**Project title:** Airlines/Airport performance-based Analysis

**Team members:** Parag Siddamsettiwar, Abhishek Upadhayay

**Abstract:** Analysis into flight data on US air traffic to better understand and suggest a carrier or airport.

**Introduction (motivation, background, and objectives)**

In today’s busy world, time is a very precious commodity. Whether you are travelling for work or leisure, mode of transportation plays an important role in deciding everything from time to cost estimation. Flights are the most convenient and fastest way to travel. In fact, frequent travelers prefer living in a city connected by airport. Airlines and airports are becoming critical necessity for every business, as we are moving towards collaborative global economy.

More and more people are taking flights and according to IATA, around 8.2 billion people will be travelling by flights by 2037. [1] This trend will continue to grow until new faster mode of transportation is invented and implemented which will require at least another lifetime.

Apart from cost, major factor, that is becoming more and more important for choosing airlines or nearby airports, is reliability. Flight delays and cancellations are a major cost escalator for businesses, passengers and Airlines. In 2010, passengers lost as much as $16.7 billion due to schedule buffer, delayed flights, flight cancellations and missed connections. [2] According to a research, an average Briton wastes 16 days of their holiday time waiting for their flight to take off. [3] Whether you are going on a business trip or going on a vacation, planning and booking a flight that will make the entire trip as planned and pleasant is very important.

Remember the time when your flight was delayed or cancelled? The hassle to adjust the entire trip or adjust the bookings takes costly toll on everyone including passengers, airlines and businesses.

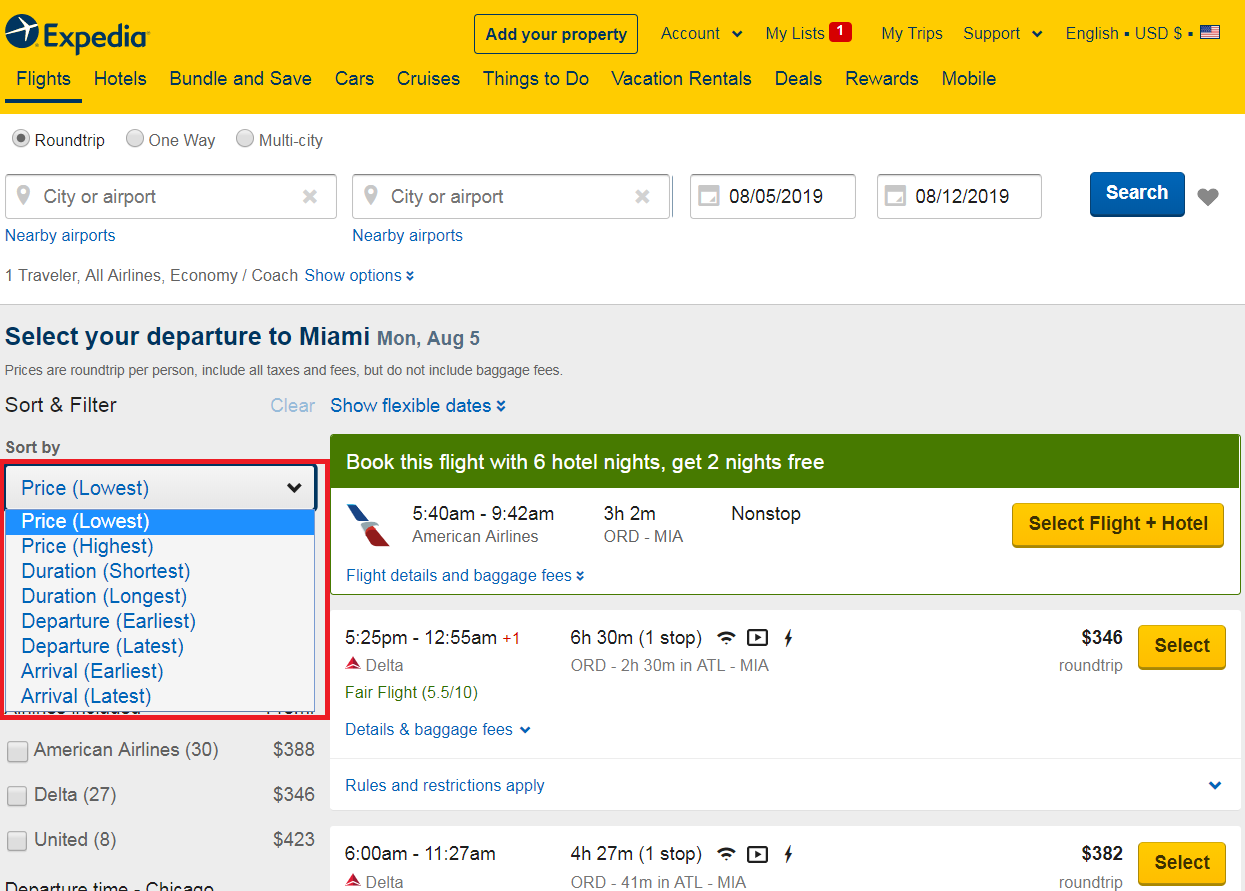
Some existing travel websites or journals do provide high-level information in form of visualizations (FlightAware, 2019) and reports (Worldwide, 2018) of delays for any particular airline at any particular airport for specific historical dates. However, none of these data visualization provides any comparison or recommend one airline/airport over the other in a single visualization or single page. e.g. Most travel websites will show you the best deal by price or journey time (Frontier, 2019) but they won’t show you usual delays and cancellations where our visualization and analysis will help the user in making an informed decision.

Now let us explore some of the flight booking websites like Expedia, Kayak or google flights.

