Submitted by: Balram Sidh

Stackline Data Engineer Assessment

Source Data

There are three source data files:

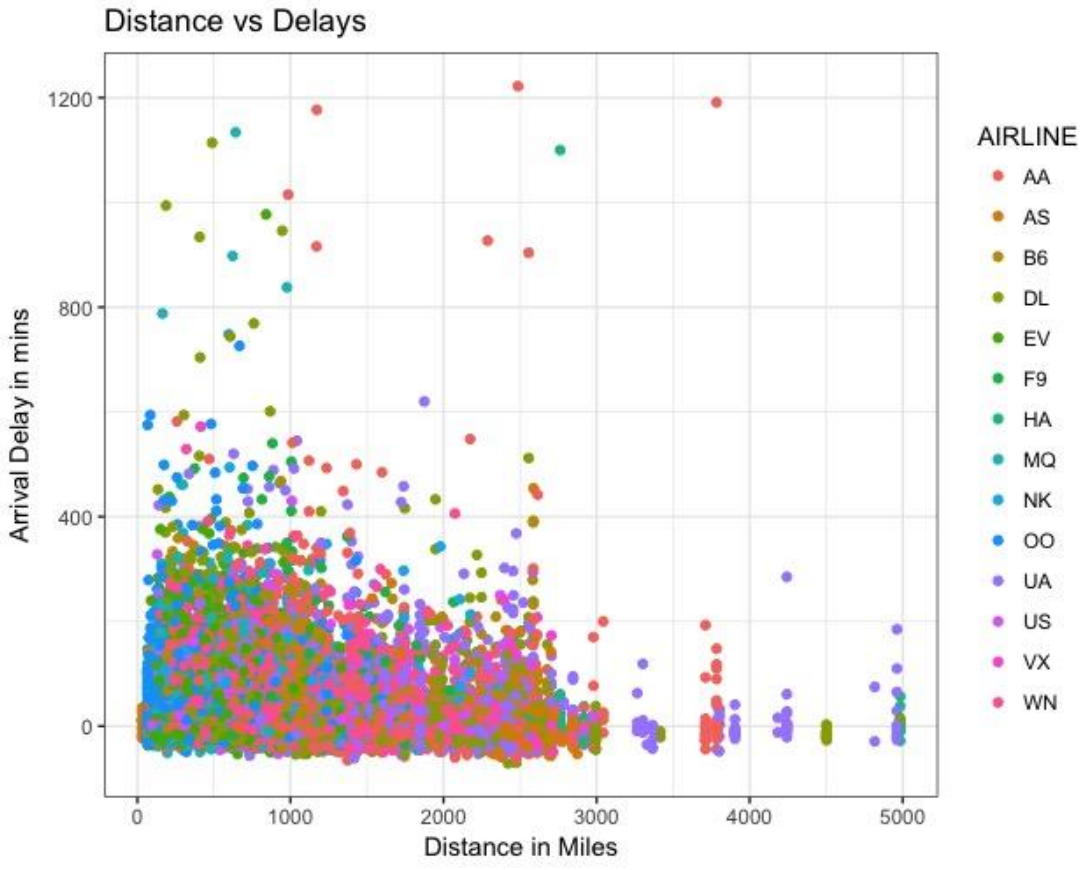
1. Flights: contains details of all flights that happened in 2015 in USA.
2. Airports: Details of all the airports involved.
3. Airline: Details of all the airline involved.

Questions

1. Correlation between distance and flight delay.

There are two delays observed in the dataset, one delay at the departure which can occur due to various reasons and other is arrival delay. Distance of the flight can one of the reasons which can affect the delay at the arrival.

The below scatter plot shows that there is no significant relationship:



Verifying the results with doing a Pearson correlation test:

