Airline Delays Demystified

How airlines make decisions based on their business models and how delays play spoilsport

Team Data Pirates

Abhishek Ghosh

Moumita Mukherjee

Sukanto Guha

Index

Introduction

The upsurge in the urban middle class population is driving more and more people to air travel in the United States (US). The aviation industry has been riding its highest wave since 9/11, with airlines returning record revenues. We analyze the data for all of 2018 and half of 2019, which was part of this golden period for airlines.

The journey towards high revenues was a combination of many factors, one of which was the vicious cost-cutting to present better fares to end customers and increase flight occupancy - something we'll touch upon in parts of our analysis.

The majority of the data provided is related to another issue which costs airlines - and hence consumers - money and annoyance. This is the issue of **flight delays**. Delays cost airlines about $8 billion a year, and customers about $17 billion. This link also gives a breakup of the costs, in particular, many of the costs are amortized costs like crew salaries or aircraft maintenance - so reducing flight delays has to be achieved by other modes which we discuss below.

In the US alone, there are 17 different carriers that fly to domestic destinations. There is a steep competition between these airlines, and they have kept their net income margin low to compete for low prices in their respective **categories** - we will be discussing more on how these categories are created below. Essentially, low-cost carriers do not compete on the same factors against full-fare carriers like Delta, and this trend causes us to analyze the delays of these airlines keeping in mind their modus operandi.

Satisfaction With Specific Aspects of Flying Experience, by Frequency of Air Travel, 2007 Gallup Poll

% Satisfied

	Frequent traveler	Less frequent traveler
	%	%
Courtesy of flight attendants	92	92
Courtesy of check-in/gate agents	86	91
On-time performance	65	73
Price of tickets	65	65
Schedules	75	82
Luggage systems	69	80
Comfort of seats	42	52
Security procedures	63	74
Efforts to deal with delays and cancellations	52	60

GALLUP POLL

Figure 1

Finally, to gain a competitive advantage in the market, an airline needs to cater to the needs of the customer. According to the Gallup poll shown above, 35% of frequent customers and 27% of infrequent travelers were dissatisfied by delays - and additionally, more than 40% with the *efforts* taken by airlines to combat these delays - which lead us to believing the cause behind delays requires significant in-depth analysis.

Our predictions, models and analysis hold for regular business years.
For anomalous times like the current Coronavirus pandemic which has ground the industry to a halt, different models would have to be made - which would be an interesting exercise by itself, but it is out of the scope of our report.

Executive Summary

Problem Statement

We will work on the problems stated below and provide data driven recommendations based on our predictive models, providing insight which may benefit airlines but will definitely benefit consumers in the airline industry:

- Forecasting delays for Q3 2019 and beyond based on data available **before** a flight leaves
- Analysis of when to take flights on which airline
- Analysis of airline routes and their implications on flight delay

Data Collection

Data Collection

The dataset that we are analyzing is from the Bureau of Transportation Statistics which includes all the flight details that have flown in the United States from January 2018 to June 2019.
The data set consists of four data files:

File Name	Size
FlightDelays	10915495 x 51
Routes	6684 x 10
AirFares	80344 x 20
Airports	363 x 3

Table 1

FlightDelays is the file which we used most heavily, as it comprises the bulk of delay data.
It's columns comprise the majority of features which we use to predict delays of an airline. All the details of each flight - when it took off, when it taxied, when it arrived, what kinds of delay it faced, is given. Additionally, this csv also has some information which is airline specific like revenues and income.

Routes is a supplemental csv, containing information about the origin, destination and length of each airplane route in the US. We mainly used Routes to cross check and add information to our two main data files - FlightDelays and AirFares.

AirFares contains the information about the average airfare on each route for each quarter. Additionally it contains the fares charged by the smallest and largest airline on a route. We use this to see how routes with a lot of competition behave in comparison to monopolistic ones and how that affects delays.

Airports lists all the US airports. This file was a supplement to FlightDelays.

After initial analysis, we found out that about 2 million of the 10 million flights were delayed more than 15 minutes. This is in line with what we expected. There are 5 different types of delays namely NAS, Security, Late Aircraft, Carrier and Weather. These delays result in Departure and Arrival delay.

Data Preprocessing

Data Cleaning

YEAR	0
QUARTER	0
MONTH	0
DAY_OF_MONTH	0
DAY_OF_WEEK	0
FL_DATE	0
CARRIER	0
FL_NUM	0
Route	0
ORIGIN	0
DEST	0
DEST_CITY	0
DEST_STATE	0
CRS_DEP_TIME	0
DEP_TIME	195282
DEP_DELAY	200219
DEP_DELAY_NEW	200219
DEP_DEL15	200219
DEP_DELAY_GROUP	200219
DEP_TIME_BLK	0
TAXI_OUT	200782
WHEELS_OFF	200781
WHEELS_ON	206390
TAXI_IN	206390
CRS_ARR_TIME	0
ARR_TIME	206389
ARR_DELAY	232638
ARR_DELAY_NEW	232638
ARR_DEL15	232638
ARR_DELAY_GROUP	232638
ARR_TIME_BLK	0
CANCELED	0
CANCELLATION_CODE	10713410
DIVERTED	0
CRS_ELAPSED_TIME	144
ACTUAL_ELAPSED_TIME	230040
AIR_TIME	230040
DISTANCE	0
CARRIER_DELAY	8832232
WEATHER_DELAY	8832232
NAS_DELAY	8832232
SECURITY_DELAY	8832232
LATE_AIRCRAFT_DELAY	8832232
PASSENGERS	15418
EMPFULL	15418
EMPPART	15418
EMPTOTAL	15418
EMPFTE	15418
NET_INCOME	15418
OP_REVENUES	15418

dtype: int64

Missing Values:

For *FlightDelays*, there were missing values for some of the delays which is comprehensible. For example, the delays have null value so we assumed that those flights did not incur any delay and arrival delay ,origin, destination were removed which reduced the number of observations to 10,709,106.

