DDS Honeypots Data Analysis

Ayşe Simge ÖZGER - Cyber Security Engineer

Emre ÖVÜNÇ - Cyber Security Engineer

Umut BAŞARAN - Software Engineer

05.06.2017

Content

1. Introduction

1.1. What is Honeypot?

Honeypot is a set of activities aimed at gathering information from the cyber criminals of the internet world. By imitating a real system, it collects information from infiltrators and uses this information to close the vulnerabilities in the existing system. It both plays a very important role in analyzing their aggressive behavior and collecting other important data. It collects important information such as malicious IP addresses, network protocols and port information. Honeypots are also very difficult to detect.

Some features of Honeypots;

- It is both reliable and fast because it integrates with real systems.
- Cost is low because it is a completely virtual system.
- They are compatible with almost all operating systems.
- They support all network protocols.
- They can intercept or route network traffic at any time.

1.2. Background of Data & Attributes

Data is a CSV file from a collection of AWS honeypot with both long int and string IPv4 adresses and full geolocation information. It has 19 attributes and there are over 450.000 observations. The most important attributes in our data are date, hour, day, month, daytime, location informations, host and destination & source ports. Generally, types of attributes are categorical and nominal.

Categorical	
Nominal	**Ordinal**
datetime	hour
host	latitude
src	longtitude
proto	daytime
type	month
spt	day
dpt	
srcstr	
cc	
country	
locale	
localeabbr	
postalcode	

Name	Description
datetime	Packet Arrival Date (YYYY-MM-DD)
hour	Packet Arrival Hour (HH:MM:SS)
host	Honeypot Server
src	Packet Source
proto	Packet Protocol Type
type	Packet Type
spt	Source Port
dpt	Destination Port
srcstr	Source IP Address
cc	Source Country Code
country	Source Country
locale	Source Location
localeabbr	Locale Abbr.
postalcode	Postal Code
latitude	Source Latitude
longtitude	Source Longitude
daytime	Period of Arrival Hour
month	Packet Arrival Month (1-12)
day	Packet Arrival Day (1-31)

1.3. What is our Aim?

Due to the increasing number of cyber attacks nowadays, honeypots have become increasingly important. [1] (Passeri,2017) It is thought that cyber warfare could lead to the end of the world and even lead to World War III. Data collected from Honeypots can be used to investigate aggressive profiles and for threat intelligence purposes. In this way, real systems can be better protected against attackers. It contributes to the preparation of a better defense by giving important intelligence information to the most attacking countries and the protocols they use.

We have also examined the data gathered from the honeypots and reached the details of the aggressive information and behavior. [2] (Jacobs, DDS Dataset Collection, 2014) Thanks to this project we have emphasized the significance of the project (network security tool) as a senior project. Given that there are no trained people in our country, we have received critical information, such as the cyber terrorist attacks on the countries that are threatening our country and the attack rates of services that are open to the whole world, thinking that more research should be done on data collected from honeypots. As a result of this research, the data obtained were also shared with the institutions required to prepare a better infrastructure. We have identified areas that need better protection by knowing in advance, such as methods used by attackers.

1.4. Methodology

Subjects

We look for data about security threats, network attacks to analyze situation of cyber threats and identify some characteristics. While we search for our dataset we used mainly internet research. We find some honeypot project reports and choose which is exactly has attributes that we searched.

Data Collection

Our dataset is in form of CSV. After, we added and modified some attributes according to our analysis aim. We converted format of hour (HH:MM:SS) to hour (HH) attribute. Then we grouped hour attribute and create categorical daytime attribute (daytime has night, morning, afternoon, evening according to hours). Also, we classify day and month attributes. To make some tests, we converted protocols (TCP, UDP and ICMP) to numbers (1,2 and 3).

Data Analysis

While we choose our data analysis methods, firstly we looked at dataset's measurements (nominal & ordinal, interval & ratio, scale …) and normality of distribution of our attributes (to choose which parametric tests that we use). Also, we had to choose methods according to our hypothesis. We will perform quantitative research which is objective, generalizable and tested, has dependent and independent variables. That's why we will perform statistical analysis.

Level of confidence determines how we can sure actual results of population. We will use Frequency table to analyze categorical variables (daytime). From this table, we obtain proportions and percentages to proof our hypothesis. In our data set, it does not contain numerical variables, that is why, we do not use test techniques which use descriptives (mean, median, mode... etc.).

