

Task 01

Data Science Internship – Assignment

Introduction

This project aims to analyze supermarket transaction data collected over a two-year period across multiple branches. The transactions are categorized into four item types (Type 1 to Type 4) and span two key provinces. By leveraging Data Analytics and Machine Learning techniques, this analysis seeks to provide actionable business insights that can enhance decision-making.

The primary objective of this project is to clean, normalize, and transform the datasets into Python-compatible formats for analysis. After data preparation, the focus shifts to developing business-valued solutions, such as forecasting future sales trends and assessing the impact of promotions.

Data Description

Item.csv

This dataset contains detailed information about the items in supermarkets.

(Code, Description, Type, Brand, Size)

Sales.csv

This dataset records transaction details for sales made across all supermarkets.

(Code, Amount, Units, Time, Province, CustomerId, Supermarket No, Basket, Day, Voucher)

Promotion.csv

This dataset includes information about promotional activities across different supermarkets.

(Code, Supermarkets, Week, Feature, Display, Province)

Supermarkets.csv

This dataset provides details about the supermarket branches.

(Supermarket No, post-code)

Cleaning Datasets

Supermarket Dataset

- Sort the 'supermarket_No' column in ascending order.

Item Dataset

1. Size Column Cleaning:

- Converted all values in the 'size' column to uppercase and removed leading/trailing spaces.
- Replaced 'OUNCE' with 'OZ' and removed unwanted characters (e.g., 'FL OZ').
- Kept only numerical values and 'OZ' or 'LB' units.
- Handled fractions by converting values like '6 1/2 OZ' to '6.5 OZ.'
- Converted mixed measurements (Ex: '6LB 11OZ') by calculating total ounces (1 LB = 16 OZ).
- Returned numerical values for sizes in both 'OZ' and 'LB' formats.

2. Handling Missing or Invalid Sizes:

- Applied filtering to identify rows where the 'cleaned_size' column is null after cleaning.
- Checked and handled missing values in the 'cleaned_size' column.

3. Description Column Cleaning:

- Removed keywords such as 'Type 1', 'Type 2', 'Type 3', and 'Type 4' from the 'description' column.
- Extracted numerical values (size information) from the 'description' column to fill in missing sizes in the 'cleaned_size' column.

4. Filling Missing Values:

- Used extracted size values from the 'description' column to fill in null values in the 'cleaned_size' column.
- Converted the 'cleaned_size' column to a float type for further numerical analysis.

5. Handling Missing Values Using Brand Averages:

- Calculated the average 'cleaned_size' for each brand.
- Replaced missing 'cleaned_size' values with the corresponding brand's average size.

- Used the overall average size to fill in any remaining missing values where brand-specific averages were unavailable.

6. Final Column Adjustments:

- Dropped the original 'size' column after the cleaning was completed.
- Renamed the 'description' column to 'description' and the 'cleaned_size' column to 'cleaned_size_in_OZ'.

Promotion Dataset

- Rename the 'supermarkets' column to 'supermarket_No'.

Sales Dataset

1. Renaming Columns:

- Renamed the 'supermarket' column to 'supermarket_No'.

2. Time Column Transformation:

- Ensured that the 'time' column was converted to string format for proper manipulation.
- Padded the 'time' values with leading zeros to ensure they followed a four-digit format (Ex: '1130' became '1130', '930' became '0930').
- Converted the 'time' column from a string format (HHMM) to a proper time format (HH).

3. Handling Negative Values:

- Converted negative values in the 'amount' column to positive values by taking the absolute value.

Normalization

- The supermarket dataset is already normalized.

Item Dataset

- Extracted unique values from the 'type' and 'brand' columns to create separate data frames for item types and brands.
- Assigned unique identifiers (type_id and brand_id) to each type and brand by indexing these values.
- Merged the type_id and brand_id back into the original dataset to associate each item with its corresponding type and brand identifiers.
- Selected key columns (code, description, type_id, brand_id, and cleaned_size_in_OZ) from the cleaned dataset to create a normalized item data frame.