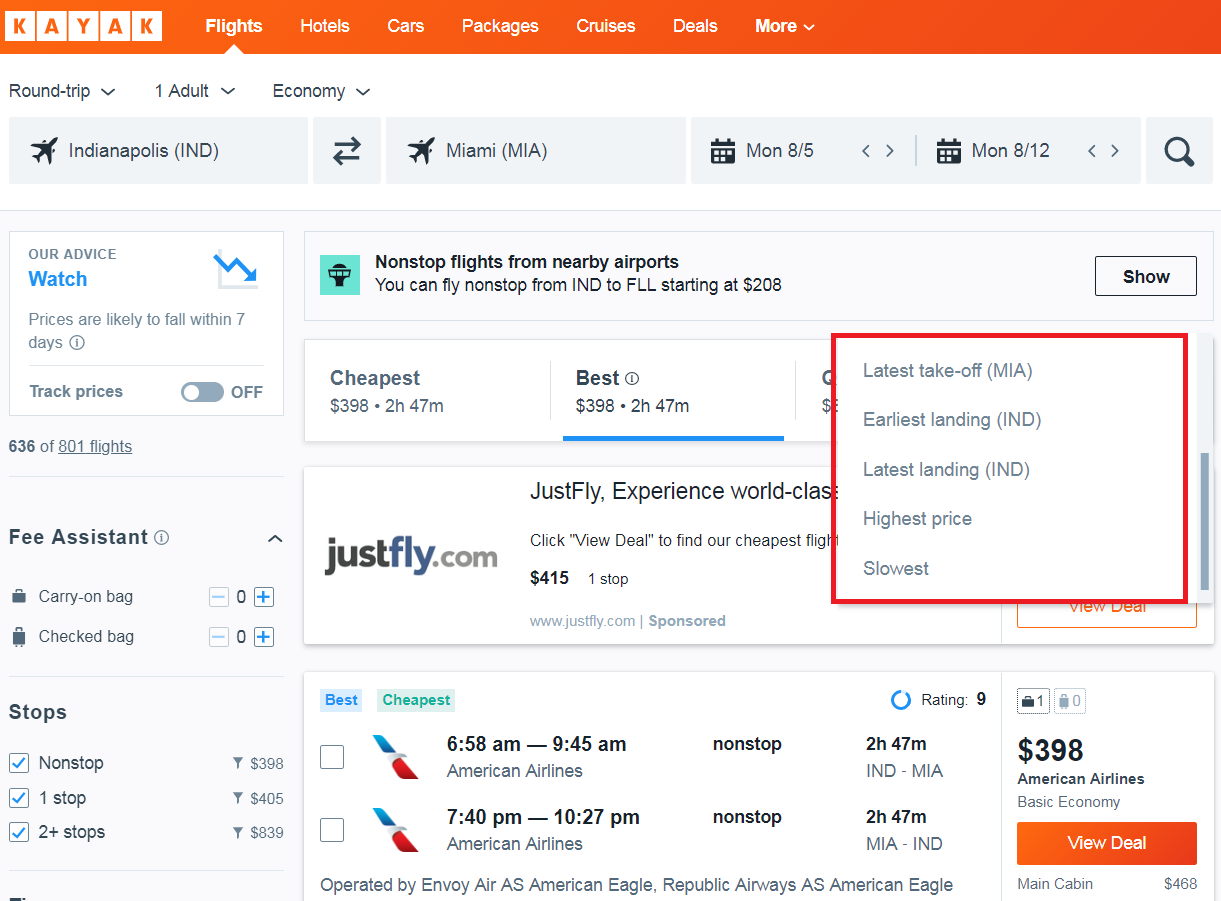
All websites that help us book flights only compares flights across different airlines by cost. You have an ability to filter out flights based on flight time, departure or arrival tie, non-stop, one stop or multiple stops etc. Due to this, even if a particular airline has a very high rate of cancellation or delay of flight from a particular airport but if the cost of flight is low, people will tend to book that flight only to discover the hassle they will end up having to deal with later. The only option available is to check for customer reviews and make a decision based on that but that is definitely not a good source of data to decide something this important.

For example:

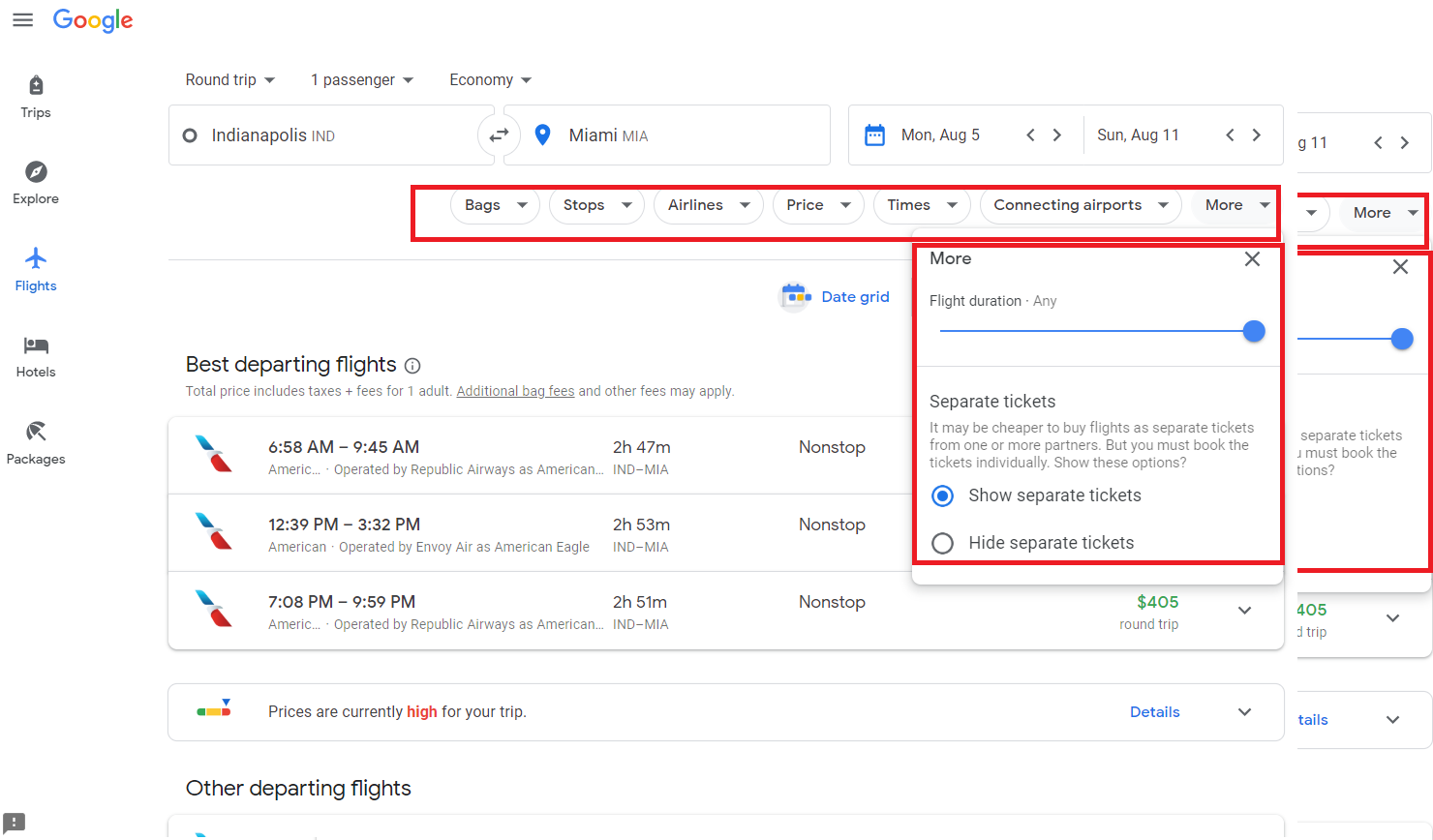
Expedia:



Kayak



Google Flights:



It would have really helped if we can analyze and visualize the delayed / cancelled / diverted flight data to get a better understanding of Airlines and Airports to make in informed choice before finalizing our travel plans!

This brings to a very important topic, is it possible to visualize airlines and airports based on delays, cancellation or diversions?

Is it possible to use these factors by using reliable and correct data source to help passengers make informed decision?

Different countries have different airline authorities that control airline traffic for their respective countries. Usually these authorities hosts the data of these airlines including delay information.

For United States, U.S. Department of Transportation, a government agency, maintains and make this data public. Thanks to Bureau of Transportation Statistics, which is part of DOT, we have found the source for reliable, robust and continuous data of flights including delays, cancellation and diversion information.

