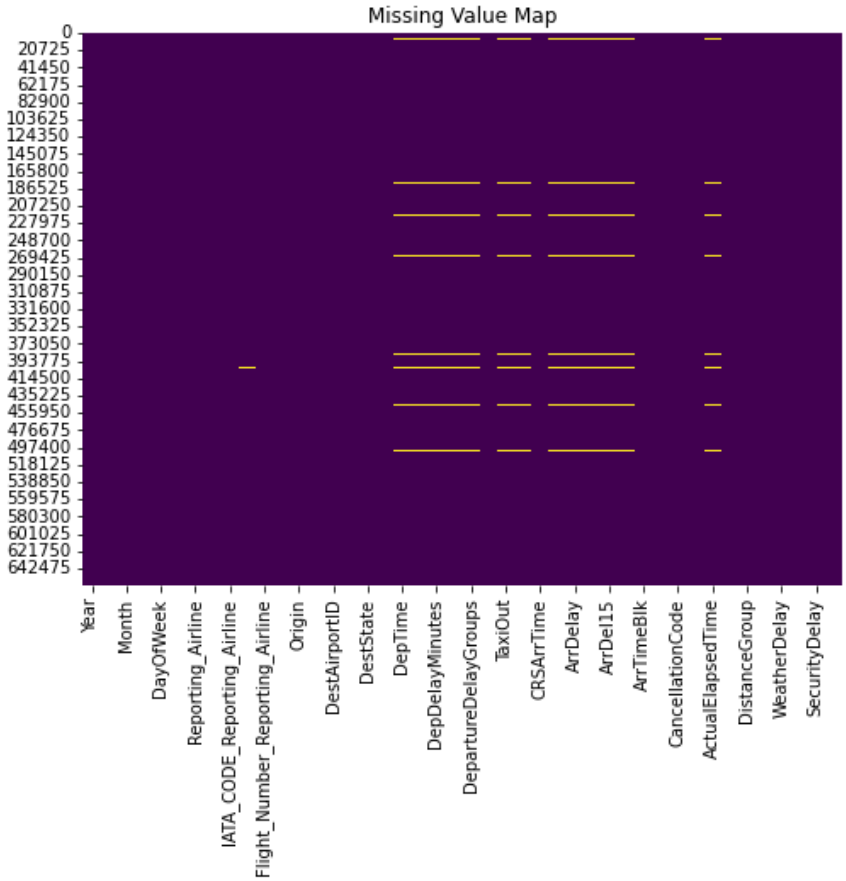


Figure 5.3: Missing Value Map

Following the visualization, rows with missing values are identified and subsequently removed to ensure a clean and complete dataset for analysis (13,463 rows). Temporary variables used in the analysis are cleaned up for a more organized

working environment and save memory (Figure 5.4).

```
# Create a boolean mask for missing values (True for missing, False for non-missing)
missing_mask = df.isna()  # Can also use df.isnull()

# Count the number of rows with missing values
num_rows_with_missing = missing_mask.any(axis=1).sum()
print(f"Number of rows with missing values: {num_rows_with_missing}")

#remove all rows with missing values
df = df.dropna()

#clean up the Variable Explorer
del missing_mask
del num_rows_with_missing
del column_names
```

Figure 5.4: Python code to count rows with missing values and drop it.

We proceed to analyze the correlation among numerical variables using Seaborn's heatmap. This visual representation assists in identifying potential relationships between different numerical features. Columns with correlations higher than 70% are printed out and removed, keeping only one of the two (Figure 5.5, 5.6).

