Final Report for FE 541

Twitter Influence on Crypto Currencies

- Akshat Goel
- Jugal Vaidya
- Sishir Yerra Doddi

Preliminary Research Question

Demand for cryptocurrency is growing worldwide with the hope of earning unprecedented profits. It can be argued that, in the future cryptocurrencies can be used as alternative payment channel along with the fiat currency. We believe that cryptocurrencies such as Bitcoin, Dogecoin, or Ethereum are drastically being influenced by social media platforms such as Twitter, Reddit, or private Discord groups. For instance, Elon Musk, the richest man in the world, tweeted "#bitcoin," which influenced bitcoin market capitalization by $110 million in a matter of hours. We believe that cryptocurrency is controlled by a massive group of tweets, not just by one. We have decided to research the impact of social media platforms on the Cryptocurrency market. To verify our hypothesis, we have decided to test a sample of data that could be used to generalize pricing behavior for all cryptocurrencies in general.

For this project, we will research effects of Twitter activity on Ethereum and predict how Ethereum prices are influenced over time. We also want to look at the different lag periods to verify social media's trailing effect on the price of Cryptocurrency.

Q1: What influences cryptocurrency prices?

Q2: What is the correlation between social media mentions of cryptocurrency and its price?

Relevant Literature

- Application of Causality detection on social media sentiment on cryptocurrency prices[Link]

  The paper studies statistical methods for causality detection between time series. Then studies the application of the measures to detect statistical causality between social sentiment changes and cryptocurrency returns.

- Understanding the social factors affecting the Crypto Currency Market[Link]

  The paper studies the impact of social and government factors on the crypto currency market. This paper identifies and discusses the important factors that govern the cryptocurrency market and analyzes the impact of these factors. A pilot user survey has also been presented at the end of this paper to understand and demonstrate the societal view of the acceptance of cryptocurrencies.

Data

Data Sources

- The data for this project was collected by creating a web scraper in python which used twitter API to download tweets which mentioned Ethereum
- The cryptocurrency prices was downloaded from CoinMarketCap.com

Data Description

Y: Ethereum Daily Close Price

X1: Number of Daily Tweets mentioning Ethereum

X2: Number of Retweets Daily mentioning Ethereum

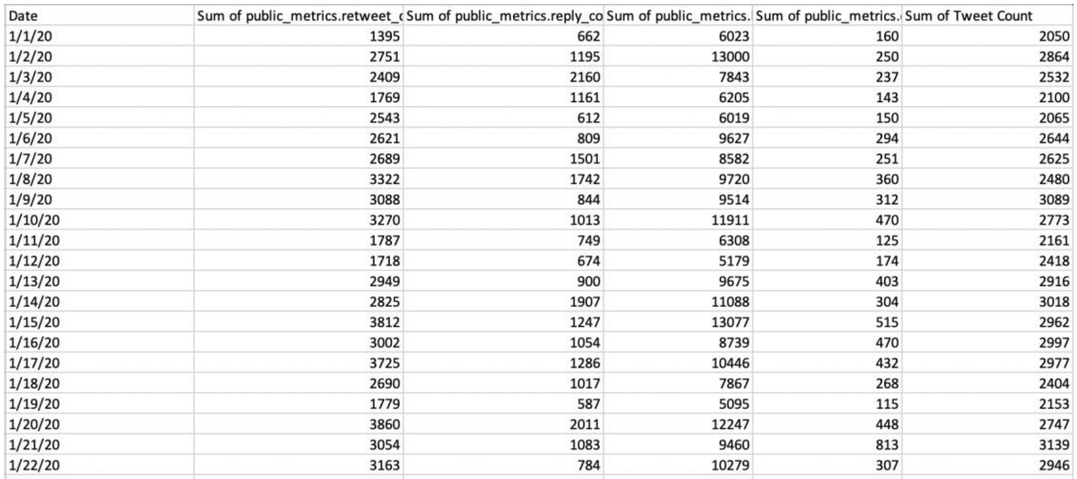
X3: Number of Replies Daily on tweets mentioning Ethereum

X4: Number of likes on tweets mentioning Ethereum

X5: Number of quotes on tweets mentioning Ethereum

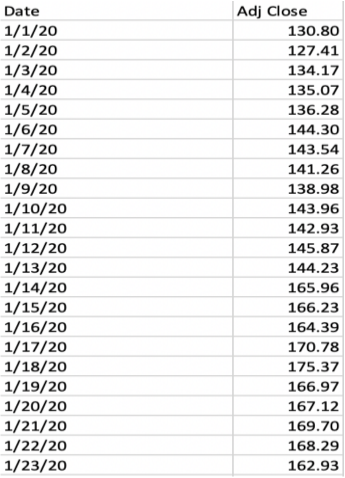
Data Snapshot

We analyzed CryptoCurrency Ethereum for this project. Ethereum prices and tweets mentioning ethereum were downloaded. Below is a snapshot of consolidated twitter data and daily ethereum prices.



