Ether Historical Data by Richard James Lopez

Project Details

This report details interesting exploratory data points for a Udacity Nanodegree for Data Analysis project. This is a perfect opportunity to explore data related to the cryptoassset Ether and its blockchain of Ethereum.

R libraries used

```r
# Useful libraries for project
library(gridExtra)
library(knitr)
library(ggplot2)
library(grid)
library(tidyr)
library(dplyr)
library(xlsx)
library(GGally)
library(RColorBrewer)
library(anytime)
library(scales)
library(ggthemes)
```

Files loaded

```r
address <- read.csv("address_v2.csv", header = TRUE)
all_data <- read.csv("all_data_v2.csv", header = TRUE)
Block_Size <- read.csv("block_size_v2.csv", header = TRUE)
block_difficulty <- read.csv("block_difficulty_v2.csv", header = TRUE)
etherprice <- read.csv('price_v2.csv', header = TRUE)
ethersupplygrowth <- read.csv('supply_v2.csv', header = TRUE)
hashrate <- read.csv('hashrate_v2.csv', header = TRUE)
marketcap <- read.csv('market_cap_v2.csv', header = TRUE)
tx <- read.csv('transactions_v2.csv', header = TRUE)
```

Data Overview

The dataset was featured on Kaggle and utilizes an etherscan API. It features a time series of data points for technological measurements like the average hashrate byday as well as economic measures like total supply of Ether coin in circulation. I discovered the dataset here https://www.kaggle.com/kingburrito666/ethereum-historical-data and got the latest data from https://etherscan.io/charts.

Content Overview

The Ethereum blockchain gives a revolutionary way of decentralizing applications and provides its own cryptocurrency. Ethereum is a decentralized platform that runs smart contracts: applications that run exactly as programmed without any possibility of downtime, censorship, fraud or third party interference. These apps run on a custom built blockchain, an enormously powerful shared global infrastructure that can move value around and represent the ownership of property. This enables developers to create markets, store registries of debts or promises, move funds in accordance with instructions given long in the past (like a will or a futures contract) and many other things that have not been invented yet, all without a middle man or counterparty risk.

Univariate Plots Section

These charts and tables start to explore simple counts of the variables. There are also data summaries and definitions that help inform what the data set variables mean in real terms and in relation to each other.

Basic Stats

To start, a count of the # of observations and variables for the complete dataset.

```
## [1] 834  12
```

Also included is an overview of the variables.

```
## 'data.frame':    834 obs. of  12 variables:
##  $ Date            : Factor w/ 834 levels "1/1/16","1/1/17",..: 634 637 652 685 718 727 730 733 736 7
##  $ Day             : Factor w/ 7 levels "Fri","Mon","Sat",..: 5 1 3 4 2 6 7 5 1 3 ...
##  $ Unix_Time_Stamp : int  1438214400 1438300800 1438387200 1438473600 1438560000 1438646400 143873280
##  $ Transactions    : int  8893 0 0 0 0 0 0 0 2050 2881 ...
##  $ Address         : int  9205 9361 9476 9563 9639 9696 9749 9790 10314 10730 ...
##  $ Price           : num  0 0 0 0 0 0 0 0 3 1.2 ...
##  $ Price_Change    : num  0 0 0 0 0 0 0 0 0 -0.6 ...
##  $ Supply          : num  72049307 72085498 72113204 72141428 72169404 ...
##  $ Hashrate        : num  23.8 48.2 55.3 64.2 69.9 ...
##  $ Block_Difficulty: num  0.121 0.603 0.887 1.02 1.126 ...
##  $ Block_Size      : int  644 582 575 581 587 587 579 584 633 668 ...
##  $ Market_Cap      : num  0 0 0 0 0 ...
```

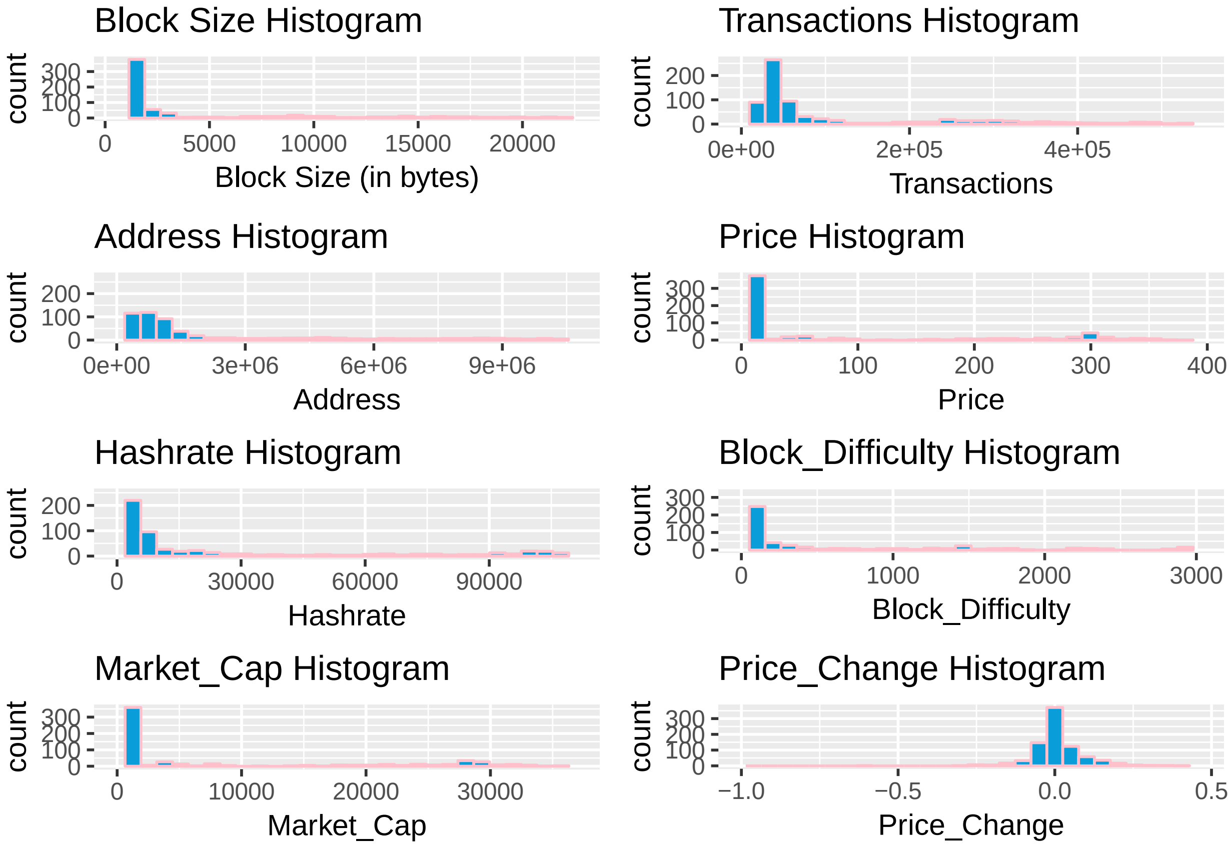
For a less formal view of the variable, here are a set of definitions to describe the variables in laymen terms.

- Date - date for the corresponding row of data.
- Day - day of week that the date represents.
- Unix_Time_Stamp - the time stamp of each day. Taken at 12 AM (midnight) each day.
- Transactions - # of orders successfully recorded on the Ethereum blockchain.
- Address - The number of public addresses key that are on the Ethereum blockchain.
- Price - value of Ether to USD for the day at the Unix_Time_Stamp.
- Price_Change - Simple change of price from previous day measured in % terms.
- Supply - amount of Ether coins in circulation.
- Hashrate - The number of hash calculations measured in GH/sec (for this dataset, the rate is a network hash rate for the whole ethereum blockchain in particular) - http://ethdocs.org/en/latest/glossary.html. It is representative of the combined power of the mining computers connected to the network. https://www.amazon.com/Cryptoassets-Innovative-Investors-Bitcoin-Beyond/dp/1260026671
- Block_Difficulty - In very general terms, the amount of effort required to mine a new block. Note that the difficulty algorithm is subject to change as what has happened with the launch of Homestead on March 16, 2016. - http://ethdocs.org/en/latest/glossary.html

- Block__Size - Ethereum's block size is based on complexity of contracts being run – it's known as a Gas limit per block, and the maximum can vary slightly from block to block. This data set of the average blocksize for the day measured in bytes. https://bitsonblocks.net/2016/10/02/a-gentle-introduction-to-ethereum/
- Market__Cap - Value of ether multiplied by the supply of Ether coins at the time of the valuation (in USD). A rough proxy for how much value the total outstanding amount of Ether represents in global markets. Can be compared to the Market Cap of other cryptoassets.

Univariate Charts

The simplicity of univariate charts is to help us identify general ideas about variables. Running a set of histograms is a first step in examining the distribution, spotting outliers and making changes to the dataset before starting the analysis.

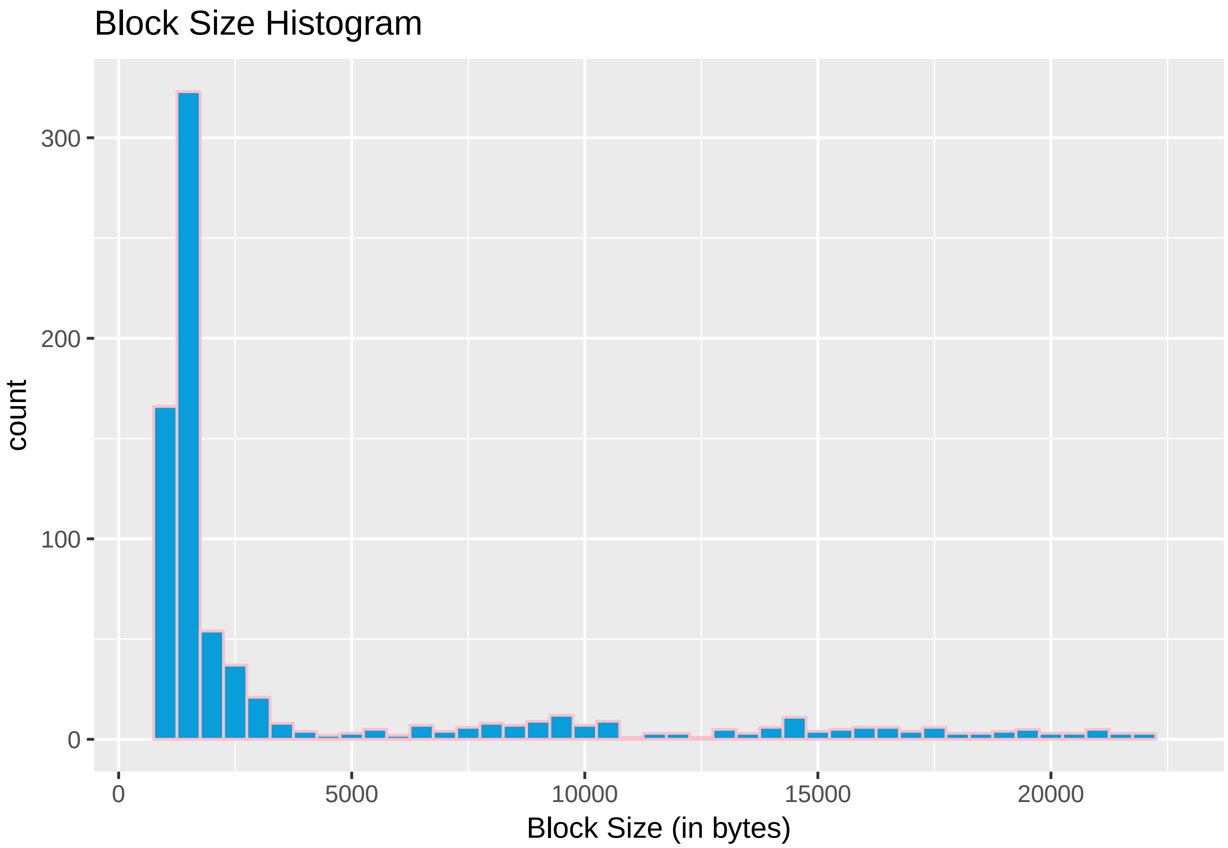


Based on the charts above, the variable "Block_Size" is convenient to examine. Of all the variables, there are the fewer preconceived notions about it given that is a fairly obscure measurement. It also appears to be extremely skewed to the left. Bringing up a summary of the stats is in order.

