Multiple Linear Regression

Regression Analysis of Factors Affecting Price of Bitcoin

Mrudul Banka

Contents

I. Proposal:

Description of the Problem:

The model presented in the proposal aims to evaluate fluctuations in the price points of bitcoin. The above model helps us understand the nature of Bitcoin (cryptocurrency). Main link to understand the data: https://coinmetrics.io/on-data-and-certainty/.

Description of variables:

We use Multiple Linear Regression Analysis in which we determine the statistical relation between the variables presented. The study provides us with the relation between the response (dependent) variable (Y) and predictor(independent) variables (Xs). The table below shows us the abbreviations for variables used to test linear relationship with Bitcoin Price.

Y	Dependent (Response variable)	Price (USD)
X_1	Independent (Predictor variable)	Exchange Volume (USD)
X_2	Independent (Predictor variable)	Fees (BTC)
X_3	Independent (Predictor variable)	Total Count
X_4	Independent (Predictor variable)	Coins Generated

Matrix scatter plot of the variables and Pearson Correlation Coefficients:

Pearson Correlation Coefficients, N = 180					
	price_USD_	**exchangeVolume_USD_**	**generatedCoins**	**fees**	**txCount**
price_USD_ price(USD)	1.00000	0.91851	0.22639	0.71483	0.54815
exchangeVolume_USD_ exchangeVolume(USD)	0.91851	1.00000	0.17794	0.78383	0.49498
generatedCoins generatedCoins	0.22639	0.17794	1.00000	-0.01929	0.43893
fees fees	0.71483	0.78383	-0.01929	1.00000	0.42858
txCount txCount	0.54815	0.49498	0.43893	0.42858	1.00000

SAS Output 1: Pairwise correlation matrix of the dataset.

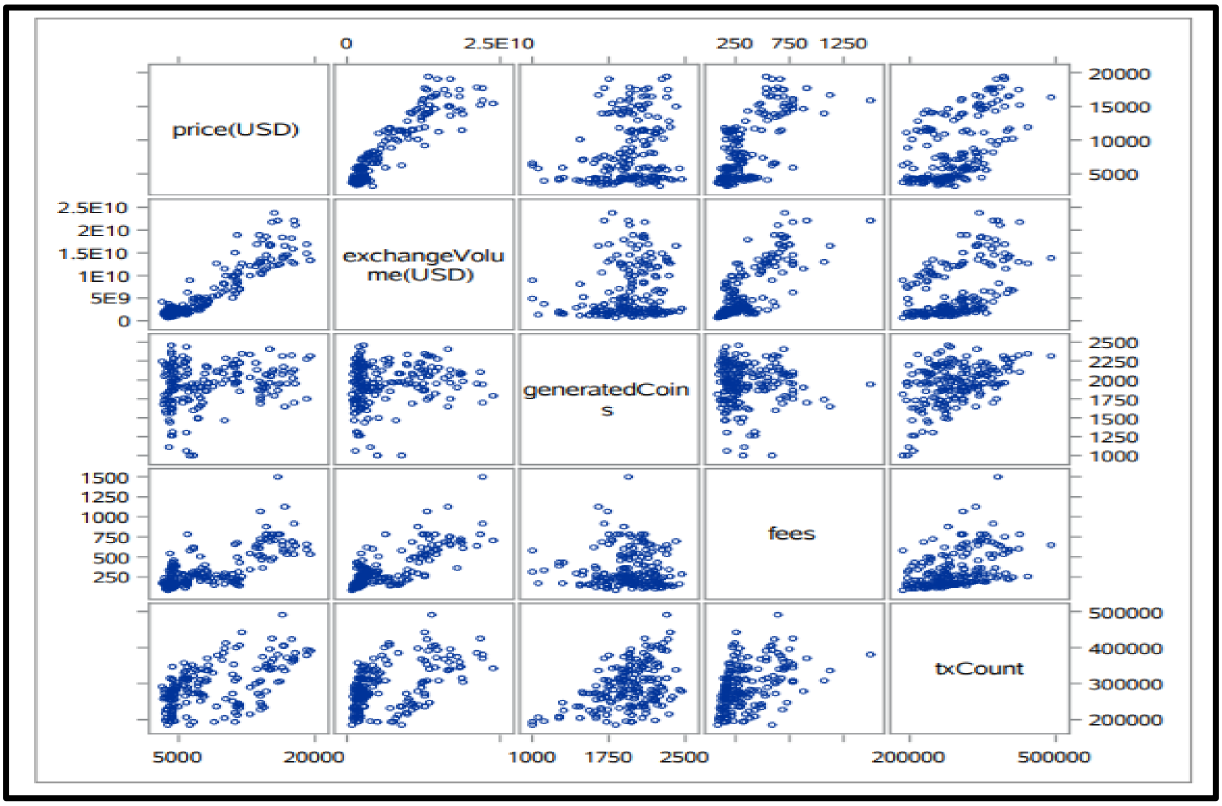

Figure 1: Matrix Scatter Plot of Variables.

Response Variable Vs Predictor Variable: Price of Bitcoin (USD) vs Exchange Volume (USD), We observe an upward trend and high correlation between the response variable Price and the predictor variable Exchange volume with a correlation value of 0.91851 (SAS Output 2). Price of Bitcoin (USD) vs Generated coins, we observe no trend and a low correlation between the response variable Price and the predictor variable generated coins with a correlation value of 0.22639(SAS Output 2). Price of Bitcoin (USD) vs fees for coin generation, we observe an upward trend and high correlation between the response variable Price and the predictor variable fees for coin generation with a correlation value of 0.71483(SAS Output 2). Price of Bitcoin (USD) is the response variable vs Total count of transactions, we observe moderate correlation between the response variable Price and the predictor variable total count of transactions with a correlation value of 0.54815(SAS Output 2).

Predictor Variable Vs Predictor Variable: Price Exchange Volume (USD) v/s Generated Coins, trend is not visible with correlation value of 0.17794(SAS Output 2). Therefore, it is useful to add Generated Coins as a predictor variable. No serious multicollinearity is observed. Exchange Volume (USD) v/s Fees for coin generation, upward trend is visible with correlation value of 0.78383(SAS Output 2). Therefore, it is not useful to add Fees for coin generation as a predictor variable as serious multicollinearity is observed. Exchange Volume (USD) v/s Total count of Transaction, trend is not visible with correlation value of 0.49498(SAS Output 2). Therefore, it is useful to add Generated Coins as a predictor variable. No serious multicollinearity is observed. Generated Coins v/s Fees for coin generation trend is not visible with correlation value of -0.01929(SAS Output 2). Serious multicollinearity is not observed. Generated Coins v/s Total count of Transaction trend is not visible with correlation value of 0.43893(SAS Output 2). Serious multi-collinearity is not observed. Fees for coin generation v/s Total count of Transaction trend is not visible with correlation value of 0.42858(SAS Output 2). Serious multi-collinearity is not observed.

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Variance Inflation
Intercept	Intercept	1	845.77343	1044.51435	0.81	0.4192	12887895622	0
exchangeVolume_USD_	exchangeVolume(USD)	1	6.920089E-7	3.911745E-8	17.69	<.0001	3244591565	2.94334
generatedCoins	generatedCoins	1	0.34255	0.56532	0.61	0.5453	15736565	1.38312
fees	fees	1	-0.33660	0.96800	-0.35	0.7285	288389	2.90493
txCount	txCount	1	0.00884	0.00284	3.11	0.0022	30710661	1.66362

SAS Output 2: Variance Inflation.

Potential Complications:

From the Variance Inflation Table (SAS Output 2), we can see that inflation values for Exchange Volume (USD), Generated Coins, Fees for coin generation and Total count of transaction are less than 5.0 and the average variance inflation is 2.22. So, serious multicollinearity is not evident. In the predictor vs predictor plots we do not see any curvilinear relationships.

II. Preliminary Multiple Linear Regression Analysis.

Preliminary Model:

The preliminary multiple regression model form is given by,

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon_i$

where, Y_i = Price of Bitcoin

X_1 = Exchange Volume

X_2 = Fees

X_3 = Total Count

X_4 = Generated Coins

ε_i = Error

The preliminary model is given by,

Price of the Bitcoin **($\hat{Y}$) = 845.77343 + 0.00000069(Exchange Volume) – 0.33660(Fees) + 0.00884(Total Count) + 0.34255(Generated Coins). (Refer SAS Output 2)**

Model Assumption:

The model assumptions to be verified are

- Multiple Linear Regression model form is reasonable.
- Residuals have constant variance.
- Residuals are normally distributed.
- Residuals are uncorrelated.
- No outliers.
- The predictors are not highly correlated with each other.

1. **To check reasonability of the model:** From the below Residual v/s Predictor variable (Figure 2), we can interpret that no curvature is observed in any of the residual v/s predictor variable plot. Therefore, we can conclude that Multiple Linear Regression model is reasonable.

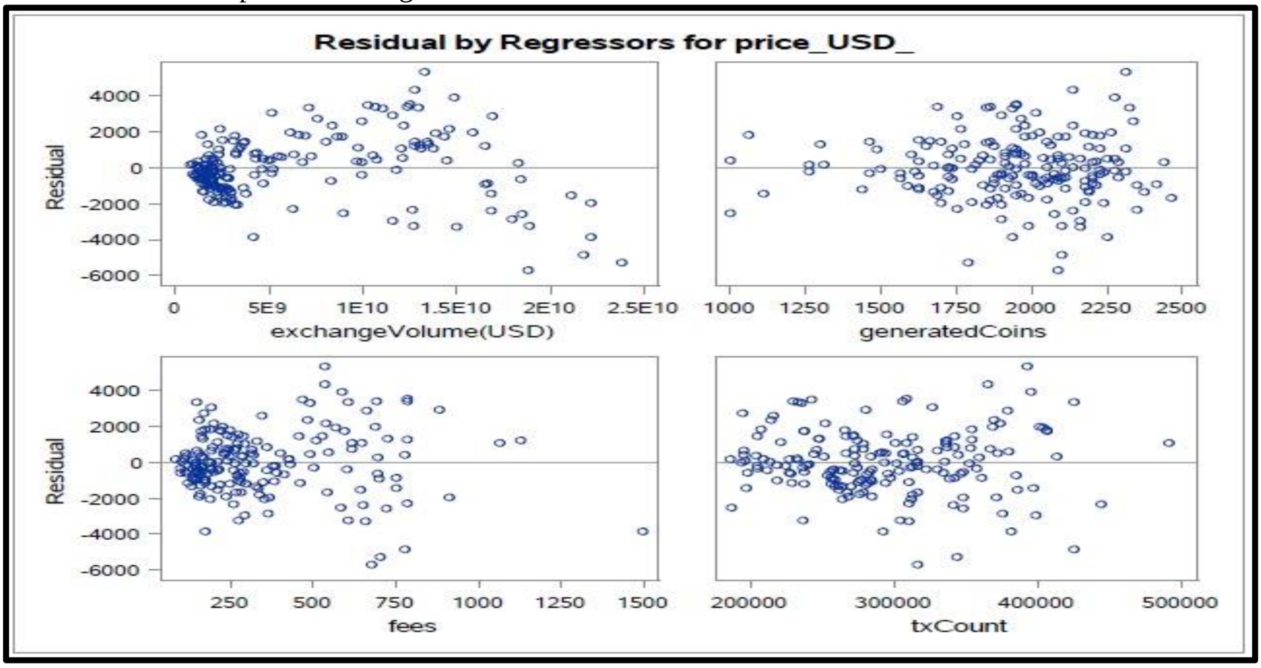

Figure 2: Residual V/s Predictor Variable.

2. **To check that residuals, have constant variance:** From the below, Residual v/s Predicted value of the price of bitcoin (Figure 3), we can observe a funnel shape. Hence it can be interpreted that residual do not have constant variance.

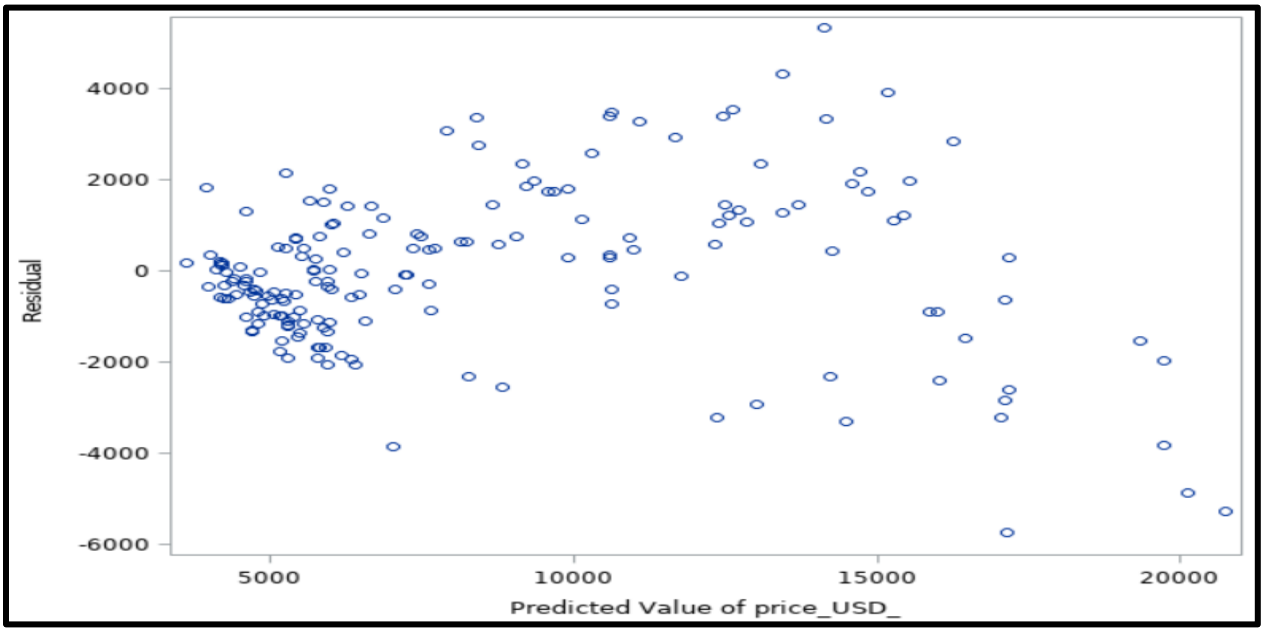

Figure 3: Residual v/s Predicted value of the price of bitcoin.

To check the result of the Residual v/s Predicted value of the price of bitcoin plot (Figure 3), we perform the modified Levene Test. The entire data is divided into two sets below and above average.

Hypothesis:
H_0: $\sigma_{d1} = \sigma_{d2}$.
v/s
H_1: $\sigma_{d1} \neq \sigma_{d2}$.
Decision rule: Reject H_0 if standard deviation of both sets is not equal.
The data is divided into two groups of predicted value of price of bitcoin.
i. Group 1: It has data points whose predicted value of the price of bitcoin is less than 6361.27
ii. Group 2: It has data points whose value for predicted value of the price of bitcoin is greater than 6361.27.
From the below SAS Output 3, it can be inferred that p value (0.0001) is less than 0.05 and hence the reject H_0, indicating the means of both the groups are not same.
From the below SAS Output 3, for T test under unequal variances it can be inferred that p value (0.0001) is less than 0.05 and hence the reject H_0, indicating the variance of both the groups are not constant. The result of the modified Levene test (SAS Output 3) and the Residual v/s Predicted value of the price of bitcoin (Figure 3) aligns indicating that **variance is not constant.**

group	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		90	726.0	591.2	62.3223	0.3044	2615.5
2		90	1711.1	1472.8	155.2	9.5543	6364.2
Diff (1-2)	Pooled		-985.2	1122.2	167.3		
Diff (1-2)	Satterthwaite		-985.2		167.3		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		726.0	602.1	849.8	591.2	515.7	692.9
2		1711.1	1402.6	2019.6	1472.8	1284.6	1726.1
Diff (1-2)	Pooled	-985.2	-1315.3	-655.0	1122.2	1016.8	1252.3
Diff (1-2)	Satterthwaite	-985.2	-1316.5	-653.9			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	178	-5.89	<.0001
Satterthwaite	Unequal	116.96	-5.89	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	89	89	6.21	<.0001

SAS Output 3: Modified Levene Test.

3. **<u>To check that normality of the residuals:</u>** From Normal Probability Plot (Figure 4), we can observe short tail on left-side, but overall normality is satisfied.

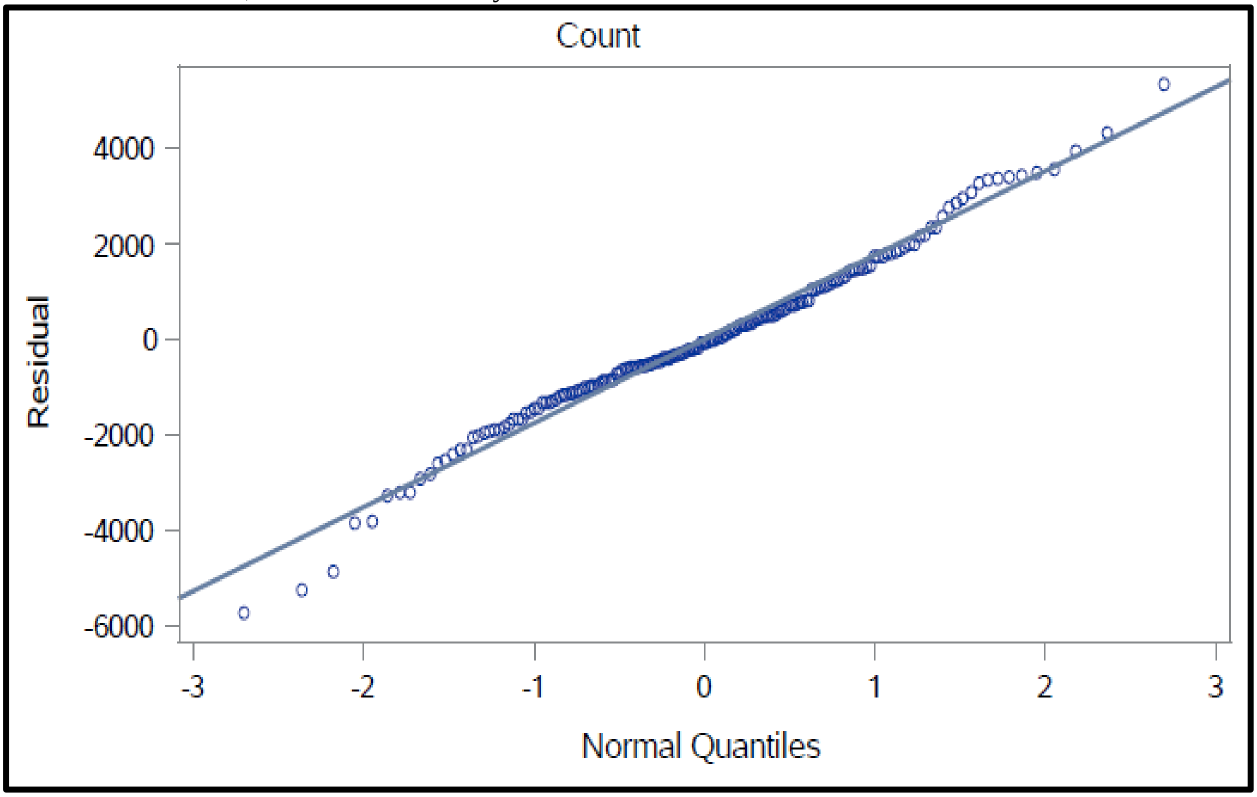

Figure 4: Normal Probability Plot.

