# Data Science With R
# Project 2: Comcast Telecom Consumer Complaints

**Date : 14- 07- 2020**

## DESCRIPTION

Comcast is an American global telecommunication company. The firm has been providing terrible customer service. They continue to fall short despite repeated promises to improve. Only last month (October 2016) the authority fined them a $2.3 million, after receiving over 1000 consumer complaints.

The existing database will serve as a repository of public customer complaints filed against Comcast.
It will help to pin down what is wrong with Comcast's customer service.

### Analysis Task :

- Import data into R environment.
- Provide the trend chart for the number of complaints at monthly and daily granularity levels.
- Provide a table with the frequency of complaint types.

- Which complaint types are maximum i.e., around internet, network issues, or across any other domains.
- Create a new categorical variable with value as Open and Closed. Open & Pending is to be categorized as Open and Closed & Solved is to be categorized as Closed.
- Provide state wise status of complaints in a stacked bar chart. Use the categorized variable from Q3. Provide insights on:

- Which state has the maximum complaints
- Which state has the highest percentage of unresolved complaints
- Provide the percentage of complaints resolved till date, which were received through the Internet and customer care calls.

- The analysis results to be provided with insights wherever applicable.

---

## ANALYSIS

### # Install libraries

```
library(dplyr)

library(ggplot2)

library(stringi)

library(lubridate)

library(ggpubr)
```

### # Import Dataset

```
comcast_data<- read.csv("Comcast Telecom Complaints data.csv",header=TRUE)

comcast_data
```

# Data Science With R
# Project 2: Comcast Telecom Consumer Complaints

**Date : 14- 07- 2020**

## # For viewing structure of the data

str(comcast_data)

```
tibble [2,224 x 12] (S3: grouped_df/tbl_df/tbl/data.frame)
 $ Ticket                   : chr [1:2224] "250635" "223441" "242732" "277946
" ...
 $ CustomerComplaint        : chr [1:2224] "Comcast Cable Internet Speeds" "P
ayment disappear - service got disconnected" "Speed and Service" "Comcast I
mposed a New Usage Cap of 300GB that punishes streaming." ...
 $ Date                     : Date[1:2224], format: "2015-04-22" "2015-08-04"
...
 $ Time                     : chr [1:2224] "3:53:50 PM" "10:22:56 AM" "9:55:4
7 AM" "11:59:35 AM" ...
 $ ReceivedVia              : chr [1:2224] "Customer Care Call" "Internet" "I
nternet" "Internet" ...
 $ City                     : chr [1:2224] "Abingdon" "Acworth" "Acworth" "Ac
worth" ...
 $ State                    : chr [1:2224] "Maryland" "Georgia" "Georgia" "Ge
orgia" ...
 $ Zipcode                  : int [1:2224] 21009 30102 30101 30101 30101 3010
1 30101 49221 94502 94501 ...
 $ Status                   : chr [1:2224] "Closed" "Closed" "Closed" "Open"
...
 $ FilingonBehalfofSomeone: chr [1:2224] "No" "No" "Yes" "Yes" ...
 $ ComplaintType            : chr [1:2224] "Internet" "Others" "Others" "Othe
rs" ...
 $ ComplaintStatus          : chr [1:2224] "Closed" "Closed" "Closed" "Open"
...
 - attr(*, "groups")= tibble [77 x 3] (S3: tbl_df/tbl/data.frame)
  ..$ State          : chr [1:77] "Alabama" "Alabama" "Arizona" "Arizona" .
..
  ..$ ComplaintStatus: chr [1:77] "Closed" "Open" "Closed" "Open" ...
  ..$ .rows          : list<int> [1:77]
  .. ..$ : int [1:17] 136 599 745 909 910 911 912 913 1191 1309 ...
  .. ..$ : int [1:9] 493 694 746 762 801 914 1190 1361 2081
  .. ..$ : int [1:14] 1488 2058 2059 2060 2061 2062 2064 2065 2066 2067 ...

  .. ..$ : int [1:6] 2063 2068 2072 2073 2074 2076
  .. ..$ : int [1:6] 950 1152 1153 1154 1155 1432
  .. ..$ : int [1:159] 9 44 45 54 56 194 207 216 218 347 ...