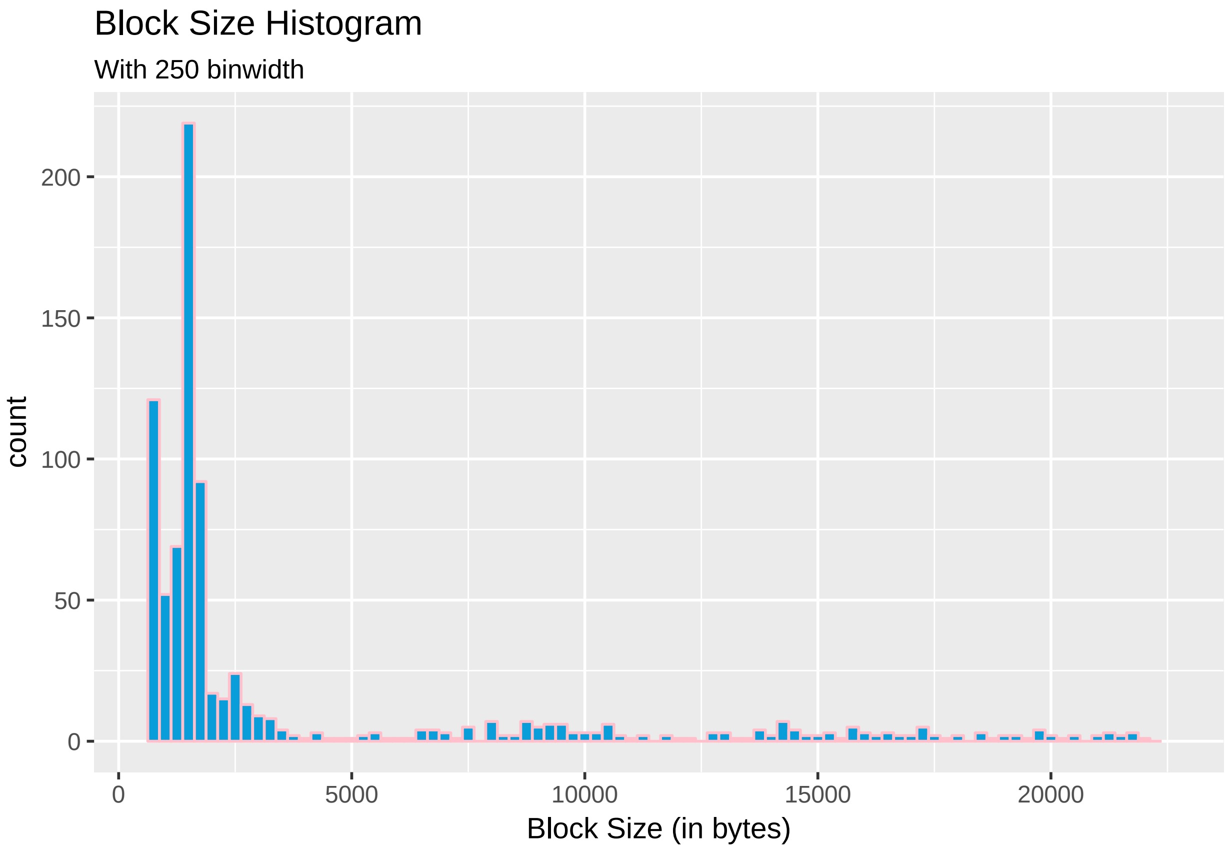
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     575    1252    1568    3893    2785   22600
```

Block Size Histogram

Digging further into the distribution allows us to shine light on how the 4th quartile appears to be so far from the 1st.

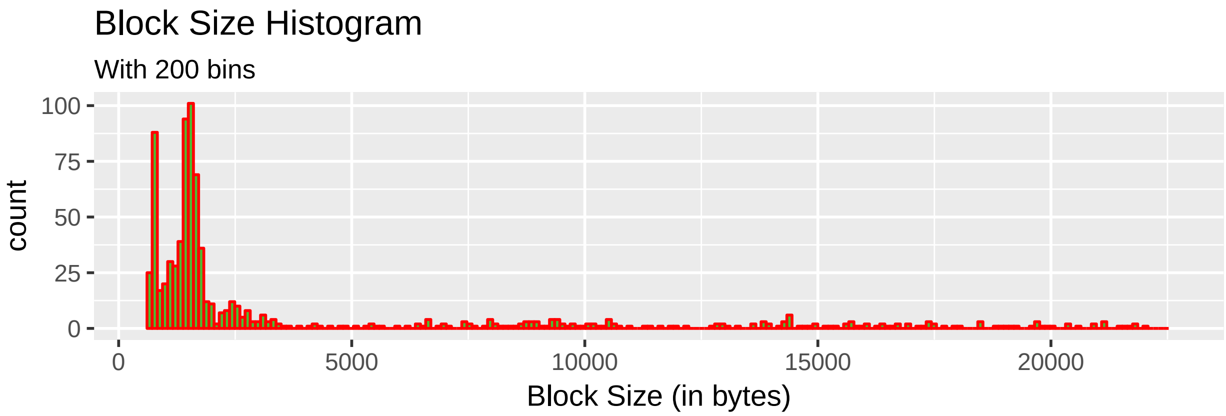


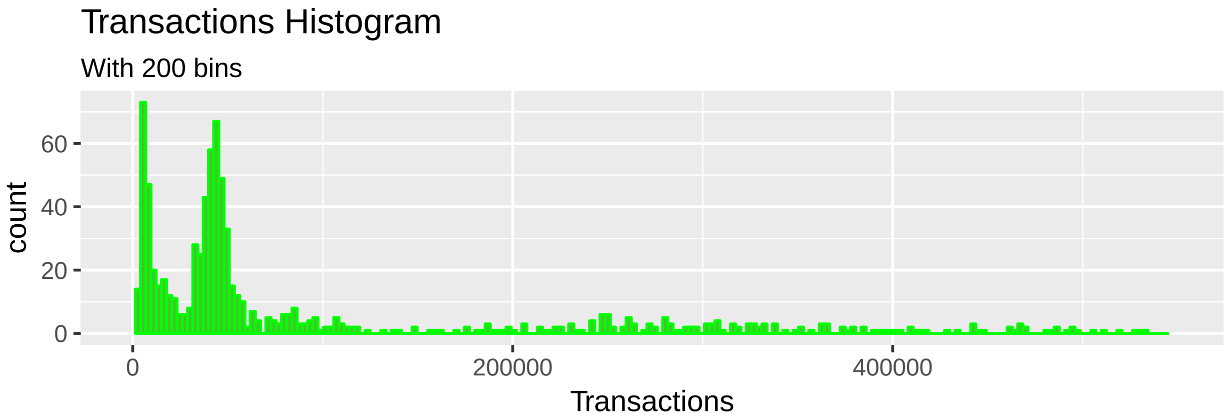
By halving the binwidth to 250 there is a more nuanced view that decouples some of those lesser Block Size values. This is evident by the two horned values at the start of the chart to the left.



Univariate__Plots__Transactions

To test the intuition behind these variables, a histogram with Transactions instead of Block Size should have similar dimensions. Logically, the amount of Transactions should help constitute how large the Block Size and we can test that logic here. Below are two histograms side by side with 100 bins.

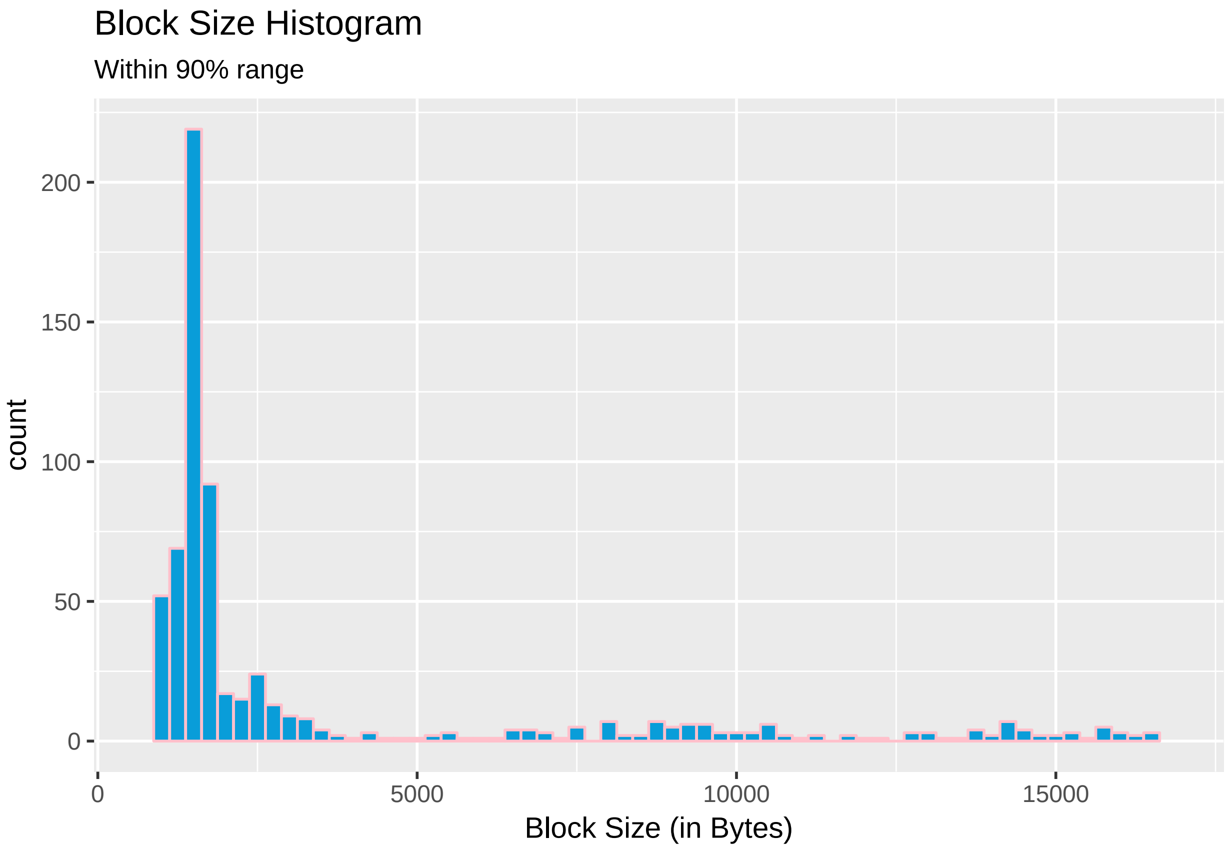




The chart does appear to have a similar distribution, but it is hardly exact. There will be more analysis with Block_Size and Transactions in more of the Bivariate and Multivariate charts too.

Univariate_Plots_Block_Size_90%Range

The Block_Size also seems to have a tail of large Block Size values occurring less frequently. By cutting off 5% of each tail, a new picture could emerge. This trimming removed 84 of the rows containing the smallest and largest BlockSizes values.



This range of values changes the distribution slightly (keeping binwidth of 250), and the left peak that existed earlier in chart Block_Size Histogram V2, but there is still the lingering values of big block sizes that exist on the right side of the chart.

Univariate Analysis

Below are a set of standard questions and answers for the univariate analysis.

What is the structure of your dataset?

> The structure of my data is a time series that tracks the features of the cryptoasset Ether.

What is/are the main feature(s) of interest in your dataset?

> The main features are possible relationships between variables of Ether. The price of Ether is relatively volatile compared to other assets and so I would like to explore relationships that may uncover behavior in Ether. The lesser known technical variables like "Block_Size", "Block_Difficulty", "Hashrate", and "Transactions" can inform an intuition on what relationships exist in Ethereum.

What other features in the dataset do you think will help support your investigation into your feature(s) of interest?

> The more technical variables interest me but I will also more common place measurements like the "Day" of the week and "Price_Change".

Did you create any new variables from existing variables in the dataset?

> Only the "Price_Change". This was possible because the Price values were collected at the same time of every day. Caveat emptor because the price volatility throughout the day consequently changes the Price_Change measured. For this reason, I would take this Price_Change as an

absolute metric because the change is relative to when a user might consider a good time to measure the price. That being said, since Ether trades 24/7, I don't think there is an absolute time to measure Price_change (unless you were measuring in real time as most websites do).

Also, technically I created the "Day" variable although that was done with a simple Excel formula.

Of the features you investigated, were there any unusual distributions?
Did you perform any operations on the data to tidy, adjust, or change the form
of the data? If so, why did you do this?

There was some unusual behavior in the 2 weeks of the Ethereum blockchain. For example, there are transactions on the first day with a zero price happening on the first day of trading. Also, once the price is a positive value, it goes back to 0 on the 12th day. This behavior required that I adjust my formula for Price_Change when compiling the data and creating this variable. You can see in row 12 of the data that there was a Price anomaly from the complete dataset.

