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.

1. **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. The code to combine all the files is available on github: <https://github.com/dmitriykats1/Springboard/blob/master/Capstone1/create_csv.ipynb>

In summary, the following steps were taken:

1. Create a list of files in each year directory via gob method.
2. Loop through each file, unzip, and read into a Pandas Dataframe via pd.read\_csv, utilizing the “compression=’zip’” parameter.
3. Since each year’s file will be ~3GB in memory, convert the datatypes in each column to reduce size.
4. Repeat the process for each year by looping through year folders.
5. Concatenate all csv files into one via shell and upload to Google BigQuery.
6. **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.

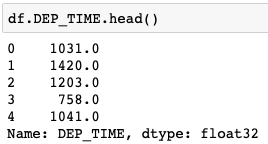


Figure 2.0 – Example of Departure Time representation in the dataset (1031.0 should be 10.52 as a float)

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.

The modeling notebook can be found here: <https://github.com/dmitriykats1/Springboard/blob/master/Capstone1/Modeling.ipynb>

1. **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: <https://github.com/dmitriykats1/Springboard/blob/master/Capstone1/EDA.ipynb>

Let’s begin with the most important feature, delay time at the destination, represented as ARR\_DELAY 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 of this feature:

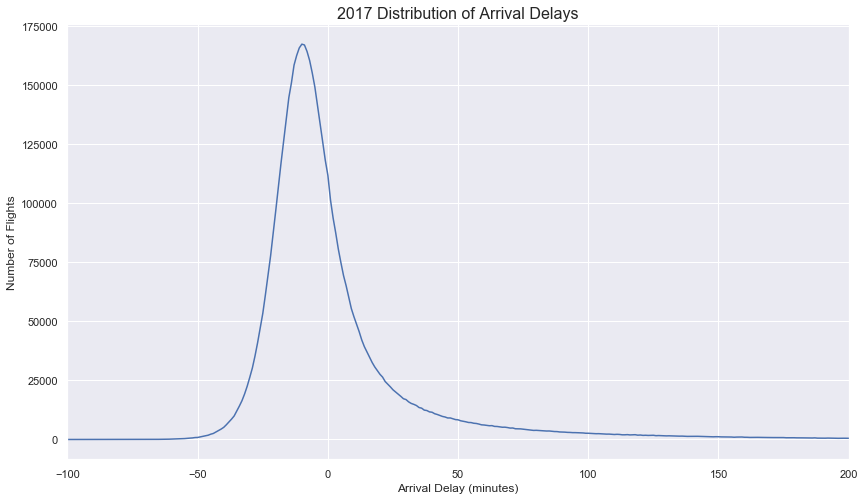


Figure 3.0 – Distribution of delays for all US flights in 2007

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.

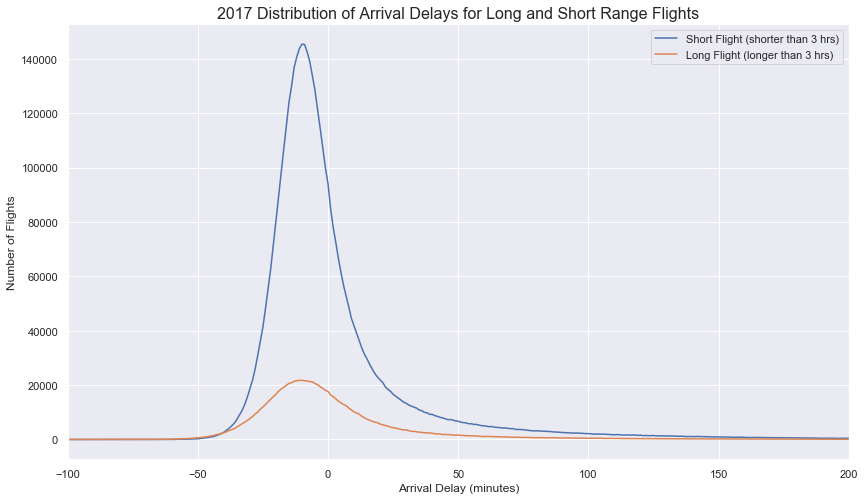


Figure 3.1 – Distribution of delays for all US flights in 2007 for short and long rage flights

