FLIGHTS ANALYSIS PROJECT

MARYANN MEILIKA

ANALYTICS PROJECT

The aviation areas generate huge amounts of data

- Weather, flight information and sensor data can be used to develop predictive model using big data analytics

- Influence of weather on the flight delays can lead to helpful insights for the airport and airline operations to enhance flight safety, improve customer satisfaction and reduce flight delays

- Taxi Out time plays a role for the companies owning the planes. So companies try to minimize their taxi time as both convenience to the customer and themselves by minimizing the amount of time the passenger spends in the plane but also saving cost due to the less fuel burn.

- Airline companies can make proper arrangements for the passengers if the flight is getting delayed and make any schedule change for connecting flights

- Citizens can arrange for change in their plans after the flight in case the flight is getting delayed

View flights dataset

final.R × mydata ×

Filter

	Airport.Code	Month.Name	Origin	Dest	TailNum	Distance	TaxiIn	TaxiOut	Delays.Carrier	Delays.Late.Aircraft	ArrDelay	Delays.Security	Delays.Weather
44	LAX	April	ABQ	DEN	N727SW	349	5	8	255	348	333	3	
45	LAX	May	ABQ	MDW	N628SW	223	3	7	1326	1515	5402	2	
46	LAX	May	ABQ	MDW	N324SW	759	3	7	435	662	1295	4	
47	LAX	May	ABQ	MDW	N232WN	759	3	11	393	795	730	7	
48	LAX	May	ABQ	MDW	N506SW	759	3	6	262	424	556	1	
49	LAX	May	ABQ	LAS	N213WN	478	4	8	323	328	713	2	
50	LAX	May	ABQ	LAS	N256WN	487	4	10	631	525	774	3	
51	LAX	May	ABQ	LAS	N685SW	487	3	10	926	930	1724	2	
52	MDW	May	ABQ	LAS	N318SW	487	4	8	496	550	1094	3	
53	PHX	June	ABQ	LAS	N270WN	487	3	10	853	1343	1214	21	
54	SAN	June	ABQ	LAS	N265WN	487	4	9	1213	1217	1614	21	
55	PHL	June	ABQ	LAS		487	NA	NA	423	603	1212	5	
56	PHX	June	ABQ	LAX	N209WN	677	10	8	565	812	818	6	
57	SAN	June	ABQ	LAX	N390SW	677	9	14	257	763	892	11	
58	PHL	June	ABQ	LAX	N636WN	677	6	10	348	382	554	2	
59	DAL	June	ABQ	LAX	N329SW	677	27	9	712	650	1150	14	
60	DAL	June	ABQ	PHL	N512SW	718	2	7	1269	2422	5016	7	
61	DAL	June	ABQ	PHL	N378SW	718	3	13	284	426	306	6	

Showing 44 to 61 of 158 entries, 19 total columns

exploring dataset

```r
#exploring dataset
nrow(mydata)
names(mydata)
summary(mydata)
str(mydata)
head(mydata)
tail(mydata)
```

```
> #exploring dataset
> nrow(mydata)
[1] 158
> names(mydata)
 [1] "Airport.Code"         "Month.Name"           "Origin"               "Dest"                 "TailNum"
 [6] "Distance"             "TaxiIn"               "TaxiOut"              "Delays.Carrier"       "Delays.Late.Aircraft"
[11] "ArrDelay"             "Delays.Security"      "Delays.Weather"       "Carriers.Total"       "Flights.Cancelled"
[16] "Flights.Delayed"      "Flights.Diverted"     "Flights.On.Time"      "Flights.Total"
> summary(mydata)
 Airport.Code        Month.Name           Origin             Dest             TailNum            Distance
 Length:158         Length:158         Length:158         Length:158         Length:158         Min.   : 148.0
 Class :character   Class :character   Class :character   Class :character   Class :character   1st Qu.: 337.2
 Mode  :character   Mode  :character   Mode  :character   Mode  :character   Mode  :character   Median : 580.0
                                                                                                 Mean   : 636.8
                                                                                                 3rd Qu.: 947.0
                                                                                                 Max.   :1670.0

     TaxiIn          TaxiOut   Delays.Carrier   Delays.Late.Aircraft    ArrDelay      Delays.Security  Delays.Weather
 Min.   : 2.000   8      :30   Min.   : 177.0   Min.   : 233.0       Min.   : 157.0   Min.   : 0.000   Min.   :  8.00
 1st Qu.: 3.000   7      :28   1st Qu.: 349.5   1st Qu.: 427.8       1st Qu.: 468.8   1st Qu.: 4.000   1st Qu.: 41.25
 Median : 4.000   9      :24   Median : 460.5   Median : 600.0       Median : 780.0   Median : 7.000   Median : 71.00
```

data structures

a. Vector
b. Matrix
c. Data frame

```r
#Creating Vectors of days
days <- c("sunday", "Monday", "Tuesday", "Wednesday", "Thursday", "Friday", "saturday" )

#MATRIX
# create a matrix of origin and dest and air port code from a series of vectors

Origin<-c("OAK","RDU","SLC","TPA","TUS","ABQ","ALB","AMA","AUS")
Dest <-c("PHL","PHX","MDW", "SAN","BWI","DAL","DEN","LAS","LAX")
Airport.Code <-c("PHL","PHX","SAN","DEN","DAL","MDW","BWI","MCO","LAX")


# Create box_office
box_office <- c(Origin, Dest,Airport.Code)

#and now the matrix
flight_matrix <- matrix(box_office , byrow = TRUE,nrow = 3 )


# Vectors titles, used for naming
titles <- c("Origin", "Dest", "Airport.Code")

# Name the rows with titles
rownames(flight_matrix) <- titles
flight_matrix
```

```r
# Create a data frame of delayed
# Definition of vectors
Delays.Security <- c(15,2,3,0,1,14,17)
Delays.Weather  <- c(599,85,60,47,76,171,413)
Delays.Carrier <- c(1302,450,371,303,360,850,1078)
TaxiIn  <- c(5,5,5,3,2,4,4)
TaxiOut <- c(21,6,5,6,14,6,7)

delayed_df <- data.frame(Delays.Security,Delays.Weather,Delays.Carrier,TaxiIn,TaxiOut)
delayed_df
```

Using the DPLYR library

1- Select including sub setting

GGplot tutorial updated 2021-06-16.R | final.R | new.R* | bigdata2_project.R* | assigment week5.R | lab11.R | R Plotting exercise.R

