Report

Part 1

Penguin data set:

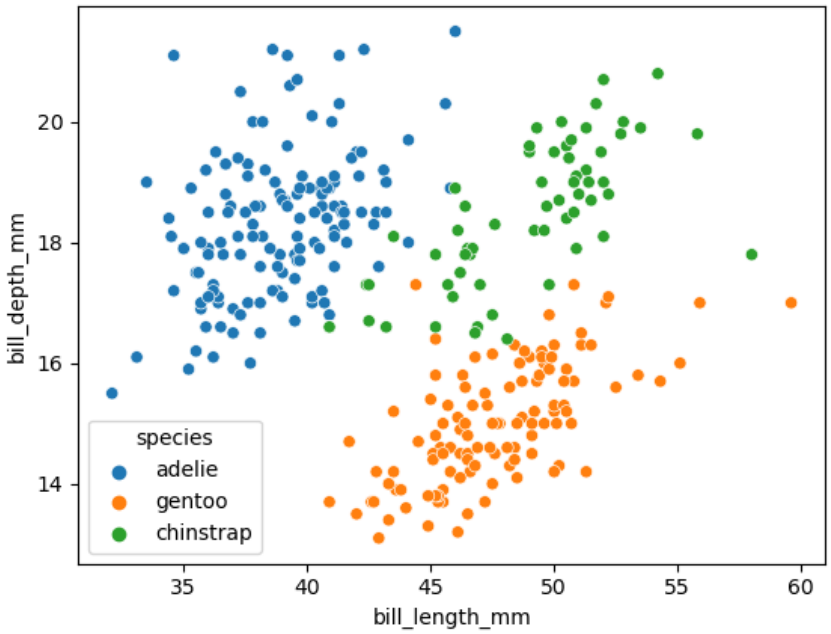
i) Penguin data set has 344 rows and 10 columns i.e., features, which includes both numerical and categorical. List of features and its short description in data set.
Species: Type of species penguin is classified into i.e., Adelie, Chinstrap, Gentoo
Island: On which island the penguin is found i.e., Torgersen, Biscoe, Dream
Calorie Requirement: Number of calories required by a penguin.
Average sleep duration: Duration of sleep.
Bill_length_mm : Bill length of penguin in millimeters.
Bill_depth_mm: Bill depth of penguin in millimeters.
Flipper_length_mm : Flipper length of penguin in millimeters.
Body_mass_g: mass of penguin in grams.
Gender: Sex of penguin
Year: Year in which data is collected.

Domain: Mostly this dataset falls in the domain of biology but can be useful in some other domains like animal statistics, etc.

ii) **Handled mismatched string formats** : merged all the duplicates for instance, 'Adelie' is merged with 'ADELIE'.
Imputing null values: 39 rows out them have null values which is more than 10 percent of dataset, so it is important to impute some of the values instead of omitting all the values. In the start I thought to impute the numerical values using mode/median/mean but after looking at rows with null values these methods seem good according to specific species categories but not as a whole dataset, so I have used interpolate method to fill all the numerical values. Additionally, I have filled species and island column of some of the rows using mode. For instance, Torgersen Island has only Adelie species.
Removing null values: This way I brought down the number from 39 to 25 rows and omitted them as nothing seemed appropriate to impute values.
Handling Outliers: On the other hand, I have also removed some outliers using boxplots which brought the dataset further down to 310 rows.
Encoding Categorical Values: Using one hot encoding I have encoded categorical values i.e., Gender, Species and Island
Normalization: Normalized all the columns manually.
Removing columns: Using correlation matrix I have removed the features which are not needed (my target is gender)

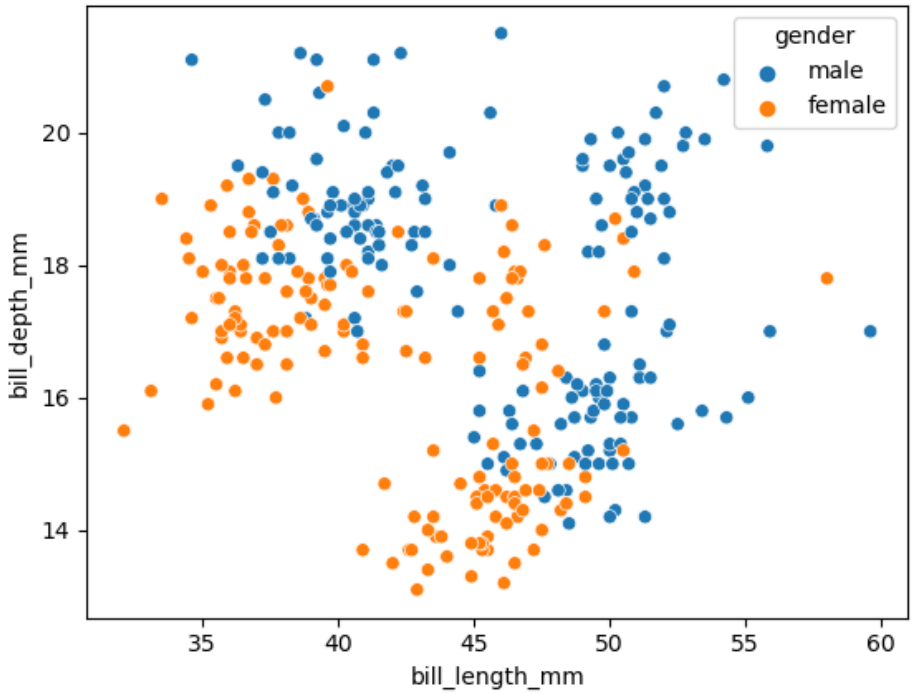
iii) Graphs and their description:

a)



