Summary of Work Done

1. **Data Preparation**:
    - Loaded and merged datasets.
    - Dropped duplicates and renamed columns for consistency.
    - Handled missing values and zeroed out markdown columns where necessary.

```
In [14]: df_features.head()
```

Out[14]:

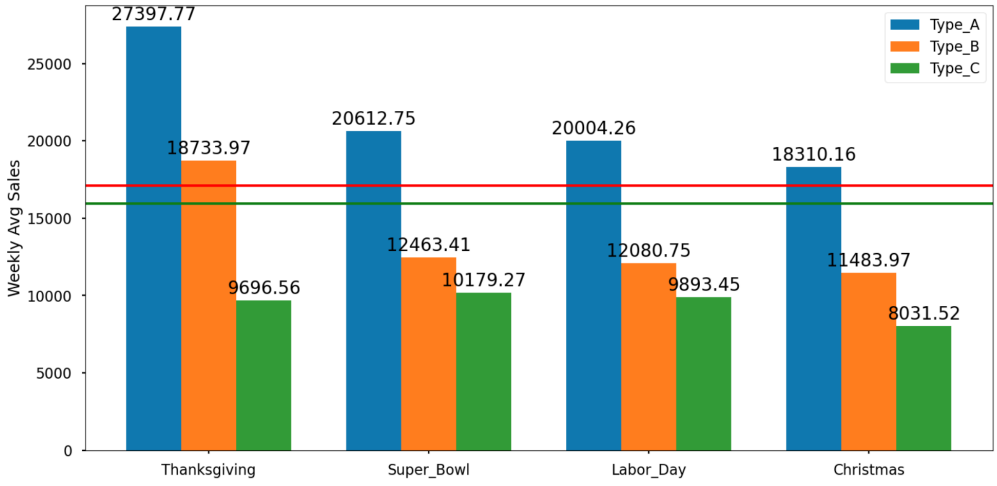
	Store	Date	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Unemployment	IsHoliday
0	1	2010-02-05	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	8.106	False
1	1	2010-02-12	38.51	2.548	NaN	NaN	NaN	NaN	NaN	211.242170	8.106	True
2	1	2010-02-19	39.93	2.514	NaN	NaN	NaN	NaN	NaN	211.289143	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

```
In [15]: # merging 3 different sets
df = df_train.merge(df_features, on=['Store', 'Date'], how='inner').merge(df_store, on=['Store'], how='inner')
df.head(5)
```

Out[15]:

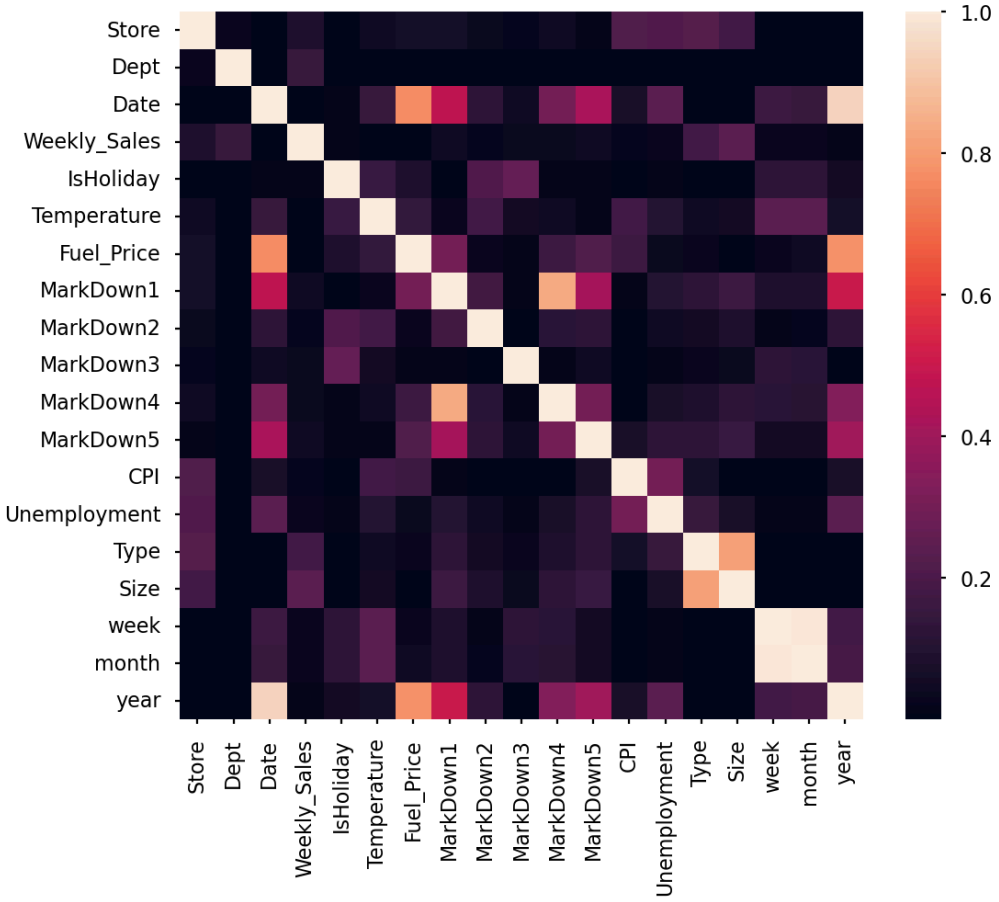
	Store	Dept	Date	Weekly_Sales	IsHoliday_x	Temperature	Fuel_Price	MarkDown1	MarkDown2	MarkDown3	MarkDown4	MarkDown5	CPI	Une
0	1	1	2010-02-05	24924.50	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	
1	1	2	2010-02-05	50605.27	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	
2	1	3	2010-02-05	13740.12	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	
3	1	4	2010-02-05	39954.04	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	
4	1	5	2010-02-05	32229.38	False	42.31	2.572	NaN	NaN	NaN	NaN	NaN	211.096358	

2. **Exploratory Data Analysis (EDA)**:
    - Analysed average weekly sales across stores and departments.
    - Investigated the impact of holidays on sales and examined sales patterns for special events like holidays.
    - Explored the effect of store types and sizes on sales.
    - Conducted time-based analysis, including examining sales trends by week, month, and year.

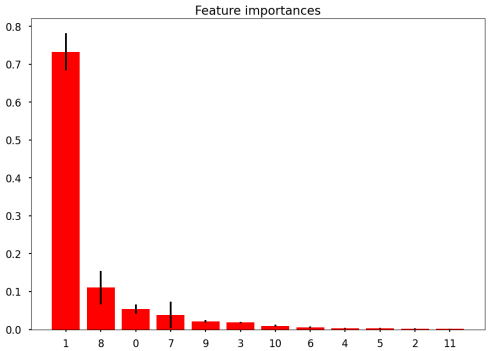


It is seen from the graph that, highest sale average is in between holidays. And, for all holidays Type A stores has highest sales.

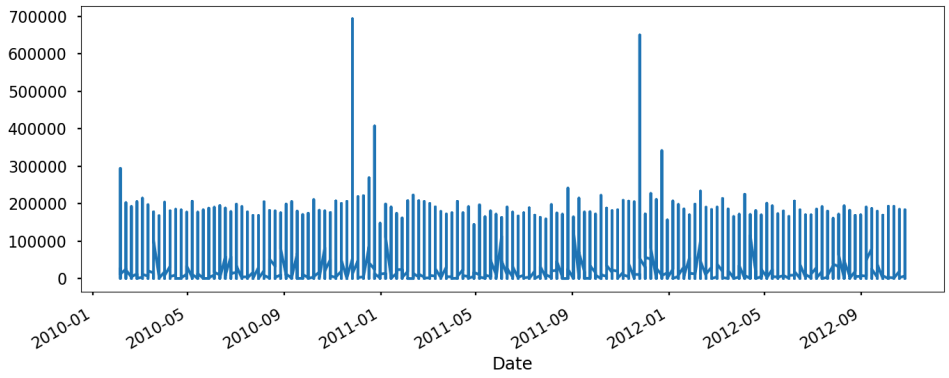
3. **Feature Engineering**:
    - Encoded categorical variables and handled boolean features.
    - Analyzed correlations between features and decided on dropping some due to high correlation or low importance.



