جامعة
الملك سعود
King Saud University

HOUSING PRICES PREDICTION

USING DATA MINING

OBJECTIVES

1. **PROBLEM**

2. **DATA**
   data overview , and data graphs

3. **DATA PREPROCESSING**
   data cleaning , and data transformation

4. **DATA MINING TASK**
   data mining Techniques and findings

PROBLEM:

Since housing is one of the most significant investments in a person's life, using data mining techniques can help predict housing prices and make informed decisions.

In our project, we analyzed a housing dataset, which provided valuable insights into the factors influencing housing prices. This data enabled us to identify key patterns and challenges, such as multicollinearity, and apply effective techniques to improve prediction accuracy.

OUR DATA

- **We applied our data mining tasks on data set consisting of :**

- **Number of tupels : 545**
- **13 attibutes which are:**

Price

Area

bedrooms

bathrooms

stories

mainroad

guestroom

basement

parking

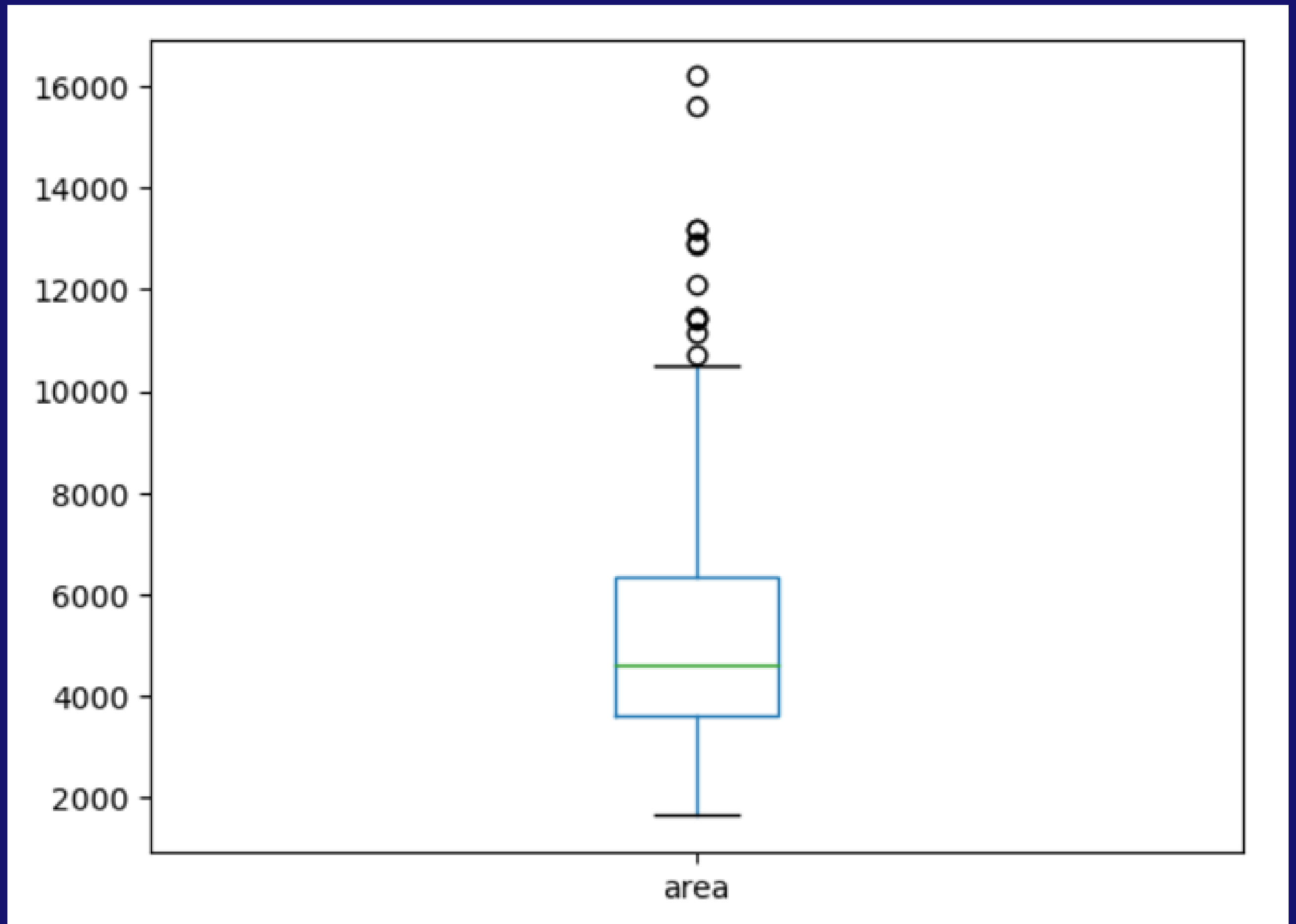
prefarea

Hotwater heating

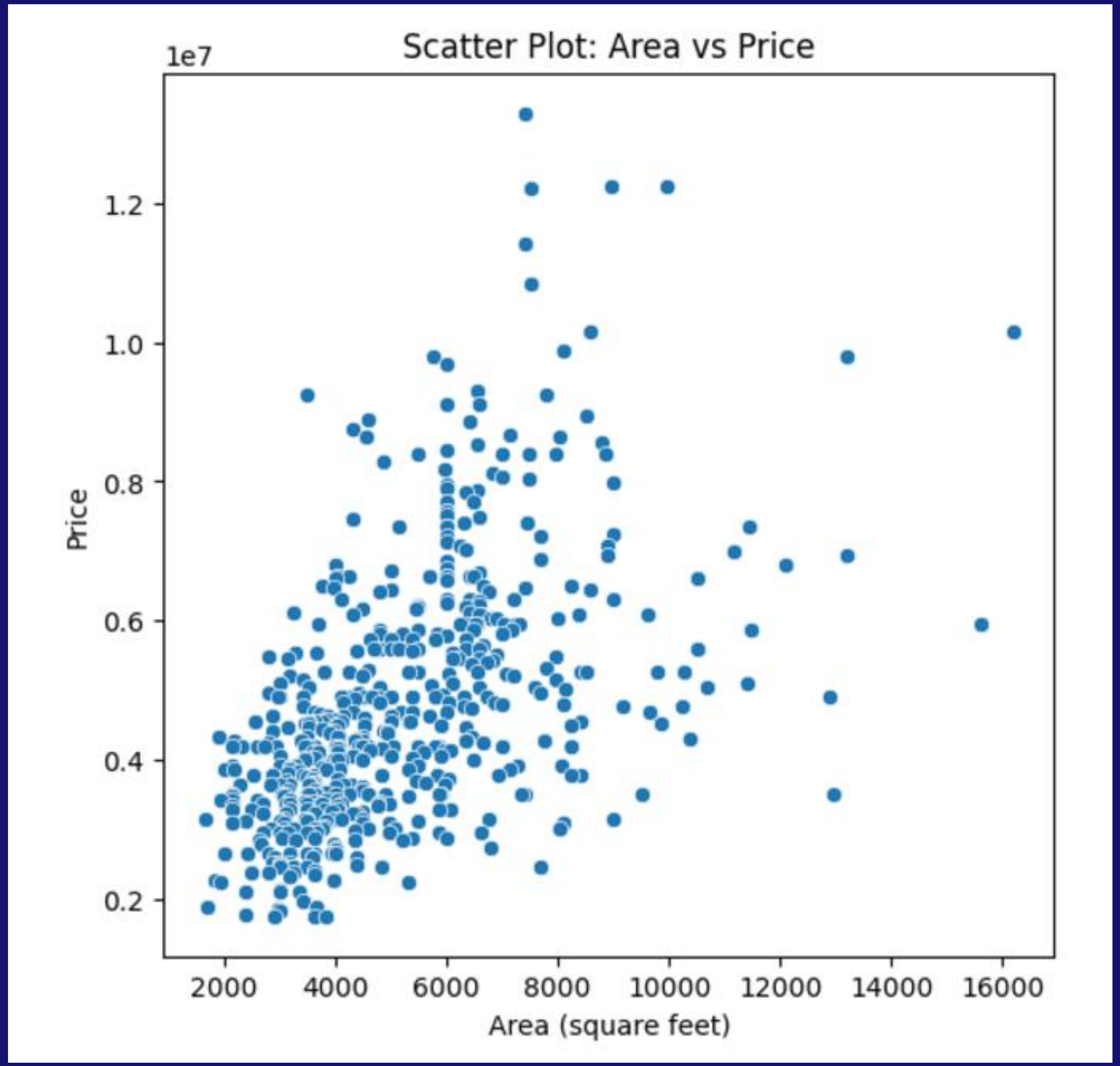
Air conditioning

Furnishing status

DATA GRAPHS:

