Time Series Sales Forecasting

Introduction to Data Management and Processing -DS 5110
Fall 2021
Course Project - Group 8

Presented By:

- Apoorva Surendra Malemath
- Ashwin Sateesh Kumar
- Barkha Saxena
- Basil Varghese
- Sravya Burugu

Northeastern
Khoury College of
Computer Sciences

Outline

- Introduction
- Understanding the Data
- Exploratory Data Analysis
- Methodology
    - ✓ Statistical Forecasting Methods - ARIMA and SARIMAX
    - ✓Machine Learning Models
- Results
- Conclusion and Future Scope

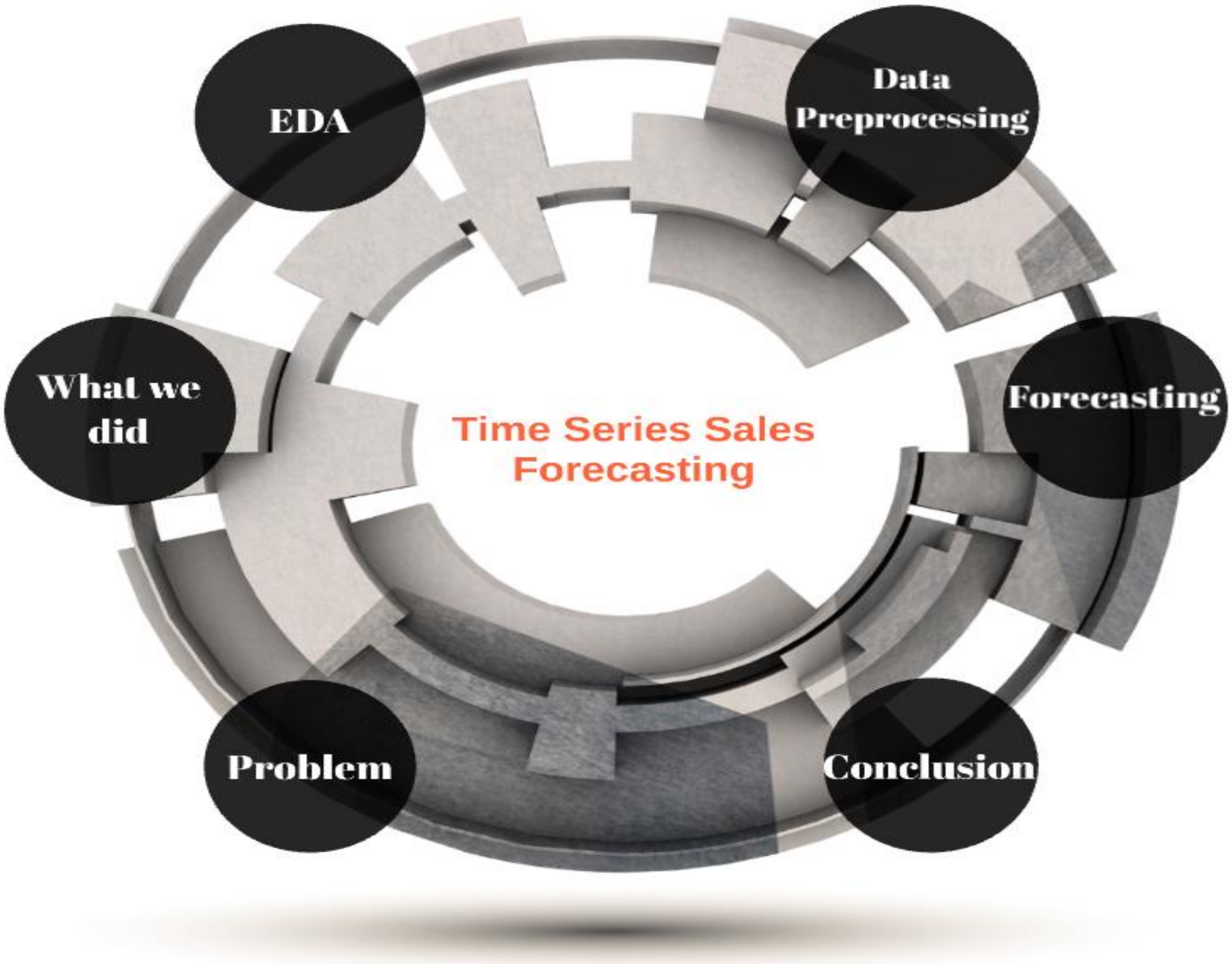
Introduction

- This project entails the sales forecasting for Corporacion Favorita, an Ecuador based grocery retailer.

- Forecasting allows businesses set reasonable and measurable goals based on current and historical data.

- Forecasting is helpful in inventory planning management, estimating revenue of an organization, demand forecasting resource allocation and supply chain management.

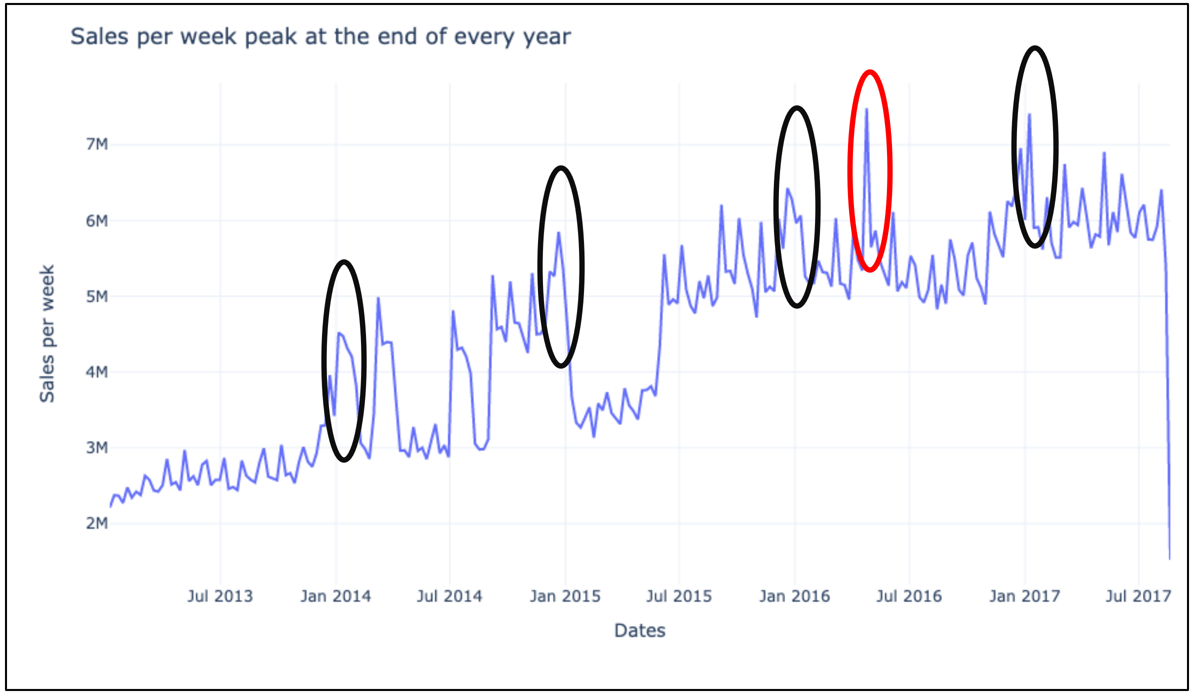
- The dataset consists of 4 years sales data, at a date-store-product level, along with information on promotions run on particular days.

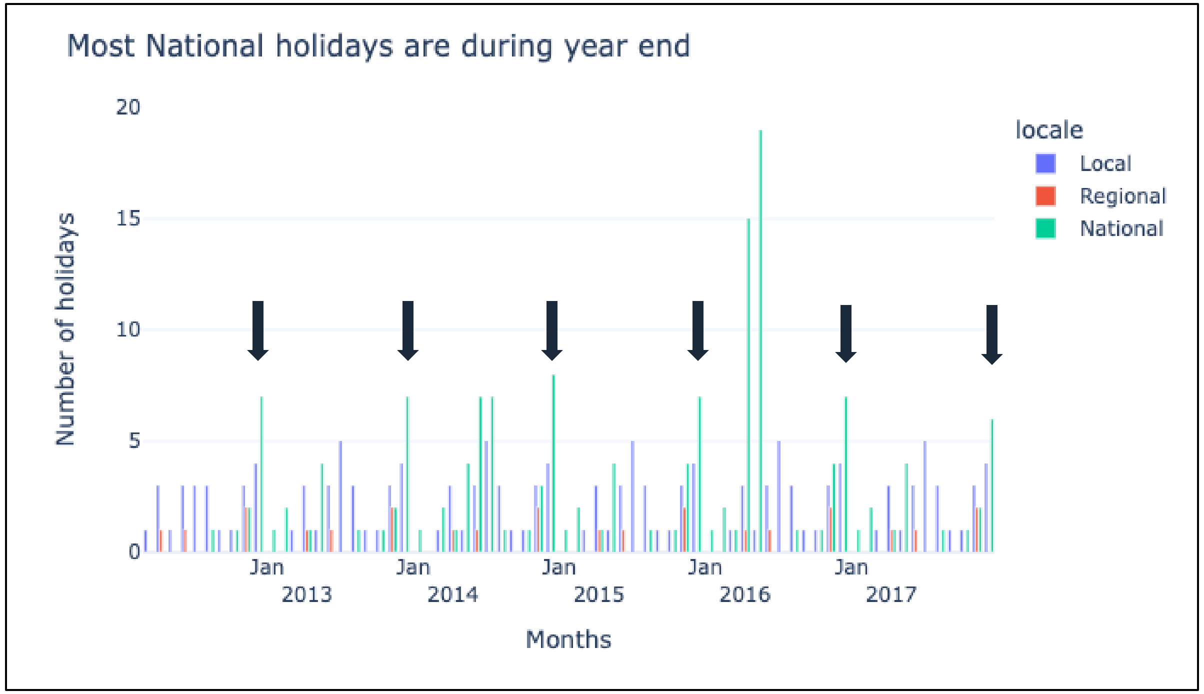

EDA
Data Preprocessing
What we did
Time Series Sales Forecasting
Forecasting
Problem
Conclusion

Understanding the Data

- Data consists of 4 years of sales information or Corporacion Favorita, an Ecuador based grocery retailer
- Data comprises of sales information for 33 families of products across 54 stores
- The stores are spread across 22 cities of 16 states in Ecuador.
- About 3 million tuples
- Additional information comprises of Promotion Information, National Holiday Details

