PROJECT TITLE:

ROSSMANN COMPANY SALES FORECASTING AND BUSINESS STRATEGY OPTIMIZATION

INTRODUCTION

In the competitive landscape of retail, accurate sales forecasting is vital for optimizing inventory management, improving customer satisfaction, and maximizing profitability. Rossmann, a leading European drug store chain, operates hundreds of stores across various regions, each with unique sales patterns influenced by factors such as promotions, holidays, and competition. This project aims to leverage historical sales data to develop a predictive model that not only forecasts future sales but also provides actionable insights for business strategy optimization.

RESEARCH QUESTION

1. **Sales Prediction:**
   - What are the key factors influencing sales at Rossmann stores?
   - How can we accurately predict daily sales at individual stores using historical data and external variables?
2. **Stores Analysis:**
   - How do sales trends vary across different regions and store types?
   - What factors contribute to the differences in sales performance across various regions?
3. **Promotion Effectiveness:**
   - Which promotional strategies have the most significant impact on sales?
   - How can Rossmann optimize its promotional activities to align with predicted sales trends?
4. **External Factors:**
   - How do holidays and other external factors impact sales performance?
   - Can weather data be integrated to improve forecasting accuracy?
5. **Time Series Analysis :**
   - Can we forecast future sales accurately using historical data?
   - Develop predictive models to forecast sales for upcoming months or years.

OBJECTIVES:

- **Primary Objective**:

- Develop an accurate sales forecasting model for Rossmann using the provided training dataset, which includes historical sales data and store-specific features.

- **Secondary Objectives**:
  - Analyze key drivers of sales to identify trends and patterns.
  - Provide recommendations for inventory optimization and promotional strategies based on forecasted sales.

  Evaluate the impact of external factors (such as holidays and promotions) on sales performance.

TOOLS AND TECHNOLOGIES:

Programming Language:

Python for data preprocessing, modeling, and analysis using libraries such as pandas, scikit-learn, statsmodels, and TensorFlow/Keras for deep learning models.

- Data Mining Techniques:

Exploratory Data Analysis (EDA): To uncover trends, seasonal patterns, and correlations.

Time Series Forecasting: Utilizing models such as SARIMA, Prophet, or machine learning approaches like XGBoost and LSTM for predictions.

Feature Engineering: Creating relevant features from the datasets to enhance model performance.

- Visualization Tools:

Tableau or Python libraries (Matplotlib, Seaborn) for visualizing sales trends and insights.

- Evaluation Metrics:

Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) for assessing model performance.

METHODOLOGY:

- Data Source and Collection: *Rossmann Sales Dataset* (from Kaggle).
- Data Preprocessing: Perform data cleaning, handle missing values, and ensure data quality.
- Exploratory Data Analysis (EDA): Use statistical methods to uncover patterns, correlations, and distributions in sales and profit data.
- Feature Engineering: Create new features such as holiday effects, promotional periods, and rolling averages to enhance the model.
- Model Development: Train multiple forecasting models (ARIMA, Prophet, LSTM, XGBoost) and tune their hyperparameters for optimal performance.
- Model Evaluation
- Business Strategy Recommendations

- Visualization: Use Tableau to create dashboards that display sales trends, forecast results, and profitability metrics.

EXPECTED OUTCOME:

Accurate Sales Forecasts: A predictive model capable of forecasting daily sales for Rossmann stores with high accuracy.

Insights on Sales Drivers: Identification of key factors affecting sales and how they interact with each other.

Actionable Business Strategies: Recommendations for optimizing inventory management and promotional activities based on forecasted sales.

Visualization of Insights: Clear visual representations of trends, forecasts, and strategies to present to stakeholders effectively.

