Detecting Ransomware Addresses on the Bitcoin Blockchain

HarvardX Final Capstone CYO Project

Kaylee Robert Tejeda

10/31/2021

Abstract

Abstracts are hard to write. I could have said less if I had more room to say it.

Contents

Introduction

Definitions and motivations. Try to complete one section per day. Turn this in sometime in the next week.

Data

Cite original paper that data is from. Specifically, describe how each variable is defined.

```
## Loading required package: tidyverse

## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

## Loading required package: caret

## Loading required package: lattice

##
## Attaching package: 'caret'

## The following object is masked from 'package:purrr':
##
##     lift

## Rows: 2916697 Columns: 10

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (2): address, label
## dbl (8): year, day, length, weight, count, looped, neighbors, income

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Goal

What is the goal of this paper? Besides graduating?

Outline of Steps Taken

1)
2)
3) ...

Exploration & Visualization

I need better graphs. I have plenty, but I need them to look better and/or have more labels, etc.

Data Analysis

Preparation

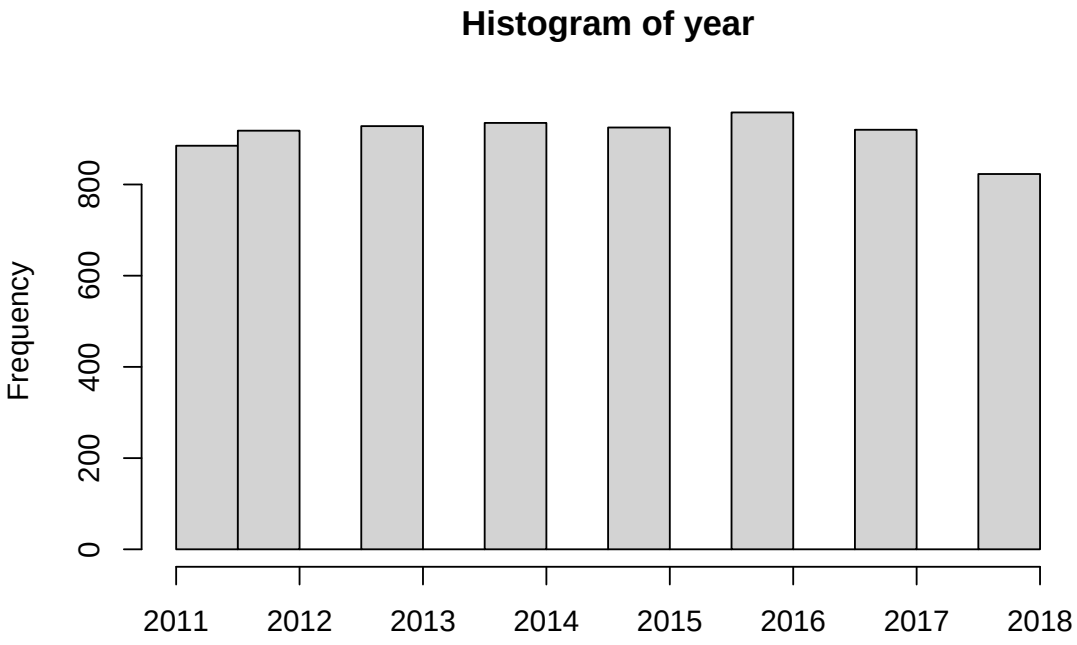
What did I do to prepare the data?

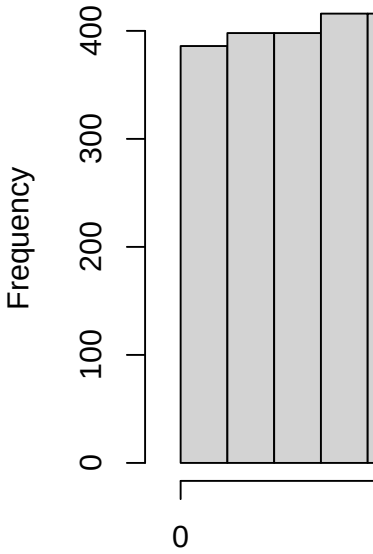
Exploration and Visualization

