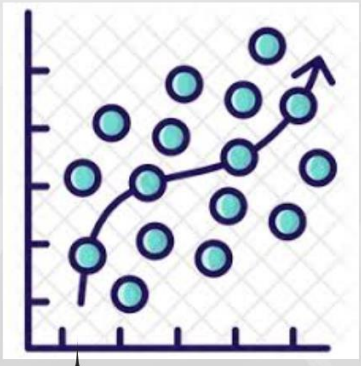
REGRESSION ANALYSIS



Presented By

SHARIF KAMBOJ

SONAM

SUJAN MANDAL

Guided By

PROF.SIULI MUKHOPADHYAY

SUBHAM NIPADHKAR

REGRESSION ANALYSIS

Acknowledgment

We Would like to express our deepest gratitude to Professor **Siuli Mukhopadhyay** Maam for contributing their valuable time and efforts in helping us out with this project. We would like to express our special thanks to Sir Shubham Nipadhkar and all our friends Their suggestions and feedback have helped us a lot in improving the quality and effectiveness of the project. Lastly, We like to thank all our supporters who have motivated us to fulfill this project before the timeline

1 Objectives

1.1 Introduction

CHAPTER-1

Objective

The primary objective of this regression analysis was to construct robust predictive models capable of accurately estimating the relationship between the predictor variables and the target variable. Through a methodical exploration of the data, meticulous model development, thorough diagnostics, and rigorous validation procedures, our aim was to establish regression models that exhibit reliability and predictive power. By adhering to established methodologies and rigorously testing our models, we aimed to uncover meaningful insights and facilitate informed decision-making. Through this process, we sought to contribute valuable knowledge and understanding to the domain under study, ultimately driving toward actionable outcomes and improvements.

CHAPTER-2

Dataset Description

About the Dataset: The dataset contains 522 rows and 13 columns out of which are 3 columns are numerical and rest all are categorical. The first five rows of the dataset are given below:

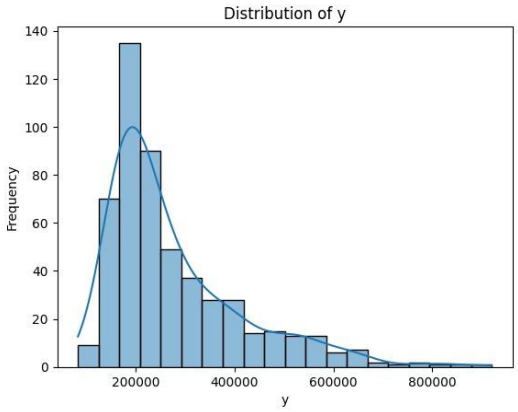
	id\t	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
0	1	360000	3032	4	4	1	2	0	1972	2	1	22221	0
1	2	340000	2058	4	2	1	2	0	1976	2	1	22912	0
2	3	250000	1780	4	3	1	2	0	1980	2	1	21345	0
3	4	205500	1638	4	2	1	2	0	1963	2	1	17342	0
4	5	275500	2196	4	3	1	2	0	1968	2	7	21786	0

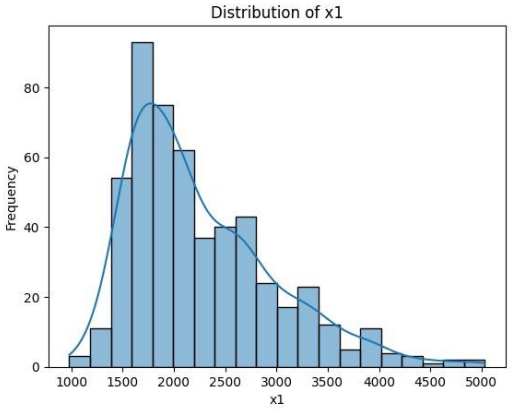
About the columns: We can see that most of the columns are Categorical and 3 columns seem to be numerical. We don't have a description of the data so we cannot infer anything about the columns except their data types.

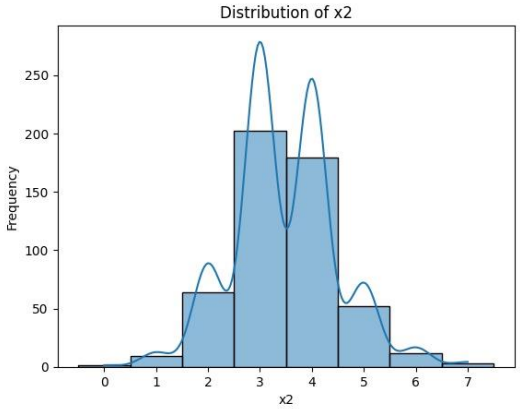
columns	Type	Category
Id/t	Integer	Numerical
x1	Integer	Numerical
x2	Integer	Categorical
x3	Integer	Categorical
x4	Integer	Categorical
x5	Integer	Categorical
x6	Integer	Categorical
x7	Integer	Numerical
x8	Integer	Categorical
x9	Integer	Categorical
x10	Integer	Numerical
x11	Integer	Categorical
y	Integer	Numerical

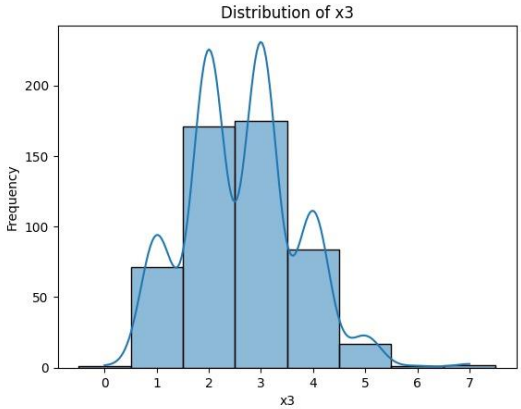
	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
count	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000	522.000000
mean	277894.147510	2260.626437	3.471264	2.641762	0.831418	2.099617	0.068966	1966.904215	2.183908	3.344828	24369.704981	0.021073
std	137923.397269	711.065933	1.014358	1.064169	0.374742	0.653970	0.253639	17.637924	0.641413	2.562812	11684.075549	0.143765
min	84000.000000	980.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1885.000000	1.000000	1.000000	4560.000000	0.000000
25%	180000.000000	1701.250000	3.000000	2.000000	1.000000	2.000000	0.000000	1956.000000	2.000000	1.000000	17204.750000	0.000000
50%	229900.000000	2061.000000	3.000000	3.000000	1.000000	2.000000	0.000000	1966.000000	2.000000	2.000000	22200.000000	0.000000
75%	335000.000000	2636.250000	4.000000	3.000000	1.000000	2.000000	0.000000	1981.000000	3.000000	7.000000	26786.750000	0.000000
max	920000.000000	5032.000000	7.000000	7.000000	1.000000	7.000000	1.000000	1998.000000	3.000000	11.000000	86830.000000	1.000000

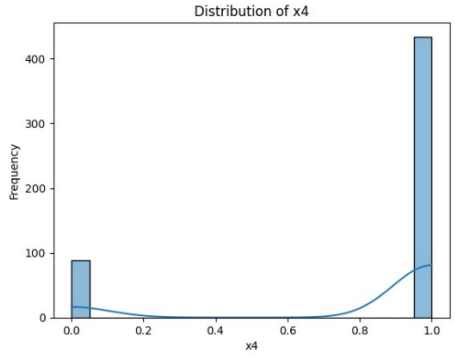
Univariate Data Analysis:

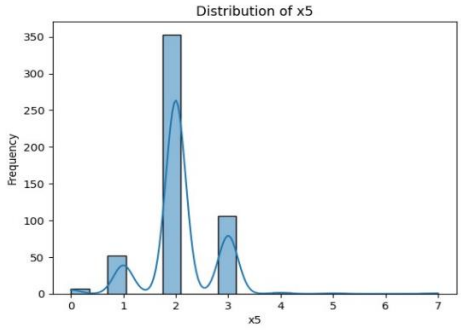


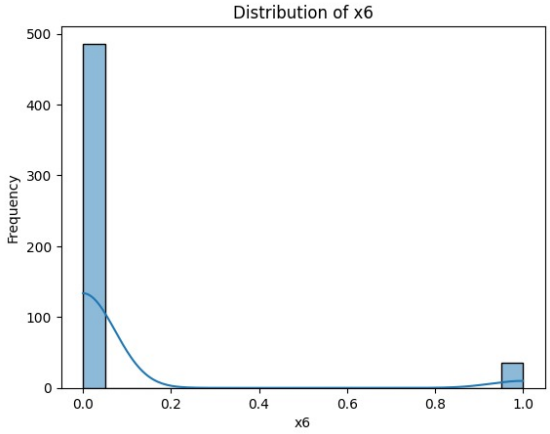


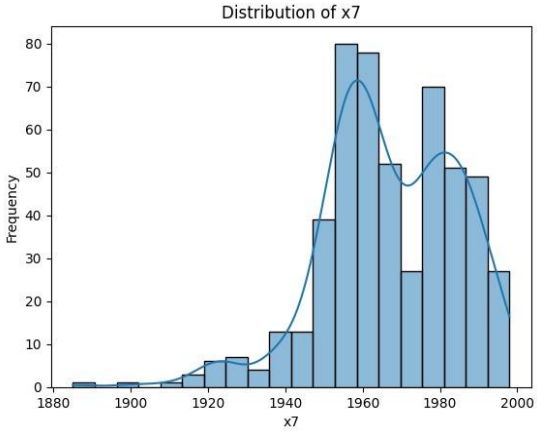


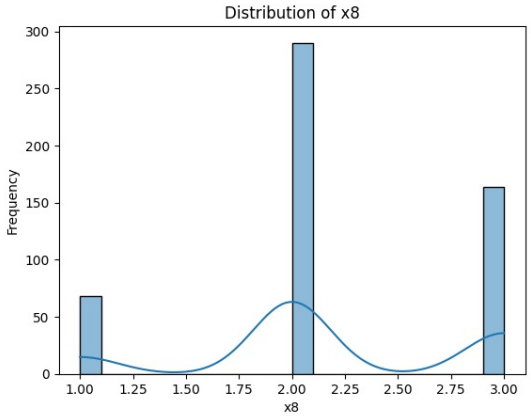


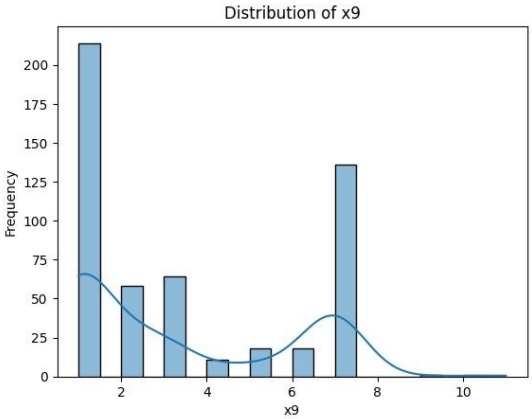


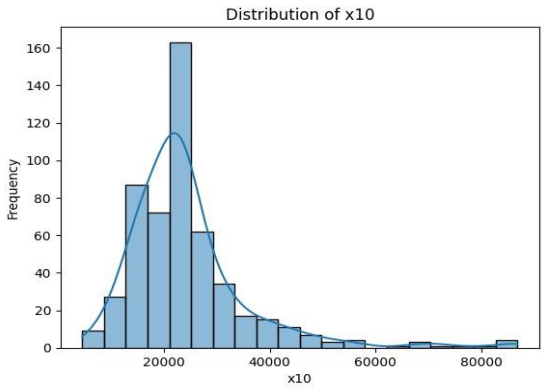


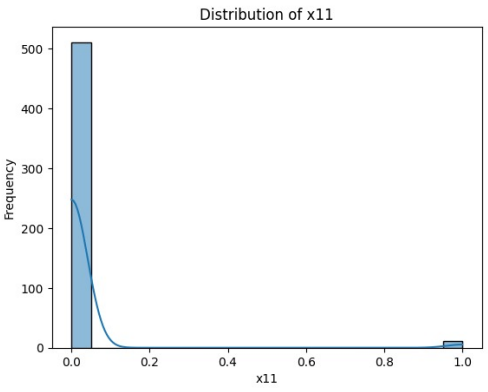

Distribution of x6
Frequency
x6


Distribution of x7
Frequency
x7


Distribution of x8
Frequency
x8


Distribution of x9
Frequency
x9


Distribution of x10
Frequency
x10


Distribution of x11
Frequency
x11

Numerical Data: y,x1,x7,x10

Numerical data typically shows a continuous distribution in the histogram, without clear separations between bars.The histogram have a smooth curve or shape, indicating the distribution of values across a range.There may be some variation in the height of bars, but overall the distribution appears continuous

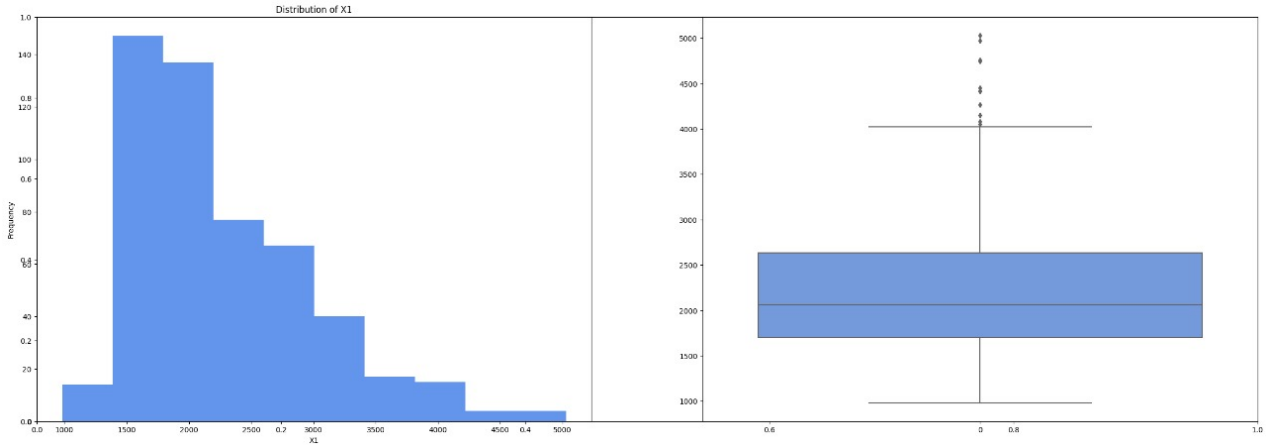
Categorical Data: x2,x3,x4,x5,x6,x8,x9,x11

Categorical data typically shows distinct bars or spikes in the histogram, indicating different categories or levels.

The histogram have a discrete distribution with clear separation between the bars. Each bar represents the frequency or count of observations in each category

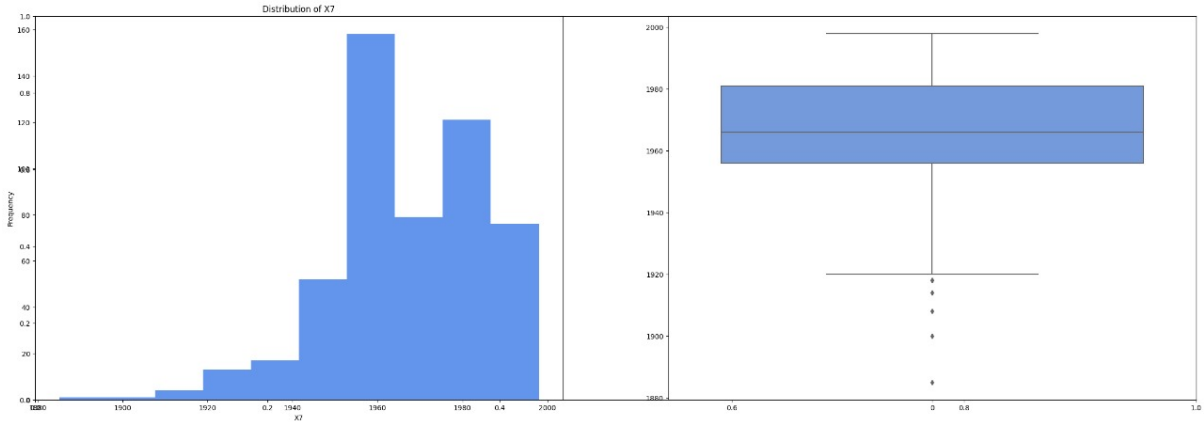
Numerical

X1



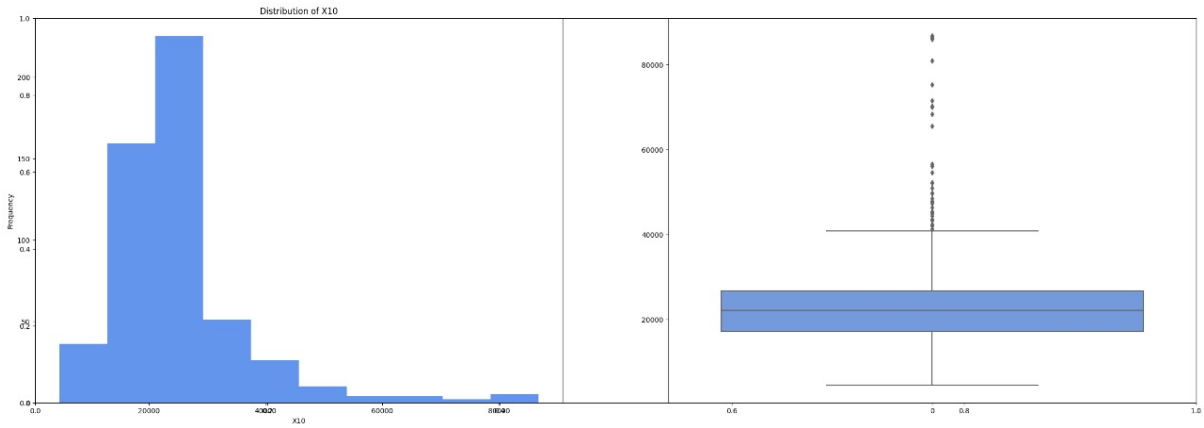
Variable x1 exhibits a right-skewed distribution as evidenced by the histogram. Additionally, the presence of outliers, as indicated by the box plot, suggests the existence of extreme values beyond the typical range of the data

X7



Variable x7 exhibits a left-skewed distribution, as evidenced by the histogram. Additionally, the presence of outliers on the lower end, as depicted in the box plot, suggests potential anomalies or extreme values in the dataset.

