Analyzing On-Time Performance of Commercial Flights in the United States Using MapReduce

Jason D. Stock

October 29, 2019

Question 1 & 2

↪ *What is the best/worst time-of-the-day/day-of-week/time-of-year to fly to minimize delays?*

For each sample in the dataset, the flight delay can be calculated by summing the arrival (`ArrDelay`) and departure (`DepDelay`) delay together. These could then be aggregated and averaged in terms of time-of-day, day-of-week, or month-of-year, to find best and worst times. In order to get an accurate representation for the scheduled time-of-day, various time intervals were considered for rounding times, including; 5, 10 and 15 minute intervals. As such, the most favorable results were seen using 15 minute intervals.

Parameter	**Best Time to Minimize Delays**	**Value**	**Worst Time to Minimize Delays**	**Value**
Time-of-Day	03:45	-8.285	20:00	26.186
Day-of-Week	Saturday	10.814	Friday	19.356
Month-of-Year	September	8.578.	December	20.612

Table 1: The best and worst time to minimize flight delays at different view of the year.

The results shown in the table not only agree with the calculations, but logically make sense. December is a busy month with various holidays, which is often an incentive for traveling. On the contrary, the number of flights and airport traffic early in the morning are often minimal. One might even depart early!

Question 3

↪ *What are the major hubs (busiest airports) in continental U.S.? Has there been a change over the 21-year period covered by this dataset?*

Major airports are considered to be busier if they have a greater number of incoming *and* outgoing flights than a prior. Therefore, it is possible to accumulate the count for each origin and departure IATA over the entire corpus. The top major hubs collected over all years ranked Chicago O'Hare International Airport (ORD) as the most trafficked, with Hartsfield-Jackson Atlanta International (ATL) and Dallas Fort Worth International (DFW) trailing close behind (Figure 1).

It was possible to collect each data sample as an `IATA_year` key pair to get an accumulated IATA count over individual years. Looking at the trends over the last 21-year period show more revealing results about the data. Specifically, the collection from 1987 and 2008 both show lower results, by a factor of roughly five, as compared to the next and previous years respectively. Additionally, ORD shows a steady lead over all years until 2003 when ATL is seen with an increase in the number of recorded flights. (Figure 2).

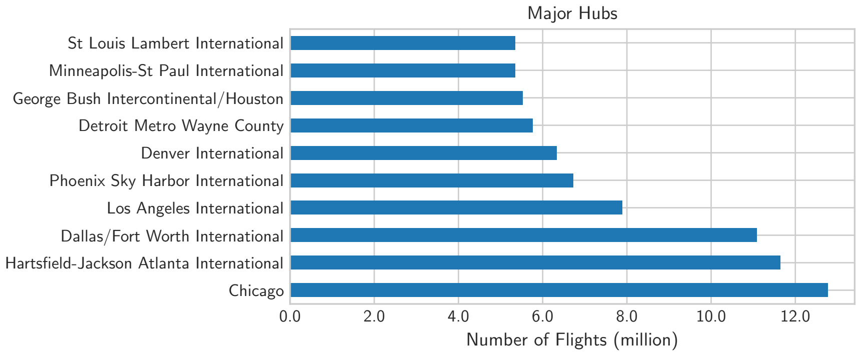

Figure 1: Top 10 major hubs in the United State.

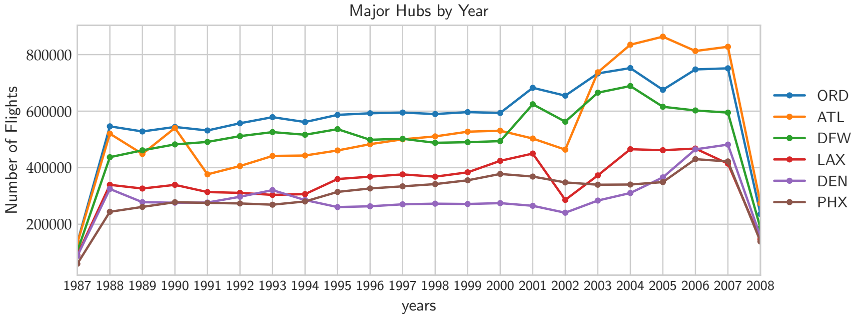

Figure 2: Top major hubs in the United State per year over a 21-year period.

