

Data Exploring and Data Wrangling - NYCFlights13 Dataset

Vaibhav Walvekar

```
# Load standard libraries  
library(tidyverse)  
library(nycflights13)
```

```
## Warning: package 'nycflights13' was built under R version 3.3.2
```

Importing and Inspecting Data:

```
# Get details about nycflights13 dataset  
?nycflights13  
ls("package:nycflights13")  
?flights
```

```
# Load different data points from the nycflights13 library  
airlines_data <- airlines  
airports_data <- airports  
flights_data <- flights  
planes_data <- planes  
weather_data <- weather
```

- The nycflights13 dataset is a collection of data pertaining to different airlines flying from different airports in NYC, also capturing flight, plane and weather specific details during the year of 2013. The data was collected into these five different branches. This method of collecting data helps us to work on individual aspects of the whole large dataset and also we can combine together multiple aspects to do some complex data analysis. There are also 3-4 database versions of nycflights13 dataset which cache the data from nycflights13 database in a local database, helping in joining tables on natural keys efficient. The source of flights dataset is RITA, Bureau of transportation statistics, http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=236.

The variables in flights dataset represent as below:

```
# Variables in flights dataset  
?flights
```

year,month,day - Date of departure

dep_time,arr_time - Actual departure and arrival times, local tz.

sched_dep_time,sched_arr_time - Scheduled departure and arrival times, local tz.

dep_delay,arr_delay - Departure and arrival delays, in minutes. Negative times represent early departures/arrivals.

hour,minute - Time of scheduled departure broken into hour and minutes.

carrier - Two letter carrier abbreviation. See airlines to get name

tailnum - Plane tail number

flight - Flight number

origin,dest - Origin and destination. See airports for additional metadata.

air_time - Amount of time spent in the air

distance - Distance flown

time_hour - Scheduled date and hour of the flight as a POSIXct date. Along with origin, can be used to join flights data to weather data.

```
# Inspecting flights dataset  
sapply (flights_data, class)
```

```
## $year  
## [1] "integer"  
##  
## $month  
## [1] "integer"  
##  
## $day  
## [1] "integer"  
##  
## $dep_time  
## [1] "integer"  
##  
## $sched_dep_time  
## [1] "integer"  
##  
## $dep_delay  
## [1] "numeric"  
##  
## $arr_time  
## [1] "integer"  
##  
## $sched_arr_time  
## [1] "integer"  
##  
## $arr_delay  
## [1] "numeric"  
##  
## $carrier  
## [1] "character"  
##  
## $flight  
## [1] "integer"  
##  
## $tailnum  
## [1] "character"  
##  
## $origin  
## [1] "character"  
##  
## $dest
```

```
## [1] "character"
##
## $air_time
## [1] "numeric"
##
## $distance
## [1] "numeric"
##
## $hour
## [1] "numeric"
##
## $minute
## [1] "numeric"
##
## $time_hour
## [1] "POSIXct" "POSIXt"
```

```
head(flights_data)
```

```
## # A tibble: 6 × 19
##   year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>   <int>         <int>         <dbl>   <int>
## 1  2013     1     1     517             515           2     830
## 2  2013     1     1     533             529           4     850
## 3  2013     1     1     542             540           2     923
## 4  2013     1     1     544             545          -1    1004
## 5  2013     1     1     554             600          -6     812
## 6  2013     1     1     554             558          -4     740
## # ... with 12 more variables: sched_arr_time <int>, arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,
## #   time_hour <dtm>
```

```
tail(flights_data,5)
```

```
## # A tibble: 5 × 19
##   year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>   <int>         <int>         <dbl>   <int>
## 1  2013     9    30     NA             1455           NA     NA
## 2  2013     9    30     NA             2200           NA     NA
## 3  2013     9    30     NA             1210           NA     NA
## 4  2013     9    30     NA             1159           NA     NA
## 5  2013     9    30     NA             840            NA     NA
## # ... with 12 more variables: sched_arr_time <int>, arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,
## #   time_hour <dtm>
```

```
flights_newdata <- flights_data[order(flights_data$month,flights_data$day),]
tail(flights_newdata,5)
```

```
## # A tibble: 5 × 19
```

```
##   year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>   <int>         <int>         <dbl>   <int>
## 1  2013   12   31     NA           705           NA     NA
## 2  2013   12   31     NA           825           NA     NA
## 3  2013   12   31     NA          1615           NA     NA
## 4  2013   12   31     NA           600           NA     NA
## 5  2013   12   31     NA           830           NA     NA
## # ... with 12 more variables: sched_arr_time <int>, arr_delay <dbl>,
## #   carrier <chr>, flight <int>, tailnum <chr>, origin <chr>, dest <chr>,
## #   air_time <dbl>, distance <dbl>, hour <dbl>, minute <dbl>,
## #   time_hour <dtm>
```

```
dim(flights_data)
```

```
## [1] 336776    19
```

```
summary(flights_data)
```

```
##      year      month      day      dep_time
## Min.   :2013   Min.   : 1.000   Min.   : 1.00   Min.   : 1
## 1st Qu.:2013   1st Qu.: 4.000   1st Qu.: 8.00   1st Qu.: 907
## Median :2013   Median : 7.000   Median :16.00   Median :1401
## Mean   :2013   Mean   : 6.549   Mean   :15.71   Mean   :1349
## 3rd Qu.:2013   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1744
## Max.   :2013   Max.   :12.000   Max.   :31.00   Max.   :2400
##
## NA's   :8255
## sched_dep_time  dep_delay      arr_time  sched_arr_time
## Min.   : 106   Min.   : -43.00   Min.   : 1      Min.   : 1
## 1st Qu.: 906   1st Qu.: -5.00   1st Qu.:1104   1st Qu.:1124
## Median :1359   Median : -2.00   Median :1535   Median :1556
## Mean   :1344   Mean   : 12.64   Mean   :1502   Mean   :1536
## 3rd Qu.:1729   3rd Qu.: 11.00   3rd Qu.:1940   3rd Qu.:1945
## Max.   :2359   Max.   :1301.00   Max.   :2400   Max.   :2359
##
## NA's   :8255   NA's   :8713
##   arr_delay      carrier      flight      tailnum
## Min.   : -86.000   Length:336776   Min.   : 1      Length:336776
## 1st Qu.: -17.000   Class :character 1st Qu.: 553     Class :character
## Median : -5.000   Mode  :character  Median :1496     Mode  :character
## Mean   :  6.895                                     Mean   :1972
## 3rd Qu.: 14.000                                     3rd Qu.:3465
## Max.   :1272.000                                    Max.   :8500
## NA's   :9430
##   origin      dest      air_time      distance
## Length:336776   Length:336776   Min.   : 20.0   Min.   : 17
## Class :character  Class :character 1st Qu.: 82.0   1st Qu.: 502
## Mode  :character  Mode  :character  Median :129.0   Median : 872
##
## Mean   :150.7   Mean   :1040
## 3rd Qu.:192.0   3rd Qu.:1389
## Max.   :695.0   Max.   :4983
## NA's   :9430
##   hour      minute      time_hour
## Min.   : 1.00   Min.   : 0.00   Min.   :2013-01-01 05:00:00
## 1st Qu.: 9.00   1st Qu.: 8.00   1st Qu.:2013-04-04 13:00:00
```

```
## Median :13.00 Median :29.00 Median :2013-07-03 10:00:00
## Mean :13.18 Mean :26.23 Mean :2013-07-03 05:02:36
## 3rd Qu.:17.00 3rd Qu.:44.00 3rd Qu.:2013-10-01 07:00:00
## Max. :23.00 Max. :59.00 Max. :2013-12-31 23:00:00
##
```

```
str(flights_data)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame': 336776 obs. of 19 variables:
## $ year : int 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 2013 ...
## $ month : int 1 1 1 1 1 1 1 1 1 1 ...
## $ day : int 1 1 1 1 1 1 1 1 1 1 ...
## $ dep_time : int 517 533 542 544 554 554 555 557 557 558 ...
## $ sched_dep_time: int 515 529 540 545 600 558 600 600 600 600 ...
## $ dep_delay : num 2 4 2 -1 -6 -4 -5 -3 -3 -2 ...
## $ arr_time : int 830 850 923 1004 812 740 913 709 838 753 ...
## $ sched_arr_time: int 819 830 850 1022 837 728 854 723 846 745 ...
## $ arr_delay : num 11 20 33 -18 -25 12 19 -14 -8 8 ...
## $ carrier : chr "UA" "UA" "AA" "B6" ...
## $ flight : int 1545 1714 1141 725 461 1696 507 5708 79 301 ...
## $ tailnum : chr "N14228" "N24211" "N619AA" "N804JB" ...
## $ origin : chr "EWR" "LGA" "JFK" "JFK" ...
## $ dest : chr "IAH" "IAH" "MIA" "BQN" ...
## $ air_time : num 227 227 160 183 116 150 158 53 140 138 ...
## $ distance : num 1400 1416 1089 1576 762 ...
## $ hour : num 5 5 5 5 6 5 6 6 6 6 ...
## $ minute : num 15 29 40 45 0 58 0 0 0 0 ...
## $ time_hour : POSIXct, format: "2013-01-01 05:00:00" "2013-01-01 05:00:00" ...
```

```
unique(flights_data$carrier)
```

```
## [1] "UA" "AA" "B6" "DL" "EV" "MQ" "US" "WN" "VX" "FL" "AS" "9E" "F9" "HA"
## [15] "YV" "O0"
```

```
length(unique(flights_data$carrier))
```

```
## [1] 16
```

```
unique(flights_data$origin)
```

```
## [1] "EWR" "LGA" "JFK"
```

```
length(unique(flights_data$origin))
```

```
## [1] 3
```

```
unique(flights_data$dest)
```

```
## [1] "IAH" "MIA" "BQN" "ATL" "ORD" "FLL" "IAD" "MCO" "PBI" "TPA" "LAX"
## [12] "SFO" "DFW" "BOS" "LAS" "MSP" "DTW" "RSW" "SJU" "PHX" "BWI" "CLT"
## [23] "BUF" "DEN" "SNA" "MSY" "SLC" "XNA" "MKE" "SEA" "ROC" "SYR" "SRQ"
## [34] "RDU" "CMH" "JAX" "CHS" "MEM" "PIT" "SAN" "DCA" "CLE" "STL" "MYR"
## [45] "JAC" "MDW" "HNL" "BNA" "AUS" "BTV" "PHL" "STT" "EGE" "AVL" "PWM"
## [56] "IND" "SAV" "CAK" "HOU" "LGB" "DAY" "ALB" "BDL" "MHT" "MSN" "GSO"
## [67] "CVG" "BUR" "RIC" "GSP" "GRR" "MCI" "ORF" "SAT" "SDF" "PDX" "SJC"
## [78] "OMA" "CRW" "OAK" "SMF" "TUL" "TYS" "OKC" "PVD" "DSM" "PSE" "BHM"
## [89] "CAE" "HDN" "BZN" "MTJ" "EYW" "PSP" "ACK" "BGR" "ABQ" "ILM" "MVY"
## [100] "SBN" "LEX" "CHO" "TVC" "ANC" "LGA"
```

```
length(unique(flights_data$dest))
```

```
## [1] 105
```

```
# Number of departures getting cancelled
sum(is.na(flights_data$dep_time))
```

```
## [1] 8255
```

