ADVANCED ANALYSIS OF GEMSTONE PRICE PREDICTION

Group 02

ST 3082 - Statistical Learning I
Project 1

s16644 - Imesh Chavindu
s16690 - Gajanan Umasuthan
s16835 - Thabeetha Shenali

INTRODUCTION

🎯BJECTIVE

- Identifying most significant influence on diamond price
- Estimate the price of a stone using its characteristics?



ABOUT THE DATASET

U.S. is the most significant consumer of diamonds, accounting for 54% of the global demand
(which totaled approximately $87 billion in 2021).

Because of this market dominance, the dataset provides detailed pricing and attribute information for 26,967 diamonds specifically located in the United States.

FROM kaggle

VARIABLES OF THE DATASET

Varible	Type	Description
Independent Variables		
Carat	Numerical	Carat weight of the diamond.
Cut	Categorical	Quality of the cut. (Fair, Good, Very Good, Premium, Ideal).
Color	Categorical	Color of the Diamond, from D (best) to J (worst).
Clarity	Categorical	A measurement of how clear the diamond is (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best)).
Depth	Numerical	Depth percentage, calculated as the ratio of z (height) to the mean of x and y dimensions.
Table	Numerical	Width of the top of the diamond relative to the widest point (also called the table percentage)
X	Numerical	Length of the diamond in mm.
Y	Numerical	Width of the diamond in mm.
Z	Numerical	Height of the diamond in mm.
Target Variable		
Price	Numerical	The Price of the Diamond. (US $)

DATA PREPROCESSING

REMOVING DUPLICATES

64 observations are duplicate.

Removal of Faulty Dimensional Values

9 observations contain at least one dimension with a value of zero.

Handle Missing Values

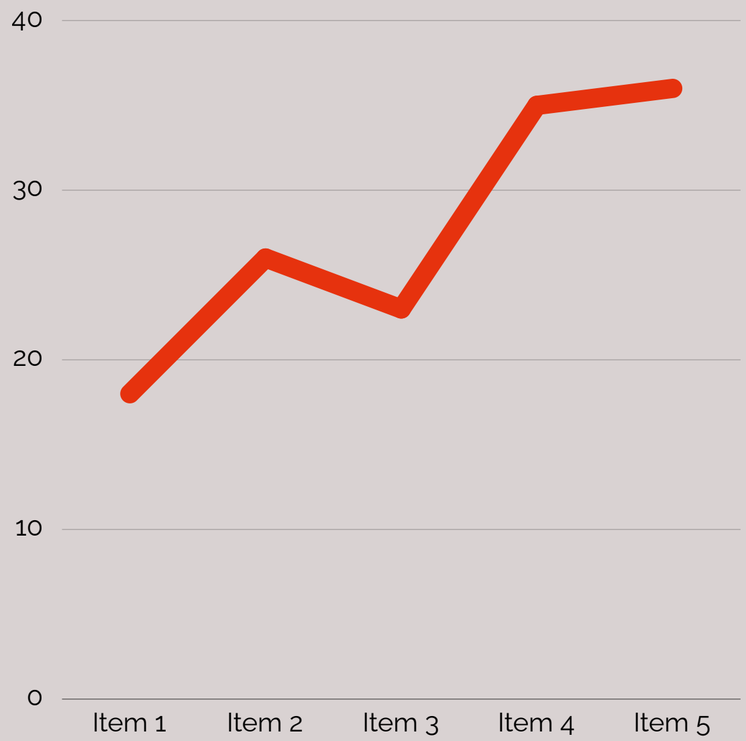
697 observations were missing in the Depth column. We removed them.

Handle Outliers

Didn’t remove

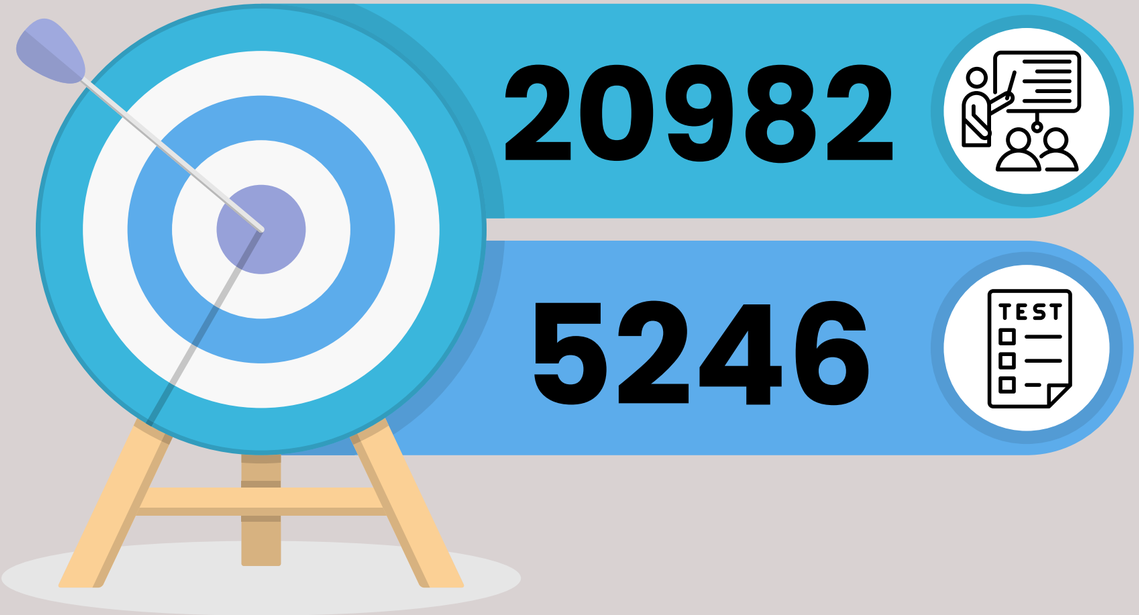
Feature enginnering

Ordinal encoding
Standard scale
log transform price variable



- **Volume:** volume=x*y*z
- **Average girdle diameter:** avg_girdle_diameter=(x+y)/2
- **Average table diameter:** avg_table_diameter= (table*avg_girdle_diameter)/100
- **Density:** density=carat/volume
- **Length to width ratio:** length2width_ratio=max(x,y)/min(x,y)

