Customer Segmentation using KMeans Clustering

Course: Principles of Data Science

Semester: Spring 2025

**Team Members:**

Priyanka Adusumilli

Dheeraj Golla

Sai Revanth Patchala

Hari Krishna Prasad Ravella

# 1. Introduction

In today’s competitive market, understanding customer behavior is crucial for business success. Customer segmentation helps companies categorize their client base into distinct groups based on shared characteristics, allowing for more personalized marketing strategies and service offerings. This project aims to build a segmentation model using unsupervised learning — specifically the KMeans clustering algorithm — on mall customer data to uncover meaningful customer segments for better business decision-making.

# 2. Related Work

Traditional customer segmentation methods rely on manual heuristics or rule-based classifications. However, with the advent of machine learning, data-driven approaches such as clustering algorithms have gained prominence. KMeans remains a widely used method due to its simplicity and scalability, especially when clusters are spherical and well-separated. Other clustering algorithms like DBSCAN and Hierarchical Clustering offer alternatives in cases with arbitrary-shaped clusters or varying densities but may be more complex or sensitive to hyperparameters. Recent works in retail analytics have employed KMeans for segmenting customers by purchasing behavior, income brackets, and lifetime value.

# 3. Methodology

The approach followed in this project is outlined below:  
- Dataset loaded from a CSV file.  
- Column names standardized to lowercase.  
- Dynamic column matching for robustness.  
- Gender encoded to binary (Male = 1, Female = 0).  
- Selected features: Gender, Age, Annual Income, Spending Score.  
- Applied StandardScaler to normalize data.  
- Used Elbow Method to choose optimal k.  
- Implemented KMeans clustering with k=5.  
- Built an interactive Streamlit dashboard for real-time predictions and visualizations.

A screenshot of a screen

Description automatically generated

# 4. Results and Discussion

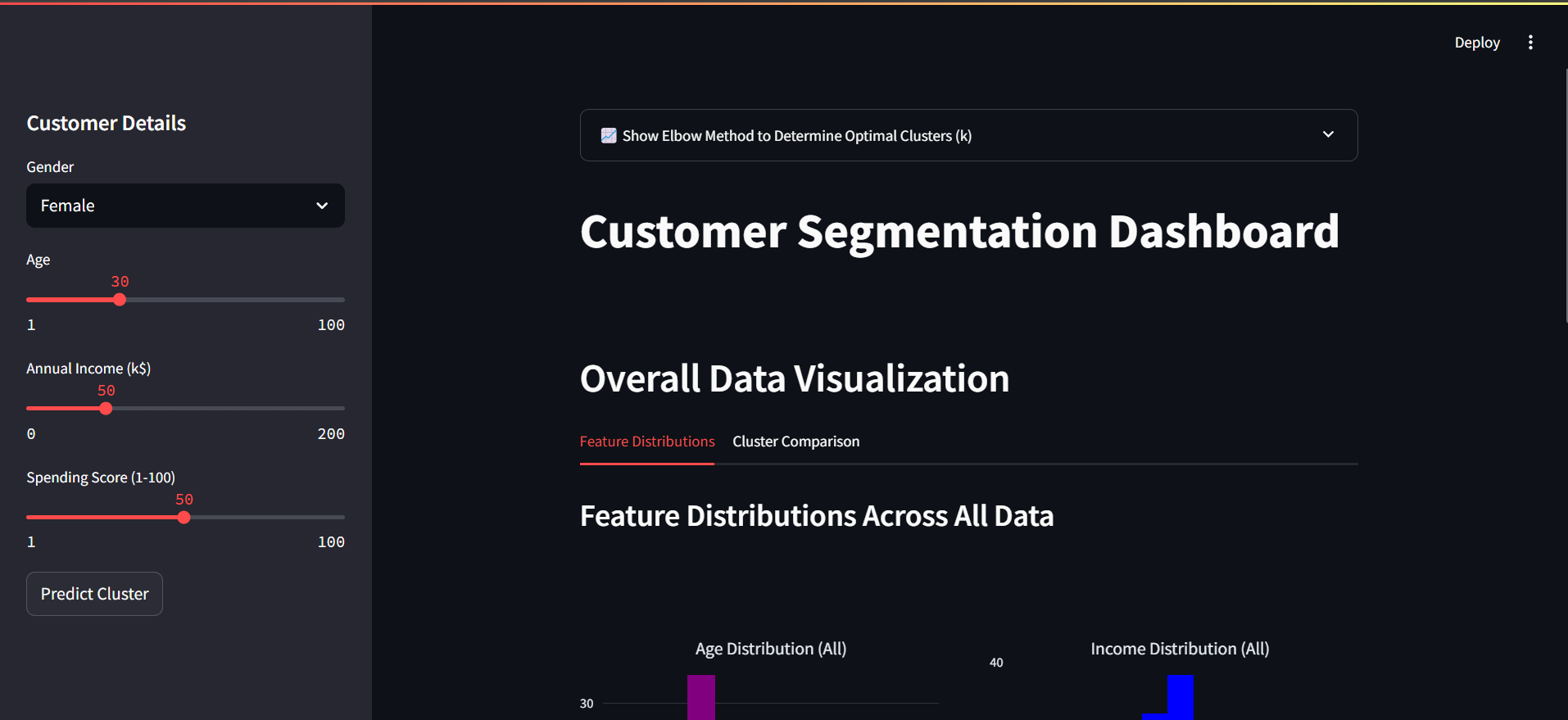
The Elbow Method indicated that 5 clusters yield optimal performance. The clusters show meaningful separation between customer types, such as:  
- High income, high spenders  
- Low income, low spenders  
- Young average spenders  
- Balanced segment  
  
KMeans was chosen over DBSCAN and Hierarchical Clustering due to its efficiency and the dataset's clean, well-separated nature. Limitations include sensitivity to outliers and the need to predefine k.

