ASSIGNMENT 1

Full Name	UBIT Name	UB Number
Dharma Acha	Dharmaac	50511275
Adarsh Reddy Bandaru	adarshre	50527208

Part 1

Data Cleaning and EDA for Penguins Dataset

Domain:

The dataset talks about penguins and their physical characteristics, behavioral aspects and geographical aspects.

Number of samples and features in the dataset:

Number of Samples: 344
Number of Features: 10

Features Overview:

Feature	**Feature Type**	**Missing values**
species	Categorical	11
island	Categorical	10
calorie requirement	Numerical	0

average sleep duration	Numerical	0
bill_length_mm	Numerical	7
bill_depth_mm	Numerical	11
flipper_length_mm	Numerical	8
body_mass_g	Numerical	5
gender	Categorical	17
year	Categorical	2

Basic Statistics of Numerical Columns:

Feature	Mean	Median	Min	Max	Standard Deviation
calorie requirement	5270.00	5106.5	3504.00	7197.00	1067.95
average sleep duration	10.44	10.00	7.00	14.00	2.26
bill_length_mm	45.49	45.10	32.10	124.30	10.81
bill_depth_mm	18.01	17.30	33.10	127.26	9.24
flipper_length_mm	197.76	197.00	10.00	231.00	27.76
body_mass_g	4175.46	4050.00	882.00	6300.00	858.71

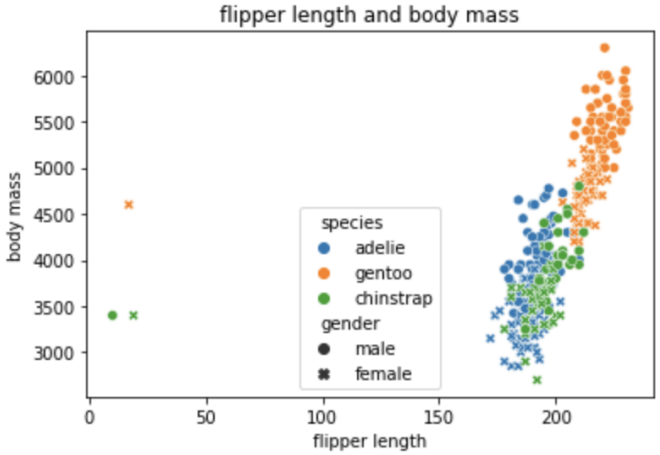
Data Cleaning and preprocessing

Handling mismatched string formats:

When we checked for distinct values in categorical columns, we observed string formatting issues in species, island and gender columns. This column had string casing issues which led to having multiple values for the same entity. We removed it by simply converting all strings to lowercase letters.

Handling Missing values:

First, we have detected rows containing null values in different features. Based on the columns we have missing values, we separated the missing value rows into four separate dataframes to handle them with different approaches and prioritised the retention of data where possible. We have used the below visualisation to fill in missing values with more precision instead of blindly filling it with mean and median.



The four different dataframes created are:

- Rows containing information on species and body_mass_g, but do not have gender columns populated.
- Rows containing information on species and gender.
- Rows that have null values in the species column
- Rows that can’t be filled in strategically.

Step 1:

We first took rows with species and body_mass_g present and no gender and created conditional statements to fill in missing values of gender. From the above graph, we can see a specific pattern in gender based on species and body mass. The below table consists of the conditions used to fill in missing gender values.

Species	Body_Mass	Gender
adelie	<= 3800	Female
	> 3800	Male
gentoo	<= 5100	Female
	> 5100	Male
chinstrap	<= 3800	Female
	> 3800	Male

Step 2:

To handle rows with missing information from columns such as bill_length, bill_depth, flipper_length, and body_mass, We calculated the mean of these columns for each species and gender and replaced the null entries in these columns with the computed mean values for the corresponding species and gender to maintain accuracy and consistency within species and gender groups.

Step 3:

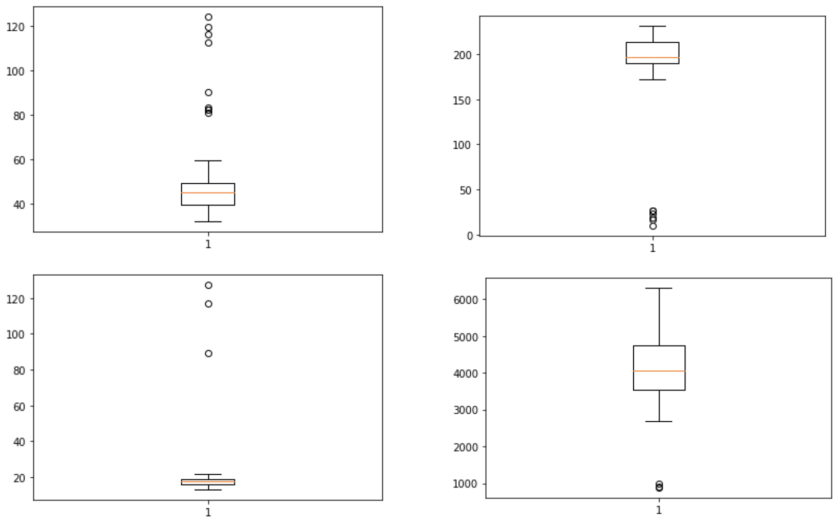
While handling missing values of rows that have null values in species, we checked the body mass and gender and assigned corresponding species names. As Adelie and Chinstrap follow a similar pattern, we only filled in Adelie when their body mass was less than 4500 and Gento otherwise. We avoided Chinstrap as we cannot separate them from Adelie as they have very similar characteristics. We chose Adelie over Chinstrap when filling in as the count of Adelie penguins in the dataset is much higher than Chinstrap.

Step 4:

Lastly, we filled in the remaining 2 rows mode for categorical columns and mean for numerical columns.

Handling Outliers:

We used boxplots to identify outliers in different columns and found out that 'bill_length_mm', 'bill_depth_mm', 'flipper_length_mm', and 'body_mass_g' have outliers in them.

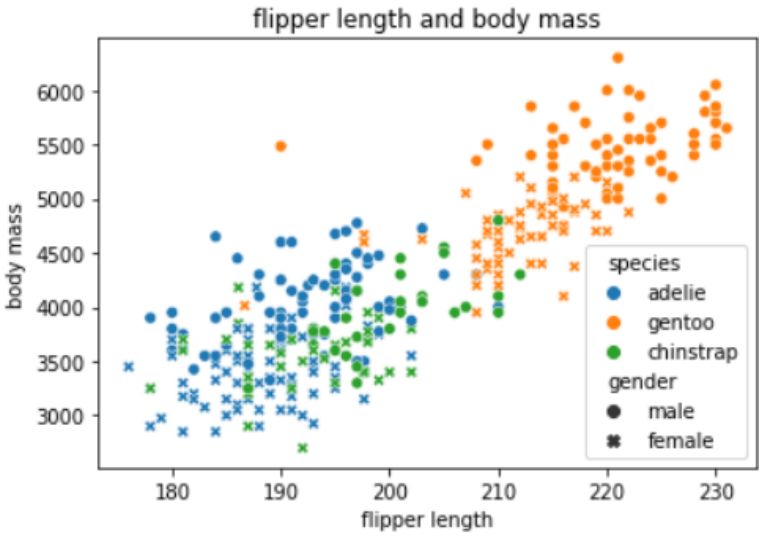


As the outliers were clear, we used the conditional statements to replace them with the mean values of respective columns.

Visualization:

Plot 1:

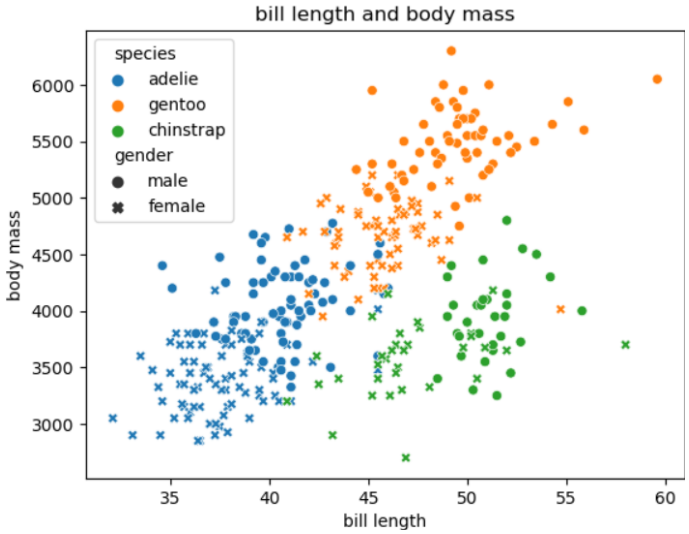
The plot represents a scatter plot between flipper length and body mass of different species based on gender classification.



