PROJECT - III REPORT

INSY 5378

AIRLINE PERFORMANCE INSIGHTS & RANKING

Using SQL & Python

by

Amal dev Thomas
Arun Tom
Namrata Patil
Smitha T Kumaraswamy
Under the Guidance of
Asst. Prof. Gene Moo Lee
Department of
Information Systems and Operations Management

May 2017

DEPARTMENT OF INFORMATION SYSTEMS AND OPERATIONS MANAGEMENT

UNIVERSITY OF TEXAS ARLINGTON

COLLEGE OF BUSINESS

1. PROBLEM STATEMENT

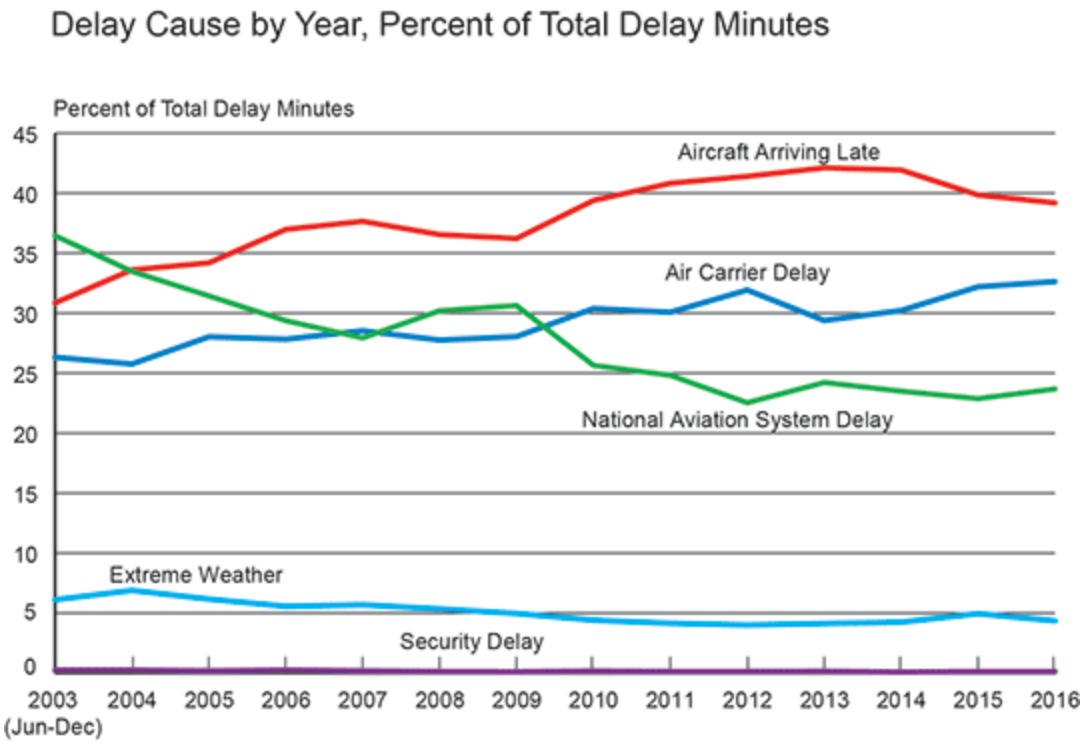
Insights on causes of flight delays and implementation of Airline carrier ranking system based on data on flights to and from Texas from Bureau of Transportation during Jan 2017.

2. MOTIVATION

Growing delays threaten the competitiveness of the U.S. in the world economy by limiting the ability of the air transport system to serve the needs of the U.S. economy. In addition to improving business performance generally, air transport impacts the economy through the jobs and revenue it directly creates in air transport-related industries, the expenditures of air travelers on auxiliary goods and services, and the secondary impacts that result as these dollars recycle throughout the economy.

Flight delay is a serious and widespread problem in the United States. Increasing flight delays place a significant strain on the U.S. air travel system and cost airlines, passengers and society many billions of dollars each year. The $8.3 billion airline component consists of increased expenses for crew, fuel and maintenance, among others. The $16.7 billion passenger component is based on the passenger time lost due to schedule buffer, delayed flights, flight cancellations and missed connections. The $2.2 billion cost from lost demand is an estimate of the welfare loss incurred by passengers who avoid air travel as the result of delays.

Since June 2003, the airlines that report on-time data also report the causes of delays and cancellations to the Bureau of Transportation Statistics. Reported causes of delay are available from June 2003 to the most recent month. Getting Insights about the causes and improve the quality of service is highly significant in every aspect.



3. TECHNOLOGY USED

Ggplot

It is a plotting system for Python based on R's ggplot2 and the Grammar of Graphics. It is built for making professional looking, plots quickly with minimal code.

Seaborn

It is a Python visualization library based on matplotlib. It provides high-level interface for drawing attractive statistical graphics.

Geoplotlib

It is an open source python toolbox for visualizing geographical data. It supports the development of hardware-accelerated interactive visualizations in pure python, and provides implementations of dot maps, kernel density estimation, spatial graphs, Voronoi tesselation, shapefiles and many more common spatial visualizations. We describe geoplotlib design, functionalities and use cases.

Matplotlib

It is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.

Pandas dataframe

It is a 2-dimensional labeled data structure with columns of potentially different types. It is like a spreadsheet or SQL table, or a dict of a series objects. It is generally the most commonly used pandas object.

Pandasql

It allows you to query pandas DataFrames using SQL syntax. It works similarly to sqldf in R. pandasql seeks to provide a more familiar way of manipulating and cleaning data for people new to Python or pandas.

min_max scaler

Transforms features by scaling each feature to a given range. This estimator scales and translates each feature individually such that it is in the given range on the training set between 0 and 1.

SQL

It is a standard computer language for relational database management and data manipulation. SQL is used to query, insert, update and modify data. An **SQL JOIN** clause is used to combine rows from two or more tables, based on a common field between them

4. DATA BACKGROUND

BACKGROUND

A Flight delay are caused when an airline flight takes off and/or lands later than its scheduled time. It is a frequent challenging problem encountered by the airline passengers. It leads to financial losses and negative impacts on the airline's business operation.

LITERATURE VIEW ON DELAY COSTS

In 2015, the direct costs for the aircraft block time for U.S passenger airlines was $65.43 per minute, nearly 16 percent less than in 2014.

Calendar Year 2015	**Direct Aircraft Operating Cost per Block Minute (in dollars)**	**Δ vs. 2014 (in %)**
Fuel	22.62	-39.1
Crew	19.54	11.9
Maintenance	11.63	-3.1
Aircraft Ownership	8.80	6.4
Other	2.85	-2.3
Total Direct Operating Costs	65.43	-15.9

DATASET AND ATTRIBUTES:

Our dataset includes 80 fields spanning over 11 categories which are,

- Time Period
- Airline
- Origin
- Destination
- Departure Performance
- Arrival Performance
- Cancellation and Diversion
- Flight Summaries
- Cause of delay
- Gate return information at origin airport
- Diverted airport information

From the original Data set, we have extracted data that are significant for our analysis, after that the reduced data set contained, 80988 Rows and 26 Fields.

FIELDS

Attribute Name	Description
FL_DATE	Flight Date (MM/DD/YYYY)
CARRIER	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.
FL_NUM	Flight Number
ORIGIN	Origin Airport
ORIGIN_CITY	Origin City
ORIGIN_CITY_NAME	Origin Airport, City Name
DEST	Destination Airport
DEST_CITY_NAME	Destination Airport, City Name
DEP_TIME	Actual Departure Time (local time: hhmm)
DEP_DELAY	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers
ARR_TIME	Actual Arrival Time (local time: hhmm).
ARR_DELAY	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
CANCELLED	Cancelled Flight Indicator (1=Yes)
CANCELLATION_CODE	Specifies the Reason For Cancellation
DIVERTED	Diverted Flight Indicator (1=Yes)
CRS_ELAPSED_TIME	CRS Elapsed Time of Flight, in Minutes
ACTUAL_ELAPSED_TIME	Elapsed Time of Flight, in Minutes
AIR_TIME	Flight Time, in Minutes
DISTANCE	Distance between airports (miles)
TAXI_IN	Taxi In Time, in Minutes
TAXI_OUT	Taxi Out Time, in Minutes
CARRIER_DELAY	Carrier Delay, in Minutes
WEATHER_DELAY	Weather Delay, in Minutes
NAS_DELAY	National Air System Delay, in Minutes
SECURITY_DELAY	Security Delay, in Minutes
LATE_AIRCRAFT_DELAY	Late Aircraft Delay, in Minutes

TYPES OF DELAY

ASPM records minutes of delay for five possible causes of flight arrival delays: carrier, weather, NAS, security, and late arrival. The data are provided by the Bureau of Transportation Statistics (BTS) for ASQP flights only. These causes of delay were determined by the Department of Transportation.