These government agencies also regularly publish data via their open data Data.gov platform

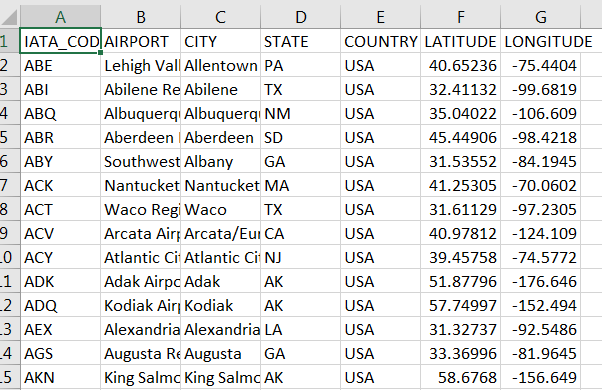
We will be pulling on time performance dataset of all the flights of United States. Such airline dataset is generally associated to IATA\_CODE which is nothing but a unique code given to every airline and airport. It is used to across databases for easy cross-reference.

Let us look at couple of basic datasets:

**Airlines:** This is needed to map IATA\_CODE with actual airline names



Airports: This is needed to map IATA\_CODE to actual names of airports.



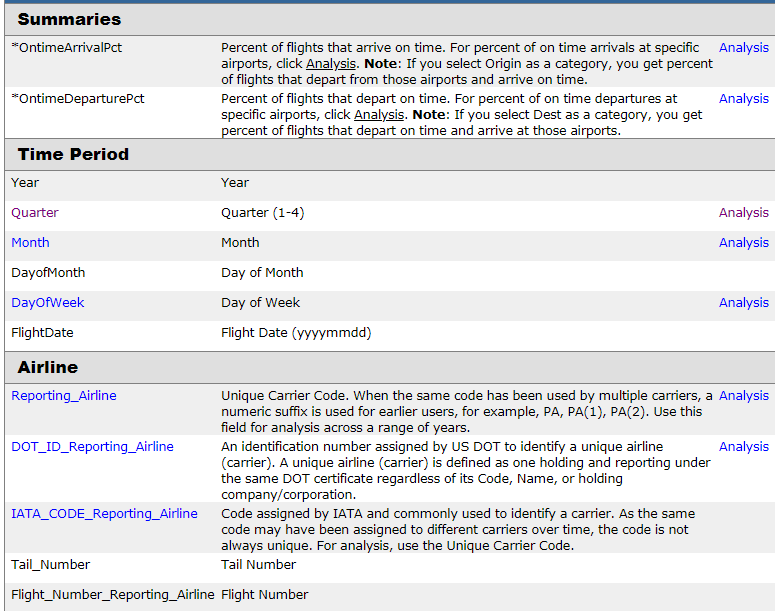
On-time performance dataset of USA domestic flights. This dataset is regularly updated and published which can be downloaded from following URL: <https://www.transtats.bts.gov/Tables.asp?DB_ID=120&DB_Name=Airline%20On-Time%20Performance%20Data&DB_Short_Name=On-Time>

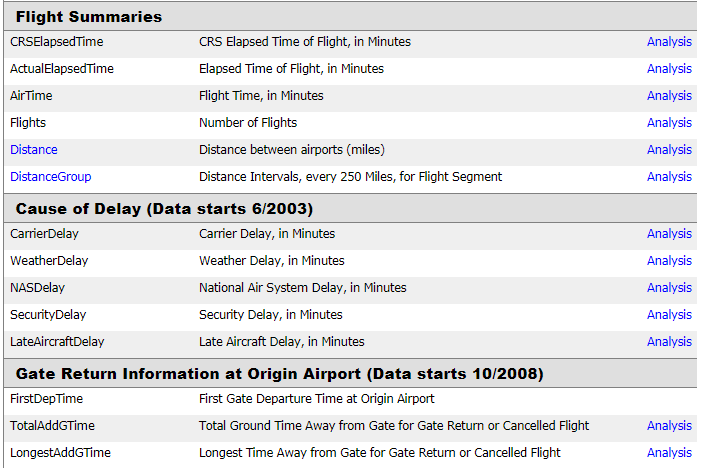
This is a large dataset with millions of records depending on how much historical data we download.

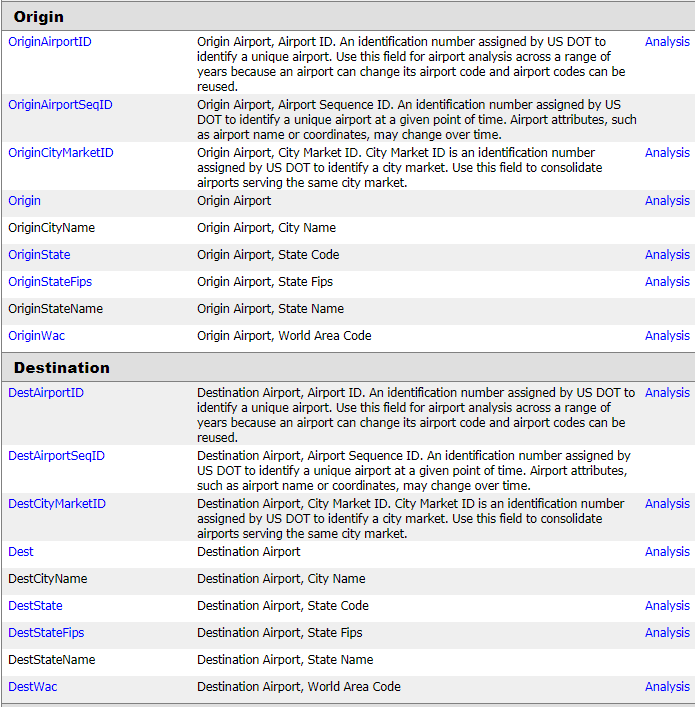
For showing high-level information, header is shown below to list all the dataset column names:

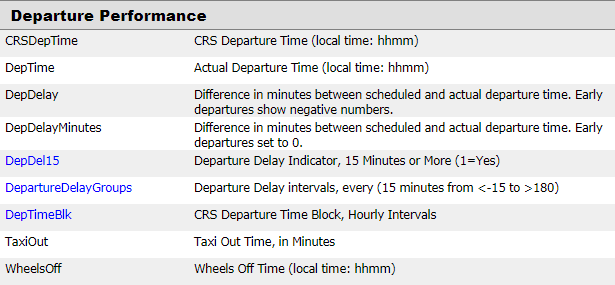
Index(['YEAR', 'MONTH', 'DAY', 'DAY\_OF\_WEEK', 'AIRLINE\_CODE', 'FLIGHT\_NUMBER','TAIL\_NUMBER', 'ORIGIN\_AIRPORT', 'DESTINATION\_AIRPORT','SCHEDULED\_DEPARTURE', 'DEPARTURE\_TIME', 'DEPARTURE\_DELAY', 'TAXI\_OUT','WHEELS\_OFF', 'SCHEDULED\_TIME', 'ELAPSED\_TIME', 'AIR\_TIME', 'DISTANCE','WHEELS\_ON', 'TAXI\_IN', 'SCHEDULED\_ARRIVAL', 'ARRIVAL\_TIME','ARRIVAL\_DELAY', 'DIVERTED', 'CANCELLED', 'CANCELLATION\_REASON',’AIR\_SYSTEM\_DELAY', 'SECURITY\_DELAY', 'AIRLINE\_DELAY','LATE\_AIRCRAFT\_DELAY', 'WEATHER\_DELAY', 'DATE', 'IATA\_CODE', 'AIRLINE','AIRPORT', 'CITY', 'STATE', 'COUNTRY', 'LATITUDE', 'LONGITUDE'],dtype='object')