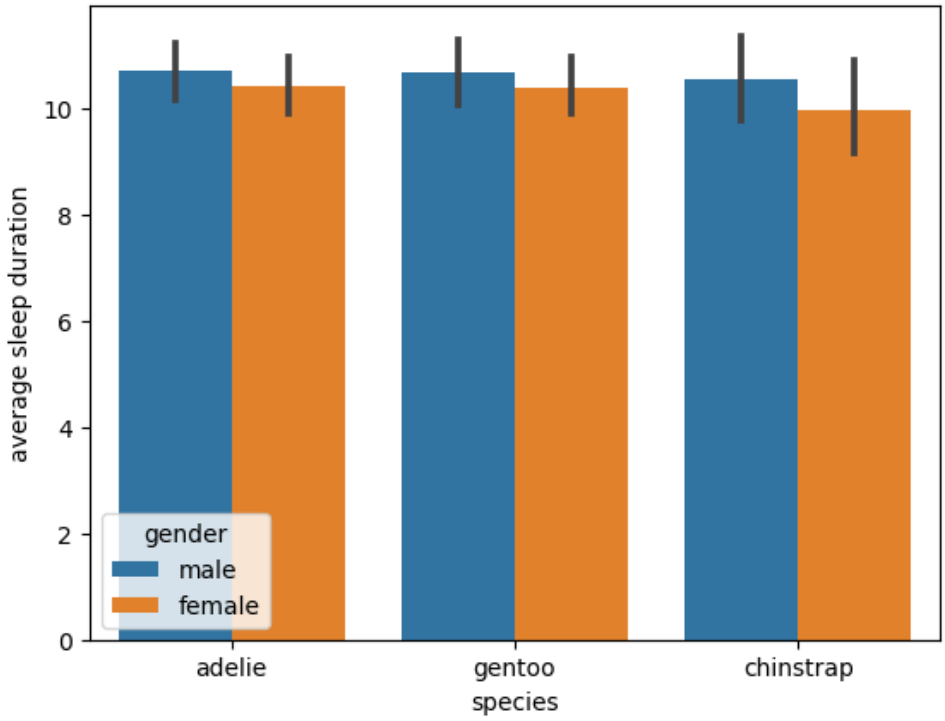
From this graph we can clearly observe the formations of species clusters when bill length and depth is considered.

b)



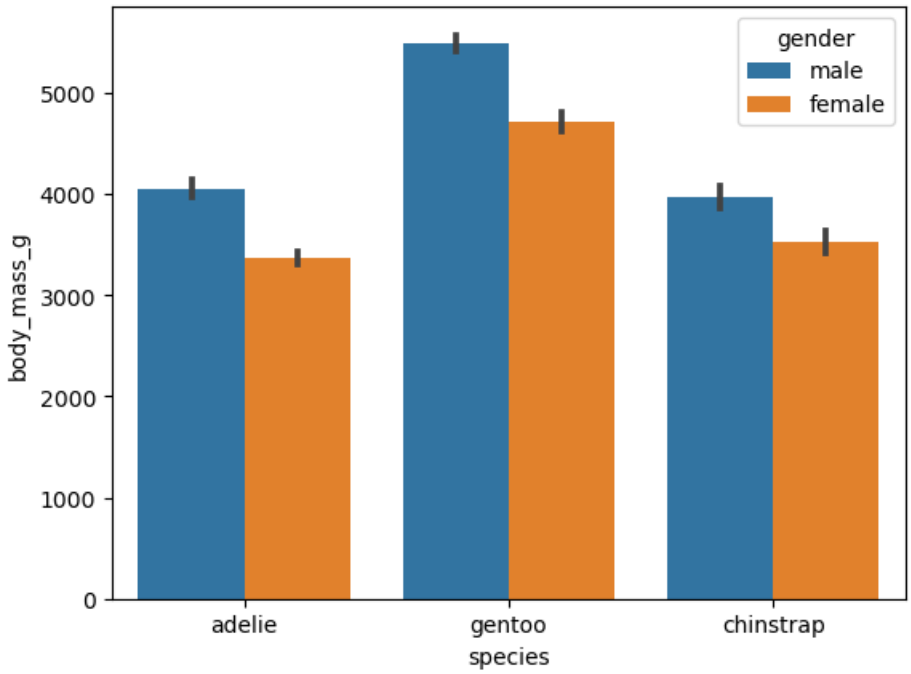
The above graph is plotted for gender with same features. Here, there are clusters formed but they are not as good when compared to species clusters so, we can say gender is correlated with bill length and depth but with less coefficient.

c)