DATA COLLECTION AND DATA CLEANING :

```
import numpy as np
import pandas as pd
from datetime import datetime
train = pd.read_csv("C:/Users/yugas/OneDrive/Desktop/rohan final/train.csv")
test = pd.read_csv("C:/Users/yugas/OneDrive/Desktop/rohan final/test.csv")
store = pd.read_csv("C:/Users/yugas/OneDrive/Desktop/rohan final/store.csv")
```

```
print(store.head())
```

```
   Store StoreType Assortment  CompetitionDistance  CompetitionOpenSinceMonth  \
0      1         c          a               1270.0                        9.0
1      2         a          a                570.0                       11.0
2      3         a          a              14130.0                       12.0
3      4         c          c                620.0                        9.0
4      5         a          a              29910.0                        4.0

   CompetitionOpenSinceYear  Promo2  Promo2SinceWeek  Promo2SinceYear  \
0                    2008.0       0              NaN              NaN
1                    2007.0       1             13.0           2010.0
2                    2006.0       1             14.0           2011.0
3                    2009.0       0              NaN              NaN
4                    2015.0       0              NaN              NaN

     PromoInterval
0              NaN
1  Jan,Apr,Jul,Oct
2  Jan,Apr,Jul,Oct
3              NaN
4              NaN
```

```
print(train.head())
```

```
   Store  DayOfWeek        Date  Sales  Customers  Open  Promo StateHoliday  \
0      1          5  2015-07-31   5263        555     1      1            0
1      2          5  2015-07-31   6064        625     1      1            0
2      3          5  2015-07-31   8314        821     1      1            0
3      4          5  2015-07-31  13995       1498     1      1            0
4      5          5  2015-07-31   4822        559     1      1            0

   SchoolHoliday
0              1
1              1
2              1
3              1
4              1
```

```
store.describe()
```

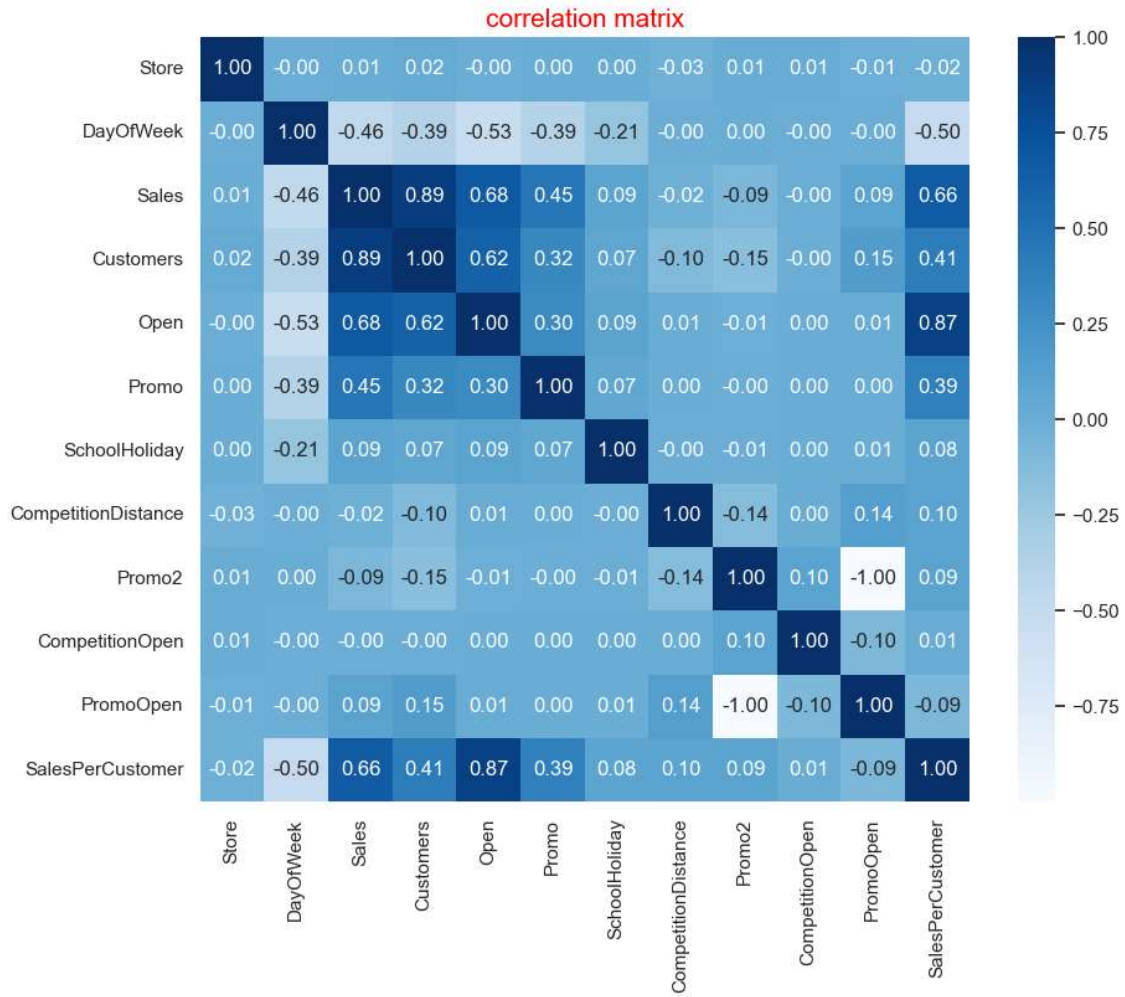
	Store	CompetitionDistance	CompetitionOpenSinceMonth	CompetitionOpenSinceYear	Promo2	Promo2SinceWeek	Promo2SinceYear
count	1115.00000	1112.000000	761.000000	761.000000	1115.000000	571.000000	571.000000
mean	558.00000	5404.901079	7.224704	2008.668857	0.512108	23.595447	2011.763573
std	322.01708	7663.174720	3.212348	6.195983	0.500078	14.141984	1.674935
min	1.00000	20.000000	1.000000	1900.000000	0.000000	1.000000	2009.000000
25%	279.50000	717.500000	4.000000	2006.000000	0.000000	13.000000	2011.000000
50%	558.00000	2325.000000	8.000000	2010.000000	1.000000	22.000000	2012.000000
75%	836.50000	6882.500000	10.000000	2013.000000	1.000000	37.000000	2013.000000
max	1115.00000	75860.000000	12.000000	2015.000000	1.000000	50.000000	2015.000000

```
print(train.isnull().sum())
print(store.isnull().sum())

merged_data = train.merge(store, on='Store', how='left')

merged_data.head()
```

