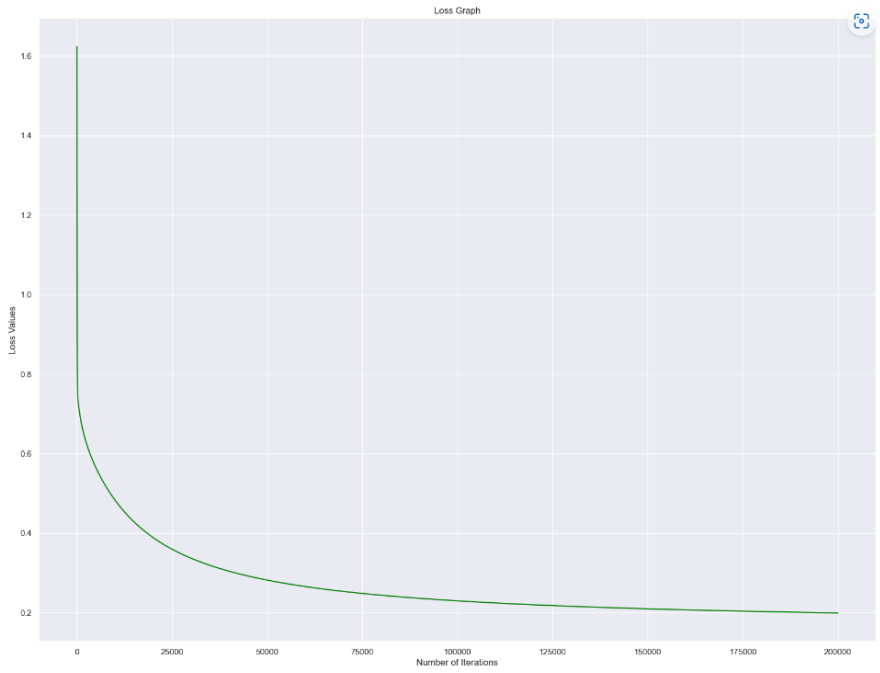
CSE 4/574: Introduction to Machine Learning

Assignment-1 Report

Logistic Regression Method

1. **Provide your best accuracy.**
   - From the analysis and design of the logistic regression model, the best accuracy obtained by using Penguin's dataset is 89.55 percentage.

2. **Include loss graph and provide a short analysis of the results.**
   Loss Graph:



   Short Analysis:

   From the above graph we can observe that as the loss value decreases exponentially as the weights are updated using the gradient descent method. We can observe that the loss value has decreased from 1.6 to 0.199.

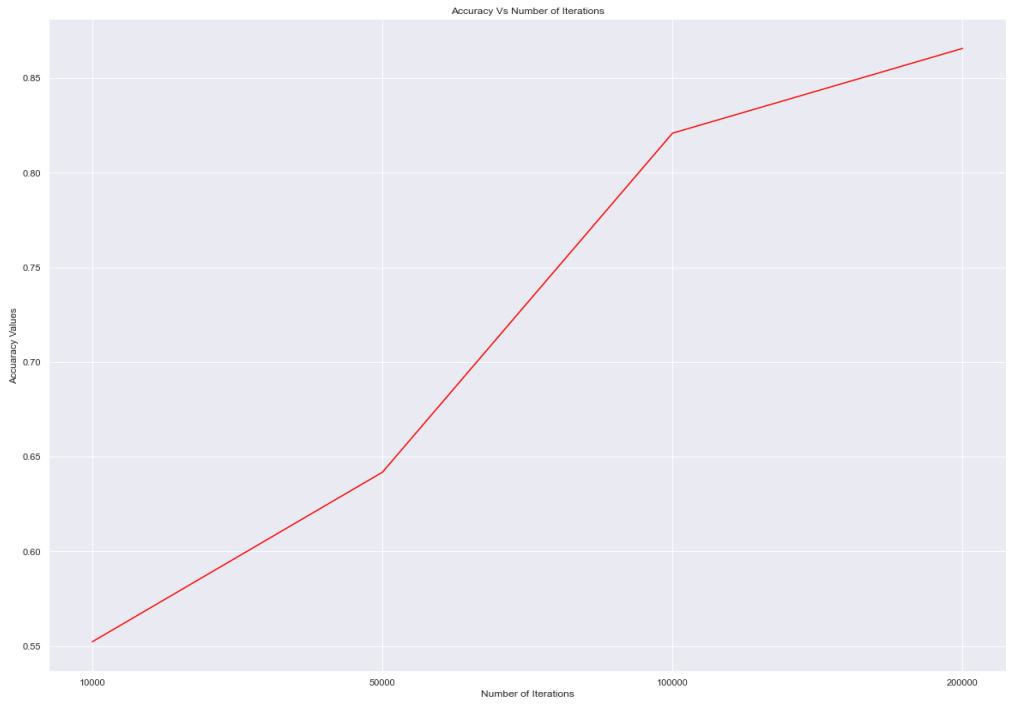
3. **Explain how hyperparameters influence the accuracy of the model. Provide at least 3 different setups with learning rate and #iterations and discuss the results along with plotting of graphs.**
   - The accuracy of the model is influenced by hyperparameters in following ways:
     a) The process of learning and the behavior of the training model are controlled by hyperparameter.
     b) The other specifications of the model are affected by values of hyperparameter, resulting in a direct impact on the performance of the model.
     Example: The other specifications can be biases and weights.

