Comcast Telecom Consumer Complaints

Business Analytic Foundation with R Tools- Question

Abstract

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 and needs help to pin down what is wrong with their customer service

Presented by: Ankita Agarwal

Problem Statement:

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.

Detailed description of the given dataset:

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

To Analyze:

Import data into R environment.

1. Provide the trend chart for the number of complaints at monthly and daily granularity levels.

2. Provide a table with the frequency of complaint types.
    a. Which complaint types are maximum i.e., around internet, network issues, or across any other domains.

3. 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.
    a. Provide state wise status of complaints in a stacked bar chart. Use the categorized variable from Q3. Provide insights on:
    b. Which state has the maximum complaints
    c. Which state has the highest percentage of unresolved complaints

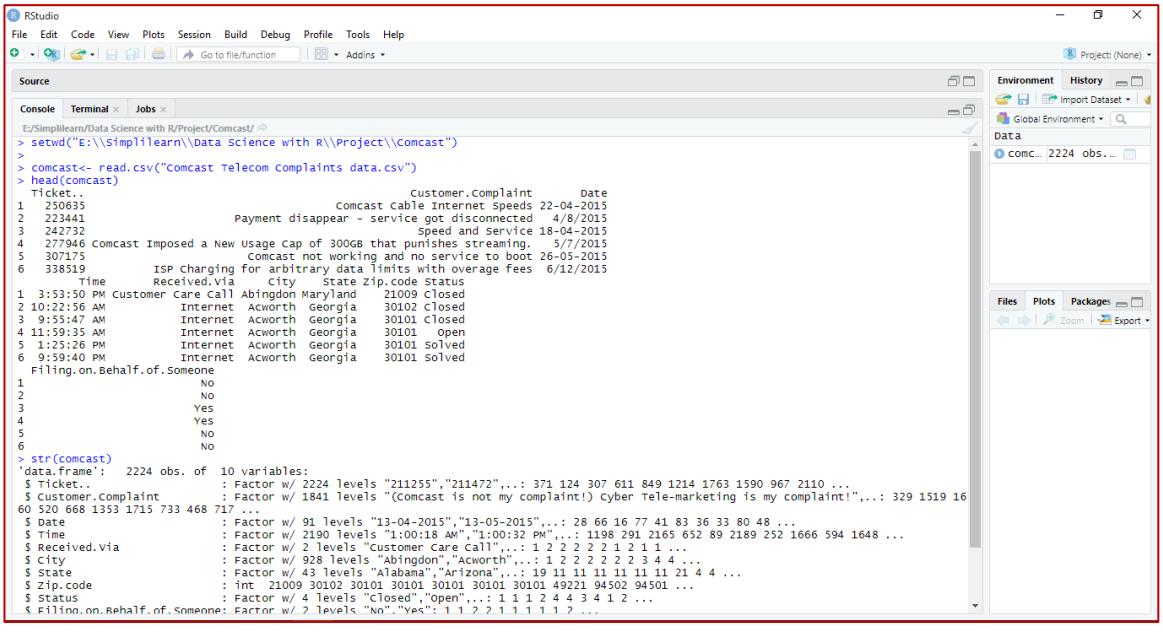
4. Provide the percentage of complaints resolved till date, which were received through the Internet and customer care calls.

The analysis results to be provided with insights wherever applicable.

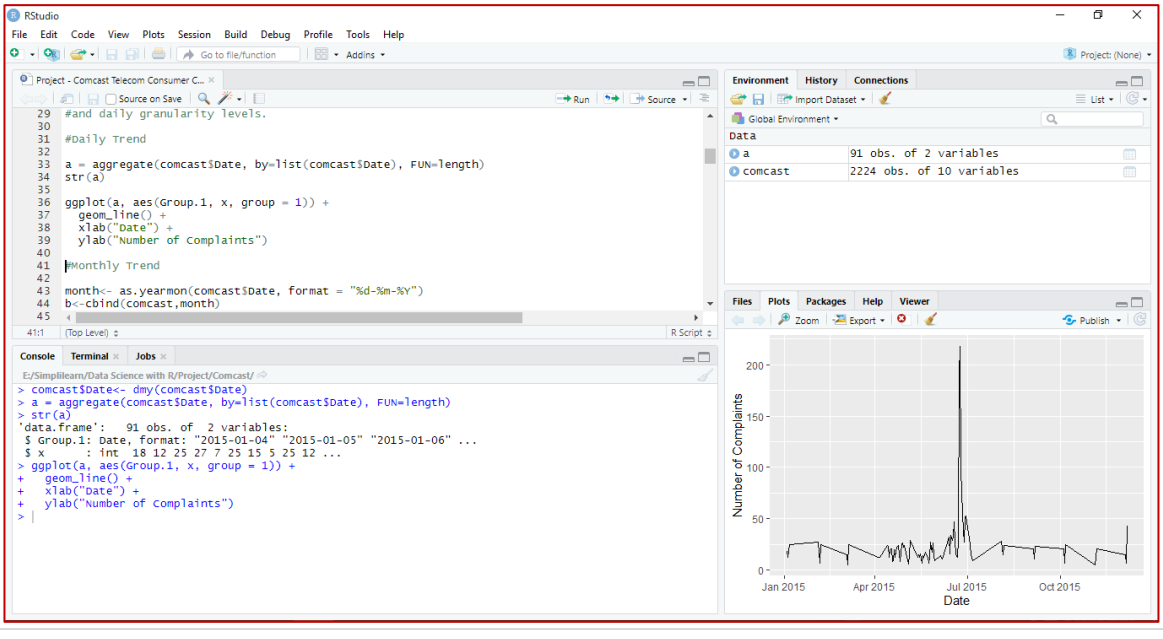
Analysis and Interpretations:

1. Provide the trend chart for the number of complaints at monthly and daily granularity levels.

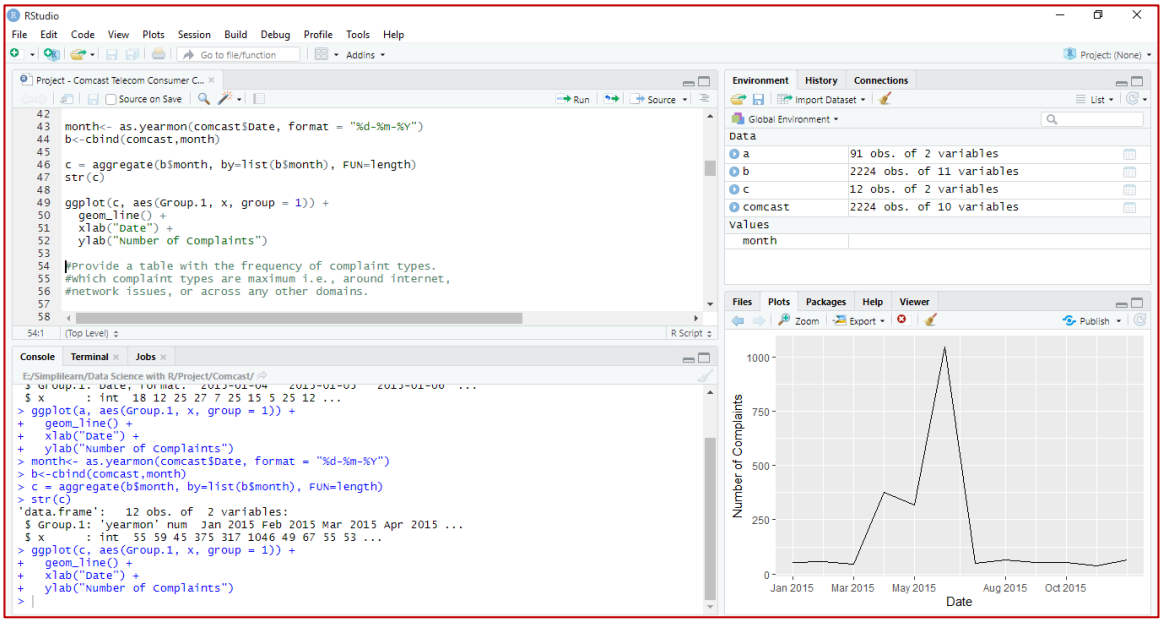
As per the business problem, I have loaded the dataset to the environment and did a basic study of it to exclude any null values from the dataset.

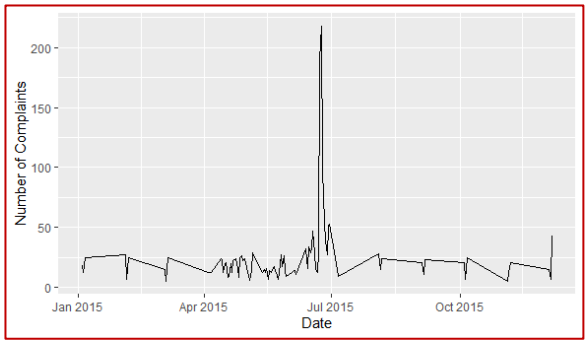


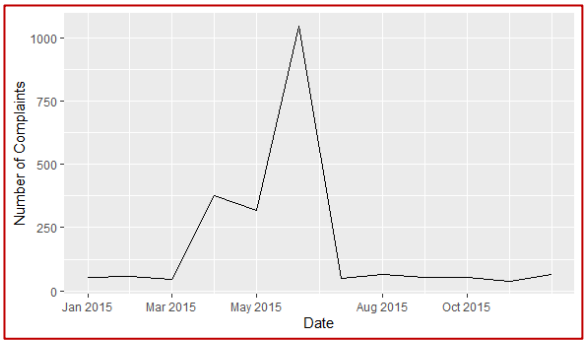
For Daily Trend, since the date in the dataset was in varied format, it was first changed into a common readable format. Then using the aggregate, we found the frequency of complaints on daily basis.



