**DATA SCIENCE WITH PYSPARK**

EDA and Build Clustering Models for Customer Segmentation with PySpark and Spark

**Table of Contents**

[**1. Introduction**](#_1c43hyin7zbf) **4**

[1.1. Project Objectives](#_y4lxrmlke5fb) 4

[1.2. Scope of Study](#_ubhgl2ngg3hz) 4

[1.3. Project Significance](#_cyjy5oc2r8w2) 5

[**2. Requirement Analysis**](#_n0mqalac5nlk) **5**

[2.1. Functional Requirements](#_ls978wmayic3) 5

[2.2. Non-Functional Requirements](#_dweg5nvy1w1i) 6

[**3. Data Collection and Processing**](#_2sc17ldrk5b3) **6**

[3.1. Data Collection](#_4kdx3pqz1smv) 6

[Collection Methods: Using PySpark to connect and extract data from these sources. Sales data is exported as CSV from ERP, weather data is fetched via API, and market data is collected from converted PDF reports.](#_64tqtb7fx6h3) 6

[3.2. Data Processing with PySpark](#_61rc90y430wk) 6

[Data Cleaning:](#_8nqjdfo3oyov) 6

[Data Transformation:](#_7an2hcyzsc2j) 7

[**4. Data Exploration and Analysis**](#_po785fxl0aib) **7**

[4.1. Exploratory Data Analysis (EDA)](#_pbe2jieyq2dw) 7

[Visualization:](#_kjm09j5mcph5) 8

[4.2. Relationship Analysis Between Variables](#_nu1i4x6b0ytj) 9

[Correlation Analysis](#_py7eql3mba3k) 9

[Insights from the Correlation Analysis](#_hmj4w44y0kdw) 9

[Multivariate Analysis](#_gmpf86wbg80b) 10

[Summary of Insights from Multivariate Analysis](#_l1yp9zsgqqdo) 10

[**5. Feature Engineering**](#_pw0d2s2ct5or) **11**

[5.1. Creating New Features](#_2iqkajgkij5v) 11

[5.2. Feature Selection](#_4oy807nob7wn) 12

[**6. Machine Learning Model Development and Evaluation**](#_rs5fjhgwa4e1) **13**

[6.1. Splitting Data into Training and Testing Sets](#_zt6wlw5bma1) 13

[6.2. Model Building with MLlib](#_vp0w012uybsb) 13

[6.3. Model Evaluation](#_cfwfp074tpcs) 14

[**7. User Interface Deployment with Gradio**](#_uw6p55qwrlfl) **14**

[7.1. UI Design](#_1csexfo2acjb) 14

[7.2. Model Integration with Gradio](#_27fxtgqabsma) 14

[7.3. Result Visualization](#_ba3drg7uke4l) 14

[**8. Results and Discussion**](#_81iet3kdw4xc) **15**

[8.1. Achieved Results](#_7mnbq0tkfh15) 15

[8.2. Model Comparison](#_udt2jimdsbk1) 15

[8.3. Discussion of Results](#_busz2wkaky6h) 15

[**9. Conclusion and Future Development**](#_6zo6d3sm3tbs) **16**

[9.1. Achieved Results](#_7vl6ipkhx2by) 16

[9.2. Limitations](#_arv6ubxgmj4o) 16

[9.3. Future Development Directions](#_o2cj8ftm952z) 16

[**10. References**](#_nis6vw6dljq4) **17**

[**11. Conclusion**](#_ae735uod8gk) **17**

# 1. Introduction

## 1.1. Project Objectives

The objective of this e-commerce analysis project is to gain a deep understanding of customer behavior, including purchase patterns, spending habits, and transaction frequency. By segmenting customers based on factors such as buying behavior, location, and spending trends, the project aims to identify high-value customers and optimize marketing strategies. Leveraging these insights, businesses can develop targeted marketing campaigns, personalized promotions, and dynamic pricing models to enhance customer engagement and drive sales. Additionally, the analysis will help improve product offerings by identifying bestsellers and underperforming items, aiding in inventory management and product innovation. Furthermore, the project seeks to enhance customer experience by identifying factors that influence satisfaction and retention, ultimately improving customer service and the overall shopping experience. Through these efforts, the project aims to optimize revenue strategies, maximize profitability, and create a more efficient, customer-centric e-commerce operation.

