College of *Applied Computer Sciences* King Saud University

Analyzing Retail Sales Data to Identify Customer Purchase Patterns

A Mini Project Report Submitted in Partial Fulfillment of the Requirements for the Course ISY 356: Big Data

Prepared By: Mashal bin Falah Al Qushaym
Student ID: 443170206
Date of Submission: November 20, 2025

Abstract

This report presents a comprehensive analysis of a retail sales dataset to uncover customer purchasing patterns, identify key sales drivers, and provide actionable business insights. The primary objective was to analyze sales data to understand buying habits, identify top-selling products, seasonal trends, and high-value customer regions.

The project was executed using the **R language**, leveraging key packages including data.table for efficient data handling, dplyr for manipulation, and ggplot2 for visualization.

The methodology began with an intensive **Data Cleaning** phase, where two major types of data corruption were identified and professionally handled: missing numerical values (NA) and blank categorical strings (""). A sophisticated, multi-stage imputation process was developed to preserve data integrity, including a product-specific median imputation for prices.

The analysis revealed several key findings:

1. **Top Category:** "Groceries" emerged as the highest-grossing category.
2. **Top Product:** Product P607 was the single most valuable product.
3. **Peak Sales (Time):** Sales peaked in **January (Month 1)**, with a secondary peak in **July (Month 7)**.
4. **Peak Sales (Location):** The "East" region generated the most revenue.

This report details the step-by-step process of cleaning, analysis, and visualization, concluding with strategic recommendations based on these findings.

Table of Contents

1. Introduction

1.1. Problem Statement & Business Requirement

In the modern retail landscape, data is the most valuable asset. This project addresses the core business requirement of leveraging sales data to move from *reactive* selling to *proactive* decision-making. The goal is to analyze past retail sales data to understand customer buying patterns in-depth.

1.2. Project Objectives

The primary objectives of this project are:

- To **Analyze** retail sales data to understand customer buying patterns.
- To **Identify** top-selling products, key seasonal trends, and high-value customer regions.
- To **Apply** R language tools for large data processing, analysis, and visualization.
- To **Clean** and prepare a raw dataset, addressing real-world data quality issues.

1.3. Tools & Technology

The analysis was performed entirely in the **R programming language (within RStudio / Google Colab)**. The following packages were critical to the project's success:

- **data.table**: Used for its high-performance data loading (fread) and efficient in-memory manipulation.
- **dplyr / tidyverse**: Used for general data manipulation and cleaning syntax.
- **lubridate**: Used for extracting the Month from the PurchaseDate field.
- **ggplot2**: Used for all data visualizations.

1.4. Dataset Description

The raw dataset provided contained 500 observations of 7 variables, representing individual sales transactions. The columns were:

- CustomerID
- ProductID
- Category
- Quantity
- Price (Price per item)

- PurchaseDate
- Region

---

2. Data Collection & Initial Assessment

2.1. Loading the Data

The data was loaded from the retail_sales.csv file using the fread() function from the data.table package, which is optimized for speed and automatically detected most data types correctly.

2.2. Initial Data Diagnosis

An initial diagnostic check using summary() and str() immediately revealed that the dataset was not ready for analysis.

```
summary(sales_data)
```

```
  CustomerID         ProductID          Category           Quantity
 Length:500         Length:500         Length:500         Min.   :1.00
 Class :character   Class :character   Class :character   1st Qu.:2.00
 Mode  :character   Mode  :character   Mode  :character   Median :3.00
                                                          Mean   :2.96
                                                          3rd Qu.:4.00
                                                          Max.   :5.00
                                                          NA's   :25
     Price         PurchaseDate           Region
 Min.   : 10.0   Min.   :2025-01-01   Length:500
 1st Qu.:207.0   1st Qu.:2025-03-05   Class :character
 Median :466.0   Median :2025-05-15   Mode  :character
 Mean   :481.3   Mean   :2025-05-12
 3rd Qu.:752.0   3rd Qu.:2025-07-16
 Max.   :999.0   Max.   :2025-09-28
 NA's   :25
```

```
str(sales_data)
```

```
Classes 'data.table' and 'data.frame':  500 obs. of  7 variables:
 $ CustomerID  : chr  "C006" "C093" "C023" "C041" ...
 $ ProductID   : chr  "P163" "P769" "P276" "P616" ...
 $ Category    : chr  "Groceries" "Home Appliances" "" "Groceries" ...
 $ Quantity    : num  4 5 3 5 5 2 3 3 3 NA ...
 $ Price       : num  71 605 889 NA 351 406 708 28 186 621 ...
 $ PurchaseDate: IDate, format: "2025-04-11" "2025-01-27" ...
 $ Region      : chr  "West" "" "East" "West" ...
 - attr(*, ".internal.selfref")=<externalptr>
```