c) By adjusting these values of hyperparameters, accuracy can be improved.
d) The values of hyperparameters are given by the user.
e) Optimization of the model with different values of hyperparameters can help yield better accuracy.
f) Examples of hyperparameters are learning rate, number of iterations, size of batch (Number of Samples), clusters and branches.

- The hyperparameters considered in this model are learning rate and number of iterations.
- Following are considered values of hyperparameters and obtained plots:

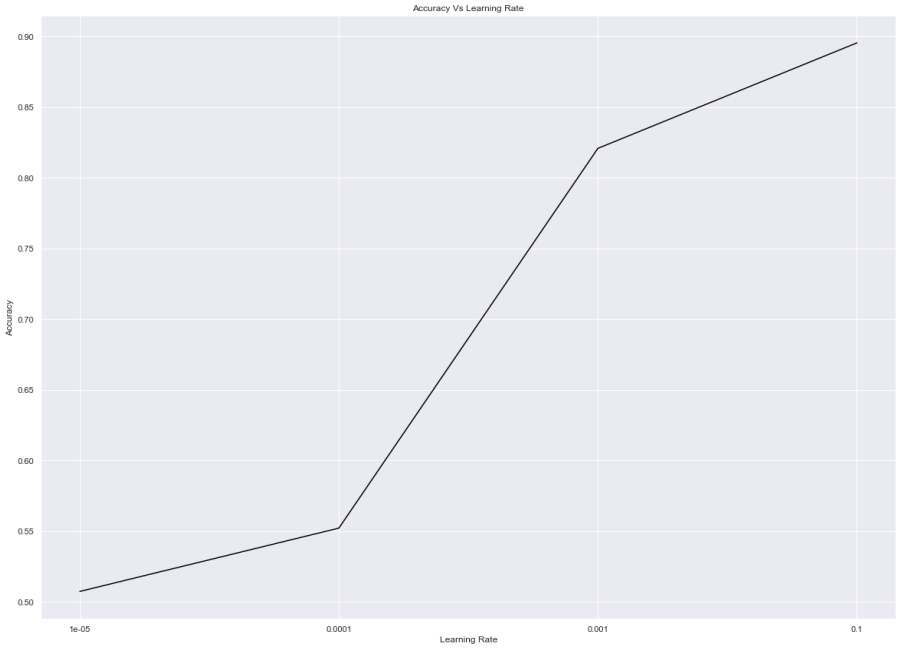
a) Learning Rate = 1e-3, Number of Iterations = 100000
b) Learning Rate = 0.001, Number of Iterations = 10000
c) Learning Rate = 0.01, Number of Iterations = 200000

a) Plot of Accuracy Vs Number of Iterations:



The above plots describe how the accuracy varies as we change the number of iterations. For this plot we have used three different iteration values 10000, 50000, 1000000, 2000000. We can observe that the accuracy increases as the number of iterations are increased.

b) Plot of Accuracy Vs Learning_Rate:



The above plots describe how the accuracy varies as we change the learning rate. For this plot we have used three different learning rate values 0.00001, 0.0001, 0.001, 0.1. We can observe that the accuracy increases as the learning rate is increased.

4. Discuss the benefits/drawbacks of using a Logistic Regression model.

Advantages:

a) Interpretation and implementation are easy, and the model can be trained efficiently.
b) In feature space, no predefined assumptions are made by the model regarding the distribution of classes.
c) Multinomial regression can be obtained by extending the logistic regression model.
d) Provides direction of association along with how appropriate the coefficient size is.
e) The speed of the model is fast in classifying unknown records.
f) The performance of the model is good when the dataset is linearly separable and provides good accuracy for simple datasets.
g) Model coefficients can be interpreted by the model as feature importance indicators.

