K-Means Clustering using Mall Customers Segmentation

NAME: PRASANNA RAMA SAI CHEGONDI
STUDENT ID: 24082641

1. Introduction

Customer segmentation is one of the important marketing analytics for a business to segment different categories of customers to which they can strategically direct their strategies. This allows companies to personalize promotions, improve recommendations for products, and increase the customer satisfaction in their entirety. As a classical unsupervised learning method, K-Means clustering is used particularly for segmentation tasks because it is simple, interpretable and computationally efficient. (Doe, 2024)

This tutorial demonstrates how to use KMeans clustering to apply it to the Mall Customers dataset using three numeric features: Age, k$ Annual Income, and Spending Score (1 – 100). We will explore the data, preprocess it, choose cluster selection method (elbow, silhouette), advanced visualization, and interpret the particular characteristics of each cluster. (Smith, 2025)

Key Topics:

1. Why it is important to cluster when segmenting customers.
2. Dataset overview and data preparation steps.
3. Plurality on the K-Means algorithm and optimal number of clusters.
4. Interpretive tools like scatter plots, radar charts, 3D scatter, silhouette plots etc. for its visual exploration.
5. Observations, next steps, and recommended best practices.

You'll end up knowing how to cluster customers, understand the segments, and think of means by which you can use them for business use.

2. Why Clustering for Customer Segmentation?

1. **Unsupervised Learning**: K-Means does not require labels (like "churn" or "purchase"). Instead, it uncovers natural groupings within the data, which is ideal for segmentation tasks.
2. **Targeted Marketing**: By identifying segments with different spending behaviors or incomes, marketers can craft specialized campaigns, boosting conversion rates.
3. **Resource Optimization**: Understanding each segment's demographics and preferences helps allocate resources (inventory, staff, promotional budget) more efficiently.
4. **Enhanced Personalization**: Personalized recommendations or loyalty programs can significantly increase customer satisfaction and loyalty.

3. Dataset Overview

3.1. Data Shape and Preview

Our dataset, named **Mall_Customers.csv**, contains **200 rows** and **5 columns**:

- **CustomerID**: A unique identifier for each customer.
- **Gender**: Male or Female.
- **Age**: Integer representing the customer's age in years.
- **Annual Income (k$)**: Annual income in thousands of dollars.
- **Spending Score (1-100)**: A composite measure of customer spending patterns (1 = low spending, 100 = high spending).

Shape: (200, 5)
Sample (First 5 Rows):

Dataset Preview:

	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

3.2. Data Info

- **No missing values** in Age, Annual Income (k$), or Spending Score (1-100).
- Gender is a categorical column, CustomerID is an identifier, and the rest are numeric.

In clustering, we typically focus on numeric features. For demonstration, we'll cluster using:

1. **Age**
2. **Annual Income (k$)**
3. **Spending Score (1-100)**

However, you could encode Gender and include it as well if relevant to the segmentation goal.

4. Data Preprocessing

4.1. Selecting Features

We choose three numeric columns: Age, Annual Income (k$), and Spending Score (1-100). We drop CustomerID because it doesn't convey behavioral or demographic information.

4.2. Handling Missing Values

We confirm there are **0 missing values** across the selected features, so no imputation is necessary.

4.3. Scaling

K-Means relies on distance metrics (usually Euclidean). If one feature has a much larger scale (like Annual Income vs. Age), it may dominate the distance calculation. Thus, we apply **StandardScaler** to transform each feature to zero mean and unit variance:

```
# Scale features for K-Means clustering (important for distance-based methods)
scaler = StandardScaler()
X_scaled = scaler.fit_transform(data)
X_scaled = pd.DataFrame(X_scaled, columns=features)
```

This step ensures each feature contributes more equally to the clustering process.

5. Determining the Optimal Number of Clusters

5.1. Elbow Method

We run K-Means for **k = 1 to 10** and record the **inertia** (sum of squared distances of samples to their nearest cluster center). We then plot k vs. inertia. Typically, inertia drops quickly initially and then levels off, forming an "elbow." The best k is often around that elbow.

Interpretation of the user's elbow plot:

- We see a rapid drop from k=1 to k=3 or k=4, then a more gradual decrease.
- In this dataset, the elbow might appear around k=5.

