JEWELRY MANUFACTURER PRICE PREDICTION

Table of Contents

1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning

1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

1.4 Inference: Basis on these predictions, what are the business insights and recommendations. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present

Problem 2: Holiday Package Analysis
2.1. Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, write an inference on it. Perform Univariate and Bivariate Analysis. Do exploratory data analysis.

2.2. Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis).

2.3. Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

2.4. Inference: Basis on these predictions, what are the insights and recommendations. Please explain and summarize the various steps performed in this project. There should be proper business interpretation and actionable insights present.

Jewelry Manufacturer Profitability Analysis

Summary:

You are hired by a company Gem Stones co ltd, which is a cubic zirconia manufacturer. You are provided with the dataset containing the prices and other attributes of almost 27,000 cubic zirconia (which is an inexpensive diamond alternative with many of the same qualities as a diamond). The company is earning different profits on different prize slots. You have to help the company in predicting the price for the stone on the bases of the details given in the dataset so it can distinguish between higher profitable stones and lower profitable stones so as to have better profit share. Also, provide them with the best 5 attributes that are most important.

Data Description:

- Carat -- Carat weight of the cubic zirconia.
- Cut -- Describe the cut quality of the cubic zirconia. Quality is increasing order Fair, Good, Very Good, Premium, Ideal.
- Color -- Colour of the cubic zirconia.With D being the worst and J the best.
- Clarity -- Clarity refers to the absence of the Inclusions and Blemishes. (In order from Worst to Best in terms of avg price) IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1
- Depth -- The Height of cubic zirconia, measured from the Culet to the table, divided by its average Girdle Diameter.
- Table -- The Width of the cubic zirconia's Table expressed as a Percentage of its Average Diameter.
- Price -- the Price of the cubic zirconia.
- X -- Length of the cubic zirconia in mm.
- Y -- Width of the cubic zirconia in mm.
- Z -- Height of the cubic zirconia in mm.

Sample of the Dataset:

	Unnamed: 0	carat	cut	color	clarity	depth	table	x	y	z	price
0	1	0.30	Ideal	E	SI1	62.1	58.0	4.27	4.29	2.66	499
1	2	0.33	Premium	G	IF	60.8	58.0	4.42	4.46	2.70	984
2	3	0.90	Very Good	E	VVS2	62.2	60.0	6.04	6.12	3.78	6289
3	4	0.42	Ideal	F	VS1	61.6	56.0	4.82	4.80	2.96	1082
4	5	0.31	Ideal	F	VVS1	60.4	59.0	4.35	4.43	2.65	779

Figure 1. Cubic_Zirconia_Sample

- Here we drop the ‘Unnamed’ column as it’s not needed for our analysis

Questions:

1.1. Read the data and do exploratory data analysis. Describe the data briefly. (Check the null values, Data types, shape, EDA, duplicate values). Perform Univariate and Bivariate Analysis.

Exploratory Data Analysis:

Let us check the basic info of the data frame.

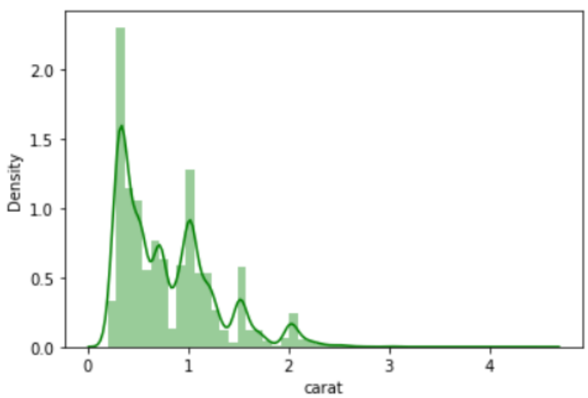
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 26967 entries, 0 to 26966
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype  
---  ------   --------------  -----  
 0   carat    26967 non-null  float64
 1   cut      26967 non-null  object 
 2   color    26967 non-null  object 
 3   clarity  26967 non-null  object 
 4   depth    26270 non-null  float64
 5   table    26967 non-null  float64
 6   x        26967 non-null  float64
 7   y        26967 non-null  float64
 8   z        26967 non-null  float64
 9   price    26967 non-null  int64  
dtypes: float64(6), int64(1), object(3)
memory usage: 2.1+ MB
```

Figure 2. Cubic_Zirconia_Info

Univariate Analysis

This analysis will display the statistical description of the numeric variable to view 5-point summary, histogram or distplot to view the distribution and the box plot to view outliers if any

```
Description of carat
------------------------------------------------
count    26933.00
mean         0.80
std          0.48
min          0.20
25%          0.40
50%          0.70
75%          1.05
max          4.50
Name: carat, dtype: float64 Distribution of carat
------------------------------------------------
```




```
BoxPlot of carat
---------------------------------------------
```

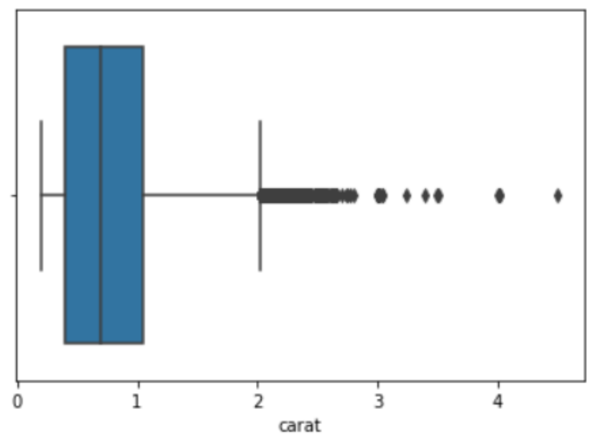


Figure 3. carat_summary