Question 4

↪ *Which cities experience the most weather related delays?*

Accounting for whether or not a city had a weather delay for a specific flight was determined as true if that flight had a positive value in the `WeatherDelay` column. Each flight could then be reduced by the origin IATA, and merged with the `airport.csv` file to retrieve the name of the city and the summation for the total number of delays. The top 10 results are shown in Figure 3.

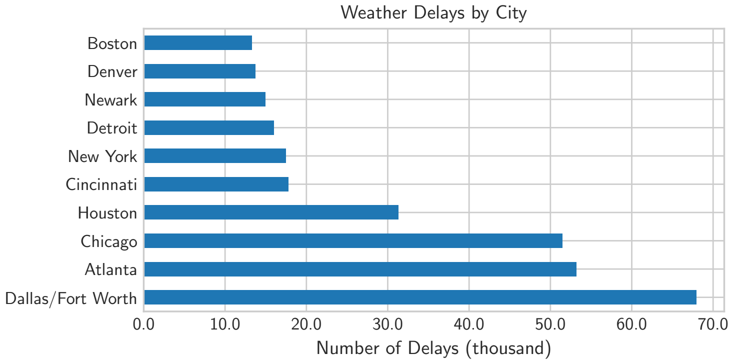

Figure 3: Top 10 cities in the United States with the highest number of weather related delay.

Question 5

↪ *Which carriers have the most delays? You should report on the total number of delayed flights and also the total number of minutes that were lost to delays. Which carrier has the highest average delay?*

Analyzing this question is three part, but it can be achieved using one MapReduce job. This is done by accumulating the count and occurrences of the carrier related flight delays. Only valid, non-empty, entries are considered for each carrier. A combiner is used to perform computations on intermediary outputs from the map phase. The reducer logic is similar in that it further aggregates the number of delays and delay time for data items with the same key.

For each of the three metrics, the top 10 carriers are shown. Firstly, the total number of delays are shown in Figure 4 with Southwest Airlines Co. seen with nearly 2× more delays over the others. However, being the top carrier is not exactly a great thing as it means the company is frequently reporting on delays. Although, this could be due to the large number of flights taken by this carrier.

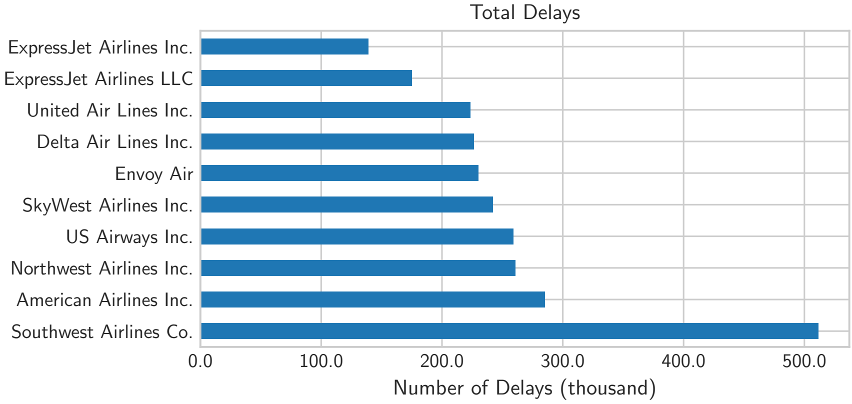

Figure 4: Top 10 carriers with highest number of delays.

It becomes difficult to determine what is meant by having a large number of delays for a specific carrier. For example, Northwest Airlines Inc. has fewer delays than Southwest Airlines Co., but what if these delays were all 5× the length? Therefore, the total delay times can also be quantified (Figure 5).

