EXPLORATORY DATA ANALYSIS OF WEEKLY RETAIL SALES USING R FOR A MULTI-STORE GROCERY CHAIN IN THE U.S.

1. Background

A regional grocery chain with 120 stores across four states in the Midwest wanted to analyze historical transaction data to inform stocking decisions and promotional strategies. The leadership suspected inefficiencies in inventory cycles and a lack of alignment between category performance and discounting efforts.

We were hired to perform an exploratory data analysis (EDA) using R, with the goal of identifying patterns, outliers, and product-category trends across regions and seasons.

2. Objective

- To explore weekly retail transaction data and uncover patterns in product demand
- To identify high-performing and low-performing product categories based on sales, volume, and margins
- To guide strategic decisions in stocking, pricing, and seasonal planning

3. Data Used

Source: Internal POS (Point-of-Sale) and SKU-level inventory logs (2022–2023)

Dataset Details:

- 1.3 million rows of transaction-level data
- Key fields:
  - Date, Store_ID, Region, Product_ID, Category, Subcategory, Units_Sold, Revenue, Cost, Promotion_Flag

Data Preparation Steps:

- Cleaned using dplyr (e.g., handling missing values, merging reference tables)
- Created derived fields: Margin = Revenue - Cost, Week, Season
- Aggregated data by week, store, and product for high-level trend analysis

4. EDA Methodology

4.1 Univariate Analysis

- **Distribution of Units_Sold** by category using ggplot2::geom_histogram()
- Identified SKU-level outliers using boxplots for weekly sales

4.2 Bivariate & Time Series Analysis

- **Revenue vs. Promotion_Flag** comparisons using grouped bar charts
- Time series plots of Units_Sold by Week and Season
- Correlation matrix for Units_Sold, Revenue, and Promotion_Flag

4.3 Regional Comparison

- Created facet_wrap dashboards by Region for each category
- Used group_by() and summarize() to extract top-performing SKUs by revenue and margin

5. Key Findings

Insight Area	**Discovery**
Seasonality	Sales of bakery and beverages spike by ~35% during Q4 (holiday effect)
Category Outliers	Frozen snacks had the most sales variance; flagged for pricing review
Promotion Impact	Promotions increased unit sales by 52% on average but reduced margin by 18%
Regional Patterns	Dairy products sell 21% better in Minnesota vs. other states
Product Gaps	Organic items underperform in urban stores but thrive in suburban markets

6. Interpretation and Recommendations

- **Align promotions with seasonal uplift**: Run beverage and snack discounts in Q4 to amplify natural demand
- **Adjust inventory per region**: Increase dairy supply in MN; reduce organic stock in dense metro zones

- **Review frozen snack pricing strategy** due to high sales inconsistency
- Implement **cluster-based demand planning** based on store performance groups (urban, suburban, rural)

7. Reporting Output

- **R Markdown Report (PDF, 28 pages)**:
  - Interactive histograms, time series, and boxplots
  - KPI tables: Top 10 SKUs by revenue and margin
  - Code appendix with reusable R scripts
- **Excel Summary File**:
  - Weekly and monthly pivot tables
  - Regional performance dashboards
  - Filterable product-level metrics
- **Optional R Shiny App**:
  - Store selector with category drill-down
  - Auto-updating charts for revenue and units sold
  - Export feature for weekly reports

8. Operational Impact

- Improved **in-stock rates by 11%** during peak holiday periods
- Reduced overstock in low-demand urban zones by **\$170K over two quarters**
- Refined pricing and promotion calendar based on cluster-specific demand curves
- EDA dashboard integrated into merchandising team's **monthly review toolkit**