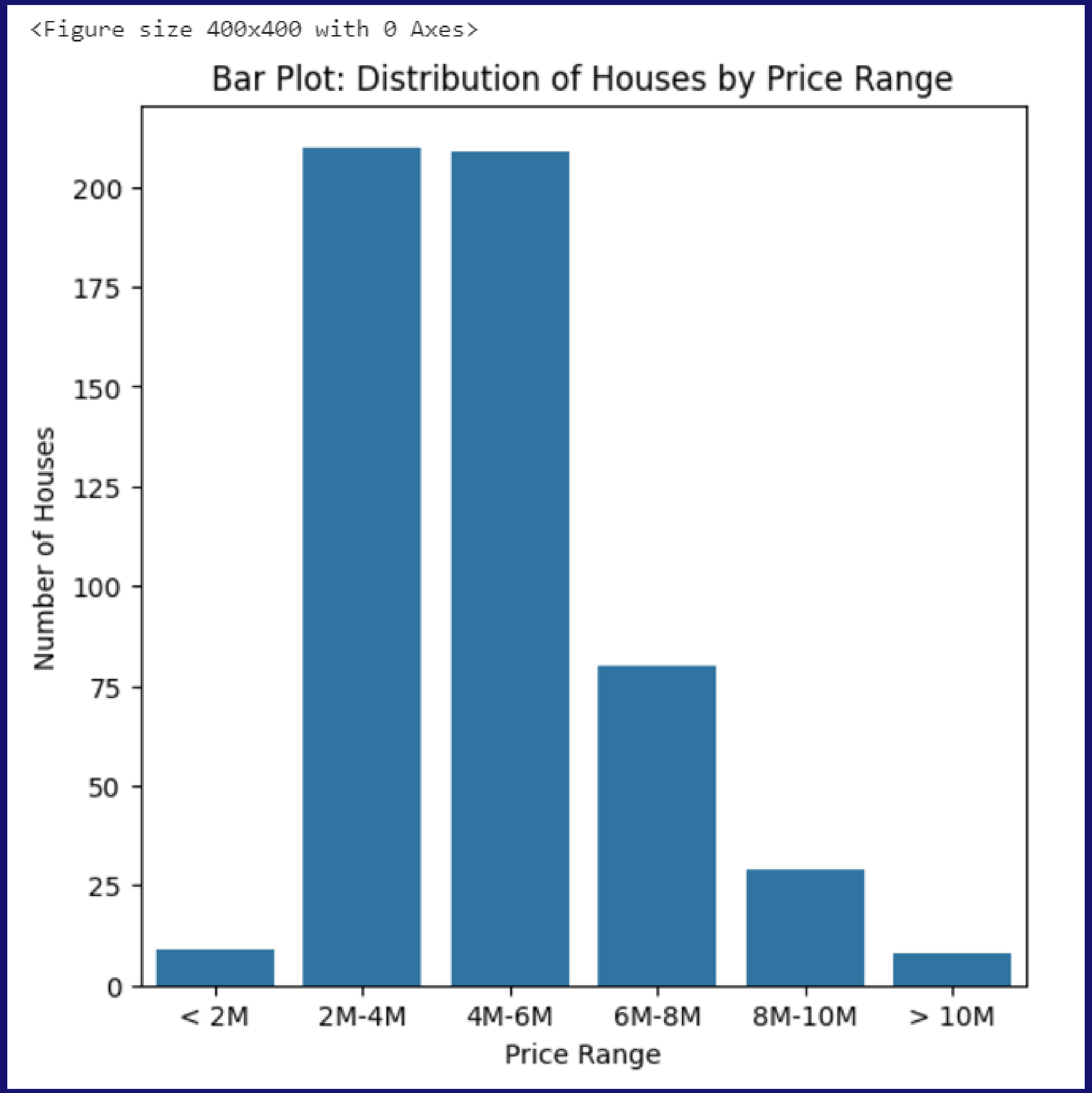


Box plot



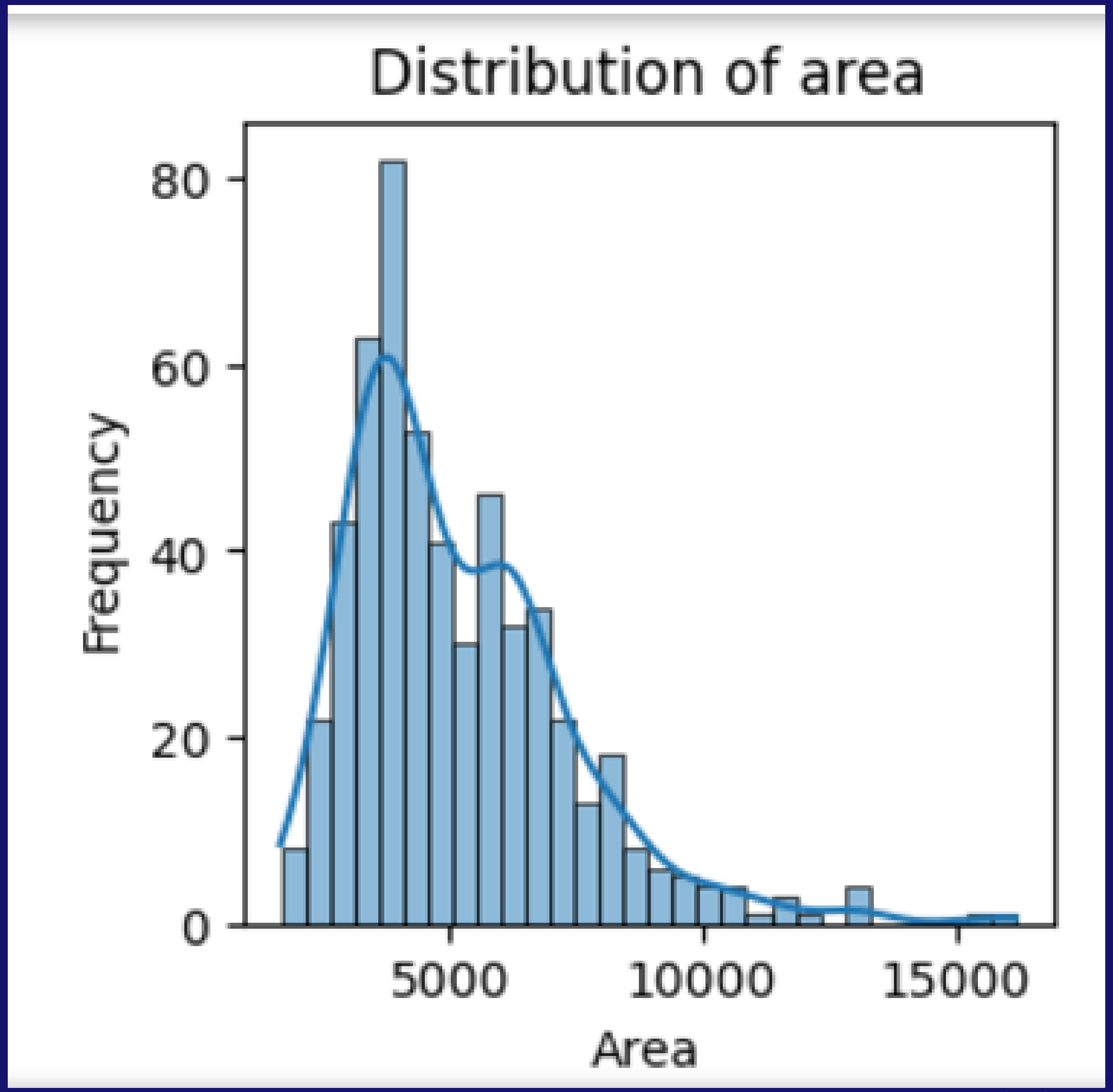
Scatter plot

DATA GRAPHS:



```
<Figure size 400x400 with 0 Axes>
```

Bar plot



Histogram

DATA PREPROCESSING:

- To achieve optimal accuracy, various preprocessing techniques were applied to enhance data efficiency, including:

1

DATA CLEANINIG

2

DATA TRANSFORMATION

Data cleaning:

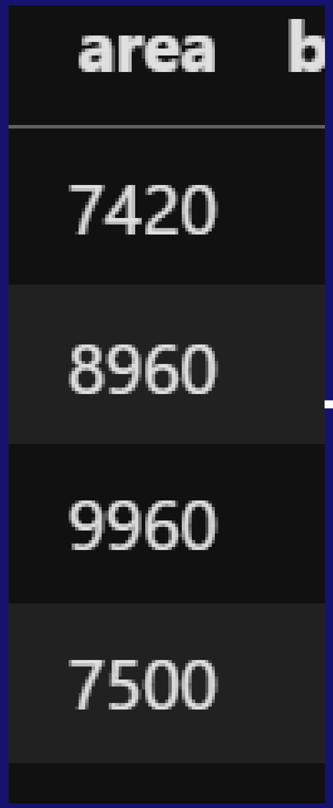
- We checked our data for missing or null values and found none.
- We identified outliers in the numeric attributes and removed the rows containing these outliers.