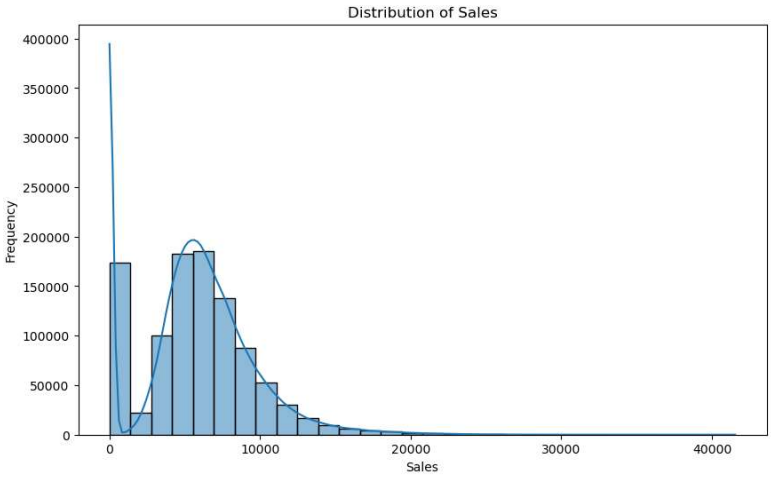
```
Store            0
DayOfWeek        0
Date             0
Sales            0
Customers        0
Open             0
Promo            0
StateHoliday     0
SchoolHoliday    0
dtype: int64
Store                        0
StoreType                    0
Assortment                   0
CompetitionDistance          3
CompetitionOpenSinceMonth  354
CompetitionOpenSinceYear   354
Promo2                       0
Promo2SinceWeek            544
Promo2SinceYear            544
PromoInterval              544
dtype: int64
```