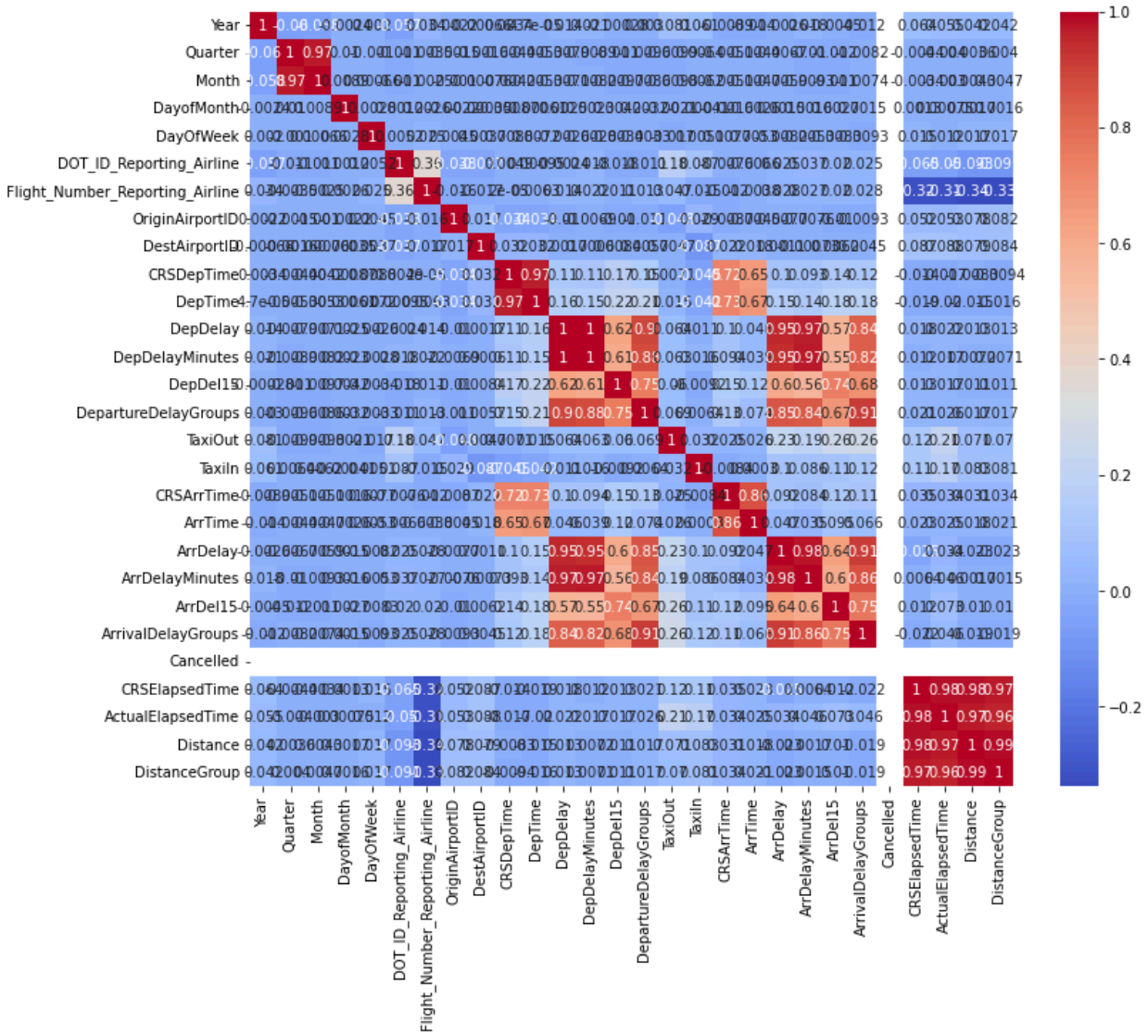


Figure 5.5: Heat map of correlation between numerical variables.

```
# In[30]: Heat map for numerical variables
# correlation analysis for numerical variables
import seaborn as sns
import matplotlib.pyplot as plt

#numerical dataframe
num_df = df[['Year', 'Quarter', 'Month', 'DayofMonth', 'DayOfWeek',
'DOT_ID_Reporting_Airline', 'Flight_Number_Reporting_Airline', 'OriginAirportID',
```

```
'DestAirportID', 'CRSDepTime', 'DepTime', 'DepDelay', 'DepDelayMinutes', 'DepDel15',
'DepartureDelayGroups', 'TaxiOut', 'TaxiIn', 'CRSArrTime', 'ArrTime', 'ArrDelay',
'ArrDelayMinutes', 'ArrDel15', 'ArrivalDelayGroups', 'Cancelled', 'CRSElapsedTime',
'ActualElapsedTime', 'Distance', 'DistanceGroup']]

correlation_matrix = num_df.corr()
plt.figure(figsize=(12, 10))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm")
plt.show()

# Set the correlation threshold
threshold = 0.7  # You can change this threshold as needed

# Create empty lists to store variable pairs with strong correlations
strong_positive_correlations = []
strong_negative_correlations = []

# Loop through the correlation matrix and identify strong correlations
for i in range(len(correlation_matrix.columns)):
    for j in range(i+1, len(correlation_matrix.columns)):
        if abs(correlation_matrix.iloc[i, j]) > threshold:
            if correlation_matrix.iloc[i, j] > 0:
                strong_positive_correlations.append((correlation_matrix.columns[i],
correlation_matrix.columns[j], correlation_matrix.iloc[i, j]))
            else:
                strong_negative_correlations.append((correlation_matrix.columns[i],
correlation_matrix.columns[j], correlation_matrix.iloc[i, j]))

# Print or display the results
print("Strong Positive Correlations:")
for variable1, variable2, correlation in strong_positive_correlations:
    print(f"{variable1} and {variable2}: {correlation:.2f}")

print("\nStrong Negative Correlations:")
```

```
for variable1, variable2, correlation in strong_negative_correlations:
    print(f"{variable1} and {variable2}: {correlation:.2f}")

# In[41]: Dropping numerical variables
# Drop a variable (column) by specifying its name
columns_to_drop = ['Quarter', 'DepDelay',
'DepartureDelayGroups','ArrDelay','ArrivalDelayGroups','ActualElapsedTime','Distance','Distan
ceGroup']
df = df.drop(columns=columns_to_drop)
