Customer Segmentation for an Online Store using K-Means Clustering

PRESENTED BY

MD LIMON MIA
Roll:08

ANANNA RANI DASH
Roll:09

DEPARTMENT OF CSE
DHAKA INTERNATIONAL UNIVERSITY

08/08/25

Introduction:

- Growth of e-commerce → more customer data
- Importance of understanding customer behavior
- Goal: Use ML to segment customers for better marketing

Problem Statement:

- Businesses struggle to market effectively to a diverse customer base
- Need to group similar customers **Customer Segmentation**
- Traditional methods are limited Use **K-Means Clustering**

Objectives:

- Segment customers based on income and spending
- Identify patterns to support targeted marketing
- Improve business decision-making using data insights

Dataset Overview:

- Source: Kaggle (E-Commerce Customer Data)

- Key Features:
  - Gender
  - Age
  - Annual Income (k$)
  - Spending Score (1–100)

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...

Data Preprocessing:

- Checked for missing values (none)
- Selected features: **Annual Income** and **Spending Score**
- Applied **StandardScaler** for normalization

```
[ ] df.isnull().sum()
```

	0
CustomerID	0
Gender	0
Age	0
Annual Income (k$)	0
Spending Score (1-100)	0

dtype: int64

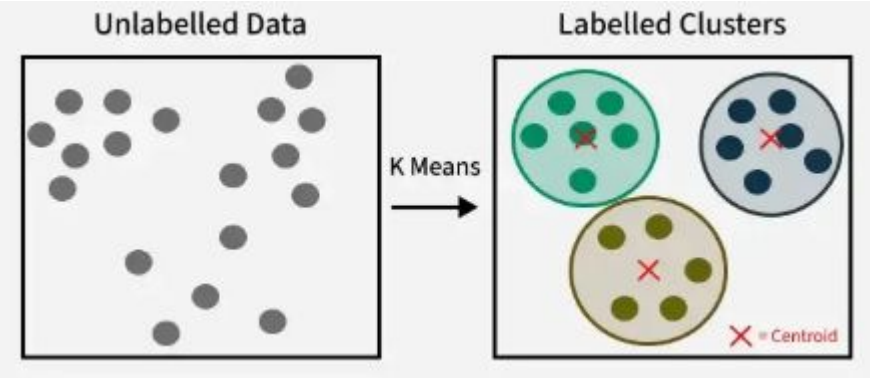
```
[ ] x = df[['Annual Income (k$)', 'Spending Score (1-100)']].copy()
```

```
[ ] x
```

	Annual Income (k$)	Spending Score (1-100)
0	15	39
1	15	81
2	16	6
3	16	77
4	17	40
...	...	...
195	120	79
196	126	28
197	126	74
198	137	18
199	137	83

200 rows × 2 columns

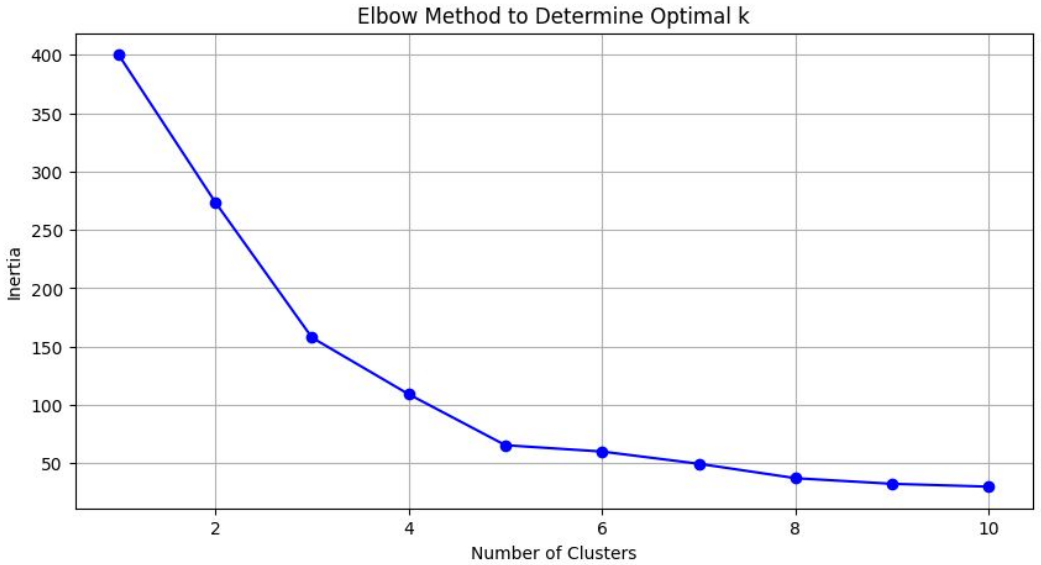
K-Means Clustering Algorithm:



1. Choose number of clusters **K**
2. Randomly initialize **K centroids**
3. Assign each data point to the **nearest centroid** $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
4. Recompute centroids as the **mean of assigned points**
5. Repeat steps 3–4 until centroids don’t change (converge)

Choosing Optimal Clusters:

- Used **Elbow Method**
- Plotted Inertia vs. Number of Clusters
- Optimal **k = 5** (based on elbow curve)

