Assignment 2

Mohamed Megahed

4/16/2020

```
library(ggplot2)
library(dplyr)

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

Reading Dataset

```
Data <- read.csv("C:/E/MSBA/Spring 2020/Analytics in Practice/Assignment 2/Sample_CA_airtraffic_delays.csv")

#Selecting the needed variables from the dataset
Data_Delays <- Data %>%
select("ArrTime","DepTime","DayofMonth","DayOfWeek","Origin","OriginAirportID","DepDelay","DepDelayMinutes",
"DepDel15","Dest","ArrDelay","ArrDelayMinutes","ArrDel15","Distance","CarrierDelay","WeatherDelay","DestStateName","Cancelled","Diverted","CRSDepTime","CRSArrTime")
```

Data Understanding

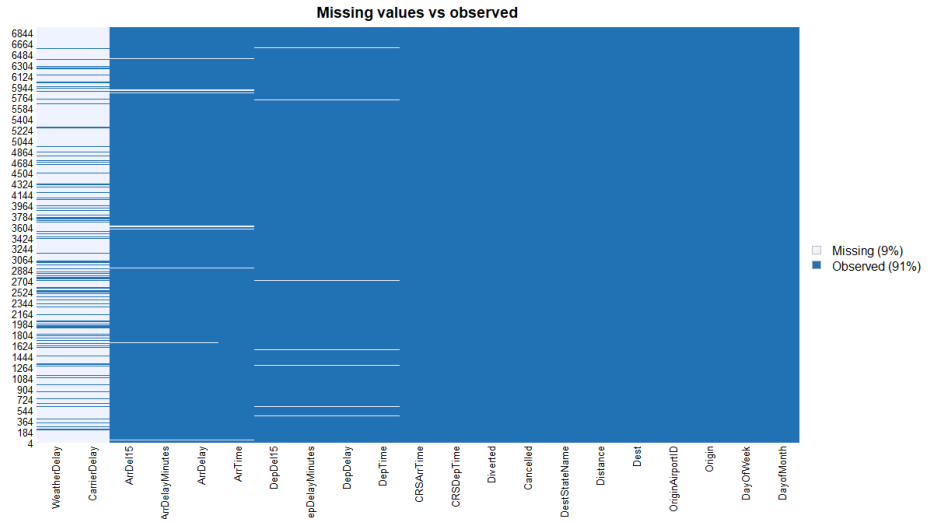
```
summary(Data_Delays)

##     ArrTime        DepTime       DayofMonth      DayOfWeek     Origin
##  Min.   :   1   Min.   :   1   Min.   : 1.00   Min.   :1.000   LAX:2747
##  1st Qu.:1055   1st Qu.: 911   1st Qu.: 8.00   1st Qu.:2.000   SAN:1032
##  Median :1514   Median :1305   Median :16.00   Median :4.000   SFO:1984
##  Mean   :1467   Mean   :1332   Mean   :15.83   Mean   :3.742   SJC: 623
##  3rd Qu.:1925   3rd Qu.:1747   3rd Qu.:24.00   3rd Qu.:5.000   SMF: 548
##  Max.   :2359   Max.   :2400   Max.   :31.00   Max.   :7.000
##  NA's   :153    NA's   :86
##  OriginAirportID    DepDelay        DepDelayMinutes      DepDel15
##  Min.   :12892   Min.   :-292.000   Min.   :   0.00   Min.   :0.0000
##  1st Qu.:12892   1st Qu.:  -6.000   1st Qu.:   0.00   1st Qu.:0.0000
##  Median :14679   Median :  -3.000   Median :   0.00   Median :0.0000
##  Mean   :14028   Mean   :   7.686   Mean   :  11.46   Mean   :0.1688
```

