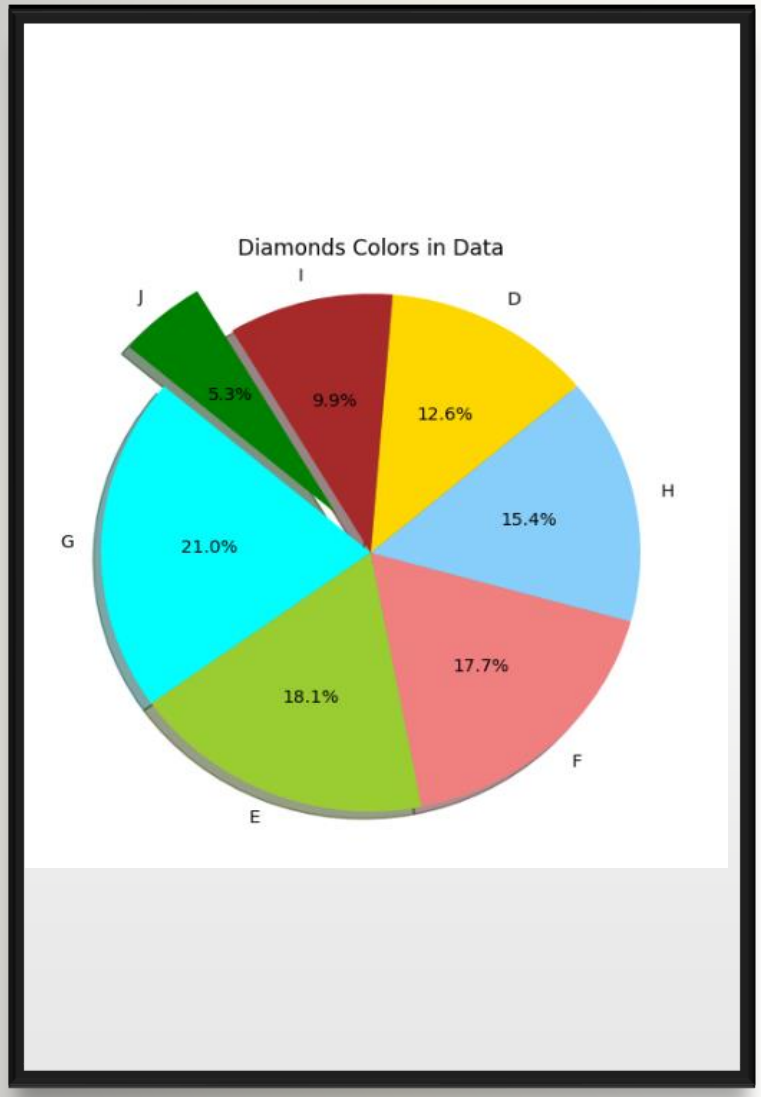
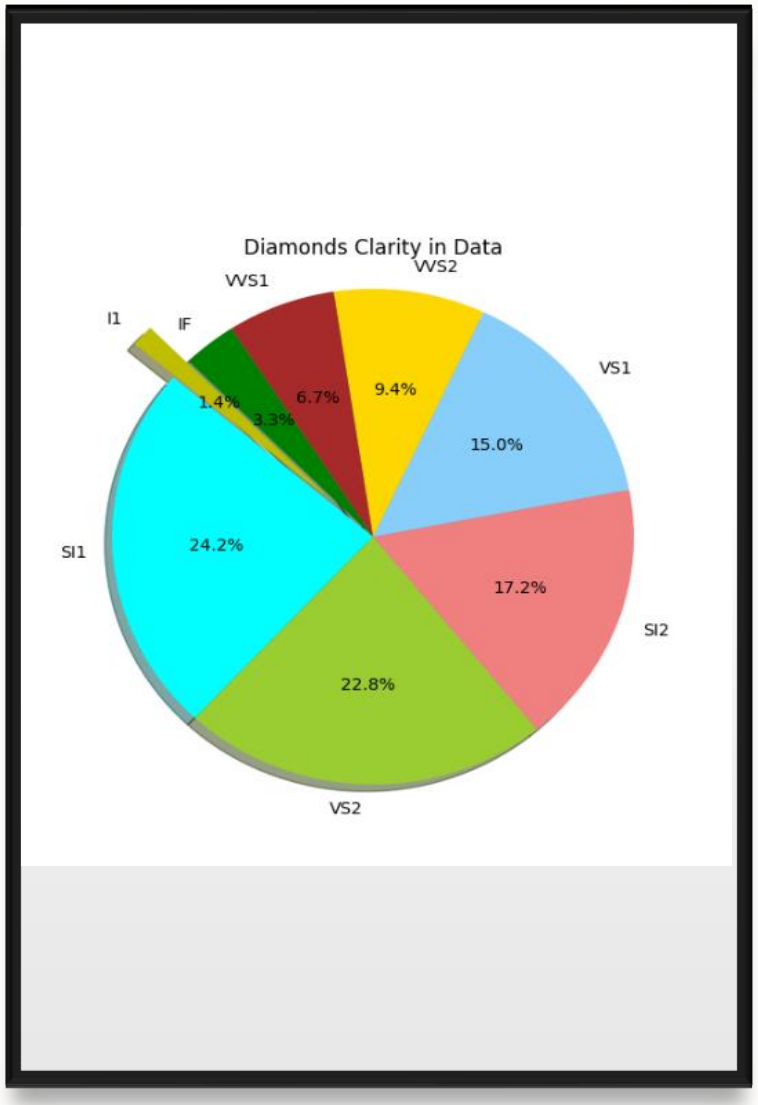
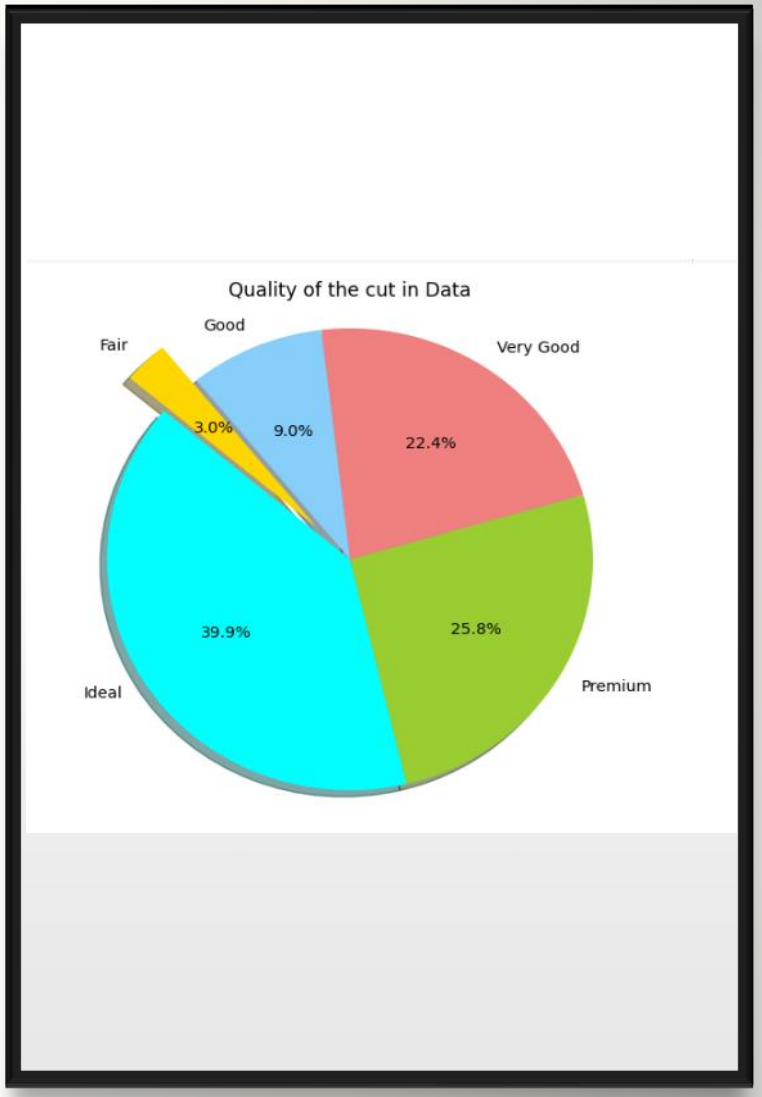

