Airlines and Airports
Punctuality Performance Analysis

Parag Siddamsettiwar
Data Science
Indiana University
psiddam@iu.edu

Abhishek Upadhayay
Data Science
Indiana University
abupad@iu.edu

Abstract

Flight data visual analysis of US air traffic to compare carriers and airports based on punctuality, delays, cancelations and diversions.

Terms

IATA – International Air Transport Association

BTS – Bureau of Transportation Statistics

DOT – Department of Transportation

1 Introduction

In today's busy world, time is a very precious commodity. Whether you are travelling for work or leisure, mode of transportation plays a very important role in your trip's success. Flights are the most convenient and fastest way to travel. In fact, frequent travelers prefer living in a city well connected by air transport. Airlines and airports are becoming critical necessity for every individual, as we are moving towards a more collaborative global economy.

1.1 Motivation

More and more people are taking flights and according to a report by IATA, around 8.2 billion people will be travelling by flights by 2037 (IATA, 2018). This trend will continue to grow due to growing economy and a need to save time by using the fastest mode of Transportation.

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 (Airlines for America, 2010).

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. The hassle to adjust the entire trip or adjust the bookings takes a toll on everyone including passengers, airlines and businesses. According to a research, an average Briton wastes 16 days of their holiday time waiting for their flight to take off (Andrew Hough, 2010).

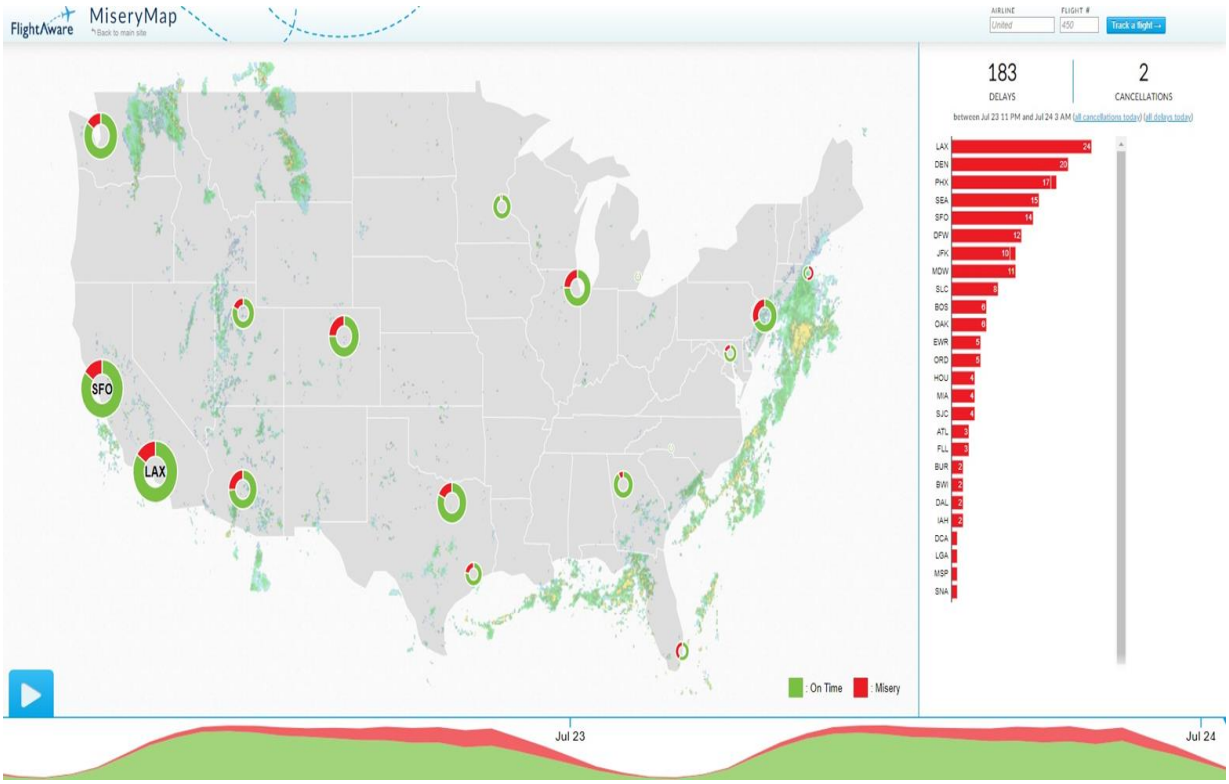
It will really help if we can analyze and visualize the flight data to get a better understanding of Airlines and Airports to make an informed choice before finalizing our travel plans

1.2 Existing work

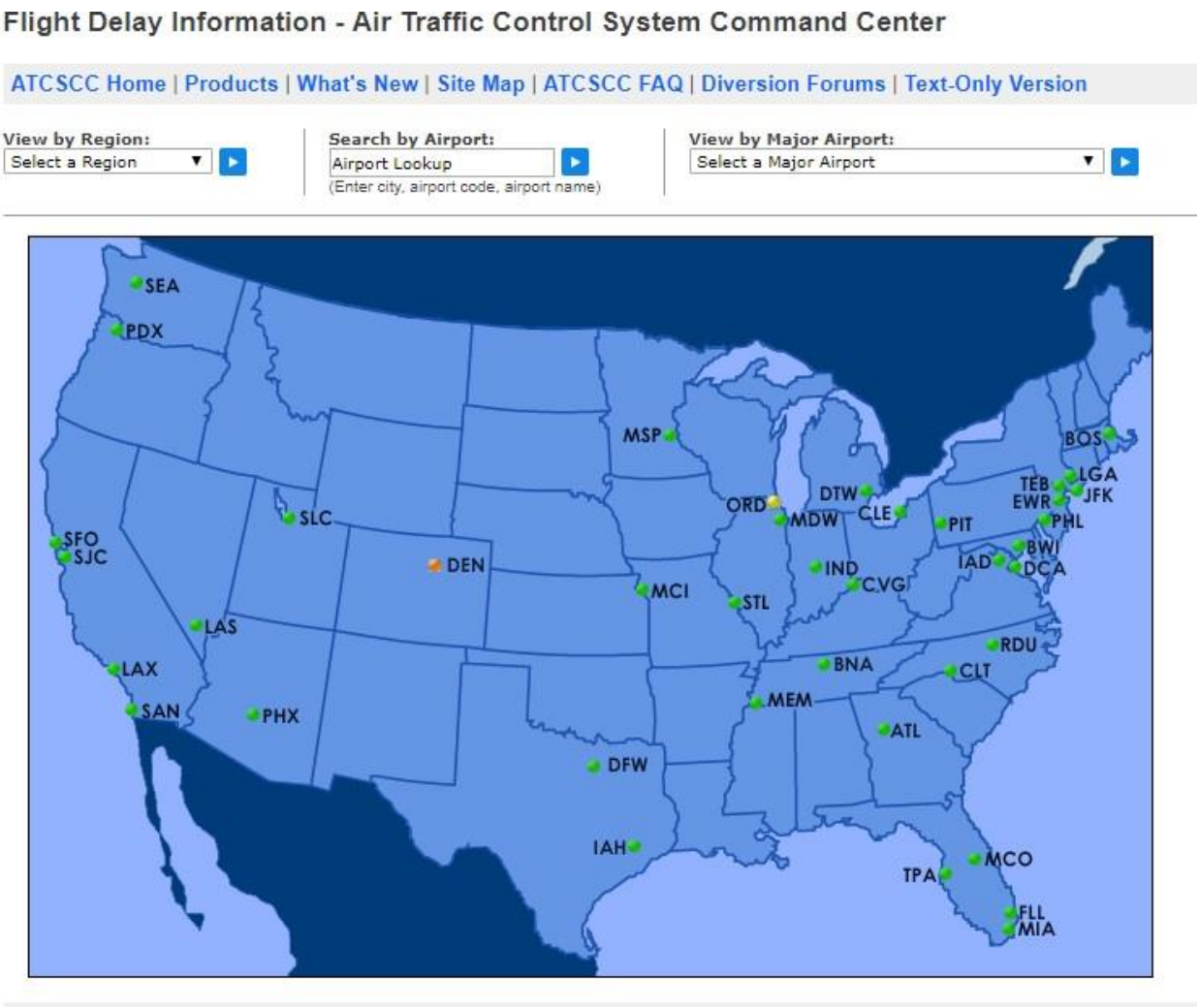
Even though there are tons of travel websites to book flights, there are only a handful websites that provide an analysis of punctuality of airlines and airports.

Even the ones that are available are either on current data (real time or past few hours) and not an analysis on historical data for a proper analysis of trends or behavior.

FlightAware's flight delay visualization (FlightAware, 2019)
This visualization shows some big Airports only, no information is available for Airlines and no option to choose other airports e.g. Indianapolis, Cincinnati etc. More importantly, this is based on current data (last few hours) and not an analysis on a historical dataset.



Flight Delay Information - Air Traffic Control System Command Center (FAA, 2019)
This visualization also shows airports and it's also based on current data (last few hours) and not an analysis on a historical dataset.
Its interactive and lets user select airport or flights and provides current situation for that.



There is not enough detail available that can help us in making decision for future travels from airports and airlines based on past trends and airlines behavior.

Even though a flight may show on time on these current data visualizations, there is a possibility that the flights of an airline are regularly late due to unavailability of aircrafts or pilots. We may want to avoid taking that Airline

1.3 Contribution and objectives

Our main objective is to analyze past years data to compare various airlines and airports based on punctuality.

We want to Create a series of visualizations in order to help a traveler find best choices of carrier and airports in terms of punctuality and certainty.

The visualizations that we will create, should be easy to understand and pleasing to eyes.

2 Dataset

Each country has an authority that controls airline traffic for them. Usually these authorities host the flight data including delay information.

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

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

We have downloaded on time performance data of all the US domestic flights. Such airline dataset is generally associated to IATA_CODE which is a unique code given to every airline and airport. It is used across databases for easy cross-reference.

Let us look at main data constituents:

Airlines

1	IATA_CODE	AIRLINE
2	UA	United Air Lines Inc.
3	AA	American Airlines Inc.
4	US	US Airways Inc.
5	F9	Frontier Airlines Inc.
6	B6	JetBlue Airways
7	OO	Skywest Airlines Inc.
8	AS	Alaska Airlines Inc.
9	NK	Spirit Air Lines
10	WN	Southwest Airlines Co.
11	DL	Delta Air Lines Inc.