Date	Sum of public_metrics.retweet_	Sum of public_metrics.reply_co	Sum of public_metrics.	Sum of public_metrics.	Sum of Tweet Count
1/1/20	1395	662	6023	160	2050
1/2/20	2751	1195	13000	250	2864
1/3/20	2409	2160	7843	237	2532
1/4/20	1769	1161	6205	143	2100
1/5/20	2543	612	6019	150	2065
1/6/20	2621	809	9627	294	2644
1/7/20	2689	1501	8582	251	2625
1/8/20	3322	1742	9720	360	2480
1/9/20	3088	844	9514	312	3089
1/10/20	3270	1013	11911	470	2773
1/11/20	1787	749	6308	125	2161
1/12/20	1718	674	5179	174	2418
1/13/20	2949	900	9675	403	2916
1/14/20	2825	1907	11088	304	3018
1/15/20	3812	1247	13077	515	2962
1/16/20	3002	1054	8739	470	2997
1/17/20	3725	1286	10446	432	2977
1/18/20	2690	1017	7867	268	2404
1/19/20	1779	587	5095	115	2153
1/20/20	3860	2011	12247	448	2747
1/21/20	3054	1083	9460	813	3139
1/22/20	3163	784	10279	307	2946

Fig. Snapshot of Twitter Data



Date	Adj Close
1/1/20	130.80
1/2/20	127.41
1/3/20	134.17
1/4/20	135.07
1/5/20	136.28
1/6/20	144.30
1/7/20	143.54
1/8/20	141.26
1/9/20	138.98
1/10/20	143.96
1/11/20	142.93
1/12/20	145.87
1/13/20	144.23
1/14/20	165.96
1/15/20	166.23
1/16/20	164.39
1/17/20	170.78
1/18/20	175.37
1/19/20	166.97
1/20/20	167.12
1/21/20	169.70
1/22/20	168.29
1/23/20	162.93

Fig. Snapshot of Ethereum Prices

Graphical Representation

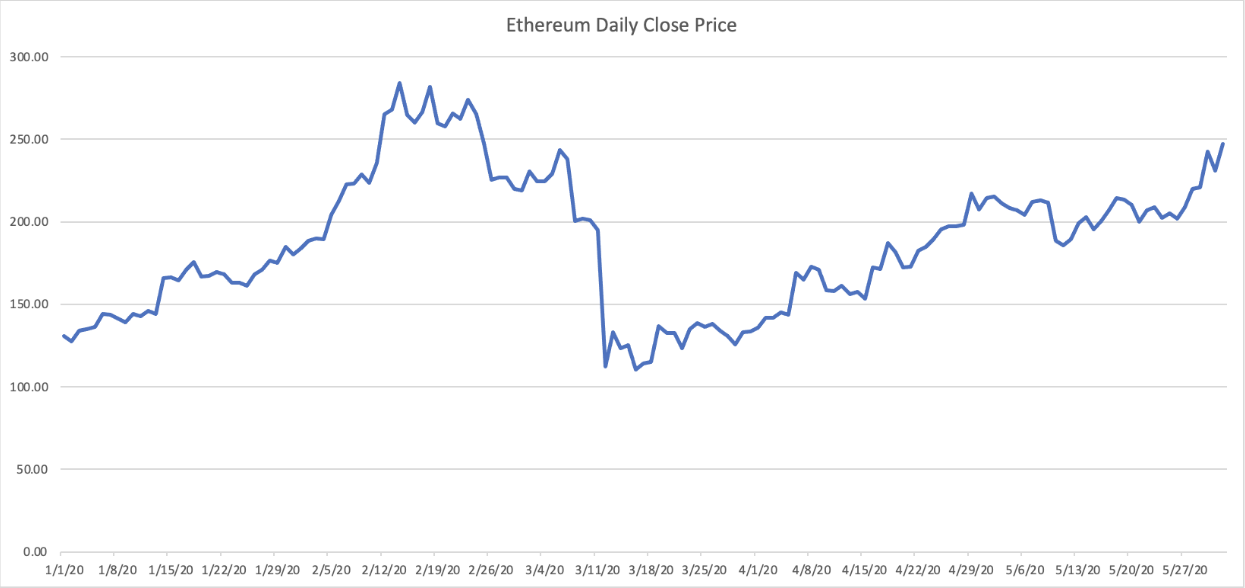


Fig. Graphical Representation of Ethereum Daily Price

Methodology

We followed the following steps for Multiple Linear Regression:

- Basic scatter plot
- Initial Model fit
- Model diagnostic
- Outliers Detection
- Model fit results after power transformation
- Model diagnostic comparison before and after transformation
- Multicollinearity Check
- Variable selection
- Added Variable Plots
- Partial F-Test
- Marginal Model Plots

