**Public Awareness about Cybersecurity**

****

**Sanchita Ajay Gawand**

**MS: Information Systems**

**California State University Los Angeles**

**CIS 5270: Business Intelligence**

**Submitted to: Dr. Shilpa Balan**

**Table of Contents:**

|  |  |  |
| --- | --- | --- |
|  | **Topic** | **Pg No** |
| 1 | **Objective of Study** | 3 |
| 2 | **Dataset** | 4 |
| 3 | **Data cleaning** | 5 |
| 4 | **Analysis and Visualizations** | 14 |
| 5 | **Statistics** | 34 |
| 6 | **User Defined Functions** | 37 |
| 7 | **Data cleaning codes** | 46 |
| 8 | **Visualization codes** | 53 |
| 9 | **Statistical/user defined codes** | 60 |
| 10 | **References** | 62 |

**OBJECTIVE OF STUDY**

There is no denial that cyber security and its part in protecting the data and privacy is in the headlines every week. However, the average American’s understanding and paying heed to this news, plays a vital role in protecting their information, social content and devices at home and work. Multiple surveys have shown reported that half respondents have been a victim of identity thefts and nearly half of them have been a victim of some sort of a phishing attack. Hence its very crucial to understand cyber security and its influence of day to day life. An awareness needs to be spread among people and organizations as the adoption of smart devices and social networking are at its pinnacle.

**DATASET**

**Data set URLs:**

<http://www.pewinternet.org/dataset/june-2016-cybersecurity-knowledge/>

**Data set file format: CSV file**

About the dataset:

This dataset is about survey conducted by Pew research center titled, “What the Public Knows about Cybersecurity” which reflects the survey of 1,055 adults last year on their understanding of online security and the results were very mixed. This survey asked thirteen questions about cyber security.Dataset has more than 25 columns and 1055 rows.

**Internet –** How often people use internet?

**https-** What does https mean?

**PhishingAttack-** How many people are aware of Phishing Attack?

**Computersnetworked-** What is group of computers gathered together called?

**Wifi-** Is Wifi encrypted by default?

**Authentication-** Identify secured passowrds

**CriminalAcess-** How do criminals access personal files?

**PrivateBrowsing**

**GPS-** Does GPS trcaks the location?

**Email-** Are emails encrypted by default?

**Hacking Tool-** Which is the hacking tool used by hackers?

**VPN-** Virtual Private Network

**INTFREQ-** Frequency of usage

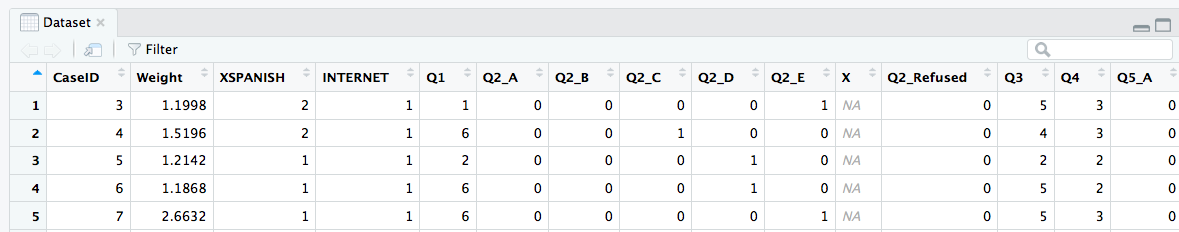
**PPGENDER-** Gender

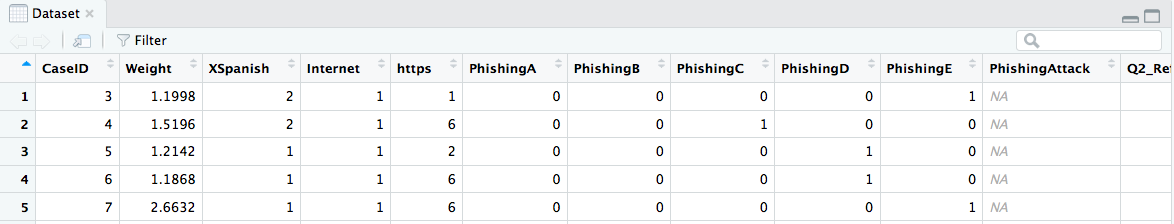
**DATA CLEANING**

**Data Cleaning 1: Renaming Column Names**

All columns were renamed.

**Before:**

**After:**



**Code Screenshot:**



**Code:**

> setwd("~/Desktop")

> Dataset<-read.csv("CybersecurityKnowledgeDS.csv",header=T,sep=",")

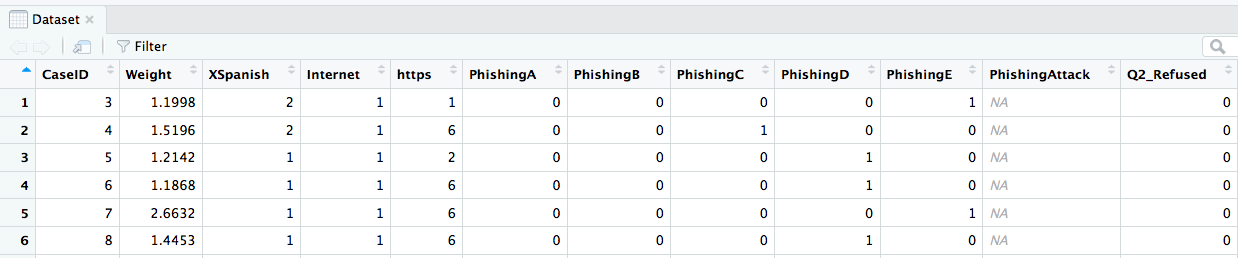
> View(Dataset)

>names(Dataset)<-c("CaseID","Weight","XSpanish","Internet","https","PhishingA","PhishingB","PhishingC","PhishingD","PhishingE","PhishingAttack","Q2\_Refused","Computersnetworked","Wifi","AuthenticationA","AuthenticationB","AuthenticationC","AuthenticationD","AuthenticationE","AuthenticationF","Authentication","Q5\_Refused","CriminalAcess","PrivateBrowsing","GPS","Email","FreeCredit","Wifi","Risks","INTMOB","INTFREQ","SMART1","SNSINT2","PPAGE","PPEDUCAT","PPETHM","PPGENDER","PPINCMP","PPMSACAT","PPSTATEN","PPWORK")

**DATA CLEANING 2: Combining columns**

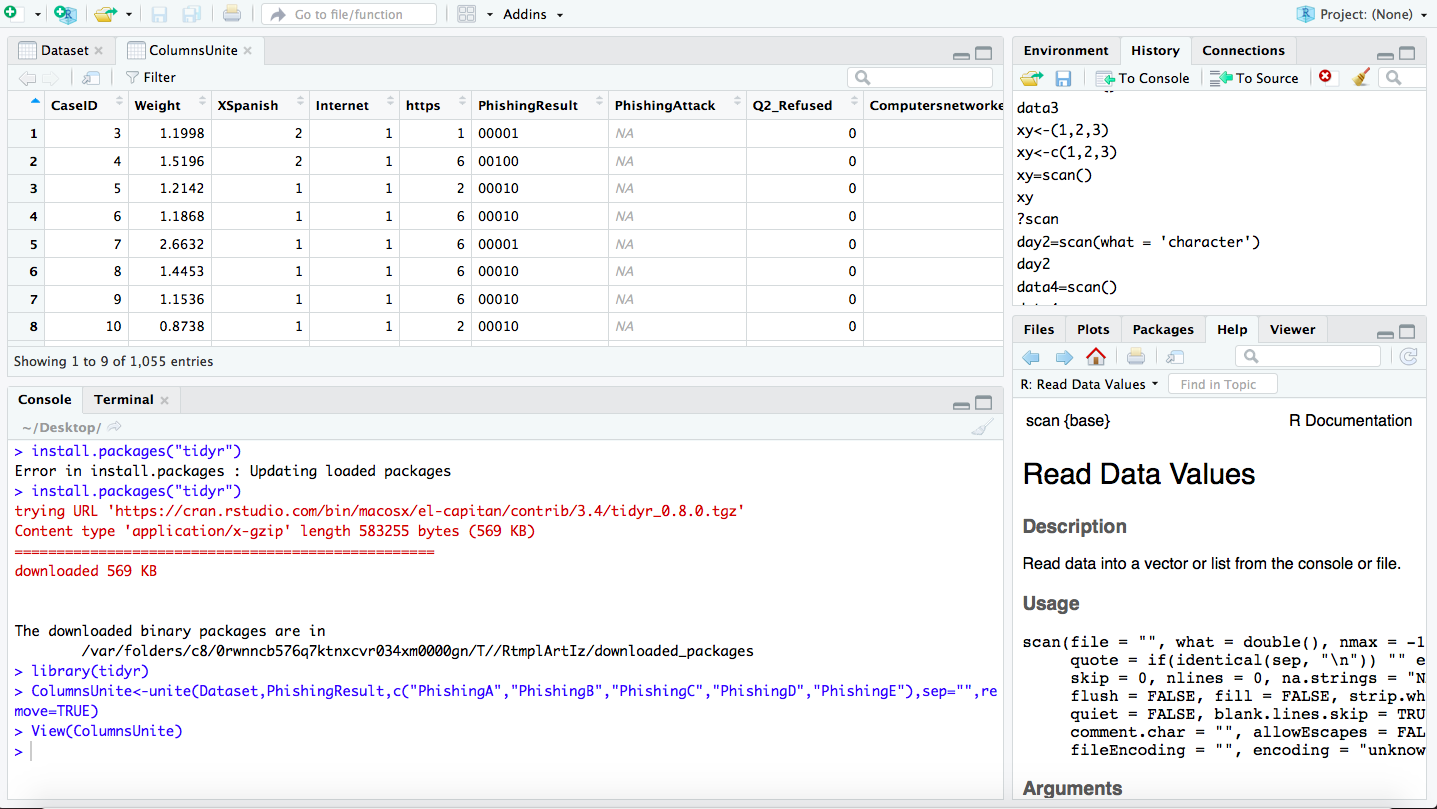
Columns PhishingA, PhishingB, PhishingC, PhishingD, PhishingE were combined to phishing Result

**Before:**



**After:**



**CodeScreenshot:** 

**Code:**

> install.packages("tidyr")

Error in install.packages : Updating loaded packages

> install.packages("tidyr")

trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.4/tidyr\_0.8.0.tgz'

Content type 'application/x-gzip' length 583255 bytes (569 KB)

==================================================

downloaded 569 KB

The downloaded binary packages are in

/var/folders/c8/0rwnncb576q7ktnxcvr034xm0000gn/T//RtmplArtIz/downloaded\_packages

> library(tidyr)

>ColumnsUnite<-unite(Dataset,PhishingResult,c("PhishingA","PhishingB","PhishingC","PhishingD","PhishingE"),sep="",remove=TRUE)

>View(ColumnsUnite)

>ColumnsUnite1<-unite(Dataset,AuthenticationResult,c("AuthenticationA","AuthenticationB","AuthenticationC","AuthenticationD","AuthenticationE","AuthenticationF"),sep="",remove=TRUE)

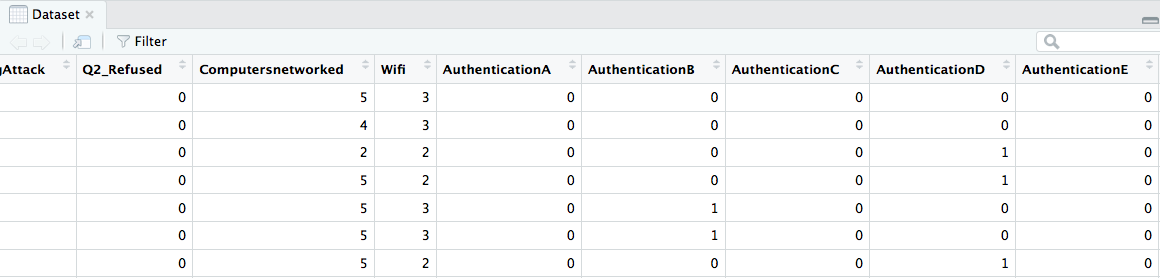
> ColumnsUnite1

**DATA CLEANING 3: Removing Irrelevant Columns**

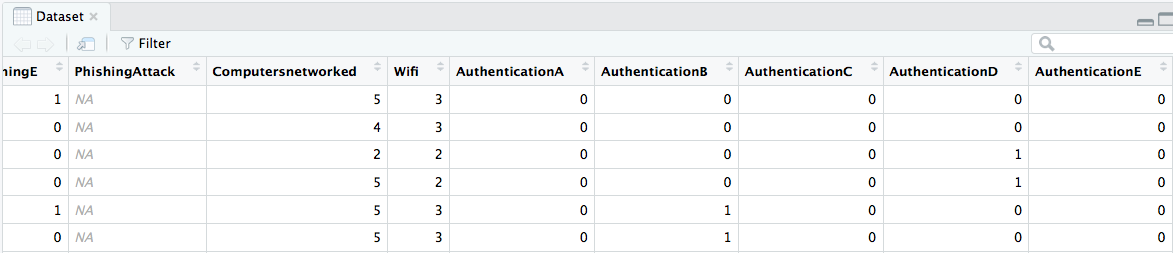
Column Q2\_Refused was removed.