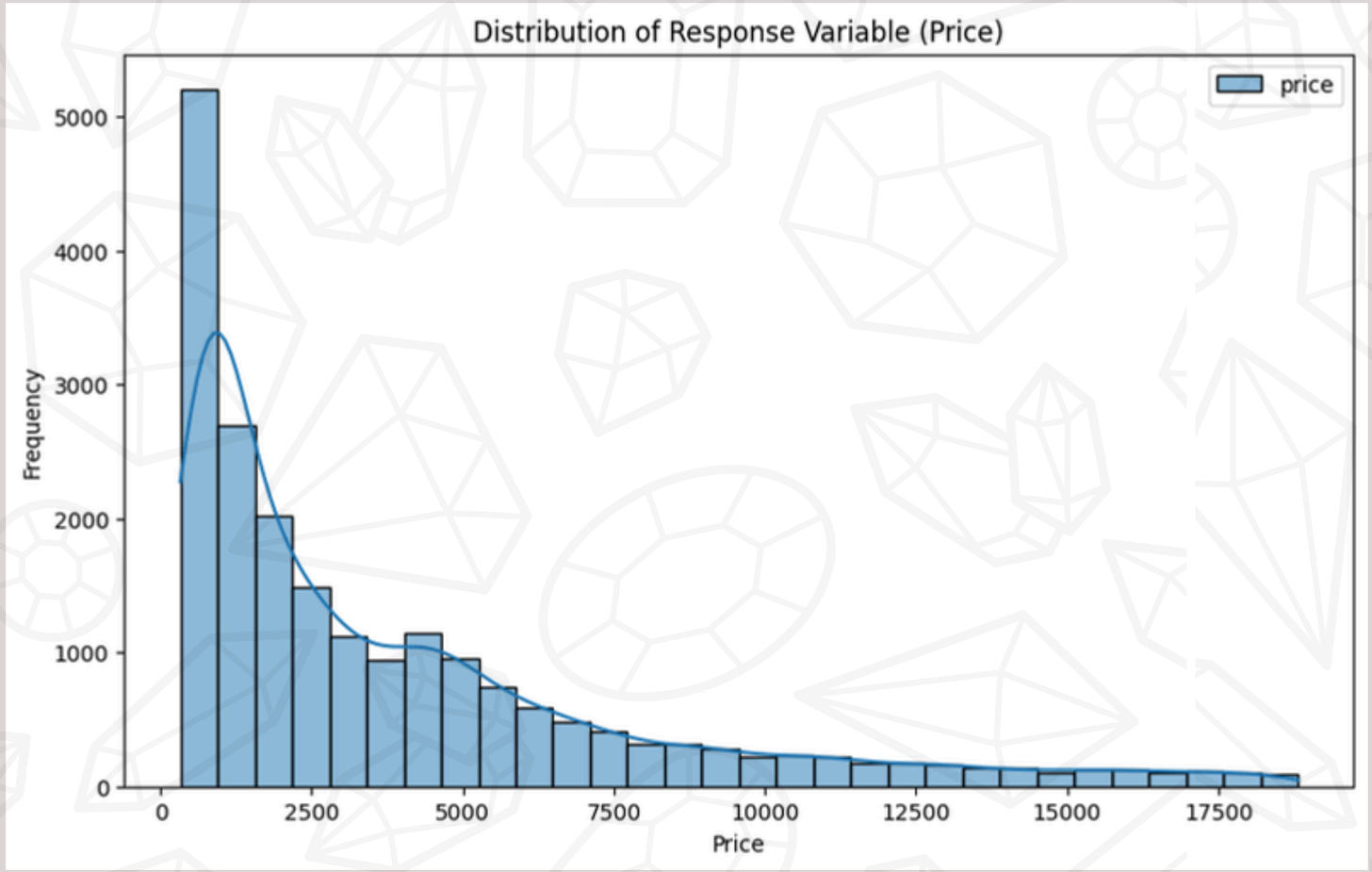
Finally 26228 Gems



20982

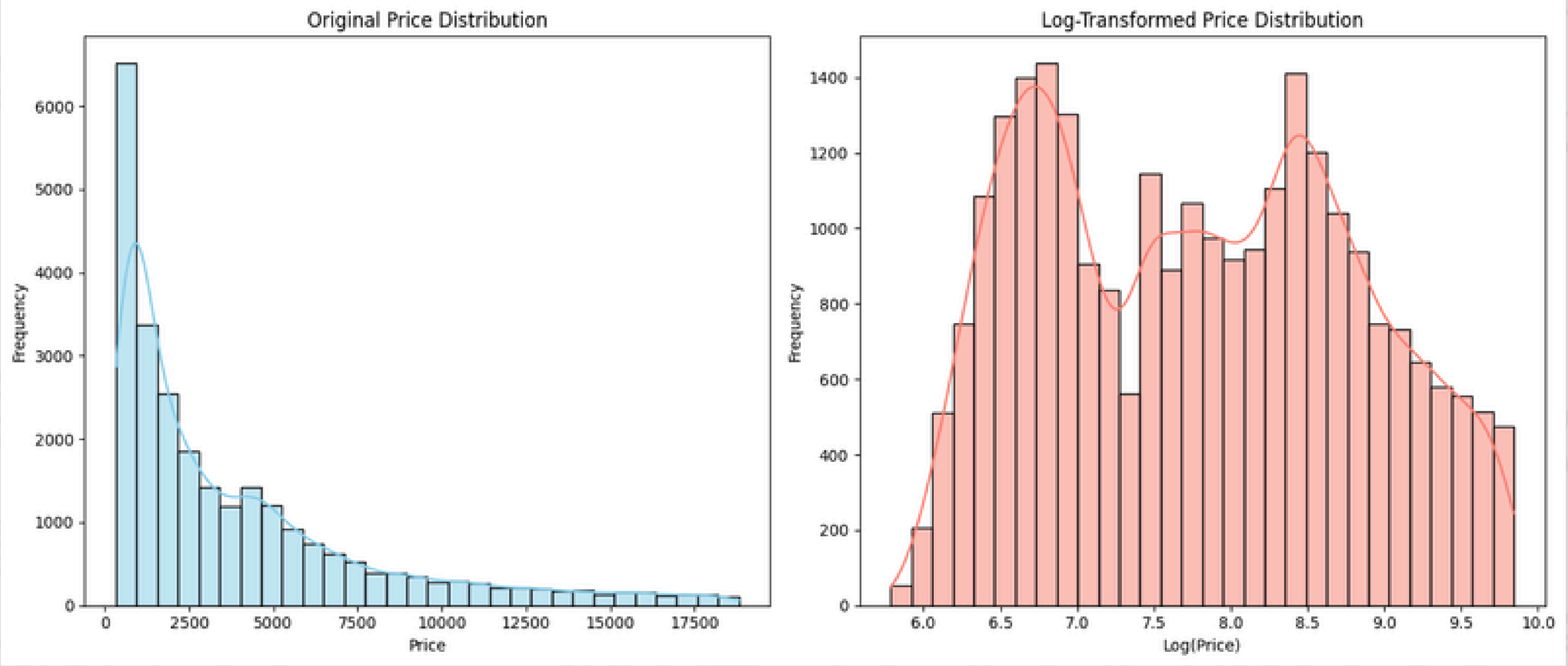
5246

DISTRIBUTION OF PRICE



Distribution of price is positively skewed

DISTRIBUTION OF PRICE AFTER LOG TRANSFORMATION



Choosing the Optimal Number of Clusters (k)

Elbow Method
Silhouette Analysis

Elbow Method (Left Plot)

- Plots Inertia (Within-cluster sum of squares) vs Number of clusters (k)
- As k increases, inertia decreases
- Look for the “elbow point” where the decrease slows down
- In this plot,
  The elbow appears at k = 3
- Beyond this point, adding more clusters gives diminishing improvement

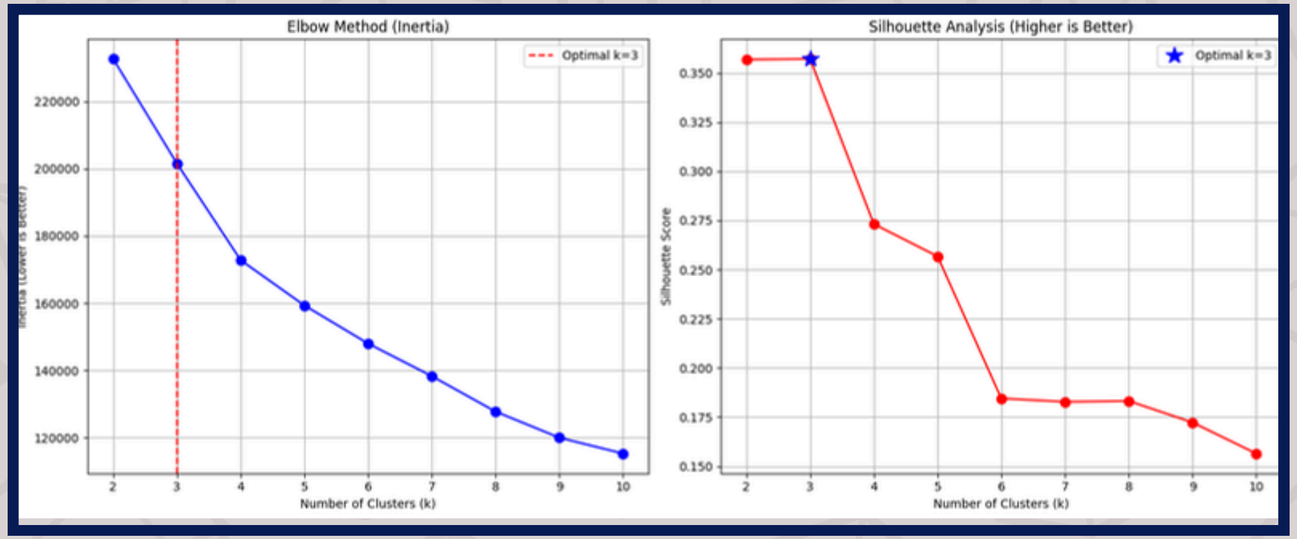
Silhouette Analysis (Right Plot)

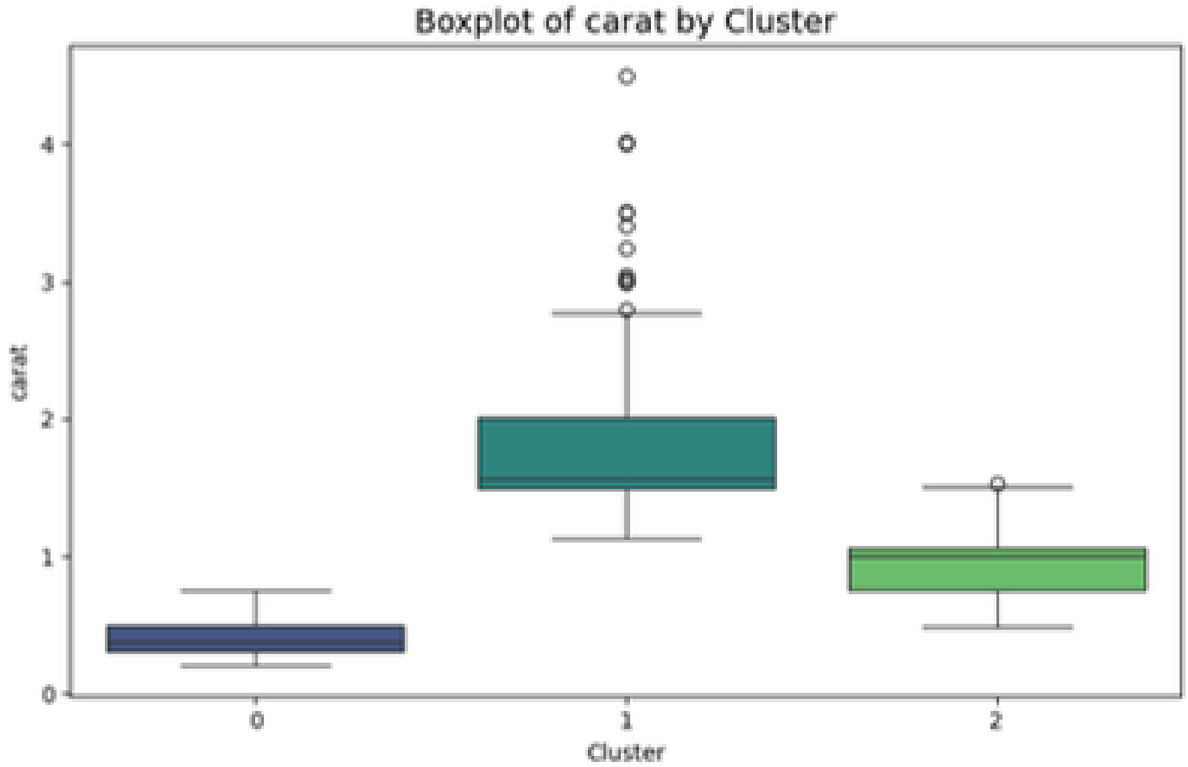
- Measures how well data points fit within their cluster
- Silhouette score ranges from –1 to +1
  Higher value → better clustering
- Choose the k with the highest silhouette score
- Maximum silhouette score occurs at k = 3

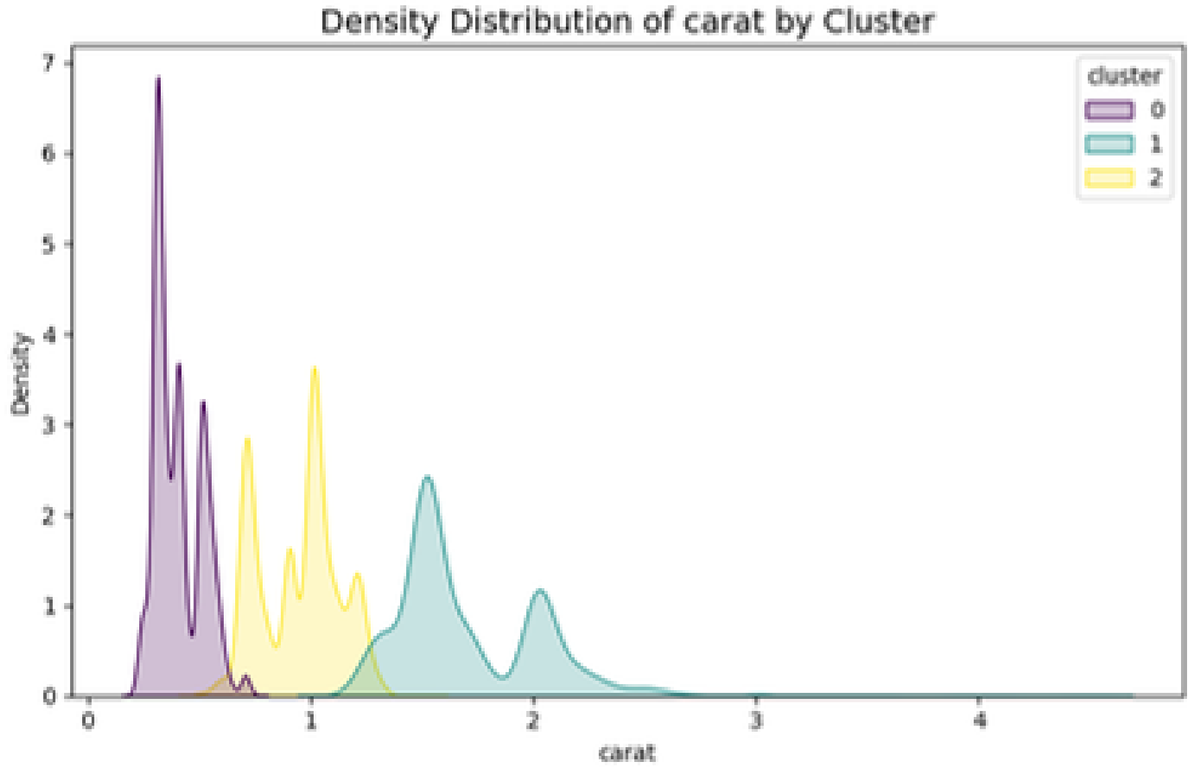
Final Decision

- Both methods agree on k = 3
- Therefore, 3 clusters is chosen as the optimal number

