Major Project Report
Name: Kinshuk Gupta
Roll Number: 12341190
Course Name: Data Analytics and Visualization
Course Code: DSL251
Institute: IIT Bhilai

Data Collection, Cleaning, and Pre-Processing (Data Abstraction)

- **Dataset Used**: *E-commerce Customer Behavior*
  Includes demographics, browsing time, and purchase behavior.
- **Missing Values**: Handled using mean/median imputation or dropped.
- **Encoding**: Applied label encoding and one-hot encoding for categorical variables.
- **Scaling**: Used StandardScaler on numerical features.
- **Feature Engineering**:
    - Created engagement scores.
    - Derived behavior-based cluster features.

Data Visualization and Model Insights

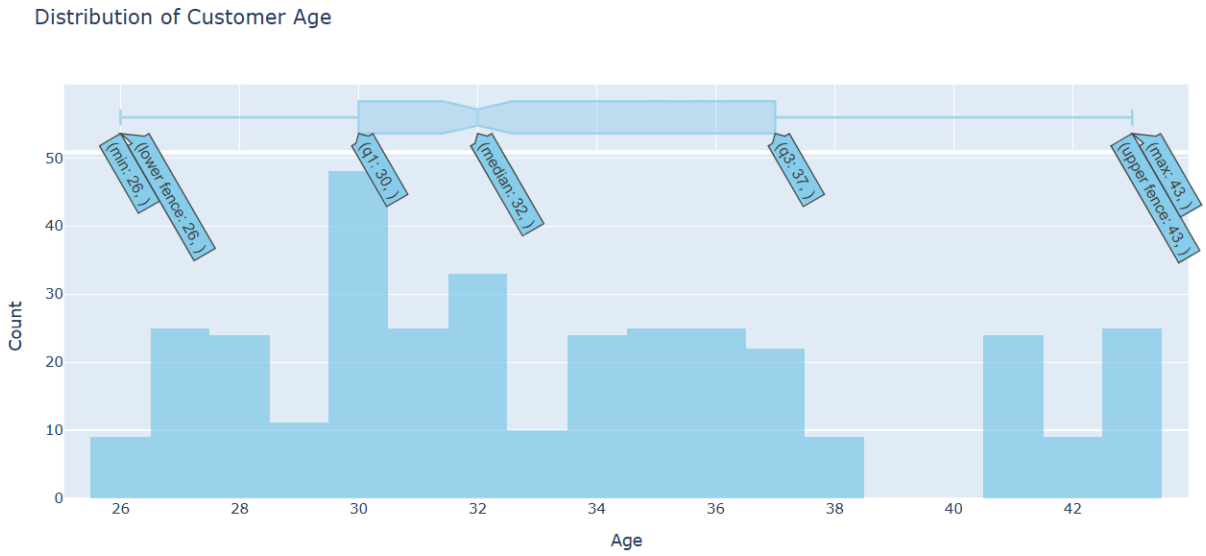
i. Task Abstraction (Validated with 5 Users)

- **Objective**: Segment customers and predict purchase behavior.
- **User Feedback**:
    - Useful for targeted marketing.

