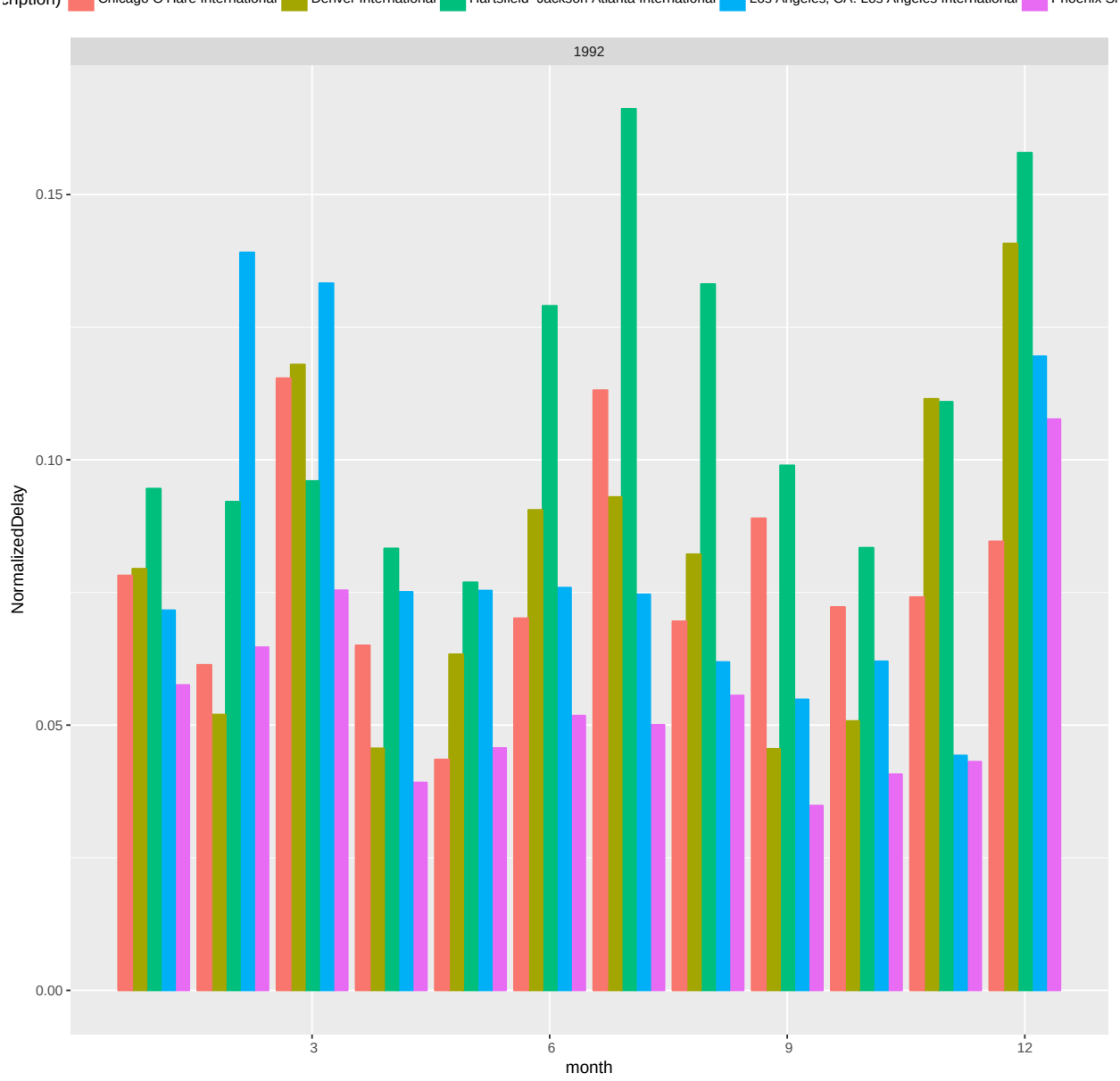