For *AirFares*, there was some data with NaN values, belonging to Q3 of 2019. We set these to the values obtained from the actual 2019 Q3 data online. Airlines with fewer than 10 data points - namely, Cape Air(9K) and Virgin Air(VK) were removed from analysis as there wasn't sufficient data to form any insight for them.

Datatype:

Numeric: We had numerical data types for the most part and we did not change it. For example, all the delays were stored as float, month and year were stored as int.

Categorical: We optimized because it creates a list of pointers to the memory address for each value in the column. For example, origin, destination and carrier had low cardinality so we converted it into category type by mapping it to integers

Categorical Encoding:

The purpose of this was to ensure our model can handle categorical variables since they will not work unless it is converted to numerical values.

Label Encoding was done for ordinal datatypes (For example, Scheduled Arrival time, Scheduled Departure Time) and interval datatypes (Example, Wheels_on, Wheels_off, Taxi_in and Taxi_out)

Binary Encoding was done for nominal values (For example, Carrier, Origin, Destination)

Feature Engineering

For *FlightDelays*, we identified that every airline has a slightly different business strategy. We have bucketed them into 4 categories - the ultra low cost, the low cost, the regional and the traditional full service carriers. We have then analyzed delays for groups and tried to find a specific pattern for each of the delays.

For *AirFares*, we added airline revenue and income for both the smallest and largest airline on a route to see its effect on fares.

Data Integrity

We found there are 5 types of delays that result in Arrival Delay. In order to corroborate our assumption, we filtered the flights which showed delay greater than 15 minutes. It consisted of 2083263 rows. When we checked the filtered rows for null values, the 5 attributes had none. On the other hand, the five attributes were null for all those flights where arrival delay was 0 or less than 0.

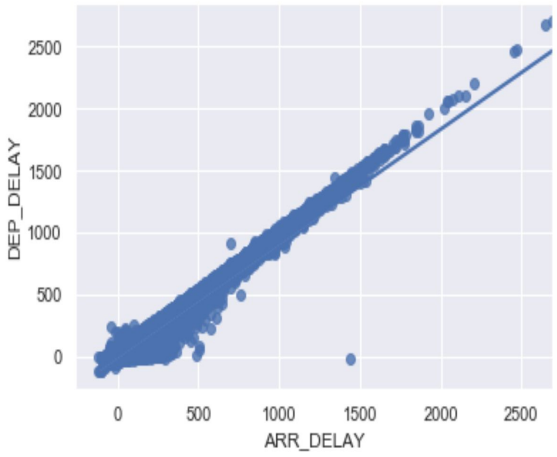
The Carrier, Weather, NAS, Security or Late Aircraft delays are not null only for those flights which have some amount of arrival delay as shown in the figure. Thus, we can safely assume that the data is credible and we can derive arrival delays from the data set.

CRS_ARR_TIME	0
ARR_TIME	0
ARR_DELAY	0
ARR_DELAY_NEW	0
ARR_DEL15	0
ARR_DELAY_GROUP	0
ARR_TIME_BLK	0
CANCELED	0
CANCELLATION_CODE	2083263
DIVERTED	0
CRS_ELAPSED_TIME	0
ACTUAL_ELAPSED_TIME	0
AIR_TIME	0
DISTANCE	0
CARRIER_DELAY	0
WEATHER_DELAY	0
NAS_DELAY	0
SECURITY_DELAY	0
LATE_AIRCRAFT_DELAY	0
PASSENGERS	2799

