Capstone Project (007)

Eeshan Gautam

MS Business Analytics **(M14828650)**

Optimizing Supply Chain Operations by predicting Store-level Weekly Sales

1. Introduction
2. Objective
3. Data Exploration
4. Methodology
5. Discussion
6. References

LETS GET STARTED

Introduction

- Accurate sales forecasting is Crucial for effective supply chain management in the retail industry.
- This research project will utilize the "Walmart Sales Forecasting" dataset, which includes historical sales data for Walmart stores, to develop a comprehensive solution that enhances demand forecasting accuracy and enables optimized decision-making for inventory management, and related supply chain operations.

Objectives

1. Demand Forecasting

2. Inventory Optimization

1. Demand Forecasting:

- Develop and compare demand forecasting models using time series analysis, regression, and machine learning algorithms.

- While considering factors such as store location, item attributes, temporal patterns, promotions, and external events to accurately predict future sales for different stores and items.

- Continued...

Objectives

1. Demand Forecasting

2. Inventory Optimization

2. Inventory Optimization:

- Utilize the demand forecasting models to optimize inventory management decisions, which can further help to achieve Inventory Optimization. And improving the Supply Chain Management at Walmart.

- Consider store-specific demand variability, lead times, service level requirements, and cost constraints to minimize inventory holding costs while ensuring sufficient stock availability.

Data Exploration

Raw Data is stored in these files from these data sources

- walmart
- stores
- features

```
walmart.head(5)
```

	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

```
store.head()
```

	Store	Type	Size
0	1	A	151315
1	2	A	202307
2	3	B	37392
3	4	A	205863
4	5	B	34875

```
feature.head()
```

	Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	2010-02-05	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	False
1	1	2010-02-12	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	True
2	1	2010-02-19	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	8.106	False
3	1	2010-02-26	46.63	2.561	NaN	NaN	NaN	NaN	NaN	211.319643	8.106	False
4	1	2010-03-05	46.50	2.625	NaN	NaN	NaN	NaN	NaN	211.350143	8.106	False

Data Exploration

Combined dataframe having all information from sales-data, stores-data, and features-data.

- Weekly Sales is the Target Variable
- No missing data in any column except Markdowns 1-5
- Description of Variables
  1. **Store** : anonymized stores of Walmart (45 nos.)
  2. **Date** : Date of Friday, when sales data is collected for the week
  3. **Weekly Sales** : aggregated sales for that week in a store
  4. **Type** : anonymized category of Walmart Store (as A,B,C)
  5. **Size** : Area of the Walmart Store
  6. **Temperature** : Avg. Temp. for Week around that store
  7. **Fuel Price** : Avg. Price for Fuel around that store
  8. **Markdowns** : Related to promotional markdowns being run
  9. **CPI** : Avg. CPI (Consumer Price Index) around that store
  10. **Unemployment** : Avg. Unemployment Rate around that
  11. **Is Holiday** : Tells whether it was a Week with a Holiday

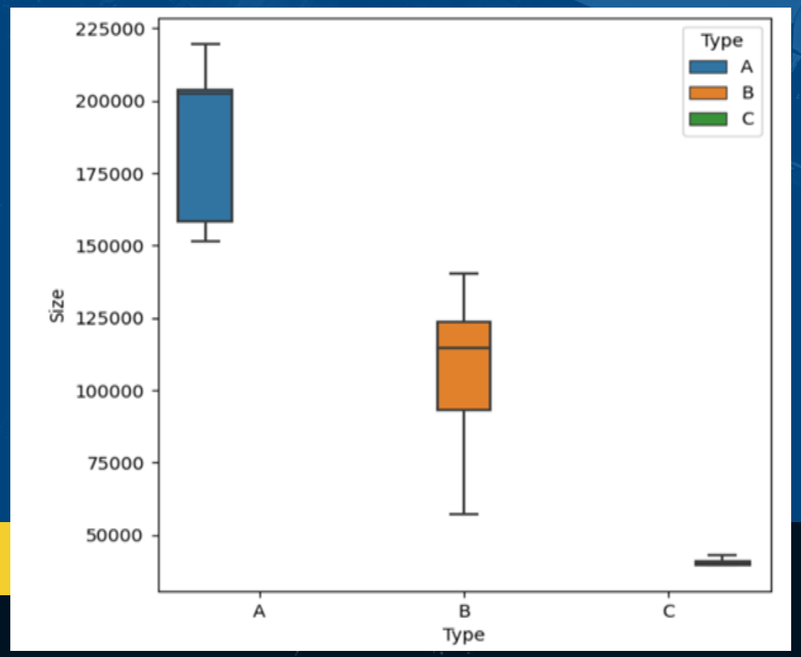
```
print(data.shape)
data.info()
```

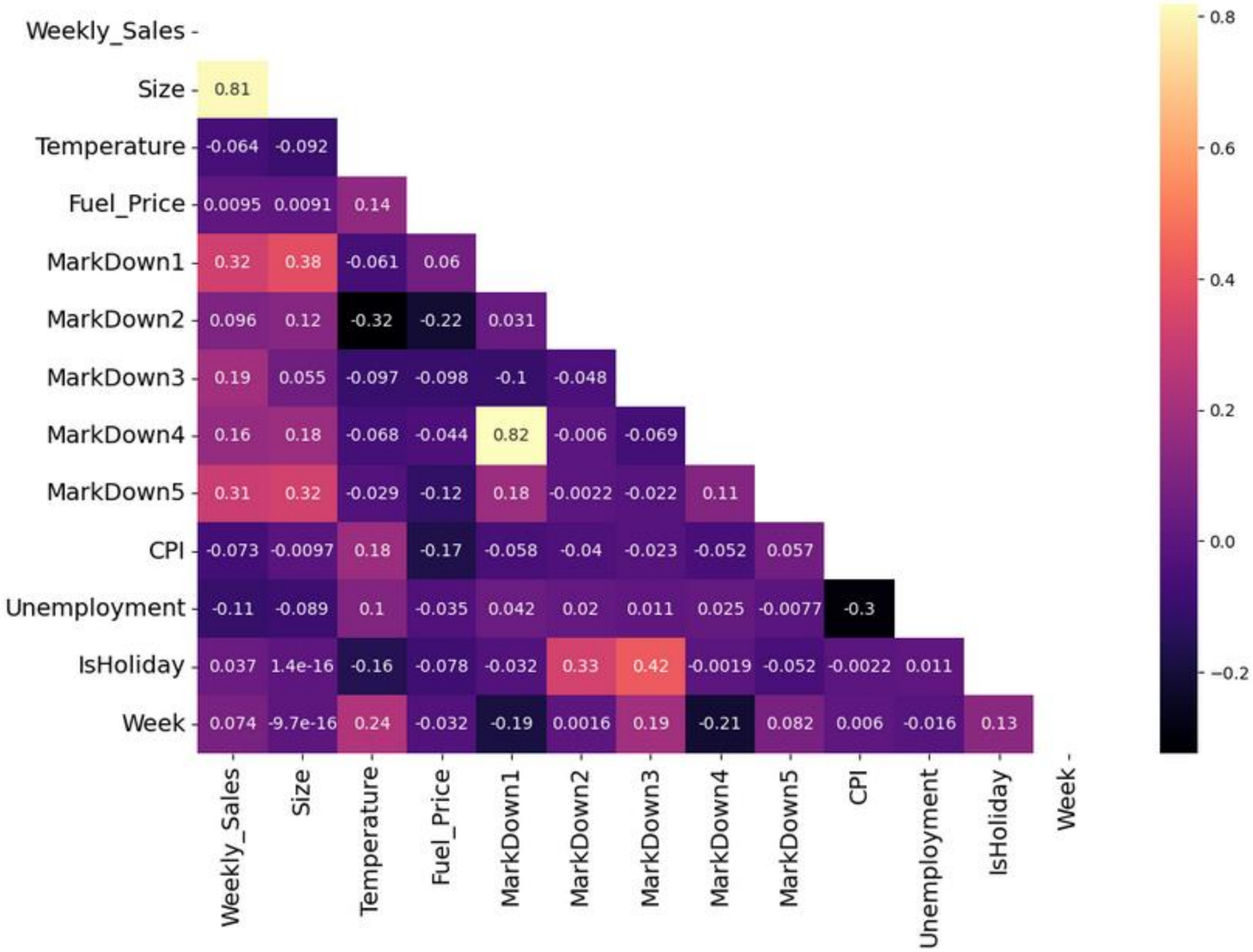
```
(6435, 15)
<class 'pandas.core.frame.DataFrame'>
Int64Index: 6435 entries, 0 to 6434
Data columns (total 15 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   Store          6435 non-null   int64
 1   Date           6435 non-null   object
 2   Weekly_Sales   6435 non-null   float64
 3   Type           6435 non-null   object
 4   Size           6435 non-null   int64
 5   Temperature    6435 non-null   float64
 6   Fuel_Price     6435 non-null   float64
 7   MarkDown1      2280 non-null   float64
 8   MarkDown2      1637 non-null   float64
 9   MarkDown3      2046 non-null   float64
 10  MarkDown4      1965 non-null   float64
 11  MarkDown5      2295 non-null   float64
 12  CPI            6435 non-null   float64
 13  Unemployment   6435 non-null   float64
 14  IsHoliday      6435 non-null   bool
dtypes: bool(1), float64(10), int64(2), object(2)
memory usage: 760.4+ KB
```

