Coulton Theuer
Rohit Maramraju
GitHub
March 14, 2023

SI618: Airline Data Report

Motivations

The COVID-19 pandemic had a profound impact on the global airline industry, leading to unprecedented disruptions in air travel since its onset in early 2020[1]. Widespread travel restrictions, border closures, and lockdown measures imposed by governments across the world to curb the spread of the virus resulted in a drastic decline in demand for air travel[2]. According to the International Air Transport Association (IATA), passenger traffic decreased by 65.9% in 2020 compared to 2019, marking the sharpest decline in aviation history[3]. These obstacles forced airlines to adapt quickly by implementing stringent health and safety protocols, reducing capacities, and even ceasing operations altogether in some cases[4]. The pandemic's long-lasting effects on the airline industry continue to reshape the landscape of air travel, with implications for consumer behavior, business models, and industry recovery[5]. The COVID-19 pandemic has not only led to a significant decline in passenger traffic and airline bankruptcies but has also had numerous other consequences on the airline industry. Despite the financial stimulus provided to the airline industry to buoy them during the pandemic, the magnitude of disruption has led to the bankruptcy of 64 airline companies[6]. Given the important role of the airline industry on modern life, it is crucial to examine the sector's recovery in detail. Our objective is to evaluate the industry before the pandemic and the changes that occurred as a result of the pandemic; by analyzing airline data, we aim to understand how the pandemic has influenced delays, cancellations due to staffing challenges, and other contributing factors. Additionally, we will look to compare the stock prices of various airlines throughout this period to gain further insights into the industry's financial resilience.

[1] Statista. (2021). Impact of the coronavirus pandemic on the global aviation industry. Retrieved from https://www.statista.com/topics/6178/coronavirus-impact-on-the-aviation-industry-worldwide/
[2] Statista. (2021). Impact of the coronavirus pandemic on the global aviation industry. Retrieved from https://www.statista.com/topics/6178/coronavirus-impact-on-the-aviation-industry-worldwide/
[3] International Air Transport Association (IATA). (2021). Air passenger market analysis. https://www.iata.org/en/iata-repository/publications/economic-reports/air-passenger-monthly-analysis---december-2020/
[4] CNBC. (2020). Airlines slash flights, take other drastic measures to combat coronavirus. https://www.cnbc.com/2020/03/16/coronavirus-makes-airlines-consider-chances-for-a-halt-to-us-flights.html
[5] "COVID-19's Impact on the Global Aviation Sector | McKinsey," accessed March 20, 2023, https://www.mckinsey.com/industries/travel-logistics-and-infrastructure/our-insights/taking-stock-of-the-pandemics-impact-on-global-aviation.
[6] CNN. (2023). How the pandemic killed off 64 airlines. https://www.cnn.com/travel/article/pandemic-airline-bankruptcies/index.html

Data Sources

We will be using two datasets relating to US flight delays and US airline stock prices from 2016 to 2021 that were provided by the Ross Business+Tech Club for the 2023 Datathon Competition[7]. These datasets are also publicly available for download from Yahoo Finance, the Bureau of Transportation Statistics' website, and the Federal Aviation Administration's website[8]. More detailed descriptions of this data are shown in Figures 1 and 2.

We also have three secondary (metadata) datasets that include abbreviation/full-name pairs for airlines, abbreviation/full-name pairs for airports, and variable descriptions for the flight delay table. These datasets were also provided by the Datathon Competition, and are publicly available at the same resources mentioned above. These dataset are described in Figures 3, 4, and 5.

#	Column	Non-Null Count	Dtype
0	year	101629 non-null	int64
1	month	101629 non-null	int64
2	carrier	101629 non-null	object
3	carrier_name	101629 non-null	object
4	airport	101629 non-null	object
5	airport_name	101629 non-null	object
6	arr_flights	101457 non-null	float64
7	arr_del15	101264 non-null	float64
8	carrier_ct	101457 non-null	float64
9	weather_ct	101457 non-null	float64
10	nas_ct	101457 non-null	float64
11	security_ct	101457 non-null	float64
12	late_aircraft_ct	101457 non-null	float64
13	arr_cancelled	101457 non-null	float64
14	arr_diverted	101457 non-null	float64
15	arr_delay	101457 non-null	float64
16	carrier_delay	101457 non-null	float64
17	weather_delay	101457 non-null	float64
18	nas_delay	101457 non-null	float64
19	security_delay	101457 non-null	float64
20	late_aircraft_delay	101457 non-null	float64

Figure 1. 'Flight_Delay_2016_2021'
This data is stored in a csv file hosted in a publicly accessible dropbox folder[9]. The data covers the time period 2016 to 2021, and this table has 101629 entries and 21 columns. Additionally, there are 365 rows in this table that contain null values.

