Final Report for:

Investigation of Correlations of Three Airline Companies with Quantities and Types of Delays at Major US Airports Including Airline Financial Data and Consideration of COVID

Industrial School of Systems and Engineering, Georgia Tech OMSA

Management 6203: Data Analytics in Business

Thuy Thi Dinh, Eleibny Feliz Santana, Spencer Allgaier, Leland Bolleter

Team: 13 aka Team Definitely Vibrant (TDV)

November 20th, 2022

*** Note ***:* Because of the large number of acronyms, that are fully defined in the paper, bibliographic citations are *italicized*. A full bibliography and literature review is at the end of this report.

I. Choice of Topic

Team 13, aka Team Definitely Vibrant (TDV), developed and discussed ten initial project proposals. Those ideas were narrowed to four selections with each team member researching a candidate topic. We debated, and democratically chose, studying airline flight delays and correlations to weather variables.

TDV expanded the initial proposal topic from:

- Optimization Input for Flight Schedules to Minimize Departure and Arrival Delays in Coordination with Risk Ranking of a Major US Airport Based on Past Flight Delays

to:

- Investigation of Correlations of Three Airline Companies with Quantities and Types of Delays at Major US Airports Including Airline Financial Data and Consideration of COVID

The TDV project hypothesis: Increased delays have a negative relationship to Airline financials.

We chose the topic based on value of business applicability, data availability, current importance, interest, and depth of modeling available. The wealth of The Bureau of Transportation Statistics (BTS) data was a compelling factor in the choice of TDV's topic. Detail on BTS and financial data are covered in section IV. Understanding the Data and Data Wrangling.

II. Business Justification / Objective

According to the industry group Airlines for America, air travel delays cost U.S. airlines and passengers an estimated $28 billion in 2018 *(A4A, 2022).* TDV and our customer determined that an investigation of the existence of correlations between an airline corporation(s), delay types, airports, and airline-financial data would provide beneficial-operational insight. Additional beneficiaries of the project include industry organizations along with the entire air-travel-business ecosystem.

TDV's report, models, and visualizations were developed on a subset of US air travel data. Although intended to be a stand-alone project, TDV's work would also be considered a proof of concept of a broader operational-intelligence system with more comprehensive US (and possibly global) air travel and financial information. A subsequent US-full-scale project would include all airlines, all public airports, delays, and select financial data for airlines.

III. Problem Statement

Our project was developed as a review of key airline-delay challenges along with possible correlations to the airline corporation's financial health. The effect of the COVID pandemic severely impacted airline passenger travel after March 2020. The Federal Aviation Administration (FAA) reported that in 2019 air carriers served 1,057.6 million passengers compared to a 2021 figure of 597.9 million passengers *(FAA, 2022).* COVID has been considered by TDV as an additional compelling topic to model. Data implications of COVID were discussed by the group and accommodated. Details of data accommodations are covered in this progress report in section IV. Understanding the Data and Data Wrangling.

Shifts in a business landscape have been more easily navigated by businesses with data and analytics prowess *(Provost et al., 2013).* In conjunction with the evolution of the airline-business landscape and COVID onset, there have been inherent-historical challenges in the airline industry. Inherent challenges include a capital-intensive industry, an environment typically expectant of a high level of customer service, a highly regulated industry, the existence of nation-owned airline competitors, and a requirement of having highly skilled pilots and mechanics. A pilot error or maintenance error has the potential consequence of a large sensational loss of life that would be detrimental to the airline corporation's reputation and financials.

IV. Understanding the Data and Data Wrangling

At the beginning of the project TDV realized that the scope of data required considerable attention. TDV chose a subset of three airlines with large market capitalizations: Delta, United, and Southwest. Delta and United were large carriers focused heavily on acquiring business customers while Southwest was known to be more of an economical airline focused the consumer market.

TDV acquired the arrival flight data for 40 major United States airport from BTS. BTS is an agency within The Department of Transportation. The data accessed included arrival flight counts, delay counts and cancellation counts. It also broke down delays into the same categories by total minutes attributed to each delay type.

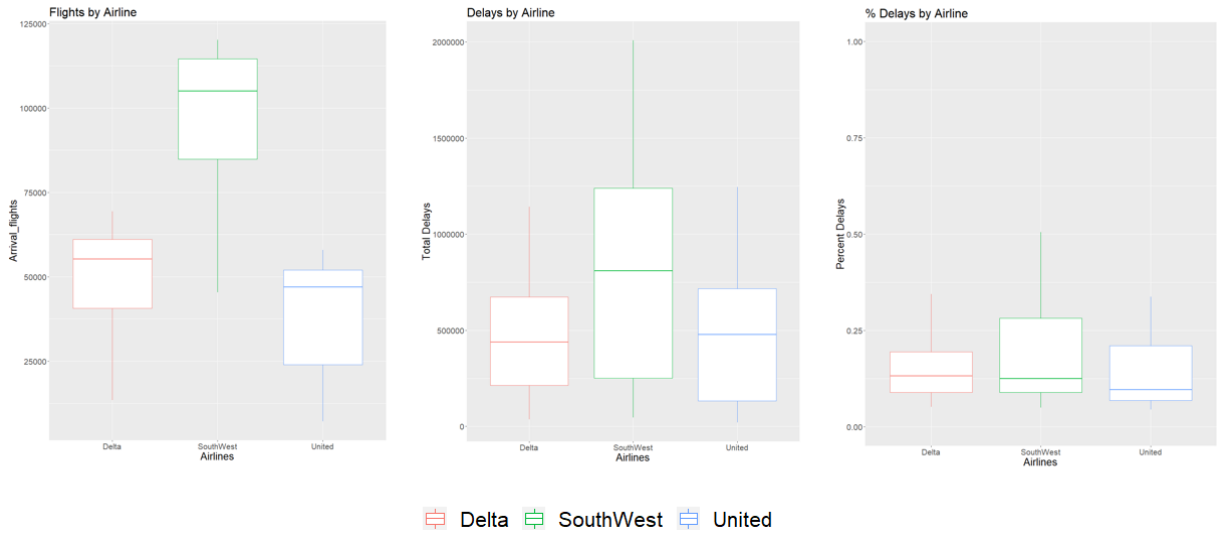
A period of 9/1/2017 to 9/30/2022 was chosen to help facilitate COVID modeling – 61 months. The US COVID-19 pandemic started in 3/2022 and is considered by TDV as a transition month. There are 30 months in each period from 9/1/2017 to 2/29/2020 and from 4/1/2020 to 9/30/2022.

