Detecting Fraud in BTC Blockchain
Sprint 2

Yael Brown

Overview

- I explained plan for acquiring data, merge data, and did preliminary analysis on all the data sets I was planning on working on.

- Settled with using 2 datasets; heist (ransomware) and transaction data (from Harvard)

- Collected data, filtered the data, then analyzing the sample subset.
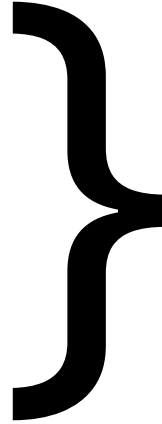
Data Collection

290 files
~2 Gb per file
390 Gb total

Date range: 2009-2020

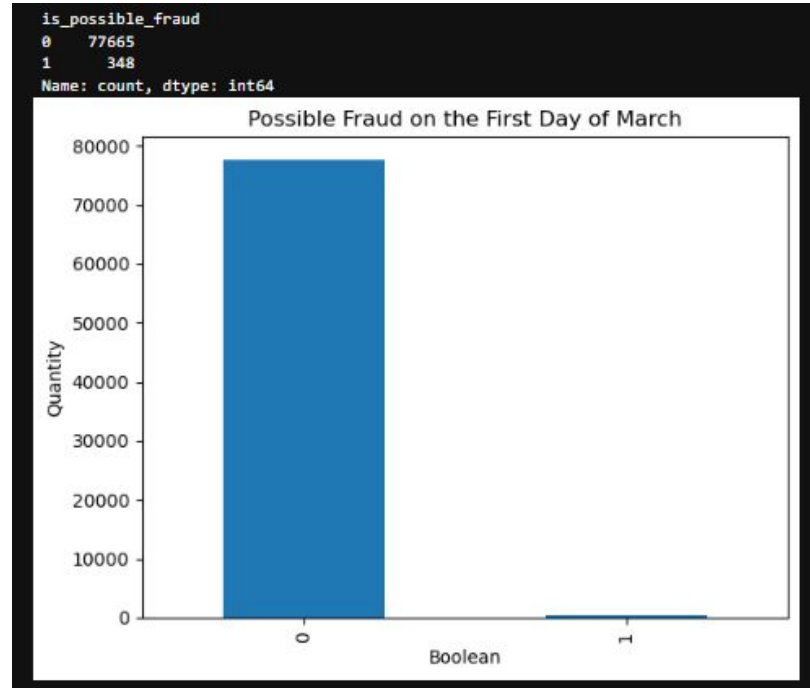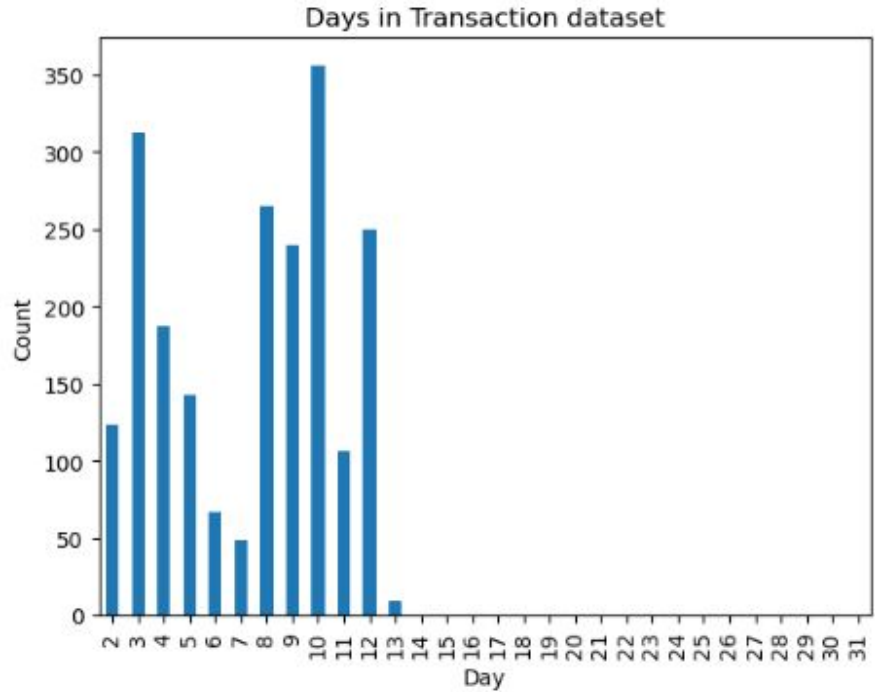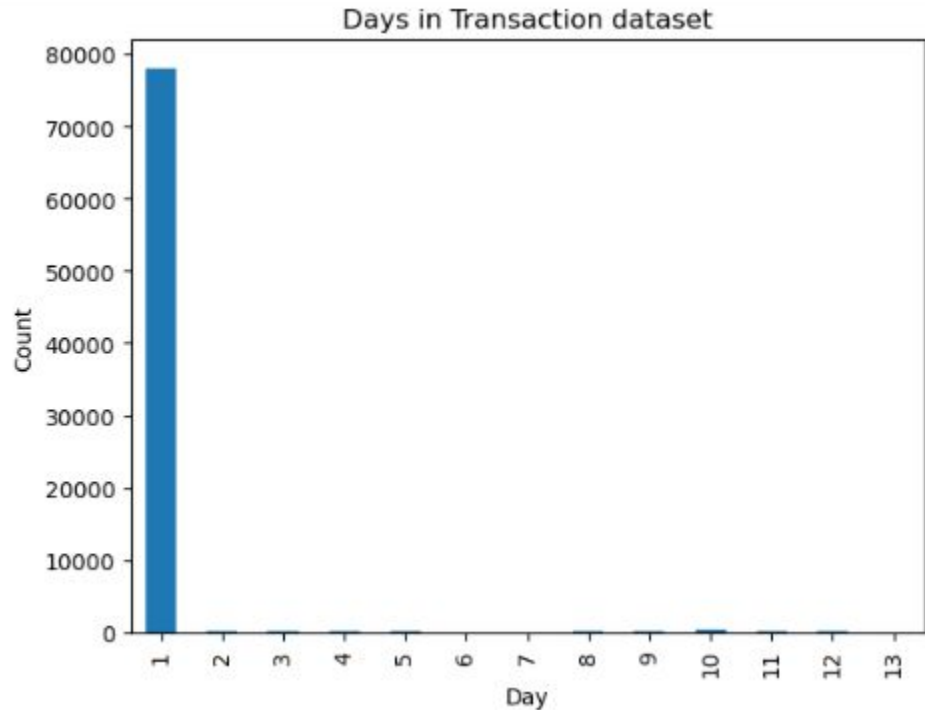Heist dataset is from 2016-2018.