To check the result of the Normal Probability Plot (Figure 4), we perform the test of Normality.

Pearson Correlation Coefficients, N = 180		
	e	z
e Residual	1.00000	0.99187
z Rank for Variable e	0.99187	1.00000

SAS Output 4: Pearson Correlation Plot.

H_0: Normality is satisfied.
V/s
H_1: Normality is violated.
Decision Rule: If $\rho < C(\alpha; n)$ then we reject H_0.
Where $\alpha = 0.1$ and $n = 180$.
The cut-off value $C(\alpha, n) = (0.1,180) = 0.99$.
Decision Rule: We reject if $\rho < C(0.1, 180) = 0.99187$ (Refer SAS Output 4) < 0.99. We fail to reject H_0, hence normality is satisfied at a significance level of 0.1. The normality test confirms the outcome of NPP Plot.

4. **To check the correlation between error.**

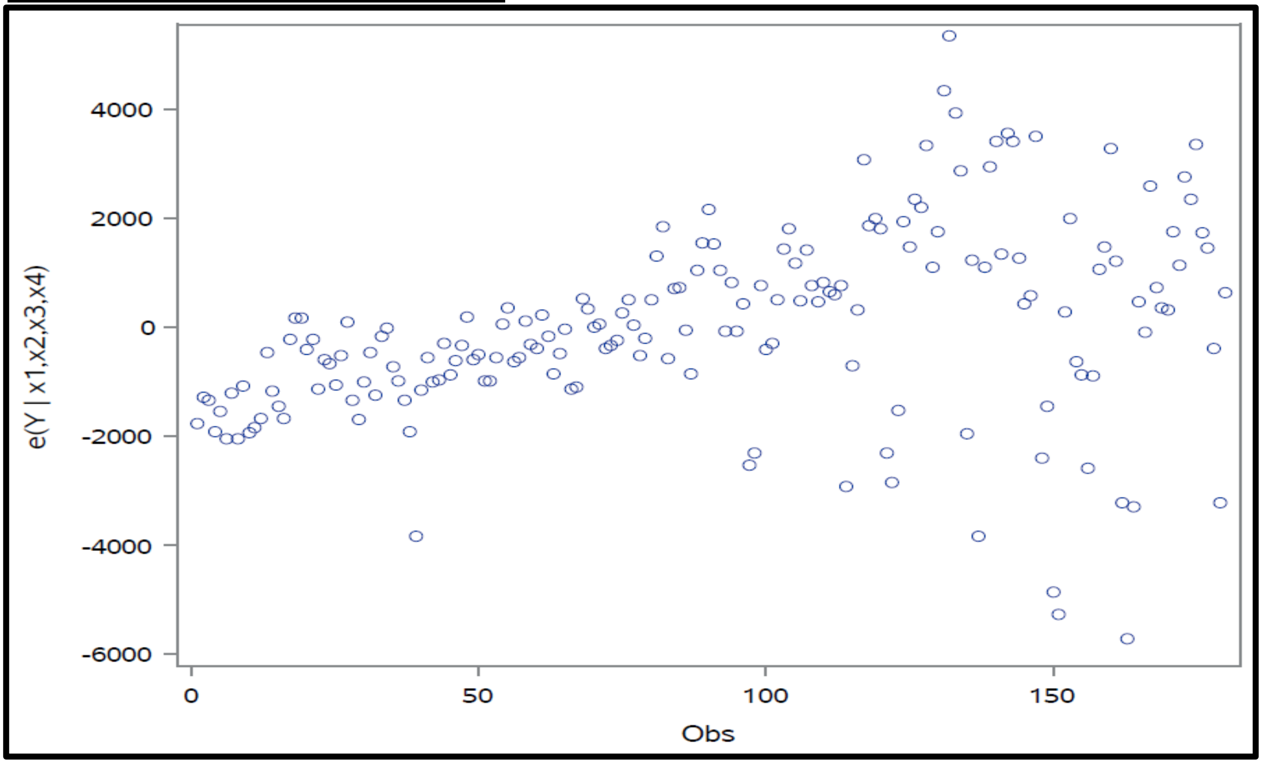

Figure 5: Time Series Plot.

From the above Time Series Plot (Figure 5) we can infer the following: there is random jaggedness upto 110 observations and upward linear trend therefore, some errors are correlated for which remedy is required.

5. **<u>To check the whether any interaction term can be added :</u>** From the below Residual v/s Interaction terms plot (Figure 6), we observe that data points are randomly scattered in all the plot and no trend is visible. Hence, we do not add any interaction term to our model.

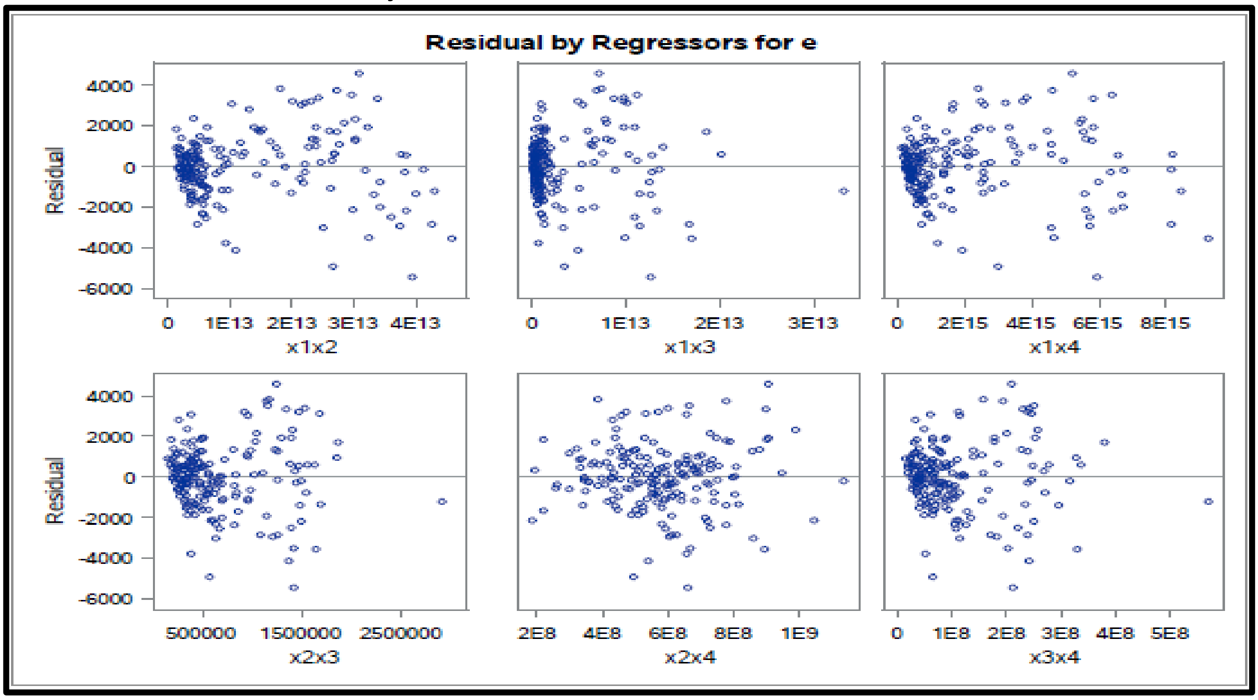

Figure 6: Residual v/s Interaction terms plot.

MODEL DIAGONISTICS

1. **<u>Test for outlier.</u>**

OBS. No	Cookdi	R-Student	H_{ii}	DFFITS	Intercept	Exchange Volume	Coins Generated	Fees	Total Count
Cut-off	0.87	3.29	0.0556	0.3333	DFBETAS = 0.1491				
15	0.00851	-0.8411	**0.0567**	-0.206	**-0.1932**	-0.022	**0.16**	0.015	0.022
16	0.00788	-0.9661	0.0405	-0.198	0.0724	**0.174**	-0.05	**-0.166**	-0.022
29	0.00814	-0.9688	0.0416	-0.202	0.1244	0.085	**-0.180**	-0.081	0.076
39	0.01526	-2.1996	0.0159	-0.279	0.113	0.000	**-0.175**	0.058	0.039
82	0.01659	1.0685	**0.068**	0.288	**0.267**	0.042	**-0.249**	-0.082	0.028
96	0.00097	0.2447	**0.074**	0.069	0.0666	0.017	-0.06	-0.014	-0.004
97	0.03922	-1.493	**0.081**	**-0.444**	**-0.388**	-0.076	**0.301**	-0.040	0.112
98	0.02141	-1.3394	**0.057**	-0.328	0.0376	**0.235**	-0.01	**-0.295**	-0.007
114	0.02282	-1.6905	0.0388	-0.340	0.0561	**-0.164**	0.06	**0.220**	**-0.216**
119	0.00933	1.1369	0.0349	0.216	-0.1001	-0.006	0.01	-0.075	**0.164**
120	0.00837	1.0284	0.0381	0.205	-0.0761	0.016	-0.02	-0.091	**0.160**
121	0.02611	-1.3518	**0.067**	**-0.362**	0.111	**-0.153**	0.03	**0.223**	**-0.240**
122	0.04519	-1.6696	**0.076**	**-0.478**	-0.1265	**-0.397**	**0.209**	**0.366**	**-0.166**
123	0.00913	-0.8805	0.0556	-0.214	-0.0302	**-0.164**	0.07	0.100	-0.057
128	0.02563	1.9177	0.0342	**0.361**	**-0.242**	-0.034	0.08	0.054	**0.204**
129	0.00578	0.635	**0.067**	0.170	-0.0956	-0.021	0.00	0.009	0.134
132	0.04625	3.1125	0.0244	**0.493**	**-0.300**	0.085	**0.155**	-0.018	**0.192**
133	0.02644	2.2609	0.0258	**0.368**	**-0.194**	0.095	0.09	-0.016	0.135
134	0.02157	1.6461	0.0386	0.330	0.0806	**0.158**	**-0.192**	-0.080	**0.156**
135	0.01586	-1.1359	**0.058**	-0.282	-0.0691	-0.134	0.12	0.000	-0.047
136	0.00889	0.712	**0.080**	0.211	0.0029	-0.063	-0.02	**0.160**	-0.010
137	0.21427	-2.3837	**0.162**	**-1.049**	**0.264**	**0.358**	**-0.169**	**-0.879**	0.111

138	0.00737	0.6368	**0.083**	0.192	-0.0145	-0.089	0.01	**0.170**	-0.033
139	0.03357	1.7029	0.0553	**0.412**	-0.0522	**-0.168**	0.12	**0.360**	**-0.163**
140	0.02881	1.9627	0.0366	**0.383**	0.1298	-0.024	0.01	**0.223**	**-0.237**
142	0.02545	2.0457	0.03	**0.360**	-0.0587	-0.098	0.09	**0.279**	-0.098
143	0.02281	1.9615	0.0292	**0.341**	-0.0156	-0.089	0.03	**0.257**	-0.077
147	0.01606	2.0026	0.02	0.286	0.0669	0.098	0.09	0.041	**-0.221**
150	0.08709	-2.8632	0.0524	**-0.673**	0.1036	**-0.355**	0.11	0.130	**-0.257**
151	0.16028	-3.1619	**0.078**	**-0.918**	**-0.319**	**-0.783**	**0.333**	**0.408**	-0.006
160	0.01746	1.879	0.0245	0.298	0.0772	0.113	0.08	0.038	**-0.235**
162	0.02861	-1.8577	0.0403	**-0.381**	-0.0802	**-0.310**	0.01	0.116	0.115
163	0.07946	-2.3699	0.0357	**-0.649**	-0.0076	**-0.431**	-0.11	0.060	**0.218**
167	0.02602	1.4957	0.0553	**0.362**	-0.0473	0.112	**0.253**	0.008	**-0.294**
173	0.02077	1.587	0.0399	0.324	0.1278	**0.197**	0.07	-0.133	**-0.235**
174	0.01405	1.3438	0.0376	0.266	0.1053	**0.189**	0.03	-0.147	**-0.159**
175	0.01977	1.9252	0.0264	0.317	**0.180**	**0.229**	-0.05	**-0.217**	-0.123
177	0.00796	0.8348	0.0539	0.199	**0.170**	0.140	-0.11	-0.127	-0.050
179	0.03449	-1.8679	0.0477	**-0.418**	-0.098	**-0.329**	-0.10	**0.208**	**0.242**

SAS Output 5: Outliers.

Bonferroni Outlier test:

X outlier: Since we have a large data set, we use the cut off value of 2(p/n) where n = 180 and p = 5. Therefore, we get the value of 0.0556. Now, we find the values of Hii (Refer SAS output 5) which are greater than the cut off value of 0.0556. We find that observation numbers 15, 82, 96, 97, 98, 121, 122, 129, 135, 136, 137, 138 and 151 have value greater than 0.0556.

Y outlier: Since we have a large data set, we use the cut off t(1-α/2n; n-p-1) value where α = .10, n = 180, p=5. Therefore, we get the value of t (0.99972; 175) = 3.290. Now, we find that no values of R-student (Refer SAS output 5) are greater than the cut off value of 3.290. Hence, there are no y outliers present.

To check the influence of Outliers on our model.

An observation is said to be an influential observation when its exclusion causes major change in fitted regression function. Following are the three measures of influence:

<u>Influence on Single fitted Value – DFFITIS.</u>

A case is considered influential if the absolute value of DFFITIS exceeds $2\sqrt{\frac{p}{n}}$ for large dataset. $2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{5}{180}}$ = 0.3333. In our case observations 97, 121, 122, 137 and 151. (Refer SAS output 5). These observations have absolute DFFITIS value greater than 0.3333. Therefore, these outliers can be considered as influential outliers.

<u>Influence on ALL fitted Value – Cook's Distance.</u>

A case is considered influential if the Cooke's Distance - Di is greater than F (0.5; p; n-p) = F (0.5; 5; 175) =0.870. In our case, no observations exceed the cutoff cook's distance (Refer SAS output 5). Therefore, there are no outliers on all fitted values.

<u>Influence on Regression Coefficient – DFBETAS</u>.

A case is considered influential if the absolute value of DFBETAS exceeds $2/\sqrt{n}$ for large dataset. $2/\sqrt{n} = 2/\sqrt{180}$ = 0.1491. In our case observation 15, 82, 97, 98, 121, 122, 136, 137, 138 and 151 (Refer SAS output 5). These observations have absolute DFBETAS value greater than 0.1491. Therefore, these outliers can be considered as influential outliers.

From the measures discussed above, we see that observation number 15, 82, 97, 98, 121, 122, 13, 137, 138 and 151 have influence and observations 96, 129 and 1359 are not influential. Remedial measures should be taken for outliers that are influential.

2. **To check (VIF) Variance inflation**.

Parameter Estimates							
Variable	**Label**	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	Intercept	1	845.77343	1044.51435	0.81	0.4192	0
exchangeVolume_USD_	exchangeVolume(USD)	1	6.920089E-7	3.911745E-8	17.69	<.0001	2.94334
generatedCoins	generatedCoins	1	0.34255	0.56532	0.61	0.5453	1.38312
fees	fees	1	-0.33660	0.96800	-0.35	0.7285	2.90493
txCount	txCount	1	0.00884	0.00284	3.11	0.0022	1.66362

SAS Output 6: Variance Inflation Table.

From the table above (SAS Output 6), we see that each individual predictor VIF value is less than 5 and the VIF bar is less than 5. So, we can say that serious multicollinearity is not an issue.

Model Assumptions analysis:
1. The residuals do not have constant variance.
2. The residuals are normally distributed.
3. Some residuals are correlated.
4. Some x outliers are observed which have influence.
To stabilize variance, **Weighted Least Square** Transformation is to be performed.

Weighted Least Square:

Weighted least square (WLS) transformation is used to stabilize the variance when transformation on dependent variable is difficult because normality in residual analysis is satisfied.

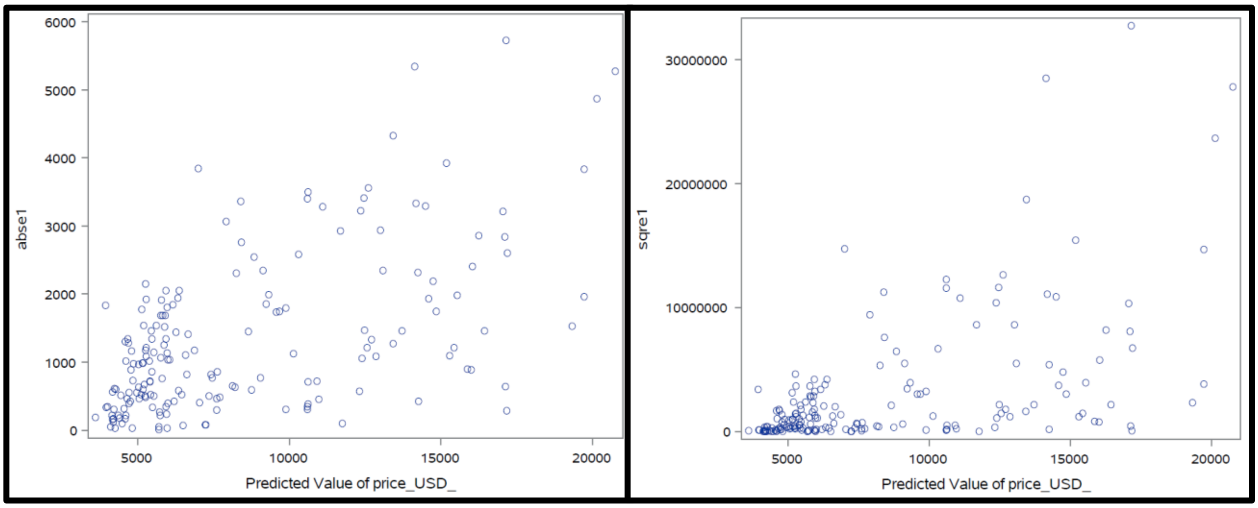

Figure 7: Absolute Residual V/s$\hat{Y}$. & Squared Residual V/s $\hat{Y}$.