```
##  3rd Qu.:14771   3rd Qu.:   4.000   3rd Qu.:   4.00   3rd Qu.:0.0000
##  Max.   :14893   Max.   :1178.000   Max.   :1178.00   Max.   :1.0000
##                  NA's   :86         NA's   :86        NA's   :86
##       Dest          ArrDelay         ArrDelayMinutes     ArrDel15
##  LAX    : 432   Min.   :-64.0000   Min.   :  0.00   Min.   :0.000
##  SEA    : 380   1st Qu.:-19.0000   1st Qu.:  0.00   1st Qu.:0.000
##  LAS    : 347   Median :-10.0000   Median :  0.00   Median :0.000
##  SFO    : 325   Mean   :  0.4813   Mean   : 11.67   Mean   :0.156
##  PHX    : 295   3rd Qu.:  3.0000   3rd Qu.:  3.00   3rd Qu.:0.000
##  DEN    : 274   Max.   :851.0000   Max.   :851.00   Max.   :1.000
##  (Other):4881   NA's   :160        NA's   :160      NA's   :160
##     Distance      CarrierDelay     WeatherDelay        DestStateName
##  Min.   :  31   Min.   :  0.00   Min.   :  0.000   California:2121
##  1st Qu.: 401   1st Qu.:  0.00   1st Qu.:  0.000   Texas     : 565
##  Median : 794   Median :  0.00   Median :  0.000   Nevada    : 414
##  Mean   : 979   Mean   : 21.44   Mean   :  5.167   Washington: 403
##  3rd Qu.:1400   3rd Qu.: 20.00   3rd Qu.:  0.000   Arizona   : 356
##  Max.   :4962   Max.   :773.00   Max.   :850.000   Colorado  : 340
##                 NA's   :5877     NA's   :5877      (Other)   :2735
##    Cancelled          Diverted          CRSDepTime     CRSArrTime
##  Min.   :0.00000   Min.   :0.000000   Min.   :   5   Min.   :   1
##  1st Qu.:0.00000   1st Qu.:0.000000   1st Qu.: 907   1st Qu.:1103
##  Median :0.00000   Median :0.000000   Median :1304   Median :1514
##  Mean   :0.02149   Mean   :0.001586   Mean   :1332   Mean   :1474
##  3rd Qu.:0.00000   3rd Qu.:0.000000   3rd Qu.:1738   3rd Qu.:1923
##  Max.   :1.00000   Max.   :1.000000   Max.   :2359   Max.   :2359
##
```

Data Pre-Processing

** checking if there any missing data using "Amelia package"

```
library(Amelia)

## Loading required package: Rcpp

## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.7.6, built: 2019-11-24)
## ## Copyright (C) 2005-2020 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##

missmap(Data_Delays, main = "Missing values vs observed",legend = TRUE)
```




```
anyNA(Data_Delays)

## [1] TRUE

sum(is.na(Data_Delays))

## [1] 12731
```

Replacing NA values with 0

```
Data_Cleaned <- Data_Delays
Data_Cleaned[is.na(Data_Cleaned)] <- 0

# converting time and date variables to factors
Data_Cleaned$DayofMonth <- as.factor(Data_Cleaned$DayofMonth)
#Data_Cleaned$DayOfWeek <- as.factor(Data_Cleaned$DayOfWeek)
#Data_Cleaned$DepDel15 <- as.factor(Data_Cleaned$DepDel15)

str(Data_Cleaned)

## 'data.frame':    6934 obs. of  21 variables:
##  $ ArrTime        : num  2021 1559 1359 2323 2031 ...
##  $ DepTime        : num  2213 1241 2007 1057 20 ...
##  $ DayofMonth     : Factor w/ 31 levels "1","2","3","4",..: 16 21 21 5 12
4 4 17 1 8 ...
##  $ DayOfWeek      : int  2 7 7 5 5 4 4 3 1 1 ...
##  $ Origin         : Factor w/ 5 levels "LAX","SAN","SFO",..: 3 1 1 1 1 2 2
1 1 1 ...
##  $ OriginAirportID: int  14771 12892 12892 12892 12892 14679 14679 12892
12892 12892 ...
##  $ DepDelay       : num  -7 -4 2 -3 0 -4 -2 -5 -4 11 ...
##  $ DepDelayMinutes: num  0 0 2 0 0 0 0 0 0 11 ...
##  $ DepDel15       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Dest           : Factor w/ 95 levels "ABQ","ACV","ANC",..: 40 6 82 86
56 8 87 88 47 83 ...
```

