**NYC flight Data Analysis**

**Description**

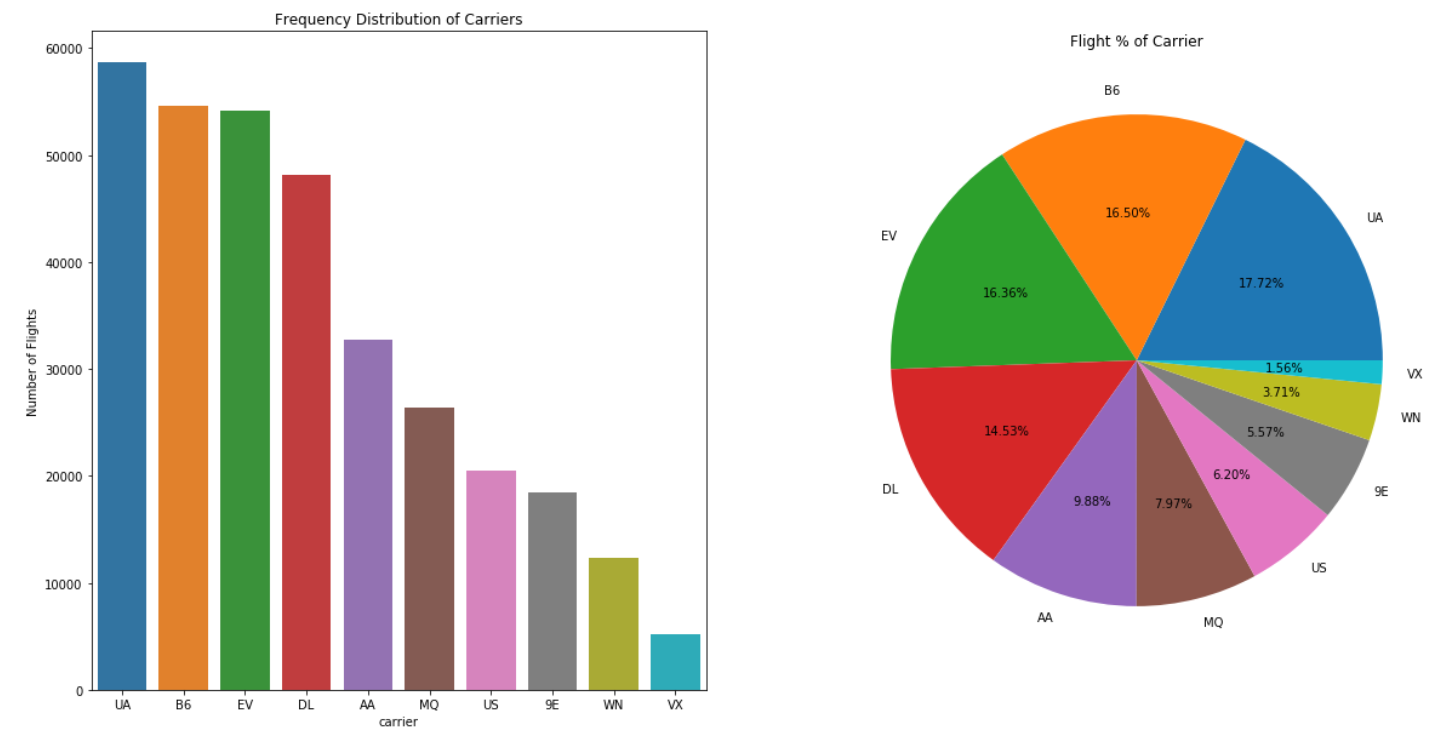
This notebook is aimed toward data analysis of flight delay for NYC in 2013. The purpose is not to obtain the best possible prediction but rather to emphasize the various steps needed to build such a model. In order to meet the possible outcomes, I worked with the central limit, Interquartile, skewness, kurtosis and correlation methods to cross-validate the results. Necessary insights were discussed during the analysis. This dataset contains information about all flights that departed from NYC (e.g. EWR, JFK and LGA) in 2013: 336,776 flights in total. The dataset misses the data related to the cause of delays. If the causes of delays were available in data set than it would be easy to pinpoint the delays, which, can be eliminated or minimised by carriers.