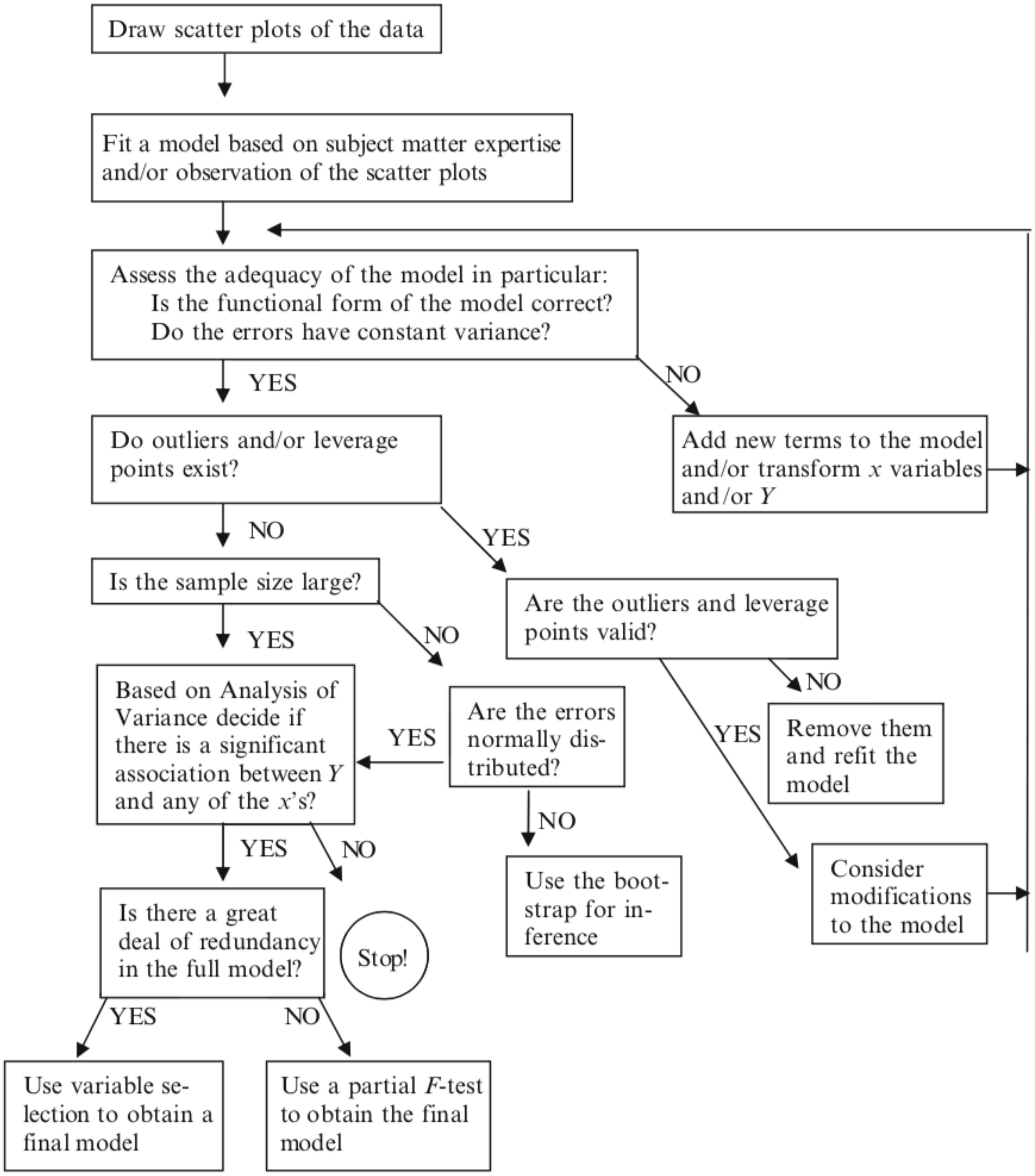
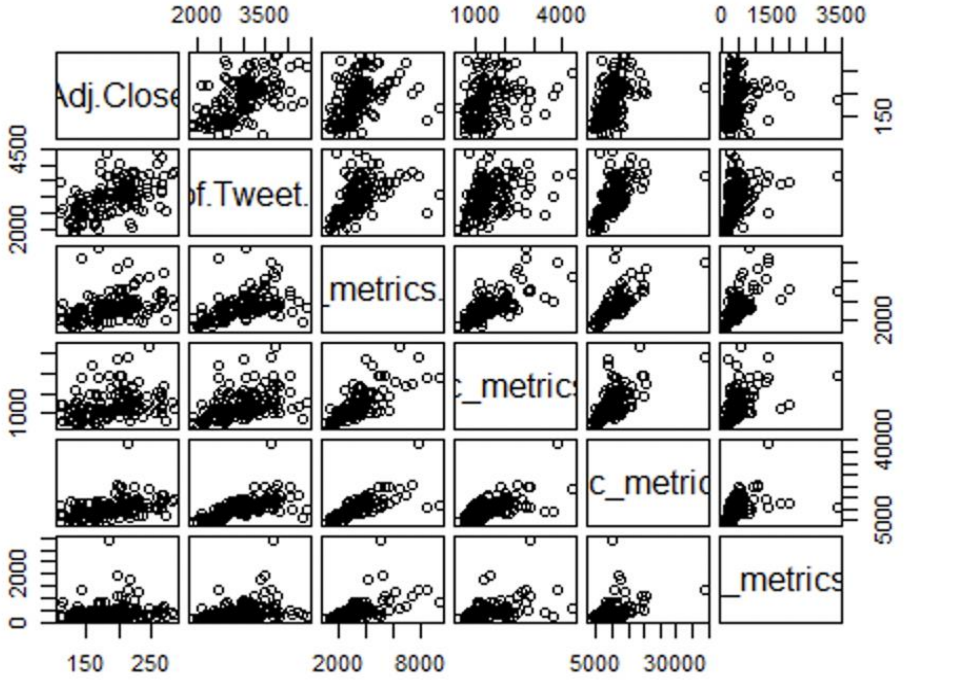

Draw scatter plots of the data
Fit a model based on subject matter expertise and/or observation of the scatter plots
Assess the adequacy of the model in particular:
Is the functional form of the model correct?
Do the errors have constant variance?
YES
NO
Add new terms to the model and/or transform x variables and/or Y
Do outliers and/or leverage points exist?
NO
YES
Is the sample size large?
YES
NO
Are the outliers and leverage points valid?
NO
Remove them and refit the model
YES
Consider modifications to the model
Based on Analysis of Variance decide if there is a significant association between Y and any of the x's?
YES
Are the errors normally distributed?
NO
Use the boot-strap for inference
YES
NO
Is there a great deal of redundancy in the full model?
Stop!
YES
NO
Use variable selection to obtain a final model
Use a partial F-test to obtain the final model

Fig. Decision Making Process for Multiple Linear Regression
Source: A Modern Approach to Regression with R, Simon J. Sheather

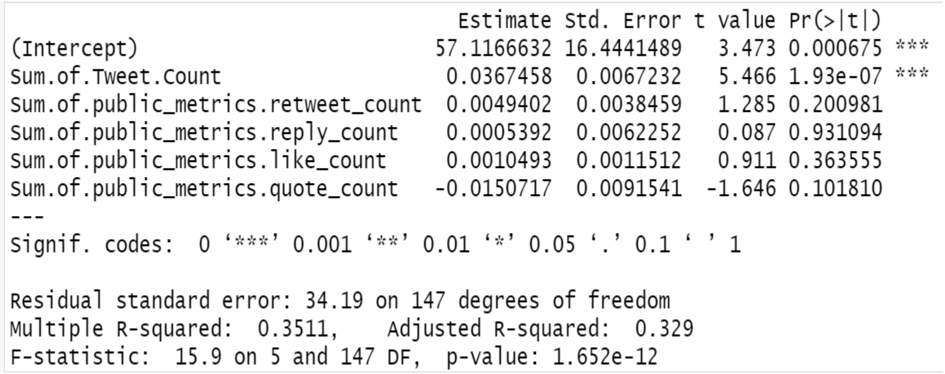
Scatter Plot



It is visible from scatter plots that-

- The data here is highly concentrated
- Some of the X variables are have correlation between themselves