From Plots in Figure 7, we see a straight edge and linear trend in the spread for the **Squared Residual V/s $\hat{Y}$.** So, we use the variance function ($\hat{v}$) (squared residuals to estimate unknown error variances) to carry out WLS transformation.

Model Assumption for WLS Transformation model:

We will follow the same model assumption as in above (Refer page 5)

1. **To check reasonability of the model**: From the Residual v/s Predictor variable (Figure 8), we can interpret that curvature is not observed in any of the Residual v/s Predictor variable plot. Therefore, we can conclude that Multiple Linear Regression model is reasonable.

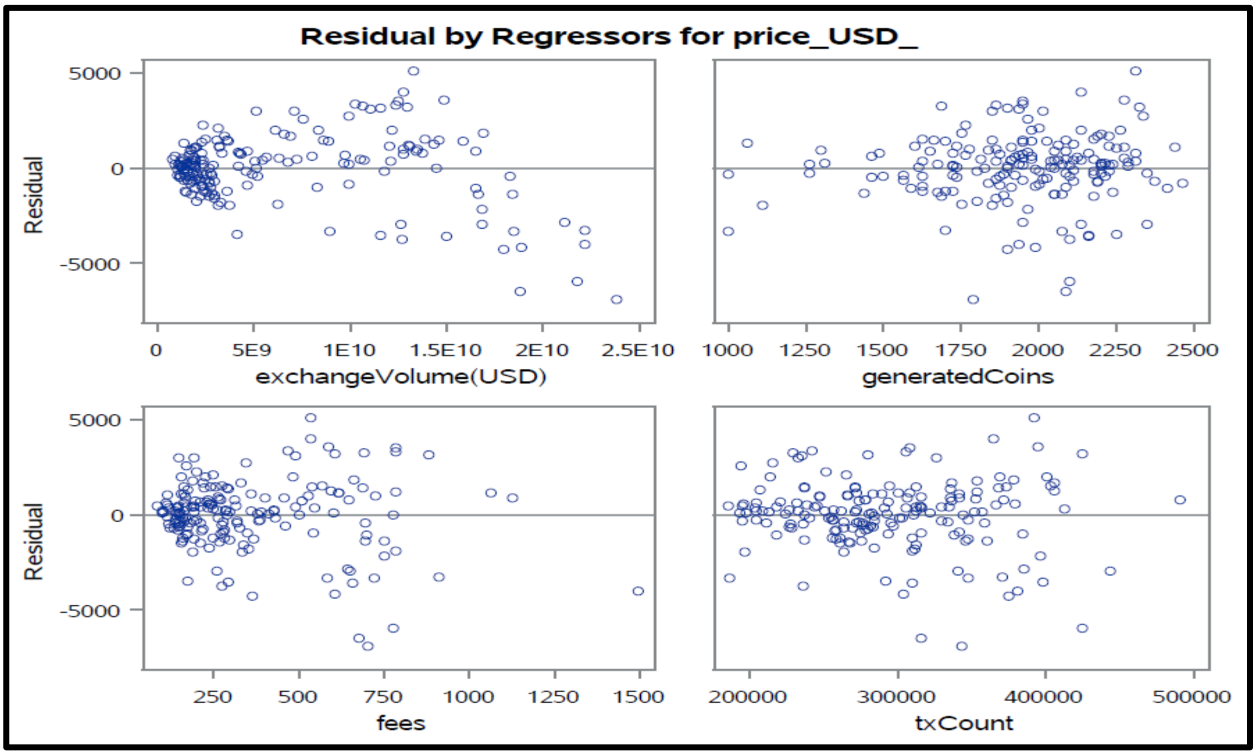

Figure 8: Residual V/s Predictor Variable.

2. **To check constant variance**. From Residual v/s Predicted value of the price of bitcoin (Figure 9), we can observe a funnel shape. Hence it can be interpreted that residual do not have constant variance.

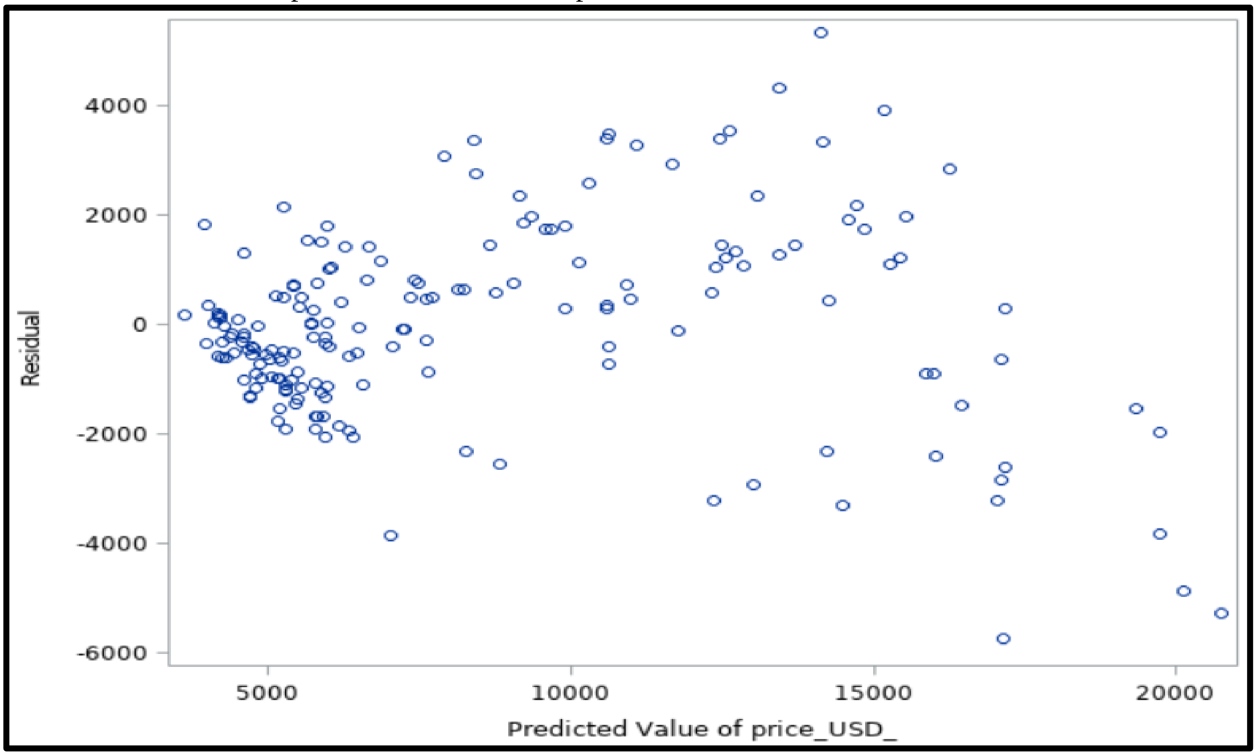

Figure 9: Residual v/s Predicted value of the price of bitcoin.

To check the result of the Residual v/s Predicted value of the price of bitcoin plot (Figure 9), we perform the modified Levene Test. The entire data is divided into two sets below and above average.

Hypothesis:
H_0: $\sigma_{d1} = \sigma_{d2}$.
v/s
H_1: $\sigma_{d1} \neq \sigma_{d2}$.
Decision rule: Reject H_0 if standard deviation of both sets is not equal.
The data is divided into two groups of exchange volume.
i. Group 1: It has data points whose predicted value of the price of bitcoin is less than 6275.82.
ii. Group 2: It has data points whose value for predicted value of the price of bitcoin is greater than 6275.82
From the figure SAS Output 6, it can be inferred that p value (0.0001) is less than 0.05 and hence we reject H_0, indicating the means of both the groups are not same. From the SAS Output 6, for T test under unequal variances it can be inferred that p value (0.0001) is less than 0.05 and hence the reject H_0, indicating the variance of both the groups are not constant. The result of the modified Levene test (SAS Output 6) and the Residual v/s Predicted value of the price of bitcoin (Figure 9) aligns indicating that variance is not constant.

group	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		88	724.3	585.5	62.4113	21.8229	2593.3
2		92	1715.1	1455.1	151.7	17.9496	6336.7
Diff (1-2)	Pooled		-990.8	1118.0	166.7		
Diff (1-2)	Satterthwaite		-990.8		164.0		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		724.3	600.2	848.3	585.5	509.9	687.5
2		1715.1	1413.8	2016.5	1455.1	1271.0	1702.2
Diff (1-2)	Pooled	-990.8	-1319.8	-661.8	1118.0	1013.0	1247.6
Diff (1-2)	Satterthwaite	-990.8	-1315.6	-666.0			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	178	-5.94	<.0001
Satterthwaite	Unequal	120.79	-6.04	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	91	87	6.18	<.0001

SAS Output 6: Modified Levene Test

3. **To check that normality of the residuals**: we can observe short tail on left-side, Normality is violated. To check the result of the Normal Probability Plot (Figure 10), we perform the test of Normality.

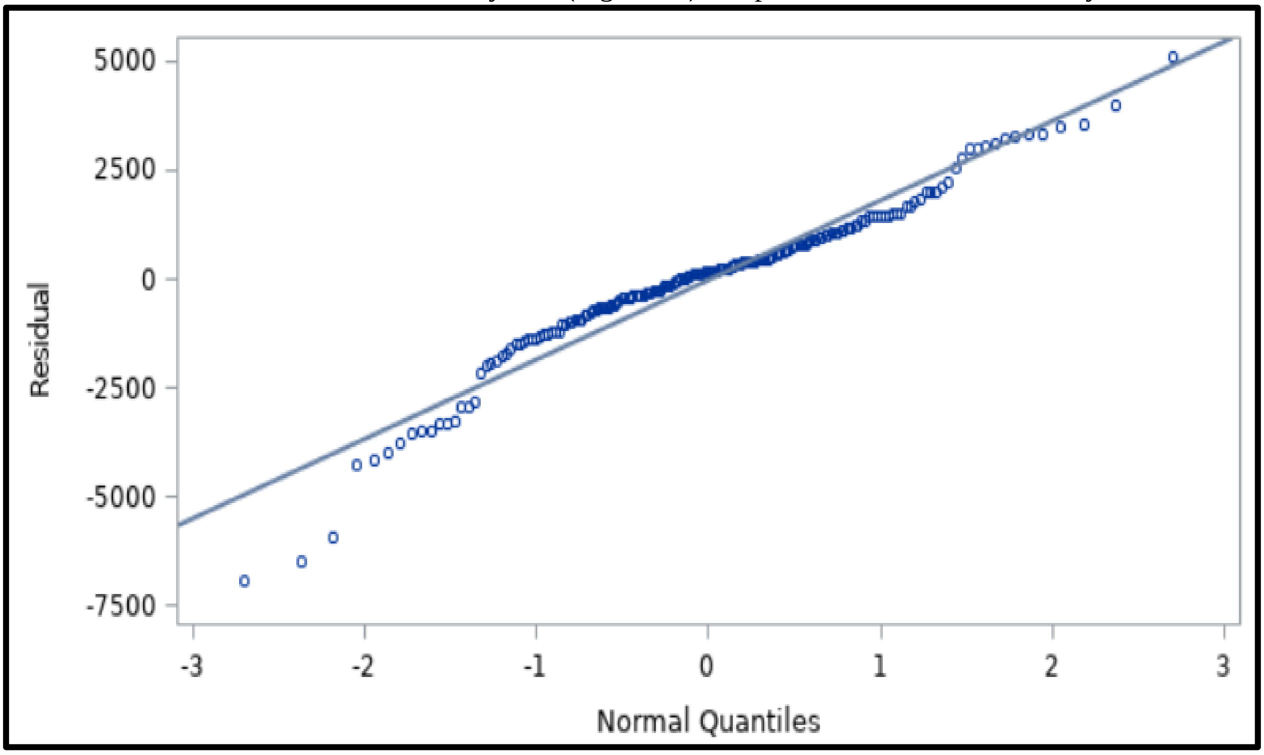

Figure 10: Normal Probability Plot.

Pearson Correlation Coefficients, N = 180		
	e2	enrm2
e2 e(Yw | x1,x2,x3,x4)	1.00000	0.97271
enrm2 Normal Scores	0.97271	1.00000

SAS Output 7: Normal Probability Plot.

H_0: Normality is satisfied.
V/s
H_1: Normality is violated.
Decision Rule: If $\rho < C(\alpha;n)$ then we reject H_0.
Where $\alpha = 0.1$ and $n = 180$.
The cut-off value C $(\alpha, n) = (0.1,180) = 0.99$
Decision Rule: We reject if $\rho <$ C $(0.1, 180) = 0.97271$ (SAS Output 7) < 0.99
We Reject H_0, hence normality is violated at a significance level of 0.1. The normality test confirms the outcome of NPP Plot.

4. Time Series Plot

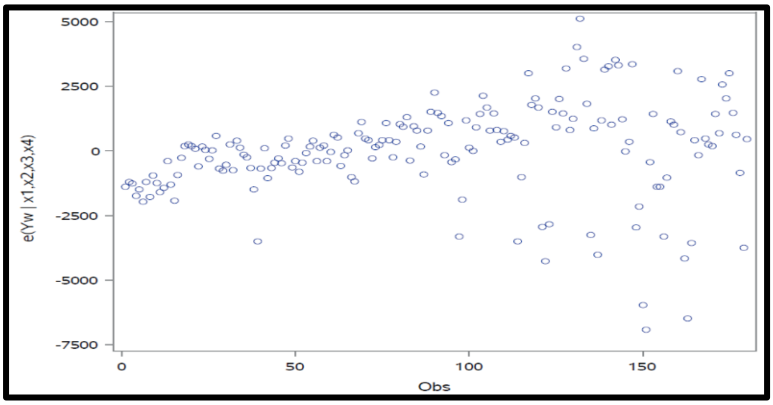

Figure 11: Time Series Plot.

From the Time Series Plot (Figure 11) we can infer the following: there is random jaggedness upto 110 observations and upward linear trend. Therefore, some errors are correlated for which remedy is required.

5. Interaction Terms

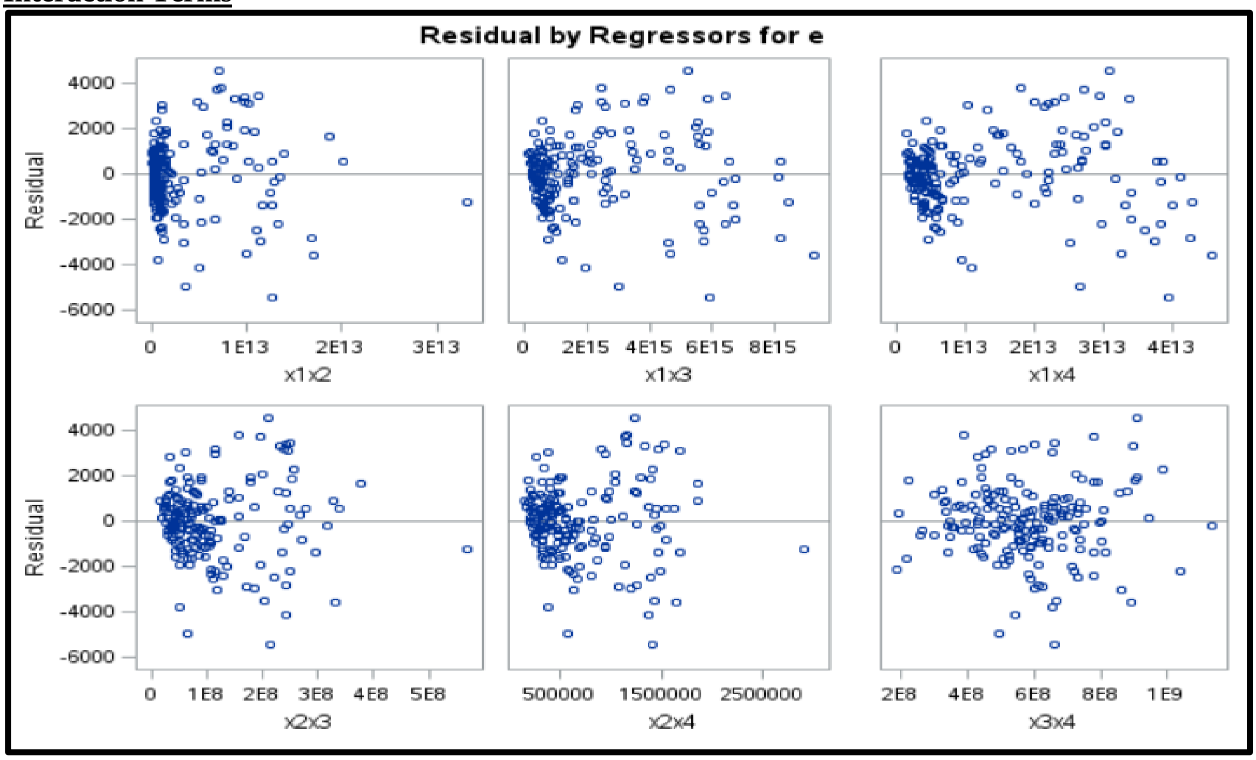

Figure 12: Residual V/s Interaction Terms.

From the Residual v/s Interaction terms plot (Figure 12) we observe that data points are randomly scattered in all the plot and no trend is visible. Hence, we do not add any interaction term to our model.

Model assumptions Analysis
1. The residual does not have constant variance.
2. The residuals are normally distributed.
3. Some residuals are correlated.
4. Some x outliers are observed which have influence.

MODEL DIAGNOSTICS.

1. **Test for Outliers**.

