Exploratory data analysis on

Airline data

##clear Environment  
rm(list=ls())  
  
#Packages to install  
  
library(lubridate)

library(ggplot2)  
library(dplyr)

library(MUCflights)

library(tidyr)

1. Importing and Exploring data

fdata=read.csv("flight\_data.csv",header=TRUE,stringsAsFactors = FALSE)  
  
#dimention of data   
dim(fdata)

## [1] 336776 19

#Summary of dataframe  
summary(fdata)

## 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 Length:336776   
## 1st Qu.: 9.00 1st Qu.: 8.00 Class :character   
## Median :13.00 Median :29.00 Mode :character   
## Mean :13.18 Mean :26.23   
## 3rd Qu.:17.00 3rd Qu.:44.00   
## Max. :23.00 Max. :59.00   
##

#Details of each column and type of it   
str(fdata)

## 'data.frame': 336776 obs. of 19 variables:  
## $ year : int 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 : int 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 : int 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 : int 227 227 160 183 116 150 158 53 140 138 ...  
## $ distance : int 1400 1416 1089 1576 762 719 1065 229 944 733 ...  
## $ hour : int 5 5 5 5 6 5 6 6 6 6 ...  
## $ minute : int 15 29 40 45 0 58 0 0 0 0 ...  
## $ time\_hour : chr "01-01-2013 05:00" "01-01-2013 05:00" "01-01-2013 05:00" "01-01-2013 05:00" ...

#Number of na's in each column   
data.frame(colSums(is.na(fdata)))

## colSums.is.na.fdata..  
## year 0  
## month 0  
## day 0  
## dep\_time 8255  
## sched\_dep\_time 0  
## dep\_delay 8255  
## arr\_time 8713  
## sched\_arr\_time 0  
## arr\_delay 9430  
## carrier 0  
## flight 0  
## tailnum 2512  
## origin 0  
## dest 0  
## air\_time 9430  
## distance 0  
## hour 0  
## minute 0  
## time\_hour 0

#There are missing values in dep\_time,dep\_delay,arr\_time,arr\_delay,tailnum and air\_time

#Number of origin i.e 3 origins  
levels(factor(fdata[,'origin']))

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

#Number of Destination i.e 105 destination  
levels(factor(fdata[,'dest']))

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

#data.frame(table(fdata[,'flight']))

####################2.Data Cleaning

# filling Missing values  
head(fdata)

## year month day dep\_time sched\_dep\_time dep\_delay arr\_time sched\_arr\_time  
## 1 2013 1 1 517 515 2 830 819  
## 2 2013 1 1 533 529 4 850 830  
## 3 2013 1 1 542 540 2 923 850  
## 4 2013 1 1 544 545 -1 1004 1022  
## 5 2013 1 1 554 600 -6 812 837  
## 6 2013 1 1 554 558 -4 740 728  
## arr\_delay carrier flight tailnum origin dest air\_time distance hour  
## 1 11 UA 1545 N14228 EWR IAH 227 1400 5  
## 2 20 UA 1714 N24211 LGA IAH 227 1416 5  
## 3 33 AA 1141 N619AA JFK MIA 160 1089 5  
## 4 -18 B6 725 N804JB JFK BQN 183 1576 5  
## 5 -25 DL 461 N668DN LGA ATL 116 762 6  
## 6 12 UA 1696 N39463 EWR ORD 150 719 5  
## minute time\_hour  
## 1 15 01-01-2013 05:00  
## 2 29 01-01-2013 05:00  
## 3 40 01-01-2013 05:00  
## 4 45 01-01-2013 05:00  
## 5 0 01-01-2013 06:00  
## 6 58 01-01-2013 05:00

attach(fdata)

## The following object is masked from package:lubridate:  
##   
## origin

#2.1 filling arr\_delay  
# arr\_delay can be filled by subtracting arr\_time and sched\_arr\_time   
# To calculate arr\_delay need to condider timezone offset from UTC,daylight and name of timezone  
  
#olson is name of timezone is obtained by airportdata for origin and destination  
airport\_data = read.csv("airport\_data.csv",sep=",",stringsAsFactors = FALSE ,col.names = c("AirportID","Name","City","Country","IATA","ICAO","Latitutde","Longtitude","Altitude","Timezone","DST","Olson","Type","Source"))  
  
#Destination  
#Finding Olson name(timezone name) by inner join of fdata and airport\_data  
  
dtz= inner\_join(data.frame(IATA=fdata$dest),airport\_data[,c("IATA","Olson","Latitutde","Longtitude")],by = "IATA")

## Warning: Column `IATA` joining factor and character vector, coercing into  
## character vector

colnames(dtz)=c("IATA","olson","dest\_lati","dest\_long")  
str(dtz)

## 'data.frame': 336776 obs. of 4 variables:  
## $ IATA : chr "IAH" "IAH" "MIA" "BQN" ...  
## $ olson : chr "America/Chicago" "America/Chicago" "America/New\_York" "America/Puerto\_Rico" ...  
## $ dest\_lati: num 30 30 25.8 18.5 33.6 ...  
## $ dest\_long: num -95.3 -95.3 -80.3 -67.1 -84.4 ...