Source on Save | Run | Source

```r
#Select  columns
mydata7 <- mydata %>% select( "Month.Name", "TailNum","TaxiIn", "TaxiOut","Origin", "Dest" ,"Distance")
mydata7
```

92:8 (Top Level) R Script

Console | Terminal | Jobs

R 4.2.0 · ~/

```
> #Select  columns
> mydata7 <- mydata %>% select( "Month.Name", "TailNum","TaxiIn", "TaxiOut","Origin", "Dest" ,"Distance")
> mydata7
   Month.Name TailNum TaxiIn TaxiOut Origin Dest Distance
1     January  N752SW      5      21    OAK  DEN      180
2     January  N388SW      5       6    RDU  DEN      632
3     January  N341SW      5       5    RDU  DEN      632
4     January  N701GS      3       6    RDU  PHL      336
5     January  N295WN      2      14    RDU  PHL      336
6     January  N483WN      4       6    RDU  PHL      336
7     January  N214WN      4       7    RDU  PHL      336
8     Febuary  N345SA      5      10    SLC  PHX      507
9     Febuary  N369SW      4       8    SLC  PHX      507
10    Febuary  N283WN      6      12    SLC  PHX      507
11    Febuary  N355SW      8       8    SLC  PHX      507
12    Febuary  N494WN      4      10    SLC  PHX      507
13    Febuary  N405WN      7      10    SLC  PHX      507
14    Febuary  N605SW      6       9    SLC  PHX      507
15    Febuary  N704SW      5      25    SLC  DAL      422
16    Febuary  N351SW      7      12    SLC  DAL      422
17    Febuary  N398SW      5      15    SLC  DAL      422
```

Arrange

```r
#Arrange rows with arrange()
mydata6 <-mydata %>% arrange(Flights.Diverted, Flights.Delayed)
mydata6

#Select  columns
```

```
> #Arrange rows with arrange()
> mydata6 <-mydata %>% arrange(Flights.Diverted, Flights.Delayed)
> mydata6
   Airport.Code Month.Name Origin Dest TailNum Distance TaxiIn TaxiOut Delays.Carrier Delays.Late.Aircraft ArrDelay
1           MDW      March    TPA  PHX  N273WN      835      3       8            413                  589      406
2           MDW       July    ABQ  LAX  N437WN      993      6      16            293                  406      210
3           BWI  September    AMA  MCO  N412WN      277      4       6            219                  304      399
4           PHX     August    ALB  BWI  N763SW      288      2       7            300                  420      289
5           SAN      March    TUS  SAN  N348SW      367     NA       7            220                  462      390
6           BWI    October    AUS  BNA  N203WN      756      4       8            619                  412      431
7           PHX      April    ABQ  DAL  N302SW      580      3       9            185                  233      157
8           PHX      March    TUS  PHX  N352SW      451      5      10            223                  437      402
9           PHX   November    AUS  DAL  N486WN      189      5       5            377                  604      299
10          PHL   November    AUS  ELP  N378SW      528      3      10            442                  502      440
11          PHL   December    AUS  LAS  N603SW     1090      4       9            528                  365      565
12          BWI    October    AUS  BWI  N324SW     1342      2       8            253                  313      202
13          MCO   November    AUS  ELP  N359SW      148      3      18            240                  234      293
14          LAX      April    ABQ  DEN  N727SW      349      5       8            255                  348      333
15          SAN      March    TUS  PHX  N278WN      451     10      12            318                  296      428
16          LAX      April    ABQ  DEN  N777QC      349      7       9            549                  284      628
17          PHX    January    RDU  DEN  N388SW      632      5       6            450                  699      642
18          BWI  September    ALB  MCO  N724SW     1130      2      24            278                  385      304
19          DAL       June    ABQ  PHL  N378SW      718      3      13            284                  426      306
20          DAL       June    ABQ  PHL  N238WN      718      7      10            471                  624      467
21          DAL       June    ABQ  MDW  N478WN     1121      4       7            617                  630     1318
```

Filter

```r
#filter rows which origin= SLC and dest = DAL and using Pipe
mydata4 <- mydata %>% filter(Origin == "SLC")%>% filter(Dest == "DAL")
mydata4
```

```
43          2155
44          1291
45         14204
46          3784
47          2767
48          1847
49          2189
50          2786
 [ reached 'max' / getOption("max.print") -- omitted 108 rows ]
> #filter rows which origin= SLC and dest = DAL and using Pipe
> mydata4 <- mydata %>% filter(Origin == "SLC")%>% filter(Dest == "DAL")
> mydata4
  Airport.Code Month.Name Origin Dest TailNum Distance TaxiIn TaxiOut Delays.Carrier Delays.Late.Aircraft ArrDelay
1          SAN    Febuary    SLC  DAL  N704SW      422      5      25           339                 416      302
2          PHL    Febuary    SLC  DAL  N351SW      422      7      12           340                 534      411
3          PHX    Febuary    SLC  DAL  N398SW      422      5      15           463                 388     1447
  Delays.Security Delays.Weather Carriers.Total Flights.Cancelled Flights.Delayed Flights.Diverted Flights.On.Time
1               4             38            13                65           1100                9           5451
2              21             26            13               112           1334               10           6558
3              12             53            12               205           2366               11           7445
  Flights.Total
1          6625
2          8014
3         10027
```

```r
#filter rows of Distance =422 and  ArrDelay = 302
mydata5 <- mydata %>% filter(Distance == 422 , ArrDelay == 302)
mydata5
```

```
> #filter rows of Distance =422 and  ArrDelay = 302
> mydata5 <- mydata %>% filter(Distance == 422 , ArrDelay == 302)
> mydata5
  Airport.Code Month.Name Origin Dest TailNum Distance TaxiIn TaxiOut Delays.Carrier Delays.Late.Aircraft ArrDelay
1          SAN    Febuary    SLC  DAL  N704SW      422      5      25           339                 416      302
  Delays.Security Delays.Weather Carriers.Total Flights.Cancelled Flights.Delayed Flights.Diverted Flights.On.Time
1               4             38            13                65           1100                9           5451
  Flights.Total
1          6625
>
```

