PROJECT TOPIC: ANALYSIS OF DATA BREACHES IN 2018

CIS 5270-01: BUSINESS INTELLIGENCE PROJECT



Siddhi Udani

Maitri Shah

MS in Information Systems

California State University, Los Angeles

Submitted to: Dr Shilpa Balan

**Table of Contents**

|  |  |  |
| --- | --- | --- |
|  | **Topic** | **Page No** |
| 1 | **Objective of Study** | 3 |
| 2 | **Dataset** | 4 |
| 3 | **Data cleaning** | 5 |
| 4 | **Analysis and Visualizations** | 13 |
| 5 | **Statistics** | 24 |
| 6 | **User Defined Functions** | 27 |
| 7 | **Data cleaning codes** | 38 |
| 8 | **Visualization codes** | 41 |
| 9 | **Statistical/user defined codes** | 43 |
| 10 | **References** | 49 |

**OBJECTIVE OF STUDY**

A data breach is a security incident in which information is accessed without authorization. Data breaches can hurt businesses and consumers in a variety of ways. They are a costly expense that can damage lives and reputations and take time to repair.

The threat landscape has continued to evolve throughout the year, with hackers ramping up targeted, sophisticated attacks. Ransomware continued to plague the healthcare sector, while phishing attacks and insider errors led to some of the biggest breaches in 2018.  However, resources and staffing gaps [continue](https://healthitsecurity.com/news/biggest-challenges-lessons-learned-from-health-cybersecurity-in-2018) to be problematic. And hackers will continue to pummel the sector with targeted attacks [through 2019](https://healthitsecurity.com/news/ai-iot-medical-devices-top-health-cybersecurity-predictions-for-2019) and beyond[1].

Although such incidents pose the risk of [identity theft](https://en.wikipedia.org/wiki/Identity_theft) or other serious consequences, in most cases there is no lasting damage; either the breach in security is remedied before the information is accessed by unscrupulous people, or the thief is only interested in the hardware stolen, not the data it contains. Nevertheless, when such incidents become publicly known, it is customary for the offending party to attempt to mitigate [damages](https://en.wikipedia.org/wiki/Damages) by providing to the victim's subscription to a [credit reporting agency](https://en.wikipedia.org/wiki/Credit_reporting_agency), for instance, new credit cards, or other instruments. [2]

The goal of this study is to analyze and develop comprehensive visualizations of the data breaches that occur in various industry with respect to the source of their location (globally), the number of records being breached as well as their risk score. These trends regarding data breaches look grim, but experts are working on ways to stop these breaches.

**DATASET URL AND DESCRIPTION**

**Dataset url**: The dataset for this study has been taken from BREACH LEVEL INDEX, the link to which has been mentioned below:

1. <https://breachlevelindex.com/data-breach-database>

**Dataset information:** This dataset has been archived from a centralized, global database of data breaches named: Breach Level Index, displaying 8 columns and 1484 rows. It shows that data breaches are very much a growing threat for organizations. The number of records compromised is remarkable, considering the lengths many organizations go to in order to protect their data.

**Size of the dataset: 709KB.**

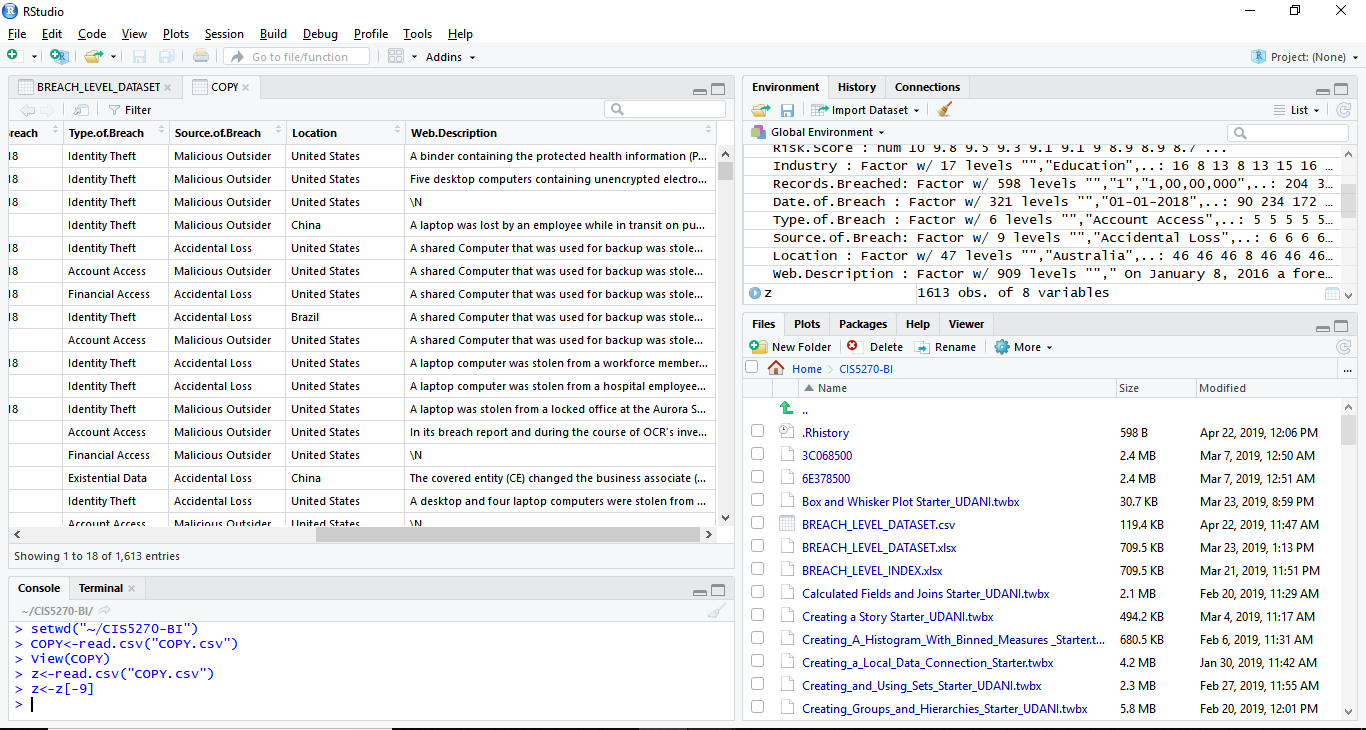
|  |  |  |
| --- | --- | --- |
| COLUMN NAME | DESCRIPTION | TYPE |
| Rank | Determines the Rank of the breach. | Number |
| Risk score | Determines the Risk score/Risk level of the breach. | Number |
| Industry | Shows the most common industries that suffer the breach. | Plain Text |
| Records Breached | States the number of records breached. | Number |
| Date of Breach | Shows the date when the breach occurred. | Date |
| Type of Breach | Tells us about the type of data breach that took place. | Plain Text |
| Source of Breach | Determines the source/reason of the actual data breach. | Plain Text |
| Location | States the location (country wise) where the breach happened. | Plain Text |

**DATA CLEANING**

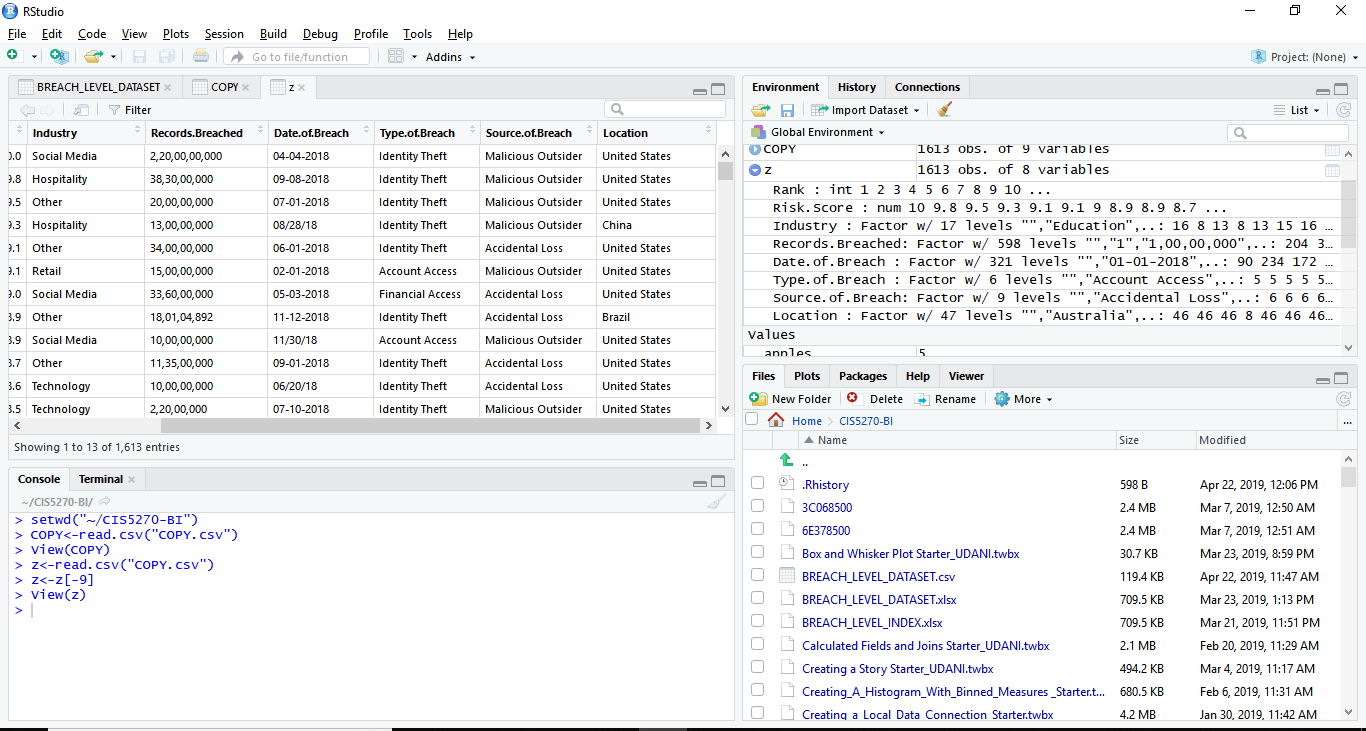
Data  cleaning  is especially  required when integrating data sources and should  be addressed  together with schema-related  data transformations. Data Cleansing or data scrubbing is the process of identifying and correcting inaccurate data from a data set.

1. **Deleting/Removing irrelevant column:** The dataset had a 9TH column named “Web Description” which had details about the url of the breach and moreover a long story of the breach that was not needed for the visualization purpose as well as did not add much value to the entire dataset since the column did not have specific/definite values, thus we removed the column from the dataset.

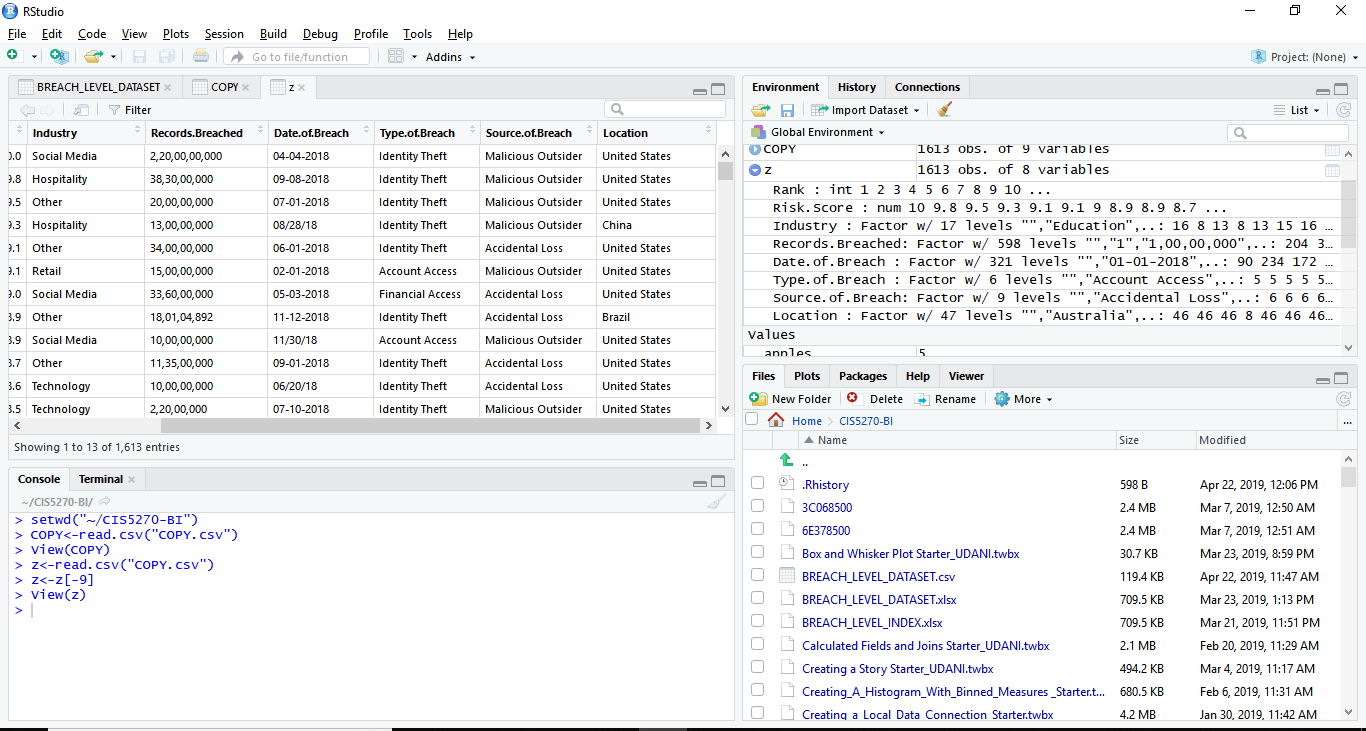
**Before:**



**After:**



**Code Screenshot:**



**R code:**

> setwd("~/CIS5270-BI")