OBS. No	Cookdi	R-Student	H_{ii}	DFFITS	Intercept	Exchange Volume	Coins Generated	Fees	Total Count
Cut-off	0.87	3.29	0.0556	0.3333	DFBETAS = 0.1491				
10	0.0156	-1.335	0.042	-0.280	**0.224**	**0.157**	-0.106	-0.145	-0.1169
15	0.0184	-1.651	0.033	-0.305	**-0.265**	-0.098	**0.182**	0.012	0.0839
16	0.0367	-1.264	**0.103**	**-0.429**	**0.234**	**0.290**	-0.131	**-0.379**	-0.053
17	0.0024	-0.400	**0.069**	-0.109	-0.068	0.017	0.037	-0.047	0.037
18	0.0040	0.359	**0.134**	0.141	0.069	-0.042	-0.040	0.081	-0.0408
19	0.0253	0.667	**0.221**	**0.355**	0.134	-0.143	-0.111	**0.218**	-0.0491
41	0.0001	0.095	**0.060**	0.024	0.003	0.002	0.017	-0.001	-0.019
44	0.0089	-0.824	**0.061**	-0.211	**-0.163**	-0.036	0.138	0.101	-0.0157
47	0.0008	0.255	**0.058**	0.063	0.012	0.003	0.040	-0.005	-0.048
48	0.0068	0.675	**0.070**	0.185	0.100	0.033	0.059	-0.045	-0.1405
59	0.0139	-0.866	**0.085**	-0.263	-0.108	0.007	**0.188**	0.116	-0.1372
65	0.0000	0.040	**0.057**	0.010	0.001	-0.001	-0.005	-0.004	0.0063
69	0.0076	1.077	0.032	0.195	-0.097	-0.031	**0.166**	0.014	-0.0557
73	0.0004	0.178	**0.058**	0.044	-0.029	-0.015	0.005	-0.003	0.031
81	0.0327	1.434	**0.074**	**0.406**	**0.291**	0.087	**-0.345**	**-0.193**	0.1211
82	0.3672	2.856	**0.190**	**1.382**	**1.216**	**0.320**	**-1.113**	**-0.509**	0.0571
84	0.0154	1.392	0.038	0.278	-0.093	-0.121	-0.078	0.028	**0.2064**
89	0.0119	1.718	0.020	0.245	0.034	-0.032	**-0.162**	-0.019	**0.1605**
92	0.0163	1.585	0.032	0.287	**-0.152**	-0.123	-0.037	0.052	**0.2164**
97	0.0173	-1.714	0.029	-0.295	**-0.195**	-0.141	0.111	-0.045	0.1162
99	0.0255	1.274	**0.073**	**0.358**	-0.059	**-0.187**	0.023	**0.326**	-0.0205
114	0.0095	-1.509	0.021	-0.219	0.027	**-0.158**	0.056	0.120	-0.1016
122	0.0133	-1.463	0.030	-0.259	-0.046	**-0.237**	0.085	0.148	-0.0499
132	0.0131	2.058	0.016	0.258	-0.086	**0.158**	0.015	-0.021	0.06
137	0.0176	-1.306	0.049	-0.297	0.070	-0.050	-0.041	**-0.191**	0.0359
139	0.0110	1.391	0.028	0.236	-0.035	0.023	0.070	**0.168**	-0.0857
150	0.0196	-1.885	0.027	-0.315	0.026	**-0.237**	0.049	0.043	-0.0533
151	0.0287	-2.077	0.033	**-0.383**	-0.080	**-0.346**	0.077	0.109	0.0227
162	0.0093	-1.382	0.024	-0.217	-0.038	**-0.192**	0.004	0.044	0.0493
163	0.0216	-2.173	0.023	-0.332	-0.026	**-0.272**	-0.020	0.028	0.0779
173	0.0076	1.282	0.023	0.196	0.090	**0.152**	0.046	-0.079	-0.1311
175	0.0085	1.578	0.017	0.208	0.106	**0.184**	-0.007	-0.131	-0.0807
179	0.0103	-1.472	0.023	-0.227	-0.071	**-0.206**	-0.035	0.095	0.1063

SAS Output 8: Outliers

Bonferroni Outlier test:
X outlier: Since we have a large data set, we use the cut off value of 2(p/n) where n = 180 and p = 5. Therefore, we get the value of 0.0556. Now, we find the values of Hii (Refer SAS output 8) which are greater than the cut off value of 0.0556. We find that observation numbers 16, 17, 18, 19, 41, 44, 47, 48, 59, 65, 73, 81, 82 and 99 have value greater than 0.0556.

Y outlier: Since we have a large data set, we use the cut off t(1-α/2n; n-p-1) value where α = .10, n = 180, p=5. Therefore, we get the value of t (0.99972; 175) = 3.290. Now, we find that no values of R-student (Refer SAS output 8) are greater than the cut off value of 3.290. Hence, there are no y outliers present.

To check the influence of Outliers on our model.
An observation is said to be an influential observation when its exclusion causes major change in fitted regression function. Following are the three measures of influence:

Influence on Single fitted Value – DFFITIS

A case is considered influential if the absolute value of DFFITIS exceeds $2\sqrt{\frac{p}{n}}$ for large dataset. $2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{5}{180}}$ = 0.3333. In our case observations 16, 19, 81, 82 and 99 (Refer SAS output 8). These observations have absolute DFFITIS value greater than 0.3333. Therefore, these outliers can be considered as influential outliers.

Influence on ALL fitted Value – Cook's Distance.
A case is considered influential if the Cooke's Distance - D_i is greater than F (0.5; p; n-p) = F (0.5; 5; 175) =0.870. In our case, no observations exceed the cutoff cook' distance (Refer SAS output 8). Therefore, there are no outliers on all fitted values.

Influence on Regression Coefficient – DFBETAS.
A case is considered influential if the absolute value of DFBETAS exceeds $2/\sqrt{n}$ for large dataset.
$2/\sqrt{n} = 2/\sqrt{180}$ = 0.1491. In our case observations 16, 19, 44, 59, 81, 82 and 99 (Refer SAS output 8). These observations have absolute DFBETAS value greater than 0.1491. Therefore, these outliers can be considered as influential outliers.
From the measures discussed above, we see that observation number 16, 19, 44, 59, 81, 82 and 99 have influence and observation 17, 18, 41, 47, 48, 65 and 73 is not influential. Remedial measures should be taken for outliers that are influential

2. **To check (VIF)Variance inflation.**
From the table (SAS output 9) , we see that each individual predictor VIF value is less than 5 and the VIF bar is less than 5. So, we can say that serious multicollinearity is not an issue.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	2299.68164	583.56221	3.94	0.0001	0
exchangeVolume_USD_	exchangeVolume(USD)	1	8.013763E-7	3.723111E-8	21.52	<.0001	1.89634
generatedCoins	generatedCoins	1	-0.76246	0.31552	-2.42	0.0167	1.54628
fees	fees	1	-1.80800	0.74223	-2.44	0.0159	1.87212
txCount	txCount	1	0.01054	0.00214	4.93	<.0001	1.63525

SAS Output 9: Variance Inflation Table.

Complications to our model:

To make variance constant, we perform the log transformation on the dependent variable of our transformed model. On performing log transformation, we get satisfactory model assumptions but there is a curvilinear spread of points on the residuals vs exchange volume(X_1). To remedy this curvilinear spread, we add a quadratic term by squaring the exchange volume predictor variable (X_1^2). On addition of the quadratic term to our model, we observe that model assumptions are satisfied with serious multicollinearity issues and a high average VIF of 13. Thus, we use the weighted least square transformed model with no serious multicollinearity issues for further analysis.

Preliminary model:
The preliminary multiple regression model form is given by,
$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon_i$.
where, Y_i = Price of Bitcoin.
X_1 = Exchange Volume.
X_2 = Fees.
X_3 = Total Count.
X_4 = Generated Coins.
εi = Error. Independent ~ Normal $(0, \sigma_i^2)$; squared residuals are used to estimate σ_i^2
The preliminary model is given by,
Change
Price of the Bitcoin **($\hat{Y}$) = 845.77343 + 0.00000069(Exchange Volume) – 0.33660(Fees) + 0.00884(Total Count) + 0.34255(Generated Coins).**

Explained Variability: Variability in our model can be explained by how well the data fits the model. R^2 = 83.73% and R_a^2 = 83.36 %.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	882.24614	220.56153	225.13	<.0001
Error	175	171.44521	0.97969		
Corrected Total	179	1053.69134			

Root MSE	0.98979	**R-Square**	0.8373
Dependent Mean	5498.02637	**Adj R-Sq**	0.8336
Coeff Var	0.01800		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	2299.68164	583.56221	3.94	0.0001	0
exchangeVolume_USD_	exchangeVolume(USD)	1	8.013763E-7	3.723111E-8	21.52	<.0001	1.89634
fees	fees	1	-1.80800	0.74223	-2.44	0.0159	1.87212
txCount	txCount	1	0.01054	0.00214	4.93	<.0001	1.63525
generatedCoins	generatedCoins	1	-0.76246	0.31552	-2.42	0.0167	1.54628

SAS Output 10: ANOVA Table for Weighted Least Square Model.

<u>F- Test on significance of Regression</u>:
H_0: β_1, β_2, β_3, β_4 = 0
V/s
H_1: β_1, β_2, β_3, β_4 (at least one β_k) $\neq 0$
Decision Rule: Reject H_0 if F* > F (1-α; p-1, n-p)
(Refer SAS Output 10) F* = MSR/MSE = 220.56153/0.97969 = 225.13
At a significance level of α=0.1, we have F (0.9,4,175) = 1.96858
Where α = 0.1 and n = 180.
Since the value of F* = 225.13> 1.96858, We Reject H_0.
Hence, we can conclude that the regression is significant.

To check the significance of individual predictors: (Refer SAS Output 10)
P-value of x1 (exchange volume) = 0.0001
P-value of x2 (Fees) = 0.0159
P-value of x3 (Tx Count) = 0.0001
P-value of x4 (Generated Coins) =0.0167

Since the p-values of all predictors are less than α=0.1, All predictor variables are significant in our model.

I. Exploration of Interaction Terms.

Partial Regression : Partial regression plots are plotted as an alternative to regular residual plots when the predictor variables are correlated. From the below (Fig 13) we see that there are serious multicollinearity issues for interactions terms with predictor variables. Hence, we use partial Regression plots to find if any interaction terms can be added to our model.
From the (Fig 13) the graph for $e(Y_w)$ Vs e(x1x2) we observe a downward trend. Hence, we add interaction term x1x2 to preliminary our model. All other partial regression plots do not show any trend, so we do not add other interactions terms to our model.

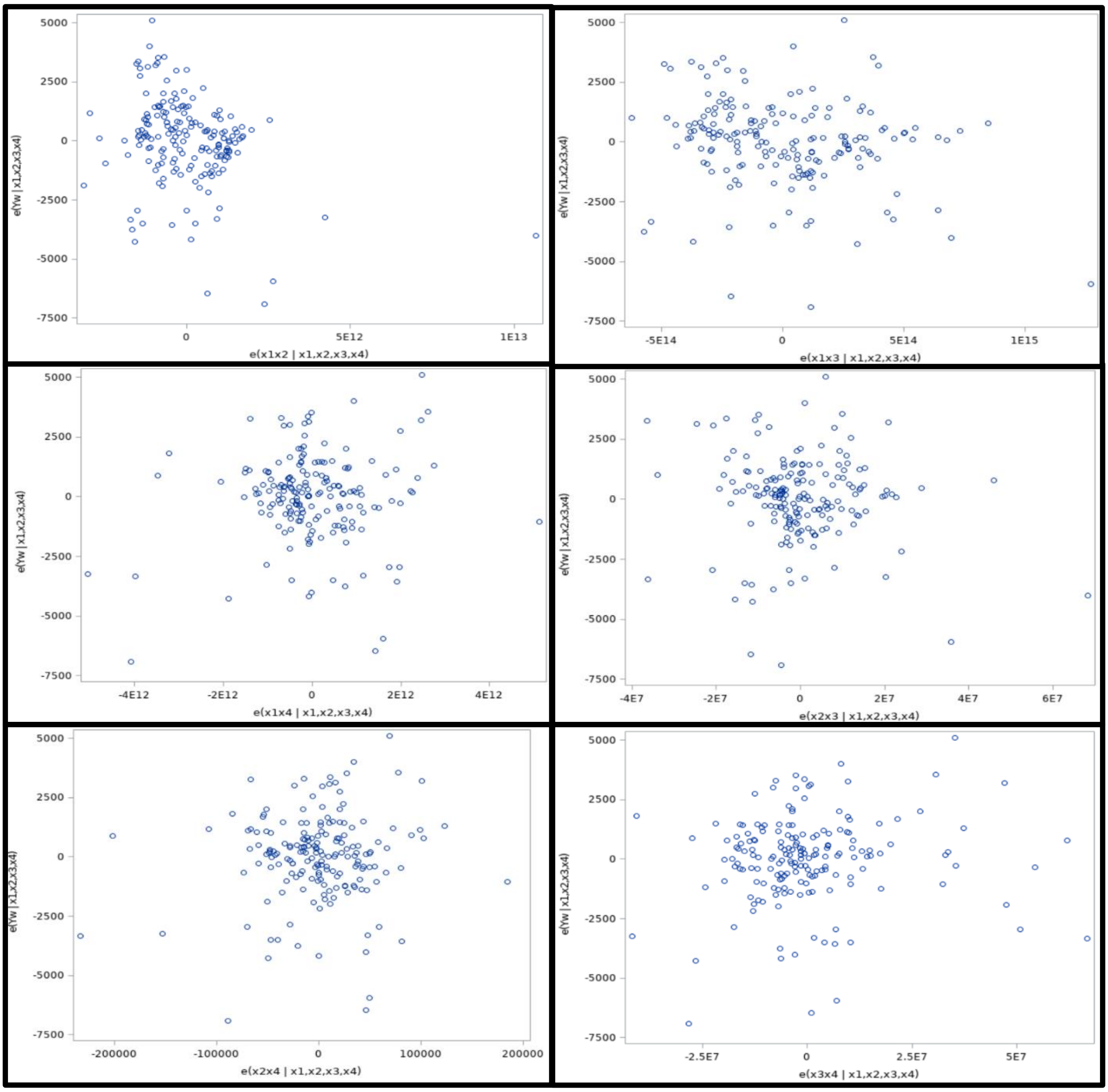

Figure 13: Residual V/s Interaction Model**.**

Pearson Correlation Coefficients, N = 180											
	price_USD_	exchangeVolume_USD_	fees	txCount	generatedCoins	x1x2	x1x3	x1x4	x2x3	x2x4	x3x4
price_USD_ price(USD)	1.00000	0.91851	0.71483	0.54815	0.22639	0.79240	0.90018	0.91979	0.75751	0.76270	0.49947
exchangeVolume_USD_ exchangeVolume(USD)	0.91851	1.00000	0.78383	0.49498	0.17794	0.90205	0.97188	0.98828	0.81175	0.81690	0.43581
fees fees	0.71483	0.78383	1.00000	0.42858	-0.01929	0.91057	0.77344	0.74810	0.96438	0.97396	0.29654
txCount txCount	0.54815	0.49498	0.42858	1.00000	0.43893	0.46152	0.64244	0.53879	0.61133	0.52206	0.92188
generatedCoins generatedCoins	0.22639	0.17794	-0.01929	0.43893	1.00000	0.08025	0.22393	0.28609	0.09970	0.17824	0.74117
x1x2	0.79240	0.90205	0.91057	0.46152	0.08025	1.00000	0.90099	0.87228	0.93056	0.91307	0.36104
x1x3	0.90018	0.97188	0.77344	0.64244	0.22393	0.90099	1.00000	0.97180	0.85163	0.82083	0.56721
x1x4	0.91979	0.98828	0.74810	0.53879	0.28609	0.87228	0.97180	1.00000	0.79428	0.81175	0.51781
x2x3	0.75751	0.81175	0.96438	0.61133	0.09970	0.93056	0.85163	0.79428	1.00000	0.97207	0.48282
x2x4	0.76270	0.81690	0.97396	0.52206	0.17824	0.91307	0.82083	0.81175	0.97207	1.00000	0.45048
x3x4	0.49947	0.43581	0.29654	0.92188	0.74117	0.36104	0.56721	0.51781	0.48282	0.45048	1.00000

SAS Output 11: Pearson Correlation for Interaction Terms.

From table shown above (SAS Output 11): We see that there is high correlation between interaction terms with correlation value > 0.7 for majority of interaction terms with predictor variables, which shows there is serious multicollinearity issues. To decrease the correlation, we standardize our interaction terms. The table for standardized interaction terms is shown below (SAS Output 12).

Pearson Correlation Coefficients, N = 180											
	price_USD_	exchangeVolume_USD_	fees	txCount	generatedCoins	stdx1x2	stdx1x3	stdx1x4	stdx2x3	stdx2x4	stdx3x4
price_USD_ price(USD)	1.00000	0.91851	0.71483	0.54815	0.22639	0.47820	0.41261	0.11414	0.36256	0.08638	0.03756
exchangeVolume_USD_ exchangeVolume(USD)	0.91851	1.00000	0.78383	0.49498	0.17794	0.61562	0.50972	0.09897	0.44035	0.05413	0.01386
fees fees	0.71483	0.78383	1.00000	0.42858	-0.01929	0.69600	0.42441	0.05357	0.42946	-0.07479	0.04079
txCount txCount	0.54815	0.49498	0.42858	1.00000	0.43893	0.31103	0.39202	0.01992	0.17629	0.05923	0.05235
generatedCoins generatedCoins	0.22639	0.17794	-0.01929	0.43893	1.00000	0.03057	0.01551	-0.30455	0.04735	-0.12014	-0.45083
stdx1x2	0.47820	0.61562	0.69600	0.31103	0.03057	1.00000	0.61238	-0.08334	0.72900	-0.04613	-0.09373
stdx1x3	0.41261	0.50972	0.42441	0.39202	0.01551	0.61238	1.00000	0.27355	0.77093	0.23267	0.23341
stdx1x4	0.11414	0.09897	0.05357	0.01992	-0.30455	-0.08334	0.27355	1.00000	0.13464	0.77170	0.52120
stdx2x3	0.36256	0.44035	0.42946	0.17629	0.04735	0.72900	0.77093	0.13464	1.00000	0.25478	-0.02006
stdx2x4	0.08638	0.05413	-0.07479	0.05923	-0.12014	-0.04613	0.23267	0.77170	0.25478	1.00000	0.26749
stdx3x4	0.03756	0.01386	0.04079	0.05235	-0.45083	-0.09373	0.23341	0.52120	-0.02006	0.26749	1.00000

SAS Output 12: Pearson Correlation for Standardized Interaction Terms

From the table (SAS Output 12) we see that correlation value of majority standardised interaction terms is below 0.7. Thus, serious multicollinearity is not a problem.