Disadvantages:

a) This results in overfitting when the number of features is greater than the number of observations.
b) Have linear boundaries.
c) The model assumes linearity between independent and dependent variables, which is one of the major limitations.
d) Only discrete functions can be predicted by logistic regression; hence, there is a limitation of the dependent variable.
e) As the model has a linear decision surface, non-linear problems cannot be analyzed.
f) Relationships that are complex are tough to obtain by using this model.
g) There is a necessity for linearity between log odds and independent variables.

Contributions:

Team Member	Assignment Part	Contribution (%)
Sai Rohit Uddagiri	Model building and report	50
Bhavana Yetinthala	Data pre-processing and report	50

Linear Regression Method

1. **Provide brief details about the nature of your dataset. What is it about? What type of data are we encountering? How many entries and variables does the dataset comprise?**

- We have chosen the diamond data set for the Assignment Part -2. The data is about the prices and other attributes of almost 54,000 diamonds. The variables are as follows:
    - **price** price in US dollars (326$− 18,823$)
    - **carat** weight of the diamond (0.2--5.01)
    - **cut** quality of the cut (Fair, Good, Very Good, Premium, Ideal)
    - **color** diamond color, from D (best) to J (worst)
    - **clarity** a measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
    - **x** length in mm (0--10.74)
    - **y** width in mm (0--58.9)
    - **z** depth in mm (0--31.8)
    - **depth** total depth percentage = z / mean (x, y) = 2 * z / (x + y) (43--79)
    - **table** width of top of diamond relative to widest point (43--95)

  The data frame consists of 53940 rows and 10 variables. The columns cut, color, clarity are categorical features whereas carat, depth, table, x, y, z are continuous features.

2. **Provide the main statistics about the entries of the dataset (mean, std, number of missing values, etc.)**

- The dataset does not contain any null values or duplicate values, but it contains a few zero values for the x, y, z columns. The statistics of the dataset are described as below:

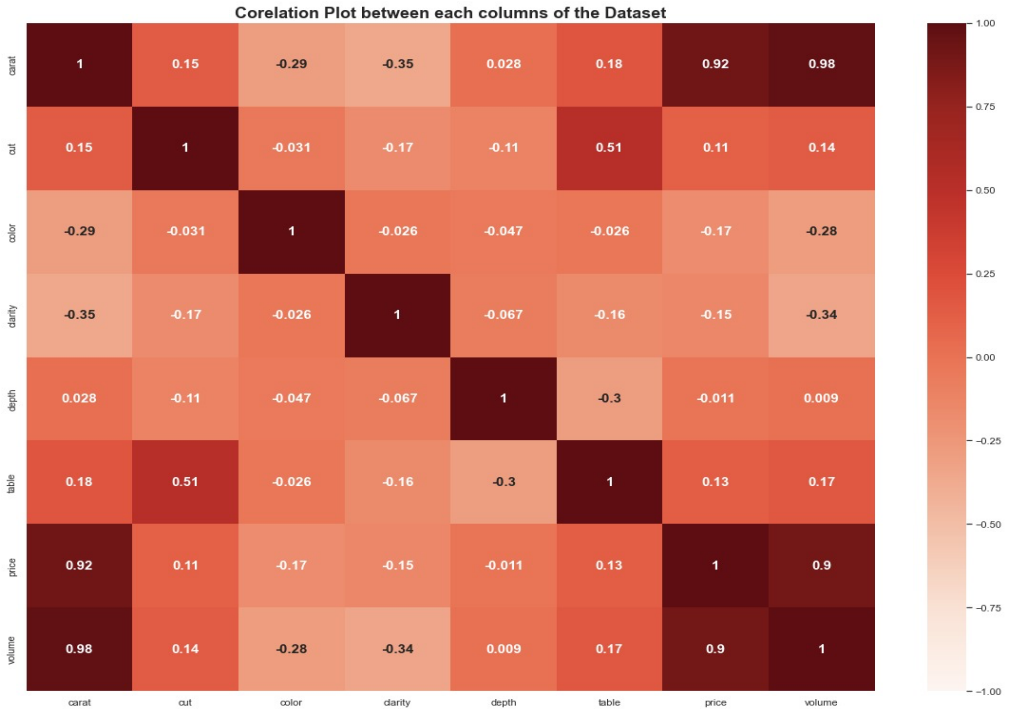
```
diamond_data.describe()
```

	Unnamed: 0	carat	depth	table	price	x	y	z
count	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000	53940.000000
mean	26970.500000	0.797940	61.749405	57.457184	3932.799722	5.731157	5.734526	3.538734
std	15571.281097	0.474011	1.432621	2.234491	3989.439738	1.121761	1.142135	0.705699
min	1.000000	0.200000	43.000000	43.000000	326.000000	0.000000	0.000000	0.000000
25%	13485.750000	0.400000	61.000000	56.000000	950.000000	4.710000	4.720000	2.910000
50%	26970.500000	0.700000	61.800000	57.000000	2401.000000	5.700000	5.710000	3.530000
75%	40455.250000	1.040000	62.500000	59.000000	5324.250000	6.540000	6.540000	4.040000
max	53940.000000	5.010000	79.000000	95.000000	18823.000000	10.740000	58.900000	31.800000

- The columns depth, table, x, y, z are normally distributed as the mean values of those columns is approximately same as the median value.