## 1.2. Scope of Study

This study focuses on analyzing customer behavior and sales trends in an e-commerce business using transaction data from **01/12/2010 to 09/12/2011**. The analysis covers key aspects such as **customer segmentation**, **purchase patterns**, **product performance**, and **pricing strategies** to provide actionable insights for optimizing marketing and sales efforts. The dataset includes essential variables like **InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country**, which help in identifying buying behaviors and customer preferences.

The study primarily applies **data-driven techniques**, including **exploratory data analysis (EDA), customer segmentation (RFM analysis and clustering), and visualization tools** to uncover trends. It also focuses on understanding **customer retention, revenue generation, and factors affecting customer satisfaction**. However, the study does not include real-time sales predictions, sentiment analysis, or advanced AI-based recommendation systems. The findings are limited to the dataset provided and may not fully capture evolving market trends or external economic factors influencing e-commerce performance.

## 1.3. Project Significance

This project holds significant value for e-commerce businesses seeking to enhance their **customer engagement, sales strategies, and overall business performance**. By analyzing customer purchase behavior, businesses can gain valuable insights into spending patterns, product preferences, and market demand, enabling them to make data-driven decisions. The **customer segmentation** approach allows companies to tailor their marketing efforts, personalize promotions, and enhance customer retention strategies, leading to improved customer satisfaction and loyalty.

Furthermore, the study helps businesses **optimize pricing strategies**, identify high-value customers, and improve inventory management by highlighting best-selling and underperforming products. These insights contribute to more effective resource allocation, boosting overall profitability. Additionally, the findings can serve as a foundation for developing future **predictive models, AI-driven recommendations, and targeted marketing campaigns**, giving businesses a competitive edge in the rapidly evolving e-commerce landscape. Ultimately, this project provides practical, data-driven solutions to improve decision-making, drive growth, and strengthen the long-term sustainability of an e-commerce business.

# 2. Requirement Analysis

## 2.1. Functional Requirements

* **Customer Segmentation** – The system should categorize customers into different segments based on purchase behavior, spending patterns, and frequency of transactions.
* **Sales and Product Analysis** – The system should analyze sales trends, identify best-selling products, and detect underperforming items to optimize inventory and marketing strategies.
* **Revenue and Pricing Insights** – The system should evaluate the impact of pricing on sales performance, helping businesses develop dynamic pricing strategies to maximize revenue.

## 2.2. Non-Functional Requirements

* **Performance Efficiency** – The system should process large volumes of e-commerce transaction data efficiently, ensuring quick retrieval and analysis of insights.
* **Scalability** – The system should be able to handle an increasing number of transactions and customers as the business grows without significant performance degradation.
* **Data Security & Privacy** – The system must ensure that customer data is securely stored and protected, complying with data privacy regulations such as GDPR to prevent unauthorized access.

# 3. Data Collection and Processing

## 3.1. Data Collection

**Data Sources:** The dataset for this project is collected from Kaggle, a popular platform for open datasets and data science competitions. It includes e-commerce transaction records spanning from 01/12/2010 to 09/12/2011, covering various aspects of customer purchases. The dataset contains key attributes such as InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country, which provide valuable insights into customer behavior, product performance, and sales trends.

##### **Collection Methods:** Using PySpark to connect and extract data from these sources. Sales data is exported as CSV from ERP, weather data is fetched via API, and market data is collected from converted PDF reports.

## 3.2. Data Processing with PySpark

### Data Cleaning:

* **Removing missing values:** Using dropna() in PySpark.
* **Handling outliers:** Detecting and removing outliers using the Interquartile Range (IQR) method.

### Data Transformation:

**Handle Missing Values:**

* Drop rows with missing CustomerID (since they cannot be assigned to a segment).
* Fill missing values in Description with "Unknown" if necessary.

**Remove Duplicates:**

* Drop duplicate rows to ensure data consistency.

**Fix Data Entry Errors:**

* Remove rows with negative Quantity or UnitPrice, as they indicate errors or refunds.
* Convert InvoiceNo to string format, ensuring all values are correctly encoded.

**Handle Canceled Orders:**

* Remove transactions where InvoiceNo starts with "C" (indicating cancellations).

**Data Integration:** Merging Customer Transaction Data with Product Information – Combine e-commerce transaction records with additional product details to enhance insights on product performance, pricing strategies, and customer preferences.