- After basic inspection of the dataset we can understand that flights dataset has 19 different variables with 336776 rows. Inspecting the head of the dataset, we understand that there are flights arriving and departing on the same day or either just arriving or departing on a given day. As from the tail of the dataset we can see that data isn't sorted, thus I have sorted the dataset based on the month and the day of the year. From further inspection we can find that there are 16 different carriers flying out of NYC airports. NYC has 3 different airports. There are 105 different destination locations to which flights fly out of NYC airports. 8255 flights departures were cancelled as the data has NA.

Formulating Questions:

1. Is there any particular trend of delays at all the airports or is it randomized?

- I think this question is interesting as it will help us understand if there are any particular pattern about the delays. By knowing the delay pattern we can try to address the systemic causes for such delays. If there is no pattern we can at least identify about some anomaly that at once caused a delay. We can also gauge the performance of airports across the 12 months.
- I plan to answer this question firstly by filtering out relevant data. Using this filtered data, I will group by month for count of delayed flights, which will help us know about any particular trends. This graphic will also help us understand the absolute numbers with regards to delays. I also plan to understand percent of delays per month for 2013 for each of the airports. From this visual we can make direct comparisons between performances of airports across months.
- So I will filter out data according to departures from particular airports. Further, remove data about the cancelled flights and also the flights that didn't have any delay. Thus the data will be split into three different sets (3 airports) with only details about flights which were delayed. Further by grouping on monthly basis and taking the count of delays we can plot for all three airports and see the trends. I also plan to plot the percentage of flight delays across months, this will give a clearer picture.

```
#Finding total count of flights flying out of all three airports on monthly basis
#This is required to find percentage of delayed flights
byMon_EWR_total <- group_by(flights_data[flights_data$origin == "EWR",],month)
( sumMon_EWR_total <- summarize(byMon_EWR_total,count=n()) )
```

```
## # A tibble: 12 × 2
##   month count
##   <int> <int>
## 1     1  9893
## 2     2  9107
## 3     3 10420
## 4     4 10531
## 5     5 10592
## 6     6 10175
## 7     7 10475
## 8     8 10359
## 9     9  9550
## 10    10 10104
## 11    11  9707
## 12    12  9922
```

```
colnames(sumMon_EWR_total) <- c("month", "TotalCount")
```

```
byMon_LGA_total <- group_by(flights_data[flights_data$origin == "LGA",],month)
( sumMon_LGA_total <- summarize(byMon_LGA_total,count=n()) )
```

```
## # A tibble: 12 × 2
##   month count
##   <int> <int>
## 1     1  7950
## 2     2  7423
## 3     3  8717
## 4     4  8581
## 5     5  8807
## 6     6  8596
## 7     7  8927
## 8     8  8985
## 9     9  9116
## 10    10  9642
## 11    11  8851
## 12    12  9067
```

```
colnames(sumMon_LGA_total) <- c("month", "TotalCount")
```

```
byMon_JFK_total <- group_by(flights_data[flights_data$origin == "JFK",],month)
( sumMon_JFK_total <- summarize(byMon_JFK_total,count=n()) )
```

```
## # A tibble: 12 × 2
##   month count
##   <int> <int>
## 1     1  9161
## 2     2  8421
## 3     3  9697
## 4     4  9218
## 5     5  9397
## 6     6  9472
## 7     7 10023
## 8     8  9983
```

```
## 9      9  8908
## 10     10  9143
## 11     11  8710
## 12     12  9146
```

```
colnames(sumMon_JFK_total) <- c("month", "TotalCount")
```

```
#Filtering data to capture only specific airport and delayed flight details
#Cancelled flights and on time departure flights have been omitted
EWR_data = filter(flights_data, flights_data$origin == "EWR" & flights_data$dep_delay>0)
LGA_data = filter(flights_data, flights_data$origin == "LGA" & flights_data$dep_delay>0)
JFK_data = filter(flights_data, flights_data$origin == "JFK" & flights_data$dep_delay>0)
```