3. Provide at least 5 visualization graphs with a brief description for each graph, e.g. discuss if there are any interesting patterns or correlations.

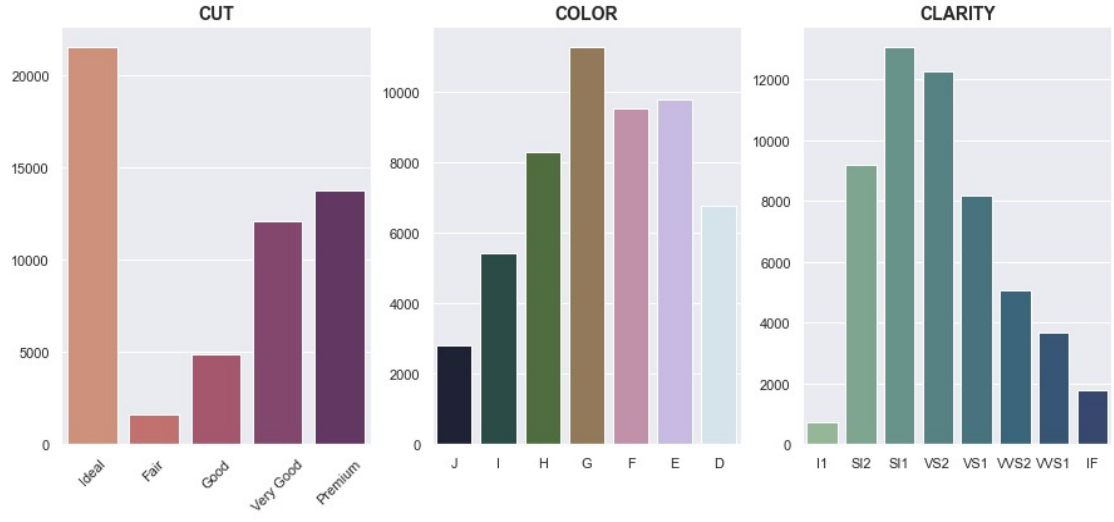
Visualization Graph1:



Description:

- The Correlation plot for the dataset describes about the correlation of different columns with other columns. From the plot, we identify that the target column ‘price’ has a high correlation with the features ‘carat’ and ‘volume’.

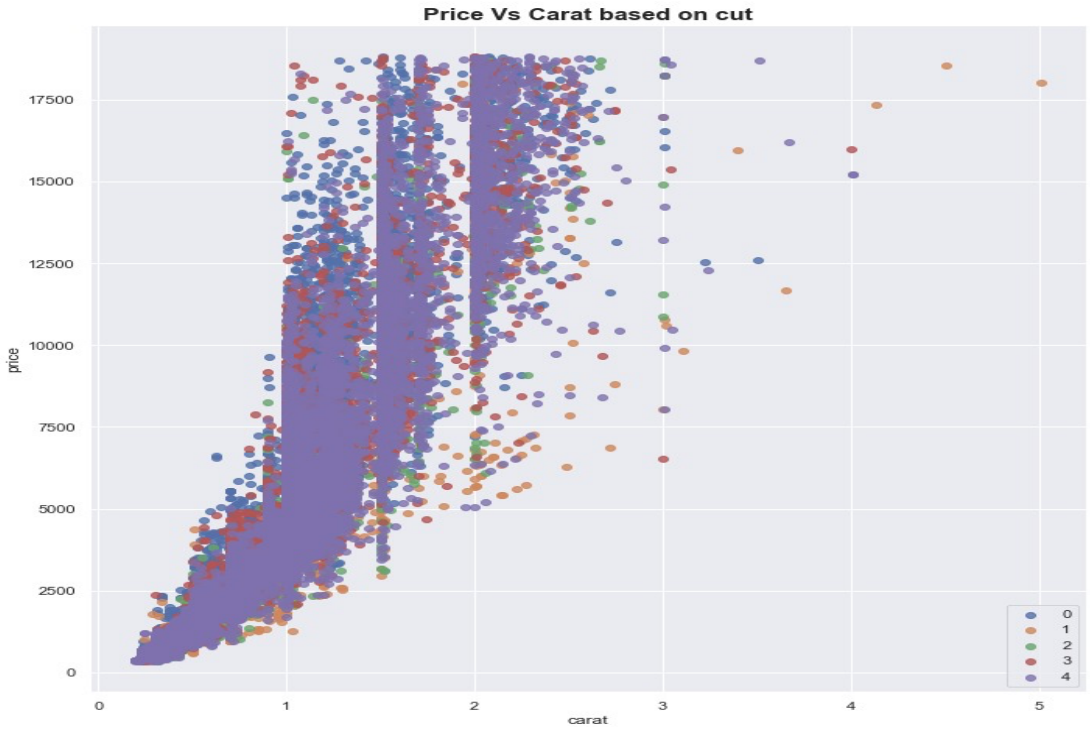
Visualization Graph2:



Description:

- The bar plot shows the classification of diamonds based on categorical features in the dataset. From the bar plot we identify the following:
    - The Ideal cut is the most common cut.
    - Diamonds of color type ‘G’ are more in the dataset.
    - SI1 is the common clarity type in the dataset.

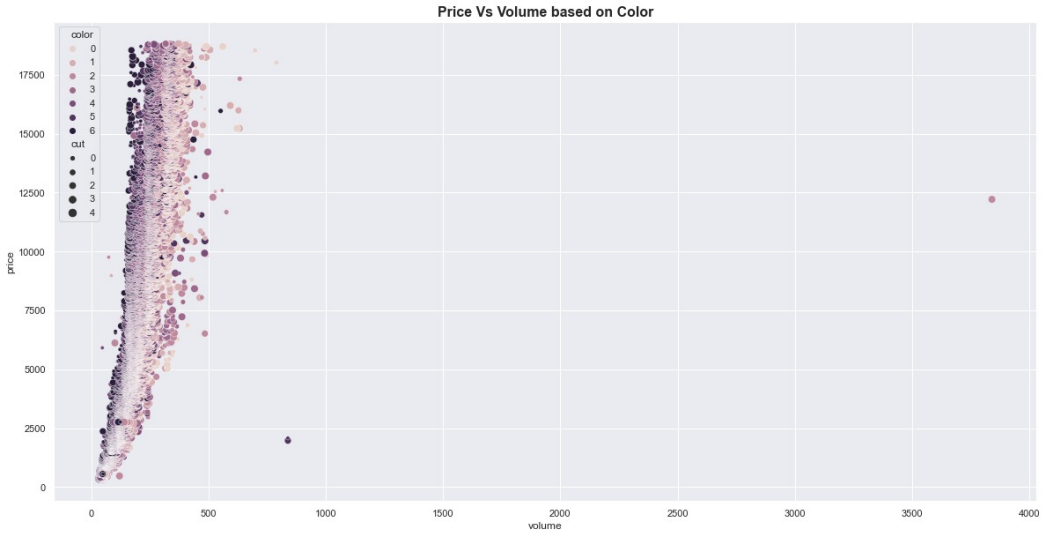
Visualization Graph3:



Description:

- The scatter plot describes the variation of price with respect to carat for each cut type. We can identify a linear relationship between these two columns as the carat value increases, the price values also increase.

Visualization Graph4:



Description:

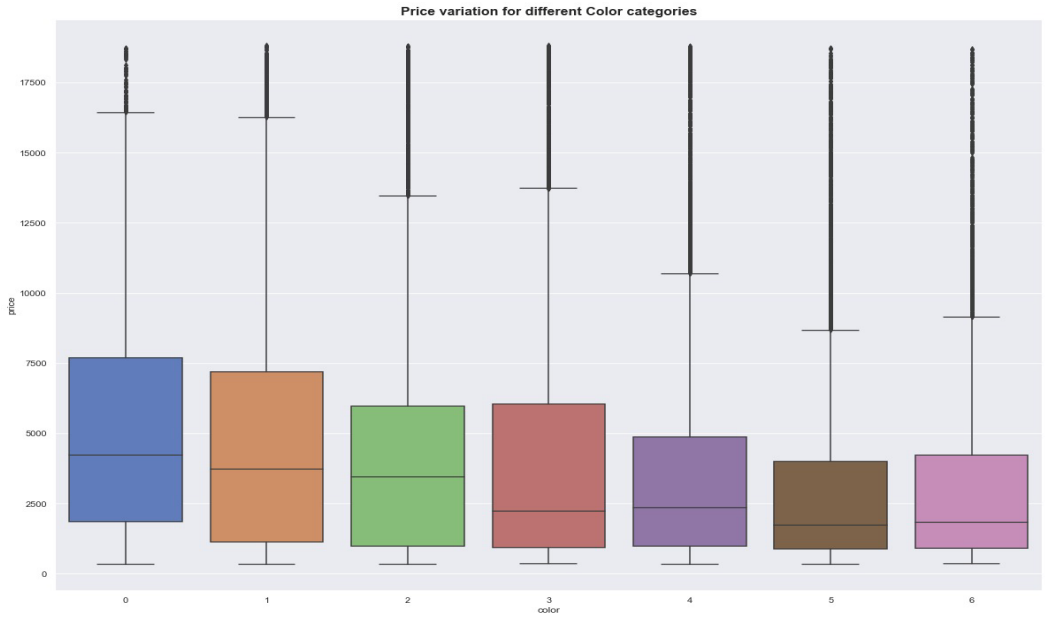
- The above graph describes the relationship between price and volume features. We identify a few outliers from this plot which can be resolved by using normalization techniques.

