MULTIVARIATE

CUSTOMER SEGMENTATION

CLUSTERING, PCA, AND CHURN PREDICTION

AKSHAT KUMAR : 24060641006
NITIKA SONY : 24060641028

INTRODUCTION

In today’s competitive business environment, understanding customer behavior is crucial for growth and sustainability. This project focuses on analyzing customer purchasing patterns using clustering techniques and predicting customer churn through machine learning models. By leveraging Principal Component Analysis (PCA) for dimensionality reduction, K-Means and Agglomerative Clustering for customer segmentation, and Logistic Regression for churn prediction, we aim to provide data-driven insights that help businesses optimize retention strategies. The goal is to categorize customers based on spending habits, identify high-risk churn groups, and suggest personalized retention strategies to improve customer engagement and loyalty. Through this analysis, businesses can enhance decision-making and maximize profitability.

Objectives

Customer Segmentation:

To categorize customers into distinct groups based on their purchasing behavior using clustering techniques like K-Means and Agglomerative Clustering.

Churn Prediction:

To develop a predictive model using logistic regression that identifies customers at risk of leaving, enabling businesses to take proactive retention measures.

Business Strategy:

To provide data-driven recommendations on customer engagement and retention strategies based on clustering and churn analysis, helping businesses maximize revenue and customer loyalty.

MULTIVARIATE

Techniques Applied

PCA

Principal Component Analysis (PCA) is a dimensionality reduction technique used to simplify complex data while retaining most of its important information. In this project, PCA helps reduce the number of features while maintaining the key variations in customer spending behavior. By transforming the original data into new variables (principal components), we can visualize customer patterns in a two-dimensional space, making clustering more effective and computationally efficient.

K-Means Clustering

K-Means Clustering is a centroid-based partitioning method where data points are grouped into K clusters by minimizing the within-cluster sum of squares. The algorithm iteratively assigns points to the nearest cluster center and updates centroids until convergence. The optimal number of clusters is determined using the Elbow Method, and clusters are formed based on their proximity to the centroids.

Agglomerative Hierarchical

Agglomerative Hierarchical Clustering is another clustering technique that builds a hierarchy of clusters. It starts by treating each customer as its own cluster and then merges the closest clusters step by step until all customers are grouped into a hierarchy. Unlike K-Means, this method does not require specifying the number of clusters beforehand. It is useful for understanding hierarchical relationships among customers and identifying natural groupings based on spending patterns.

MULTIVARAIATE

METHODOLOGY

The Project begins with data preprocessing where in missing and invalid values are removed, and a new column is created for Total Spend.

To reduce the complexity of the dataset while retaining its original information **Principle Component Analysis (PCA)** is applied, to perform PCA the data was standardized, this is necessary because variables like Total Spend and Quantity have very different scales, the Z-score normalization technique was used:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of each feature. The variance explained by each Principle Component was analyzed using a Scree plot.

Moving forward we used two clustering methods to segment customers based on their purchasing behavior: **K-means Clustering** and **Agglomerative Hierarchical Clustering.** For K-means the optimum number of clusters were selected using Elbow method and for Hierarchical Clustering Ward's method was used, which minimizes variance within clusters.

Clusters from both the techniques were plotted and compared and evaluated.

MULTIVARAIATE

METHODOLOGY

The evaluation criterion used was Silhouette Score, which measures how well-separated the clusters are based on the following metric:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

Here a(i) is the average intra-cluster distance and b(i) is the average nearest-cluster distance, these were calculated using Euclidean distance. A higher silhouette score indicates better-defined clusters, scores for both the models were compared.

To identify customer churn a logistic regression model was developed. Customers were labeled as churned (1) or retained (0) based on their last purchase date. To increase model performance several new columns were introduced: Transaction, Recency and monetary. The model performance was evaluated using Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) scores.

Based on clustering and churn prediction results a targeted customer retention strategies was developed, for example - offering personalized discounts to moderate spenders.