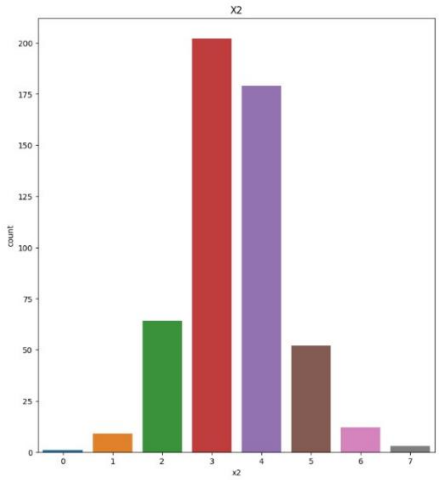
X10

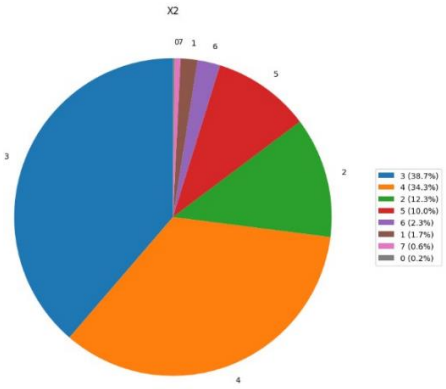


The histogram suggests a slight right skew, indicating that the majority of the data is concentrated towards the lower values but extends to higher values. Additionally, the box plot shows some outliers on the upper side, indicating potential extreme values beyond the typical range. This suggests that while most values cluster towards the lower end, there are some higher values that deviate significantly.

Categorical

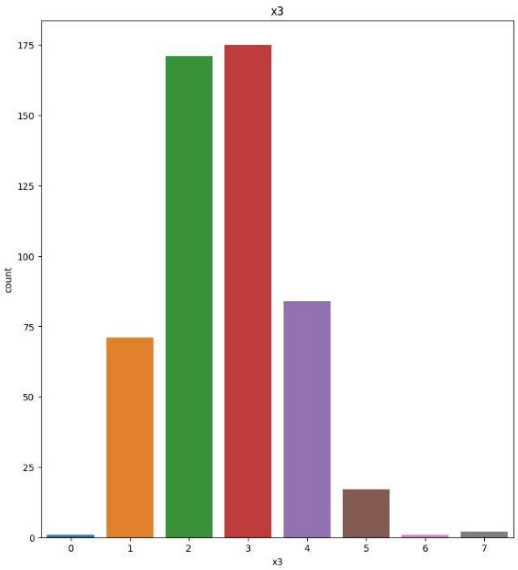
X2

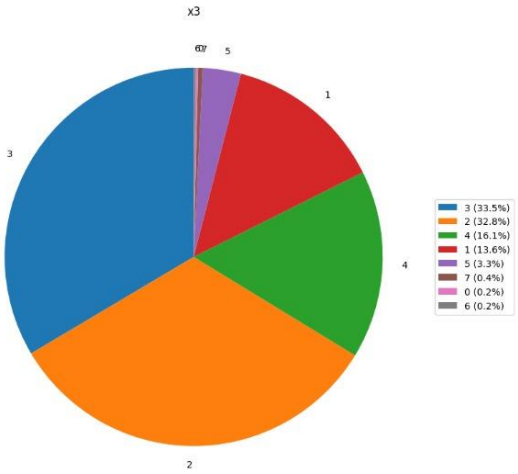




based on the countplot and pie chart, variable x2 exhibits a clear distribution with categories 3 and 4 being the most prevalent, followed by 2 and 5, and then 1 and 6. Categories 0 and 7 are the least frequent. Notably, categories 3 and 4 combined account for the majority of occurrences (73%), indicating a significant presence in the dataset.

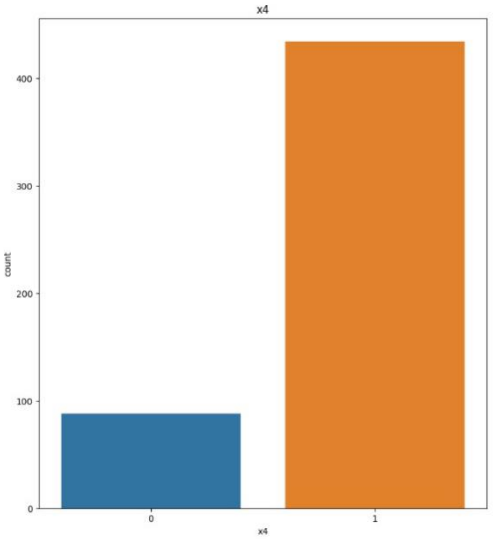
X3

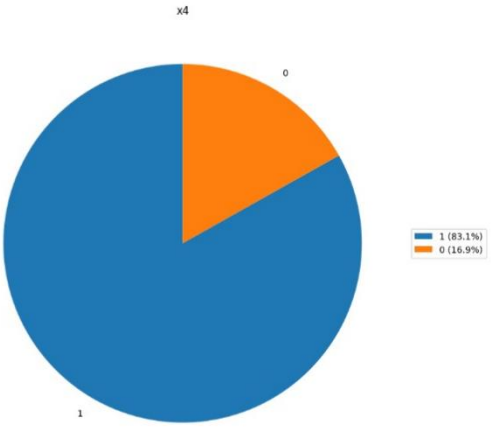




Based on the countplot and pie chart, variable x3 exhibits varied distribution among its categories. Categories 2 and 3 have the highest occurrences, representing 32.8% and 33.5% respectively. Categories 1 and 4 follow, with 13.6% and 16.1% respectively. Categories 0,6 and 7 have minimal representation, each below 1%. This suggests that x3 is not evenly distributed among its categories, with categories 2 and 3 being the most prevalent, while categories 0, 6, and 7 are rare.

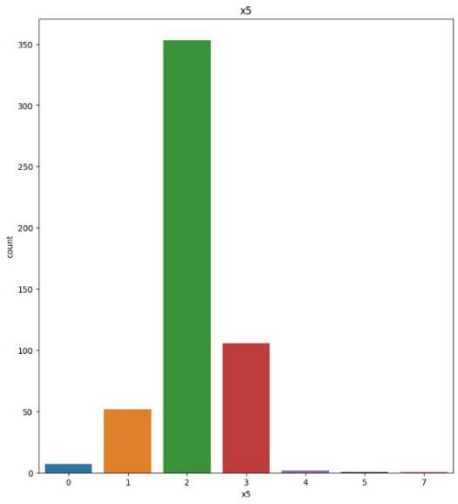
X4

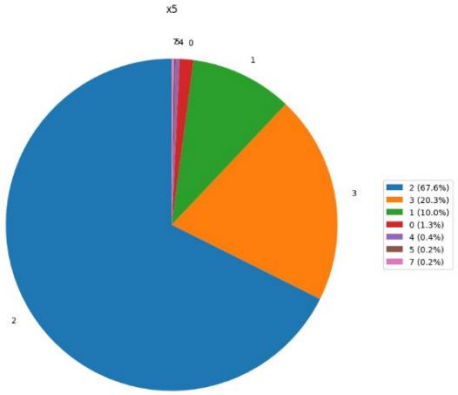




Based on the countplot and pie chart, variable x4 is heavily skewed towards category 1, which constitutes the majority at 83.1%. Category 0, on the other hand, is much less common, accounting for only 16.9% of the observations. This suggests that the majority of instances fall into category 1, indicating a significant imbalance in the distribution of this variable.

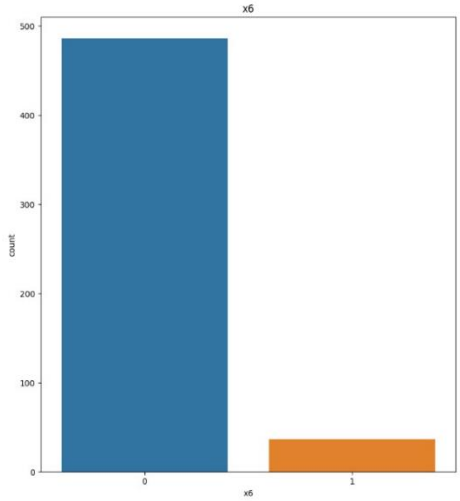
X5

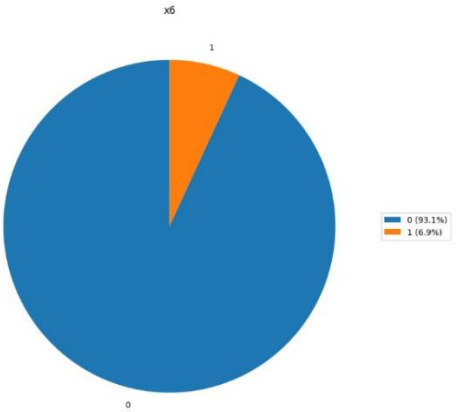




Based on the countplot and pie chart, it's evident that the variable x5 is predominantly composed of categories 2 and 3, which together account for the majority of occurrences (around 87.9%). Categories 0 and 1 make up a smaller portion, while categories 4, 5, and 7 are relatively rare. This suggests that x5 is skewed towards categories 2 and 3, with categories 4, 5, and 7 being outliers or uncommon occurrences in the dataset.

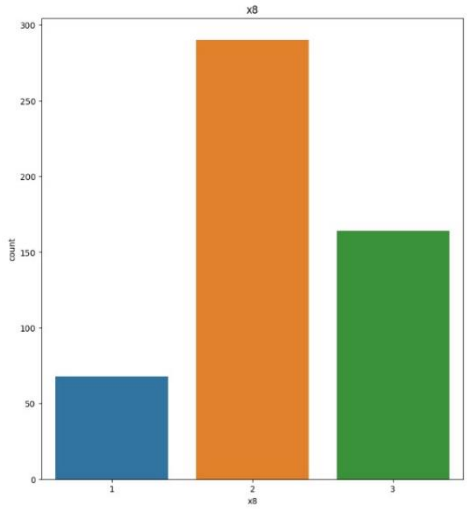
X6

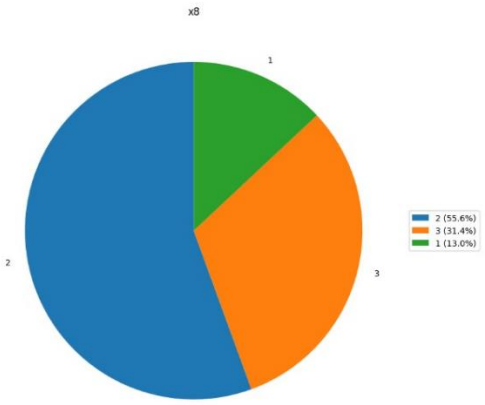




In interpreting the variable x6, it's evident that the majority (93.1%) falls into one category (0), while a small proportion (6.9%) falls into the other category. This suggests a significant class imbalance, with category 0 dominating the dataset. This insight could inform further analysis or modeling strategies to address the class imbalance issue.

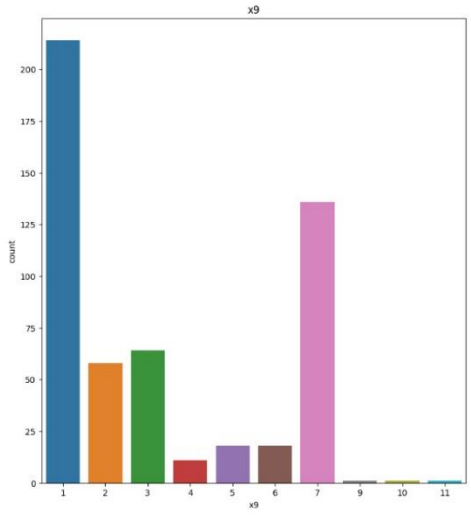
X8

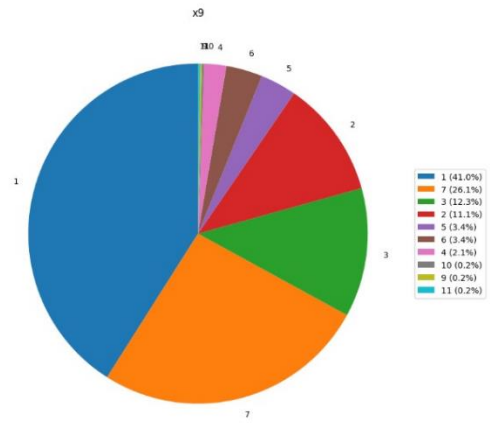




variable x8, it appears to be a categorical variable with three distinct values: 1, 2, and 3. The countplot and pie chart reveal that the majority of observations fall into categories 2 and 3, with category 2 being the most frequent at 55.6% and category 3 following at 31.4%. Category 1 is the least frequent at 13.0%. This suggests that categories 2 and 3 are more prevalent in the dataset compared to category 1.

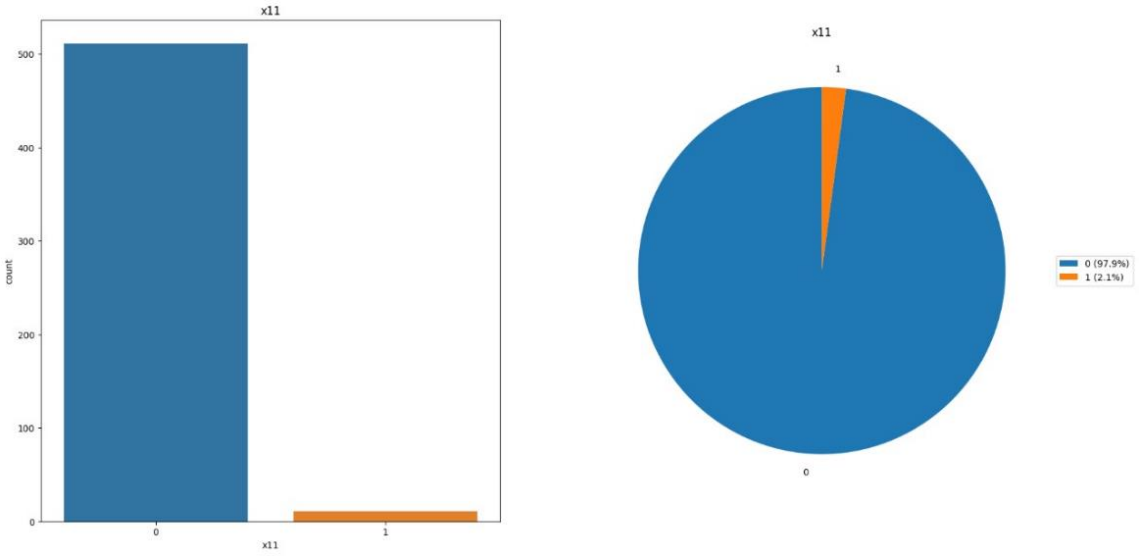
X9





Based on the countplot and pie chart, variable x9 exhibits uneven distribution, with the majority of occurrences found in categories 1, 7, 3, and 2, in descending order. Categories 5, 6, 4, 10, 9, and 11 are less prevalent, with 9 and 11 being the least frequent.

