Comcast Telecom Consumer Complaints Analysis

Abhilash Roy

3rd April 2021

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 to be done ->

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 to tell which complaint types are maximum i.e., around internet, network issues, or across any other domains.
3. Provide state wise status of complaints while providing insights on:
    - Which state has the maximum complaints?
    - 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.

Data Dictionary ->

S.No.	Column name	Explanation
1	Ticket #	Ticket number assigned to each complaint
2	Customer Complaint	Description of complaint
3	Date	Date of complaint
4	Time	Time of complaint
5	Received Via	Mode of communication of the complaint
6	City	Customer city
7	State	Customer state
8	Zipcode	Customer zip
9	Status	Status of complaint
10	Filing on behalf of someone	Complaint done on behalf of someone

NOTE : In this data set, factors like population, gender and age were not considered. This was a closed data project, and analysis had to done only with the data provided. Description of steps taken have also been provided for easy understanding.

All the observations/ insights have been highlighted by the color **"Violet"**

And the steps taken will be highlighted by the color **"Brown"** *for easy differentiation.*

=> In the beginning, load all required libraries and then check working directory of system.

```
getwd()
```

```
## [1] "/Users/abhilashroy/Desktop/RStudio | github/Comcast_Complaints_Analysis"
```

=> As the working directory is correct, save csv file into r in a new variable.

```
Complaints <- read.csv("Comcast Telecom Complaints Data.csv")
```

=> Check the structure of data

```
str(Complaints)
```

```
## 'data.frame':    2224 obs. of  10 variables:
##  $ Ticket..                   : chr  "250635" "223441" "242732" "277946" ...
##  $ Customer.Complaint         : chr  "Comcast Cable Internet Speeds" "Payment disappear - service go
##  $ 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" ...
```

This data set contains 2224 rows and 10 columns.

=> Now, to check whether all columns contain null values.

```
length(which(is.na(Complaints$Ticket..)))
```

```
## [1] 0
```

```
length(which(is.na(Complaints$Customer.Complaint)))
```

```
## [1] 0
```

```
length(which(is.na(Complaints$Date)))
```

```
## [1] 0
```

```
length(which(is.na(Complaints$Time)))
```

```
## [1] 0
```

```
length(which(is.na(Complaints$Received.Via)))
```

```
## [1] 0
```

```
length(which(is.na(Complaints$City)))
```

```
## [1] 0
```

```
length(which(is.na(Complaints$State)))
```

```
## [1] 0
```

```
length(which(is.na(Complaints$Zip.code)))
```

```
## [1] 0
```

```
length(which(is.na(Complaints$Status)))
```

```
## [1] 0
```

```
length(which(is.na(Complaints$Filing.on.Behalf.of.Someone)))
```

```
## [1] 0
```

None of the columns contain any null values.

=> Change data type of State, City, Status and Received.Via from Character to Factor

```
Complaints <- Complaints %>% mutate(City = factor(City))
Complaints <- Complaints %>% mutate(State = factor(State))
Complaints <- Complaints %>% mutate(Status = factor(Status))
Complaints <- Complaints %>% mutate(Received.Via = factor(Received.Via))
```

=> As some dates contain the character "/"and some contain "-", so using lubridate package to convert them into date type and similar format. Also convert time to time data type.

```
Complaints <- Complaints %>% mutate(Date = dmy(Date))
Complaints <- Complaints %>% mutate(Time = hms(Time))
```

=> Check the structure of data

```
str(Complaints)
```

```
## 'data.frame':    2224 obs. of  10 variables:
##  $ Ticket..                   : chr  "250635" "223441" "242732" "277946" ...
##  $ Customer.Complaint         : chr  "Comcast Cable Internet Speeds" "Payment disappear - service go
##  $ Date                       : Date, format: "2015-04-22" "2015-08-04" ...
##  $ Time                       :Formal class 'Period' [package "lubridate"] with 6 slots
##   .. ..@ .Data : num  50 56 47 35 26 40 55 14 30 31 ...
##   .. ..@ year  : num  0 0 0 0 0 0 0 0 0 0 ...
##   .. ..@ month : num  0 0 0 0 0 0 0 0 0 0 ...
##   .. ..@ day   : num  0 0 0 0 0 0 0 0 0 0 ...
##   .. ..@ hour  : num  3 10 9 11 1 9 10 6 11 6 ...
##   .. ..@ minute: num  53 22 55 59 25 59 13 56 46 46 ...
##  $ Received.Via               : Factor w/ 2 levels "Customer Care Call",..: 1 2 2 2 2 2 1 2 1 1 ...
##  $ City                       : Factor w/ 928 levels "Abingdon","Acworth",..: 1 2 2 2 2 2 2 3 4 4 ..
##  $ State                      : Factor w/ 43 levels "Alabama","Arizona",..: 19 11 11 11 11 11 11 21 4
##  $ Zip.code                   : int  21009 30102 30101 30101 30101 30101 30101 49221 94502 94501 ...
##  $ Status                     : Factor w/ 4 levels "Closed","Open",..: 1 1 1 2 4 4 3 4 1 2 ...
##  $ Filing.on.Behalf.of.Someone: chr  "No" "No" "Yes" "Yes" ...
```

```
levels(Complaints$State)
```

```
##  [1] "Alabama"              "Arizona"              "Arkansas"
##  [4] "California"           "Colorado"             "Connecticut"
##  [7] "Delaware"             "District of Columbia" "District Of Columbia"
## [10] "Florida"              "Georgia"              "Illinois"
## [13] "Indiana"              "Iowa"                 "Kansas"
## [16] "Kentucky"             "Louisiana"            "Maine"
## [19] "Maryland"             "Massachusetts"        "Michigan"
## [22] "Minnesota"            "Mississippi"          "Missouri"
```

```
## [25] "Montana"              "Nevada"              "New Hampshire"
## [28] "New Jersey"           "New Mexico"          "New York"
## [31] "North Carolina"       "Ohio"                "Oregon"
## [34] "Pennsylvania"         "Rhode Island"        "South Carolina"
## [37] "Tennessee"            "Texas"               "Utah"
## [40] "Vermont"              "Virginia"            "Washington"
## [43] "West Virginia"
```

