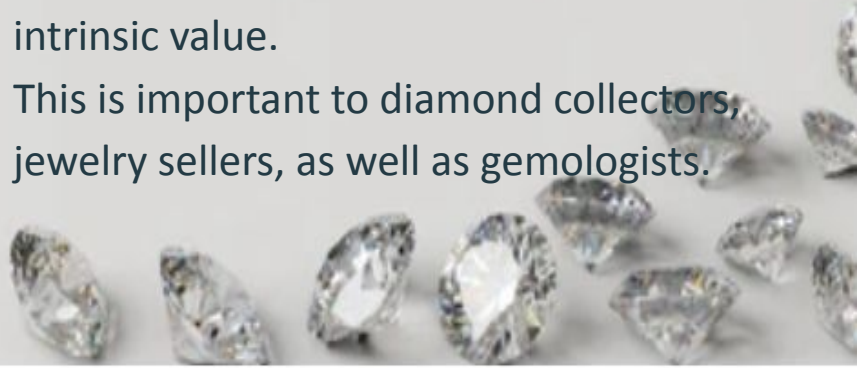
Predicting The Price Of Diamonds

Team 11

Prateek, Hector, Anusha, Kaan

Problem Statement

- Diamond is one of the most prestigious luxury gemstones.
- We would like to build a machine learning model to predict the price of diamonds, based on a set of sample parameters.
- Match its physical characteristics to its intrinsic value.
- This is important to diamond collectors, jewelry sellers, as well as gemologists.



CARAT	CLARITY	COLOUR	CUT
1.25 CTS	FL — FLAWLESS	COLOURLESS: D, E, F	ROUND
1 CT	IF — INTERNALLY FLAWLESS	NEAR COLOURLESS: G, H, I, J	PRINCESS
0.75 CT	VVS_1, VVS_2 — VERY, VERY SLIGHTLY INCLUDED	FAINT: K, L, M	MARQUISE
0.50 CT	VS_1, VS_2 — VERY, SLIGHTLY INCLUDED	VERY LIGHT: N, O, P, Q, R	PEAR
0.25 CT	SI_1, SI_2 — SLIGHTLY INCLUDED	LIGHT: S, T, U, V, W, X, Y, Z	EMERALD
0.10 CT	I_1, I_2, I_3 — INCLUDED		
0.05 CT			

Data Source

- Link to our data from Kaggle.com:

https://www.kaggle.com/code/karnikakapoor/diamond-price-prediction/data

- https://www.diamonds.pro/education/4cs-diamonds/
- Definition of criteria according to

Gemological Institute of America (GIA)

Brief Description of our Data

Data Frame with 53940 rows and 11 columns

Numeric

Index: the unique index counter of each diamond sample

Price: the price (in USD) of each diamond sample

Carat: the weight of diamond sample in carats(1 carat = ⅕ gram)