Airports

	A	B	C	D	E	F	G
1	IATA_COD	AIRPORT	CITY	STATE	COUNTRY	LATITUDE	LONGITUDE
2	ABE	Lehigh Vall	Allentown	PA	USA	40.65236	-75.4404
3	ABI	Abilene Re	Abilene	TX	USA	32.41132	-99.6819
4	ABQ	Albuquerq	Albuquerq	NM	USA	35.04022	-106.609
5	ABR	Aberdeen	Aberdeen	SD	USA	45.44906	-98.4218
6	ABY	Southwest	Albany	GA	USA	31.53552	-84.1945
7	ACK	Nantucket	Nantucket	MA	USA	41.25305	-70.0602
8	ACT	Waco Regi	Waco	TX	USA	31.61129	-97.2305
9	ACV	Arcata Air	Arcata/Eu	CA	USA	40.97812	-124.109
10	ACY	Atlantic Ci	Atlantic Ci	NJ	USA	39.45758	-74.5772
11	ADK	Adak Airpo	Adak	AK	USA	51.87796	-176.646
12	ADQ	Kodiak Air	Kodiak	AK	USA	57.74997	-152.494
13	AEX	Alexandria	Alexandria	LA	USA	31.32737	-92.5486
14	AGS	Augusta Re	Augusta	GA	USA	33.36996	-81.9645
15	AKN	King Salmo	King Salmo	AK	USA	58.6768	-156.649

***On-time performance Dataset of USA domestic flights*. (Bureau of Transportation Statistics, 2013)**

This dataset is regularly updated and published; it 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.

Below are all the dataset columns downloaded by us:

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 columns can be found on the TranStats website (Bureau of Transportation Statistics, 2013), e.g. below screenshots:

Summaries	
*OntimeArrivalPct	Percent of flights that arrive on time. For percent of on time arrivals at specific airports, click Analysis. **Note**: If you select Origin as a category, you get percent of flights that depart from those airports and arrive on time.
*OntimeDeparturePct	Percent of flights that depart on time. For percent of on time departures at specific airports, click Analysis. **Note**: If you select Dest as a category, you get percent of flights that depart on time and arrive at those airports.
Time Period	
Year	Year
Quarter	Quarter (1-4)
Month	Month
DayofMonth	Day of Month
DayOfWeek	Day of Week
FlightDate	Flight Date (yyyymmdd)
Airline	
Reporting_Airline	Unique Carrier Code. When the same code has been used by multiple carriers, a numeric suffix is used for earlier users, for example, PA, PA(1), PA(2). Use this field for analysis across a range of years.
DOT_ID_Reporting_Airline	An identification number assigned by US DOT to identify a unique airline (carrier). A unique airline (carrier) is defined as one holding and reporting under the same DOT certificate regardless of its Code, Name, or holding company/corporation.
IATA_CODE_Reporting_Airline	Code assigned by IATA and commonly used to identify a carrier. As the same code may have been assigned to different carriers over time, the code is not always unique. For analysis, use the Unique Carrier Code.
Tail_Number	Tail Number
Flight_Number_Reporting_Airline	Flight Number

Flight Summaries	
CRSElapsedTime	CRS Elapsed Time of Flight, in Minutes
ActualElapsedTime	Elapsed Time of Flight, in Minutes
AirTime	Flight Time, in Minutes
Flights	Number of Flights
Distance	Distance between airports (miles)
DistanceGroup	Distance Intervals, every 250 Miles, for Flight Segment
Cause of Delay (Data starts 6/2003)	
CarrierDelay	Carrier Delay, in Minutes
WeatherDelay	Weather Delay, in Minutes
NASDelay	National Air System Delay, in Minutes
SecurityDelay	Security Delay, in Minutes
LateAircraftDelay	Late Aircraft Delay, in Minutes
Gate Return Information at Origin Airport (Data starts 10/2008)	
FirstDepTime	First Gate Departure Time at Origin Airport
TotalAddGTime	Total Ground Time Away from Gate for Gate Return or Cancelled Flight
LongestAddGTime	Longest Time Away from Gate for Gate Return or Cancelled Flight

3 Process and Results

After carefully studying the data columns available to us, we decided to download the ones that will be useful in our analysis.

We downloaded 10 years of data and tried to load it into python for exploratory analysis. Obviously, this data was very large and we have limited memory available to us on our local computers, we noticed extremely slow responsive behavior of python program. It was taking a lot of time to load data and process it.

Our main focus was to creat visualization and not how much data we can bring in, so we decided to restrict our dataset to 1 year. After using only 1 year of data, we were able to load and process it successfully.

Winter months generally have lots of disruption in air traffic due to unusual weather, so we decided to focus on the month of January.

Top 20 delayed flights

Since our focus is on delays, as a starting point begin by getting a list of flights with highest delays.

ARR_DELAY column shows delay in minutes. Applying SORT_VALUES method on this column gave the output we wanted. For display purpose, we restricted the results to top 20 delayed flights.

	FL_DATE	OP_UNIQUE_CARRIER	ARR_DELAY	ORIGIN_CITY_NAME	DEST_CITY_NAME	CARRIER_DELAY	WEATHER_DELAY
472366	1/11/2018	AA	2023.0	Eagle, CO	New York, NY	2007.0	0.0
483314	1/2/2018	AA	1778.0	Norfolk, VA	Dallas/Fort Worth, TX	1752.0	0.0
141684	1/3/2018	HA	1717.0	Honolulu, HI	New York, NY	35.0	1682.0
498049	1/3/2018	AA	1648.0	Salt Lake City, UT	Dallas/Fort Worth, TX	0.0	0.0
468392	1/9/2018	AA	1537.0	Kona, HI	Los Angeles, CA	1531.0	0.0
479862	1/9/2018	AA	1515.0	Washington, DC	Dallas/Fort Worth, TX	1515.0	0.0
485341	1/11/2018	AA	1510.0	Dayton, OH	Dallas/Fort Worth, TX	1510.0	0.0
389147	1/4/2018	YV	1486.0	Fresno, CA	Phoenix, AZ	1469.0	0.0
83493	1/6/2018	B6	1473.0	Aguadilla, PR	New York, NY	1125.0	0.0
104363	1/8/2018	EV	1454.0	Roswell, NM	Dallas/Fort Worth, TX	1454.0	0.0
210032	1/3/2018	OO	1449.0	Meridian, MS	Chicago, IL	2.0	0.0
225663	1/7/2018	OO	1433.0	Salt Lake City, UT	Chicago, IL	2.0	0.0
221370	1/20/2018	OO	1432.0	Burbank, CA	Salt Lake City, UT	1420.0	0.0
210031	1/3/2018	OO	1431.0	Dallas/Fort Worth, TX	Hattiesburg/Laurel, MS	1416.0	0.0
31082	1/8/2018	UA	1429.0	Lihue, HI	San Francisco, CA	39.0	0.0
210033	1/3/2018	OO	1422.0	Hattiesburg/Laurel, MS	Meridian, MS	1226.0	0.0
253161	1/12/2018	OO	1390.0	Birmingham, AL	Chicago, IL	1390.0	0.0
233957	1/22/2018	OO	1383.0	Aspen, CO	Denver, CO	197.0	0.0
259079	1/15/2018	OO	1380.0	Durango, CO	Denver, CO	1361.0	0.0
254010	1/13/2018	OO	1378.0	Grand Junction, CO	Phoenix, AZ	1378.0	0.0

The most delayed flight was of American Airlines, which was more than 2000 minutes delayed, almost 1.5 days. That made us to check how many overall flights were delayed in January month irrespective of carrier.

Before that some calculations:

In [8]:
```
# now check the percentage of flights delayed in Jan 2018

not_delayed = flights['DELAYED'].value_counts()[0] # first value of the result above
delayed = flights['DELAYED'].value_counts()[1] # second value of the result above
total_flights = not_delayed + delayed # total count of flights
float(delayed) / total_flights
```

Out[8]: 0.3147348443655524

More than 31 % of flights were delayed in January of 2018. That is a big chunk of flights being delayed.

So how about cancellations?

```
# now chec how many were cancelledm using cancled column id dataframe with values 0 and 1

not_canceled, canceled = flights['CANCELLED'].value_counts()
float(canceled) / len(flights)
```

Flight cancellation in January is not as scary as delays. But still a significant number, around 3 % of flights were cancelled.

On time vs delayed ratio of flights per airline

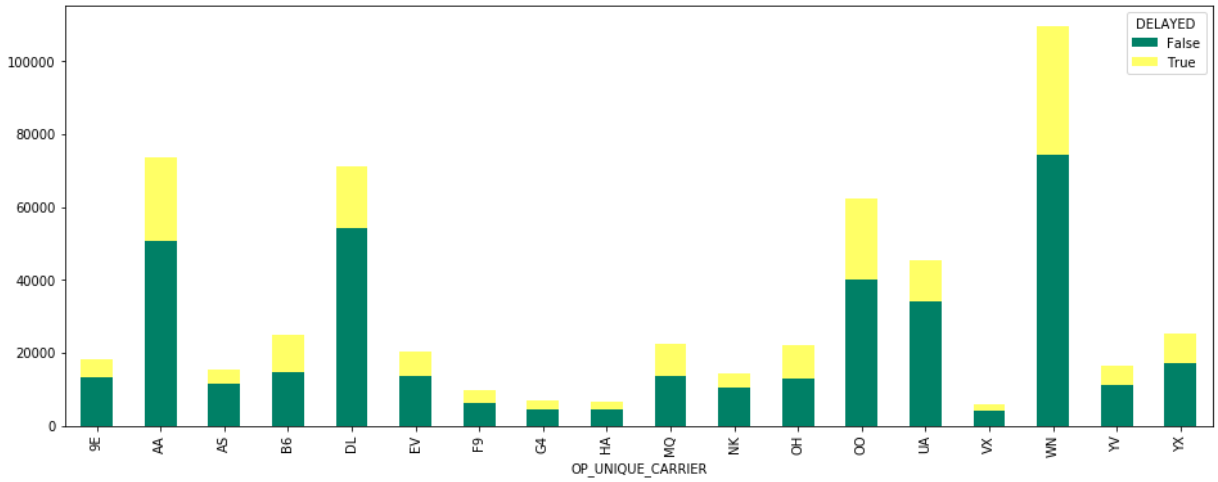
Next, we were interested in understanding number of flights operated by each carrier in January. Also, we want to compare total number of flights for each carrier by ‘delayed’ and ‘on-time’.