This Column was Removed

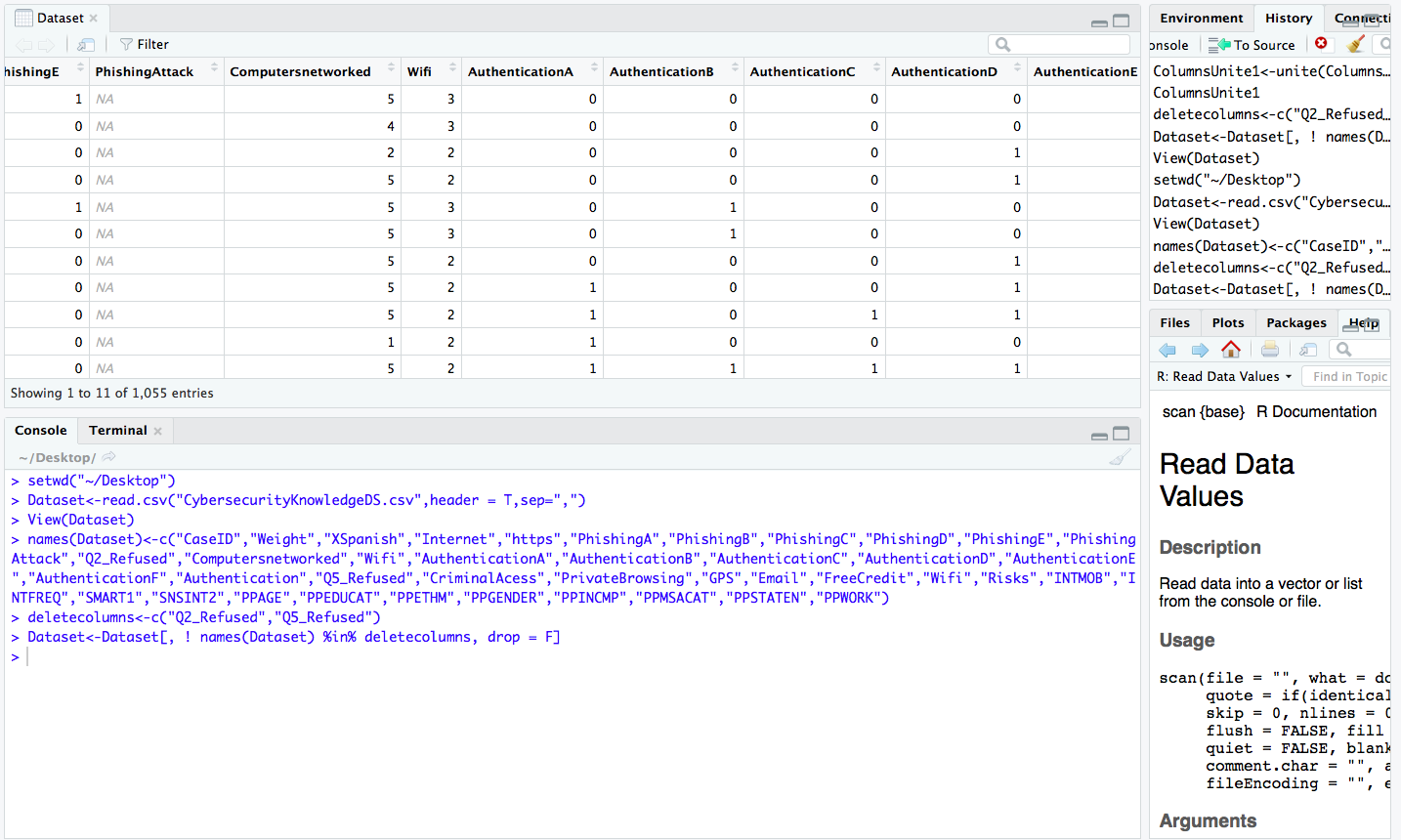
**Before:**

****

**After:**

****

**Code Screenshot:**



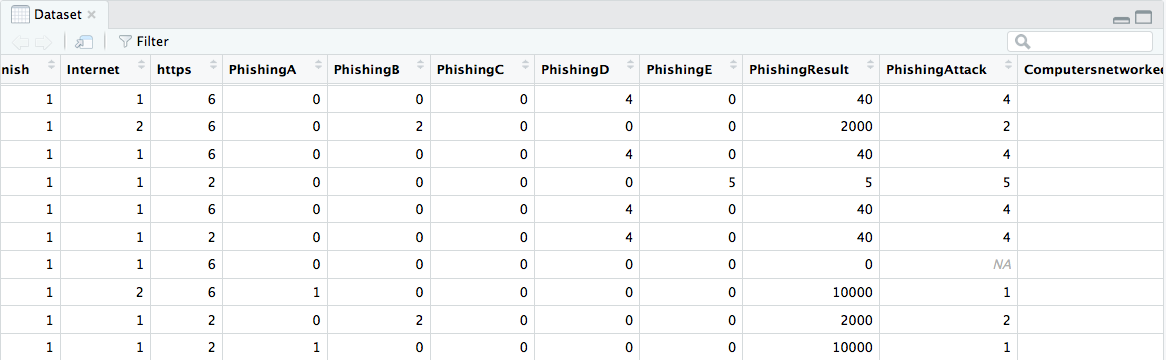
**Code:**

> deletecolumns<-c("Q2\_Refused","Q5\_Refused")

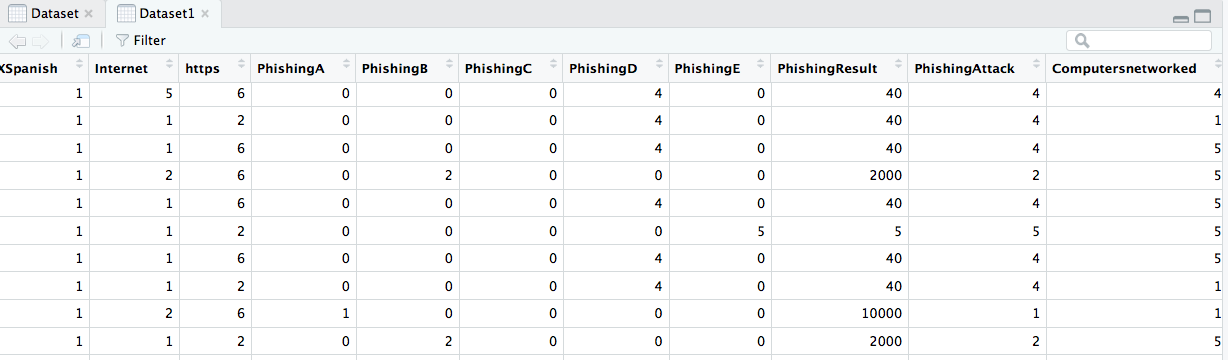
> Dataset<-Dataset[, ! names(Dataset) %in% deletecolumns, drop = F]

**DATA CLEANING 4: Removing Rows where values are not available (NA)**

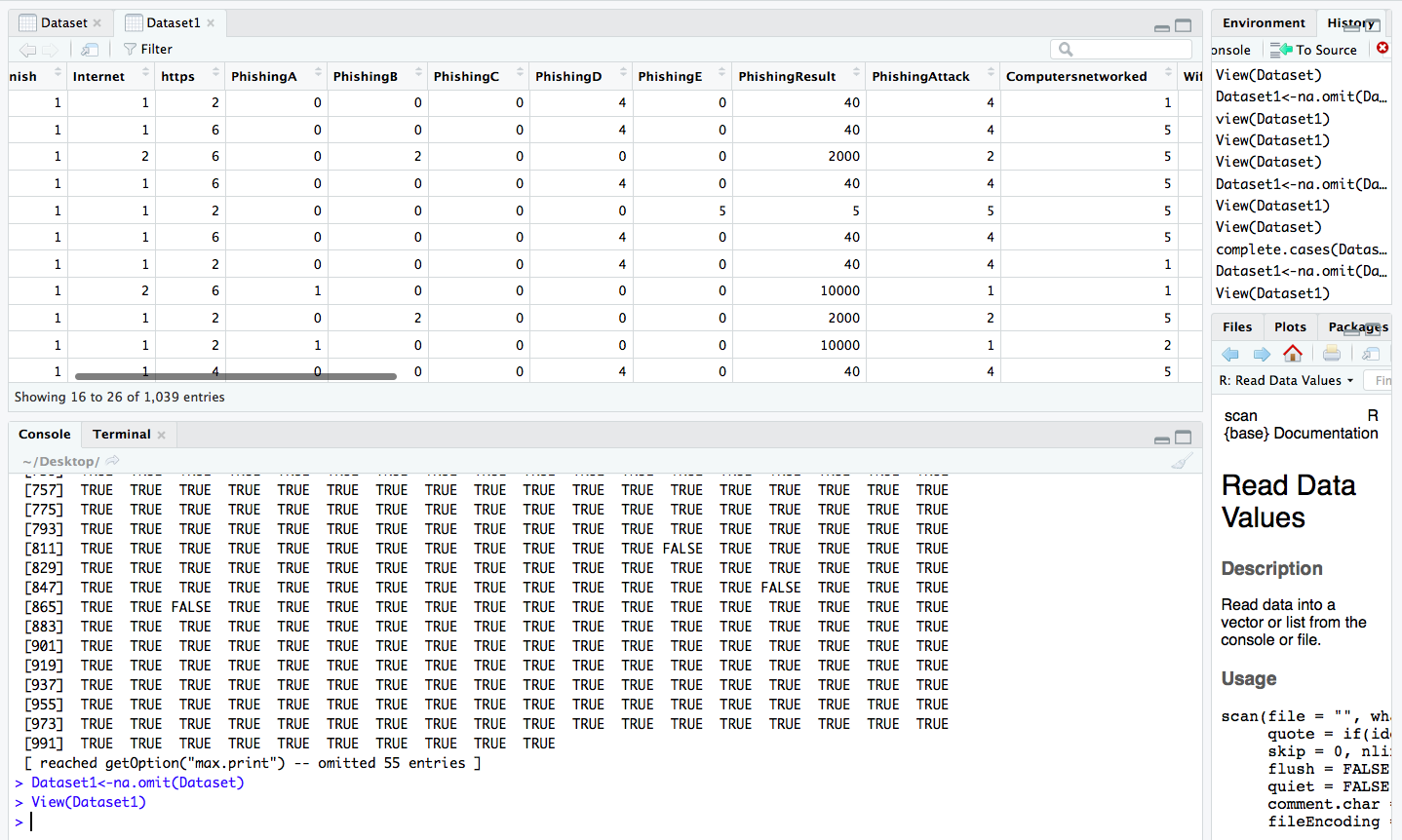
**Before Cleaning:**

****

**After Cleaning:**

****

**Code Screenshot:**

****

**Code:**

> View(Dataset)

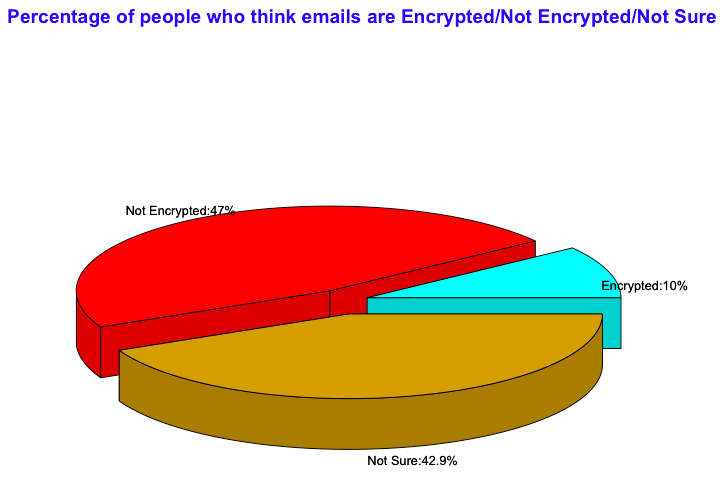
> complete.cases(Dataset)

[ reached getOption("max.print") -- omitted 55 entries ]

> Dataset1<-na.omit(Dataset)

> View(Dataset1)

**ANALYSIS/VISUALIZATIONS**  
**Q.1. Percentage of people who think emails are encrypted by default?**



**Visualization:** Pie chart 3D

**Functions Used:** Colors, Pie3d.labels, title, table, round, paste

**ANALYSIS/DESCRIPTION:**

As per the Pie chart above, people who took survey, were asked if emails are encrypted by default. Around 47% of the crowd feel that emails are not encrypted, 10% feel that emails are encrypted by default and 42.9% do not about these. Although we use emails on daily basis these days, it has been seen that more than 60 % do not how secure they are to use considering its confidentiality. This is a matter of concern for safety issues and awareness should be spread. In reality emails are not encrypted by default.

