Seasonal Sales Forecasting:

Walmart Sales

Table of Contents

Background

Problem

Leveraging Data for Enhanced Retail Strategy - The Walmart

- Retail companies leverage data as a pivotal asset to predict consumer behavior
    - Useful for predicting buying patterns, future sales, and formulating promotional plans
- Data analysis facilitates the prediction of future sales and understanding customer behavior
- Strategic planning based on data insights enhances profitability and competitiveness

Goal

- Predict sales for Walmart based on historical data (2010-2013)
    - One of world's largest retail companies
- Incorporate additional factors like **Walmart department type, CPI, temperature, fuel price, holiday time period, promotional markdowns, annual season, and unemployment rate** to understand their impact on weekly sales.
- Dive into potential relationship between *season of the year* and sales
- Overall objective is to apply analysis techniques to provide data-driven insights for predictive sales modeling for Walmart.

Walmart

Hypothesis: We hypothesize that there will be an increased number of sales during the winter season at Walmart. This expectation is driven by factors such as heightened consumer demand and spending influenced by the winter holiday season and strategic holiday promotions.

Dataset

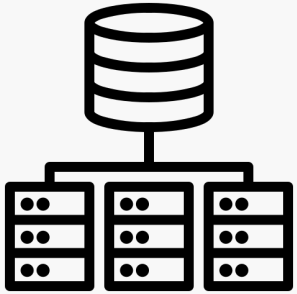
- "**Walmart Store Sales Forecasting**" - Kaggle Dataset
  - Collected between 2010-2013 for 45 stores
  - Train.csv
  - Features.csv
  - Stores.csv
  - Test.csv
- Contained many features allowing us to accurately draw conclusions and understand the relationship between variables that can affect sales
  - ex: unemployment and cpi
- Well-structured data

Raw Training Data

	Store	Dept	Date	Weekly_Sales	IsHoliday
0	1	1	2010-02-05	24924.50	False
1	1	1	2010-02-12	46039.49	True
2	1	1	2010-02-19	41595.55	False
3	1	1	2010-02-26	19403.54	False
4	1	1	2010-03-05	21827.90	False

Shape: (421570 rows, 5 columns)

Dataset



- Features
  - **Date** : categorical, 02/2010 - 11/2012
  - **Store:** categorical, 45 stores total
  - **Weekly_Sales** : numerical continuous
  - **Department:** categorical
  - **IsHoliday :** binary - is true if there is a holiday within the week
  - **Season:** categorical, Winter, Spring, Summer, Fall
  - **Temperature:** numerical continuous
  - **Fuel_Price:** numerical
  - **MarkDown1-5:** numerical continuous
  - **CPI:** numerical continuous
  - **Unemployment:** numerical continuous
  - **Type:** categorical, Superstore, Discount Store, Neighborhood Market
  - **Size:** numerical continuous

ML Applications of this Dataset

- **Human Management/Resource Allocation**:
  - Insights from this study enable strategic resource allocation based on regional demand and profitability consideration
  - Also allows stores to increase staffing during busy periods
- **Seasonal Sales Analysis:**
  - Studying sales patterns based on the season can allow for more promotional offers
- **Revenue Protection:**
  - Helps the company reach predicted seasonal targets which can have a positive effect on stock prices
- **Inventory Management**
  - Insights into store size assist in optimizing inventory levels to meet demand.
  - Facilitates better control over stock, minimizing both excess and stockouts.
- **Customer Retention**
  - Understanding store sales in creating personalized shopping experiences.
  - Optimizing inventory based on store size contributes to customer satisfaction and retention.

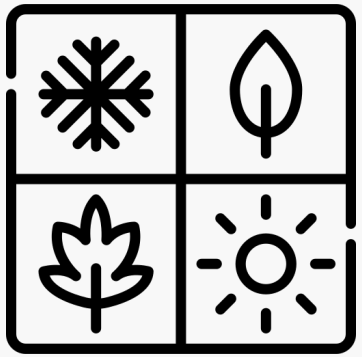
Data Cleaning

Data Cleaning - Preprocessing

- **Type Column:**
  - Create a mapping for the 'Type' column where "A" is mapped to 3, "B" to 2, and "C" to 1.
- **Date Column:**
  - Convert the 'Date' column in the DataFrame to datetime format using the specified format '%Y-%m-%d'.
- **Extract Year Column:**
  - Extract the year from the 'Date' column and create a new 'Year' column.
- **Normalization:**
  - Before running the model used the StandardScaler from scikit-learn to standardize the numerical columns

Data Cleaning - Season

- **Generate Season Column:**
  - Extract the month from the 'Date' column and create a new **'Season'** column:
    - *Spring*: March, April, May
    - *Summer*: June, July, August
    - *Autumn*: September, October, November
    - *Winter*: December, January, February



- **One-Hot Encode Season Column:**
  - perform one-hot encoding on the 'Season' column using pd.get_dummies() to create binary columns for each season.