MULTIVARIATE

Data Pre-processing

Missing Values:

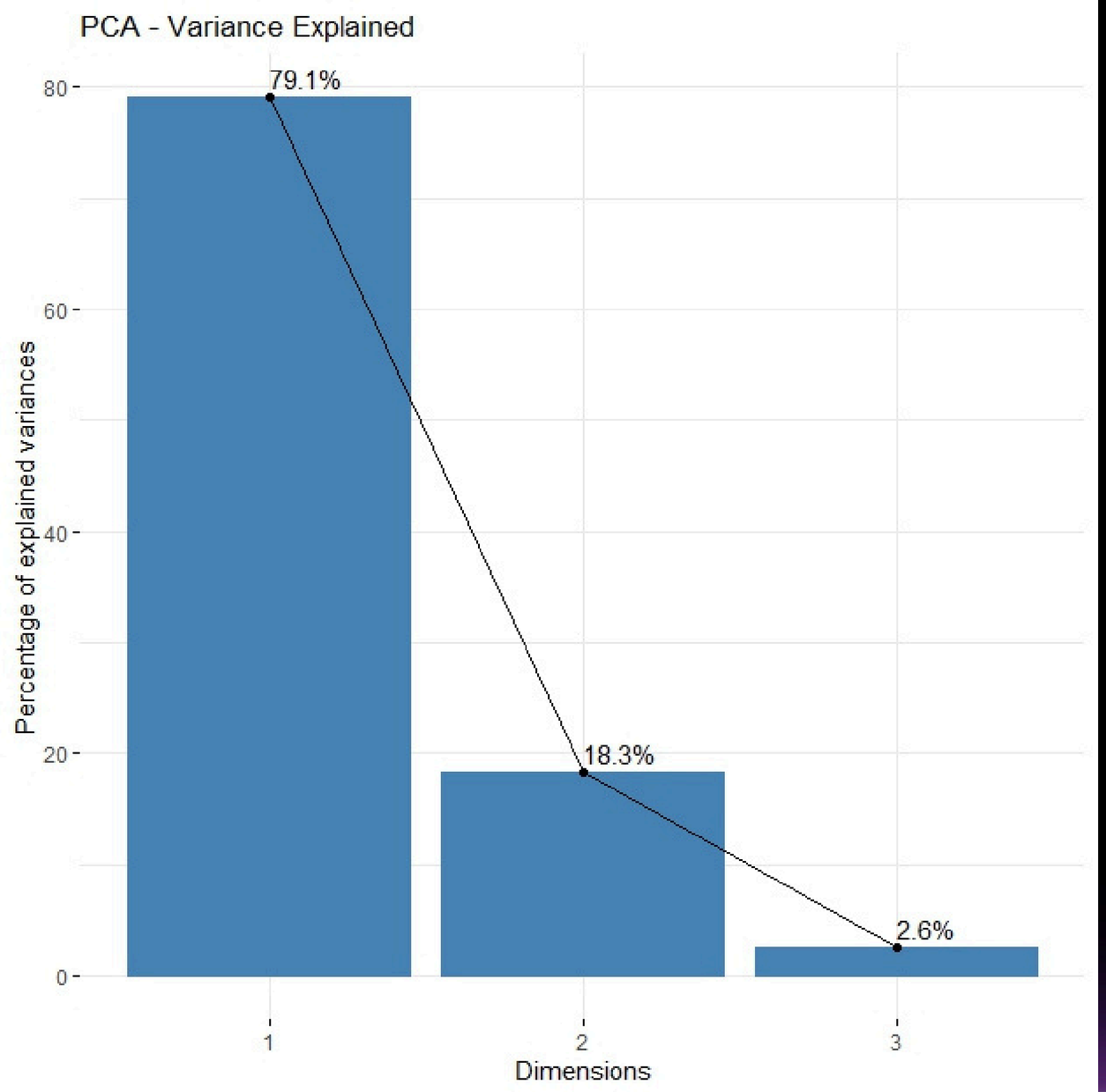
The CustomerID column has missing values (135080 entries). The dataset is cleaned using drop_na() to remove rows with missing values.

Scaling the data:

The customer-level data (excluding CustomerID) is standardized using scale(X), ensuring all features have a mean of 0 and a standard deviation of 1.

Feature Engineering:

New columns are created:
Transaction: The number of unique purchases a customer has made
Recency : Records how recently a customer made a purchase.
Monetary: Records the average spend per transaction
Total_Spend: Quantity * UnitPrice to represent the total amount spent by each customer.