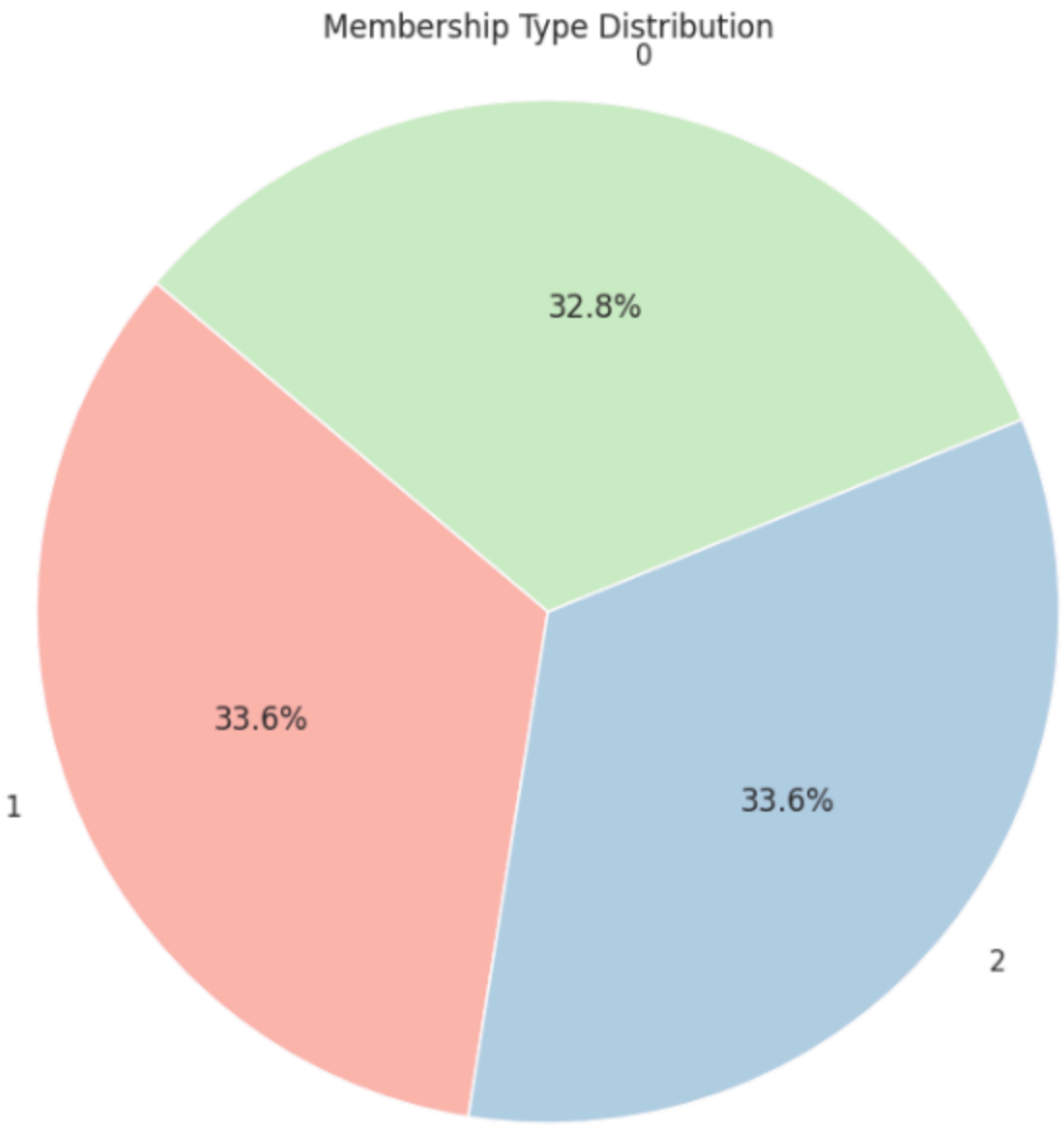
- Helps in building recommendation systems.
- Assists UX research on browsing vs. buying gap.