Due to categorical variables, we have chosen hypotheses with possibilities and probabilities. To test these hypotheses, we use cross-tables, frequencies tables and to test whether our hypothesis true or not we performed chi-square tests. Also, to get some informations from data, we use pie charts, line graphs and bar charts.

1.5. Research Questions & Hypotheses

Research Questions

1) Is there a relationship between Time of Attack and Used Protocol Type?
2) Is there a relationship between Time of Month and Target Host?
3) Is there a relationship between Time of Target Host and Used Protocol Type?
4) Is there a relationship between Time of Month and Time of Attack?
5) Is there a relationship between Time of Day and Used Protocol Type?
6) Is there a relationship between Time of Time of Attack and Attacker Country?
7) Is there a relationship between Time of Used Protocol Type and Attacker Country?
8) What is the probability between countries to occur?

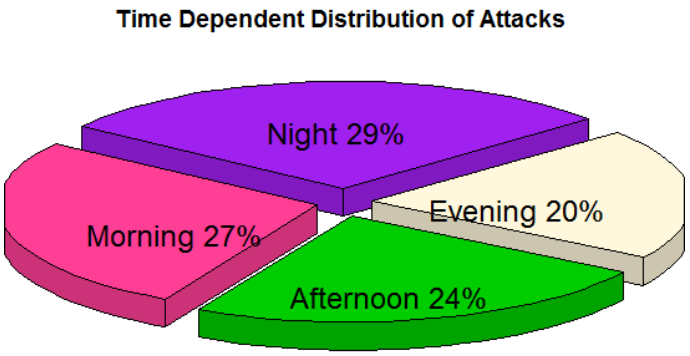
Hypotheses

1) There is a relationship between Time of Attack and Used Protocol Type.
2) There is a relationship between Month and Target Host.
3) There is a relationship between Target Host and Used Protocol Type.
4) There is a relationship between Month and Time of Attack.
5) There is a relationship between Day and Used Protocol Type.
6) There is a relationship between Time of Attack and Attacker Country.
7) There is a relationship between Used Protocol Type and Attacker Country.
8) Countries have different probability to occur.

2. Analysis

2.1. Overview of Data

a) Time Dependent Distribution of Attacks

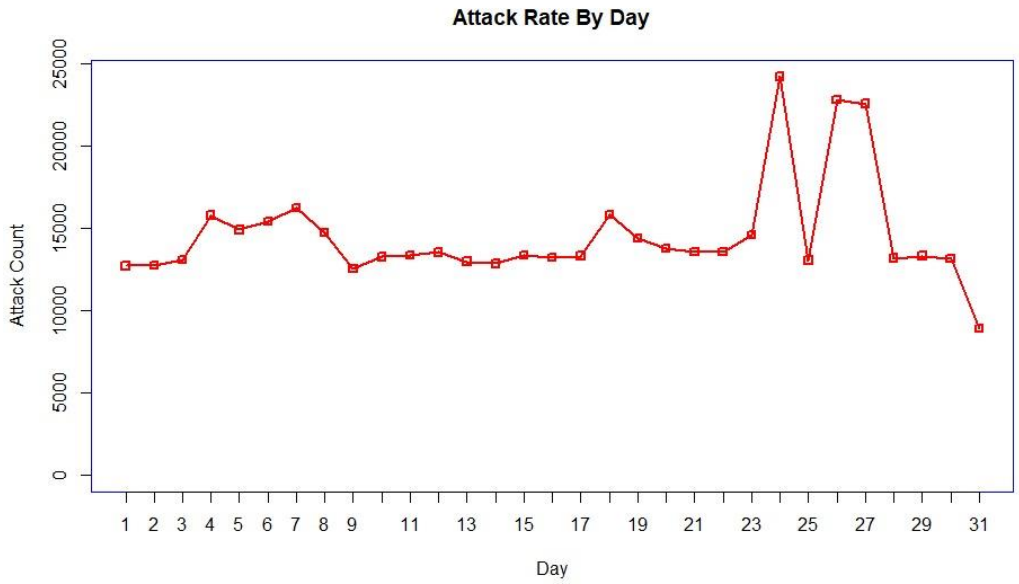


Time	Freq
Night	131332
Morning	121301
Afternoon	108517
Evening	90430

```
    Night   Morning Afternoon   Evening
0.2908278 0.2686146 0.2403052 0.2002524
```

According to Pie Chart and Frequency Table we can obviously see that the most of attacks occur at night. We think the reason of this situation is people at work, school or busy except nights.