For Monthly Trend, the date was converted into months. Then using the aggregate, we found the frequency of complaints on monthly basis.







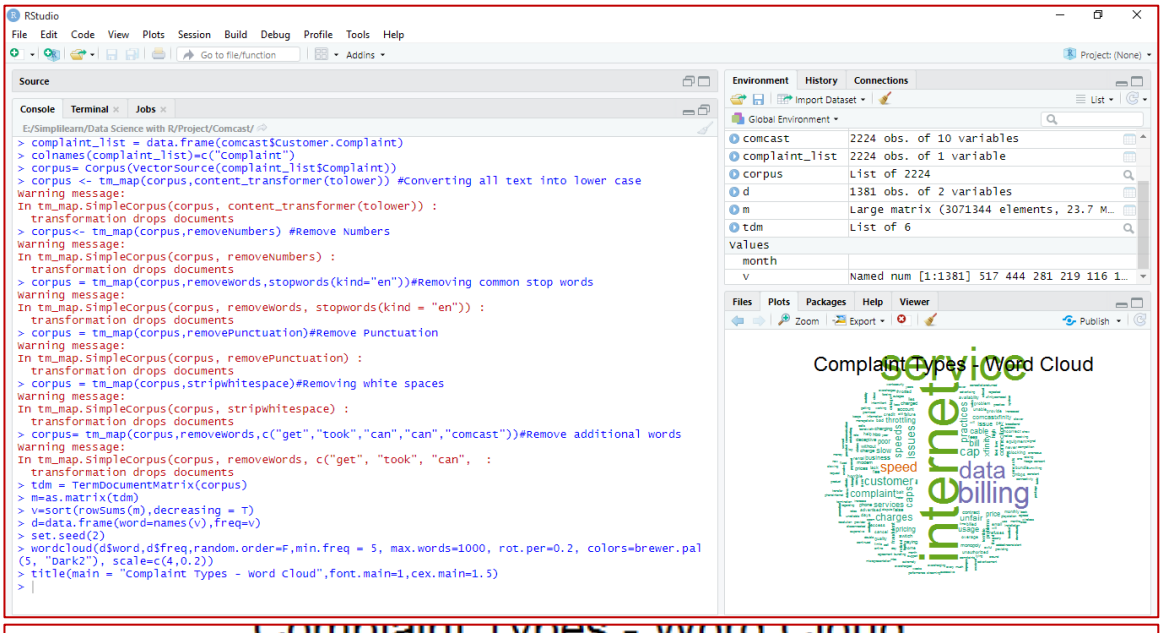
Interpretation:

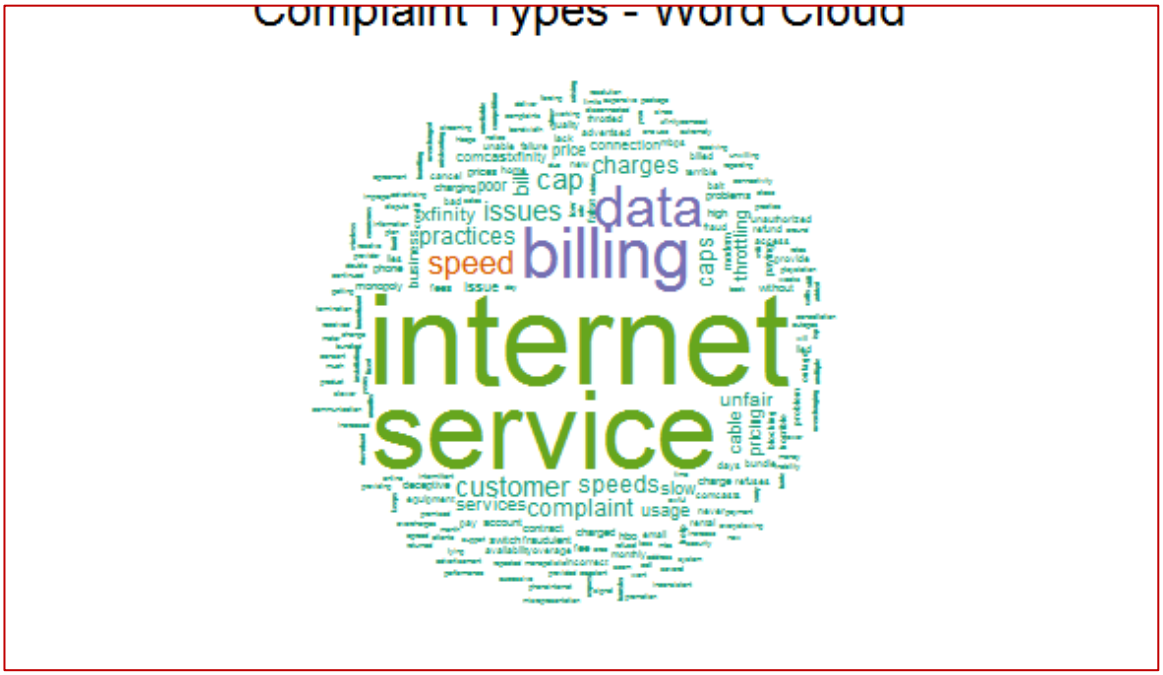
As can be clearing seen the June has a major spike in complaints with last week of June increasing the complaints to the maximum.

