Project on Data Science with R Programming

Name of the project:

Comcast Telecom Consumer Complaints

Project on Data Science with R Programming

Project 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.

Data Dictionary

- Ticket #: Ticket number assigned to each complaint
- Customer Complaint: Description of complaint
- Date: Date of complaint
- Time: Time of complaint
- Received Via: Mode of communication of the complaint
- City: Customer city
- State: Customer state
- Zipcode: Customer zip
- Status: Status of complaint
- Filing on behalf of someone
-

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.

Source Code with graphs an insights:

```
#loading the file in R enviroment
tele<-read.csv('Comcast Telecom Complaints data.csv')
View(tele)
str(tele)
names(tele)
sum(is.na(tele))
```

Output:

```
> tele<-read.csv('Comcast Telecom Complaints data.csv')
> View(tele)
> str(tele)
'data.frame':   2224 obs. of  10 variables:
 $ Ticket..                : chr  "250635" "223441" "242732" "277946" ...
 $ Customer.Complaint      : chr  "Comcast Cable Internet Speeds" "Payment disappear - service got disconnected" "Speed and Service" "Comcast Imposed
a New Usage Cap of 300GB that punishes streaming." ...
 $ Date                    : chr  "22-04-2015" "4/8/2015" "18-04-2015" "5/7/2015" ...
 $ Time                    : chr  "3:53:50 PM" "10:22:56 AM" "9:55:47 AM" "11:59:35 AM" ...
 $ Received.Via            : chr  "Customer Care Call" "Internet" "Internet" "Internet" ...
 $ City                    : chr  "Abingdon" "Acworth" "Acworth" "Acworth" ...
 $ State                   : chr  "Maryland" "Georgia" "Georgia" "Georgia" ...
 $ Zip.code                : int  21009 30102 30101 30101 30101 30101 30101 49221 94502 94501 ...
 $ Status                  : chr  "Closed" "Closed" "Closed" "Open" ...
 $ Filing.on.Behalf.of.Someone: chr  "No" "No" "Yes" "Yes" ...
> names(tele)
 [1] "Ticket.."                  "Customer.Complaint"        "Date"
 [4] "Time"                      "Received.Via"              "City"
 [7] "State"                     "Zip.code"                  "Status"
[10] "Filing.on.Behalf.of.Someone"
> sum(is.na(tele))
[1] 0
>
```

Insights: This shows that there is no missing values in dataset, so now data is tidy and available to process further or do EDA based on requirement. • Processing Date.

```
#converting the required variables as factor
tele$Received.Via=as.factor(tele$Received.Via)
tele$City=as.factor(tele$City)
tele$State=as.factor(tele$State)
tele$Status=as.factor(tele$Status)

tele$Filing.on.Behalf.of.Someone=as.factor(tele$Filing.on.Behalf.of.Someone)

#converting the date column to in date format
library(lubridate)
tele$Date<-dmy(tele$Date)

#Creating the chart for the number of complaints at monthly and daily granularity levels.
library(dplyr)
monthly_data<-summarise(group_by(tele,month=as.integer(month(Date))),Count=n())
monthly_data
```

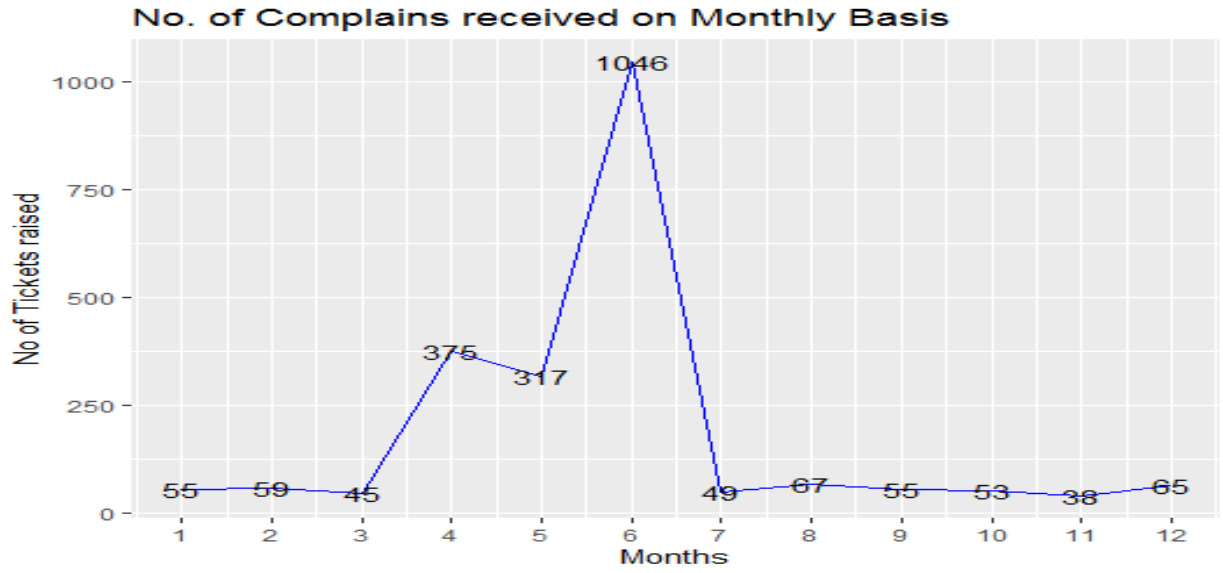
daily_data<-summarise(group_by(tele,Date),Count=n())
daily_data