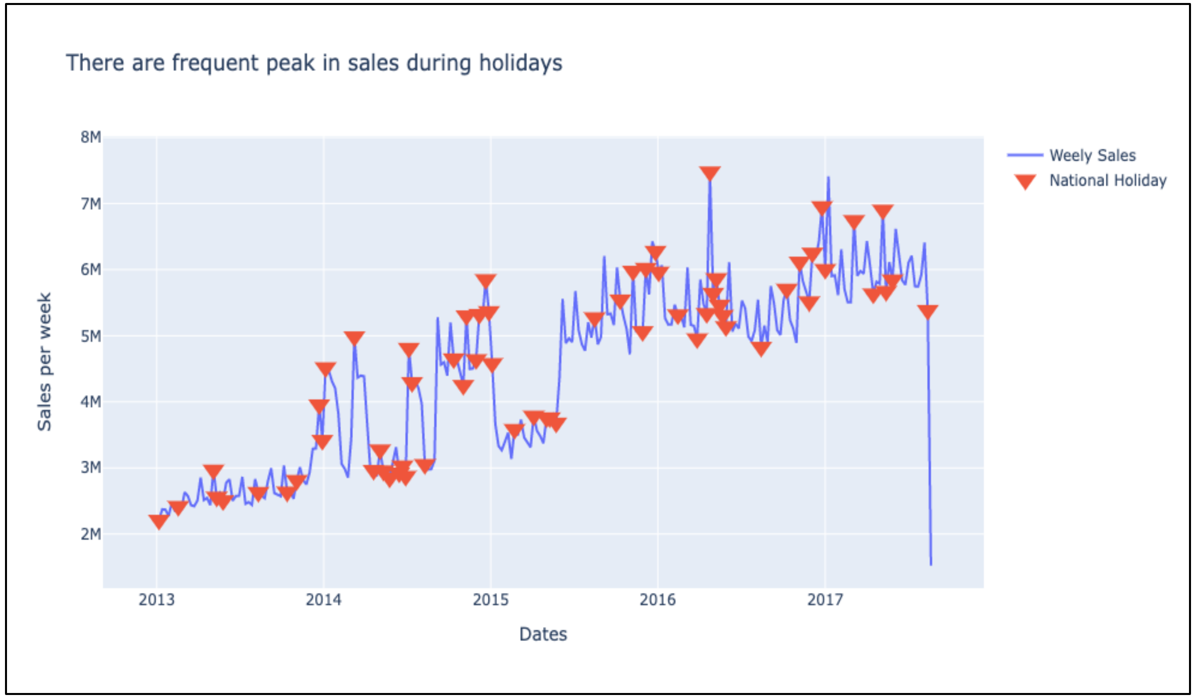
Exploratory Data Analysis

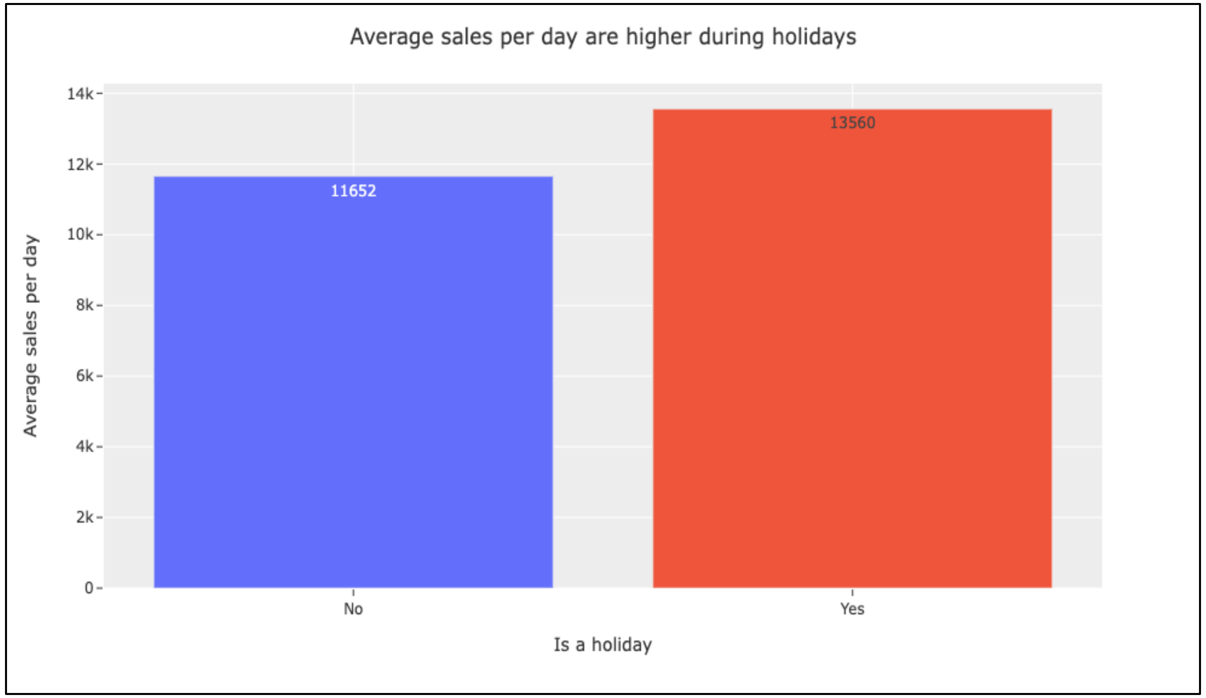




- There is a peak in sales at the end of every year, during December – January
- Most national holidays are during the end of the year
- There is an increasing trend of sales from 2013 – 2017
- An anomalous peak in sales is observed during Apr – May 2016
  This is attributed to a magnitude 7.8 earthquake that struck Ecuador on April 16, 2016

Exploratory Data Analysis



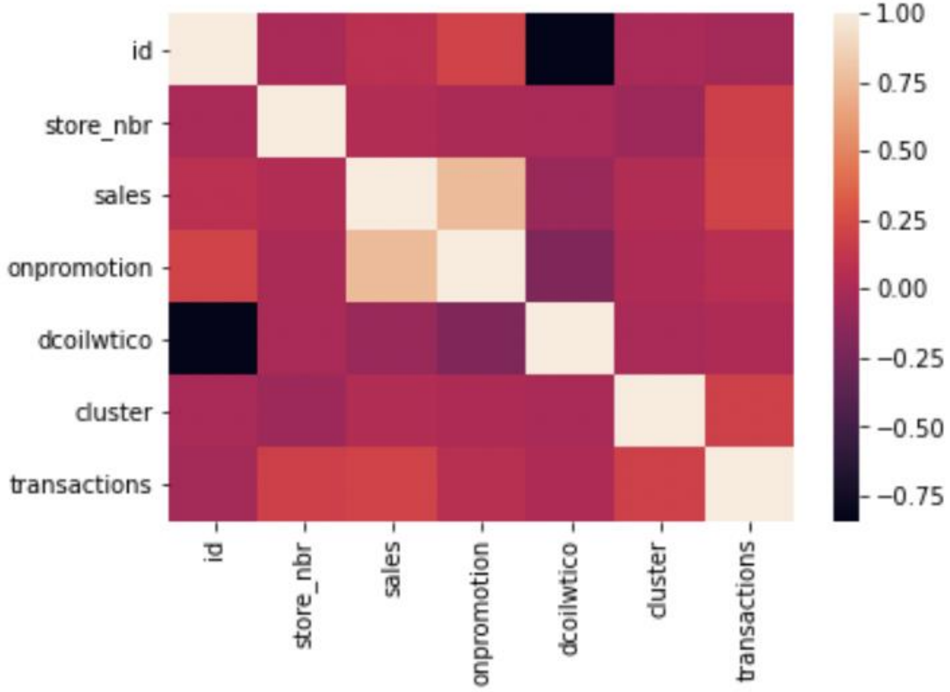


- There is a peak in sales corresponding to most national holidays
- Average sales per day during weeks with holidays are than weeks without holidays
- There are certain national holidays during which sales are low
- More promotions can be given during those holidays