```
## Loading required package: matrixStats

##
## Attaching package: 'matrixStats'

## The following object is masked from 'package:dplyr':
##
##     count

##  [1] "address"   "year"      "day"       "length"    "weight"    "count"
##  [7] "looped"    "neighbors" "income"    "label"     "grey"

## spec_tbl_df [2,916,697 x 11] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ address  : chr [1:2916697] "111K8kZAEnJg245r2cM6y9zgJGHZtJPy6" "1123pJv8jzeFQaCV4w644pzQJzVWay2zc
##  $ year     : num [1:2916697] 2017 2016 2016 2016 2016 ...
##  $ day      : num [1:2916697] 11 132 246 322 238 96 225 324 298 62 ...
##  $ length   : num [1:2916697] 18 44 0 72 144 144 142 78 144 112 ...
##  $ weight   : num [1:2916697] 0.008333 0.000244 1 0.003906 0.072848 ...
##  $ count    : num [1:2916697] 1 1 1 1 456 ...
##  $ looped   : num [1:2916697] 0 0 0 0 0 0 0 0 0 0 ...
##  $ neighbors: num [1:2916697] 2 1 2 2 1 1 2 2 2 1 ...
##  $ income   : num [1:2916697] 1.00e+08 1.00e+08 2.00e+08 7.12e+07 2.00e+08 ...
##  $ label    : Factor w/ 29 levels "montrealAPT",..: 27 28 27 27 28 28 27 27 27 28 ...
##  $ grey     : Factor w/ 2 levels "black","white": 1 1 1 1 1 1 1 1 1 1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   address = col_character(),
##   ..   year = col_double(),
##   ..   day = col_double(),
##   ..   length = col_double(),
##   ..   weight = col_double(),
##   ..   count = col_double(),
##   ..   looped = col_double(),
```

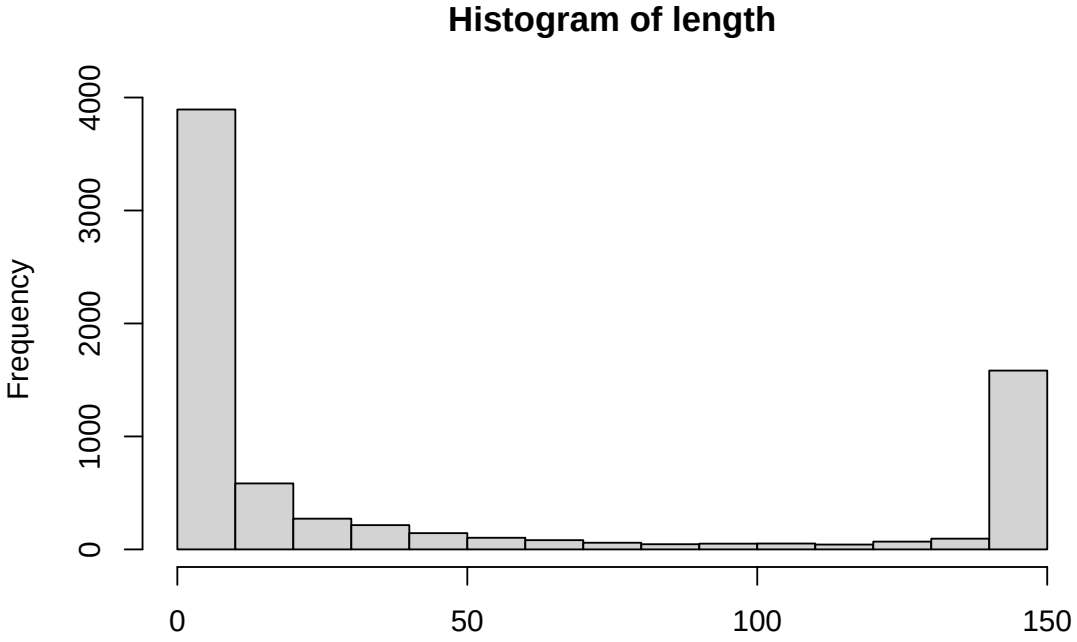
```
##   ..   neighbors = col_double(),
##   ..   income = col_double(),
##   ..   label = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

