Student ID: 2204473
Yashkumar Rajubhai Prajapati

CSC_555 Project Report

Segmenting Beverage Consumers Based on Transaction Patterns

Introduction

It is essential in today's business to understand how customers act so they can be retained more effectively, marketing can be maximized, and revenue can be increased. Customer segmentation is the focus of this project through analysis of a large-scale simulated beverage sales dataset. The primary aim is to identify patterns in customers' purchase behaviour and categorize similar customers into groups based on prominent transaction features. By doing so, businesses can target customer groups more efficiently using personalized offerings, promotions, and communication.

The input used is over one million transaction records, providing depth context for behaviour analysis and clustering. Through distributed data processing, aggregation, and machine learning, the project builds an end-to-end pipeline for deriving actionable insights from raw transactional data.

Objectives

- To clean, aggregate, and transform a large-scale sales dataset
- To engineer useful customer-level behaviour features (e.g., total spend, frequency, use of discount)
- To divide customers into several distinct segments employing unsupervised machine learning techniques
- To visualize and comprehend the characteristics of each segment
- To establish the effectiveness of a scalable data pipeline to customer profile
- Technologies used: S3, PySpark (EMR), Athena, MLlib

Dataset Description

The data used in this project consists of a set of sample soft drink sales data, with detailed transactional data of purchases by the customers across numerous categories of soft drinks. All columns of each record are Order_ID, Customer_ID, Product, Category, Region, Quantity, Unit_Price, Discount, Total_Price, and Order_Date. Both the B2B and B2C purchasing habits have been addressed in the data, with a gigantic universe to work at the customer level.

The data consists of 1,048,575 rows and is over 100MB in size, thus meeting the minimum requirement in terms of data volume required by the project. It was stored in Amazon S3 and processed through distributed computing power to deal with its volume effectively.

The reasons why this dataset has been selected are:

- It contains real customers' transactional behaviour that is segment able.
- It is a categorical and numerical dataset with a high level of feature engineering that is possible on it.
- Because of its heterogeneity and size, it was ideal to pilot an elastic data processing and machine learning pipeline.

Lastly, the goal was to take advantage of this data to derive customer-level features and identify lucrative buying behaviour patterns that will be used to make effective target marketing and business strategy.

Data Transformation and Processing

One million plus raw transactional data was read from Amazon S3 in CSV format. Apache Spark on EMR was utilized to perform required preprocessing tasks like null value drop, record-level aggregation at the customer level, and feature engineering like creation of new features like Total_Orders, Total_Quantity, Total_Spend, Avg_Discount, Unique_Products, and Unique_Categories. Region data was retained by joining with a pre-processed Region data set.

The processed data was once more loaded in Parquet format to S3 for performance tuning and registered as an external Amazon Athena table. Several Athena queries were executed to extract insights such as top spenders, discount recipients, and product engagement levels. This allowed for structured querying with SQL for analysis and machine learning preparation.

Machine Learning Model Development

- KMeans clustering, a machine learning non-supervised learning algorithm, was used for customer segmentation based on their consumption patterns. Apache Spark MLlib was utilized because it is scalable and efficient with big data.
- Seven normalized customer features were used as model input features: Total_Orders, Total_Quantity, Total_Spend, Avg_Discount, Unique_Products, Unique_Categories, and Region_Count. These features measured frequency as well as variety of purchases.
- The Elbow Method was used to find the best value of k (number of clusters). WSSSE values were computed for k=2 to k=10, and a clear inflection at k=4 indicated the best clustering balance.
- The last model was trained for k=4 and labelled each customer by cluster. Cluster centroids were explored for finding representative behaviour types, such as high-value buyers, discount shoppers, and low-engagement buyers.
- PCA was used for projecting the feature vectors of high dimensions to two dimensions for visualization. It showed well-separated clusters in a scatter plot of clusters, confirming that the model is strong in customer segment separation.
- The clustering process transformed raw transaction data to structured segments where more targeted customer engagement and retention strategies could be developed.

Evaluation and Results

Elbow Method was used to determine the best value of k (k=number of clusters) by estimating WSSSE (Within Set Sum of Squared Errors) between k=2 to k=10. Elbow point was clearly seen at k=4 when WSSSE reduced significantly from previous values and began levelling off and indicated that four clusters produced an optimal model with acceptable intra-cluster distance and acceptable complexity.

