MA541 Course Project

Sophia Roper
Deep Amish Shah
Tanmay Sule
Manoj Babu Kosaraju

1.0 Introduction

Bitcoin, the groundbreaking digital currency introduced in 2009 by an anonymous entity known as Satoshi Nakamoto, represents a paradigm shift in the realm of finance. Built upon blockchain technology, Bitcoin operates on a decentralized network, allowing for peer-to-peer transactions without the need for intermediaries like banks or governments. This decentralized nature ensures transparency, security, and censorship resistance, making Bitcoin transactions immutable and publicly accessible on the blockchain ledger.

The allure of Bitcoin lies in its promise of financial sovereignty, offering users greater control over their funds and the ability to transact globally with minimal fees and delays. As a result, Bitcoin has garnered widespread adoption and recognition as a legitimate form of currency, with an ever-expanding ecosystem of merchants, investors, and enthusiasts.

1.1 About Bitcoin Transactions

Bitcoin transactions are the cornerstone of the cryptocurrency's decentralized financial system, enabling peer-to-peer transfers of digital currency without the need for intermediaries. These transactions are recorded on a public ledger known as the blockchain, which maintains a transparent and immutable record of all Bitcoin activity. Each transaction comprises various components, including sender and recipient addresses, transaction amounts, and transaction timestamps.

The dataset utilized in this analysis contains a comprehensive record of Bitcoin transactions from January 2009 to December 2018, encompassing millions of transaction instances. Key features of these transactions, such as Length, Weight, Count, and Neighbors, provide insights into transaction patterns and behaviors. For instance, Length quantifies the mixing rounds on Bitcoin, while Weight and Count quantify transaction merge behavior and patterns.

1.2 Data Description

Datasource: - https://archive.ics.uci.edu/dataset/526/bitcoinheistransomwareaddressdataset

This dataset contains the daily transactions from the Bitcoin transaction graph from January, 2009 to December, 2018. Any network edges that transfer less than B0.3 were filtered out, and there are no missing values. It includes 2916697 transaction instances, each comprising of 10 features:

- Address, Year, Day

  These variables indicate the Bitcoin address, and the Date and Year that the transaction took place.

- Length, Weight, Count, Looped, Neighbors, Income

  These variables help quantify the patterns of the transactions. Looped counts the number of transactions, Weight quantifies the merge behavior, Count quantifies the merging pattern, Length quantifies the mixing rounds on Bitcoin, Neighbors is the

number of neighbors for the given node in the Bitcoin graph, and Income is the Satoshi amount of money sent in the transaction.

- Label

  This variable indicates the ransomware family associated with the transaction. If there is no ransomware associated with the transaction, then the label is white, otherwise it is the name of the ransomware family.

1.3 Goal

To build a classification model that can predict if a Bitcoin transaction is a ransomware payment based on its features, aiding in the detection of ransomware attacks.

2.0 Methods

In this project, we will use Python to analyze the data and build our classification model. In our statistical analysis, we visualize the data to form conclusions about the variables' distribution and relationships. We will also explore the centrality and variability within the dataset to form further conclusions. In our inferential analysis, we will use hypothesis testing to form conclusions about the correlation between the key features and the target variable (Label). After these analyses, we will develop our classification model by comparing the performances of both single-variable and multiple-variable logistic regression models. The statistical results of each model will be examined to determine their efficacy.

3.0 Statistical Analysis

3.1 Descriptive Analysis

To simplify our computation, we selected the variables Length, Weight, Count, and Neighbors as our key features to examine. The Label column was also converted into a binary variable for future computational purposes. Any labels marked as ‘white’ were replaced with 0 and the ransomware labels were replaced with 1.