Initial Model Fit

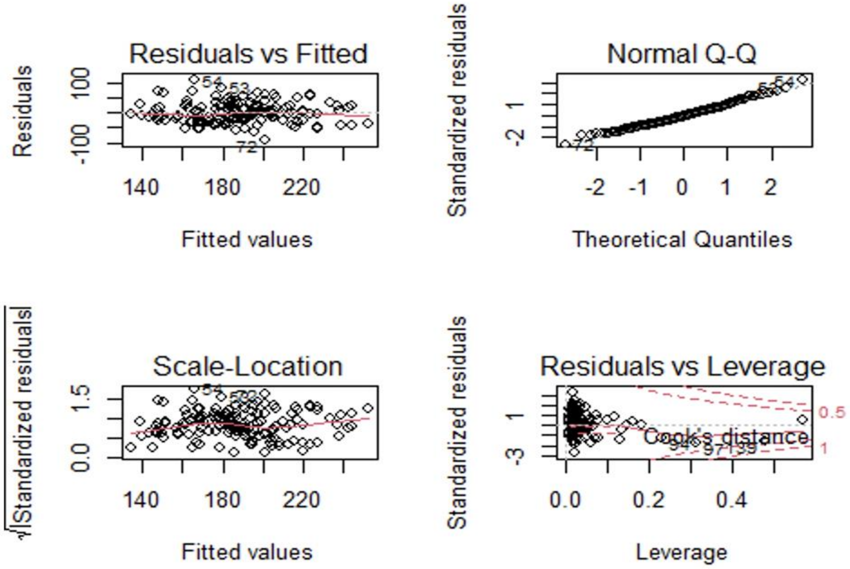


```
                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                        57.1166632 16.4441489   3.473 0.000675 ***
Sum.of.Tweet.Count                  0.0367458  0.0067232   5.466 1.93e-07 ***
Sum.of.public_metrics.retweet_count  0.0049402  0.0038459   1.285 0.200981    
Sum.of.public_metrics.reply_count    0.0005392  0.0062252   0.087 0.931094    
Sum.of.public_metrics.like_count     0.0010493  0.0011512   0.911 0.363555    
Sum.of.public_metrics.quote_count   -0.0150717  0.0091541  -1.646 0.101810    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.19 on 147 degrees of freedom
Multiple R-squared:  0.3511,	Adjusted R-squared:  0.329 
F-statistic:  15.9 on 5 and 147 DF,  p-value: 1.652e-12
```

Regression results of the original model with all variables

- Only one variable among all the variables is significant here, which is Tweet Counts.
- Also, only 33% of the variability in the Ethereum Price is being explained by this model.

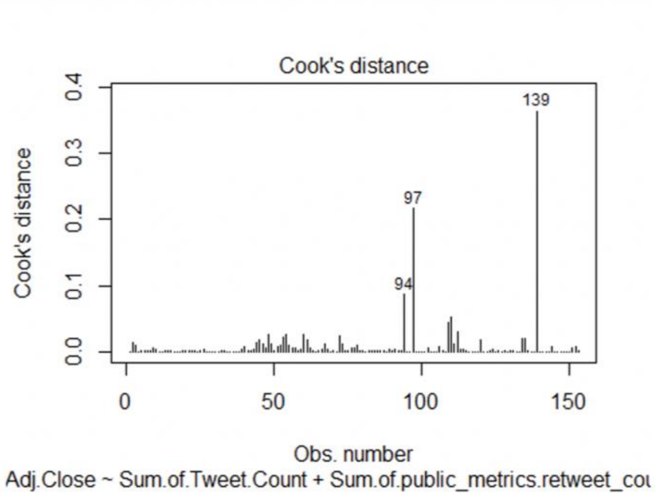
Model Diagnostic



We will verify the assumptions of linear model here-

- Looking at the Scale-Location plot, we can see that the red line is not exactly horizontal, therefore error terms might not be independently distributed here
- Also, there are some outliers which are visible though Residuals vs Leverage plot

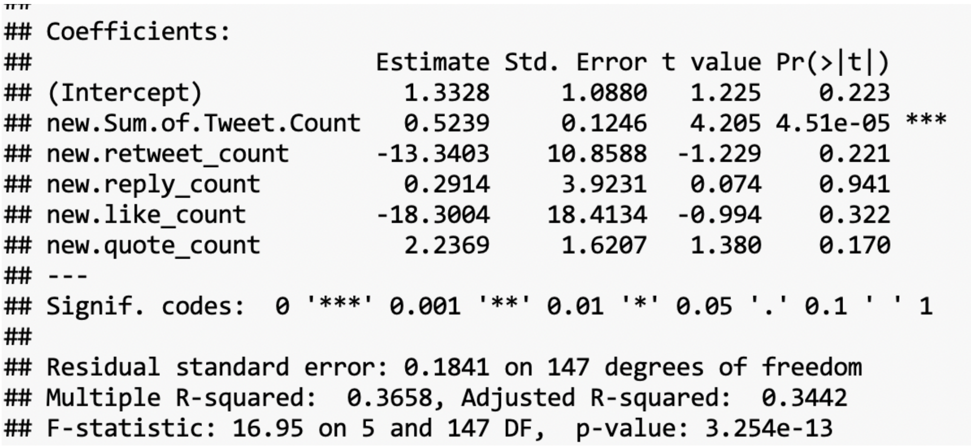
Outliers



Dealing with the outliers-

- There are three outliers in our dataset. They are the influential points; hence affect the results of our model.
- We did not have any solid justification to remove these outliers. Therefore, we have not removed them from the model.