Data Cleaning – isHoliday

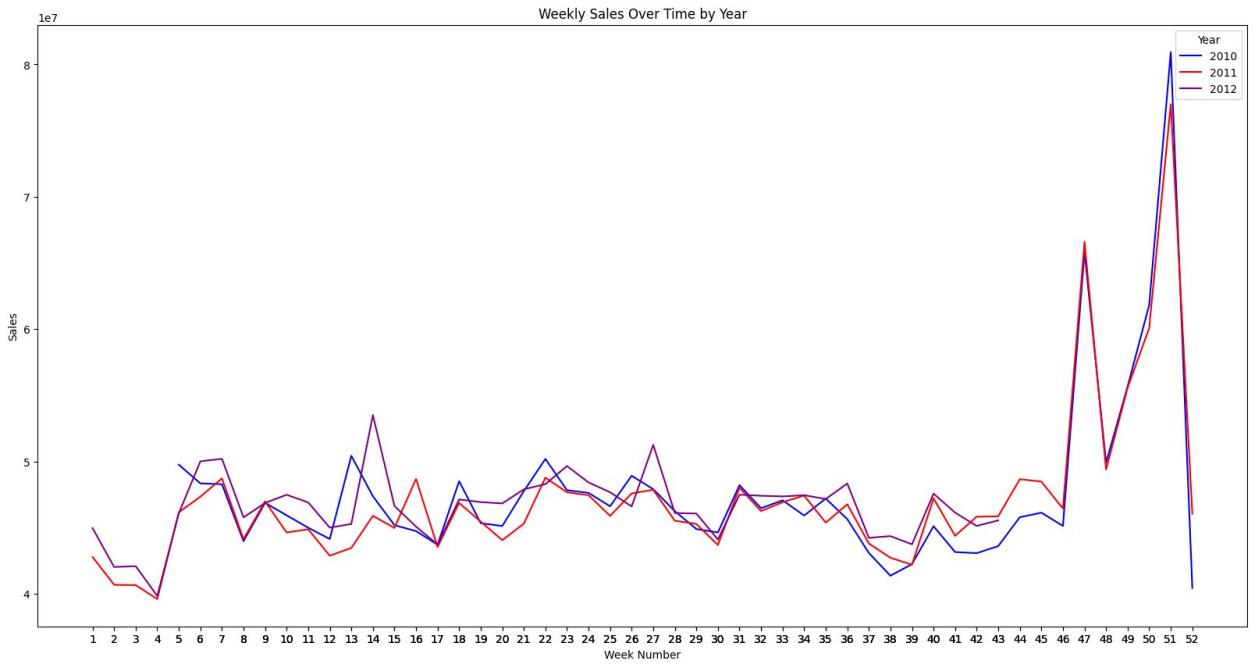
- **Currently only 4 holidays included:**
  - Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13
  - Labor Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13
  - Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13
  - Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13
- **Added other commonly known promotional days such as:**
  - July 4th
  - Valentine's Day
  - New Years
  - Back to School
  - Mother's Day
  - Halloween
  - Hanukkah

Visualizations

Exploratory Data Analysis

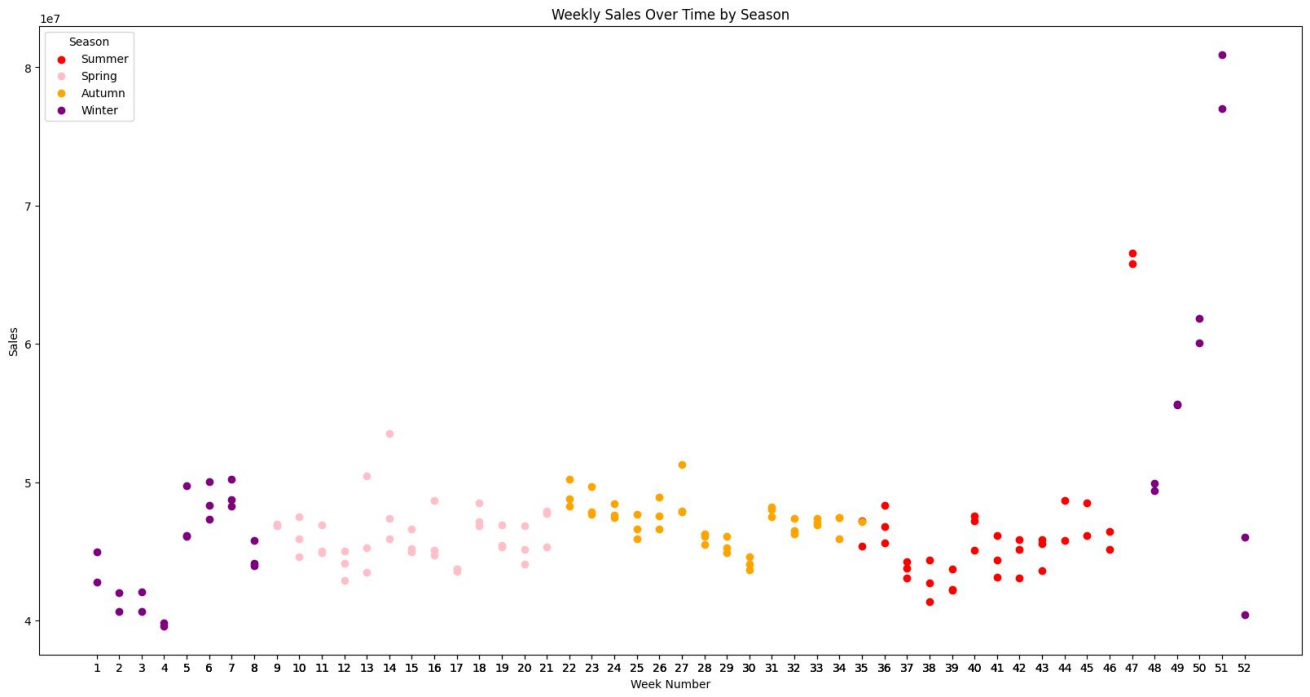
- **Goal**: Understand the correlation between different features and Walmart sales in the dataset through EDA visualizations.
- Exploration Highlights:
  - **Yearly Sales Overview**
    - Weekly sales across the different years for general understanding
  - **Seasonal Insights**
    - Weekly sales per season
  - **Store Type**
    - Impact of store type on average weekly sales
  - **Holiday Sales Growth**
    - Impact of holiday on sales growth
  - **Unemployment Rate Impact on Sales**
    - Impact of unemployment rate on sales patterns
  - **Correlation Analysis**
    - Overall correlation matrix to view relationships between all features

