**Researching the Regularity and**

**Elements of Flight Delays**

Apr 21st, 2021

University of Texas Dallas

BUAD 6356\_Project team 2 report

Group Member: Sameeullah Babar, Ye Young Kim, Keren Huang, Zhenkun Zang

**Executive Summary**

Flight delays and cancellations have become an important topic and issue for air transport systems around the world. The aviation industry continues to suffer economic losses associated with flight delays. About 20% of U.S. flights were delayed in 2018, according to data from the U.S. Transportation Bureau (BTS). These flight delays have serious economic implications of $31.2 billion per year in the US. In this paper, we examined the factors impacting aircraft delays and cancellations, and if winter season specifically has a higher impact on flight delays. For this analysis, we used a large database set which is airline delay and cancellation data from 2016 to 2018, and US weather events from 2016 to 2018, extracted from Kaggle. A number of multivariate statistical techniques were used for our analysis. We concluded that NAS delays, which is the type of delay that is within the control of the National Airspace System, such as non-extreme weather conditions, airport operations, heavy traffic volume or air traffic control, has the highest impact for flight delays and winter season is not a key factor for causing the flight delays.

**A.Introduction**

Based on the Bureau of Transportation Statistic’s report, flight delays are caused by carrier delay, national aviation system delay (NAS) and extreme weather. In this project, we pulled out delayed flight information to find out which component most causes the flight delay and cancellation. Our hypothesis was that flights that take place in winter will experience longer arrival delays. The aim of this analysis is to predict the length of flight arrival delays in minutes based on various factors and especially due to winter weather patterns. This analysis can help travelers to predict flight delays in real time and will allow for air traffic control and airport personnel to better allocate air space and runway time.

**B.Data Description**

Flight delays and cancellation data has been collected to incorporate multi-year data to offer additional time series insights. This dataset contains 2.6 millions records between 2016 and 2018 and includes information on 28 variables related to each delayed flight, including flight data, origin, destination, departure time, departure delay, arrival time, and arrival delay, and so on. In addition to flight information, weather data has been collected using historical weather reports that were collected from 2,071 airport-based weather stations.

**C.Preprocessing Data**

When we started analyzing, first we cleaned up “NA” and missing data. Then we deleted the variables listed below before starting our data analysis.

CRS\_DEP\_TIME: Schedule departure time

DEP\_TIME: Actual depart time

WHEELS\_OFF: Wheel up undercarriage

WHEELS\_ON: Wheel down undercarriage(landing)

CRS\_ARR\_TIME: Schedule arrival time

ARR\_TIME: Actual arrival time

CANCELLED: 0 means not canceled, 1 means canceled

DIVERTED: Diverted to a different airport

CRS\_ELAPSED\_TIME: Schedule flying time in the air

SECURITY\_DELAY: delay caused by security reason

We used the rbind function to merge 2016, 2017 and 2018 data for flight delays. We then created a new column in our data table named IS\_WINTER, which detailed whether the flight in question took place in winter. This variable is determined based on the flight date not taking place from March 15th to October 15th of that year. While Winter is generally considered to be from December through February, winter weather begins to present itself in October and will usually persist through March. Including IS\_WINTER, we used the following variables:

