Project (SI -424)

On

Deployment of various Model using Regression Analysis

Submitted as a course project (Regression Analysis SI -424)
Department of Mathematics (IIT- Bombay)

By

Gautam Patel (22N067)

Shiv Yadav (22N0064)

Utkarsh (22N0063)

Under the guidance of

Prof. Siuli Mukhopadhyay

Department of Mathematics

Indian Institute of Technology Bombay

April, 2023

Certificate

This is to certify that the project report entitled as **"Deployment of various Models using Regression Analysis"** was prepared by Gautam Patel, Shiv Yadav and Utkarsh student of M.Sc ASI, Department of Mathematics, IIT Bombay under my supervision and guidance.

This report is his original work and up to standard and is approved for submission.

Date: 26 / April /2023

Prof. Siuli Mukhopadhyay,

Department of Mathematics,

IIT Bombay

DECLARATION

I declare that this written submission represents my ideas in my words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will cause disciplinary action by the Institute and can also evoke penal action from the sources, which have thus not been properly cited, or from whom proper permission has not been taken when needed.

Place:

Date:

Gautam Patel(22N0067)
Shiv Yadav(22N0064)
Utkarsh(22N0063)

INTRODUCTION

Regression analysis is a statistical method for estimating the relationship between two or more variables. It is a widely used technique in data analysis to identify the strength of the relationship between a dependent variable and one or more independent variables. In this project report, we will analyze a dataset using regression analysis and discuss the findings in detail.

Regression analysis is a powerful statistical tool that allows us to explore the relationship between a dependent variable and one or more independent variables. High level regression analysis is an advanced statistical technique that is used to model complex relationships and identify significant predictors that can help explain and predict the behavior of the dependent variable.

The purpose of this project report is to present the results of a high level regression analysis that was conducted to investigate the relationship between a set of independent variables and a dependent variable of interest. The report will provide a detailed description of the data used in the analysis, the statistical methods employed, and the findings of the analysis.

Through this analysis, we hope to gain a better understanding of the underlying factors that influence the behavior of the dependent variable, and to identify potential areas for further research or intervention. The findings of this analysis may have important implications for a variety of fields, including business, finance, healthcare, and social science.

Overall, this project report represents an important contribution to the field of statistical analysis and provides valuable insights into the complex relationships that exist between variables in our world.

OBJECTIVE

The main objective of this project is to perform a regression analysis on a dataset to identify the relationship between the dependent and independent variables and deploy the various models using Regression techniques. We aim to accomplish the following objectives:

- **Identify the variables that are most strongly correlated with the dependent variable.**
- **Determine the significance of the relationship between the dependent and independent variables.**
- **Develop a predictive model for the dependent variable using the independent variables.**
- **Evaluate the model's performance and accuracy.**

Dataset:

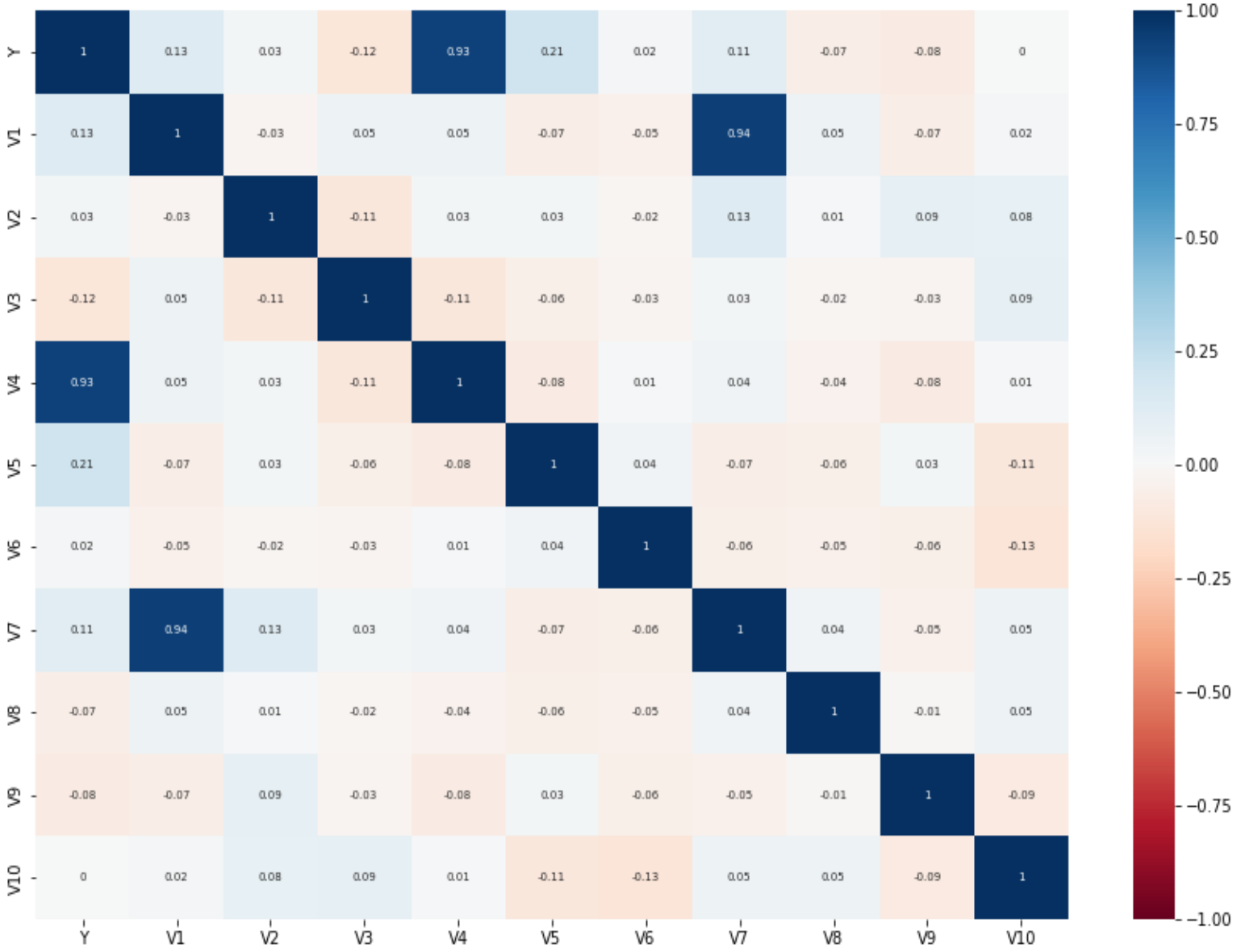
	Unnamed: 0	Y	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
0	1	3.955299e+07	2.705616	8.010346	4.124158	6.085618	2.720406	6.666423	7.604752	6.165214	1	0
1	2	1.113885e+08	5.221836	10.097714	0.800022	8.442870	0.409000	7.120131	10.637852	6.953741	1	0
2	3	4.091115e+07	4.288222	8.762010	3.751944	5.939178	2.666111	6.113596	9.573163	4.890891	1	0
3	4	4.923284e+07	4.061546	9.595422	4.872827	6.791508	0.049650	6.198104	9.166444	4.608687	1	0
4	5	1.040508e+08	4.788482	8.548239	1.529381	8.023423	2.282913	8.070690	9.851151	6.101992	1	0
5	6	6.113850e+07	4.720057	10.504637	3.037831	6.239784	5.318873	6.756406	9.795869	4.379351	1	0
6	7	1.045229e+08	3.675749	7.640511	1.190308	7.759436	5.106688	6.262776	8.463551	6.540107	1	0
7	8	4.473807e+07	4.685888	8.652104	1.228506	5.967075	3.211363	4.979370	9.500369	6.960119	1	0
8	9	4.396231e+07	5.373471	10.213209	0.474869	6.435069	0.233425	5.840306	10.898799	7.403333	1	0
9	10	8.731712e+07	4.852295	9.006450	8.674073	7.896944	0.415595	5.918222	9.794931	6.711999	1	0
10	11	4.009556e+07	4.057582	8.615965	2.389965	6.340041	0.360051	6.425868	9.227637	4.070758	1	0
11	12	8.813539e+07	5.152161	9.532239	3.382079	7.611699	2.120072	6.014867	10.825422	3.883686	1	0
12	13	5.136089e+07	4.763663	9.081523	18.895992	6.603088	1.294280	5.787510	10.154710	6.579959	1	0
13	14	7.362738e+07	4.896709	9.461965	4.747575	7.237305	2.044667	6.303415	9.844133	6.685294	1	0
14	15	4.891150e+07	5.585205	9.507028	0.646727	6.394218	1.398805	5.814855	10.672135	3.943432	1	0

Methodology:

We will use the following methodology to perform the regression analysis:

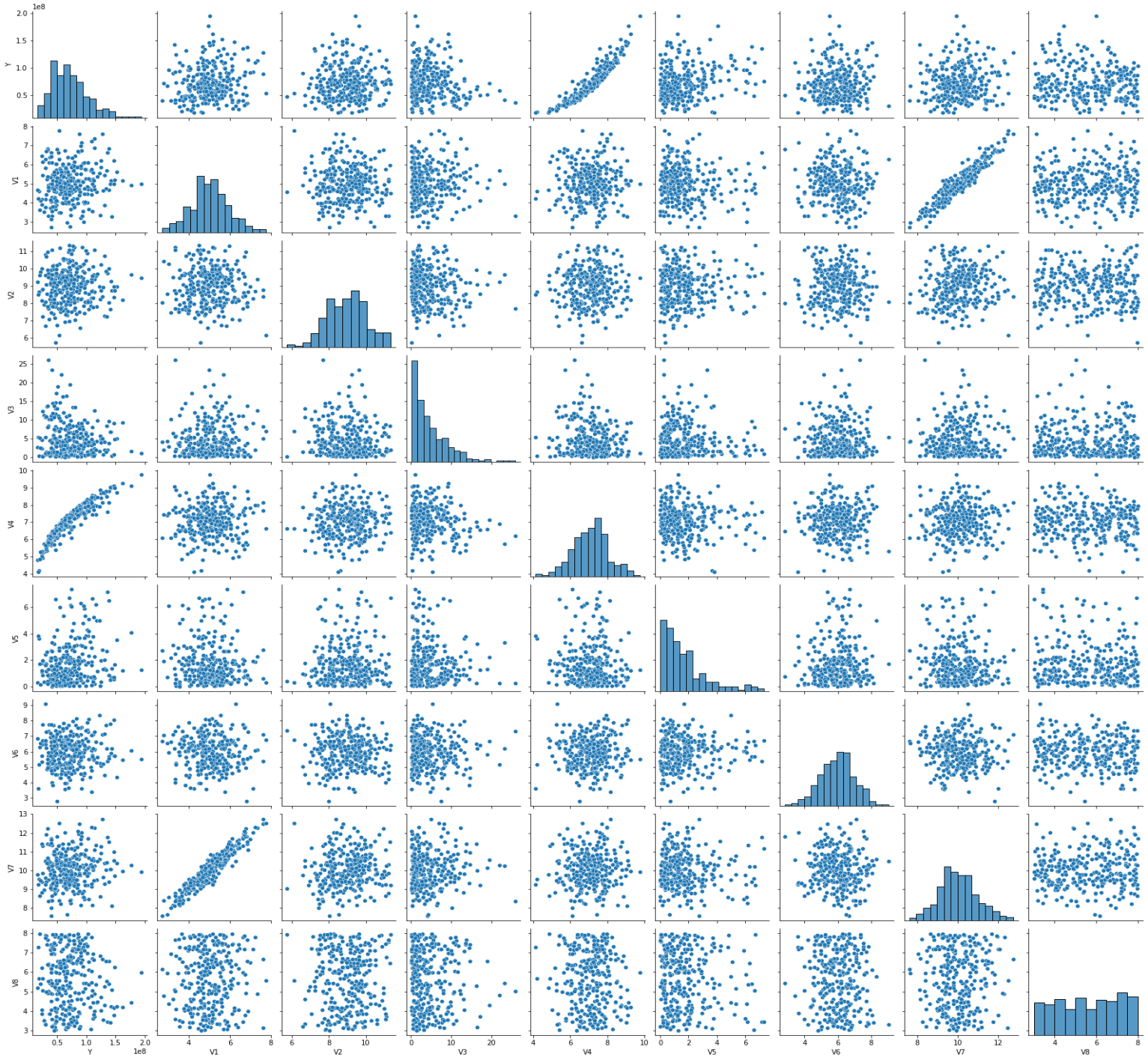
- **Data cleaning and preprocessing** : We will remove any missing or duplicate data and convert categorical variables to numerical values using one-hot encoding.
- **Exploratory data analysis**: We will analyze the dataset using graphs and descriptive statistics to identify any patterns or trends in the data.
- **Correlation analysis**: We will calculate the correlation coefficient between the dependent variable and each independent variable to identify the variables that are most strongly correlated with the dependent variable.
- **Regression analysis**: We will perform multiple linear regression analysis to develop a predictive model for the dependent variable using the independent variables.
- **Model evaluation**: We will evaluate the model's performance and accuracy using various metrics such as R-squared, mean squared error, and root mean squared error.

Correlation Heatmap:



Comment: From the above Correlation Heat Map one can easily interpret that most of the independent variables are INDEPENDENT with the Response Variable, while there is a strong Correlation between Response Variable(Y) and Regressor Variable (V4) and also there is high correlation between independent variables V1 and V7.

Pair plot with density curve:



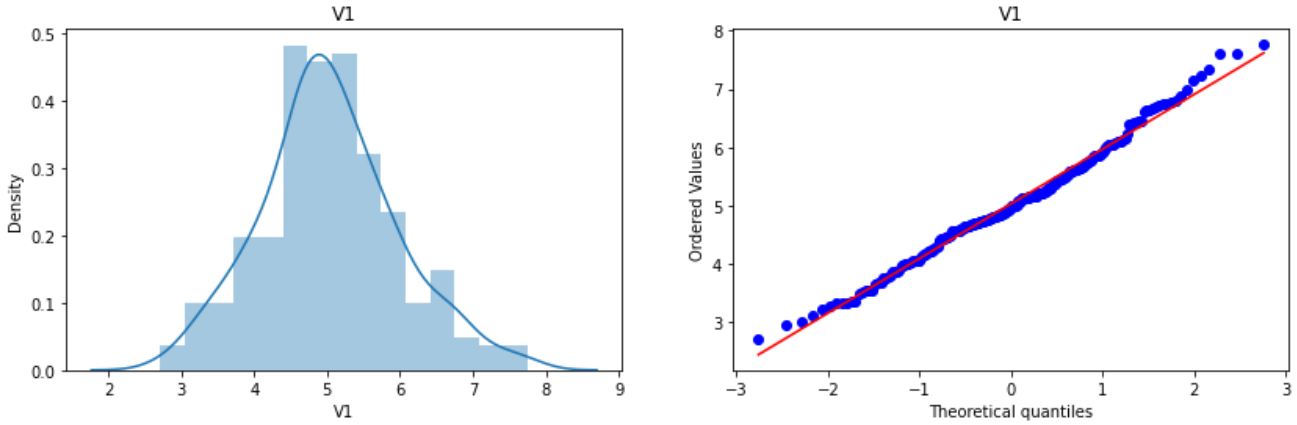
Comment : From the above Pair plot(scatter plot of the variable) , it can be easily interpreted that most the variables are approximately normally distributed except the variable V3 and V5 (which are slightly positively skewed) while the variable V8 is Approximately uniformly distributed

