Purpose

The purpose of assignment two in Part 1: From Data to Graphs is to try to understand chapter 4, "From Data to Graphs", in the textbook, Applied Security Visualization, by Raffael Marty. Chapter in the textbook describes what to do after data are collected, describes how often to take data from different sources in combining them, and describes hard to understand data. To demonstrate all of that, come up with an example to do data analysis. Such data analysis must follow the six steps of the Information Visualization Process, as outlined in chapter 4 of the textbook. Furthermore, such data analysis must define a problem to solve and a question that requires an answer. As well, such data analysis must choose a problem with some complexity to compare data from at least two different sources.

Requirements

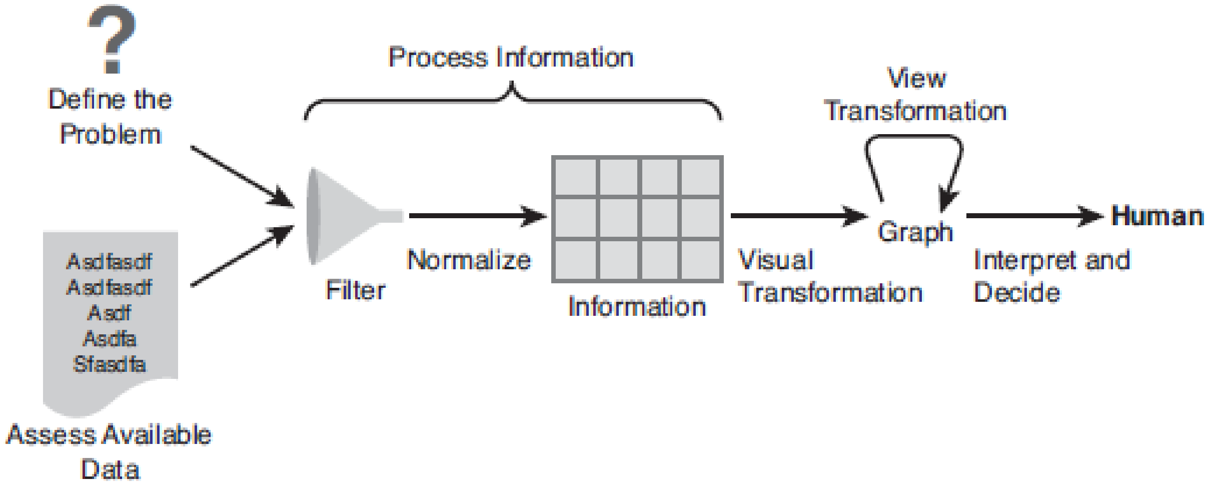
- Windows-based operating system — current version 10
- RStudio version 1.1.423
- R statistical computing programming language version 3.4.3

Objectives

- Show stops that are involved to transform textual data into graphical representations
- Introduce a process to systematically define a problem, choose needed dataset, transform dataset into something be meaningful, and arrive a graphical representation of a problem

Information Visualization Process

Visualization of data must start with a clear objective and a clearly defined problem with various decisions that can be made. To generate useful graphs, plots, and diagrams, these must be meaningful to address a clear objective and to help solve a clearly defined problem. Below is a diagram of an information visualization process, as outlined in the textbook. Starting from the left in below diagram has a problem that is identified. Based from that identified problem, specific data sources are identified to answer questions being posed from that identified problem. Then various decisions are made to eventually arrive graphical representations that are interpretable and decidable.



Abstract

Step 1: Define the Problem

Visualization of data is always case-driven rather than data-driven. Firstly, when analyzing a large dataset, a normal person might not understand all fields in that large dataset; instead, focus on a clear objective. Focus only on those fields in a large dataset to answer a problem. Secondly, when analyzing a large dataset, try to identify anomalies. Try to define as many cases and to verify those cases by visualization. In this step of an information visualization process, the following questions are tried to be answered in defining a problem.

- What is it that you are interested in?
- What questions need to be answered by the graph that you are about to generate?
- What do you expect to see?
- What would you like to see?
- Are there any anomalies?

While extracting malware host detections from a Comma-Separated Values (csv) dataset, there are several interesting points of view that can be looked at for security data analysis. These are malware hosts with highest number of malware hosts detection, malware hosts with different threat prevention programs being employed of them, and malware hosts with highest number of threat preventions. Importantly, if there are generated graphs, these graphs need to answer how many malware hosts have been detected by at least one threat prevention program. Converting these interesting points of view into questions, some questions that are likely to be answered by plotting graphs are:

- Which malware hosts are most identified to be malign?
- List at least 10 different threat prevention programs that show up often.
- Which malware hosts have most number of threat prevention programs be associated?
- How many malware hosts are there with at least one threat prevention program being associated? List some of them with same threat prevention programs.

Some expectations of malware host detections from a dataset are the following. There are going to be many malware hosts, including outliers in having no count of threat prevention programs being associated. Even if all threat prevention programs are identified, some of those threat prevention programs are not being associated as much as other threat prevention programs with malware hosts. A greater number of fewer threat prevention programs between one and three are being associated with malware hosts.

Supposedly, there should be no outliers, such that there should be no none and no zero count of threat prevention programs being associated with malware hosts. Dataset of malware host detections could identify which threat prevention program has higher priority or is more effective against each malware host being associated. Also, dataset of malware host detections has some malware hosts as domain names or has some malware hosts as Internet Protocol (IP) addresses; instead, dataset of malware host detections could have both domain names and IP addresses of malware hosts. To add more details, dataset of malware host detections could provide physical locations of identified malware hosts by longitude and latitude. If possible, include port numbers of identified malware hosts.

So far, from inspecting dataset of malware host detections, there seems to be no obvious anomalies besides having zero or none count of threat prevention programs being associated with malware hosts.

Description of Environment

Step 2: Assess Available Data

In this step two, first thing is to find out what type of data is needed to answer those posed questions in step one. An important question that arises in this step is the following to reside to other methods in addressing problems in step one.

- What pieces of data do you need?

In addressing such question to figure out what types of malware hosts are being associated with threat prevention programs, logs from network-based intrusion detection systems and intrusion prevention systems are deployed for monitoring malware hosts. Although dataset of malware hosts detection does not offer types of protocols or types of flags that are being involved with malware hosts, threat prevention programs could indicate whether those identified malware hosts are malign or benign. Dataset of malware host detections also do not offer extra information, such as whether threat prevention programs are identifying malware hosts behind a firewall, through a web proxy server, or under a specific network infrastructure. However, the following question is answerable.

- What type of data do we have available now?

Right now, dataset of malware hosts detection is a table of three columns. Column one is a list of identified malware hosts, which can be duplicated. Column two is a list of threat prevention programs being associated with those malware hosts. For each malware host, there can be more than one threat prevention program. Column three is a list of count on number of threat prevention programs for each malware host.

Step 3: Process Information

At this step three, problems have been identified as well as having available data. Essentially, this step three is to format those data, so that visualization tools are used to generate graphical representations, thereby transforming data into information. Transforming data into information adds metadata to data themselves in giving explanations, meanings, or semantics to data, known as parsing. Parsing basically takes raw data files and extracts those data files to be put into individual fields of meaningful information. Some examples of parsers for various types of raw data files are Snort alert logs, PF firewall logs, TCPdump, and so forth. Commercial parsers are also available to extract raw data files, such as Security Information Management or Enterprise Security Management.

After parsing or extracting raw data into information, some additional data might be included. Such additional data includes geographical locations of identified malware hosts or Domain Name System of malware hosts to enable understanding of raw data better and to help in reading graphs easier. Additional data includes role of machines to take raw data files or to extract those data files in outputting dataset of malware hosts detection. Possible roles of machines include finance, research and development, Web server, and so on to come up with identified malware hosts being associated with threat prevention programs. Parsing and adding additional data to dataset of malware hosts detection are used to prepare for meaningful information and for plotting graphs. Some additional data is time sensitive; for example, Internet Protocol (IP) address of identified malware host might resolve to a different IP address on the next day. Thus, dataset of malware hosts detection should have timestamp of data and time when malware hosts are first identified. Moreover, since identified malware hosts must use Domain Name System (DNS) lookup, such DNS lookup would likely alert attackers, especially if attackers operate on DNS server, thereby causing attackers to obfuscate their origin. If so, dataset of malware hosts detection might become outdated, such that an update is required to identify malware hosts in being associated with threat prevention programs.

