Data Analysis with R

Comcast Telecom Consumer Complaints.

Introduction

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 $2.3 million, after receiving over 1000 consumer complaints.

Data Dictionary

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.

- ➔ 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
- ➔ Zip Code: Customer zip
- ➔ Status: Status of complaint
- ➔ Filing on behalf of someone

Expectations/Tasks

- Import data into the R environment.
- Provide the trend charts 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 theInternet and customer care calls.

R-Script

```
# Analysis for Comcast Telecom Consumer Complaints.

# Problem :- The firm has been providing terrible customer service.
# They continue to fall short despite repeated promises to improve.
# They need help to pin down what is wrong with Comcast's customer
# service.

library(lubridate)
library(ggplot2)
library(dplyr)

# Task - Import data into the R environment.
# Set a working directory where data is stored.
setwd(choose.dir())
Comcast_Complaints <- read.csv('Comcast Telecom Complaints
data.csv',
                               sep = ',',
                               strip.white = TRUE,
                               stringsAsFactors = FALSE)
# Check table structure
str(Comcast_Complaints)

# The column names have some dots (..), rename the column names.
Comcast_Complaints <- rename(Comcast_Complaints,
                             Ticket_Id = Ticket..,
                             Customer_Complaint =
Customer.Complaint,
                             Received_Via = Received.Via,
                             Zip_Code = Zip.code,
                             Representative =
Filing.on.Behalf.of.Someone)
```

```
# Convert columns into data structures.
Comcast_Complaints$Received_Via <-
as.factor(Comcast_Complaints$Received_Via)
Comcast_Complaints$City <- as.factor(Comcast_Complaints$City)
Comcast_Complaints$State <- as.factor(Comcast_Complaints$State)
Comcast_Complaints$Zip_Code <-
as.factor(Comcast_Complaints$Zip_Code)
Comcast_Complaints$Status <- as.factor(Comcast_Complaints$Status)
Comcast_Complaints$Representative <-
as.factor(Comcast_Complaints$Representative)

# Check table structure to be in-place to proceed further.
str(Comcast_Complaints)

# Date variable looks to be using different formats.
# Considering date as dd/mm/yy so convert format to make date
consistent.
Comcast_Complaints$Date[grepl('/', Comcast_Complaints$Date)] <-
sapply(
        Comcast_Complaints$Date[grepl('/',
Comcast_Complaints$Date)],
          function(date) {
                  paste0(unlist(strsplit(date, '/'))[c(1,2,3)],
                         collapse = '-')
                  }, USE.NAMES = FALSE)
Comcast_Complaints$Date <- as.Date(Comcast_Complaints$Date,
"%d-%m-%y")

# Check for empty values.
empty_values <- sum(is.na(Comcast_Complaints))
if(empty_values != 0) print(empty_values)

# Add a new categorical variable to categorize complaint type
# Domains of complaints
domains <- c(Charges =
c("arge","price","hike","bill","saction","fun"),
             Service = c("custom","serv"),
             Network = c("network", "call", "signal"),
             Usage= c("data","cap","usage"),
             Internet = c("internet", "speed"),
             Ethics = c("fraud","rac","mono", "not"))

complaint_type <- sapply(domains,
                         grepl,
                         Comcast_Complaints$Customer_Complaint,
                         ignore.case = TRUE)
```

```
complaint_type <- apply(complaint_type, 1, function(r)
names(which(r)))
complaint_type <- sapply(complaint_type,
                         function(s) if (length(s) == 0) "Other"
                         else names(which.max(table(
                                 unlist(strsplit(
                                         gsub('[[:digit:]]+', '',
s), " "))))))
                         )

Comcast_Complaints$Complaint_Domain <- as.factor(complaint_type)

# Check table structure
str(Comcast_Complaints)

# 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.

# Checks status and returns updated status as Open or Closed.
resolution <- function(status) {
        status <- ifelse(status == "Solved" | status == "Closed",
                         "Closed", "Open")
        return(as.factor(status))
}

# Add a new variable to categorize resolution status as
Open/Closed.
Comcast_Complaints <- transform(Comcast_Complaints,
                                        Resolution =
resolution(Status))

# Check table structure
str(Comcast_Complaints)

# Task:- Provide the trend chart for the number of complaints at
monthly
# and daily granularity levels.
Month_Wise_Data <- transform(Comcast_Complaints,
                             Month = month(Date, label = TRUE))
monthly_count <- table(Month_Wise_Data$Month)
daily_count <- table(Comcast_Complaints$Date)

# Check monthly mean and median data to know some details.
mean(monthly_count) # 185.3333
median(monthly_count) # 57
```

```
mean(daily_count) # 24.43956
median(daily_count) # 17

# Visualize month wise complaints
ggplot(as.data.frame(monthly_count),
       aes(Var1, Freq, group=1, label=Freq)) +
        geom_line(size=1.2) +
        geom_point(size = 2) +
        geom_text(nudge_y = 50) +
        labs(y = "Number of complaints",
             x="",
             title = "Monthly granularity level") +
        theme_minimal()

# Visualize date wise complaints
ggplot(as.data.frame(daily_count),
       aes(Var1, Freq, group=1, label=Freq)) +
        geom_line(size=1) +
        geom_point(size = 2) +
        geom_text(nudge_y = 5) +
        labs(y = "Number of complaints",
             x="",
             title = "Daily granularity level") +
        theme(axis.text.x = element_text(angle = 90, size = 6))

# Insights. Reason behind spike in June
Month_Wise_Data <- filter(Month_Wise_Data, Month == "Jun")

# Function to get insights by states on the sample under test.
sample_insights_by_state <- function(sample_df) {
        sample_df <- group_by(sample_df, State, Complaint_Domain)
        sample_df <- summarise(sample_df, Count = n())
        sample_df <- group_by(sample_df, State) %>%
                mutate(Distribution = round(Count*100/sum(Count),
digits=0))
        ggplot(sample_df,
               aes(x= Count, y = reorder(State, Count),
                   fill = Complaint_Domain)) +
                geom_bar(position = "stack", stat = "identity") +
                geom_text(size=2.5, position = position_stack(vjust
= 0.5),
                          aes(label = Distribution))+
                scale_fill_brewer(palette="Paired") +
                labs(x = "Number of complaints",
                     y = "",
                     title = paste("State wise complaint
(percentages) ")) +
```

