A4

Abhishek Ravi Chandran, Chinmayee Vaidya

February 12, 2016

Design

In this assignment we perform simple linear regression with the predictor as time to try and find out the cheapest price of a carrier for a certain time N.

We run 2 jobs one after the other to perform the analysis. The first job groups all the flights together by year and carrier and get the values scheduled elapsed time and price. This is then passed to the reducer, where we apply simple regression by calculating slope and intercept for the data. The we get the linear regression for N by calculating intercept+slope*N.

This is then output to a file.

Then we have a java method that reads the ouput files from s3 to compare all the flights every year. The flight that has the lowest price the most number of times for all the years is returned and passed to the second job.

In the second job we only filter out the records for the flight given be the above method. Then we calculate the median for every week. We used our fast median method that was implemented last time.

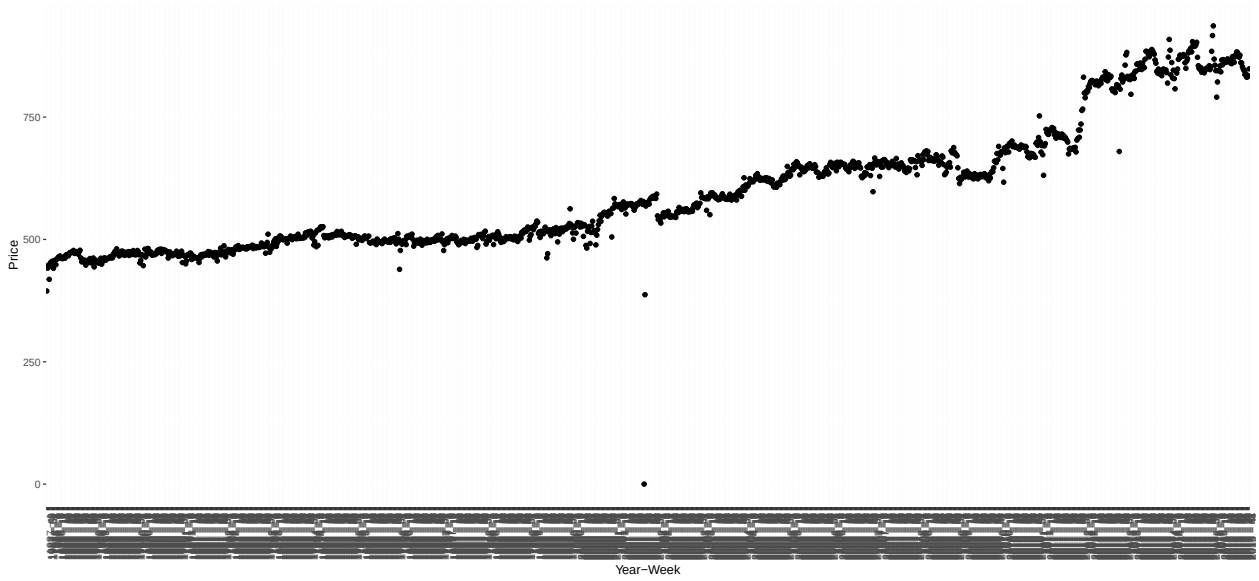
The output is then downloaded from s3, which is then read by R to produce the report.

The graph below shows the cheapest fligts for N=1 and N=200

Median Graph:

X-Scale - Weeks

Carrier: UA For N=1



Median Graph:

X-Scale - Weeks

Carrier: AS For N=200