Select relevant entries from identified malware hosts detection. By identifying objectives, as stated in step one, make implicit choices about which relevant entries are interesting. Visualizing all entries of malware hosts detection is not always necessary; sometimes, visualize a specific subset of entries to fulfill a certain criterion is sufficient. In other words, select information from malware hosts detection that can express interesting properties in answering a defined problem and a question. Be careful filtering too much information loses the meaning of that defined problem.

Aggregation involves summarization of data to simplify a large dataset, so that patterns of identified malware hosts detection can be seen in a graph. For example, when aggregating a graph that is based on threat prevention programs, display those malware hosts detection that use same threat prevention program the most. Obviously, there would probably be a handful of malware hosts detection; if so, choose the top of malware hosts detection with most threat prevention programs being attached to them. Another simpler example of aggregation is to identify the top five threat prevention programs that have most frequency to be used by malware hosts detection.

There are a couple of challenges that are associated with data processing. First, adding additional data to identified malware hosts detection, such as number of extra columns be needed for extra data, anomalies to be dealt with extra data, classification for extra data, etc. Second, semantics of fields in malware hosts detection can change accordingly. If every dataset conforms to a common standard of definition on what must be considered malware hosts detection, there would be needed to deal with these challenges.

Step 4: Visual Transformation

Output from the previous step three is contained into a large dataset that have all the information for visualization, such as identified malware hosts detection, count, and threat prevention programs. Parsed output should be available in a comma-separated format, .csv, file. File with .csv makes further processing be simpler to work with data. In step four, mapping data into some visual structure to produce a graphic representation.

The first decision to make is what data dimension is the primary one, which is a dimension to be analyzed or to compare distribution of data on other dimensions. If multiple data dimensions are important, type of chosen graph is selected to represent data. Various charts are indifferent of assigning preference to data dimension, such as link graphs, scatterplots, and parallel coordinates. For other graph types, decide which data dimension to use as the focus. Once decided on primary dimension, identify whether all data dimensions need to be explicitly displayed for graph attributes, such as size, color, and shape.

Step 5: View Transformation

After having a first visual representation of data on identified malware hosts detection from step four, if careful on what to display and how to display things, there would be a graph that is useful. Otherwise, if there is no one graph that is useful, restart the process from step three after gaining knowledge on what to correct for new graphs. First graph after regenerating graphs from step three provides some initial insight into entries within dataset. These entries provide context and help determine which graph components should be filtered out to highlight interesting parts of these entries. Be cautious when filtering entries in dataset so that critical data do not get removed accidentally.

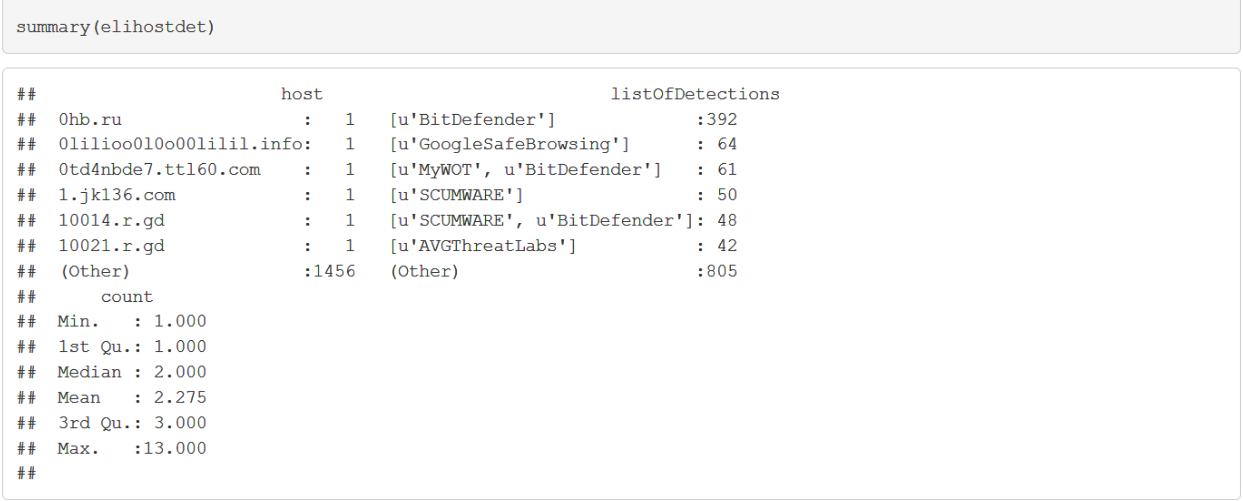
Step 6: Interpret and Decide

Step five results in multiple graphs to focus on those graphs for relevant data. Have at least one graph to be used in satisfying a defined problem to be answered. Be careful when looking at graphs. Verify things to make sure what is displayed on graphs are what is interpreted. A strong method that is useful when analyzing and interpreting graphs is a comparison against other graphs. These graphs can either be generated over same type of data but at a previous point in time or be a reference graph for that type of visualization. If a graph is expected to be display with a certain pattern, that same graph suddenly does not show up anymore pattern is when an anomaly is found. Assuming generated graphs are displaying what is expected from them, go back to problem definition to answer these questions:

- What were you looking for?
- Why did you generate the graph?

Based on a defined problem to be solved and a question to be answered, make an informed decision about how to solve that problem and to answer that question by generating graphs and analyzing those graphs.

Data Dictionary

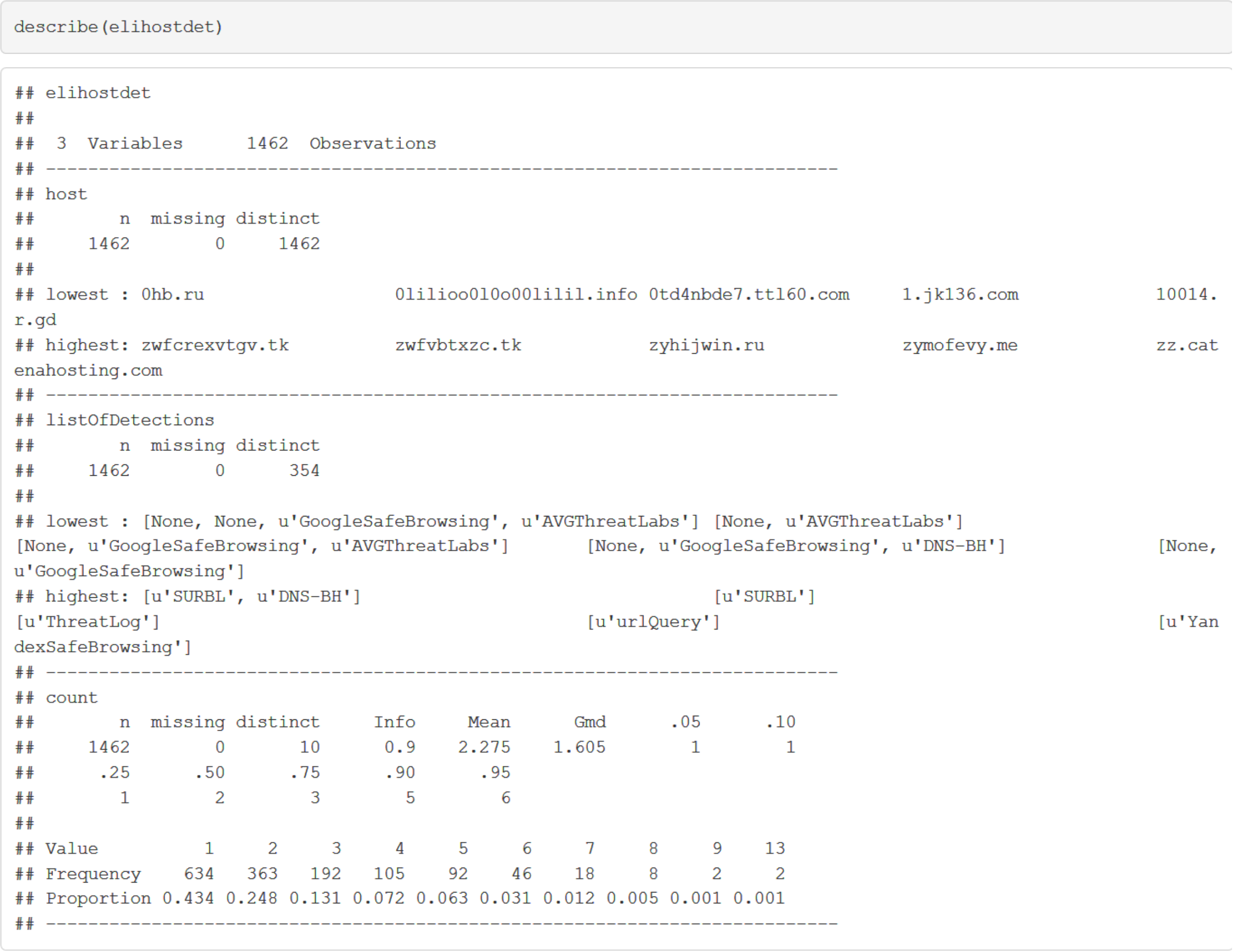


```
summary(elihostdet)
```

```
##                       host                             listOfDetections
##  0hb.ru                  :   1   [u'BitDefender']               :392
##  0lilioo0l0o00lilil.info:   1   [u'GoogleSafeBrowsing']        : 64
##  0td4nbde7.ttl60.com     :   1   [u'MyWOT', u'BitDefender']     : 61
##  1.jk136.com             :   1   [u'SCUMWARE']                  : 50
##  10014.r.gd              :   1   [u'SCUMWARE', u'BitDefender']: 48
##  10021.r.gd              :   1   [u'AVGThreatLabs']             : 42
##  (Other)                 :1456   (Other)                        :805
##      count
##  Min.   : 1.000
##  1st Qu.: 1.000
##  Median : 2.000
##  Mean   : 2.275
##  3rd Qu.: 3.000
##  Max.   :13.000
##
```