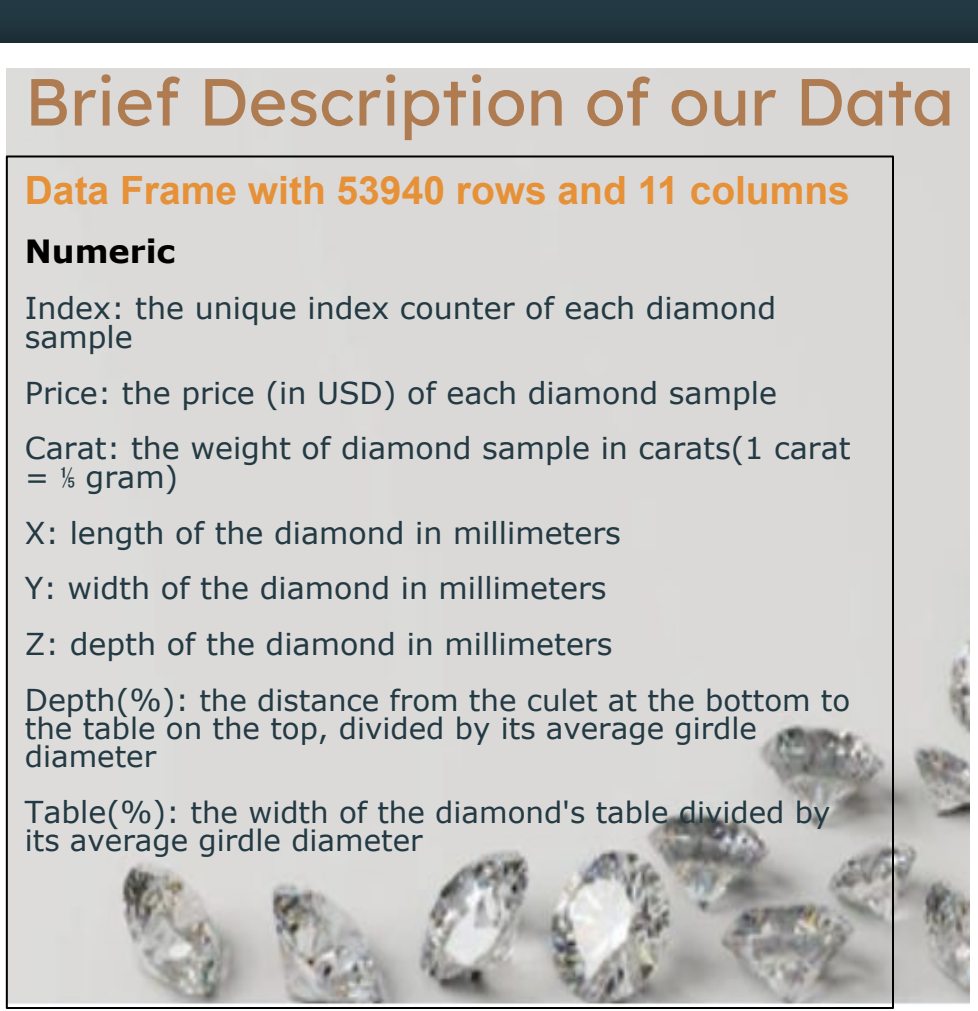
X: length of the diamond in millimeters

Y: width of the diamond in millimeters

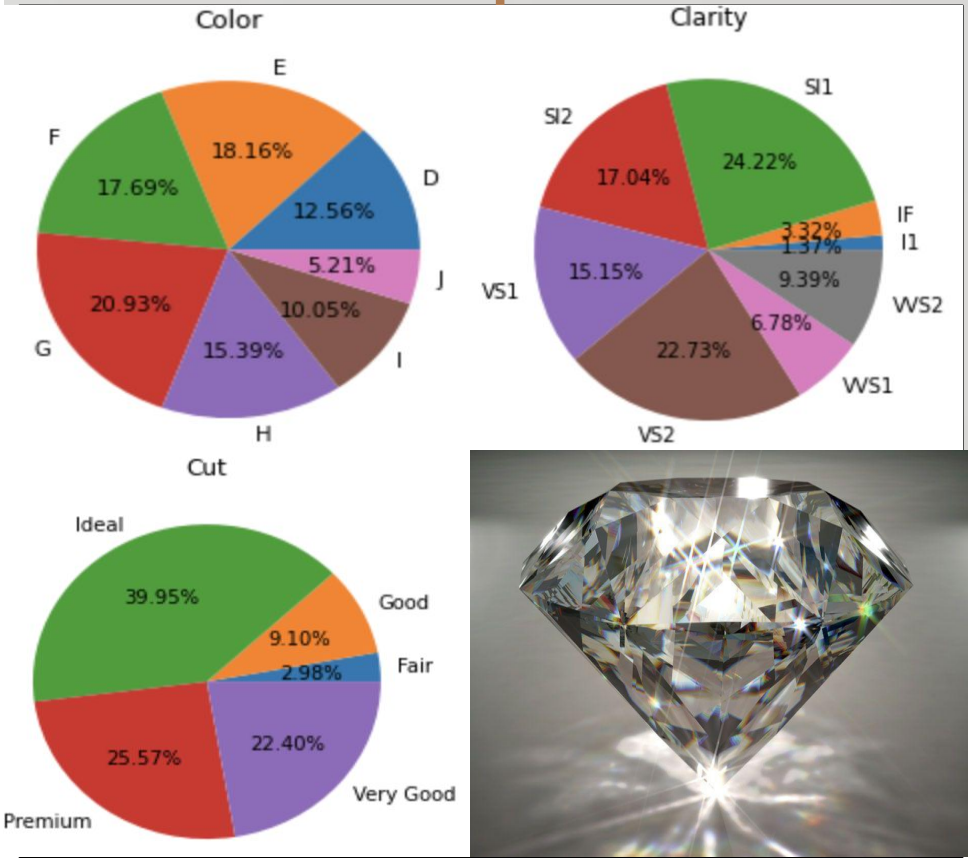
Z: depth of the diamond in millimeters

Depth(%): the distance from the culet at the bottom to the table on the top, divided by its average girdle diameter

