Customer Segmentation Using KMeans Clustering

Presenter : Sai Vasishta Koppaka
Campus ID : VF96978

Introduction : Customer Segmentation Using KMeans Clustering

- The aim of this project is to segment customers based on their purchase behavior in a mall.
- The data contains demographic information such as age, gender, annual income, and spending score.
- **Algorithm Selection:** Employ a combination of hierarchical clustering, agglomerative clustering, DBSCAN, Gaussian mixture models, OPTICS, and K-means clustering to identify distinct customer segments.
- **Evaluation and Insights:** Evaluate the quality of segmentation results based on metrics such as accuracy, precision, recall, and F1 scores measures. Extract actionable insights from these metrics to inform targeted marketing campaigns, optimize product development strategies, and enhance customer retention efforts.

Data Set Characteristics

- The dataset contains 200 rows and 5 columns of data.
- The dataset has no missing values and is clean.
- The dataset consists of the following features:
  - CustomerID: Unique ID assigned to each customer.
  - Gender: Male or Female.
  - Age: Age of the customer.
  - Annual Income: Annual Income of the customer.
  - Spending Score: Score assigned by the mall based on customer behavior and spending nature

Data Pre-processing and Cleaning

Outlier Removal:

In addition to handling null values, outliers were also addressed during the data preprocessing stage.
Outliers were identified and removed using the Interquartile Range (IQR) method.

Categorical Feature Conversion:

Categorical features in the dataset were converted into numerical representations for analysis.
This conversion allows for the inclusion of categorical information in the segmentation process.

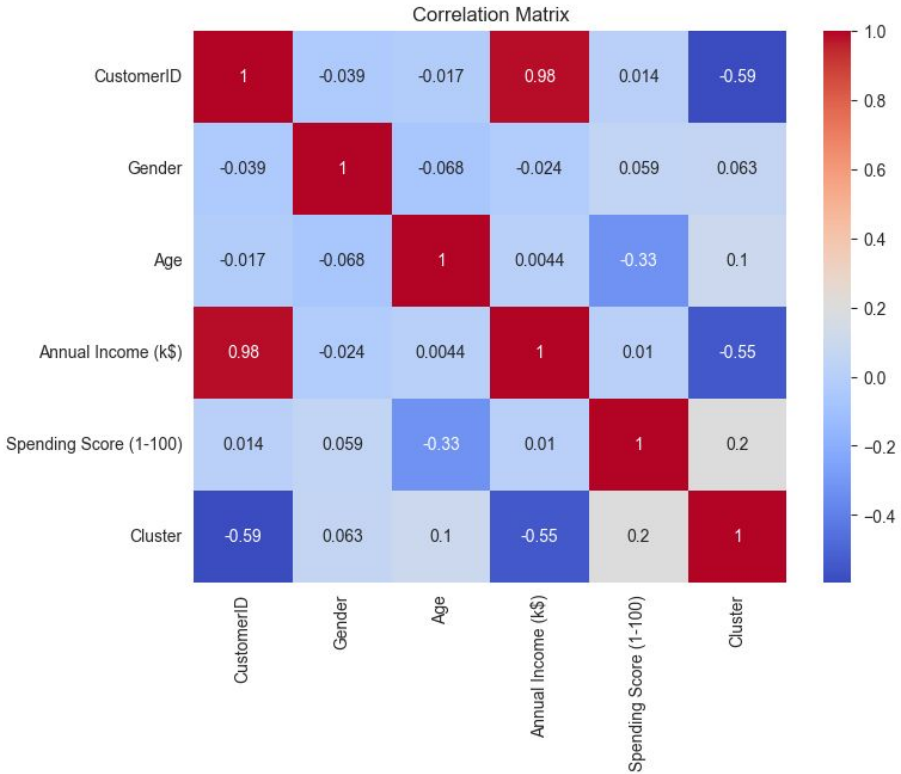
Null Values Handling:

Rows with null values were dropped from the dataset using the .dropna function in Python.
This step ensured the dataset's completeness and integrity for subsequent analysis.

