CSL 325
Artificial Intelligence Lab

Department of Computer Science
Bahria University, Islamabad

Instructor: Dr. Arshad Farhad

Lab # 11: Agglomerative Clustering & PCA

Objectives:

- Understand and apply clustering algorithms (Agglomerative clustering & PCA).

==

SECTION 1 — DATA PRE-PROCESSING

==

2.1 Load and Inspect Dataset

```
df = pd.read_csv("customers_messy.csv")
df.info()
df.isna().sum()
```

Observations (Fill below):

- Number of rows loaded: _20_______
- Number of missing values:
  - Age: ___1___
  - MonthlyIncome: _3_____
  - SpendingScore: __1____
  - VisitsPerMonth: __2____
- Dirty strings found: ☑ ? ☑ unknown ☑ NA □ Other: ___________

2.2 Cleaning Dirty Values

What values did you convert to NaN?

?, unknown, NA

2.3 Data Type Conversion

After conversion, data types are:

Column	Type Before	Type After
Age	object	Float64
MonthlyIncome	object	Float64
SpendingScore	object	Float64
VisitsPerMonth	object	Float64

2.4 Strategy for Handling Missing Values

Tick the method used:

- □ Dropped rows
- ☑ Filled with median
- □ Filled with mean
- □ Other: ___________________________

Why did you choose this method?

I filled missing values with the median because the dataset contains extreme outliers (like MonthlyIncome = 25000 and SpendingScore = 110).
Median is not affected by outliers, so it gives more reliable and stable values than the mean.

==

SECTION 2 — OUTLIER REMOVAL (IQR METHOD)

==

3.1 IQR Calculations

Feature	Q1	Q3	IQR	Lower Bound	Upper Bound
Age	_23.00__	_35.75___	__12.75__	___3.875_	___54.875_
MonthlyIncome	___1775.00_	__2025.00__	___250.00_	___1400.00_	__2400.00__
SpendingScore	__47.50__	__76.25__	___28.75_	_4.375___	__119.375__
VisitsPerMonth	__2.75__	__4.25__	__1.50__	__0.50__	__6.50__

3.2 Rows Removed by IQR

- Original rows: __20____
- Rows after IQR removal: __17____
- Total removed: ___3___

Are the removed entries real outliers?

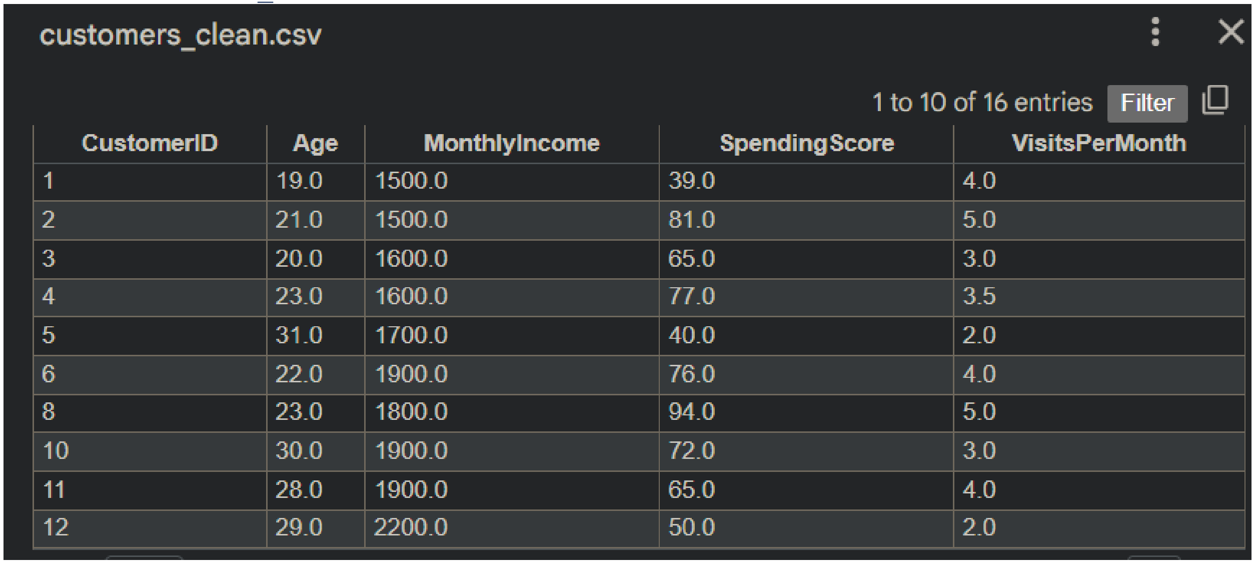
Yes. The removed rows contained extreme values such as MonthlyIncome = 25000, SpendingScore = 110 or -5, and inconsistent visit patterns. These do not match the general distribution of the dataset and are treated as real outliers.

