A4-pranav-shreyasa-tirthraj

pranav-shreysa-tirthraj

October 12, 2017

Introduction

Map reduce is a software framework for easily writing applications which process vast amounts of data in parallel on large clusters. A typical MapReduce job usually splits the input data set into independent chunks which are then worked upon by map tasks run in parallel.

The Map tasks take in a Key-Value input and produces an intermediate Key-Value output. This output is typically consumed by the reduce jobs which then merges all the values associated with the same Key. gives a output by reducing the outputs of the individual mappers.MapReduce automatically parallelizes the program while taking care of input data partitioning,scheduling program execution and handling machine failures.

Typically,large complex problem sets require multiple map-reduce tasks run either in sequence or parallel which would produce a resultant output for the input query.

Code Overview

In this problem statement,we are calculating the mean delay of the 5 most active airlines and the 5 most active airports in the country based on an input dataset given to us.Based on it,we will be plotting the mean delay of the top 5 airports and the top 5 airlines.

The system uses 3 mappers and 2 reducers.

The first mapper is used to parse the text and clean the data. In this step,the stated at http://janvitek.org/pdpmr/f17/task-a4-delay.html are performed to clean the input data as specified. The output of the mapper is then used to calculate the mean delay of the flights and the busiest airports for every month,in every year.

The second mapreduce job calculates the mean delay for each airline by taking the input from the first mapper output. takes the output of the first map job and finds out the mean delay for each airline per month,per year.

The third mapreduce job calculates the mean delay for every airport coming to a particular destination from the output of map job 1 per month,per year.

The second and the third mapreduce jobs run in parallel whereas the data cleaning job and other mapreduce jobs execute in serial.

The second and third map outputs are used for analysis of the mean delays of the top 5 airlines and the top 5 busiest airports. The graph plots of this analysis per year,instead of a cumulative aggregate helps in segregating the recent statistics with the old statistics.There are several reasons for grouping per year instead of overall group- 1. A current statistical analysis is easy to obtain and thus helpful for the near future for possible scenarios like diverting flights to other moderately loaded airports. 2. The turn of the century injected with it a burst of technology in aviation industry which makes the data from the late 20th century obsolete with its outdated technology.

