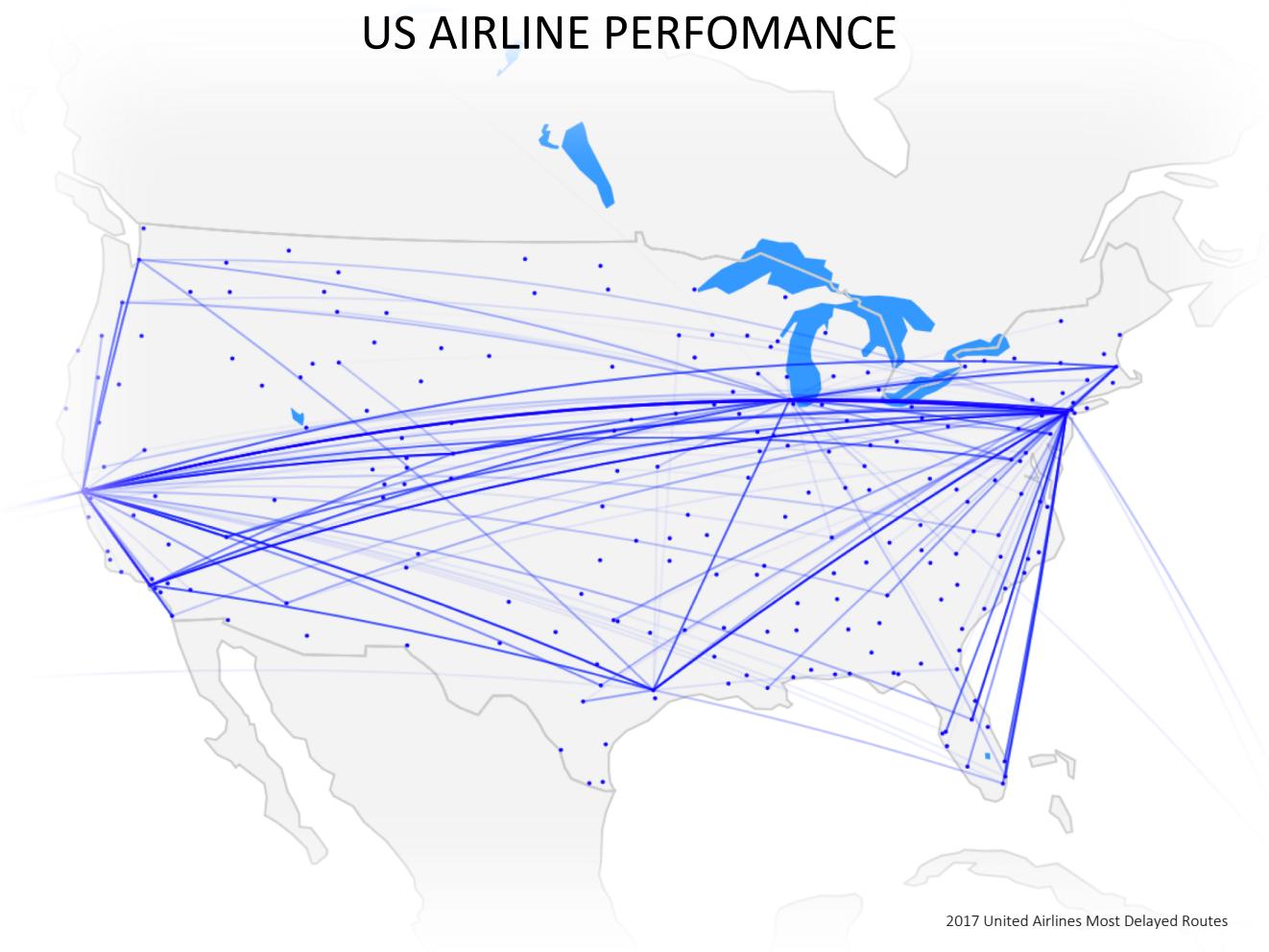


US AIRLINE PERFORMANCE



Dmitriy Kats
May 2019

Table of Contents

1.	<i>Introduction and Executive Summary</i>	3
2.	<i>Acquiring and Aggregating the Data</i>	3
2.1.	Adding Weather Data	3
2.2.	Adding Airport Coordinates	4
3.	<i>Cleaning the data</i>	5
4.	<i>Exploratory Data Analysis (EDA)</i>	6
4.1.	Distribution of Arrival Delays	6
4.2.	Airport and Airline On-Time Performance	8
4.3.	Regional Analysis.....	10
4.4.	Regional Traffic Volume and Delays	13
4.5.	Individual Route Analysis	15
5.	<i>EDA Conclusions</i>	17
6.	<i>Modeling</i>	18
7.	<i>Feature Engineering</i>	19
8.	<i>Conclusions and Next Steps</i>	20

1. Introduction and Executive Summary

Air travel has always been and still is a headache for many travelers. The unknowns of delays and cancellations are some of the biggest contributors to the stress. In this analysis we'll attempt to shine a light on the unknowns and try to predict the probability of delay and even the delay length of a given future flight. We'll take a look at 15 years of airline performance data, containing over 75 million flights in a dataset available from US Department of Transportation.

2. Acquiring and Aggregating the Data

The data was obtained from United States Department of Transportation. Unfortunately, the data are only available by month in a zipped file, which presents a challenge when it comes to downloading and concatenating. Additionally, the website utilizes a webapp to click on checkboxes and buttons to activate the download, which makes scraping impossible. As such, the data was acquired manually. Due to limitations in available data in years prior to 2004, only data for 2004 - 2018 was obtained for a total of 12 zipped files per year (for a total of 12 files x 15 years = 180 files). The data was then combined into a set of files organized by year and then into one complete csv file. The dataset will be uploaded to Google BigQuery for further analysis. For this analysis we will be using the 2017 Dataset. The code to combine all the files is available on github.

In summary, the following steps were taken:

- a. Create a list of files in each year directory via glob method.
- b. Loop through each file, unzip, and read into a Pandas Dataframe via pd.read_csv, utilizing the "compression='zip'" parameter.
- c. Since each year's file will be ~3GB in memory, convert the datatypes in each column to reduce size.
- d. Repeat the process for each year by looping through year folders.
- e. Concatenate all csv files into one via shell and upload to Google BigQuery.

2.1. Adding Weather Data

Additional data were obtained during the feature engineering stage in order to compute precipitation information for a specific airport and departure time. The precip_sum column was created in the flight data frame. This column includes the sum of four hours' worth of precipitation data prior to the departure hour.

	station	network	valid	precip_in		CRSElapsedTime	Distance	DistanceGroup	precip_sum
0	LAX	CA_ASOS	2016-12-31 22:00:00	0.0		120.0	574.0	3	0.0
1	LAX	CA_ASOS	2016-12-31 23:00:00	0.0		115.0	574.0	3	0.0
2	LAX	CA_ASOS	2017-01-01 00:00:00	0.0		250.0	1865.0	8	0.0
3	LAX	CA_ASOS	2017-01-01 01:00:00	0.0		150.0	999.0	4	0.0
4	LAX	CA_ASOS	2017-01-01 02:00:00	0.0		85.0	391.0	2	0.0

Figure 1 Mapping Precipitation Data to the Main Flight DataFrame

2.2. Adding Airport Coordinates

We're also going to need airport coordinates in order to create extra features for the modeling portion of the project. We'll first import a csv file with all US airport latitude and longitude data and join it with our flights data frame.

locationID	Latitude	Longitude	start_lat	start_lon	end_lat	end_lon	airline	airport1	airport2
YUM	32.6686	-114.5991	39.8617	-104.6731	29.9844	-95.3414	UA	DEN	IAH
MQT	46.3497	-87.3873	33.9425	-118.4072	29.9844	-95.3414	UA	LAX	IAH
SCE	40.8500	-77.8487	33.9425	-118.4072	38.9475	-77.4600	UA	LAX	IAD
ECP	30.3553	-85.7991	32.7336	-117.1897	39.8617	-104.6731	UA	SAN	DEN
ADK	51.8781	-176.6461	38.8522	-77.0378	29.9844	-95.3414	UA	DCA	IAH

Figure 2 Mapping Airport Coordinates to the Main DataFrame

Spot checking DEN & IAH (Denver International & Houston International):