Fl\_Date: flight date

OP\_Carrier: Flight company such as UA, AA, Delta, etc

OP\_CARRIER\_FL\_NUM: Flight number

Origin: Departure airport

DEST: Arriving airport

DEP\_DELAY: Departure delay time

TAXI\_OUT: From gate to runway

TAXI\_IN: Runway to gate

ARR\_DELAY: Arrival delay time

ACTUAL\_ELAPSED\_TIME: Actual flying time in the air

AIR\_TIME: After wheel off and before wheel on

DISTANCE: Flight route

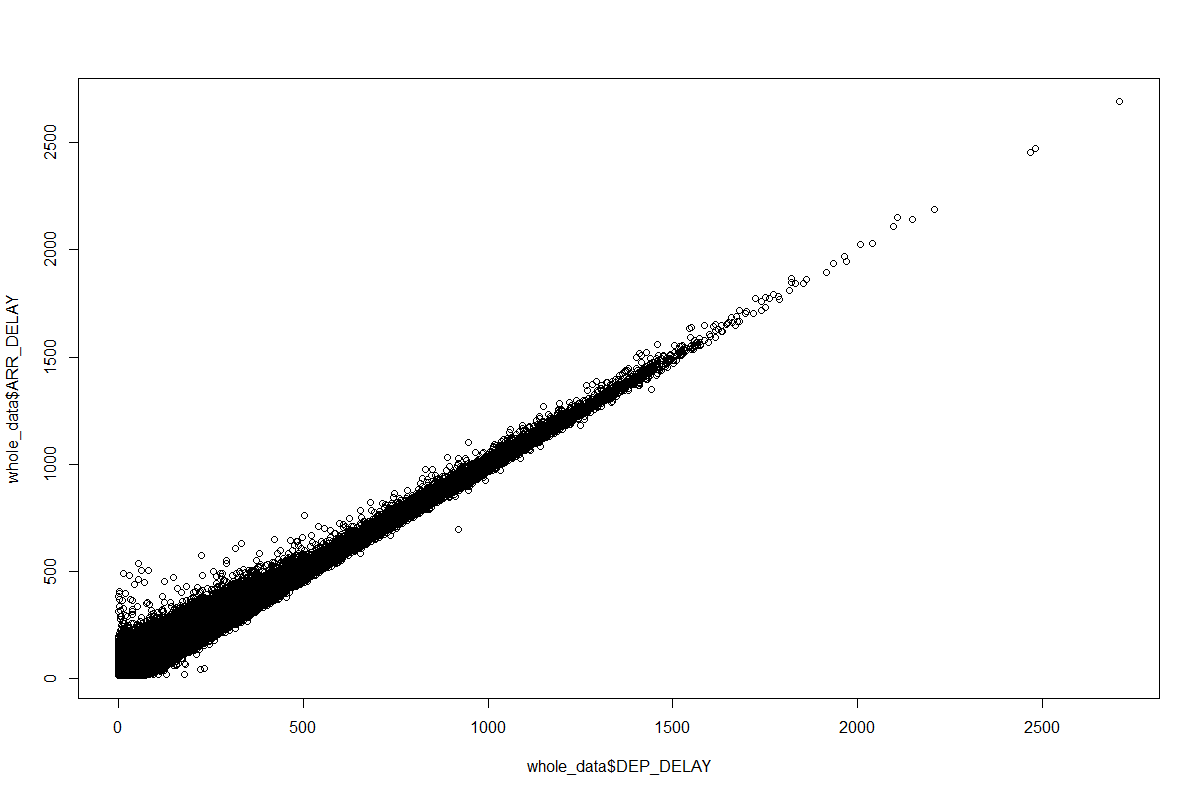
CARRIER\_DELAY: delay caused by airline’s fault