Data Exploration

- Correlation Matrix with all variables

- Store Type is a function of Store Size



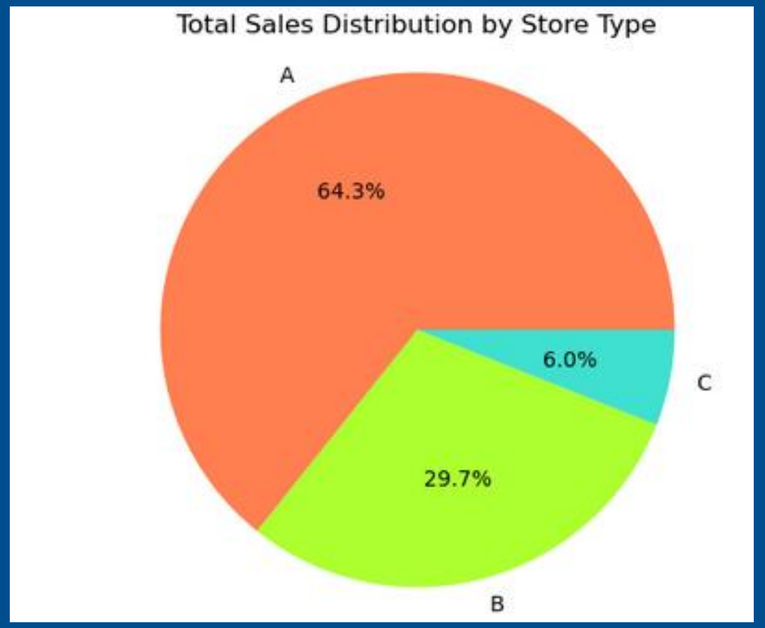


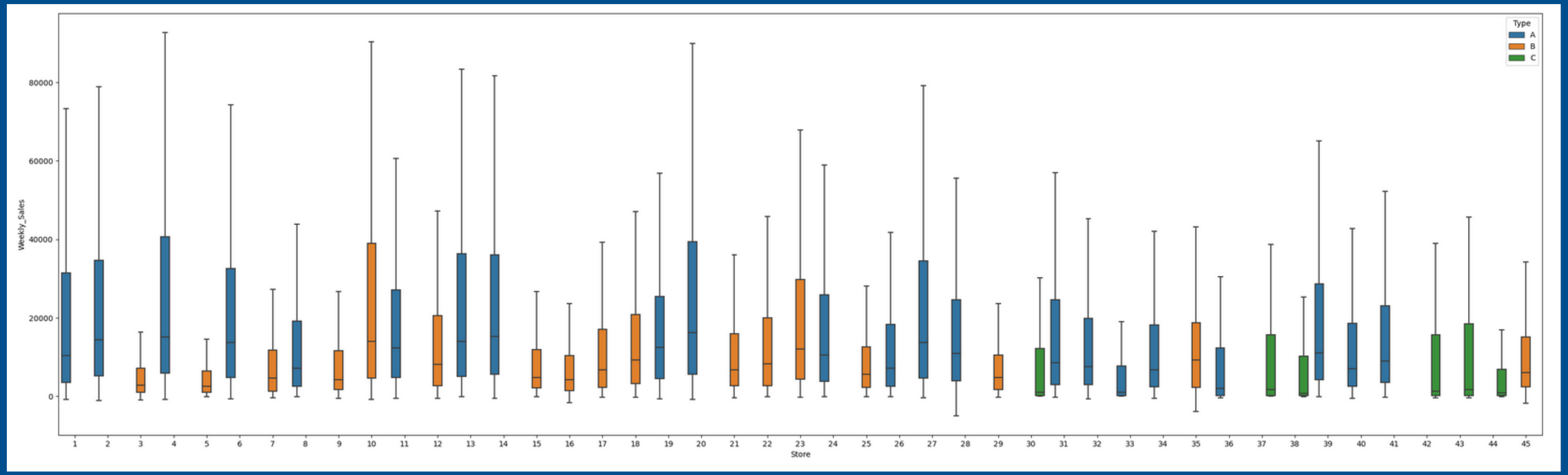
Data Exploration

Variation in Weekly Sales with the Type of Store

- Typically Stores of type A see more Weekly_Sales than B and C types have the least