	Store	DayOfWeek	Date	Sales	Customers	Open	Promo	StateHoliday	SchoolHoliday	StoreType	Assortment	CompetitionDistance	CompetitionOpenSinc
0	1	5	2015-07-31	5263	555	1	1	0	1	c	a	1270.0	
1	2	5	2015-07-31	6064	625	1	1	0	1	a	a	570.0	
2	3	5	2015-07-31	8314	821	1	1	0	1	a	a	14130.0	
3	4	5	2015-07-31	13995	1498	1	1	0	1	c	c	620.0	
4	5	5	2015-07-31	4822	559	1	1	0	1	a	a	29910.0	

CORRELATION MATRIX

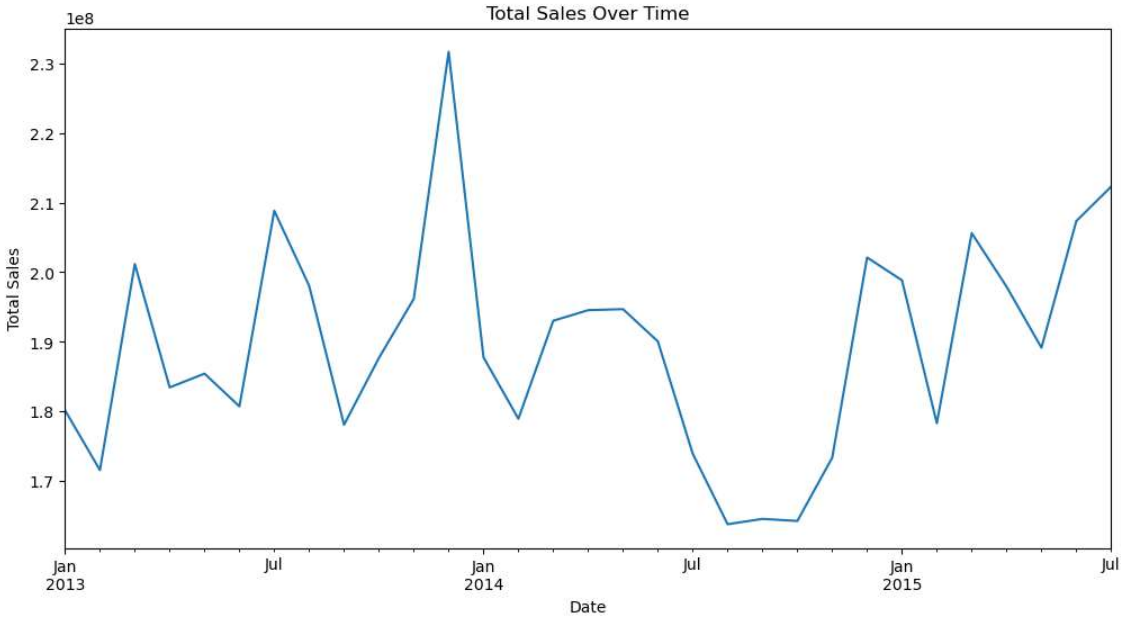


SALES PREDICTION:



The histogram plot shows the distribution of sales, with a significant concentration at lower sales values and a long tail towards higher values. This right-skewed distribution is common in sales data, where a few days might see exceptionally high sales, but most days have lower sales.

This visualization is useful because it confirms the variability in daily sales and highlights the importance of considering outliers in our analysis



Overall Trend:

- Positive Growth: There's a noticeable upward trend in total sales from January 2013 to July 2015. This indicates that Rossmann has been experiencing overall growth in its sales over the period.

Seasonal Patterns:

- Seasonal Peaks and Troughs: The plot shows recurring peaks around January and mid-year (July), suggesting that Rossmann's sales are influenced by seasonal factors. These peaks could correspond to specific sales events, holidays, or promotional periods when consumer spending typically increases.

Significant Fluctuations:

- High Sales Periods: There are notable peaks, particularly around January 2014 and January 2015, where total sales exceed 2.3e8. These spikes could be linked to significant promotions, new product launches, or high consumer demand during the holiday season.