Table(%): the width of the diamond's table divided by its average girdle diameter



Brief Description of our Data



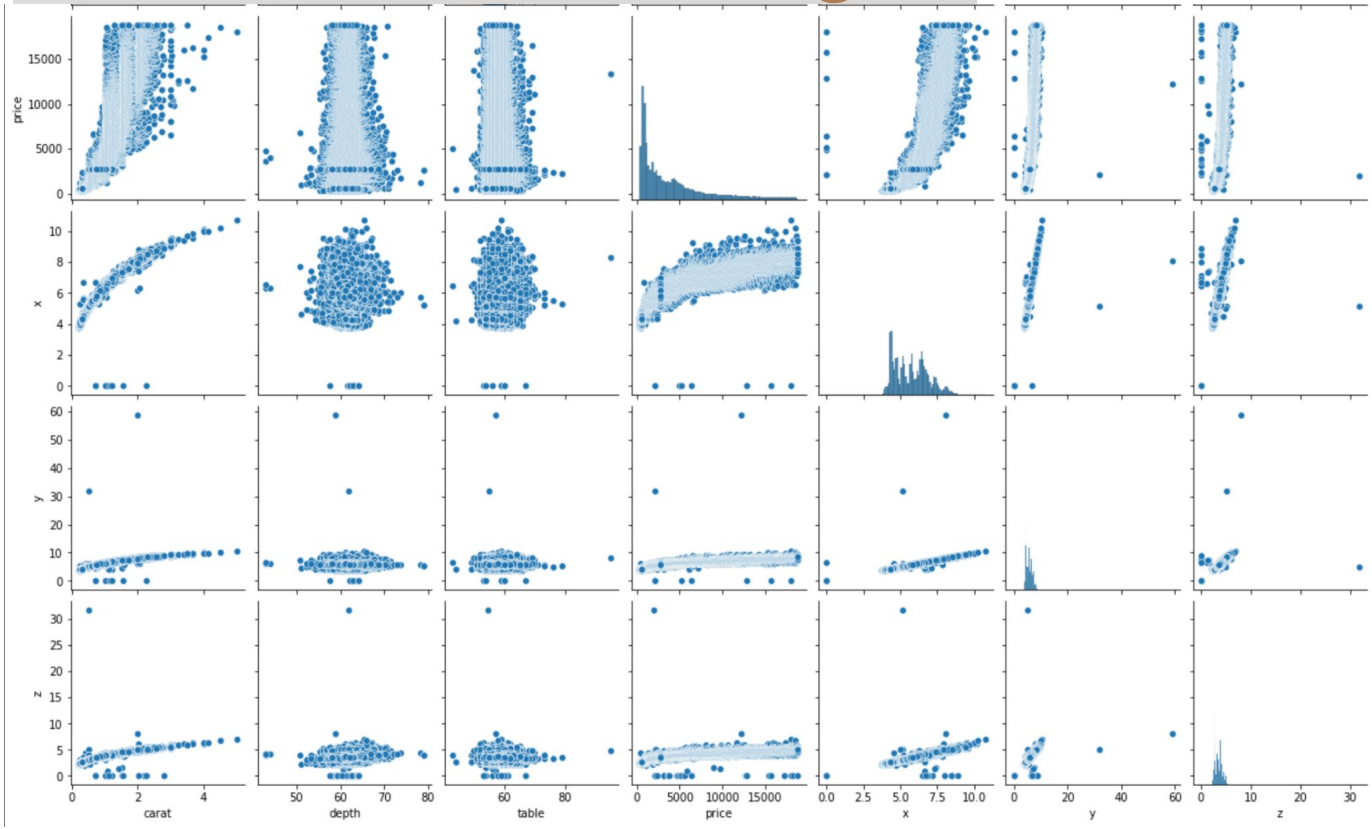
Categorical

Cut: Describe cut quality of the diamond. Quality in increasing order Fair, Good, Very Good, Premium, Ideal