CARRIER DELAY

Carrier delay is within the control of the air carrier. Examples of occurrences that may determine carrier delay are: aircraft cleaning, aircraft damage, awaiting the arrival of connecting passengers or crew, baggage, bird strike, cargo loading, catering, computer, outage-carrier equipment, crew legality (pilot or attendant rest), damage by hazardous goods, engineering inspection, fueling, handling disabled passengers, late crew, lavatory servicing, maintenance, oversales, potable water servicing, removal of unruly passenger, slow boarding or seating, stowing carry-on baggage, weight and balance delays.

LATE ARRIVAL DELAY

Arrival delay at an airport due to the late arrival of the same aircraft at a previous airport. The ripple effect of an earlier delay at downstream airports is referred to as delay propagation.

NAS DELAY

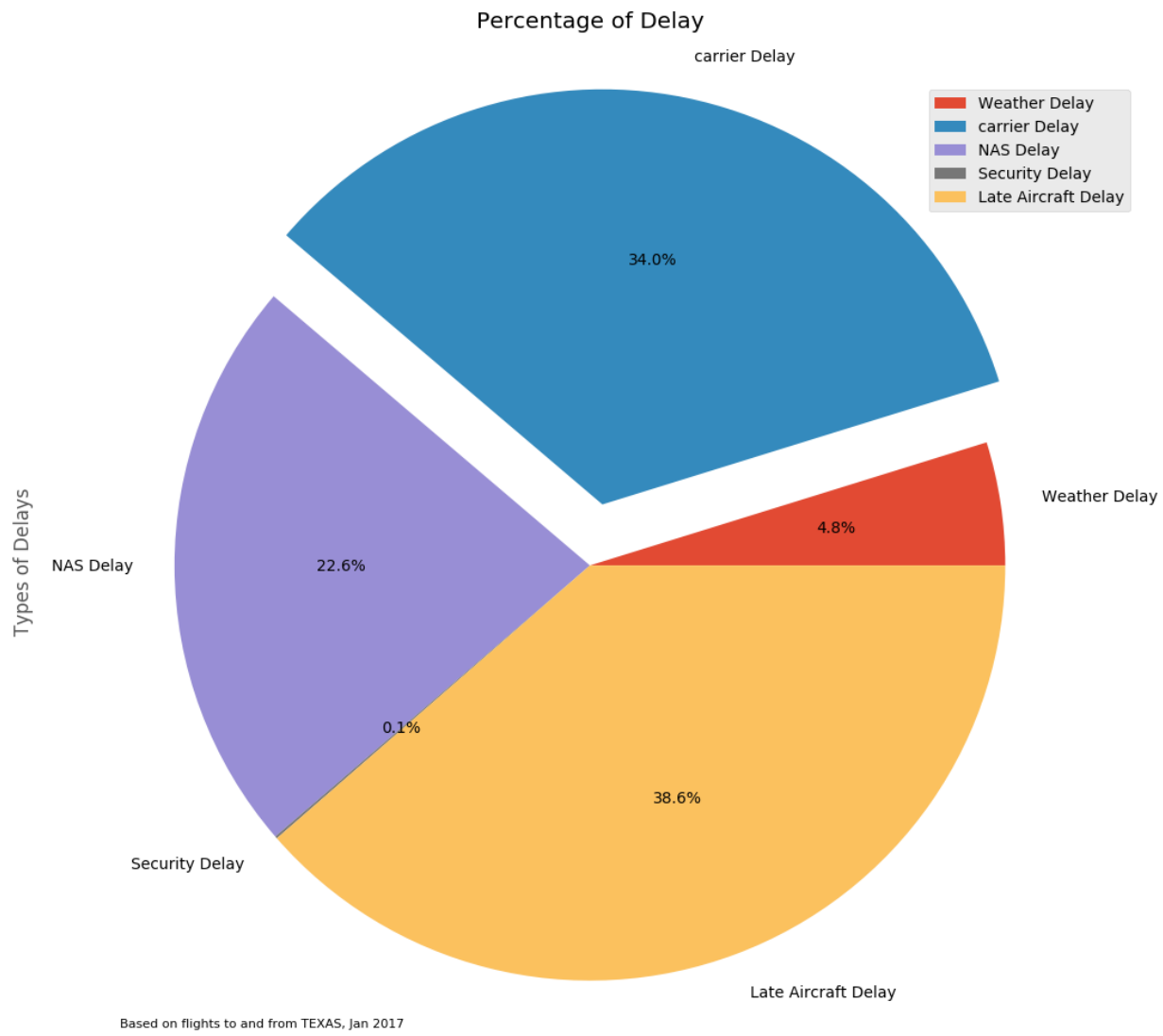
Delay that is within the control of the National Airspace System (NAS) may include: non-extreme weather conditions, airport operations, heavy traffic volume, air traffic control, etc. Delays that occur after Actual Gate Out are usually attributed to the NAS and are also reported through OPSNET.

SECURITY DELAY

Security delay is caused by evacuation of a terminal or concourse, re-boarding of aircraft because of security breach, inoperative screening equipment and/or long lines more than 29 minutes at screening areas.

WEATHER DELAY

Weather delay is caused by extreme or hazardous weather conditions that are forecasted or manifest themselves on point of departure, enroute, or on point of arrival.



The above pie chart depicts the effect of each type of delays in our data set. Based on the above analysis, the **late aircraft delay (38.6%)**, the **Carrier delay (34%)** and **NAS delay (22.6%)** consists of most of the delays. It is important that the delays due to weather conditions is 4.8% which is significantly less. Both Late Aircraft delay and Carrier delays are directly connected to the performance of carrier, so it is evident that this analysis can provide insights for the evaluation of carrier and can be used for further improvement.

5. IMPLEMENTATION

Our entire dataset was downloaded from the website of Bureau of transportation Statistics, and included 80,998 rows of data. The dataset contained large number of missing values, these missing fields were filled using **fillna** function which populated the fields with 0's according to the situation. Our main focus during the data collection was on the flights to and from Texas in the month of Jan 2017.

Our objective was to gain insight about the causes of airline delays and how they would in turn affect the ranking of an aircraft carrier and based on this insight we assigned a performance rank for each carrier. The dataset contained Carrier code. The name of the airlines was in a different csv file. We opened our data as a pandas dataframe from the original csv file. Two files were opened this way- One for the main data and one for the carrier names which were abbreviated. We had used sql to join the name of airlines and the corresponding abbreviations. Then we used adhoc sql queries on the joined dataframe and got the resultant dataframes. Modules like **matplotlib**, **geoplotlib**, **seaborn** etc. were used for creating meaningful visualization.

For plotting data on the map using geoplotlib we required the latitude and longitude of each city. For that we used geocoder from **geopy** module. After iterating through the destination and source cities the accurate latitude and longitudes were obtained. We used google service for getting the location in geocoder. Even though openstreet map service was used for this purpose, due to inaccuracy we changed to google service. For using the google map api we created an api key. The contents of dataframe were written to csv and using this data we obtained a meaningful visualization in geoplotilb.

RANKING SYSTEM

To obtain a meaningful and efficient ranking system the following data were considered:

- Number of airlines for each carrier (flight volume).
- Speed of the aircrafts in mph.
- Ratio between operated flights and scheduled flights.
- Taxi in and Taxi out time (time to leave and enter the gate).
- The average arrival delay.

We have not included avg. departure delay because usually it depends on the departure airport. The above data was used to obtain a ratio which indicated a score for the aircrafts rankings. The aircrafts airtime was used to obtain the speed of flight. After analyzing the speed of all the aircrafts, we found their averages to be between 400 and 500 mph. The distance between source and destinations were plotted using the arrival delay and departure delays. As an inference, we came to understand that delay was comparatively higher during short distance travels. Long distance flights experienced shorter delays. To an extent distance did not affect the flight delay.

After analyzing the causes of delay -weather delays were almost 4.8% and was the least. The maximum delays were caused by late aircraft delay and carrier delay (due to aircraft maintenance) and was near to 70% together.