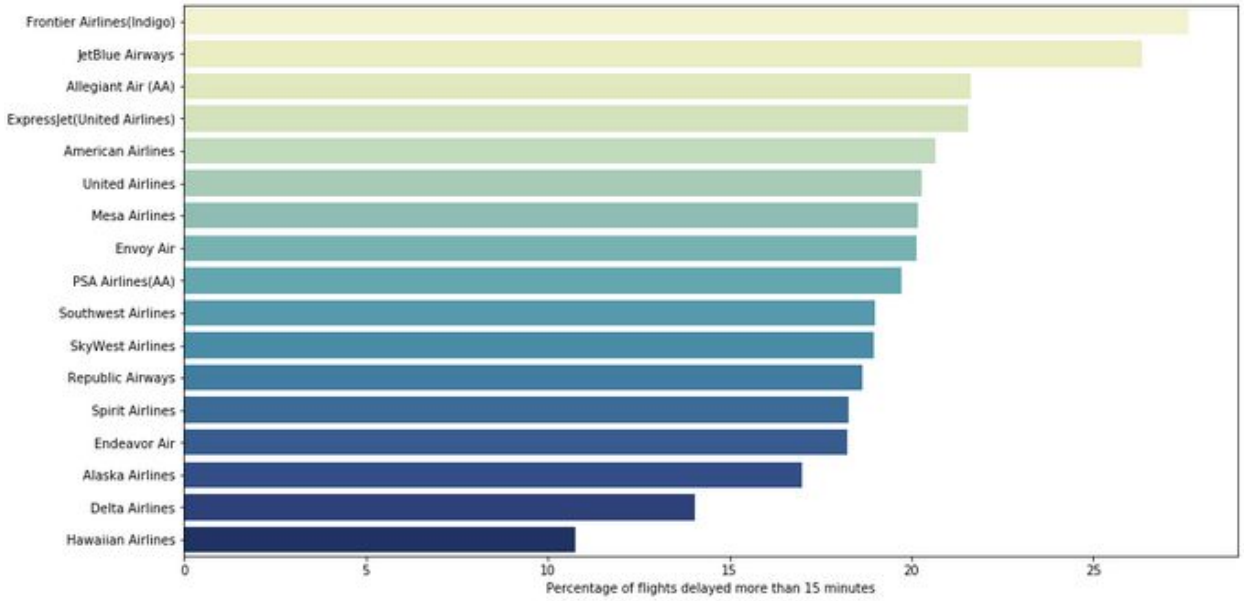
Exploratory Data Analysis

Arrival vs Departure Delay

The departure delay is positively correlated with the arrival delay, predicting one will predict the other. Throughout our analysis, we focus on the Arrival Delay parameter for our analysis unless stated otherwise.



Most Delayed Airline?



A first look at all the airlines and their delays shows that Frontier Airlines has the highest average delay while Hawaiian has the lowest. Does it mean that:

- Are these really the best and the worst performing airlines?
- Is there any external factor that influences the delay or is it solely based on the operational strategy of the airlines?

Data Analysis II - A Deeper Dive into the Data

Categorization of Airlines

There are broadly two categories of systems followed by airlines in the US:

1) **Hub**: It involves a system of connections to destinations around a single *hub*, generally cities that are the largest or most economically viable in their area. With this model, airlines will require that you stop in their hub to connect between two cities, creating the *spokes* of the system.
2) **Point to point**: It emphasizes flying between two cities directly, regardless of size.

The second categorization is based on the the cost of flying and business strategy of an airline:

1) **Ultra Low Cost Carrier (ULCC)**: Cost per mile is $0.1/mile on average. They fly point-to-point and offer the lowest fare to consumers. This group contains: Frontier Airlines(F9), Allegiant Air(G4) and Spirit Airlines(NK).

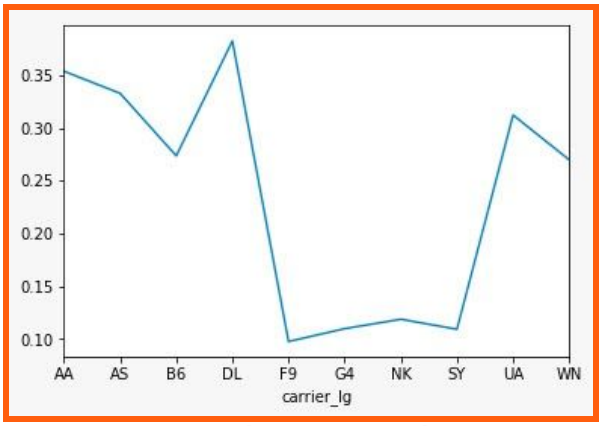


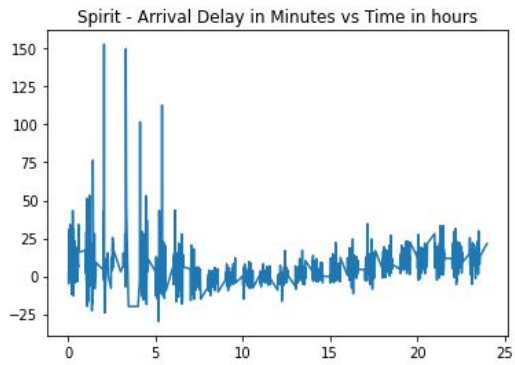
Figure: Airfare per mile vs Carrier

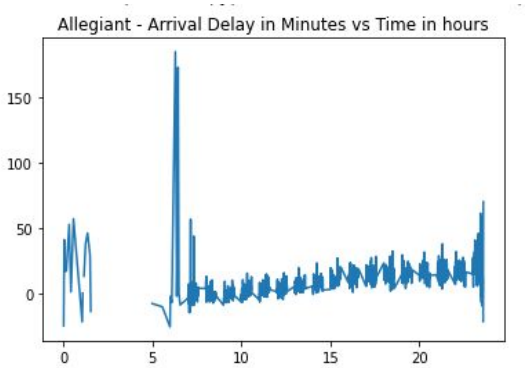
2) **Low Cost Carrier(LCC):** Cost per mile is between $0.25/mile and $0.31/mile on average. Their networks may be either hub and spoke or point to point depending on if they started out as ULCC or FFC. They offer some services over ULCCs and charge more. This group consists of United(UA), Southwest Airlines(WN) and JetBlue(B6).
3) **Full Cost Carrier(FCC):** Full fare airlines, cost per mile > $0.33/mile. Exclusively hub and spoke, consists of Delta(DL), Alaska(AS), Hawaain Airlines(HA) and American Airlines(AA).
4) **Regional Airlines:** These airlines have smaller planes, fewer passengers and fewer flights than the other categories. These are American Eagle(+subsidiaries),Envoy Air, SkyWest and Endeavour Air. We analyze them only in one category where they are relevant to delay.

