Diamonds Price Prediction

End_To_End Project
SHAI For AI

Main Goal:

The main goal of the project is to try to predict the price of the Diamond using the features Provided from the Dataset .

The predicted price will be evaluated Using RMSE.

Discovering The Data:

The Data Contains 11 Features:

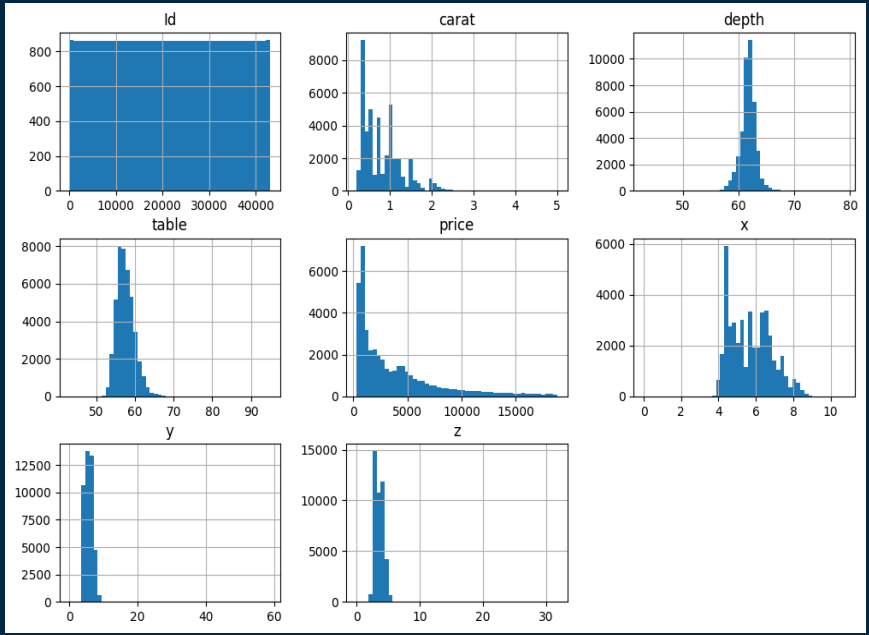
- Numerical Features = {Id , Carat , Table , Depth , X , Y , Z , Price }

- Categorical Features = { Color , Clarity , Cut }

  - The Price Feature is the Target Feature.

  - There are no missing values in all the columns.

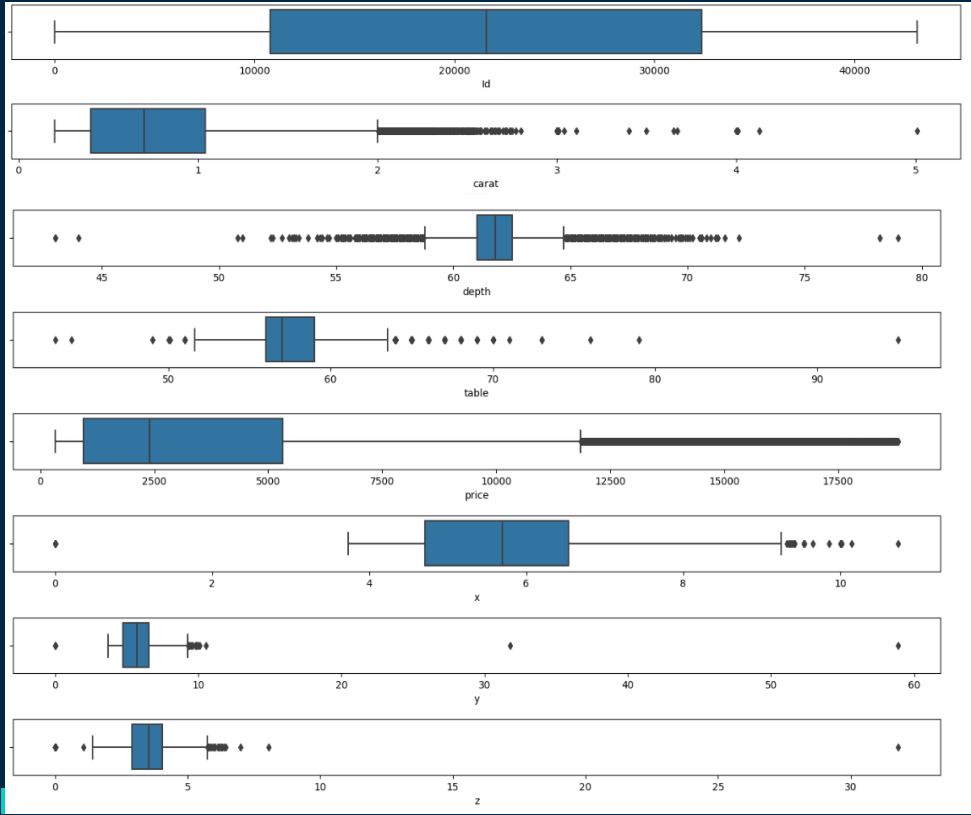
Data Visualization :