Min-Max Normalization:

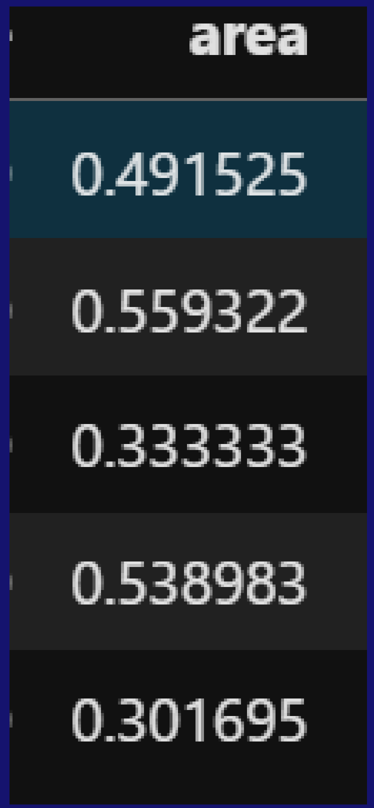
- The area attribute was scaled to a range between 0 and 1 to standardize its influence in modeling.

- This ensures all features have equal weight during analysis and prevents any feature from dominating due to its scale.

Min-Max Normalization:



area	b
7420	
8960	
9960	
7500	



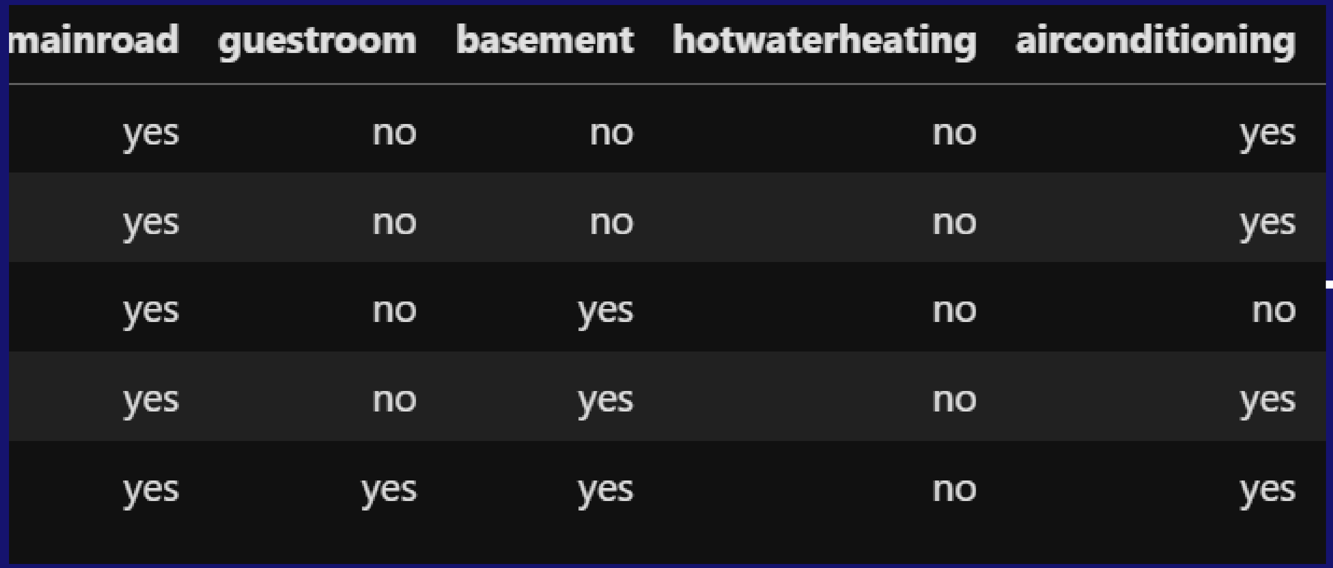
area
0.491525
0.559322
0.333333
0.538983
0.301695

Encoding:

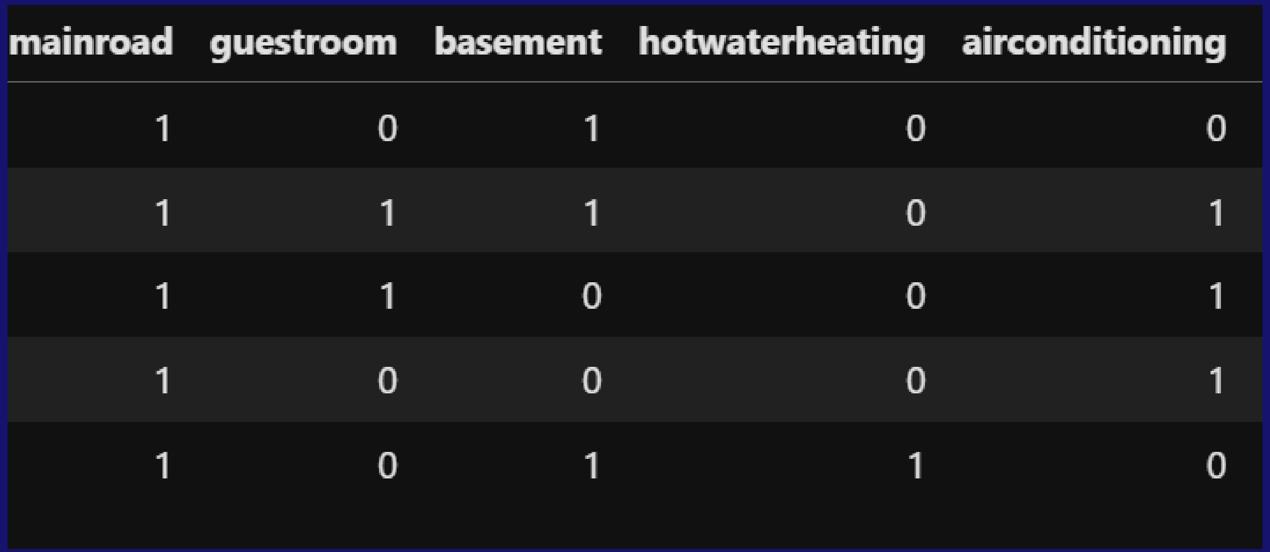
- In our dataset, we transformed categorical attributes into numerical representations to make them usable for machine learning models:

- Attributes like mainroad, guestroom were encoded as 0 for "No" and 1 for "Yes".these outliers.

Encoding:



mainroad	guestroom	basement	hotwaterheating	airconditioning
yes	no	no	no	yes
yes	no	no	no	yes
yes	no	yes	no	no
yes	no	yes	no	yes
yes	yes	yes	no	yes



mainroad	guestroom	basement	hotwaterheating	airconditioning
1	0	1	0	0
1	1	1	0	1
1	1	0	0	1
1	0	0	0	1
1	0	1	1	0

DATA MINING TASK : CLASSIFICATION

Classification is a supervised learning technique, which means it requires a class label to classify objects.

We trained our model to be able to predict if the price was high or medium or low using (price) class labe

DATA MINING TASK : CLASSIFICATION

To build our model We used a decision tree algorithm based on (IG and gini index)which is a recursive algorithm produces a tree with a leaf nodes representing the final decisions.

The final decision for class price is either true or false. the prediction is made on the rest attributes.
('area', 'bedrooms', 'bathrooms', 'airconditioning', 'prefarea')

DATA MINING TASK : CLASSIFICATION

- This technique includes dividing the dataset into two sets:

1. **TRAINING DATASET:** USED FOR BUILDING THE DECISION TREE.
2. **TESTING DATASET:** USED TO EVALUATE THE CONSTRUCTED MODEL.

- We tried 3 different sizes of training and testing data to get the best result for construction and evaluation.

1. 90% TRAINING 10% TESTING
2. 80% TRAINING 20% TESTING
3. 70% TRAINING 30% TESTING

ACCURACY COMPARISON

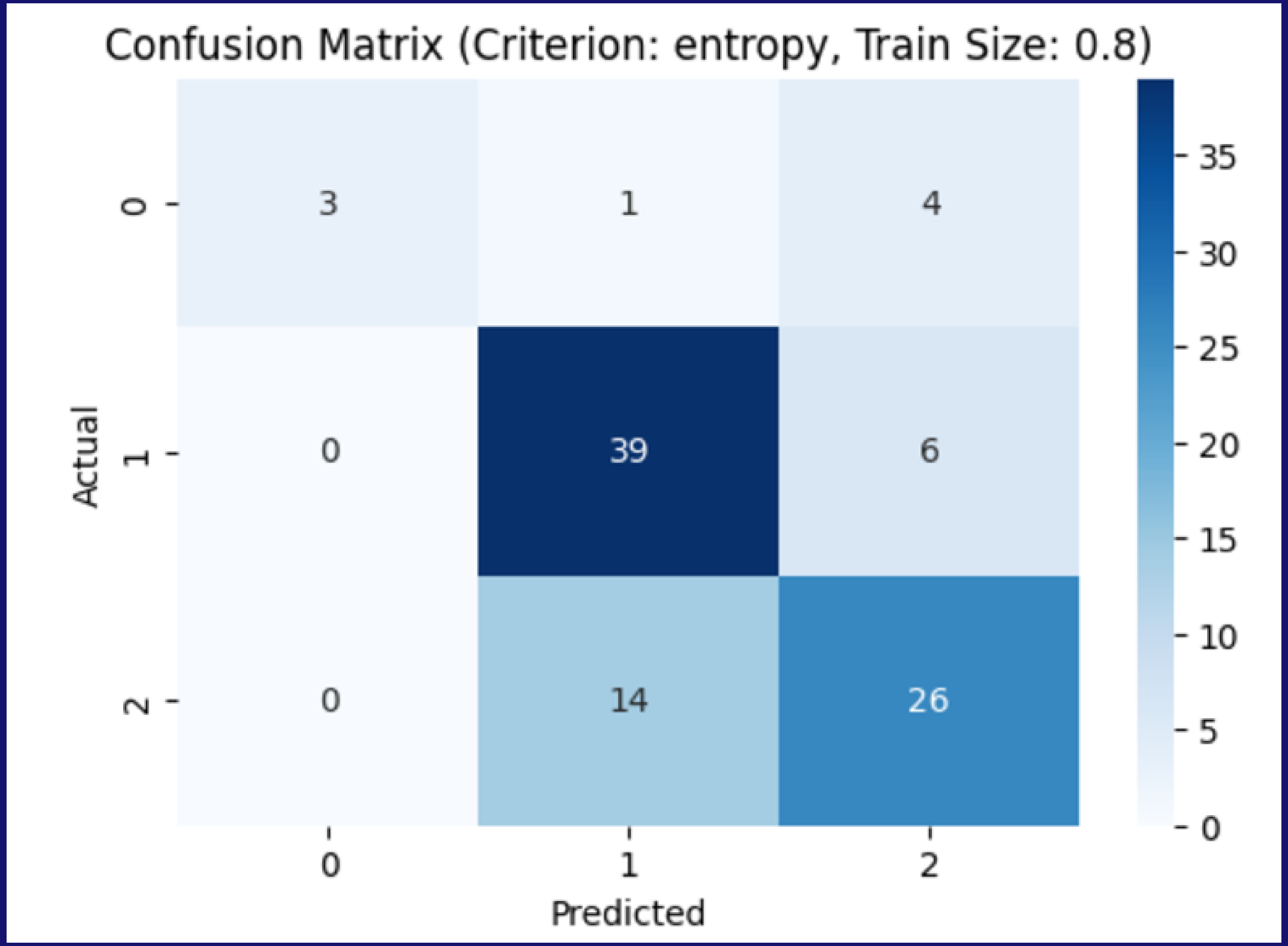
	90-10		80-20		70-30	
	IG	Gini	IG	Gini	IG	Gini
Accuracy	0.72	0.68	0.73	0.68	0.58	0.66

