

Gem Stones

Analysis

Name Bhushan Rai

PGP-DSBA Online January’ 21

Date: 01/08/2021

Table of Contents

Contents

[Executive Summary 3](#_bookmark0)

[Introduction 3](#_bookmark1)

[Data Description 3](#_bookmark2)

[Sample of the dataset](#_bookmark3) 4

[Exploratory Data Analysis 4](#_bookmark4)

[Missing](#_bookmark5) Values.......................................................................................................................................5

[Descriptive Statistics](#_bookmark7) 6

Find and remove [duplicate Values](#_bookmark8) 6

Univariate Analysis................................................................................................................................7

Check Outliers......................................................................................................................................8

Skewness..............................................................................................................................................8

Treated Outliers...................................................................................................................................8

[Count plot](#_bookmark10) 9

Bivariate Analysis................................................................................................................................10

Bivariate Analysis................................................................................................................................11

Corrrelation.........................................................................................................................................12

Convert categorical..............................................................................................................................13

2.2 Data Split: Split the data into test and train(1 pts), build classification model CART (1.5 pts),

Random Forest (1.5 pts), Artificial Neural Network(1.5 pts). Object data should be

converted into categorical/numerical data to fit in the models.

(pd.categorical().codes(), pd.get\_dummies(drop\_first=True)) Data split, ratio defined for the split,

train-test split should be discussed. Any reasonable split is acceptable.

Use of random state is mandatory. Successful implementation of each model.

Logical reason behind the selection of different values for the parameters involved in

each model. Apply grid search for each model and make models on best\_params.

Feature importance...........................................................................................................................14

2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using

Accuracy (1 pts), Confusion Matrix (2 pts), Plot ROC curve and get ROC\_AUC score for

each model (2 pts), Make classification reports for each model. Write inferences on

each model (2 pts). Calculate Train and Test Accuracies for each model.

Comment on the validness of models (overfitting or underfitting)

Build confusion matrix for each model. Comment on the positive class in hand.

Must clearly show obs/pred in row/col Plot roc\_curve for each model.

Calculate roc\_auc\_score for each model.

Comment on the above calculated scores and plots. Build classification reports for each model.

Comment on f1 score, precision and recall, which one is important here. ...................................14

Cart Model...........................................................................................................................................14

AUC for training and test data.............................................................................................................15

Cart model score .................................................................................................................................16

Conclusion of cart model.....................................................................................................................17

Random Forest model..........................................................................................................................17

RF Conclusion.......................................................................................................................................18

ANN model.............................................................................................................................................19

Comparison of performance metrics.....................................................................................................20

Reccomendations...................................................................................................................................21

Reccomendations....................................................................................................................................22

List of Figures

|  |
| --- |
| Fig.1 – Distplot 8 |
| Fig.2 – Boxplot 9 |
| Fig.3 – Count plot...........................................................................................................................................10 |
| Fig.4 – Countplot............................................................................................................................................11 |
| Fig 5- Pairplot...............................................................................................................................................12 |
| Fig 6- Heatmap.............................................................................................................................................13 |

List of Tables

|  |
| --- |
| Table 1. Sample of data and types of data......................................................................................................5 |
| Table 2. missing values....................................................................................................................................6 |
| Table 3. descriptive statistics and duplicate values.........................................................................................7 |
| Table 4. skewness ...9 |
| Table 5. correlation........................................................................................................................................12 |
| Table 6. proportions of normalize data.........................................................................................................13 |
| Table 7. Cart model variable importance.......................................................................................................14 |
| Table 8. predicted class..................................................................................................................................15 |
| Table 9. Cart train and test.............................................................................................................................16 |
| Table 10. Cart precision and f1 score..............................................................................................................17 |
| Table 11. RF train and test...............................................................................................................................17 |
| Table 12. RF test data precision score and F1 score........................................................................................18 |
| Table 13: RF variable importance.....................................................................................................................18 |
| Table 14. ANN train data score.........................................................................................................................19 |
| Table 15.ANN train precision and F1 score.......................................................................................................20 |
| Table 16. Comparison of 3 models....................................................................................................................20 |

# Executive 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.

# 

# Introduction

# The purpose of the project is to predict the price of the stone on the basis of the details given in dataset using Linear Regression

