Store Sales Time Series Prediction Report

Abstract

This project focuses on developing a machine learning model to forecast daily unit sales for Corporación Favorita, a major grocery retailer in Ecuador. Accurate sales predictions are crucial for:

- Reducing overstocking and understocking
- Minimizing food waste
- Enhancing customer satisfaction

Using historical sales data, store metadata, promotional activities, oil prices, and holiday information, several models were tested, including:

- **Linear Regression**
- **ARIMA**
- **LSTM**

The **Boosted Hybrid model** (Linear Regression + Random Forest) performed best, achieving an RMSLE of **0.6841**, indicating strong potential for future forecasting tasks.

Data Wrangling

The data wrangling process was critical to ensure the datasets were correctly aligned and ready for analysis. Several challenges were encountered, and the following strategies were employed to address them:

Data Discrepancies

- **Challenge**: The original datasets, such as sales data, holiday events, and oil prices, had inconsistencies in formats, including different date formats and misaligned keys.
- **Solution**:
  - To standardize the data, all date fields were converted to a uniform format, ensuring consistency across datasets.
  - The datasets were then merged using common keys such as store IDs and dates, ensuring proper alignment.
  - Categorical variables, such as store type and holiday type, were encoded consistently, ensuring they could be merged accurately without loss of information.

Missing Values

- **Challenge**: Some datasets, especially the sales data, contained missing values. Leaving these values unaddressed could have negatively impacted the models' accuracy.
- **Solution**:
    - Missing sales data was imputed using the mean sales for the corresponding product categories and days. This approach helped maintain the statistical properties of the dataset without introducing bias.
    - Missing values in other datasets, such as holidays and oil prices, were handled on a case-by-case basis, ensuring all necessary data was either appropriately filled or removed to prevent data inconsistencies.

Feature Engineering

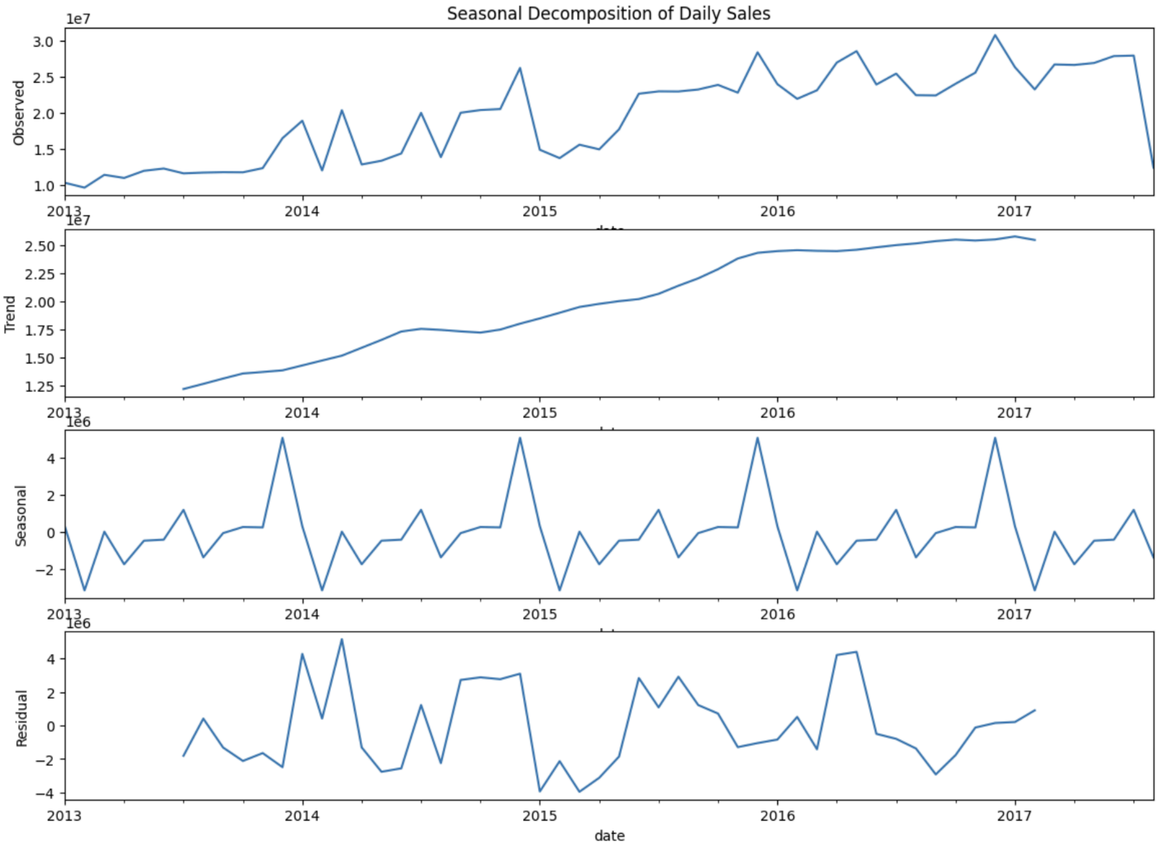
- **Challenge**: The raw data lacked key features related to time, which are essential for capturing trends and seasonality in sales forecasting.
- **Solution**:
    - New features were engineered based on the date information, such as the day of the week, month, and quarter. These features provided more granular insights into temporal patterns and enhanced the model's ability to learn from the time-based dynamics of the data.