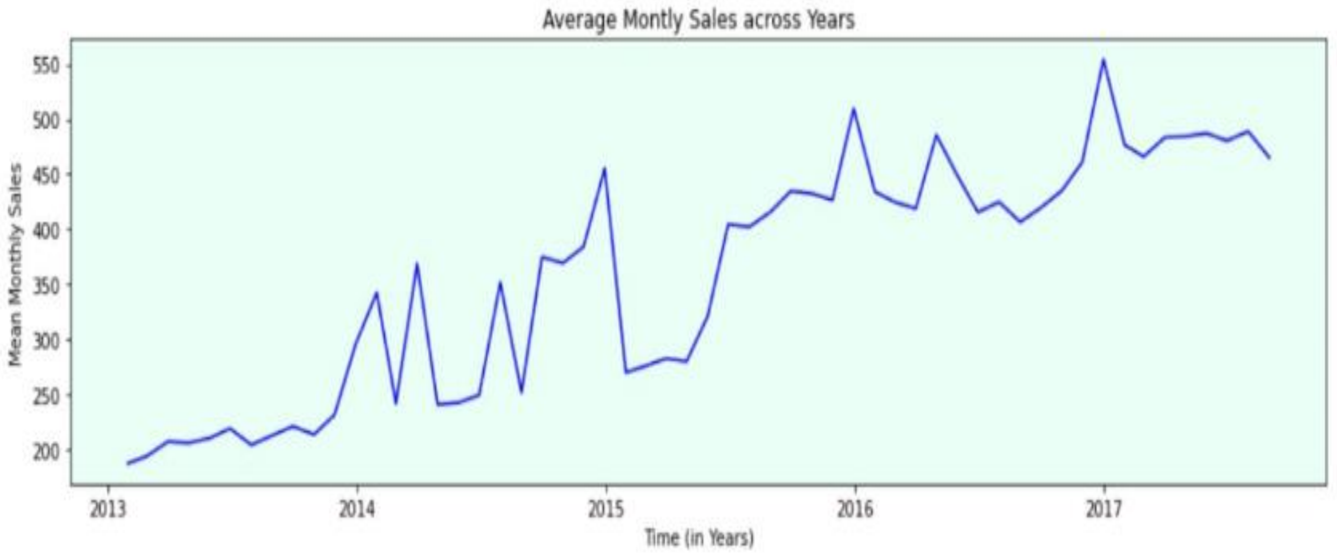
LVX
VERITAS
VIRTVS

Exploratory Data Analysis

Correlation Matrix

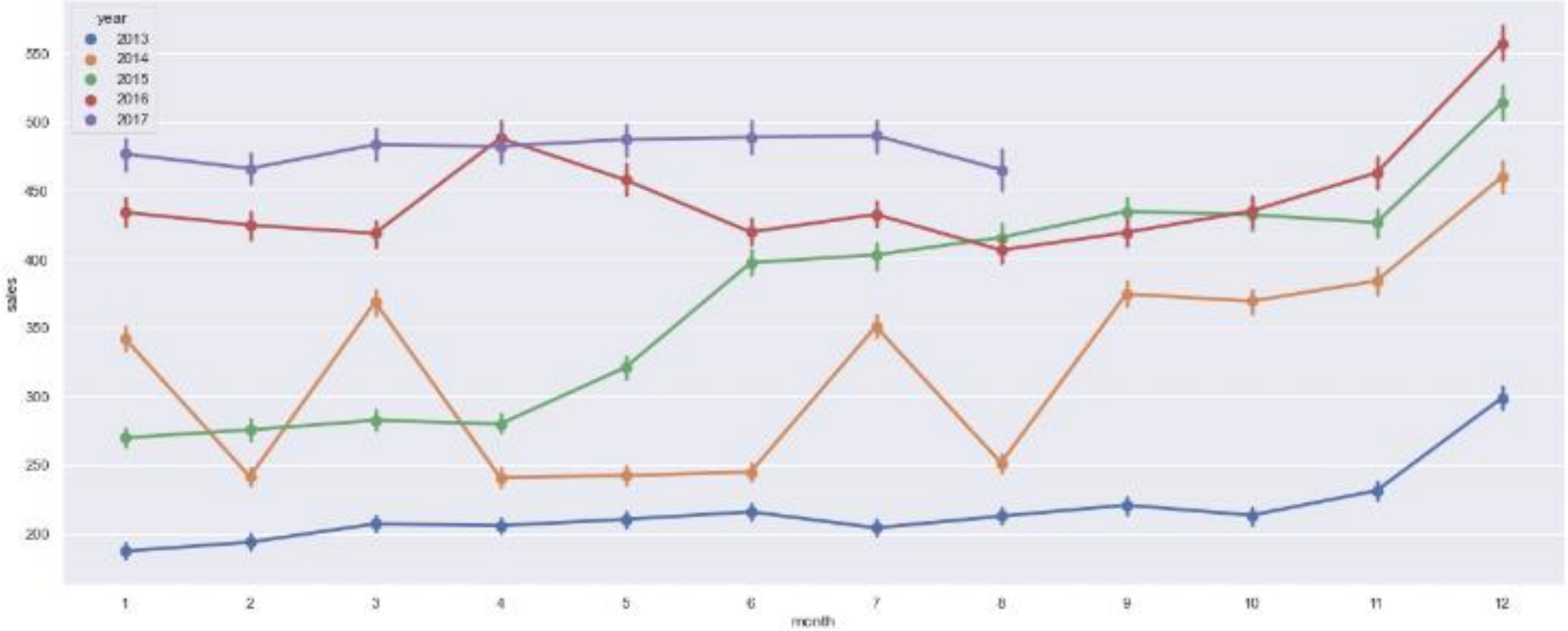


Average monthly sales across years



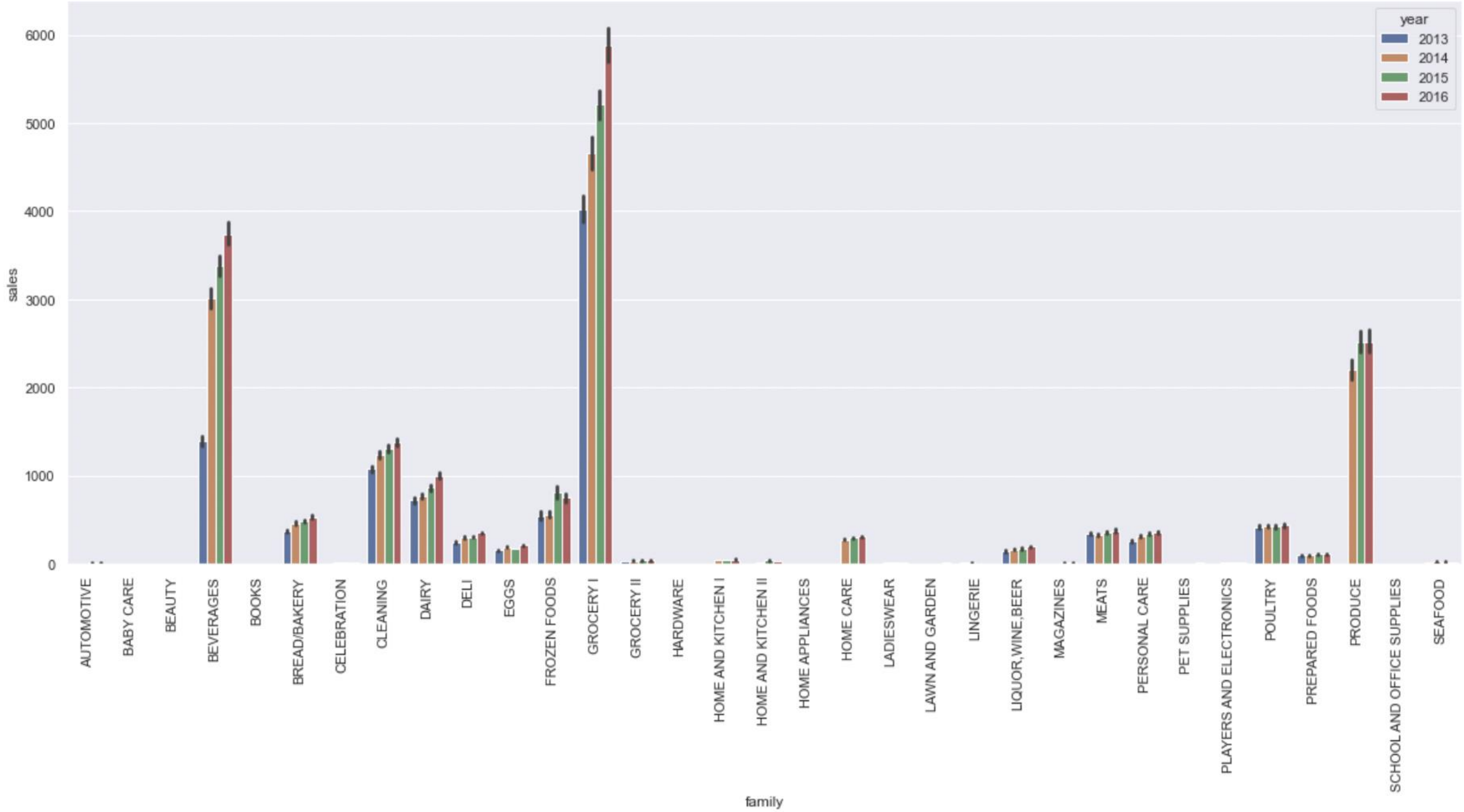
Exploratory Data Analysis

Total sales per month for each year



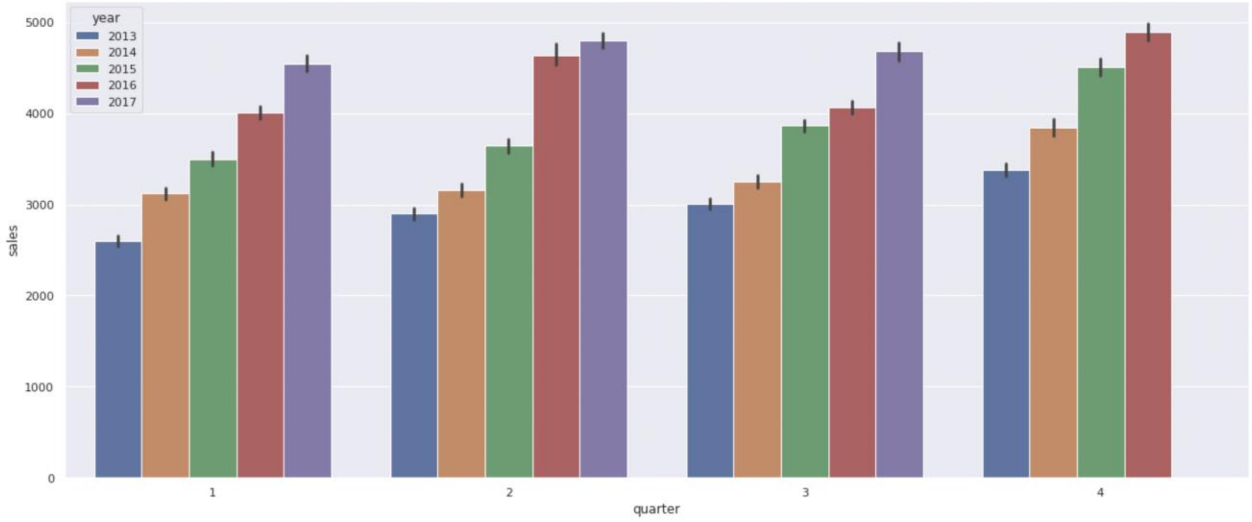
Exploratory Data Analysis

Total sales of each family of products for each year

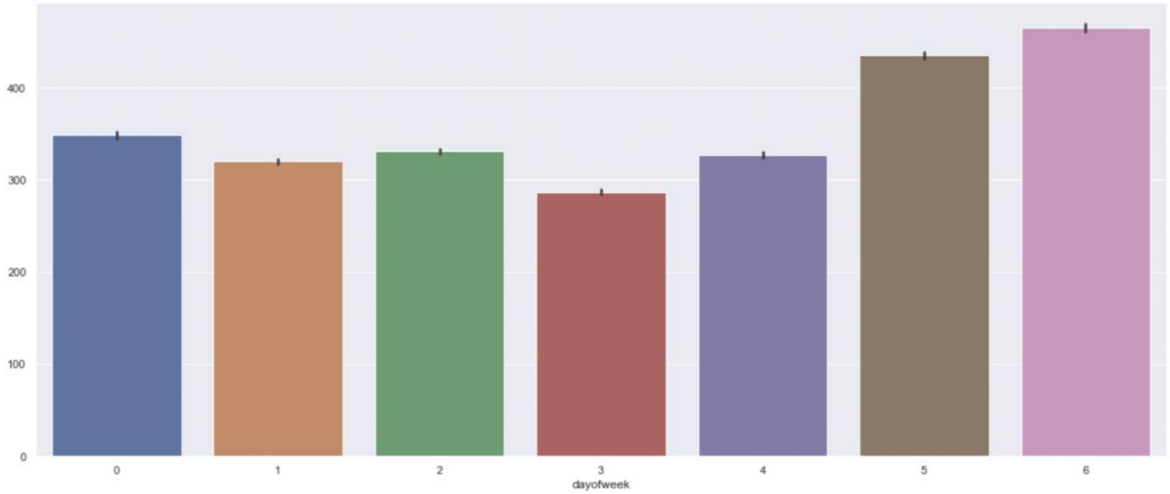


Exploratory Data Analysis

Quarterly sales Analysis

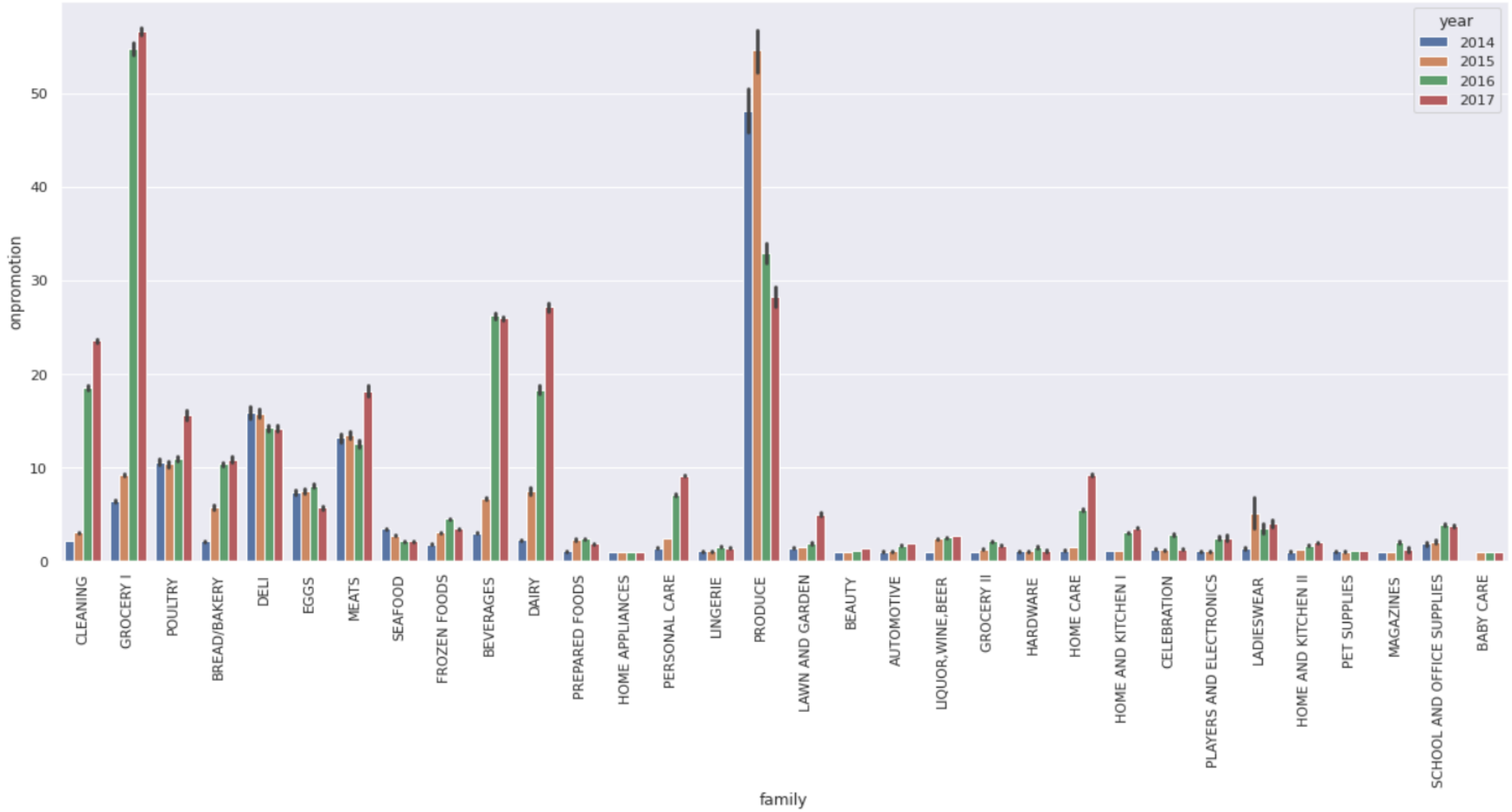


Weekly sales Analysis



Exploratory Data Analysis

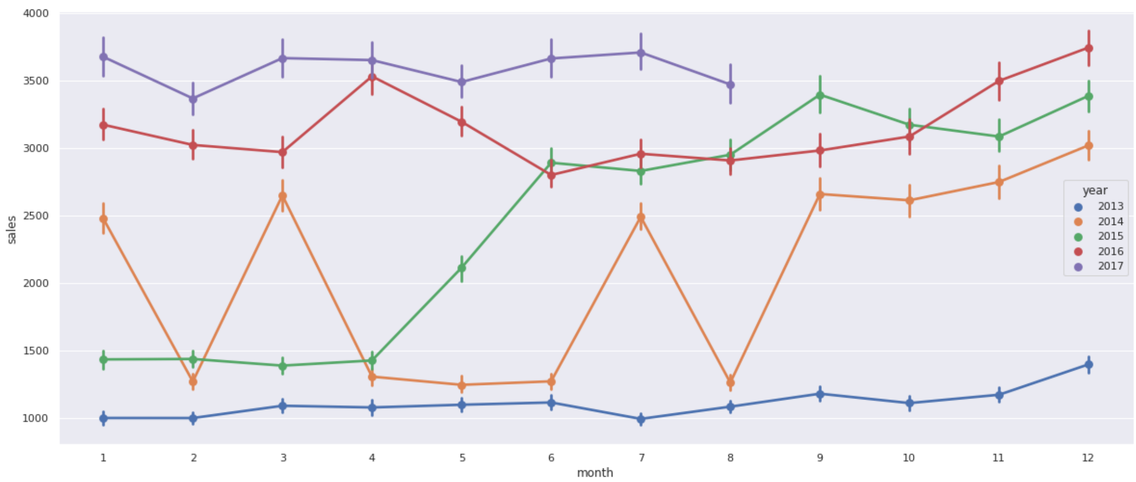
Number of Promotions on various family of Products