- Low Sales Periods: There are dips observed, for instance around mid-2014, where sales drop below 1.7e8. These troughs could be due to off-season periods, stockouts, or reduced consumer spending.

Implications for Business Strategy:

- Inventory Management: Understanding these patterns can help Rossmann ensure they have sufficient stock during peak periods to meet consumer demand, avoiding stockouts and maximizing sales.

- Marketing Campaigns: Targeted promotions can be planned around these high-sales periods to capitalize on increased consumer activity. Similarly, efforts can be made to boost sales during lower periods.

- Staffing: Adequate staffing during peak times ensures better customer service and efficient operations, while optimized staffing during low periods can help control costs.

Predicting Sales Using Random Forest Model

Model Selection and Hyperparameter Tuning

After extensive analysis, Random Forest was selected as the most appropriate model for predicting Rossmann's sales. The model was fine-tuned using hyperparameter tuning through Grid Search, resulting in optimal performance.

Model Performance Metrics

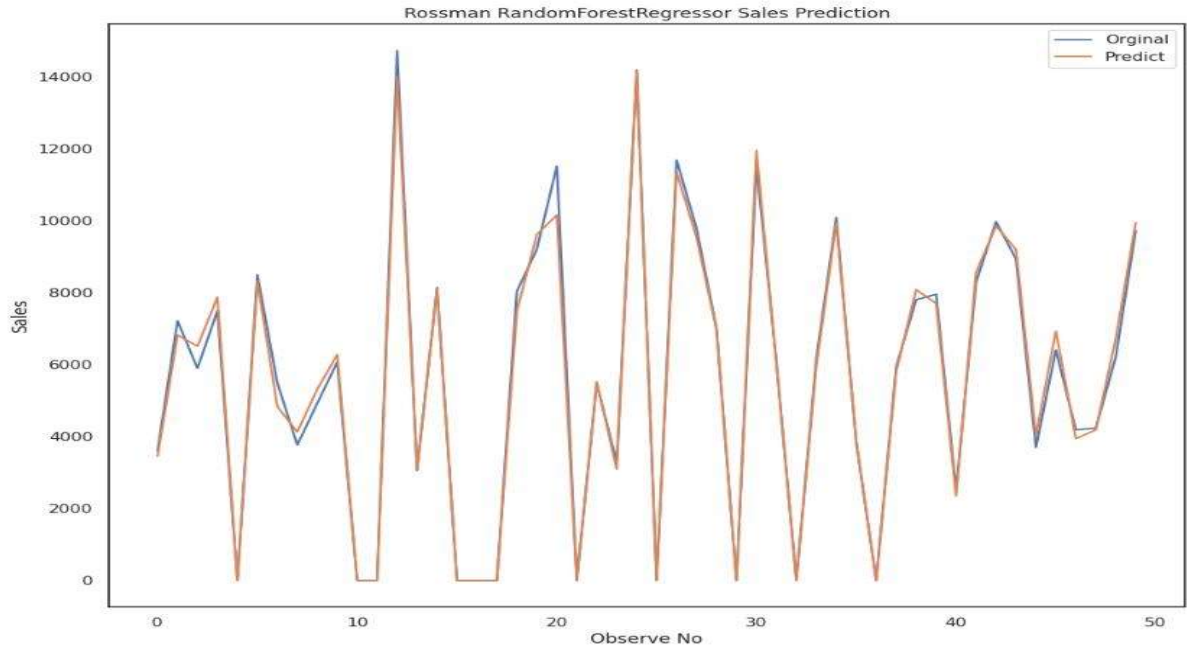
R^2 Score: 0.987

The model explains approximately 98.7% of the variance in the sales data, indicating strong predictive power.

Root Mean Squared Error (RMSE): 438.48

On average, the model's predictions deviate from the actual sales figures by ±438 units, demonstrating high accuracy given the scale of sales.

Graphical Analysis



Original Sales (Blue Line): Represents the actual sales data.

Predicted Sales (Orange Line): Represents the sales data predicted by the Random Forest model.

