Flight Delays Data Story

Contents

[Background and Motivation 3](#_Toc462567408)

[Client Benefits 3](#_Toc462567409)

[Data 3](#_Toc462567410)

[What does the data tell you 4](#_Toc462567411)

[Data at a glance – 2008 flights 4](#_Toc462567412)

# Background and Motivation

Every year approximately 20% of airline flights are delayed or cancelled, resulting in significant costs to both travelers and airlines. As a frequent traveler, we are always apprehensive about my flight getting delayed and we keep wondering if only we knew what causes the delay and predict all the factors with reasonable confidence, then not only the airlines will have more control over their planning and mitigate all the frustrations caused to the traveler. Historically,

Flight delays cost the airline industry $8 billion a year due to increased spending on crews, fuel and maintenance. (Source: FAA 2010)

Delay cost passengers nearly $17 billion. (Source FAA 2010)

## Client Benefits

Accurately predicting airline delay will allow

* Client Airline to proactively identify potential causes and find ways to alleviate such causes
* Passengers to be mentally prepared and reduce stress and anxiety due to uncertainty of delay

Before we assess the client benefits, we would like to see what the dataset chosen for this analysis tells you.

# Data

We used four different data sources

1. The official flight database for every domestic flight in the US
2. Historical weather data
3. Airport information with geodata and names (e.g. for visualization and interpretation of results)
4. Information about aircraft models

**Data Sources:**

**We have downloaded flight data from the website below for 2008 as our dataset for analysis**

* The 2008 departure and arrival data is from the American Statistical Association
  + <http://stat-computing.org/dataexpo/2008/the-data.html>

# What does the data tell you

## Data at a glance – 2008 flights

Total flights in 2008 – 7 million

Average daily flights – 19,204

Average flight time – 2 hours 7 minutes

Average flight distance – 726 miles

Airline with highest daily operations – 3252 flights (SouthWest)

Airline with highest departure delay – 14.1 minutes (United)

Airline with highest arrival delay – 12.6 minutes (American)

Let us now explore some of the potential causes of delays.

* + Do some airports systematically generate delays independently of the company?
  + Are some airlines just less timely than other airlines?
  + What is the influence of the time of day?
  + Are some airports more prone to weather delays than others?

Below, see the data visualizations and the explanation next to it.

