INTELLIHACK 5.0

MACHINE LEARNING HACKATHON

Q2. Customer Segmentation

The Code Rushers

1. Introduction

Customer segmentation is a critical process in understanding customer behavior and tailoring marketing strategies to different groups. This report focuses on segmenting customers of an e-commerce platform based on their behavior, using clustering techniques. The dataset contains six features: **customer_id**, `total_purchases`, `avg_cart_value`, `total_time_spent`, `product_click`, and `discount_count`. The goal is to identify three distinct customer segments: **Bargain Hunters**, **High Spenders**, and **Window Shoppers**.

2. Approach Taken

The following steps were taken to achieve the goal:

1. **Exploratory Data Analysis (EDA):** Understand the dataset, check for missing values, and visualize distributions and relationships.
2. **Model Selection:** Apply clustering algorithms (K-Means, Agglomerative Clustering) to identify customer segments.
3. **Model Evaluation:** Use metrics like Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score to evaluate the models.
4. **Model Performance Visualization:** Use PCA to reduce dimensionality and visualize the clusters in 2D.
5. **Cluster Analysis:** Interpret the clusters and assign customer types based on their characteristics.

3. Exploratory Data Analysis

Exploratory data analysis (EDA) is a critical initial step in the machine learning workflow. It involves using Python libraries to inspect, summarize, and visualize data to uncover trends, patterns, and relationships. Here's a breakdown of the key steps in performing EDA with Python:

1. Importing Libraries

The following Python libraries were used for EDA:

- `pandas (pd)`: For data manipulation and analysis.
- `scikit-learn (sklearn)`: For an efficient clustering and analysis tasks
- `numPy (np)`: For numerical computations.
- `matplotlib.pyplot (plt)`: For basic plotting functionalities.
- `seaborn (sns)`: A high-level visualization library built on top of Matplotlib.

2. Loading the Data

The dataset was loaded using pandas from csv file to a Dataframe:

```
df_init = pd.read_csv('customer_behavior_analytcis.csv')
```

3. Initial Inspection

The dataset was initially inspected using the following methods:

- `df.info()`: Provided an overview of the dataset, including column names, data types, and non-null counts.
- `df.head()`: Displayed the first few rows of the dataset.
- `df.tail()`: Displayed the last few rows of the dataset.
- `df.describe()`: Summarized the statistical properties of a DataFrame
- `df.dtypes`: Checked the data types of each column.

2. Data Reduction

The last column named **customer_id** dropped beccause it doesn't add value to our analysis:

```
df = df_init.drop(columns=['customer_id'])
```

5. Data Cleaning

The dataset was cleaned to handle missing values and duplicates:

- Missing values were identified using `df.isnull().sum()`.

  We encountered 20 missing values in columns `total_purchases`, `avg_cart_value` and `product_click`.
- Missing values were filled with the mean of each feature using `df.fillna(df.mean())`.
- Duplicates were checked using `df.duplicated().sum()`.

  No duplicates found.

6. Univariate Analysis

Univariate analysis was performed to understand the distribution of individual features:

- Histograms and box plots were created to visualize the distribution of numerical features.
- Density plots were used to analyze the shape of distributions.

Histogram Analysis

Figure 1 shows the histograms for key features such as *total_purchases, avg_cart_value, total_time_spent, product_click*, and *discount_counts*. From the plots, we observe:

- Most variables are **right-skewed**, indicating that the majority of users have lower values while a few have significantly high values.
- The *product_click* feature exhibits multiple peaks, indicating **possible clusters** in user behavior.

