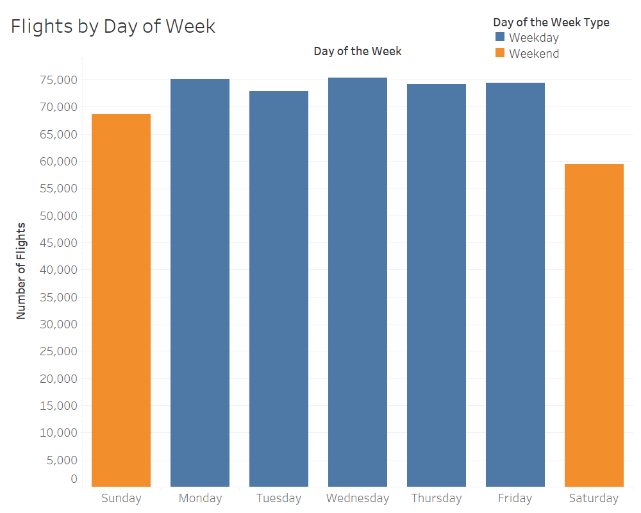
Blog Post

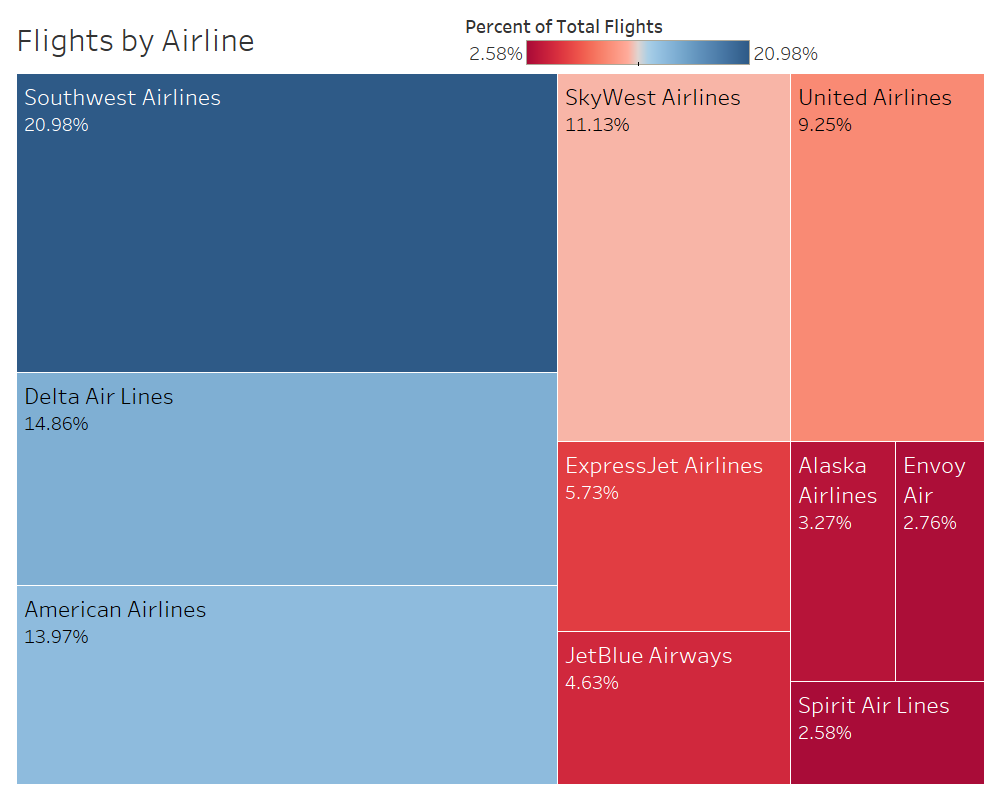
Everybody has had at least one bad experience at an airport. It doesn’t matter if you’re a seasoned professional taking your regular morning flight to Los Angeles or a desperate college student trying to board the last flight back to school after break; the airport will, at any given time, be prepared to ruin your day. Personally, I love airports. The energy, the purpose, the excitement, the atmosphere: airports are truly unlike any other place in the world in that regard. But even I have been cursed by their fickle ways, and I thought it would be interesting to study these occurrences to inform myself and lessen the chances of unfortunate events when flying. I took five years of Bureau of Transportation Statistics air traffic data and analyzed various facets of the flight experience to formulate the best strategies for air travel. Some of my results are mostly expected due to the nature of the business, but I gleaned a few important and intriguing insights that could help shape my and your travel activity in the future. Perhaps the conclusion of the greatest impact: despite holding 15% of all flights in my dataset, second-most among all airlines, Delta Air Lines is among the best carriers at pushing out scheduled flights on time.