So, we decided to create a bar chart.

As we can see, above chart is not showing any delay information it is just displaying the total number of flights. The x axis is also not informative, the numbers don’t make any sense.

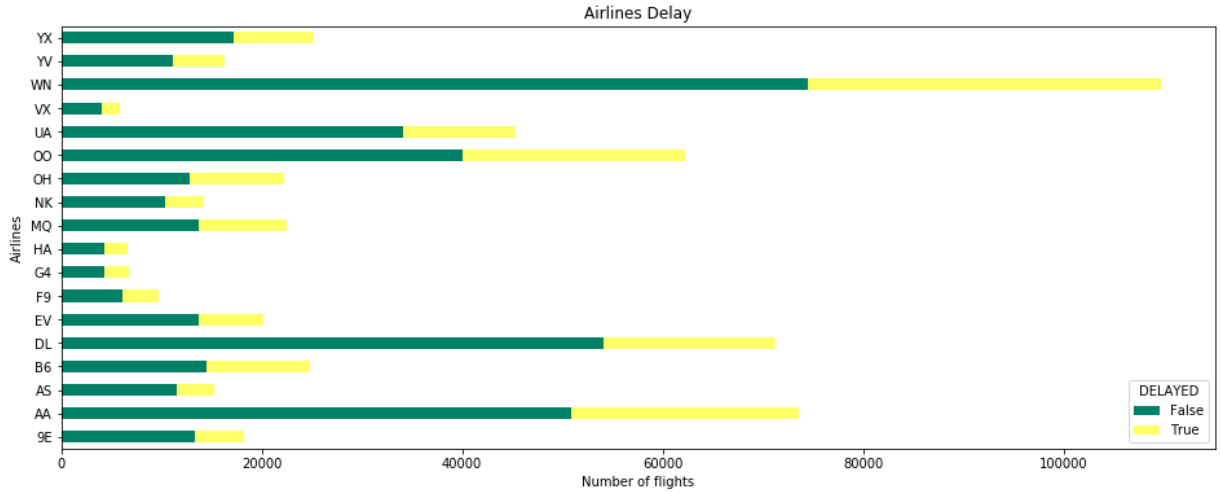
How about stacked bar chart, which shows delay in a color?

The first version:



Even this doesn’t look good due to unpleasant x-axis.

How about flipping the axis, and make it a horizontal bar chart?



Much better, now we know which airlines are the most delayed, but it's still based on absolute numbers. The flights with the greatest number of flights will have longer bar and this can be misleading.

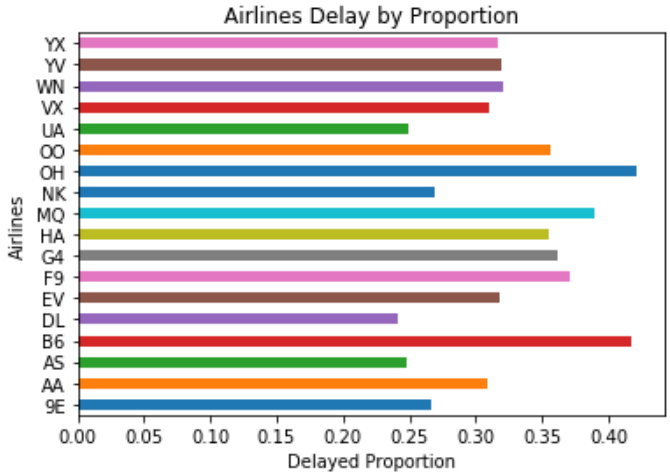
To avoid this, we should check the proportions instead of absolute numbers.

Airline carrier delays proportion

We had to create a new column to visualize the proportion of delayed flights across airlines.

```
# create a proportion coloumn
delayed_by_carrier['PROPORTION_DELAYED'] = delayed_by_carrier[True] / delayed_by_carrier['FLIGHTS_COUNT']
delayed_by_carrier[:100]
```

```
ax = delayed_by_carrier.plot(x='OP_UNIQUE_CARRIER', y='PROPORTION_DELAYED', kind = 'barh')
plt.ylabel('Airlines')
plt.xlabel('Delayed Proportion')
plt.title('Airlines Delay by Proportion')
ax.legend().set_visible(False)
```



Clearly the ones that looked like they had more delays in previous visualization are actually better when compared in proportions. E.g. Southwest (WN) and Delta (DL).

Now we can make some sense out of it.
Delta (DL) seems to fare much better as only 28% of flights had some delay as compared to 43% of PSA Airlines (OH) and JetBlue (B6).

Still this visualization is based on just the flight numbers that got delayed, even if its 1 min.

We can do much better if we take minutes of delays in consideration to assess the severity of delay.

For that, let us pick one airline for a detailed analysis, we picked Southwest as it is our most preferred airline.

Southwest

As a starting point it is always good to understand the distribution, so we used DESCRIBE function to get all the important statistics.

```
1 southwest = flights[flights['OP_UNIQUE_CARRIER'] == 'WN']['ARR_DELAY']
2 southwest.describe()
```

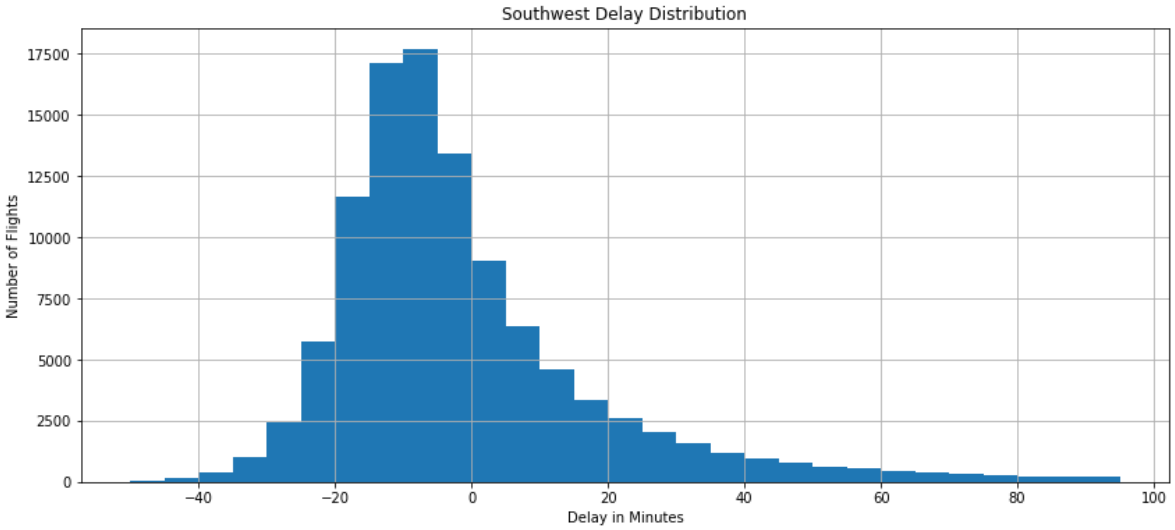
```
count    107023.000000
mean          0.282154
std          27.337047
min         -73.000000
25%         -14.000000
50%          -6.000000
75%           6.000000
max         550.000000
Name: ARR_DELAY, dtype: float64
```

Average delay is just 0.28 minutes, while longest was 550 minutes. Interesting insights here is that IQR (Inter quartile range) for Southwest airlines delay is actually negative. That means most of the time southwest airlines arrived before time.

Let us verify this by checking its distribution in a histogram.

Since some delays are in negative (DESCRIBE helped), we need to make some adjustments for bin sizes and values. From above statistics, we can see that most of the data should be captured between -50 and 100 minutes delay, so we used that for bin values.

```
1 bin_values = np.arange(start=-50, stop=100, step=5)
2 southwest.hist(bins=bin_values, figsize=[14,6])
3 plt.xlabel('Delay in Minutes')
4 plt.ylabel('Number of Flights')
5 plt.title('Southwest Delay Distribution')
```

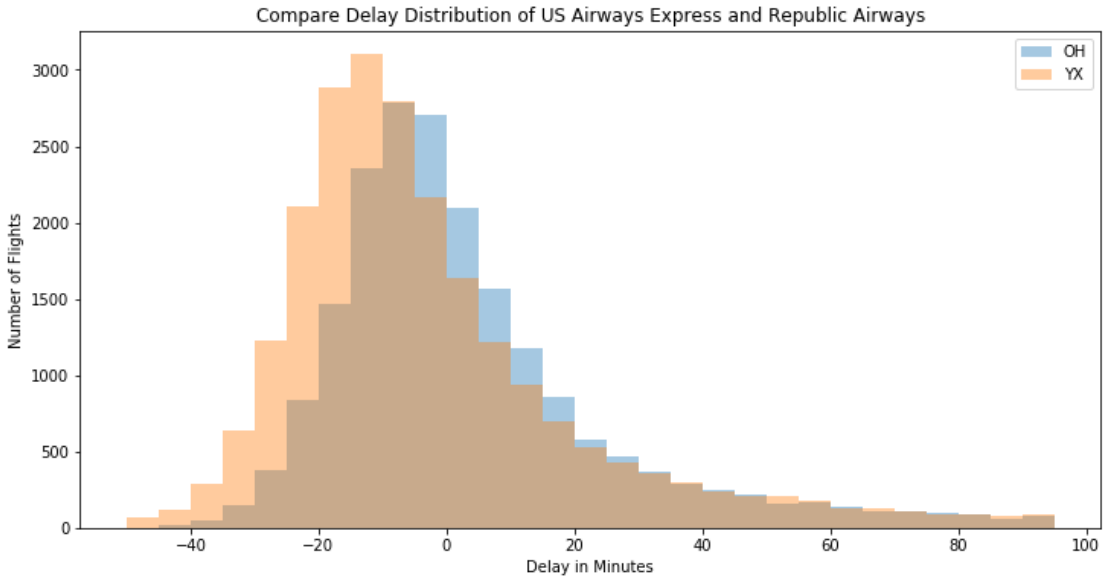


Looks like southwest is mostly before time or on-time even during January's stormy winters. So, its not a bad airline to book your travel.