```
Description of depth
------------------------------------------------
count    26933.00
mean        61.75
std          1.39
min         50.80
25%         61.10
50%         61.80
75%         62.50
max         73.60
Name: depth, dtype: float64 Distribution of depth
------------------------------------------------
```

BoxPlot of depth

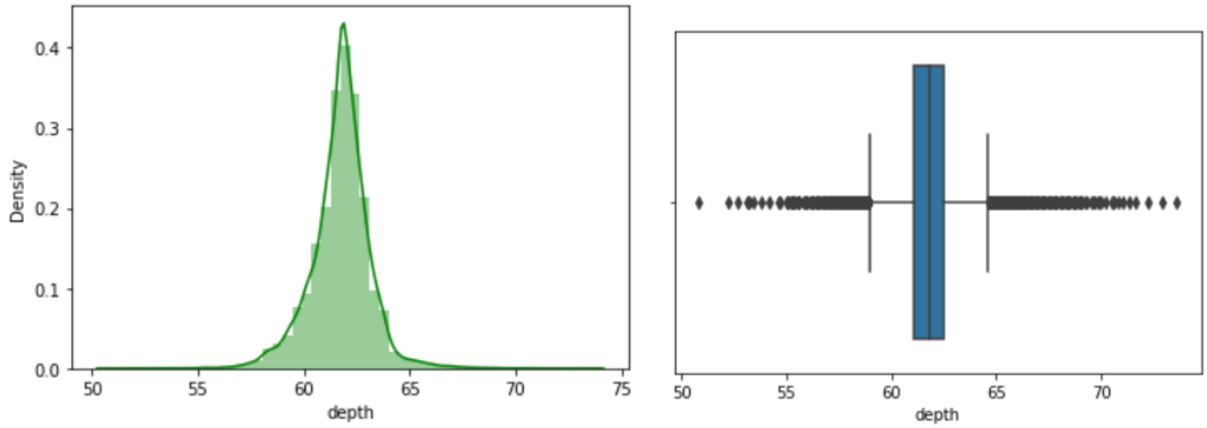


Figure 4. depth_summary

```
Description of table
--------------------------------------------------
count     26933.00
mean         57.46
std           2.23
min          49.00
25%          56.00
50%          57.00
75%          59.00
max          79.00
Name: table, dtype: float64 Distribution of table
--------------------------------------------------
```

BoxPlot of table

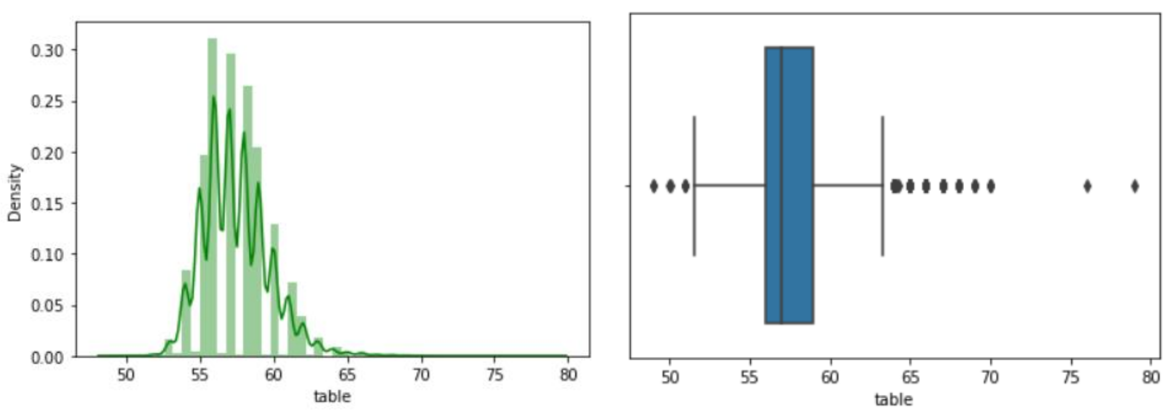


Figure 5. table_summary

```
Description of x
----------------------------------------
count    26933.00
mean         5.73
std          1.13
min          0.00
25%          4.71
50%          5.69
75%          6.55
max         10.23
Name: x, dtype: float64 Distribution of x
----------------------------------------
```

```
BoxPlot of x
----------------------------------------
```

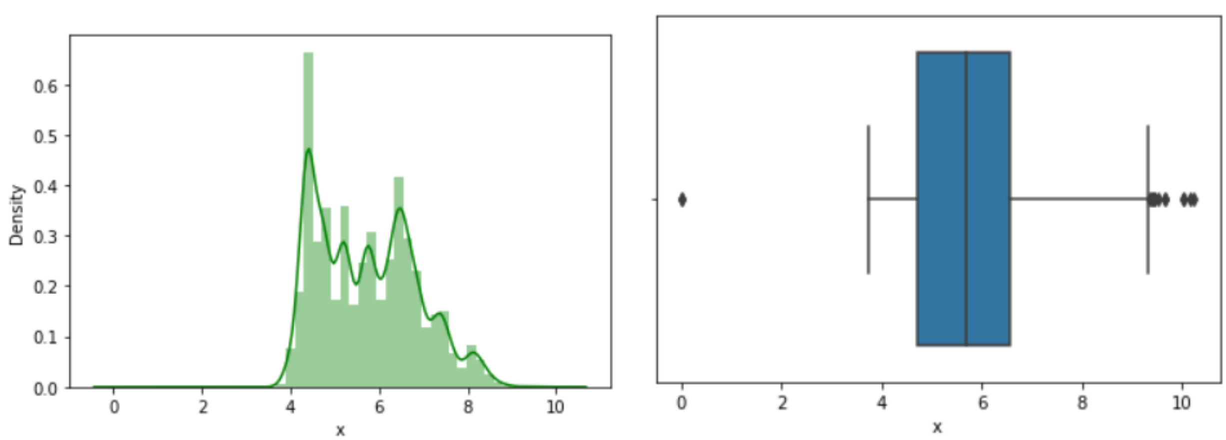


Figure 6. x_summary

```
Description of y
----------------------------------------
count    26933.00
mean         5.73
std          1.17
min          0.00
25%          4.71
50%          5.70
75%          6.54
max         58.90
Name: y, dtype: float64 Distribution of y
----------------------------------------
```