We are most interested in the 'year', 'month', 'carrier', 'airport', 'arr_flight', and '*_ct' columns. The 'year' and 'month' columns describe the date that the airport/airline data was collected. The 'carrier' column indicates the airline that was tracked. The 'airport' column indicates the airport that was tracked. The 'arr_flight' indicates the number of flights from the specified airline that arrived at the specified airport. The '*_count' columns indicate the frequency of a specific type of delay (like weather or carrier/airline issues). The '*_count' columns are decimal numbers because a single delay can be explained by multiple factors; for instance, the captain could be late (a carrier delay) and the weather could be horrible (a weather), resulting in one delay split of .5 for carrier_ct and .5 for weather_ct. These variables together describe the number of different types of delays that happened on a monthly basis at each airport for each carrier (airline). We will use this information in combination with stock return data later on.

Figure 2. 'US_Airlines_StockPrice_2016_2021'
This data is stored in a csv file hosted in a publicly accessible dropbox folder[10]. The data covers the time period 2016 to 2021, and this table has 9060 entries and 8 columns. There are no null values.
We are most interested in the 'Airline', 'Date', and 'Adj Close' columns. The 'Airline' column is the stock ticker abbreviation for

#	Column	Non-Null Count	Dtype
0	Airline	9060 non-null	object
1	Date	9060 non-null	object
2	Open	9060 non-null	float64
3	High	9060 non-null	float64
4	Low	9060 non-null	float64
5	Close	9060 non-null	float64
6	Adj Close	9060 non-null	float64
7	Volume	9060 non-null	int64

[7] "Datathon," U-M Ross Business+Tech, accessed March 18, 2023, ht
[8] "Yahoo Finance - Stock Market Live, Quotes, Business & Finance News," accessed March 18, 2023, https://finance.yahoo.com/; "Bureau of Transportation Statistics," accessed March 18, 2023, https://www.bts.gov/; "Aviation Data & Statistics | Federal Aviation Administration," accessed March 18, 2023, https://www.faa.gov/data_research/aviation_data_statistics.
[9] "4. Flight_Delay_2016_2021.Csv," accessed March 18, 2023, https://www.dropbox.com/h?preview=4.+Flight_Delay_2016_2021.csv.
[10] "2. US_Airlines_StockPrice_2016_2021.Csv," accessed March 18, 2023, https://www.dropbox.com/h?preview=2.+US_Airlines_StockPrice_2016_2021.csv.

a particular airline. The 'Date' column is a date that an airline stock was publicly traded. The 'Adj Close' column is the adjusted closing valuation of a stock after considering things like stock splits, dividends, etc. This contrasts with the 'Close' column, which is just the price of the stock at the end of the day. We can use this information to calculate a daily return for the different airlines.

	Code	Description
366	BAS	Baranautica Air Service LLC
921	LDM	Laudamotion GmbH
106	4C	Aerovias de Intergracian Regional
1483	TXA	Texas National Airlines
15	0Q	Flying Service N.V.

Figure 3. 'Airlines'
This data is stored in a csv file hosted in a publicly accessible dropbox folder[11]. It stores abbreviation/full-name pairs for airlines. This table has 1712 entries and 2 columns. There is one null entry.

	Code	Description
2567	ITM	Osaka, Japan: Osaka International
2623	JBK	Oakland, CA: Berkeley Municipal Heliport
720	BIP	Bulimba, Australia: Bulimba Airport
4053	NY1	Kingston, NY: Kingston Ulster
4321	PAZ	Poza Rica, Mexico: Tajin

Figure 4. 'Airports'
This data is stored in a csv file hosted in a publicly accessible dropbox folder[12]. It stores abbreviation/full-name pairs for airlines. This table has 6654 entries and 2 columns. There are no null values.

```
{'year': 'Year of the data',
 'month': 'Month of the data',
 'carrier': 'Abbreviation of airline',
 'carrier_name': 'Name of airline',
 'airport': 'Airport code',
```

Figure 5. 'Metadata_Flight_Delay_2016_2021'
This data is stored in a csv file hosted in a publicly accessible dropbox folder[13]. It has been converted into a dictionary in our code, with the keys corresponding to columns in the delay table. The values are detailed descriptions of the corresponding delay table variable.

Data Manipulation Methods

Our initial EDA methodology was to familiarize ourselves with our datasets of our Airline stock data as well as our flight delay data. With our stock data, we looked to convert the stock tickers to create a new column called 'carrier_name' using the replace function and a dictionary of full airline names. This will enable us to join the stock data table and the airline delay table later on the carrier_name. We were able to identify that the delay data and the stock data only had a few airlines that overlapped, these were the tickers that we looked to convert. The overlapping airlines included American Airlines, Delta Airlines, United Airlines, Southwest Airlines, Spirit Airlines and JetBlue Airlines. We then got an understanding of the stock price metrics offered to us such as the 'Open', 'High', 'Low', 'Close' and 'Adj Close'. These will enable us to create a new column to calculate the returns of the stock price every day and aggregated to every month to then compare to our delay data. Since we have monthly delay data for every airline, we are planning on using a monthly stock return column. Using a pairplot to identify any possible correlations, we were able to pick out that 'Close' and 'Adj Close' were understandably very similar; the latter only changed marginally after post trading hours calculations of stock splits, dividends etc. were calculated. We then looked at a correlation matrix which led us to identify that we do not have to worry about multicollinearity in our data.