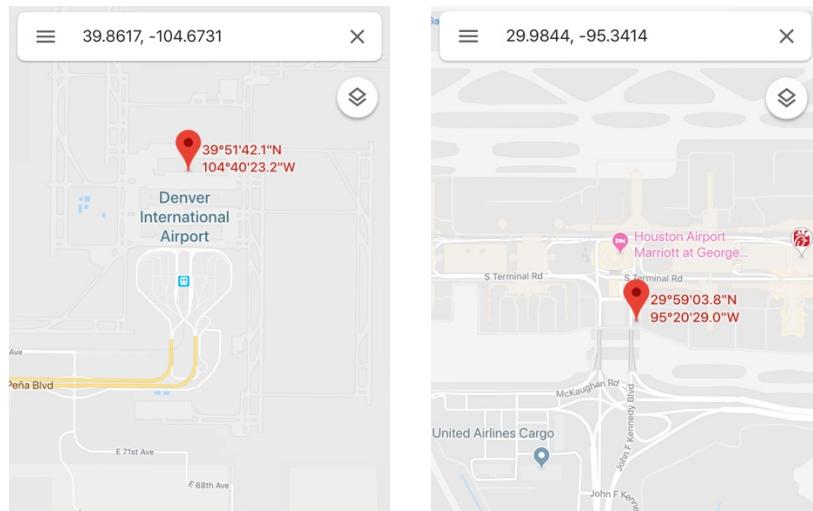


Figure 3 Denver and Houston Aiprots Check Against DataFrame Information

3. Cleaning the data

The dataset is relatively clean from the perspective of missing data. Only 1.5% of rows have missing values, which can be removed without significant impact on the analysis. As we go through EDA and ML portions of this analysis, the data will need to be formatted to fit the needs of the approach. As an example, time of day is presented in the dataset as a float-type number, ranging from 00 to 59 mins, which leaves 60 to 99 blank. This can be problematic when visualizing the data or converting to a timestamp. This will be addressed through a custom method which will convert the float-type number to a proper format.

```
df.DEP_TIME.head()  
0    1031.0  
1    1420.0  
2    1203.0  
3     758.0  
4    1041.0  
Name: DEP_TIME, dtype: float32
```

Table 1 Example of Departure Time representation in the dataset (e.g. 1031.0 should be 10.52 as float)

Furthermore, we are going to be dealing with a lot of categorical data, such as airline names and airport codes, which will have an impact on predicting performance. In order apply modeling techniques these data will need to be converted into numerical values, as the solvers rely on mathematical algorithms. When encoding these values, we'll have to be mindful of memory space required as there are over 300 unique origins and destinations and over 10 unique airlines. A possible solution to this is the use of a sparse matrix.

4. Exploratory Data Analysis (EDA)

In this section we'll explore the data statistically and visually to understand any trends or features we should focus on during the modeling stage. The code for this section is available on github with a sample data set on Google drive:

Code:

<https://github.com/dmitriykats1/Springboard/blob/master/Capstone1/EDA-2017.ipynb>

Data:

<https://drive.google.com/file/d/15PXxxTY9X4w0exxu6vMEReJm3Pwcn4TE/view?usp=sharing>

4.1. Distribution of Arrival Delays

Let's begin with the most important feature, delay time at the destination, represented as ArrDelay in our data set. This feature set has both negative and positive values, representing early and late arrival times, respectively. Let's take a look at the distribution:

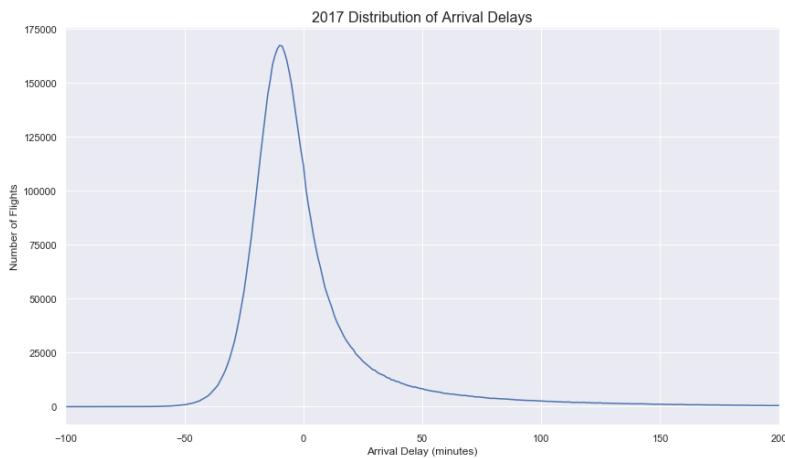


Figure 4 Distribution of delays for all US flights in 2017

From the above we see that majority of flights being on time or early, and the distribution is also skewed and non-normal. With the mean arrival delay of 4.3mins., median of -6.0mins, and a standard deviation of 45.5mins. The next obvious questions we can ask is: Do long and short flights have same distributions? To answer this, we'll split the dataset into long flights with durations longer than 3hrs and remainder will be short flights.

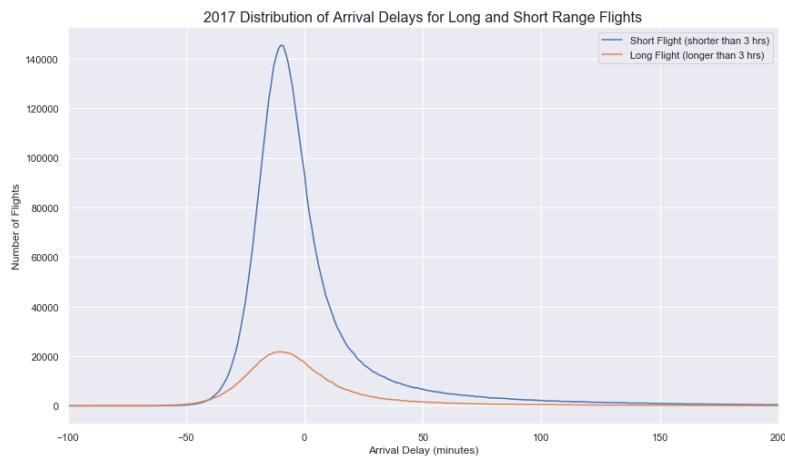


Figure 5 Distribution of delays for all US flights in 2017 for short and long range flights

The short- and long-range flights seem to have the same distributions with slightly different means which we can investigate use to investigate any statistical differences, with long range flights having a mean delay of 3.8mins and short-range flights having a mean delay of 4.4mins. Since majority of flights are on time or early, taking an average of arrival delay does not paint a full picture due to the skew. This can be better visualized below:

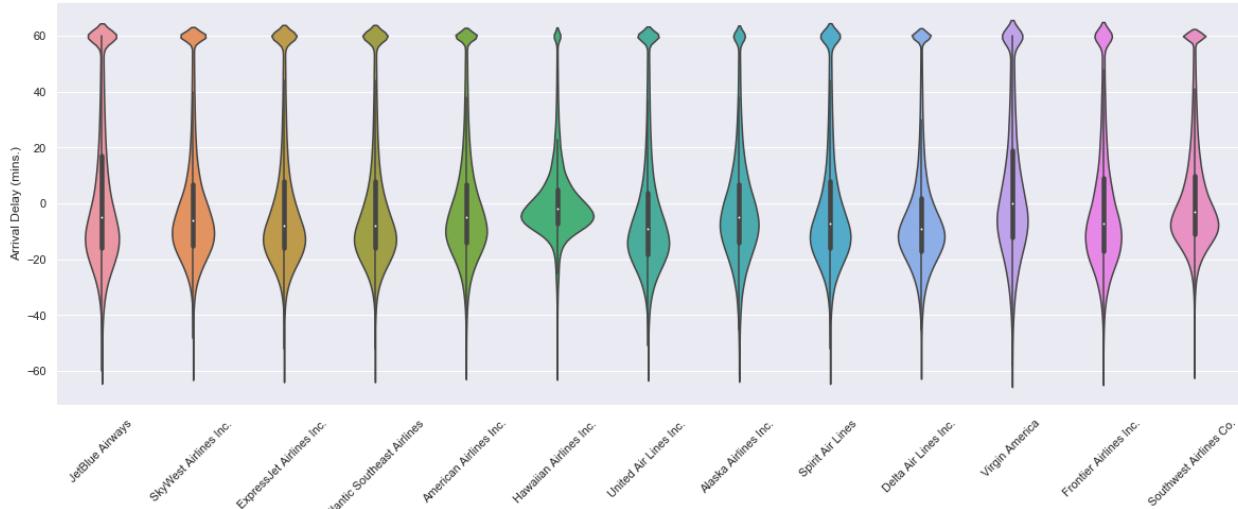


Figure 6 Arrival delay distribution split by airlines

Next, we can take a look at ONLY delayed flights, or flights that are delayed 15 mins or more as defined by the FAA. One of the questions we can ask is if there is any correlation between arrival delays and other features. First let's look at departure delays and arrival delays:

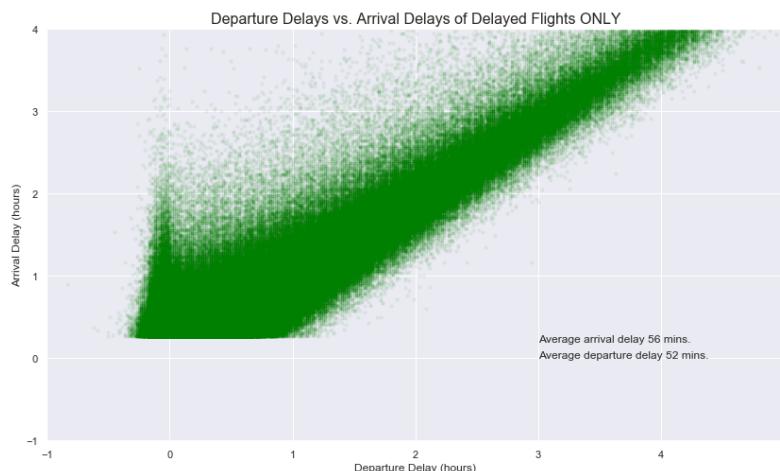


Figure 7 Correlation of departure and arrival times for all delayed flights in 2017

There is a positive linear relationship between departure delays and arrival delays, this makes sense intuitively and we'll need to keep this in mind when modeling our data. We do notice a small spike in arrival delays around the 0-minute mark for departure delays. This may be due to airlines trying to get out on time and closing the aircraft door in order to have an on-time departure. Only finding themselves waiting in taxi lines to depart and subsequently be delayed on arrival.

4.2. Airport and Airline On-Time Performance

Another set of features that may impact on-time performance are the airlines and airports. Let's take a look at the summary of top airlines and on-time performance:

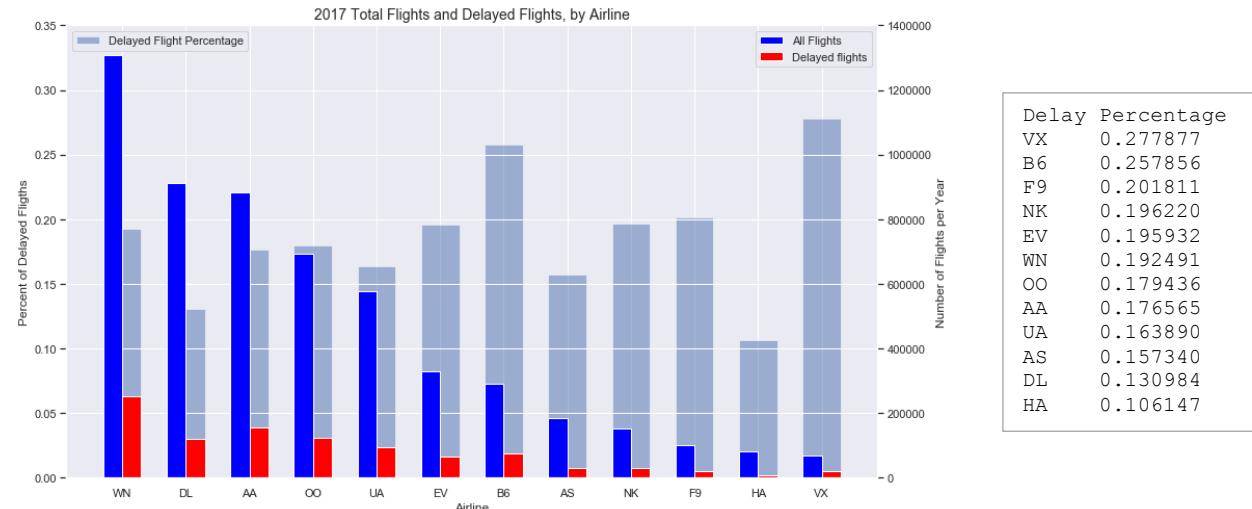


Figure 8 Summary of Airline Performance for 2017

Above figure gives us a glimpse of how each airline performed in 2017, the data above is sorted by flight volume. It's interesting that airlines with most flights don't necessarily have the worst delay record, as we can see from the delay percentage bars. Delay percentages will certainly change as we break this data down by origin airport. Let's see if there is a list of airports that have consistently high delays when broken down by airline.

Initial analysis looked at Southwest, American, and United Airlines (which make up 50% of all US flights) and delay percentages as a function of origin airport.

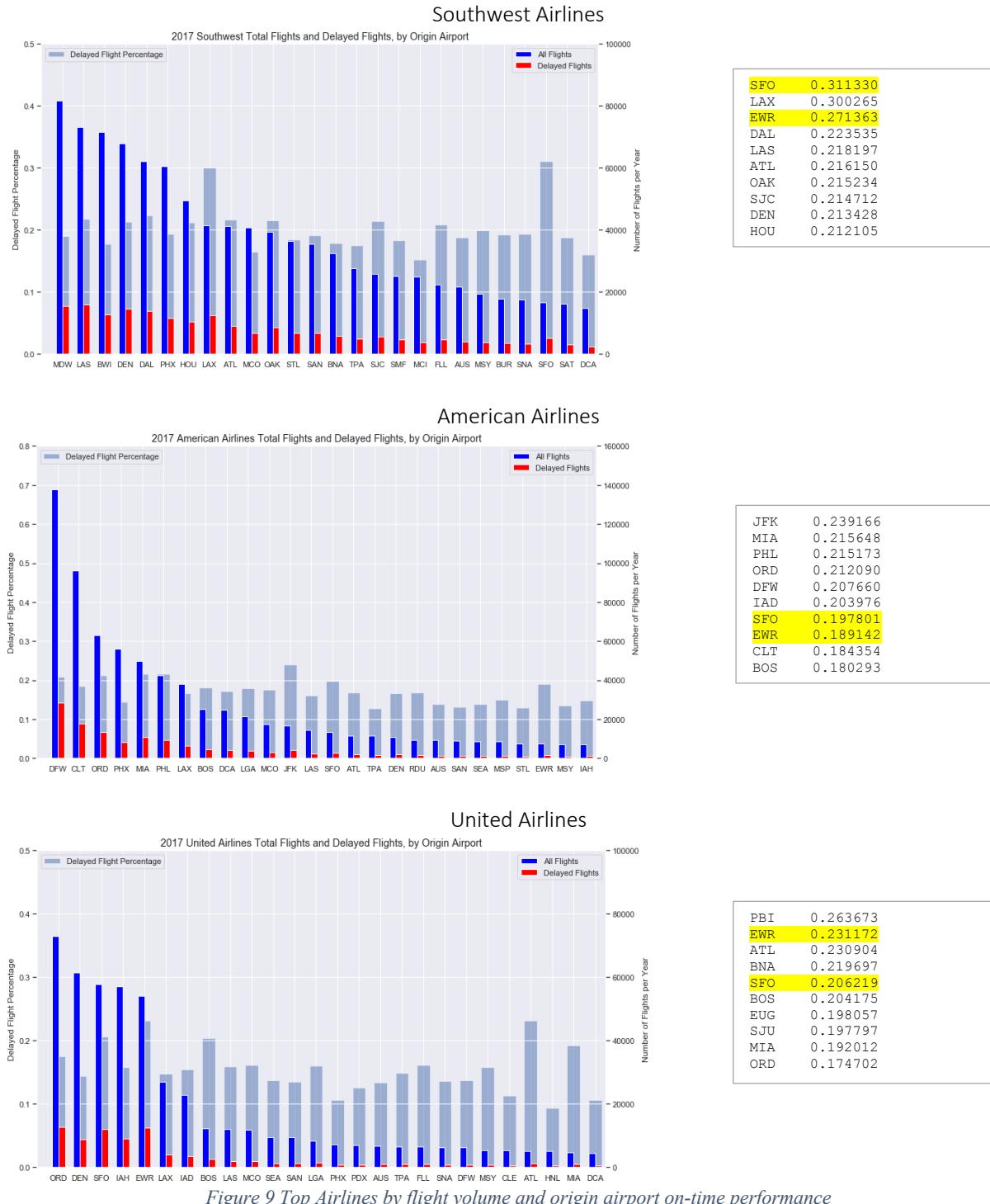


Figure 9 Top Airlines by flight volume and origin airport on-time performance

We see EWR and SFO appear on the list for all three airlines. We can further investigate these airports and surrounding airports to see if regional air congestion impacts other airports in the area. But first let's confirm these airports appear on the top delayed airports list.