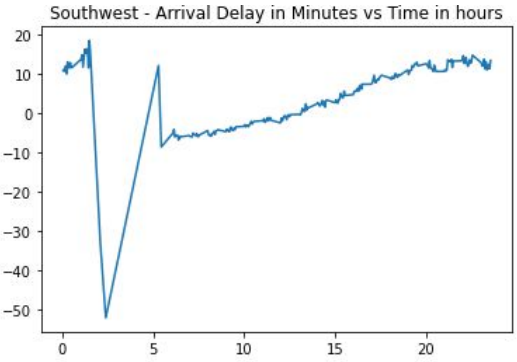
Delays by Time of Day and Category

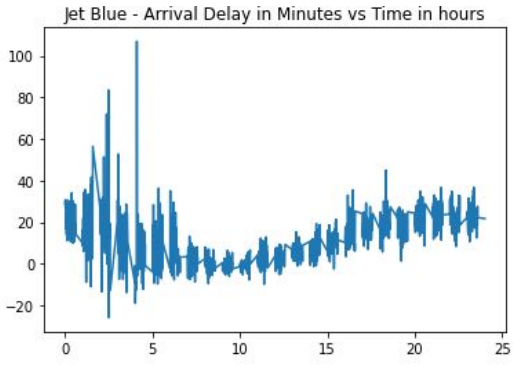
Now that we have categorized our airlines based on our data, we analyze delay patterns over the course of a day for each category.

ULCCs and LCCs:

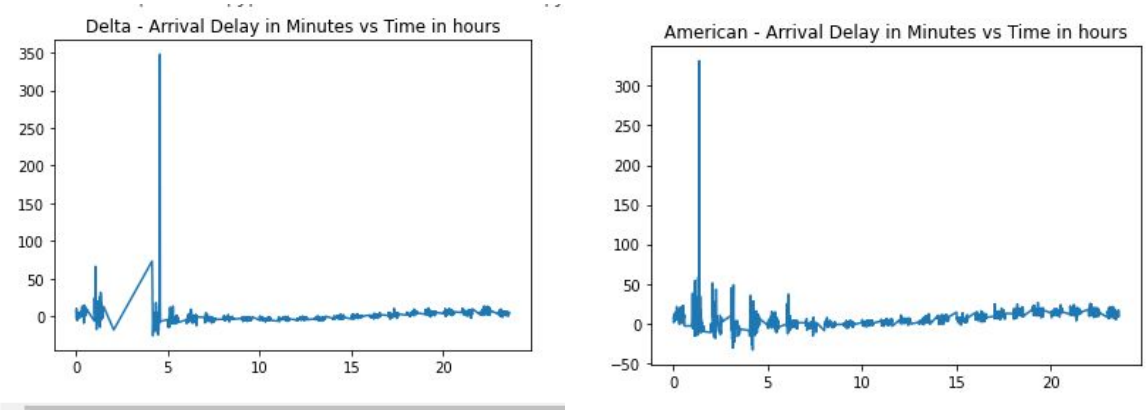




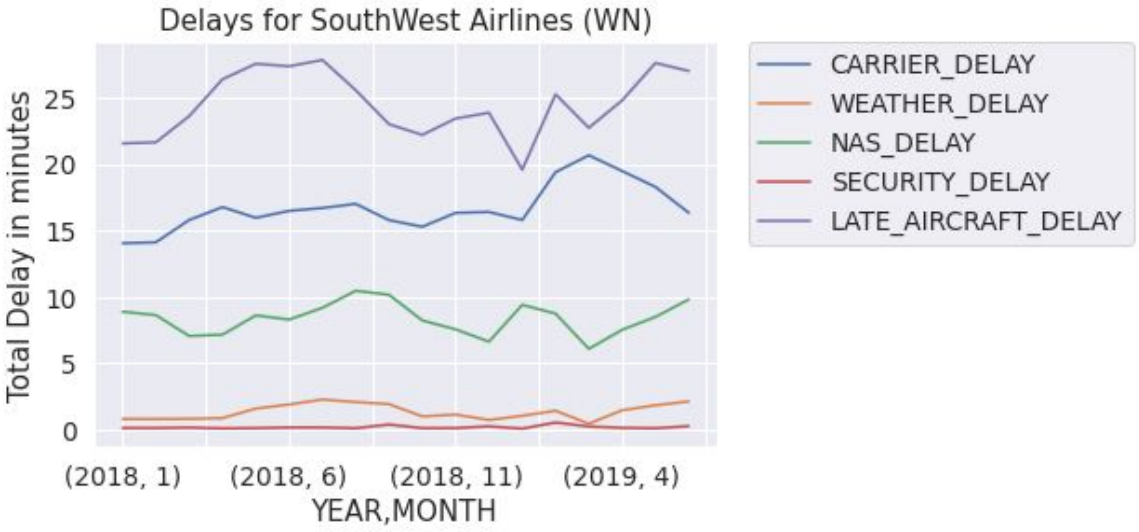




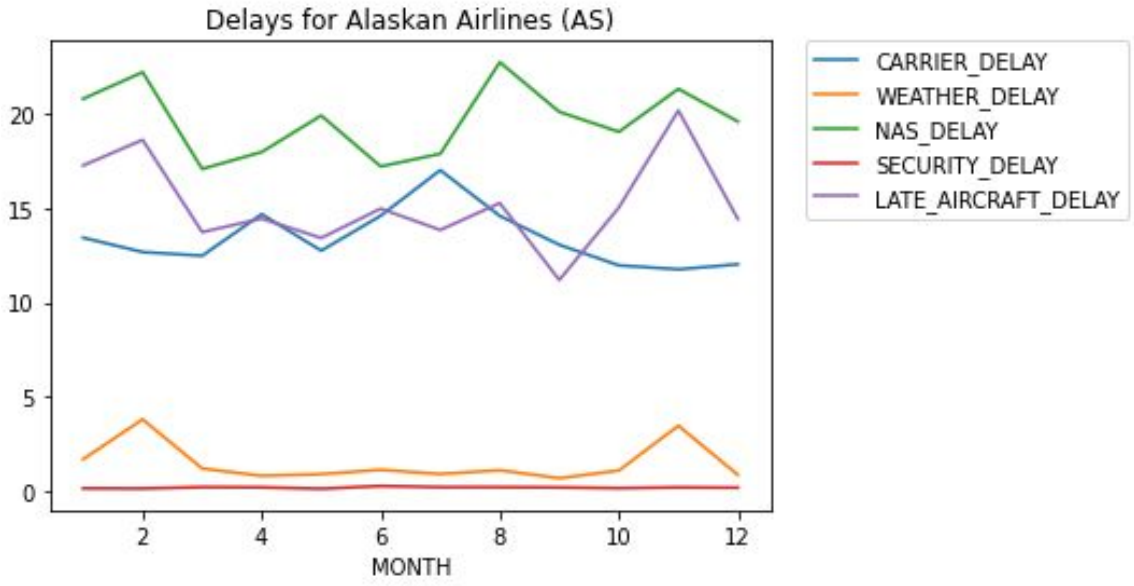
FCCs



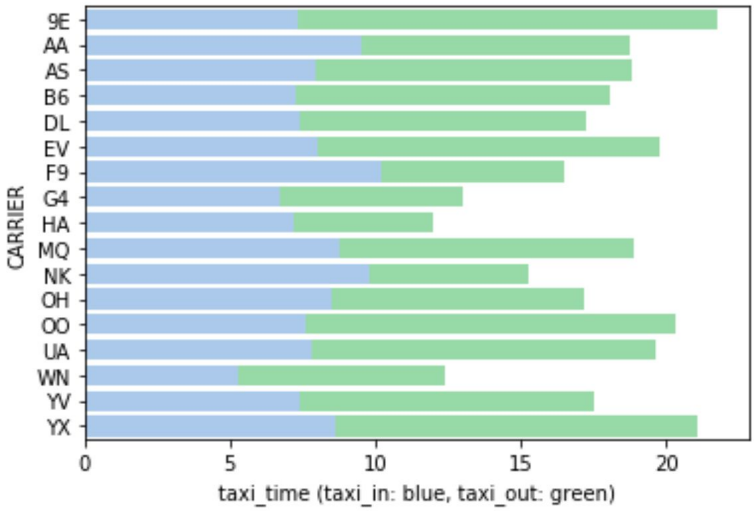
We observe that flying gets progressively more risky on all ULCCs and LCCs as the day progresses, but not on full fare airlines. Our thinking is that the point to point model and rapid turnover of ULCCs causes systemic delays which amplify as the day progresses and peaks with the final flight of the day, before starting over at negative values again.



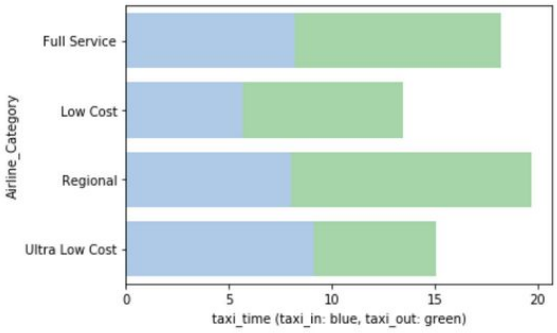
We back our observations above by studying the field LATE_ARRIVAL_DELAYS in our input. We observe that for LCCs and ULCCs, the percentage of late aircraft delay in the total delay is significantly higher than FCCs, as can be seen in the Southwest graph (above) and the Alaskan graph(below).



Delays by Taxi Time and Category



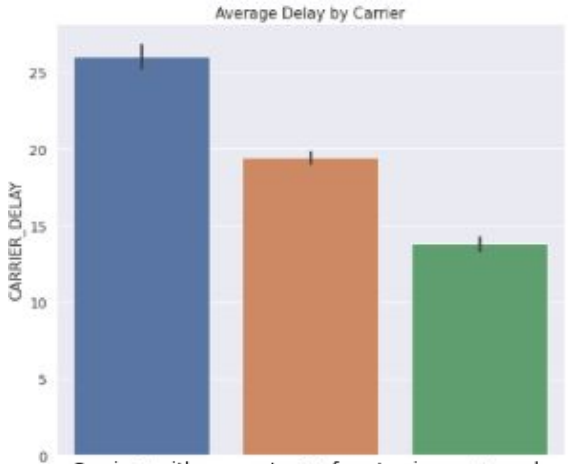
We observe that ULCCs are excellent at saving time here, especially the leaders of their respective groups - for example, Delta spends the least time taxiing - but it falls short when compared to any member of the ULCC category.



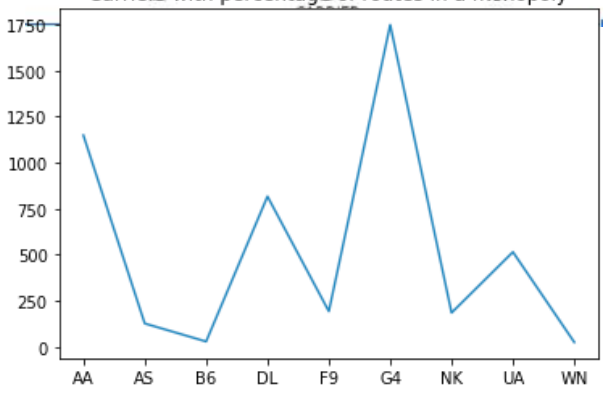
The maximum delay is faced by customers of regional airlines. We hypothesize this is because regional airlines often do not have competition on the routes they fly. We analyze a bit further in the following subsection with the curious case of Allegiant Air.