[11] "1b. Airlines.Csv," accessed March 18, 2023, https://www.dropbox.com/h?preview=1b.+Airlines.csv.
[12] "1a. Airports.Csv," accessed March 18, 2023, https://www.dropbox.com/h?preview=1a.+Airports.csv.
[13] "4a. Metadata_Flight_Delay_2016_2021.Csv," Dropbox, accessed March 18, 2023, https://www.dropbox.com/s/10sqlmn93lm79b3/4a.%20Metadata_Flight_Delay_2016_2021.csv?dl=0.

Next, we turned our attention to addressing null values in our datasets. While our stock data had no null values, our delay data had approximately 365 null values, with 193 rows containing one null value and 172 rows containing 15 nulls. With a total dataset of 101,629 rows, we felt that we had enough data to glean insights out of but we still wanted to be thorough to determine the validity of these null values. To determine if nulls were coming from a specific year, we split our delay data into 2017 and 2021 delay data. Upon further inspection, we found that the number of null values in our airlines of interest was not significant. We did a value count for null rows and discovered that the airlines we were focused on did not have a substantial amount of null values. As such, we will proceed by replacing missing values with 0.

Continuing with our EDA, we identified the most important columns in our delay data set to be ‘year’, ‘month’, ‘carrier’, ‘airport’, ‘arr_flight’, and ‘*_ct’. The first four are self-explanatory while the arr_flight refers to the number of flights from that specific airline that landed at that airport and the _ct column is a count column which indicates the particular delay and the overall count of said delay. We were initially confused about the partial values in the count column but upon closer inspection, we were able to identify that delays could be attributed to multiple causes. For example, a delay due to weather could also be precipitated by a delay from the carrier, such that this delay would be counted as .5 for each column. This makes sense intuitively as it eliminates double counting of a single delay of a flight.

After cleaning the delay data, we continued to visualize the data in a pairplot. Understanding that this would take quite a while to visualize over 100,000 rows we found that it was ultimately useful. Through this pairplot, we were able to identify that we had a tough task on our hands in that we had different units for certain delays. For instance, 'late_aircraft_ct' and 'late_aircraft_delay' appear to be highly correlated, but they measure the same event (one airplane delaying another airplane) in a slightly different way. 'late_aircraft_ct' is the count of aircraft delays due to another flight on the same aircraft being delayed, and 'late_aircraft_delay' is the number of minutes of delay due to another flight on the same aircraft being delayed. Also through this visualization, we determined that the distribution of our variables resembled a power-law distribution which indicates that we must log-normalize these variables to properly understand them. Similar to the stock data, we also used a correlation matrix which then revealed that there are quite a few correlated features.

We have a few other tables of data that also helped in our processing of our data and these were metadata which served as indicator tables for our columns. They included the full names of our aircraft delays and aircraft terminology which helped us situate ourselves with our delay tables. We set our index to these field names and then converted their descriptions to a dictionary which allowed us to view the field names as keys and their full names as the values. Similarly, we also had an airline and airports database which worked in a similar fashion to helping us understand the meaning of the abbreviations of each airline and each airport. We simply read these in and took samples of each database to get an understanding of their contents.

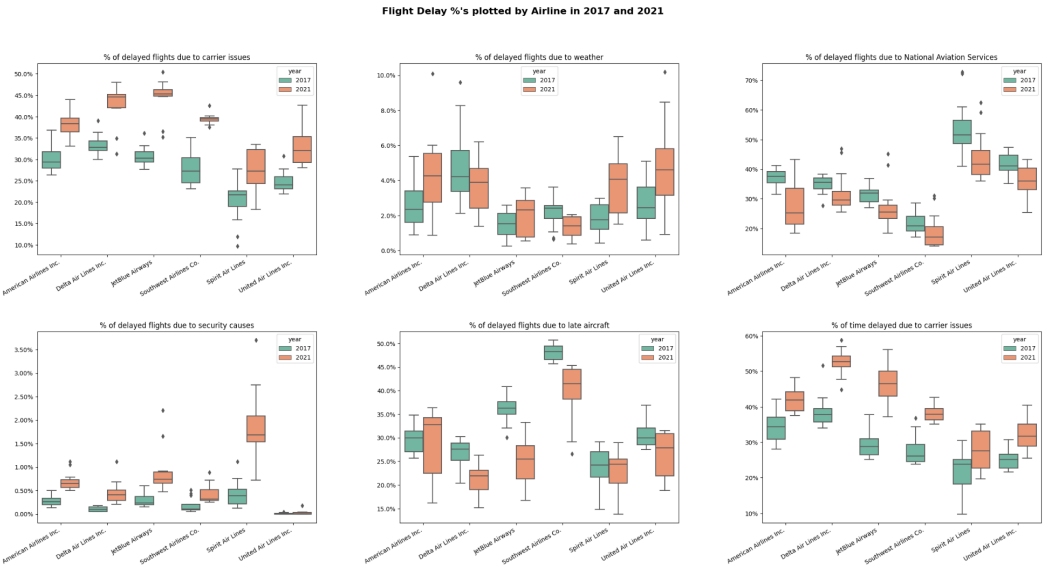
Moving forward to manipulating the data and understanding how these two databases interacted, we aggregated returns of the stock data from which we can take the log of the returns which helps us identify the movement of the stock price through each year. We then got the brunt of our manipulation when we created two dataframes of the cleaned data for 2017 and 2021 stock data. To prepare our delay database for merging, we dropped carrier, airport and airport name and grouped each delay by carrier_name, year and month. We then merged our stock and delay database on carrier_name, year and month with an inner merge to keep only the common rows of data between the two. We then split apart our final dataframes into data17 and data21 which contained airline delay data along with stock data for each respective year. The final part of our manipulation involved the normalization of