#Origin  
otz= inner\_join(data.frame(IATA=fdata$origin),airport\_data[,c("IATA","Olson","Latitutde","Longtitude")],by = "IATA")

## Warning: Column `IATA` joining factor and character vector, coercing into  
## character vector

colnames(otz)=c("IATA","olson","origin\_lati","origin\_long")  
str(dtz)

## 'data.frame': 336776 obs. of 4 variables:  
## $ IATA : chr "IAH" "IAH" "MIA" "BQN" ...  
## $ olson : chr "America/Chicago" "America/Chicago" "America/New\_York" "America/Puerto\_Rico" ...  
## $ dest\_lati: num 30 30 25.8 18.5 33.6 ...  
## $ dest\_long: num -95.3 -95.3 -80.3 -67.1 -84.4 ...

#Creating two column of dest and origin Olson names   
fdata$olson\_dest = dtz$olson  
fdata$olson\_origin = otz$olson  
str(fdata)

## 'data.frame': 336776 obs. of 21 variables:  
## $ year : int 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 : int 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 : int 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 : int 227 227 160 183 116 150 158 53 140 138 ...  
## $ distance : int 1400 1416 1089 1576 762 719 1065 229 944 733 ...  
## $ hour : int 5 5 5 5 6 5 6 6 6 6 ...  
## $ minute : int 15 29 40 45 0 58 0 0 0 0 ...  
## $ time\_hour : chr "01-01-2013 05:00" "01-01-2013 05:00" "01-01-2013 05:00" "01-01-2013 05:00" ...  
## $ olson\_dest : chr "America/Chicago" "America/Chicago" "America/New\_York" "America/Puerto\_Rico" ...  
## $ olson\_origin : chr "America/New\_York" "America/New\_York" "America/New\_York" "America/New\_York" ...

make\_time = function(y,m,d,h,min,timez){  
 date=paste(y,m,d,sep = "-")  
 time=paste(h,min,00,sep=":")  
 dt=paste(date,time,sep = " ")  
 t=as.POSIXct(dt,format="%Y-%m-%d %H:%M:%S",tz="UTC")  
 force\_tz(t,timez)  
}  
make\_time(2018,1,23,14,38,"Asia/Calcutta")

## [1] "2018-01-23 14:38:00 IST"

#Updating dep\_time,sched\_dep\_time,sched\_arr\_time and arr\_time  
fdata$dep\_time<-make\_time(year,month,day,floor(fdata$dep\_time/100),fdata$dep\_time%%100,fdata$olson\_origin)  
fdata$sched\_dep\_time <- make\_time(year,month,day,floor(fdata$sched\_dep\_time/100),fdata$sched\_dep\_time%%100,fdata$olson\_origin)  
fdata$sched\_arr\_time <- make\_time(year,month,day,floor(fdata$sched\_arr\_time/100),fdata$sched\_arr\_time%%100,fdata$olson\_dest)  
fdata$arr\_time<- make\_time(year,month,day,floor(fdata$arr\_time/100),fdata$arr\_time%%100,fdata$olson\_dest)  
  
#filling nas with value in arr\_delay  
fdata$arr\_delay= as.numeric(difftime(fdata$arr\_time,fdata$sched\_arr\_time,units = "mins"))  
  
#Creating new columns of utc time  
fdata$dep\_timeUTC = with\_tz(fdata$dep\_time,"UTC")   
fdata$sched\_dep\_timeUTC = with\_tz(fdata$sched\_dep\_time,"UTC")  
fdata$arr\_timeUTC = with\_tz(fdata$arr\_time,"UTC")  
fdata$sched\_arr\_timeUTC=with\_tz(fdata$sched\_arr\_time,"UTC")  
  
data.frame(colSums(is.na(fdata)))

## colSums.is.na.fdata..  
## year 0  
## month 0  
## day 0  
## dep\_time 8255  
## sched\_dep\_time 0  
## dep\_delay 8255  
## arr\_time 8715  
## sched\_arr\_time 1  
## arr\_delay 8715  
## carrier 0  
## flight 0  
## tailnum 2512  
## origin 0  
## dest 0  
## air\_time 9430  
## distance 0  
## hour 0  
## minute 0  
## time\_hour 0  
## olson\_dest 0  
## olson\_origin 0  
## dep\_timeUTC 8255  
## sched\_dep\_timeUTC 0  
## arr\_timeUTC 8715  
## sched\_arr\_timeUTC 1

summary(fdata)