From the plot we can infer that the species Gentoo which are male have the highest flipper length and body mass when compared with other species. Female Gentoo also have large flipper length and high body mass. Most of the Chinstrap species's flipper length and body mass lies between 190 mm-210 mm and 3000g - 4500g. The species Adelie flipper length and body mass lies between 180mm- 200mm and 3000g-4500g respectively.

Plot 2:

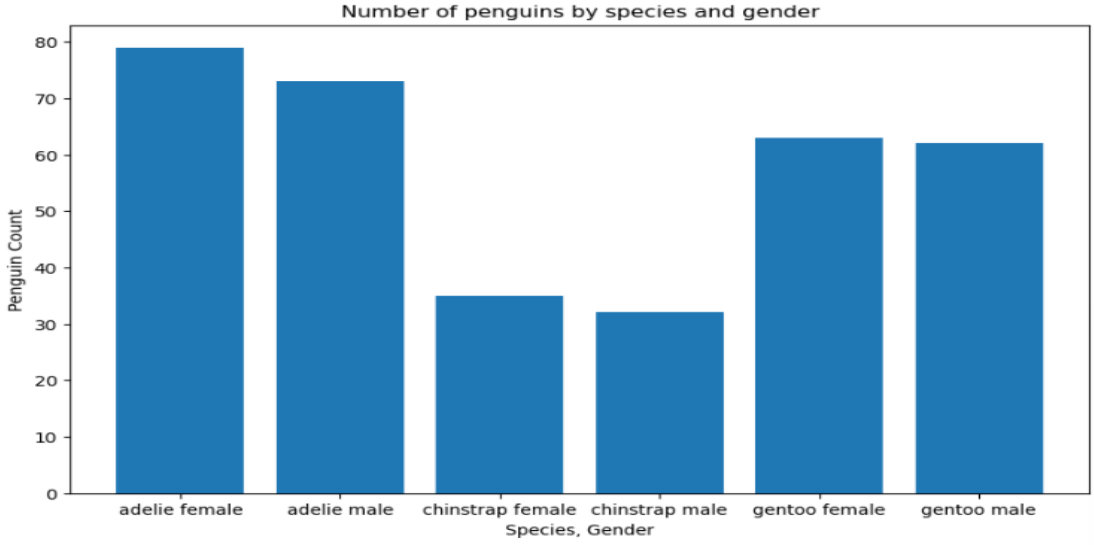
The plot represents a scatter plot between bill length and body mass of different species based on gender classification.



The plot describes the relation between bill length and body mass among all the species based on their gender. From the graph it is clear that bill length and body mass of gentoo is dominating other species whose bill length and body mass ranges from 45-55mm and 5000 -6000 g respectively. On other hand most of the Adelie species penguins has less bill length and body mass whose range varies from 35 -45 mm and 3000 -4500g.

Plot 3:

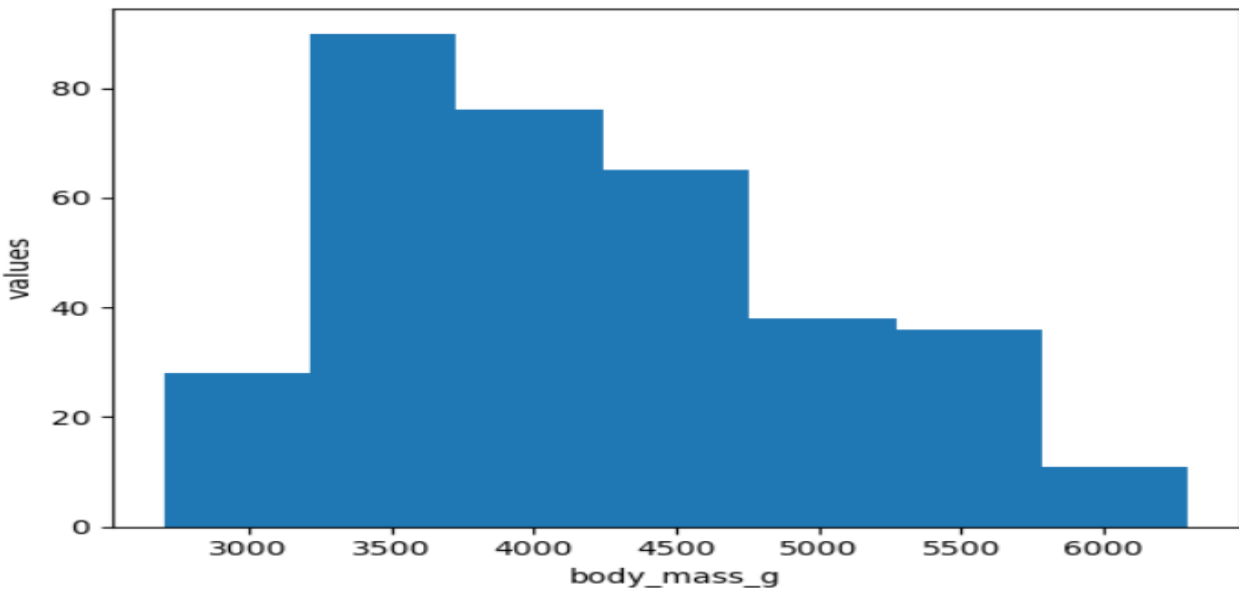
The bar graph describes the total number of penguins by gender for each species.



In every species female count is always greater than male count. Overall the Species named Adelie has more numbers which approximately equals to 79 in female and 75 in male. The count Chinstrap species is half of the Adelie species with respected male and females. Gentoo has more or less equal number of male and female penguins.

Plot 4 :

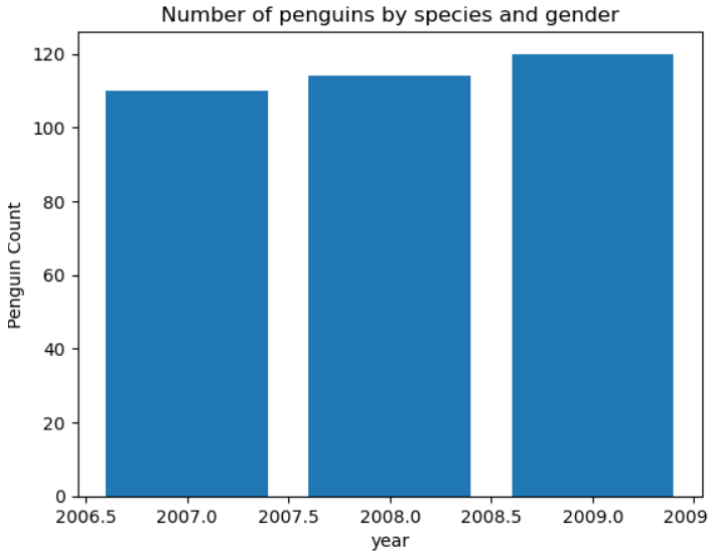
The hist plot of body mass of given penguin species ranges from 2000 to 7000.



The frequency of body mass that ranges from 3000- 4000 is high. That means there are a high number of species whose body mass belongs to that range, whereas the species whose body mass is greater than 5500 are less in number.

Plot 5:

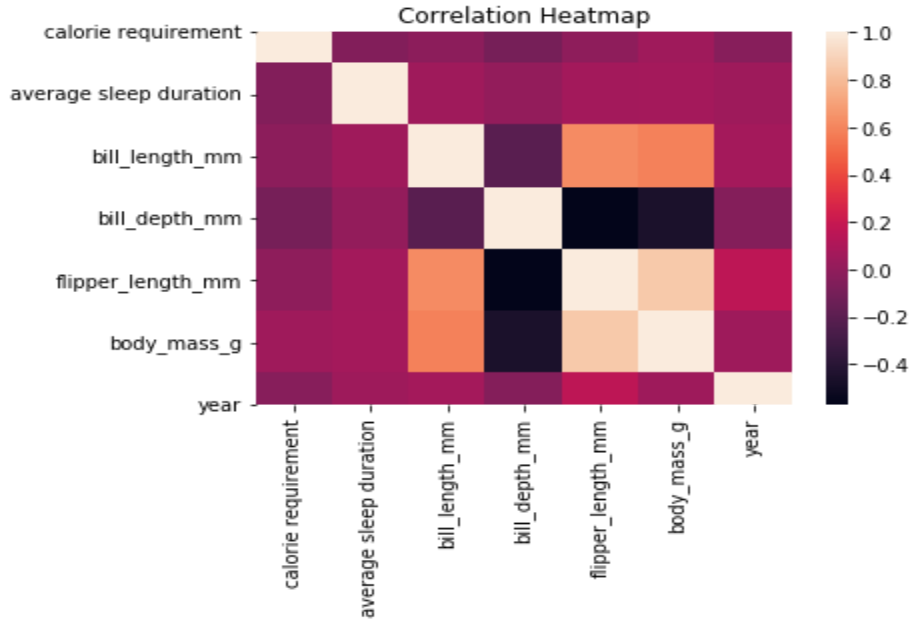
The graph shows count of species by gender in a particular year



Overall there is no greater difference in the number of penguins between the years. All the penguins count lies between 100- 120. As we look deeper we can see a narrow difference between the number of penguins in respective years and the total number of penguins has slightly increased as the year increases. In the year 2009 finally the count of penguins settled at slightly below 120.