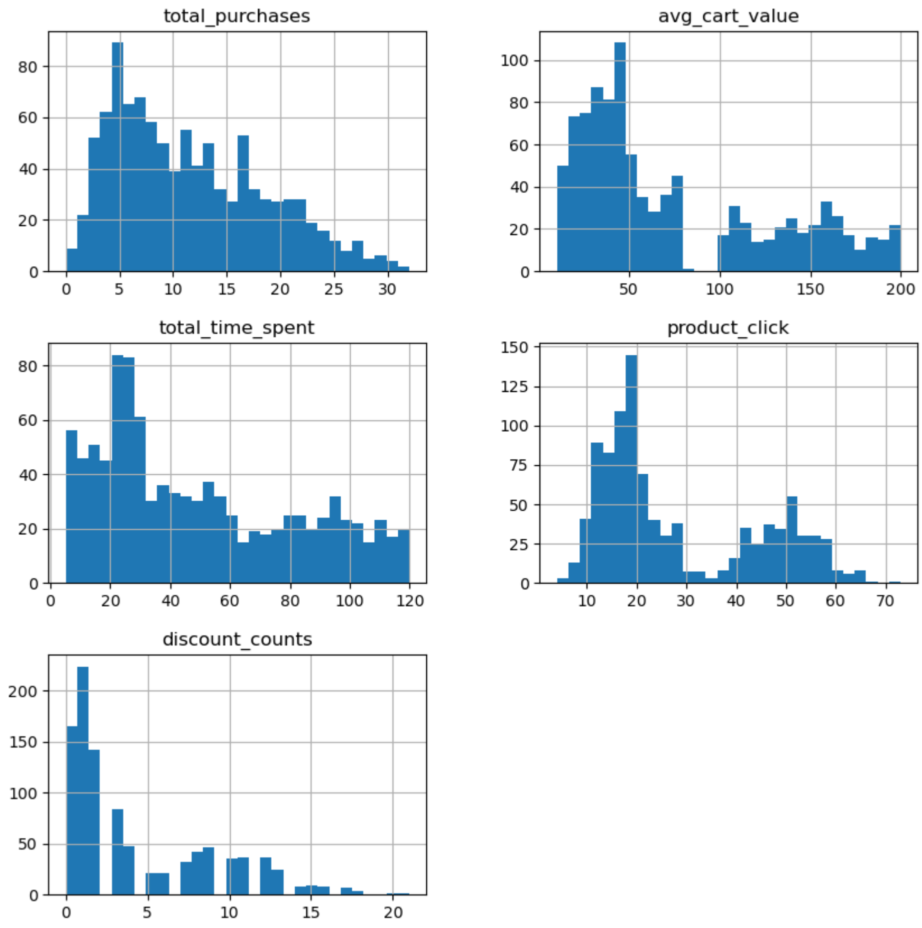

Figure 1: Histograms of Key Features

Box Plot Analysis

Figure 2 illustrates the box plots of the dataset. Key observations include:

- Significant **outliers** are present in almost all features, particularly in *avg_cart_value* and *total_time_spent*, which suggests that a small percentage of users exhibit extreme behavior.

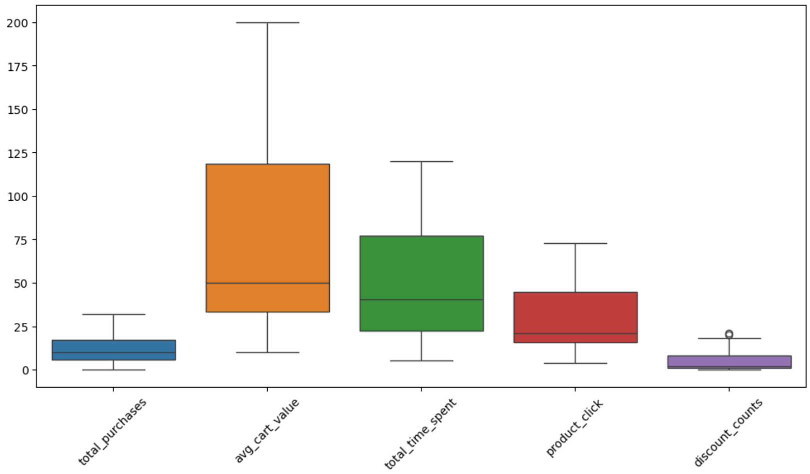

Figure 2: Box Plot Analysis of Key Features

Density Plot Analysis

The density plots in Figure 3 highlight:

- Multiple peaks in *product_click* and *total_purchases*, which could indicate natural segmentation in user behavior.

