Analysis and Classification of Bitcoin Ransomwares (Intermediate track of DataHacks 2021)

Kevin Zhao, Yunyi Huang, Alan Zhang

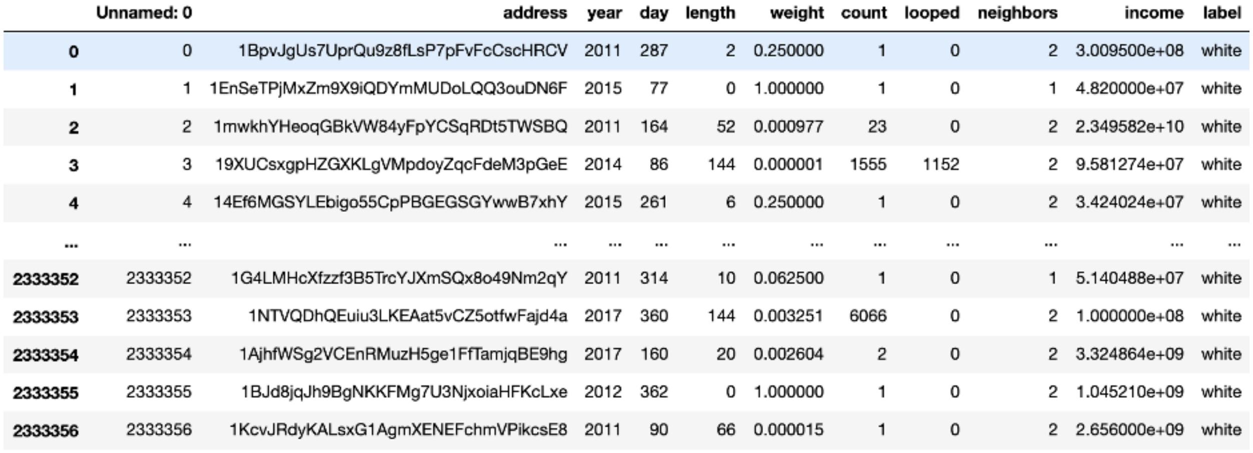
April 11, 2021

Introduction

Ransomware has been prevalent since the day computers were commercialized. The invention of Bitcoin and other cryptocurrency supported by the blockchain technology has sparked a whole new currency revolution. Cryptocurrency democratized currency in the hands of the public free from government regulation but also became another asset of the dark economy in courtesy of its anonymity. Bitcoin is the most popular form of cryptocurrency used by ransomware families to collect ransom from its victims anonymously. Based on the transaction history of Bitcoins from January of 2009 to December of 2018, we attempted to create machine learning models to classify ransomware transactions from regular transactions and predict whether a future transaction is ransom or not and classify them into a ransomware family.