```
##  $ ArrDelay       : num  -24 -19 -23 -17 -12 -21 -6 -8 -27 -18 ...
##  $ ArrDelayMinutes: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ArrDel15       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Distance       : int  820 1009 1024 224 1956 1065 1635 909 200 2454 ...
##  $ CarrierDelay   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ WeatherDelay   : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ DestStateName  : Factor w/ 36 levels "Alaska","Arizona",..: 24 31 34 4
17 30 32 4 6 4 ...
##  $ Cancelled      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Diverted       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ CRSDepTime     : int  2220 1245 2005 1100 20 1525 620 2035 1603 1355
...
##  $ CRSArrTime     : int  2045 1618 1422 2340 2043 1141 1717 1353 2040 2202
...
```

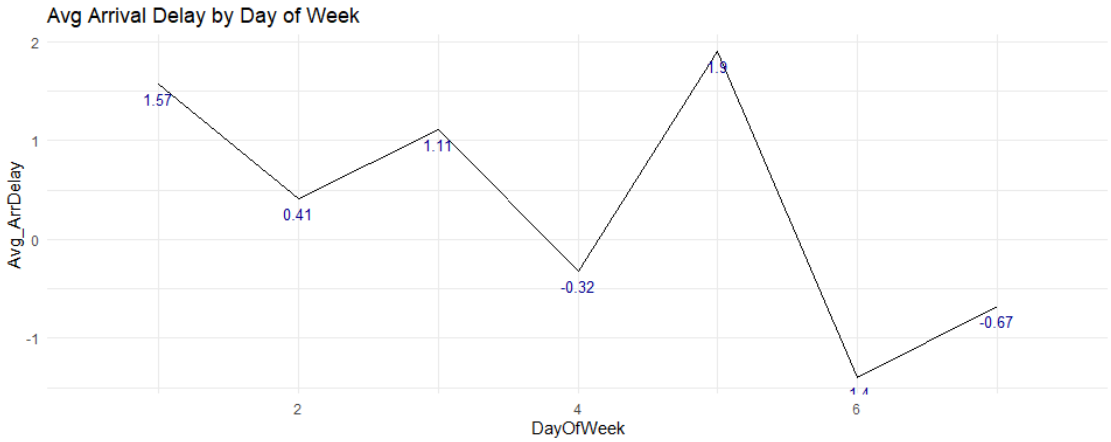
```
head(Data_Cleaned)
```

```
##   ArrTime DepTime DayofMonth DayOfWeek Origin OriginAirportID DepDelay
## 1    2021    2213         16         2    SFO           14771       -7
## 2    1559    1241         21         7    LAX           12892       -4
## 3    1359    2007         21         7    LAX           12892        2
## 4    2323    1057          5         5    LAX           12892       -3
## 5    2031      20         12         5    LAX           12892        0
## 6    1120    1521          4         4    SAN           14679       -4
##   DepDelayMinutes DepDel15 Dest ArrDelay ArrDelayMinutes ArrDel15 Distance
## 1               0        0  JFK      -24               0        0      820
## 2               0        0  AUS      -19               0        0     1009
## 3               2        0  SEA      -23               0        0     1024
## 4               0        0  SJC      -17               0        0      224
## 5               0        0  MSP      -12               0        0     1956
## 6               0        0  BNA      -21               0        0     1065
##   CarrierDelay WeatherDelay DestStateName Cancelled Diverted CRSDepTime
## 1            0            0      New York         0        0       2220
## 2            0            0         Texas         0        0       1245
## 3            0            0    Washington         0        0       2005
## 4            0            0    California         0        0       1100
## 5            0            0     Minnesota         0        0         20
## 6            0            0     Tennessee         0        0       1525
##   CRSArrTime
## 1       2045
## 2       1618
## 3       1422
## 4       2340
## 5       2043
## 6       1141
```

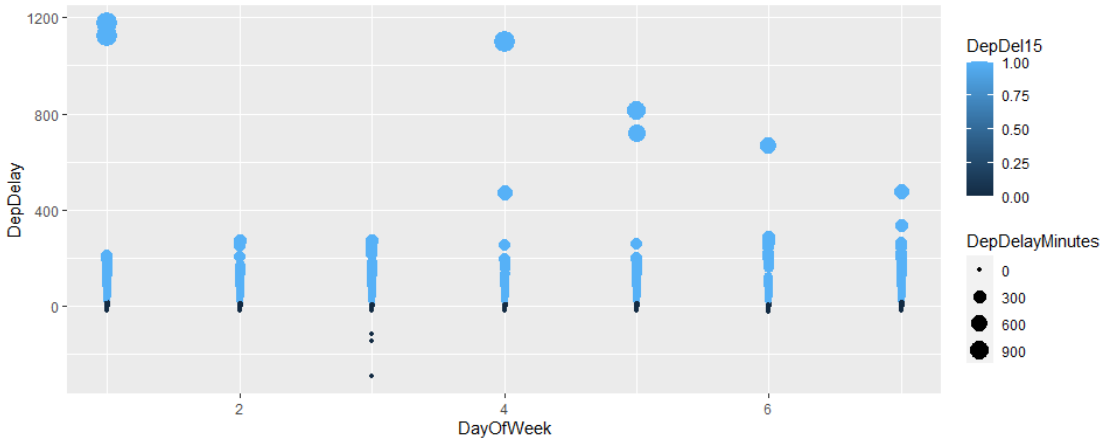
Q1 and 2