ii. Visual Encoding

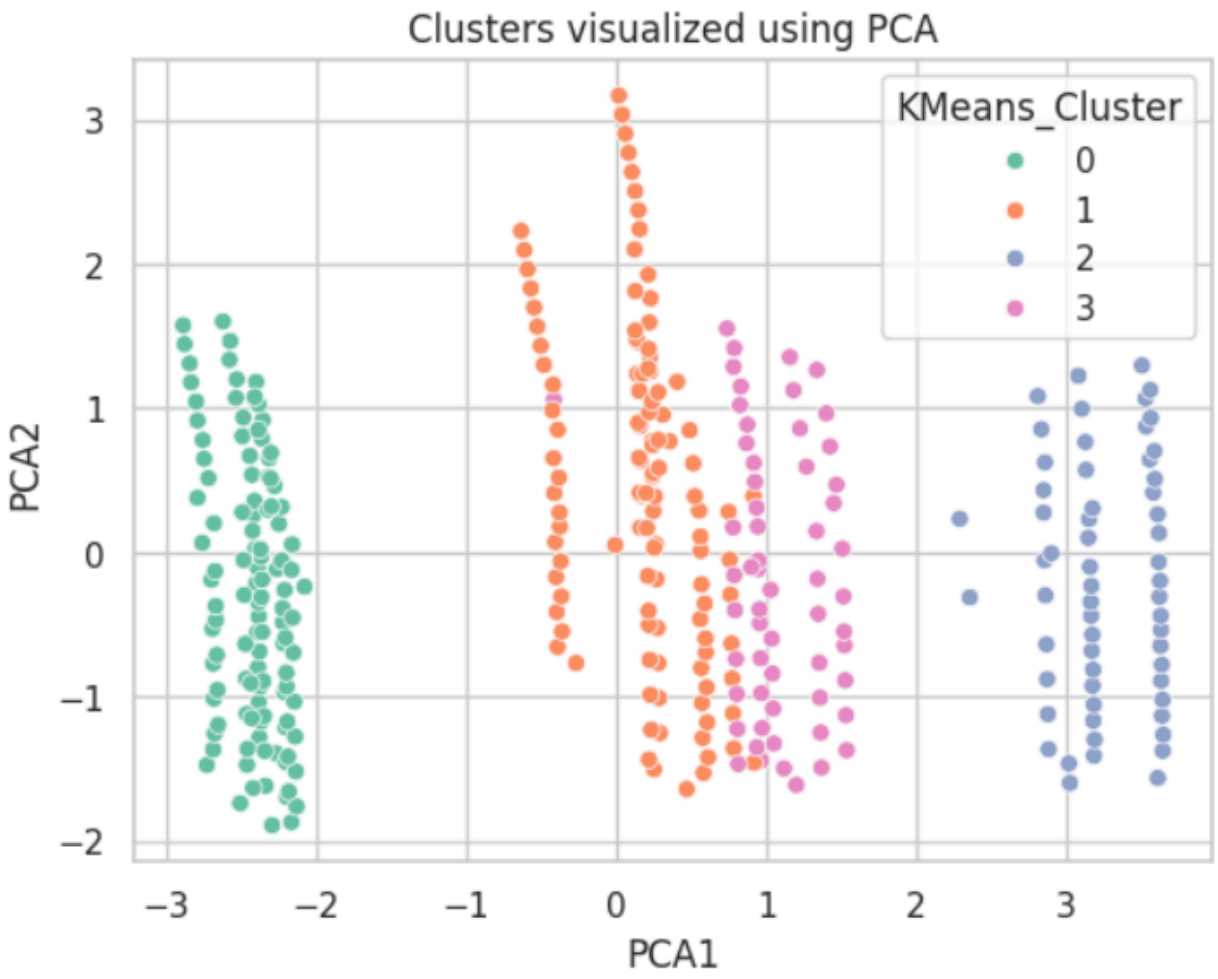
- **Histograms/Bar Charts**: For demographics and spending patterns.