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:

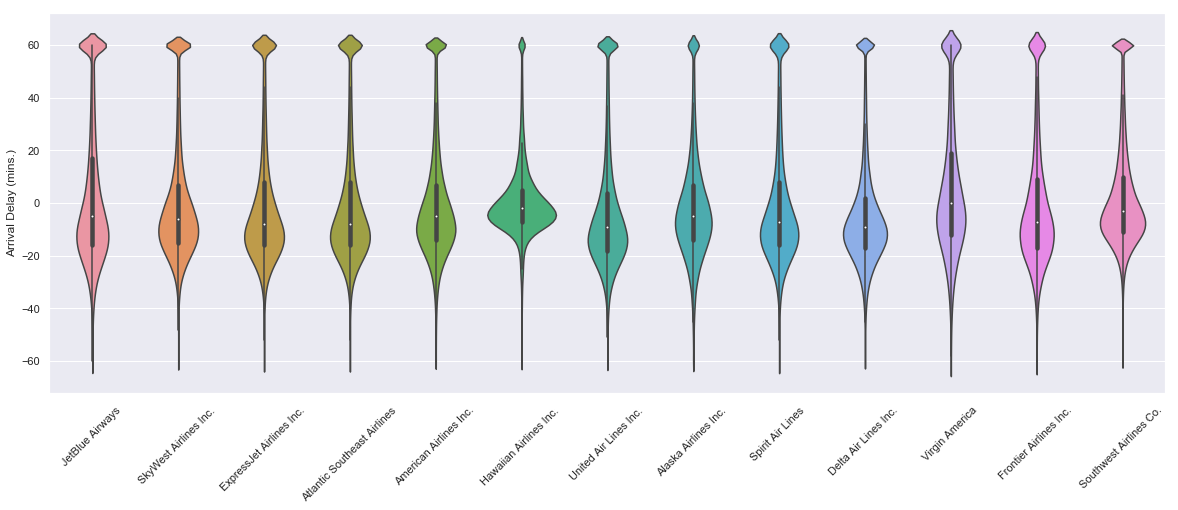


Figure 3.2 – Arrival delays split by airlines

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:

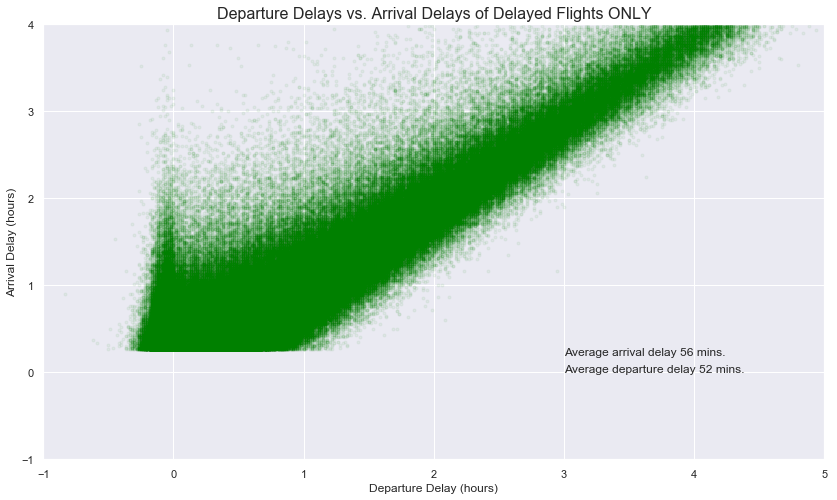
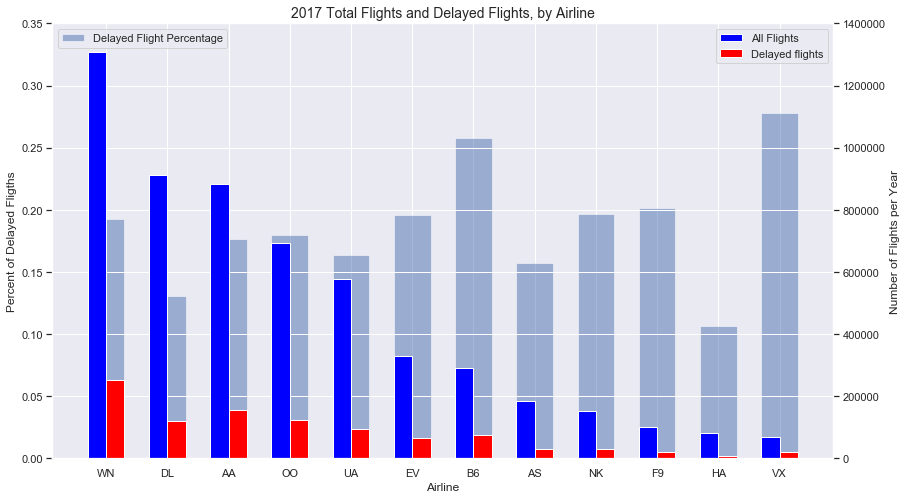