Model run with k=4 on WSSSE 238.59. Seven features were normalized and employed in assigning each customer to a cluster label. Centroid analysis revealed that:

- Cluster 0 were low-spending low-frequency customers.
- Cluster 1 were mid-value diversity moderate buyers with no discounts.
- Cluster 2 were high-value diversity frequent users with high utilization levels of discounts.
- Cluster 3 were medium-level customers with moderate frequency regularity and discounts.

The two principal PCA features were graphed to observe a scatter plot that took visually the clusters into account. Cluster 0 was well differentiated, and Clusters 1–3 were proximate but separable clusters. This confirmed that the model was effective in segregating primary customer clusters based on behaviour features. These clusters form a valid justification for target, retention, and segmentations-based business decisions, deriving business value by unleashing the power of Spark MLlib with cloud data infrastructure.

Challenges:

- Installed Python packages manually (i.e., NumPy, seaborn) on EMR.
- Athena permission problems; fixed using session tokens and boto3.
- Plot rendering caused initial slowdown in visualization.

Benefits:

- 1M+ record scalability processing with Spark's distributed engine.
- Very fast turnaround of aggregations and complex transformations.
- End-to-end automation through seamless integration across S3, Athena, and Spark.

Future Improvements:

- Include real-time data streaming (e.g., Kinesis).
- Attempt DBSCAN or GMM for adaptive clustering.
- Deploy model with Sage Maker for prod inference.

Diagram 1. End-to-End Data Processing and Clustering Pipeline Architecture

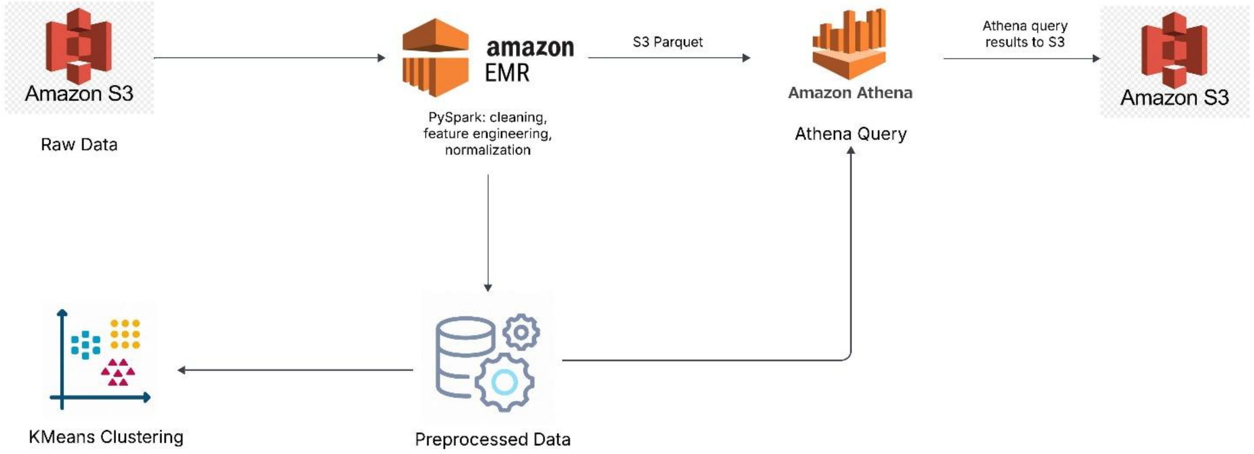
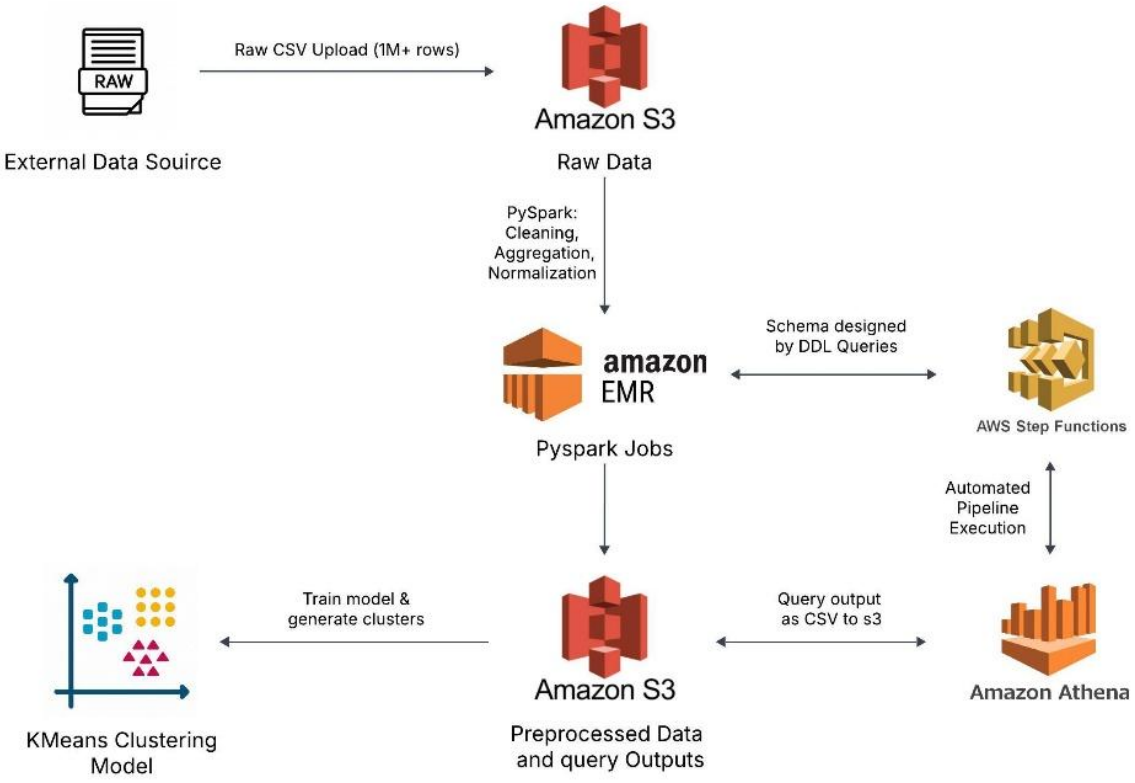