Promotion Dataset

- Extracted unique values from the 'feature' and 'display' columns to create separate data frames for features and displays.
 - Assigned unique identifiers (feature_id and display_id) to each feature and display by indexing these values.
 - Merged the feature_id and display_id back into the original promotion dataset to associate each promotion record with its corresponding feature and display identifiers.
 - Selected essential columns (code, supermarket_No, week, feature_id, display_id, and province) to create a normalized promotion data frame.
-
- The Sales dataset is normalized.

Business Valued Solution 01

Analyze the effectiveness of promotions on items

While analyzing the sales and promotion datasets, I noticed that sales data spanned from week 1 to week 28, while promotions started from week 43 and continued until week 104. Based on this observation, I hypothesized that promotions were likely applied to items with lower sales during the first 28 weeks. To test this assumption, I aimed to analyze the impact of promotions on sales by comparing pre-promotion sales with forecasted post-promotion sales for the same items.

Steps

1. Loaded and Cleaned Data

The relevant sales, promotion, and item datasets were loaded. After cleaning and normalizing, I prepared the data for analysis.

2. Calculated Weekly Sales (Weeks 1-28)

I calculated the total weekly sales per item for the initial 28 weeks (pre-promotion period). This helped establish a baseline of how much each item sold before any promotions were applied.

	code	week	total_sales	date
0	111112360	13	5.59	2022-04-02
1	111112360	23	4.99	2022-06-11
2	111112360	25	5.99	2022-06-25
3	111112360	26	5.99	2022-07-02
4	111112360	27	46.32	2022-07-09
...
14164	9999985766	24	62.66	2022-06-18
14165	9999985766	25	48.90	2022-06-25
14166	9999985766	26	65.61	2022-07-02
14167	9999985766	27	62.11	2022-07-09
14168	9999985766	28	19.90	2022-07-16

[14169 rows x 4 columns]

3. Forecasted Post-Promotion Sales (Weeks 29-104)

Using SARIMAX, I forecasted the sales of each item from week 29 to week 104, the period when promotions were applied. This forecast assumed that promotions would increase sales.

```

Name: predicted_mean, Length: 76, dtype: float64, 9999985217: 29      23.76
30      23.76
31      23.76
32      23.76
33      23.76
...
100     39.65
101     39.65
102     39.65
103     39.65
104     47.52
Name: predicted_mean, Length: 76, dtype: float64, 9999985260: 29      99.33
30      99.33
31      99.33
32      99.33
33      99.33
...
100     69.95
101     69.95
102     69.95
103     69.95
104    198.66
Name: predicted_mean, Length: 76, dtype: float64, 9999985261: 29    201.78
30    201.78
31    201.78
32    201.78
33    201.78

```

4. Converted Week Numbers to Dates

To improve readability and clarity, I converted the week numbers into actual dates, assuming week 1 started on January 1, 2022. This made the results easier to interpret and visualize.

5. Calculated Total Pre-Promotion Sales:

I aggregated the total sales for each item during weeks 1-28 to represent the pre-promotion sales.

```

      code  pre_promo_sales
0    111112360         74.67
1    566300023        103.39
2    566300028       1023.93
3    566300029         1.49
4    566300035         17.52
..      ...
777  9999985217        541.59
778  9999985260       4871.07
779  9999985261       7901.14
780  9999985488       1231.69
781  9999985766       1822.00

[782 rows x 2 columns]

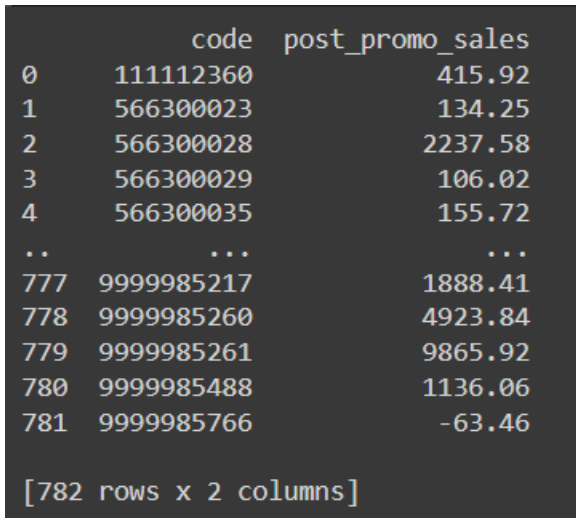
```

6. Filtered Items with Promotions:

I filtered the items that received promotions during weeks 43-104, as I was interested in analyzing the impact of promotions on these specific items.