By addressing these challenges methodically, a clean and well-structured dataset was created, laying a strong foundation for further exploratory analysis and model building.

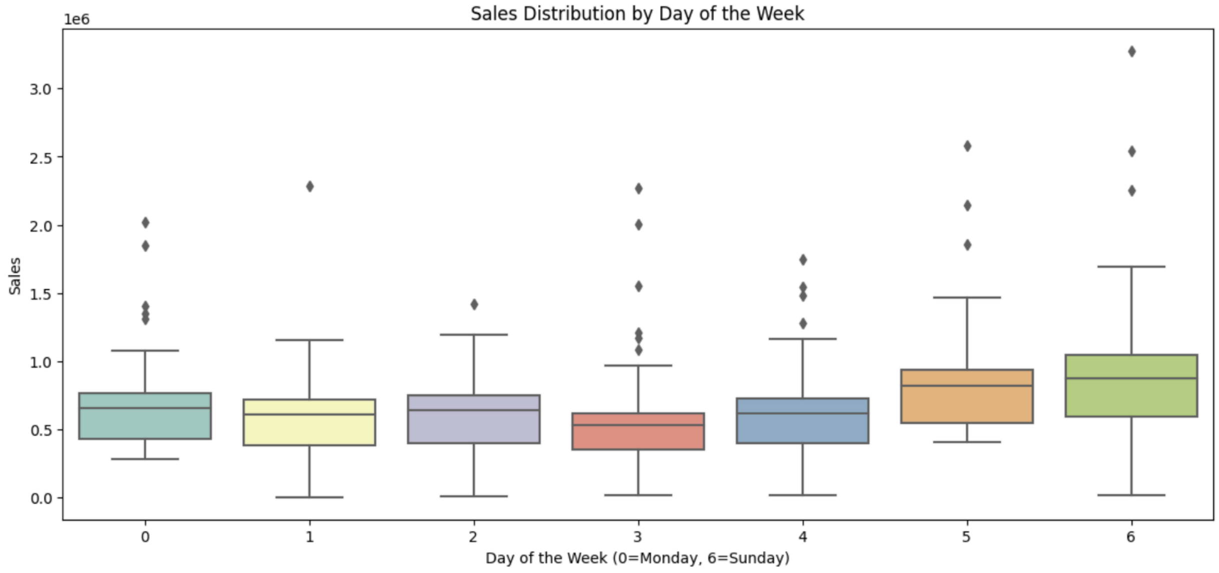
Exploratory Data Analysis (EDA)

