Members: Deep Amish Shah, Manoj Babu Kosaraju, Sophia Roper, Tanmay Sule

Overview:

This dataset contains the entire Bitcoin transaction graph from January 2009 to December 2018. The researchers extracted daily transitions on the network and formed the graph using a time interval of 24 hours. Any network edges that transfer less than B0.3 were filtered out, and there are no missing values.

The variables of this dataset include:

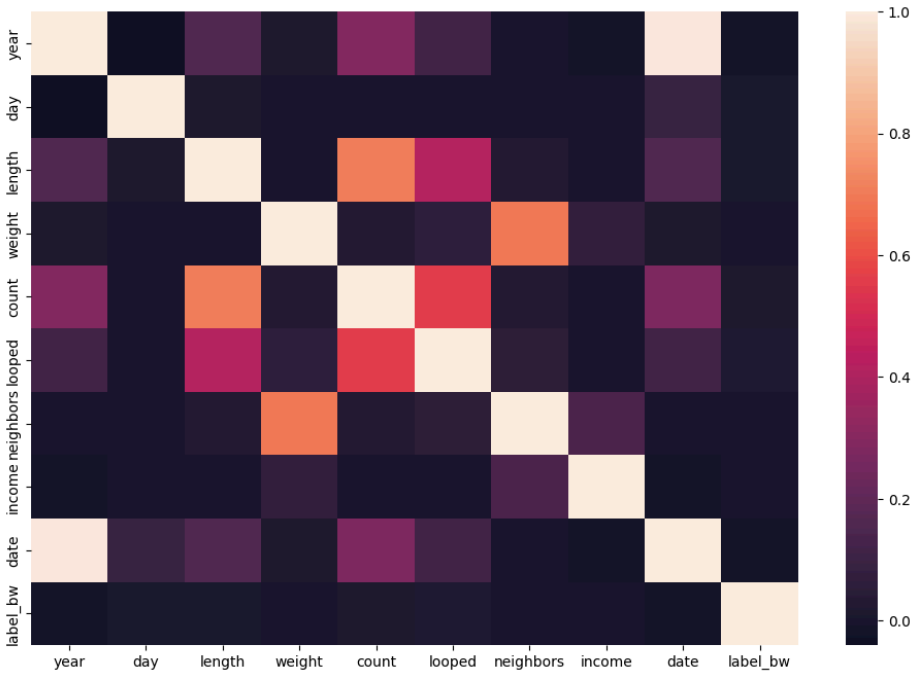
- Address (Text)
- Year (Number)
- Day (Number)
- Length (Number)
- Weight (Number)
- Count (Number)
- Looped (Number)
- Neighbors (Number
- Income (Number)
- Label (Text)

Data visualization and summary statistics:

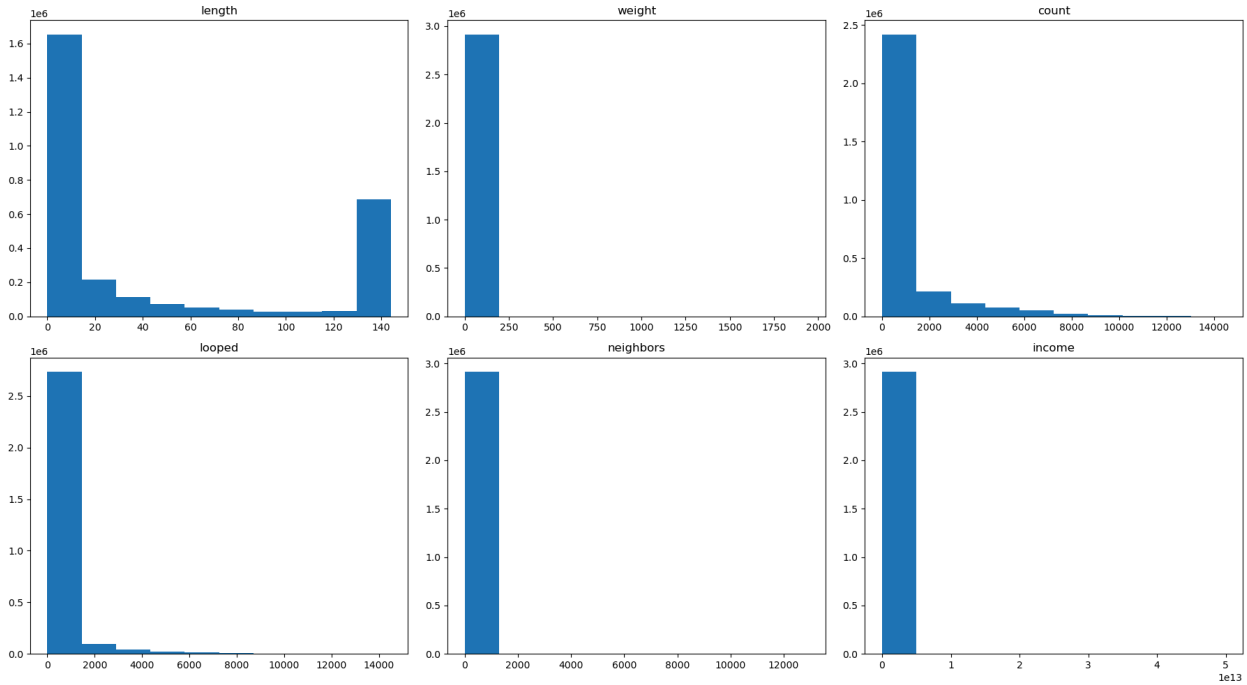
Mean year = 2014.475011288454, year standard deviation = 2.2573971651095768
Mean day = 181.457211016434, day standard deviation = 104.0118179365244
Mean length = 45.00859293920486, length standard deviation = 58.98235218811939
Mean weight = 0.5455192341640024, weight standard deviation = 3.6742546257308684
Mean count = 721.6446428957139, count standard deviation = 1689.6755041726624
Mean looped = 238.50669884461772, looped standard deviation = 966.3215201658936
Mean neighbors = 2.206516137946451, neighbors standard deviation = 17.918762006490027
Mean income = 4464889007.186174, income standard deviation = 162685932780.61386

