COSC 6397: Big Data Analytics

Instructor: Dr. Edgar Gabriel

Homework 1

==========

Flights Data Analysis

==========

project report

by

Venkata Yeshes Meka
PSID: 1141507
vmeka@uh.edu

Contents

1. Introduction

Goal:

1. To implement a MapReduce job which determines the percentage of delayed flights per Origin Airport
2. To implement a MapReduce job which determines the percentage delayed flights per Origin Airport and Month.

2. Experimental Setup

Tools Used:

1. Cluster provided to the course(Shark Cluster) configured to run distributed processing of data.
2. Cloudera virtual machine with Hadoop eco-system installed -
3. Java 1.6 or higher.
4. Apache Hadoop 2.2.0 - is an open-source software framework for storage and large scale processing of data-sets on clusters of commodity hardware.
5. Eclipse - is an integrated development environment (IDE). It contains a base workspace and an extensible plug-in system for customizing the environment. Written mostly in Java,
6. Fugu – is a graphical frontend for the text-based Secure File Transfer Protocol (SFTP) client that ships with Mac OS X. SFTP is similar to FTP, but the entire session is encrypted, meaning nothing, including passwords, is sent in the clear.
7. SSH – built in the command lines tools of the Mac OSX operating system to remotely login to the cluster.
8. Microsoft Excel for generating plots to visualize data.

Dataset:

The dataset is the data of flights obtained from "http://stat-computing.org/dataexpo/2009/the-data.html"

The following shows the sample from the dataset
2008,1,3,4,NA,905,NA,1025,WN,469,,NA,80,NA,NA,NA,LAX,SFO,337,NA,NA,1,A,0,NA,NA,NA,NA,NA
Size of the dataset ~ 700 MB with information about 7 million flights recorded from 1987 – 2008.

3. Solution Strategy

Assumptions about the dataset with respect to the features:

The delay percentage is computed taking the departure delay field in the dataset. It is calculated using the following formula:

$$\textbf{Delay Percentage} = \frac{\textbf{Total number of flights that have departure delays from the airport}}{\textbf{Total number of flights that take of from the airport}}$$

The dataset contains "NA" value in the departure delay field, so the assumption is that any value in the departure delay with positive values or "NA" are considered as the delays.

Part 1:

In order to find the delay percentage per origin airport , using MapReduce,

Mapper:

The Mapper generates the intermediate key-value pairs of (origin airport ID , the flag which indicates if that flight is delayed or not). Here the airport ID is of the type "Text" and the flag is of the type "IntWritable".

Reducer:

The Reducer then computes the aggregation of total number of delayed flights using the flag counter and the total number of flights from that particular airport and map these to the key i.e., the respective origin airport id. Here the output key i.e., origin airport Id is of "Text" type and the output value of the delay percentage is of "IntWritable " type.

Thus the output from reducer gives the desired results of delay percentage by the origin airport ID.

Part 2:

In order to find the delay percentage by origin airport Id per month using MapReduce

Mapper:

The Mapper generates the intermediate key-value pairs with the key being a combination of (origin airport id and the month). The month is available in the dataset under the month field. The value here is the flag counter to represent if that particular flight is delayed or not. Here the airport ID is of the type "Text" and the flag is of the type "IntWritable".

Reducer:

The Reducer aggregates the total number of delayed flights and the total number of flights from that particular airport Id. It aggregates the total considering the number of flights from the airport per month basis. Finally it calculates the delay percentage. The output key i.e., the combination of

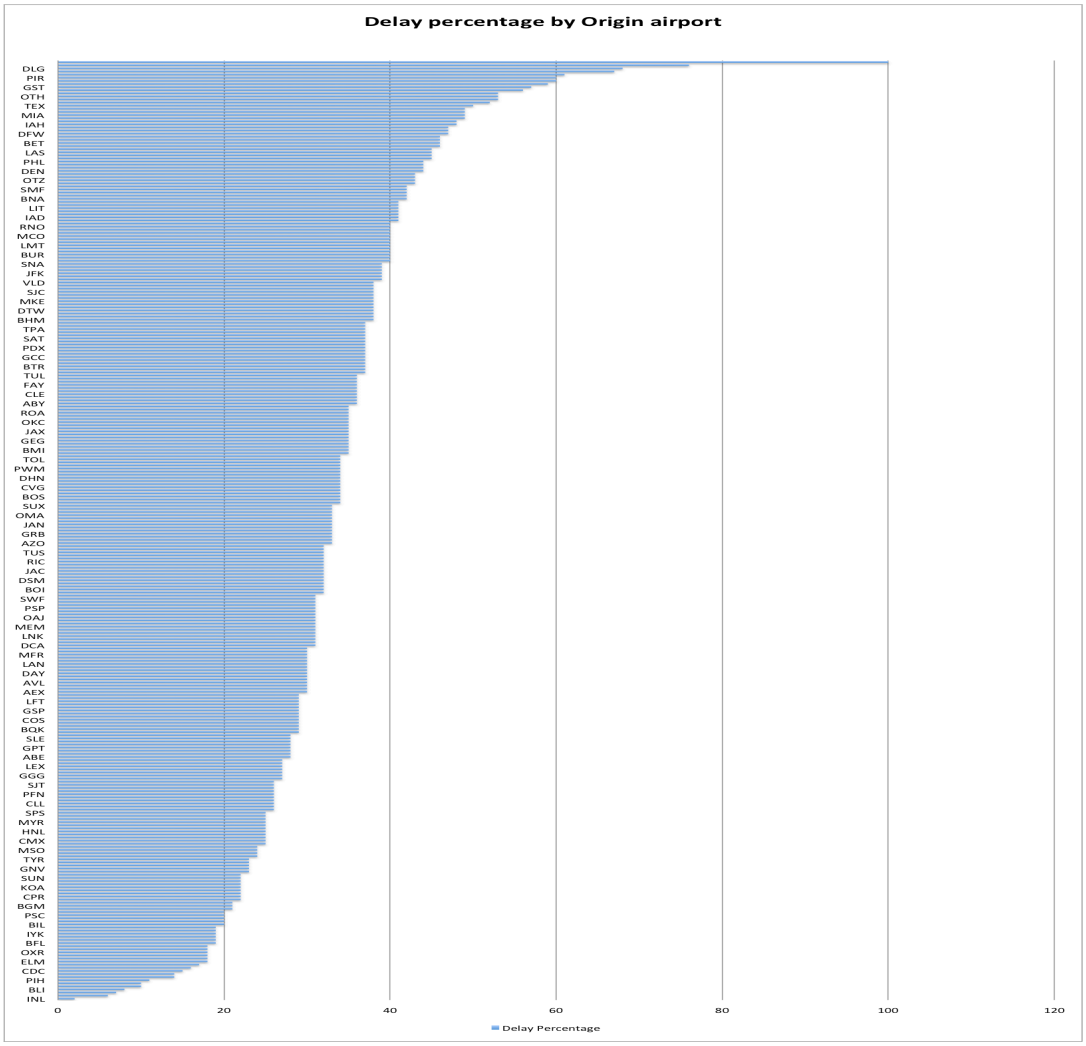
airport id and month is of "Text" type and the output value i.e., delay percentage by month per month is of type "IntWritable".

Initially the code is tested using a sample dataset containing 17,000 flights data locally using a Cloudera Virtual Machine with Hadoop eco-system installed. Then the code is deployed onto the shark cluster provided to the class and tested on the actual dataset with 7 million flights data. The results are then analyzed to record the findings.

4. Results & Observations

Part 1:

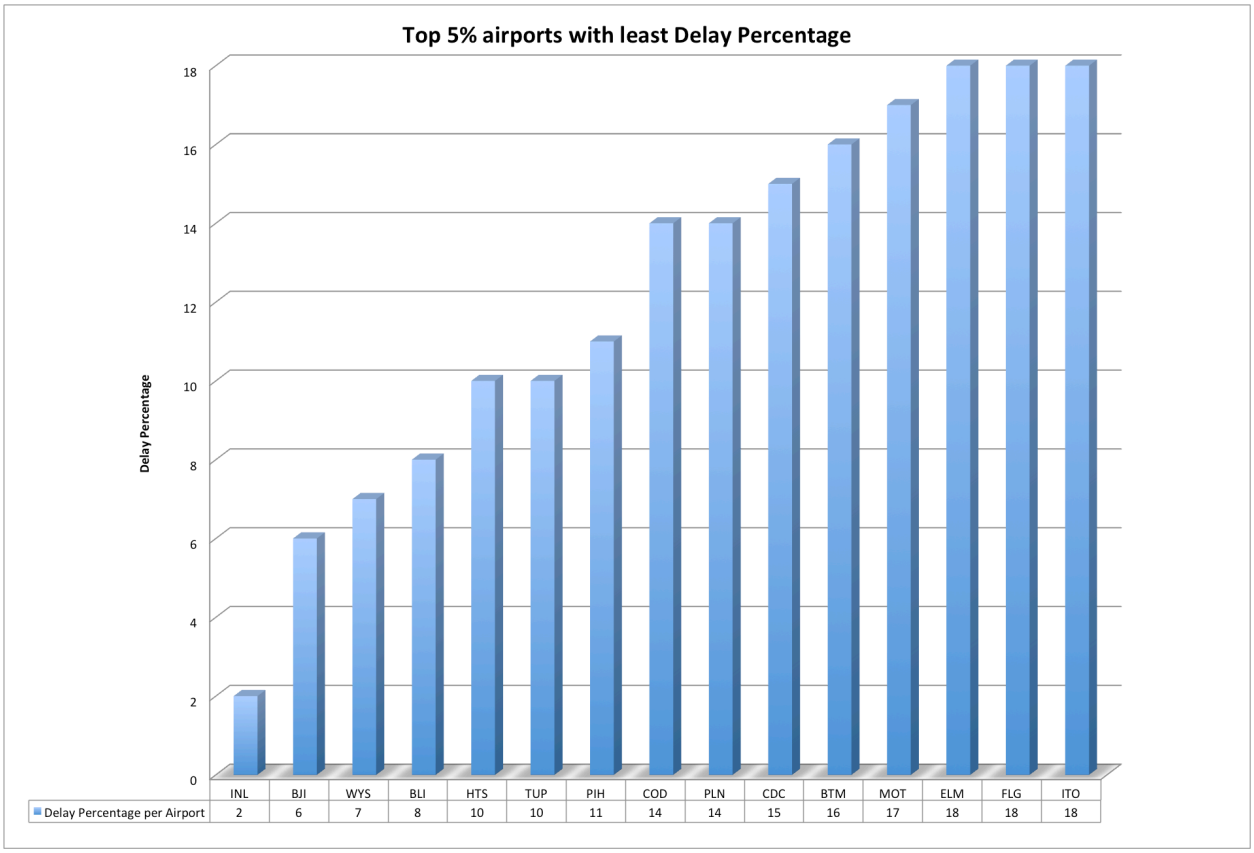
a. Plot showing the delay percentage by the origin airport for all the airports in the dataset.



The above plot shows how the delay percentage is spread out over ~300 airports given in the dataset. Considering the data by origin airport
The airport with least delay - Falls International airport, Minnesota

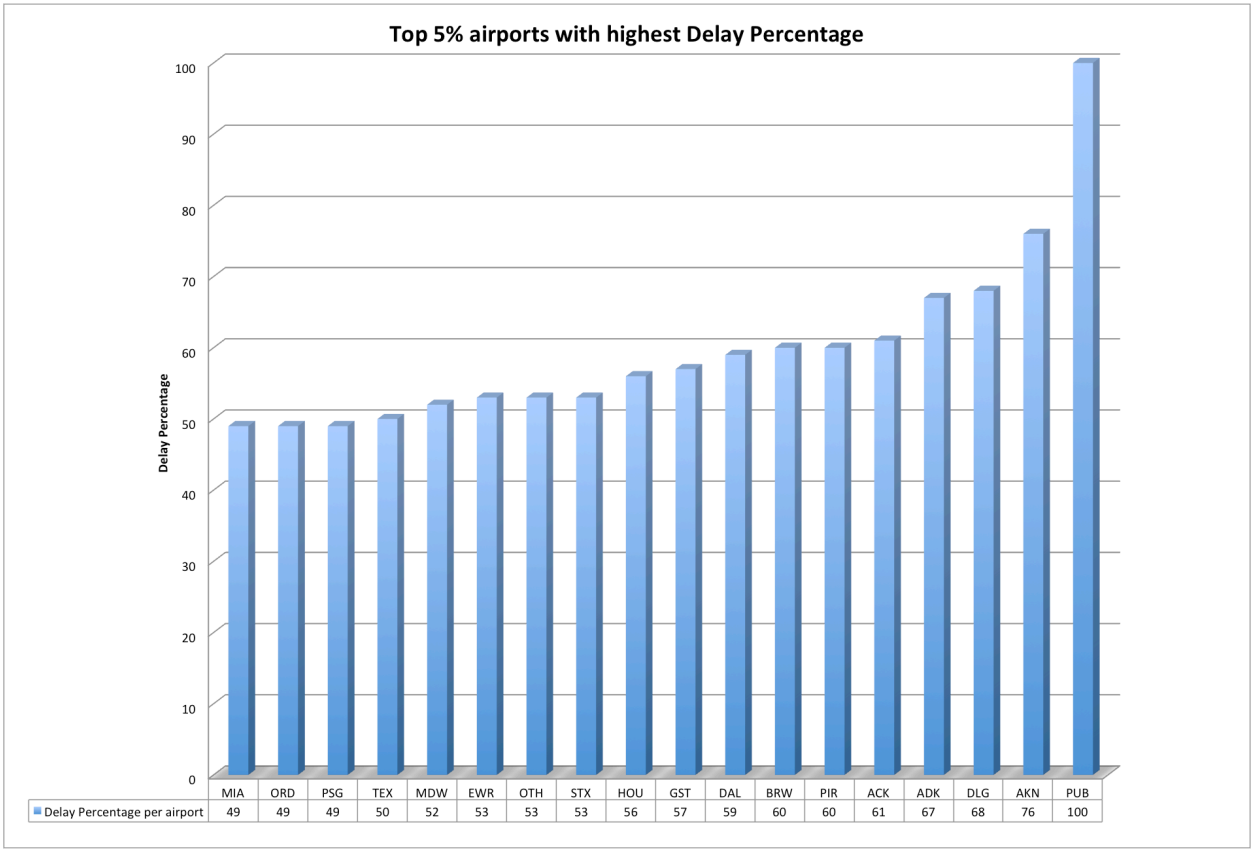
The airport with highest delay – Pueblo Memorial airport, Colorado

b. Plot showing the top 5% of airports with least delay percentage recorded.



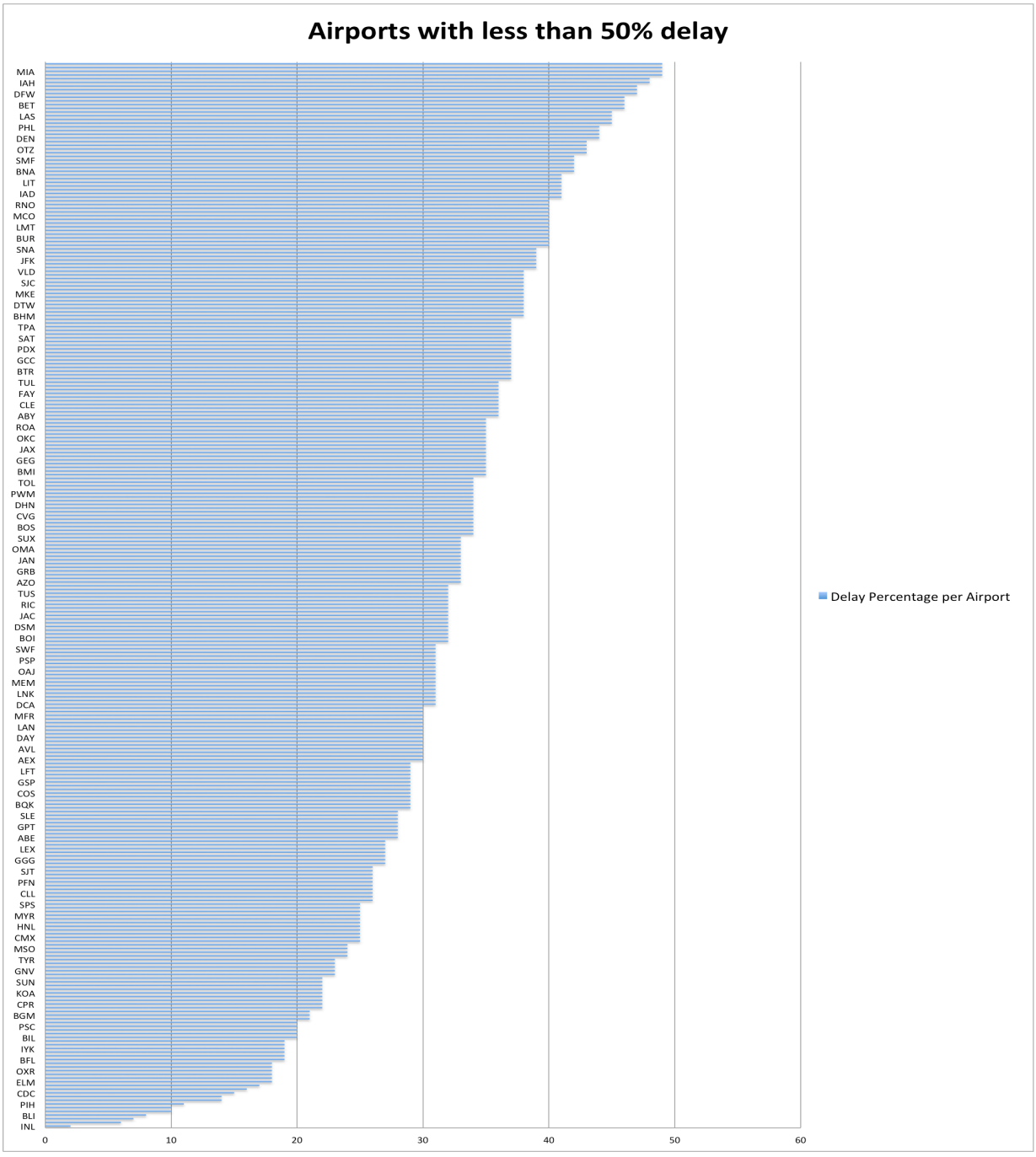
About 15 airports fall in the 5% of airports which recorded the least delay percentage overall with respect to the origin airport with INL – Fall International airport, Minnesota being the airport with least delay of 2%.

c. Plot showing the top 5% of airports with highest delay percentage



About 18 airports fall in the 5% of airports which recorded the highest delay percentage overall with respect to the origin airport with PUB – Pueblo Memorial airport, Colorado being the airport with least delay of 100%.

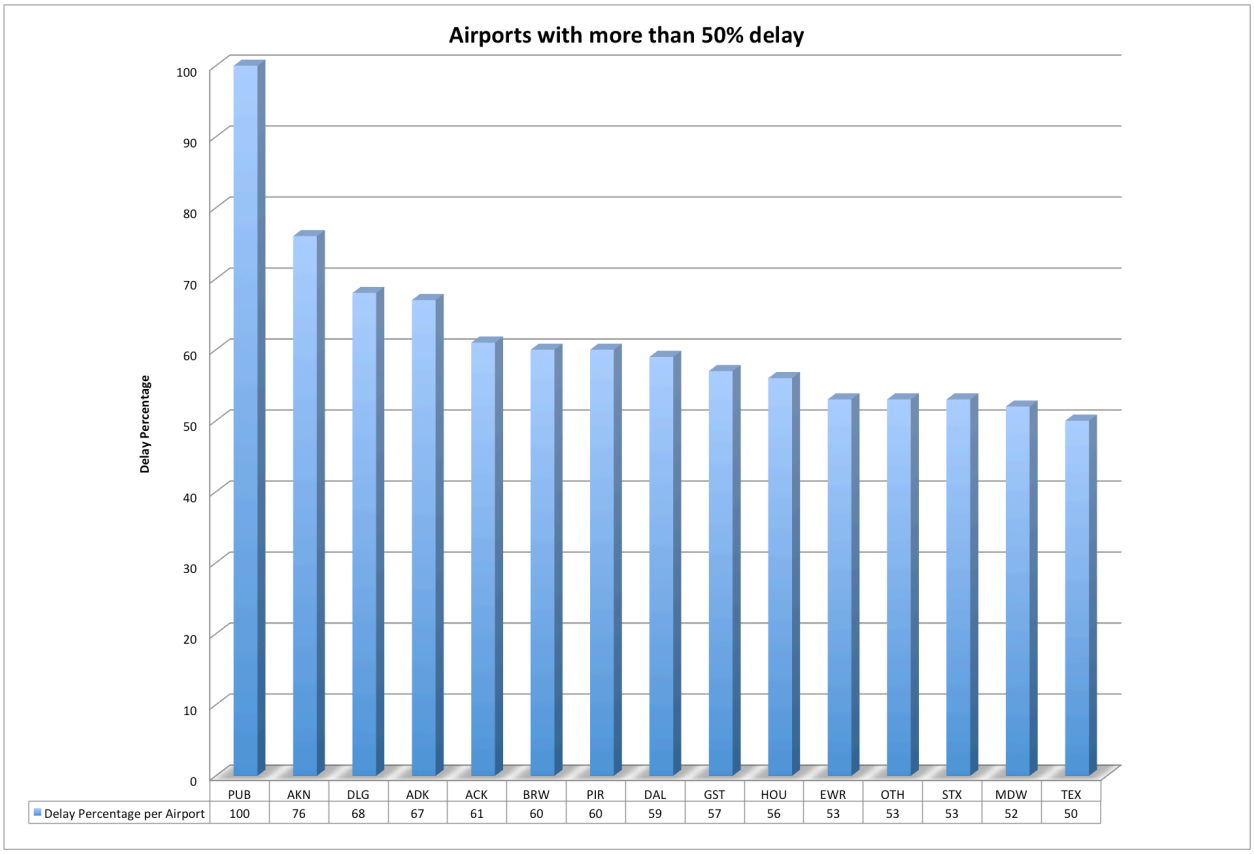
d. Plot showing the airports with delay percentage less than 50%



About 286 airports fall in the airports which recorded the delay percentage overall less than 50% with respect to the origin airport with INL – Falls

International airport, Minnesota being the airport with least delay of 2% and PSG - James C. Johnson Petersburg, AK with 49 % delay.

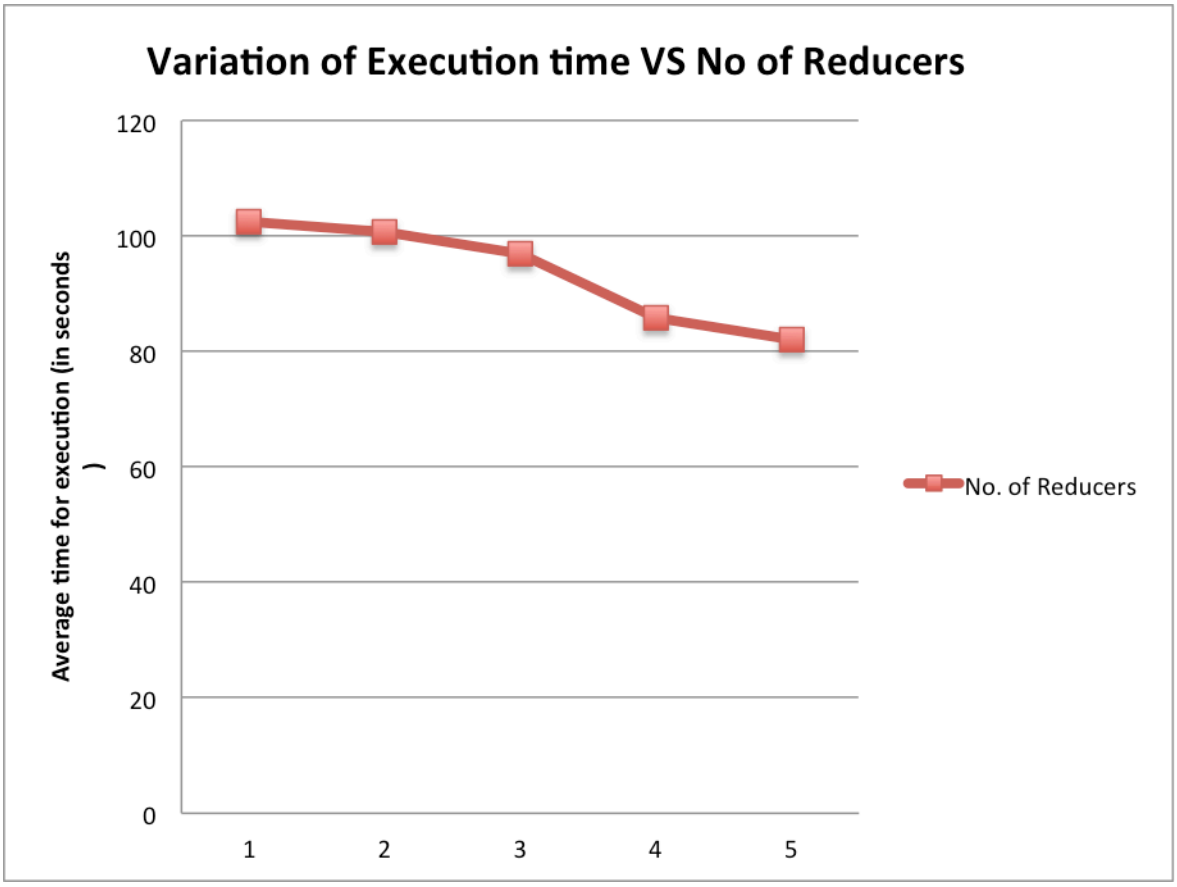
e. Plot showing the airports with delay percentage more than 50%



About 16 airports fall in the airports which recorded the delay percentage overall more than 50% with respect to the origin airport with TEX – Telluride Regional, Colorado being the airport with delay of 50% and PUB - Pueblo Memorial airport, Colorado with 100% delay.

f. Table showing the variation of average time taken for the job to complete with variation in the number of reducers

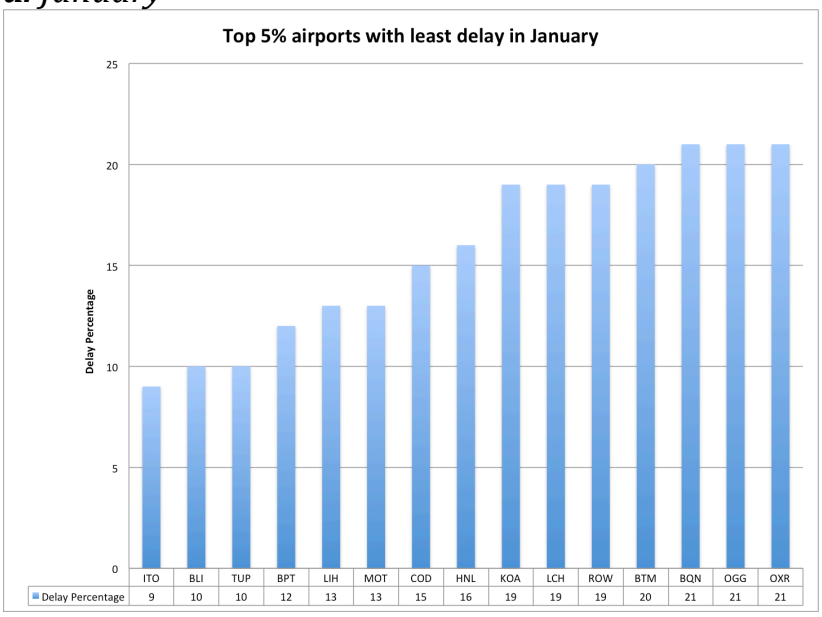
SL. No	Job Name	No. of Reduce Tasks	Average Time taken(in seconds)
1	Flights by airport	1	102.523155433
2	Flights by airport	2	100.661065544
3	Flights by airport	4	96.862429866
4	Flights by airport	6	85.71118128
5	Flights by airport	8	82.05924305

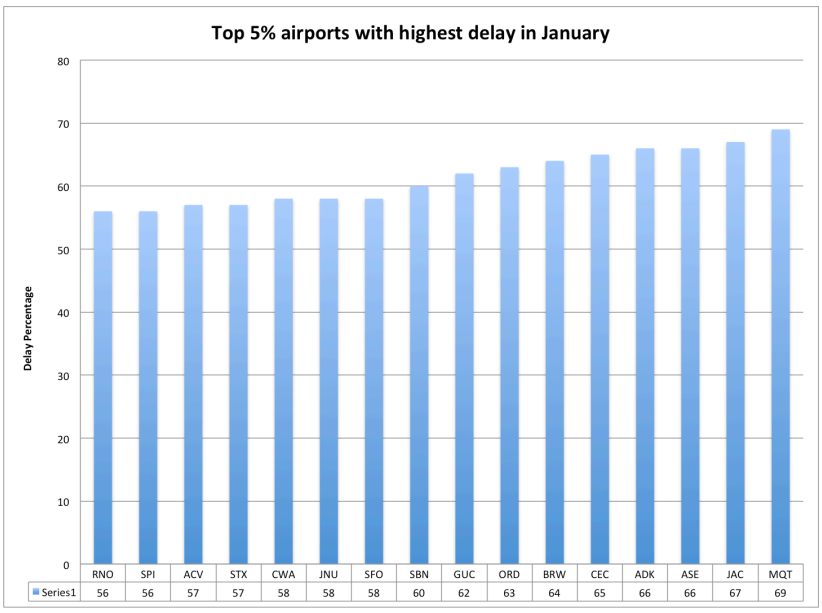


With the increase in number of reducers, the time of execution of job decreases i.e., no of reducers is inversely proportional to time of execution of jobs.

Part 2:
Plots showing the airports with top 5% least and highest delay percentages per month

a. January

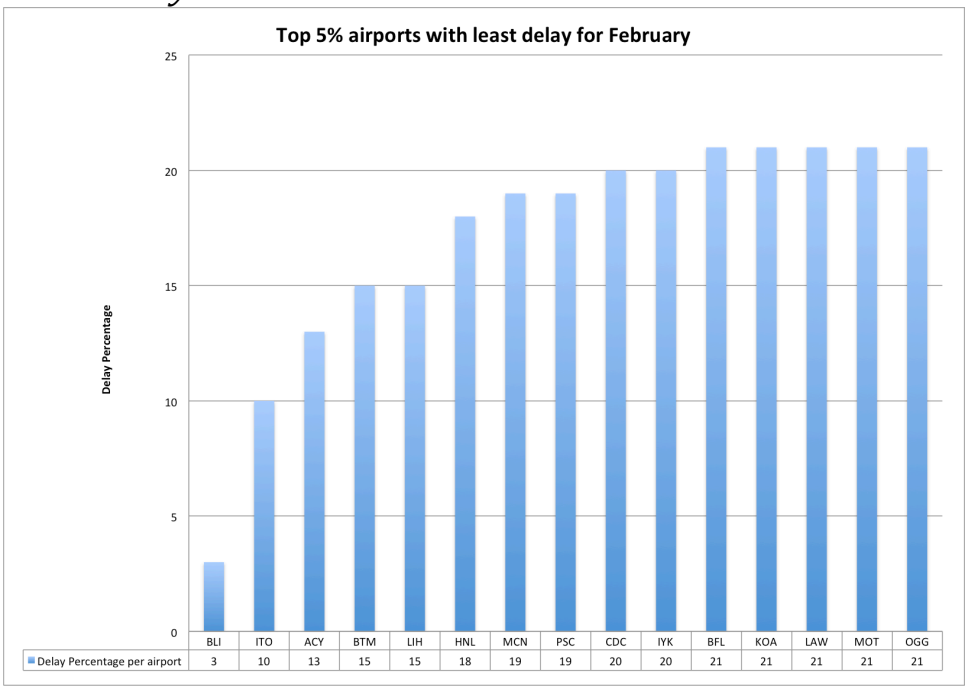


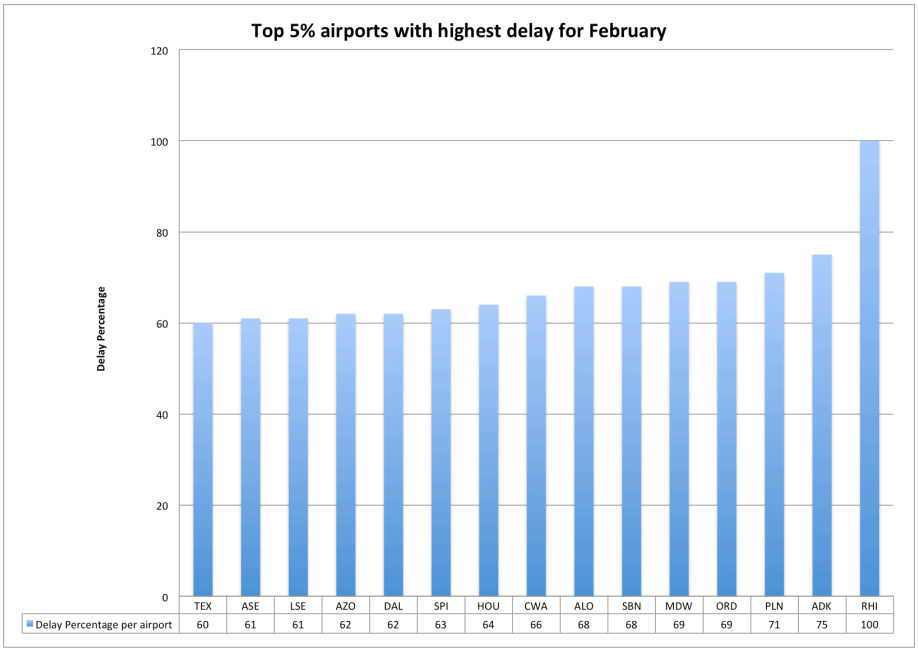


The above plots show the delay percentage by origin airport per month values for January. The airport with least delay in this month is ITO - Hilo International, Hawaii.

The airport with highest delay in this month is MQT – Sawyer International Airport, Michigan

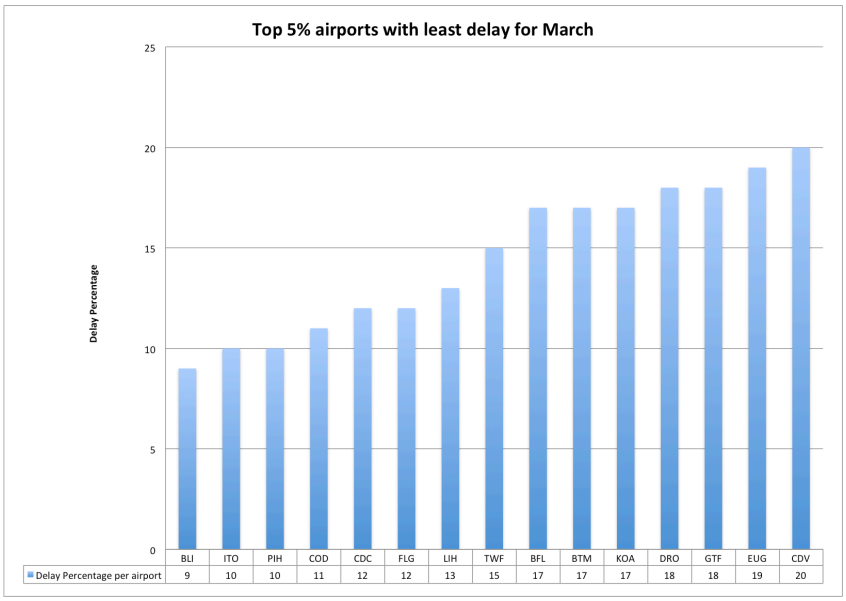
b. February

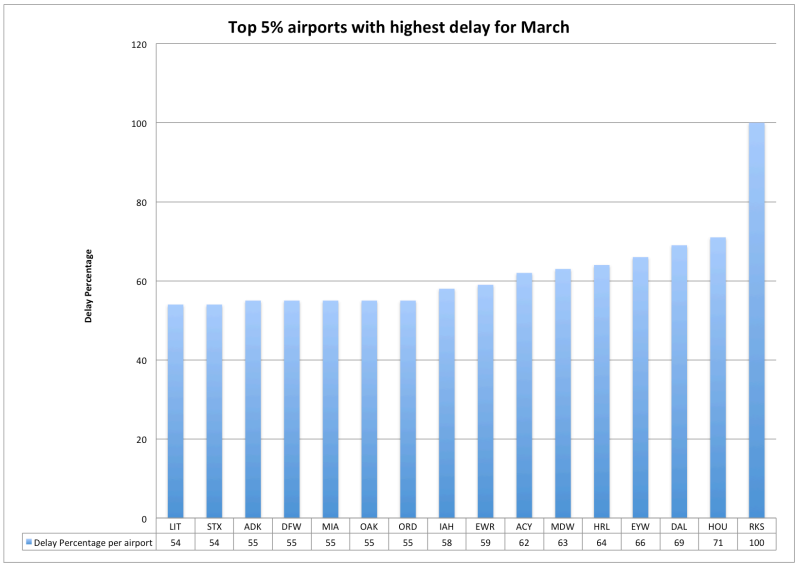




The above plots show the delay percentage by origin airport per month values for February.
The airport with least delay in this month is BLI – Bellingham international airport, Washington with 3% delay. The airport with highest delay in this month is RHI – Rhinelander-Oneida County, Wisconsin with 100% delay.

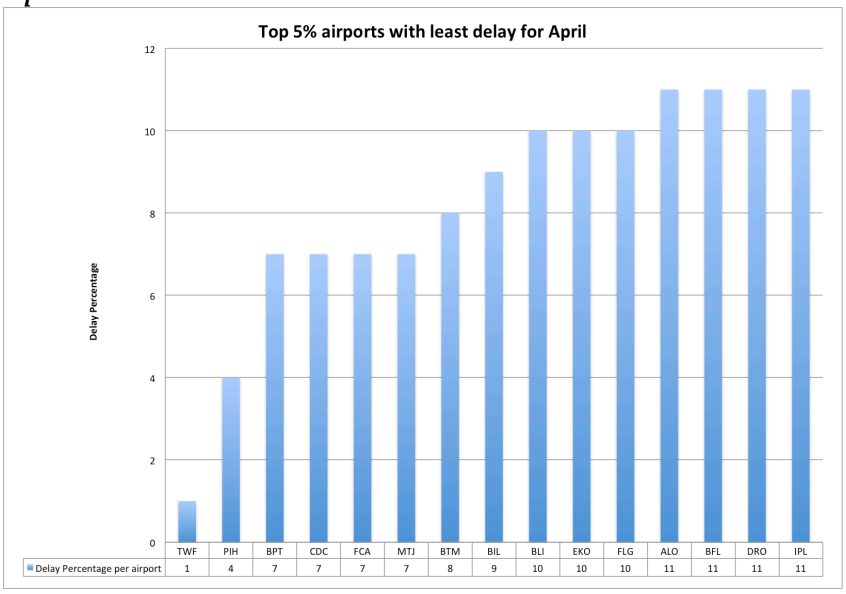
March

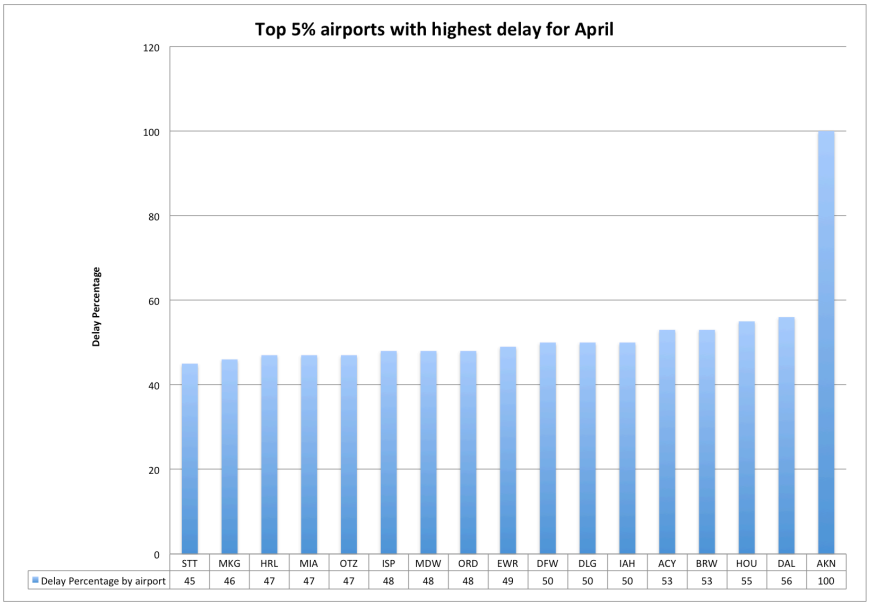




The above plots show the delay percentage by origin airport per month values for March. The airport with least delay in this month is BLI – Bellingham international airport, Washington with 9% delay. The airport with highest delay in this month is RKS – Rock Springs-Sweetwater County, Wyoming with 100% delay.

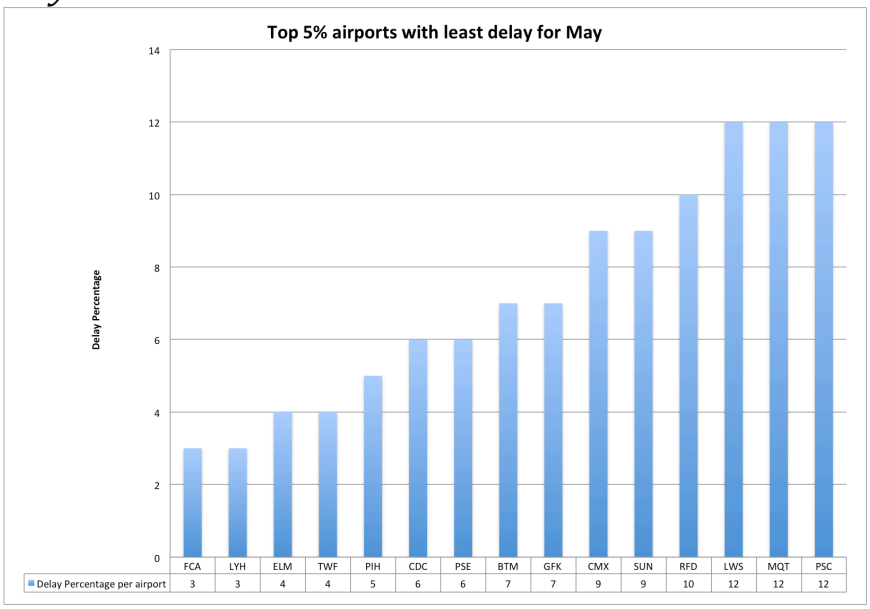
April

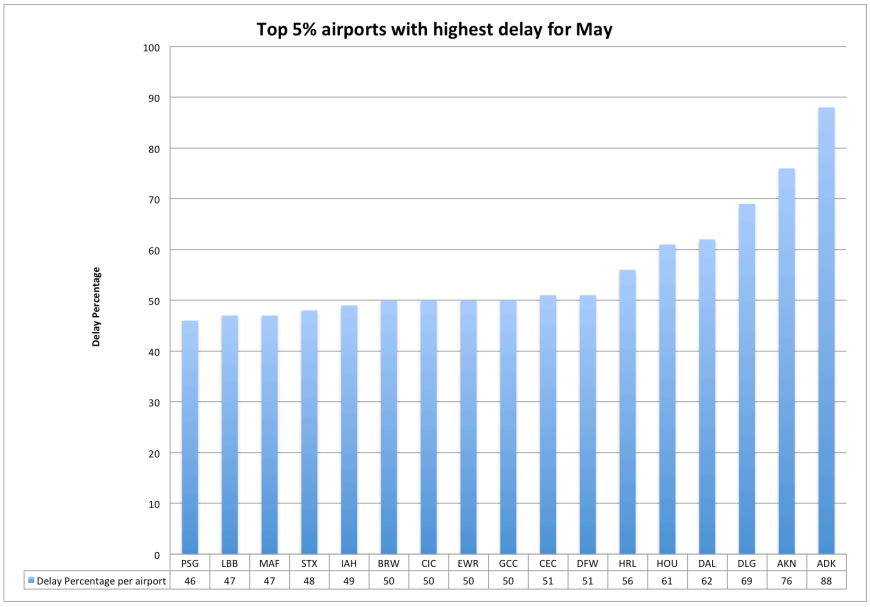




The above plots show the delay percentage by origin airport per month values for April. The airport with least delay in this month is TWF– Joslin Field - Magic Valley airport, Idaho with 1% delay. The airport with highest delay in this month is AXN – Chandler, Minnesota, with 100% delay.

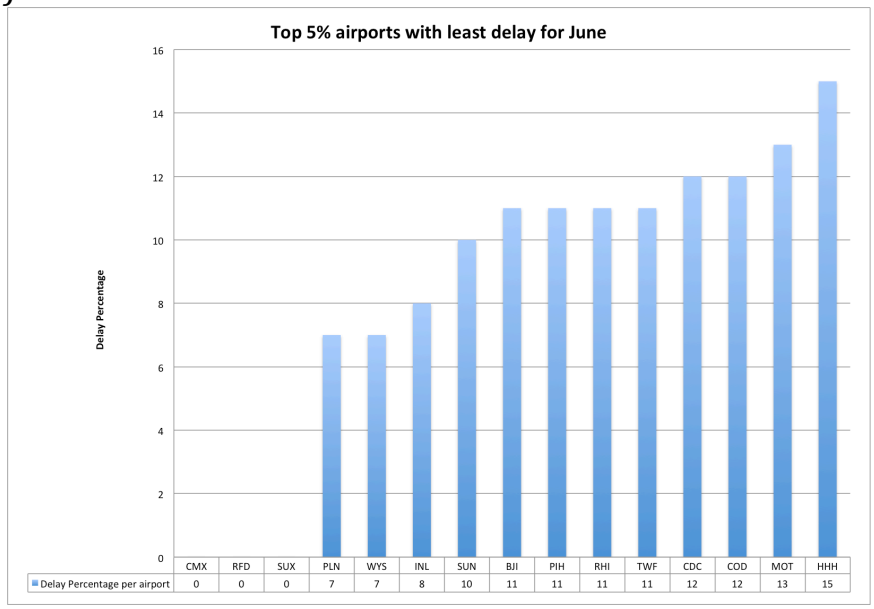
May

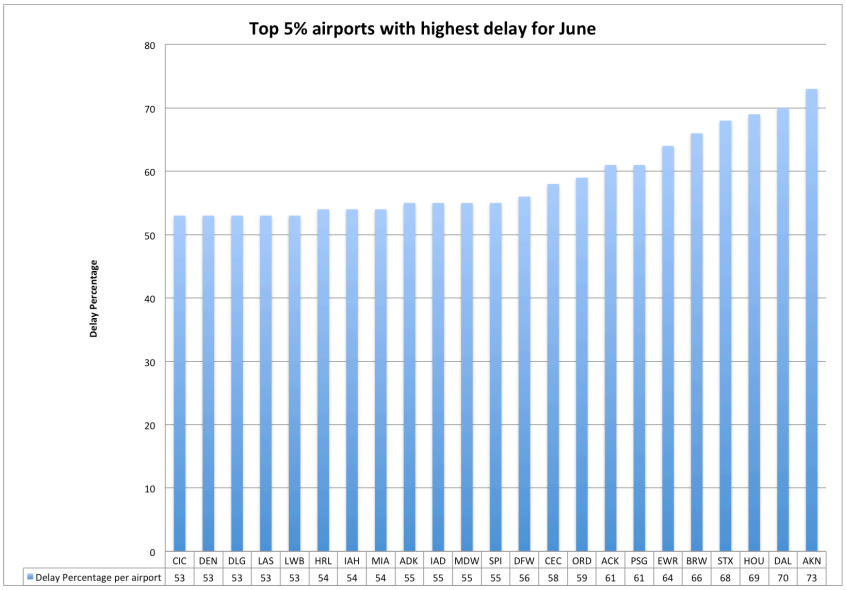




The above plots show the delay percentage by origin airport per month values for May.
The airport with least delay in this month is FCA– Glacier Park Intl, Montana with 3% delay.
The airport with highest delay in this month is ADK – Adak, Alaska with 88% delay.