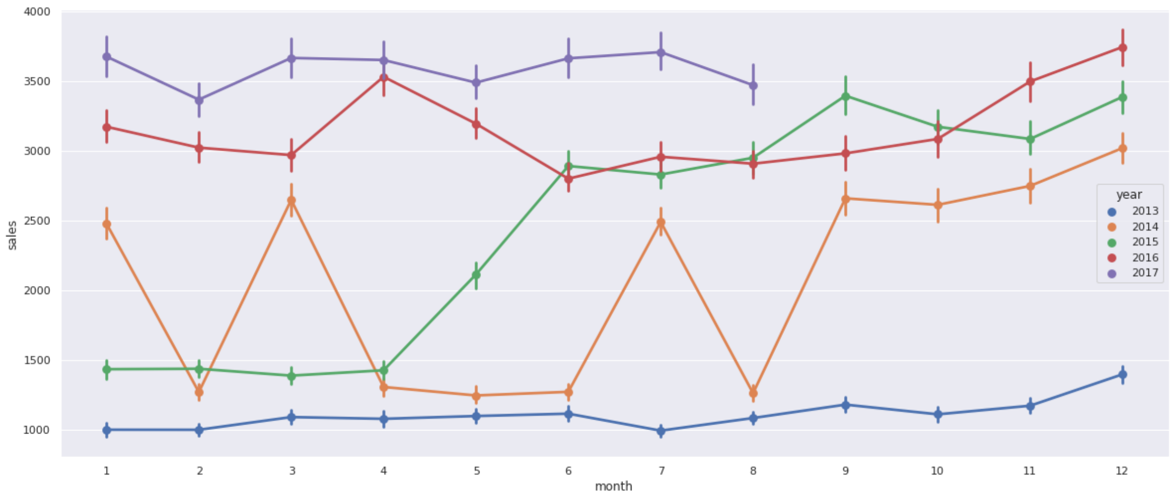
LVX VERITAS VIRTVS

Exploratory Data Analysis

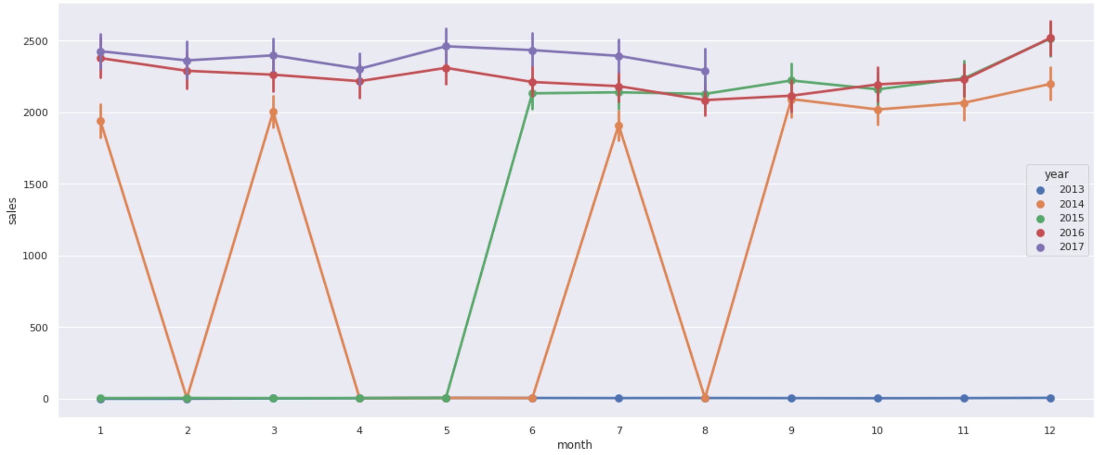
Trend in sales of Grocery I



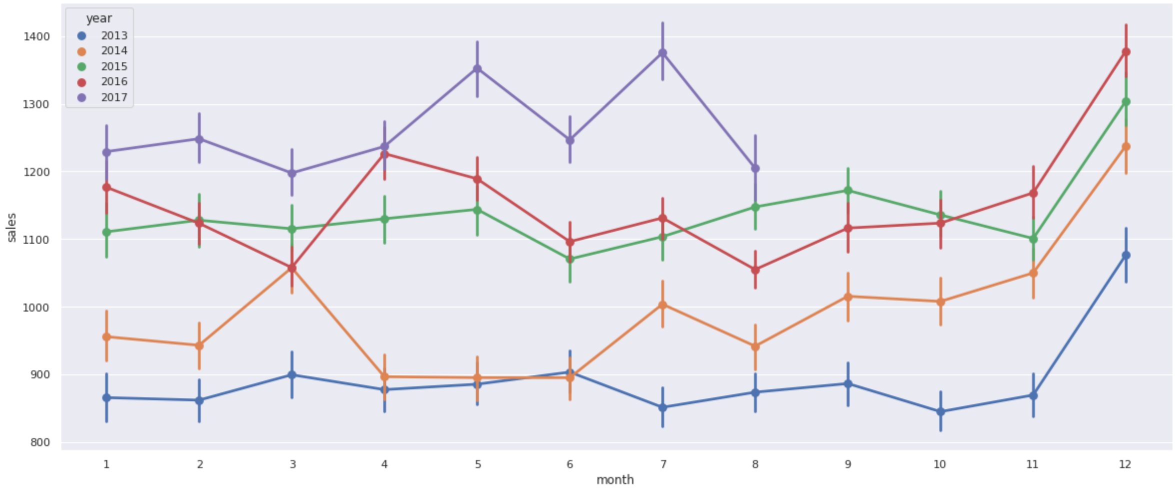
Trend in sales of Beverages



Trend in sales of Produce

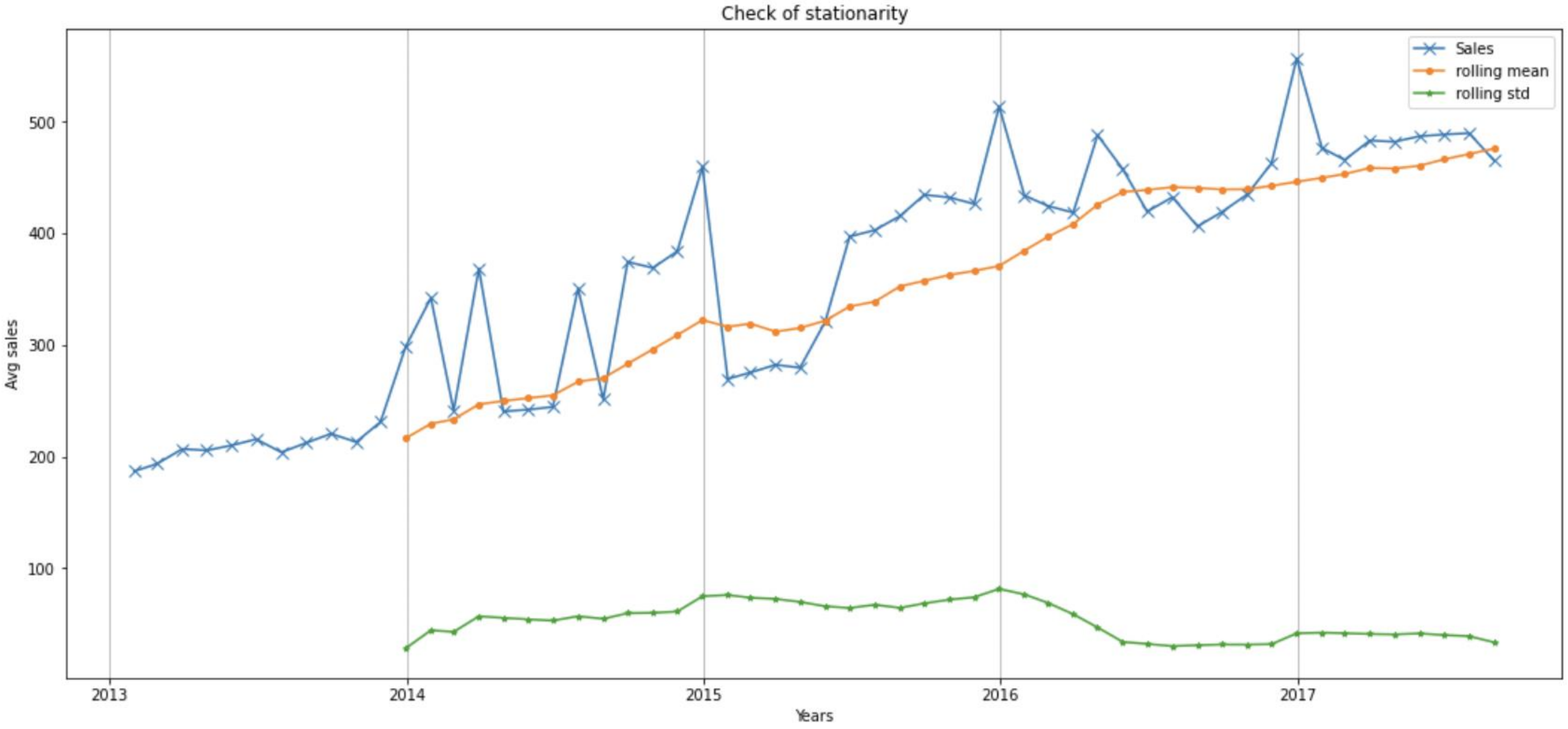


Trend in sales of Cleaning Supplies



Exploratory Data Analysis

Stationarity Check



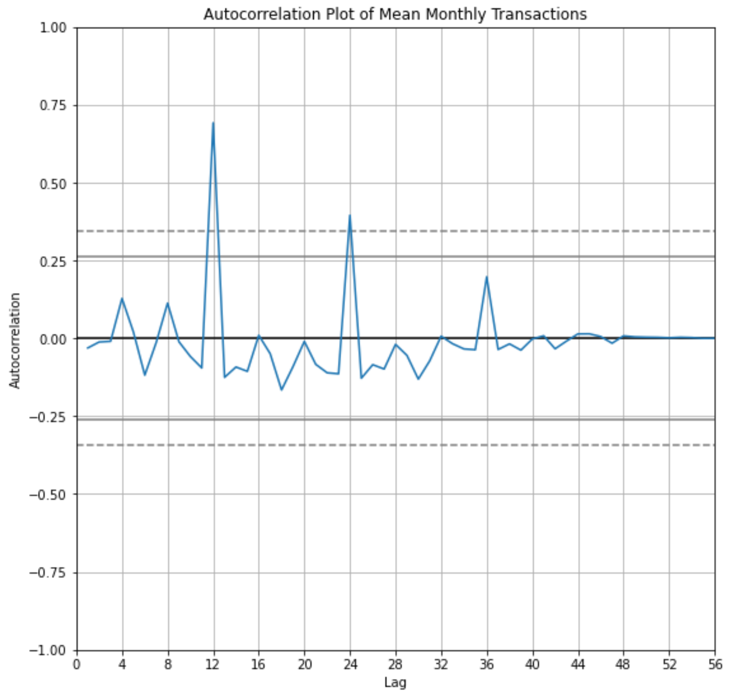
Before:
test statistic: -1.581928187307198
P-value: 0.492664903858226
Critiical Values {'1%': -3.55770911573439, '5%': -2.9167703434435808, '10%': -2.59622219478738}

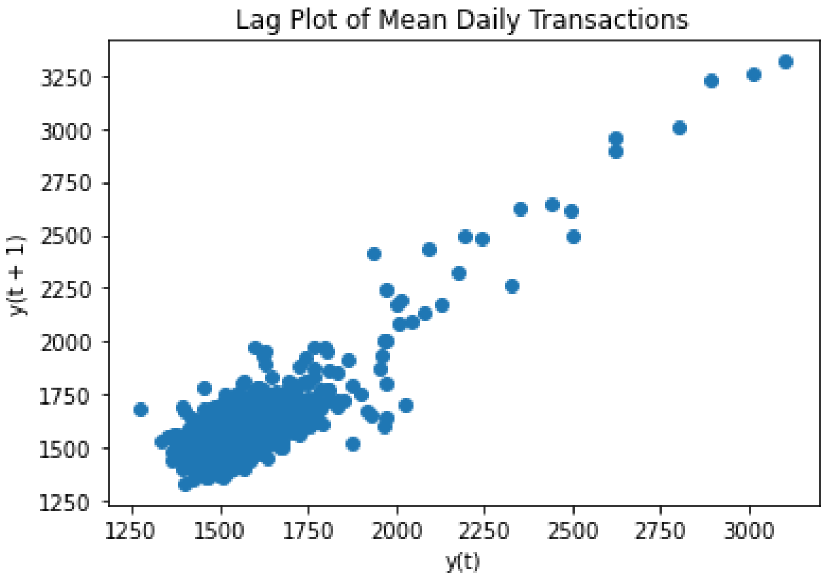
With 99%, 95% and 90% data is not stationary.

After:
test statistic: -3.7341645625933024
P-value: 0.0036574713013068556
Critical Values {'1%': -3.6055648906249997, '5%': -2.937069375, '10%': -2.606985625}