Comparison

Now compare two airlines histograms to verify the severity concept.

We have to pick airlines with similar number of total flights to compare properly. We picked US Airways Express 'OH' and Republic Airways 'YX'



The darker area is where the distribution overlaps. The distribution looks similar, but they are not the same as US Airways Express has more shift to the right because of more delays.

This was evident from the earlier analysis of proportion of delays/total as well. But histogram is confirming that the severity of delays is also higher.

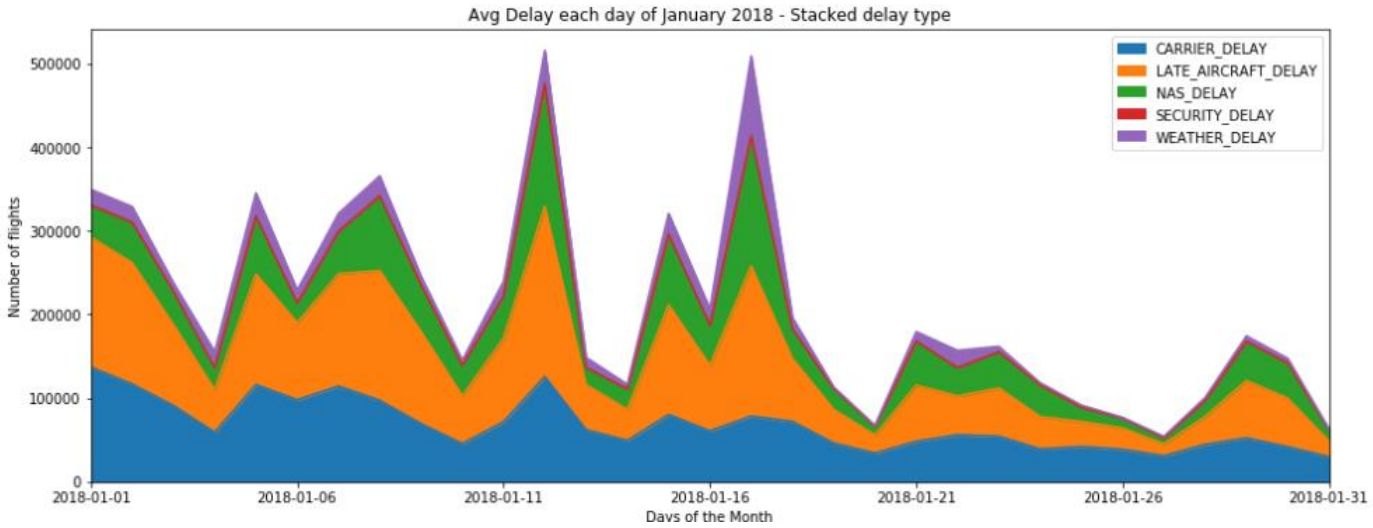
Flight delay reason stacked area chart

Now we have good insight about delays in January, but we should also look at reasons of delays if it provides a new insight.

Our dataset already had 5 types of delays. Carrier delay, weather delay, late aircraft delay, NAS (National Air System) delay and Security delay.

We thought of comparing different types of delays to see the contribution of each to total delay.

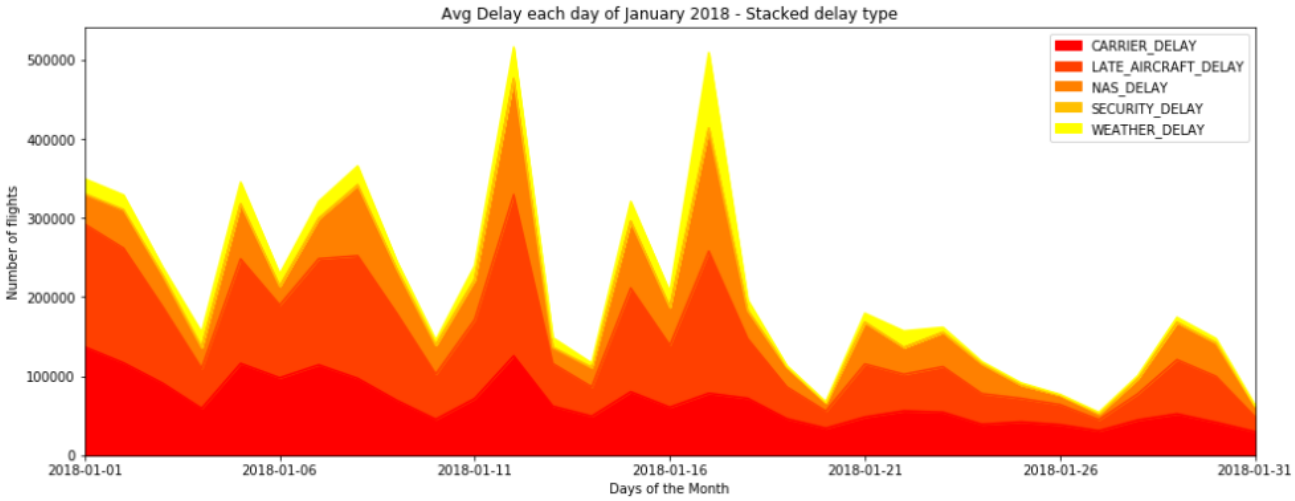
So, we tried stacked area chart.



As you can see it's a great visualization, showing total number of flights delayed with delay reasons stacked on top of each other for the month of January.

This can be more appealing if we can color code it differently. We decided to use sequential color map. 'autumn' color map was a good option as it had an effect of heat map on the area with larger share shown with a darker color.

We can clearly see that the main reasons for delay are Carrier delay and Late Aircraft delay.

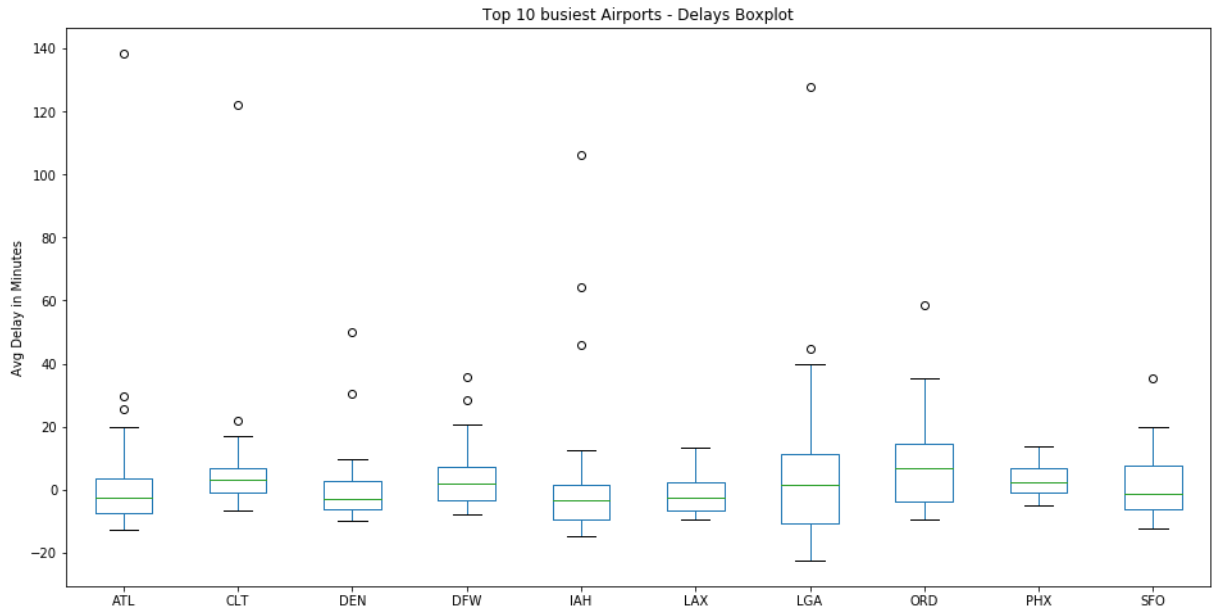


Apart from carriers, can we do some exploratory analysis on airports delays?

Top 10 Busiest airports and their delays

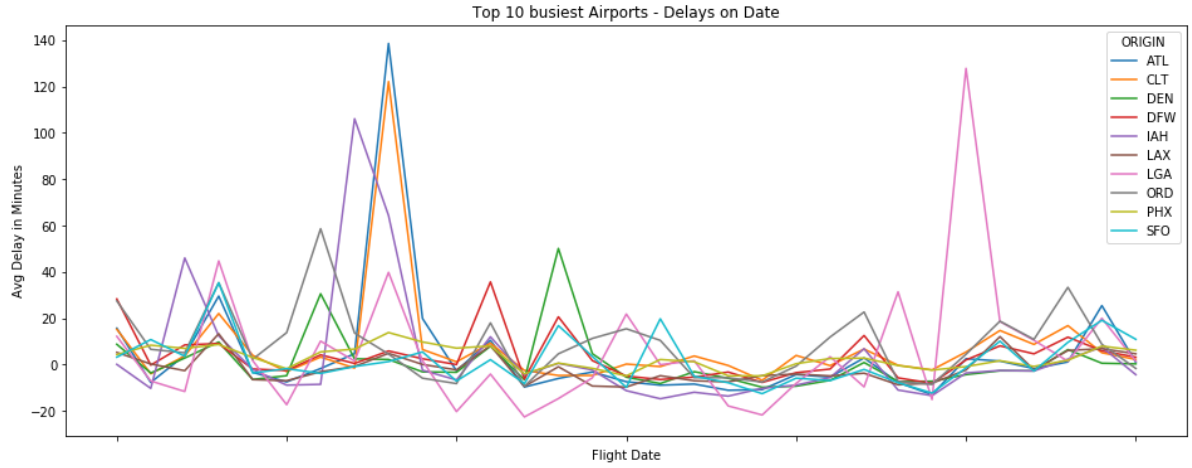
For a start, let us find top busiest 10 airports and analyze where we can avoid going for a vacation in January (based on delays)

For such comparison we can use box plots very effectively as it can help us compare airports side by side. It also helps understand the distribution by avoiding outliers.