Format

Data frame with columns

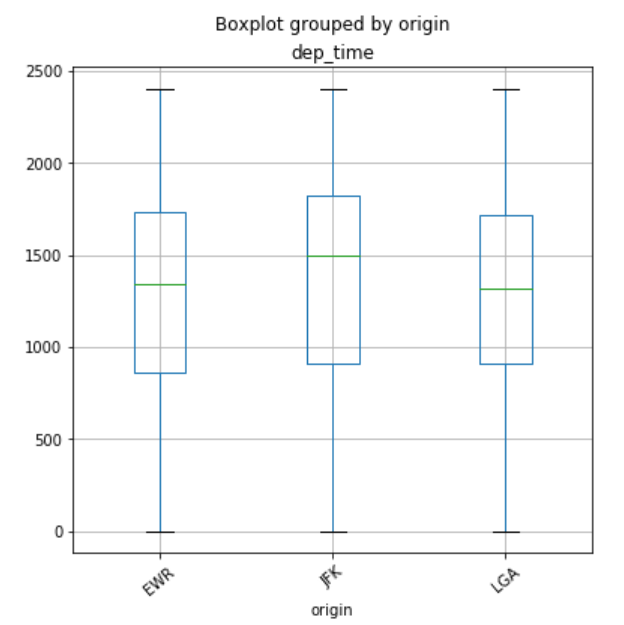
* **year, month, day** Date of departure.
* **dep\_time, arr\_time** Actual departure and arrival times (format HHMM or HMM), local tz.
* **sched\_dep\_time, sched\_arr\_time** Scheduled departure and arrival times (format HHMM or HMM), local tz.
* **dep\_delay, arr\_delay** Departure and arrival delays, in minutes. Negative times represent early departures/arrivals.
* **carrier** Two letter carrier abbreviation
* **flight** Flight number.
* **Tailnum** Plane tail number
* **Origin, dest** Origin and destination.
* **air\_time** Amount of time spent in the air, in minutes.
* **distance** Distance between airports, in miles.
* **hour, minute** Time of scheduled departure broken into hour and minutes.
* **time\_hour** Scheduled date and hour of the flight

**Visualizations**

**Graphical representation of carriers scheduled flights in numbers and %**

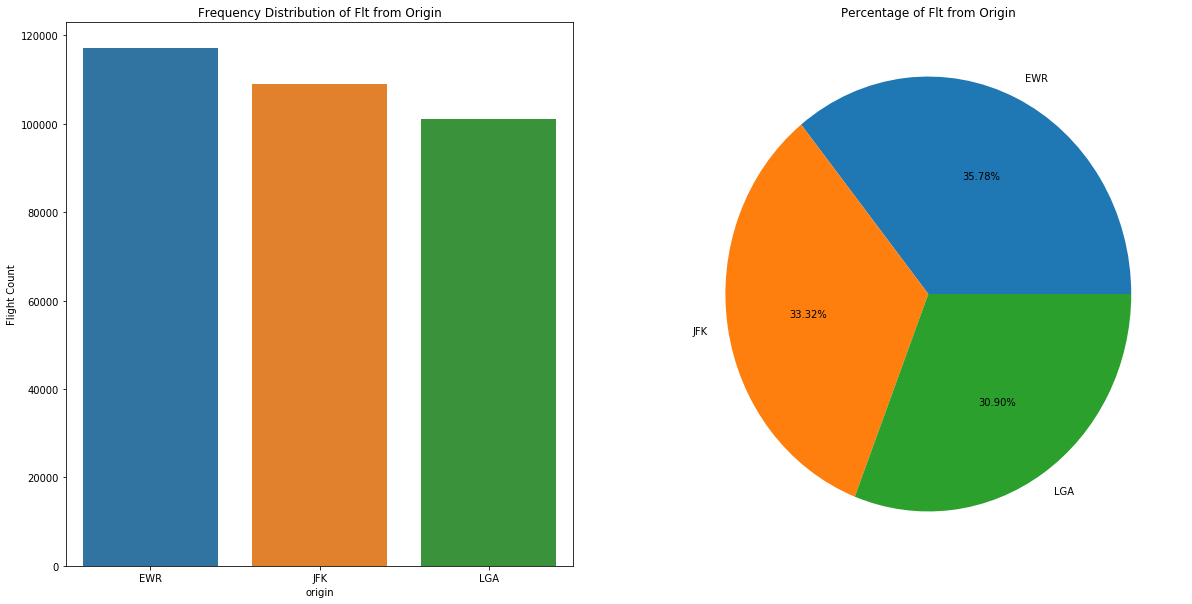
* The number of flights of each carrier that depart from NYC. The United Airlines has nearly 60K flights, then by Jetblue trailing in around 55K flights.
* 50% of the flights that depart from NYC are from three carriers - United Airline (17 %), JetBlue (16.5%) and ExpressJet (16%)

**Overall pattern of departure time from NYC airports**

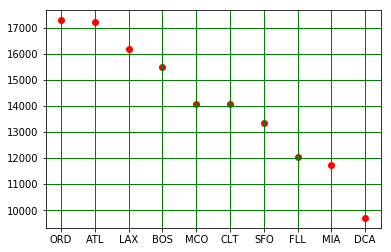


* Overall pattern of departure time from NYC airports The highest IQ range (900 to 1800) is for JFK, where 75% flights are falling under the departure time of 1800 and middle data point for departure is 1500
* The total number of destination flight from NYC is 104:

1. **EWR 117127**
2. **JFK 109079**
3. **LGA 101140**

****

The number of flights from three different airports in New York. All three have quite huge number of flights departures all around the year but EWR airport comes first with almost 120K flights.