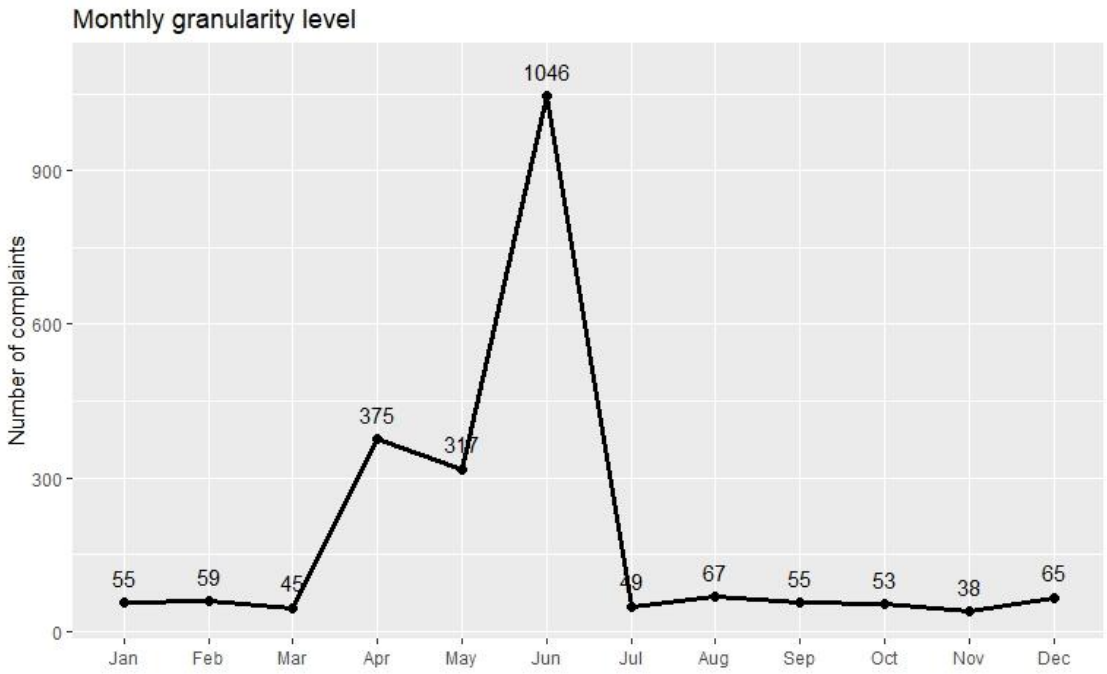
```
                theme_minimal()
}

sample_insights_by_state(Month_Wise_Data)

# Task:- 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.
table(Comcast_Complaints$Complaint_Domain)
# Above table defines that maximum complaints are around any other
# domains (580) followed by Charges (527) and Internet(476).

# Task:- Provide state wise status of complaints in a stacked bar
# chart. Use the categorized variable from Q3.
# Create a new data frame which specifies resolution status by
states
# in percentages.
Grouped_By_State <- group_by(Comcast_Complaints, State, Resolution)
Grouped_By_State <- summarise(Grouped_By_State, Count = n())
Grouped_By_State <- group_by(Grouped_By_State, State) %>%
        mutate(Distribution = round(Count*100/sum(Count),
digits=0))
ggplot(Grouped_By_State,
       aes(x= Count, y = reorder(State, Count),
           label = paste(Distribution, "% (", Count,")"),
           fill = Resolution)) +
        geom_bar(position = "stack", stat = "identity") +
        geom_text(size = 3, position = position_stack(vjust = 0.8))
+
        scale_fill_brewer(palette="Pastel1", direction = -1) +
        labs(x = "Number of complaints", y = "",
             title = "State wise complaint resolution") +
        theme(legend.position = "top")

# Provide insights on:
# Which state has the maximum complaints
# Ans - Top 3 states as per graph are Georgia, Florida and
California.

# Which state has the highest percentage of unresolved complaints
Grouped_by_Unresolved <- Grouped_By_State %>% filter(Resolution ==
"Open")
total_unresolved <- sum(Grouped_by_Unresolved$Count)
Grouped_by_Unresolved <- Grouped_by_Unresolved %>% mutate(
        Distribution = round(Count*100/total_unresolved, digits=0))
head(arrange(Grouped_by_Unresolved, desc(Distribution)))
```

```
# Ans- Georgia(80) with 15%, California(61) with 12% and
Tennessee(47) with 9%.
# Insights on open complaints w.r.t complaint domains
Grouped_by_Unresolved <- Comcast_Complaints %>% filter(Resolution
== "Open")
sample_insights_by_state(Grouped_by_Unresolved)
ggplot(as.data.frame(table(Grouped_by_Unresolved$Complaint_Domain))
,
       aes(x = "", y = Freq, fill = Var1,
           label = paste(round(Freq*100/sum(Freq), digits = 0),
"%"))) +
        geom_bar(stat = "identity", width = 1) +
        geom_text(size=5, position = position_stack(vjust = 0.5),
                  show.legend = FALSE,
                  aes(label = paste(round(Freq*100/sum(Freq),
                                          digits = 0), "%")))+
        coord_polar("y", start = 0) +
        scale_fill_brewer(palette="Paired", direction = -1) +
        labs(x = "", y = "", fill = "Complaint Types",
             title = "Domain distribution for unresolved
complaints") +
        theme_minimal()
# Internet, usage and charges related complaints are unresolved.
# This might be due to the lack of technical support from the
company.

# Task :- Provide the percentage of complaints resolved till date,
# which were received through the Internet and customer care calls.
# Create a new data frame which specifies resolution status by
# received through the Internet and customer care calls in
percentages.
Grouped_By_Type <- summarise(group_by(Comcast_Complaints,
                                      Received_Via, Resolution),
Count = n())
Grouped_By_Type <- group_by(Grouped_By_Type, Received_Via) %>%
        mutate(Distribution = round(Count*100/sum(Count),
digits=0))

# Visualize data.
ggplot(Grouped_By_Type,
       aes(x = Received_Via, y = Count, fill = Resolution,
           label = paste(Distribution, "%"))) +
        geom_bar(position = "stack", stat = "identity", width =
0.5) +
        geom_text(size = 4, position = position_stack(vjust = 0.5))
+
        scale_fill_brewer(palette="Pastel1", direction = -1) +
```

```
        labs(x = "", y = "", title = "Complaint Resolution Status")
+
        theme_minimal()
# The percentage of complaints resolved via. Customer Care Calls
are
# found to be slightly better as compared to the Internet.
```