Plotting the distribution plots without any transformation

Univariate Regreession

1. Y vs V1

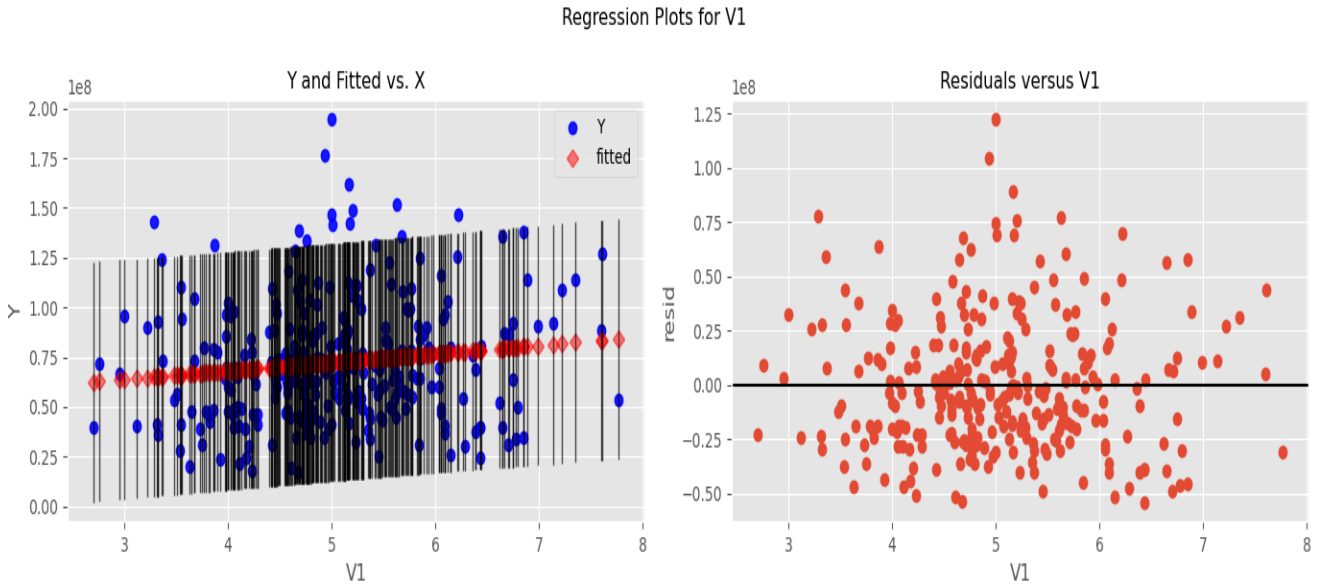
Distribution plot of V1



Fitting simple linear regression and model summary

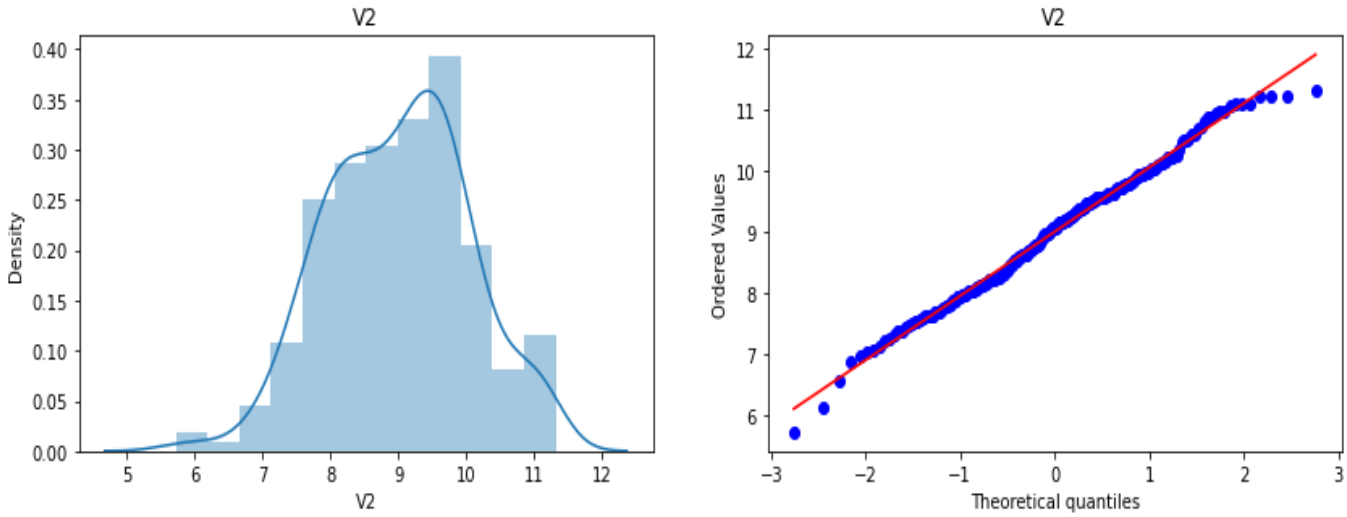
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.017
Model:                            OLS   Adj. R-squared:                  0.013
Method:                 Least Squares   F-statistic:                     5.056
Date:                Tue, 25 Apr 2023   Prob (F-statistic):             0.0253
Time:                        14:48:58   Log-Likelihood:                -5592.7
No. Observations:                 300   AIC:                         1.119e+04
Df Residuals:                     298   BIC:                         1.120e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   5.092e+07   9.72e+06      5.241      0.000    3.18e+07       7e+07
V1          4.278e+06    1.9e+06      2.248      0.025    5.34e+05    8.02e+06
==============================================================================
Omnibus:                       33.613   Durbin-Watson:                   2.200
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               42.405
Skew:                           0.819   Prob(JB):                     6.19e-10
Kurtosis:                       3.843   Cond. No.                         29.4
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
eval_env: 1
```

Regression plot for V1



2. Y vs V2

Distribution plot of V2



Fitting simple linear regression and model summary (Y vs V2)

OLS Regression Results

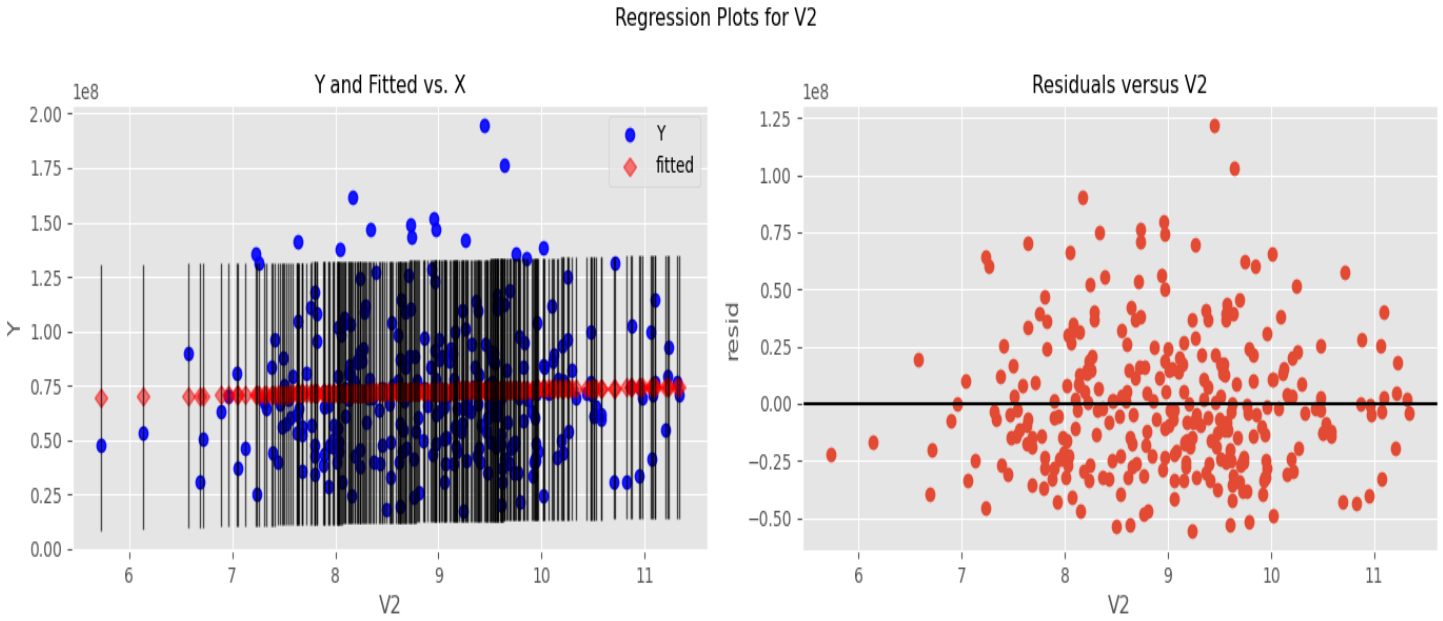
Dep. Variable:	Y	R-squared:	0.001
Model:	OLS	Adj. R-squared:	-0.002
Method:	Least Squares	F-statistic:	0.2690
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	0.604
Time:	14:49:22	Log-Likelihood:	-5595.1
No. Observations:	300	AIC:	1.119e+04
Df Residuals:	298	BIC:	1.120e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	6.451e+07	1.53e+07	4.204	0.000	3.43e+07	9.47e+07
V2	8.8e+05	1.7e+06	0.519	0.604	-2.46e+06	4.22e+06

Omnibus:	33.802	Durbin-Watson:	2.205
Prob(Omnibus):	0.000	Jarque-Bera (JB):	42.362
Skew:	0.832	Prob(JB):	6.33e-10
Kurtosis:	3.788	Cond. No.	79.6

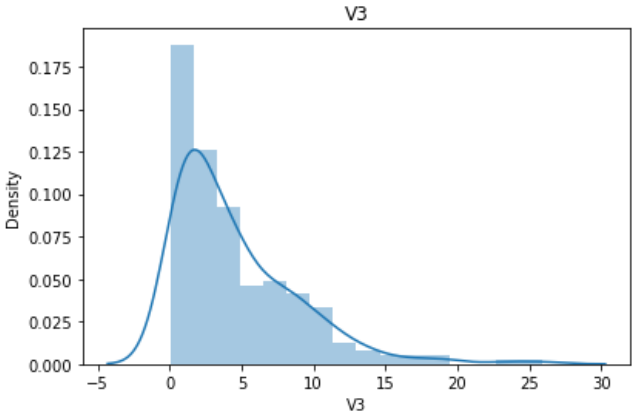
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
eval_env: 1

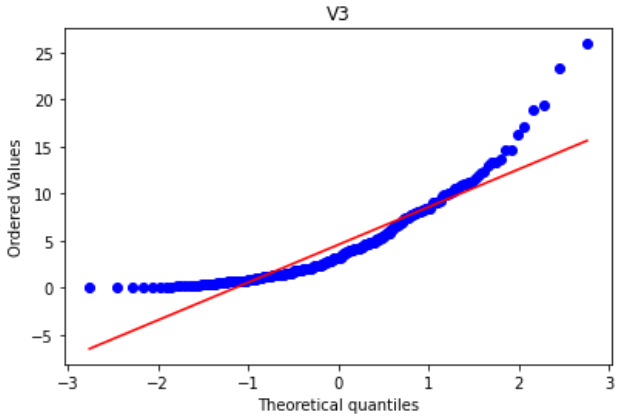
Regression plot for V2



3. Y vs V3

Distribution plot of V3





Fitting simple linear regression and model summary (Y vs V3)

OLS Regression Results

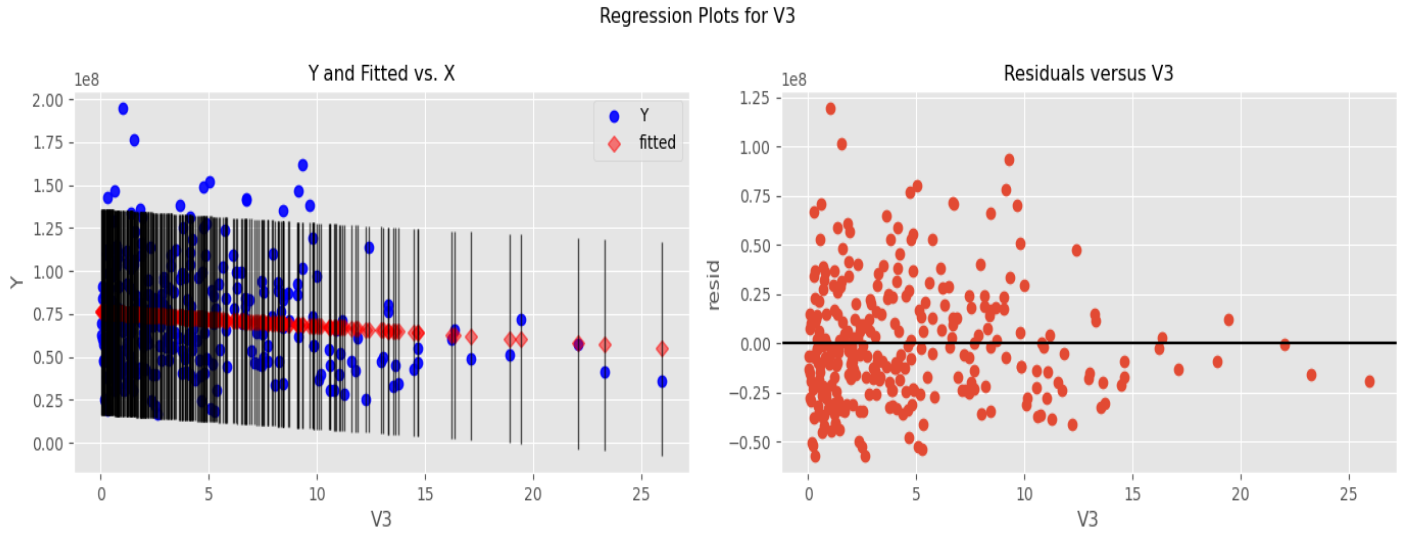
Dep. Variable:	Y	R-squared:	0.014
Model:	OLS	Adj. R-squared:	0.011
Method:	Least Squares	F-statistic:	4.328
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	0.0384
Time:	14:49:27	Log-Likelihood:	-5593.1
No. Observations:	300	AIC:	1.119e+04
Df Residuals:	298	BIC:	1.120e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.62e+07	2.53e+06	30.161	0.000	7.12e+07	8.12e+07
V3	-8.248e+05	3.96e+05	-2.080	0.038	-1.6e+06	-4.45e+04

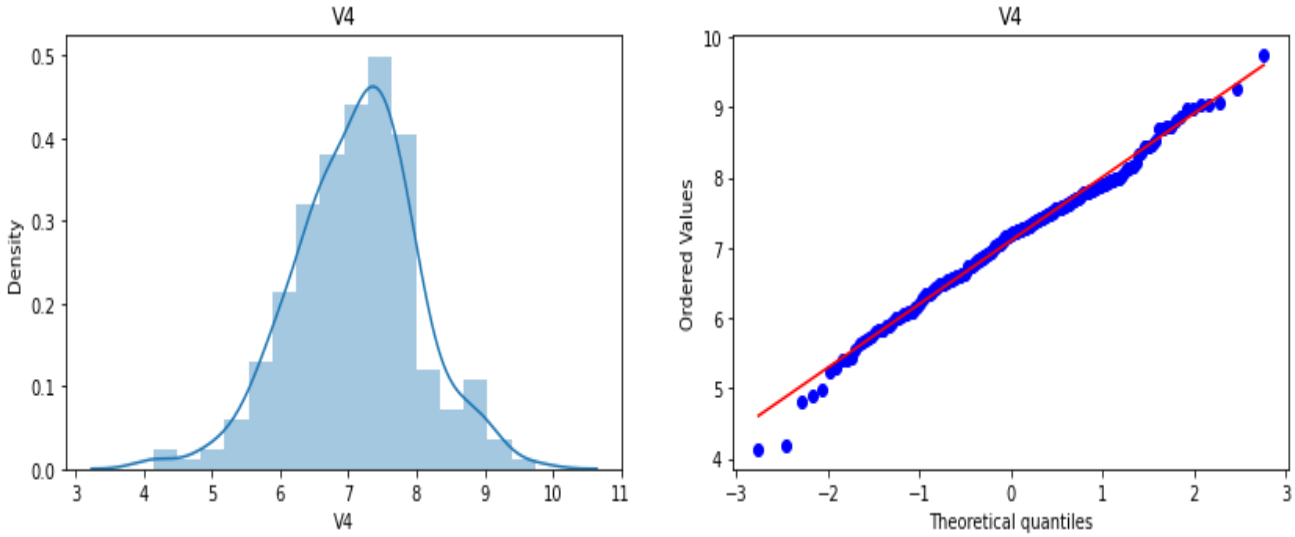
Omnibus:	33.784	Durbin-Watson:	2.169
Prob(Omnibus):	0.000	Jarque-Bera (JB):	42.571
Skew:	0.824	Prob(JB):	5.70e-10
Kurtosis:	3.829	Cond. No.	9.31

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
eval_env: 1

Regression plot for V3



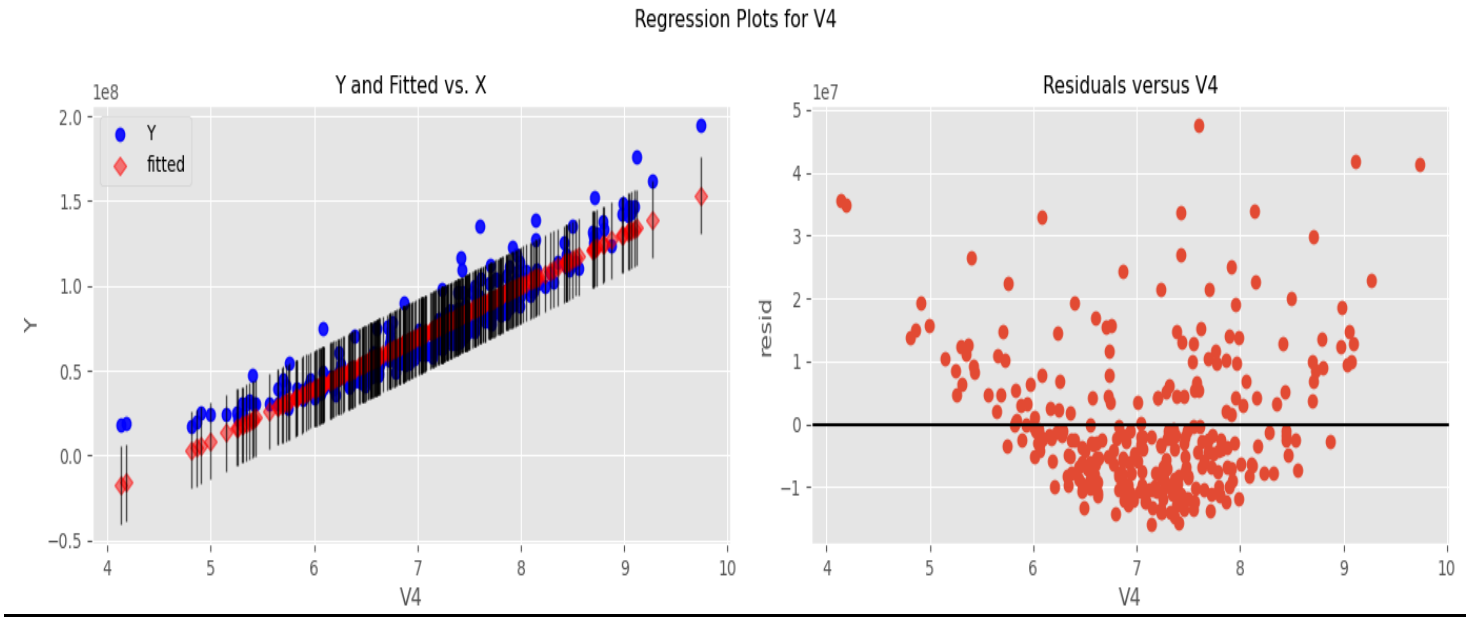
4. Y vs V4



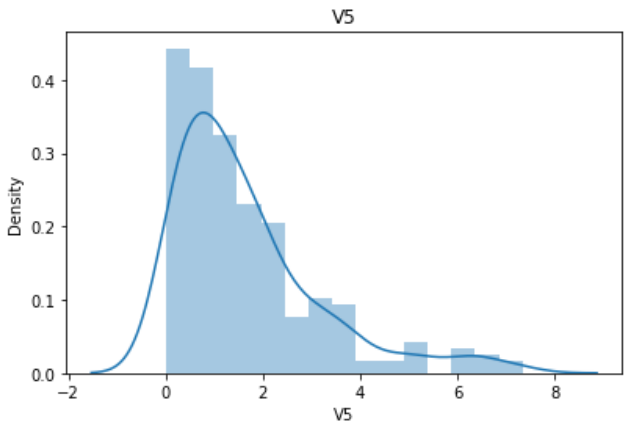
Fitting simple linear regression and model summary (Y vs V4)

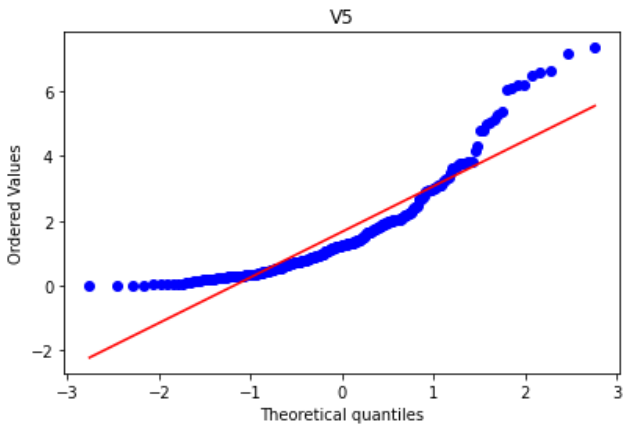
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.862
Model:                            OLS   Adj. R-squared:                  0.862
Method:                 Least Squares   F-statistic:                     1865.
Date:                Tue, 25 Apr 2023   Prob (F-statistic):          2.64e-130
Time:                        14:49:32   Log-Likelihood:                -5297.9
No. Observations:                 300   AIC:                         1.060e+04
Df Residuals:                     298   BIC:                         1.061e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept  -1.433e+08   5.04e+06    -28.447      0.000   -1.53e+08   -1.33e+08
V4          3.045e+07   7.05e+05     43.189      0.000    2.91e+07    3.18e+07
==============================================================================
Omnibus:                       78.424   Durbin-Watson:                   1.915
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              150.018
Skew:                           1.383   Prob(JB):                     2.65e-33
Kurtosis:                       5.086   Cond. No.                         56.0
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
eval_env: 1
```

Regression plot for V4



5. Y vs V5

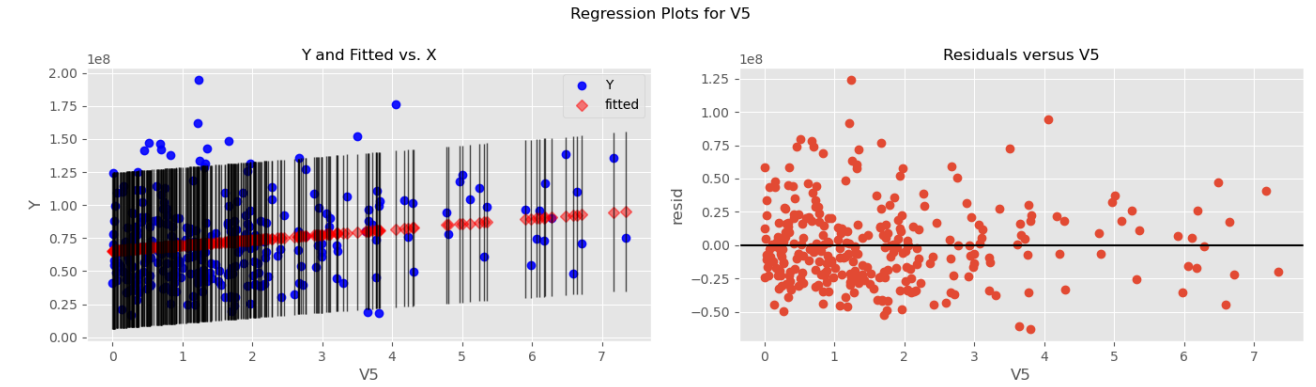




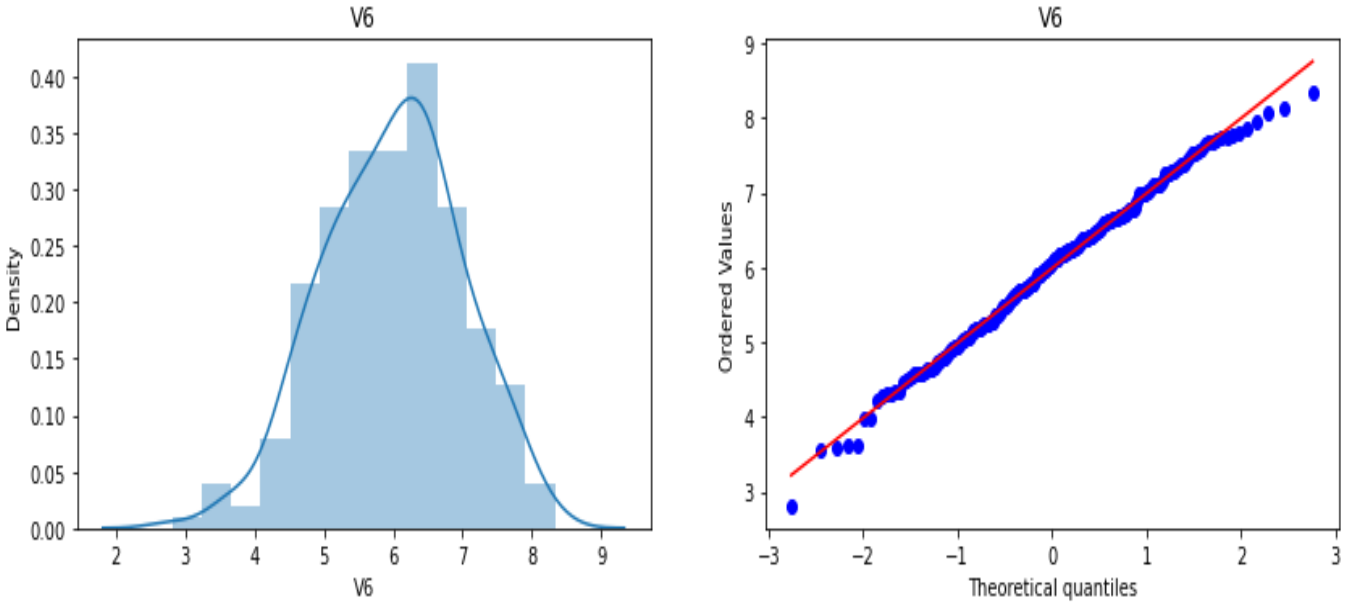
Fitting simple linear regression and model summary (Y vs V5)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.042
Model:                            OLS   Adj. R-squared:                  0.039
Method:                 Least Squares   F-statistic:                     13.13
Date:                Tue, 25 Apr 2023   Prob (F-statistic):           0.000341
Time:                        14:49:41   Log-Likelihood:                -5588.8
No. Observations:                 300   AIC:                         1.118e+04
Df Residuals:                     298   BIC:                         1.119e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   6.553e+07   2.57e+06     25.532      0.000    6.05e+07    7.06e+07
V5          4.045e+06   1.12e+06      3.624      0.000    1.85e+06    6.24e+06
==============================================================================
Omnibus:                       37.030   Durbin-Watson:                   2.196
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               49.136
Skew:                           0.849   Prob(JB):                     2.14e-11
Kurtosis:                       4.025   Cond. No.                         3.80
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
eval_env: 1
```

Regression plot for V5



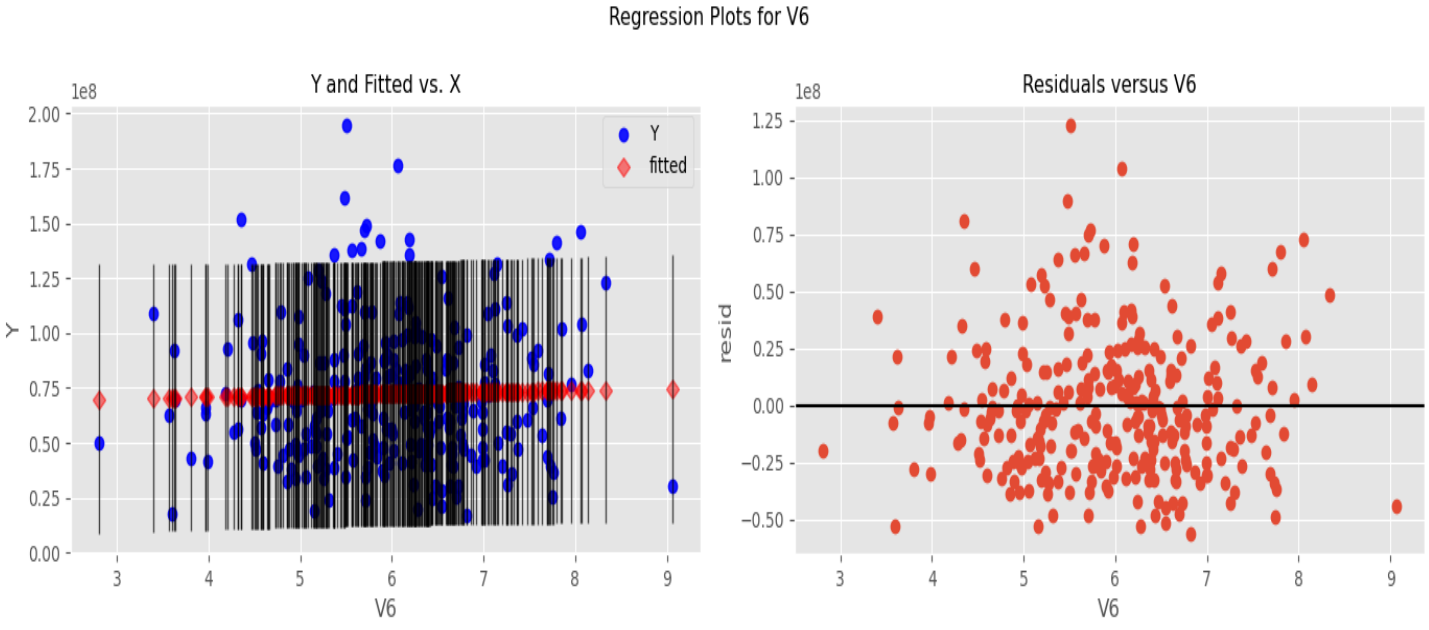
6.) Y vs V6



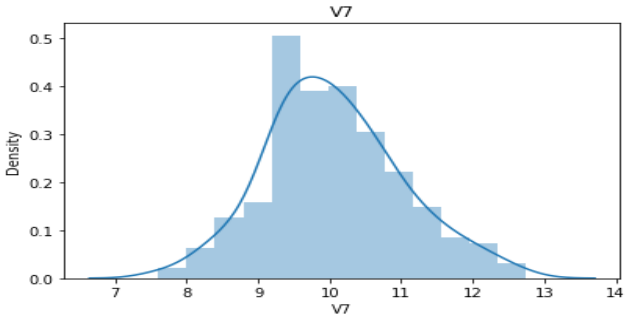
Fitting simple linear regression and model summary (Y vs V6)

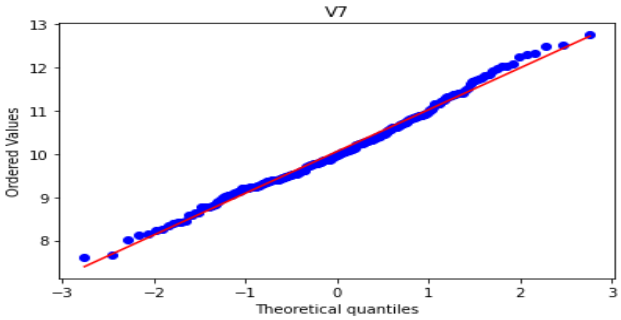
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                 -0.003
Method:                 Least Squares   F-statistic:                     0.1796
Date:                Tue, 25 Apr 2023   Prob (F-statistic):              0.672
Time:                        14:49:47   Log-Likelihood:                -5595.2
No. Observations:                 300   AIC:                         1.119e+04
Df Residuals:                     298   BIC:                         1.120e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   6.797e+07   1.06e+07      6.392      0.000     4.7e+07    8.89e+07
V6          7.441e+05   1.76e+06      0.424      0.672   -2.71e+06     4.2e+06
==============================================================================
Omnibus:                       33.590   Durbin-Watson:                   2.200
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               42.099
Skew:                           0.826   Prob(JB):                     7.22e-10
Kurtosis:                       3.799   Cond. No.                         37.5
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
eval_env: 1
```

Regression plot for V6



7.) Y vs V7





Fitting simple linear regression and model summary (Y vs V7)