Financial data was collected from Yahoo Finance during the period of 9/1/2017 to 9/30/2022. Availability of granular financial data was limited; therefore, the smallest available time span was opted for each financial variable. The finance dataset included monthly data for Market Cap and stock value. Quarterly data was used for gross income, net profit, and revenue.

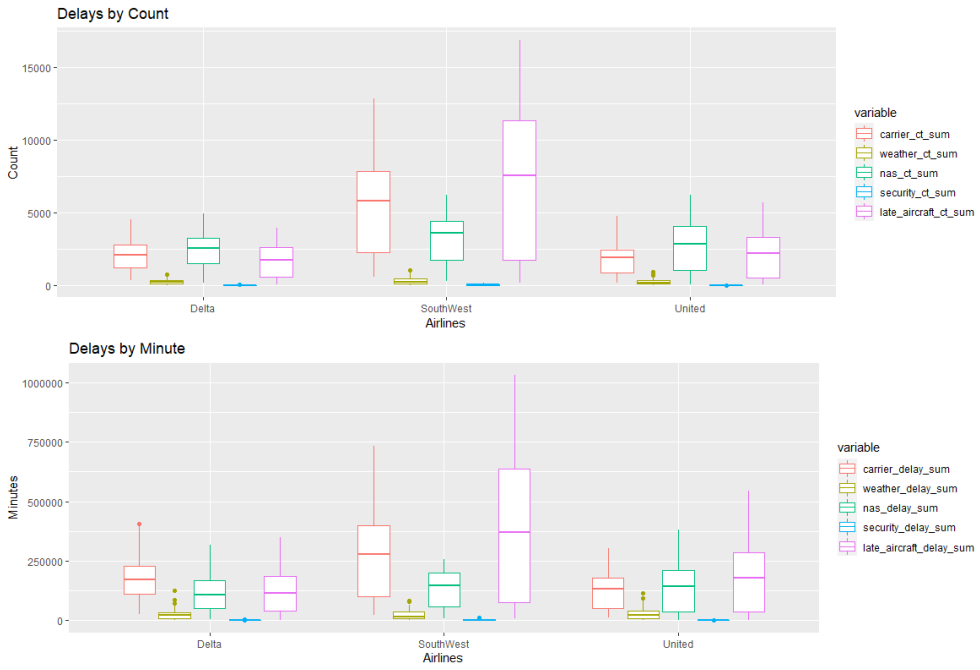
During data cleaning, the daily flight data was aggregated to monthly and quarterly sums based on the adjoining financial variables. Afterwards, non-critical variables and alias columns were identified and removed. To allow for the use of a log() scale in graphs and visuals, zero values were replaced with 0.0001 to eliminate R related handling errors. After all aggregation was completed, a binary Covid value was added to indicate if the data point was considered pre-Covid onset or post-Covid onset.

V. Exploratory Data Analysis:

During exploratory data analysis, TDV compared flight data between all three carries to observe factors including trends and differences. First, analysis was performed to determine the difference in number of flights by airlines and number of delays by Airline. That analysis showed that Southwest had significantly higher traffic during the explored time frame. This was not a full picture however, as it only captures the view of the top 40 busiest airports in the United States, while both Delta and United have a much larger presence internationally. Understandably, Southwest also had a higher count of delays when compared to Delta and United. However, when broken down into percentage, percent delays were relatively standard across the board at approximately 20%. This showed that none of the observed carries had more history of significantly more delays when compared to their competitors.



To identify critical variables, TDV analyzed the breakdown of types of delays between carriers. Box plots were created with each delay type, both in count and in minutes, to determine which types of delays had the biggest impact on total delays. Carrier delays, National Airspace System ((NAS), where the FAA agency resides) delays, and late aircraft were the largest contributors to delay issues while security delays and weather delays had minimal effect in comparison. It was also noted that count of delays did not always correlate to impact of delays. For instance, looking at the United information, even though NAS had a greater delay count compared to late aircraft delays, late aircraft contributed to more grounded flight time. Since the difference is not significant, for the purpose of this project, we focused mainly on number of flight delays over total minutes of flight delays.



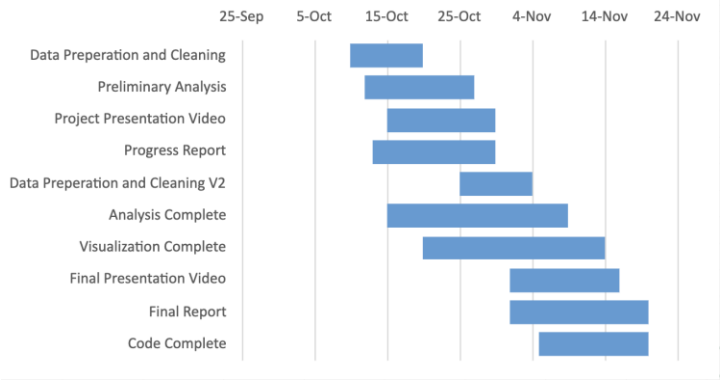
VI. Approach/Methodology, Level of R code, and Programming

TDV employed a standard data-science methodology. The team did research on the airline industry related to delays. We shared those findings on the TDV Slack collaboration site. We also researched plausible sites for data. The BTS data became a core resource for our work. Initially TDV collected BTS and NOAA weather. After work with the NOAA data, we expanded the project topic to include financial data. This is an example of one of the iterations TDV did for an enhanced analytic effort. TDV collected financial data from Yahoo Finance and Microtrends.

Our current core data includes sets aggregated by day, month, and quarter. TDV also uses airline carrier and airport as top-level values. With airline carrier and airport, we have monthly and quarterly data for delay types attributed to the airline carrier, weather, NAS, security, and late aircraft. The project financial data includes stock price, revenue, gross profit, and net income.

TDV completed R code for data collation, investigation, and modeling. R code for data and modeling are hosted on the Georgia Tech GitHub TDV folder (Team-13). TDV has comprehensive and cleaned data sets of the data fields above. TDV used the standard iterative process of modeling <-> evaluation <-> feedback (from team), and back to modeling.