```
BoxPlot of y
--------------------------------------------
```

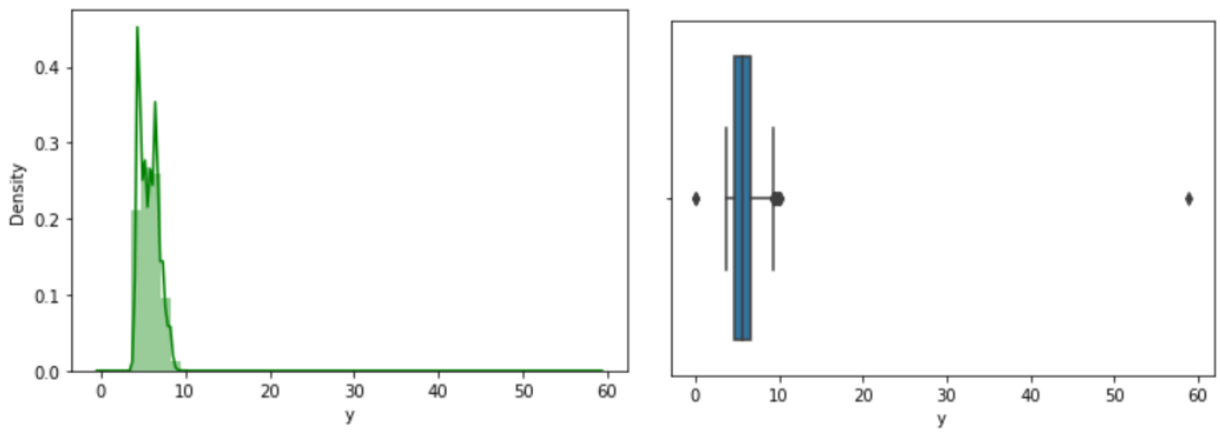


Figure 7. y_summary

```
Description of z
------------------------------------------
count    26933.00
mean         3.54
std          0.72
min          0.00
25%          2.90
50%          3.52
75%          4.04
max         31.80
Name: z, dtype: float64 Distribution of z
------------------------------------------
```

```
BoxPlot of z
--------------------------------------------
```

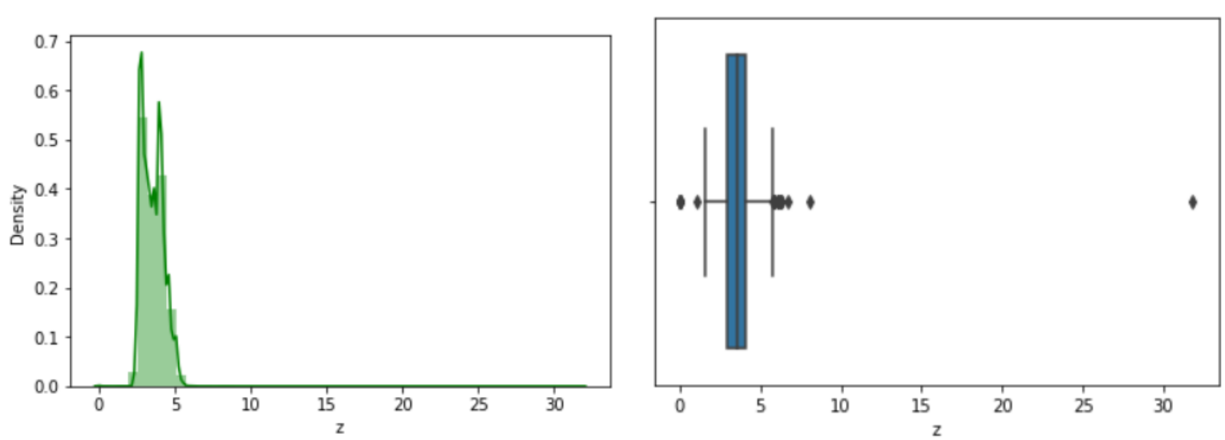


Figure 8. z_summary