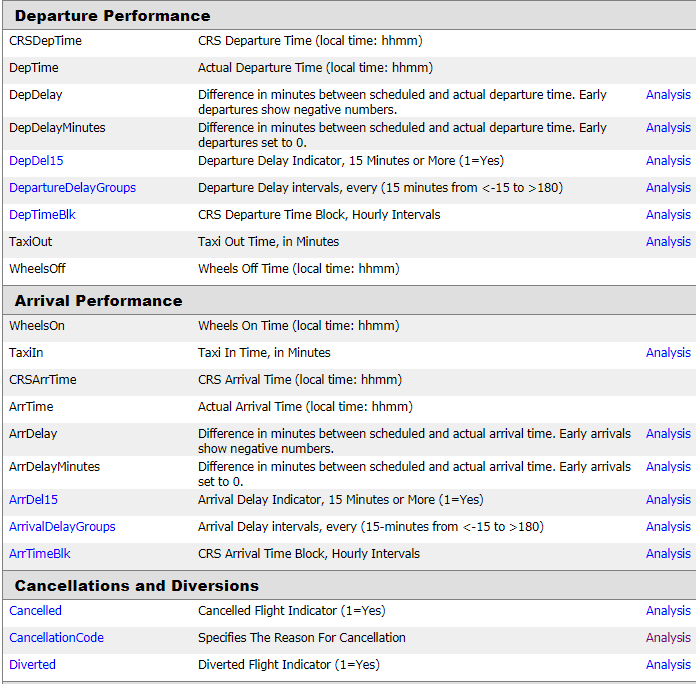
Additional details of these can be found on the transtats website which is as follows:

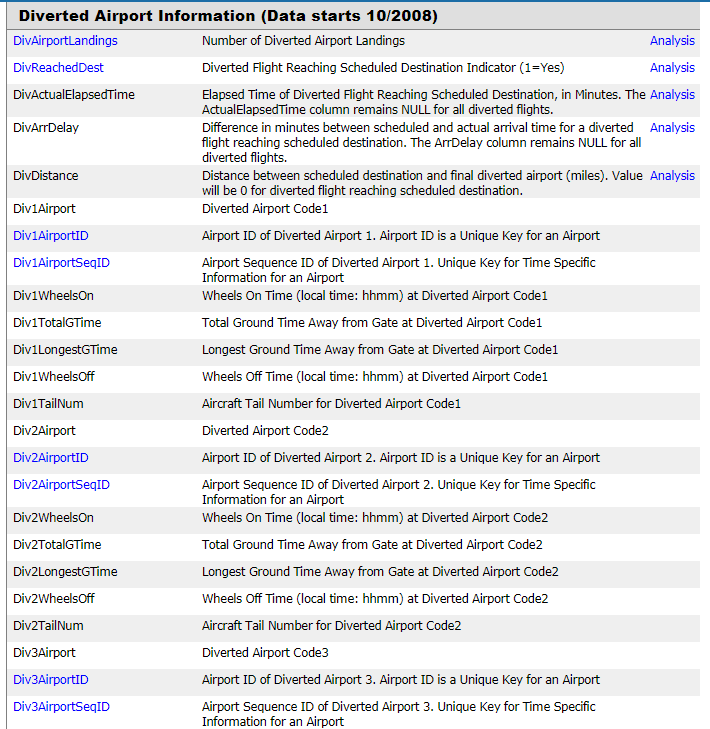














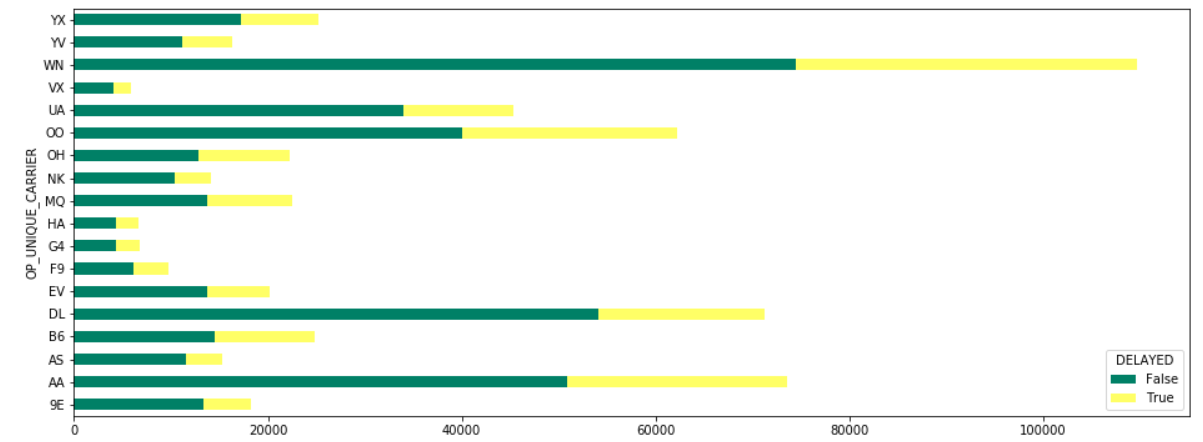
After carefully studying the data columns available to us, we have decided on downloading few of these that makes more sense in current analysis. We downloaded 10 years of data and tried to load it into python for doing exploratory analysis. As this data se very large and we have limited memory available on our local computers, we ended up noticing extremely slow responsive behavior of python program. It was taking way too more time to load data, combine it. We wanted to focus our work towards creating visualization and not on how much data we can bring in so we decided to restrict our dataset to 1 year of flights data. After using only 1 year of data, we were able to load it comparatively quickly.