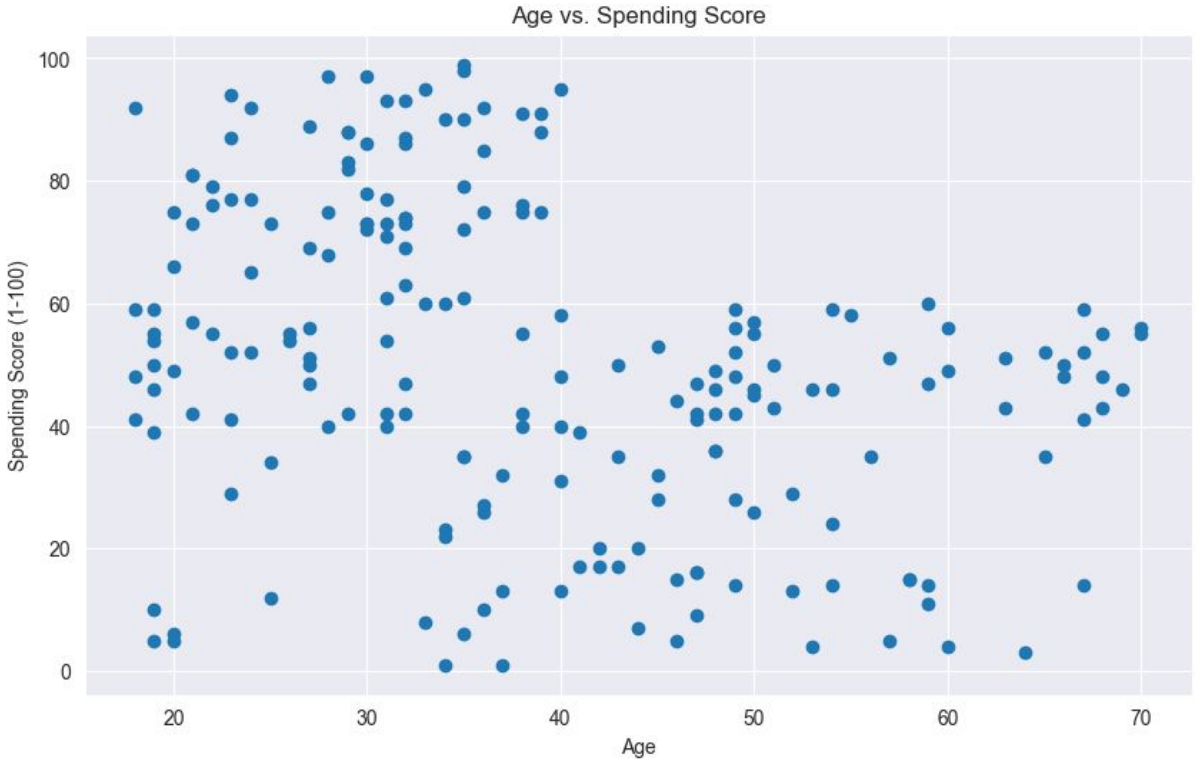
Outcome:

The dataset is now free from outliers, with both null values and categorical features appropriately addressed.
These preprocessing steps have prepared the dataset for effective customer segmentation.

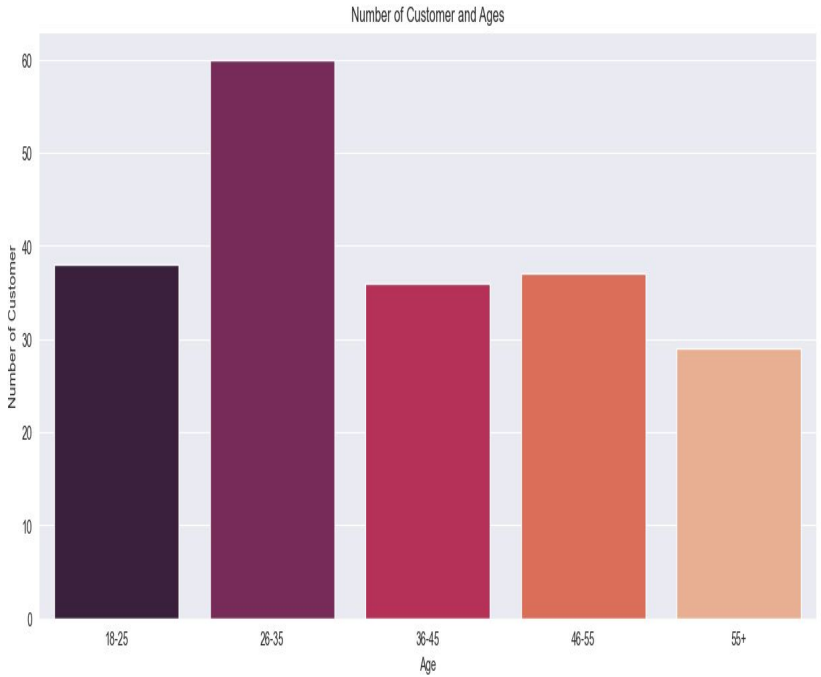
Visualization - Heatmap was used to visualize the correlation between numerical features