VII. Project Timeline



VI. Model 1 – Airline Revenue, Gross Profit, and Net Income Correlation to Delays

<u>Data Set Description Overview</u>

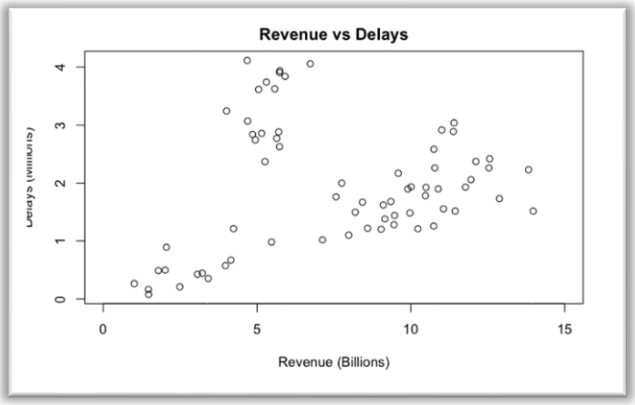
Financial data for this model was obtained through publicly available quarterly financial reports at macrotrends.net from January 2017 to September 2022. Delay counts were obtained from the BTS to represent the corresponding count of unique delays across all airlines in the top 40 busiest domestic airports for each quarter.

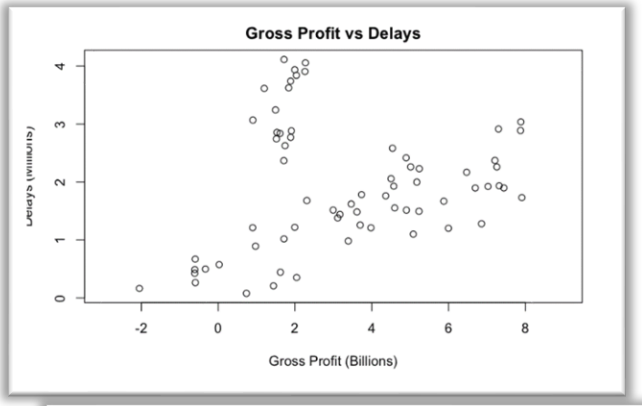
<u>EDA Work Overview</u>

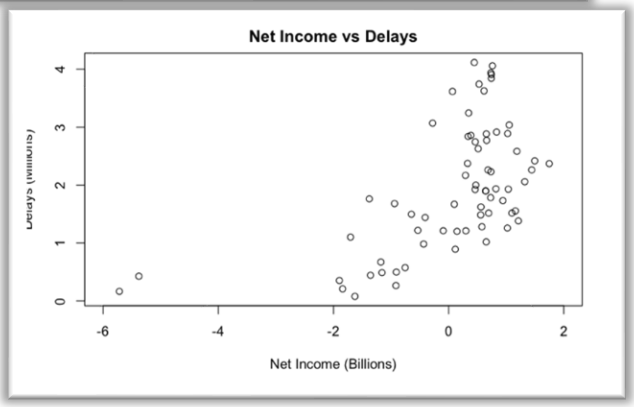
Our initial hypothesis was that there would be a significant negative relationship between delays and any of the three indicators, but most significantly in gross profit. The reason being delays should incur costs on a business' operations and gross profit is the indicator that most purely represents that number without accounting for taxes/insurance/other expenses like net income.

Using scatterplots to visualize potential linear relationships and outliers, two observations were immediately recognized. First, there was an identifiable *positive* relationship between delays and financial success as represented in the three indicators. The trend showed clear signs of heteroscedasticity, but the relationship was contrary to our initial hypothesis. Second, two outliers from 2020 Q2 and Q3 stood out from the rest of the data,

particularly from Delta's Net Income values. As part of our model prep, the two outliers were removed after determining that those two quarters were not representative of the rest of the time-period. Even though they are included in the post-COVID era of our research, there was a clear indicator that the first two quarters of the pandemic had too much change to be representative of the time-period.







Model Overview

After the relationships were visible and the outliers were removed, it was still important to recognize the difference in the industry that the COVID pandemic had brought on. As a result, the data was blocked into two different time dependent sections consistent with the rest of the research, pre-COVID (Q1 2020 and earlier) and post-COVID (Q2 2020 and later). A simple linear regression using the formula (DelayCount ~ Revenue + GrossProfit + NetIncome) was then applied to answer how delays affect financial performance.

```
Call:
lm(formula = arr_delay_sum ~ Revenue + GrossProfit + NetIncome,
    data = precovid_combined)

Residuals:
     Min       1Q   Median       3Q      Max
-1698748  -296052   -29950   387070   908950

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.877e+06  3.756e+05  10.321 1.04e-11 ***
Revenue     -3.044e-04  7.110e-05  -4.281 0.000158 ***
GrossProfit  1.155e-04  7.763e-05   1.487 0.146696
NetIncome    9.122e-04  3.241e-04   2.814 0.008296 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 607400 on 32 degrees of freedom
Multiple R-squared:  0.4752,    Adjusted R-squared:  0.426
F-statistic: 9.659 on 3 and 32 DF,  p-value: 0.0001087
```

```
Call:
lm(formula = arr_delay_sum ~ Revenue + GrossProfit + NetIncome,
    data = postcovid_combined)

Residuals:
     Min       1Q   Median       3Q      Max
-1273500  -697820  -214841   585328  1835863

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.156e+06  5.335e+05   4.041 0.000545 ***
Revenue     -6.238e-05  1.051e-04  -0.594 0.558839
GrossProfit  5.554e-05  1.542e-04   0.360 0.722143
NetIncome    7.170e-04  2.571e-04   2.788 0.010714 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 993500 on 22 degrees of freedom
Multiple R-squared:  0.2764,    Adjusted R-squared:  0.1778
F-statistic: 2.802 on 3 and 22 DF,  p-value: 0.0637
```

The result was two model outputs showing the statistical relationships during both pre-COVID and post-COVID eras. There was a positive relationship, but revenue and gross profit failed to show statistical significance across both eras. Net income was the only indicator that showed a significant relationship at a 95% confidence level both pre-COVID and post-COVID. The results were surprising at first and we thought of ways to adapt the model

to account for any non-representative data or factors not already mentioned. It was decided that shifting the financials into the future by a quarter or two would not be appropriate since it would not accurately reflect the cost incurred by a delay. Fixing delays to a rate and switching to a logistic regression model would also be incorrect since it did not align as accurately with our research question. There did not appear to be any corrections that needed to be made with this model.