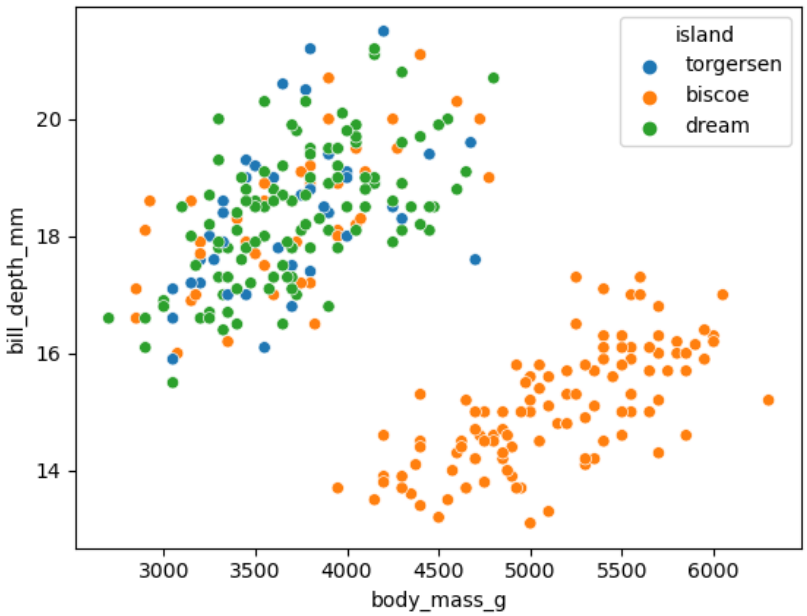
From the above graph we can clearly infer that average sleep duration is a bit shorter for female penguins.

d)



The same applies to weight i.e., generally, female weighs a bit shorter than male in all species categories.

e)



Clusters with specific set of body mass and bill depth are only on Biscoe Island.

Diamond data set:

i) Diamond data set is huge dataset compared to penguins which has 53940 rows and 13 features comprised of both numerical and categorical data.

Carat : Feature representing the weight of diamond
Cut: Says how good the cut of diamond is i.e., Ideal, Good, fair, Very Good, Premium
Color: Grading of diamond color in terms of Alphabets.
Clarity: How clear the diamond is represented by unique code comprising of alphabets and numbers.
Average Us salary: Average salary of the person who bought the diamond (People of United States)
Number of diamonds mined : Total number of diamonds mined in millions.
Depth: depth of diamond (percentage relative to its width) .
Table: width (top) of diamond (percentage relative to its width).
Price: Cost of diamond
x: X dimension of diamond
y: Y dimension of diamond
z: X dimension of diamond

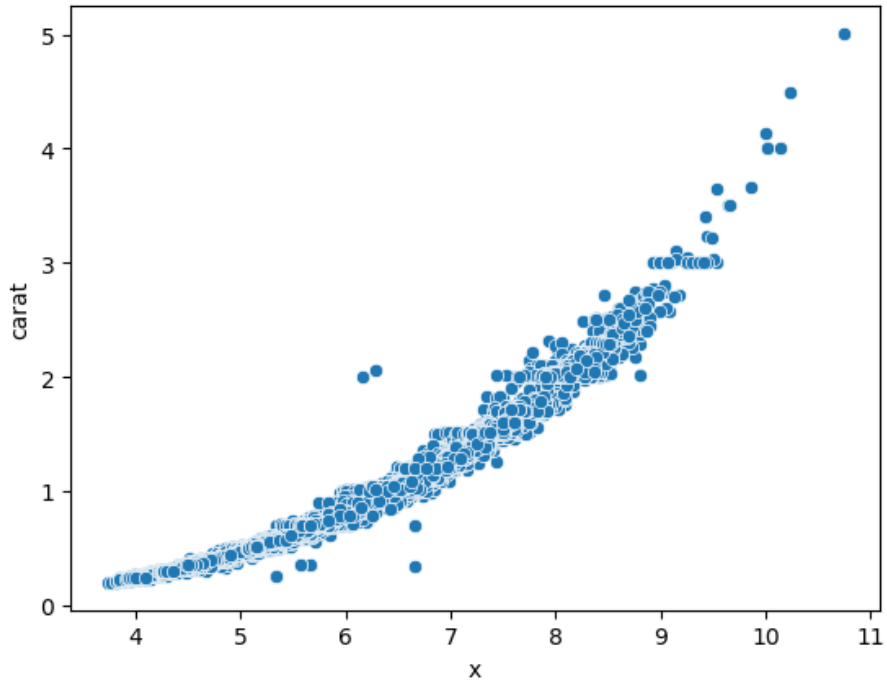
Domain: Mostly falls in Economics domain.

ii) **Removing unnecessary column:** Dropped Unnamed column as it is the replica of index column starting from 1.

Removing improper data and changing data types: After going through information of all the columns the data type is different than expected so investigated furthermore and found that all the columns has data like True, False, Maybe which is irrelevant. Therefore, I have dropped all rows with irrelevant data and changed the data type of each column accordingly.
Mismatched strings are combined: All the data of categorical data with same category is combined as one for example IDEAL, Ideal and ideal are changed to 'ideal'
Dropped the null values: As we are having sufficiently large data compared to number of null values and filling the data seemed inappropriate, I have dropped all of them.
Handled Outliers: I have visualized most of the data with box plots and removed most of the outliers.