```
##       Date Day Unix_Time_Stamp Transactions Address Price Price_Change
## 1  7/30/15 Thu      1438214400         8893    9205  0.00    0.0000000
## 2  7/31/15 Fri      1438300800            0    9361  0.00    0.0000000
## 3   8/1/15 Sat      1438387200            0    9476  0.00    0.0000000
## 4   8/2/15 Sun      1438473600            0    9563  0.00    0.0000000
## 5   8/3/15 Mon      1438560000            0    9639  0.00    0.0000000
## 6   8/4/15 Tue      1438646400            0    9696  0.00    0.0000000
## 7   8/5/15 Wed      1438732800            0    9749  0.00    0.0000000
## 8   8/6/15 Thu      1438819200            0    9790  0.00    0.0000000
## 9   8/7/15 Fri      1438905600         2050   10314  3.00    0.0000000
## 10  8/8/15 Sat      1438992000         2881   10730  1.20   -0.6000000
## 11  8/9/15 Sun      1439078400         1329   11004  1.20    0.0000000
## 12 8/10/15 Mon      1439164800         2037   11679  0.00   -1.0000000
## 13 8/11/15 Tue      1439251200         4963   13576  0.99    0.0000000
## 14 8/12/15 Wed      1439337600         2036   13913  1.29    0.3030303
##      Supply Hashrate Block_Difficulty Block_Size Market_Cap
## 1  72049307  23.7569            0.121        644    0.00000
## 2  72085498  48.1584            0.603        582    0.00000
## 3  72113204  55.2709            0.887        575    0.00000
## 4  72141428  64.1779            1.020        581    0.00000
## 5  72169404  69.8559            1.126        587    0.00000
## 6  72197883  76.6115            1.217        587    0.00000
## 7  72225411  81.9449            1.328        579    0.00000
## 8  72252487  82.9366            1.381        584    0.00000
## 9  72279925  89.6063            1.471        633  216.83977
## 10 72307868  97.6083            1.586        668   86.76944
## 11 72335046 102.5407            1.709        618   86.80206
## 12 72362864 113.1109            1.838        631    0.00000
## 13 72390891 126.6631            2.036        692   71.66698
## 14 72418262 132.7661            2.207        653   93.41956
```

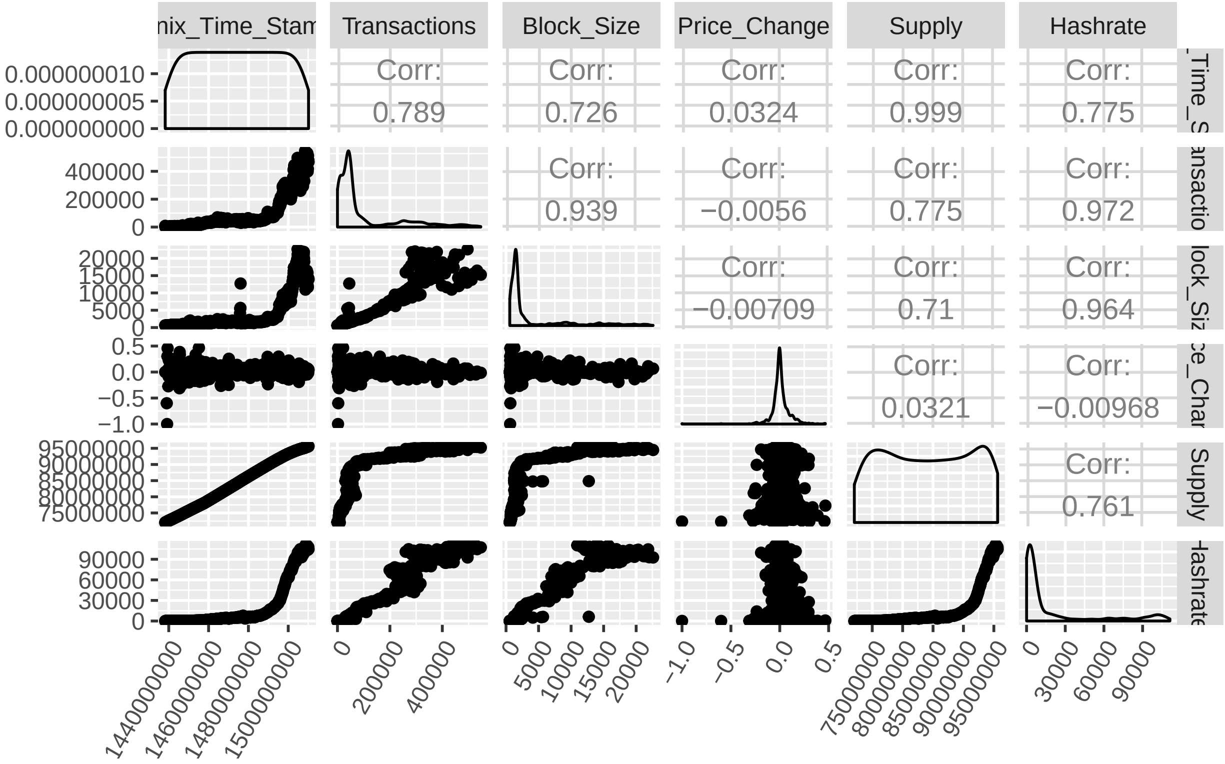
Bivariate Plots Section

These charts and tables dig further into the relationships between the variables. When particular variables are coupled together, it is apparent that subsetting as well as cutting off the tails for some distributions are worthwhile activities to understand the context of the data time series.

Basic Stats

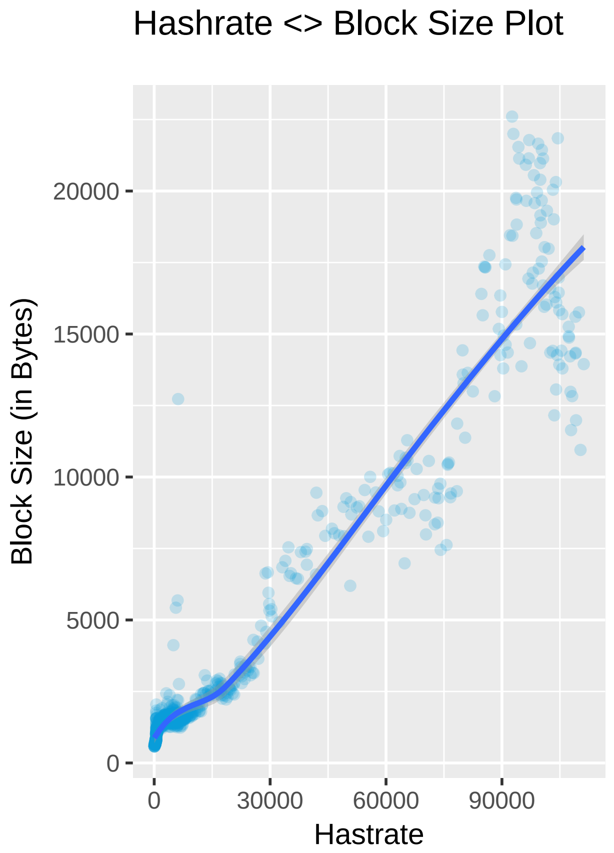
To start out examing variables, a ggppairs chart may give some initial ideas for Bivariate plots to examine.

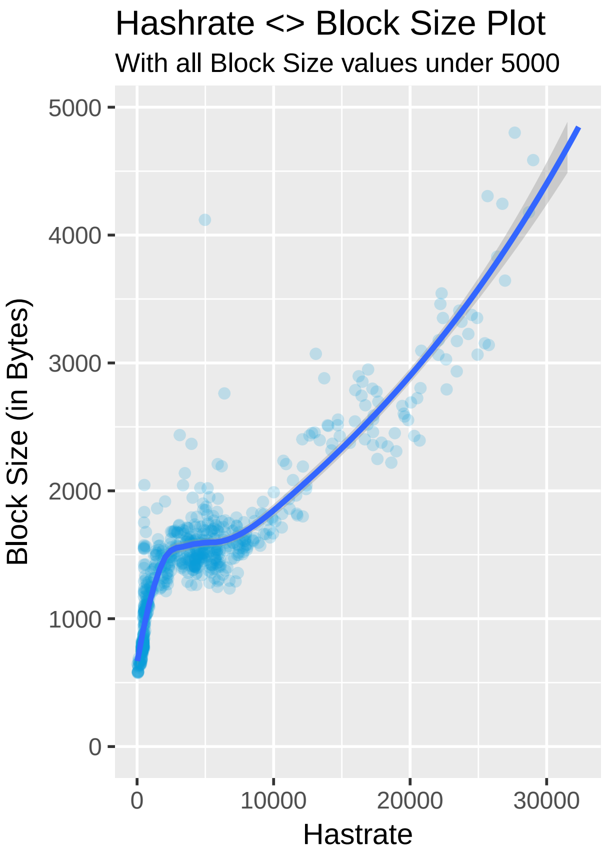
```
## [1] "Unix_Time_Stamp" "Transactions"    "Block_Size"      "Price_Change"
## [5] "Supply"          "Hashrate"
```



Block Size <> Hashrate

Based on the Correlation charts above, there is a strong (0.964) correlation coefficient between the Block_Size and Hashrate. This makes sense technologically as more hashes would create a larger block. However, the technology for writing hashes has changed over the course of this time series. Visualizing the relationship can show the story the correlation coefficient is telling.

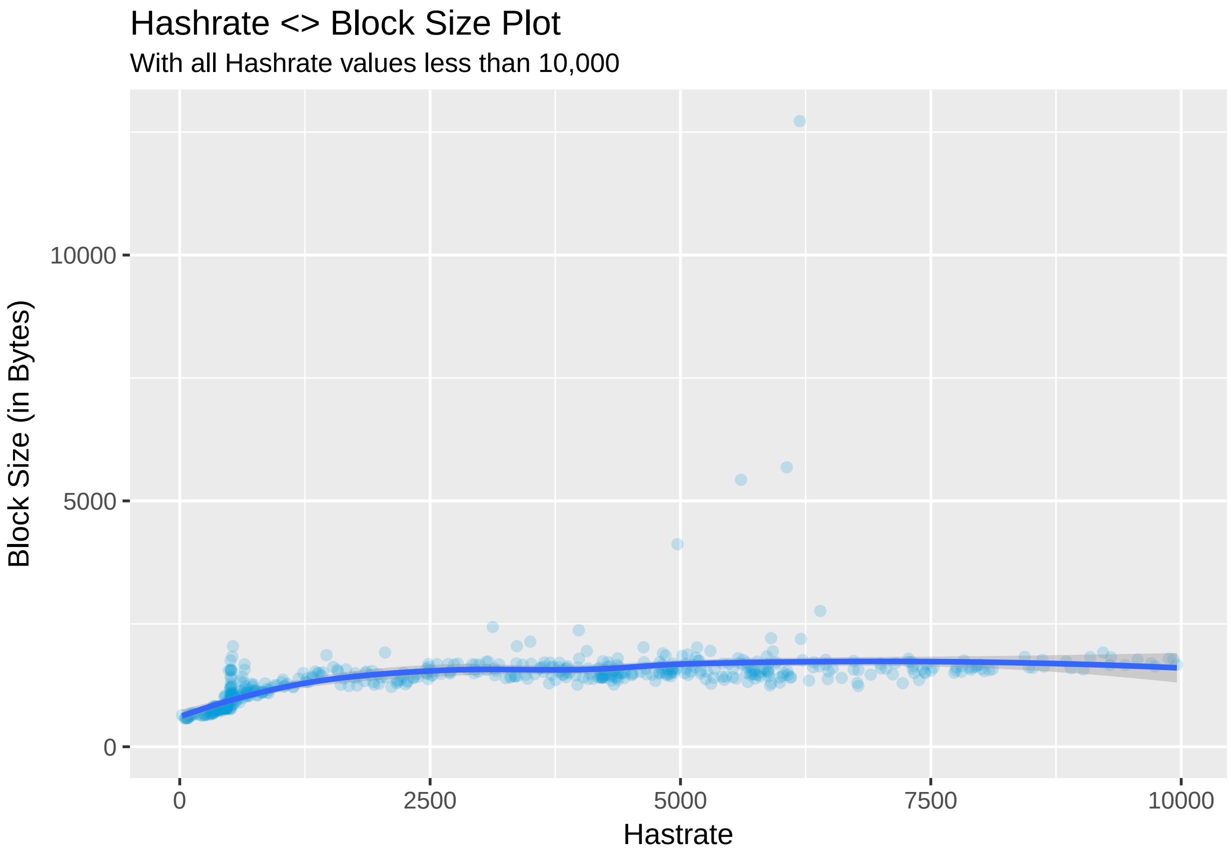




The above chart to the left is interesting because there seems to be two data regimes.

A block size of up to 5000 Bytes appears to have tight positive correlation to hashrate (relative to blocks larger than 5000 Bytes). Placing another scatter plot to the right of it with the smaller subset of data will focus on the smaller block size.

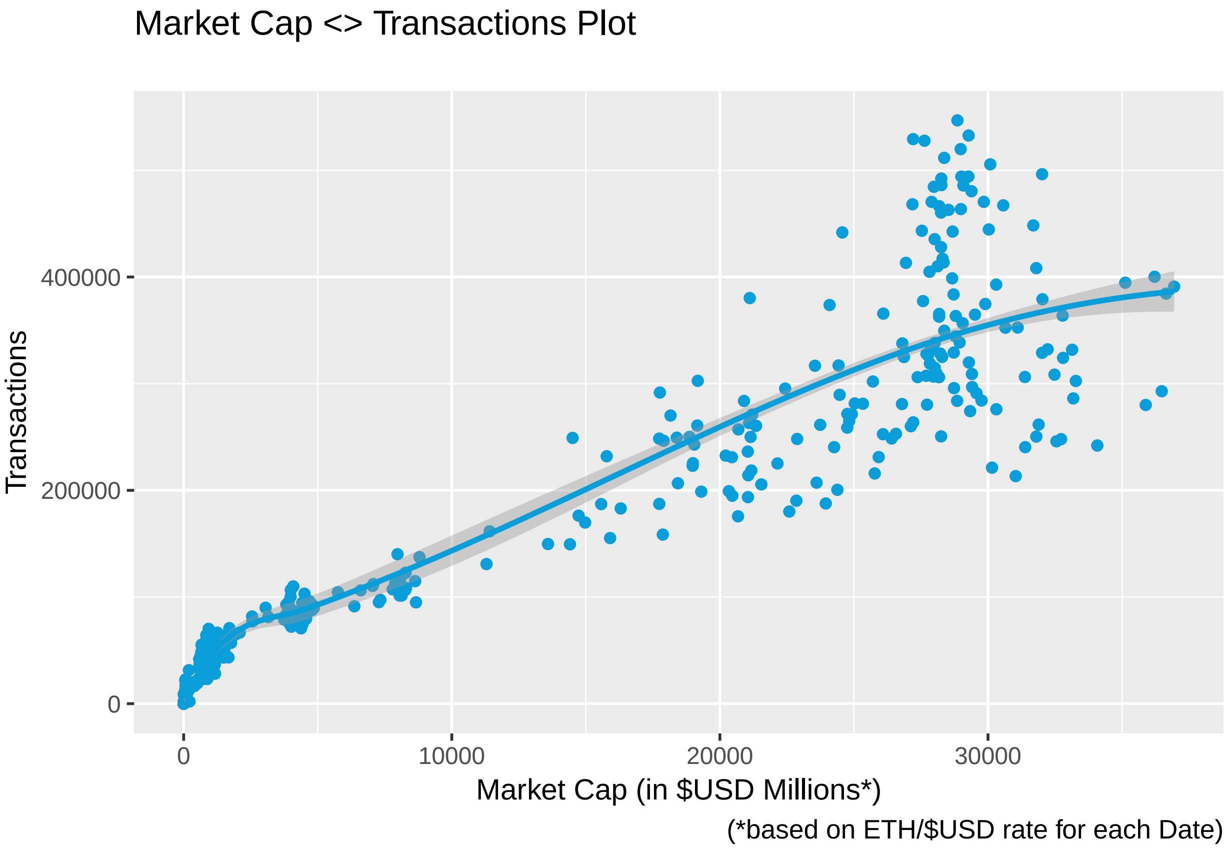
From this chart, maybe the Block_Size is not the main departure of behavior. Perhaps it has to do with blocks that have a hashrate of less than 10,000 Bytes.



This visualization shows that for relatively small Block Sizes, hashrate is uncorrelated. There is a range of a hashrate of 0 to 10000 Bytes. This may be due to upgraded technological protocol, the decreasing size of transaction value that may affect the hashrate, or it may just be hard to scale block size as it get larger. This visualization begets new questions that weren't initially visible.

Market Cap <> Transactions

The relationship betweenMarket Cap and Transactions is remarkably similar to the relationship between Block Size and Hash Rate. It stands to reason that as the Market Cap increases, there will be more liquidity (Transactions per day as a proxy).

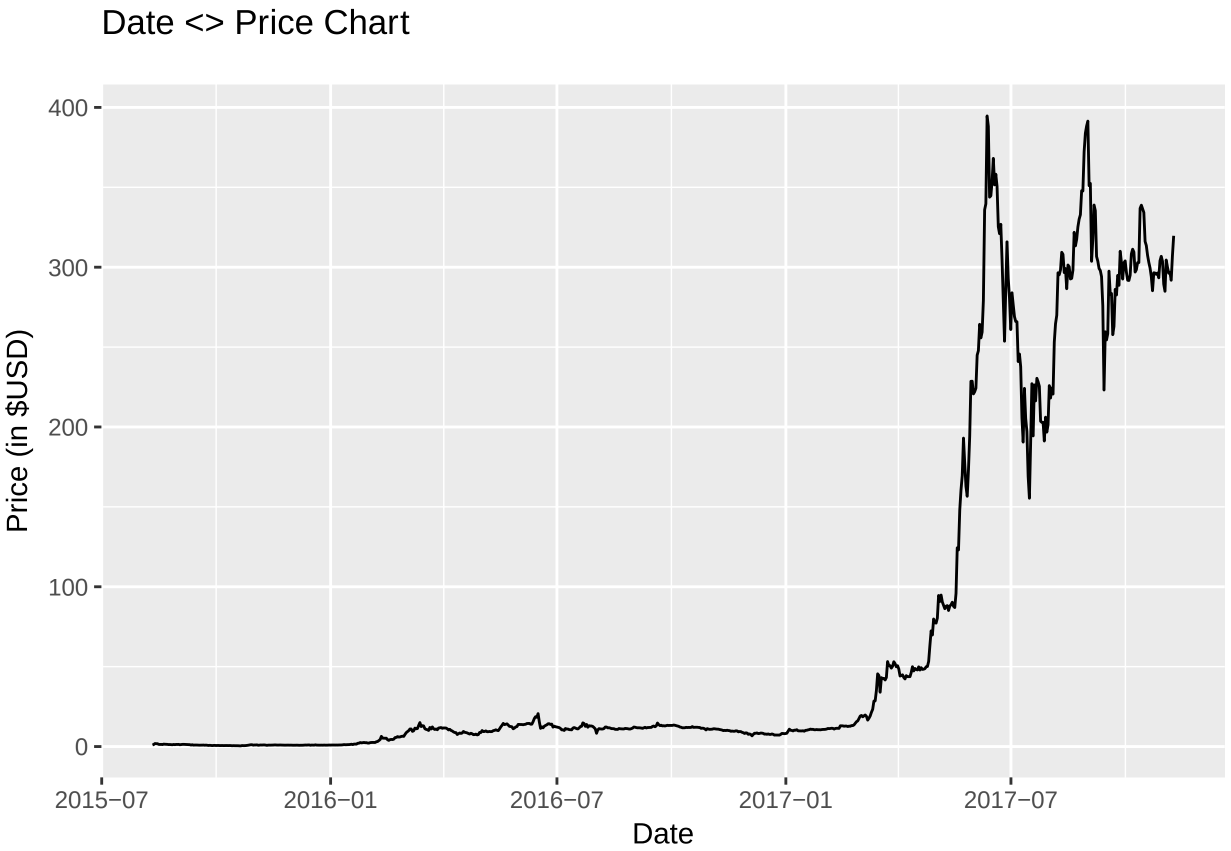


Just as increasing Hashrates helps to prove the robustness of the Block_Size, Market Cap increases indicates that Transactions will probably be more frequent. While not necessarily true (you can have a large Market Cap with just a handful of Ether holders who do not making regular transactions), this chart appears to resemble that of the Block_Size <> Hashrate chart.

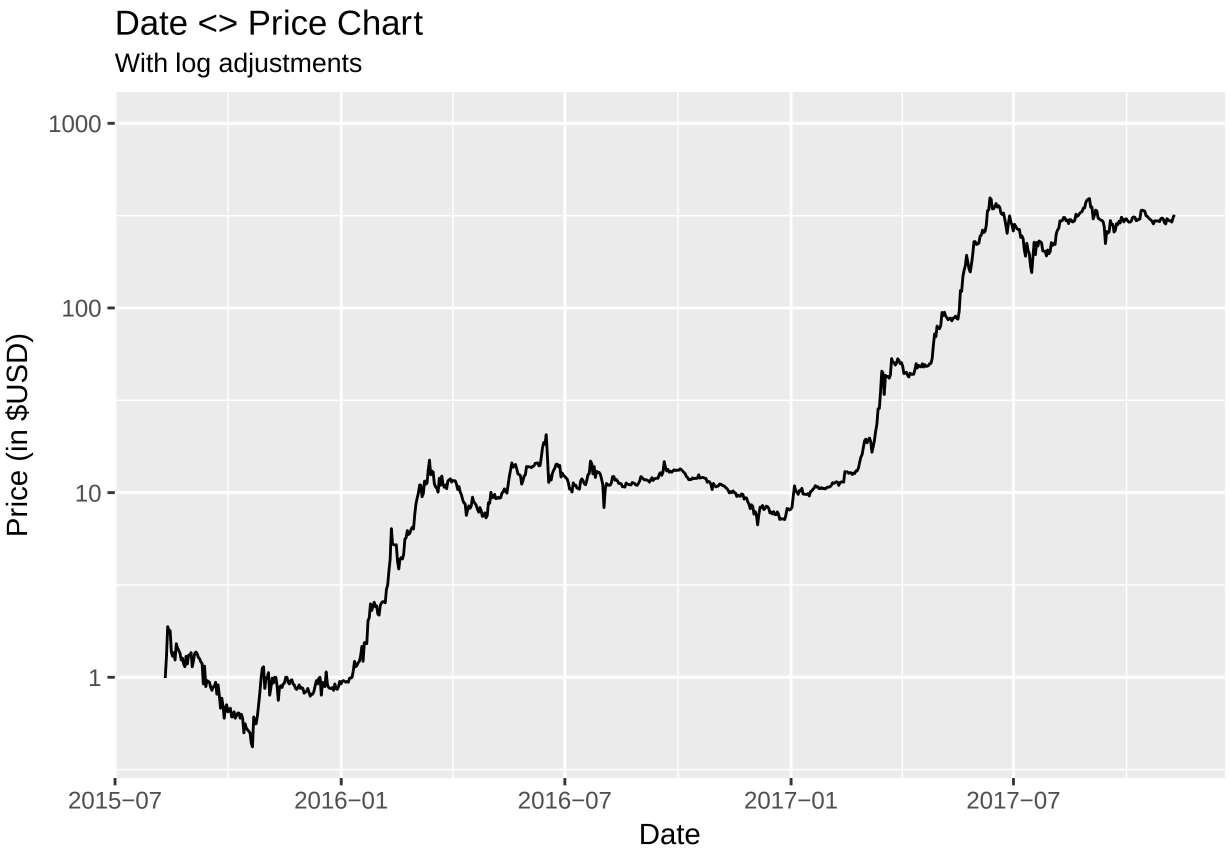
Price <> Time

One of the more vanilla relationships will be to examine Price behavior.