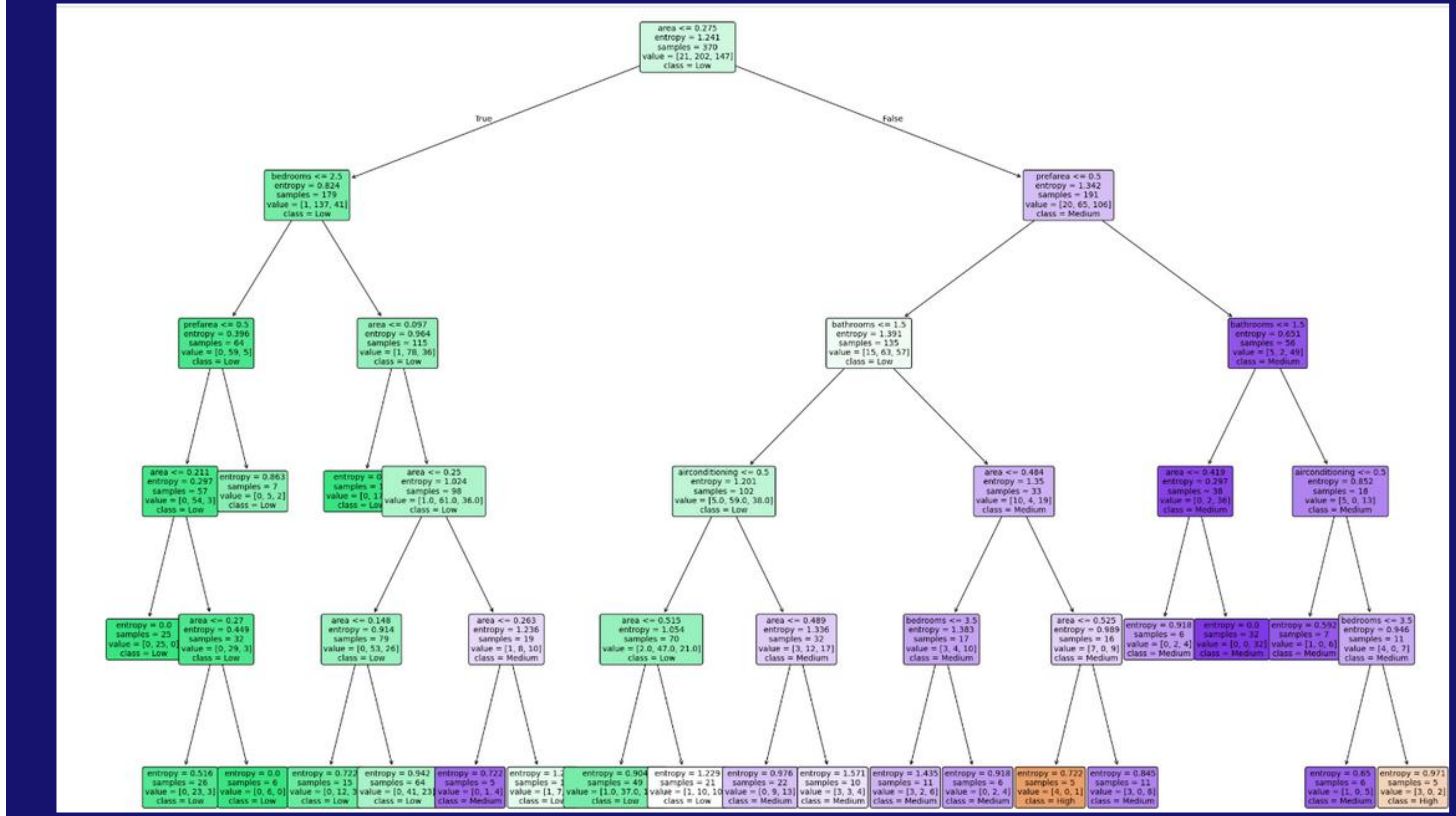
Train Size	Splitting Criterion	Class Label	Precision	Recall	Specificity
0.7	entropy	High	30.8	28.6	92.8
		Low	62.4	80.3	56.2
		Medium	56.1	39	77.5
	gini	High	100	14.3	100
		Low	67.1	83.3	63
		Medium	63.6	59.3	75
0.8	entropy	High	100	37.5	100
		Low	72.2	86.7	68.8
		Medium	72.2	65	81.1
	gini	High	66.7	50	97.6
		Low	66.7	84.4	60.4
		Medium	70	52.5	83
0.9	entropy	High	100	33.3	100
		Low	69.2	81.8	68
		Medium	75	68.2	80
	gini	High	25	33.3	93.2
		Low	66.7	90.9	60
		Medium	84.6	50	92

FINDINGS

We evaluate our model by measuring the accuracy, Precision, Recall and Specificity measures of the testing dataset

overall the best Best Train Size is 0.8 (80% training, 20% testing) and the Best Algorithm is Entropy Criterion (Information gain)