Output:

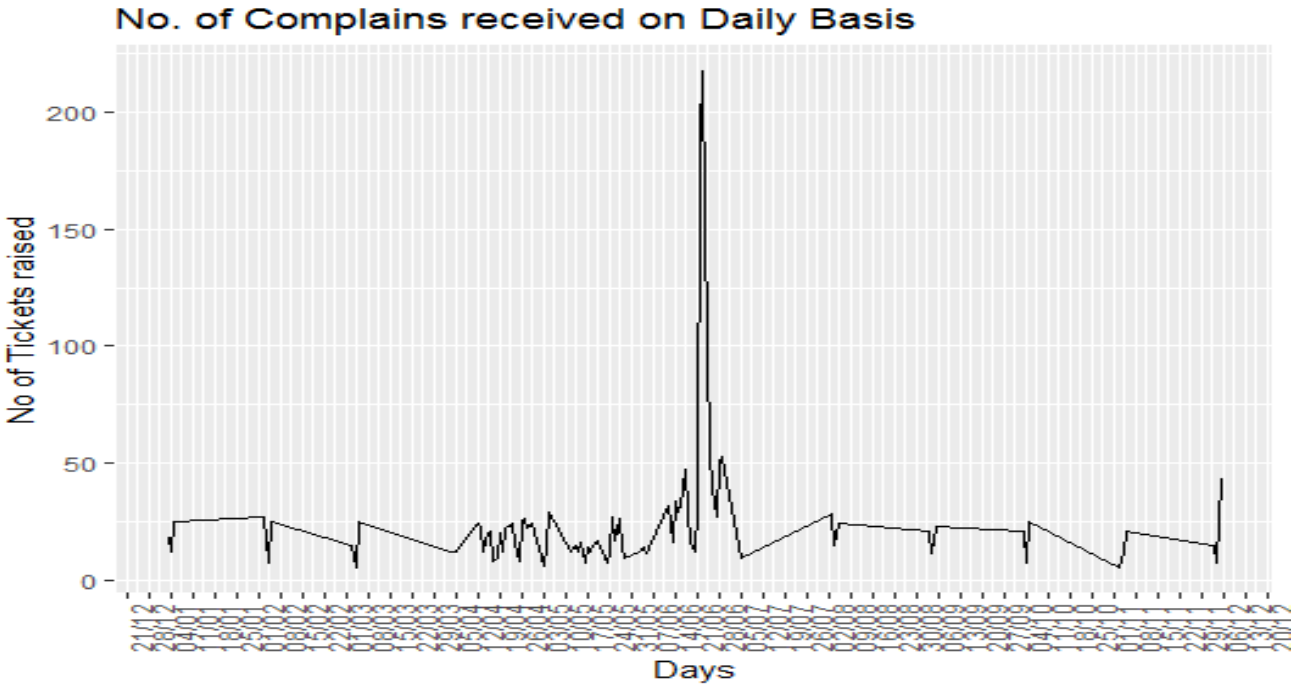
```
> library(dplyr)
> monthly_data<-summarise(group_by(tele,month=as.integer(month(Date))),Count=n())
> monthly_data
# A tibble: 12 x 2
   month Count
   <int> <int>
 1     1    55
 2     2    59
 3     3    45
 4     4   375
 5     5   317
 6     6  1046
 7     7    49
 8     8    67
 9     9    55
10    10    53
11    11    38
12    12    65
>
> daily_data<-summarise(group_by(tele,Date),Count=n())
> daily_data
# 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
```

```
library(ggplot2)

ggplot(monthly_data,aes(x=month,y=Count,label=Count)) + geom_line(colour='blue',linetype=1)+
  geom_text()+
  scale_x_continuous(breaks = monthly_data$month)+
  labs(title = "No. of Complains received on Monthly Basis",x='Months',y="No of Tickets raised")

 ggplot(daily_data,aes(x=as.POSIXct(Date),y=Count))+geom_line()+
    theme(axis.text.x  =element_text(angle=90) )+
    scale_x_datetime(breaks = "1 weeks",date_labels='%d/%m')+
    labs(title ='No. of Complains received on Daily Basis',x='Days',y="No of Tickets raised" )
```