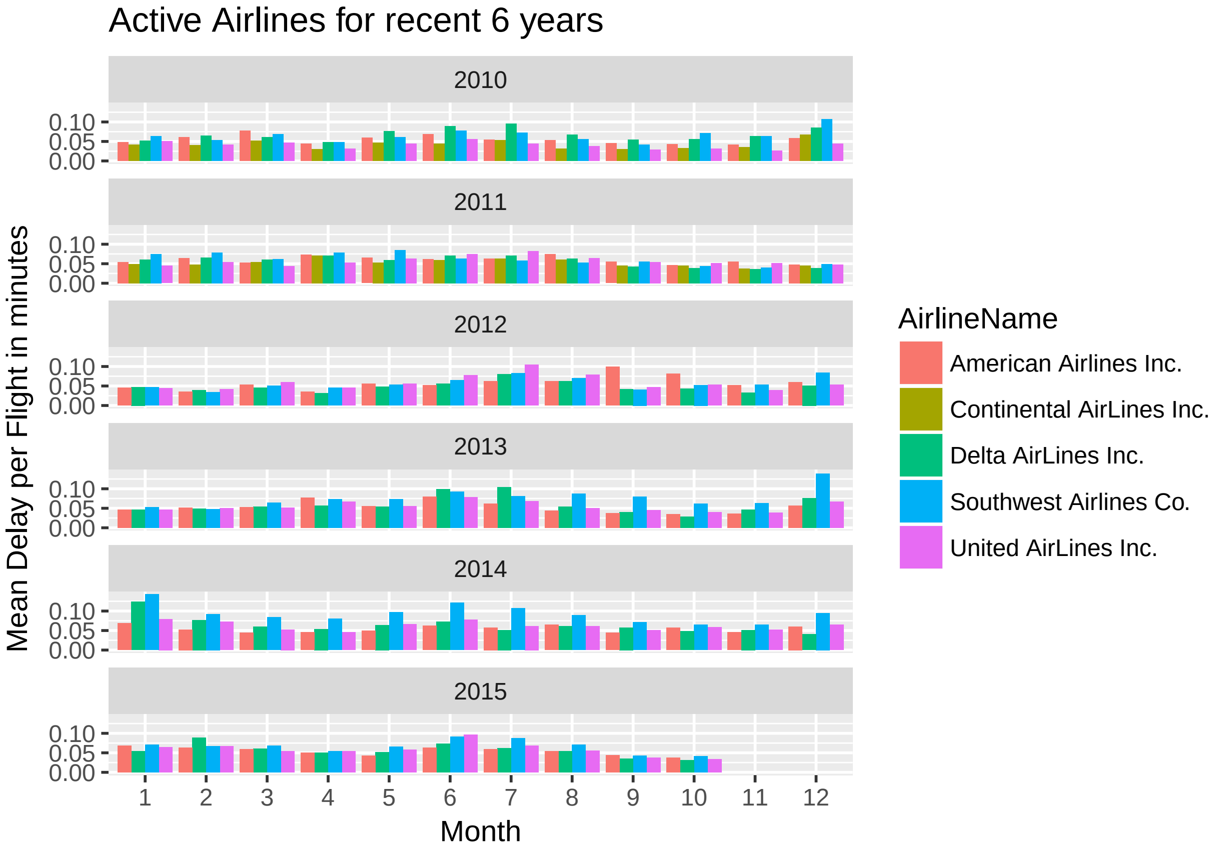
Execution Details

1. The target configuration is as follows- The EMR cluster configuration on which the job was run was as follows - 4 vCPU,

8 GiB memory,
EBS only storage EBS Storage:32 GiB

2. The output graphical representations are as follows

Plot 1: Most Active Airlines in the period 2011-2015 with their Mean Delay



Plot 1 graphically represents the most actives airlines : American Airline, Continental Airline, Delta Airline, Southwest Airline and United Airlines in the period 2010 to 2015 with each graph representing data for months 1-12 but the last graph that displays data for months 1-10.