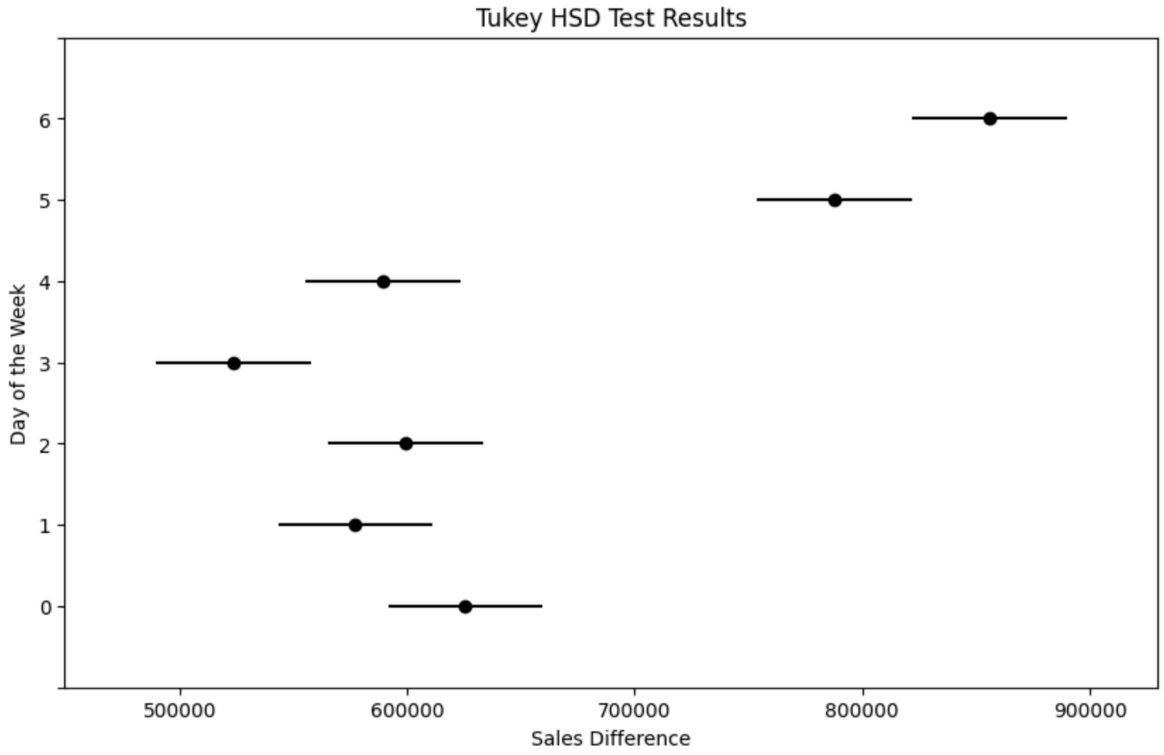
Sales Trends Analysis

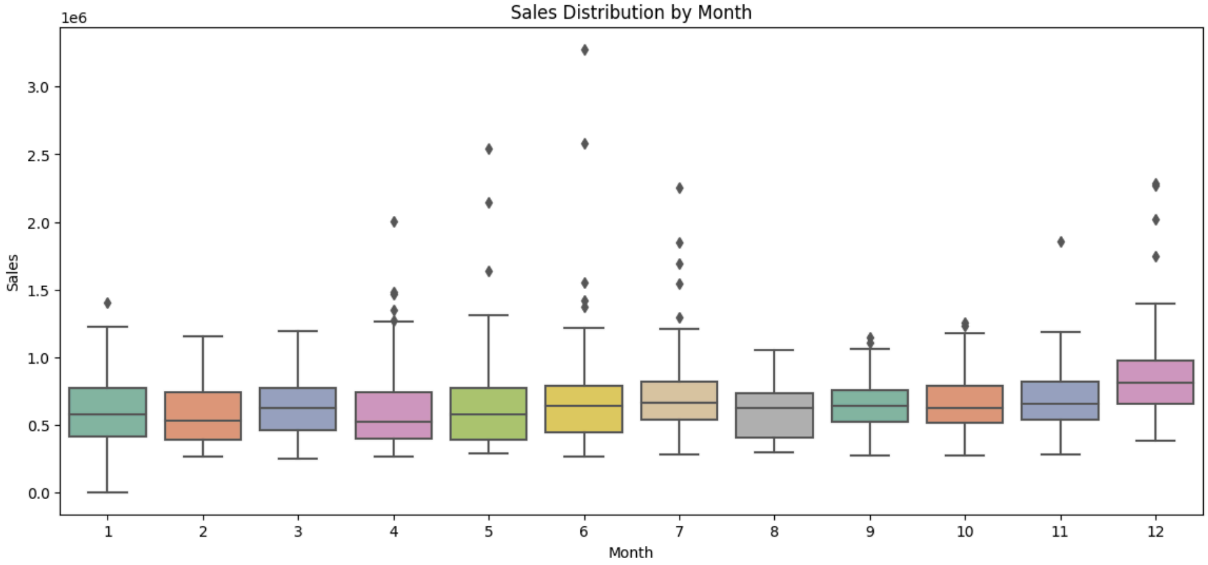
The sales data from 2013 to 2017 reveals a steady upward trend, indicating strong business growth and successful market expansion. There are clear seasonal fluctuations, with notable sales peaks at the end of each year. However, the residuals suggest the presence of other unexplained factors affecting sales, highlighting the need for further investigation into potential external influences such as local events or market dynamics.