DIAMOND PRICE
PREDICTOR
SHAI for AI

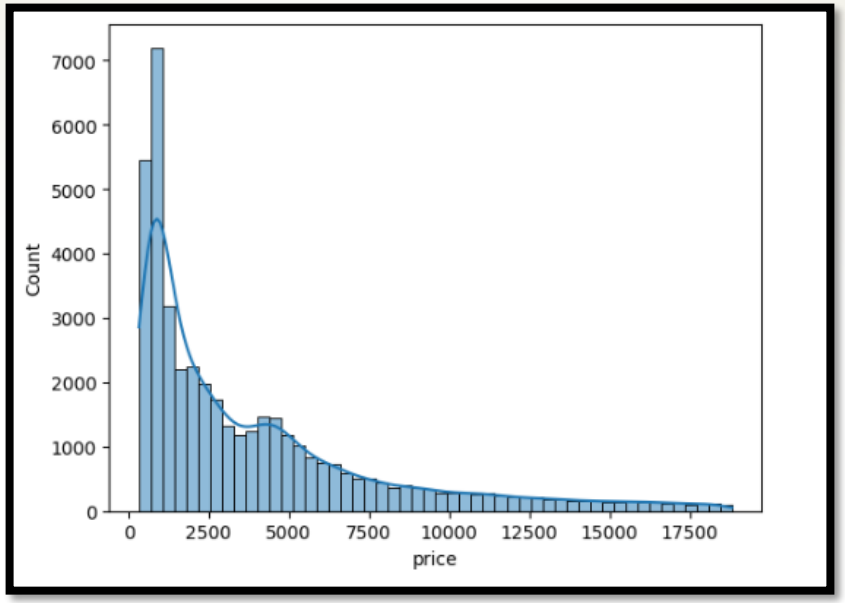
EXPLOTERAY DATA ANALYSIS

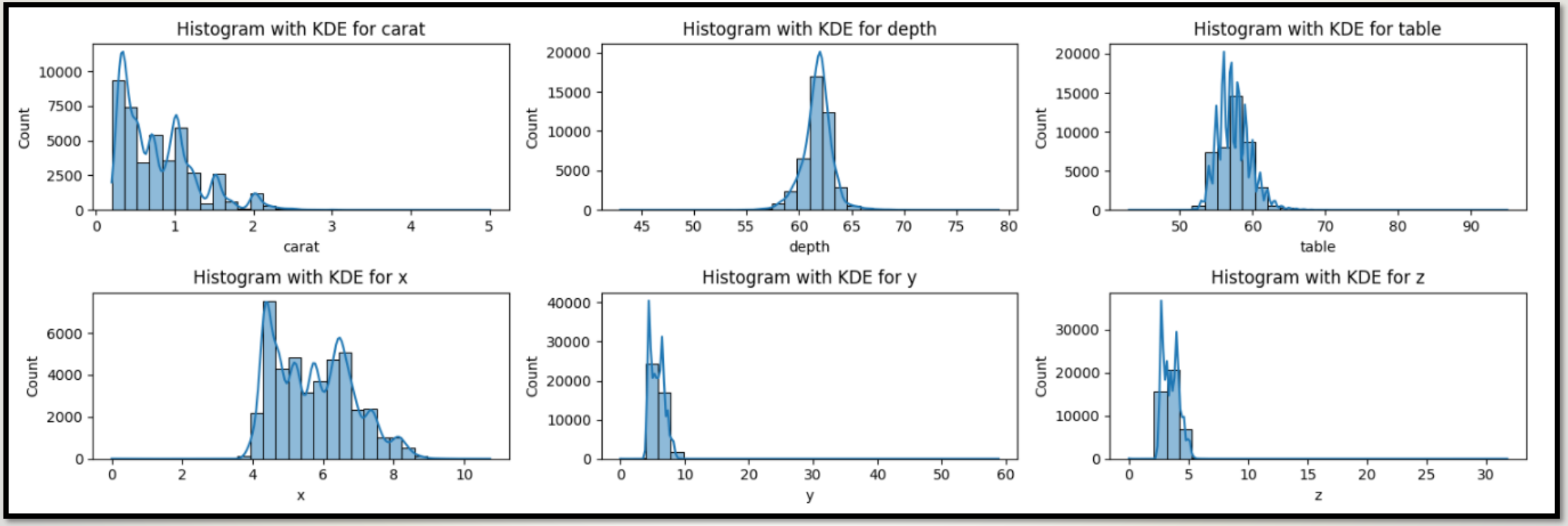
DATA VISUALIZATION

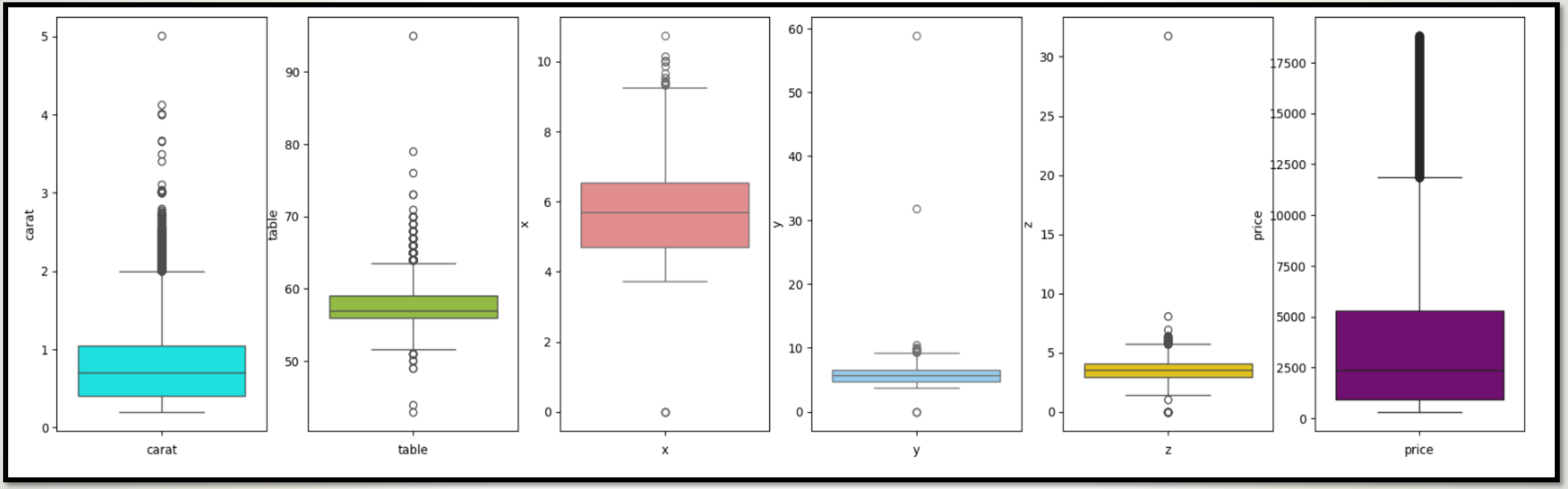

Diamonds Colors in Data
J
I
D
H
G
F
E
5.3%
9.9%
12.6%
15.4%
21.0%
17.7%
18.1%


Diamonds Clarity in Data
VVS1
VVS2
I1
IF
VS1
SI1
SI2
VS2
1.4%
3.3%
6.7%
9.4%
15.0%
24.2%
17.2%
22.8%


Quality of the cut in Data
Fair
Good
Very Good
Ideal
Premium
3.0%
9.0%
22.4%
39.9%
25.8%