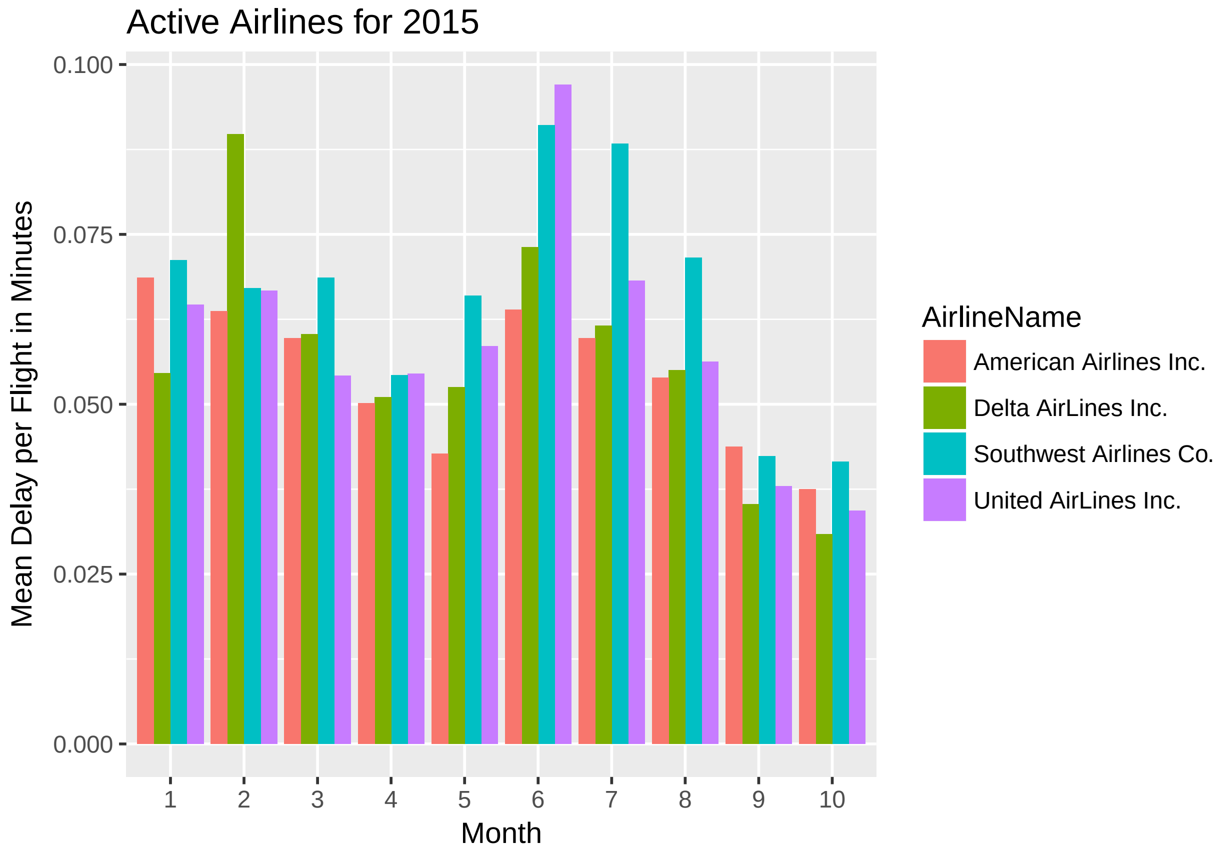
The top 2 graphs in Plot 1 represent the years 2010 and 2011. It shows presence of 5 airlines with their mean delay in minutes(mean delay has been normalised by the following formula: Arrival Delay/Scheduled Length).

It is interesting to note that after year 2011, Continental Airline disappear from the graph as the airline is probably discontinued with reasons out of scope.

In the year 2010, Southwest airline has the maximum delay of 0.106 in the month of December. In the year 2011, Southwest airline has the maximum delay of 0.085 in the month of May. In the year 2012, United airline has the maximum delay of 0.104 in the month of July. In the year 2013, Southwest airline has the maximum delay of 0.138 in the month of July. In the year 2014, Southwest airline has the maximum delay of 0.142 in the month of January. In the year 2015, United Airline has the maximum mean delay of 0.097 in the month of June