2. Provide a table with the frequency of complaint types.

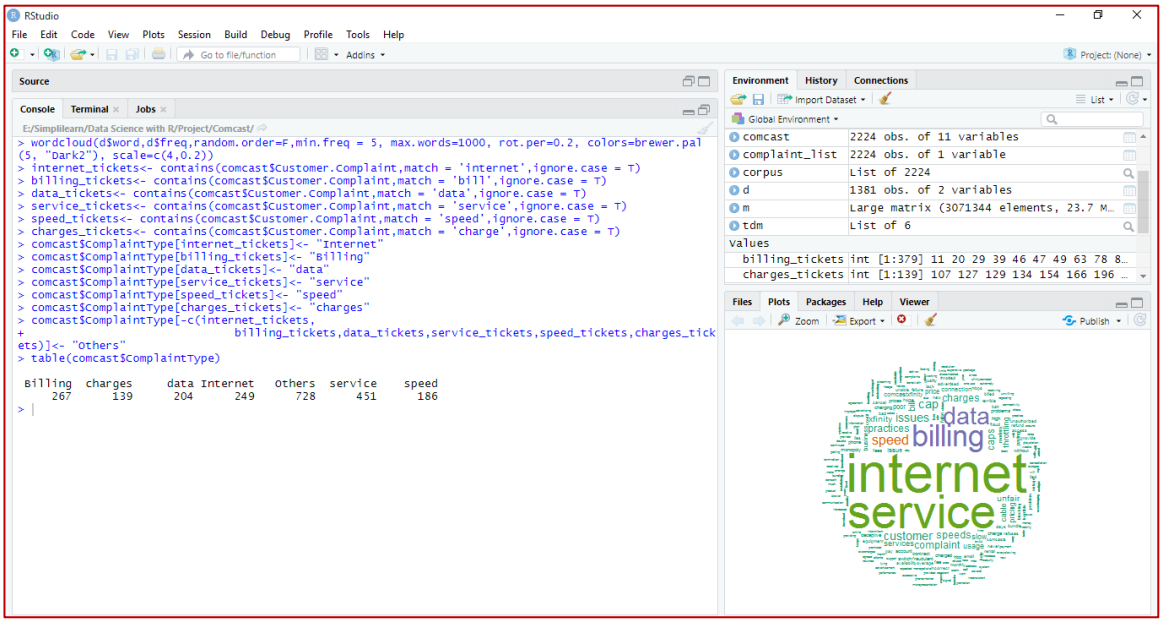
a. Which complaint types are maximum i.e., around internet, network issues, or across any other domains.

In order to understand the complaint types, first a world cloud was created using text mining techniques. All unwanted text was removed from the complaint type column to understand what exactly were the issues that the consumer faced.





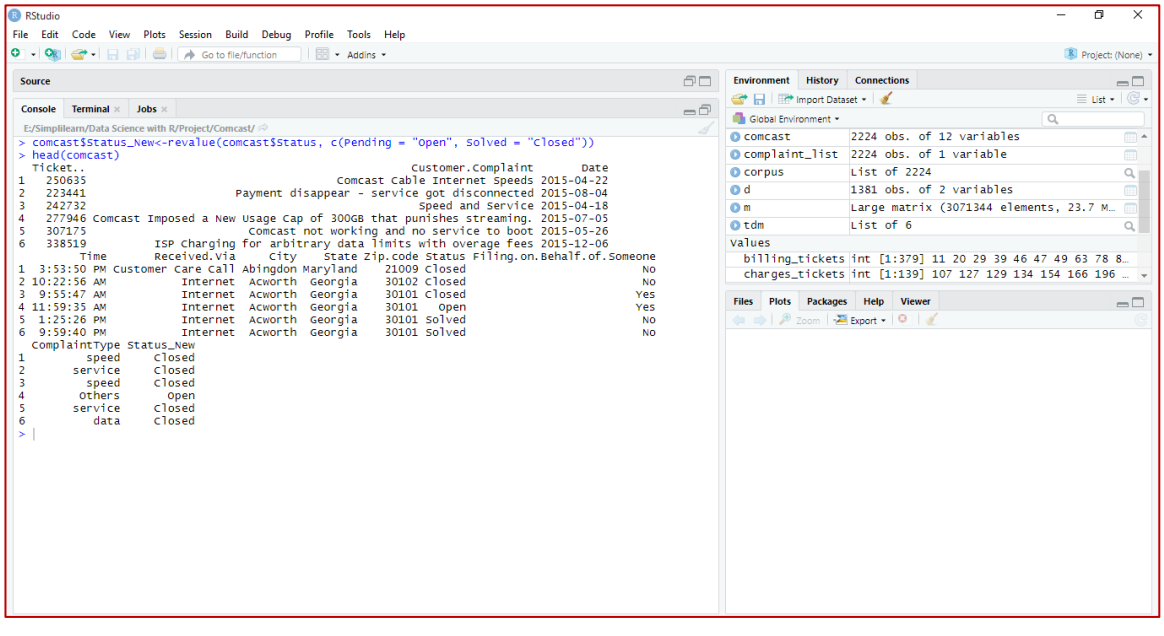
As can be clearing seen from the word cloud, most complaint types relate to Data, Billing, Internet, Service or speed. In order to check the frequency of these compliants a table was created that points to frequency of these types of complaints.