Weekly Sales Over Time by Year



- Increase in sales around **weeks 47 to 51** (potentially due to Thanksgiving, Black Friday, and Christmas shopping)
- Insufficient data during first few weeks of year 2010

Weekly Sales by Season



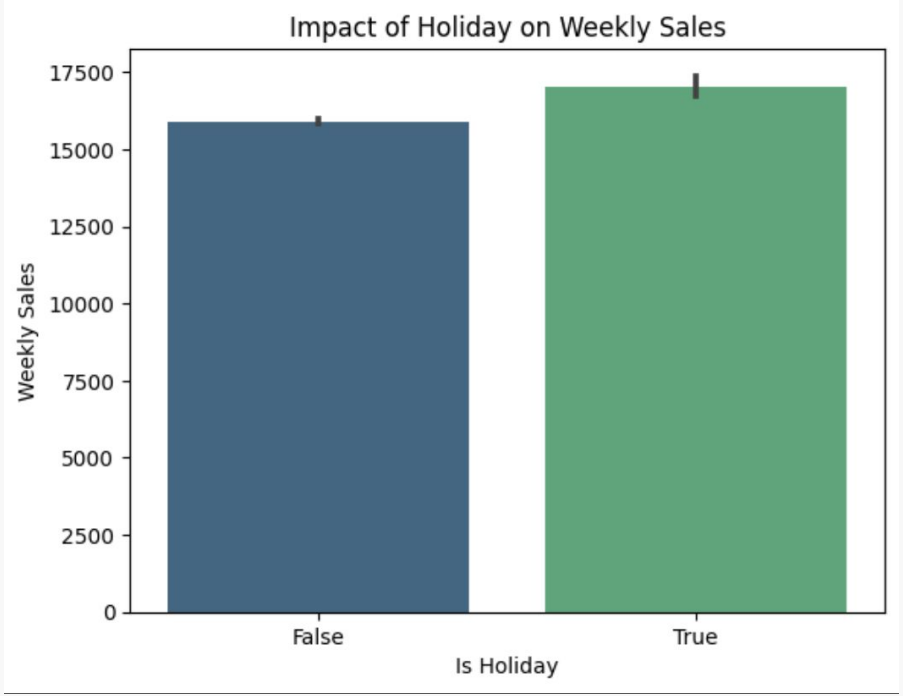
Type of Store Impact on Sales

Store Type Assumption:

- **Type A:** Walmart Superstore
- **Type B:** Walmart Discount Center
- **Type C:** Walmart Neighborhood Market

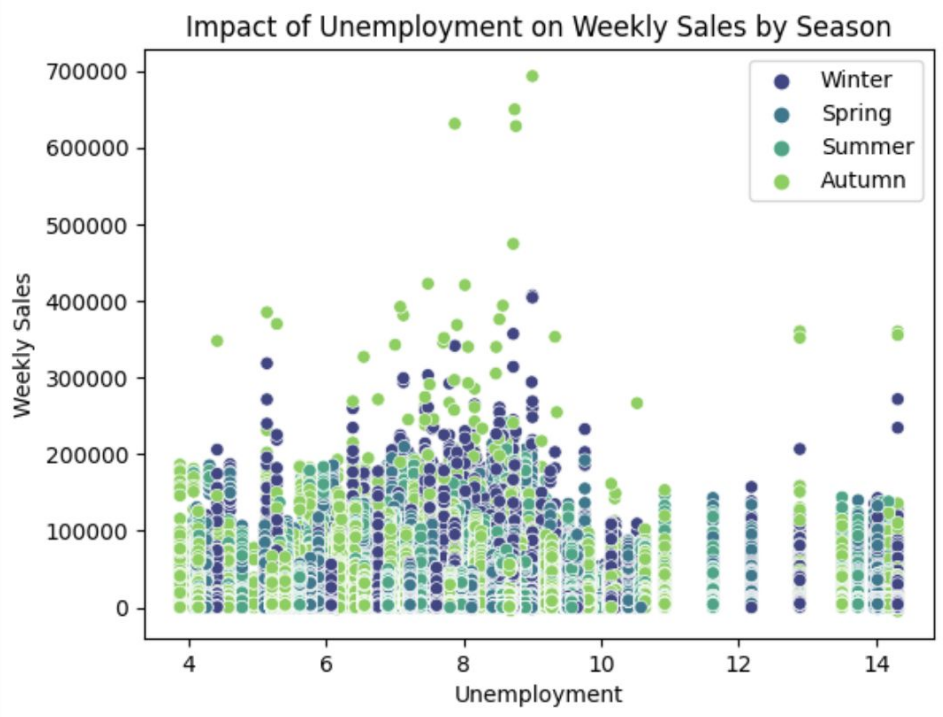
Conclusion from graph: Walmart Supercenter seems to historically have the highest number of sales.

