# DS 730: Final Project

Problem 3

Eli Bolotin

April 4th, 2019

## Introduction to the data

The dataset I used for this problem "consists of flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008" [1]. I originally found this dataset from hadoop illuminated [3].

All of the source data files are compressed CSV files. There are 22 of them in total. Compressed, their collective size is about 1.6gb. Uncompressed, their size is about 12 gb. I downloaded them onto EC2, uncompressed them, and then transferred these files to S3. The core data was accompanied by supplementary files that I used for mapping (airports and carriers).

## S3 storage

The data is stored on S3 at the following locations:

- **Flight data**: s3a://ebfp3/flight_data/
- **Supplemental data** (flights and carriers): s3a://ebfp3/supp_data/

## Data sources


- [1] Data source, http://stat-computing.org/dataexpo/2009/the-data.html
- [2] Data info, http://stat-computing.org/dataexpo/2009/
- [3] Hadoop Illuminated, https://hadoopilluminated.com/hadoop_illuminated/Public_Bigdata_Sets.html

# Questions and Answers

## Question 1

- Since 1987, which airline the highest frequency of the most departure delays per year? Said differently: which airline was the most late the most often?

## Answer 1

```
+--------------------+-----+
|        Description |Count|
+--------------------+-----+
|Southwest Airline...|    9|
|Delta Air Lines Inc.|    7|
|United Air Lines ...|    3|
|US Airways Inc. (...|    2|
|Continental Air L...|    1|
+--------------------+-----+
```

Southwest Airlines had the highest frequency of "most departure delays per year". The reverse analysis of most frequently having the *least delays per year* was Alaska Airlines.

## Question 2

- For each year, which airline flew the least of amount of miles compared to the yearly average (of miles flown by all airlines). Display the year, month, carrier, and % difference.

## Answer 2

```
+----+-----+--------------------+------------+
|Year|Month|         Description|percent_diff|
+----+-----+--------------------+------------+
|1987|   10|Pan American Worl...|      -0.837|
|1988|    4|Pacific Southwest...|     -0.9357|
|1989|    5|Eastern Air Lines...|     -0.9588|
|1990|    2|Alaska Airlines Inc.|     -0.8394|
|1991|   11|Pan American Worl...|     -0.9001|
|1992|    2|Alaska Airlines Inc.|     -0.8515|
|1993|    2|Alaska Airlines Inc.|      -0.877|
|1994|    2|Alaska Airlines Inc.|     -0.8293|
|1995|    2|Alaska Airlines Inc.|     -0.7818|
|1996|   11|Alaska Airlines Inc.|     -0.7592|
|1997|    2|Alaska Airlines Inc.|     -0.7733|
|1998|    2|Alaska Airlines Inc.|     -0.7729|
```

```
|1999|    2|Alaska Airlines Inc.|     -0.7658|
|2000|   11| Aloha Airlines Inc.|     -0.9768|
|2001|    2| Aloha Airlines Inc.|     -0.9708|
|2002|    2|Alaska Airlines Inc.|     -0.7308|
|2003|   11|Hawaiian Airlines...|     -0.8977|
|2004|   10|Hawaiian Airlines...|     -0.9022|
|2004|   11|Hawaiian Airlines...|     -0.9022|
|2004|    9|Hawaiian Airlines...|     -0.9022|
|2005|    2|Hawaiian Airlines...|     -0.9066|
|2006|    4| Aloha Airlines Inc.|     -0.9308|
|2007|    2| Aloha Airlines Inc.|     -0.9333|
|2008|    2| Aloha Airlines Inc.|     -0.9275|
+----+-----+--------------------+------------+
```

**Interpretting the results**: in 1987, Pan American flew 83.7% less miles than the average mileage flown by all airlines for that year. That figure is 90% for Pan American in 1991. Not surprisingly - both airlines went out of business in by the end of 1991.

Note that Alaska Airlines also had an underutilized fleet in the 90s. However, this airline changed it's pricing structure to increase competitiveness (by discounting fares drastically), and began to win over market share as a result.