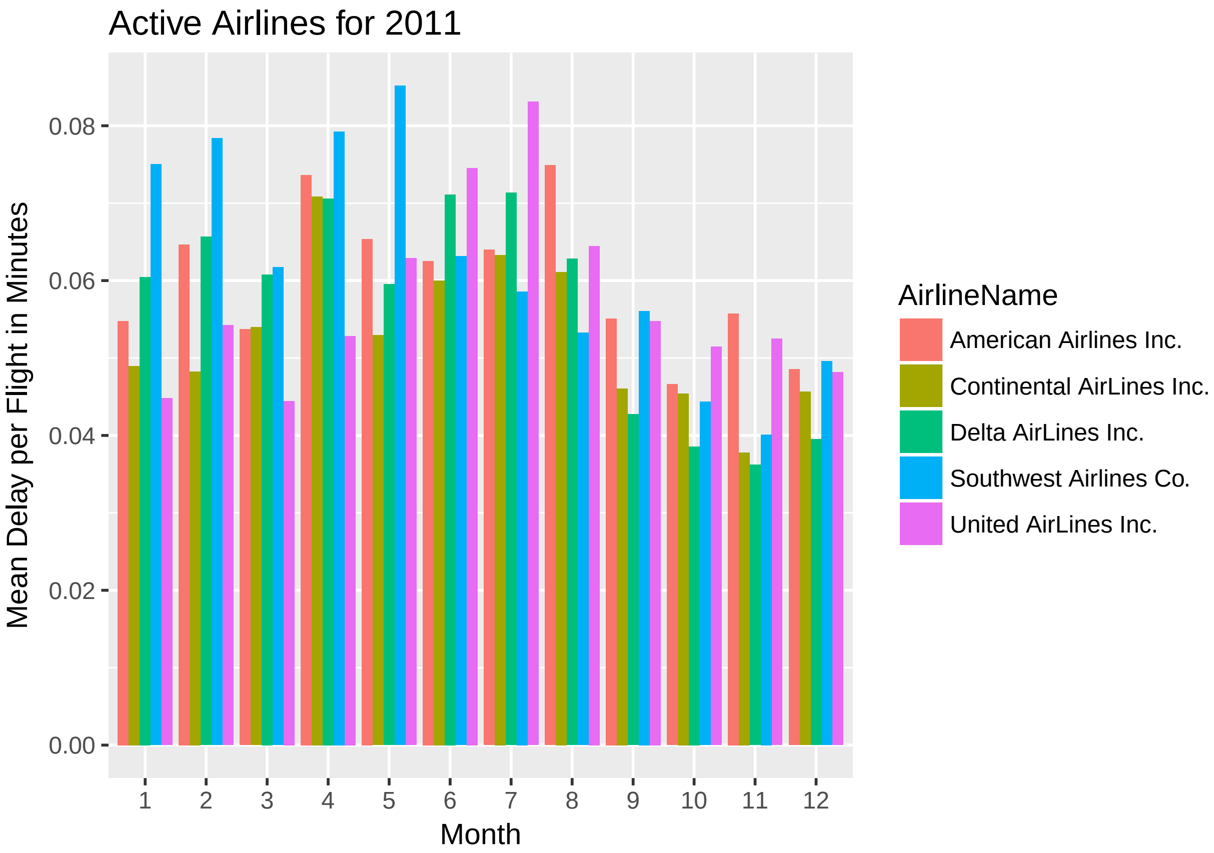
After studying the data it was observed that Southwest airline is the most delayed airline over the period of 2010 to 2015.