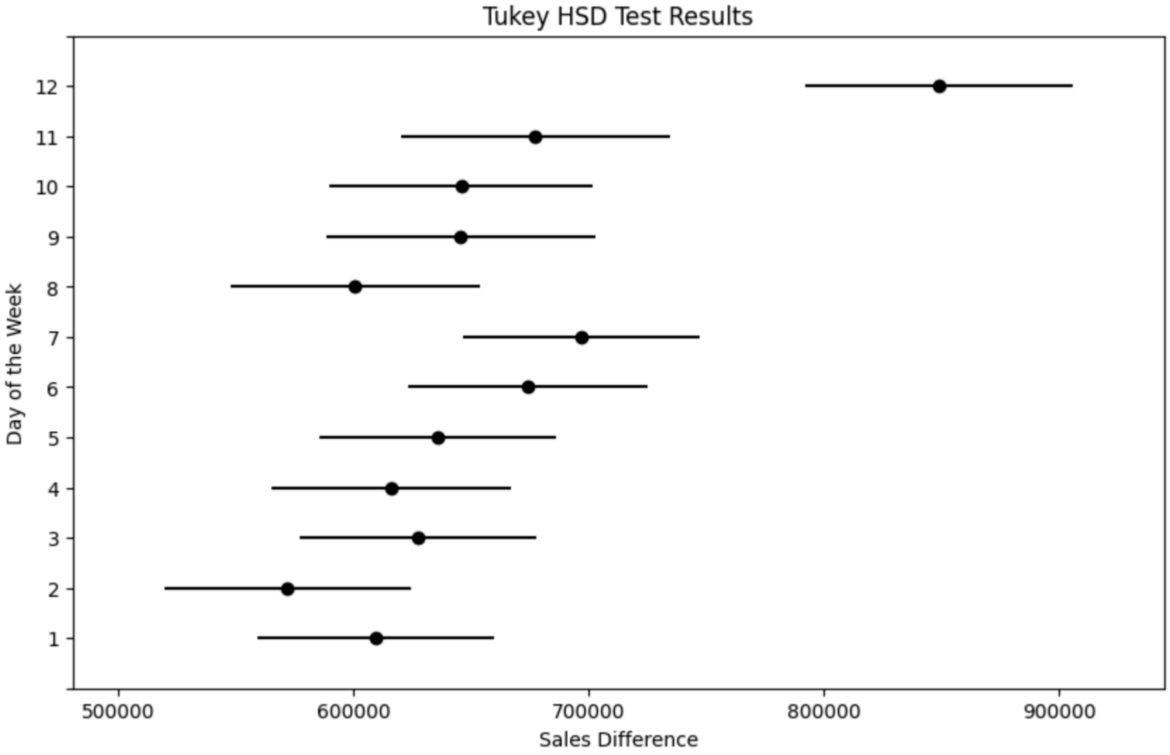


The statistical analysis, including ANOVA and Tukey's HSD, reveals that weekend sales, particularly on Saturdays and Sundays, are significantly higher than weekday sales, with Thursday showing notably lower sales. Additionally, December emerges as the month with significantly higher sales compared to all other months.



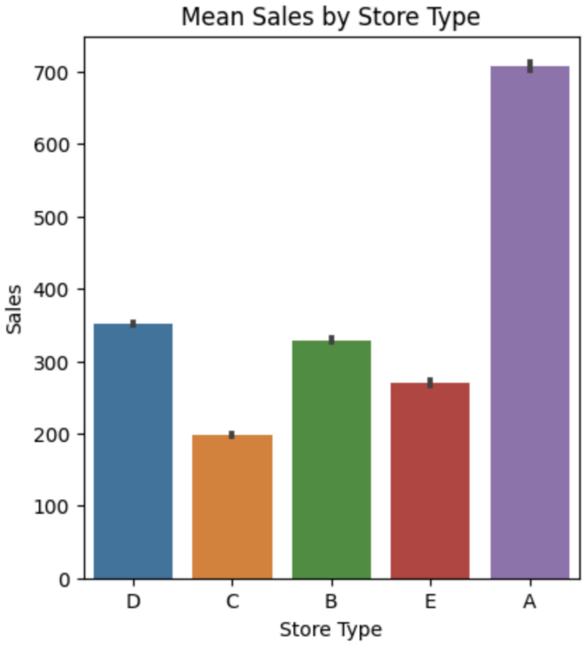

Tukey HSD Test Results
Day of the Week
0
1
2
3
4
5
6
500000
600000
700000
800000
900000
Sales Difference

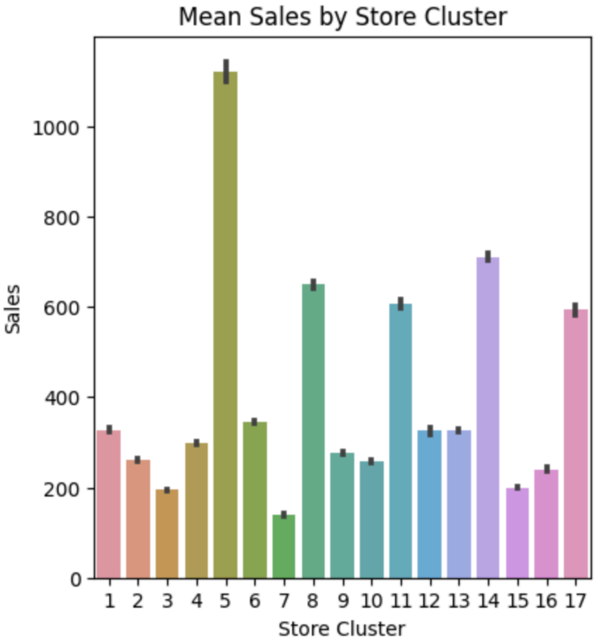

Sales Distribution by Month
1e6
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Sales
1
2
3
4
5
6
7
8
9
10
11
12
Month