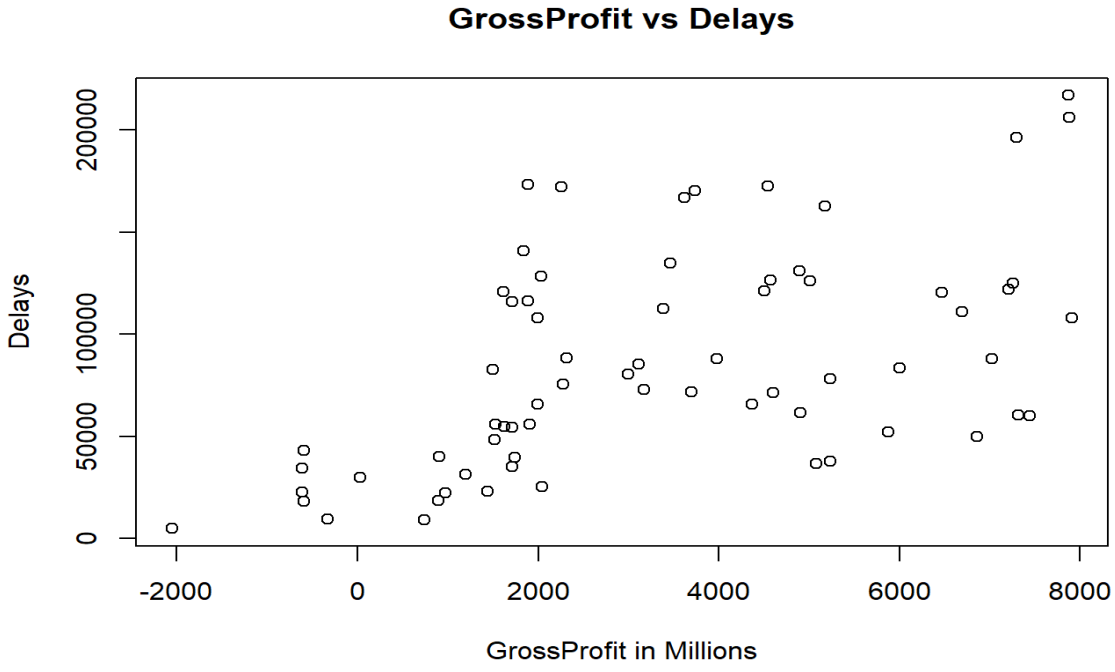
After the results were produced and no other modifications were necessary, it made more sense when we remembered that delay rate across airlines is roughly consistent across airlines at about 20% more delays. That in mind, more delays often relate to more flights sold, higher revenue, higher gross profit, and higher net income. Conclusively, net income on a company's quarterly financial statement is most closely aligned with delay count, but with a positive relationship and it is still difficult to make predictions on delays using these factors alone.

VII. Model 2 – Gross Profit Correlated to Delays

Dataset description Overview

For models 2, 2a and 2b, we used the financial data of year 2017 to date (2022) for Delta, United and Southwest airline. This data was combined with weather delays data for the same period.

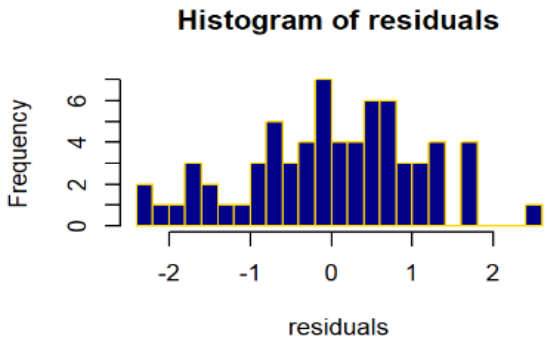
The weather data was aggregated to quarters to make and then merged with the financial data before starting EDA. We use scatterplot and boxplots to understand the data and identify relationships between predictor and response. We also identified and eliminated alias columns to minimize the chances of multicollinearity.

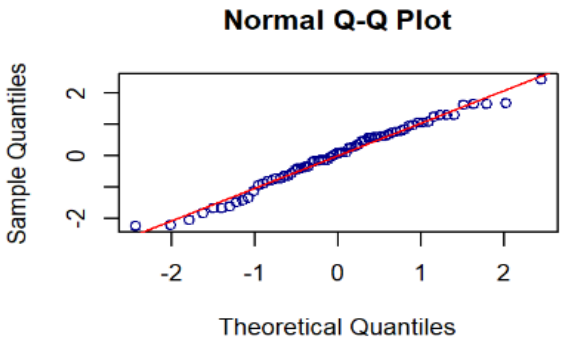


Overview

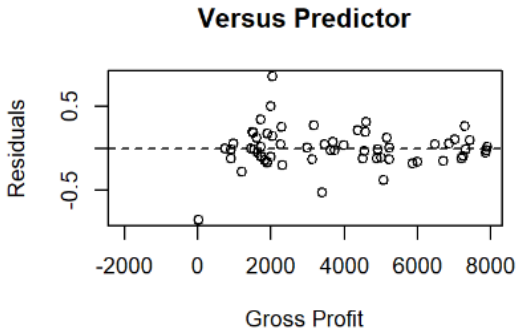
Three techniques were used to model the data: linear regression, log-linear transformation, and SQRT, determined after using Box Cox. We will explain each.

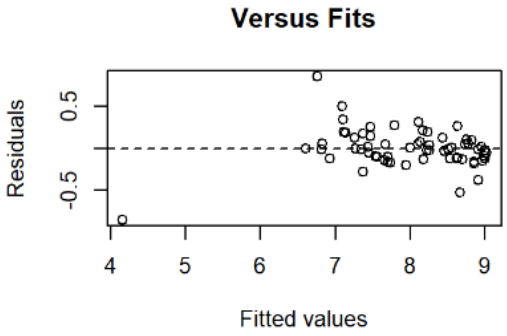
In model 2, use Linear regression to predict the gross profit of the airlines based on delays data.
After performing residual analysis on the regression results, we observed that the normality assumption is violated, indicating that the model is not a good fit for the data.