Pipe

```r
#filter rows using Pipe that CarrierDelay !=NA and select some columns and arrange them by descending of distance
mydata8 <-mydata %>% filter(Delays.Carrier !="NA")%>%
  select( "Month.Name", "TailNum","TaxiIn", "TaxiOut","Origin", "Dest" ,"Distance") %>%
  arrange(desc(Distance))%>%group_by(TailNum)
mydata8
```

```
> #filter rows using Pipe that CarrierDelay !=NA and select some columns and arrange them by descending of distance
> mydata8 <-mydata %>% filter(Delays.Carrier !="NA")%>%
+   select( "Month.Name", "TailNum","TaxiIn", "TaxiOut","Origin", "Dest" ,"Distance") %>%
+   arrange(desc(Distance))%>%group_by(TailNum)
> mydata8
# A tibble: 158 × 7
# Groups:   TailNum [135]
   Month.Name TailNum TaxiIn TaxiOut Origin Dest  Distance
   <chr>      <chr>    <int>   <int> <chr>  <chr>    <int>
 1 March      N281WN       4       8 ABQ    BWI       1670
 2 March      N789SW       3       8 ABQ    BWI       1670
 3 March      N751SW       6      14 TUS    MDW       1440
 4 March      N408WN       5      10 TUS    MDW       1440
 5 October    N324SW       2       8 AUS    BWI       1342
 6 November   N679AA       3      10 AUS    BWI       1342
 7 December   N479WN       5       9 AUS    MDW       1242
 8 December   N489WN       3      10 AUS    MDW       1242
 9 September  N724SW       2      24 ALB    MCO       1130
10 June       N478WN       4       7 ABQ    MDW       1121
# … with 148 more rows
>
```

Mutate

```r
mydata1

#adding delay_result column
mydata2 <- mutate(mydata, DELAY_RESULT = Flights.Delayed  + ArrDelay )
mydata2

#filter rows which origin= RDU and dest = PHL
```

72:8 (Top Level) R Script

Console Terminal Jobs

R 4.2.0 · ~/

```
37                45            587                2            3978          4612
38                65           1935               12            7037          9049
39               150           1991               20           11957         14118
40                40           1019               28            5870          6957
41                48           1187               16            7402          8653
42                73           1496                5            9183         10757
43                79           1527                3            9463         11072
44                 4            958                3            4900          5865
45               457           8802               72           25565         34896
46               225           2489                9            8262         10985
47                49           2037               37            6779          8902
48                80           1291                7            7479          8857
49               138           1476               25            5976          7615
50               106           2012               49           10653         12820
   DELAY_RESULT
1         12316
2          2521
3          2019
4          1976
5          1759
6          3533
7          8448
8          1610
```

Rank

```r
#Calculate Rank for Variables

# calculate rank for variables from Flights.On.Time to Flights.Delayed.
mydata9 = mutate_at(mydata, vars(Flights.On.Time:Flights.Delayed), funs(Rank=min_rank(.)))
mydata9
```

```
38              0            33            13            65          1935            12          7037
39              9            43            15           150          1991            20         11957
40              5            28            13            40          1019            28          5870
41             11            17            14            48          1187            16          7402
42              9            27            12            73          1496             5          9183
43             17            52            12            79          1527             3          9463
44              3            18            11             4           958             3          4900
45              2           556            11           457          8802            72         25565
   Flights.Total Flights.On.Time_Rank Flights.Diverted_Rank Flights.Delayed_Rank
1          34338                 156                    89                   151
2          10133                  75                    12                    71
3           8632                  50                    93                    56
4           9099                  67                    85                    41
5           7531                  38                    58                    30
6          12719                 106                    46                   106
7          29315                 146                   145                   147
8           7241                  39                    80                    21
9           5356                  14                    22                    27
10         11745                  95                    93                   109
11         30970                 141                    93                   157
12          4315                   3                    37                     2
13          8645                  51                    37                    58
14         13652                 122                    64                   112
15          6625                  26                    48                    20
```

```r
mydata9
#Reverse Rank  by highest number
mydata13 = mutate_at(mydata, vars(Flights.On.Time:Flights.Delayed), funs(Rank=min_rank(desc(.))))
mydata13
```

```
32             16            23            12            69          1552            19          5415
33              8            56            14            84          2281            12         11069
34             11           104            13           228          2556             9         16704
35              1           127            12           368          2375            84          8095
36             12            43            15            14          1410             5          7698
37              4             8            10            45           587             2          3978
38              0            33            13            65          1935            12          7037
39              9            43            15           150          1991            20         11957
40              5            28            13            40          1019            28          5870
41             11            17            14            48          1187            16          7402
42              9            27            12            73          1496             5          9183
43             17            52            12            79          1527             3          9463
44              3            18            11             4           958             3          4900
45              2           556            11           457          8802            72         25565
   Flights.Total Flights.On.Time_Rank Flights.Diverted_Rank Flights.Delayed_Rank
1          34338                   3                    67                     8
2          10133                  84                   142                    88
3           8632                 109                    62                   103
4           9099                  92                    72                   118
5           7531                 121                    96                   129
6          12719                  53                   112                    53
7          29315                  13                    14                    12
8           7241                 120                    79                   138
9           5356                 145                   128                   132
10         11745                  64                    62                    50
```