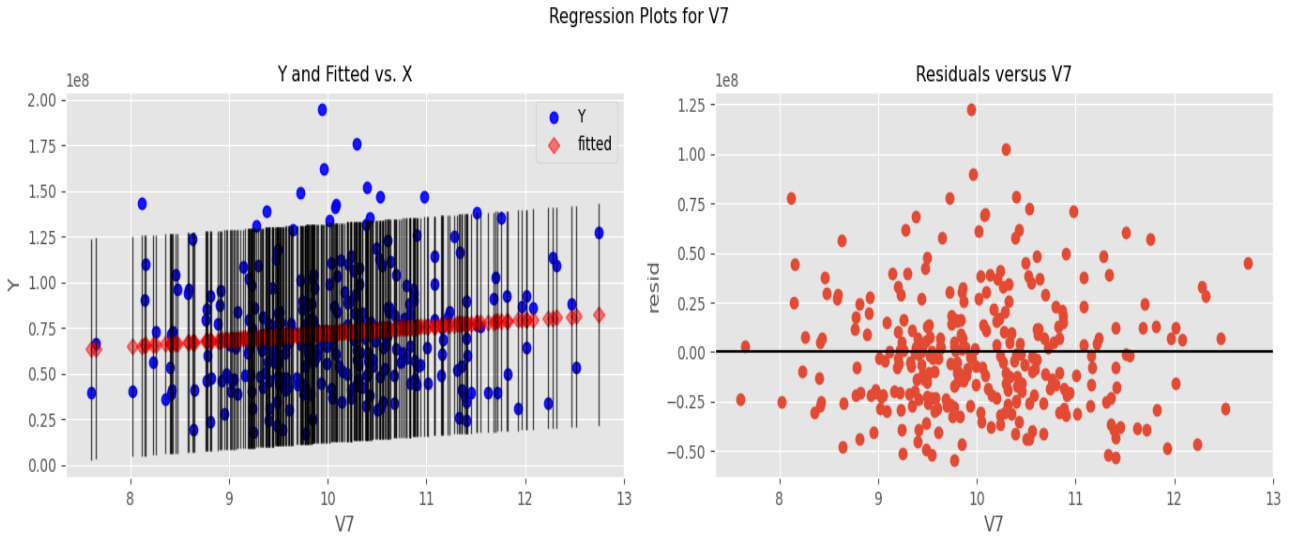
OLS Regression Results

Dep. Variable:	Y	R-squared:	0.013
Model:	OLS	Adj. R-squared:	0.009
Method:	Least Squares	F-statistic:	3.792
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	0.0524
Time:	14:49:51	Log-Likelihood:	-5593.4
No. Observations:	300	AIC:	1.119e+04
Df Residuals:	298	BIC:	1.120e+04
Df Model:	1		
Covariance Type:	nonrobust		

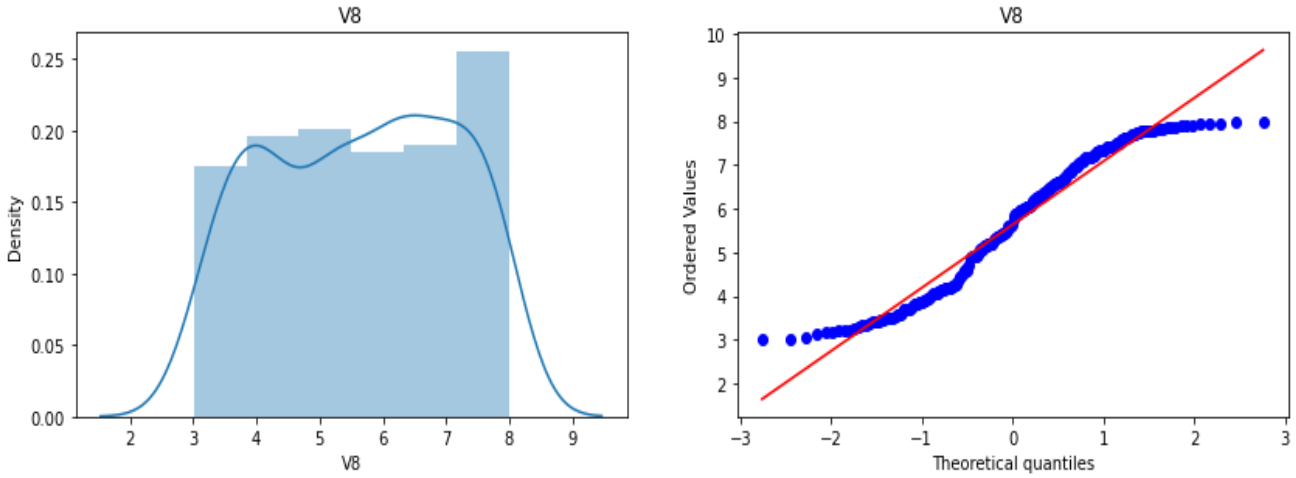
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.576e+07	1.89e+07	1.891	0.060	-1.45e+06	7.3e+07
V7	3.653e+06	1.88e+06	1.947	0.052	-3.86e+04	7.34e+06

Omnibus:	33.491	Durbin-Watson:	2.200
Prob(Omnibus):	0.000	Jarque-Bera (JB):	42.129
Skew:	0.819	Prob(JB):	7.11e-10
Kurtosis:	3.828	Cond. No.	110.

Regression plot for V7



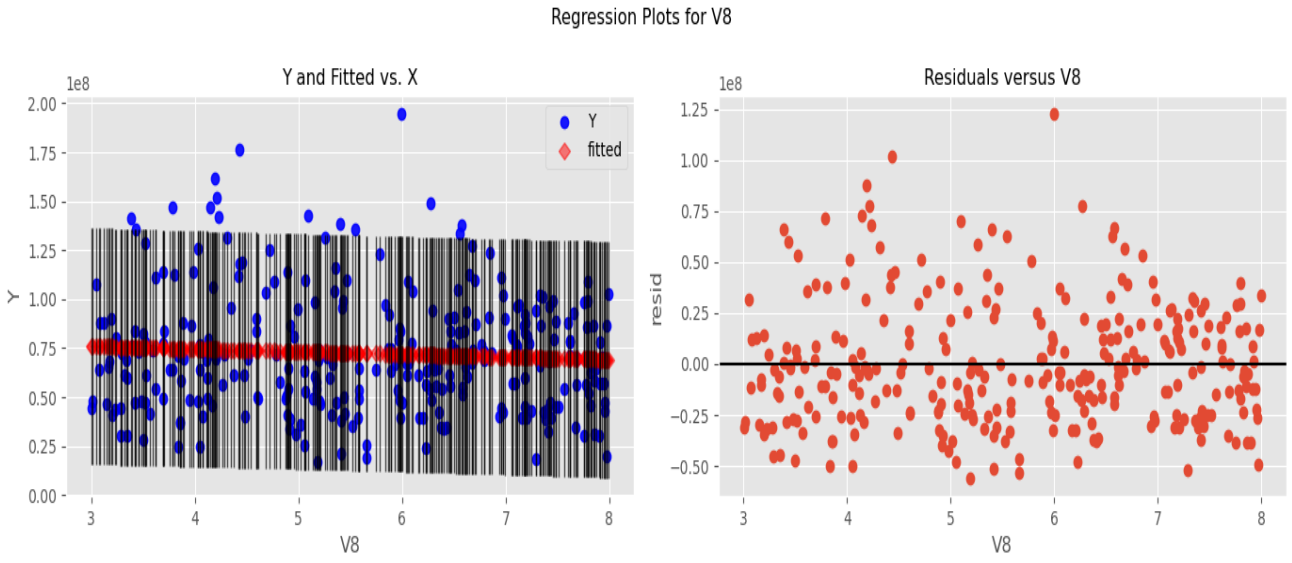
8.) Y vs V8



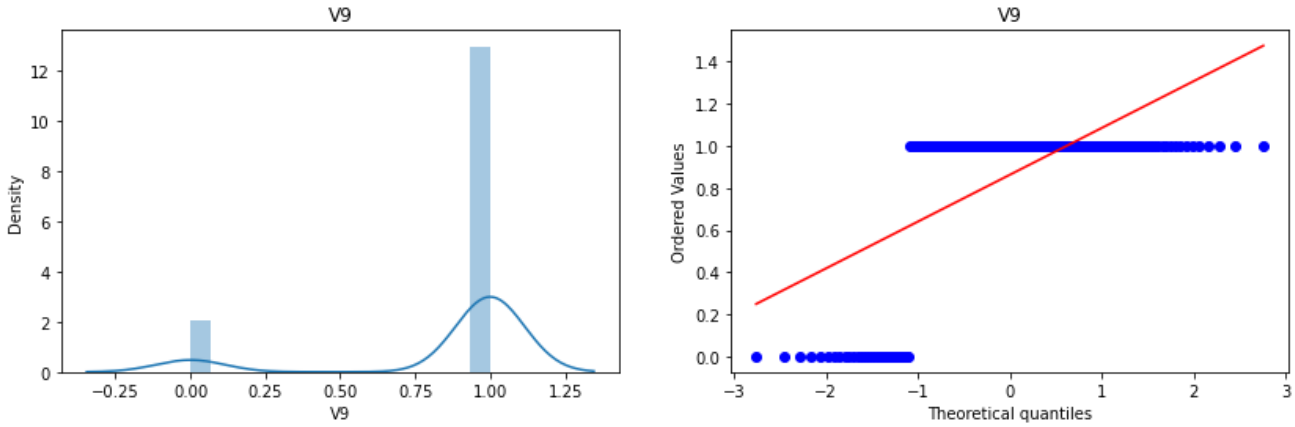
Fitting simple linear regression and model summary (Y vs V8)