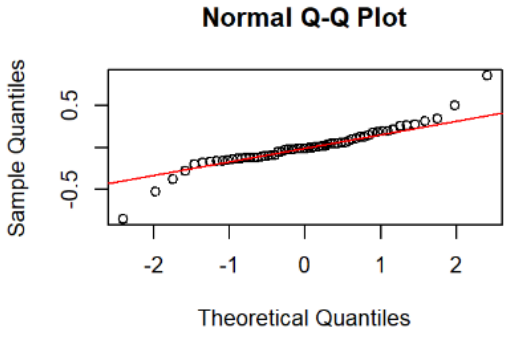


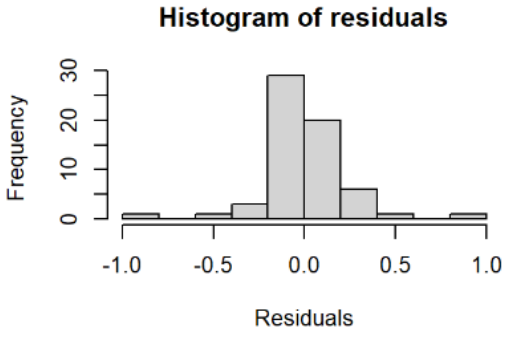


On model 2a, we used the log linear transformation method. We see some improvements in terms of normality. We now see a single peak on the histogram. However, the distribution is not symmetric. In terms of constant variance and linearity, reflected on the top plots, we observe clusters of data. The residuals are not equally or randomly distributed around the zero line, indicating that this model is not a good fit for the data.







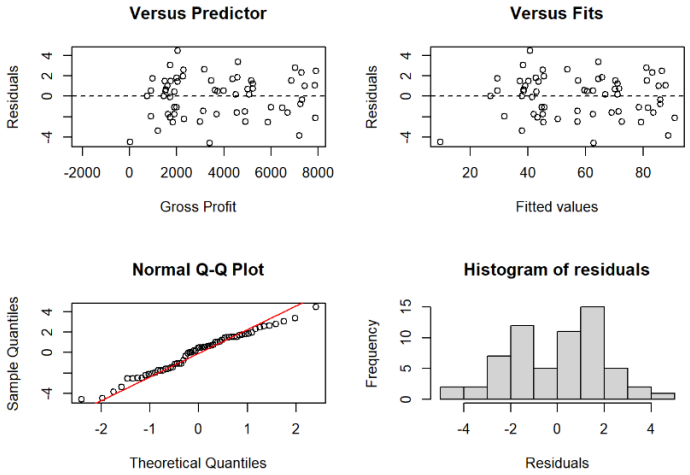


2b)When the normality or constant variance assumptions do not hold, we often we use a transformation that normalizes or stabilizes the response variable. We used the Box Cox technique to determine the ideal Lambda. In this case, we are using the SQRT function to transform the response and fit a new model and proceed with the residual analysis to evaluate the assumptions.

We observe that the transformation has improved the model significantly. However, we can see that the normality assumption is still violated.

The implication of the normality assumption violation is that the uncertainty in predicting the gross profit would be higher than estimated using this model. The model might not contain as many significant factors to accurately predict gross profit.

Thus, we conclude that weather delays alone do not statistically affect airlines gross profit.



VIII. Model 3 - Market Cap and Delays:

The model 3 dataset comprises of all the flight data from Delta, United and Southwest aggregated on a monthly basis with Market Cap values. In each of the Market Cap modes, number of delays are insignificant when modeling Market Cap. Opposite to our hypothesis we did not see an impact of total delays on Market Cap. In the total model, we saw that delays were only significant at the 90% confidence interval. In individual Airline models, we saw no significant relationship at all. It was seen that number of flight has a positive relationship to market cap and cancelations has a negative impact on Market Cap.

Total Market:

```
Call:
lm(formula = MarketCap ~ Airline + arr_flights_sum + arr_del15_sum +
    arr_cancelled_sum, data = comb)

Residuals:
       Min         1Q     Median         3Q        Max
-1.159e+10 -3.349e+09  3.957e+08  3.523e+09  1.696e+10

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.720e+10  1.277e+09  13.469  < 2e-16 ***
AirlineSouthWest  -1.112e+10  1.194e+09  -9.309  < 2e-16 ***
AirlineUnited     -9.717e+09  8.488e+08 -11.447  < 2e-16 ***
arr_flights_sum    2.675e+05  3.052e+04   8.767 4.06e-16 ***
arr_del15_sum     -1.511e+05  8.140e+04  -1.856   0.0647 .
arr_cancelled_sum -4.923e+05  9.130e+04  -5.393 1.71e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.871e+09 on 231 degrees of freedom
Multiple R-squared:  0.6841,    Adjusted R-squared:  0.6772
F-statistic:   100 on 5 and 231 DF,  p-value: < 2.2e-16
```

United:

```
Call:
lm(formula = MarketCap ~ arr_flights_sum + arr_del15_sum + arr_cancelled_sum,
    data = UA)

Residuals:
       Min         1Q     Median         3Q        Max
-8.696e+09 -2.130e+09 -7.425e+07  2.327e+09  6.344e+09

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       7712989621 1205970206   6.396 1.13e-08 ***
arr_flights_sum       225947      58422   3.868 0.000228 ***
arr_del15_sum         100616     209893   0.479 0.633035
arr_cancelled_sum    -898225     231992  -3.872 0.000225 ***
```

Delta:

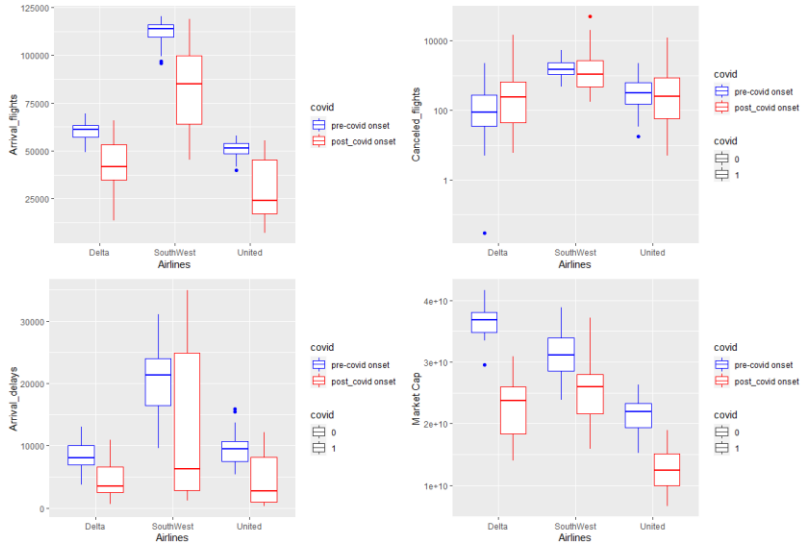
```
Call:
lm(formula = MarketCap ~ arr_flights_sum + arr_del15_sum + arr_cancelled_sum,
    data = DL)

Residuals:
       Min         1Q     Median         3Q        Max
-1.331e+10 -3.556e+09  8.524e+08  2.809e+09  1.005e+10

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       7014750761 2305497605   3.043 0.003245 **
arr_flights_sum       504608      75688   6.667 4.08e-09 ***
arr_del15_sum        -351686     326430  -1.077 0.284815
arr_cancelled_sum    -1292062     317184  -4.074 0.000115 ***
```

SouthWest:

```
Call:
lm(formula = MarketCap ~ arr_flights_sum + arr_del15_sum + arr_cancelled_sum,
    data = SW)

Residuals:
       Min         1Q     Median         3Q        Max
-9.353e+09 -3.001e+09 -2.249e+08  2.377e+09  1.112e+10

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        1.908e+10  2.964e+09   6.437 1.08e-08 ***
arr_flights_sum    9.421e+04  4.119e+04   2.287  0.02506 *
arr_del15_sum      5.131e+04  8.900e+04   0.577  0.56603
arr_cancelled_sum -2.979e+05  9.568e+04  -3.114  0.00263 **
```

Covid Impact: In the Covid models, there was approximately a 4.6 Billion dollar, 9.6 Billion, 6.1 Billion Dollar impact on Market Cap for Southwest, Delta and United, respectively. There was a significant decrease in arrival flights as consumers reduce unnecessary travel which also lead to a decrease in delays. Cancelations also increased during Covid for some carriers. Both these variables of decreased

flight and increase cancelations, had a significant impact on the Market cap which was expected from our original models. Deltas increased cancellations may be why we see more of a differential between the Market Cap before/ after Covid.



Results: Delays do not have a strong correlation to Market Cap. The greatest impact of market cap from our model was Arrival flights and Cancelations. What was noted was the correlations between Arrival flights and delay counts. As seen in our early evaluations there is relationship between number of flight and number of total delays. Both the graph and the linear model show approximately a 21% relationship. This meant that as flights increased, delays also increased. So there is multiple collinearity between delays and arrival flights. A VIF was conducted to verify this assumption. To improve our model, it would be necessary to remove delays as a dependent variable.

```
##                      GVIF Df GVIF^(1/(2*Df))
## Airline          3.781668  2        1.394507
## arr_flights_sum  8.527725  1        2.920227
## arr_del15_sum    4.214406  1        2.052902
## arr_cancelled_sum 1.154047  1        1.074266
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## Airline          1.617644  2        1.127771
## arr_del15_sum    1.524305  1        1.234628
## arr_cancelled_sum 1.078044  1        1.038289
```

```
Call:
lm(formula = arr_del15_sum ~ arr_flights_sum, data = comb)

Residuals:
     Min       1Q   Median       3Q      Max
-15631.1  -2208.4    277.2   2042.8  15555.5

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.708e+03  6.628e+02  -5.595 6.12e-08 ***
arr_flights_sum  2.188e-01  9.583e-03  22.831  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4467 on 235 degrees of freedom
Multiple R-squared:  0.6893,    Adjusted R-squared:  0.6879
F-statistic: 521.2 on 1 and 235 DF,  p-value: < 2.2e-16
```

IX. Model 4 – United Airlines Stock Price Correlated to Delays

Data Set Description Overview

For this model data sets for airline delays and stock price came from BTS and Yahoo Finance websites. BTS data is aggregated monthly. Although daily BTS airline delay data was desired only summarized monthly data was available from BTS. Data files were kept separate per airline since another EDA effort captured inter-airline relationships. Yahoo Finance monthly stock price data was available for the market opening price for the 1st day of the month. To rationalize the combination of the 2 data sets, the BTS delay data was assigned to the monthly opening day stock price from the first trading day of the next month. The BTS data summary was all major United States (US) airports used by Delta, Southwest, and United Airlines.

EDA Work Overview and Relation to Domain Knowledge

In our EDA and domain knowledge work COVID provided interesting challenges. COVID onset in the US caused a decrease in flights and an initial large increase in cancelled flights. COVID again caused problems with arguably overwhelming demand in the runup to the US air travel rebound in June 2021 and on into 2022.

The BTS data required considerable review of their data dictionary (Appendix 1), individual data points, and overall data sets. EDA work was interspersed with work on collecting and combining the data sets. Because of careful work described above for the data sets much of the "mid-cycle" EDA work was combined. EDA work continued iteratively, initially, with modeling work described below.

A major distinction in the data sets related to the BTS fields. BTS provided two groups of data. One group of data contains flight count information by delay type with a summary by all delay types field. The second group of data contains the minutes of delay attributed to delay type along with a summary by all delay types field. Linear regression models used either flight delay counts or flight delay minutes separately. No models used combined flight delay counts and flight delay minutes data.