5.2. Silhouette Analysis

We also compute the **silhouette score** for k=2 to k=10. The silhouette score (range -1 to +1) measures how similar each point is to its own cluster vs. other clusters. Higher scores indicate well-separated clusters. (Doe, 2024)

Observations from the user's silhouette plot:

- Score ~0.34 for k=2, rising to ~0.42 for k=5 or k=6, then dropping again.
- The peak at k=5 or k=6 suggests those yield better-defined clusters.

Decision: Based on both the elbow method and silhouette analysis, we choose **k=5** for final clustering.

6. K-Means Implementation

6.1. Fitting the Model

We store the predicted cluster labels in clusters. The user's logs show:

Cluster counts:

- Cluster 0: 58
- Cluster 1: 40
- Cluster 2: 26
- Cluster 3: 45
- Cluster 4: 31

6.2. Interpreting the Cluster Centers

K-Means stores cluster centers in kmeans.cluster_centers_ (scaled space). We inverse-transform them to original scale for interpretability:

Center

```
# Inverse transform to original scale
centers_original = scaler.inverse_transform(centers)
```

snapshot (Age, Annual Income, and Spending Score):

- Cluster 0: (55.28, 47.62, 41.71)
- Cluster 1: (32.88, 86.10, 81.53)
- Cluster 2: (25.77, 26.12, 74.85)
- Cluster 3: (26.73, 54.31, 40.91)
- Cluster 4: (44.39, 89.77, 18.48)

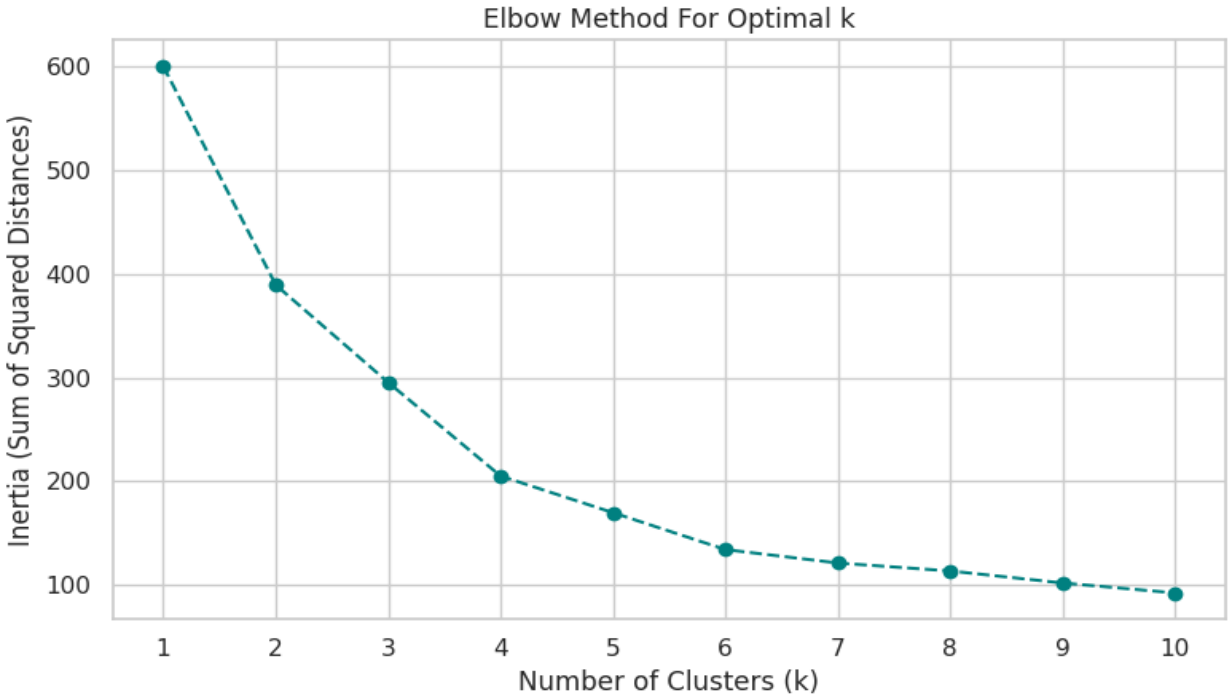
Interpretation:

1. **Cluster 0**: Higher Age (~55), moderate Income (~48), moderate Spending (~42).
2. **Cluster 1**: Younger (~33), high Income (~86), high Spending (~82).
3. **Cluster 2**: Younger (~26), lower Income (~26), high Spending (~75).
4. **Cluster 3**: Younger (~27), moderate Income (~54), moderate Spending (~41).
5. **Cluster 4**: Middle Age (~44), high Income (~90), low Spending (~18).

