Analyzing flight disruption patterns using Spark framework

Abstract

This report aims to characterize how levels of activity in air travel correlates on-time performance of aircraft in the US. It is discovered that the (1) seasonality factor is a contributor to delays, with lower average delays during the post-summer low season; (2) levels of activity on a particular route and airport contributes to the probability of delay as characterized by a positive relationship between probability of delays and flights per route/per airport. (3) It is worth noting that although larger airports experience a higher chance of delay, the delays are on average shorter than at smaller airports.

Background

According to the US Bureau of Transportation Statistics (BTS), about 20% of all US domestic flights in the past 5 years have been delayed for over 15 minutes or more[1]. This phenomenon is estimated to have caused an aggregated loss in GDP of $31.2 billion in 2007 alone in a study commissioned by the Federal Aviation Administration[2]. In light of such a situation, flight on-time performance should be analysed for patterns leveraging relevant data-sets.

Currently, OTP data of 11 major airlines are logged by the BTS on a minute by minute basis and the full dataset is published regularly[3]. However, existing portals do not allow for any cross examination of flight performance, for example in minutes delayed by the level of activity in an airport. These features and their derivations can be insightful and useful in future predictive model building, and should be opened to the public. This project therefore aims to:

1) Analyse the OTP dataset using Spark to assess and identify how the different levels of activities in airspace, airports and routes affect the on-time performance of aircrafts

2) Explain the results and compare the relevant performance of these metrics

Technical Framework

Data

2 sources of data are used in this project, namely (1) Flight on-time performance data from the Bureau of Transportation Statistics, with a total 5,660,970 records from the period between October 2016 to September for airport and route analysis; and 11,436,737 records from the period between January 2015 to December 2016 for seasonal time analysis. (2) Airport data from openflights.org, which includes 7184 records of airport information from across the world.

Development Environment

Due to challenges in finding a platform that would allow both easy experimentation as well as production scalability, the team has settled on using the free-tier online Spark platform provided by Databricks in the data discovery phase, and a custom Spark cluster deployed on Amazon EC2 for production using the Flintrock (0.6.0)[4] deployment tool and Jupyter notebook.

[1] https://www.transtats.bts.gov/HomeDrillChart.asp
[2] http://www.nextor.org/pubs/TDI_Report_Final_11_03_10.pdf
[3] https://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236&DB_Short_Name=On-Time
[4] https://github.com/nchammas/flintrock/tree/v0.6.0

Storage-wise, although 6GB local storage is provided by Databricks for free tier users; and most instances on EC2 offer storage considerably larger than the data source. The team has elected to host the datasets on an Amazon S3 bucket for consistency and availability across the different platforms.

Table 1 Specification of platforms used

Specifications	Databricks Community	EC2 Cluster
Spark Version	2.2.0	2.2.0
Workers	1	8
Memory per Worker (GB)	6.0	12.2
Instance Type	-	m3.xlarge
Availability Zone	us-west-2c	ap-northeast-1a
AMI	-	ami-3bd3c45c

A number of additional libraries were also installed/used in this project, they include (1) the *plotly*/matplotlib API for certain tables and graphs; (2) the *numpy* library for generating multi-dimensional arrays; and (3) the *scipy* library for general statistics and (4) the *Datetime* library for timestamp-string conversation to modify the date for comparison.

Results and Discussion

Overall Delay Performance by Time

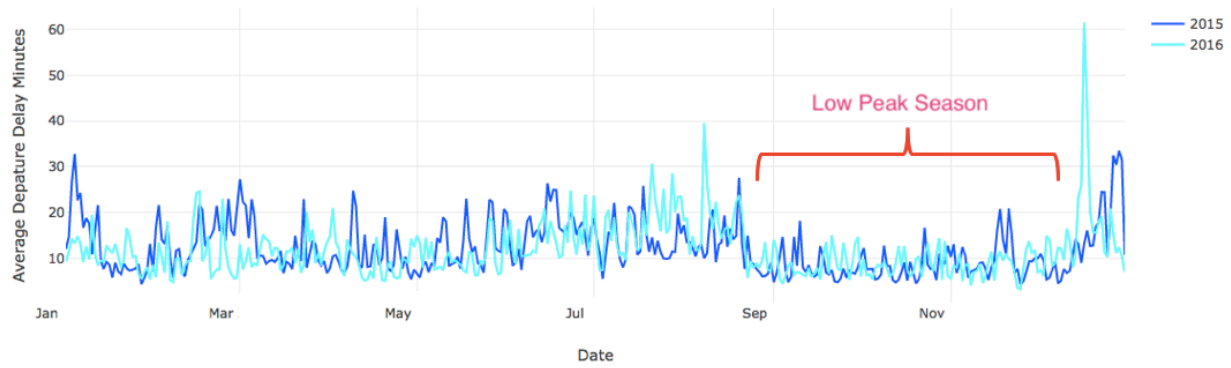

Figure 1 Average delay time per day

Figure 1 shows how delay performance varied between 2015 and 2016. It can be observed that the average flight delay is between 10-20 minutes per flight for most of the year, and between the start of September and the end of November, a lower average period can be observed with daily averages no higher than 18 minutes. In addition, it is observed that the average delay times increase rapidly during the end/start of year periods.