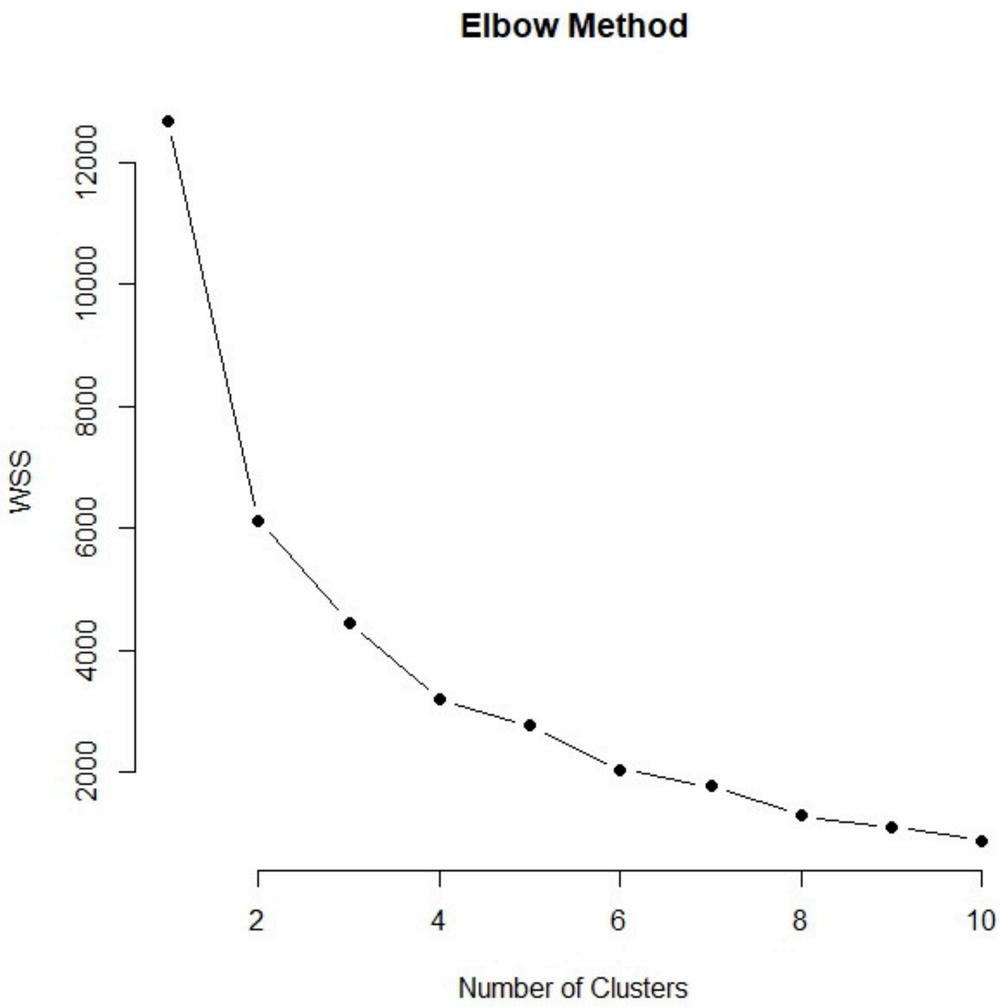
PCA INSIGHTS

To reduce dimensionality and improve clustering performance, PCA (Principal Component Analysis) was applied to customer-level data.

- PCA Summary:
  - PC1 (Principal Component 1) explains 79.1% of the total variance.
  - PC2 (Principal Component 2) explains 18.3% of the total variance.
  - The first two components together capture 97.4% of the variance in the dataset.

A scree plot confirmed that two principal components are sufficient for effective clustering.

The dataset thereafter was transformed into a lower-dimensional space using the first two principal components.