# 5. Conclusion and Future Work

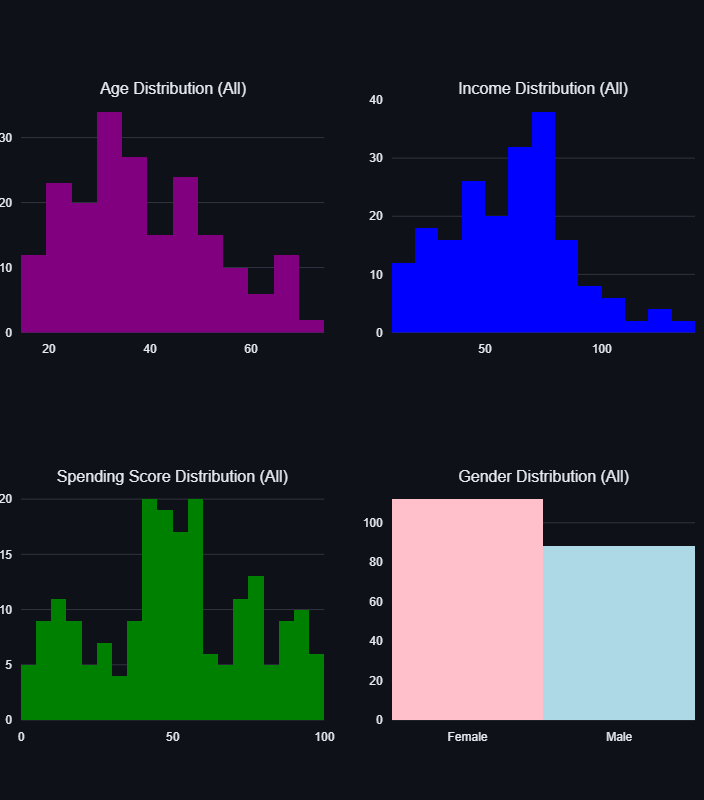
The project effectively demonstrates the use of KMeans clustering for customer segmentation. Preprocessing and visualization steps ensured accurate clustering. Future work can explore alternate algorithms like DBSCAN or Gaussian Mixture Models, including richer behavioral features, and deploy the dashboard online.

# Appendix: Visual Outputs

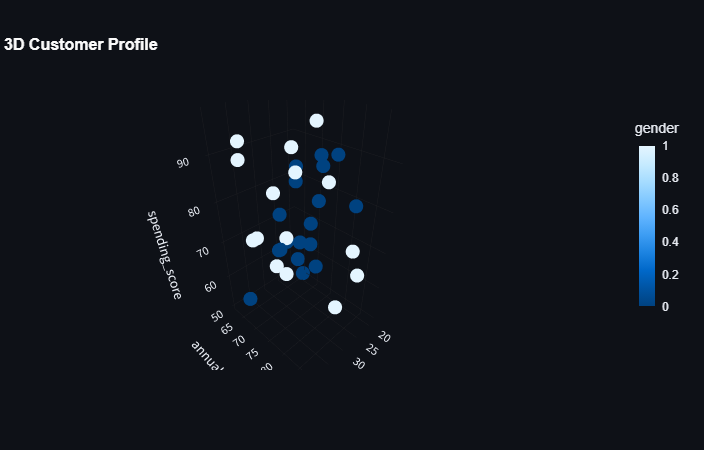
## Streamlit Dashboard UI



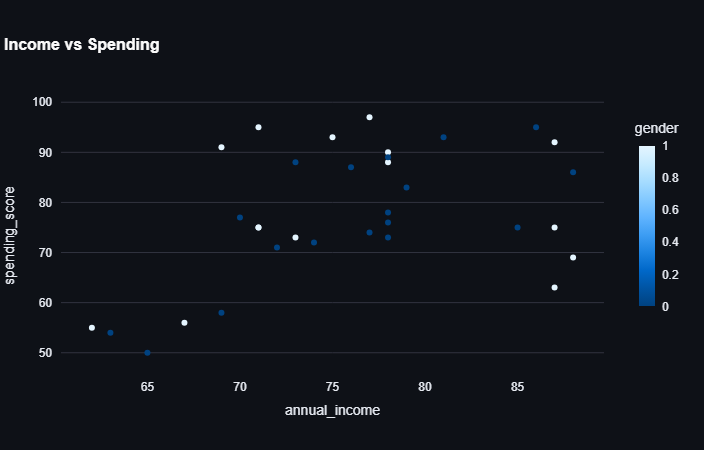
## Feature Distribution Histograms



## 3D Customer Profile View



## Income vs Spending Scatter Plot



## Elbow Method Plot