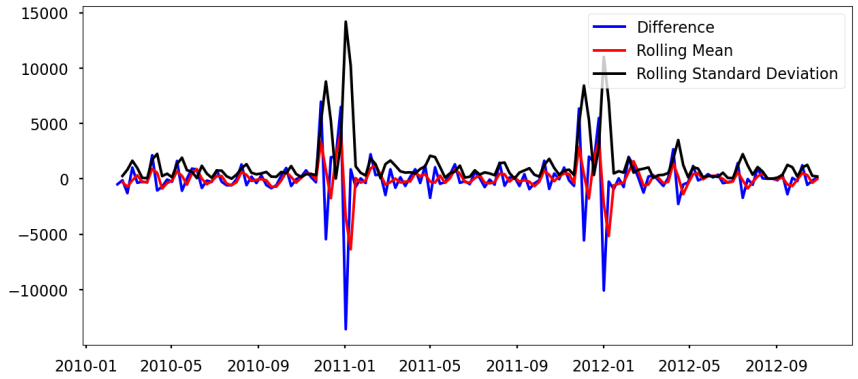
4. **Modeling**:
    - Built and tuned a RandomForestRegressor model, using WMAE (Weighted Mean Absolute Error) as a metric.
    - Performed feature importance analysis to refine the model.

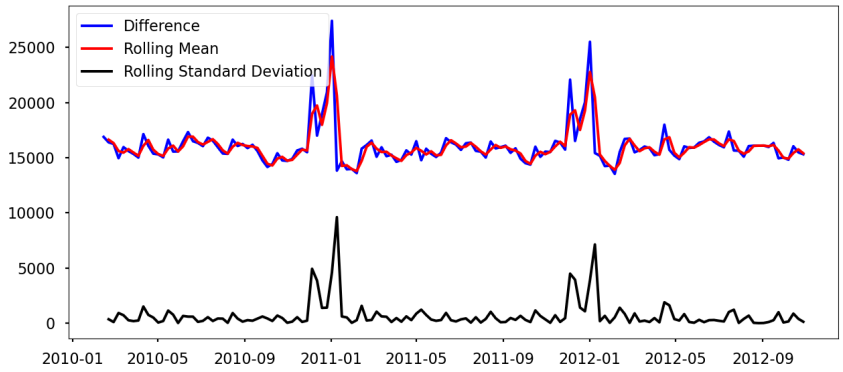


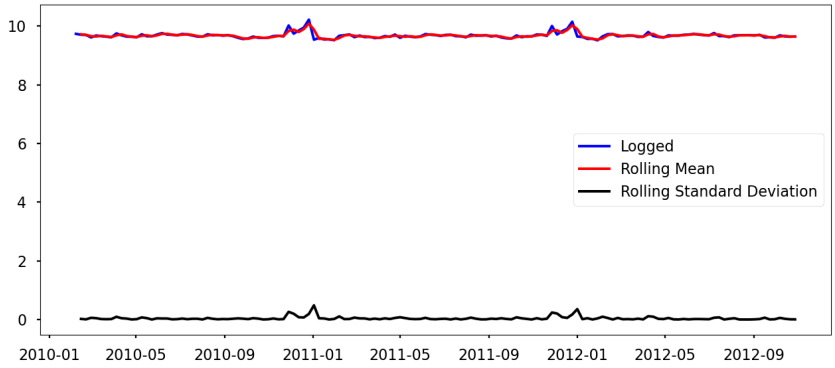
5. **Time Series Analysis**:
    - Converted data to time series format and resampled it to weekly and monthly frequencies.
    - Plotted and analyzed sales data to understand seasonal patterns.



    - Attempted to make the data more stationary using differencing and shifting techniques







Auto-ARIMA MODEL

I tried my data without any changes, then tried with shifting, taking log and difference version of data. Differenced data gave best results. So, I decided to take difference and use this data.