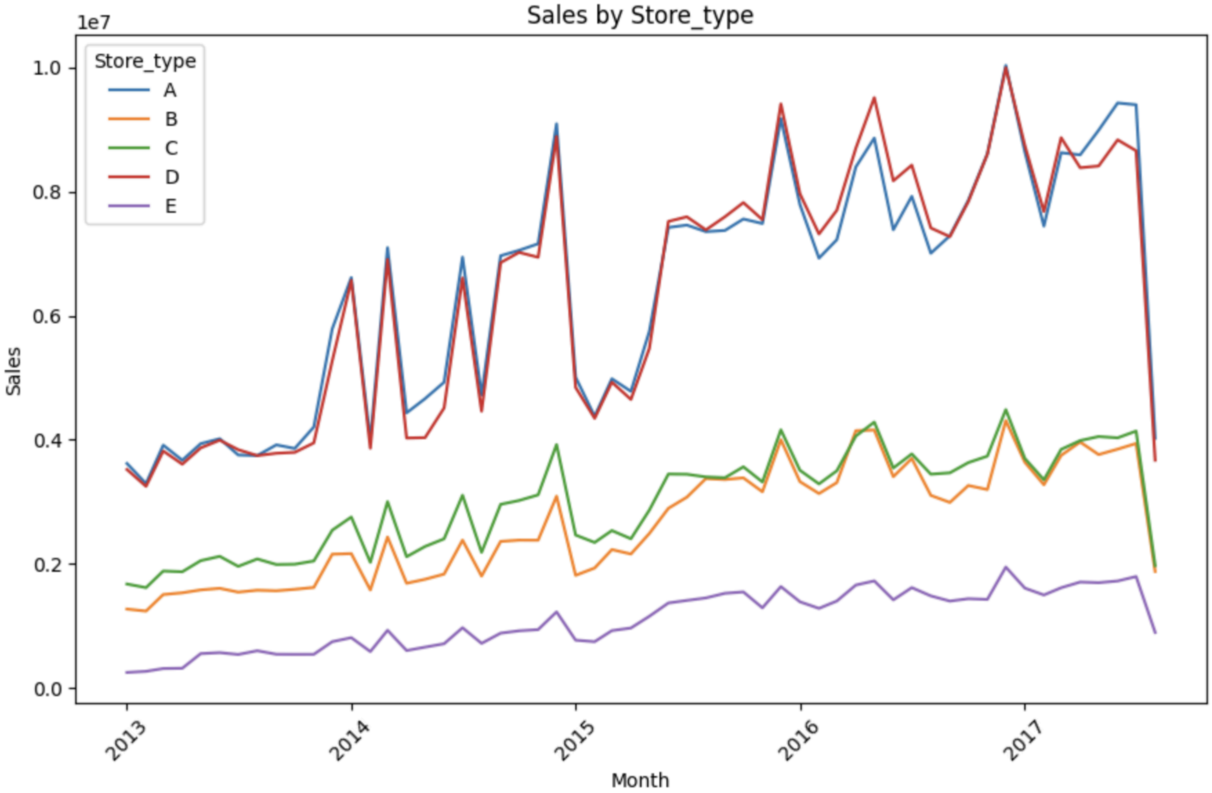


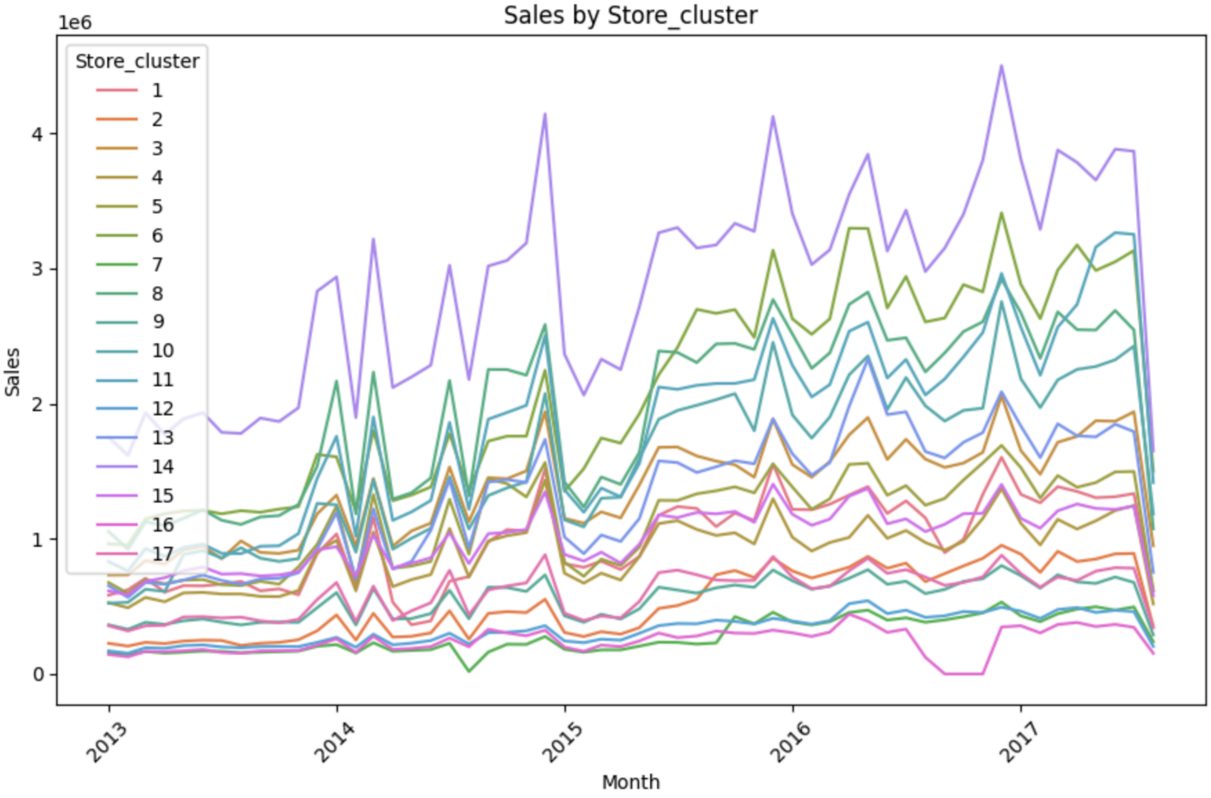
Sales distributions by categorical features