As shown by image above, there are ties for identified malware hosts detection with first six being 0hb.ru, 0lilioo0l0o00lilil.info, 0td4nbde7.ttl60.com, 1.jk136.com, 10014.r.gd, and

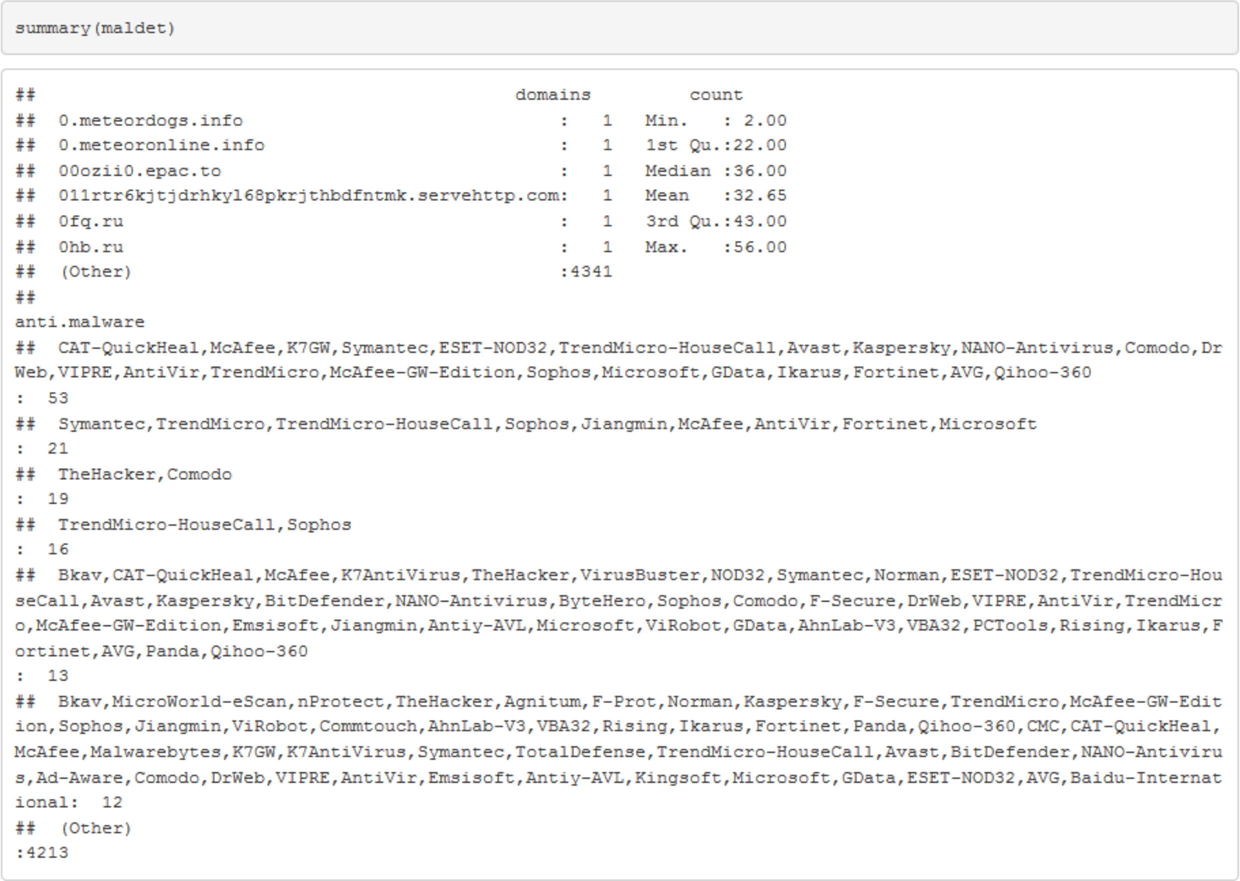
10021.r.gd, to have a count of one for each of malware hosts detection. In fact, all malware hosts detections have a count of one meaning each malware hosts detection has only one set of threat prevention programs. A total of 1462 observations of malware hosts detection has been found. Also, as shown by image at the bottom of the previous page, the top sets of threat prevention programs with the most counts that are used against malware hosts detection are BitDefender, GoogleSafeBrowsing, MyWOT, SCUMWARE, and AVGThreatLabs. For counts of threat prevention programs, malware hosts detection can have at most thirteen different threat prevention programs; on average, most malware hosts detections have two to three threat prevention programs in them.



```
describe(elihostdet)
```

```
## elihostdet
##
##  3  Variables      1462  Observations
## ---------------------------------------------------------------------------
## host
##        n  missing distinct
##     1462        0     1462
##
## lowest : 0hb.ru                  0lilioo010o00lilil.info 0td4nbde7.ttl60.com     1.jk136.com             10014.
r.gd
## highest: zwfcrexvtgv.tk          zwfvbtxzc.tk            zyhijwin.ru             zymofevy.me             zz.cat
enahosting.com
## ---------------------------------------------------------------------------
## listOfDetections
##        n  missing distinct
##     1462        0      354
##
## lowest : [None, None, u'GoogleSafeBrowsing', u'AVGThreatLabs'] [None, u'AVGThreatLabs']
[None, u'GoogleSafeBrowsing', u'AVGThreatLabs']       [None, u'GoogleSafeBrowsing', u'DNS-BH']              [None,
u'GoogleSafeBrowsing']
## highest: [u'SURBL', u'DNS-BH']                                 [u'SURBL']
[u'ThreatLog']                                        [u'urlQuery']                                         [u'Yan
dexSafeBrowsing']
## ---------------------------------------------------------------------------
## count
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     1462        0       10      0.9    2.275    1.605        1        1
##      .25      .50      .75      .90      .95
##        1        2        3        5        6
##
## Value          1     2     3     4     5     6     7     8     9    13
## Frequency    634   363   192   105    92    46    18     8     2     2
## Proportion 0.434 0.248 0.131 0.072 0.063 0.031 0.012 0.005 0.001 0.001
## ---------------------------------------------------------------------------
```

As shown by above image, dataset of identified malware hosts detection has 3 variables being three columns and 1462 observations being that many rows. Same as before, each observation or row has distinct malware hosts detection without any repeats; also, first five (lowest) and last five (highest) of malware hosts detection are shown. Similarly, there are 1462 observations for set of threat prevention programs but only 354 of which are distinctive without any repeats; also, first five (lowest) and last five (highest) set of threat prevention programs are shown. When dealing with real numbers, R statistical computing programming language shows several information, such as, Gini mean difference, Fisher information, quartiles, frequency and proportion of each real number, etc., as shown by above image.