Analyzing flight volume / delay volume percentages can give us a better picture of how the delays are weighted.

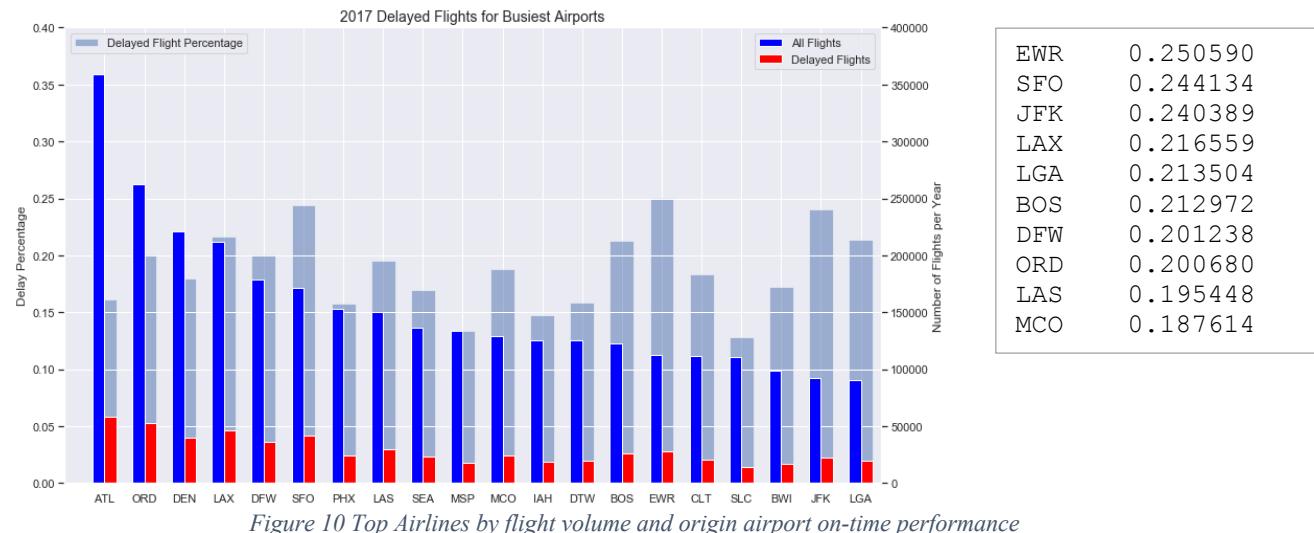


Figure 10 Top Airlines by flight volume and origin airport on-time performance

We can see, that EWR and SFO are top offending airports in 2017. But this list is sorted by flight volume, so we don't get the full picture. Surrounding airports may be impacted. We can also note that all three major airports in the NYC area appear on the list: EWR, JFK, and LGA. Let's look at EWR and airports within 30-mile radius and compare the performance to the national average.

4.3. Regional Analysis

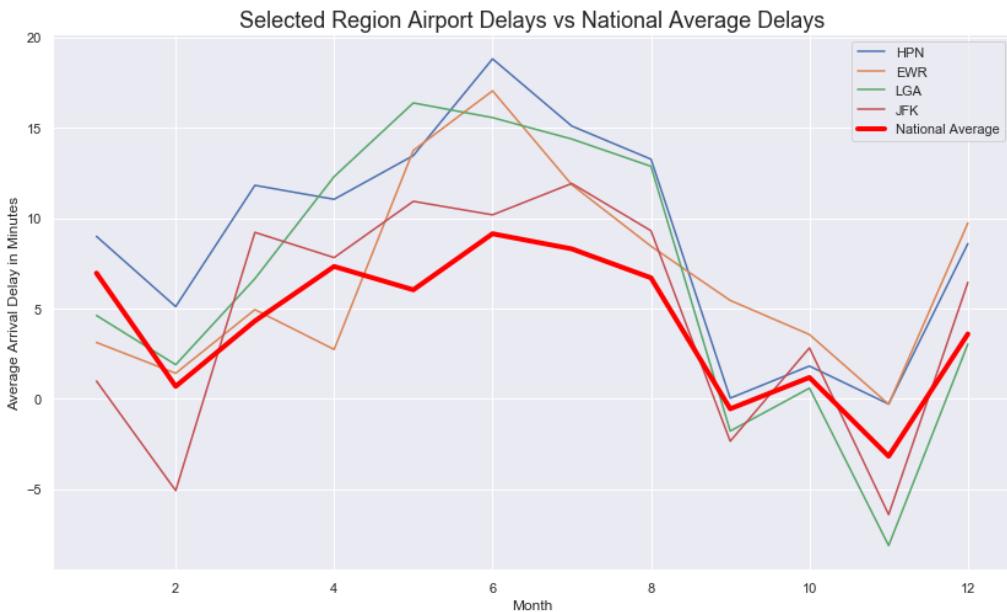


Figure 11 NYC Area average delays compared to a national average, by month

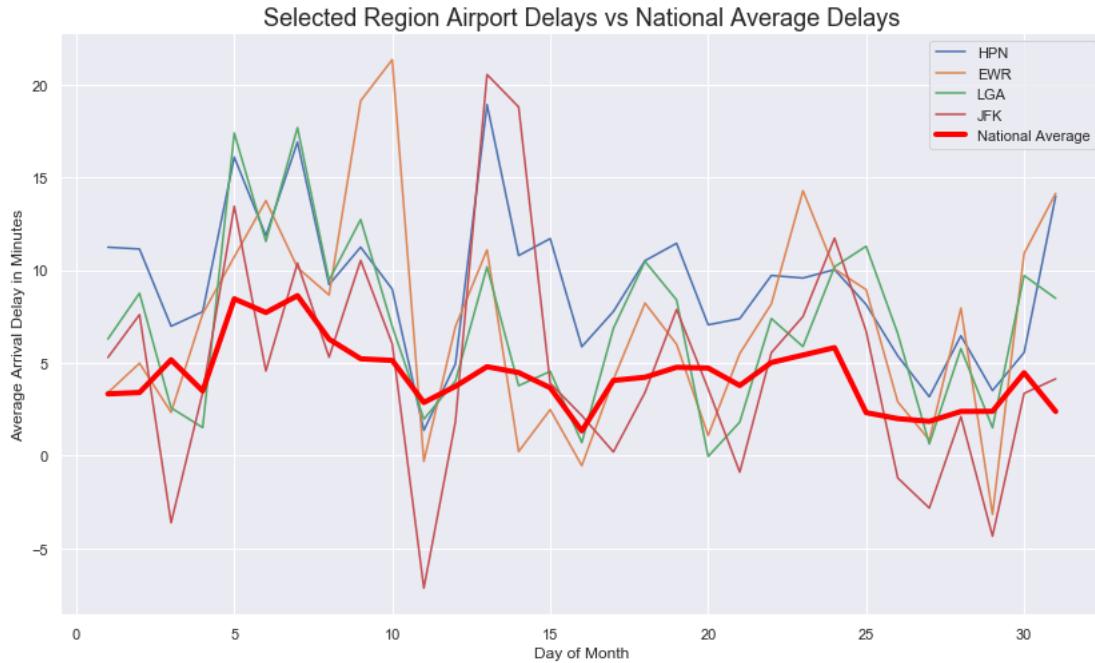


Figure 12 NYC Area average delays compared to a national average, by day of month