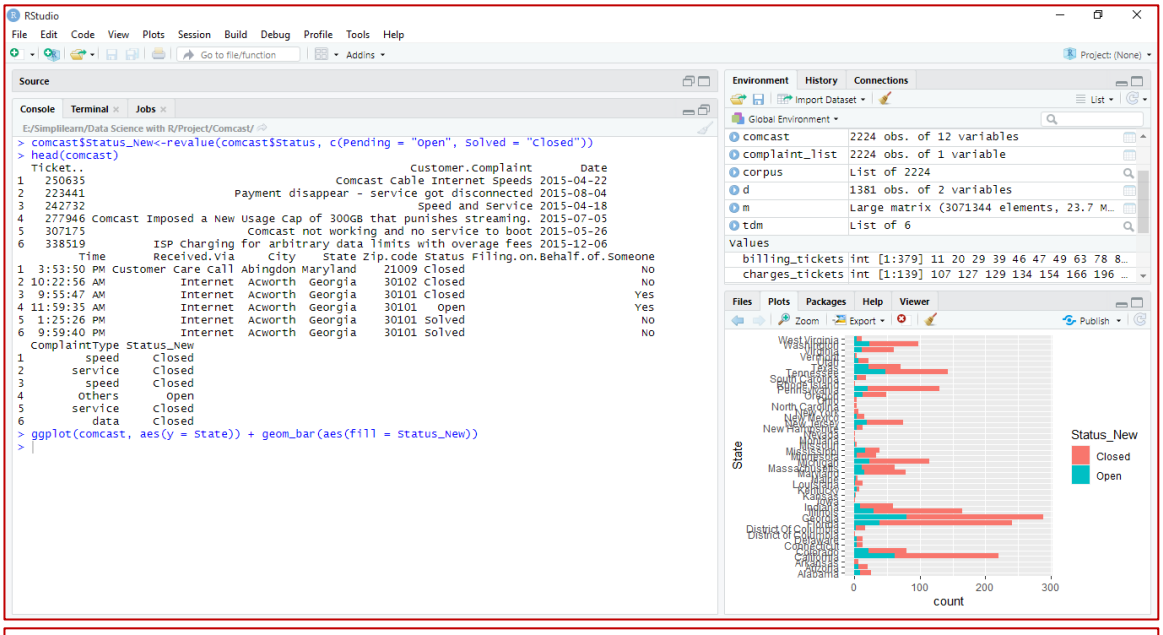
Interpretation:

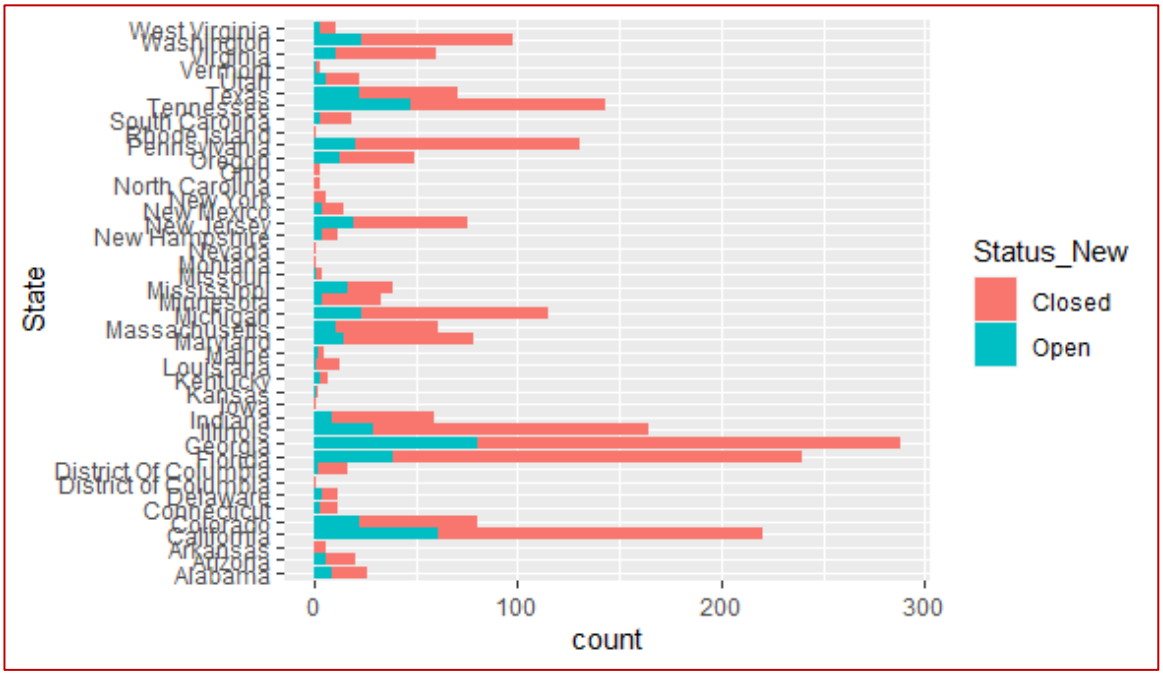
Others are basically, all complaints that do not fall under the major complaint types. Apart from that, the most complaint type by consumers as was seen and now verified is that of Internet and the service provided by the company.