K MEANS

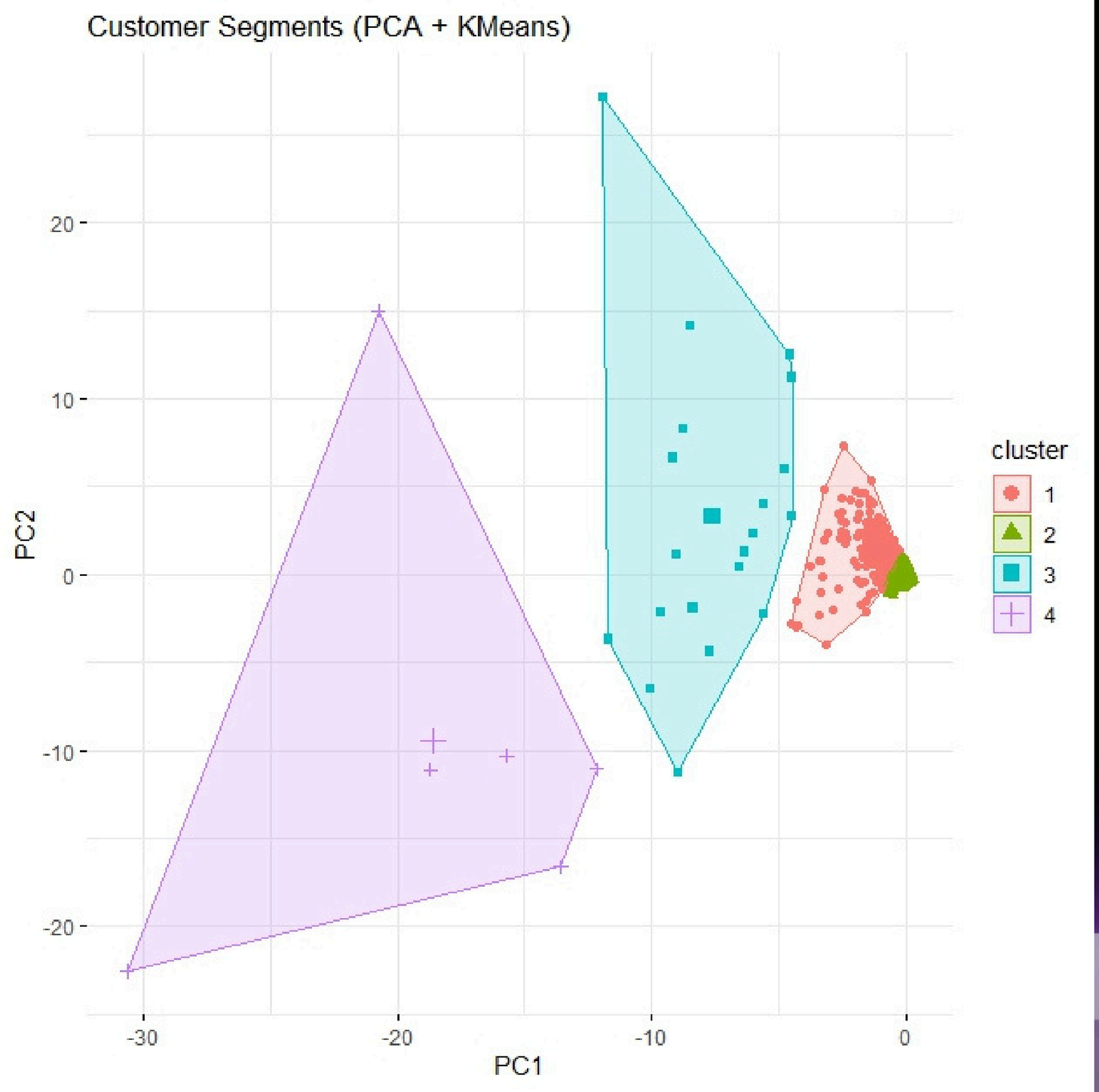
The Elbow Method was used to determine the optimal number of clusters by analyzing the rate of decrease in the within-cluster sum of squares (WSS).