Identify uncorrelated or unrelated features:

To understand the dependence of one column on another and to find any linear patterns, we used a correlation heatmap.



We see a strong positive correlation between body mass and flipper length. We also observe that bill length and flipper length have a moderate positive correlation and bill depth and flipper length have a moderate negative correlation. The column's average sleep duration, calorie requirement and year do not have any noticeable correlation.

Normalize non-categorical features

Using a min-max scaler we have normalized all the numerical columns “calorie requirement”, “average sleep duration”, “bill_length_mm”, “bill_depth_mm”, “flipper_length_mm”, “body_mass_g”. This will be quite useful for increasing the performance of an algorithm that uses gradient descent like logistic regression, linear regression and neural networks.

Converting categorical to the binary column using one hot encoding

We used **pd.get_dummies()** on the categorical columns to convert them into binary columns. The resulting data frame now contains binary values for each column. This processing step will be later useful for making the machine learning model.

Data Cleaning and EDA for Diamonds Dataset

Domain:

The Diamond dataset describes vivid characteristics of diamond such as weight of the diamond in carats, type of cut used for diamond, clarity of diamond that tells about absence of inclusions and blemishes, Total number of diamonds mined. And other characters like length, height, breadth of the diamond

Sample and Feature in the dataset:

Number of Samples: 53941
Number of Features: 12

Features Overview:

Feature	Feature Type	Missing values
carat	Numerical	1510

cut	Categorical	1293
color	Categorical	1512
clarity	Categorical	353
average us salary	Numerical	0
number of diamonds mined	Numerical	0
depth	Numerical	694
table	Numerical	1542
price	Numerical	1583
x	Numerical	1526
y	Numerical	1221
z	Numerical	1433

Basic Statistics of Numerical Columns:

Feature	Mean	Median	Min	Max	Standard Deviation
carat	0.736	0.70	0.20	1.79	0.3766
average us salary	39506.756	39528.00	30000	48999	5485.5500
number of diamonds mined (millions)	2.901.	2.91	0.60	5.20	1.326515
depth	61.883	61.80	60.00	63.90	0.820

table	57.348	57.00	54.00	62.00	1.88
price	1874.58	2401.00	326.00	4000	882.27
x	5.72	5.70	3.73	8.99	1.101
y	5.49	5.71	4.00	7.00	0.802
z	3.49	3.53	2.06	5.00	0.629

Data Cleaning and preprocessing

First we dropped the unnamed column from the dataset.

Handling missing values:

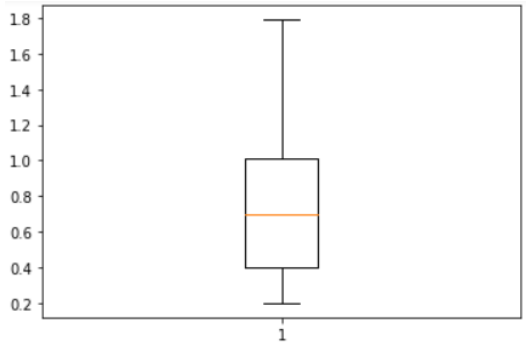
We used the method interpolate to fill in missing values from numerical columns in order to retain the linearity of the dataset. We then replaced missing values in categorical columns with the mode of their respective columns.

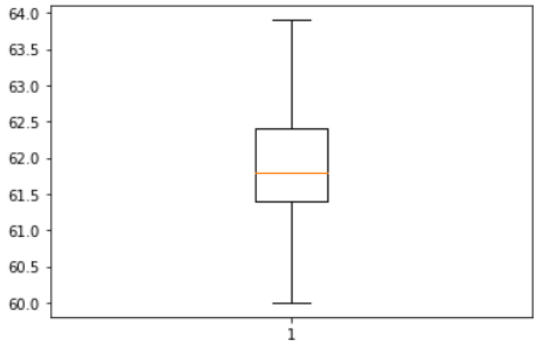
Handling String Formats:

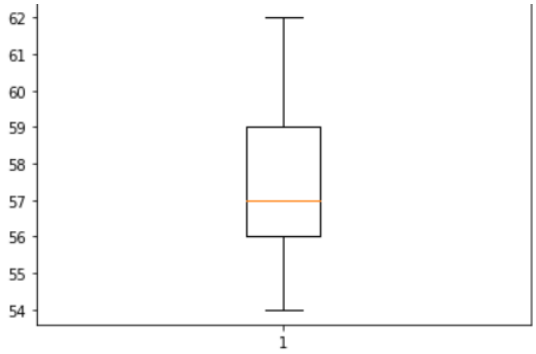
We observed the columns " clarity", "cut", " color" have different string formats. So these strings are formatted to upper case using **str.upper()** method

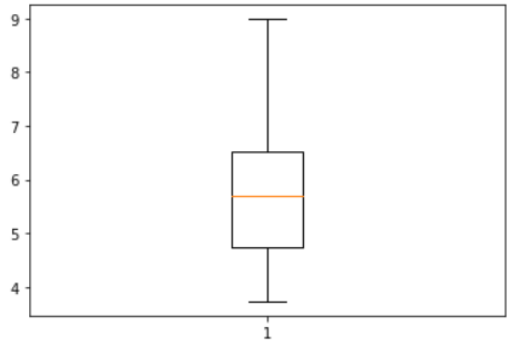
Handling outliers:

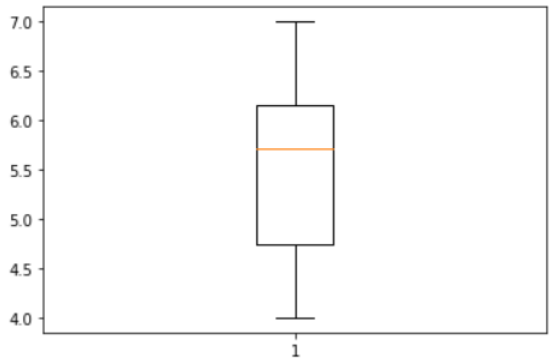
Plotting the box plot for numerical columns we came to know the columns "carat", "depth", "table","price","x","y","z" have outliers. So these outliers are removed using **numpy where condition** by substituting them with mean**.** Here are the results we obtained for respective columns.