> COPY<-read.csv("COPY.csv")

> View(COPY)

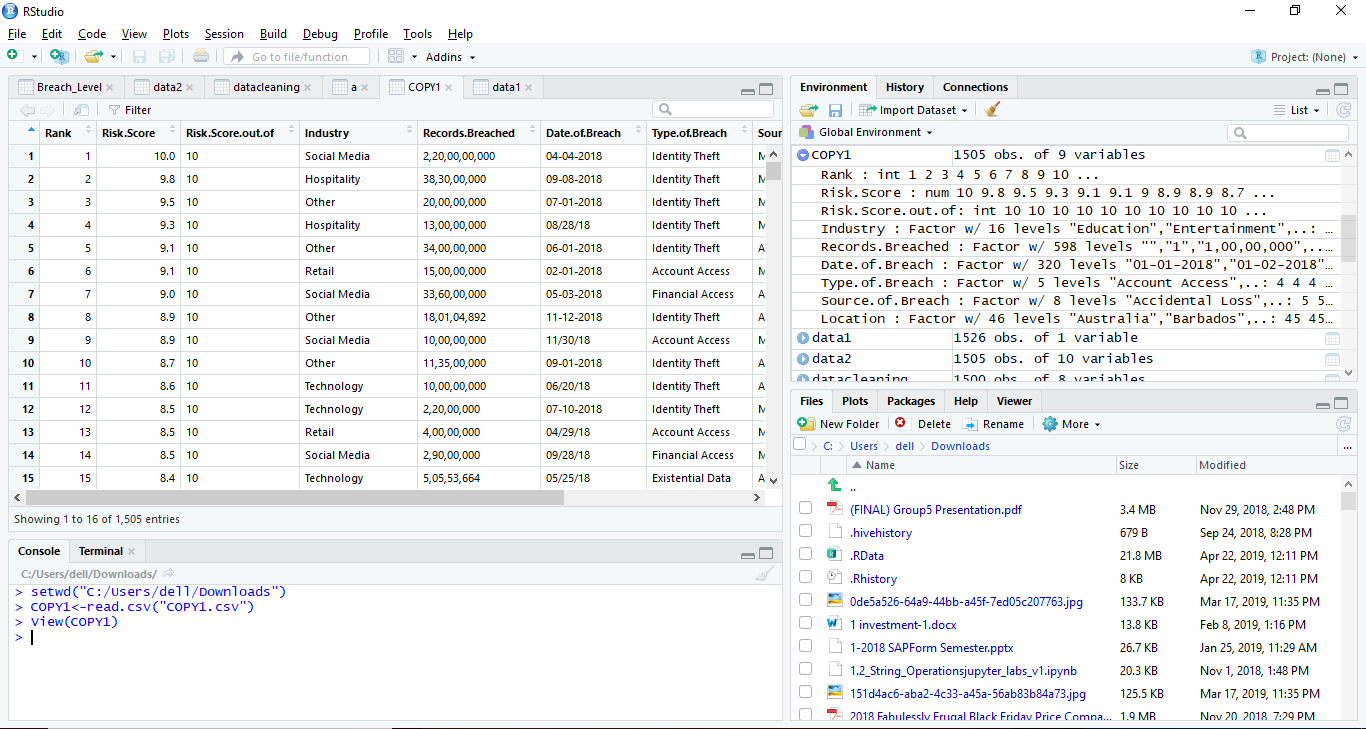
> z<-read.csv("COPY.csv")

> z<-z[-9]

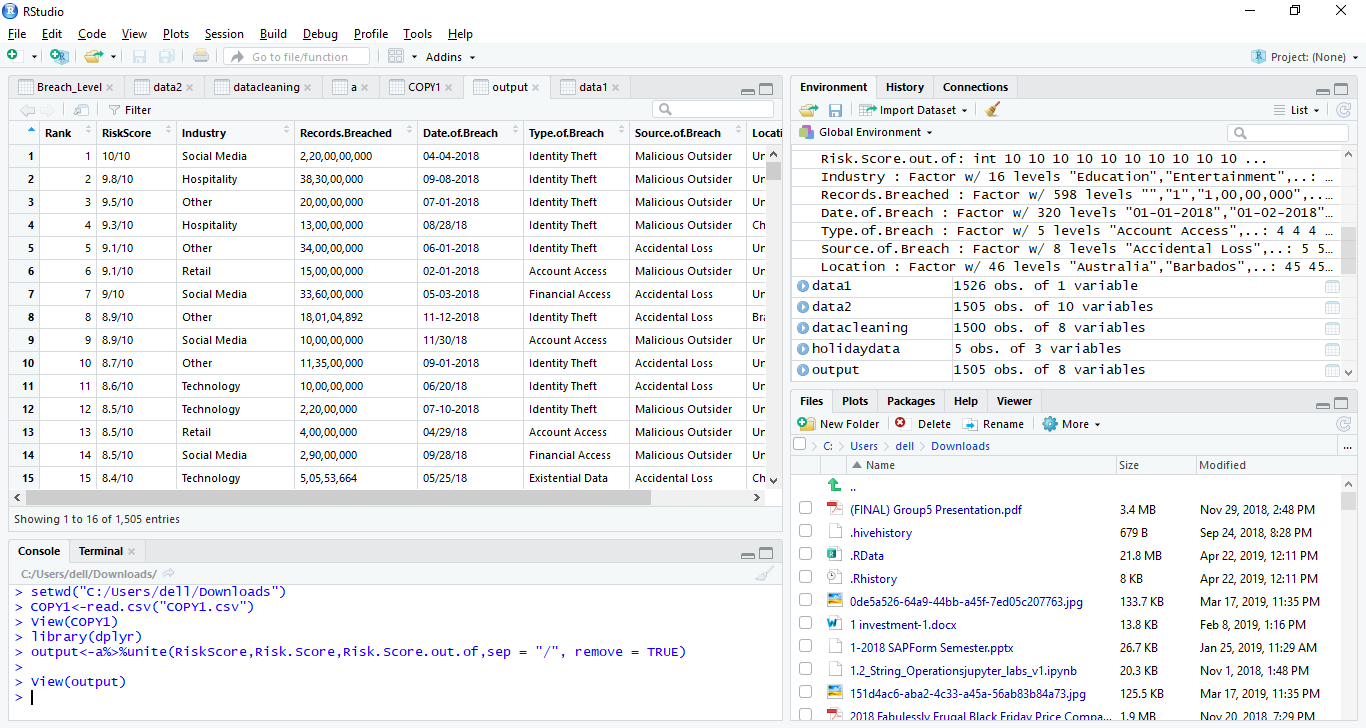
> View(z)

1. **Combining columns**: The information related to the risk score of the data breaches was split into 2 columns, however it made more sense to have that data together into one column which indicated the risk score for each data breach out of 10. Thus we decided to combine the 2 separate columns and have a merged column named as RiskScore.

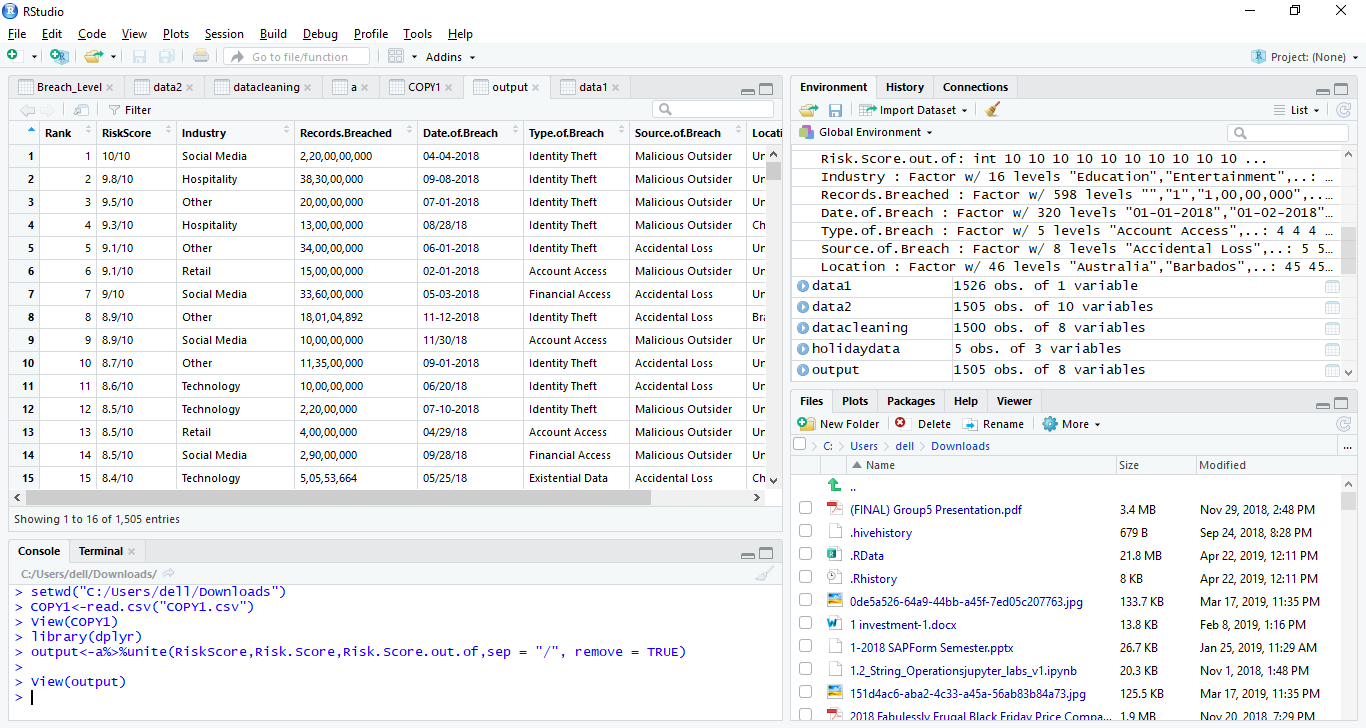
**Before**:



**After:**



**Code Screenshot:**



**R-code:**

>setwd("C:/Users/dell/Downloads")

> COPY1<-read.csv("COPY1.csv")

> View(COPY1)

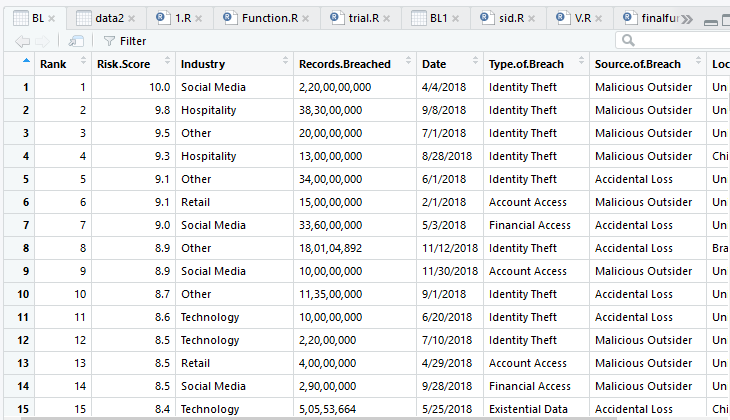
> library(dplyr)

> output<-a%>%unite(RiskScore,Risk.Score,Risk.Score.out.of,sep = "/", remove = TRUE)

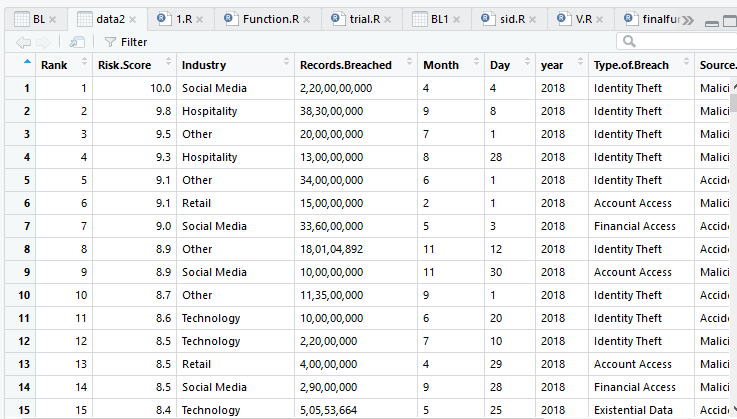
> View(output)

1. **Separating Date.of.breach column into ‘year’, ‘month’ and ‘day’:** We had a column named ‘date.of.breach’ which was split into three separate columns named: ‘year’, ‘month’ and ‘day’. This was done to simplify the data, so that we could have greater insight into the data by looking separately at the month and day when the breach occurred.