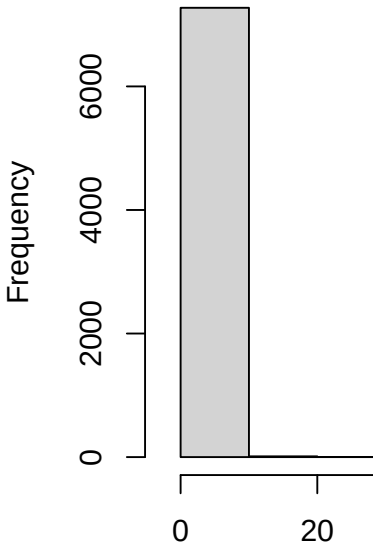
```
## [1] 0.01426221
```



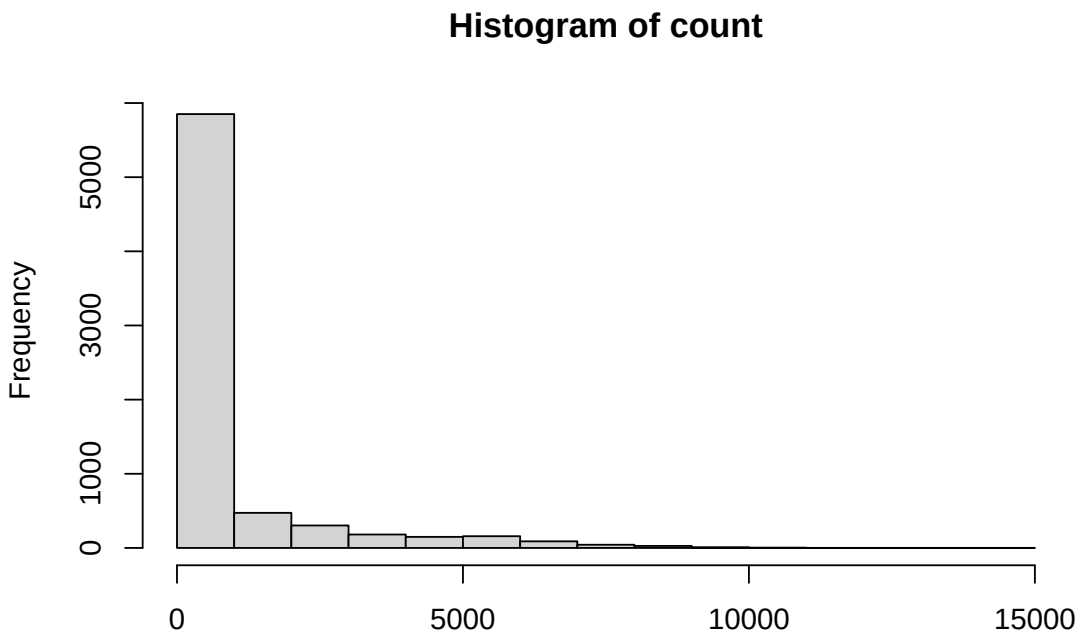


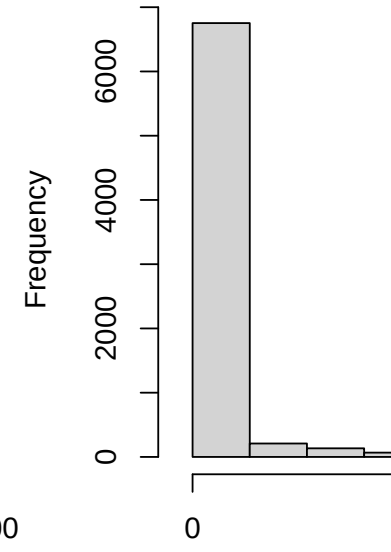
.