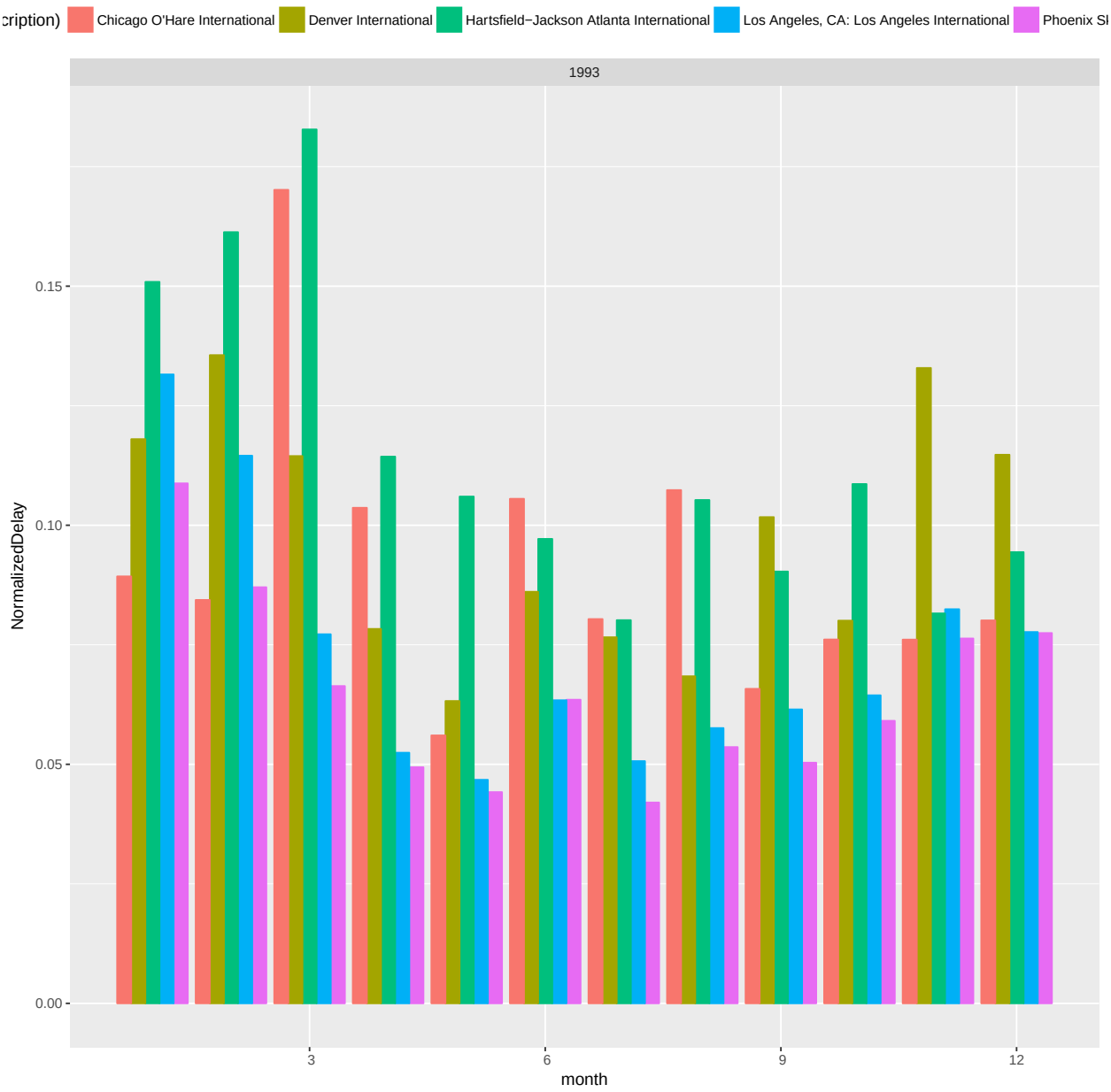
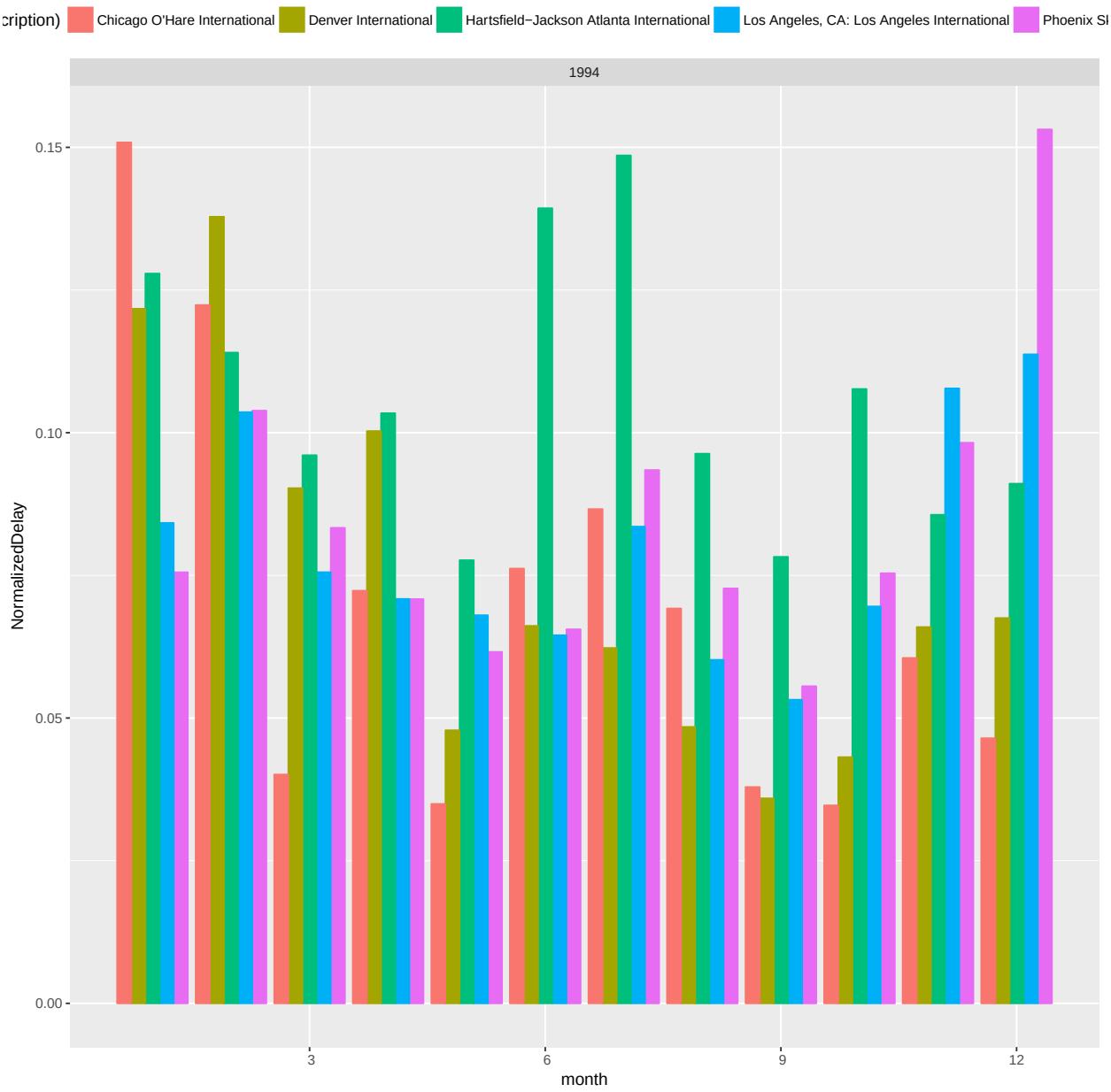