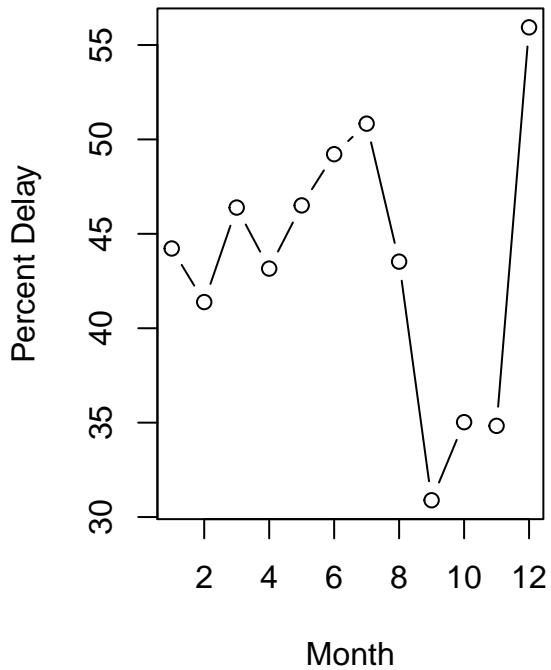
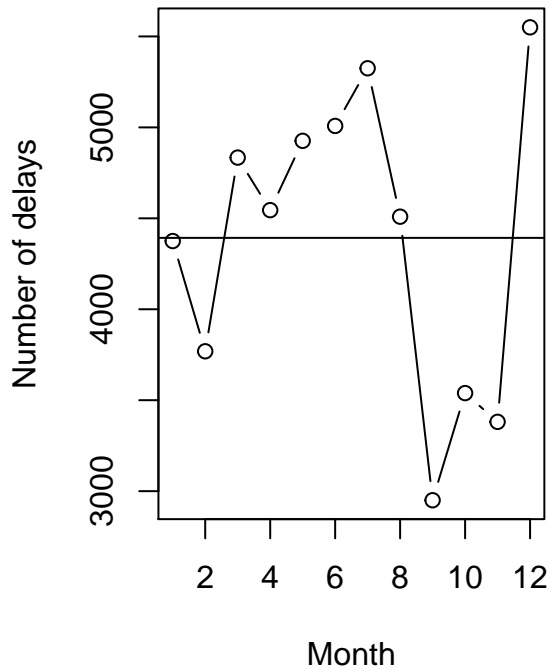
```
#Grouping by delay of flights on monthly basis for EWR Airport
#Plotting count of delayed flights and percentage delayed flights per month for EWR airport
byMon_EWR <- group_by(EWR_data,month)
( sumMon_EWR <- summarize(byMon_EWR,count=n()) )
```

```
## # A tibble: 12 × 2
##   month count
##   <int> <int>
## 1     1  4375
## 2     2  3769
## 3     3  4834
## 4     4  4545
## 5     5  4926
## 6     6  5008
## 7     7  5325
## 8     8  4509
## 9     9  2950
## 10    10  3539
## 11    11  3381
## 12    12  5550
```

```
par(mfrow=c(1,2))
plot(sumMon_EWR, type='b', ylab = 'Number of delays', xlab = 'Month')
abline(h=mean(sumMon_EWR$count))
```

```
sumMon_EWR_final = merge(x = sumMon_EWR, y = sumMon_EWR_total, by = "month", all = TRUE)
sumMon_EWR_final$percent_delay <- with(sumMon_EWR_final, (count/TotalCount)*100)
plot(x=sumMon_EWR_final$month,y=sumMon_EWR_final$percent_delay,
     ylab='Percent Delay', xlab='Month', type = 'b')
mtext('Monthly trend of delays at EWR Airport', side = 1, line = -21, outer = TRUE)
```


Monthly trend of delays at EWR Airport



```
#Grouping by delay of flights on monthly basis for LGA Airport
#Plotting count of delayed flights and percentage delayed flights per month for LGA airport
par(mfrow=c(1,2))
byMon_LGA <- group_by(LGA_data,month)
( sumMon_LGA <- summarize(byMon_LGA,count=n()) )
```

```
## # A tibble: 12 × 2
##   month count
##   <int> <int>
## 1     1  2193
## 2     2  2225
## 3     3  2842
## 4     4  2692
## 5     5  2812
## 6     6  3318
## 7     7  3539
## 8     8  3075
## 9     9  2216
## 10    10  2611
## 11    11  2471
## 12    12  3696
```

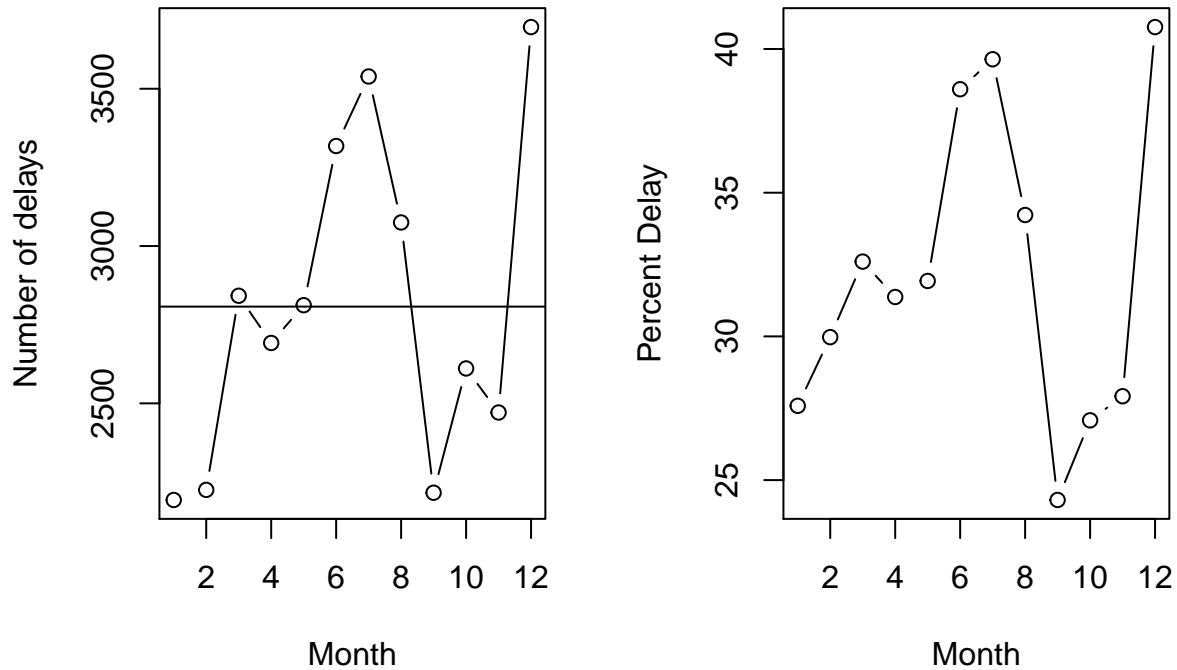
```
plot(sumMon_LGA, type='b', ylab = 'Number of delays', xlab = 'Month')
abline(h=mean(sumMon_LGA$count))
```

```

sumMon_LGA_final = merge(x = sumMon_LGA, y = sumMon_LGA_total, by = "month", all = TRUE)
sumMon_LGA_final$percent_delay <- with(sumMon_LGA_final, (count/TotalCount)*100)
plot(x=sumMon_LGA_final$month,y=sumMon_LGA_final$percent_delay,
     ylab='Percent Delay', xlab = 'Month', type = 'b')
mtext('Monthly trend of delays at LGA Airport', side = 1, line = -21, outer = TRUE)

```

Monthly trend of delays at LGA Airport



```

#Grouping by delay of flights on monthly basis for JFK Airport
#Plotting count of delayed flights and percentage delayed flights per month for JFK airport
par(mfrow=c(1,2))
byMon_JFK <- group_by(JFK_data,month)
( sumMon_JFK <- summarize(byMon_JFK,count=n()) )