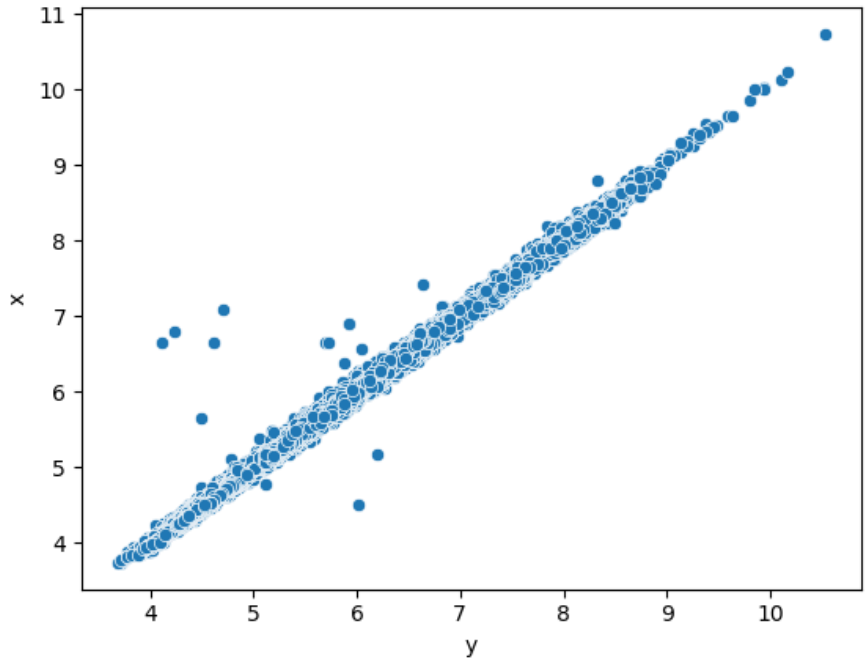
iii) Visualization:

a)



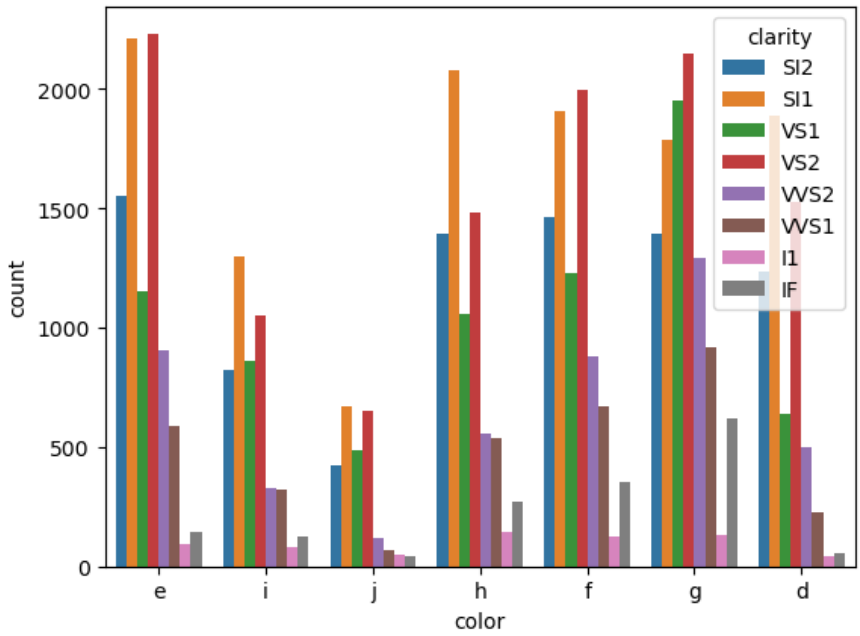
Scatter plot is visualized with carat on y-axis and dimension x on x-axis. From the plot we can clearly identify that carat value increases along with x while it may not be exactly linear it seem to fit a proper curve similar to parabola.

b)



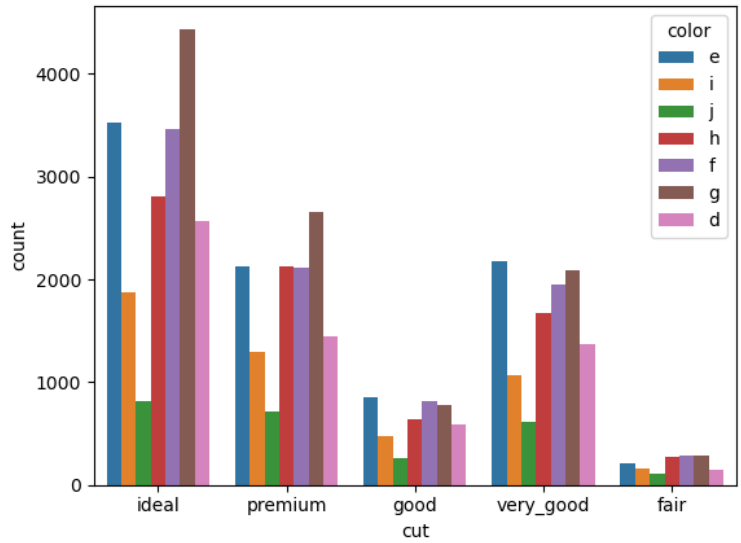
Similarly, a scatter plot is drawn between two dimensions x and y which shows a clear linear relationship among the dimensions with some minimal outliers.

c)