Filter Data

Btc_filtered_data = {
    btc_tx_data[date] >= March 1, 2016 ∧
    Btc_tx_data[date] <= March 31, 2016
    }

43 Files from 290

Check for poss. Fraud Transactions.
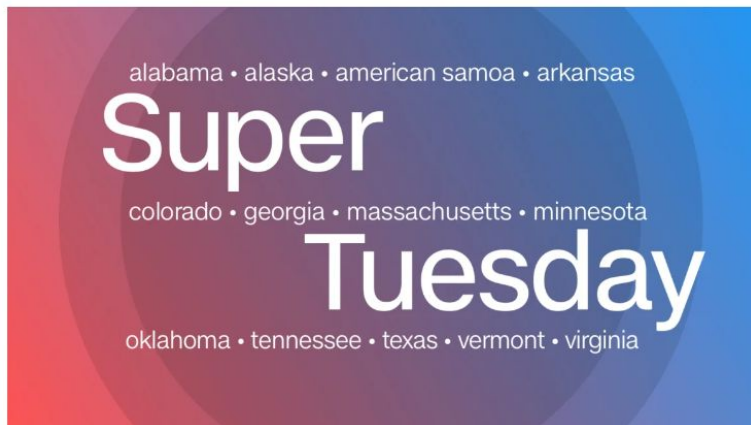
Checking Dates of Poss. Fraud.

What happened March 1, 2016 (cont)