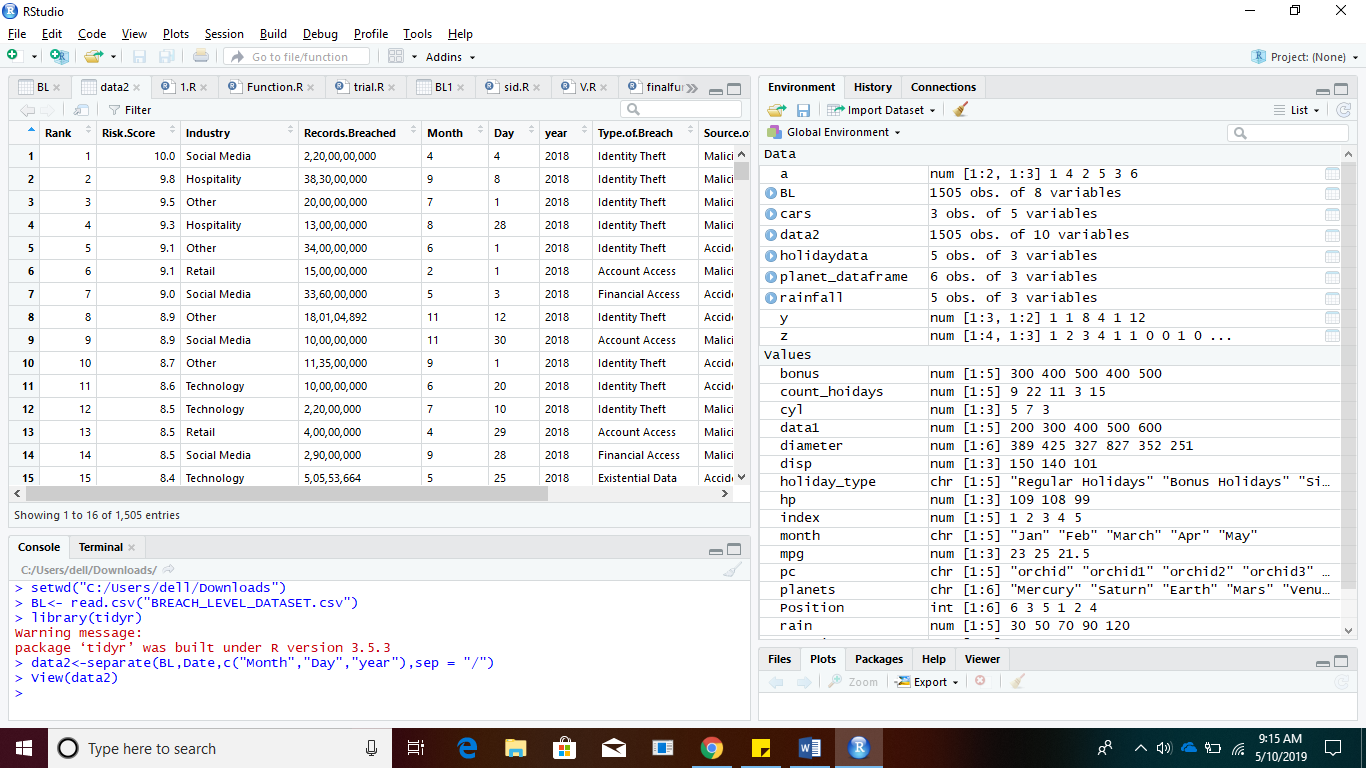
**Before:**



**After:**



**Code screenshot:**



**R code:**

>setwd("C:/Users/dell/Downloads")

> Breach\_Level<-read.csv("BREACH\_LEVEL.csv")

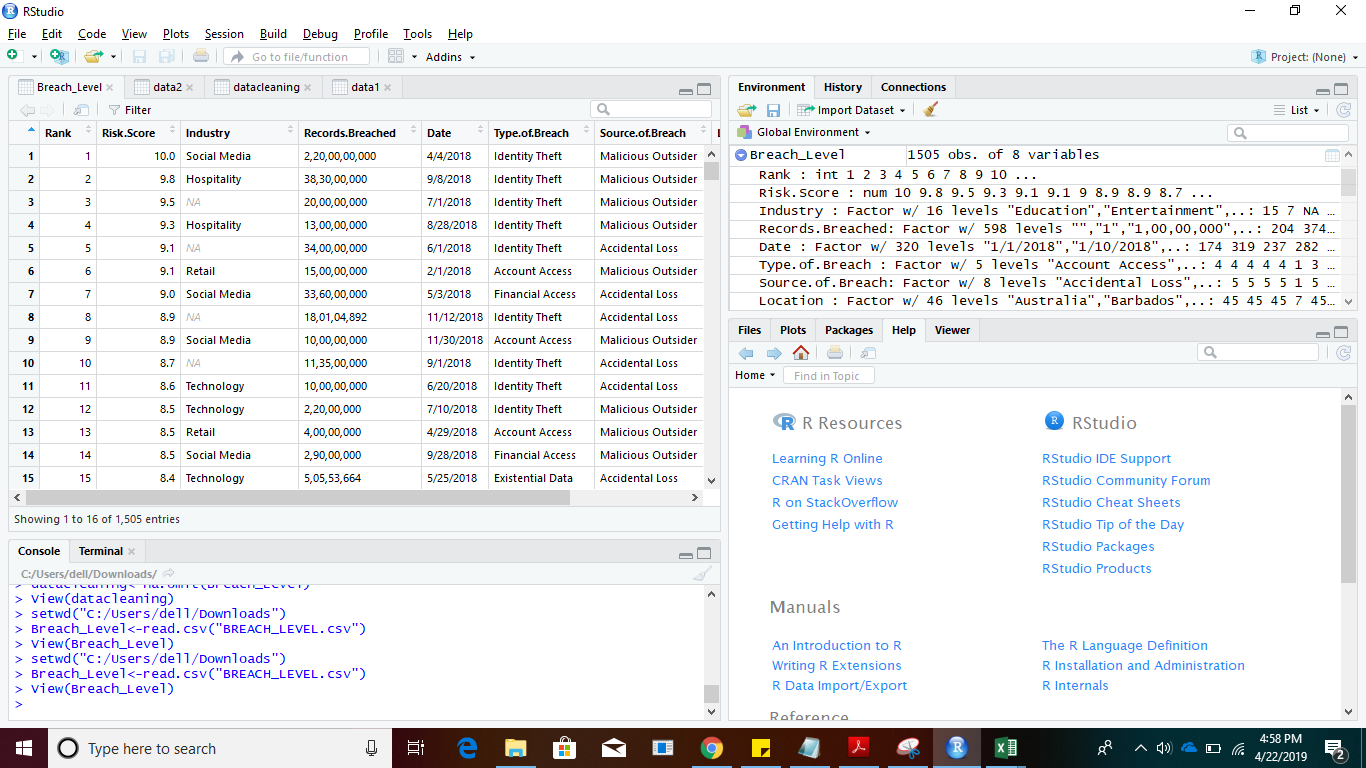
> View(Breach\_Level)

> data2<-separate(Breach\_Level,Date,c("year","month","day"),sep = "/")

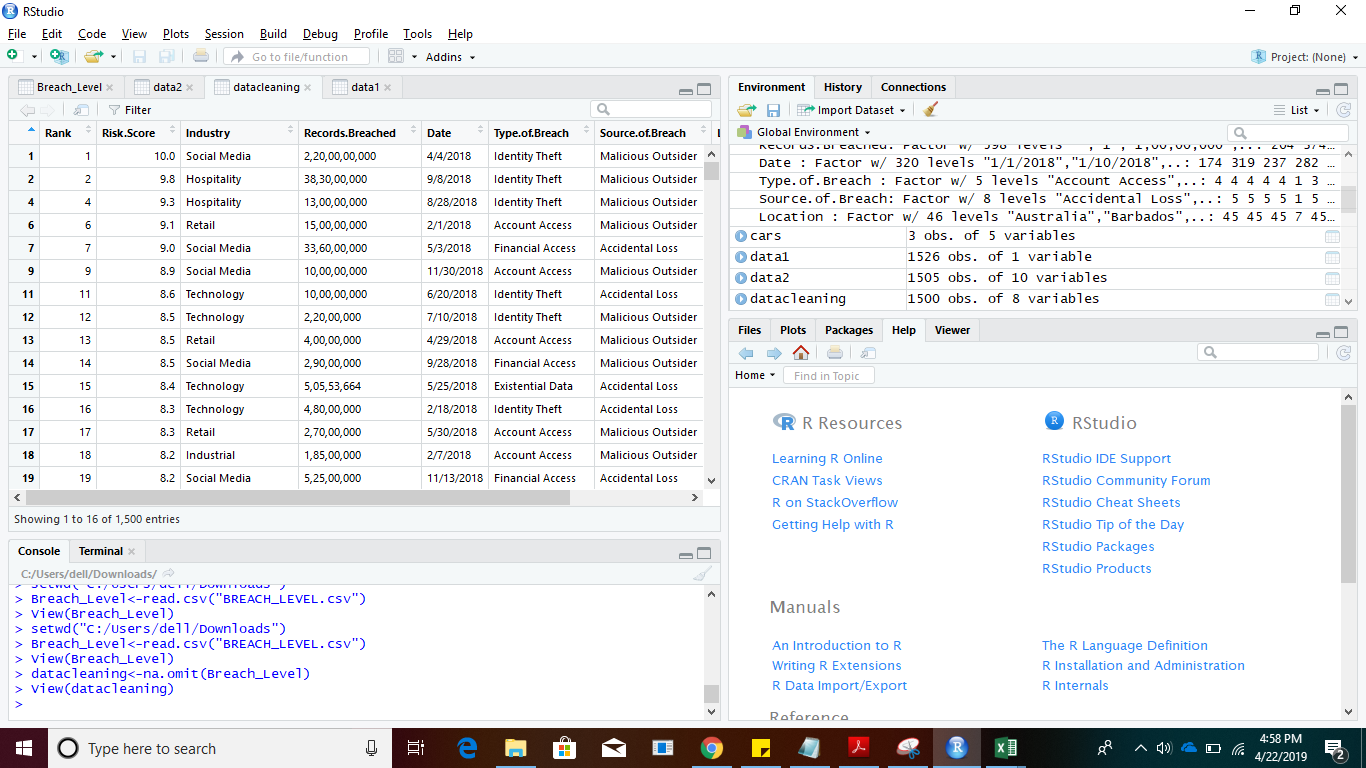
> View(data2)

1. **Removing NULL values:** The dataset contained NULL values in the ‘Industry’ column which were eliminated since they were not required and could act as misleading values for the visualizations.

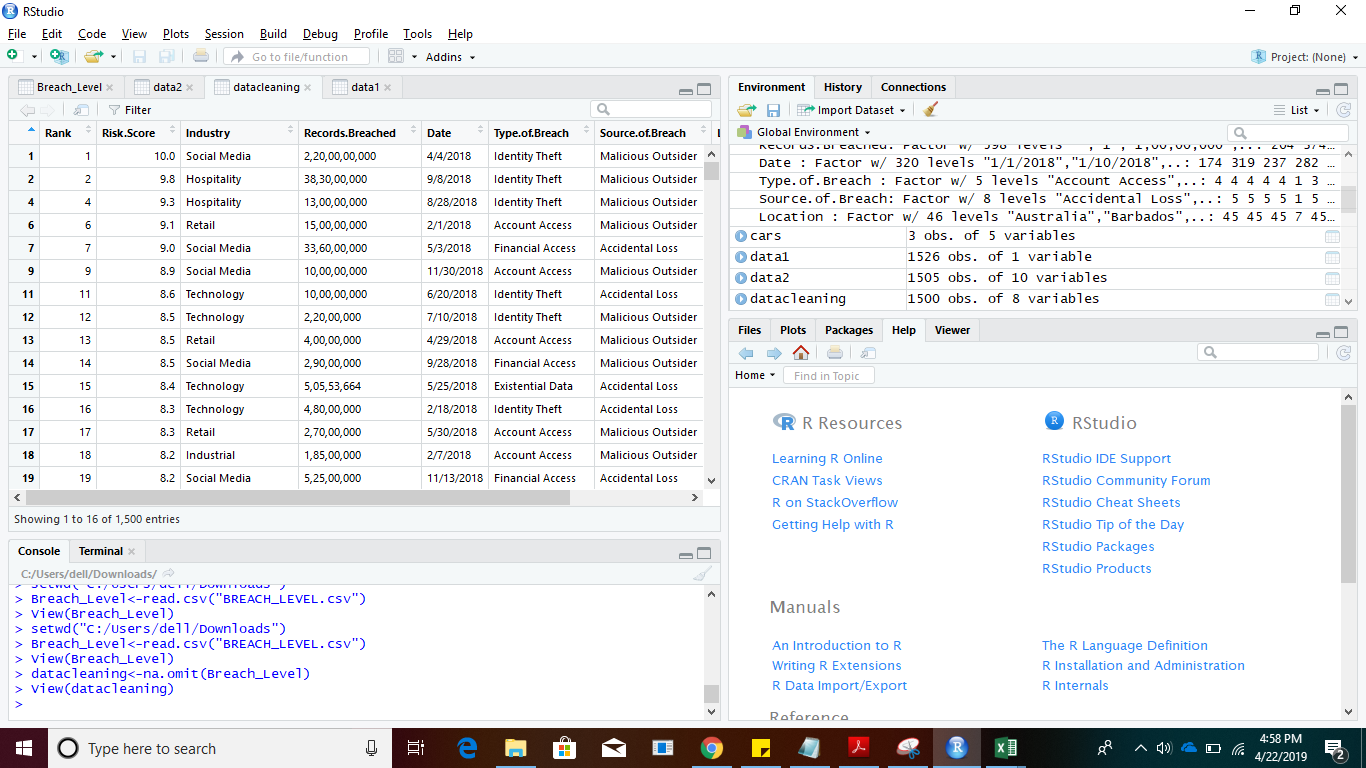
**Before:**



**After:**



**Code Screenshot:**



**R code:**

>setwd("C:/Users/dell/Downloads")

> Breach\_Level<-read.csv("BREACH\_LEVEL.csv")

> View(Breach\_Level)

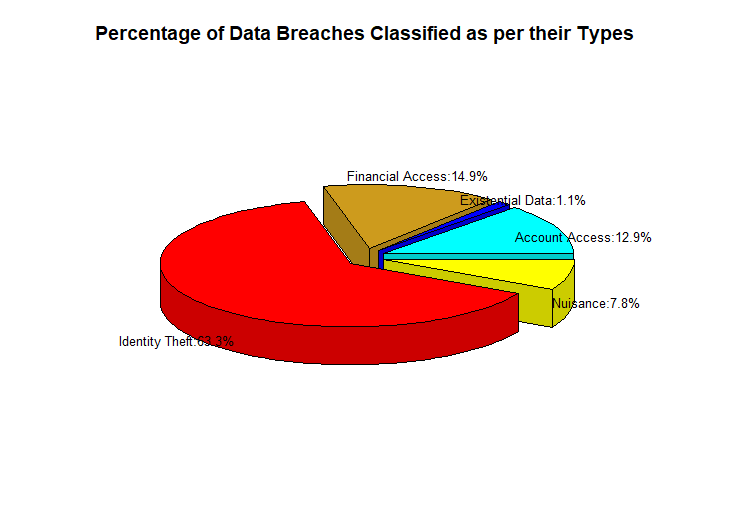
> datacleaning<-na.omit(Breach\_Level)