****

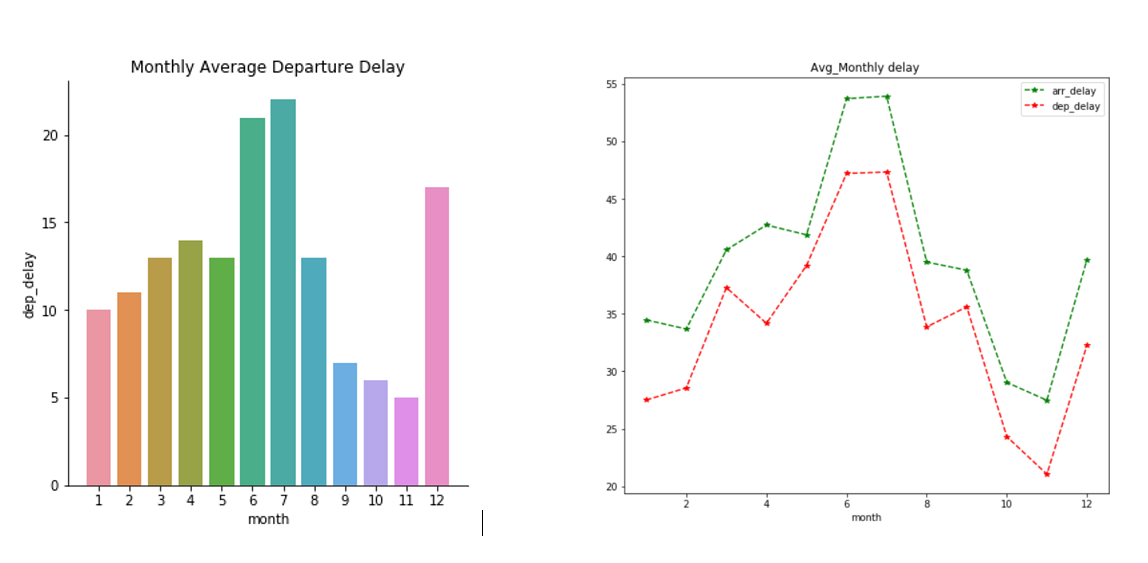
Most flights destinations are to Chicago, Atlanta and Los Angeles. Together the number of flights are nearly 50K. Then comes Boston, Orlando and others.

**Total number of unique Airline headed to BOS from NYC**

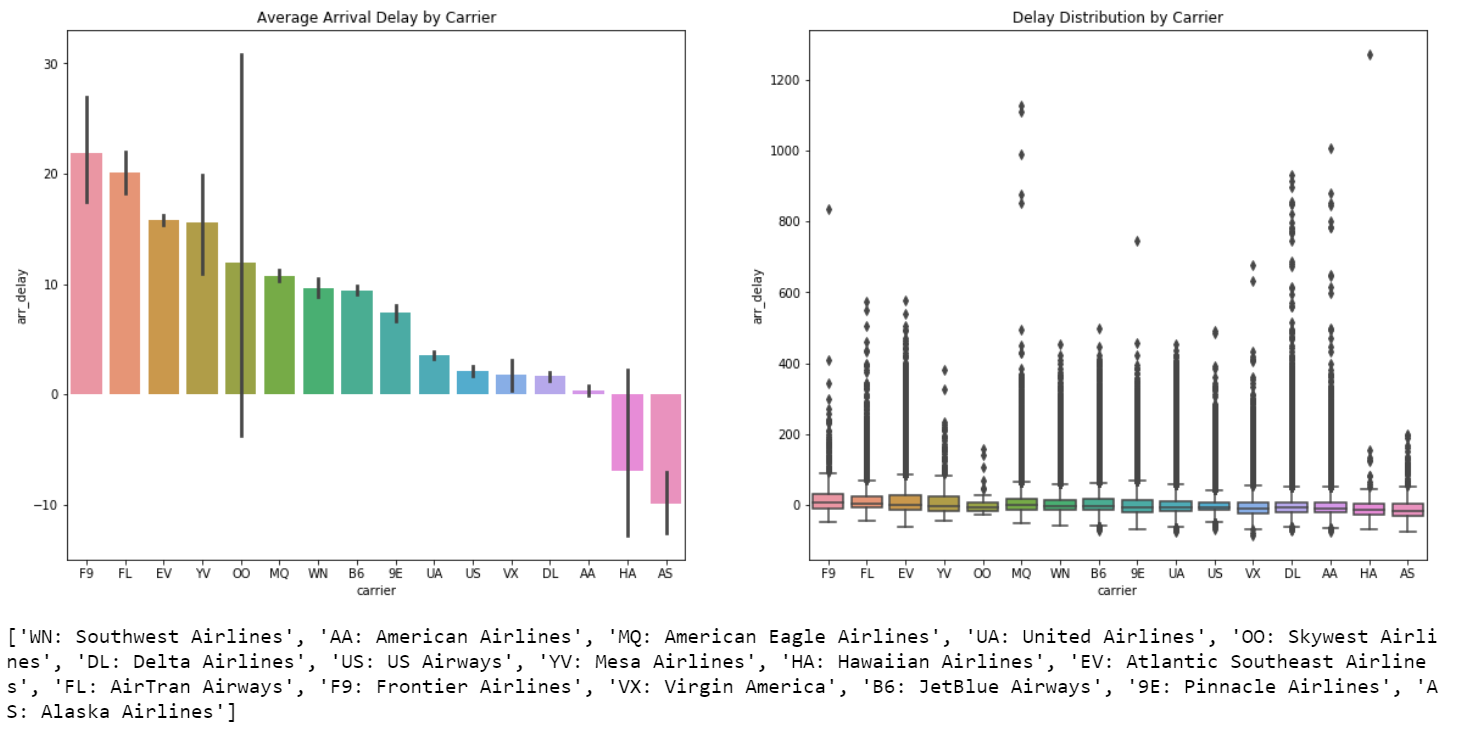
* **Carrier fly to BOS ['B6' 'AA' 'DL' 'UA' 'US' '9E' 'EV']**
* **Total number of Carrier headed to 'BOS' from NYC is 7**
* **Total unique aircraft headed to 'BOS' from NYC is 1308**