We can see clear dispersion in data at LaGuardia NYC and Ohare' Chicago. While other airports delays are evenly distributed around 0. Based on 2018 data, we can say that try to avoid these two airports during January.

Let us try to visualize the same using line chart instead of boxplots but date on x axis.

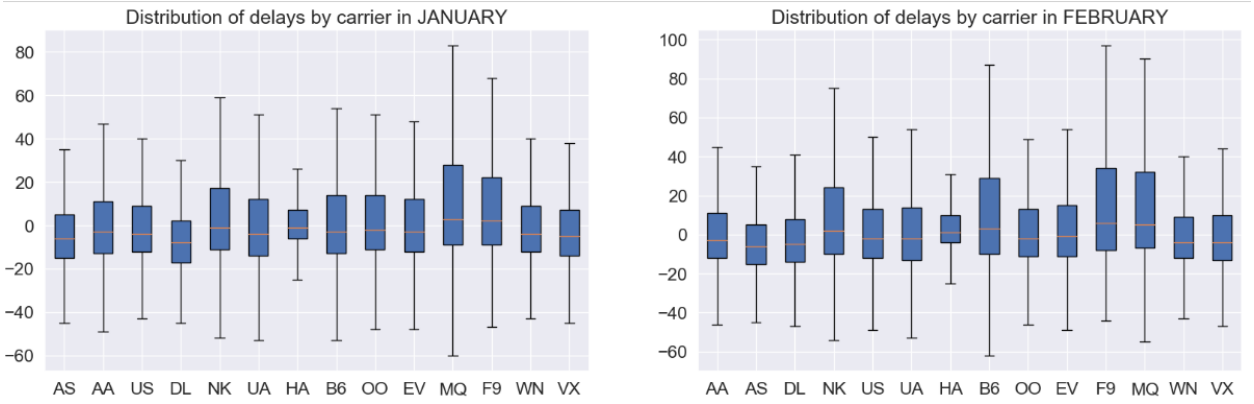


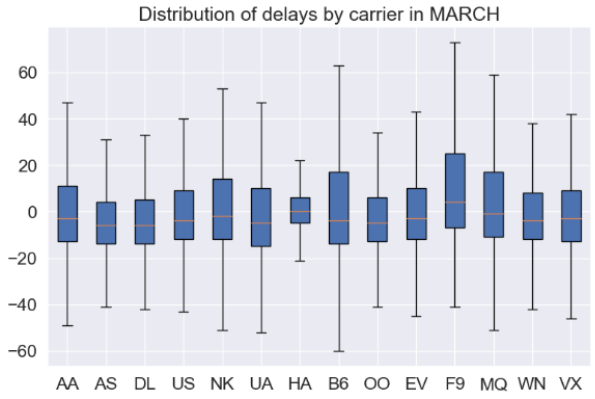
Clearly, some days have spikes for most of the carriers suggesting weather related delays on those dates. Suggesting delays at one large airport has a chain reaction effect on other airport delays.

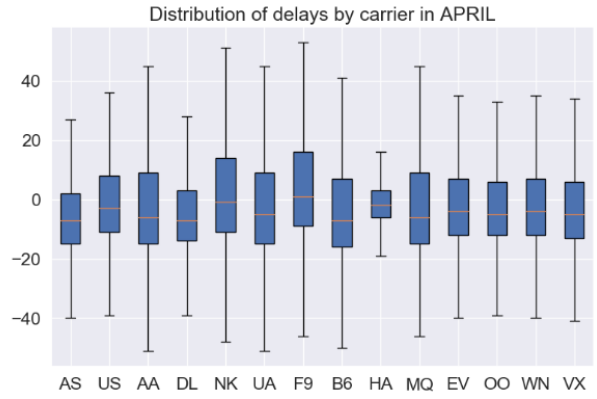
It was time we look at complete year data and try to find some interesting insights into flight data and delays.

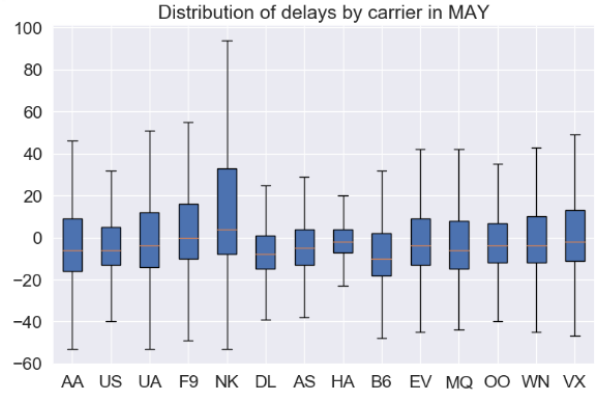
Distribution of delays by month

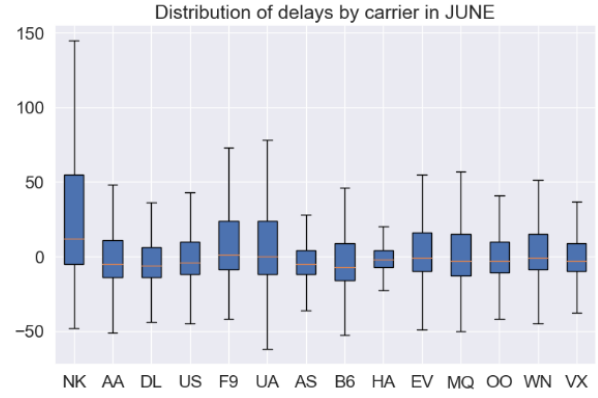
As we are looking at all 12 months data, first let us compare all the airlines and their delays in each month. We used sub-plots for this.

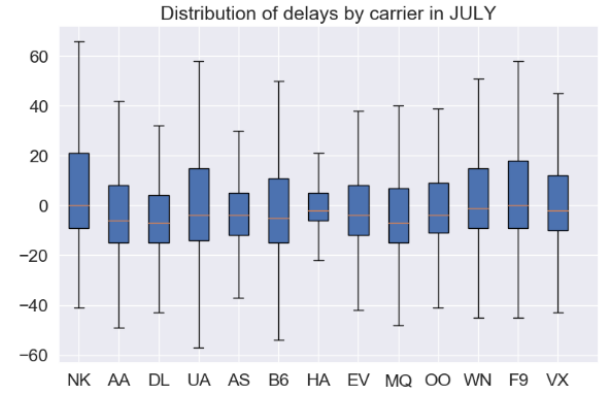


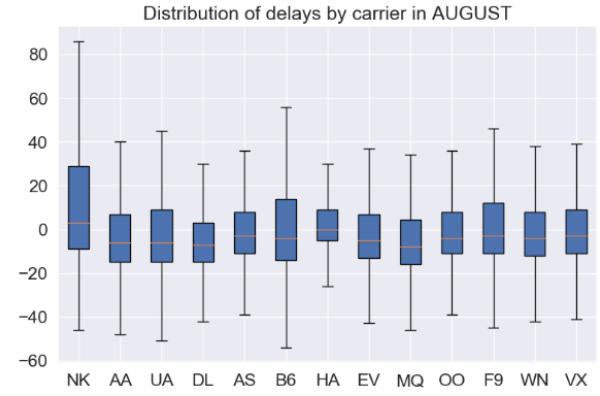

Distribution of delays by carrier in MARCH
AA AS DL US NK UA HA B6 OO EV F9 MQ WN VX

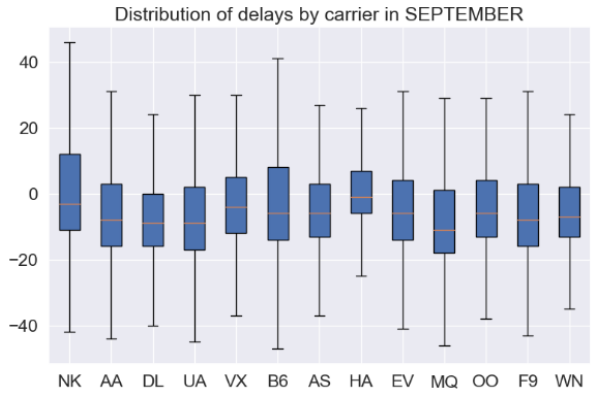

Distribution of delays by carrier in APRIL
AS US AA DL NK UA F9 B6 HA MQ EV OO WN VX

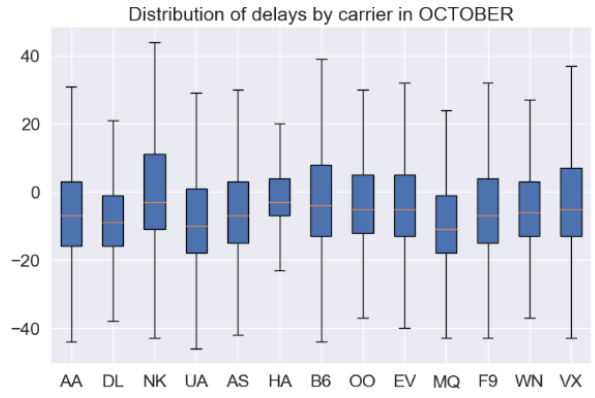

Distribution of delays by carrier in MAY
AA US UA F9 NK DL AS HA B6 EV MQ OO WN VX

