MOHAB KAMAL &
YOUSSEF TAHER

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

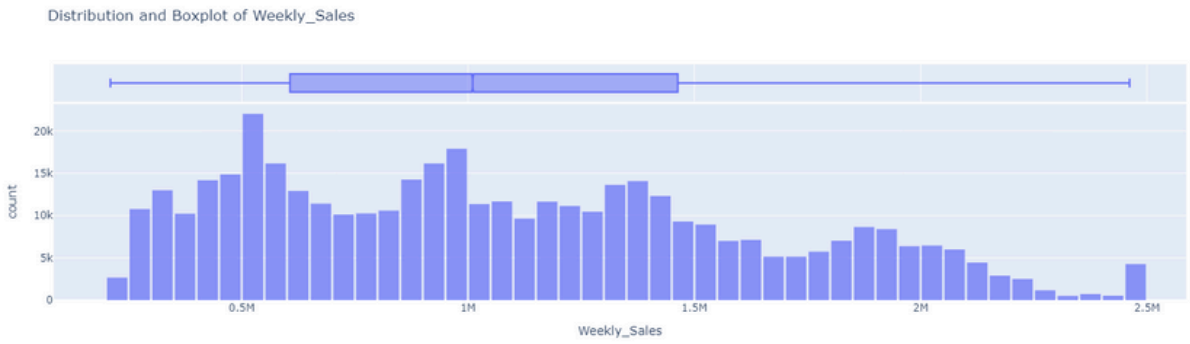
- Dataset: walmart_cleaned.csv
- Features: Mixture of numerical and categorical columns.
- Missing values: Already handled in Milestone 1.

| 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

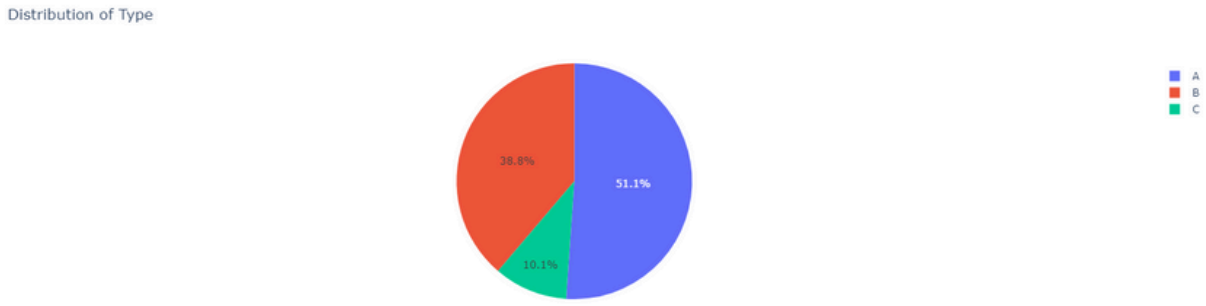
## Univariate Analysis

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**

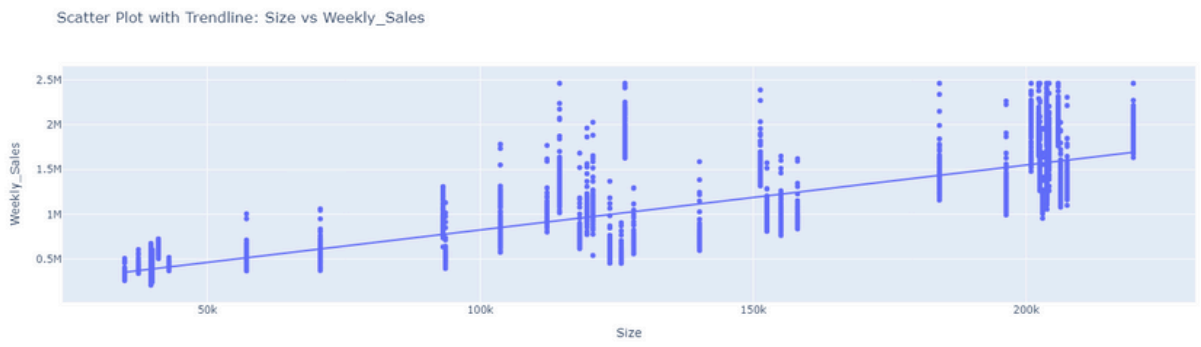
a. Value counts were computed for all categorical variables.
b. Visualizations included:
c. Pie charts for features with ≤4 unique values.
d. Bar charts for those with more.
e. Only the top categories (top 5) were displayed for better clarity.



# Exploratory Data Analysis

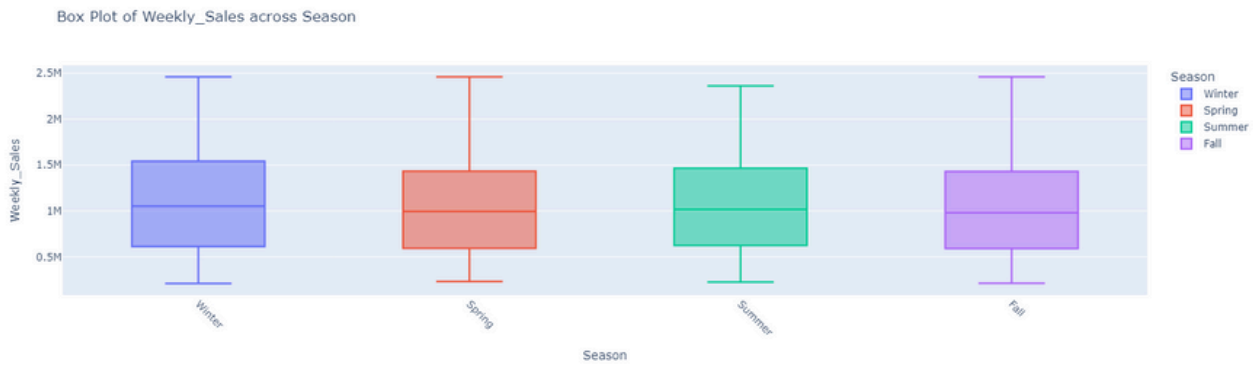
## Bivariate Analysis

1. **Numerical vs Numerical**
    - Used scatter plots with trendlines (via OLS regression).
    - Also calculated and printed Pearson correlation values.
    - This helped highlight linear relationships between continuous variables.




2 .**Numerical vs Categorical**

- Boxplots were used to visualize how numerical data varies across categories.
- A warning was displayed for categorical features with >10 unique values to avoid clutter.



# 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.
- Dash, Dash Bootstrap Components for dash.

# 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