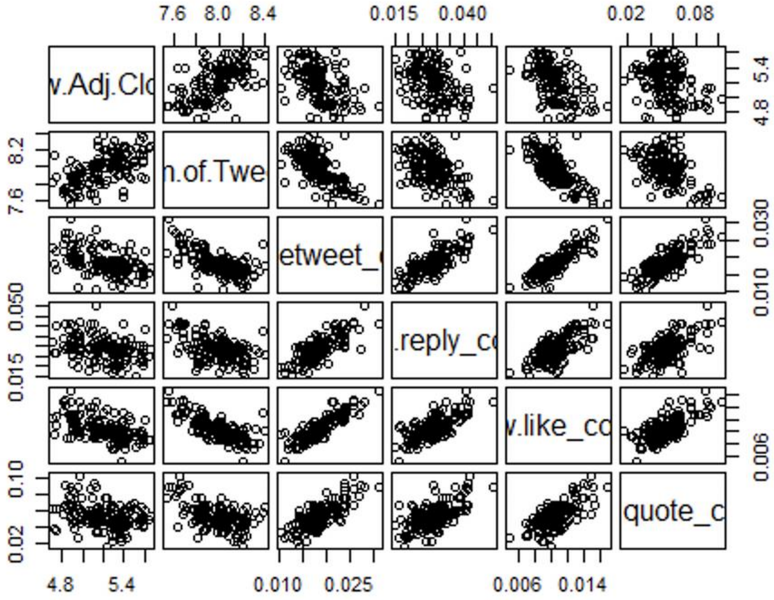
Model Fit results after Power Transformation



```
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)    
## (Intercept)               1.3328     1.0880   1.225    0.223    
## new.Sum.of.Tweet.Count    0.5239     0.1246   4.205 4.51e-05 ***
## new.retweet_count       -13.3403    10.8588  -1.229    0.221    
## new.reply_count           0.2914     3.9231   0.074    0.941    
## new.like_count          -18.3004    18.4134  -0.994    0.322    
## new.quote_count           2.2369     1.6207   1.380    0.170    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1841 on 147 degrees of freedom
## Multiple R-squared:  0.3658, Adjusted R-squared:  0.3442 
## F-statistic: 16.95 on 5 and 147 DF,  p-value: 3.254e-13
```

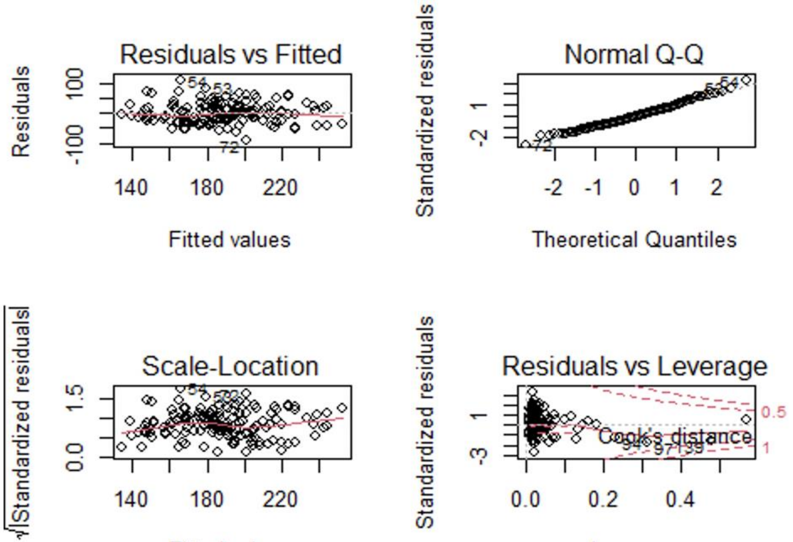
- Here, we have transformed the variables using the lambda values which we have obtained using Power Transform function.
- Even after applying the transformation, only the Tweet Count variable is significant and there is not much improvement in R-squared.

Scatter Plot after Power Transformation

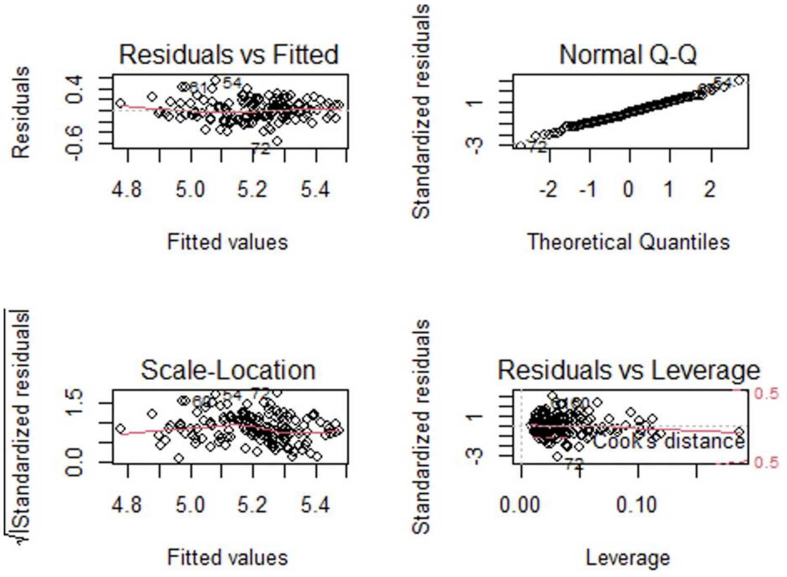


- After applying the Transformation we can see that the data is not as concentrated as before and it seems to be spread out.
- Issue of Multicollinearity is more visible now.