3. 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.



a. Provide state wise status of complaints in a stacked bar chart.



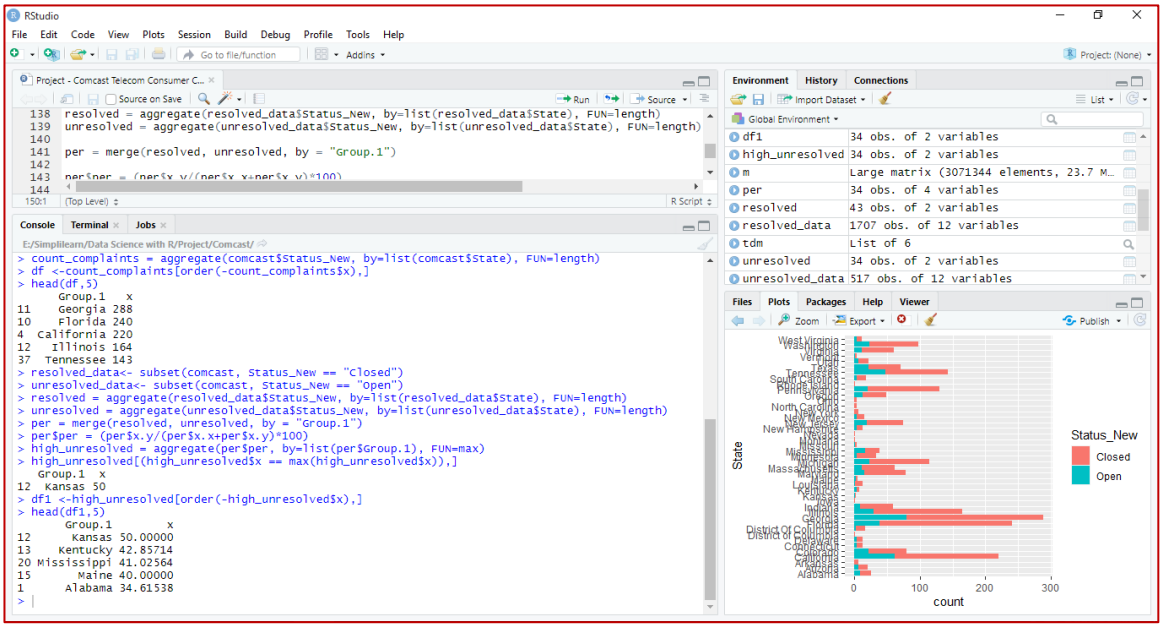


b. Use the categorized variable from Q3. Provide insights on: - Which state has the maximum complaints.
c. Which state has the highest percentage of unresolved complaints.

To find maximum complaints in a State, aggregation was done to calculate the frequency of complaint types based on State.

To find unresolved complaints, two subsets were created, one with Open cases and other with Closed. Then aggregation was done as done for maximum complaints. We

then merged the two dataframes and calculated the percentage on unresolved complaints over total complaints.

