2025

ANALYSIS AND VISUALIZATION REPORT

Table of Contents

Introduction

This milestone focuses on performing exploratory data analysis (EDA) and visualizations for the Walmart dataset. The goal is to uncover hidden patterns, trends, and relationships that will guide machine learning development in the next phase.

We aim to understand how features behave, detect outliers, check data quality, and identify relevant variables for predictive modeling. Interactive and static visual tools (like Plotly, Seaborn, and Matplotlib) were used to support this analysis.

Objectives

- Analyze the distribution of numerical and categorical features.
- Identify relationships between key variables.
- Detect potential outliers and data imbalances.
- Provide visual insights to guide feature selection for ML models.

Data Overview

- Dataset: walmart_cleaned.csv
- Features: Mixture of numerical and categorical columns.
- Missing values: Already handled in Milestone 1.

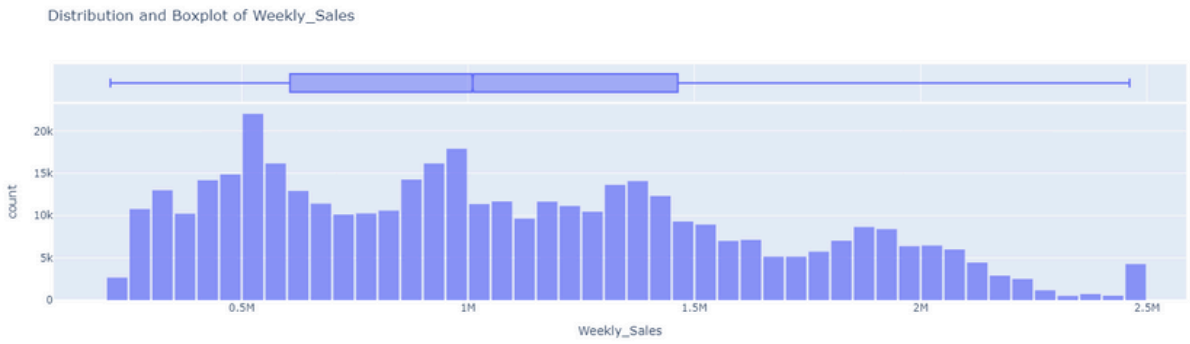
Store	Dept	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	Type	Size
1	1	1643690.90	0	42.31	2.572	211.096358	8.106	A	151315
1	1	1641957.44	1	38.51	2.548	211.242170	8.106	A	151315
1	1	1611968.17	0	39.93	2.514	211.289143	8.106	A	151315
1	1	1409727.59	0	46.63	2.561	211.319643	8.106	A	151315
1	1	1554806.68	0	46.50	2.625	211.350143	8.106	A	151315

Exploratory Data Analysis

Univariate Analysis

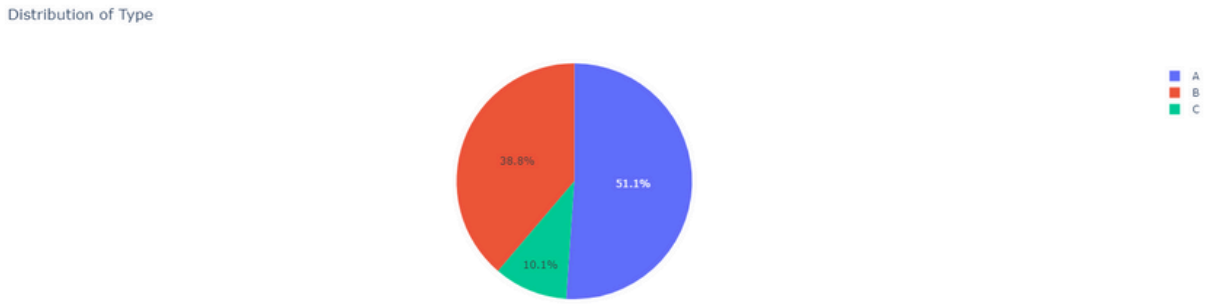
1. Numerical Features

- Histograms and boxplots were generated using Plotly to explore the distribution of numerical variables.
- Summary statistics were also printed (mean, median, std, etc.).
- This helped identify skewness, peaks, and outliers.