```

Figure 5.6: Python code for identifying and dropping highly correlated numerical variables.

We now move on to selecting categorical variables. First, we removed any reason for delay because we were trying to predict the delay. Moreover, we also see that *IATA_CODE_Reporting_Airline* and *Reporting_Airline* have the same values; therefore, we would also remove *IATA_CODE_Reporting_Airline (Figure 5.7)*.

```
# Drop a variable (column) by specifying its name

columns_to_drop = ['IATA_CODE_Reporting_Airline', 'CancellationCode', 'CarrierDelay',

'WeatherDelay', 'NASDelay', 'SecurityDelay', 'LateAircraftDelay']

df = df.drop(columns=columns_to_drop)
```

Figure 5.7: Python code for removing unnecessary variables.

To go any further, we must create dummy variables for the remaining categorical variables. Additionally, we employ random shuffling of the dataset to facilitate the subsequent selection of a subset for assessing the correlation among categorical variables.

```
# Shuffling the DataFrame

shuffled_df = df.sample(frac=1)

print(shuffled_df)

# Selecting the categorical columns

categorical_columns_names = ['FlightDate', 'Reporting_Airline', 'Tail_Number', 'Origin',

'OriginState', 'Dest', 'DestState', 'DepTimeBlk', 'ArrTimeBlk']

# Using get_dummies() to convert categorical variables into dummy/indicator variables

dummies = pd.get_dummies(shuffled_df, columns=categorical_columns_names)
```

Figure 5.7: Python code for creating dummy variables.

Because our dataset is too big to generate the correlation between the categorical variables, we had to choose a smaller random subset to test for multicollinearity. We randomly sampled approximately 2% of the data frame to assess multicollinearity. Everything with a correlation score of more than 70% is then printed out and evaluated (Figure 5.8).

```
# Creating a new DataFrame with n lines

dummy_subset = dummies.head(6500)

# Compute the correlation matrix

correlation_matrix = dummy_subset.corr()

# Display the correlation matrix

print("Categories with correlation greater than 0.7:")
```

```
for col in correlation_matrix.columns:
    for index, value in correlation_matrix[col].items():
        if col != index and abs(value) > 0.7:
            print(f"{col} - {index}")
```

Figure 5.8: Python code for checking correlation of all categorical subsets.

