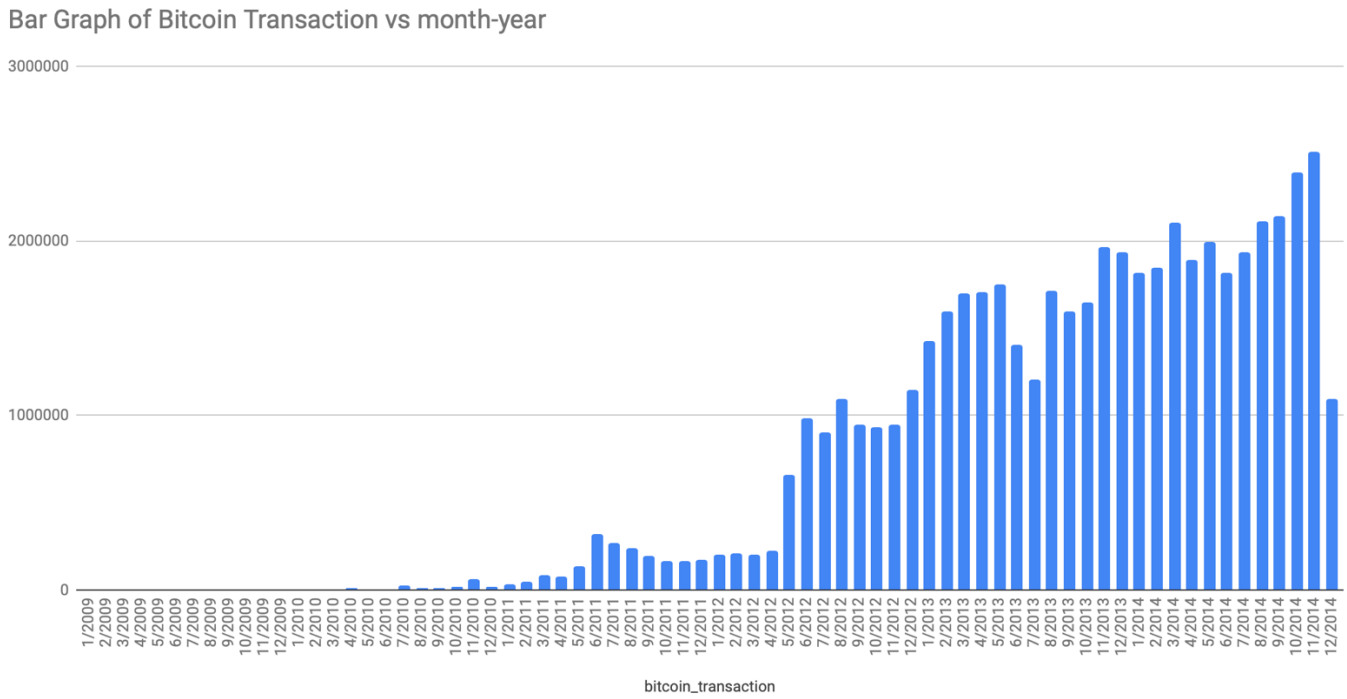
COURSEWORK 1: ANALYSIS OF BITCOIN TRANSACTIONS

PART A. TIME ANALYSIS (30%)

Create a bar plot showing the number of transactions which occurred every month between the start and end of the dataset. What do you notice about the overall trend in the utilisation of bitcoin?

Code Snippet: For this task we will use Hadoop cluster Map Reduce. We read in our textfile of all transactions, which at this point only contains lines i.e. transactions.csv . We will use MRjob for Map reduce. We will then compute mapper where first we will split data by comma delimeter. Afterwards we use a filter, to check if the data is in correct format or not, since it can always happen that there are errors in the data. Then using the tx_time which is in unixtime, the 3rd feature of the file of each transaction we extract that as month and year format. Then setting a filter of year 2009 to 2014 we will make a key of(year,month) and value (1)and yield it. Then these data is used in combiner and reducer to sum of values on the basis of key. It will gives us the combined transactions by adding them up as taking unique key as year and month. Then at last we can save our file in our HDFS system.



If we see the bar graph above the Bitcoin existence was started in January 2009. If we see the whole timeline frame from 2009-2014 as follows:

Jan 2009 – Mar 2010 : Bitcoin first transaction was started in Jan 2009 as we can see above . Till March 2010 there is very minimal transaction or we can say no transaction at that time the price of Bitcoin was also almost none. So, only the person who have hobbies in cryptocurrency used to do transactions.