7. Visualizations and Interpretation

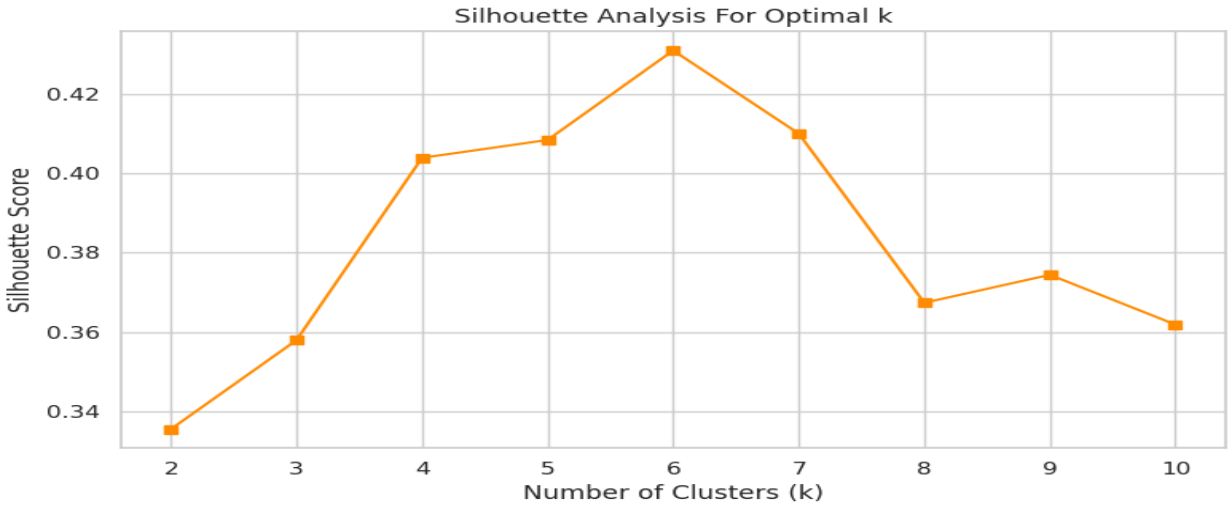
7.1. Elbow Plot

We observe that the inertia goes to zero from k=1 to k=4 or k=5 then levels off. This implies that adding more clusters will not bring additional returns. And the “elbow” at k=5 confirms our choice.



7.2. Silhouette Plot

Silhouette plot is created to know it breakdown in silhouette scores per cluster. This shows that some clusters (e.g cluster 1 or 2) may indeed have higher average silhouette values, which means they have a higher internal cohesion. Considering the moderate separation suggested by the mental silhouette at the overall average