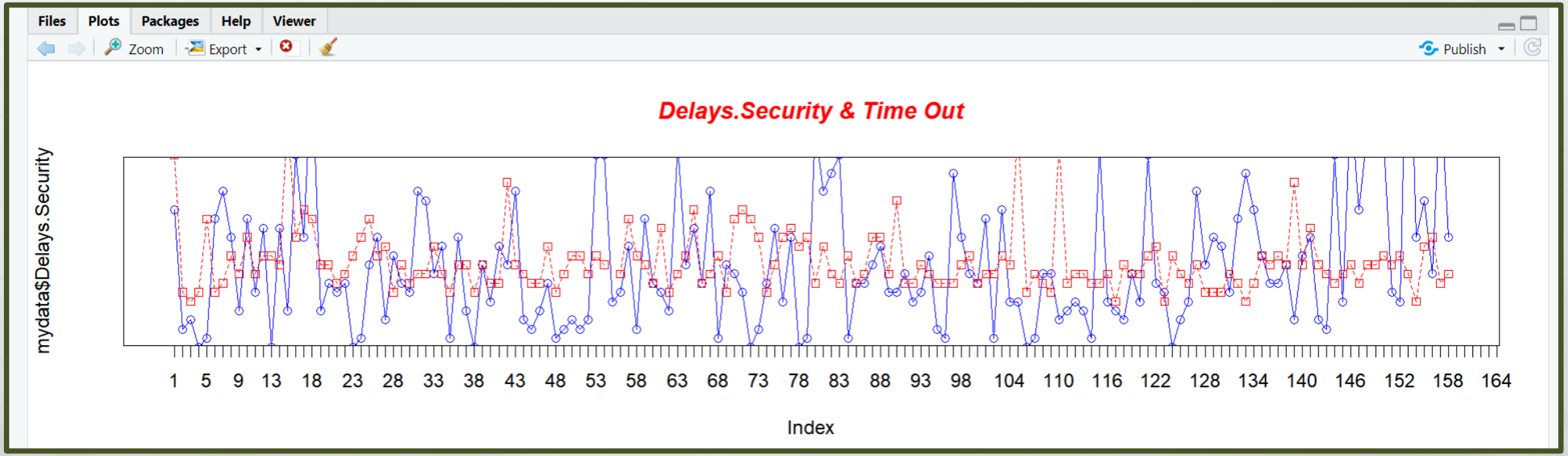
Charting

1- Plots

```
# plot
plot(mydata$Delays.Security)

# Graph Delays.Security with blue line and circle points
plot(mydata$Delays.Security, type="o", col="blue", ylim=c(1,20) , axes=FALSE, ann=TRUE)


# Graph TaxiOut with red dashed line and square points
lines(mydata$TaxiOut, type="o", pch=22, lty=2, col="red")
# Make x axis
axis(1, at=1:180)

axis(2, las=1, at=4*0:g_range[2])

# Make y axis with horizontal labels
axis(2, las=1, at=4*0:g_range[2])


# Create a title with a red, bold/italic font
title(main="Delays.Security & Time Out", col.main="red", font.main=4)

# Create a legend at (1, g_range[2]) that is slightly smaller
legend(1, g_range[2], c("Delays.Security","TaxiOut"), cex=0.8,
       col=c("blue","red"), pch=21:22, lty=1:2);

# Create box around plot
box()
```

Charting

1- Plots



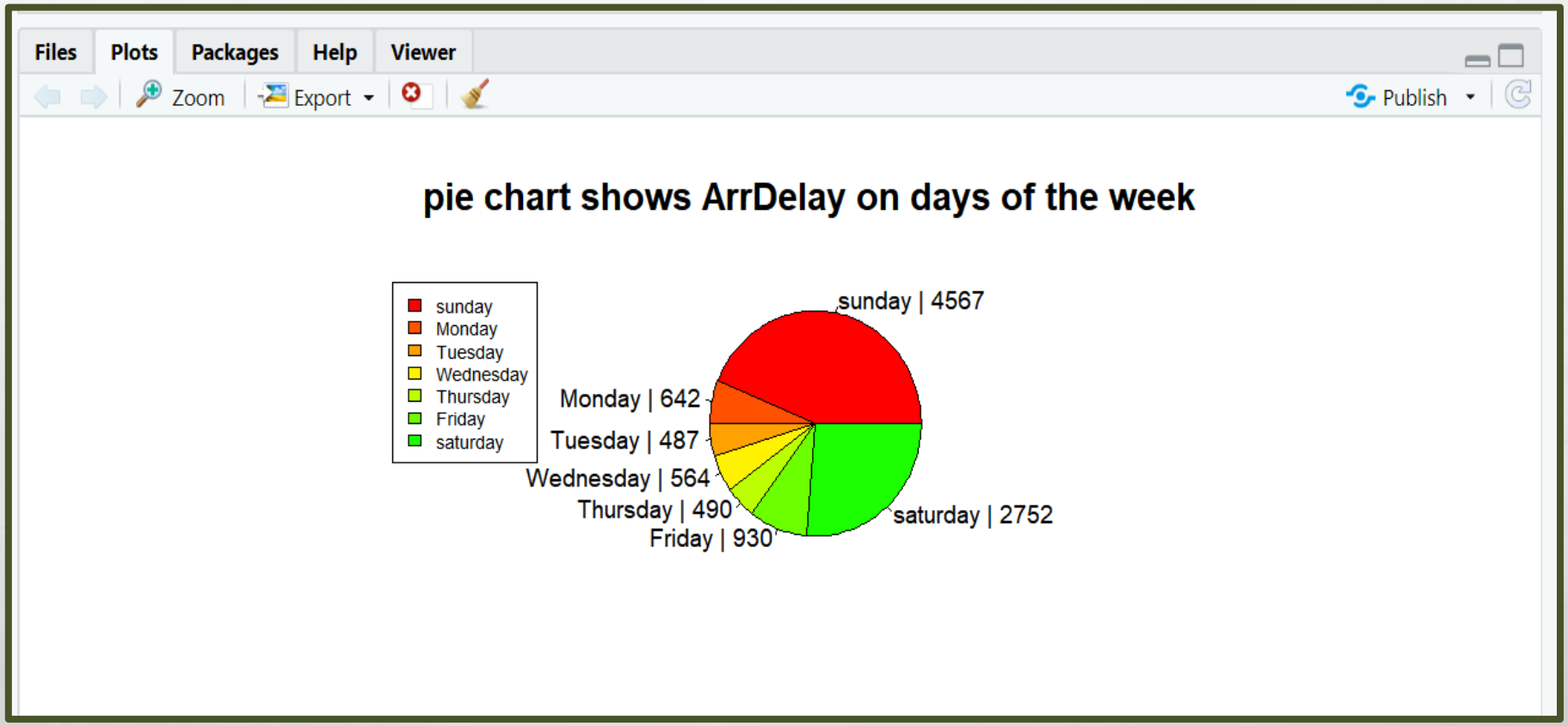
Charting

2- Pie

```r
## Selecting 7 rows
values_7<- head(mydata,7)

## Fetching Names for Labels


## Getting all the values of  ArrDelay  col
ArrDelay <- c(values_7[,11])


## Concatenation the values
v<-paste(days,"|",ArrDelay)


## create pie chart
pie(values_7$ArrDelay, main="pie chart shows ArrDelay on days of the week ",col=rainbow(length(values_7)),labels=c(v), cex=0.8)

legend(-3.2, 1.0, c(days), cex=0.6, fill=rainbow(length(values_7)))