Mar 2010 – May 2010 : Bitcoin transaction just got double from the past data , as in result the Bitcoin gained some value in real world with less than ~0.01 $.

May 2010 – July 2010 : In May and June the transaction was minimal but suddenly the transaction got four times from previous month which was ~6678 to 26,488 . It leads to increase in the price of Bitcoin as well to 1000% which was before ~ $ 0.01 to ~ $ 0.08.

July 2010 – Dec 2010 : July onwards till Oct 2010 the transaction was steady , but in Nov it got raised double than in July to ~63,000. And than again the transaction got decreased to ~ 17,000. So not much price fluctuation in this time frame.

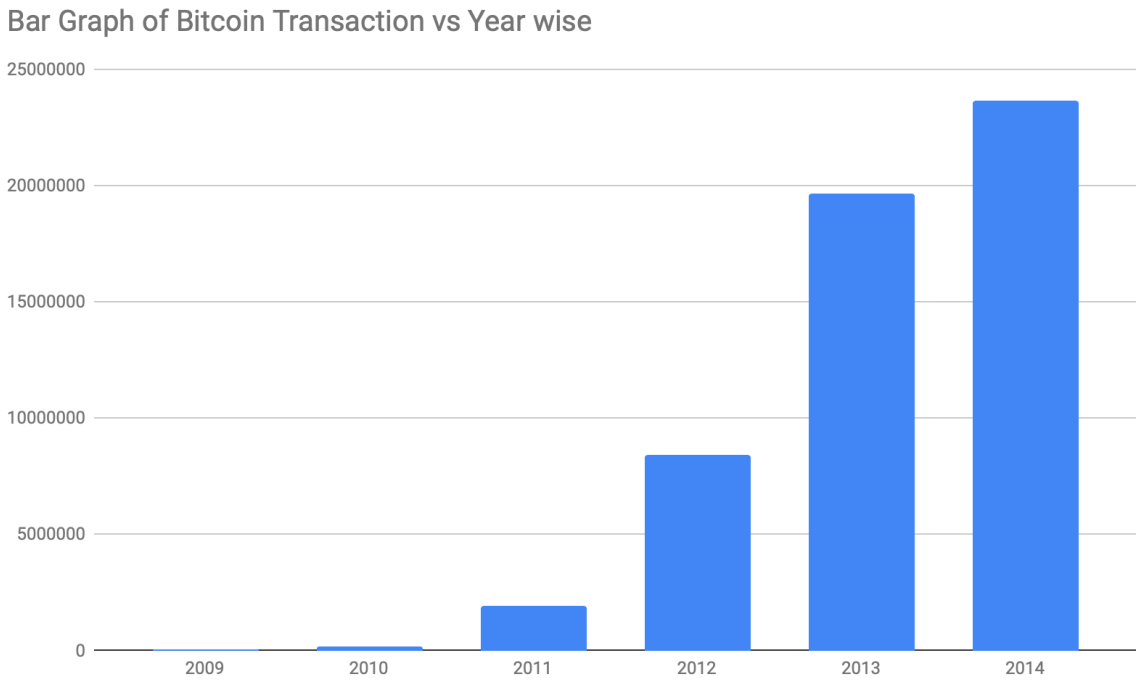
Jan 2011 – April 2011 : The transaction from Jan 2011 gradually and steadily increasing from ~35,000 to 47,168 in Feb to ~83,222 in Mar 2011. So if we compare the transaction of Mar 2010 to Mar 2011 the transaction got almost 16 times . Now, the bitcoin started gaining some popularity. The price of Bitcoin also rose to $1 in these months.

May 2011 – April 2012 : The transaction gradually increased and get doubled in this time frame from past of 80,000 to mean of 200,000 transactions per month. People started to recognise Bitcoin in this time frame and started using them in US.

May 2012 – Dec 2012 : In May the transaction got triple from before and it get rise in the next month in the same rate till Dec 2012. During the months of May and
June of 2012, the bit coin market witnessed a high spike which crossed the threshold of 10 million.

Feb 2013 – Dec 2013 : The transaction got triple almost from last period and it was rising exponentially in this period but lost it's transactions in May 2013 , but get come back again in next month rose until 2013. This exponential increase in transaction of Bitcoin by peoples all over the world resulted to increase it's price from $13 from Feb 2013 to ~$800.