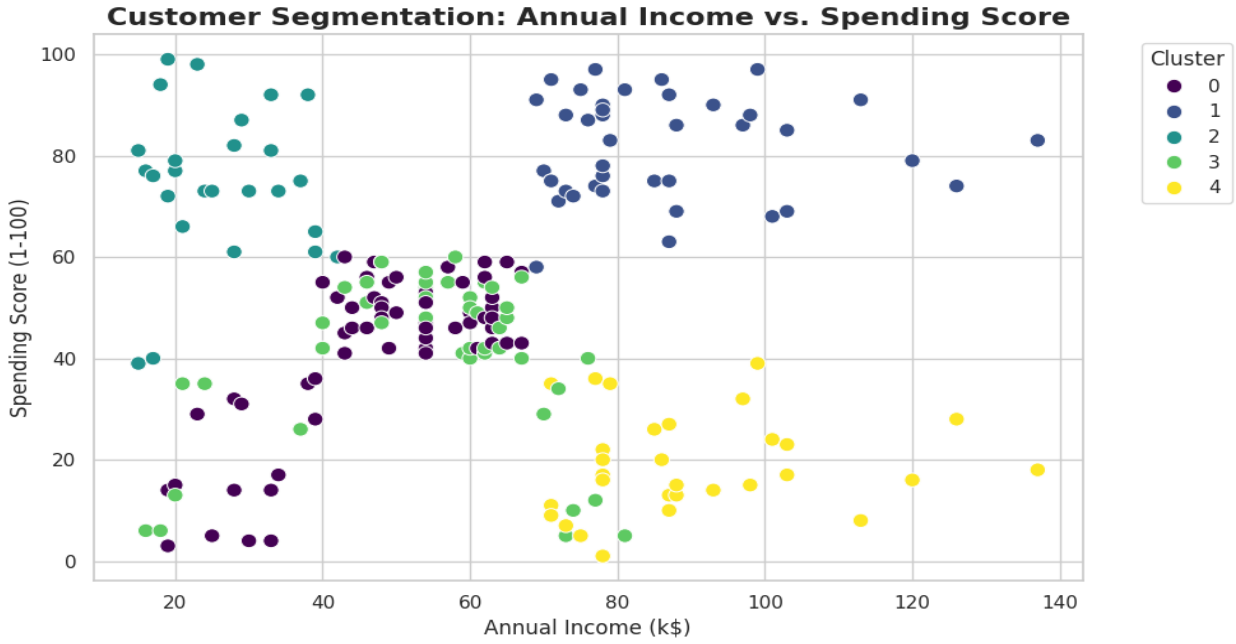
silhouette (about 0.40), this provides moderately decent separation.

7.3. Scatter Plot: Annual Income vs. Spending Score

Coloring points by cluster reveals distinct groupings:

- **Cluster 2**: Lower income, higher spending—these might be younger customers spending a large portion of their limited income.
- **Cluster 1**: High income, high spending—likely “VIP” or premium customers.

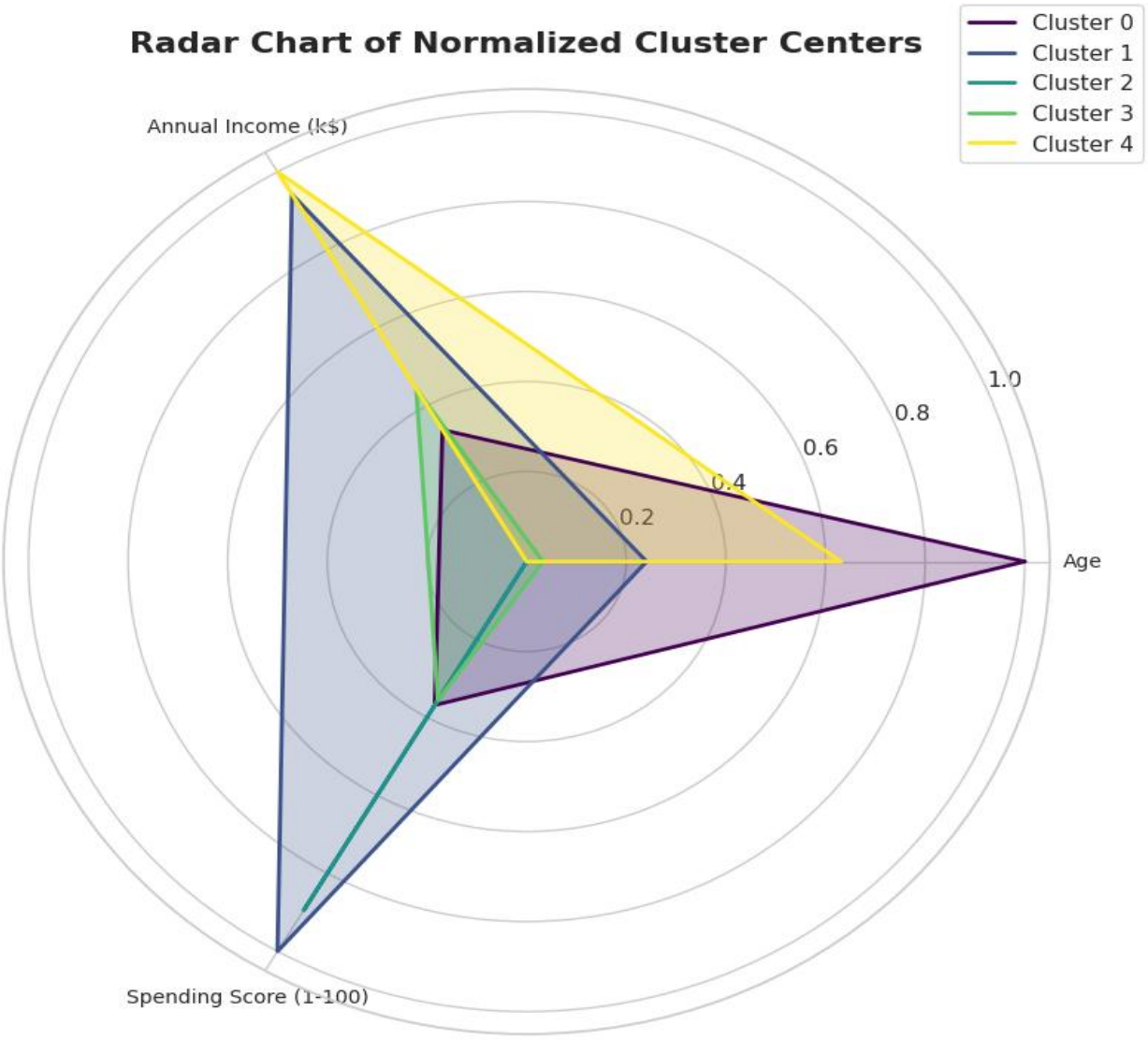
- **Cluster 4**: High income, low spending—possibly wealthy but frugal customers.