Confusion Matrix (Criterion: entropy, Train Size: 0.8)
Actual
Predicted
0
1
2
0
1
2
3
1
4
0
39
6
0
14
26
35
30
25
20
15
10
5
0


area <= 0.275 entropy = 1.241 samples = 370 value = [21, 202, 147] class = Low
True
False
bedrooms <= 2.5 entropy = 0.824 samples = 179 value = [1, 137, 41] class = Low
prefarea <= 0.5 entropy = 1.342 samples = 191 value = [20, 65, 106] class = Medium
prefarea <= 0.5 entropy = 0.396 samples = 64 value = [0, 59, 5] class = Low
area <= 0.097 entropy = 0.964 samples = 115 value = [1, 78, 36] class = Low
bathrooms <= 1.5 entropy = 1.391 samples = 135 value = [15, 63, 57] class = Low
bathrooms <= 1.5 entropy = 0.651 samples = 56 value = [5, 2, 49] class = Medium
area <= 0.211 entropy = 0.297 samples = 57 value = [0, 54, 3] class = Low
entropy = 0.863 samples = 7 value = [0, 5, 2] class = Low
entropy = 0 samples = 1 value = [0, 17 class = Low
area <= 0.25 entropy = 1.024 samples = 98 value = [1.0, 61.0, 36.0] class = Low
airconditioning <= 0.5 entropy = 1.201 samples = 102 value = [5.0, 59.0, 38.0] class = Low
area <= 0.484 entropy = 1.35 samples = 33 value = [10, 4, 19] class = Medium
area <= 0.419 entropy = 0.297 samples = 38 value = [0, 2, 36] class = Medium
airconditioning <= 0.5 entropy = 0.852 samples = 18 value = [5, 0, 13] class = Medium
entropy = 0.0 samples = 25 value = [0, 25, 0] class = Low
area <= 0.27 entropy = 0.449 samples = 32 value = [0, 29, 3] class = Low
area <= 0.148 entropy = 0.914 samples = 79 value = [0, 53, 26] class = Low
area <= 0.263 entropy = 1.236 samples = 19 value = [1, 8, 10] class = Medium
area <= 0.515 entropy = 1.054 samples = 70 value = [2.0, 47.0, 21.0] class = Low
area <= 0.489 entropy = 1.336 samples = 32 value = [3, 12, 17] class = Medium
bedrooms <= 3.5 entropy = 1.383 samples = 17 value = [3, 4, 10] class = Medium
area <= 0.525 entropy = 0.989 samples = 16 value = [7, 0, 9] class = Medium
entropy = 0.918 samples = 6 value = [0, 2, 4] class = Medium
entropy = 0.0 samples = 32 value = [0, 0, 32] class = Medium
entropy = 0.592 samples = 7 value = [1, 0, 6] class = Medium
bedrooms <= 3.5 entropy = 0.946 samples = 11 value = [4, 0, 7] class = Medium
entropy = 0.516 samples = 26 value = [0, 23, 3] class = Low
entropy = 0.0 samples = 6 value = [0, 6, 0] class = Low
entropy = 0.722 samples = 15 value = [0, 12, 3 class = Low
entropy = 0.942 samples = 64 value = [0, 41, 23] class = Low
entropy = 0.722 samples = 5 value = [0, 1, 4] class = Medium
entropy = 1.2 samples = 1 value = [1, 7, class = Low
entropy = 0.904 samples = 49 value = [1.0, 37.0, 1 class = Low
entropy = 1.229 samples = 21 value = [1, 10, 10 class = Low
entropy = 0.976 samples = 22 value = [0, 9, 13] class = Medium
entropy = 1.571 samples = 10 value = [3, 3, 4] class = Medium
entropy = 1.435 samples = 11 value = [3, 2, 6] class = Medium
entropy = 0.918 samples = 6 value = [0, 2, 4] class = Medium
entropy = 0.722 samples = 5 value = [4, 0, 1] class = High
entropy = 0.845 samples = 11 value = [3, 0, 8] class = Medium
entropy = 0.65 samples = 6 value = [1, 0, 5] class = Medium
entropy = 0.971 samples = 5 value = [3, 0, 2] class = High

DATA MINING TASK : CLUESTRING

Clustering is an unsupervised machine learning technique used to group data points into clusters based on their similarity and dissimilarity.