Jan 2014 – Dec 2014 : This year transaction were going down from last year 2013 as many hackers were stealing Bitcoins in the 2013 year. So it led to decrease in use of Bitcoin so the transaction got decreased in Jan to 1,817,513 compared to 1,935,103 but it started to rise again but the transaction again got fall to minimum in Jun 2014, than it rose to maximum to double again almost in Nov to 2,512,559 but the downfall was near and it fell to 1,094,574 to almost the transaction of 2012.



Overall if we summarize the Bitcoin graph the transactions of Bitcoins increases exponentially from the start of the year in 2009 to 2014. As I have included the year vs Bitcoin transaction graph as well.

The code and output file for this Task is linked in the folder

- BitCoinTaskA.py
- Result.txt

PART B. TOP TEN DONORS (40%)

Obtain the top 10 donors over the whole dataset for the Wikileaks bitcoin address: {1HB5XMLmzFVj8ALj6mfBsbifRoD4miY36v}. Is there any information on who these wallets belong to? Can you work out how much was being donated if converted into pounds?

Code Snippet : First SparkContext is generated, it allows our application to access the Spark cluster. For this task we need to explore in 2 datasets

- vout.csv (destination wallets) , and
- vin.csv (source transactions of the coins).

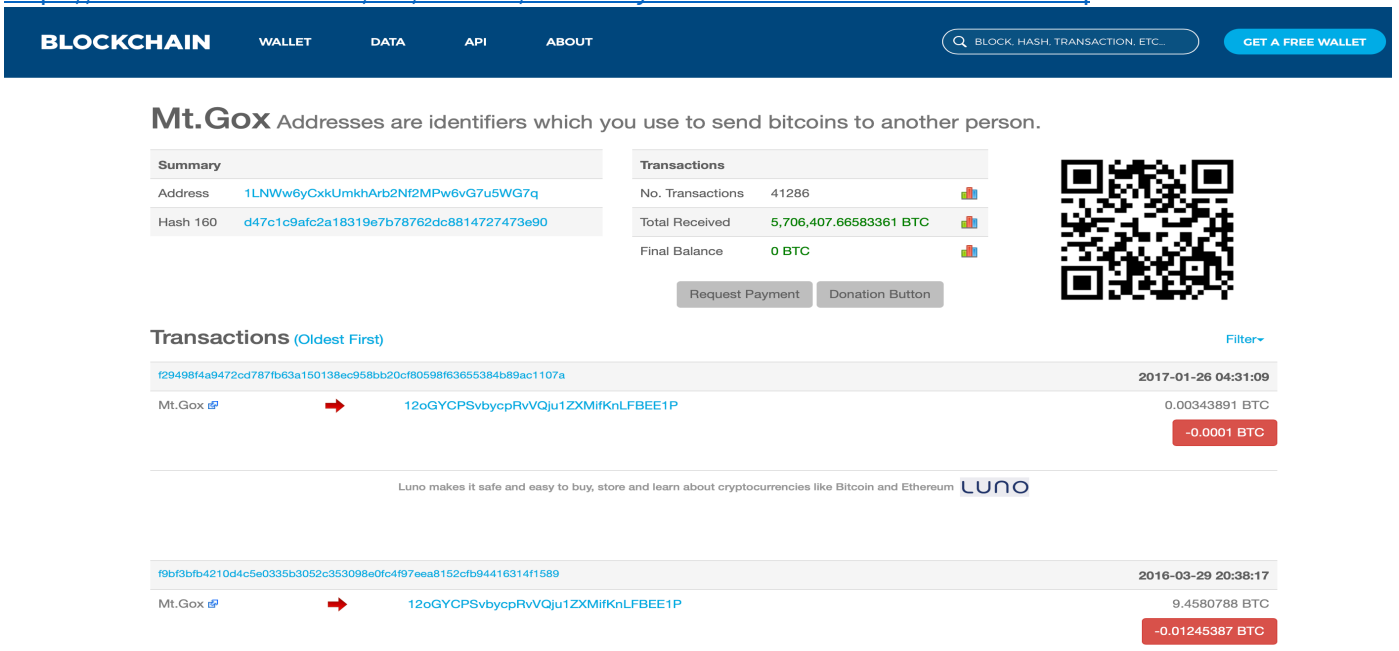
To achieve this task I used dataframes of pyspark. First we need to filter the data from whole vout files which are based on wikileaks bitcoin address: {1HB5XMLmzFVj8ALj6mfBsbifRoD4miY36v}. After getting the filtered data of wikileaks address we made our first join of this filtered data with vin using **txid with hash**. Now we got the transactions that has been involved with wikileaks donation but still don't know which wallet is donating to wikileaks. For that we have to again join this data with the whole vout.csv file using tx_hash with hash and n with vout of vout.csv file with first join data obtained. Now we have the data with wallet which were donating to the wikileaks but the data is still spreaded maybe 1 wallet is used to do multiple time of donation to the wikileaks. To aggregate the amount donated by the wallet to the wikileak we will use groupBy of public keys or wallet address and sum the values. And at last we can just sort the result by descending.
This result we can save it in file in HDFS or show it in pyspark shell.

