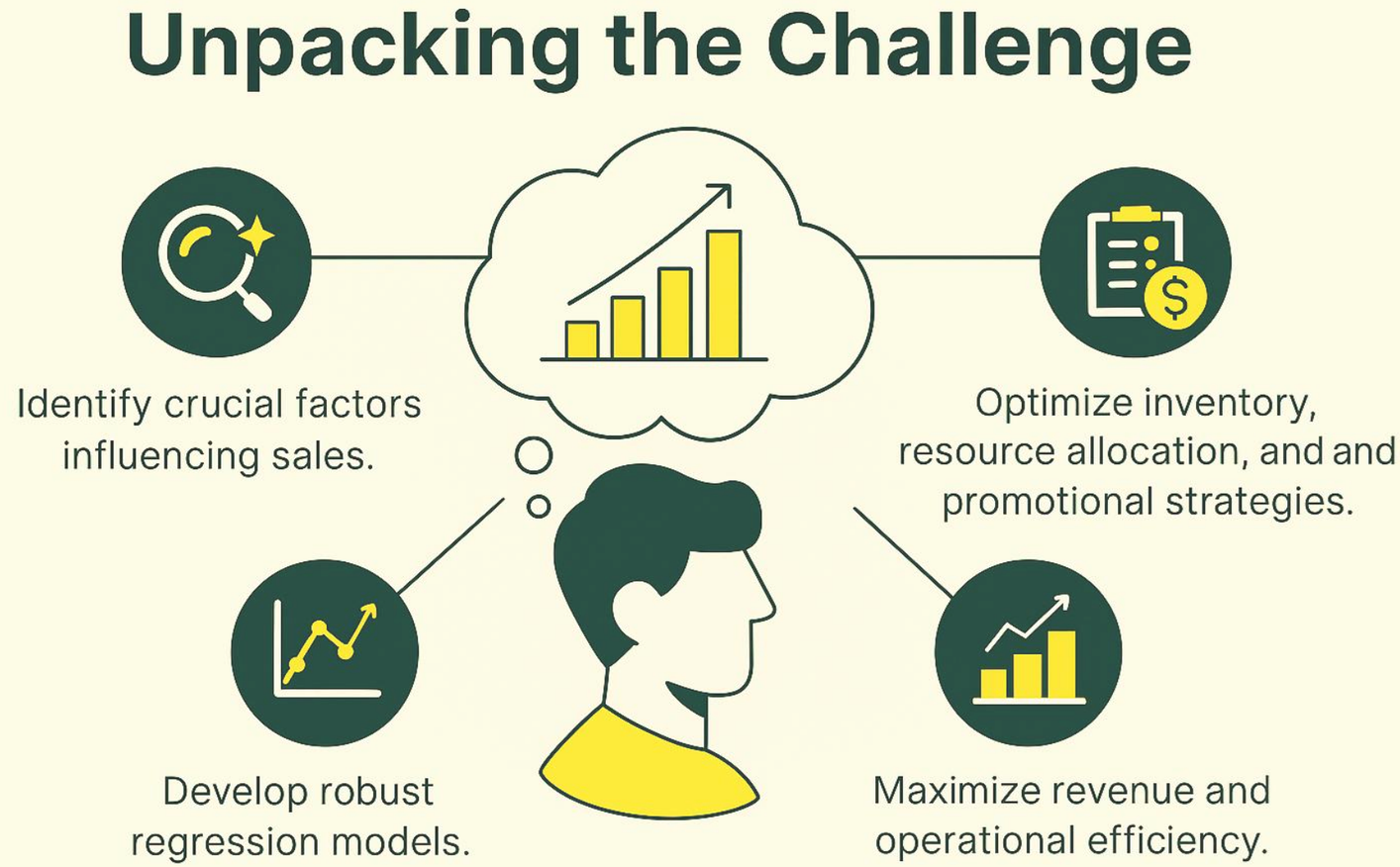
Store Sales Forecasting Project

Welcome to the Store Sales Forecasting Project report. This presentation outlines our approach, findings, and the robust predictive model developed to optimize sales strategies.

Project Overview

Our primary objective was to build a predictive framework for weekly sales across stores. By leveraging historical data and store- specific details, we aimed to:



Data Foundation

The Datasets: A Closer Look

The project utilized three core datasets, each providing unique insights into store operations and sales performance.