# 4. Data Exploration and Analysis

## 4.1. Exploratory Data Analysis (EDA)

**Statistical Analysis:** Focuses on understanding customer behavior through various analytical techniques. Descriptive statistics are applied to key features such as Quantity and UnitPrice, summarizing central tendencies and variations within the dataset. Additionally, a correlation analysis is performed on Recency, Frequency, and Monetary (RFM) scores to identify relationships in customer spending behavior, with results visualized through a heatmap. The study also includes customer segmentation using clustering algorithms such as K-Means, DBSCAN, and Hierarchical Clustering, ensuring a structured approach to customer categorization. The effectiveness of these clustering techniques is evaluated using Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score, allowing for the selection of the most suitable segmentation model. Finally, WSSSE (Within-Cluster Sum of Squared Errors) is calculated to assess the compactness of the K-Means clusters. These insights provide valuable data-driven recommendations for optimizing marketing strategies, improving customer engagement, and enhancing overall business performance.

### Visualization:

**Missing Values Analysis**

* A **bar chart** showing the percentage of missing values for each feature.

**Descriptive Statistics Visualization**

* **Boxplots** for UnitPrice and Quantity to detect outliers and distribution patterns.

**Customer Purchase Behavior**

* A **histogram** to visualize the distribution of the number of purchases per customer.

**Sales Trends Over Time**

* **Line chart** showing the number of invoices per month and cancellation rates.

**Geographic Customer Analysis**

* **Bar chart** displaying the number of unique customers per country.
* **Comparison of revenue from domestic (UK) vs. foreign customers** using a bar chart.

**Correlation Heatmap**

* A **heatmap** displaying correlations between **Recency, Frequency, and Monetary (RFM) scores** to understand customer spending behavior.

**Customer Segmentation & Clustering**

* **KDE (Kernel Density Estimation) plots** for Recency, Frequency, and Monetary distributions.
* **3D scatter plot** visualizing customer segments identified by K-Means clustering.
* **Bar charts comparing clustering evaluation metrics** such as **Silhouette Score, Davies-Bouldin Score, and Calinski-Harabasz Score**.

These visualizations provide valuable insights into **customer purchasing patterns, sales performance, and optimal customer segmentation strategies**.

## 4.2. Relationship Analysis Between Variables

### Correlation Analysis

* **Data Preparation**
* The dataset is cleaned by **removing the CustomerID column**, keeping only numerical features (Recency, Frequency, and Monetary).
* These columns are **transformed into a feature vector** using **PySpark's VectorAssembler**, which is necessary for computing correlations in Spark.
* **Correlation Computation**
* The correlation matrix is computed using **PySpark's Correlation.corr() function**, which calculates **Pearson correlation coefficients** for the numerical variables.
* The correlation results are converted into a **Pandas DataFrame** for easier interpretation.
* **Heatmap Visualization**
* The correlation matrix is **visualized using Seaborn's heatmap**, displaying the strength and direction of relationships between the RFM metrics.
* The color gradient **ranges from -1 to 1**, where:
  + **1.0** indicates a perfect positive correlation.
  + **-1.0** indicates a perfect negative correlation.
  + **0** means no correlation.

### Insights from the Correlation Analysis

* **Recency vs. Frequency:** A moderate to strong **negative correlation**, meaning customers who purchase frequently tend to have lower recency (i.e., they bought more recently).
* **Frequency vs. Monetary:** A strong **positive correlation**, indicating that customers who purchase more frequently also tend to spend more.
* **Recency vs. Monetary:** A weaker correlation, suggesting that recent purchases do not always translate into higher spending.

### Multivariate Analysis

* **Correlation Analysis (Multivariate Relationships)**
* A **correlation matrix** is computed for Recency, Frequency, and Monetary (RFM scores) using **PySpark’s Correlation.corr()** function.
* The correlation results are **visualized using a heatmap** to show the strength and direction of relationships between these variables.
* **Kernel Density Estimation (KDE) Plots for RFM Features**
* The distribution of Recency, Frequency, and Monetary is analyzed using **Seaborn’s KDE plots** to identify patterns in customer behavior.
* These plots help detect **skewness and outliers** in spending and purchasing frequency.
* **Feature Engineering for Multivariate Analysis**
* A **VectorAssembler** is used in **PySpark** to combine Recency, Frequency, and Monetary into a single feature vector for modeling.
* This step is crucial for clustering and machine learning applications.
* **Clustering Analysis (K-Means and Other Methods)**
* Customers are segmented based on their **RFM features** using clustering techniques like **K-Means, DBSCAN, and Hierarchical Clustering**.
* A **3D scatter plot** is generated to visualize customer segments in terms of their **Recency, Frequency, and Monetary values**.
* **Customer Segmentation Model Deployment**
* A **pipeline** is built using PySpark’s Pipeline() function, which integrates **VectorAssembler, StandardScaler, and K-Means** for multivariate analysis and clustering.
* A **Gradio app** is implemented to allow users to input RFM values and predict customer segments in real time.

