###########################################

###########################################

### Comcast Telecom Consumer Analysis ###

### ###

### By: Sushrut Kalyan Patnaik ###

###########################################

###########################################

# 1. Import data into R environment:

# Install the package if not not available.

# install.packages("dplyr", dependencies = TRUE)

# install.packages("ggplot2", dependencies = TRUE)

# install.packages("stringi", dependencies = TRUE)

# install.packages("lubridate", dependencies = TRUE)

# install.packages("tidyverse", dependencies = TRUE)

# install.packages("plotrix", dependencies = TRUE)

# install.packages("wordcloud", dependencies = TRUE)

# install.packages("tm", dependencies = TRUE)

# Use the required library/packages for your code.

library(dplyr) # Data frame manipulation

library(ggplot2) # Data visualization package

library(stringi) # String/text/natural language processing

library(tidyverse) # Collection of required Data Science R packages

library(lubridate) # Deal with Date-Time data

library(plotrix) # 3D Pie Chart

library(wordcloud)

library(tm)

# load the data set

# comcast\_data = read.csv("C:/Comcast Telecom Complaints data.csv") # One way

# or

# You can read you source file from anywhere from you PC. I prefer to use below.

comcast\_data = read.csv(choose.files()) #File: Comcast Telecom Complaints data.csv

# View sample or complete data and Structure of the data set

head(comcast\_data) # First 5 record from data set

tail(comcast\_data) # Last 5 records from data set

View(comcast\_data) # Complete data set like Tablular view

str(comcast\_data) # Structure of the data set

# Check if there is any missing data point

any(is.na(comcast\_data)) # False is NA/blank is available and vice versa

# Convert the DATE to one format

comcast\_data$Date = dmy(comcast\_data$Date)

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

# Filter for observations daily

comcast\_daily = dplyr::summarise(group\_by(comcast\_data,Date), Count = n())

View(comcast\_daily)

# Filter for observations monthly

comcast\_monthly = dplyr::summarise(group\_by(comcast\_data, Month = as.integer(month(Date))), Count = n())

comcast\_monthly.Total = arrange(comcast\_monthly, Month)

View(comcast\_monthly)

# Renaming the months to factor

month.name = c("Jan", "Feb", "Mar",

"Apr", "May", "Jun",

"Jul", "Aug", "Sep",

"Oct", "Nov", "Dec")

comcast\_monthly$Month = month.name[comcast\_monthly$Month]

comcast\_monthly$Month <- as.character(comcast\_monthly$Month)

comcast\_monthly$Month <- factor(comcast\_monthly$Month , levels = comcast\_monthly$Month )

# Display the monthly complaints

View(comcast\_monthly)

# Plotting the number of complaints per day

ggplot\_daily = ggplot(comcast\_daily, aes(x = comcast\_daily$Date, y = comcast\_daily$Count)) +

geom\_point(col = "red", size = 1.5) +

geom\_line(col = 'blue', linetype = "dashed", size = .75) +

xlab("Date")+

ylab("No. of Complaints") +

ggtitle("Number of complaints per Day") +

theme(plot.title = element\_text(hjust = 0.5))

# Plotting the number of complaints per Month

ggplot\_monthly = ggplot(comcast\_monthly, aes(x = comcast\_monthly$Month, y = comcast\_monthly$Count, group = 1)) +

geom\_point(col = "red", size = 1.5) +

geom\_line(col = 'blue', linetype = "dashed", size = .75) +

xlab("Months")+

ylab("No. of Complaints") +

ggtitle("Number of complaints per Month") +

theme(plot.title = element\_text(hjust = 0.5))

# 3. Provide a table with the frequency of complaint types:

# Find complaint type from "Customer Complaint"

# Here ignore.case If TRUE, the default, ignores case when matching names.

names(comcast\_data)<-gsub(pattern = '\\.',replacement = "",x=names(comcast\_data))

names(comcast\_data)

network\_complaint = contains(comcast\_data$CustomerComplaint, match = 'network', ignore.case = T)

internet\_complaint = contains(comcast\_data$CustomerComplaint, match = 'internet', ignore.case = T)

bill\_complaint = contains(comcast\_data$CustomerComplaint, match = 'bill', ignore.case = T)

email\_complaint = contains(comcast\_data$CustomerComplaint, match = 'email', ignore.case = T)

charge\_complaint = contains(comcast\_data$CustomerComplaint, match = 'charge', ignore.case = T)

speed\_complaint = contains(comcast\_data$CustomerComplaint, match = 'speed', ignore.case = T)

#datacap\_complaint = contains(comcast\_data$CustomerComplaint, match = 'data cap', ignore.case = T)

data\_complaint = contains(comcast\_data$CustomerComplaint, match = 'data', ignore.case = T)

comcast\_data$ComplaintType[network\_complaint] = "Network"

comcast\_data$ComplaintType[internet\_complaint] = "Internet"

comcast\_data$ComplaintType[bill\_complaint] = "Bill"

comcast\_data$ComplaintType[email\_complaint] = "e-Mail"

comcast\_data$ComplaintType[charge\_complaint] = "Charge"

comcast\_data$ComplaintType[speed\_complaint] = "Speed"