```
Description of price
--------------------------------------------------
count    26933.00
mean      3937.53
std       4022.55
min        326.00
25%        945.00
50%       2375.00
75%       5356.00
max      18818.00
Name: price, dtype: float64 Distribution of price
--------------------------------------------------
                                          BoxPlot of price
                                          ----------------------------------------------
```

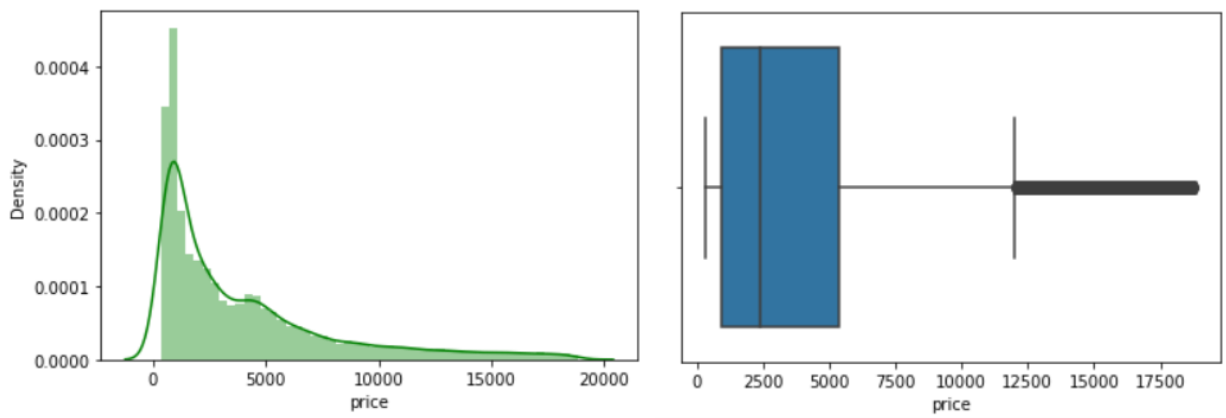


Figure 9. price_summary

Multivariate Analysis

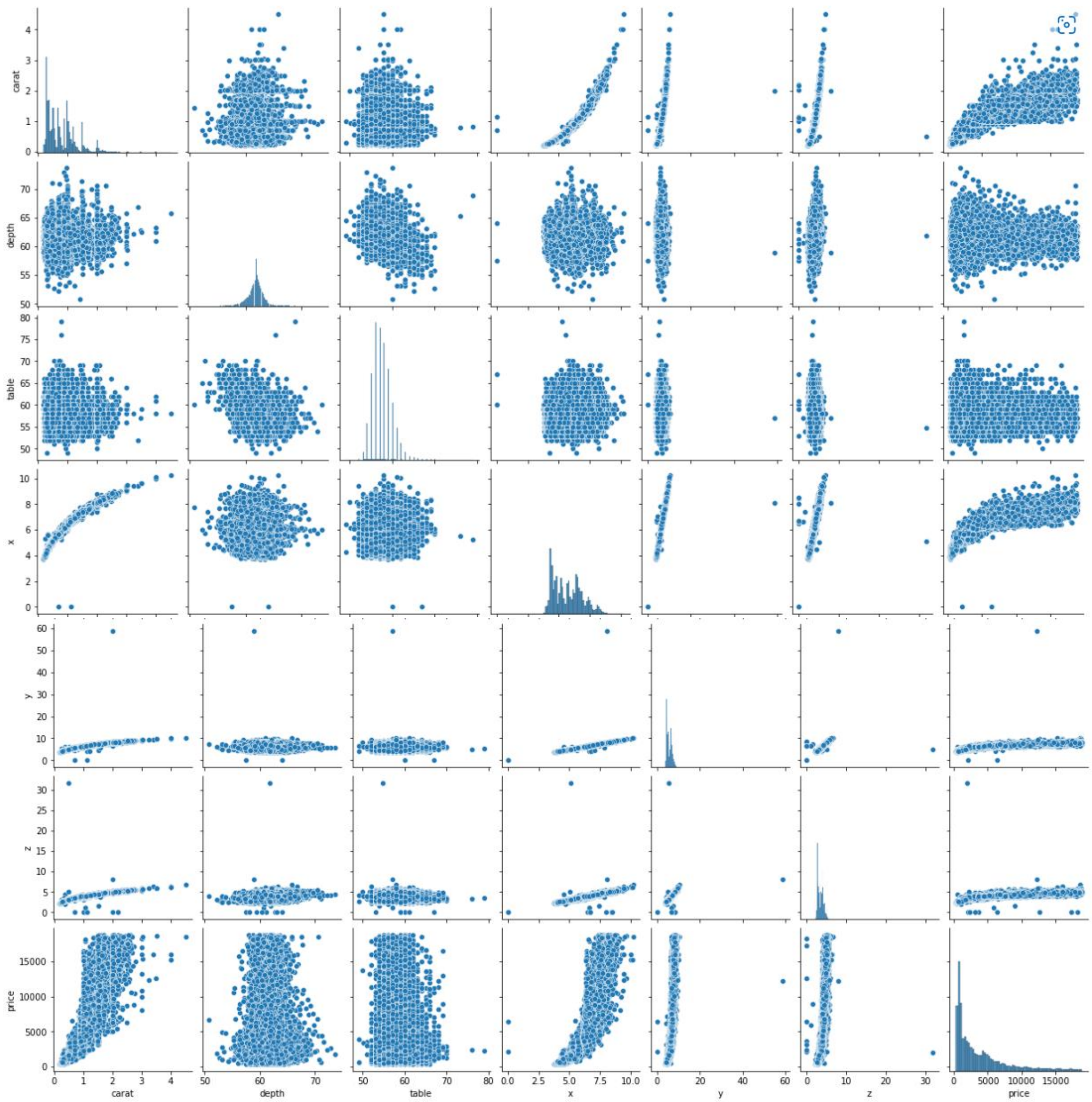


Figure 10. Cubic_Zirconia_Pairplot

Correlation Heatmap

Figure 11. Cubic_Zirconia_Correlation

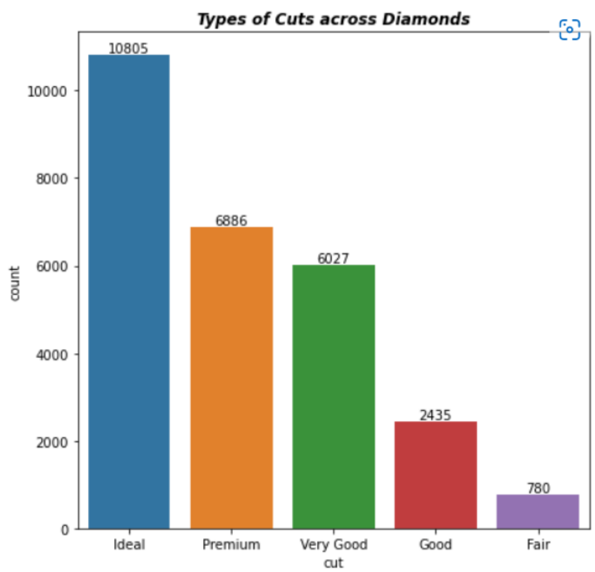


Figure 12. Cuts vs Diamond price

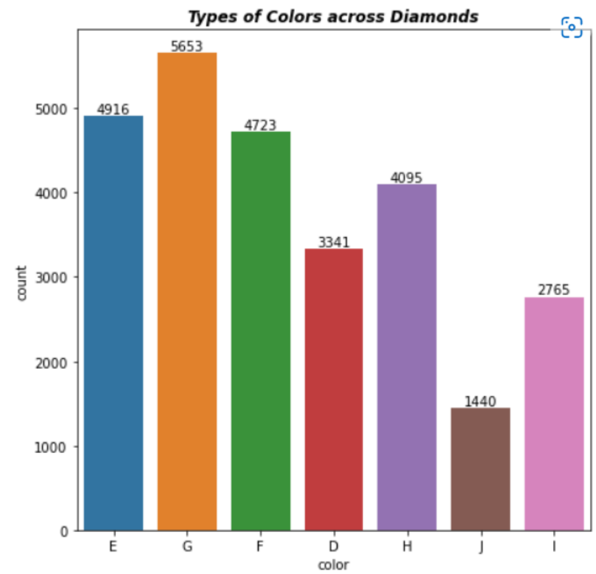


Figure 13. Color vs Diamond price

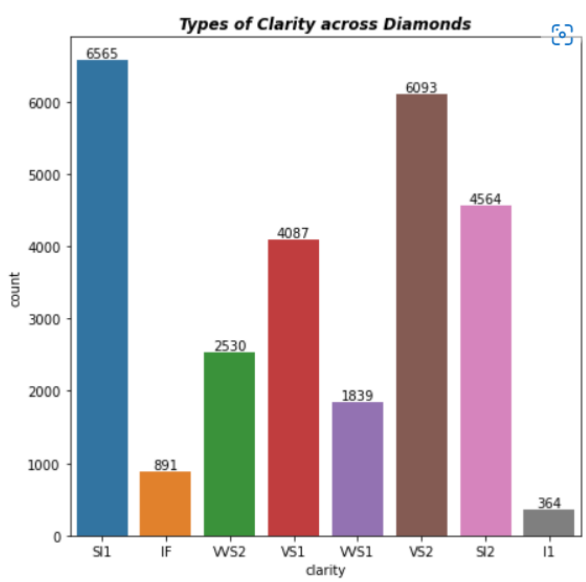


Figure 14. Clarity vs Diamond price

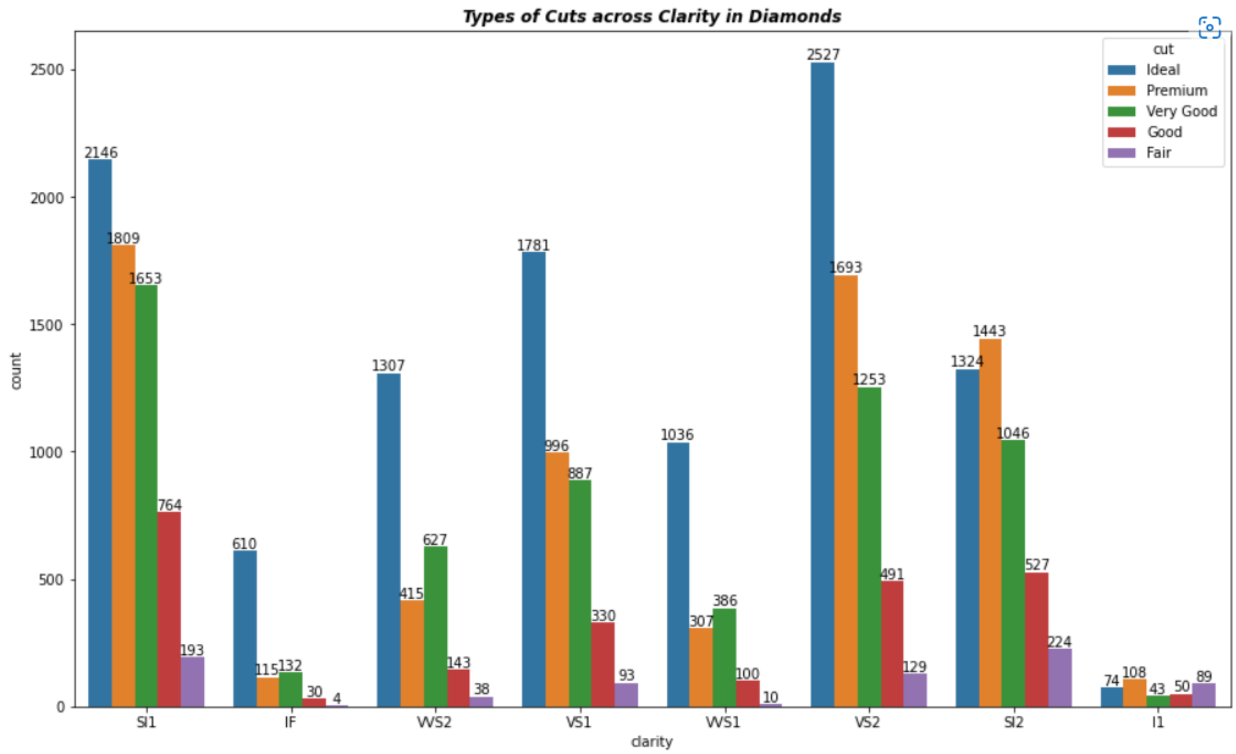


Figure 15. Cuts across Clarity in Diamonds

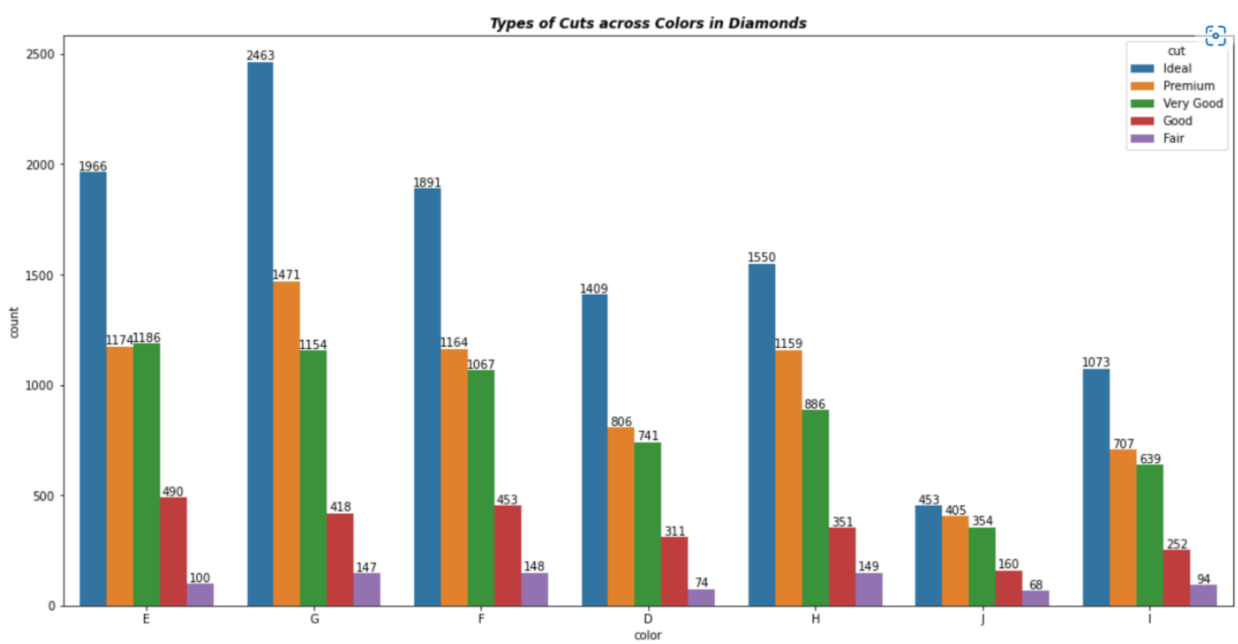


Figure 16. Cuts across Color in Diamonds