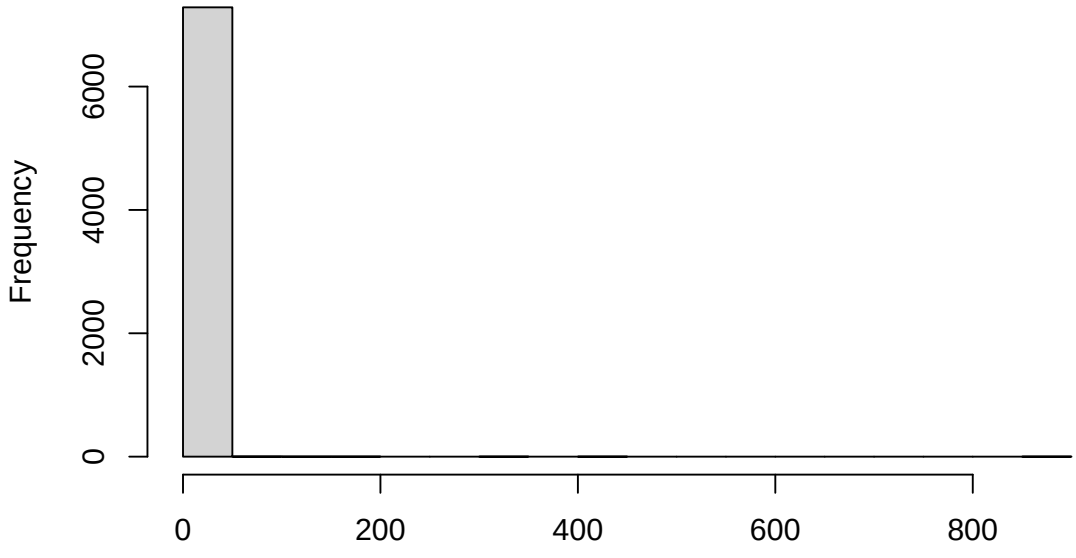


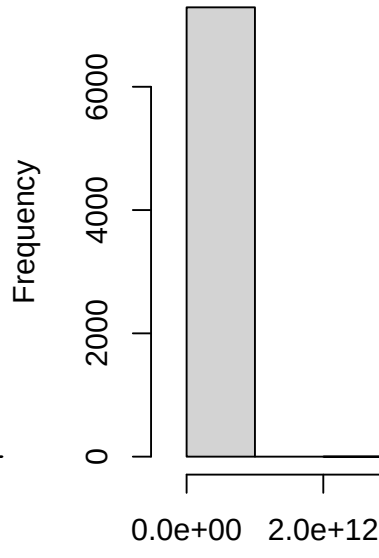


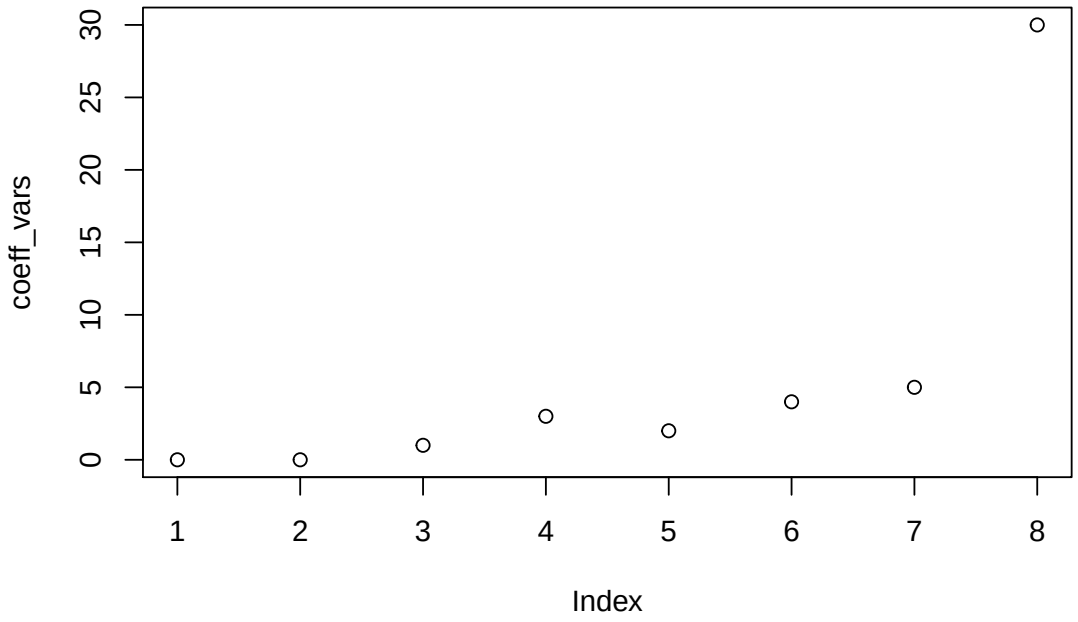
.