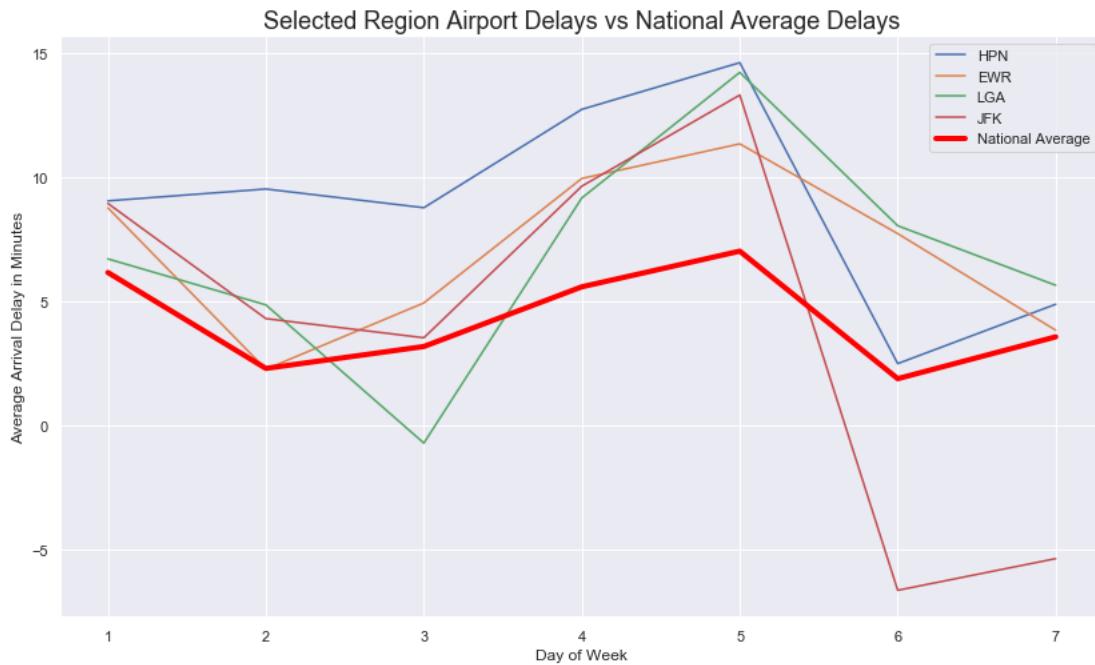


Figure 13 NYC Area average delays compared to a national average, by day of week

We can see that all airports in the area tracks along with the national average on all charts, which was expected. However, regional airports tend to perform worse, on average. Note, that this analysis shows delays by minutes whereas last set of charts focused on percent of flights that were delayed. Next let's take a look at how the delays break down, what is causing these delays?

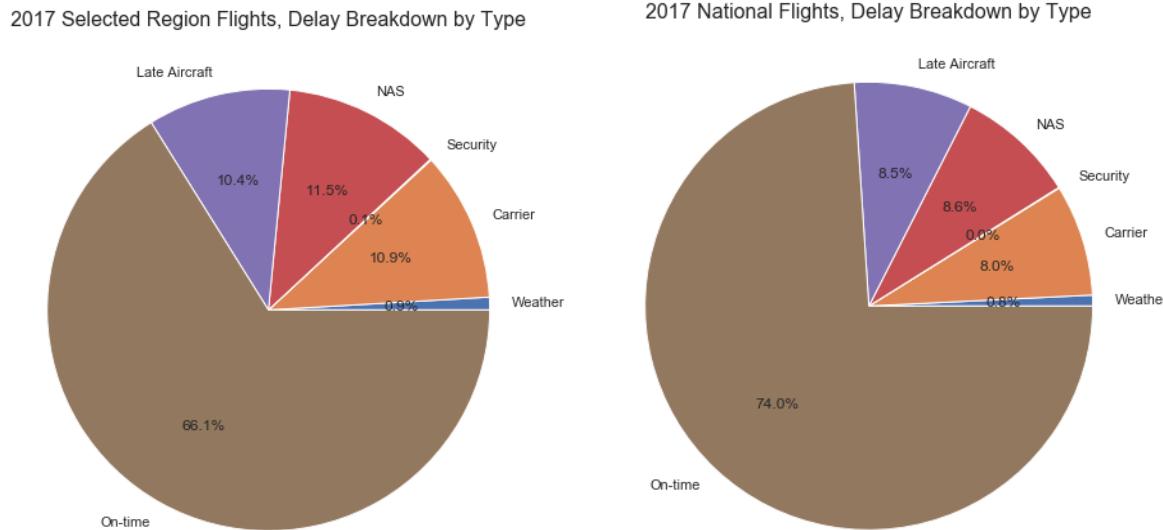
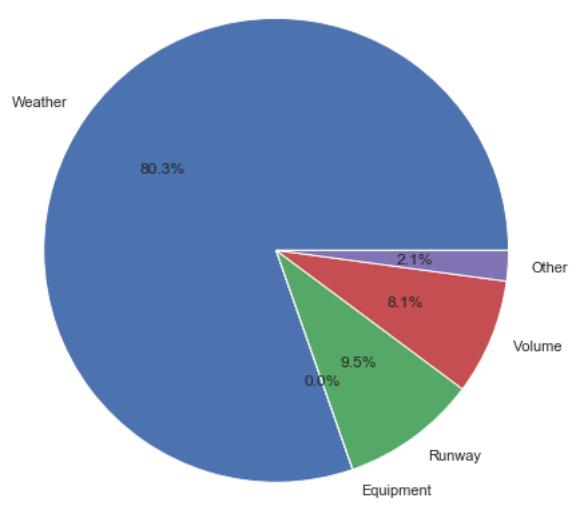


Figure 14 NYC Area vs National delay breakdown by cause

Since NAS data is further broken down into sub-categories, we can download the actual data from FAA. Breaking down the FAA data (NAS delays section from above):

2017 Selected Region Flights, NAS Delay Breakdown by Type



NAS - We can continue to look at patterns for heavy traffic or trends in increasing air traffic patterns areas and look for consistencies. Additionally, extreme events are not considered here, normal weather patterns that cause air traffic slowdowns are. Looking at weather patterns may be beneficial. FAA has a database breaking down the NAS delays by cause. (FAA OPSNET)

Carrier - carrier performance can be further analyzed by location and see if there are patterns

Late Aircraft - This can be handled with arrival delay information. This field highly depends on the other delay causes.

Weather - only extreme weather events are considered here. These events are rare and would result in region-wide cancellations, as such, these can be ignored as outliers.

Figure 15 NYC Area delay breakdown of NAS delay data

NYC Area had an on-time performance of only 66% in 2017 (as measured by arrival delay), which is lower than the national average of 74%. We can note that late aircraft, carrier, and NAS delays made up an overwhelming majority of delays. We'll keep this in mind for further EDA and feature engineering stages of this project.

4.4. Regional Traffic Volume and Delays

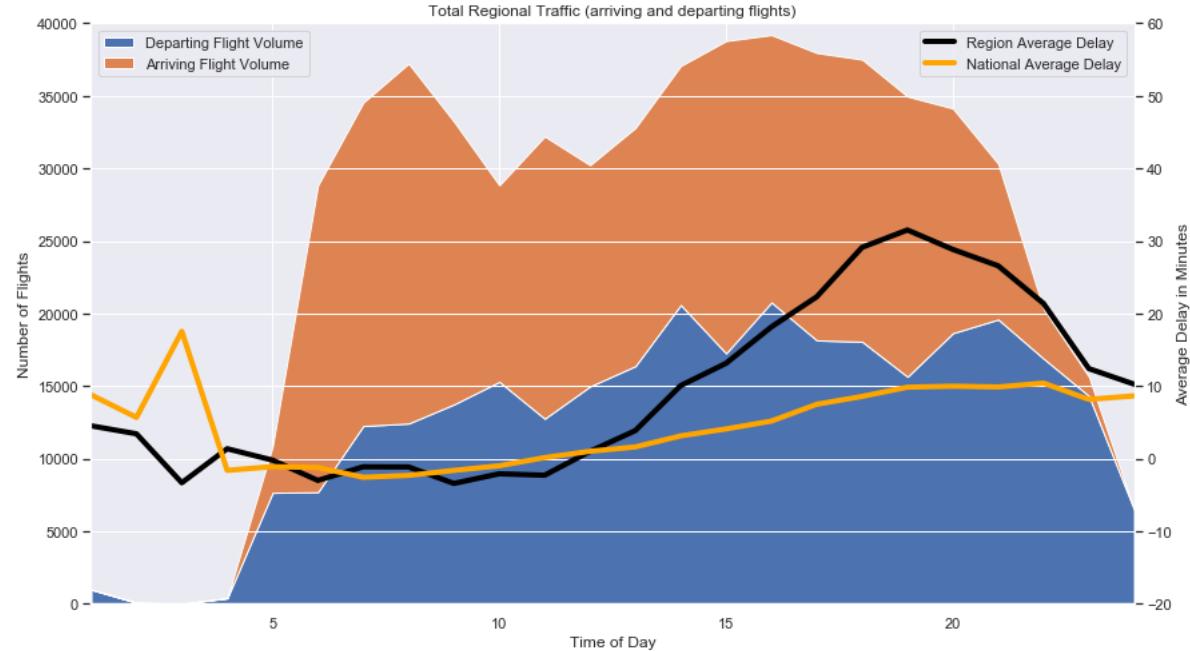


Figure 16 Air traffic volume and average delay time, throughout the day

Taking into account arriving and departing flights in the NYC Area, we can see that air traffic is relatively flat throughout the day, however the average delay in minutes tends to increase towards the afternoon and evening hours. Although, the same can be seen on a national scale, the delays peak at ~30mins for NYC area.

