Predicting Flight Delays: A Comprehensive Analysis of US Flight Performance Data

Xiaodan Lu, Hongxin Wu, Yunhong Yang, Xinran Zhang

Abstract

The airline industry faces the challenge of providing timely services, especially during the holiday season when passenger traffic peaks. This study aims to predict flight delays in December 2022, using data from the US Transportation Department's performance dataset for on-time, delayed, or canceled flights. The project employs machine learning techniques to identify variables contributing to delays and explore patterns in flight performance across airlines, days of the week, and city pairs. Complementary to predictive modeling, Tableau is used to create dashboards and other visualization tools to facilitate a comprehensive understanding of the data for both technical and non-technical audiences. The project offers insights that can help airlines optimize their operations, enhance the travel experience for passengers, and promote efficient resource allocation during the busy holiday season.

[Keywords] Flight Delays, Machine Learning, Data Visualization, Airline Industry, Predictive Modeling, Tableau, Holiday Season, Travel Experience, Resource Allocation, Performance Optimization

Introduction

Flight delays and cancellations cause significant inconvenience for passengers and disrupt the smooth functioning of the airline industry, particularly during the busy holiday season. In December, when many people take time off for Christmas vacations, the timely performance of flights becomes a critical factor in ensuring a positive travel experience. To address this issue, the project focuses on predicting flight delays using real data from the US Transportation Department's performance dataset for on-time, delayed, or canceled flights in December 2022.

The project employs a combination of machine learning models to identify key variables that influence flight performance and predict delays with the highest accuracy. The project is structured around the prediction, process, and results, with an emphasis on understanding the data and making it accessible to both technical and non-technical audiences. To achieve this, leverage data visualization tools such as Tableau are used to create informative dashboards that clearly convey patterns in flight performance across airlines, days of the week, and city pairs.

The study aims to provide valuable insights into the airline industry, enabling better decision-making and resource allocation during the holiday season. By identifying the factors that contribute to flight delays, airlines can implement targeted strategies to minimize disruptions and improve the overall travel experience for passengers. Furthermore, the project showcases the potential of combining machine learning and data visualization to address complex problems and make data-driven decisions in a rapidly evolving industry.

About the data

This project utilizes data from the U.S. Department of Transportation, which includes information about all flights that have taken place in the United States since 1987 up until the present day. The data has undergone rigorous cleaning processes, including filtering out insignificant variables and removing any missing values. The final selection of data pertains to aircraft flights that took place in December 2022 and contain approximately 576,827 rows of information across 48 variables. These variables cover essential details such as the aircraft's origin airport, destination, date of the flight, taxi out time, taxi in time, whether it arrived on time or not, and the number of minutes of delay it experienced (Figure 1 provides a basic explanation of the proper nouns for aircraft routes).

In total, the dataset includes information about 576,827 flights that had 355 different destinations. On average, the departure delay for each flight was approximately 20.81 minutes, while the arrival delay averaged 20.89 minutes. Furthermore, it is noteworthy that 5.385% of all flights in the dataset were canceled.

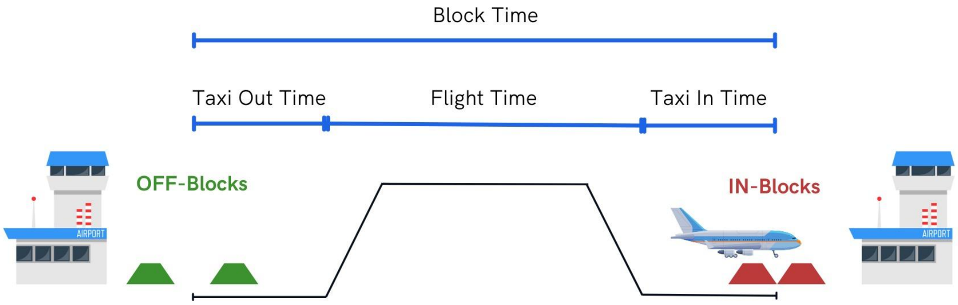

Fig. 1: Taxi Out and Taxi In variables Explanation

Exploratory Data Analysis

During the exploratory data analysis phase, the project delves deep into the dataset to understand the underlying patterns, trends, and relationships between different variables. Various visualizations are created using Python's Plotly library and Tableau to aid in this process. Key findings from the EDA include:

1. Flight delay patterns: Identified days of the week and specific time periods with a higher incidence of flight delays. This information can be helpful in understanding the temporal factors contributing to delays. Based on the heatmap (Figure 9) analysis, it can be observed that Hawaii Airlines (HA) tends to experience significant delays on Mondays, while Delta Airlines (DL) consistently maintains the lowest average delay time throughout the entire week. Concurrently, Frontier Airlines (F9) encounters substantial flight delays over the weekend. Therefore, it is advisable to avoid traveling with Frontier Airlines during weekends for a more punctual travel experience.
2. Airline performance: Analyzed the performance of different airlines with respect to flight delays. Discovered variations in delay patterns across carriers, which can be indicative of differences in operational efficiency. The findings indicate that there are indeed variations in airline performance. Specifically, Figure 8 illustrates that Frontier Airlines (F9) has a

higher average departure and arrival time compared to other airlines. Thus, with such a long delay time, F9 Airlines need to make changes to reduce the delay time.

3. Airport Performance: Analyze the relationship between different airports and aircraft delays to determine if some airports need to improve efficiency. According to a packed bubble chart (Figure 5), Eastern Sierra Regional Airport (BIH) in California has the longest average delay of 96.28 minutes. Therefore, it is imperative for BIH airport to improve efficiency given the significant delay.
4. Distance and delay: Carried out the relationship between flight distance and delay duration, uncovering insights into how distance may impact the severity of delays. From the scatter plot (Figure 7) comparing flight distances and delays, it becomes evident that long-distance flights tend to experience fewer delays compared to short-distance flights. The most significant delays are observed within the 0-1000 mile flight distance range.
5. Reasons for delays: Examined the various reasons for flight delays, such as weather, mechanical issues, and air traffic control restrictions. This analysis can help determine the most common causes of delays and inform potential mitigation strategies. By calculating the total delay time for each of the five types of delays (weather delays, safety delays, NAS delays, aircraft delays, and carrier delays), it is discovered that the highest frequency of delays is caused by late aircraft and carrier issues (Figure 4).
6. State-wise average delay: Explored the geographical distribution of average flight delays by state, Visualized this data using a choropleth map, revealing regional patterns and potential factors contributing to delays in specific areas. The choropleth map (Figure 6) displaying state-wise averages indicates that the Midwest's northern states encounter lengthier average delays, whereas states along the East Coast tend to have shorter delay durations on average.

The insights gained from the EDA phase play a critical role in guiding the subsequent steps of the project, including feature selection and model selection.

Methodology

To predict flight delays, there are two tendencies to construct the proposed model, which are classification and regression methods. In the classification method, multinomial logistic regression, k-nearest neighbors, and random forest are attempted, and “DepartureDelayGroups” is chosen as the dependent variable. The “DepartureDelayGroup” is the departure delay interval every 15 minutes, which is from -15 to 180. The negative number means that the airplane departs earlier than the scheduled time. In the regression method, linear regression, ridge regression, LASSO regression, decision tree, and boosting are attempted, and the airplane departure delay minute is chosen as the dependent variable. This dependent variable is calculating the difference in minutes between the scheduled and actual departure time.

After pre-training and testing all models, none of the classification models performs well in predicting the delay groups of flights. The random forest classifier is unable to run due to the less capacity of the computer. The best model in this trend is multinomial-logistic regression, which only has a 0.52 accurate rate in the prediction. Changing parameters in each model also does not improve the performance of prediction. The main possible problem to cause this situation is that many factors in the data set are highly unbalanced. For example, the numbers of airline carriers’ records are different. There are 138,009 American Airlines records in the data set, but there are only 6623 records about Hawaii Airlines. The highly-unbalanced factors may be the main factor influencing the performance of classification models.

Multinomial-Logistic	KNN	CNN	Random Forest
52%	43%	16%	/

Table 1: Accuracy of Classification Models

Previewing the output of the regression models in Figure 2, Boosting is the outstanding regression model among all regression models. The root mean square error(RMSE) of Boosting is much lower than the other regression models. Therefore, Boosting becomes the main proposed model in the further analysis.

For the boosting model, Gaussian distribution is used to estimate the result with the parameter "n.trees" to be 100, the "interaction.depth" to be 4, the "shrinkage" to be 0.2, and "verbose" to be False. With all these parameters together to train the model, the RMSE for this settlement is 28.65. From the result, the boosting seems to perform better and can be used in future analysis.