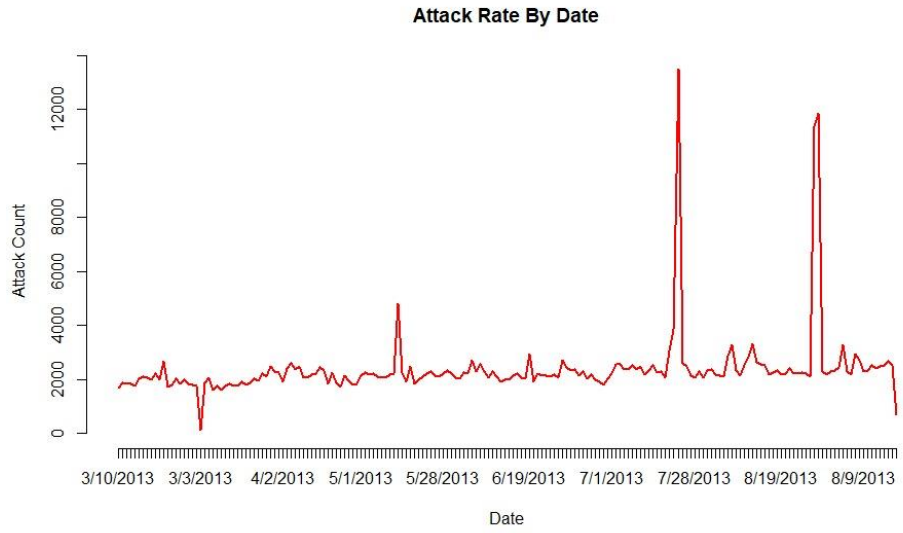
b) Attack Rate by Day



In this line graph, we learn that the days at the end of the month attack rates increasing.

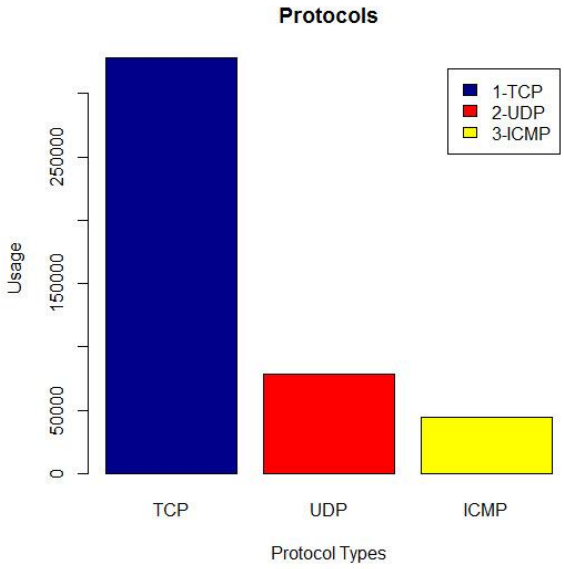
The reason of this at the end of the month people complete their jobs and attacker can be online more time.

c) Attack Rate by Date



According to this graph, we can see that some time periods attack rates increases or decreases visibly.