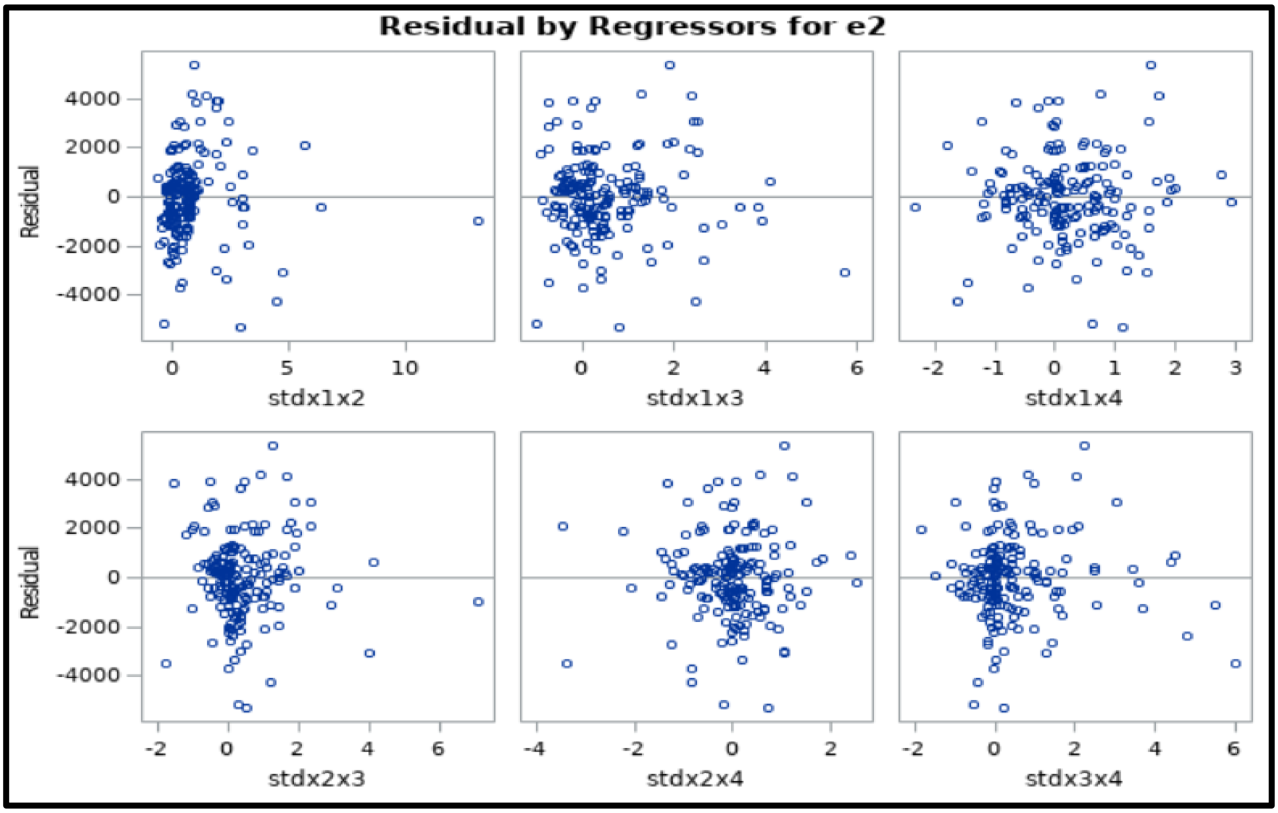

Figure 14: Residual V/s Standardized Interaction Terms.

Now, we plot the residual vs standardised interaction terms (Figure 14). The graph for Residual vs stdx1x2, Residual vs stdx1x3, Residual vs stdx2x3 we observe a downward trend. Hence, we add stdx1x2 stdx1x3 and stdx2x3 interaction terms to our model. All other standardised interaction plots do not show any trend, so we do not add other standard interactions terms to our model.

II. Model Search

For selecting the best models, model search is now performed by using backward elimination, stepwise regression and best subsets methods.

1. **Backward Elimination Method**:

The backward elimination search procedure begins with the model containing all potential predictor variables and identifies the one with the largest P-value. If the maximum P-value is greater than predetermined limit $\alpha = 0.10$, that predictor variable is dropped. The model with the remaining variable is then fitted, and the next candidate for drop. The process continues until no further predictor variables can be dropped. We remove terms one by one and not at once due to multicollinearity issue.

Backward Elimination: Step 5

Variable stdx3x4 Removed: R-Square = 0.8773 and C(p) = 4.6416

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	3373898830	674779766	248.79	<.0001
Error	174	471923225	2712202		
Corrected Total	179	3845822055			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	345.05564	674.66194	709460	0.26	0.6097
exchangeVolume_USD_	7.62875E-7	2.957888E-8	1804119489	665.19	<.0001
txCount	0.01253	0.00248	69125464	25.49	<.0001
stdx1x2	-632.40263	144.22344	52148012	19.23	<.0001
stdx1x3	-683.17616	204.76852	30189860	11.13	0.0010
stdx2x3	788.66182	235.46249	30427080	11.22	0.0010

Bounds on condition number: 3.7075, 64.83

All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination								
Step	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	stdx1x4		9	0.0000	0.8799	9.0050	0.00	0.9438
2	stdx2x4		8	0.0000	0.8799	7.0369	0.03	0.8579
3	generatedCoins	generatedCoins	7	0.0001	0.8797	5.2080	0.17	0.6780
4	fees	fees	6	0.0008	0.8790	4.3005	1.11	0.2934
5	stdx3x4		5	0.0017	0.8773	4.6416	2.38	0.1249

SAS Output 13: Backward Deletion Step 5 and Summary

2. **Best Subset**:

The REG Procedure
Model: MODEL1
Dependent Variable: price_USD_

Adjusted R-Square Selection Method

Number of Observations Read	180
Number of Observations Used	180

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
1	0.8428	0.8437	45.4570	2707.8758	2714.26169	exchangeVolume_USD_
1	0.6258	0.6279	351.1001	2863.9649	2870.35083	x1x2

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
2	0.8543	0.8559	30.0906	2695.1761	2704.75502	exchangeVolume_USD_ stdx1x2
2	0.8536	0.8552	31.0520	2696.0221	2705.60097	exchangeVolume_USD_ txCount

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
3	0.8655	0.8677	15.3715	2681.7914	2694.56323	exchangeVolume_USD_ txCount stdx1x2
3	0.8607	0.8630	22.0520	2688.0973	2700.86913	exchangeVolume_USD_ txCount x1x2

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
4	0.8665	0.8694	14.9482	2681.4482	2697.41302	exchangeVolume_USD_ txCount stdx1x2 stdx2x3
4	0.8664	0.8694	15.0355	2681.5333	2697.49804	exchangeVolume_USD_ txCount stdx1x2 stdx1x3

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
5	0.8738	0.8773	5.8280	2672.2866	2691.44434	exchangeVolume_USD_ txCount stdx1x2 stdx1x3 stdx2x3
5	0.8677	0.8714	14.1450	2680.6992	2699.85691	fees txCount stdx1x2 stdx2x3 x1x2

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
6	0.8743	0.8785	6.0764	2672.4636	2694.81429	fees txCount stdx1x2 stdx1x3 stdx2x3 x1x2
6	0.8743	0.8785	6.0764	2672.4636	2694.81429	exchangeVolume_USD_ txCount stdx1x2 stdx1x3 stdx2x3 x1x2

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
7	0.8736	0.8786	8.0000	2674.3836	2699.92730	generatedCoins fees txCount stdx1x2 stdx1x3 stdx2x3 x1x2
7	0.8736	0.8786	8.0000	2674.3836	2699.92730	exchangeVolume_USD_ generatedCoins txCount stdx1x2 stdx1x3 stdx2x3 x1x2

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
6	0.8743	0.8785	6.0764	2672.4636	2694.81429	fees txCount stdx1x2 stdx1x3 stdx2x3 x1x2
6	0.8743	0.8785	6.0764	2672.4636	2694.81429	exchangeVolume_USD_ txCount stdx1x2 stdx1x3 stdx2x3 x1x2
6	0.8743	0.8785	6.0764	2672.4636	2694.81429	exchangeVolume_USD_ fees txCount stdx1x2 stdx1x3 stdx2x3
6	0.8743	0.8785	6.0764	2672.4636	2694.81429	exchangeVolume_USD_ fees txCount stdx1x3 stdx2x3 x1x2
5	0.8738	0.8773	5.8280	2672.2866	2691.44434	exchangeVolume_USD_ txCount stdx1x2 stdx1x3 stdx2x3
7	0.8736	0.8786	8.0000	2674.3836	2699.92730	generatedCoins fees txCount stdx1x2 stdx1x3 stdx2x3 x1x2
7	0.8736	0.8786	8.0000	2674.3836	2699.92730	exchangeVolume_USD_ generatedCoins txCount stdx1x2 stdx1x3 stdx2x3 x1x2
7	0.8736	0.8786	8.0000	2674.3836	2699.92730	exchangeVolume_USD_ generatedCoins fees txCount stdx1x2 stdx1x3 stdx2x3
7	0.8736	0.8786	8.0000	2674.3836	2699.92730	exchangeVolume_USD_ generatedCoins fees txCount stdx1x3 stdx2x3 x1x2
6	0.8734	0.8776	7.3331	2673.7734	2696.12407	exchangeVolume_USD_ generatedCoins txCount stdx1x2 stdx1x3 stdx2x3

SAS Output 14: Best Subsets.

The above SAS output 14, shows two best models with increment in number of parameter and best ten models with combinations of all predictor terms. To select best subset model, main priority is given to model with Mallows C_p close to number of parameters p, minimum Akaike's Information Criterion(AIC), Schwarz' Bayesian Criterion(SBC), and equalization of R-square and Adjusted R-Square. We select models 3,5,6 and 7.

Potential best models we selected are: Model with 3 predictors contains Exchange Volume, Total Count and stdx1x2, Model with 5 predictors contains Exchange Volume, Total Count, stdx1x2, stdx1x3 and stdx2x3, Model with 6 predictors contains Fees, Total Count, stdx1x2, stdx1x3, stdx2x3 and x1x2, Model with 7 predictors contains Generated Coins, Fees, Total Count, stdx1x2, stdx1x3, stdx2x3 and x1x2.

3. **Stepwise Regression**:

   In Stepwise Regression method, each predictor or interaction term is added or removed based on its least p-value compared to alpha. In each step, as shown in above figures the R-square value increases as a parameter is added and the mallows C_p decreases. This process takes place till the R-square value is highest with the number of parameters added. In this case, in step 3 Total Count is added that makes total of 4 predictors having least P-values. This last step (step3) (SAS Output 15) concludes the final model which will be the model to be considered.

Stepwise Selection: Step 3

Variable txCount Entered: R-Square = 0.8677 and C(p) = 14.0926

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	3337129915	1112376638	384.87	<.0001
Error	176	508692140	2890296		
Corrected Total	179	3845822055			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	1157.83333	656.58669	8987736	3.11	0.0796
exchangeVolume_USD_	7.497946E-7	3.022691E-8	1778441522	615.31	<.0001
txCount	0.00959	0.00242	45390461	15.70	0.0001
stdx1x2	-461.11703	113.15111	48000645	16.61	<.0001

Bounds on condition number: 1.9267, 14.585

All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Selection									
Step	Variable Entered	Variable Removed	Label	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	exchangeVolume_USD_		exchangeVolume(USD)	1	0.8437	0.8437	43.9456	960.59	<.0001
2	stdx1x2			2	0.0123	0.8559	28.6976	15.06	0.0001
3	txCount		txCount	3	0.0118	0.8677	14.0926	15.70	0.0001

SAS Output 15: Stepwise Regression.

4. **Selection of Potential Best Models**: We have selected total 4 best models by doing various procedures, out of which we selected 2 potential best models - Model with 5 predictors: Exchange Volume, Total Count, stdx1x2, stdx1x3 and stdx2x3. We get this model as best model from Best Subset and Backward Deletion and Model with 3 predictors: Exchange Volume, Total Count and stdx1x2. We get this model as best from Stepwise Deletion and Best Subset.

   We rejected model with 6 predictor variables and model with 7 predictor variables because of serious multicollinearity issue.

III. Model Selection.

From the model search phase, two potential best models are:

First Potential Best Model:
$Y_i = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \varepsilon_i$
Where, Y_i = Price of Bitcoin
X_1 = Exchange Volume
X_2= Total Count
X_3= stdx1x2
X_4 = stdx1x3
X_5 = stdx2x3
ε_i = Error

Second Potential Best Model:
$Y_i = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \varepsilon_i$
Where, Y_i = Price of Bitcoin
X_1 = Exchange Volume
X_2= Total Count
X_3= stdx1x2
ε_i = Error

1. **Matrix Plot.**

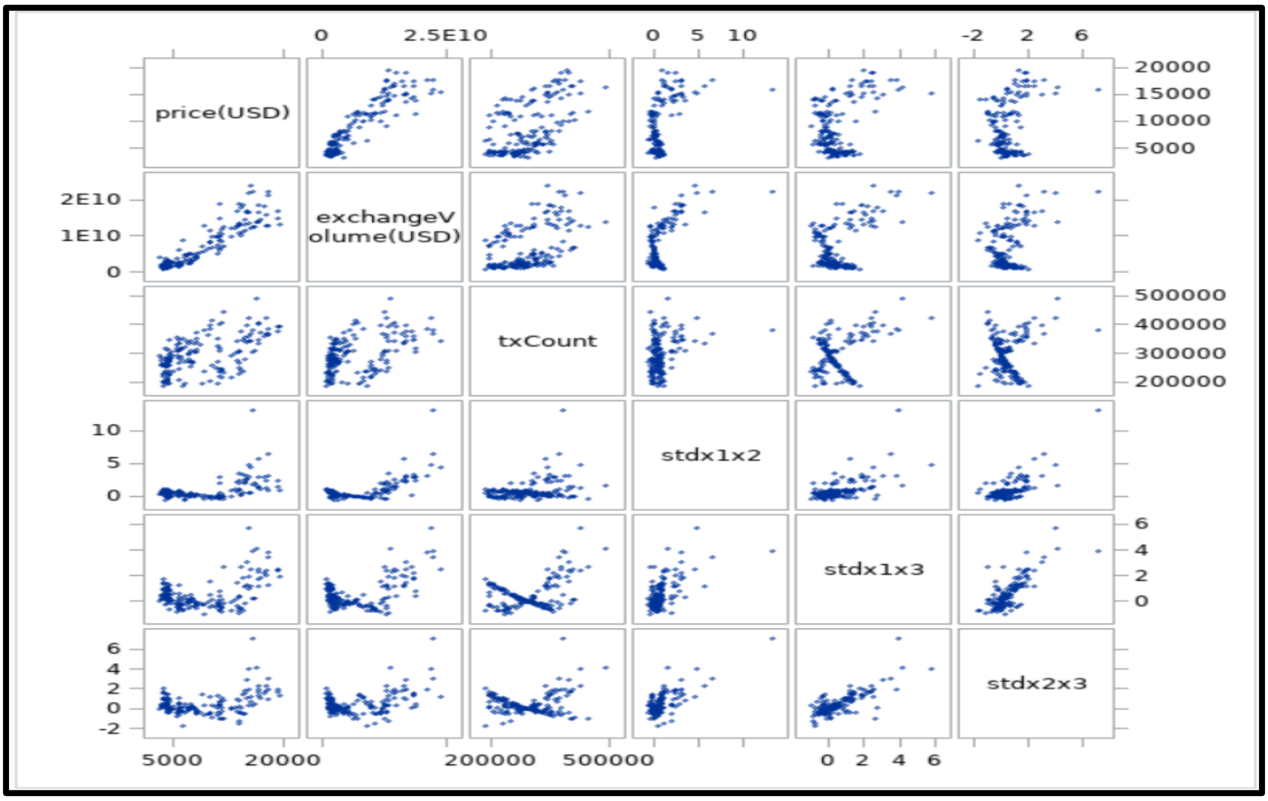

Figure 15: Matrix Scatter Plot for Best Model.

In first potential best solution (Refer Fig 15), We observe that there are no serious multicollinearity issues except for stdx1x3 vs stdx2x3 which shows an upward trend. In second potential best solution (Refer Fig 16) we observe that there are no serious multicollinearity issues. All predictor vs predictor plots do not depict significant trend.

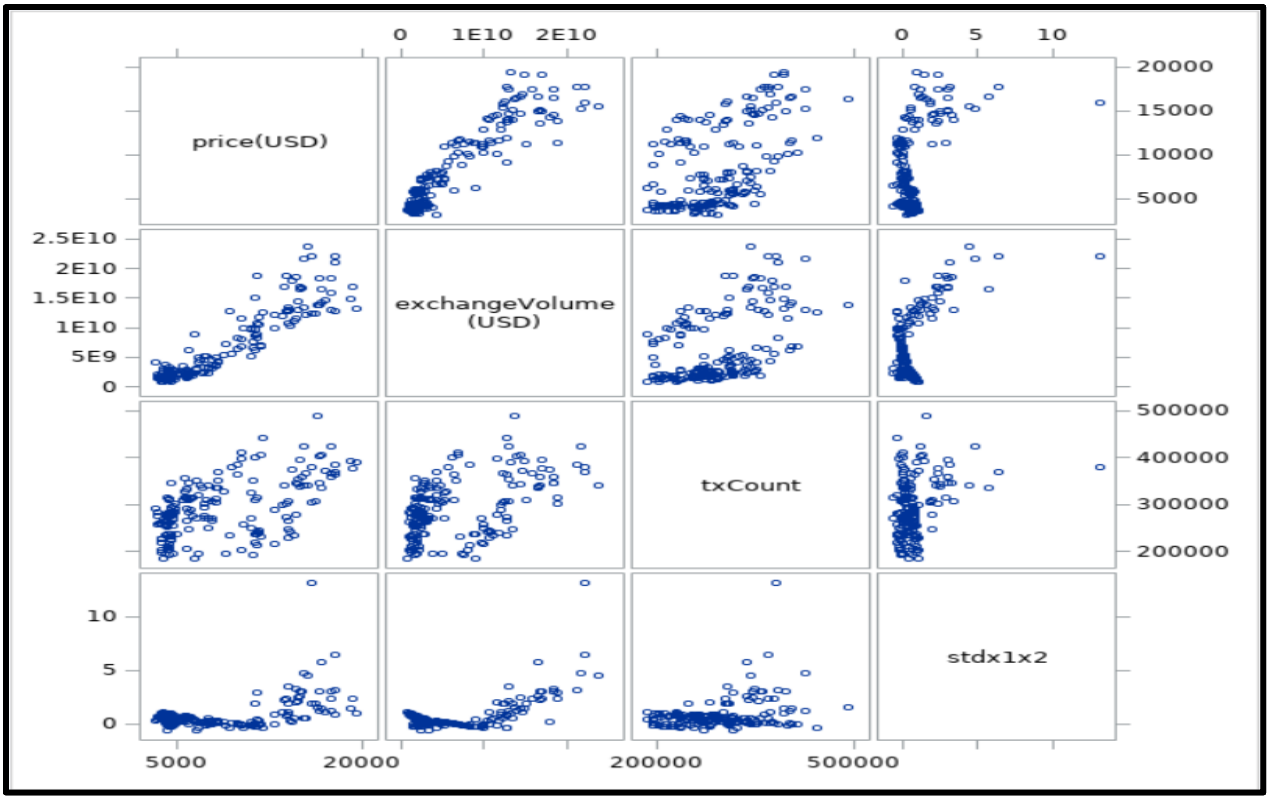

Figure 16: Matrix Scatter Plot for Second best model.

