**Predicting High-Selling Products Using Machine Learning on Retail and Warehouse Sales Data**

**Kisan Nallusamy Dhanabackiam (KN316)**

**Risvan Alangadan (RA725)**

**Introduction**

Retail and warehouse operations are a key point of leverage for supply chain success and overall performance in the consumer market. Organizations rely on knowing product performance to guide their decisions on inventory, logistics and marketing. With the increased visibility of transactional data, "data analytics" and more specifically, "machine learning" offer new ways to surface patterns and make predictions to enhance operational sustainability and profitability.

You are provided with a dataset containing over 300,000 records of monthly retail and warehouse sales, specifically noting product and supplier descriptions, item types, total sales, and transfer activity. The records consist of many years of data allowing the reader to identify trends and potentially forecast performance.

The main purpose of this report is to learn if machine-learning models, and more specifically logistic regression, can predict if a product is going to high seller months. The analysis will include warehouse sales, retail product transfers, item type, and seasonality and will be used to outline the features that are most closely associated with strong product performance. This analysis will start with exploratory data analysis (EDA) which will provide broader understanding of the overall nature of the dataset and identify patterns, then will move on to predictive modelling and unsupervised learning.

Besides classification, the report employs clustering methods to cluster products or suppliers with similar characteristics, and helpful insights for businesses. The end goal is to not only compute the performance of the predictive models but also give useful interpretations that a business could act on, including stocking options, supplier locations, and seasonal sales.

**Exploratory Data Analysis**   
To provide a model, the one could start with an initial data investigation into the data structure, trends, and data issues or anomalies, such as missing values or the data of interest has a skewed distribution.

**Data Overview**

The dataset contains over 307,000 records across multiple years and includes:

* Product-level sales data for each month
* Item types such as Wine, Beer, Spirits, etc.
* Sales figures at both retail and warehouse levels
* Transfers between retail locations
* Supplier information

**Key Observations**

* Missing values: Missing values for retail sales in a record were interpreted as months relating to no units sold, and replaced with 0.
* No sales records: Some months had no retail or warehouse sales of most items, meaning these are likely slow movers or seasonal product sales.
* Sales skew: A small contribution of products accounted for the vast majority of total sales volume (typical to retail sales datasets.
* Seasonality: There also seemed to be growth in retailing activity, near the end of the last few years, more notably in November and December, which may be attributed to shopping during the Christmas holiday season.