7. Calculated Total Forecasted Post-Promotion Sales:

Summing the forecasted sales from weeks 29 to 104 for each item gave me the projected sales during the promotion period.



	code	post_promo_sales
0	111112360	415.92
1	566300023	134.25
2	566300028	2237.58
3	566300029	106.02
4	566300035	155.72
..
777	9999985217	1888.41
778	9999985260	4923.84
779	9999985261	9865.92
780	9999985488	1136.06
781	9999985766	-63.46

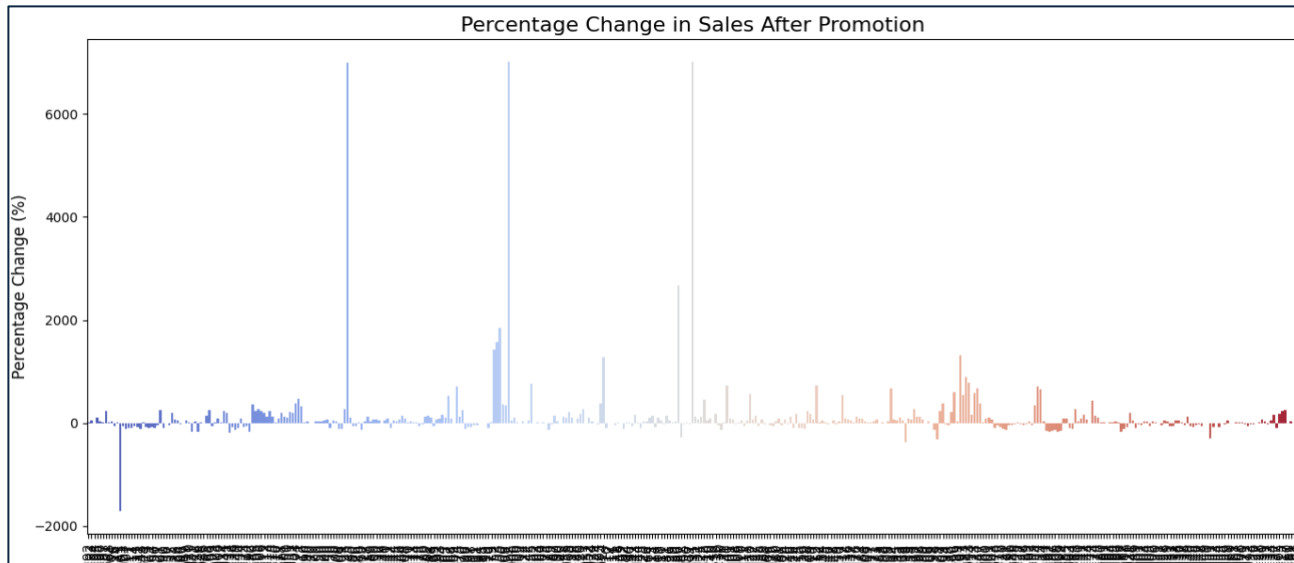
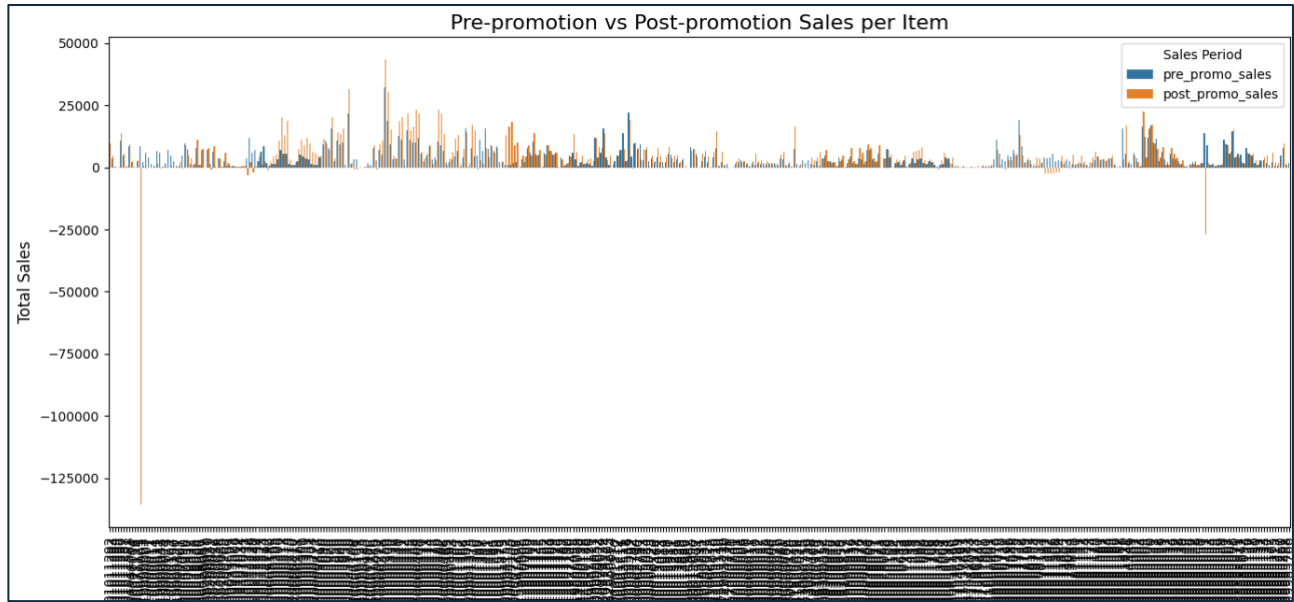
[782 rows x 2 columns]