> View(datacleaning)

**ANALYSIS AND VISUALIZATIONS**

1. What is the percentage wise classification of the data breaches as per their type?

**Visual:**



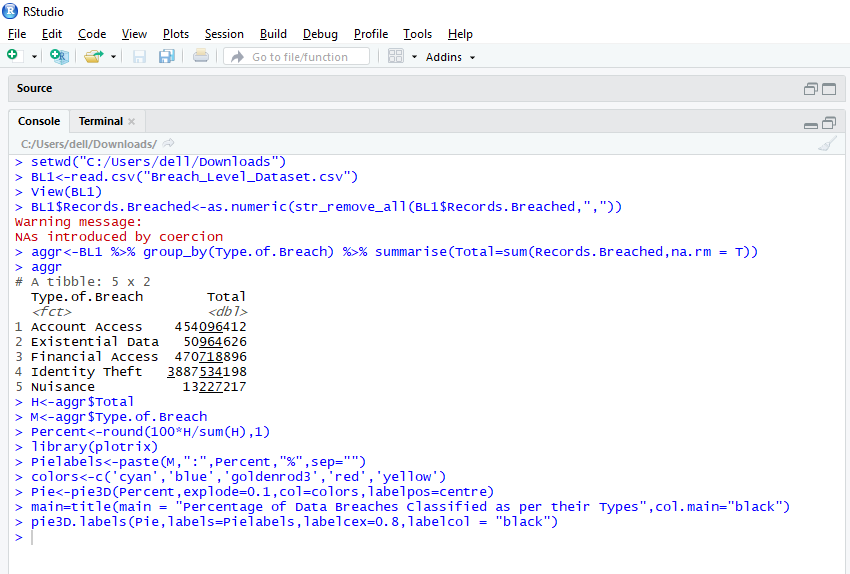
**Visualization**: Pie chart - 3D

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

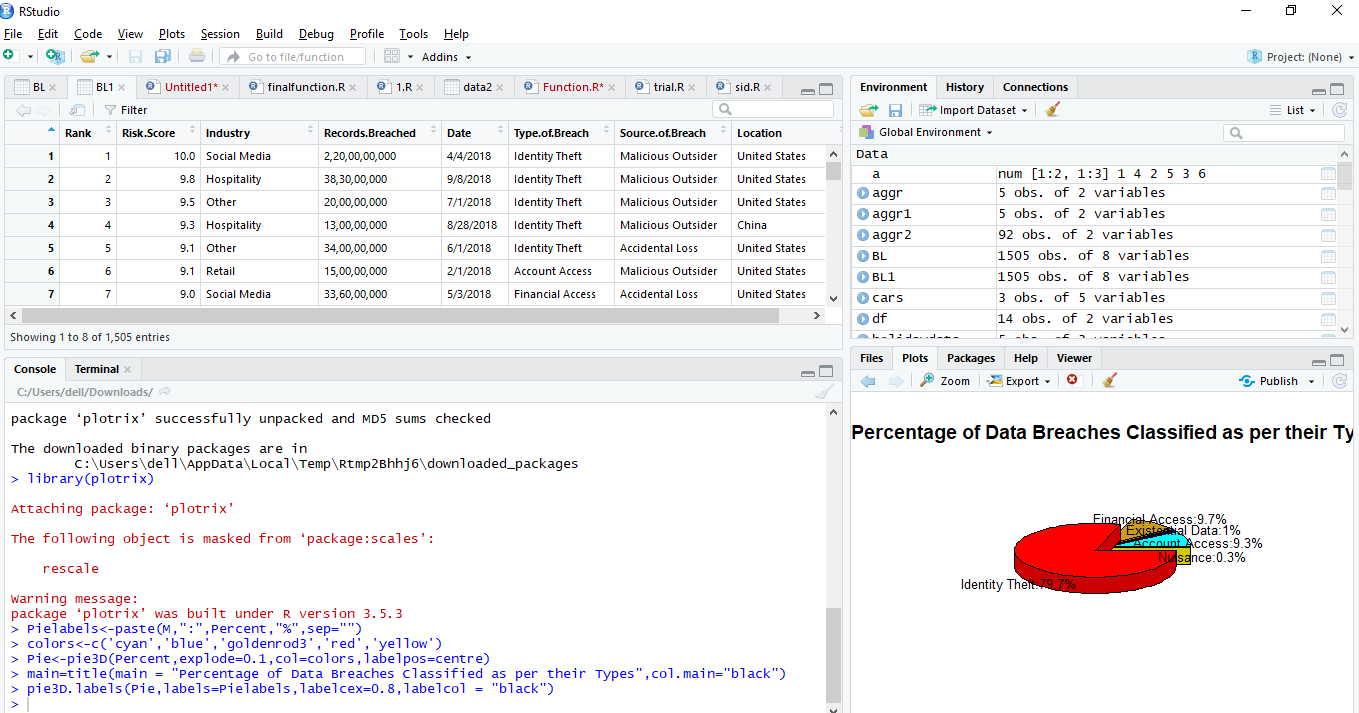
**Package/library**: plotrix

**Insight**: A number of important data breach trends emerged in the first half of 2018. **Identity theft** was once again the most prevalent data breach type tracked by the Breach level index. It accounted for approximately **63.3%** of the compromised records breached in 2018. Identifying any malicious or hidden code within incoming data files whether on your network or in a cloud is now a cyber imperative. What you don’t know or cannot see can harm you.[3]. This type of data breach can be caused by improper disposal of sensitive information, or simply leaving a confidential document in plain sight with an unknown individual or improperly guided access. The next most prevalent types of data breaches were financial access and account access. Nuisance and Existential data saw a comparatively decrease in the year 2018 counting for 7.8% and 1.1% respectively as a whole.

**R code:**

****

**R code with output:**



**R script:**

> setwd("C:/Users/dell/Downloads")

> BL1<-read.csv("Breach\_Level\_Dataset.csv")

> View(BL1)

> BL1$Records.Breached<-as.numeric(str\_remove\_all(BL1$Records.Breached,","))

Warning message:

NAs introduced by coercion

> aggr<-BL1 %>% group\_by(Type.of.Breach) %>% summarise(Total=sum(Records.Breached,na.rm = T))

> aggr

# A tibble: 5 x 2

 Type.of.Breach        Total

 <fct>                 <dbl>

1 Account Access    454096412

2 Existential Data   50964626

3 Financial Access  470718896

4 Identity Theft   3887534198

5 Nuisance           13227217

> H<-aggr$Total

> M<-aggr$Type.of.Breach

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

> library(plotrix)

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

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

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

> main=title(main = "Percentage of Data Breaches Classified as per their Types",col.main="black")

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

1. What was the impact of the breach sources on various industries in 2018?

**Visual:**

|  |
| --- |
| https://lh6.googleusercontent.com/uQktkt44DDKSUcoqe7ME9wXUqCGRQ_O48f8ibFcB90ExzNtJEU51IRyqDc-SSzLw0sWQ_UqDZ3dsl25WcnGSBSCAipVL51OCnU9ta-U5OF7nnDncpL6el2fDCNIDdf5wS3_MKc81 |

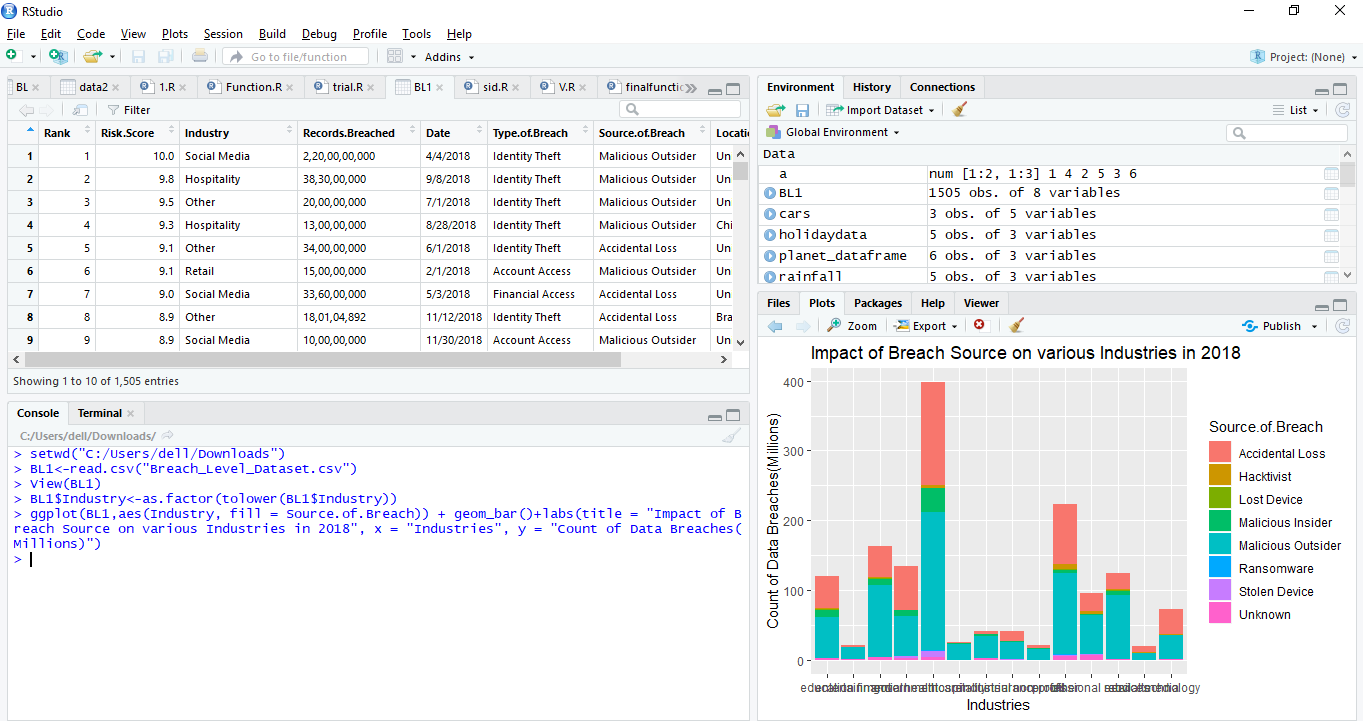
**Visualization**: Stacked Bar Chart

**Functions Used**: ggplot, aes, geom\_bar(), title

**Package/library**: ggplot2

**Insight**: Any company possessing sensitive data is under threat of being breached. Hackers can obtain any personal information, from names to heart rate data. One of the largest data breaches of 2018 proves this. In this visual we try to explain how the source of the data breaches have created an impact on various industries affected. The industry that was targeted the most in 2018 was the Healthcare industry, where roughly 400 Million records were breached. Healthcare organizations tend to [suffer from insider threats more than organizations in any other sector.](https://www.ekransystem.com/en/blog/healthcare-data-protection-solutions-monitor-and-audit-your-software) Financial, education, government and other industries were the next ones that were highly affected due to these breaches. An almost common trend that we’ve observed in all the bars is that, most of these breaches have been sourced from: Accidental Loss followed by Malicious outsider and various other sources with these two being the top.

**R code with output:**



**R script:**

>setwd("C:/Users/dell/Downloads")

> BL1<-read.csv("Breach\_Level\_Dataset.csv")

> View(BL1)

>BL1$Industry<-as.factor(tolower(BL1$Industry))

> install.packages("ggplot2")

> library(ggplot2)

>ggplot(BL1,aes(Industry, fill = Source.of.Breach)) + geom\_bar()+labs(title = "Impact of Breach Source on various Industries in 2018", x = "Industries", y = "Count of Data Breaches(Millions)")

1. **How are the records breached distributed as per their Risk Score?**

**Visual:**

|  |
| --- |
| **https://lh3.googleusercontent.com/2JFCrbTztK8m6vh3t2hiSY7wp97AAcoAtly82qVcoBjunQCttJVL1FaMEsF5-EH60VE9IsqqBZXNSdrPYdRmCz3p3OM9XuWmoZC8cbxnseS8Q4nQgKKdK4jcXrf5o8bRDhaGkAGD** |

**Visualization**: Scatter plot

**Functions Used**: ggplot, subset, plot, color, title

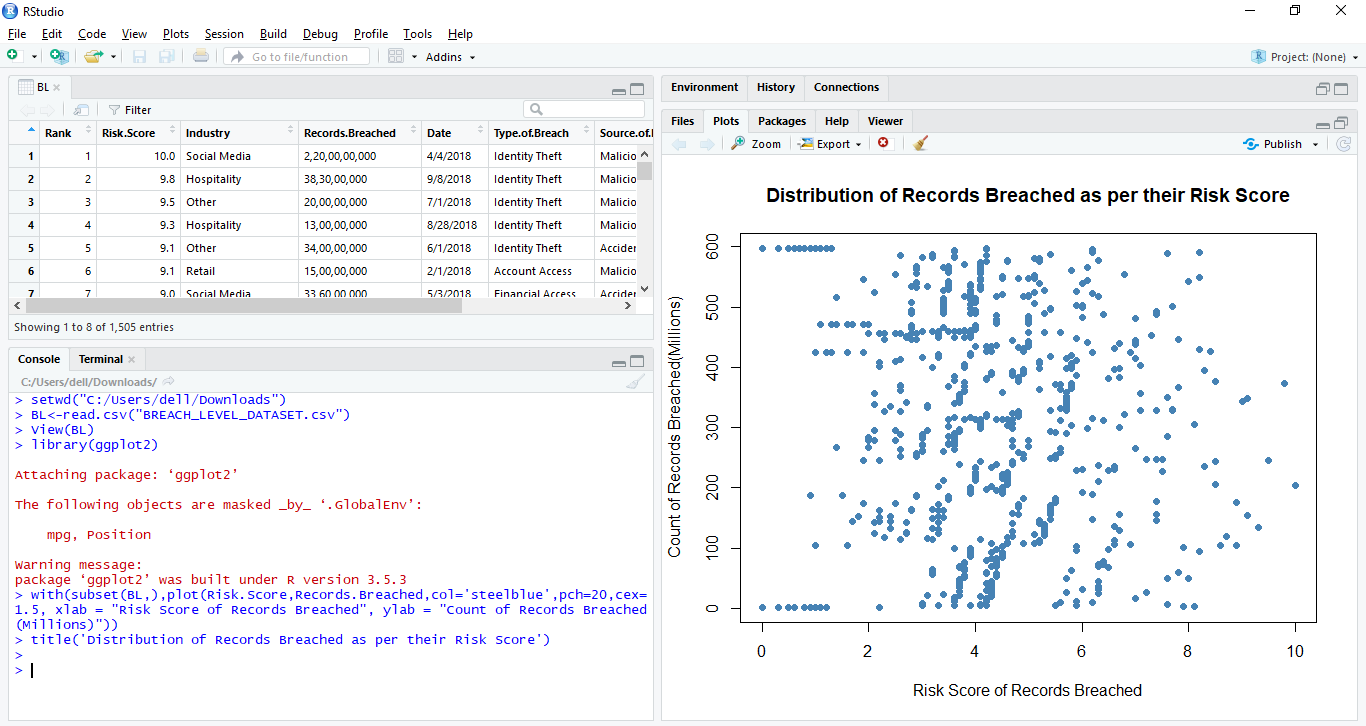
**Package/library**: ggplot2

**Insight:** Risks to data security include malicious programs, viruses, worms, hackers, identity theft,etc. Most people are aware of threats to their computers. These risk scores for the records breached are calculated to determine how severely your organization could be impacted if a security breach were to occur or has occurred. The visual indicates that the maximum records were breached with a risk score from 2 to 7, which indicate that most of them were really critical.