The data from the BTS carries individual flight data for every reporting flight carrier from 1987 to May 2020. Information included ranges from arrival and departure statistics to demographic and geographic data to airline and plane descriptors. The entire dataset includes almost two hundred million records. To complete my analysis, I sampled the middle month from each of the four quarters (February, May, August, November) for five years from 2015 to 2019. I figured this would allow me to avoid the anomalies stemming from the impact of COVID – 19 on the airline industry and let me investigate the data from different periods of the year. My final dataset included 500,000 observations, with 125,000 points from each quarter.

I conducted some exploratory analysis on my sample, and a few discoveries stood out. Figure 1 below shows number of flights by day of the week.



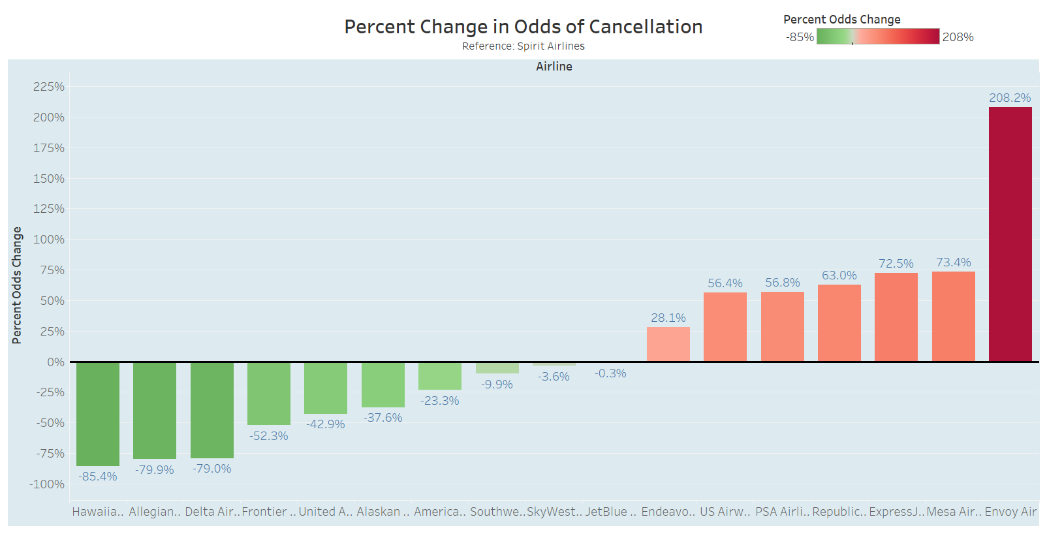
It is obvious that there is a significantly lower number of flights on weekends in the United States, especially on Saturday compared to all other days of the week. An ANOVA analysis fortified this statement, concluding that although weekdays did not have a significantly different number of flights, Saturday and Sunday were characterized by statistically significant differences both from weekdays and from each other. In addition, an exploration of flights by airline produced Figure 2 below.



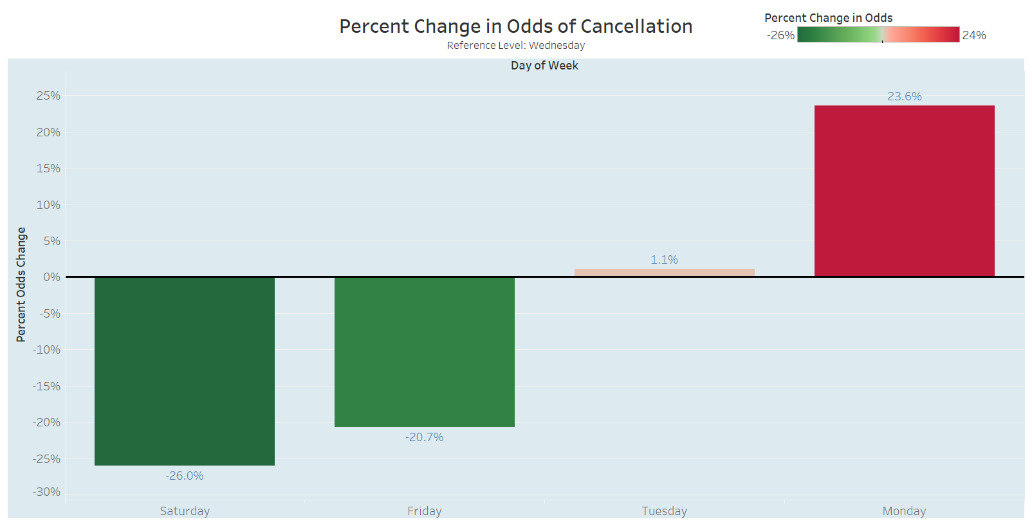
This table shows that Southwest Airlines, Delta Air Lines, and American Airlines are the most utilized airlines in the United States, although only Southwest Airlines can claim a market share of greater than 20%.