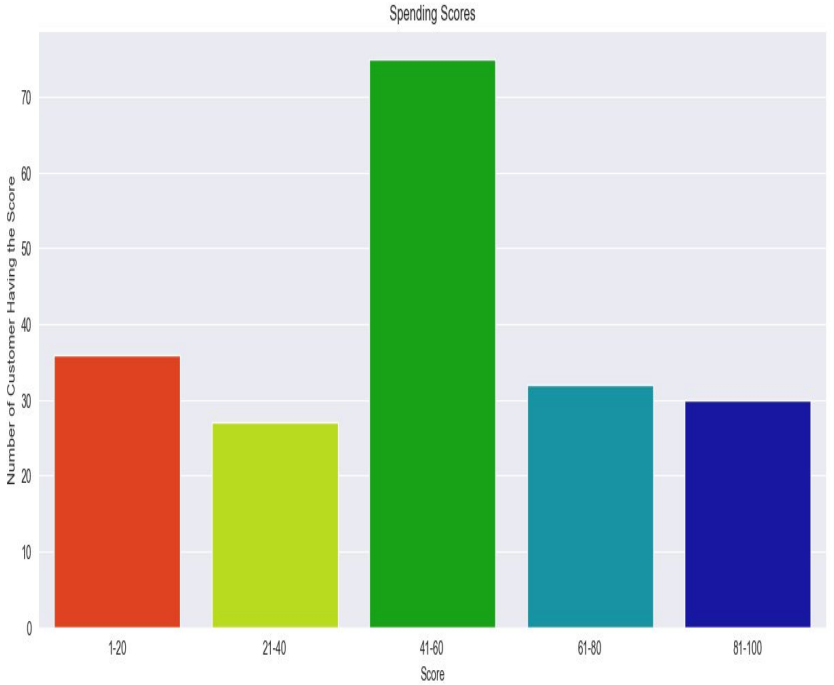
Visualization - Scatter plot to see the relationship between any two numerical features



