**Project DM**

**Ali Gamal Mohamed Salem id:210318**

**Mohamed Gomaa Moustafa id:2103129**

**Marwan Khaled id:** **22030157**

**Introduction:**

The objective of this project is to analyze online retail data using data mining techniques. The dataset used in this project is sourced from an Excel file containing information about online retail transactions.

**Problem Statement:**

The primary problem addressed through data mining techniques in this project is to understand customer behavior and segment customers based on their purchasing patterns. Additionally, the project aims to identify outliers and canceled orders to improve data quality for further analysis.

**Screenshots of Code:**

**How the data looks:**

**A screenshot of a computer

Description automatically generated**

Awe have 541909 row and 8 columns**A screenshot of a computer

Description automatically generated**

**Data cleaning:**

**1)remove null values**

**A screenshot of a computer

Description automatically generated**

**2)Detect outliersA screenshot of a computer

Description automatically generated**

A screenshot of a computer program

Description automatically generated

**Plots:**

A screenshot of a graph

Description automatically generated

**A black rectangular object with a white border

Description automatically generated**

**A graph of a distribution of quantity

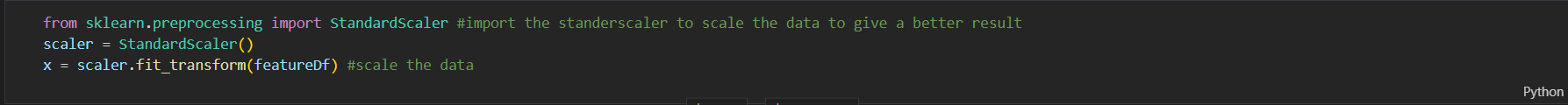
Description automatically generated**

**3) feature extraction:**

**A screenshot of a computer program

Description automatically generated**

**4) the model**

****

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A computer screen shot of a program

Description automatically generated

**5) the results**

A diagram of a clustering structure

Description automatically generated

A graph with a line

Description automatically generated

A graph with colored dots

Description automatically generated

**Conclusion of Project Results:**

1. **Data Cleaning:**
   * Missing values in the "Description" and "CustomerID" columns were handled by replacing missing descriptions with "missing" and removing rows with missing customer IDs.
   * Negative values in the "Quantity" column were investigated to differentiate between canceled orders and outliers. Canceled orders were identified and removed, while outliers were dropped.
2. **Feature Engineering:**
   * Total purchase amount was calculated by multiplying quantity and unit price.
   * Invoice date was transformed into separate features for year, month, day, hour, and minute.
   * Features including recency, frequency, and monetary were engineered for further analysis. These features provide valuable insights into customer behavior and purchasing patterns, enabling businesses to understand their customers better and tailor marketing strategies accordingly.
3. **Clustering Analysis:**
   * Hierarchical clustering and the Elbow method were utilized to determine the optimal number of clusters.
   * KMedoids clustering algorithm was applied with three clusters based on the Elbow method's recommendation.
   * Customer segments were visualized based on their purchasing behavior, providing insights for targeted marketing strategies.

**Conclusion:**

This project successfully analyzed online retail data, identified customer segments, and provided actionable insights for improving business strategies. By leveraging data mining techniques, businesses can better understand customer behavior and tailor marketing efforts to specific customer segments.