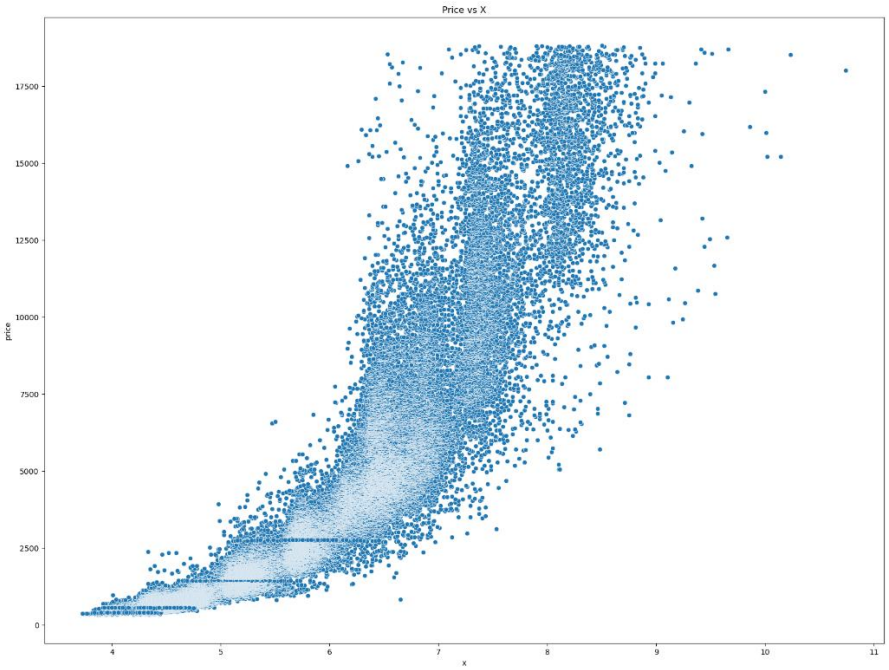
A count plot from seaborn library is used to count the clarity of diamonds among various types of color grading. In all the color grades I1 and IF clarity seem to have less number of diamonds while SI2 and VS2 being the highest.

d)



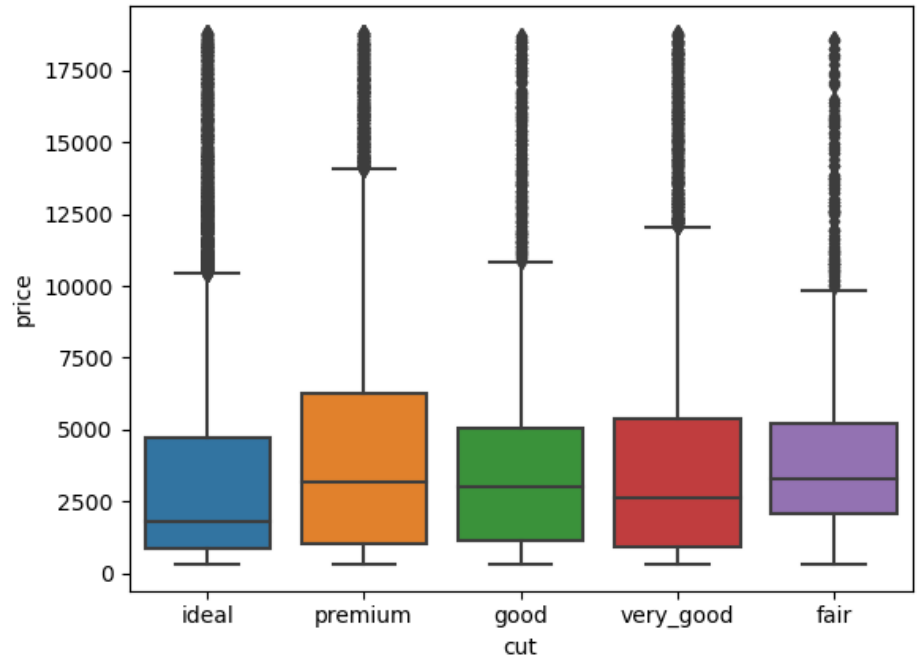
Similarly, count plot of cut is visualized with color as category among each type of cut. Distribution of color seem to be similar when comparing each cut category and the whole dataset with some minor exceptions i.e., very_good cut has a greater number of 'e' graded diamonds while 'g' being the highest for all other categories.

e)



Scatter plot between price and X is visualized and the increasing relation can be seen clearly.

f)

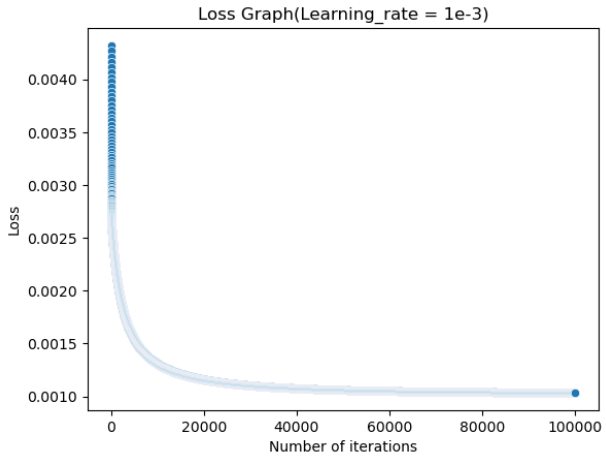


Plotted Box plot of price with various cut categories. All the cut categories seem to have approximately the same statistics i.e., median and interquartile range.