### Summary of Insights from Multivariate Analysis

* The correlation matrix indicates that **Frequency and Monetary** have a **strong positive relationship**, meaning customers who buy more frequently also tend to spend more.
* The **KDE plots reveal skewed distributions**, suggesting that most customers have **low Frequency and low Monetary values**, while a few high-value customers make large purchases.
* The **clustering results show distinct customer segments**, allowing businesses to target high-value customers differently from low-spending or inactive customers.

# 5. Feature Engineering

## 5.1. Creating New Features

**Date-Based Features**

Date-related features were created to analyze shopping trends over time.

**Customer Segmentation Features (RFM Model)**

To categorize customers based on their shopping behavior, three key features were used:

* **Recency (R)**: Measures how recently a customer made a purchase.
* **Frequency (F)**: Counts how often a customer makes purchases.
* **Monetary (M)**: Represents the total amount a customer has spent.

These features were used for K-Means clustering to segment customers into groups such as VIP, Potential, Churned, and New Customers, helping optimize targeted marketing campaigns.

**Revenue and Customer Value Features**

* **Total Revenue per Order**: Calculates the revenue generated by each transaction, helping identify high-value orders.
* Customer Lifetime Value (CLV): Estimates the total revenue a customer is expected to generate over their lifetime, aiding in customer retention strategies.

These features helped businesses prioritize high-value customers by offering exclusive promotions and loyalty rewards.

**Purchase Behavior and Seasonal Trends**

* Monthly Purchase Behavior: Analyzes sales trends over time, revealing that transactions peak in November and decline in December due to holiday shopping and seasonal shifts.
* Customer Segments by Purchase Frequency: Categorizes customers into New, Regular, Loyal, and VIP groups based on purchasing habits, enabling personalized engagement strategies.

**Domestic vs. International Market Analysis**

* Domestic vs. Foreign Spending: Differentiates between UK-based and international customers, showing that while domestic customers contribute the most revenue, international markets offer growth opportunities.
* Market Expansion Potential: Suggests businesses should focus on international marketing efforts, particularly in countries with existing customer engagement, such as Germany and France.

**Data Transformation for Model Optimization**

* Log Transformation: Applied to Recency, Frequency, and Monetary values to correct skewed data distributions, ensuring better machine learning model performance.
* Feature Scaling: Standardized different feature scales (e.g., Recency in days, Frequency as count, Monetary in currency) to improve clustering accuracy.

**Final Insights**

The newly engineered features provided valuable insights into customer behaviors, spending patterns, and market trends, allowing businesses to optimize marketing strategies and decision-making. The RFM model and customer segmentation helped tailor engagement efforts, while CLV and revenue metrics ensured that high-value customers were prioritized. Seasonal trends and international vs. domestic spending insights further guided expansion strategies, ensuring data-driven business growth.

## 5.2. Feature Selection

**Selection of Key Features for Customer Segmentation**

To effectively segment customers, only numerical features were retained, including:

* Recency: Time since the last purchase.
* Frequency: Number of transactions per customer.
* Monetary: Total spending of each customer.

Categorical variables such as InvoiceNo, StockCode, and Description were excluded, as they were not useful for clustering models.

**Handling Redundant and Correlated Features**

* CustomerID was removed since it is a unique identifier and does not contribute to customer behavior analysis.
* A correlation matrix was computed to check for multicollinearity between Recency, Frequency, and Monetary (RFM) scores:
  + Recency and Frequency had a moderate negative correlation (-0.26).
  + Frequency and Monetary were strongly correlated (0.56).
* Despite the correlation, all three features were retained, as they provided distinct information for clustering.

Standardization and Transformation

* Feature Scaling: Since RFM values had different ranges, StandardScaler was applied to ensure balanced weightage in clustering.
* Log Transformation: Applied to Recency, Frequency, and Monetary to correct right-skewed distributions, making them more normally distributed.

**Feature Selection for Clustering**

* VectorAssembler combined Recency, Frequency, and Monetary into a single feature vector for clustering.
* Features were used in K-Means, DBSCAN, and Hierarchical Clustering models, with k=4 being the optimal number of clusters.
* Feature selection played a key role in defining VIP customers, potential customers, churned customers, and new customers.