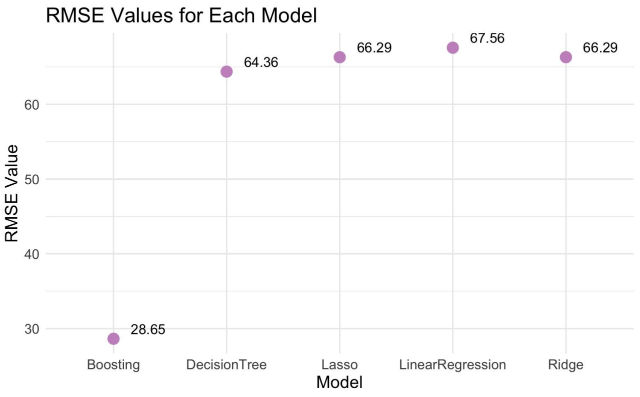

Fig. 2: RMSE of Regression Model

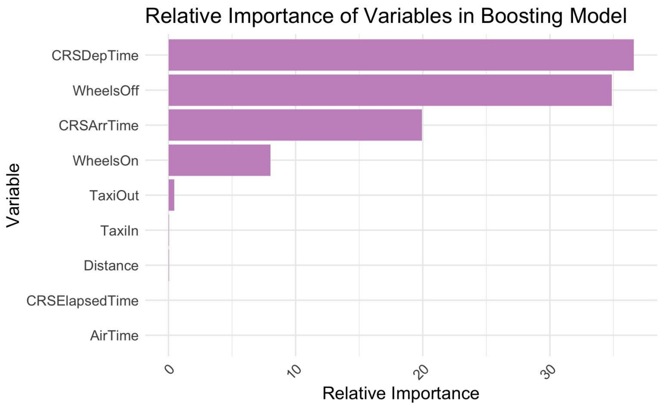

Fig. 3: Relative Importance in Boosting Model

Results

Glancing all models constructed by classification and regression methods, the regression method is the deciding part of the prediction. The boosting with regression methods performs better and is used as the main model. The following figure is the relative importance of the final model. From the result of the boosting regression model, the variables "CRSDepTime", "WheelsOff", "CRSArrTime", and "WheelsOn" are the four most important variables that contribute to predicting departure delay minutes. Thus, the original estimated departure and arrival time of a flight can greatly influence how long the flight would likely be delayed. Also, the duration of wheels on and wheels off can respectively affect the delay of a flight. When booking flights for trips, these can all be factors that need to be taken into consideration.

Discussion

The main purpose of this project is to have a brief overview of the causes of flight delays. The advantage of this project is containing several pre-training models with diverse methods. Working through all methods, boosting regression is the optimal solution to train this kind of data set. In further analysis and advanced model training, balancing factors is one of the significant points to improve model performance. Another possible way for improvement is to subset the data with certain weights to make factors balanced. For classification models, constructing subsets of factors and applying them to each classification model is also a worthy attempt, which is quite similar to the best subset selection in regression models. To improve the performance of the boosting model, adding more factors, such as departure airport and arrival airport, may be worth trying in the future.

Conclusions

In summary, the research aims to uncover the key factors contributing to flight delays, with a particular focus on the bustling holiday season when flight demand surges. Extensive exploratory data analysis is carried out, employing a diverse range of visualization techniques to gain a comprehensive understanding of flight delay patterns and trends across various regions, airlines, and timeframes.

The analysis reveals that certain geographical areas, such as the Midwest, experience a higher frequency of flight delays compared to others, like the east coast. Moreover, the predictive models identify original estimated departure and arrival times, as well as the duration of wheels-on and wheels-off events, as critical factors in determining flight delays.

This study offers valuable insights for travelers by shedding light on the factors that influence flight delays, thereby enabling them to make more informed decisions when booking flights and organizing trips. Furthermore, the research highlights potential avenues for future investigation, such as refining model performance and incorporating supplementary factors, including departure and arrival airports. Ultimately, the findings contribute to a deeper understanding of the root causes of flight delays and may prove instrumental in assisting airlines and airports in identifying areas for improvement, enhancing the overall travel experience for passengers.

References

"Bureau of Transportation Statistics." United States Department of Transportation, https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGK&QO_fu146_anzr=b0-gvz.