From the levels of State column, it can easily be seen that "District of Columbia has been mentioned twice with different spellings. One as "District of Columbia" and other as "District Of Columbia"

=> Will change that below.

```
Complaints <- Complaints %>%
  mutate(State = gsub("District Of Columbia","District of Columbia",State)) %>%
  mutate(State = as.factor(State))
```

=> Changed data type to factor again, as changing data coverts type to character. And now, check again.

```
levels(Complaints$State)
```

```
##  [1] "Alabama"              "Arizona"              "Arkansas"
##  [4] "California"           "Colorado"             "Connecticut"
##  [7] "Delaware"             "District of Columbia" "Florida"
## [10] "Georgia"              "Illinois"             "Indiana"
## [13] "Iowa"                 "Kansas"               "Kentucky"
## [16] "Louisiana"            "Maine"                "Maryland"
## [19] "Massachusetts"        "Michigan"             "Minnesota"
## [22] "Mississippi"          "Missouri"             "Montana"
## [25] "Nevada"               "New Hampshire"        "New Jersey"
## [28] "New Mexico"           "New York"             "North Carolina"
## [31] "Ohio"                 "Oregon"               "Pennsylvania"
## [34] "Rhode Island"         "South Carolina"       "Tennessee"
## [37] "Texas"                "Utah"                 "Vermont"
## [40] "Virginia"             "Washington"           "West Virginia"
```

=> All good then :D

Complaints in this data set are from 42 different states.

=> Complains are currently divided in four different types, as mentioned below. (Total no. of complaints in each type is also given)

Status	length(Status)
Closed	734
Open	363
Pending	154
Solved	973

=> Create a new categorical variable with value as Open and Closed. Open and Pending is to be categorized as Open, wheras Closed and Solved is to be categorized as Closed.

```
Complaints <- Complaints %>%
  mutate(Status = ifelse((Status=='Open'|Status=='Pending'),"Open","Closed"))

Complaints %>% group_by(Status) %>% summarize(length(Status))
```

```
## # A tibble: 2 x 2
##   Status `length(Status)`
##   <chr>             <int>
## 1 Closed             1707
## 2 Open                517
```

```
Complaints <- Complaints %>% mutate(Status = as.factor(Status))

str(Complaints)
```

```
## 'data.frame':    2224 obs. of  10 variables:
##  $ Ticket..                   : chr  "250635" "223441" "242732" "277946" ...
##  $ Customer.Complaint         : chr  "Comcast Cable Internet Speeds" "Payment disappear - service go
##  $ Date                       : Date, format: "2015-04-22" "2015-08-04" ...
##  $ Time                       :Formal class 'Period' [package "lubridate"] with 6 slots
##   .. ..@ .Data : num  50 56 47 35 26 40 55 14 30 31 ...
##   .. ..@ year  : num  0 0 0 0 0 0 0 0 0 0 ...
##   .. ..@ month : num  0 0 0 0 0 0 0 0 0 0 ...
##   .. ..@ day   : num  0 0 0 0 0 0 0 0 0 0 ...
##   .. ..@ hour  : num  3 10 9 11 1 9 10 6 11 6 ...
##   .. ..@ minute: num  53 22 55 59 25 59 13 56 46 46 ...
##  $ Received.Via               : Factor w/ 2 levels "Customer Care Call",..: 1 2 2 2 2 2 1 2 1 1 ...
##  $ City                       : Factor w/ 928 levels "Abingdon","Acworth",..: 1 2 2 2 2 2 2 3 4 4 ..
##  $ State                      : Factor w/ 42 levels "Alabama","Arizona",..: 18 10 10 10 10 10 10 20 4
##  $ Zip.code                   : int  21009 30102 30101 30101 30101 30101 30101 49221 94502 94501 ...
##  $ Status                     : Factor w/ 2 levels "Closed","Open": 1 1 1 2 1 1 2 1 1 2 ...
##  $ Filing.on.Behalf.of.Someone: chr  "No" "No" "Yes" "Yes" ...
```

=> Data Wrangling Done.

