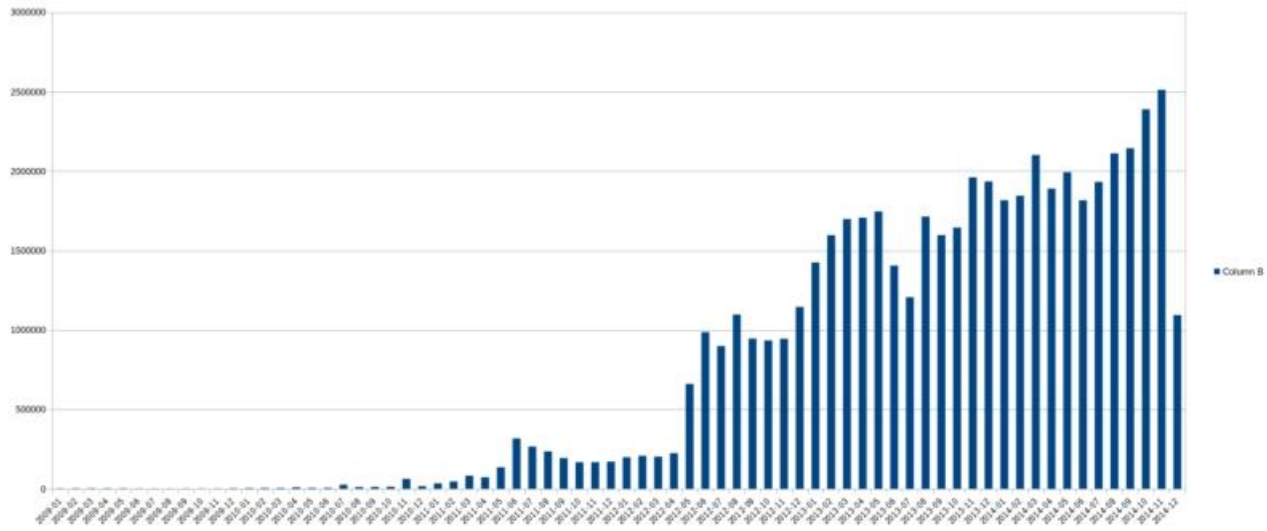
Batch Processing (Hadoop & Spark)

Part A:

Below is the bar plot showing the number of transactions that occurred every month from January 2009 till December 2014.



time passes. There is a sudden increase of bit coin utilisation during June of 2011, where this value goes up to approximately above 3 million. There was a stable situation for around one year till May 2012, after which there was a sudden rise in the utilisation of bit coin. During the months of May and June of 2012, the bit coin market witnessed a high spike which crossed the threshold of 10 million.
From May 2012 till end of 2014, there was an overall increase in the utilisation of bitcoin. We can see sometimes there is decline in its popularity as well, like you can see in June July of 2013, the graph went down. It then rose again, and in the end of 2014, it was touching 25million transactions.

Part B:

In this part we have extracted top 10 donors who have donated to WikiLeaks the most. Below is the procedure that is followed to get the top 10 donors and in the end, it is displayed who those donors are and how many pounds have been donated to WikiLeaks.

Procedure:

Cleaning:
In this pre-computational part, we have cleaned the data by removing header file and checked transaction's length of fields. In this way, we have a clean and filtered data of vin and vout.

Filtering vout:
In this part, we have filtered vout.csv by just extracting those transactions that have WikiLeaks id with it. We checked publicKey value of vout.csv in such a way this must be the same as of WikiLeaks bitcoin address. In short, receiving wallet was WikiLeaks in this case.

First Join:
We implemented first join on filtered vout (WikiLeaks only) and cleaned vin to get all details of donors that have donated to WikiLeaks. Join was done on "txid" of vin and "hash" of vout. In this way, we have these primary keys and after successful join on these keys, we now have all the fields from vin and details of wallet of Wikileaks from vout.

Second Join:
Result of first join was then joined again with full vout so basically, we now have all the details of donor from the result of the join. In this wat, we now have details of wallets of the donors as well as the details of WikiLeaks and their transaction details.

Sort:
In the end, we have sorted the results of second join using the values we have from that join. We extracted top 10 donors from this sorted list and those donors can be seen in below table.

Donor Wallet ID	Company/Owner	Bitcoins	Value in £
17B6mtZr14VnCKaHkvzqpkuxMYKTvezDcp	-	46515.189	144092517.04
19TCgtx62HQmaaGy8WNhLvoLXLr7LvaDYn	-	5770.0	17874028.7
14dQGpcUhejZ6QhAQ9UGVh7an78xoDnfap	-	1931.482	5983252.12
1LNWw6yCxkUmkhArb2Nf2MPw6vG7u5WG7q	Mt. Gox*	1005.303	3114179.32
1L8MdMLrgkCQJ1htiGRAcP11eJs662pYSS	-	806.134	2497203.16
1LNWw6yCxkUmkhArb2Nf2MPw6vG7u5WG7q	Mt. Gox*	806.099	2497094.74
1ECHwzKtRebkymjSnRKLqhQPkHCdDn6NeK	-	648.520	2008954.09
18pcznb96bbVE1mR7Di3hK7oWKsA1fDqhJ	-	637.043	1973401.19
19eXS2pE5f1yBggdwhPjauqCjS8YQCmnXa	-	576.835	1786891.74
1B9q5KG69tzjhqq3WSz3H7PAxDVTAwNdbV	-	556.7	1724518.51

*We have searched on blockchain.com these Wallet IDs to get company/owner name.