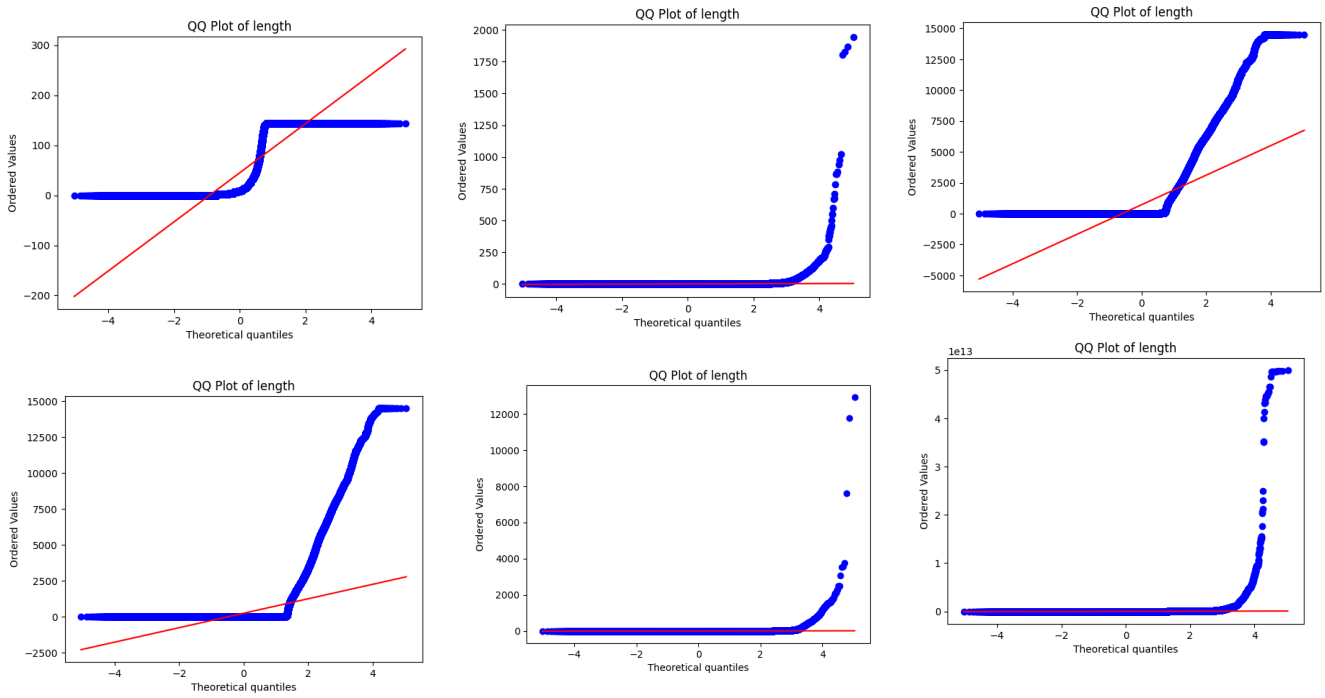
The label column, which indicates the ransomware associated with the data, overwhelmingly contained the label ‘white,’ which means there was not known to be ransomware with the transaction. For the correlation, we replaced the white labels with ‘1’ and the non-white labels with ‘0’ to observe the correlation between the other variables and the label as well.