**The Breach Level Index determines the Breach Level security as follows:**

|  |  |  |
| --- | --- | --- |
| **Risk score** | **Severity** | **Effect** |
| 1-2.9 | Minimal | * A breach with no material effect. * Usually less than one thousand records. * Breach notification required, but little damage done. |
| 3-4.9 | Moderate | * A breach with low long-term business impact. * Usually involves the loss of several thousands of records of semi sensitive information. * Limited breach notification and financial exposure. |
| 5-6.9 | Critical | * A breach with likely short to midterm exposure to business. * Legal and/or regulatory impact. * Usually tens of thousands of records of moderate sensitive information involved. * Some breach notification and financial loss. |
| 7-8.9 | Severe | * A breach with likely short to midterm exposure to business. * Legal and/or regulatory impact. * Usually tens of thousands of records of moderate sensitive information involved. * Some breach notification and financial loss. |
| 9-10 | Catastrophic | * Breach with immense long term impact on breached ogranization, customers and/or partners. * Very large amount of highly sensitive information lost (usually 10-100+ million records). * Massive notification process. * Potentially existential financial loss for breached organization in remediation and related costs. * Use of lost sensitiv |

**R code with output:**

****

**R script:**

> setwd("C:/Users/dell/Downloads")

> BL<-read.csv("BREACH\_LEVEL\_DATASET.csv")

> View(BL)

> library(ggplot2)

> with(subset(BL,),plot(Risk.Score,Records.Breached,col='steelblue',pch=20,cex=1.5, xlab = "Risk Score of Records Breached", ylab = "Count of Records Breached(Millions)"))

> title('Distribution of Records Breached as per their Risk Score')

**Statistical Summary and Functions.**

**R Script for Summary:** The summary for our dataset has been shown below. It indicates that we have in all 1505 rows with 8 unique variables that have been used to draw various visualizations. It also specifies the data type for each variable, indicating whether it is int, num, char, factor, date format, etc. It also states the minimum value, maximum value, quartile range, median and mean for all the variables.

> str(BL1)

'data.frame': 1505 obs. of  8 variables:

$ Rank            : int 1 2 3 4 5 6 7 8 9 10 ...

$ Risk.Score      : num 10 9.8 9.5 9.3 9.1 9.1 9 8.9 8.9 8.7 ...

$ Industry        : Factor w/ 14 levels "education","entertainment",..: 13 6 10 6 10 12 13 10 13 10 ...

$ Records.Breached: num  2.20e+09 3.83e+08 2.00e+08 1.30e+08 3.40e+08 ...

$ Date            : Date, format: "2018-04-04" "2018-09-08" "2018-07-01" ...

$ Type.of.Breach  : Factor w/ 5 levels "Account Access",..: 4 4 4 4 4 1 3 4 1 4 ...

$ Source.of.Breach: Factor w/ 8 levels "Accidental Loss",..: 5 5 5 5 1 5 1 1 5 1 ...

$ Location        : Factor w/ 46 levels "Australia","Barbados",..: 45 45 45 7 45 45 45 4 45 45 ...

> summary(BL1)

     Rank        Risk.Score       Industry Records.Breached         Date

Min.   : 1.0  Min. : 0.000   healthcare:398 Min.   :1.000e+00 Min. :2018-01-01

1st Qu.: 376.8   1st Qu.: 1.200 other     :224 1st Qu.:1.140e+02 1st Qu.:2018-03-01

Median : 752.5   Median : 3.300 financial :163   Median :8.610e+02 Median :2018-05-01

Mean   : 752.9   Mean : 3.103   government:135 Mean   :4.405e+06 Mean :2018-05-14

3rd Qu.:1129.2   3rd Qu.: 4.500 retail    :124 3rd Qu.:8.000e+03 3rd Qu.:2018-07-11

Max.   :1505.0  Max. :10.000   education :121 Max.   :2.200e+09 Max. :2018-12-30

NA's   :1                   (Other) :340 NA's :398

Type.of.Breach           Source.of.Breach  Location

Account Access  :211 Malicious Outsider:834     United States :947

Existential Data: 15    Accidental Loss :506  Australia :331

Financial Access:212    Malicious Insider : 79  United Kingdom: 64

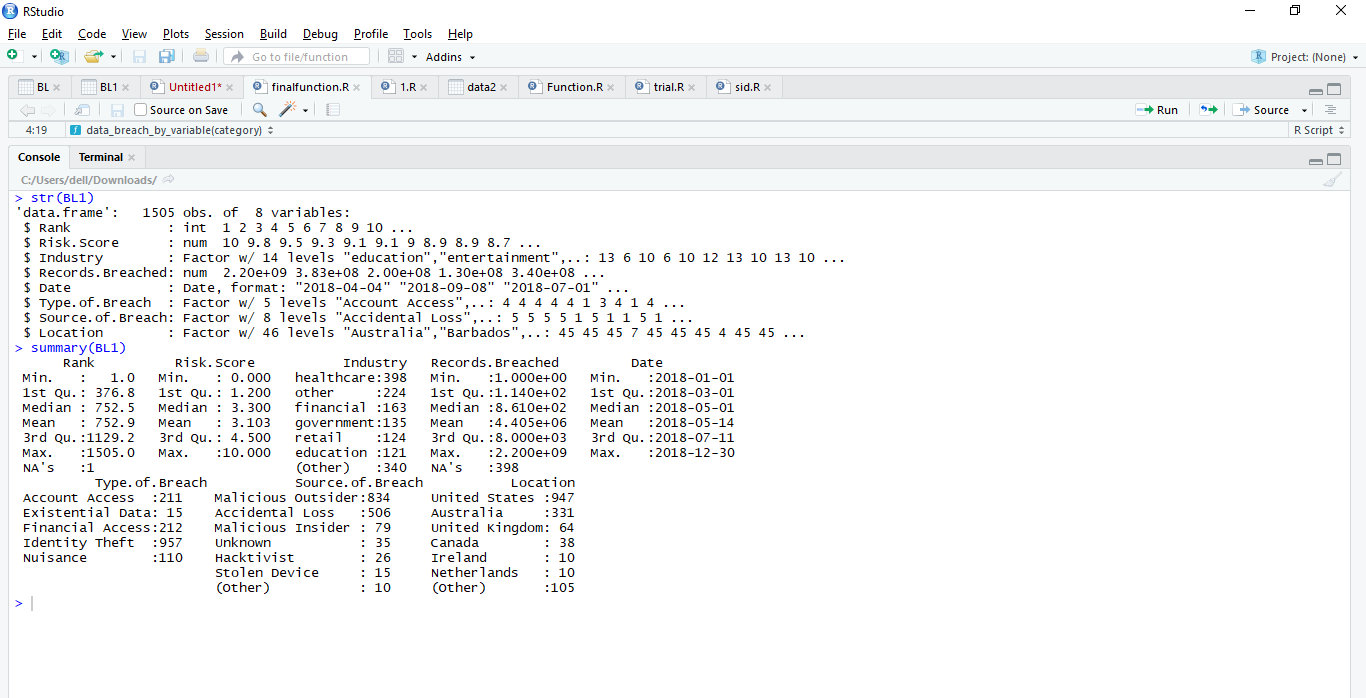
Identity Theft  :957 Unknown         : 35 Canada : 38

Nuisance        :110 Hacktivist        : 26 Ireland : 10

                        Stolen Device : 15  Netherlands : 10

                        (Other) : 10  (Other) :105

**R code with output:**

****

**Statistical summary:** In order to show the usage of the statistical functions, we’ve run the following script that demonstrates how to calculate the Summary, Mean, Median, Standard deviation, Mode, Minimum, Maximum, Range and Quantile for a specific variable: Risk Score.

> summary(BL$Risk.Score)

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

 0.000   1.200 3.300   3.103 4.500 10.000

> mean(BL$Risk.Score,na.rm=TRUE)

[1] 3.103389

> sd(BL$Risk.Score,na.rm=TRUE)

[1] 2.012741

> median(BL$Risk.Score, na.rm = TRUE)

[1] 3.3

> mode(BL$Risk.Score)

[1] "numeric"

> mode(as.numeric(BL$Risk.Score))

[1] "numeric"

> min(BL$Risk.Score)

[1] 0

> max(BL$Risk.Score)

[1] 10

> range(BL$Risk.Score)

[1]  0 10

> quantile(BL$Risk.Score,c(0.25,0.75))

25% 75%

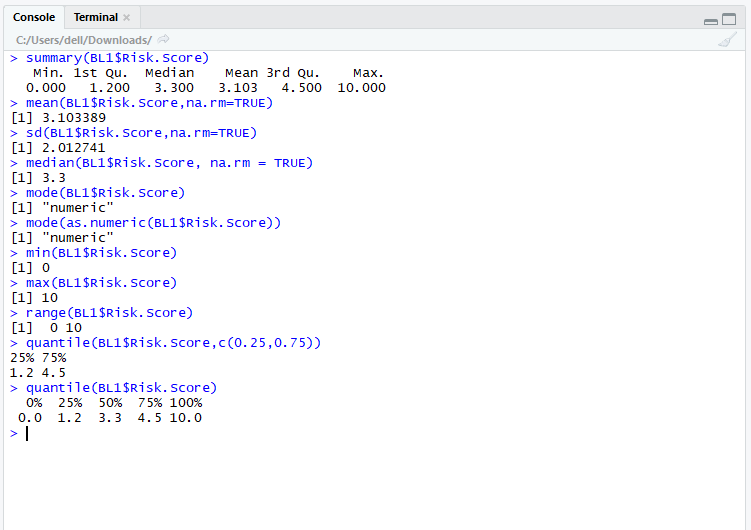
1.2 4.5

> quantile(BL$Risk.Score)

 0%  25% 50%  75% 100%

1. 1.2 3.3  4.5 10.0

**R code with output:**



**User defined function[1]:** The following user defined function has been made in accordance with the dataset: Breach Level Index.

**Question: What is the count of the number of records breached ‘Category’ wise?**

Here the category includes the 3 variables of the dataset, that is:

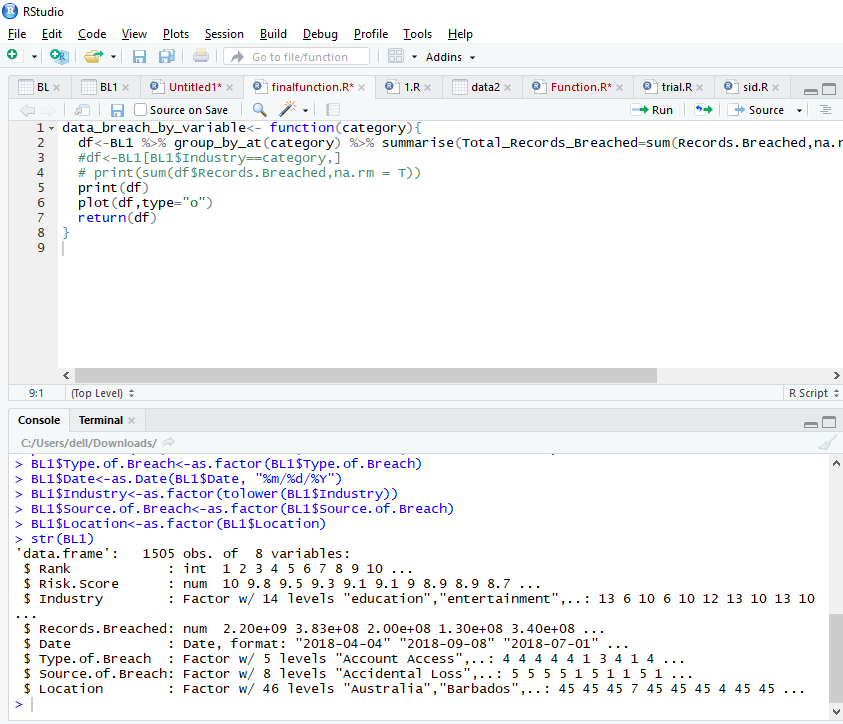
- Type of Breach

- Location

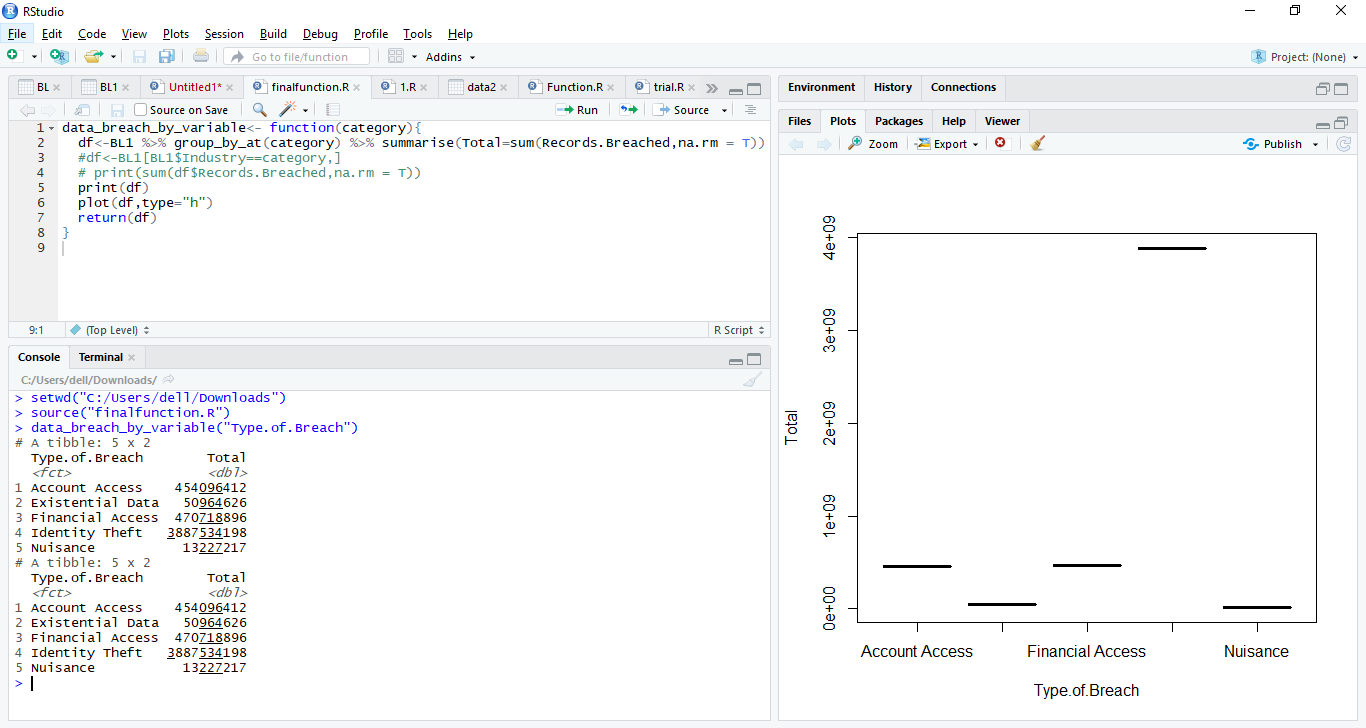
- Source of breach

(Say for example, if the user directly wants to know how many records have been Breached Industry wise( or any category) he can directly get an answer to it, that is, the script would take the category as an input from the User himself.)

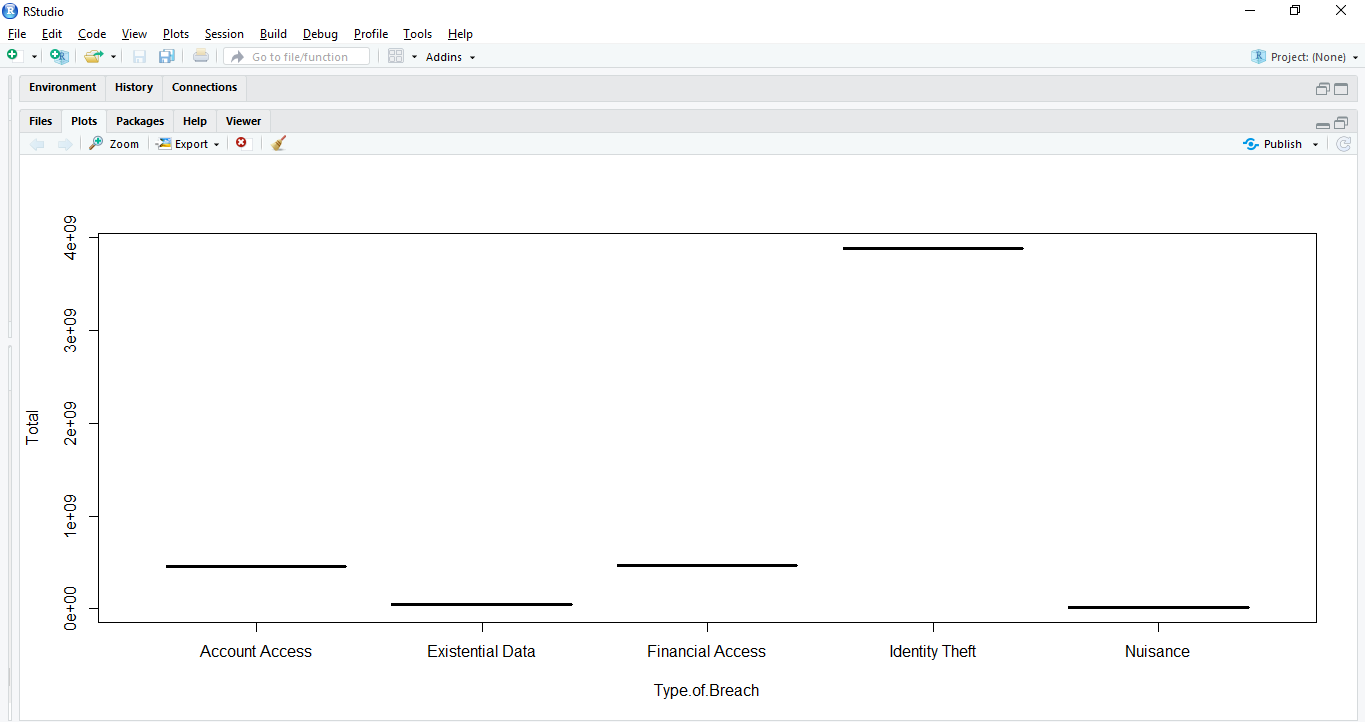
**R code [1]:**

****

**R code with Output: category: ‘type of breach’**

****

**Visual:**

****

**Insight:** As per the visual, when we classify the records on the basis of their types, it is seen that ‘Identity Theft’ is the main type of data breach that took place in 2018 and this clearly reflects from our very first visualization of pie chart. It caused around 40 Million data breaches. After ‘Identity Theft’, the other types of Breach followed are ‘ Account Access and Financial Access’**.**

**R script for the function:**

data\_breach\_by\_variable<- function(category){

df<-BL1%>%group\_by\_at(category)%>% summarise(Total\_Records\_Breached=sum(Records.Breached,na.rm = T))

 #df<-BL1[BL1$Industry==category,]

 # print(sum(df$Records.Breached,na.rm = T))

 print(df)

 plot(df,type="o")

 return(df)

}

**R code for running the user defined function:**

> setwd("C:/Users/dell/Downloads")

> source("finalFunction.R")

> data\_breach\_by\_variable("Type.of.Breach")

# A tibble: 5 x 2

 Type.of.Breach   Total\_Records\_Breached

 <fct>                             <dbl>

1 Account Access                454096412

2 Existential Data               50964626

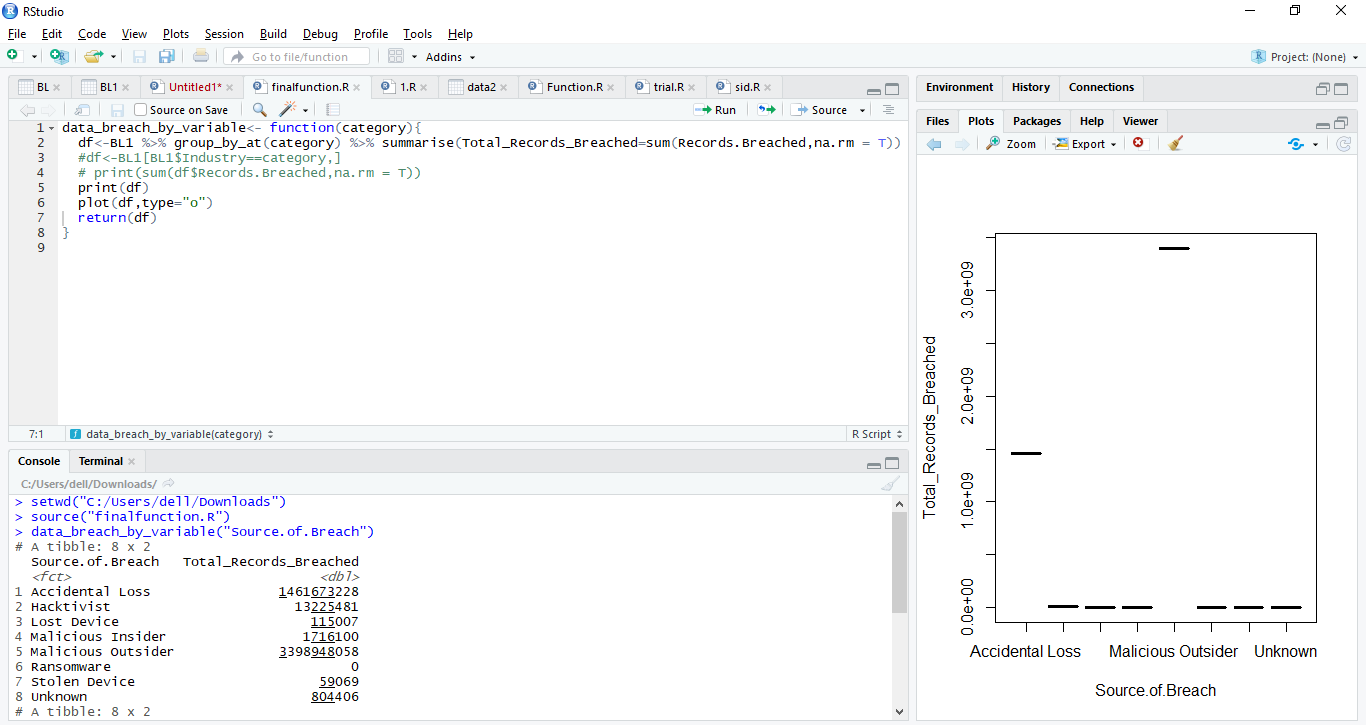
3 Financial Access              470718896

4 Identity Theft               3887534198

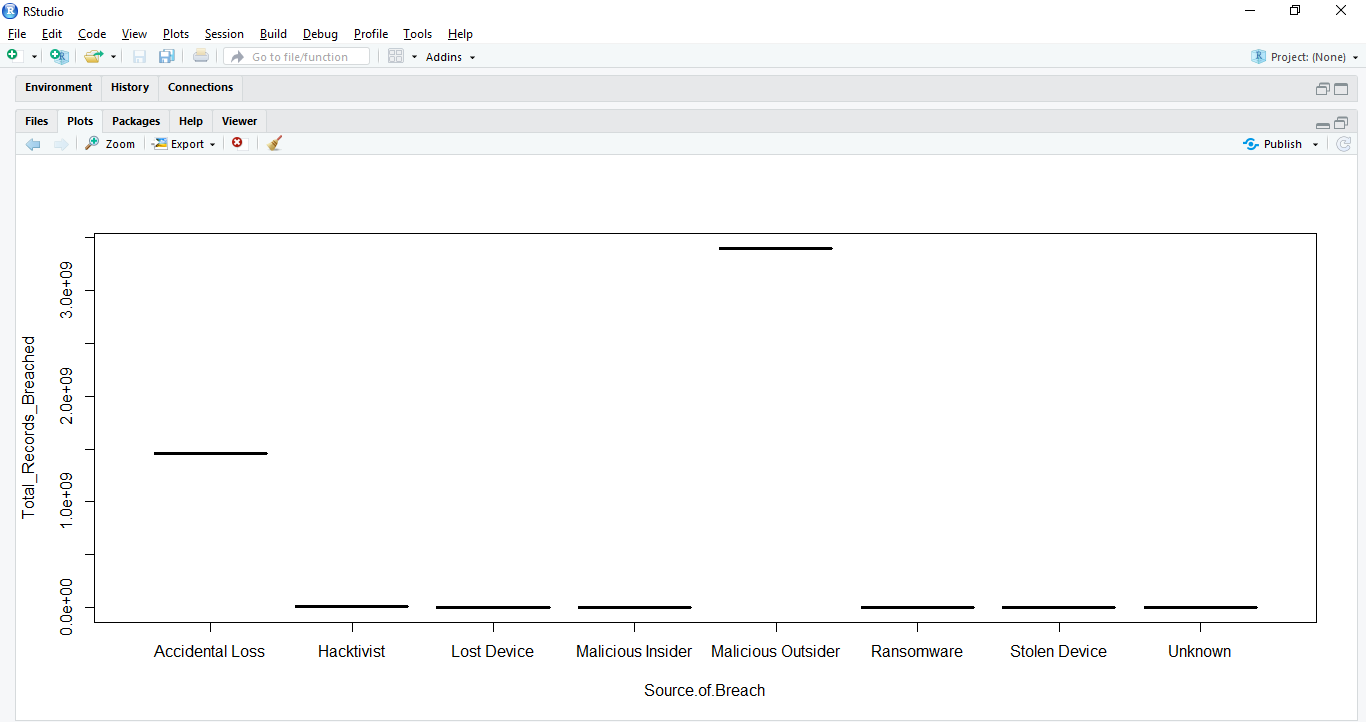
5 Nuisance                       13227217

**R code [2]:**

**R code with Output: category: ‘Source of breach’**

****

**Visual:**

****

**Insight:** From the visual it is clearly understandable that the data breaches that make the news are typically carried out by Malicious outsider, followed by Accidental data loss. Malicious Outsiders set up a website capable of exploiting any computer that browses to it, then they send emails to the insiders that entice them to click a link to that site. Mostemployees will not take the bait, but it just takes one person to give in to curiosity and click the link.