Visualization Graph5:

Description:

- The above graph is between the price and volume columns. Here the volume column is normalized by applying root transformation. We identify that the linear relationship between the price and volume has improved by using this technique.

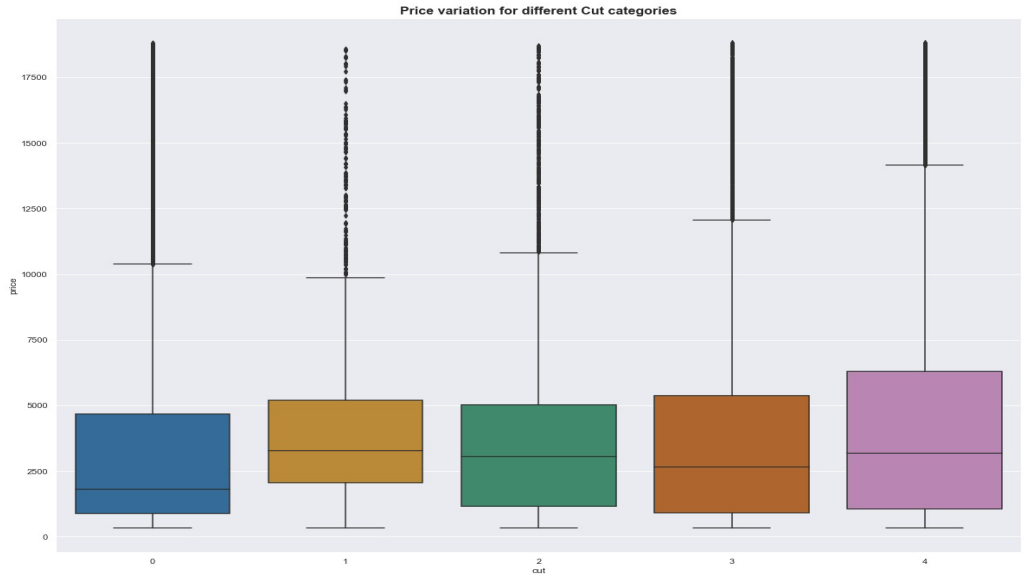
Visualization Graph6:



Description:

- The above box plot gives details about the price variation for different categories of color. We observe that color ‘G’ has higher median price compared to other colors.

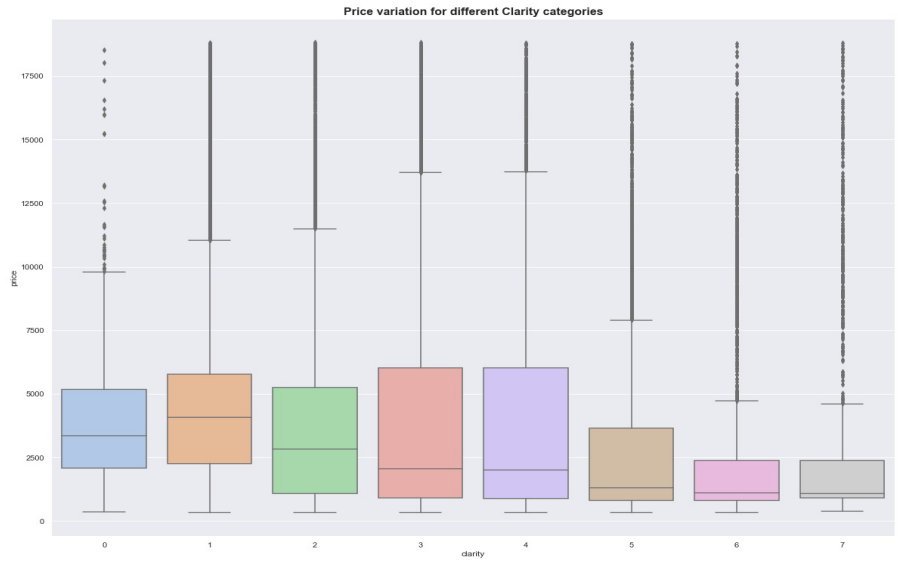
Visualization Graph7:



Description:

- The above box plot gives details about the price variation for different categories of cut. We observe that premium cut has higher median price compared to other colors.