```

Charting

2- Pie



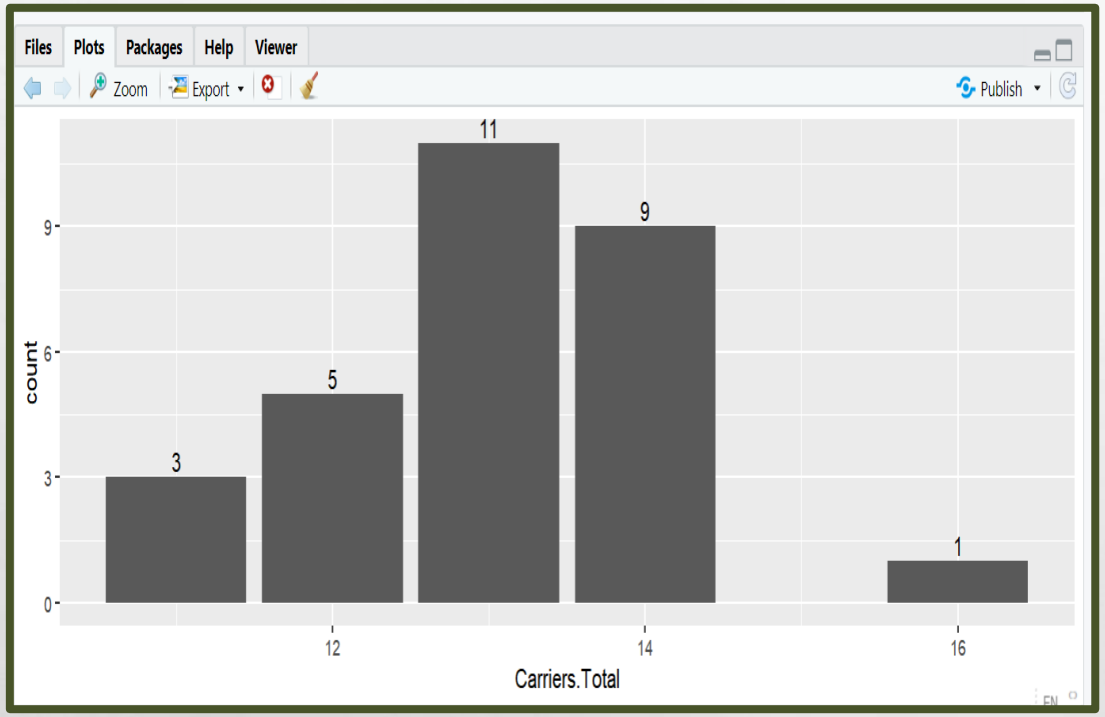
Charting

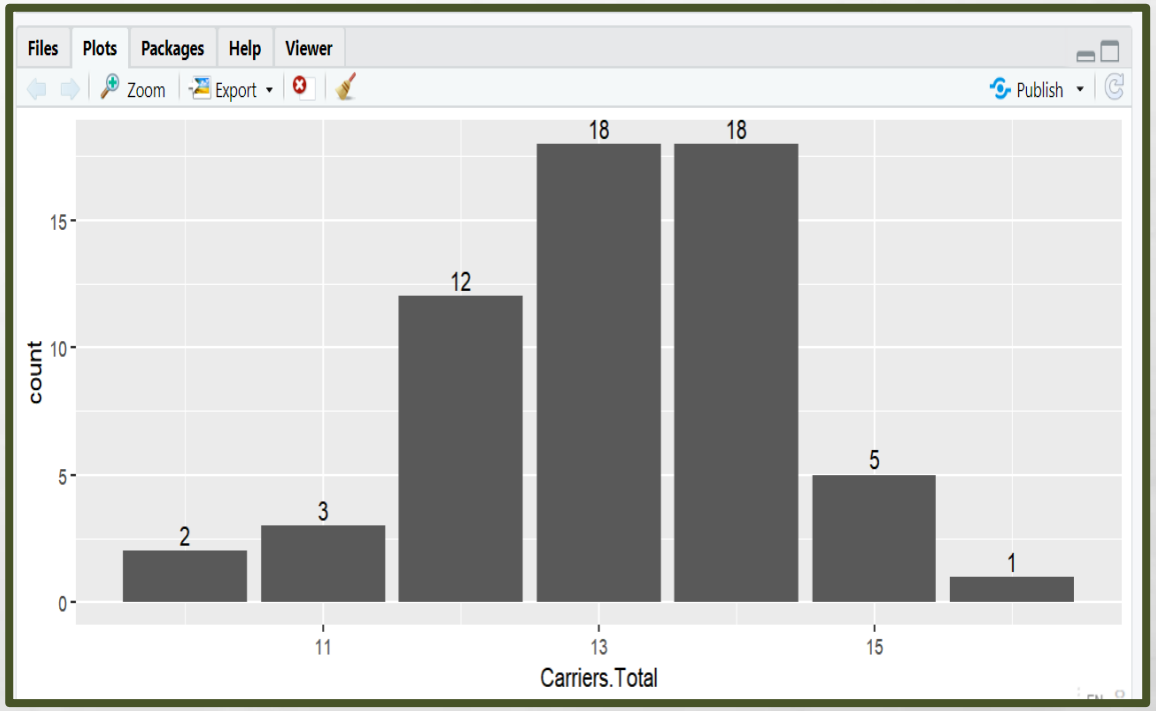
3- Bar

```
#creating a bar graph,that shows Total of carriers in my data
# with a subset of Delays.Weather>=138, and  the count.
ggplot(data= subset(mydata,Delays.Weather>=138), aes(x =Carriers.Total )) + geom_bar( )+
  geom_text(stat='count', aes(label=..count..),vjust=-0.3)


#creating a bar graph,that shows Total of carriers in my data
# with a subset of Delays.Security>=10, and  the count.
ggplot(data= subset(mydata,Delays.Security>=10), aes(x =Carriers.Total )) + geom_bar( )+
  geom_text(stat='count', aes(label=..count..),vjust=-0.3)
```