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.004
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     1.308
Date:                Tue, 25 Apr 2023   Prob (F-statistic):              0.254
Time:                        14:49:54   Log-Likelihood:                -5594.6
No. Observations:                 300   AIC:                         1.119e+04
Df Residuals:                     298   BIC:                         1.120e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   8.005e+07    6.9e+06     11.593      0.000    6.65e+07    9.36e+07
V8         -1.36e+06   1.19e+06     -1.144      0.254    -3.7e+06     9.8e+05
==============================================================================
Omnibus:                       31.166   Durbin-Watson:                   2.202
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               38.042
Skew:                           0.798   Prob(JB):                     5.49e-09
Kurtosis:                       3.705   Cond. No.                         23.4
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
eval_env: 1
```

Regression plot for V8



9.) Y vs V9



Fitting simple linear regression and model summary (Y vs V9)

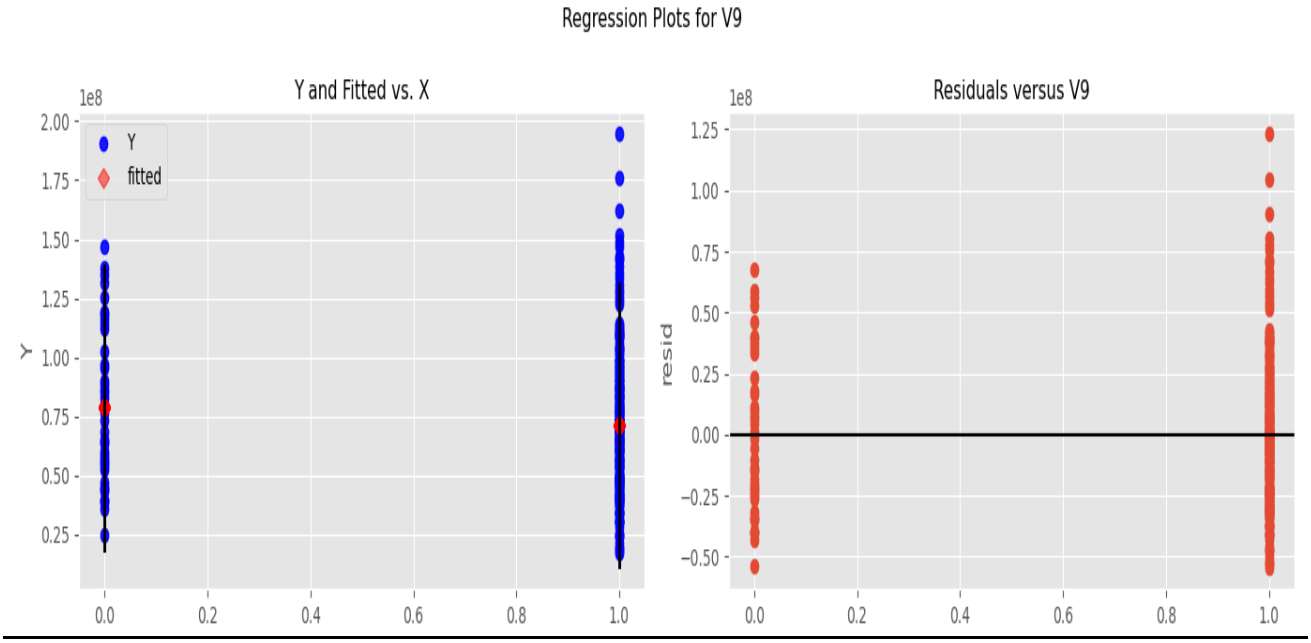
OLS Regression Results

Dep. Variable:	Y	R-squared:	0.007
Model:	OLS	Adj. R-squared:	0.004
Method:	Least Squares	F-statistic:	2.123
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	0.146
Time:	14:49:59	Log-Likelihood:	-5594.2
No. Observations:	300	AIC:	1.119e+04
Df Residuals:	298	BIC:	1.120e+04
Df Model:	1		
Covariance Type:	nonrobust		

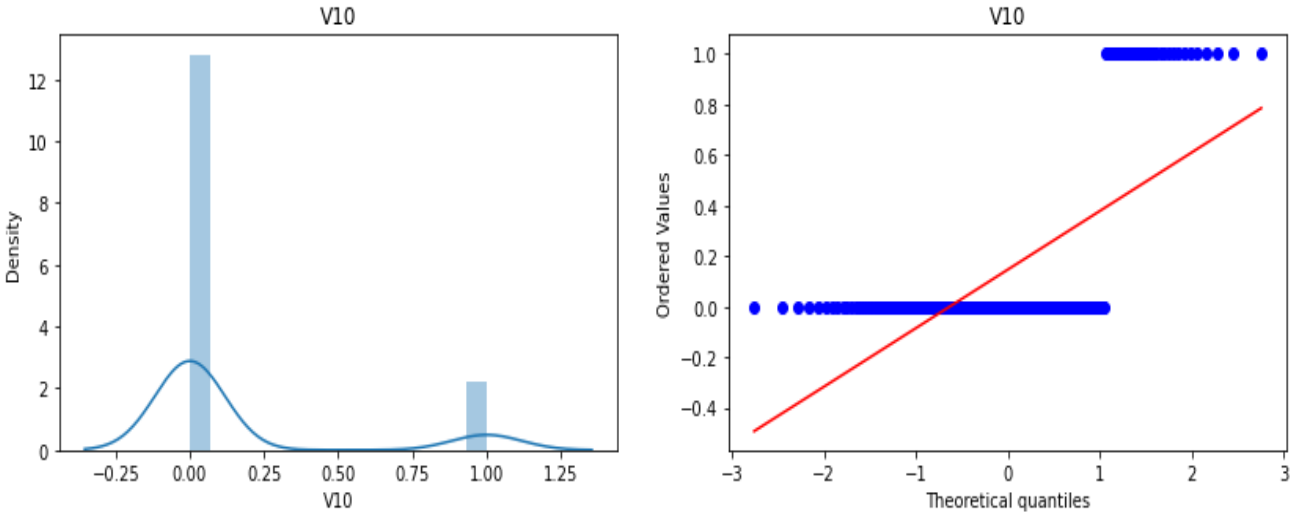
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.885e+07	4.76e+06	16.578	0.000	6.95e+07	8.82e+07
V9	-7.459e+06	5.12e+06	-1.457	0.146	-1.75e+07	2.61e+06

Omnibus:	33.668	Durbin-Watson:	2.203
Prob(Omnibus):	0.000	Jarque-Bera (JB):	42.427
Skew:	0.821	Prob(JB):	6.13e-10
Kurtosis:	3.834	Cond. No.	5.23

Regression plot for V9



Y vs V10



Fitting simple linear regression and model summary (Y vs V10)

OLS Regression Results

Dep. Variable:	Y	R-squared:	0.000
Model:	OLS	Adj. R-squared:	-0.003
Method:	Least Squares	F-statistic:	0.0003251
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	0.986
Time:	14:50:02	Log-Likelihood:	-5595.3
No. Observations:	300	AIC:	1.119e+04
Df Residuals:	298	BIC:	1.120e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	7.24e+07	1.9e+06	38.048	0.000	6.87e+07	7.61e+07
V10	9.17e+04	5.09e+06	0.018	0.986	-9.92e+06	1.01e+07

Omnibus:	33.510	Durbin-Watson:	2.201
Prob(Omnibus):	0.000	Jarque-Bera (JB):	41.884
Skew:	0.828	Prob(JB):	8.03e-10
Kurtosis:	3.781	Cond. No.	2.95

Regression plot for V10

Multivariate Analysis

APPLYING MULTIPLE LINEAR REGRESSION BLINDELY ON THE WHOLE DATA

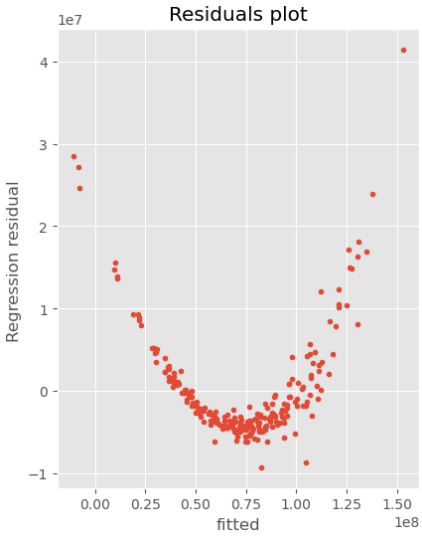
OLS Regression Results

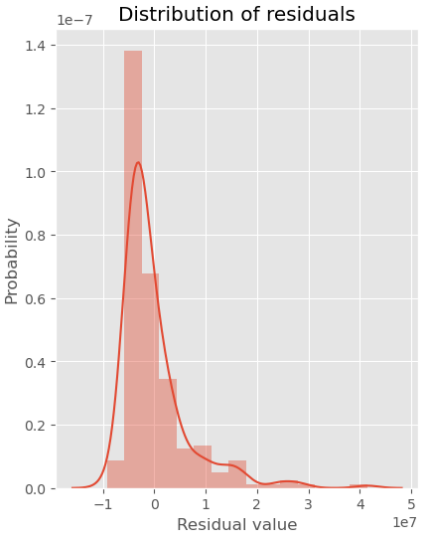
Dep. Variable:	Y	R-squared:	0.949
Model:	OLS	Adj. R-squared:	0.947
Method:	Least Squares	F-statistic:	430.1
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	2.41e-142
Time:	20:07:29	Log-Likelihood:	-4112.5
No. Observations:	240	AIC:	8247.
Df Residuals:	229	BIC:	8285.
Df Model:	10		
Covariance Type:	nonrobust		

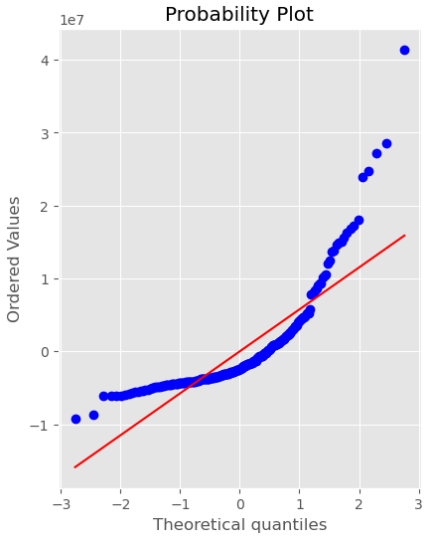
	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.736e+08	9.78e+06	-17.751	0.000	-1.93e+08	-1.54e+08
V1	2.871e+06	1.73e+06	1.660	0.098	-5.36e+05	6.28e+06
V2	-1.789e+05	4.94e+05	-0.362	0.717	-1.15e+06	7.94e+05
V3	668.3932	1.04e+05	0.006	0.995	-2.04e+05	2.06e+05
V4	3.096e+07	5.01e+05	61.836	0.000	3e+07	3.2e+07
V5	5.693e+06	2.95e+05	19.289	0.000	5.11e+06	6.27e+06
V6	-2.251e+04	4.53e+05	-0.050	0.960	-9.15e+05	8.7e+05
V7	5.072e+05	1.7e+06	0.299	0.765	-2.84e+06	3.85e+06
V8	-3.255e+05	3.05e+05	-1.069	0.286	-9.26e+05	2.75e+05
V9	5.683e+05	1.3e+06	0.436	0.663	-2e+06	3.14e+06
V10	1.79e+06	1.29e+06	1.389	0.166	-7.49e+05	4.33e+06

Omnibus:	153.399	Durbin-Watson:	1.963
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1035.251
Skew:	2.583	Prob(JB):	1.58e-225
Kurtosis:	11.766	Cond. No.	421.

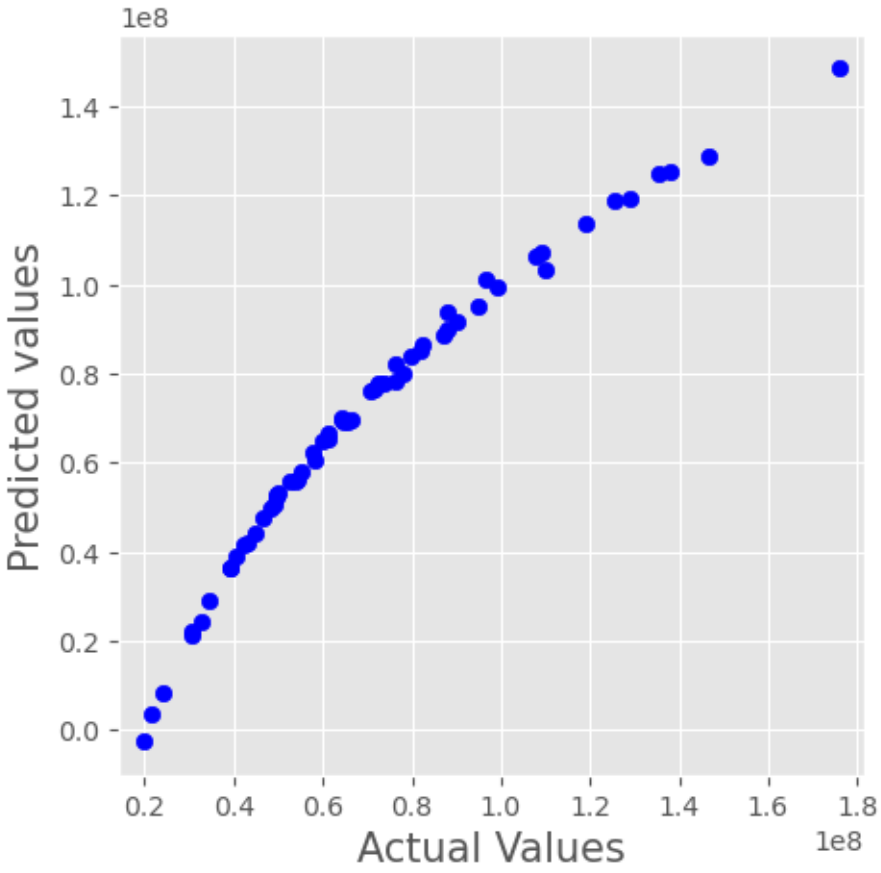
Model validation







Making predictions

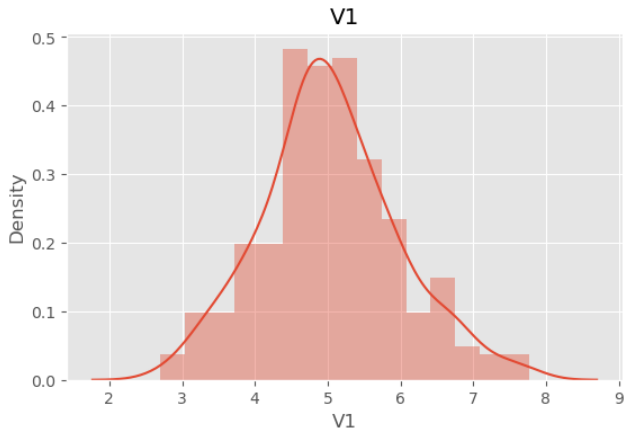


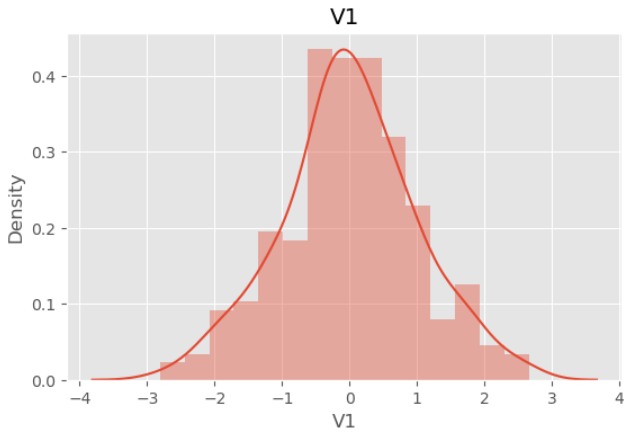
Transformation

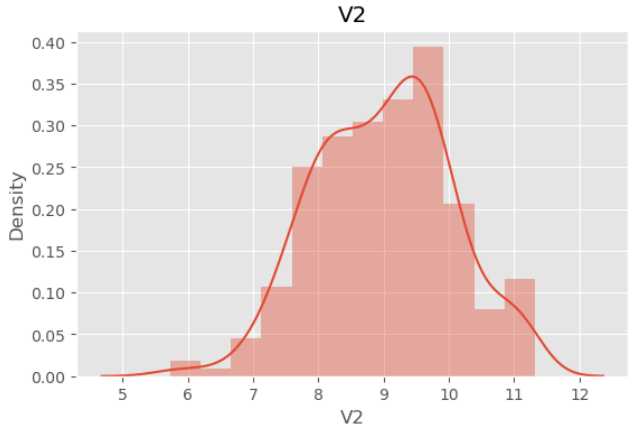
Applying Box-Cox Transform

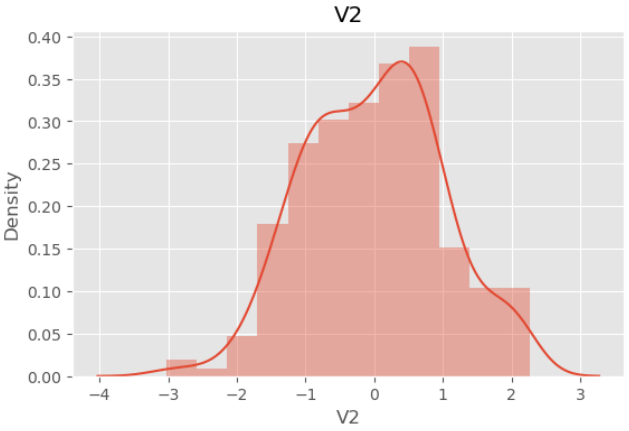
	cols	box_cox_lambdas
0	V1	0.517748
1	V2	1.218425
2	V3	0.274555
3	V4	1.490793
4	V5	0.307718
5	V6	1.359079
6	V7	-0.013576
7	V8	0.935454
8	V9	0.523669
9	V10	-0.492590

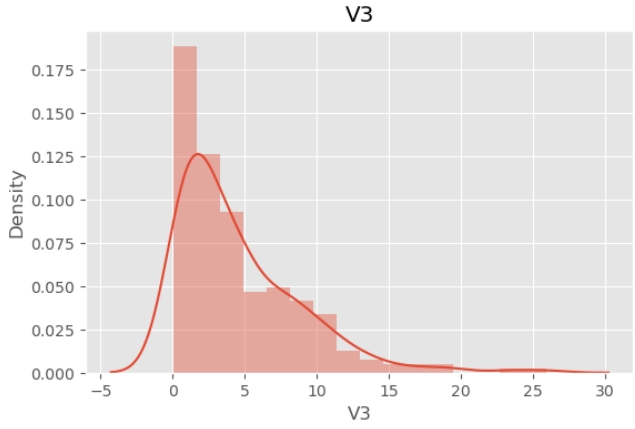
Before and after comparision for Box-Cox Plot

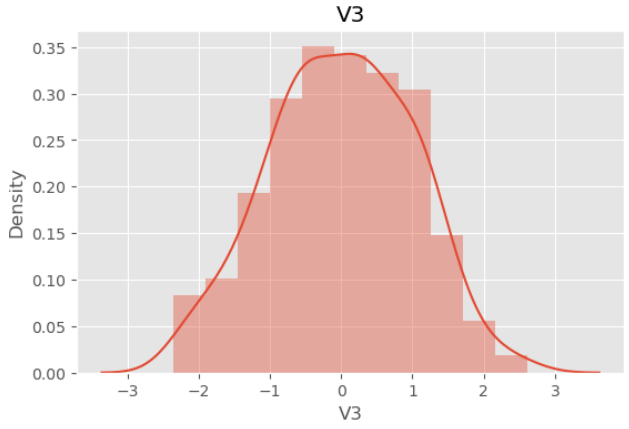


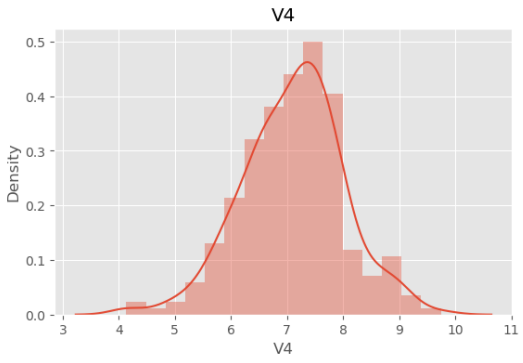


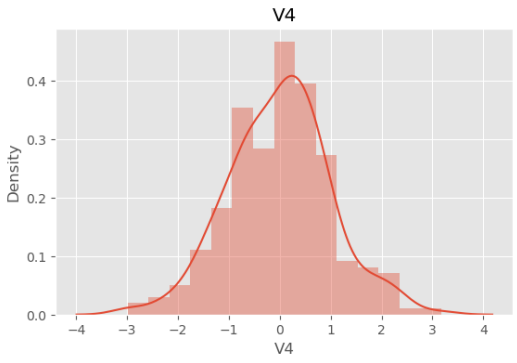


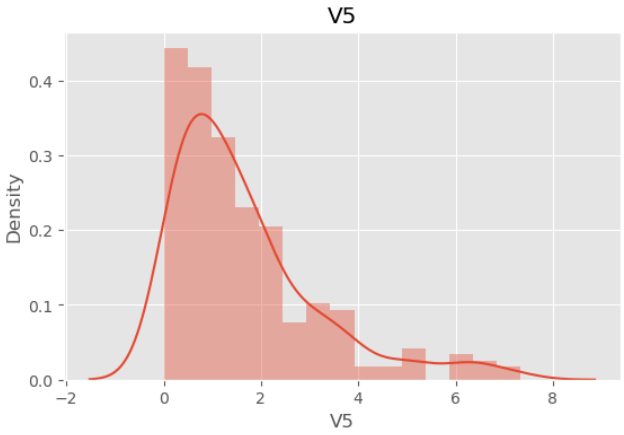


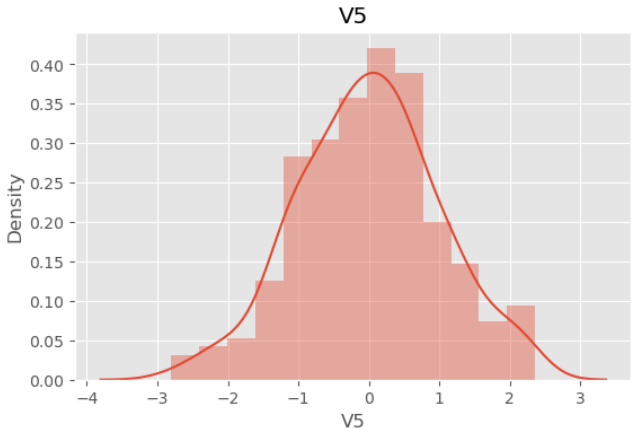


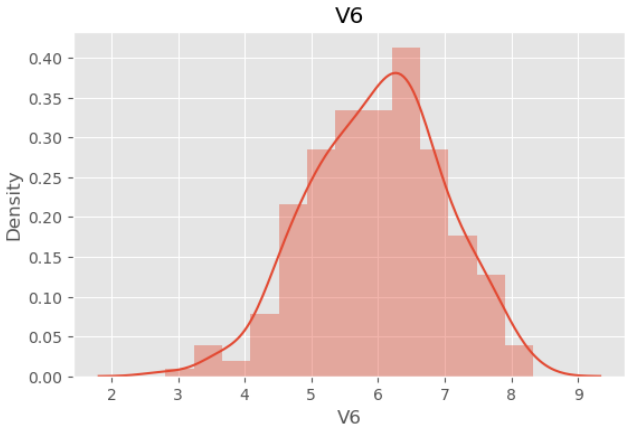


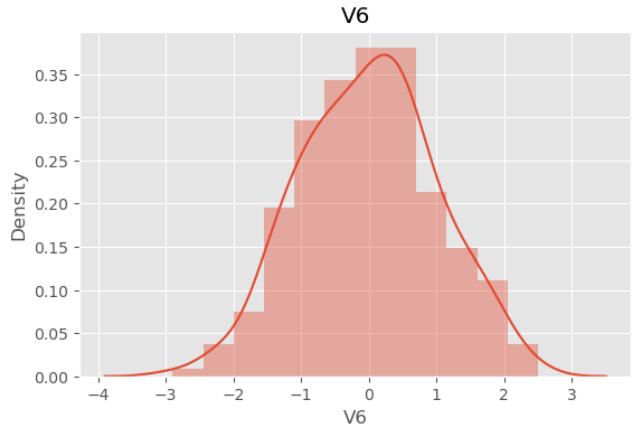


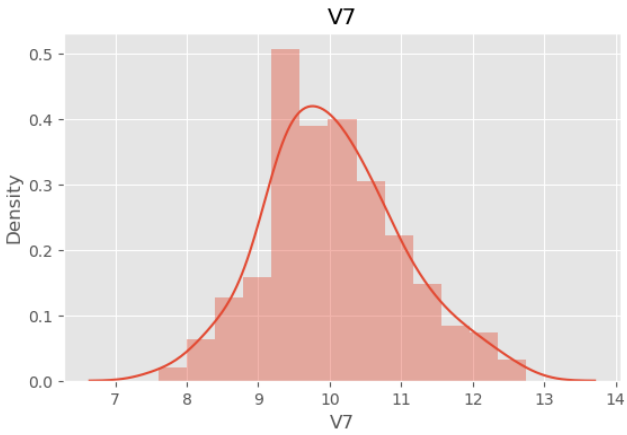


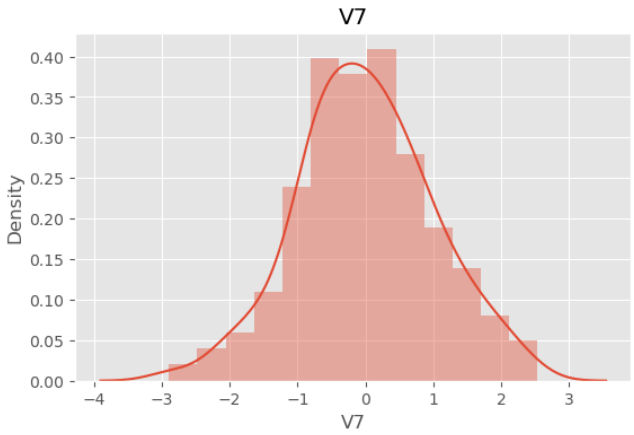

V5
Density
V5

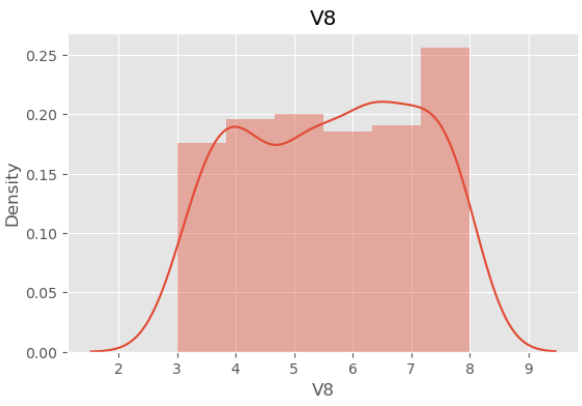

V5
Density
V5

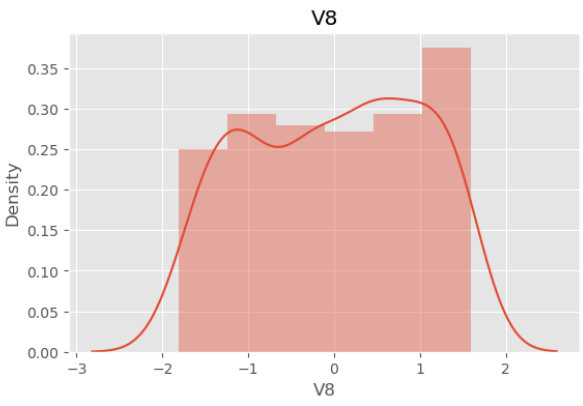

V6
Density
V6

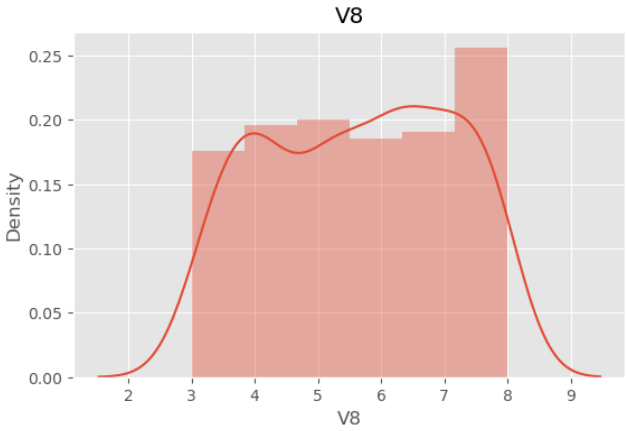

V6
Density
V6

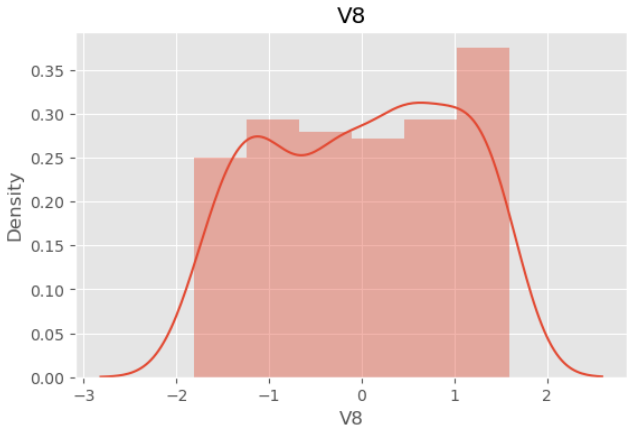

V7
Density
V7

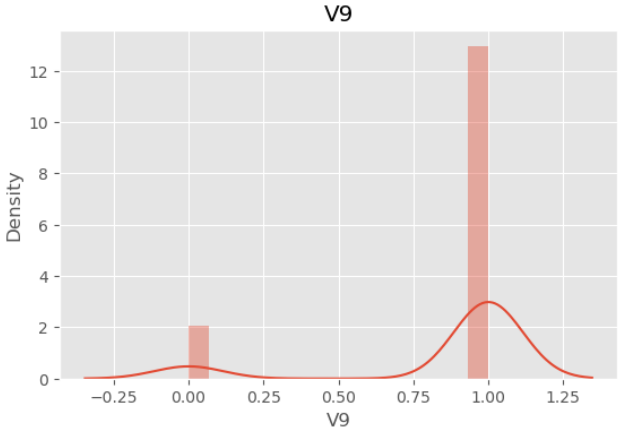

V7
Density
V7

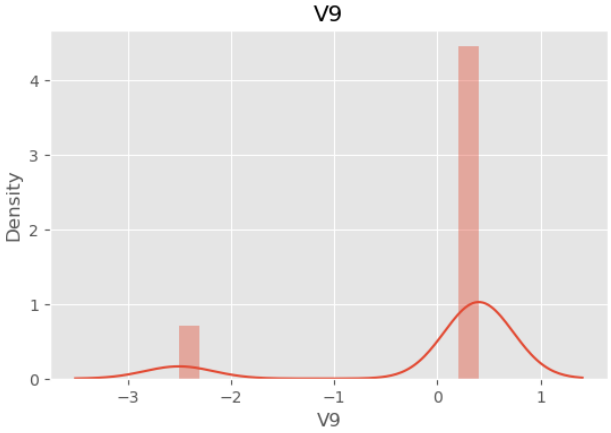

V8
Density
V8

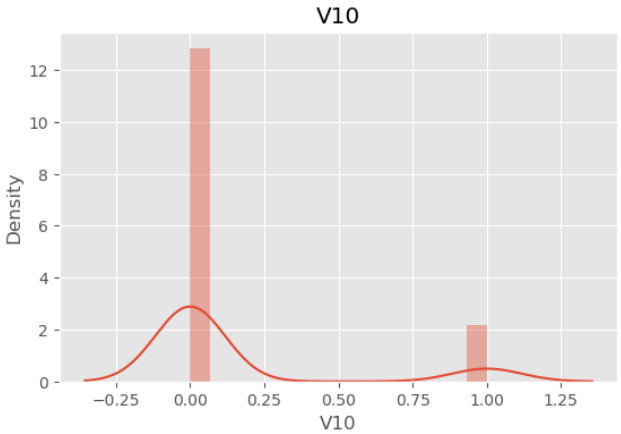

V8
Density
V8

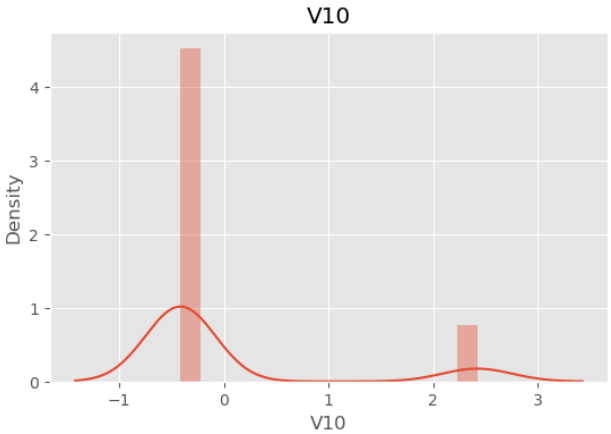

V8
Density
V8


V8
Density
V8


V9
Density
V9


V9
Density
V9


V10
Density
V10


V10
Density
V10

Selecting best model

Best subset selection

To perform best selection, we fit separate models for each possible combination of the n predictors and then select the best subset. That is we fit:

- All models that contains exactly one predictor
- All models that contain 2 predictors at the second step: (nC2)
- Until reaching the end point where all n predictors are included in the model

This results in 2n possibilities as this is a power set problem. In our case there are 2^10=1024 possible combinations

Algorithm

- Let M_0 denote the null model which contains no predictors, this model simply predicts the sample mean of each observation
    - For k=1 ,2,...,n
        - Fit all (nCk) models that contain exactly k predictors
        - Pick the best among these (nk) models, and call it M_k. Here the best is defined as having the smallest RSS, or an equivalent measure
    - Select the single best model among M_0,M_1,...,M_n using cross validated predicton error, Cp, BIC, adjusted R2 or any other method.

Implementing Best subset selection

Finding the best subsets for each number of features

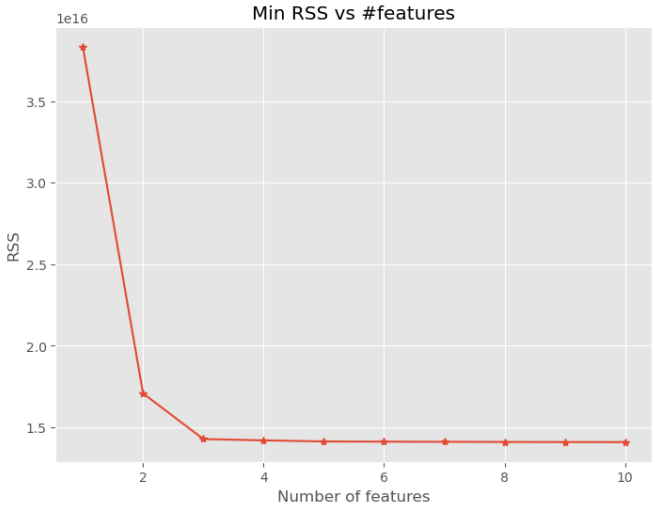
1.) **Using the smallest RSS Value**

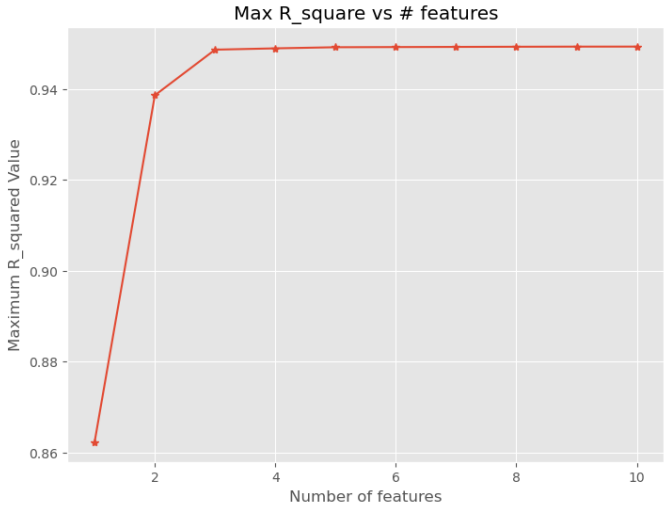
	numb_features	RSS	R_squared	features
3	1	3.834547e+16	0.862249	(V4,)
34	2	1.707693e+16	0.938654	(V4, V5)
70	3	1.427903e+16	0.948705	(V1, V4, V5)

2.) **Using the largest R^2 squared Value**

	numb_features	RSS	R_squared	features
3	1	3.834547e+16	0.862249	(V4,)
34	2	1.707693e+16	0.938654	(V4, V5)
70	3	1.427903e+16	0.948705	(V1, V4, V5)

Plotting the minimum RSS and Maximum R_square Vs Number of features

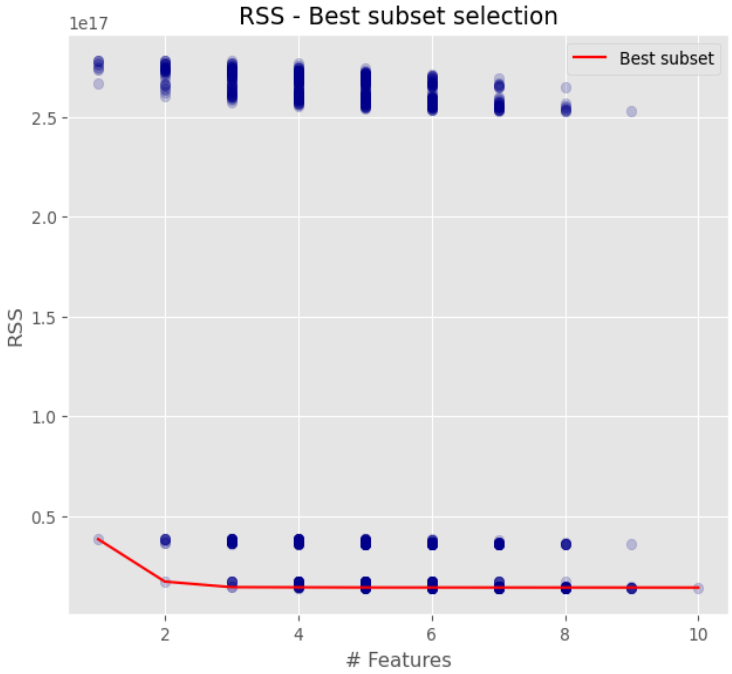


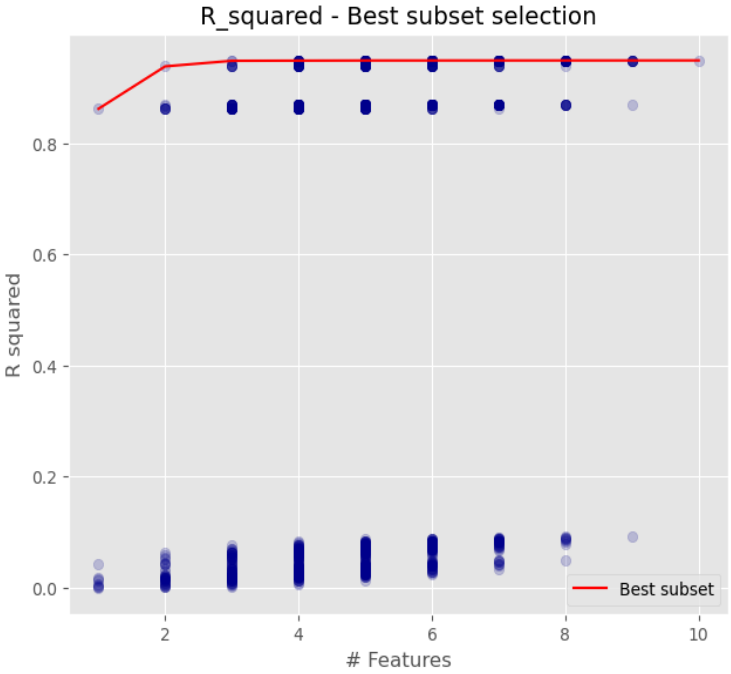


Adding columns to the data frame with RSS and R squared values of the best subset

	numb_features	RSS	R_squared	features	min_RSS	max_R_squared
0	1	2.737246e+17	0.016682	(V1,)	3.834547e+16	0.862249
1	1	2.781174e+17	0.000902	(V2,)	3.834547e+16	0.862249
2	1	2.743839e+17	0.014314	(V3,)	3.834547e+16	0.862249
3	1	3.834547e+16	0.862249	(V4,)	3.834547e+16	0.862249
4	1	2.666196e+17	0.042206	(V5,)	3.834547e+16	0.862249

Plotting the best subset selection process





Forward stepwise selection

For computational reasons, the best subset cannot be applied for any large n due to the $2\text{^}n$ complexity. Forward Stepwise begins with a model containing no predictors, and then adds predictors to the model, one at the time. At each step, the variable that gives the greatest additional improvement to the fit is added to the model.

Algorithm

Let M_0 denote the null model which contains no predictors

- For k=1,2,...,$n-1$
  - Consider all $n-k$ models that augment the predictors in M_k with one additional predictor
  - Choose the \text it{best} among these $n-k$ models, and call it M_k+1
  - Select the single best model among $M_0,M_1,...,M_n$ using cross validated prediction error, C_p, BIC ,$adjustedR\text{^}2$ or any other method.

- **Forward stepwise subset selection**
- Number of features | Features | RSS
- (1, ['V4'], 38345471133521288),
- (2, ['V4', 'V5'], 17076929672279732),
- (3, ['V4', 'V5', 'V1'], 14279029676255524),
- (4, ['V4', 'V5', 'V1', 'V10'], 14198449815008740),
- (5, ['V4', 'V5', 'V1', 'V10', 'V8'], 14129682186328508),
- (6, ['V4', 'V5', 'V1', 'V10', 'V8', 'V9'], 14119167377711116),
- (7, ['V4', 'V5', 'V1', 'V10', 'V8', 'V9', 'V6'], 14110386286093154),
- (8, ['V4', 'V5', 'V1', 'V10', 'V8', 'V9', 'V6', 'V2'], 14102288584789658),
- (9, ['V4', 'V5', 'V1', 'V10', 'V8', 'V9', 'V6', 'V2', 'V7'],140929044563301
- (10,['V4', 'V5', 'V1', 'V10', 'V8', 'V9', 'V6', 'V2', 'V7', 'V3'],140909182

Comparing models: AIC, BIC, Mallows'CP

The training set Mean Squared Error (MSE) is generally an underestimate of the test MSE. This is because when we fit a model to the training data using least squares, we specifically estimate the regression coefficients such that the training RSS is minimized. In particular, the training RSS decreases as we add more features to the model, but the test error may not. Therefore the training RSS and R^2 may not be used for selecting the best model unless we adjust for this underestimation.

Mallow's Cp

Mallow's C_p is named after Colin Lingwood Mallows and is defined as:
$C_p=(1/m)*(RSS+2d\sigma'^2)$
where σ'^2 is an estimate of the variance of the error ϵ associated with each response measurement.
Typically σ'^2 is estimated using the full model containing all predictors.

Akaike's Information Criteria (AIC)

The AIC criterion is defined for a large class of models fit by maximum likelihood. In the case of a linear model with Gaussian errors, MLE and least squares are the same thing and the AIC is given by
$AIC=(1/(m\sigma'^2))*(RSS+2d\sigma'^2)$

Bayesian Information Criteria (BIC)

BIC is derived from a Bayesian point of view, and looks similar to the C_p and AIC- it is defined (up to irrelevant constants) as:
$BIC=(1/m\sigma'^2)(RSS+log(m)d\sigma'^2)$
Like C_p and AIC, the BIC will tend to take small values for a model with low test error.

Adjusted R2

Since the $R2$ always increases as more variables are added, the adjusted $R2$ accounts for that fact and introduces a penalty. The intuition is that once all the correct variables have been included in the model, additional noise variables will lead to a very small decrease in RSS, but an increase in k and hence will decrease the adjusted $R2$. In effect, we pay a price for the inclusion of unnecessary variables in the model.