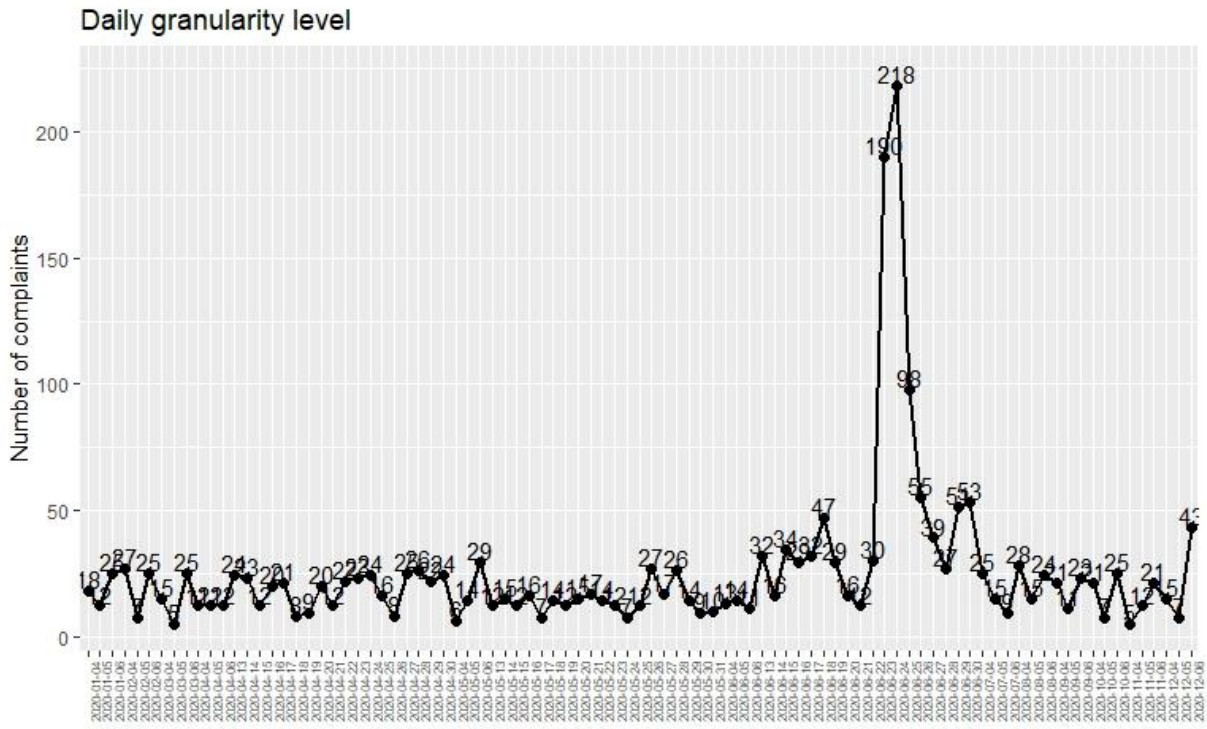
Results

- Provide the trend charts for the number of complaints at monthly and daily granularity levels.



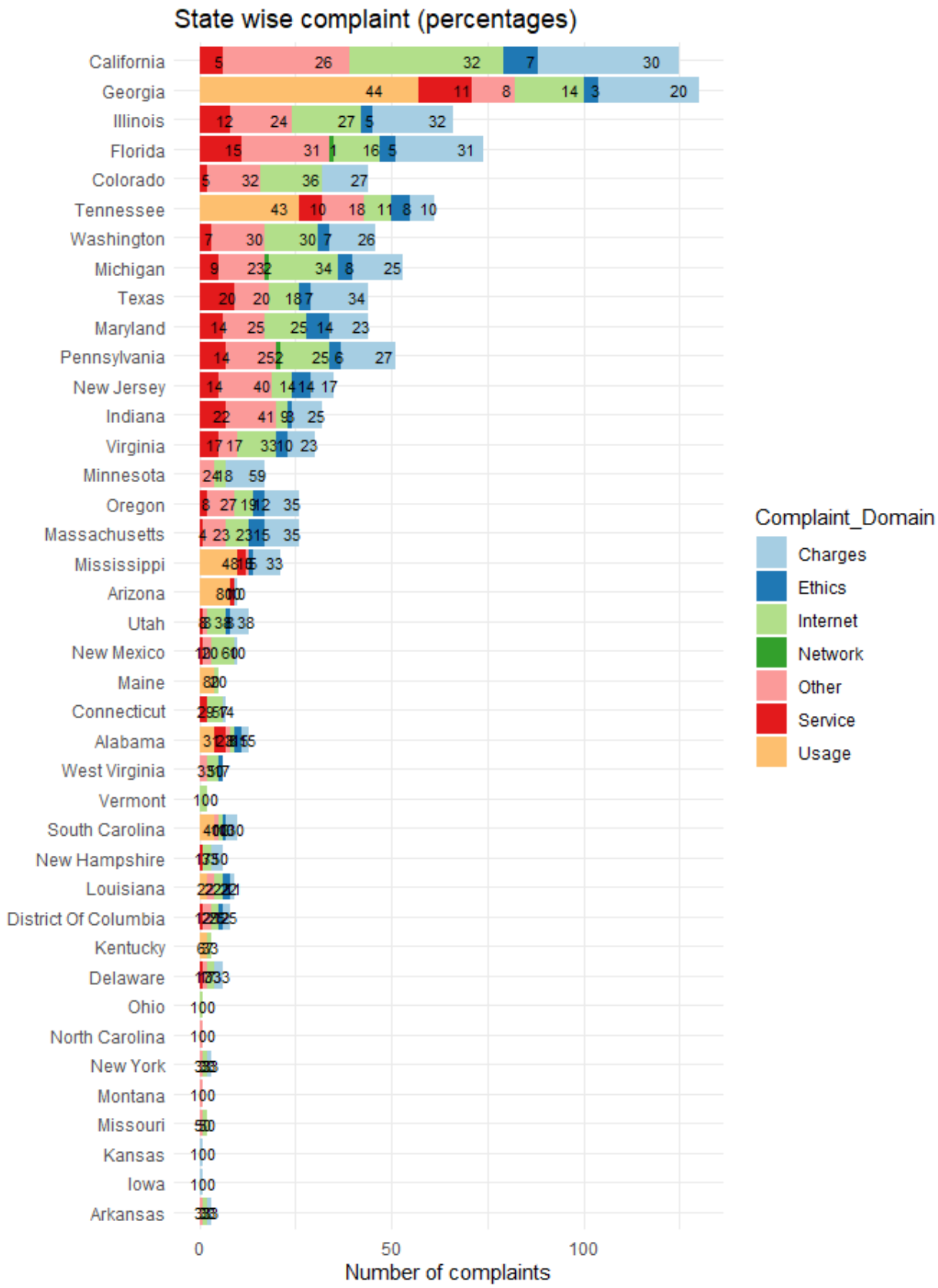
  - Looking at the charts above, a sudden surge in complaints can be seen in between the first quarter of the financial year, and especially in the month of June.
  - The average recorded cases per month is 185 with a median of 57. This indicates that some issue has occurred (as in June month) which moved the mean cases to be three times of the median.