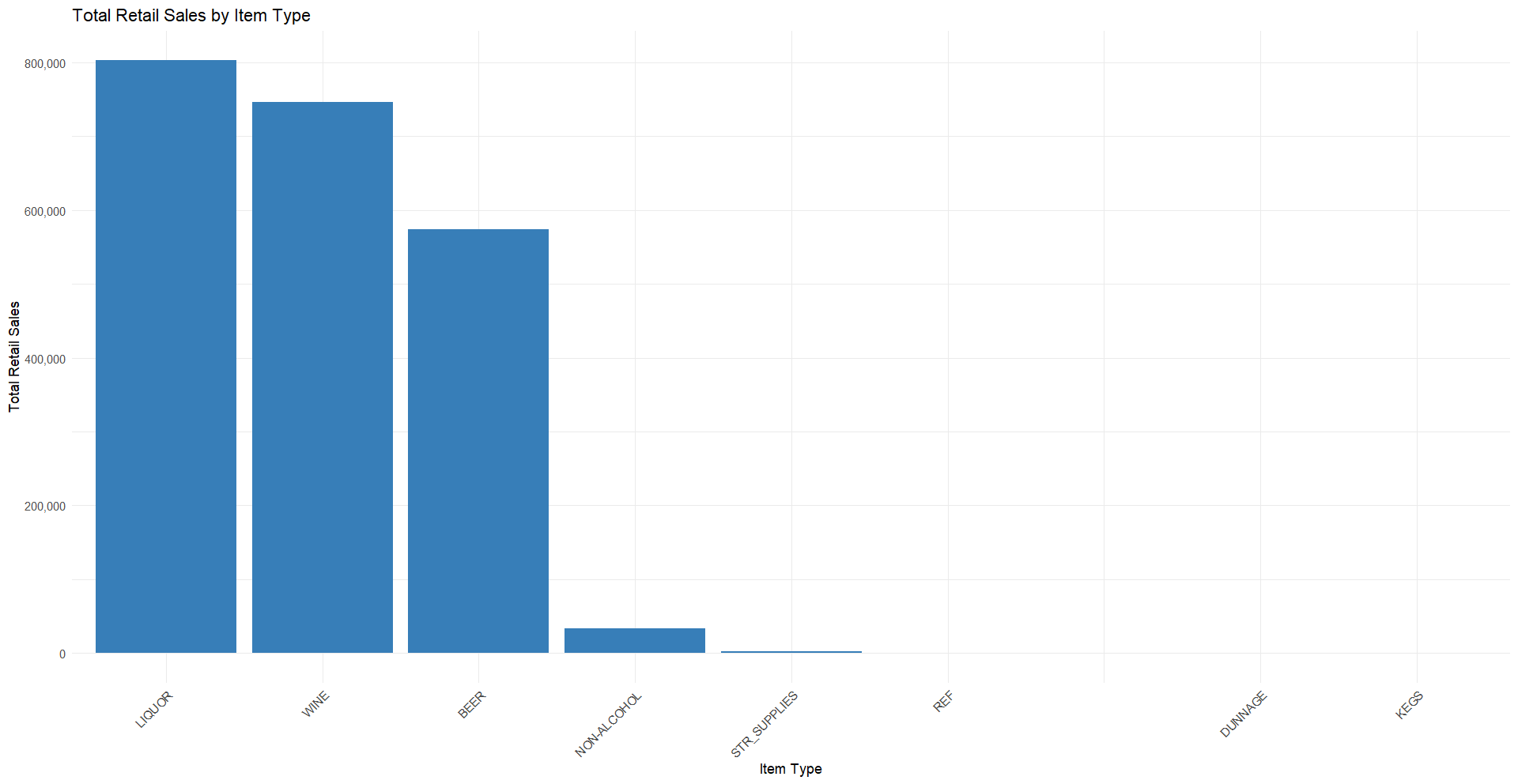
**Visual Summary**

A graph of sales trends

AI-generated content may be incorrect.

**Figure 1:** This line graph shows total monthly sales from retail stores and warehouses. A clear seasonal pattern is visible, with retail sales spiking during the end-of-year holiday period, and warehouse sales increasing slightly earlier, suggesting inventory preparation.

**Total Retail sales by Item Type**

**Figure 2:** The chart shows total retail sales by item type. LIQUOR, WINE, and BEER are the top-selling categories. Some types like REF and DUNNAGE have very low or no sales. This highlights the focus on alcoholic products in retail.

**Methodology**

This section outlines the approach taken to prepare the dataset, use machine learning algorithms, and measure the algorithms' success classifying high volume sellers and product behavior patterns.

**Data Preparation**

Data was cleaned by replacing missing data in the RETAIL.SALES, RETAIL.TRANSFERS, and WAREHOUSE.SALES columns with a value of zero under the assumption that the missing data reveals no activity was recorded. A new target field, High\_Seller (1 = retail sales for a product for a month that was above the median of the entire dataset and zero otherwise) was created. The attributes that we selected to emulate are WAREHOUSE.SALES, RETAIL.TRANSFERS, MONTH, and ITEM.TYPE (categorical variables were one-hot encoded).

**Logistic Regression**

Logistic regression serves as the default binary classification model due to its suitability with binary outcomes, interpretability, and ability to be fitted to predict the probability a product is an "excellent" seller with the selected features. The data was randomly split into held-out training and testing datasets to characterize the generalization performance of the model. The hourly regressions were inspected to try to characterize the impact of some of the selected features on the probability of high sales.

**Clustering (K-Means)**  
K-means cluster analysis was used as an unsupervised learning technique to characterize the natural groupings between the products based on average monthly WAREHOUSE.SALES and RETAIL.TRANSFERS. All the features selected were standardized prior to clustering. The optimal number of clusters was determined using the Elbow Method. This analysis included additional business considerations regarding the similarity between products, any seasonality observed, and potential segmentation by means of inventory or marketing planning.