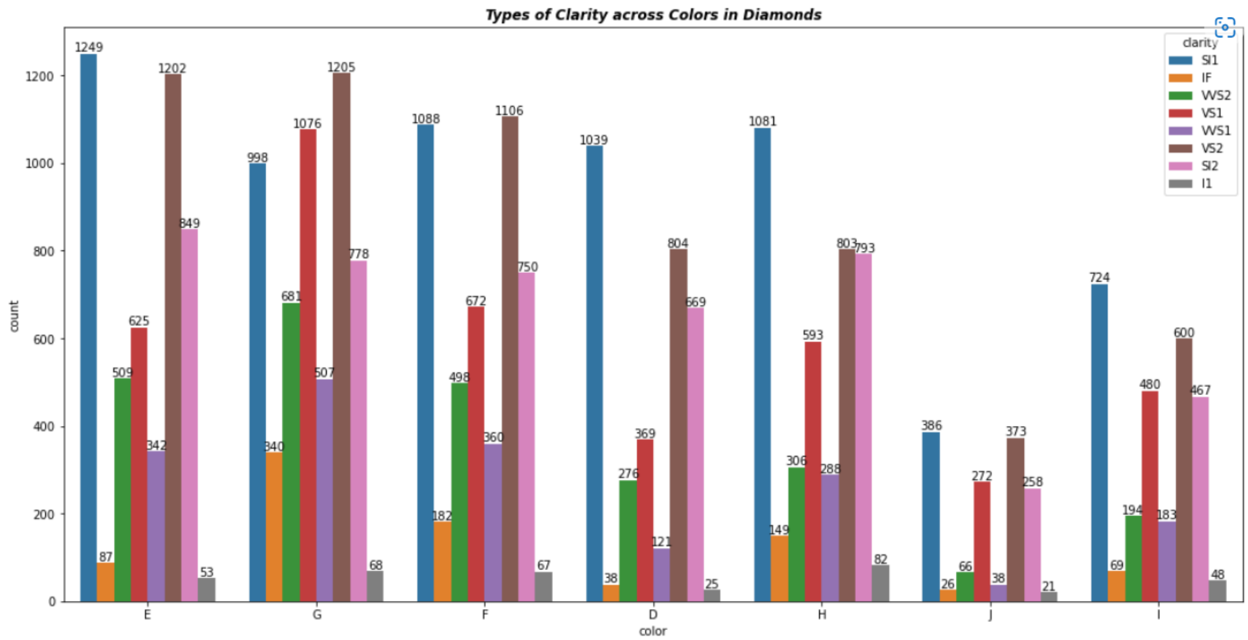


Figure 17. Clarity across Color in Diamonds

Bi-Variate Analysis with Target variable

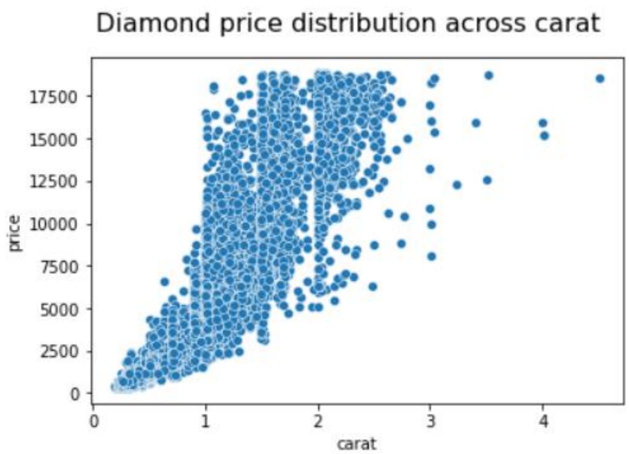


Figure 18. Diamond price across Carat

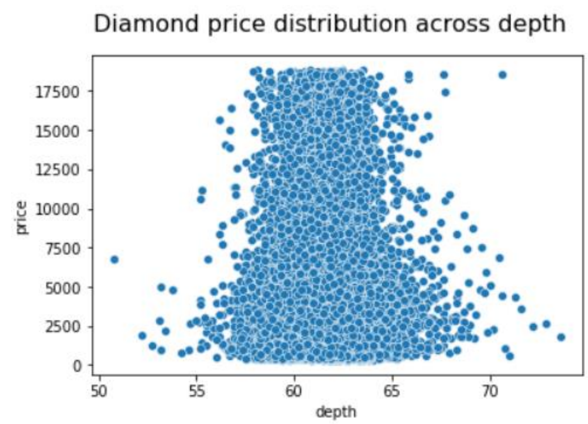


Figure 19. Diamond price across Depth

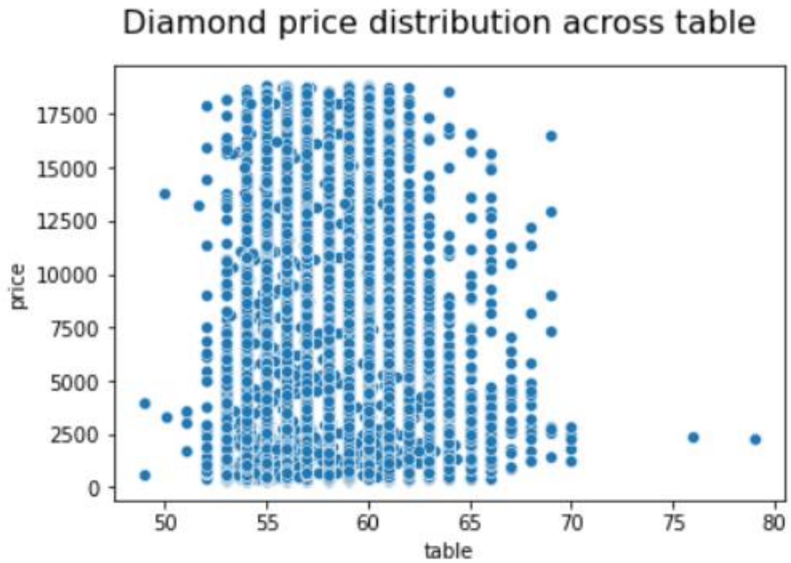


Figure 20. Diamond price across Table

Insights:--

- The dataset has 26,967 rows and 11 columns
- We've dropped the 'Unnamed' column as it's not useful for our analysis
- 6 columns are of the 'float', 1 of 'integer' and 3 of 'object' data type
- There are 34 duplicate rows which were dropped as they are only 0.12% of the dataset (34/26967)
- There are 697 null values in the 'depth' column, which we have imputed with the 'median' as it's a continous variable
- There are outliers in the below 2 features as indicated by the box-plots
    - Carat – above 2
    - Table – above 65
    - However we'll not be treating the datset for these outliers as these are legit data points carrying valuable information
    - This is represented in the 2 scatter plots in Bi-Variate Analysis with Target variable, higher the weight (carat) and width (table), higher the diamond price
- We can see from the pairplot and the heatmap, the variables 'x', 'y' and 'z' are highly correlated with each other and with the target variable – 'price'
    - The lowest value is 0.85, which indicates a high degree of correlation
    - Thereby, we'll be removing them from the dataset
- We've plotted 3 graphs depicting the various types of 'Cut', 'Color' and 'Clarity' across the diamonds and plotted 3 more, which are a combination of the above
    - Cuts across Clarity
    - Cuts across Colors
    - Clarity across Colors

- We’ve also plotted the target variable, ‘price’ with the other numeric data types – carat (weight), depth (height) and table (width)
    - As we can see, higher the value of the features, higher the price of the diamonds

1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of combining the sub levels of a ordinal variables and take actions accordingly. Explain why you are combining these sub levels with appropriate reasoning.