Figure 3.3 – Correlation of departure and arrival times for all delayed flights in 2017

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.

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:



VX 0.277877

B6 0.257856

F9 0.201811

NK 0.196220

EV 0.195932

WN 0.192491

OO 0.179436

AA 0.176565

UA 0.163890

AS 0.157340

DL 0.130984

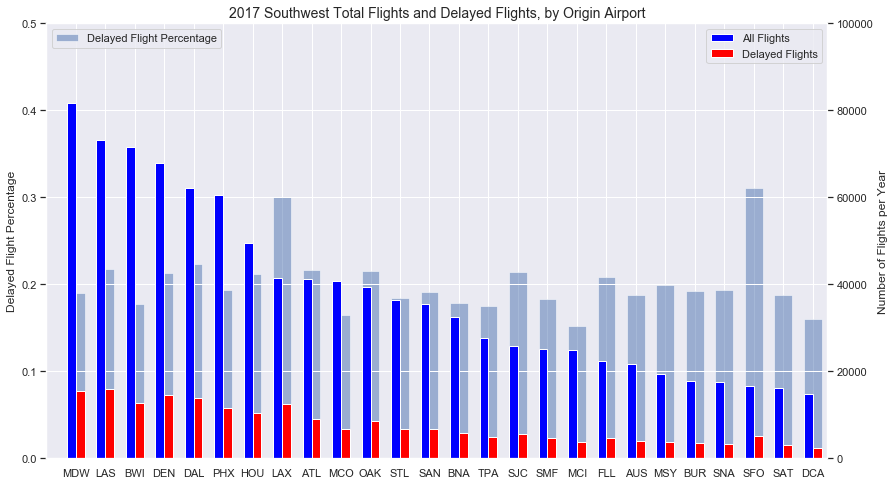
HA 0.106147

Figure 3.4 – 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.