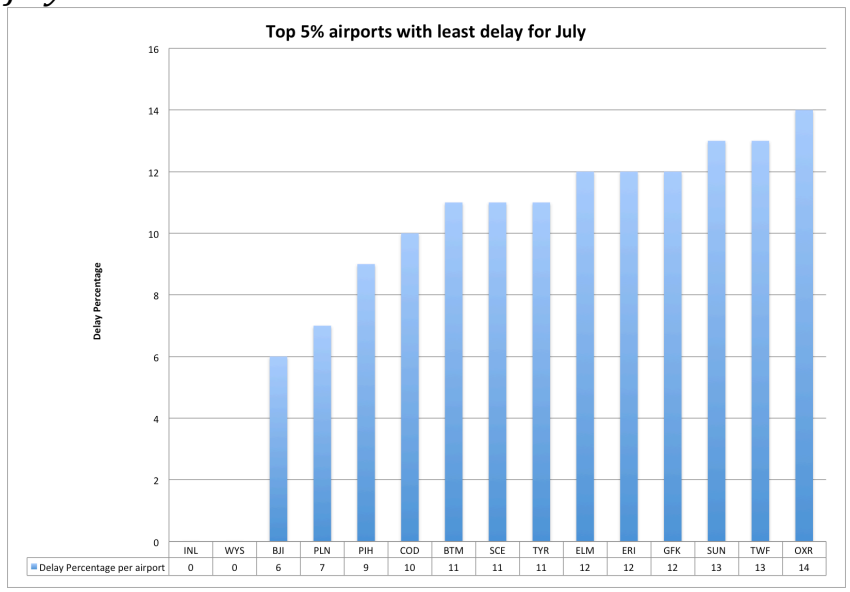
June

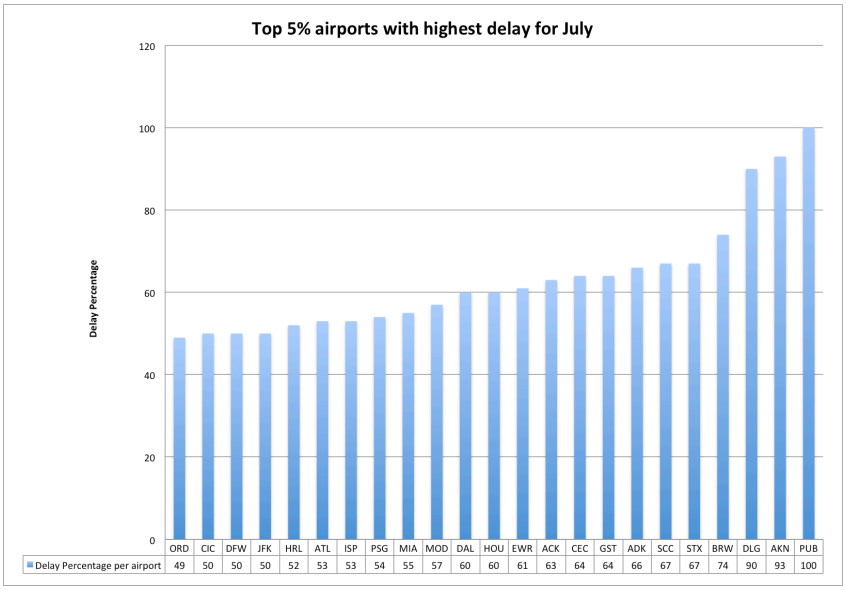




The above plots show the delay percentage by origin airport per month values for June. The airports with least delay in this month is CMX– Houghton County Memorial, Michigan, RFD- Greater Rockford, Illinois, SUX - Sioux Gateway, Iowa with 0% delay. The airport with highest delay in this month is AXN – Chandler, Minnesota with 73% delay.

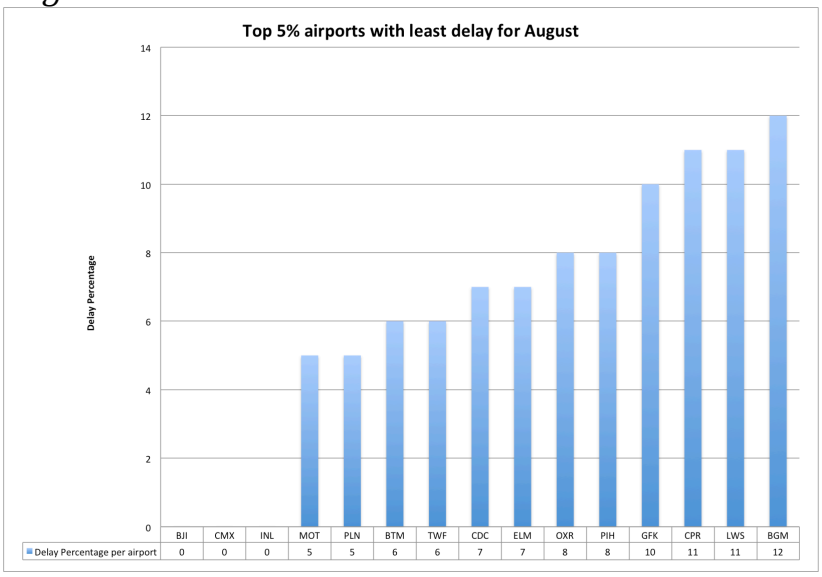
July

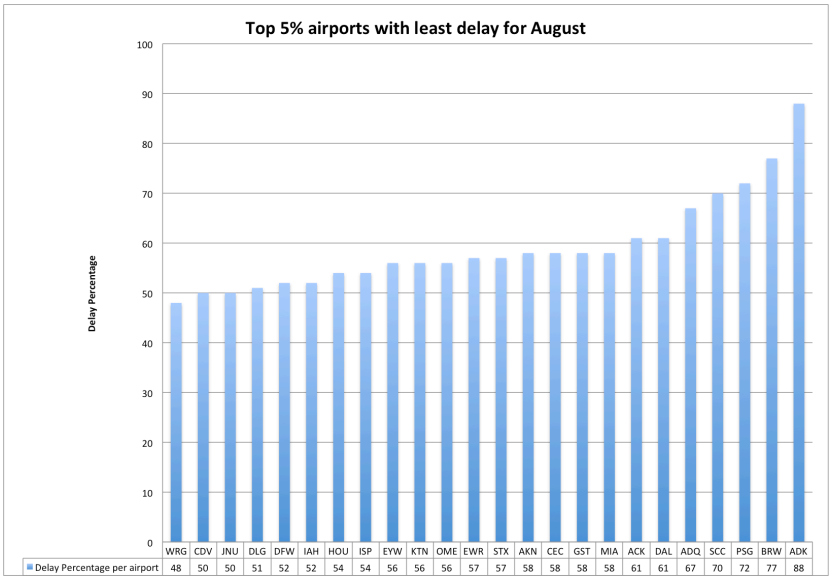




The above plots show the delay percentage by origin airport per month values for July. The airports with least delay in this month is INL–Falls International Airport, Minnesota, WYS- Yellowstone, Montana with 0% delay. The airport with highest delay in this month is PUB – Pueblo Memorial airport, Colorado with 100% delay.