  .. ..$ : int [1:61] 10 217 277 329 349 360 362 451 514 515 ...

  .. ..$ : int [1:58] 69 148 149 150 263 264 286 301 303 306 ...
  .. ..$ : int [1:22] 157 265 266 287 302 304 305 460 464 576 ...
  .. ..$ : int [1:9] 275 290 1192 1287 1407 1408 2138 2139 2156
  .. ..$ : int [1:3] 1224 1402 1887
  .. ..$ : int [1:8] 1290 1413 1864 1907 2181 2182 2183 2184
  .. ..$ : int [1:4] 602 1289 1300 2185
  .. ..$ : int 2123
  .. ..$ : int [1:14] 2117 2118 2119 2120 2121 2124 2125 2126 2127 2128 ...
  .. ..$ : int [1:2] 2122 2132
  .. ..$ : int [1:201] 135 241 242 243 244 247 248 269 452 453 ...
  .. ..$ : int [1:39] 245 246 249 250 251 551 561 711 742 798 ...
  .. ..$ : int [1:208] 2 3 5 6 31 32 33 34 35 36 ...
  .. ..$ : int [1:80] 4 7 83 85 89 90 91 97 107 113 ...
  .. ..$ : int [1:135] 67 68 189 211 212 231 232 240 267 268 ...
  .. ..$ : int [1:29] 26 151 152 153 186 233 234 318 367 402 ...
  .. ..$ : int [1:50] 17 198 235 236 237 473 474 499 621 630 ...
  .. ..$ : int [1:9] 137 184 648 924 929 934 1160 1236 2158
  .. ..$ : int 2143
  .. ..$ : int 1467
  .. ..$ : int 1468
  .. ..$ : int [1:4] 364 698 1498 1499
  .. ..$ : int [1:3] 647 1500 1651
  .. ..$ : int [1:12] 1260 1317 1318 1319 1894 1895 1896 1897 1898 2159 ...
  .. ..$ : int 2162
  .. ..$ : int [1:3] 309 1042 1926
  .. ..$ : int [1:2] 190 2050
  .. ..$ : int [1:63] 1 51 165 167 169 171 172 173 174 175 ...
  .. ..$ : int [1:15] 166 168 170 176 177 446 727 744 749 756 ...
  .. ..$ : int [1:50] 28 252 253 255 256 257 258 259 260 294 ...
  .. ..$ : int [1:11] 41 158 271 325 533 682 981 1405 1649 1650 ...
  .. ..$ : int [1:92] 8 46 47 48 49 50 202 208 288 289 ...
  .. ..$ : int [1:23] 138 228 443 591 593 751 791 812 899 1106 ...
  .. ..$ : int [1:29] 43 53 238 239 315 316 376 492 634 862 ...
  .. ..$ : int [1:4] 636 1738 1971 1986
  .. ..$ : int [1:23] 272 273 274 344 816 817 819 830 846 863 ...
  .. ..$ : int [1:16] 818 831 864 948 949 1259 1471 1474 1523 1524 ...
  .. ..$ : int [1:3] 917 1086 1231
  .. ..$ : int 209
  .. ..$ : int 270
  .. ..$ : int 1113
  .. ..$ : int [1:8] 478 802 1126 1199 1200 1258 1740 2188
  .. ..$ : int [1:4] 42 603 604 656
  .. ..$ : int [1:56] 185 224 225 227 279 281 282 283 285 352 ...
  .. ..$ : int [1:19] 192 193 254 280 584 628 633 799 985 988 ...
  .. ..$ : int [1:11] 11 12 14 15 1111 1112 1829 1830 1832 1833 ...
  .. ..$ : int [1:4] 13 16 1173 1831
  .. ..$ : int [1:6] 164 734 1261 1602 1646 1916
  .. ..$ : int [1:3] 1652 1867 1954
  .. ..$ : int [1:3] 472 631 2113
  .. ..$ : int [1:36] 29 348 488 489 660 661 662 663 664 665 ...
  .. ..$ : int [1:13] 30 196 197 490 491 786 808 1015 1128 1618 ...
  .. ..$ : int [1:110] 52 140 159 162 163 195 199 213 214 215 ...
  .. ..$ : int [1:20] 27 39 40 480 607 805 811 815 944 1034 ...
  .. ..$ : int 1720
  .. ..$ : int [1:15] 368 369 370 371 374 770 771 995 1341 1342 ...
  .. ..$ : int [1:3] 372 373 1425
  .. ..$ : int [1:96] 55 71 187 188 278 331 456 457 483 484 ...
  .. ..$ : int [1:47] 486 549 623 624 720 721 723 724 758 759 ...
  .. ..$ : int [1:49] 38 518 539 754 865 866 867 868 869 870 ...
  .. ..$ : int [1:22] 156 874 875 880 881 882 885 886 889 890 ...

  .. ..$ : int [1:16] 494 1165 1204 1746 1747 1748 1749 1751 1752 1753 ...
  .. ..$ : int [1:6] 847 1360 1750 1810 1905 2137
  .. ..$ : int [1:2] 313 1301
  .. ..$ : int 1735
  .. ..$ : int [1:49] 18 19 20 21 22 23 25 57 58 59 ...
  .. ..$ : int [1:11] 24 62 66 383 697 1664 1672 1673 1675 1677 ...
  .. ..$ : int [1:75] 142 143 160 203 204 205 206 226 261 262 ...
  .. ..$ : int [1:23] 141 161 191 619 620 635 674 688 1017 1030 ...

  .. ..$ : int [1:8] 314 681 907 1229 1230 1328 1656 2146
  .. ..$ : int [1:3] 220 908 1327
  .. ..@ ptype: int(0)
  ..- attr(*, ".drop")= logi TRUE
```