```

```

## # A tibble: 12 × 2
##   month count
##   <int> <int>
## 1     1  3094
## 2     2  3130
## 3     3  3533
## 4     4  3306
## 5     5  3553
## 6     6  4329
## 7     7  5045
## 8     8  4129
## 9     9  2649
## 10    10  2572

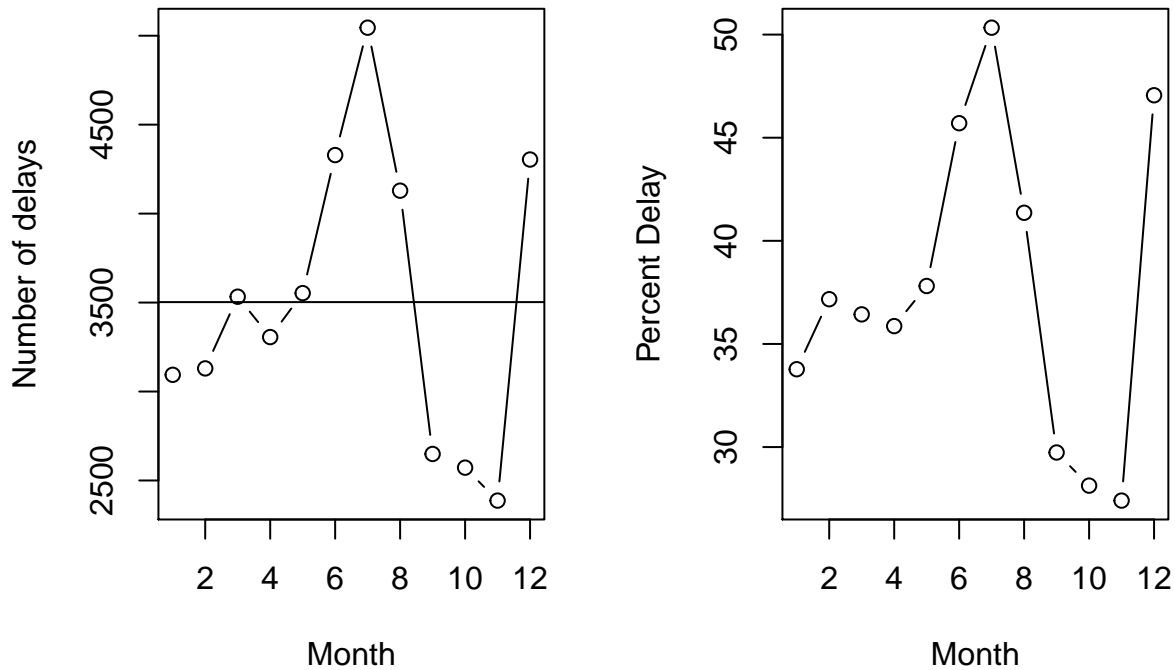
```

```
## 11    11  2387
## 12    12  4304
```

```
plot(sumMon_JFK, type='b', ylab = 'Number of delays', xlab = 'Month')
abline(h=mean(sumMon_JFK$count))

sumMon_JFK_final = merge(x = sumMon_JFK, y = sumMon_JFK_total, by = "month", all = TRUE)
sumMon_JFK_final$percent_delay <- with(sumMon_JFK_final, (count/TotalCount)*100)
plot(x=sumMon_JFK_final$month,y=sumMon_JFK_final$percent_delay,
     ylab='Percent Delay', xlab='Month', type = 'b')
mtext('Monthly trend of delays at JFK Airport', side = 1, line = -21, outer = TRUE)
```

Monthly trend of delays at JFK Airport



- So by looking at the above visualizations, we can conclude that number of delays are highest in the month of December and lowest in the months of Sept, Oct and November for all three airports. Thus there is a trend which tells us that during the holiday season the delays are higher and they are lower just before that holiday period.
- We can also observe that Airport LGA and JFK perform better than Airport EWR in terms of the average number of delays per month. This observation can be reasoned out as EWR flies out more flights than LGA or JFK.
- Another specific thing to note from the visuals is even though there is a dip in the number of delays we observe increase in percentage of delay in flights and vice versa for some months. Example of this can be seen in “Monthly trend of delays at LGA Airport”, for month 10 to 11, the number of delays decrease though there is an increase in percentage of delays during the same period as observed in the right visual. Couple of similar instances are observed.

- Now to further my analysis and understand the reason for the observed trend I came up with the Number of Carrier flying from specific Airports visualization which is as below. In this, we see that in general the number of carrier flying from each of the airports do not change much, hence the increase in delays in December or decrease in other months cannot be answered. We would need to do further data exploration analysis to resolve the reason behind the trend we observe.

```
#Plotting number of different carriers flying out of each Airport
par(mfrow=c(1,3))
( CarrierMon_EWR <- summarize(byMon_EWR,CarrierCount = length(unique(carrier)) ) )
```

```
## # A tibble: 12 × 2
##   month CarrierCount
##   <int>         <int>
## 1     1             10
## 2     2             10
## 3     3             10
## 4     4             11
## 5     5             11
## 6     6             12
## 7     7             11
## 8     8             11
## 9     9             11
## 10    10             11
## 11    11             12
## 12    12             11
```

```
barplot(CarrierMon_EWR$CarrierCount, main="Number of Carrier flying from EWR",
        xlab="Month", ylab="Carrier Count", names.arg=CarrierMon_EWR$month,
border="black", density=CarrierMon_EWR$CarrierCount,cex.names = 0.5)
```

```
( CarrierMon_LGA <- summarize(byMon_LGA,CarrierCount = length(unique(carrier)) ) )
```

```
## # A tibble: 12 × 2
##   month CarrierCount
##   <int>         <int>
## 1     1             13
## 2     2             12
## 3     3             12
## 4     4             12
## 5     5             12
## 6     6             12
## 7     7             12
## 8     8             13
## 9     9             13
## 10    10             12
## 11    11             13
## 12    12             12
```