2 .Categorical Features

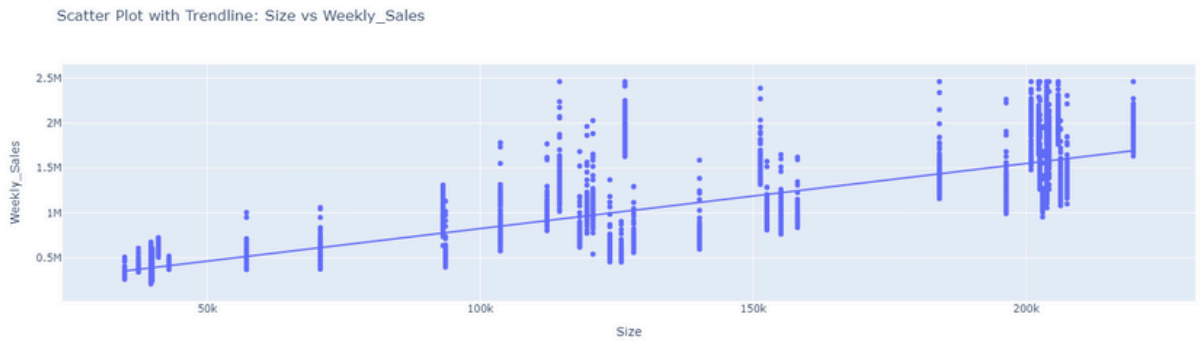
a. Value counts were computed for all categorical variables.
b. Visualizations included:
c. Pie charts for features with ≤4 unique values.
d. Bar charts for those with more.
e. Only the top categories (top 5) were displayed for better clarity.



Exploratory Data Analysis

Numerical vs Numerical

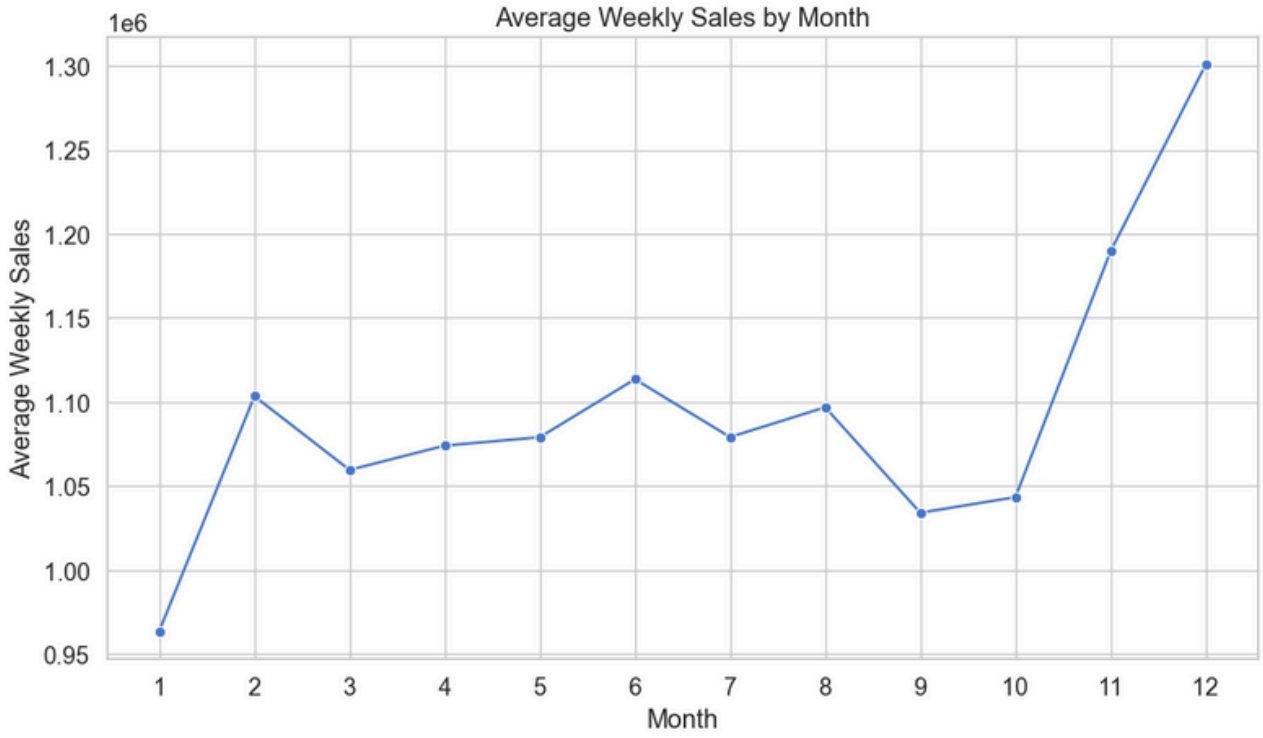
- **Scatter Plot with Trendline: Size vs. Weekly Sales**
  - Used scatter plots with trendlines (via OLS regression).
  - Also calculated and printed Pearson correlation values.
  - This helped highlight linear relationships between continuous variables.
  - Larger store sizes (e.g., 200k) tend to correlate with higher weekly sales (up to $2.5 million).
  - The trendline indicates a general positive relationship between size and sales.