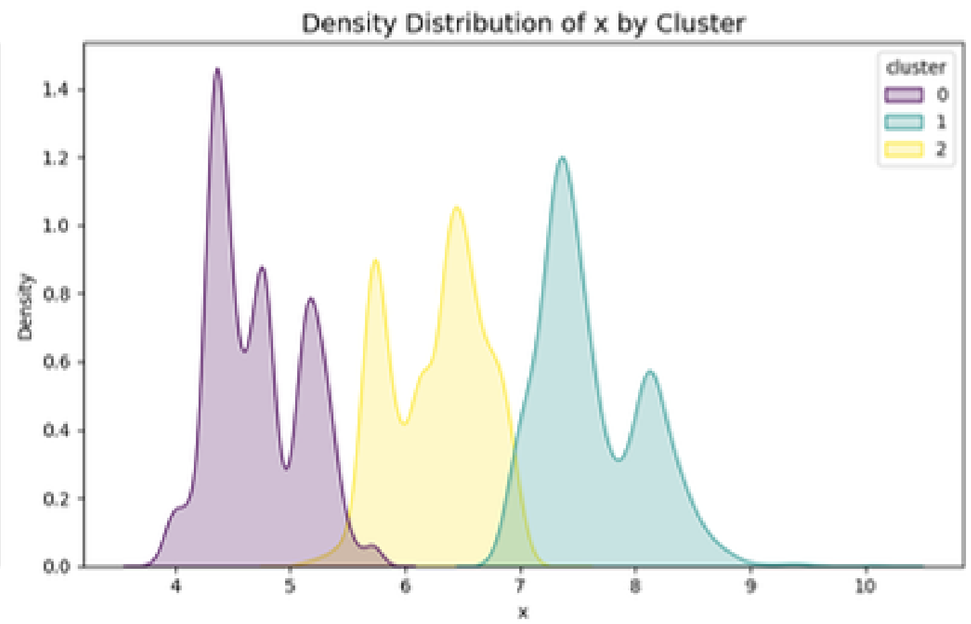
The optimal number of clusters was selected by combining the Elbow Method and Silhouette Analysis, both indicating k = 3

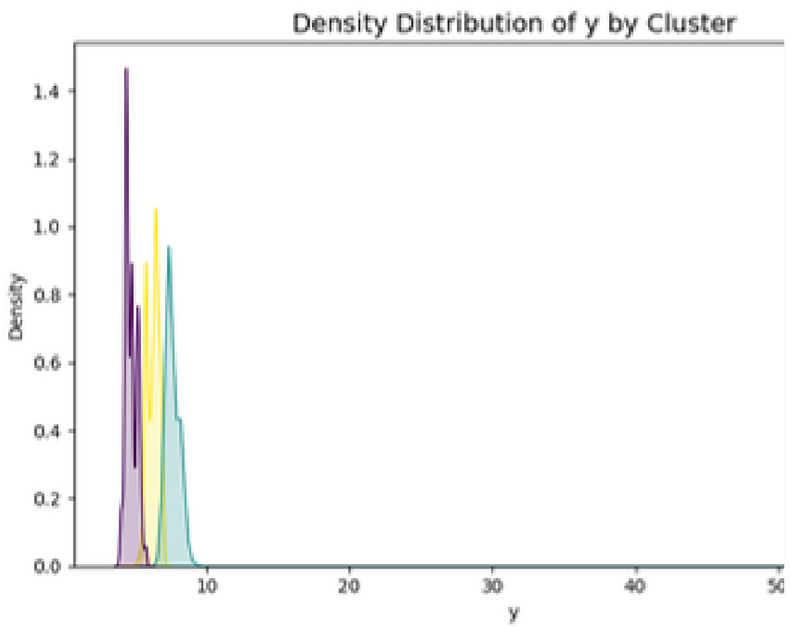


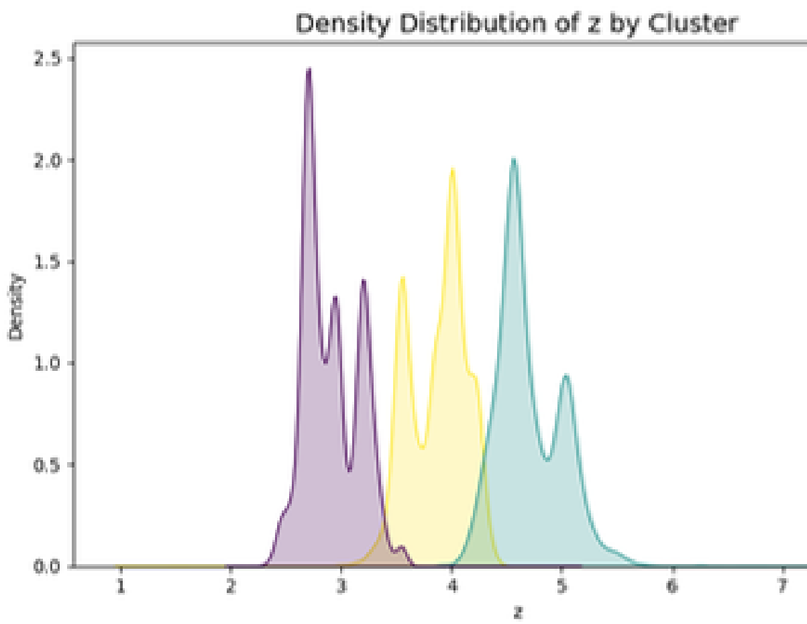




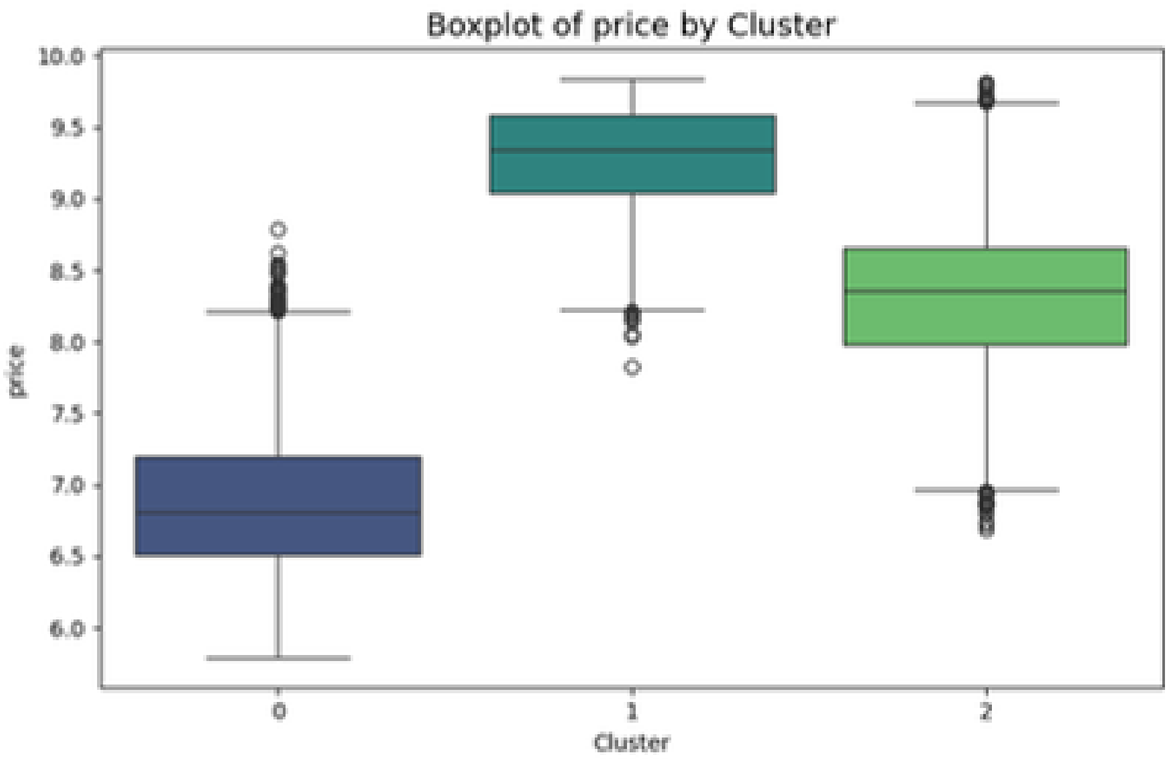
We can see that carat values are not equal between clusters