Relation of Delays and Monopolies - Allegiant Air

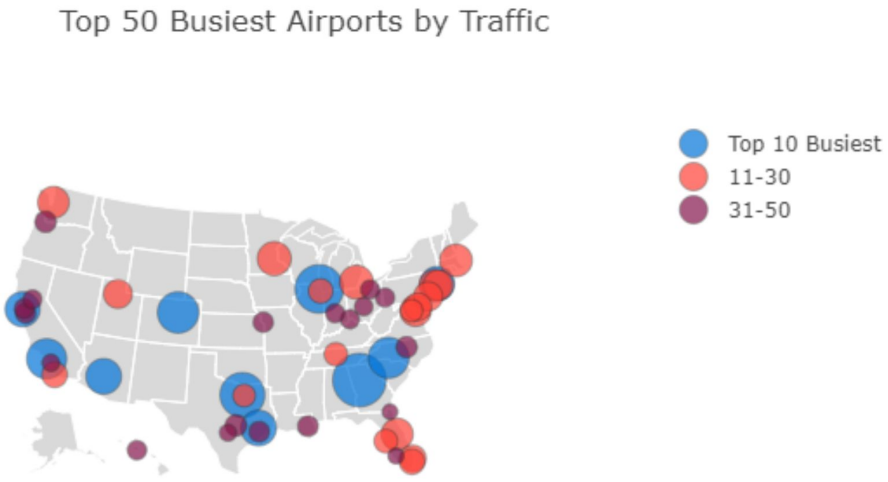
The large delays faced by regional airlines caused us to analyze why their customers would be willing to fly them. Because of the lack of data for all but one regional airline, we focused our attention on Allegiant Air - as it is a ULCC which flies mostly regional routes. Allegiant's delays are far greater compared to other ULCCs as seen on the right, in addition to the fact that it flies the fewest people among ULCCs.



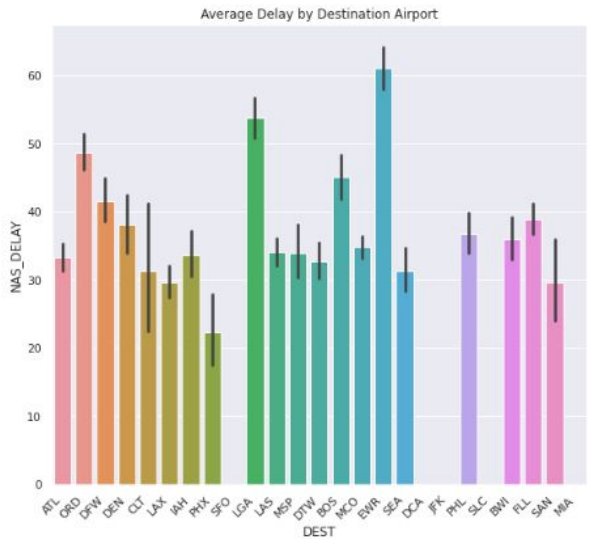
The reasoning behind this could be found in another part of the data: the number of routes which Allegiant flies where it is the sole operator is extremely high as can be seen from the graph on the right. We analyzed from AirFares.csv that lack of competition raises fares on a route. Our analysis shows that the average fare per mile charged by the largest carrier on a route is $0.25/mile, and the smallest carrier can charge $0.2/mile. The only carrier in a monopoly, on the other hand, can charge $0.26/mile. This is in spite of the fact that monopolies are usually shorter (average route length 462 miles vs 1125 miles for 2 or more airlines). This, in addition to US government programs like the [Essential Air Service] which provide incentives to fly routes which are unprofitable make airlines like Allegiant tick - customers often have no choice.