- **Weekly Sales During Promo Weeks (Holiday Trend)**
  - Sales peak during weeks 2010-W47 and 2011-W47, reaching around $1.5 million, likely due to holiday promotions.
  - There are noticeable dips, such as in 2010-W52 and 2011-W36, with sales dropping to approximately $1.1 million.

Exploratory Data Analysis

- **Average Weekly Sales by Month**
  - Sales are lowest in January (around $0.95 million) and peak in November and December (around $1.3 million), reflecting a strong year-end trend, possibly due to holiday shopping.

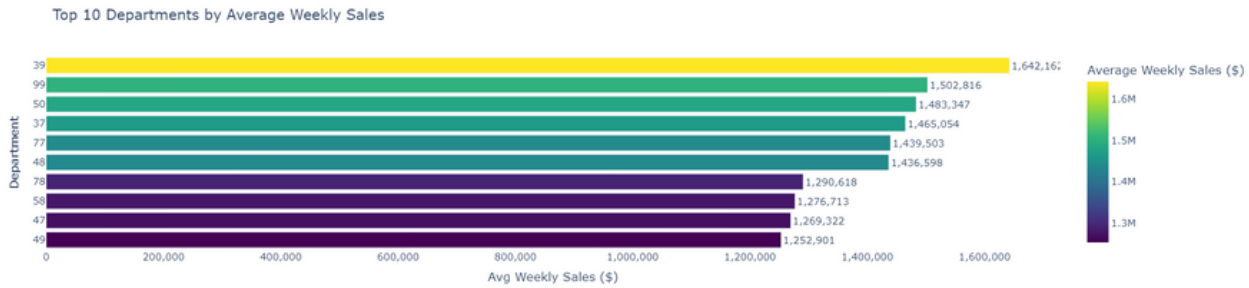


Numerical vs Categorical

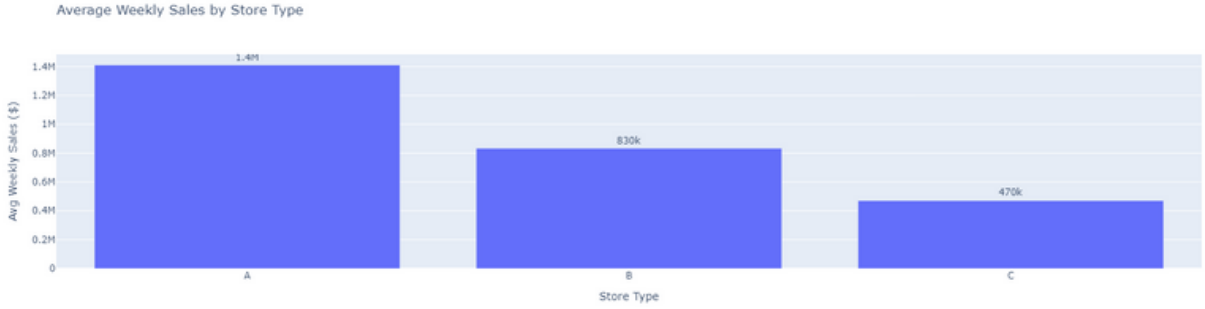
- **Top 10 Stores by Total Sales:**
  - Store 4 tops the list with $21.25 billion in total sales.
  - Store 20 is a close second with $21.23 billion.
  - Sales drop to $14.66 billion for Store 19, showing significant variation in total sales performance.

Exploratory Data Analysis

- **Top 10 Departments by Average Weekly Sales:**
  - Department 39 leads with the highest average weekly sales at $1.642 million.
  - Department 99 follows closely with $1.502 million.
  - Sales decrease progressively, with Department 49 at the bottom of the top 10 with $1.252 million.
  - There is a clear gradient, indicating a range of performance among departments.