With 99%,95%,90% confidence the data is stationary

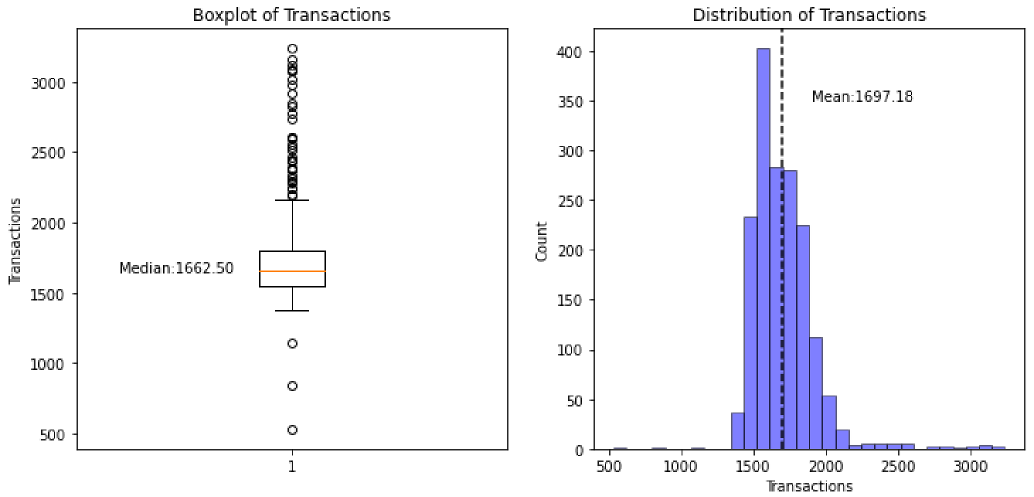
Exploratory Data Analysis

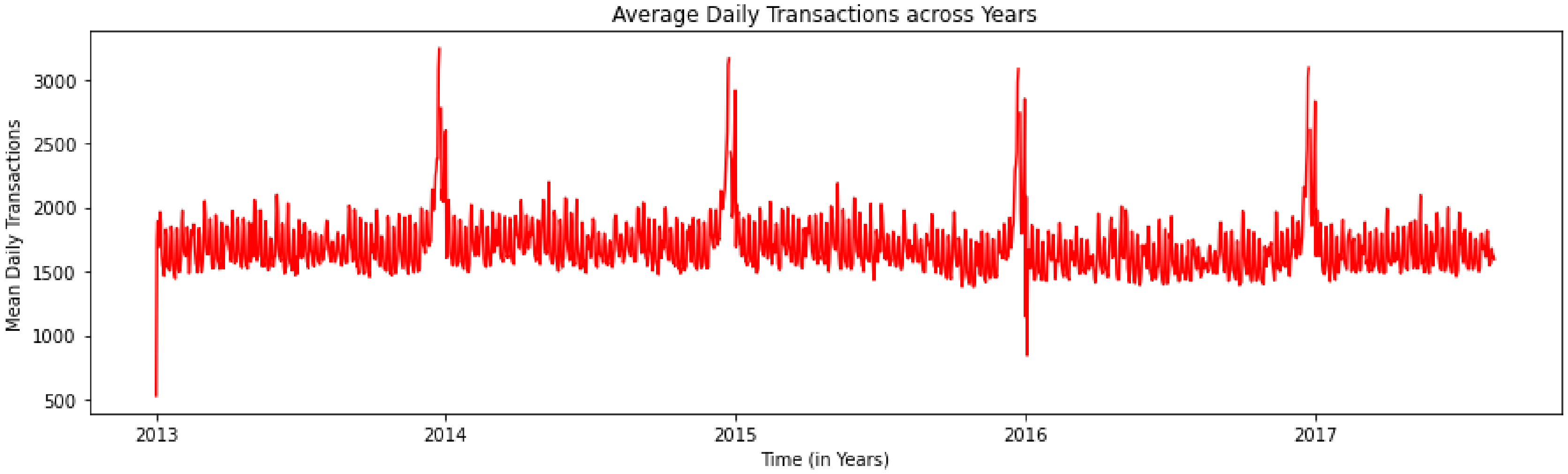




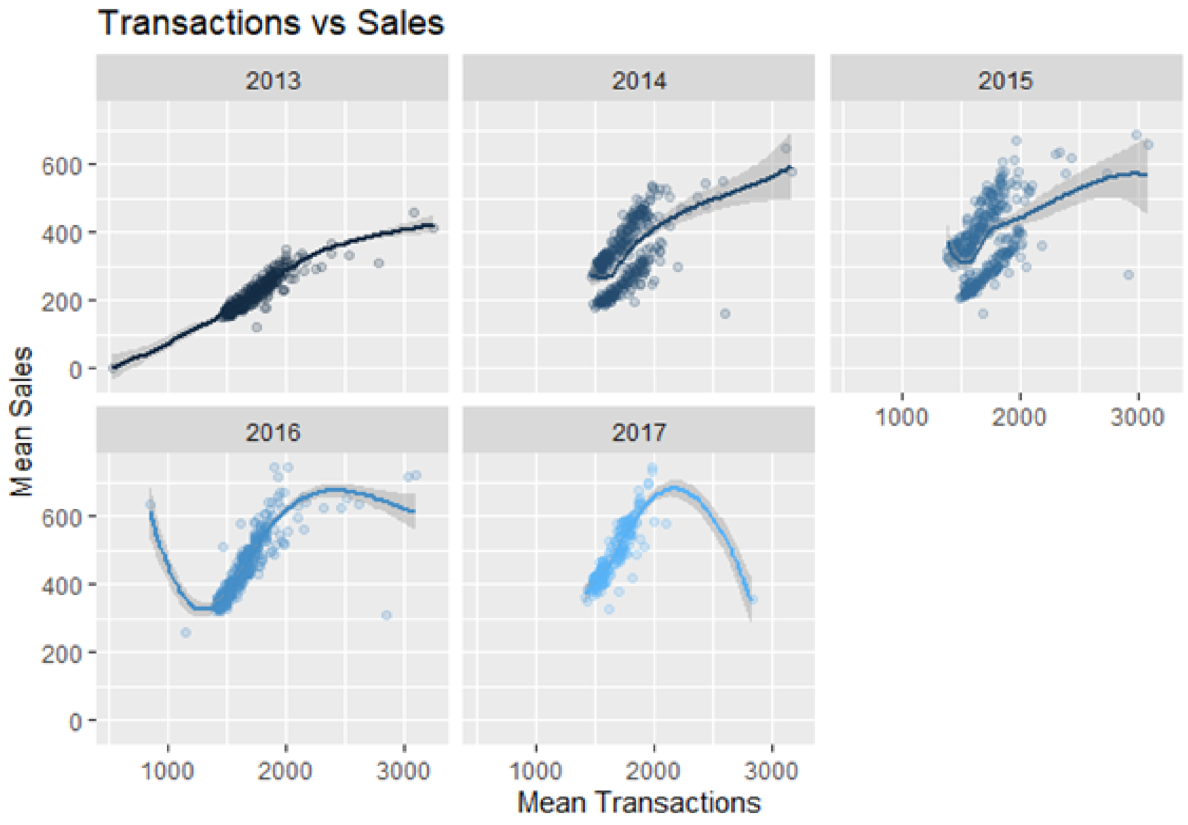
- There is a peak in the transactions for every 4 months, indicating there is correlation of transactions every 4 months.
- *The lag plot shows linearity between a transactions on current day and transaction on the previous day.*
- *This confirms that we can impute the null values in the transactions with mean transaction of previous day*

Exploratory Data Analysis





Exploratory Data Analysis



- The sales and transactions have a positive relationship across the years
- The gradient of the plot has increased over the years

Methodology - Statistical Forecasting Methods

Holt- Winters Exponential Smoothing

Objective here: Predict average monthly sales of the franchise over a period of 1 year.

Preprocessing: Resampling to monthly average

Approach: Please find the code in here.

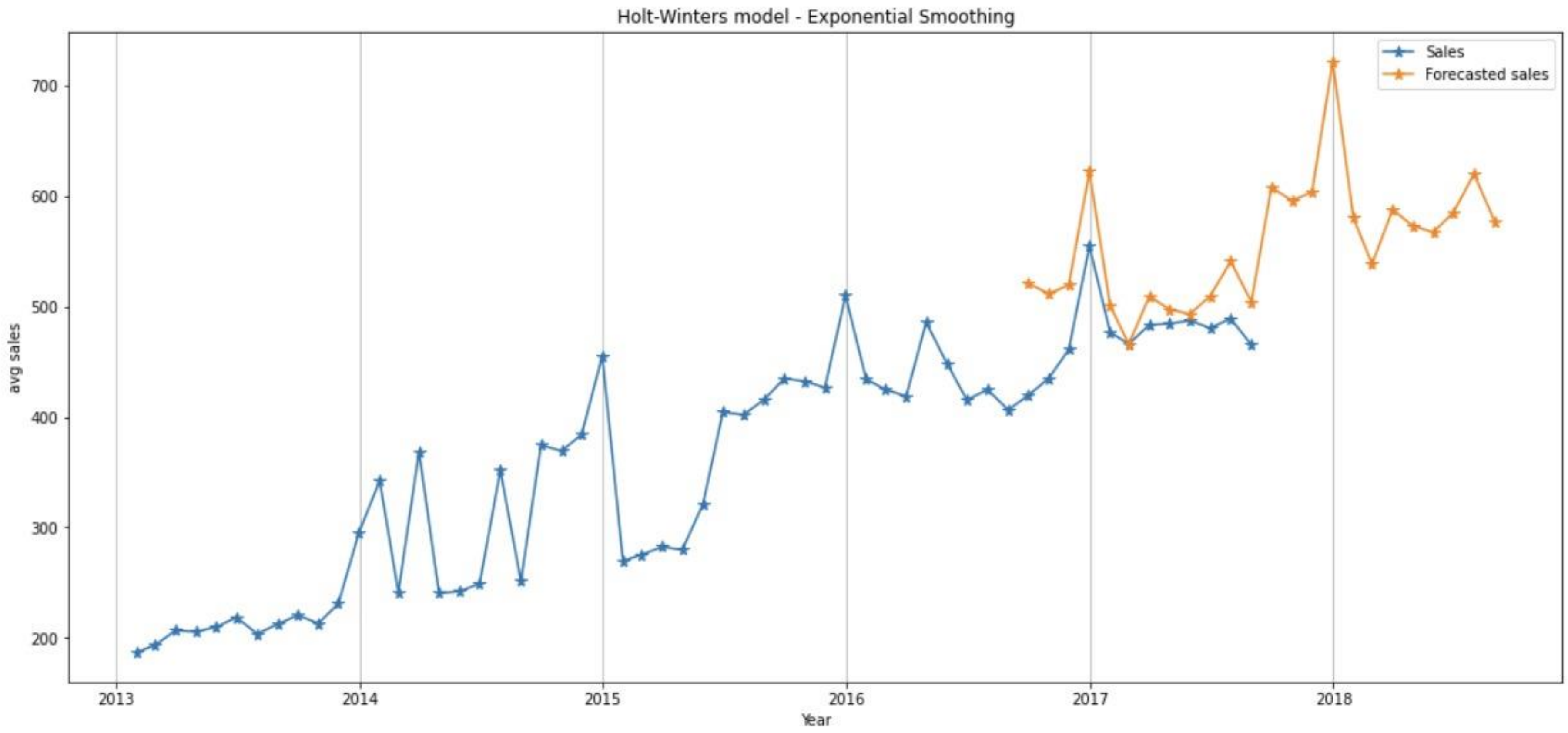
Step-1:Split the data into train and test data with 3:1 ratio.

Step-2: Modelling

Step-3: Evaluation Metric (MAPE)

Methodology - Statistical Forecasting Methods

Holt- Winters Exponential Smoothing



Evaluation Metric, MAPE : 7.05

Methodology - Statistical Forecasting Methods

Objective here:
- Forecast sales for 'Grocery I' family of products.
- Forecast sales for 'Store 44'

As they have the highest sales.

Data: Features engineered/used:
- P : Auto Regression model for lag
- D : Differencing
- Q : Moving Average – Biggest lag after which other lags are not significant
- **Exogenous feature**: holidays
- Conversion of date format.
- Re-sampling and taking the mean.

Approach: Please find the code in here

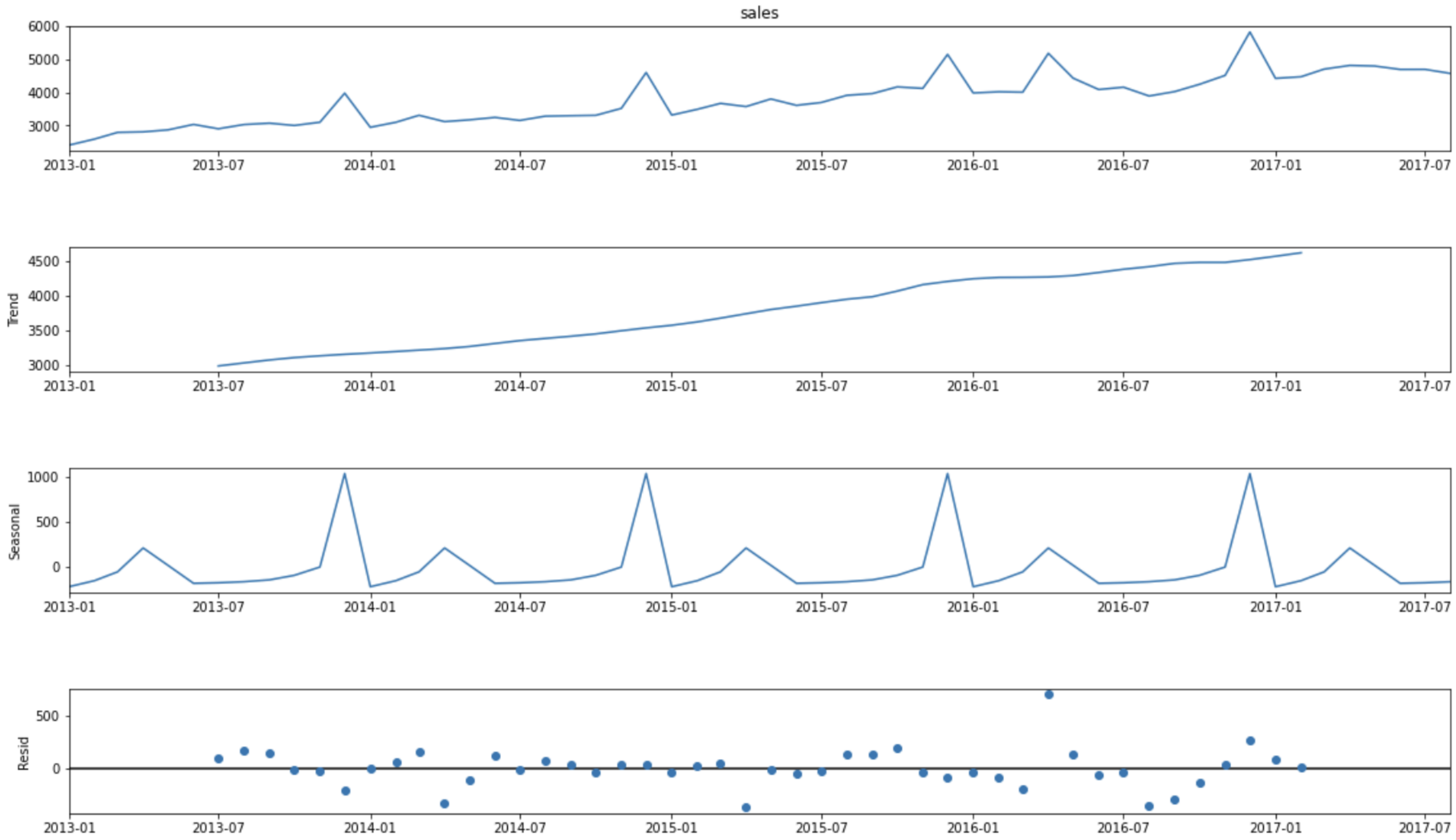
Comparison of different methods using MAPE.
- ARIMA using monthly mean sales data.
- ARIMA using log transformed daily sales data.
- SARIMAX using daily sales data.
- SARIMAX with exogenous feature for daily sales data.

Methodology - Statistical Forecasting Methods

Performing forecasting for 'GROCERY I' family of products.