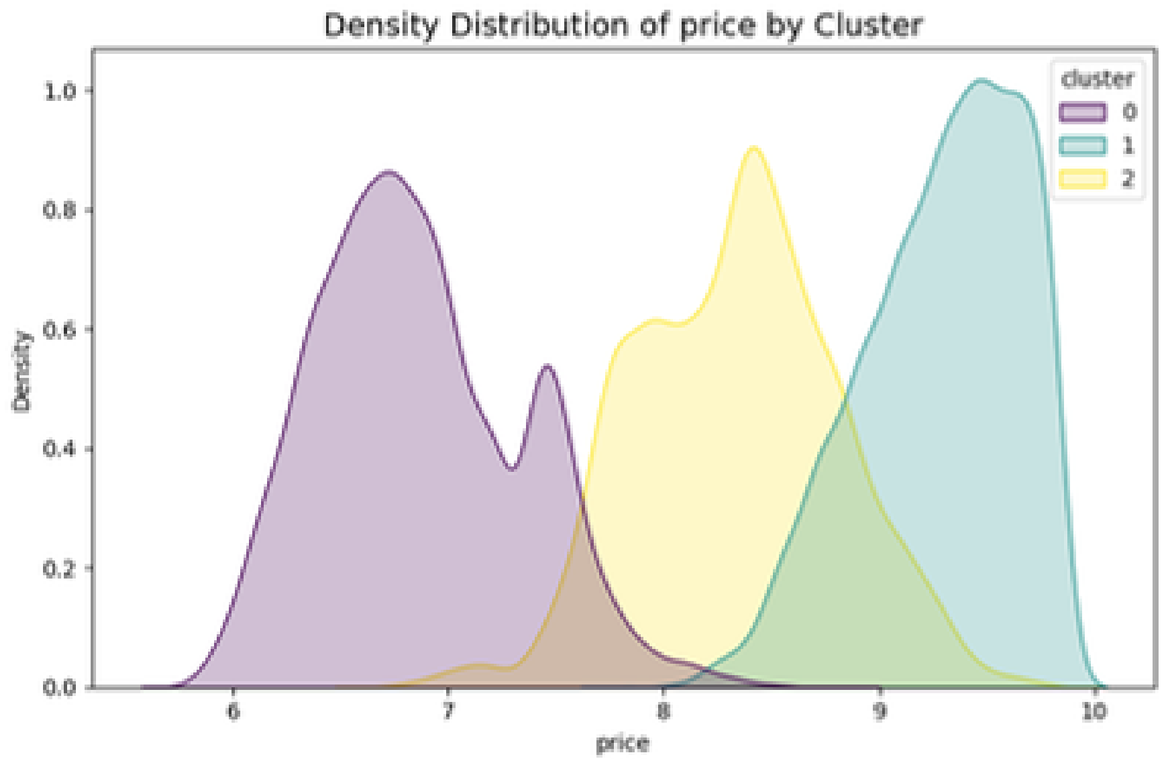






X, Y, Z variables are also different between clusters which indicates volume also plays a part in clusters





Our predictor variable price is different between each clusters

We can see that we obtain different range of values for each clusters

ADVANCED ANALYSIS

We did both!!

Cluster-wise Modeling

✔

Global Modeling

✔

- Multiple Linear
- Ridge, Lasso, Elastic Net
- Partial Least Square
- Random Forest
- XG Boost

07 Predictive Models

01. Multiple Linear Regression (MLR)

Why use

- Simple, interpretable baseline model
- Good for understanding relationships between variables.

When to use

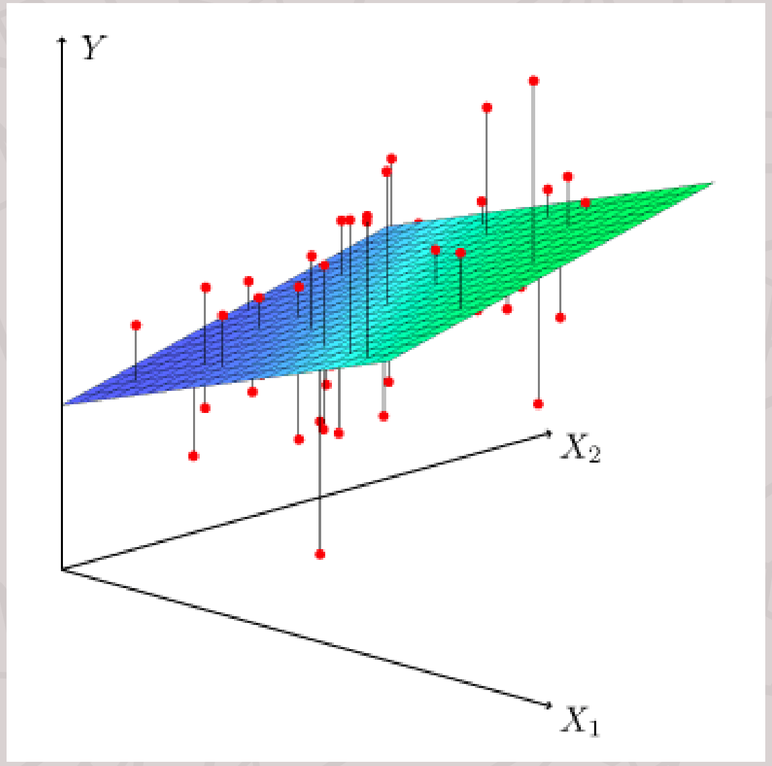
- Linear relationship between X and Y
- Low multicollinearity
- Number of predictors ≪ number of observations

Limitations / Disadvantages

- Sensitive to multicollinearity
- Overfitting when predictors are many
- Poor performance on nonlinear data
- Sensitive to outliers.

Tuning parameters

- None (closed-form solution)



02. Ridge Regression (L2 Regularization)

Why use

- Reduces overfitting
- Handles muticollinearity well

When to use

- Many correlated predictors
- All predictors are believed to be important

Limitations / Disadvantages

- Does not perform feature selection
- Less interpretable than MLR

Tuning parameters	
PARAMETER	**MEANING**
α (lambda)	Strength of penalty (higher → more shrinkage)

GRIDSEARCH CV

03. Lasso Regression (L1 Regularization)

Why use

- Performs automatic feature selection
- Produces sparse models

When to use

- High-dimensional data
- Only few predictors are truly important

Limitations / Disadvantages

- Unstable when predictors are highly correlated
- Can arbitrarily drop important variables

Tuning parameters

PARAMETER	MEANING
α (lambda)	Controls sparsity (higher → more coefficients = 0)

GRIDSEARCH CV

04. Elastic Net (L1 + L2)

Why use

- Combines strengths of Lasso + Ridge
- Stable feature selection with correlated predictors

When to use

- High dimensional data
- Strong multicollinearity
- Need feature selection + stability

Limitations / Disadvantages

- More complex to tune
- Less interpretable than Lasso alone

GRIDSEARCH CV

Tuning parameters	
PARAMETER	**MEANING**
α	Overall penalty strength
L1_ratio	*Balance between Lasso (1) and Ridge (0)*

05. Partial Least Squares Regression (PLS)

Why use

- Handles multicollinearity + dimensionality reduction
- Supervised alternative to PCA regression

When to use

- Predictors >> observations
- Strong correlation among predictors

Limitations / Disadvantages

- Components are hard to interpret
- Linear model (cannot capture nonlinearities)

TUNING PAPAMETERS	
PARAMETER	**MEANING**
n_components	Number of latent components

06. Random Forest Regression

Why use

- Strong nonlinear modeling
- Robust to outliers and noise
- Handles interactions automatically

When to use

- Complex nonlinear relationships
- Medium-large datasets
- Feature importance is needed

Limitations / Disadvantages

- Computationally expensive
- Less interpretable
- Slow training on large data

TUNING PARAMETERS	
PARAMETER	**MEANING**
n_estimators	Number of trees
max_depth	Tree depth
min_samples_split	Minimal samples to split
max_features	Features considered at each split

GRIDSEARCH CV

07. XGBoost Regression

Why use

- State-of-the-art performance
- Handles missing values
- Strong regularization

When to use

- High predictive accuracy is required
- Structured/tabular data
- Competitions & production models

Limitations / Disadvantages

- Sensitive to hyperparameters
- Risk of overfitting if not tuned
- Computationally intensive

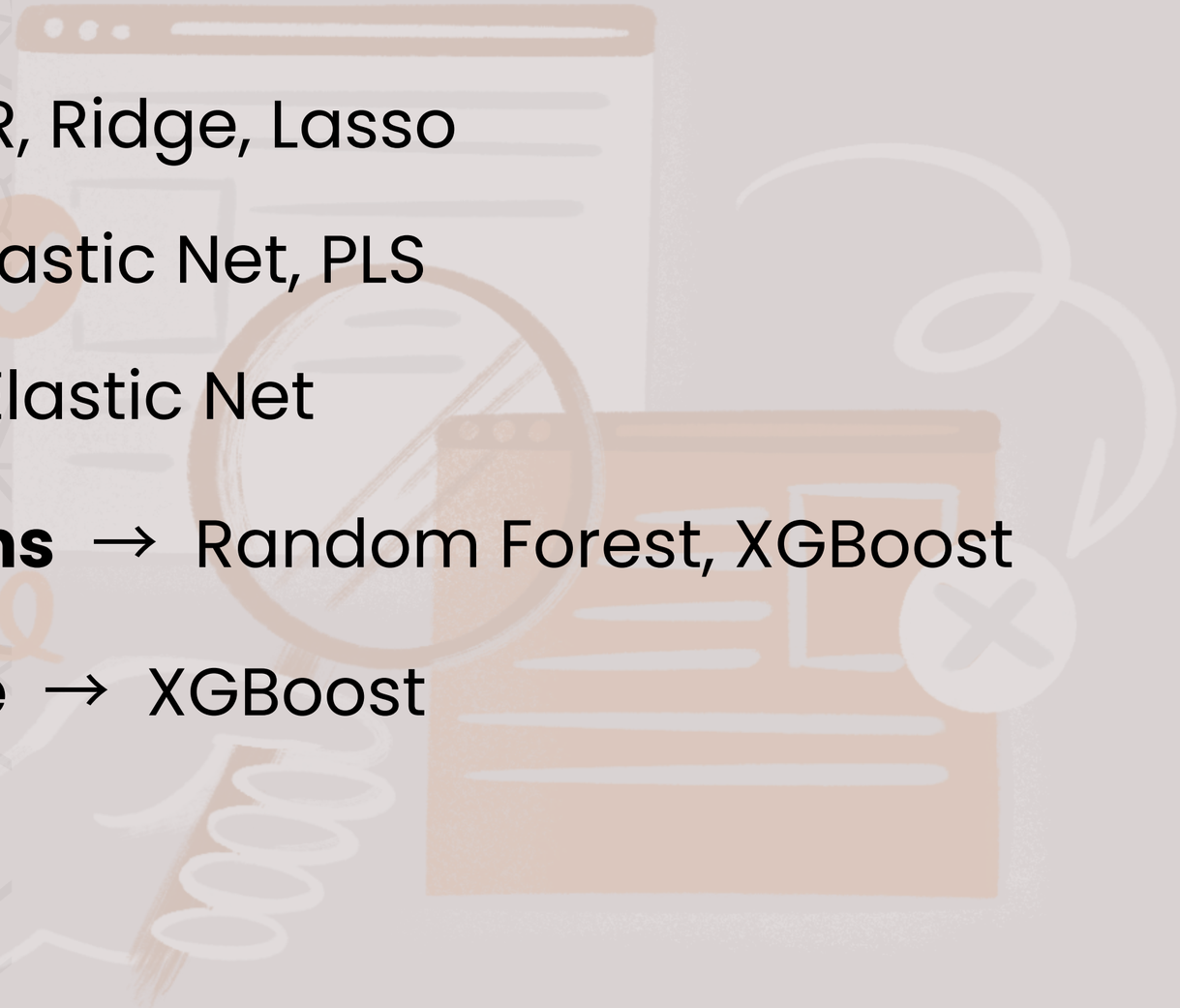
GRIDSEARCH CV

TUNING PARAMETERS	
PARAMETER	**MEANING**
n_estimators	Number of boosting rounds
learning_rate	Step size shrinkage
max_depth	Tree depth
subsample	Row sampling
colsample_bytree	Feature sampling
reg_alpha	L1 regularization
reg_lambda	L2 regularization