Store_Details

- 45 rows, 6 columns
- Unique store attributes (ID, Type, Address, Area Code, Location, Size).

Business_Data

- 8,190 rows, 15 columns
- Weekly records: Temperature, Fuel Price, MarkDowns 1-5, CPI, Unemployment Rate, Holiday status, and engineered time features.

Sales_History

- 421,570 rows, 8 columns
- Transactional data: Store, Department, Date, Total Sales, Holiday, Year, Month, Weekday. Advanced features like lags and rolling means were engineered.

Addressing Missing Values & Outliers

Missing Values

- Store_Details
- Business Data: Filled with zero
- No missing values in Sales Histstory

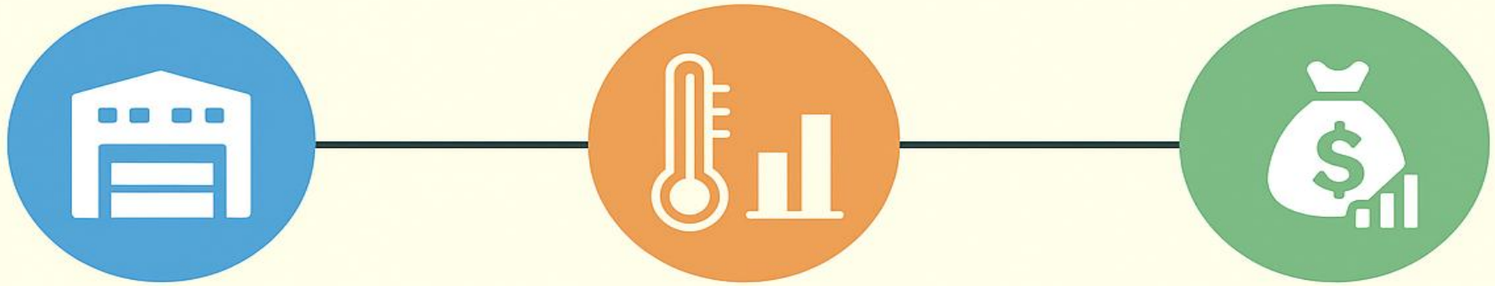
Outliers

- Outliers: Sales, Temperature, Fuel Price, CPI, Unemployment Ratc.
- Fixed with IQR Method

Exploratory Data Analysis

Understanding the raw data's characteristics provided a foundation for feature engineering and model selection.

Key Insights – Descriptive Statistics



Store_Details

- 22/45 are E-Commerce
- Size: 34.8k – 219.6k sq ft
- Avg: 126.5k sq ft

Business_Data

- Temp: 4°F – 102°F (Avg 59°F)
- Discounts rare (mostly zero)
- CPI Avg: 173 (126–229)
- Unemp: 7.7% (4.3–11%)