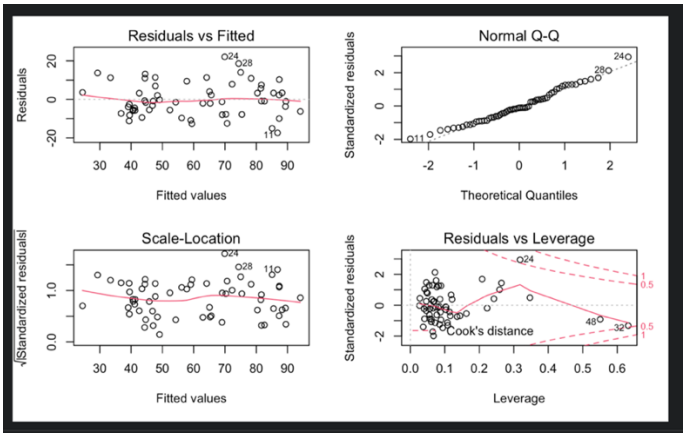
Model Overview

Model 4 is one of approximately 40 models developed that focused on stock price and delays. Model 4 focuses on United airlines and counts of delayed flights. Model 4 did find correlation with a dependent variable of the month end opening United stock price (Open.Stock.Price) to:

- The sum of monthly carrier delays attributed to United (Sum.of.carrier_ct)
- The sum of monthly weather delays for United (Sum.of.weather_ct)
- The sum of monthly United Airlines cancellations (Sum.of.arr_cancelled)

The independent variables were significant and correctly attribute a negative correlation to United's stock price. We also had R-squared values well above the financial industry standard of 0.7 at 0.8. This shows that 80% of the variability observed in United's stock price is explained by the regression model. Delta and Southwest airlines were determined by this study to not have correlation between stock price and delays.

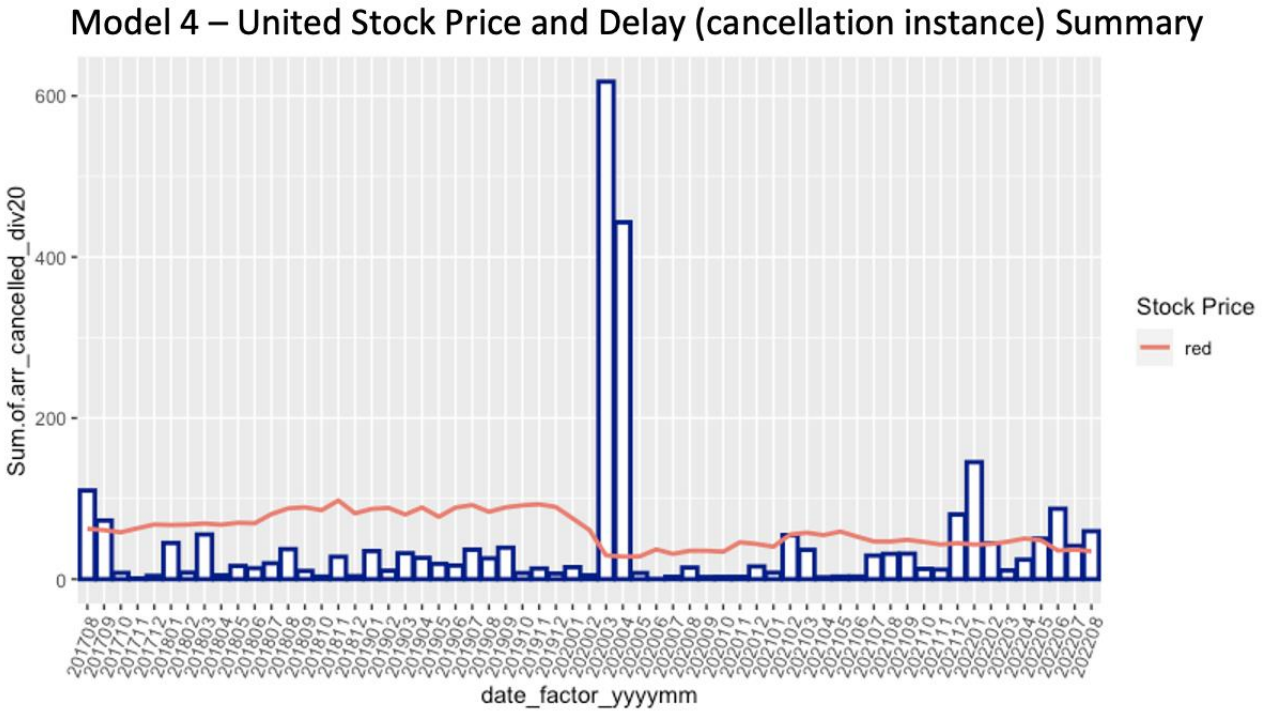
```
Call:
lm(formula = Open.Stock.Price ~ Sum.of.carrier_ct + Sum.of.weather_ct +
    Sum.of.nas_ct + Sum.of.security_ct + Sum.of.late_aircraft_ct +
    Sum.of.arr_cancelled, data = ual_data)

Residuals:
    Min      1Q  Median      3Q     Max
-17.400  -5.901  -1.010   5.740  22.119

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          40.2938833  3.2669519  12.334  < 2e-16 ***
Sum.of.carrier_ct    -0.0097907  0.0021966  -4.457 4.23e-05 ***
Sum.of.weather_ct    -0.0168770  0.0082410  -2.048  0.04544 *
Sum.of.nas_ct         0.0154198  0.0020535   7.509 6.15e-10 ***
Sum.of.security_ct   -0.5181368  0.3495411  -1.482  0.14406
Sum.of.late_aircraft_ct  0.0009860  0.0030146   0.327  0.74487
Sum.of.arr_cancelled    -0.0016752  0.0006269  -2.672  0.00995 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.089 on 54 degrees of freedom
Multiple R-squared:  0.8269,    Adjusted R-squared:  0.8076
F-statistic: 42.99 on 6 and 54 DF,  p-value: < 2.2e-16
```



The Residuals vs Fitted plot indicates a linear model absent of non-linear shapes and minor clustering. The Normal Q-Q is good with tails trailing off slightly. The Spread-Location plot is good with residuals spread equally. The Residuals vs Leverage plot is a scattered somewhat but no top right or bottom right possible high-influence

variables. The worst data point in the Residuals vs Leverage is 32 which is the beginning of COVID onset in the United States in March 2020.

X. Analysis Supporting Project Objective, Supporting Visuals, and Results



This is a graph of United cancellations by month divided by a factor of 20 depicted by the blue bars. The red line overlay is stock price. With careful review the stock price trends inversely with cancellations. It is also interesting to point out in the graph that stock price declined sharply in January 2020 and the huge number of cancellations in March and April 2020 with US COVID onset.