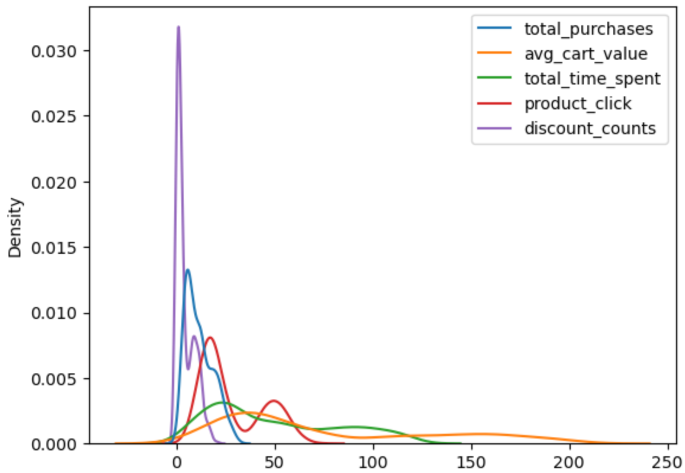

Figure 3: Density Plots of Key Features

Overall Key Insights from Univariate Analysis

- The dataset exhibits **skewness and outliers**, suggesting that some users behave significantly differently from the majority.
- The presence of multiple peaks in certain features indicates the possibility of **natural segmentation**, making clustering a suitable approach.
- **Outliers** should be handled appropriately to prevent bias in clustering results.

7. Bivariate Analysis

Bivariate analysis was conducted to explore relationships between pairs of features.

- Scatter plots were used to identify trends and potential correlations.
- Pair plots were created to visualize relationships between numerical features.
- Correlation heatmaps were created to visualize correlation matrices.

Scatter Plot Analysis

Figure 4 presents scatter plots for key feature pairs, helping us identify relationships and clustering patterns.

- There is a **non-linear relationship** between *total_purchases* and *avg_cart_value*, indicating that higher cart values do not always translate to more purchases.
- Distinct clusters are observed in *product_click* vs. *avg_cart_value*, suggesting user segmentation based on engagement and spending behavior.

Figure 4: Scatter Plot Analysis of Feature Pairs

Pair Plot Analysis

The pair plot visualization in Figure 5 provides insights into the distribution and interactions between multiple features simultaneously.

- *Total_purchases* and *total_time_spent* exhibit a extbfnegative correlation, indicating that users who spend less time tend to make more purchases quickly.
- The relationship between *discount_counts* and *avg_cart_value* is **weak**, suggesting that discounts may not significantly influence higher spending users.
- Some feature pairs, such as *product_click* and *discount_counts*, show extbfscattered clustering, reinforcing potential user segmentation.