**R script for the function:**

data\_breach\_by\_variable<- function(category){

df<-BL1%>%group\_by\_at(category)%>% summarise(Total\_Records\_Breached=sum(Records.Breached,na.rm = T))

 #df<-BL1[BL1$Industry==category,]

 # print(sum(df$Records.Breached,na.rm = T))

 print(df)

 plot(df,type="o")

 return(df)

}

**R code for running the user defined function:**

> setwd("C:/Users/dell/Downloads")

> source("finalFunction.R")

> data\_breach\_by\_variable("Source.of.Breach")

# A tibble: 8 x 2

 Source.of.Breach   Total\_Records\_Breached

<fct>                               <dbl>

1 Accidental Loss                1461673228

2 Hacktivist                       13225481

3 Lost Device                        115007

4 Malicious Insider                 1716100

5 Malicious Outsider             3398948058

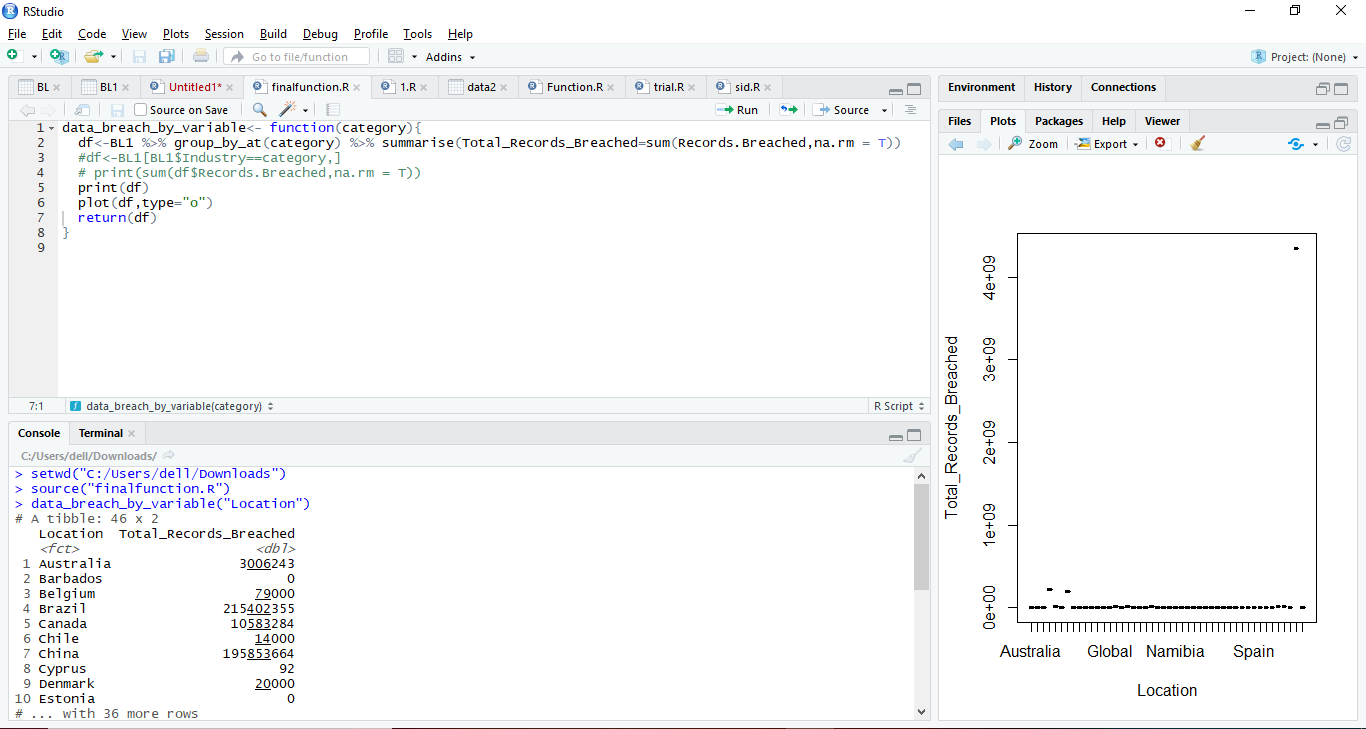
6 Ransomware                              0

7 Stolen Device                       59069

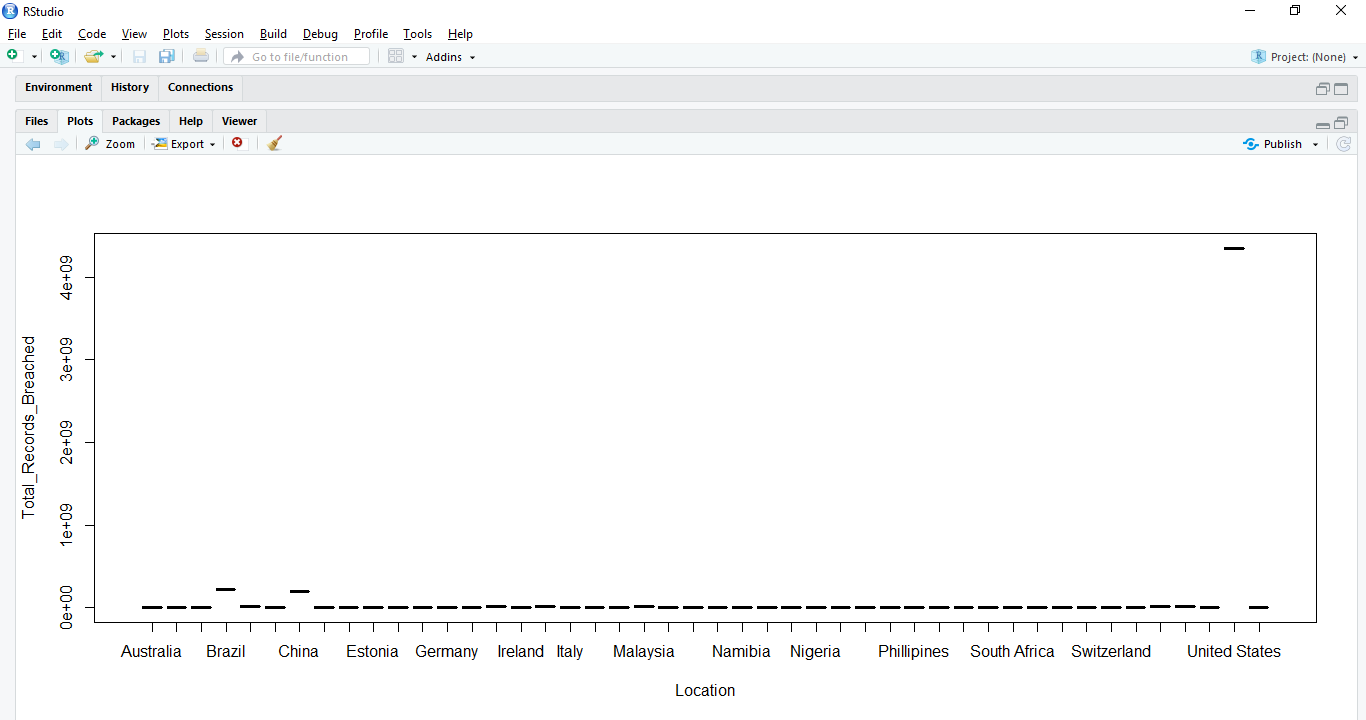
8 Unknown                            804406

**R code [3]:**

**R code with Output: category: ‘Location’**

****

**Visual:**

****

**Insight:** Data security is one of the largest concerns impacting the world today. The number I’d lost or stolen data records varies around the world.Broken down by country, United States led the way in the number of both compromised records and security incidents. The most biggest U.S companies affected by these data breaches were Yahoo, Marriott International, EBay, Equifax, Target Stores, Uber, etc. The count of these records breach crosses approximately 50 Millions in 2018. Following US, the other countries affected were Brazil and China and so on.

**R script for the function:**

data\_breach\_by\_variable<- function(category){

df<-BL1%>%group\_by\_at(category)%>% summarise(Total\_Records\_Breached=sum(Records.Breached,na.rm = T))

 #df<-BL1[BL1$Industry==category,]

 # print(sum(df$Records.Breached,na.rm = T))

 print(df)

 plot(df,type="o")

 return(df)

}

**R code for running the user defined function:**

> setwd("C:/Users/dell/Downloads")

> source("finalFunction.R")

> data\_breach\_by\_variable("Location")

# A tibble: 46 x 2

  Location  Total\_Records\_Breached

  <fct>                      <dbl>

1 Australia                3006243

2 Barbados                       0

3 Belgium                    79000

4 Brazil                 215402355

5 Canada                  10583284

6 Chile                      14000

7 China                  195853664

8 Cyprus                        92

9 Denmark                    20000

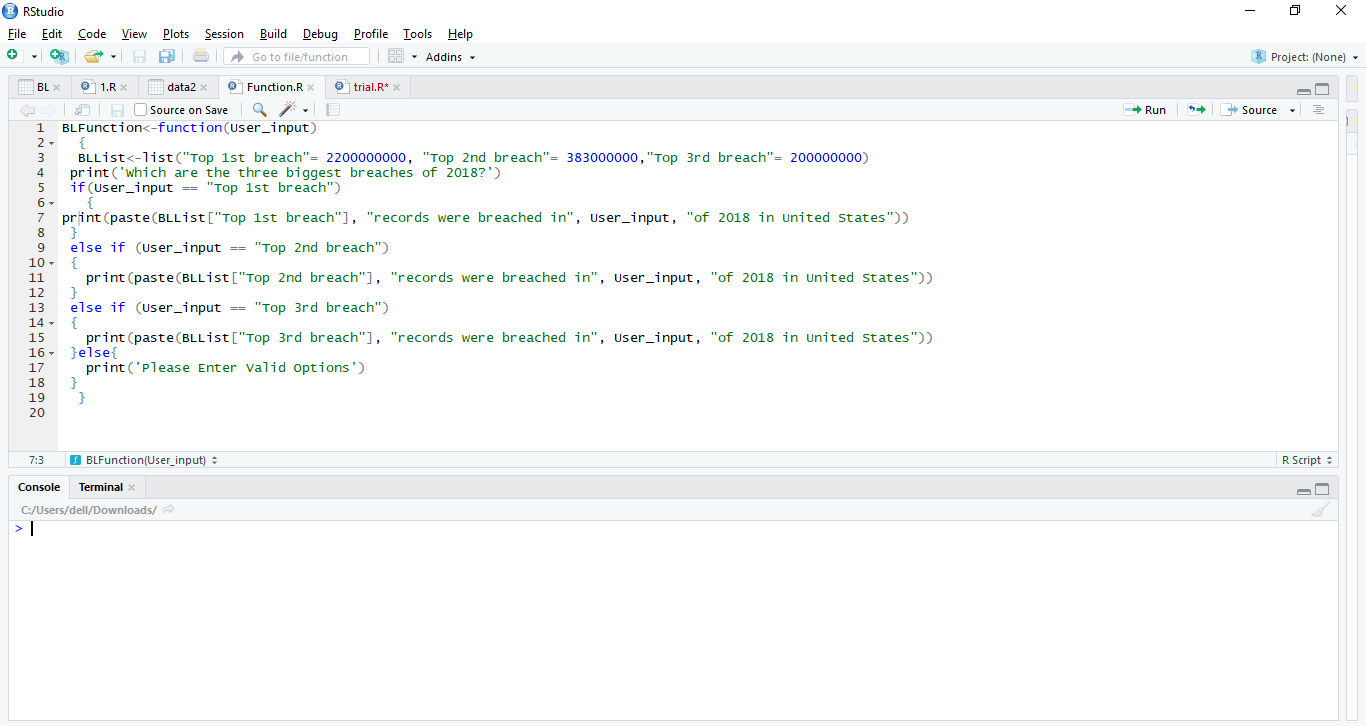
10 Estonia                        0

**User defined function [2]:** Since the previous visual indicated that the highest number of records were breached in **United States,** we executed this user defined function to state the top 3 breaches of the United States in 2018 as per Breach Leve Index.