```
Month <- Data_Cleaned %>% group_by( DayofMonth ) %>%
    summarise( Count = n(), Avg_ArrTime=mean(ArrTime),
Avg_ArrDelay=mean(ArrDelay),  Avg_DepTime=mean(DepTime),
```

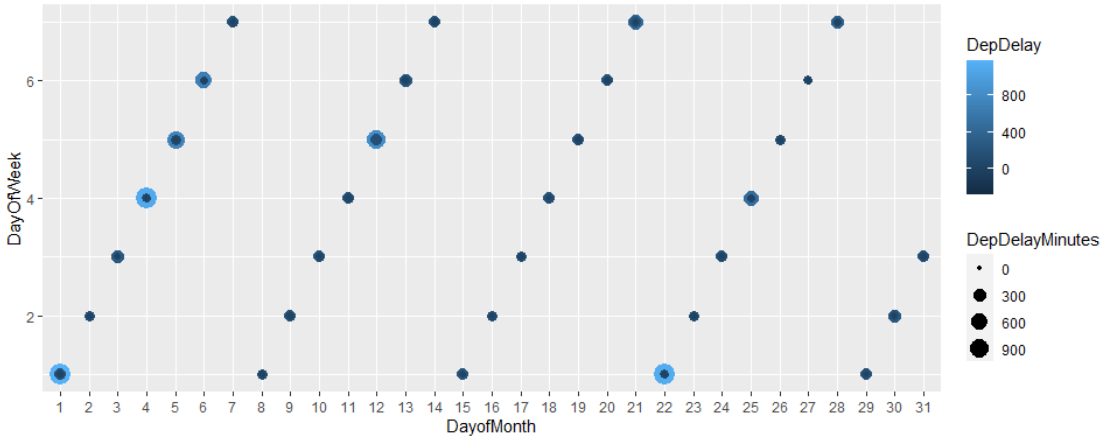
```
Avg_DepDelay=mean(DepDelay))

Week <- Data_Cleaned%>% group_by( DayOfWeek ) %>%
    summarise( Count = n(), Avg_ArrTime=mean(ArrTime),
Avg_ArrDelay=mean(ArrDelay),  Avg_DepTime=mean(DepTime),
Avg_DepDelay=mean(DepDelay))

ggplot(data=Week, aes(x=DayOfWeek, y=Avg_ArrDelay))
+geom_line(stat="identity", fill="steelblue")+
  geom_text(aes(label=round(Avg_ArrDelay, 2)), vjust=1.6, color="darkblue",
  position = position_dodge(0.9),
size=3.5)+scale_fill_brewer(palette="Paired")+
  theme_minimal()+labs(title="Avg Arrival Delay by Day of Week")