Output as graphs:



Insights: As we can clearly see from the graph that the complaint was increased drastically in the month of June due to some reason.



Insight: With the help of the daily report chart it is clearly observed the complaint increased in the second half of the month of June.

```
 #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
library(tidyverse)
network_tickets<-contains(tele$Customer.Complaint,match='network',ignore.case = T)
internet_tickets<-contains(tele$Customer.Complaint,match ='internet',ignore.case = T)
bill_tickets<-contains(tele$Customer.Complaint,match='bill',ignore.case = T)
email_tickets<-contains(tele$Customer.Complaint,match="email",ignore.case = T)
charge_tickets<-contains(tele$Customer.Complaint,match='charge',ignore.case = T)

tele$ComplaintType[network_tickets]<-'Network'

tele$ComplaintType[internet_tickets]<-'Internet'
tele$ComplaintType[bill_tickets]<-'Billing'
tele$ComplaintType[email_tickets]<-'Email'
tele$ComplaintType[charge_tickets]<-'Charges'
tele$ComplaintType[-c(network_tickets,internet_tickets,bill_tickets,email_tickets,charge_tickets)]<-"Others"


View(tele)
table(tele$ComplaintType)
```

Output:

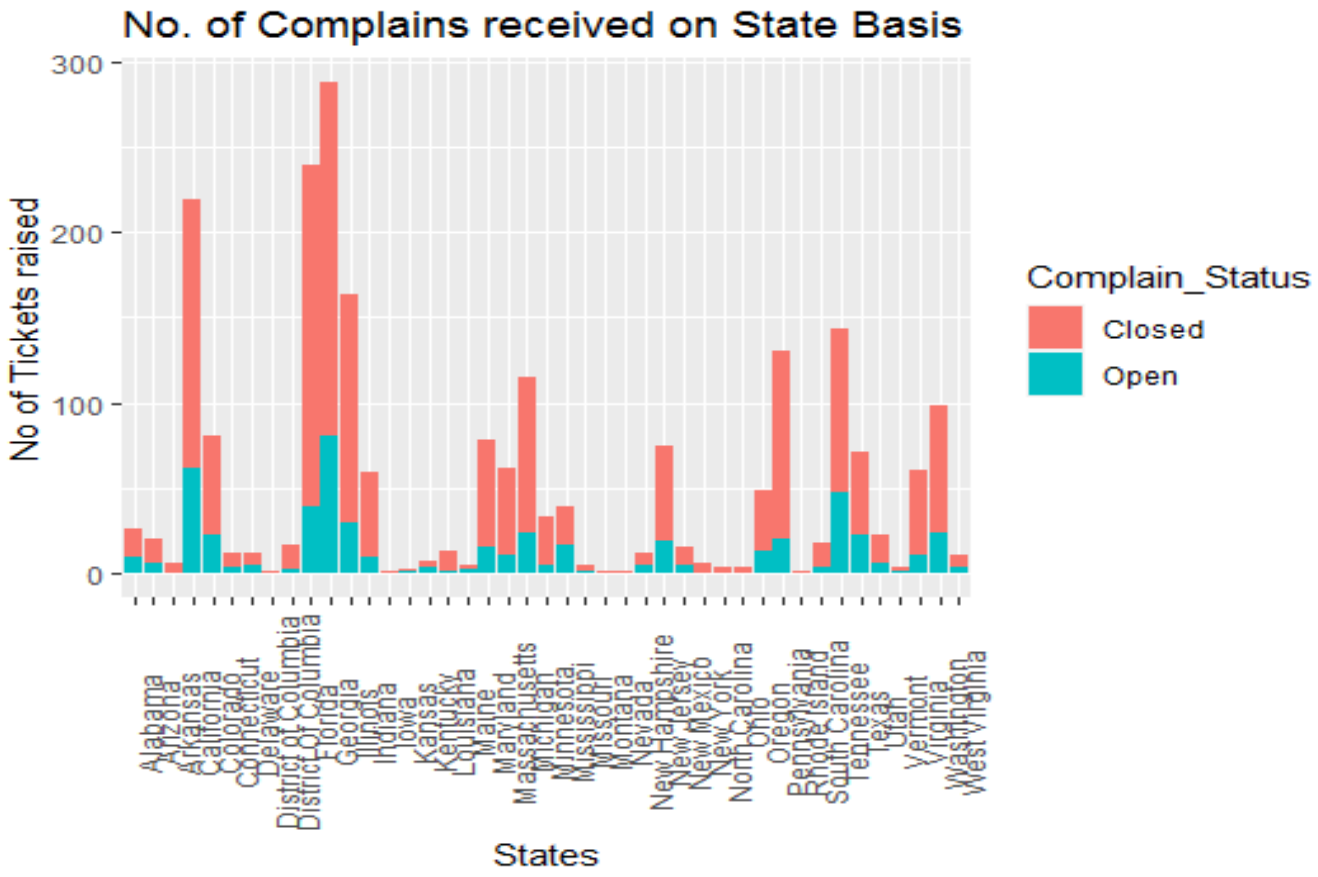
```
>
>  View(tele)
>  table(tele$ComplaintType)

 Billing  Charges    Email Internet   Others
     363      139       16      472     1233
>
```