our delays data. We have 5 types of delays, carrier_ct, weather_ct, nas_ct, security_ct, late_aircraft_ct, these then get added to create a arr_del15. To normalize each of the delays, we divide each of the delays by the total arr_del15 to create a proportion of each delay. To confirm that our normalization went to plan, we added each of the delay ratios up to confirm that they all add to one. In a similar manner, we have the minutes delay that each of these delays created for each airline. We normalized each delay in minutes over the total number of minutes of delays which was displayed in the arr_delay column. With all this manipulation in place, we then moved to create visualizations and analyze their results.

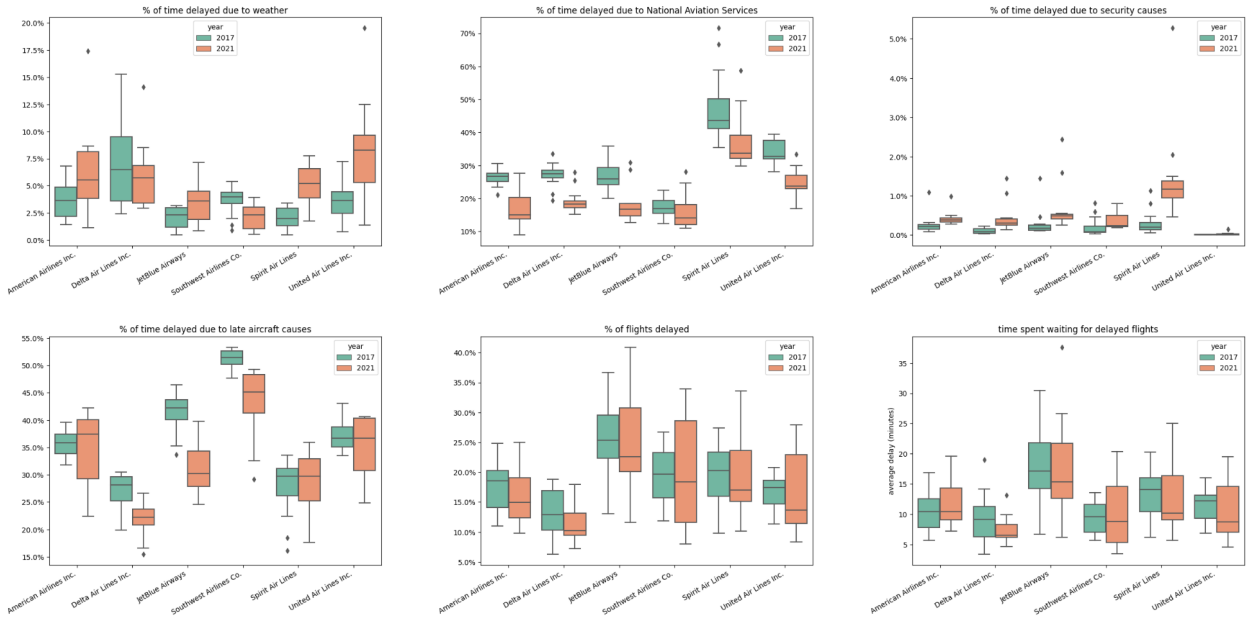
Visualizations and Analysis

Before we start our analysis we are working with the assumption that our 2017 data is representative of the airline industry pre pandemic and the 2021 data is more representative of the pandemic conditions. This is a limitation of our dataset without comprehensive airline delay and stock data through the pandemic. One of the main questions we wanted to answer was how did the pandemic affect airline delays and cancellations and was there a change in the type of delays as a result of the pandemic. A delay is an arrival that is more than 15 minutes past its expected arrival. From these delays, they are then categorized in a reason for each delay, these delays are a delay due to carrier, a weather delay, a security delay and a NAS delay (National Airspace System). Plotting these normalized ratios and grouping by carrier, we were able to compare how delay distribution changed from 2017 to 2021.

Comparing each type of delay between our airlines and plotting 2017 and 2021 values, we're able to find some trends in this data. Looking at the figure below,, we see that there is a potentially significant increase in the % of flights delayed due to carrier issues. This meshes up with the news reporting in 2021 when there were almost daily stories about carriers having issues due to staffing or lack of demand. Other trends that we were able to include Airline quality in the context of delays. For example, we see that Southwest Airlines had the highest percentage of flights delayed due to late aircraft in both 2017 and 2021 which indicates that it's a systemic airline issue for Southwest. Similarly, we can identify Spirit Airlines highest average flights delayed due to National Aviation Service issues both in 2017 and 2021. This does not bode well if Spirit Airlines flights are being held up by the NAS due to issues.



