Map Reduce Assignment 4 Report

Authors: Akash Singh, Surekha Jadhwani

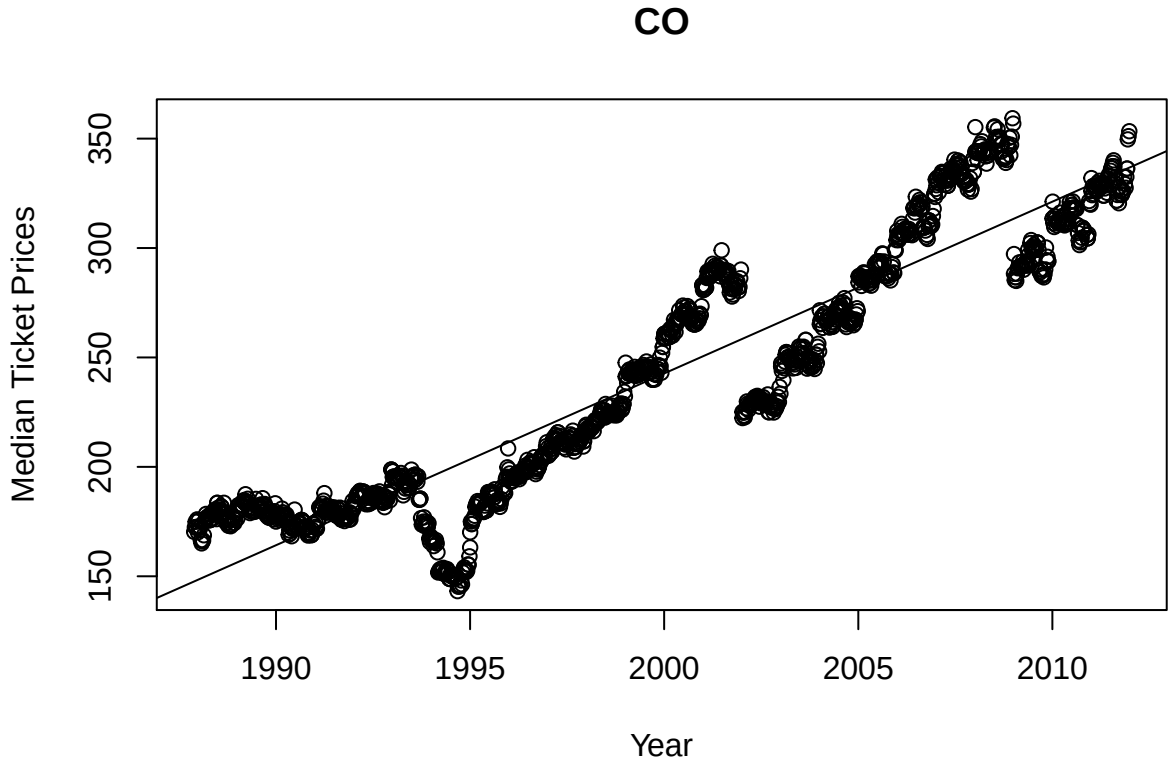
Description:

The aim of this assignment is to figure out the cheapest airline by finding the relation between average ticket prices and scheduled flight minutes over a number of years using the Bureau of Transport Statistics' On-time Performance (OTP) dataset which has over 27 years of air travel information about flights in the USA. In our program, sane records with Average ticket price missing or negative or higher than 100000 will not be used for any calculation.

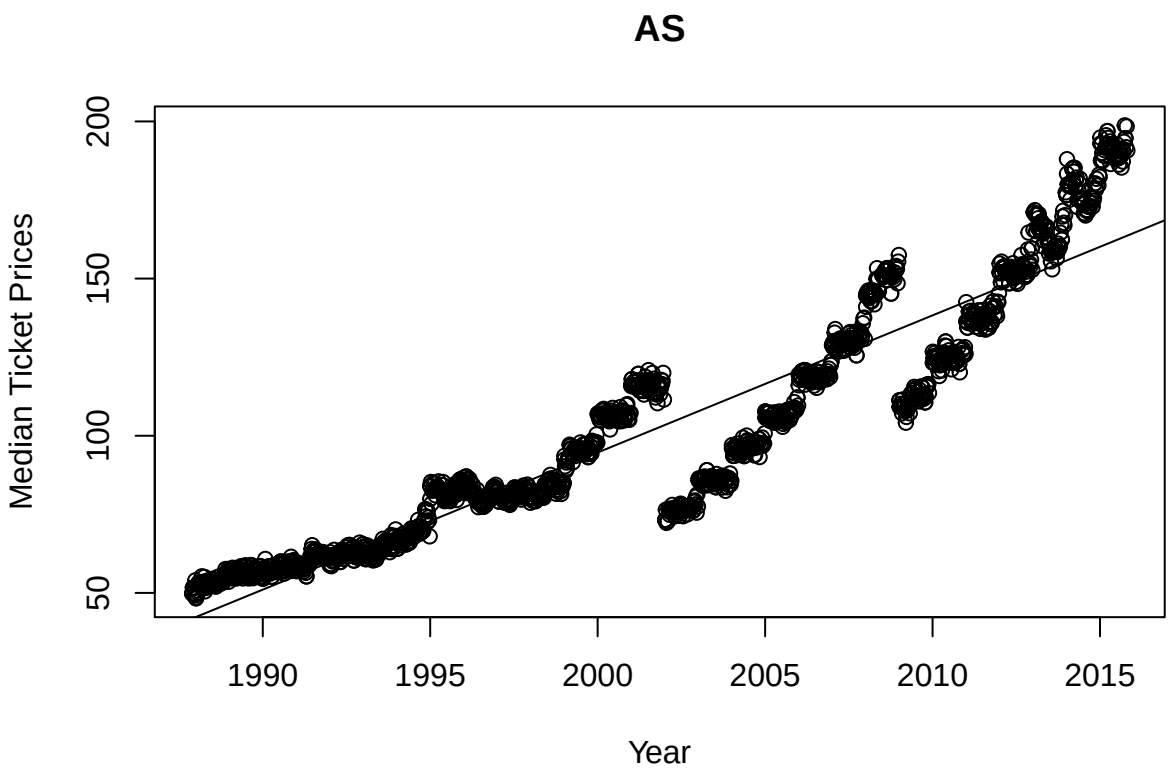
Implementation:

We are first using MapReduce program to extract carriers, their average ticket prices and scheduled flight minutes for all the years in the given dataset. The output of this program is then used by R program to compute intercept and slope of simple linear regression model. Output of R program is later used by another MapReduce program which finds the least expensive carrier(carrier with lowest price at N for the most years.) and computes its week-wise median prices for all the years. This is finally used for plotting the graph.

The graph of the median price of the least expensive carrier for each week of the entire dataset, for N=1 using R Script is shown below:



The graph of the median price of the least expensive carrier for each week of the entire dataset, for N=200 using R Script is shown below:



Analysis:

We observed that as the scheduled flight minutes increased, the median ticket prices decreased. Another observation is that the ticket prices have increased over the period of many years without considering inflation.