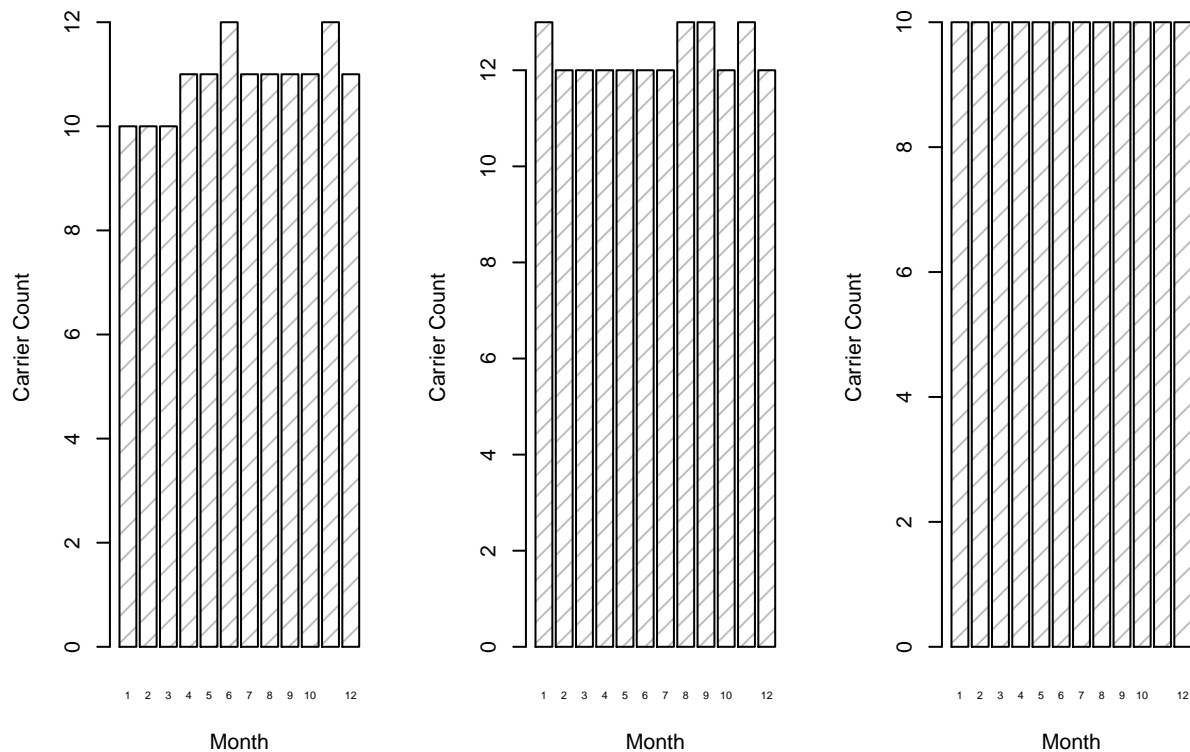
```
barplot(CarrierMon_LGA$CarrierCount, main="Number of Carrier flying from LGA",
        xlab="Month", ylab="Carrier Count", names.arg=CarrierMon_LGA$month,
border="black", density=CarrierMon_LGA$CarrierCount,cex.names = 0.5)
```

```
( CarrierMon_JFK <- summarize(byMon_JFK,CarrierCount = length(unique(carrier)) ) )
```

```
## # A tibble: 12 × 2
##   month CarrierCount
##   <int>     <int>
## 1     1         10
## 2     2         10
## 3     3         10
## 4     4         10
## 5     5         10
## 6     6         10
## 7     7         10
## 8     8         10
## 9     9         10
## 10    10         10
## 11    11         10
## 12    12         10
```

```
barplot(CarrierMon_JFK$CarrierCount, main="Number of Carrier flying from JFK",
        xlab="Month", ylab="Carrier Count", names.arg=CarrierMon_JFK$month,
        border="black", density=CarrierMon_JFK$CarrierCount,cex.names = 0.5)
```

Number of Carrier flying from EI Number of Carrier flying from L Number of Carrier flying from JI



2. Which carriers have been the top and the bottom performers in 2013?

- I think this question will help us identify the carriers which have been performing badly through out the year. By knowing this we can help the general public to avoid commuting by this carrier.

- I feel that to answer this question we would have to look at the number of flights departing delayed and also arriving delayed. I plan to ignore the carriers which departed delayed though arrived on or before time as in all the time was covered by the carrier during flight. Although there is a ethical promise that a carrier makes to start on scheduled time, I plan to ignore this concern in my below analysis.

```
Carrier_data = filter(flights_data, flights_data$dep_delay>0 & flights_data$arr_delay>0)
bycarrier <- group_by(Carrier_data, carrier)
( sumCarrier <- summarize(bycarrier, count=n()) )
```

```
## # A tibble: 16 × 2
##   carrier count
##   <chr> <int>
## 1     9E  5055
## 2     AA  6668
## 3     AS   125
## 4     B6 16436
## 5     DL 10126
## 6     EV 19183
## 7     F9   256
## 8     FL  1386
## 9     HA    37
## 10    MQ  6944
## 11    OO    8
## 12    UA 16606
## 13    US  3765
## 14    VX  1217
## 15    WN  4293
## 16    YV   198
```

```
Total_data = filter(flights_data, !is.na(flights_data$dep_time))
bycarrier_total <- group_by(Total_data, carrier)
( sumCarrier_total <- summarize(bycarrier_total, count=n()) )
```

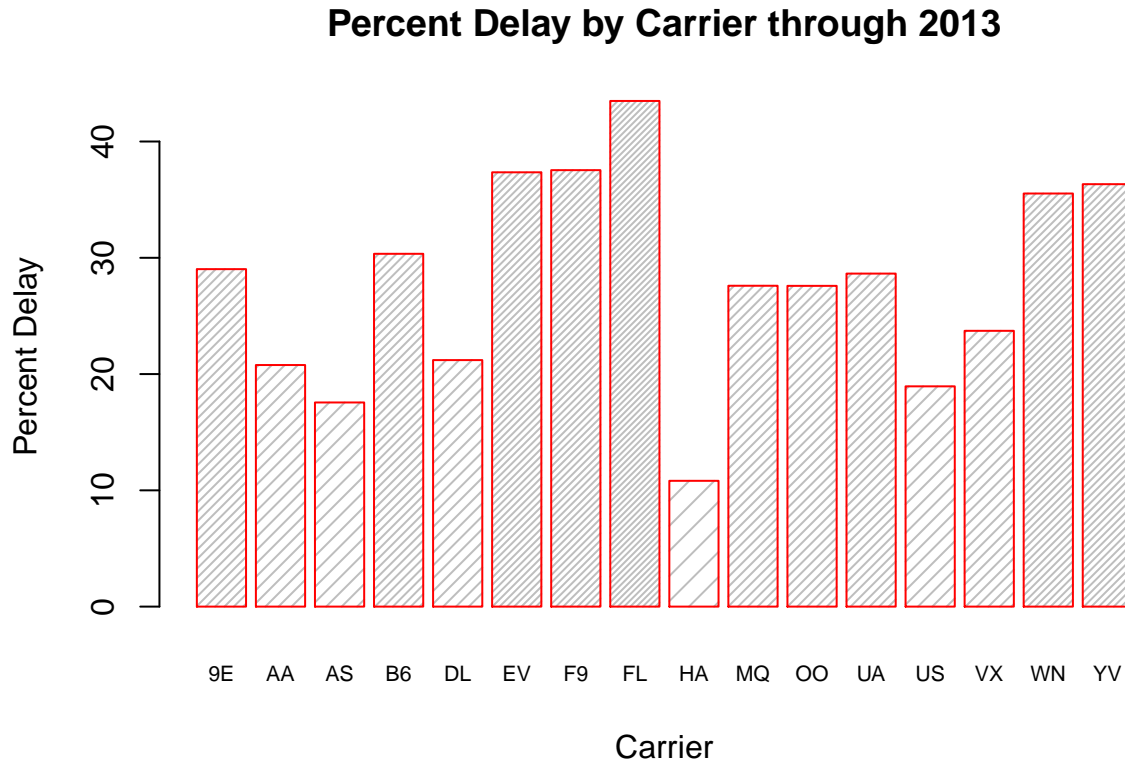
```
## # A tibble: 16 × 2
##   carrier count
##   <chr> <int>
## 1     9E 17416
## 2     AA 32093
## 3     AS   712
## 4     B6 54169
## 5     DL 47761
## 6     EV 51356
## 7     F9   682
## 8     FL  3187
## 9     HA   342
## 10    MQ 25163
## 11    OO   29
## 12    UA 57979
## 13    US 19873
## 14    VX  5131
## 15    WN 12083
## 16    YV   545
```

```

colnames(sumCarrier_total) <- c("carrier", "TotalCount")

sumCarrier_final = merge(x = sumCarrier, y = sumCarrier_total, by = "carrier", all = TRUE)
sumCarrier_final$percent_delay <- with(sumCarrier_final, (count/TotalCount)*100)
par(mfrow=c(1,1))
barplot(sumCarrier_final$percent_delay, main="Percent Delay by Carrier through 2013",
        xlab="Carrier", ylab="Percent Delay", names.arg=sumCarrier_final$carrier,
        border="red", density=sumCarrier_final$percent_delay,cex.names = 0.7)

```



```

(meanCarrier <- summarize(bycarrier,mean=mean(arr_delay)))