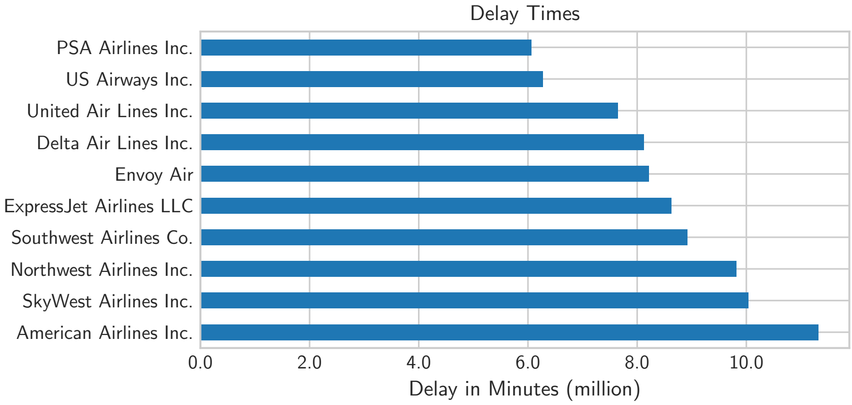

Figure 5: Top 10 carriers with the highest accumulative delay added up over a 21-year period

To further measure the delays of various carriers, the average delay time is accounted for over all delayed flights (Figure 6). This provides an interesting view into the carriers, especially Southwest Airlines Co. This is because they no longer appear in the top 10, and a list of other airlines do. It is possible that the scale of the airline influences the range of delay time. Whereas, another airline such as Mesa Airlines Inc. has fewer total, but significantly longer delays.

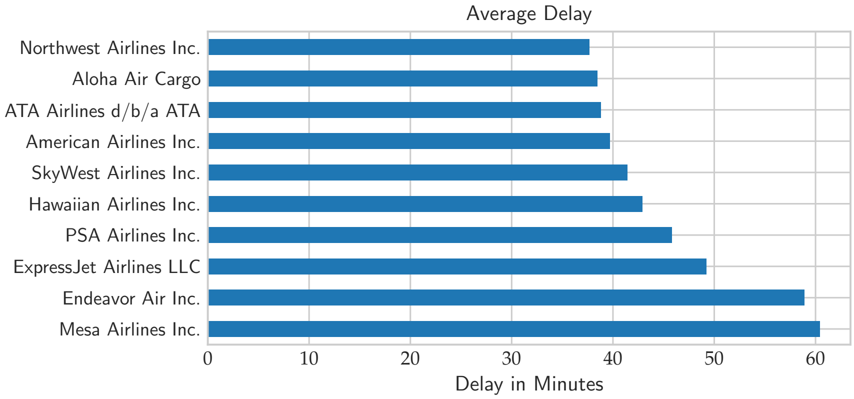

Figure 6: Top 10 carriers with the highest average delay time.

Over all evaluations, the following carriers appear: Northwest Airlines Inc., American Airlines Inc., SkyWest Airlines Inc., and ExpressJet Airlines LLC. As a result, it would be recommended to avoid these airlines when flying to minimize the probability of having a delayed flight.

Question 6

↪ *Do East or West coast airports have more delays?*

An airport is considered to be in a coastal region if it resides in a state bordering the Atlantic or Pacific Ocean. The states that have a shoreline on the east coast include: Maine, New Hampshire, Massachusetts, Rhode Island, Connecticut, New York, New Jersey, Delaware, Maryland, Virginia, North Carolina, South Carolina, Georgia, Florida, and Washington DC. Whereas the west coast only include four states, including: Alaska, Washington, Oregon, and California.

Luckily, the state abbreviation (e.g., NY, FL, AK, CA, etc.) can be used to check if an airport exists in a state that is found on the east coast, west coast, or neither. Then one of these associated labels can be tied to each of the IATA abbreviation in the `airport.csv` dataset. This allows the data to be reduced by IATA, and the total number of delayed flights to accumulate for each coastal region. The accumulation of occurrences follow a particular pattern based on the reason for delay. Specifically, flights with a departure delay (`DepDelay`) contribute to the origin airport, whereas the arrival delay (`ArrDelay`) would increase the total delay count for the destination airport.