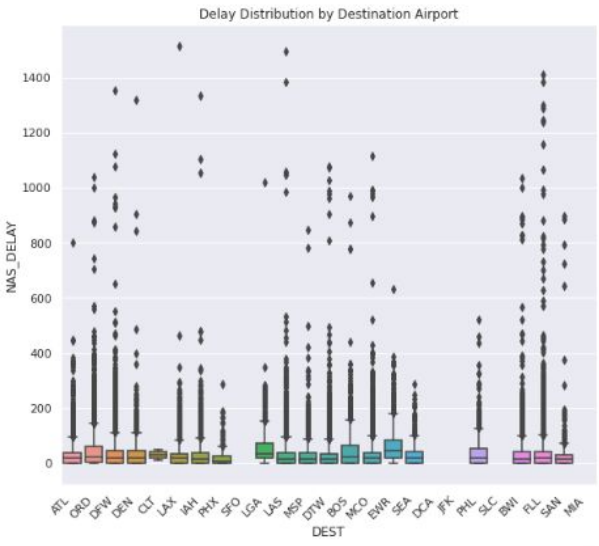


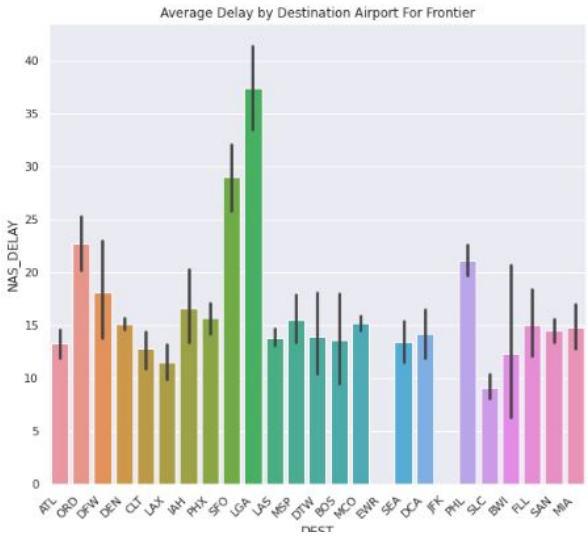
Delays by Airports

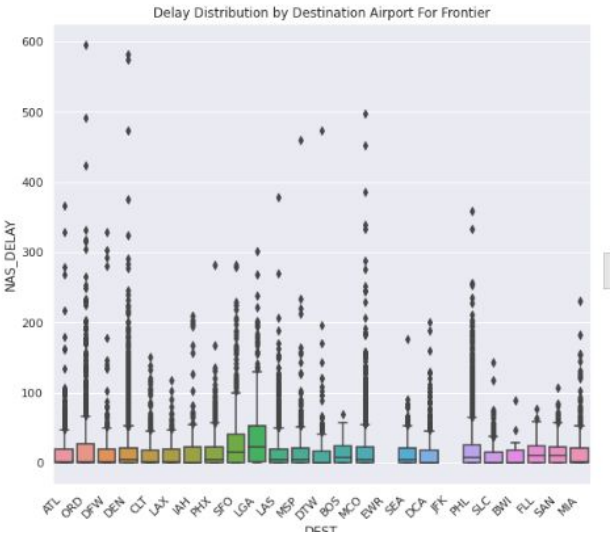


ULCCs travel to most of the busiest airports in the US. Avoiding the general New York area makes sense, however, we observe that customer demand leads to the ULCCs flying to the busiest airports anyway. With the exception of Southwest, most ULCCs do not try to save turnaround time by flying to secondary airports, contrary to the general theory of low cost carriers elsewhere around the world. We present the results of both Frontier and Spirit, which fly to all but 2 of the 25 busiest airports in the US.









Forecasting delays

An explainable predictor based on a Random Forest model

For the ease of analysis, we divided the data for initial analysis and then built a machine learning model on this divided data. We also took additional data from IATA for latitude and longitude, mapped airline codes to airline names .

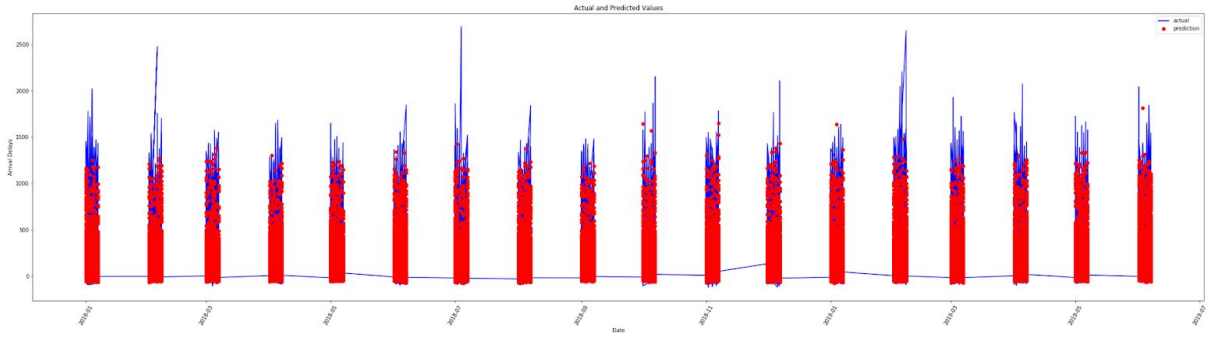
Random Forest to build our model, based on the non-linear relationship of independent variables. Random Forest (RF) is an ensemble classifier that consists of many decision trees, and outputs the mode of the classes output by individual trees. It combines the concept of bagging with the random selection of variables at each tree split. We chose Random Forest because of the flexibility to achieve variable importance, their low sensitivity to outliers in the training data, and their good performance in cases where the number of variables are large, similar to our dataset.

Our delay prediction models use a combination of categorical and continuous explanatory variables. Initially, we started with 43 features to train our model on the entire dataset. However, the computational time was slow and we built our models specific to each airline.

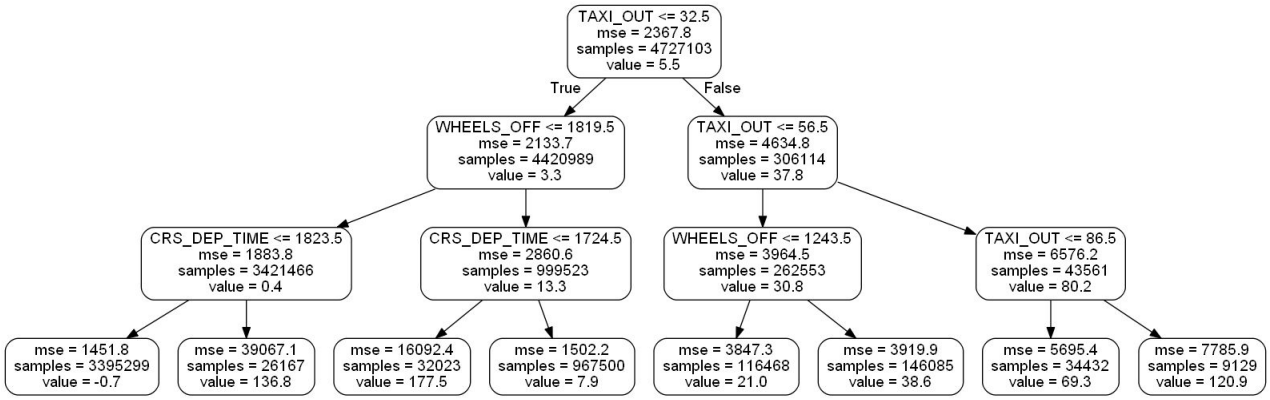
Model Performance

We ran our model on each airline and achieved an average training accuracy score from 90%-95% and average testing accuracy score of 87%-95%.