WEATHER\_DELAY: delay caused by bad weather

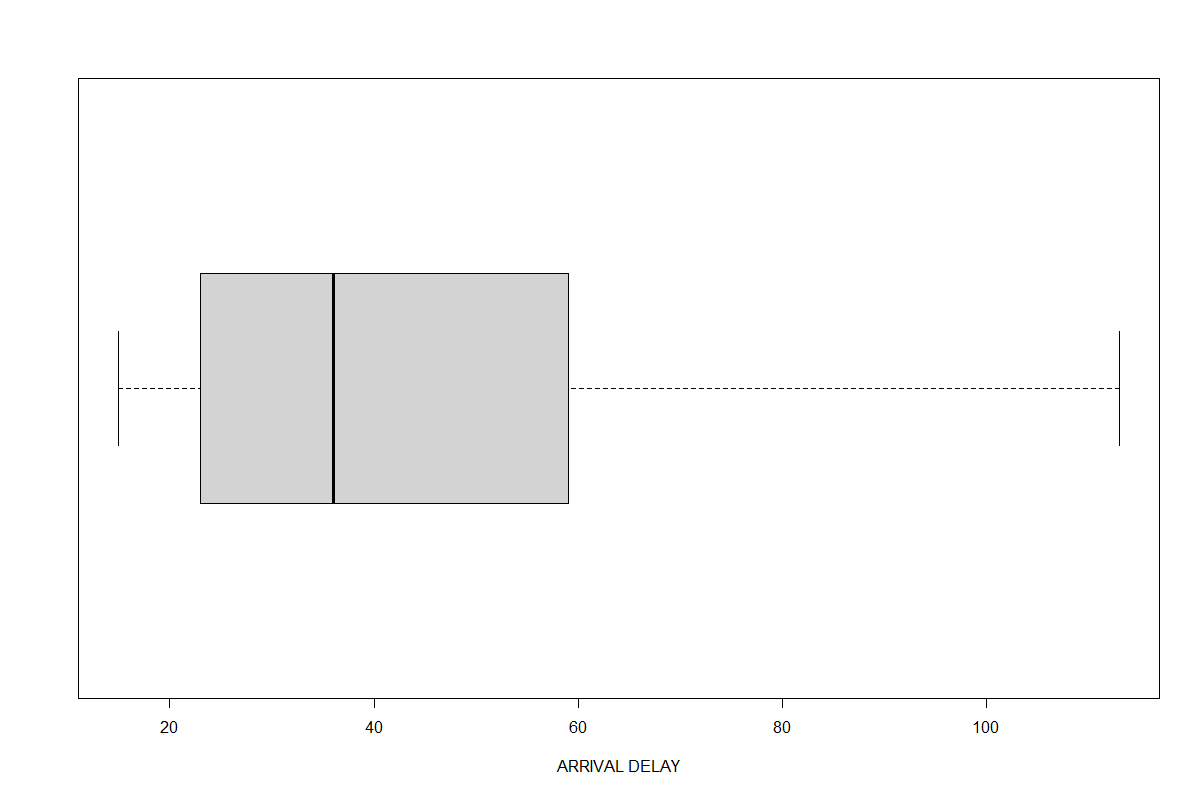
NAS\_DELAY: delay caused by air traffic control

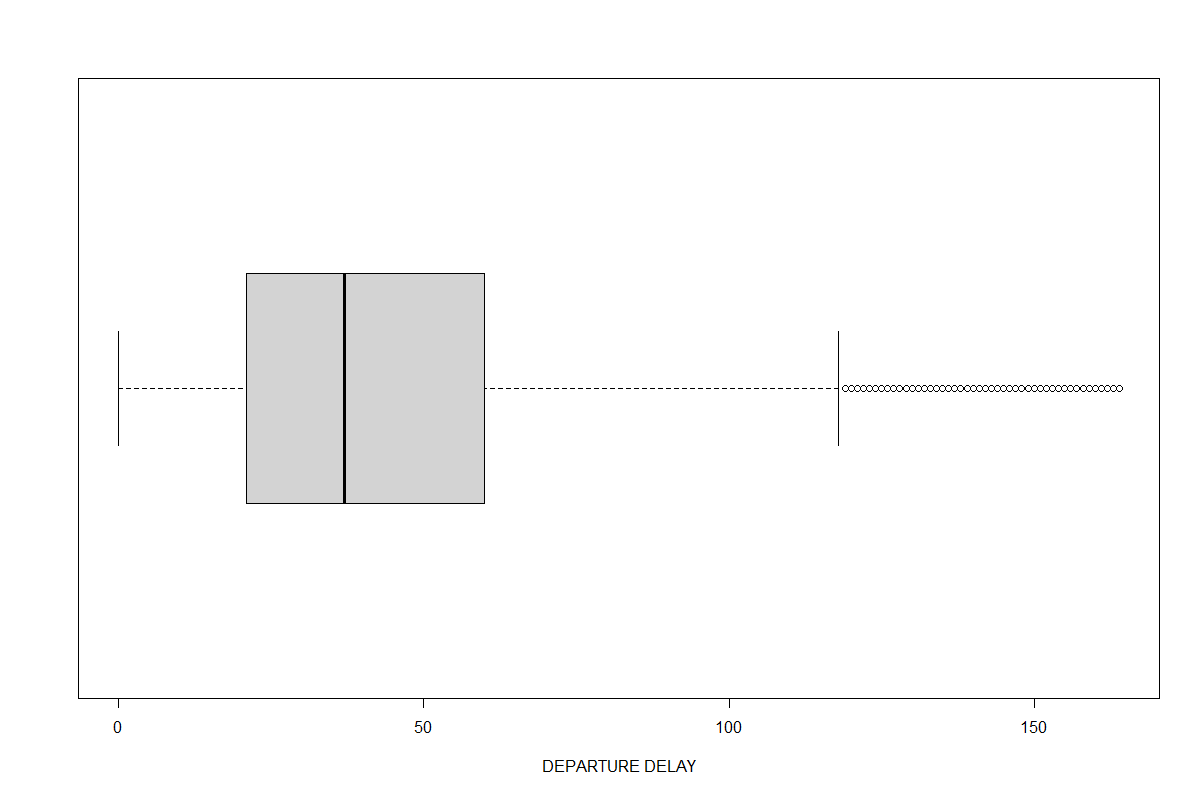
LATE\_AIRCRAFT\_DELAY: delay caused by the airplane arrives late