Data Cleaning

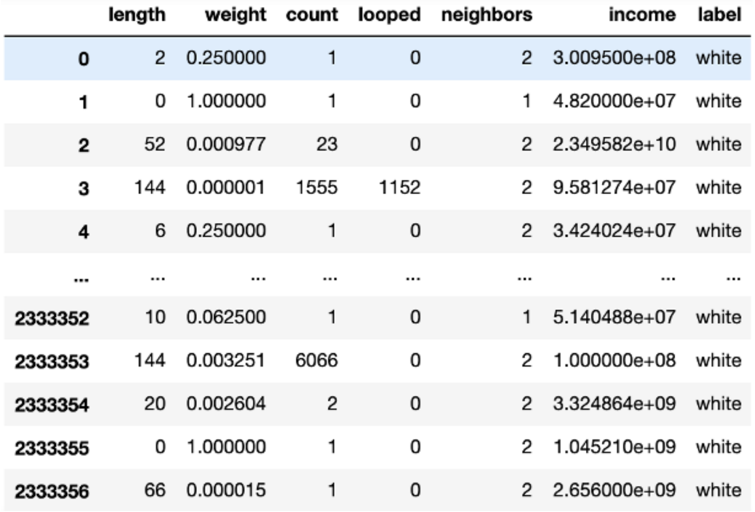
Original Dataset



	Unnamed: 0	address	year	day	length	weight	count	looped	neighbors	income	label
0	0	1BpvJgUs7UprQu9z8fLsP7pFvFcCscHRCV	2011	287	2	0.250000	1	0	2	3.009500e+08	white
1	1	1EnSeTPjMxZm9X9iQDYmMUDoLQQ3ouDN6F	2015	77	0	1.000000	1	0	1	4.820000e+07	white
2	2	1mwkhYHeoqGBkVW84yFpYCSqRDt5TWSBQ	2011	164	52	0.000977	23	0	2	2.349582e+10	white
3	3	19XUCsxgpHZGXKLgVMpdoyZqcFdeM3pGeE	2014	86	144	0.000001	1555	1152	2	9.581274e+07	white
4	4	14Ef6MGSYLEbigo55CpPBGEGSGYwwB7xhY	2015	261	6	0.250000	1	0	2	3.424024e+07	white
...	...	...	...	...	...	...	...	...	...	...	...
2333352	2333352	1G4LMHcXfzzf3B5TrcYJXmSQx8o49Nm2qY	2011	314	10	0.062500	1	0	1	5.140488e+07	white
2333353	2333353	1NTVQDhQEuiu3LKEAat5vCZ5otfwFajd4a	2017	360	144	0.003251	6066	0	2	1.000000e+08	white
2333354	2333354	1AjhfWSg2VCEnRMuzH5ge1FfTamjqBE9hg	2017	160	20	0.002604	2	0	2	3.324864e+09	white
2333355	2333355	1BJd8jqJh9BgNKKFMg7U3NjxoiaHFKcLxe	2012	362	0	1.000000	1	0	2	1.045210e+09	white
2333356	2333356	1KcvJRdyKALsxG1AgmXENEFchmVPikcsE8	2011	90	66	0.000015	1	0	2	2.656000e+09	white

We have dropped the column of “unnamed:0,” “address,” “year,” and “day” since these are either not useful or irrelevant for our later analysis. There are no missing values in the data frame. However, we have noticed that there are many labels which only occur a small number of times. Therefore, we replaced all labels that occurred under 1000 times to a single label “other” for better interpretation.