Count
price

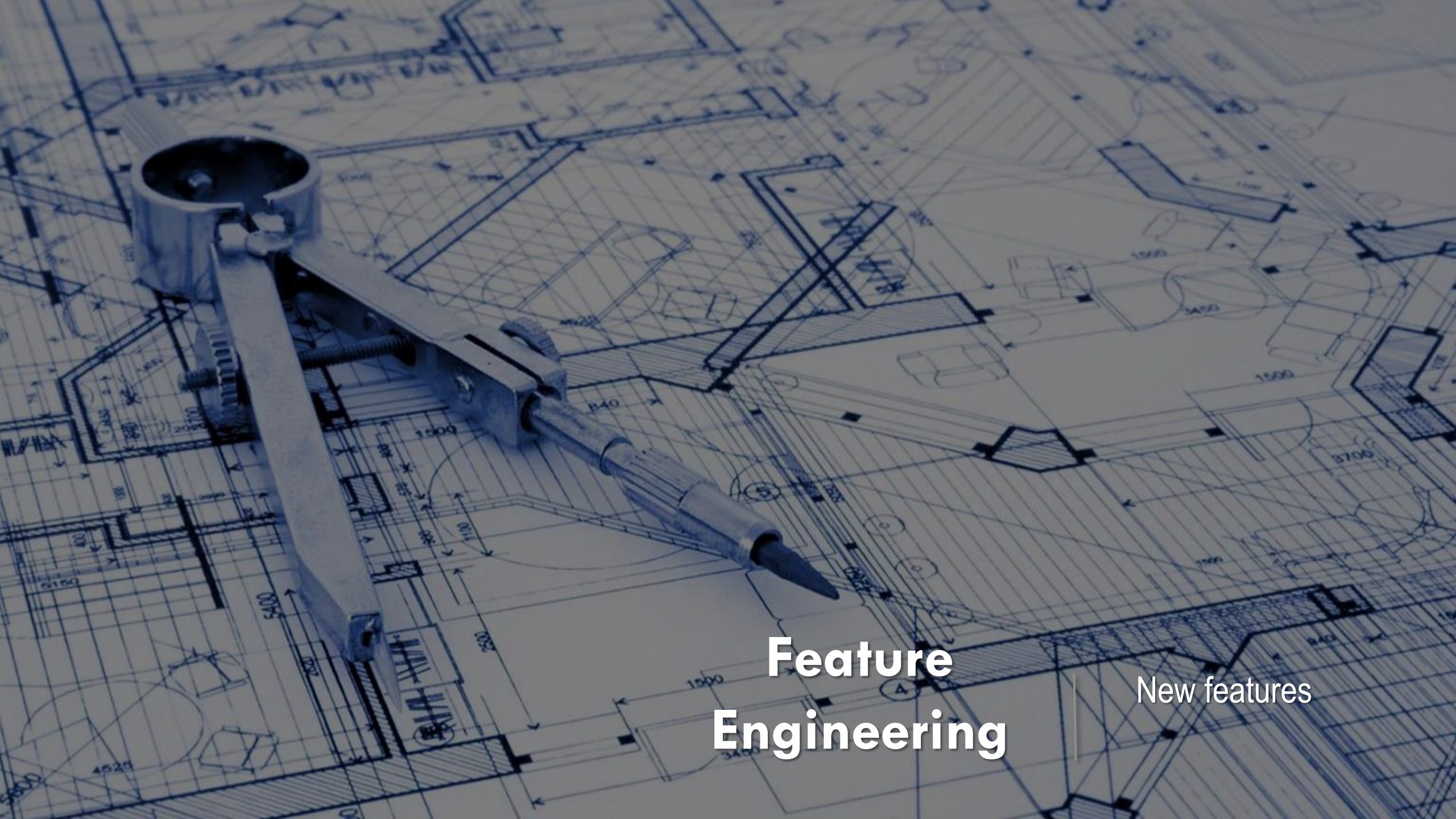

Histogram with KDE for carat
Count
carat
Histogram with KDE for depth
Count
depth
Histogram with KDE for table
Count
table
Histogram with KDE for x
Count
x
Histogram with KDE for y
Count
y
Histogram with KDE for z
Count
z


carat
table
x
y
z
price


Correlation Matrix
carat 1.00 -0.14 -0.29 -0.35 0.02 0.18 0.92 0.98 0.95 0.95
cut -0.14 1.00 0.02 0.19 -0.22 -0.43 -0.06 -0.13 -0.12 -0.15
color -0.29 0.02 1.00 -0.03 -0.05 -0.02 -0.17 -0.27 -0.26 -0.27
clarity -0.35 0.19 -0.03 1.00 -0.06 -0.16 -0.15 -0.37 -0.36 -0.37
depth 0.02 -0.22 -0.05 -0.06 1.00 -0.30 -0.01 -0.03 -0.03 0.09
table 0.18 -0.43 -0.02 -0.16 -0.30 1.00 0.13 0.20 0.18 0.15
price 0.92 -0.06 -0.17 -0.15 -0.01 0.13 1.00 0.89 0.86 0.86
x 0.98 -0.13 -0.27 -0.37 -0.03 0.20 0.89 1.00 0.97 0.97
y 0.95 -0.12 -0.26 -0.36 -0.03 0.18 0.86 0.97 1.00 0.94
z 0.95 -0.15 -0.27 -0.37 0.09 0.15 0.86 0.97 0.94 1.00
carat cut color clarity depth table price x y z
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

Notes

- Category features are not balanced.
- Some features, such as price, are highly skewed.
- Features have different ranges.
- Features contain many outliers.
- x, y, z, and carat are highly correlated with each other.
- Cut and depth features are not correlated with price.