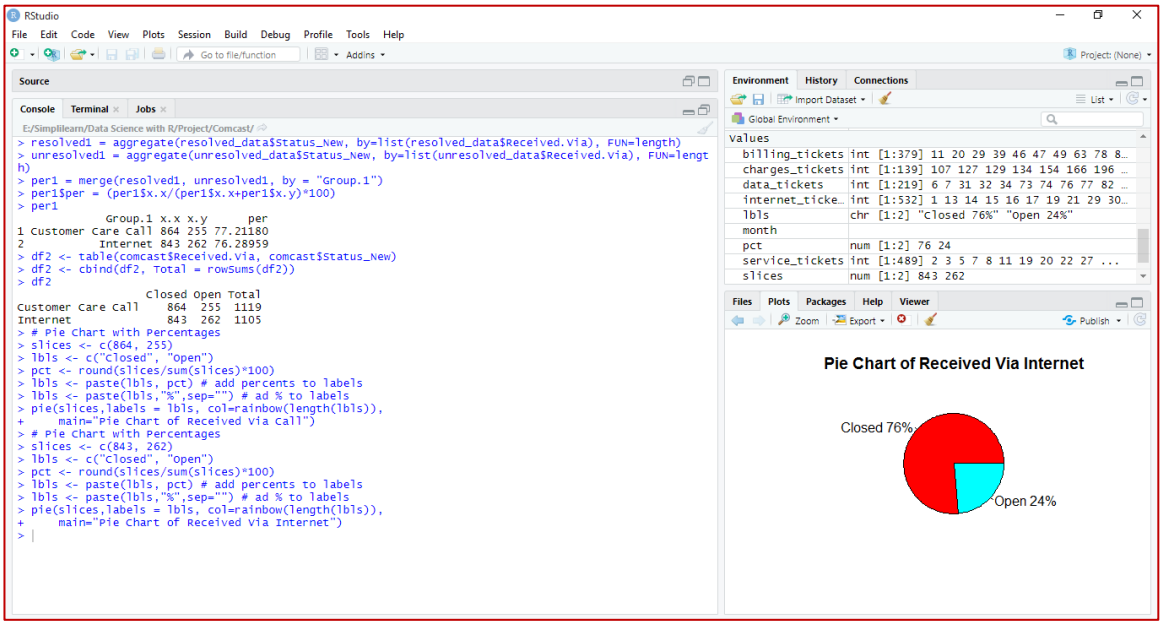


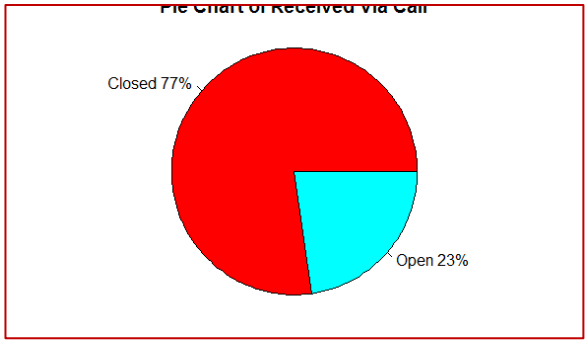
Interpretation:

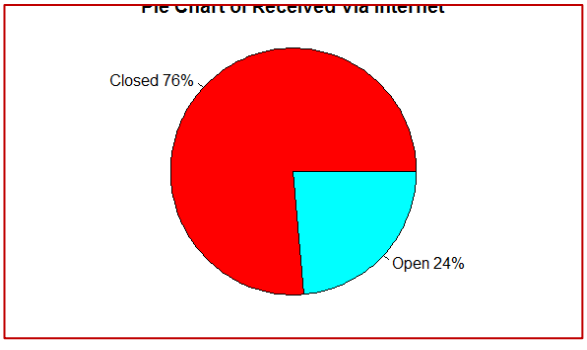
Georgia has the maximum complaints followed by Florida and California. However, Kansas has the most unresolved cases which amounts to 50%.

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

Using the above subsets and aggregating the complaint type by received via call or internet, we get the percentage of total resolved complaints. Pie chart has been created to show the difference as well.







Interpretation:

A total of 76% and 77% have been resolved for complaints received via Internet and Call respectively.

