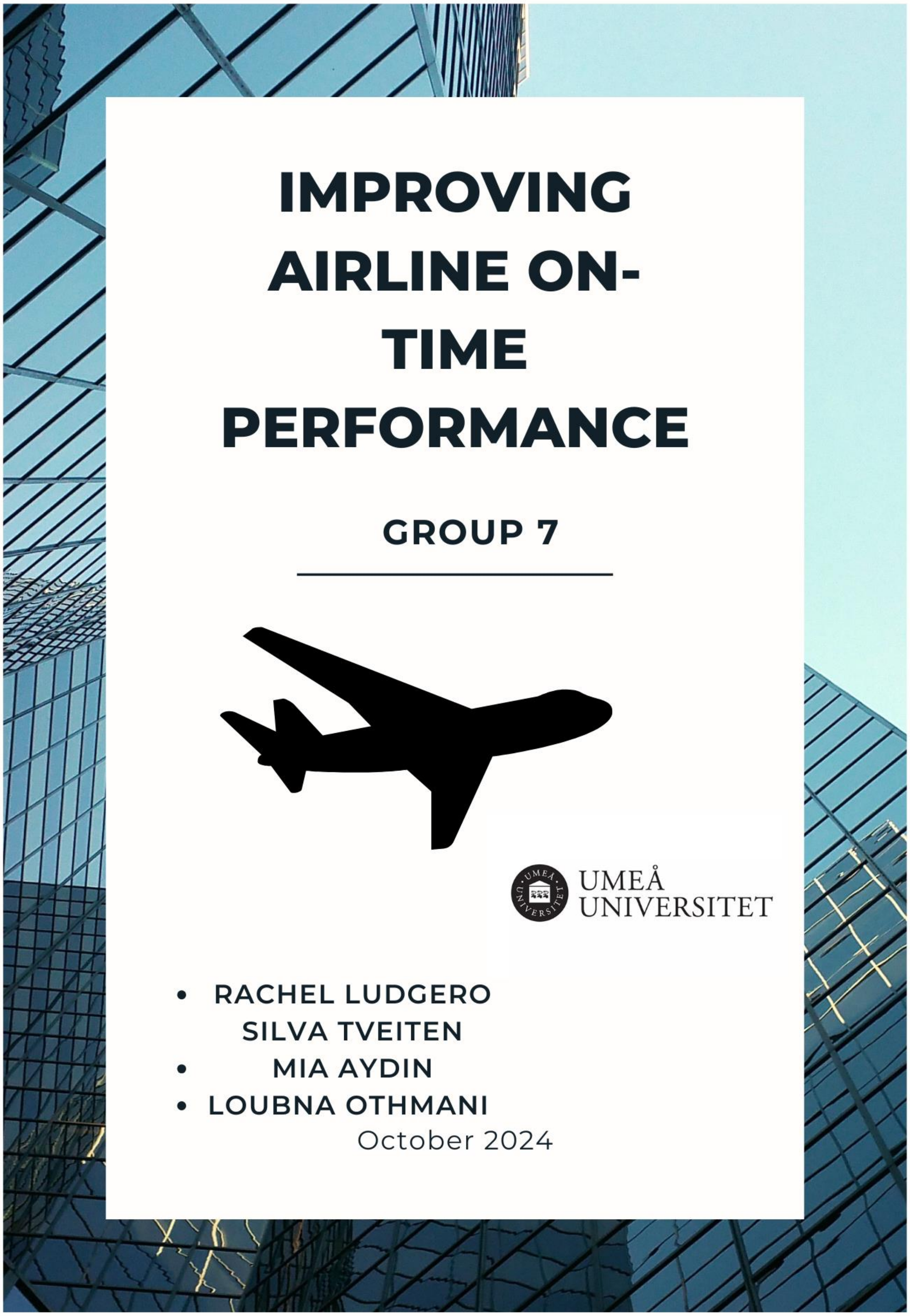


IMPROVING AIRLINE ON-TIME PERFORMANCE

GROUP 7

UMEÅ UNIVERSITET

- RACHEL LUDGERO SILVA TVEITEN
- MIA AYDIN
- LOUBNA OTHMANI

October 2024

Introduction

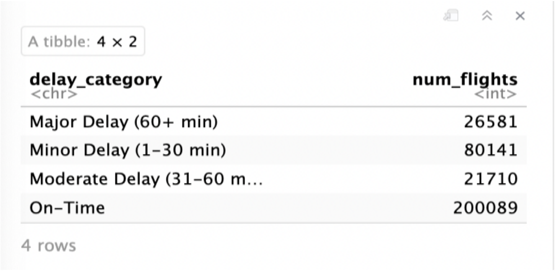
As our case study, we decided to work on the “flights" dataset from the nycflights13 package which provides data on all flights departing from New York City´s three major airports in 2013. It includes information such as departure and arrival times, delays, carrier details and more. Exploring the flights dataset allows us to delve into various aspects of air travel from NYC in 2013. By analyzing departures and arrivals delays, we can identify trends and factors contributing to flight punctuality. The dataset also offers insights into the performance of different airlines and the impact of external factors such as weather conditions, technical and operational issues.

In this project, we will explore flight data by first summarizing departure and arrival delays. There are 4 152 200 departures delays and 2 257 174 arrivals delays in 2013. We will categorize delays as significant (more than 1 hour) or reasonable (1-30 mins). Next, we will have some descriptive analysis which will include the month, day of week, time of the day and airline with the most delays. After analyzing these patterns, we will examine correlations between variables such as airtime, distance, and hour, then draw a regression model to find the best-fit model. Finally, we will predict a model to prevent delays and provide recommendations to improve airline performance.

Descriptive Analysis