Sales_History

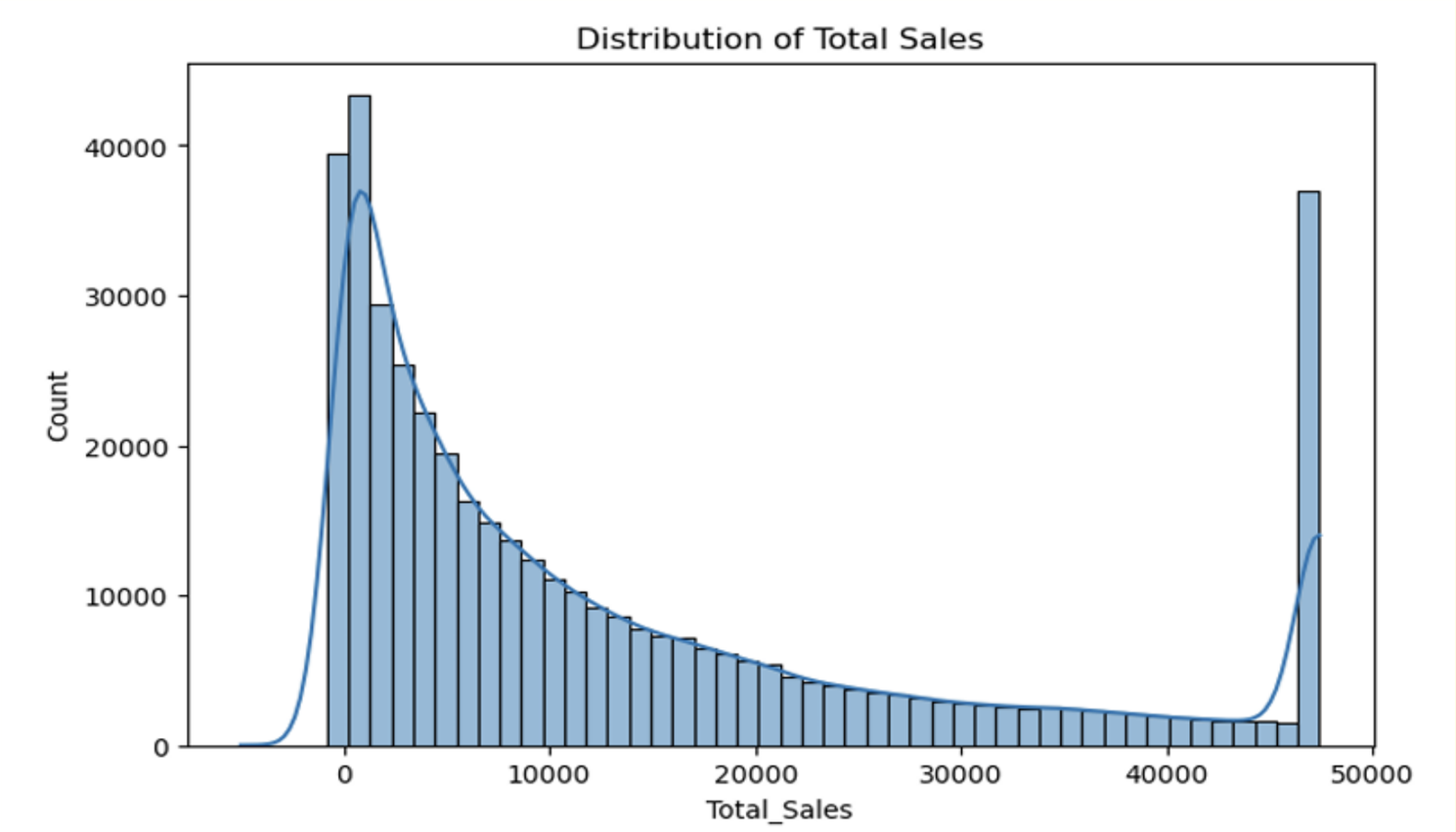
- Weekly Sales: -5k to 4l k (Avg 13.6k)
- 81 Depts, 45 Stores
- Mostly non-holiday weeks

Visual Discoveries

Sales Patterns & Relationships

Visualizations helped us uncover critical sales dynamics and variable correlations.

- **Sales distribution:** Most weeks had low or medium sales, but there were some big spikes during holidays or promotions.