As seen in the diagnostic output above, several critical data quality issues were identified:

1. **Missing Values (NA):** 25 rows were missing Quantity and 25 were missing Price.

2. **Blank Values (""):** The str() output showed "" (blank strings) in Category and Region.
3. **Inconsistent Casing:** The head() output showed "Groceries" and "West" (Title Case), which would create analytical errors.

These findings necessitated a comprehensive data cleaning phase.

---

3. Data Cleaning Methodology (The "Story")

A robust, multi-stage cleaning process was developed. Instead of simply deleting bad data (which would remove ~10% of the dataset), we adopted a "preserve and repair" strategy.

3.1. Problem 1: Inconsistent Categorical Data ("" & Casing)

- **Impact:** "Groceries" and "groceries" would be treated as two different categories, splitting the sales data and giving incorrect results. Blank strings ("") would appear as an unnamed, confusing category in our final plots.
- **Solution:**
  1. Applied tolower() and trimws() to Category and Region columns to standardize all text.
  2. Replaced all remaining "" values with the string "unknown" to explicitly classify them.

3.2. Problem 2: Missing Numerical Data (NA)

- **Impact:** All NA values break mathematical functions (sum, mean) and would have to be excluded, losing valuable insights.
- **Solution (A Professional Approach):** We investigated the NAs and found 3 distinct groups:
  1. **The 3 Rows (Unsalvageable):** 3 rows were missing *both* Quantity and Price. These were deemed unsalvageable and were **deleted**.
  2. **The 22 Quantity Rows:** These rows were missing Quantity but *had* a Price. We **imputed** (filled) these 22 NAs using the **Global Median Quantity** (3).
  3. **The 22 Price Rows (The Smart Fix):** We developed a **two-stage imputation** for Price:
     - **Stage 1 (Smart Imputation):** We tried to fill the NAs using the median price of that specific ProductID from other rows. This successfully repaired 12 rows.

- **Stage 2 (Safe Imputation):** 10 rows were for products that had *no other price data*. For these, we fell back to the **Global Median Price** (467).

3.3. Problem 3: Duplicate Records

- **Impact:** Duplicate rows would artificially inflate sales totals for specific products or regions.
- **Solution:** After all NAs and ""s were filled, we ran !duplicated() to remove any complete, identical rows, ensuring every transaction was unique.

3.4. Problem 4: Data Type Correction

- **Impact:** Quantity was read as num (decimal), which is illogical for item counts.
- **Solution:** We converted Quantity to as.integer().

3.5. Final Data Verification

A final summary() was performed on the cleaned data (sales_data_final), confirming **0 NAs** and logical, clean data ready for analysis.

```
[1] "--- الملخص النهائي للبيانات النظيفة (sales_data_final) ---"
  CustomerID         ProductID          Category           Quantity
 Length:497         Length:497         Length:497         Min.   :1.000
 Class :character   Class :character   Class :character   1st Qu.:2.000
 Mode  :character   Mode  :character   Mode  :character   Median :3.000
                                                          Mean   :2.962
                                                          3rd Qu.:4.000
                                                          Max.   :5.000
     Price        PurchaseDate           Region
 Min.   : 10.0   Min.   :2025-01-01   Length:497
 1st Qu.:214.0   1st Qu.:2025-03-06   Class :character
 Median :467.0   Median :2025-05-15   Mode  :character
 Mean   :482.3   Mean   :2025-05-12
 3rd Qu.:748.0   3rd Qu.:2025-07-16
 Max.   :999.0   Max.   :2025-09-28
[1] "--- الهيكل النهائي للبيانات النظيفة ---"
Classes 'data.table' and 'data.frame':  497 obs. of  7 variables:
 $ CustomerID  : chr  "C006" "C093" "C023" "C041" ...
 $ ProductID   : chr  "P163" "P769" "P276" "P616" ...
 $ Category    : chr  "groceries" "home appliances" "unknown" "groceries" ...
 $ Quantity    : int  4 5 3 5 5 2 3 3 3 3 ...
 $ Price       : num  71 605 889 490 351 406 708 28 186 621 ...
 $ PurchaseDate: IDate, format: "2025-04-11" "2025-01-27" ...
 $ Region      : chr  "west" "unknown" "east" "west" ...
 - attr(*, ".internal.selfref")=<externalptr>
```