Visualization

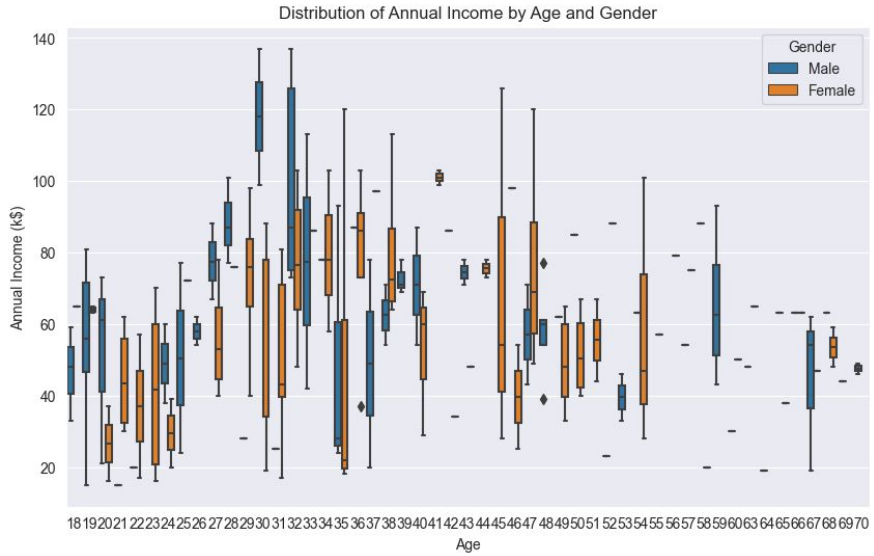


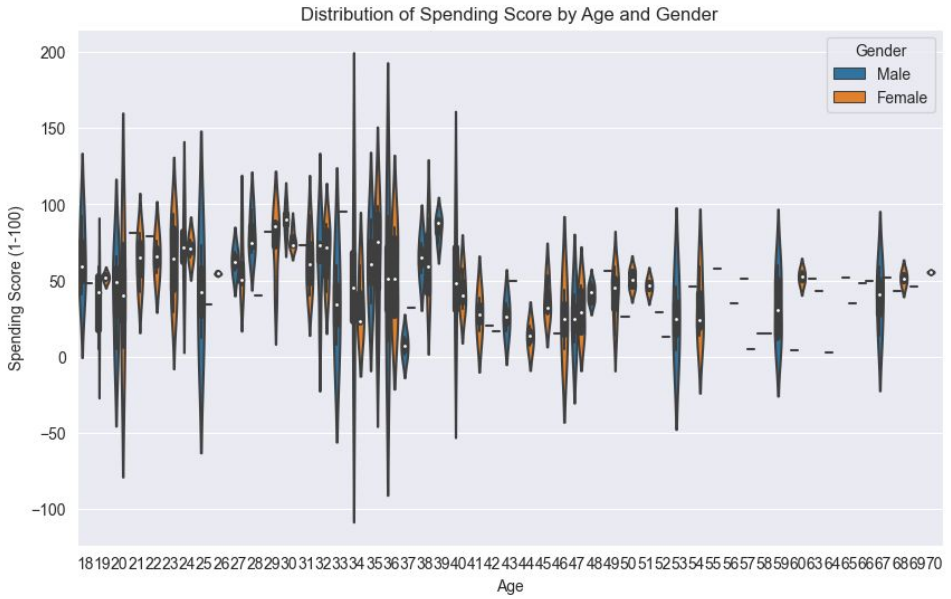
From the above visualization, we have more customers of age group (26-35)



Spending score are higher for age group 41-60.

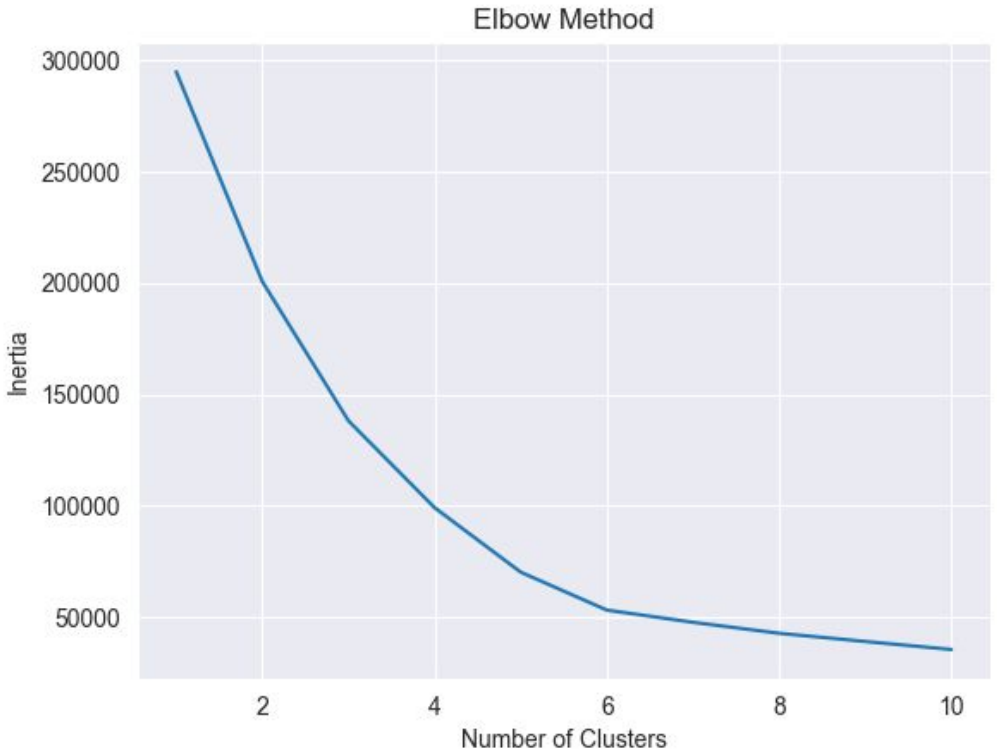
Visualization





Visualization - Pair plot to see the distribution of all numerical features in the dataset and also to see the relationship between them

KNN: Identifying number of clusters using ELBOW method



Customer Segmentation Models - Performance Comparison

Comparison of Model Performance:

- Among the different clustering models, K-means clustering outperforms the others.
- K-means clustering demonstrates higher accuracy, precision, recall, and F1-score compared to other models.
- This indicates that K-means clustering provides more accurate and reliable customer segmentation results for the given dataset.