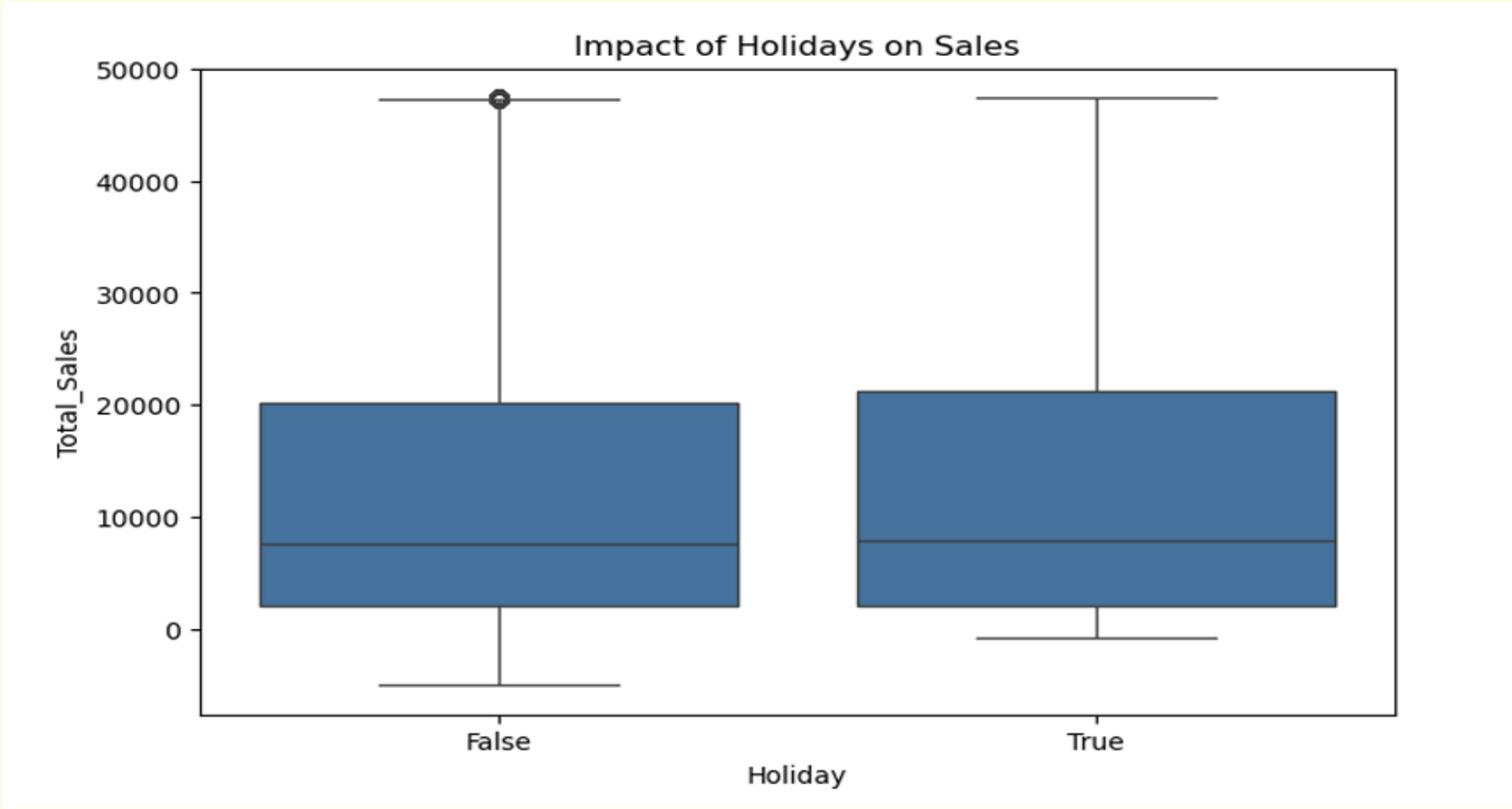


Visual Discoveries

Sales Patterns & Relationships

Visualizations helped us uncover critical sales dynamics and variable correlations.

- **Holiday effect:** Sales were much higher on holidays compared to normal weeks.