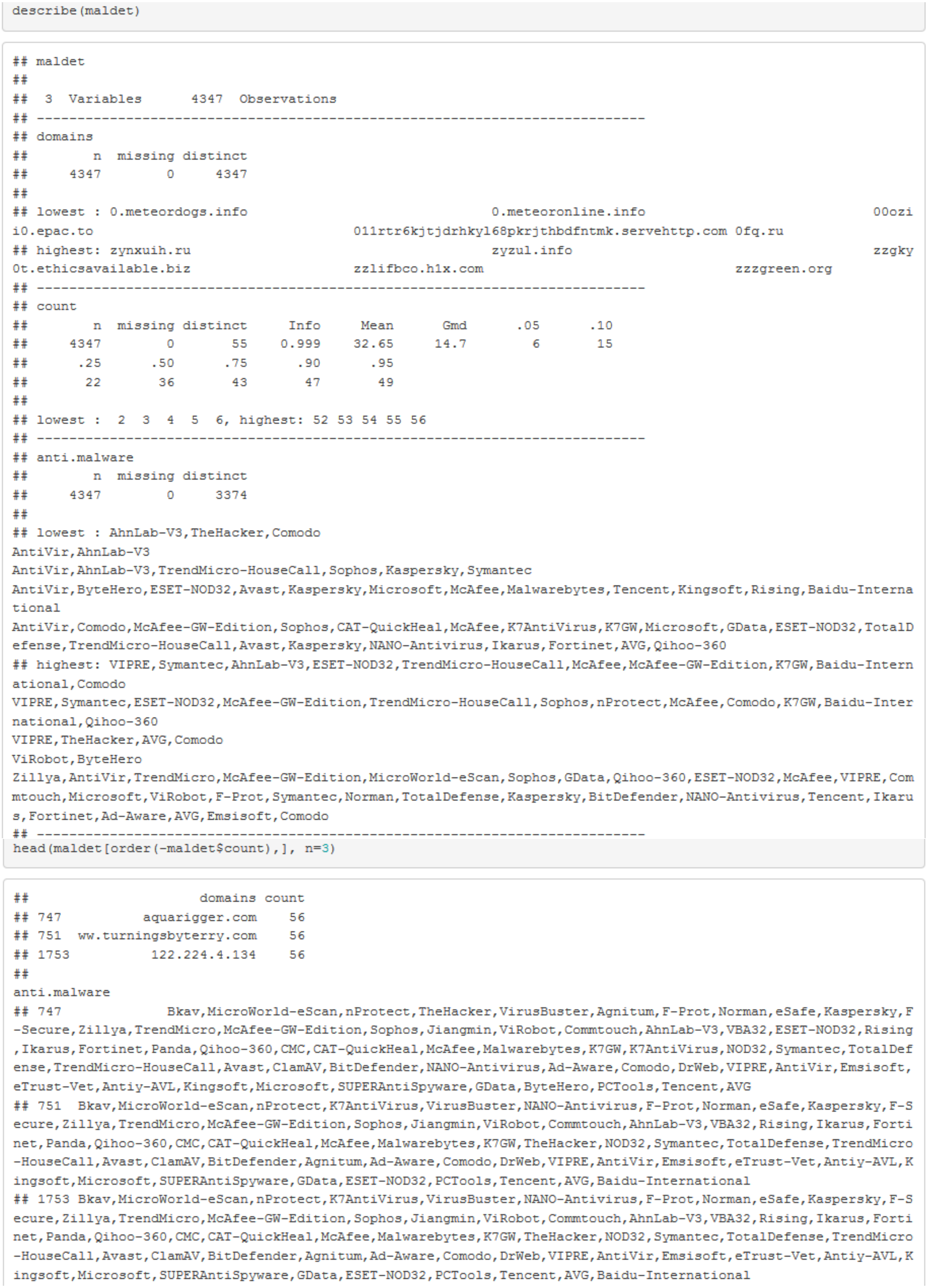


```
summary(maldet)
```

```
##                                                  domains         count
##  0.meteordogs.info                               :   1   Min.   : 2.00
##  0.meteoronline.info                             :   1   1st Qu.:22.00
##  00ozii0.epac.to                                 :   1   Median :36.00
##  011rtr6kjtjdrhkyl68pkrjthbdfntmk.servehttp.com:   1   Mean   :32.65
##  0fq.ru                                          :   1   3rd Qu.:43.00
##  0hb.ru                                          :   1   Max.   :56.00
##  (Other)                                         :4341
##
anti.malware
##  CAT-QuickHeal,McAfee,K7GW,Symantec,ESET-NOD32,TrendMicro-HouseCall,Avast,Kaspersky,NANO-Antivirus,Comodo,Dr
Web,VIPRE,AntiVir,TrendMicro,McAfee-GW-Edition,Sophos,Microsoft,GData,Ikarus,Fortinet,AVG,Qihoo-360
:  53
##  Symantec,TrendMicro,TrendMicro-HouseCall,Sophos,Jiangmin,McAfee,AntiVir,Fortinet,Microsoft
:  21
##  TheHacker,Comodo
:  19
##  TrendMicro-HouseCall,Sophos
:  16
##  Bkav,CAT-QuickHeal,McAfee,K7AntiVirus,TheHacker,VirusBuster,NOD32,Symantec,Norman,ESET-NOD32,TrendMicro-Hou
seCall,Avast,Kaspersky,BitDefender,NANO-Antivirus,ByteHero,Sophos,Comodo,F-Secure,DrWeb,VIPRE,AntiVir,TrendMicr
o,McAfee-GW-Edition,Emsisoft,Jiangmin,Antiy-AVL,Microsoft,ViRobot,GData,AhnLab-V3,VBA32,PCTools,Rising,Ikarus,F
ortinet,AVG,Panda,Qihoo-360
:  13
##  Bkav,MicroWorld-eScan,nProtect,TheHacker,Agnitum,F-Prot,Norman,Kaspersky,F-Secure,TrendMicro,McAfee-GW-Edit
ion,Sophos,Jiangmin,ViRobot,Commtouch,AhnLab-V3,VBA32,Rising,Ikarus,Fortinet,Panda,Qihoo-360,CMC,CAT-QuickHeal,
McAfee,Malwarebytes,K7GW,K7AntiVirus,Symantec,TotalDefense,TrendMicro-HouseCall,Avast,BitDefender,NANO-Antiviru
s,Ad-Aware,Comodo,DrWeb,VIPRE,AntiVir,Emsisoft,Antiy-AVL,Kingsoft,Microsoft,GData,ESET-NOD32,AVG,Baidu-Internat
ional:  12
##  (Other)
:4213
```

As shown by image above, there are same number of identified malware domains detection with first six being 0.meteordogs.info, 0.meteoronline.info, 00ozii0.epac.to, 011rtr6kjtjdrhkyl68pkrjthbdfntmk.servehttp.com, 0fq.ru, and 0hb.ru, to have a count of one for each of malware domains detection. In fact, all malware domains detections have a count of one meaning each malware domains detection has only one set of anti-malware programs. A total of 4347 observations of malware domains detection has been found. Also, as shown by image above, the top sets of anti-malware programs are shown, which would need further filtering. For counts of anti-malware programs, malware domains detection can have at most fifty-six different ant-malware programs; on average, most malware domains detections have 32 to 33 anti-malware programs in them.

Image at the top of the next page shows dataset of identified malware domains detection has 3 variables being three columns and 4347 observations being that many rows. Same as before, each observation or row has distinct malware domains detection without any repeats; also, first five (lowest) and last five (highest) of malware domains detection are shown. Similarly, there are 4347 observations for set of anti-malware programs but only 3374 sets are unique; also, first five (lowest) and last five (highest) set of anti-malware programs are shown. When dealing with real numbers, R shows several information, such as, Gini mean difference, Fisher information, quartiles, etc., as shown by image at the top of the next page. Furthermore, image at the bottom of the next page shows first three malware domains detection with most count being fifty-six different anti-malware programs for each malware domains detection, which are 122.224.4.134, aquarigger.com, and www.turningsbyterry.com.



```
describe(maldet)
```