Color: Color of the diamond, with D being the best and J the worst

Clarity: How obvious inclusions are within the diamond(in order from best to worst, FL = flawless, I3= level 3 inclusions), FL, IF, VVS1, VVS2, VS1, VS2, SI1, SI2

Data Pre-Processing of Numeric Variables



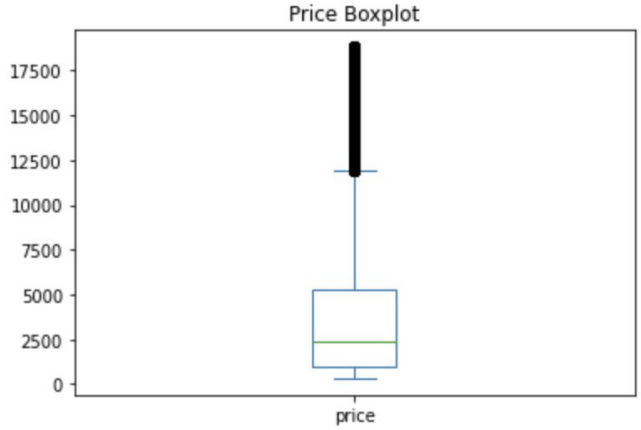
- Acquire the dataset.
- Import crucial dependencies, & dataset.
- Handling missing values & outliers
- Encoding the categorical data.

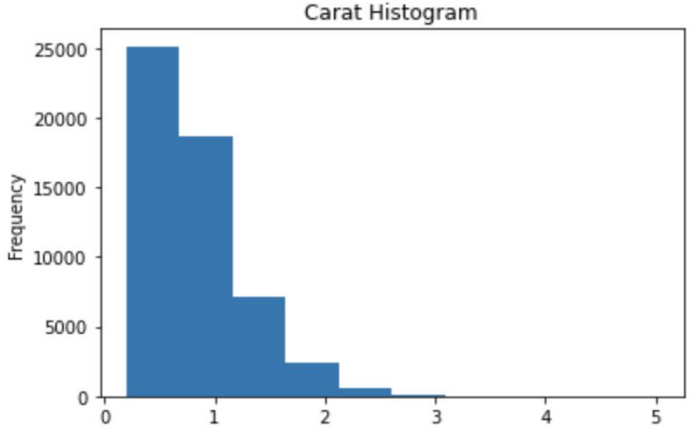
	count	mean	std	min	25%	50%	75%	max
carat	53940.0	0.797940	0.474011	0.2	0.40	0.70	1.04	5.01
depth	53940.0	61.749405	1.432621	43.0	61.00	61.80	62.50	79.00
table	53940.0	57.457184	2.234491	43.0	56.00	57.00	59.00	95.00
price	53940.0	3932.799722	3989.439738	326.0	950.00	2401.00	5324.25	18823.00
x	53940.0	5.731157	1.121761	0.0	4.71	5.70	6.54	10.74
y	53940.0	5.734526	1.142135	0.0	4.72	5.71	6.54	58.90
z	53940.0	3.538734	0.705699	0.0	2.91	3.53	4.04	31.80