Charting

3- Bar





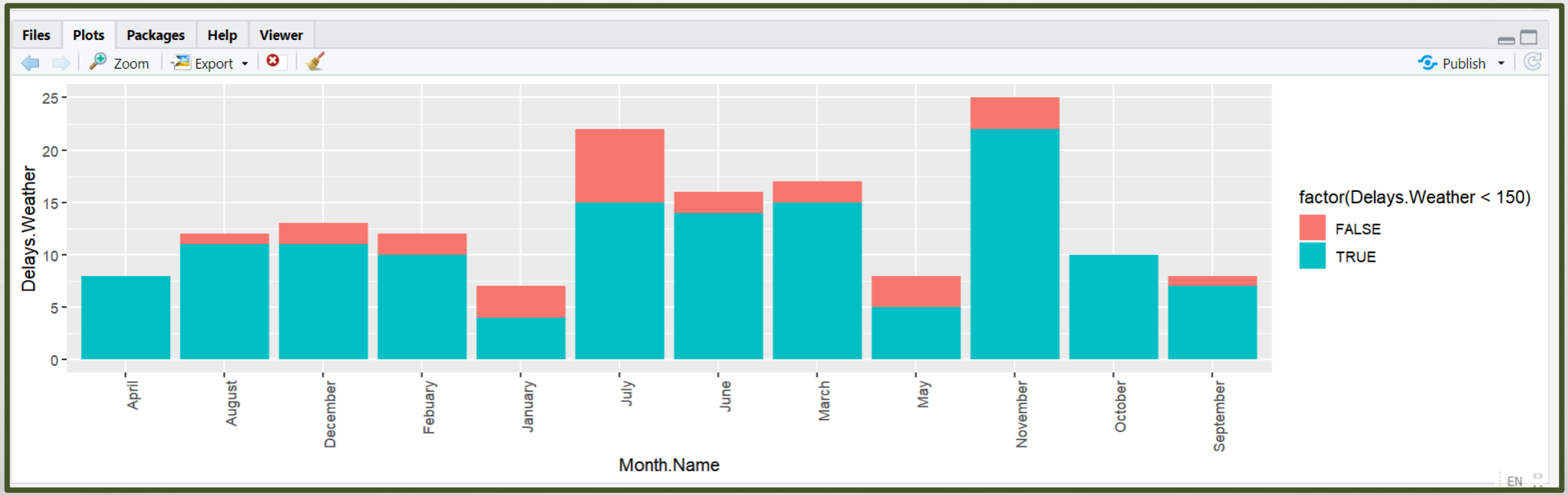
shows Total of carriers in my data with a subset of Delays.Weather>=138, and the count.

bar graph,that shows Total of carriers in my data with a subset of Delays.Security>=10, and the count.

4-stacked bar

```
#creating a stacked bar graph of the Delays.Weather and
# divided out byMonth.Name with  fill=factor
ggplot(data= mydata,aes (x = factor (Month.Name),fill =factor(Delays.Weather<150))) +
  xlab("Month.Name") +
  ylab("Delays.Weather") +
  geom_bar() + theme(axis.text.x = element_text(angle = 90, hjust = 1))
```

4-stacked bar

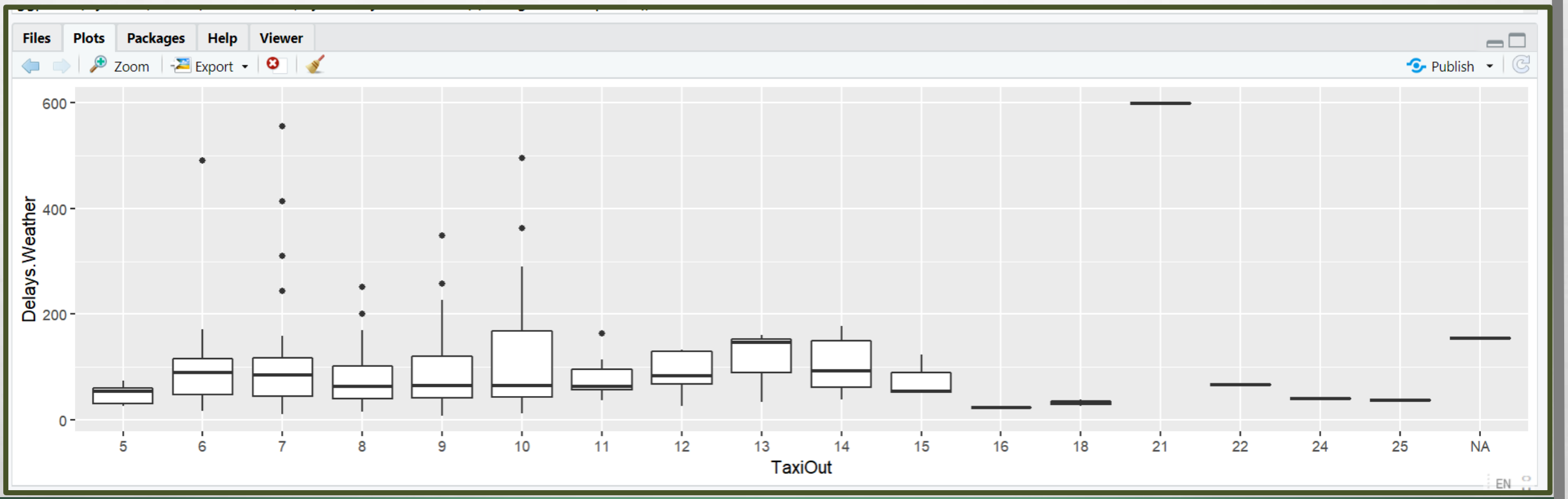


The graph Shows the Delays.Weather that<150 and divided out by Month.Name with fill=factor