When identifying the questions I wanted to ask in this analysis, I aimed my thoughts towards queries that could result in actionable insights that could allow a person to change their behavior and optimize their air travel. Thus, I settled on three main topics. The bulk of the analysis revolved around the study of cancellations and departure delays in the dataset. In addition, I explored departures across airports in cities and/or metropolitan areas with more than one airport within an hour’s drive of each other. By looking into flight outcomes across these different characteristics, I hoped to get a better idea of what a successful or unsuccessful flight looks like in the United States.

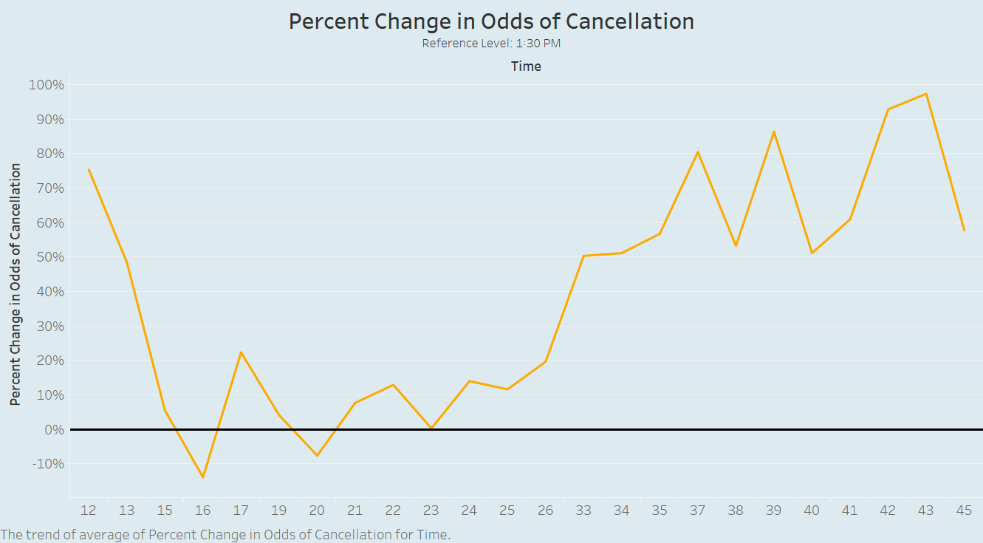
To start, I will detail the analysis of flight cancellations in the United States. In perusing the dataset for possibly relevant factors, I decided to focus on decisions that individuals make when planning their air travel. Thus, I honed in on three main predictors: scheduled departure time, departure day, and airline. My process included studying each predictor variable individually through tests of association and simple logistic regressions. It resulted that airline, departure day, and scheduled departure time all had significant associations with the cancellation dummy outcome. Logistic regression models with various significance levels were utilized to create the following odds ratio graphs for the different predictors.



This graph shows that Hawaiian Airlines, Allegiant Air, and Delta Airlines have significantly lower odds of cancelling a flight compared to Spirit Air Lines. On the other hand, Envoy Airlines have almost double the odds of cancelling a flight compared to Spirit Air Lines.