There is a significant variation in sales performance across different store types and clusters. Additionally, the sales trajectories show distinct differences, with certain clusters and types consistently outperforming others. Moreover, all charts reveal distinct seasonal trends, with sales peaking at similar times each year, indicating the influence of seasonal shopping behaviors on sales performance.

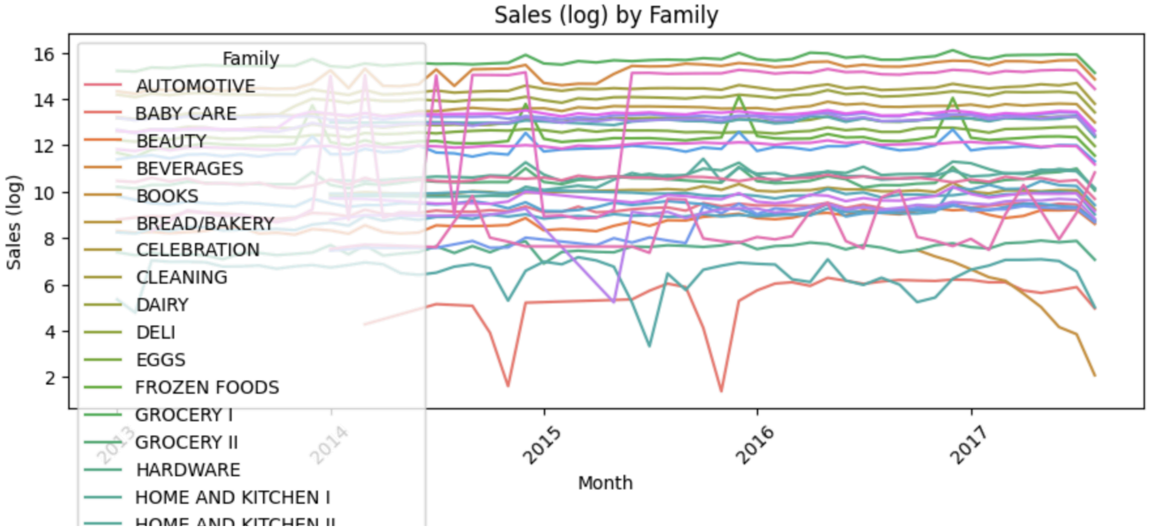

Mean Sales by Store Type
Sales
0
100
200
300
400
500
600
700
D
C
B
E
A
Store Type


Mean Sales by Store Cluster
Sales
0
200
400
600
800
1000
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Store Cluster