Part 2

i) Best accuracy: 0.919355
The corresponding weights are stored in the pickle file

ii)



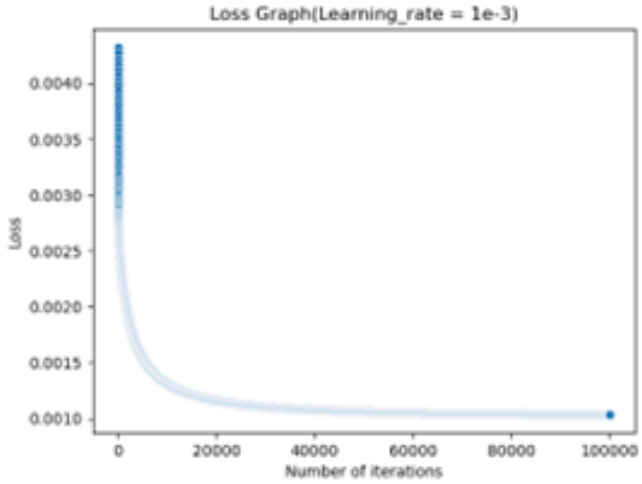
I have visualized the variation of loss along with number of iterations provided learning rate = 1e-3

Brief Analysis: Initially the loss of model is decreasing drastically which means model is learning very fast to a certain number of iterations and then it starts to flatten out and may overfit the data with further iterations.

iii) **Set up1:**

Learning rate = 1e-3 and number of iterations = 100000

Accuracy = 0.919355

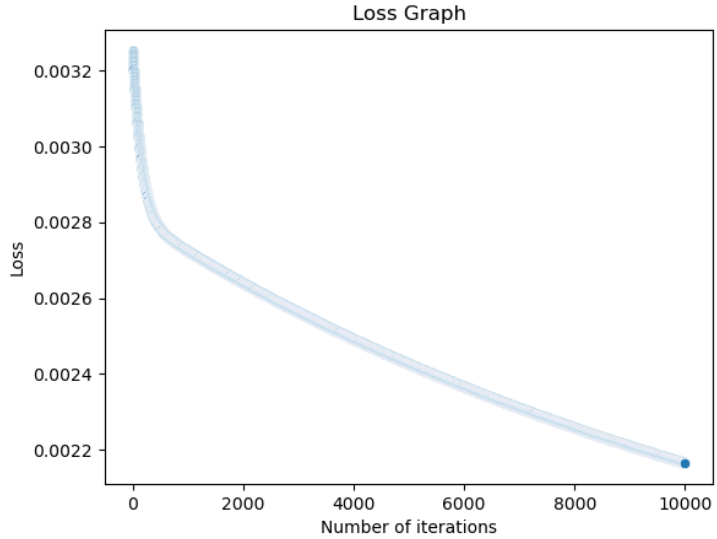


The slope of end point of the graph seems to flatten out which says the model is converged

Set up2:

Learning rate = 1e-5 and number of iterations = 10000

Accuracy = 0.8387

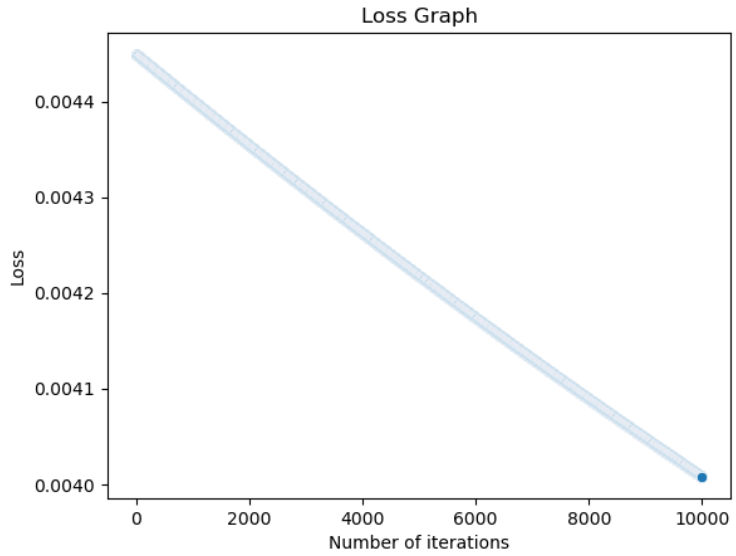


With decreasing the learning rate of the curve and number of iterations the accuracy of model is dropped a bit. The slope of the end point of the graph suggests that the model is not yet converged as it is leaning slowly.

Set up3:

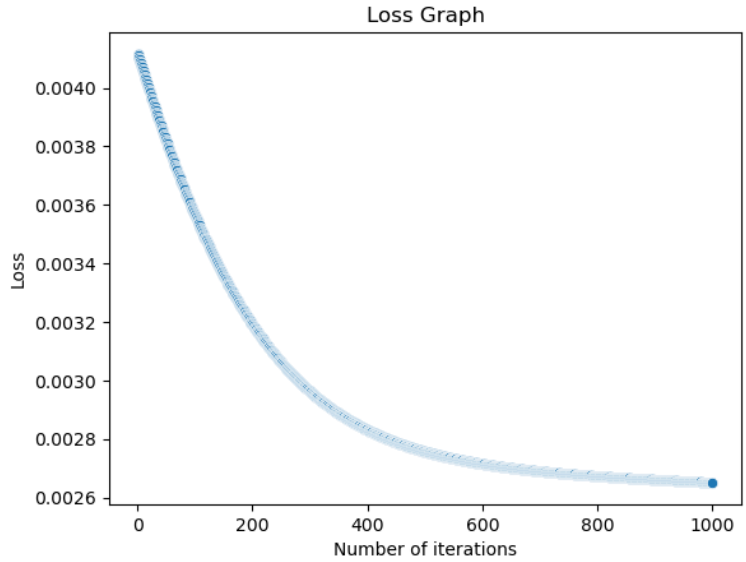
Learning rate = 1e-10 and number of iterations = 10000

Accuracy = 0.53226



Now with the same number of iterations and decreased learning rate reduced the accuracy of model drastically and the graph also suggests that the model is not yet converged.