Insight: The close alignment between the original and predicted sales lines confirms the high R^2 score and low RMSE, validating the model's accuracy.

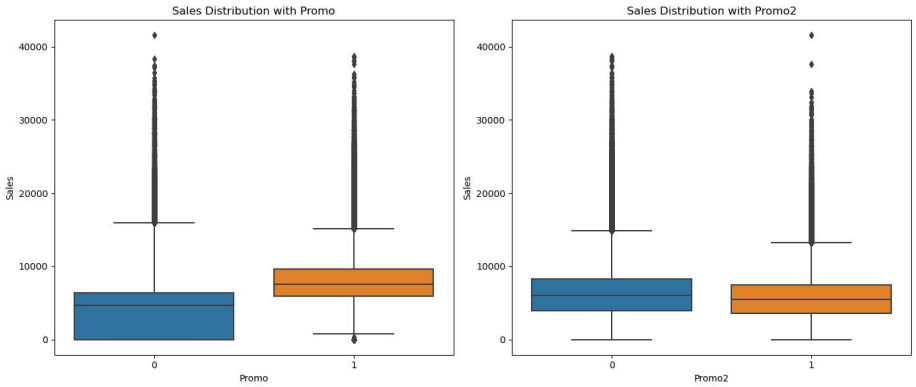
Interpretation and Conclusion

The Random Forest model exhibits exceptional performance in predicting Rossmann's sales, as evidenced by the high R^2 score and low RMSE.

This robust model can be leveraged for:

- Accurate Sales Forecasting: Predicting future sales trends to inform inventory management and strategic planning.
- Optimizing Promotions: Identifying periods for targeted promotions to maximize sales.
- Strategic Decision Making: Providing insights to support data-driven decisions in various aspects of store operations.

PROMOTION EFFECTIVENESS:




```
avg_sales_promo = merged_data.groupby('Promo')['Sales'].mean().reset_index()
avg_sales_promo2 = merged_data.groupby('Promo2')['Sales'].mean().reset_index()

print("Average Sales with Promo:")
print(avg_sales_promo)

print("Average Sales with Promo2:")
print(avg_sales_promo2)
```

```
Average Sales with Promo:
   Promo        Sales
0      0  4406.050805
1      1  7991.152046
Average Sales with Promo2:
   Promo2        Sales
0       0  6124.710161
1       1  5423.718218
```

```
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error

merged_data = pd.get_dummies(merged_data, columns=['StoreType', 'Assortment', 'StateHoliday', 'PromoInterval'])

X = merged_data[['Promo', 'Promo2', 'CompetitionDistance', 'CompetitionOpenSinceMonth', 'CompetitionOpenSinceYear']]
y = merged_data['Sales']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = LinearRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)

mse = mean_squared_error(y_test, y_pred)
rmse = mse ** 0.5
print('Root Mean Squared Error:', rmse)

coefficients = pd.DataFrame(model.coef_, X.columns, columns=['Coefficient'])
print(coefficients.sort_values(by='Coefficient', ascending=False))
```

```
Root Mean Squared Error: 3412.878666156424
                          Coefficient
Promo                     3590.487890
CompetitionOpenSinceYear     3.667155
CompetitionDistance         -0.016795
CompetitionOpenSinceMonth  -35.362893
Promo2                    -732.037052
```

Analysis of Promo and Promo2:

Average Sales:

- Promo (Standard Promotion):
  - Without Promo: Average Sales = 4406.05
  - With Promo: Average Sales = 7991.15
  - Impact: Sales significantly increase during promotional periods, nearly doubling with promotions.
- Promo2 (Extended Promotion):
  - Without Promo2: Average Sales = 6124.71
  - With Promo2: Average Sales = 5423.72
  - Impact: Contrary to expectations, sales appear to decrease slightly during extended promotions.