```
##       Date Day Unix_Time_Stamp Transactions Address Price Price_Change
## 1  7/30/15 Thu      1438214400         8893    9205  0.00    0.0000000
## 2  7/31/15 Fri      1438300800            0    9361  0.00    0.0000000
## 3   8/1/15 Sat      1438387200            0    9476  0.00    0.0000000
## 4   8/2/15 Sun      1438473600            0    9563  0.00    0.0000000
## 5   8/3/15 Mon      1438560000            0    9639  0.00    0.0000000
## 6   8/4/15 Tue      1438646400            0    9696  0.00    0.0000000
## 7   8/5/15 Wed      1438732800            0    9749  0.00    0.0000000
## 8   8/6/15 Thu      1438819200            0    9790  0.00    0.0000000
## 9   8/7/15 Fri      1438905600         2050   10314  3.00    0.0000000
## 10  8/8/15 Sat      1438992000         2881   10730  1.20   -0.6000000
## 11  8/9/15 Sun      1439078400         1329   11004  1.20    0.0000000
## 12 8/10/15 Mon      1439164800         2037   11679  0.00   -1.0000000
## 13 8/11/15 Tue      1439251200         4963   13576  0.99    0.0000000
## 14 8/12/15 Wed      1439337600         2036   13913  1.29    0.3030303
##      Supply Hashrate Block_Difficulty Block_Size Market_Cap
## 1  72049307  23.7569            0.121        644    0.00000
## 2  72085498  48.1584            0.603        582    0.00000
## 3  72113204  55.2709            0.887        575    0.00000
## 4  72141428  64.1779            1.020        581    0.00000
## 5  72169404  69.8559            1.126        587    0.00000
## 6  72197883  76.6115            1.217        587    0.00000
## 7  72225411  81.9449            1.328        579    0.00000
## 8  72252487  82.9366            1.381        584    0.00000
## 9  72279925  89.6063            1.471        633  216.83977
## 10 72307868  97.6083            1.586        668   86.76944
## 11 72335046 102.5407            1.709        618   86.80206
## 12 72362864 113.1109            1.838        631    0.00000
## 13 72390891 126.6631            2.036        692   71.66698
## 14 72418262 132.7661            2.207        653   93.41956
```



The price above shows the linear modeling of the Price which is generally not recommended for asset price. The plot below models regular behavior for asset prices by making sure to take the log10 value of the 10 for the scale. You will find popular websites like coinmarketcap.com also have this log scale option on their price charts.



Price Change <> Transactions

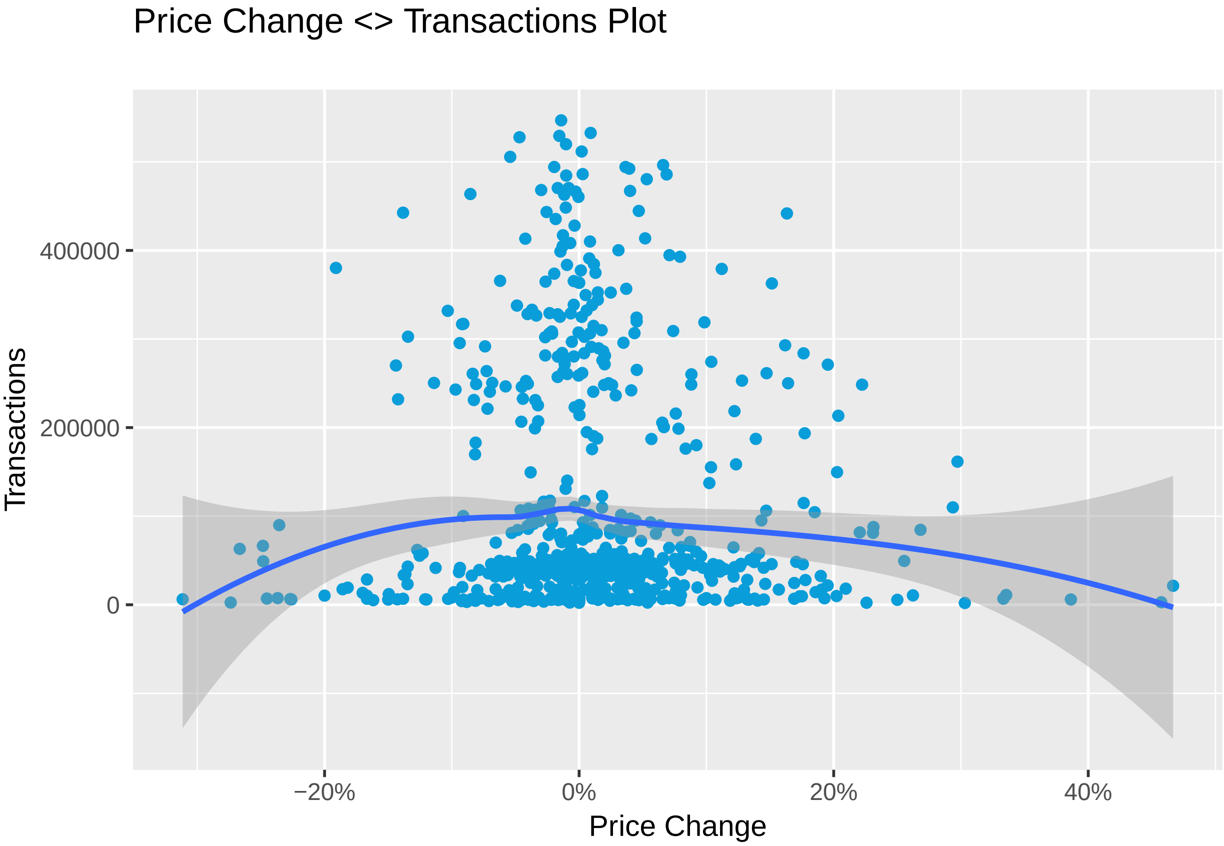
There also may be a relationship between Price Change and the # of transactions (perhaps as an unadjusted proxy for volume). To do this, one can subset the data so that it removes the beginning time period where there were price values of zero.

Viewing the first rows of the data identifies when the Price have a continuous positive series. 8/11/15 is a proper date to start.

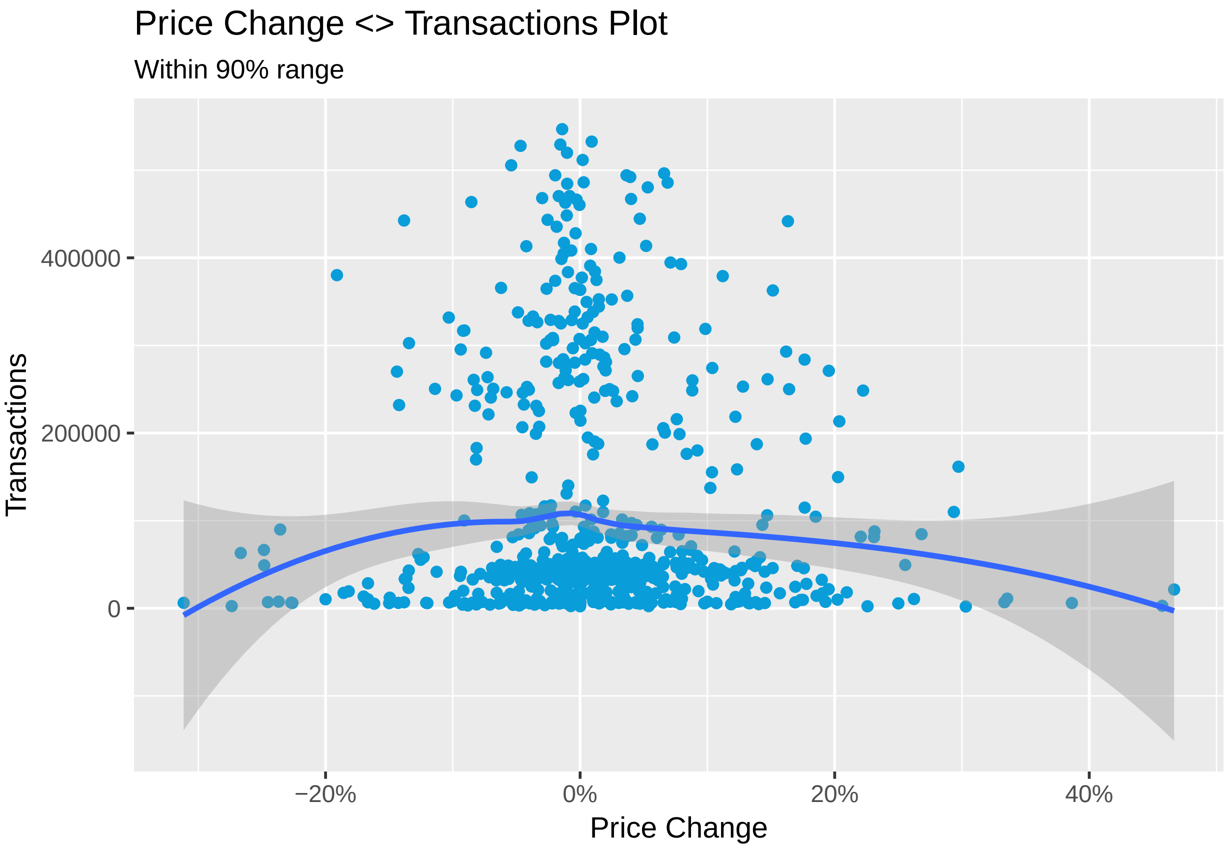
```
##       Date Day Unix_Time_Stamp Transactions Address Price Price_Change
## 1  7/30/15 Thu      1438214400         8893    9205  0.00    0.0000000
## 2  7/31/15 Fri      1438300800            0    9361  0.00    0.0000000
## 3   8/1/15 Sat      1438387200            0    9476  0.00    0.0000000
## 4   8/2/15 Sun      1438473600            0    9563  0.00    0.0000000
## 5   8/3/15 Mon      1438560000            0    9639  0.00    0.0000000
## 6   8/4/15 Tue      1438646400            0    9696  0.00    0.0000000
## 7   8/5/15 Wed      1438732800            0    9749  0.00    0.0000000
## 8   8/6/15 Thu      1438819200            0    9790  0.00    0.0000000
## 9   8/7/15 Fri      1438905600         2050   10314  3.00    0.0000000
## 10  8/8/15 Sat      1438992000         2881   10730  1.20   -0.6000000
## 11  8/9/15 Sun      1439078400         1329   11004  1.20    0.0000000
## 12 8/10/15 Mon      1439164800         2037   11679  0.00   -1.0000000
## 13 8/11/15 Tue      1439251200         4963   13576  0.99    0.0000000
## 14 8/12/15 Wed      1439337600         2036   13913  1.29    0.3030303
##      Supply Hashrate Block_Difficulty Block_Size Market_Cap
## 1  72049307  23.7569            0.121        644    0.00000
## 2  72085498  48.1584            0.603        582    0.00000
## 3  72113204  55.2709            0.887        575    0.00000
## 4  72141428  64.1779            1.020        581    0.00000
## 5  72169404  69.8559            1.126        587    0.00000
## 6  72197883  76.6115            1.217        587    0.00000
## 7  72225411  81.9449            1.328        579    0.00000
## 8  72252487  82.9366            1.381        584    0.00000
## 9  72279925  89.6063            1.471        633  216.83977
## 10 72307868  97.6083            1.586        668   86.76944
## 11 72335046 102.5407            1.709        618   86.80206
## 12 72362864 113.1109            1.838        631    0.00000
## 13 72390891 126.6631            2.036        692   71.66698
## 14 72418262 132.7661            2.207        653   93.41956
```