```
carat         0
cut           0
color         0
clarity       0
depth       697
table         0
x             0
y             0
z             0
price         0
dtype: int64
```

Figure 21. Cubic_Zirconia_Null

- There are 697 null values in depth and none in the other variables
- Central tendency measures such as mean, median, or mode is considered for imputation
- Mean is the average of all values in a set, Median is the middle number in a set of numbers sorted by size, and Mode is the most common numerical value
- As we’ve seen previously in the box-plot, depth seems normally distributed
- For a symmetric and numeric data distribution, one can use the Mean or Median value for imputing missing values
- However, Mean imputation does not preserve the relationships among variables
- Also, as it’s a continuous variable, we’ll be imputing it with the ‘Median’
- Dropping the null values isn’t an option in this scenario due to the above reasons and the statistical analysis remains unbiased
- Combining the sub-levels of ordinal variables (cut, color, clarity) is not possible in this scenario
    - Cut has 5 sub-levels (Fair, Good, Very Good, Premium, Ideal) in increasing order
    - Color has 7 sub-levels (D, E, F, G, H, I, J) with D being the worst and J the best
    - Clarity has 8 sub-levels (IF, VVS1, VVS2, VS1, VS2, Sl1, Sl2, l1) in order from worst to best
    - All we know from the sub-levels is they are ranked from worst to best; we do not know what is the differentiating criteria to combine / split them
    - For e.g., we know Sl2 is better than Sl1 in terms of clarity; however, we do not by how much
    - This makes it very confusing and difficult to combine the sub-levels and hence not considered

1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

- Most machine learning models only accept numerical variables, hence, preprocessing the categorical variables becomes a necessary step
- We need to convert these categorical variables to numbers such that the model is able to understand and extract valuable information
- There are many encoding techniques; however, we’ll be using the One Hot Encoding method for our dataset as the features (cut, color, clarity) are nominal (do not have any order)
    - In one hot encoding, for each level of a categorical feature, we create a new variable
    - Each category is mapped with a binary variable containing either 0 or 1. Here, 0 represents the absence, and 1 represents the presence of that category
    - These newly created binary features are known as Dummy variables
    - The number of dummy variables depends on the levels present in the categorical variable

	carat	depth	table	price	cut_Fair	cut_Good	cut_Ideal	cut_Premium	cut_Very Good	color_D	...	color_I	color_J	clarity_I1	clarity_IF	clarity_SI1	clarity_SI2	clarity_VS1	clarity_VS2	clarity_VVS1	clarity_VVS2
0	0.30	62.1	58.0	499	0	0	1	0	0	0	...	0	0	0	0	1	0	0	0	0	0
1	0.33	60.8	58.0	984	0	0	0	1	0	0	...	0	0	0	1	0	0	0	0	0	0
2	0.90	62.2	60.0	6289	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	1
3	0.42	61.6	56.0	1082	0	0	1	0	0	0	...	0	0	0	0	0	0	1	0	0	0
4	0.31	60.4	59.0	779	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	1	0

Figure 22. Cubic_Zirconia_One Hot Encoding

- Once the encoding and the split has been performed, we use 4 models -- ANN, Decision Tree, Random Forest, and Linear Regression to compare which one yields the best result

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	1159.773507	1151.312687	0.916329	0.919295
Decision Tree Regressor	34.179254	770.398133	0.999927	0.963864
Random Forest Regressor	221.820611	590.918411	0.996939	0.978740
ANN Regressor	570.188703	582.542869	0.979776	0.979338

Figure 22. Cubic_Zirconia_Model Comparison

- Root Mean Square Error (RMSE) is the standard deviation of the residuals (prediction errors)
- Residuals are a measure of how far from the regression line data points are
- RMSE is a measure of how spread out these residuals are. It tells us how concentrated the data is around the line of best fit
- Decision Tree, Random Forest and ANN have Overfitting issues as the Test RMSE is greater than the Training RMSE
- Overfitting is a modeling error in statistics that occurs when a function is too closely aligned to a limited set of data points
- As a result, the model is useful in reference only to its initial data set, and not to any other data sets
- In terms of Accuracy, all 4 models perform well on both the training and test dataset with the test score being close to the training score
- Only Linear Regression has a slightly higher Test score compared to the Training score
- Considering both metrics, I'd select the ANN model as it's Train and Test RMSE are quite close by and its test accuracy is the highest amongst all
- Now, let's use the Grid Search for hyperparameter tuning to find the optimum parameters to be used in our models