Model Coefficients:

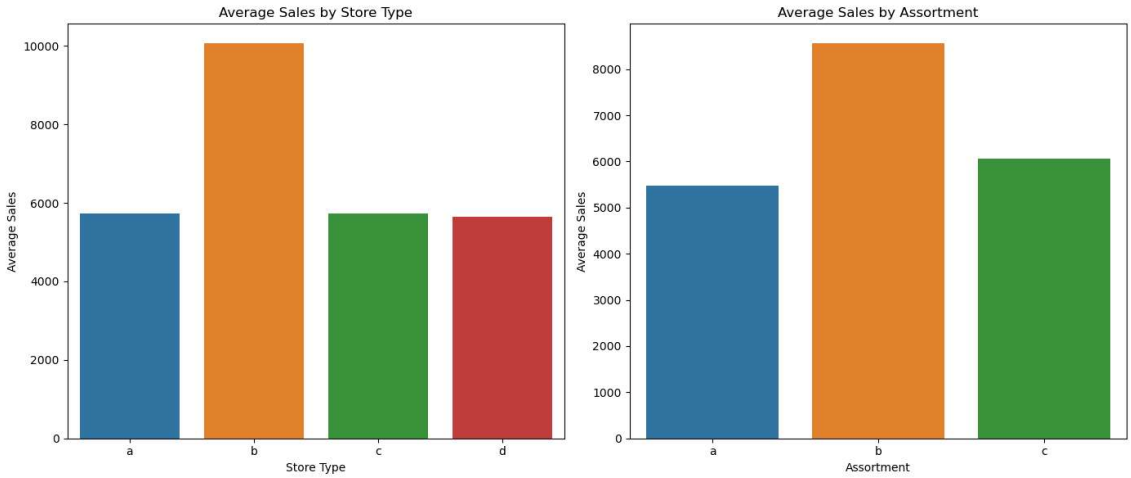
- Promo: The positive coefficient (3590.49) indicates a substantial increase in sales during promotional periods.
- CompetitionOpenSinceYear: Slight positive coefficient (3.67), suggesting a negligible impact from competition duration.
- CompetitionDistance: Negative coefficient (-0.02), indicating distant competitors slightly reduce sales.
- CompetitionOpenSinceMonth: Negative coefficient (-35.36), suggesting an older competition start date might negatively impact sales.
- Promo2: Negative coefficient (-732.04), indicating a decrease in sales during extended promotions.

Interpretation and Insights:

1. Promotions Drive Sales: Standard promotions (Promo) significantly boost sales. Businesses should leverage this by planning regular promotional campaigns to maximize sales.
2. Extended Promotions: Promo2 seems to have a less favorable impact. It might be wise to re-evaluate the strategy behind extended promotions, as they could lead to consumer fatigue or diminished returns.
3. Competitive Landscape: Proximity and duration of competition have a noticeable impact. Understanding competitors' influence can help in strategic planning.
4. Actionable Strategies:

- Regular Promotions: Continue with regular promotions to maintain high sales.
- Optimize Extended Promotions: Reassess extended promotional strategies to ensure they effectively drive sales.
- Competitor Analysis: Keep track of competitors' activities and adjust strategies accordingly to mitigate negative impacts.

STORES ANALYSIS:




```
store_type_sales = merged_data.groupby('StoreType')['Sales'].mean().reset_index()

print(store_type_sales)
```

```
  StoreType          Sales
0         a    5738.179710
1         b   10058.837334
2         c    5723.629246
3         d    5641.819243
```

Average Sales by Store Type:

Store A: $5,738.18

Store B: $10,058.84

Store C: $5,723.63

Store D: $5,641.82

- **Store Type 'a':** Average sales are around 6000.
- **Store Type 'b'**: Average sales peak at approximately 10000, making it the highest among the store types.
- **Store Type 'c'**: Average sales are similar to Store Type 'a', around 6000.
- **Store Type 'd'**: Average sales are the lowest, at around 5000.

Average Sales by Assortment:

- **Assortment 'a'**: Average sales are about 5000.
- **Assortment 'b':** Average sales significantly increase to around 8000.
- **Assortment 'c'**: Average sales are around 6000, higher than 'a' but lower than 'b'.