Plot 2: Most Active Airlines in the year 2015 with their Mean Delay



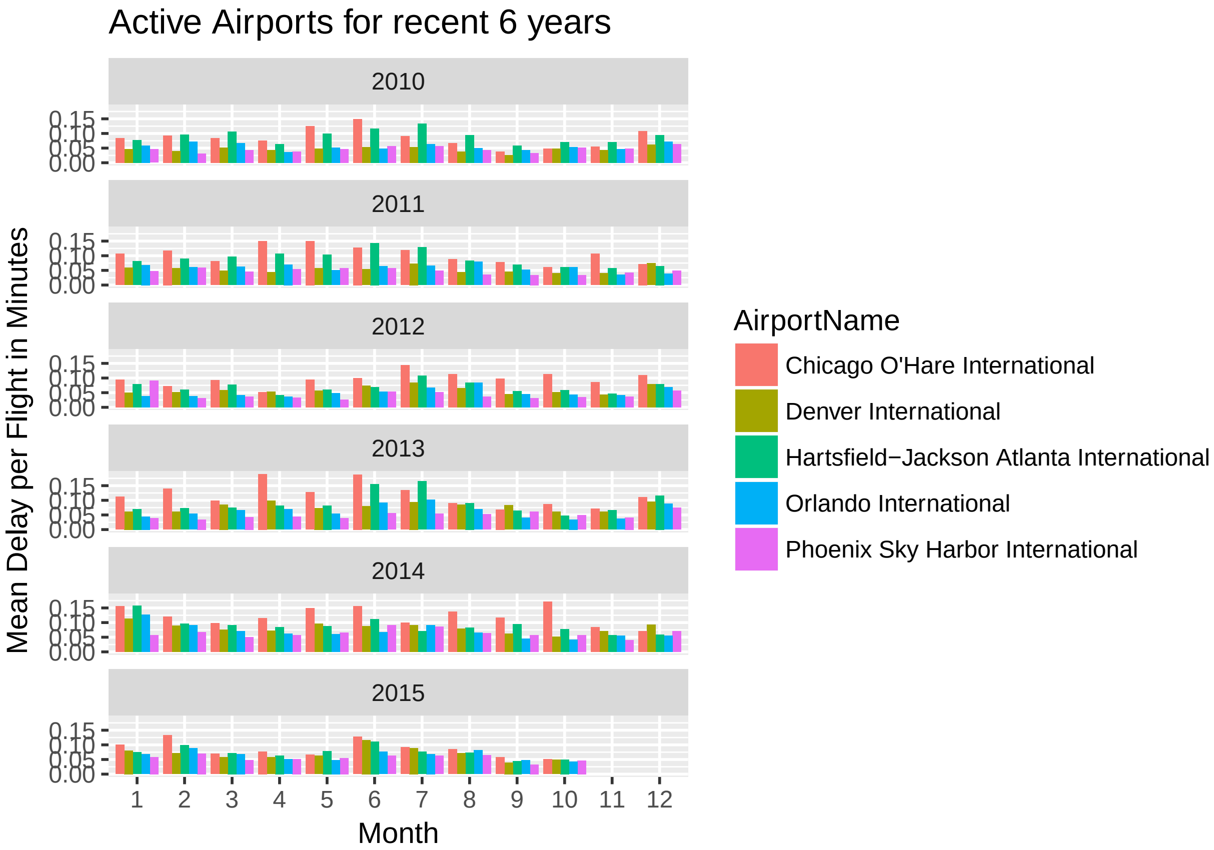
Plot 2 graphically represents the 4 most actives airlines : American Airline, Delta Airline, Southwest Airline and United Airlines in the year 2015. United Airline has the most mean delay of 0.097 in the month of June and Southwest Airline with the second most mean delay of 0.083 in the same month. Out of these airlines shown, Delta Airlines has the least mean delay of 0.030 in the month of October.

Plot 3: Most Active Airlines in the year 2011 with their Mean Delay



Plot 3 graphically represents the 5 most actives airlines : American Airline, Continental Airlines, Delta Airline, Southwest Airline and United Airlines in the year 2011. Southwest airline has the maximum mean delay of 0.085 in the month of May and United Airline with the second most mean delay of 0.091 in July. Out of these airlines shown, Delta Airlines has the least mean delay of 0.036 in November.