d) Attack Rate by Protocols

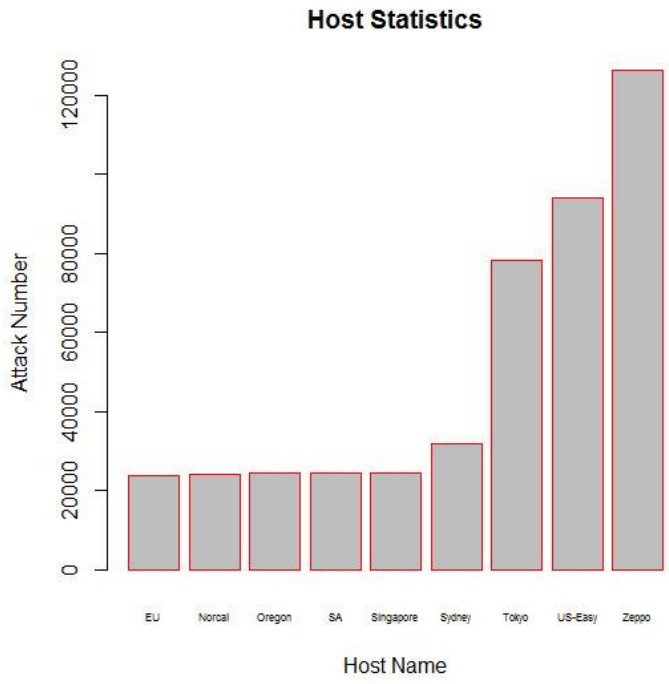


Proto	Freq
TCP	327991
UDP	78779
ICMP	44811

TCP	UDP	ICMP
0.72631709	0.17445154	0.09923137

In this bar chart and frequency table, we can see that especially TCP protocol used for attacks. According to this information precautions can be taken.

e) Attack Rate by Host Names

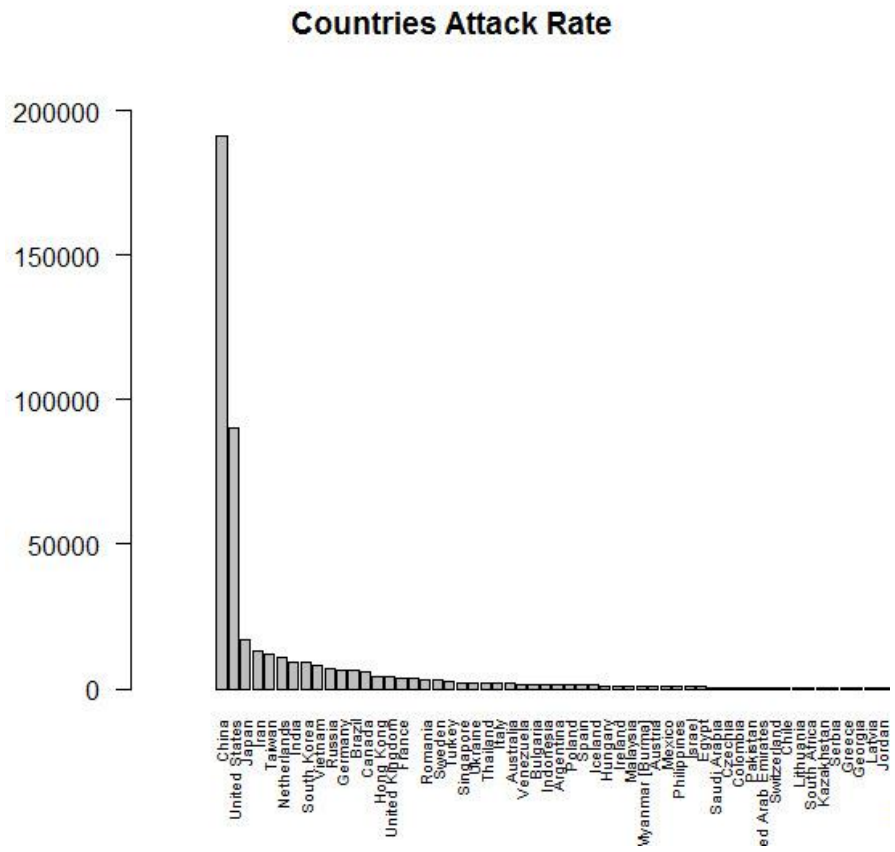


Host	Freq
Norcal	24566
Oregon	94076
SA	24316
Singapore	78151
Sydney	24456
Tokyo	126189
US-Easy	31779
Zeppo	24094

```
      Zeppo          EU       Norcal       Oregon           SA    Singapore       Sydney       Tokyo     US-Easy
5.335465e-02 5.304463e-02 5.439987e-02 2.083254e-01 5.384626e-02 1.730605e-01 5.415628e-02 2.794376e-01 7.037260e-02
```

According to Host Statistics and frequency table, visibly Zeppo is most used target for attacks. To take needed precautions, we can look for Zippo's information.

f) Attack Rate by Countries



In this Bar Chart, we can see that China has more attack rate than the other countries. According to this chart, people must be careful from IP addresses which come from China, United States, Japan etc.

2.2. Hypothesis 1

H_0: There is no relationship between Time of Attack and Used Protocol Type.