```

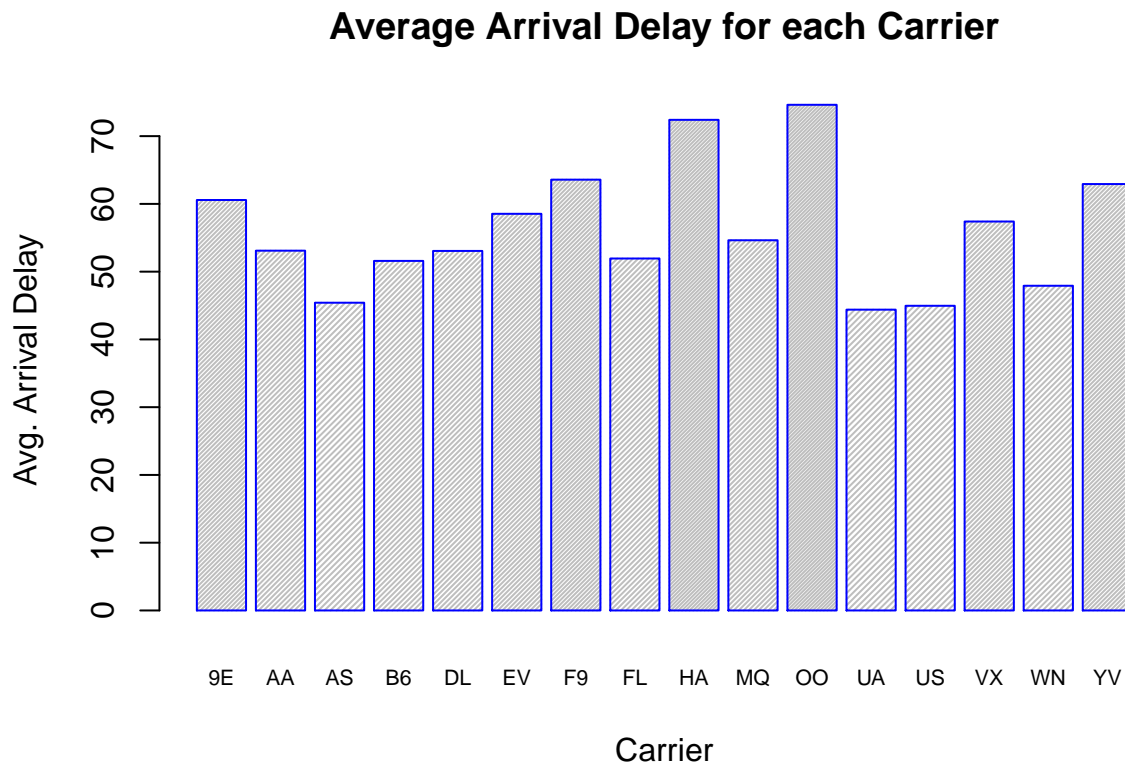
```

## # A tibble: 16 × 2
##   carrier      mean
##   <chr>      <dbl>
## 1     9E 60.57626
## 2     AA 53.09988
## 3     AS 45.41600
## 4     B6 51.58761
## 5     DL 53.05777
## 6     EV 58.53662
## 7     F9 63.57812
## 8     FL 51.94300
## 9     HA 72.40541
## 10    MQ 54.63292
## 11    OO 74.62500

```

```
## 12    UA 44.39474
## 13    US 44.96441
## 14    VX 57.39852
## 15    WN 47.92057
## 16    YV 62.92929
```

```
barplot(meanCarrier$mean, main="Average Arrival Delay for each Carrier",
        xlab="Carrier", ylab="Avg. Arrival Delay", names.arg=meanCarrier$carrier,
        border="blue", density=meanCarrier$mean,cex.names = 0.7)
```



- The performance of the carrier can be gauged by (1) what percentage of flights of a particular carrier are delayed in departure and also delayed in arrival and (2) what is the average delay in arrival time for each of the carrier over the year of 2013.

- Firstly, looking at the visualization (Percent Delay by Carrier through 2013), we observe that carrier FL has the highest delay %, thus making it the least performer among other carriers. Carrier HA has the best performance in terms of delay %.
- Secondly, looking at the visualization (Average Arrival Delay for each Carrier), we observe that OO and HA have higher arrival delays among other carriers. UA and US carriers perform best when looking from this perspective. I have considered average arrival delay because I feel that in all for a traveller the delay in reaching a particular point is more significant than delay in departure.

Data Wrangling

- How many flights were there from NYC airports to Seattle in 2013?


```
#Finding airport code for Seattle
Sea_airport_filter = filter(airports,grepl("Seattle",airports$name))
Sea_airport_code = Sea_airport_filter$faa
#Filtering flights for Seattle from the flights dataset
filter(flights_data,flights_data$dest == Sea_airport_code)
```

```
## # A tibble: 3,923 × 19
##   year month   day dep_time sched_dep_time dep_delay arr_time
##   <int> <int> <int>   <int>         <int>         <dbl>   <int>
## 1  2013     1     1     724             725           -1    1020
## 2  2013     1     1     743             730            13    1059
## 3  2013     1     1     857             851             6    1157
## 4  2013     1     1    1418            1419            -1    1726
## 5  2013     1     1    1421            1355            26    1735
## 6  2013     1     1    1730            1729             1    2039
## 7  2013     1     1    1808            1815             -7    2111
## 8  2013     1     1    1824            1830             -6    2203
## 9  2013     1     1    1826            1830             -4    2154
## 10 2013     1     1    1952            1930             22    2257
## # ... with 3,913 more rows, and 12 more variables: sched_arr_time <int>,
## #   arr_delay <dbl>, carrier <chr>, flight <int>, tailnum <chr>,
## #   origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>, hour <dbl>,
## #   minute <dbl>, time_hour <dtm>
```

To find the number of flights from NYC to Seattle, firstly, I have found the airport code of Seattle airport using airports dataset. Using this I have filtered the flights dataset based on the flights flying to SEattle as the destination. Thus the total number of flights from NYC to Seattle in 2013 is 3923.

- How many airlines fly from NYC to Seattle?

```
Sea_bound = filter(flights_data,flights_data$dest == Sea_airport_code)
#Calculating number of unique carriers to Seattle
length(unique(Sea_bound$carrier))
```

```
## [1] 5
```

Here, I have again filtered the flights dataset to find all flights flying to Seattle and then I found the unique carriers. Thus there are 5 carriers who fly from NYC to Seattle.

- How many unique air planes fly from NYC to Seattle?

```
#Calculating unique number of air planes to Seattle by using tailnum
length(unique(Sea_bound$tailnum))
```

```
## [1] 936
```

To find the unique airplanes, the distinguishing factor is the tailnum. Thus using tailnum as the distinguishing factor, there are 936 unique airplanes that fly between NYC to Seattle.

```
#Calculating unique number of air planes to Seattle by using flight
length(unique(Sea_bound$flight))
```

```
## [1] 166
```

Another logic to find the number of air planes could knowing how many unique flight numbers are arriving at Seattle from NYC. By that logic the count would be 166.

- What is the average arrival delay for flights from NYC to Seattle?

```
#Only considering flights that were delayed
Sea_bound_filter = filter(Sea_bound,Sea_bound$arr_delay>0)
summarize(Sea_bound_filter,mean = mean(arr_delay))
```

```
## # A tibble: 1 × 1
##   mean
##   <dbl>
## 1 39.79984
```

I have filtered out only the flights that had arrival delay at the Seattle airport. Thus to find the average arrival delay, I am not considering the flights that were on time or reached before time. Thus the arrival delay for flights from NYC to Seattle is 39.79984 minutes.

```
#Considering all flights
summarize(Sea_bound,mean = mean(arr_delay,na.rm = TRUE))
```

```
## # A tibble: 1 × 1
##   mean
##   <dbl>
## 1 -1.099099
```

If we take all the flights landing at SEattle from NYC, then the average arrival delay decreases to -1.09 minutes.