## year month day   
## Min. :2013 Min. : 1.000 Min. : 1.00   
## 1st Qu.:2013 1st Qu.: 4.000 1st Qu.: 8.00   
## Median :2013 Median : 7.000 Median :16.00   
## Mean :2013 Mean : 6.549 Mean :15.71   
## 3rd Qu.:2013 3rd Qu.:10.000 3rd Qu.:23.00   
## Max. :2013 Max. :12.000 Max. :31.00   
##   
## dep\_time sched\_dep\_time   
## Min. :2013-01-01 05:17:00 Min. :2013-01-01 05:15:00   
## 1st Qu.:2013-04-05 06:33:00 1st Qu.:2013-04-04 13:29:00   
## Median :2013-07-04 09:24:00 Median :2013-07-03 10:30:00   
## Mean :2013-07-03 17:57:32 Mean :2013-07-03 05:49:08   
## 3rd Qu.:2013-10-01 16:38:00 3rd Qu.:2013-10-01 07:00:00   
## Max. :2013-12-31 23:56:00 Max. :2013-12-31 23:59:00   
## NA's :8255   
## dep\_delay arr\_time   
## Min. : -43.00 Min. :2013-01-01 00:03:00   
## 1st Qu.: -5.00 1st Qu.:2013-04-05 07:14:00   
## Median : -2.00 Median :2013-07-04 11:07:00   
## Mean : 12.64 Mean :2013-07-03 19:33:39   
## 3rd Qu.: 11.00 3rd Qu.:2013-10-01 19:19:00   
## Max. :1301.00 Max. :2014-01-01 00:00:00   
## NA's :8255 NA's :8715   
## sched\_arr\_time arr\_delay carrier   
## Min. :2013-01-01 00:05:00 Min. :-1486.00 Length:336776   
## 1st Qu.:2013-04-04 15:08:30 1st Qu.: -18.00 Class :character   
## Median :2013-07-03 12:22:00 Median : -6.00 Mode :character   
## Mean :2013-07-03 07:46:01 Mean : -18.34   
## 3rd Qu.:2013-10-01 09:22:30 3rd Qu.: 12.00   
## Max. :2013-12-31 23:59:00 Max. : 1496.00   
## NA's :1 NA's :8715   
## flight tailnum origin dest   
## Min. : 1 Length:336776 Length:336776 Length:336776   
## 1st Qu.: 553 Class :character Class :character Class :character   
## Median :1496 Mode :character Mode :character Mode :character   
## Mean :1972   
## 3rd Qu.:3465   
## Max. :8500   
##   
## air\_time distance hour minute   
## Min. : 20.0 Min. : 17 Min. : 1.00 Min. : 0.00   
## 1st Qu.: 82.0 1st Qu.: 502 1st Qu.: 9.00 1st Qu.: 8.00   
## Median :129.0 Median : 872 Median :13.00 Median :29.00   
## Mean :150.7 Mean :1040 Mean :13.18 Mean :26.23   
## 3rd Qu.:192.0 3rd Qu.:1389 3rd Qu.:17.00 3rd Qu.:44.00   
## Max. :695.0 Max. :4983 Max. :23.00 Max. :59.00   
## NA's :9430   
## time\_hour olson\_dest olson\_origin   
## Length:336776 Length:336776 Length:336776   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## dep\_timeUTC sched\_dep\_timeUTC   
## Min. :2013-01-01 10:17:00 Min. :2013-01-01 10:15:00   
## 1st Qu.:2013-04-05 10:33:00 1st Qu.:2013-04-04 17:29:00   
## Median :2013-07-04 13:24:00 Median :2013-07-03 14:30:00   
## Mean :2013-07-03 21:57:32 Mean :2013-07-03 09:49:08   
## 3rd Qu.:2013-10-01 20:38:00 3rd Qu.:2013-10-01 11:00:00   
## Max. :2014-01-01 04:56:00 Max. :2014-01-01 04:59:00   
## NA's :8255   
## arr\_timeUTC sched\_arr\_timeUTC   
## Min. :2013-01-01 06:03:00 Min. :2013-01-01 06:05:00   
## 1st Qu.:2013-04-05 12:14:00 1st Qu.:2013-04-04 20:08:30   
## Median :2013-07-04 16:07:00 Median :2013-07-03 17:22:00   
## Mean :2013-07-04 00:33:39 Mean :2013-07-03 12:46:01   
## 3rd Qu.:2013-10-02 00:19:00 3rd Qu.:2013-10-01 14:22:30   
## Max. :2014-01-01 06:00:00 Max. :2014-01-01 05:59:00   
## NA's :8715 NA's :1

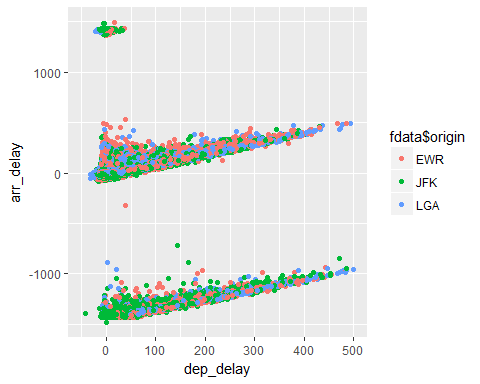
#2.2  
attach(fdata)

table(origin)