Insight: As we can observe that there are some complaints from different-different categories and we combined them into one, i.e.- others. So most of the complaints are related to Internet issue.

```
 #creating a new column for  Open & Pending as Open and Closed & Solved as Closed.

tele$Complain_Status<-ifelse(tele$Status == "Open"| tele$Status =="Pending",'Open','Closed')

#creating the stacked bar for the no of complains received on the basis of cities

Complain<-summarise(group_by(tele,State,Complain_Status),Count=n())
ggplot(Complain,aes(x=State,y=Count,fill=Complain_Status))+geom_col()+
  theme(axis.text.x = element_text(angle = 90))+
```

labs(title ='No. of Complains received on State Basis',x='States',y="No of Tickets raised")

Output as graph:



Insight: Now it`s clearly shown that the highest number of complaints recorded from the state Georgia and the second highest number of complaints recorded from the state Florida.

#Calculating the State with highest unclosed_data cases

Complain %>% filter(Complain_Status=='Open')%>%arrange(desc(Count)) %>% head(1)

Output:

```
# Groups:   State [1]
  State   Complain_Status Count
  <fct>   <chr>           <int>
1 Georgia Open               80
> |
```

Insight: As seen from the table the maximum number of open cases is from the state of Georgia.

#Calculating the percentage of complaints closed_data till date,

#which were received through the Internet and customer care calls.
resolved_data <- group_by(tele,Complain_Status)
total_resloved<- summarise(resolved_data ,percentage =(n()/nrow(resolved_data))*100)
total_resloved
resolved_data1 <- group_by(tele,Received.Via,Complain_Status)
Category_resloved<- summarise(resolved_data1 ,percentage =(n()/nrow(resolved_data1))*100)
Category_resloved

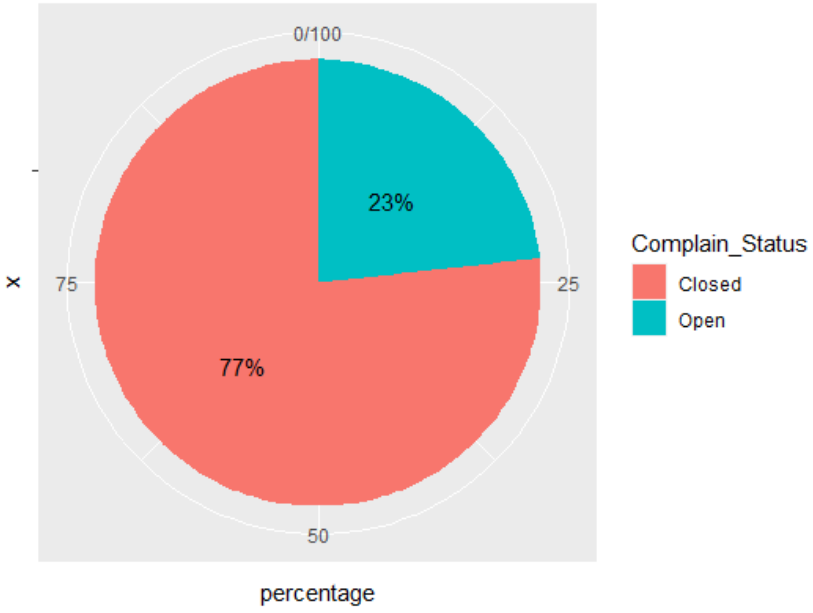
Output:

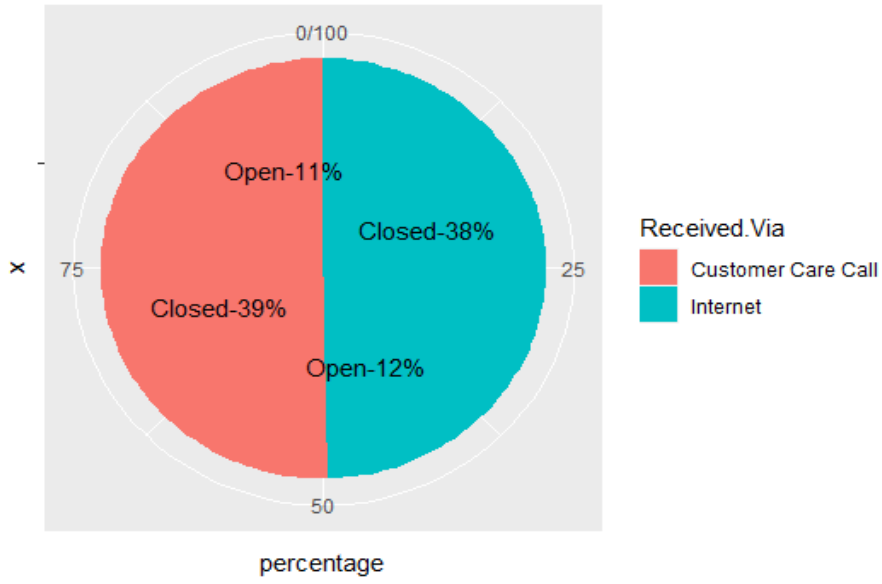
```
# A tibble: 2 x 2
  Complain_Status percentage
  <chr>                <dbl>
1 Closed                76.8
2 Open                  23.2
> resolved_data1 <- group_by(tele,Received.Via,Complain_Status)
> Category_resloved<- summarise(resolved_data1 ,percentage =(n()/nrow(resolved_data1))*100)
`summarise()` has grouped output by 'Received.Via'. You can override using the `.groups` argument.
> Category_resloved
# A tibble: 4 x 3
# Groups:   Received.Via [2]
  Received.Via        Complain_Status percentage
  <fct>               <chr>                <dbl>
1 Customer Care Call  Closed                38.8
2 Customer Care Call  Open                  11.5
3 Internet            Closed                37.9
4 Internet            Open                  11.8
>
```

Ploting Pie Chart for Total Resolved Vs Category Resolved

```
total<-ggplot(total_resloved,
              aes(x= "",y =percentage,fill = Complain_Status))+
  geom_bar(stat = "identity",width = 1)+
  coord_polar("y",start = 0)+
  geom_text(aes(label = paste0(round(percentage),"%")),
            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())
total
# Pie Chart for Category wise Ticket Status
category<-ggplot(Category_resloved,
                 aes(x= "",y =percentage,fill = Received.Via))+
  geom_bar(stat = "identity",width = 1)+
  coord_polar("y")+
  geom_text(aes(label = paste0(Complain_Status,"-",round(percentage),"%")),
            position = position_stack(vjust = 0.5))
category
```

Output as graphs:





With the help of above Chart of Total Resolved we can conclude that the total resolved complaints are 77% and open cases are 23%.From the Category Resolved the complaint received through Internet and Customer care call are same of 50%-50% in which closed case through Internet is 38% and open is 12% in case of Customer Care Call the closed cases 39% and open cases are 11%.

Overall Insights From the Project:

As per the above analysis we observe that in the 2nd half of the June mont h Comcast received high amount of complaints in which most of the compla ints are related to internet service issue and the highest amount of complai nts are received from the state Georgia. The highest unresolved complaints are related from the state Georgia and the total amount of resolved compla ints are 77% in which 38% are received the internet and 39% are from the customer care calls.