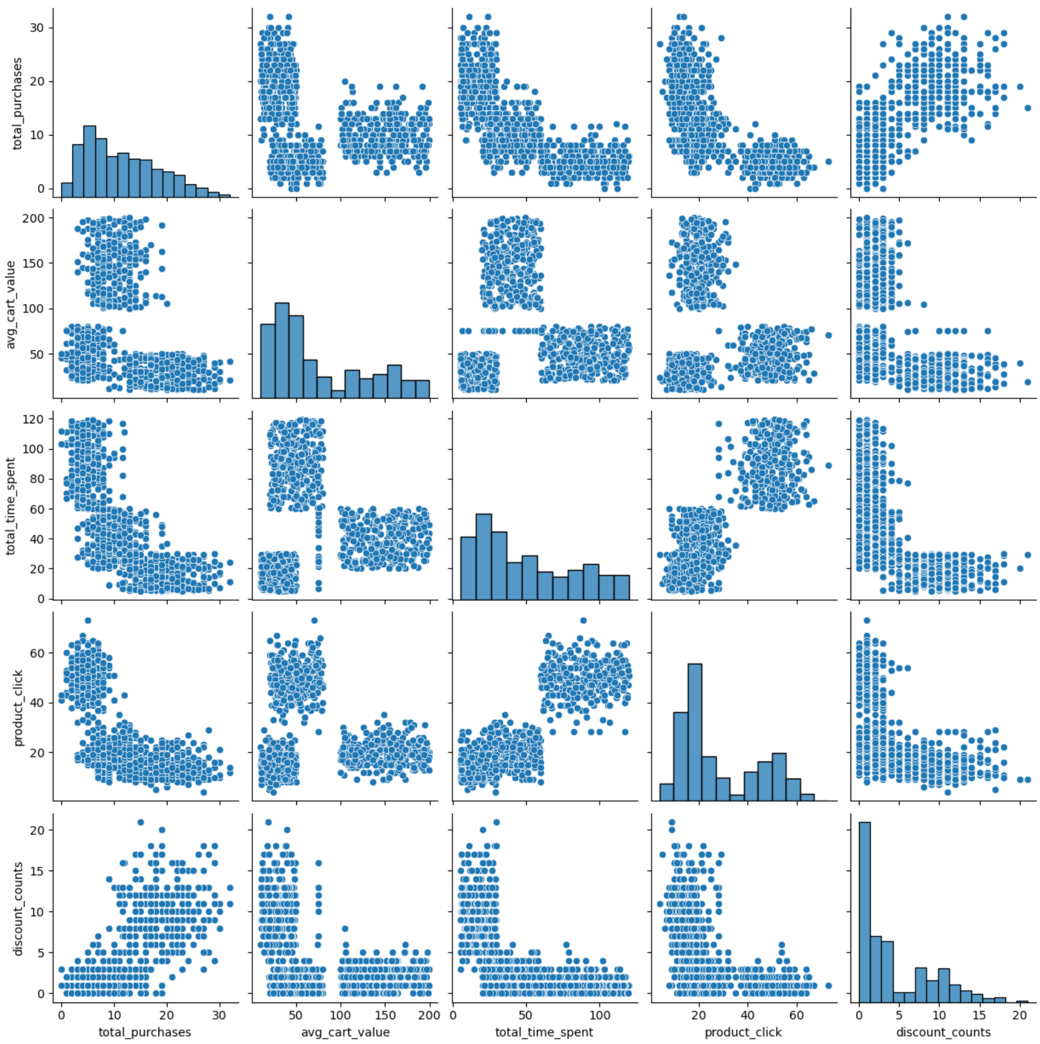

Figure 5: Pair Plot Analysis of Key Features

Correlation Map Analysis

The correlation heatmap (Figure 6) provides an overall view of feature relationships.

- *Total_time_spent* and *product_click* have a extbfstrong positive correlation, reinforcing the pattern observed in scatter plots.
- *Total_purchases* and *discount_counts* have a extbfweak correlation, meaning discount usage does not significantly affect the number of purchases.
- Features with moderate correlation indicate potential extbfmulti-dimensional clustering patterns.

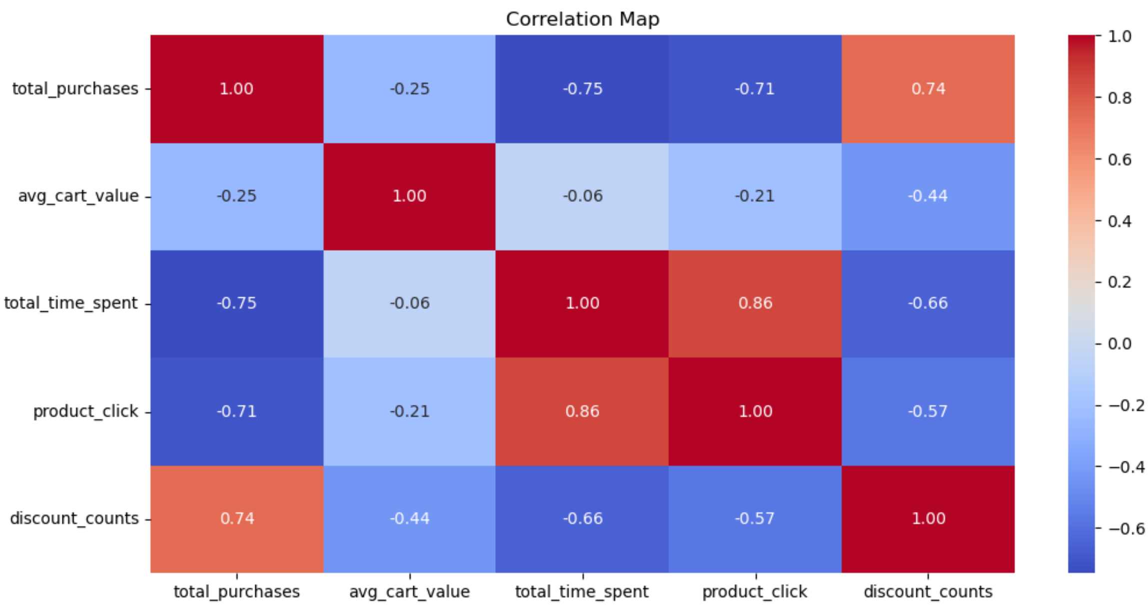

Figure 6: Correlation Matrix of Key Features

Overall Key Insights from Bivariate Analysis

- Some features exhibit **strong correlations**, which can be used as primary attributes for clustering.
- The presence of **scattered clusters** in scatter plots suggests extbfnatural segmentation among users.
- Relationships between engagement metrics and spending behavior provide valuable insights for **personalized marketing and recommendation systems**.