Primarily, it is crucial to know how important the delay was. In order to quantify these delays, we categorize them as seen below in Table 1.

Table 1: Flight Delay Summary



A tibble: 4 × 2

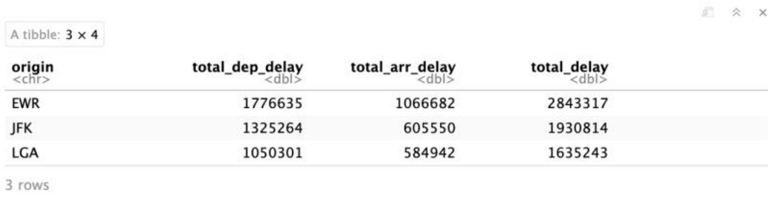
delay_category <chr>	num_flights <int>
Major Delay (60+ min)	26581
Minor Delay (1–30 min)	80141
Moderate Delay (31–60 m…	21710
On–Time	200089

4 rows

Flights are classified as having a major delay if the delay exceeds 60 minutes, a moderate delay if it ranges from 31 to 60 minutes, and a minor delay if it falls between 1 and 30 minutes. Out of 336 776 flights, 26 581 experienced delays of over an hour while 200 089 flights were on time, representing about 59%. This illustrates that although the majority of flights are relatively punctual, there is still considerable room for improvement, given that 40% of flights experienced delays.

Furthermore, it is important to see which of the airports has the most delays, which we can see in the table below.

Table 2: Airport with the most delays



A tibble: 3 × 4

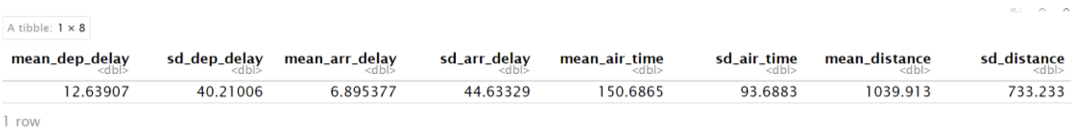
origin	total_dep_delay	total_arr_delay	total_delay
<chr>	<dbl>	<dbl>	<dbl>
EWR	1776635	1066682	2843317
JFK	1325264	605550	1930814
LGA	1050301	584942	1635243

3 rows

As the table indicates, Newark (EWR) airport experiences the most delays. This can be attributed to several factors such as weather conditions and the lack of parallel runways, which can significantly delay flights.

The table below summarizes key statistics on the average delays of flights, highlighting important metrics (mean and standard deviation) to understand the overall performance and reliability of air travel.