List of Top 10 Donors: For the confirmation of wallet the check has been done in https://www.blockchain.com with the transactions that this wallet is donating actually to wikileak or not and how much they have donated :

(To convert BTC to GBP https://www.coingecko.com/en/price_charts/bitcoin/gbp)

Rank	Wallet Address	Amount Donated (BitCoin)	Amount Donated GBP
1	{17B6mtZr14VnCKaHkvzqpkuxMYKTvezDcp}	46515.1894803	146441445.2813545
2	{19TCgtx62HQmaaGy8WNhLvoLXLr7LvaDYn}	5770.0	18165402.5
3	{14dQGpcUhejZ6QhAQ9UGVh7an78xoDnfap}	1931.482	6080788.2065
4	**{1LNWw6yCxkUmkhArb2Nf2MPw6vG7u5WG7q}**	**1894.37418624**	**5963963.53183008**
5	{1L8MdMLrgkCQJ1htiGRAcP11eJs662pYSS}	806.13402728	2537911.4513842603
6	{1ECHwzKtRebkymjSnRKLqhQPkHCdDn6NeK}	648.5199788	2041703.0232571
7	{18pcznb96bbVE1mR7Di3hK7oWKsA1fDqhJ}	637.04365574	2004986.609020174
8	{19eXS2pE5f1yBggdwhPjauqCjS8YQCmnXa}	576.835	1815490.10055
9	{1B9q5KG69tzjhqq3WSz3H7PAxDVTAwNdbV}	556.7	1752118.611
10	{1AUGSxE5e8yPPLGd7BM2aUxfzbokT6ZYSq}	500.0	1573665

I found 1 wallet owner that belongs to **Mt. Gox**
that's address:{1LNWw6yCxkUmkhArb2Nf2MPw6vG7u5WG7q} highlighted in the table. Using
https://www.blockchain.com/btc/address/1LNWw6yCxkUmkhArb2Nf2MPw6vG7u5WG7q



This wallet has donated to wikileaks till 2014 is 1894.37418624 BitCoins.
Rest wallet users are anonymous. I tried to backtrack them in this website but unable to do that.

