# Pharmaceutical Sales prediction across multiple stores

## An End-to-End Machine Learning Project

- **NAME: MINAL DEVIKAR**
- **INSTITUTION: DIGICHROME ACADEMY**

# BUSINESS NEED

**OBJECTIVE**

**Sales Forecasting**:

Predict daily sales for Rossman Pharmaceuticals across multiple store locations for the next six weeks.

**IMPORTANCE**

•**Informed Decision-Making**:

- **Financial Planning**: Enables realistic budgets and revenue forecasts.
- **Inventory Management**: Optimizes stock levels to reduce costs associated with overstocking or stockouts.
- **Resource Allocation**: Strategically allocates marketing resources and staffing based on anticipated customer traffic.

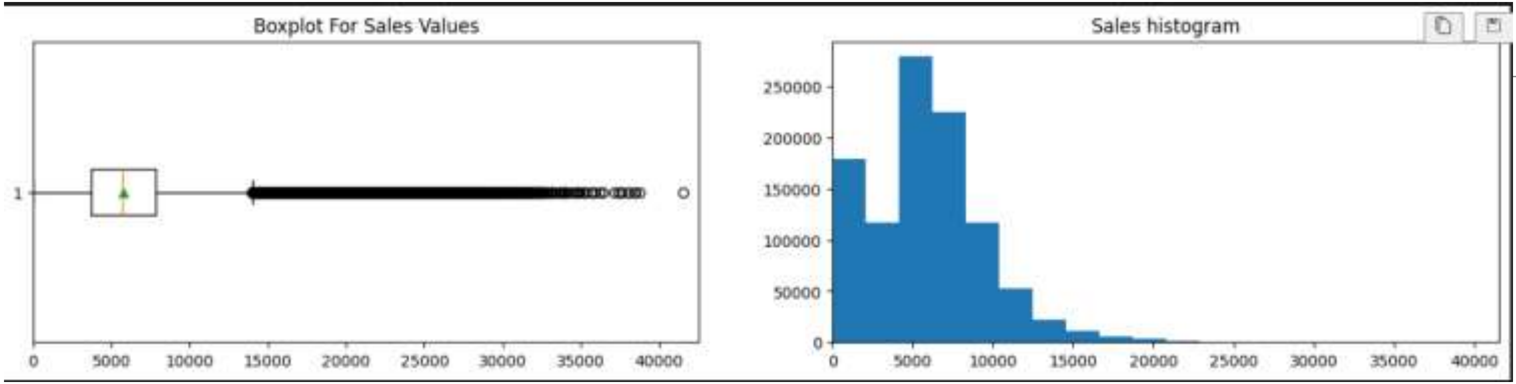
Photo by Peels

# Data Overview

- **Dataset Description**: Daily sales data from multiple stores.

- **Key Features**:
  - Id
  - Store
  - Sales
  - Customers
  - Open (store status)
  - StateHoliday
  - SchoolHoliday
  - StoreType
  - Assortment
  - CompetitionDistance
  - Promo

# Task 1 - Exploration of customer purchasing behaviour

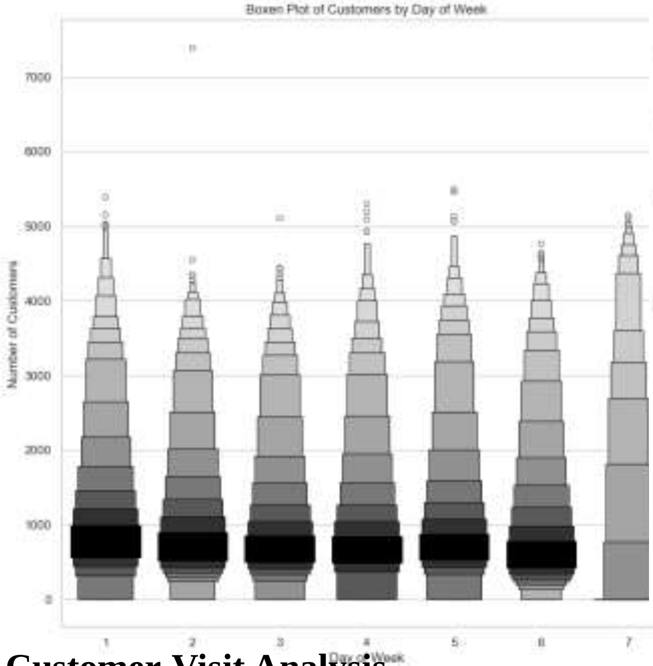


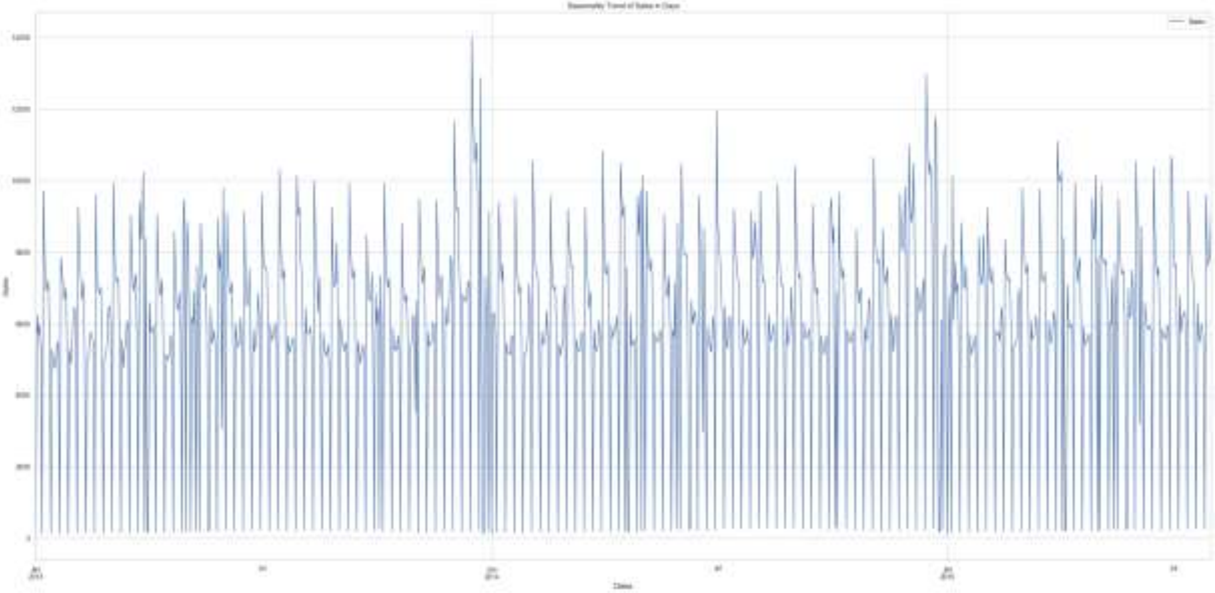
**Boxplot**

- **Sales Concentration**: Most sales are below 10,000.
- **Outliers**: Many outliers up to 40,000.
- **Median**: Near 5,000–6,000 range.