Observations:

- A sharp decline in WSS was observed as the number of clusters increased.
- The "elbow point" at k = 4 indicated the optimal number of clusters, beyond which the reduction in WSS became marginal.
- Increasing the number of clusters beyond 4 provided diminishing returns, making 4 the ideal choice for segmentation.

```
> cat("K-Means Silhouette Score: ", mean(silhouette_kmeans[, 3]), "\n")
K-Means Silhouette Score:  0.7753565
```

The clustering structure was validated using the silhouette score, which was 0.7753, suggesting well-separated and well-defined clusters.
The clustering structure is strong, as indicated by the high silhouette score



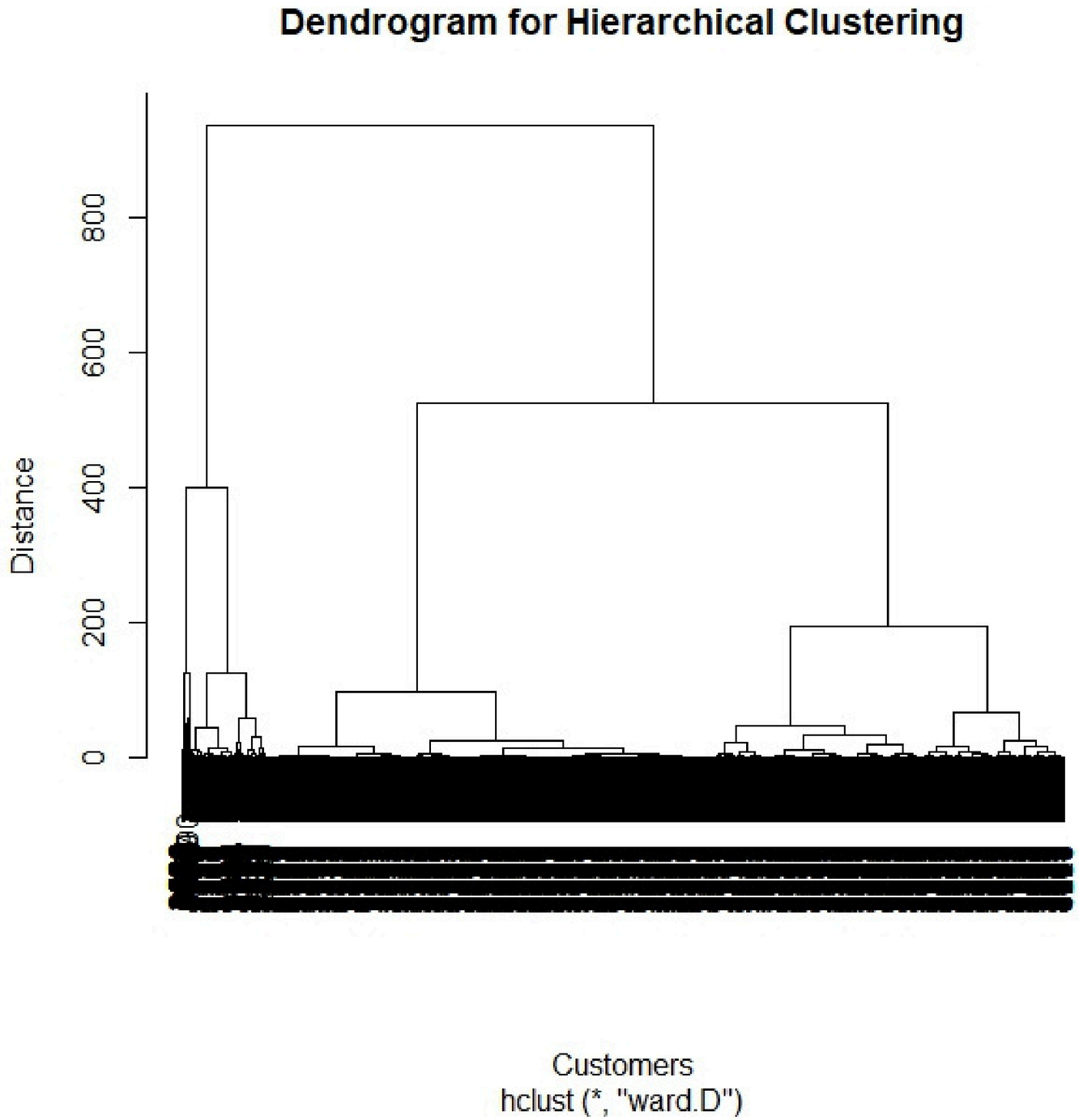
The graph depicts four distinct customer segments identified through K-Means clustering on PCA-reduced data.