4. Model Selection

4.1 K-Means Clustering

- K-Means was chosen due to its simplicity, efficiency and after **evaluation scores**.
- The number of clusters (`K=3`) was verfied by using the **Elbow Method**.
  - The plot of inertia vs. number of clusters showed a clear elbow at `K=3`, confirming the optimal number of clusters.

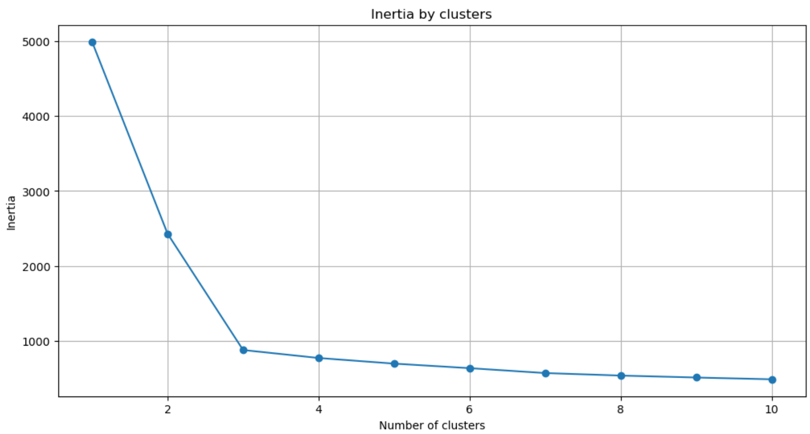

Figure 7: Elbow Method Plot

Why we choosed **K-Means Clustering** Model:

- Handles Numerical & Continuous Data Effectively
- Scalability for Large Datasets
- Clear & Distinct Customer Groups
- Straightforward Implementation & Interpretation
- Works Well with Feature Scaling
- Optimizable with the Elbow & Silhouette Methods

4.2 Agglomerative Clustering

- Agglomerative Clustering was also applied for comparison.
- The same number of clusters (`K=3`) was used.

5. Model Evaluation

5.1 Silhouette Score

- **K-Means:** Silhouette Score = 0.61329 (Well separation between clusters).
- **Agglomerative Clustering:** Silhouette Score = 0.61212 (slightly worse than K-Means).

5.2 Davies-Bouldin Score

- **K-Means:** Davies-Bouldin Score = 0.56779 (lower is better, indicating well separation).
- **Agglomerative Clustering:** Davies-Bouldin Score = 0.56768

5.3 Calinski-Harabasz Score

- **K-Means:** Calinski-Harabasz Score = 2341.25 (higher is better, indicating dense and very well-separated clusters).
- **Agglomerative Clustering:** Calinski-Harabasz Score = 2327.75

6. insights gained from model evaluation

1. **Silhouette Score (Higher is Better)**
   Both models perform similarly, but K-Means slightly outperforms Agglomerative Clustering in terms of well-separated clusters.
2. **Davies-Bouldin Score (Lower is Better)**
   Both models have almost identical Davies-Bouldin Scores, indicating that the clusters are well-separated in both cases.
3. **Calinski-Harabasz Score (Higher is Better)**
   K-Means has a slightly higher Calinski-Harabasz Score, suggesting that its clusters are more compact and well-separated compared to Agglomerative Clustering.

- K-Means is the better choice for customer segmentation based on the evaluation metrics. It produces well-separated, compact clusters with slightly better performance than Agglomerative Clustering across all three metrics. Additionally, K-Means is computationally efficient, making it more scalable for large datasets.