**Histogram**

- **Distribution**: Most sales between 0 and 10,000.
- **Trend**: Sharp decline after 10,000.
- **Skewness**: Right-skewed, indicating few stores have much higher sales.





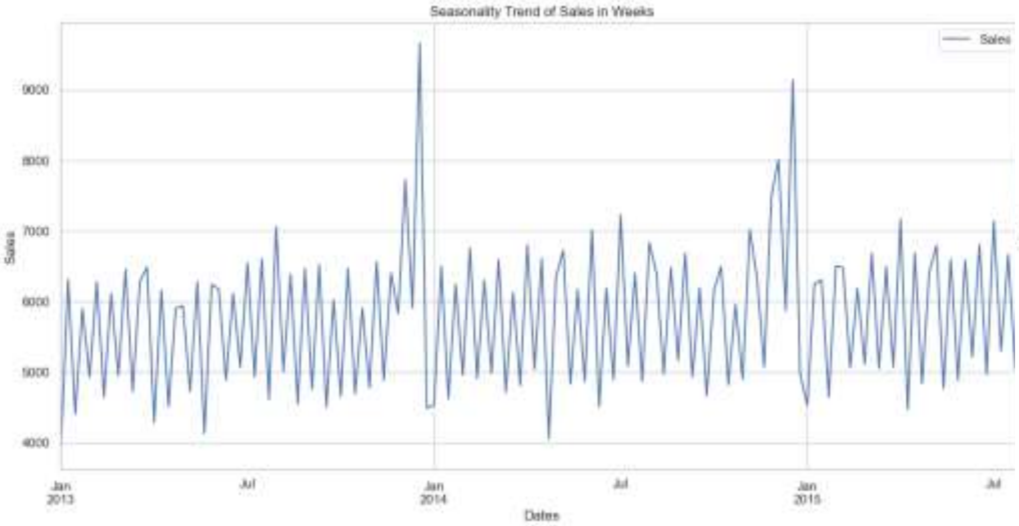
### Customer Visit Analysis

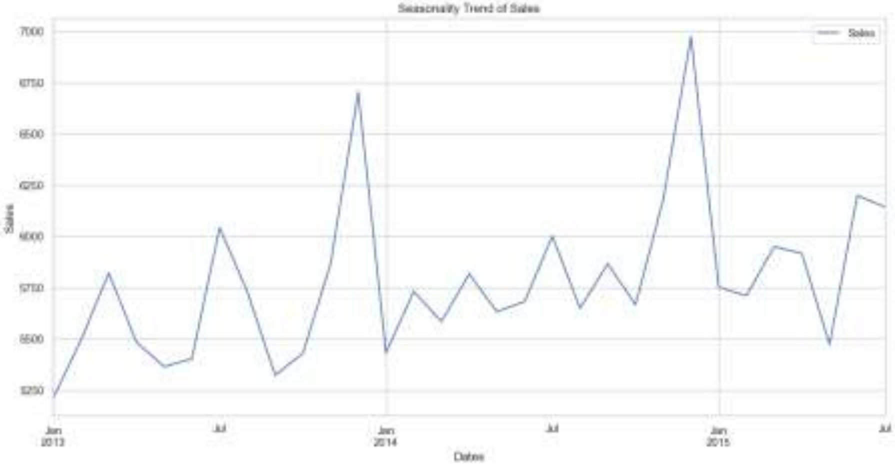
•**Higher Traffic**: Weekends (days 6 and 7) attract more customers than weekdays (days 1-5), with some outliers likely due to promotions.

•**Actionable Insights**: Consider increasing staffing on busy days and implementing promotions on slower days to boost sales.

### Sales Data Patterns

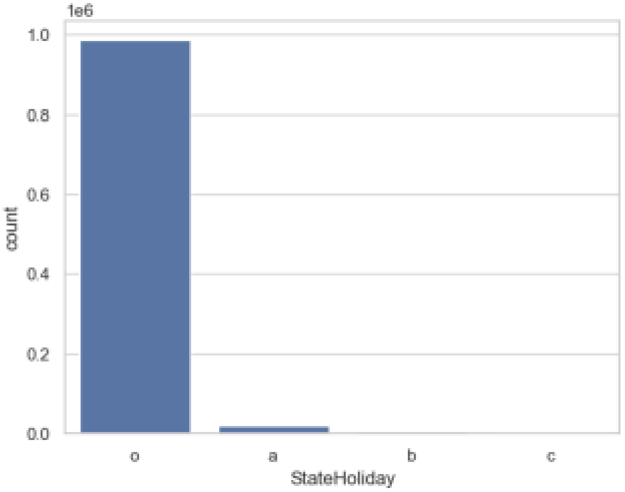
•**Sales Analysis**: The data reveals a clear weekly sales cycle with peaks of 8,000 to 10,000 units, occasional spikes above 12,000, and consistent low points near zero, indicating possible non-business days.





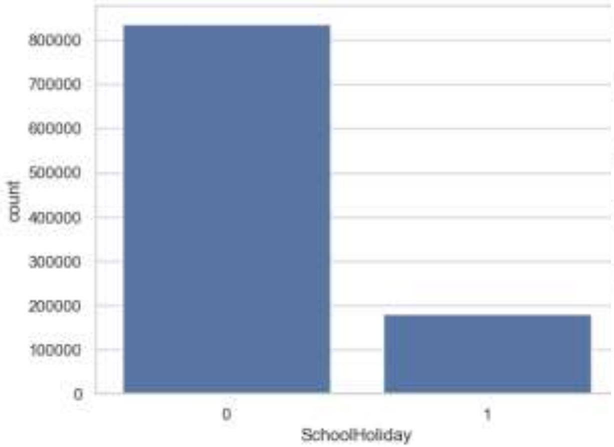
**Strategic Opportunities**: Leverage seasonal trends for inventory optimization, enhance forecasting for financial planning, analyze market fluctuations for deeper insights, and engage customers with targeted marketing during peak sales periods.