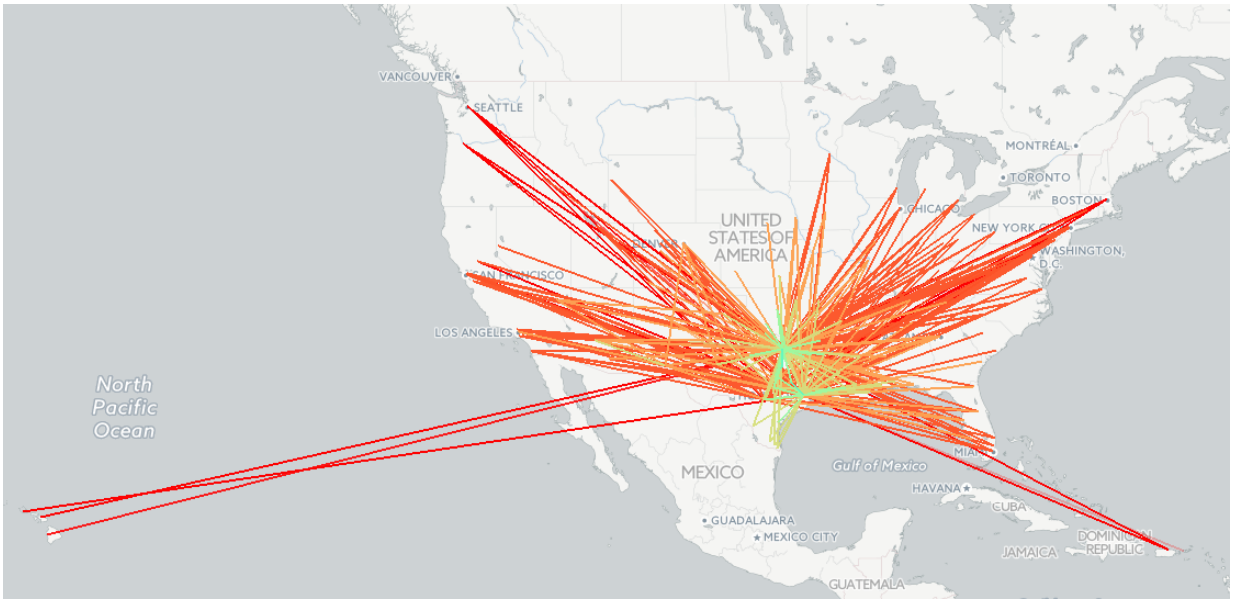
All the values were rescaled and normalized using min_max scaler and thus all values were between 0 and 1. Due to properties of our ranking system we rescaled between 1 and 2. To obtain the final ranking, we used a score variable comparing 6 variables which was in turn used to check the ranking status. I.e. a higher score would lead to a higher ranking which means the scheduled flight was operated correctly without any delays.

Score= a /(1+b)

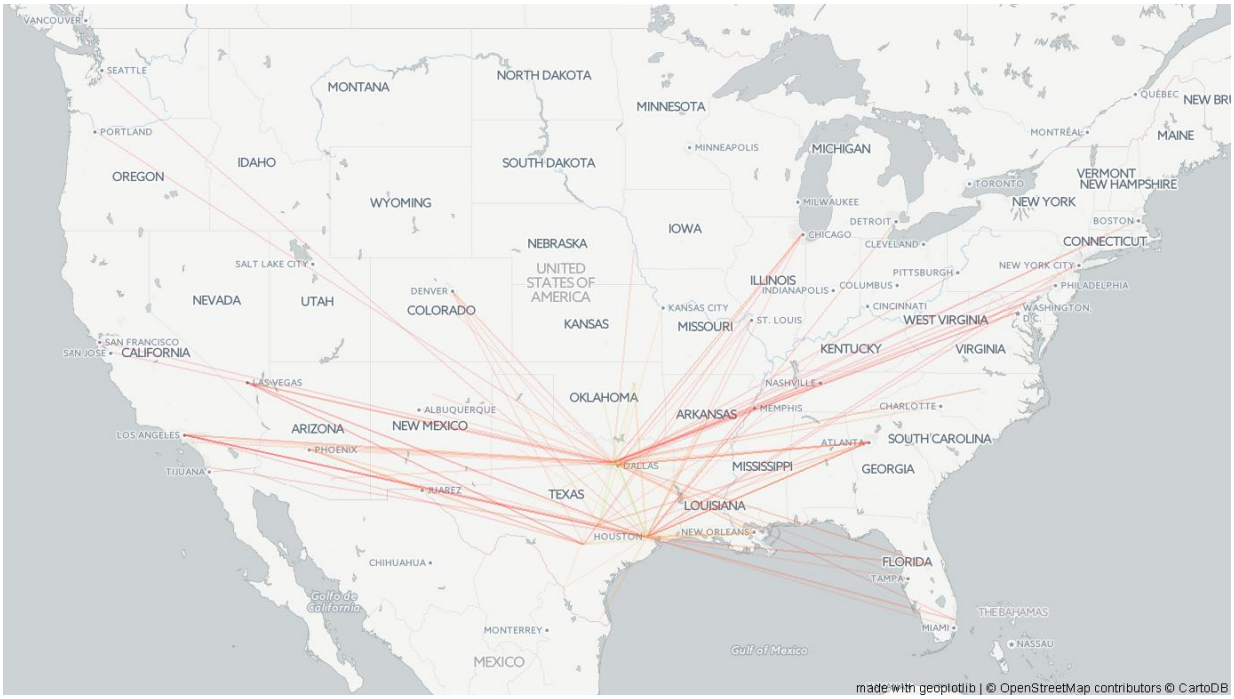
where a is the product of operated flights, total volume of flights and the flight speed and b is the product of arrival delay, average taxi in and taxi out times. The values of variable **a** were directly proportional to Score and the values of **b** were inversely proportional to score.

6. INSIGHTS AND ANALYSIS

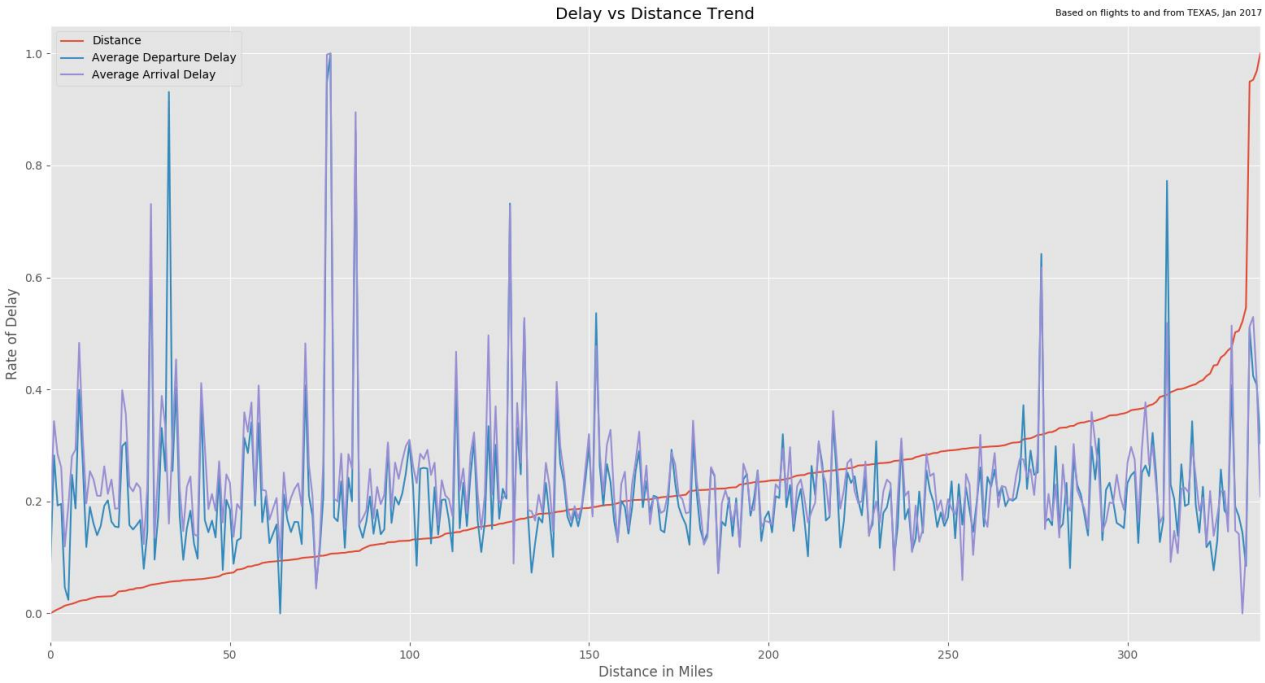
PLOTTING AIRLINE ROUTE OF ENTIRE DATASET