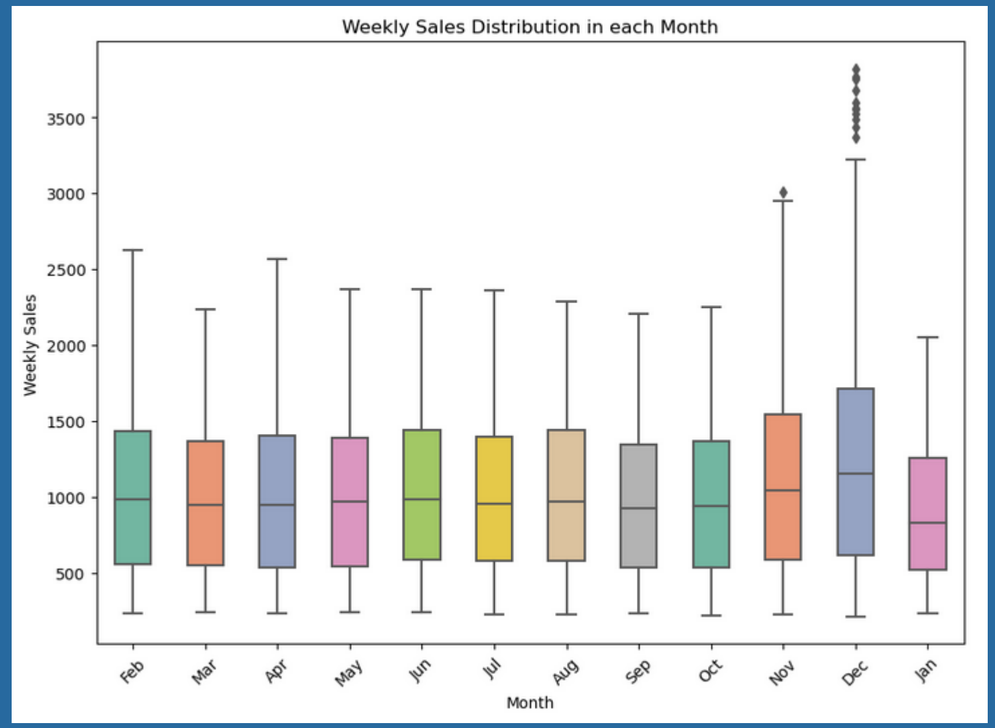
- Weekly Sales for all stores



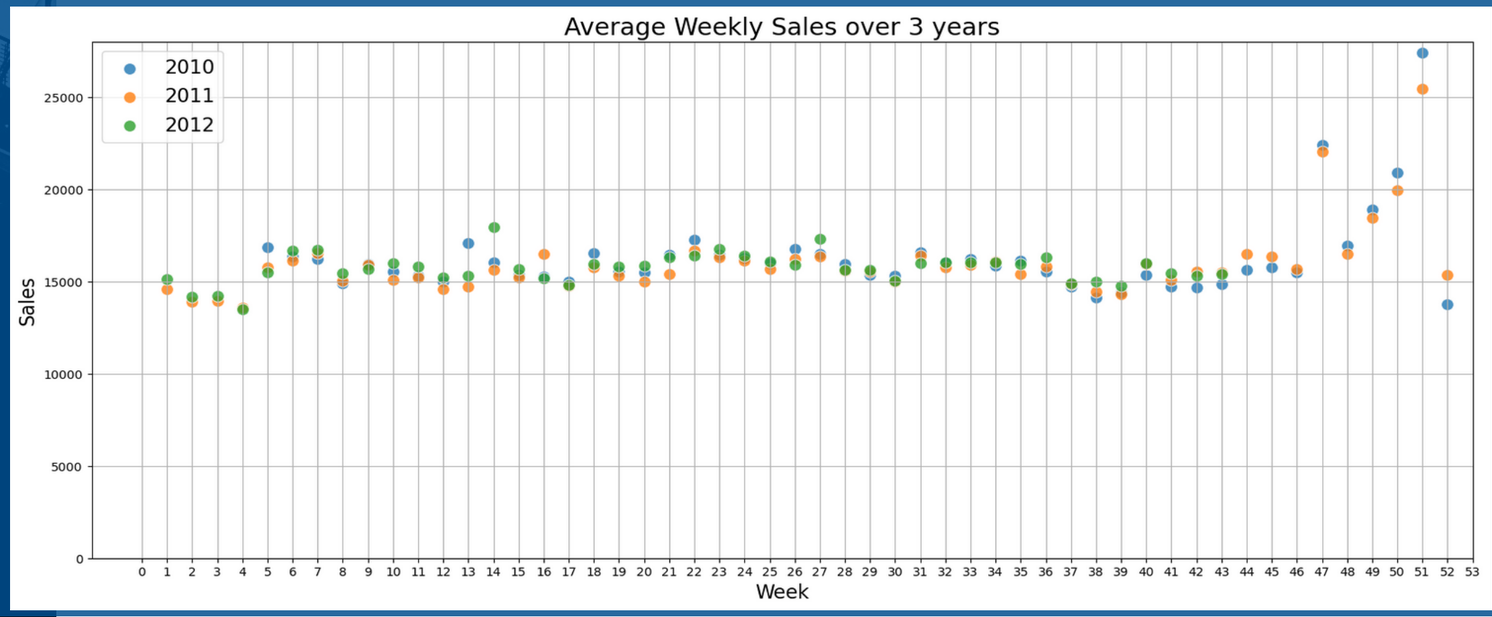


Data Exploration

1. Mean Weekly Sales by Month

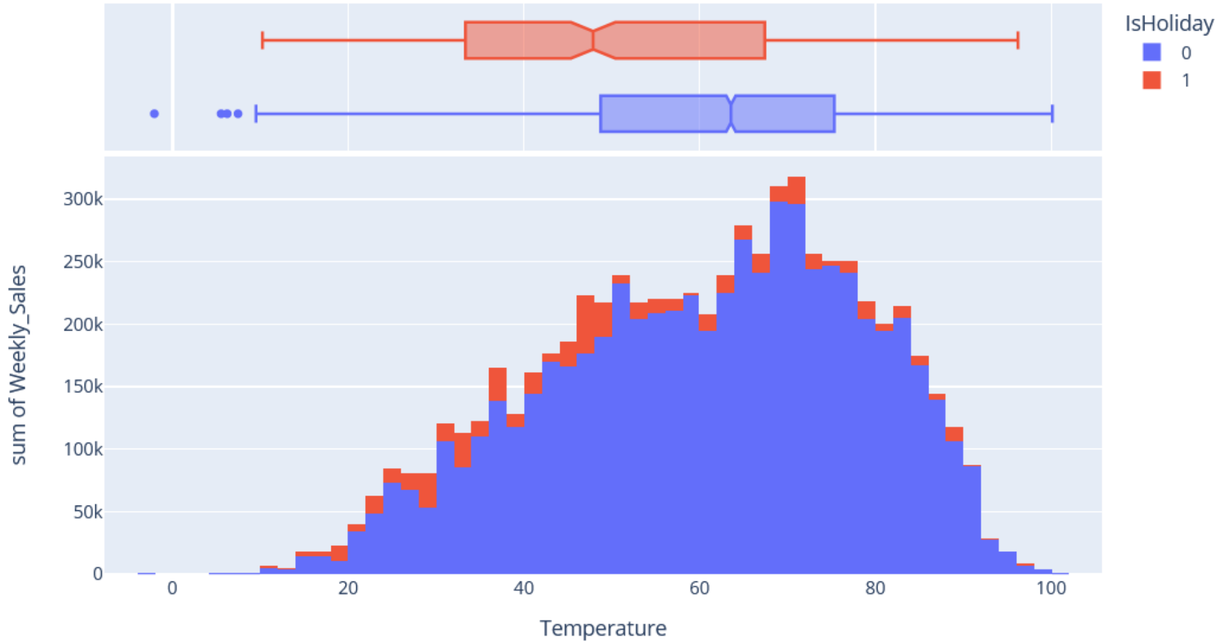


1. Combined Average Weekly Sales of all stores across every week for the years: 2010 to 12

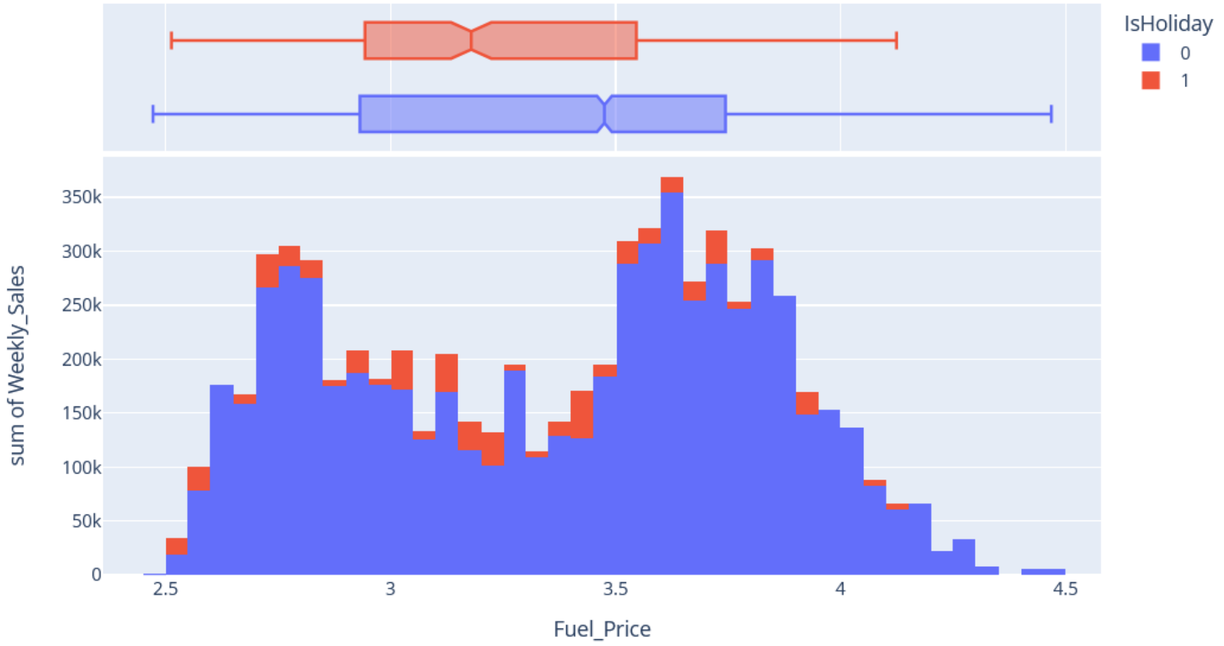


Data Exploration

1. Variation of Weekly Sales with Temperature conditional to Holiday week



2. Variation of Weekly Sales with Fuel Price keeping conditional to Holiday week