MODEL	WHY USE	WHEN TO USE	LIMITATIONS/ DISADVANTAGES	KEY TUNING PARAMETERS	WHAT THE PARAMETERS DO
Multiple Linear Regression (MLR)	Simple, interpretable baseline	Linear relationship, low multicollinearity, small–medium p	Sensitive to outliers & multicollinearity, overfits when p is large, cannot model nonlinearity	None	No regularization or tuning
Ridge Regression	Reduces overfitting, handles multicollinearity	Many correlated predictors, all variables relevant	No feature selection, less interpretable	α (lambda)	Controls strength of L2 penalty (shrinks coefficients)
Lasso Regression	Feature selection + regularization	High dimensional data, sparse true model	Unstable with correlated variables, may drop important predictors	α (lambda)	Controls L1 penalty, forces coefficients to zero
Elastic Net	Combines Ridge + Lasso benefits	High multicollinearity +feature selection needed	More complex tuning, less transparent	α, l1_ratio	α: penalty strength, l1_ratio: L1 vs L2 balance

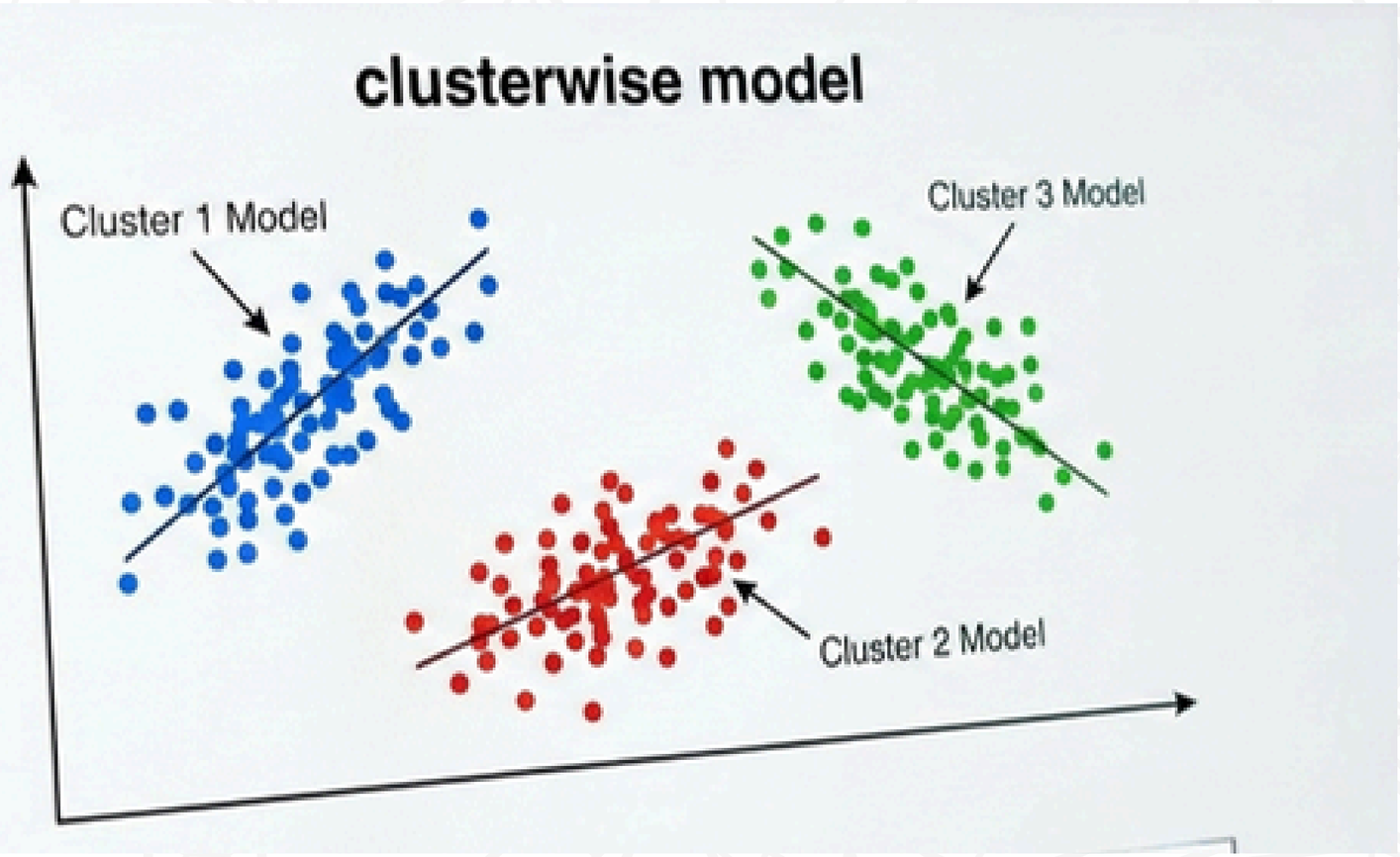
MODEL	WHY USE	WHEN TO USE	LIMITATIONS/ DISADVANTAGES	KEY TUNING PARAMETERS	WHAT THE PARAMETERS DO
Partial Least Squares (PLS)	Handles multicollinearity & p > n	Strongly correlated predictors, small samples	Components hard to interpret, linear only	n_components	Number of latent components used
Random Forest Regression	Strong nonlinear modeling, robust	Nonlinear relationships, interactions, noisy data	Slow training, low interpretability	n_estimators, max_depth, max_features, min_samples_split	Controls number, depth & randomness of trees
XGBoost Regression	High accuracy, strong regularization	Large datasets, performance-critical tasks	Sensitive to tuning, risk of overfitting	n_estimators, learning_rate, max_depth, subsample, colsample_bytree, reg_alpha, reg_lambda	Controls boosting speed, complexity, and regularization

QUICK INTERPRETATION GUIDE



- **Interpretability focus** → MLR, Ridge, Lasso
- **Multicollinearity** → Ridge, Elastic Net, PLS
- **Feature selection** → Lasso, Elastic Net
- **Nonlinear & complex patterns** → Random Forest, XGBoost
- **Best predictive performance** → XGBoost

Cluster-wise Modelling



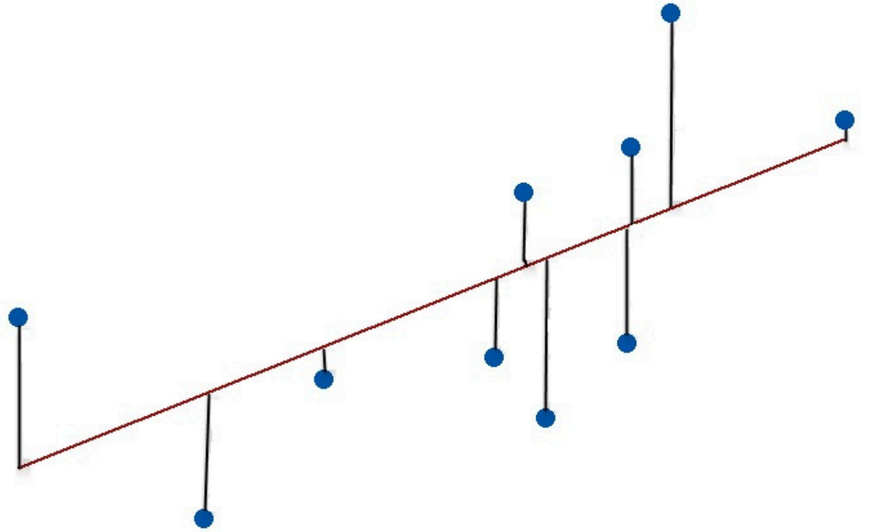
Separate Cluster Comparison

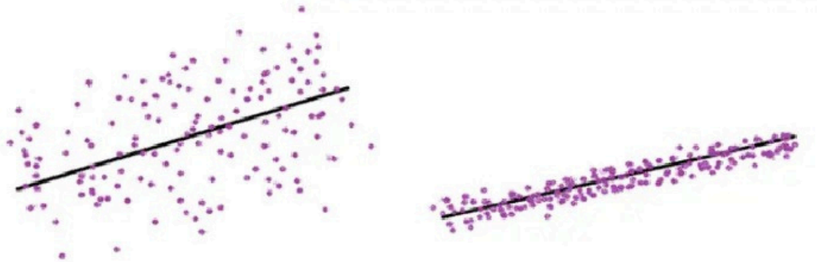
What Is RMSE?

RMSE is the square root of the average of squared differences between observed and predicted values. It's a widely used regression metric that tells us how much error to expect from our predictions, on average.

The mathematical formula for calculating RMSE is:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$





R-squared : 17%

R-squared : 83%

What Is R-Squared?

R-squared, denoted as R^2, is a statistical measure of goodness of fit in regression models. It tells us how much of the variation in the dependent variable can be explained by the model.

The value of R-squared lies between 0 and 1:

- $R^2 = 0$ means the model explains none of the variability;
- $R^2 = 1$ means the model explains all of it.

