# Analysis on Flights Cancellation and Delays

**Objective**

The project report is intended to perform basic data exploratory analysis on the Flights Cancellation dataset published by US Department of Transportation. The analysis aims at revealing hidden data trends and also facts like most visited Destination, common Cancellation reasons, dependencies among various variables if any, most reliable carrier etc.

**Data Source**

The U.S. Department of Transportation’s (DOT) Bureau of Transportation Statistics tracks the on-time performance of domestic flights operated by large air carriers. The [Flights delay and cancellation](https://www.kaggle.com/usdot/flight-delays) data was collected and published by the DOT’s Bureau of Transportation Statistics which included summary information on the number of on-time, delayed, canceled, and diverted flights etc.

url: <https://www.kaggle.com/usdot/flight-delays/data>

**Data Format**

The following times are in the xx:yy - hour:minute format (e.g. 1536 means 3:36pm, 345 means 3:45am, 16 means 00:16am)

* scheduled\_departure – scheduled departure time from Origin Airport
* departure\_time- Actual departure time from Origin Airport
* scheduled\_arrival- Scheduled arrival time at destination airport
* arrival\_time – Actual arrival time at destination airport
* wheels\_off – Time stamp at take-off during flight
* wheels\_on – Time stamp when wheels touched the ground

The following times are in minutes format (negatives mean actual\_time is ahead of scheduled\_time for the absolute value of that negative number)

* arrival\_delay – Delay in arrival at Destination
* departure\_delay – Delay in departure at Origin
* taxi\_in – Time taken at Origin airport
* taxi\_out – Time take at Destination airport
* scheduled\_time – Scheduled time taken by flight
* elapsed\_time – Elapsed time taken by flight
* air\_time – Time in Air for a flight

*Distance is in miles*

Data Cleaning and Preparation

The process of data preparation involved following steps:

* Data was imported from three files named ‘Airports’, ‘Airlines’ and ‘Flights’.
* The data was then examined for missing values.
* In Flights, there were 12 columns for which the values were NA.
* A refined dataset was created by omitting NA values from 8 out of 12 columns. The reasons for exclusion of the 4 columns were first that, the values were missing in these for more than 80% of the rows and secondly, the columns were like AIR\_SYSTEM\_DELAY, SECURITY\_DELAY which can hold NA values in case there is no delay from the original schedule.
* It was further observed that majority of values for Cancellation reasons were empty.
* Data for Airlines and Airports were also examined for NA values and it was observed that Airport had 3 NA values for latitude and longitude. These values were not removed as latitude and longitude were not required and also can be restored by searching for airports.

Normalization

Data was further normalized through logical design and the following tables were created:

1. Dbo.FactFlights- containing all the measures
2. Dbo.DimAirlines- Dimension containing all the values for Airline carriers
3. Dbo.DimAirports- Dimension containing information about Airports
4. Dbo.DimDate- Dimension containing values pertaining to time.

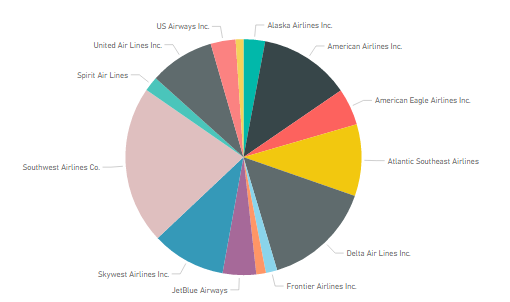
Following stored procedures were created to normalize the data in above four tables:

1. Dbo.uspCreatePopulateFactFlights
2. Dbo.uspCreatePopulateDimAirlines
3. Dbo.uspCreatePopulateDimAirport
4. Dbo.uspCreatePopulateDimDate

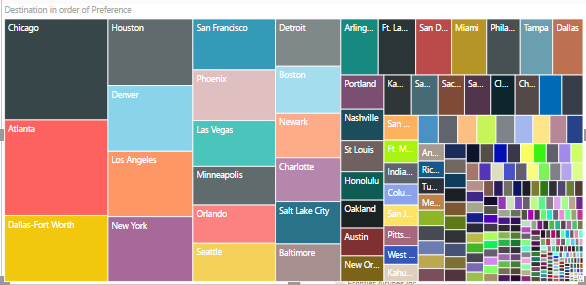
**Exploratory Data analysis**

The data was then explored on following aspects (refer to SQL Code attached in the appendix):

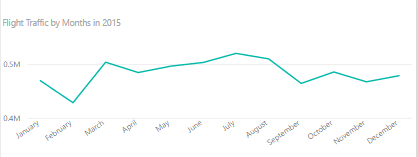
1. Maximum Flights by Airline Carrier:



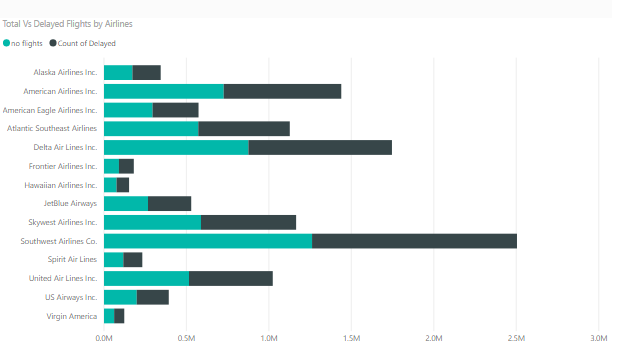
1. Most visited Destinations:



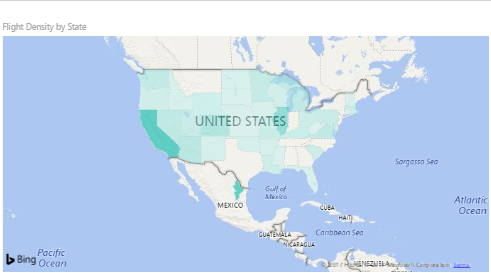
1. Flight Traffic by months:



1. Total vs Delayed flights by Carrier



1. Flight Density by state



**Summary**

After the above data analysis, the following insights were obtained from the Flight cancellation and delays data for the year 2015:

* 1. Chicago was one of the busiest of airports catering to maximum number of flights.
  2. Southwest Airlines Co. was the Airline with maximum number of flights.
  3. Chicago and Seattle were areas with high flight density.
  4. Almost 40 % of the flights were delayed from their scheduled arrival.
  5. February was the month with least air traffic.
  6. ‘San Francisco to Los Angeles’ and ‘New York to Chicago’ were among the busiest routes.
  7. Phoenix and Charlotte were the airports with maximum security delay at airport.
  8. ‘Chicago to Los Angeles’ was the route with maximum weather delay.

Please find attached the SQL code, R code and Power BI report used for the above analysis.