KTH Electrical Engineering

# Flight delay data analysis

Individual project

INDRE TAUROSEVICIUTE AND SHIVAM SAINI

Stockholm 2019

ID2221 HT19-1 Data-Intensive Computing
School of Electrical Engineering and Computer Science
Royal Institite of technology

# Flight Delays in the US

Indre Tauroseviciute and Shivam Saini

October 31, 2019

**Abstract**

## 1 The problem

Since the invention of an airplane, people got a chance to travel around the globe fast and comfortable. Usually trips are planned in advance and flight time has an impact on booking for accommodation or personal schedule. This means that, in the event of a delay, it costs money and frustration rearranging the stay. This can be caused by a number of factors like carriers, weather, aircraft, etc. which leads to a certain time of a delay. Sadly, very few airlines inform about the cause and so passengers are left waiting in uncertainty.

By analyzing flight data from the recent years, assumptions for the reliability of airlines can be made.

## 2 The solution

There are thousands of flights in US every day and so to analyze it big data processing algorithms should be used. For this particular case we chose Spark to work the data flow of flight on-time performance and causes in the US. The descriptive analysis shows flight delay patterns within years and airlines.

### 2.1 The dataset

The data selected for this project is taken from The U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics (BTS) [2]. It has a data collection since June 2003 and updates it every month as soon as the Air Travel Consumer Report is released. It includes information about the date of flight, origin airport, destination airport time delay, time performance, destination and causes. Data for each month is stored in a CSV file and full dataset is roughly 5.5 GB in size [3]. We will be using the data for the previous 5 years (since 2014) for this project.

For the purposes to understand the data and demonstrate the outcome in comprehensive way, additional dataset was used. It contains all the IDs from the US airlines corresponding to their names.

### 2.2 Analysis methods

Our goal was to get useful insights from the data and to understand it. The flight performance data from the previous five years, brought us some questions that we focused on:

- When is the best time of day to fly with minimise delays?
- Which is the best day of the week to fly with minimise delays?
- Which is the best day of the month to fly with minimise delays?
- Which is the best month of the year to fly with minimise delays?
- Which airlines were the most/least reliable throughout the year?

Since the dataset used was big in size CSV file, analyses required open-source distributed data streaming framework. We chose Spark for that reason. Meanwhile, the output is also CSV file, but small in size, so we used Excel to make visualizations. To conduct the experiment, the analyses was divided into following steps.

1. The datasets as CSV files were fetched and fed into the Spark. Cassandra and Kafka were considered, but Spark seemed to be enough.
2. The dimension reduction across the data was run to find the desired results. There were 8 methods read in DataFrame used:
    - time_of_day(flightDF) - times were ranged within an hour and number of delays (smaller than 30min and bigger than 30min)calculated;
    - day_of_week(flightDF) - all delays per week day calculated;
    - day_of_month(flightDF) - all delays per month day calculated;
    - month_of_year(flightDF) - all delays per month, per year calculated;
    - dep_performance_airline(flightDF) - all departure delays per airline calculate;
    - dep_performance_airport(flightDF) - all departure delays per airport calculate;
    - arr_performance_airline(flightDF) - all arrival delays per airline calculate;
    - arr_performance_airport(flightDF) - all arrival delays per airport calculate.
3. The results were written in CSV and visualized, to compare them and to get useful insights. This was done by using simple MS Excel functions.

## 3 The results

The results of the analysis will be given accordingly to the research questions from the section 2.2.

## 3.1 The time of the day

The research showed that the time of the day with the least chance of the delay since 2014 is the morning, specifically at 4am to 6am. The number of departure delay increases throughout the day and reaches a peak at night between 11pm and 3am (fig 1).