Identifying that carrier related delays account for the highest proportion of delays in 2021, we wanted to then determine if the increase from 2017 to 2021 in carrier related delays was statistically significant.



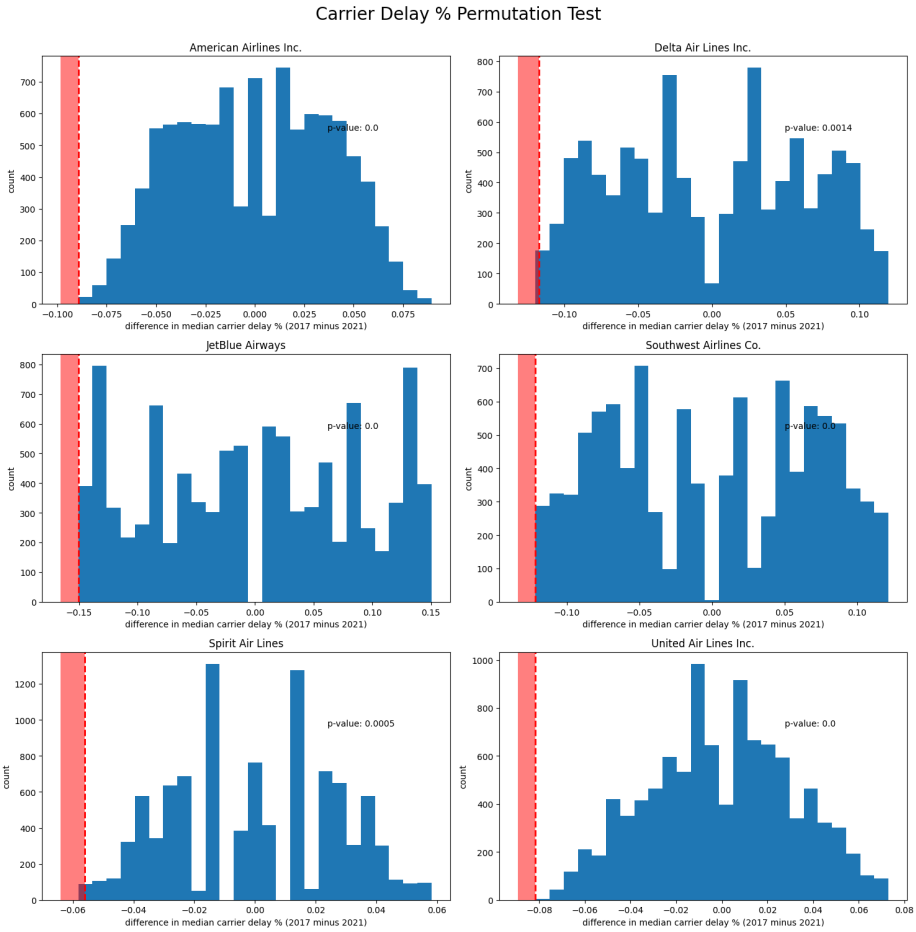
Before continuing to our significance testing, this boxplot above plots out the time delays between each of the airlines comparing 2017 and 2021. Comparing the amount of time that each delay incurs, we find that there are not many differences between delay times for 2017 and 2021. It looks as if there is a larger range of time delays since the height of our 2021 boxplots are much larger than the 2017 boxplots. Picking out two particular observations, we see that there is a potentially significant increase in delay time for Spirit airlines due to security causes; the second potential item of note is an increase in United AIrlines increase in time delay due to weather related delays. As these two items only account for 10% of delays in total, they are of note but potentially not important for the overall dataset.

To determine if the increase in carrier delays between 2017 and 2021 was statistically significant, we ran a permutation test on the carrier delay % numbers between 2017 and 2021 to get the p-value of each airline between these variables. We did this through a user defined function called “get_permutation” which takes in a two column dataframe and returns the difference between the medians between the two columns. We then use this function in another user defined function of “get_permutation_list” which takes in a dataframe and an n number of iterations to generate. The latter function uses list comprehension and get_permutation to loop through n number of times and returns a dataframe of the results with columns ‘diff’, ‘median_1’, and ‘median_2’.