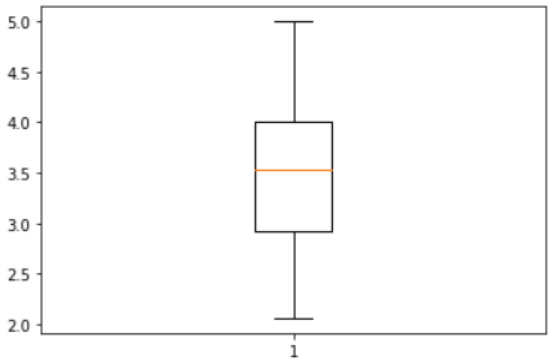






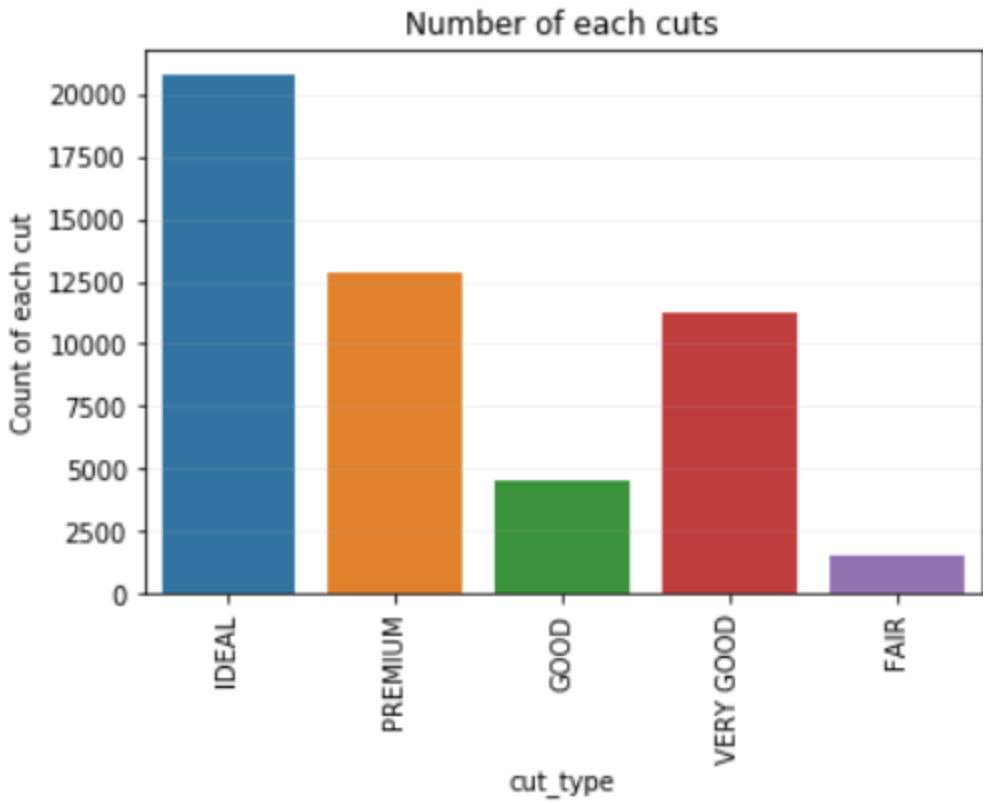




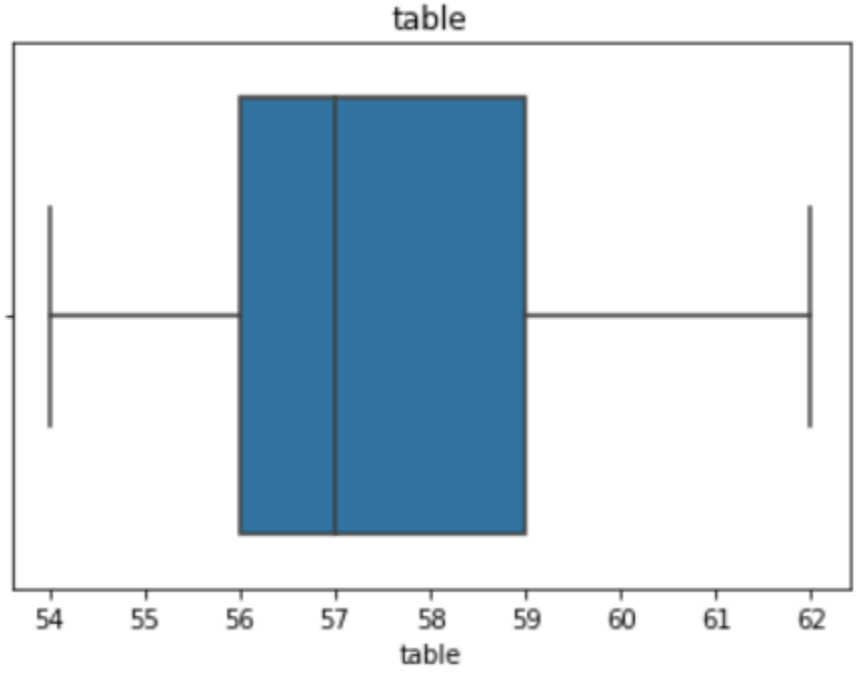


Visualization graphs

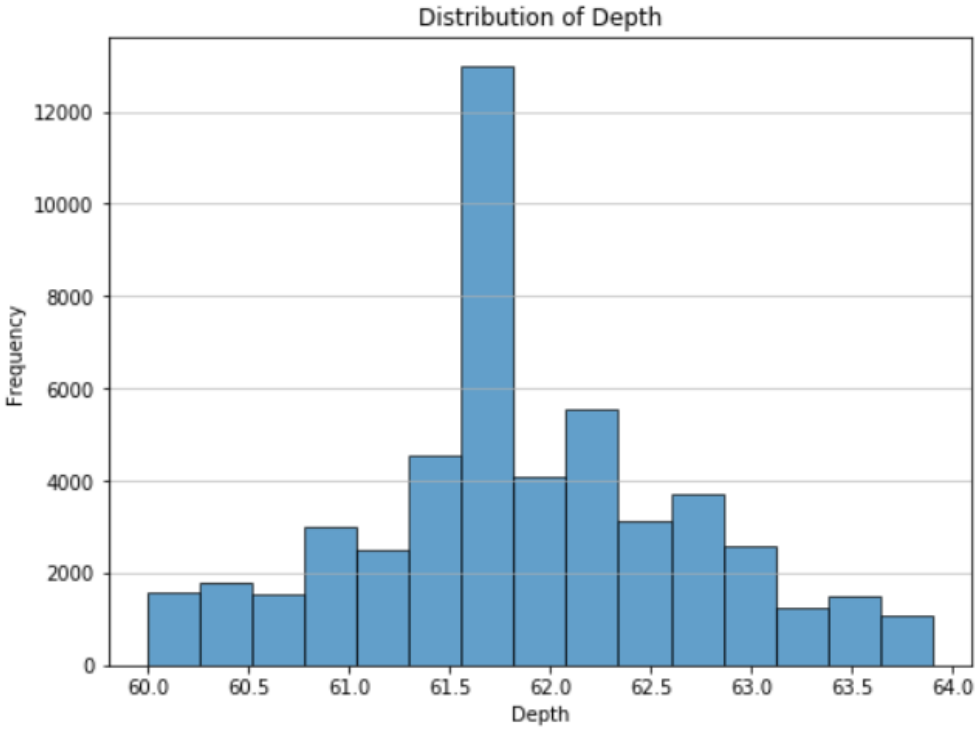
We imported a seaborn library to plot visualization graphs. The plot cut vs count represents the count of each cut type of the diamond.



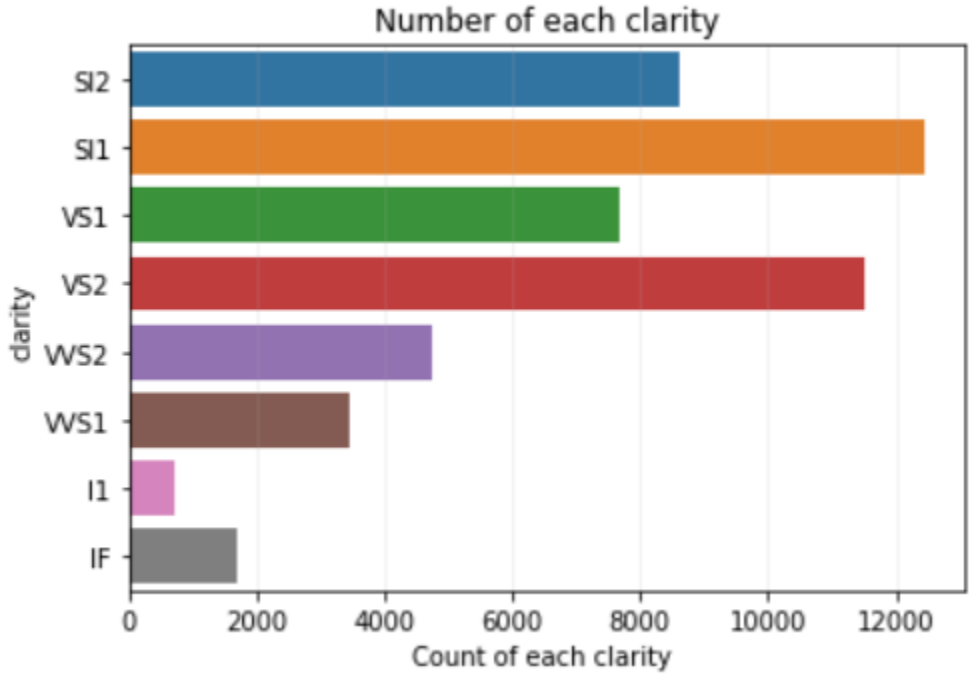
The graph describes an ideal cut that is far greater than other types of cuts, On other hand premium cut is the second highest cut which is equal to 12500. The count of cut named Very good is approximately half of the ideal cut which is slightly less than 12500. As observed from the graph the fair cut is the least one.



The above graph represents the box plot of the table. As we can see there are no outliers. And the median value is settled at 57 and all the values lies between 56 -59.