We implemented a box plot and QQ plot visualization to depict the distribution of these variables. A QQ plot compares the distribution of two sets of quantiles, and it forms a roughly linear scatter plot if the sets come from the same distribution. Using a normal QQ plot provides insight into the normality of the distribution for each variable. A box plot summarizes a set of data by visualizing its minimum, first quartile, median, third quartile, and maximum. This plot can help show the distribution of data and gives information about a dataset’s symmetry, skewness, variance, and outliers. From Figure 1, we observed that all four variables are heavily skewed to the right, with the weight, count, and neighbors variables appearing to be heavily impacted by outliers. To confirm this observation, we conducted the D'Agostino-Pearson and Shapiro-Wilk tests for normality on each of the variables.

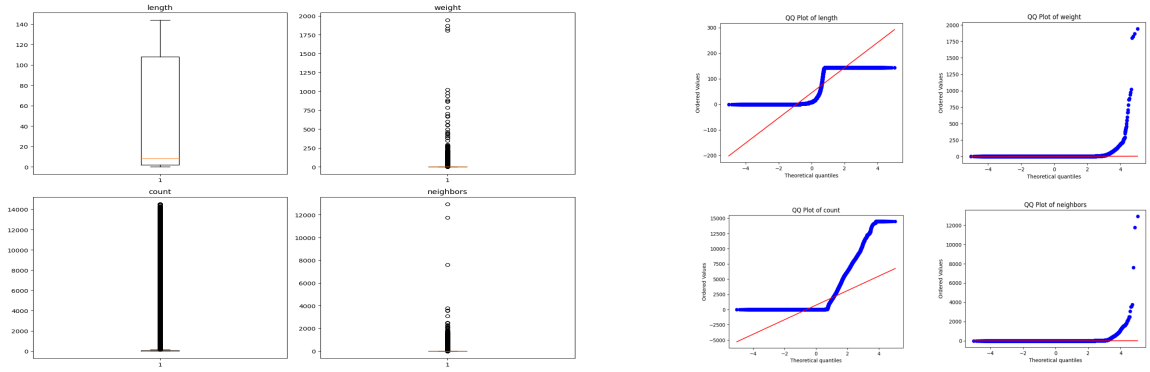

Figure 1. Illustration of the distribution of features through box plots and QQ plots

The D'Agostino-Pearson uses the skewness and kurtosis of data to determine how far a dataset's distribution is from the normal distribution while the Shapiro-Wilk test orders and standardizes the sample data to see if it follows the normal distribution. The results from conducting these tests on the Weight, Length, Count, and Neighbors variables can be seen in Figure 2. These results confirm our observations that none of the four variables follow a normal distribution. Each of the results show that the distribution of each variable is significantly different from a normal distribution.

```
Weight: NormaltestResult(statistic=15083010.11489625, pvalue=0.0)
Length: NormaltestResult(statistic=864404.313353176, pvalue=0.0)
Count: NormaltestResult(statistic=1771811.6898896296, pvalue=0.0)
Neighbors: NormaltestResult(statistic=15891400.930463219, pvalue=0.0)
```

```
Weight: ShapiroResult(statistic=0.045362114906311035, pvalue=0.0)
Length: ShapiroResult(statistic=0.7001427412033081, pvalue=0.0)
Count: ShapiroResult(statistic=0.5025076866149902, pvalue=0.0)
Neighbors: ShapiroResult(statistic=0.012168943881988525, pvalue=0.0)
```

Figure 2. The results of the D'Agostino-Pearson and Shapiro-Wilk normality tests

We also created a table with the full statistical description of each of the variables to provide further insight into the visualizations and normality tests. Each of the variables

appear to have a significant variability that explains their skewed distribution. Weight has a minimum value of 3.6065e-94, a maximum value of 1943.749, and a mean value of 0.5455, which explains its highly skewed behavior. Similarly, Neighbors has a minimum value of 1, a maximum value of 12920, and a mean value of 2.2065. The comparison between median and means for the Length and Count also confirms that their distribution is skewed as a result of outliers. Length has a mean value of 45.0086 and a median value of 8, and Count has a mean value of 721.6446 and a median value of 1.