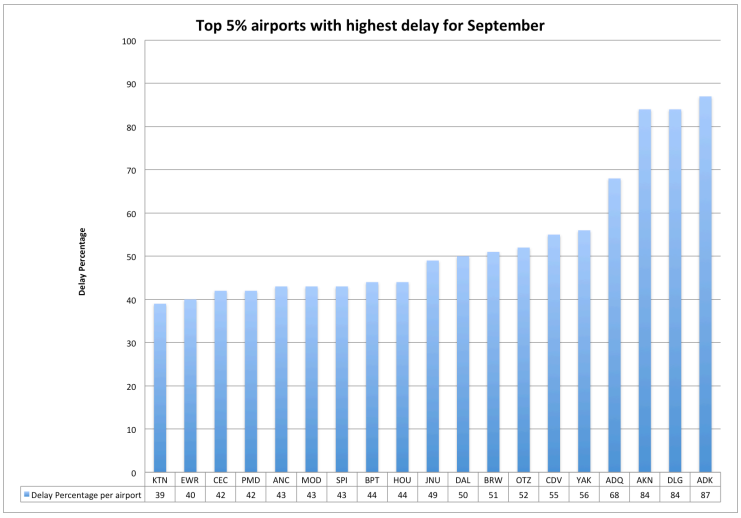
August





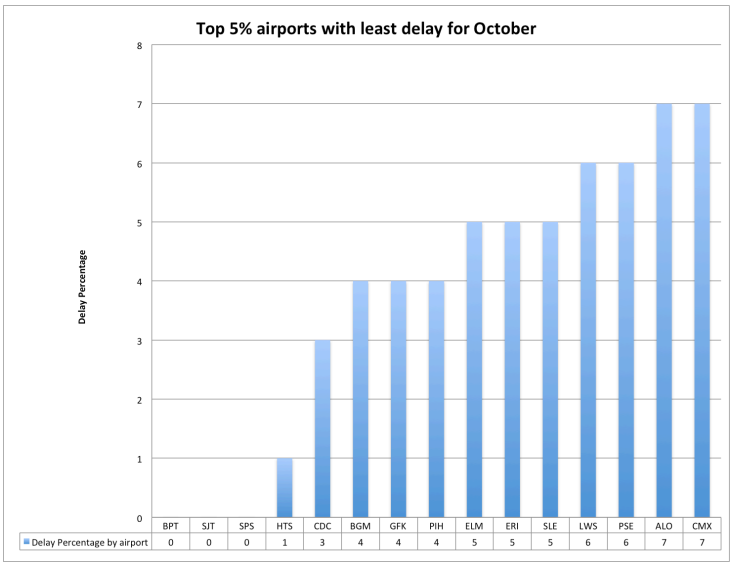
The above plots show the delay percentage by origin airport per month values for August. The airports with least delay in this month is CMX– Houghton County Memorial, Michigan, BJI - Bemidji-Beltrami County, Minnesota, INL – Falls International Airport, Minnesota with 0% delay. The airport with highest delay in this month is ADK – Adak, Alaska with 88% delay.

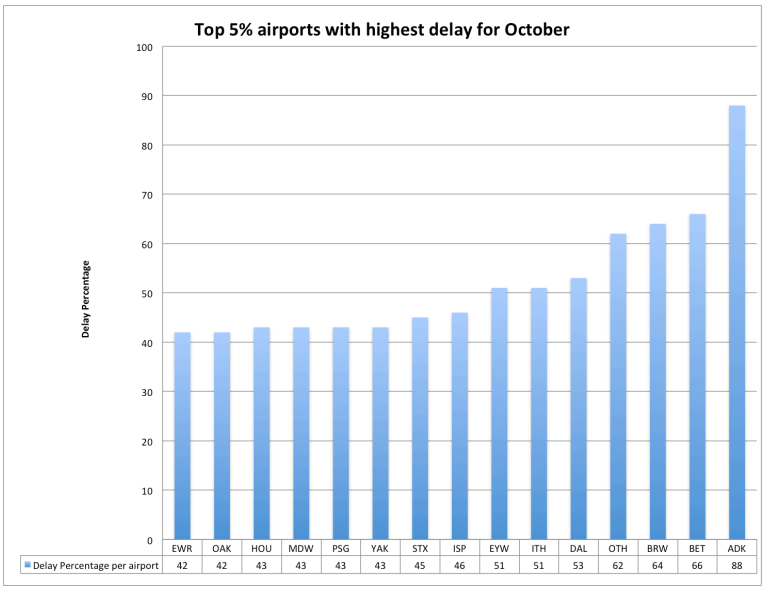
September



The above plots show the delay percentage by origin airport per month values for September.
The airport with least delay in this month is IYK– Inyokern airport, California with 5% delay. The airport with highest delay in this month is ADK –Adak, Alaska with 87% delay.

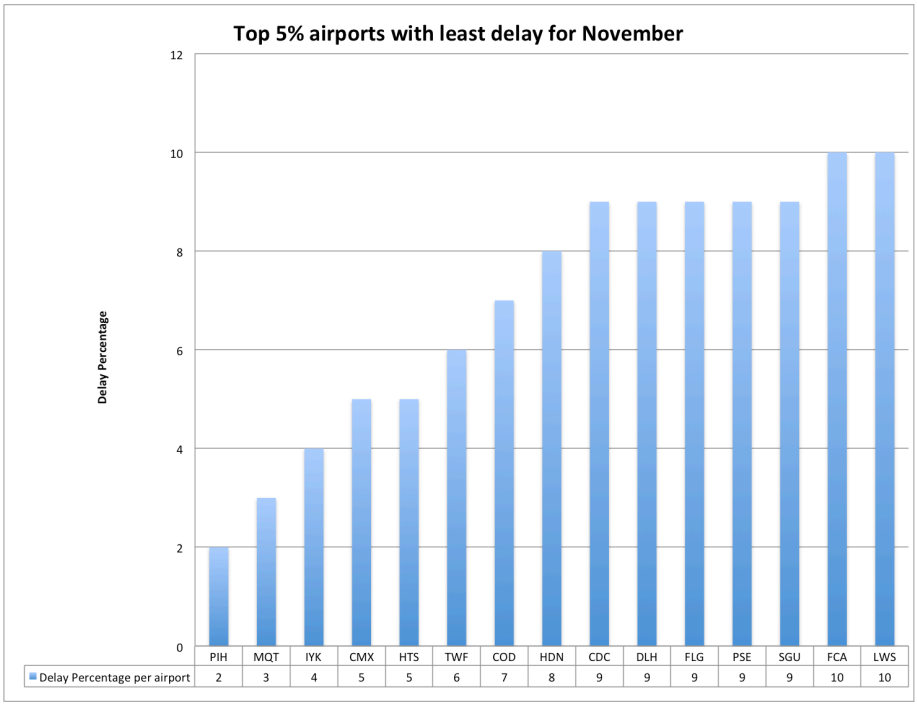
October

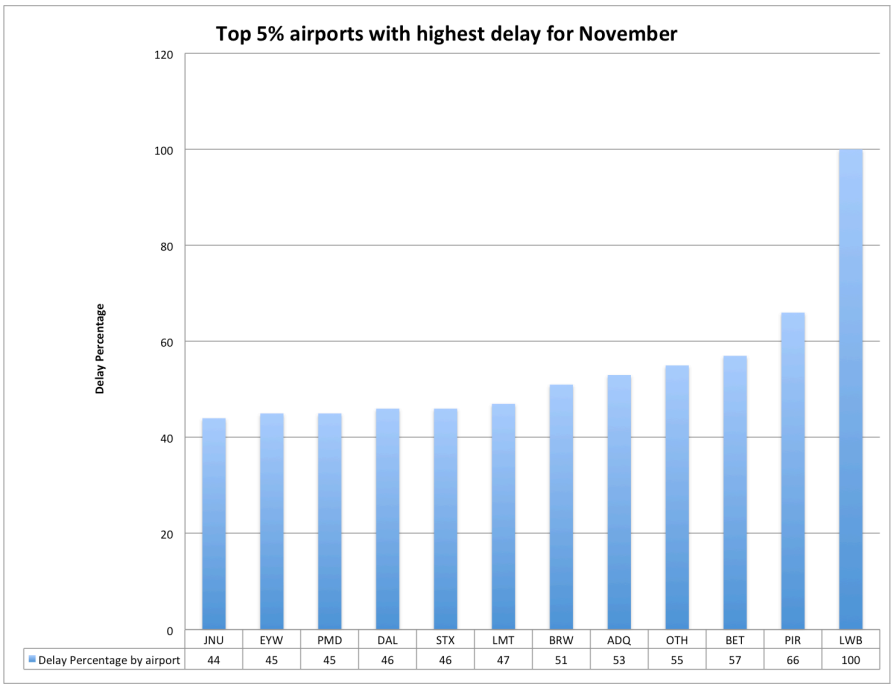




The above plots show the delay percentage by origin airport per month values for October. The airports with least delay in this month is BPT– Southeast Texas Regional, Texas, SJT - San Angelo Regional /Mathis, Texas, SPS – Sheppard AFB/Wichita Falls Municipal airport, Texas with 0% delay. The airport with highest delay in this month is ADK – Adak, Alaska with 88% delay.