We can further breakdown the delay times by day and hour to get a full picture of performance of all airlines in the NYC Area:

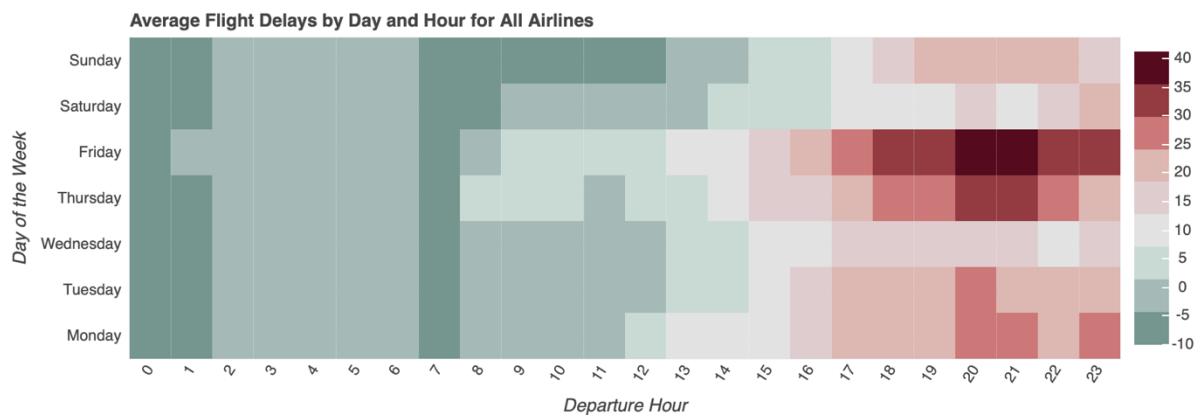


Figure 17 Heatmap of arrival delays broken out by hour and day of week

It looks like NYC Area sees an increased delays during Thursday and Friday afternoons/evenings. This could be due to increase in passenger loading during those hours and days or more flights. We can look into this next.

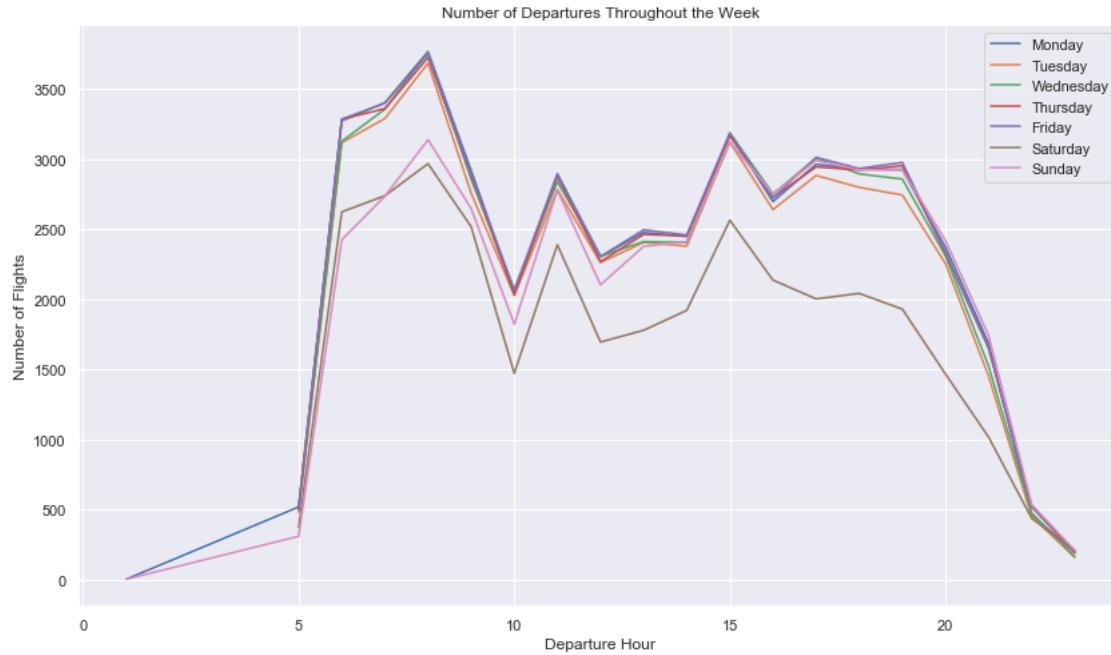


Figure 18 Number of Departing Flights in the NYC Area Throughout the Week

There does not seem to be any spikes during evening / night hours for Thursday and Friday. Scheduling seems relatively consistent.

Let's take a look at the passenger loading data by month and see if there is a trend relative to delays:

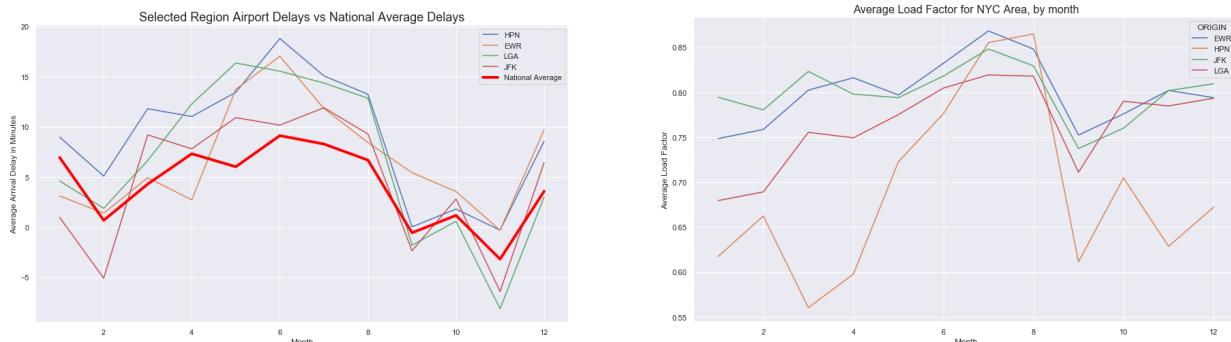


Figure 19 Average Delay (left) and Passenger Load Factor (right), by month

There seems to be a trend between delay and passenger loading, the more passengers fly, the higher the average delay. Unfortunately, only monthly aggregated Load Factor data are available, and we cannot compare loading data to hourly performance.

4.5. Individual Route Analysis

Now we can take a look at specific routes and their historic performance throughout the year. Are there routes that are prone to delays and are there ones that are always on-time? We'll first filter and create a data frame that we can use to answer our questions:

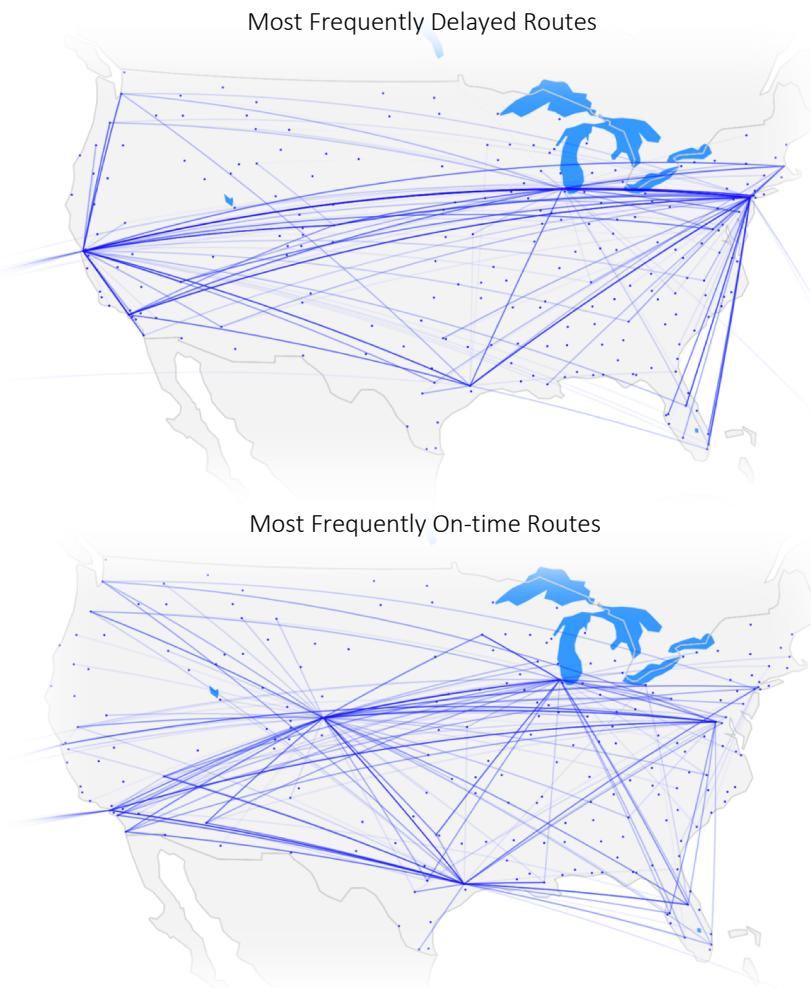
First we'll group the data by airline, origin, and destination airport, to see what are the most flown routes for the specific airline and then add the average delay time for each route, we can do this for on-time flights, as well:

airline	airport1	airport2	cnt	avg_del
UA	EWR	SFO	4885	10.128557
UA	ORD	SFO	4397	14.556516
UA	DEN	SFO	3912	14.054448
UA	ORD	EWR	3151	15.589971
UA	EWR	BOS	3069	14.160964

airline	airport1	airport2	cnt	avg_del
UA	DEN	IAH	3625	-3.664552
UA	LAX	IAH	3556	-3.771372
UA	LAX	IAD	2457	-6.881563
UA	SAN	DEN	1557	-8.271676
UA	DCA	IAH	1478	-8.714479

Figure 20 Most Delayed(left) and On-time(right) Routes for United Airlines

Now we can merge the data with latitude and longitude information as described in section 2.2 and visualize what this looks like on a map.



United Airlines Most Delayed(top) and Most On-time(bottom) flight routes in 2017.

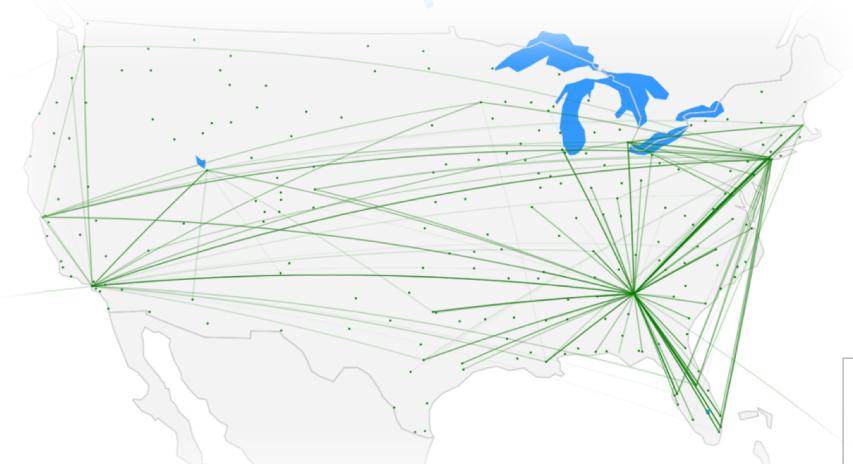
The averages are weighted based on number of flights for the particular route. Darker lines represent higher delay frequency.

There is a clear difference in performance based on the route chosen. Coast to coast or long-haul flights tend to represent routes that are most delayed. This will be considered during the modeling stage. As simple average delay will not isolate problematic routes. This can be seen in Figure 3.1 and the distribution of arrival delays based on length of flight.

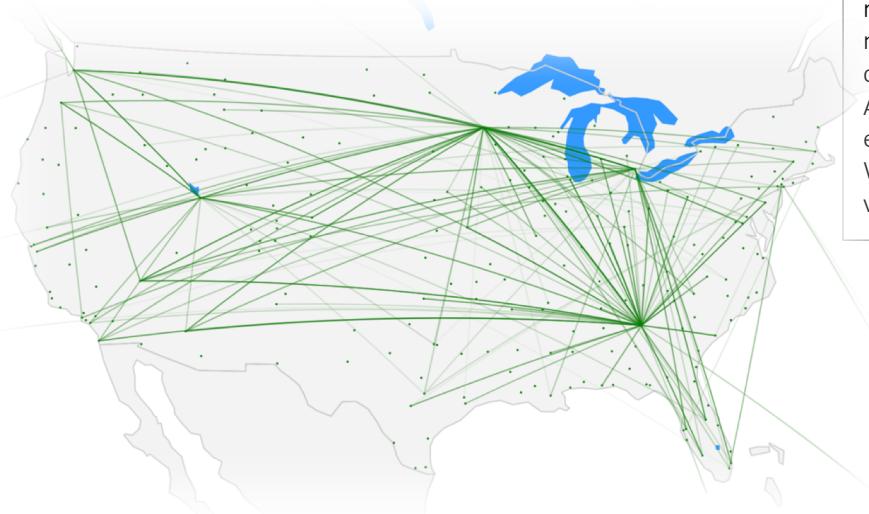
These routes were somewhat expected as the airports that appear in the delayed routes list are problematic for all airlines as can be seen in Figure 10.

Delta Airlines

Most Frequently Delayed Routes



Most Frequently On-time Routes



Delta Airlines Most Delayed (top) and Most On-time (bottom) flight routes in 2017

We see the same pattern for Delta Airlines, long-haul flights seem to dominate most delayed routes. And although Atlanta is not one of the most frequently delayed airports, routes that connect to JFK and LAX, for example, are. Additionally, Delta has three hubs with exceptional performance: MSP, DTW, and SLC. Which do not appear on the delayed route visualization.

5. EDA Conclusions

The intent of the EDA was to find any patterns in the data by creating visualizations which can later be used to create additional features and consequently a more robust model. We found a lot of trends and insights during this EDA but when we find these insights it's important to stop and ask: why are we seeing these trends? What is the underlying cause? We can't just apply this to future predicting without knowing the cause.

Possible questions to answer from EDA:

From route analysis:

1. Is there a link between origin airports / destination airports and these delayed routes?
2. Do other airlines have similar performance for a given route?
3. Does a given dominate a given airport?

From flight delay percentage comparison between Bay Area and National:

1. Why does NYC area have more delays than the national average?
2. Is there a statistically significant difference between number of delays in the NYC vs National average?
3. What region has lowest percentage of flight delays?

From the delayed percentage chart by airport:

1. Why do EWR, SFO, and JFK have the highest percentage of delayed flights?
2. Do these airports have most coast to coast flights? If so, why would these flights be delayed?
3. International traffic was not considered for this analysis which could have a huge impact on airport loading, especially major airport hubs such as EWR, SFO, and JFK.

Datasets

Loading Data

https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=311

Complete Datasets

https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236

Weather:

<https://mesonet.agron.iastate.edu/ASOS/>

Airline Info (automatic download)

https://www.transtats.bts.gov/Download_Lookup.asp?Lookup=L_UNIQUE_CARRIERS

Airport Coordinates:

<https://drive.google.com/file/d/1bMVXqd8Tm30RwmYLBgKk8RMr786luDoK/view?usp=sharing>

FAA OPSNET:

<https://aspm.faa.gov/opsnet/sys/Delays.asp>

6. Modeling

We'll initially approach this problem using RandomForestClassifier with an 80/20 test/train split to predict if flight is delayed or not, along with the probability. Random Forest was chosen for its relatively fast training algorithm and accuracy. It is also relatively simple to tune and use, and since there are a large number of trees, overfitting tends to be less of an issue.

We start our run with minimal data, only including the features that will be available at the time of purchasing the ticket: date, carrier, origin, destination, and time. Original run is on the entire 2017 dataset, with no feature engineering, only binning departure and arrival hour blocks:

MONTH	DAY_OF_MONTH	DAY_OF_WEEK	OP_CARRIER	ORIGIN	DEST	STATUS	ARR_HOUR	DEP_HOUR
0	3	21	2	2	39	81	0	11
1	3	21	2	2	145	39	0	16
2	3	21	2	2	39	146	0	13
3	3	21	2	2	39	236	0	9
4	3	21	2	2	110	26	0	13

Table 2 DataFrame Used for Initial Model

Using n_estimators=25, Initial results were favorable, as the AUC is significantly higher than chance, at 0.737 with an accuracy of 82.5%. This is not bad for first run but leaves room for improvement.