$R2a=1-(RSS/(m-k-1))/(TSS/(m-1))=1-((1-R2)(m-1)/(m-k-1))$

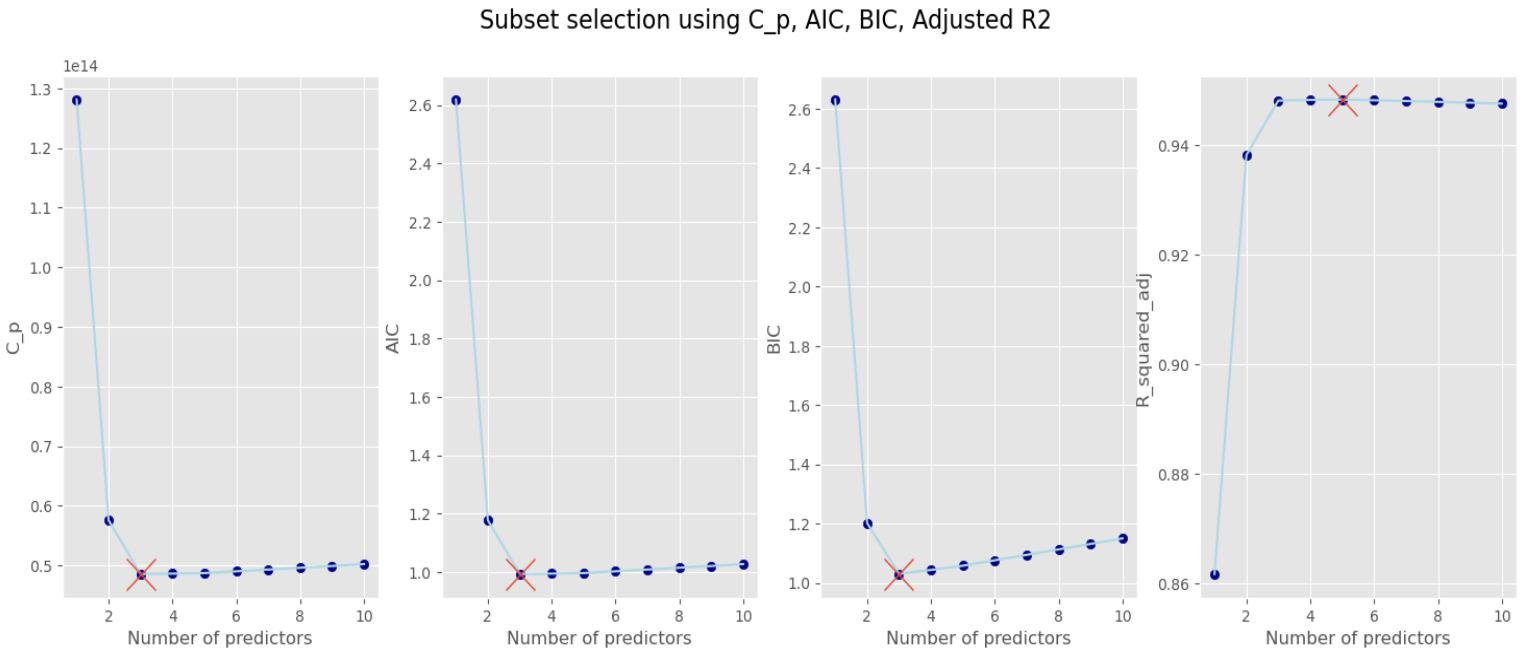
Theoretical justification

Cp, AIC, BIC all have rigorous theoretical justification that rely on asymptotic arguments, i.e. when the sample size m grows very large, whereas the adjusted $R2$, although quite intuitive, is not as well motivated in statistical theory.

Computing the Cp, AIC, BIC and R-square adjusted

	features	RSS	R_squared	numb_features	C_p	AIC	BIC	R_squared_adj
1	[V4]	3.834547e+16	0.862249	1	1.281444e+14	2.619107	2.631453	0.861787
2	[V4, V5]	1.707693e+16	0.938654	2	5.757546e+13	1.176768	1.201460	0.938240
3	[V4, V5, V1]	1.427903e+16	0.948705	3	4.857530e+13	0.992817	1.029855	0.948185
4	[V4, V5, V1, V10]	1.419845e+16	0.948994	4	4.863288e+13	0.993994	1.043377	0.948302
5	[V4, V5, V1, V10, V8]	1.412968e+16	0.949241	5	4.872983e+13	0.995975	1.057705	0.948378
6	[V4, V5, V1, V10, V8, V9]	1.411917e+16	0.949279	6	4.902096e+13	1.001925	1.076001	0.948240
7	[V4, V5, V1, V10, V8, V9, V6]	1.411039e+16	0.949310	7	4.931787e+13	1.007994	1.094415	0.948095
8	[V4, V5, V1, V10, V8, V9, V6, V2]	1.410229e+16	0.949339	8	4.961706e+13	1.014109	1.112876	0.947947
9	[V4, V5, V1, V10, V8, V9, V6, V2, V7]	1.409290e+16	0.949373	9	4.991195e+13	1.020136	1.131250	0.947802
10	[V4, V5, V1, V10, V8, V9, V6, V2, V7, V3]	1.409091e+16	0.949380	10	5.023147e+13	1.026667	1.150126	0.947629

Plotting the computed values as a function of number of features



Model 1: Based on Adjusted R^2

features = [V4, V5, V1, V10, V8]

X train dataset

	const	V4	V5	V1	V10	V8
203	1.0	6.393998	6.708380	4.492399	0	7.927108
266	1.0	8.794487	0.828996	6.851933	0	6.578809
152	1.0	8.285431	0.692489	5.682339	0	3.050583
9	1.0	7.896944	0.415595	4.852295	0	6.711999
233	1.0	6.702799	4.222613	6.358627	0	7.112359

Model Building

APPLYING MULTIPLE LINEAR REGRESSION ON THE SELECTED DATASET

OLS Regression Results

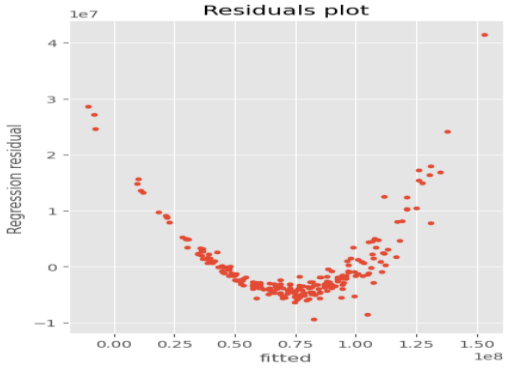
Dep. Variable:	Y	R-squared:	0.949
Model:	OLS	Adj. R-squared:	0.948
Method:	Least Squares	F-statistic:	877.7
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	2.29e-149
Time:	20:09:10	Log-Likelihood:	-4112.7
No. Observations:	240	AIC:	8237.
Df Residuals:	234	BIC:	8258.
Df Model:	5		
Covariance Type:	nonrobust		

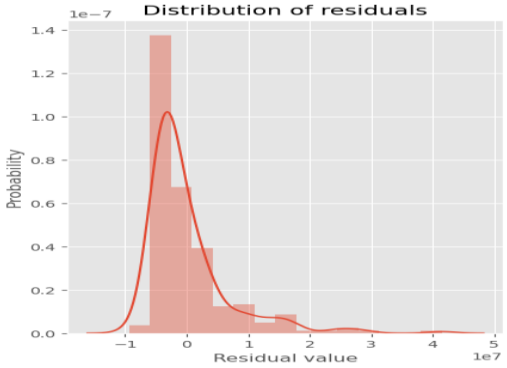
	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.719e+08	4.53e+06	-37.978	0.000	-1.81e+08	-1.63e+08
V4	3.093e+07	4.87e+05	63.493	0.000	3e+07	3.19e+07
V5	5.701e+06	2.92e+05	19.550	0.000	5.13e+06	6.28e+06
V1	3.361e+06	4.71e+05	7.130	0.000	2.43e+06	4.29e+06
V10	1.794e+06	1.25e+06	1.437	0.152	-6.66e+05	4.25e+06
V8	-3.265e+05	3e+05	-1.087	0.278	-9.18e+05	2.65e+05

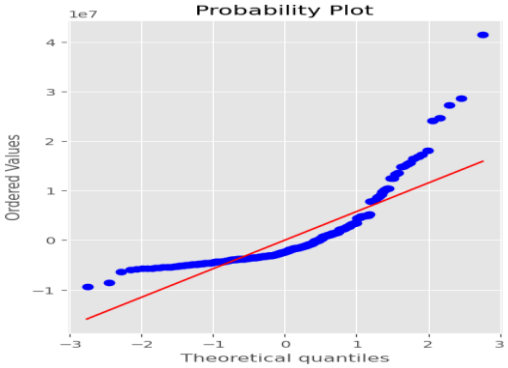
Omnibus:	153.969	Durbin-Watson:	1.969
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1046.881
Skew:	2.592	Prob(JB):	4.71e-228
Kurtosis:	11.821	Cond. No.	110.

Checking the Assumptions of the error terms

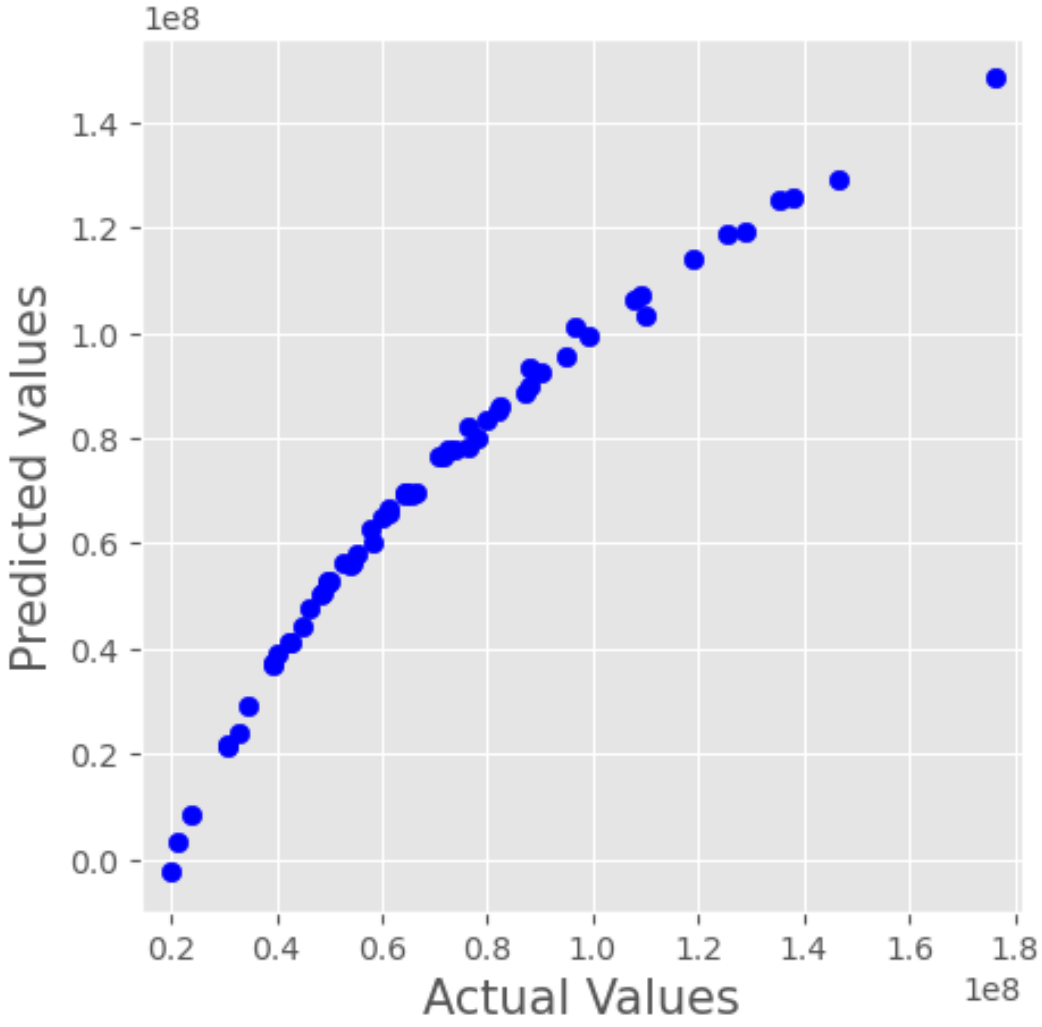
In linear regression we assume that the error term follows normal distribution. So we have to check this assumption before we can use the model for making predictions. We check this by looking at the histogram of the error term visually, making sure that the error terms are normally distributed around zero and that the left and right side are broadly similar.







Making predictions



Now based on the plot of the residual: Transformation in the response variable: Y=$log(Y)$ and $V4$=$sqrt(V4)$

	const	V4	V5	V1	V10	V8
232	1.0	2.772754	1.653905	4.711091	0	4.921883
59	1.0	2.620883	0.953635	5.256652	0	5.172716
6	1.0	2.785576	5.106688	3.675749	0	6.540107
185	1.0	2.449669	1.935401	4.163253	0	7.201239
173	1.0	2.639514	0.080228	5.630239	0	7.942917

Fitting the transformed model

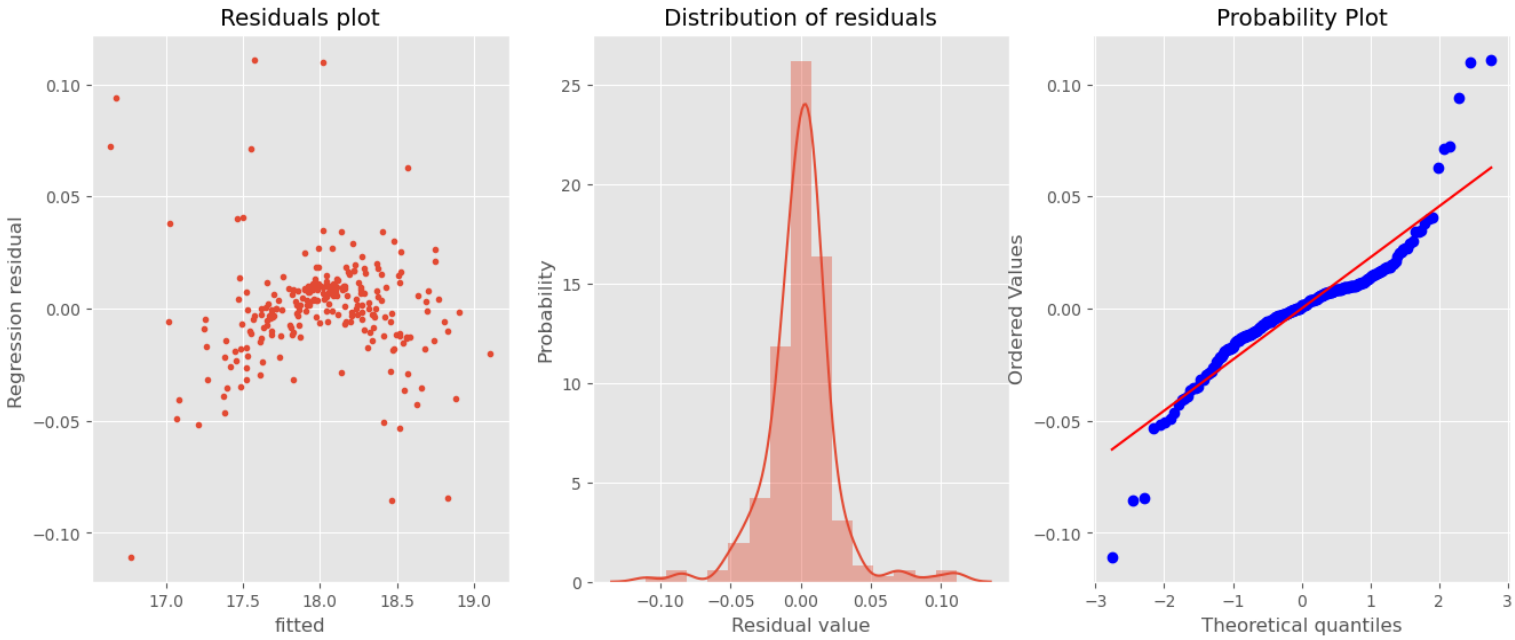
OLS Regression Results

Dep. Variable:	Y	R-squared:	0.997
Model:	OLS	Adj. R-squared:	0.997
Method:	Least Squares	F-statistic:	1.442e+04
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	4.46e-289
Time:	20:10:38	Log-Likelihood:	552.28
No. Observations:	240	AIC:	-1093.
Df Residuals:	234	BIC:	-1072.
Df Model:	5		
Covariance Type:	nonrobust		

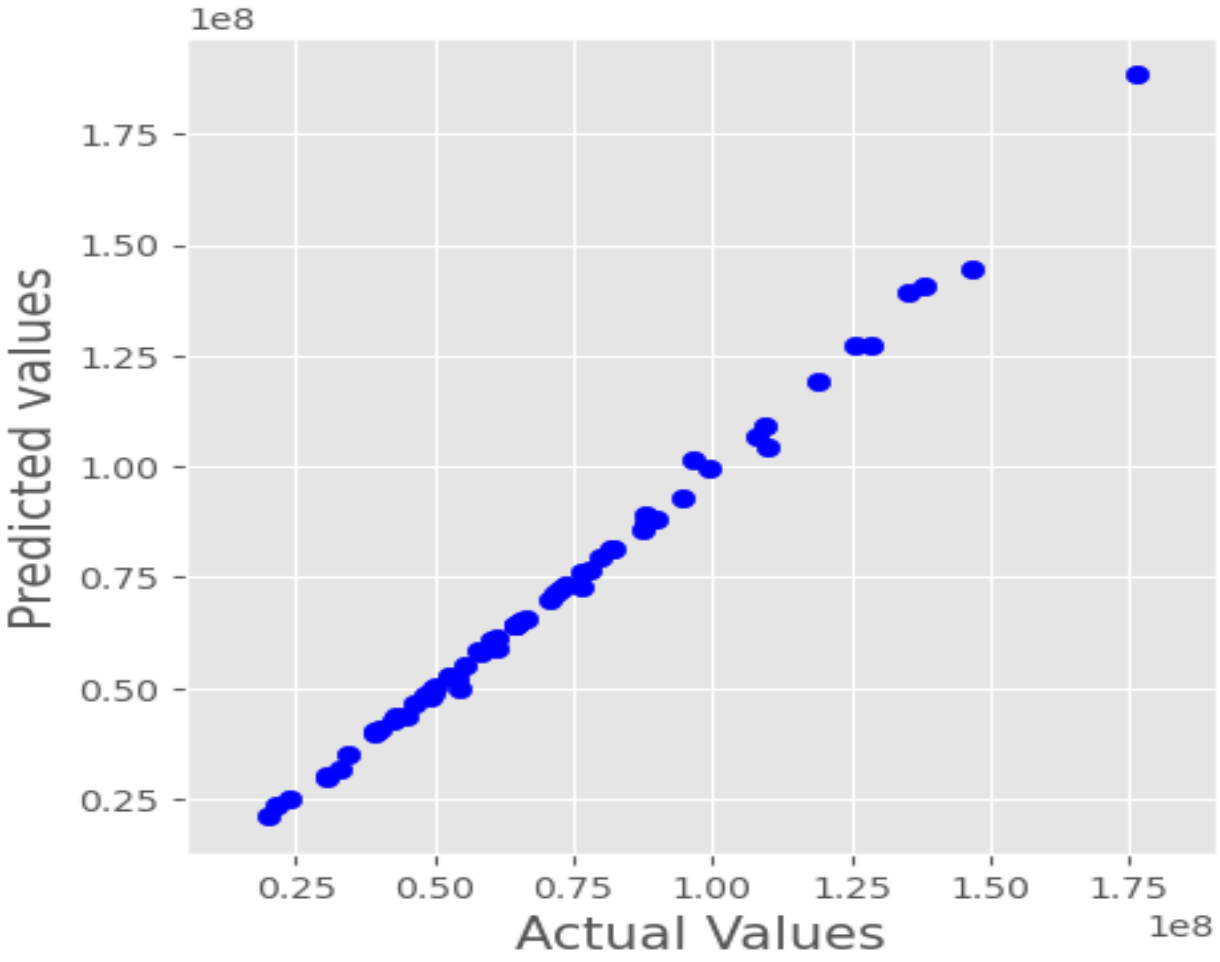
	coef	std err	t	P>|t|	[0.025	0.975]
const	11.2681	0.027	421.431	0.000	11.215	11.321
V4	2.3983	0.009	259.316	0.000	2.380	2.417
V5	0.0773	0.001	73.283	0.000	0.075	0.079
V1	0.0455	0.002	26.677	0.000	0.042	0.049
V10	0.0228	0.005	5.046	0.000	0.014	0.032
V8	0.0011	0.001	1.023	0.307	-0.001	0.003

Omnibus:	44.928	Durbin-Watson:	2.095
Prob(Omnibus):	0.000	Jarque-Bera (JB):	412.012
Skew:	0.313	Prob(JB):	3.41e-90
Kurtosis:	9.388	Cond. No.	148.

Checking the assumption of residuals



Making prediction



Making Box-cox transformation

	cols	box_cox_lambdas
0	V4	1.490793
1	V5	0.307718
2	V1	0.517748
3	V10	-0.492590
4	V8	0.935454

Model Building

Fitting the transformed model [V1, V4, V5, V8, V10] (Box-Cox Transform)

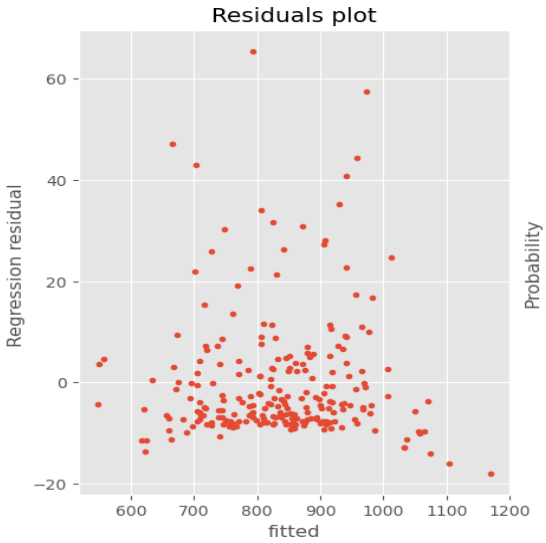
OLS Regression Results

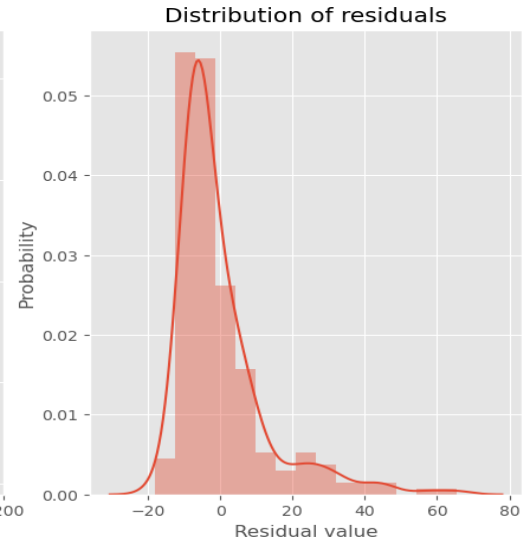
Dep. Variable:	Y	R-squared:	0.986
Model:	OLS	Adj. R-squared:	0.986
Method:	Least Squares	F-statistic:	3418.
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	1.77e-216
Time:	20:13:39	Log-Likelihood:	-946.57
No. Observations:	240	AIC:	1905.
Df Residuals:	234	BIC:	1926.
Df Model:	5		
Covariance Type:	nonrobust		

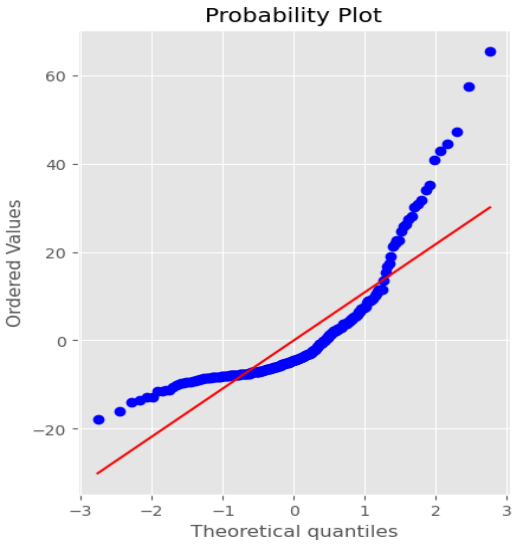
	coef	std err	t	P>|t|	[0.025	0.975]
const	79.2493	10.466	7.572	0.000	58.630	99.868
V4	29.9387	0.235	127.398	0.000	29.476	30.402
V5	84.3130	2.480	33.994	0.000	79.427	89.199
V1	47.1805	3.691	12.781	0.000	39.908	54.453
V10	4.3528	2.326	1.871	0.063	-0.231	8.936
V8	-0.8234	0.667	-1.235	0.218	-2.137	0.490

Omnibus:	132.428	Durbin-Watson:	2.118
Prob(Omnibus):	0.000	Jarque-Bera (JB):	614.646
Skew:	2.315	Prob(JB):	3.40e-134
Kurtosis:	9.326	Cond. No.	265.

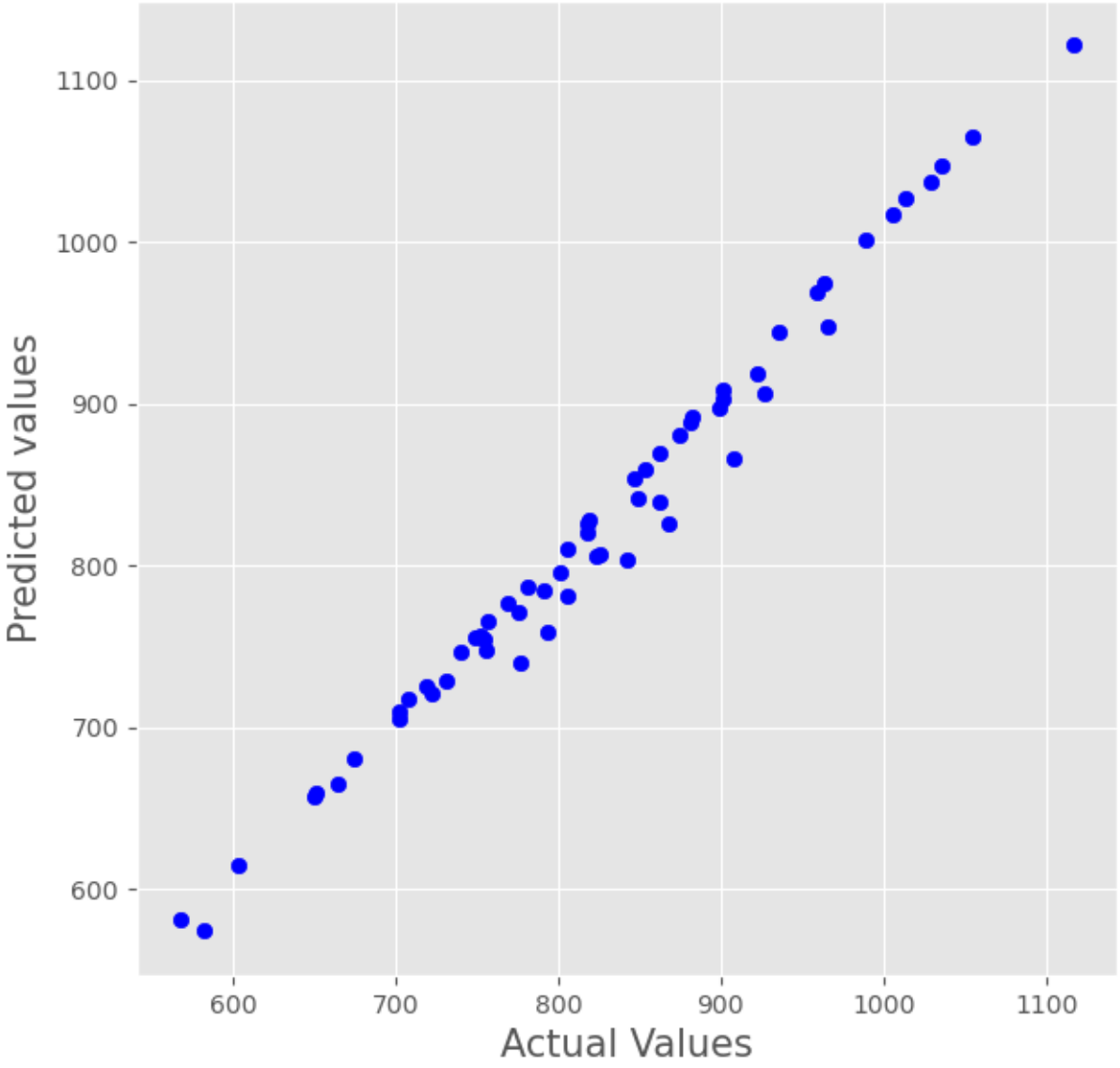
Checking the assumption of residuals







Making prediction



Analysis for Model selsected Based on AIC,BIC, And Mallos Cp

Regressors = [V4, V5, V1] **Response = Y**

	const	V4	V5	V1
232	1.0	7.688166	1.653905	4.711091
59	1.0	6.869028	0.953635	5.256652
6	1.0	7.759436	5.106688	3.675749
185	1.0	6.000877	1.935401	4.163253
173	1.0	6.967032	0.080228	5.630239

Model Building

APPLYING MULTIPLE LINEAR REGRESSION ON THE SELECTED DATASET

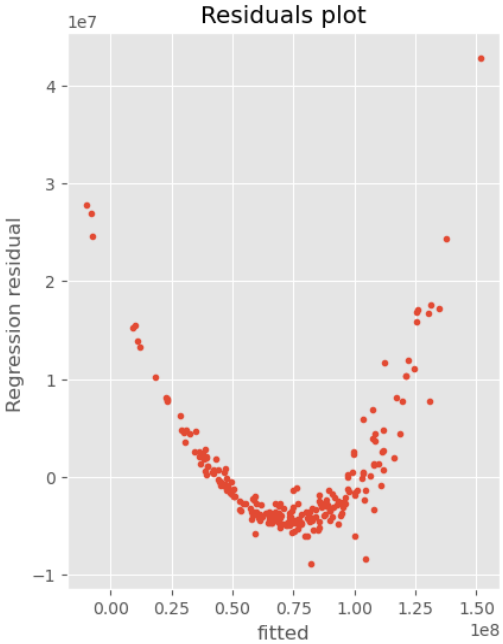
OLS Regression Results

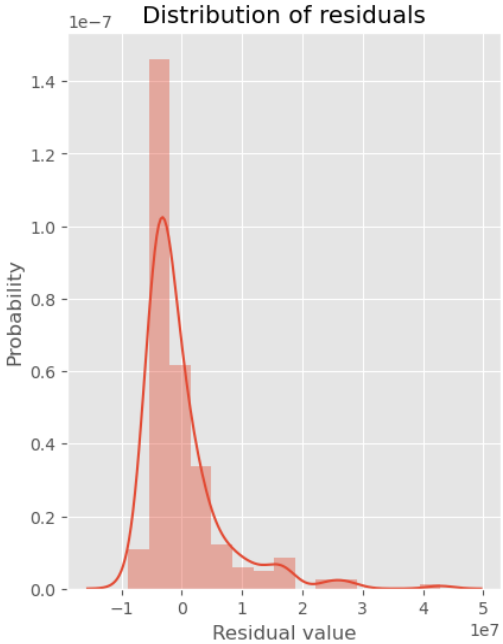
Dep. Variable:	Y	R-squared:	0.949
Model:	OLS	Adj. R-squared:	0.948
Method:	Least Squares	F-statistic:	1455.
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	7.29e-152
Time:	20:15:22	Log-Likelihood:	-4114.3
No. Observations:	240	AIC:	8237.
Df Residuals:	236	BIC:	8250.
Df Model:	3		
Covariance Type:	nonrobust		

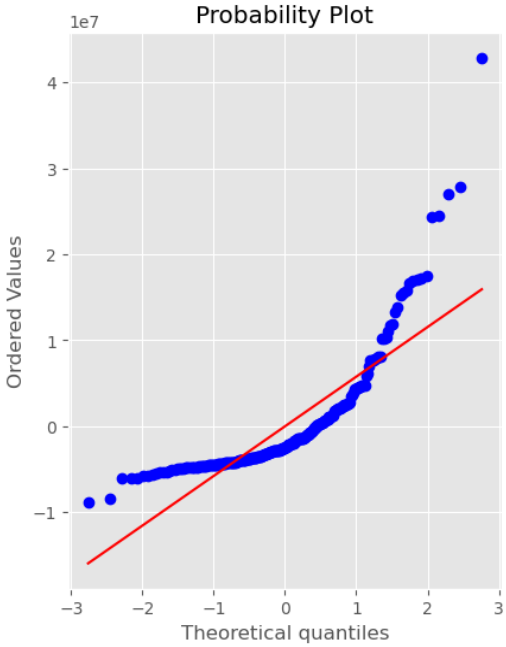
	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.734e+08	4.24e+06	-40.914	0.000	-1.82e+08	-1.65e+08
V4	3.094e+07	4.88e+05	63.373	0.000	3e+07	3.19e+07
V5	5.683e+06	2.9e+05	19.583	0.000	5.11e+06	6.25e+06
V1	3.322e+06	4.71e+05	7.046	0.000	2.39e+06	4.25e+06

Omnibus:	156.946	Durbin-Watson:	1.968
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1129.067
Skew:	2.630	Prob(JB):	6.70e-246
Kurtosis:	12.232	Cond. No.	87.1

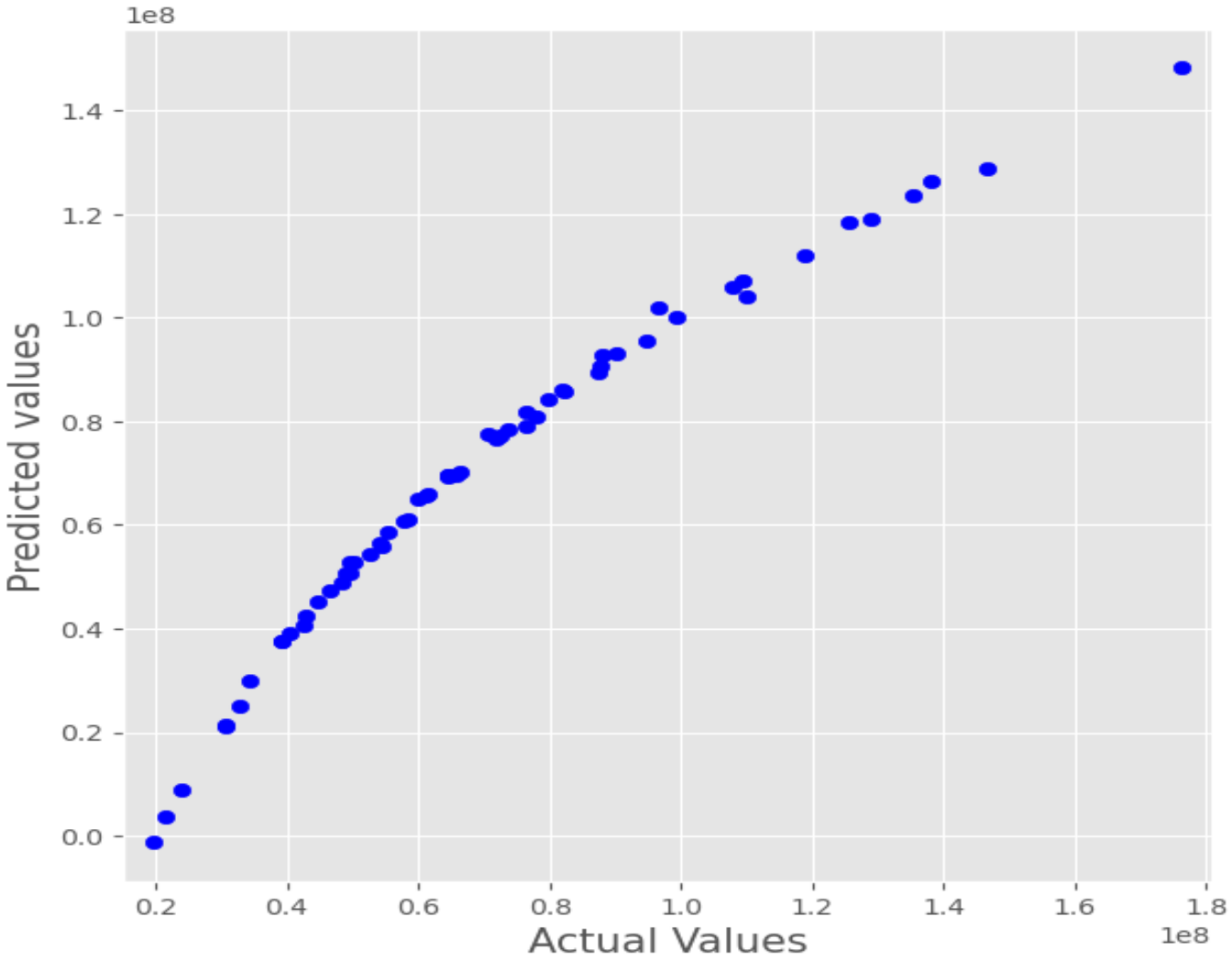
Checking for assumption of errors







Making prediction



Applying Box-Cox Transformation on selected features

	const	V4	V5	V1
232	1.0	20.920804	1.167454	2.231040
59	1.0	17.686327	0.985498	2.361271
6	1.0	21.210582	1.651609	1.962035
185	1.0	14.459627	1.225306	2.092715
173	1.0	18.063827	0.460088	2.446717

Model Building

Fitting the transformed model

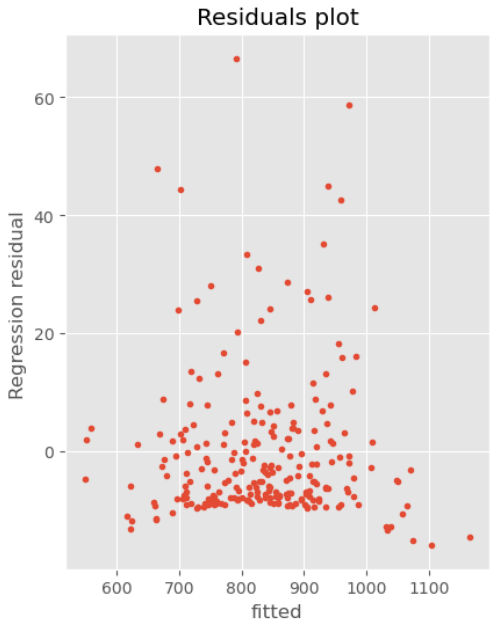
OLS Regression Results

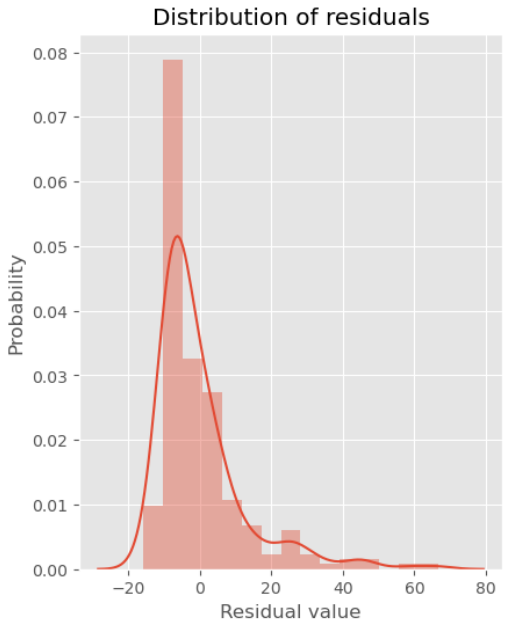
Dep. Variable:	Y	R-squared:	0.986
Model:	OLS	Adj. R-squared:	0.986
Method:	Least Squares	F-statistic:	5630.
Date:	Tue, 25 Apr 2023	Prob (F-statistic):	3.34e-219
Time:	20:16:14	Log-Likelihood:	-949.00
No. Observations:	240	AIC:	1906.
Df Residuals:	236	BIC:	1920.
Df Model:	3		
Covariance Type:	nonrobust		

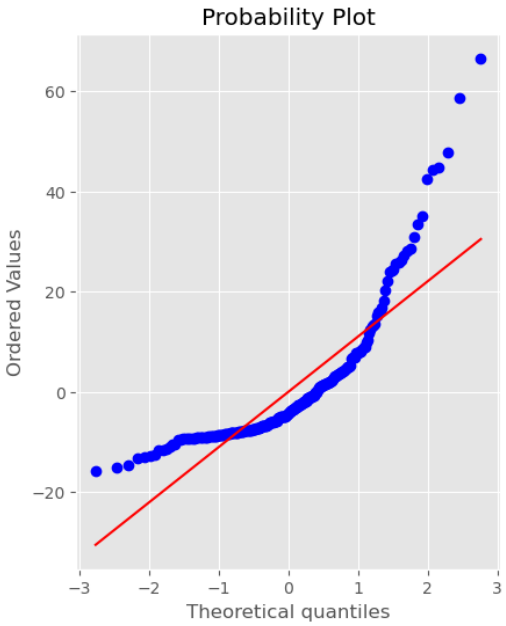
	coef	std err	t	P>|t|	[0.025	0.975]
const	76.6850	10.094	7.597	0.000	56.799	96.571
V4	29.9428	0.236	126.686	0.000	29.477	30.408
V5	84.0446	2.484	33.839	0.000	79.152	88.938
V1	46.8604	3.704	12.650	0.000	39.563	54.158

Omnibus:	133.107	Durbin-Watson:	2.078
Prob(Omnibus):	0.000	Jarque-Bera (JB):	633.016
Skew:	2.316	Prob(JB):	3.49e-138
Kurtosis:	9.468	Cond. No.	248.

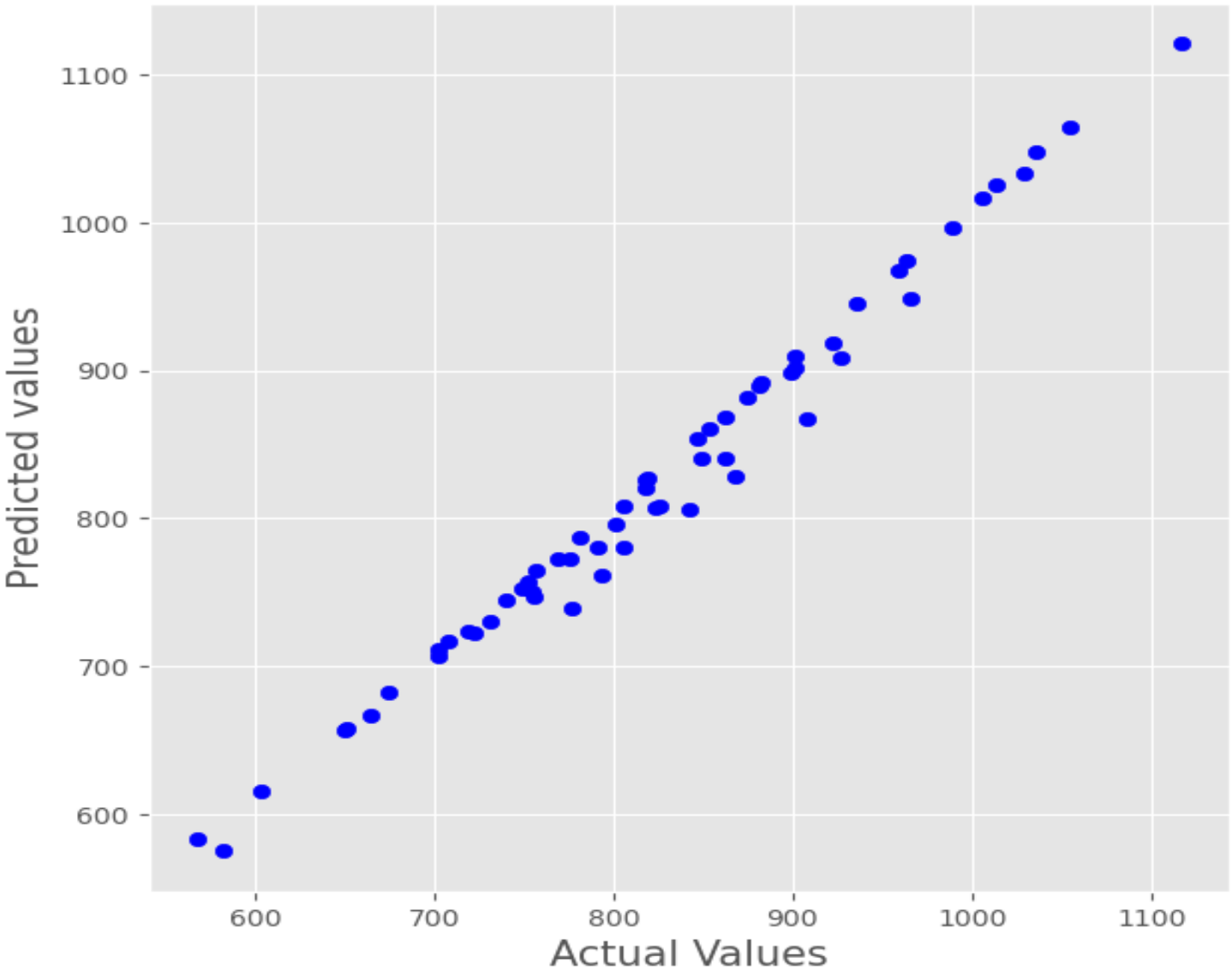
Checking for model assumptions





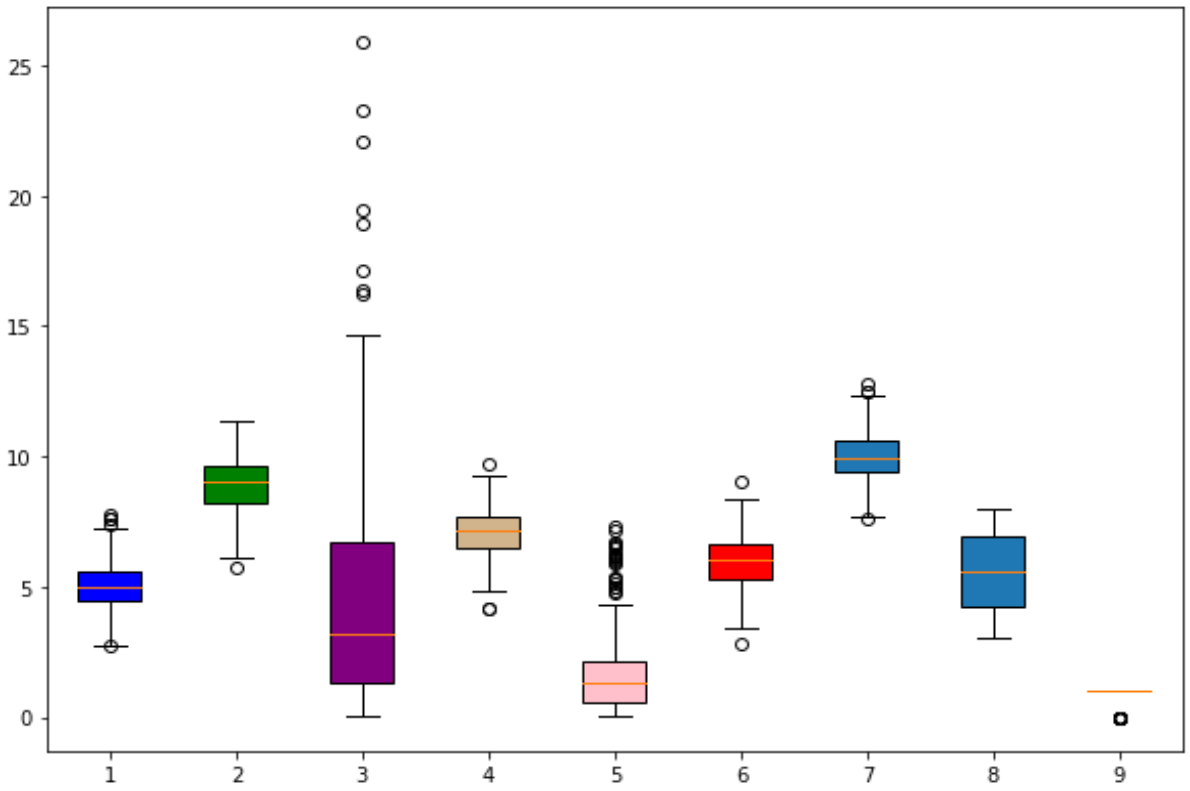


Making predictions



TESTING FOR INFLUENTIAL OBSERVATIONS for MODEL based on Adjusted R^2

Box Plot of all the features

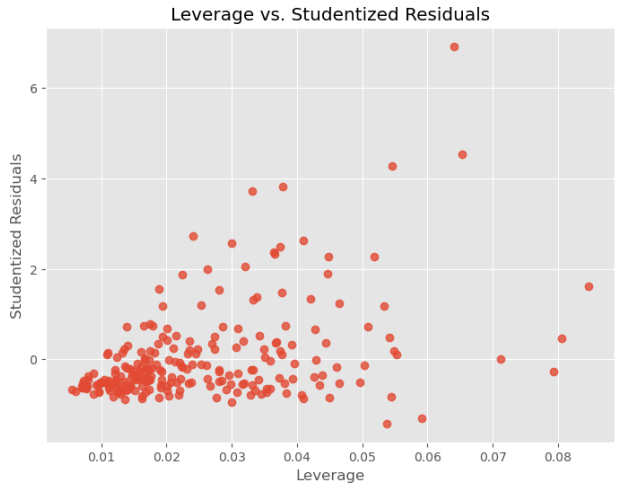


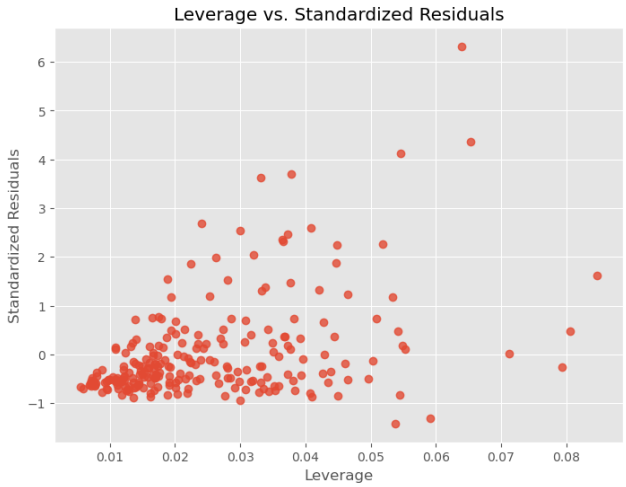
Identifying Outlying Y Observations-Studentized Deleted Residuals

Summary of the Influentials

	standard_resid	hat_diag	student_resid	e_i
232	-0.452036	0.007983	-0.451266	-3.054031e+06
59	-0.596600	0.006992	-0.595777	-4.032741e+06
6	-0.430597	0.039282	-0.429847	-2.862927e+06
185	0.421475	0.020034	0.420733	2.830208e+06
173	-0.511221	0.020557	-0.510413	-3.431940e+06

Plotting Leverage V|s Studentized residuals





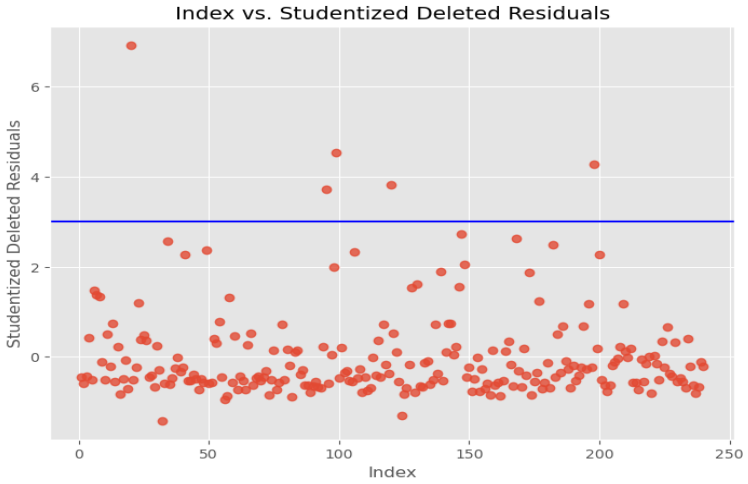
Calculating Studentized Deleted Residuals

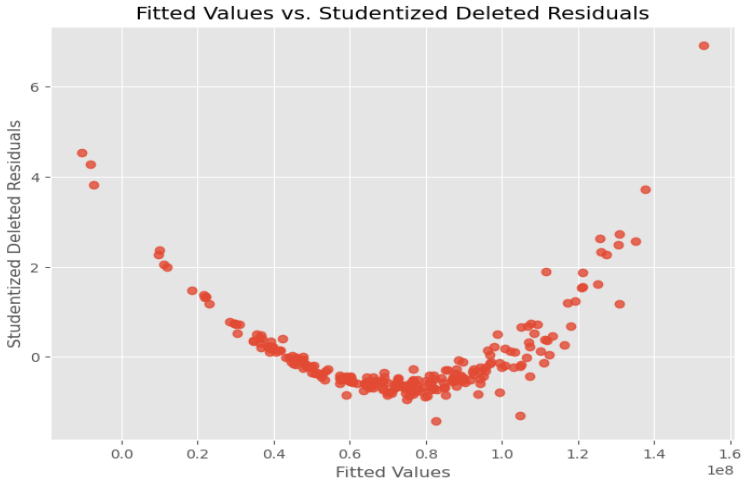
	dfb_const	dfb_V4	dfb_V5	dfb_V1	dfb_V10	dfb_V8	cooks_d	standard_resid	hat_diag	dffits_internal	student_resid	dffits
232	0.001121	-0.019030	0.002192	0.009781	0.011594	0.012775	0.000274	-0.452036	0.007983	-0.040551	-0.451266	-0.040482
59	-0.015481	0.011285	0.020396	-0.009110	0.016959	0.013441	0.000418	-0.596600	0.006992	-0.050062	-0.595777	-0.049992
6	0.011592	-0.024508	-0.063188	0.039072	0.007120	-0.025615	0.001264	-0.430597	0.039282	-0.087071	-0.429847	-0.086919
185	0.029727	-0.032432	0.002925	-0.026549	-0.012692	0.031644	0.000605	0.421475	0.020034	0.060262	0.420733	0.060156
173	0.014438	0.006369	0.031685	-0.015665	0.019779	-0.049495	0.000914	-0.511221	0.020557	-0.074062	-0.510413	-0.073945

Residuals, Diagonal Elements of the Hat Matrix, and Studentized Deleted Residuals of predictor Variables.

	standard_resid	hat_diag	student_resid	e_i	di
232	-0.452036	0.007983	-0.451266	-3.054031e+06	-0.451266
59	-0.596600	0.006992	-0.595777	-4.032741e+06	-0.595777
6	-0.430597	0.039282	-0.429847	-2.862927e+06	-0.429847
185	0.421475	0.020034	0.420733	2.830208e+06	0.420733
173	-0.511221	0.020557	-0.510413	-3.431940e+06	-0.510413

Plotting Studentized Deleted Residuals V|S Index





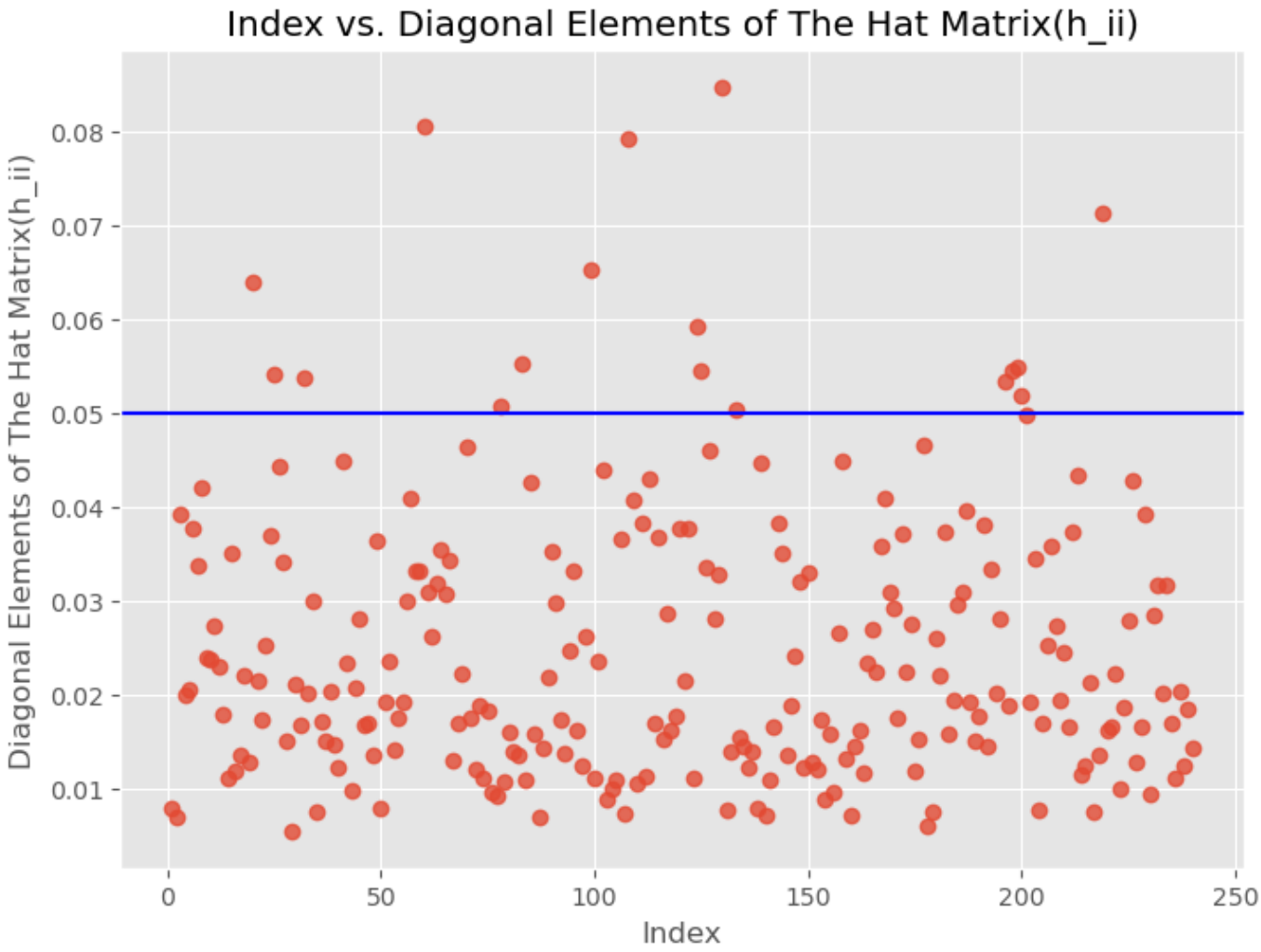
Identifying the Index of Observation in Y

Criteria = If Studentized Deleted Residuals >=3 then it said to be outlier

Filter the column to only include values greater than 3