Set up4:
Learning rate = 1e-5 and number of iterations = 1000
Accuracy = 0.629



With the learning rate remaining same and a smaller number of iterations the accuracy of the model still dropped as it is not iterated enough to converge.

iv) Benefits: The model is both simple and fast to implement. Additionally, it is very effective if our data is linearly separable.
Drawbacks: Does not perform very well on non-linearly separable data and is sensitive to outliers. On the other hand, the model will not work properly if there are many features and not processed properly.

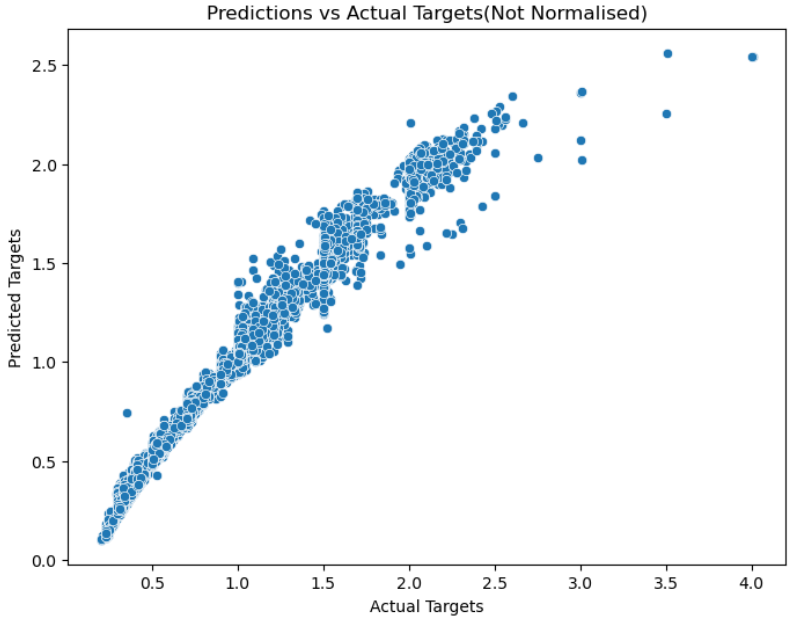
Part 3:

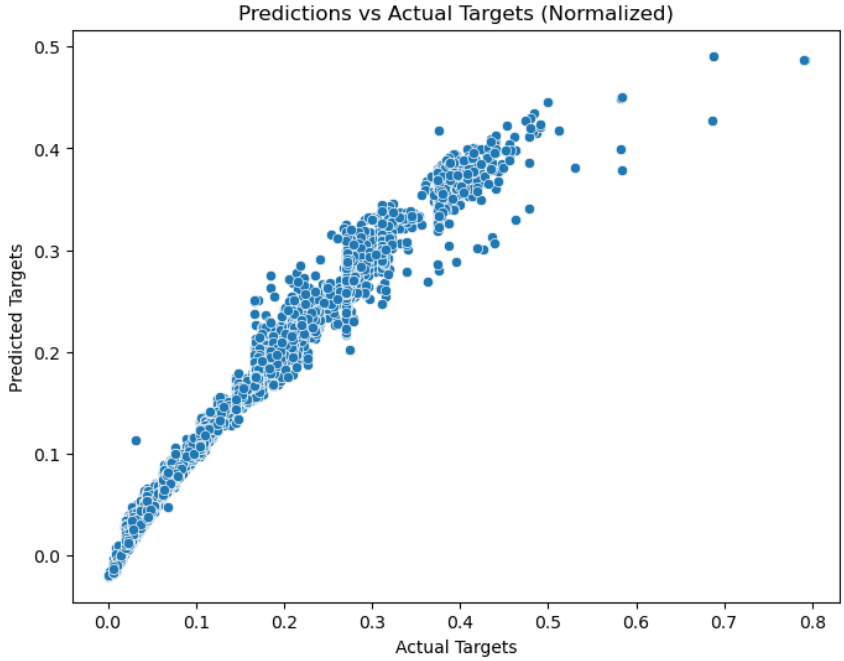
i) We have used the previously processed data set (diamond.csv)

ii) Loss value = 0.0002513(Mean Squared Loss)

iii)

- I have done all the processes after normalizing the data, after prediction I have included graphs with both normalization and without normalization. Removed normalization using original_value = normalized_value *(max_value – min_value) + min_value
- As I have included more than 2 features it is not possible to visualize the graph between data points , targets, and predictions I have instead visualized a scatter plot between predicted values versus actual values.





iv) OLS is simple and straightforward way to implement and get the corresponding weights of input features and will output best linear unbiased estimators. On the other hand, the performance is really poor if there are many dependent variables and single independent variable(cannot capture complex relation between data). Additionally, OLS is sensitive to outliers.

v)

Benefits

- Linear Regression is an easy to implement and efficient way to predict continuous data if the relationships are linear.
- Coefficients of features are easy to interpret