Plot 4: Most Active Aiports in the period 2010-2015 with their Mean Delay



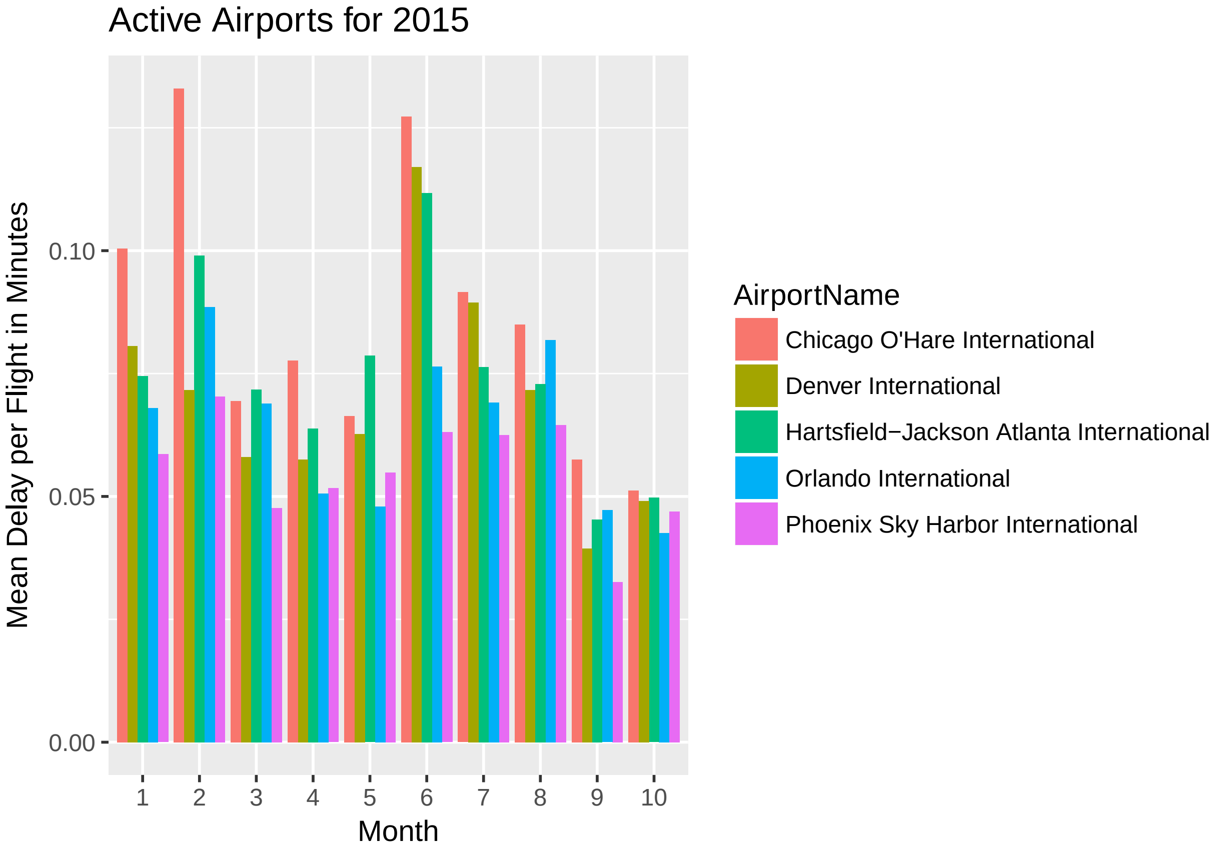
Plot 4 graphically represents the most actives airports : HartsField Jackson Atlanta International Airport, Chicago O'Hare International Airport, Denver International Airport, Orlando International Airport and Phoenix Sky Harbor International Airport in the period 2010 to 2015 with each graph representing data for months 1-12 but the last graph that displays data for months 1-10.

It shows presence of 5 airports with their mean delay in minutes(mean delay has been normalised by the following formula: Arrival Delay/Scheduled Length).

In the year 2010, Chicago O'Hare International Airport has the maximum delay of 0.148 in June. In the year 2011, Chicago O'Hare International Airport has the maximum delay of 0.15 in May. In the year 2012, Chicago O'Hare International Airport has the maximum delay of 0.143 in July. In the year 2013, Chicago O'Hare International Airport has the maximum delay of 0.189 in April. In the year 2014, Chicago O'Hare International Airport has the maximum delay of 0.171 in October In the year 2015, Chicago O'Hare International Airport has the maximum mean delay of 0.133 in June