Cluster Characteristics:

- Cluster 1 (Red, Circles): Densely packed, indicating a homogeneous customer group.
- Cluster 2 (Green, Triangles): Slightly overlaps with Cluster 1, suggesting related but distinct behaviors.
- Cluster 3 (Blue, Squares): Well-separated with higher dispersion, representing diverse customer behaviors.
- Cluster 4 (Purple, Crosses): Highly isolated, indicating a unique or niche customer segment.

Separation & Similarity:

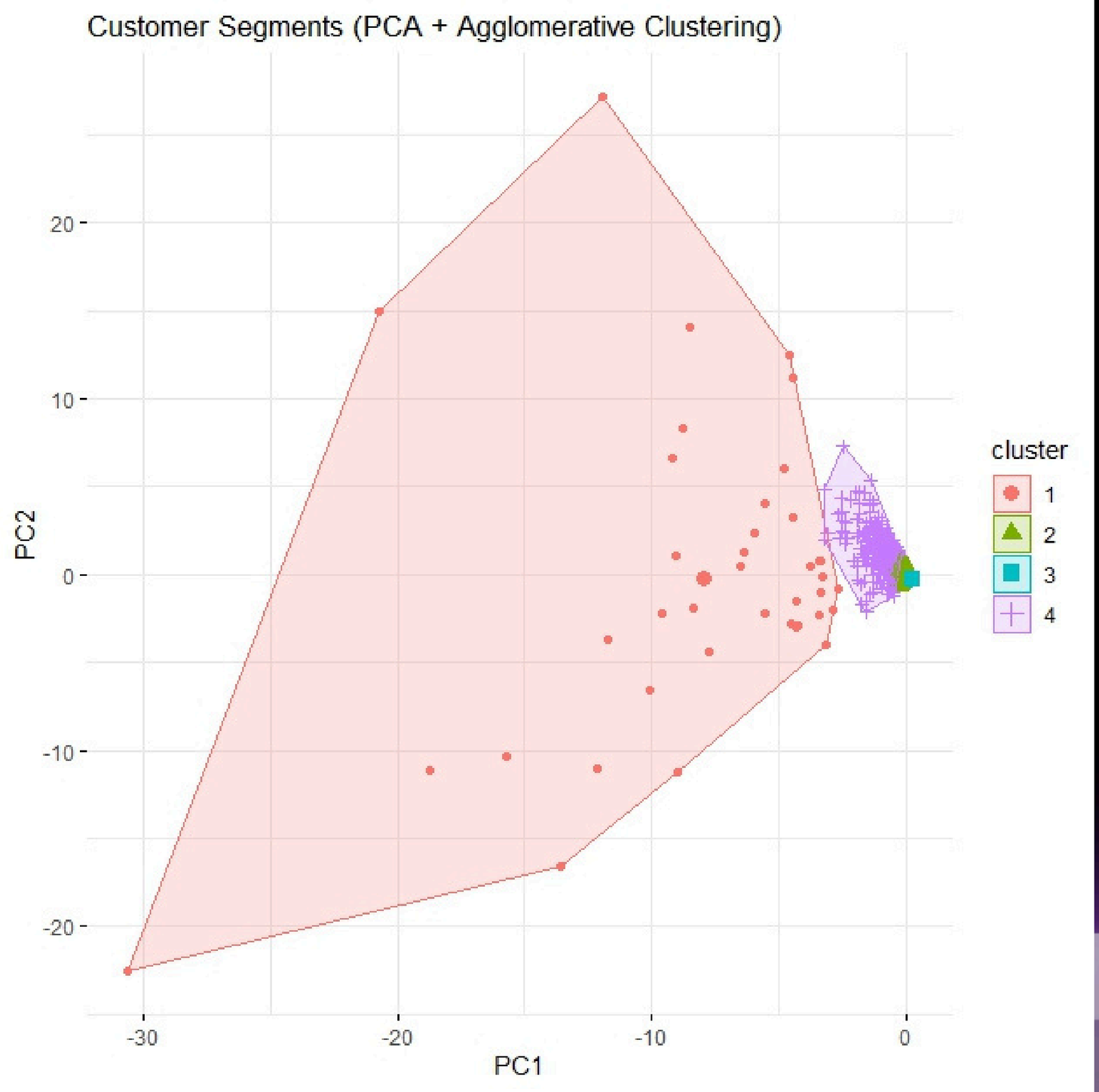
- Clusters 1 & 2 are close, implying behavioral similarities.
- Cluster 3 is distinctly separate, suggesting different purchasing patterns.
- Cluster 4 is the most isolated, possibly an outlier or a specialized customer group.