5- boxplots & outlier

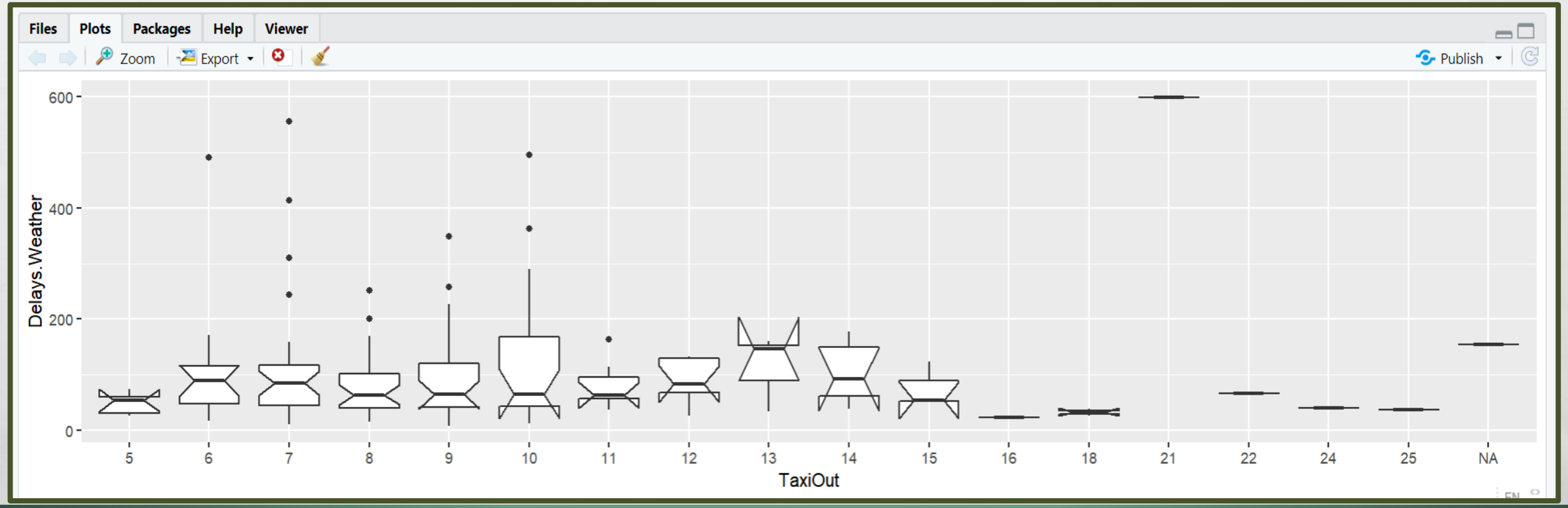
```
#To create different boxplots for Delays.Weather for different levels of x= TaxiOut
mydata$TaxiOut = factor(mydata$TaxiOut)
ggplot(mydata, aes(x=TaxiOut, y=Delays.Weather)) + geom_boxplot()

#shows outlier for Delays.Weather.
ggplot(mydata, aes(x=TaxiOut, y=Delays.Weather)) + geom_boxplot(notch = TRUE)
```

5- boxplots



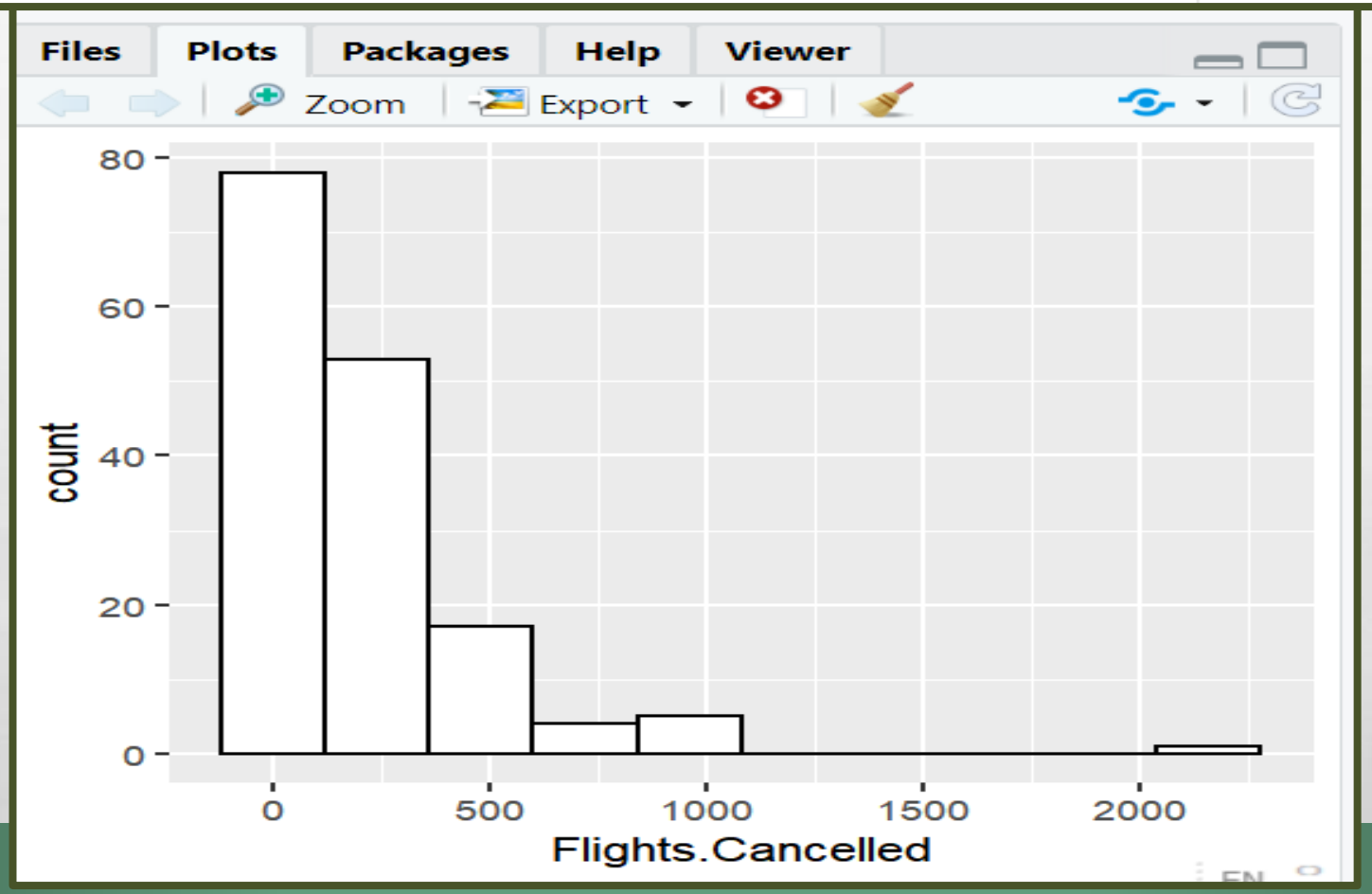
create different boxplots for Delays.Weather for different levels of x= TaxiOut