- **Model Diagnostic Comparison**



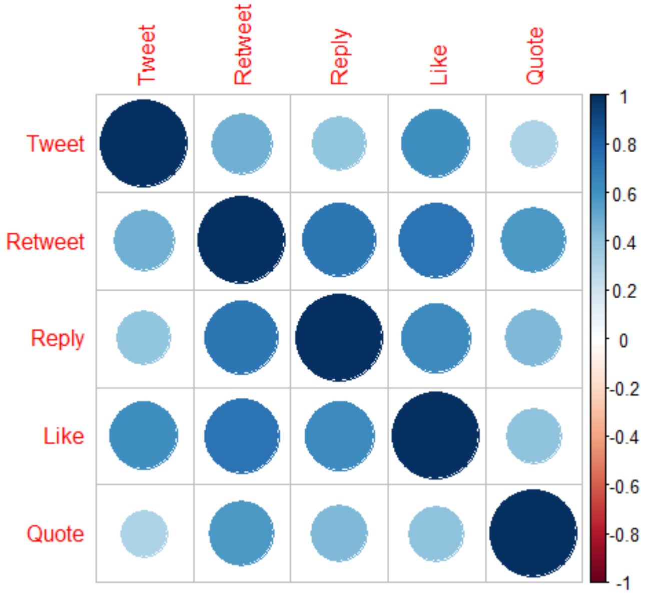
Before Power Transformation



After Power Transformation

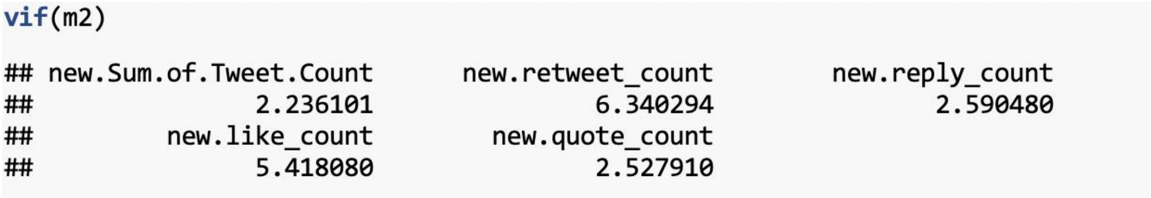
- We can see some improvement in the Scale-Location plot. The red line is more horizontal after transformation, indicating that assumption of non-constant variance holds. Also, Residuals vs Leverage plot has improvement in terms of outliers.

Model Building- Multicollinearity



- From the Correlation Plot, it is clear that many of the X-variables are correlated.
- Some of the coefficients of X-variables were negative and standard error for those Coefficients was also high very high. Therefore, it is certainly a case of multicollinearity and we should consider dropping some of the correlated variables.

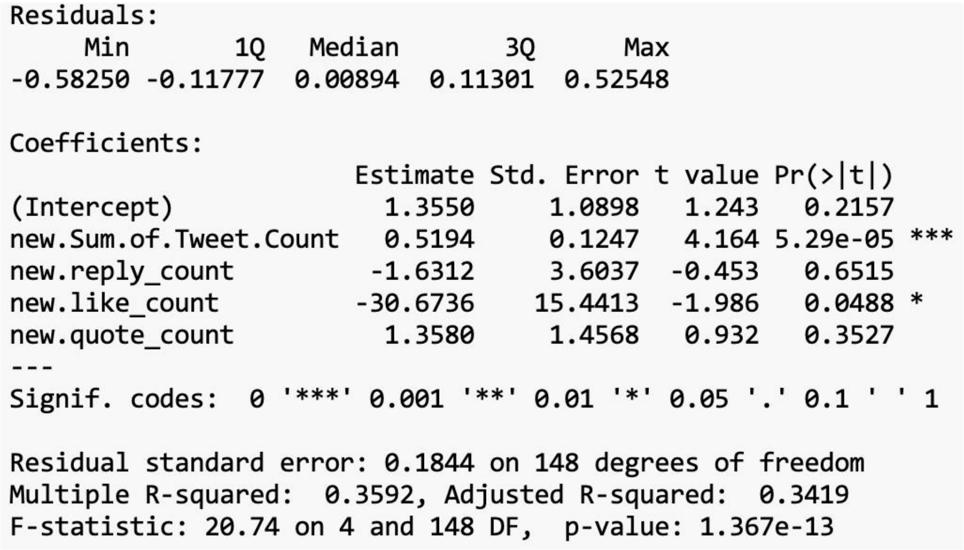
Variable Selection



```
vif(m2)

## new.Sum.of.Tweet.Count       new.retweet_count         new.reply_count
##               2.236101                6.340294                2.590480
##         new.like_count         new.quote_count
##               5.418080                2.527910
```

Cutoff for removing the variable in the case of VIF is 5. Therefore, we will drop the variable ‘Retweet Count’ and run the model again.