```
{'max_depth': 15, 'min_samples_leaf': 3, 'min_samples_split': 15}
```

Figure 23. Decision Tree

```
{'max_depth': 10, 'max_features': 6, 'min_samples_leaf': 3, 'min_samples_split': 30, 'n_estimators': 300}
```

Figure 24. Random Forest

```
{'activation': 'relu', 'hidden_layer_sizes': (100, 100), 'solver': 'adam'}
```

Figure 25. Artificial Neural Network

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	1159.773507	1151.312687	0.916329	0.919295
Decision Tree Regressor	473.264505	659.511967	0.986067	0.973517
Random Forest Regressor	994.244191	1060.524446	0.938509	0.931521
ANN Regressor	572.455629	577.646930	0.979615	0.979684

Figure 26. Cubic_Zirconia_Grid Search_Model Comparison

- We can see the results once we've applied the optimum parameters through the grid search process
- Decision Tree still has Overfitting issues as the Test RMSE is still higher than the Training RMSE
- Random Forest and ANN are similar, however within our range of +/- 10%
    - Among the two, ANN has the higher accuracy in both training and test datasets
- Linear Regression still has a higher test accuracy than training
- Overall, like before, ANN is the best model as it has
    - Lowest Test RMSE
    - Training and Test RMSE are very close and the difference is negligible
    - Highest accuracy in both training and test datasets

1.4 Inference: Basis on these predictions, what are the business insights and recommendations.

- We have our equation for the Linear regression

price = (-790.53) * Intercept + (8929.5) * carat + (-18.44) * depth + (-24.24) * table + (-742.35) * cut_Fair + (-161.59) * cut_Good + (102.29) * cut_Ideal + (6.18) * cut_Premium + (4.94) * cut_Very_Good + (728.6) * color_D + (536.25) * color_E + (411.15) * color_F + (207.12) * color_G + (-278.18) * color_H + (-767.81) * color_I + (-1627.67) * color_J + (-3774.61) * clarity_I1 + (1502.72) * clarity_IF + (-349.59) * clarity_SI1 + (-1324.26) * clarity_SI2 + (613.14) * clarity_VS1 + (296.37) * clarity_VS2 + (1172.26) * clarity_VVS1 + (1073.44) * clarity_VVS2

price = (-790.53) * Intercept + (8929.5) * carat + (-18.44) * depth + (-24.24) * table + (-742.35) * cut_Fair + (-161.59) * cut_Good + (102.29) * cut_Ideal + (6.18) * cut_Premium + (4.94) * cut_Very_Good + (728.6) * color_D + (536.25) * color_E + (411.15) * color_F + (207.12) * color_G + (-278.18) * color_H + (-767.81) * color_I + (-1627.67) * color_J + (-3774.61) * clarity_I1 + (1502.72) * clarity_IF + (-349.59) * clarity_SI1 + (-1324.26) * clarity_SI2 + (613.14) * clarity_VS1 + (296.37) * clarity_VS2 + (1172.26) * clarity_VVS1 + (1073.44) * clarity_VVS2

Figure 27. Cubic_Zirconia_Linear Regression Equation

- Our intercept term is -790.53, which means it is the mean diamond price when all the features are zero
- When cut_Ideal increases by 1 unit, price increases by 102.29 units, keeping all other predictors constant. etc.
- There are also some negative co-efficient values, for instance, cut_Fair has its corresponding co-efficient as -742.35. This implies, when the cut type is Fair, the price decreases by 742.35 units, keeping all other predictors constant. etc.

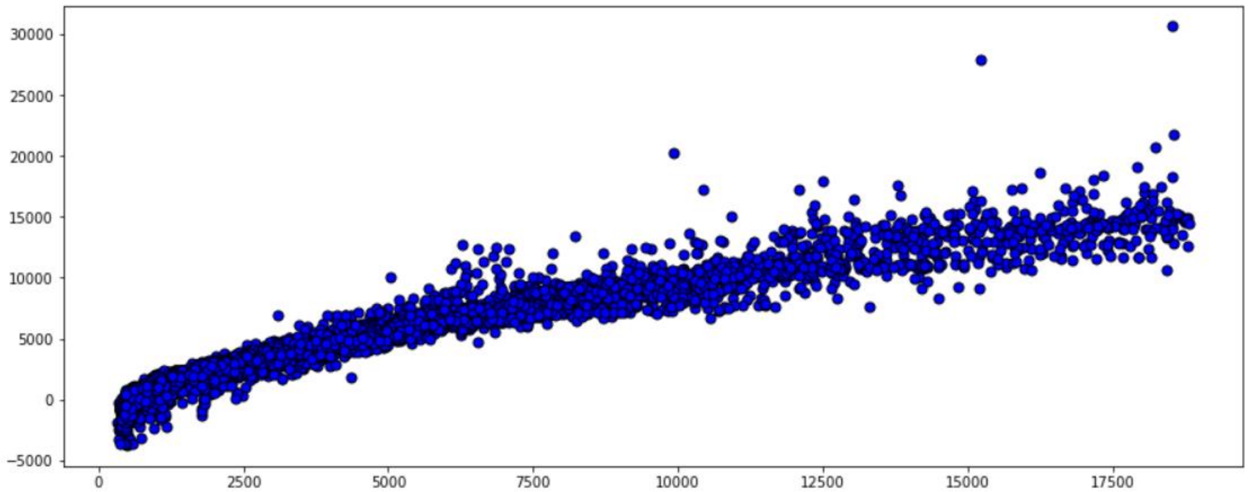


Figure 28. Cubic_Zirconia_Test vs Predicted

- We've plotted the predicted y value vs actual y values for the test data
- A good model's prediction will be close to actual leading to high R2 values (0.916 for our model)

Business Insights and Recommendations: --

- Carat is the most important contributor of the price as an increase of 1-unit results in an increase of 8,929.5 units in the price
- Clarity (5 positives out of 8 types), Cut (3 positives out of 5 types) and Color (4 positives out of 7 types) in that order are the next 3 determinants of price
- Depth and Table have a negative impact on the price
- Clarity types IF (1,503), VVS1 (1,172) and VVS2 (1,073) are the highest contributors to price
- Cut types Ideal (102), Premium (6) and Very Good (5) are the highest contributors to price
- Colors D (729), E (536), F (411) are the highest contributors to price
- The company should focus on the above combinations to
    - optimize its costs
    - forecast its ideal price and
    - maximize its profits
- Clarity types I1 (-3,775) and SI2 (-1,324) are the highest detractors of price
- Cut types Fair (-742) and Good (-162) are the highest detractors of price
- Colors J (-1,628) and I (-768) are the highest detractors of price
- For the above combinations, the company should
    - look into the cost of manufacturing and decide if they'd still like to continue offering these options to its customers
    - analyze their current inventory and see how best to clear the stocks quickly; perhaps some offers or discounts can help