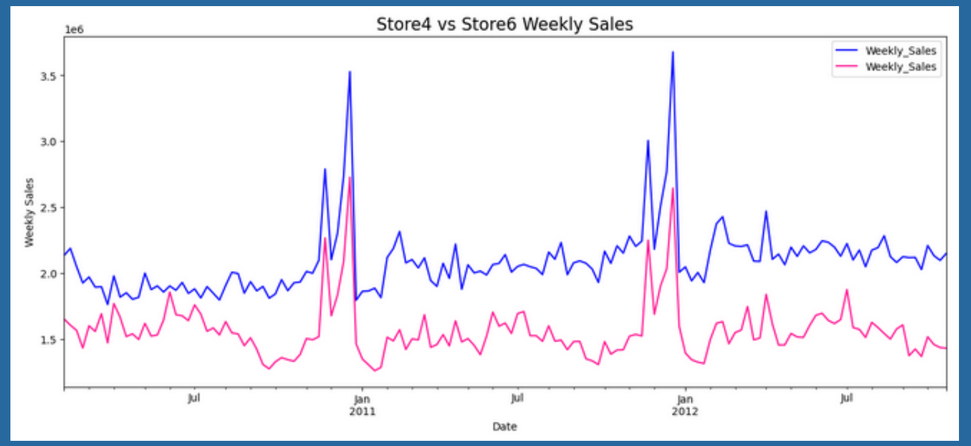
Methodology

1. Time Series
   - Time Series at Store-level
   - Time Series Decomposition

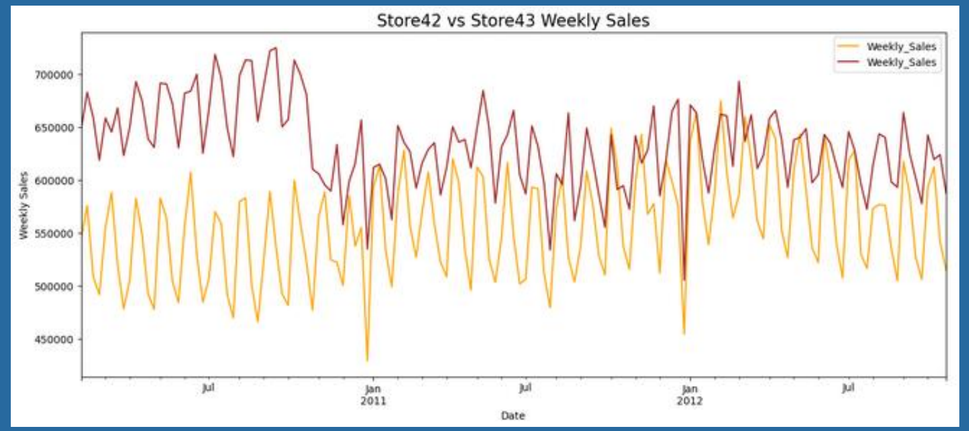
2. Modelling and Machine Learning
3. Improving Model
4. Final Model Selection

1. Time Series

- Weekly Sales of Store 4 is following the same trend as store 6.



- Store 42 is following the same trend as store 43.



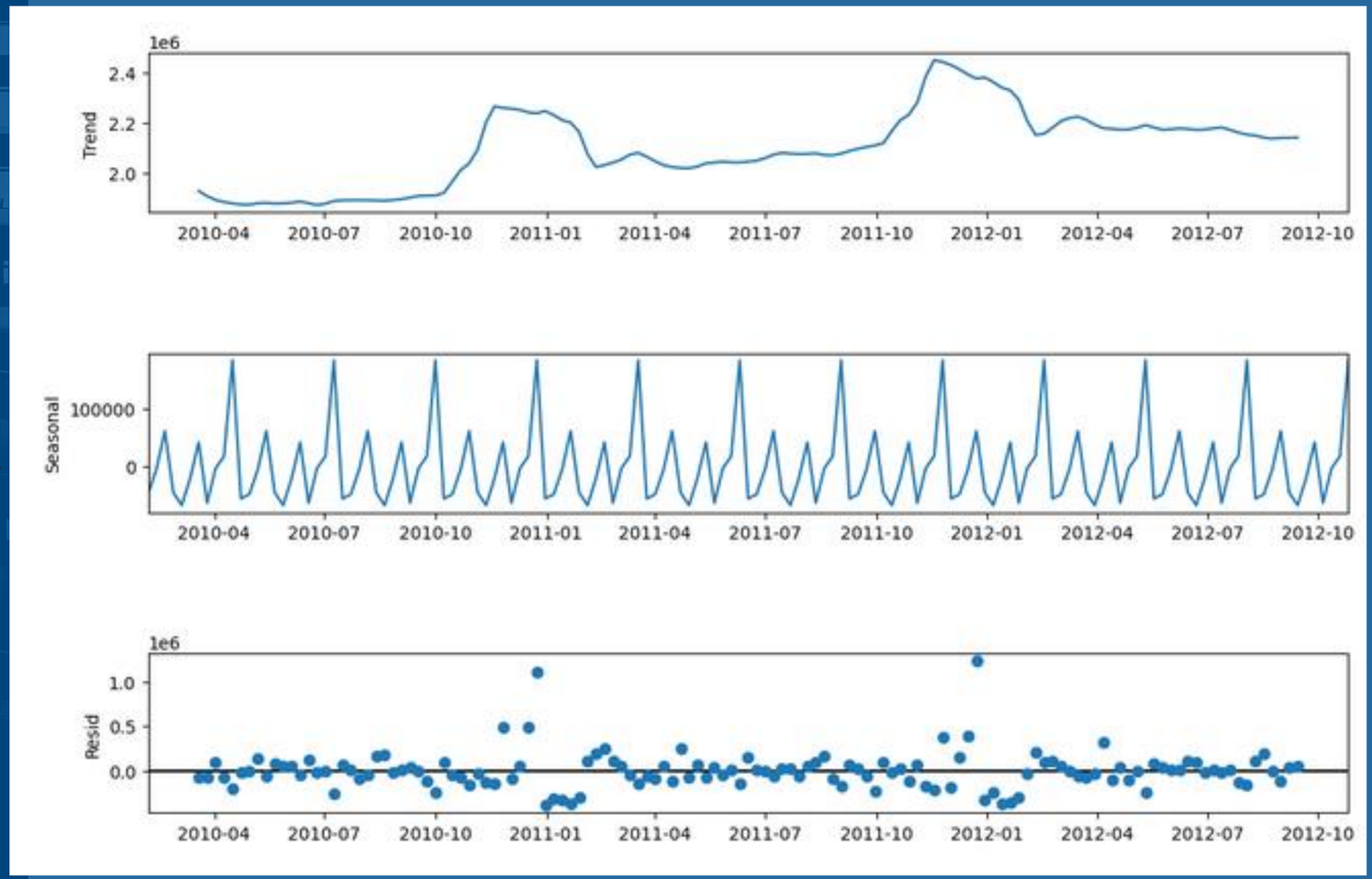
Methodology

1. **Time Series**
   - Time Series at Store-level
   - Time Series Decomposition