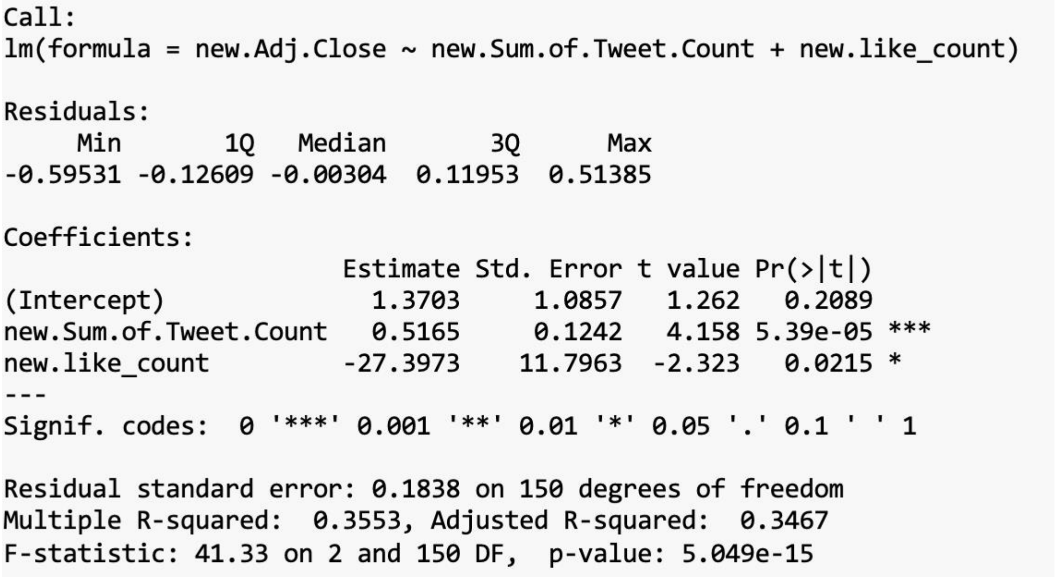
Model results after dropping ‘Retweet Count’ Variable.



```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.58250 -0.11777  0.00894  0.11301  0.52548 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)              1.3550     1.0898   1.243   0.2157    
new.Sum.of.Tweet.Count   0.5194     0.1247   4.164 5.29e-05 ***
new.reply_count         -1.6312     3.6037  -0.453   0.6515    
new.like_count         -30.6736    15.4413  -1.986   0.0488 *  
new.quote_count          1.3580     1.4568   0.932   0.3527    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1844 on 148 degrees of freedom
Multiple R-squared:  0.3592,	Adjusted R-squared:  0.3419 
F-statistic: 20.74 on 4 and 148 DF,  p-value: 1.367e-13
```

- In this new model, there are two independent variables which are significant and R-squared is almost the same.

Model fit results after using AIC

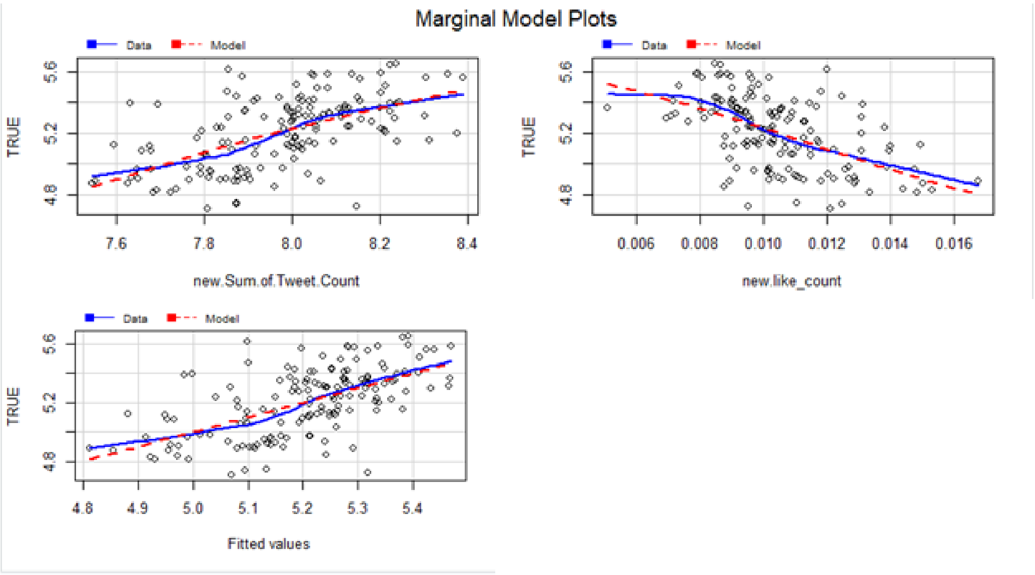


```
Call:
lm(formula = new.Adj.Close ~ new.Sum.of.Tweet.Count + new.like_count)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.59531 -0.12609 -0.00304  0.11953  0.51385 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)              1.3703     1.0857   1.262   0.2089    
new.Sum.of.Tweet.Count   0.5165     0.1242   4.158 5.39e-05 ***
new.like_count         -27.3973    11.7963  -2.323   0.0215 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1838 on 150 degrees of freedom
Multiple R-squared:  0.3553,	Adjusted R-squared:  0.3467 
F-statistic: 41.33 on 2 and 150 DF,  p-value: 5.049e-15
```

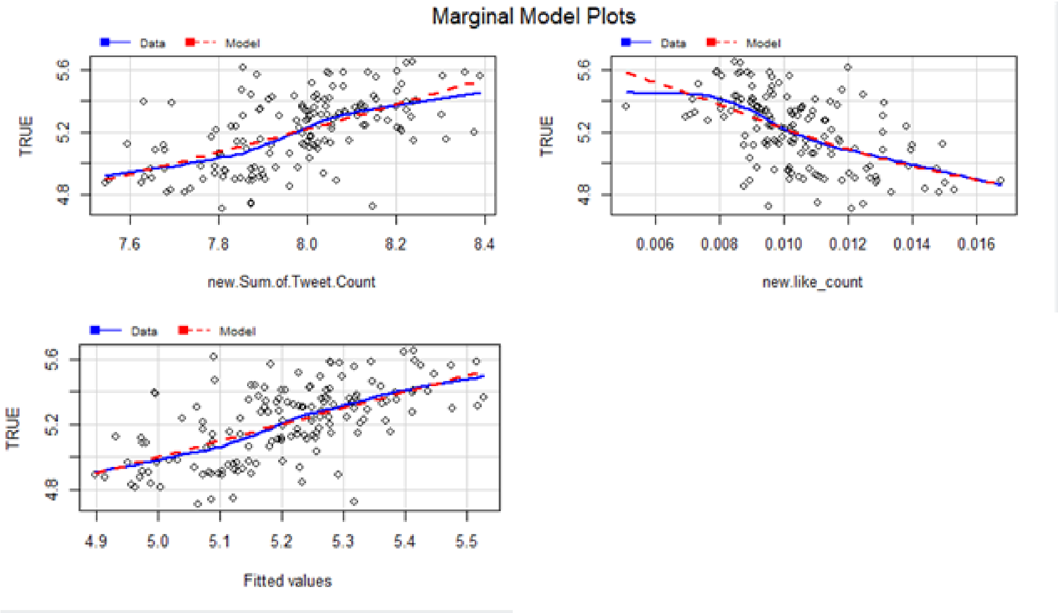
- We used AIC to get rid of redundant x variables and come up with a simpler model.