```
Index
288    6.927341
280    3.720156
198    4.545367
141    3.813676
80     4.275942
```

Identifying Outlying X Observations-Hat Matrix: Leverage ValuesUse of Diagonal Elements of The Hat matrix : h_ii



Thresold Value= $2p/n=2*6/140=0.05$

Potential Influential observation in X (h_ii > 0.05)

Index	
288	0.063997
79	0.054111
101	0.053710
206	0.080539
176	0.050810
177	0.055285
198	0.065343
254	0.079281
272	0.059152
100	0.054441
36	0.084737
236	0.050316
103	0.053333
80	0.054523
205	0.054885

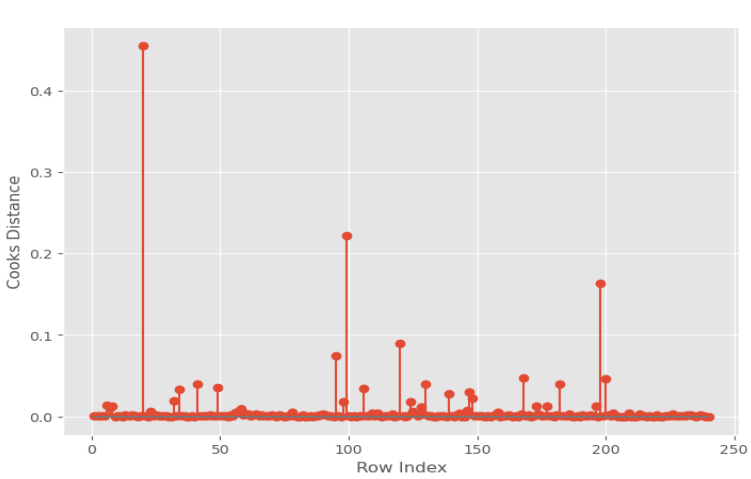
```
34      0.051869
48      0.071211
```

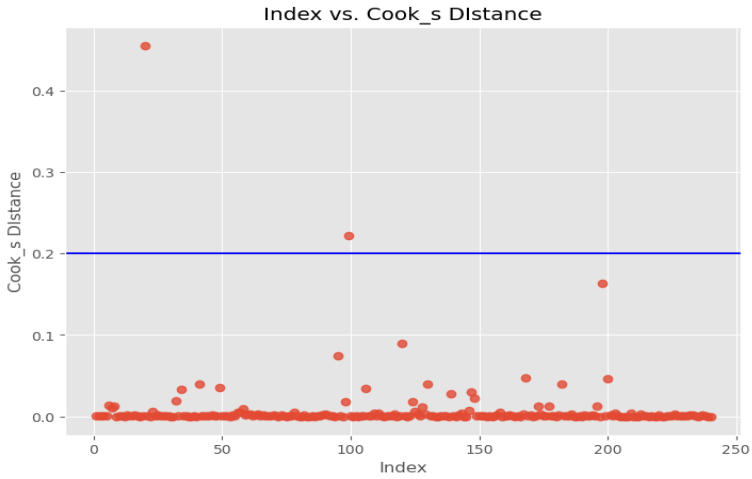
Identifying Influential Cases-DFFITS, *Cook's* Distance and DFBETAS Measures

1. ***Cook's*** Distance

	dfb_const	dfb_V4	dfb_V5	dfb_V1	dfb_V10	dfb_V8	cooks_d	dffits
232	0.001121	-0.019030	0.002192	0.009781	0.011594	0.012775	0.000274	-0.040482
59	-0.015481	0.011285	0.020396	-0.009110	0.016959	0.013441	0.000418	-0.049992
6	0.011592	-0.024508	-0.063188	0.039072	0.007120	-0.025615	0.001264	-0.086919
185	0.029727	-0.032432	0.002925	-0.026549	-0.012692	0.031644	0.000605	0.060156
173	0.014438	0.006369	0.031685	-0.015665	0.019779	-0.049495	0.000914	-0.073945

Plotting Cooks distance



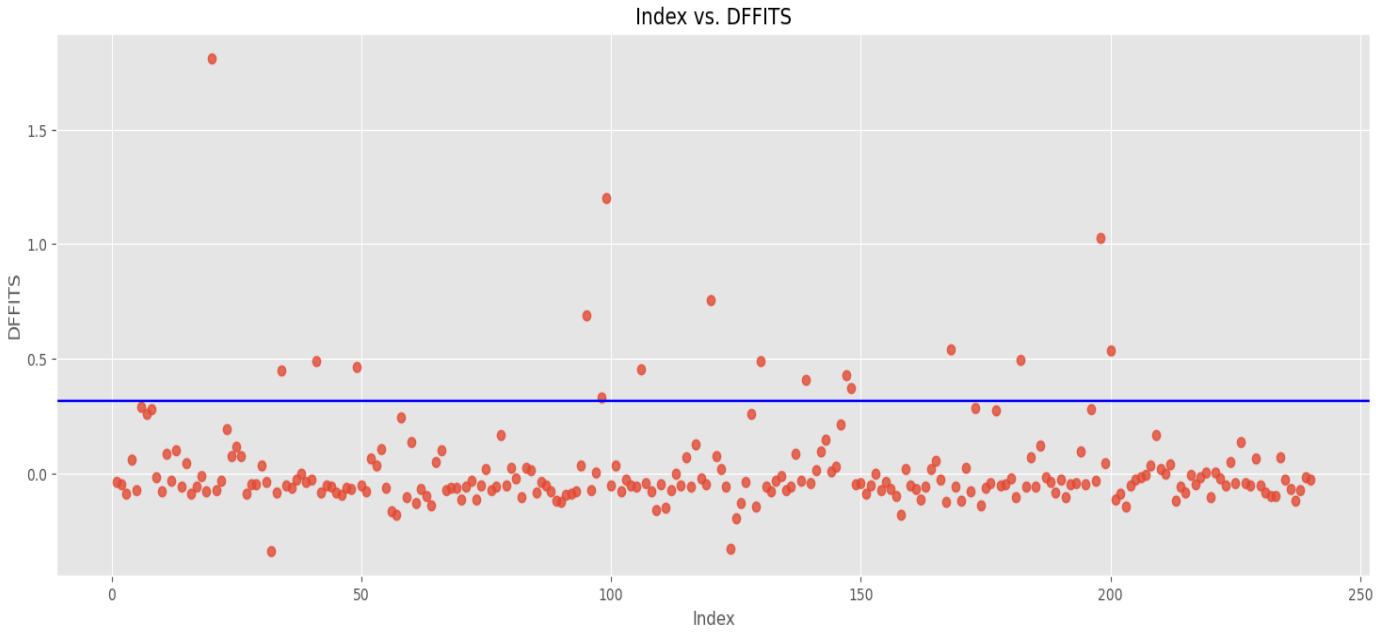


Filtered column cook distance [Cooks distance > 0.20}