2. Modelling and Machine Learning
3. Improving Model
4. Final Model Selection

Time Series Decomposition (A & M)

- (Additive) Decomposition of Weekly Sales for Store 4 gives these components:



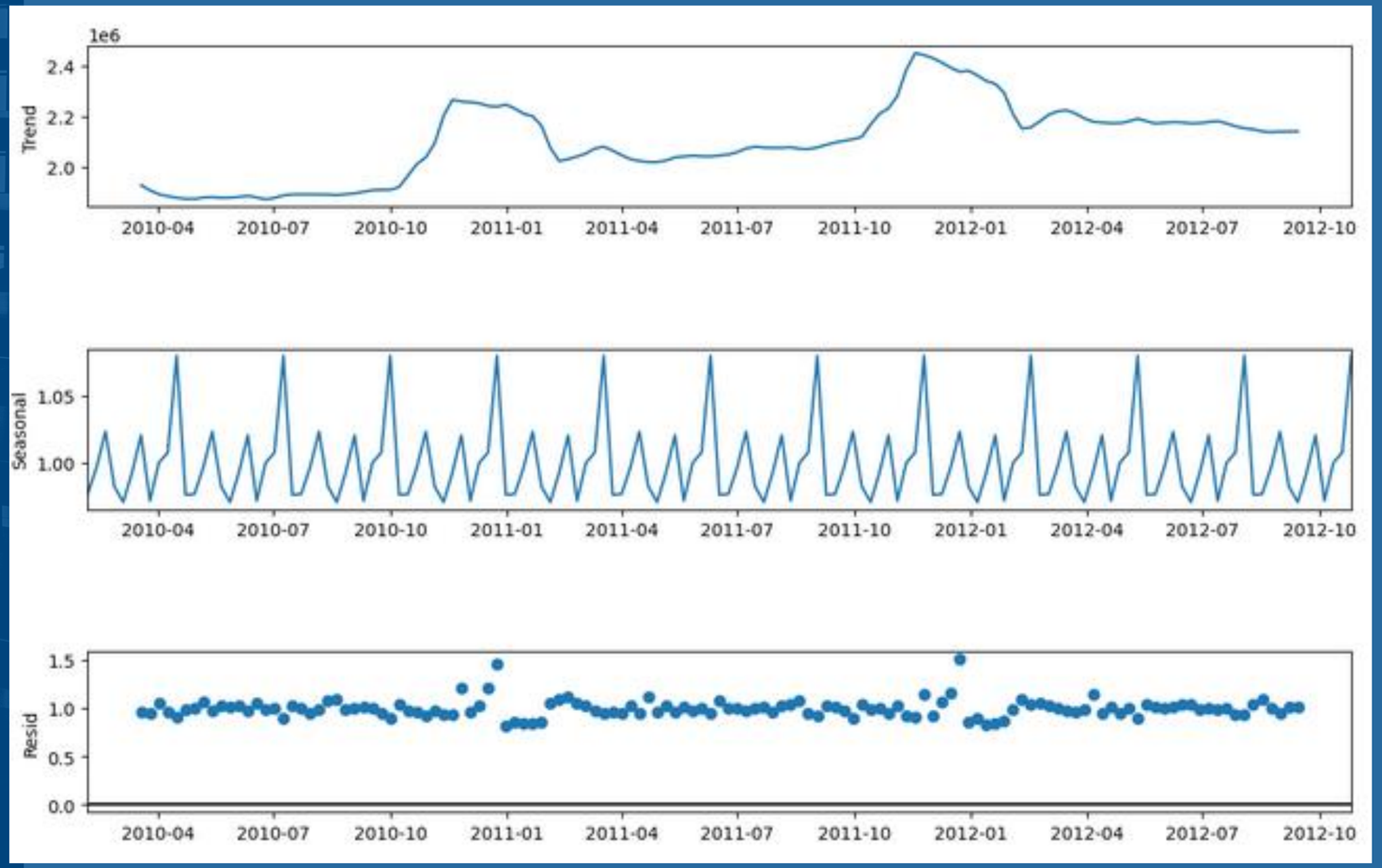
Methodology

1. **Time Series**
   - Time Series at Store-level
   - Time Series Decomposition

2. Modelling and Machine Learning
3. Improving Model
4. Final Model Selection

Time Series Decomposition (A & M)

- (Multiplicative) Decomposition of Weekly Sales for Store 4 gives these components:



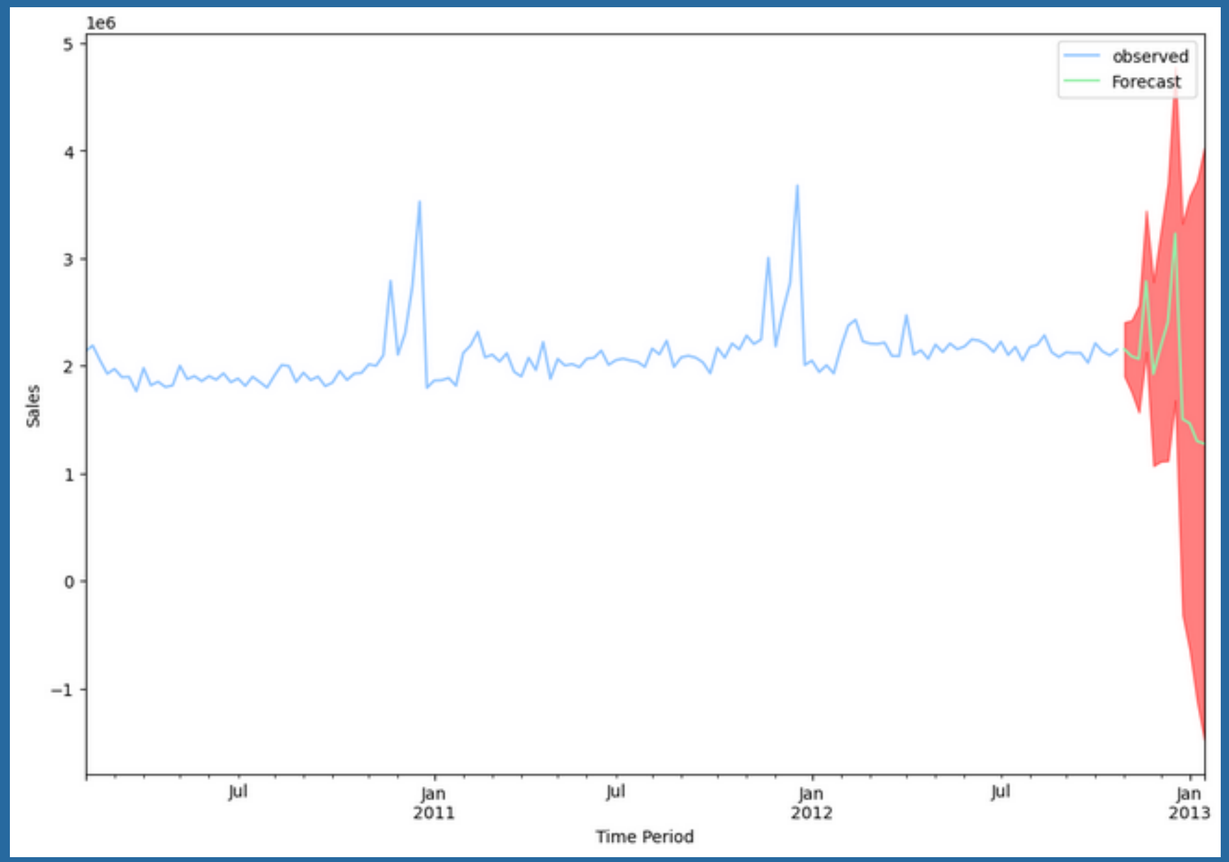
Methodology

1. **Time Series**
   - Time Series at Store-level
   - Time Series Decomposition