Southwest Airlines



SFO 0.311330

LAX 0.300265

EWR 0.271363

DAL 0.223535

LAS 0.218197

ATL 0.216150

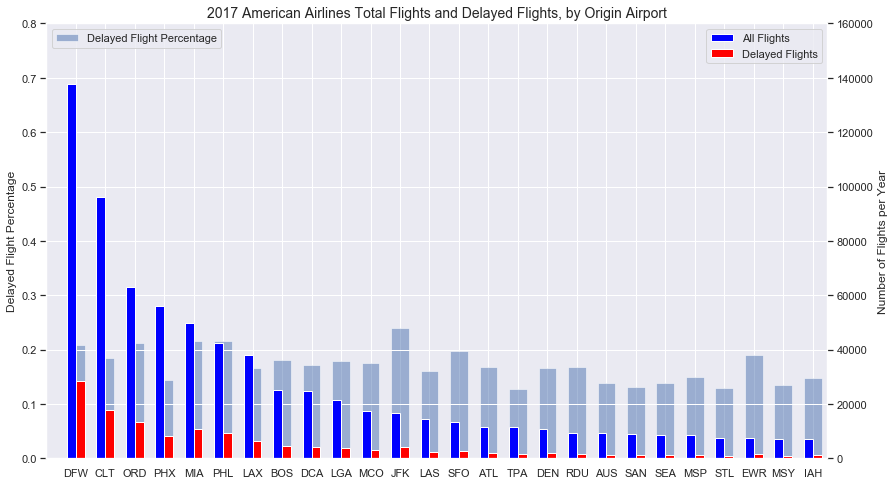
OAK 0.215234

SJC 0.214712

DEN 0.213428

HOU 0.212105

American Airlines



JFK 0.239166

MIA 0.215648

PHL 0.215173

ORD 0.212090

DFW 0.207660

IAD 0.203976

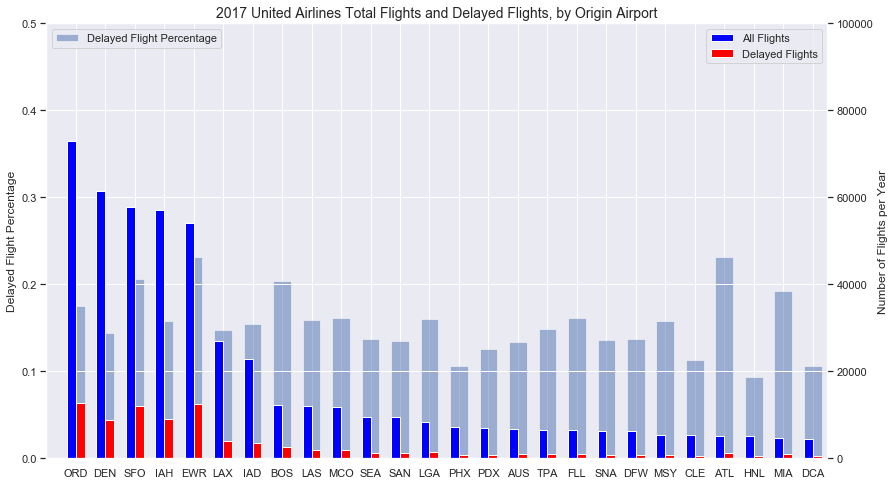
SFO 0.197801

EWR 0.189142

CLT 0.184354

BOS 0.180293

United Airlines



PBI 0.263673

EWR 0.231172

ATL 0.230904

BNA 0.219697

SFO 0.206219

BOS 0.204175

EUG 0.198057

SJU 0.197797

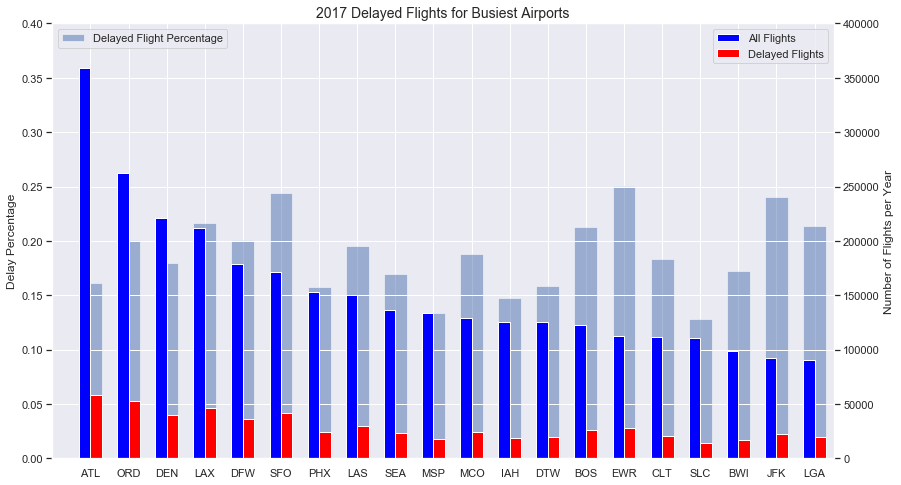
MIA 0.192012

ORD 0.174702

Figure 3.5 – 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.