- The dendrogram represents the hierarchical structure of customer segments, where customers with similar characteristics are merged first before forming broader clusters.
- The height of the branches indicates the dissimilarity between clusters. Higher branches suggest more distinct clusters.
- The thick black lines at the bottom represent individual customers that are very close in similarity.

```
> cat("Agglomerative Clustering Silhouette Score: ", m
Agglomerative Clustering Silhouette Score:  0.4898739
```

- The Silhouette Score (0.4899) indicates moderate clustering quality, meaning some clusters may have slight overlaps.
- Hierarchical clustering captures customer relationships better but may struggle with larger datasets.



Agglomerative Clustering Visualization :

- Customers are grouped into four distinct clusters based on PCA-reduced data.

- Cluster Distribution:
  - Cluster 1 (Red, Circles): The most spread-out group, suggesting diverse customer behavior.
  - Cluster 2 (Green, Triangles): Small and compact, indicating high similarity among customers.
  - Cluster 3 (Blue, Squares): Clearly distinct from others, representing a unique segment.
  - Cluster 4 (Purple, Crosses): Densely packed and slightly overlapping with other clusters.

Logistic Model Interpretation

The goal of churn prediction is to identify customers that a likely to leave and stop purchasing in the future, for this a logistic regression model was developed:

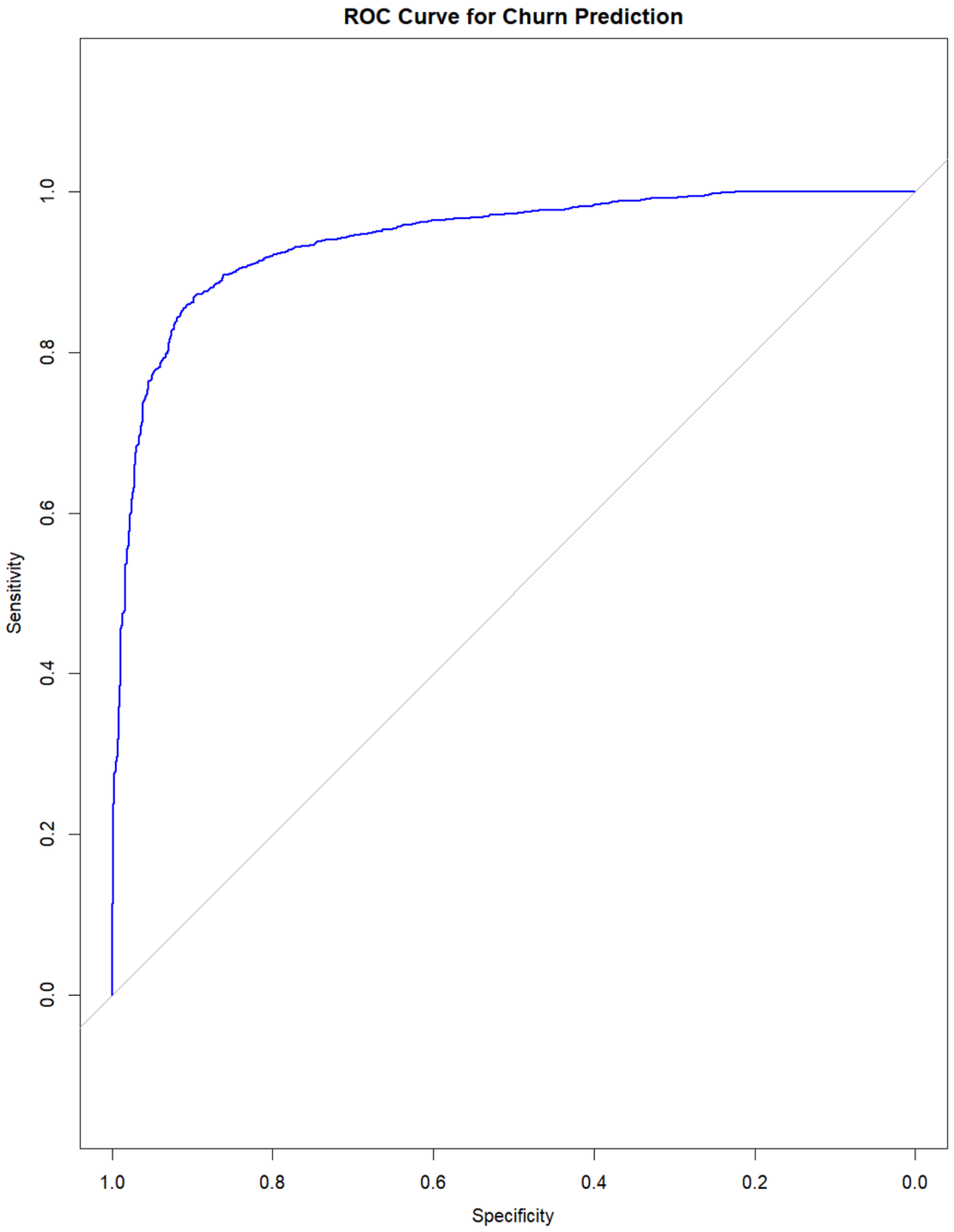
```
# Fit logistic regression model
churn_model <- glm(Churn_Label ~ Total_Spend + Transactions + Total_Quantity + Recency + Monetary,
                   data = customer_df, family = binomial)
```