ARIMA

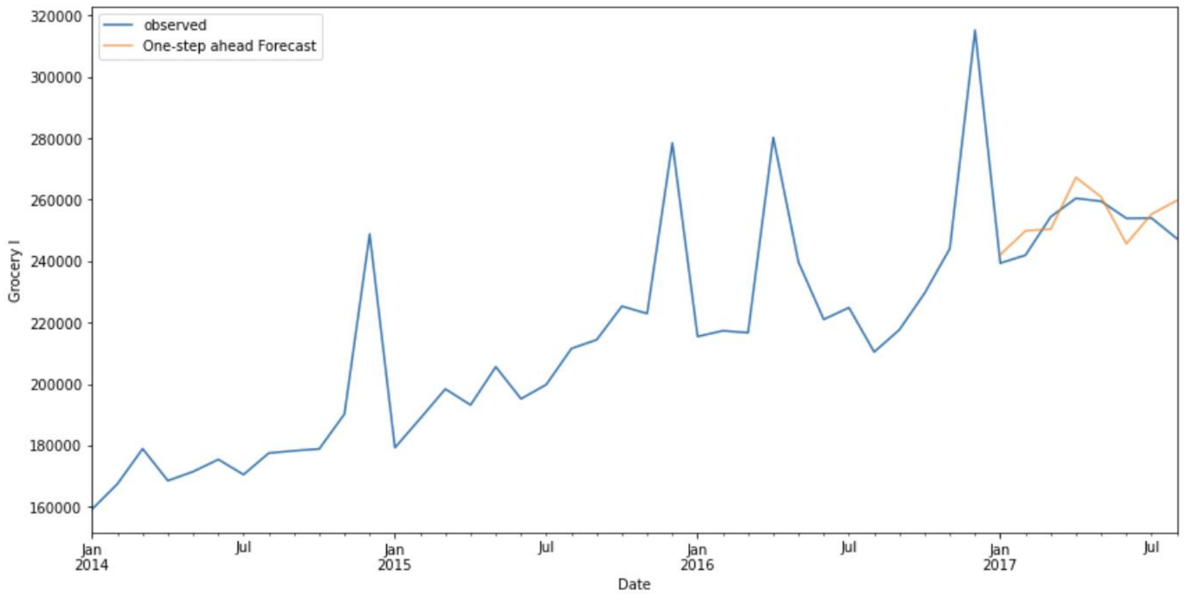
- Seasonality
- Trend
- Stationary



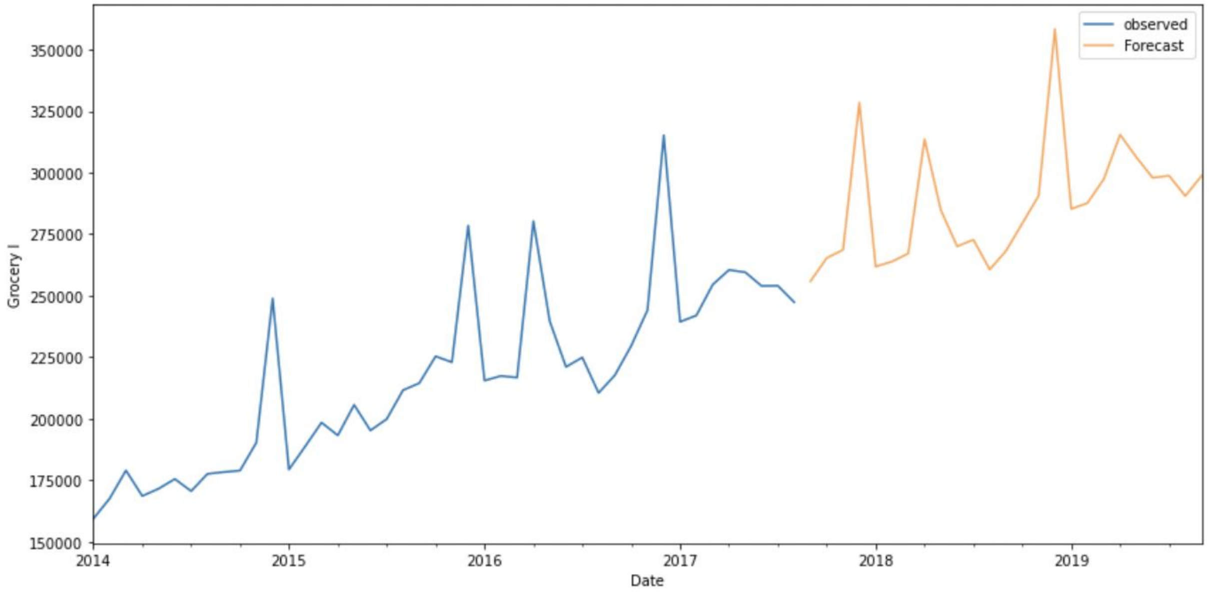
Methodology - Statistical Forecasting Methods

Performing forecasting for 'GROCERY I' family of products.

Performing validation by forecasting for 6 months.



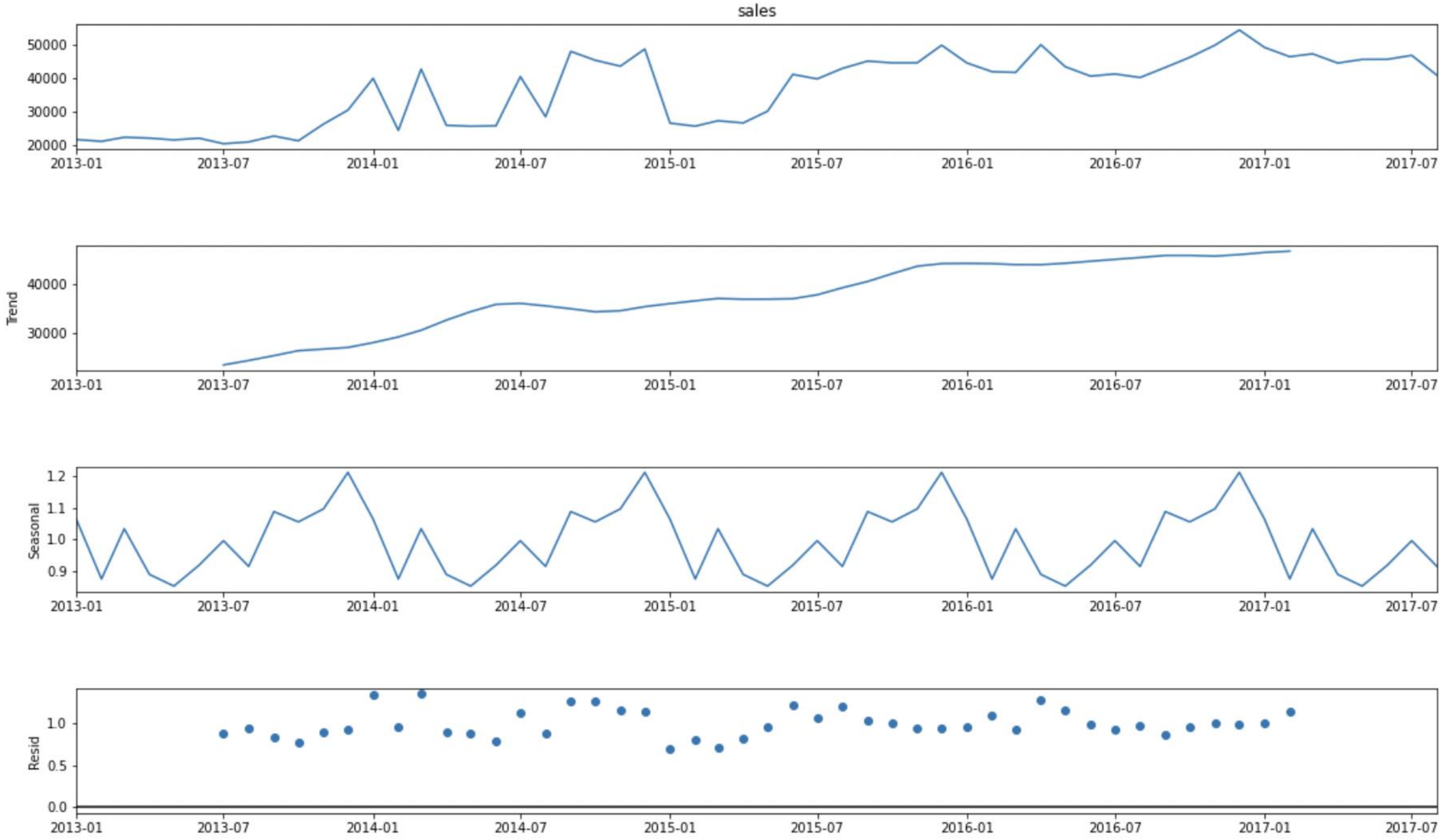
Performing Future Forecast



Evaluation Metric, MAPE: 2.25

Methodology - Statistical Forecasting Methods

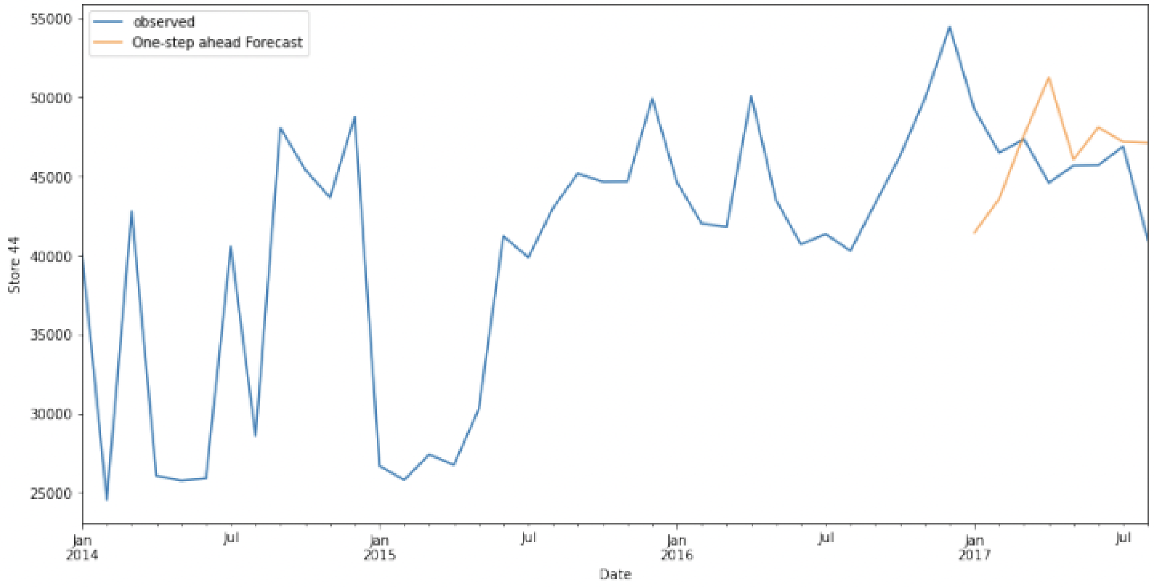
Performing forecasting for ‘STORE 44’.



Methodology - Statistical Forecasting Methods

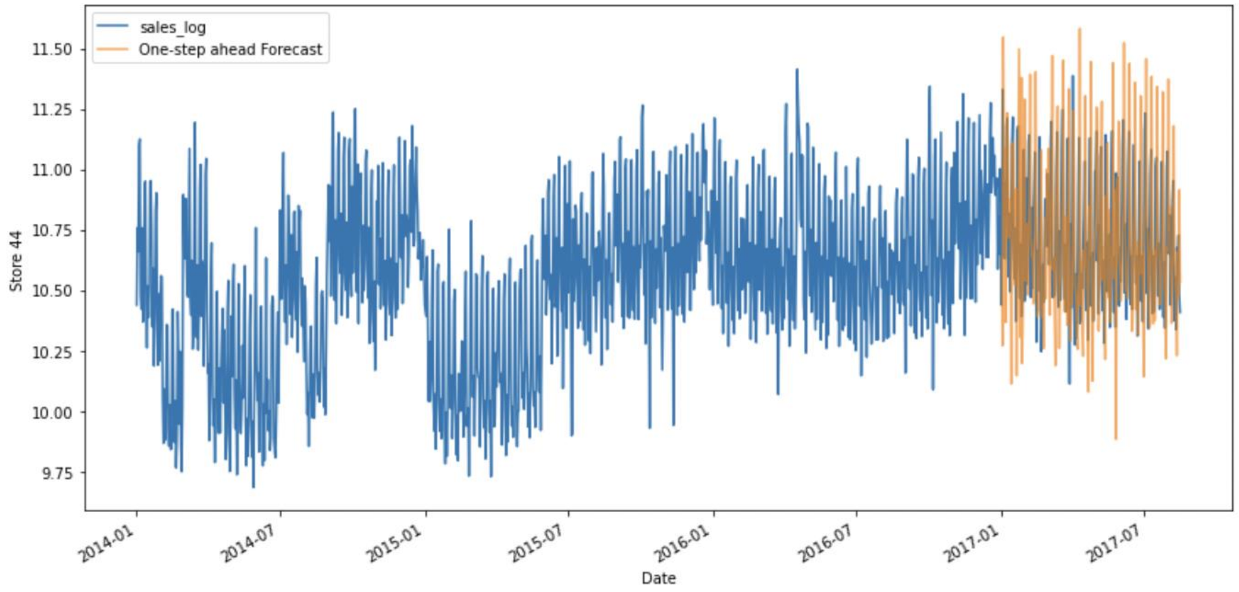
Performing forecasting for ‘STORE 44’.

Forecasts using mean of sales grouped by month



Evaluation Metric, MAPE: 7.42

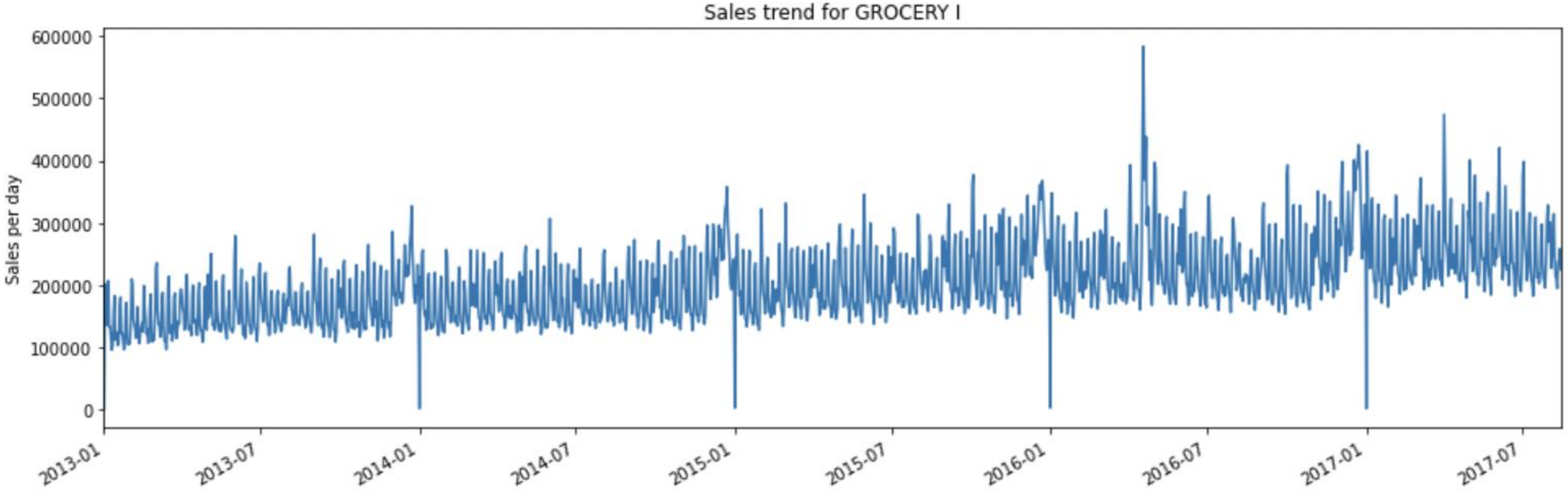
Forecasting using Log Transformation on number of sales