**Code:**

> setwd("~/Desktop")

> Dataset<-read.csv("CybersecurityKnowledgeDS.csv",header=T,sep=",")

> View(Dataset)

>names(Dataset)<-c("CaseID","Weight","XSpanish","Internet","https","PhishingA","PhishingB","PhishingC","PhishingD","PhishingE","PhishingResult","PhishingAttack","Computersnetworked","Wifi","AuthenticationA","AuthenticationB","AuthenticationC","AuthenticationD","AuthenticationE","AuthenticationF","AuthenticationResult","Authentication","CriminalAcess","PrivateBrowsing","GPS","Email","FreeCredit","Wifi","Risks","INTMOB","INTFREQ","SMART1","SNSINT2","PPAGE","PPEDUCAT","PPETHM","PPGENDER","PPINCMP","PPMSACAT","PPSTATEN","PPWORK")

> attach(Dataset)

The following objects are masked from Dataset (pos = 3):

   Authentication, AuthenticationA,

   AuthenticationB, AuthenticationC,

   AuthenticationD, AuthenticationE,

   AuthenticationF, AuthenticationResult,

   CaseID, Computersnetworked, CriminalAcess,

   Email, FreeCredit, GPS, https, Internet,

   INTFREQ, INTMOB, PhishingA, PhishingAttack,

   PhishingB, PhishingC, PhishingD, PhishingE,

   PhishingResult, PPAGE, PPEDUCAT, PPETHM,

   PPGENDER, PPINCMP, PPMSACAT, PPSTATEN,

   PPWORK, PrivateBrowsing, Risks, SMART1,

   SNSINT2, Weight, Wifi, Wifi, XSpanish

> table(Email)

Email

 1  2 3

106 496 453

> x<-table(Email)

> labels<-c("Encrypted","Not Encrypted","Not Sure")

> Percent<-round(100\*x/sum(x),1)

> Pielabels<-paste(labels,":",Percent,"%",sep="")

> colors<-c('cyan','red','goldenrod3')

> library(plotrix)

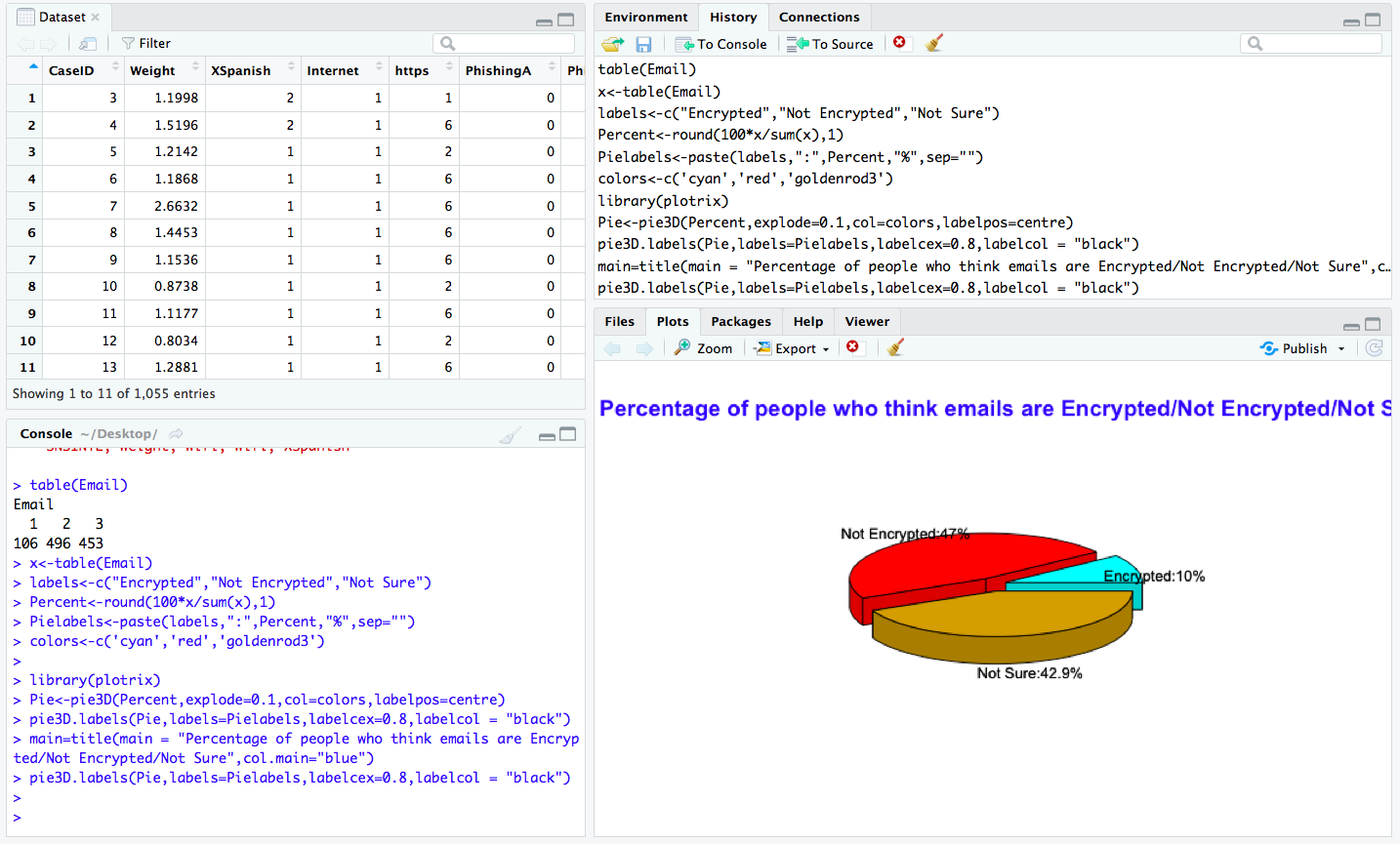
> Pie<-pie3D(Percent,explode=0.1,col=colors,labelpos=centre)

> pie3D.labels(Pie,labels=Pielabels,labelcex=0.8,labelcol = "black")

> main=title(main = "Percentage of people who think emails are Encrypted/Not Encrypted/Not Sure",col.main="blue")

> pie3D.labels(Pie,labels=Pielabels,labelcex=0.8,labelcol = "black")

**Screenshot:**

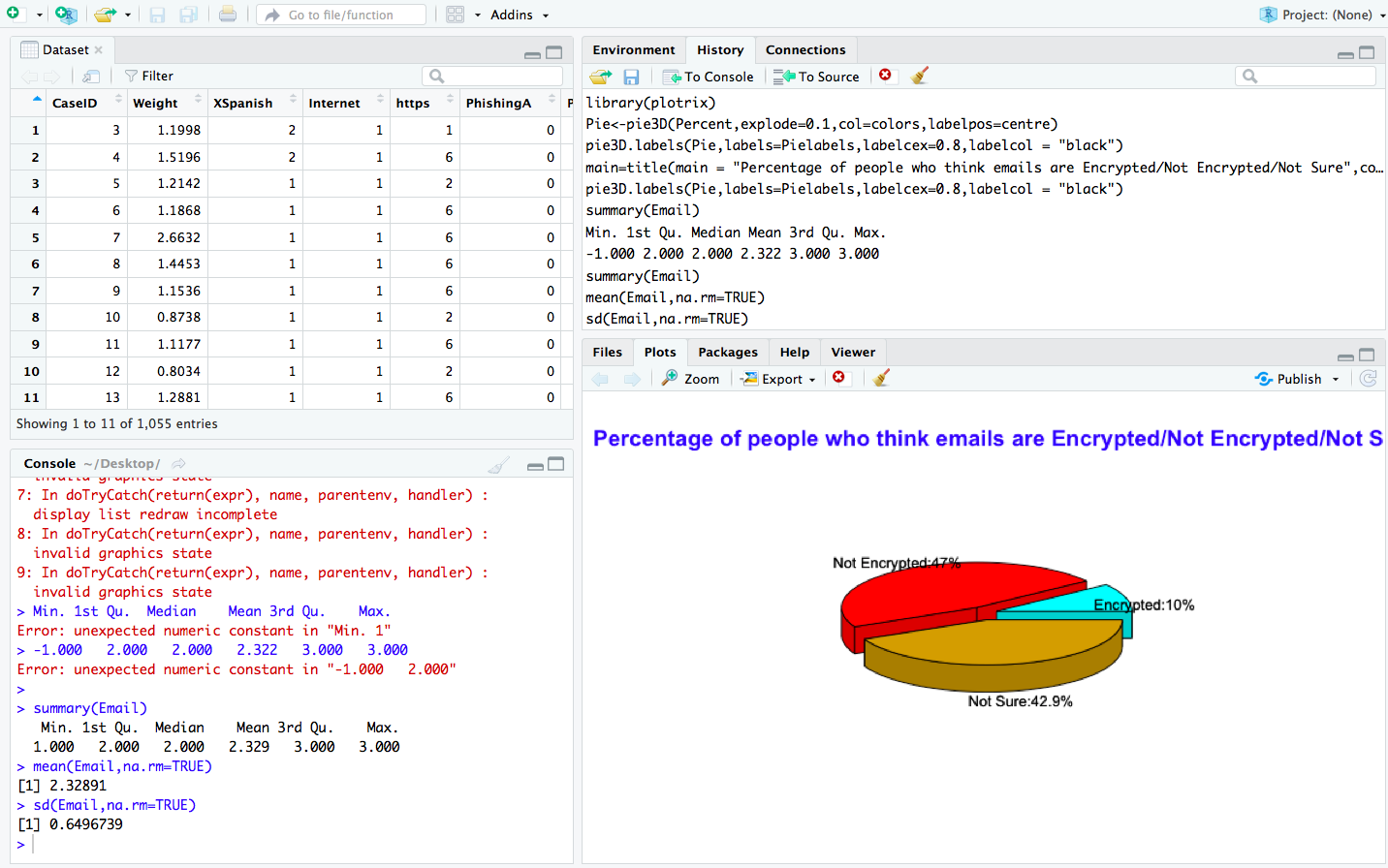


**Complete Code Screenshot:**





**Statistics:**



**Code:**

**Summary of Email is shown below:**

> summary(Email)

  Min. 1st Qu.  Median Mean 3rd Qu.    Max.