=> In the beginning, I will check the no. of complaints at daily granularity level. Before that, will have to create a data frame for positioning of text on the plot. Also check on which date max no. of complaints were made.

```
Complaints %>%
  group_by(Date) %>% mutate(as.numeric(Date)) %>%
  summarise(No_of_Complaints = n()) %>% top_n(10)
```

```
## Selecting by No_of_Complaints
```

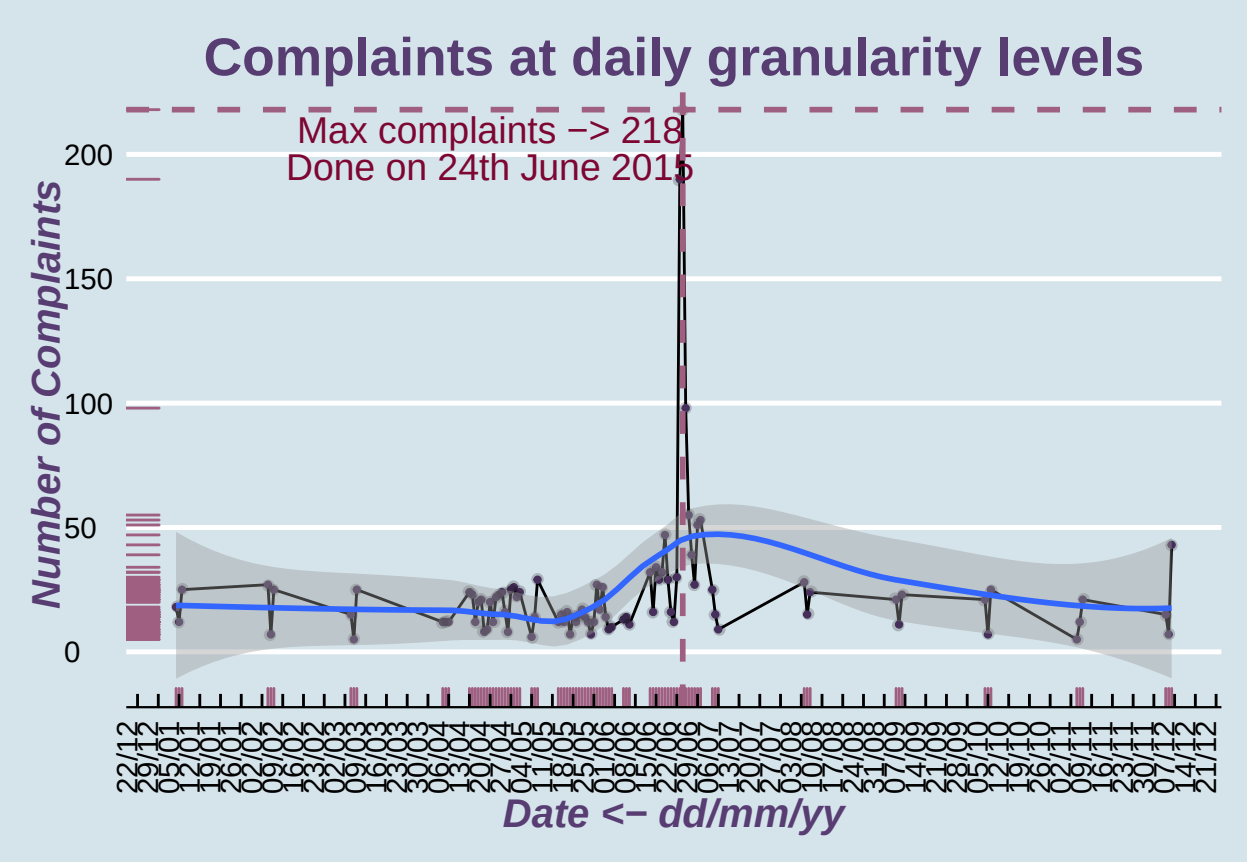
```
## # A tibble: 10 x 2
##    Date       No_of_Complaints
##    <date>                <int>
##  1 2015-06-15               34
##  2 2015-06-18               47
##  3 2015-06-23              190
##  4 2015-06-24              218
##  5 2015-06-25               98
##  6 2015-06-26               55
##  7 2015-06-27               39
##  8 2015-06-29               51
##  9 2015-06-30               53
## 10 2015-12-06               43
```

```
max_complaints <-
  data_frame(max_complaints = c("Max complaints -> 218",
             "Done on 24th June 2015"),
             x = as.Date(c("2015-04-20", "2015-04-20")), y = c(210,195) )
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## Please use `tibble()` instead.
```

```
Complaints %>%
  group_by(Date) %>% summarise(No_of_Complaints = n()) %>%
  ggplot(aes(Date, No_of_Complaints)) +
  geom_line() + scale_x_date(date_breaks = "1 week", date_labels = "%d/%m") +
  geom_point(col = "#583d72", size = 0.8) + geom_jitter(alpha = 0.2) +
  geom_rug(color = "#9f5f80") + geom_smooth() +
  geom_text(data = max_complaints, aes(x,y, label = max_complaints),
            size = 5, color = "#7d0633") + theme_economist() +
  xlab("Date <- dd/mm/yy") + ylab("Number of Complaints") +
  ggtitle("Complaints at daily granularity levels") +
  theme(axis.text.x = element_text(hjust = 0.5, angle = 90)) +
  theme(plot.title = element_text(hjust = 0.5, colour = "#583d72", size = 20)) +
  theme(axis.title.x = element_text(size = 15, color = "#583d72",
                                    face = "bold.italic")) +
  theme(axis.title.y = element_text(size = 15, color = "#583d72",
                                    face = "bold.italic")) +
  geom_hline(yintercept = 218, color = "#9f5f80", lty = 2, size = 1) +
  geom_vline(aes(xintercept=as.Date("2015-06-24")),
             color = "#9f5f80", lty = 2, size = 1)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

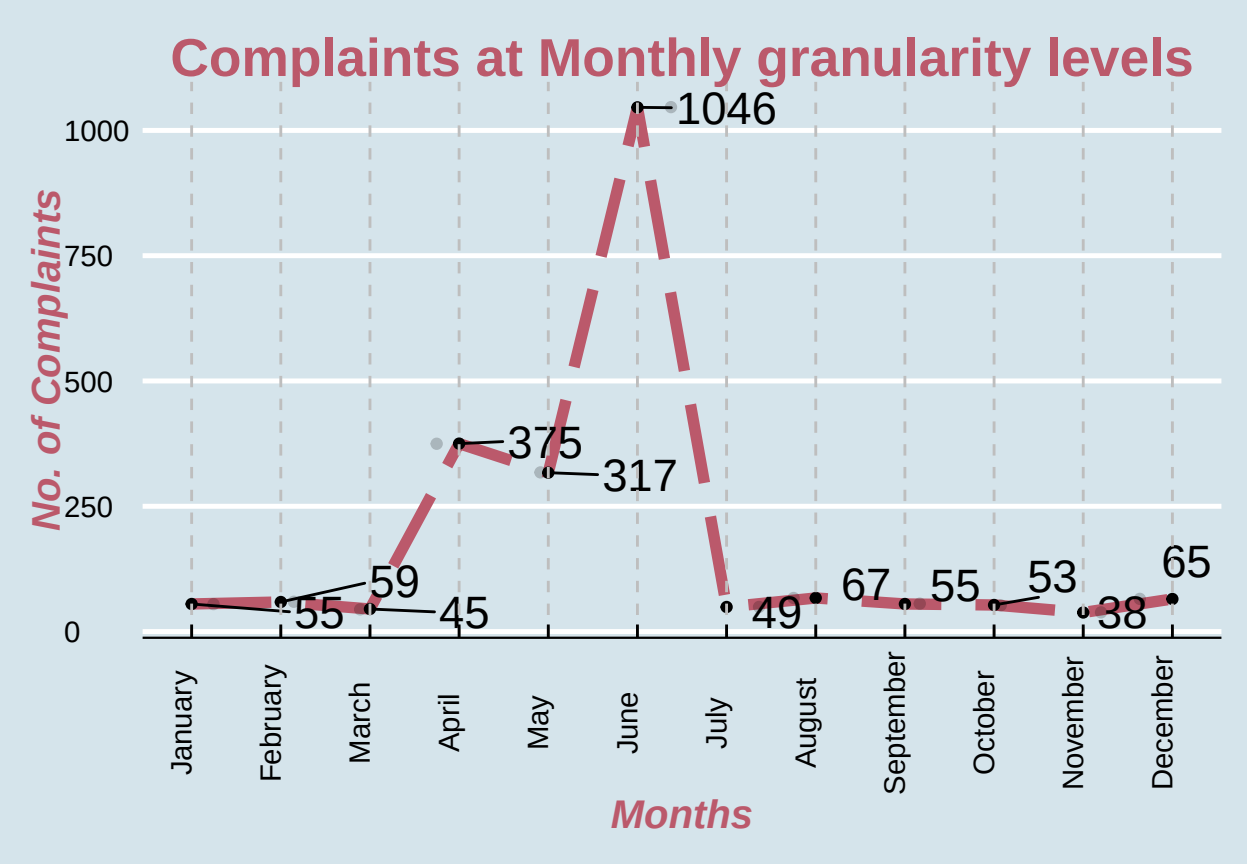


Through the above plot, it can be seen that max no. of complaints on a single day were 218 and were made on 24th June 2015.

=> Even though the above plot gives decent overview on daily granularity levels, but even then its not going to be easy to estimate monthly complaints from the same. So, generating a plot for Complaints at monthly granularity levels.