	year	day	length	weight	count	looped	neighbors	income	date	label_bw
year	1.000000	-0.040307	0.163101	0.011827	0.285415	0.113164	-0.000876	-0.020535	0.992071	-0.021367
day	-0.040307	1.000000	0.011919	0.000864	-0.002538	0.002566	0.000181	0.002628	0.085585	0.008097
length	0.163101	0.011919	1.000000	0.000228	0.703467	0.411609	0.031523	0.000488	0.164124	0.006860
weight	0.011827	0.000864	0.000228	1.000000	0.022313	0.061646	0.691963	0.069774	0.011904	-0.002676
count	0.285415	-0.002538	0.703467	0.022313	1.000000	0.560370	0.025441	-0.003635	0.284270	0.008654
looped	0.113164	0.002566	0.411609	0.061646	0.560370	1.000000	0.052826	0.002551	0.113153	0.017810
neighbors	-0.000876	0.000181	0.031523	0.691963	0.025441	0.052826	1.000000	0.138966	-0.000851	0.000872
income	-0.020535	0.002628	0.000488	0.069774	-0.003635	0.002551	0.138966	1.000000	-0.020147	0.002716
date	0.992071	0.085585	0.164124	0.011904	0.284270	0.113153	-0.000851	-0.020147	1.000000	-0.020282



It appears that the length and count columns as well as the weight and neighbors had the strongest correlation out of the column pairs, with coefficients of 0.7034 and 0.6919, respectively.





After generating the histograms and qq plots, the distribution of the columns all appear to be heavily skewed to the right. None of the columns appear to follow a normal distribution. To confirm our hypotheses about the columns not following a normal distribution, we decided to run a normality test. And for more manageability, we decided to proceed with just the weight, length, count, and neighbors column for the rest of the analysis.

We conducted the D'Agostino-Pearson test for normality on each of our four variables with the following results:

```
NormaltestResult(statistic=15083010.11489625, pvalue=0.0)
NormaltestResult(statistic=864404.313353176, pvalue=0.0)
NormaltestResult(statistic=1771811.6898896296, pvalue=0.0)
NormaltestResult(statistic=15891400.930463219, pvalue=0.0)
```

These results indicate that it is extremely unlikely that any of the columns follow a normal distribution.

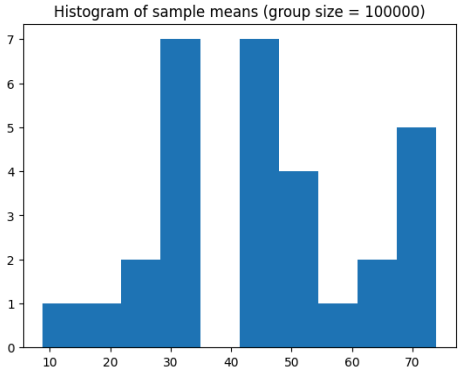
We also conducted the Shapiro-Wilk test for normality on each of the four variables with the following results:

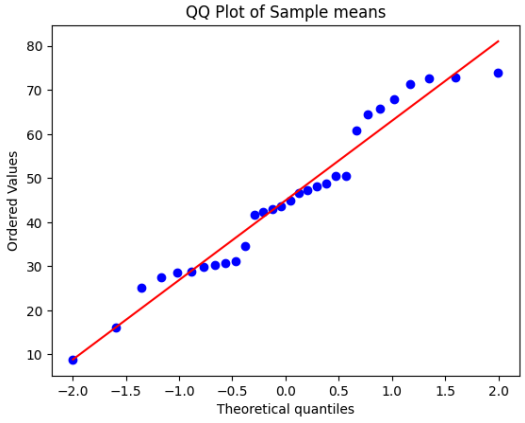
```
ShapiroResult(statistic=0.045362114906311035, pvalue=0.0)
ShapiroResult(statistic=0.7001427412033081, pvalue=0.0)
ShapiroResult(statistic=0.5025076866149902, pvalue=0.0)
ShapiroResult(statistic=0.012168943881988525, pvalue=0.0)
```

Like the previous results, these indicate that it is extremely unlikely that any of the columns follow a normal distribution.