What happened March 1, 2016

What happened March 1, 2016 (cont)
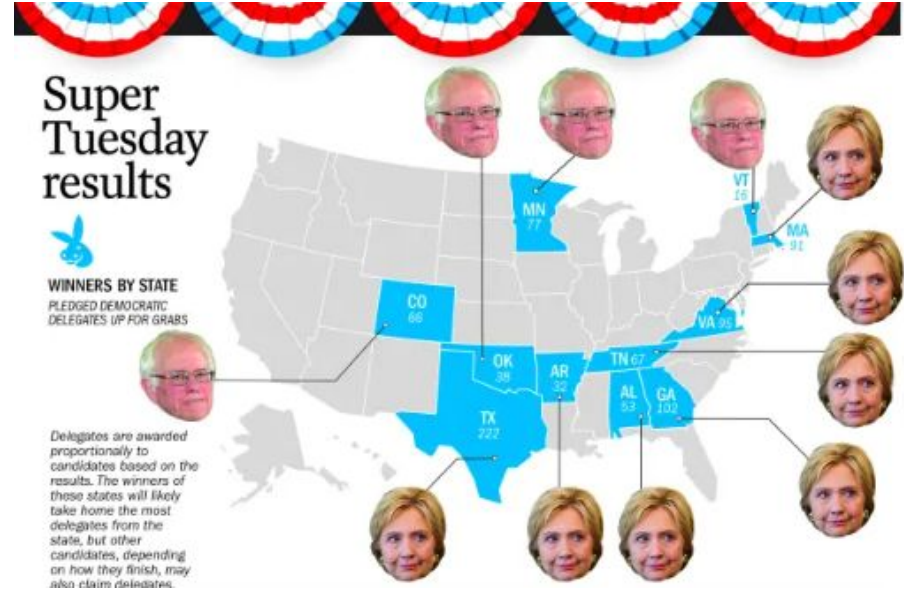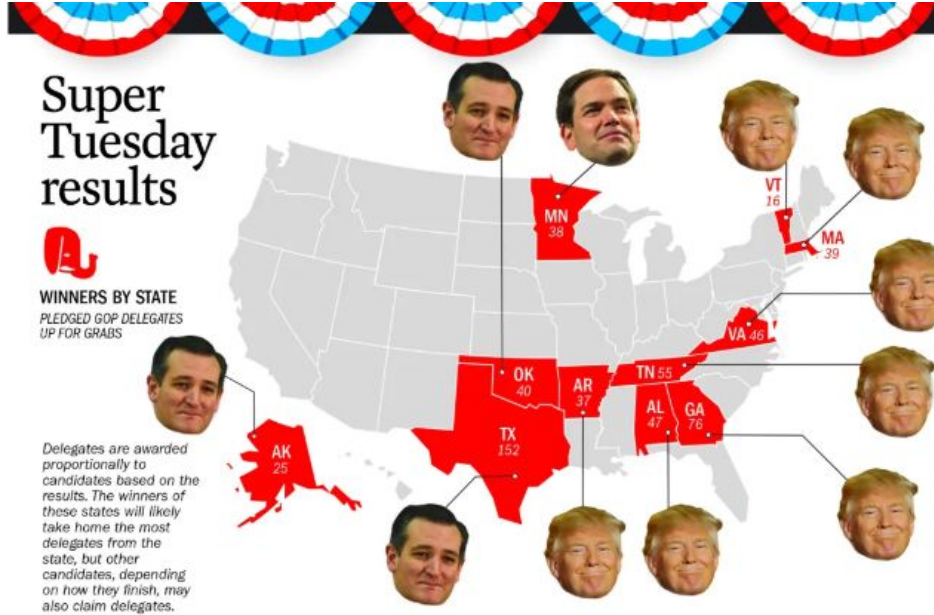