```
Month_Count <-
  Complaints %>%
  group_by(MonthsName = as.integer(month(Date)))

Month_Count %>%  group_by(MonthsName) %>%
  summarize(No_of_Complaints = length(MonthsName)) %>%
  ggplot(aes(MonthsName, No_of_Complaints, label = No_of_Complaints)) +
  geom_line(color = "#bb596b", size = 2, lty = 5) +  geom_point() +
  geom_text_repel(nudge_x = 1, size = 6) + geom_jitter(alpha = 0.2) +
  scale_x_continuous(breaks = c(1:12),
                     labels = c("January", "February", "March", "April", "May",
                                "June", "July", "August", "September", "October",
                                "November", "December")) +
  theme_economist() + xlab("Months") + ylab("No. of Complaints") +
  theme(axis.text.x = element_text(angle = 90, hjust = 0.5)) +
```

```
  ggtitle("Complaints at Monthly granularity levels") +
  theme(plot.title = element_text(hjust = 0.5, color = "#bb596b", face = "bold",
                                  size = 20)) +
  theme(axis.title.x = element_text(size = 15, color = "#bb596b",
                                    face = "bold.italic")) +
  theme(axis.title.y = element_text(size = 15, color = "#bb596b",
                                    face = "bold.italic")) +
  geom_vline(xintercept = c(1:12), lty = 2, color = "grey")
```



Through this, it can be intercepted that highest number of complaints were done in June, whereas the lowest number of complaints were filed in November.

As it can be seen from the data, there are in total 2224 customer complaints. We need to see, which types of words have most frequently been used.