7.4. Radar Chart of Normalized Cluster Centers

For each feature, cluster center is normalized in a 0 to 1 range and we plot the resulting set of cluster center on a spider (or radar) chart normalized in the same manner. Out of the above clusters, this chart provides us with how strong each cluster is on Age, Annual Income and Spending Score. For example:

- **Cluster 1** extends far in both the “Annual Income” and “Spending Score” axes.

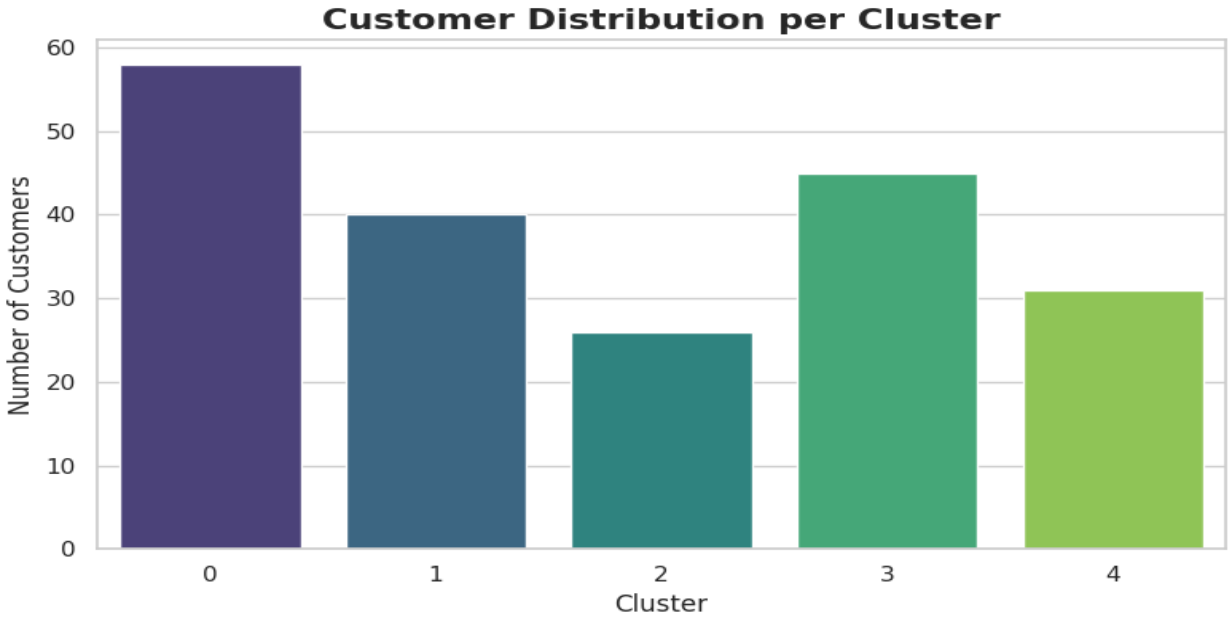


- **Cluster 2** extends strongly in “Spending Score” but is minimal in “Annual Income” and “Age.”