THE WINNER

FOR CLUSTER 0

XG BOOST

Cluster 0	TRAIN		TEST	
Model	R^2	RMSE	R^2	RMSE
MULTIPLE LINEAR	0.9349	0.1205	0.9365	0.1215
LASSO	0.9192	0.1342	0.9193	0.1371
RIDGE	0.9345	0.1208	0.9359	0.1222
ELASTIC NET	0.9206	0.1331	0.9220	0.1347
PARTIAL LEAST SQUARE	0.9082	0.1431	0.9107	0.1442
RANDOM FOREST	0.9915	0.0436	0.9634	0.0922
XG BOOST	0.9861	0.0556	0.9694	0.0844

WINNER

FOR CLUSTER 1

XG BOOST

Cluster 1	TRAIN		TEST	
Model	R^2	RMSE	R^2	RMSE
MULTIPLE LINEAR	0.8300	0.1530	0.8156	0.1530
LASSO	0.7579	0.1826	0.7386	0.1822
RIDGE	0.8273	0.1543	0.8114	0.1548
ELASTIC NET	0.7812	0.1736	0.7635	0.1733
PARTIAL LEAST SQUARE	0.7156	0.1979	0.6918	0.1978
RANDOM FOREST	0.9780	0.0551	0.8948	0.1156
XG BOOST	0.9619	0.0725	0.9053	0.1097

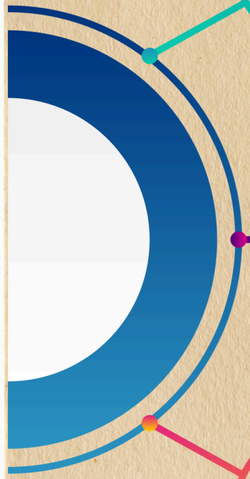
FOR CLUSTER 2

XG BOOST

Cluster 2	TRAIN		TEST	
Model	R^2	RMSE	R^2	RMSE
MULTIPLE LINEAR	0.9309	0.1237	0.9334	0.1225
LASSO	0.9158	0.1366	0.9159	0.1377
RIDGE	0.9300	0.1245	0.9328	0.1230
ELASTIC NET	0.9170	0.1356	0.9179	0.1360
PARTIAL LEAST SQUARE	0.9096	0.1415	0.9098	0.1426
RANDOM FOREST	0.9885	0.0504	0.9598	0.0952
XG BOOST	0.9751	0.0743	0.9661	0.0874

Cluster-wise Modelling

The Winner is



C 0 XG Boost

C 1 XG Boost

C 2 XG Boost

Global Modelling

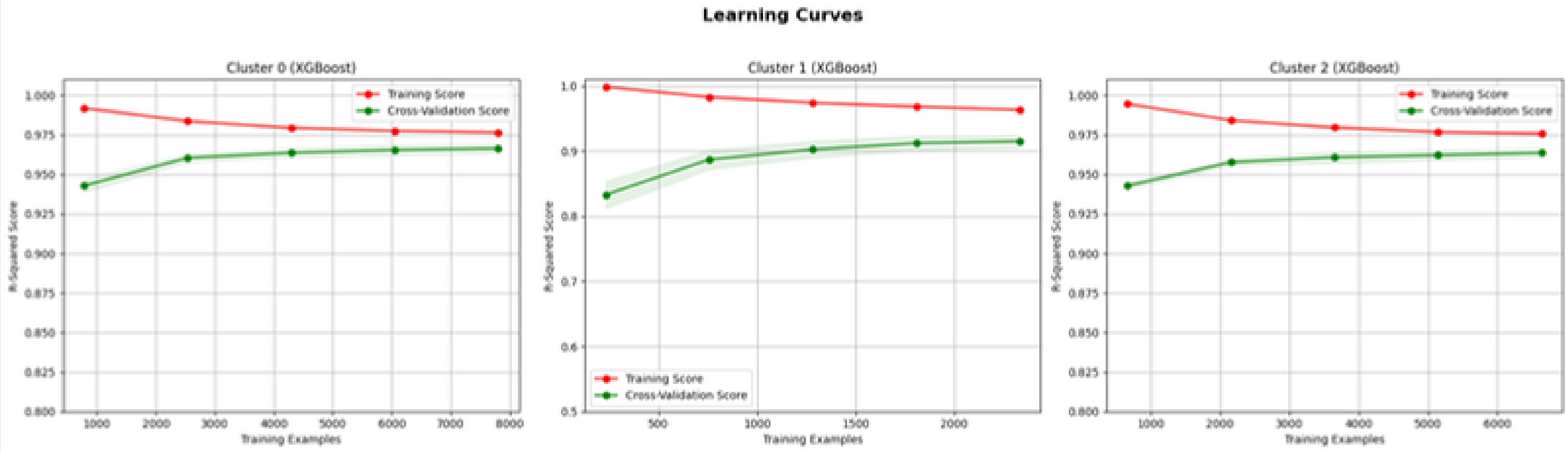
Global Modelling

Winner

XG Boost!

Model	TRAIN		TEST	
	R^2	RMSE	R^2	RMSE
MULTIPLE LINEAR	0.9813	0.1395	0.9810	0.1399
LASSO	0.9810	0.1405	0.9807	0.1409
RIDGE	0.9813	0.1396	0.9810	0.1400
ELASTIC NET	0.9810	0.1405	0.9807	0.1409
PARTIAL LEAST SQUARE	0.9811	0.1401	0.9809	0.1402
RANDOM FOREST	0.9988	0.0357	0.9910	0.0964
XG BOOST	0.9958	0.0657	0.9924	0.0882

Learning Curves

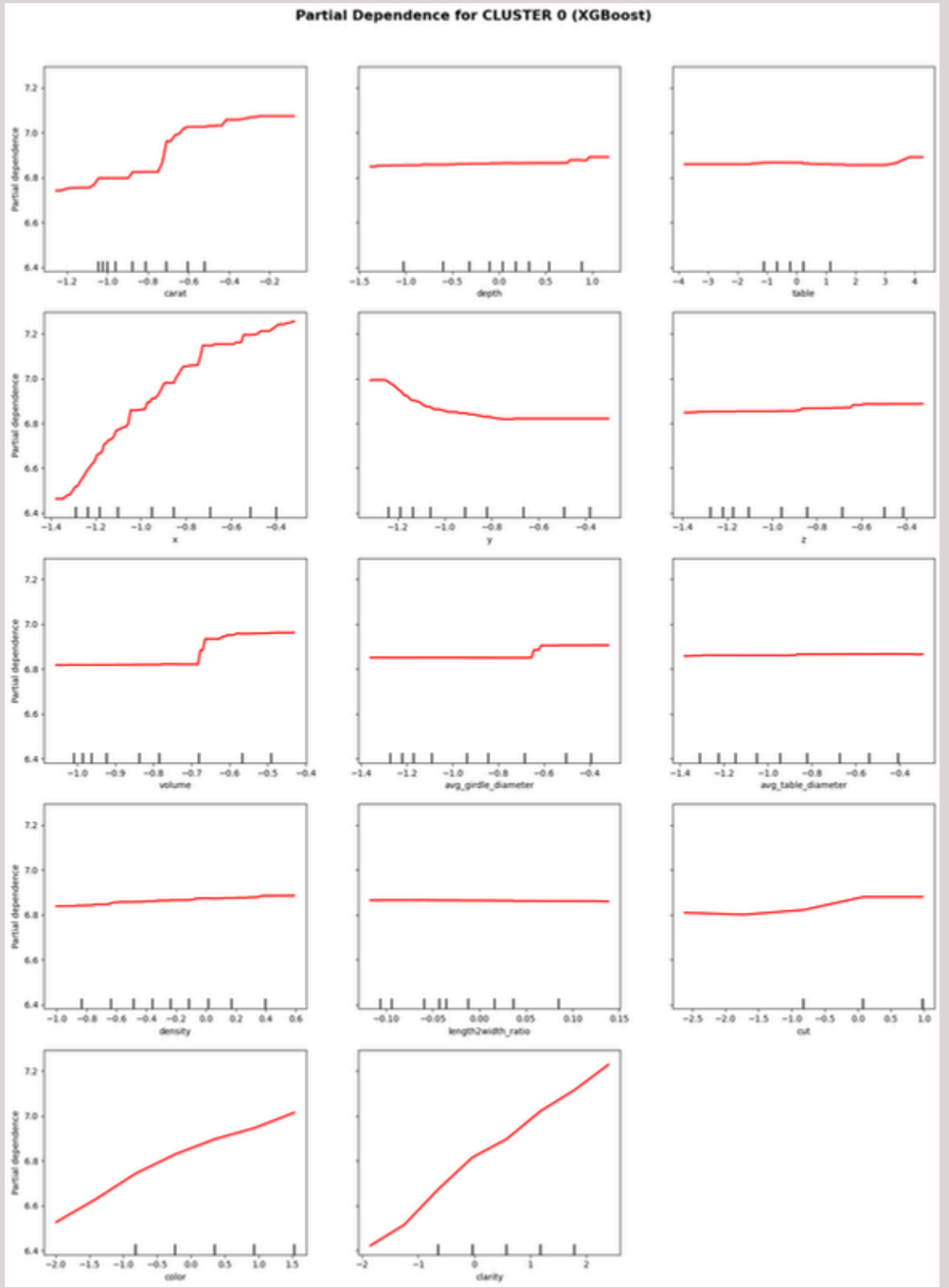


In summary, these curves give us the green light to proceed. The models are healthy, stable, and ready to explain the feature importance, which I will show on the next slide

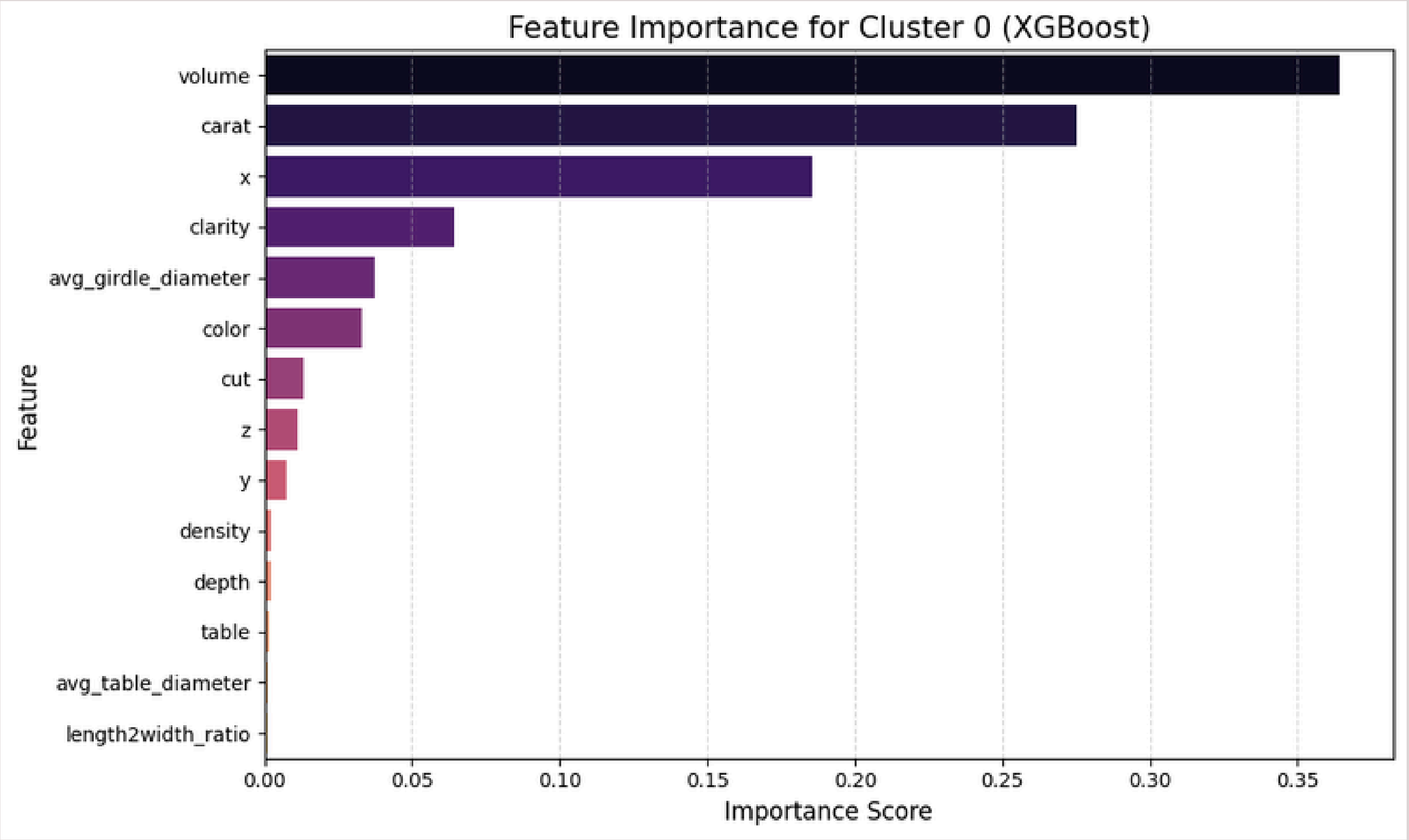
Cluster 0

- Size attributes (**carat, volume, x-dimension) are primary price drivers**
- Quality attributes show non-linear effects
- Biggest gains from improving clarity and color
- Proportion metrics (depth, table, ratios) have limited impact
- Several engineered features contribute minimal marginal value
- Model behavior is consistent with domain knowledge (validates clustering)

price predictions are driven mainly by size and clarity, while fine geometric proportions play a secondary role.