**Overview**: Sales from January 2013 to July 2015 show an upward trend with seasonal peaks during holidays; recommend implementing seasonal promotions, adjusting inventory, enhancing customer engagement during low periods, continuous data analysis for emerging trends, and utilizing historical data for accurate forecasting.

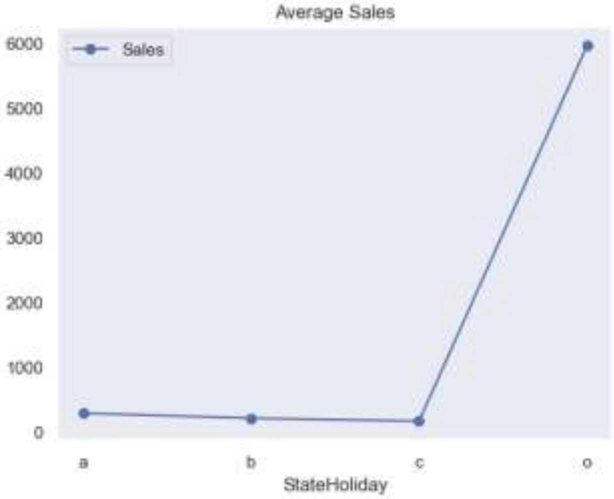


## Holiday Insights

•**Key Findings**: Holiday 'o' dominates with 986,105 occurrences, indicating a regular transaction period; businesses should focus resources during this time while leveraging holidays 'a', 'b', and 'c' for targeted marketing and growth opportunities through themed promotions.

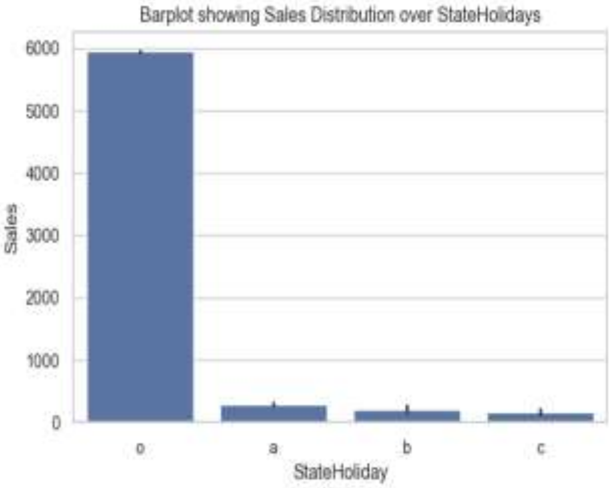


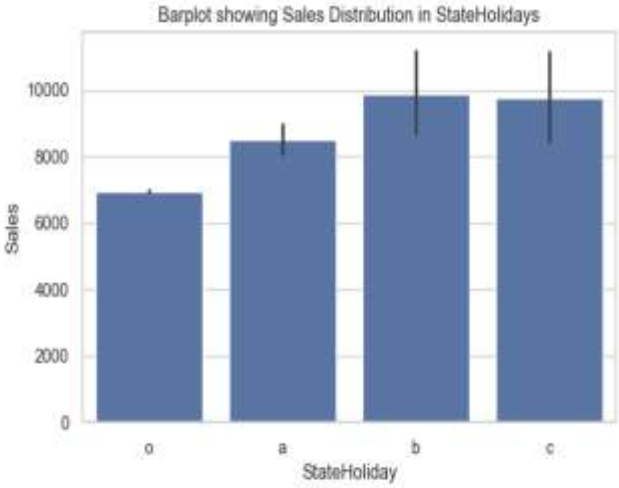
Holiday 'o' indicates a regular transaction period; targeted marketing opportunities exist for holidays 'a', 'b', and 'c', necessitating resource alignment primarily with holiday 'o' while preparing for potential spikes and enhancing lesser holidays through strategic initiatives.

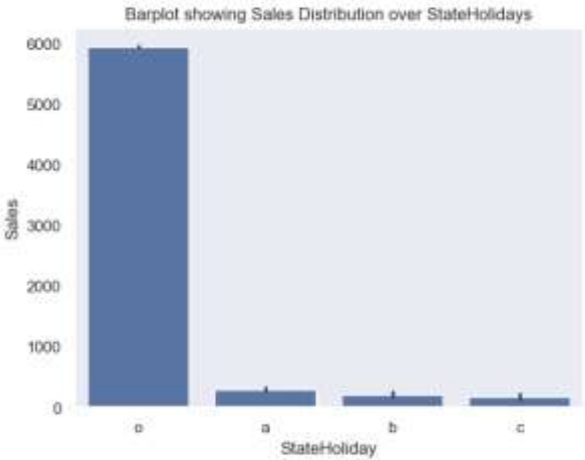


**Seasonal Purchases Behaviour**

Holiday 'o' is the most common period indicating normal activities, while holidays 'a', 'b', and 'c' present unique marketing opportunities; prioritize resources for holiday 'o', enhance lesser holidays through strategic plans, and analyze consumer behavior for deeper insights.