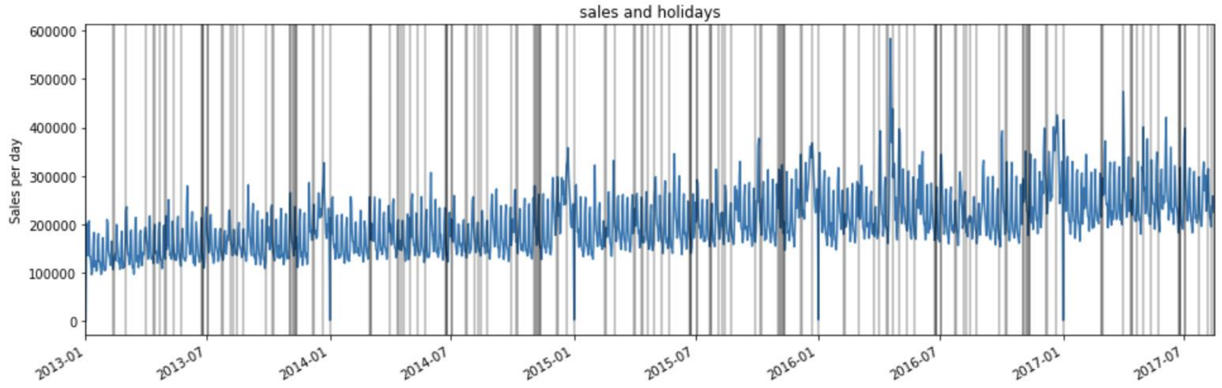
Evaluation Metric, MAPE: 3.1

Methodology - Statistical Forecasting Methods

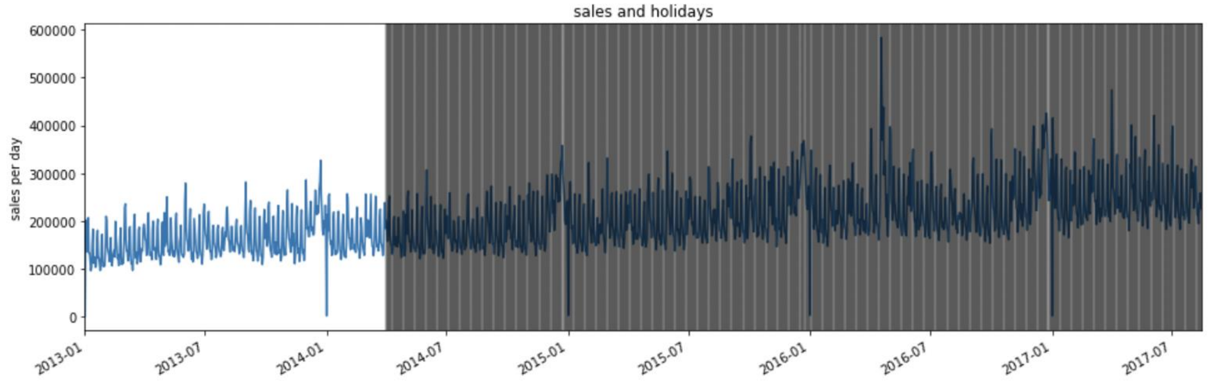
Performing forecasting for ‘Grocery I’ using SARIMAX.



Sales trends with Holidays



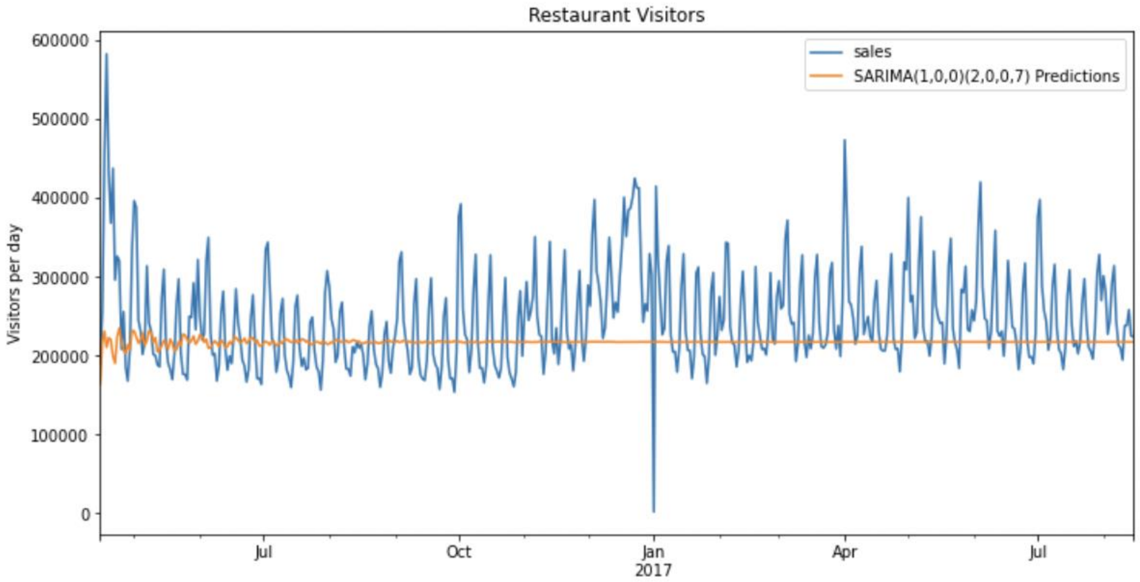
Sales trends with Promotions



Methodology - Statistical Forecasting Methods

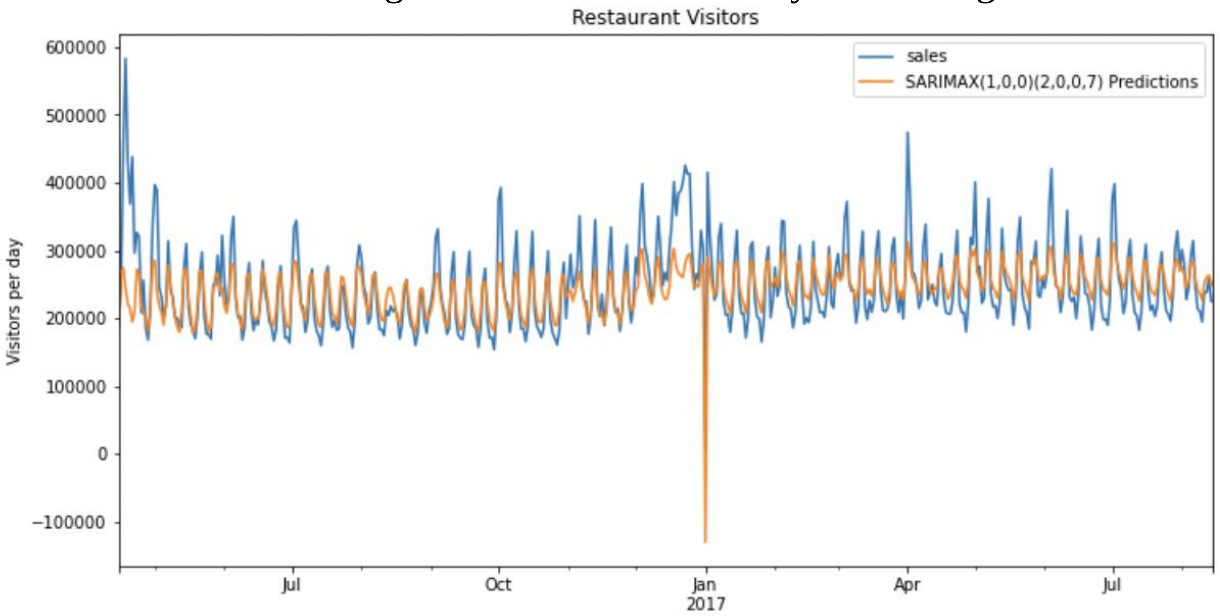
Performing forecasting for ‘Grocery I’ using SARIMAX.

Sales trends using SARIMAX



Evaluation Metric, MAPE: 67.34

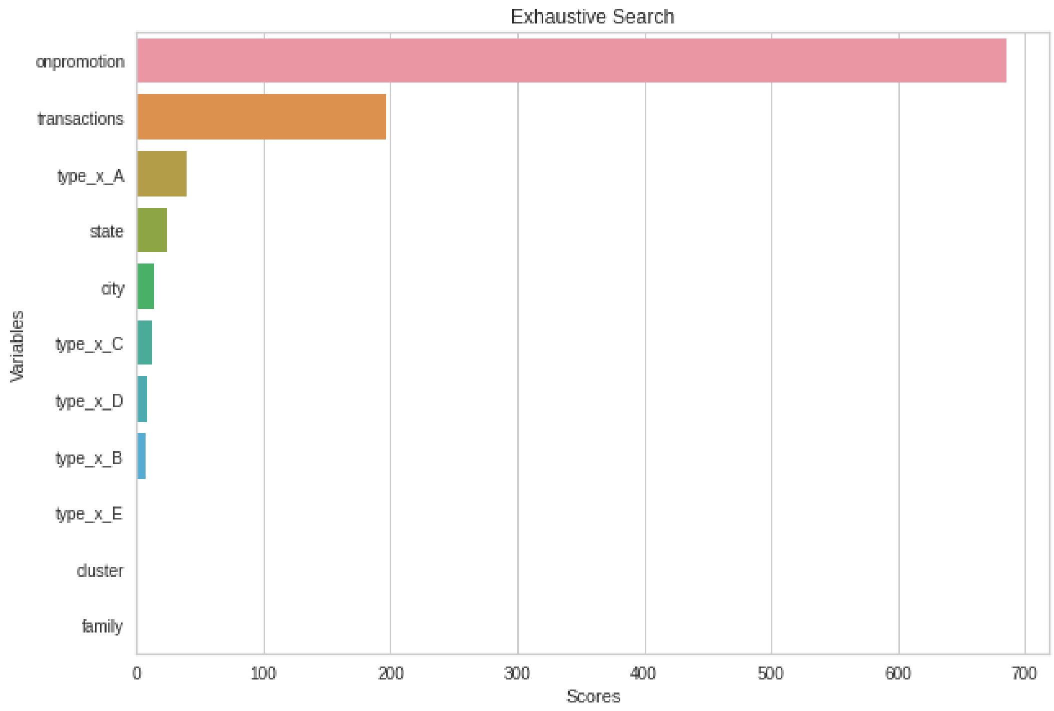
Sales trends using SARIMAX and Holiday as the exogenous variable.



Evaluation Metric, MAPE: 24.12

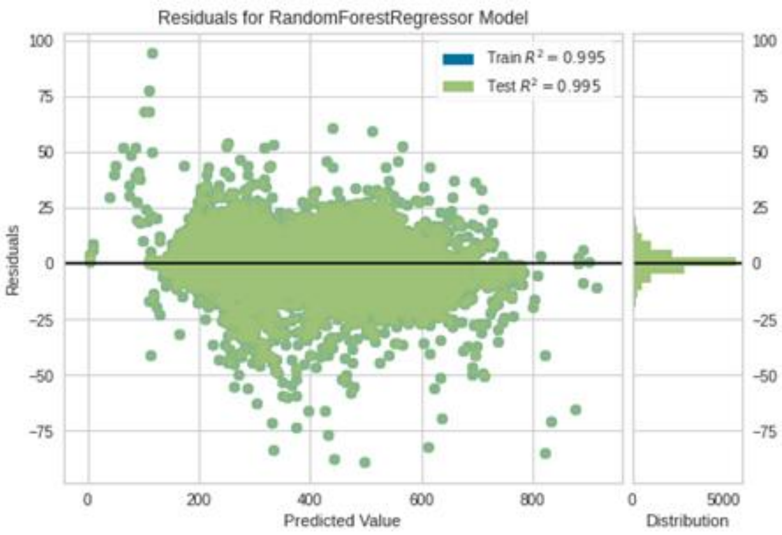
Methodology- Feature Selection

- Feature selection - Exhaustive Search is used to find the best features with their respective scores.
- F_regression is used as a scoring function.
- The top 9 features were selected from results found