2. Modelling and Machine Learning
3. Improving Model
4. Final Model Selection

Time Series

- Time Series Model's prediction for the next 13 weeks is as shown.
- But the Error Interval is really large, and keeps increasing for future weeks.
- Hence TIME SERIES MODEL IS NOT CHOSEN



Methodology

1. Time Series
2. Modelling and Machine Learning
   - Encoding Categorical data
   - Feature Engineering
3. Improving Model
4. Final Model Selection
5. Improving Model

Encoding Categorical data:

- Creating dummy variables for all categorical variables 'Type', 'Store' ,'Month' ,'Year' ,'Day' , 'Week'.
- Resulting dataset has 71 columns

Imputation for missing values:

- Markdowns 1 to 5 have missing values ranging from 65% to 50%
- No data available from Feb 2010 to Nov 2011, and not available for all stores.

Methodology

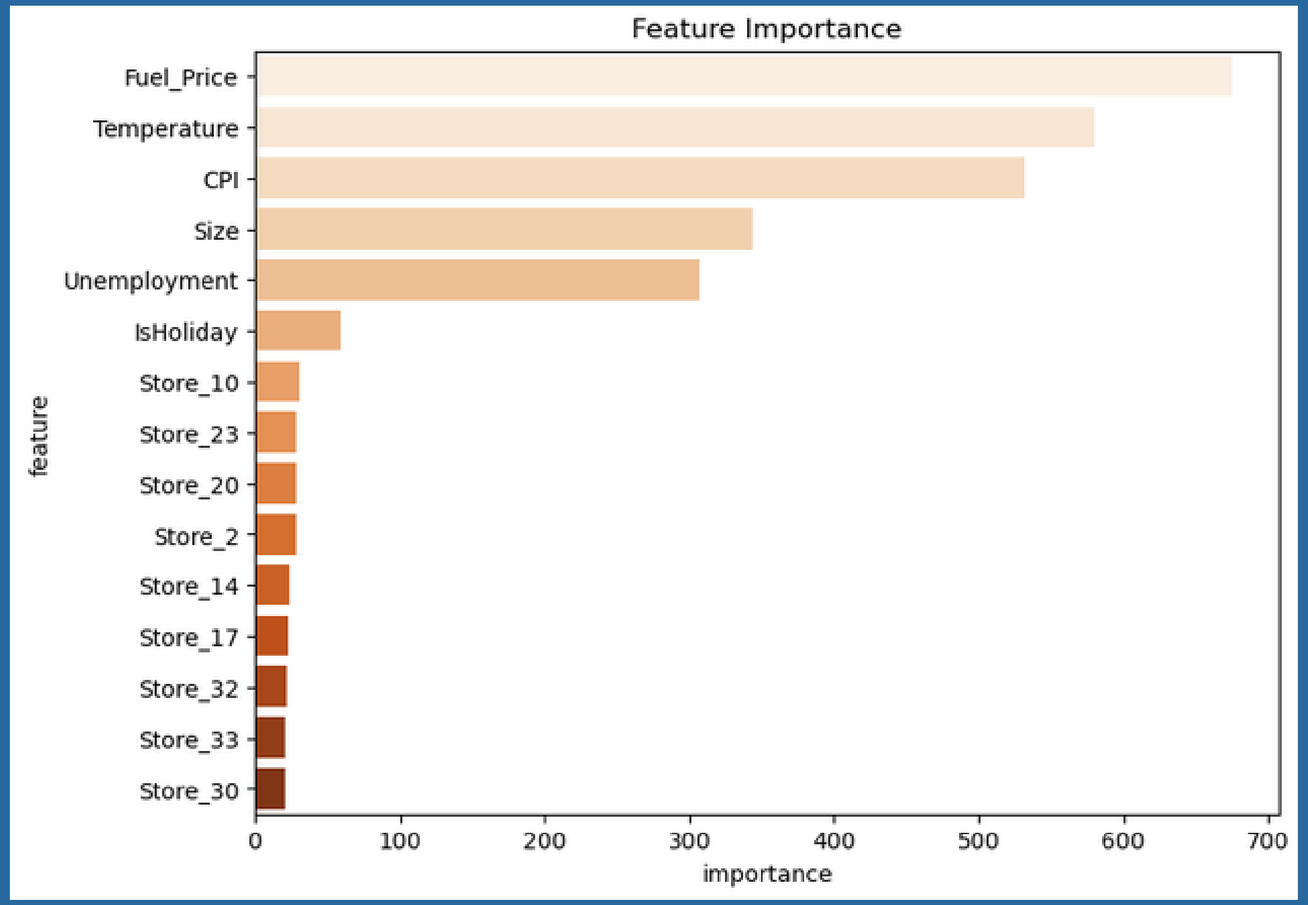
1. Time Series

2. Modelling and Machine Learning

3. Improving Model
   - Light GBM Regressor
   - LASSO Variable Selection
4. Final Model Selection

Light Gradient Boosting Method:

- LGBM Regressor allows to choose the best features to include in the model.
- It chooses leaf with max. delta to grow.