of day.jpg

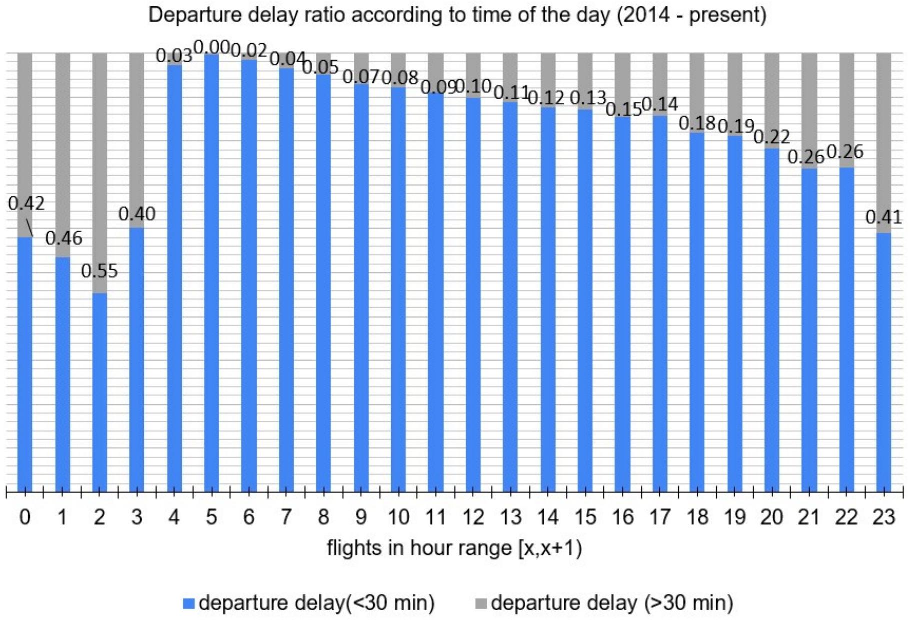

Figure 1: The times of the day with flight delay ratio

## 3.2 The day of the week

The analysis showed that there is no significant day of the week when the flight departure delay chances increases. However, Saturday seem to be slightly more reliable and Friday - the least reliable (fig 2).

## 3.3 The day of the month

When it come to the day of the month, there is no particular time when the flight is expected to be least late. However, there is a slight chance, that if there were more delays in the first half of the month, in the second half, there will be less and vice-versa. This is just an assumption which should be researched in further analysis (table 1).

## 3.4 The month of the year

Meanwhile, there is a pattern in flight delays when it comes about months themselves. Study showed that it is less likely for the departure delay in October and November in the US, while the number of delays increases in the summer (June and July) (table 2).

of the week.jpg

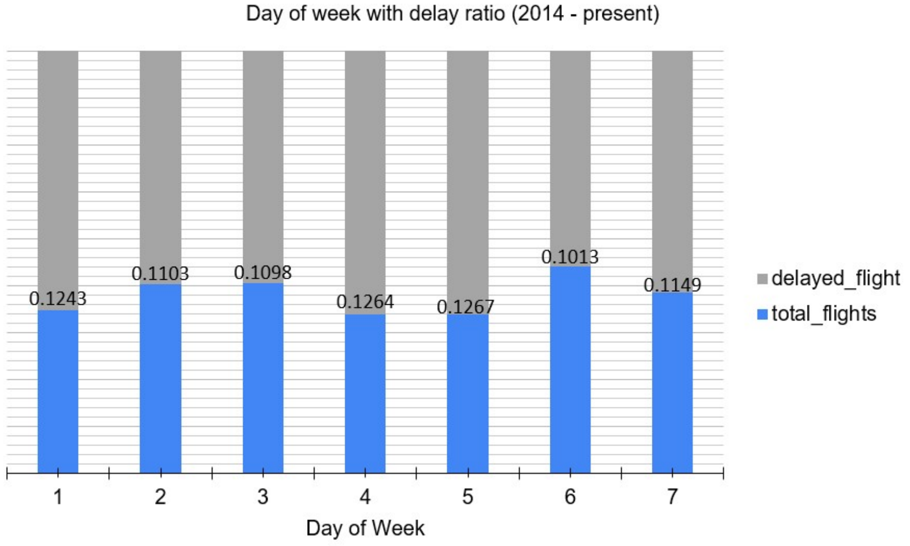

Figure 2: The days of the week with flight delay ratio

| Month | Delay ratio | The worst day | Delay Ratio | The best day |
|---|---|---|---|---|
| 1 | 0.195 | 5 | 0.069 | 27 |
| 2 | 0.157 | 20 | 0.053 | 29 |
| 3 | 0.165 | 1 | 0.081 | 19 |
| 4 | 0.140 | 3 | 0.077 | 23 |
| 5 | 0.143 | 19 | 0.080 | 7 |
| 6 | 0.193 | 19 | 0.102 | 4 |
| 7 | 0.187 | 23 | 0.080 | 4 |
| 8 | 0.172 | 1 | 0.072 | 30 |
| 9 | 0.115 | 18 | 0.056 | 15 |
| 10 | 0.129 | 9 | 0.057 | 17 |
| 11 | 0.137 | 26 | 0.063 | 8 |
| 12 | 0.209 | 27 | 0.068 | 6 |