- What proportion of flights to Seattle come from each NYC airport?

```
#Grouping by origin
by_origin = group_by(Sea_bound,origin)
#Calculating proportions
summarize(by_origin,count = n(),prop=n()/nrow(Sea_bound))
```

```
## # A tibble: 2 × 3
##   origin count    prop
##   <chr> <int>    <dbl>
## 1 EWR 1831 0.4667346
## 2 JFK 2092 0.5332654
```

Firsly, I have grouped the flights by their origin, thus EWR and JFK are the only two origins for flights from NYC to Seattle. Then to find the proportion of flights from each of the airport, using the number of flights from each airport, I have divided each by the total number of flights from NYC to Seattle. Thus there are 46.67% flights to Seattle are from EWR airport and 53.33% flights are from JFK airport flying out of NYC.

Study Flight Delays with Weather data

```
#Filtering only delayed flights from all airports
flights_delayed <- filter(flights_data,dep_delay>0)
flights_not_delayed <- filter(flights_data,dep_delay<=0)
```

Above, I have loaded the flights and weather dataset and also filtered data according to delayed or not delayed as that will help in comparison when combined with the weather dataset. By filtering out the delayed flights, I plan to study average time delay and number of delays per some of the variables (visib, wind_speed, wind_gust) in the weather dataset. If we consider the whole dataset, without filtering, then due to averaging out we could miss out on some of the specific flights that were actually delayed because there are flights which have departed early. Thus to avoid such a miss, I have considered only delayed flights for analysis. Also some of the plane models might not be affected by weather and hence might takeoff on or before time, to remove those biases, I consider only delayed flights.

```
#Grouping by origin and time hour, as analysis would be at the granularity
#of the weather dataset
by_time_hour_airport = group_by(flights_delayed,origin,time_hour)

#Calculating the average time delay per airport per time_hour and
#also calculating the number of flights per airport per time_hour
sum_delay_count = summarize(by_time_hour_airport,totaldelay = mean(dep_delay),
                             count = n())
```

As the granularity of analysis would of the weather dataset, I have grouped the flight_delayed dataset by origin and time_hour bringing it to similar granularity. By grouping, I have calculated the average delay time at a particular time_hour and airport and also calculated total count of delays at a particular time_hour and airport.

```
#Joining the above output with the weather dataset
#This is an inner join and the time_hour for which data is not present in weather
#dataset are omitted.
#The time_hour in sum_delay_count df is in GMT and time_hour in weather_data df
#is in PST or PDT
#While joining these columns, they take care of the timezone, thus we dont
#have to change anything
combine_df = merge(sum_delay_count, weather_data,by=c("origin","time_hour"))
```

Above, I have merged the weather and the grouped dataset so that it will help in analysis. The merging is done on origin and time_hour columns.

```
#Working on the combined df, grouping by visibility to see trends between
#delays and the weather variables
by_visib = group_by(combine_df,visib)

#Calculating average delay in time per visib
avg_delay_v = summarize(by_visib,avg_dep_delay_time = mean(totaldelay))

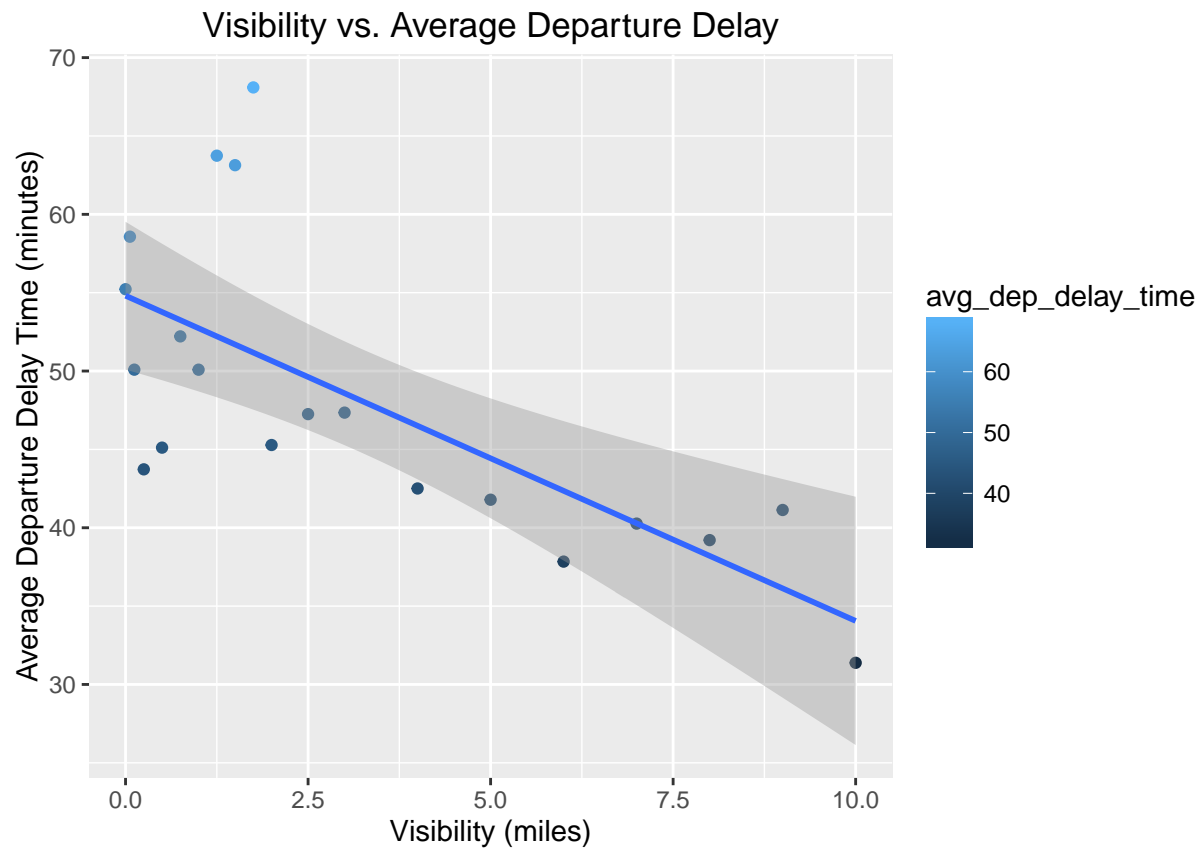
#Calculating average dep_delay count per visib
number_of_delay_per_visib = summarize(by_visib,Avg_Delay_Count_Per_Visib = mean(count))
```

To analyse the flight delays based on visibility variable, I have grouped by visibility and calculated the average departure delay time per visibility and also calculated average flights delayed per visibility. I plot these two metrics as below:

```

#Plotting scatter plots for Visibility vs. Average Departure Delay and
#Visibility vs. Average Number of Delays
#along with the regression line, which basically shows the trend.
p1 <- ggplot(avg_delay_v, aes(x = visib, y = avg_dep_delay_time, color=avg_dep_delay_time))
p1 + geom_point() + geom_smooth(method = "lm") +labs(x = "Visibility (miles)",
                                                    y="Average Departure Delay Time (minutes)",
                                                    title = "Visibility vs. Average Departure Delay")