 1.000   2.000 2.000   2.329 3.000 3.000

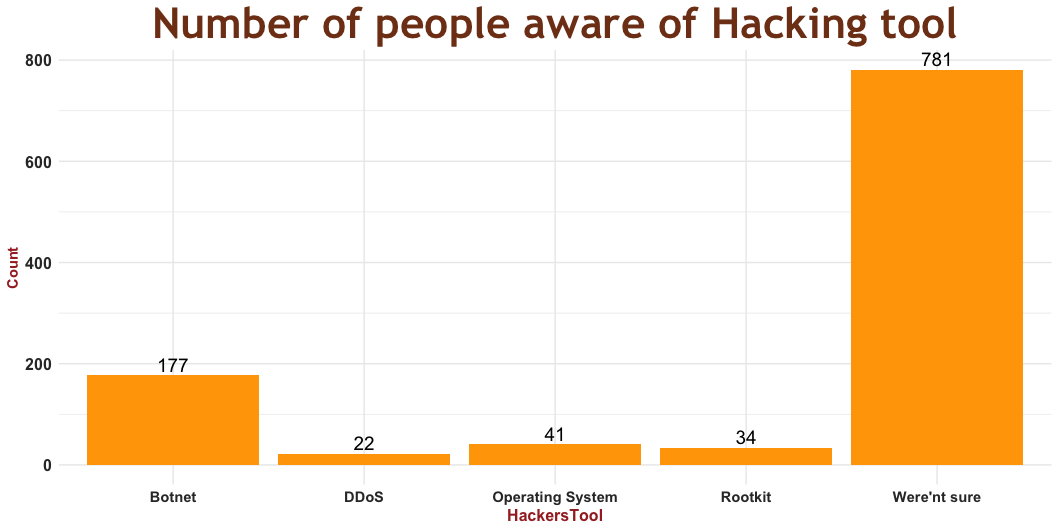
> mean(Email,na.rm=TRUE)

[1] 2.32891

> sd(Email,na.rm=TRUE)

[1] 0.6496739

**Q.2. Number of people who are aware of hacking tool which are used mostly these days to hack our computers?**



**Visualizations:** Bar Plot with labels

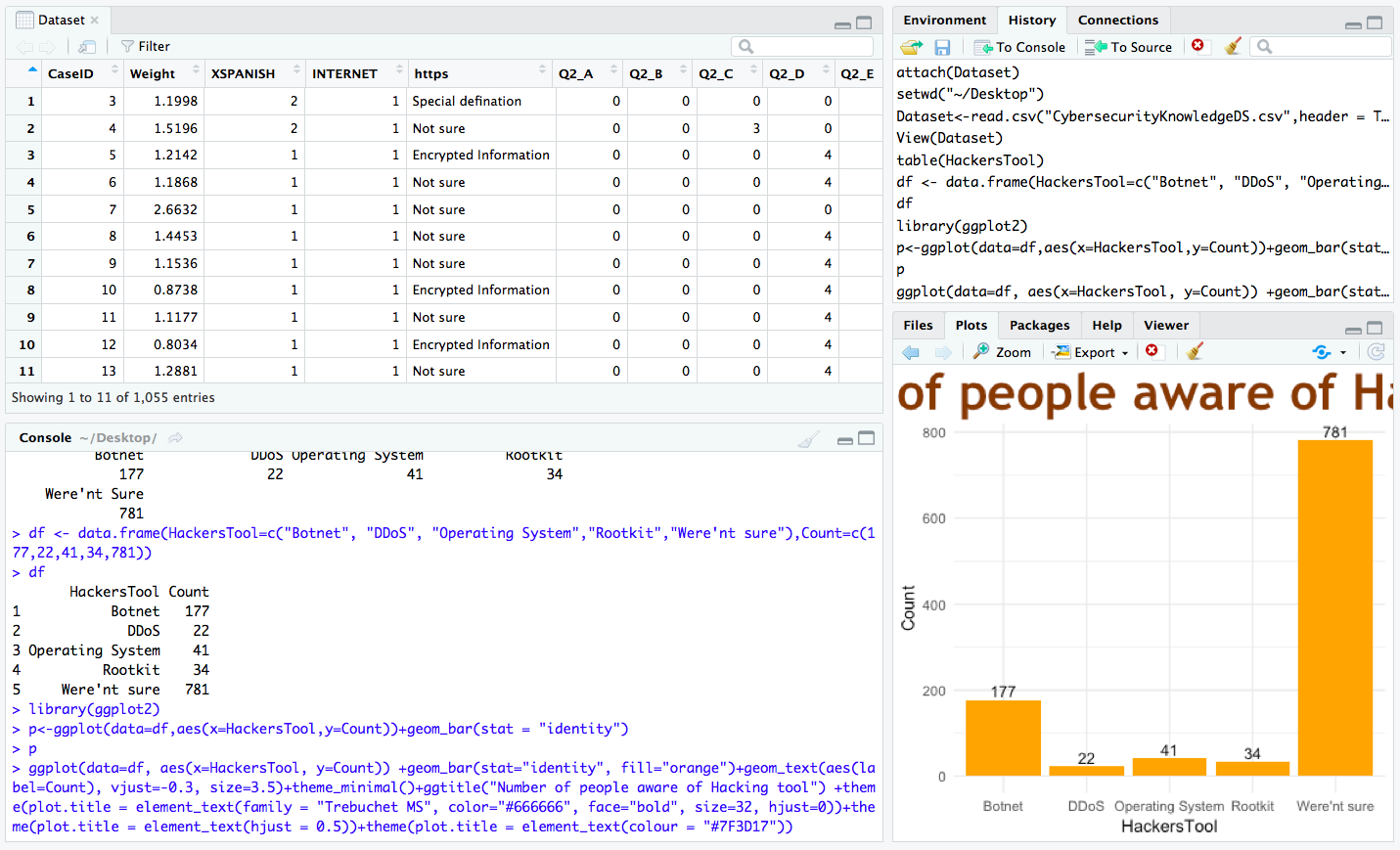
**Functions Used:** labels, ggplot2, geom\_bar, geom\_text, theme, color

**Packages/library:** ggplot2

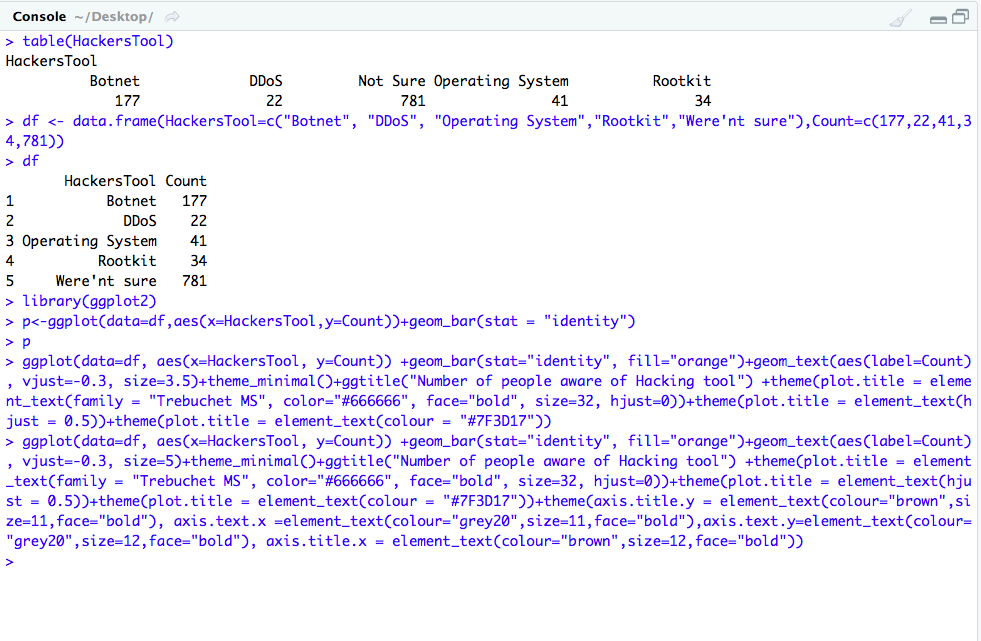
**Analysis/Description:**

In the survey, people were asked about which hacking tool is used by the hackers to hack your confidential files, accounts, emails. Five options were given; ie Botnet, DDoS, Operating System, Rootkit, Not Sure. Out of 1055 people, 781 people answered that they are not aware about the tool or have any knowledge regarding it. Only 177 people answered Botnet which is the tool used for hacking. Labels on the top of each bar indicates the number of people who feel that is the hacking tool used.

**Screenshot:**



**Code Screenshot:**

****

**Code:**

> setwd("~/Desktop")

> Dataset<-read.csv("CybersecurityKnowledgeDS.csv",header = T,sep=",")

> View(Dataset)

> table(HackersTool)

HackersTool

         Botnet         DDoS Operating System          Rootkit

            177     22 41               34

   Were'nt Sure

            781

> df <- data.frame(HackersTool=c("Botnet", "DDoS", "Operating System","Rootkit","Were'nt sure"),Count=c(177,22,41,34,781))

> df

      HackersTool Count

1           Botnet 177

2             DDoS 22

3 Operating System    41

4          Rootkit 34

5     Were'nt sure   781

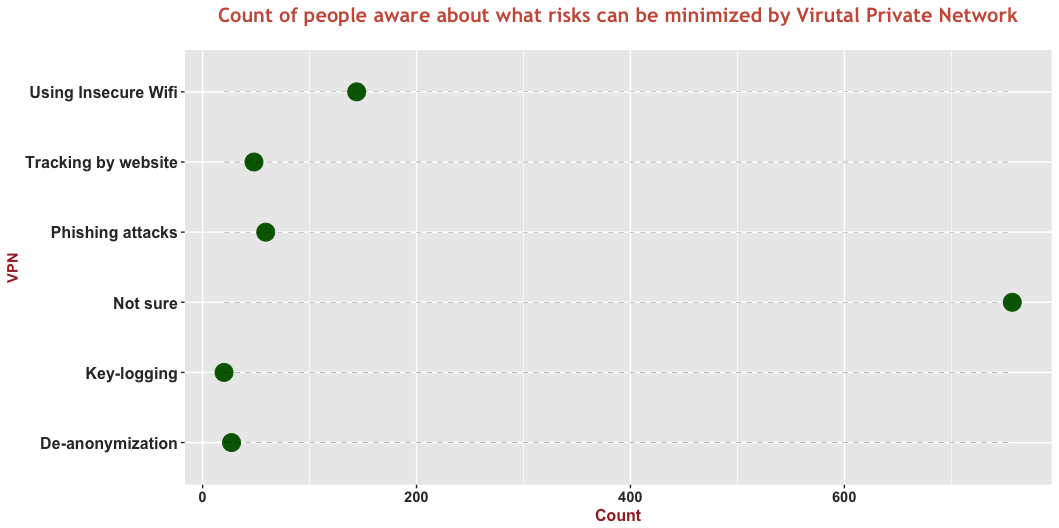
> library(ggplot2)

> p<-ggplot(data=df,aes(x=HackersTool,y=Count))+geom\_bar(stat = "identity")

> p

> ggplot(data=df, aes(x=HackersTool, y=Count)) +geom\_bar(stat="identity", fill="orange")+geom\_text(aes(label=Count), vjust=-0.3, size=3.5)+theme\_minimal()+ggtitle("Number of people aware of Hacking tool") +theme(plot.title = element\_text(family = "Trebuchet MS", color="#666666", face="bold", size=32, hjust=0))+theme(plot.title = element\_text(hjust = 0.5))+theme(plot.title = element\_text(colour = "#7F3D17"))+theme(axis.title.y = element\_text(colour="brown",size=11,face="bold"), axis.text.x=element\_text(colour="grey20",size=11,face="bold"),axis.text.y=element\_text(colour="grey20",size=12,face="bold"), axis.title.x = element\_text(colour="brown",size=11,face=”bold”)