H_1: There is a relationship between Time of Attack and Used Protocol Type.

```
> chisq.test(sn$daytime, sn$proto)

        Pearson's Chi-squared test

data:  sn$daytime and sn$proto
X-squared = 5634.7, df = 6, p-value < 2.2e-16
```

```
> CrossTable(sn$daytime, sn$proto)

   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  451580
```

sn$daytime	sn$proto TCP	UDP	ICMP	Row Total
Night	93800	22576	14956	131332
	26.456	4.902	283.963	
	0.714	0.172	0.114	0.291
	0.286	0.287	0.334	
	0.208	0.050	0.033	
Morning	87216	23275	10810	121301
	8.928	211.150	125.054	
	0.719	0.192	0.089	0.269
	0.266	0.295	0.241	
	0.193	0.052	0.024	
Afternoon	85465	12115	10937	108517
	560.620	2454.062	2.642	
	0.788	0.112	0.101	0.240
	0.261	0.154	0.244	
	0.189	0.027	0.024	
Evening	61509	20813	8108	90430
	264.977	1608.457	83.480	
	0.680	0.230	0.090	0.200
	0.188	0.264	0.181	
	0.136	0.046	0.018	
Column Total	327990	78779	44811	451580
	0.726	0.174	0.099	

Conclusion: According to Chi-square test results p-value (2.2e-16) <0.05 and we can REJECT the null hyphothesis.

With %95 confidence, we have sufficient evidence to conclude that there is a relationship between Time of Attack and Used Protocol Type.

2.3. Hypothesis 2

H_0: There is no relationship between Month and Target Host.

H_1: There is a relationship between Month and Target Host.

```
> chisq.test(sn$month, sn$host)

        Pearson's Chi-squared test

data:  sn$month and sn$host
X-squared = 12700, df = 48, p-value < 2.2e-16
```

```
> CrossTable(sn$host, sn$month, prop.r = 'False', prop.c = 'False')

   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  451581
```

sn$host	sn$month 3	4	5	6	7	8	9	Row Total
28142724	0	0	0	0	0	0	0	0
	NaN	NaN	NaN	NaN	NaN	NaN	NaN	
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	
groucho-eu	2928	3135	3035	3423	4652	5077	1704	23954
	2.982	15.296	142.102	0.085	0.464	2.058	559.797	
	0.006	0.007	0.007	0.008	0.010	0.011	0.004	
groucho-norcal	3614	3626	3556	3437	4793	4324	1216	24566
	171.132	9.226	24.373	2.347	1.024	118.901	50.261	
	0.008	0.008	0.008	0.008	0.011	0.010	0.003	
groucho-oregon	13533	15370	15430	15710	15592	15322	3119	94076
	514.928	355.718	27.447	358.057	344.561	911.211	122.474	
	0.030	0.034	0.034	0.035	0.035	0.034	0.007	
groucho-sa	3085	3385	3996	3513	4832	4319	1186	24316
	14.755	0.223	7.780	0.125	5.247	106.084	42.134	
	0.007	0.007	0.009	0.008	0.011	0.010	0.003	
groucho-singapore	9104	12736	13860	11363	14094	13554	3440	78151
	2.390	285.028	200.913	1.733	57.872	442.369	25.211	
	0.020	0.028	0.031	0.025	0.031	0.030	0.008	
groucho-sydney	3824	3333	3685	3141	4219	4783	1471	24456
	297.763	2.868	6.702	39.233	49.672	17.374	235.889	
	0.008	0.007	0.008	0.007	0.009	0.011	0.003	
groucho-tokyo	9789	14388	20019	16485	27571	34287	3650	126189
	1776.052	623.023	1.573	147.944	450.992	2487.323	411.723	
	0.022	0.032	0.044	0.037	0.061	0.076	0.008	
groucho-us-east	4006	3942	3929	4484	6287	8052	1079	31779
	15.761	60.150	228.273	1.398	5.108	318.832	32.757	
	0.009	0.009	0.009	0.010	0.014	0.018	0.002	
zeppo-norcal	3582	3461	3498	3297	4788	4086	1382	24094
	186.498	1.873	22.292	7.699	5.207	168.707	171.349	
	0.008	0.008	0.008	0.007	0.011	0.009	0.003	
Column Total	53465	63376	71008	64853	86828	93804	18247	451581

Conclusion: According to Chi-square test results p-value (2.2e-16) <0.05 and we can REJECT the null hyphothesis.