Explanation-wise, the low period coincides with the post-summer low-season and there is no national holiday in US between Labor Day and Thanksgiving Day which is on early September and late November every year[5]. For spikes in delay times, the long holiday of Christmas and New

[5] https://www.officeholidays.com/countries/usa/2015.php

Years are likely causes for the increase in average delay times. Therefore, although inconclusive, it is highly likely that length of delay correlates with air-travel seasonality.

Delay Performance by Route

Every route consumes air capacity. This consumption could lead to the saturation of airspace and delays might result. This part aims to study the effect of route traffic on delay.

Per Route Levels of Activity

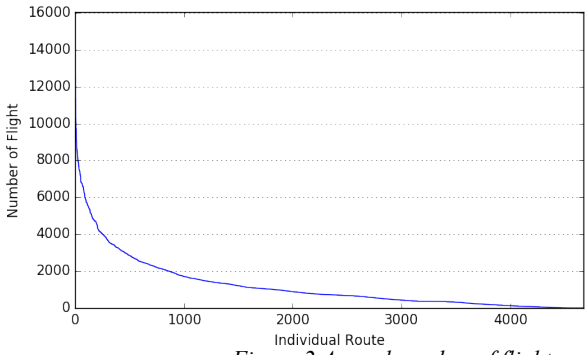

Figure 2 Annual number of flights per route

Table 2 Top ten busiest routes

Route	Flight per Year
SFO-LAX	16038
LAX-SFO	15715
JFK-LAX	12632
LAX-JFK	12629
LAX-LAS	11332
LAS-LAX	11328
SEA-LAX	10262
LAX-SEA	10254
LGA-ORD	10139
ORD-LGA	9983

Figure 2 shows the distribution of the number of flight for different unique routes. As shown in the figure, over 3000 routes have less than 2000 flights, for which the average percentage of delay would become very diverged due to small sampling size. Therefore, only the top 500 data will be selected below.

On-time performance to Levels of Activity

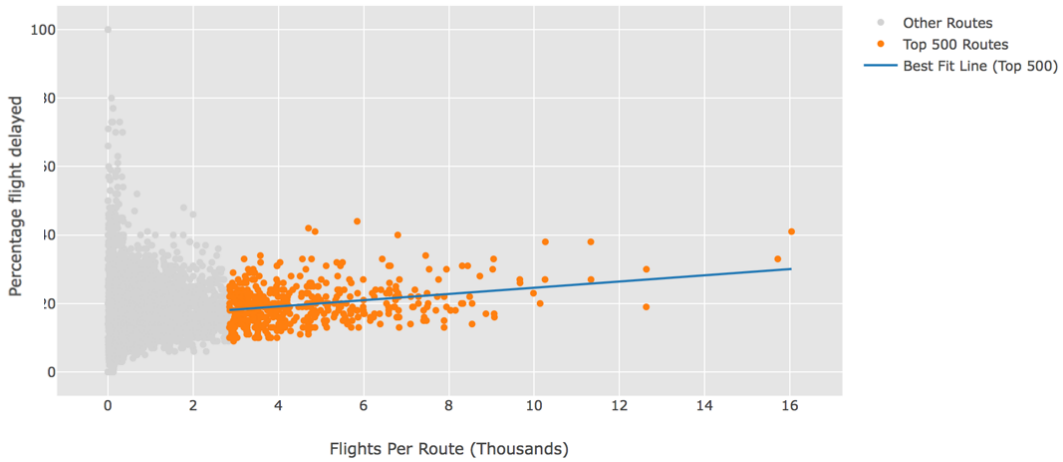

Figure 3 Flight per route and percentage flights delayed for the top 500 routes

Figure 3 shows the average delay of the 4671 routes during the flight period. The *x*-axis reflects the levels of activity of each route in terms of flights flown while the *y*-axis reflects the probability

of that route experiencing delays. After applying least square method to the top 500 data and finding the best fitting curve, the R is found to be 0.295562 while the slope is found to be 0.908696. The slope represents that for every thousand flights increase in the route, the chance of having delay would increase by 1% although the correlation is not completely clear as characterized by a low to medium R value.

Delay Performance by Airport* - *Bigger airport – Bigger delays?