Methodology

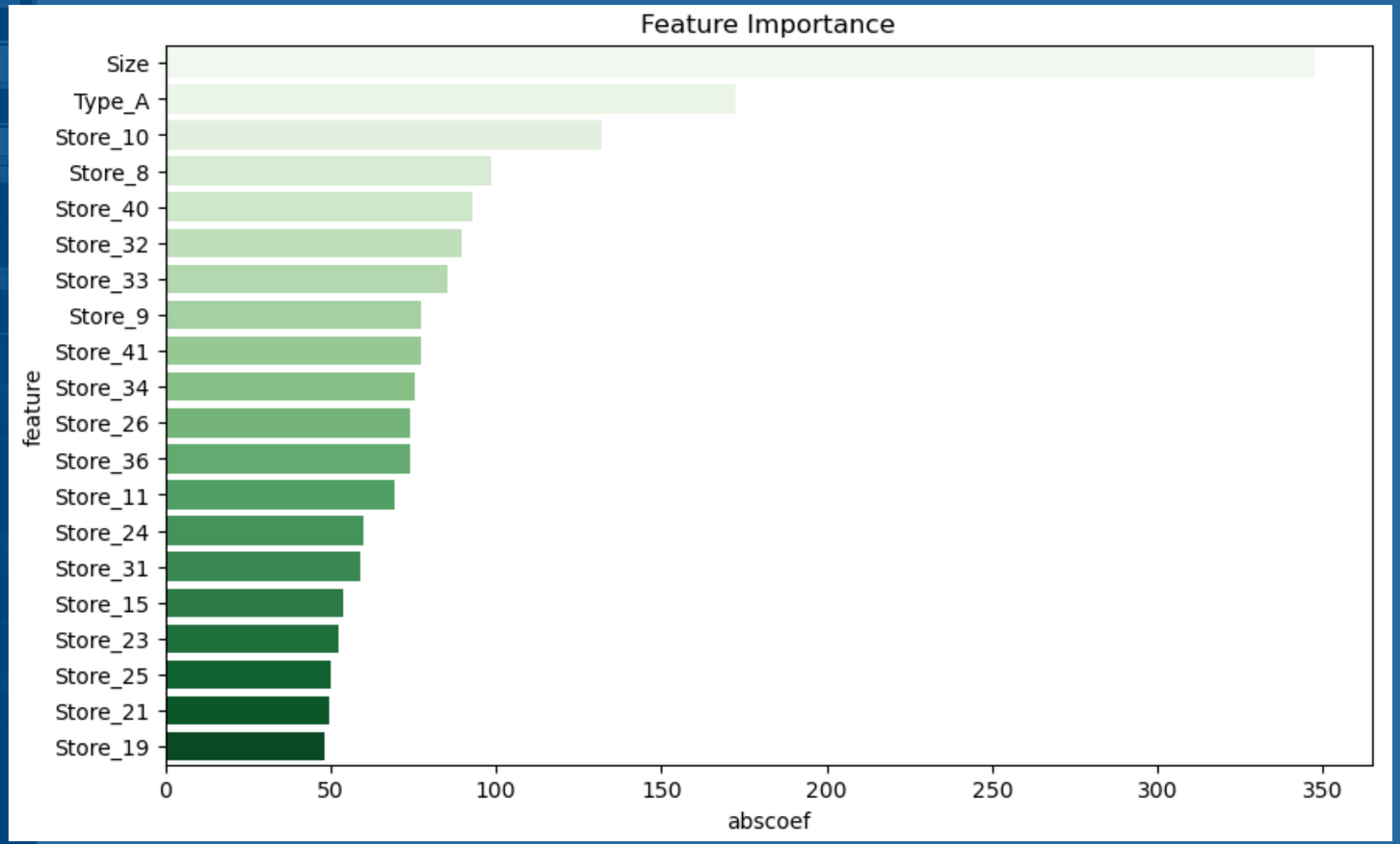
1. Time Series

2. Modelling and Machine Learning

3. Improving Model
   - Light GBM Regressor
   - LASSO Variable Selection
4. Final Model Selection

LASSO Variable Selection:

- LASSO does both, variable selections and regularization to enhance Prediction accuracy.
- Cross validated score of 0.93 makes it a good choice



Methodology

1. Time Series

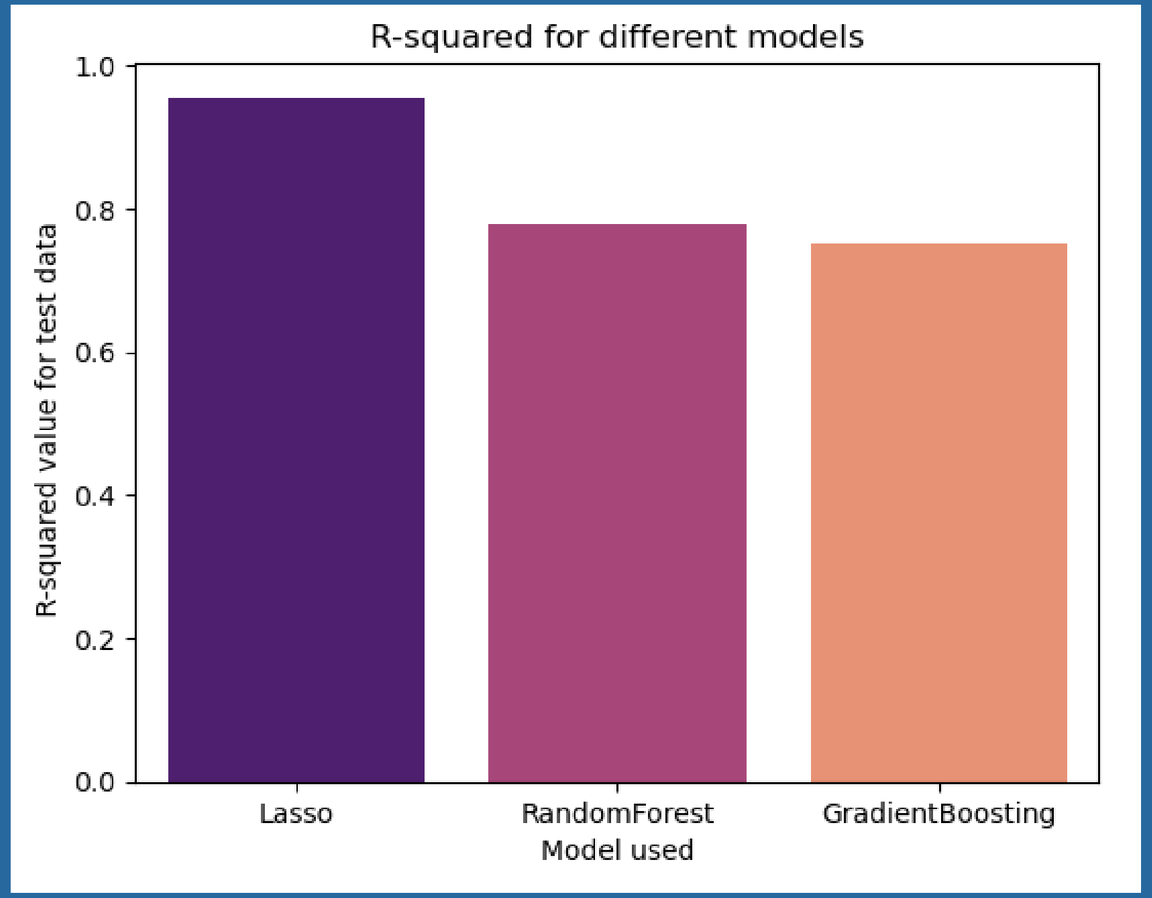
2. Modelling and Machine Learning

3. Improving Model

4. Final Model Selection
   - Random Forest Regression
   - Gradient Boosting Regression
   - LASSO Cross Validation

Final Model Selection:

- R-squared among the 3 models comes out to be highest for LASSO.