```
Complaints %>%
  select(Customer.Complaint) %>%
  mutate(Customer.Complaint = removePunctuation(Customer.Complaint)) %>%
  mutate(Customer.Complaint = tolower(Customer.Complaint)) %>%
  mutate(Customer.Complaint = stripWhitespace(Customer.Complaint)) %>%
  unnest_tokens(word, Customer.Complaint) %>%
  count(word) %>% arrange(desc(n)) %>%
  filter(!word %in% stop_words$word) %>%
```

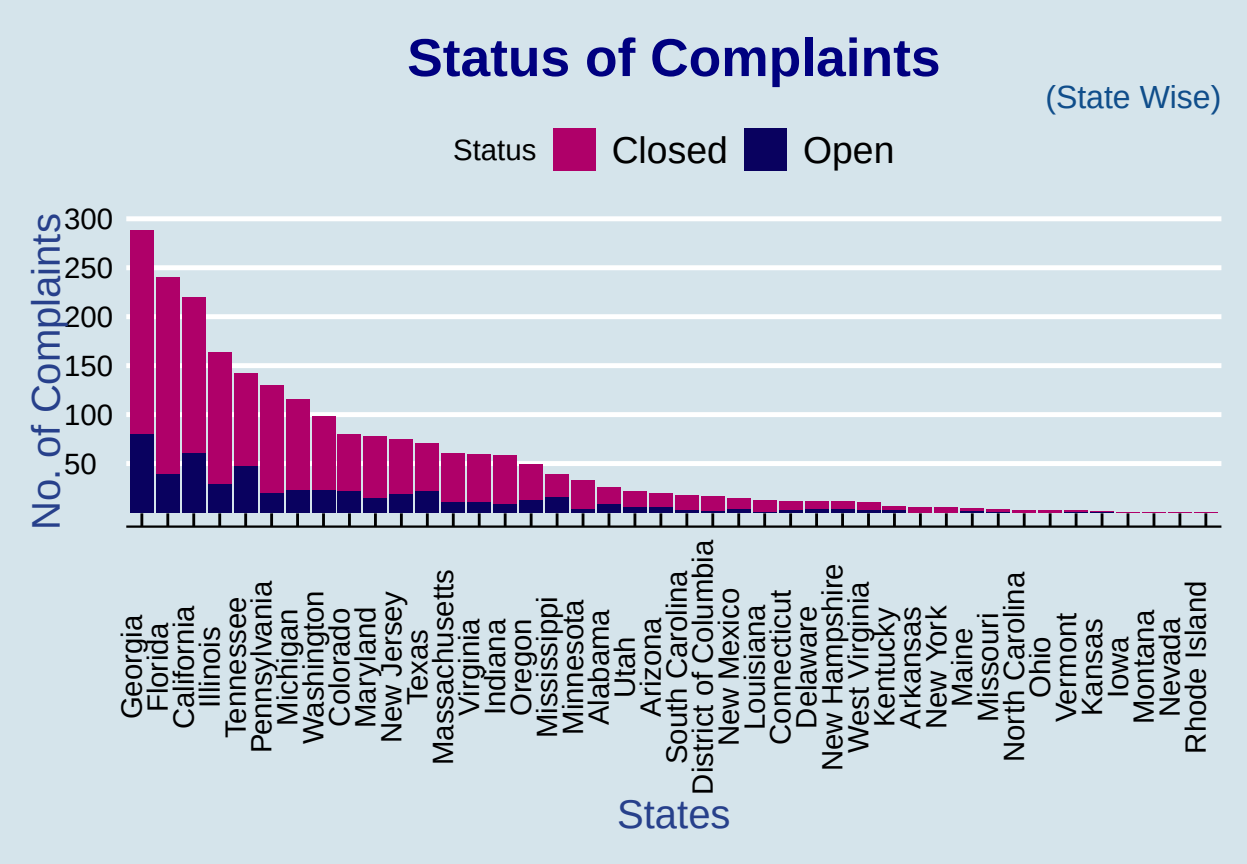
```
  filter(word != "comcast") %>%
  head(20)
```

```
##          word   n
## 1    internet 517
## 2     service 444
## 3     billing 281
## 4        data 219
## 5       speed 116
## 6         cap 103
## 7      issues  90
## 8    customer  88
## 9        caps  78
## 10  practices  73
## 11    charges  71
## 12     speeds  71
## 13  complaint  70
## 14       bill  64
## 15    xfinity  62
## 16     unfair  58
## 17 throttling  53
## 18   services  52
## 19      cable  50
## 20       slow  49
```

It can be easily seen that internet has been the most used word for complaints followed by service, billing, data and speed. Through this its safe to assume that lots of people have issue with Internet and its related aspects.

=> To see the Status of Complaints in all the states. Below code/plot.