After checking the correlation, we decided to drop “OriginState”, “DestState”, and “Tail_Number” (Fifure 5.9).

```
# Drop a variable (column) by specifying its name
columns_to_drop = ['OriginState', 'DestState', 'Tail_Number']
df = df.drop(columns=columns_to_drop)
print(df)
```

Figure 5.9: Python code for dropping categorical variables.

Machine Learning Models

Regression

We utilized scikit-learn library in Python to perform the multiple regression. The dataset, represented by the variables X (independent variables) and y (dependent variable), is split into 70% training and 30% testing sets using the train_test_split function. To ensure that the optimization process is not skewed toward features with larger magnitudes, the features in X are then standardized using StandardScaler. An SGDRegressor model is created and fitted using the training data, and predictions are

made on the scaled test data. The code extracts and displays the significant variables based on the coefficients obtained from the model. The significance threshold is set at 1% (alpha), and only variables with coefficients greater than 0.01 or less than -0.01 are considered. Finally, the code evaluates the model's performance by calculating and printing the Mean Absolute Error between the actual and predicted values on the test set using metrics from scikit-learn. This process provides insights into the most influential variables and assesses the model's predictive accuracy (Figure 5.10).

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import SGDRegressor
from sklearn.preprocessing import StandardScaler
from sklearn import metrics
from sklearn.metrics import mean_squared_error


# Dataset
X = dummies.drop(columns=['ArrDelayMinutes'], axis=1)
y = dummies['ArrDelayMinutes']
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
# Scale the features (Standardization)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

```
# Create and fit the SGDRegressor model
model = SGDRegressor(max_iter=1000, tol=1e-3, random_state=42)
model.fit(X_train_scaled, y_train)
# Make predictions
y_pred = model.predict(X_test_scaled)
# Extract and display the significant variables
coefficients = pd.Series(model.coef_, index=X.columns)
significant_variables = coefficients[coefficients.abs() > 0.01]  # Alpha = 1%
significant_variables_sorted = significant_variables.abs().sort_values(ascending=False)
print("Most significant variables:")
print(significant_variables_sorted)
# Evaluating the model
print("Mean Absolute Error:", metrics.mean_absolute_error(y_test, y_pred))
# Calculate the mean squared error
mse = mean_squared_error(y_test, y_pred)
print(f"Mean Squared Error: {mse}")
```

Figure 5.10: Python code for Regression method.

Decision Tree

We employ a Decision Tree Regressor from the scikit-learn library to create and evaluate a regression model. It begins by importing necessary libraries for data manipulation, model creation, and performance evaluation. The dataset, represented by features (X) and target variable (y), is then split into training and testing sets using the

train_test_split function. A Decision Tree Regressor is instantiated, trained on the training set using the fit method, and subsequently used to predict the target variable for the test set with the prediction method. Model evaluation metrics are then computed and printed, including the Mean Absolute Error (MAE) using metrics.mean_absolute_error and the Mean Squared Error (MSE) using mean_squared_error. These metrics provide insights into the accuracy and performance of the Decision Tree Regressor on predicting the arrival delay minutes in the test dataset. The model's effectiveness can be further analyzed based on the obtained error metrics (Figure 5.11).

```
# In[20]: Decision Tree

# Importing required libraries

from sklearn.model_selection import train_test_split

from sklearn.tree import DecisionTreeRegressor

from sklearn import metrics

from sklearn.metrics import mean_squared_error



# Dataset

X = dummies.drop(columns=['ArrDelayMinutes'])

y = dummies['ArrDelayMinutes']

# Splitting the dataset into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Creating a decision tree regressor
```

```
regressor = DecisionTreeRegressor()

# Training the decision tree regressor

regressor.fit(X_train, y_train)

# Making predictions on the test data

y_pred = regressor.predict(X_test)

# Evaluating the model

print("Mean Absolute Error:", metrics.mean_absolute_error(y_test, y_pred))

# Calculate the mean squared error

mse = mean_squared_error(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
```

Figure 5.11: Python code for Decision Tree method.