Top 20 delayed flights

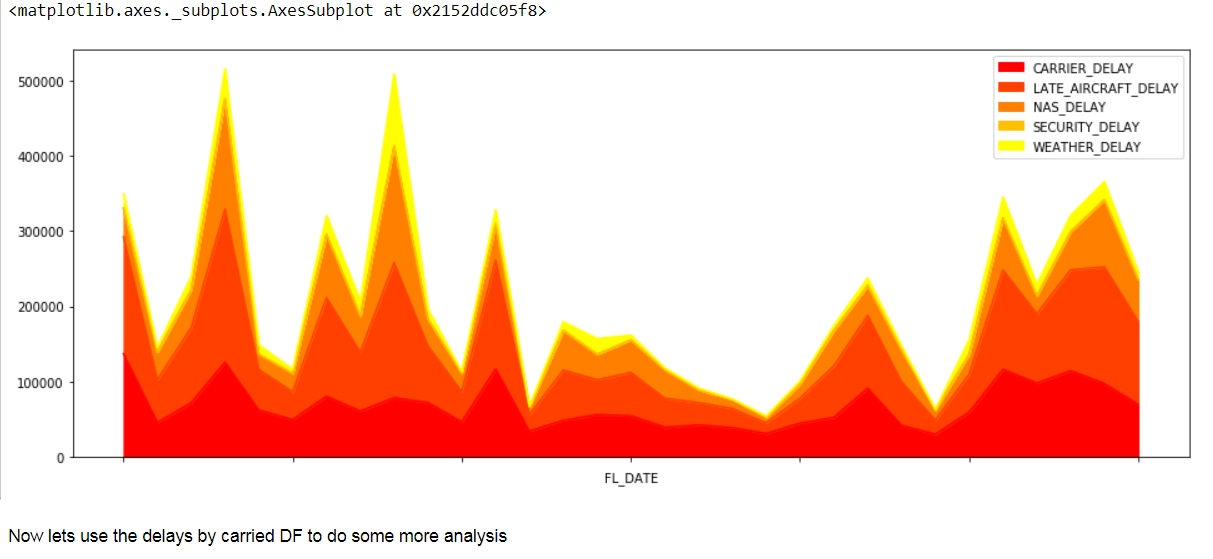


The most delayed flight was almost 2000 minutes delayed, almost a day

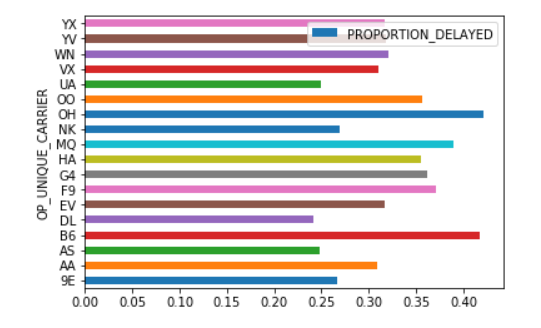
On time vs delayed ration of flights per airline