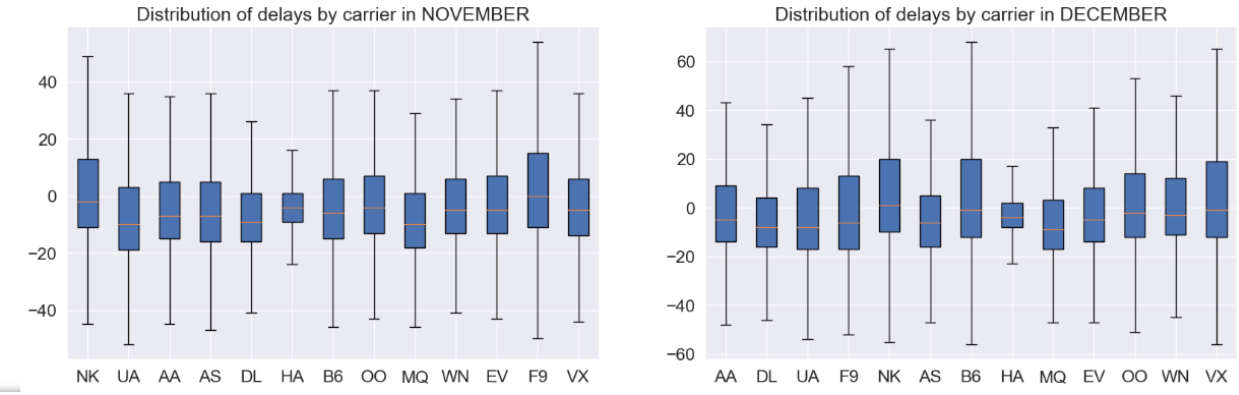

Distribution of delays by carrier in JUNE
NK AA DL US F9 UA AS B6 HA EV MQ OO WN VX


Distribution of delays by carrier in JULY
NK AA DL UA AS B6 HA EV MQ OO WN F9 VX


Distribution of delays by carrier in AUGUST
NK AA UA DL AS B6 HA EV MQ OO F9 WN VX


Distribution of delays by carrier in SEPTEMBER
NK AA DL UA VX B6 AS HA EV MQ OO F9 WN


Distribution of delays by carrier in OCTOBER
AA DL NK UA AS HA B6 OO EV MQ F9 WN VX

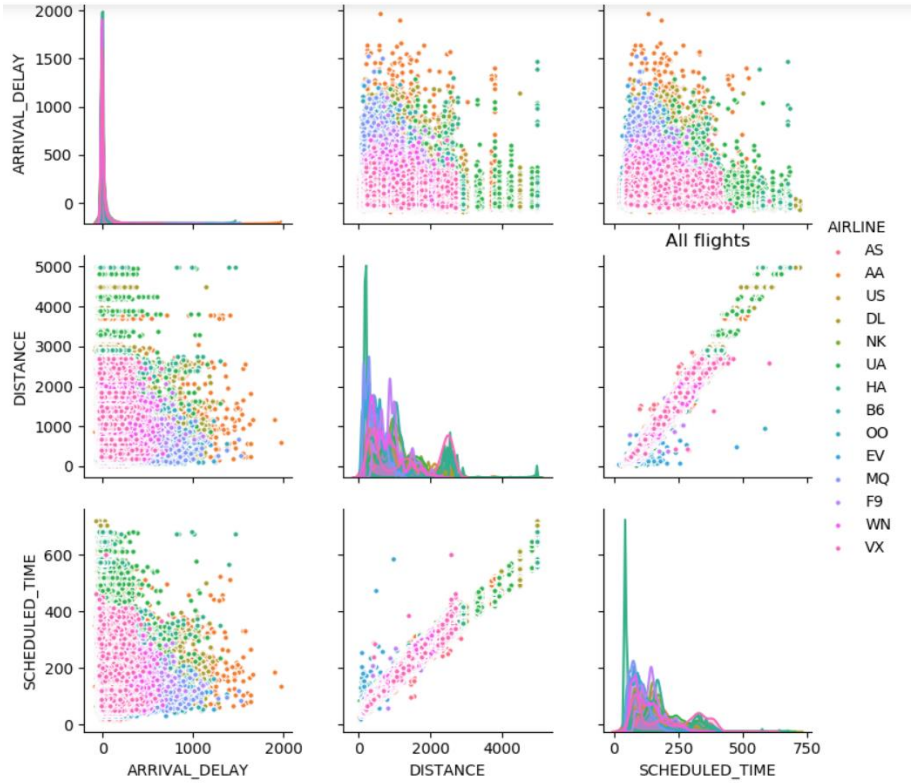


There is good information that can be extracted by studying each month and comparing each airline carrier. Still this visualization is not user friendly. It has lot of information and it will take even an expert a significant amount of time to derive any insight.

Comparing airline, arrival delay, distance and scheduled time using pair-plot

Moving on with the exploratory analysis for whole year data. Let us try pair-plot and see if we can derive any useful insights.

```
plt.style.use('default')
sns.pairplot(flights.loc[:,['ARRIVAL_DELAY', 'DISTANCE', 'SCHEDULED_TIME', 'AIRLINE']], hue='AIRLINE', plot_kws={'s':14})
plt.title('All flights')
```



This one is also impacted by huge amount of data, so not much can be derived out of it. Also, our color coding is not helping to distinguish the airlines.

But we still got a good information out of it. The Distance vs Arrival Delay graph is suggesting that as the distance increases, the number of flights with large delays decreases. It may be due to the long-haul flights covering the delay on their way or airlines may simply cancel significantly delayed long haul flights.

Choropleth state wise maps

Is there a way we can visualize cancellation, diversion and delay data by every state of USA?

We have airport information like city, state, latitude and longitude available in our dataset.

So, let's try to use the number of cancellation of flights for all the airports for a particular state and try to visualize that using map. We attempted it by referring the visualization shown on (Altair, 2019).

The biggest problem we ran into while trying this is the state coding done in the vega_dataset

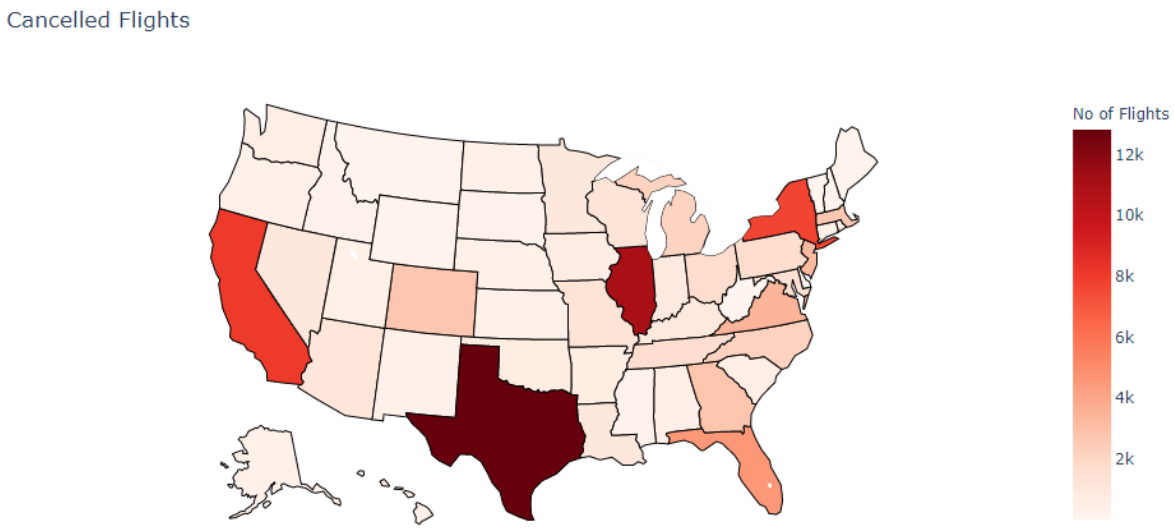
```
import altair as alt
from vega_datasets import data

airports = data.airports.url
states = alt.topo_feature(data.us_10m.url, feature='states')
```

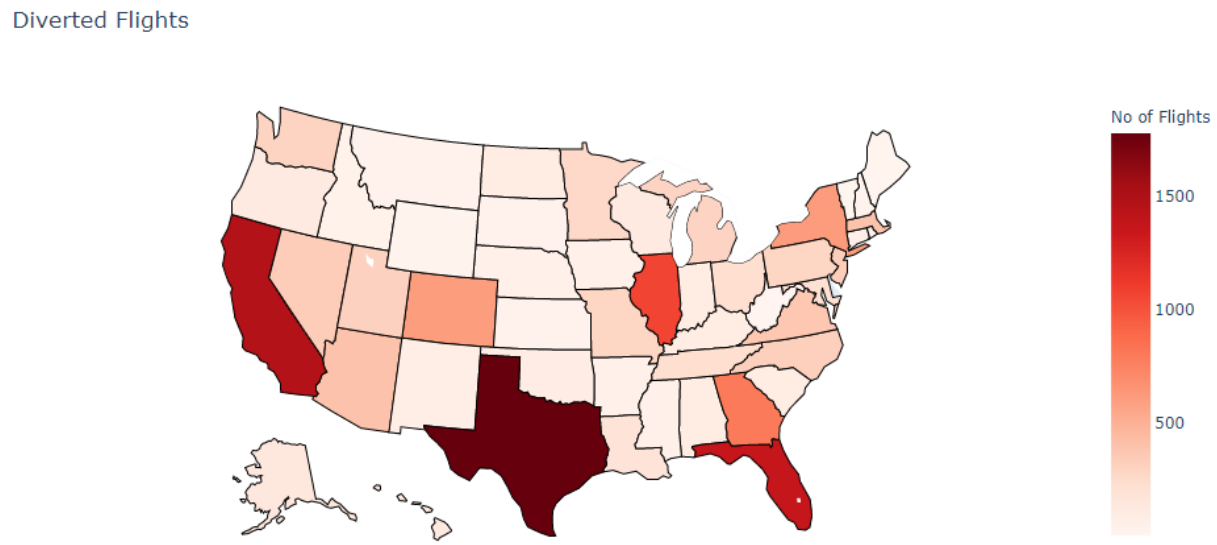
We were not able to map the state ID returning from states dataset and match them with state codes like TX, IN, CA that is present in our data set.

We tried for several days but we were not successful. Finally, we dumped idea of using Altair for this purpose and started looking into another python library which can help do this fairly quickly. We were able to plot Choropleth by referring to (Plotly, n.d.)