**Q.3. Show the count of people who answered correctly about which risks can be minimized by Virtual Private Network?**



**Visualizations:** Dot Plot

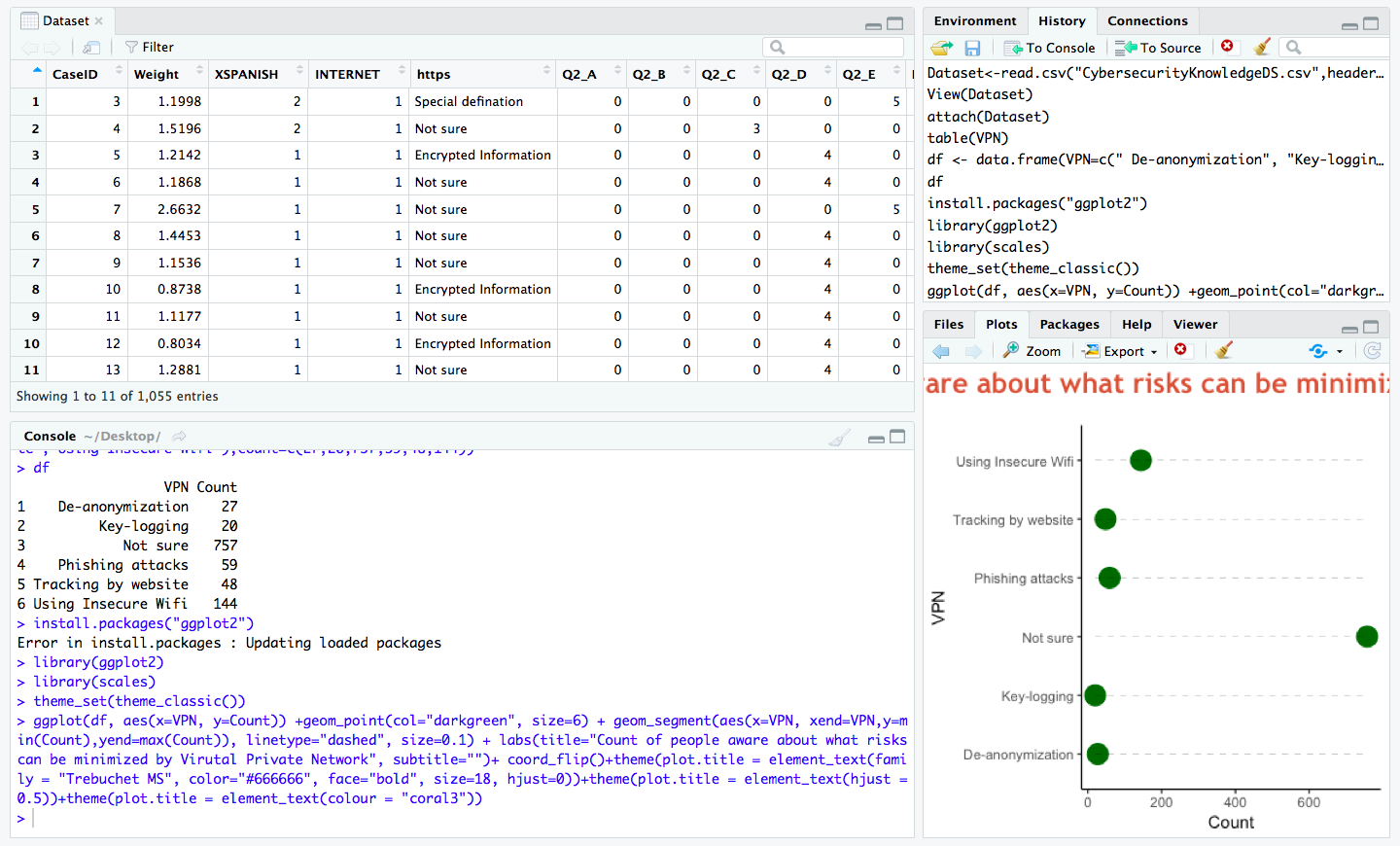
**Functions Used:** Color, plot, Coord\_flip(), aes, geom\_point

**Packages/Library:** ggplot2, scales, theme

**Analysis/Description:**

As per Dot plot visualized above, out of 1055 people who have taken survey only 144 people are aware about what risks can be minimized by Virtual Private Network. Huge count of people ie around 757 people are not aware about these risks at all. Green Dots on the graph indicates the count of people who have answered each question.

**Screenshot:**



**Code Screenshot:**



**Code:**

> setwd("~/Desktop")

> Dataset<-read.csv("CybersecurityKnowledgeDS.csv",header = T,sep = ",")

> View(Dataset)

> attach(Dataset)

The following objects are masked from Dataset (pos = 5):

 Authentication, AuthenticationResult, CaseID, GENDER, HackersTool,

   https, INTERNET, INTFREQ, INTMOB, PhishingAttack, PhishingResult,

   PPAGE, PPEDUCAT, PPETHM, PPINCIMP, PPMSACAT, PPSTATEN, PPWORK, Q10,

   Q11, Q12, Q13, Q2\_A, Q2\_B, Q2\_C, Q2\_D, Q2\_E, Q4, Q5\_A, Q5\_B, Q5\_C,

   Q5\_D, Q5\_E, Q5\_F, Q7, Q9, SMART1, SNSINT2, VPN, Weight, XSPANISH

> table(VPN)

VPN

  De-anonymization         Key-logging Not sure    Phishing attacks

                27   20 757                  59

Tracking by website Using Insecure Wifi

                48  144

> df <- data.frame(VPN=c(" De-anonymization", "Key-logging", "Not sure","Phishing attacks","Tracking by website","Using Insecure Wifi"),Count=c(27,20,757,59,48,144))

> df

                 VPN Count

1    De-anonymization    27

2         Key-logging    20

3            Not sure 757

4    Phishing attacks    59

5 Tracking by website    48

6 Using Insecure Wifi   144

> install.packages("ggplot2")

Error in install.packages : Updating loaded packages

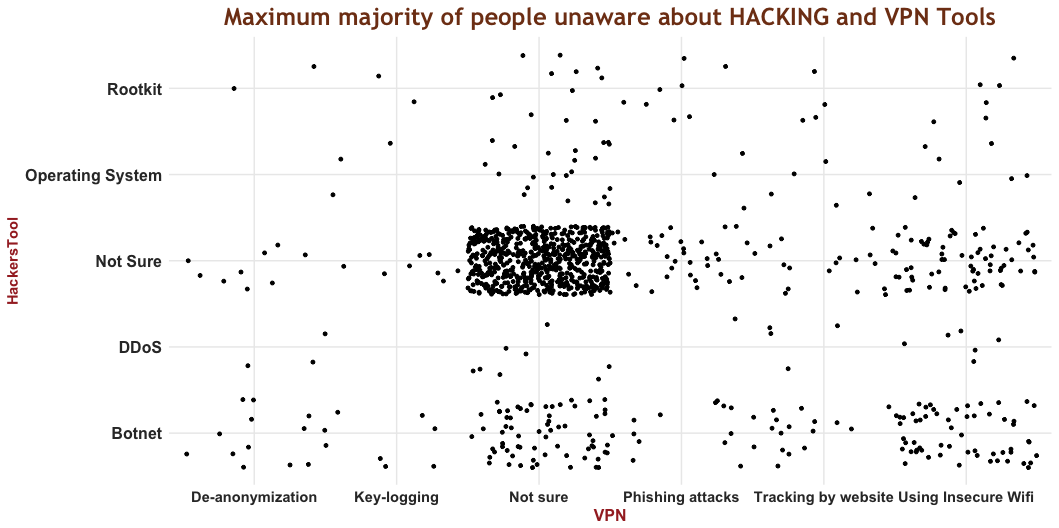
> library(ggplot2)

> library(scales)

> theme\_set(theme\_classic())

> ggplot(df, aes(x=VPN, y=Count)) +geom\_point(col="darkgreen", size=6) + geom\_segment(aes(x=VPN, xend=VPN,y=min(Count),yend=max(Count)), linetype="dashed", size=0.1) + labs(title="Count of people aware about what risks can be minimized by Virutal Private Network", subtitle="")+ coord\_flip()+theme(plot.title = element\_text(family = "Trebuchet MS", color="#666666", face="bold", size=18, hjust=0))+theme(plot.title = element\_text(hjust = 0.5))+theme(plot.title = element\_text(colour = "coral3")) +theme(plot.title = element\_text(colour = "coral3"))+theme(axis.title.y = element\_text(colour="brown",size=11,face="bold"), axis.text.x =element\_text(colour="grey20",size=11,face="bold"),axis.text.y=element\_text(colour="grey20",size=12,face="bold"), axis.title.x = element\_text(colour="brown",size=12,face="bold"))

**Q.4. Show the count of people who are unaware about hacking tool as well are VPN Tools which they should be aware of?**