EWR 0.250590

SFO 0.244134

JFK 0.240389

LAX 0.216559

LGA 0.213504

BOS 0.212972

DFW 0.201238

ORD 0.200680

LAS 0.195448

MCO 0.187614

Figure 3.6 – Top Airlines by flight volume and origin airport on-time performance

We can see, in fact 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. Let’s look at SFO and airports within 50-mile radius and compare the performance to the national average.

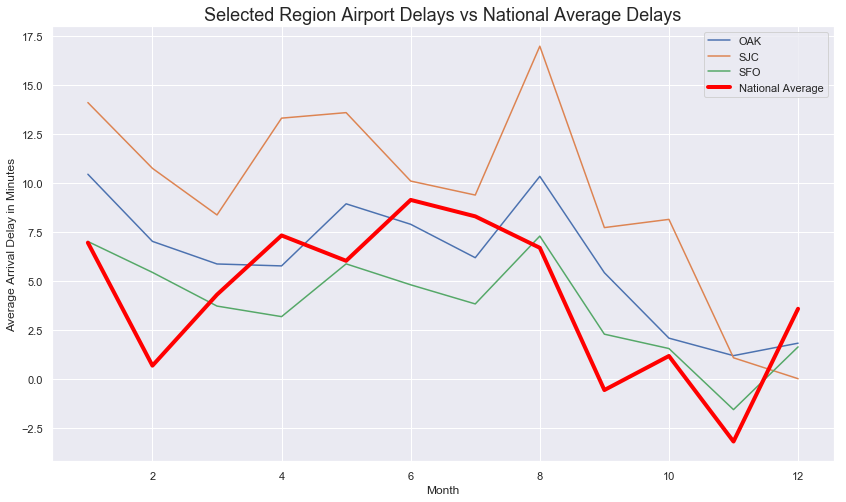


Figure 3.7 – Bay Area average delays compared to a national average, by month

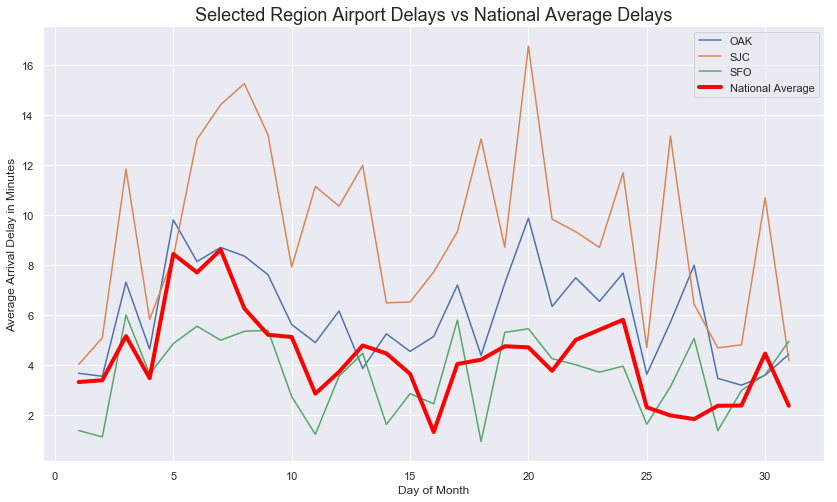


Figure 3.8 – Bay Area average delays compared to a national average, by day of month