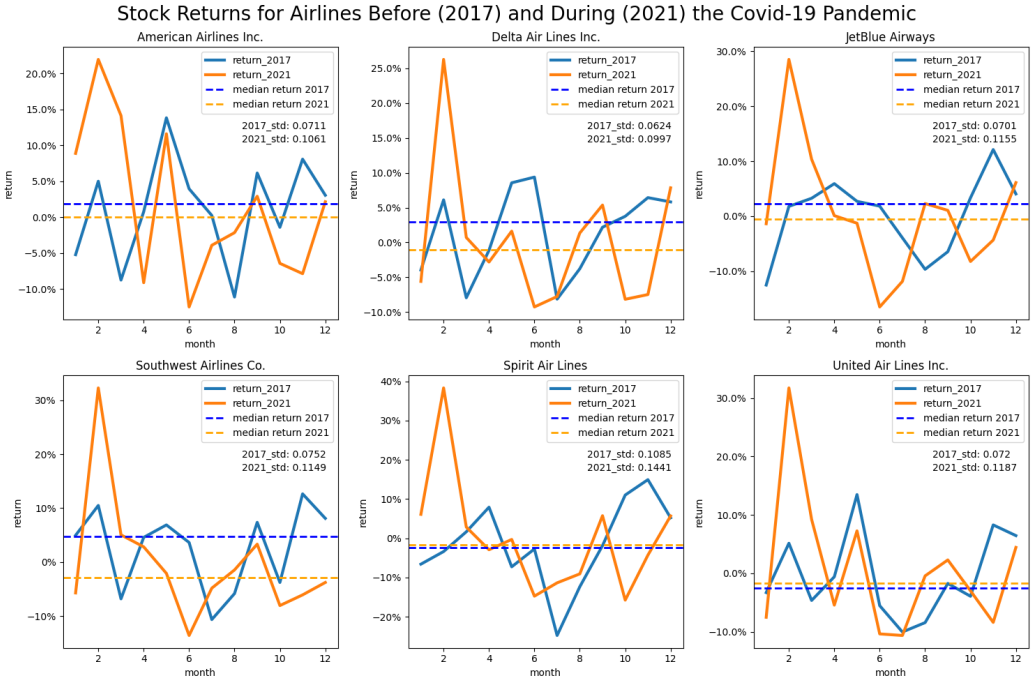
We used these two functions with a dataframe of combined, normalized values of 2017 and 2021 delay data. We dumped these results to pickle for easy retrieval, as it takes a long time to generate 10,000 permutations for different 6 airlines. We finally then called our final user defined function plot_permutation_test which creates a 3x2 boxplot and displays the p-value as a cutoff of the plotted data. Comparing the p-values of between the 2017 and 2021 delay data, we find significant differences between every single airline. We can conclude that the pandemic did indeed cause a significant impact of delays due to carrier related issues. This analysis is plotted on page 7 with a histogram with a p-value designation showing the significance of each carrier delay comparison between 2017 and 2021 and the reflection of all p-values meeting the threshold of .05 for significance.

After determining this significant impact of delays, we then wanted to compare stock return data to see if there were similarly any significant losses in return due to the pandemic. We plotted the percentage returns of each

airline in 2017 and 2021 in one graph to view any graphical differences between the two. We were able to identify that the median monthly stock return for each airline was lower in 2021 versus 2017 for 4 of the 6 airlines. Additionally, all airlines in 2021 had a negative median monthly stock return, whereas in 2017 only 2 airlines had negative median monthly stock returns. For every airline in 2021, the standard deviation of the monthly stock return was higher than in 2017, which means that the airline stocks were more volatile overall in pandemic times. The figure at the bottom of this page shows this analysis. One other point of interest is the spike in return in the airline industry in February. These returns seem industry related as this peak is reflected in all of the airlines. Even with this massive peak, we see that the returns are still lagging behind 2017 median returns. To see if these results were statistically significant, we will perform a permutations test similar to the one that we performed above with the 2017 and 2021 carrier delay data. In this test, we will be comparing the monthly returns of each of our six airlines in 2017 and 2021.