|  |  |
| --- | --- |
|  | **2008 Departure Delay and Arrival Delay Distribution**  We notice a much higher probability of short delays - actually negative, so advances - for departure delays and a wider distribution (in minutes) for arrivals. Notice the long right-hand tails. Some flights are delayed for very long times, over two hours. On the other hand, the delays are centered on just below zero. In both cases, the mode of the distribution is less than zero, meaning most of the flights leave from gate and arrive at gate even before the published schedule time of departure and arrival. As we will show below, the longer delays cancel out the shorter negative delays (advances), leading to average delays that are above zero.  The x-axis for the two plots are to scale. As a result, we can see that the arrival delay distribution, compared with the departure delay distribution, leans toward left. A flight delay is defined by the schedule time of an event compared against the actual time of the event. Airlines usually put extra buffer time in a flight to ensure on-time arrival. Therefore, the departure delay and arrival delay distributions difference indicates that some departure delays are recovered during the flights due to the extra amount of time embedded in the flight time between two airports.  **Next**, we consider the impact of month on the delays. We would expect that winter months have the longest delay. A column chart with departure and arrival delay in minutes plotted by month is the most effective way to see the potential effects of month. |
|  | **Delays by Month**  For both departures and arrivals, the impact of December is clear - the highest delays are in that month. On the other hand, September, October and November are the months with the least amount of delay. For the summer, June and July are marked by higher delays. Also, February posts high delay values as well. The reason for winter's high delay values is probably because of snowstorms in the northeast of the US. Also, in summer, thunderstorms in Dallas Forth Worth (DFW) and Chicago areas can cause high delay impact to the rest of country. A snowstorm/storm may only affect operations at an airport or two. However, delay propagation, which marks as the major contributor for flight delay, can cause ripple effects on delay to downstream flight operations.  **Next**, we also think that the time of day should have an impact. Normally, flight delays cumulate throughout the day through a knock-on effect, where delayed flights provoke other delays because of tight schedules and runway congestion. We plot the mean delay by hour of day in a column chart. |
|  |  |
|  | **Delays by Hour**  We see a marked "V" shaped decline in delay with the lowest delays in early morning hours. Both departure and arrival delays accumulate from earlier morning reaching their peaks in the evening hours. For departure, the highest mean delay is during prime-time of 18:00 to 21:00, and for arrivals, it is slightly later (the average flight duration is a few hours) and peaks at around 22:00. The increasing of flight delay by the hours of the day is mainly caused by flight delay propagation. Although a flight is built with scheduled buffer time for unforeseeable flight delay during the flight operations, it is not sufficient to cover all types of delay. As a result, if a flight is delayed, the next flight has to wait for the late arrival flight to be ready before it can be operated. Hence, flight delays for both departure and arrival flights do increase over time.  **The variation in mean delay by hour that we see implies that Hour of Day should be a good predictor of flight delay.** |
|  | **Delays by Month for select Airports**  We see differences between airports - Chicago O'Hare and San Francisco are similar to the overall profile for mean delay at all airports, with higher delays in December and January and a midsummer bump. On the other hand, Boston Logan shows lower mean delays in the beginning of the year and more delays in all 3 summer holiday months. New York LaGuardia has more delay in the springtime, with February, March and June with a higher mean delay than in December. In all locations, December is a month with higher than normal delays.  Given that Northeast Corridor do receives significant amount of rain/snow, further analysis is required for examining the cauase and distributions (temporal and spatial relationship) of the flight departure delay. |
|  | For arrival delays, we see four distinct peak months for Chicago O'Hare. December, January, February and June. The latter may be because of holiday travel. The former three may be entirely weather related since these three months have the worst climatic conditions. San Francisco is similar to Chicago again. The peak delay is in January possibly weather related (bad winter conditions in January 2008). Again the springtime peak in New York shows the longest delays from February to June. In all locations, the end of year only has minimal delays in October and November, with delays rising back up in December.  Next we would like to consider the impact of Hour of Day across the airports. We repeat the analysis for departure and arrival delay by hour. |
|  | **Delays by Hour for select Airports**  It is apparent that the early morning (late night) hours are linked with longer departure delays for Chicago and Boston. Midnight and 1:00 am have large delays at ORD, and for BOS, the delay gets worse until 3:00 am and then reverses. Generally, in all four airports, we see the delays accumulating throughout the day until 19:00 r 20:00 and then dropping off. This profile is most apparent for Chicago and New York. On the other hand, in San Francisco, the delays start mid-morning and remain stable throughout the afternoon. |
|  | The early morning hours have peak delays again for Chicago and Boston. In both cases, this occurs at 2:00am, with delayed flights arriving beyond schedule. Generally, as the day wears on, arrival delays increase up until 22:00. It is unclear weather schedule constraints cause the early morning delays (less staff in the early morning) since delays drops off before peaking again. Only San Francisco has a distinct profile with more stable delays throughout the day, with a morning dip from 5:00 amd to 9:00am.  It is obviously that the departure and arrival delay profiles for an airport follow similar shapes. However, the arrival delay distribution shifts to the right for around 2 hours from the departure delay. The average US scheduled flight duration, from 2008 data, is 128 minutes with an average flight distance of 726 miles. The 128 minutes flight duration explains the reason why arrival delay distribution is roughly 2 hours beyond the departure distribution.  Next we look generally at delays linked to carriers, for all airports. |
|  | **Delays by Carrier**  For flights with one of 20 unique carriers, average flight delays vary considerably. The analysis is affected by the number of flights for each carrier. Some, with a small number of flights, like Aloha Airlines (AQ) or Hawaiian Airlines (HA), have lowest mean delays. Both AQ and HA mostly operate flight between Hawaii and Continental United States. Since the flights are mostly 5-6 hours, these flights are usually exempt from Ground Delay Programs (GDPs) when an airport encounter inclement weather conditions. This means that no delay is imposed by air traffic management system. Therefore, the average delays for these two carriers can be quite low.  What if we look at just one airport - Chicago? Does carrier lead to better prediction in this case? |
|  |  |
|  | **Delays by Carrier in Chicago**  Here we see that certain carriers, the small ones, have large departure and arrival delays, such as 9E B6, CO, EV and YV. Mainstream carriers like Delta, Northwestern and US Air have smaller delays. Also apparent is that the profile is similar for both departure and arrival delays.  Let's look at the situation for Boston. |
|  | **Delays by Carrier in Boston**  The departure and arrival delay for XE and YV is much longer. With the threshold of 15 minutes mean delay, you can expect a delayed fligh if you are travelling wih CO, XE or YV. In contrast, travelling with EV or US means you can expect to be on time.  However, data exploration may only reveal partial fact at different airports. However, the true causality still requires additional analysis and evaluation using domain judgement. |

|  |  |
| --- | --- |
|  | **Correlations:**  Here we look a little more closely at the attributes themselves and how they are related to the classes. We used scatterplots and plot each variable against delay. Then, since we have a large number of attributes, the best would be a parallel coordinates plot of the attributes, coloring by class.  **Results:**  The scatterplots does not show a clear correlation. However they need close reading.   * The highest correlation coefficient is for **ArrivalHour**, because of the time of day effect that was noted earlier. This could be compounded by the early morning delays - which cause a spike at the lower end in the scatterplot. * **Carrier** and **Month** show some correlation which confirms what we say in the column plots above on for the full dataset. * The bottom two plots reveal that the **Duration** and **Distance** characteristics. These may be correlated to delay, but in fact negatively - because with longer distance there is less delay since the flights are prioritized for take-off and the longer duration allow them to make up tardy departures. So we see that long-haul is less likely to be delayed. |

# Conclusions:

From the above visualizations, we can make several hypothesis as below to assess what features should be included in predicting flight delays.

* Hour of day
* Month
* Carrier
* Origin
* Destination
* Distance
* Duration