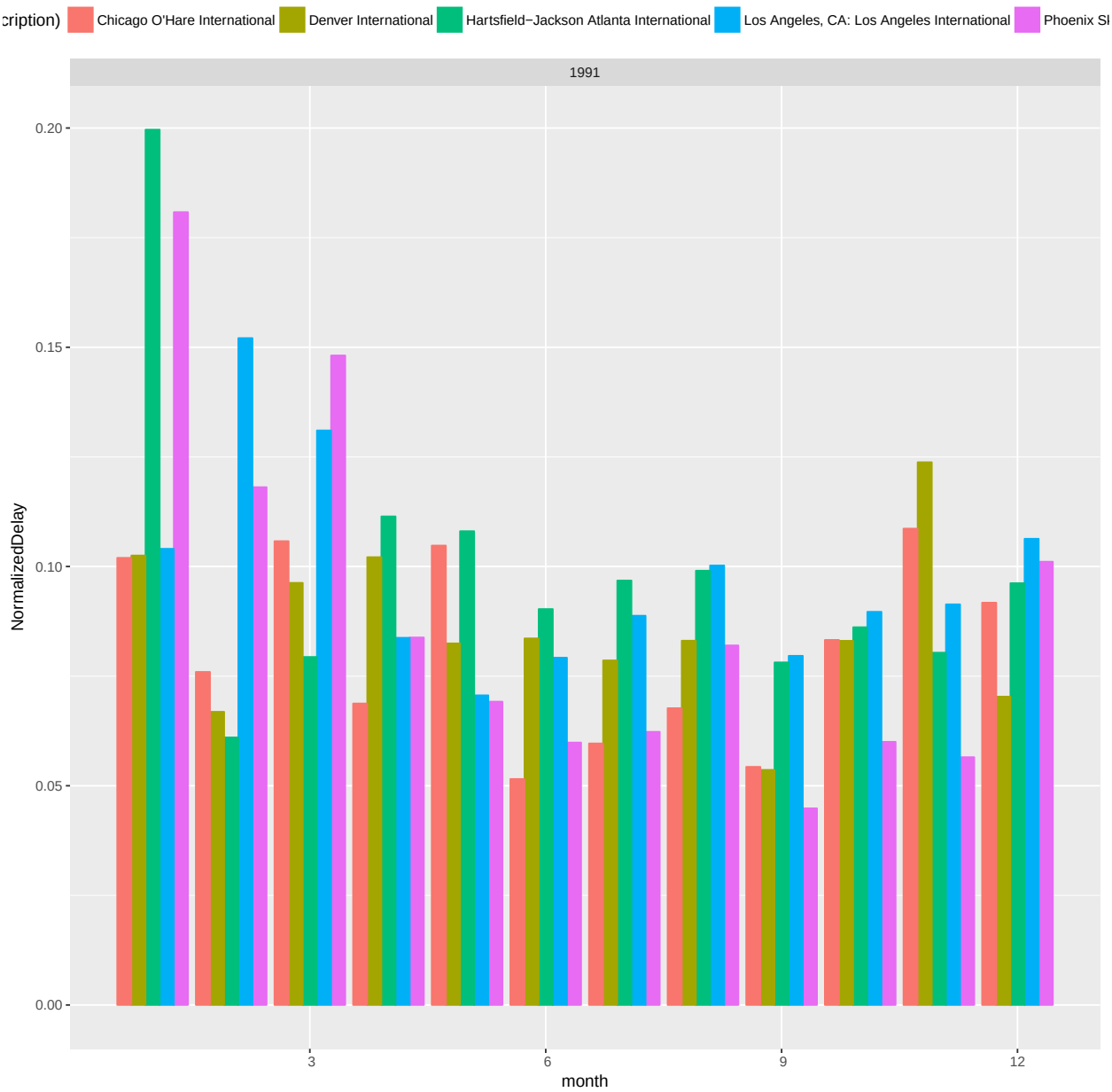
Program Design