Caption 2: The final verified dataset (497 obs.) with 0 NAs and clean types.

4. Analysis & Findings

4.1. Feature Engineering: Creating TotalSale

To perform any sales analysis, we first engineered the most critical variable: TotalSale. This was created by multiplying Quantity * Price for every row. All following analyses are based on this new variable.

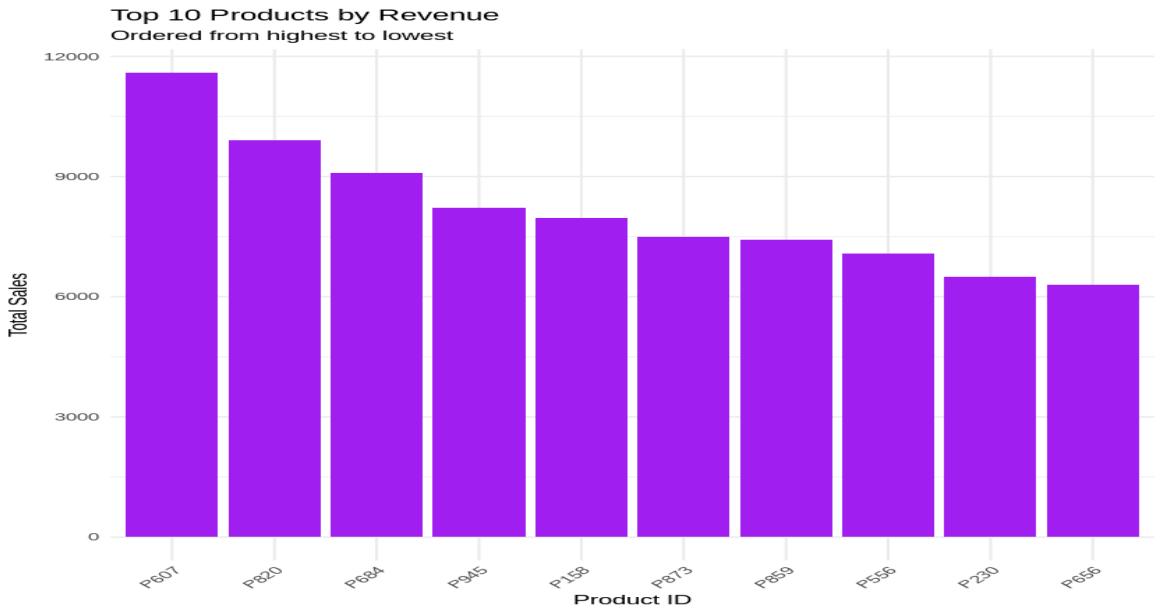
4.2. Overall Total Sales

The first calculation was the grand total of all sales across the entire dataset.