```

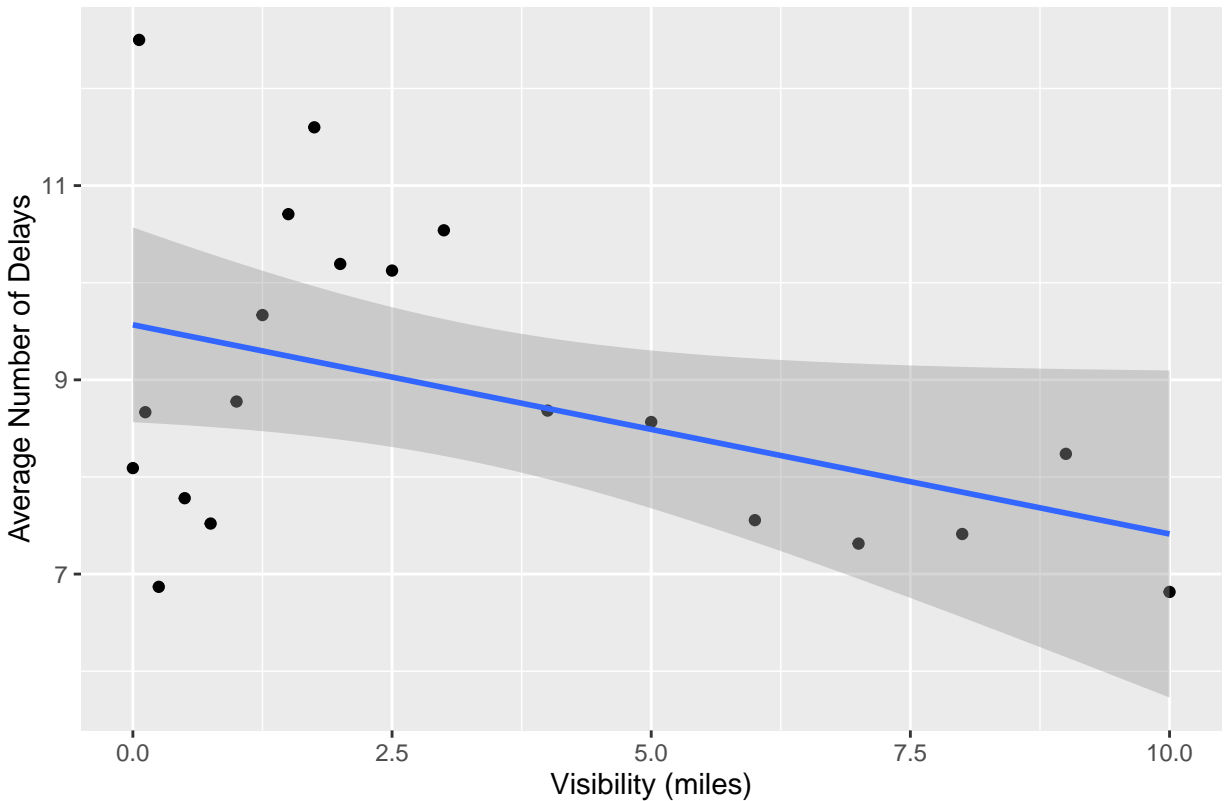


```

p2 <- ggplot(number_of_delay_per_visib, aes(x = visib, y = Avg_Delay_Count_Per_Visib))
p2 + geom_point()+ geom_smooth(method = "lm") +labs(x = "Visibility (miles)",
                                                    y="Average Number of Delays",
                                                    title = "Visibility vs. Average Number of Delays")

```

Visibility vs. Average Number of Delays

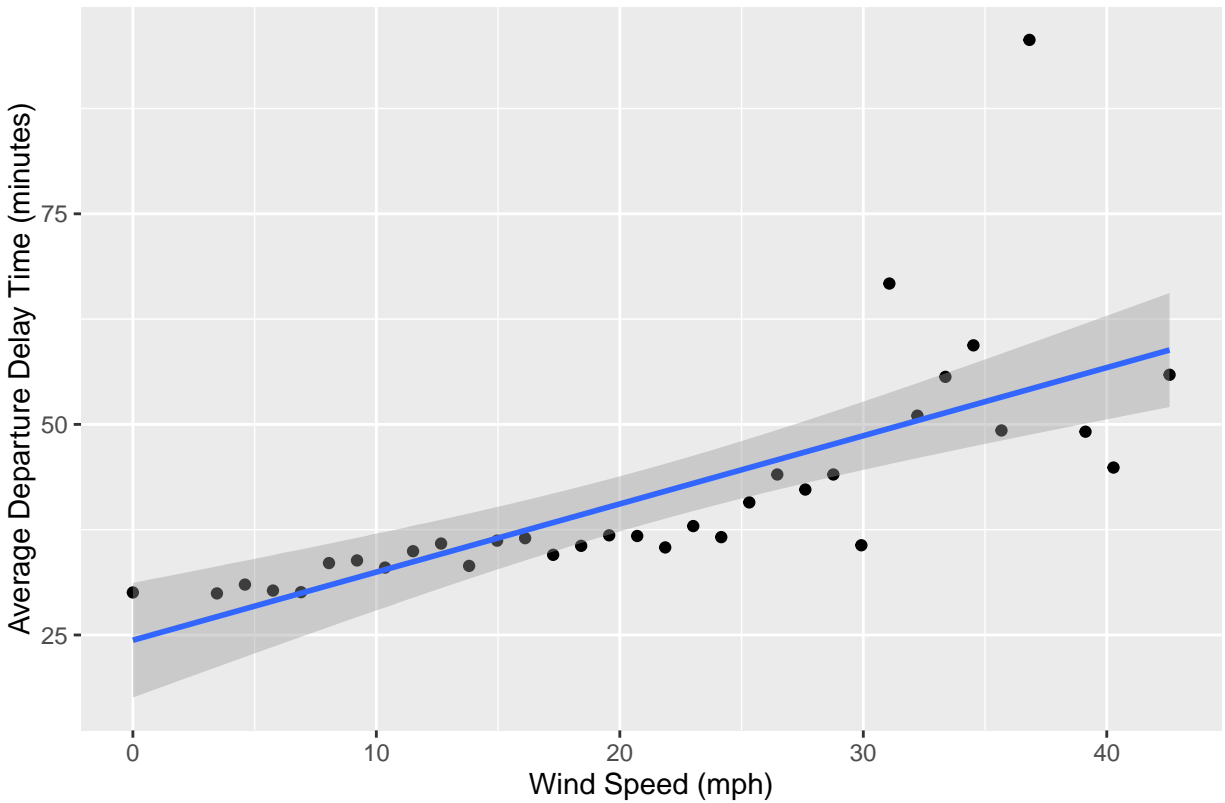


From the above graphics we can see that lower the visibility higher are the average departure delay time and average count of number of delays. This proves that one of the weather variable like the visibility adversely impacts the flights from NYC. To explore more below we can look at the impact of wind_speed on the flight delays

```
#Removing incorrect data - The data in row 677 seems incorrect because
#the value of wind_speed = 1048.36
#and wind_gust = 1206.43, which is a variation of almost 100 times when compared
#to other values in the data set
#Also omitting NA's from wind_speed column
combine_df_new = combine_df[-677,]
combine_df_new = combine_df_new[!is.na(combine_df_new$wind_speed),]
#Working on the combined df, grouping by wind_speed to see trends between
#delays and the weather variables
by_wind_speed = group_by(combine_df_new,wind_speed)
avg_delay_ws = summarize(by_wind_speed,avg_dep_delay_time = mean(totaldelay))

p3 <- ggplot(avg_delay_ws, aes(x = wind_speed, y = avg_dep_delay_time))
p3 + geom_point() + geom_smooth(method = "lm") +labs(x = "Wind Speed (mph)",
                                                    y="Average Departure Delay Time (minutes)",
                                                    title = "Average Departure Delay vs. Wind Speed")
```

Average Departure Delay vs. Wind Speed



Here, I have removed one row from the dataset as the information in the tuple seems to be incorrect. Values of wind_speed and wind_gust are 100 times greater than the other values in the same column. I have also omitted NA's from wind_speed column. I have grouped on wind_speed and calculated the average departure delay in time per value of wind speed (as there are only specific values of wind speed observed in the dataset, it actually is very much a continuous variable). The above graphic depicts that as the wind speed increases the average departure delay time increases. Thus wind_speed also impacts flights from NYC.