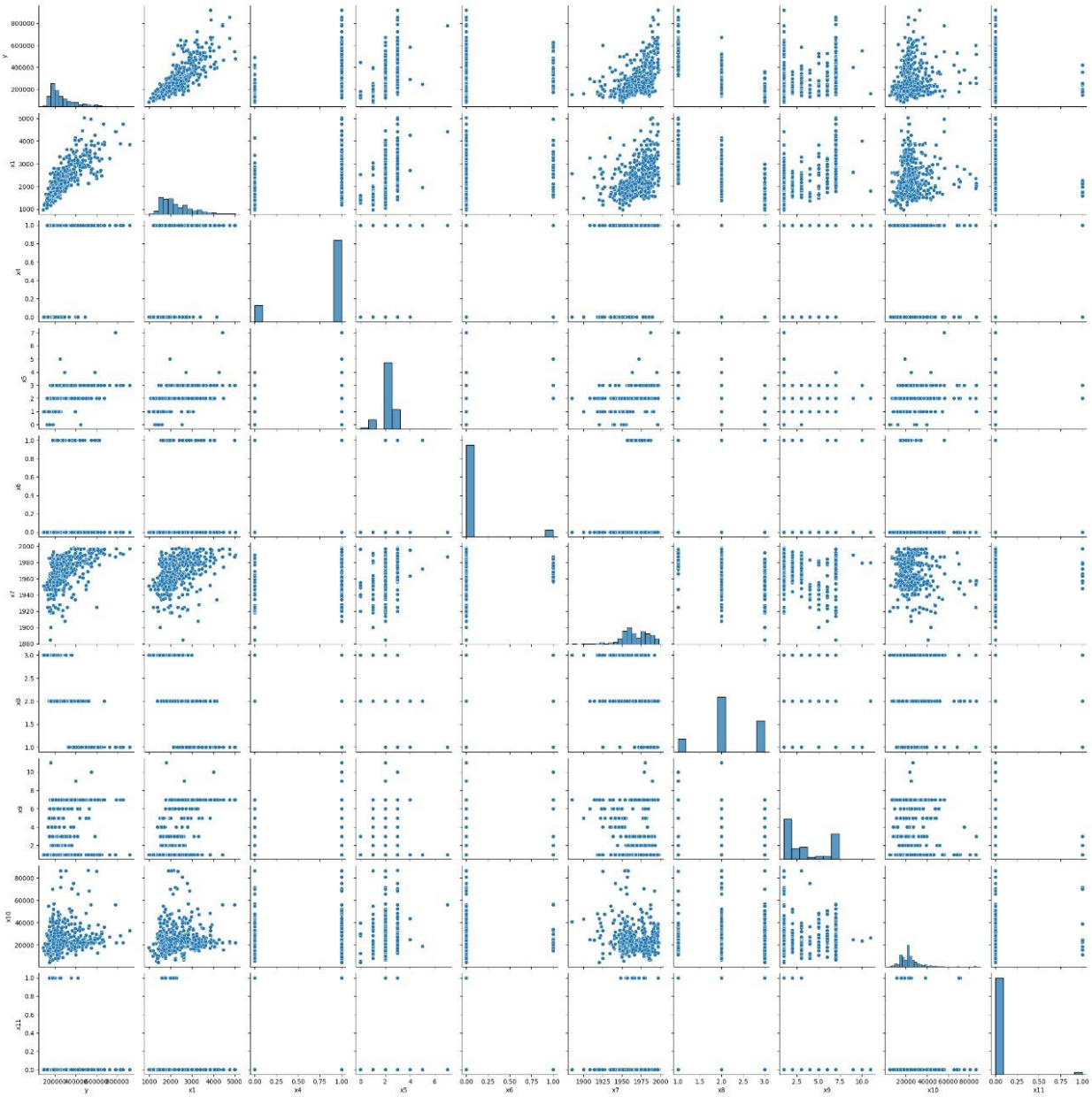
X11



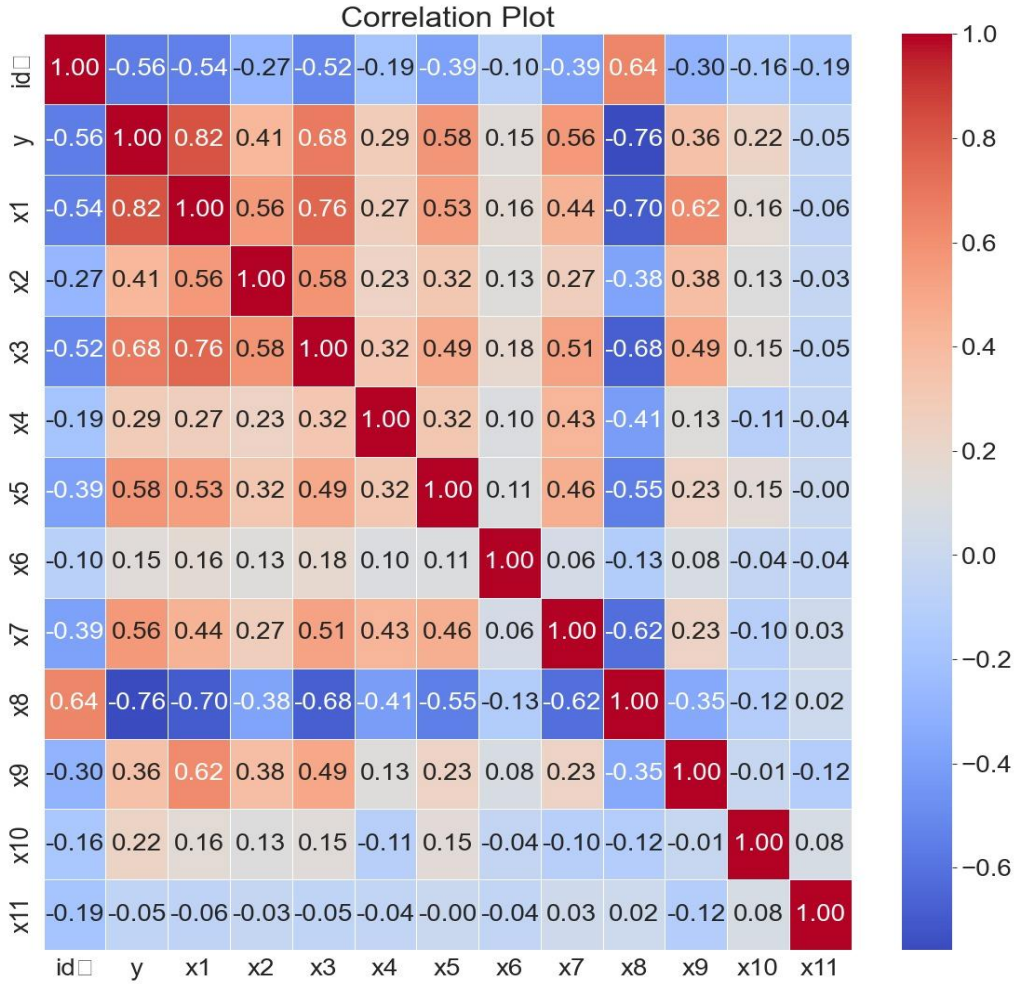
In interpreting variable x11, it's evident that the majority of the data falls into category 0, comprising approximately 97.9% of the observations, while category 1 represents a much smaller proportion, around 2.1%.

Bivariate Data Analysis

- **Pairplot:**



From the correlation matrix provided, we can draw several inferences:



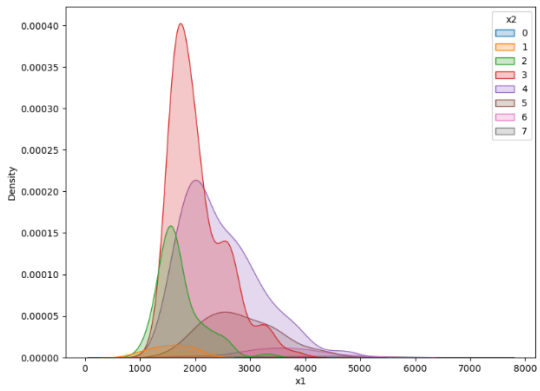
- Strong Positive Correlations: The target variable y shows strong positive correlations with variables x1, x3, x5, x7, and x9. This indicates that as the values of these variables increase, the value of y tends to increase as well.
- Strong Negative Correlations: There is a strong negative correlation between y and x8, indicating that as the value of x8 increases, the value of y tends to decrease significantly.
- Moderate Correlations: Variables x2, x4, and x10 exhibit moderate positive correlations with y, suggesting that they have some influence on y but not as strong as the variables mentioned above.
- Weak Correlations: There are weak correlations between y and x6, and y and x11. This suggests that these variables have minimal influence on y.Correlation between Explanatory Variables: Some explanatory variables show correlations with each other. For example, x1 and x3 have a strong positive correlation, indicating that they might be measuring related aspects of the data. Similarly, x7 and x8 exhibit a moderate negative correlation, suggesting that they might be inversely related.

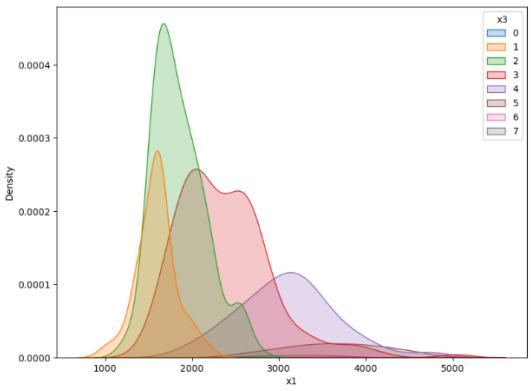
- No Linear Relationship: It's important to note that correlation measures linear relationships between variables. So, if there are non-linear relationships, they might not be captured by correlation coefficients.These inferences provide insights into the relationships between the variables in the dataset, which can be valuable for further analysis and modeling. However, it's essential to remember that correlation does not imply causation, and further analysis may be required to establish causal relationships.

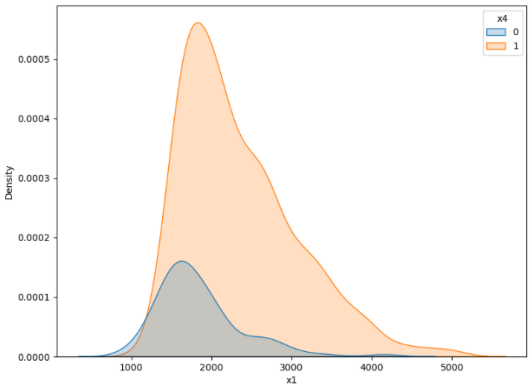
Bivariate Analysis of Continuous vs Categorical Variables

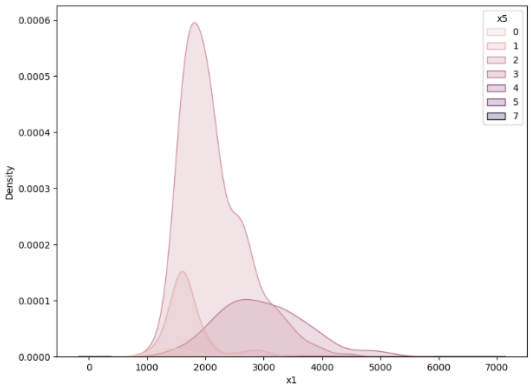
A density plot visually represents the distribution of a continuous variable, estimating the probability density function. When categories are involved, it displays the proportion of each category's data within intervals of the continuous variable. This allows for easy comparison of distributions across different groups in the data, aiding in identifying patterns and differences.