## origin  
## EWR JFK LGA   
## 120835 111279 104662

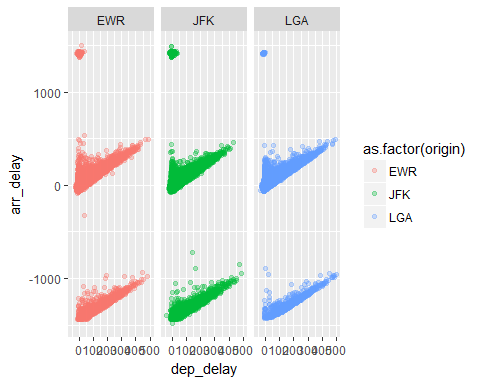
#departure delay vs arrival delay  
p=ggplot(fdata,aes(dep\_delay,arr\_delay))+xlim(-50,500)  
p+geom\_point(aes(color=fdata$origin))

## Warning: Removed 8774 rows containing missing values (geom\_point).



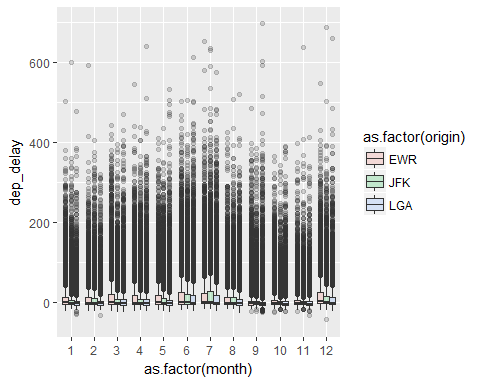
#From observation of plots :-  
#1. In all three origins there is outliers  
#2. Seeing to plot there is positive correlation between dep\_delay and arr\_delay  
#3. Some flights departured exact time but arrived early  
#4. Only one flight departed early i.e -43  
#5. JFK origin flights arrived early even though departed late  
#6. Most of the LGA origin flights reached destination in time i.e arrival delay is less.  
#7. LGA time management is good   
  
p+geom\_point(alpha=0.3,aes(color=as.factor(origin)))+facet\_grid(.~origin)

## Warning: Removed 8774 rows containing missing values (geom\_point).



#Again LGA has good time management   
  
#Departure delay month wise  
ggplot(fdata,aes(as.factor(month),dep\_delay))+  
 geom\_boxplot(alpha=0.2,aes(fill=as.factor(origin)))+ylim(-50,700)

## Warning: Removed 8285 rows containing non-finite values (stat\_boxplot).



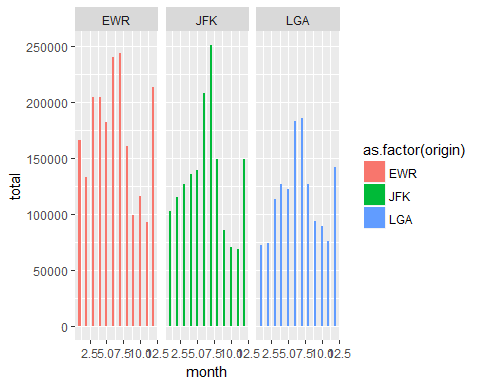
#From plot   
#1. June and july(6 and 7) more number flights departed from all 3 origins  
#2. In month of 3,4,5,6,7 and 12 more number of flights departed from origin EWR  
#3. In month 9,10 and 11 less number of flights departed   
  
  
  
  
#departure delay in each origin   
fdata %>% select(dep\_delay,origin)%>%filter(dep\_delay>0) %>% group\_by(origin) %>% summarise(sum(dep\_delay))

## # A tibble: 3 x 2  
## origin `sum(dep\_delay)`  
## <chr> <int>  
## 1 EWR 2055092  
## 2 JFK 1599144  
## 3 LGA 1402547

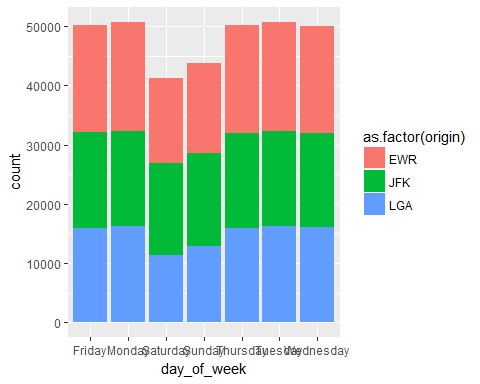
#EWR has more departure delay  
#In which month more departure delays  
mm=fdata %>% select(dep\_delay,origin,month,day) %>%   
 filter(dep\_delay>0) %>%   
 group\_by(month,origin)%>%   
 summarise(total=sum(dep\_delay))  
  
mm%>% spread(origin,total) %>%  
 mutate(total=sum(EWR,JFK,LGA))%>%  
 mutate(EWR\_per = (EWR/total)\*100,JFK\_per = (JFK/total)\*100,LGA\_per = (LGA/total)\*100)