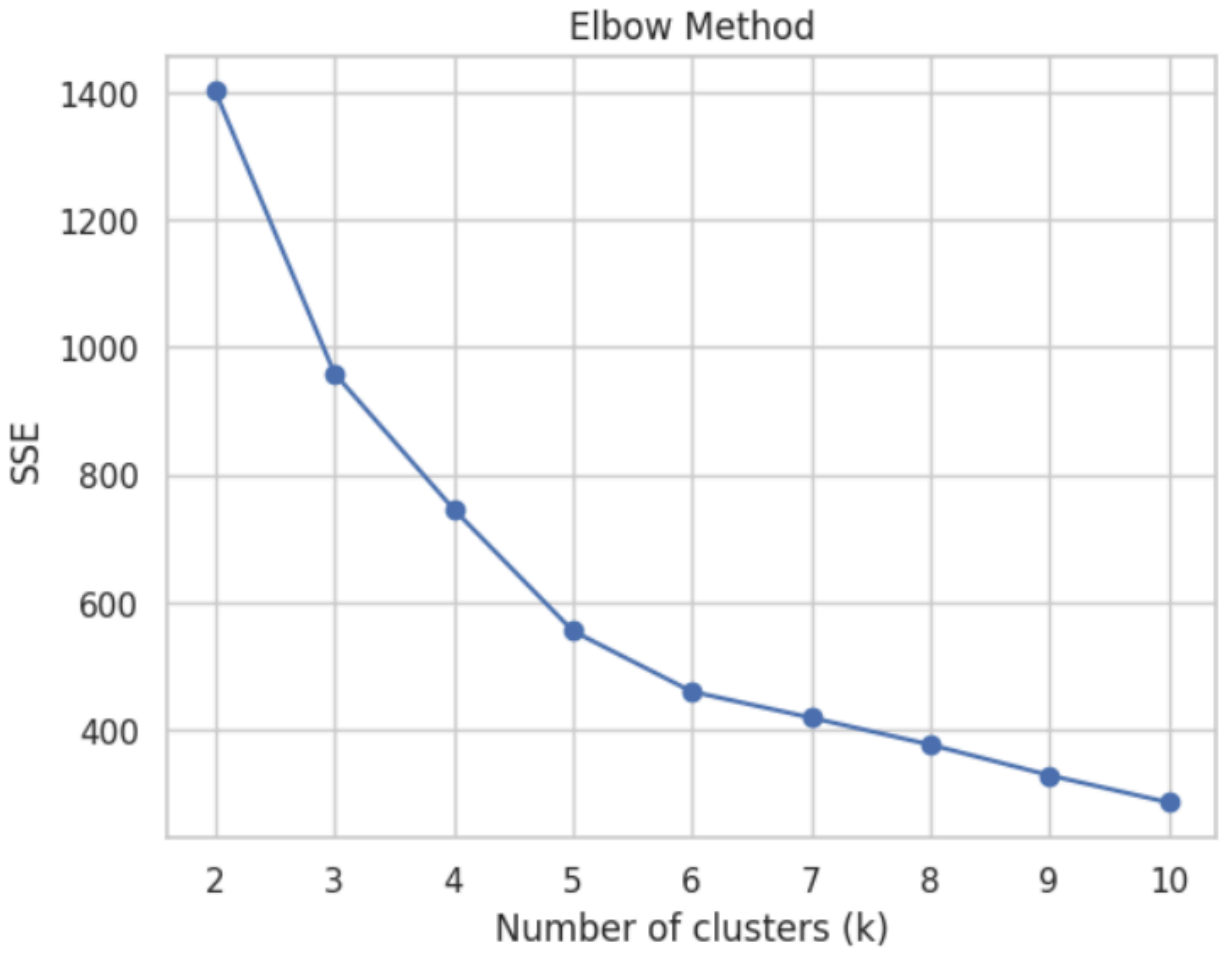
- **Pie Charts**: Gender and product category distribution.



- **Scatter Plots (PCA)**: Display clustering (KMeans k=5) visually.