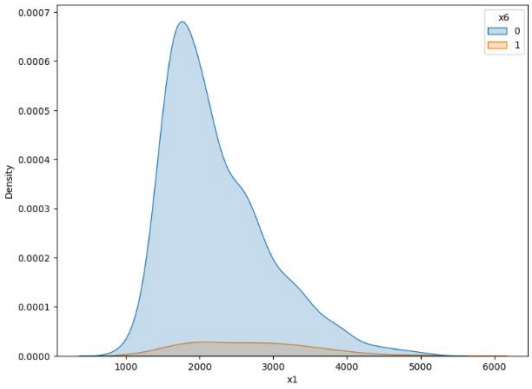
X1

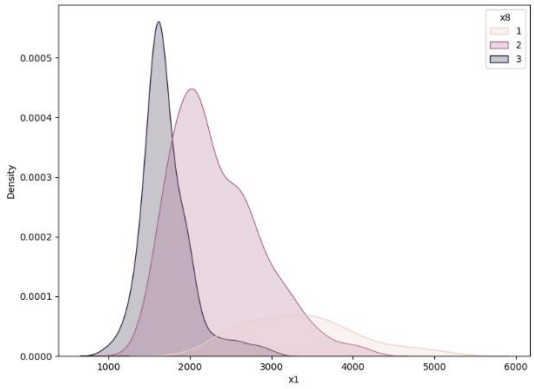

x2
0
1
2
3
4
5
6
7
Density
x1

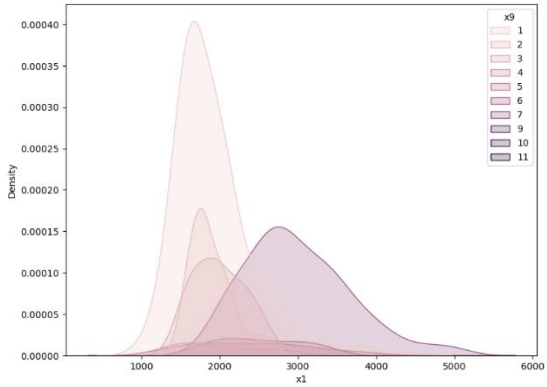

x3
0
1
2
3
4
5
6
7
Density
x1

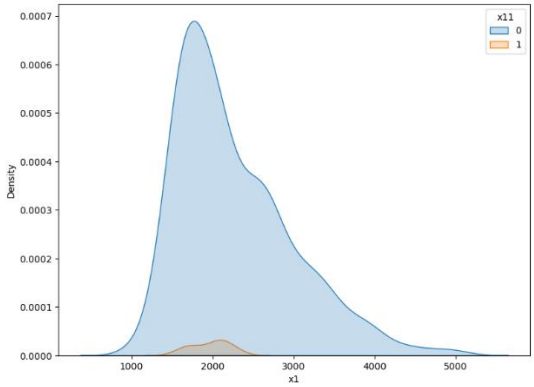

x4
0
1
Density
x1


x5
0
1
2
3
4
5
7
Density
x1

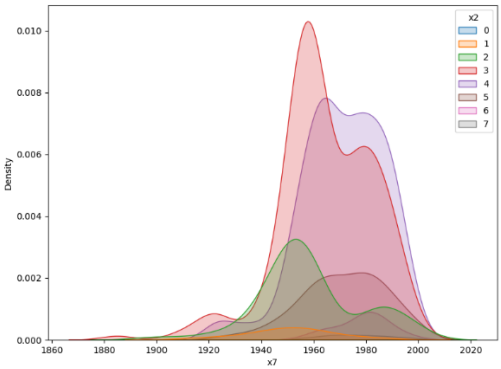

x6
0
1
Density
x1

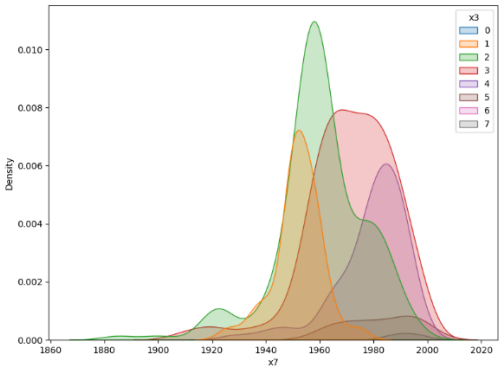

x8
1
2
3
Density
x1

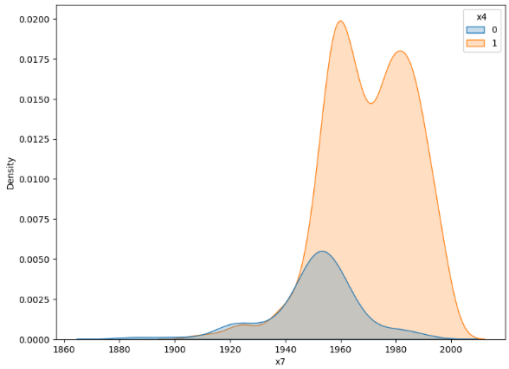

x9
1
2
3
4
5
6
7
9
10
11
Density
x1

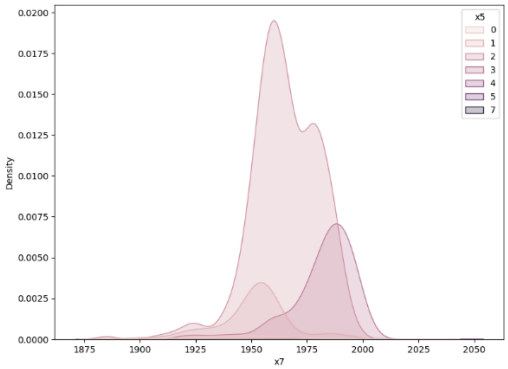

x11
0
1
Density
x1

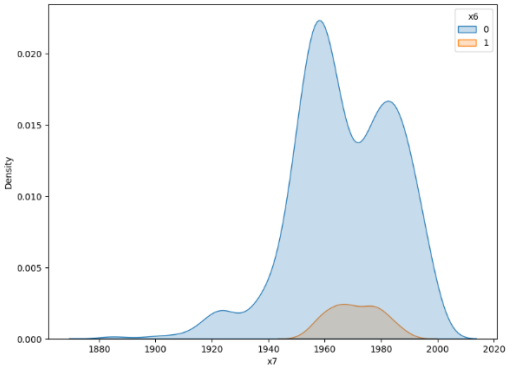
X7

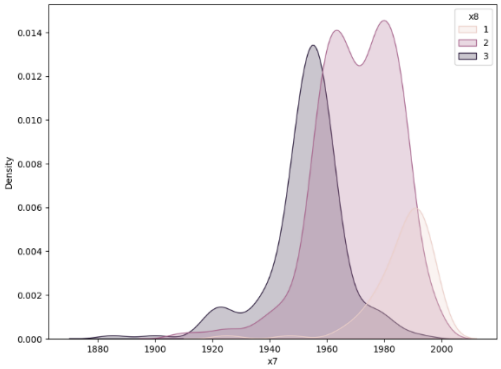

x2
0
1
2
3
4
5
6
7
Density
x7

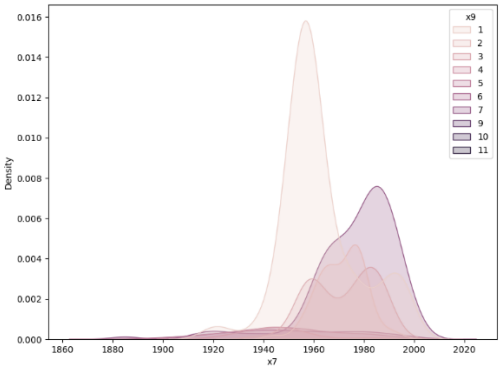

x3
0
1
2
3
4
5
6
7
Density
x7

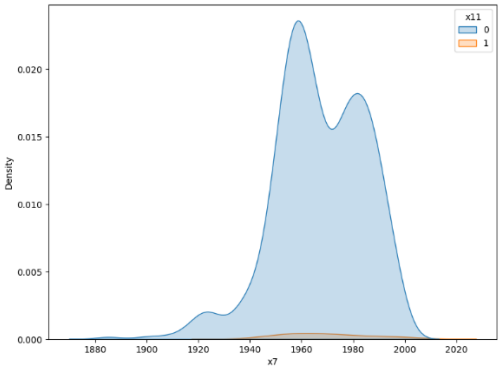

x4
0
1
Density
x7


x5
0
1
2
3
4
5
7
Density
x7

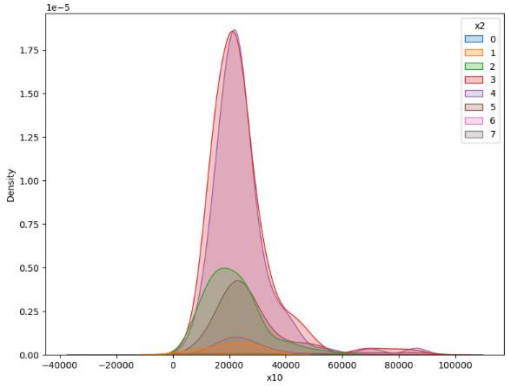

x6
0
1
Density
x7

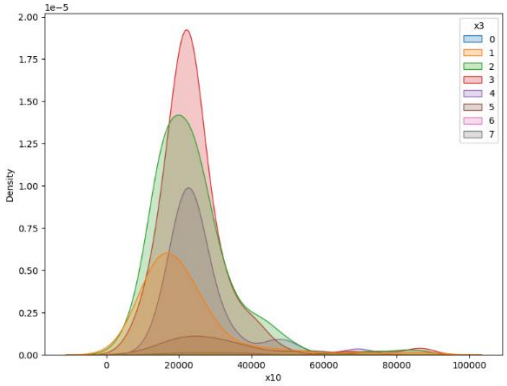

x8
1
2
3
Density
x7

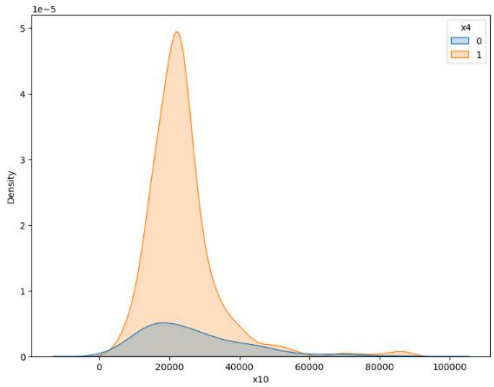

x9
1
2
3
4
5
6
7
9
10
11
Density
x7

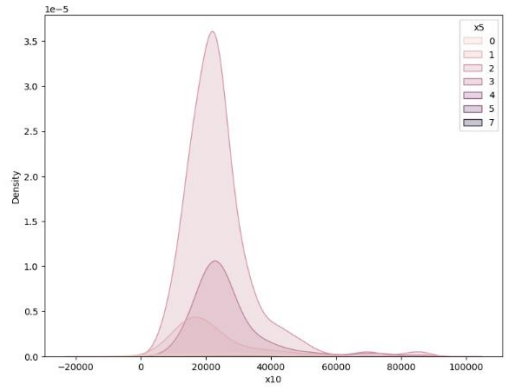

x11
0
1
Density
x7

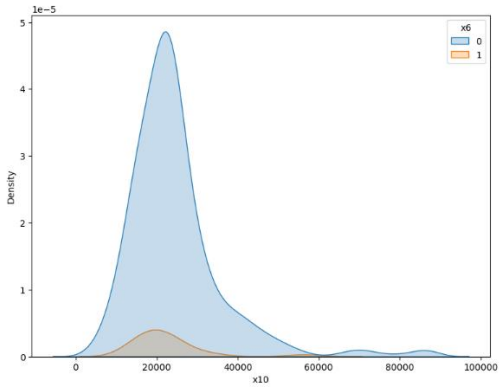
X10

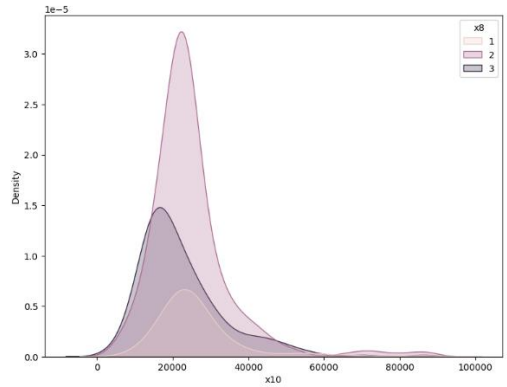

1e−5
x2
0
1
2
3
4
5
6
7
Density
x10

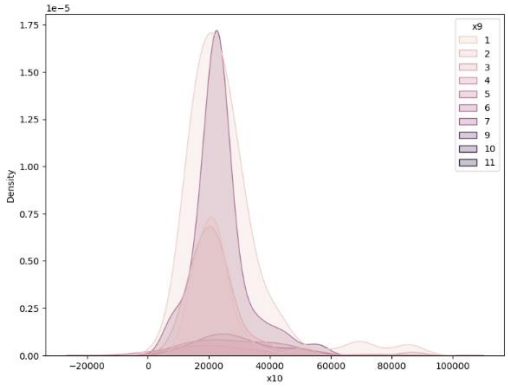

1e−5
x3
0
1
2
3
4
5
6
7
Density
x10

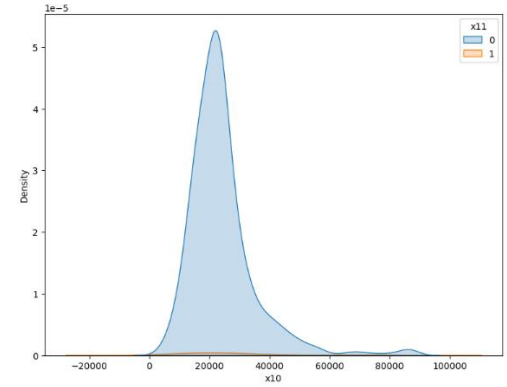

1e−5
x4
0
1
Density
x10


1e−5
x5
0
1
2
3
4
5
7
Density
x10


1e−5
x6
0
1
Density
x10


1e−5
x8
1
2
3
Density
x10


1e−5
x9
1
2
3
4
5
6
7
9
10
11
Density
x10


1e−5
x11
0
1
Density
x10

CHAPTER 3

Univariate Linear Regression

Y vs x1

OLS Regression Results

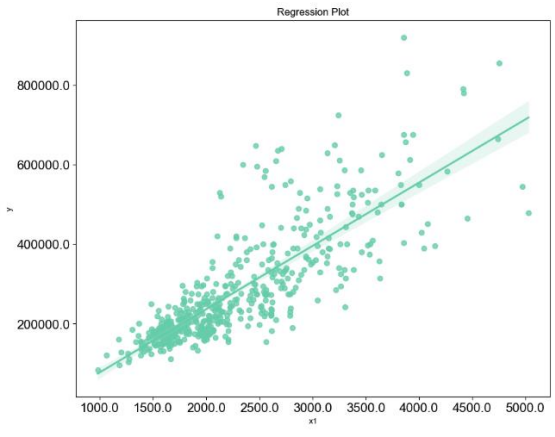
Dep. Variable:	y	R-squared:	0.672
Model:	OLS	Adj. R-squared:	0.671
Method:	Least Squares	F-statistic:	1063.
Date:	Fri, 26 Apr 2024	Prob (F-statistic):	8.28e-128
Time:	11:08:07	Log-Likelihood:	-6627.2
No. Observations:	522	AIC:	1.326e+04
Df Residuals:	520	BIC:	1.327e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8.143e+04	1.16e+04	-7.049	0.000	-1.04e+05	-5.87e+04
x1	158.9502	4.875	32.605	0.000	149.373	168.527

Omnibus:	145.431	Durbin-Watson:	1.358
Prob(Omnibus):	0.000	Jarque-Bera (JB):	499.032
Skew:	1.267	Prob(JB):	4.33e-109
Kurtosis:	7.065	Cond. No.	7.90e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 7.9e+03. This might indicate that there are strong multicollinearity or other numerical problems.

An R-squared of 0.672 indicates 67.2% of y's variability is explained by x1. The coefficient for x1 (158.9502) suggests its impact on y. Both constant and x1 coefficients are significant ($p < 0.05$), indicating a relationship. TheF-statistic (1063.0) and p-value (8.28e-128) confirm overall significance.



quite linear but has a megaphone-like structure

Y vs x7

OLS Regression Results

Dep. Variable:	y	R-squared:	0.309
Model:	OLS	Adj. R-squared:	0.307
Method:	Least Squares	F-statistic:	232.1
Date:	Fri, 26 Apr 2024	Prob (F-statistic):	1.34e-43
Time:	11:08:08	Log-Likelihood:	-6821.5
No. Observations:	522	AIC:	1.365e+04
Df Residuals:	520	BIC:	1.366e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8.266e+06	5.61e+05	-14.739	0.000	-9.37e+06	-7.16e+06
x7	4343.9758	285.137	15.235	0.000	3783.813	4904.138

Omnibus:	131.152	Durbin-Watson:	0.931
Prob(Omnibus):	0.000	Jarque-Bera (JB):	276.537
Skew:	1.350	Prob(JB):	8.93e-61
Kurtosis:	5.328	Cond. No.	2.20e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.2e+05. This might indicate that there are strong multicollinearity or other numerical problems.

An R-squared of 0.309 indicates 30.9% of y's variability is explained by x7. The coefficient for x7 (4343.9758) suggests its impact on y. Both constant and x7 coefficients are significant ($p < 0.05$), indicating a relationship. The F-statistic (232.1) and p-value (1.34e-43) confirm overall significance.

Quite linear but not that much

Y vs x10

OLS Regression Results

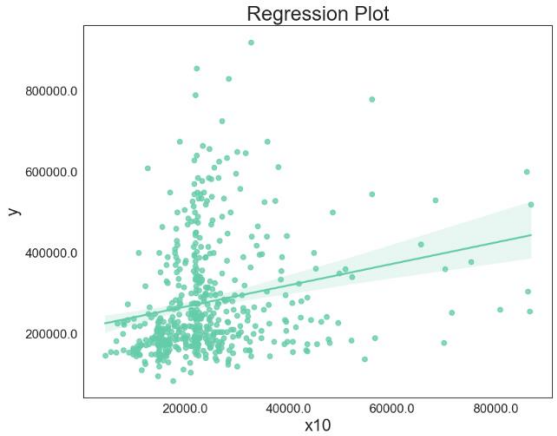
Dep. Variable:	y	R-squared:	0.050
Model:	OLS	Adj. R-squared:	0.048
Method:	Least Squares	F-statistic:	27.51
Date:	Fri, 26 Apr 2024	Prob (F-statistic):	2.27e-07
Time:	11:08:10	Log-Likelihood:	-6904.3
No. Observations:	522	AIC:	1.381e+04
Df Residuals:	520	BIC:	1.382e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.134e+05	1.36e+04	15.655	0.000	1.87e+05	2.4e+05
x10	2.6462	0.504	5.245	0.000	1.655	3.637

Omnibus:	151.365	Durbin-Watson:	0.385
Prob(Omnibus):	0.000	Jarque-Bera (JB):	339.761
Skew:	1.531	Prob(JB):	1.67e-74
Kurtosis:	5.500	Cond. No.	6.26e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 6.26e+04. This might indicate that there are strong multicollinearity or other numerical problems.

This ANOVA table reflects a univariate linear regression. The R-squared of 0.050 shows that only 5% of y's variability is explained by x10. The coefficient for x10 (2.6462) suggests its effect on y. Both constant and x10 coefficients are significant ($p < 0.05$), indicating a relationship. The F-statistic (27.51) and p-value (2.27e-07) confirm overall significance.



Not linear

CHAPTER 4

Preprocessing of the Data

- No missing values.
- Dropping id/t as it was irrelevant.
- One-hot encoding: Convert categorical variables into binary vectors where each category is represented by a binary feature.
- **Grouping:** When dealing with categorical columns containing a large number of categories, grouping can help manage the dimensionality of the data and improve model performance

  Grouped categories based on their frequency or occurrence in the dataset. Combine infrequent or rare categories into a single group to reduce noise and improve model generalization**.**
- One Hot Encoding of a categorical variable: Notably, we didn't need to create dummy variables for X4, X6, as they were already in binary form.

  The variable X8 is categorical and it takes values 1,2,3 . We are using dummy variables for X8 and dropped the first dummy variable to prevent multicollinearity.

	id\t	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
0	1	360000	3032	4	4	1	2	0	1972	2	1	22221	0
1	2	340000	2058	4	2	1	2	0	1976	2	1	22912	0
2	3	250000	1780	4	3	1	2	0	1980	2	1	21345	0
3	4	205500	1638	4	2	1	2	0	1963	2	1	17342	0
4	5	275500	2196	4	3	1	2	0	1968	2	7	21786	0

	y	x1	x2	x3	x4	x5	x6	x7	x9	x10	x11	x8_2	x8_3
0	360000	3032	0	1	1	1	0	1972	1	22221	0	1	0
1	340000	2058	0	0	1	1	0	1976	1	22912	0	1	0
2	250000	1780	0	1	1	1	0	1980	1	21345	0	1	0
3	205500	1638	0	0	1	1	0	1963	1	17342	0	1	0
4	275500	2196	0	1	1	1	0	1968	0	21786	0	1	0

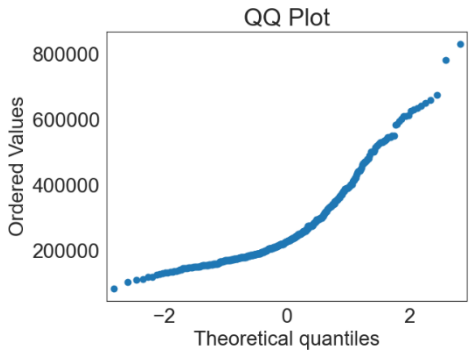
- Train-test split: Train Test and Split: Now, we split data into train data and test data for model validation. We split the data into 20% test data and 80% train data denoted by X_test and X_train. We will use this training data to build the model and then for validation we will use the test data. Now, we will move on to the Model selection

CHAPTER 5

Multivariate Linear Regression

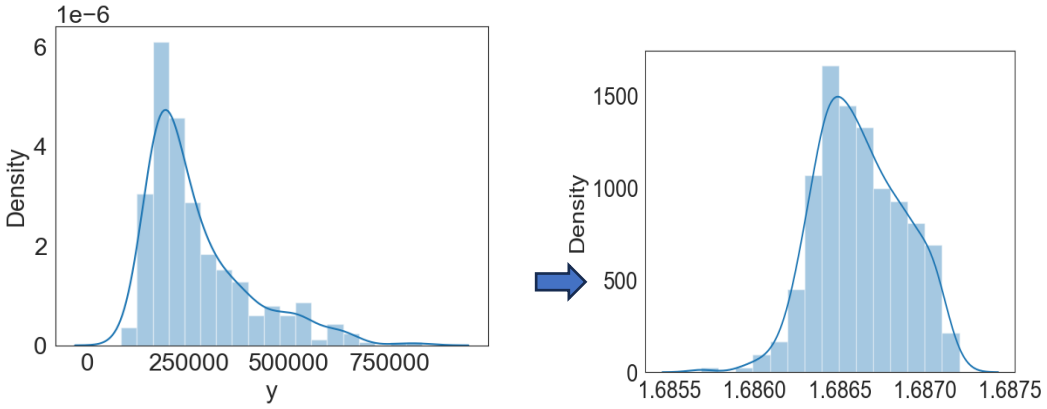
Assumptions Checking

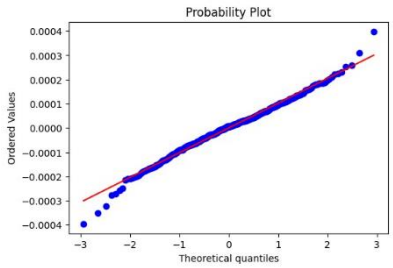
NORMALITY: the quantiles of the residuals are plotted against the quantiles of a theoretical normal distribution. If the residuals are normally distributed, the points on the plot will fall approximately along a straight line.



Here, not normal

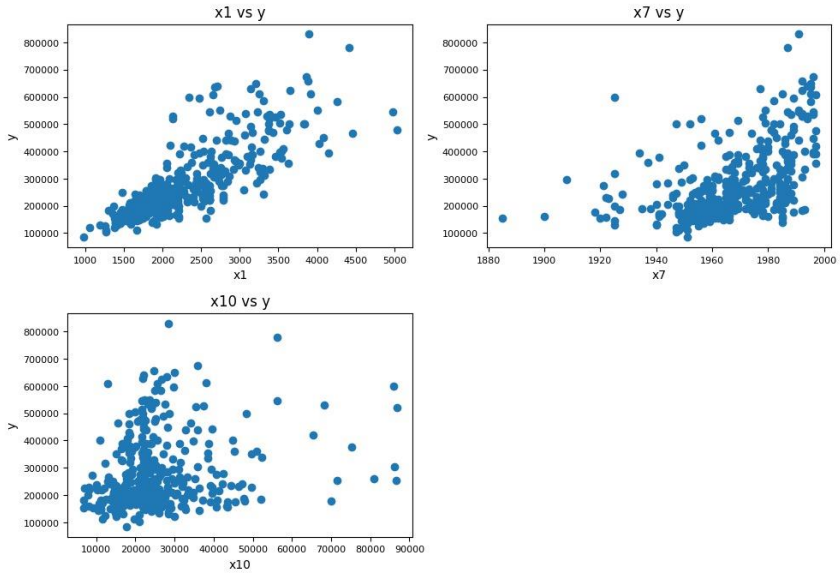
REMEDY USED: Box-cox transformation: The Box-Cox transformation is a statistical method used to stabilize the variance and make the data more normally distributed. It involves raising the variable to a power, typically denoted by lambda (λ). This transformation can be beneficial for improving the performance of statistical models, particularly when the data violate assumptions of normality and constant variance.



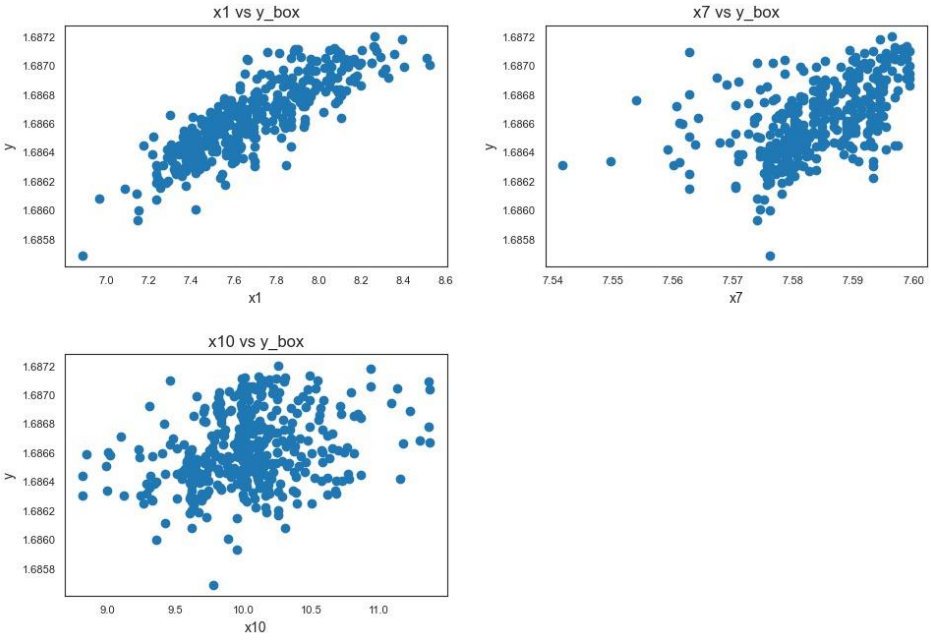


After the Box-Cox transformation, the histogram shows a slight improvement towards normality, but a slight rightward shift suggests some non-normality, possibly due to outliers or other factors influencing the upper end of the distribution.