POINTS TO NOTE :-

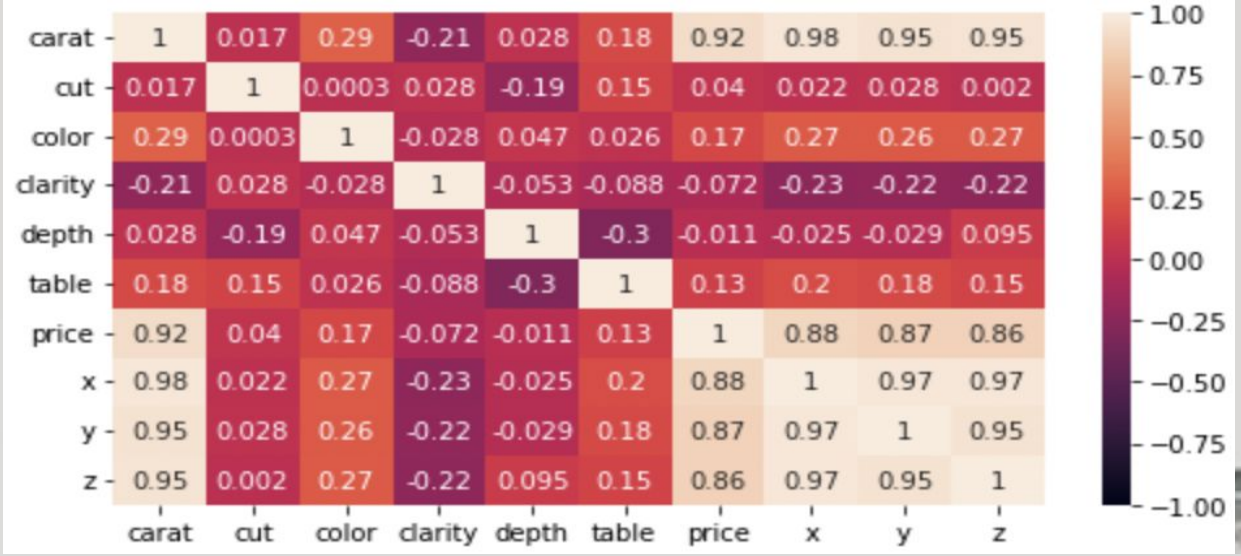
Its was noticed that some features like x,y,z dimensions, required to be preprocessed for our ML model.

- Min value of “x”, “y”, “z” are dimensionless.
- Outliers Existence in dimensions, table & depth.





Correlation Heatmap



- Categorical variables (cut, color, clarity) is not significantly correlated with price
- Dimensions (x,y,z) is positively correlated with price
- Carat, as expected, positively correlated with price

Anticipated Results & Implications

Purpose: Helping jewelrist and potential jewelry buyers to determine the value of diamonds that have differing characteristics.

Actions:

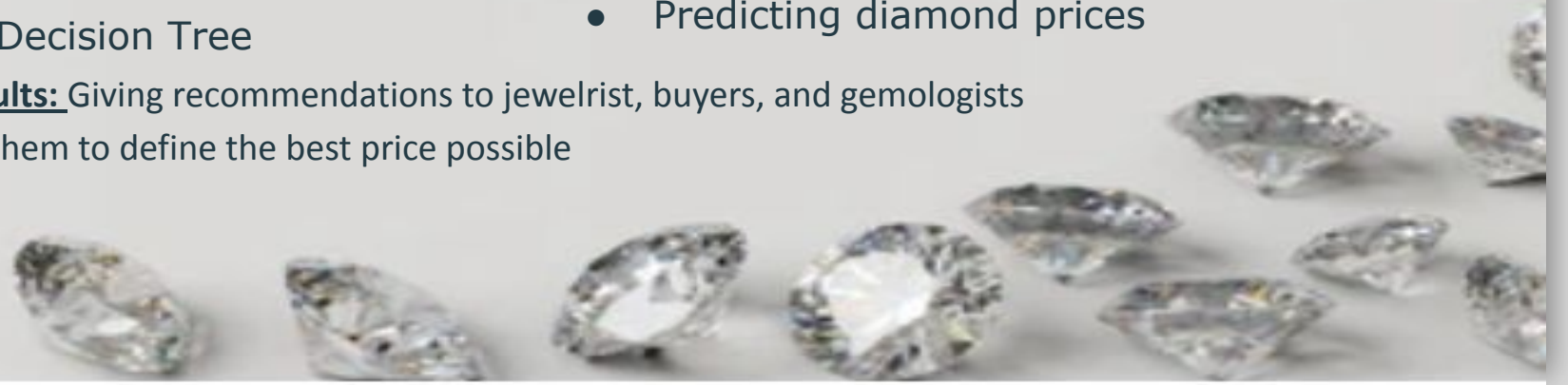
a) ML Models:

- Linear Regression
- KNN
- Decision Tree

b) Prediction:

- Evaluating coefficients and significant variables
- Predicting diamond prices

Results: Giving recommendations to jewelrist, buyers, and gemologists for them to define the best price possible




Thank you for
listening :)