	weight	length	count	neighbors
count	2.916697e+06	2.916697e+06	2.916697e+06	2.916697e+06
mean	5.455192e-01	4.500859e+01	7.216446e+02	2.206516e+00
std	3.674255e+00	5.898236e+01	1.689676e+03	1.791877e+01
min	3.606469e-94	0.000000e+00	1.000000e+00	1.000000e+00
25%	2.148438e-02	2.000000e+00	1.000000e+00	1.000000e+00
50%	2.500000e-01	8.000000e+00	1.000000e+00	2.000000e+00
75%	8.819482e-01	1.080000e+02	5.600000e+01	2.000000e+00
max	1.943749e+03	1.440000e+02	1.449700e+04	1.292000e+04

Table 1. Analysis of the key features with important statistical measures

3.2 Inferential Analysis

3.2.1 Correlation Heat Map

To better understand the relationship between the variables, we created a heatmap to visualize the correlation between each of the features, including the label column. As seen in Figure 3, the variable pairs Weight-Neighbors and Length-Count are highly correlated with each other, with a coefficient of 0.7. The remaining variable pairs do not appear to be correlated with each other.

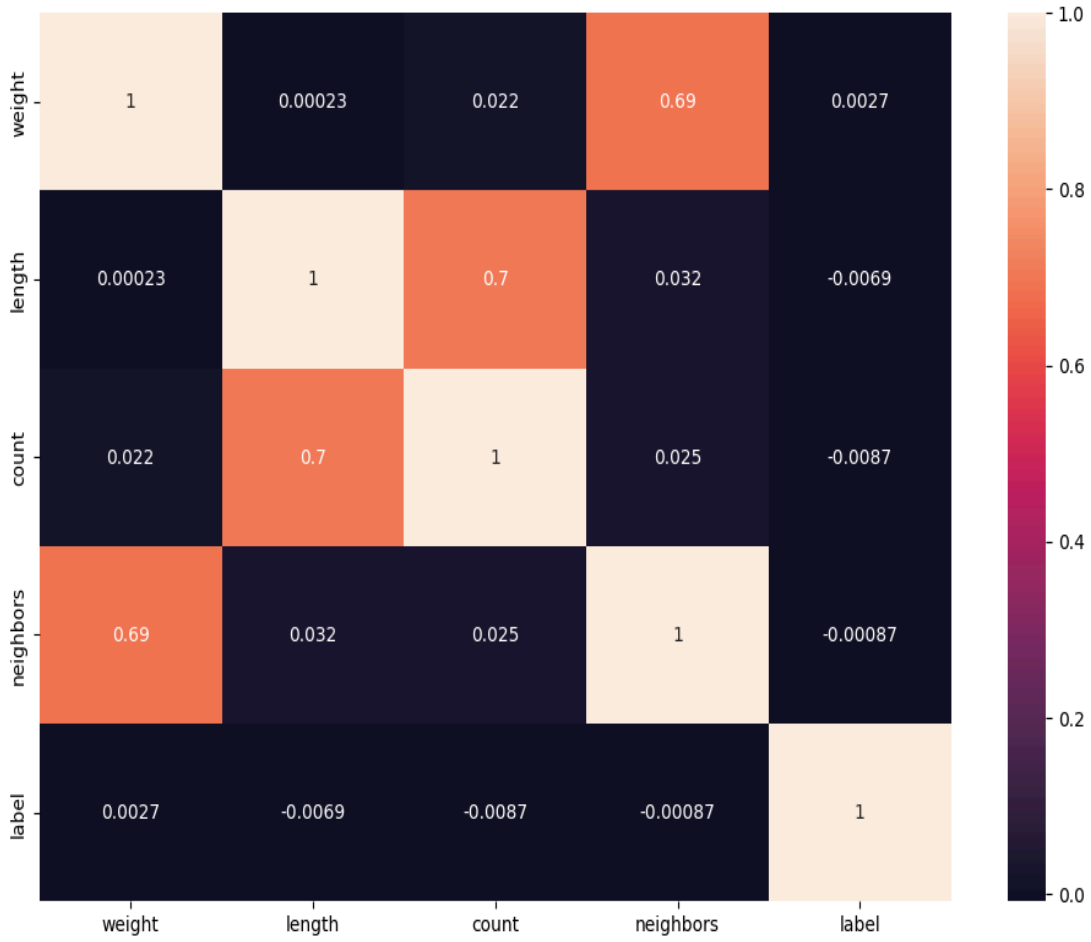

Figure 3. A heatmap that displays the correlation coefficients for each pair of variables

3.2.2 Central Limit Theorem

We then used the Length variable to delve into the Central Limit Theorem (CLT). The Central Limit Theorem states that a sample variable approximately follows a normal distribution regardless of the distribution of the population given that a sufficiently large sample size is used. We randomly sampled 10% of the total dataset, for a sample size of 291670, to test if this theorem holds for our dataset.

First, we partitioned the Length variable, the population in this case, into 30 sequential groups, each containing 100000 values. As seen in Figure 4, while the histogram does not clearly imitate a normal distribution, the QQ plot shows a relatively linear scatterplot. This roughly normal appearance appears to follow the core principles of the Central Limit Theorem.