Random Forest

This Python code implements a Random Forest Regressor using scikit-learn for a regression task. After importing the necessary libraries and loading the dataset, the code splits it into training and testing sets. A Random Forest Regressor is then instantiated with specific parameters, including the number of trees in the forest (n_estimators=100), and is trained on the training data. The model is used to predict the target variable for the test set, and the Mean Absolute Error (MAE) and Mean Squared Error (MSE) are computed and printed for model evaluation. These metrics serve as indicators of the accuracy and performance of the Random Forest Regressor in predicting arrival delay minutes, offering valuable insights into the model's effectiveness for the given regression task (Figure 5.12).

```
# In[30]: Random forest

import numpy as np

import pandas as pd

from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestRegressor  # For regression tasks

from sklearn import metrics

from sklearn.metrics import mean_squared_error



# Dataset

X = dummies.drop(columns=['ArrDelayMinutes'])

y = dummies['ArrDelayMinutes']

# Split the data into training and testing sets

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Create a Random Forest Regressor (or Classifier) instance

model = RandomForestRegressor(n_estimators=100, n_jobs=-1, random_state=42)

# Fit the model on the training data

model.fit(X_train, y_train)

# Predict the target variable for the test set

y_pred = model.predict(X_test)

# Evaluating the model
```

```
print("Mean Absolute Error:", metrics.mean_absolute_error(y_test, y_pred))

# Calculate the mean squared error

mse = mean_squared_error(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
```

Figure 5.12: Python code for Random Forest method.

The code for all of the methods is performed on both “ArrDelayMinutes” and “DepDelayMinutes” for comparison.

Analysis

When considering regression, decision tree, and random forest models, each exhibits distinct advantages and drawbacks. Linear regression offers interpretability and computational efficiency but assumes a linear relationship between variables, limiting its ability to capture complex, non-linear patterns. Decision trees, on the other hand, excel in handling non-linearity and are easy to interpret. However, they are prone to overfitting and instability, making them sensitive to small changes in the data. Random forest, an ensemble of decision trees, mitigates the overfitting issue by combining multiple trees and often provides superior predictive accuracy. Yet, the trade-off includes increased complexity, reduced interpretability, and potential computational intensity. Random Forest struck the balance between the accuracy and complexity of the model.

Using Regression analysis, we have MAE for departure and arrival delay of 13 to 14 million minutes, which is a very big number indicating that Regression Analysis is not at all suitable for predicting our flight delays (Appendix 2).

For the Decision Tree, the MAE for Departure delay is roughly 3.5 minutes and 1.6 minutes for Arrival delay. These results are really good. Our predictions are accurate for up to 3 minutes, which is very usable. But we still have one more method to consider, and that is Random Forest (Appendix 3).

With the Random Forest method, we incrementally increase the number of random trees from 1 to 100 until it surpasses our decision tree results. It beat the Decision Tree method after 10 random trees. However, only 10 trees are not a forest yet, so we go up to 100 random trees. It's a good balance between the complexity of the model and the computational power. It produces a Mean Absolute Error (MAE) of 2.5 minutes for departure delays and 1.2 minutes for arrival delays (Appendix 4).

In conclusion, our Random Forest model with 100 trees performed best in predicting both departure and arrival delays. The results show promising accuracy, paving the way for further optimization and application in real-world scenarios.

Recommendation

Using this accurate machine learning algorithm, here are some of our recommendations for decision-makers:

1. **Integration with Operations:** To deliver precise and timely information regarding probable aircraft delays, include the prediction model into regular operational procedures. Make sure that the forecasts are accessible to the appropriate parties, including airport employees and airline operations teams.
2. **Resource Allocation:** Make use of the forecasts to allocate resources as efficiently as possible. For example, send more workers or equipment to areas

where delays are more likely. By taking a proactive stance, airport and airline operations delays can be lessened.

3. **Strategic Planning:** Include the forecasts in strategic planning procedures so that decision-makers may foresee any interruptions and make necessary adjustments ahead of time. This may entail modifying flight schedules, optimizing routes, and improving overall operational efficiency.
4. **Communication with Passengers:** Create a communication plan to alert travelers to any potential delays using the model's projections. Passengers may better organize and manage their expectations when there is clear and timely information.