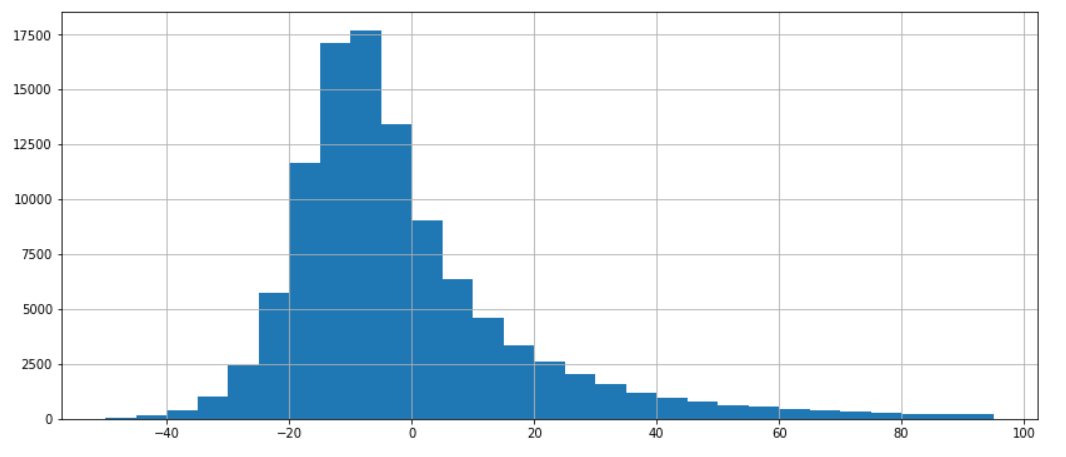
Flight delay reason heatmap



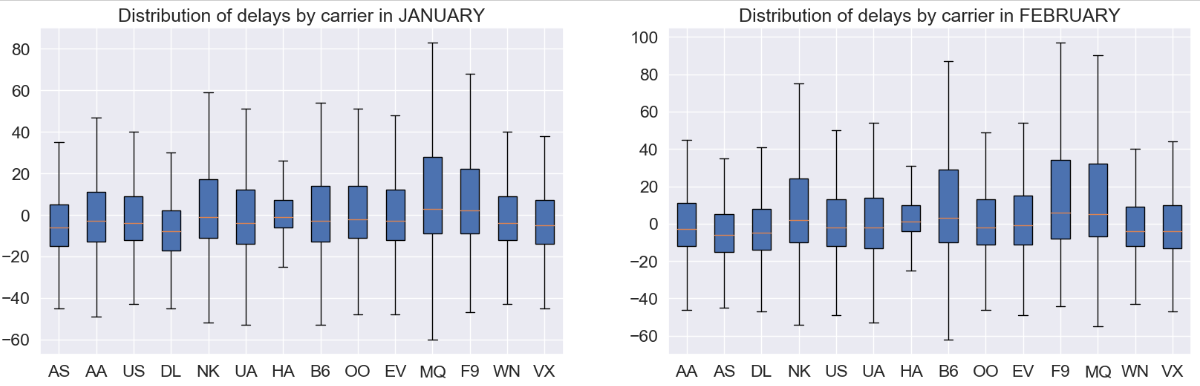
Delay proportion

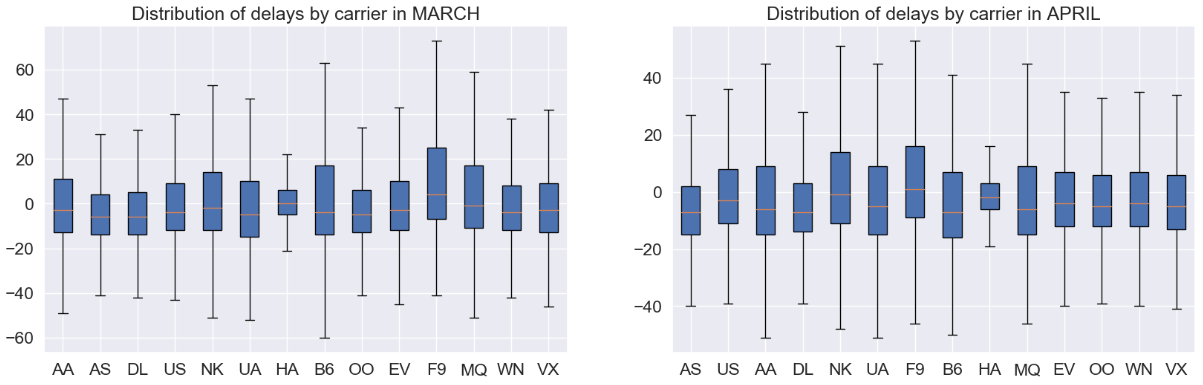


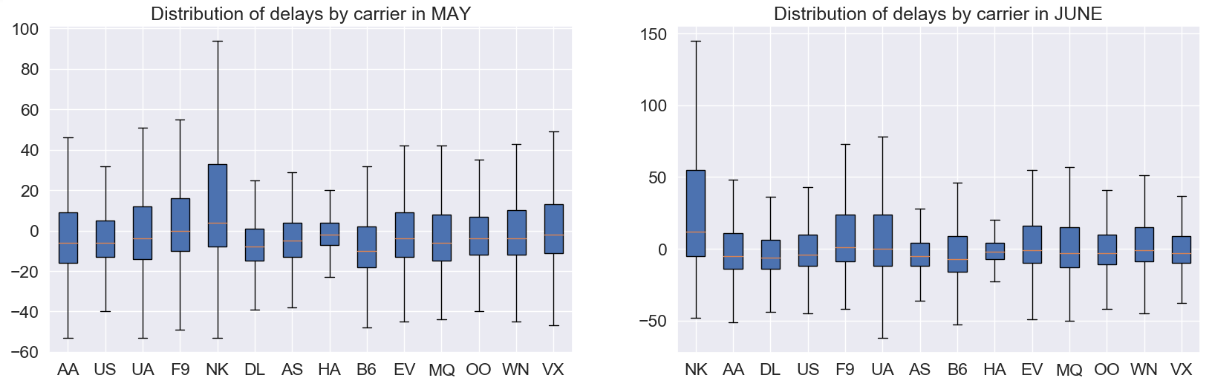
Southwest delay distribution

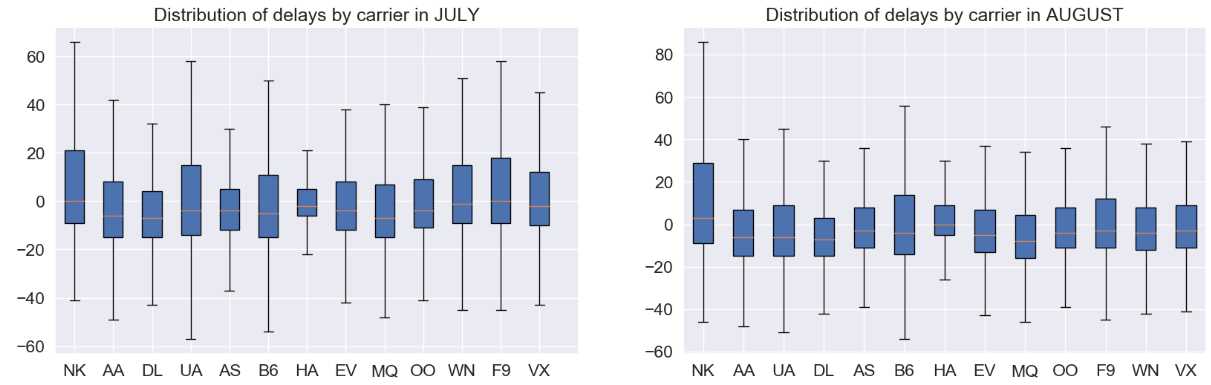


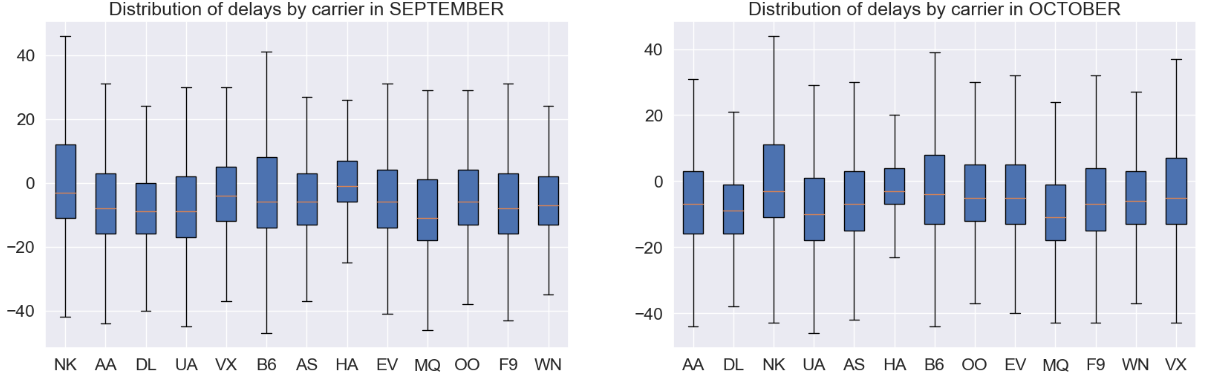
Distribution of delay by carrier