8. Compared Pre- and Post-Promotion Sales:

I merged the pre-promotion sales with the forecasted post-promotion sales to compare how the sales of each item changed over time. The percentage change was calculated to quantify the impact of promotions.

9. Visualized Results:

I created bar plots to compare pre-promotion and post-promotion sales for each item. Additionally, a percentage change bar plot was generated to highlight which items experienced the most significant changes in sales due to promotions.

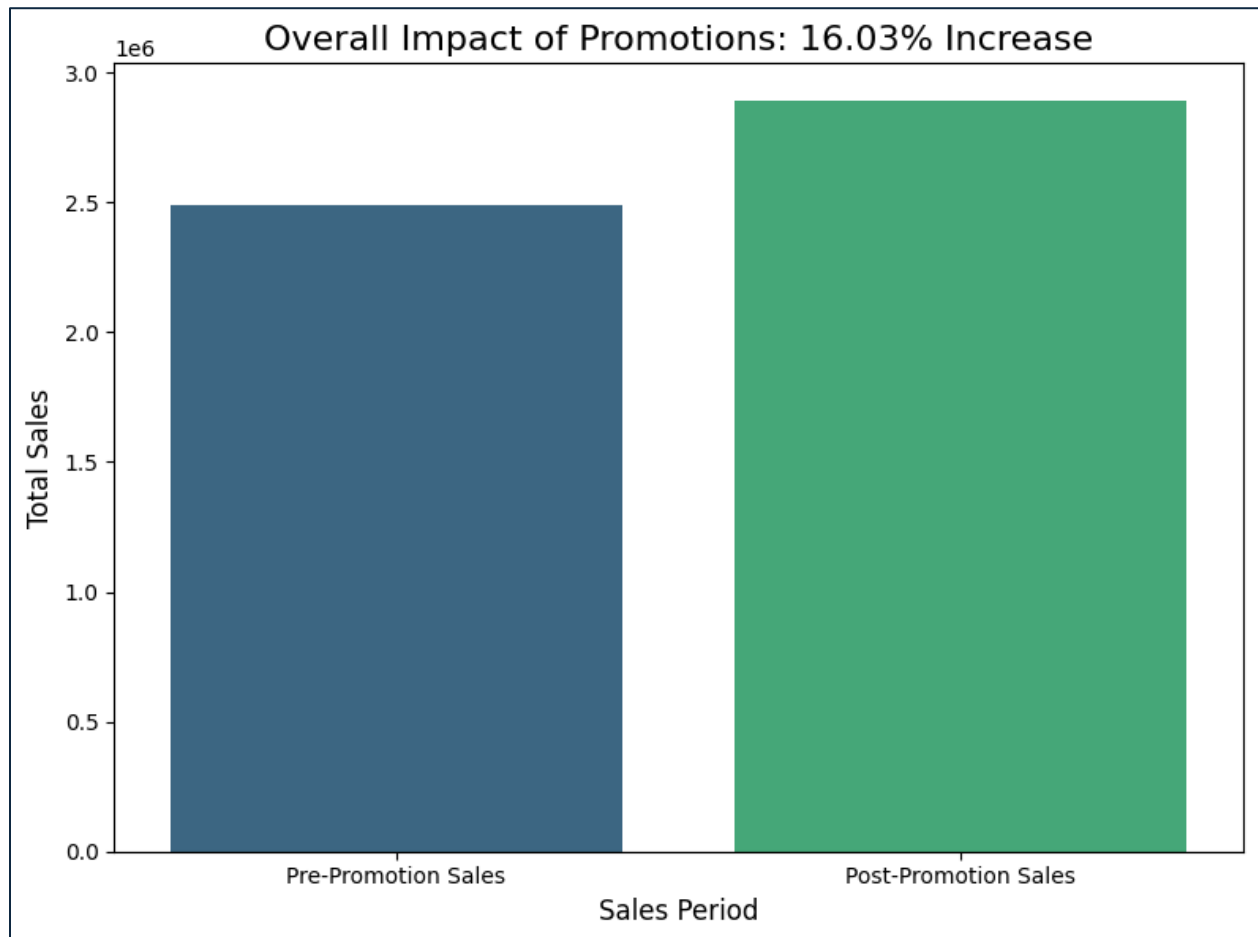


10. Overall Analysis:

I calculated the total pre-promotion sales (weeks 1-28) and total forecasted post-promotion sales (weeks 29-104) to measure the overall impact. A final bar plot was created, summarizing the total sales before and after promotions for promoted items, along with the overall percentage change.

Final Output

The visualizations showed clear differences in sales before and after promotions, with some items experiencing a significant increase in sales.



Conclusion

By analyzing pre-promotion and post-promotion sales, I was able to quantify the effect of promotions on item sales. The results showed a positive impact of promotions, with a notable percentage increase in total sales after the promotion period.

Business Valued Solution 02

Effectiveness of Features and Displays on Sales

I assumed that certain promotion features and display types could have varying impacts on sales. By analyzing sales data grouped by these categories, I aimed to identify the most effective promotion strategies.

Steps

1. I began by merging the promotion_df_normalized dataset with two other datasets—df_feature (containing details about promotional features) and df_display (containing details about promotional displays).

	code	supermarket_No	week	feature_id	display_id	province	\
0	2700042240	285	91	1	1	2	
1	2700042292	285	92	2	2	2	
2	2700042274	285	92	2	2	2	
3	2700042273	285	92	2	2	2	
4	2700042254	285	92	2	2	2	

	promo_date	feature	display
0	2023-09-30	Not on Feature	Mid-Aisle End Cap
1	2023-10-07	Interior Page Feature	Not on Display
2	2023-10-07	Interior Page Feature	Not on Display
3	2023-10-07	Interior Page Feature	Not on Display
4	2023-10-07	Interior Page Feature	Not on Display

2. Next, I merged the promotion data with our pre-promotion vs. post-promotion sales comparison, ensuring that each promotion is properly associated with the respective sales data.
3. After merging the datasets, I grouped sales data by promotion feature and display type, calculating the total sales both before (pre-promotion) and after (post-promotion) for each category.
4. To quantify the impact, I calculated the percentage change in sales for both feature types and display types, which allowed me to measure the effectiveness of different promotional strategies.