Table 3: Key statistic about the average delays



A tibble: 1 × 8

mean_dep_delay	sd_dep_delay	mean_arr_delay	sd_arr_delay	mean_air_time	sd_air_time	mean_distance	sd_distance
<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
12.63907	40.21006	6.895377	44.63329	150.6865	93.6883	1039.913	733.233

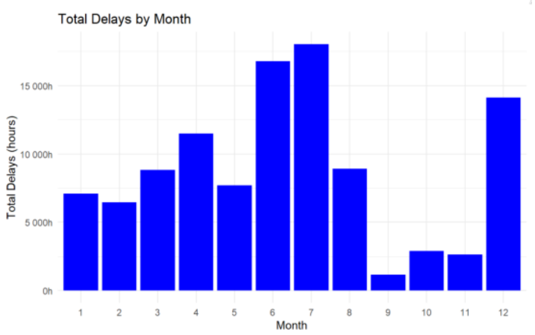
1 row

As the table indicates, the average departure delay time across all flights is 12.63 minutes and the average arrival delay is 6.89 minutes. The average distance is 1039 miles (about 1672.11 km), and the average airtime is 150 minutes (about 5 hours).

Delay Analysis

After reviewing a snapshot of our data, it is essential to analyze how delays fluctuate over time. This will help us identify patterns and understand the factors influencing these variations.

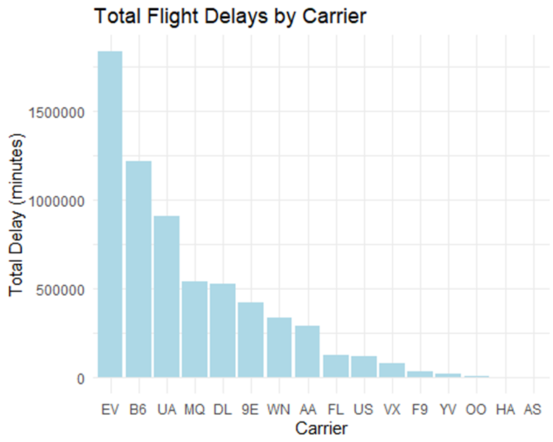
First, we will identify the month with the highest hours of delays. This analysis will help us determine any seasonal trends and how delays fluctuate throughout the year.



Graph 1: Total delays (in hours) by month

As the graph shows, June and July have the most delays, suggesting a seasonal pattern in flight delays. Moreover, June and July are peak travel months with many people flying for vacations which can overload the system leading to delays. On the other hand, September has fewer delays, suggesting that there are fewer travelers which means less disruptions compared to the busy summer months. It also indicates a more efficient period for travel and potentially smoother operations for airlines.

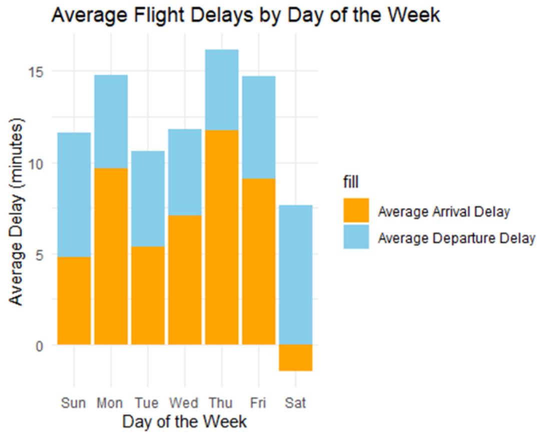
Another point to consider when it comes to delays is which airline is the most likely to be late.



Graph 2: Total Delays (in minutes) by airline

As we can see, EV is by far the highest carrier with the highest total delays (in minutes) reflecting significant operational challenges (weather, technical issues, airport infrastructures, etc.). However, AS airlines has the lowest total delays (in minutes), indicating more efficient management and smoother operations.