The goal of this assignment was to perform provide visualization mean delay of the five most active airlines and for the five most active airports in the country. I have used `1 Map-Reduce job` that does the initial cleaning of the airline dateset and then gets the normalized delay(with respect to the scheduled length of each flight) and mean delay for each month for a particular year. The output is a results.csv file placed in the `output` folder that has airportId and year or airlineId and year as key and value as NormalizedDelay for the year and for each month seperated by `","` starting from Jan. ##### Plot 1: Most Active Airlines in the period 2011-2015 with their Mean Delay




cription)
Chicago O'Hare International
Denver International
Hartsfield−Jackson Atlanta International
Los Angeles, CA: Los Angeles International
Phoenix SI
1993
NormalizedDelay
0.15
0.10
0.05
0.00
3
6
9
12
month


cription)
Chicago O'Hare International
Denver International
Hartsfield−Jackson Atlanta International
Los Angeles, CA: Los Angeles International
Phoenix SI
1994
0.15
0.10
0.05
0.00
NormalizedDelay
3
6
9
12
month




```
grp_by_airline <- aggregate(airline_data$count,by=list(airlineId=airline_data$airlineId), FUN=sum)
dfMeanSortedAfter<-arrange(grp_by_airline, desc(x))
TopAirlineAfter=dfMeanSortedAfter[1:6,]
active_airlines_tableAfter <- sqldf("select F.airlineId, F.year,F.month, F.NormalizedDelay,T.x as count
active_airlinesAfter <- sqldf("select N.Description, F.airlineId, F.year, F.month, F.NormalizedDelay, F
```

The Plots 1-? represent the Normalized delay for the Top 5 airlines. The available data is from Dec 1987 to Oct 2015. For the period 1988 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 follwed by Southwest and US airways. It is interesting to note that **Northwest Airlines** was the least delayed among these 5 top airlines from 1987 to 2004 but after 2004 it topped the list for 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 2010 Jan due to bankcruptcy and closed down operations as Northwest in 2009 Dec.

After 1998 Southwest has topped the list for delays among airlines while American Airlines which appeared

in the top 5 busiest airlines after NorthWest stopped its services has the least delay among the set.

It was also noticed that after 2011 the mean delays for all the airlines have reduced considerably.

The Plots 1-? represent the Normalized delay for the Top 5 airports. The available data is from Dec 1987 to Oct 2015. For the period 1988 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 follwed by Southwest and US airways. It is interesting to note that **Northwest Airlines** was the least delayed among these 5 top airlines from 1987 to 2004 but after 2004 it topped the list for delays till the time it was merged with Delta airlines in 2010 Jan due to bankcruptcy and closed down operations as Northwest in 2009 Dec. After 1998 Southwest has topped the list for delays among the other 4 airlines. Another point to note is that the airlines are the most delayed during the month of december (this could be because of holiday season).

Conclusion

The exercise helped in observing few interesting facts.

It was noticed that the maximum delays in all the airlines are in the month of December which could primarily be, 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.