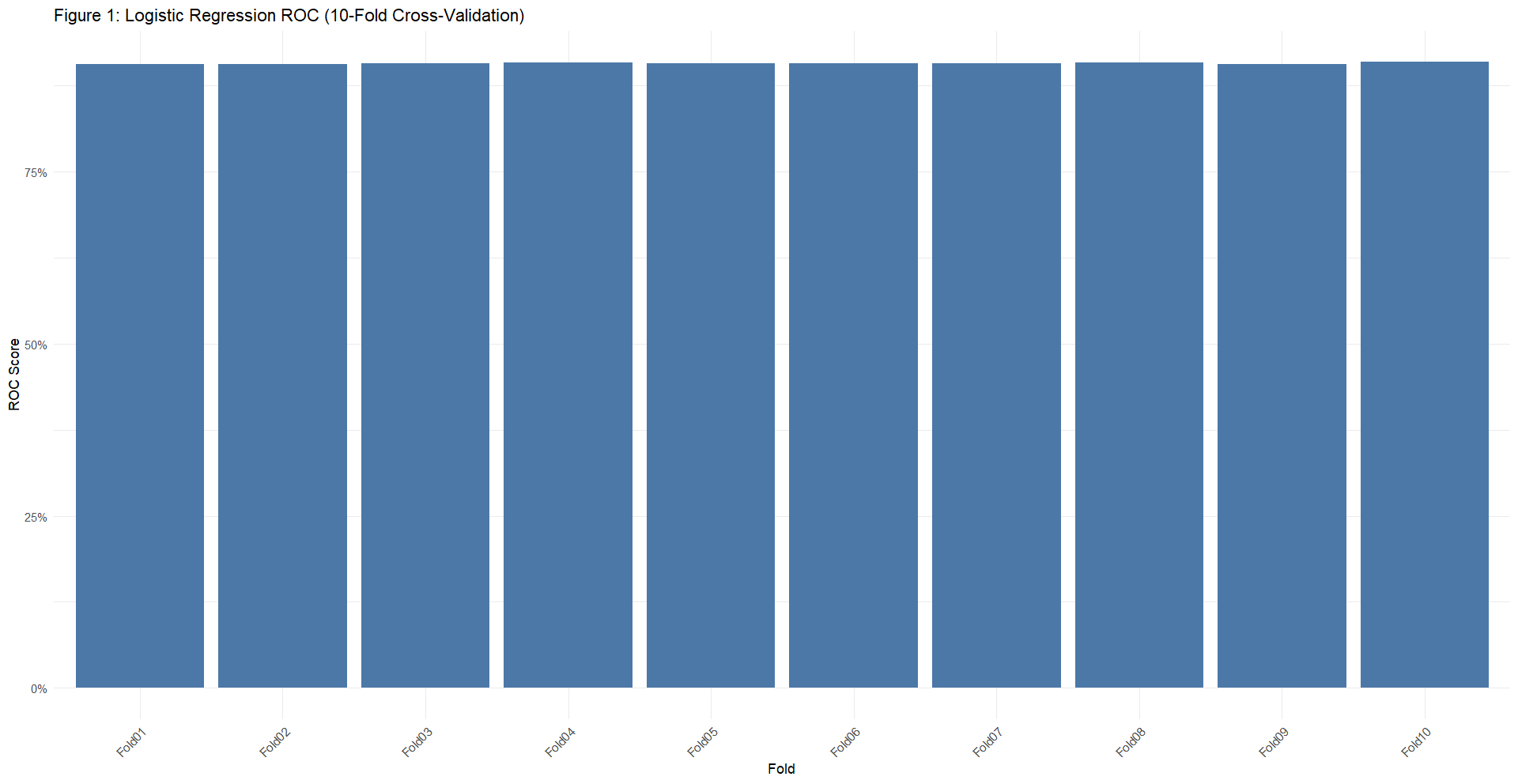
**Cross Validation**  
10-fold cross-validation was used during training to ensure that the model was being correctly estimated. It divides data into 10 parts, trains on 9 parts and tests on the remaining part, and repeats the same 10 times. This drill attempts to prevent overfitting and provides a more accurate estimate of model performance.

**Results**

This section presents the outcomes of the machine learning models applied to the sales data. The primary model used was logistic regression to predict whether a product would be a high seller (above median retail sales). The results were evaluated using accuracy, a confusion matrix, and 10-fold cross-validation. A comparison with a standard decision tree model was also included.