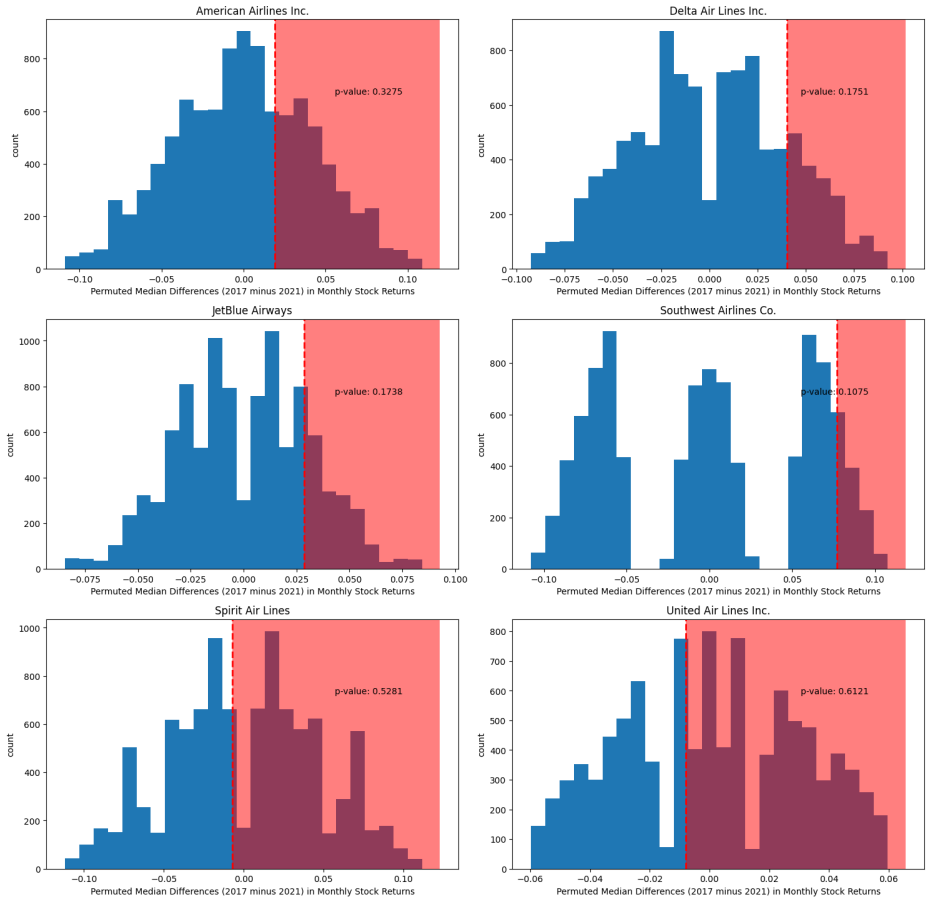


We first generated 10000 permutations of the median monthly stock return for each airline's monthly return and plotted the distribution. We then evaluated the statistical significance of the generated median stock returns. The plots on the next page plot out the permutations and the red line marks the p-value. Looking at



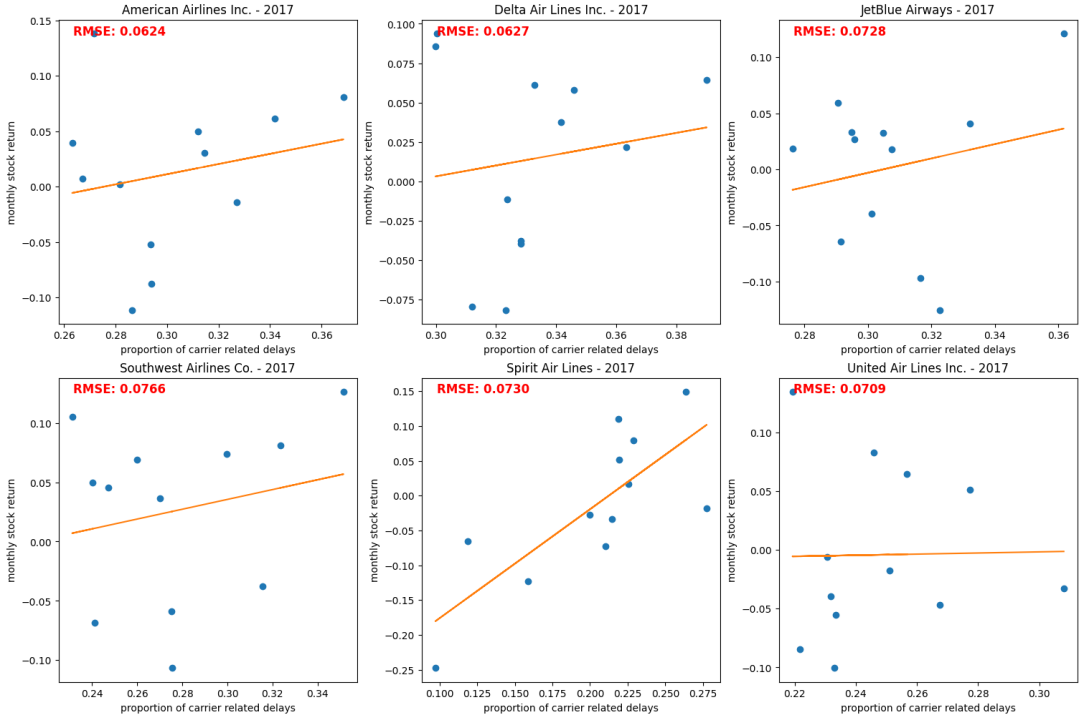
the plots, we find that our p-values are large and that the differences between the stock returns in 2017 and 2021 could just be due to chance as well. Even though we identified that the medians for the returns were lower in multiple airlines, we find that the differences were not statistically significant to draw any conclusions from.

Permutation Test for Stock Returns for Airlines Before (2017) and During (2021) the Covid-19 Pandemic