With %95 confidence, we have sufficient evidence to conclude that there is a relationship between Month and Target Host.

2.4. Hypothesis 3

H_0: There is no relationship between Target Host and Used Protocol Type.

H_1: There is a relationship between Target Host and Used Protocol Type.

```
> chisq.test(sn$host, sn$proto)

        Pearson's Chi-squared test

data:  sn$host and sn$proto
X-squared = 36246, df = 16, p-value < 2.2e-16
```

```
> CrossTable(sn$host, sn$prot, prop.c = 'False', prop.r='False')

   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  451581
```

sn$host	sn$prot TCP	UDP	ICMP	Row Total
groucho-eu	17405	4380	2169	23954
	0.003	9.686	18.199	
	0.039	0.010	0.005	
groucho-norcal	16421	4823	3322	24566
	113.281	67.394	320.773	
	0.036	0.011	0.007	
groucho-oregon	84179	7755	2142	94076
	3676.656	4566.163	5542.776	
	0.186	0.017	0.005	
groucho-sa	17112	4429	2775	24316
	17.074	8.247	54.337	
	0.038	0.010	0.006	
groucho-singapore	61024	6165	10962	78151
	319.951	4091.331	1326.191	
	0.135	0.014	0.024	
groucho-sydney	17177	4479	2800	24456
	19.320	10.595	57.391	
	0.038	0.010	0.006	
groucho-tokyo	72809	37285	16095	126189
	3874.440	10593.667	1019.573	
	0.161	0.083	0.036	
groucho-us-east	24663	4643	2473	31779
	108.343	146.398	146.836	
	0.055	0.010	0.005	
zeppo-norcal	17201	4820	2073	24094
	5.105	90.501	42.264	
	0.038	0.011	0.005	
Column Total	327991	78779	44811	451581

Conclusion: According to Chi-square test results p-value (2.2e-16) <0.05 and we can REJECT the null hyphothesis.

With %95 confidence, we have sufficient evidence to conclude that there is a relationship between Target Host and Used Protocol Type.

2.5. Hypothesis 4

H_0: There is no relationship between Month and Time of Attack.

H_1: There is a relationship between Month and Time of Attack.

```
> chisq.test(sn$month, sn$daytime)

        Pearson's Chi-squared test

data:  sn$month and sn$daytime
X-squared = 8007.2, df = 18, p-value < 2.2e-16
```

```
> CrossTable(sn$month, sn$daytime, prop.r = 'False', prop.c = 'False')

   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  451580

             | sn$daytime
   sn$month |       Night |     Morning |   Afternoon |     Evening | Row Total |
------------|-------------|-------------|-------------|-------------|-----------|
           3 |       15575 |       13530 |       14400 |        9960 |     53465 |
             |       0.043 |      48.140 |     187.499 |      52.049 |           |
             |       0.034 |       0.030 |       0.032 |       0.022 |           |
------------|-------------|-------------|-------------|-------------|-----------|
           4 |       18245 |       16946 |       16252 |       11933 |     63376 |
             |       1.887 |       0.355 |      68.639 |      45.296 |           |
             |       0.040 |       0.038 |       0.036 |       0.026 |           |
------------|-------------|-------------|-------------|-------------|-----------|
           5 |       19551 |       17954 |       17980 |       15523 |     71008 |
             |      58.603 |      65.741 |      49.217 |     119.487 |           |
             |       0.043 |       0.040 |       0.040 |       0.034 |           |
------------|-------------|-------------|-------------|-------------|-----------|
           6 |       18409 |       17650 |       16791 |       12003 |     64853 |
             |      10.835 |       3.024 |      93.402 |      74.552 |           |
             |       0.041 |       0.039 |       0.037 |       0.027 |           |
------------|-------------|-------------|-------------|-------------|-----------|
           7 |       21692 |       30999 |       19342 |       14795 |     86828 |
             |     501.883 |    2526.095 |     111.199 |     386.551 |           |
             |       0.048 |       0.069 |       0.043 |       0.033 |           |
------------|-------------|-------------|-------------|-------------|-----------|
           8 |       32023 |       19456 |       19371 |       22954 |     93804 |
             |     824.330 |    1308.107 |     445.958 |     925.492 |           |
             |       0.071 |       0.043 |       0.043 |       0.051 |           |
------------|-------------|-------------|-------------|-------------|-----------|
           9 |        5837 |        4766 |        4381 |        3262 |     18246 |
             |      53.047 |       3.726 |       0.003 |      42.014 |           |
             |       0.013 |       0.011 |       0.010 |       0.007 |           |
------------|-------------|-------------|-------------|-------------|-----------|
Column Total |      131332 |      121301 |      108517 |       90430 |    451580 |
------------|-------------|-------------|-------------|-------------|-----------|
```

Conclusion: According to Chi-square test results p-value (2.2e-16) <0.05 and we can REJECT the null hyphothesis.