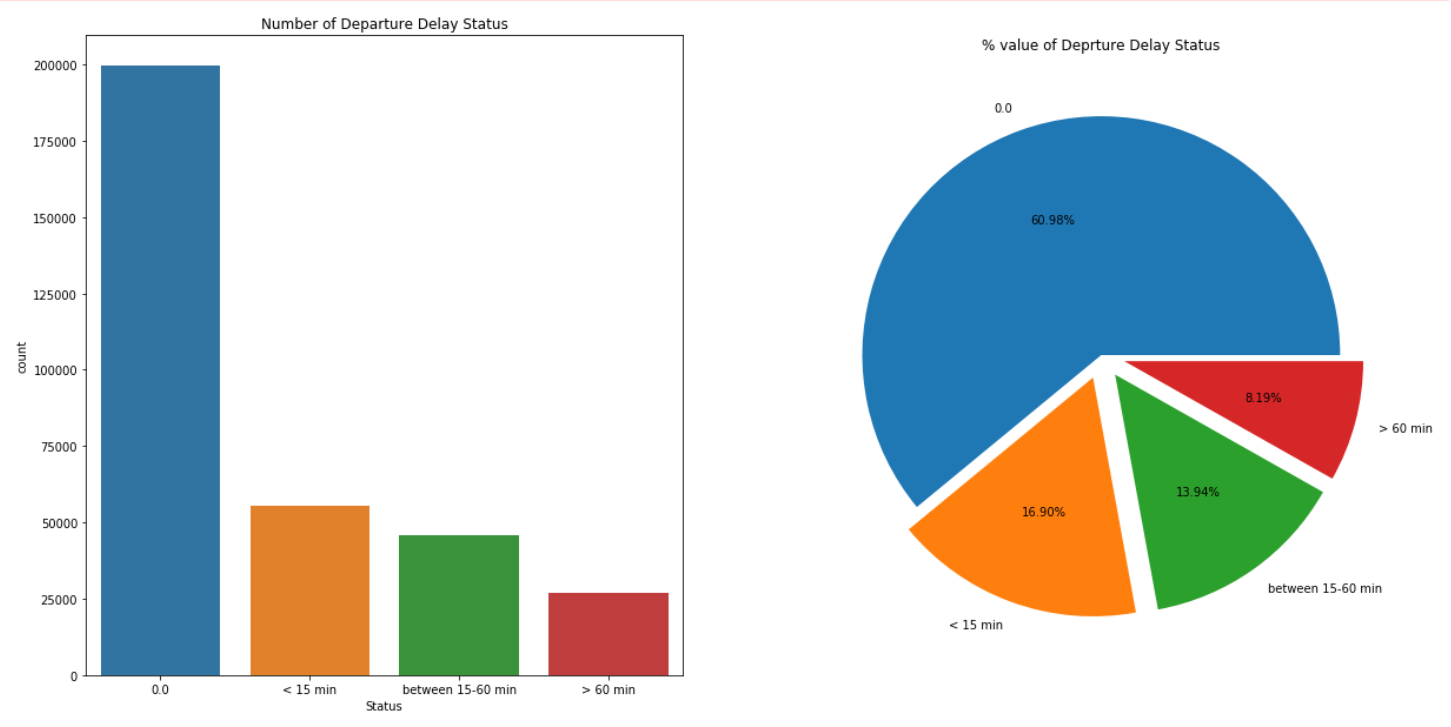
The average monthly delay time for One air carrier. Most delays are in May, June and July. The least delay months are in September, October, and November.