7. Cluster Analysis

7.1 Cluster Characteristics

- **Cluster 0 (Window Shoppers):**
  - Low `total_purchases`.
  - Moderate `avg_cart_value`.
  - High `total_time_spent` and `product_click`.
  - Low `discount_count`.
- **Cluster 1 (Bargain Hunters):**
  - High `total_purchases`.
  - Low `avg_cart_value`.
  - Moderate `total_time_spent` and `product_click`.
  - High `discount_count`.
- **Cluster 2 (High Spenders):**
  - Moderate `total_purchases`.
  - High `avg_cart_value`.
  - Moderate `total_time_spent` and `product_click`.
  - Low `discount_count`.

7.2 Customer Type Mapping

- Customers were mapped to segments based on cluster characteristics:
  - **Cluster 0:** Window Shoppers.
  - **Cluster 1:** Bargain Hunters.
  - **Cluster 2:** High Spenders.

8. Visualization

8.1 PCA for Dimensionality Reduction

- PCA was used to reduce the data to 2 dimensions for visualization.
- A scatter plot of the first two principal components showed clear separation between clusters.

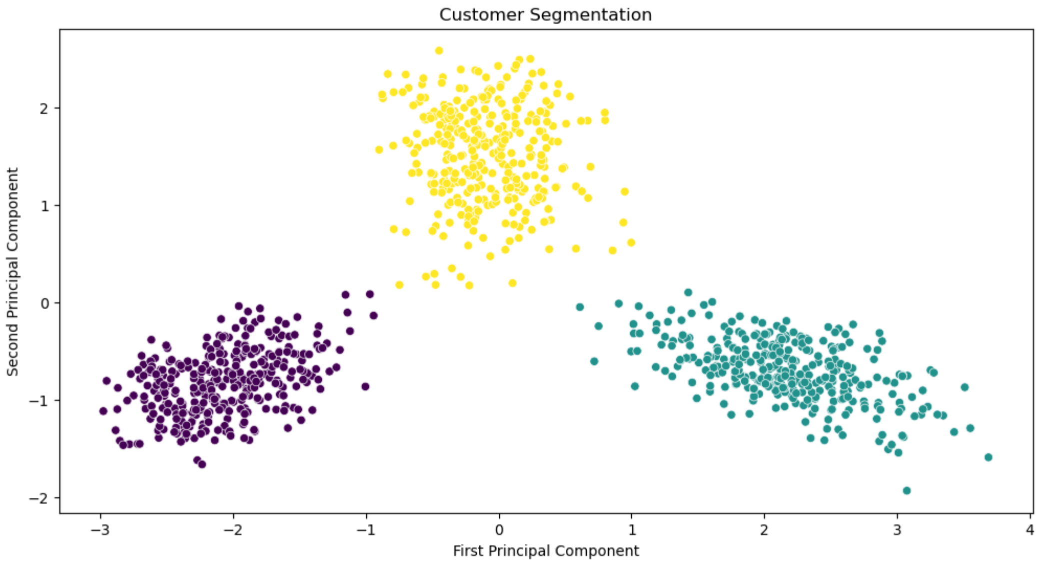

Figure 8: PCA Plot

9. Challenges Faced

1. **Outliers:** Outliers in `total_time_spent` and `avg_cart_value` affected clustering.
2. **Feature Scaling:** Ensuring all features were on the same scale was crucial for accurate clustering.
3. **Cluster Interpretation:** Assigning meaningful labels to clusters required careful analysis of feature means.
4. **Model Selection:** Since K-Means and Agglomerative Clustering clashes with similar scores, after spending so much time with analysis and research selected **K-means** as our model.

10. Suggestions for Improvement

1. **Feature Scaling and Engineering:** Create new features like `purchase_frequency` or `discount_usage_rate` to improve clustering.
2. **Outlier Handling:** Use robust scaling or remove outliers to improve model performance.
3. **Larger Dataset:** Use a larger dataset to validate the stability of the clusters.
4. **Use Different Distance Metrics** K-Means relies on Euclidean distance, which may not work well for high-dimensional data. Try Manhattan distance or cosine similarity if the dataset has sparse or categorical features

11. Conclusion

The analysis successfully identified three distinct customer segments: **Bargain Hunters**, **High Spenders**, and **Window Shoppers**. K-Means clustering provided the best results, with clear separation between clusters. The insights gained can help the e-commerce platform tailor marketing strategies to each segment, improving customer engagement and revenue.