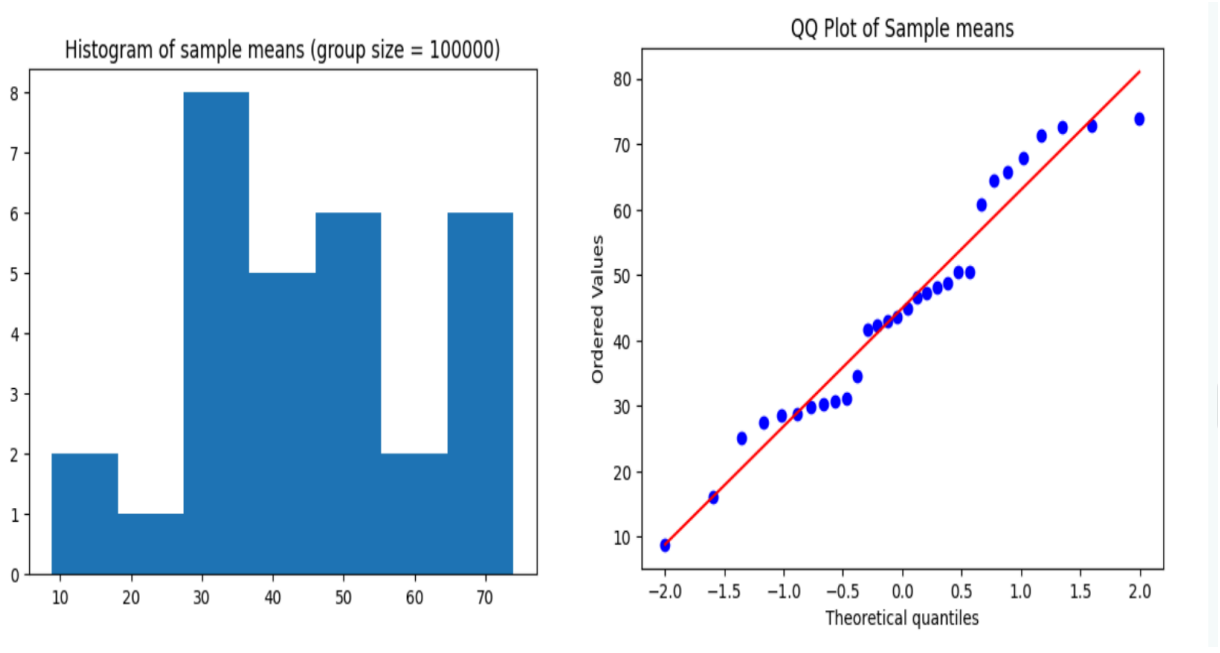

Figure 4. Histogram and QQ plot for the distribution of the 30 sample group

To further corroborate these findings, we also computed the requisite statistics. The mean of the sample means (44.9516) was approximately equal to the population mean (45.0086). The standard deviation of the sample means was calculated to be 17.4123 and the population standard deviation divided by the square root of n (100000), which indicates the expected standard error of the sample means, was calculated to be 0.1865. The standard deviation of the sample means is significantly lower than the population standard deviation, which indicates that the sample means have a decreased variability. The standard deviation of sample means being larger than the population standard deviation divided by the square root of n is understandable and expected due to the variability of the process, especially because the Length variable is highly varied. These statistical results help to confirm our observation that the sample means are approximately normally distributed, and support the principles of the Central Limit Theorem.