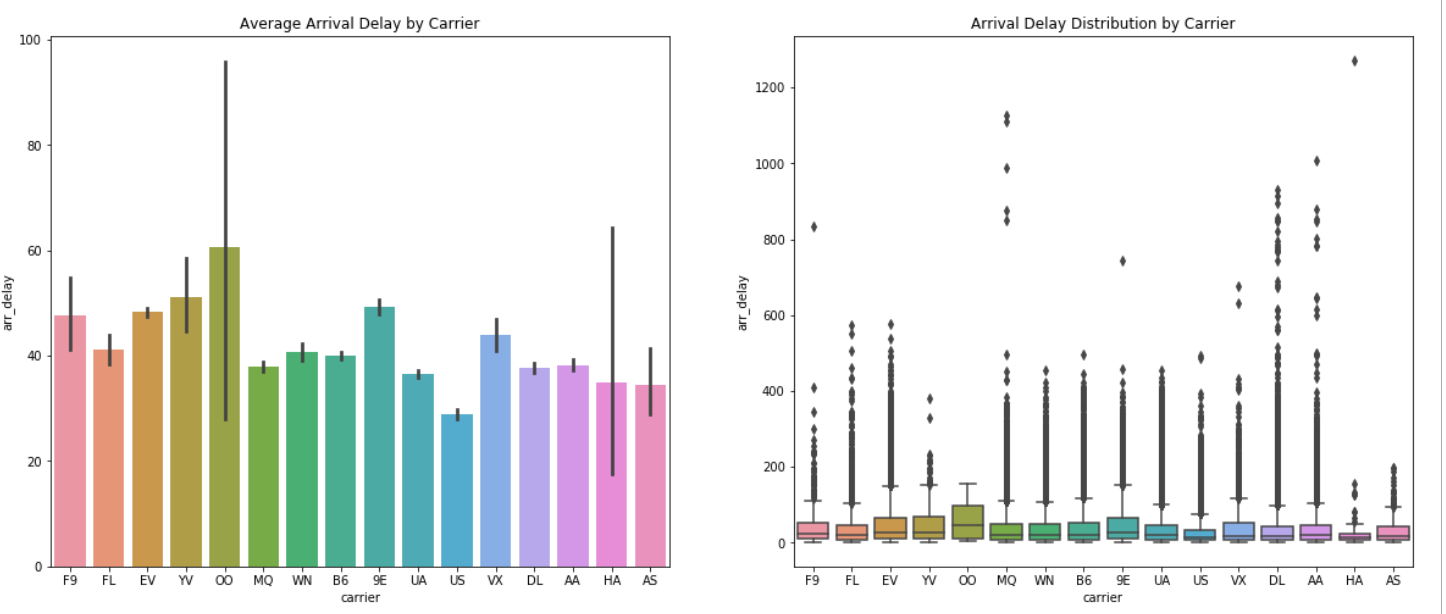


The Average Monthly Departure Delays are High in the month of June and July. The most probable reason is the weather. The delay in December might be due to large crowd travelling across.

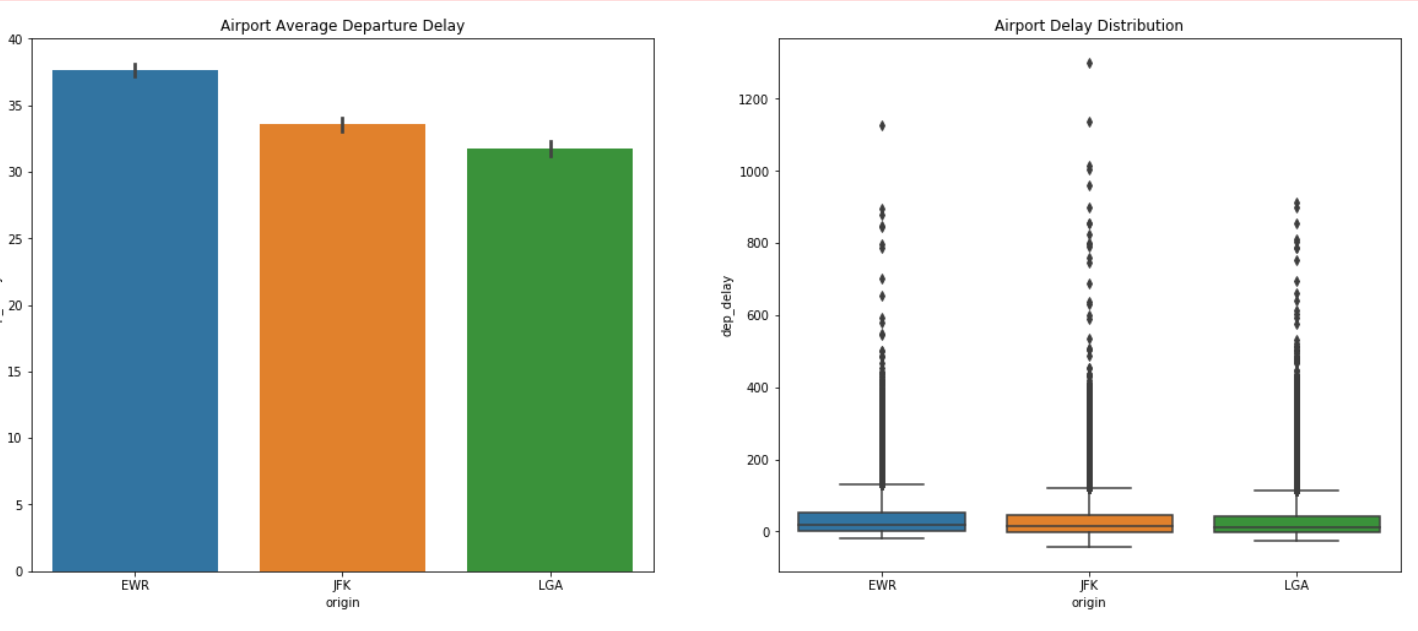


The Average delay for Arrival is high with Frontier and AirTran Airways. The best carrier to reach destination are Hawaiian Airlines and Alaska Airlines (Pilots reach holiday destinations faster)

The Number of flights that gets delayed. Up to 60% flights do not have any delays and just 9% flights gets delayed for more than an hour.