**Visualizations:** Scatter Plot

**Functions Used:** geom\_jitter, gtitle, Colors

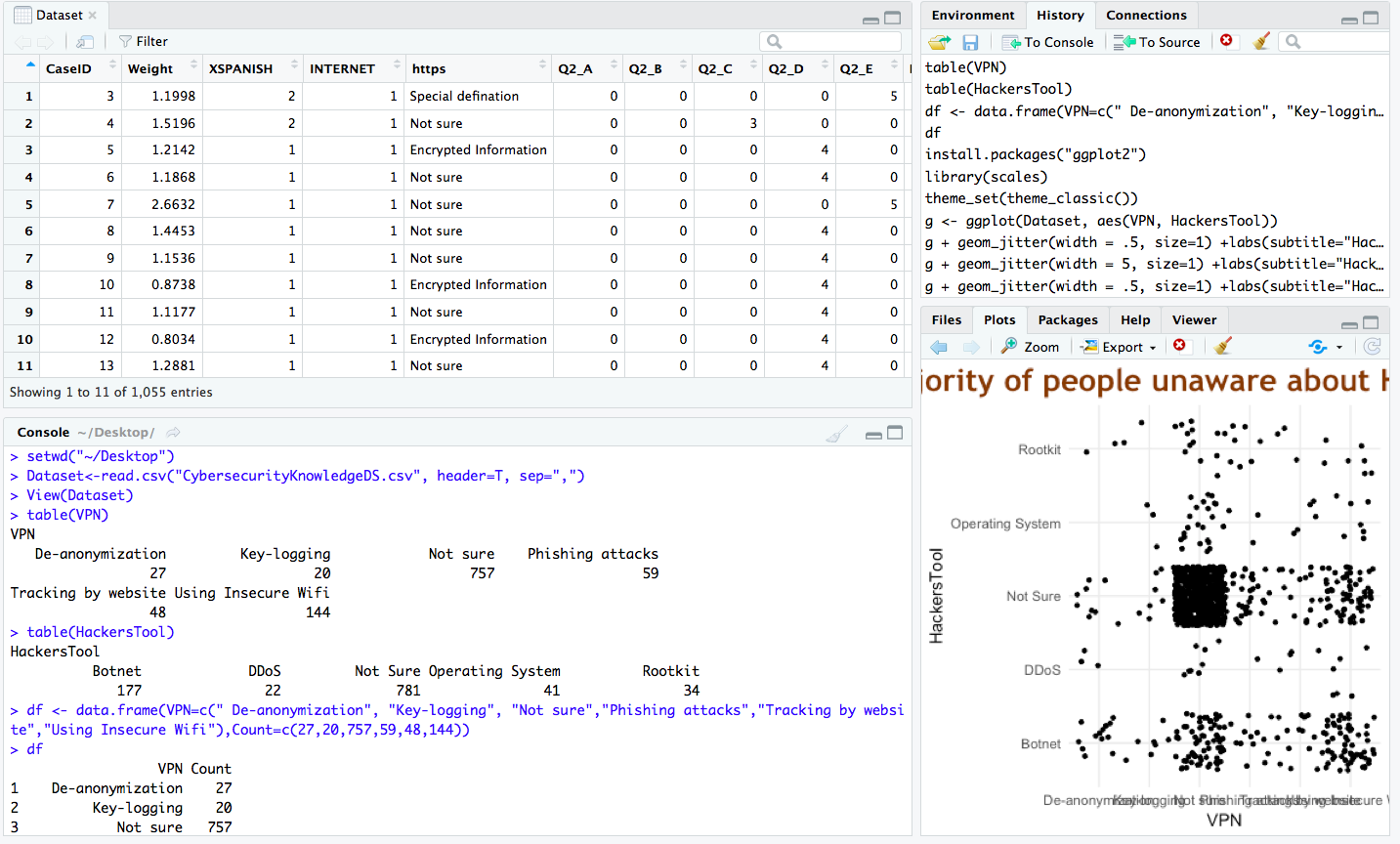
**Dataframe**

**Packages/Library:** ggplot2

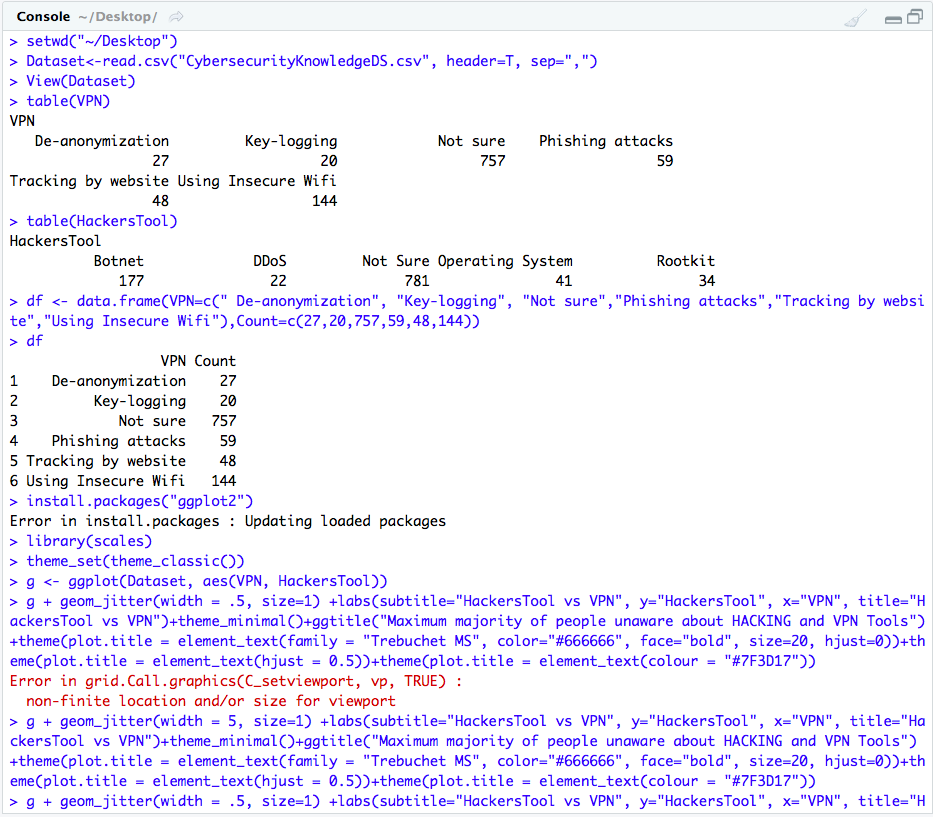
**Analysis/Description:**

As per the visualization shown above majority of people are unaware about hacking as well as VPN Tools. There are around 757 people unaware about Hacking Tool and 781 people unaware about The Virtual Private Network risks. About rest options people had mix response. Using insecure Wifi minimizes the risks of VPN is the correct answer which was answered by only 144 people and 177 people know about Botnet which is tool used by hackers.

**SCREENSHOT:**



**CODE SCREENSHOT:**



**CODE:**

> setwd("~/Desktop")

> Dataset<-read.csv("CybersecurityKnowledgeDS.csv", header=T, sep=",")

> View(Dataset)

> table(VPN)

VPN

  De-anonymization         Key-logging Not sure    Phishing attacks

                27   20 757                  59

Tracking by website Using Insecure Wifi

                48  144

> table(HackersTool)

HackersTool

         Botnet         DDoS Not Sure Operating System          Rootkit

            177     22 781               41 34

> df <- data.frame(VPN=c(" De-anonymization", "Key-logging", "Not sure","Phishing attacks","Tracking by website","Using Insecure Wifi"),Count=c(27,20,757,59,48,144))

> df

                 VPN Count

1    De-anonymization    27

2         Key-logging    20

3            Not sure 757

4    Phishing attacks    59

5 Tracking by website    48

6 Using Insecure Wifi   144

> install.packages("ggplot2")

Error in install.packages : Updating loaded packages

> library(scales)

> theme\_set(theme\_classic())

> g <- ggplot(Dataset, aes(VPN, HackersTool))

> g + geom\_jitter(width = .5, size=1) +labs(subtitle="HackersTool vs VPN", y="HackersTool", x="VPN", title="HackersTool vs VPN")+theme\_minimal()+ggtitle("Maximum majority of people unaware about HACKING and VPN Tools") +theme(plot.title = element\_text(family = "Trebuchet MS", color="blue", face="bold", size=18, hjust=0))+theme(plot.title = element\_text(hjust = 0.5))+theme(plot.title=element\_text(colour="#7F3D17"))+theme(axis.title.y= element\_text(colour="brown",size=11,face="bold"),axis.text.x=element\_text(colour="grey20",size=11,face="bold"),axis.text.y = element\_text(colour="grey20",size=12,face="bold"), axis.title.x = element\_text(colour="brown",size=12,face="bold"))

**Statistics**

**Statistical Summary of Dataset:**

> summary(Dataset)

    CaseID           Weight XSPANISH        INTERNET

Min.   : 3.0  Min. :0.2626   Min. :1.000 Min.   :1.00

1st Qu.: 285.5   1st Qu.:0.8005 1st Qu.:1.000   1st Qu.:1.00

Median : 580.0   Median :0.9325 Median :1.000   Median :1.00

Mean   : 580.3  Mean :1.0000   Mean :1.045 Mean   :1.29

3rd Qu.: 878.5   3rd Qu.:1.1483 3rd Qu.:1.000   3rd Qu.:1.00

Max.   :1161.0  Max. :2.6632   Max. :2.000 Max.   :5.00

                  https Q2\_A             Q2\_B Q2\_C

Does not exist       : 89 Min. :0.0000  Min. :0.000 Min. :0.0000

Encrypted Information:363   1st Qu.:0.0000 1st Qu.:0.000   1st Qu.:0.0000

Limited Acces        : 15 Median :0.0000  Median :0.000 Median :0.0000

New Site Version     : 10 Mean :0.1005  Mean :0.237 Mean :0.1536

Not sure             :562 3rd Qu.:0.0000  3rd Qu.:0.000 3rd Qu.:0.0000

Special defination   : 16 Max. :1.0000  Max. :2.000 Max. :3.0000

     Q2\_D          Q2\_E PhishingResult  PhishingAttack

Min.   :0.000 Min.   :0.000 Min. :    0 Min. : 1.000

1st Qu.:0.000   1st Qu.:0.000 1st Qu.:   40 1st Qu.: 4.000

Median :4.000   Median :0.000 Median :   40 Median : 4.000

Mean   :2.165 Mean   :1.185 Mean : 1280   Mean : 5.582

3rd Qu.:4.000   3rd Qu.:0.000 3rd Qu.:   40 3rd Qu.: 5.000

Max.   :4.000 Max.   :5.000 Max. :12300   Max. :123.000

                                                NA's :5

          HackersTool        Q4 Q5\_A             Q5\_B

Botnet          :177 Min. :1.000   Min. :0.0000 Min. :0.0000

DDoS            : 22 1st Qu.:2.000   1st Qu.:0.0000 1st Qu.:0.0000

Not Sure        :781 Median :2.000   Median :0.0000 Median :0.0000

Operating System: 41   Mean :2.335 Mean :0.3261   Mean :0.5573

Rootkit         : 34 3rd Qu.:3.000   3rd Qu.:1.0000 3rd Qu.:2.0000

                       Max. :3.000 Max. :1.0000   Max. :2.0000

     Q5\_C          Q5\_D Q5\_E              Q5\_F

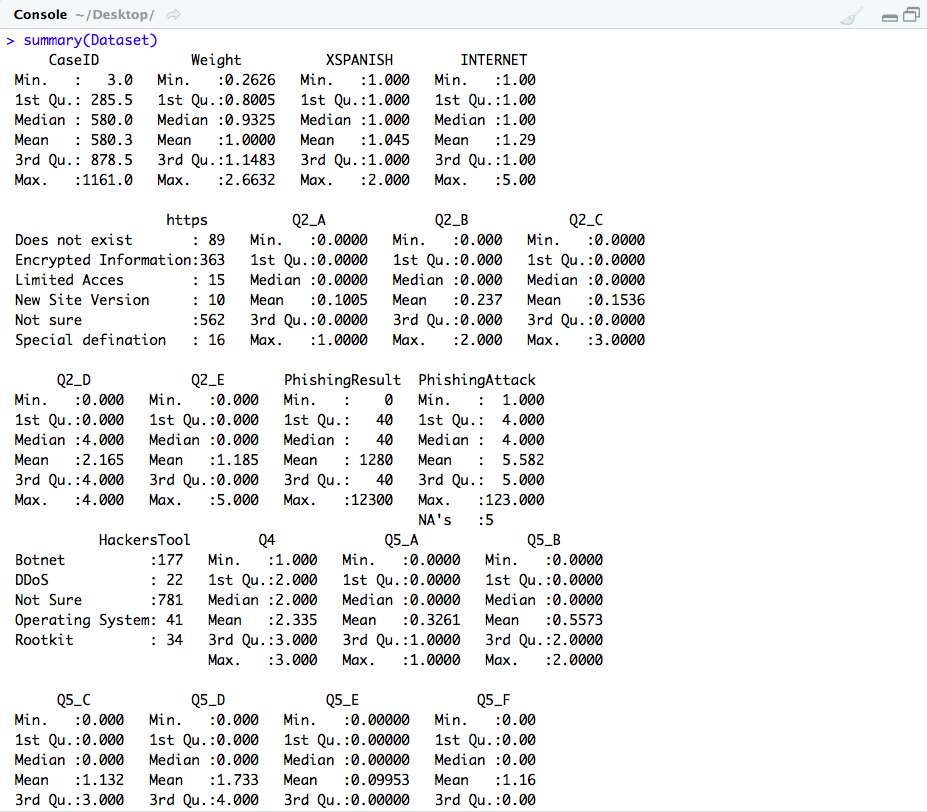
Min.   :0.000 Min.   :0.000 Min. :0.00000   Min. :0.00

1st Qu.:0.000   1st Qu.:0.000 1st Qu.:0.00000   1st Qu.:0.00

Median :0.000   Median :0.000 Median :0.00000   Median :0.00

Mean   :1.132 Mean   :1.733 Mean :0.09953   Mean :1.16

3rd Qu.:3.000   3rd Qu.:4.000 3rd Qu.:0.00000   3rd Qu.:0.00



**USER DEFINED FUNCTIONS:**

**Q) All Wi-Fi traffic is encrypted by default on all wireless routers.**

1. **True**
2. **False**
3. **Not Sure**

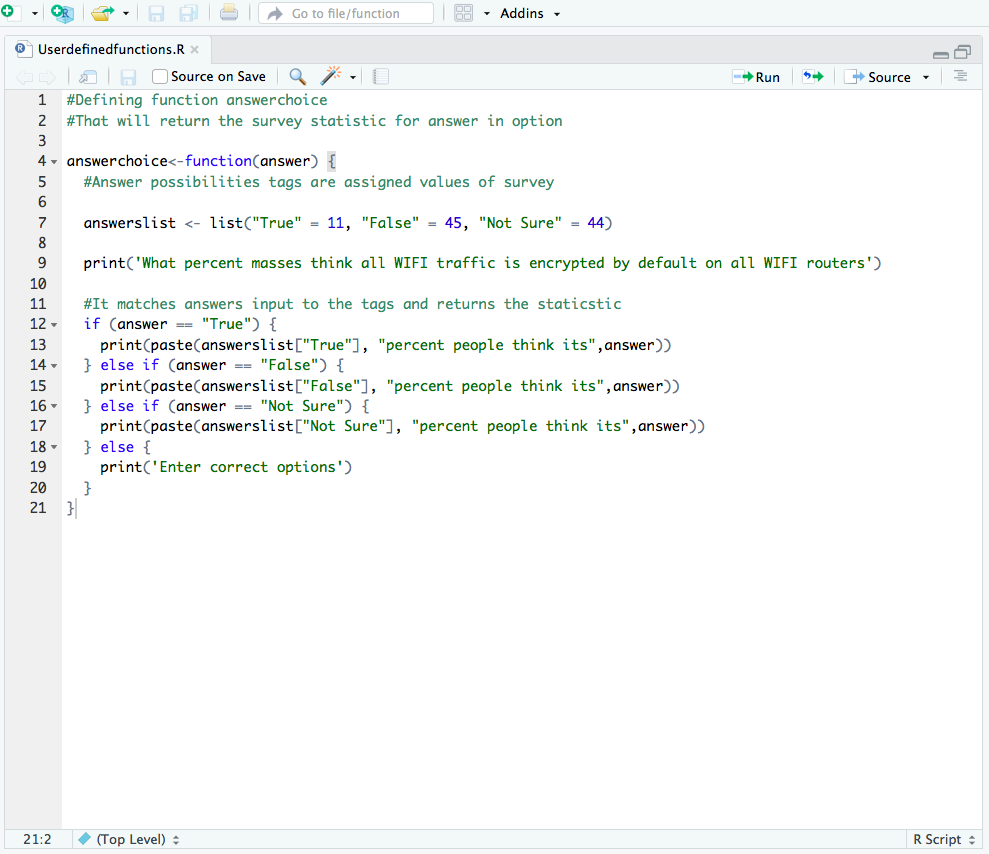
People were surveyed with the above question if all Wi-Fi traffic is encrypted by default.