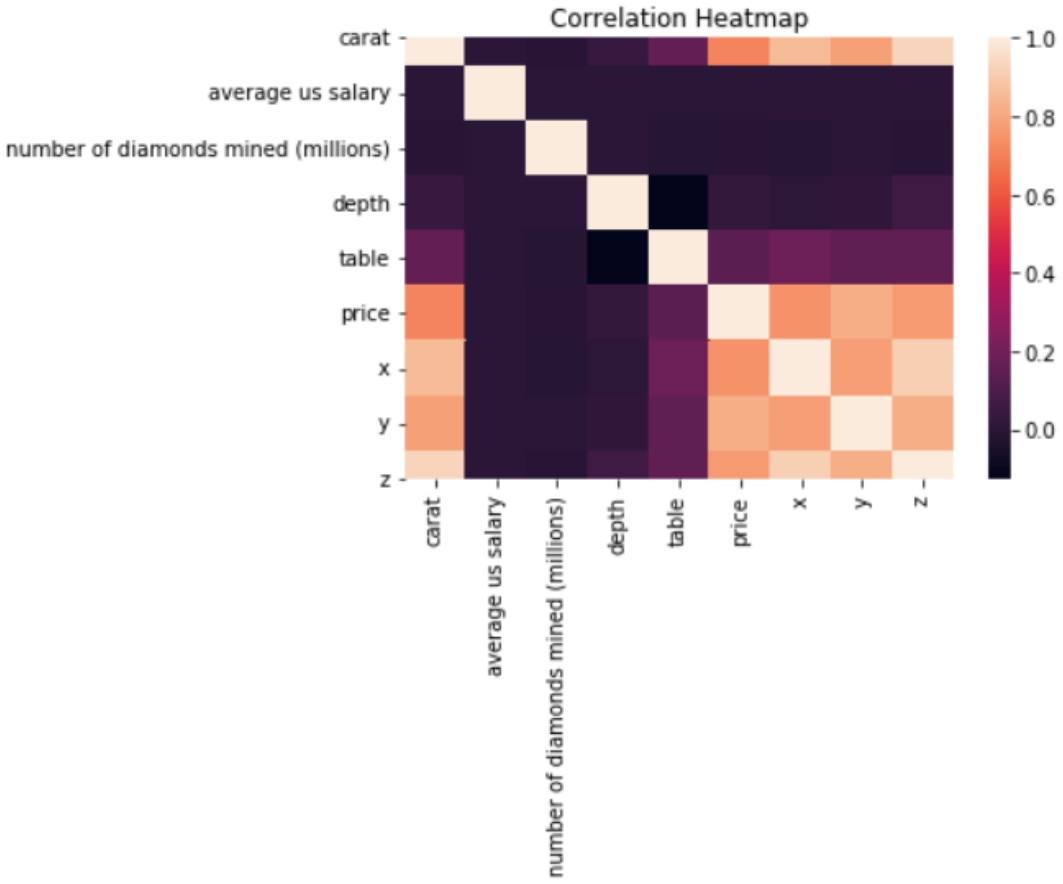


The graph depicts histplot of the depth column. The x axis is the values ranging from 60.0 to 64.0. The Y-axis represents the count of the values. Histograms are divided into bins, each bin represents the frequency of the data points. This Hist Plot is unimodal which consists of 1 peak.



The above graph describes the count of each clarity. The clarity SI 1 is the highest clarity whose count is above 12000. When compared with all other clarities I1 clarity is the least one. Clarity VS2 is the second highest one which equals to slightly greater than 11000. VS1, VVS1, IF are record least count.

Heatmap is used to know the dependency of one column with other columns.



From the above correlation heat map we can observe there is strong positive correlation between carat and other columns. And we see that there is moderate positive correlation between carat and price columns. Also there is no correlation between some of the columns for example between number of diamonds mined and y data points.

One-hot encoding

It is a technique which is used to convert categorical values to numerical format. The categorical columns in the dataset are “cut”, “color”,”clarity”. The method **pd.get_dummies()** is used to convert into numerical values. These numerical values are generally binary.

Normalization:

It is a data processing technique which is used to scale and transform the data into common range. This technique scales numerical columns “depth”,”table”,”price”,”x”,”y”,”z” by using min-max scaling.

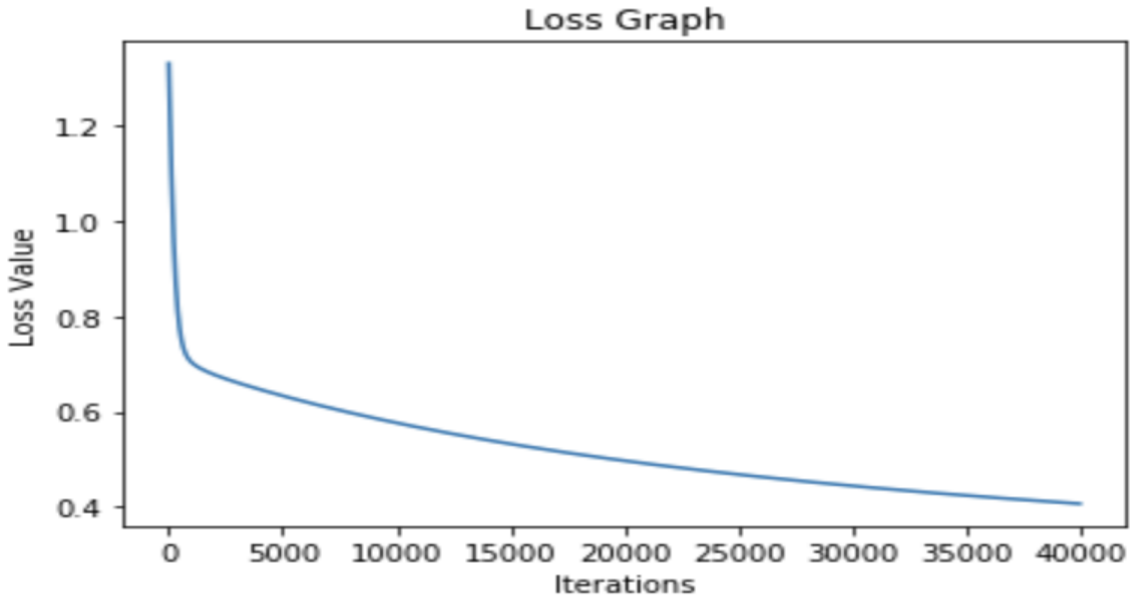
Part 2 - Logistic Regression Implementation

We have chosen Gender_Male as our target variable. Then we split the X and Y dataset into train and test in 80:20 ratio. Then we used the formulas and code structure provided to create a model class for logistic regression. We also added bias by concatenating 1’s at the end of X.

The best accuracy after trying out multiple setups is **86.95%**.

Loss Graph of the model with best parameters (iterations:40000, learning rate:0.004)

The below loss graph follows a decreasing loss indicating that the model adjusted the weights of the model in order to reduce the difference between actual values and predicted values. We see that by the end of iterations the loss value reached the lowest point and shows the model is learning from the loss.

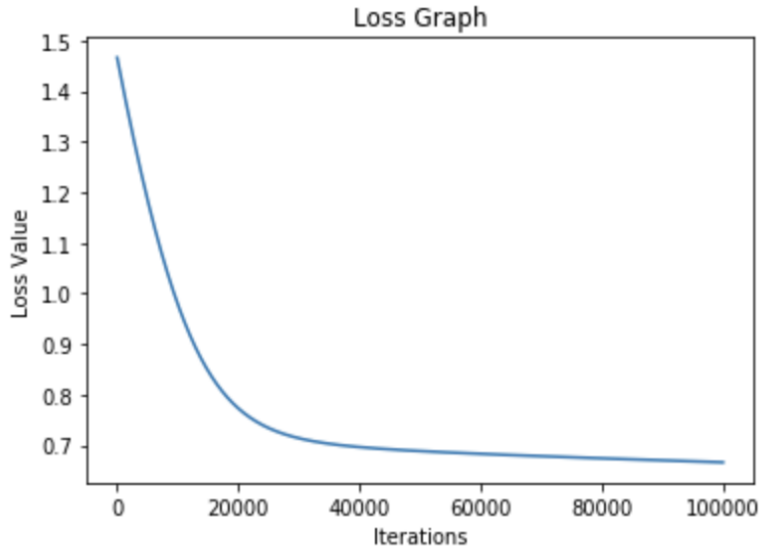


Testing with different parameters:

I started from a learning rate of 0.0001 and iterations of 100000 and kept increasing the learning rate and decreasing the iterations.