* The top 5 US airline (American Airlines (AA), Southwest Airlines (WN), Delta Air Lines (DL), United Airlines (UA), Alaska Airlines (AS)generate an average delay of 37.2 minutes. Alaska Airlines, with an 34 minutes per flight, the second lowest of all the carriers.
* Carriers with higher average delay generation are Skywest Airlines(OO) with 60 minutes per flight, Mesa Airlines (YV) with 50 minutes per flight, and Pinnacle Airlines (9E) with 49 minutes per flight. The error bar provide the insight that airlines with low number of flights having higher standard deviation distribution from the mean (OO, HA, YV, F9, AS); so it seems like size matters
* The boxplot shows, airlines with higher number of flights results having a higher chance of extreme waiting situation. American Eagle Airlines (MQ), American Airlines(AA), Delta Airline(DL) registered the maximum Carrier Delay for 2013 with an exception of Hawaiian Airlines (HA).



* There seems to be a correlation between the number of flights operated and departure delay. A descending pattern can be seen from barplot for average departure delay per flight. Considering, the assumption that JFK being a busiest airport among 3 due to international flights; so the maximum departure delay for 2013 is registered by JFK.

**Performance Analysis**

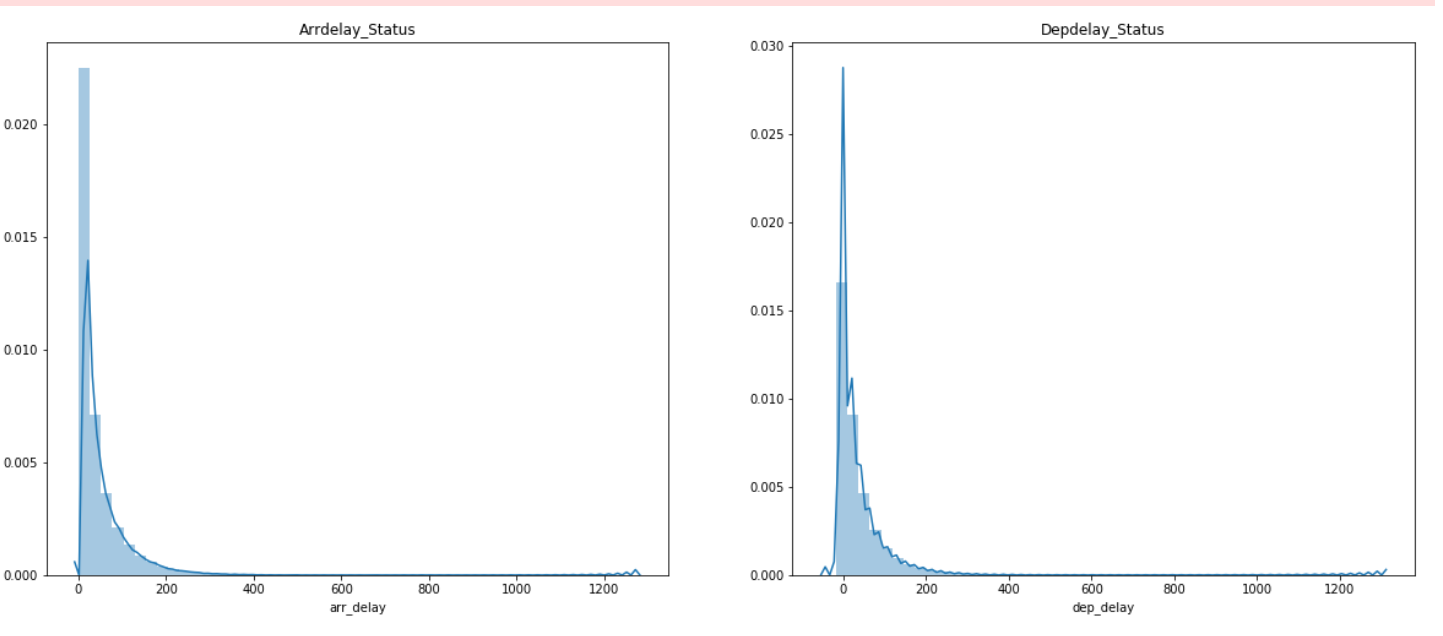
* Sched\_dep: 336776
* Operated: 328521
* Cancelled: 8255
* Ratio operated flights over scheduled flights: 97.54881583010666
* Ratio of cancelled flights: 2.451184169893338