- A sudden spike is seen in between June 22 and June 24 (from below plot) where the number of recorded complaints were seen to be 5 times more than normal recorded daily cases.
- The average recorded cases per day is 24 with a median of 17. This indicates the surge in complaint in between some dates i.e. between June 22 to June 24.



- <u>Analysis of complaints w.r.t states in the month of June is as shown</u>

  Major rises can be seen in Georgia with usage (44%) domains followed by the internet and charges related complaints in Georgia and California. In the month of June, there might be chances that data was heavily deducted in Georgia. The Internet and charges related issues can be seen almost in every state. Below, visualization depicts the state wise percent analysis of complaints in the month of June.


State wise complaint (percentages)
California
Georgia
Illinois
Florida
Colorado
Tennessee
Washington
Michigan
Texas
Maryland
Pennsylvania
New Jersey
Indiana
Virginia
Minnesota
Oregon
Massachusetts
Mississippi
Arizona
Utah
New Mexico
Maine
Connecticut
Alabama
West Virginia
Vermont
South Carolina
New Hampshire
Louisiana
District Of Columbia
Kentucky
Delaware
Ohio
North Carolina
New York
Montana
Missouri
Kansas
Iowa
Arkansas
0
50
100
Number of complaints
Complaint_Domain
Charges
Ethics
Internet
Network
Other
Service
Usage