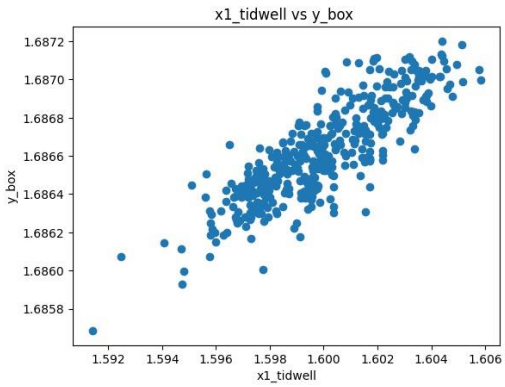
- **Linearity:** We will check Linearity of numerical columns with y and then with y transformed by box-cox

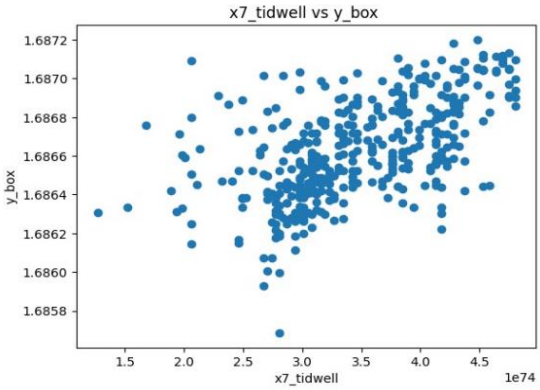


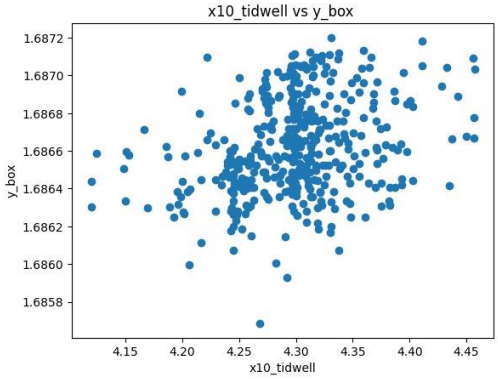
since, there is no linearity between numerical x's and y do, We need to transform x using Box Tidwell Test.



REMEDY USED: Box-Tidwell transformation: The Box-Tidwell transformation extends the Box-Cox transformation to handle situations where there is a linear relationship between the predictors (independent variables). It involves transforming the predictors to improve linearity and address issues like heteroscedasticity and nonlinearity.



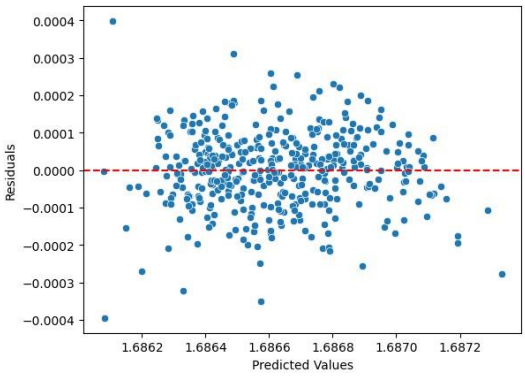




After transformation, X1 and x7 got linear while not that much change in x10.

It's possible that the relationship between our response variable y and x10 is inherently nonlinear. In such cases, a simple transformation may not adequately address the nonlinearity.

- **Homoscedasticity:** It refers to the assumption that the variance of the errors (or residuals) in a regression model is constant across all levels of the independent variables. In simpler terms, it means that the spread of the residuals is consistent across the range of predicted values.



Our assumption of homoscedasticity gets satisfied.

- **Autocorrelation of Residuals:**

autocorrelation of residuals violates the assumption of independence of observations, which is essential for accurate parameter estimation and hypothesis testing.

The **Durbin-Watson test** is a statistical test used to detect the presence of autocorrelation in the residuals of a regression analysis.

A Durbin-Watson test statistic value of 1.98 is generally considered to be close to 2, which is often interpreted as indicating little to no autocorrelation in the residuals

- **Multicollinearity:**

Multicollinearity in regression analysis arises when two or more independent variables are so closely intertwined that they essentially convey redundant information, complicating the interpretation and reliability of the model's predictions. Hence, The assumption should be checked for no multicollinearity.

The Variance Inflation Factor (VIF) measures the extent of multicollinearity among the independent variables in a regression model. A high VIF value indicates that an independent variable is highly correlated with other independent variables, potentially leading to unreliable coefficient estimates and inflated standard errors.

x1: A VIF of 2032.67 indicates extremely high multicollinearity with other independent variables. This suggests that x1 is highly correlated with other predictors in the model.

x2: A VIF of 2.60 suggests low to moderate multicollinearity with other independent variables.

x3: A VIF of 4.75 suggests moderate multicollinearity with other independent variables.

x4: A VIF of 8.17 indicates relatively high multicollinearity with other independent variables.

x5: A VIF of 4.07 suggests moderate multicollinearity with other independent variables.

x6: A VIF of 1.13 indicates low multicollinearity with other independent variables.

x7: A VIF of 2446.42 indicates extremely high multicollinearity with independent variables.

x9: A VIF of 3.50 suggests moderate multicollinearity with other independent variables.

x10: A VIF of 636.61 indicates extremely high multicollinearity with independent variables.

x11: A VIF of 1.05 indicates low multicollinearity with other independent variables.

x8_2: A VIF of 8.74 indicates relatively high multicollinearity with other independent variables.
x8_3: A VIF of 8.59 indicates relatively high multicollinearity with other independent variables
Interpretation: Independent variables are highly multicorrelated. We will try to remove this after the best subset model selection

CHAPTER 6

Model Selection

We are selecting subsets with the following metrics:
1) Adjusted R square or MSE (mean squared error)
2) Mallows Cp
3) Akaike's Information Criterion (AICp)
4) Bayesian Information Criterion (BICp/SBCp)
We calculate the above metrics for all the possible combinations of Variables, since our total number of explanatory variables after one hot encoding is 8 so the total possible combinations are 2^12 = 4096

	features	p	SSEp	R_squared	Adj_R_squared	AIC	BIC	Cp
0	(x1,)	1	0.000007	0.725114	0.724452	-6253.626968	-6245.560796	300.205007
1	(x2,)	1	0.000024	0.112804	0.110666	-5765.023449	-5756.957276	1893.331624
2	(x3,)	1	0.000014	0.481240	0.479990	-5988.795872	-5980.729700	934.724512
3	(x4,)	1	0.000022	0.180216	0.178241	-5797.976976	-5789.910804	1717.936732
4	(x5,)	1	0.000025	0.051055	0.048769	-5736.965773	-5728.899600	2053.991131
...	...	...	...	...	...	...	...	...
4090	(x1, x2, x3, x5, x6, x7, x9, x10, x11, x8_2, x...	11	0.000004	0.835375	0.830903	-6447.415865	-6399.018831	23.005489
4091	(x1, x2, x4, x5, x6, x7, x9, x10, x11, x8_2, x...	11	0.000004	0.829971	0.825353	-6433.947022	-6385.549987	36.727218
4092	(x1, x3, x4, x5, x6, x7, x9, x10, x11, x8_2, x...	11	0.000004	0.839735	0.835382	-6458.608153	-6410.211118	11.935439
4093	(x2, x3, x4, x5, x6, x7, x9, x10, x11, x8_2, x...	11	0.000007	0.731928	0.724647	-6244.092791	-6195.695756	285.671543
4094	(x1, x2, x3, x4, x5, x6, x7, x9, x10, x11, x8_...	12	0.000004	0.840497	0.835759	-6458.596192	-6406.166071	11.000000

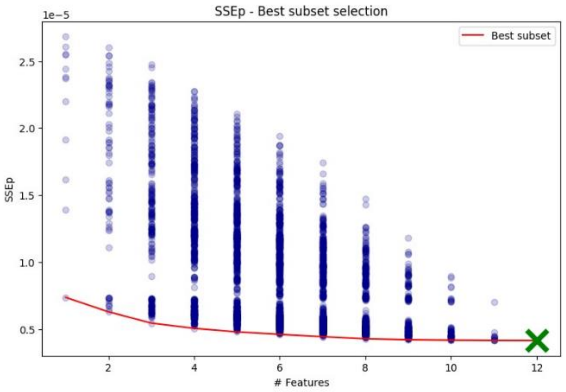
The above table shows the few rows of the data frame which store the values of 5 metrics corresponding to each possible combination. Here p denotes the number of variables in the model. We will use each criterion to select a subset.

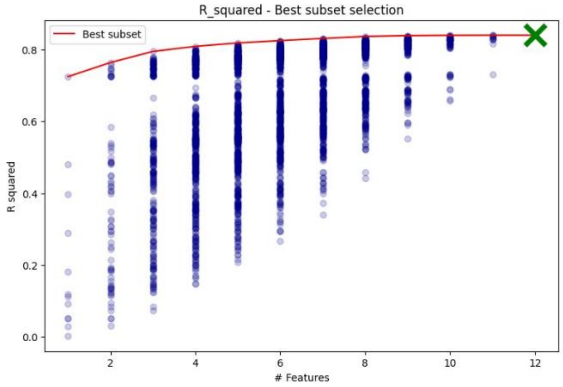
After selecting a subset based on each criterion we will use stepwise regression to get another subset.

From the obtained 5 subsets we will select the best subset to build our model. There is no hard and fast rule to select the best subset, we will observe the subsets and based on our observations we will select on

Best Subset from R square / SSEp

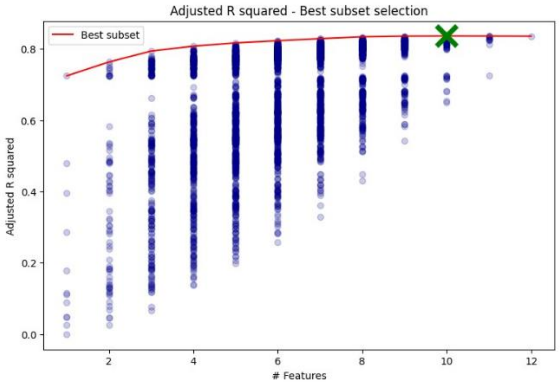
The R square criterion is equivalent to using the error sum of squares SSE, R square (SSEp) can never decrease(increase) as additional X variables are included in the model. The intent in using the R square criterion is to find the point where adding more X variables is not worthwhile because it leads to a very small increase in R square





Best Subset from Adjusted R Square

For using the adjusted R square criterion, we seek to find a subset for which the adjusted R square is at the maximum so that adding more variables is not worthwhile.



+8

The red line in the above figure shows the maximum value corresponding to each p, and the cross on the line is the maximum value of the adjusted R square.

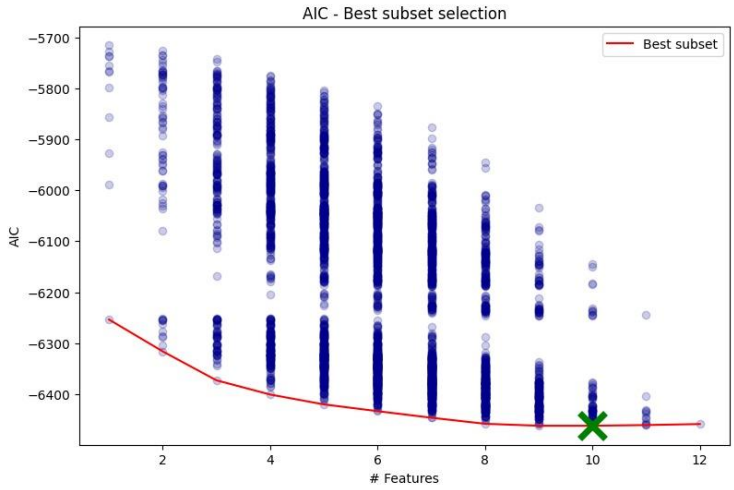
The maximum value of the adjusted R square is at the value corresponding to p=10. The best subset obtained by the adjusted R square criterion is (x4, x7, x8, x9, x10, x11, x13, x14, x15_4). So, this can be one of the possible sets of explanatory variables that fits an optimal model to our dataset. Now, we look for the other possible model with the other subset selection criterion.

Best Subset from AICp

We search for models that have small values of AICp where these criteria are given:

$$AICp = n * ln(SSEp) - n * ln(n) + 2p$$

An AIC score is a number used to determine which variables can be used to get the optimal fit for the model. AIC is most useful when we are working with a small data set. The lower the AIC score the better it is for the subset selection



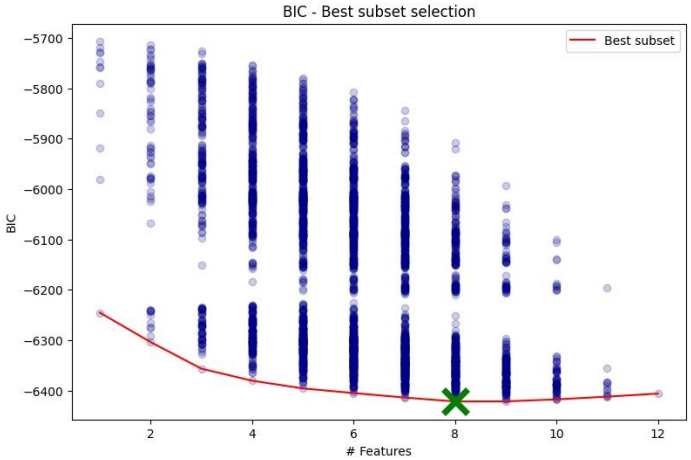
The red line in the above figure shows the minimum value corresponding to each p, and the cross on the line is the minimum value of the AICp.

Best Subset from BICp

We search for models that have small values of BICp where these criteria are given:

$$BICp = n * ln(SSEp) - n * ln(n) + 2p$$

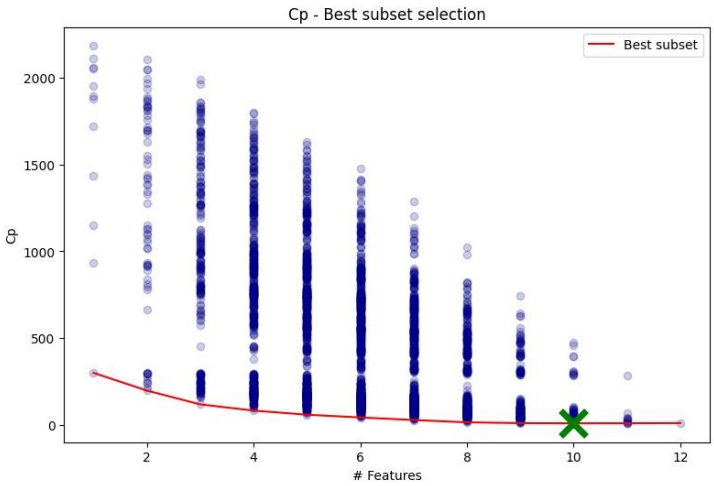
A BIC score is a number used to determine which variables can be used to get the optimal fit for the model. The lower the BIC score the better it is for the subset selection.



The red line in the above figure shows the minimum value corresponding to each p, and the cross on the line is the minimum value of the BICp.

Best Subset from Mallows Cp

In using the Cp criterion, we seek to identify subsets of X variables for which the Cp value is small and the Cp value is near p. Subsets with small Cp values have a small total mean squared error, and when the Cp value is also near p, the bias of the regression model is small.



The red line in the above figure shows the minimum value corresponding to each p, and the cross on the line is the minimum value of the Cp

Even though computerized approaches can be very helpful in identifying appropriate subsets for detailed, final consideration, the process of developing a useful regression model must be pragmatic and needs to utilize large doses of subjective judgment. Building the Regression Modell: Model Selection and Validation variables that are considered essential should be included in the regression model before any computerized assistance is sought. Further, computerized approaches that identify only a single subset of explanatory variables as ''best'' need to be supplemented so that additional subsets are also considered before the final regression model is decided upon.

Result of Subset selections by various criterion:

	Parameter	Value	p	Features
0	R_squared	0.840497	12	(x1, x2, x3, x4, x5, x6, x7, x9, x10, x11, x8_...
1	Adj_R_squared	0.836324	10	(x1, x2, x3, x4, x6, x7, x9, x10, x8_2, x8_3)
2	AIC	-6461.972660	10	(x1, x2, x3, x4, x6, x7, x9, x10, x8_2, x8_3)
3	BIC	-6421.772628	8	(x1, x3, x4, x7, x9, x10, x8_2, x8_3)
4	Cp	9.607538	10	(x1, x2, x3, x4, x6, x7, x9, x10, x8_2, x8_3)

We go with the subsets: (x1, x2, x3, x4, x6, x7, x9, x10, x8_2, x8_3) we can see that at p=10, Cp and AIC is minimum and R_adjusted is maximum and it is around 10.

Best Subset :(x1, x2, x3, x4, x6, x7, x9, x10, x8_2, x8_3).

Another good subsets: 1) (x1, x3, x4, x7, x9, x10, x8_2, x8_3) corresponding p=8 and using BIC.
2)the full model using R Square.