Methodology

1. Time Series

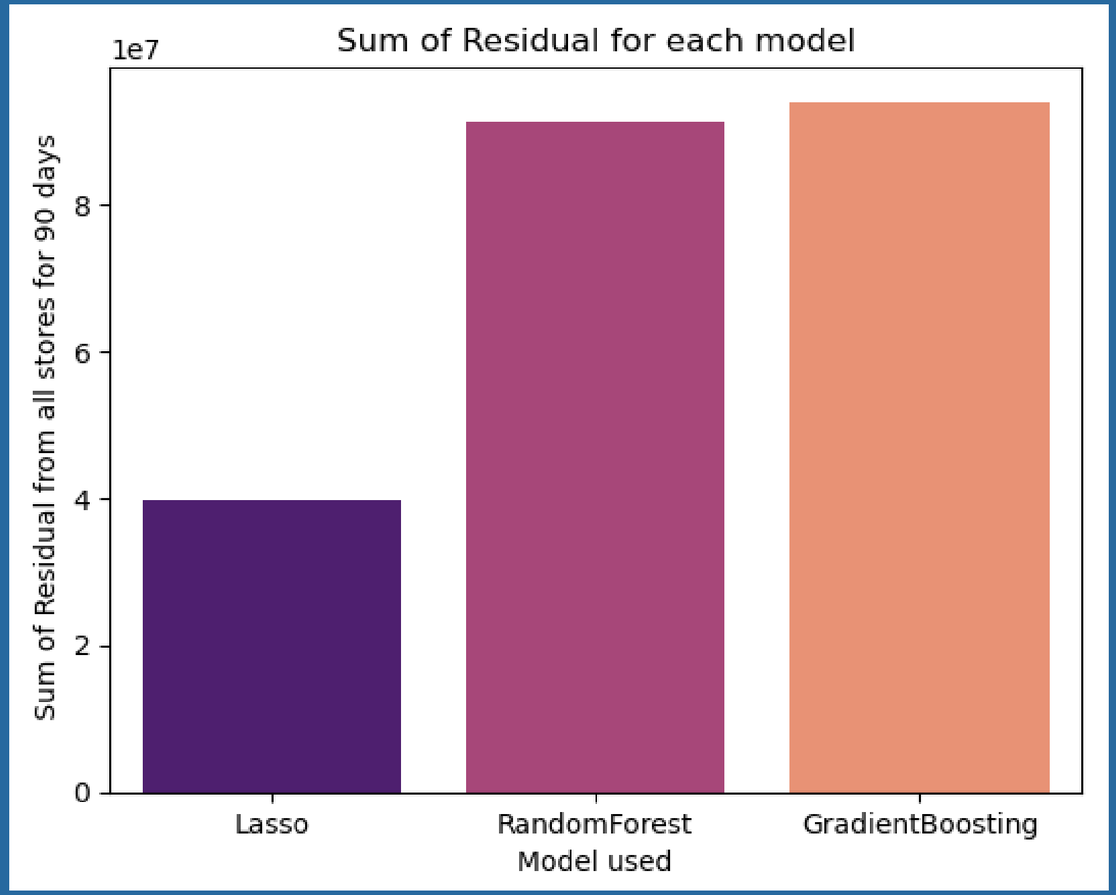
2. Modelling and Machine Learning

3. Improving Model

4. Final Model Selection
   - Random Forest Regression
   - Gradient Boosting Regression
   - LASSO Cross Validation

Final Model Selection:

- Sum of Residues among the 3 models comes out to be least for LASSO.



Discussion : Limitations

- **Actual Testing data is unseen (kaggle)**
  - ML model used for the project, varies from the actual project on Kaggle.
- **Limited Covariates and lagged variables**
  - The availability and selection subject to certain limitations.
  - Important Time Lags were included but additional variables that enhance accuracy
- **Time Constraints**
  - It is not possible to thoroughly explore ALL aspects and alternative modelling approaches, and there may be potential areas for improvement that were not fully explored
- **Computational Constraints:**
  - 536k records at departmental level, without aggregation of departments to single store it takes really long times to run even simple ML models
- **Missing data from Markdown events**

Discussion : Conclusions

- Data has components of Time Series, yet Time Series modelling was not the best to predict the sales.
- Sales start to increase in Nov, and reach maximum in Dec and fall in Jan.
- Size of store is a huge predictor of sales.
- Holiday week tends to see more sales than the usual weeks, yet not all Holidays means increase in Sales, and therefore it is not a good predictor of Sales.
- Sales tend to follow a general pattern that is true throughout all stores, the magnitude keeps changing but the trend is followed.

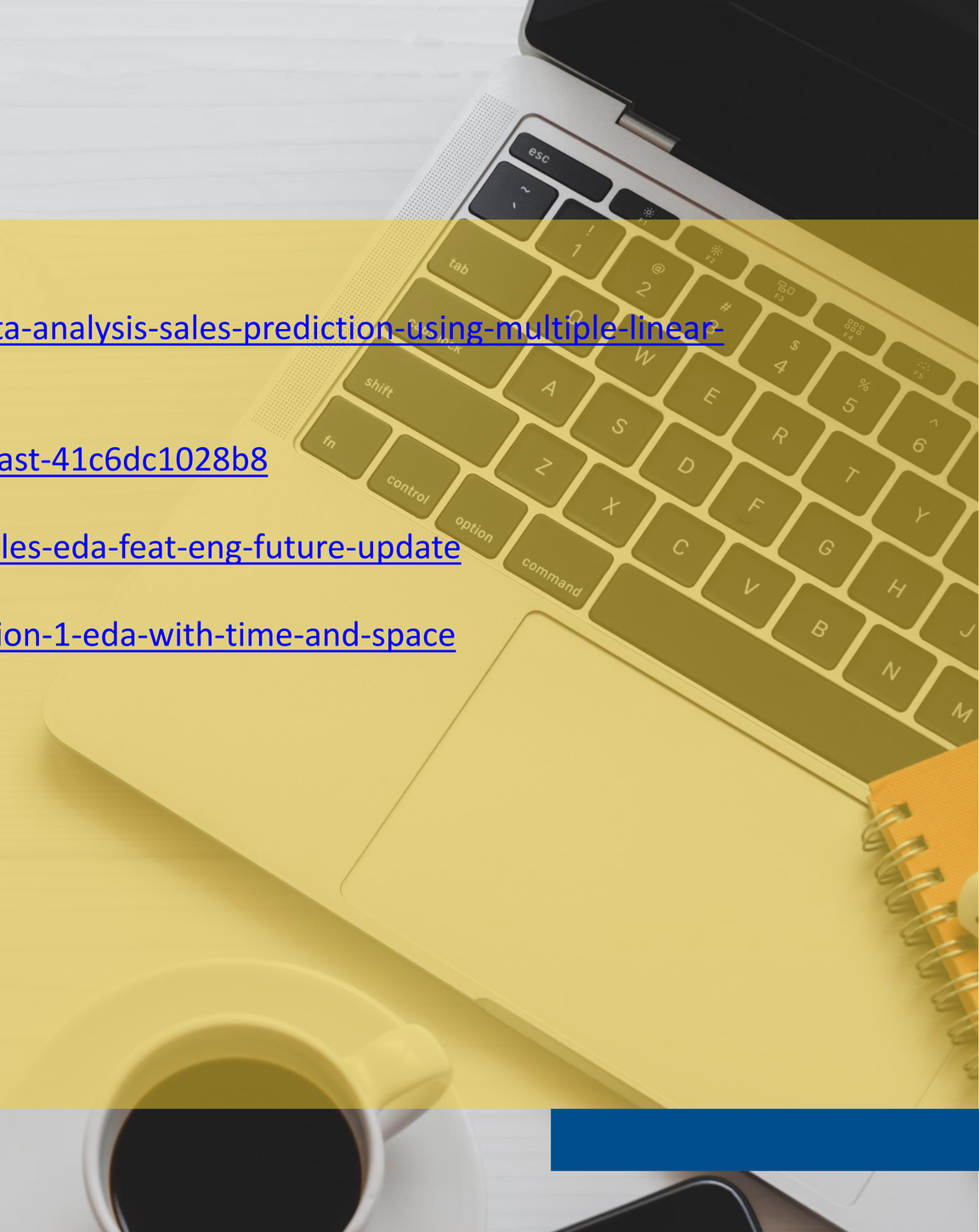
Discussion : Possible Future Work

- **Zoom-In at Departmental level**
  - More granular analysis at departmental level, reduce error
- **Simplified Modeling Approach**
  - Combining Time Series and Regression provides a practical solution, but this approach might have limitations
- **Transportation Routing**
  - Integrate demand forecasts with transportation optimization
  - Determine the most efficient routes for delivering products to Walmart stores

References:

- https://medium.datadriveninvestor.com/walmart-sales-data-analysis-sales-prediction-using-multiple-linear-regression-in-r-programming-adb14afd56fb
- https://medium.com/analytics-vidhya/walmart-sales-forecast-41c6dc1028b8
- https://www.kaggle.com/code/maxdiazbattan/wallmart-sales-eda-feat-eng-future-update
- https://www.kaggle.com/code/yepp2411/walmart-prediction-1-eda-with-time-and-space
- https://yanyudm.github.io/Data-Mining-R/

- Theme : SlidesCarnival



Thank you!

Appendix

- **HTML – Output**

Walmart_Sales_Forecast_Final.html

- **Python Notebook**

Walmart_Sales_Forecast_Final.ipynb

- **Original Data Zipped**

walmart-recruiting-store-sales-forecasting.zip