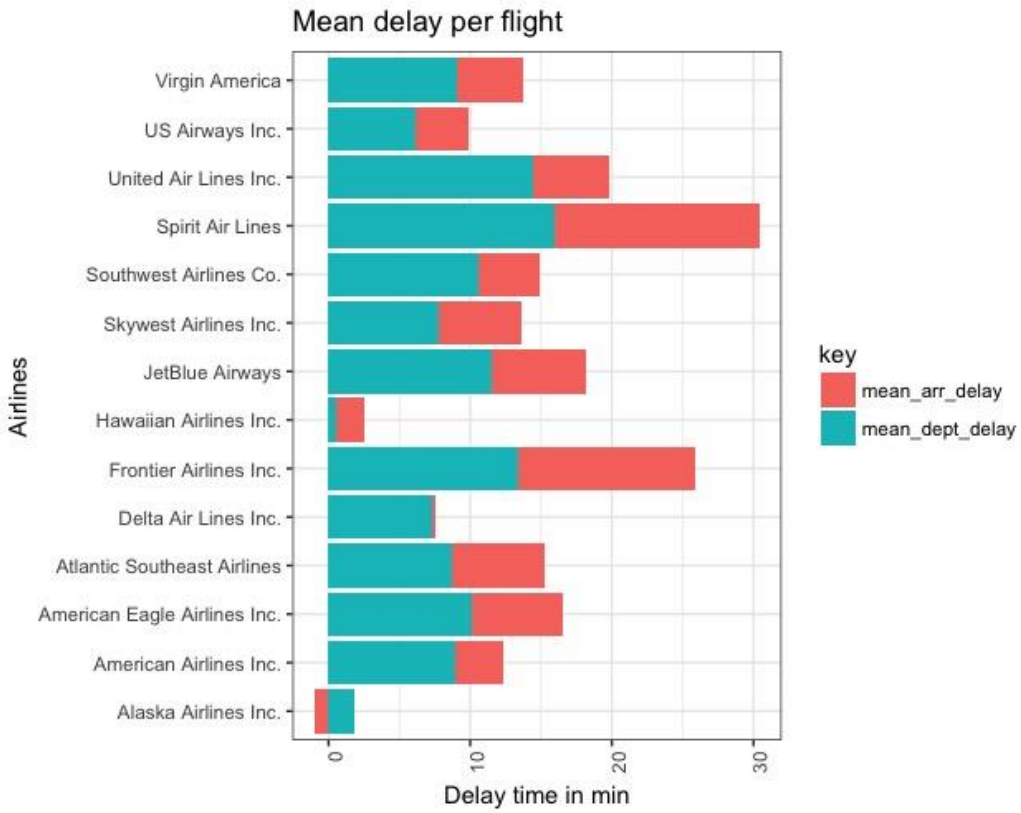
```
> cor.test(df_flights$DISTANCE, df_flights$ARRIVAL_DELAY)

        Pearson's product-moment correlation

data:  df_flights$DISTANCE and df_flights$ARRIVAL_DELAY
t = -60.84, df = 5714000, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.02626311 -0.02462430
sample estimates:
        cor
-0.02544372
```

The coefficient is -0.025 which verifies that there is no significant correlation.

One interesting fact here is that arrival delay for most flights (I don't know what Hawaiian airlines is upto) is less than the departure delay, which can mean that airlines are recovering from departure delay during the flight. So, in a way arrival distance can have a relation to the combination of departure delay and the flights distance. This relationship can be further explored by doing a linear regression.