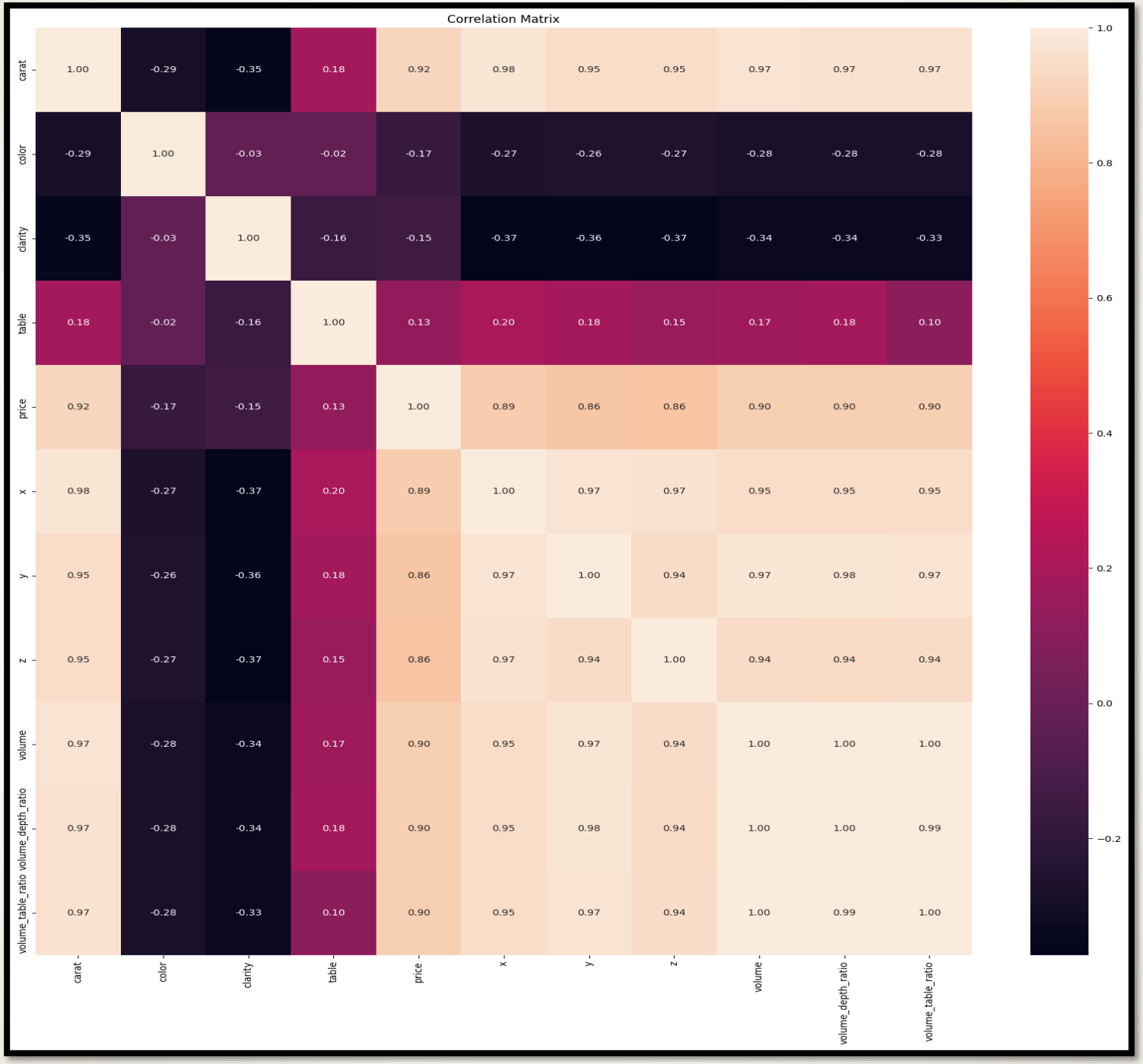


Feature Engineering

New features

Feature Engineering

- **Volume:** Calculated as x × y × z.
- **Volume to Depth Ratio:** Ratio of volume to depth, indicating diamond compactness.
- **Volume to Table Ratio:** Ratio of volume to table, adding analytical dimensions.
- **Removed Uncorrelated Features:** Dropped "depth" and "cut" due to weak price correlation.