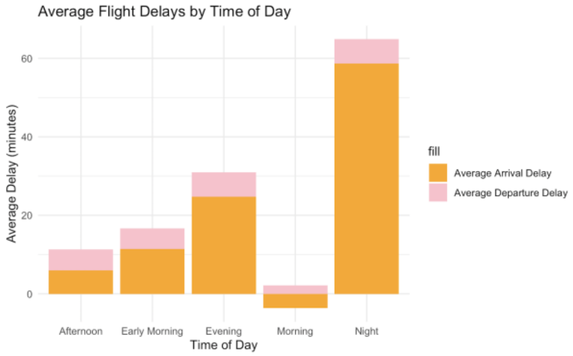
Additionally, after identifying the month with the most delays, it is also important to examine delays by day of the week. It will identify the optimal days for travel against those with the highest delays.



Graph 3: Average flight delay (in minutes) by day of the week

By looking at the bar for each day, we can identify which days of the week experience the highest and the lowest delays. For example, the bars for Thursday and Friday are higher, it could suggest that these days have more delays, possibly due to higher travel as people travel for the weekend or travel for work, for example. On the other hand, the average arrival delays for Saturday are below 0, suggesting that on average, flights are arriving ahead of schedule which means that the flights are landing before their scheduled arrival time.

Presented below are the distribution of flight delays across various times of the day, showing us peak hours for delays and fluid periods for travel.



Graph 4: Average delay (in minutes) by time of the day

The graph shows a comparison between departure and arrival days for each time of the day. If the bar for average departure delay is taller than that for average arrival delay, it suggests that flights tend to depart late more frequently than they arrive late and vice versa. The average delay during the night is high, it might indicate late-night congestion or operational challenges affecting these flights. On the other hand, the average delay during the morning is relatively low and specifically the average arrival delay is below 0 (negative delay) indicating that flights tend to arrive earlier than scheduled in the morning. A negative delay indicates that the flight arrived ahead of time which could be explained for several reasons such as morning flights benefit from reduced air traffic (less congestion) as well as the fact that morning flights are the first for both aircrafts and crews, therefore they are less likely to be affected by delays from earlier flights.

Regression Model

Then we draw a simple linear regression to examine the relationship between arrival delays and distance:

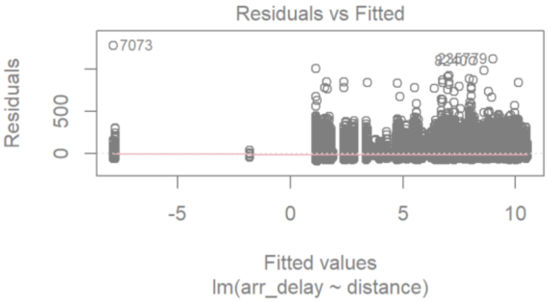
$$\text{Arrival Delay} = 10.82 - 0.003 * \text{Distance}$$

H0: There is no relationship between arrival delay and distance

HA: There is a relationship between arrival delay and distance

Where –0.003 suggests that there is a negative relationship between distance and arrival delays. When the flight's distance increases by 1 mile, the expected arrival delay decreases by about 0.003

minutes. Also, the p-value of distance is 2.2e-16 which is an extremely small p-value indicating that there is a statistically significant relationship between distance and arrival delay. Furthermore, the R-squared is 0.003 which is extremely low and highlights the absence of more variables. Since the p-value is much smaller than the typical threshold of 0.05, we reject the null hypothesis. When plotting the fitted values against the residuals, we obtain this plot:



Plot 1: Residuals VS Fitted Values for the simple linear model.

As the plot illustrates, the residuals are concentrated around 0 for most of the fitted values but there are a few extreme outliers (above 1000). In terms of homoscedasticity, the variance of the residuals is not constant across all values which suggests that one of the key assumptions for a linear model is not fulfill. Therefore, this plot shows that there is heteroscedasticity and to find our best fit model by adding more predictors of arrival delays (airtime, departure delay, etc.)