PLOTTING OF ROUTES WITH NO DELAYS

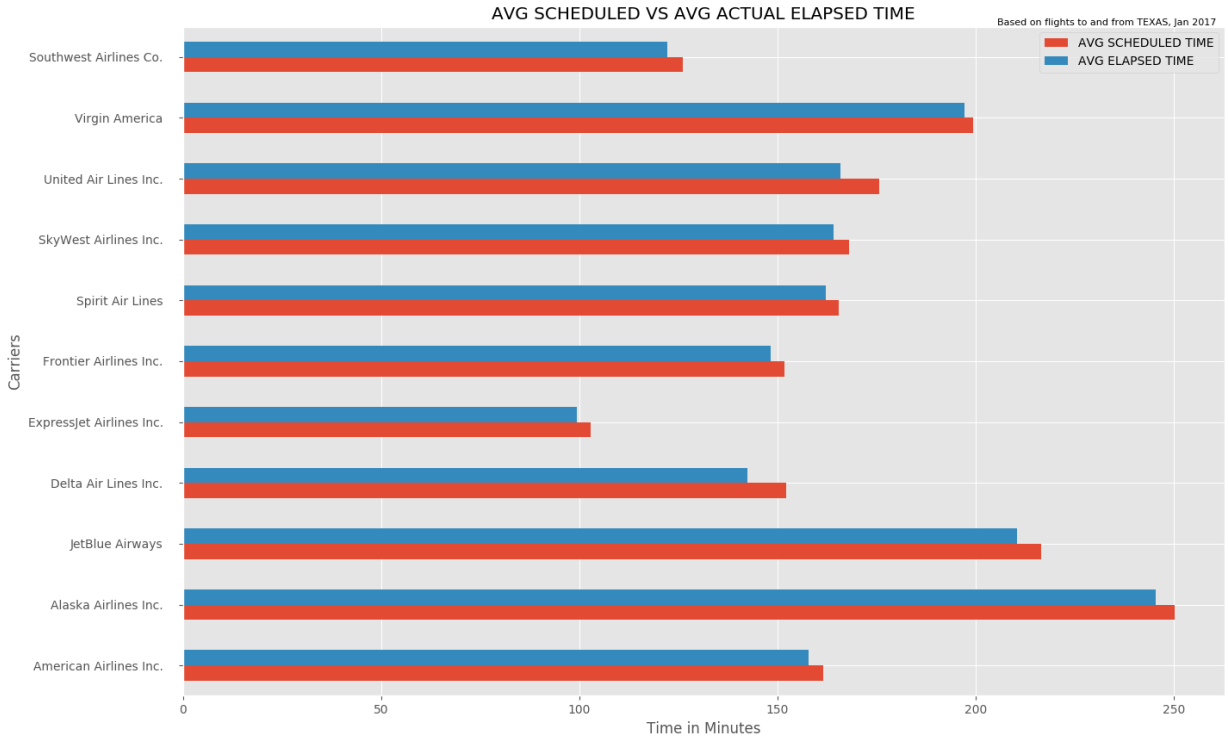


DISTANCE AGAINST ARRIVAL AND DEPRATURE DELAYS



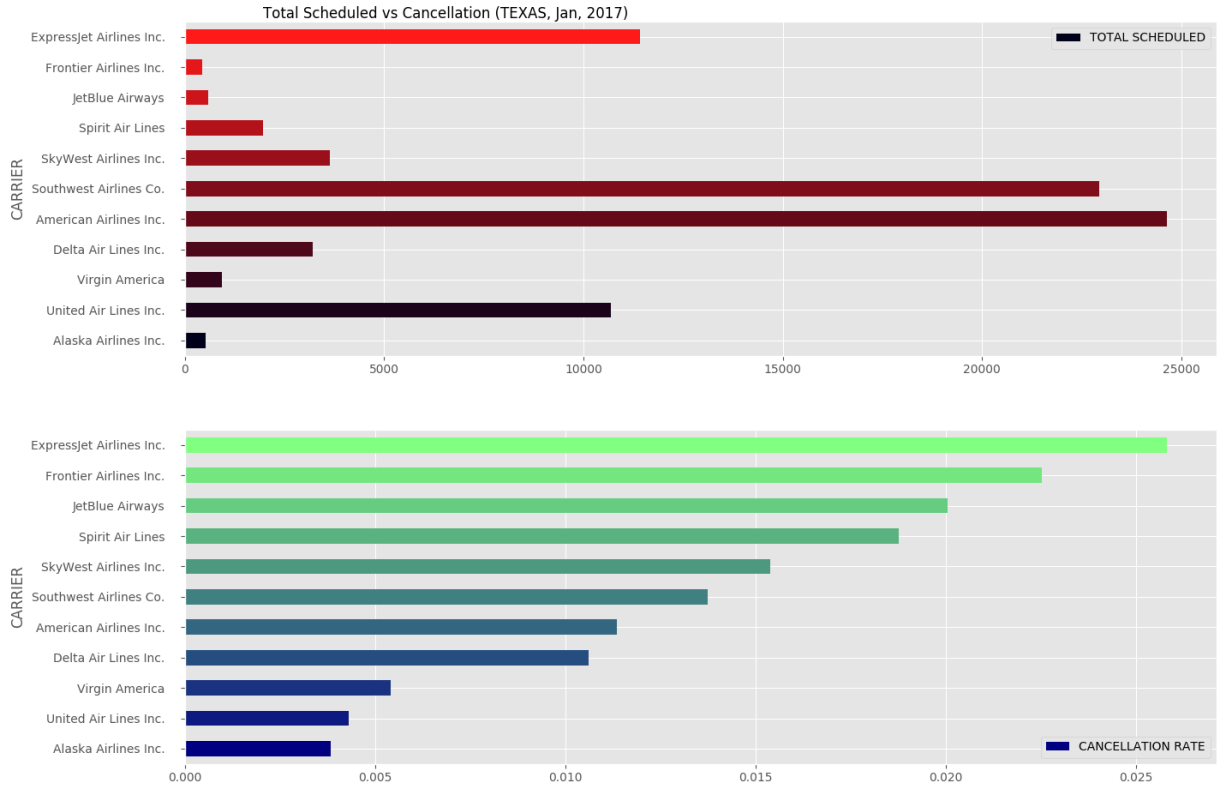
From the graph, it is evident that distance of flight doesn't have much impact on the delays, even though lower distance flight shows relatively higher delay. There for, we can assume that longer flights might have makeup the delayed minutes during flight.

SCHEDULED VS ELAPSED TIME OF FLIGHT



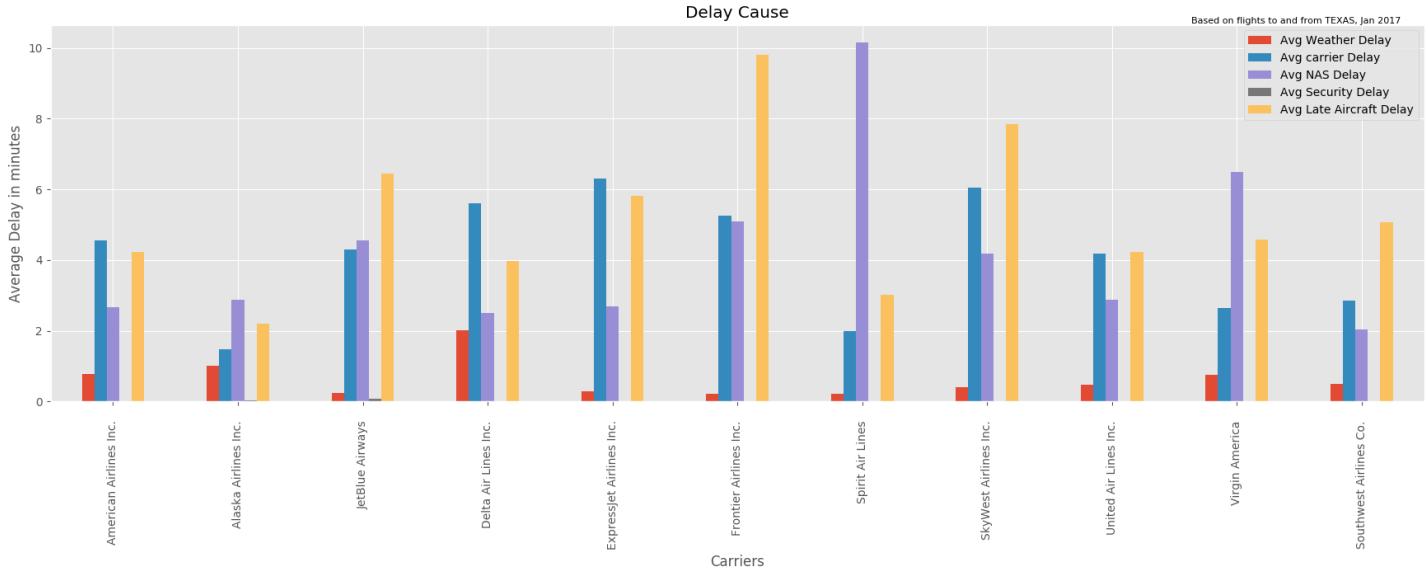
From the Graph, we can see that all carriers have completed the flight before scheduled time.