Around 11% people answered True, 45% people answered False, 44% people answered Not Sure.

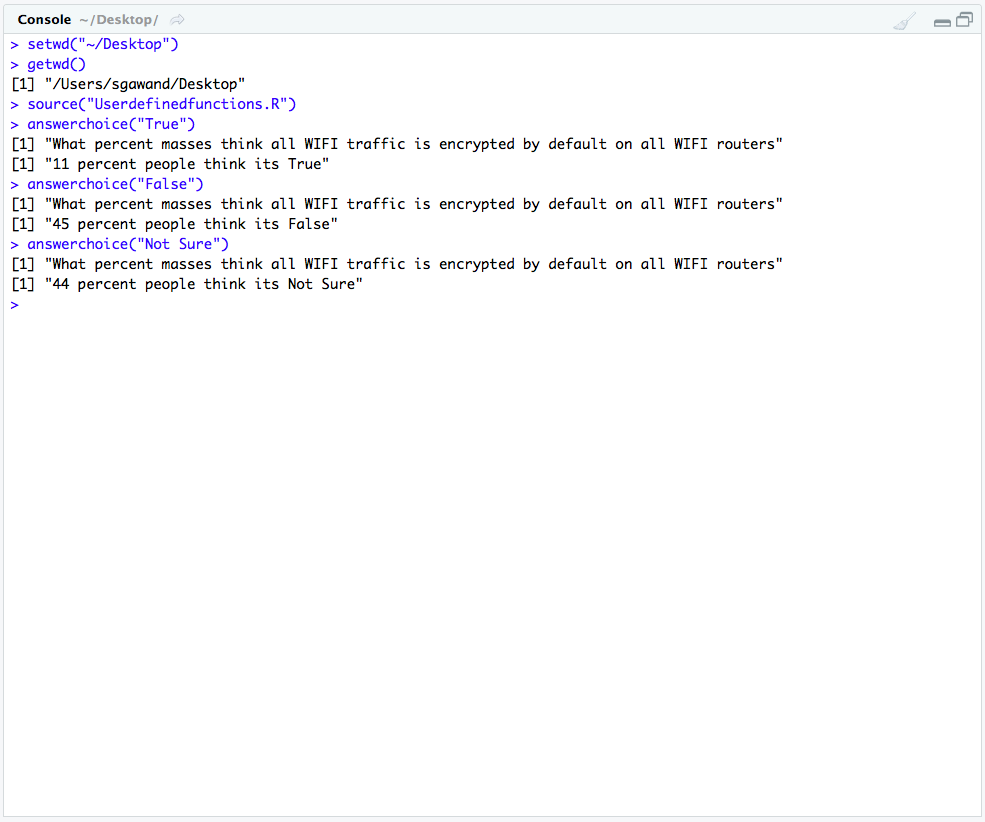
Funtion answerchoice is created that accepts a user definied argument. In this case it’s the answer or opinion of the survey. Function returns the statistic associated to it. A list of tags and its corresponsding values are saved in answerlist. The value corresponding to the user inputted argument is returned as the answer. Function also handles incorrect argument case.

****

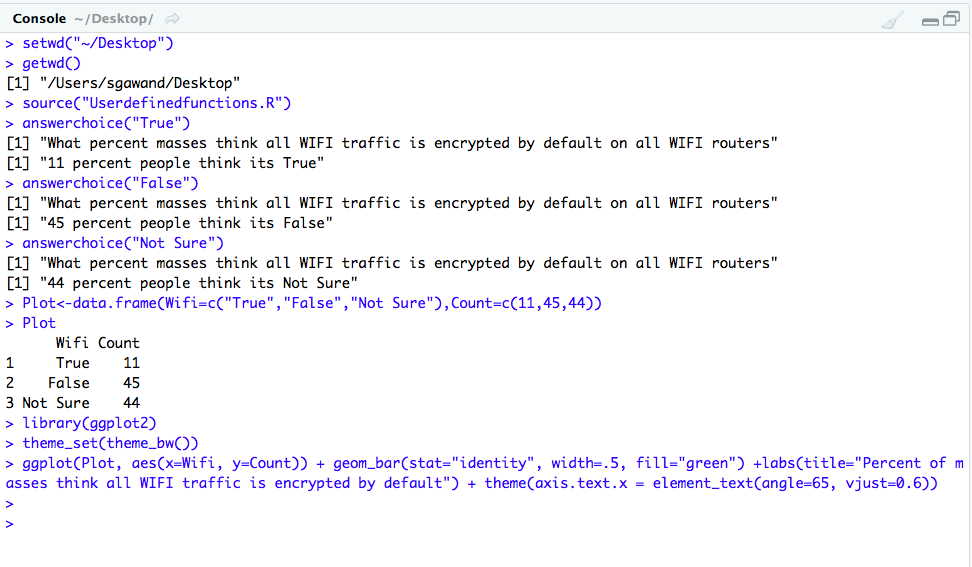
**CODE:**

****

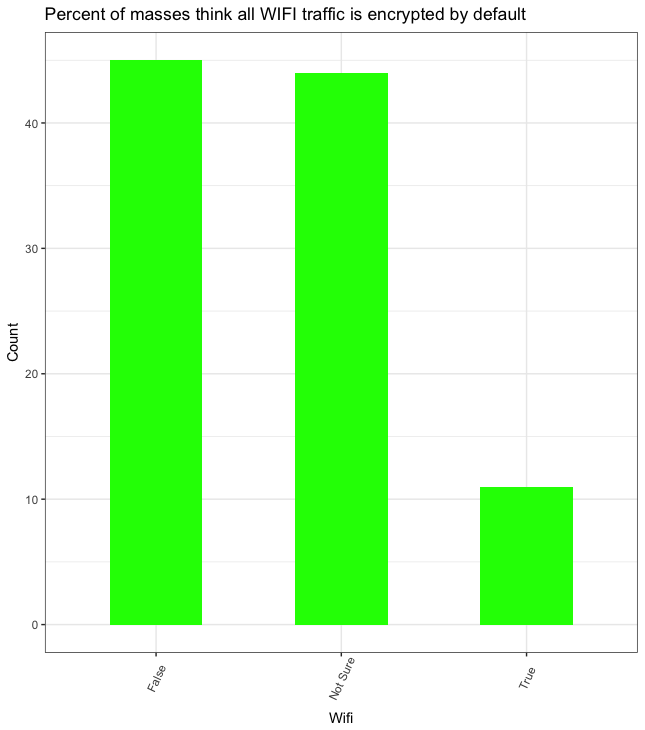
**EXECUTION/OUTPUT SCREENSHOT:**

****

**EXECUTION WITH VISUALIZATION:**

****

**Visualization:**



**Code:**

#Defining function answerchoice

#That will return the survey statistic for answer in option

answerchoice<-function(answer) {

#Answer possibilities tags are assigned values of survey

answerslist <- list("True" = 11, "False" = 45, "Not Sure" = 44)

print('What percent masses think all WIFI traffic is encrypted by default on all WIFI routers')

#It matches answers input to the tags and returns the staticstic

if (answer == "True") {

print(paste(answerslist["True"], "percent people think its",answer))

} else if (answer == "False") {

print(paste(answerslist["False"], "percent people think its",answer))

} else if (answer == "Not Sure") {

print(paste(answerslist["Not Sure"], "percent people think its",answer))

} else {

print('Enter correct options')

}

}

**Execution Result:**

> setwd("~/Desktop")

> getwd()

[1] "/Users/sgawand/Desktop"

> source("Userdefinedfunctions.R")

> answerchoice("True")

[1] "What percent masses think all WIFI traffic is encrypted by default on all WIFI routers"

[1] "11 percent people think its True"

> answerchoice("False")

[1] "What percent masses think all WIFI traffic is encrypted by default on all WIFI routers"

[1] "45 percent people think its False"

> answerchoice("Not Sure")

[1] "What percent masses think all WIFI traffic is encrypted by default on all WIFI routers"

[1] "44 percent people think its Not Sure"

**For Visualization:**

> Plot<-data.frame(Wifi=c("True","False","Not Sure"),Count=c(11,45,44))

> Plot

Wifi Count

1 True 11

2 False 45

3 Not Sure 44

> library(ggplot2)

> theme\_set(theme\_bw())

> ggplot(Plot, aes(x=Wifi, y=Count)) + geom\_bar(stat="identity", width=.5, fill="green") +labs(title="Percent of masses think all WIFI traffic is encrypted by default") + theme(axis.text.x = element\_text(angle=65, vjust=0.6))

**CODES:**

**DATA CLEANING CODES**

**1) Renaming Column Names:**

> setwd("~/Desktop")

> Dataset<-read.csv("CybersecurityKnowledgeDS.csv",header=T,sep=",")

> View(Dataset)

>names(Dataset)<-c("CaseID","Weight","XSpanish","Internet","https","PhishingA","PhishingB","PhishingC","PhishingD","PhishingE","PhishingAttack","Q2\_Refused","Computersnetworked","Wifi","AuthenticationA","AuthenticationB","AuthenticationC","AuthenticationD","AuthenticationE","AuthenticationF","Authentication","Q5\_Refused","CriminalAcess","PrivateBrowsing","GPS","Email","FreeCredit","Wifi","Risks","INTMOB","INTFREQ","SMART1","SNSINT2","PPAGE","PPEDUCAT","PPETHM","PPGENDER","PPINCMP","PPMSACAT","PPSTATEN","PPWORK")

**2) Combining Columns:**

> install.packages("tidyr")

Error in install.packages : Updating loaded packages

> install.packages("tidyr")

trying URL 'https://cran.rstudio.com/bin/macosx/el-capitan/contrib/3.4/tidyr\_0.8.0.tgz'

Content type 'application/x-gzip' length 583255 bytes (569 KB)

==================================================

downloaded 569 KB

The downloaded binary packages are in

/var/folders/c8/0rwnncb576q7ktnxcvr034xm0000gn/T//RtmplArtIz/downloaded\_packages

> library(tidyr)

>ColumnsUnite<-unite(Dataset,PhishingResult,c("PhishingA","PhishingB","PhishingC","PhishingD","PhishingE"),sep="",remove=TRUE)

>View(ColumnsUnite)

>ColumnsUnite1<-unite(Dataset,AuthenticationResult,c("AuthenticationA","AuthenticationB","AuthenticationC","AuthenticationD","AuthenticationE","AuthenticationF"),sep="",remove=TRUE)

> ColumnsUnite1

**3) Removing Irrelevant Columns**

> deletecolumns<-c("Q2\_Refused","Q5\_Refused")

> Dataset<-Dataset[, ! names(Dataset) %in% deletecolumns, drop = F]

**4) Removing Rows where values are not available (NA)**

> View(Dataset)

> complete.cases(Dataset)

  [1]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

 [19]  TRUE TRUE  TRUE TRUE FALSE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

 [37]  TRUE TRUE  TRUE TRUE TRUE  TRUE FALSE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

 [55]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

 [73]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

 [91]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE FALSE  TRUE TRUE TRUE TRUE

[109]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[127]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[145]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[163]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[181]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[199]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[217]  TRUE TRUE  TRUE TRUE TRUE FALSE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE TRUE  TRUE