Part C:

In this part I have done first task to explore data by means of exchange as opposed to long term investments. I have explored the dataset by looking at transactions as if the bitcoins were respent on some other time or were those still present and not spent again (parked).

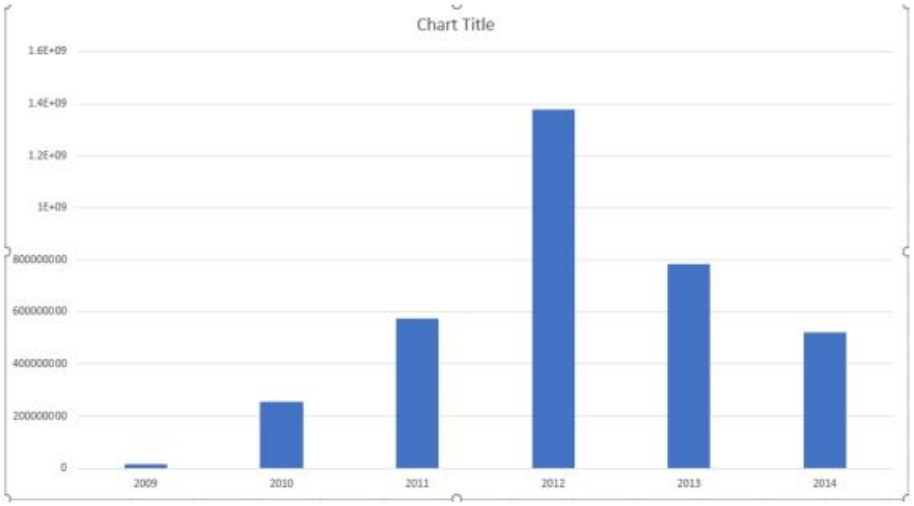
Logic behind this thing is we check "n" feature on vout to see if this is respent somewhere.

We joined the vin and vout datasets by mapping "hash" and "txid" of vout and vin respectively. In this way, we now have all the transations together. Then a join is used on feature "n" of vout with feature "vout" of vin, to get those transactions merged and that we have "n" and "vout" features.

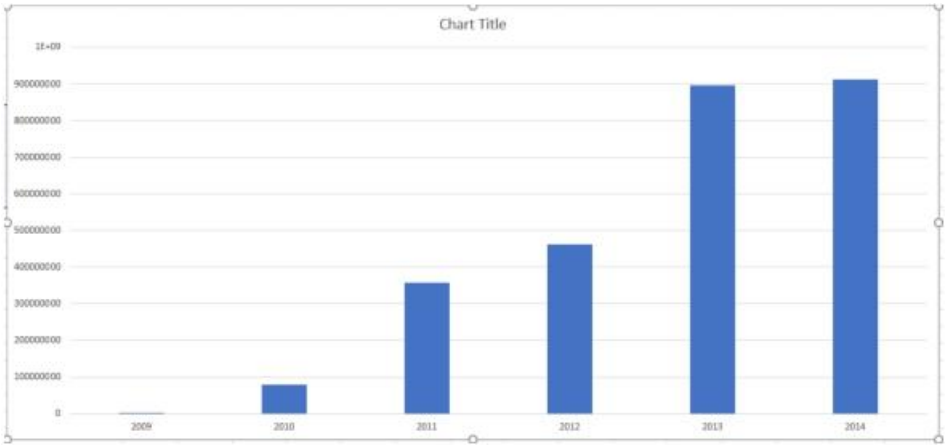
If both these features match, then we can realize that the bitcoin has been respent and that this is of small transaction(exchange). While all others must be parked coins (large purchases).

We then group these small and large transactions by year, so we can see them in grouping using different charts and graphs.

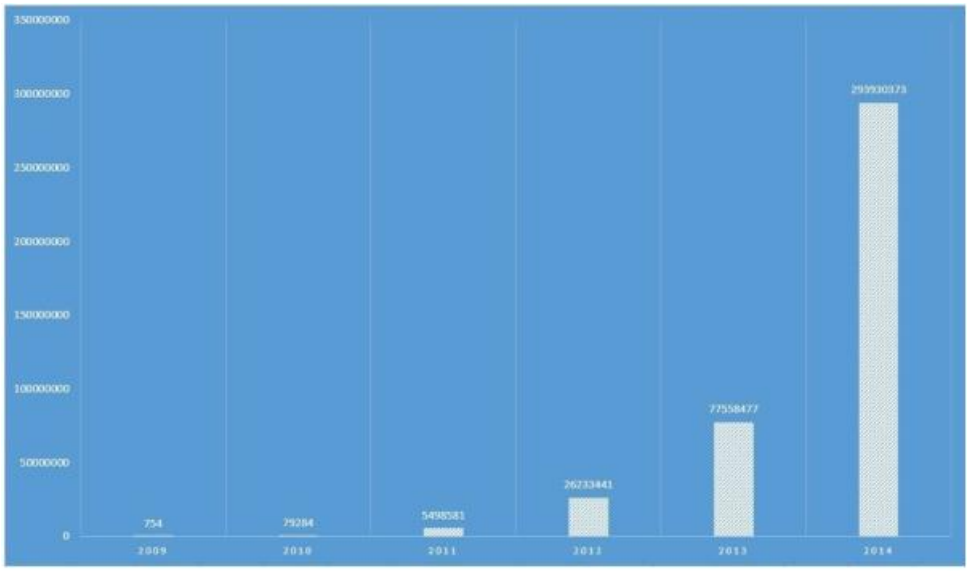
For small transactions, we have used bar chart to show sum of values of transactions grouped by year:



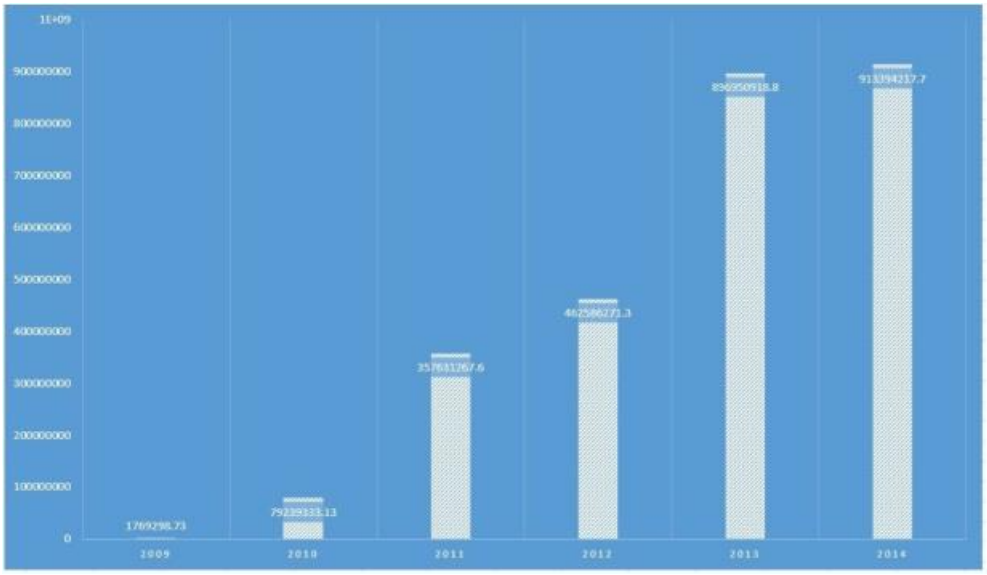
For large transactions, we have used bar chart to show sum of values of transactions grouped by year:



For small count transactions, we have used bar chart to show number of transactions grouped by year:



For large count transactions, we have used bar chart to show number of transactions grouped by year:



For total all transactions, we have used pie chart to show total number of transactions grouped by year:

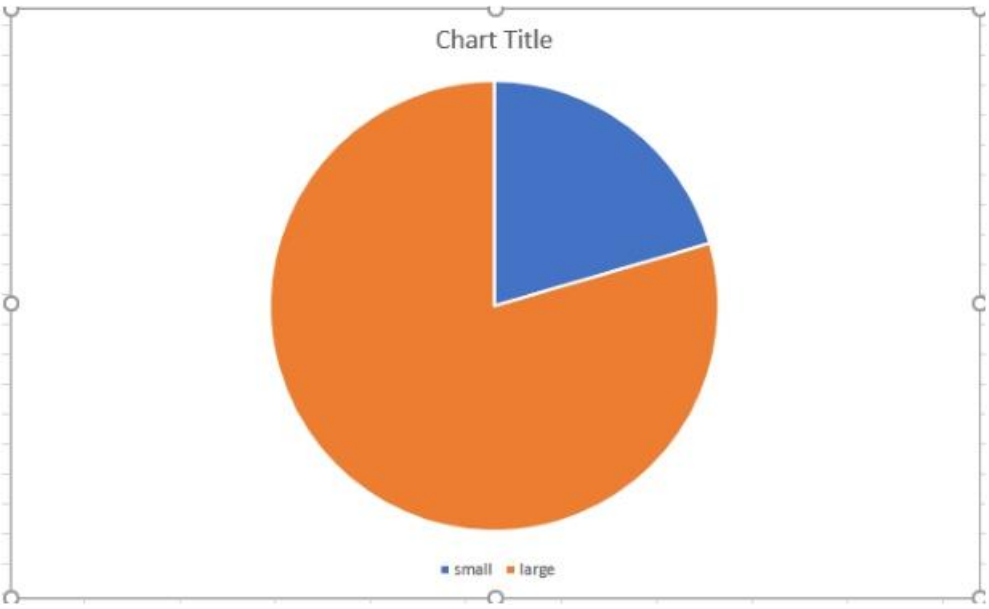


Chart Title
small
large