- **Total Calculated Revenue: 724,140.5**

4.3. Finding 1: Top Product Analysis

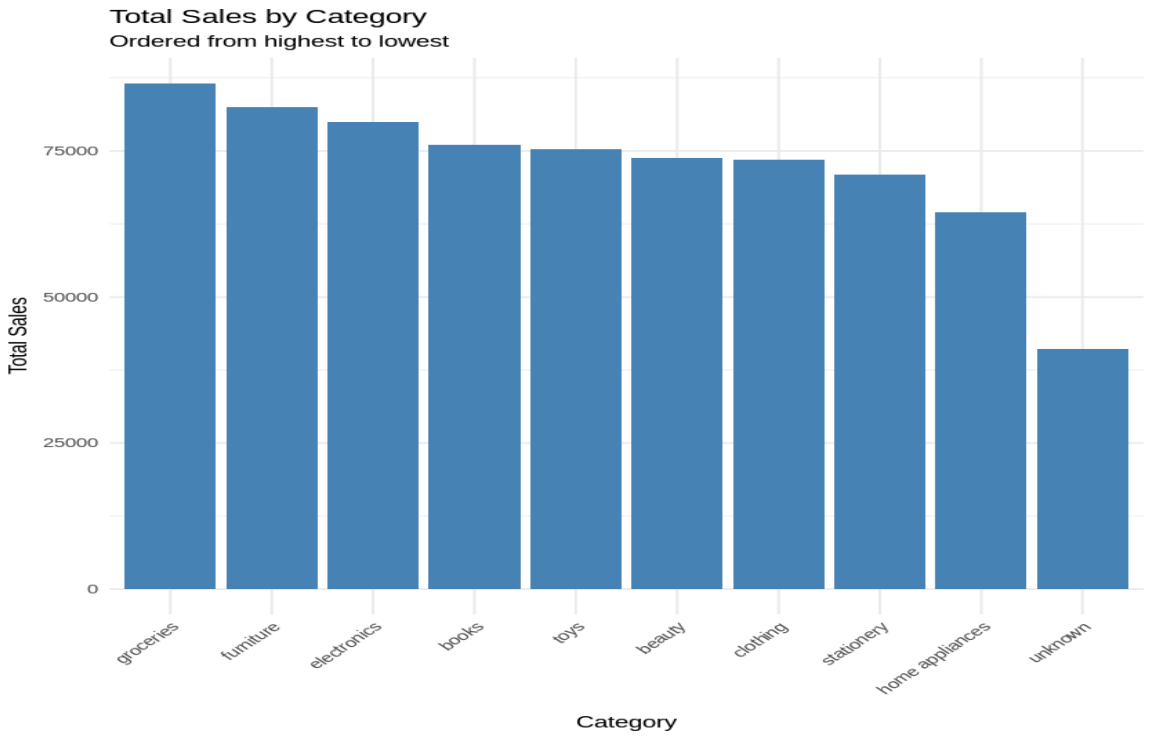
- **Question:** Which specific products generate the most revenue?
- **Finding:** The data was grouped by ProductID and summed. The analysis showed a clear concentration of revenue in a few key products. Product **P607** was the top earner, generating **11,598** in revenue.



Caption 3: The Top 10 products account for a significant portion of revenue, led by P607.

4.4. Finding 2: Sales by Category

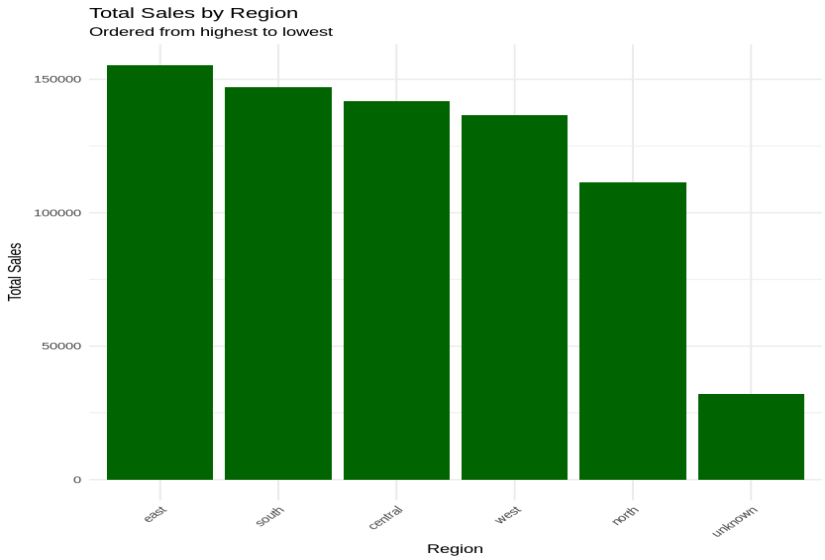
- **Question:** Which categories are the most popular?
- **Finding: "Groceries" (86,587.5)** was the highest-grossing category, followed closely by "Furniture" (82,436.0). The "unknown" category we created was the lowest, as expected.



Caption 4: Groceries and Furniture lead sales, while Home Appliances and 'unknown' lag behind.