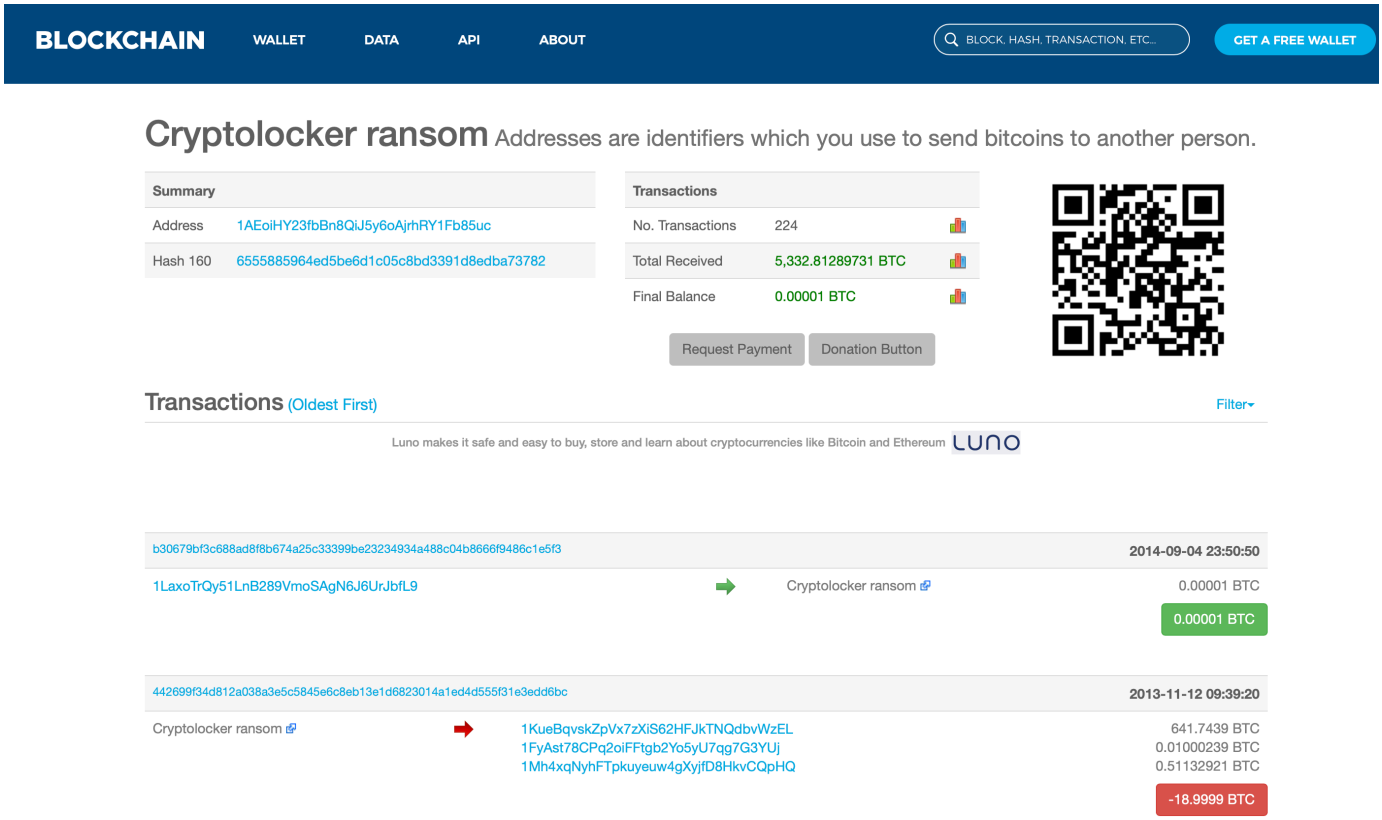
For this my code and output file is in the linked folder

- partBDataFrame.py
- resultBTop10.txt

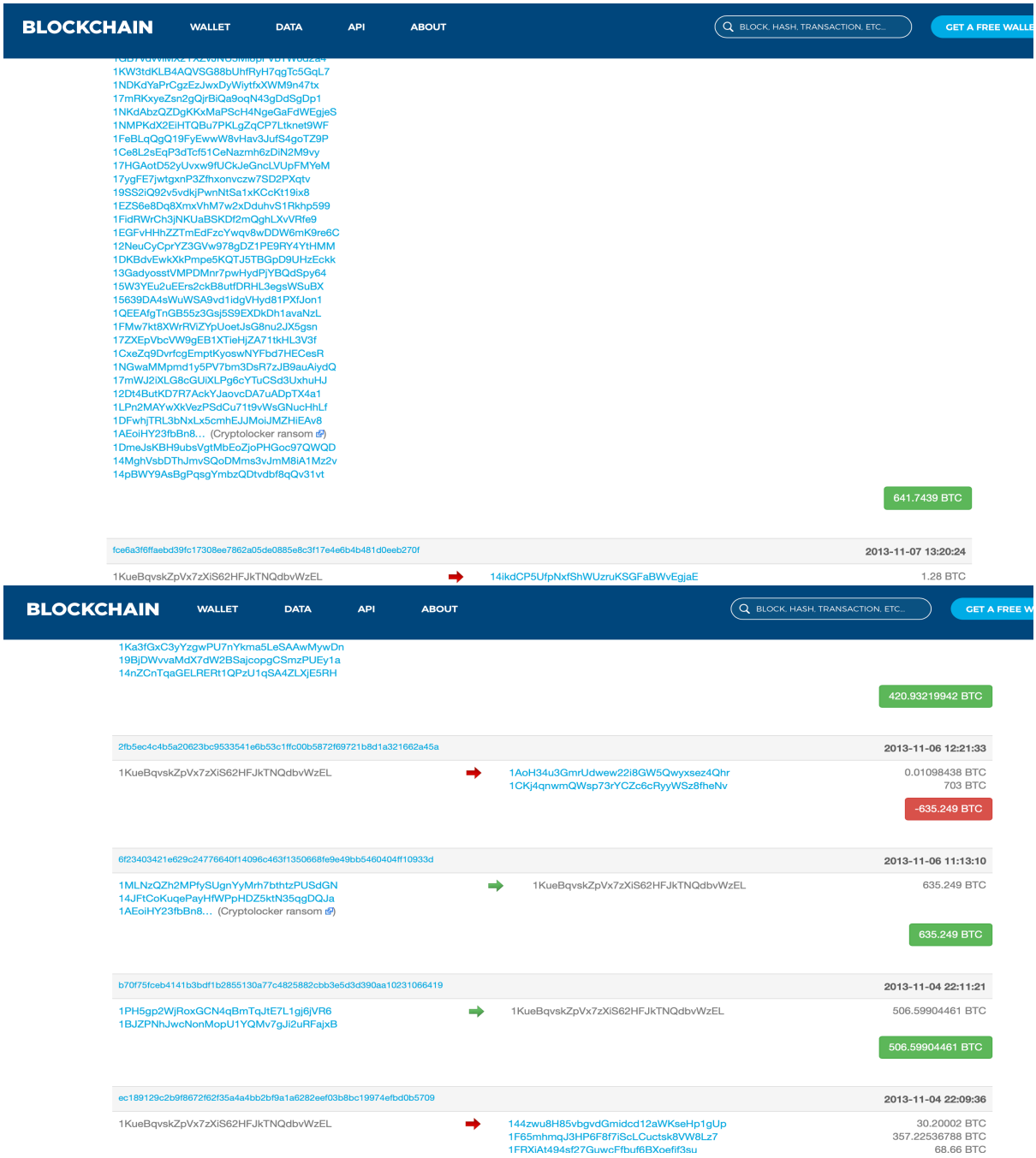
PART C. DATA EXPLORATION (30%)