**Final** **Insights**

Feature selection in both files optimized customer segmentation by reducing unnecessary data, improving model performance, and ensuring efficient clustering. The final features—Recency, Frequency, and Monetary—proved to be the most effective for customer classification, leading to better marketing strategies and business insights.

# 6. Machine Learning Model Development and Evaluation

## 6.1. Splitting Data into Training and Testing Sets

* Splitting data into **80% training and 20% testing** using random sampling.

## 6.2. Model Building with MLlib

* **VectorAssembler** to combine Recency, Frequency, and Monetary (RFM) features.
* **StandardScaler** for feature normalization.
* **K-Means Clustering** to segment customers into groups.
* **Silhouette Score Evaluation** to assess cluster quality.
* **Gradio-based Web Interface** for real-time customer segmentation.

## 6.3. Model Evaluation

* **Within-Cluster Sum of Squared Errors (WSSSE)** to measure clustering compactness.
* **Silhouette Score** to assess the separation between clusters.
* **Davies-Bouldin Index** to evaluate cluster similarity.
* **Calinski-Harabasz Score** to determine clustering performance.
* **3D Scatter Plot Visualization** to analyze customer segmentation.

# 7. User Interface Deployment with Gradio

## 7.1. UI Design

* **Input Fields:** Recency (Days), Frequency (Number of Purchases), Monetary (Total Spend in USD).
* **Prediction Button:** Generates the customer segment.
* **Results Section:** Displays the predicted customer segment based on K-Means clustering.

## 7.2. Model Integration with Gradio

* **Loading the trained K-Means model and scaler.**
* **Preprocessing Function:**
  + Applies **log transformation** and **scaling** to RFM data.
* **Prediction Function:**
  + Uses the **K-Means model** to classify customers into one of the **4 segments**:
    1. New Customers - First-time buyers, need engagement strategies.
    2. Potential Customers - Spend high, need retention strategies.
    3. Churned Customers - Inactive, require re-engagement marketing.
    4. VIP Customers - Loyal high spenders, need exclusive benefits.

## 7.3. Result Visualization

* **Gradio UI:**
  + Users enter **Recency, Frequency, Monetary** values.
  + A **prediction result** box displays the customer segment.
* **Interactive Deployment:**
  + **Gradio Interface (gr.Interface)** with labeled input fields.
  + **Launches a Web UI** where users can test predictions.

# 8. Results and Discussion

## 8.1. Achieved Results

* **Cluster Performance:** The K-Means clustering model successfully segmented customers into **four distinct groups** based on Recency, Frequency, and Monetary (RFM) scores.
* **Evaluation Metrics:** The model achieved a **Silhouette Score of 0.56**, indicating a well-structured clustering solution with minimal overlap between segments.
* **Key Insights:**
  + **VIP customers** have the highest **Frequency and Monetary values**, making them the most valuable segment.
  + **Potential customers** show high spending potential but require engagement strategies to increase purchase frequency.
  + **Churned customers** exhibit **low purchasing activity**, indicating a need for re-engagement efforts.

## 8.2. Model Comparison

* **K-Means vs DBSCAN:** K-Means provided **more stable and well-separated clusters**, whereas DBSCAN struggled with noise and required extensive parameter tuning.
* **K-Means vs Hierarchical Clustering:** K-Means was significantly **faster and more scalable**, making it a better choice for handling large datasets efficiently.
* **Silhouette Score Analysis:** The **elbow method and silhouette analysis** confirmed that **k=4** was the optimal number of clusters, balancing intra-cluster compactness and inter-cluster separation.
* **Computational Efficiency:** K-Means outperformed other clustering algorithms in terms of speed and interpretability, making it ideal for real-time customer segmentation.

## 8.3. Discussion of Results

* **VIP Customers contribute the majority of revenue**, emphasizing the need for targeted loyalty programs, exclusive deals, and personalized marketing to **retain and further engage** this segment.
* **Churned customers require re-engagement strategies**, such as email marketing campaigns, personalized discounts, and remarketing ads to **encourage them to return**.
* **Potential customers show high spending potential** but are not frequent buyers, indicating an opportunity to **increase purchase frequency** through targeted promotions and incentives.
* **New customers need onboarding strategies** to convert them into loyal buyers. Offering **first-time purchase discounts, personalized recommendations, and follow-up communication** can improve retention.
* **Overall, customer segmentation using RFM and K-Means provides actionable insights** that can drive data-driven marketing strategies, optimize resource allocation, and improve customer satisfaction.