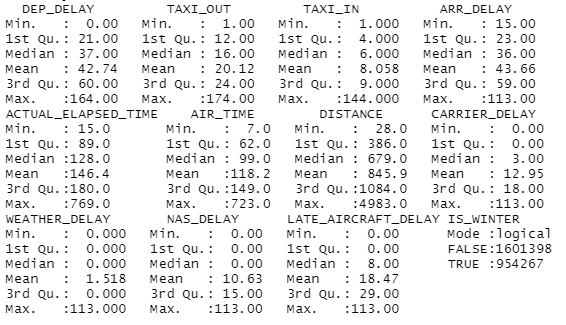
**D.Exploratory Data Analysis**



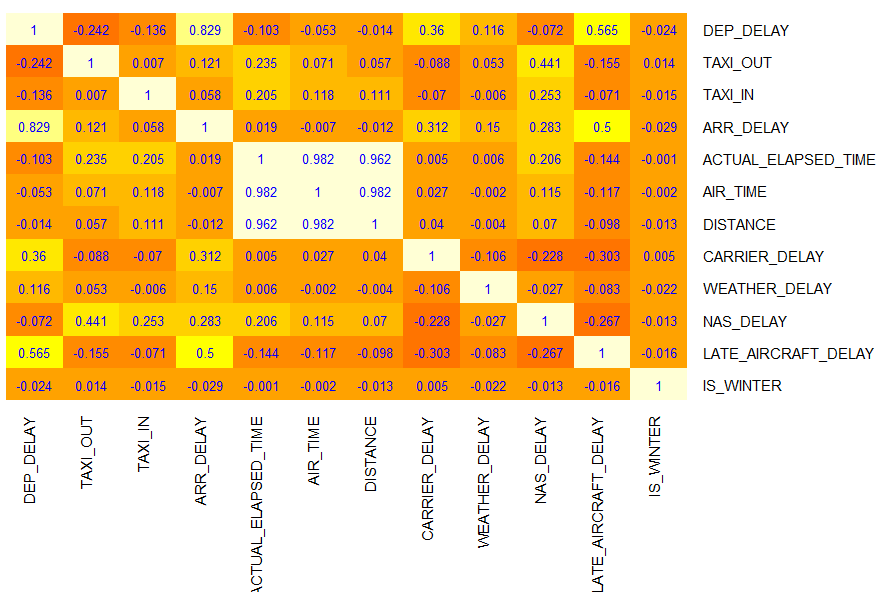
The unfiltered scatterplot of Departure Delay and Arrival Delay show a mostly linear relationship. The data is mostly homoskedastic. There are a lot of noteworthy outliers with delays going into the thousands of minutes. We decided to make the outlier cutoff the sum of the Interquartile ranges of departure delay and 1.5 times the mean. This decision was made due to our desire to predict more realistic delays as opposed to significant weather events such as blizzards or hurricanes.

The boxplots below show the outlier-free distributions of Departure delays and Arrival delays.