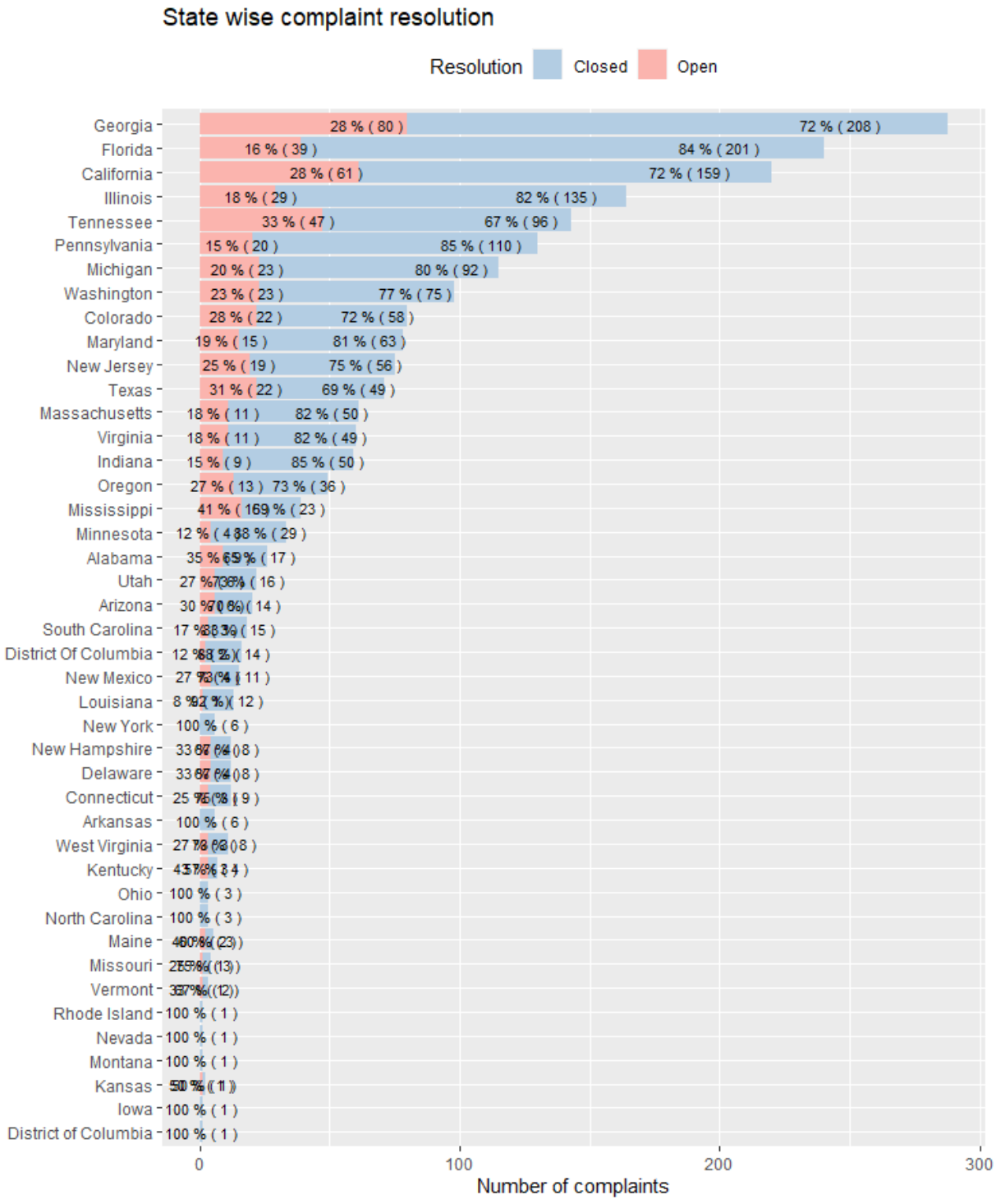
- Provide a table with the frequency of complaint types.

Domain	NETWORK	ETHICS	USAGE	SERVICE	INTERNET	CHARGES	OTHER
Count	11	149	227	260	476	527	574

  - Which complaint types are maximum i.e., around the internet, network issues, or across any other domains.
    Looking at the table above, it looks like most of the complaints are not present in any of the specific domains but is a cluster of non-specific domains with a count of 574 which is followed by charges(527), internet(476) and service(260) domain related issues. If we neglect other domains, majority complaints are in the domain of the internet and charges which include billing, price and packs. Comcast can check these domain areas to have formal audits in order to reduce complaints.

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

  Below charts provide visualization with the number of cases by states and resolution distribution in percentages.