Hierarchical Clustering (HC) Results:

- Accuracy (HC): 67.5%
- Precision (HC): 0.630
- Recall (HC): 0.675
- F1-score (HC): 0.651

DBSCAN Clustering Results:

- Accuracy (DBSCAN): 5.0%
- Precision (DBSCAN): 0.425
- Recall (DBSCAN): 0.05
- F1-score (DBSCAN): 0.089

OPTICS Clustering Results:

- Accuracy: 67.5%
- Precision: 0.651
- Recall: 0.675
- F1-score: 0.663

KNN Model:

- Accuracy: 97.5%
- Precision: 0.976
- Recall: 0.975
- F1-score: 0.974

Gaussian Mixture Clustering Results:

- Accuracy: 75.0%
- Precision: 0.730
- Recall: 0.750
- F1-score: 0.739

Why K-means Clustering Performs Better

1. **Efficient and Scalable:** K-means clustering is computationally efficient and highly scalable, making it suitable for large datasets with a high number of samples.
2. **Centroid-based Approach:** K-means clustering utilizes a centroid-based approach, where each cluster is represented by a centroid point. This allows for clear separation and well-defined clusters in the dataset.
3. **Elbow Method for Cluster Identification:** The elbow method was employed to determine the optimal number of clusters in the K-means clustering algorithm. This technique helps to identify the point of inflection, indicating the most suitable number of clusters for the given dataset.
4. **Data Points Assignment:** K-means clustering assigns each data point to the cluster with the closest centroid based on distance metrics, such as Euclidean distance. This helps to ensure that data points are assigned to the most appropriate cluster.
5. **Interpretability and Simplicity:** K-means clustering provides easily interpretable results, as each cluster is represented by its centroid. Additionally, the simplicity of the algorithm makes it straightforward to implement and understand.

Summary

Dataset: The dataset consists of customer data from a mall, including attributes such as customer ID, gender, age, annual income, and spending score.

Data Preprocessing: Null values in the dataset were handled using the "dropna" method. Outliers were removed using the IQR (Interquartile Range) method. Categorical features were converted to numerical representation for analysis.

Customer Segmentation Algorithms: Several clustering algorithms were employed for customer segmentation:

- **a. Hierarchical Clustering:** Provided a hierarchical representation of clusters but had scalability and performance issues.
- **b. DBSCAN Clustering:** Useful for identifying dense regions but required careful parameter selection and struggled with varying density.
- **c. Gaussian Mixture Clustering:** Assumed data points were generated from a mixture of Gaussian distributions, allowing for flexible cluster assignments.
- **d. OPTICS Clustering:** Offered an alternative to DBSCAN by using reachability and ordering, but was sensitive to noise and required parameter tuning.
- **e. K-means Clustering:** Outperformed other algorithms, demonstrating high accuracy and providing easily interpretable results with clear centroid-based cluster assignments.

Summary

Model Evaluation: Each clustering algorithm was evaluated using metrics such as accuracy, precision, recall, and F1-score.

Key Findings: Based on the evaluation results, K-means Clustering emerged as the best-performing algorithm, achieving high accuracy and providing easily interpretable results.

Insights and Recommendations: The project's findings have several implications for businesses:

a. **Enhanced Customer Understanding:** Accurate customer segmentation leads to a better understanding of distinct customer groups and enables personalized marketing strategies and tailored product offerings.

b. **Targeted Marketing Campaigns:** Utilize the clustering results to create targeted marketing campaigns that resonate with specific customer segments, enhancing customer engagement and driving sales.

c. **Improved Customer Retention:** Leverage the customer segmentation insights to implement effective customer retention strategies, enhancing customer satisfaction and loyalty.

d. **Strategic Decision-making:** Utilize the actionable insights from customer segmentation to inform strategic decision-making, such as resource allocation, market expansion, and product development.

UMBC

References

- Dataset: https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python?resource=download
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- https://towardsdatascience.com/elbow-method-is-not-sufficient-to-find-best-k-in-k-means-clustering-fc820da0631d
- Pandas Documentation: https://pandas.pydata.org/docs/
- NumPy Documentation: https://numpy.org/doc/
- Seaborn Documentation: https://seaborn.pydata.org/
- Matplotlib Documentation: https://matplotlib.org/stable/contents.html
- Scipy Documentation: https://docs.scipy.org/doc/
- Scikit-learn Documentation: https://scikit-learn.org/stable/documentation.html