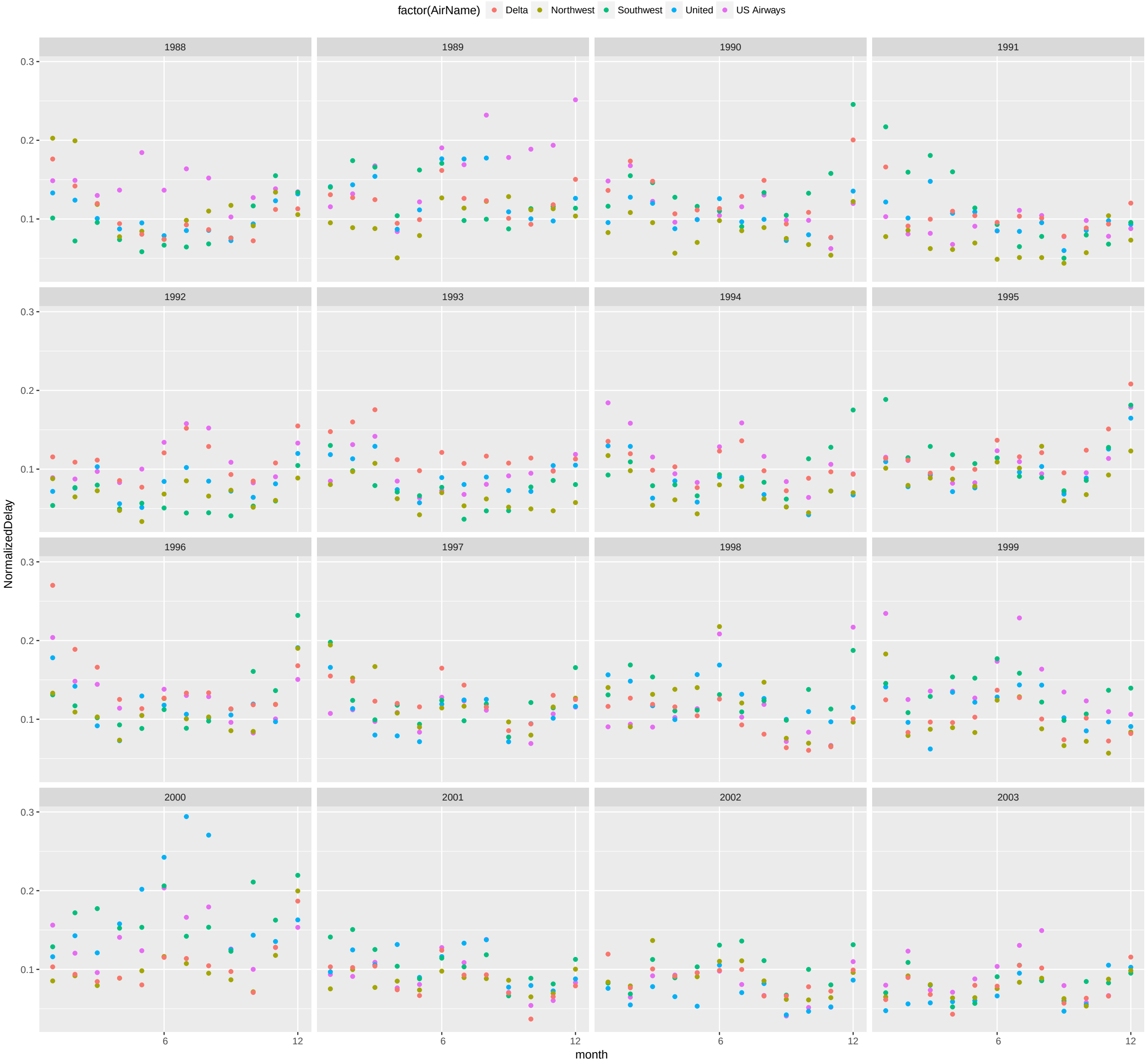
Shreysa A8 Assignment

Summary of the Program Design

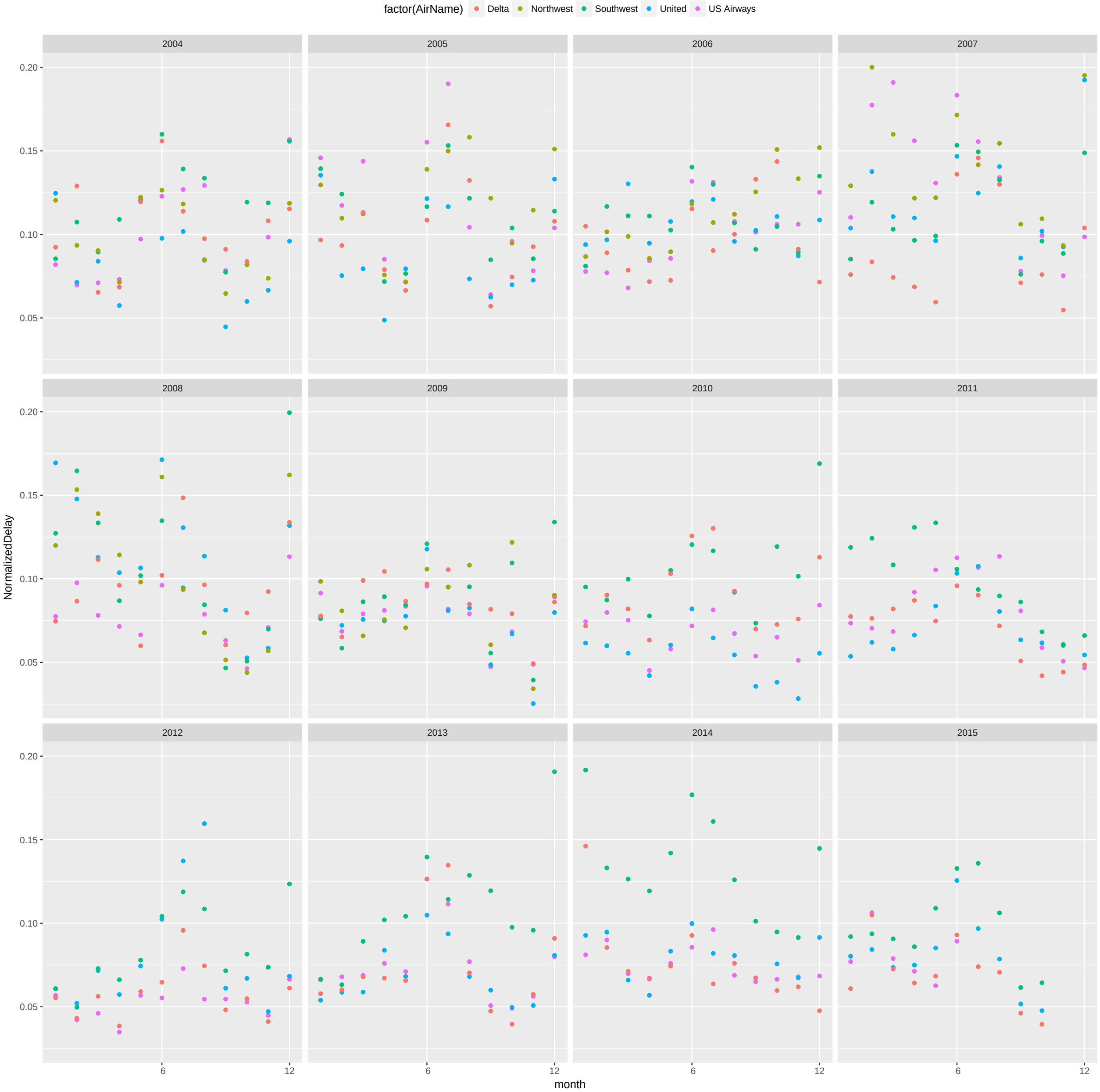
The goal of this assignment was to provide visualization of mean delays of the five most active airlines and for the five most active airports in the country. I used `1 Map-Reduce job` that does the initial cleaning of the airline dateset and also outputs the **airline/airport_identifier, airlineid/airportid, year and month as key and normalized delay as value** . The reducer then calculates the normalized mean delay and outputs **airlineId/airportId, year, month as key and airline/airport_identifier, normalized mean delay and number of records for each key as value**. The output is saved in a results.csv file. This output is then loaded into R, seperated into airline and airport data. The top 5 most active airlines and airports are picked and the results are generated for visualization in R.