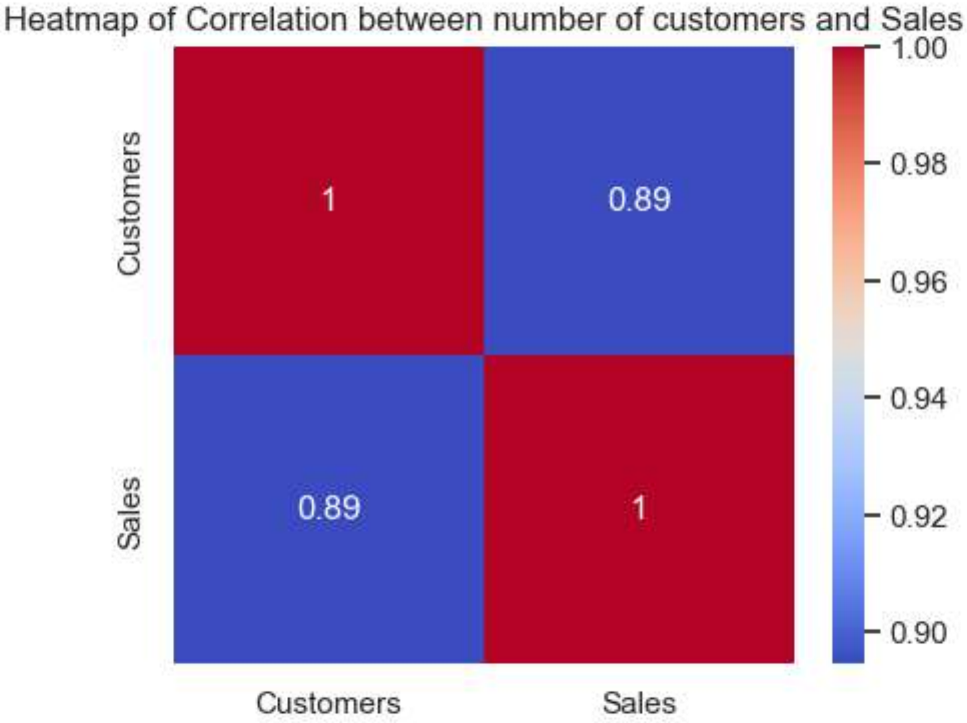




Holiday 'o' dominates regular activities; leverage holidays 'a', 'b', and 'c' for targeted marketing, prioritize resource allocation for holiday 'o', explore growth potential through strategic initiatives for lesser holidays, and analyze consumer behavior for better insights.

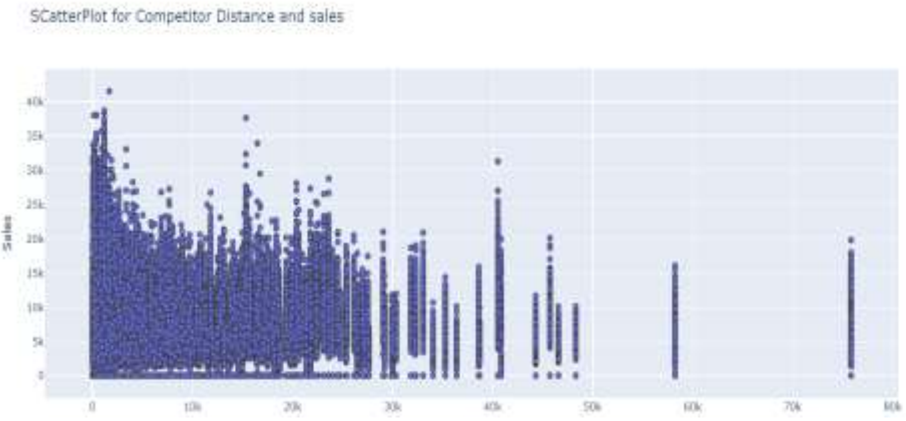
**Sales behavior before, during and after holidays**
Holiday 'o' represents regular periods of high activity; leverage lesser holidays ('a', 'b', 'c') for targeted marketing, focus resource allocation on holiday 'o', explore growth potential for lesser holidays, and analyze consumer behavior to optimize strategies.

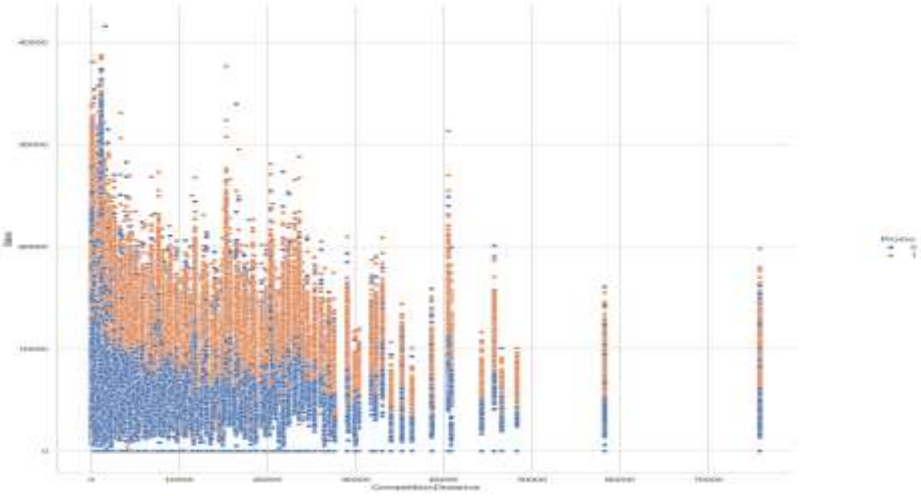
# Heatmap Correlation between number of customers and Sales



**Customer Insights & Recommendations**

•**Key Findings**: A strong positive correlation (0.8947) between customers and sales indicates that increasing customer numbers drives sales; recommend investing in customer acquisition, enhancing retention strategies, utilizing data for targeted marketing, improving sales forecasting, and enhancing the overall customer experience.





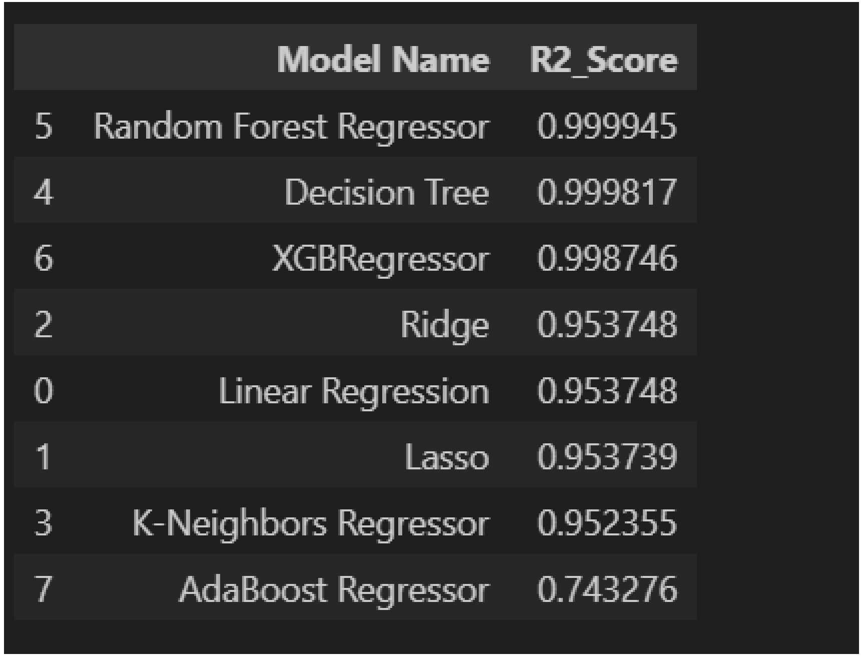
**Competitor Distance Insights** Most data points cluster within 0-30,000 meters competitor distance, with sales generally ranging from 0 to 40,000 units; a slight negative correlation suggests that sales tend to decrease with increased distance from competitors, peaking at close proximity (0-5,000 meters), while variability decreases with distance and outliers are present at longer distances.