```
all_data_post_zero <- subset(all_data,
                             Unix_Time_Stamp >= 1439251200)
```

Now that the data is subset, below is the relationship between the Price Change (post-1st 2 weeks) and the # of Transactions.



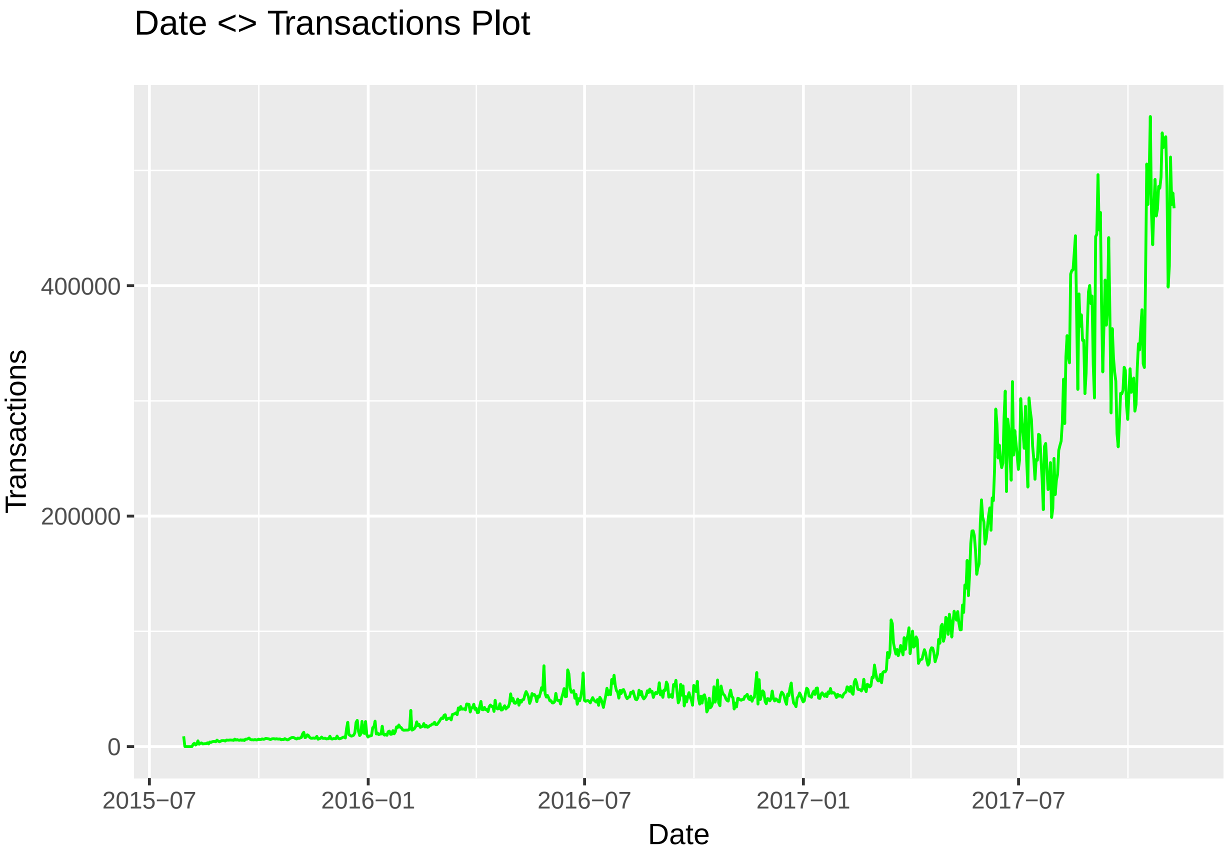
These charts seem to frame a cluster of data points between in single digits Price_Changes. This makes sense for normal asset price behavior. However, I don't want to use this somewhat arbitrary cutoff. Instead I will cut off the tails by 5% among the distribution.



Within this range, I see that a majority of the data points are less than +/- 10% (Note: 82 observations were removed from the dataset when the tails of the distribution were filtered out). The pattern that stands out is how there is a loose cluster of data points with high # of transactions that exists over a gap above a tighter cluster of low # of transactions. There could be several explanations for this. My intuition is that there is a significant uptick of transactions along our time series data. There are probably two difficult time periods to these data points.

I can examine this future with a simple plot of the # of transactions over time.

```
options(scipen=10000)
ggplot(data = all_data,
       aes(x = anytime(Unix_Time_Stamp), y = Transactions)) +
  geom_line(color = 'green') +
  xlab('Date') +
  ggtitle('Date <> Transactions Plot',
          subtitle = '')
```



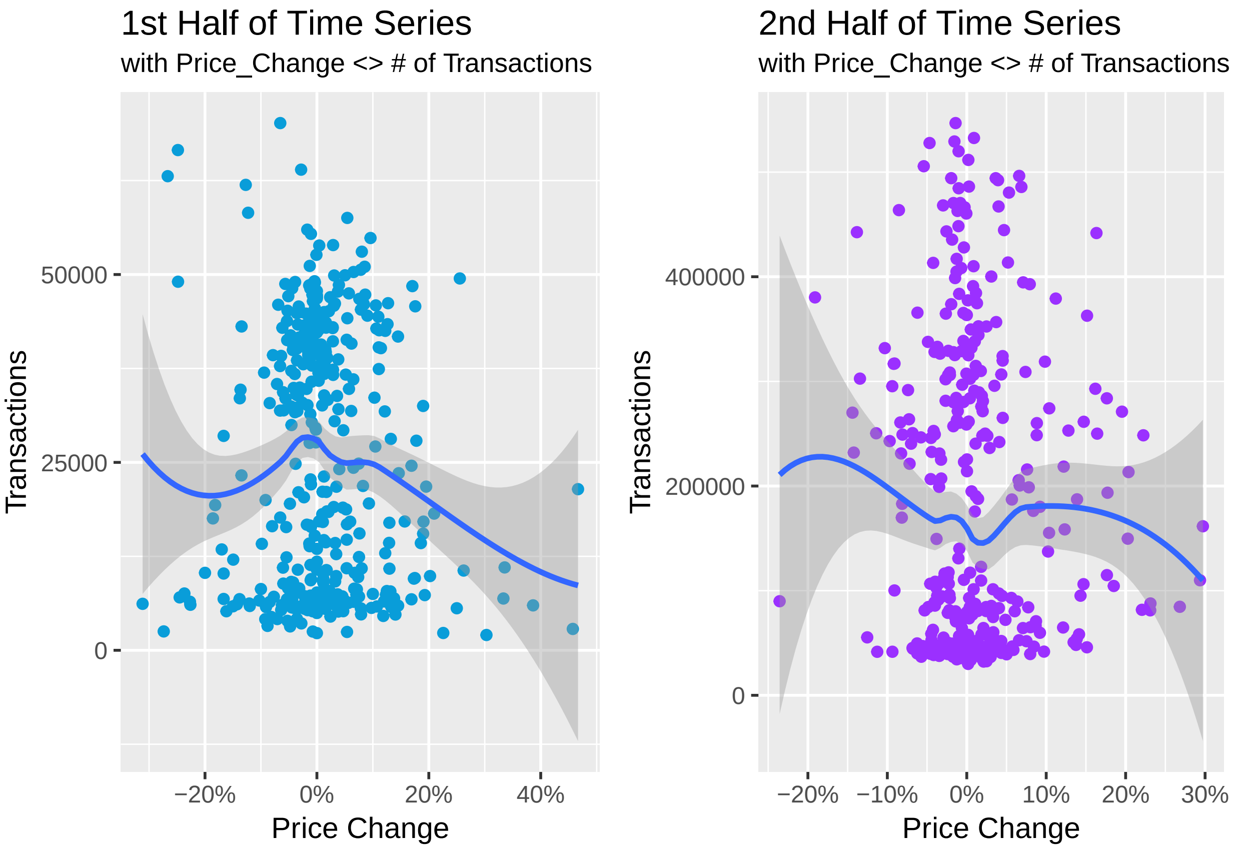
I contemplated cherry-picking what would be a good cutoff point. However for future rendering of this data, I simply decided to choose the middle point of the time series and approximately slice the data into two halves.

With a summary I can find the median point.

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
## 1438000000 1456000000 1474000000 1474000000 1492000000 1510000000
```

```
all_data_Transactions_1st <- subset(all_data_post_zero,
                                    Unix_Time_Stamp <= 1.474e+09 )
all_data_Transactions_2nd <- subset(all_data_post_zero,
                                    Unix_Time_Stamp > 1.474e+09 )
```

Now, let me create new charts with these two subsets.