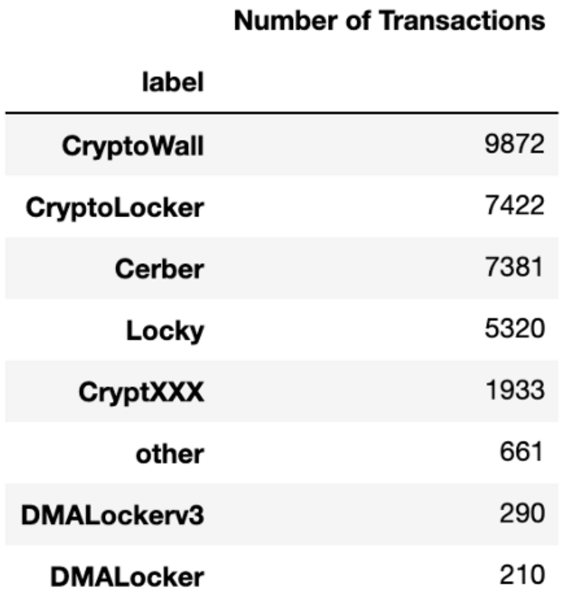
Modified Dataset



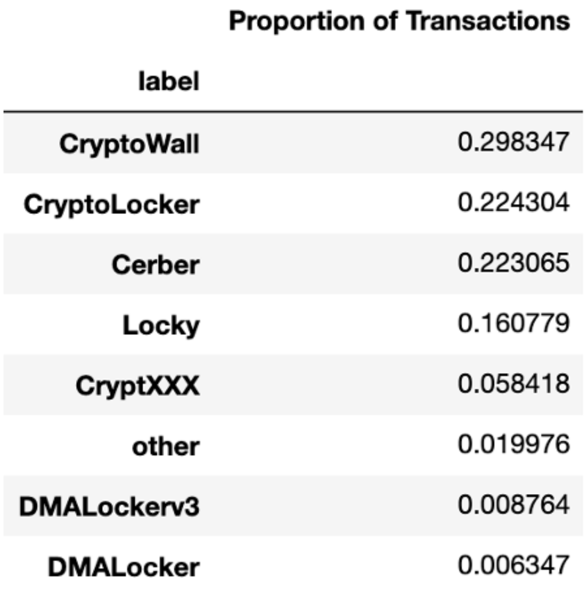
	length	weight	count	looped	neighbors	income	label
0	2	0.250000	1	0	2	3.009500e+08	white
1	0	1.000000	1	0	1	4.820000e+07	white
2	52	0.000977	23	0	2	2.349582e+10	white
3	144	0.000001	1555	1152	2	9.581274e+07	white
4	6	0.250000	1	0	2	3.424024e+07	white
...	...	...	...	...	...	...	...
2333352	10	0.062500	1	0	1	5.140488e+07	white
2333353	144	0.003251	6066	0	2	1.000000e+08	white
2333354	20	0.002604	2	0	2	3.324864e+09	white
2333355	0	1.000000	1	0	2	1.045210e+09	white
2333356	66	0.000015	1	0	2	2.656000e+09	white

Descriptive Analysis

According to the prompt, the variable “count” represents the number of transactions. **Based on the variable of “count,” we have identified that the top three ransom labels that have the most ransom transactions are CryptoWall, CryptoLocker, and Cerber.**



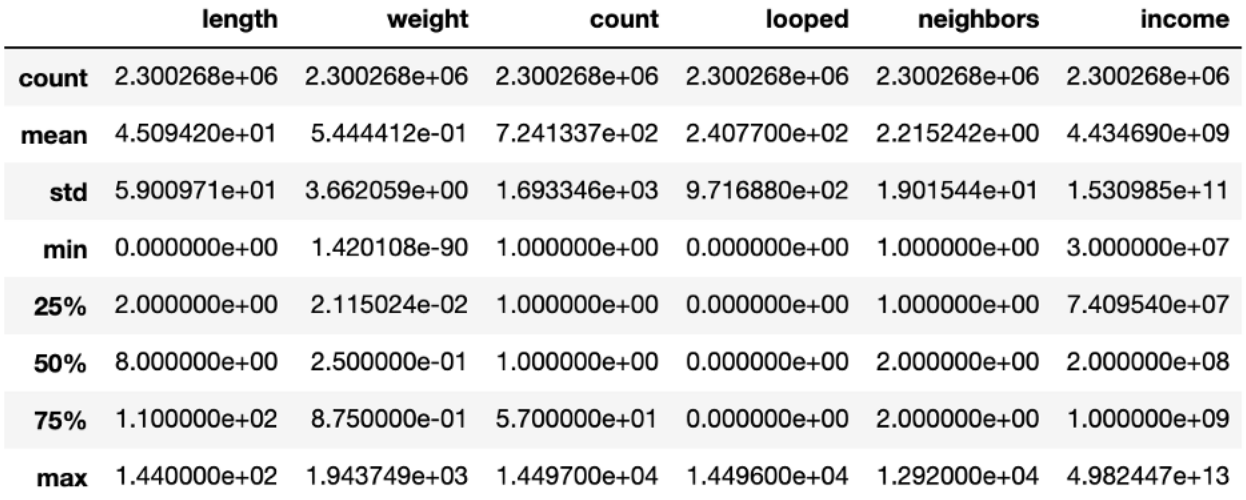
label	Number of Transactions
CryptoWall	9872
CryptoLocker	7422
Cerber	7381
Locky	5320
CryptXXX	1933
other	661
DMALockerv3	290
DMALocker	210



label	Proportion of Transactions
CryptoWall	0.298347
CryptoLocker	0.224304
Cerber	0.223065
Locky	0.160779
CryptXXX	0.058418
other	0.019976
DMALockerv3	0.008764
DMALocker	0.006347