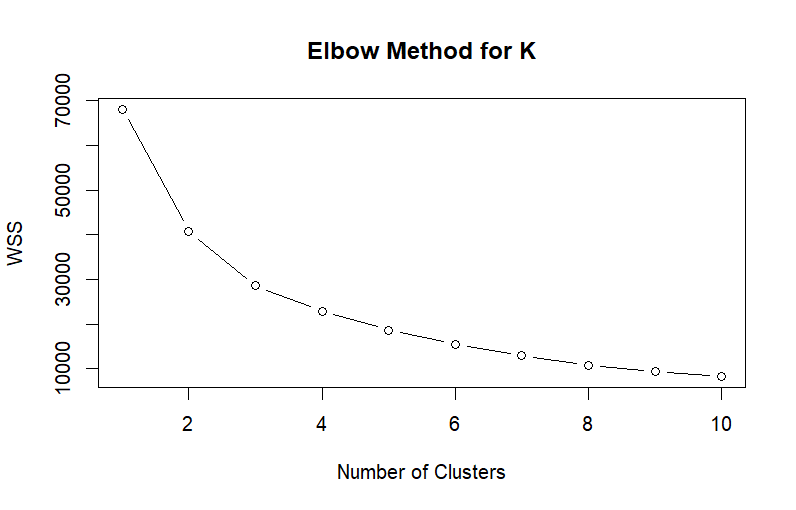
**Logistic Regression Performance**

By 10-fold cross-validation, the logistic regression model achieved 83.7% accuracy. The accuracy indicates a highly reliable capacity to forecast high-selling products.  
The ROC area was computed as 0.91, reflecting very good power to discriminate between high- and low-selling products. The model had high sensitivity of 92% (i.e., it can predict high-sellers very well) but only moderate specificity of 75% (i.e., the model had moderate power to identify low-sellers).

A Kappa value of 0.674 (substantial agreement) confirms the predictive capability of the model as the predicted and actual outcomes are in substantial agreement.  
Confusion matrix also depicts that the model predicts most accurately for products with low false positives and false negatives.  
  
  
**Figure 3:** This bar chart displays the logistic regression model’s accuracy score across 10 different validation folds. The accuracy remains consistent, indicating the model generalizes well to unseen data.ROC scores across 10-fold cross-validation folds for the logistic regression model. All folds consistently achieve ROC values above 90%, indicating strong model performance in distinguishing high-selling products.

**K-Means Clustering**

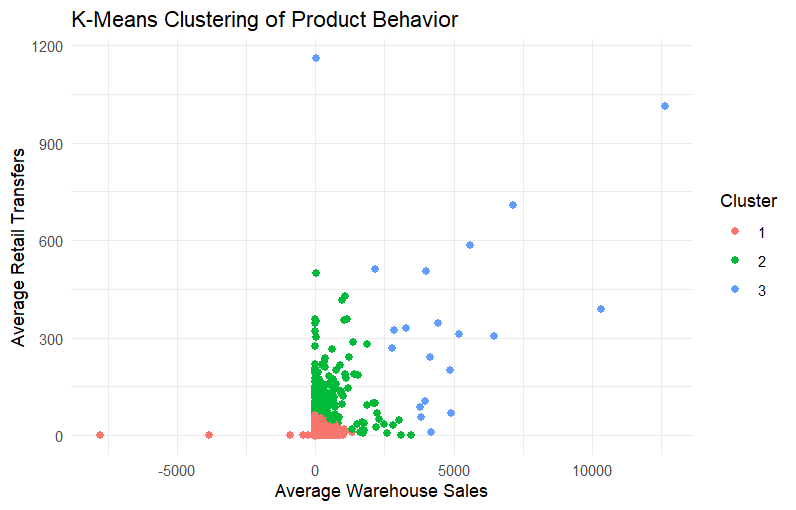
In addition to classification, K-means clustering was applied to explore patterns in product sales behavior. The features used for clustering were **average monthly warehouse sales** and **retail transfers**, which were standardized before training.



**Figure 4:**The Elbow Method suggested that **3 clusters** were optimal. These clusters revealed meaningful differences in product behavior:

* **Cluster 1**: Low in both warehouse and retail sales
* **Cluster 2**: High warehouse sales but low retail sales (potential overstock)
* **Cluster 3**: High in both (consistent top-sellers)

This segmentation can help businesses tailor marketing, inventory, or pricing strategies for different product groups.