Interpretation:

1. Store Types:

- Store Type 'b' is clearly the top performer, with the highest average sales. This suggests that stores categorized under 'b' have characteristics or operational strategies that drive higher sales.

- Store Types 'a' and 'c' have similar sales performance, indicating that they might share common features or market conditions.

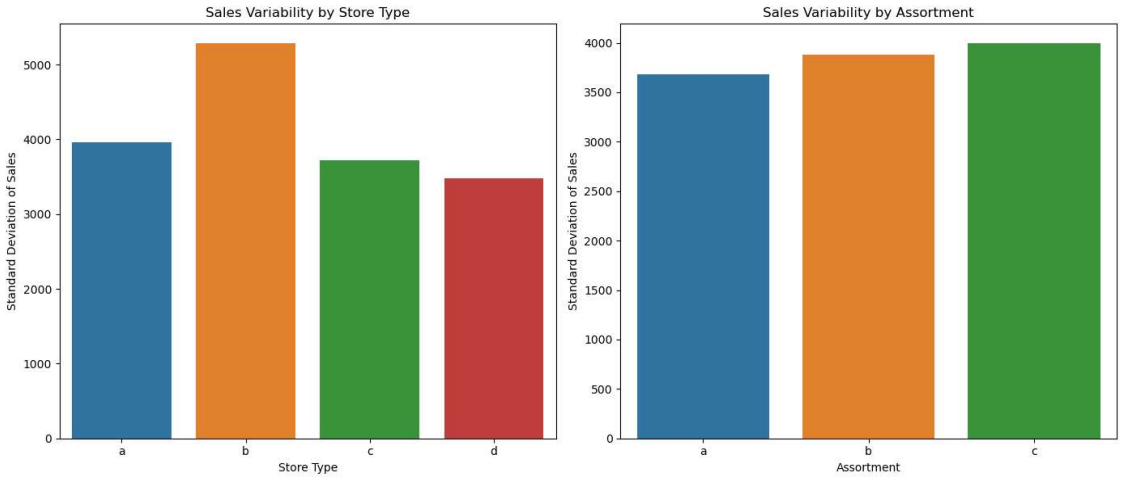
- Store Type 'd' has the lowest average sales, suggesting these stores might need a different strategy or support to boost performance.

2. Assortments:

- Assortment 'b' outperforms the others, indicating that the product mix or variety in this assortment resonates more with customers, driving higher sales.

- Assortment 'c' performs better than 'a', suggesting some added value but not as impactful as 'b'.

- Assortment 'a' shows the lowest average sales, which could mean that this assortment might not fully meet customer preferences or demand.



Sales Variability by Store Type:

- **Store Type 'b'**: Highest variability and sales. Focus on managing stock during peaks.
- **Store Types 'a' and 'c'**: Moderate variability. Aim for consistency.
- **Store Type 'd'**: Lowest variability, consistent but lower sales.

<u>Factors Contributing to Sales Performance Across different Stores:</u>

Store Types and Assortments:

- Store Type 'b' has the highest sales, suggesting it's optimized for performance, likely due to its size, location, or specific characteristics.

- Assortment 'b' drives the most sales, indicating strong customer preference for the products in this assortment.

Promotions:

- Promo: Significantly boosts sales across all regions, demonstrating the effectiveness of regular promotions.

- Promo2: Has mixed results, suggesting extended promotions may need regional tailoring for better effectiveness.

Competition:

- Competition Distance: Slightly negative impact on sales, indicating closer competitors might drive more foot traffic and sales due to consumer preference for proximity.

Seasonal and Holiday Impact:

- State Holidays: Certain holidays negatively affect sales, implying regional holidays have a considerable impact on consumer behavior.

Actionable Insights:

- Store Strategy: Focus on the successful elements of Store Type 'b' and replicate or adapt these strategies across other store types to boost sales.
- Product Mix: Analyze what makes Assortment 'b' successful and consider expanding or adapting these products across more stores.
- Support for Underperformers: Investigate challenges faced by Store Type 'd' and Assortment 'a', and provide targeted support or changes to improve their performance.