Visualization Graph8:



Description:

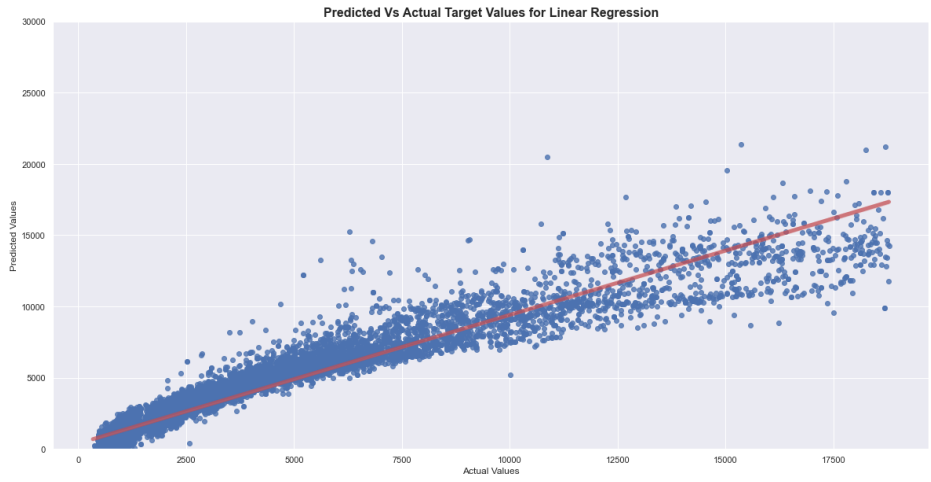
- The above box plot gives details about the price variation for different categories of clarity. We observe that color SI2 has higher median price compared to other colors.

4. Provide your loss value.

- From the analysis and design of the linear regression model, the loss value obtained by using diamond dataset is 32598096141.657288.

5. Show the plot comparing the predictions vs the actual test data.

- Plot comparing the predictions and actual test data:



6. Discuss the benefits/drawbacks of using OLS estimate for computing the weights.

Benefits:

a) When the held assumptions are true, the ordinary least squares model is the most efficacious estimator and a comparatively easy and approachable method.
b) The ordinary least squares model gives data regarding cost structures.
c) Can differentiate between the roles of different variables that are responsible for affecting output.
d) Values of coefficients can be analyzed either in terms of how inputs help to obtain the output response or as cost drivers.
e) Omission of variables and errors in data can be used for the interpretation of data noise.
f) A hypothesis test can be performed on the model specification and for important variables.

Drawbacks:

a) The main drawback of an ordinary least squares estimate is that there is a presumption that weights should be known exactly.
b) Feasible results are not obtained when the dataset is small, which is good for a large dataset.
c) If the dataset contains abnormally high or low values, the results are misinterpreted.
d) If there is non-linearity, weight values may be inexact.
e) If the error term is not properly analyzed, the results of the model are inclined toward functional form, depending on the initial design of the regression.
f) Assumptions on the error term directly impact inefficiency and noise division.

7. Discuss the benefits/drawbacks of using a Linear Regression model.

Benefits:

a) Implementation is simple, and the interpretation of output coefficients is easy.
b) When compared to other models, it is the best algorithm to implement as it is less complex when there is a linear relationship between dependent and independent variables.
c) The model can be used with both categorical and continuous data.