1st Half of Time Series

with Price_Change <> # of Transactions

2nd Half of Time Series

with Price_Change <> # of Transactions

Based on these charts above, one can observe that the scale of transaction numbers is bigger for the 2nd half of the history. However, there appears to be a similar pattern of transactions to price change. While it would be fruitful to examine the periods, no divergent patterns jump off the pageat first glance.

Bivariate Analysis

Below are a set of standard questions and answers for the bivariate analysis.

Talk about some of the relationships you observed in this part of the investigation. How did the feature(s) of interest vary with other features in the dataset?

> The parallel between Block Size <> Hashrate & Transactions <> Market Cap was something that I didn't originally suspect. The visualization helped play outsome logical relationships that I could posit, but couldn't substantiate till I explored this data.

Did you observe any interesting relationships between the other features (not the main feature(s) of interest)?

> I observed that there appears to be two different data clusters of Ether trading. Based on the chart titled "Price_Change <> Transactions". This may be due to a function of time or perhaps a technological protocol update.

What was the strongest relationship you found?

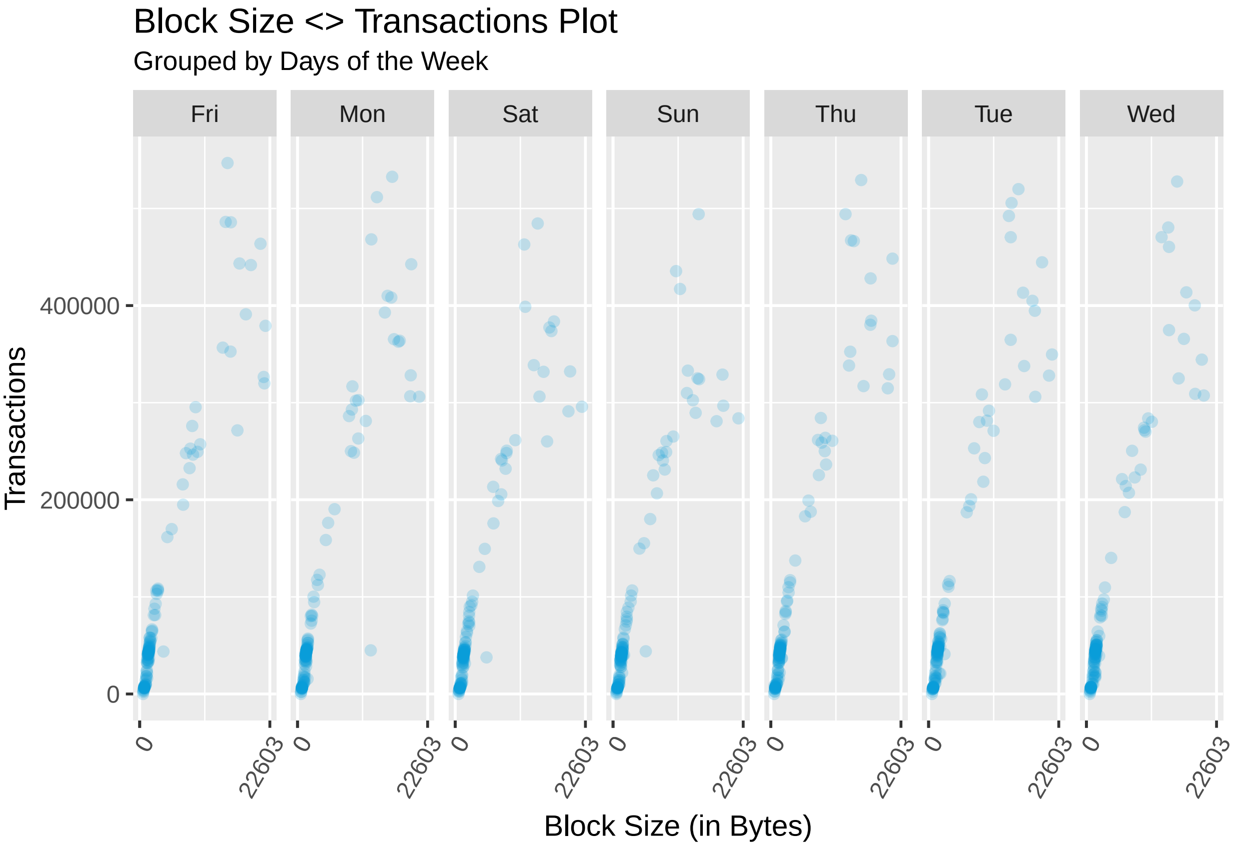
> Unix_Time_Stamp and Supply. This makes sense as the supply of Ether coins was designed to be released on a publicized and regular schedule.

Multivariate Plots Section

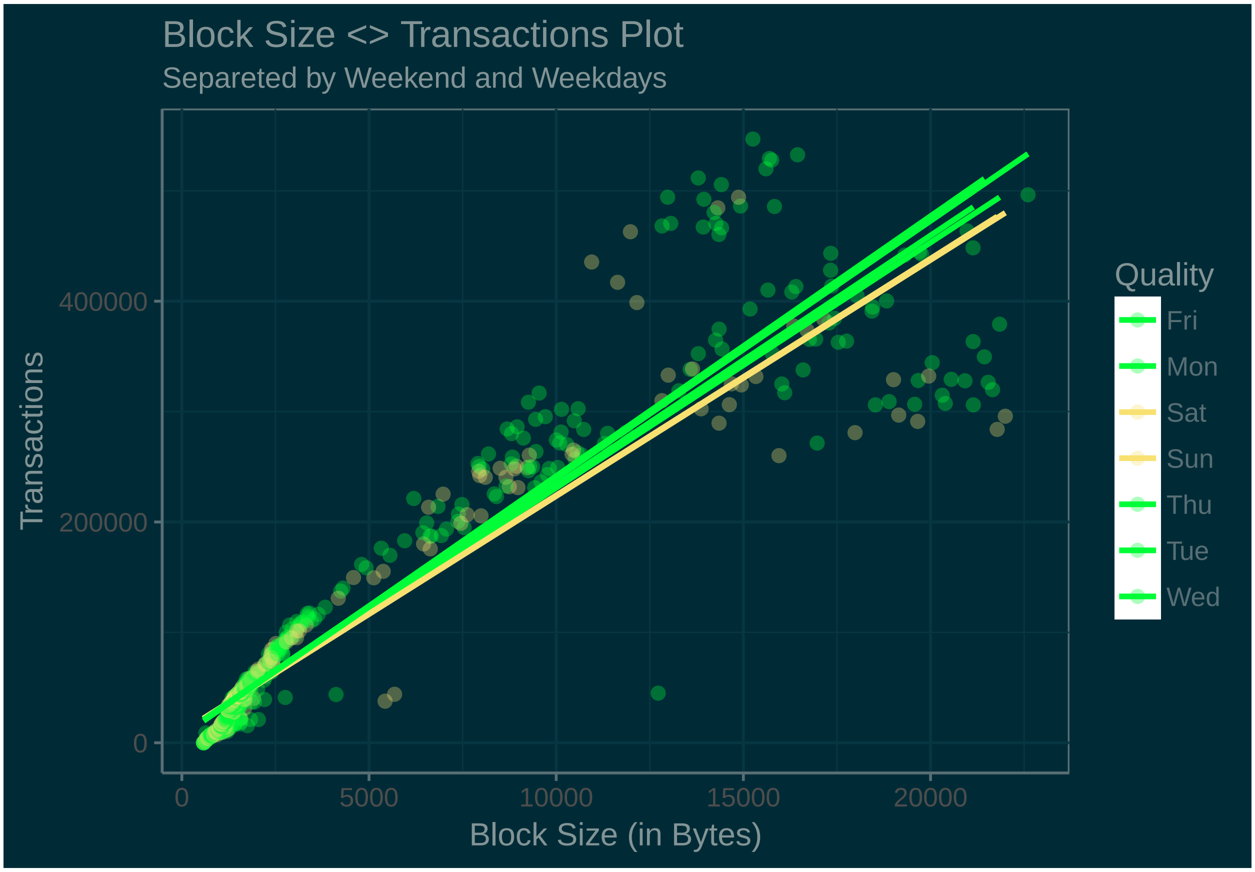
These charts are fewer but required considerable amount of time to decompose. The relationships are a little more nuanced and can hinge on a multiple logical insights.

Block_Size <> Day of the Week

The variable "Day" is unique to this dataset as most asset classes do not trade on the weekend. Ether does and so it is worth seeing how much it is traded on individual days.



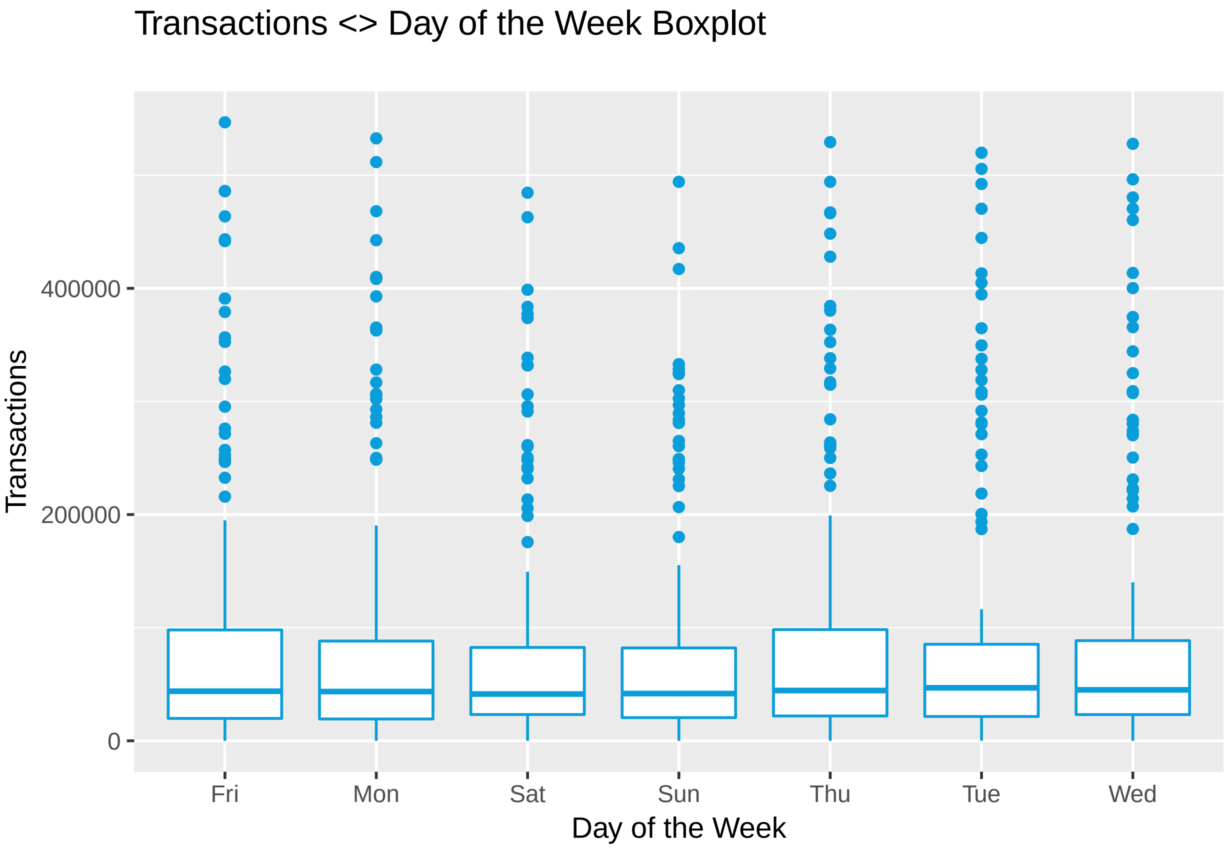
The alpha parameter here is set to 1/5 meaning, there is a distinct blue dot when 5 observations overlap. This is meant to contrast how frequent there are low transactions days vs. high transactions days. However, there it is not convenient having them on separate plots.