```
Complaints %>%
  ggplot(aes(x = fct_infreq(State), fill = Status)) +
  geom_bar(stat = 'count') +
  scale_y_continuous(breaks = c(50,100,150,200,250,300)) +
  xlab("States") +ylab("No. of Complaints") + theme_economist() +
  theme(axis.text.x = element_text(hjust = 0.5, angle = 90)) +
  ggtitle("Status of Complaints", subtitle = "(State Wise)") +
  theme(plot.title = element_text(hjust = 0.5, color = "navy", size = 20)) +
  theme(plot.subtitle = element_text(hjust = 1, color = "dodgerblue4",
                                     size = 12)) +
  theme(axis.title.x = element_text(color = "royalblue4", size = 15)) +
  theme(axis.title.y = element_text(color = "royalblue4", size = 15)) +
  scale_fill_manual(values = c("#af0069","#09015f"))
```



This plot clearly shows that Georgia has the highest number of combined complaints. But doesn't portray properly, on which states has the minimum no. of complaints and how many complaints does top 10 in the list have.

```
Complaints %>%
  group_by(State) %>%
  summarise(length(State)) %>%
  arrange(`length(State)`) %>%
  top_n(10)
```

```
## Selecting by length(State)
```

```
## # A tibble: 10 x 2
##    State        `length(State)`
##    <fct>                  <int>
##  1 Maryland                  78
##  2 Colorado                  80
##  3 Washington                98
##  4 Michigan                 115
##  5 Pennsylvania             130
##  6 Tennessee                143
##  7 Illinois                 164
##  8 California               220
```

```
##  9 Florida                  240
## 10 Georgia                  288
```

```
Complaints %>%
  group_by(State) %>%
  summarise(length(State)) %>%
  arrange(`length(State)`) %>%
  top_n(-10)
```

```
## Selecting by length(State)
```

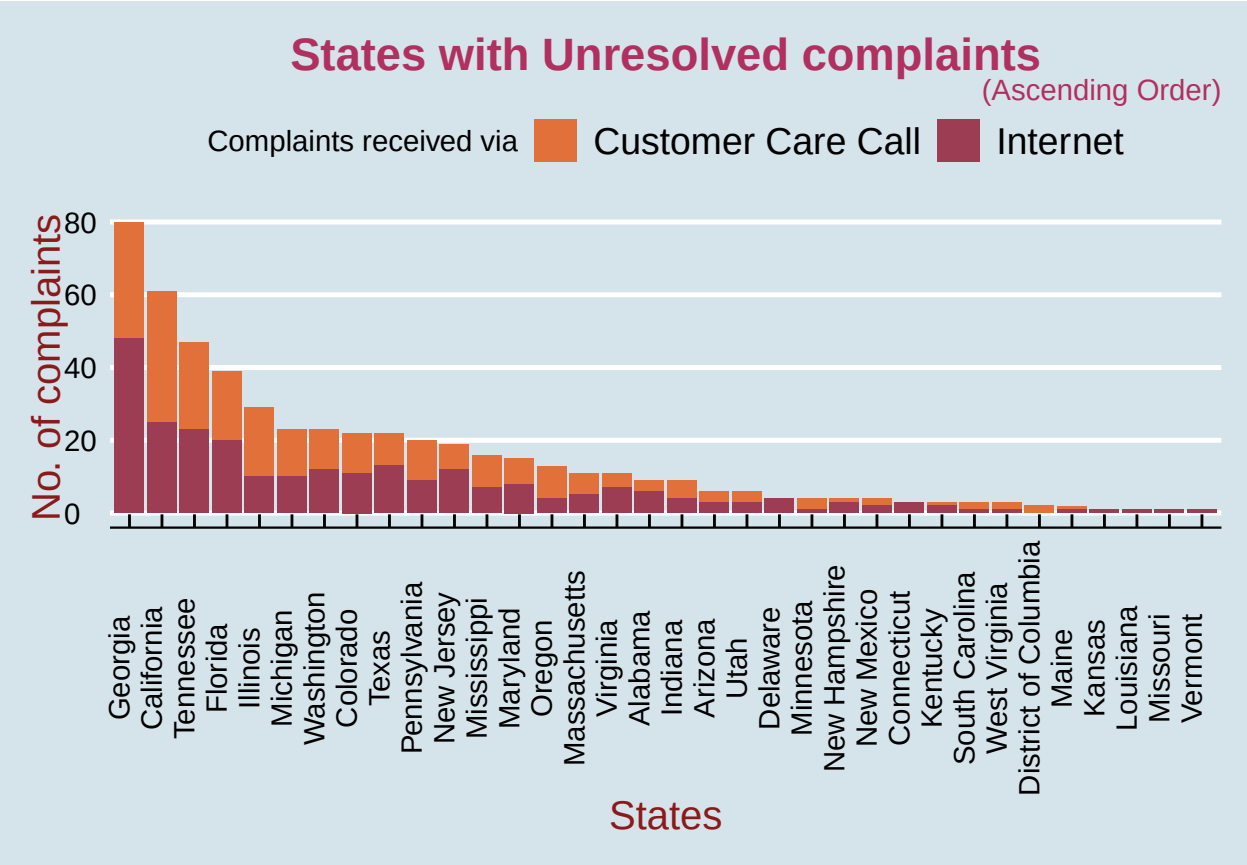
```
## # A tibble: 10 x 2
##    State          `length(State)`
##    <fct>                    <int>
##  1 Iowa                         1
##  2 Montana                      1
##  3 Nevada                       1
##  4 Rhode Island                 1
##  5 Kansas                       2
##  6 North Carolina               3
##  7 Ohio                         3
##  8 Vermont                      3
##  9 Missouri                     4
## 10 Maine                        5
```

Through above codes, we can easily say that minimum number of complaints made in any state is 1, and this has been achieved by 4 states. Which are Iowa, Montana, Nevada and Rhode Island. And, maximum no. of complaints have been done in Georgia. A total of 288 complaints.