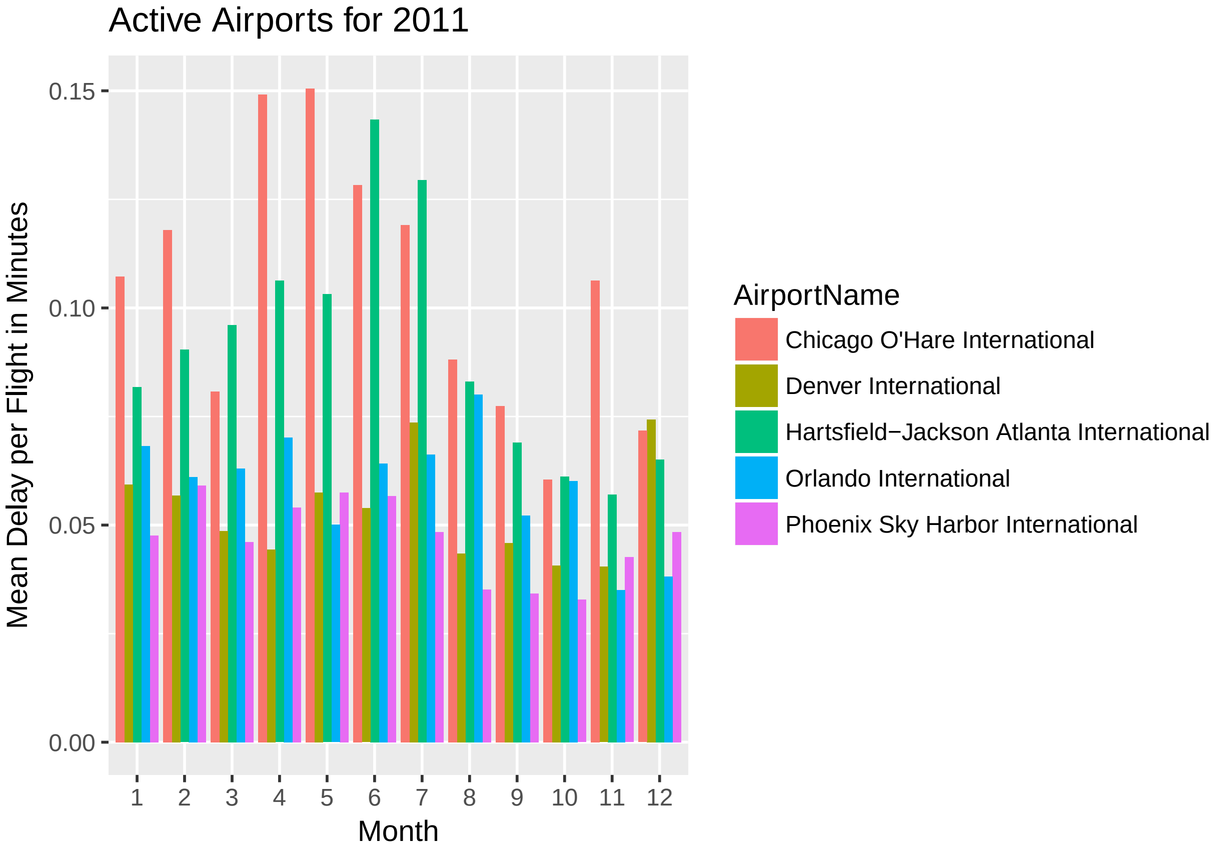
After studying the data it was observed that Chicago O'Hare International Airport is the most over-extended airport over the period of 2010 to 2015.

Plot 5: Most Active Airports in the year 2015 with their Mean Delay



Plot 5 graphically represents the 5 most actives airports : HartsField Jackson Atlanta International Airport, Chicago O'Hare International Airport, Denver International Airport, Orlando International Airport and Phoenix Sky Harbor International Airport in 2015. Chicago O'Hare International Airport has the most mean delay of 0.133 in June and Denver International Airport with the second most mean delay of 0.11 in the same month. Out of the airports shown, Phoenix Sky Harbor International Airport has the least mean delay of 0.032 in September out of the shown airports.

Plot 6: Most Active Airports in the year 2011 with their Mean Delay



Plot 6 graphically represents the 5 most actives airports : HartsField Jackson Atlanta International Airport, Chicago O'Hare International Airport, Denver International Airport, Orlando International Airport and Phoenix Sky Harbor International Airport in 2011. Chicago O'Hare International Airport has the most mean delay of 0.15 in May and HartsField Jackson Atlanta International Airport with the second most mean delay of 0.143 in June. Out of the airports shown, Phoenix Sky Harbor International Airport has the least mean delay of 0.032 in October out of the shown airports.

3. The execution times for the runs are as follows- The longest time the job took to run was 32 minutes on the EMR cluster.The cleaning of the data took the longest time with 22-24 minutes on average.

Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.