## # A tibble: 12 x 8  
## # Groups: month [12]  
## month EWR JFK LGA total EWR\_per JFK\_per LGA\_per  
## <int> <int> <int> <int> <int> <dbl> <dbl> <dbl>  
## 1 1 166240 102466 72704 341410 48.7 30.0 21.3  
## 2 2 133044 115264 73765 322073 41.3 35.8 22.9  
## 3 3 203944 126772 113344 444060 45.9 28.5 25.5  
## 4 4 204031 135272 126542 465845 43.8 29.0 27.2  
## 5 5 182115 138951 122051 443117 41.1 31.4 27.5  
## 6 6 239994 207728 182382 630104 38.1 33.0 28.9  
## 7 7 243737 250043 185088 678868 35.9 36.8 27.3  
## 8 8 160809 148795 126990 436594 36.8 34.1 29.1  
## 9 9 99517 85985 93312 278814 35.7 30.8 33.5  
## 10 10 115964 70253 89055 275272 42.1 25.5 32.4  
## 11 11 92437 68561 75521 236519 39.1 29.0 31.9  
## 12 12 213260 149054 141793 504107 42.3 29.6 28.1

#Seeing to this LGA is best airport interms of departure delay  
  
ggplot(mm,aes(month,total,fill=as.factor(origin)))+geom\_col(width = 0.3)+facet\_grid(.~origin)



# From visulaization   
# 1. Month 6 and 7 has more delay compare to other months for all origins  
# 2. EWR has more delay   
  
#Reasons for more delay on month 6 and 7 ?  
  
#Add new column for fdata i.e days\_of\_week  
fdata$day\_of\_week = weekdays(as.POSIXct(fdata$sched\_dep\_timeUTC))  
#departure with repect to weekdays and weekends  
ggplot(fdata,aes(day\_of\_week,fill=as.factor(origin)))+geom\_bar()

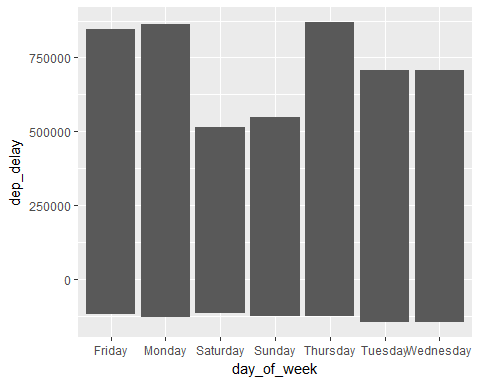


#Surprised weekends are less departured than weekdays  
#Tuesdays has more number of departure  
  
  
##Does days of a week matter for departure delay/ arrival delay  
fdata %>%   
 select(month,day,day\_of\_week,origin,dep\_delay) %>% filter(dep\_delay>0) %>%  
 group\_by(day\_of\_week) %>% summarise(total=sum(dep\_delay))

## # A tibble: 7 x 2  
## day\_of\_week total  
## <chr> <int>  
## 1 Friday 845394  
## 2 Monday 862660  
## 3 Saturday 515515  
## 4 Sunday 549778  
## 5 Thursday 870680  
## 6 Tuesday 705893  
## 7 Wednesday 706863

ggplot(fdata,aes(day\_of\_week,dep\_delay))+geom\_col()

## Warning: Removed 8255 rows containing missing values (position\_stack).



#Friday more delay does it mean more number of flights on thursday  
fdata %>% select(day\_of\_week)%>% group\_by(day\_of\_week) %>% table() %>% sort()

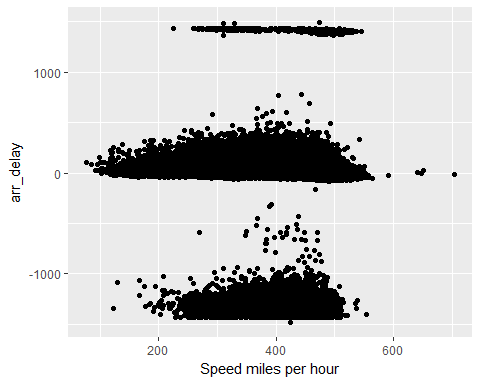
## .  
## Saturday Sunday Wednesday Thursday Friday Tuesday Monday   
## 41278 43796 49998 50137 50153 50705 50709

# does speed matter for departure delay  
#Travel duration  
time = as.double(difftime(fdata$arr\_time,fdata$dep\_time,units = "hour"))   
summary(time)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## -22.750 2.517 3.217 2.517 3.867 12.500 8715

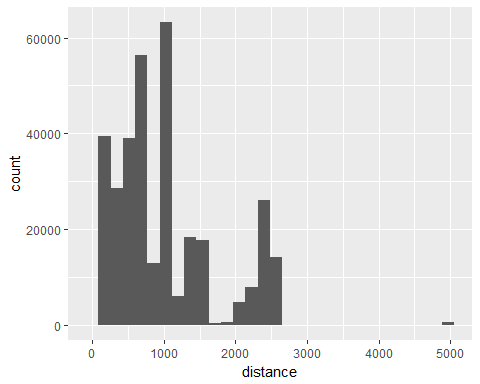
#To Crosscheck   
#y= data.frame(time,fdata$arr\_time,fdata$dep\_time)  
#some time difference is negative due to timezones  
  