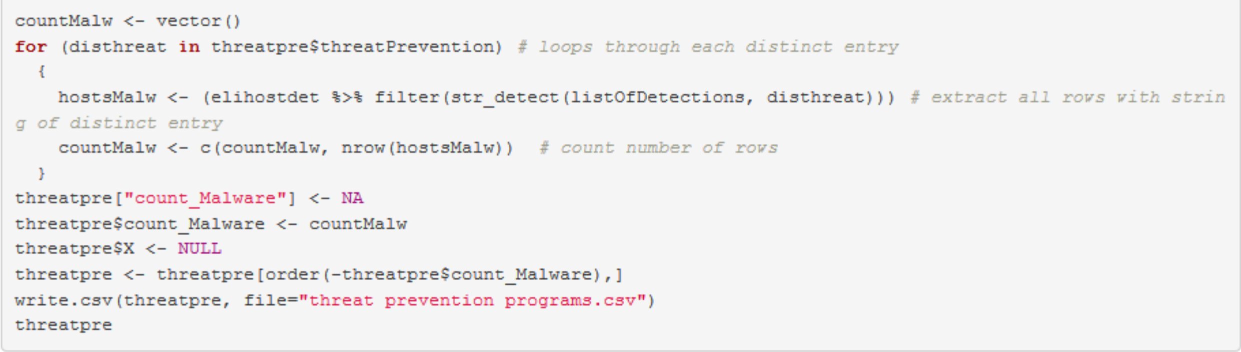
```
## maldet
##
##  3  Variables      4347  Observations
## ---------------------------------------------------------------------------
## domains
##        n  missing distinct
##     4347        0     4347
##
## lowest : 0.meteordogs.info                                0.meteoronline.info                             00ozi
i0.epac.to                              011rtr6kjtjdrhkyl68pkrjthbdfntmk.servehttp.com 0fq.ru
## highest: zynxuih.ru                                       zyzul.info                                      zzgky
0t.ethicsavailable.biz                  zzlifbco.h1x.com                               zzzgreen.org
## ---------------------------------------------------------------------------
## count
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##     4347        0       55    0.999    32.65     14.7        6       15
##      .25      .50      .75      .90      .95
##       22       36       43       47       49
##
## lowest :  2  3  4  5  6, highest: 52 53 54 55 56
## ---------------------------------------------------------------------------
## anti.malware
##        n  missing distinct
##     4347        0     3374
##
## lowest : AhnLab-V3,TheHacker,Comodo
AntiVir,AhnLab-V3
AntiVir,AhnLab-V3,TrendMicro-HouseCall,Sophos,Kaspersky,Symantec
AntiVir,ByteHero,ESET-NOD32,Avast,Kaspersky,Microsoft,McAfee,Malwarebytes,Tencent,Kingsoft,Rising,Baidu-Interna
tional
AntiVir,Comodo,McAfee-GW-Edition,Sophos,CAT-QuickHeal,McAfee,K7AntiVirus,K7GW,Microsoft,GData,ESET-NOD32,TotalD
efense,TrendMicro-HouseCall,Avast,Kaspersky,NANO-Antivirus,Ikarus,Fortinet,AVG,Qihoo-360
## highest: VIPRE,Symantec,AhnLab-V3,ESET-NOD32,TrendMicro-HouseCall,McAfee,McAfee-GW-Edition,K7GW,Baidu-Intern
ational,Comodo
VIPRE,Symantec,ESET-NOD32,McAfee-GW-Edition,TrendMicro-HouseCall,Sophos,nProtect,McAfee,Comodo,K7GW,Baidu-Inter
national,Qihoo-360
VIPRE,TheHacker,AVG,Comodo
ViRobot,ByteHero
Zillya,AntiVir,TrendMicro,McAfee-GW-Edition,MicroWorld-eScan,Sophos,GData,Qihoo-360,ESET-NOD32,McAfee,VIPRE,Com
mtouch,Microsoft,ViRobot,F-Prot,Symantec,Norman,TotalDefense,Kaspersky,BitDefender,NANO-Antivirus,Tencent,Ikaru
s,Fortinet,Ad-Aware,AVG,Emsisoft,Comodo
## ---------------------------------------------------------------------------
```

```
head(maldet[order(-maldet$count),], n=3)
```

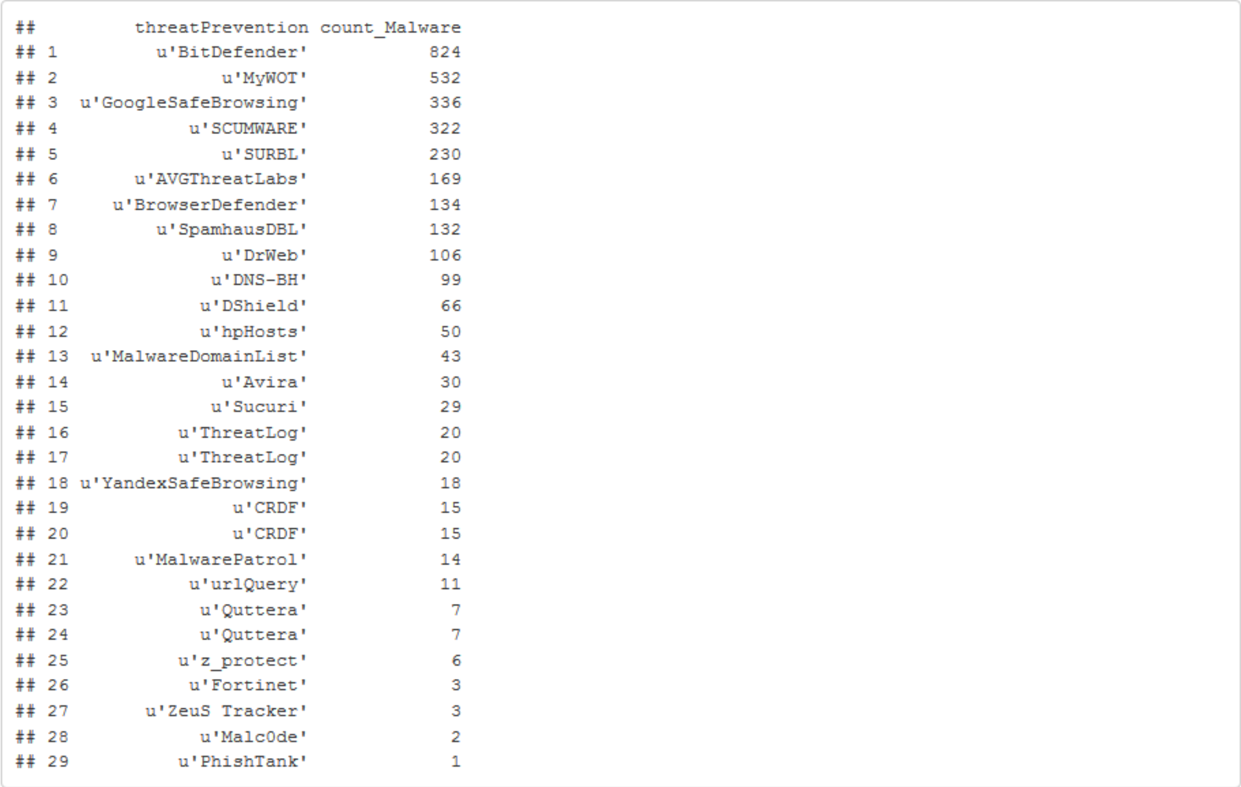
```
##                     domains count
## 747          aquarigger.com    56
## 751  ww.turningsbyterry.com    56
## 1753           122.224.4.134    56
##
anti.malware
## 747              Bkav,MicroWorld-eScan,nProtect,TheHacker,VirusBuster,Agnitum,F-Prot,Norman,eSafe,Kaspersky,F
-Secure,Zillya,TrendMicro,McAfee-GW-Edition,Sophos,Jiangmin,ViRobot,Commtouch,AhnLab-V3,VBA32,ESET-NOD32,Rising
,Ikarus,Fortinet,Panda,Qihoo-360,CMC,CAT-QuickHeal,McAfee,Malwarebytes,K7GW,K7AntiVirus,NOD32,Symantec,TotalDef
ense,TrendMicro-HouseCall,Avast,ClamAV,BitDefender,NANO-Antivirus,Ad-Aware,Comodo,DrWeb,VIPRE,AntiVir,Emsisoft,
eTrust-Vet,Antiy-AVL,Kingsoft,Microsoft,SUPERAntiSpyware,GData,ByteHero,PCTools,Tencent,AVG
## 751  Bkav,MicroWorld-eScan,nProtect,K7AntiVirus,VirusBuster,NANO-Antivirus,F-Prot,Norman,eSafe,Kaspersky,F-S
ecure,Zillya,TrendMicro,McAfee-GW-Edition,Sophos,Jiangmin,ViRobot,Commtouch,AhnLab-V3,VBA32,Rising,Ikarus,Forti
net,Panda,Qihoo-360,CMC,CAT-QuickHeal,McAfee,Malwarebytes,K7GW,TheHacker,NOD32,Symantec,TotalDefense,TrendMicro
-HouseCall,Avast,ClamAV,BitDefender,Agnitum,Ad-Aware,Comodo,DrWeb,VIPRE,AntiVir,Emsisoft,eTrust-Vet,Antiy-AVL,K
ingsoft,Microsoft,SUPERAntiSpyware,GData,ESET-NOD32,PCTools,Tencent,AVG,Baidu-International
## 1753 Bkav,MicroWorld-eScan,nProtect,K7AntiVirus,VirusBuster,NANO-Antivirus,F-Prot,Norman,eSafe,Kaspersky,F-S
ecure,Zillya,TrendMicro,McAfee-GW-Edition,Sophos,Jiangmin,ViRobot,Commtouch,AhnLab-V3,VBA32,Rising,Ikarus,Forti
net,Panda,Qihoo-360,CMC,CAT-QuickHeal,McAfee,Malwarebytes,K7GW,TheHacker,NOD32,Symantec,TotalDefense,TrendMicro
-HouseCall,Avast,ClamAV,BitDefender,Agnitum,Ad-Aware,Comodo,DrWeb,VIPRE,AntiVir,Emsisoft,eTrust-Vet,Antiy-AVL,K
ingsoft,Microsoft,SUPERAntiSpyware,GData,ESET-NOD32,PCTools,Tencent,AVG,Baidu-International
```