shows outlier for Delays Weather



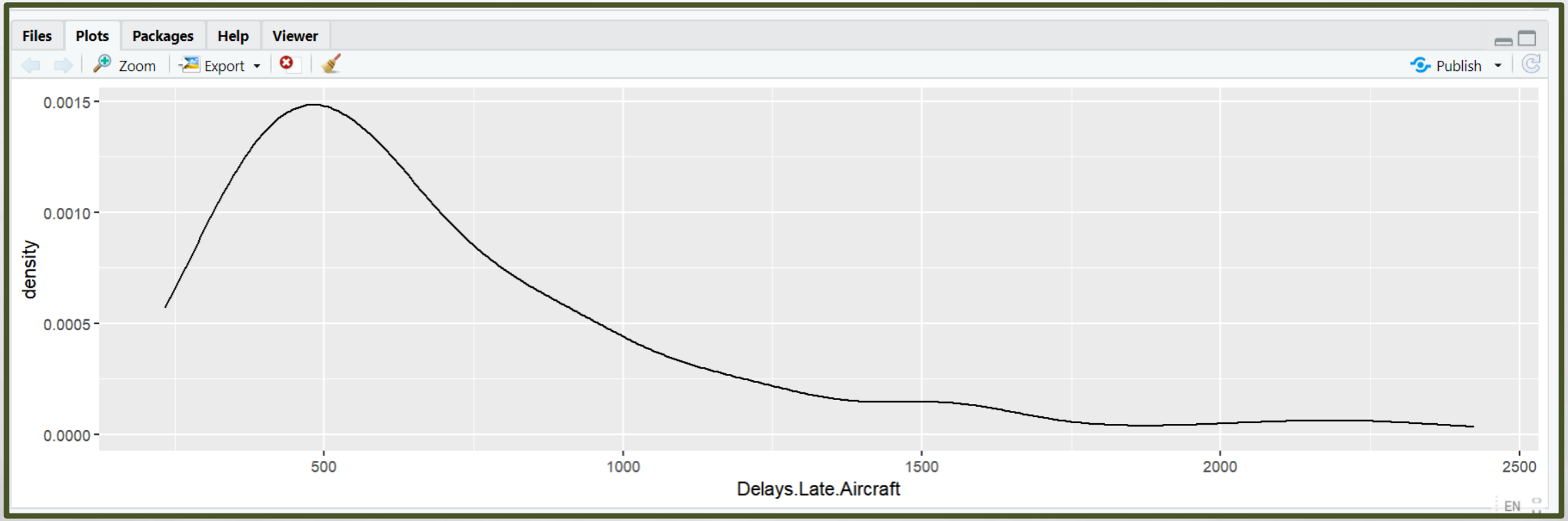
6- histogram

```
# Creating a histogram and define the number of bins
ggplot(data = mydata , aes( x = Flights.Cancelled)) + geom_histogram(color="black", fill="white", bins = 10)
```



7- Density Plot

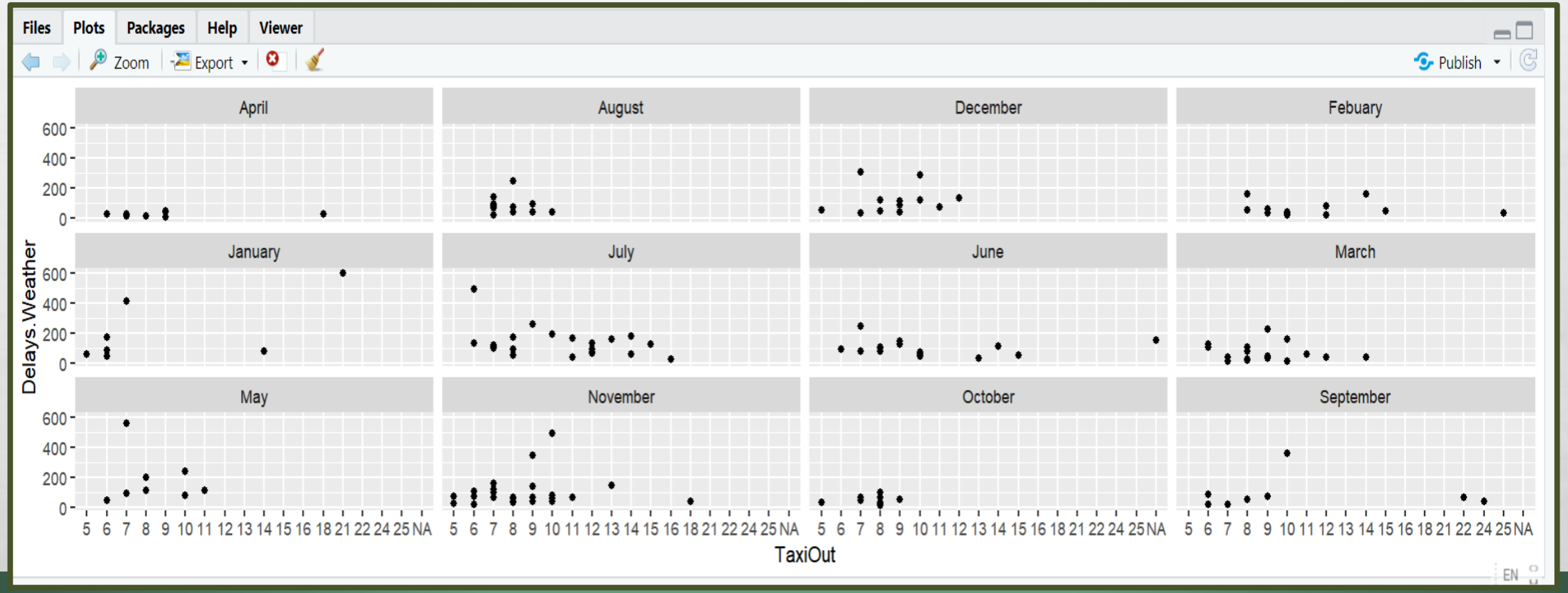
```
#Creating Density Plot to present the distribution of a Delays.Late.Aircraft.
ggplot(mydata, aes( x = Delays.Late.Aircraft)) + geom_density( )
```



presents the distribution of a Delays Late Aircraft.

8- Faceting

```
#Faceting
#Faceting for Month.Name
ggplot(mydata, aes(TaxiOut, Delays.Weather)) +  geom_point() +  facet_wrap(~Month.Name,nrow = 3)
```



ggplot(mydata, aes(TaxiOut, Delays.Weather)) + geom_point() + facet_wrap(~Month.Name,nrow = 3)

Thank you