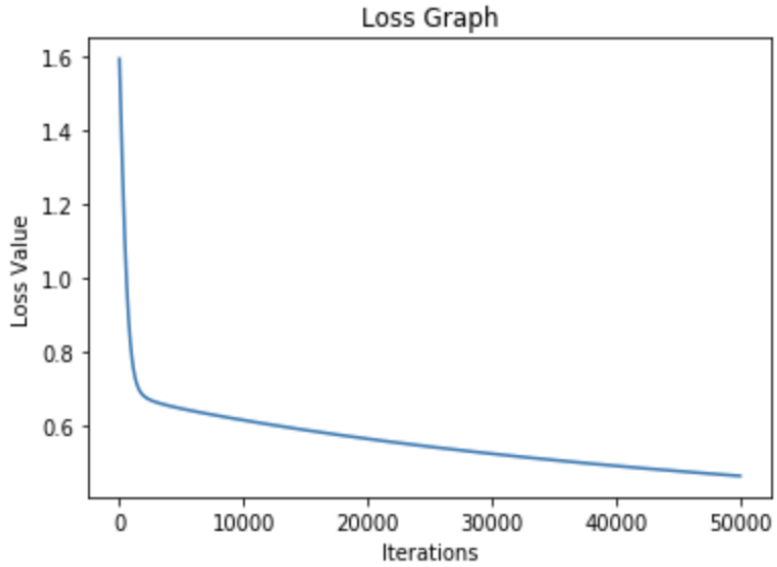
- Model 1 setup (learning_rate: 0.0001, iterations: 100000):

```
Accuracy:  52.17391304347826
```



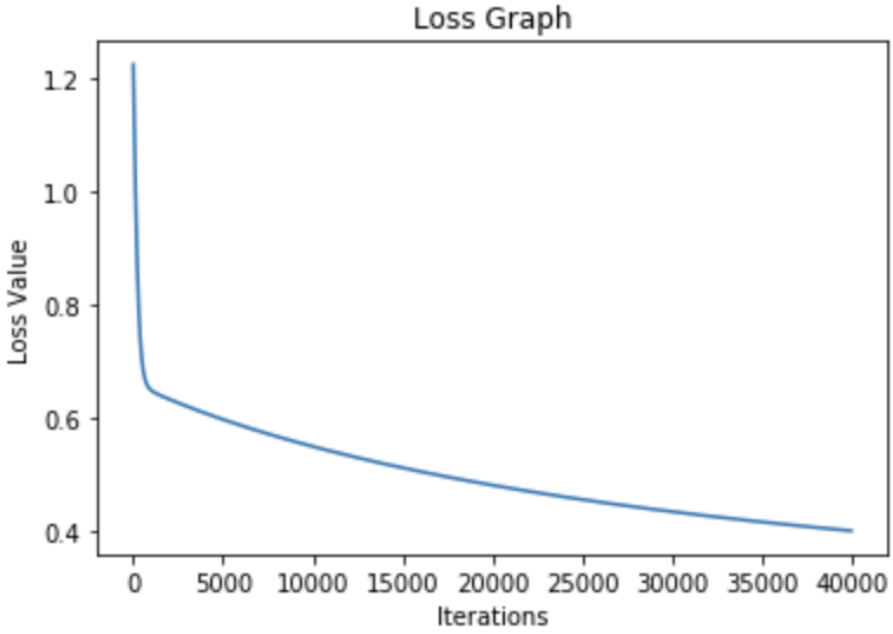
- Model 2 setup (learning_rate: 0.002, iterations: 50000):

```
Accuracy:  75.36231884057972
```



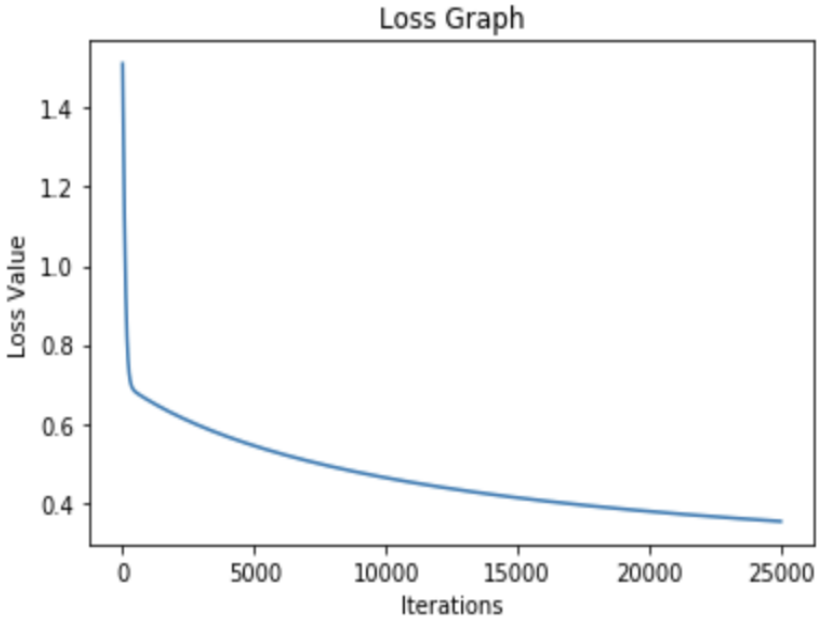
- Model 3 setup (learning_rate: 0.004, iterations: 40000):

```
Accuracy:  86.95652173913044
```



- Model 4 setup (learning_rate: 0.01, iterations: 25000):

```
Accuracy:  84.05797101449275
```



Observation:

We see that as we increased the learning rate, the curve got steeper at the start showing the model adjusted the weights aggressively at the start. But once we increased the learning rate to 0.01, the accuracy decreased. Increasing it further may cause overfitting.

Benefits of using logistic regression:

1.Logistic regression is simple and easy to understand
2. It performs well for small datasets and works well with penguins dataset for the same reason.
3. Since it is a quick and simple method it requires less computational power.

Drawbacks:

1. Sensitive to outliers which affects model's coefficients.
2. The basic idea of logistic regression is used only for implementation of binary classification
3. It is a linear model. So it cannot handle non-linearity between independent and dependent variables

Part 3 - Linear Regression Implementation

We have used the diamonds dataset for analysis and model building of part 3.

Domain:

The Diamond dataset describes vivid characteristics of diamond such as weight of the diamond in carats, type of cut used for diamond, clarity of diamond that tells about absence of inclusions and blemishes, Total number of diamonds mined. And other characters like length, height, breadth of the diamond

Sample and Feature in the dataset:

Number of Samples: 53941
Number of Features: 12

Features Overview:

Feature	Feature Type	Missing values
carat	Numerical	1510
cut	Categorical	1293
color	Categorical	1512
clarity	Categorical	353
average us salary	Numerical	0
number of diamonds mined	Numerical	0
depth	Numerical	694
table	Numerical	1542
price	Numerical	1583
x	Numerical	1526
y	Numerical	1221
z	Numerical	1433

Basic Statistics of Numerical Columns:

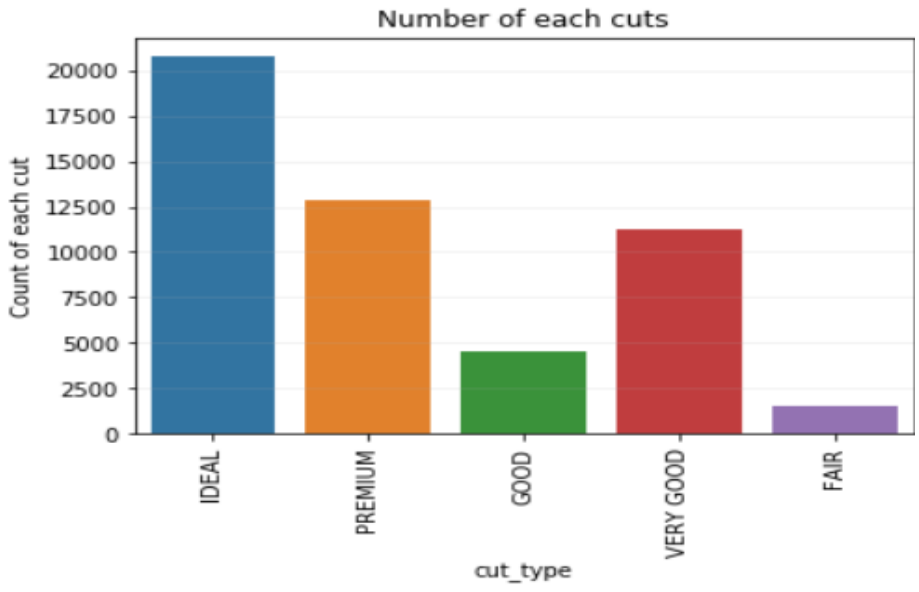
Feature	Mean	Median	Min	Max	Standard Deviation
carat	0.736	0.70	0.20	1.79	0.3766
average us salary	39506.756	39528.00	30000	48999	5485.5500

number of diamonds mined (millions)	2.901.	2.91	0.60	5.20	1.326515
depth	61.883	61.80	60.00	63.90	0.820
table	57.348	57.00	54.00	62.00	1.88
price	1874.58	2401.00	326.00	4000	882.27
x	5.72	5.70	3.73	8.99	1.101
y	5.49	5.71	4.00	7.00	0.802
z	3.49	3.53	2.06	5.00	0.629