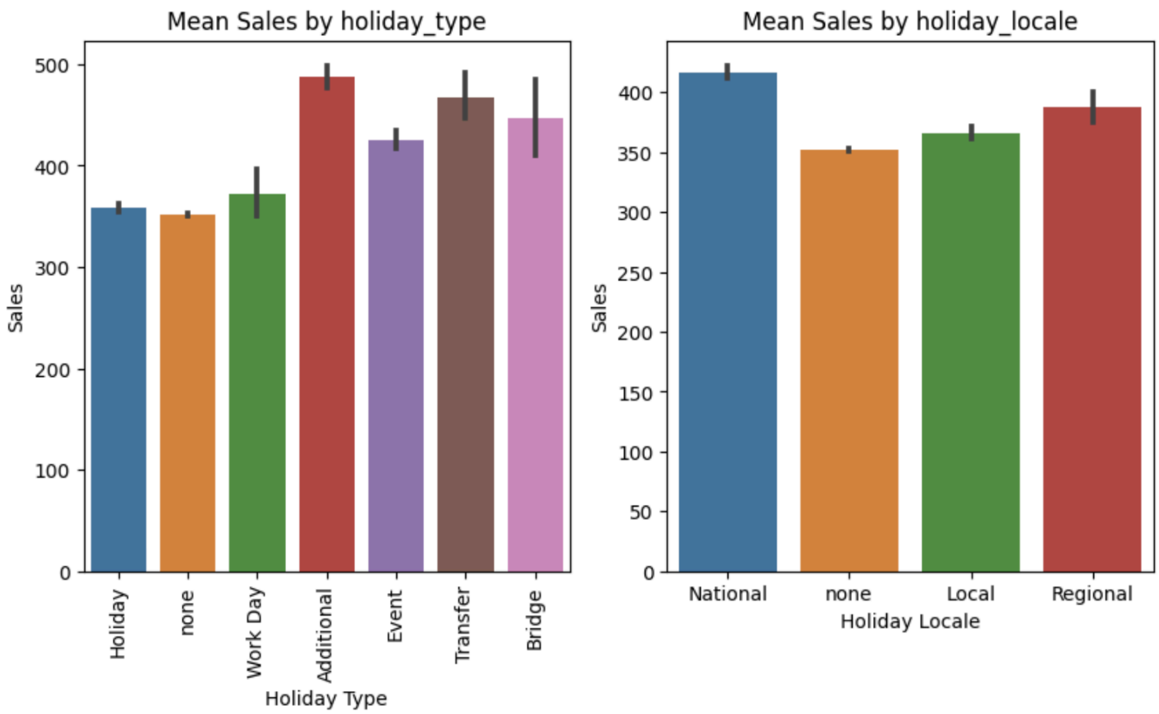

Sales by Store_type
1e7
Sales
0.0
0.2
0.4
0.6
0.8
1.0
Store_type
A
B
C
D
E
2013
2014
2015
2016
2017
Month

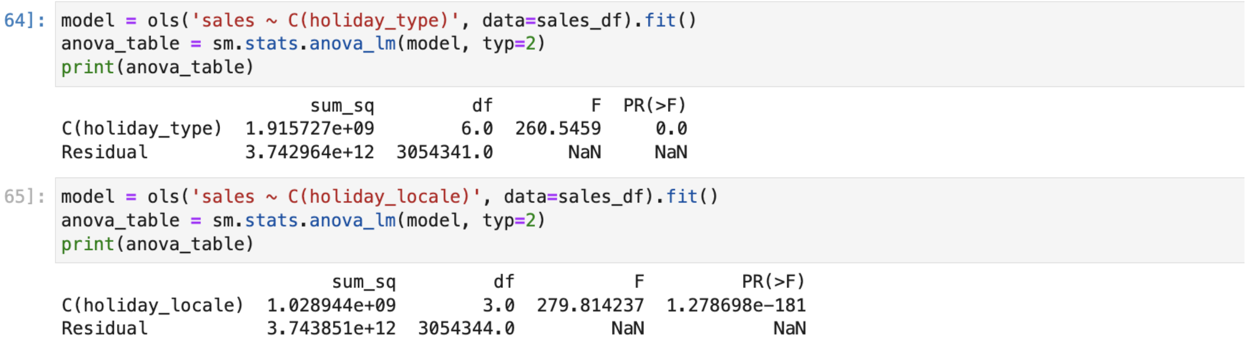


There is a significant variation in sales performance across different families. Some show strong seasonal trends while others don't.



Based on ANOVA test, holiday type, and locale has a significant impact on sales.





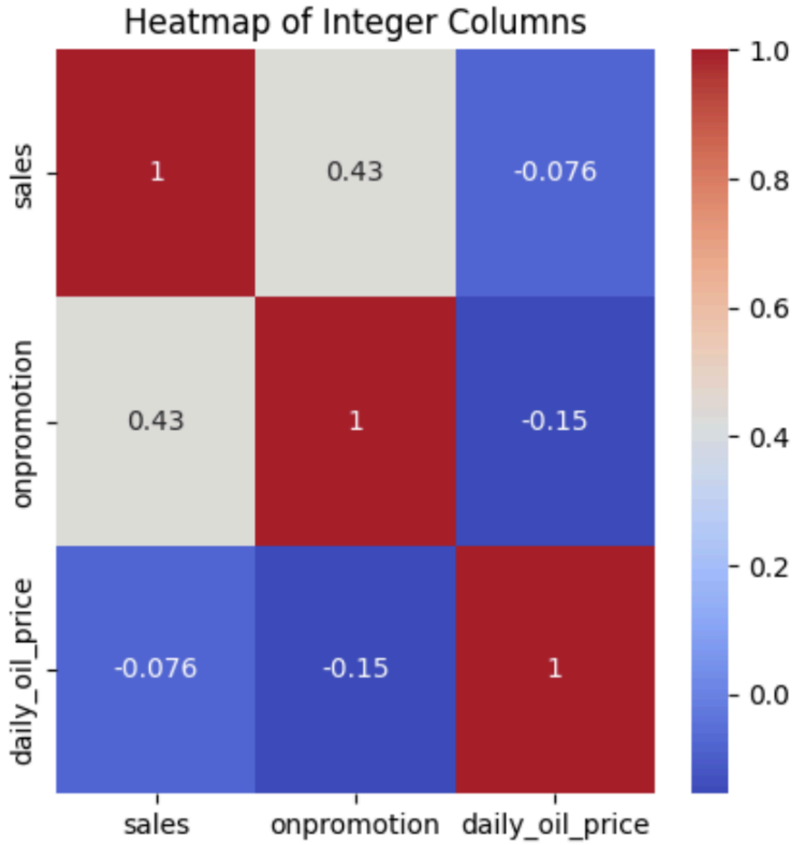
```
64]: model = ols('sales ~ C(holiday_type)', data=sales_df).fit()
     anova_table = sm.stats.anova_lm(model, typ=2)
     print(anova_table)

                           sum_sq         df           F  PR(>F)
     C(holiday_type)  1.915727e+09        6.0  260.5459     0.0
     Residual         3.742964e+12  3054341.0       NaN     NaN
```