From the previous Histogram , We can Notice some insights :

- The Id Column is not useful , It is just a counter Identification .
- The Depth is normally Distributed which is Kinda Good.
- The Price & Carat Columns is very Skewed to the right , and this should be Fixed

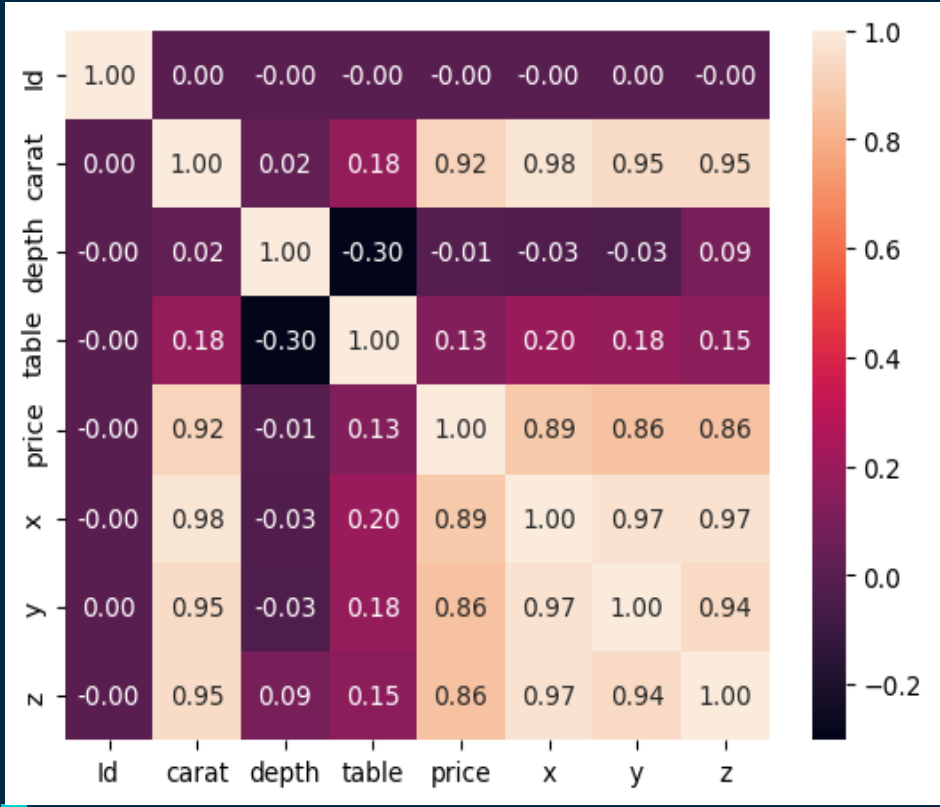
Data Visualization :



From the previous Boxplots , We can Notice some insights :

- In the {x,y,z} Columns we have some zero values which make the object 2D or 1D, so We should drop them.
- There are a lot of Outliers in Most of the columns , So I will try to get rid of them