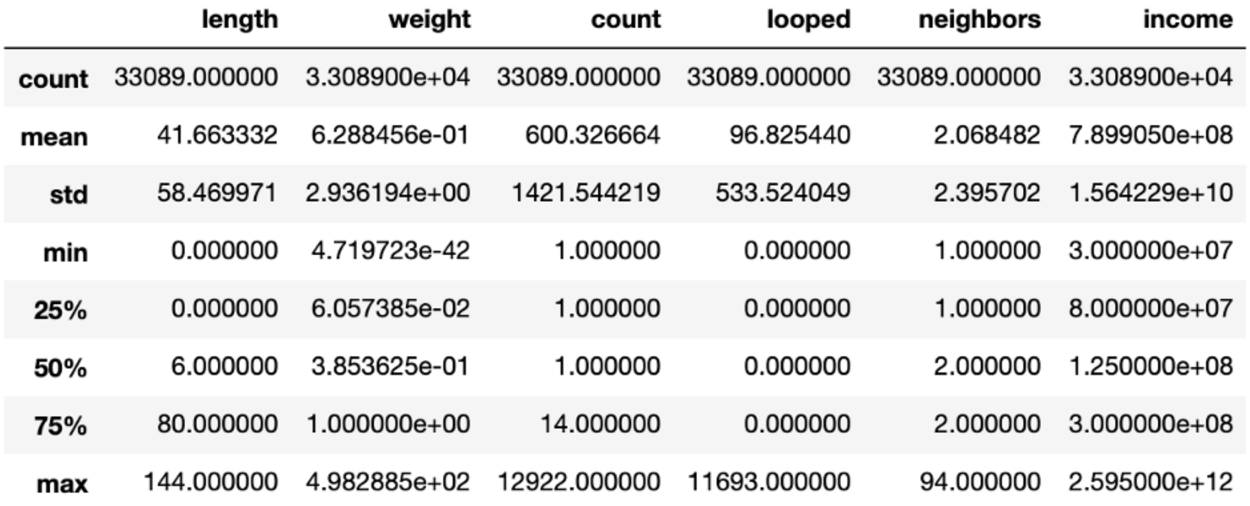
For convenience, we divided the data into white-labeled data and nonwhite-labeled data.

White-labeled data summary:



	length	weight	count	looped	neighbors	income
count	2.300268e+06	2.300268e+06	2.300268e+06	2.300268e+06	2.300268e+06	2.300268e+06
mean	4.509420e+01	5.444412e-01	7.241337e+02	2.407700e+02	2.215242e+00	4.434690e+09
std	5.900971e+01	3.662059e+00	1.693346e+03	9.716880e+02	1.901544e+01	1.530985e+11
min	0.000000e+00	1.420108e-90	1.000000e+00	0.000000e+00	1.000000e+00	3.000000e+07
25%	2.000000e+00	2.115024e-02	1.000000e+00	0.000000e+00	1.000000e+00	7.409540e+07
50%	8.000000e+00	2.500000e-01	1.000000e+00	0.000000e+00	2.000000e+00	2.000000e+08
75%	1.100000e+02	8.750000e-01	5.700000e+01	0.000000e+00	2.000000e+00	1.000000e+09
max	1.440000e+02	1.943749e+03	1.449700e+04	1.449600e+04	1.292000e+04	4.982447e+13

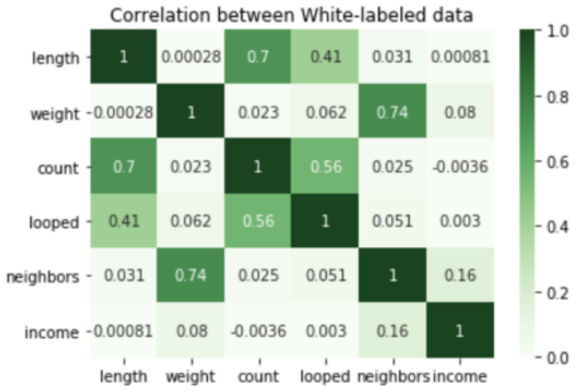
Non-, white-labeled data summary:

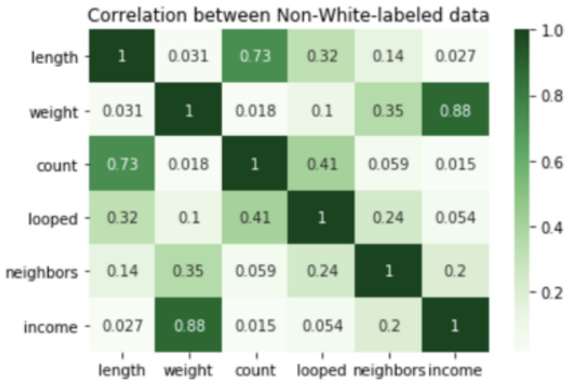


	length	weight	count	looped	neighbors	income
count	33089.000000	3.308900e+04	33089.000000	33089.000000	33089.000000	3.308900e+04
mean	41.663332	6.288456e-01	600.326664	96.825440	2.068482	7.899050e+08
std	58.469971	2.936194e+00	1421.544219	533.524049	2.395702	1.564229e+10
min	0.000000	4.719723e-42	1.000000	0.000000	1.000000	3.000000e+07
25%	0.000000	6.057385e-02	1.000000	0.000000	1.000000	8.000000e+07
50%	6.000000	3.853625e-01	1.000000	0.000000	2.000000	1.250000e+08
75%	80.000000	1.000000e+00	14.000000	0.000000	2.000000	3.000000e+08
max	144.000000	4.982885e+02	12922.000000	11693.000000	94.000000	2.595000e+12

Data Visualizations

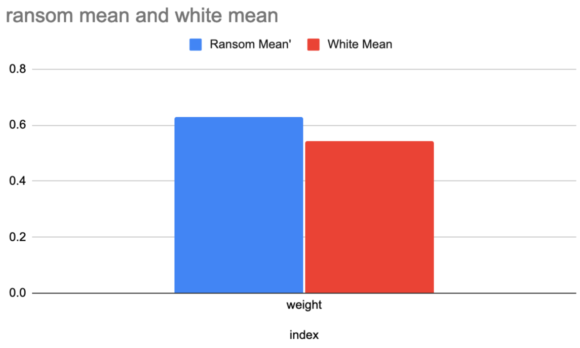
Finding 1:

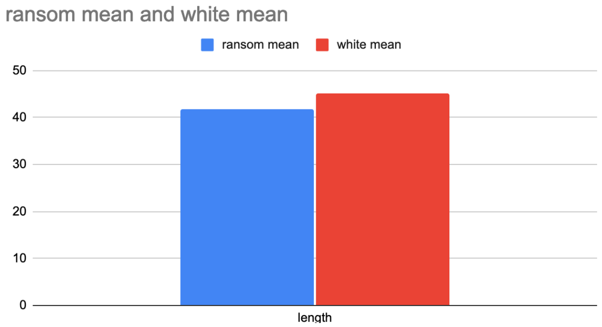


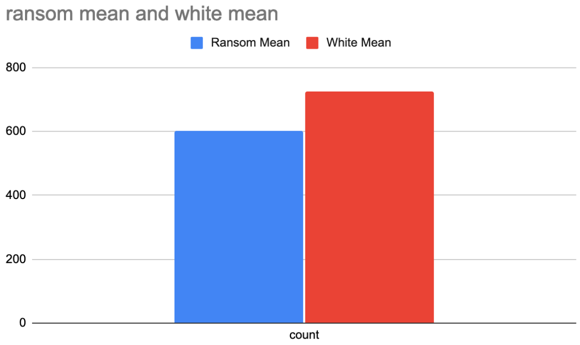


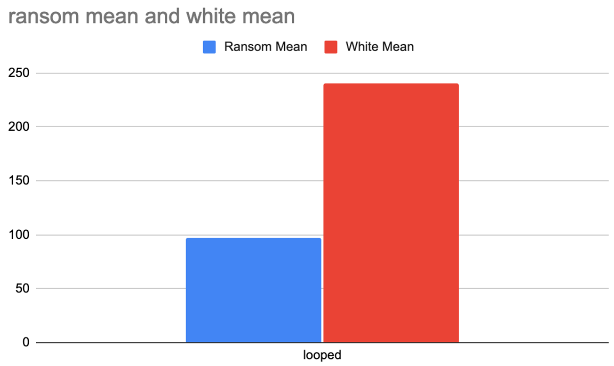
In the heatmaps above, we are showing the correlation matrix for both white-labeled data and non-white-labeled data and we have noticed some significant difference. As the graphs suggest, weight and income have low correlation (0.08) in white-labeled data, but high correlation (0.88) in non-white-labeled data. Also, weight and neighbors have high correlation (0.74) in white-labeled data, but low correlation (0.35) in non-white-labeled data. These might be helpful for later analysis.