# Data Description

# 

## 

## Sample of the dataset:

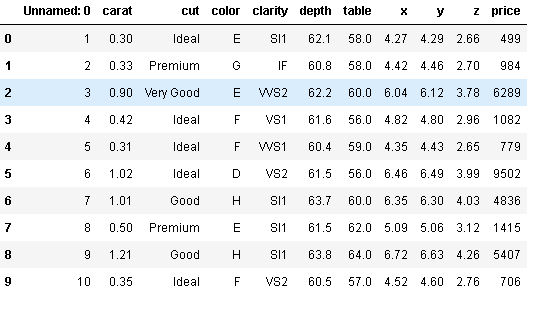


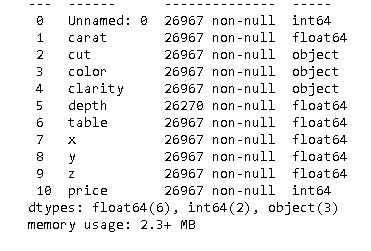
Table 1. Dataset Sample

# Exploratory Data Analysis

# Let us check the shape of the dataset

# 

# Let us check the info of the dataset



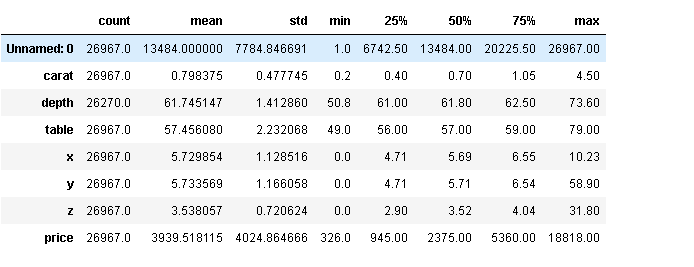
## Check for data types

## 

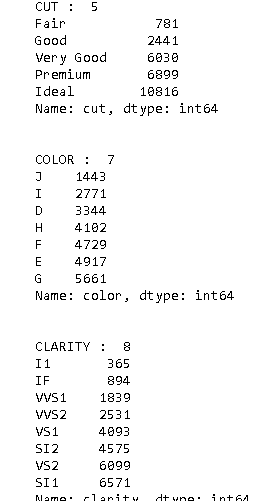
## Check for missing values in the dataset:

## 

# Descriptive Statistics:

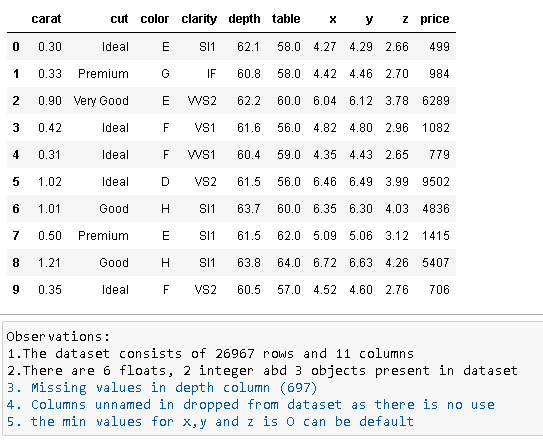


Unique Categorical values



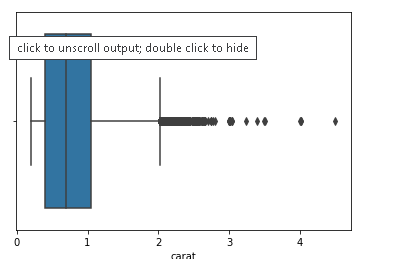
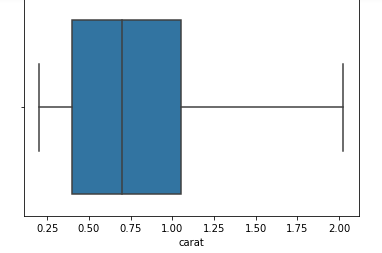
Check Duplicate values:

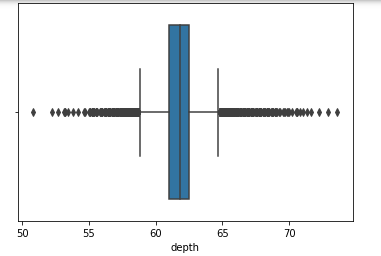
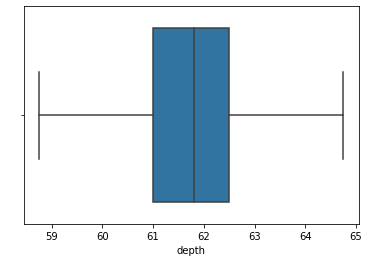