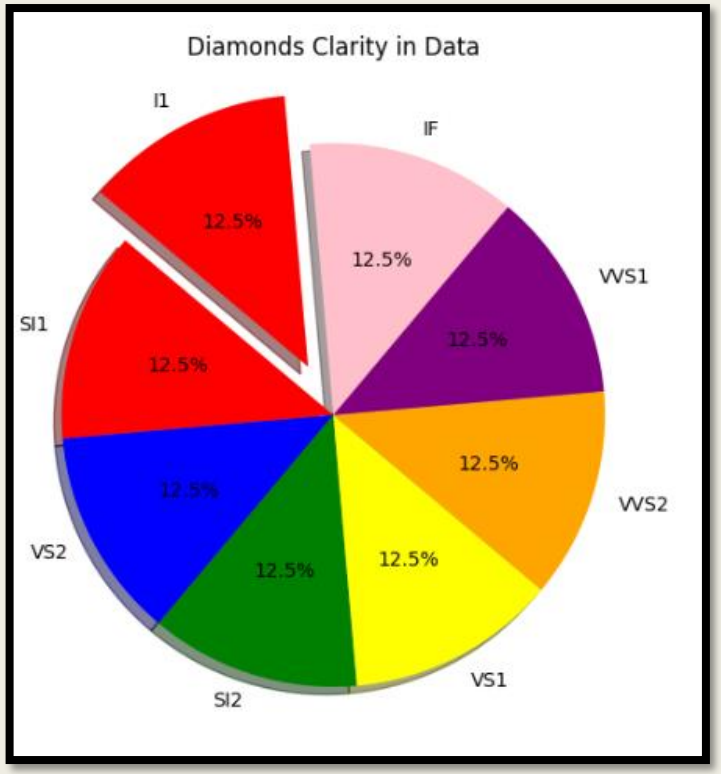

Correlation Matrix
carat 1.00 -0.29 -0.35 0.18 0.92 0.98 0.95 0.95 0.97 0.97 0.97
color -0.29 1.00 -0.03 -0.02 -0.17 -0.27 -0.26 -0.27 -0.28 -0.28 -0.28
clarity -0.35 -0.03 1.00 -0.16 -0.15 -0.37 -0.36 -0.37 -0.34 -0.34 -0.33
table 0.18 -0.02 -0.16 1.00 0.13 0.20 0.18 0.15 0.17 0.18 0.10
price 0.92 -0.17 -0.15 0.13 1.00 0.89 0.86 0.86 0.90 0.90 0.90
x 0.98 -0.27 -0.37 0.20 0.89 1.00 0.97 0.97 0.95 0.95 0.95
y 0.95 -0.26 -0.36 0.18 0.86 0.97 1.00 0.94 0.97 0.98 0.97
z 0.95 -0.27 -0.37 0.15 0.86 0.97 0.94 1.00 0.94 0.94 0.94
volume 0.97 -0.28 -0.34 0.17 0.90 0.95 0.97 0.94 1.00 1.00 1.00
volume_depth_ratio 0.97 -0.28 -0.34 0.18 0.90 0.95 0.98 0.94 1.00 1.00 0.99
volume_table_ratio 0.97 -0.28 -0.33 0.10 0.90 0.95 0.97 0.94 1.00 0.99 1.00
carat color clarity table price x y z volume volume_depth_ratio volume_table_ratio
1.0 0.8 0.6 0.4 0.2 0.0 −0.2