Table 1: The day of the month having a flight with highest and lowest delay possibility since 2014

| Year | Delay ratio | The best month | Delay Ratio | The worst month |
|---|---|---|---|---|
| 2014 | 0.093 | 9 | 0.171 | 1 |
| 2015 | 0.070 | 10 | 0.150 | 6 |
| 2016 | 0.076 | 11 | 0.153 | 7 |
| 2017 | 0.066 | 11 | 0.144 | 6 |
| 2018 | 0.092 | 10 | 0.146 | 7 |
| 2019 | 0.106 | 3 | 0.161 | 6 |

Table 2: The month of the year having a flight with highest and lowest delay possibility since 2014

### 3.5 The most/least reliable airlines

When analysing reliability of airlines, it was noticed that there is no significant difference them being late in departure and arrival. In other words, the same airlines showed to be most / least reliable in both arrival and departure. All 6 years, Hawaiian Airlines were on time in most of the flights(table 3), meanwhile few airlines were competing for the title of being the most unreliable (table 4).

| Year | Arrival delay Ratio | Departure delay Ratio | Carrier name |
|---|---|---|---|
| 2014 | 0.028 | 0.022 | Hawaiian Airlines Inc. |
| 2015 | 0.041 | 0.031 | Hawaiian Airlines Inc. |
| 2016 | 0.030 | 0.024 | Hawaiian Airlines Inc. |
| 2017 | 0.046 | 0.038 | Hawaiian Airlines Inc. |
| 2018 | 0.042 | 0.034 | Hawaiian Airlines Inc. |
| 2019 | 0.043 | 0.035 | Hawaiian Airlines Inc. |

Table 3: Airline the best arrival performances each year since 2014

| Year | Arrival delay Ratio | Departure delay Ratio | Carrier name |
|---|---|---|---|
| 2014 | 0.166 | 0.163 | ExpressJet Airlines LLC |
| 2015 | 0.192 | 0.181 | Spirit Air Lines |
| 2016 | 0.154 | 0.153 | Spirit Air Lines |
| 2017 | 0.184 | 0.187 | JetBlue Airways |
| 2018 | 0.196 | 0.203 | Frontier Airlines Inc. |
| 2019 | 0.195 | 0.200 | JetBlue Airways |

Table 4: Airline the worst arrival performances each year since 2014

### 3.6 Other studies

By using this dataset, we were willing to discover main delay reasons for the flights, however, it did not work, because of the data containing large number of empty cells or being violated. This disrupted the results and so was decided not to show them in the report.

Furthermore, we had data about the airports and were planning to provide analysis regarding this parameter too. This did not work because of different airport sizes. Some airports in the US are a lot smaller than others, therefore the results seem to be violated. With more aggregation and research, answers to these questions could be found..

## 4 Running the code

The code for this research was written in Scala using IntelliJ IDEA platform to run the code. This interface also supports Spark, and dependencies for spark were added through sbt build tool. To run the code, user needs to add the dataset(link here [1]) into the project's resource folder first. The code for the research can be found in the attachment.

## References

[1] Link to data and the code. https://drive.google.com/file/d/1sqPIx9ni1zc6_PvXKDEifUXCUANl_tA4/view?usp=sharing.

[2] BUREAU OF TRANSPORTATION STATISTICS. Airline on-time statistics and delay causes. https://www.transtats.bts.gov/OT_Delay/OT_DelayCause1.asp.

[3] BUREAU OF TRANSPORTATION STATISTICS. Data. https://www.transtats.bts.gov/Fields.asp?Table_ID=236.