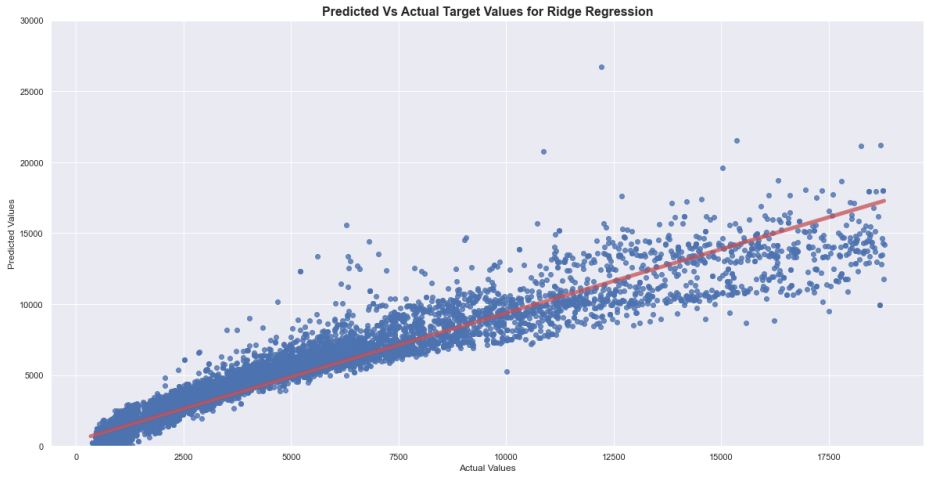
Drawbacks:

a) Limits to linear boundaries.
b) The technique assumes a straight-line (linear) relationship between independent and dependent variables.
c) Linear regression can be impacted if the dataset contains abnormally high or low values.
d) Over-fitting can occur when the number of independent variables is too large.
e) Does not provide complete information regarding relationships among variables.
f) There's a problem of multicollinearity, as linear regression assumes that independent variables are not highly correlated.

Ridge Regression Method

1. **Provide your loss value.**
   - From the analysis and design of the linear regression model, the loss value obtained by using diamond dataset is 33117030916.51486.

2. **Show the plot comparing the predictions vs the actual test data.**
   - Plot comparing predictions vs actual test data:



3. **Discuss the difference between Linear and Ridge regressions. What is the main motivation for using l2 regularization?**

Linear Regression:

- The method examines best-fit straight line relationships in an effort to explain the link between dependent and independent variables.
- The rule of least squares is used, and the error terms are assumed to have a normal distribution.

Ridge Regression:

- This method aims to lessen the impact of strongly correlated independent variables.
- By this, coefficients are shrunk to zero, and a penalty term is introduced to the least squares approach to prevent the over-fitting issue.

Differences:

Name of Difference	Linear Regression	Ridge Regression
The Problem of Multicollinearity	a) Assumes multicollinearity is not present among independent variables. b) As a result, it's possible to get incorrect findings.	a) Multicollinearity problems can be solved by reducing the coefficients to zero.
The Problem of Regularization	a) Regularization term is not present.	a) Helps to avoid overfitting issue by including a regularization component in the objective function.

Comparison of bias and variance tradeoffs	a) Has high variance and low bias.	a) Has low variance and has a bit of high bias.
Problem of Coefficients Interpretation	a) Interpretation is easy, since they show the variation in the dependent variable brought on by a change of one unit in the independent variable.	a) Interpretation is not easy, as the coefficients are modified to lessen the issue of multicollinearity and overfitting.
Complexity Problem	a) Implementation is easy and design is simple.	a) The model is slightly more challenging than linear regression since parameter selection is necessary.

Main Motivation for using L2 Regularization:

- The main motivation for using L2 regularization is to decrease the chances of overfitting a model.
- In essence, reducing the issue of overfitting can increase the performance of linear regression.
- In this, values of small coefficients are obtained by adding term of penalty to the function of cost.

4. Discuss the benefits/drawbacks of using a Ridge Regression model.

Benefits:

a) One of the key benefits of the Ridge Regression Method is that it prevents the model from becoming overfit.
b) In comparison to least squares regression, the model can produce findings with a reduced test mean square error when multicollinearity is present.
c) The model works well when there is a sizable multivariate dataset (number of observations is fewer than number of predictions).
d) There is no requirement for unbiased estimators.

Drawbacks:

a) One of the biggest drawbacks is that the final model includes all the predictors.
b) When there are numerous predictor factors, it becomes challenging to interpret coefficients.
c) This technique has no ability to perform feature selection.
d) There's a need to take a proper value for the regularization parameter; otherwise, it may lead to overfitting and underfitting problems.

Contributions:

Team Member	Assignment Part	Contribution (%)
Sai Rohit Uddagiri	Data pre-processing, Data Visualization and report	50
Bhavana Yetinthala	Model building and report	50

References:

- **https://www.geeksforgeeks.org/advantages-and-disadvantages-of-logistic-regression/**
- **https://www.kaggle.com/discussions/general/352871**
- **https://www.ib-net.org/benchmarking-methodologies/performance-benchmarking/statistical-techniques/**
- **https://www.geeksforgeeks.org/ml-advantages-and-disadvantages-of-linear-regression/**