When adding departure delay and airtime to our model it gives us:

Arrival Delay = -15.91 - 0.08 * distance + 1.01* dep_delay + 0.68* air_time

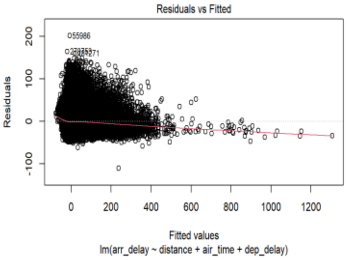
H0: One of the predictors influence arrival delays

HA: No predictors influence arrival delays

Where –0.08 suggests a negative relationship between distance and arrival delays, once again. However, there is a strong relationship between arrival delay and departure delay, if the flight leaves late, it is likely to arrive late. For every additional minute of airtime, the arrival delay increases by 0.68 minutes which implies that longer flights, in terms of airtime, tend to have a higher arrival delay due to weather conditions for instance.

When looking at the statistics of this model, we have an adjusted R-Squared of 0.8773 which indicates that approximately 87.73% of the variability in the arrival delay can be explained by the variables in our model (either distance, departure time or airtime). This is a relatively high R-Squared that could underline that our model fits the data well. In this case, the p-value is still the same which is less than common significance levels (0.05), we would reject the null hypothesis. This means we have convincing evidence to conclude that at least one of the predictor variables significantly influences the arrival delay.

The plot 2 below shows a skewed relationship to the right. There are major outliers that influence the line. For homoscedasticity, the variance does not seem to be constant for all values. Therefore, logistic regression will be analyzed and used.



Plot 2: Residuals VS Fitted Values for multiple regression model

Table 4: Logistic regression

Model	AIC	Log likelihood
Model 1: Distance as explanatory variable	442 011	-221 003.5
Model 2: Distance + Dep_delay + Air_time as explanatory variables	255 141	-127 566.4
Model 3: Dep_delay + air_time + distance + Origin as explanatory variables	254 769	-127 378.6

In logistic regression, the key statistics when it comes to choosing the better fit model, we look at the AIC criterion and the likelihood. As the table shows above, in terms of AIC the better model is the one with the lower value which is model 3. on the other hand, when it comes to log likelihood, we chose the one with the higher value which is the model 3.

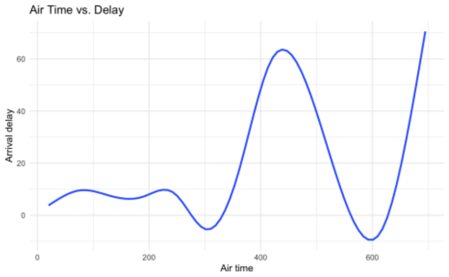
Therefore, our best fit model is:

Arrival Delay = Dep_delay + Air_time+ Distance + Origin

Correlation Analysis

In the nycflights13 dataset, understanding the relationship between key variables can provide valuable insights into flights performances and delay patterns. Correlation analysis helps us measure the relationship between different variables.

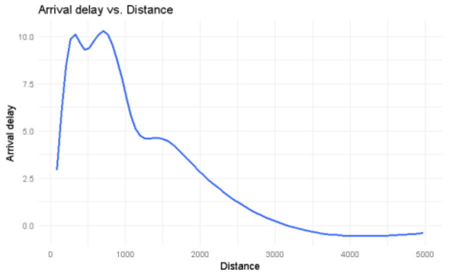
First, we looked at the correlation between the time spend the air and the arrival delay, as shown below:



Graph 5: Time spend in air vs arrival delay

In the above graph we are analyzing the numerical variables of arrival delay (response) and airtime (explanatory)which we can observe as arrival delay increases so does the airtime. The airplane will spend more time in the air, and we have drops in the arrival delay variable. However, we cannot observe a clear specific pattern with these variables. The correlation between these two variables came up to a value of -0.03529709. The value indicates a non-linear relationship and a number remarkably close to 0 where we have a situation of no correlation.

Secondly, we look at the relationship between arrival delay and distance to see if there are any patterns.



Graph 6: Arrival delay vs distance

In the above graph we can observe a very weak relationship between the arrival delay variable on the y-axis and distance on the x-axis. Again, we have a non-linear relationship, but there is a negative and no correlation to the dataset. The correlation we observed is approximately -0.06186776, being close to 0 in this scenario. Additionally, since we have a non-linear relationship, we cannot say that distance influences arrival delays of flights. The y-axis the arrival delay is measured in minutes, and the distance x-axis is measured in miles. By observing the graph arrival delay increases when we have shorter distances, however when distance increases the arrival delay tends to be lower. The reason for this could be that an airline will use more fuel and use more speed when travelling to longer destinations.