Best Subset from Stepwise Regression

The forward stepwise regression procedure is probably the most widely used of the automatic search methods. It was developed to economize on computational efforts. as compared with the various all-possible-regressions procedures. Essentially, this search method develops a sequence of regression models, at each step adding. or deleting an X variable. The criterion for adding or deleting an X variable can be stated equivalently in terms of error sum of squares reduction, coefficient of partial correlation, t* statistic, or F* statistic.
Stepwise regression resulted in the following output –

```
Add  x7                             with p-value 0.0
Add  x8_3                           with p-value 3.97408e-28
Add  x4                             with p-value 2.56704e-09
Add  x10                            with p-value 1.87426e-07
Add  x8_2                           with p-value 5.66021e-09
Add  x11                            with p-value 0.0183369
Add  x5                             with p-value 0.0289354
Add  x1                             with p-value 0.0369417
Add  x3                             with p-value 0.0231816
resulting features:
['x7', 'x8_3', 'x4', 'x10', 'x8_2', 'x11', 'x5', 'x1', 'x3']
```

Regression Results from the good subset –

OLS Regression Results

Dep. Variable:	y_box	R-squared:	0.838
Model:	OLS	Adj. R-squared:	0.834
Method:	Least Squares	F-statistic:	210.3
Date:	Sat, 27 Apr 2024	Prob (F-statistic):	9.82e-154
Time:	01:15:31	Log-Likelihood:	3239.3
No. Observations:	417	AIC:	-6457.
Df Residuals:	406	BIC:	-6412.
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.6504	0.006	277.644	0.000	1.639	1.662
x1	0.0005	2.98e-05	16.331	0.000	0.000	0.001
x2	-1.693e-05	1.11e-05	-1.520	0.129	-3.88e-05	4.97e-06
x3	8.148e-05	1.55e-05	5.263	0.000	5.1e-05	0.000
x4	6.354e-05	1.63e-05	3.898	0.000	3.15e-05	9.56e-05
x6	5.158e-05	2.05e-05	2.519	0.012	1.13e-05	9.18e-05
x7	0.0041	0.001	5.337	0.000	0.003	0.006
x9	4.808e-05	1.25e-05	3.859	0.000	2.36e-05	7.26e-05
x10	0.0001	1.31e-05	7.688	0.000	7.5e-05	0.000
x8_2	-0.0001	2.02e-05	-5.398	0.000	-0.000	-6.93e-05
x8_3	-0.0001	2.81e-05	-4.741	0.000	-0.000	-7.8e-05

Omnibus:	13.345	Durbin-Watson:	1.976
Prob(Omnibus):	0.001	Jarque-Bera (JB):	23.084
Skew:	-0.183	Prob(JB):	9.71e-06
Kurtosis:	4.093	Cond. No.	1.75e+04

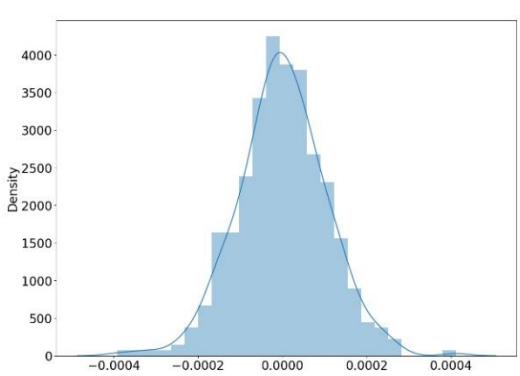
We can see that the Adjusted R square is very high which indicates that this subset is good for building our model. Also, the p-values of all variables are less than 0.05 which concludes that all the variables are significant.

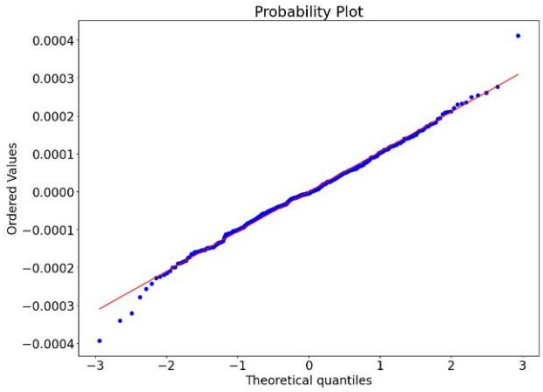
CHAPTER 7

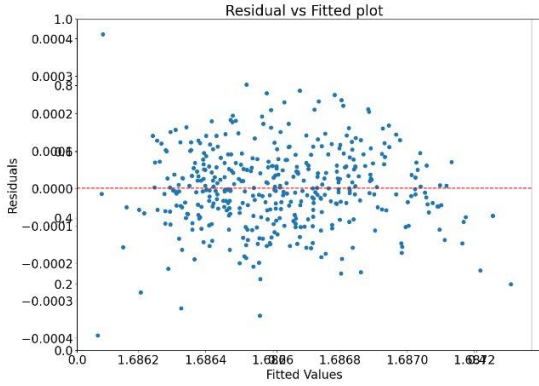
Diagnostics and Remedial Measures

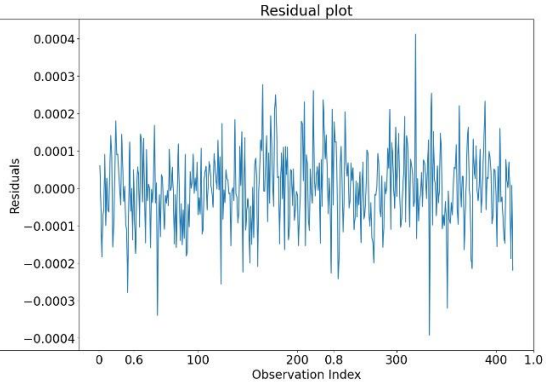
Model Diagnostics :

Residual analysis (Regression Plots)



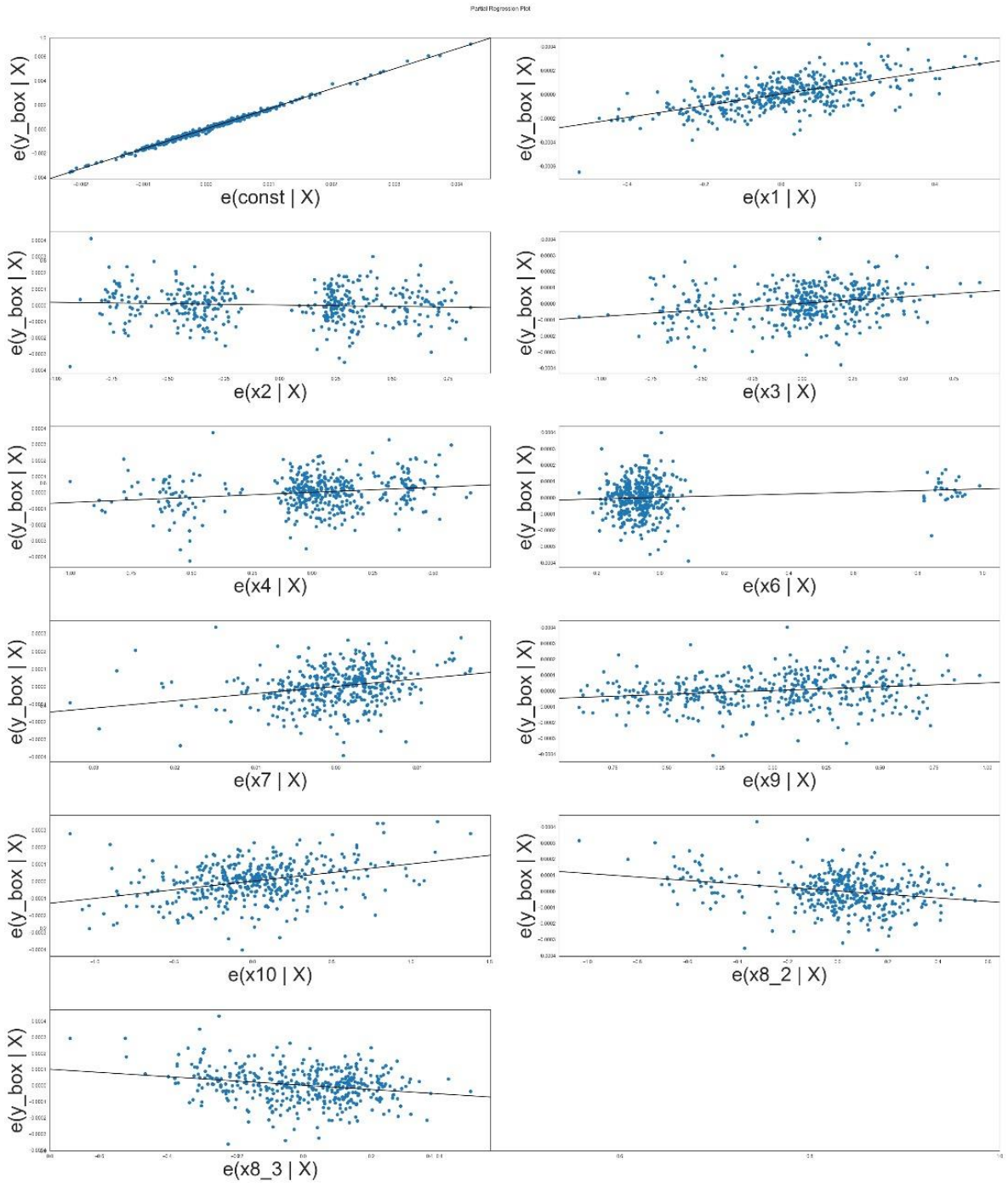




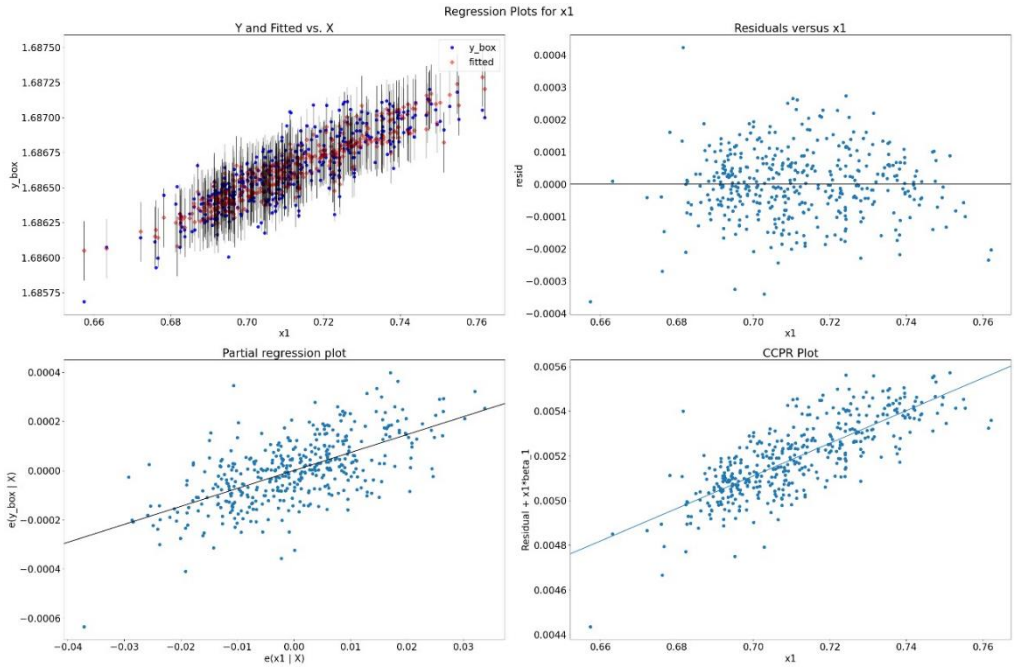


- From the Residual density plot we can see that it is bell-shaped and symmetric around 0.
- From the normal probability plot we can see that there is not much departure from normality except for some extreme values which we will deal with later
- The Residual vs Fitted plot shows a random scattering of points around the zero line, indicating that the model's predictions are unbiased on average across the range of fitted values.
- Error terms seem to have constant variance except for some extreme values

Added Variable plots



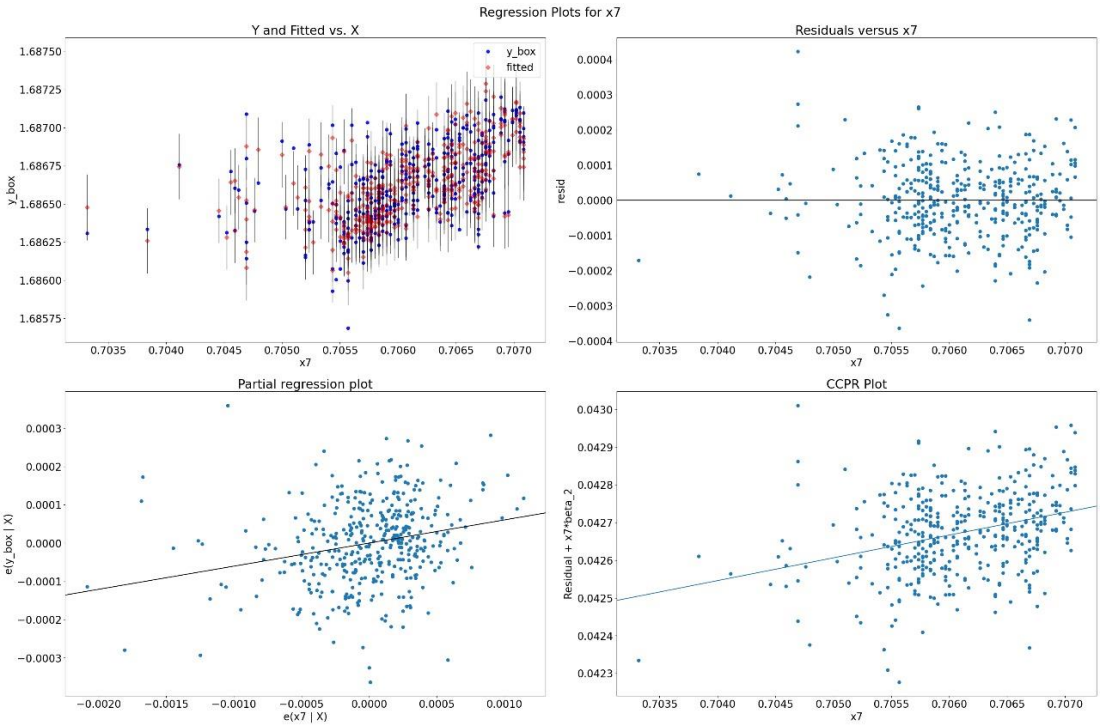
Variable X1



This residual ·plot suggests that a linear relation for X1 is appropriate in the model already containing other variables. From the Partial regression plot, we can observe a positive linear relation for variable X1.

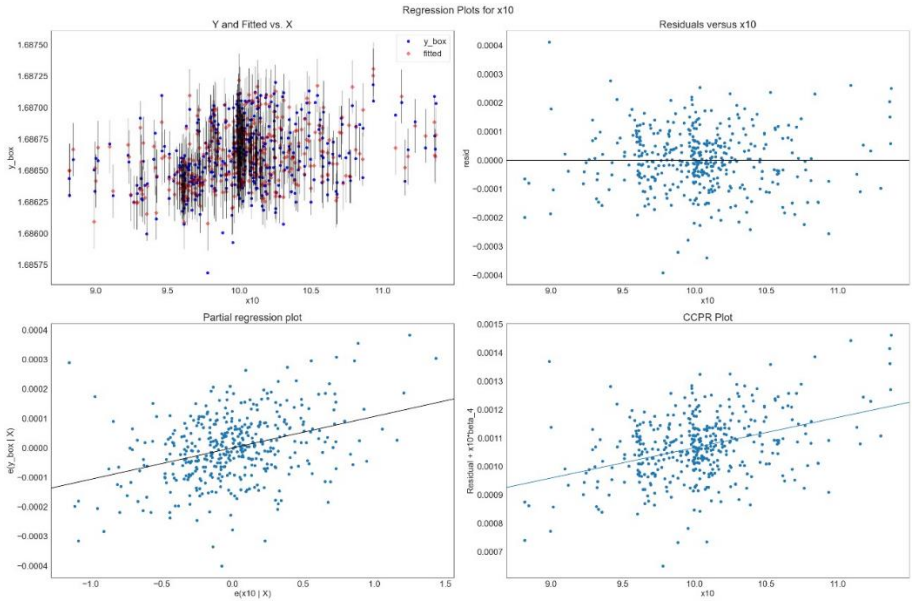
CCPR plots are valuable diagnostic tools in regression analysis, providing insights into the relationship between individual variables and the response variable while considering the effects of other variables in the model.

Variable X7



This residual ·plot suggests that a linear relation for X7 is appropriate in the model already containing other variables. From the Partial regression plot, we can observe a positive linear relation for variable X7.

Variable X10



This residual ·plot suggests that a linear relation for X10 is appropriate in the model already containing other variables. From the Partial regression plot, we can observe a positive linear relation for variable X10.

- For categorical predictor variables, added variable plots are not typically used because the concept of adding one unit of a categorical variable does not apply in the same way as it does for continuous variables.