The plot above tries to plot the points for the days of the week on the same plane. To highlight more of the contrast, you can play with the background color (solarized in this example) and even assign the same colors to weekdays and weekends (set to alpha of 3 observations per full point). A regression line helps make a clear distinction that the mean on weekend days between the two factors resides beneath that of the weekdays. However, this statistic are fraught because the regression is measuring two variables, instead of just Transactions. Furthermore, the difference between weekends and weekdays is not compelling and visually easy to see. A box plot is a lot simpler and may actually illustrate more in this example.

Transactions <> Day of the Week

This chart simplifies the relationship by removing the Block Size variable and focusing on Transactions.

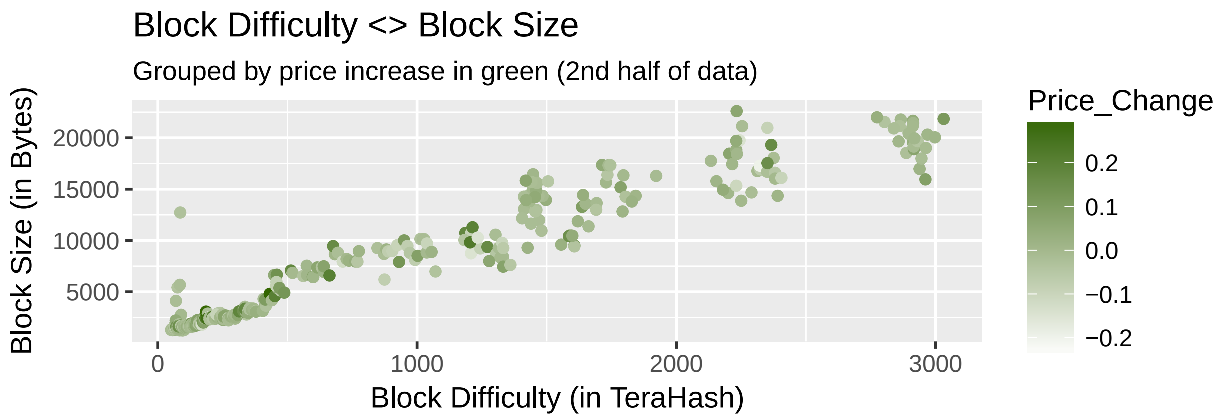


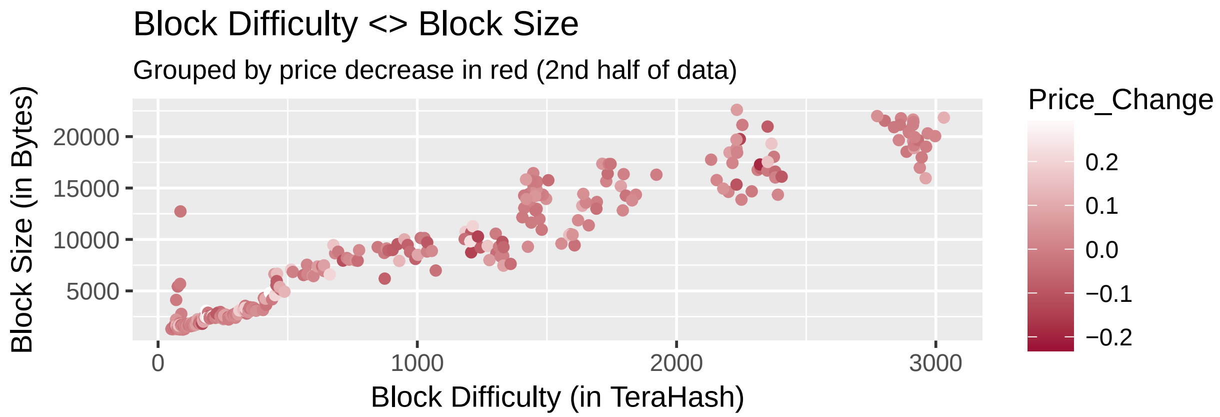
This plot further corroborates 2 things > 1) the 75th quantile is lowest for Saturday and Sunday, and > 2) the highest amount of transactions for both Saturday and Sunday are lower than the highest amount of transactions for all the other days of the week.

We can make other observations about the busiest days being Thursday and Friday based on the 75th quantile points as well as the 2 data points with the most transactions also being on both of these days.

Price Change <> Block Factors

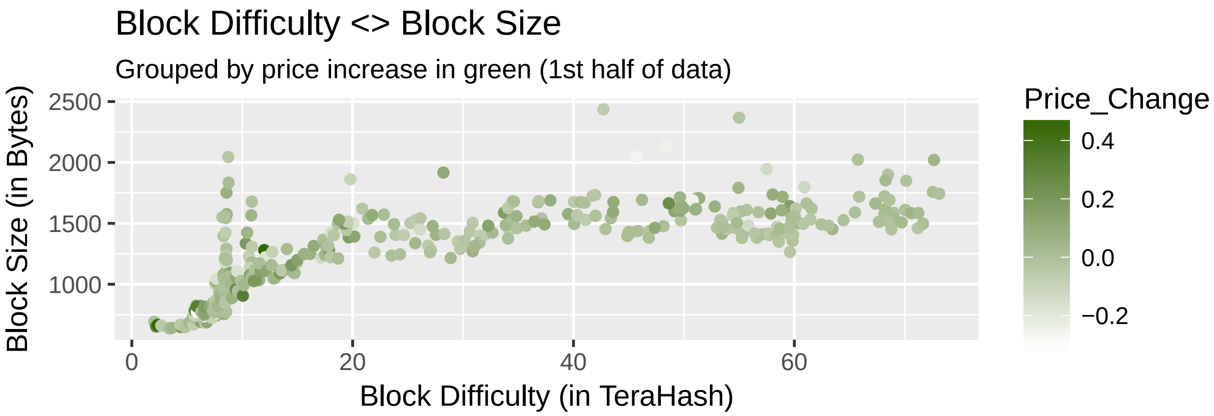
Another factor that may affect volatility is the ability to fulfill a block. To test under what Block creation factors cause the greatest rise and fall in price, one canadd a color scale to both the Block Size and Block Difficult variables. One may see visible differences in looking into a 3rd variable - Price Change in this case. Continuing the subset work done earlier, the first charts start off with the most recent time period (the second half of the non-zero Price time series) and made 2 charts. The one with the green highlights the conditions under which the biggest positive Price_Changes were observed. The one with red highlights the biggest negative Price_Changes.



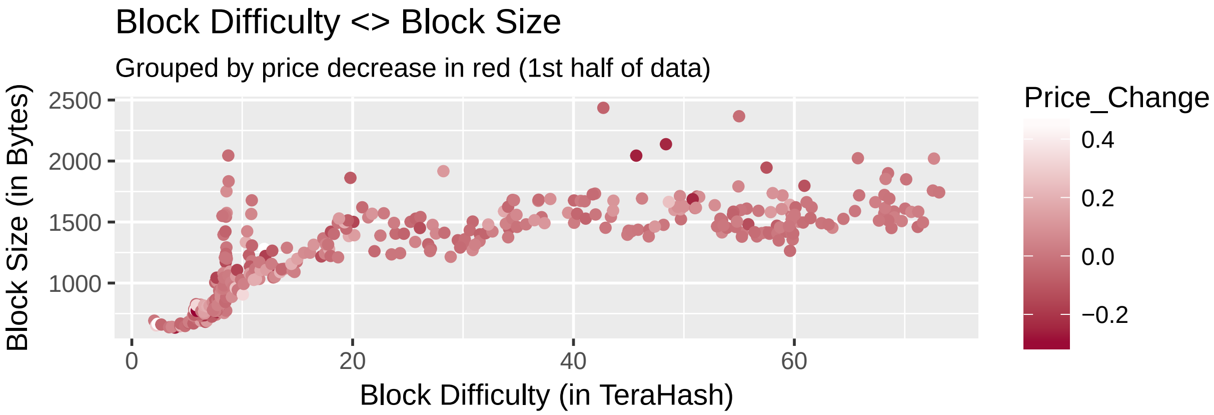


The charts above do not appear to support the thesis that the Price Changes will be muted if there is more activity on the block chain. The darker red data points can be found along the spectrum. It is hard to discern a clear signal from this chart.

Below is the same analysis with the original data set from the 1st half of transactions.


Block Difficulty <> Block Size

Grouped by price increase in green (1st half of data)


Block Difficulty <> Block Size

Grouped by price decrease in red (1st half of data)

This chart above gives the same mixed message about Price Changes based on anytime of Block chain dynamics.

Multivariate Analysis

Talk about some of the relationships you observed in this part of the
investigation. Were there features that strengthened each other in terms of
looking at your feature(s) of interest?

> Partially. There were some features that made perfect sense like # of Transactions to the Day of the week. However, when I considered the Block Size for Day of the week, there appeared to be noise and it didn't make sense in this rendering. By doing too much, it helped me suss out what a clear relationship and remove what was unclear.

Were there any interesting or surprising interactions between features?

> There were clusters of Block Difficulty and Block Size observations that coalesced around regular intervals of Block Difficulty for the 2nd subset of data (the more recent half of data). I couldn't see this in earlier scatter plots. However, once I created the same plot for the 1st half of data, the clusters at regular intervals dispersed. This leads me to the conclusion that the Block Difficulty is a more discrete function in the more recent period, perhaps due to technological protocol. For the 1st half of the data set, the Block Difficulty seemed to evolve in a such a way that there weren't large gaps in the data.
>
> Of note, this observation was a tangential insight from the main relationship I was trying to explore. The unintended discovery of it was one of the reasons why I left this chart as part of the analysis.

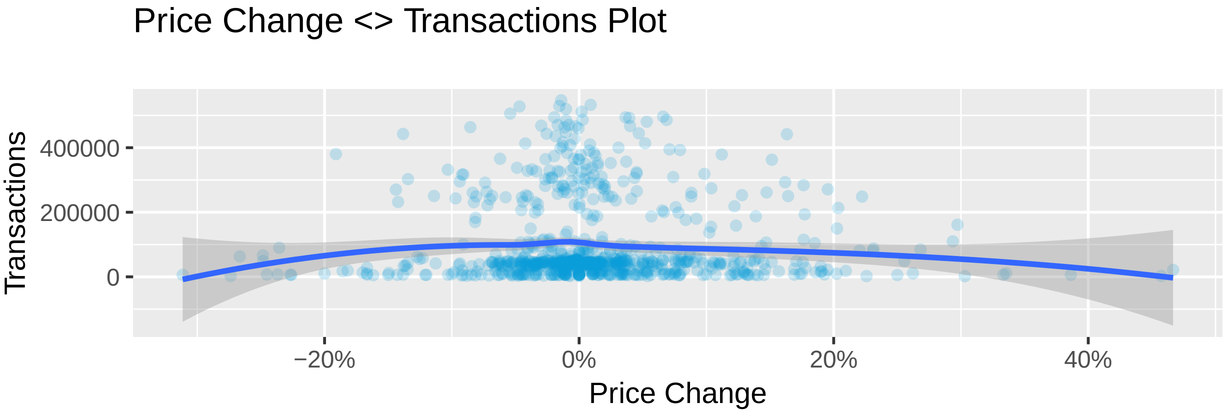
OPTIONAL: Did you create any models with your dataset? Discuss the strengths and limitations of your model.

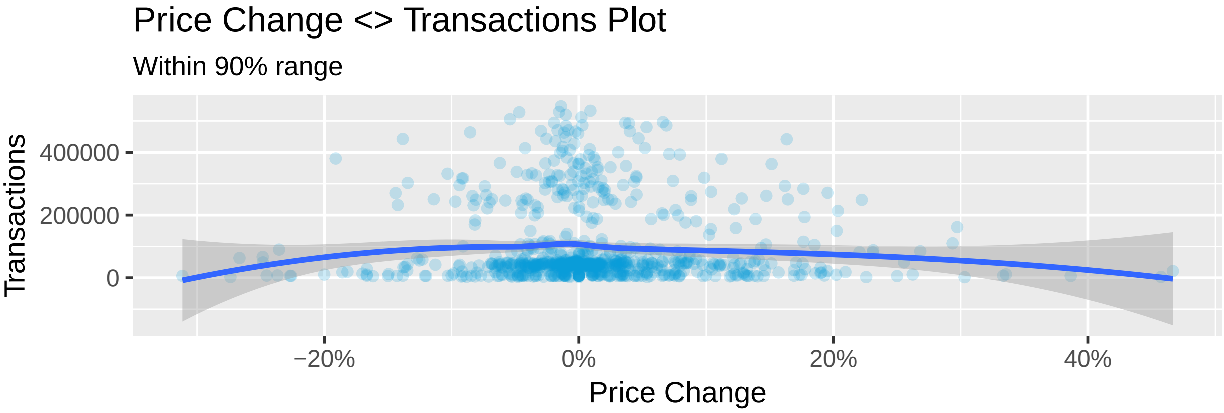
> I tried to expand the model to incorporate the Price_Change variable. This proved to supply me with mixed visualizations that were hard to untangle. Especially when you plot Price Changes, adjusting the alpha parameters creates problems as one large Price Change can look a lot like many small Price Change as they equivocate to the same weight. I tried to create a couple of Multivariate plots with these parameters and ultimately had to remove them.
>
> Going forward, the Block Size and Block Difficulty probably won't be easy to model in terms of Price Change. However, other simple insights like trading on the weekend to look for larger Price Changes are valuable.
>
> I also looked at the Hashrate logic for Block Size for future modeling. I observed a lack ofcorrelation between Hashrate and Block Size as it applies to small block sizes. This is a sort of permissible dynamic for the infancy period of a Cryptoasset as developers work to make a robust blockchain. As the Block Sizes increase, the hashrates adds more complexity and they work hand in hand towards sustaining a healthy, longlasting Cryptoasset. This contextualization of when a Cryptoasset is still in its infancy and when it is mature with a necessary amount of complexity is a key relationship I'd model in the future.