Results and Observations

Malware Hosts Detection



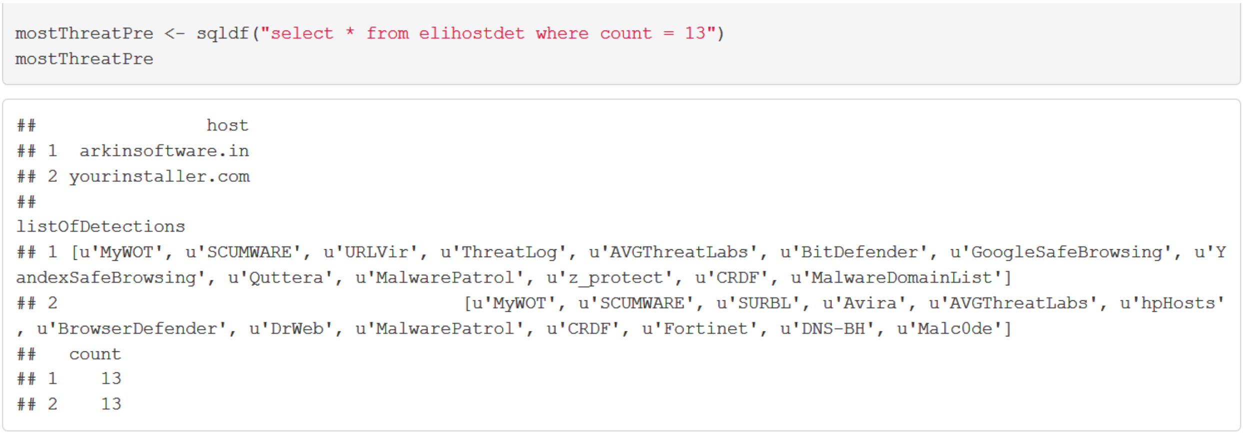
```
countMalw <- vector()
for (disthreat in threatpre$threatPrevention) # loops through each distinct entry
  {
    hostsMalw <- (elihostdet %>% filter(str_detect(listOfDetections, disthreat))) # extract all rows with strin
g of distinct entry
    countMalw <- c(countMalw, nrow(hostsMalw))  # count number of rows
  }
threatpre["count_Malware"] <- NA
threatpre$count_Malware <- countMalw
threatpre$X <- NULL
threatpre <- threatpre[order(-threatpre$count_Malware),]
write.csv(threatpre, file="threat prevention programs.csv")
threatpre
```



```
##          threatPrevention count_Malware
## 1         u'BitDefender'           824
## 2               u'MyWOT'           532
## 3  u'GoogleSafeBrowsing'           336
## 4            u'SCUMWARE'           322
## 5               u'SURBL'           230
## 6         u'AVGThreatLabs'           169
## 7       u'BrowserDefender'         134
## 8          u'SpamhausDBL'          132
## 9               u'DrWeb'           106
## 10             u'DNS-BH'            99
## 11            u'DShield'            66
## 12            u'hpHosts'            50
## 13  u'MalwareDomainList'            43
## 14              u'Avira'            30
## 15             u'Sucuri'            29
## 16          u'ThreatLog'            20
## 17          u'ThreatLog'            20
## 18 u'YandexSafeBrowsing'            18
## 19               u'CRDF'            15
## 20               u'CRDF'            15
## 21      u'MalwarePatrol'            14
## 22           u'urlQuery'            11
## 23            u'Quttera'             7
## 24            u'Quttera'             7
## 25          u'z_protect'             6
## 26           u'Fortinet'             3
## 27       u'ZeuS Tracker'             3
## 28            u'Malc0de'             2
## 29          u'PhishTank'             1
```

Dataset of identified malware hosts detection contains set of threat prevention programs for each malware hosts detection. Image above shows a table of extracted threat prevention programs from dataset of malware hosts detection. As shown by above image, the top five threat prevention programs that are used the most against malware hosts detection are BitDefender with 824, MyWOT with 532, GoogleSafeBrowsing with 336, SCUMWARE with 322, and SURBL with 230 against malware hosts detections.

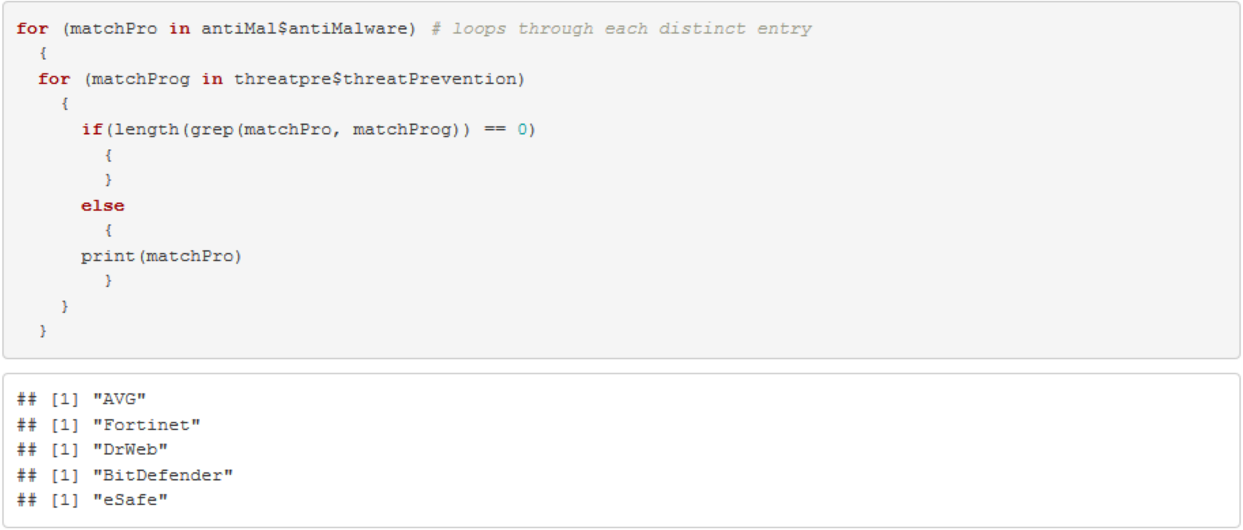
Image at the top of the next page shows two identified malware hosts detections with most number of threat prevention programs. These two malware hosts detections are arkinsoftware.in and yourinstaller.com having a total of 13 threat prevention programs with each of them. Having that many threat prevention programs, these two malware hosts detections are identified as malign hosts. Together, these two malware hosts detections have a total of 21 distinctive threat prevention programs, as shown image at the top of the next page.



```
mostThreatPre <- sqldf("select * from elihostdet where count = 13")
mostThreatPre
```

```
##                 host
## 1  arkinsoftware.in
## 2 yourinstaller.com
##
listOfDetections
## 1 [u'MyWOT', u'SCUMWARE', u'URLVir', u'ThreatLog', u'AVGThreatLabs', u'BitDefender', u'GoogleSafeBrowsing', u'Y
andexSafeBrowsing', u'Quttera', u'MalwarePatrol', u'z_protect', u'CRDF', u'MalwareDomainList']
## 2                                   [u'MyWOT', u'SCUMWARE', u'SURBL', u'Avira', u'AVGThreatLabs', u'hpHosts'
, u'BrowserDefender', u'DrWeb', u'MalwarePatrol', u'CRDF', u'Fortinet', u'DNS-BH', u'Malc0de']
##   count
## 1    13
## 2    13
```