## Warning: Ignoring unknown parameters: fill
```



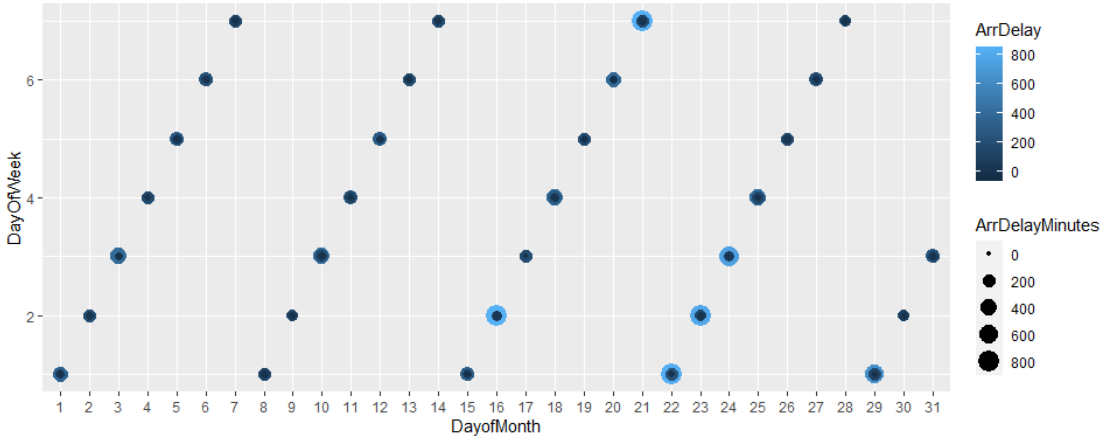

```
ggplot(Data_Cleaned, aes(x=DayOfWeek, y= DepDelay, color = DepDel15, size =
DepDelayMinutes)) + geom_point()
```



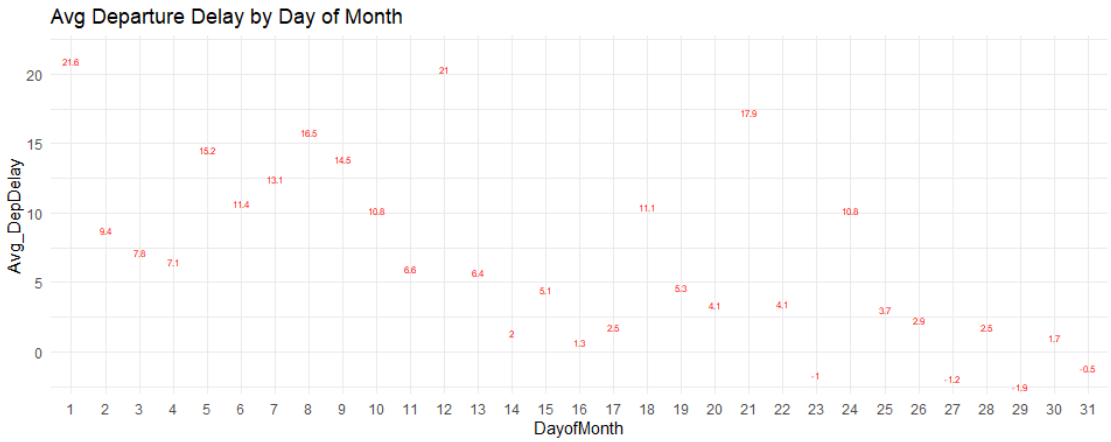

```
# Arrival and Departure Delays per week of the month and days
ggplot(Data_Cleaned, aes(x=DayofMonth, y= DayOfWeek, color =DepDelay , size
=DepDelayMinutes)) + geom_point()
```



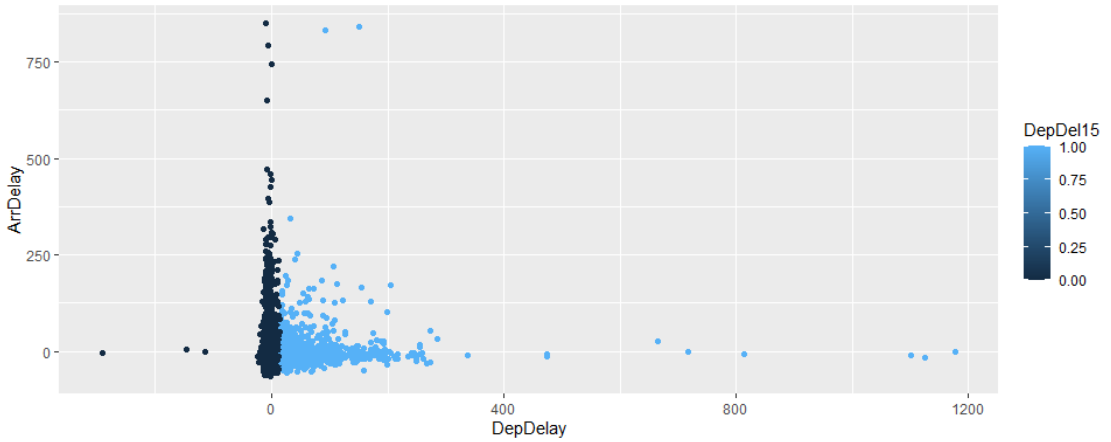

```
ggplot(Data_Cleaned, aes(x=DayofMonth, y= DayOfWeek, color =ArrDelay , size
=ArrDelayMinutes)) + geom_point()
```




```
ggplot(data=Month, aes(x=DayofMonth, y=Avg_DepDelay))
+geom_line(stat="identity", fill="steelblue")+
  geom_text(aes(label=round(Avg_DepDelay, 1)), vjust=1.6, color="red",
  position = position_dodge(0.9),
size=2)+scale_fill_brewer(palette="Paired")+
  theme_minimal()+labs(title="Avg Departure Delay by Day of Month")