As we explored the Central Limit Theorem, we also partitioned the Length population into 10 sequential groups, each containing 300000 values. As seen in Figure 5, the histogram is tri-modal and follows a normal distribution less closely than the previous samples. However the QQ plot is more linear when compared to the previous samples. Like the previous example, the approximately normal QQ plot is consistent with the principles of the Central Limit Theorem.

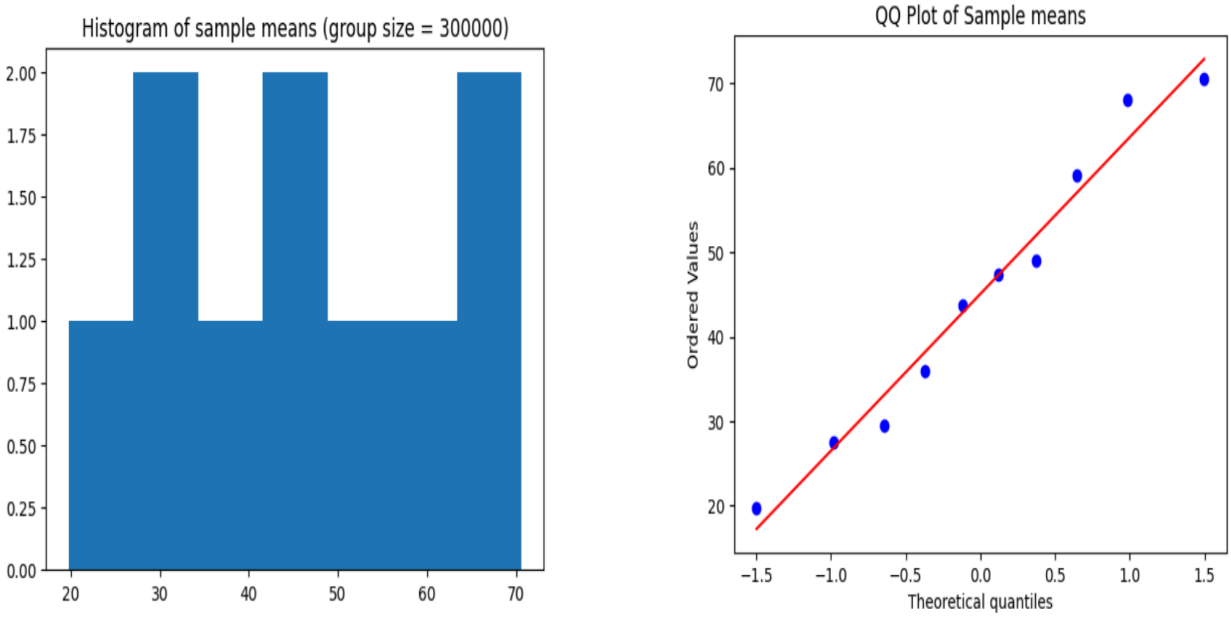

Figure 5. Histogram and QQ plot for the distribution of the 10 sample group

The population mean (45.0086) and mean of sample means (45.0737) remained consistently similar, which aligns with the Central Limit Theorem. The population standard deviation divided by the square root of n (300000) was calculated to be 0.1077, which is slightly lower than the value for the previous samples because of the larger sample size. The standard deviation of the sample means was calculated to be 16.3465, which again is higher than the population standard deviation divided by the square root of n. Like the smaller samples, this variability of the sample means is an explanation for this disparity. Nevertheless, both samples' approximately linear QQ plot and similar population mean support the principles of the Central Limit Theorem.