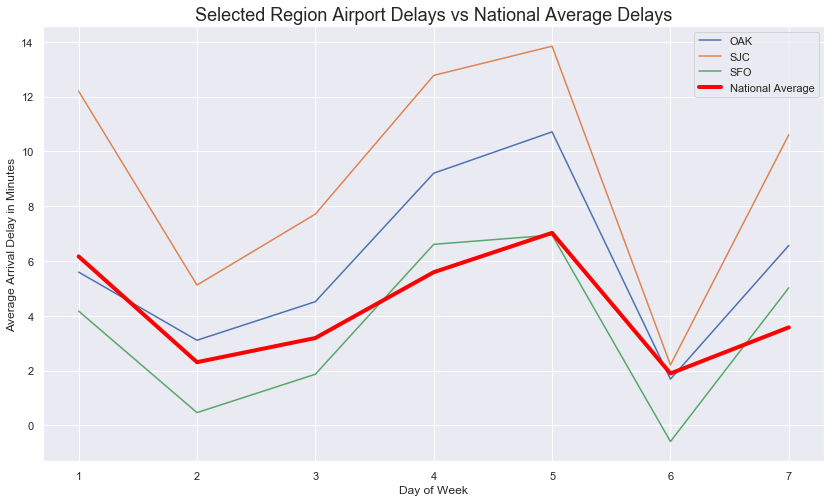
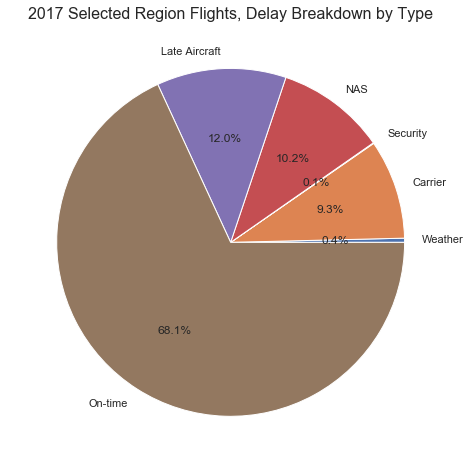


Figure 3.9 – Bay Area average delays compared to a national average, by day of week

We can see that SFO tracks along with the national average on all charts, which was unexpected. SJC seems 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. So even though SJC had roughly 13% of all flights delayed, those flights were delayed longer than other airports in the area. Next let’s take a look at how the delays break down, what is causing these delays?



NAS - We can continue to look at patterns for heavy traffic or trends in increasing air traffic patterns areas and look for constancies. 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 3.10 – Bay Area delay breakdown by cause

Breaking down the FAA data (NAS delays section from above):

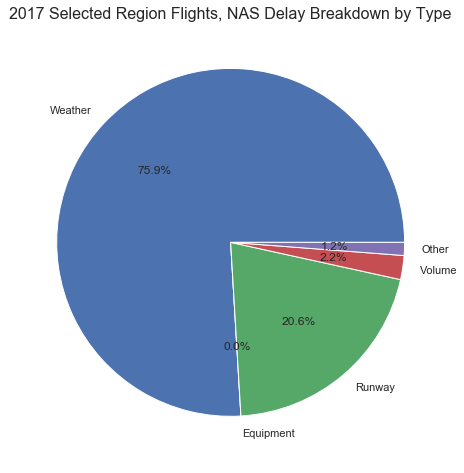


Figure 3.11 – Bay Area delay breakdown of NAS delay data

Bay Area had an on-time performance of only 68% in 2017, which is lower than the national average of 82%. We can note that late aircraft, carrier, and NAS delays made up an overwhelming majority of delays.