## Warning: Ignoring unknown parameters: fill

## geom_path: Each group consists of only one observation. Do you need to
adjust
## the group aesthetic?
```



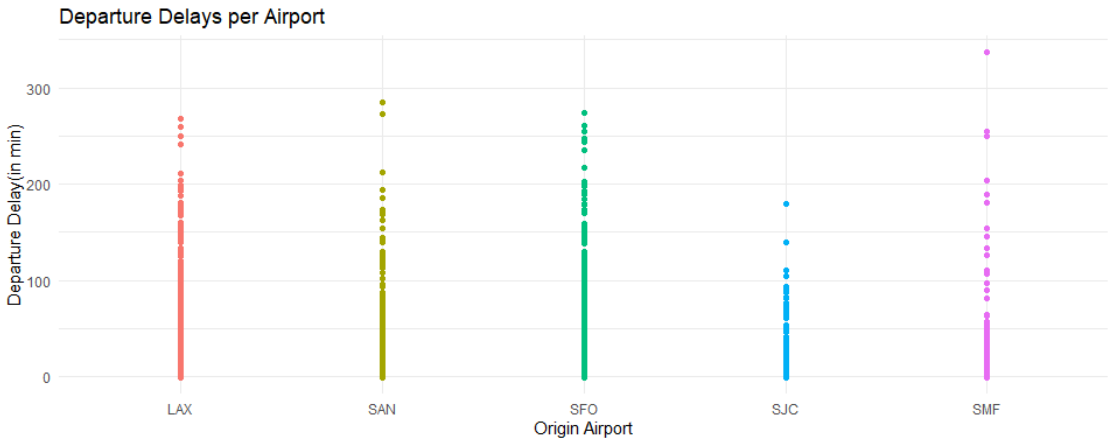

```
ggplot(Data_Cleaned, aes(x=DepDelay, y= ArrDelay, color = DepDel15, sahpe =
DepDelayMinutes)) + geom_point()
```



Q3

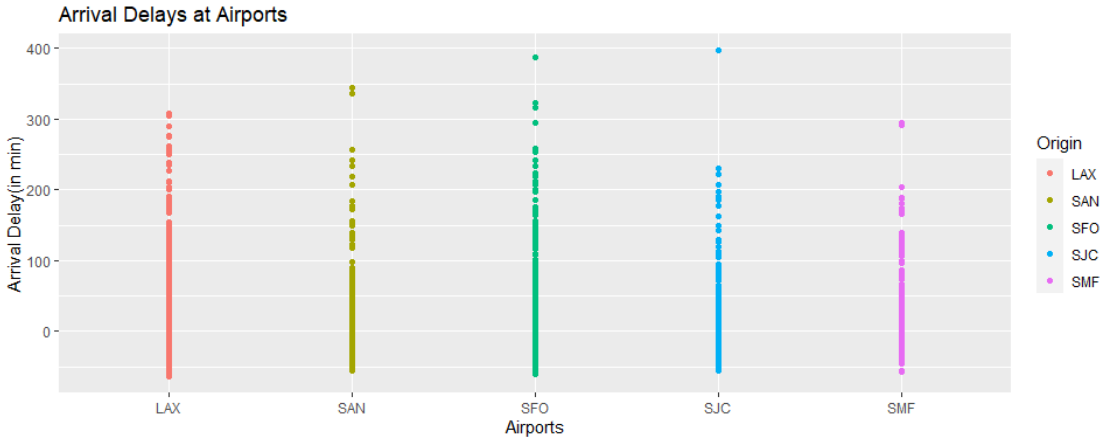
Which Airport ('Origin Airport') has highest departure delay?

```
Data_Cleaned %>%
 filter(DepDelay >= -1L & DepDelay <= 403L) %>%
 ggplot() +aes(x = Origin, y = DepDelay, color = Origin) +
 geom_point() +
 scale_fill_hue() +
 labs(x = "Origin Airport", y = "Departure Delay(in min)", title = "Departure
Delays per Airport") +
 theme_minimal() +
 theme(legend.position = "none")
```



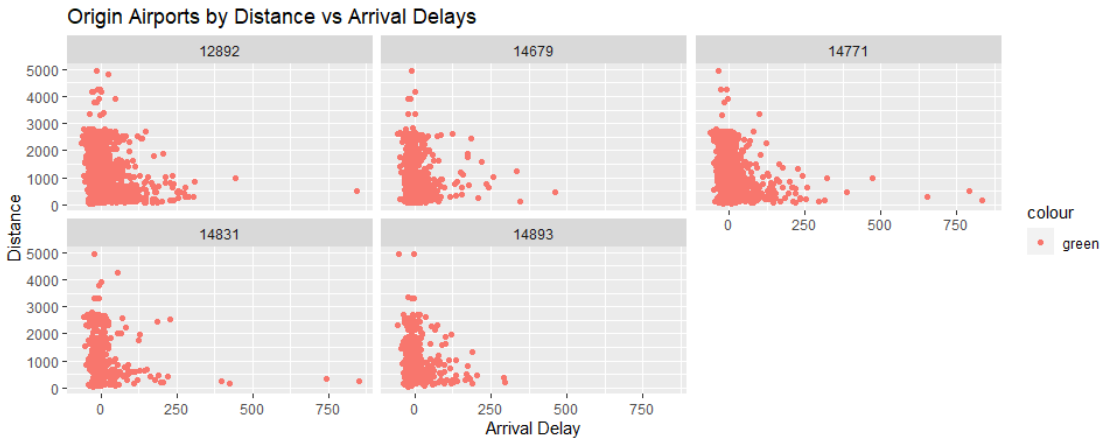
Q4 Which Airport has highest Arrival delay?