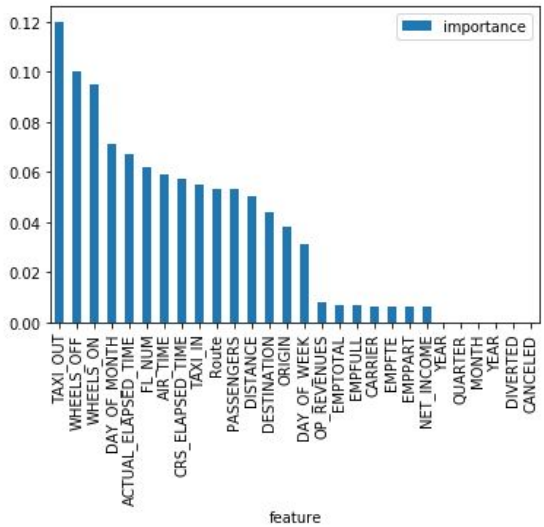
Figure : Actual vs Predicted flight delays till the 3rd quarter of 2019



Actual: Blue
Prediction:Red



Feature Importance



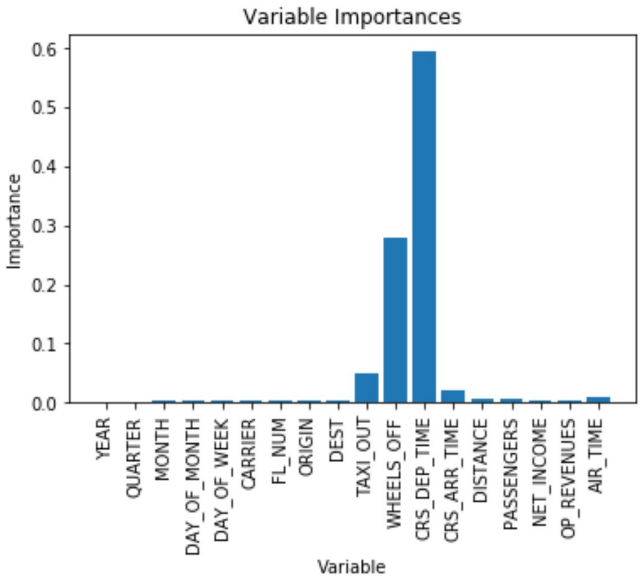
During the first phase of the project, we identified these attributes as the most important features. After the analysis of data by different correlation methods, we selected the final 18 sets of features for a single model.

Finally, we used 18 sets of features for our single model irrespective of airlines.

```
Variable: CRS_DEP_TIME         Importance: 0.5944
Variable: WHEELS_OFF           Importance: 0.28046
Variable: TAXI_OUT             Importance: 0.04951
Variable: CRS_ARR_TIME         Importance: 0.02185
Variable: AIR_TIME             Importance: 0.01083
Variable: DISTANCE             Importance: 0.00646
Variable: PASSENGERS           Importance: 0.00538
Variable: FL_NUM               Importance: 0.00481
Variable: ORIGIN               Importance: 0.00464
Variable: DAY_OF_MONTH         Importance: 0.00404
Variable: DEST                 Importance: 0.00325
Variable: OP_REVENUES          Importance: 0.00295
Variable: NET_INCOME           Importance: 0.00274
Variable: MONTH                Importance: 0.0026
Variable: DAY_OF_WEEK          Importance: 0.00256
Variable: CARRIER              Importance: 0.00236
Variable: QUARTER              Importance: 0.00068
Variable: YEAR                 Importance: 0.00048
```

Since we are predicting arrival delay, we tried to build a model using departure specific features such as Wheels off and taxi out. We omitted the Taxi in and Wheels on features from the model.

Ultimately, this is the final features that we are using



Recommendations

Modelling of future delays

We predicted delays based on the most important features related to an airline. Our model uses the features which are known to us before the arrival of the flight. So, our model can predict delays in advance since we can precompute the parameters beforehand rather than on-the-go and hence is fast and efficient. The results can be shown to customers in real-time.

Customer Insights

- Customers should consider the time of the day according to the type of airline they are choosing. For example, if travelling low-cost, it is far more important to travel earlier in the day.
- Be wary of routes which have monopolistic tendencies. Usually, all airlines - even ULCCs - fly to all the big airports - so it might be worth driving down to a larger city than taking a flight from a regional airport.
- Low cost airlines may often be a better choice earlier in the day over full fare airlines because they are streamlined for quick turnovers - so choose them in case travelling without much luggage.

Appendix

The code for the flight delay prediction can be found at GitHub in the below link:

Data Pirates GitHub: https://github.com/sukantoguha/Data-Pirates

References:

https://scikit-learn.org/

https://towardsdatascience.com/

https://heartbeat.fritz.ai/