Malware Domains Detection

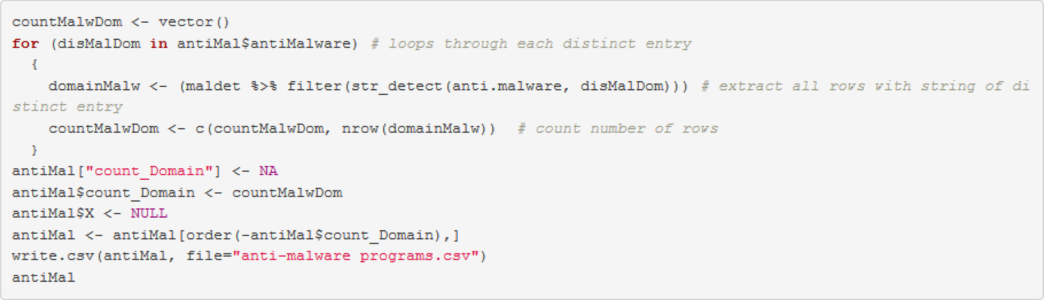


```
for (matchPro in antiMal$antiMalware) # loops through each distinct entry
  {
  for (matchProg in threatpre$threatPrevention)
    {
      if(length(grep(matchPro, matchProg)) == 0)
        {
        }
      else
        {
      print(matchPro)
        }
    }
  }
```

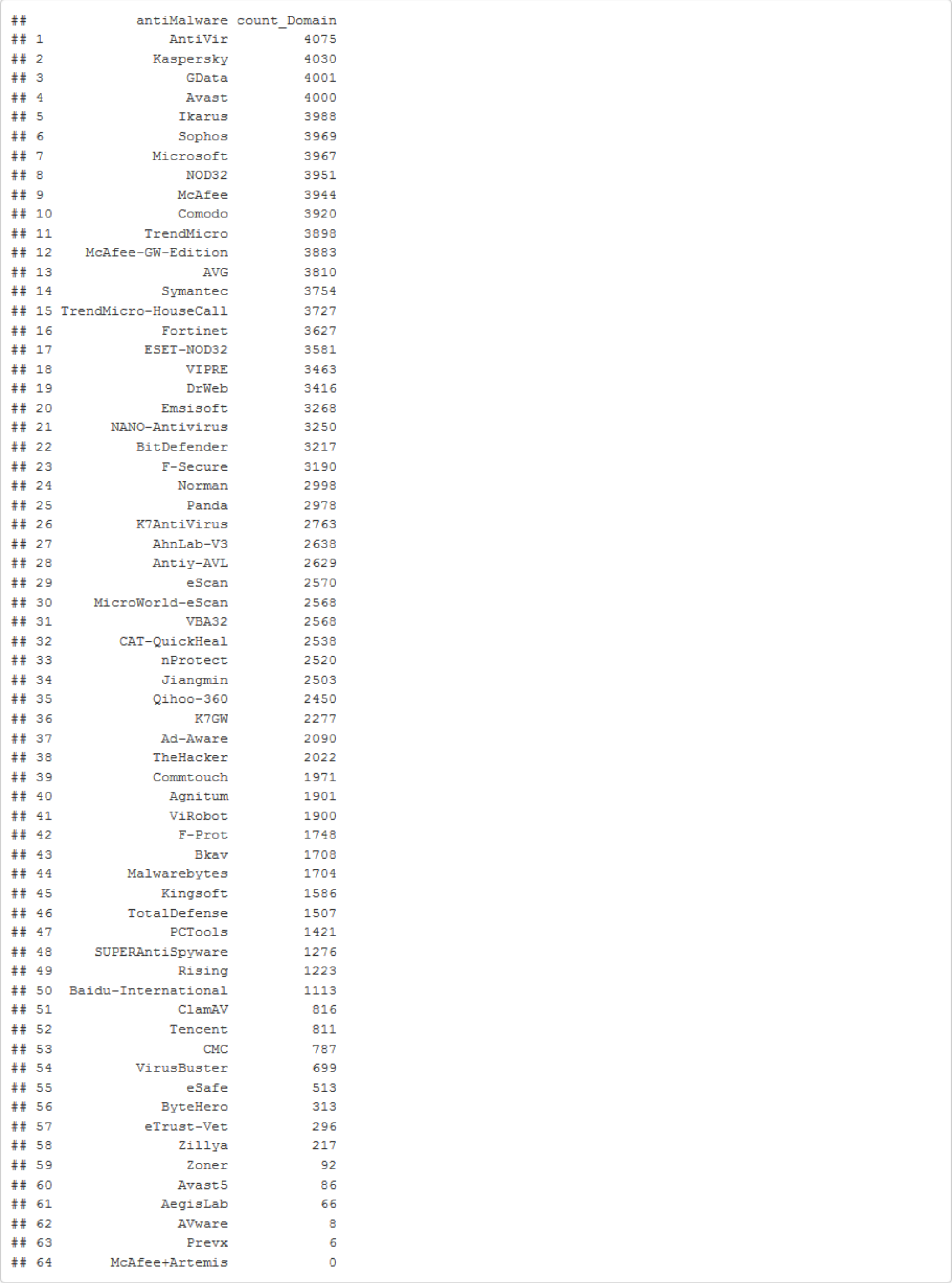
```
## [1] "AVG"
## [1] "Fortinet"
## [1] "DrWeb"
## [1] "BitDefender"
## [1] "eSafe"
```

Image above shows a comparison of string of anti-malware programs in identified malware domains detection against substring of threat prevention programs in identified malware hosts detection in a double for loop. Even though there are five results (i.e., AVG, Fortinet, DrWeb, BitDefender, and eSafe) to have matchings of anti-malware programs and threat prevention programs, only Fortinet, DrWeb, and BitDefender are anti-malware programs or threat prevention programs that are useful against malware domains detection and against malware hosts detection. Anti-malware programs of AVG and eSafe in malware domains detection show up as AVGThreatLabs and GoogleSafeBrowsing threat prevention programs in malware hosts detection, which are not exact strings.

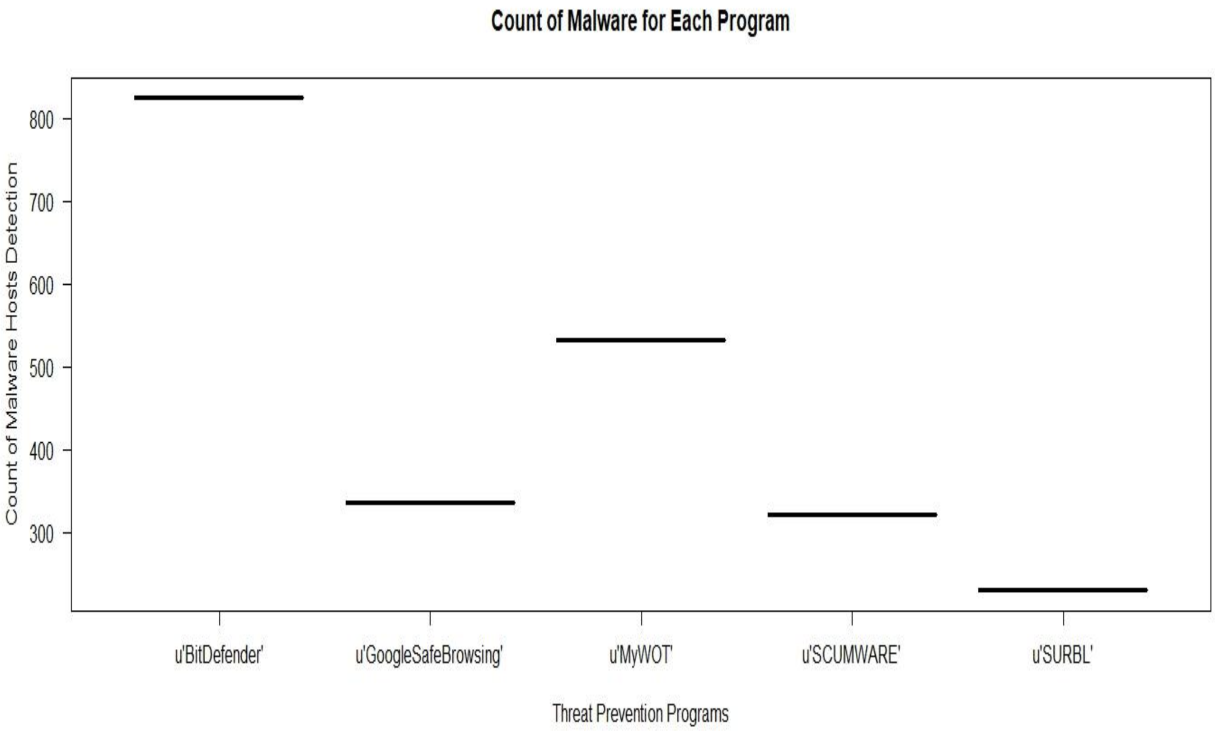
Image at the next page shows a table of count of anti-malware programs being used for all identified malware domains detections. As shown by that image, the top five anti-malware programs are AntiVir with 4075, Kaspersky with 4030, GData with 4001, Avast with 4000, and Ikarus with 3988 against malware domains detections. An anomaly is shown for anti-malware program of McAfee+Artemis with zero count against malware domains detection is probably given as an outlier; that is, McAfee+Artemis anti-malware program has no malware domains detection to begin with in dataset.