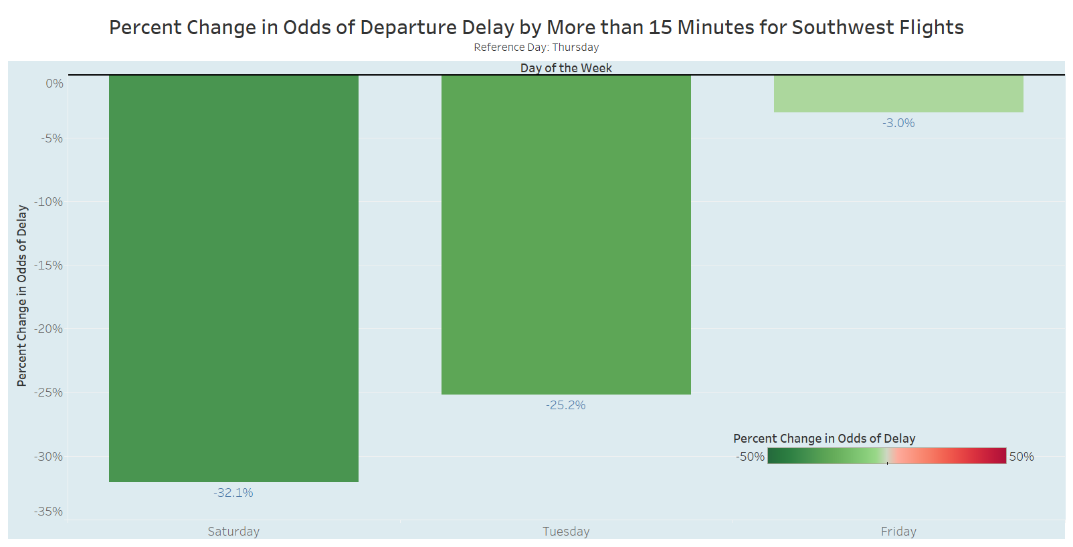


Looking now at the day of the week, the graph shows that the later in the week that you schedule your flight, the lower the odds of cancellation are. Additionally, the chart below plotting percent change in odds against departure time shows that odds of cancellation increase as the day goes on, as expected.

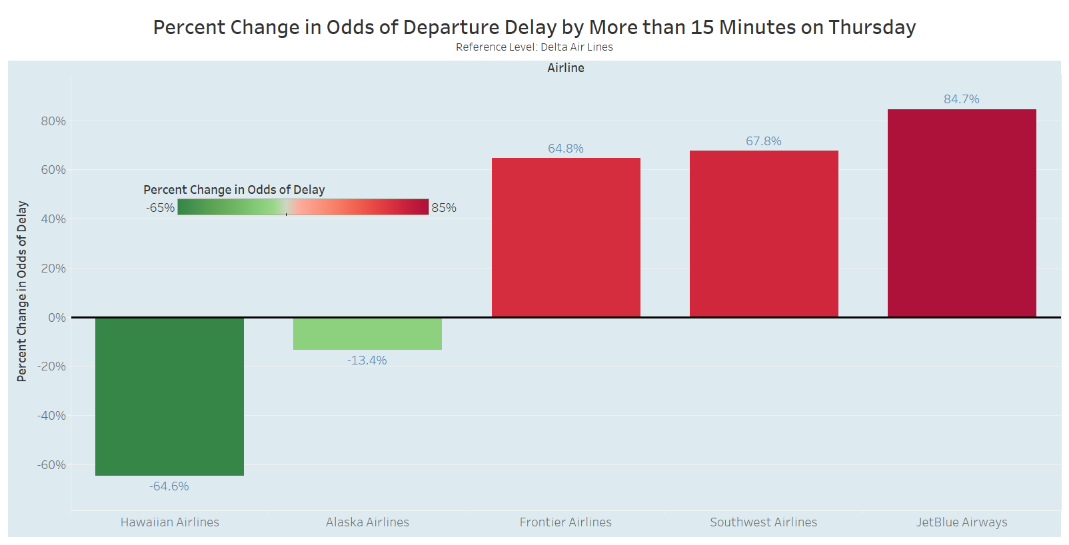


The analysis of flight cancellations in the United States results in mostly logical conclusions; it makes sense that odds of cancellation would be lower later in the week as flights are more expensive and fewer people are flying, and it follows that as the day passes and more and more flights are pushed back all over the country, the odds of cancellation of a given flight would increase. However, the airline analysis is quite intriguing, as it grants insights into the carriers that can be best relied upon to fly their scheduled flights.