Visualization graphs

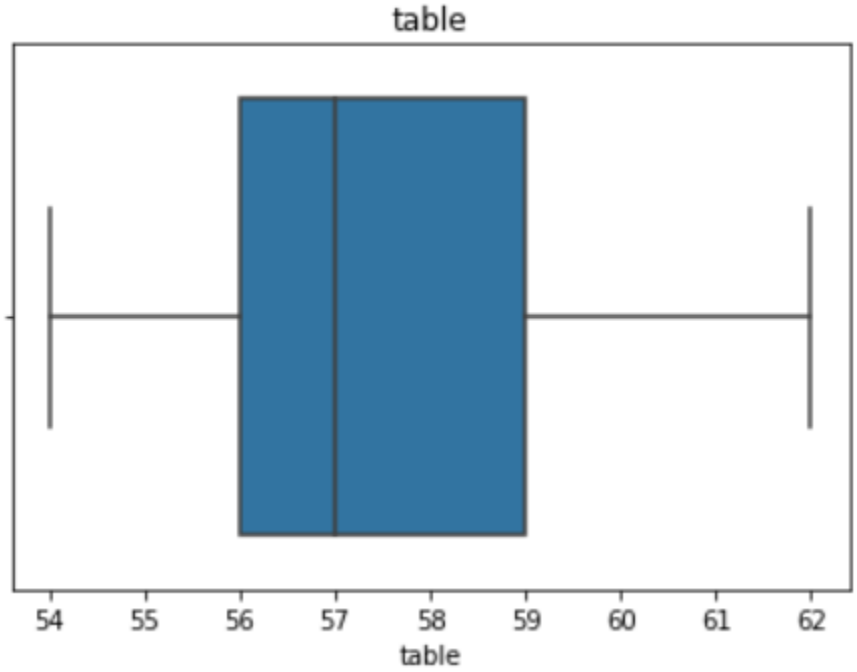
Plot 1:

We imported a seaborn library to plot visualization graphs. The plot cut vs count represents the count of each cut type of the diamond.



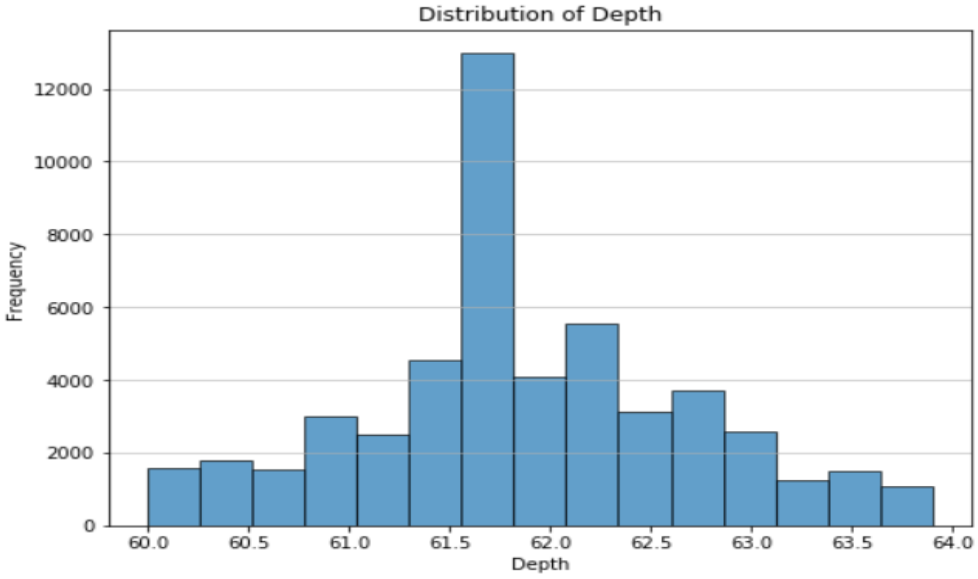
The graph describes an ideal cut that is far greater than other types of cuts, On other hand premium cut is the second highest cut which is equal to 12500. The count of cut named Very good is approximately half of the ideal cut which is slightly less than 12500. As observed from the graph the fair cut is the least one.

Plot 2:



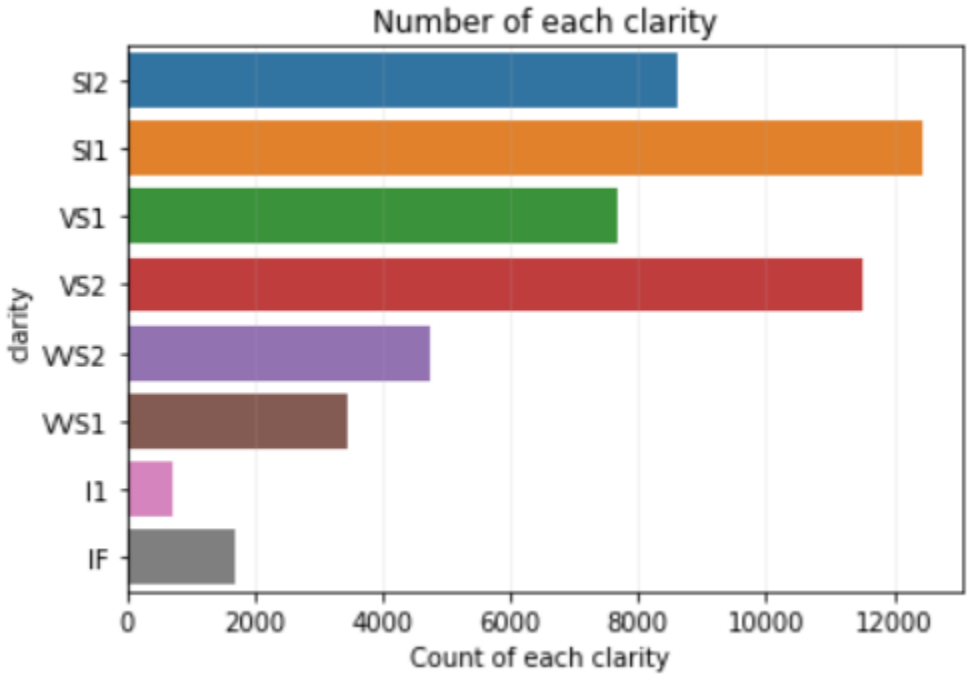
The above graph represents the box plot of the table. As we can see there are no outliers. And the median value is settled at 57 and all the values lies between 56 -59.

Plot 3



The graph depicts histplot of the depth column. The x axis is the values ranging from 60.0 to 64.0. The Y-axis represents the count of the values. Histograms are divided into bins, each bin represents the frequency of the data points. This Hist Plot is unimodal which consists of 1 peak.

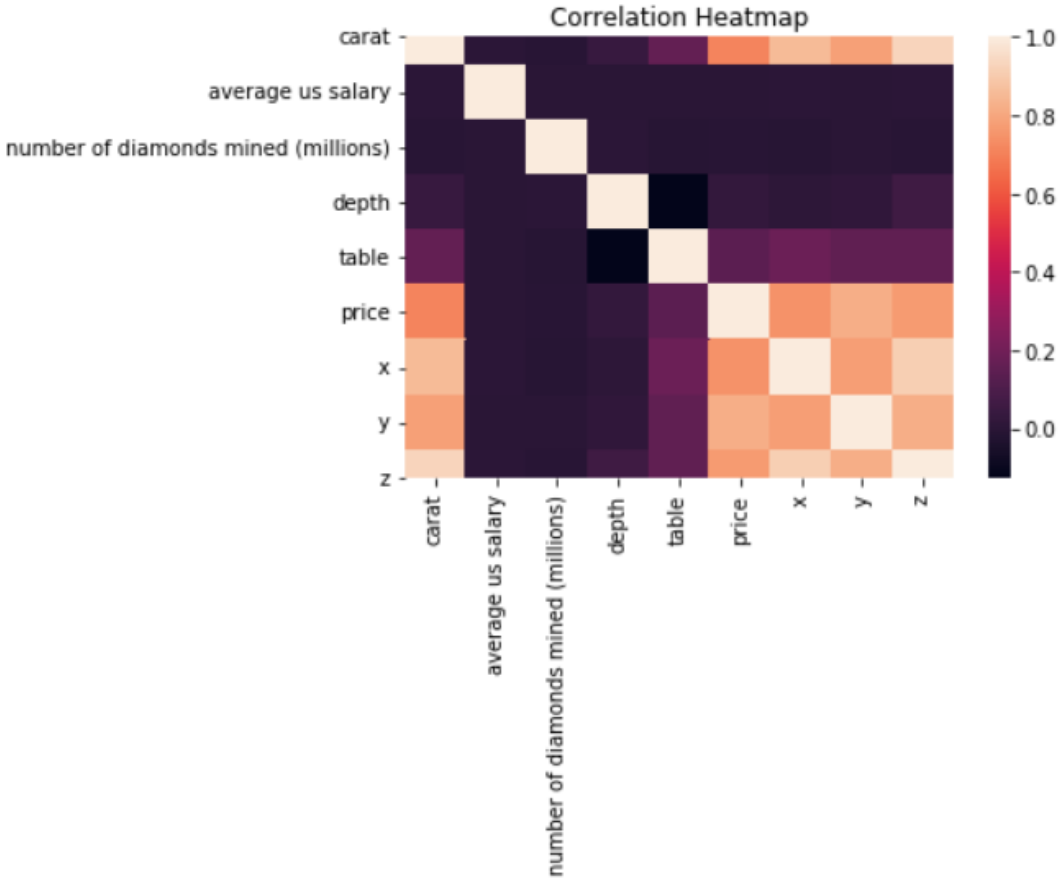
Plot 4



The above graph describes the count of each clarity. The clarity SI 1 is the highest clarity whose count is above 12000. When compared with all other clarities I1 clarity is the least one. Clarity VS2 is the second highest one which equals to slightly greater than 11000. VS1, VVS1, IF are record least count.