Predictive Modelling

We used the predict () function to generate predictions for our model. Firstly, we want to predict the predicted arrival delays (in minutes) for the distances (in miles), and we decided to take 500, 1000, 1500 and 2000.

Arrival Delay = 10.82 - 0.003 * 500 = 8.95 minutes

Arrival Delay = 10.82 - 0.003 * 1000= 7.08 minutes

Arrival Delay = 10.82 - 0.003 * 1500 =5.20 minutes

Arrival Delay = 10.82 - 0.003 * 2000 = 3.32 minutes

Here, we see that as the distance increases, the predicted arrival delay decreases which aligns with the distance coefficient of our model (-0.08).

Secondly, we want to predict the predicted arrival of (in minutes) for the distance (in miles), departure delays (in minutes) and the airtime (in minutes). We use distances of 500 and 1000 miles (about 1609.34 km), departure delays of 10 and 20 minutes and airtime of 50 and 100 minutes (about 3 hours).

Arrival Delay = -15.91942 -0.08919* distance + 1.0195*dep_delay +0.68*air_time

Arrival Delay = -15.91942 -0.08919* 500 + 1.0195*10+0.68*50 = -15.96983

Arrival Delay = -15.91942 -0.08919* 1000 + 1.0195*20+0.68*100 = -16.02025

For first prediction, with a distance of 500 miles, a departure delay of 10 minutes and airtime of 50 minutes, the model estimates an arrival delay of -15.91 minutes. This translates to the flight arriving 16.36 minutes early. For the second prediction, with 1000 miles (about 1609.34 km), a departure delay of 20 minutes and airtime of 100 minutes (about 3 hours), the model estimates an arrival delay of –16.02025. This means the flight is expected to arrive about 16.02 minutes early.

Despite the departure delay and airtime increase, the negative coefficient for distance (-0.08) plays a vital role in decreasing the predicted arrival as the distance increases. The negative predicted arrival delay means that the flight is expected to arrive earlier than scheduled. From both predictions, the flights are expected to arrive approximately 16 minutes early. The dominant factor is the variable distance and its negative coefficient, which leads to a decrease in predicted delays and results in earlier arrivals as the distance increases.

Recommendations

There could be many more underlying factors that would lead to delays that are not exactly presented in the data, for example, we do not exactly know how many employees work, or how many passengers are on each flight.

Recommendations for improving on-time performance. Enough tools as well as equipment, and knowledgeable employees to make sure to control the flights successfully within the airports, as well as importance of using technology and communication (Bowen 2023). Monitoring technologies is also an effective way of reducing delays, which gives the employees time to come up with a plan for any upcoming delays (Lisbona & Smale 2022).

Conclusion

In conclusion, the report analyses the business question, which is the underlying factor for delays. The report shows various numerical and categorical variables plotted against average delay time. First, the report summarizes the flight delay summary and how much delay time makes it a major, moderate, and minor delay, and shows the airport with the most delays and concludes that EWR airport had the most delays. The analysis also summarizes key findings such as the mean and the standard deviation of departure delay, arrival delay, airtime, and distance. The data also summarizes total delays by month, and concludes that July has the highest total delay, and total delays by carriers. The analysis plots 2 different residual plots, one with a smaller model and one with a larger model with more than 1 explanatory variable, and both findings were presented with a small p-value, leading to the rejection of the null hypothesis. However, to find a better fit, logistic regression was used, and table 4 shows the summary. In the correlation analysis, arrival delay is plotted against airtime, and distance. Furthermore, in the predictive model to predict arrival delay different values of distance were added. Lastly, the recommendation which gives an idea about using technologies and equipment.

References:

Bowen. (2023). *Assessing the Impact of On-Time Performance for Airports.* https://blog.aci.aero/safety-and-operations/assessing-the-impact-of-on-time-performance-for-airports/ (Accessed 2024-10-21).

Lisbona. N, Smale. W. (2022). *The airport tech helping to prevent delayed flights.* https://www.bbc.com/news/business-60228430 (Accessed 2024-10-21).