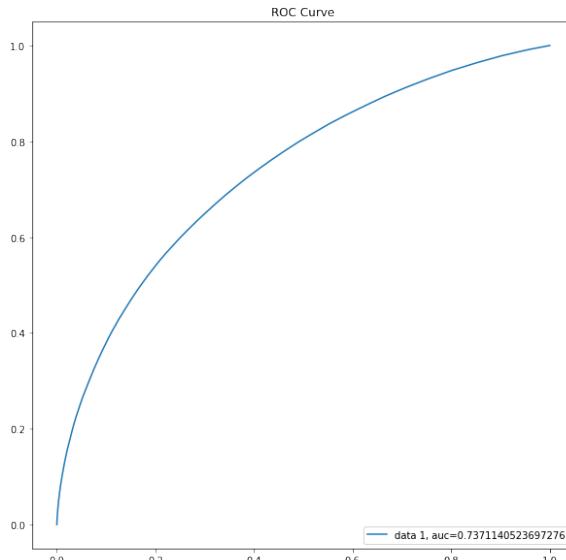


Figure 21 ROC Curve for Initial Run

		Actual		
		Positive	Negative	
Predicted	Positive	874,265	36,085	910,350
	Negative	158,869	46,663	205,532
		1,033,134	82,748	

Table 3 Confusion Matrix for Initial Run

Interpreting the confusion matrix, we can see that our predicted delays were much lower than actual, 46k vs 205k, respectively. This may not be a bad thing as this is a more conservative approach to predicting. However, we also didn't predict all on-time flights accurately, only 874k out of 1.02M were predicted accurately to be on-time.

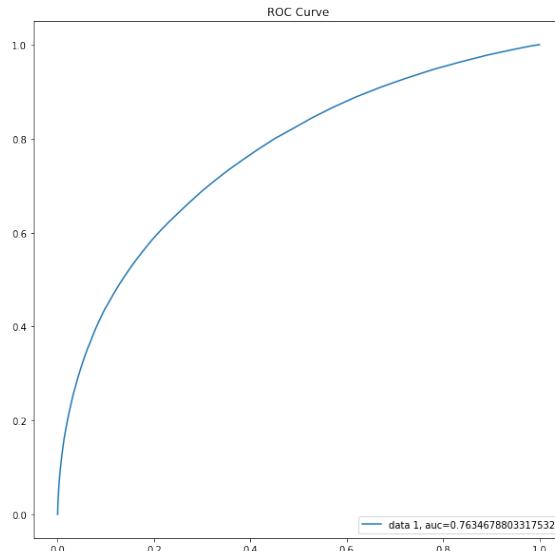
7. Feature Engineering

In order to improve our model, we'll add extra features to our dataset.

Flight number / plane tail number were added back into the dataset. Note that this is not an engineered feature but provides historical significance to the model if certain flight numbers are historically delayed.

We'll also add direction of flight, as flights heading West are typically susceptible to the jet stream and may be subject to more delays. Then we limited the dataset to only include five major airlines and 20 major airports, by volume. And the date features were OneHotEncoded in order to reduce dimensional importance.

Using n_estimators=100, results improved to AUC of 0.763 and Accuracy of 85.8%.



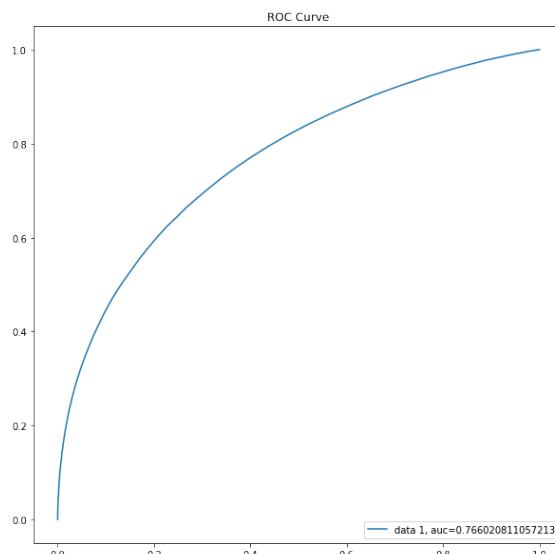
		Actual		
		Positive	Negative	
Predicted	Positive	409,244	5,010	414,254
	Negative	64,382	11,527	75,909
		473,626	16,537	

Again, we see that we under-predicted delayed flights by approximately 65 thousand flights.
Note, the totals are different due to reduction in modeled data.

Figure 22 ROC Curve and Confusion Matrix - Run 2

Next, we added precipitation data as a cumulative sum for the past 4 hours prior to scheduled departure. This was done to improve the model due to weather delays. This was not expected to be a large contributor, as weather delays are relatively rare. We also added a congestion component by adding a feature containing total number of scheduled flights for any given hour at a particular airport.

Using n_estimators=100, results improved to AUC of 0.766 and Accuracy of 86.1%.

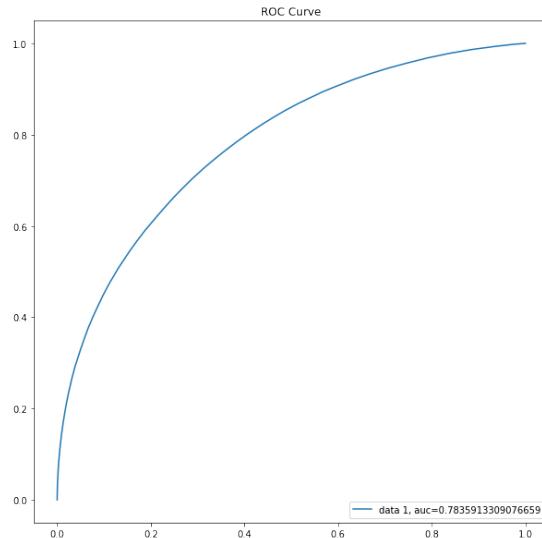


		Actual		
		Positive	Negative	
Predicted	Positive	408,622	5,632	414,254
	Negative	62,604	13,305	75,909
		471,226	18,937	

Again, we see that we under-predicted delayed flights but by a smaller amount. Approximately 62 thousand flights were misclassified as delayed.

Figure 23 ROC Curve and Confusion Matrix - Run 3

Next step was to add historic delay data based on routes and aggregate it on a monthly basis. We know from our EDA that certain airports are prone to delays and we can take advantage of this by creating a feature that takes the combination of origin and destination into consideration. We also added feature scaling and increased the n_estimators to 150. We saw increases in accuracy to 86.1% and AUC to 0.784.



		Actual		
		Positive	Negative	
Predicted	Positive	406,653	7,601	414,254
	Negative	60,396	15,513	75,909
		467,049	23,114	

Again, we see that we over-predicted delayed flights but by a smaller amount. Approximately 60 thousand flights were misclassified as delayed.

Interpreting the results, we can calculate some key metrics:

Recall / Sensitivity: When the flight is actually on-time, how often does our model predict that it is on-time?

$$\text{Recall} = 406,653 / 414,254 = 0.982$$

Precision: When the flight is predicted to be on-time, how often is the model predicting it to be on-time correctly?

$$\text{Precision} = 406,653 / 467,049 = 0.871$$

False Positive Rate: When the flight is actually delayed, how often does it predict that it's not?

$$\text{False Positive Rate} = 7,601 / 75,909 = 0.100$$

True Negative Rate: When the flight is actually delayed, how often is the model correctly predicting that the flight is delayed?

$$\text{True Negative Rate} = 15,513 / 75,909 = 0.204$$

Our model does not predict delayed flights very accurately and due to the fact that the data is imbalanced, our overall performance doesn't look too bad. However, our ultimate goal is to predict delayed flights more accurately.

8. Conclusions and Next Steps

There are three major contributors to delayed flights, as we found in the EDA, Figure 14: NAS (mostly weather), Carrier, and Late Aircraft.

We found a lot of correlations in our data with delayed flights, including certain airports and flight routes, load factor, and weather. However, due to limitations in the load factor data, we were not able to capitalize on the finding. We were only able to obtain monthly averages and our analysis boiled down to hourly variations in flight delays. If we can add load factor data on an hourly basis, the intuition is that the predicted delayed flights will increase. Additionally, weather data only included precipitation information and not wind or other events. This can be further improved with better datasets. The other key contribution to delays is Carrier performance, which can be explored via airline data on maintenance records and even age of aircraft based on the tail number and FAA data. And finally, Late Aircraft delays can be traced using the tail number data and origin / destination information to better predict overall delays. Due to the time limitations and the scope of this project, some of these options were not explored at this time.