Comparing the total accumulated values give some indication for which coast has more delays, but what if one coast has twice the number of flights as the other? This can be solved by computing the average number of delayed flights as the number of delays over the total number of flights for each region (Table 2).

Region	Total Number of Delayed Flights	Total Number of Flights	Percentage of Flights Delayed %
East Coast	36,739,109	80,308,569	45.75
West Coast	15,815,751	36,156,127	43.74

Table 2: Comparison of flight delays in the east coast versus west coast.

It is evident that the total number of delayed flights is nearly doubled in the east coast, but so does the total number of flights. However, as a percentage, it can be seen that the delayed flights on the east coast still exceed over the west.

Question 7

↪ *What are the aspects that you can infer from the LateAircraftDelay field?*

There are no clear links between flights in the dataset. Specifically, a flight sample does not contain information to the connecting flights either before or after. This is interesting because the `LateAircraftDelay` field relates to a flight being delayed due to the previous flight being delayed upon arrival. Therefore, the question that comes to mind is "where are these flights coming from?" For example, what are the top prior origins that fly into Denver International Airport that cause outgoing flights to be delayed, and how many flights get delayed?

Several observations and constraints are made to narrow down the scope of this question. Firstly, it can be observed that there are no data samples with the `LateAircraftDelay` field populated before 2003. This means only data from 2003 - 2008 is considered. Secondly, the amount of time an aircraft was delayed due to a late aircraft is not considered. Only the observed count is summated with empty entries resembling a non-delayed flight. Lastly, it can be assumed that an arrival delay is equally likely from any origin to a destination airport.

A connecting flight is delayed when a late aircraft from a prior origin arrives behind its scheduled arrival time. Therefore, the `ArrDelay` between the origin and destination can indicate how often the origin contributes to the destination having a late aircraft delay. However, not every flight that comes into an airport is subject to be a connecting flight. For example, there is a total of 656,566 flights from all origins into the destination Dallas/Fort Worth International Airport (DFW) with an arrival delay. Whereas, the number of flights out of DFW delayed due to a late aircraft totals to 194,947. This suggests that not every flight that was delayed upon arrival was a connecting flight.

With the assumption that arrival delays are equally considered between airports, it can be said that number of arrival delays to some destination can give insight to the number of connecting late aircraft delays. This can be quantified by taking a ratio of the arrival delay at a single origin to a destination over the total arrival delays for that destination. This will give a percentage, namely a contribution, between zero and one for each prior origin flying to a given destination. Thereafter, this contribution can be multiplied by the total number of late aircraft delays at the destination to understand the number of flights that get delayed as a result of incoming flights from that origin.

For this experiment, the top five airports to have late aircraft delays are analyzed, and the top five contributing airports are shown for each (Table 3).

Origin	Delays Due to Late Aircrafts	Total Arrival Delays	Top 5 Prior Origins	Contribution %	# Late Aircrafts
ORD	275,448	860,325	LGA	3.037	8,365
			MSP	2.823	7,777
			ATL	2.546	7,015
			DFW	2.497	6,879
			LAX	2.465	6,790
ATL	194,947	974,977	DFW	2.546	4,964
			LGA	2.532	4,936
			EWR	2.234	4,356
			MCO	2.152	4,195
			ORD	2.020	3,939
DFW	163,729	656,566	ATL	4.297	7,036
			ORD	3.301	5,404
			DEN	2.720	4,453
			LAX	2.423	3,968
			IAH	2.395	3,922
LAS	115,735	401,334	LAX	7.721	8,936
			PHX	6.673	7,723
			ORD	3.994	4,623
			DEN	3.903	4,517
			DFW	3.847	4,453
DEN	101,693	425,860	DFW	4.407	4,482
			ORD	4.041	4,110
			PHX	4.040	4,108
			LAS	3.683	3,745
			ATL	3.664	3,726

Table 3: Top prior origins that contribute towards an airports total number of late aircraft delays.