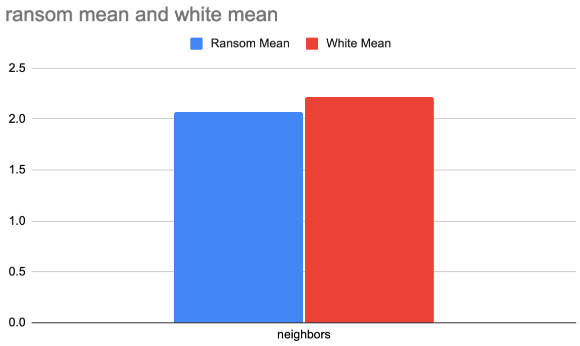
Finding 2:

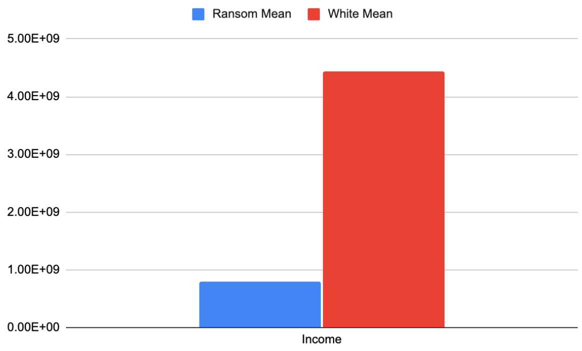






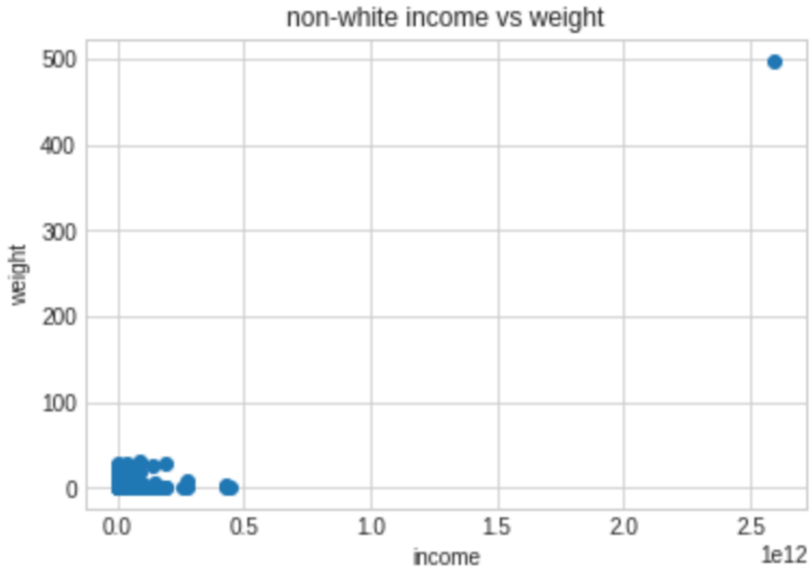






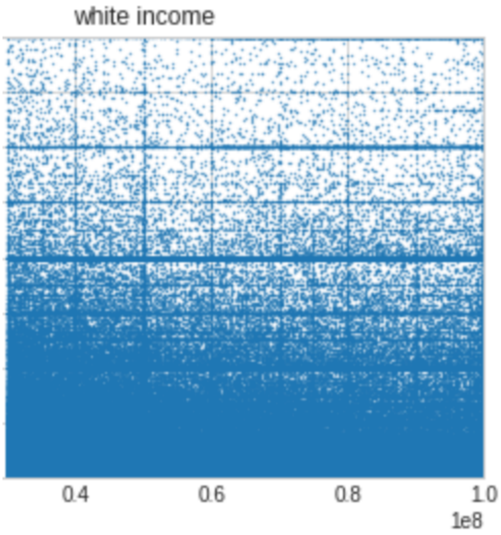
In general, we find that the mean statistics for all statistics that we're keeping track of to be lower than the white label bitcoin transactions. We expect this to be because white label bitcoins change hands more than ransom bitcoin do since ransom bitcoin will end up in an accumulation wallet after sending it through multiple wallets to hide the origin. This is why the weight mean is higher than the white mean while the length mean of the ransom transaction is smaller. The mean income of ransom bitcoin is a very small percentage of all bitcoins traded, about 1.5% so the total mean is significantly lower than the mean of the white label transactions.