# Data Science With R
# Project 2: Comcast Telecom Consumer Complaints

**Date : 14- 07- 2020**

## # Summary of the data

summary(comcast_data)

```
    Ticket          CustomerComplaint       Date           
 Length:2224        Length:2224        Min.   :2015-01-04  
 Class :character   Class :character   1st Qu.:2015-05-06  
 Mode  :character   Mode  :character   Median :2015-06-20  
                                       Mean   :2015-06-09  
                                       3rd Qu.:2015-06-25  
                                       Max.   :2015-12-06  
     Time           ReceivedVia            City          
 Length:2224        Length:2224        Length:2224       
 Class :character   Class :character   Class :character  
 Mode  :character   Mode  :character   Mode  :character  



    State              Zipcode         Status         
 Length:2224        Min.   : 1075   Length:2224       
 Class :character   1st Qu.:30057   Class :character  
 Mode  :character   Median :37211   Mode  :character  
                    Mean   :47994                     
                    3rd Qu.:77059                     
                    Max.   :99223                     
 FilingonBehalfofSomeone ComplaintType      ComplaintStatus   
 Length:2224             Length:2224        Length:2224       
 Class :character        Class :character   Class :character  
 Mode  :character        Mode  :character   Mode  :character  
```

## # manipulating Columns names

names(comcast_data) <- stri_replace_all(regex= "\\.", replacement = " ", str = names(comcast_data))

head(comcast_data)

```
# A tibble: 6 x 12
# Groups:   State, ComplaintStatus [3]
  Ticket CustomerComplai~ Date       Time  ReceivedVia City  State Zipcode
  <chr>  <chr>            <date>     <chr> <chr>       <chr> <chr>   <int>
1 250635 Comcast Cable I~ 2015-04-22 3:53~ Customer C~ Abin~ Mary~   21009
2 223441 Payment disappe~ 2015-08-04 10:2~ Internet    Acwo~ Geor~   30102
3 242732 Speed and Servi~ 2015-04-18 9:55~ Internet    Acwo~ Geor~   30101
4 277946 Comcast Imposed~ 2015-07-05 11:5~ Internet    Acwo~ Geor~   30101
5 307175 Comcast not wor~ 2015-05-26 1:25~ Internet    Acwo~ Geor~   30101
6 338519 ISP Charging fo~ 2015-12-06 9:59~ Internet    Acwo~ Geor~   30101
# ... with 4 more variables: Status <chr>, FilingonBehalfofSomeone <chr>,
#   ComplaintType <chr>, ComplaintStatus <chr>
```

**Now , Comcast dataset  is loaded into R, it is available to process further.**

# Data Science With R

# Project 2: Comcast Telecom Consumer Complaints

**Date : 14- 07- 2020**

**# Finding NA's in dataset (missing number)**

```
na_vector<- is.na(comcast_data)

length(na_vector[na_vector==TRUE])
```

```
[1] 0
```

- Here we identify there is no any missing number in data.