Holiday Impact on Sales



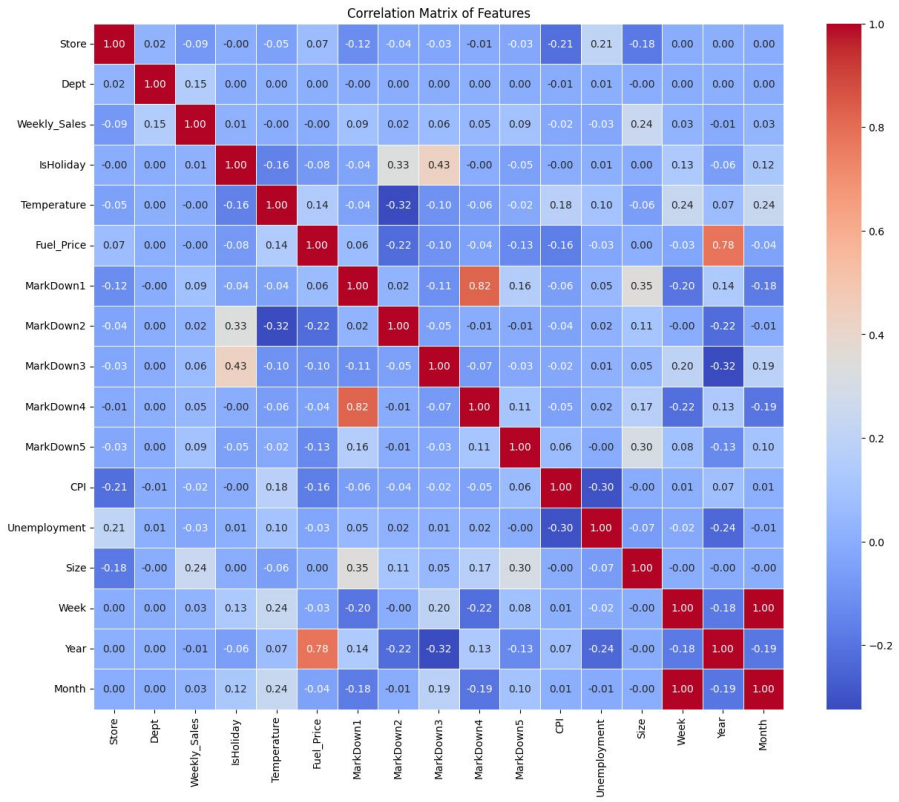
- Holiday sales seem to be slightly higher than non-holiday sales
  - Confirms the idea that customers are more likely to shop during holiday weeks
  - Possible contributors include promotional activities, festive discounts, and seasonal marketing campaigns.

Unemployment Rate Impact on Sales



- No strong conclusions can be drawn from the graph
- No clear trend with data

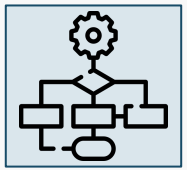
Correlation Matrix



- Pearson's Correlation Coefficient:

$$r = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

- Can see correlation between each feature and sales (only applicable non-categorical variables)
  - *Weekly_Sales* and *Size* have greatest correlation



Model

WMAE - Weighted Mean Absolute Error

$$\text{WMAE} = \frac{1}{\sum w_i} \sum_{i=1}^{n} w_i \, |y_i - \hat{y}_i|$$

- w_i are weights, w = 5 if holiday and 1 if not holiday
- y_i - Actual Weekly Sales
- $\hat{y}_i$ -Predicted Weekly Sales
- n - number of observations
- **WHY WMAE?**
  - Commonly used as an error metric for forecasting
  - Handles seasonality/special events
  - Can align with business objective such as when to set promotions
  - Works well on Skewed Data Distribution

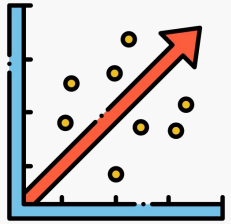
RMSE – Root Mean Squared Error

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

- yi - actual sales
- $\hat{y}$ i - predicted sales
- n - number of observations
- **WHY RMSE as our metric?**
  - Widely accepted in regression, ensuring industry compatibility
  - Sensitive to large errors, crucial for applications where outliers significantly impact results
  - Mathematically straightforward and Similar Unit of Measurement

Linear Regression

- Used for numeric prediction when the relationship between the dependent variable and multiple independent variables is assumed to be linear and representable as a linear combination
- Parameters:
  - fit_intercept = True
  - positive = False
- Not Efficient Model:
  - The WMAE values for the validation data is extremely high
  - Impacted by outlier data



WMAE:

- All Seasons: 14564
- Spring: 14107
- Autumn: 14464
- Summer: 14474
- Winter: 15267

RMSE:

- All Seasons: 21786
- Spring: 20540
- Autumn: 22210
- Summer: 20739
- Winter: 23018

Bayesian Ridge Regression

- Statistical method used for numeric prediction modeling
- Uses regularization, it is robust to outliers, and incorporates prior information
- Parameters:
  - alpha_1=0.1,
  - lambda_2=0.1
- Not Efficient Model:
  - Behavior is identical to Linear Regression
  - dependent on linearity and linear combination of factors is causing high errors

WMAE:

- All Seasons: 14564
- Spring: 14106
- Autumn: 14463
- Summer: 14474
- Winter: 15266

RMSE:

- All Seasons: 21786
- Spring: 20540
- Autumn: 22210
- Summer: 20739
- Winter: 23018

Random Forest Regression

- Ensemble method with high predictive accuracy that can
  - Handle Missing Values and outliers
  - scale well with new features or samples
  - understand non-linear relationships
- The parameters are: n_estimators=100, max_depth = 20
- Good Prediction Model:
  - Fine-tuned with GridSearch
  - Better at understanding the non-linear relationships between Weekly Sales and the independent factors

WMAE:

- All Seasons: 1905
- Spring: 1506
- Autumn: 1644
- Summer: 1571
- Winter: 2703

RMSE:

- All Seasons: 5201
- Spring: 3405
- Autumn: 4483
- Summer: 3488
- Winter: 6571

Extreme Gradient Boosting Regression

- **XG Boosting**
  - Ensemble method with high predictive accuracy
  - correct predictions are assigned a lower weight, while the incorrect predictions are given a larger weight
    - focuses on the points that were incorrectly predicted
  - uses regularization and hyper parameter tuning
  - can handle non-linearities, outliers, and missing data
- The parameters are: n_estimators=1000, max_depth=12, learning_rate = 0.1
- Good Prediction Model:
  - The weighting mechanism helps in generating more accurate predictions

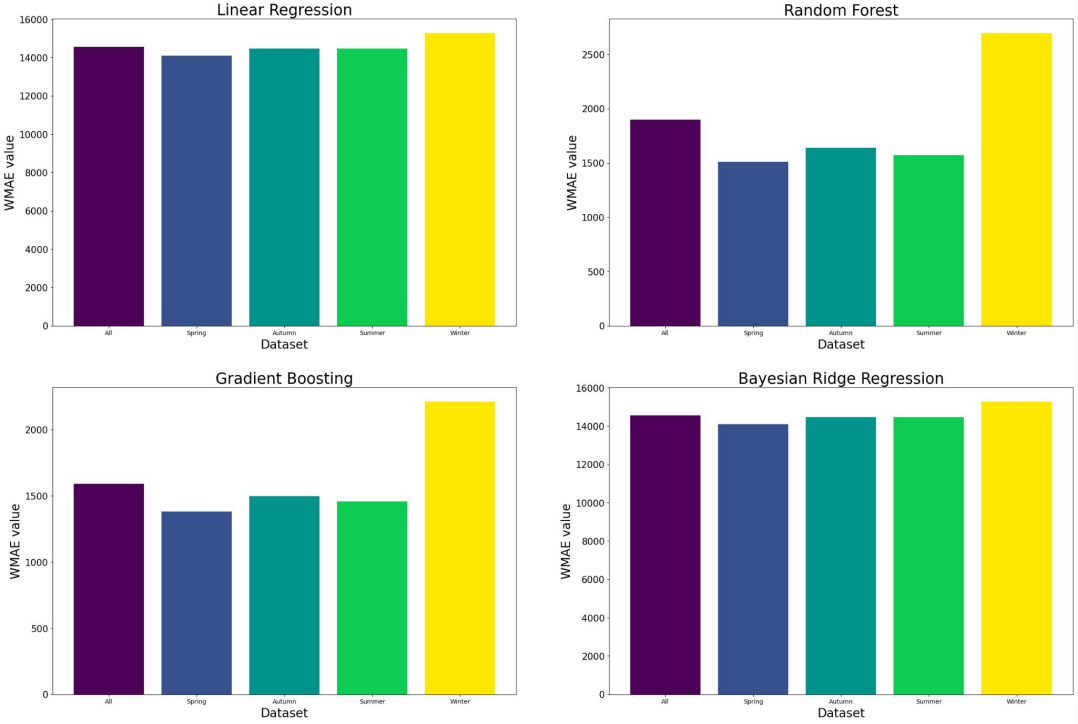
WMAE:

- All Seasons: 1589
- Spring: 1367
- Autumn: 1498
- Summer: 1457
- Winter: 2216

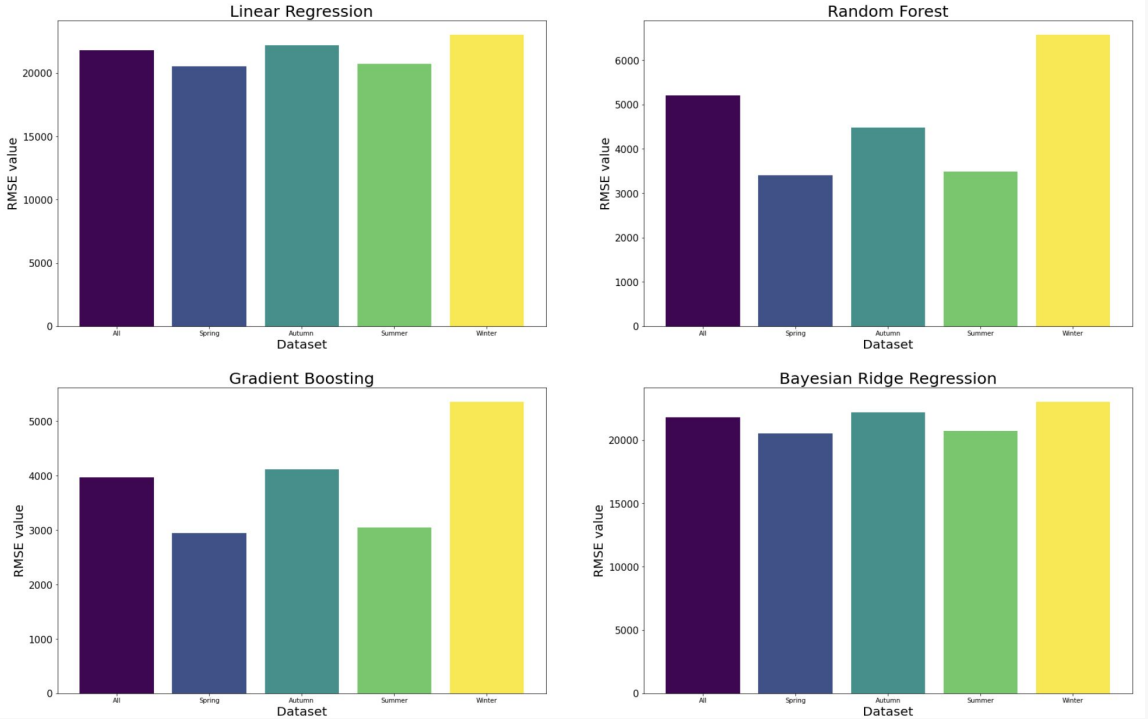
RMSE:

- All Seasons: 3971
- Spring: 2950
- Autumn: 4122
- Summer: 3045
- Winter: 5355

WMAE Plots


Linear Regression
WMAE value
Dataset
All
Spring
Autumn
Summer
Winter
Random Forest
Gradient Boosting
Bayesian Ridge Regression

RMSE Plots


Linear Regression
RMSE value
Dataset
All
Spring
Autumn
Summer
Winter
Random Forest
Gradient Boosting
Bayesian Ridge Regression

Results – Selecting the Best Model

- Lowest WMAE and RMSE score model - **XGBoost Regressor**
  - Better at capturing non-linearities
  - Ensemble learning method
  - Robust to outliers
  - Provides feature importance scores allowing retailers to understand the impact of each feature

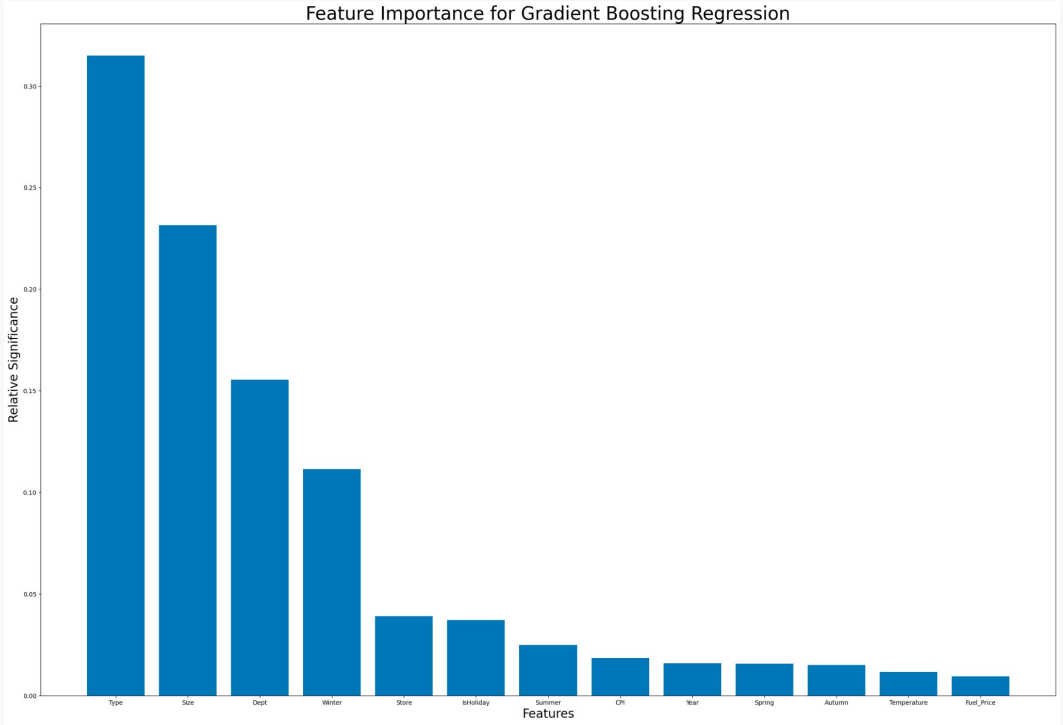
- Best Model if the date is in Spring - **XGBoost trained on only Spring data**

- Best Model if the date is in Summer - **XGBoost trained on only Summer data**

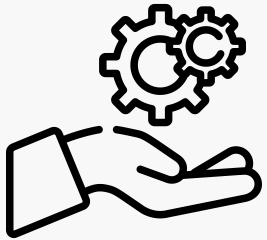
- Best Model for other Seasons **- XGBoost trained on entire dataset**