Second-Order model with Possible Interaction terms

Here we are considering all the interactions terms and second degree terms and using stepwise selection algorithm we will analyse all first order possible interactions including second degree terms in the model.
Before doing this we will transform the variables using correlation transformation as follows:

$$Y_i^* = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i - \bar{Y}}{s_Y}\right) \qquad s_Y = \sqrt{\frac{\sum_i (Y_i - \bar{Y})^2}{n-1}}$$

$$X_{ik}^* = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik} - \bar{X}_k}{s_k}\right) \qquad s_k = \sqrt{\frac{\sum_i (X_{ik} - \bar{X}_k)^2}{n-1}}$$

OLS Regression Results

Dep. Variable:	y_std	R-squared (uncentered):	0.857
Model:	OLS	Adj. R-squared (uncentered):	0.843
Method:	Least Squares	F-statistic:	61.59
Date:	Sat, 27 Apr 2024	Prob (F-statistic):	2.31e-137
Time:	01:22:21	Log-Likelihood:	1071.3
No. Observations:	417	AIC:	-2069.
Df Residuals:	380	BIC:	-1919.
Df Model:	37		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x1	-167.9797	116.141	-1.446	0.149	-396.338	60.379
x2	0.4971	2.074	0.240	0.811	-3.581	4.575
x3	-2.7201	2.886	-0.942	0.347	-8.395	2.955
x4	-2.9574	3.705	-0.798	0.425	-10.243	4.328
x6	4.9913	7.140	0.699	0.485	-9.048	19.031
x7	278.1319	132.754	2.095	0.037	17.107	539.156
x9	-2.8660	2.321	-1.235	0.218	-7.430	1.697
x10	-67.8983	56.793	-1.196	0.233	-179.566	43.770
x8_2	-7.4133	4.219	-1.757	0.080	-15.708	0.882
x8_3	-13.2283	6.018	-2.198	0.029	-25.061	-1.395
x1^2	-2.1187	8.967	-0.236	0.813	-19.751	15.513
x7^2	-744.3295	353.466	-2.106	0.036	-1439.323	-49.336
x10^2	1.4995	1.703	0.880	0.379	-1.849	4.848
x1*x2	0.0827	0.175	0.474	0.636	-0.260	0.426
x1*x3	-0.1801	0.343	-0.525	0.600	-0.855	0.494
x1*x4	-0.3903	0.291	-1.340	0.181	-0.963	0.182
x1*x6	-0.0969	0.366	-0.265	0.791	-0.816	0.622
x1*x9	-0.2127	0.224	-0.950	0.343	-0.653	0.227
x1*x8_2	1.8142	0.435	4.170	0.000	0.959	2.669
x1*x8_3	2.1071	0.702	3.002	0.003	0.727	3.487
x7*x2	-1.4782	5.626	-0.263	0.793	-12.540	9.583
x7*x3	7.2820	7.757	0.939	0.348	-7.970	22.534
x7*x4	8.3942	9.949	0.844	0.399	-11.168	27.956
x7*x6	-13.1958	19.326	-0.683	0.495	-51.195	24.804
x7*x9	7.8531	6.229	1.261	0.208	-4.394	20.101
x7*x8_2	17.5321	11.177	1.569	0.118	-4.444	39.509
x7*x8_3	32.7650	15.834	2.069	0.039	1.633	63.897
x10*x2	0.0396	0.115	0.344	0.731	-0.186	0.266
x10*x3	0.1938	0.142	1.367	0.173	-0.085	0.473
x10*x4	-0.0134	0.144	-0.093	0.926	-0.296	0.269
x10*x6	-0.0714	0.339	-0.211	0.833	-0.738	0.595
x10*x9	0.0683	0.142	0.480	0.631	-0.212	0.348
x10*x8_2	0.3011	0.207	1.454	0.147	-0.106	0.708
x10*x8_3	0.3900	0.277	1.410	0.159	-0.154	0.934
x1*x7	226.6657	154.673	1.465	0.144	-77.456	530.787
x1*x10	1.7829	3.278	0.544	0.587	-4.663	8.228
x7*x1	226.6657	154.673	1.465	0.144	-77.456	530.787
x7*x10	87.4518	75.636	1.156	0.248	-61.266	236.169
x10*x1	1.7829	3.278	0.544	0.587	-4.663	8.228
x10*x7	87.4518	75.636	1.156	0.248	-61.266	236.169

Result from Stepwise Regression including second degree and interaction terms –

```
Add  x8_3                       with p-value 2.77515e-30
Add  x1*x3                      with p-value 2.01598e-41
Add  x3                         with p-value 6.05992e-33
Add  x1*x8_3                    with p-value 7.5847e-16
Add  x8_2                       with p-value 1.4539e-16
Add  x10^2                      with p-value 5.93475e-09
Add  x1*x8_2                    with p-value 1.64804e-20
Drop x1*x3                      with p-value 0.564178
Add  x7*x8_3                    with p-value 1.29545e-06
Add  x7*x8_2                    with p-value 0.000119883
Add  x10*x9                     with p-value 0.00261394
Add  x10*x4                     with p-value 0.0044034
Add  x10*x8_2                   with p-value 0.00216371
Add  x6                         with p-value 0.00440038
resulting features:
['x8_3', 'x3', 'x1*x8_3', 'x8_2', 'x10^2', 'x1*x8_2', 'x7*x8_3', 'x7*x8_2', 'x10*x9', 'x10*x4', 'x10*x8_2', 'x6']
```

Final_good_features :
['x8_3', 'x3', 'x1*x8_3', 'x8_2', 'x10^2', 'x1*x8_2', 'x7*x8_3', 'x7*x8_2', 'x10*x9', 'x10*x4', 'x10*x8_2', 'x6']

Regression Results from the model using above features –

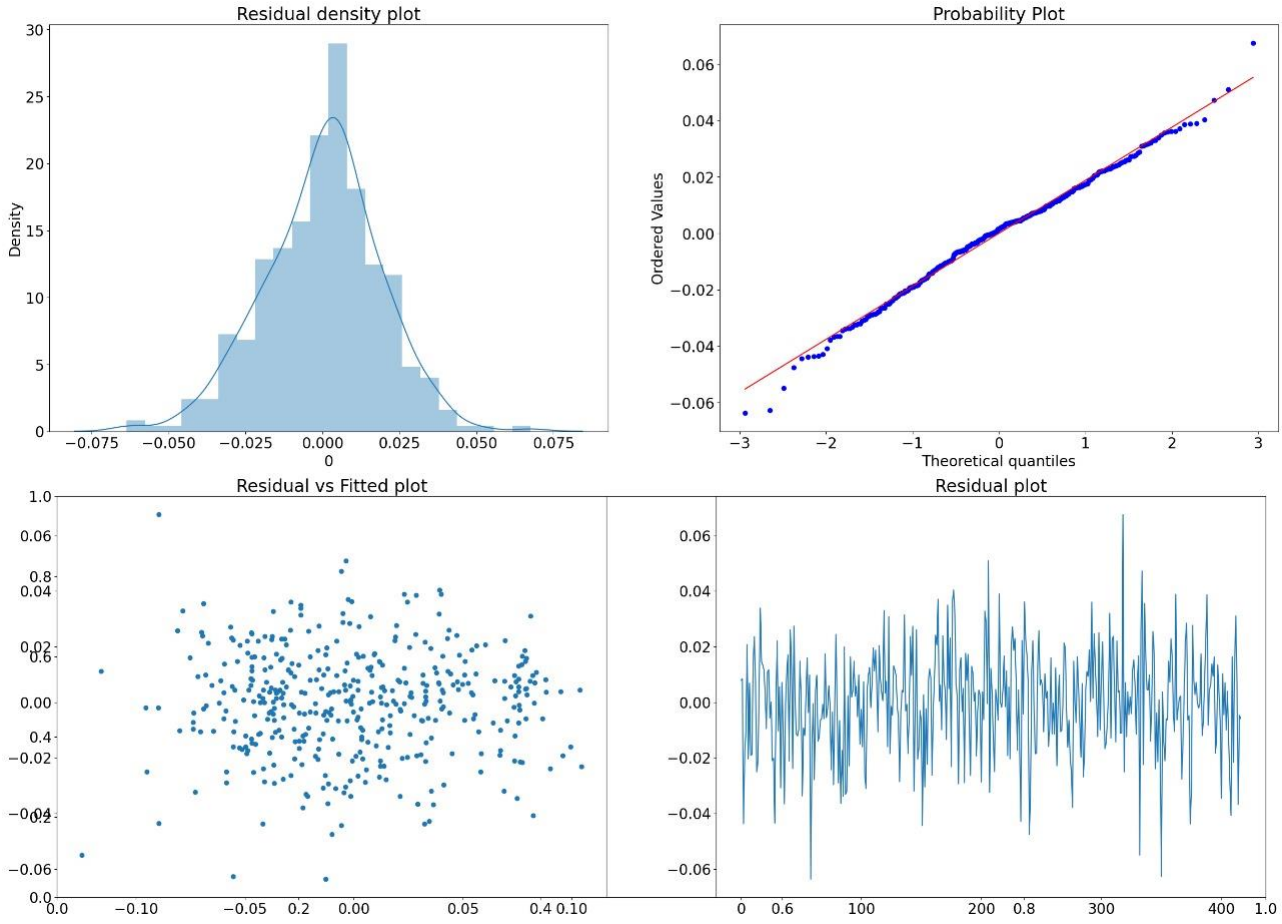
OLS Regression Results

Dep. Variable:	y_std	R-squared (uncentered):	0.849
Model:	OLS	Adj. R-squared (uncentered):	0.845
Method:	Least Squares	F-statistic:	190.4
Date:	Sat, 27 Apr 2024	Prob (F-statistic):	4.39e-158
Time:	01:23:06	Log-Likelihood:	1060.4
No. Observations:	417	AIC:	-2097.
Df Residuals:	405	BIC:	-2048.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x8_3	-7.3945	1.967	-3.759	0.000	-11.261	-3.528
x3	0.0145	0.003	5.086	0.000	0.009	0.020
x1*x8_3	2.4525	0.215	11.418	0.000	2.030	2.875
x8_2	-7.1551	1.401	-5.108	0.000	-9.909	-4.402
x10^2	0.2016	0.020	10.050	0.000	0.162	0.241
x1*x8_2	2.0584	0.130	15.789	0.000	1.802	2.315
x7*x8_3	17.2173	5.247	3.282	0.001	6.903	27.531
x7*x8_2	16.6697	3.747	4.449	0.000	9.304	24.035
x10*x9	0.0138	0.005	2.968	0.003	0.005	0.023
x10*x4	0.0209	0.006	3.303	0.001	0.008	0.033
x10*x8_2	0.2275	0.069	3.308	0.001	0.092	0.363
x6	0.0109	0.004	2.864	0.004	0.003	0.018

Omnibus:	12.075	Durbin-Watson:	1.916
Prob(Omnibus):	0.002	Jarque-Bera (JB):	20.878
Skew:	-0.150	Prob(JB):	2.93e-05
Kurtosis:	4.054	Cond. No.	7.12e+03

We can clearly notice an increase in adjusted R square after adding interaction and 2nd order terms. Adjusted R square was 0.834 when there were no interaction terms and 2nd-degree terms in the model
and after adding significant interaction and 2nd-degree terms in the model adjusted R square increased 0.845.



- From the Residual density plot we can see that it is bell-shaped and symmetric around 0.
- From the normal probability plot we can see that there is not much departure from normality except for some extreme values which we will deal with later
- The Residual vs Fitted plot shows a random scattering of points around the zero line, indicating that the model's predictions are unbiased on average across the range of fitted values.
- Error terms seem to have constant variance except for some extreme values

Identifying Influential Cases

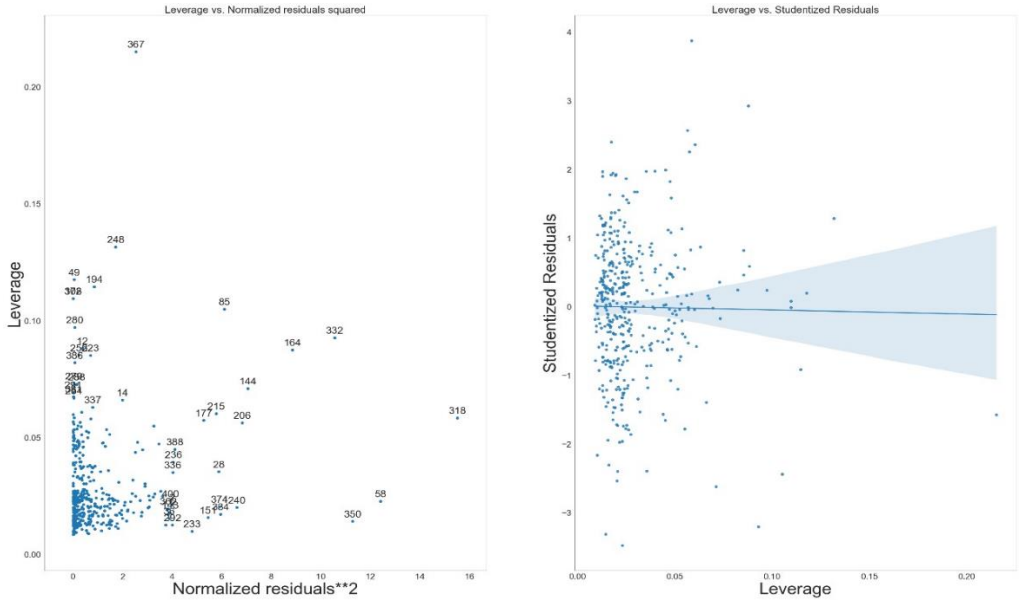
We have calculated studentised residuals, leverage values, Cook's distance and DFFITS for identifying outlying and influential observations.

	standard_resid	student_resid	hat_diag	cooks_d	dffits_internal	dffits
0	0.422580	0.422151	0.020765	3.155581e-04	0.061536	0.061474
1	-0.038638	-0.038591	0.052362	6.874356e-06	-0.009083	-0.009071
2	-1.994529	-2.001921	0.021381	7.242891e-03	-0.294813	-0.295906
3	-0.805918	-0.805568	0.016634	9.155380e-04	-0.104816	-0.104771
4	-0.576718	-0.576242	0.024955	7.093833e-04	-0.092264	-0.092188
...	...	...	...	...	...	...
412	1.892889	1.898969	0.024886	7.620192e-03	0.302394	0.303366
413	0.281996	0.281675	0.021834	1.479204e-04	0.042131	0.042083
414	-1.755352	-1.759891	0.023359	6.141462e-03	-0.271473	-0.272175
415	-0.259862	-0.259563	0.015442	8.826070e-05	-0.032544	-0.032507
416	-0.005383	-0.005376	0.022649	5.595585e-08	-0.000819	-0.000818

417 rows × 6 columns

Use of hat matrix for identifying outlying observations

Considering the observations with large residuals and high leverage values as influential. We plotted the flowing plots to visualize the influential observations.



Observations with both high leverage and large normalized residuals are potential outliers that exert a significant influence on the regression model's fit.

We can clearly notice many outlying cases with high leverage values in plots. Any observation with a leverage value greater than 2p/n is considered to be outlying in their X values.

Given below are the values and index with leverage values greater than 2p/n :

	y	hat_diag
12	0.056101	0.088251
14	0.011474	0.066160
49	0.026443	0.117646
85	-0.035421	0.105051
108	-0.091658	0.109598
144	0.003452	0.071055
164	0.034401	0.087630
194	-0.003284	0.114673
223	0.018092	0.085310
248	-0.055140	0.131688
254	-0.034215	0.066962
255	-0.005489	0.085332
279	0.056570	0.073198
280	-0.003731	0.097200
288	-0.105004	0.072842
291	0.008654	0.069447
332	-0.179878	0.092811
337	-0.059196	0.063052
367	-0.060235	0.215203
372	0.044431	0.109594
381	-0.038716	0.067586
386	0.009606	0.082141

Influence on Single Fitted Values – DFFITS

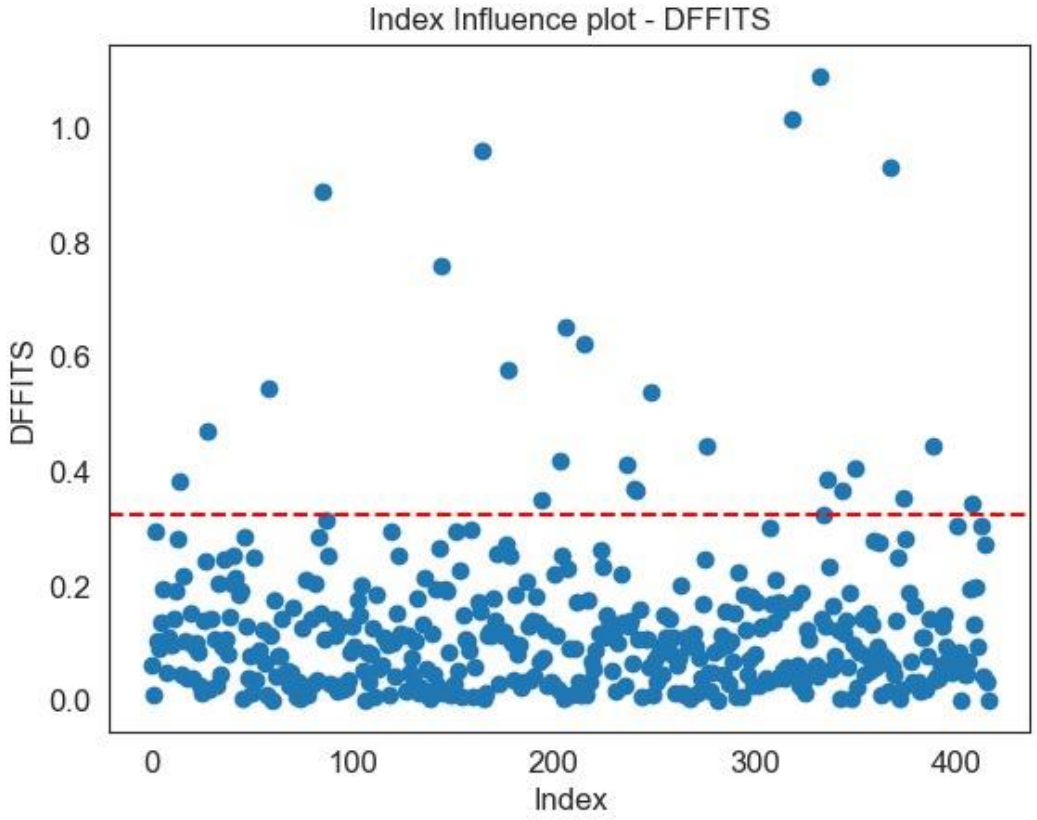
A useful measure of the Influence that case i has on the fitted value Yi, is given by:

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

It can be shown that the DFFITS values can be computed by using only the result from fitting the entire data set, as follows:

$$(DFFITS)_i = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}\right]^{1/2} \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2} = t_i \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

As a guideline for identifying influential cases, we suggest considering a case influential if the absolute value of DFFITS exceeds 1 for small to medium data sets and $2\sqrt{\left(\frac{P}{n}\right)}$ for large data sets. Given below are the values and index with DFFITS greater than $2\sqrt{\left(\frac{P}{n}\right)}$ –



Observations above the red line are Influential cases according to DFFITS.