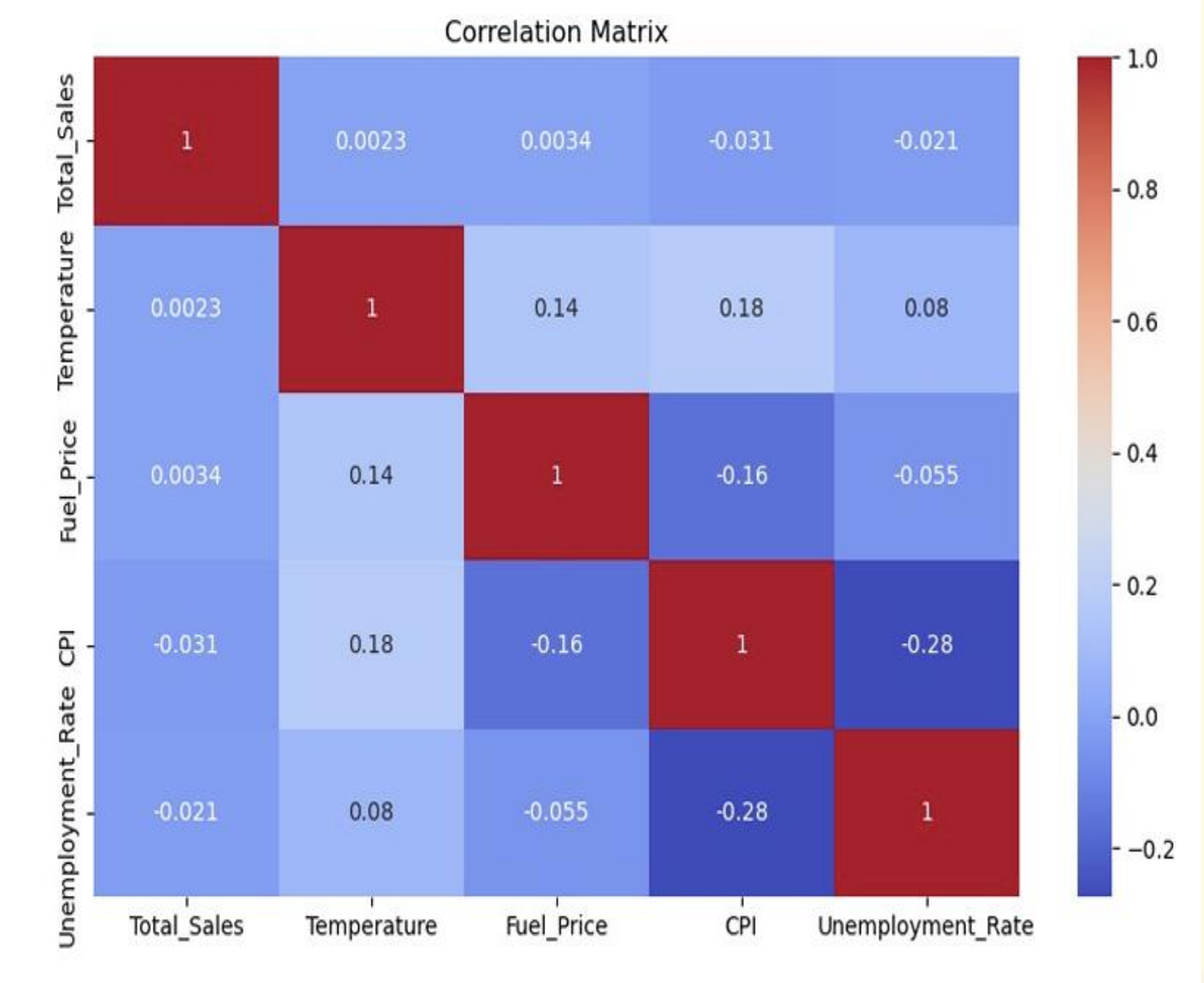
Visual Discoveries

Sales Patterns & Relationships

Visualizations helped us uncover critical sales dynamics and variable correlations.

- **Correlation findings:**

  - Bigger stores → more sales (moderate).
  - Discounts, CPI, unemployment → weak effect.
  - MarkDown1 & MarkDown4 → strongly linked.