Histogram of count
Frequency
0
1000
3000
5000
0
5000
10000
15000


Frequency
0
2000
4000
6000
0


Histogram of neighbors
Frequency
0
2000
4000
6000
0
200
400
600
800


Frequency
0
2000
4000
6000
0.0e+00
2.0e+12

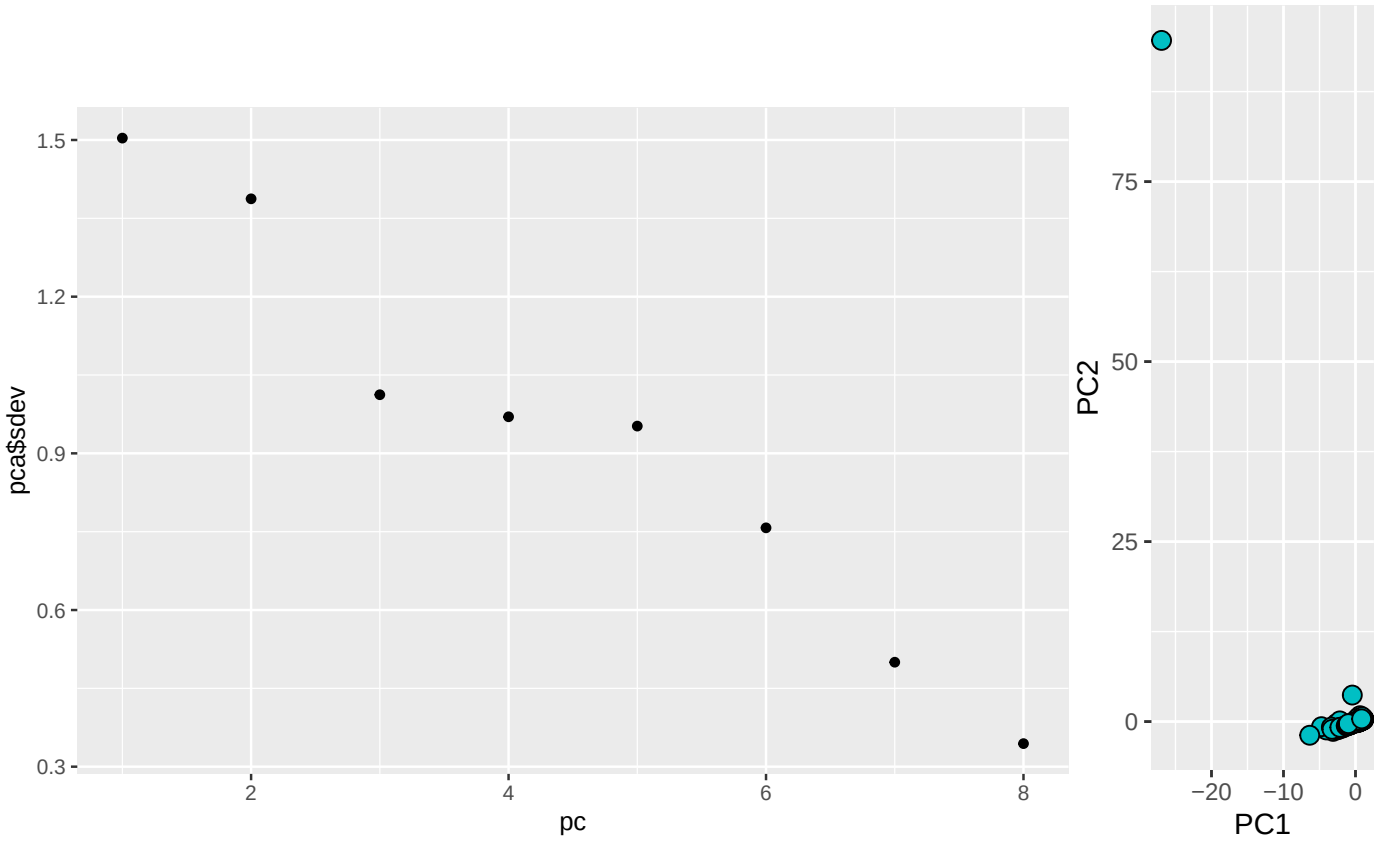



```
##      year       day    length    weight     count    looped neighbors    income
##         0         0         1         3         2         4         5        30
```

```
## Standard deviations (1, .., p=8):
## [1] 1.5036299 1.3873661 1.0122377 0.9700048 0.9521794 0.7573684 0.5000553
## [8] 0.3441947
##
## Rotation (n x k) = (8 x 8):
##                   PC1          PC2           PC3         PC4         PC5
## year      -0.23580472 -0.108225866  0.4048339909 -0.28922012 -0.80803821
## day       -0.01534363  0.002087438 -0.8886984955  0.05059436 -0.45467926
## length    -0.51977190 -0.199569330 -0.0338858389  0.23872673  0.08260547
## weight    -0.19451510  0.366114843 -0.1374452268 -0.77282334  0.18035054
## count     -0.57696908 -0.201095979 -0.0001687107  0.09481824  0.01614863
## looped    -0.48386150 -0.107019035 -0.1173165200 -0.07536481  0.28205469
## neighbors -0.22499302  0.652784772  0.0194992349  0.03670568 -0.03083644
## income    -0.13610522  0.579994986  0.1101887875  0.49348548 -0.14221297
##                   PC6          PC7          PC8
## year       0.12542996 -0.128086724 -0.008934577
## day        0.02574126 -0.002799138 -0.003425822
## length    -0.54406462 -0.570113970 -0.063329595
## weight    -0.23123693  0.019909028 -0.354377613
## count     -0.14434794  0.770896314  0.047940942
## looped     0.77268180 -0.241460002 -0.009328160
## neighbors -0.04098522 -0.054316519  0.718290816
## income     0.12072534  0.051237722 -0.593285144
```