```
Data_Cleaned %>%
 filter(ArrDelay >= -64L & ArrDelay <= 405L) %>%
 ggplot() +aes(x = Origin, y = ArrDelay, color = Origin) +
 geom_point() +
 scale_fill_hue() +
 labs(x = "Airports", y = "Arrival Delay(in min)", title = "Arrival Delays at
Airports")
```



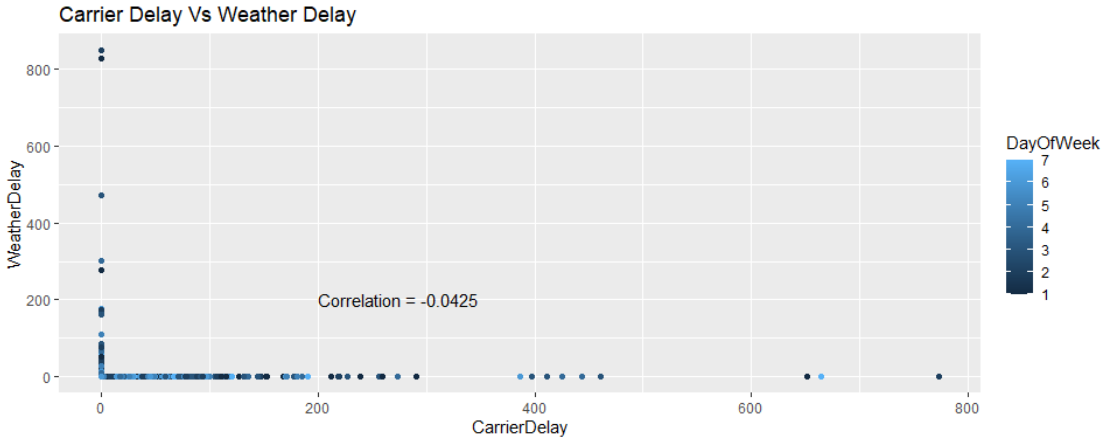
Q5 How do you relate the delay pattern to the distance travelled?

```
ggplot(Data_Cleaned, aes(ArrDelay, Distance)) +
  geom_point(aes(colour= "green"))+
  facet_wrap(~OriginAirportID)+labs(title="Origin Airports by Distance vs
Arrival Delays")+xlab("Arrival Delay") + ylab("Distance")
```



Q6 Is there any correlation between weather delay and carrier delay?

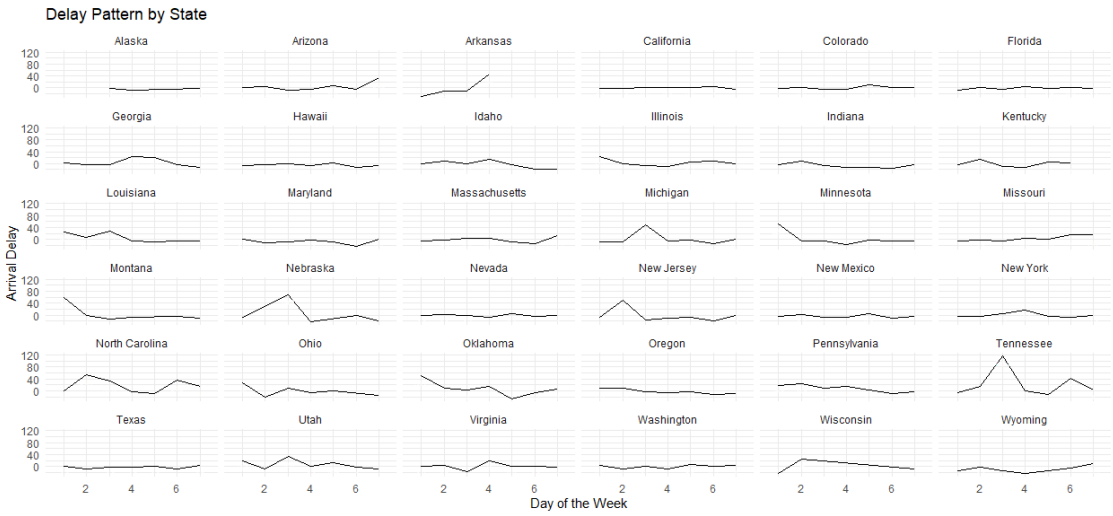
```
ggplot(Data_Cleaned, aes(x=CarrierDelay, y= WeatherDelay, color = DayOfWeek))
+ geom_point()+labs(title = "Carrier Delay Vs Weather Delay")+
  annotate(geom = "text",x=200,y=200,label="Correlation = -
0.0425",hjust="left")
```



Q7 What is the delay pattern you can find in respective states?