#calculating speed  
fdata$speed = fdata$distance\*60 / air\_time  
  
ggplot(fdata,aes(speed,arr\_delay)) + geom\_point()+xlab("Speed miles per hour")

## Warning: Removed 9432 rows containing missing values (geom\_point).



avgSpeed=fdata %>% select(speed,tailnum) %>% group\_by(tailnum) %>% summarise(avg\_speed=mean(speed,na.rm = TRUE))  
  
  
ggplot(fdata,aes(distance))+geom\_histogram()

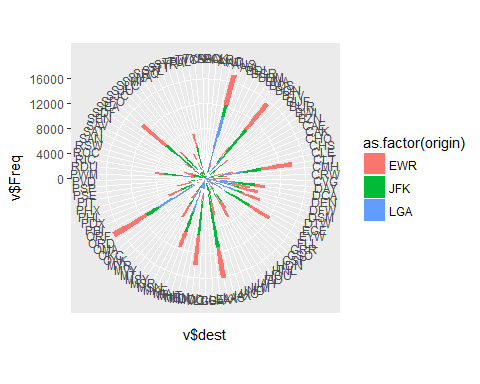
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



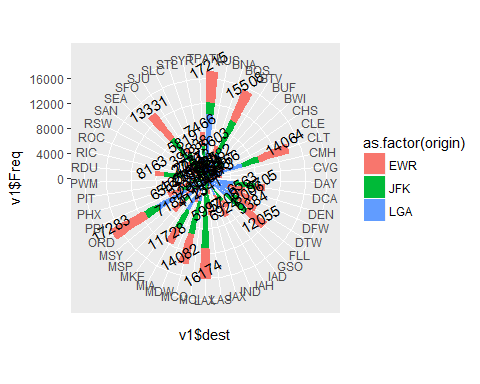
#From visulaization  
# 1. Speed is negative value i.e we have to consider timezone to calculate speed   
# 2. so convert local time to standard time i.e UTC timezone

####################3.Exploring Destination

#Explore with destination   
# v=data.frame(fdata%>% select(dest) %>% table())  
# dim()=c("names","count")  
#   
# ggplot(as.data.frame(v),aes(v)) + geom\_bar() + geom\_count()  
v=fdata%>% select(dest,origin) %>% table()  
v= data.frame(v)   
  
  
#Number of flights for each destination from each origin  
ggplot(v,aes(x=v$dest,y=v$Freq,fill=as.factor(origin))) + geom\_bar(stat="identity")+coord\_polar(theta = "x",direction = 1)



v1=v%>% filter(Freq >1000)  
  
label\_data= v1 %>% select(dest,Freq) %>% group\_by(dest) %>%summarise(total=sum(Freq))  
  
ggplot(v1,aes(x=v1$dest,y=v1$Freq,fill=as.factor(origin))) + geom\_bar(stat="identity")+coord\_polar() +  
 stat\_summary(fun.y=sum,aes(label=..y..,group=dest),geom="text",vjust=-0.2,angle=30)



#From visulaization  
#1. More number of flights for destination ORD  
#2. More number of flights are from EWR

####################4.Exploring with airlines

#Exploring with airline data   
airline\_data = read.csv("airlines\_data.csv",stringsAsFactors = FALSE)  
#select columns which are required for analysis  
str(airline\_data)

## 'data.frame': 6161 obs. of 8 variables:  
## $ X.1 : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Unknown: chr "Private flight" "135 Airways" "1Time Airline" "2 Sqn No 1 Elementary Flying Training School" ...  
## $ X.N : chr "\\N" "\\N" "\\N" "\\N" ...  
## $ X. : chr "-" "" "1T" "" ...  
## $ N.A : chr "N/A" "GNL" "RNX" "WYT" ...  
## $ X.N.1 : chr "" "GENERAL" "NEXTIME" "" ...  
## $ X.N.2 : chr "" "United States" "South Africa" "United Kingdom" ...  
## $ Y : chr "Y" "N" "Y" "N" ...

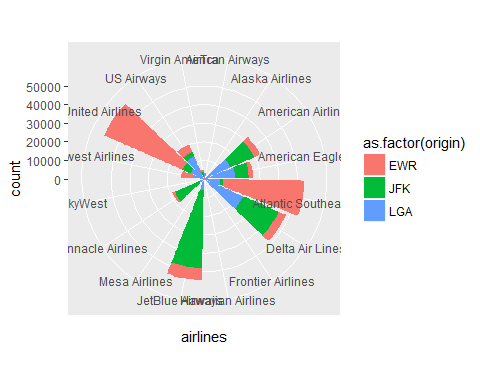
airline\_data = airline\_data[,c(2,4,7)]  
colnames(airline\_data) = c("Airlines\_name","carrier","Country")  
#appending 3 columns Airlines name and country with respect carrier code   
  
ad= inner\_join(data.frame(carrier=fdata$carrier),airline\_data,by="carrier")