**# Processing Date**

```
comcast_data$Date<- dmy(comcast_data$Date)

comcast_data
```

```
# A tibble: 2,224 x 12
# Groups:   State, ComplaintStatus [77]
   Ticket CustomerComplai~ Date       Time  ReceivedVia City  State
   <chr>  <chr>            <date>     <chr> <chr>       <chr> <chr>
 1 250635 Comcast Cable I~ 2015-04-22 3:53~ Customer C~ Abin~ Mary~
 2 223441 Payment disappe~ 2015-08-04 10:2~ Internet    Acwo~ Geor~
 3 242732 Speed and Servi~ 2015-04-18 9:55~ Internet    Acwo~ Geor~
 4 277946 Comcast Imposed~ 2015-07-05 11:5~ Internet    Acwo~ Geor~
 5 307175 Comcast not wor~ 2015-05-26 1:25~ Internet    Acwo~ Geor~
 6 338519 ISP Charging fo~ 2015-12-06 9:59~ Internet    Acwo~ Geor~
 7 361148 Throttling serv~ 2015-06-24 10:1~ Customer C~ Acwo~ Geor~
 8 359792 Comcast refuses~ 2015-06-23 6:56~ Internet    Adri~ Mich~
 9 318072 Comcast extende~ 2015-01-06 11:4~ Customer C~ Alam~ Cali~
10 371214 Comcast Raising~ 2015-06-28 6:46~ Customer C~ Alam~ Cali~
# ... with 2,214 more rows, and 5 more variables: Zipcode <int>,
#   Status <chr>, FilingonBehalfofSomeone <chr>, ComplaintType <chr>,
#   ComplaintStatus <chr>
```

**# Extracting monthly ticket count and daily ticket count**

```
monthly_count<-summarise(group_by(comcast_data,Month=as.integer(month(Date))),Count=n())

monthly_count

daily_count<-summarise(group_by(comcast_data,Date),Count=n())

daily_count

monthly_count<-arrange(monthly_count,Month)

monthly_count
```

# Data Science With R
# Project 2: Comcast Telecom Consumer Complaints

**Date : 14- 07- 2020**

**monthly_count**

```
# A tibble: 12 x 2
   Month Count
   <int> <int>
 1     1    55
 2     2    59
 3     3    45
 4     4   375
 5     5   317
 6     6  1046
 7     7    49
 8     8    67
 9     9    55
10    10    53
11    11    38
12    12    65
```

**Daily_count**

```
# A tibble: 91 x 2

   Date       Count
   <date>     <int>
 1 2015-01-04    18
 2 2015-01-05    12
 3 2015-01-06    25
 4 2015-02-04    27
 5 2015-02-05     7
 6 2015-02-06    25
 7 2015-03-04    15
 8 2015-03-05     5
 9 2015-03-06    25
10 2015-04-04    12
# ... with 81 more rows
```

**monthly_count**

```
# A tibble: 12 x 2
   Month Count
   <int> <int>
 1     1    55
 2     2    59
 3     3    45
 4     4   375
 5     5   317
 6     6  1046
 7     7    49
 8     8    67
 9     9    55
10    10    53
11    11    38
12    12    65
```