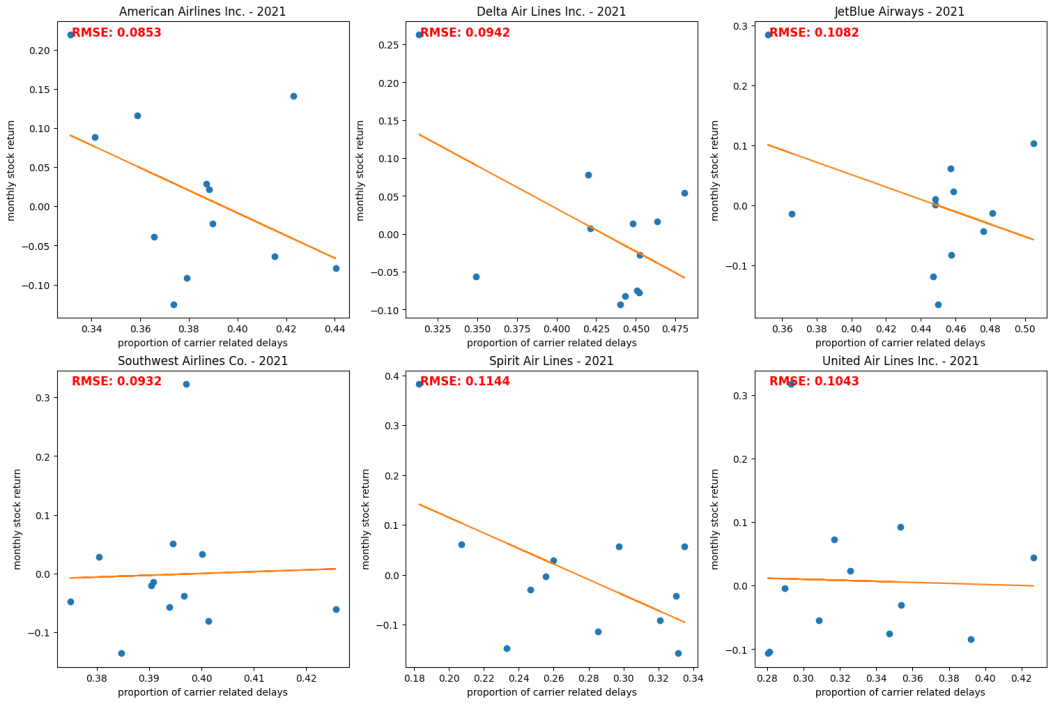
In particular, the lowest p-value was .1075 for Southwest Airlines; which means that given random permutations of the data between the years 2017 and 2021 for Southwest Airlines, there is a 10.75% chance that the difference in medians between the years 2017 and 2021 is 0.077091 or more greater. We do not have sufficient evidence to say that the median monthly stock returns for each airline are different in 2017 and 2021. Although there was significant change between the returns for 2017 and 2021, the financial markets are complex and may not be strictly correlated with raw flight delay data. Presumably, delay data plays a part in the overall picture of an airline's stock price and return but perhaps not enough of a part to correlate with strictly delay data.

Fully embracing the data scientist part of our identity, we wanted to take our analysis a step further to analyze if our one significant feature, delays due to carrier issues, can predict a stock's return. We first started by defining two functions, the first of which performed linear regression with 5 fold cross validation and returns the average root mean squared error. The second function plots the regression with a fitted line and displays the RMSE for each airline.



We knowingly undertook this exercise with the understanding that we distilled our data down to 12

points (monthly return of a year) which would significantly decrease the weight of our experiment and decrease its predictive power. Looking at the results, we find “positive” correlations in monthly stock returns with proportion of carrier delays. This once again is counterintuitive to common sense in that we expect that with a great proportion of carrier delays, we expect returns to decrease. We think this because with increased carrier delays, you expect increased media coverage and social media negativity with individuals being affected by delays that can be pointedly attributed to each airline. The best performing model was Delta Air Lines Inc, with an RMSE of 0.0627. This means that the model's predicted values were on average .0627 away from the actual monthly stock return for the associated carrier-related delay value. Given that Delta Airlines monthly stock return ranged from -.075 to .1, this is not a good model.



Looking at 2021 data, we find that our initial assumption is reflected in that a correlation of increased carrier delays results in negative correlation with stock returns. Although the graph may have reflected our assumption, looking at the RMSE, we find that they look even worse for this set of data. The best performing model was American Airlines Inc for 2021, with a RMSE of 0.0853. The stock return values ranged from about -.1 to .2, and on average the model's predicted values were .0853 away from the actual monthly stock return for the associated carrier-related delay value. This is also not a good model.

Conclusions

In conclusion, the COVID-19 pandemic has undoubtedly transformed the landscape of air travel and the airline industry. Our study utilized two datasets of US flight delays and US airline stock prices to reveal if there were any major differences in the airline industry before and after the pandemic. We conducted extensive EDA to familiarize ourselves with the datasets, address null values, calculate daily stock returns, and identify potential correlations between variables. We normalized the delay data to properly understand the context between different types of delays.

Our analysis of the combined data revealed a statistically significant increase in delays due to carrier-related issues in 2021, likely driven by staffing challenges and reduced demand because of the pandemic. We also looked at stock returns for the airlines in these years and found that there was no significant difference in the monthly stock returns between 2017 and 2021. With a lack of statistical significance, this suggests that there are other factors that might be influencing stock prices in the complex financial market. Additionally, our attempt to predict stock returns based on the proportion of carrier delays was not successful, as the models generated had high root mean squared errors, indicating poor predictive power. This further emphasizes the complex nature of financial markets and the multitude of factors that can influence stock prices. Overall, our work was based on comprehensive datasets, thorough data processing, and tried to offer insights into the impact of the pandemic on airline delays. Further research is needed to better understand how airline delay data along with other myriad factors influence the financial resilience and recovery of the industry.

Statement of Work

Coulton Theuer was responsible for most of the coding portion of this project. Rohit Maramraju was responsible for most of the written portion of this project with final editing being split between the two individuals. To improve collaboration in the future, we need to identify a platform to enable both individuals to code alongside each other efficiently. We tried different methods such as GitHub Codespaces, VS Code collaboration but neither of these platforms could handle our complex dataset and setup.