CLT:

We then investigated the central limit theorem for the length column by generating a histogram, qq-plot and statistical measure comparison for 30 samples divided sequentially from the population.

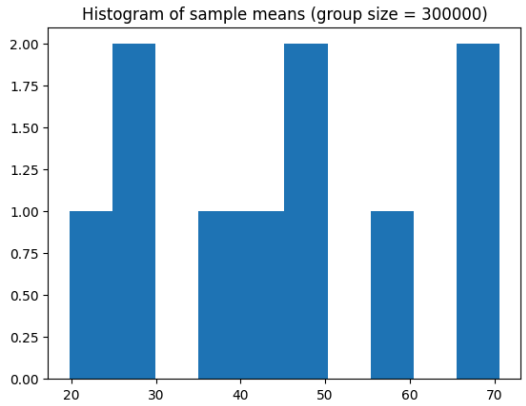


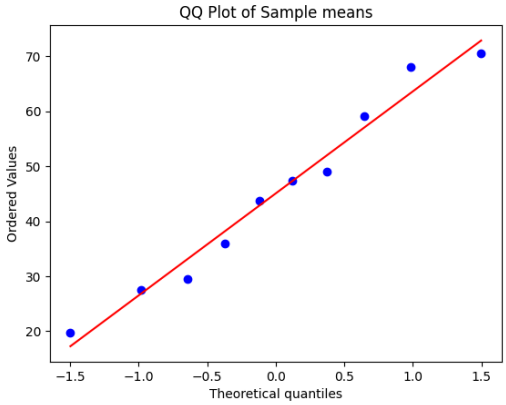



```
Population mean: 45.00859293920486
Mean of sample means: 44.951606648739315
Population standard deviation: 58.98235218811939
Standard deviation of sample means: 17.412318473464754
Population standard deviation divided by sqrt(n): 0.034536349493286635
```

While the comparison between the population and mean of sample means follows the CLT, the relationship between the standard deviation of the sample means and the population standard deviation divided by sqrt(n) is not approximately equal and does not follow the CLT.

We then did the same as above with 10 samples divided sequentially from the population.






```
Population mean: 45.00859293920486
Mean of sample means: 45.07368511257962
Population standard deviation: 58.98235218811939
Standard deviation of sample means: 16.34651349643786
Population standard deviation divided by sqrt(n): 0.034536349493286635
```

Like the previous example, while the comparison between the population and mean of sample means follows the CLT, the relationship between the standard deviation of the sample means and the population standard deviation divided by sqrt(n) is not approximately equal and does not follow the CLT. However, because our previous tests showed that the population does not follow a normal distribution, these results are expected.

Confidence Interval:

We then selected one random sample from the 10 sample and 30 sample cases and created a 95% confidence interval of the mean.

```
95% Confidence Interval for 10 sample case: (43.58853432700981, 44.013385672990196)
95% Confidence Interval for 30 sample case: (44.61371891914963, 45.352561080850364)
```

The second interval is the only one that captures the population mean of 45.0737 and is the more accurate of the two.

Hypothesis Testing:

We then used the random samples from above to perform a two-tailed t-test between the population mean and the mean of each sample.

For the ten sample case:

H0: μ = 45.0737

HA: μ != 45.0737

x_bar = 43.9009, s = 59.3632, n = 300000, a = .05
Mean T-statistic: -11.74309390632738, P-value: 7.787233742908089e-32
Reject null hypothesis

The null hypothesis is rejected because the p-value is < a.

H0: σ = 58.9824
HA: σ != 58.9824

x_bar = 43.9009, s = 59.3632, n = 300000, a = .05

Chi-square statistic: 21600000.0, P-value: 0.0
Reject null hypothesis

The null hypothesis is rejected because the p-value is < a

For the thirty sample case:

H0: μ = 45.0737
HA: μ != 45.0737

X_bar = 44.9831, s = 59.6027, n = 100000, a = .05
Mean T-statistic: -0.48047200340394564, P-value: 0.6308928569981553
Fail to reject null hypothesis

We do not reject the null hypothesis because the p-value is > a.

H0: σ = 58.9824
HA: σ != 58.9824

X_bar = 44.9831, s = 59.6027, n = 100000, a = .05

Chi-square statistic: 7200000.0, P-value: 0.0
Reject null hypothesis

The null hypothesis is rejected because the p-value is < a

We perform independent t-tests to compare the means of the numerical columns between two groups defined by the label column.

H0: There is not a difference between the means of the two groups (white vs non-white)
HA: There is a difference between the means of the two groups.

```
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
```

Our tests show that there was a significant difference between the means of the two groups for length and count.