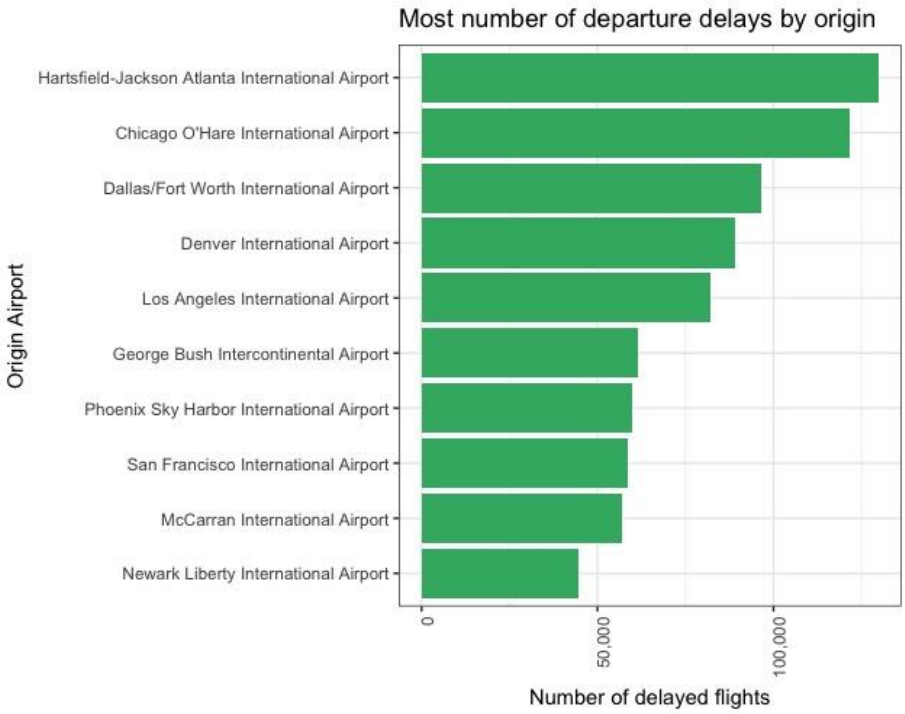


Submitted by: Balram Sidh

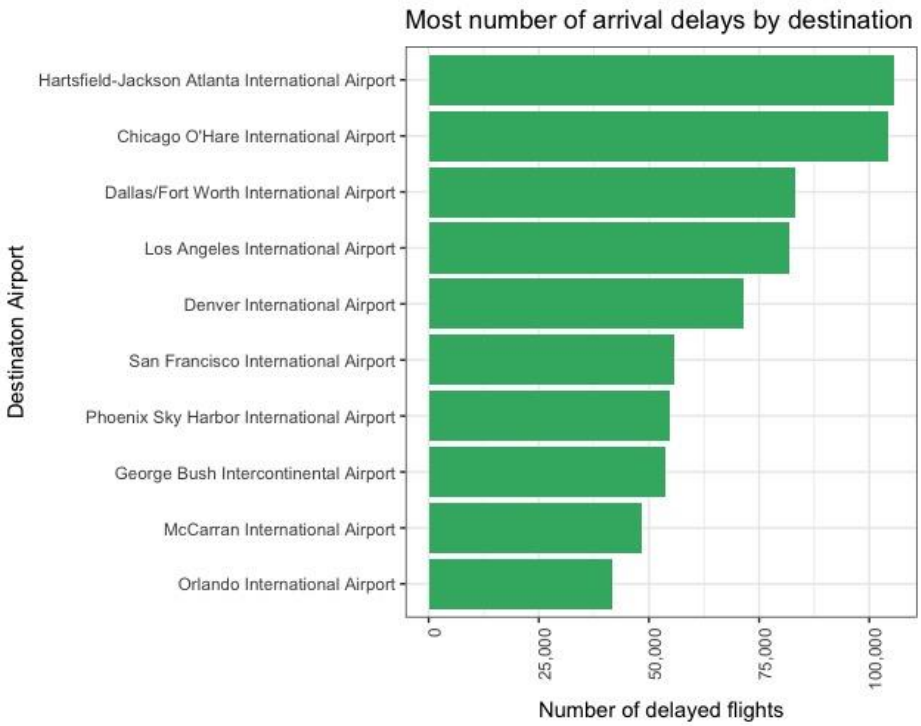
2. Which airports are part of the most delayed flights (either source or destination)?

Note: I am considering airports for which **most number** of flights got delayed.

Case one: Origin Airports where most number of departure delay has occurred.



case two: Destination airports where most number of arrival delay as occurred.



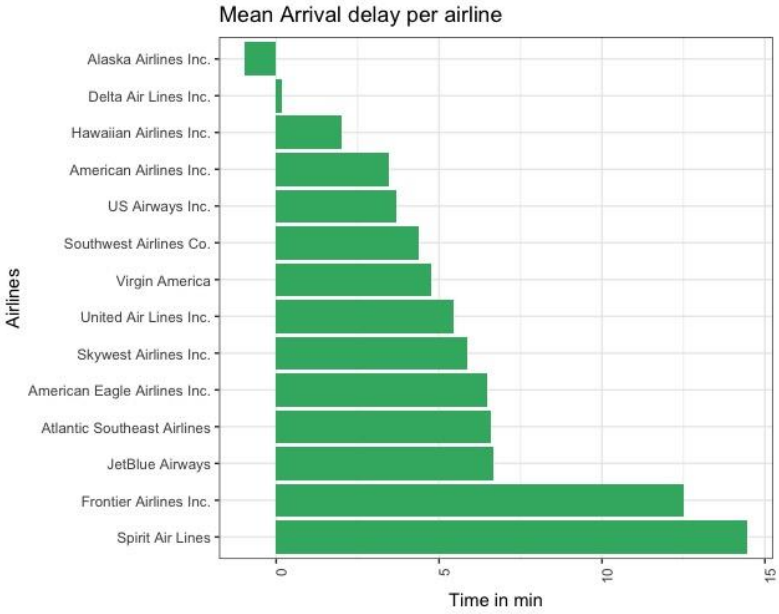
Submitted by: Balram Sidh

From the above two graphs it is evident that most number of flights are getting delayed at:

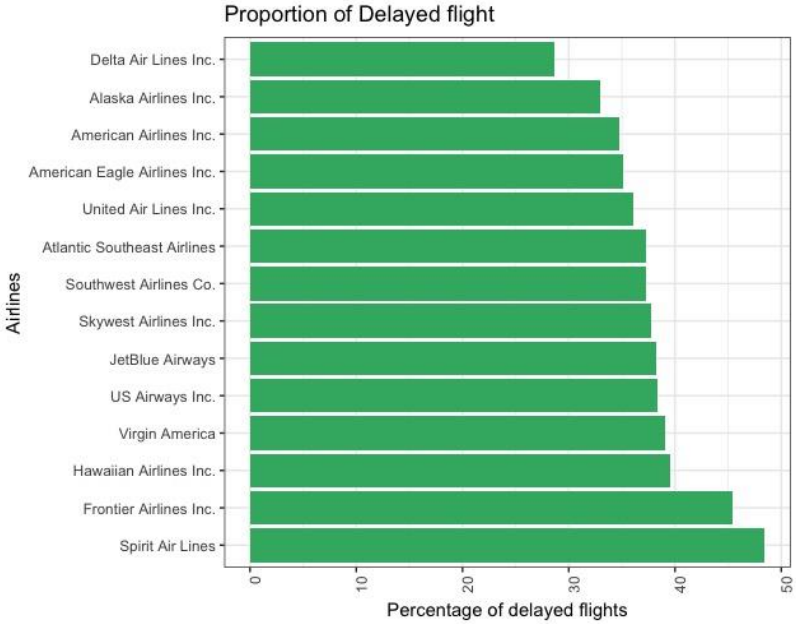
1. Hartsfield-Jackson Atlanta International Airport
2. Chicago O'Hare International Airport
3. Dallas/Fort Worth International Airport

3. Which airline should you fly on to avoid significant delays?

From the below graph, it is evident that Alaska airline is the safest best when it comes to avoiding delays.



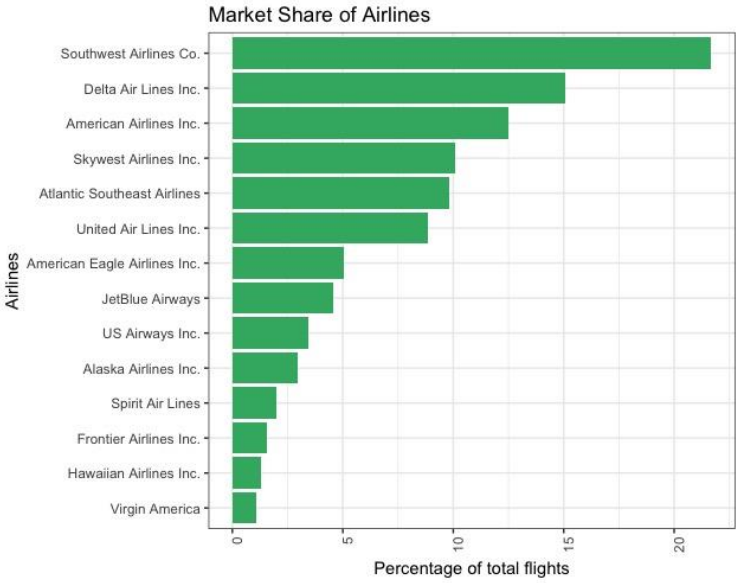
To further strengthen the argument, below graph shows the proportion of number of delayed flights and total flights per airline, again Alaska is a close second to Delta airlines.



Submitted by: Balram Sidh

4. Are there any interesting observations you can make from the dataset?

Market share or airlines:



Busiest route: