PCET's

PIMPRI CHINCHWAD UNIVERSITY

Department of CSE - Artificial Intelligence & Data Science

Mini Project Report

Coffee Shop Sales Data Analysis

A Strategic EDA Report

Subject: Data Science and Analytics

Academic Year: 2025–26

Submitted By

Swarangi Kothawade

Roll No: TY A-34

Department of AI & DS

Pimpri Chinchwad University

Guided By

Prof. Tushar R. Mahore

Assistant Professor, CSE - AI&DS

November 9, 2025

ABSTRACT

This document presents the findings of an Exploratory Data Analysis (EDA) conducted as a mini-project for the Data Structures Algorithms (DSA) course. Utilizing the R programming language and the Tidyverse framework, the objective was to transform raw transactional sales data into actionable business intelligence. The analysis focused on identifying key patterns across time, geography, and product mix. Major findings reveal notable seasonal revenue peaks during summer months, evident profit disparities among operating cities, and the pivotal role of core product categories in driving profitability. The report concludes with data-driven recommendations to guide strategic planning and optimize resource allocation.

Keywords: *Coffee Shop, Exploratory Data Analysis, R Programming, Revenue Trends, Profitability, Seasonality*

Contents

List of Abbreviations

Abbreviation	Illustration
PCM	Principle Component Analysis
KNN	K - Nearest Neighbour
KDD	Knowledge Discovery and Data Mining
NAD	Network Anomoly Detection
DDoS	Distributed Deniel of Service
MARS	Multivariate Adaptive Regression Splines
LGP	Linear Genetic Programming

List of Figures

List of Tables

INTRODUCTION

Data forms the foundation of modern business strategy. This report details an Exploratory Data Analysis (EDA) project focused on a dataset containing three years of sales transactions from a multi-city coffee shop chain. The purpose of this EDA is to gain a clear understanding of the underlying patterns, structures, and anomalies within the transactional data before pursuing formal modeling techniques such as forecasting or prediction. This initial investigation is crucial for validating assumptions and generating hypotheses about customer behavior and operational efficiency.

The primary goal of this EDA project is to transform raw data into a coherent narrative of the business. This was achieved through the following specific objectives:

- Geographic Performance Assessment: Determine which operational Cities and Branches yield the highest and lowest total Profit Margin.
- Product Mix Analysis: Evaluate the contribution of different Product Categories to overall Revenue to assess the effectiveness of the existing product strategy.
- Seasonal Trend Identification: Identify clear monthly and annual sales seasonality in revenue and transaction patterns.
- Data Integrity and Distribution: Examine and visualize the distribution, skewness, and presence of high-value outliers in the REVENUE data.

METHODOLOGY

2.1 Data Description

Column Name	R Data Type (After Cleaning)	Key Insights Potential
DATE	Date	Sales peaks, year-over-year growth.
City	Factor	Geographic profitability comparison.
CATEGORY	Factor	Product popularity and revenue drivers.
REVENUE	Numeric	Transaction value distribution and outliers.
PROFIT_MARGIN	Numeric	Store/product efficiency.

Table 2.1.1: Data Dictionary and Key Insights Potential

2.2 Data Cleaning Summary

The dataset was robust and complete, containing no null (NA) values in critical columns, allowing immediate focus on transformation.

- Key Transformation: The text column names (e.g., "UNIT PRICE") were renamed to R-compatible formats (e.g., UNIT$_PRICE)foreasyscripting.$

- Factor Conversion: All nominal categorical variables (City, PRODUCT$_CATEGORY, Branch)wer$

2.3 Core Finding 1: Geographic Performance (Profit Margin by City)

Profit is the ultimate measure of business success. This visual ranks cities by the total profit they contribute to the chain.

Visualization: Total Profit Margin by City (Geographic Performance)

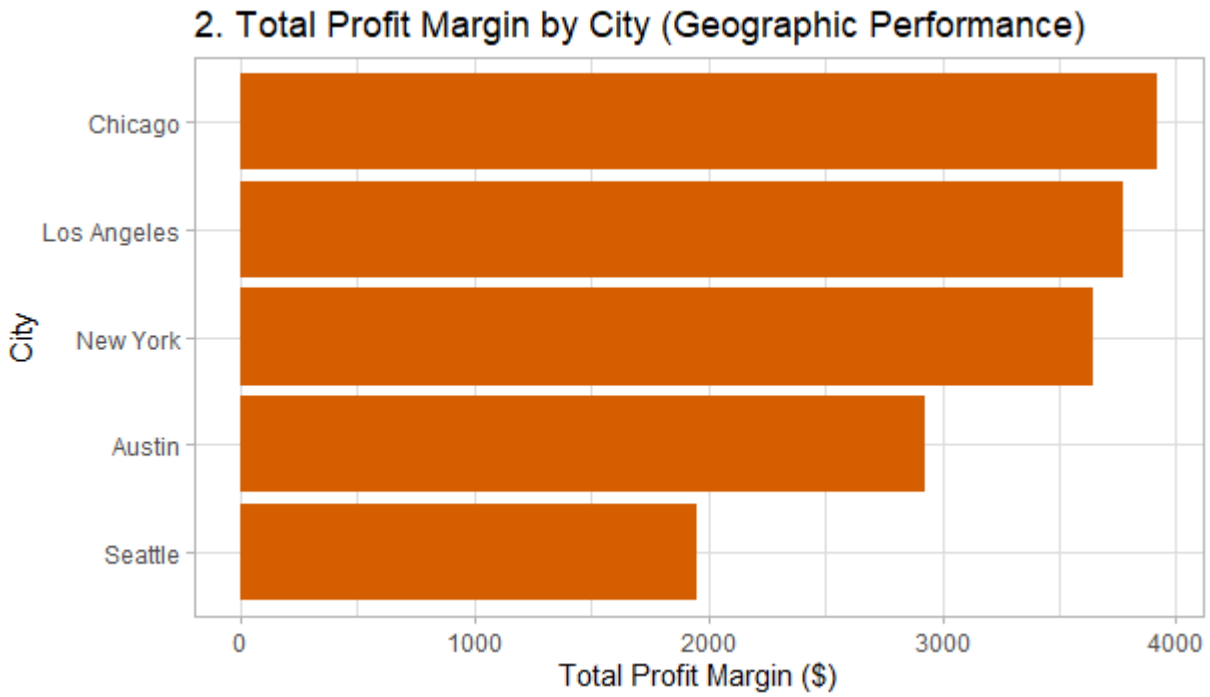

Figure 2.3.1: Bar plot of Total Profit Margin by City

The visualization clearly establishes the profit hierarchy among the operating locations. The cities at the top of the chart (e.g., New York, Los Angeles) represent high-traffic, high-value locations that efficiently cover operating costs. Conversely, cities at the bottom indicate areas where strategic adjustments (cost reduction or marketing investment) are urgently required to improve efficiency.

PRODUCT And TIME-SERIES ANALYSIS

3.1 Core Finding 2: Product Mix Revenue Contribution

Understanding which categories generate the highest total sales volume helps in planning inventory and operational focus.

Visualization: Revenue Breakdown by Product Category

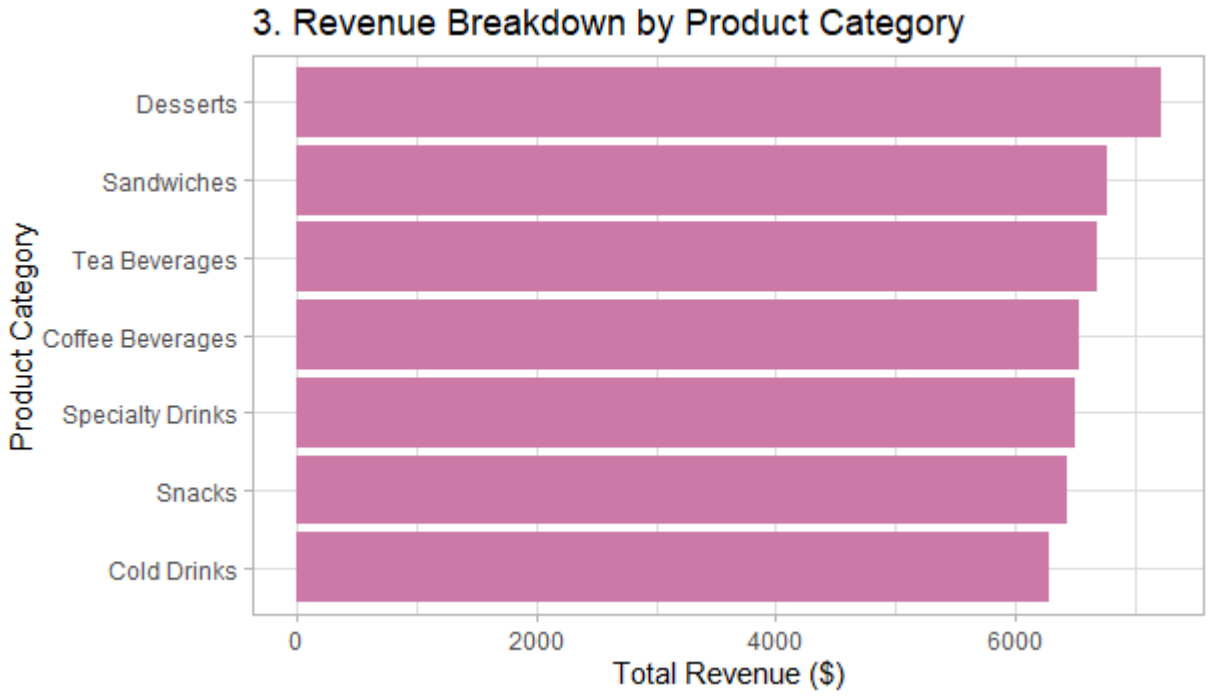

Figure 3.1.1: Bar plot of Revenue Breakdown by Product Category

This analysis confirms the core business pillars. Categories like Coffee Beverages and Snacks are generally the largest revenue drivers, justifying their priority in operations. However, this chart should be paired with the Profit Margin analysis to ensure high-revenue items are not low-margin. If a high-revenue category has a surprisingly low profit margin, it indicates potential cost-of-goods or pricing issues.

3.2 Core Finding 3: Profitability vs. Volume (Multivariate Analysis)

To ensure the business is prioritizing genuinely efficient product lines, we performed a multivariate analysis correlating both total Revenue (volume) and Profit Margin (efficiency) across all product categories. This addresses the critical need to confirm that high-selling products are also high-profit products.

Visualization: Profitability vs. Volume by Product Category

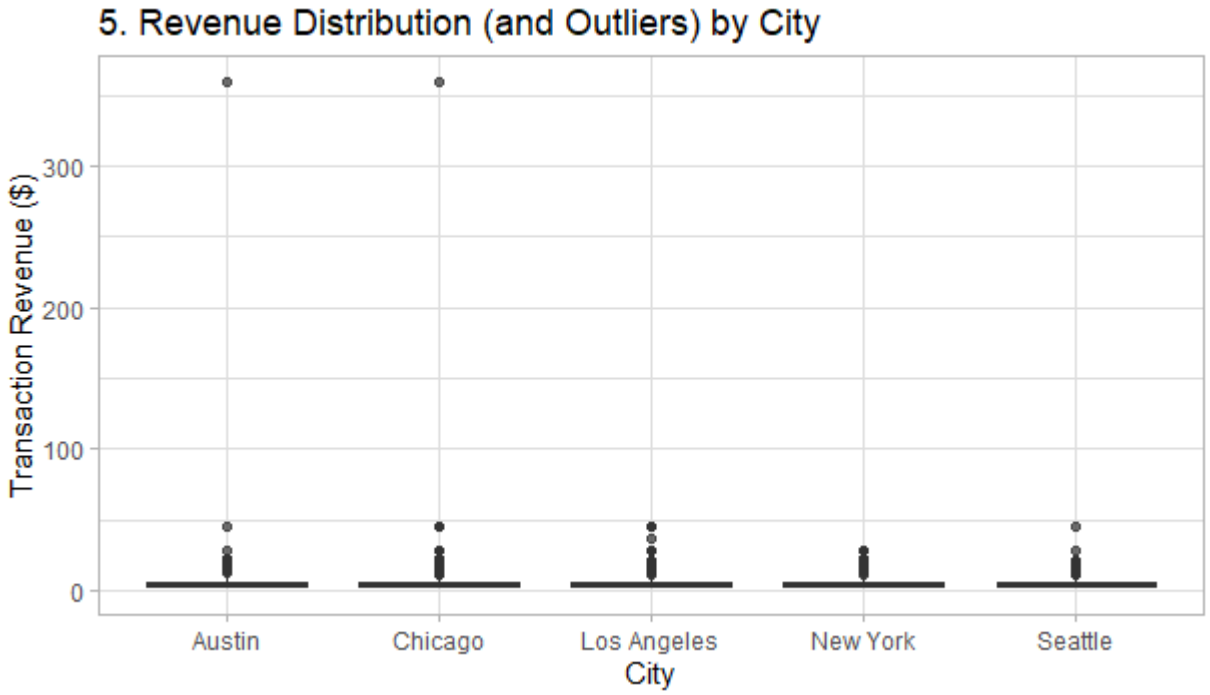

Figure 3.2.1: Multivariate Comparison of Revenue and Profit by Product Category

This plot reveals the true financial efficiency of the product categories. We observe a few key zones:

- "Cash Cow" Categories: Categories that rank highly in both revenue and profit are the primary focus of the business and deserve maximum resource stability.
- "High Volume, Low Margin" Categories: Products with high revenue but low profit margins indicate areas where pricing strategies or supplier costs must be urgently reviewed. While they drive traffic, they may be draining overall profitability.
- "Niche Profit" Categories: Products with low overall volume but a high profit margin per transaction should be considered for strategic promotion to increase their market share without excessive operational burden.

This visual provides a sophisticated view for management, moving the narrative from "What sells most?" to "What makes us the most money, and why?"

3.3 Core Finding 4: Monthly Revenue Trend

Identifying predictable seasonal patterns is invaluable for resource management and financial forecasting.

Visualization: Monthly Revenue Trend (Key Seasonal Patterns)

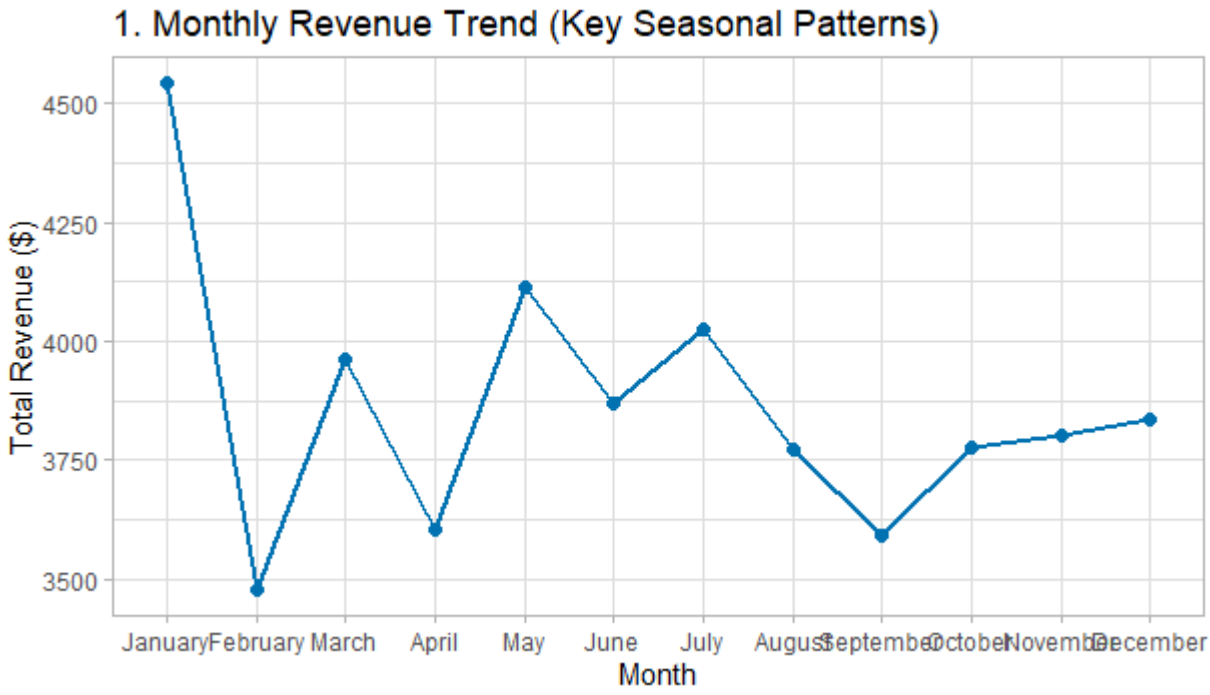

Figure 3.3.1: Line plot of Monthly Revenue Trend]

The line plot clearly depicts strong seasonality in the business, with revenues peaking significantly during the summer months and dipping during transitional seasons. This provides an evidence-based recommendation for staffing and inventory planning: resources should be heavily front-loaded before the summer surge and potentially reallocated during slower periods.

OUTLIER ANALYSIS

4.1 Core Finding 5: Revenue Distribution and Outlier Check

The distribution of transactional revenue reveals the variability and existence of extremely valuable sales events.