## Question 3

- For every airport, which months of the year have the most delays with exception of November and December? Delays are defined as CarrierDelay + WeatherDelay + NASDelay + SecurityDelay + LateAircraftDelay.
  - Output the airport, city, month, sum of delays.
  - Output only top 20 airports

## Answer 3

```
+--------------------+----------------+-----+----------+
|             airport|            city|Month|sum_delays|
+--------------------+----------------+-----+----------+
|William B Hartsfi...|         Atlanta|    7|   4389913|
|Chicago O'Hare In...|         Chicago|    7|   3670572|
|Dallas-Fort Worth...|Dallas-Fort Worth|    6|   2815249|
|         Newark Intl|          Newark|    7|   1856703|
|George Bush Inter...|         Houston|    6|   1621504|
|         Denver Intl|          Denver|    6|   1407076|
| John F Kennedy Intl|        New York|    7|   1400956|
|   Philadelphia Intl|    Philadelphia|    7|   1374650|
|           LaGuardia|        New York|    7|   1208416|
|Detroit Metropoli...|         Detroit|    6|   1168501|
|Los Angeles Inter...|     Los Angeles|    7|   1163292|
```

```
|McCarran Internat...|        Las Vegas|    7|   1146545|
|Gen Edw L Logan Intl|           Boston|    7|   1138799|
|Phoenix Sky Harbo...|          Phoenix|    7|   1087779|
|Washington Dulles...|        Chantilly|    6|   1049662|
|Minneapolis-St Pa...|      Minneapolis|    6|    989554|
|Orlando Internati...|          Orlando|    7|    962422|
|Charlotte/Douglas...|        Charlotte|    6|    961091|
|Baltimore-Washing...|        Baltimore|    6|    896289|
|San Francisco Int...|    San Francisco|    6|    841711|
+--------------------+-----------------+-----+----------+
```

Clearly, the summer months of June and July are the busiest time of year to travel (with exception of Nov, Dec), as evidenced by the number of delays. Hartsfield Intl. airport, being the busiest airport in the world by passenger traffic, has the most delays.

## Question 4

- What is the most popular route of each airline? Route is defined as a combo of origin and destination. The time frame includes all years present in data.
  - Output airline, origin, destination, and count of flights.

## Answer 4

```
+--------------------+------+----+-------------+
|             Airline|Origin|Dest|Count_flights|
+--------------------+------+----+-------------+
|Southwest Airline...|   HOU| DAL|       230971|
|United Air Lines ...|   SFO| LAX|       191983|
|American Airlines...|   ORD| DFW|       136731|
|Continental Air L...|   BOS| EWR|       111611|
|Delta Air Lines Inc.|   ATL| LGA|       107065|
|US Airways Inc. (...|   BOS| PHL|       106837|
|Alaska Airlines Inc.|   ANC| SEA|       104826|
|Northwest Airline...|   MSP| DTW|       102806|
|America West Airl...|   LAS| PHX|        89209|
|American Eagle Ai...|   SAN| LAX|        62283|
|Trans World Airwa...|   MCI| STL|        49293|
|Skywest Airlines ...|   SAN| LAX|        41894|
|Hawaiian Airlines...|   OGG| HNL|        37185|
|AirTran Airways C...|   MCO| ATL|        24152|
|     JetBlue Airways|   JFK| FLL|        23936|
| Aloha Airlines Inc.|   OGG| HNL|        22031|
|Expressjet Airlin...|   IAH| DAL|        21904|
|Pan American Worl...|   LGA| BOS|        21891|
|Atlantic Southeas...|   ATL| PFN|        16099|
```

| | | | |
|---|---|---|---|
| Comair Inc. | DCA | BOS | 12449 |
| Mesa Airlines Inc. | PHX | TUS | 11110 |
| Eastern Air Lines... | BOS | DCA | 10494 |
| ATA Airlines d/b/... | LGA | MDW | 10425 |
| Independence Air | JFK | IAD | 8909 |
| Frontier Airlines... | DEN | LAS | 8655 |
| Piedmont Aviation... | CLT | GSO | 6256 |
| Pacific Southwest... | SFO | LAX | 4149 |
| Pinnacle Airlines... | CVG | DTW | 3856 |
| Midway Airlines I... | DTW | MDW | 2539 |