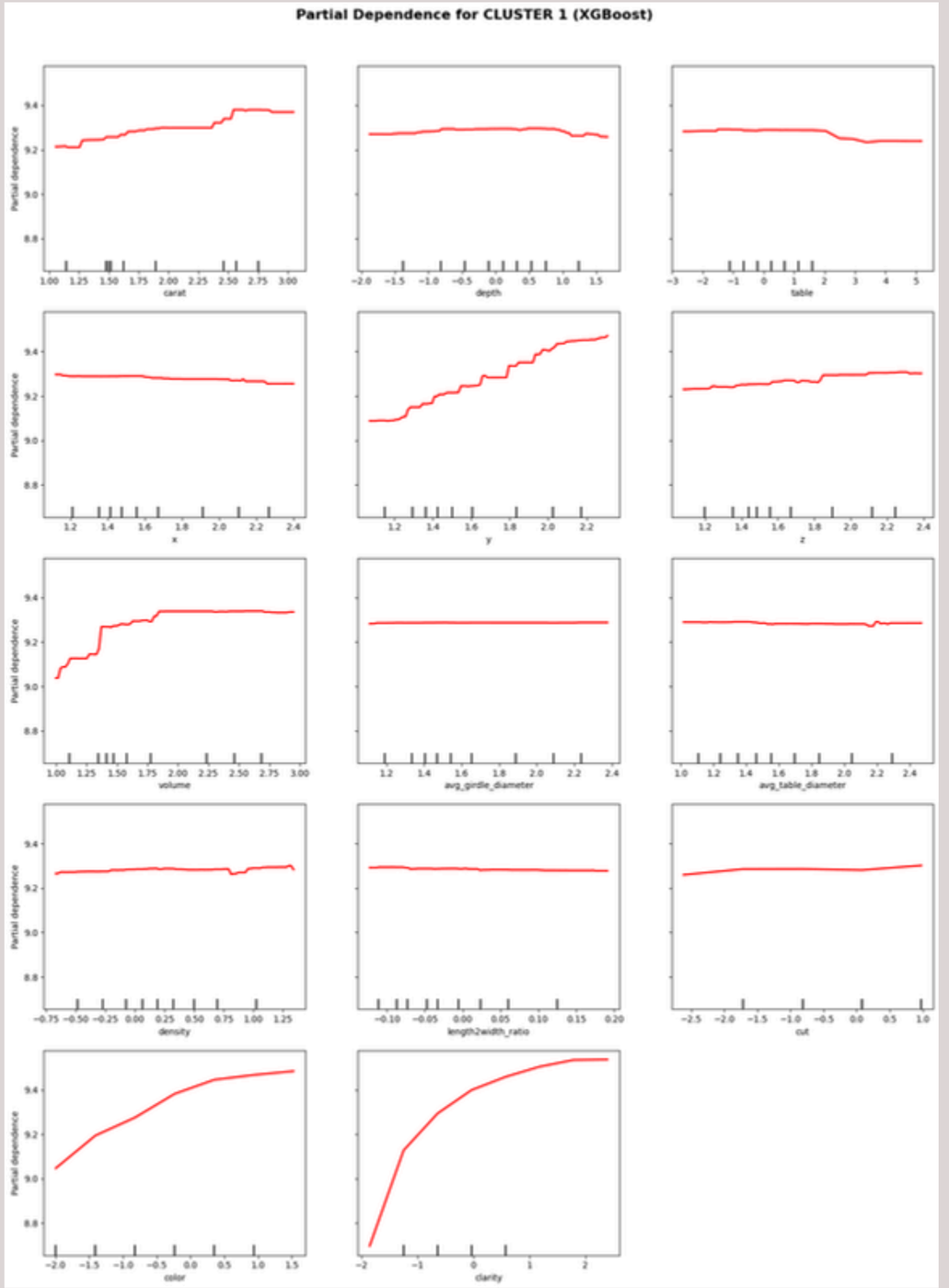


Cluster 0

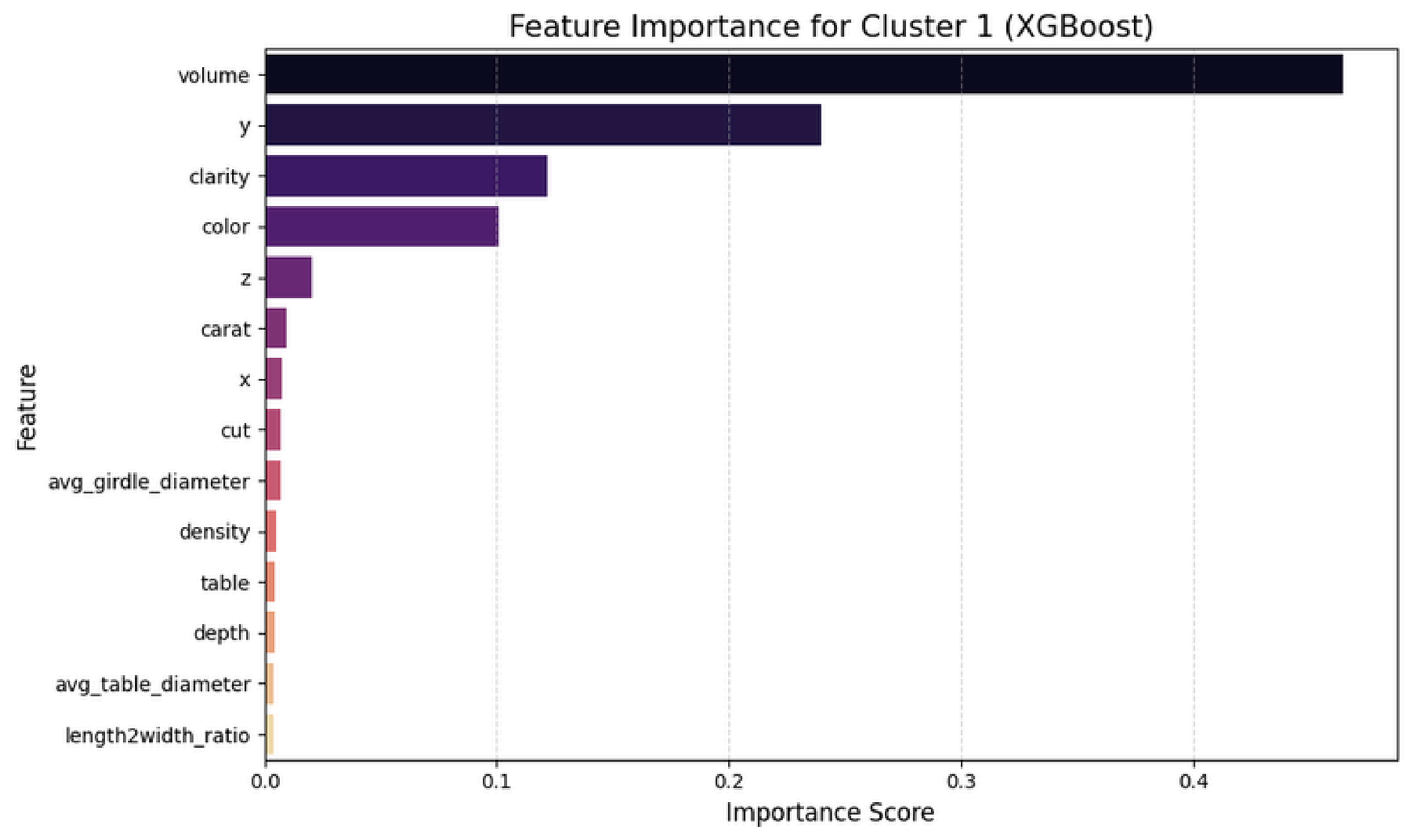

Feature Importance for Cluster 0 (XGBoost)
volume
carat
x
clarity
avg_girdle_diameter
color
cut
z
y
density
depth
table
avg_table_diameter
length2width_ratio
Feature
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
Importance Score

Cluster 1

- **Primary Drivers: Clarity and Color** exhibit the steepest slopes, acting as the most significant predictors of price.
- **Volume ,Y-dimension ,Carat displays "step-like" jumps** at specific market weight thresholds.
- The model shows **high sensitivity to clarity at lower grades**, but price gains plateau as quality reaches the highest levels.
- Physical metrics like **depth, table, ratios and cut have nearly flat profiles,** contributing negligible marginal value to the prediction.
- Secondary Cut Influence: **The 'Cut' grade shows a relatively flat response,** indicating it is a secondary factor compared to raw size and internal quality.