Appendix

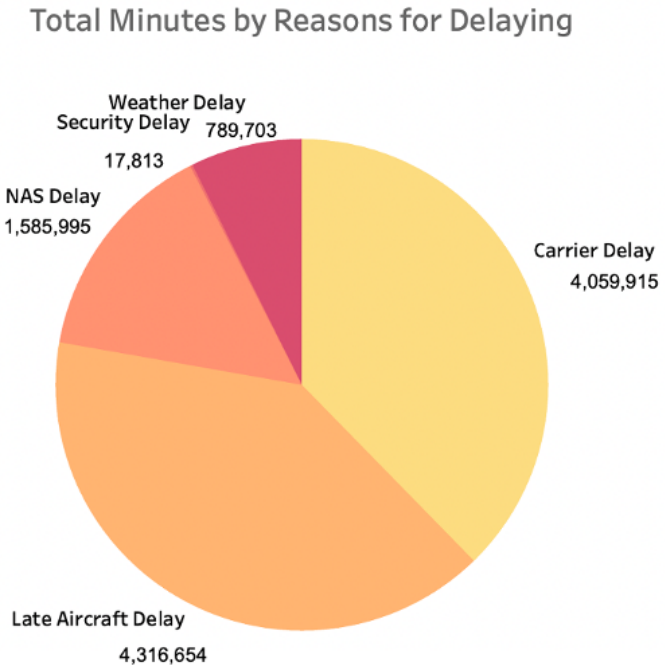

Fig. 4: Total Minutes by Reasons for Delaying

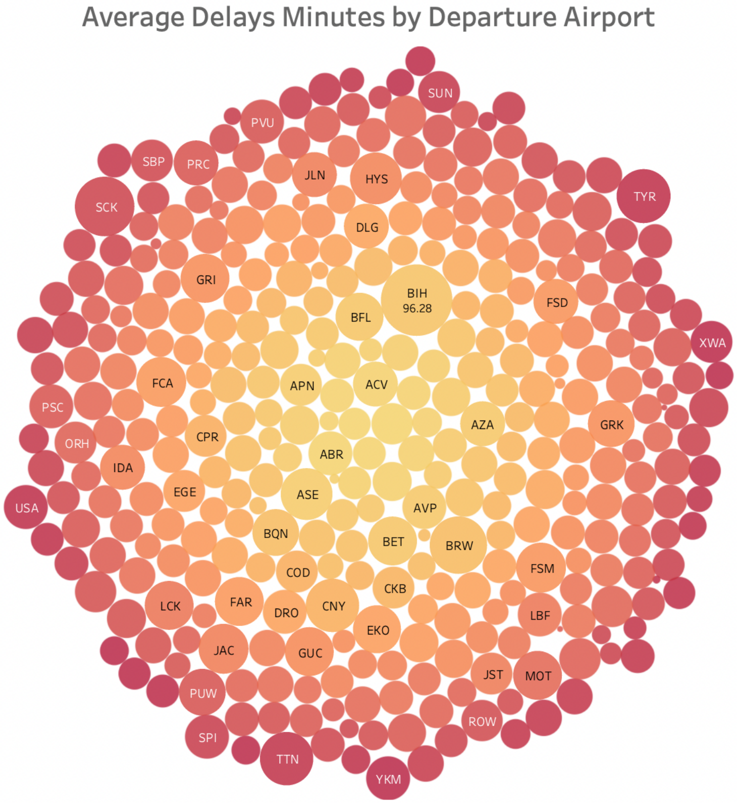

Fig. 5: Average Delays Minutes by Departure Airport