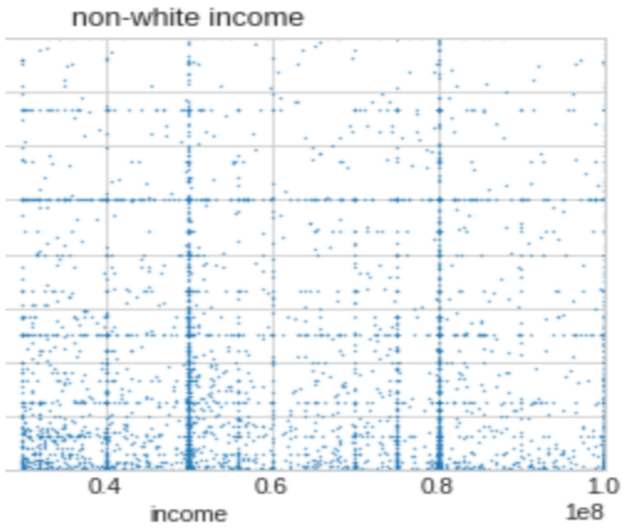
Finding 3:



Previously, we have found that there is high correlation (0.88) between weight and income in the non-white-labeled data. However in the scatterplot above, we have noticed that there is an outlier in the up-right corner, which may influence the correlation.

Finding 4:





For non-white transactions, incomes are likely to concentrate around **0.5** and **0.8.** But this pattern is not observed in white transactions

Analysis

Due to our lack of advanced analysis and hands-on experience, we were not able to find significant patterns and trends in this dataset. This explains why blockchain technology is so good at hiding the origin and purpose of a transaction. Even with ample research into the blockchain technology and background on cryptocurrency heist, it was hard to find correlation between any of the variables tracked in the database. All we could find was a difference in the mean statistics between the white and non-white labels. Standardizing the dataset was to no avail due to the huge dispersion of data.

Machine Learning Model

Due to the similarity between white and non-white transactions, we built a **deep neural network model** with 4 layers and **Batch Normalization.** Our hope is that the deep neural network can figure out some hidden patterns on its own that can't be easily discovered by humans. We also chose the **PCA model** because it's able to tackle the problem of high dimensionality. To our expectation, these two models yielded the best results, achieving **98%** accuracy. We also attempted the **K-Nearest-Neighbor model** with **GridSearchCV**, achieving 92% accuracy. This makes sense to us because KNN isn't good at learning underlying structures. The **Random Forest model** performed the worst with only 33% accuracy.

We expect that none of the models generated here are accurate models for labeling future transactions because we believe the high accuracy is due to over-fitting the dataset. Over 98% of the data are white labels so a model with accuracy around 98% most likely means that it just guessed white for everything and got a 98% accuracy without taking into account other features.

Conclusion

After looking at the dataset, we were unable to find any missing or null values in the dataset as everything was generated by the computer, so there were no human mistakes by negligence. We were unable to find any patterns and trends from the dataset or any significant correlation between any two variables. For machine learning, we applied multiple models, including deep neural networks, K-Nearest-Neighbor model, and Random Forest model, etc. However, all of these models ended up with accuracy in the 95% range, which we suspect to be models overfitted to the training data. This means that these models may not be a predictor for unknown values. We believe that a simpler model may be a better predictor than the traditional models we used in this project. However, due to time constraints, we were not able to develop our own model. Overall, our group thinks that this is a very interesting project. If anyone was able to discover a good model for predicting suspicious transactions, it would dramatically help law enforcement agencies in cracking down ransomware families and tracking down crimes. We look forward to seeing other group’s findings and models for this project.