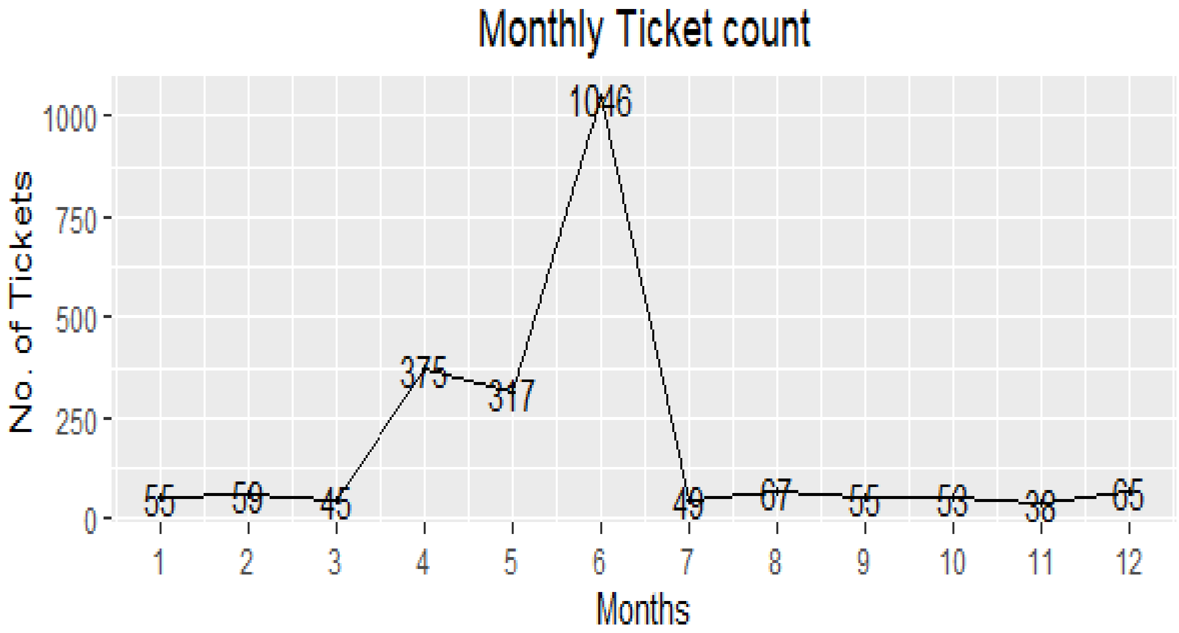
**Data Science With R**

**Project 2: Comcast Telecom Consumer Complaints**

**Date : 14- 07- 2020**

```
## Comparing Monthly complaints and Daily complaints

# Monthly complaints
```

```
ggplot(data = monthly_count,aes(Month,Count,label= Count))+
 geom_line()+
 geom_point(size= 0.6)+
 geom_text()+
 scale_x_continuous(breaks = monthly_count$Month)+
 labs(title= "Monthly Ticket count", x= "Months",y="No. of Tickets")+
 theme(plot.title = element_text(hjust = 0.5))
```



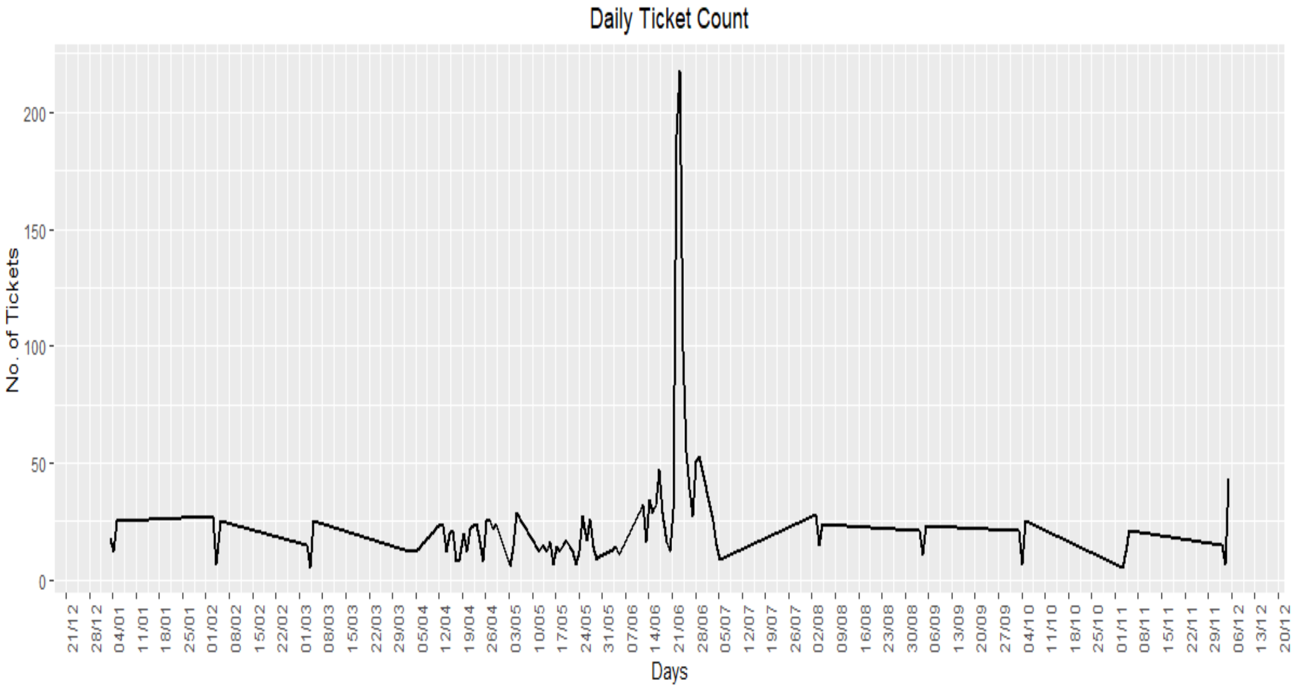
- As we can see in the month of April and May number of tickets are increases but in the month of June it increases more.
- So, there may be some reason for which they received high amount of tickets.