```
State_df<-Data_Cleaned %>%
    group_by( DestStateName, DayOfWeek ) %>%
    summarise( Count = n(), Avg_ArrTime=mean(ArrTime),
Avg_ArrDelay=mean(ArrDelay),  Avg_DepTime=mean(DepTime),
Avg_DepDelay=mean(DepDelay))
ggplot(State_df, aes(DayOfWeek, Avg_ArrDelay)) +geom_line(stat="identity",
fill="steelblue")+
  geom_text(aes(label=round(Avg_ArrDelay, 1)), vjust=1.6, color="white",
            position = position_dodge(0.9), size=2)+
  scale_fill_brewer(palette="Paired")+
  theme_minimal()+
  facet_wrap(~DestStateName)+labs(title="Delay Pattern by State")+xlab("Day
of the Week") + ylab("Arrival Delay")
```

```
## Warning: Ignoring unknown parameters: fill
```



Q8 How many delayed flights were cancelled?

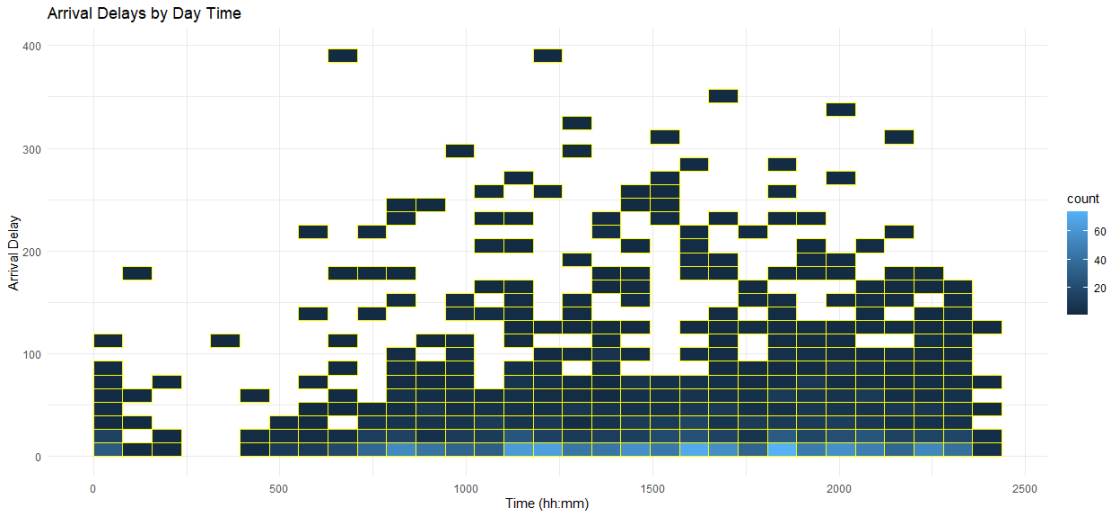
```
cancelled_df<-Data_Cleaned %>%
    filter(DepDelay>0) %>% filter(Cancelled==1) %>%
    summarise( Count = n())
cancelled_df

##   Count
## 1    44
```

Q9 How many delayed flights were diverted?

```
diverted_df<-Data_Cleaned%>%
    filter(DepDelay>0) %>% filter(Diverted==1) %>%
    summarise( Count = n())
diverted_df

##   Count
## 1     3
```

Q10 What time of the day do you find Arrival delays?

```
Data_Cleaned %>%
 filter(ArrDelay >= 0L & ArrDelay <= 410L) %>%
 ggplot() + aes(x = CRSArrTime, y = ArrDelay) +
 geom_bin2d(size = 0.5, colour = "yellow") +
 labs(x = "Time (hh:mm)", y = "Arrival Delay", title = "Arrival Delays by Day
Time") +
 theme_minimal()
```



Q11 What time of the day do you find Departure delays?

```
Data_Cleaned %>%
 filter(DepDelay >= -1L & DepDelay <= 417L) %>%
 ggplot() + aes(x = CRSDepTime, y = DepDelay) +
 geom_bin2d(size = 0.5, colour = "red") +
 labs(x = "Time (hh:mm)", y = "Arrival Delay", title = "Departure Delays by
Day Time") +
 theme_minimal()
```