Programming Codes:

```
#Reading Comcast Data and loading libraries

rm(list=ls())

library(dplyr)
library(ggplot2)
library(lubridate)
library(plyr)
library(zoo)
library(NLP)
library(tm)
library(RColorBrewer)
library(wordcloud)
library(gridExtra)

setwd("E:\\Simplilearn\\Data Science with R\\Project\\Comcast")

comcast<- read.csv("Comcast Telecom Complaints data.csv")
head(comcast)
str(comcast)

sum(is.na(comcast))

# Cleaning the date

comcast$Date<- dmy(comcast$Date)

#Provide the trend chart for the number of complaints at monthly and daily granularity levels.

#Daily Trend

a = aggregate(comcast$Date, by=list(comcast$Date), FUN=length)
str(a)

ggplot(a, aes(Group.1, x, group = 1)) +
  geom_line() +
  xlab("Date") +
  ylab("Number of Complaints")

#Monthly Trend
```

```
month<- as.yearmon(comcast$Date, format = "%d-%m-%Y")
b<-cbind(comcast,month)

c = aggregate(b$month, by=list(b$month), FUN=length)
str(c)

ggplot(c, aes(Group.1, x, group = 1)) +
  geom_line() +
  xlab("Date") +
  ylab("Number of Complaints")

#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.

complaint_list = data.frame(comcast$Customer.Complaint)
colnames(complaint_list)=c("Complaint")

corpus= Corpus(VectorSource(complaint_list$Complaint))

#Text Cleaning

corpus <- tm_map(corpus,content_transformer(tolower)) #Converting all text into lower case
corpus<- tm_map(corpus,removeNumbers) #Remove Numbers
corpus = tm_map(corpus,removeWords,stopwords(kind="en"))#Removing common stop words
corpus = tm_map(corpus,removePunctuation)#Remove Punctuation
corpus = tm_map(corpus,stripWhitespace)#Removing white spaces
corpus= tm_map(corpus,removeWords,c("get","took","can","can","comcast"))#Remove additional words

#Create Term Document Matrix (TDM)

tdm = TermDocumentMatrix(corpus)
m=as.matrix(tdm)
v=sort(rowSums(m),decreasing = T)

#List with Frequency of Compliant Types

d=data.frame(word=names(v),freq=v)

#word cloud

set.seed(2)
```

```
wordcloud(d$word,d$freq,random.order=F,min.freq = 5, max.words=1000, rot.per=0.2,
colors=brewer.pal(5, "Dark2"), scale=c(4,0.2))
title(main = "Complaint Types - Word Cloud",font.main=1,cex.main=1.5)

# Complaint Type Processing as seen from word cloud

internet_tickets<- contains(comcast$Customer.Complaint,match = 'internet',ignore.case = T)
billing_tickets<- contains(comcast$Customer.Complaint,match = 'bill',ignore.case = T)
data_tickets<- contains(comcast$Customer.Complaint,match = 'data',ignore.case = T)
service_tickets<- contains(comcast$Customer.Complaint,match = 'service',ignore.case = T)
speed_tickets<- contains(comcast$Customer.Complaint,match = 'speed',ignore.case = T)
charges_tickets<- contains(comcast$Customer.Complaint,match = 'charge',ignore.case = T)

comcast$ComplaintType[internet_tickets]<- "Internet"
comcast$ComplaintType[billing_tickets]<- "Billing"
comcast$ComplaintType[data_tickets]<- "data"
comcast$ComplaintType[service_tickets]<- "service"
comcast$ComplaintType[speed_tickets]<- "speed"
comcast$ComplaintType[charges_tickets]<- "charges"

comcast$ComplaintType[-c(internet_tickets,
                         billing_tickets,data_tickets,service_tickets,speed_tickets,charges_tickets)]<- "Others"

table(comcast$ComplaintType)

#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.

comcast$Status_New<-revalue(comcast$Status, c(Pending = "Open", Solved = "Closed"))
head(comcast)

#Provide state wise status of complaints in a stacked bar chart. Use the categorized variable from Q3.
Provide insights on:

ggplot(comcast, aes(y = State)) + geom_bar(aes(fill = Status_New))

#Which state has the maximum complaints

count_complaints = aggregate(comcast$Status_New, by=list(comcast$State), FUN=length)
df <-count_complaints[order(-count_complaints$x),]
head(df,5)

#Which state has the highest percentage of unresolved complaints
```

```
resolved_data<- subset(comcast, Status_New == "Closed")
unresolved_data<- subset(comcast, Status_New == "Open")

resolved = aggregate(resolved_data$Status_New, by=list(resolved_data$State), FUN=length)
unresolved = aggregate(unresolved_data$Status_New, by=list(unresolved_data$State), FUN=length)

per = merge(resolved, unresolved, by = "Group.1")

per$per = (per$x.y/(per$x.x+per$x.y)*100)

high_unresolved = aggregate(per$per, by=list(per$Group.1), FUN=max)
high_unresolved[(high_unresolved$x == max(high_unresolved$x)),]
df1 <-high_unresolved[order(-high_unresolved$x),]
head(df1,5)

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

resolved1 = aggregate(resolved_data$Status_New, by=list(resolved_data$Received.Via), FUN=length)
unresolved1 = aggregate(unresolved_data$Status_New, by=list(unresolved_data$Received.Via),
FUN=length)

per1 = merge(resolved1, unresolved1, by = "Group.1")

per1$per = (per1$x.x/(per1$x.x+per1$x.y)*100)
per1

df2 <- table(comcast$Received.Via, comcast$Status_New)
df2 <- cbind(df2, Total = rowSums(df2))
df2

# Pie Chart with Percentages
slices <- c(864, 255)
lbls <- c("Closed", "Open")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
   main="Pie Chart of Received Via Call")

# Pie Chart with Percentages
slices <- c(843, 262)
```

```
lbls <- c("Closed", "Open")
pct <- round(slices/sum(slices)*100)
lbls <- paste(lbls, pct) # add percents to labels
lbls <- paste(lbls,"%",sep="") # ad % to labels
pie(slices,labels = lbls, col=rainbow(length(lbls)),
   main="Pie Chart of Received Via Internet")
```

----------The End----------