3.2.3 Confidence Interval

After examining the Central Limit Theorem, we used the generated samples to examine confidence intervals for the population mean of the Length variable. Confidence intervals give a range of values that you expect to hold your estimate with a given level of confidence. A 95% confidence interval would mean you expect the range of values to hold your estimate 95% of the time. We selected a single sample from the collection of 30 samples that each contain 100000 values that was defined above. The lower bound value was calculated to be 43.5885 and the upper bound value was calculated to be 44.0134. These values mean we are 95% confident that the true population mean falls within the range 43.5885 to 44.0134. We also selected a single random sample from the collection of 10 samples that each contained 300000 values. The lower bound value was calculated to be 44.6137 and the upper bound value was calculated to be 45.3536. Only the second interval actually contained the population mean (45.0086), which indicates that the population mean is more likely to fall within this interval compared to the other interval.

3.2.4 Independent Two-Sample T-Test

Before we started forming the classification model, we more closely examined the relationship between the Label and other variables. We conducted multiple independent two-sample t-tests to determine if there was a statistically significant difference between the means of the two Label groups (white vs non-white) for each of the quantitative variables. This test uses the Null Hypothesis - There is not a difference between the means of the two groups - and the Alternative Hypothesis - There is a difference between the

means of the two groups. It generates a t-statistic, which is calculated by dividing the difference between the two sample means by the estimated standard error, and an associated p-value. If the t-statistic is greater than the critical value, then the null hypothesis is rejected, and, if the t-statistic is less than the critical value, the null hypothesis is not rejected. If the p-value is less than the given alpha, the null hypothesis is rejected, and if the p-value is greater than or equal to the alpha, the null hypothesis is not rejected. We used an alpha level of 0.05 in our t-test, and the results are given in Figure 6.

```
T-test for 'length' and 'label':
T-statistic: -2.3758356634709794, P-value: 0.017509901031010345
Reject null hypothesis

T-test for 'weight' and 'label':
T-statistic: 0.3943609569198237, P-value: 0.6933148531783684
Fail to reject null hypothesis

T-test for 'count' and 'label':
T-statistic: -4.5447109981816265, P-value: 5.503271072713147e-06
Reject null hypothesis

T-test for 'neighbors' and 'label':
T-statistic: -0.31529196433179285, P-value: 0.75254021120546
Fail to reject null hypothesis
```

Figure 6: Results of the independent two-sample t-test

The results of the test show that there is a significant difference in mean length between white and non-white labels, and there is a significant difference in mean count between white and non-white labels. However, there is not a significant difference in mean weight between the white and non-white labels and there is not a significant difference in mean neighbors between the white and non-white labels.

4.0 Classification Model

4.1 Simple Logistic Regression

In our first logistic regression models, we used each of the individual variables as input and the Label variable as the output. Table 2 shows the regression coefficient, standard error, z-score, p-value, lower confidence interval, and upper confidence interval for each of the four models.

