CS 4048 – Data Science
Final Project
BCS7C-BSE7A

Instructions:

- **Maximum of 3 students per group are allowed**
- **Dataset (electronics.json) for the project is uploaded on the portal.**
- **The codes must be in running form.**
- **You can use built-in libraries.**
- **You need to preprocess the dataset before using it.**
- **You need to visualize your findings.**
- **Your approach to the problems will highly be seen.**
- **You must complete the project before the deadline (30th December, 2023)**
- **Late submissions won't be entertained.**
- **Plagiarism will not be entertained.**
- **AI generated codes will be dealt with strict actions.**

Introduction:

Imtiaz Mall, a renowned department store chain, is experiencing declining sales and a significant number of non-recurring customers in its electronics section. To address this challenge, you, the newly appointed Senior Data Scientist, have been tasked with conducting a comprehensive analysis of the electronics section data and developing data-driven strategies for customer retention and sales growth. This project focuses on the initial steps of this analysis, specifically exploring the data through various techniques and comparing the results of three clustering algorithms: K-Means, DBSCAN, and K-Means++.

Module 1: Data Acquisition and Preprocessing:

1. **Data Acquisition:**
    - Download the provided historical sales data for the electronics section.
    - Ensure the data includes customer demographics, purchase history, product details, spending amounts, and dates of transactions.
2. **Data Cleaning:**
    - Identify and handle missing values using appropriate techniques like mean/median imputation or dropping rows/columns with excessive missingness.
    - Analyze outliers and determine whether to retain or remove them based on their impact on the analysis.
    - Address inconsistencies in data format and encoding.
3. **Data Transformation:**

- Create new features that provide deeper insights into customer behavior, such as:
  - Average spending per purchase
  - Purchase frequency per month
  - Brand affinity score (based on product brand preferences)
  - Product category preferences (e.g., TVs, smartphones, laptops)
- Standardize or normalize numeric features to ensure they contribute equally to the clustering algorithms.

Module 2: Exploratory Data Analysis (EDA):

1. **Univariate Analysis:**
   - Analyze the distribution of key features like customer age, purchase amount, and purchase frequency using histograms, boxplots, and descriptive statistics.
   - Identify potential skewness or outliers in the data.
2. **Bivariate Analysis:**
   - Utilize scatterplots and heatmaps to explore relationships between different features, such as purchase amount vs. income level, brand affinity vs. product category, and purchase frequency vs. age.
   - Investigate the presence of correlations and identify any impactful relationships.
3. **Temporal Analysis:**
   - Analyze trends in customer behavior over time, including changes in purchase frequency, average spending, and product preferences.
   - Identify seasonal variations or any significant shifts in customer behavior patterns.

Module 3: Clustering Analysis:

This phase involves implementing and comparing three clustering algorithms: K-Means, DBSCAN, and K-Means++.

A. K-Means Clustering:

1. Define the number of clusters (k):
   - Analyze the elbow plot to determine the optimal number of clusters based on the sum of squared distances within each cluster.
   - Consider silhouette analysis to evaluate the quality of clusters formed at different k values.

2. Apply K-Means algorithm:
    - Implement K-Means with the chosen k value to segment customers into distinct clusters based on their purchase behavior and preferences.
3. Analyze cluster characteristics:

    Investigate the key features of each cluster, such as average purchase amount, brand affinity, and product category preferences.

    - Identify significant differences and similarities between the clusters.

B. DBSCAN Clustering:

1. Define eps and MinPts parameters:
    - Experiment with different values of eps (neighborhood radius) and MinPts (minimum number of points within eps) to identify the best configuration for clustering.
    - Utilize silhouette analysis or other cluster quality metrics to evaluate different parameter combinations.
2. Apply DBSCAN algorithm:
    - Implement DBSCAN with the chosen parameters to discover clusters of customers based on their density and spatial distribution.
3. Analyze cluster characteristics:
    - Explore the characteristics of each cluster, including size, density, and distribution within the data space.
    - Compare the clusters formed by DBSCAN with those identified by K-Means.

C. K-Means++ Clustering:

1. Apply K-Means++ algorithm:
    - Implement K-Means++ with the desired number of clusters to initialize centroids strategically for improved clustering performance.
2. Compare results to K-Means:
    - Analyze the resulting clusters formed by K-Means++ and compare them to those generated by regular K-Means.
    - Evaluate the effectiveness of K-Means++ in achieving better cluster quality and convergence speed.

Module 4: Comparison and Conclusion:

1. **Compare the results of all three clustering algorithms:**
    - Analyze the similarity and differences in the clusters formed by K-Means, DBSCAN, and K-Means++.
    - Evaluate the effectiveness of each algorithm in identifying distinct customer segments based on their purchase behavior and preferences.

- Consider metrics such as cluster silhouette score, Calinski-Harabasz score, and Davies-Bouldin index to compare the overall quality of clustering results.
- Discuss the advantages and disadvantages of each algorithm in the context of Imtiaz Mall's specific needs and data characteristics.

2. **Draw conclusions and recommendations:**
   - Based on the findings of the EDA and clustering analysis, provide insights into the customer segments within the electronics section.
   - Identify the key factors that differentiate the customer segments and explain their purchasing behavior patterns.
   - Recommend data-driven strategies for customer retention and sales growth based on the identified segments.
   - Suggest potential applications of the clustering results, such as personalized product recommendations, targeted marketing campaigns, and tailored loyalty programs.
   - Propose further analysis and investigations to enhance the understanding of customer behavior and optimize the performance of the electronics section.