Influence on all Fitted Values – Cook's Distance

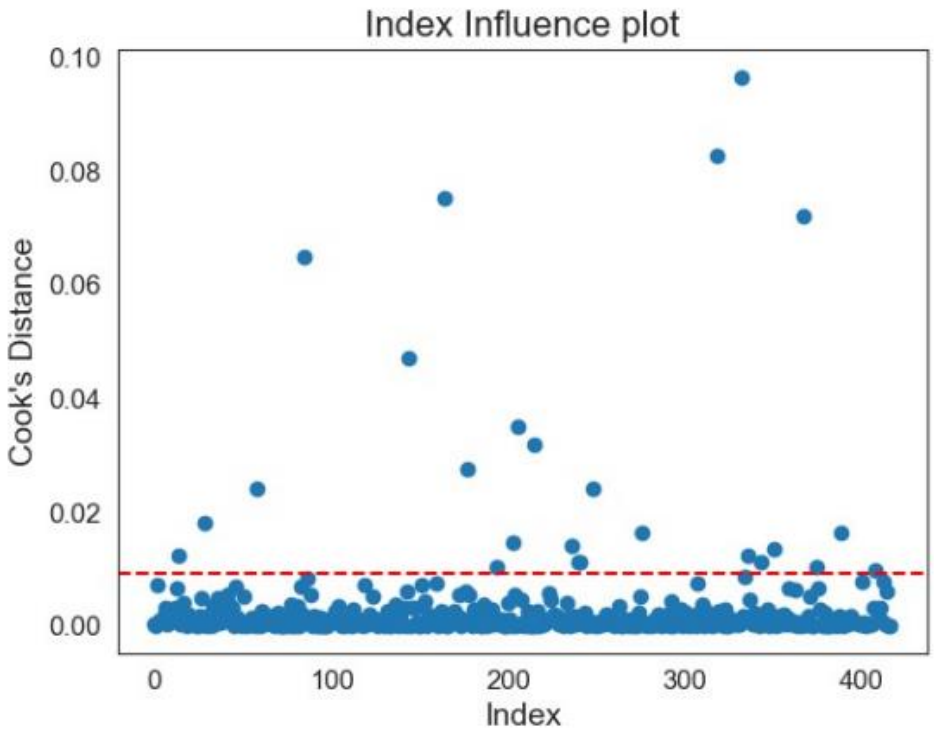
In contrast to the DFFITS measure, which considers the influence of the ith case on the fitted value Yi for this case, Cook's distance measure considers the influence of the ith case on all n fitted values. Cook's distance measure, denoted by Di , is an aggregate influence measure, showing the effect of the ith case on all n fitted values

$$D_i = \frac{\sum_{j=1}^{n} \left(\hat{Y}_j - \hat{Y}_{j(i)}\right)^2}{pMSE}$$

Fortunately, Cook's distance measure Di can be calculated without fitting a new regression function each time a different case is deleted. An algebraically equivalent expression is:

$$D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

We have used the rule of thumb that cook's distance greater than 3 times the mean is a possible outlier. Following are the cases identified as outlying by Cook's Distance.



- We can see the outlying cases given by Cook's distance are common in the cases given by DFFITS and high leverage values. So, we will remove these cases from our model.

Model - After removing outlying cases

OLS Regression Results

Dep. Variable:	y_std	R-squared (uncentered):	0.888
Model:	OLS	Adj. R-squared (uncentered):	0.885
Method:	Least Squares	F-statistic:	252.1
Date:	Sat, 27 Apr 2024	Prob (F-statistic):	1.43e-172
Time:	01:24:32	Log-Likelihood:	1067.8
No. Observations:	392	AIC:	-2112.
Df Residuals:	380	BIC:	-2064.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
x8_3	-8.9296	2.061	-4.333	0.000	-12.981	-4.878
x3	0.0138	0.003	5.407	0.000	0.009	0.019
x1*x8_3	2.6196	0.218	12.030	0.000	2.191	3.048
x8_2	-8.1356	1.264	-6.434	0.000	-10.622	-5.649
x10^2	0.2112	0.017	12.172	0.000	0.177	0.245
x1*x8_2	2.1403	0.113	18.954	0.000	1.918	2.362
x7*x8_3	21.1833	5.516	3.841	0.000	10.338	32.028
x7*x8_2	19.2276	3.387	5.676	0.000	12.567	25.888
x10*x9	0.0160	0.004	4.000	0.000	0.008	0.024
x10*x4	0.0160	0.006	2.847	0.005	0.005	0.027
x10*x8_2	0.2235	0.066	3.377	0.001	0.093	0.354
x6	0.0110	0.003	3.457	0.001	0.005	0.017

Omnibus:	1.504	Durbin-Watson:	2.021
Prob(Omnibus):	0.471	Jarque-Bera (JB):	1.553
Skew:	-0.148	Prob(JB):	0.460
Kurtosis:	2.910	Cond. No.	8.74e+03

Multicollinearity Remedial measure

We will be using VIF, i.e., Variance inflation factor for the detection of possible multicollinearity in the model. We obtain the following values of the VIF:

	Var	Vif
0	x8_3	2008491.62
6	x7*x8_3	1986220.71
3	x8_2	1358710.37
7	x7*x8_2	1349106.68
2	x1*x8_3	2972.42
5	x1*x8_2	1557.23
10	x10*x8_2	900.91
4	x10^2	26.66
9	x10*x4	9.67
1	x3	5.24
8	x10*x9	3.35
11	x6	1.13

We will use Ridge regression which can remedy the multicollinearity problem.

Multicollinearity Remedial Measure - Ridge Regression

Ridge regression is one of several methods that have been proposed to remedy multicollinearity problems by modifying the method of least squares to allow biased estimators of the regression coefficients. When an estimator has only a small bias and is substantially more precise than an unbiased estimator, it may well be the preferred estimator since it will have a larger probability of being close to the true parameter value.

Estimator b is unbiased but imprecise, whereas estimator b is much more precise but has a small bias. The probability that b^R falls near the true value is much greater than that for the unbiased estimator b.

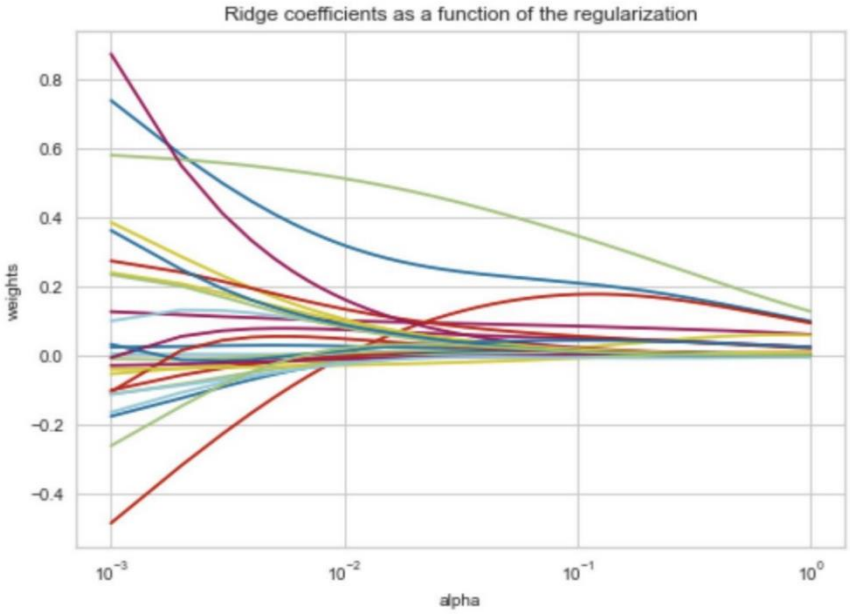
- Ridge Estimated Standardized Regression Coefficients for Different Biasing Constants 'c' –

	C	x8_3	x3	x1*x8_3	x8_2	x10^2	x1*x8_2	x7*x8_3	x7*x8_2	x10*x9	x10*x4	x10*x8_2	x6
0	0.00	-9.263151	0.014135	2.560481	-8.043547	0.268754	2.133641	22.142830	19.063643	0.015912	0.016875	0.167018	0.010981
1	0.05	-0.128019	0.030660	0.173700	-0.245657	0.301656	0.537005	-0.044651	-0.083929	-0.004121	0.031457	0.032078	0.010712
2	0.10	-0.096784	0.032945	0.082287	-0.167057	0.239907	0.300535	-0.034625	-0.058531	-0.005404	0.030437	0.035769	0.010647
3	0.15	-0.085689	0.033933	0.049666	-0.133503	0.199339	0.205299	-0.031000	-0.047307	-0.005695	0.029805	0.032619	0.010558
4	0.20	-0.079953	0.034502	0.032762	-0.114335	0.170695	0.154030	-0.029105	-0.040796	-0.005755	0.029357	0.028362	0.010481
5	0.25	-0.076420	0.034880	0.022376	-0.101755	0.149344	0.122055	-0.027928	-0.036481	-0.005743	0.029013	0.024294	0.010415
6	0.30	-0.074006	0.035154	0.015331	-0.092792	0.132794	0.100236	-0.027119	-0.033388	-0.005707	0.028735	0.020690	0.010359
7	0.35	-0.072240	0.035365	0.010233	-0.086049	0.119578	0.084412	-0.026523	-0.031049	-0.005666	0.028501	0.017562	0.010312
8	0.40	-0.070881	0.035535	0.006373	-0.080773	0.108777	0.072419	-0.026063	-0.029213	-0.005624	0.028299	0.014857	0.010270
9	0.45	-0.069796	0.035676	0.003348	-0.076523	0.099782	0.063020	-0.025693	-0.027729	-0.005585	0.028120	0.012510	0.010233
10	0.50	-0.068904	0.035797	0.000915	-0.073018	0.092172	0.055460	-0.025388	-0.026502	-0.005550	0.027960	0.010464	0.010200
11	0.55	-0.068154	0.035901	-0.001084	-0.070073	0.085651	0.049249	-0.025131	-0.025470	-0.005519	0.027813	0.008669	0.010170
12	0.60	-0.067509	0.035994	-0.002754	-0.067561	0.079999	0.044058	-0.024908	-0.024587	-0.005490	0.027678	0.007085	0.010142
13	0.65	-0.066946	0.036077	-0.004169	-0.065389	0.075053	0.039655	-0.024714	-0.023823	-0.005465	0.027552	0.005679	0.010117
14	0.70	-0.066448	0.036153	-0.005383	-0.063491	0.070688	0.035876	-0.024541	-0.023154	-0.005443	0.027433	0.004423	0.010093
15	0.75	-0.066003	0.036222	-0.006435	-0.061816	0.066808	0.032597	-0.024386	-0.022563	-0.005424	0.027321	0.003298	0.010071
16	0.80	-0.065599	0.036286	-0.007354	-0.060326	0.063336	0.029725	-0.024245	-0.022037	-0.005407	0.027214	0.002283	0.010051
17	0.85	-0.065231	0.036346	-0.008164	-0.058989	0.060211	0.027191	-0.024116	-0.021564	-0.005391	0.027112	0.001364	0.010031
18	0.90	-0.064892	0.036402	-0.008882	-0.057783	0.057382	0.024938	-0.023997	-0.021137	-0.005378	0.027014	0.000529	0.010012
19	0.95	-0.064577	0.036454	-0.009523	-0.056688	0.054811	0.022923	-0.023887	-0.020749	-0.005366	0.026920	-0.000233	0.009994

- VIF Values for Regression Coefficients and R2 for Different Biasing Constants ‘c’ –

C	x8_3	x3	x1*x8_3	x8_2	x10^2	x1*x8_2	x7*x8_3	x7*x8_2	x10*x9	x10*x4	x10*x8_2	x6	R2
0.00	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.888498
0.05	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.789409
0.10	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.766250
0.15	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.754747
0.20	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.747534
0.25	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.742468
0.30	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.738664
0.35	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.735683
0.40	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.733271
0.45	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.731273
0.50	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.729588
0.55	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.728144
0.60	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.726892
0.65	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.725793
0.70	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.724821
0.75	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.723953
0.80	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.723173
0.85	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.722467
0.90	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.721825
0.95	2.008492e+06	5.242935	2972.423801	1.358710e+06	26.663947	1557.230608	1.986221e+06	1.349107e+06	3.347916	9.669872	900.911245	1.12706	0.721236

Ridge Trace –



CHAPTER 8

Model Validation and Final Model Selection

OLS Regression Results

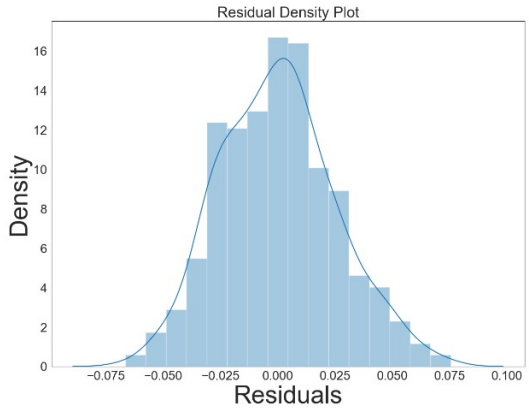
Dep. Variable:	y_std	R-squared:	0.888
Model:	OLS	Adj. R-squared:	0.885
Method:	Least Squares	F-statistic:	251.7
Date:	Sat, 27 Apr 2024	Prob (F-statistic):	3.53e-172
Time:	02:20:39	Log-Likelihood:	1068.2
No. Observations:	392	AIC:	-2110.
Df Residuals:	379	BIC:	-2059.
Df Model:	12		
Covariance Type:	nonrobust		

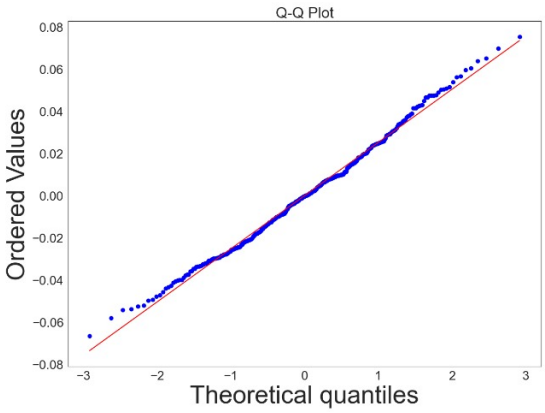
	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0153	0.017	-0.891	0.374	-0.049	0.019
x8_3	-9.2632	2.095	-4.422	0.000	-13.382	-5.144
x3	0.0141	0.003	5.478	0.000	0.009	0.019
x1*x8_3	2.5605	0.228	11.245	0.000	2.113	3.008
x8_2	-8.0435	1.269	-6.339	0.000	-10.539	-5.548
x10^2	0.2688	0.067	4.020	0.000	0.137	0.400
x1*x8_2	2.1336	0.113	18.848	0.000	1.911	2.356
x7*x8_3	22.1428	5.621	3.939	0.000	11.090	33.196
x7*x8_2	19.0636	3.393	5.618	0.000	12.392	25.736
x10*x9	0.0159	0.004	3.966	0.000	0.008	0.024
x10*x4	0.0169	0.006	2.958	0.003	0.006	0.028
x10*x8_2	0.1670	0.092	1.822	0.069	-0.013	0.347
x6	0.0110	0.003	3.453	0.001	0.005	0.017

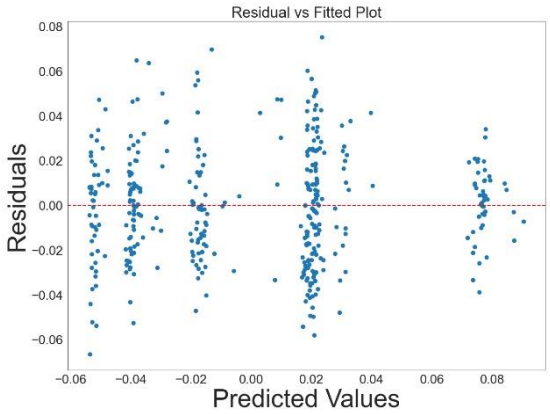
Omnibus:	1.736	Durbin-Watson:	2.024
Prob(Omnibus):	0.420	Jarque-Bera (JB):	1.799
Skew:	-0.157	Prob(JB):	0.407
Kurtosis:	2.893	Cond. No.	1.12e+04

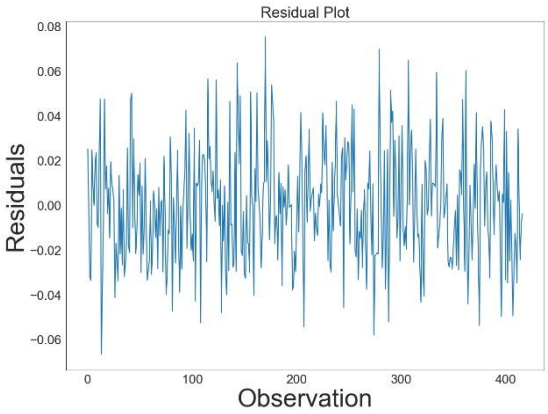
Using Ridge regression Our Model:

x8_3+x3+x1*x8_3+x8_2+x10^2+x1*x8_2+x7*x8_3+x7*x8_2+x10*x9+x10*x4+x10*x8_2+x6









Conclusion

1) Performed EDA on the dataset consisting of 522 rows and 13 columns
2) Detected high leverage and outlying cases, and identified the influential cases among
them from the model using Cook's Distance and removed them.
3) Detected Multicollinearity and performed remedial measures to remove multicollinearity from the data.
4) Implemented best subset selection using different criteria – Adj R2, AIC, BIC, Mallow's
Cp.
5) Implemented Stepwise Feature Selection Algorithm and compared the obtained subsets from the previously obtained subsets based on Adj R2, AIC, BIC, Mallow's Cp.
6) Checked the assumptions and validated the model on test dataset.
7) Achieved adjusted R squared of 0.885 on the test dataset in the final model