```
## Importance of components:
##                             PC1    PC2    PC3    PC4    PC5    PC6     PC7
## Standard deviation     1.5036 1.3874 1.0122 0.9700 0.9522 0.7574 0.50006
## Proportion of Variance 0.2826 0.2406 0.1281 0.1176 0.1133 0.0717 0.03126
## Cumulative Proportion  0.2826 0.5232 0.6513 0.7689 0.8822 0.9539 0.98519
##                            PC8
```

```
## Standard deviation     0.34419
## Proportion of Variance 0.01481
## Cumulative Proportion  1.00000
```




```
## Warning in rm(pca, x, coeff_vars, d, means, pc, sds): object 'x' not found

## Warning in rm(pca, x, coeff_vars, d, means, pc, sds): object 'd' not found
```

Insights Gained from Exploration

Modeling approach

An overview of why I picked the methods that I did. Based on suggestions from original paper, that Random Forests were hard to apply here, and that it was all topological data to begin with, hence that lead me to SOMs. Also, describe the reasoning behind the binary approach. Describe what you learned about SOMs.

Random Forests

Self Organizing Maps

Method 1: Binary Random Forests

If we ask a simpler question, is this a useful approach?

Method 2: Binary SOMs

If we ask the same question to a more sophisticated and ### Method 3: Categorical SOMs

Final Method: Combined Methods 1 and 3

Results & Performance

Results

Performance

In terms of what? Time? RAM?

Summary

Comparison to original paper and impact of findings

Limitations

Future Work

Conclusions

Get Monero!