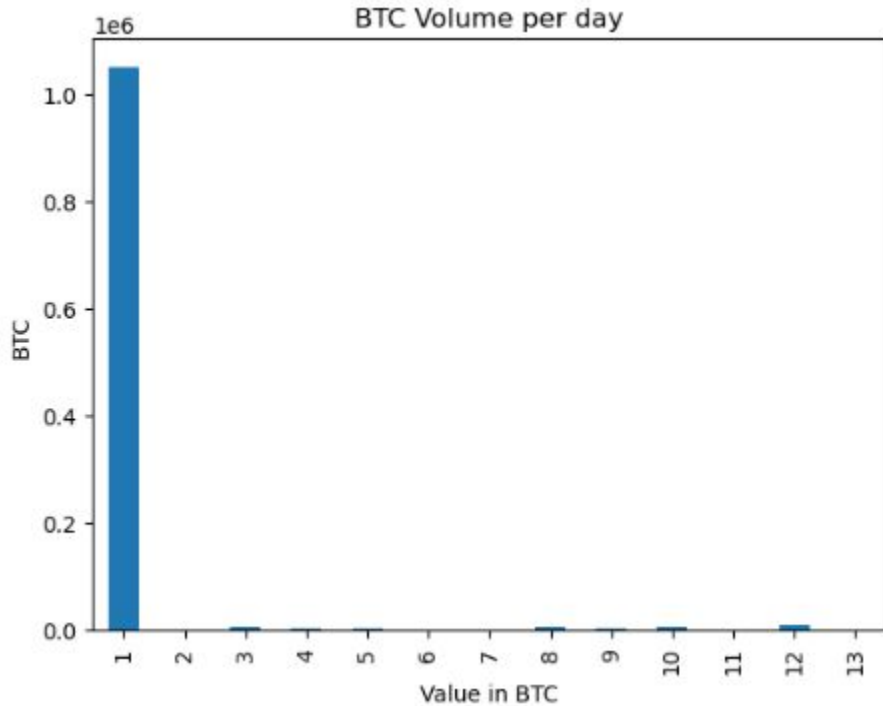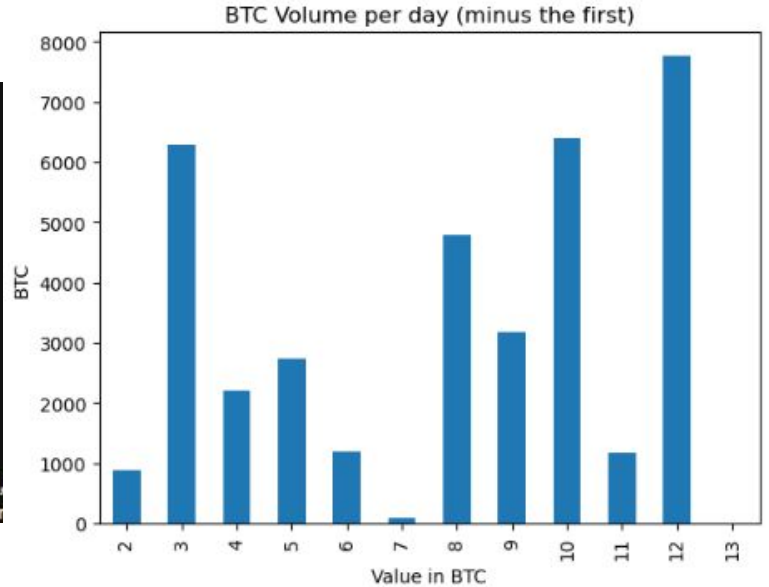# Explore Volume in BTC