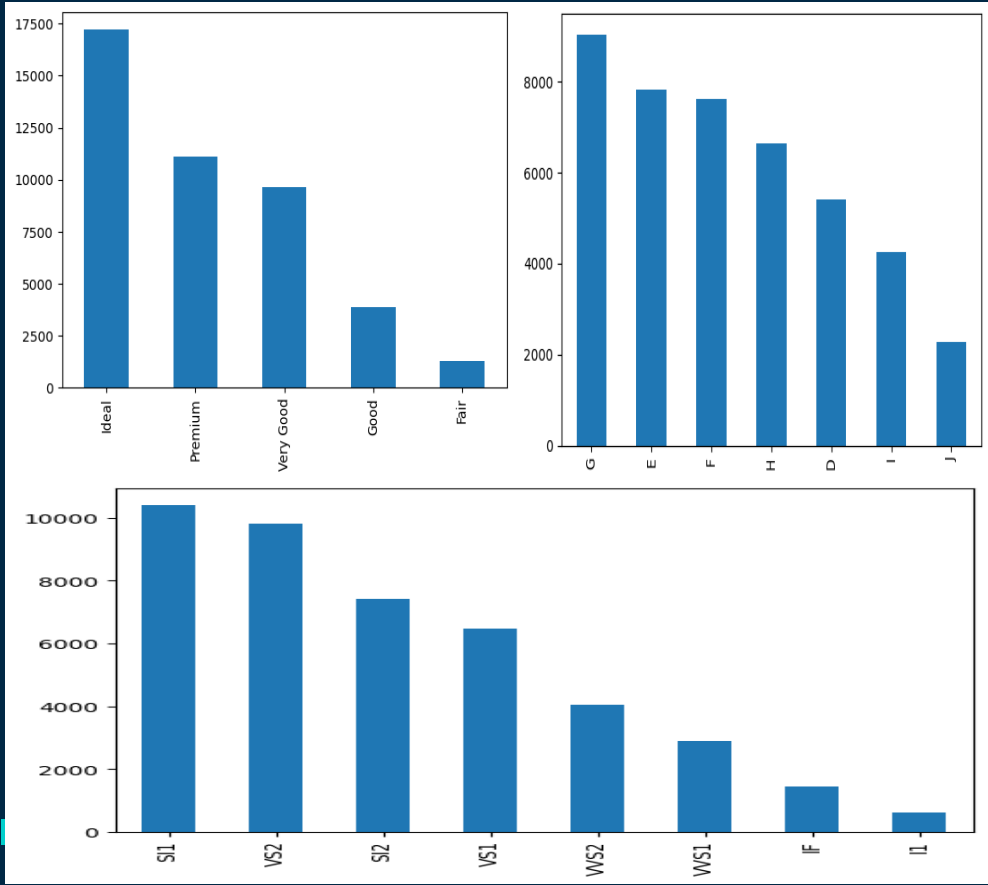
Data Visualization :



From the previous Heatmap , We Can Notice some insights :

- The { x , y , z } features have a very postitive Linear Corrleation between each other and with some other features like { Carat , Price} .
- From the previous insight , We see that it is good to apply feature Combining on these 3 features.

Data Visualization :



From the previous Barplots for Categorical Columns{ Cut , Color , Clarity }, We Can Notice some insights :

- The Disparity between The Occurrence of each categories in the Cut , Clarity Columns , its unbalanced , So We have to fix it .

Data Preprocessing :

The Preprocessing stage went through many steps :

- Handling the 1D , 2D Diamond.
- Handling Outliers for each feature .
- Handling Duplicate Records/Rows.
- Handling the skewed distribution features ,by apply the log function on them.
- Handling the unbalanced categories ,by applying Manual Method / Feature Hasher.

Data Preparing :

The main goal of this stage , is to prepare the data to training by applying some Feature Transformation and Scaling methods.

For Numerical Feature :

- Standard Scaler
- Robust Scaler (Good For features that have many outliers)

For Categorical Feature:

- One Hot Encoder
- Ordinal Encoder
- Manual Encoder (Function I code it myself,its like baseline)

Model Training :

In this stage, I applied different Approaches,which are combined pipelines .
I will mention some of them :

- Manual feature Hasher + Ordinal Encoder + Robust Scaler

- Manual feature Hasher + One Hot Encoder + Standard Scaler

And I applied Each Approach on many models , I will mention some of them :

Linear Regression , SVM , XGBoost Regressor , CatBoost Regressor ,
Random Forest and Decision Tree

Also I applied Some methods , Like Random Search and Grid Search to get the best parameters.

Model Evalution :

I will mention some RMSE Results for some Approaches/Models:

Approach + Model	RMSE
Standad Scaler +One Hot Encoder+ XGBoost	543
Robust Scaler + Ordinal Encoder + Randomized Search + CatBoost	560
Robust Scaler + OneHotEncoder + ManualHasher + CatBoost	651

THANKS