Next, a similar investigation was conducted on departure delays in U.S. air traffic. The same three factors (airline, departure day, departure time) were studied together and separately in a logistic regression model with a departure time dummy indicating whether or not a flight was delayed by over 15 minutes used as the response variable. Using departure delay as the response, I was able to fit a logistic regression model with airline and departure day and their interaction effect, and I used this model to identify some interesting relationships and select relevant odds ratios. Due to the presence of significant interactions, the analysis was subset by airline and day of the week, and further effects were analyzed within those categorical variables. For instance, the interaction between Southwest Airlines and day of the week demonstrated a very low p-value, so the graph below was constructed comparing odds of departure delay, extracted from a logistic regression model on the restricted data, for Southwest flights across different days of the week.



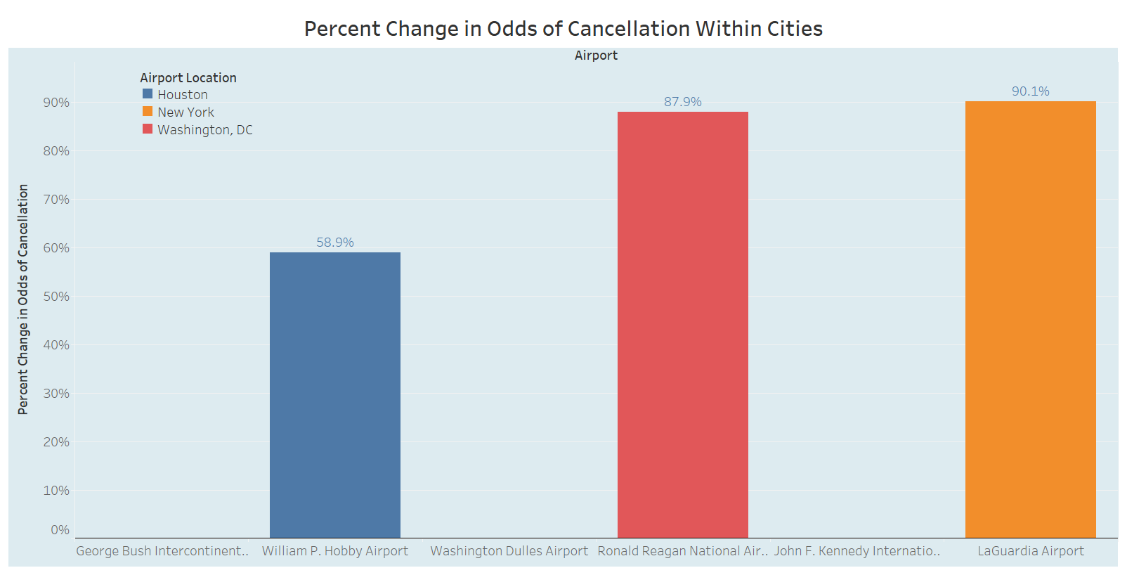
From this graph, we can conclude that Saturday and Tuesday are the most likely days for on-time departures when flying Southwest. Additionally, I studied the effect of airline on departure delay on Thursdays due to the relative prevalence of low p-values in the full logistic regression model.