The table below shows the summary distributions of our variables. The mean Departure Delay is 42.74 minutes and the mean Arrival Delay is 43.66 minutes. The mean weather delay is 1.5 minutes, while the mean NAS delay is 10.63 minutes. 

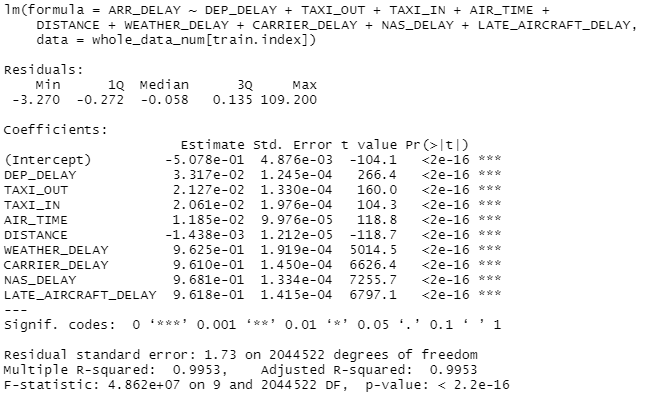
Below is a heatmap showing the correlations among all the variables.



The lighter the color is, the two variables have a stronger positive correlation. The darker the color is, the two variables have a stronger negative correlation. For example, if taxi\_out decreases, then the Dep\_delay would also decrease. If LATE\_AIRCRAFT\_DELAY increases, then DEP\_DELAY would increase as well. DISTANCE and AIR\_TIME have the highest positive correlation of 0.962. LATE\_AIRCRAFT\_DELAY and CARRIRIE\_DELAY have the lowest negative correlation of -0.303.

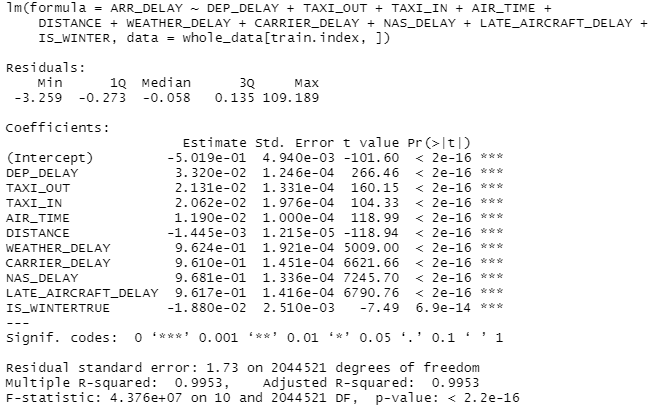
**E.Empirical Analysis**

We ran a Sequential Linear Regression model using Arrival Delay as our result variable. First, we made a Linear Regression model with Departure Delay, TAXI\_OUT, TAXI\_IN, Air Time, Weather Delay, Carrier Delay, NAS Delay, and Late Aircraft Delay as our predictors. The data was split into training and validation data at a ratio of 80:20 respectively. Our model showed that NAS Delay has the biggest effect on the result variable, with the coefficients for Weather Delay, Carrier Delay, and Late Aircraft Delay being very slightly smaller. Weather Delay specifically has a coefficient of 0.9625, meaning that for every minute a flight is delayed due to weather, the delay in arrival will increase by 0.925 minutes(57.7 seconds). All predictor variables were significant well past the 99 percent level and the model is significant with an adjusted R-Squared of 0.9953. Using the accuracy function and our validation data to determine model accuracy, we found that our model had a Root Mean Squared Error of 1.756 and a Mean Absolute Error of 0.338.