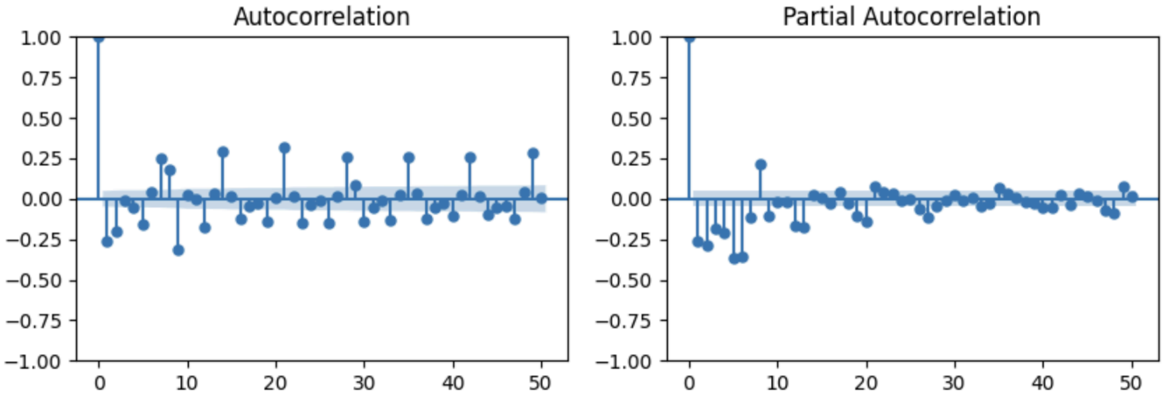
```
65]: model = ols('sales ~ C(holiday_locale)', data=sales_df).fit()
     anova_table = sm.stats.anova_lm(model, typ=2)
     print(anova_table)

                             sum_sq         df           F         PR(>F)
     C(holiday_locale)  1.028944e+09        3.0  279.814237  1.278698e-181
     Residual           3.743851e+12  3054344.0         NaN            NaN
```

Sales show a strong positive correlation with promotions. Additionally, there is a moderate negative correlation with daily oil prices.



The ACF and PACF plots indicate that the time series exhibits significant autocorrelation over multiple lags, suggesting the presence of a trend or seasonality. The PACF plot shows that the influence of subsequent lags diminishes quickly after the first few lags.



Model Preprocessing

- **Standardization**: Data was standardized to have a mean of zero and unit variance, essential for PCA.
- **Feature Engineering**:
    - **Fourier Terms**: Captured yearly, quarterly, and weekly seasonality.
    - **Categorical Encoding**: One-hot encoding for variables like product families, store types, and holidays, enabling models to utilize these features effectively.

Modeling

Model Selection

Three primary models were explored:

- **Linear Regression**: Baseline model with an RMSLE of **0.7743**.
- **ARIMA**: Captured time series characteristics but did not outperform the baseline.
- **LSTM**: Still in progress, but showed potential for capturing long-term dependencies.

Best Model: Boosted Hybrid

- **Boosted Hybrid (Linear Regression + Random Forest)** delivered the best performance, achieving an RMSLE of **0.6841**.
- This model captured complex feature interactions, particularly around promotions and holidays.

Future Scope of Work

Integration of Additional Data Sources

Future work could involve integrating more external data sources, such as weather conditions and competitor pricing. These factors could further improve the accuracy of sales forecasts by accounting for additional external influences on consumer behavior.

Refinement of LSTM Model

The LSTM model, still in its developmental stages, could be refined by tuning its hyperparameters, increasing the look-back period, or using more advanced architectures like bidirectional LSTMs. This would likely improve its ability to model complex temporal patterns in the sales data.

Advanced Validation Techniques

To ensure the robustness of the models, time series cross-validation should be implemented. This approach would better assess model performance and generalizability, especially in the context of predicting future sales.