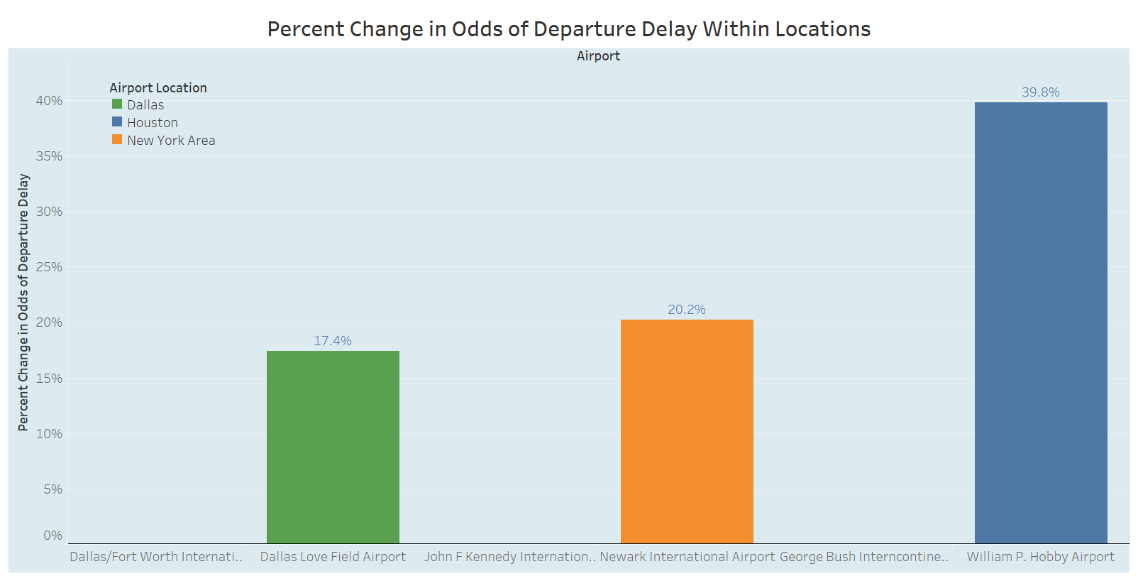
I decided to use Delta Air Lines as the reference level due to its impressive performance in the above model for cancellations and relatively large market share. It makes sense that smaller, more niche carriers such as Hawaiian Airlines and Alaska Airlines would have lower odds of delay; however, flights from Delta Air Lines are far more likely to depart on time on Thursdays than other large providers such as Frontier Airlines, Southwest Airlines, and JetBlue Airways.

Due to the presence of significant interaction effects in the full logistic model with airline and day of the week, the analysis of departure delay was slightly more restricted and less generalizable. However, as demonstrated by the two figures above, it is still possible to gather impactful conclusions regarding the best carriers to fly on specific days and best days to fly with specific carriers.

Lastly, I wanted to investigate the differences in flight outcomes for airports that reside in the same metropolitan statistical area, hoping to outline some statistical differences that could help someone decide which airport to fly out of in specific cities. I created various data tables containing only flights with origins in the same MSA and ran various tests of association and logistic regression models (when appropriate) to verify and quantify relationships. The graph below shows differences in cancellation outcomes.



We can see that certain airports in Houston, New York City, and Washington, DC are far better at keeping their schedules than others in the same areas. Specifically, flights departing from LaGuardia Airport in the New York area are 90.1% times more likely to be cancelled than flights departing from John F. Kennedy in the New York area, and flights departing from Ronald Reagan International Airport in the DC area are 87.9% more likely to be cancelled than flights departing from Washington Dulles Airport in the DC area. It is possible this phenomenon could be attributed to the volume of flights stemming from each airport; for instance, Ronald Reagan International Airport records twice as many flights as Washington Dulles in my dataset. However, there is only a five percent difference in departures from the two airports in the New York Area, so it is possible the efficiency of the airports could have an impact as well. In the same way, we look at departure delay differences across airports in the same region.



Looking at the graph above, we note that the change in odds are not quite as drastic as the change in odds for cancellations; this could be credited to the relatively high frequency of delays to outright cancellations. Dallas shows up in this graph, and it appears that departures from Dallas Love Field Airport are 17.4% more likely to be delayed, which is quite significant considering that Dallas Love pushes out about a quarter of the flights that Dallas/Fort Worth does. From the New York section of the graph, we can see that John F. Kennedy International Airport appears to be the most efficient in that area; flights from that airport are less likely to be cancelled than those at LaGuardia and less likely to be delayed than those at Newark International Airport. Lastly, it appears that William P. Hobby Airport is significantly less efficient than George Bush Intercontinental Airport, despite pushing out a third of the number of flights that the latter pushes out.

From this analysis, I discovered interesting relationships and made actionable conclusions about air traffic in the United States utilizing tests of association and logistic regressions. For example, I uncovered and quantified the quality of Delta Air Lines when compared to other players in the industry. I established that Tuesday, Wednesday and Saturday were the best days to fly to avoid those panicky moments when you’re running all over the airport trying to figure out if and when you are going to make it off the ground. I pinpointed significant differences in location efficiency among airports in large metropolitan areas such as Houston, New York, Dallas, and Washington, DC. In further analysis, it would be interesting to investigate other factors that could affect cancellations and departure delays, such as average price of a seat on the flight, which could have an important effect on both consumer decision-making and flight outcomes across different days of the week and airports. However, I hope to have enlightened you as I enlightened myself on some different facets of air travel across different dimensions. Now to figure out which flights Delta has scheduled next Thursday.