2. **To check reasonability of the model:**
From the Residual V/s Predictor variable for both models (Refer Figure 17, Figure 18), we can interpret that curvature is not observed in any of the residual v/s predictor variable plot. Therefore, we can conclude that both potential best models are reasonable.

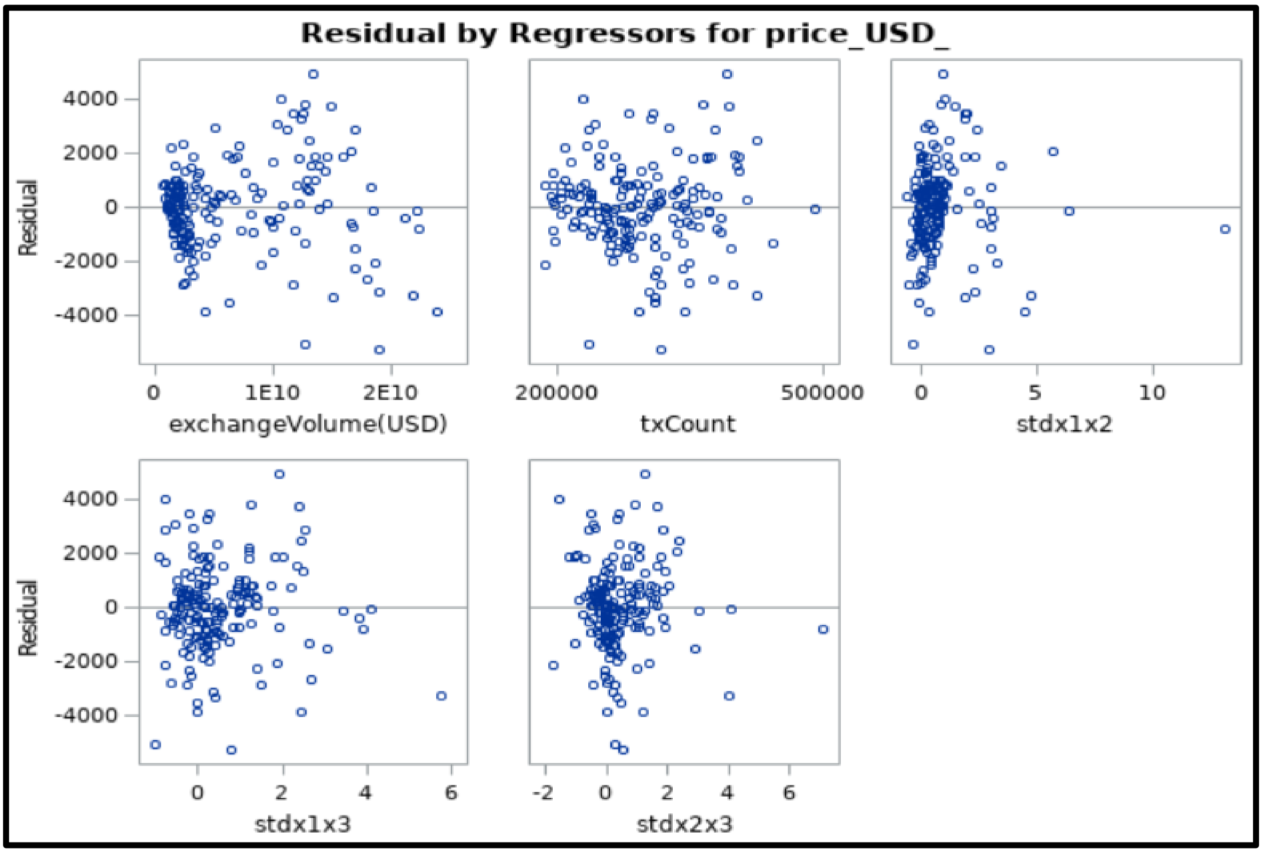

Figure 17 : Residual v/s Predictor variable for Best model.

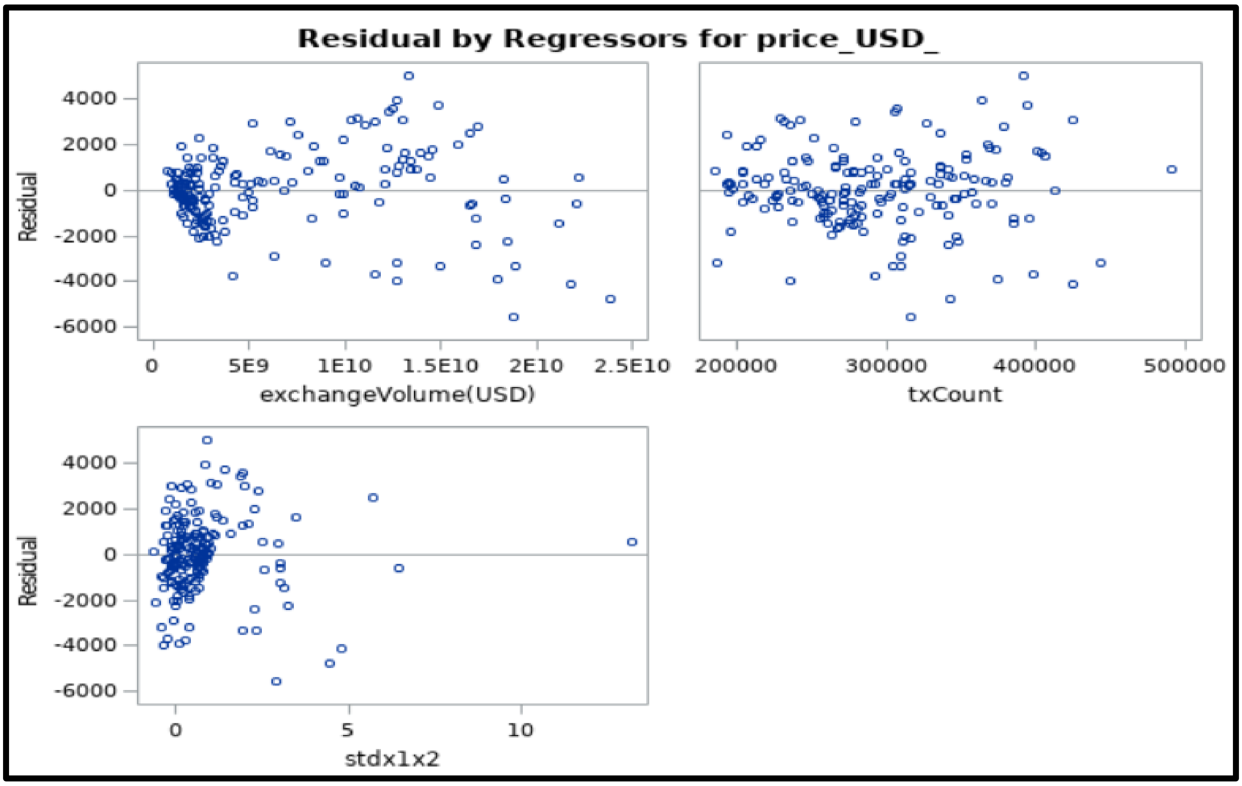

Figure 18 : Residual v/s Predictor variable for Second Best model.

3. **<u>To check that residuals, have constant variance.</u>**
From Residual v/s Predicted value of the price of bitcoin (Figure 19 – first potential best model on left, second potential best model on right), we can observe a funnel shape in both cases. Hence it can be interpreted that residual does not have constant variance.

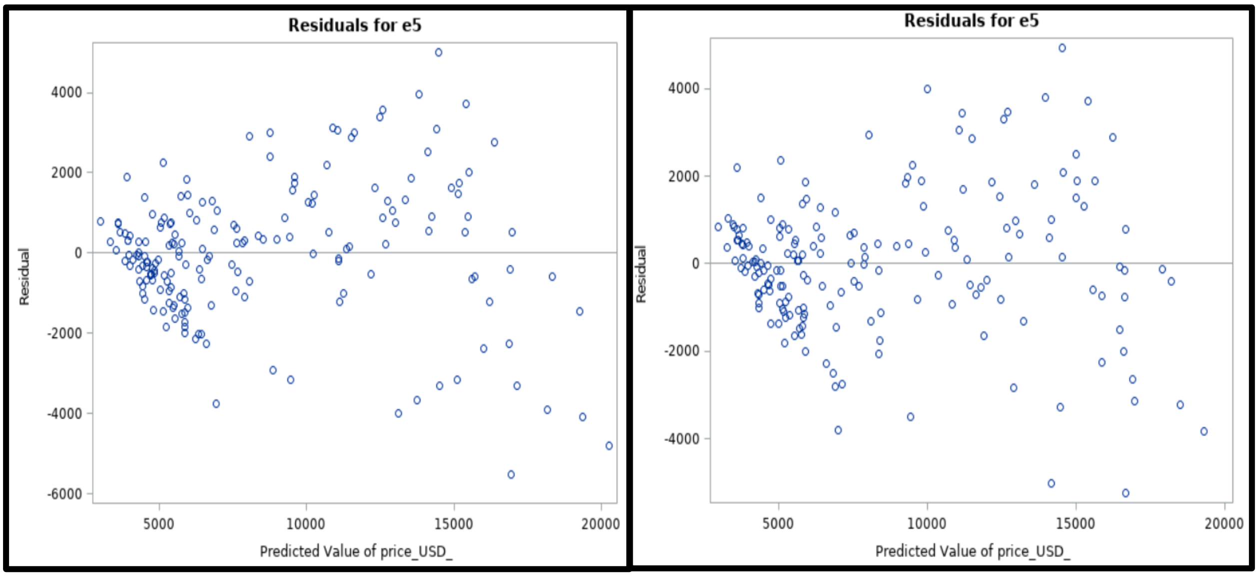

Figure 19: Residual v/s $\hat{Y}$ for Best Model and Second-best model.

The TTEST Procedure

Variable: s

group	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		90	768.2	589.9	62.1768	3.0818	2478.2
2		90	1609.1	1313.3	138.4	10.6145	5401.6
Diff (1-2)	Pooled		-840.9	1018.0	151.8		
Diff (1-2)	Satterthwaite		-840.9		151.8		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		768.2	644.7	891.8	589.9	514.5	691.3
2		1609.1	1334.1	1884.2	1313.3	1145.5	1539.2
Diff (1-2)	Pooled	-840.9	-1140.3	-541.4	1018.0	922.3	1136.0
Diff (1-2)	Satterthwaite	-840.9	-1141.3	-540.5			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	178	-5.54	<.0001
Satterthwaite	Unequal	123.5	-5.54	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	89	89	4.96	<.0001

The TTEST Procedure

Variable: s

group	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		90	762.1	582.5	61.4059	13.2590	2510.2
2		90	1636.3	1419.3	149.6	42.3754	5983.4
Diff (1-2)	Pooled		-874.3	1084.8	161.7		
Diff (1-2)	Satterthwaite		-874.3		161.7		

group	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		762.1	640.1	884.1	582.5	508.1	682.7
2		1636.3	1339.1	1933.6	1419.3	1237.9	1663.4
Diff (1-2)	Pooled	-874.3	-1193.4	-555.1	1084.8	982.9	1210.5
Diff (1-2)	Satterthwaite	-874.3	-1194.5	-554.0			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	178	-5.41	<.0001
Satterthwaite	Unequal	118.16	-5.41	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	89	89	5.94	<.0001

SAS Output 16: Modified Levene Test for First Potential Best Model (left) and Second-Best Potential (right).

Hypothesis:
H_0: $\sigma_{d1} = \sigma_{d2}$.
v/s
H_1: $\sigma_{d1} \neq \sigma_{d2}$.
Decision rule: Reject H_0 if standard deviation of both sets is not equal.
The data is divided into two groups of exchange volume.
i. Group 1: It has data points whose predicted value of the price of bitcoin is less than 6867.39.
ii. Group 2: It has data points whose value for predicted value of the price of bitcoin is greater than 6867.39
From the SAS Output 16, it can be inferred that p value (0.0001) is less than 0.05 and hence the reject H_0, indicating the means of both the groups are not same. From the SAS Output 16, for T test under unequal variances it can be inferred that p value (0.0001) is less than 0.05 and hence the reject H_0, indicating the variance of both the groups are not constant. The result of the modified Levene test (SAS Output 16) and the Residual v/s Predicted value of the price of bitcoin (Figure 19) aligns indicating that variance is not constant.

Hypothesis:
H_0: $\sigma_{d1} = \sigma_{d2}$.
v/s
H_1: $\sigma_{d1} \neq \sigma_{d2}$.
Decision rule: Reject H_0 if standard deviation of both sets is not equal.
The data is divided into two groups of exchange volume.
i. Group 1: It has data points whose predicted value of the price of bitcoin is less than 6517.83
ii. Group 2: It has data points whose value for predicted value of the price of bitcoin is greater than 6517.83
From the SAS Output 16, it can be inferred that p value (0.0001) is less than 0.05 and hence the reject H_0, indicating the means of both the groups are not same. From the SAS Output 16 for T test under unequal variances it can be inferred that p value (0.0001) is less than 0.05 and hence the reject H_0, indicating the variance of both the groups are not constant. The result of the modified Levene test (SAS Output 16) and the Residual v/s Predicted value of the price of bitcoin (Figure 19) aligns indicating that variance is not constant.

6. <u>To check that normality of the residuals.</u>

From Normal Probability Plot (Figure 20), we can observe short tail on left-side indicating normality is not satisfied.

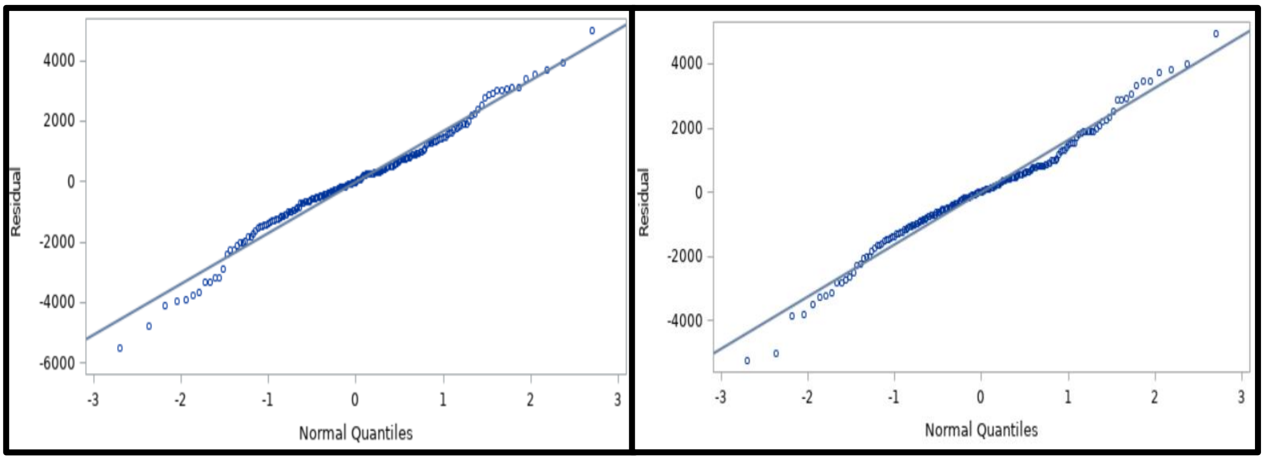

Figure 20: Modified Levene Test for Best Model (left) and Second-Best Model (right).

Pearson Correlation Coefficients, N = 180 (Best Model)

	e5	enrm5
e5 Residual	1.00000	0.98938
enrm5 Rank for Variable e5	0.98938	1.00000

Pearson Correlation Coefficients, N = 180 (Second-Best Model)

	e5	enrm5
e5 Residual	1.00000	0.98943
enrm5 Rank for Variable e5	0.98943	1.00000

SAS Output 17: Pearson Correlation Table for Best Model (left) and Second-Best Model (right)

H_0: Normality is satisfied.
V/s
H_1: Normality is violated.
Decision Rule: If $\rho < C(\alpha; n)$ then we reject H_0.
Where $\alpha = 0.1$ and $n = 180$.
The cut-off value $C(\alpha, n) = (0.1,180) = 0.99$
Decision Rule: We reject if $\rho < C(0.1, 180) = 0.98938$ (SAS Output 17, left) < 0.99. we reject H_0, hence normality is violated at a significance level of 0.1. The normality test confirms the outcome of NPP Plot.

H_0: Normality is satisfied.
V/s
H_1: Normality is violated.
Decision Rule: If $\rho < C(\alpha; n)$ then we reject H_0.
Where $\alpha = 0.1$ and $n = 180$.
The cut-off value $C(\alpha, n) = (0.1,180) = 0.99$
Decision Rule: We reject if $\rho < C(0.1, 180) = 0.98943$(SAS Output 17, right) < 0.99. we reject H_0, hence normality is violated at a significance level of 0.1. The normality test confirms the outcome of NPP Plot.

MODEL DIAGONISTICS.