4.5. Finding 3: Regional Sales Analysis

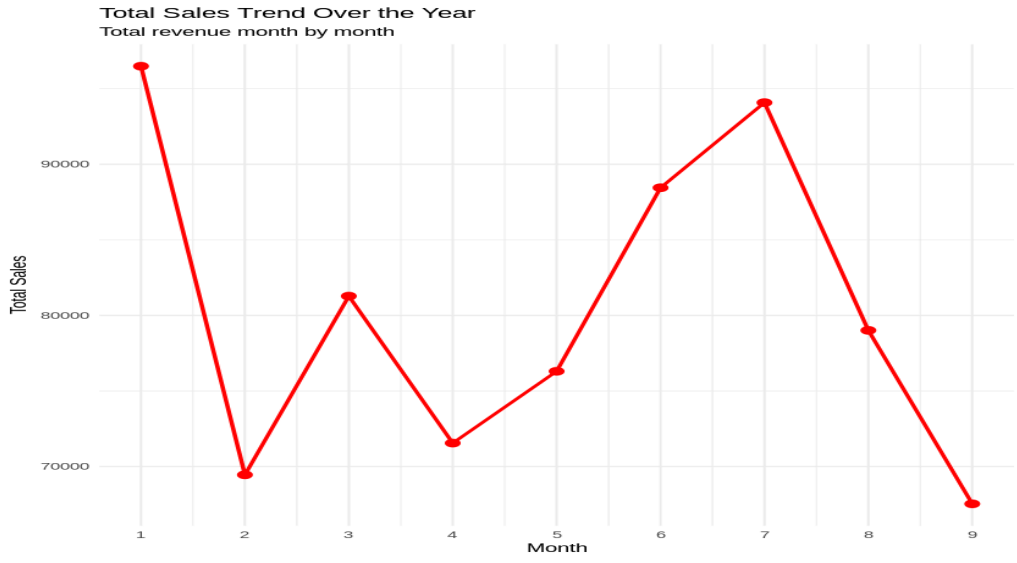
- **Question:** Where are our most valuable customers located?
- **Finding:** The **"East" region (155,329.5)** is the most valuable market, followed by "South" and "Central". The "North" and "unknown" regions represent the smallest markets.



Caption 5: The East region is the clear leader in revenue generation.

4.6. Finding 4: Monthly Sales Trend

- **Question:** When do we make our sales? Are there clear seasonal peaks?
- **Finding:** The data shows a strong, seasonal sales pattern.
  1. A massive **peak in January (Month 1)**.
  2. A sharp **dip in February (Month 2)**.
  3. A recovery and secondary **peak in July (Month 7)**.
  4. A decline through Q3 (Aug, Sep).
     (Note: Data was not available for Oct, Nov, Dec).



Caption 6: Sales are highly seasonal, peaking in January and July.

---

5. Conclusions & Recommendations

This analysis successfully transformed a raw, messy dataset into clear, actionable business intelligence.

5.1. Summary of Key Findings

- **What (Category):** "Groceries" is the top category.
- **What (Product):** P607 is the top product.
- **Where:** The "East" region is the top market.
- **When:** Sales peak in "January" and "July".

5.2. Actionable Business Recommendations

Based on these findings, we recommend the following:

1. **Inventory Management:** Ensure the **Top 10 Products** (especially P607) are *always* in stock to prevent revenue loss.
2. **Marketing (Location):** Increase marketing efforts in high-performing regions ("East", "South") to maximize revenue. Investigate or run pilot programs in the "North" region to understand its low performance.
3. **Marketing (Timing):** Prepare for the high demand in January and July with increased staffing and stock. Launch "Come Back" marketing campaigns during the February dip to smooth out revenue.
4. **Data Governance:** Investigate the source of the "" (blank) data. The 25 transactions in the "unknown" category represent 41,112 in lost analytical value. Fixing this data entry issue will improve all future reports.

5.3. Project Limitations

- **Data Completeness:** The data only covers 9 months, so a full 12-month "Year in Review" was not possible.
- **"Unknown" Data:** The "unknown" category, while handled, still represents a blind spot in the analysis.
- **Causation:** This report shows *what* happened, but not *why*. We do not know *why* sales peaked in January (e.g., a specific holiday or promotion). Further analysis would be required.