```
str(Complaints)
```

```
## 'data.frame':    2224 obs. of  10 variables:
##  $ Ticket..                   : chr  "250635" "223441" "242732" "277946" ...
##  $ Customer.Complaint         : chr  "Comcast Cable Internet Speeds" "Payment disappear - service go
##  $ Date                       : Date, format: "2015-04-22" "2015-08-04" ...
##  $ Time                       :Formal class 'Period' [package "lubridate"] with 6 slots
##   .. ..@ .Data : num  50 56 47 35 26 40 55 14 30 31 ...
##   .. ..@ year  : num  0 0 0 0 0 0 0 0 0 0 ...
##   .. ..@ month : num  0 0 0 0 0 0 0 0 0 0 ...
##   .. ..@ day   : num  0 0 0 0 0 0 0 0 0 0 ...
##   .. ..@ hour  : num  3 10 9 11 1 9 10 6 11 6 ...
##   .. ..@ minute: num  53 22 55 59 25 59 13 56 46 46 ...
##  $ Received.Via               : Factor w/ 2 levels "Customer Care Call",..: 1 2 2 2 2 2 1 2 1 1 ...
##  $ City                       : Factor w/ 928 levels "Abingdon","Acworth",..: 1 2 2 2 2 2 2 3 4 4 ..
##  $ State                      : Factor w/ 42 levels "Alabama","Arizona",..: 18 10 10 10 10 10 10 20 4
##  $ Zip.code                   : int  21009 30102 30101 30101 30101 30101 30101 49221 94502 94501 ...
##  $ Status                     : Factor w/ 2 levels "Closed","Open": 1 1 1 2 1 1 2 1 1 2 ...
##  $ Filing.on.Behalf.of.Someone: chr  "No" "No" "Yes" "Yes" ...
```

```
Complaints %>%
  filter(Status == "Open") %>%
  ggplot(aes(x = fct_infreq(State), fill = Received.Via)) +
  geom_bar() + theme_economist() +
  theme(axis.text.x = element_text(hjust = 0.5, angle = 90)) +
  xlab("States") + ylab("No. of complaints") +
  theme(axis.title.x = element_text(color = "firebrick4", size = 15)) +
  theme(axis.title.y = element_text(color = "firebrick4", size = 15)) +
  ggtitle("States with Unresolved complaints",
          subtitle = "(Ascending Order)") +
  theme(plot.title = element_text(hjust = 0.5, color = "maroon")) +
  theme(plot.subtitle = element_text(hjust = 1, color = "maroon")) +
  scale_fill_manual(name = "Complaints received via",
                    values = c("#e2703a", "#9c3d54"))
```

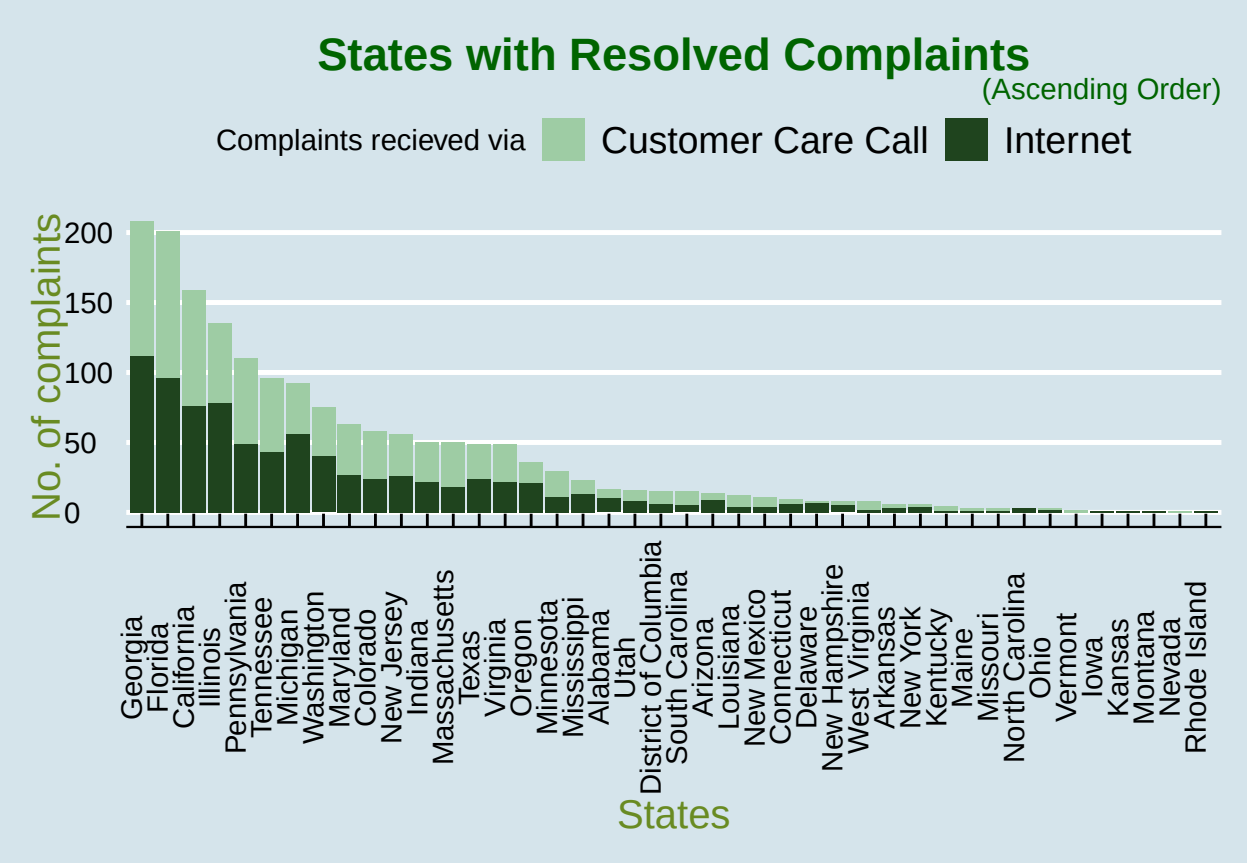


Above plot shows States having unresolved complaints in descending order. This shows that Georgia has the highest no. of unresolved complaints.

=> Now, to see the states with resolved complaints.