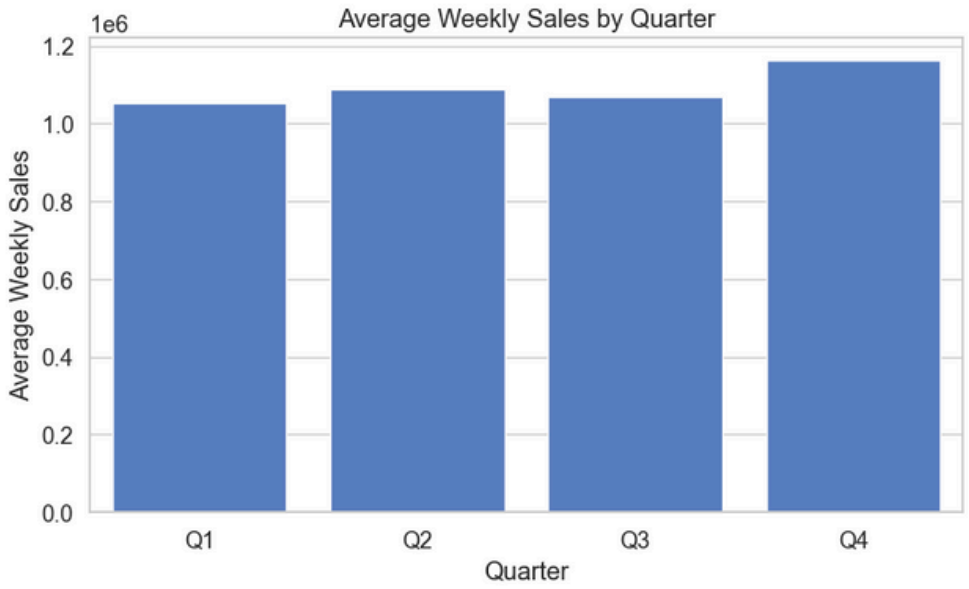


- **Average Weekly Sales by Store Type**
  - Store Type A has the highest average weekly sales at $1.4 million.
  - Store Type B follows with $830,000, and Type C is the lowest at $470,000.
  - This suggests Store Type A is the most profitable.

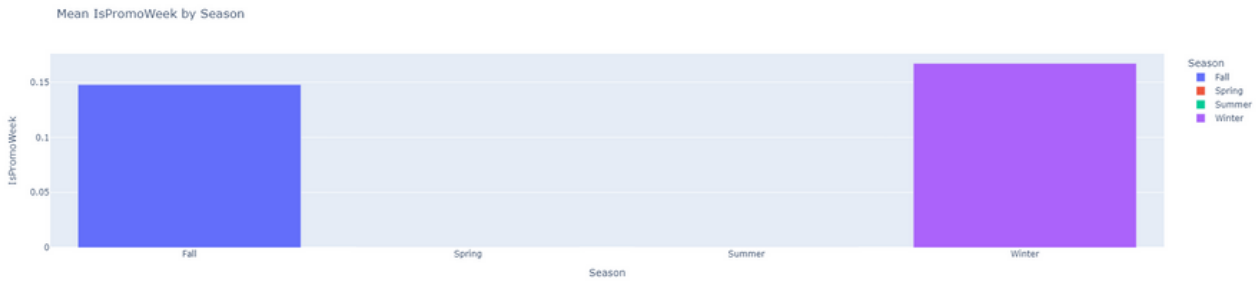


Exploratory Data Analysis

- **Average Weekly Sales by Quarter**
  - Q4 has the highest average weekly sales, slightly above $1.2 million.
  - Q1, Q2, and Q3 are relatively similar, around $1.0-$1.1 million, indicating a seasonal peak in Q4.