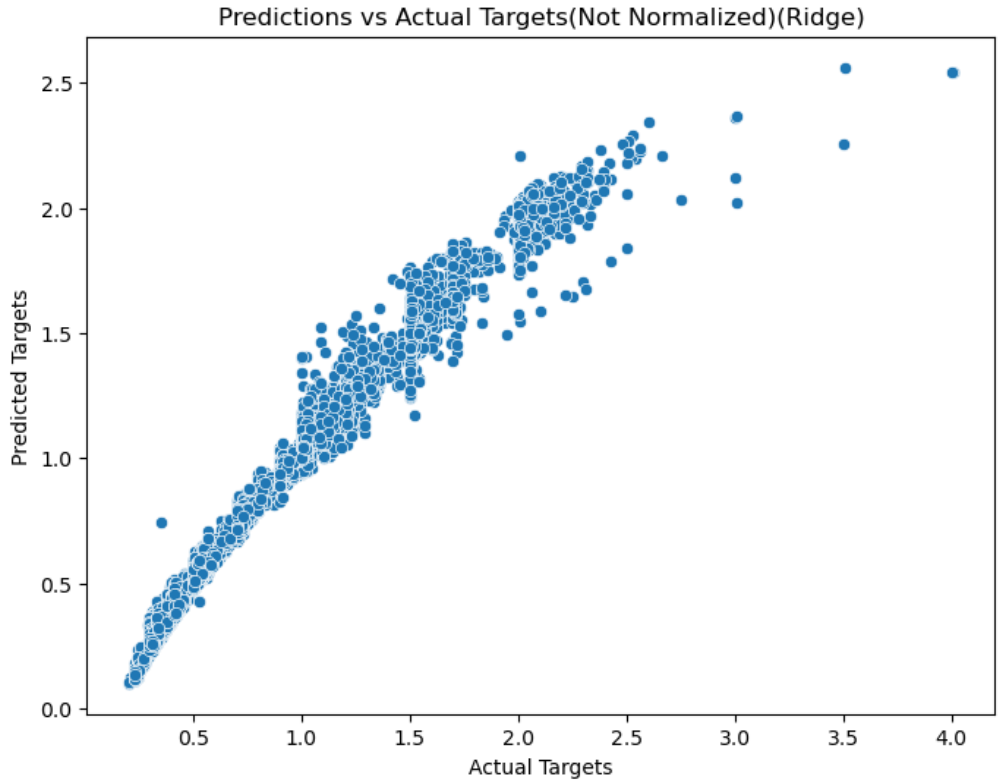
Drawbacks

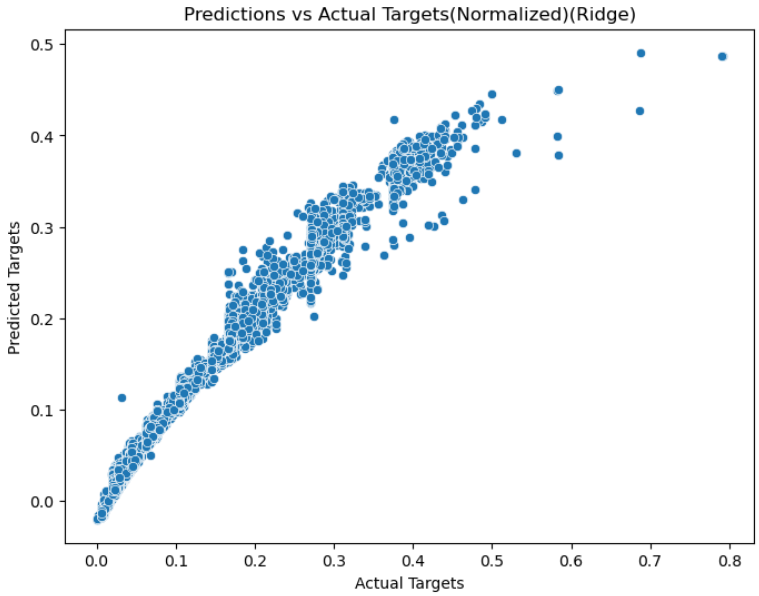
- It may not work properly with non-linear relationships.
- It may overfit the data with too many features as input (performs poorly on unseen data).

Part 4:

i) Ridge loss = 0.01119109
Mean squared loss = 0.0002514

ii) Similar to linear regression I have included graphs with both normalization and without normalization.





iii)

- Linear Regression directly predicts the target on linear bases while Ridge regression includes L2 regularization term.
- Both regressions are sensitive to outliers, but the ridge regression can have a bit less sensitivity because of L2 Regularization
- Ridge regression can have balance overfitting by adding a penalty while Linear Regression cannot(Linear regression tend to have high variance and less bias).

iv) Benefits:

- Overfitting is controlled by shrinking the estimated coefficients to zero using L2 regularization and is less likely to fit the noisy data.
- Additionally, L2 Regularization helps to balance bias and variance.

Drawbacks:

- Choosing Hyperparameter (Lambda) is generally done by cross-validation which adds to the complexity of the model.
- Like Linear Regression, Ridge also assumes that relationship between features and target is linear.
- Does not perform very well on unscaled data.
- Like Linear regression feature engineering is needed.

References

https://numpy.org/doc/stable/reference/generated/numpy.ones.html

https://numpy.org/doc/stable/reference/generated/numpy.zeros.html

https://numpy.org/doc/stable/reference/generated/numpy.concatenate.html

https://practicaldatascience.co.uk/machine-learning/how-to-save-and-load-machine-learning-models-using-pickle

ML Class Slides and lectures