# Data Science With R
# Project 2: Comcast Telecom Consumer Complaints

**Date : 14- 07- 2020**

## # Daily complaints

```
ggplot(data= daily_count,aes(as.POSIXct(Date),Count))+
 geom_line()+
 geom_line(size= 1)+
 scale_x_datetime(breaks = "1 weeks",date_labels= "%d/%m")+
 labs(title = "Daily Ticket Count",x = "Days", y="No. of Tickets")+
 theme(axis.text.x = element_text(angle= 90),
  plot.title = element_text(hjust= 0.5))
```



- Here in this graph we observe that, after the middle of June the number of tickets increases with respect to other days.

# Data Science With R
# Project 2: Comcast Telecom Consumer Complaints

**Date : 14- 07- 2020**

**## Which complaint types are maximum i.e., around internet, network issues, or across any other domains.**

**# Complaint  Type Processing**

```
network_tickets<- contains(comcast_data$CustomerComplaint,match= 'network',ignore.case = TRUE)

internet_tickets<- contains(comcast_data$CustomerComplaint,match= 'internet',ignore.case= TRUE)

billing_tickets<- contains(comcast_data$CustomerComplaint,match= 'bill',ignore.case = TRUE)

email_tickets<- contains(comcast_data$CustomerComplaint,match= 'email',ignore.case= TRUE)

charges_tickets<- contains(comcast_data$CustomerComplaint,match= 'charge',ignore.case = TRUE)


comcast_data$ComplaintType[network_tickets]<- "Network"

comcast_data$ComplaintType[internet_tickets]<- "Internet"

comcast_data$ComplaintType[billing_tickets]<- "Billing"

comcast_data$ComplaintType[email_tickets]<- "Email"

comcast_data$ComplaintType[charges_tickets]<- "Charges"


comcast_data$ComplaintType[-c(network_tickets,internet_tickets,billing_tickets,

                      email_tickets,charges_tickets)]<- "Others"


table(comcast_data$ComplaintType)
```

**table(comcast_data$ComplaintType)**

```
Billing  Charges    Email Internet  Network   Others
     363      139       16      472        1     1233
```

- Here from the above result, we can observe that there are different types of complaints like internet , network issue etc. for processing further.

- So , we can take these types of complaints from different categories and combine them into one category i.e. Others.

- The above result we can say that most of the complaints are of internet issue.

# Data Science With R
# Project 2: Comcast Telecom Consumer Complaints

**Date : 14- 07- 2020**

**## Create a new categorical variable with value as Open and Closed. Open & Pending is to be categorized as Open and Closed & Solved is to be categorized as Closed.**