- Ransomware often gets victims to pay via bitcoin. Find wallet IDs involved in such attacks and investigate how much money has been extorted. What happens to the coins afterwards? [25 Marks]

Code Snippet : To find the ransomware we need to again load the SparkContext first. Here to find the ransomware wallet and how much victims has paid to them in total. I have researched many wallets maybe they are involved in ransomware. Using my research in many websites like https://www.bitcoinabuse.com and other websites I made a list of ransomware and cryptolocker wallets which were involved in the year between 2009-2014 and also latest , but these data are not there in data provided to us.
And we will be needing here transactions.csv and vout.csv
First we will join the transactions.csv and vout.csv with tx_hash and hash to get all the transactions with time stamp and transaction details with wallet address.
Than we will use the list created and join the wallet and publickey of the first join. Then we will receive a list of transactions and bitcoin values associated with these wallets from the whole database. We still cannot decide anything , but after this join if we groupBy wallet address and sum the values, and order them descending. We will receive these addresses with amount of bitcoins transferred. It help us to find the ransomware wallet final amount of transaction value.
We have checked the wallet address : {1AEoiHY23fbBn8QiJ5y6oAjrhRY1Fb85uc}
This wallet is involved in cryptolocker ransom, as confirmed here :
https://www.blockchain.com/btc/address/1AEoiHY23fbBn8QiJ5y6oAjrhRY1Fb85uc
It total received 5,332.81289731 BitCoin
Or we can say $ 21,318,452.84 amount it extorted from the victim.



Than I counted the transactions done per wallet as well using the aggregate formula of groupby wallet address and count the values. For this wallet than I found 115 transactions have done than it means **115** victims have been affected and paid to this wallet. Than in the website the total transactions shown is **224.** If 115 is in transactions than 224 – 115 = 109 transactions are going out. So it means the wallet holder has sent to 109 wallets.
I tried to backtrack in the website what happen after victim pays to this wallet. I found that as soon as wallet holder receives the coins he split the coins in random number and send to more than 1 wallet. Than I tried to trace that track by tracking the BTC sent to wallet : 1KueBqvskZpVx7zXiS62HFJkTNQdbvWzEL
. But I got into infinite loop(maybe it is ending somewhere as hackers are too sharp they use onion peel wallet transactions to cover up there tracks). As the same process is going on again and again. As soon as wallet holder or hacker receives the coin it pass it on to another wallet with random bitcoins. But I have noticed that the hacker uses the same wallet for some transactions .



Like he is sending the BTC to the same address after receiving by randoming to other wallets as well. Here in this address he has sent many times after receiving.
So for the conclusion what happens to that bitcoin is hacker will mine that bitcoin after how many rounds we don't know maybe he is using online also , we can't track that as of now with this information and time frame. But if we track these traces maybe we can catch the hacker .

For this the file linked are :

- partCRansomware.py
- transaction.txt/csv
- walletSum.txt/csv
- walletTransactionsCount.txt/csv

Note : I submitted the code before the extended deadline on 3rd Dec