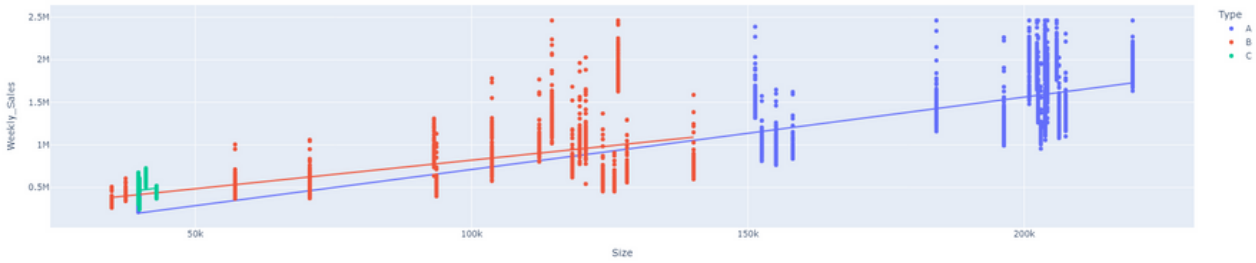


- **Mean IsPromoWeek by Season**
  - Winter and Fall have the highest mean IsPromoWeek values (around 0.15), suggesting more promotional weeks during these seasons.
  - Spring and Summer show no significant promotional activity.

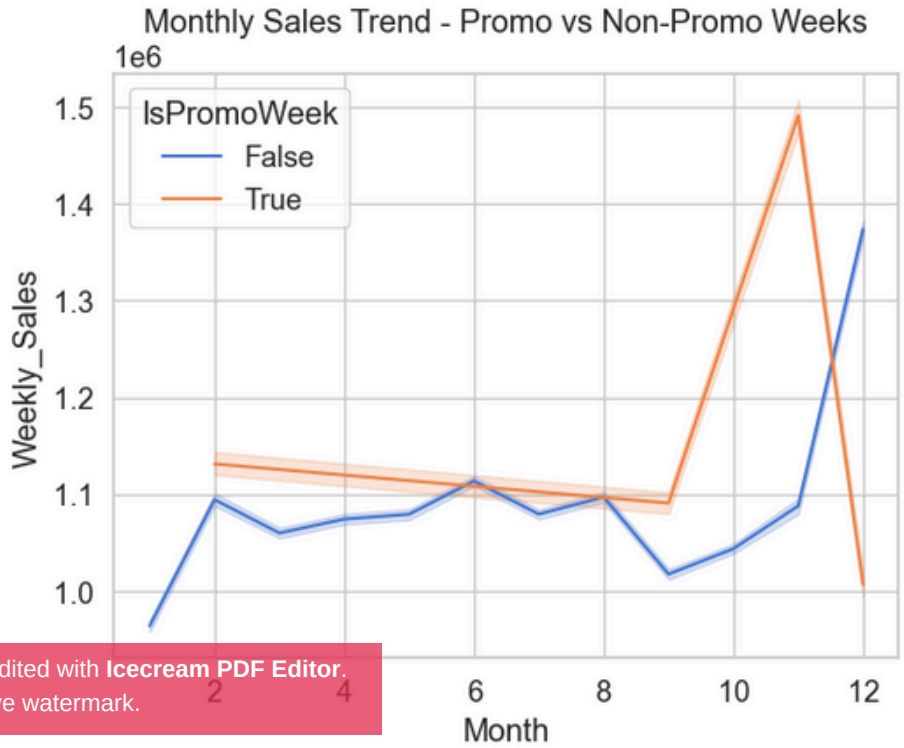


Exploratory Data Analysis

- **Scatter Plot with Trendline: Type vs Weekly Sales**
  - Type A stores show the highest sales, often exceeding $2 million.
  - Type B and C stores have lower and more variable sales, with Type C clustering around $0.5-$1 million.
  - The trendlines suggest Type A outperforms others consistently.



- **Monthly Sales Trend - Promo vs Non-Promo Weeks**
  - Promo weeks (True) show a significant spike in December, reaching $1.45 million, compared to non-promo weeks (False) at $1.3 million.
  - Non-promo weeks are more stable, ranging between $1.0-$1.1 million, while promo weeks vary widely.



Key Insights

- Some features are right-skewed and contain outliers.
- Sales tend to drop as temperature rises in some stores.
- There’s a meaningful difference in sales between holiday and non-holiday periods.
- Strong correlations were found between some variables (e.g., CPI and Fuel_Price).

Tools and Libraries

- Pandas and NumPy for data handling.
- Plotly, Seaborn, and Matplotlib for visualizations.

Summary

The EDA phase provided critical insights that will improve model accuracy in Milestone 3. We now understand which features are most useful, which need transformation, and what types of relationships exist in the data.

We thank you for your continued support