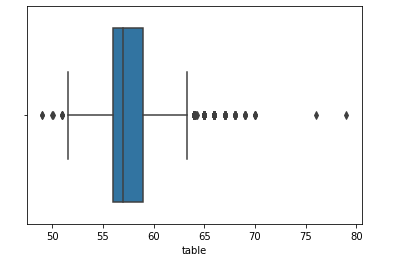
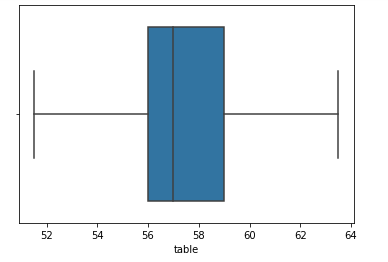


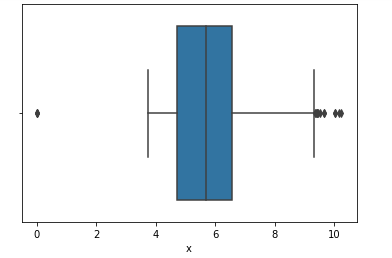
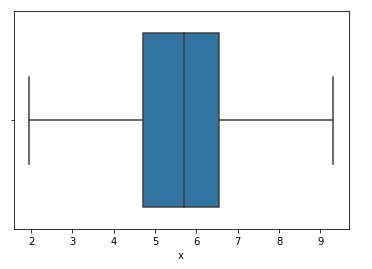
EDA:

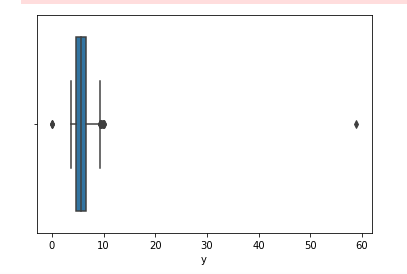
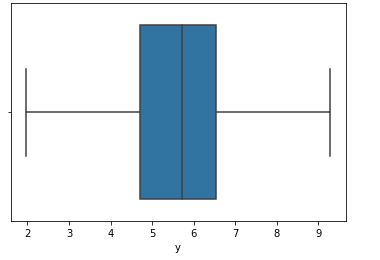
Treating outliers

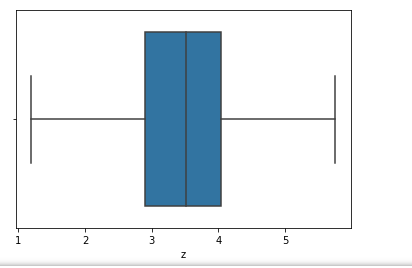
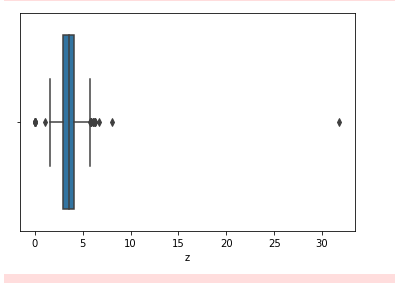
 

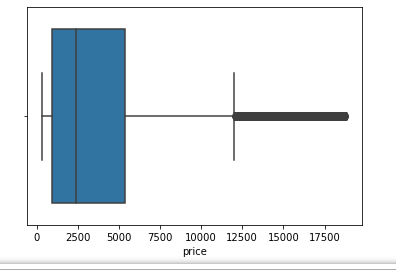
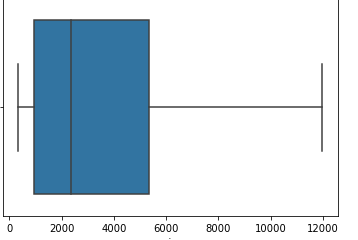
 

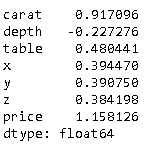
 



Skewness of data:



# Univariate Analysis:

# 

# 

# 

# Bivariate Analysis:

# 

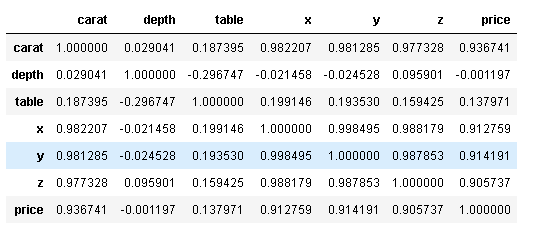
# 

# 

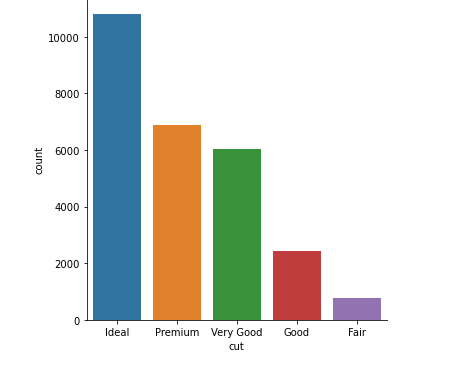
# Heat Map:

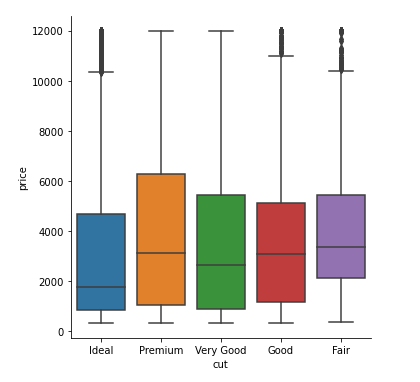
# 

Fig.2 – Heatmap

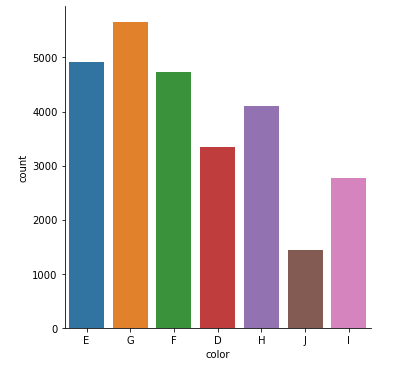


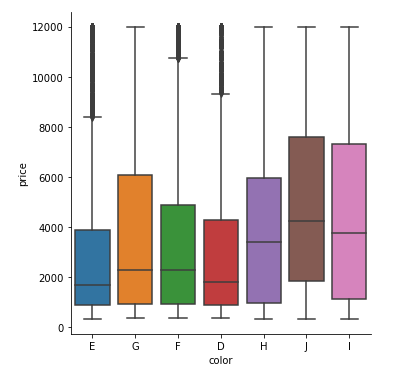
Analyzing categorical Variables:

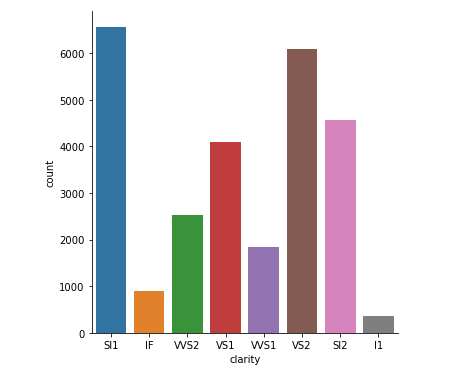


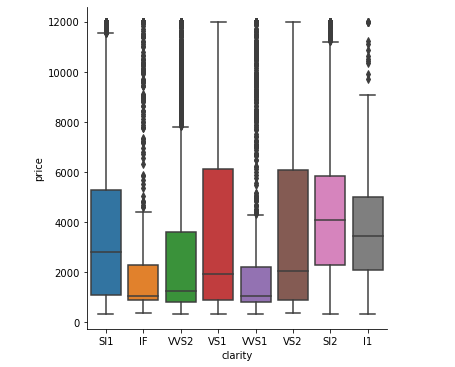


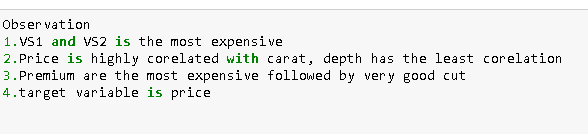
#Premium is the most expensive





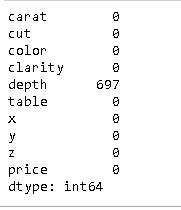






**1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Do you think scaling is necessary in this case?**

Missing values:



# 

# Median values:

# 

# Imputed missing values:

# 

# 

# Shape of data :

# 

# Sclaed and normalized:

# 

# Categorical variables:

# 

# Category cut value counts

# 

# Category color value counts:

# 

# Category clarity value counts

# 

# Converted to data type float64

# 

# Data Description:

# 

# Coefficents of each attributes:

# 

# Linear Regression using stats model

# 

# 

# 

# 

# Prediction on Test data

# 

# 

# 

# 

# 