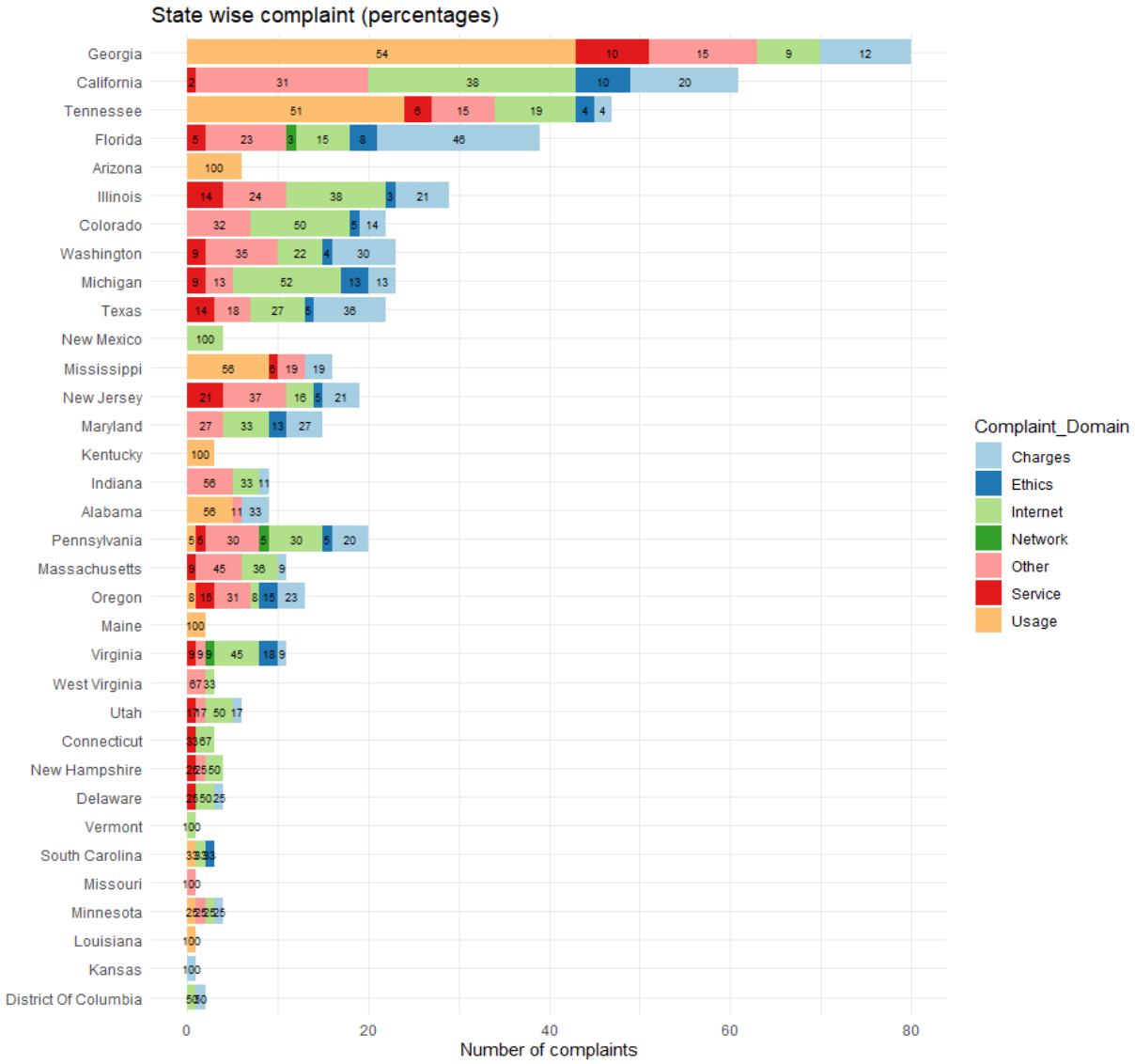


- Which state has the maximum complaints

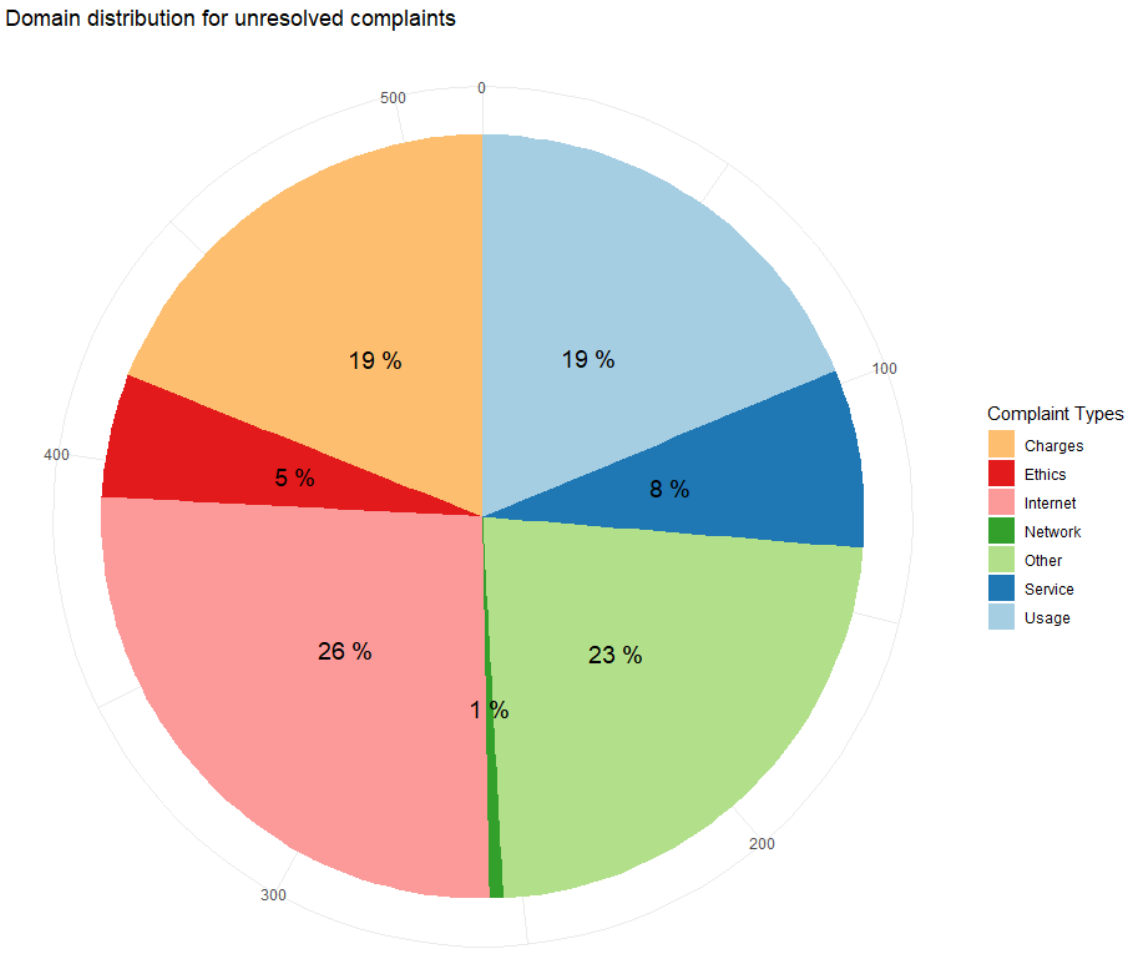
  Looking at the above chart, Georgia appears to have a maximum recorded cases followed by Florida and California.
- Which state has the highest percentage of unresolved complaints