**Competition Distance and Promotion Insights**

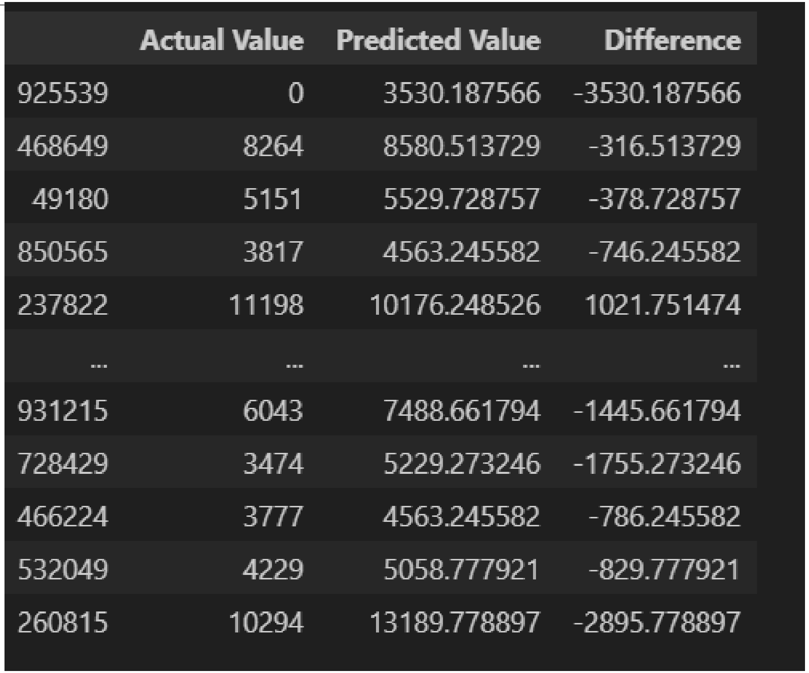
•**Overview**: Most data points cluster at lower competition distances (0-20,000), with sales values evenly distributed; promotions (Promo = 1) tend to boost sales, while a weak negative trend suggests that higher competition distances slightly decrease maximum sales values, indicating that other unmeasured factors may significantly influence sales performance.

# TASK 2 - PREDICTION OF STORE SALES

## Model R2 Score

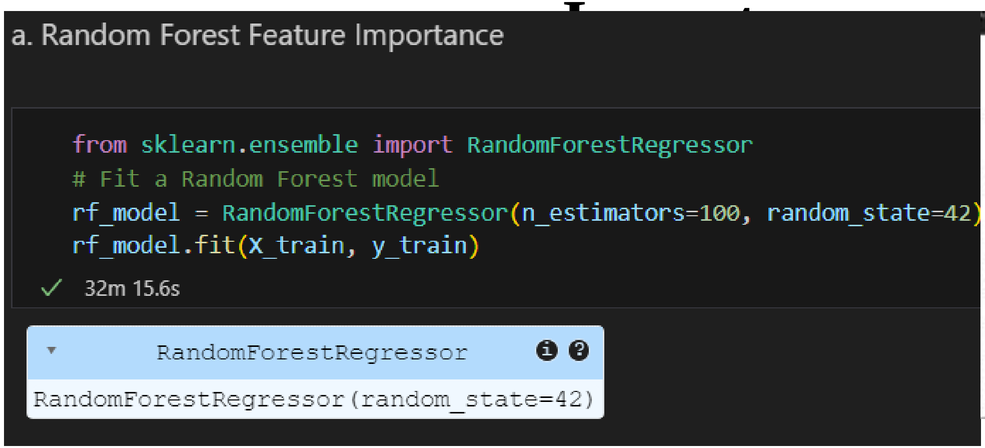


| | Model Name | R2_Score |
|---|---|---|
| 5 | Random Forest Regressor | 0.999945 |
| 4 | Decision Tree | 0.999817 |
| 6 | XGBRegressor | 0.998746 |
| 2 | Ridge | 0.953748 |
| 0 | Linear Regression | 0.953748 |
| 1 | Lasso | 0.953739 |
| 3 | K-Neighbors Regressor | 0.952355 |
| 7 | AdaBoost Regressor | 0.743276 |



| | Actual Value | Predicted Value | Difference |
|---|---|---|---|
| 925539 | 0 | 3530.187566 | -3530.187566 |
| 468649 | 8264 | 8580.513729 | -316.513729 |
| 49180 | 5151 | 5529.728757 | -378.728757 |
| 850565 | 3817 | 4563.245582 | -746.245582 |
| 237822 | 11198 | 10176.248526 | 1021.751474 |
| ... | ... | ... | ... |
| 931215 | 6043 | 7488.661794 | -1445.661794 |
| 728429 | 3474 | 5229.273246 | -1755.273246 |
| 466224 | 3777 | 4563.245582 | -786.245582 |
| 532049 | 4229 | 5058.777921 | -829.777921 |
| 260815 | 10294 | 13189.778897 | -2895.778897 |

# Random Forest Feature

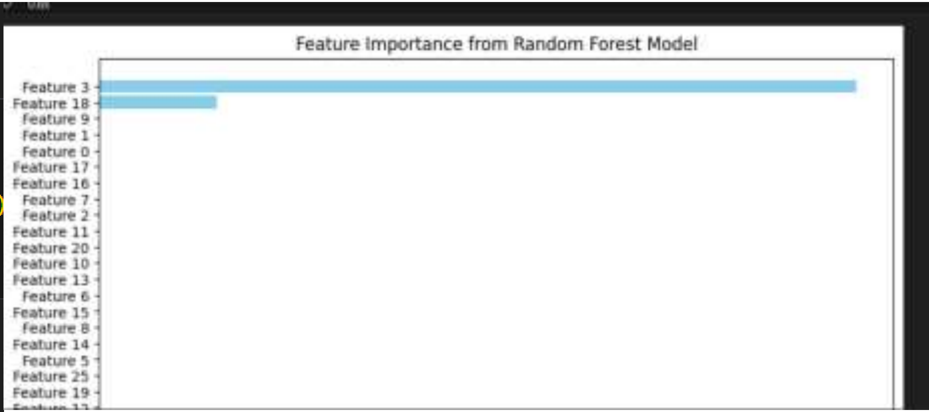


a. Random Forest Feature Importance

```
from sklearn.ensemble import RandomForestRegressor
# Fit a Random Forest model
rf_model = RandomForestRegressor(n_estimators=100, random_state=42)
rf_model.fit(X_train, y_train)
```