  
**Figure 5:** This scatter plot shows the result of K-means clustering based on average warehouse sales and retail transfers. Three clusters were identified:

* Cluster 1 (red): Low sales in both areas
* Cluster 2 (green): High warehouse but low retail sales
* Cluster 3 (blue): High sales in both categories

**Recommendations**

1. **Enhance Top-Seller Support:**Top-seller products identified through the classification model should be preferred in inventory management, promotions, and advertising to take advantage of steady demand.
2. **Prevent Overstock:**Cluster analysis revealed a group of highly sold products at the warehouse with minimal transfers to the retail point. These products need to be checked for demand mismatch and solutions like markdowns or redistribution must be suggested to prevent overstocking.
3. **Review Low-Performing Products:**Products continuously in low warehouse and retail sales clusters can be enhanced by re-pricing, packaging, or deletion, especially when they are taking up inventory space without contributing to the top line.
4. **Add More Business Data:**Future modelling efforts should add more variables (e.g., price, promotions, location) to better enable prediction and segmentation performance, especially around edge cases.

**Limitations  
  
1. Feature Scope:** The data set had predominantly sales and transfer information. The other influential factors like price, location, and promotions were not present, and this can lower the power of the models for forecasting.

**2. Missing Data Assumptions:** It was assumed that missing sales values were zero sales, something that was not universally true and could distort the classification or clustering operations.

**3. Model Constraints:** Logistic regression represents a linear relationship between predictors and the log-odds of the target class. While interpretable, this may not fully describe complex patterns such as those of more complex models.

**4. Clustering Sensitivity:** K-means clustering relies on assumptions like equal cluster size and spherical shape, which may oversimplify actual product behaviors too much.

**Conclusion**

The study illustrated the application of machine learning algorithms (K-means clustering and logistic regression), in predicting and evaluating product performance using the sales data from warehouse and retail events. Logistic regression classification was 83.7% accurate, giving good ROC values with the right recognition of best-selling products. K-means clustering provided business-meaningful segmentation of products as low-performing, overstocked product, and best-sellers, with business-meaningful results. Therefore, even though the models were very accurate, the use of additional variables and exploration of other machine learning algorithms can predict and explain better. However, this analysis opens doors to better data-driven decision making for marketing, inventory management as well as supply chain management.

Data Source Link - <https://www.kaggle.com/datasets/kirbysasuke/retail-sales?resource=download&SSORegistrationToken=CfDJ8NaUH7gUrDtFgwt-2-W4oQDH0Udofkd0g1P6lSIywTEN1-Tw4rQCi09QtnSXA1Ceo92Bh4vaM1IWJmfK2DolipyDFRimZnicn2yrCxWJUZ-f0XtiQQ4Z37gzYKRH7wocneAIGdX20eCLykFdJs3hgjk9cizO_9qC-PCJc2akF2-yMFHrYEEimwwmLWrAF2OyA1XYXVOrkg2CUVLmADb0HaAVLIqiJ3LPo3JmnPmc8axicN8-1obli6P3wICtfDQEODXpjdS7m-ia9PoCYE20-3cpWejsgI79PBmlDHbzxwfAfKg3nEBgeR2EoQPXvULwwX1Ye0TyfvUmTrSUcGR6n_yceovY&DisplayName=Risvan>

**Appendix**> # Load required libraries

> library(tidyverse)

> library(lubridate)

> library(scales)

> library(forcats)

> library(caret)

> library(ggplot2)

> library(cluster

> # Load the dataset

> data <- read.csv("C:\\Users\\RISVAN\\Downloads\\Warehouse\_and\_Retail\_Sales.csv")

> # Convert YEAR and MONTH into a proper Date column

> data$Date <- make\_date(data$YEAR, data$MONTH, 1)

> # Replace missing values in sales columns with 0

> data$RETAIL.SALES[is.na(data$RETAIL.SALES)] <- 0

> data$RETAIL.TRANSFERS[is.na(data$RETAIL.TRANSFERS)] <- 0

**Figure 1**

> # 1. Retail Sales Over Time

> monthly\_sales <- data %>%

+ group\_by(Date) %>%

+ summarise(Retail\_Sales = sum(RETAIL.SALES),

+ Warehouse\_Sales = sum(WAREHOUSE.SALES))

> ggplot(monthly\_sales, aes(x = Date)) +

+ geom\_line(aes(y = Retail\_Sales, color = "Retail Sales")) +

+ geom\_line(aes(y = Warehouse\_Sales, color = "Warehouse Sales")) +

+ labs(title = "Monthly Sales Trends (Retail vs Warehouse)",

+ x = "Date", y = "Sales Volume", color = "Sales Type") +

+ theme\_minimal()