Plot 1: Top 5 most Active Airlines in the period 1988-2003 with their Mean Delay



The Plots 1 and 2 represent the Normalized mean delays for the Top 5 busiest airlines for the period 1988 to 2015. The top 5 busiest airlines for this period were **Delta Airlines, Northwest Airlines, Southwest airlines, United airlines and US Airways**.

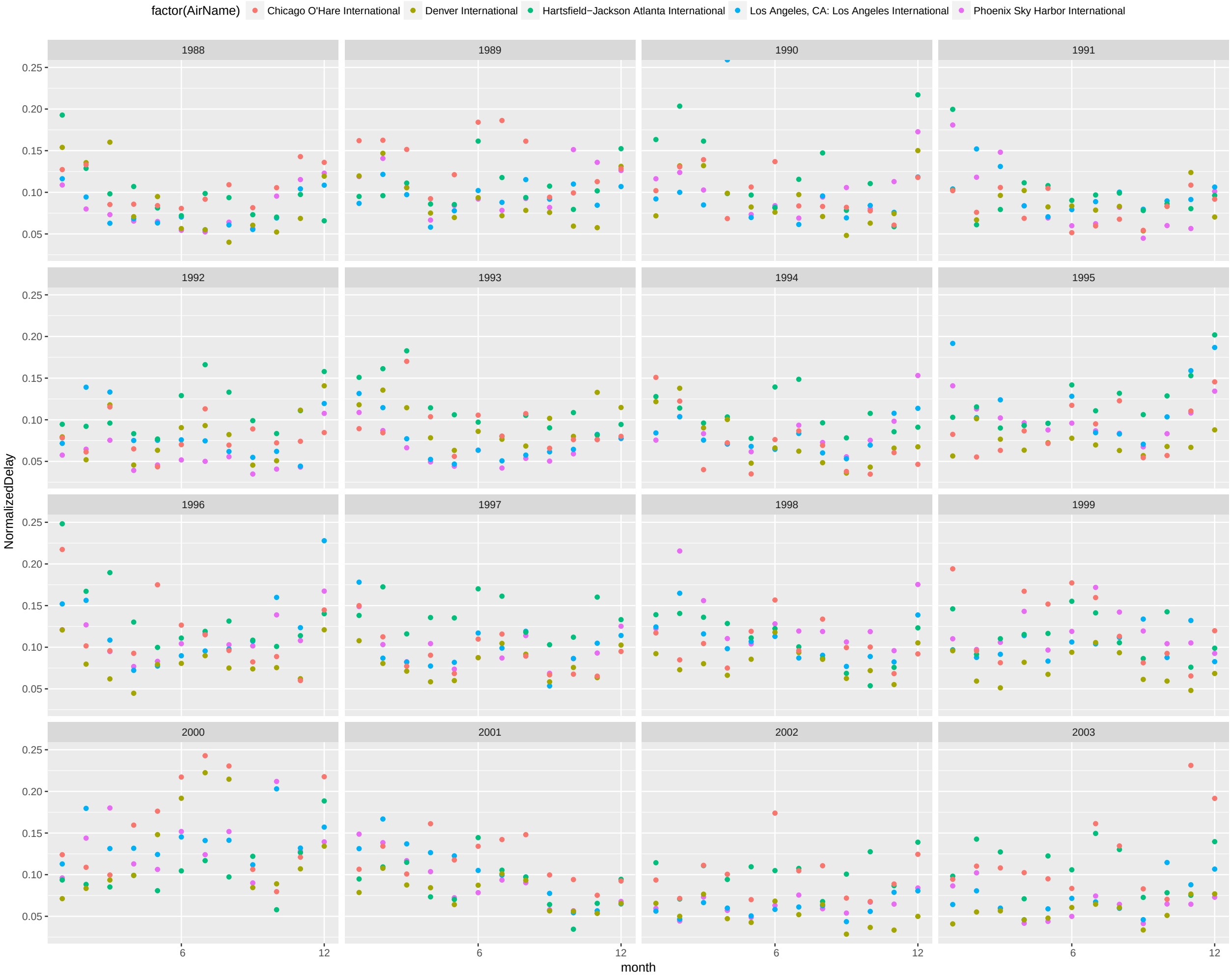
Plot 2: Most Active Airlines in the period 2004-2015 with their Mean Delay