✓ 32m 15.6s

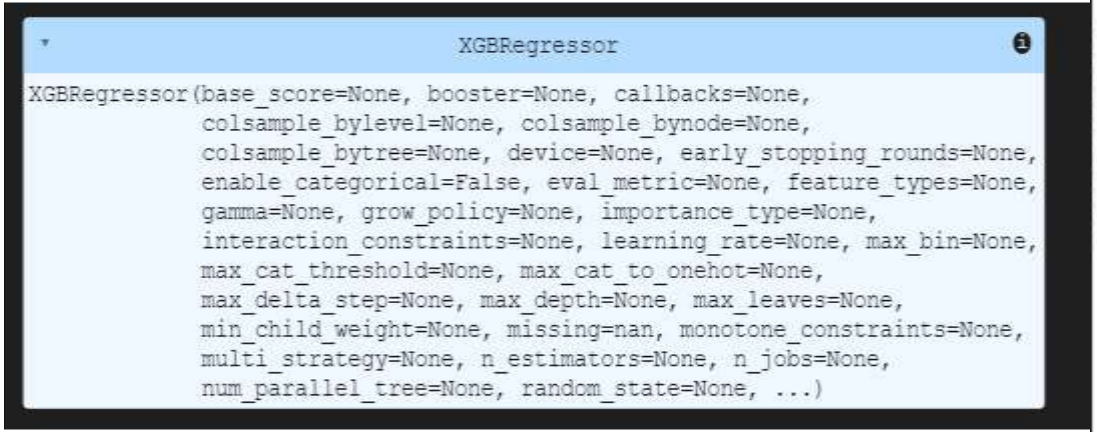
```
RandomForestRegressor
RandomForestRegressor(random_state=42)
```



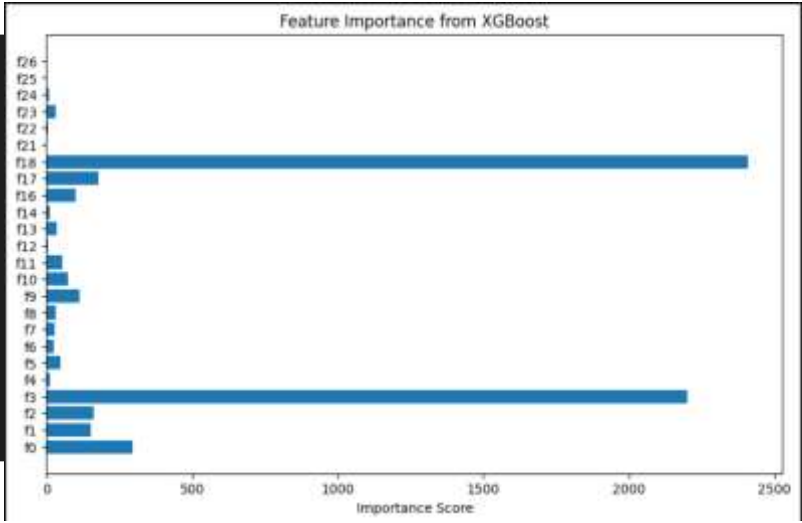
Feature Importance from Random Forest Model

**Feature Importance Insights:**
Feature 3 is the most significant contributor to model performance, followed by Feature 18, while Features 1, 0, and others show minimal impact on sales prediction.

# XGBoost Feature Importance



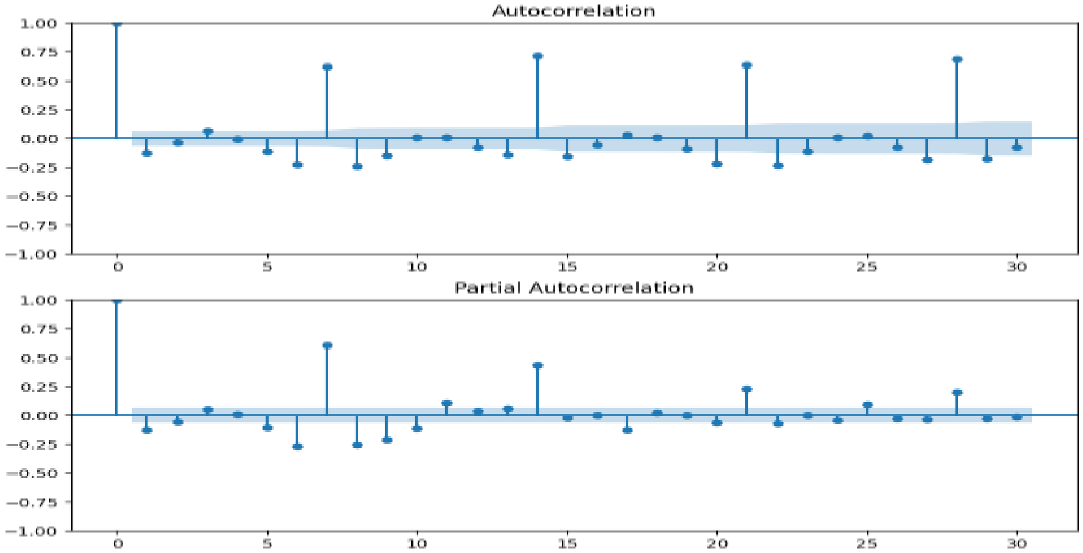
```
XGBRegressor(base_score=None, booster=None, callbacks=None,
             colsample_bylevel=None, colsample_bynode=None,
             colsample_bytree=None, device=None, early_stopping_rounds=None,
             enable_categorical=False, eval_metric=None, feature_types=None,
             gamma=None, grow_policy=None, importance_type=None,
             interaction_constraints=None, learning_rate=None, max_bin=None,
             max_cat_threshold=None, max_cat_to_onehot=None,
             max_delta_step=None, max_depth=None, max_leaves=None,
             min_child_weight=None, missing=nan, monotone_constraints=None,
             multi_strategy=None, n_estimators=None, n_jobs=None,
             num_parallel_tree=None, random_state=None, ...)
```