#comcast\_data$ComplaintType[datacap\_complaint] = "Data Cap"

comcast\_data$ComplaintType[data\_complaint] = "Data"

comcast\_data$ComplaintType[-c(network\_complaint,internet\_complaint,bill\_complaint,

email\_complaint,charge\_complaint,speed\_complaint,

data\_complaint)] = "Others"

complaint\_frequency = table(comcast\_data$ComplaintType)

complaint\_frequency

View(complaint\_frequency)

freq = c(348,122,219,15,355,2,973,190)

lab = c("Bill","Charge","Data","e-Mail","Internet", "Network", "Others", "Speed")

par(mfrow = c(1,2))

pie3D(complaint\_frequency,

labels = complaint\_frequency,

radius = 1,

height = 0.1,

border = "white",

explode = 0.2,

main = "Pie Chart of Frequency of complaint types")+

theme(plot.title = element\_text(hjust = 0.5))

pie3D(freq,

labels = lab,

radius = 1,

height = 0.1,

border = "white",

explode = 0.2)

# From the above table we can see that the Others and 2nd largest is Internet type complaints are maximum.

#4. Which complaint types are maximum i.e., around internet, network issues, or

# across any other domains:

names(comcast\_data)

# To achieve this goal we need to eliminate duplicate CustomerComplaint

ComplaintType\_lower = comcast\_data %>%

mutate(tolower(CustomerComplaint))

ComplaintType = table(ComplaintType\_lower$CustomerComplaint)

ComplaintType = data.frame(ComplaintType)

names(ComplaintType)

ComplaintType\_filter = ComplaintType %>%

rename(Complaint\_Type = Var1, Frequency = Freq)

most\_freq = ComplaintType\_filter %>% arrange(desc(Frequency))

most\_freq

wordcloud::wordcloud(words = most\_freq$Complaint\_Type, freq = most\_freq$Frequency, min.freq = 3,

max.words =300, random.order = TRUE, colors=brewer.pal(8, "Dark2"), rot.per=0.55)

## 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.

##### i. Which state has the maximum complaints

names(comcast\_data)

str(comcast\_data)

unique(comcast\_data$Status)

open <- ( comcast\_data$Status == "Open"| comcast\_data$Status =="Pending")

closed <- (comcast\_data$Status == "Closed"| comcast\_data$Status =="Solved")

comcast\_data$ComplaintStatus[open] <-"Open"

comcast\_data$ComplaintStatus[closed]<- "Closed"

comcast\_data\_1 <- group\_by(comcast\_data,State,ComplaintStatus)

status\_data<- dplyr::summarise(comcast\_data\_1,Count = n())

View(chart\_data)

# plot the Ticket Status Distribution per States chart

p1 <- ggplot(status\_data, aes(x =status\_data$State, y = status\_data$Count)) +

geom\_bar(stat="identity", color= "#60ff20", fill= "black", width = .8)+

#geom\_text(aes(y = status\_data$Count,

#label=status\_data$Count,

#group=status\_data$ComplaintStatus), vjust = 1, nudge\_y = 40,

#angle = 45, col = "blue", size = 4)+

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 = "red"),

plot.title = element\_text(hjust = 0.5))+

labs(title = "Total Number of Ticket Status Distribution per States",

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

p1

##### ii. Which state has the highest percentage of unresolved complaints

State\_by\_Complian = comcast\_data %>% filter(ComplaintStatus == "Open") %>%

group\_by(State) %>% summarise(NumberOfComplian = n())

str(State\_by\_Complian)

ggplot(State\_by\_Complian, aes(x = State\_by\_Complian$State, y = State\_by\_Complian$NumberOfComplian, group = 1))+

geom\_point(col = "red", size = 1.5) +

geom\_line(col = 'blue', linetype = "dashed", size = .75) +

xlab("State")+

ylab("No. of unresolved complaints") +

ggtitle("Highest % of unresolved complaints by State") +

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 = "red"),

plot.title = element\_text(hjust = 0.5))

# From the table generated above we can see that Georgia has maximum unresolved complaints

#5. Provide the percentage of complaints resolved till date, which were received through

# the Internet and customer care calls:

cs\_total = comcast\_data %>%

group\_by(comcast\_data$ComplaintStatus) %>%

summarise(NumOfComplaintStatus = n())

cs\_total

A = cs\_total$NumOfComplaintStatus

Percentage = round((A/sum(A))\*100,1)

lbl = paste(cs\_total$`comcast\_data$ComplaintStatus`," ", Percentage,"%", sep = " ")

pie3D(A, labels = lbl,

height = 0.1,

border = "white",

explode = 0.1,

main = "Pie Chart of Complaints resolved v/s unresolved")+

theme(plot.title = element\_text(hjust = 0.5))

# Pie chart we can clearly see that there are total 76.8% Complaints resolved.

unique(comcast\_data$ReceivedVia)

unique(comcast\_data$ComplaintStatus)

Internet = comcast\_data %>%

filter(comcast\_data$ReceivedVia == 'Internet', comcast\_data$ComplaintStatus == 'Closed') %>%

summarise(NumOfComplaints = n())

Internet\_Percent = round((Internet$NumOfComplaints/sum(cs\_total$NumOfComplaintStatus)\*100),1)

CCC = comcast\_data %>%

filter(comcast\_data$ReceivedVia == 'Customer Care Call', comcast\_data$ComplaintStatus == 'Closed') %>%

summarise(NumOfComplaints = n())

CCC\_Percent = round((CCC$NumOfComplaints/sum(cs\_total$NumOfComplaintStatus)\*100),1)

# From 76.75% resolved Complaints, 37.9% complaints are Internet type while 38.8% are Customer Care Call type.