For the period 1987 to 1990 `US Airways` topped the list for delays, however towards the end of 1990 `Southwest and Delta` were more delayed. Thereafter, for a majority period from 1991 to 1998 `Delta` was the most delayed flight, closely followed by `Southwest and US airways` but after 1998 Southwest has topped the list for delays among the top 5 busiest airlines.

Furthermore, **Northwest Airlines** was the least delayed among these 5 top airlines from 1987 to 2004 but after 2004 it topped the list in delays till 2009. It can be seen that after 2009 Northwest Airlines does not appear in the graph as it was merged with Delta airlines in Jan 2010 due to bankcruptcy and closed down operations as Northwest in 2009 Dec.

Plot 3: Top 5 most Active Airports in the period 1988-2003 with their Mean Delay

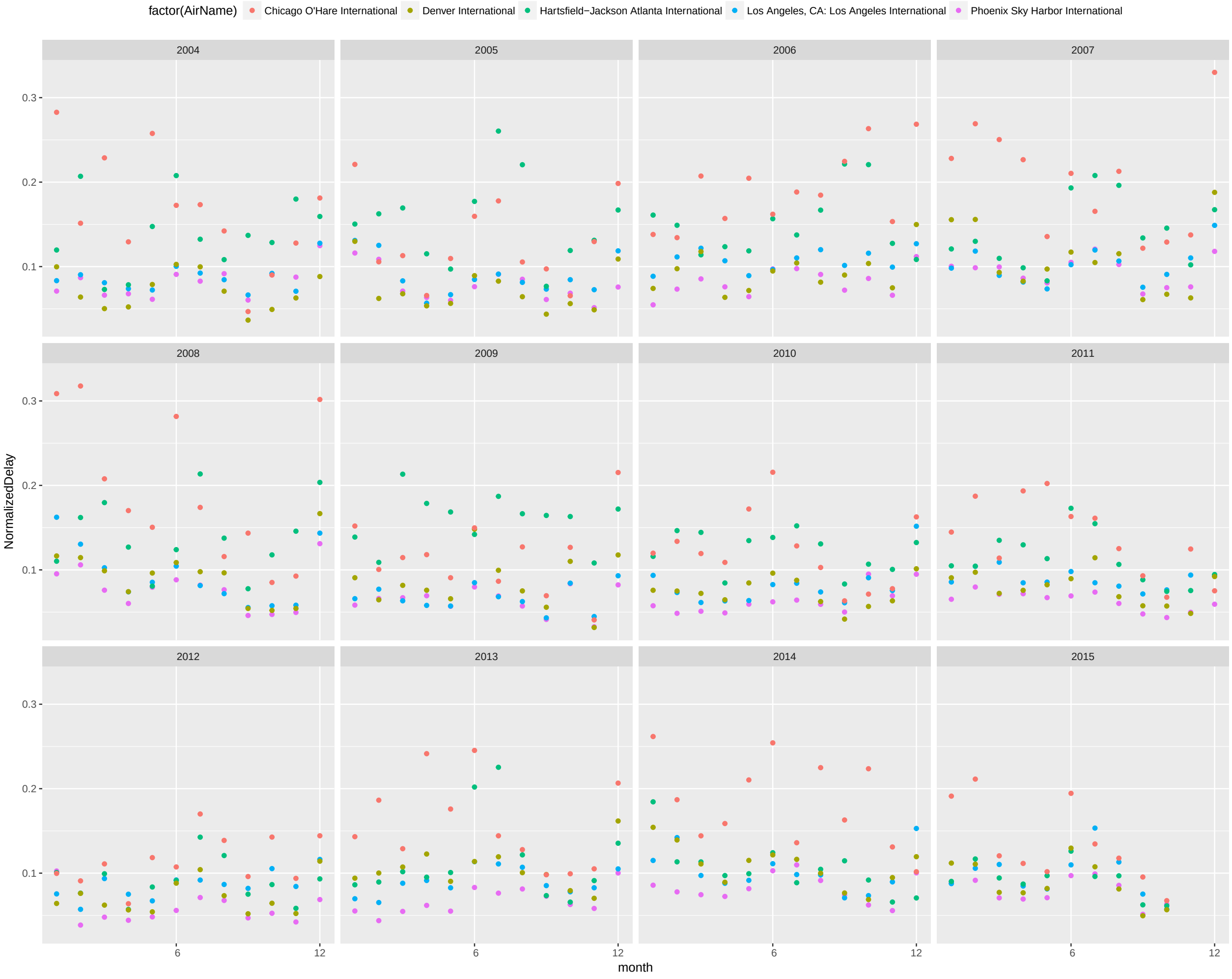


The Plots 3 and 4 represent the Normalized delay for the Top 5 airports for the period 1988 to 2015. The top 5 busiest airports are: **Hartsfield-Jackson Atlanta International, Chicago O'Hare International, Los Angeles, CA: Los Angeles International, Phoenix Sky Harbor International and Denver International Airport**. For the period 1987 to 1996 `Hartsfield-Jackson Atlanta International` Airport was the one with most delays however after 1996 `Chicago O'Hare International` Airport has been the most prominent in terms of delays out of the top 5 busiest airports.

`Denver International Airport` was the one with least delays among these till 2010 however post that `Phoenix Sky Harbor International Airport` has the minimum delays out of the set.

In the current implementation of the code, I converted the CRSDeptTime, CRSArrTime, DepTime and ArrTime from the format HHMM to minutes and got the above results. The original implementation in A4 lacked this and due to that we were getting wrong and different results for the top 5 most active airlines as well as the airports.

Plot 4: Top 5 most Active Airports in the period 2004-2015 with their Mean Delay



Conclusion

The exercise helped in observing few interesting facts:

- The job took **14.3 minutes to execute on AWS EMR c4.large** machine.
- `Delta Airlines` was the most delayed airline from 1991 to 1998 but post that `Southwest airline` has been the most delayed airline among the top 5 busiest airlines while `NorthWest` was the one with minimum delays in the set till 2004 and after it merged with delta.
- `NorthWest Airline` was one of the top 5 busiest airline and also had minimum delays among the this set but close to the year when it was shut down the mean delay of the airline grew with it being the one with most delays in the years 2005-2009.
- The mean delays for the airlines were considerably `reduced` after year `2000`.
- It was noticed that the maximum delays for all airlines and airports were in the month of `December` which could primarily be because of December being the holiday month, most people travel during that time. Additionally, `September` has the least number of delays which could be because schools reopen after summer vacations and less people travel then.