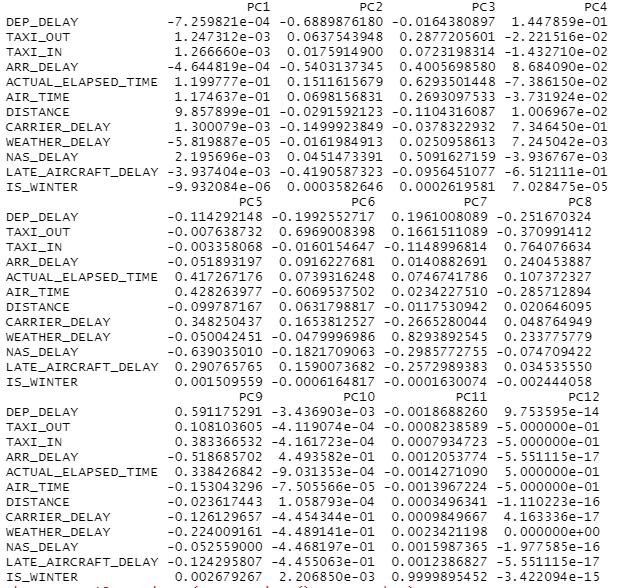
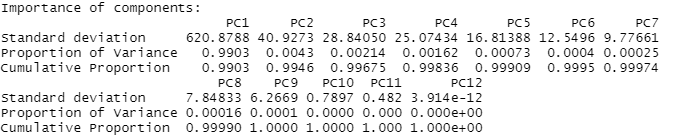
After running the aforementioned regression, we then added the variable IS\_WINTER to the model. We then used our training data and refit the data with the new predictors. We found that the coefficients for the other predictors remained the same, but when our IS\_WINTER variable was TRUE, the predicted Arrival Delay would decrease by 0.0189 minutes(1.1 seconds). IS\_WINTER was statistically significant at the 99% level. The adjusted R-Squared is the same as the previous model, 0.9953.



However, when determining the accuracy of our new model, we found that the Root Mean Squared Error was 35.851 and the Mean Absolute Error was 27.7. These error values show that our new model will generally over-estimate the error by approximately half an hour. This significant increase in error shows that our model is overfitting the data, leading to higher error.



This can be further seen in our Principal Component Analysis, where IS\_WINTER does not provide impact on the principal components until the 11th of 12 principal components. This component accounts for less than .01% of the total variance and gives credence to the argument that including IS\_WINTER in our model has led to an overfitted and less accurate model. Therefore it is prudent to exclude IS\_WINTER from our model despite its statistical significance as it only affects the result variable by 1 second and causes much more error than if it is omitted from the analysis.



**F.Conclusions**

Based on the flight delay dataset, we analyzed the biggest factors affecting flight delays. We concluded that National Air Service(NAS) delays were the most significant, with weather delays and carrier delays shortly behind. However, Winter weather is not a good predictor as it leads to much more error. Our analysis shows that airline companies need to have more effective schedules to improve airport operation controls and avoid heavy traffic volume because NAS delay can be controlled and reduced by rearranging operations.

**Sources:**

Dataset

*Airline Delay and Cancellation Data, 2009 - 2018*. (2019, August 11). [Dataset]. https://www.kaggle.com/yuanyuwendymu/airline-delay-and-cancellation-data-2009-2018

*US Weather Events (2016 - 2020)*. (2021, January 18). [Dataset]. <https://www.kaggle.com/sobhanmoosavi/us-weather-events>

2018<https://drive.google.com/file/d/1nKQvYxo_Sp_jlYn-EkDv6zyVFto1Nbtp/view?usp=sharing>

2017 <https://drive.google.com/file/d/1uoD51s_j4Do-gXIL1u72dJTmqSayGHMK/view?usp=sharing>

2016 <https://drive.google.com/file/d/1otdU120wwONCoKHXebcpoWW1fOMIuanO/view?usp=sharing>

Website

*Bureau of Transportation Statistics*, Airline Service Quality Performance 234. “On-Time Arrival

Performance National (January - December, 2018)” https://www.transtats.bts.gov/OT\_Delay/OT\_DelayCause1.asp?20=E

*Impact of Flight Disruptions*. (2019, October 24). Travel Incorporated. https://www.travelinc.com/impact-of-flight-disruptions/

Links to Data:

2018:

<https://drive.google.com/file/d/1nKQvYxo_Sp_jlYn-EkDv6zyVFto1Nbtp/view?usp=sharing> 2017:

<https://drive.google.com/file/d/1uoD51s_j4Do-gXIL1u72dJTmqSayGHMK/view?usp=sharing>

2016:

<https://drive.google.com/file/d/1otdU120wwONCoKHXebcpoWW1fOMIuanO/view?usp=sharing>