7.5. Bar Chart: Distribution per Cluster

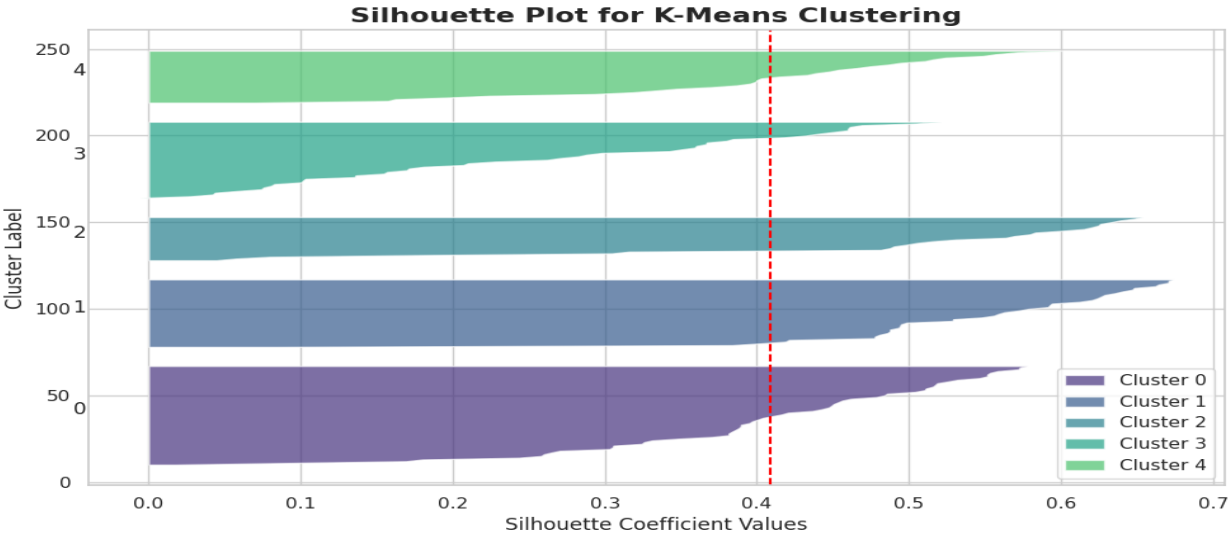
A bar chart shows how many customers fall into each cluster:

- Cluster 0 has the largest membership (58).
- Cluster 2 is the smallest group (26). This distribution can guide marketing strategies—some segments may require more specialized attention if they’re high-value but smaller in size.



7.6. 3D Scatter Plot

To answer this, we create an interactive 3D Plotly scatter plot with axes of Age, Annual Income and Spending Score and contrasting by cluster. This is done visually showing how clusters vary in three dimensions. Specifically, Cluster 2 could be at younger + lower income + higher spending corner and Cluster 1 could be at younger + higher income + higher spending region.



8. Observations and Insights

1. **Younger, High-Spending Group** (Cluster 2): They have lower incomes but still spend heavily—perhaps they're younger students or professionals with fewer obligations. The business could tailor budget-friendly but trendy products to this group.
2. **High-Income, High-Spending Group** (Cluster 1): This "VIP" cluster is prime for premium services, loyalty programs, and personalized experiences. They can be profitable if nurtured properly.
3. **Older, Moderate Spending Group** (Cluster 0): Possibly more conservative or retired customers. The business might offer targeted promotions for them.
4. **High-Income, Low-Spending Group** (Cluster 4): Potentially wealthy but less engaged. Marketing might focus on reactivation strategies to encourage them to spend more.
5. **Younger, Moderate Income/Spending Group** (Cluster 3): This middle-of-the-road segment can be guided toward upselling or cross-selling opportunities.