<u>Additional results from Model 4</u>

Data is bifurcated

- Flight count
- Minutes late
- Detail and aggregation for both types of data
- Data dictionary difficult to find and requires careful work to familiarize (Appendix 1)

Many R models developed by the group, over 100 - in GitHub Final Code folder and Code folder

Traders extensively study and use this type of data

- Probably daily or even more granular
- Probably purchased – free sources of daily data were not found

XI. Business Impact and Bottomline

TDV completed significant portions of the Final Project requirements prior to the group-project-progress-report submission. TDV completed tasks including coming up to speed on domain knowledge, finding multiple data sets, analysis of data, cleansing of data, combination of multiple data sets, modeling, and initial visualizations. The team has confidence that prior work, collaborative efforts, collaborative methods/tools, and specified-additional-projected work allow for early completion of the Final Project Report and Final Project Video.

XII. Conclusion

Taking a statistical approach to the business implications of delays, TDV provided solid evidence that delays are about equal across airlines at around 20% and that the count is difficult to predict using only financial indicators. Our research is novel in using several model transformations and combinations as outlined.

Do delays have a negative impact on Airline financial portfolios?
What we found:

- While we found some evidence for delays impacting Airline financials the results were limited to United Airlines with cancellations, United attributed delays, and weather
- Many additional models did not provide adequate results
- The data available to the team was monthly aggregate data
- Data that is at a daily or more granular level would be desirable

What we suggest:

- As a follow-on effort it would be desirable to pursue United Airlines models with more granular (and more expensive) data as a first effort in correlations of delays and financials. Additional airline studies would be considered after further positive results from the United Airlines study.

Bibliography

A4A. (2022, July 12). *U.S. Passenger Carrier Delay Costs*. Airlines for America, Retrieved October 23, 2022 from https://www.airlines.org/dataset/u-s-passenger-carrier-delay-costs/#:~:text=In%202021%2C%20the%20average%20cost,levels%20to%20%2428.14%20per%20minute

Bureau of Transportation Statistics Data (BTS) key website locations:
TranStats
https://www.transtats.bts.gov/homedrillchart.asp
Airline On-Time Statistics and Delay Causes
https://www.transtats.bts.gov/ot_delay/OT_DelayCause1.asp?20=E
Detailed Statistics Departures
https://www.transtats.bts.gov/ontime/departures.aspx
Annual Airline On-Time Rankings 2003-2020
https://www.bts.gov/topics/airlines-and-airports/annual-airline-time-rankings-2003-2020

FAA. (2022, August 31). *Air Traffic By The Numbers*. Retrieved October 24, 2022 from https://www.faa.gov/air_traffic/by_the_numbers

Provost, F, & Fawcett, (2013). *Data Science for Business: What you need to know about data mining and data analytic thinking*, Amazon Audible)

Literature Review

DOT Airline Customer Service Dashboard
https://www.transportation.gov/airconsumer/airline-customer-service-dashboard

DOT Flight Delays & Cancellations
https://www.transportation.gov/individuals/aviation-consumer-protection/flight-delays-cancellations

DOT Airline Customer Service Dashboard
https://www.transportation.gov/airconsumer/airline-customer-service-dashboard

DOT Flight Delays & Cancellations
https://www.transportation.gov/individuals/aviation-consumer-protection/flight-delays-cancellations

Largest airlines in the world
https://en.wikipedia.org/wiki/Largest_airlines_in_the_world

How Airlines Get Airport Slots & Why They Cost So Much
https://simpleflying.com/airport-slots/

Wave of Airline Flight Delays This Year Mostly Self-Inflicted

https://www.bloomberg.com/news/articles/2022-07-15/flight-delays-in-us-linked-to-airlines-more-than-government

The Summers Worst US Airports for Flight Delays and Cancellations
https://www.travelpulse.com/news/airlines/the-summers-worst-us-airports-for-flight-delays-and-cancellations.html

It's a pain to fly these days. The FAA and airlines are trying to fix that
https://www.cnbc.com/2022/06/14/faa-airlines-work-to-reduce-summer-travel-delays.html

Timeline of the COVID-19 pandemic in the United States (2020)
https://en.wikipedia.org/wiki/Timeline_of_the_COVID-19_pandemic_in_the_United_States_(2020)

List of the busiest airports in the United States
https://en.wikipedia.org/wiki/List_of_the_busiest_airports_in_the_United_States

High Customer Satisfaction with North America Airlines, J.D. Power Finds
https://www.jdpower.com/business/press-releases/2021-north-america-airline-satisfaction-study

DOT Airline Customer Service Dashboard
https://www.transportation.gov/airconsumer/airline-commitments-PDF-version

(AfA, 2022)
Airlines for America
Data and Statistics
https://www.airlines.org/dataset/u-s-passenger-carrier-delay-costs/#:~:text=In%202021%2C%20the%20average%20cost,levels%20to%20%2428.14%20per%20minute.

Statistica
American Airlines brand awareness, usage,…, 2022
https://www.statista.com/forecasts/1335247/american-airlines-airlines-brand-profile-in-the-united-states

FAA
Air Traffic by the Numbers
https://www.faa.gov/air_traffic/by_the_numbers

NOAA data
https://www.ncei.noaa.gov/cdo-web/search?datasetid=GHCND

Appendix 1 – BTS Data Dictionary

- Airline_Delay_Cause dataset:
- arr_flights: Number of flights which arrived at the airport.
- Arr_del15: Number of flights delayed (>= 15minutes late).
- carrier_ct: Number of flights delayed due to air carrier (e.g. maintenance or crew problems, aircraft cleaning, baggage loading, fueling, etc.).
- weather_ct: Number of flights delayed due to weather.
- nas_ct: Number of flights delayed due to National Aviation System (e.g. non-extreme weather conditions, airport operations, heavy traffic volume, and air traffic control).
- security_ct: Number of flights delayed due to security (e.g. evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines in excess of 29 minutes at screening areas).
- late_aircraft_ct: Number of flights delayed due to a previous flight using the same aircraft being late.
- arr_cancelled: Number of cancelled flights.
- arr_diverted: Number of diverted flights.
- arr_delay: Total time (minutes) of delayed flights.
- carrier_delay: Total time (minutes) of delayed flights due to air carrier.
- weather_delay: Total time (minutes) of delayed flights due to weather.
- nas_delay: Total time (minutes) of delayed flights due to National Aviation System.
- security_delay: Total time (minutes) of delayed flights due to security.
- late_aircraft_delay: Total time (minutes) of delayed flights due to a previous flight using the same aircraft being late.