With %95 confidence, we have sufficient evidence to conclude that there is a relationship between Month and Time of Attack.

2.6. Hypothesis 5

H_0: There is no relationship between Day and Used Protocol Type.

H_1: There is a relationship between Day and Used Protocol Type.

```
> chisq.test(sn$day, sn$proto)

        Pearson's Chi-squared test

data:  sn$day and sn$proto
X-squared = 57648, df = 60, p-value < 2.2e-16
```

Conclusion: According to Chi-square test results p-value (2.2e-16) <0.05 and we can REJECT the null hyphothesis.

With %95 confidence, we have sufficient evidence to conclude that there is a relationship between Day and Used Protocol Type.

2.7. Hypothesis 6

H_0: There is no relationship between Time of Attack and Attacker Country.

H_1: There is a relationship between Time of Attack and Attacker Country.

```
> chisq.test(sn$daytime, sn$country)

        Pearson's Chi-squared test

data:  sn$daytime and sn$country
X-squared = 37980, df = 531, p-value < 2.2e-16
```

Conclusion: According to Chi-square test results p-value (2.2e-16) <0.05 and we can REJECT the null hyphothesis.

With %95 confidence, we have sufficient evidence to conclude that there is a relationship between Time of Attack and Attacker Country.

2.8. Hypothesis 7

H_0: There is no relationship between Used Protocol Type and Attacker Country.

H_1: There is a relationship between Used Protocol Type and Attacker Country.

```
> chisq.test(sn$proto, sn$country)

        Pearson's Chi-squared test

data:  sn$proto and sn$country
X-squared = 165160, df = 354, p-value < 2.2e-16
```

Conclusion: According to Chi-square test results p-value (2.2e-16) <0.05 and we can REJECT the null hyphothesis.

With %95 confidence, we have sufficient evidence to conclude that there is a relationship between Used Protocol Type and Attacker Country.

2.9. Hypothesis 8

H_0: The categories of Country occur with equal probabilities.

H_1: Countries have different probability to occur.

The categories of Country occur with equal probabilities.	One-Sample Chi-Square Test	,000	Reject the null hypothesis.

```
> chisq.test(table(dt$country))

        Chi-squared test for given probabilities

data:  table(dt$country)
X-squared = 17679000, df = 177, p-value < 2.2e-16
```

Conclusion: According to Chi-square test results p-value (2.2e-16) <0.05 and we can REJECT the null hyphothesis.

With %95 confidence, we have sufficient evidence to conclude that all countries have different probability to occur.

3. Conclusion

In this project, we tried to investigate and deduce honeypot data. According to data, we hypothesized with some attributes and made deductions. We did this process by our specified methodologies. Generally, we used Chi-squared Test to test our hypothesis. Because our data is categorical and non-numerical, we had to use that test procedure. The aim is that finding the relation between variables or seeing the probability for only one variable.

According to the information we have obtained as a result of our analyzes;

- Many precautions must be taken against specific protocols at certain times of day.
- Further measures should be taken in target systems that have been attacked these months due to the increase in attacks that occurred some months later.
- We have seen that some protocols are used more effectively in certain systems, so systems must be configured accordingly.
- Special precautions should be taken according to the protocols used for attacks on certain parts of the month and days.
- The attacking countries vary depending on the protocols used and the time of day. Therefore, we should be more careful about IP addresses from certain countries.

4. References

[1] Passeri, P. (2017, March 20). February 2017 Cyber Attack Statistics. http://www.hackmageddon.com: http://www.hackmageddon.com/2017/03/20/february-2017-cyber-attacks-statistics/

[2] Jacobs, J. (2014, April 30). DDS Dataset Collection. Data Driven Security: http://datadrivensecurity.info/blog/pages/dds-dataset-collection.html