7. **<u>Test for outlier of Best Model:</u>**

OBS No.	Cookdi	R-Student	Hii	DFFITS	Intercept	Exchange Volume	Total Count	stdx1x2	stdx1x3	stdx1x3
Cut-off	0.87	3.29	0.0556	0.3333	DFBETAS = 0.1491					
10	0.02	-1.704	0.035	-0.323	**0.197**	0.092	**-0.244**	0.05	**0.212**	**-0.155**
16	0.01	-1.739	0.027	-0.292	0.121	0.042	**-0.165**	**0.15**	**0.159**	**-0.194**
82	0.01	1.356	0.031	0.243	**0.176**	-0.068	-0.142	-0.01	0.119	-0.005
97	0.03	-1.330	**0.099**	**-0.441**	**-0.290**	**-0.153**	**0.289**	**-0.17**	-0.128	**0.328**
98	0.01	-2.172	0.019	-0.303	0.091	-0.062	-0.119	**0.18**	**0.177**	**-0.212**
114	0.03	-1.782	0.057	**-0.438**	0.107	-0.087	-0.111	0.12	**-0.257**	**0.206**
119	0.01	1.223	0.041	0.254	**-0.151**	-0.048	**0.176**	0.03	-0.002	-0.089
120	0.01	1.175	0.045	0.256	-0.136	-0.036	**0.158**	0.03	0.039	-0.118
121	0.02	-0.868	**0.159**	**-0.378**	0.053	-0.024	-0.055	0.04	**-0.289**	**0.239**
122	0.06	-1.703	**0.107**	**-0.590**	-0.040	**-0.265**	0.074	**0.20**	**-0.417**	**0.263**
123	0.00	-0.250	**0.086**	-0.077	-0.005	-0.023	0.011	0.00	-0.055	0.029
128	0.03	1.582	**0.073**	**0.445**	**-0.250**	0.037	**0.251**	**-0.26**	-0.027	**0.266**
129	0.00	-0.044	**0.208**	-0.022	0.012	0.001	-0.012	0.01	0.000	-0.014
131	0.02	2.365	0.022	0.353	-0.143	**0.159**	0.137	**-0.20**	0.001	0.103
132	0.06	3.131	0.035	**0.598**	**-0.273**	**0.167**	**0.264**	**-0.33**	0.105	**0.162**
133	0.04	2.338	0.040	**0.478**	**-0.182**	0.132	**0.167**	**-0.23**	0.127	0.115
135	0.00	-0.079	**0.100**	-0.026	-0.001	-0.001	0.003	-0.02	-0.006	0.006
136	0.03	1.334	**0.091**	**0.423**	-0.031	-0.018	0.023	**0.30**	**-0.155**	0.026
137	0.09	-0.668	**0.539**	**-0.723**	0.081	**0.197**	-0.075	**-0.47**	**0.214**	**-0.160**
138	0.01	0.950	0.044	0.203	0.011	0.020	-0.012	**0.15**	-0.067	-0.041
139	0.03	2.162	0.043	**0.460**	0.146	0.124	-0.146	**0.29**	0.000	**-0.289**
140	0.09	2.568	**0.082**	**0.765**	**0.390**	**0.263**	**-0.394**	**0.37**	0.127	**-0.563**
141	0.02	1.192	**0.089**	**0.372**	0.145	0.141	**-0.154**	**0.21**	-0.015	**-0.226**
147	0.02	1.898	0.030	0.331	**0.165**	**0.230**	**-0.169**	-0.01	-0.073	-0.062
150	0.13	-2.139	**0.148**	**-0.892**	0.063	-0.023	-0.017	0.03	**-0.551**	0.057
151	0.10	-2.476	**0.092**	**-0.787**	**-0.181**	**-0.331**	**0.260**	**-0.36**	**-0.308**	**0.437**
160	0.02	1.783	0.039	0.359	**0.175**	**0.243**	**-0.184**	0.02	-0.087	-0.081
162	0.03	-1.965	0.050	**-0.453**	-0.101	**-0.323**	0.145	-0.12	0.044	0.148
163	0.09	-2.359	0.047	**-0.746**	-0.146	**-0.439**	**0.213**	**-0.30**	0.010	**0.289**
164	0.02	-2.036	0.025	-0.328	-0.046	**-0.216**	0.067	-0.08	0.060	0.075
167	0.01	1.056	0.042	0.222	0.100	**0.160**	-0.105	-0.07	-0.102	0.061
173	0.01	0.820	**0.070**	0.226	0.074	0.108	-0.073	-0.13	-0.118	**0.163**

175	0.01	1.396	0.035	0.266	0.077	0.127	-0.069	**-0.16**	-0.132	**0.185**
177	0.00	-0.167	**0.069**	-0.046	-0.013	-0.023	0.013	0.03	0.025	-0.033
179	0.14	-3.260	**0.077**	**-0.943**	**-0.163**	**-0.703**	**0.207**	**0.48**	**0.575**	**-0.484**
180	0.00	-0.511	**0.072**	-0.142	-0.044	-0.064	0.043	0.08	0.074	-0.107

SAS Output 18: Outliers.

Bonferroni Outlier test:

X outlier: Since we have a large data set, we use the cut off value of 2(p/n) where n = 180 and p = 6. Therefore, we get the value of 0.0667. Now, from the SAS output we find the values of H_{ii} (Refer Output 18) which are greater than the cut off value of 0.0667. We find that observation numbers 97, 121, 122, 123, 128, 129, 135, 136, 137, 140, 141, 150, 151, 173, 177, 179 and 180 have value greater than 0.0667.

Y outlier: Since we have a large data set, we use the cut off $t(1-\alpha/2n; n-p-1)$ value where $\alpha = .10$, n = 180, p=6. Therefore, we get the value of t (0.99972; 173) = 3.290. Now, from the SAS output we find that no values of R-student (Refer Output 18) are greater than the cut off value of 3.290. Hence, there are no y outliers present.

To check the influence of Outliers on our model

An observation is said to be an influential observation when its exclusion causes major change in fitted regression function. Following are the three measures of influence:

<u>Influence on Single fitted Value – DFFITIS</u>

A case is considered influential if the absolute value of DFFITIS exceeds $2\sqrt{\frac{p}{n}}$ for large dataset. $2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{6}{180}}$ = 0.3651. In our case observations 97, 121, 122,128,136, 137, 140, 141, 150, 151 and 179. (Refer Output 18) These observations have absolute DFFITIS value greater than 0.3651. Therefore, these outliers can be considered as influential outliers.

<u>Influence on ALL fitted Value – Cook's Distance.</u>

A case is considered influential if the Cooke's Distance - D_i is greater than F (0.5; p; n-p) = F (0.5; 6; 174) =0.891. In our case, no observations exceed the cutoff cook' distance (Refer Output 18) Therefore, there are no outliers on all fitted values.

<u>Influence on Regression Coefficient – DFBETAS.</u>

A case is considered influential if the absolute value of DFBETAS exceeds $2/\sqrt{n}$ for large dataset. $2/\sqrt{n}= 2/\sqrt{180}$ = 0.1491. In our case observation 97, 121, 122, 136, 137, 150, 151, 173 and 179 (Refer SAS Output 18) These observations have absolute DFBETAS value greater than 0.1491. Therefore, these outliers can be considered as influential outliers.

From the measures discussed above, we see that observation number 97, 121, 122, 136, 137, 150, 151,177 and 179 have influence and observations 123, 129, 135 and 177 are not influential. Remedial measures should be taken for outliers that are influential.

8. **<u>Test for outlier for Second Potential Best Model:</u>**

OBS. No	Cookdi	R-Student	Hii	DFFITS	Intercept	Exchange Volume	Coins Generated	Fees
Cut-off	0.839	3.29	0.0444	0.2981	DFBETAS = 0.1491			
97	0.03534	-1.913	0.0377	**-0.3788**	**-0.3283**	**-0.2321**	**0.331**	0.0917
114	0.04602	-2.2262	0.0366	**-0.4339**	**0.2548**	**-0.1524**	**-0.2578**	**0.2694**
119	0.00892	1.0365	0.0322	0.189	-0.1431	-0.0426	**0.1647**	-0.0503
121	0.06032	-1.9405	**0.0611**	**-0.495**	**0.3418**	-0.1357	**-0.3418**	**0.2874**
122	0.0785	-2.3846	**0.0537**	**-0.5678**	0.1362	**-0.4447**	-0.0846	**0.3907**

128	0.0298	1.8665	0.0335	**0.3477**	**-0.2639**	0.0452	**0.2664**	-0.0722
129	0.00558	0.5568	**0.0669**	0.1491	-0.1296	-0.0078	0.1316	-0.0131
131	0.02603	2.3731	0.0186	**0.3269**	-0.1407	**0.1629**	0.135	-0.1419
132	0.05714	3.0503	0.0251	**0.4892**	**-0.2904**	**0.1827**	**0.2829**	**-0.1845**
133	0.03205	2.2352	0.0256	**0.3621**	**-0.2036**	0.1483	**0.1893**	-0.0909
135	0.00345	-0.3641	**0.0939**	-0.1172	0.0059	-0.0091	0.0024	-0.0838
136	0.04754	1.5459	**0.0742**	**0.4378**	0.0136	-0.0476	-0.0236	**0.3663**
137	0.04436	0.4276	**0.4913**	**0.4202**	-0.0062	-0.1391	0.0024	**0.4**
140	0.02108	1.8741	0.0238	0.2924	**0.2241**	**0.193**	**-0.2301**	-0.0449
147	0.0193	1.841	0.0226	0.2797	**0.1811**	**0.2123**	**-0.1862**	-0.125
150	0.10285	-2.5209	**0.0625**	**-0.6511**	**0.25**	-0.1221	**-0.1961**	**-0.2935**
151	0.14285	-2.9746	**0.0632**	**-0.7726**	-0.0705	**-0.4406**	**0.1734**	**-0.2186**
160	0.0201	1.7193	0.0268	0.2851	**0.1914**	**0.2208**	**-0.201**	-0.114
162	0.03737	-1.9962	0.0367	**-0.3899**	-0.1038	**-0.3188**	**0.1519**	0.0429
163	0.09703	**-2.4029**	0.0343	**-0.6414**	-0.1257	**-0.4379**	**0.2**	-0.0693
164	0.01879	-1.9796	0.0191	-0.2764	-0.055	**-0.2002**	0.0772	0.0195
167	0.01501	1.3232	0.0333	0.2456	**0.1757**	**0.1828**	**-0.1809**	-0.1164
173	0.01781	1.445	0.0332	0.2677	**0.2222**	**0.1638**	**-0.2186**	-0.105
175	0.01449	1.7917	0.018	0.2422	**0.1753**	0.1424	**-0.1634**	-0.1151
179	0.07037	-2.4312	**0.0467**	**-0.5379**	**-0.2804**	**-0.4632**	**0.3123**	**0.34**

SAS Output 19: Outliers.

Bonferroni Outlier test:
X outlier: Since we have a large data set, we use the cut off value of $2(p/n)$ where n = 180 and p = 4. Therefore, we get the value of 0.0444. Now, from the SAS output we find the values of H_{ii} (Refer SAS Output 19) which are greater than the cut off value of 0.0444. We find that observation numbers 121, 122, 129, 135, 136, 137, 150, 151 and 179 have value greater than 0.0444.

Y outlier: Since we have a large data set, we use the cut off t(1-α/2n; n-p-1) value where α = .10, n = 180, p=4. Therefore, we get the value of t (0.99972; 175) = 3.290. Now, from the SAS output we find that no values of R-student (Refer SAS Output 19) are greater than the cut off value of 3.290. Hence, there are no y outliers present.

To check the influence of Outliers on our model
An observation is said to be an influential observation when its exclusion causes major change in fitted regression function. Following are the three measures of influence:

<u>Influence on Single fitted Value – DFFITIS</u>

A case is considered influential if the absolute value of DFFITIS exceeds $2\sqrt{\frac{p}{n}}$ for large dataset. $2\sqrt{\frac{p}{n}} = 2\sqrt{\frac{4}{180}}$ = 0.2981. In our case observations 121, 122, 136, 137, 150, 151 and 179. (Refer SAS Output 19) These observations have absolute DFFITIS value greater than 0.2981. Therefore, these outliers can be considered as influential outliers.

<u>Influence on ALL fitted Value – Cook's Distance.</u>
A case is considered influential if the Cooke's Distance - D_i is greater than F (0.5; p; n-p) = F (0.5; 4; 176) =0.839. In our case, no observations exceed the cutoff cook' distance (Refer SAS Output 19) Therefore, there are no outliers on all fitted values.

Influence on Regression Coefficient – DFBETAS.

A case is considered influential if the absolute value of DFBETAS exceeds $2/\sqrt{n}$ for large dataset. $2/\sqrt{n} = 2/\sqrt{180} = 0.1491$. In our case observation 121, 122, 136, 137, 150, 151 and 179 (Refer SAS Output 19) These observations have absolute DFBETAS value greater than 0.1491. Therefore, these outliers can be considered as influential outliers.

From the measures discussed above, we see that observation number 121, 122,136,137 150, 151 and 179 have influence and observations 129 and 135 are not influential. Remedial measures should be taken for outliers that are influential.

9. Time series plots

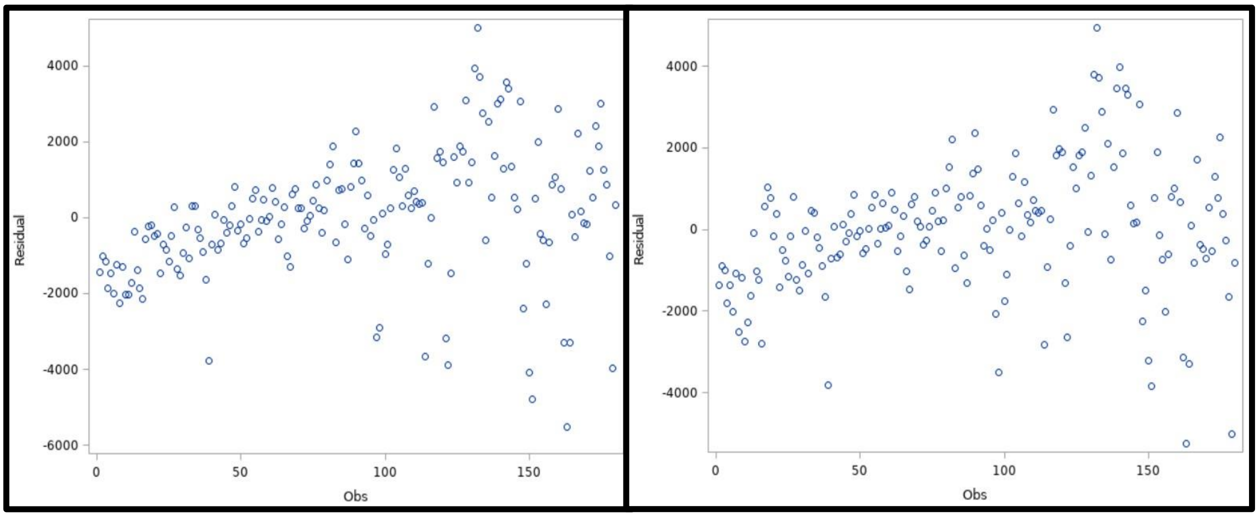

Figure 21: Time Series Plot for Best Model (left) and Second-Best Model (right).

From the Time Series Plot (Figure 21) we can infer the following: there is random jaggedness upto 110 observations and upward linear trend. Therefore, some errors are correlated for which remedy is required.

10. **To check (VIF)variance inflation**: Average VIF value for first best model = 2.58 < 5. So, we don't have any serious multicollinearity issues. Average VIF value for second best model = 1.61 < 5. So, we don't have any serious multicollinearity issues.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Variance Inflation
Intercept	Intercept	1	345.05564	674.66194	0.51	0.6097	12887895622	0
exchangeVolume_USD_	exchangeVolume(USD)	1	7.62875E-7	2.957888E-8	25.79	<.0001	3244591565	1.96607
txCount	txCount	1	0.01253	0.00248	5.05	<.0001	44537705	1.48576
stdx1x2		1	-632.40263	144.22344	-4.38	<.0001	48000645	2.78811
stdx1x3		1	-683.17616	204.76852	-3.34	0.0010	6341835	3.01854
stdx2x3		1	788.66182	235.46249	3.35	0.0010	30427080	3.70750

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Variance Inflation
Intercept	Intercept	1	1157.83333	656.58669	1.76	0.0796	12887895622	0
exchangeVolume_USD_	exchangeVolume(USD)	1	7.497946E-7	3.022691E-8	24.81	<.0001	3244591565	1.92665
txCount	txCount	1	0.00959	0.00242	3.96	0.0001	44537705	1.32462
stdx1x2		1	-461.11703	113.15111	-4.08	<.0001	48000645	1.61041

SAS Output 22: Variance Inflation for Best Model (top) and Second-Best Model (bottom)

Model Assumptions analysis for First and Second-best models
1. The residuals do not have constant variance.
2. The residuals are not normally distributed.
3. Some residuals are correlated.
4. Some x outliers are observed which have influence.

Selection of Final model:

We select First Potential Best Model: $Y_i = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \varepsilon_i$. in comparison with Second Potential Best Model: $Y_i = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \varepsilon_i$. as our final model because its C(p) is closer to number of parameters, it has a lower AIC and SBC value, a higher R^2 value (Refer SAS Output 24)

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
3	0.8655	0.8677	15.3715	2681.7914	2694.56323	exchangeVolume_USD_ txCount stdx1x2
3	0.8607	0.8630	22.0520	2688.0973	2700.86913	exchangeVolume_USD_ txCount x1x2

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
5	0.8738	0.8773	5.8280	2672.2866	2691.44434	exchangeVolume_USD_ txCount stdx1x2 stdx1x3 stdx2x3
5	0.8677	0.8714	14.1450	2680.6992	2699.85691	fees txCount stdx1x2 stdx2x3 x1x2

SAS Output 24: Comparison between Best Model (top) and Second-Best Model (bottom).

Final Model:

$Y_i = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \varepsilon_i$.

Where, Y_i = Price of Bitcoin, X_1 = Exchange Volume, X_2= Total Count, X_3= stdx1x2, X_4 = stdx1x3, X_5 = stdx2x3, ε_i = Error.

VI. Final Multiple Linear Regression Model:

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	345.05564	674.66194	0.51	0.6097
exchangeVolume_USD_	exchangeVolume(USD)	1	7.62875E-7	2.957888E-8	25.79	<.0001
txCount	txCount	1	0.01253	0.00248	5.05	<.0001
stdx1x2		1	-632.40263	144.22344	-4.38	<.0001
stdx1x3		1	-683.17616	204.76852	-3.34	0.0010
stdx2x3		1	788.66182	235.46249	3.35	0.0010

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	3373898830	674779766	248.79	<.0001
Error	174	471923225	2712202		
Corrected Total	179	3845822055			

Root MSE	1646.87657	R-Square	0.8773
Dependent Mean	8461.64406	Adj R-Sq	0.8738
Coeff Var	19.46284		

SAS Output 25: Parameter Estimates and ANOVA for Best Model.