```
index
288     0.455396
198     0.222072
```

2. *DIFITS*

THERSOLD VALUE = $2*Sqrt(p/n)=2*sqrt(6/240)=0.3162272$



Filtered column dffits [Dffits > 0.316]

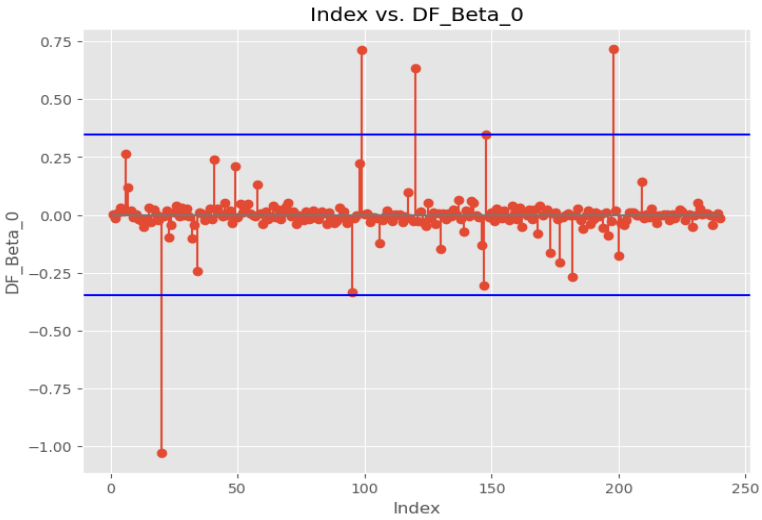
Index	
288	1.811366
101	-0.338554
224	0.449668
228	0.489979
181	0.462417
280	0.688885
65	0.328063
198	1.201825
142	0.455084
141	0.755634
272	-0.328201
36	0.491748
27	0.408947
285	0.427904
138	0.372313
81	0.542932
201	0.492066
80	1.026825
34	0.533212

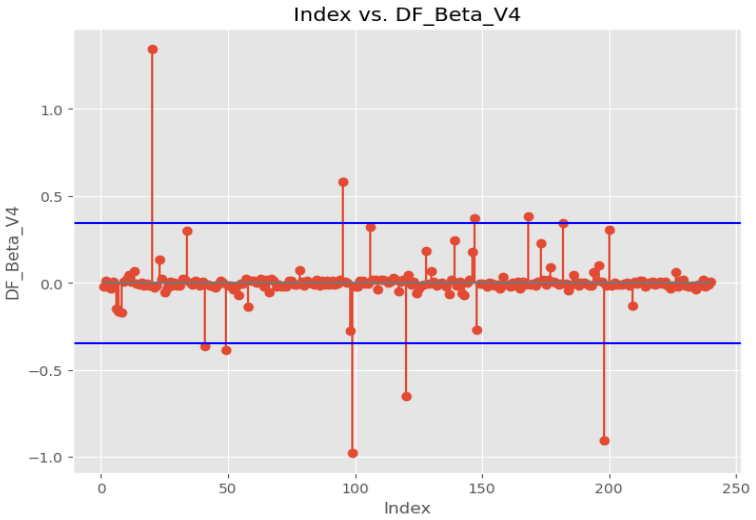
3. $DFBETA$

THERSOLD Value = $2*Sqrt(p+1)/(n-p-1)=2*sqrt(7)/233=0.3466672$

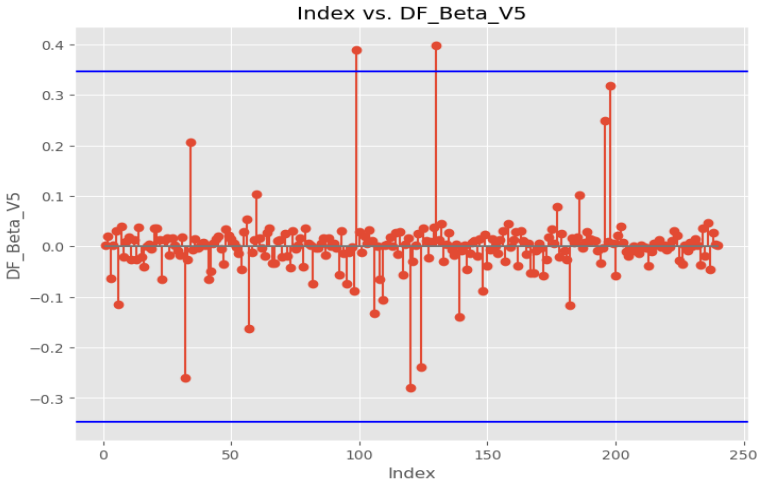
a.) For Constant and V4

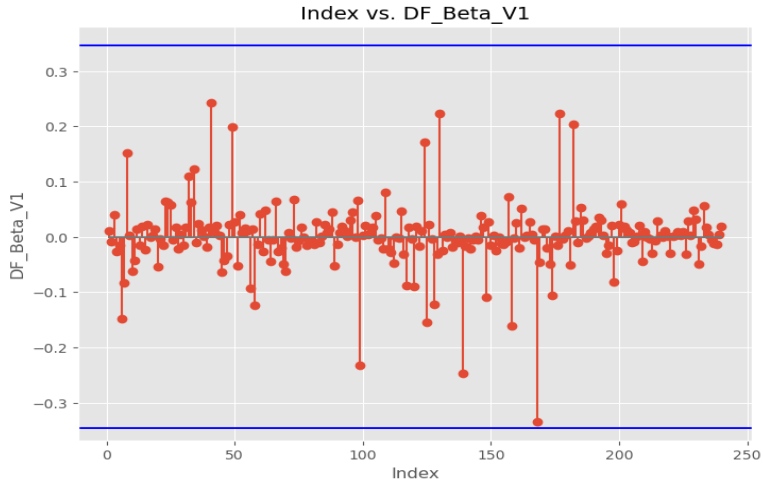
In [



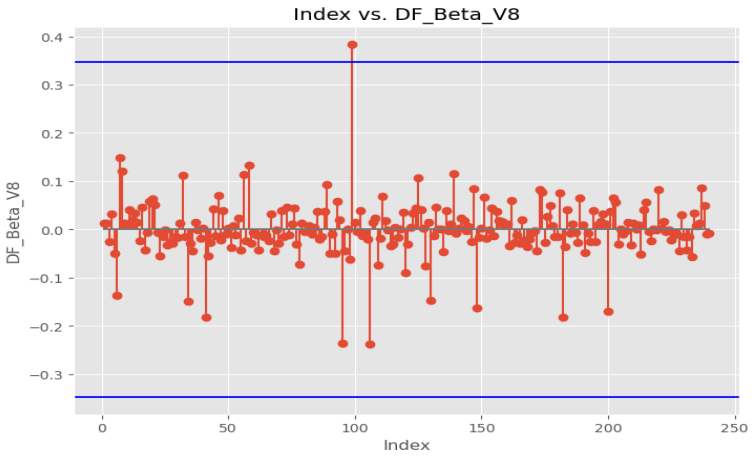


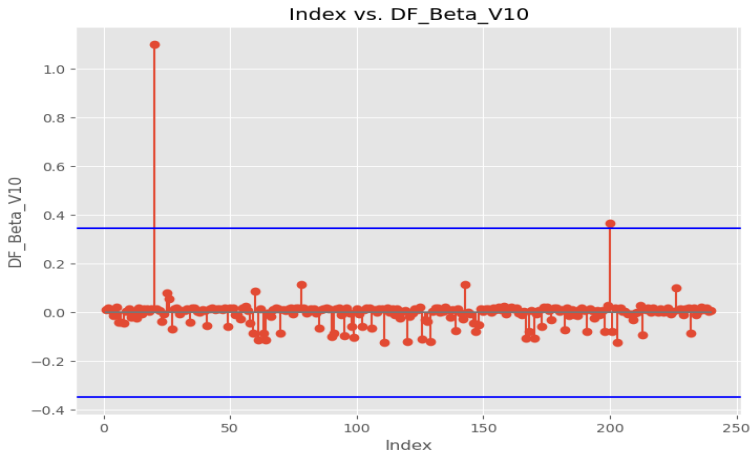
b.) V5 and V1





C,) V8 and V10





Potential influential observation in data set:

```
Index
288   -1.028718
198    0.715207
141    0.634918
138    0.348762
80     0.719053
Name: dfb const, dtype: float64 288    1.346715
228   -0.361075
181   -0.385700
280    0.580858
198   -0.974291
141   -0.647431
285    0.372216
81     0.385613
80    -0.903996
Name: dfb_V4, dtype: float64 198    0.388563
36     0.398705
Name: dfb_V5, dtype: float64 Series([], Name: dfb_V1, dtype: float64) 288    1.10124
3
34     0.365745
Name: dfb_V10, dtype: float64 198    0.383965
```

Final Index of Potential Outliers and Influential variables

$IndexN0$=288,198

So let's use the findings. I fitted two different regression models. First, I removed the observations that were deemed influential and fitted an OLS model. Second, I removed all outliers, and then fitted another OLS model.

Fitting the model after removing the common outliers in X and Y

Model Building

fitting the transformed model

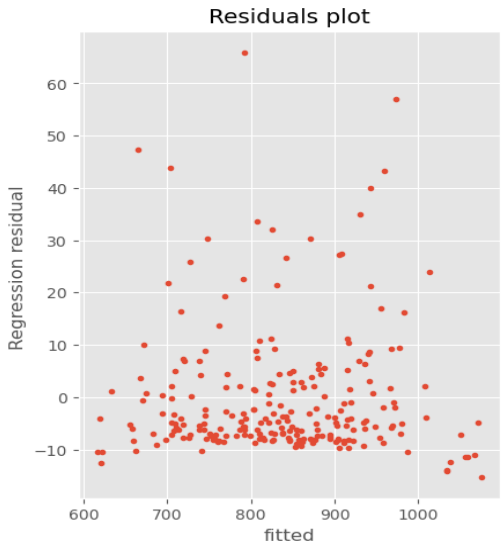
OLS Regression Results

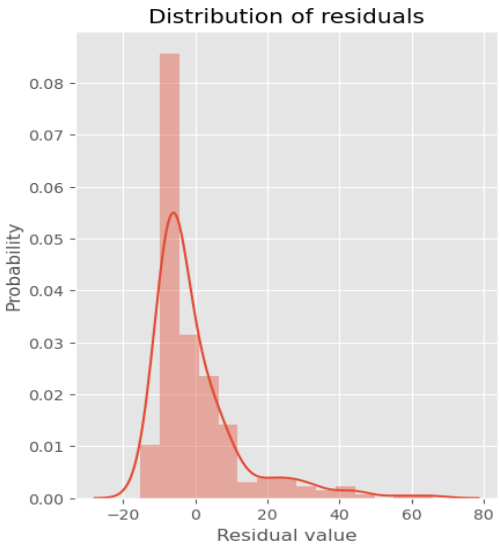
Dep. Variable:	Y	R-squared:	0.984
Model:	OLS	Adj. R-squared:	0.984
Method:	Least Squares	F-statistic:	2909.
Date:	Wed, 26 Apr 2023	Prob (F-statistic):	5.69e-205
Time:	01:44:10	Log-Likelihood:	-927.08
No. Observations:	235	AIC:	1866.
Df Residuals:	229	BIC:	1887.
Df Model:	5		
Covariance Type:	nonrobust		

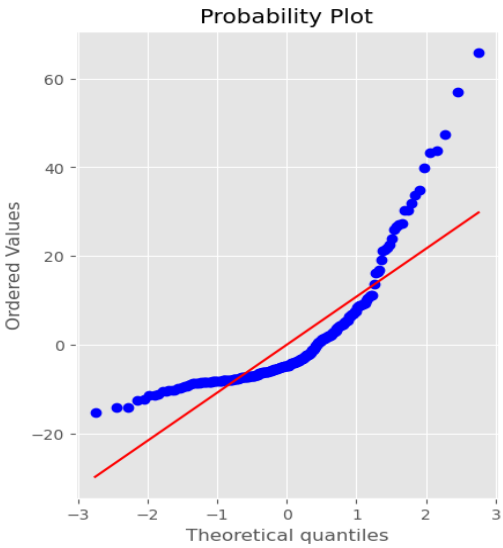
	coef	std err	t	P>|t|	[0.025	0.975]
const	76.4070	10.722	7.126	0.000	55.280	97.534
V4	30.1014	0.258	116.839	0.000	29.594	30.609
V5	84.2398	2.507	33.597	0.000	79.299	89.180
V1	47.3071	3.703	12.775	0.000	40.011	54.604
V10	4.8881	2.364	2.068	0.040	0.231	9.546
V8	-0.8972	0.671	-1.336	0.183	-2.220	0.426

Omnibus:	131.938	Durbin-Watson:	2.104
Prob(Omnibus):	0.000	Jarque-Bera (JB):	620.488
Skew:	2.350	Prob(JB):	1.83e-135
Kurtosis:	9.425	Cond. No.	268.

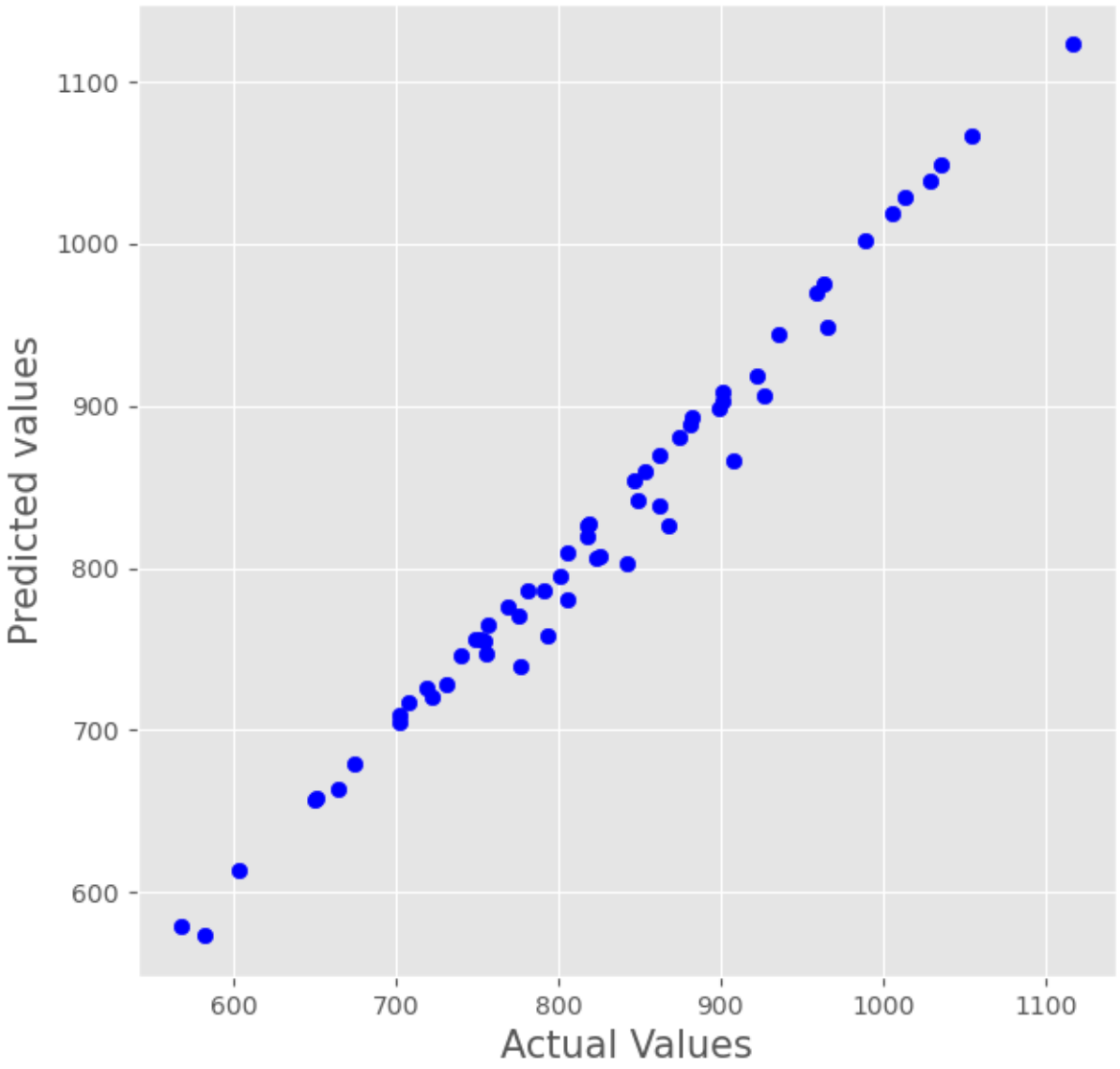
Checking for Error Assumption:







Making Prediction with Improved Model



Detecting and Removing Multicollinearity using VIF in model based on Adjusted R^2:

Regressors in original model : [V4,V5,V1,V10,V8]

VIF TABLE before removing Multicollinearity
Threshold Value of VIF = 5

:

	variables	VIF
0	V4	26.309253
1	V5	2.153232
2	V1	23.681275
3	V10	1.178003
4	V8	13.188072

VIF TABLE after removing Multicollinearity

	variables	VIF
0	V4	2.377264
1	V5	2.140486
2	V10	1.172391

Model Building

Fitting the transformed model

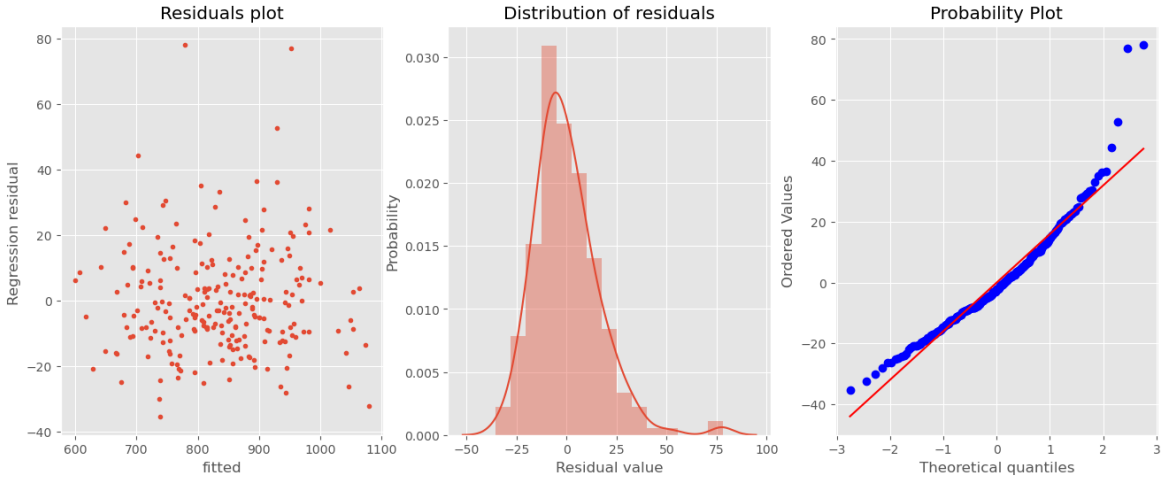
OLS Regression Results

Dep. Variable:	Y	R-squared:	0.973
Model:	OLS	Adj. R-squared:	0.973
Method:	Least Squares	F-statistic:	2822.
Date:	Wed, 26 Apr 2023	Prob (F-statistic):	1.18e-181
Time:	02:45:14	Log-Likelihood:	-990.35
No. Observations:	235	AIC:	1989.
Df Residuals:	231	BIC:	2003.
Df Model:	3		
Covariance Type:	nonrobust		

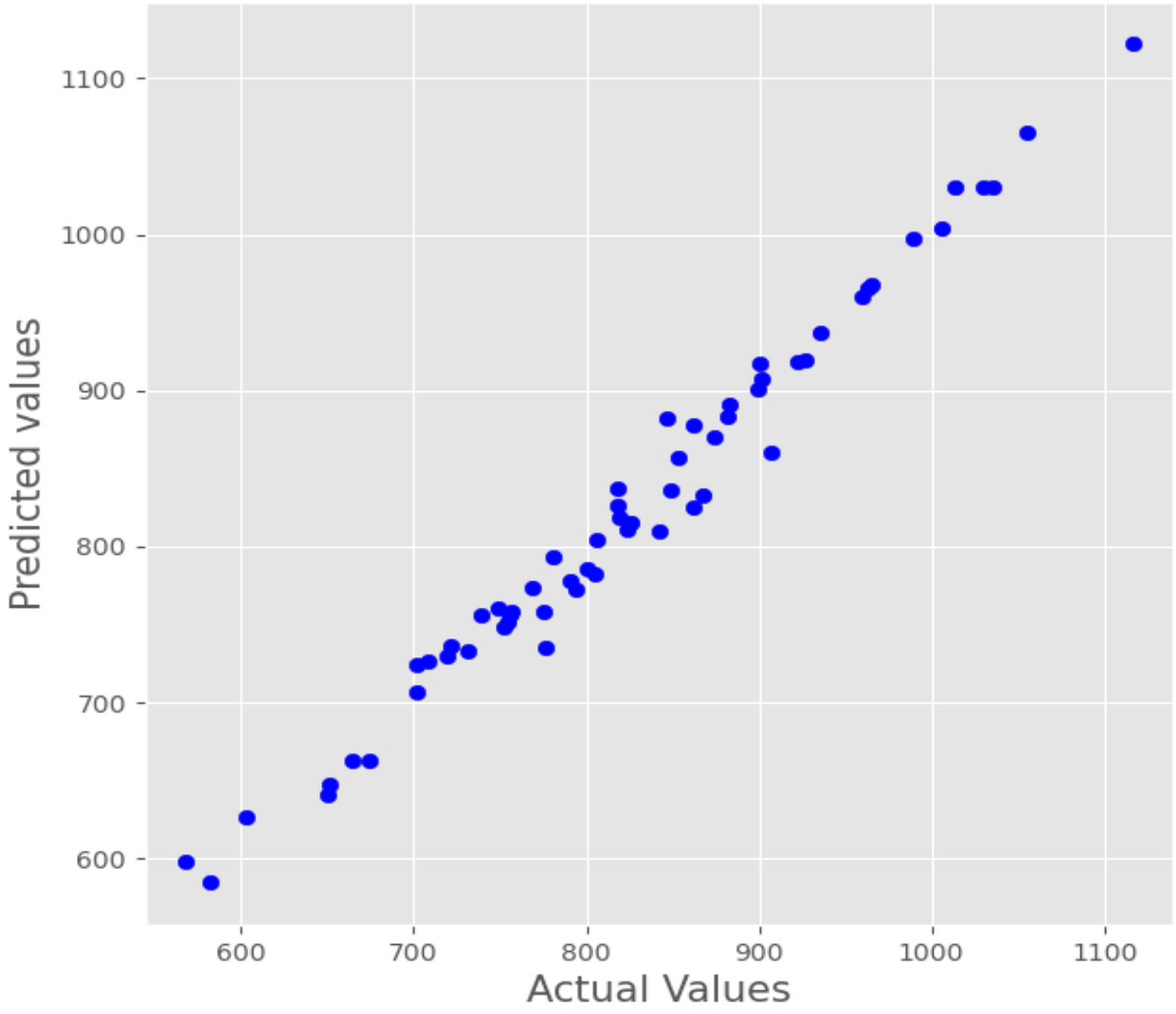
	coef	std err	t	P>|t|	[0.025	0.975]
const	182.0054	7.492	24.292	0.000	167.243	196.768
V4	30.1176	0.336	89.705	0.000	29.456	30.779
V5	82.7479	3.260	25.382	0.000	76.325	89.171
V10	4.6243	3.076	1.503	0.134	-1.436	10.685

Omnibus:	65.353	Durbin-Watson:	2.127
Prob(Omnibus):	0.000	Jarque-Bera (JB):	178.304
Skew:	1.219	Prob(JB):	1.91e-39
Kurtosis:	6.503	Cond. No.	137.

Checking for Error Assumption:



Making prediction with improved model



FINAL MODEL,

1. **BASED ON ADJUSTED R^2**
   Y= 76.4078+30.1014*V4 + 84.2398* V5 + 47.3071* V1 + 4.8881*V10 – 0.9872*V8

2. **BASED ON Cp , AIC AND BIC,**
   Y= 76.6850 + 29.9428*V4 + 84.0446*V5 + 46.8604*V1

3. **AFTER REMOVING MULTICOLLINEARITY,**
   Y= 182.0054 + 30.1176*V4 + 82.7449*V5 + 4.6243*V10

******