Plot 5:

Heatmap is used to know the dependency of one column with other columns.



From the above correlation heat map we can observe there is strong positive correlation between carat and other columns. And we see that there is moderate positive correlation between carat and price columns. Also there is no correlation between some of the columns for example between number of diamonds mined and y data points.

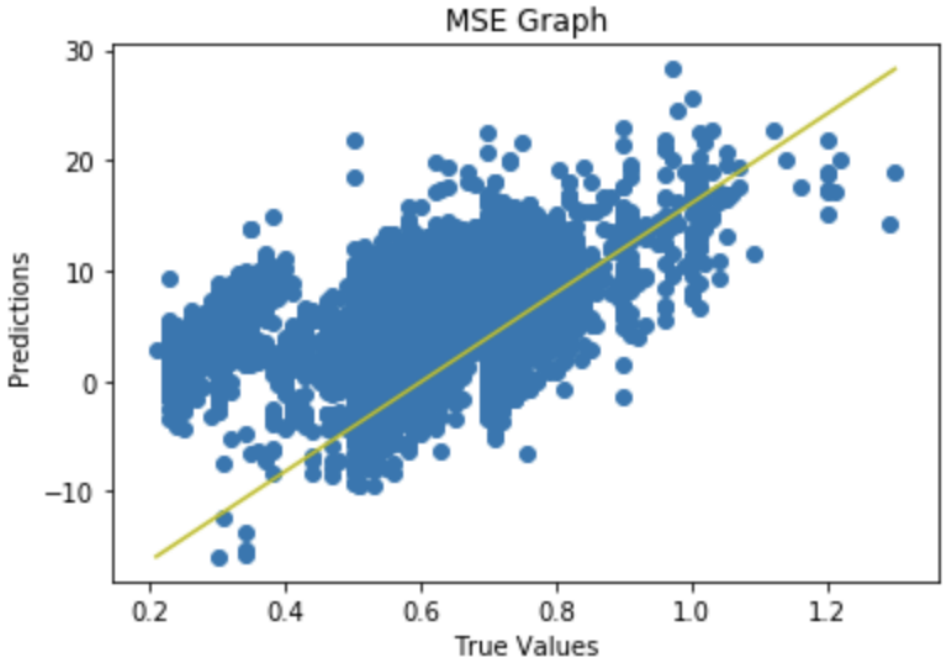
Linear Regression Model:

We have chosen Carat as our target variable. Then we split the X and Y dataset into train and test in 80:20 ratio. Then we used the formulas provided for linear regression for implementation.

The MSE value for the model is **35.29.**

Predicted vs Actual Data Graph:

```
Mean Squared Error: 35.29690095026595
```



OLS Estimation:

Benefits:

1. OLS is a widely used method because of its simplicity, efficiency, and flexibility.
2. OLS also works for non-linear models
3. It involves minimizing the sum of squared differences between the actual and predicted values, which is easy to understand.
4. It is efficient and inexpensive with respect to computation.

Drawbacks:

1. Sensitive to outliers, which can change estimations and decrease accuracy.
2. In case of on multivariate dataset that contains a single independent variables set and multiple dependent variables sets it performs poorly
3. when some points in the training data have excessively large or small values it performs poorly.

Linear regression model

Benefits:

1. It is easy to understand the relationship between variables.
2. It is computationally efficient when dealing with large datasets.
3. It provides best results when relations between dependent and independent variables are linear.
4. It is easy to interpret the output coefficients.

Drawbacks:

1.When there is high correlation between independent variables, linear regression becomes unstable. Which is the case with this data.
2. Outliers have a huge effect on regression.
3. Not suitable for modeling non-linear and complex relationships.

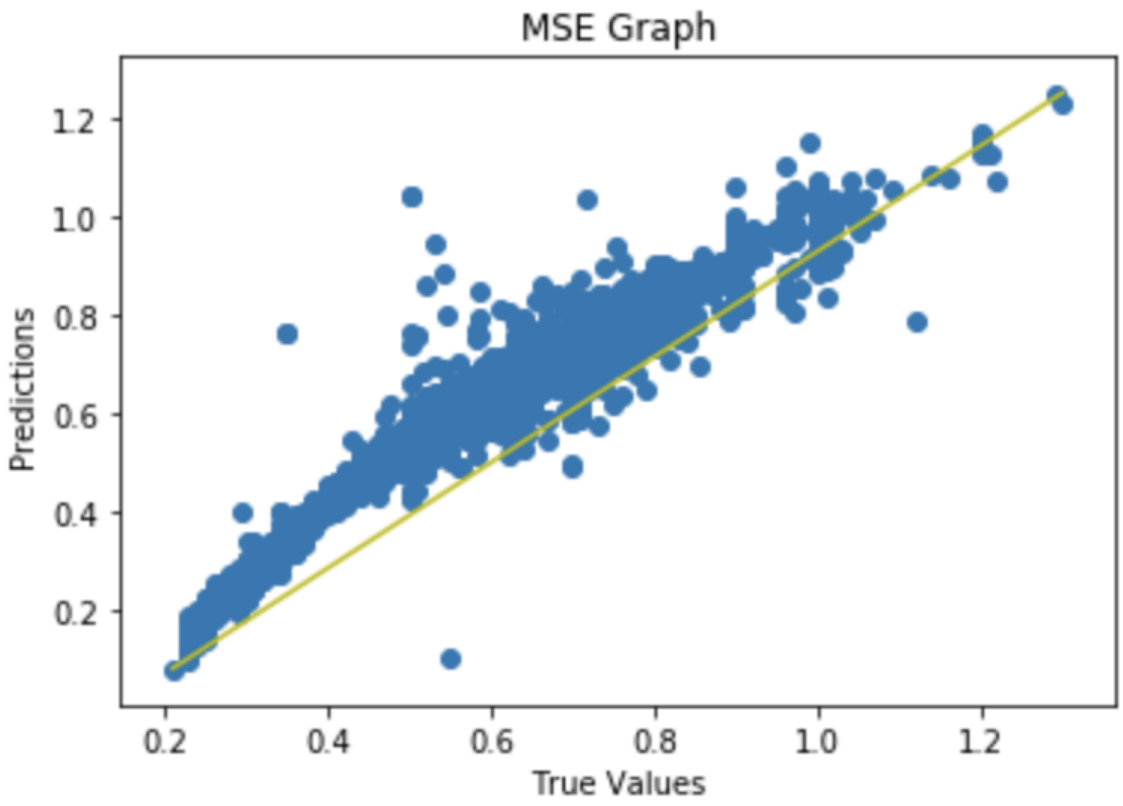
Part 4 - Linear Regression Implementation

We have chosen Carat as our target variable. Then we split the X and Y dataset into train and test in 80:20 ratio. Then we used the formulas provided for linear regression for implementation.

The MSE value for the model is **0.0034.**

Predicted vs Actual Data Graph:

```
Mean Squared Error: 0.003482340524075 01
```



The main motivation of using L2 regularization is to reduce model over fitting. It is used to have smaller coefficients.

Linear regression:	Ridge regression
It is used to find the coefficients that minimize the sum of squared differences between the true values and predicted values. It does not include regularization terms.	Ridge regression includes L2 regularization
Linear regression has small bias and high variance that makes to get overfitted	It increases the bias and decreases variance which gives good predictions
The coefficients in linear regression are relatively large	Since it uses L2 regularization coefficients gets smaller.

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

Benefits:
1. Ridge regression is used to prevent overfitting of the model.
2. It can handle even independent variables that are highly correlated to each other.
3. It increases bias and decreases variance which gives more accurate predicted values.
4, It is insensitive to outliers.

Drawbacks:
1. It doesn't perform variable selection.
2. It is used for linear relationships, which means there should be a linear relationship between dependent and independent variables.
3. It requires a lambda for controlling the strength of regularization.

References:

1. #https://seaborn.pydata.org/
2. #https://numpy.org/doc/stable/reference/arrays.html
3. #https://piazza.com/class_profile/get_resource/llr3w56262t6qi/lm7bqxowa773h7
4. #https://piazza.com/class_profile/get_resource/llr3w56262t6qi/lmgx1ewoxbi1ai

Contribution

1.Adarsh Reddy Bandaru Assignment 1 - 50

2.Dharma Acha Assignment 1 - 50