# 9. Conclusion and Future Development

## 9.1. Achieved Results

"The project successfully applied customer segmentation using **RFM analysis and K-Means clustering**, enabling businesses to identify distinct customer groups and optimize marketing strategies. The clustering model provided **valuable insights** into customer behavior, leading to improved targeting and personalized marketing campaigns."

## 9.2. Limitations

* The model relies solely on **transactional data (RFM metrics)**, **excluding other factors** like customer demographics, browsing behavior, or product preferences.
* **K-Means clustering requires a predefined number of clusters (k=4)**, which may not dynamically adapt to evolving customer behaviors.
* The analysis is **limited to historical purchase data**, meaning it does not predict future customer behaviors or emerging trends.

## 9.3. Future Development Directions

* **Enhancing model features** by incorporating **demographic, behavioral, and external factors** such as seasonality or economic trends.
* **Exploring advanced clustering techniques** like **Gaussian Mixture Models (GMM), Agglomerative Clustering, or Autoencoders** for improved segmentation accuracy.
* **Developing a real-time segmentation system**, updating customer groups dynamically based on **new transaction data**.
* **Integrating the segmentation model into business intelligence (BI) tools** such as **Tableau or Power BI** for more interactive analysis.
* **Expanding the Gradio-based web application** into a **fully functional customer analytics dashboard**, allowing businesses to analyze and act on customer insights efficiently.

# 10. References

●​ Brown, T., & Green, A. (2020). *Big Data Analytics with PySpark.* New York, NY: Data Science Press.  
●​ Pedregosa, F., et al. (2011). *Scikit-learn: Machine Learning in Python.* *Journal of Machine Learning Research*, 12, 2825-2830.  
●​ Apache Spark MLlib Documentation. Retrieved from<https://spark.apache.org/docs/latest/ml-guide.html>●​ Google Colab Documentation. Retrieved from<https://colab.research.google.com/notebooks/intro.ipynb>●​ PySpark Documentation. Retrieved from<https://spark.apache.org/docs/latest/api/python/index.html>

# 11. Conclusion

This project successfully demonstrated the power of **data-driven customer segmentation** using **RFM analysis and K-Means clustering**. By leveraging **PySpark** for large-scale data processing and machine learning, the study identified **four distinct customer segments**, each with unique purchasing behaviors and spending patterns. These segments—**VIP Customers, Potential Customers, Churned Customers, and New Customers**—offer valuable insights that businesses can use to **optimize marketing strategies, improve customer retention, and maximize revenue growth**.

The analysis revealed that **VIP customers contribute the most revenue**, requiring exclusive loyalty programs to sustain their engagement, while **Churned customers need targeted re-engagement strategies** to prevent loss of business. **Potential customers** show high spending potential but low frequency, indicating an opportunity for **personalized promotions** to encourage repeat purchases. **New customers, while promising, require onboarding strategies** such as first-time purchase incentives and personalized recommendations to ensure long-term retention.

The use of **K-Means clustering** proved effective in segmenting customers based on **Recency, Frequency, and Monetary (RFM) scores**, achieving a **Silhouette Score of 0.56**, indicating well-defined customer groups. Additionally, **feature engineering techniques, such as log transformation and standardization**, improved the accuracy and interpretability of the clustering model.

Beyond customer segmentation, the study also provided **valuable business intelligence insights** by analyzing **sales trends, purchasing behavior, and geographic distribution of customers**. The findings emphasized the importance of **seasonal demand fluctuations, the impact of pricing strategies, and differences between domestic and international markets**. These insights enable businesses to make **informed decisions on pricing, inventory management, and targeted marketing campaigns**.

While the project successfully identified customer segments, it has certain limitations. The model relies solely on **transactional data**, excluding other influential factors such as **customer demographics, website interaction data, and external economic indicators**. Additionally, **K-Means clustering requires a predefined number of clusters**, which may not dynamically adapt to evolving customer behaviors.

Despite these challenges, the project lays a strong foundation for future enhancements, including **integrating advanced clustering techniques (e.g., Gaussian Mixture Models, Hierarchical Clustering), incorporating additional customer data sources, and deploying real-time segmentation systems**. By continuously refining the segmentation approach, businesses can achieve **higher personalization, better customer engagement, and long-term competitive advantage in the e-commerce industry**.

## 

## 