Conclusions

In our project, we embarked on an extensive exploration of factors influencing airline operations. Our investigation covered a broad spectrum: we analyzed how flight patterns change with seasons and across different days of the week, delved into the timing of delays to ascertain if they predominantly occur at specific times of the day, and scrutinized the relationship between the distance of flights and the extent of their delays. Furthermore, we examined how departure delays correlate with arrival delays across various airlines. Lastly, we ventured into the realm of predictive analytics, exploring the potential of machine learning models in forecasting airline delays, opening avenues for more proactive and efficient airline management strategies.

Our analysis highlights key aspects of airline operations: Seasonal and weekly trends significantly impact flight frequency and delays, with peak delays on Mondays and extended arrival delays on Saturdays. Afternoon and evening hours see increased delays, reflecting the buildup of earlier disruptions. Flight distance influences departure delays, but not necessarily arrival times, indicating that other factors like airline-specific issues are also at play. Additionally, A strong correlation exists between departure and arrival delays, with variances in performance across airlines. Lastly, the random forest model turned out to be the most accurate in predicting the departure and arrival delays.

Appendices

Appendix 1: Correlations between numerical variables:

Strong Positive Correlations:

Quarter and Month: 0.97

CRSDepTime and DepTime: 0.97

CRSDepTime and CRSArrTime: 0.73

DepTime and CRSArrTime: 0.73

DepDelay and DepDelayMinutes: 1.00

DepDelay and DepartureDelayGroups: 0.89

DepDelay and ArrDelay: 0.95

DepDelay and ArrDelayMinutes: 0.97

DepDelay and ArrivalDelayGroups: 0.84

DepDelayMinutes and DepartureDelayGroups: 0.88

DepDelayMinutes and ArrDelay: 0.95

DepDelayMinutes and ArrDelayMinutes: 0.97

DepDelayMinutes and ArrivalDelayGroups: 0.82

DepDel15 and DepartureDelayGroups: 0.75

DepDel15 and ArrDel15: 0.74

DepartureDelayGroups and ArrDelay: 0.85

DepartureDelayGroups and ArrDelayMinutes: 0.84

DepartureDelayGroups and ArrivalDelayGroups: 0.91

CRSArrTime and ArrTime: 0.86

ArrDelay and ArrDelayMinutes: 0.98

ArrDelay and ArrivalDelayGroups: 0.91

ArrDelayMinutes and ArrivalDelayGroups: 0.86

ArrDel15 and ArrivalDelayGroups: 0.75

CRSElapsedTime and ActualElapsedTime: 0.98

CRSElapsedTime and Distance: 0.98

CRSElapsedTime and DistanceGroup: 0.97

ActualElapsedTime and Distance: 0.97

ActualElapsedTime and DistanceGroup: 0.96

Distance and DistanceGroup: 0.99

Strong Negative Correlations:

Appendix 2: The accuracy of Regression model.

Accuracy	Departure Delay	Arrival Delay
MAE	14229020283.027243 mins	13213920797.694628 mins
MSE	5.7161336475632705e+22	3.2887679338376612e+22

Appendix 3: The accuracy of Decision Tree model.

Accuracy	Departure Delay	Arrival Delay
MAE	3.491 minutes	1.6313 minutes
MSE	267.0044	273.9907
RMSE	16.34	16.55

Appendix 4: The accuracy of Random Forest model.

Accuracy	Departure Delay	Arrival Delay
MAE	2.4776 minutes	1.1896 minutes
MSE	149.8350	118.4285
RMSE	12.24	10.88

References

"IATA Code." Pegasus Airlines, n.d., https://www.flypgs.com/en/travel-glossary/iata-code#:~:text=It%20stands%20for%20the%20International,be%20a%20member%20of%20IATA..

"List of airline codes." Wikipedia, Wikimedia Foundation, 15 Oct. 2023, https://en.wikipedia.org/wiki/List_of_airline_codes.

"On-Time Performance Data." Bureau of Transportation Statistics, n.d., https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp?20=E.

Fox, A. (2022, October 27). *The airlines with the most delays this year, according to the Bureau of Transportation Statistics*. Travel + Leisure. https://www.travelandleisure.com/most-delayed-airlines-2021-2022-6814429