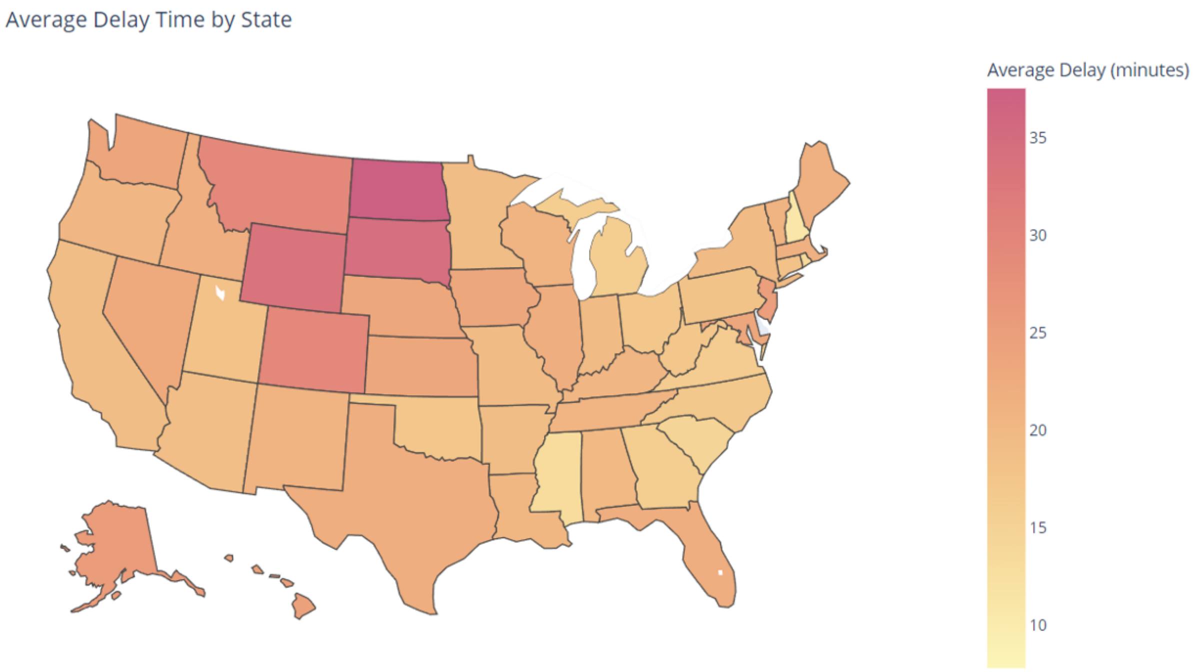

Fig. 6: Average Delay Time by State

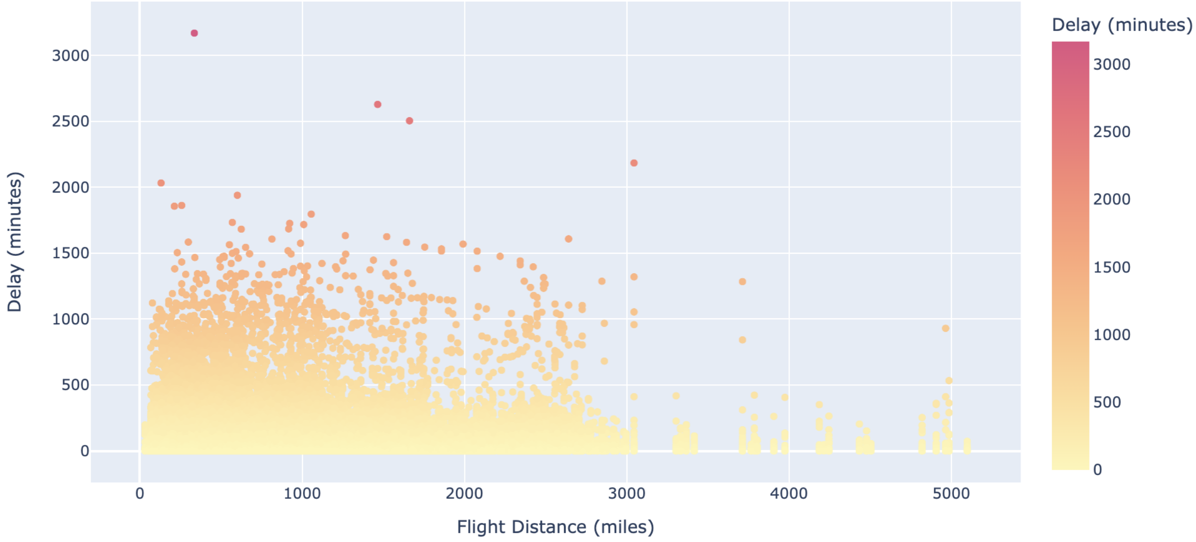

Fig. 7: Flight Distances vs. Average Delays