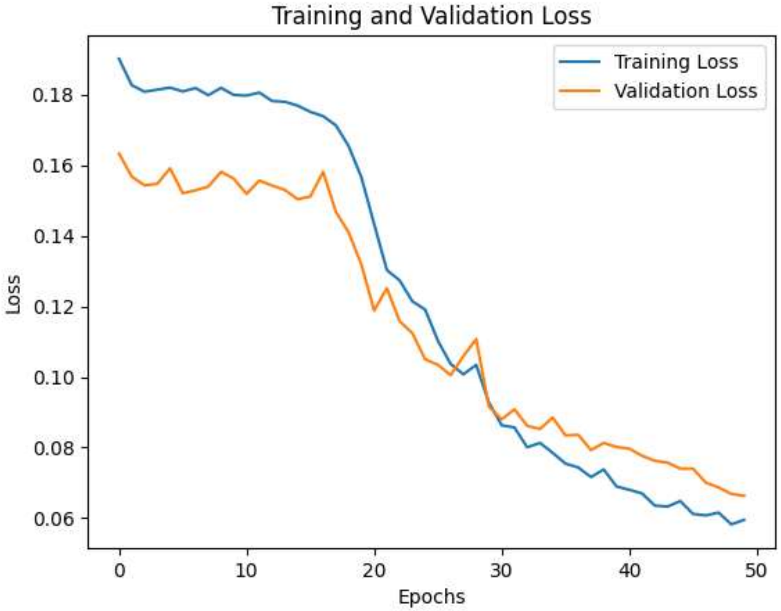


**Insights & Recommendations:**
Prioritize key features like f21 and f4, as they are the strongest sales predictors. Simplify the model by potentially removing low-impact features (e.g., f7, f8, f12). Focus on enhancing top features, and consider feature engineering for complex patterns. Regularly update the model to maintain prediction accuracy by adjusting to shifts in feature importance.

# TASK 3 - BUILDING MODEL WITH DEEP LEARNING

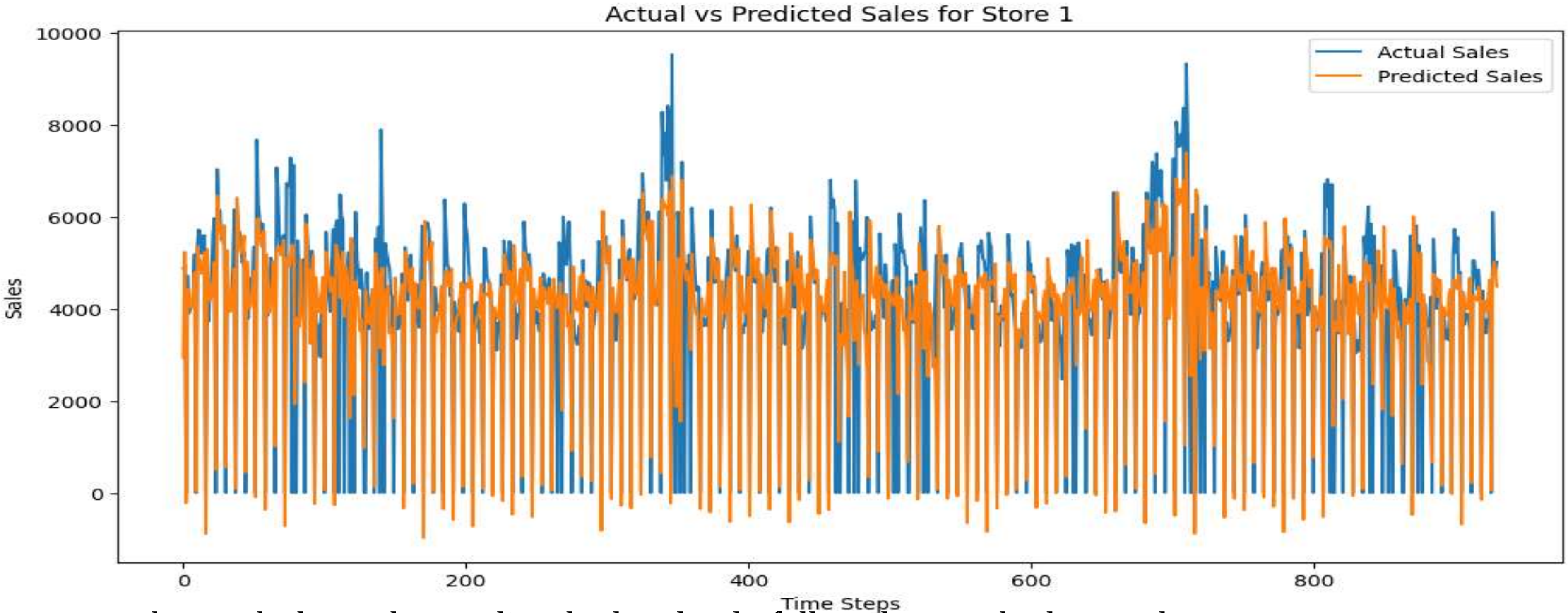
## Check for autocorrelation and partial autocorrelation





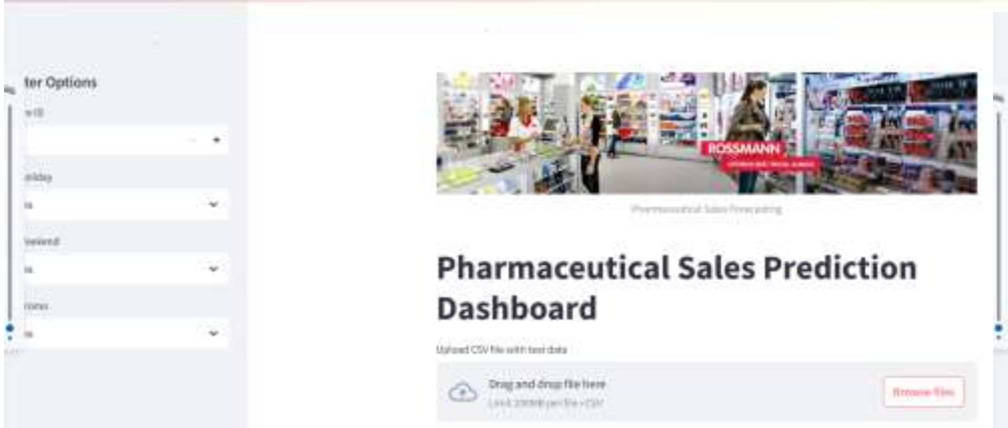
The ACF and PACF plots indicate significant lags (up to lag 5 and specific others), suggesting an ARIMA model with these parameters may effectively capture the autocorrelation pattern.