BTC Volume per day

day	
1	1050048
2	892
3	6292
4	2202
5	2726
6	1194
7	93
8	4785
9	3168
10	6387
11	1160
12	7763
13	2

Name: satoshi
Total btc's i

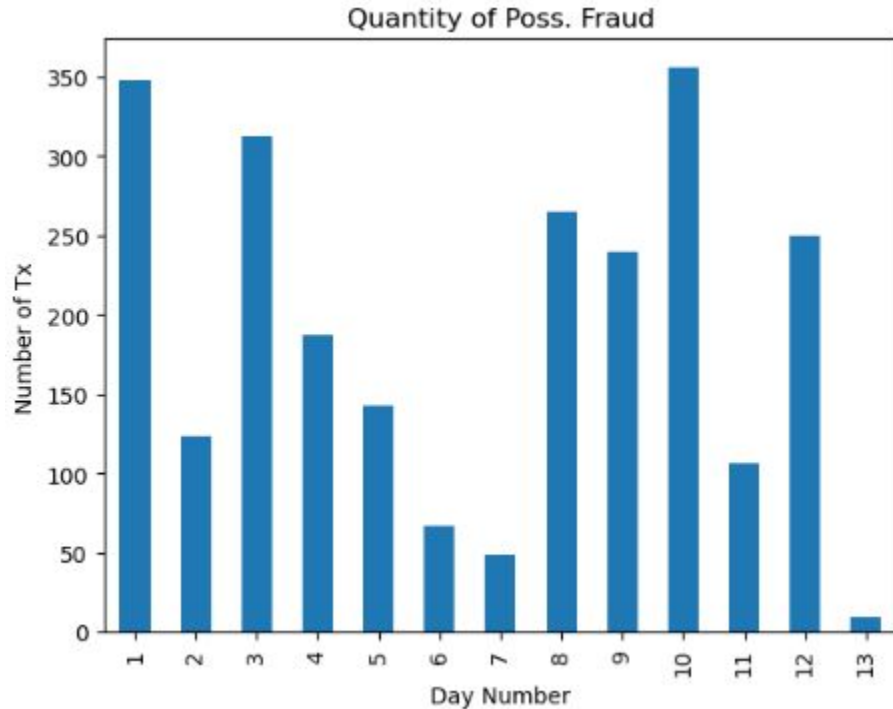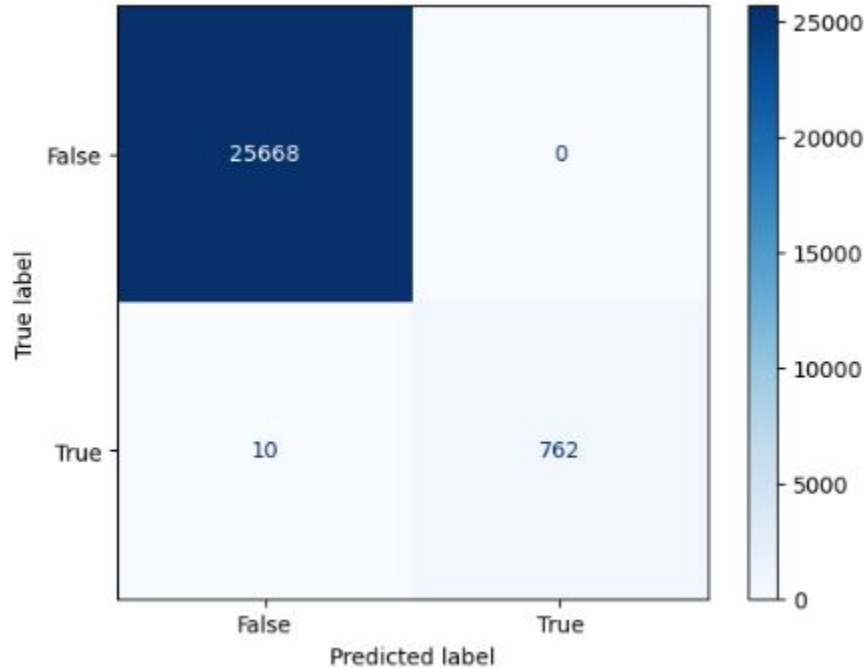BTC Volume per day (minus the first)

Quantity and Volume of Fraud
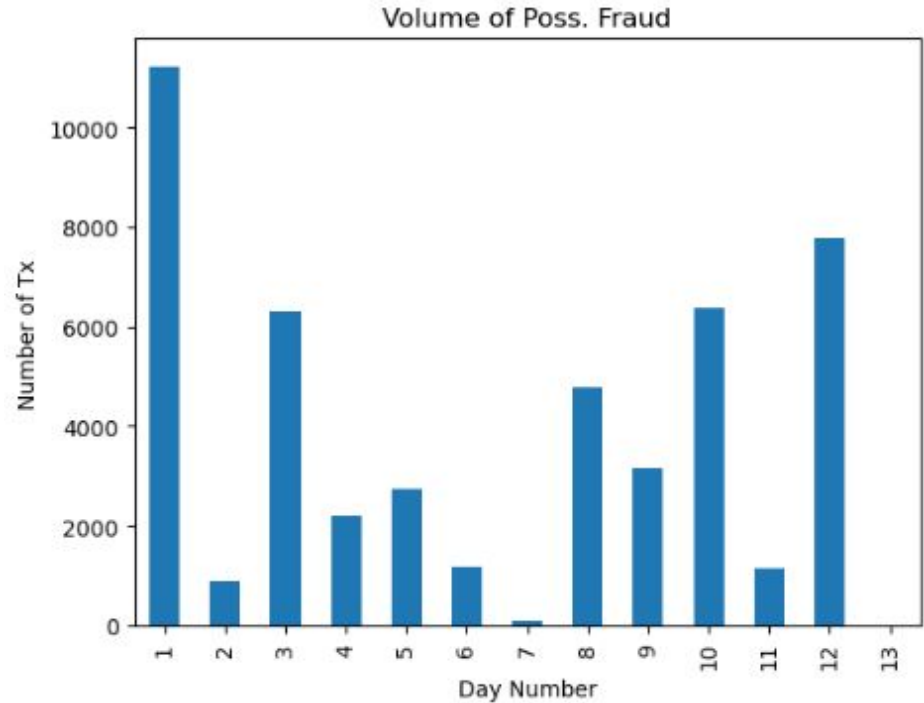
Preliminary Modeling

Basic Logistic Regression model

Y target variable is "is_possible_fraud"

Accuracy: 0.9996217851739788

May need to do multi-label classification.

Goal is to classify by both sender and receiver addresses, and value.

End