TOTAL SCHEDULED VS CANCELLED



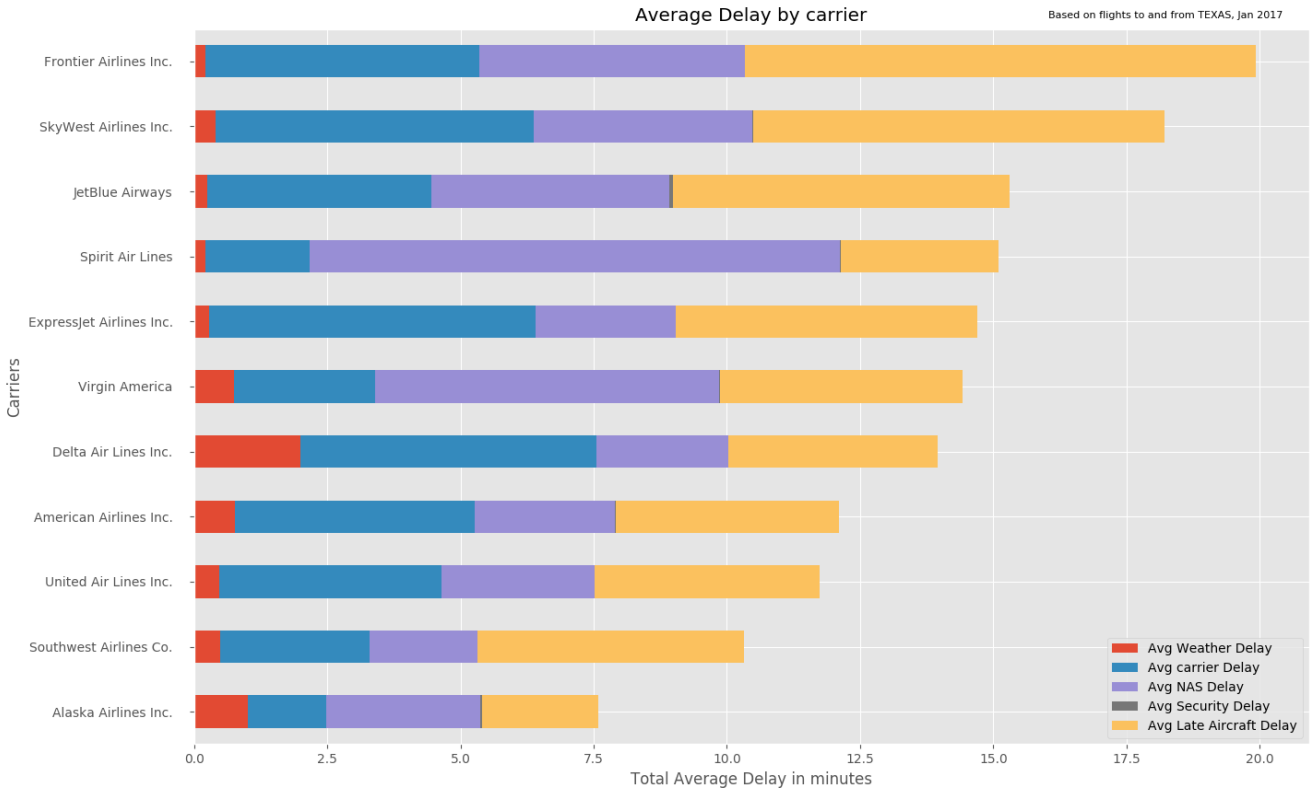
Expresjet Airline and Frontier Airlines has the highest rate of cancellation, which later going to effect on their rank. From the graph, we can see that American Airlines and Southwest have significant amount of flights and very less cancellation. From these we can say that their service quality is high and they will be securing high ranks

DELAY CATEGORIES OF EACH CARRIER

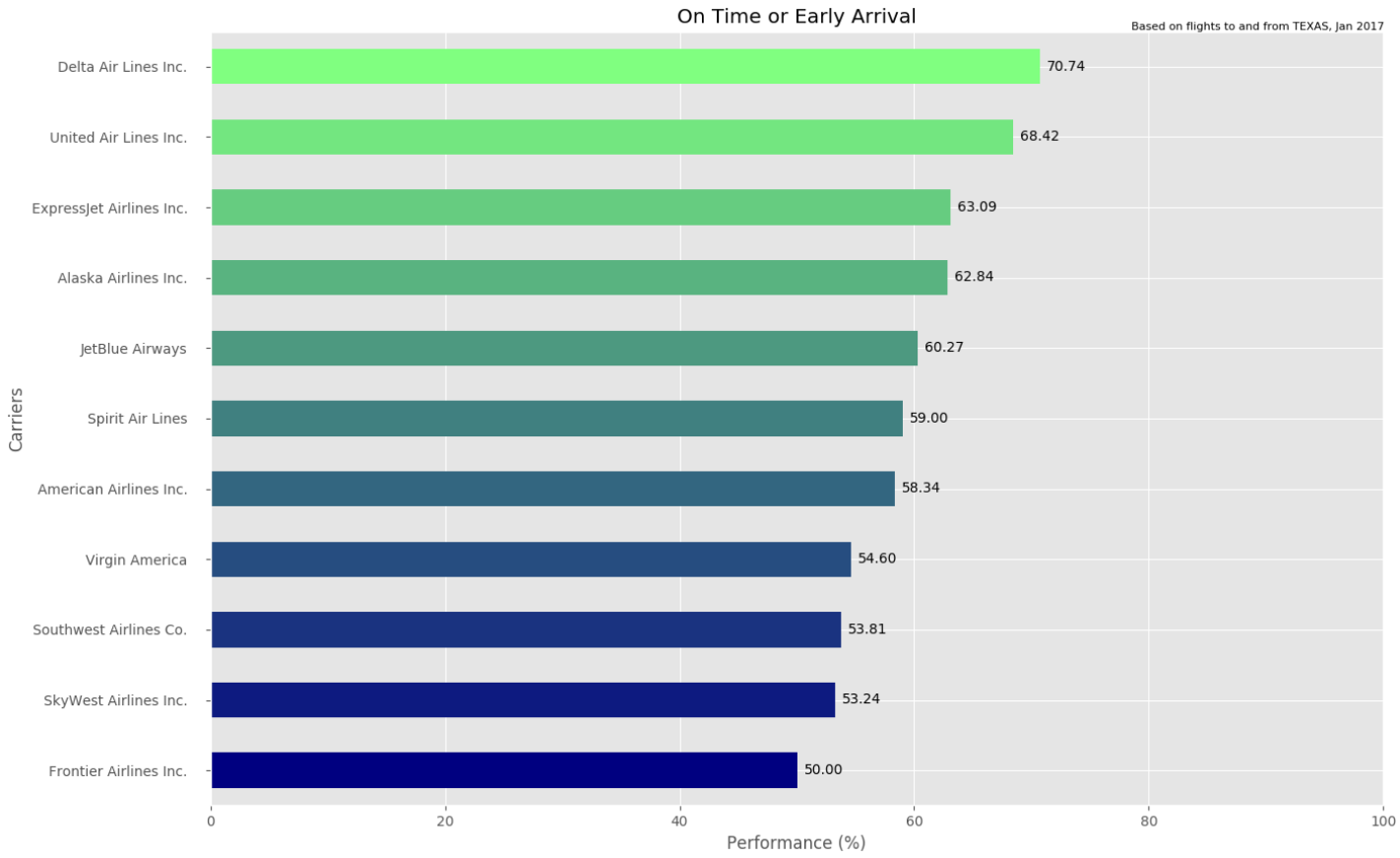


From the above bar chart, in each category of delay American Airlines, Alaskan Airlines, United Airlines and Southwest airlines has scored very less delays. Fortier Airlines has very high late aircraft delay, which should be handled and Spirit Airlines has the peek NAS delay which is very less for all other carriers (they must have investigated on this issue).

From the below graph, it is evident that Fortier airlines has the highest sum of average of all delays. And Alaskan Airlines and southwest airlines has lowest.