## Warning: Column `carrier` joining factor and character vector, coercing  
## into character vector

summary(ad)

## carrier Airlines\_name Country   
## Length:336776 Length:336776 Length:336776   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character

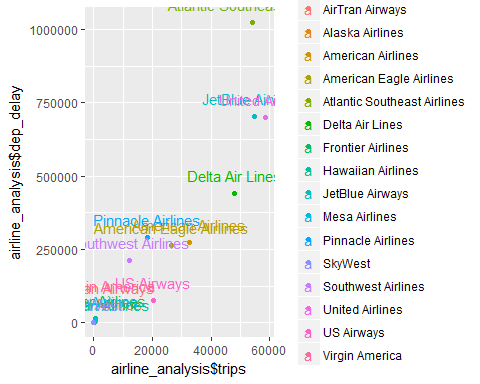
#Appending another two new columns of airline name and country of it  
fdata$airlines=ad$Airlines\_name  
fdata$country=ad$Country  
  
ggplot(fdata,aes(airlines,fill=as.factor(origin)))+geom\_bar()+coord\_polar()



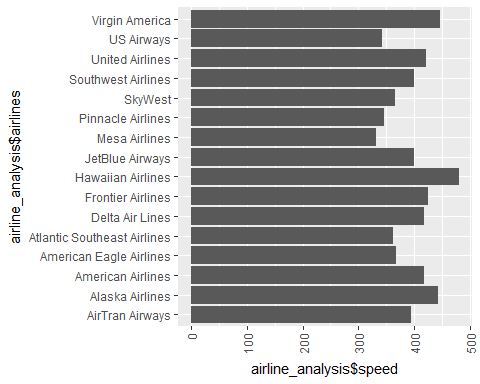
#from visulaization  
#1. United airlines and atlantic southeas are from EWR  
#2. JetBlue airways and delta air lines are from JFK  
airline\_analysis = fdata %>% select(airlines,dep\_delay,speed,distance) %>% group\_by(airlines) %>% summarise(trips=n(),dep\_delay=sum(dep\_delay,na.rm = TRUE),speed = mean(speed,na.rm = TRUE),dist\_min=min(distance),dist\_max=max(distance))  
airline\_analysis

## # A tibble: 16 x 6  
## airlines trips dep\_delay speed dist\_min dist\_max  
## <chr> <int> <int> <dbl> <dbl> <dbl>  
## 1 AirTran Airways 3260 59680 394 397 762  
## 2 Alaska Airlines 714 4133 444 2402 2402  
## 3 American Airlines 32729 275551 417 187 2586  
## 4 American Eagle Airlines 26397 265521 368 184 1147  
## 5 Atlantic Southeast Airlines 54173 1024829 363 80.0 1389  
## 6 Delta Air Lines 48110 442482 418 94.0 2586  
## 7 Frontier Airlines 685 13787 425 1620 1620  
## 8 Hawaiian Airlines 342 1676 480 4983 4983  
## 9 JetBlue Airways 54635 705417 400 173 2586  
## 10 Mesa Airlines 601 10353 332 96.0 544  
## 11 Pinnacle Airlines 18460 291296 345 94.0 1587  
## 12 SkyWest 32 365 366 229 1008  
## 13 Southwest Airlines 12275 214011 401 169 2133  
## 14 United Airlines 58665 701898 421 116 4963  
## 15 US Airways 20536 75168 342 17.0 2153  
## 16 Virgin America 5162 66033 446 2248 2586

ggplot(airline\_analysis,aes(airline\_analysis$trips,airline\_analysis$dep\_delay,color=as.factor(airlines))) + geom\_jitter(height = 2,width = 2) +geom\_text(data=airline\_analysis,label=airline\_analysis$airlines,vjust=-1)



#From visulaization  
#1.Jetblue airways,united airways,atlantic souteast and delta airlines   
#2 US Airways and delta air lines has less delay corresponding to there number of trips  
ggplot(airline\_analysis,aes(airline\_analysis$airlines,airline\_analysis$speed))+geom\_bar(stat="identity")+  
 theme(axis.text.x=element\_text(angle = 90,hjust=1,vjust=0.5))+coord\_flip()

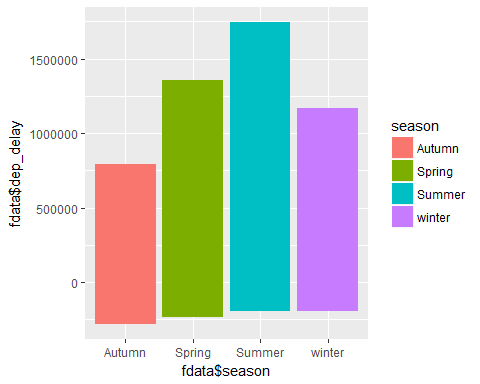


#hawalian airlines have highest speed

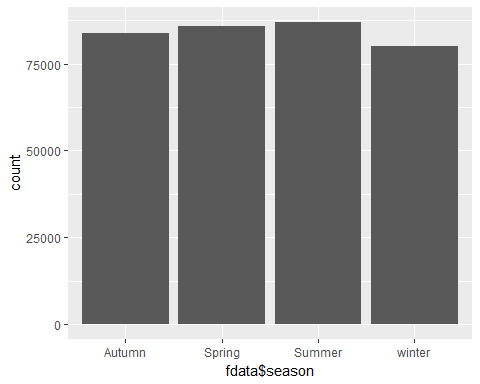
####################5.Departure delay by season