Results

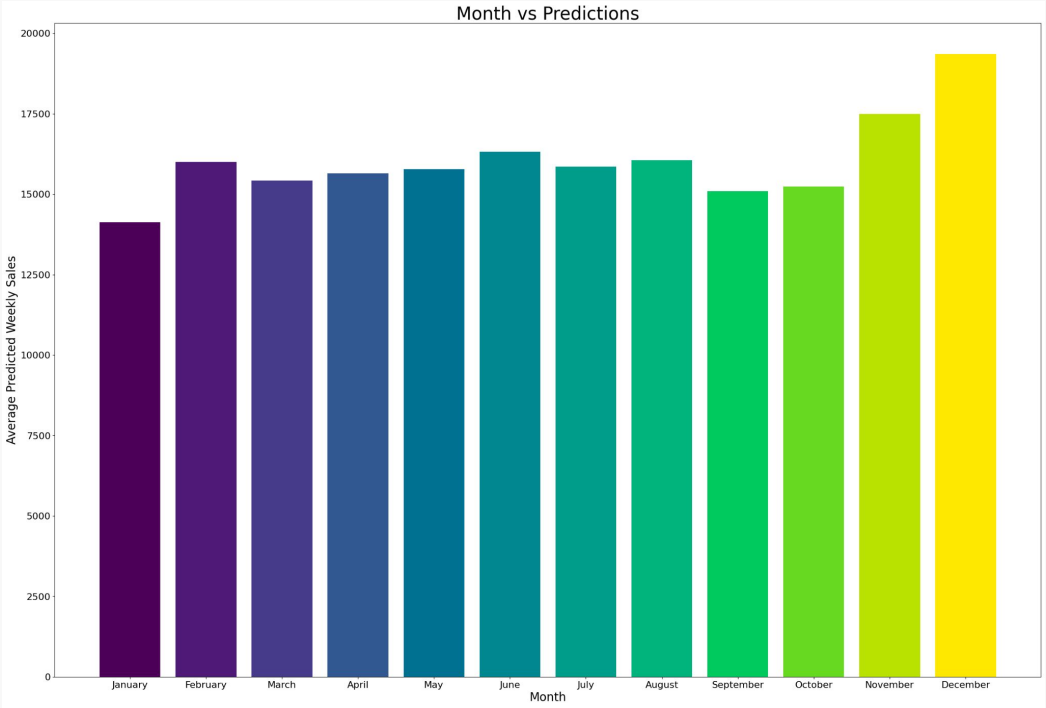
Results – Feature Importance



- **Type** of Store and **Size** have largest impact on sales
- Temperature, CPI have negligible effect as supported by initial EDA