$$Churn\_Label = \begin{cases} 1, & \text{if last purchase date } < \text{churn threshold} \\ 0, & \text{otherwise} \end{cases}$$

The churn threshold was defined as follows: if a customer has not made a purchase since September 1, 2011, they are labeled as churned (1), otherwise active (0)). 'family = binomial' specifies that the dependent variable (Churn_Label) follows a Bernoulli distribution.

Based on the above model the following results were produced:

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -3.910e+00  1.878e-01 -20.819   <2e-16 ***
Total_Spend     2.846e-04  1.582e-04   1.799   0.0721 .
Transactions    1.060e+00  6.223e-02  17.035   <2e-16 ***
Total_Quantity -8.873e-05  1.465e-04  -0.606   0.5446
Recency         3.524e-02  1.569e-03  22.460   <2e-16 ***
Monetary       -6.009e-04  2.894e-04  -2.076   0.0379 *
```



A slight increase in **total spending is associated with a higher churn** probability, but since the p-value is marginal, there is not a meaningful impact

A higher number of **transactions significantly increases** the likelihood of churn. Customers making more transactions may be less loyal and more likely to switch providers.

Total quantity is not statistically significant as its p value > 0.05, meaning total quantity purchased does not have a clear impact on churn.

A **higher recency (longer time since last purchase) increases churn probability**, as customers who haven't purchased recently are more likely to leave.

A **higher average spend per transaction (Monetary) slightly reduces churn**, therefor customers who spend more per transaction are less likely to leave.

Transactions and Recency are the strongest churn predictors as both are highly significant. Overall the model performs very well with an AUC value of 0.941

MULTIVARIATE

Business Applications

This project can be used by companies and MNCs for:

- Proactive customer retention:
  - By identifying customers likely to churn, businesses can take early action, such as offering discounts, personalized offers, or improved customer service.
- Segmentation & Targeted Marketing:
  - By Understanding which customer attributes contribute to churn helps businesses tailor marketing campaigns for different customer segments.
- Revenue Optimization:
  - Retaining an existing customer is significantly cheaper than acquiring a new one. Churn prediction helps optimize customer lifetime value by reducing revenue losses from customer attrition.

MULTIVARIATE

Conclusion

We began with customer segmentation using K-Means clustering and Agglomerative Hierarchical Clustering, for this purpose PCA was used for dimensionality reduction. The model performance were compared using the Silhouette Score.

Additionally , we developed a logistic regression model for churn prediction, after feature engineering (adding Recency, Monetary Value, and refining Transactions), the model performed well with an good AUC score of 0.941 with the key take away being:

- Customers with high recency and low transaction frequency are more likely to churn.

Combining clustering with predictive modeling enables data-driven customer retention, optimizing marketing efforts and driving business growth.

MULTIVARIATE

THANK YOU!