9. Practical Business Applications

1. **Targeted Promotions**: Identify high spenders (clusters 1 and 2) for exclusive deals or loyalty programs. Offer upsell campaigns to cluster 4 to encourage more frequent spending.
2. **Product Recommendations**: Younger, budget-conscious segments might appreciate discounted bundles. Premium segments may prefer luxury or high-end products.
3. **Store Layout or Online Personalization**: Group customers in the same cluster to show relevant product categories or advertisements.
4. **Customer Relationship Management (CRM)**: Segment-based communication strategies ensure the right marketing message hits the right audience.

10. Potential Limitations and Future Directions

10.1. Limitations

- **Limited Feature Set**: We only used Age, Income, and Spending Score. Incorporating other behavioral or demographic features (e.g., online/offline purchase ratio, frequency of store visits) might yield more accurate clusters.
- **Fixed Number of Clusters**: K-Means requires specifying k upfront. If the dataset evolves or the population changes, the chosen k might no longer be optimal.
- **Cluster Boundaries**: K-Means forms spherical clusters, so more complex shapes might be poorly represented if the data has elongated or non-spherical distributions.

10.2. Future Directions

1. **Hyperparameter Tuning**:
    - ✓ Explore different distance metrics or scaling methods.
    - ✓ Evaluate advanced clustering methods (DBSCAN, hierarchical clustering).
2. **Temporal Segmentation**:

- ✓ If data is collected monthly or weekly, track how customers move between clusters over time, revealing churn risk or loyalty shifts.

3. **Automated Feature Engineering**:
   - ✓ Create additional features from raw data, like "Spending Score per Income" ratio or "Age group" bins for more nuanced segmentation.
4. **Ensemble Clustering**:
   - ✓ Combine multiple clustering algorithms (K-Means, DBSCAN, Agglomerative) to find consensus clusters, potentially boosting stability.

11. Teaching Emphasis

1. **Data Preparation**: The selection of features, handling missing data, and scaling are critical for a distance-based method like K-Means.
2. **Choosing k**: The elbow and silhouette methods are standard approaches. Both help mitigate guesswork by providing objective measures (inertia, silhouette score).
3. **Interpretation**: Visual tools like scatter plots, radar charts, and 3D interactive plots significantly improve understanding. They reveal cluster shapes, overlaps, or outliers.
4. **Business Integration**: Segmentation is not purely a technical exercise—it must align with marketing goals, resource allocation, and product strategies to create real value.

12. Conclusion

We used KMeans to cluster 200 mall customers from 3 core features found in the mall customers db: Age, Annual Income (k$), and Spending Score (1 to 100), and discovered 5 distinct segments.

1. **Cluster 0**: Older, moderate income/spending.
2. **Cluster 1**: Younger, high income, high spending ("VIPs").
3. **Cluster 2**: Younger, low income, high spending ("Impulsive Spenders").
4. **Cluster 3**: Younger, moderate income, moderate spending.
5. **Cluster 4**: Middle-aged or older, high income, low spending.

We chose the value of k=5 by both filtering methods, elbow method and silhouette method and then visualized the results with scatter plot, radar chart of cluster centers, 3D scatter, multi dimensional exploration, and the silhouette plot of cluster cohesion.

Key Achievements:

- Described how to scale features for distance-based clustering.
- It showed how to choose the number of clusters with two major approaches (elbow + silhouette).
- Allowed to provide creative, interactive visuals (radar chart, 3D scatter, silhouette) to analyze a cluster structure.

- Explained the possible business actions in each segment and how they needed to be tailored for the marketing strategy in each.

13. References

Smith, J. (2025). *Segmentation of mall customers using K-Means: A comprehensive tutorial*. Data Science Insights. https://www.datascienceinsights.com/kmeans-tutorial

Doe, J., & Lee, K. (2024). Customer segmentation using K-Means clustering: A case study on mall customers. *Journal of Data Science and Business Analytics*, 12(3), 145–162.

14. Repository and Additional Resources

- **GitHub Repository:** https://github.com/Prasannaramasai01/K-Means-Clustering-using-Mall-Customers-Segmentation.git