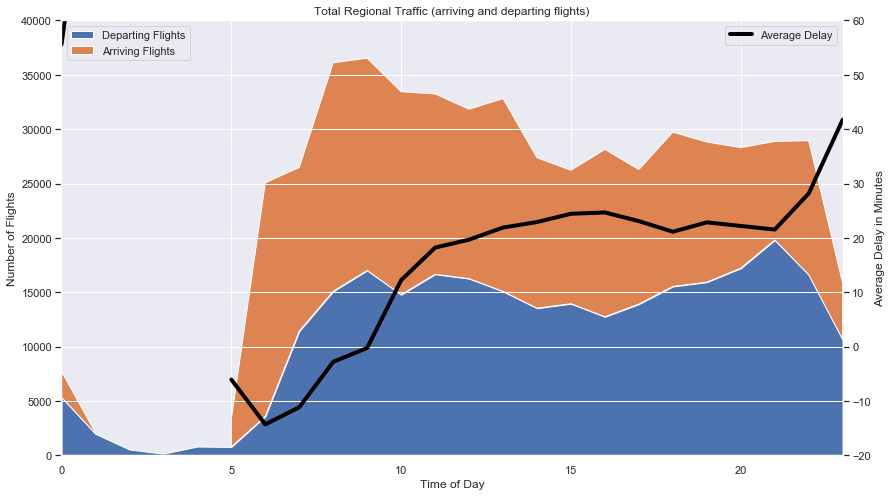


Figure 3.12 – Air traffic volume and average delay time, throughout the day

Taking into account arriving and departing flights in the Bay 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. We can breakdown the delay times by day and hour to get a full picture of performance of all airlines in the Bay Area:

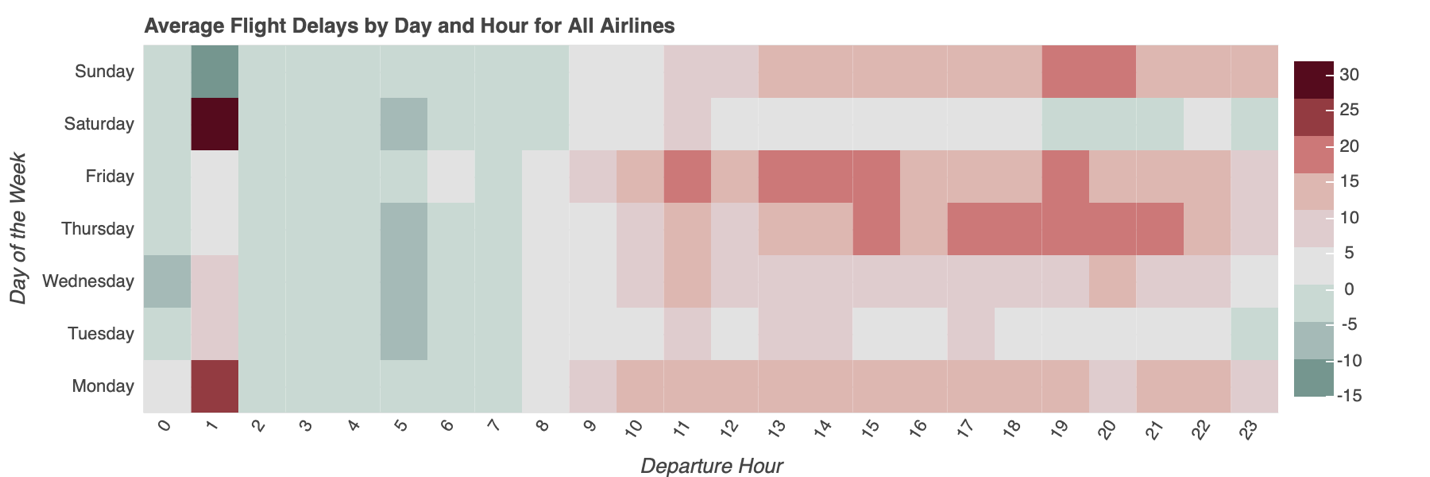


Figure 3.13 – Heatmap of arrival delays broken out by hour and day of week

It looks like Bay Area sees an increased delays during Thursday, Friday, and Sunday afternoons.