	names	coef	se	z	pval	CI[2.5%]	CI[97.5%]
0	Intercept	-4.241393	0.004959	-855.326373	0.000000	-4.251112	-4.231674
1	weight	0.001897	0.000507	3.744542	0.000181	0.000904	0.002890

	names	coef	se	z	pval	CI[2.5%]	CI[97.5%]
0	Intercept	-4.208338	0.005330	-789.610341	0.0	-4.218784	-4.197892
1	count	-0.000049	0.000003	-14.751152	0.0	-0.000055	-0.000042

	names	coef	se	z	pval	CI[2.5%]	CI[97.5%]
0	Intercept	-4.196516	0.006119	-685.813976	0.0	-4.208509	-4.184523
1	length	-0.001011	0.000086	-11.708912	0.0	-0.001180	-0.000842

	names	coef	se	z	pval	CI[2.5%]	CI[97.5%]
0	Intercept	-0.007134	0.001472	-4.847135	0.000001	-0.010018	-0.004249
1	neighbors	-0.015768	0.000461	-34.212406	0.000000	-0.016671	-0.014865

Table 2. Results from the simple logistic regression models

The statistically significant p-values in each of the model results show that all four variables are significant predictors of the Label variable.

4.2 Multiple Logistic Regression

After running the simple logistic regression models, we created a multiple logistic regression model that took all four variables as input and the Label variable as output. Table 2 shows the regression coefficient, standard error, z-score, p-value, lower confidence interval, and upper confidence interval for the multiple logistic regression model. The statistically significant p-value of each coefficient indicates that all four variables are significant predictors of the Label variable.

	names	coef	se	z	pval	CI[2.5%]	CI[97.5%]
0	Intercept	-0.016967	0.002747	-6.175650	0.000000	-0.022352	-0.011582
1	weight	-0.009592	0.002536	-3.782284	0.000155	-0.014563	-0.004622
2	length	-0.083074	0.000215	-386.311521	0.000000	-0.083496	-0.082653
3	count	0.000912	0.000004	248.326582	0.000000	0.000905	0.000919
4	neighbors	-0.029317	0.001314	-22.304165	0.000000	-0.031894	-0.026741

Table 3. Results from the multiple logistic regression model

The prediction equation obtained from the model is:

Label = -0.016967 - (0.009594 x Weight) - (0.083074 x Length) + (0.000912 x Count) - (0.029317 x Neighbors)

5.0 Summary and Conclusion

5.1 Summary

In this study, we employed Python for comprehensive data analysis and model development. Utilizing descriptive statistics such as box plots and QQ plots, we visualized variable distributions and pinpointed outliers, providing valuable insights into the dataset's characteristics, which was heavily skewed. Our inferential analysis encompassed correlation heatmaps, exploration of the Central Limit Theorem (CLT), confidence interval calculations, and independent two-sample t-tests, revealing significant associations between variables like Weight and Neighbors and shedding light on the distributional properties of key features like Length. Notably, our examination of the CLT indicated that sample means approximate normal distributions with large sample sizes, enhancing our

understanding of statistical principles in this context. Furthermore, confidence intervals offered nuanced perspectives on the population mean of the Length variable, highlighting varying degrees of accuracy in our estimates. Independent two-sample t-tests uncovered significant differences in mean length and count between transactions associated with ransomware payments and those that were not, underscoring the discriminative power of these features. Regarding model performance, simple logistic regression models exhibited statistically significant predictive capabilities for individual variables, while a multiple logistic regression model incorporating all variables demonstrated considerable efficacy in identifying ransomware-related transactions. These findings collectively contribute to a deeper understanding of Bitcoin transaction patterns and lay a foundation for enhanced cybersecurity measures in the digital realm.

5.2 Conclusion:

The analysis indicates that specific transaction features, including Length, Weight, Count, and Neighbors, demonstrate potential predictive capabilities in identifying ransomware payments within Bitcoin transactions. Moreover, the developed classification model, especially the multiple logistic regression model, emerges as a valuable asset in detecting ransomware activity within Bitcoin transactions, thereby contributing significantly to cybersecurity efforts. Further refinement and validation of this model on additional datasets hold promise for enhancing its reliability and relevance in real-world applications. In summary, this study enriches our comprehension of Bitcoin transaction patterns and offers a practical framework for detecting ransomware activities within cryptocurrency transactions, thus addressing a crucial aspect of cybersecurity.