[235]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[253]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[271]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[289]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[307]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[325]  TRUE TRUE  TRUE TRUE TRUE FALSE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE TRUE  TRUE

[343]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[361]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[379]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[397] FALSE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE FALSE TRUE TRUE

[415]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[433]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[451]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[469]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[487]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[505]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[523]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[541]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE FALSE TRUE TRUE  TRUE FALSE TRUE TRUE

[559]  TRUE TRUE  TRUE TRUE TRUE FALSE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE TRUE  TRUE

[577]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[595]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[613]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE FALSE  TRUE TRUE TRUE

[631]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[649]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[667]  TRUE TRUE FALSE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE TRUE

[685]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[703]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[721]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE FALSE  TRUE TRUE TRUE TRUE TRUE TRUE  TRUE TRUE

[739]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[757]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[775]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[793]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[811]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE FALSE TRUE  TRUE TRUE TRUE TRUE

[829]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[847]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE FALSE  TRUE TRUE TRUE

[865]  TRUE TRUE FALSE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE TRUE

[883]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[901]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[919]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[937]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[955]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[973]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE TRUE  TRUE TRUE TRUE TRUE

[991]  TRUE TRUE  TRUE TRUE TRUE  TRUE TRUE TRUE TRUE  TRUE

[ reached getOption("max.print") -- omitted 55 entries ]

> Dataset1<-na.omit(Dataset)

> View(Dataset1)

**Visualization Codes:**

**Visualization 1:**

> setwd("~/Desktop")

> Dataset<-read.csv("CybersecurityKnowledgeDS.csv",header=T,sep=",")

> View(Dataset)

>names(Dataset)<-c("CaseID","Weight","XSpanish","Internet","https","PhishingA","PhishingB","PhishingC","PhishingD","PhishingE","PhishingResult","PhishingAttack","Computersnetworked","Wifi","AuthenticationA","AuthenticationB","AuthenticationC","AuthenticationD","AuthenticationE","AuthenticationF","AuthenticationResult","Authentication","CriminalAcess","PrivateBrowsing","GPS","Email","FreeCredit","Wifi","Risks","INTMOB","INTFREQ","SMART1","SNSINT2","PPAGE","PPEDUCAT","PPETHM","PPGENDER","PPINCMP","PPMSACAT","PPSTATEN","PPWORK")

> attach(Dataset)

The following objects are masked from Dataset (pos = 3):

   Authentication, AuthenticationA,

   AuthenticationB, AuthenticationC,

   AuthenticationD, AuthenticationE,

   AuthenticationF, AuthenticationResult,

   CaseID, Computersnetworked, CriminalAcess,

   Email, FreeCredit, GPS, https, Internet,

   INTFREQ, INTMOB, PhishingA, PhishingAttack,

   PhishingB, PhishingC, PhishingD, PhishingE,

   PhishingResult, PPAGE, PPEDUCAT, PPETHM,

   PPGENDER, PPINCMP, PPMSACAT, PPSTATEN,

   PPWORK, PrivateBrowsing, Risks, SMART1,

   SNSINT2, Weight, Wifi, Wifi, XSpanish

> table(Email)

Email

 1  2 3

106 496 453

> x<-table(Email)

> labels<-c("Encrypted","Not Encrypted","Not Sure")

> Percent<-round(100\*x/sum(x),1)

> Pielabels<-paste(labels,":",Percent,"%",sep="")

> colors<-c('cyan','red','goldenrod3')

> library(plotrix)

> Pie<-pie3D(Percent,explode=0.1,col=colors,labelpos=centre)

> pie3D.labels(Pie,labels=Pielabels,labelcex=0.8,labelcol = "black")

> main=title(main = "Percentage of people who think emails are Encrypted/Not Encrypted/Not Sure",col.main="blue")

> pie3D.labels(Pie,labels=Pielabels,labelcex=0.8,labelcol = "black")

**Visualization 2:**

> setwd("~/Desktop")

> Dataset<-read.csv("CybersecurityKnowledgeDS.csv",header = T,sep=",")

> View(Dataset)

> table(HackersTool)

HackersTool

         Botnet         DDoS Operating System          Rootkit

            177     22 41               34

   Were'nt Sure

            781

> df <- data.frame(HackersTool=c("Botnet", "DDoS", "Operating System","Rootkit","Were'nt sure"),Count=c(177,22,41,34,781))

> df

      HackersTool Count

1           Botnet 177

2             DDoS 22

3 Operating System    41

4          Rootkit 34

5     Were'nt sure   781

> library(ggplot2)

> p<-ggplot(data=df,aes(x=HackersTool,y=Count))+geom\_bar(stat = "identity")

> p

> ggplot(data=df, aes(x=HackersTool, y=Count)) +geom\_bar(stat="identity", fill="orange")+geom\_text(aes(label=Count), vjust=-0.3, size=3.5)+theme\_minimal()+ggtitle("Number of people aware of Hacking tool") +theme(plot.title = element\_text(family = "Trebuchet MS", color="#666666", face="bold", size=32, hjust=0))+theme(plot.title = element\_text(hjust = 0.5))+theme(plot.title = element\_text(colour = "#7F3D17"))+theme(axis.title.y = element\_text(colour="brown",size=11,face="bold"), axis.text.x=element\_text(colour="grey20",size=11,face="bold"),axis.text.y=element\_text(colour="grey20",size=12,face="bold"), axis.title.x = element\_text(colour="brown",size=11,face=”bold”)

**Visualization 3:**

> setwd("~/Desktop")

> Dataset<-read.csv("CybersecurityKnowledgeDS.csv",header = T,sep = ",")

> View(Dataset)

> attach(Dataset)

The following objects are masked from Dataset (pos = 5):

 Authentication, AuthenticationResult, CaseID, GENDER, HackersTool,

   https, INTERNET, INTFREQ, INTMOB, PhishingAttack, PhishingResult,

   PPAGE, PPEDUCAT, PPETHM, PPINCIMP, PPMSACAT, PPSTATEN, PPWORK, Q10,

   Q11, Q12, Q13, Q2\_A, Q2\_B, Q2\_C, Q2\_D, Q2\_E, Q4, Q5\_A, Q5\_B, Q5\_C,

   Q5\_D, Q5\_E, Q5\_F, Q7, Q9, SMART1, SNSINT2, VPN, Weight, XSPANISH

> table(VPN)

VPN

  De-anonymization         Key-logging Not sure    Phishing attacks

                27   20 757                  59

Tracking by website Using Insecure Wifi

                48  144

> df <- data.frame(VPN=c(" De-anonymization", "Key-logging", "Not sure","Phishing attacks","Tracking by website","Using Insecure Wifi"),Count=c(27,20,757,59,48,144))

> df

                 VPN Count

1    De-anonymization    27

2         Key-logging    20

3            Not sure 757

4    Phishing attacks    59

5 Tracking by website    48

6 Using Insecure Wifi   144

> install.packages("ggplot2")

Error in install.packages : Updating loaded packages

> library(ggplot2)

> library(scales)

> theme\_set(theme\_classic())

> ggplot(df, aes(x=VPN, y=Count)) +geom\_point(col="darkgreen", size=6) + geom\_segment(aes(x=VPN, xend=VPN,y=min(Count),yend=max(Count)), linetype="dashed", size=0.1) + labs(title="Count of people aware about what risks can be minimized by Virutal Private Network", subtitle="")+ coord\_flip()+theme(plot.title = element\_text(family = "Trebuchet MS", color="#666666", face="bold", size=18, hjust=0))+theme(plot.title = element\_text(hjust = 0.5))+theme(plot.title = element\_text(colour = "coral3")) +theme(plot.title = element\_text(colour = "coral3"))+theme(axis.title.y = element\_text(colour="brown",size=11,face="bold"), axis.text.x =element\_text(colour="grey20",size=11,face="bold"),axis.text.y=element\_text(colour="grey20",size=12,face="bold"), axis.title.x = element\_text(colour="brown",size=12,face="bold"))

**Visualization 4:**

> setwd("~/Desktop")

> Dataset<-read.csv("CybersecurityKnowledgeDS.csv", header=T, sep=",")

> View(Dataset)

> table(VPN)

VPN

  De-anonymization         Key-logging Not sure    Phishing attacks

                27   20 757                  59

Tracking by website Using Insecure Wifi

                48  144

> table(HackersTool)

HackersTool

         Botnet         DDoS Not Sure Operating System          Rootkit

            177     22 781               41 34

> df <- data.frame(VPN=c(" De-anonymization", "Key-logging", "Not sure","Phishing attacks","Tracking by website","Using Insecure Wifi"),Count=c(27,20,757,59,48,144))

> df

                 VPN Count

1    De-anonymization    27

2         Key-logging    20

3            Not sure 757

4    Phishing attacks    59

5 Tracking by website    48

6 Using Insecure Wifi   144

> install.packages("ggplot2")

Error in install.packages : Updating loaded packages

> library(scales)

> theme\_set(theme\_classic())

> g <- ggplot(Dataset, aes(VPN, HackersTool))

> g + geom\_jitter(width = .5, size=1) +labs(subtitle="HackersTool vs VPN", y="HackersTool", x="VPN", title="HackersTool vs VPN")+theme\_minimal()+ggtitle("Maximum majority of people unaware about HACKING and VPN Tools") +theme(plot.title = element\_text(family = "Trebuchet MS", color="blue", face="bold", size=18, hjust=0))+theme(plot.title = element\_text(hjust = 0.5))+theme(plot.title=element\_text(colour="#7F3D17"))+theme(axis.title.y= element\_text(colour="brown",size=11,face="bold"),axis.text.x=element\_text(colour="grey20",size=11,face="bold"),axis.text.y = element\_text(colour="grey20",size=12,face="bold"), axis.title.x = element\_text(colour="brown",size=12,face="bold"))

**USER DEFINED CODES**

**Code:**

#Defining function answerchoice

#That will return the survey statistic for answer in option

answerchoice<-function(answer) {

#Answer possibilities tags are assigned values of survey

answerslist <- list("True" = 11, "False" = 45, "Not Sure" = 44)

print('What percent masses think all WIFI traffic is encrypted by default on all WIFI routers')

#It matches answers input to the tags and returns the staticstic

if (answer == "True") {

print(paste(answerslist["True"], "percent people think its",answer))

} else if (answer == "False") {

print(paste(answerslist["False"], "percent people think its",answer))

} else if (answer == "Not Sure") {

print(paste(answerslist["Not Sure"], "percent people think its",answer))

} else {

print('Enter correct options')

}

}

**Execution Result:**

> setwd("~/Desktop")

> getwd()

[1] "/Users/sgawand/Desktop"

> source("Userdefinedfunctions.R")

> answerchoice("True")

[1] "What percent masses think all WIFI traffic is encrypted by default on all WIFI routers"

[1] "11 percent people think its True"

> answerchoice("False")

[1] "What percent masses think all WIFI traffic is encrypted by default on all WIFI routers"

[1] "45 percent people think its False"

> answerchoice("Not Sure")

[1] "What percent masses think all WIFI traffic is encrypted by default on all WIFI routers"

[1] "44 percent people think its Not Sure"

**For Visualization:**

> Plot<-data.frame(Wifi=c("True","False","Not Sure"),Count=c(11,45,44))

> Plot

Wifi Count

1 True 11

2 False 45

3 Not Sure 44

> library(ggplot2)

> theme\_set(theme\_bw())

> ggplot(Plot, aes(x=Wifi, y=Count)) + geom\_bar(stat="identity", width=.5, fill="green") +labs(title="Percent of masses think all WIFI traffic is encrypted by default") + theme(axis.text.x = element\_text(angle=65, vjust=0.6))

**REFERENCE**

Asin N. “Pew Research Center: Internet, Science & Tech.” *Pew Research Center: Internet, Science & Tech*, 14 May 2018, [www.pewinternet.org/](http://www.pewinternet.org/)

Tutorials Point. “R Overview.” *Www.tutorialspoint.com*, Tutorials Point, 11 Apr. 2018, [www.tutorialspoint.com/r/r\_overview.htm](http://www.tutorialspoint.com/r/r_overview.htm)

“Deep Learning in R.” *An R Introduction to Statistics | R Tutorial*, [www.r-tutor.com](http://www.r-tutor.com)