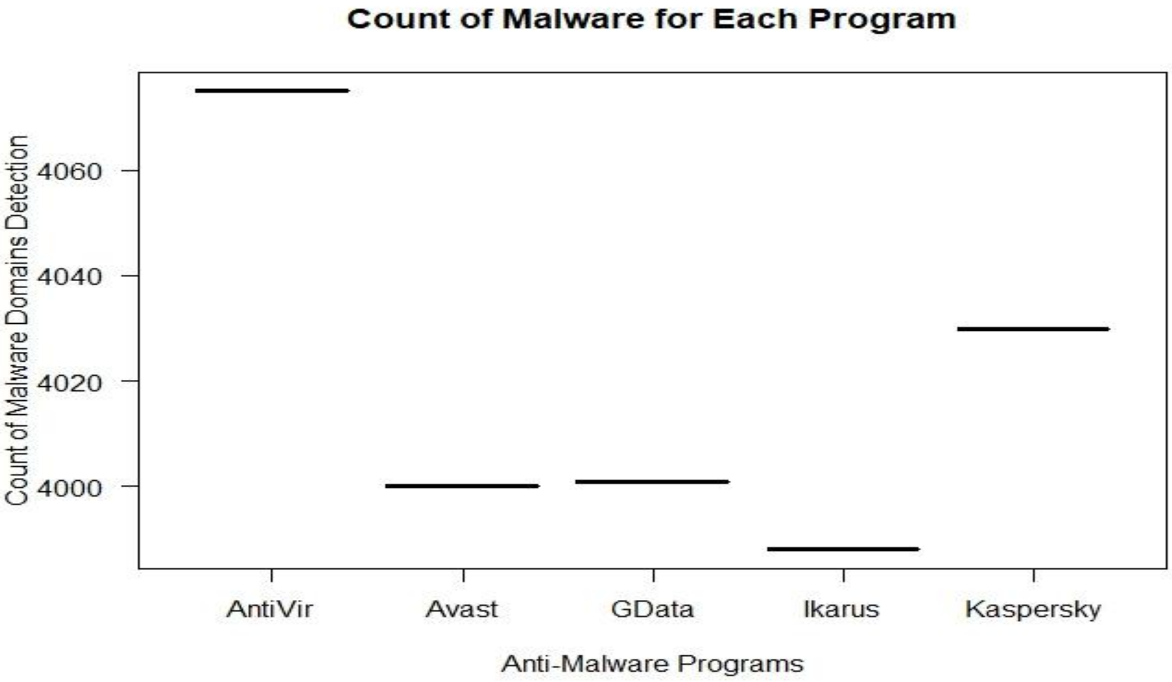
```
countMalwDom <- vector()
for (disMalDom in antiMal$antiMalware) # loops through each distinct entry
  {
    domainMalw <- (maldet %>% filter(str_detect(anti.malware, disMalDom))) # extract all rows with string of di
stinct entry
    countMalwDom <- c(countMalwDom, nrow(domainMalw))  # count number of rows
  }
antiMal["count_Domain"] <- NA
antiMal$count_Domain <- countMalwDom
antiMal$X <- NULL
antiMal <- antiMal[order(-antiMal$count_Domain),]
write.csv(antiMal, file="anti-malware programs.csv")
antiMal
```



```
##             antiMalware count_Domain
## 1               AntiVir         4075
## 2             Kaspersky         4030
## 3                 GData         4001
## 4                 Avast         4000
## 5                Ikarus         3988
## 6                Sophos         3969
## 7             Microsoft         3967
## 8                 NOD32         3951
## 9                McAfee         3944
## 10               Comodo         3920
## 11           TrendMicro         3898
## 12    McAfee-GW-Edition         3883
## 13                  AVG         3810
## 14             Symantec         3754
## 15 TrendMicro-HouseCall         3727
## 16             Fortinet         3627
## 17           ESET-NOD32         3581
## 18                VIPRE         3463
## 19                DrWeb         3416
## 20             Emsisoft         3268
## 21       NANO-Antivirus         3250
## 22          BitDefender         3217
## 23             F-Secure         3190
## 24               Norman         2998
## 25                Panda         2978
## 26          K7AntiVirus         2763
## 27            AhnLab-V3         2638
## 28            Antiy-AVL         2629
## 29                eScan         2570
## 30     MicroWorld-eScan         2568
## 31                VBA32         2568
## 32        CAT-QuickHeal         2538
## 33             nProtect         2520
## 34             Jiangmin         2503
## 35            Qihoo-360         2450
## 36                 K7GW         2277
## 37             Ad-Aware         2090
## 38            TheHacker         2022
## 39            Commtouch         1971
## 40              Agnitum         1901
## 41              ViRobot         1900
## 42               F-Prot         1748
## 43                 Bkav         1708
## 44         Malwarebytes         1704
## 45             Kingsoft         1586
## 46         TotalDefense         1507
## 47              PCTools         1421
## 48     SUPERAntiSpyware         1276
## 49               Rising         1223
## 50  Baidu-International         1113
## 51               ClamAV          816
## 52              Tencent          811
## 53                  CMC          787
## 54          VirusBuster          699
## 55                eSafe          513
## 56             ByteHero          313
## 57           eTrust-Vet          296
## 58               Zillya          217
## 59                Zoner           92
## 60               Avast5           86
## 61             AegisLab           66
## 62               AVware            8
## 63                Prevx            6
## 64       McAfee+Artemis            0
```



Above chart plots the top five threat prevention programs against number of identified malware hosts detections.



Above chart plots the top five anti-malware programs against number of identified malware domains detections.

Conclusions

In conclusion, two datasets of Comma-Separated Values (.csv) are compared. One dataset of .csv pertains a database of identified malware hosts detection with a count and with a set of threat prevention programs for each malware hosts detection. Another dataset of .csv pertains a database of identified malware domains detection with a count and with a set of anti-malware programs for each malware domains detection. For malware hosts detection, the top malware hosts detection with the most number of threat prevention programs at 13 are arkinsoftware.in and yourinstaller.com. The top five threat prevention programs that are used by most malware hosts detections are 824 instances of BitDefender, 532 instances of MyWOT, 336 instances of GoogleSafeBrowsing, 322 instances of SCUMWARE, and 230 instances of SURBL against malware hosts detections. As for malware domains detection, the top malware domains detection with the most number of anti-malware programs at 56 are 122.224.4.134, aquarigger.com, and ww.turningsbyterry.com. The top five anti-malware programs that are used by most malware domains detections are 4075 instances of AntiVir, 4030 instances of Kaspersky, 4001 instances of GData, 4000 instances of Avast, and 3988 instances of Ikarus against malware domains detections. Between those two datasets of malware hosts detection and of malware domains detection, only Fortinet, DrWeb, and BitDefender show up as threat prevention programs and as anti-malware programs.

To see full code of all plots and charts of this assignment two Part 1: From Data to Graphs, visit the following URL repository on GitHub:

https://github.com/kleung52/SRT411-Assignment-Two