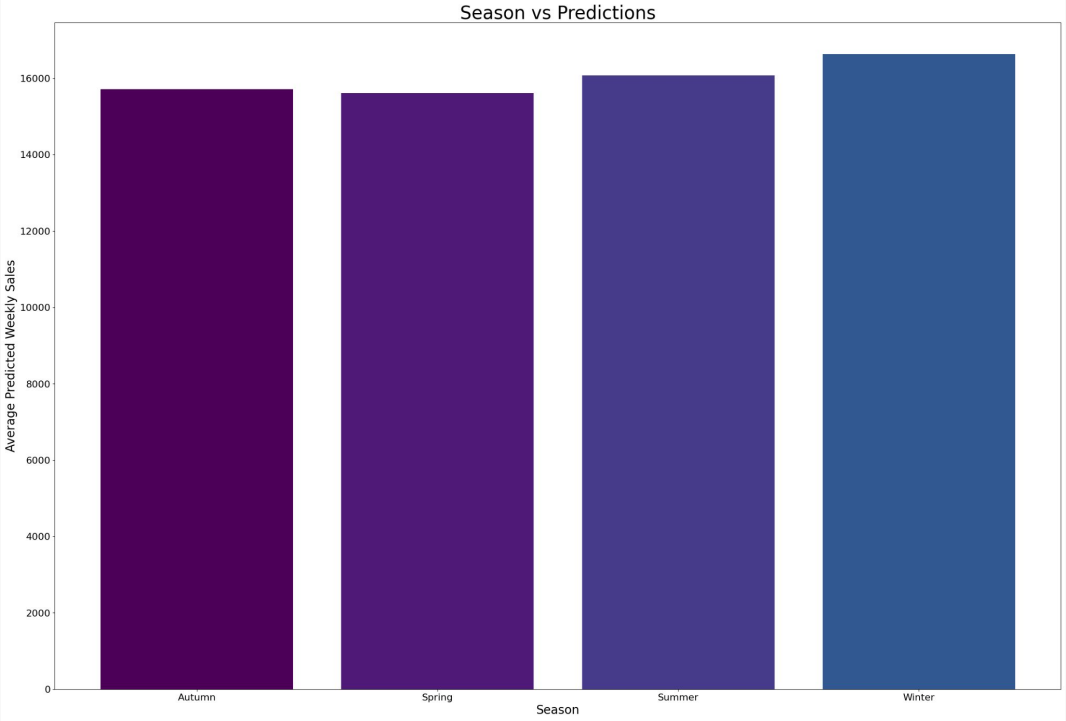


Prediction Sales per month



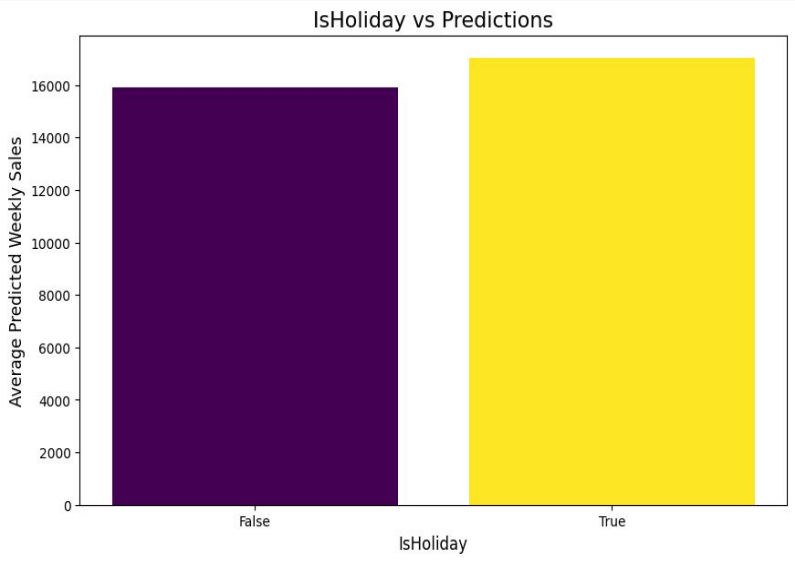
- **December** and **November** have the highest sales
- Likely because of the “Holiday season” in the US with large promotional days such as Black Friday and Christmas Deals

Predicted Sales per Season

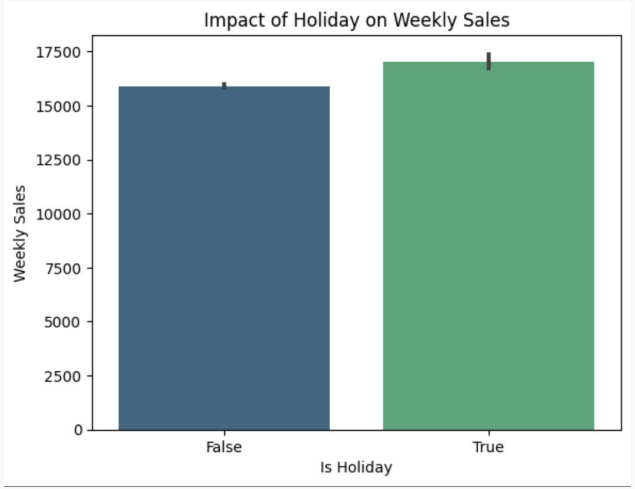


- **Winter Season** has the most impact as compared to other seasons.
    - Supports hypothesis that winter season would have the highest predicted sales at Walmart.

Holiday Impact on Predicted Sales



Original bar chart from training data

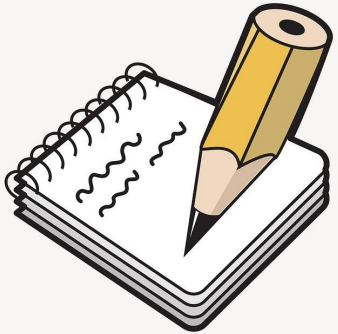


- Predicted Weekly Sales values for holiday breakdown is very similar to training data breakdown → **validates model accuracy**
- Predicting that weekly sales would be higher for holiday days compared to non-holiday days

Conclusion

Conclusion

- Hypothesis was supported by our findings
  - Sales tend to be higher in the winter months where there are more holidays
  - Can also indirectly assume sales tend to be higher when temperatures are colder
- Slight correlation between lower fuel prices and weekly sales but not enough data to make a conclusion



Limitations

- **Only had data available from 2010-2013** which made it difficult during training
  - Impact of economic conditions such as CPI and Unemployment cannot fully be studied with just 2 years of data
- **Lacks indications of any additional marketing or promotional activities**
- **Lack of Regional Information**
  - Impact of Walmarts stores in higher-income areas on sales
  - Variables like unemployment/CPI have different interpretations based on each area
- **Lack of data from online purchases**
  - website traffic and e-commerce trends.
- **No product-specific details**
  - specifications, features, and attributes

Future Work

- **Incorporate External Factors**:
  - Enhance the models by including external factors like economic indicators and competitor data
- **Anomaly Detection:**
  - Identify unusual patterns or events that might disrupt regular sales patterns allowing for proactive responses
- **Location-Based Analysis:**
  - Understand the impact of the locations of stores on sales
- **Mobile App / Online Store Analytics:**
  - Analyze and predict the sales of Walmart customers that shop online through mobile app metrics

References

- Ahmad, Arfat, and M. P. Gupta. "A Predictive Analytics Approach for Sales Prediction in Retail Industry." ResearchGate, www.researchgate.net/publication/316789653_A_Predictive_Analytics_Approach_for_Sales_Prediction_in_Retail_Industry.

- Walmart's sales data analysis - a big data analytics perspective. (2017). https://doi.org/10.1109/APWConCSE.2017.00028

- El-Khawaldeh, Ahmad, Omar Nawayseh, and Mohammad Saraee. "Sales Forecasting with Machine Learning Techniques." ResearchGate, www.researchgate.net/publication/331964026_Sales_Forecasting_with_Machine_Learning_Techniques.

- Uvarova, Olga, and Svetlana Dolganova. "Machine Learning in Retail Price Optimization: The Case of Fashion Retail." IEEE Xplore, ieeexplore.ieee.org/document/9154450.

- Vouldis, Angelos, and Dimitris Dolkas. "Sales Forecasting in Fashion Retail: A Review." SpringerLink, link.springer.com/article/10.1007/s10845-017-1340-3.