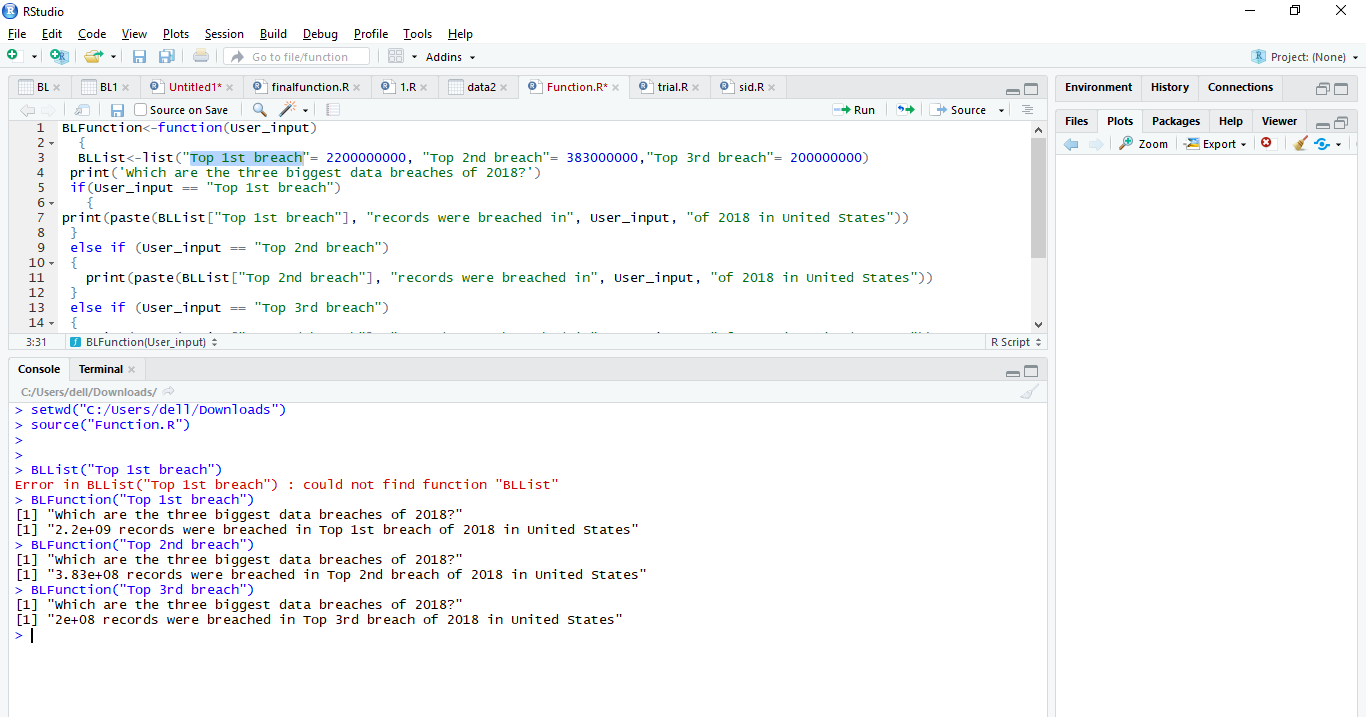
**Question: Which are the Top Three breaches of United States in 2018?**

(To display the output, we take input from the user such as: Top 1st Breach, Top 2nd Breach and Top 3rd Breach)

**R code :**

****

**R code with output:**



**R script and output:**

BLFunction<-function(User\_input)

 {

 BLList<-list("Top 1st breach"= 2200000000, "Top 2nd breach"= 383000000,"Top 3rd breach"= 200000000)

print('Which are the three biggest data breaches of 2018?')

if(User\_input == "Top 1st breach")

  {

print(paste(BLList["Top 1st breach"], "records were breached in", User\_input, "of 2018 in United States"))

}

else if (User\_input == "Top 2nd breach")

{

  print(paste(BLList["Top 2nd breach"], "records were breached in", User\_input, "of 2018 in United States"))

}

else if (User\_input == "Top 3rd breach")

{

  print(paste(BLList["Top 3rd breach"], "records were breached in", User\_input, "of 2018 in United States"))

}else{

  print('Please Enter Valid Options')

}

 }

**R code for running the user defined function:**

> setwd("C:/Users/dell/Downloads")

> source("Function.R")

> BLList("Top 1st breach")

Error in BLList("Top 1st breach") : could not find function "BLList"

> BLFunction("Top 1st breach")

[1] "Which are the three biggest data breaches of 2018?"

[1] "2.2e+09 records were breached in Top 1st breach of 2018 in United States"

> BLFunction("Top 2nd breach")

[1] "Which are the three biggest data breaches of 2018?"

[1] "3.83e+08 records were breached in Top 2nd breach of 2018 in United States"

> BLFunction("Top 3rd breach")

[1] "Which are the three biggest data breaches of 2018?"

[1] "2e+08 records were breached in Top 3rd breach of 2018 in United States"

**Complete R Script:**

* **Data cleaning codes**

**Deleting/Removing irrelevant column:**

> setwd("~/CIS5270-BI")

> COPY<-read.csv("COPY.csv")

> View(COPY)

> z<-read.csv("COPY.csv")

> z<-z[-9]

> View(z)

**Combining columns:**

>setwd("C:/Users/dell/Downloads")

> COPY1<-read.csv("COPY1.csv")

> View(COPY1)

> library(dplyr)

> output<-a%>%unite(RiskScore,Risk.Score,Risk.Score.out.of,sep = "/", remove = TRUE)

> View(output)

**Separating Date.of.breach column into ‘year’, ‘month’ and ‘day’:**

>setwd("C:/Users/dell/Downloads")

> Breach\_Level<-read.csv("BREACH\_LEVEL.csv")

> View(Breach\_Level)

> data2<-separate(Breach\_Level,Date,c("year","month","day"),sep = "/")

> View(data2)

**Removing NULL values:**

>setwd("C:/Users/dell/Downloads")

> Breach\_Level<-read.csv("BREACH\_LEVEL.csv")

> View(Breach\_Level)

> datacleaning<-na.omit(Breach\_Level)

> View(datacleaning)

* **Visualizations code:**

**Visual 1:**

> setwd("C:/Users/dell/Downloads")

> BL1<-read.csv("Breach\_Level\_Dataset.csv")

> View(BL1)

> BL1$Records.Breached<-as.numeric(str\_remove\_all(BL1$Records.Breached,","))

Warning message:

NAs introduced by coercion

> aggr<-BL1 %>% group\_by(Type.of.Breach) %>% summarise(Total=sum(Records.Breached,na.rm = T))

> aggr

# A tibble: 5 x 2

 Type.of.Breach        Total

 <fct>                 <dbl>

1 Account Access    454096412

2 Existential Data   50964626

3 Financial Access  470718896

4 Identity Theft   3887534198

5 Nuisance           13227217

> H<-aggr$Total

> M<-aggr$Type.of.Breach

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

> library(plotrix)

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

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

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

> main=title(main = "Percentage of Data Breaches Classified as per their Types",col.main="black")

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

**Visual 2:**

>setwd("C:/Users/dell/Downloads")

> BL1<-read.csv("Breach\_Level\_Dataset.csv")

> View(BL1)

>BL1$Industry<-as.factor(tolower(BL1$Industry))

> install.packages("ggplot2")

> library(ggplot2)

>ggplot(BL1,aes(Industry, fill = Source.of.Breach)) + geom\_bar()+labs(title = "Impact of Breach Source on various Industries in 2018", x = "Industries", y = "Count of Data Breaches(Millions)")

**Visual 3:**

> setwd("C:/Users/dell/Downloads")

> BL<-read.csv("BREACH\_LEVEL\_DATASET.csv")

> View(BL)

> library(ggplot2)

> with(subset(BL,),plot(Risk.Score,Records.Breached,col='steelblue',pch=20,cex=1.5, xlab = "Risk Score of Records Breached", ylab = "Count of Records Breached(Millions)"))

> title('Distribution of Records Breached as per their Risk Score')

* **Summary code:**

> str(BL1)

'data.frame': 1505 obs. of  8 variables:

$ Rank            : int 1 2 3 4 5 6 7 8 9 10 ...

$ Risk.Score      : num 10 9.8 9.5 9.3 9.1 9.1 9 8.9 8.9 8.7 ...

$ Industry        : Factor w/ 14 levels "education","entertainment",..: 13 6 10 6 10 12 13 10 13 10 ...

$ Records.Breached: num  2.20e+09 3.83e+08 2.00e+08 1.30e+08 3.40e+08 ...

$ Date            : Date, format: "2018-04-04" "2018-09-08" "2018-07-01" ...

$ Type.of.Breach  : Factor w/ 5 levels "Account Access",..: 4 4 4 4 4 1 3 4 1 4 ...

$ Source.of.Breach: Factor w/ 8 levels "Accidental Loss",..: 5 5 5 5 1 5 1 1 5 1 ...

$ Location        : Factor w/ 46 levels "Australia","Barbados",..: 45 45 45 7 45 45 45 4 45 45 ...

> summary(BL1)

     Rank        Risk.Score       Industry Records.Breached         Date

Min.   : 1.0  Min. : 0.000   healthcare:398 Min.   :1.000e+00 Min. :2018-01-01

1st Qu.: 376.8   1st Qu.: 1.200 other     :224 1st Qu.:1.140e+02 1st Qu.:2018-03-01

Median : 752.5   Median : 3.300 financial :163   Median :8.610e+02 Median :2018-05-01

Mean   : 752.9   Mean : 3.103   government:135 Mean   :4.405e+06 Mean :2018-05-14

3rd Qu.:1129.2   3rd Qu.: 4.500 retail    :124 3rd Qu.:8.000e+03 3rd Qu.:2018-07-11

Max.   :1505.0  Max. :10.000   education :121 Max.   :2.200e+09 Max. :2018-12-30

NA's   :1                   (Other) :340 NA's :398

Type.of.Breach           Source.of.Breach  Location

Account Access  :211 Malicious Outsider:834     United States :947

Existential Data: 15    Accidental Loss :506  Australia :331

Financial Access:212    Malicious Insider : 79  United Kingdom: 64

Identity Theft  :957 Unknown         : 35 Canada : 38

Nuisance        :110 Hacktivist        : 26 Ireland : 10

                        Stolen Device : 15  Netherlands : 10

                        (Other) : 10  (Other) :105

* **Statistical Summary code:**

> summary(BL$Risk.Score)

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

 0.000   1.200 3.300   3.103 4.500 10.000

> mean(BL$Risk.Score,na.rm=TRUE)

[1] 3.103389

> sd(BL$Risk.Score,na.rm=TRUE)

[1] 2.012741

> median(BL$Risk.Score, na.rm = TRUE)

[1] 3.3

> mode(BL$Risk.Score)

[1] "numeric"

> mode(as.numeric(BL$Risk.Score))

[1] "numeric"

> min(BL$Risk.Score)

[1] 0

> max(BL$Risk.Score)

[1] 10

> range(BL$Risk.Score)

[1]  0 10

> quantile(BL$Risk.Score,c(0.25,0.75))

25% 75%

1.2 4.5

> quantile(BL$Risk.Score)

 0%  25% 50%  75% 100%

1. 1.2 3.3  4.5 10.0

* **User Defined Function code[1]:**

**R code [1]:**

data\_breach\_by\_variable<- function(category){

df<-BL1 %>% group\_by\_at(category) %>% summarise(Total\_Records\_Breached=sum(Records.Breached,na.rm = T))

 #df<-BL1[BL1$Industry==category,]

 # print(sum(df$Records.Breached,na.rm = T))

 print(df)

 plot(df,type="o")

 return(df)

}

> setwd("C:/Users/dell/Downloads")

> source("finalFunction.R")

> data\_breach\_by\_variable("Type.of.Breach")

# A tibble: 5 x 2

 Type.of.Breach   Total\_Records\_Breached

 <fct>                             <dbl>

1 Account Access                454096412

2 Existential Data               50964626

3 Financial Access              470718896

4 Identity Theft               3887534198

5 Nuisance                       13227217

**R code [2]:**

data\_breach\_by\_variable<- function(category){

df<-BL1 %>% group\_by\_at(category) %>% summarise(Total\_Records\_Breached=sum(Records.Breached,na.rm = T))

 #df<-BL1[BL1$Industry==category,]

 # print(sum(df$Records.Breached,na.rm = T))

 print(df)

 plot(df,type="o")

 return(df)

}

> setwd("C:/Users/dell/Downloads")

> source("finalFunction.R")

> data\_breach\_by\_variable("Source.of.Breach")

# A tibble: 8 x 2

 Source.of.Breach   Total\_Records\_Breached

<fct>                               <dbl>

1 Accidental Loss                1461673228

2 Hacktivist                       13225481

3 Lost Device                        115007

4 Malicious Insider                 1716100

5 Malicious Outsider             3398948058

6 Ransomware                              0

7 Stolen Device                       59069

8 Unknown                            804406

**R code [3]:**

data\_breach\_by\_variable<- function(category){

df<-BL1 %>% group\_by\_at(category) %>% summarise(Total\_Records\_Breached=sum(Records.Breached,na.rm = T))

 #df<-BL1[BL1$Industry==category,]

 # print(sum(df$Records.Breached,na.rm = T))

 print(df)

 plot(df,type="o")

 return(df)

}

> setwd("C:/Users/dell/Downloads")

> source("finalFunction.R")

> data\_breach\_by\_variable("Location")

# A tibble: 46 x 2

  Location  Total\_Records\_Breached

  <fct>                      <dbl>

1 Australia                3006243

2 Barbados                       0

3 Belgium                    79000

4 Brazil                 215402355

5 Canada                  10583284

6 Chile                      14000

7 China                  195853664

8 Cyprus                        92

9 Denmark                    20000

Estonia                       0

* **User Defined Function code[2]:**

BLFunction<-function(User\_input)

 {

 BLList<-list("Top 1st breach"= 2200000000, "Top 2nd breach"= 383000000,"Top 3rd breach"= 200000000)

print('Which are the three biggest data breaches of 2018?')

if(User\_input == "Top 1st breach")

  {

print(paste(BLList["Top 1st breach"], "records were breached in", User\_input, "of 2018 in United States"))

}

else if (User\_input == "Top 2nd breach")

{

  print(paste(BLList["Top 2nd breach"], "records were breached in", User\_input, "of 2018 in United States"))

}

else if (User\_input == "Top 3rd breach")

{

  print(paste(BLList["Top 3rd breach"], "records were breached in", User\_input, "of 2018 in United States"))

}else{

  print('Please Enter Valid Options')

}

 }

> setwd("C:/Users/dell/Downloads")

> source("Function.R")

> BLList("Top 1st breach")

Error in BLList("Top 1st breach") : could not find function "BLList"

> BLFunction("Top 1st breach")

[1] "Which are the three biggest data breaches of 2018?"

[1] "2.2e+09 records were breached in Top 1st breach of 2018 in United States"

> BLFunction("Top 2nd breach")

[1] "Which are the three biggest data breaches of 2018?"

[1] "3.83e+08 records were breached in Top 2nd breach of 2018 in United States"

> BLFunction("Top 3rd breach")

[1] "Which are the three biggest data breaches of 2018?"

[1] "2e+08 records were breached in Top 3rd breach of 2018 in United States"

**References**

[1] Security Breaches in Healthcare in the Last Three Years. (2018, March 30). Retrieved March 9, 2019, from <https://www.hipaajournal.com/security-breaches-in-healthcare-in-the-last-three-years/>

[2] Wikipedia. (2019, March 29). Data breach. Retrieved from <https://en.wikipedia.org/wiki/Data_breach>

[3] Vaynberg, B. (2018, May 21). Understanding The Seven Types Of Data Breach. Retrieved from <https://chiefexecutive.net/understanding-seven-types-data-breach/>