```
Complaints %>%
  filter(Status == "Closed") %>%
  ggplot(aes(x = fct_infreq(State), fill = Received.Via)) +
```

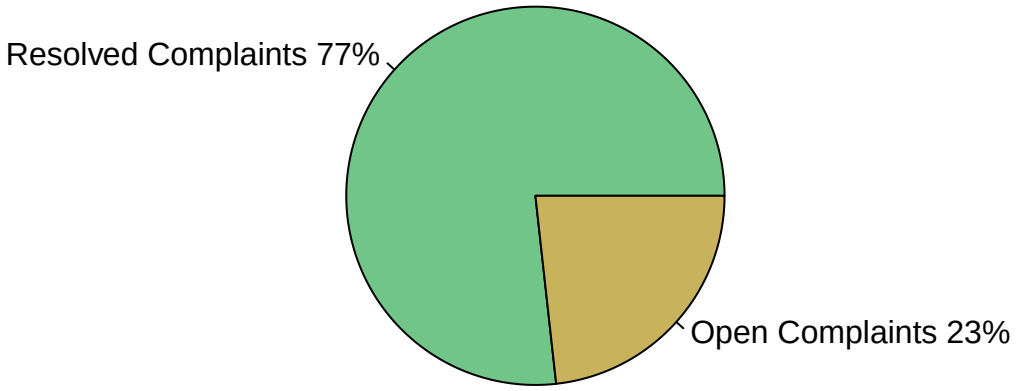
```
  geom_bar() + theme_economist() +
  theme(axis.text.x = element_text(hjust = 0.5, angle = 90)) +
  xlab("States") + ylab("No. of complaints") +
  theme(axis.title.x = element_text(color = "olivedrab4", size = 15)) +
  theme(axis.title.y = element_text(color = "olivedrab4", size = 15)) +
  ggtitle("States with Resolved Complaints",
          subtitle = "(Ascending Order)") +
  theme(plot.title = element_text(hjust = 0.5, color = "dark green")) +
  theme(plot.subtitle = element_text(hjust = 1, color = "dark green")) +
  scale_fill_manual(name = "Complaints recieved via",
                    values = c("#9ecca4","#1f441e" ))
```



This shows that Georgia has the highest no. of Resolved complaints as well.

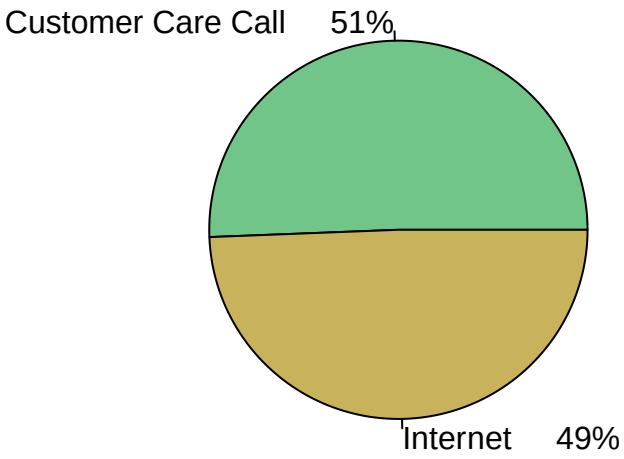
=> Now, to see the percentage of complaints resolved till date.

```
p <- Complaints %>% group_by(Status) %>% summarize(length(Status))

pie(p$`length(Status)`,
    labels = paste0(c("Resolved Complaints", "Open Complaints"),
                    " ", round(c((1707/2224)*100, (517/2224)*100)), "%"),
    col = c("#72C588", "#C8B35C"))
```



Through this we can see, that in the given time frame, 77% of the complaints were resolved, whereas 23% of the complaints are still unresolved.

```
Complaints_received_via <-
  Complaints %>% filter(Status == "Closed") %>%
  group_by(Received.Via) %>%
  summarize(length(Received.Via))

pie(Complaints_received_via$`length(Received.Via)`,
      labels = paste0(c("Customer Care Call", "Internet"),
                      "     ", round(c((864/1707)*100, (843/1707)*100)),
                      "%"), col = c("#72C588", "#C8B35C"))
```



And, out of those resolved cases 51% have been from customer care call and 49% were from internet.