Final Plots and Summary

The plots below are 3 of the most interesting as part of the analysis. In addition to the insights earlier described, I've included some commentary on the rationale of trying to create these plots.

Plot One - Price Change <> Transactions

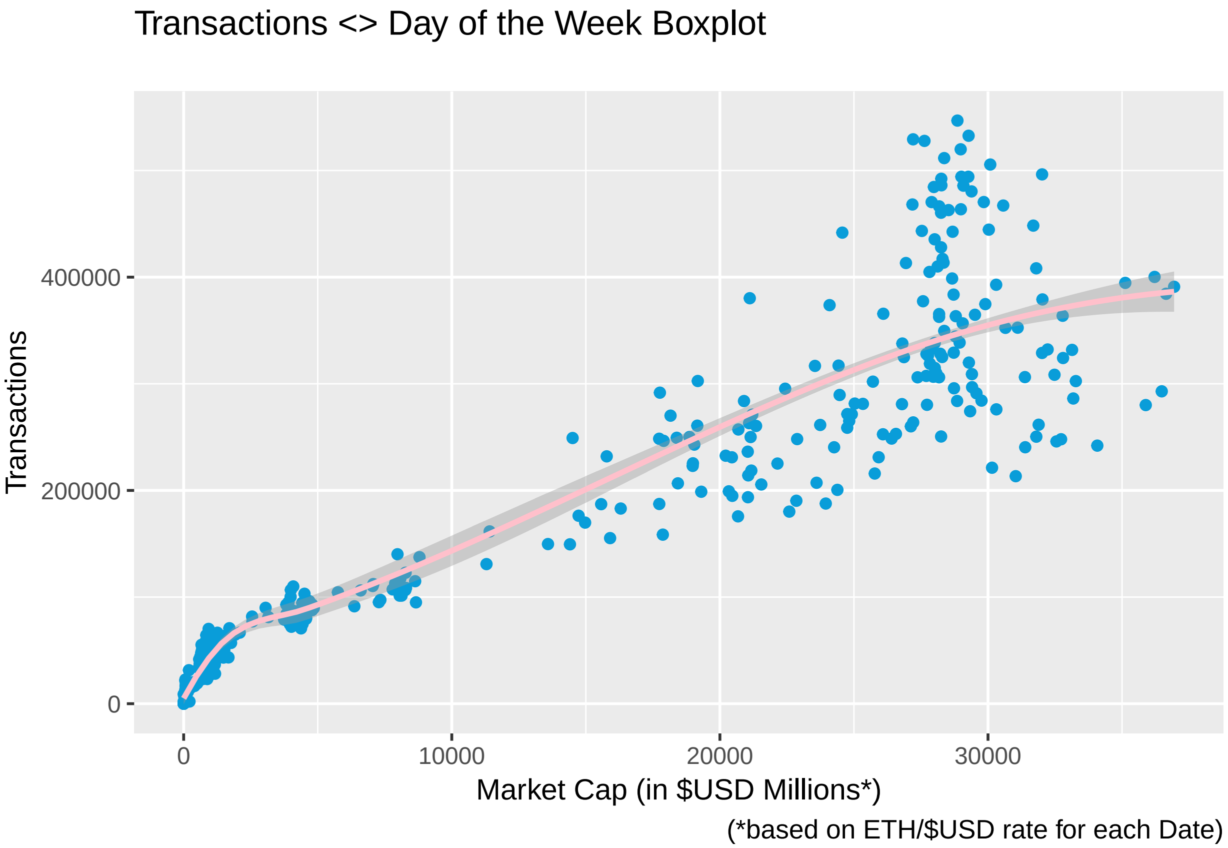




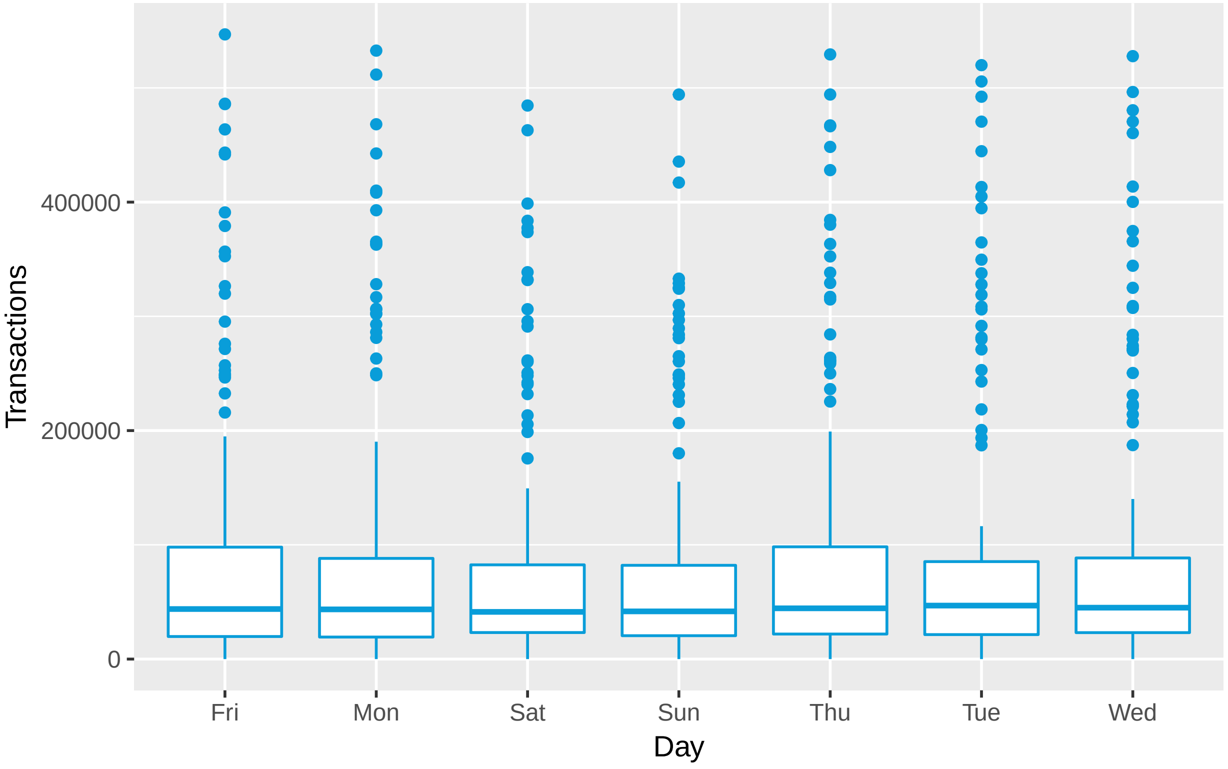
Description One

This chart is unique because there appears to be a horizontal gap where the # of transactions seems to skip. I thought about this for a long time. Eventually, I posit that since there was such an abrupt rise of Transactions for Ether in early Summer of 2017, that there was a quick leap towards new levels of activity. This leap essentially leaves a gap in data because there is such a few amount of days that had Transaction count in this period.

The chart below actually corroborates that when the Market Cap went from approximately 10,000 to 15,000, there were only a week or two of observations. It took time to arrive at this conclusion, but the visualizations helpe framed the thought process.



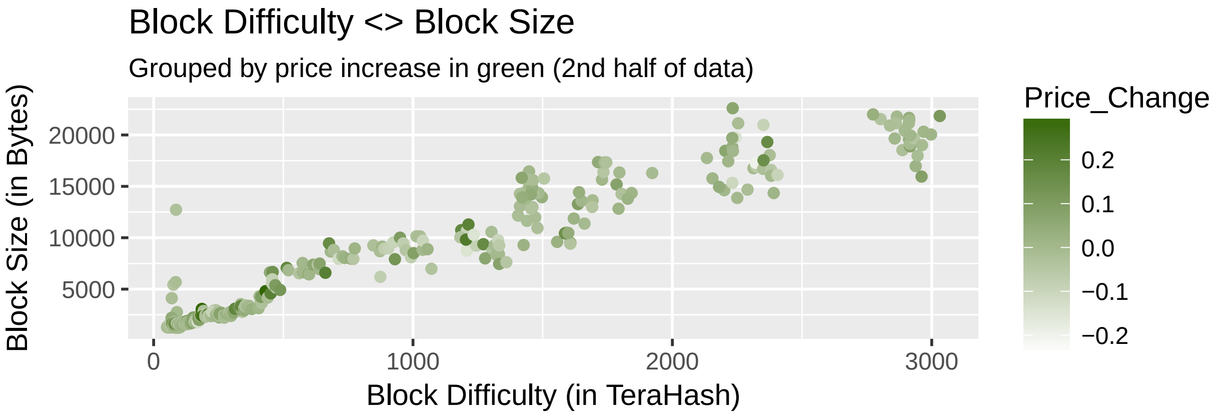
Plot Two - Transactions <> Day of the Week

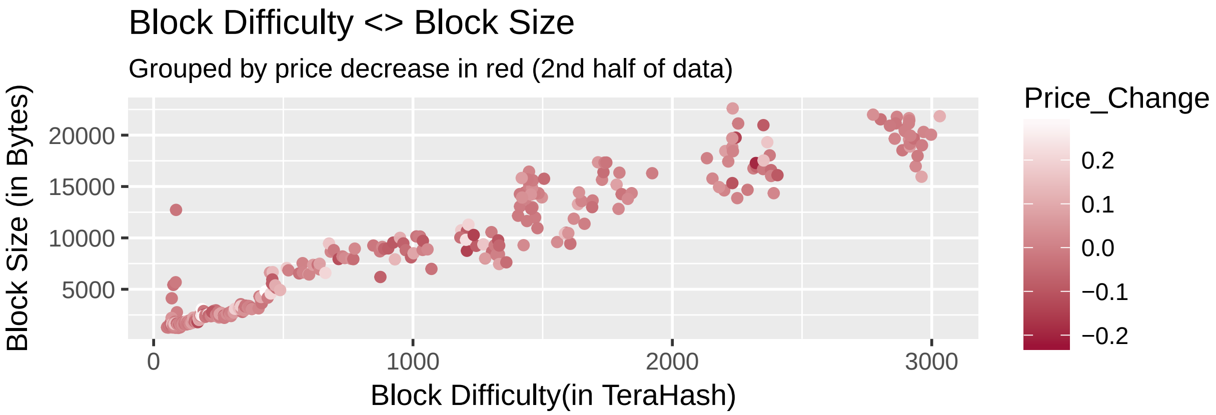


Description Two

This chart was one of the simplest to make but it told an unequivocal message. I initially tried a more nuanced scatter plot, but no relationship was discernible. Once I put it into this quick plot, a relationship was easy to see with the box plots.

Plot Three - Price Change <> Block Factors





Description Three

After adjusting the 3rd variable with many different Geom types (alpha parameters, point_size, continuous or discrete scale for color, etc.), there was no clear signal that was definitive. I decided to keep this analysis in as a testament that you can't force a message if the data doesn't support it.

References

Original Dataset - https://www.kaggle.com/kingburrito666/ethereum-historical-data Current repository of dataset - https://etherscan.io/charts Necessary primer to understanding Cryptoassets - https://www.amazon.com/Cryptoassets-Innovative-Investors-Bitcoin-Beyond/dp/1260026671 Glossary - http://ethdocs.org/en/latest/glossary.html More Technical Glossary - https://bitsonblocks.net/2016/10/02/a-gentle-introduction-to-ethereum/ Price Reference- https://coinmarketcap.com/currencies/ethereum/

Reflection

This analysis is by no means exhaustive and quite the contrary. I enjoyed thinking about this data set and what questions I could answer with it. The exploratory data analysis tested my intuition as I thought I understood some of the more technical variables. While some of the relationships played out as expected, others played out for reasons I couldn't explain. This forced me to look at the technical documentation and blogs to get explanations for some of these variables. I also ended up with more questions than I originally had.

This exercise has started to make me think about the following in particular:

- time series and whether subsetting it more would help going forward.
- the curious dynamics to the Block Size being added to the block chain. I thought there would be more variables that correlate to it, but the only variables with high correlation coefficients that I tested were the Hashrate and Transactions. There has to be more to it as I hypothesized Block Size to be a sort of proxy for complexity or a possible disruptor. I found it be a bit more random than I suspected.
- no actionable variables appeared to be correlated to Price Change with this first exploration. This stands to reason because if there was an obvious relationship, participants would probably act on it and mute its effect by focusing on that relationship. The exploration still gave hints as to which directions may have more clues to a possible variable or relationship.

I foresee updating this analysis every month or so. I am curious to see if these patterns, behavior and relationships hold over time as this data set continues on. Reach out to me at rjl2155(at)Columbia(dot)edu if you have any ideas on what I may be missing.