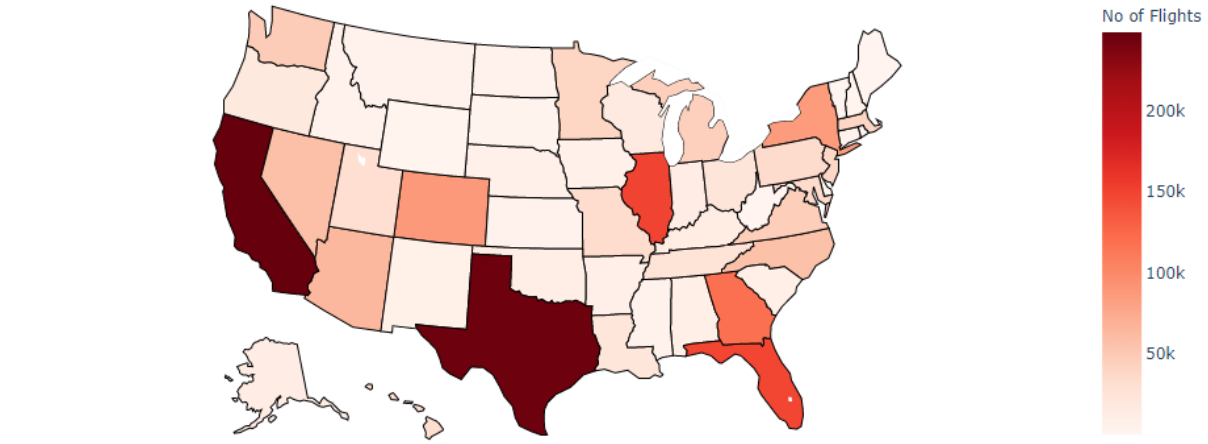
Cancelled flights



Diverted flights



Delayed flights



Even though we have some good information shown in these maps, its still misleading as it is based on absolute number of flights.

States like Texas, California and New York with large number of flights (even if they have better proportion) will be displayed in darker color in these maps.

Failed attempt at maps with longitude and latitude

We also had longitude and latitude information, we tried plotting airport delay in form of a heat map.

We failed miserably as airports have fixed coordinates and when this huge amount of data was plotted, all hundreds of dots plotted on top of each other at the airport locations. It only resulted in showing dots for all the airports. We learned our lesson that this dataset was not good for coordinate plotting in a map.

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

Since our state map based on absolute numbers was not good enough, we used percentages.

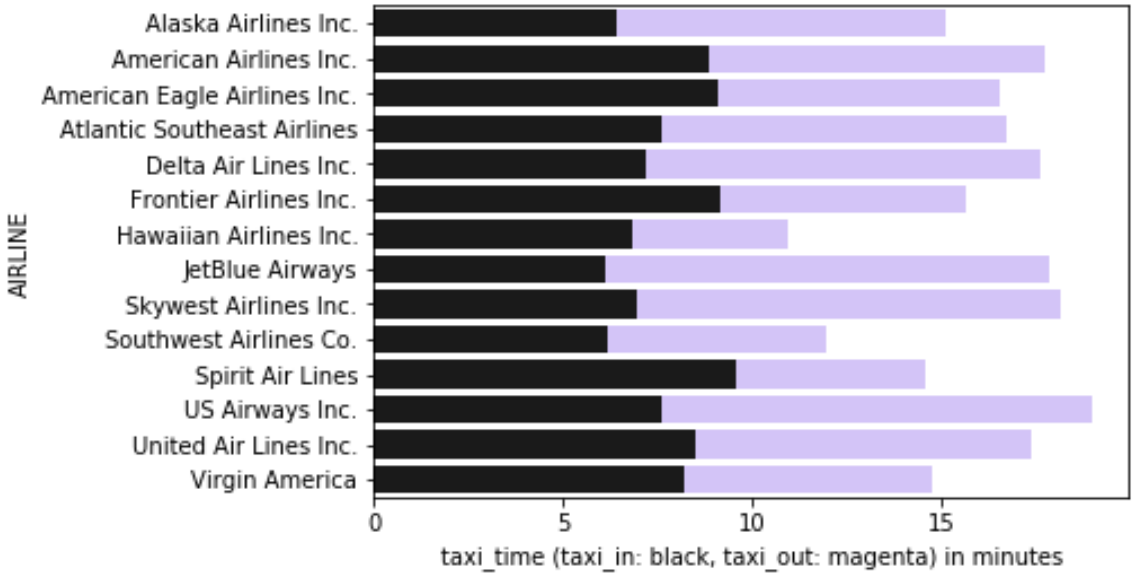
We did some calculation on the data to derive cancellation and diversion rates and calculated percentage of flights for each airline across all the flights from the entire year of data.

	AIRLINE	Flight Volume	Flight Percentage	Cancellation Percentage	Diversion Percentage
0	Virgin America	61903	1.063794	0.862640	0.195467
1	Hawaiian Airlines Inc.	76272	1.310723	0.224198	0.078666
2	Frontier Airlines Inc.	90836	1.561003	0.647320	0.173940
3	Spirit Air Lines	117379	2.017141	1.707290	0.155053
4	Alaska Airlines Inc.	172521	2.964748	0.387779	0.239391
5	US Airways Inc.	198715	3.414887	2.046650	0.213874
6	JetBlue Airways	267048	4.589180	1.601210	0.273359
7	American Eagle Airlines Inc.	294632	5.063207	5.099582	0.276956
8	United Air Lines Inc.	515723	8.862622	1.274521	0.269137
9	Atlantic Southeast Airlines	571977	9.829339	2.662869	0.348615
10	Skywest Airlines Inc.	588353	10.110758	1.692861	0.268376
11	American Airlines Inc.	725984	12.475926	1.504028	0.293395
12	Delta Air Lines Inc.	875881	15.051884	0.436589	0.203452
13	Southwest Airlines Co.	1261855	21.684789	1.271382	0.270158

This is showing that southwest is a market leader with around 22% of all the total US domestic flights while it only had 1.3 % of flights cancelled (of all US domestic flights).

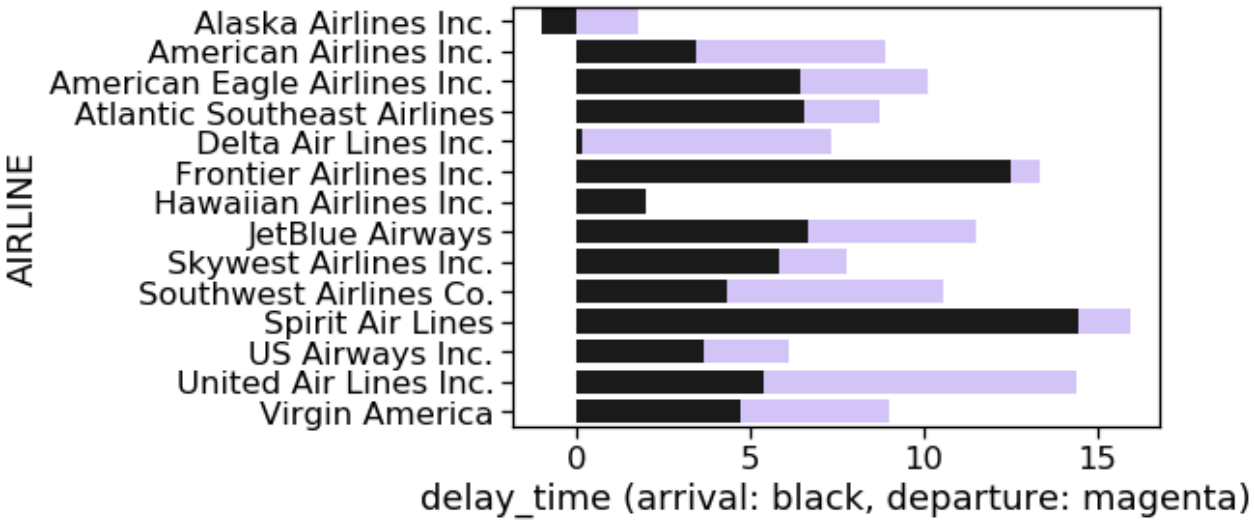
Taxi-in vs taxi-out in stacked bar chart for every airline

Another insight we wanted to check was taxi-in vs taxi-out by airline to check which airlines are the most efficient in these parameters.



It is clearly visible that US airways has largest taxi-in and taxi-out time amongst all the airlines. Hawaiian Airline flights spends least time on taxi-in and taxi-out. Good for the traveler!

Arrival vs departure delay for every airline

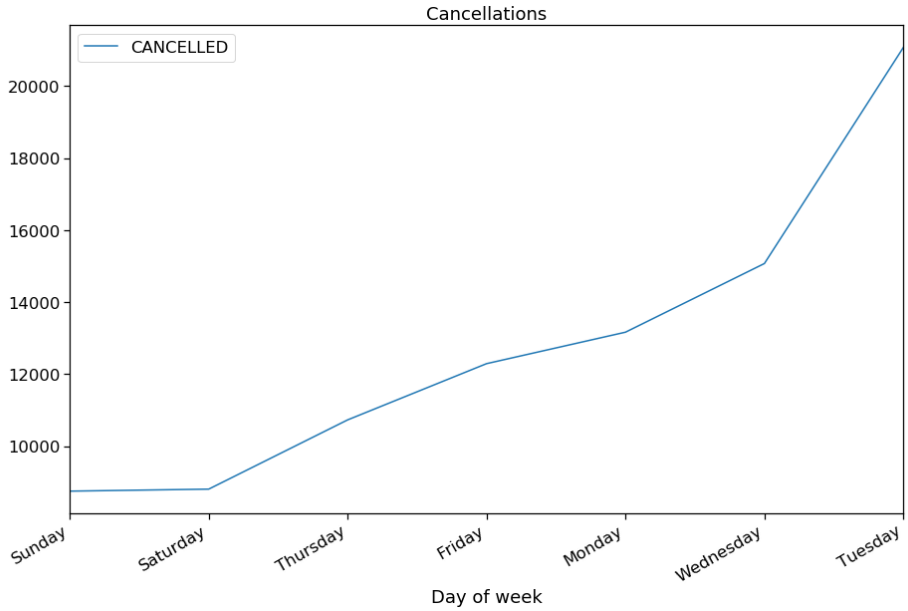


We can clearly infer from this visualization that Delta airlines almost has no arrival delay time whereas Alaska airlines has negative arrival delay. That means Alaska airlines most of the times reached before time. Spirit airlines performs worst when it comes to these delays with the highest arrival delay.