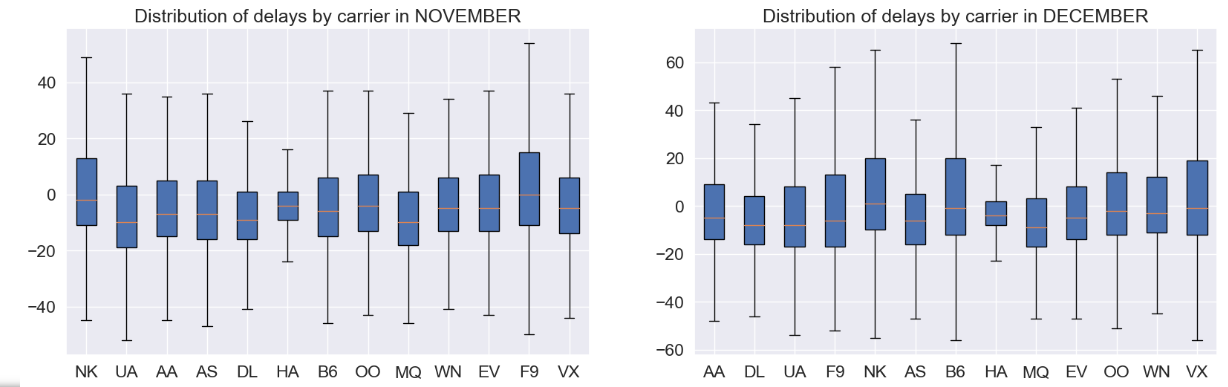




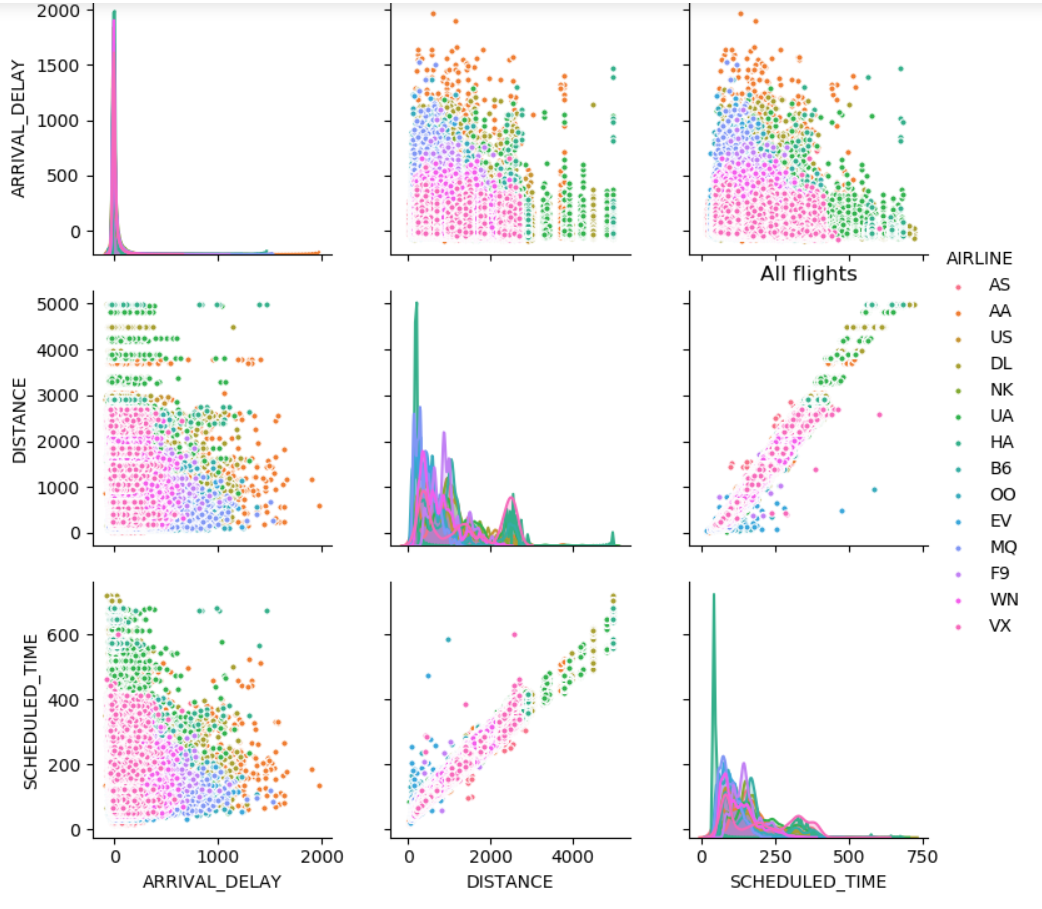


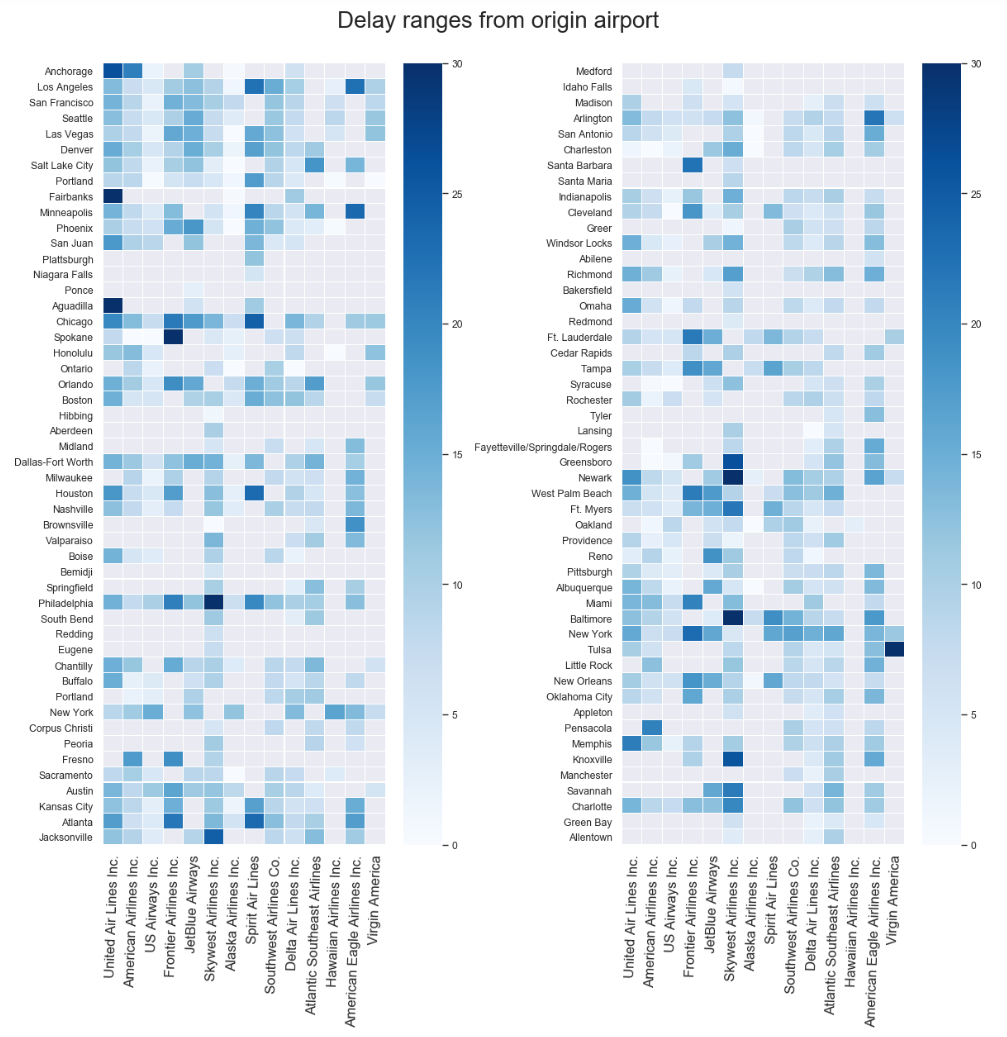






Compairing airline, arrival delay, distance and scheduled time using pairplot

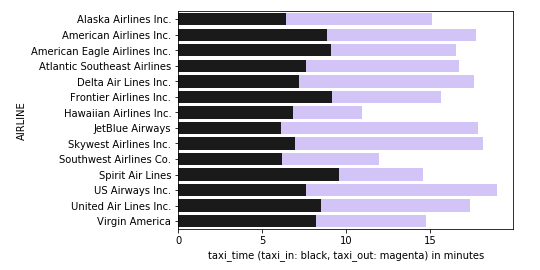




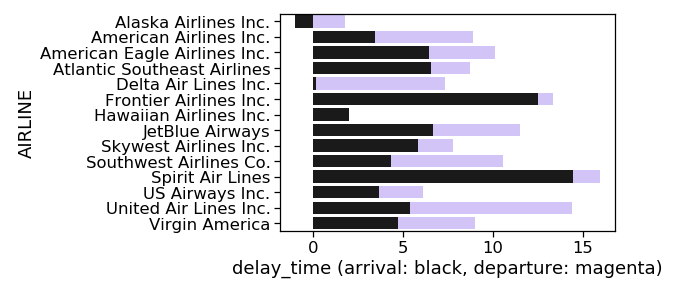
Calculating cancellation rate, diversion rate and flight percentage for each airline carrier.