season <- function(m){  
 if\_else(( m >=12 || m <= 2) ,"winter",  
 if\_else( (m>=3 && m<=5), "Spring",  
 if\_else( (m>=6 && m<=8), "Summer","Autumn")))  
   
}  
fdata$season = unlist(lapply(fdata$month,season))  
  
#departure delay with respect season  
ggplot(fdata,aes(fdata$season,fdata$dep\_delay,fill=season))+geom\_bar(stat="identity")

## Warning: Removed 8255 rows containing missing values (position\_stack).

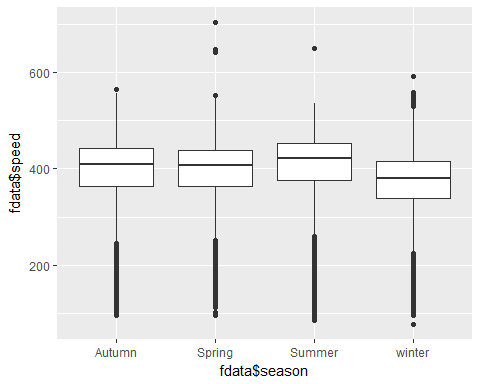


ggplot(fdata,(aes(fdata$season)))+geom\_bar(stat="count")



ggplot(fdata,aes(fdata$season,fdata$speed),fill=season)+geom\_boxplot()

## Warning: Removed 9430 rows containing non-finite values (stat\_boxplot).



#spring and summer has more delay than other seasons

6.exploring with tailnum

attach(fdata)

## The following objects are masked from fdata (pos = 4):  
##   
## air\_time, arr\_delay, arr\_time, arr\_timeUTC, carrier, day,  
## dep\_delay, dep\_time, dep\_timeUTC, dest, distance, flight,  
## hour, minute, month, olson\_dest, olson\_origin, origin,  
## sched\_arr\_time, sched\_arr\_timeUTC, sched\_dep\_time,  
## sched\_dep\_timeUTC, tailnum, time\_hour, year

## The following objects are masked from fdata (pos = 5):  
##   
## air\_time, arr\_delay, arr\_time, carrier, day, dep\_delay,  
## dep\_time, dest, distance, flight, hour, minute, month, origin,  
## sched\_arr\_time, sched\_dep\_time, tailnum, time\_hour, year

## The following object is masked from package:lubridate:  
##   
## origin

#Number of flights travelled   
length(levels(as.factor(fdata$tailnum)))

## [1] 4043

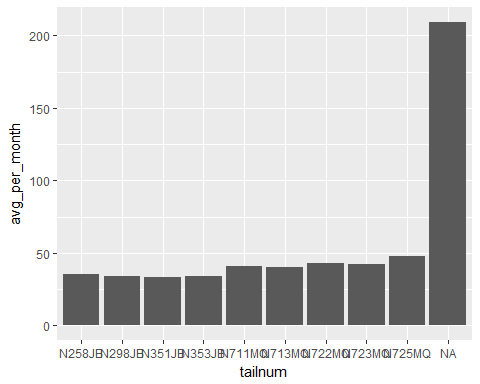
trip=fdata[] %>% select(tailnum,day,month) %>% group\_by(tailnum,month,day) %>% summarise(count = n())  
  
#maximum number of trips per month for each tailnum  
maxtrip\_per\_month=trip %>% group\_by(tailnum,month)%>% summarise(count=sum(count,na.rm = TRUE)) %>%  
 summarise(avg\_per\_month=sum(count)/12)  
head(maxtrip\_per\_month)

## # A tibble: 6 x 2  
## tailnum avg\_per\_month  
## <chr> <dbl>  
## 1 D942DN 0.333  
## 2 N0EGMQ 30.9   
## 3 N10156 12.8   
## 4 N102UW 4.00   
## 5 N103US 3.83   
## 6 N104UW 3.92

maxtrip\_v=maxtrip\_per\_month %>% filter(avg\_per\_month>33)  
head(maxtrip\_v)

## # A tibble: 6 x 2  
## tailnum avg\_per\_month  
## <chr> <dbl>  
## 1 N258JB 35.6  
## 2 N298JB 33.9  
## 3 N351JB 33.5  
## 4 N353JB 33.7  
## 5 N711MQ 40.5  
## 6 N713MQ 40.2

ggplot(maxtrip\_v,aes(tailnum,avg\_per\_month)) +geom\_bar(stat = "identity")



#from visulization N711MQ maximum trips throught the year