**# Creating new variables Complaint Status with values Open and Closed.**

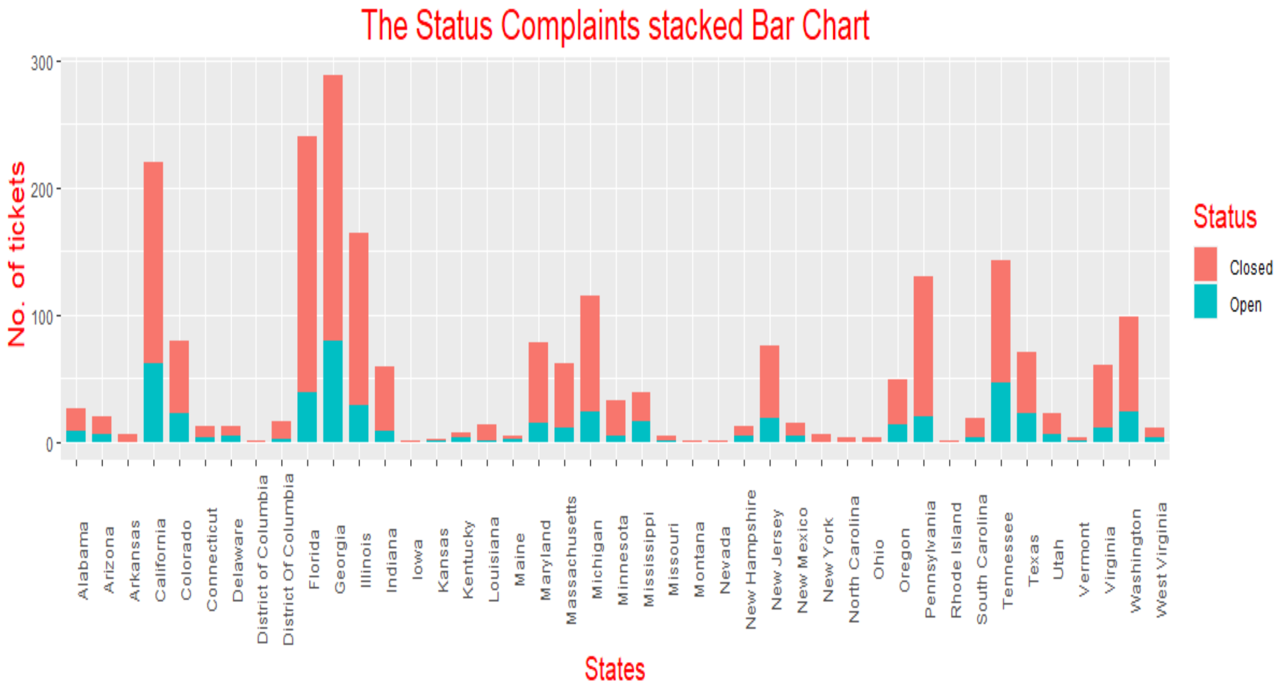
```
Open_complaints<- (comcast_data$Status=='Open' | comcast_data$Status=='Pending')

Closed_complaints<-(comcast_data$Status=='Closed' | comcast_data$Status=='Solved')

comcast_data$ComplaintStatus[Open_complaints]<- "Open"

comcast_data$ComplaintStatus[Closed_complaints]<- "Closed"
```

**# Creating Stacked barchart for complaints based on state and status**

```
comcast_data<- group_by(comcast_data,State,ComplaintStatus)

chart_data<- summarise(comcast_data,Count=n())

ggplot(as.data.frame(chart_data),mapping= aes(State,Count))+

geom_col(aes(fill= ComplaintStatus) , width=0.75)+

theme(axis.text.x = element_text(angle=90),

      axis.title.y = element_text(size= 15),

      axis.title.x = element_text(size = 15),

      title= element_text(size = 16,colour = "#FF0000"),

      plot.title= element_text(hjust=0.5))+

     labs(title ="The Status Complaints stacked Bar Chart",

     x= "States", y = " No. of tickets", fill ="Status")
```

# Data Science With R
# Project 2: Comcast Telecom Consumer Complaints

**Date : 14- 07- 2020**



- Here from the above bar chart, we can observed the Status of the States. The State ‘Georgia’ have the highest number of complaints and the State ‘Florida’ is the second highest number of complaints as compaired to the other Sates.

**# Finding State which has highest number of unresolved tickets.**

```
chart_data%>%

 filter(ComplaintStatus== "Open")-> Open_complaints

Open_complaints[Open_complaints$Count== max(Open_complaints$Count),c(1,3)]
```

```
# A tibble: 1 x 2
# Groups:   State [1]
  State    Count
  <chr>    <int>
1 Georgia     80
```

- As we can observe that State Georgia has maximum no. of unresolved Tickets and these ticket count   is 80.