November

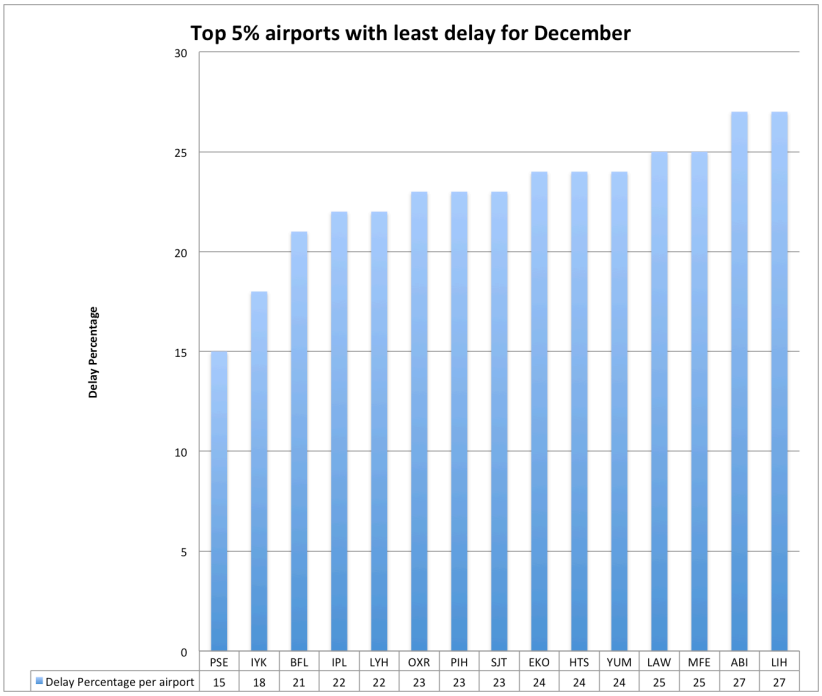


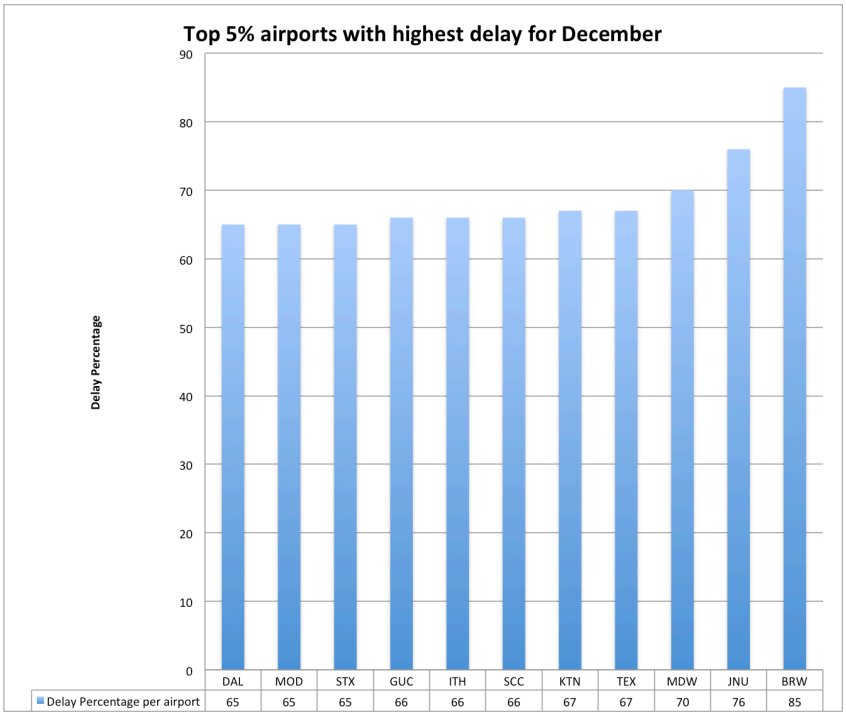


The above plots show the delay percentage by origin airport per month values for September.

The airport with least delay in this month is PIH– Pocatello Regional airport, Idaho with 2% delay. The airport with highest delay in this month is LWB – Greenbrier Valley, West Virginia with 100% delay.

December

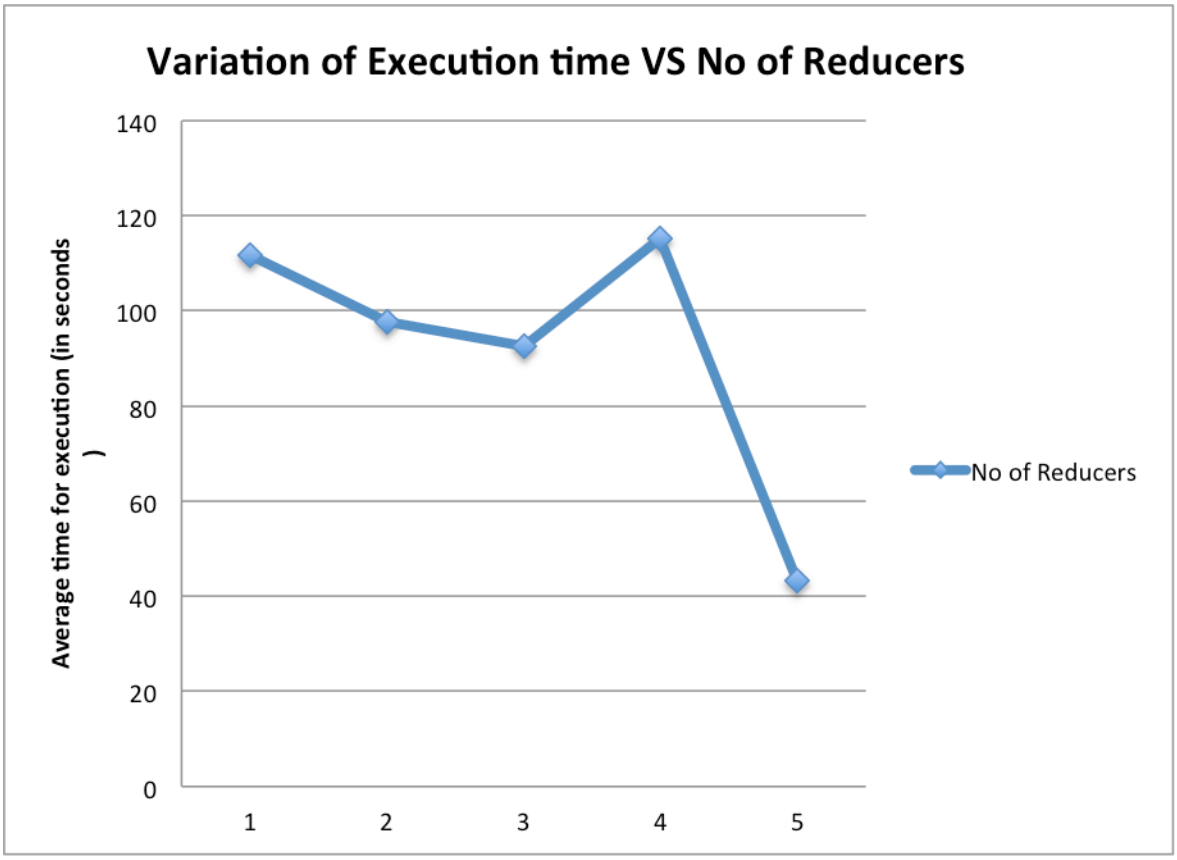




The above plots show the delay percentage by origin airport per month values for December. The airport with least delay in this month is PSE– Mercedita, Puerto Rico with 15% delay. The airport with highest delay in this month is BRW – Wiley Post Will Rogers Memorial, Alaska with 85% delay.

Table showing the variation of average time taken for the job to complete with variation in the number of reducers

SL. No	Job Name	No. of Reduce Tasks	Average Time taken(in seconds)
1	Flights by airport per month	1	111.580839591
2	Flights by airport per month	2	97.830283371
3	Flights by airport per month	4	92.716193061
4	Flights by airport per month	6	115.320081117
5	Flights by airport per month	8	43.237517395



The running time/ execution time for the MapReduce jobs with increase in number of reducers is shown above. It doesn't scale up inversely because of the issues with read/writes and shuffles amongst the nodes.

Table showing the airport with least and highest delay percentage per month

SL. No	Month	Airport with Least delay percentage	Airport with Highest delay percentage
1	January	ITO	MQT
2	February	BLI	RHI
3	March	BLI	RKS
4	April	TWF	AXN
5	May	FCA	ADK
6	June	CMX,RFD,SUX	AXN
7	July	INL, WYS	PUB
8	August	CMX, BJI, INL	ADK
9	September	IYK	ADK
10	October	BPT, SJT, SPS	ADK
11	November	PIH	LWB
12	December	PSE	BRW

SL. No	Airport Code	Airport Name
1	ADK	Adak, Alaska
2	AXN	Chandler, Minnesota
3	BJI	Bemidji-Beltrami County, Minnesota
4	BLI	Bellingham international airport, Washington
5	BPT	Southeast Texas Regional, Texas
6	BRW	Wiley Post Will Rogers Memorial, Alaska
7	CMX	Houghton County Memorial, Michigan
8	FCA	Glacier Park Intl, Montana
9	INL	Falls International, Minnesota
10	ITO	Hilo International, Hawaii
11	IYK	Inyokern airport, California
12	LWB	Greenbrier Valley, West Virginia
13	MQT	Sawyer International Airport, Michigan
14	PIH	Pocatello Regional airport, Idaho
15	PSE	Mercedita, Puerto Rico
16	PUB	Pueblo Memorial airport, Colorado
17	RFD	Greater Rockford, Illinois
18	RHI	Rhinelander-Oneida County, Wisconsin
19	RKS	Rock Springs-Sweetwater County, Wyoming

20	SJT	San Angelo Regional /Mathis, Texas
21	SPS	Sheppard AFB/Wichita Falls Municipal airport, Texas
22	SUX	Sioux Gateway, Iowa
23	TWF	Joslin Field - Magic Valley airport, Idaho
24	WYS	Yellowstone, Montana

5. Future Scope

- The results obtained can be further analyzed to find the co-location patterns and thus try to determine the best possible alternatives if a particular airport has more delays and suggest the alternative to the user using the airport which is collocated to the airport that has more delays.

- Also, the relation between location of the airport and the climate conditions can be assessed.

- Also the best airports to fly from with respect to a particular month can be determined based on the delay percentage values from the results obtained to help the user to have a smoother travel.

6. Challenges

- *The effect of read-writes and shuffles:* The effect of using more than one reducer might sometimes be tricky to understand since, if the data has to be transferred to a different node it will have the overhead of reading from the actual node on which the data is present and the writes to the node which are available which may in turn affect the performance by increasing the running time. Also with more number of reducers the overhead of reading the outputs from multiple files and then processing will also increase the running time. Also if the size of the input dataset is very large then the shuffles take more time thus effecting the performance of the jobs.

- *The effect of cluster infrastructure:* The cluster infrastructure used to perform the analytics task has a major effect on the performance of the MapReduce jobs. If there is extreme loads with comparatively less number of nodes it might affect the running time and thus the performance.

- *The effect of hypothesis with respect to the data preprocessing*: The data pre-processing also has a significant effect on the results. Since there may be multiple factors like missing values and NaN's in the dataset and the results vary depending on the choice of data preprocessing assumptions

7. References

[1] http://www.hadoopmaterial.com/2013/10/how-good-are-citys-farmers-markets.html
[2] https://www.code.google.com
[3] http://www.orzota.com/step-by-step-mapreduce-programming/