Cluster 1

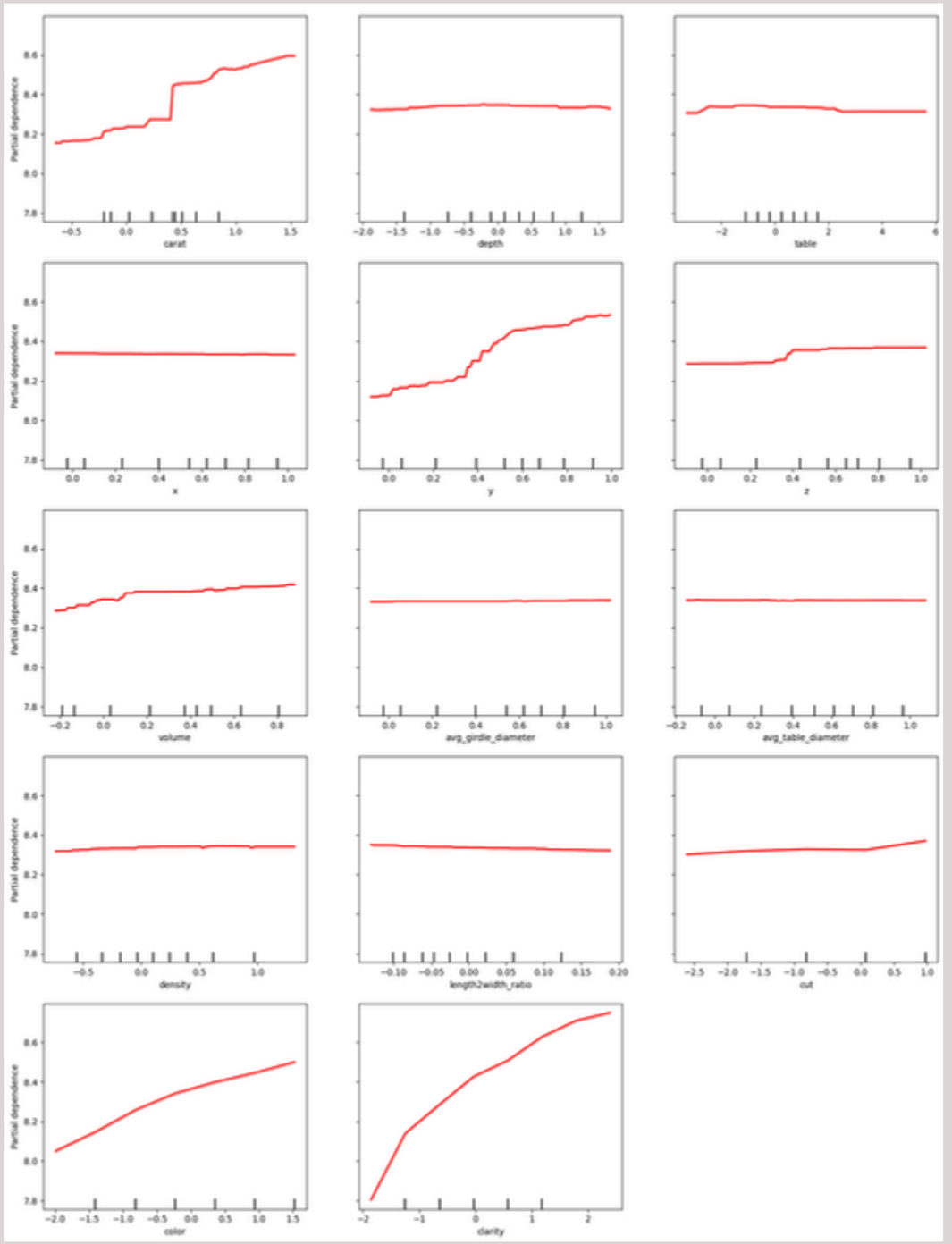

Feature Importance for Cluster 1 (XGBoost)
volume
y
clarity
color
z
carat
x
cut
avg_girdle_diameter
density
table
depth
avg_table_diameter
length2width_ratio
Feature
0.0
0.1
0.2
0.3
0.4
Importance Score

Cluster 2

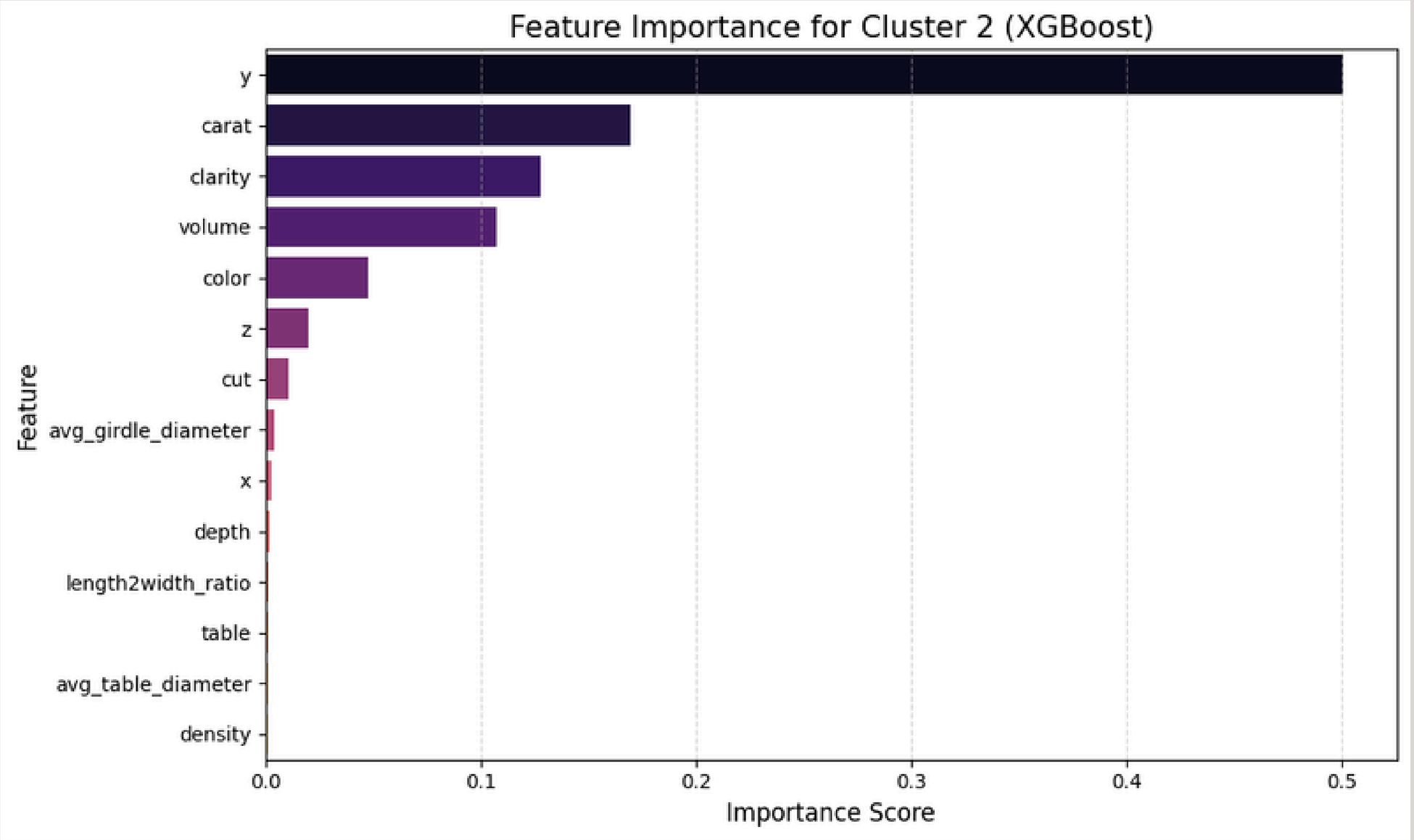
- **Clarity** and **color** show the steepest slopes, representing the **strongest impact on price.**

- Price increases significantly with **carat and y-dimension,** featuring a notable **"step-up"** price threshold for carat weight.

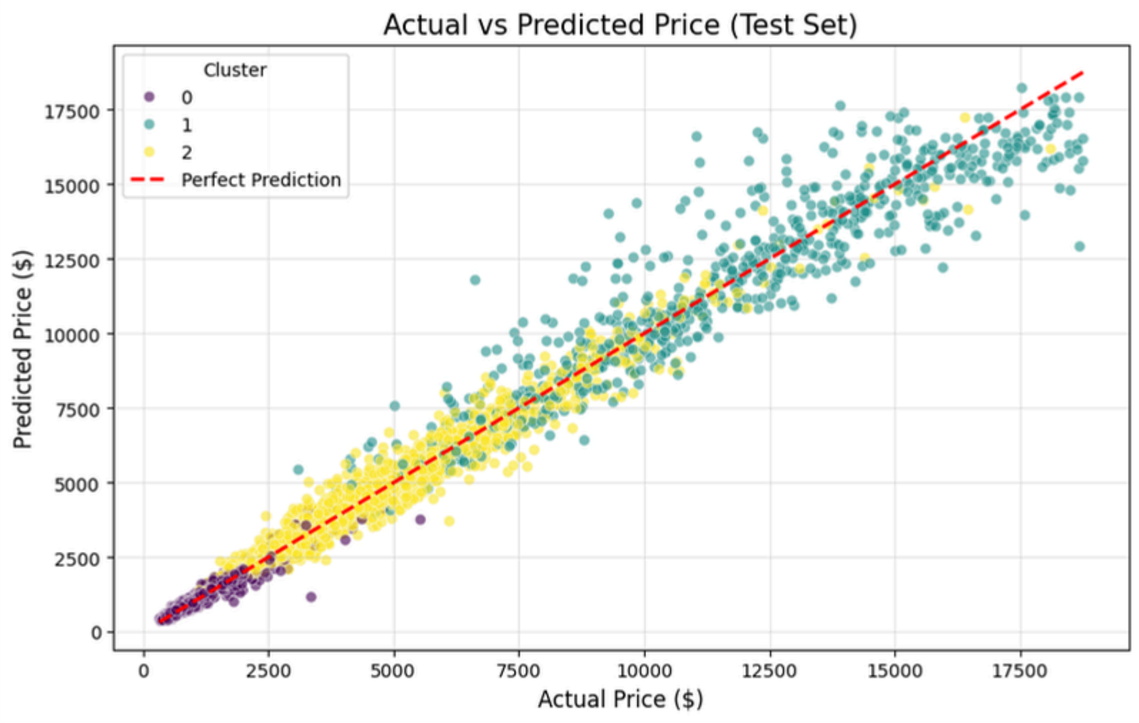
- **Price sensitivity is highest at lower clarity grades**, with gains tapering off as quality reaches the top tier.
- **Minor Geometric Impact:** Proportions like **depth, table, and length-to-width ratios** are nearly flat, offering minimal predictive value.
- **Cut grade and x-dimension show flat responses,** indicating they are secondary to weight and width for this group.



Cluster 2


Feature Importance for Cluster 2 (XGBoost)
y
carat
clarity
volume
color
z
cut
avg_girdle_diameter
x
depth
length2width_ratio
table
avg_table_diameter
density
Feature
0.0
0.1
0.2
0.3
0.4
0.5
Importance Score

Conclusion



model has predicted the price closely

The XGBoost model successfully captures the complex, non-linear relationship between gem attributes and price, confirming that size and internal quality are the dominant valuation factors across all clusters.

Thank You!

SOME PEOPLE LOSE DIAMONDS IN SEARCH OF STONES

Presented by Group 2

ST 3082 - Statistical Learning 1

s16644 - Imesh Chavindu

s16690 - Gajanan Umasuthan

s16835 - Thabeetha Shenali