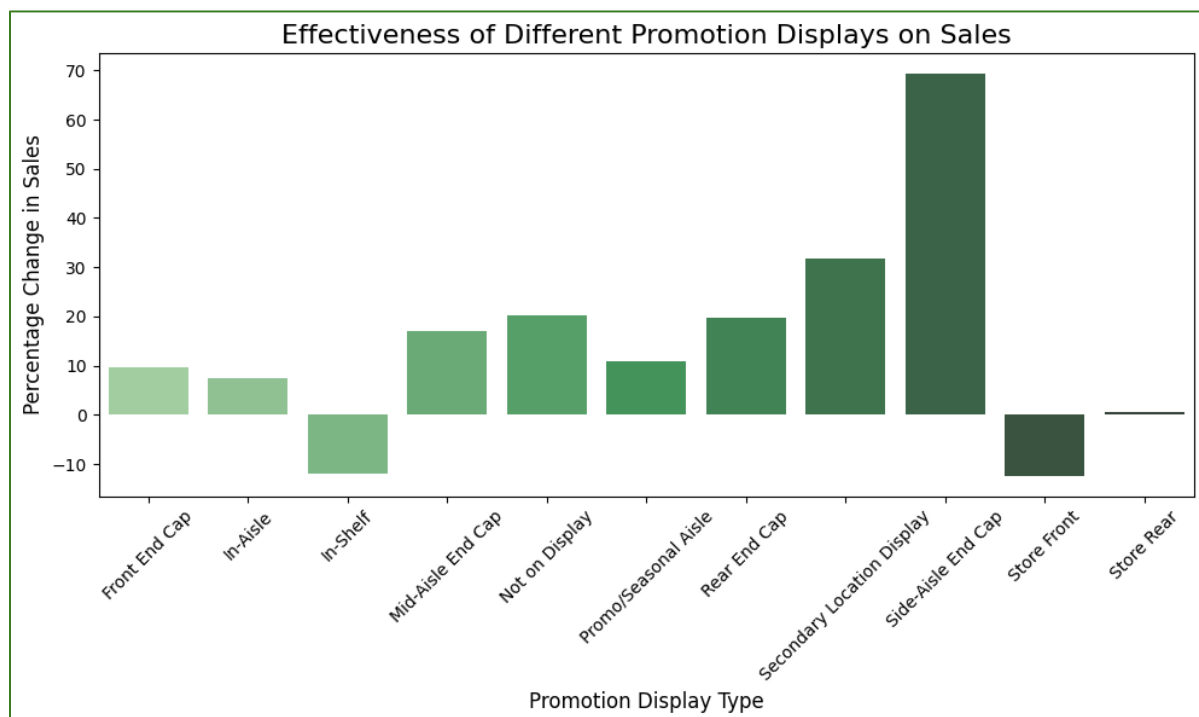
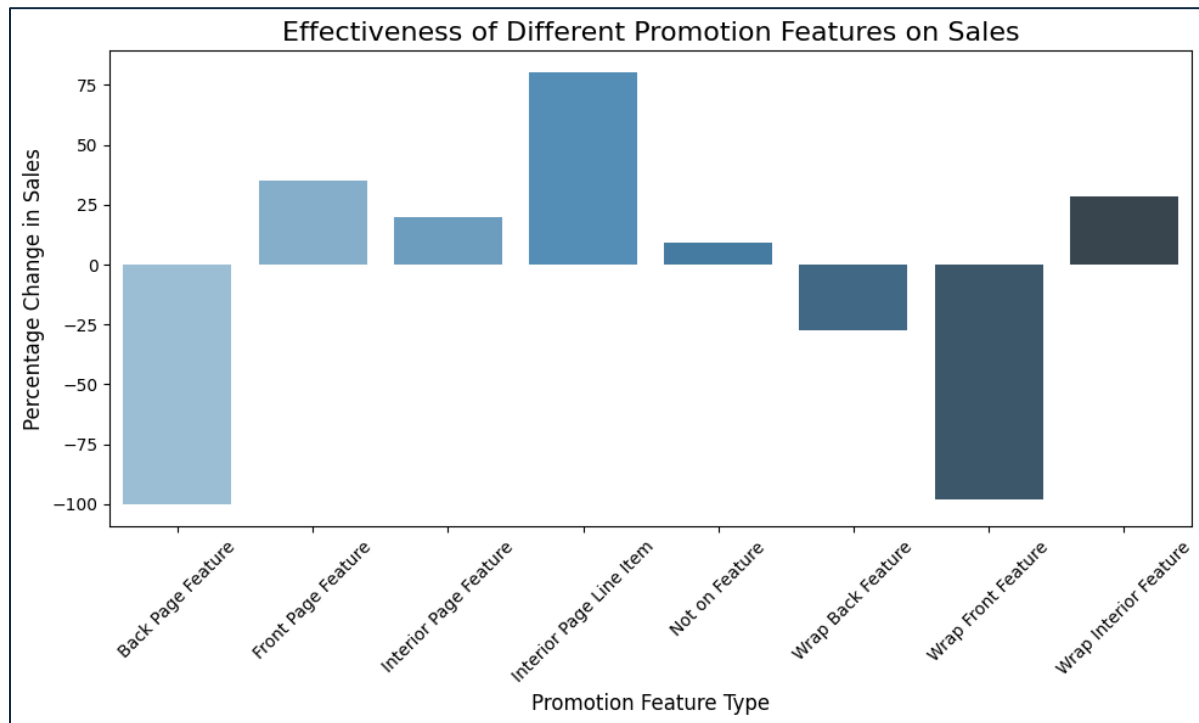
	feature	pre_promo_sales	post_promo_sales	\
0	Back Page Feature	3.653155e+07	-9.407193e+04	
1	Front Page Feature	2.340372e+08	3.158709e+08	
2	Interior Page Feature	1.137528e+09	1.361640e+09	
3	Interior Page Line Item	4.988840e+07	8.990663e+07	
4	Not on Feature	6.353129e+08	6.948932e+08	
5	Wrap Back Feature	3.960946e+07	2.879918e+07	
6	Wrap Front Feature	2.355043e+07	4.604070e+05	
7	Wrap Interior Feature	1.275950e+08	1.640571e+08	
	percentage_change			
0		-100.257509		
1		34.966143		
2		19.701612		
3		80.215520		
4		9.378091		
5		-27.292179		
6		-98.045017		
7		28.576472		