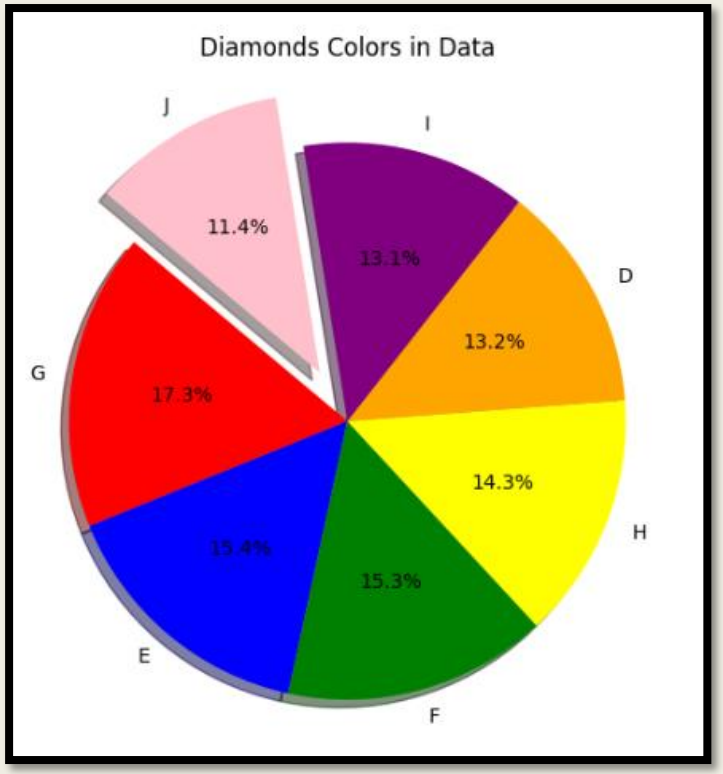
Data Processing

Fix Data

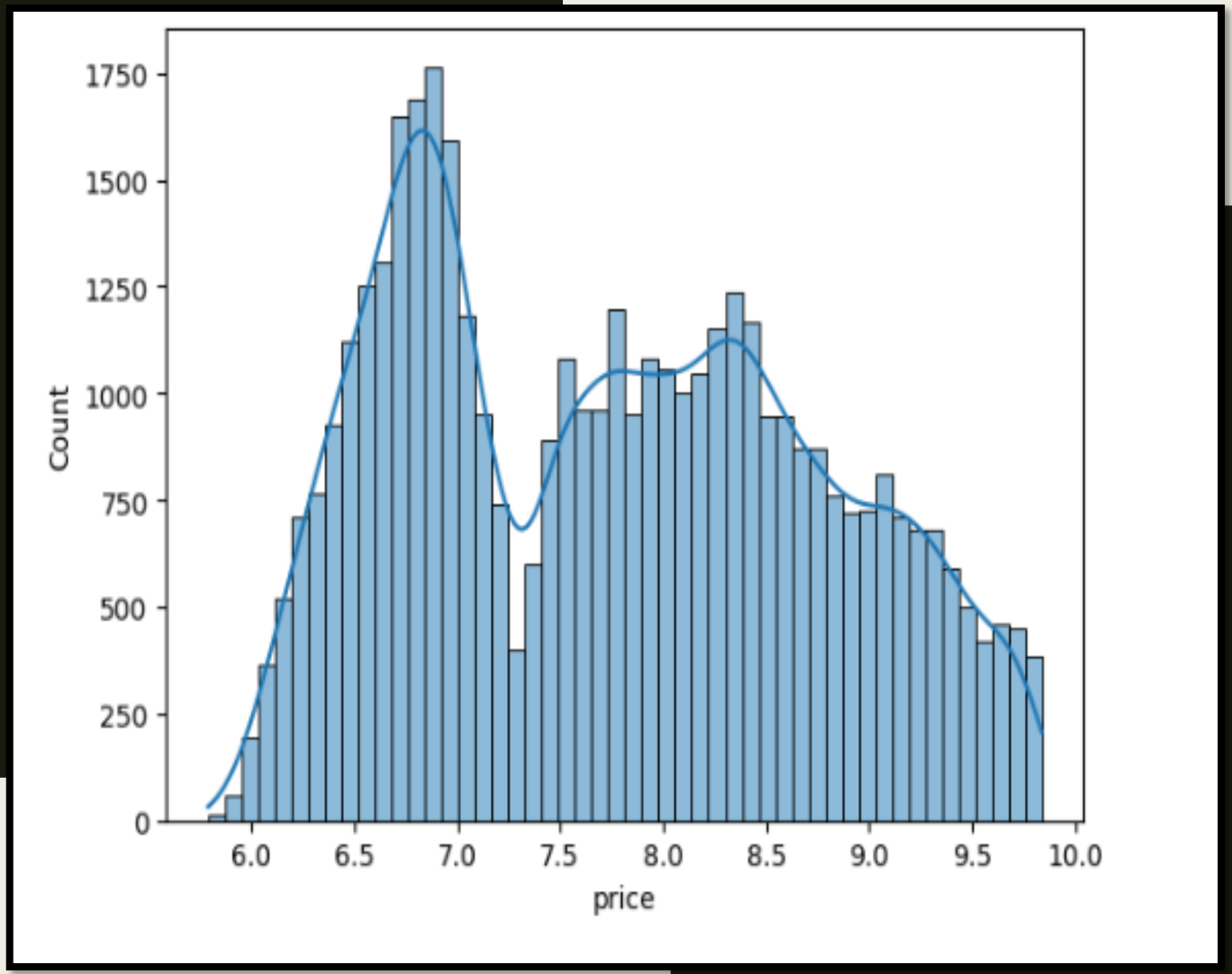
Data Processing

- **Oversampling and Undersampling:** balance categorical features.
- **Logarithm Transformation:** Normalize skewed data, equalize the range, and remove outliers.
- **Normalization:** Achieve a mean of zero before model training.