**## Provide the percentage of complaints resolved till date, which were received through the Internet and customer care calls.**

**# Calculating Total Resolved and Category Resolved**

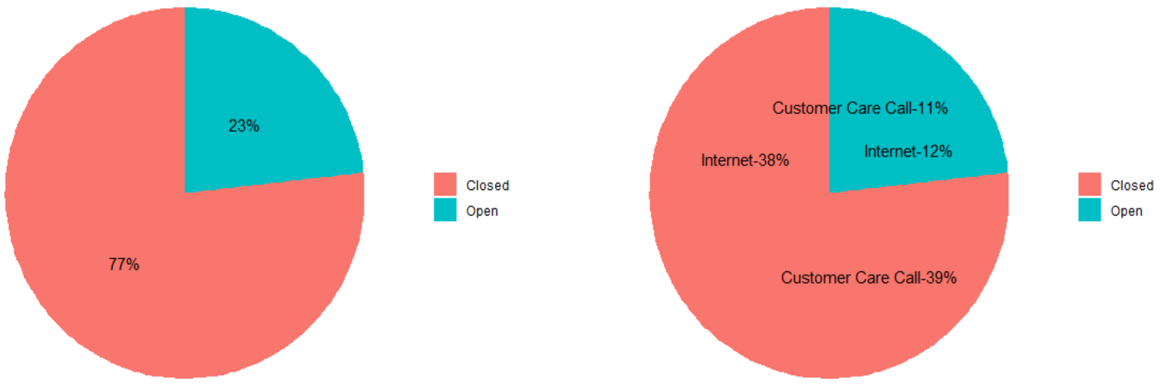
```
resolved_data<- group_by(comcast_data,ComplaintStatus)

total_resolved<- summarise(resolved_data,Percentage = (n()/nrow(resolved_data)))

resolved_data<- group_by(comcast_data,ReceivedVia,ComplaintStatus)

category_resolved<- summarise(resolved_data,Percentage = (n()/nrow(resolved_data)))
```

**## Plotting Pie Chart for Total Resolved Vs. Category Resolved**

```
par(mfrow = c(1,2))

total<- ggplot(total_resolved,

            aes(x= "" , y= Percentage, fill= ComplaintStatus))+

 geom_bar(stat= "identity" , width =1)+

 coord_polar("y", start = 0)+

 geom_text(aes(label = paste0(round(Percentage*100), "%")),

          position = position_stack(vjust= 0.5))+

 labs(x= NULL, y= NULL, fill = NULL)+

 theme_classic()+

 theme(axis.line= element_blank(),

      axis.text = element_blank(),

      axis.ticks = element_blank())
```

## Pie Chart for Category wise Ticket Status

```
category<- ggplot(category_resolved,
               aes(x="", y= Percentage, fill= ComplaintStatus))+
  geom_bar(stat= "identity", width= 1)+
  coord_polar("y", start = 0)+
  geom_text(aes(label = paste0(ReceivedVia, "-" ,round(Percentage*100), "%")),
          position = position_stack(vjust= 0.5))+
  labs(x= NULL, y = NULL , fill = NULL)+
  theme_classic()+
  theme(axis.line = element_blank(),
       axis.text = element_blank(),
       axis.ticks = element_blank())
ggarrange(total, category, nrow= 1, ncol= 2)
```



- From the above Pie Chart of the Total Resolved and Category resolved, we can conclude that, the Total Resolved Complaints are 77%, in which 38% complaints are received from Internet issue while 39% are from Customer Care Calls.

However, the Category Resolved Complaints are 23%, in which 12% are received from Internet issue while 11% are from Customer Care Calls.

# Data Science With R
# Project 2: Comcast Telecom Consumer Complaints

Date : 14- 07- 2020

## **• Insights of the Comcast Telecom Consumer Complaints Project:**

As per the above analysis, we can conclude that, we can see in the month of April and May number of tickets are increases but in the second half of the **June month** Comcast received high amount of Complaints. We can say that most of the complaints are related to **internet** issue. The highest amount of Complaints are received from State **'Georgia'** and the State 'Florida' is the second highest number of complaints as compared to the other States. As we can observe that State **Georgia** has maximum no. of unresolved Tickets and these ticket count is **80**. The **Total Resolved Complaints** are **77%**, in which **38%** complaints are received from **Internet** issue while **39%** are from **Customer Care Calls.**

However, the Category Resolved Complaints are 23%, in which 12% are received from Internet issue while 11% are from Customer Care Calls.