==

SECTION 3 — FINAL CLEANED DATASET

==

Save as: `customers_clean.csv`



CustomerID	Age	MonthlyIncome	SpendingScore	VisitsPerMonth
1	19.0	1500.0	39.0	4.0
2	21.0	1500.0	81.0	5.0
3	20.0	1600.0	65.0	3.0
4	23.0	1600.0	77.0	3.5
5	31.0	1700.0	40.0	2.0
6	22.0	1900.0	76.0	4.0
8	23.0	1800.0	94.0	5.0
10	30.0	1900.0	72.0	3.0
11	28.0	1900.0	65.0	4.0
12	29.0	2200.0	50.0	2.0

Number of final rows after cleaning: _____8___

==

SECTION 4 — AGGLOMERATIVE CLUSTERING (SCIPY)

==

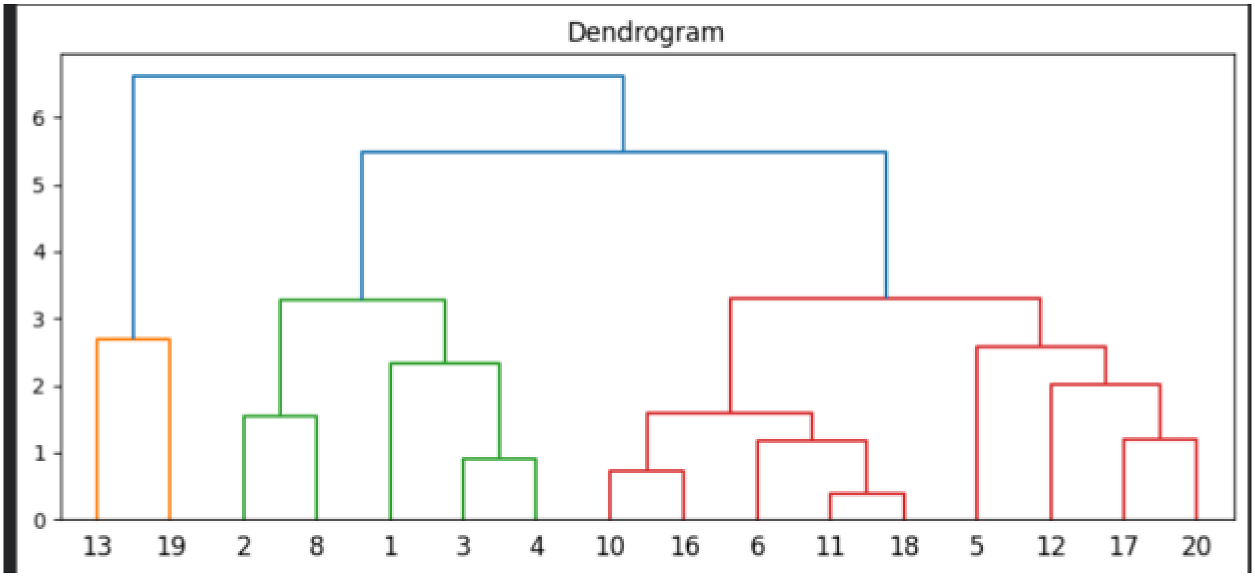
5.1 Standardization

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Why do we scale data before clustering?We scale data so that all features have equal importance in distance-based clustering. Without scaling, large-range features dominate the clustering result.

5.2 Dendrogram

Attach your dendrogram screenshot here:



5.3 Optimal Number of Clusters

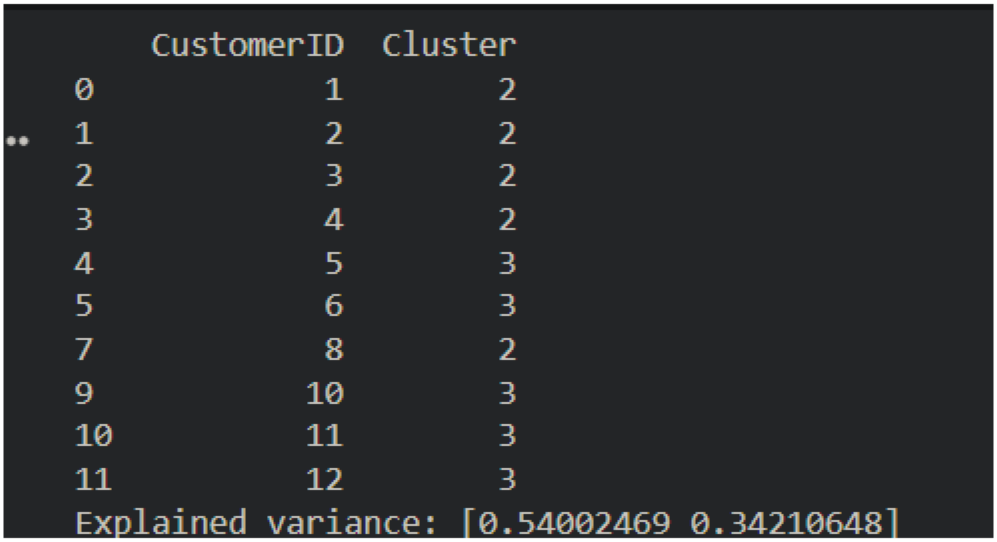
Based on dendrogram:

I choose _____**3**__ **clusters** because:

5.4 Cluster Assignment

Write first 10 entries:

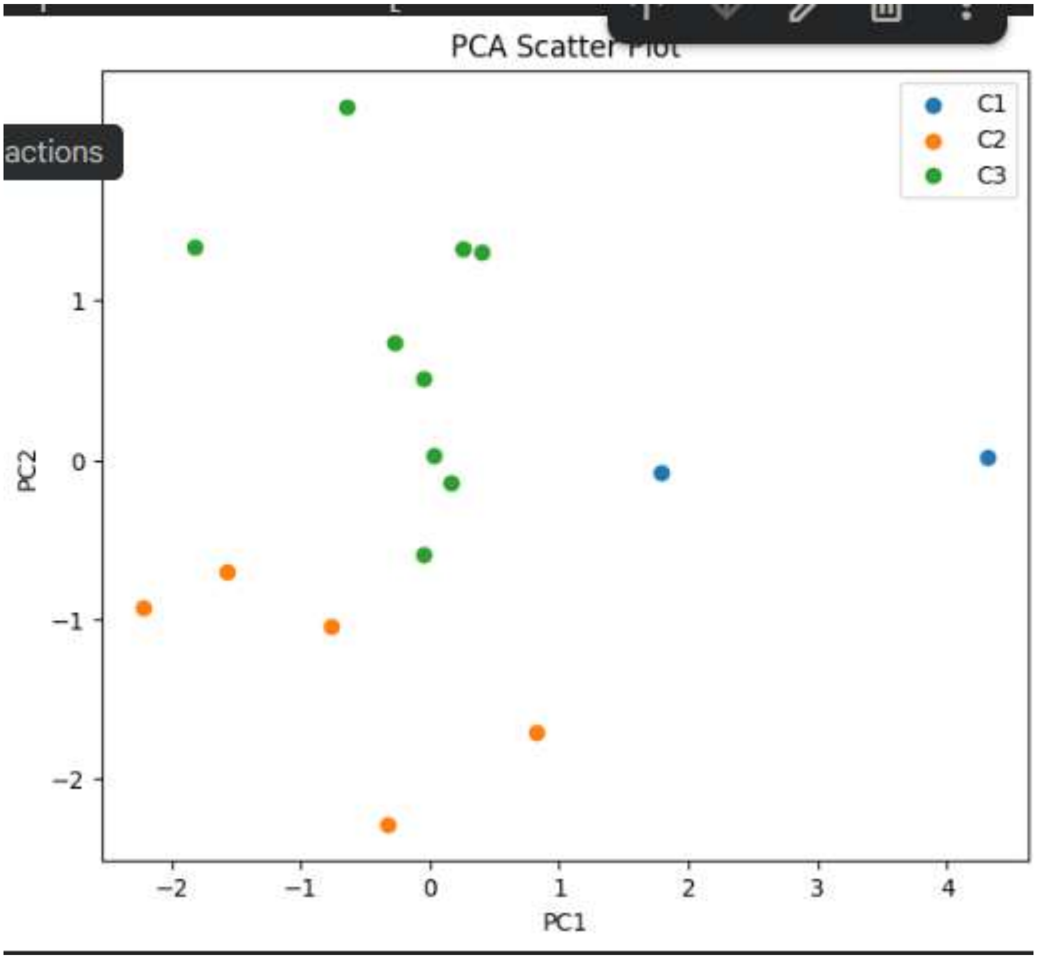
CustomerID	**Cluster**
1	____
2	____
3	____
4	____
5	____
6	____
7	____
8	____
9	____
10	____



```
      CustomerID  Cluster
0              1        2
1              2        2
2              3        2
3              4        2
4              5        3
5              6        3
7              8        2
9             10        3
10            11        3
11            12        3
Explained variance: [0.54002469 0.34210648]
```

5.5 2D Cluster Plot

Attach your visualization here:



==

SECTION 5 — PRINCIPAL COMPONENT ANALYSIS (PCA)

==

6.1 PCA Computation

Explained Variance

Component	Variance Ratio
PC1	___0.5400___
PC2	____0.3421__

Total Variance Explained = ___0.8821___ %

6.2 PCA Scatter Plot

Attach the scatter plot (PC1 vs PC2 with cluster colors):

(Insert Image)

6.3 Interpretation of PCA Components

==

SECTION 6 — LAB CONCLUSION

==

Write your conclusion:

- Did PCA help in visualizing clusters?

Yes, PCA helped a lot. After reducing the data to two components (PC1 and PC2), the clusters became easy to see on a simple 2-D scatter plot. Even though the clustering was done on the original features, PCA allowed us to clearly visualize how different groups of customers are separated.

• Did IQR removal improve results?
Yes, removing outliers using the IQR method improved the clustering. Outliers usually pull distances in the wrong direction and make clusters messy. After removing them, the dendrogram looked cleaner, and the PCA clusters became more compact and meaningful.

• What did you learn about hierarchical clustering?

I learned that hierarchical clustering builds clusters step-by-step by merging the closest points or groups. The dendrogram is the main tool to understand it. By looking at the height of merges in the dendrogram, we can decide the best number of clusters. It also does not require choosing the number of clusters at the start like K-means, which makes it useful for exploring data.

-