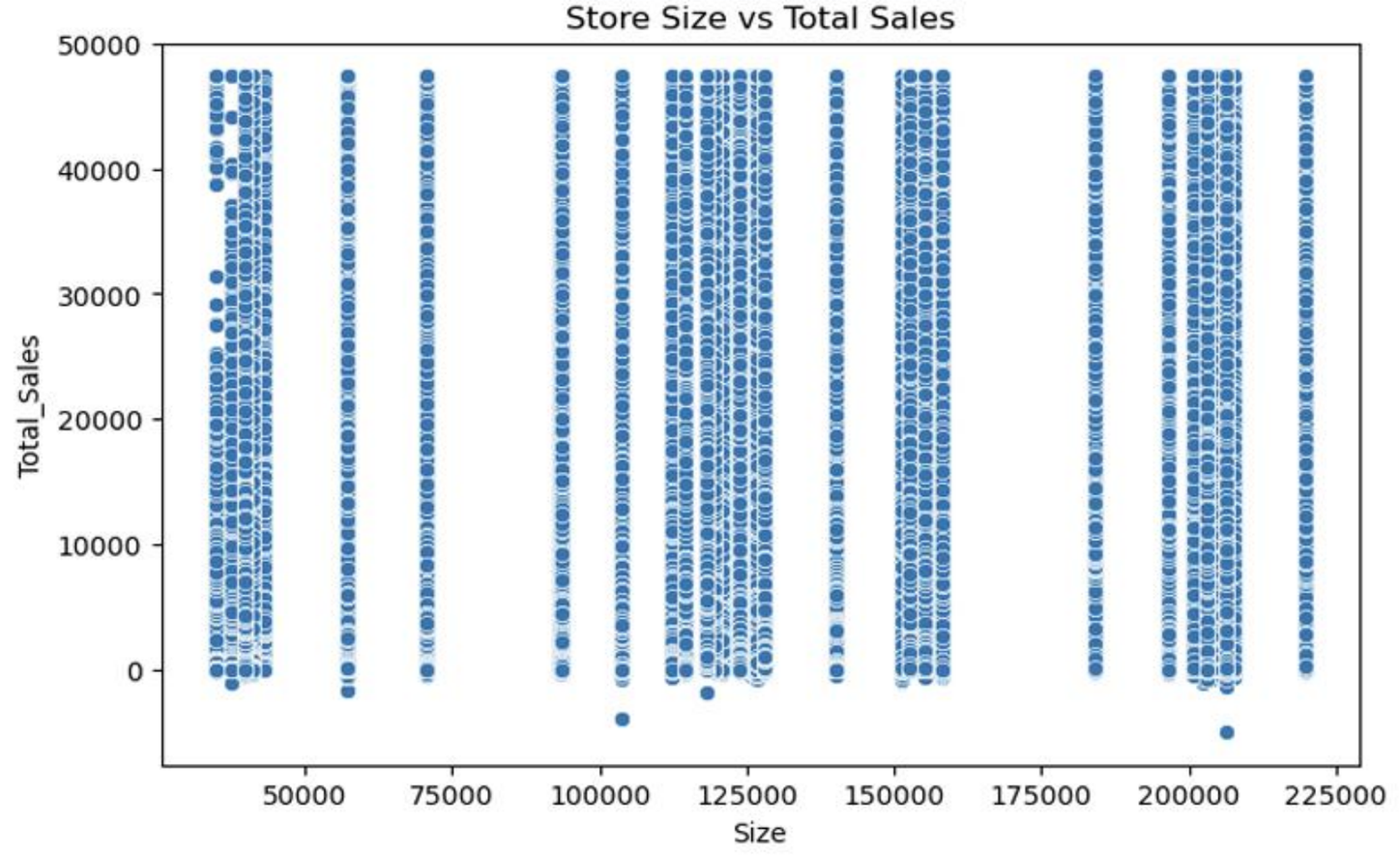


Visual Discoveries

Sales Patterns & Relationships

Visualizations helped us uncover critical sales dynamics and variable correlations.

- **Scatter plots:** Showed that large stores always earn more, while economic factors and discounts don’t show a clear impact.



Feature Importance

The Predictive Power of Variables

Information Value (IV) analysis highlighted which variables are most crucial for predicting sales.

1

Holiday

HIGH IV (>0.4)

Significant impact: Holidays increase sales.

2

Store Size

MODERATE IV (0.25–0.3)

Bigger stores consistently sell more.

3

MarkDowns

LOW IV (<0.1 individually)

Only large, infrequent discounts show notable impact.

4

Economic Variables

LOW IV

(CPI, Unemployment): Effects present but weak compared to holiday/promotions.

Model Development

Benchmarking Regression Models

We evaluated several regression models, assessing their performance using RMSE (Root Mean Squared Error) and R^2 (Coefficient of Determination).

Linear Regression	~0.45	~8,900	Baseline, interpretable
Random Forest	~0.62	~7,550	Robust, captures non-linearities
Gradient Boosting	~0.64	~7,350	Slightly better, slower
XGBoost	~0.66	~7,200	Best performing, flexible

Feature importance from tree models consistently highlighted Holiday, Store Size, and certain MarkDowns as top contributors.

The Chosen Model: XGBoost Regression

After thorough evaluation, XGBoost emerged as the optimal choice for our sales forecasting framework.

Why XGBoost?

- Superior Performance: Achieved the highest test R^2 and lowest RMSE on validation data.
- Robustness: Handles interactions and non-linear relationships better than other models.
- Flexibility: Adapts well to feature scaling, collinearity, and works effectively with engineered features.

The Chosen Model:

Summary & Outlook

Conclusion & Future Opportunities

This project successfully delivered a robust sales prediction model, offering valuable insights for Store’s strategic planning.

Key Achievements

- Integrated diverse datasets for comprehensive analysis.
- Quantitatively modeled holiday and promotional effects.
- Developed a high-accuracy XGBoost model.
- Provided actionable insights for optimizing inventory, resources, and promotions.

Challenges Overcome

- Ensured proper data merging without leakage.
- Effectively managed significant missing data in MarkDowns.
- Handled computationally intensive feature engineering.
- Judiciously treated outliers to preserve event-driven sales spikes.
- Optimized hyperparameter tuning for boosting models.

THE END