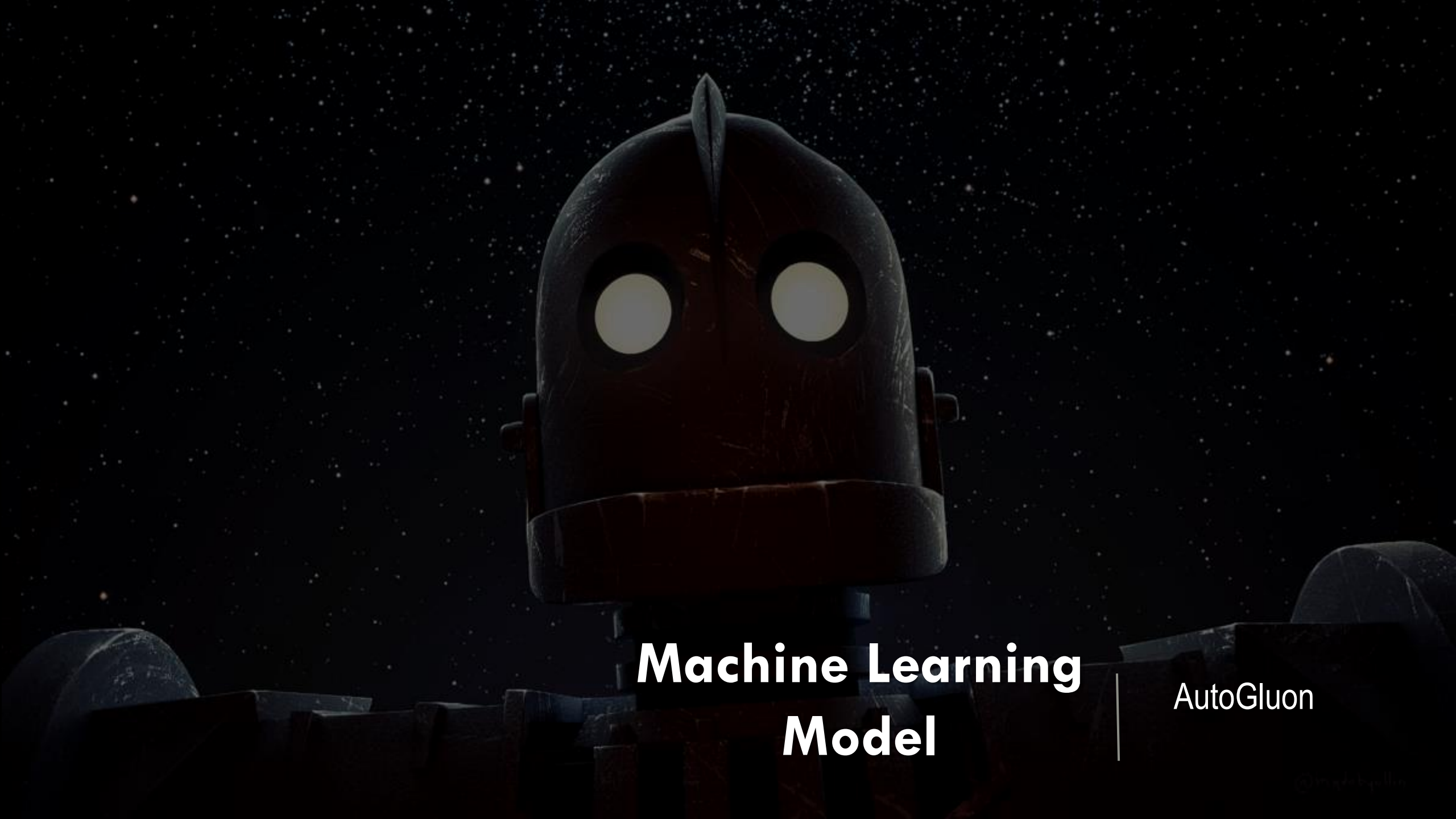
OVERSAMPLING & UNDERSAMPLING



LOGARITHM TRANSFORMATION FOR FEATURES

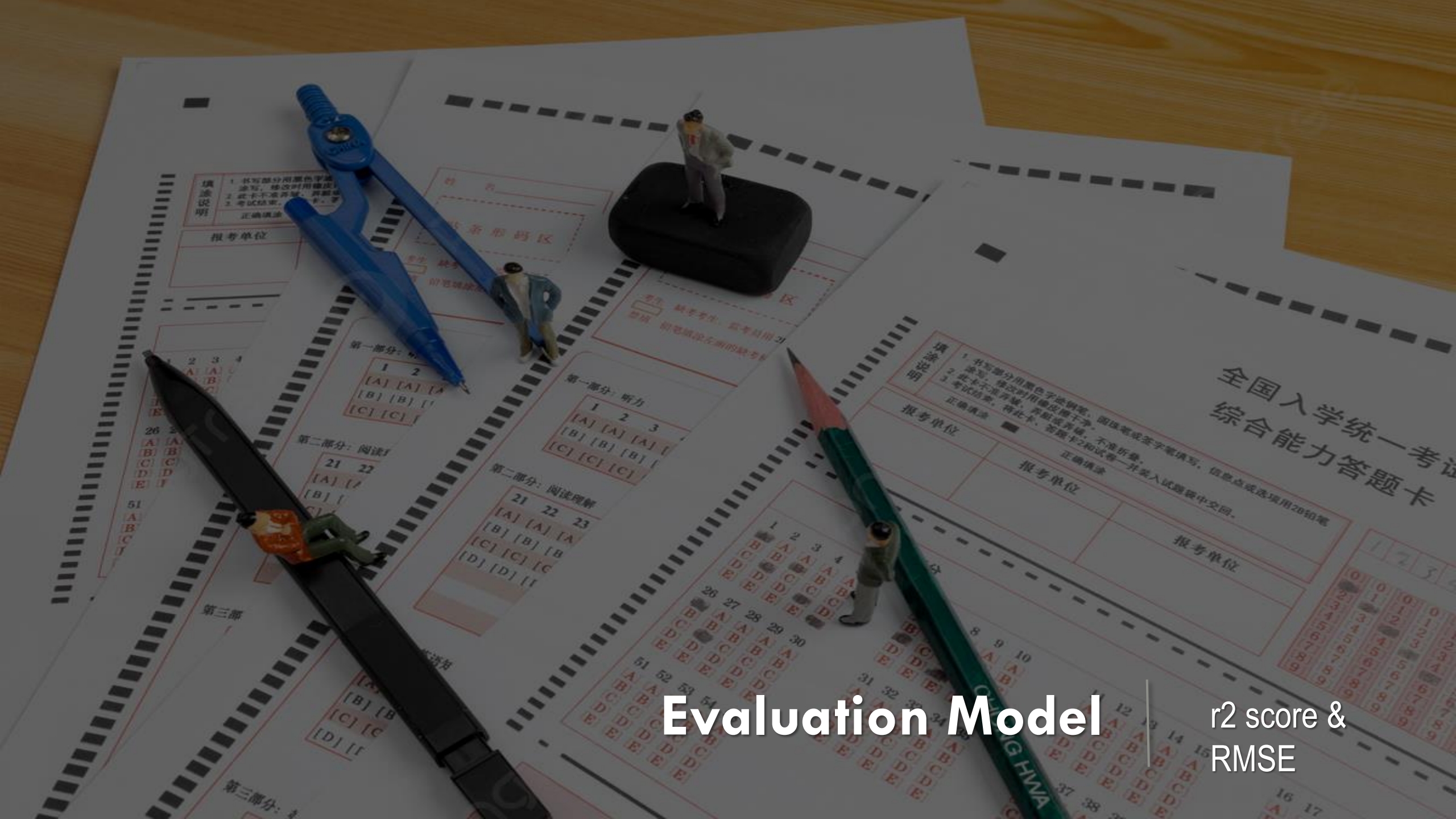
	carat	table	x	y	z	volume	volume_depth_ratio	volume_table_ratio	price
count	4.312800e+04	4.312800e+04	4.312800e+04	4.312800e+04	4.312800e+04	4.312800e+04	4.312800e+04	4.312800e+04	43128.000000
mean	-1.370737e-16	-1.948028e-15	1.133494e-15	-9.542440e-16	-5.588390e-16	6.010156e-16	4.323094e-16	1.212575e-16	7.745299
std	1.000012e+00	1.000012e+00	1.000012e+00	1.000012e+00	1.000012e+00	1.000012e+00	1.000012e+00	1.000012e+00	1.001382
min	-1.411102e+00	-7.315294e+00	-1.070661e+01	-1.081159e+01	-9.078761e+00	-7.593446e+00	-2.648100e+00	-2.722815e+00	5.789960
25%	-9.189991e-01	-6.142355e-01	-9.352538e-01	-9.351735e-01	-9.329006e-01	-9.314457e-01	-9.301276e-01	-9.382759e-01	6.852243
50%	-1.173909e-01	-1.640217e-01	2.321928e-02	3.103315e-02	3.656647e-02	5.378109e-02	-6.271268e-02	-6.648548e-02	7.723562
75%	6.497424e-01	7.135756e-01	7.835498e-01	7.827099e-01	7.423372e-01	7.716198e-01	7.407497e-01	7.351745e-01	8.530307
max	4.572998e+00	8.160699e+00	3.220093e+00	1.255471e+01	1.201249e+01	5.712088e+00	7.777952e+00	7.625078e+00	9.841399

NORMALIZATION


Machine Learning
Model
AutoGluon

Machine Learning Model

- We have used **AutoGluon** to select our machine learning model.
- It determined that the best-fitting model for our data is **WeightedEnsemble_L3**.



Evaluation Model

r2 score &
RMSE

Evaluation Model

- Our model achieved an R-squared score of 0.98 and an RMSE of 538


THANK YOU