The day and month having highest delay by average for departures

* 8/03/2013 Avg-Delay: 83.65

Day and month which have highest number of flight delay

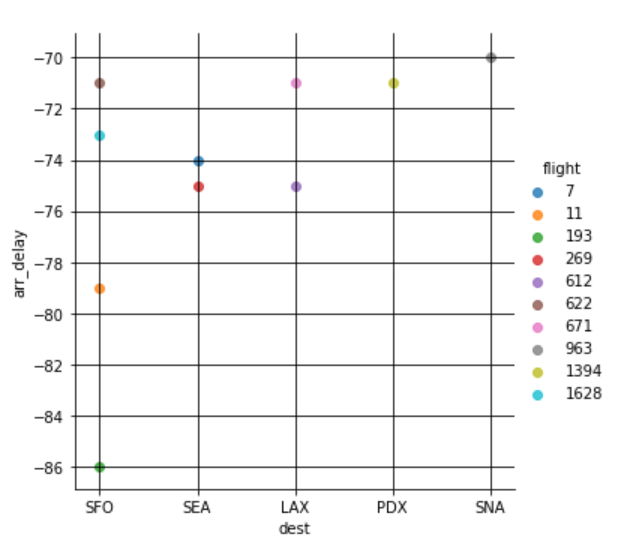
* 23/12/2013 Number of flights: 673



* Skewness\_arr: 3.433909
* Kurtosis\_arr: 24.606943
* Skewness\_dep: 3.250293
* Kurtosis\_dep: 22.521400

It can be seen on the histogram and by the skewness and kurtosis indexes, that the skewness is >1 which reflect the data distribution is highly positively skewed and Kurtosis>3 shows the leptokurtic distribution, having longer and fatter tail with a central peak higher and sharper.

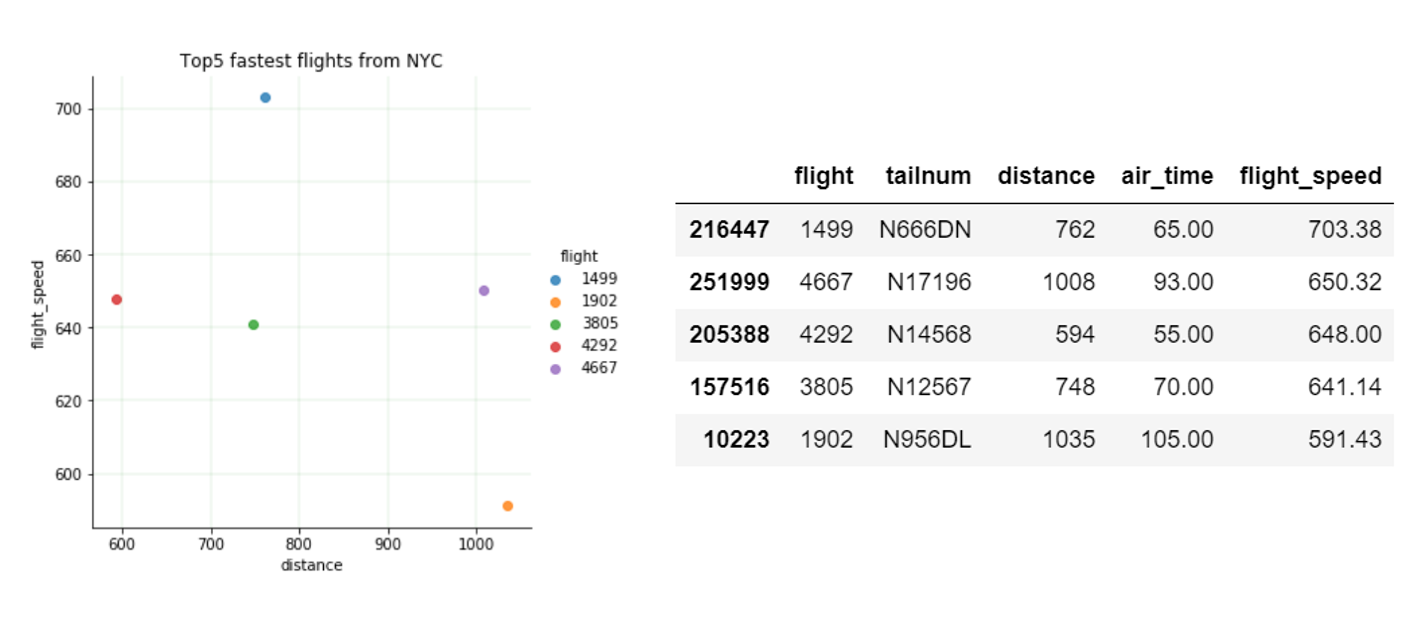
The histogram shows the delays are mostly located on the left side of the graph, with a long tail to the right. The majority of delays are short, and the longer delays, while unusual, are more heavy loaded in time.