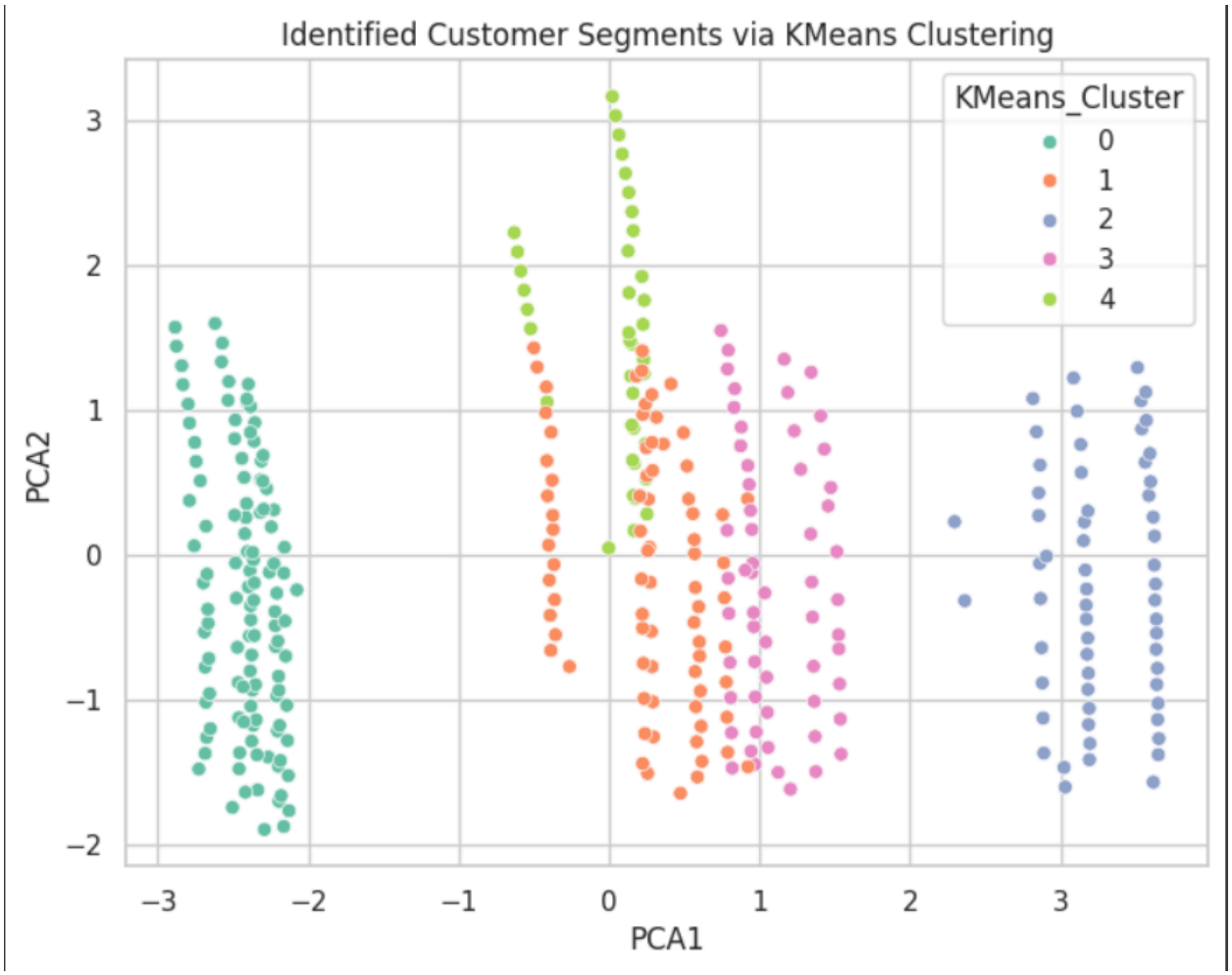
- **Elbow Curve & Silhouette Score**: Used to decide optimal cluster number.




```
Silhouette Score for KMeans using k=5: 0.48823132914102185
```

Models and Insights

Clustering Models



- **K-Means (k=5)**:
    - Cluster 0: Low spenders.
    - Cluster 1: High-income, loyal buyers.
    - Cluster 2: Browsers, rare buyers.
    - Cluster 3: Moderate spenders, upsell potential.
    - Cluster 4: Inconsistent engagement.

- **DBSCAN**:
    - Identified dense regions and outlier points.

Classification Models

- **Random Forest**:
    - Accuracy: 91.4%
    - Precision (macro): 0.745
    - Recall (macro): 0.80
    - F1-Score (macro): 0.769
- **Decision Tree**:
    - Accuracy: 91.4%
    - Precision (macro): 0.744
    - Recall (macro): 0.80
    - F1-Score (macro): 0.769

Note: The performance of both models was similar; Random Forest was slightly more robust.

Key Insights

- *Clustering helps design **custom campaigns** for each segment.*
- *Classification models aid in **purchase prediction**.*
- *Supports **personalized recommendation systems**.*
- *Highlights the gap between **browsing and actual purchasing**, guiding UX changes.*