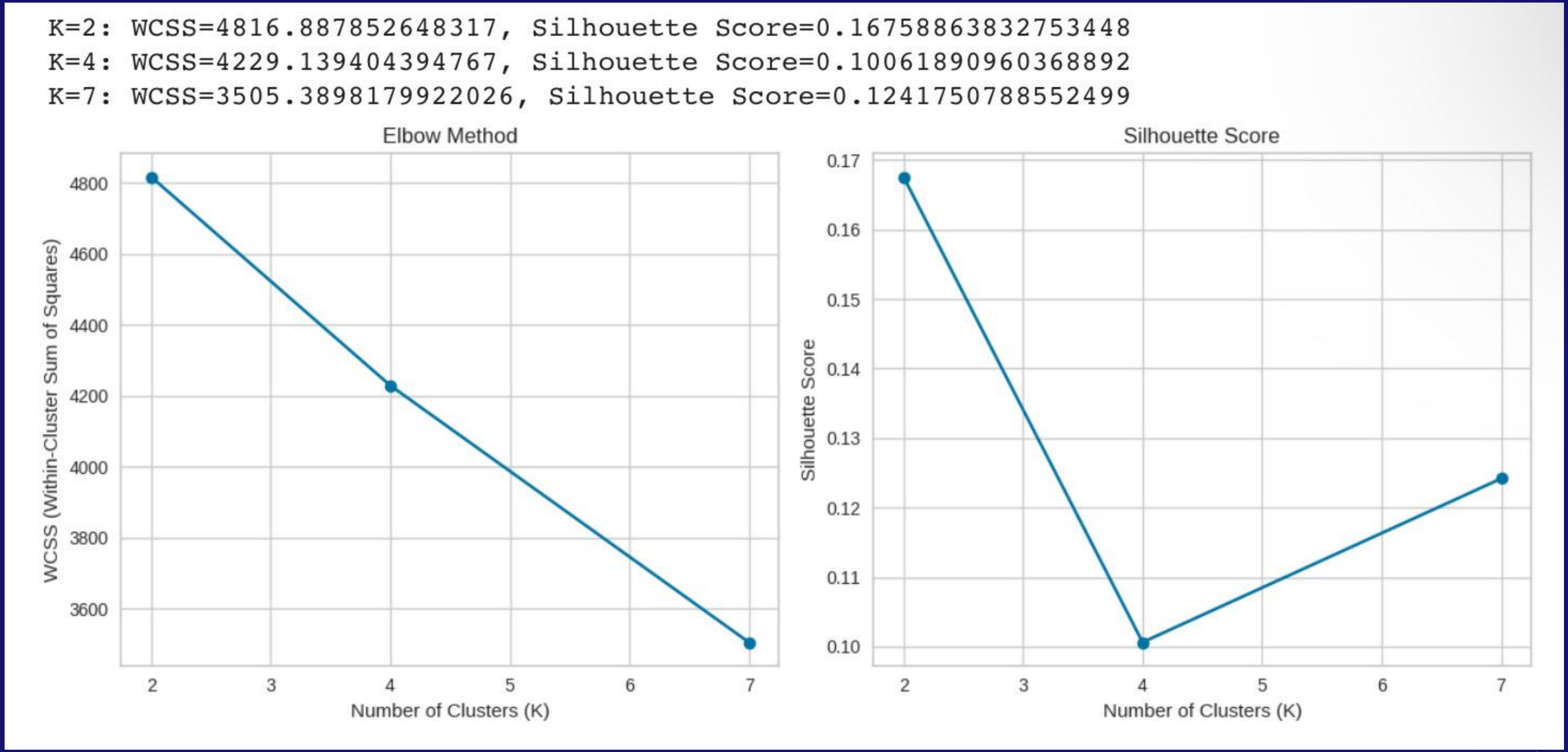
Our model will create a set of clusters for houses with similar characteristics, These clusters reveal market trends and help predict characteristics or pricing for new properties.

DATA MINING TASK : CLUESTRIN G

We used the K-Means algorithm, which groups data points into K clusters, where each cluster is represented by a central point . The algorithm assigns each data point to the cluster with the nearest centroid.

then iteratively recalculates the center and reassigns data points until the centroids no longer change, indicating that the data points are correctly grouped.

DATA MINING TASK : CLUESTRING

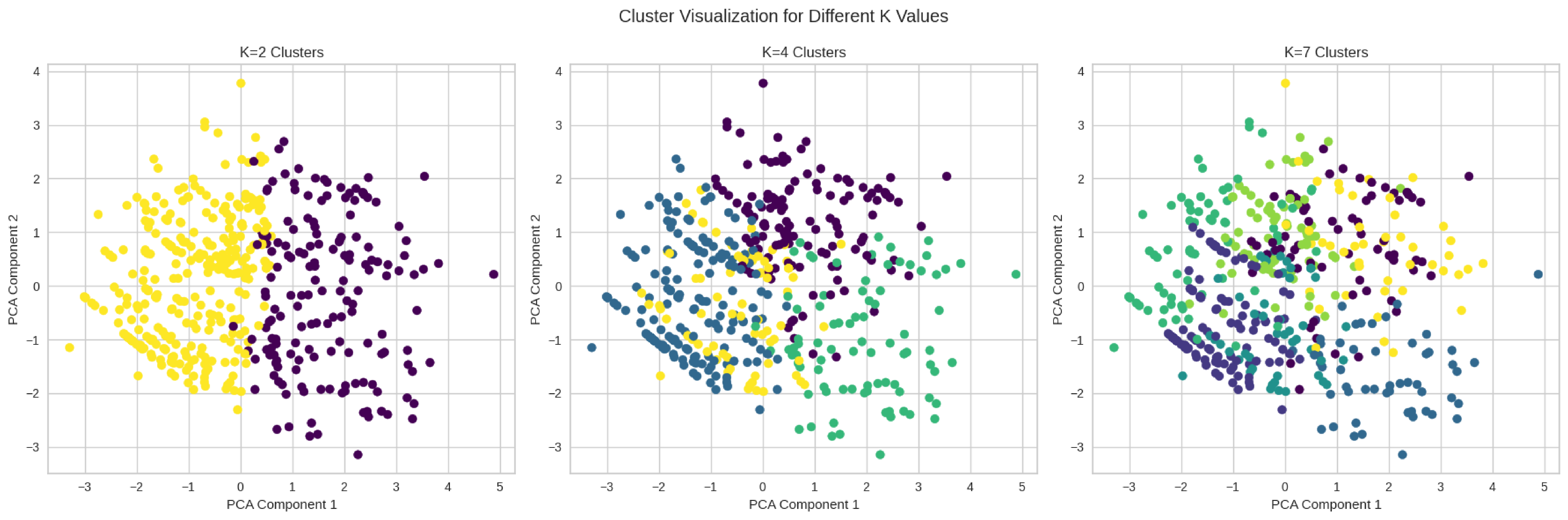


```
K=2: WCSS=4816.887852648317, Silhouette Score=0.16758863832753448
K=4: WCSS=4229.139404394767, Silhouette Score=0.10061890960368892
K=7: WCSS=3505.3898179922026, Silhouette Score=0.1241750788552499
```

FINDINGS

We tested three different values K, evaluating the models using the Silhouette Width and Within-Cluster Sum of Squares (WCSS) metrics.

The optimal model is K=4, providing a balance between compactness and meaningful segmentation of the data.



THANK YOU FOR LISTENING!

Team members:

Haifa Alsaif - Flwah Alrashed

Majd Alruways - Norah Alfaheed

Shouq Altamimi

Supervised by: Dr. Hessah Alsaaran