**Figure 2**

> # 2. Total Sales by Item Type

> item\_type\_sales <- data %>%

+ group\_by(ITEM.TYPE) %>%

+ summarise(Total\_Retail\_Sales = sum(RETAIL.SALES)) %>%

+ arrange(desc(Total\_Retail\_Sales))

> ggplot(item\_type\_sales, aes(x = reorder(ITEM.TYPE, -Total\_Retail\_Sales), y = Total\_Retail\_Sales)) +

+ geom\_bar(stat = "identity", fill = "#377eb8") +

+ labs(title = "Total Retail Sales by Item Type",

+ x = "Item Type", y = "Total Retail Sales") +

+ scale\_y\_continuous(labels = comma) +

+ theme\_minimal() +

+ theme(axis.text.x = element\_text(angle = 45, hjust = 1))

**Logistic Regression with Cross-Validation**

> # Create binary classification target: High\_Seller

> median\_sales <- median(data$RETAIL.SALES)

> data$High\_Seller <- ifelse(data$RETAIL.SALES > median\_sales, 1, 0)

> # Select features for modeling

> model\_data <- data %>%

+ select(High\_Seller, WAREHOUSE.SALES, RETAIL.TRANSFERS, MONTH, ITEM.TYPE) %>%

+ drop\_na()

> # Handle categorical variable: ITEM.TYPE

> model\_data$ITEM.TYPE <- as.factor(model\_data$ITEM.TYPE)

> model\_data$ITEM.TYPE <- fct\_lump\_n(model\_data$ITEM.TYPE, n = 5)

> # Make High\_Seller a factor with valid names ("Yes", "No")

> model\_data$High\_Seller <- factor(ifelse(model\_data$High\_Seller == 1, "Yes", "No"))

> # Set seed for reproducibility

> set.seed(123)

> # Setup 10-fold cross-validation

> train\_control <- trainControl(

+ method = "cv",

+ number = 10,

+ classProbs = TRUE,

+ summaryFunction = twoClassSummary,

+ savePredictions = "final"

+ )

> # Train logistic regression model using caret

> logit\_model <- train(

+ High\_Seller ~ .,

+ data = model\_data,

+ method = "glm",

+ family = "binomial",

+ trControl = train\_control,

+ metric = "ROC" # Evaluate based on ROC-AUC

There were 12 warnings (use warnings() to see them)

> # Print model results

> print(logit\_model)

**Cross-Validation Accuracy Plot**

**Figure 3**

> cv\_df <- logit\_model$resample

> # If it has ROC scores instead of Accuracy:

> ggplot(cv\_df, aes(x = Resample, y = ROC)) +

+ geom\_col(fill = "#4c78a8") +

+ labs(

+ title = "Figure 1: Logistic Regression ROC (10-Fold Cross-Validation)",

+ x = "Fold",

+ y = "ROC Score"

+ ) +

+ scale\_y\_continuous(labels = scales::percent\_format(accuracy = 1)) +

+ theme\_minimal() +

+ theme(axis.text.x = element\_text(angle = 45, hjust = 1))

**K-Means Clustering**

cluster\_data <- data %>%

group\_by(ITEM.CODE) %>%

summarise(Avg\_Warehouse = mean(WAREHOUSE.SALES),

Avg\_Transfers = mean(RETAIL.TRANSFERS)) %>%

drop\_na()

# Standardize

scaled\_data <- scale(cluster\_data[, -1])

**Figure 4**

# Elbow Method

wss <- sapply(1:10, function(k) {

kmeans(scaled\_data, centers = k, nstart = 10)$tot.withinss

})

plot(1:10, wss, type = "b", main = "Elbow Method for K", xlab = "Number of Clusters", ylab = "WSS")

**Figure 5**

# K-means with 3 clusters

set.seed(123)

kmeans\_model <- kmeans(scaled\_data, centers = 3, nstart = 25)

cluster\_data$Cluster <- as.factor(kmeans\_model$cluster)

# cluster plot

ggplot(cluster\_data, aes(x = Avg\_Warehouse, y = Avg\_Transfers, color = Cluster)) +

geom\_point(size = 2) +

labs(title = "K-Means Clustering of Product Behavior",

x = "Average Warehouse Sales", y = "Average Retail Transfers") +

theme\_minimal()