Week by cancellation

We wanted to see if we could come up with any interesting insights by looking at data of each day of week. We found that cancellation is highest on Tuesdays and lowest on Sunday & Saturday.

This can be due to less number of passengers travelling on weekdays just after the weekend resulting and underutilization of flights and thus higher cancellations. However, to prove this we will have to download passenger data set from transportation website and then analyze that. This is something we will try to do in future.



Delay ranges from airports

Now the final one, we knew from previous analysis that severity of delay is important. We also had good insights on airports and airlines.

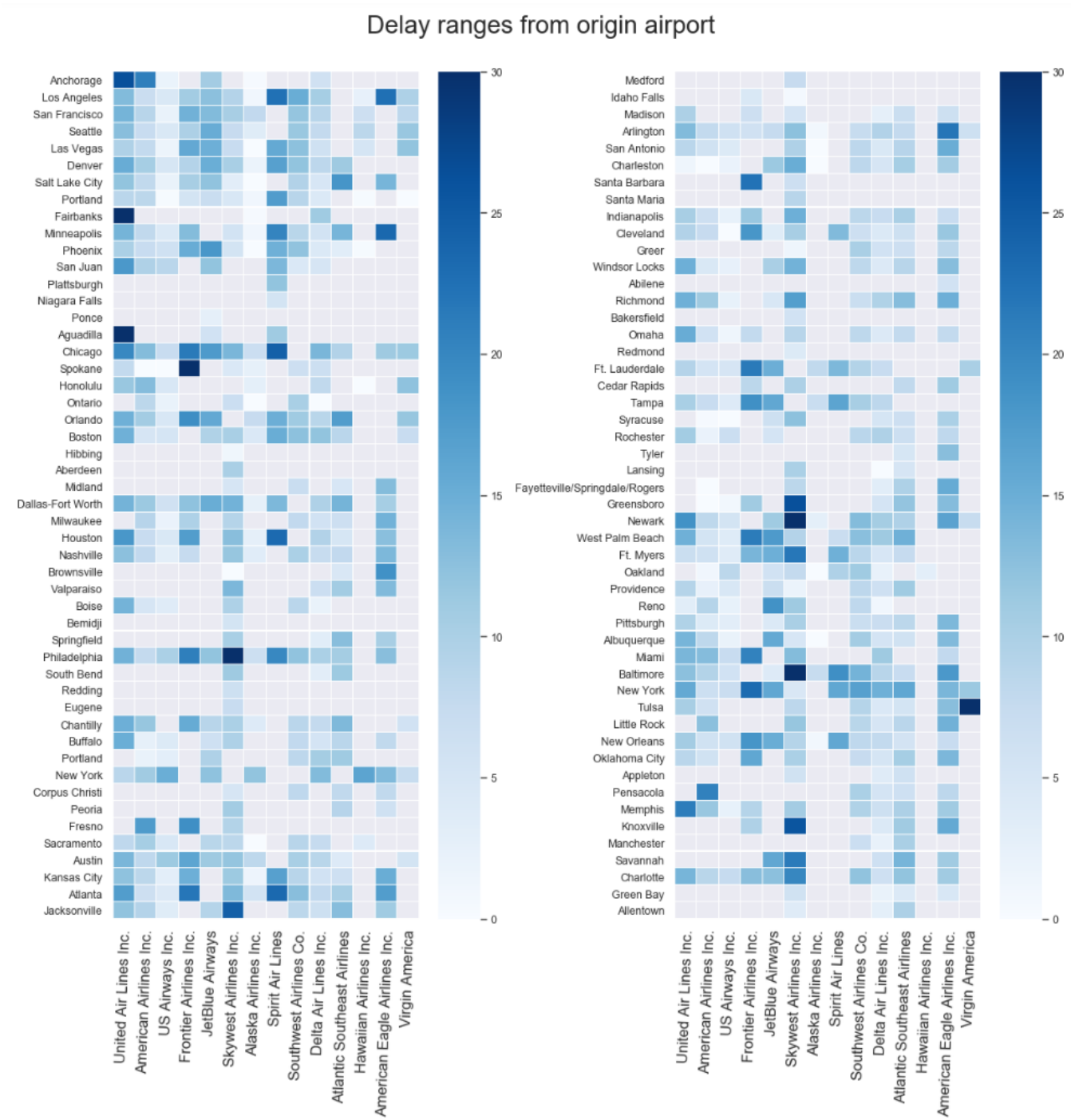
Can we do something to have all three together? If we can visualize that, it may really help a passenger to plan their travel efficiently.

We thought about a heat map that can show us delays (severity by heat) at each airport for each airline in a grid. After researching on internet, we found some interesting heat maps like the one explained in this pager (Group56, 2013) . This heat map was great, but it does no help in comparing airline carriers at each airport.

So, we looked at another example. This visualization (quarbby's, 2015) actually shows each airline against airports and gives some insights but it was very difficult to read.

Finally, we started designing our own heat map by referring (TutorialGateway, 2019) tableau heat map.

At the end, we came up with following version. In this heat map, airports are plotted on y axis and x axis shows airlines. As the shade of blue, get darker, delay increases. Shade variation and delay in minutes is listed to the right.


Delay ranges from origin airport
Anchorage
Los Angeles
San Francisco
Seattle
Las Vegas
Denver
Salt Lake City
Portland
Fairbanks
Minneapolis
Phoenix
San Juan
Plattsburgh
Niagara Falls
Ponce
Aguadilla
Chicago
Spokane
Honolulu
Ontario
Orlando
Boston
Hibbing
Aberdeen
Midland
Dallas-Fort Worth
Milwaukee
Houston
Nashville
Brownsville
Valparaiso
Boise
Bemidji
Springfield
Philadelphia
South Bend
Redding
Eugene
Chantilly
Buffalo
Portland
New York
Corpus Christi
Peoria
Fresno
Sacramento
Austin
Kansas City
Atlanta
Jacksonville
Medford
Idaho Falls
Madison
Arlington
San Antonio
Charleston
Santa Barbara
Santa Maria
Indianapolis
Cleveland
Greer
Windsor Locks
Abilene
Richmond
Bakersfield
Omaha
Redmond
Ft. Lauderdale
Cedar Rapids
Tampa
Syracuse
Rochester
Tyler
Lansing
Fayetteville/Springdale/Rogers
Greensboro
Newark
West Palm Beach
Ft. Myers
Oakland
Providence
Reno
Pittsburgh
Albuquerque
Miami
Baltimore
New York
Tulsa
Little Rock
New Orleans
Oklahoma City
Appleton
Pensacola
Memphis
Knoxville
Manchester
Savannah
Charlotte
Green Bay
Allentown
30
25
20
15
10
5
0
United Air Lines Inc.
American Airlines Inc.
US Airways Inc.
Frontier Airlines Inc.
JetBlue Airways
Skywest Airlines Inc.
Alaska Airlines Inc.
Spirit Air Lines
Southwest Airlines Co.
Delta Air Lines Inc.
Atlantic Southeast Airlines
Hawaiian Airlines Inc.
American Eagle Airlines Inc.
Virgin America

4 Conclusion

The visualizations presented clearly showed airports that one can avoid for travel during a particular time period. E.g. Avoid travelling to Chicago and LaGuardia in January (*Top 10 Busiest airports and their delays*).

Southwest is a market leader with around 22% of all the total US domestic flights (*Calculating cancellation, diversion rate and flight percentage for each airline carrier*) while it only had 1.3 % of flights cancelled. Southwest also fared well in total number of delayed flights and severity of delays. So basically, use Southwest!

The main reason for delay in January is not 'weather delay', as popular belief suggests but other delays like 'Carrier Delay' and 'Late Aircraft Delay'. (*Flight delay reason stacked area chart*)

Most number of flights are cancelled on Tuesday (probably due to empty flights), so avoid Tuesday if you want to avoid the hassle.

Finally, the visualization '*Delay Ranges from Origin Airport'* provides good heat map for a traveler to assess his/her choices to make an informed decision about travel. E.g. We will avoid Virgin airlines at Tulsa and SkyWest airlines for Newark and Baltimore due to their large delays at these airports last year.

5 Future Work

This project is just a start in our data science journey. A lot can be done on top of what we have achieved in this project.

These visualizations were a deep dive on a limited data to prove our concept, we can load historical data to come up with a trend in a future work.

We want to make our visualizations more interactive; we want to create visualizations that will allow users to select airports, airlines, states and time periods to analyze.

We can create a visualization where if you enter boarding and destination airports, it will provide best airlines based on punctuality analysis. We have to create algorithms in python to achieve this.

6 References

Airlines for America. (2010). *Annual U.S. Impact of Flight Delays (NEXTOR report)*. Retrieved from airlines.org: http://airlines.org/data/annual-u-s-impact-of-flight-delays-nextor-report/

Altair. (2019). *Locations of US Airports*. Retrieved from Altair: https://altair-viz.github.io/gallery/airports_count.html

Andrew Hough. (2010). *Holidaymakers waste 16 days waiting in airports over lifetime*. Retrieved from telegraph.co.uk: https://www.telegraph.co.uk/travel/travelnews/7546209/Holidaymakers-waste-16-days-waiting-in-airports-over-lifetime.html

FAA. (2019). *Flight Delay Information*. Retrieved from fly.faa.gov: https://www.fly.faa.gov/flyfaa/usmap.jsp

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

Group56. (2013). *Heat Mapping and Predicting Flight Delays and Their Propagations in a Real-World Air Traffic Simulation.* Stanford. Retrieved from http://snap.stanford.edu/class/cs224w-2013/projects2013/cs224w-056-final.pdf

IATA. (2018). *IATA Forecast Predicts 8.2 billion Air Travelers in 2037*. Retrieved from IATA: https://www.iata.org/pressroom/pr/Pages/2018-10-24-02.aspx

Plotly. (n.d.). *Ploty*. Retrieved from Plotly Graphing Libraries: https://plot.ly/python/choropleth-maps/

quarbby's. (2015). *Heatmap.* NA: Plotly. Retrieved from Plotly: https://plot.ly/~quarbby/58

TutorialGateway. (2019, 07 25). *TutorialGateway*. Retrieved from TutorialGateway: https://www.tutorialgateway.org/tableau-heat-map/