The model learns effectively with both losses decreasing steadily, leveling out around 30 epochs, suggesting early stopping could optimize training time.



The graph shows that predicted sales closely follow the actual sales trend, although there are several instances where the predicted values fluctuate more than actual values. This suggests the model is reasonably accurate but may sometimes overestimate or underestimate sales. To improve accuracy, consider fine-tuning the model or using additional data features.

# Task 4 - Serving predictions on a web interface



ter Options

re ID

oliday

No

Weekend

No

romo

No

Pharmaceutical Sales Forecasting

## Pharmaceutical Sales Prediction Dashboard

Upload CSV file with test data

Drag and drop file here
Limit 200MB per file • CSV

ilter Options

ore ID

1

Holiday

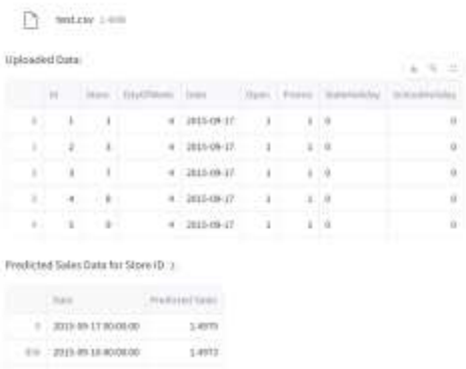
Yes

Weekend

Yes

Promo

Yes



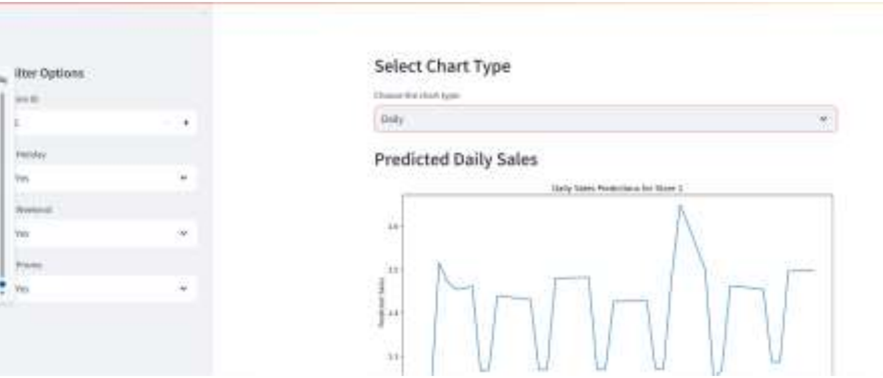
test.csv 1.4KB

Uploaded Data:

| | Id | Store | DayOfWeek | Date | Open | Promo | StateHoliday | SchoolHoliday |
|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 4 | 2015-09-17 | 1 | 1 | 0 | 0 |
| 1 | 2 | 3 | 4 | 2015-09-17 | 1 | 1 | 0 | 0 |
| 2 | 3 | 7 | 4 | 2015-09-17 | 1 | 1 | 0 | 0 |
| 3 | 4 | 8 | 4 | 2015-09-17 | 1 | 1 | 0 | 0 |
| 4 | 5 | 9 | 4 | 2015-09-17 | 1 | 1 | 0 | 0 |

Predicted Sales Data for Store ID 1:

| | Date | Predicted Sales |
|---|---|---|
| 0 | 2015-09-17 00:00:00 | 1.4975 |
| 856 | 2015-09-18 00:00:00 | 1.4972 |



ilter Options

ore ID

1

Holiday

Yes

Weekend

Yes

Promo

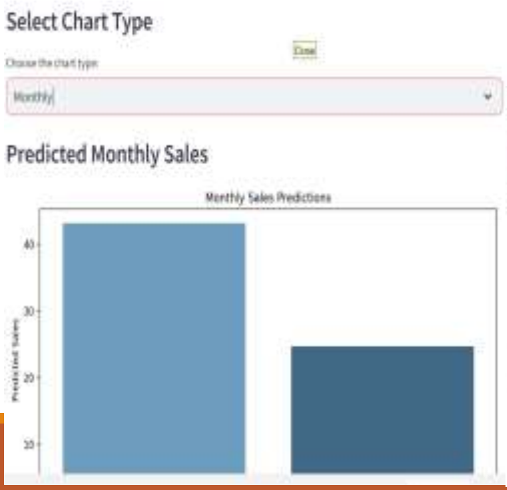
Yes

### Select Chart Type

Choose the chart type

Daily

### Predicted Daily Sales

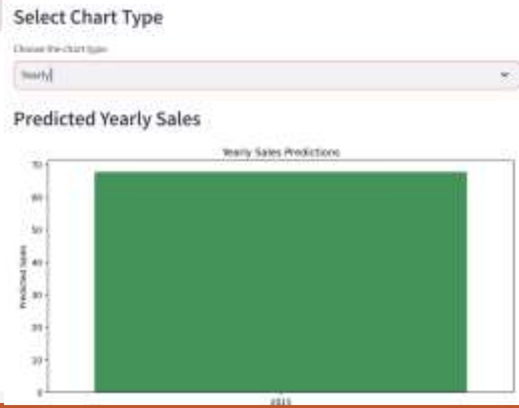


### Select Chart Type

Choose the chart type

Monthly

### Predicted Monthly Sales



### Select Chart Type

Choose the chart type

Yearly

### Predicted Yearly Sales

# Interpretation and Recommendation

## Interpretation

The predicted sales align well with actual sales patterns, capturing overall trends effectively. However, occasional deviations indicate that the model might not fully account for some influencing factors, leading to slight overestimations or underestimations in certain instances.

## Recommendations

1. **Model Fine-tuning:** Experiment with hyperparameter tuning to reduce fluctuations in predictions and increase precision.
2. **Additional Features:** Incorporate new features (e.g., economic indicators, competitor promotions) to capture factors that may influence sales.
3. **Regular Updates:** Retrain the model periodically to incorporate recent data, ensuring it adapts to evolving sales patterns.
4. **Ensemble Techniques:** Consider using ensemble methods to enhance robustness and further smooth out prediction fluctuations.

# THANKYOU