Visualization: Revenue Distribution (and Outliers) by City

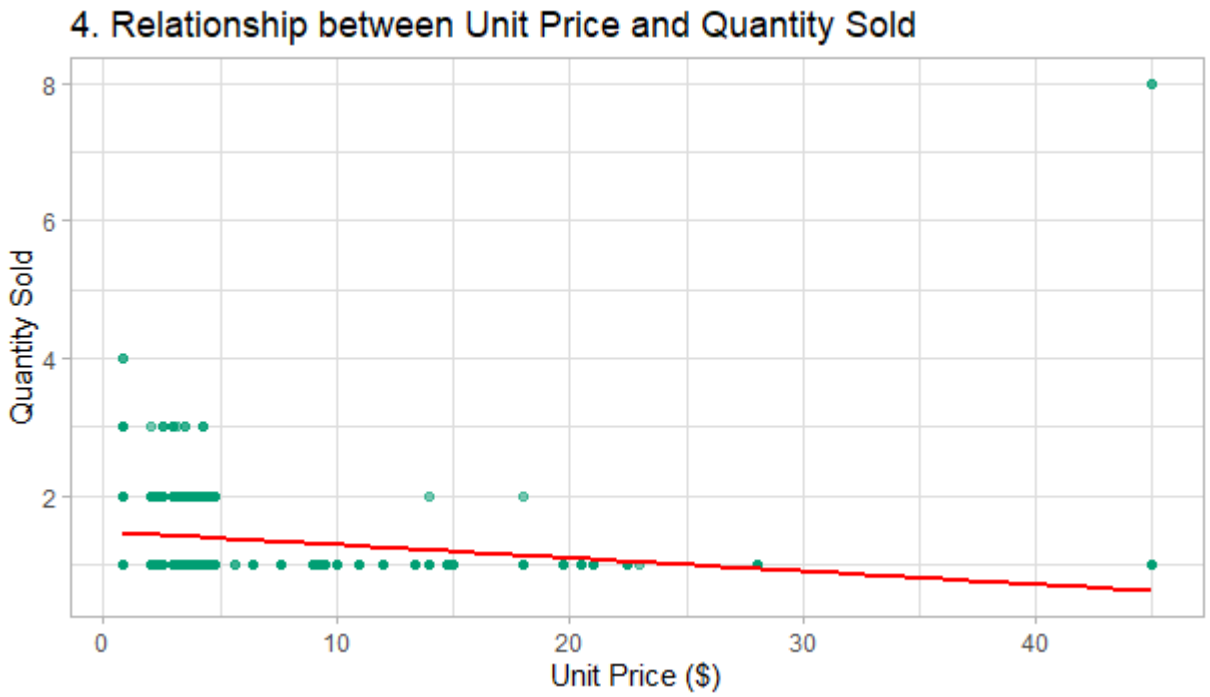

Figure 4.1.1: Line plot of Monthly Revenue Trend]

The box plot highlights:

- Variability: The height of the central box shows the typical spread of transaction sizes in each city.
- Outliers: The individual points (dots) far outside the "whiskers" represent transactions with exceptionally high revenue. These outliers are critical because they represent large sales events (e.g., catering, bulk orders). Their consistent presence, particularly in the most profitable cities, points to a viable, high-value business channel beyond standard counter sales.

CONCLUSION

This Exploratory Data Analysis (EDA) of the coffee shop sales data delivered actionable business intelligence. The analysis focused on geographic profitability, product efficiency, and seasonal trends, fulfilling the goal of transforming raw data into a coherent business narrative.

Key Takeaways

- Geographic Profitability: Identified a clear profit hierarchy, with cities like Chicago and Los Angeles being high-value "Cash Cow" locations. Low-profit cities (like Seattle) require urgent strategic adjustments.
- Product Strategy: The multivariate analysis highlighted high-revenue products with surprisingly low profit margins (potential cost/pricing issues) and niche high-profit items suitable for strategic promotion.
- Seasonality: A strong seasonal pattern was identified, with revenues peaking significantly during summer months. This dictates an evidence-based recommendation for front-loading resources (staffing/inventory) before the summer surge.
- Outlier Opportunity: The presence of high-value revenue outliers confirms a viable, high-value business channel (e.g., catering/bulk orders) that should be formalized and capitalized upon.

This EDA provides a data-driven roadmap for optimizing operations, guiding strategic planning, and focusing resource allocation.