SDAIA PROJECT ( Written Description )

**Introduction:**

In order to fully understand how the pricing system of diamonds worked, it was imperative to collect data on the features taken into account for pricing diamonds as well as their prices. Then, the dataset was further analyzed to build graphs highlighting important qualities of a diamond. Finally, a predictive model was created to predict the prices of diamonds.

**Project Target:**

Prediction How is Diamond Priced Based on its Attributes

## **Dataset Description**

#### Diamonds dataset hosted on Kaggle. and contains 53940 rows and 11 colmun.

#### **Columns represent:**

* **'Unnamed:** 0' - ID of the diamond
* **Cara**t - Measures weight of the diamond , 1 carat = 200 mg
* **x, y, z** - Length, width and height of the diamond respectively
* **Cut** - Identifies quality of the cut (Ideal to Fair)
* Color - Identifies color of the diamond (D - Best to J - Worst)
* Clarity - Measures the clarity of the diamond (IF - Best, I1 - Worst)
* **Table** - Width of top of diamond relative to its widest point
* **depth**, Value derived from width and height of the diamond
* **Price** - Price of the diamond

### **Observations for Data Cleaning:**

* There are no missing values in the dataset
* Continous/Numerical variables are: "carat","depth","table","price",'x', 'y','z'
* Categorical variables are: 'cut','color', 'clarity'
* We can delete the IDs column as it provides no useful statistics.
* In the minimum value of x ,y and z is zero and It doesn’t make any sense to have length\width\depth of a diamond to be zero, so we can drop these rows
* The Price column is the dependent column and all others are independent columns

# Exploratory Data Analysis (EDA)

#### **Numerical Analysis (Target variable vs remaing continous variables) t**

#### here is a good linear relationship between Length, Width, Height and Carat with price variables.

* Correlation coefficient between price with carat and price with dimensions are high (0.9 above). That means features carat and dimensions of the diamond highly affect the target diamond price.
* The features x,y,z also show high coorelation coefficient among each other. So they can be combined into one feature.
* The Depth is inversely related to Price. Why ? scientifically this is because if a Diamond's Depth percentage is too large or small the Diamond will become 'Dark' in appearance because it will no longer return an Attractive amount of light.
* Self Relation ie. of a feature to itself is 1 as expected

### **Categorical Analysis ( Target variable vs remain categories variables)**

**Cut type**

* Even Ideal cut type diamonds are high in the dataset, it doesn't have high price in total.
* Even Fair diamonds are least in the dataset, it secures 2nd position in total price distribution of cut category.
* Premium cut type diamonds in the dataset have high price in total.

**Color type**

* J type has least in number of diamonds in dataset, but it has high price in total price distribution.
* G type are maximum in number in dataset but it doesnot have high price in total price distribution in the dataset.

**Clarity type**

* L1 diamonds are least in number in the dataset but it considerably has high price in total price distribution of clarity type diamonds.
* Sl1 diamonds are maximum in the dataset but it doesn't have high price in total price distribution of clairty type diamonds

### **MODELLING By using Linear Regression Model:**

Score value is = 0.9172785263839015

R-squared value is 0.9166529687435219

RMSE value is = 866.0265999468861

The R\_seguerd is very close and almost identical for training data and we can say here it is just fitting the performance in taring set as just the performance in testing set

**Conclusion:**

R\_squreed is 0.91 which tells the correlation between price vs different independent variables explained by 91%

* A one unit increase in carat will increase of 47 dollars in price.holding all the other features constant
* A one unit increase in cut will increase 53.59 dollars in price,holding all the other features constant
* A one unit increase in color will decrease of 187 dollars in price, holding all the other features constant
* A one unit increase in Clarity will increase of 218.40 dollars in price, holding all the other features constant
* A one unit increase in depth will decrase of 5.32 dollars in price, holding all the other features constant
* A one unit increase in table will decrase of 5.32 dollars in price, holding all the other features constant
* A one unit increase in depth will decrase of 5.32 dollars in price, holding all the other features constant

. Finally, Our linear model is good as the r-square difference in train & test dataset is less than 5%.