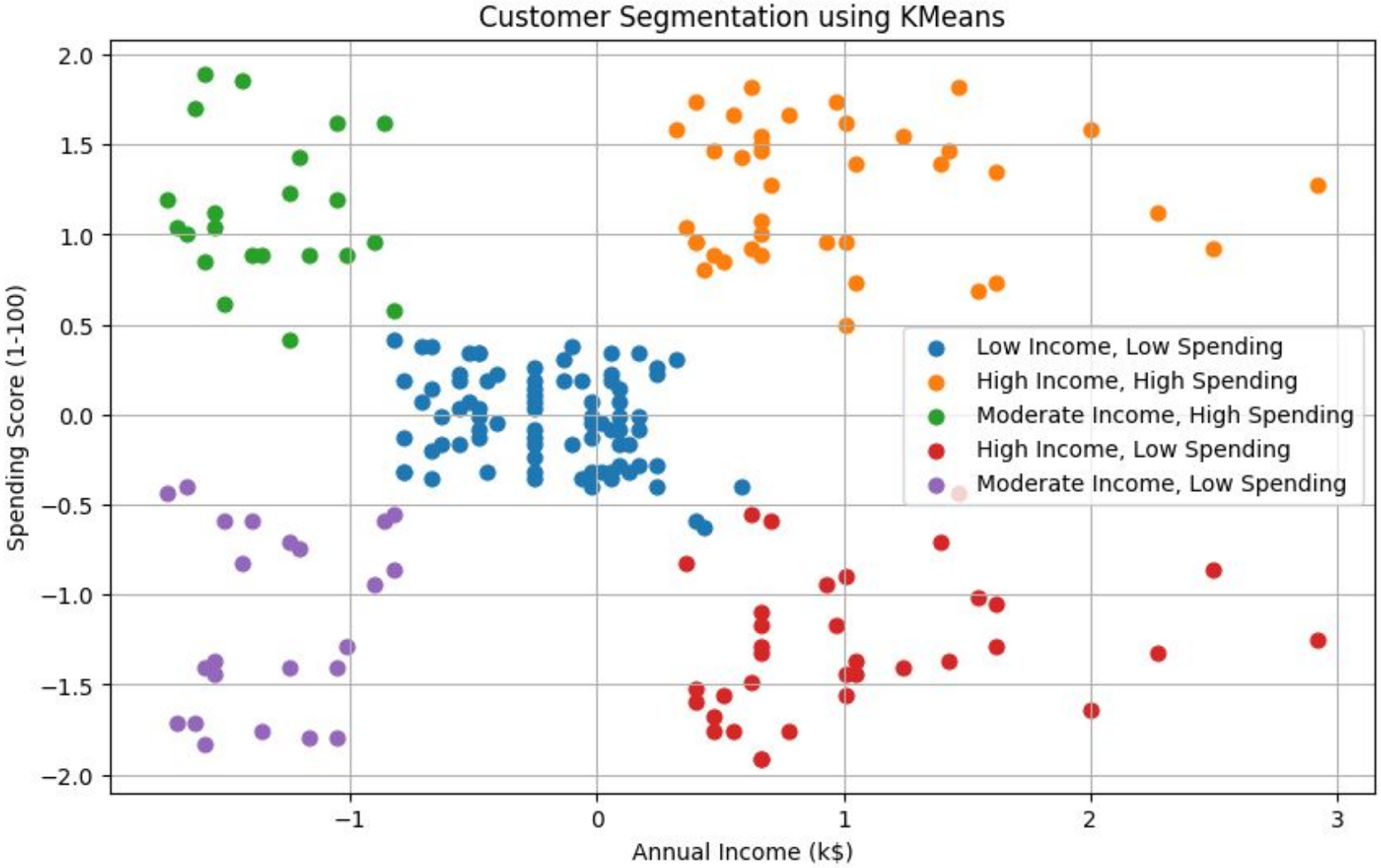


Cluster Visualization:

- 2D scatter plot using Matplotlib
- X-axis: Income (scaled)
- Y-axis: Spending Score (scaled)
- 5 Color-coded Clusters

Cluster Interpretation:

1. High Income, Low Spending
2. Low Income, Low Spending
3. High Income, High Spending *(Target)*
4. Moderate Income, High Spending *(Loyal)*
5. Low Income, High Spending *(Irregular)*


Customer Segmentation using KMeans
Spending Score (1-100)
Annual Income (k$)
Low Income, Low Spending
High Income, High Spending
Moderate Income, High Spending
High Income, Low Spending
Moderate Income, Low Spending
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−1
0
1
2
3

Cluster Summary Table:

Segment	Annual Income (k$)	Spending Score (1-100)
High Income, High Spending	86.54	82.13
High Income, Low Spending	88.20	17.11
Low Income, Low Spending	55.30	49.52
Moderate Income, High Spending	25.73	79.36
Moderate Income, Low Spending	26.30	20.91

Tools & Libraries Used:

- **Python**
- **Scikit-learn** → Clustering, scaling
- **Matplotlib** → Visualization
- **Pandas** → Data handling
- **Google Colab**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
```

Conclusion:

- K-Means helped uncover meaningful customer segments
- Business can now tailor marketing efforts more effectively
- Achieved the goal of improved customer understanding

References:

- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters.
- Vijay Choudhary, “Customer Segmentation Dataset,” Kaggle. [Online]. Available:https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python
- [3] Kotler, P., and Keller, K.L., “Marketing Management,” 15th ed., Pearson Education, 2016.
- https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/

Q&A

Open for questions

Thank You!