$\hat{Y}$ = 345.05564+0.000000762875*(Exchange Volume) + 0.01253*(Total Count) - 632.40263 *(Standardized Exchange Volume*Standardized Generated Coins) - 683.17616*(Standardized Exchange Volume * Standardized Fees) + 788.66182*(Standardized Generated Coins * Standardized Fees) (Refer SAS Output 25)

Where b_0= 345.05, b_1= 0.000000762875, b_2= 0.01253, b_3= - 632.40263, b_4= - 683.17616, b_5= 788.66182

$X_h = 159$; $\hat{Y}_h = 67475.544$

Bonferroni Joint Confidence Interval for the parameters:
Bonferroni coefficient,
$B = t\left(1 - \frac{\alpha}{2g}; n - p\right)$
$= t\left(1 - \frac{0.1}{2(4)}; 180 - 6\right)$
$= t\,(0.9875;175) = =2.326$
CI for,
$\beta_1 = b_1 \pm B\times s\{b_1\}$
$=7.62875 \times 10^{-7} \pm 2.326*2.957888 \times 10^{-8}$
$= (6.94*10^{-7}, 8.32*10^{-7})$
$\beta_2 = b_2 \pm B\times s\{b_2\}$
$= 0.01253 \pm 2.326*0.00248$
$= (6.76*10^{-3}, 0.0183)$
$\beta_3 = b_3 \pm B\times s\{b_3\}$
$=-632.40263 \pm 2.326*144.22344$
$= (-968.3, -297.4437)$
$\beta_4 = b_4 \pm B\times s\{b_4\}$
$= -683.17616 \pm 2.326*204.76852$
$= (-116.7510, -207.6013)$
$\beta_5 = b_5 \pm B\times s\{b_5\}$
$= 788.66182 \pm 2.326*235.46249$
$= (241.80, 1334.5234)$
Therefore, we are 90% confident that β_1 lies between $(6.94*10^{-7}, 8.32*10^{-7})$, β_2 lies between $(6.76*10^{-3}, 0.0183)$, β_3 lies between (-968.3, -297.4437), β_4 lies between (-116.7510, -207.6013) and β_5 lies between (241.80, 1334.5234) simultaneously.

xnew159					
159	1.2764E10	265586	0.8164687	-0.483764	-0.340233

newyhat1	syhat1	spred1	t	cl90	cu90	pl90	pu90
67475.544	1.1384E10	1.1387E10	1.653658	-108966	243917.09	-108987	243938.11

SAS Output 26

Confidence Band for mean response at X_{159}

$S\{\hat{Y}_h\} = (MSE\ (X^t_h\ (X^TX)^{-1}\ X_h))^{1/2}$ (SAS Output 26)
$= 1.1384*10^{10}$
Two-sided confidence band:
$\hat{Y}_h \pm B*s\{\hat{Y}_h\}$
$= 67475.544 \pm 2.326*\ \text{sqrt}\ (1.1384*10^{10})$
$= (0, 315650.043)$

Confidence Interval for mean response at X_{159}

t-value $= t\,(1-\alpha/2;\ n-p) = t\,(0.95,175) = 1.645$
$\hat{Y} \pm t\,(1-\alpha/2;\ n-p)\ s\{\hat{Y}\}$
$=67475.544 \pm 1.65*\text{sqrt}1.1384*10^{10}$
$= (0, 243917.09)$
The value of Price cannot be negative, so the lower boundary is taken as 0.
Thus, we are 90% confident that the Price for X_{159} lies between (0, 243917.09) USD

Prediction Interval for mean response at X_{159}

$S_{pred} = \sqrt{(s^2\{\hat{Y}\} + MSE)} = \sqrt{1.1387 \text{ x } 10\text{^}10}$ = 106709.9 (SAS Output 26)
PI for the actual observation,
$\hat{Y} \pm t\,(1\text{-}\alpha/2;\, n\text{-}p)\, S_{pred}$
= 67475.544± 1.645*106607.9
= (0,243937.2024)
The value of Price cannot be negative, so the lower boundary is taken as 0.
Thus, we are 90% confident that the actual value of the Price will lie in between the values (0,243937.2024) USD

Confidence Region for the entire regression surface at x_{159}

Working-Hotelling coefficient:
W^2 = P*F (1-α, p, n-p) = 6*F (0.90,5,175) = 6*1.85 = 11.1
Thus, W = 3.331
The confidence region = $\hat{Y}_h \pm W{*}s\{\hat{Y}_h\}$
= 67475.544 ± 3.331* sqrt (1.1384*10^{10})
= (0, 422879.3516)
The value of Price cannot be negative, so the lower boundary is taken as 0.
Thus, with a confidence level of 90% we can conclude that the entire regression surface for the mean response of Price at x_{159} will lie between the boundary values of (0,422879.3516) USD

IV. Final Discussion

In the preliminary model I have used the variables Exchange Volume (USD), Coins Generated, Fees (BTC) and Total Count of transactions to predict the price of Bitcoin. From verification of model assumptions, we see that the model has non-constant variance with normality satisfied. Hence, I performed Weighted least square transformation.

After performing weighted transformation, non-constant variance is observed, and normality is violated. Thus, we performed log transformation on dependent variable which resulted in constant variance and satisfied normality, but Residual Vs Exchange volume plot showed a curvilinear relationship. As a remedy we added a quadratic term of exchange volume to our model. Performing residual analysis for the transformed model with quadratic term, all model assumptions were satisfied except for variance inflation. The average variance inflation value being greater than 10 results in the model having serious multicollinearity issues. Of all these transformed models, we selected the model with only weighted least square transformation as our model to perform analysis on because it has no serious multicollinearity issues.

The interaction terms are identified and added to our initial model after performing partial regression and standardization. After analysing the plots of interaction terms, we decided to add three standardized interaction terms - stdx1x2 stdx1x3, stdx2x3 and a partial regression term x1x2 to our weighted least square model. The potential best model i.e. Final mode fits well with our data with $R^2 = 0.8773$ and an adjusted R^2 of 0.8738 and has variance inflation factor values (VIF) average of 2.59 < 5, thus indicating no serious multicollinearity issue. I observed some outliers which have an influence on our model addition, with non-constant variance and normality is satisfied.

Further Analysis
Log transformation and addition of the quadratic term for Exchange volume should be added to resolve the non-constant variance. Variance inflation issue could be resolved by complex Regression techniques. These suggestions might prove worthy in future research and analysis; However, they are not required to be performed in this project.

References:

Main link to understanding of data: https://coinmetrics.io/on-data-and-certainty/.
https://www.coinist.io/cryptocurrency-trading-volume/.

• https://support.bitpay.com/hc/en-us/articles/115003393863-What-are-bitcoin-miner-fees- Why-are-miner-fees-so-high

• Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). Applied Linear Statistical Models, 5th Edition. New York: McGraw-Hill Education.

Data Set:
Here, following is the raw data, where Y is the dependent variable (i.e. Price) and X1, X2, X3, and X4are independent variables - Exchange Volume, Fees, Coins Generated and Tx Count respectively.

Y	X1	X2	X3	X4
$3,370.22	$1,752,760,000	159.06	275410	2050.00
$3,420.40	$1,468,960,000	157.95	260817	1725.00
$3,341.84	$1,515,110,000	193.19	257831	1675.00
$3,373.82	$2,021,190,000	234.94	283857	1800.00
$3,650.63	$2,219,590,000	210.75	260386	1687.50
$3,880.04	$3,159,090,000	189.71	263162	1850.00
$4,066.10	$2,463,090,000	249.19	255232	1625.00
$4,326.99	$3,258,050,000	350.45	310850	1900.00
$4,200.34	$2,272,040,000	347.53	274736	1625.00
$4,384.44	$2,553,360,000	365.32	347140	2237.50
$4,324.34	$2,941,710,000	337.96	312101	1862.50
$4,137.75	$2,975,820,000	265.91	266681	1887.50
$4,189.31	$2,109,770,000	208.98	210722	1625.00
$4,090.48	$2,800,890,000	297.96	236657	1437.50
$3,998.35	$3,764,240,000	334.18	196194	1112.50
$4,089.01	$2,369,820,000	541.97	315575	1987.50
$4,137.60	$2,037,750,000	386.00	203704	1262.50
$4,332.82	$1,727,970,000	425.82	208068	1262.50
$4,372.06	$1,511,610,000	425.31	226380	1312.50
$4,345.10	$1,537,460,000	358.84	261736	1937.50
$4,384.45	$1,959,330,000	378.18	223642	1612.50
$4,389.21	$2,486,080,000	461.76	280967	1862.50
$4,570.36	$1,937,850,000	388.50	271298	2100.00
$4,555.59	$1,944,930,000	413.21	280561	2037.50
$4,701.76	$2,599,080,000	382.74	283496	2175.00
$4,901.42	$2,722,140,000	290.50	236647	2050.00
$4,585.27	$1,933,190,000	239.14	195135	1925.00
$4,591.63	$2,987,330,000	325.40	269105	2187.50
$4,228.29	$2,697,970,000	281.90	277751	2462.50
$4,376.59	$2,172,100,000	272.60	276063	2025.00
$4,589.14	$1,844,620,000	257.34	256198	2200.00
$4,605.16	$2,700,890,000	217.77	279313	2187.50
$4,229.81	$1,386,230,000	144.85	217721	2112.50
$4,229.34	$1,679,090,000	143.24	193102	1725.00
$4,122.47	$1,557,330,000	159.73	253318	2175.00

$4,168.88	$1,864,530,000	148.91	256042	2350.00
$4,131.98	$2,219,410,000	133.73	263130	2375.00
$3,875.37	$2,716,310,000	153.74	267894	2175.00
$3,166.30	$4,148,070,000	174.82	291899	2250.00
$3,637.75	$1,818,400,000	119.92	228088	2112.50
$3,606.28	$1,239,150,000	99.41	197507	2200.00
$3,591.09	$1,943,210,000	168.31	218077	1587.50
$4,073.79	$1,563,980,000	173.70	282761	1950.00
$3,916.36	$1,213,830,000	146.41	235606	1462.50
$3,901.47	$1,411,480,000	136.35	255081	2200.00
$3,628.02	$1,194,830,000	117.16	227476	1737.50
$3,629.92	$928,114,000	99.70	203695	2075.00
$3,796.15	$768,015,000	83.68	185132	1812.50
$3,681.58	$1,374,210,000	147.37	226478	1562.50
$3,928.41	$1,043,740,000	150.09	263510	1725.00
$3,892.94	$1,686,880,000	143.50	257814	1837.50
$4,197.13	$1,712,320,000	134.45	275799	2212.50
$4,171.62	$1,367,050,000	115.33	254269	2112.50
$4,166.11	$1,207,450,000	101.44	212540	1725.00
$4,341.05	$1,208,210,000	119.05	203523	1625.00
$4,395.81	$1,431,730,000	180.47	294852	1875.00
$4,408.46	$1,288,020,000	152.53	281949	2287.50
$4,319.37	$1,116,770,000	136.73	232692	1675.00
$4,229.88	$1,161,770,000	160.34	273118	1562.50
$4,324.46	$1,069,940,000	149.02	283540	1987.50
$4,369.35	$906,928,000	108.22	231068	1975.00
$4,429.67	$1,313,870,000	111.80	238203	2275.00
$4,614.52	$1,968,740,000	149.34	296785	2012.50
$4,776.21	$1,597,140,000	167.26	302941	2000.00
$4,789.25	$1,222,280,000	148.17	292320	1737.50
$4,829.58	$2,791,610,000	174.41	293011	1912.50
$5,464.16	$3,615,480,000	281.53	309683	1700.00
$5,643.53	$1,669,030,000	227.60	292999	1762.50
$5,835.96	$1,976,040,000	162.71	283278	2437.50
$5,687.57	$2,008,070,000	193.42	314551	2175.00
$5,741.58	$1,821,570,000	209.14	334274	2050.00
$5,603.82	$2,399,270,000	192.20	329428	1887.50
$5,583.74	$1,780,540,000	191.51	356807	2225.00
$5,708.11	$2,354,430,000	161.97	312230	2237.50
$5,996.79	$2,207,100,000	167.77	312104	1912.50
$6,036.66	$2,034,630,000	155.33	288949	2275.00
$6,006.00	$2,401,840,000	185.64	315925	2137.50
$5,935.52	$2,735,700,000	199.79	347053	2087.50
$5,524.60	$1,966,990,000	141.98	315230	2312.50
$5,747.95	$1,905,040,000	112.65	265647	2250.00
$5,899.74	$1,710,130,000	166.67	246588	1300.00
$5,787.82	$1,403,920,000	175.04	206981	1062.50

$5,754.44	$2,859,040,000	278.86	333871	1925.00
$6,114.85	$1,772,150,000	263.59	315930	1837.50
$6,132.02	$2,311,380,000	231.89	277397	1725.00
$6,440.97	$2,870,320,000	296.94	351373	1950.00
$6,777.77	$4,653,770,000	277.59	341324	1875.00
$7,087.53	$3,369,860,000	271.77	277479	1487.50
$7,164.48	$2,483,800,000	273.25	293991	1625.00
$7,404.52	$2,380,410,000	197.66	251587	1762.50
$7,403.22	$3,111,900,000	240.34	270896	1662.50
$7,023.10	$2,326,340,000	294.13	335480	1950.00
$7,141.38	$4,602,200,000	262.17	301202	1800.00
$7,446.83	$3,226,250,000	361.33	341128	1912.50
$7,173.73	$5,208,250,000	319.97	271625	1500.00
$6,618.61	$4,908,680,000	317.06	194554	1000.00
$6,295.45	$8,957,350,000	583.61	185886	1000.00
$5,938.25	$6,263,250,000	783.39	309159	1750.00
$6,561.48	$3,197,110,000	615.09	271867	1600.00
$6,634.76	$4,200,880,000	603.51	321636	1900.00
$7,323.24	$5,123,810,000	502.64	310244	1900.00
$7,853.57	$4,651,670,000	401.17	306450	2087.50
$7,697.21	$3,667,190,000	289.99	270738	1700.00
$7,766.03	$3,149,320,000	247.44	264695	2000.00
$8,039.07	$3,488,450,000	329.02	336029	2187.50
$8,205.74	$4,277,610,000	310.30	370918	2162.50
$8,077.95	$3,633,530,000	256.40	311885	1850.00
$8,232.38	$4,225,180,000	265.92	352050	1962.50
$8,074.02	$5,058,610,000	233.54	305586	1862.50
$8,241.71	$4,342,060,000	195.64	336533	1950.00
$8,789.04	$5,475,580,000	191.22	329524	1912.50
$9,352.72	$5,653,320,000	236.50	379086	2125.00
$9,823.43	$6,348,820,000	245.23	365821	1950.00
$10,077.40	$11,568,800,000	290.90	397917	2162.50
$9,906.79	$8,310,690,000	270.63	384219	2087.50
$10,198.60	$6,783,120,000	247.05	412725	2287.50
$10,978.30	$5,138,500,000	190.30	326193	2012.50
$11,082.70	$6,608,310,000	187.45	352868	2200.00
$11,315.40	$6,132,410,000	226.12	400505	2262.50
$11,685.70	$6,895,260,000	219.68	405531	2225.00
$11,923.40	$12,656,300,000	257.17	443399	2350.00
$14,266.10	$17,950,700,000	364.61	374765	1900.00
$17,802.90	$21,136,000,000	641.43	384936	1950.00
$16,523.30	$13,911,300,000	567.09	403225	2175.00
$15,168.40	$13,433,300,000	458.10	341256	2050.00
$15,427.40	$12,153,900,000	483.64	368427	2137.50
$16,919.80	$14,603,800,000	537.94	372821	1950.00
$17,500.00	$12,976,900,000	604.30	424393	2325.00
$16,384.60	$13,777,400,000	644.61	490459	2312.50

$16,601.30	$14,310,000,000	592.70	405507	2112.50
$17,760.30	$12,740,600,000	534.01	364051	2137.50
$19,475.80	$13,314,600,000	535.44	391725	2312.50
$19,106.40	$14,839,500,000	587.13	394057	2275.00
$19,118.30	$16,894,500,000	662.60	378482	1750.00
$17,760.30	$22,149,700,000	911.59	370141	1700.00
$16,642.40	$16,516,600,000	1128.76	335350	1650.00
$15,898.00	$22,198,000,000	1495.95	380493	1937.50
$13,948.70	$13,086,000,000	1065.62	308072	1737.50
$14,608.20	$11,572,300,000	881.12	279371	1900.00
$13,995.90	$10,664,700,000	689.39	228791	1687.50
$14,036.60	$13,454,300,000	725.62	247298	1775.00
$16,163.50	$12,487,600,000	784.82	307486	1950.00
$15,864.10	$12,336,500,000	785.83	304904	1862.50
$14,695.80	$13,025,500,000	785.80	353659	2062.50
$14,681.90	$14,452,600,000	778.57	344260	1825.00
$12,897.70	$12,136,300,000	545.14	290259	2037.50
$14,112.20	$10,291,200,000	469.41	241601	1950.00
$13,625.00	$16,846,600,000	649.16	340809	2137.50
$14,978.20	$16,871,900,000	749.69	395806	1962.50
$15,270.70	$21,783,200,000	778.26	424840	2100.00
$15,477.20	$23,840,900,000	701.36	342564	1787.50
$17,462.10	$18,314,600,000	694.48	358679	2100.00
$17,527.30	$15,866,000,000	686.11	368025	2025.00
$16,476.20	$18,413,900,000	694.53	345506	2075.00
$15,123.70	$16,660,000,000	752.42	360101	2050.00
$14,588.50	$18,500,800,000	720.33	347227	2075.00
$14,968.20	$16,534,100,000	697.44	337766	2412.50
$13,453.90	$12,065,700,000	617.64	299913	2225.00
$13,952.40	$12,763,600,000	528.47	265586	1862.50
$14,370.80	$11,084,100,000	491.32	234890	1937.50
$13,767.30	$12,750,800,000	510.12	273473	1875.00
$13,836.10	$18,853,800,000	606.38	303566	1987.50
$11,431.10	$18,830,600,000	675.10	315604	2087.50
$11,198.80	$15,020,400,000	656.90	309322	2162.50
$11,429.80	$10,740,400,000	491.49	243454	2062.50
$11,656.20	$11,801,700,000	429.40	240433	2225.00
$12,889.20	$9,935,180,000	343.36	215435	2337.50
$11,633.10	$10,537,400,000	314.61	245395	2087.50
$10,944.50	$9,660,610,000	285.96	271759	2212.50
$10,903.40	$9,940,990,000	283.50	250247	2200.00
$11,421.70	$8,873,170,000	220.85	236422	1975.00
$11,256.00	$9,746,200,000	209.13	220304	1937.50
$11,174.90	$7,583,270,000	168.25	193421	1962.50
$11,475.30	$8,350,360,000	155.34	213288	1975.00
$11,755.50	$7,107,360,000	145.75	232028	1850.00
$11,306.80	$8,637,860,000	160.30	236442	2087.50

$10,108.20	$8,041,160,000	162.21	204159	1462.50
$10,237.30	$9,959,400,000	214.03	257504	2000.00
$9,142.28	$12,726,900,000	274.36	235750	2100.00