  .Georgia seems to have the highest i.e. 15% of unresolved complaints followed by California (12%) and Tennessee (9%). This can be seen from above visualization too the count of unresolved complaints for top 3 states are 80,61 and 47 respectively.
- <u>Insights on unresolved complaints w.r.t complaint domain.</u>

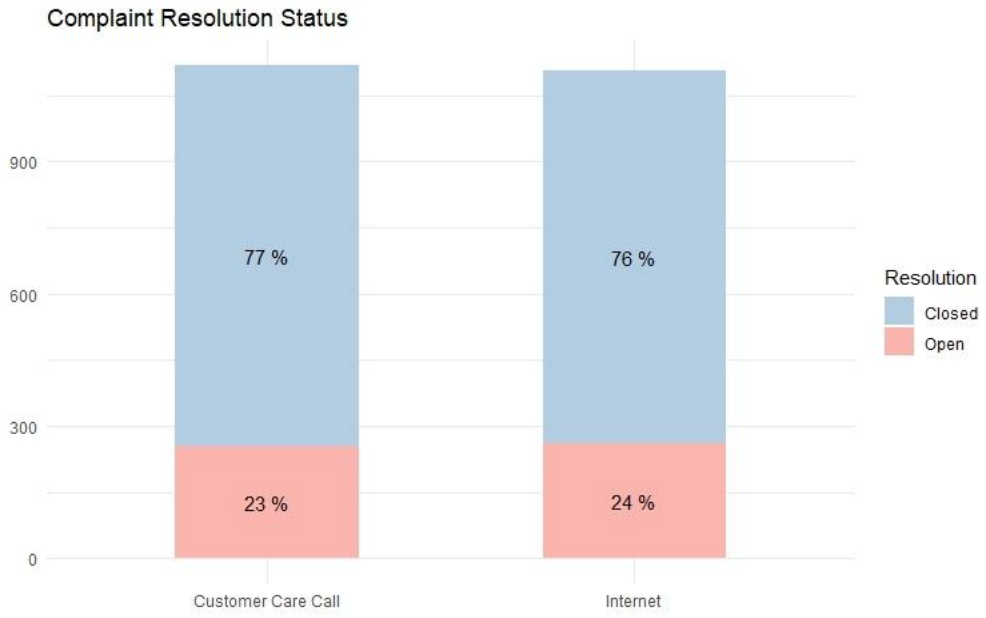
  <u>*Open complaints distribution by states :-*</u>



From the above chart, it can be seen that the top contributors to open complaints are Georgia, California and Tennessee. Georgia and Tennessee involve high stakes of data cap/usage related open complaints. From the above chart it can be seen that customer support in the top 3 states is not proper, when it comes to resolution of complaints.



- Provide the percentage of complaints resolved till date, which were received through the Internet and customer care calls.
  Looking at the below chart, it can be seen that the complaints received via Customer Care are slightly better addressed as compared to the ones received via Internet. But, the percentage itself depicts that the variation is only around 1% which must not be a big problem.



Conclusion

The Customers are facing issues related to :-

- Internet
- Usage/Data cap
- Charges

Majority stakes of complaint reside in Georgia, California, Florida, Illinois and Tennessee. Comcast needs to look at its charges, data deductions in the above mentioned states to reduce the number of complaints. Also, there are some complaints registered over company's work ethics, so this also needs to be taken care of.