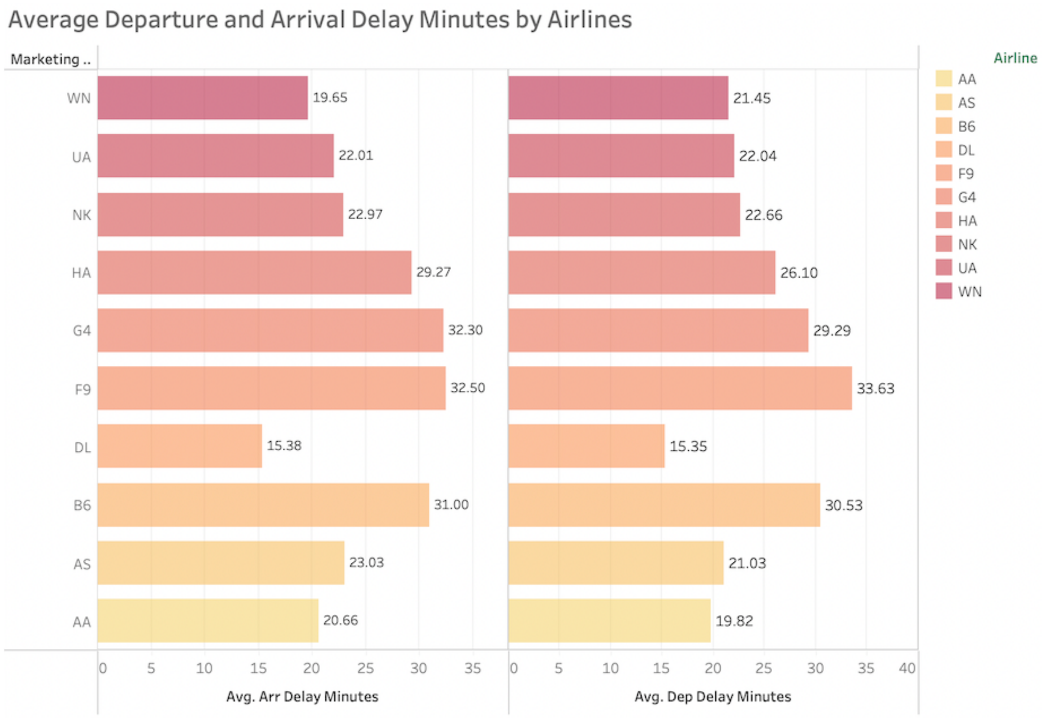

Fig. 8: Average Departure and Arrival Delay Minutes by Departure Airport

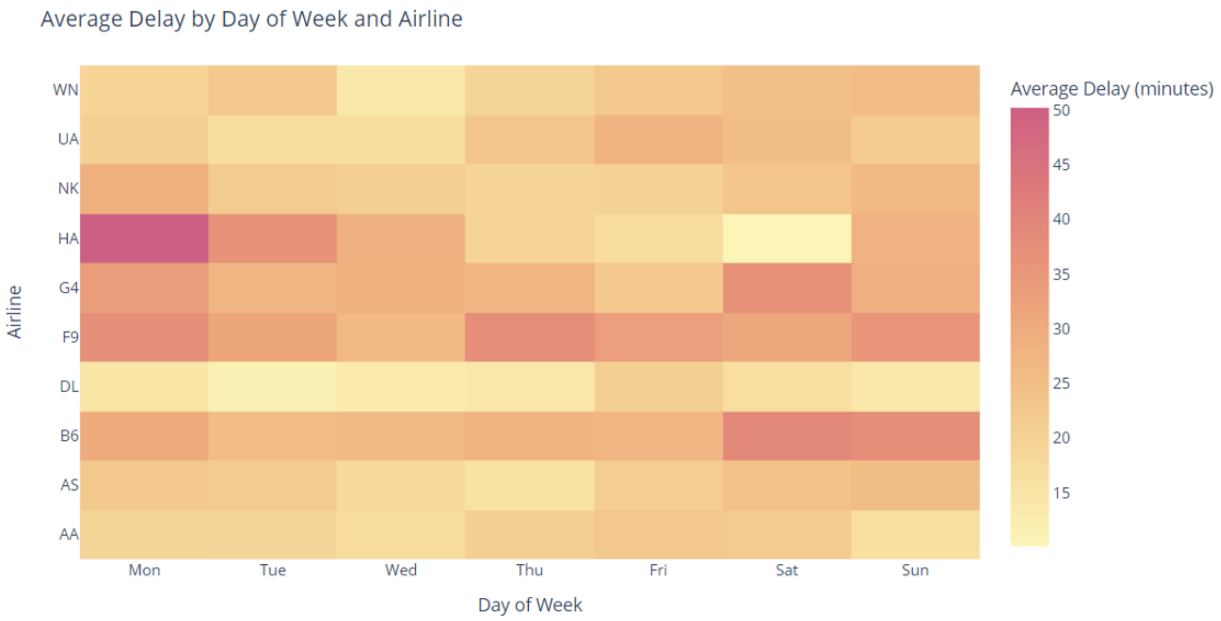

Fig. 9: Average Delay by Day of Week and Airline