	display	pre_promo_sales	post_promo_sales	\
0	Front End Cap	9.978355e+07	1.093376e+08	
1	In-Aisle	5.970409e+07	6.422458e+07	
2	In-Shelf	2.079499e+08	1.830078e+08	
3	Mid-Aisle End Cap	4.326863e+07	5.068874e+07	
4	Not on Display	1.451836e+09	1.746001e+09	
5	Promo/Seasonal Aisle	5.741602e+07	6.361938e+07	
6	Rear End Cap	2.288446e+08	2.739035e+08	
7	Secondary Location Display	8.641693e+07	1.137720e+08	
8	Side-Aisle End Cap	5.590687e+06	9.466895e+06	
9	Store Front	1.514452e+07	1.326719e+07	
10	Store Rear	2.809849e+07	2.824422e+07	
	percentage_change			
0		9.574807		
1		7.571499		
2		-11.994301		
3		17.148934		
4		20.261618		
5		10.804223		
6		19.689735		
7		31.654725		
8		69.333308		
9		-12.396057		
10		0.518614		

5. I then used Seaborn and Matplotlib to visualize the results.

Final Output

The final output provided a clear comparison of how various promotional features and displays impacted sales.



Conclusion

By comparing pre-promotion and post-promotion sales across different feature types and display types, All can identify the most successful approaches for future promotions.