Marginal Model Plots

Marginal Model Plots before adding the interaction term

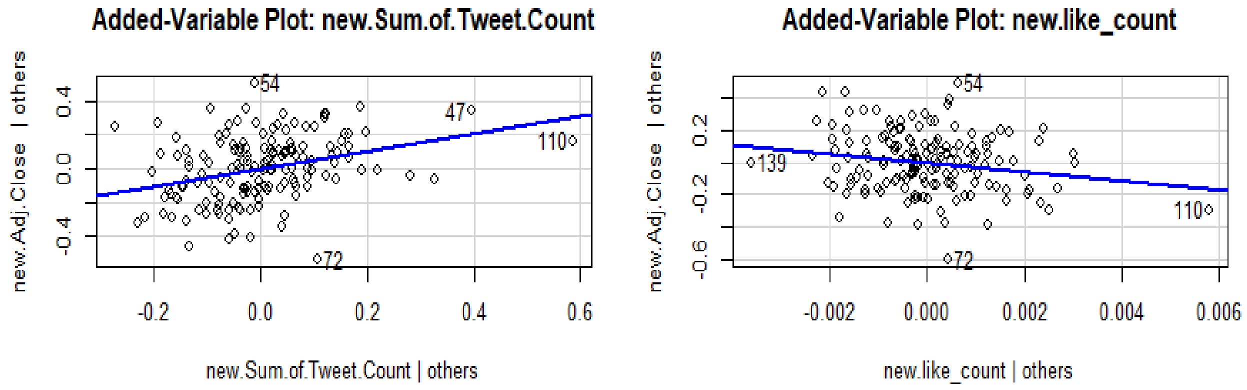


Marginal Model Plots after adding the interaction term



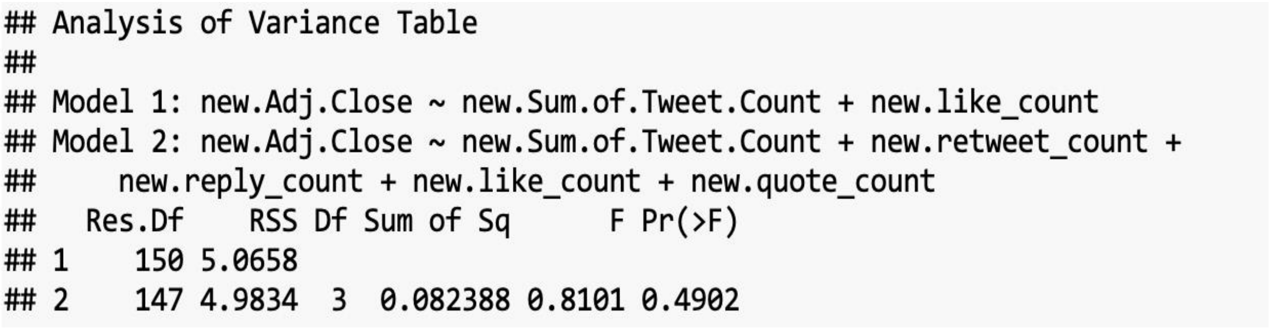
- We can observe from the graph of Y vs Y-hat, the solid curve (nonparametric regression model) is matching well with the dashed curve (the linear model).

Added Variable Plots



- This shows the relationship between the Ethereum price and one of the predictors in the regression model, after controlling for the presence of the other predictors. The blue line is not horizontal in both the graphs. Therefore, both the variables are significant in presence of other variables.

Model Comparison: Partial F-Test



```
## Analysis of Variance Table
##
## Model 1: new.Adj.Close ~ new.Sum.of.Tweet.Count + new.like_count
## Model 2: new.Adj.Close ~ new.Sum.of.Tweet.Count + new.retweet_count +
##     new.reply_count + new.like_count + new.quote_count
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    150 5.0658
## 2    147 4.9834  3  0.082388 0.8101 0.4902
```

- Given the P-value of 0.4902, we cannot support the alternate hypothesis (opting for the full model). Therefore, we are going to adopt the reduced model, which is model 1 in the analysis of variance table.

Conclusion

Despite knowing that social media influence impacts Cryptocurrency prices, our model did not have a significant prediction power. Though both of the variables in the final model are significant, the prediction power in the results is much lower than we expected. The results are strange, but we believe that adding other platforms like Facebook, Reddit, Discord could improve the prediction power.

Limitations

- Limited social media platforms were used for analysis. Gathering data for other platforms like Facebook, Reddit and discord is not technically feasible.
- Frequency of twitter data was low. We did not have consistent high frequency data available for twitter.

Future Work

- Other viral cryptocurrencies like Dogecoin could have better results as social media heavily influences their prices
- As we are using a time series data, adding lagged variables could have better results
- We can also try classification and predict the direction of cryptocurrencies price movement rather than predicting the magnitude. Here, we can try applying Random forest, SVM etc.