Per Airport Levels of Activity

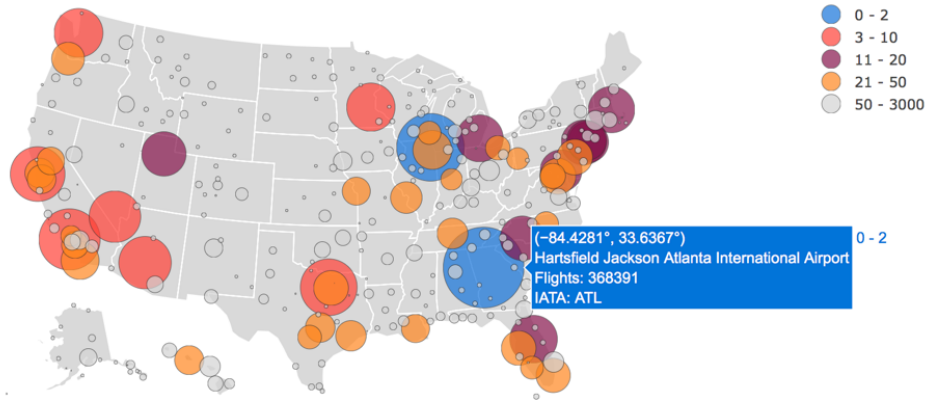

Figure 4 Number of Flights per airport between October 2016 and September 2017

In terms of airports excluding cancelled flights, the largest domestic airports in the US is Atlanta international airport with 368,391 flights, closely followed by Chicago O'Hare international airport with 257,802 flights. These statistics are within expectations as the two airports are gateways for East-West travel and Atlanta airport is the largest regional hub for the southeast region.

On-time performance to Levels of Activity

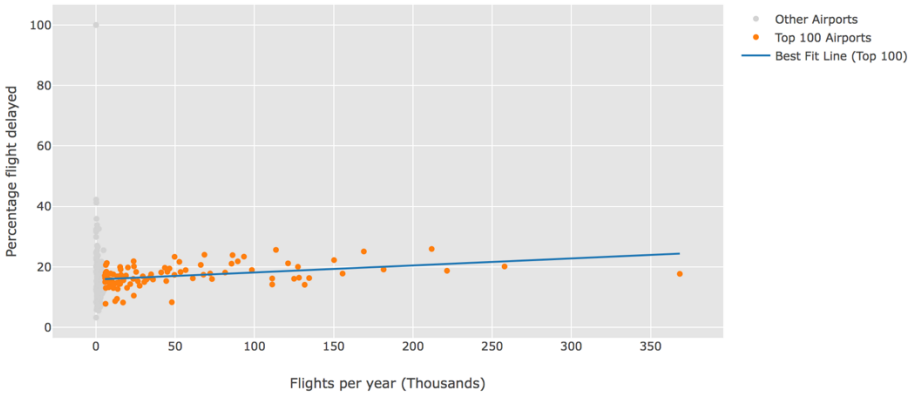

Figure 5 Flight per airport and percentage flights delayed for the top 100 airports

Figure 5 shows how the probability of flight delays at a particular airport varies with the number of flights handled per year. Since small airports with less than 10 thousand flights a year exhibits very high variance in their on-time performance, they are excluded for the analysis. The best fitting curve plotted using the least square method yielded an R of 0.401154 and a slope of 0.0232204653628, indicating that there is a loose correlation between how busy an airport is and its proportion of flights delayed. A possible explanation to this phenomenon could be that certain variables affecting the OTP of an airport do not scale as an airport grows larger, such as congestion in air space which causes delay.

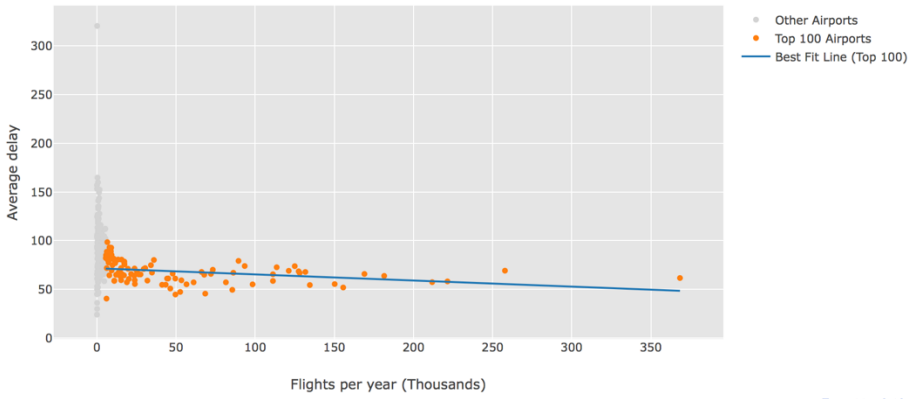

Figure 6 Domestic flight count and average delay time for the top 100 airports

Figure 6 shows how average delay time varies with the level of activity an airport. It is shown that the relationship between the two variables, as indicated by an R of -0.337584 and a very flat slope of -0.0626816674434 is weak. But an interesting explanation could be that although large airports experience a higher probability of delay, the additional capacity helps flights fly out quicker, reducing the average delay times.

Conclusion

In conclusion, the thesis that levels of activity at various stages of air travel correlates to flight delays is justified, as both seasonal factors (which is expected), to airport size and flights per route correlates positively with the delay probability. It is particularly worth noting that although flights from larger airports are more likely to be delayed, with a stronger correlation than levels of activity on the route level, the length of the delay is usually shorter than at smaller airports. Meaning that on average passengers would more likely experience a delay at O'Hare than a regional airport, but the delays are likely to be shorter.

Further work should be done to see how these factors can be further broken down to it's elements, for example on the types of delay and with the inclusion of weather factors which this report does not cover.

Appendix

Github Repository (Merged): git@github.com:sjjchung/cloud-project.git