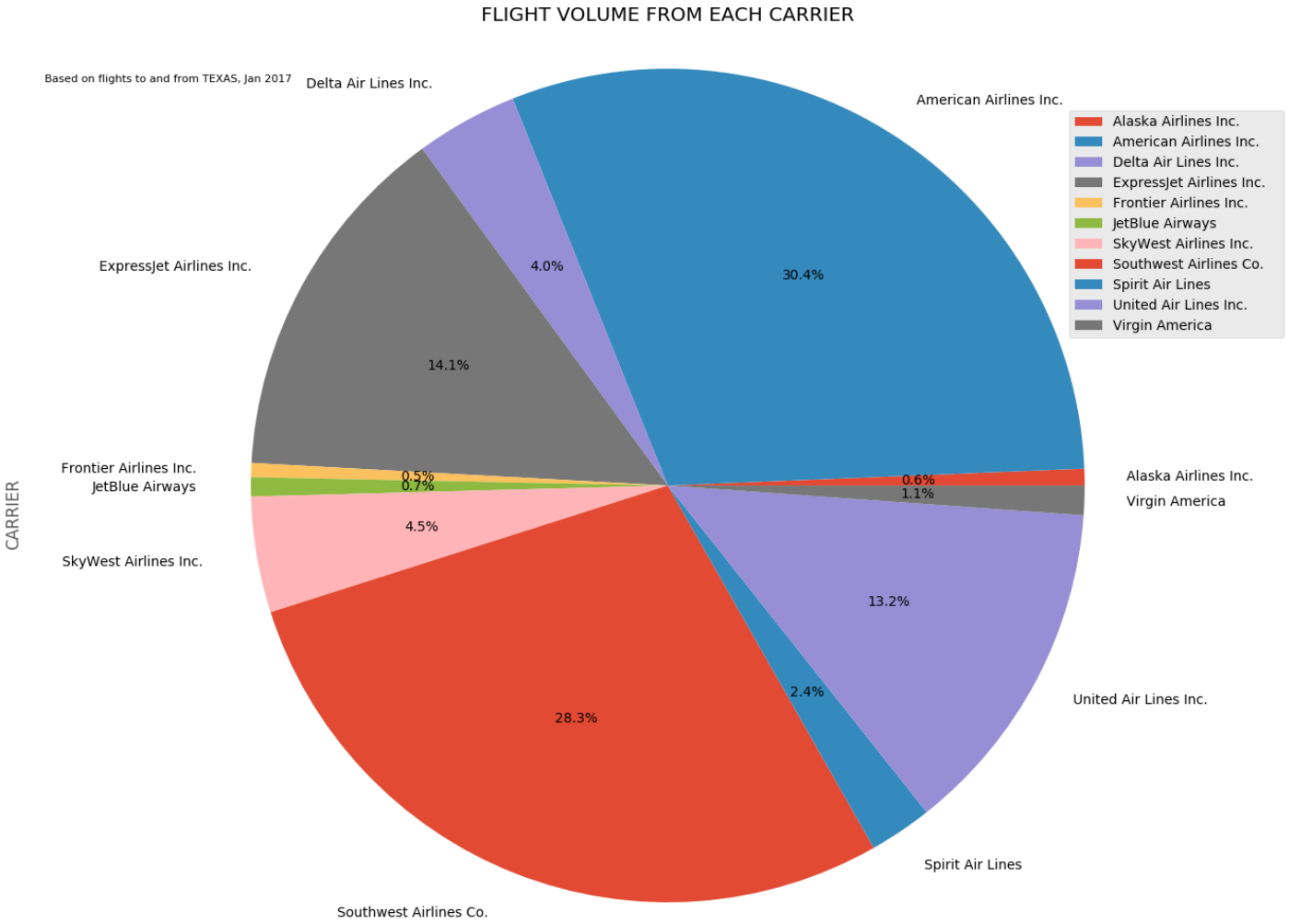


PERFORMANCE BASED ON ARRIVAL DELAY



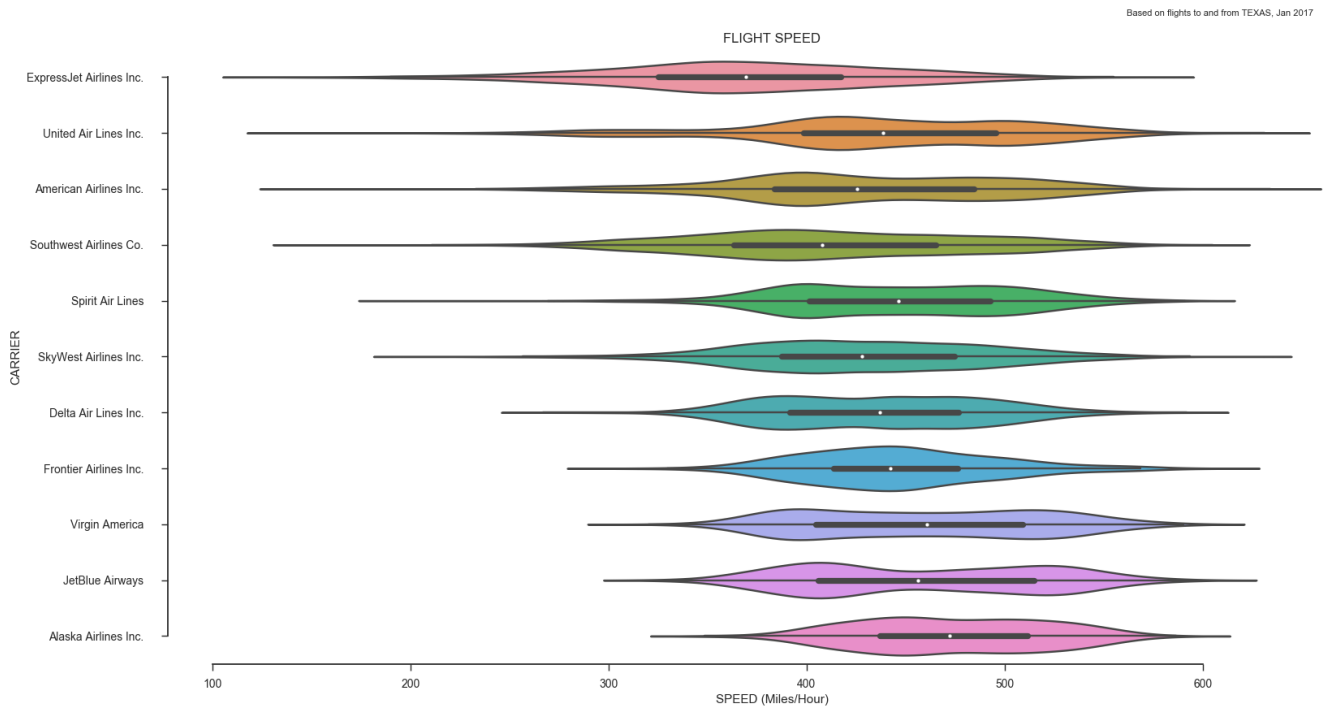
Above, performance graph show that Delta Airlines, united airlines have higher percentage of on time or early arrival with around 70%. Forntier airlines and skywest Airlines are the low scorers with around 50% of on time or early arrival.

FLIGHT VOLUME / TOTAL NUMBER OF FLIGHTS FROM EACH CARRIER



We can see that American Airlines (30.4%) and Southwest airlines (28.3 %) together have the lion share of flights to and from TEXAS. Frontier Airlines and JetBlue Airways have the least number of flights.

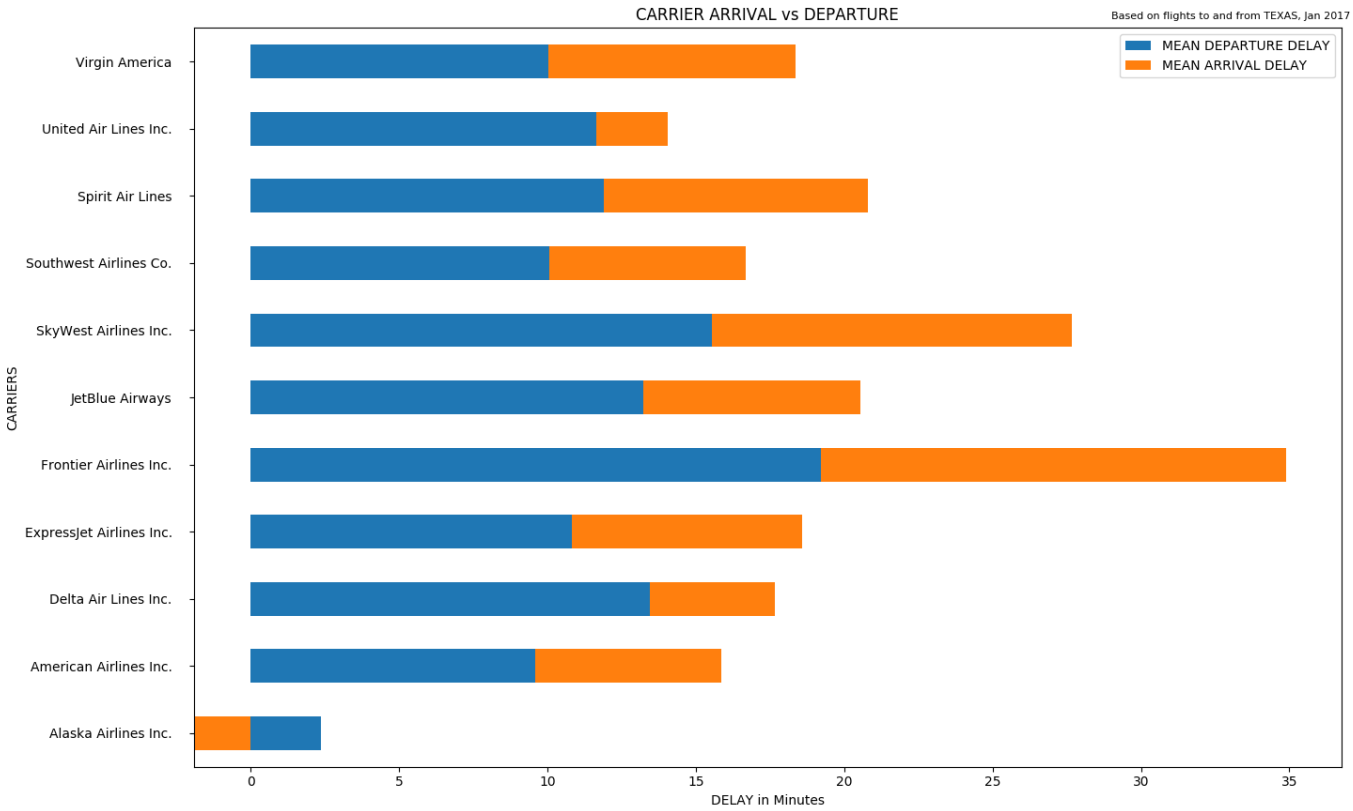
FLIGHT SPEED



Based on the violin plot, we can see that in average, most flying speed across airlines are close to 400~450 miles per hour; with the ExpressJet Airlines Inc. is the slowest airline and large variation (by simply looking at the data shape distribution). The fastest service is offered by Alaskan Airlines.

It is interesting to see that, in some rare cases, an aircraft can go as high as 800 miles per hour in average during a flight trip.

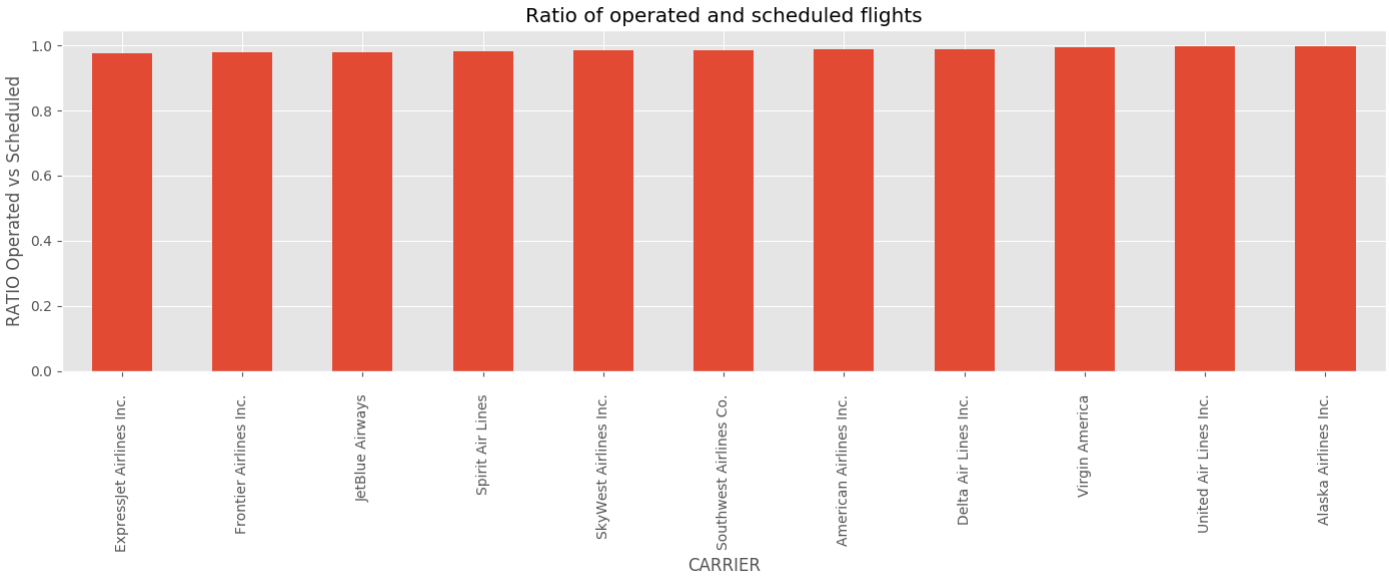
AVERAGE ARRIVAL DELAY VS DEPARTURE DELAY



Based on this analysis, we can see that all the lines have longer departure delays than arrival delays, Frontier Airlines have both delays equal and high. We can say that the flights can adjust speed to catch up time while departure delay sometimes are out of control.

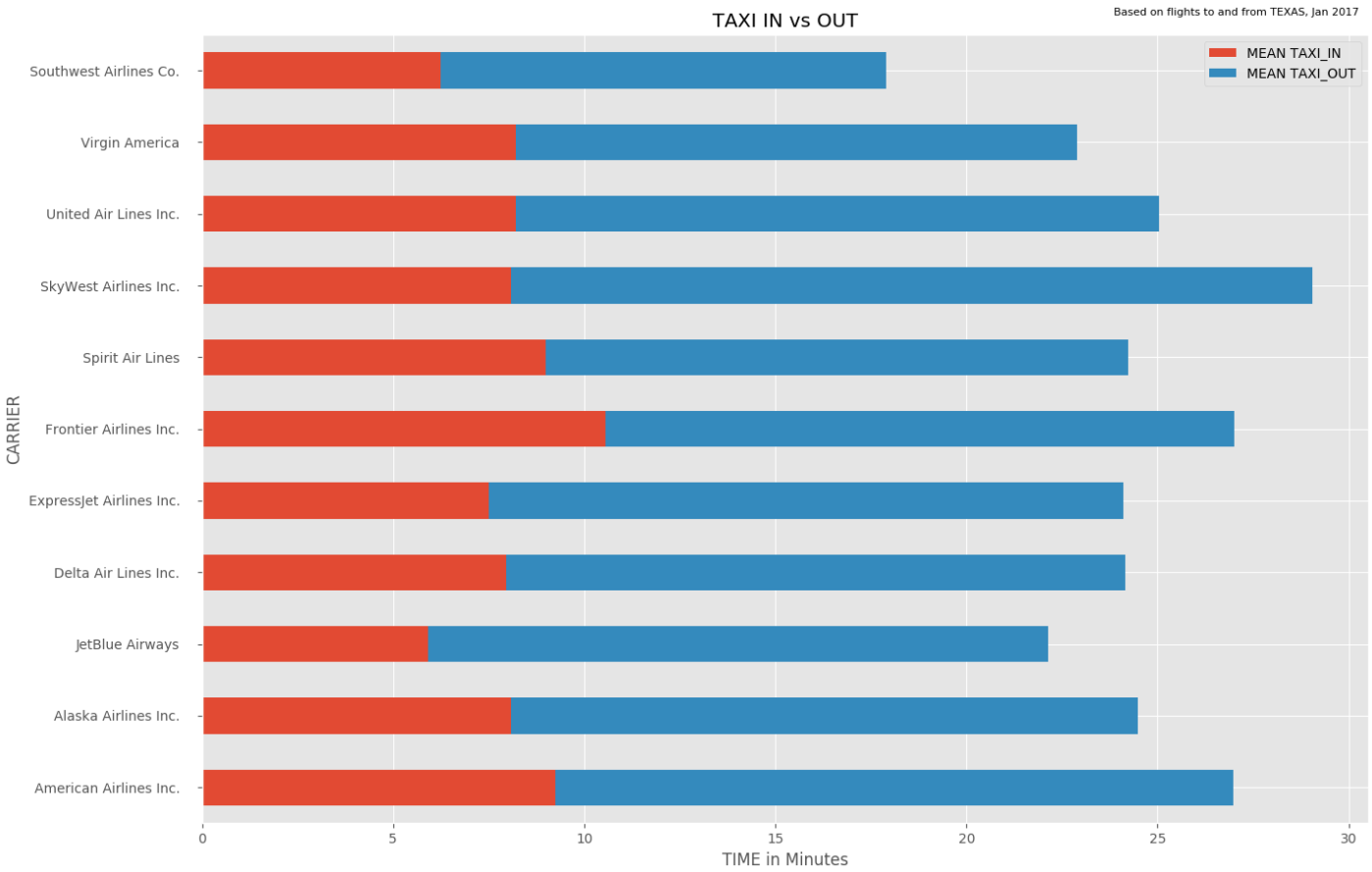
Skywest Airlines and Frontier Airlines are among the longest arrival and departure delay airlines. It is worth noting that Alaska Airlines is the only airline among all to arrive the destination earlier than scheduled in average.

RATIO OF OPERATED AND SCHEDULED FLIGHTS



From the graph, we can see that about 98% of scheduled flights are operated, with Expressjet Airlines with the lowest and Alaskan Airlines with the highest.

MEAN TAXI IN AND TAXI OUT



Interestingly, we can see that overall taxi in time is less than taxi out time for all the airlines. All airlines have an average taxi_in time less than 10 minutes except Frontier Airlines, while all taxi out time are greater than 10 minutes. Also, it seems Southwest has the shortest taxi-in and taxi-out.

7. ANALYSIS RESULT

We have 6 variables which decide score. The score is proportional to a subset (a) of the variables whereas being inversely proportional to a different subset (b) of the variables. We used the following formula for calculating score on normalized data which we scaled between 1 and 2.

CARRIER	RATIO_OP_SCH	FLIGHT_SPEED	ARRIVAL_DELAY	FLIGHTS_VOLUME	TAXI_IN	TAXI_OUT
Alaska Airlines Inc.	2	2	1	1.003223673	1.469513821	1.508629329
American Airlines Inc.	1.657281463	1.641380791	1.464607739	2	1.718027285	1.652870336
Delta Air Lines Inc.	1.691794426	1.574900906	1.349132845	1.114151099	1.442462476	1.484923569
ExpressJet Airlines Inc.	1	1	1.549748055	1.45404199	1.344297793	1.528606601
Frontier Airlines Inc.	1.149548313	1.701109817	2	1	2	1.514635009
JetBlue Airways	1.262804743	1.780685714	1.525411059	1.006406018	1	1.490665621
SkyWest Airlines Inc.	1.474325537	1.559848463	1.79949787	1.132087948	1.467007869	2
Southwest Airlines Co.	1.549244328	1.507375997	1.486462214	1.929285832	1.069609378	1
Spirit Air Lines	1.321055248	1.708516905	1.613928843	1.063192263	1.661401389	1.38520954
United Air Lines Inc.	1.978557195	1.784826635	1.24498746	1.423499752	1.495306105	1.554886169
Virgin America	1.927850085	1.80040967	1.581637335	1.019796661	1.493109681	1.323942818

Score = a/(1+b),

Where,
a = (RATIO_OP_SCH) * (FLIGHT_SPEED) * (FLIGHTS_VOLUME) and
b = (ARRIVAL_DELAY) * (TAXI_IN) * (TAXI_OUT)

A higher score indicates a better rank.

RANKING

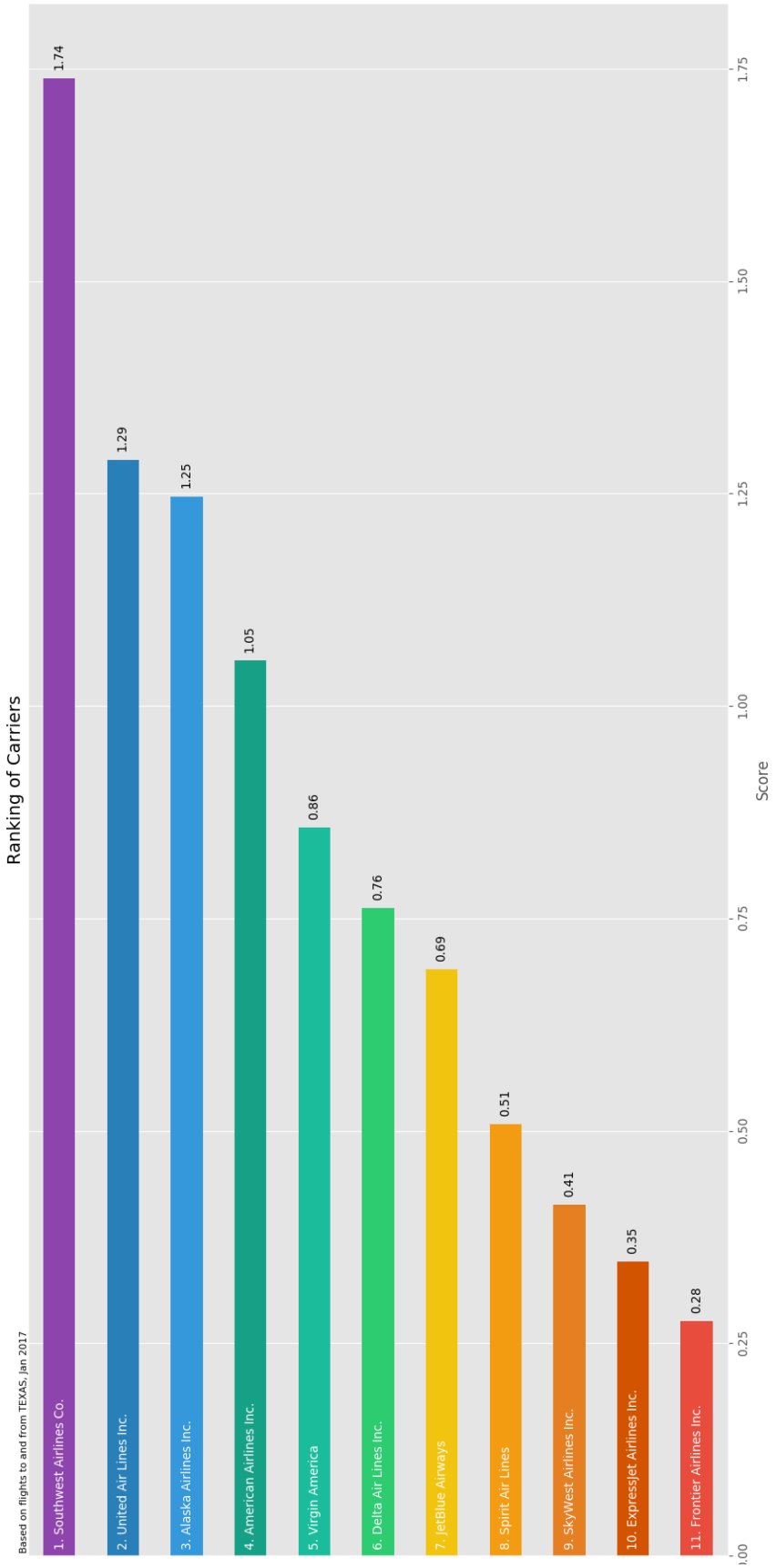
Based on the score, below are our ranking of Carriers, to and from Texas, Operated during Jan 2017.

1. Southwest Airlines Co.
2. United Air Lines Inc.
3. Alaska Airlines Inc.
4. American Airlines Inc.
5. Virgin America
6. Delta Air Lines Inc.
7. JetBlue Airways
8. Spirit Air Lines
9. SkyWest Airlines Inc.
10. ExpressJet Airlines Inc.
11. Frontier Airlines Inc.

Based on the data, Analysis and visualization we concluded that Southwest Airlines outperformed all other carriers and secured rank 1.

This lower scored carrier can improve their performance based on this insight and increase quality of service and gain customer satisfaction.

CARRIER	RATIO_OP_SCH	FLIGHT_SPEED	ARRIVAL_DELAY	FLIGHTS_VOLUME	TAXI_IN	TAXI_OUT	SCORE
Southwest Airlines Co.	1.549244328	1.507375997	1.486462214	1.929285832	1.069609378	1	1.739600009
United Air Lines Inc.	1.978557195	1.784826635	1.24498746	1.423499752	1.495306105	1.554886169	1.29072991
Alaska Airlines Inc.	2	2	1	1.003223673	1.469513821	1.508629329	1.247421513
American Airlines Inc.	1.657281463	1.641380791	1.464607739	2	1.718027285	1.652870336	1.054554628
Virgin America	1.927850085	1.80040967	1.581637335	1.019796661	1.493109681	1.323942818	0.857766733
Delta Air Lines Inc.	1.691794426	1.574900906	1.349132845	1.114151099	1.442462476	1.484923569	0.763169397
JetBlue Airways	1.262804743	1.780685714	1.525411059	1.006406018	1	1.490665621	0.691248554
Spirit Air Lines	1.321055248	1.708516905	1.613928843	1.063192263	1.661401389	1.38520954	0.509022364
SkyWest Airlines Inc.	1.474325537	1.559848463	1.79949787	1.132087948	1.467007869	2	0.414584689
ExpressJet Airlines Inc.	1	1	1.549748055	1.45404199	1.344297793	1.528606601	0.347476115
Frontier Airlines Inc.	1.149548313	1.701109817	2	1	2	1.514635009	0.277041415


Ranking of Carriers
Based on flights to and from TEXAS, Jan 2017
Score
1. Southwest Airlines Co.
2. United Air Lines Inc.
3. Alaska Airlines Inc.
4. American Airlines Inc.
5. Virgin America
6. Delta Air Lines Inc.
7. JetBlue Airways
8. Spirit Air Lines
9. SkyWest Airlines Inc.
10. Expressjet Airlines Inc.
11. Frontier Airlines Inc.
1.74
1.29
1.25
1.05
0.86
0.76
0.69
0.51
0.41
0.35
0.28
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75

8. REFERENCES

I. www.pandas.pydata.org
II. ggplot.yhathq.com
III. seaborn.pydata.org
IV. arxiv.org/abs/1608.01933
V. pypi.python.org/pypi/pandasql
VI. www.pymotw.com/2/pickle/
VII. www.matplotlib.org
VIII. www.kaggle.com
IX. www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time
X. aspmhelp.faa.gov/index.php/Types_of_Delay
XI. www.rita.dot.gov/bts/help/aviation/html/understanding.html