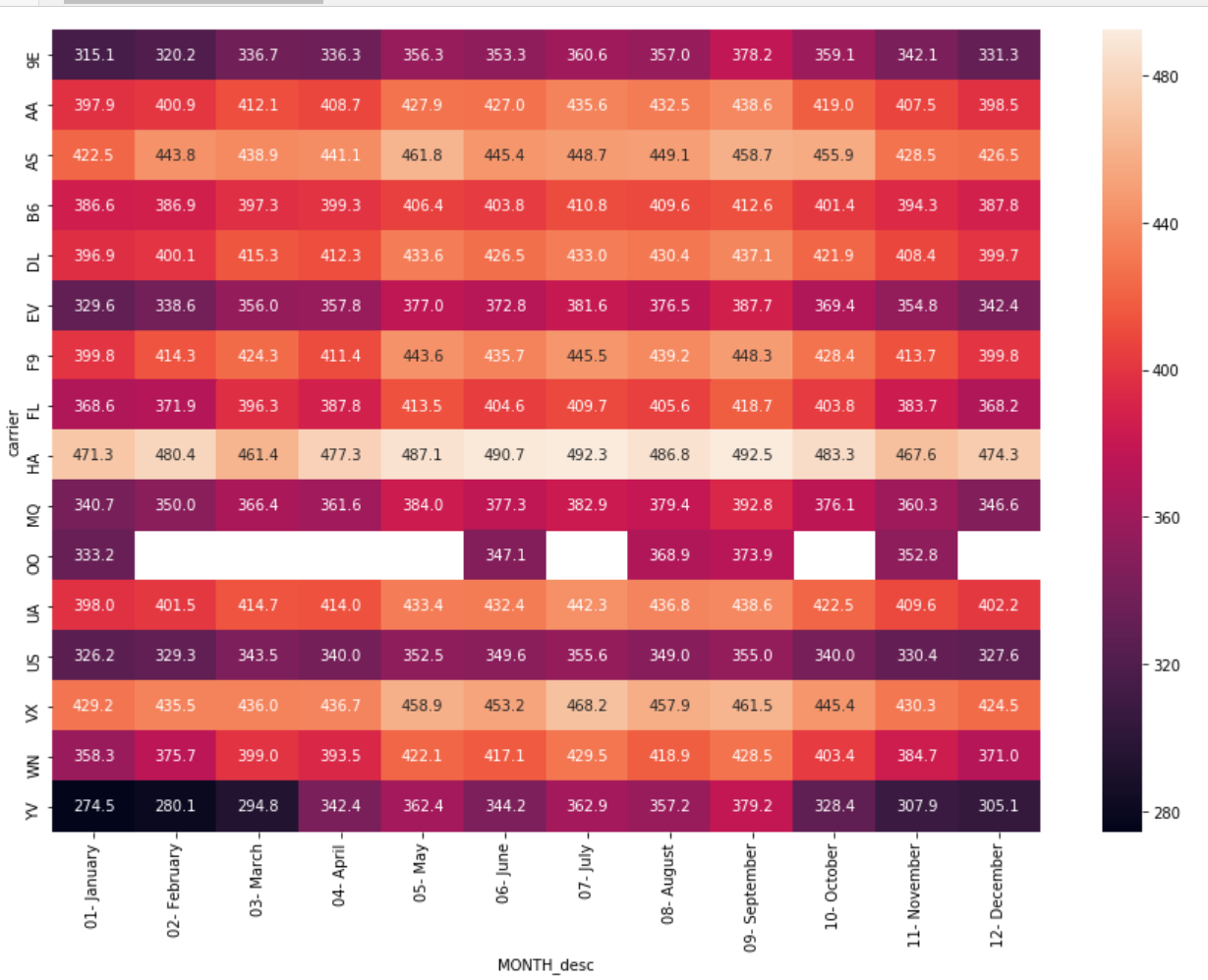


Top 5 destinations where flights arrives early than expected arrival time:

* SFO
* SEA
* LAX
* PDX
* SNA

Top 5 fastest flights from NYC





The peak season for air travel in USA is considered to be June to August and lean season is mid of January to February. The airlines operate highest number of flights and carry maximum PAX load during the summer season and vis-à-vis during lean season. The data proves that the statement is true and most of the airlines having maximum departure between May to August and minimum between January to February. From, the heatmap, it is visible that during May to August most of the airlines tend to fly faster than normal flight speed, to cover maximum departure. Whereas, it is vis-à-vis during lean season.

**Conclusion**

Though the dataset doesn't offer reasons for delays and missing important data such as taxi in and out, flight diversion, chocks on and off timing, and fuel consumption. So, it is clear that the dataset doesn't provide clear understanding of delay issues, which may be supportive to look into delays that can be controlled or reduced.

For example: If airlines permits pilots to fly aircraft at higher speed and fuel consumption on planes that departed late, the delay spread can be minimize along the flight network. This would decrease the possible delay itself and significatively reduced the number of aircraft delays.

A solution applicable to one type of delay may affect the others, resulting in a ripple effect that will allow more efficient operations; benefiting passengers, airports, carriers and even the world as a whole.

Reference: <https://www.kaggle.com/lampubhutia/nyc-flight-data-analysis>