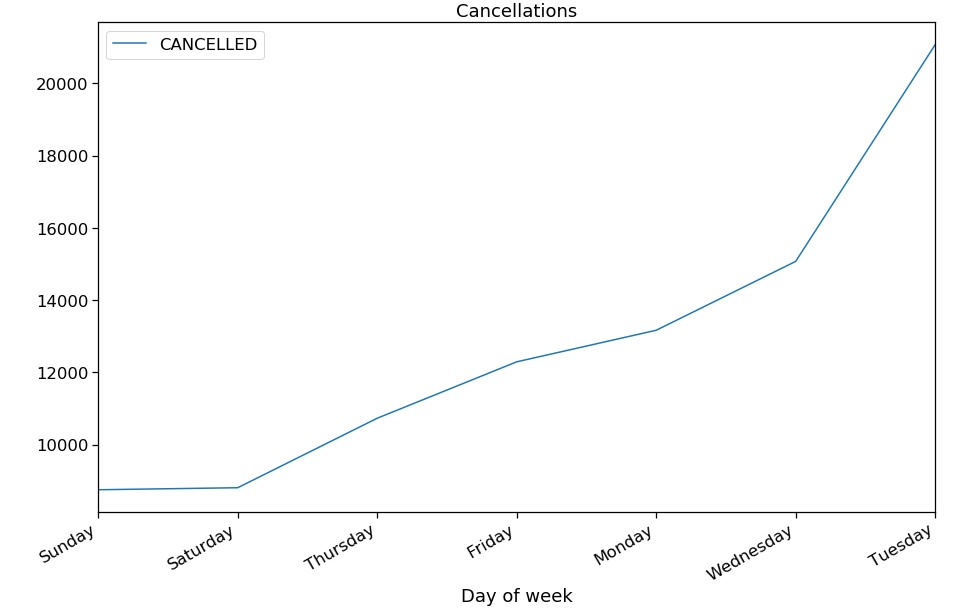
Taxi-in vs taxi-out in stacked bar chart for every airline



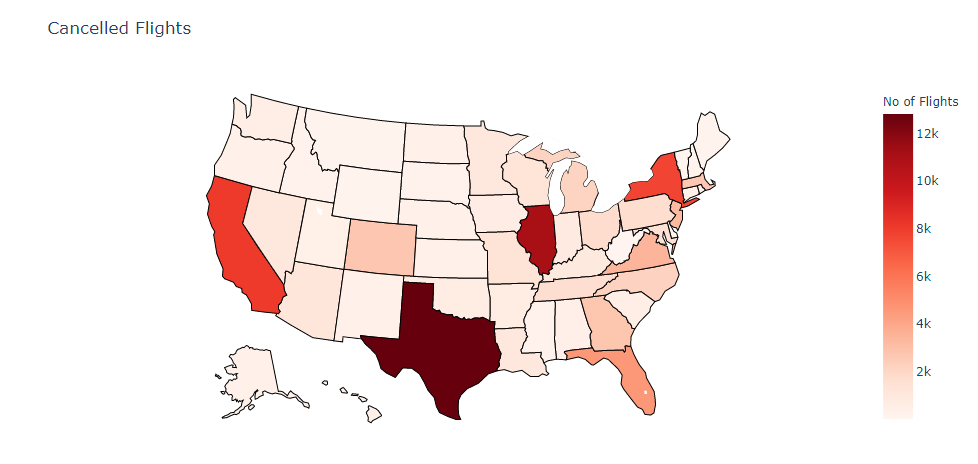
Arrival vs departure delay for every airline



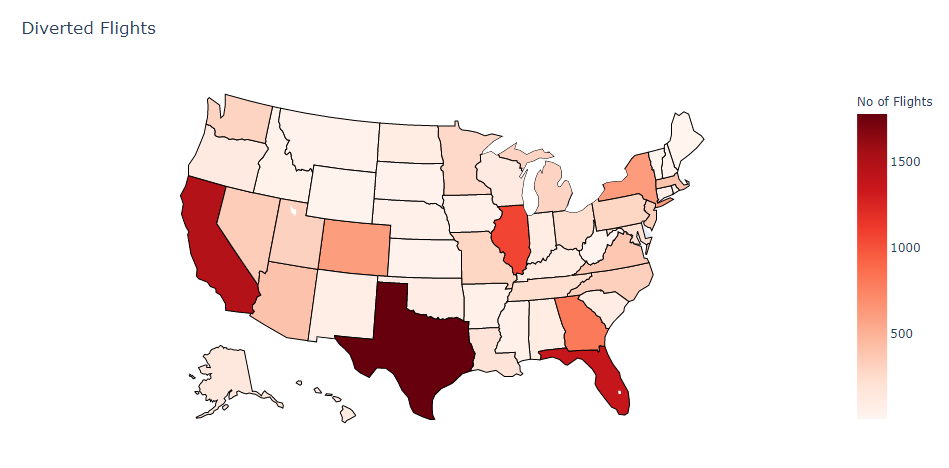
Week by cancellation



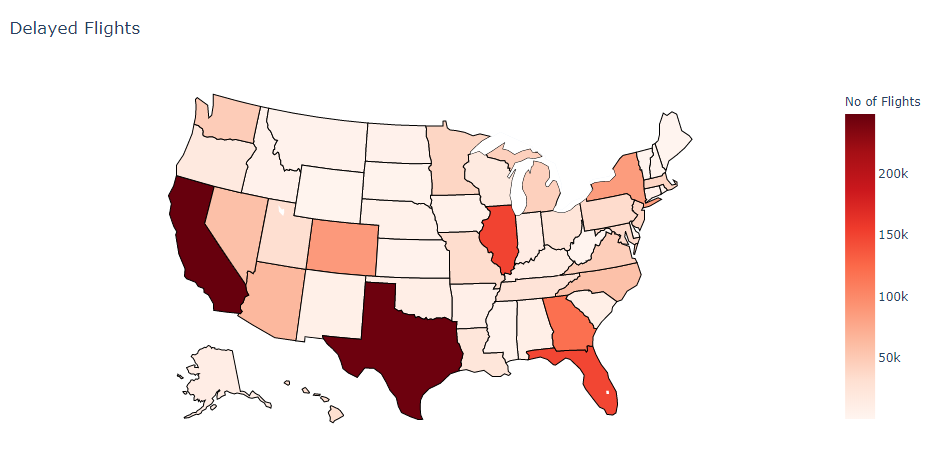
Choropleth state wise map visualization for cancelled flights



Choropleth state wise map visualization for diverted flights



Choropleth state wise map visualization for delayed flights



# References

FlightAware. (2019). *Miserymap*. Retrieved from flightaware: https://flightaware.com/miserymap/

Frontier, F. (2019). *Flight Deals*. Retrieved from Our best deals: https://www.flyfrontier.com/deals/flight-sales

Statistics, B. o. (2019, January). *Reporting Carrier On-Time Performance (1987-present)*. Retrieved from Transtats: https://www.transtats.bts.gov/DL\_SelectFields.asp

Worldwide, O. A. (2018, Jan). *OAG 2018 Punctuality report.* Retrieved from www.oag.com: https://www.oag.com/hubfs/Free\_Reports/Punctuality\_League/2018/PunctualityReport2018.pdf

Bureau of Transportation Statistics: <https://www.bts.gov/>

U.S. Department of Transportation: <https://www.transportation.gov/aviation>

Data.gov: <https://www.data.gov/>