Diagram 2. Full Pipeline Architecture with Orchestration and Scheduled Execution


RAW
External Data Souirce
Raw CSV Upload (1M+ rows)
Amazon S3
Raw Data
PySpark: Cleaning, Aggregation, Normalization
amazon EMR
Pyspark Jobs
Schema designed by DDL Queries
AWS Step Functions
Automated Pipeline Execution
KMeans Clustering Model
Train model & generate clusters
Amazon S3
Preprocessed Data and query Outputs
Query output as CSV to s3
Amazon Athena

Student ID: 2204473
Yashkumar Rajubhai Prajapati

Fig 1. Distribution of Customer Types

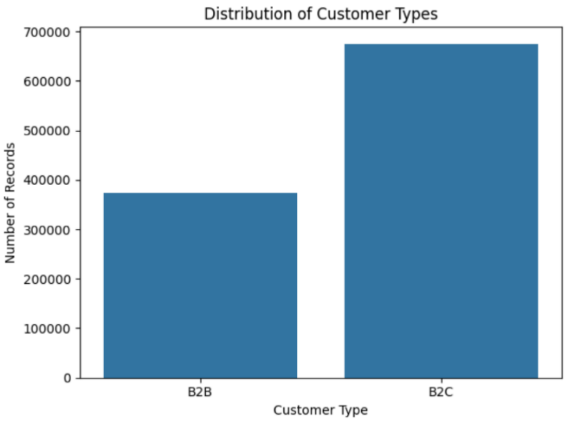


Fig 2.Total Sales by Product Category

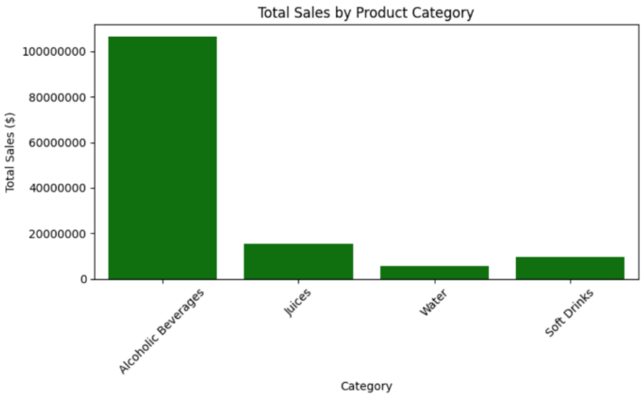


Fig 3. Top 10 Regions by Revenue

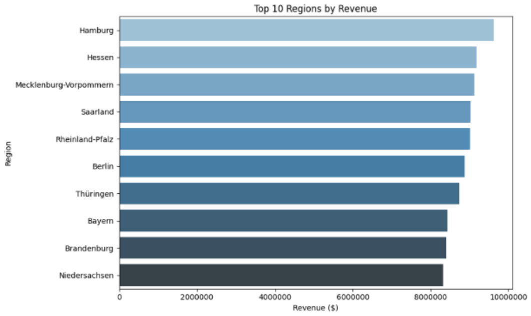


Fig 4. Monthly Sales Trend Over Time

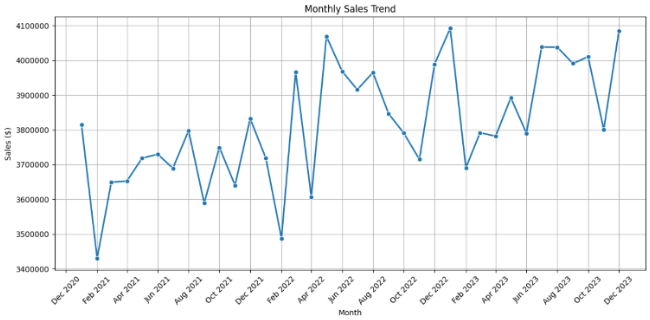


Fig 5. Elbow Method to Determine Optimal Number of Clusters (k)

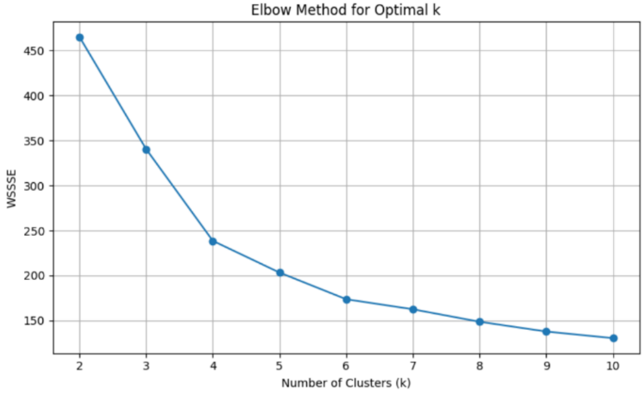
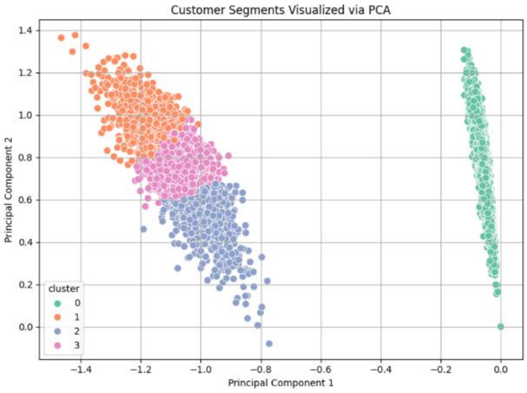


Fig 6. Customer Segments Visualized via PCA



Conclusion:

The project built a stunning end-to-end cloud-based customer segmentation pipeline with Apache Spark and AWS infrastructure. Raw transitional data were computed on EMR, transformed into actionable customer-level features, and warehoused on S3 for querying at low cost with Athena. KMeans clustering with k=4 could successfully partition customers into spending, frequency, and discount behaviour segments, and PCA visualization achieved crisp differentiation between the segments, thus establishing the effectiveness of the model. The response validates the feasibility of distributed processing in big data analytics and provides actionable insights for customer strategy for the future.