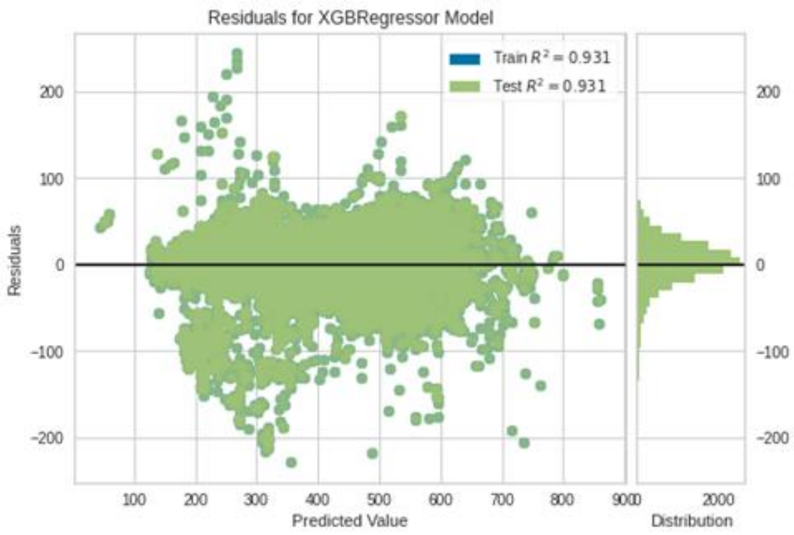
Methodology – ML Models

RANDOM FOREST REGRESSOR



RMSE = 8.732

XGBOOST REGRESSOR

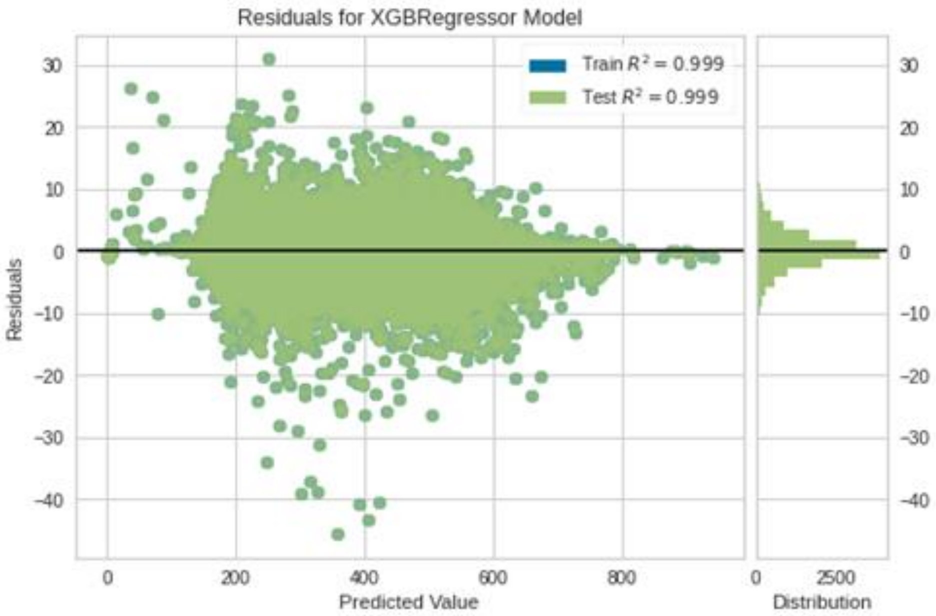


RMSE = 34.223

Methodology – Hyperparameter Tuning

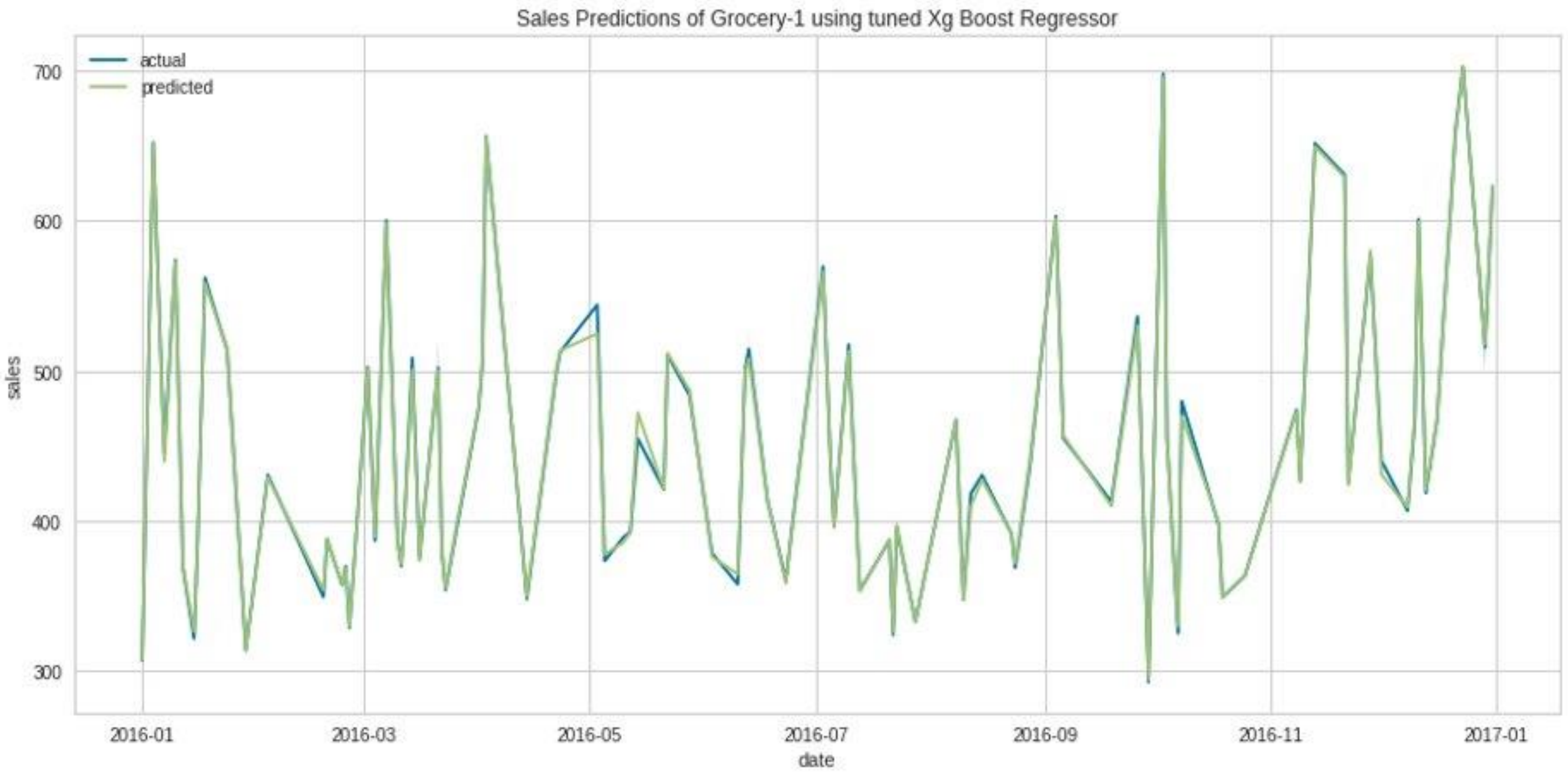
- Hyperparameter tuning is performed for XG-Boost regressor with randomized search cv
- The best parameters were chosen after 5 iterations and 10 Fold cross validation and the results are as follows:

```
#considering best estimators from the randomised search
XGB1 = XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=0.9, gamma=0.4, gpu_id=-1,
             importance_type='gain', interaction_constraints='',
             learning_rate=0.25, max_delta_step=0, max_depth=10,
             min_child_weight=7,  monotone_constraints='()',
             n_estimators=100, n_jobs=8, num_parallel_tree=1, random_state=0,
             reg_alpha=0, reg_lambda=1, scale_pos_weight=1, subsample=1,
             tree_method='exact', validate_parameters=1, verbosity=None)
```



RMSE = 4.143

Methodology- Hyperparameter Tuning



The performance was evaluated on the sales of Grocery 1 for the year 2016 with tuned Xgboost Model.

Methodology – ML Models

Objective here: Predict sales of ***per unit per day per store*** over a period of 1 year (demo for Grocery-I for cluster 5 of stores)

Data: Features engineered/used:

a. **Lag Feature**: Previous year sales
b. **Standard time series features**: day, week, month, year etc
c. **Exogenous feature**: store related, holidays related (transferred, Local)

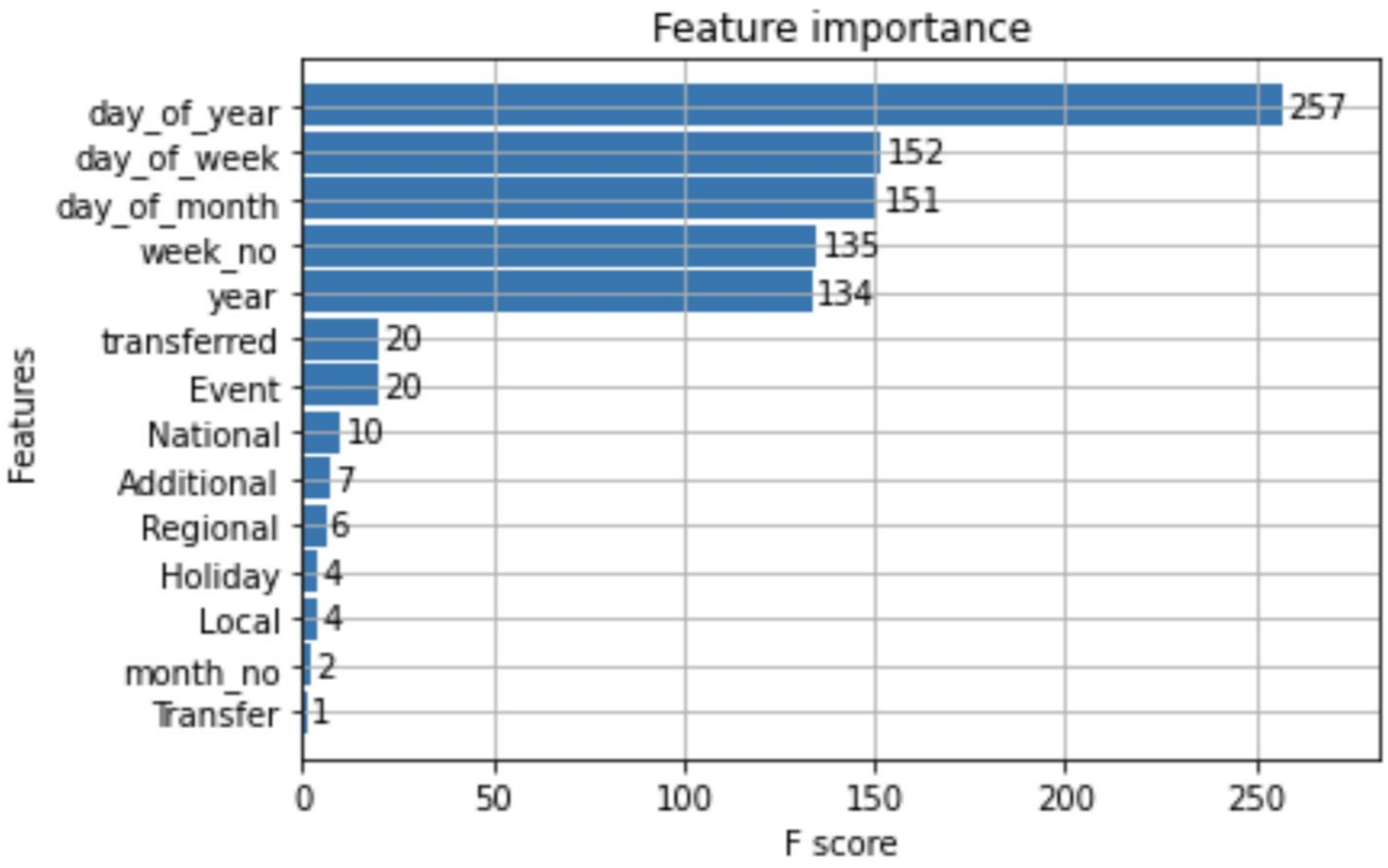
Problems: While modeling time series as an ML problem, issues and their resolutions:

a. feature selection (as values of features might not be available in future) – time invariant features
b. feature engineering (models understand only numerical data) – one hot encoding
c. Null values/missing values/zero values as at daily level – been removed to avoid introducing bias

Approach: Please find the code here and in the colab notebook here.

Step-1: Data Imports and EDA
Step-2: Data Pre-processing (data cleaning, feature engineering, one-hot encoding)
Step-3: Test-Train Split (Out of 4 years of daily data, testing on last 1 year)
Step-4: Modelling (XGBoost and Random Forest)
Step-5: Evaluation Metric (MAPE)

Methodology - ML Models



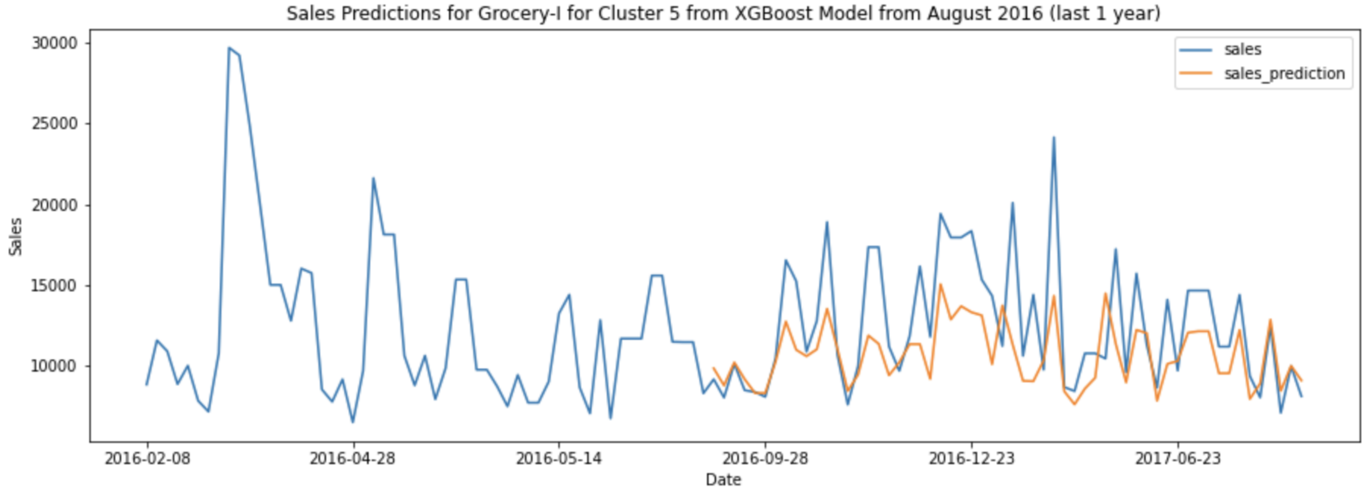
XGBoost

```
[61] subset_sales[["store_nbr","day_of_week","Holiday","Work Day","sales","sales_prediction","Error_perc"]].tail()
```

date	store_nbr	day_of_week	Holiday	Work Day	sales	sales_prediction	Error_perc
2017-07-25	44	2	0	0	8047.0	8913.0	10.76
2017-08-05	44	6	1	0	12463.0	12855.0	3.15
2017-08-10	44	4	1	0	7097.0	8453.0	19.11
2017-08-11	44	5	0	0	9979.0	10005.0	0.26
2017-08-15	44	2	1	0	8123.0	9098.0	12.00

```
plt.figure()
subset_sales[['sales','sales_prediction']].plot(figsize=(15, 5))
plt.xlabel("Date")
plt.ylabel("Sales")
plt.title("Sales Predictions for Grocery-I for Cluster 5 from XGBoost Model from August 2016 (last 1 year)")
plt.show()
```

```
<Figure size 432x288 with 0 Axes>
```




```
[50] mape(y_test, y_pred)
```

```
26.4065130478286
```

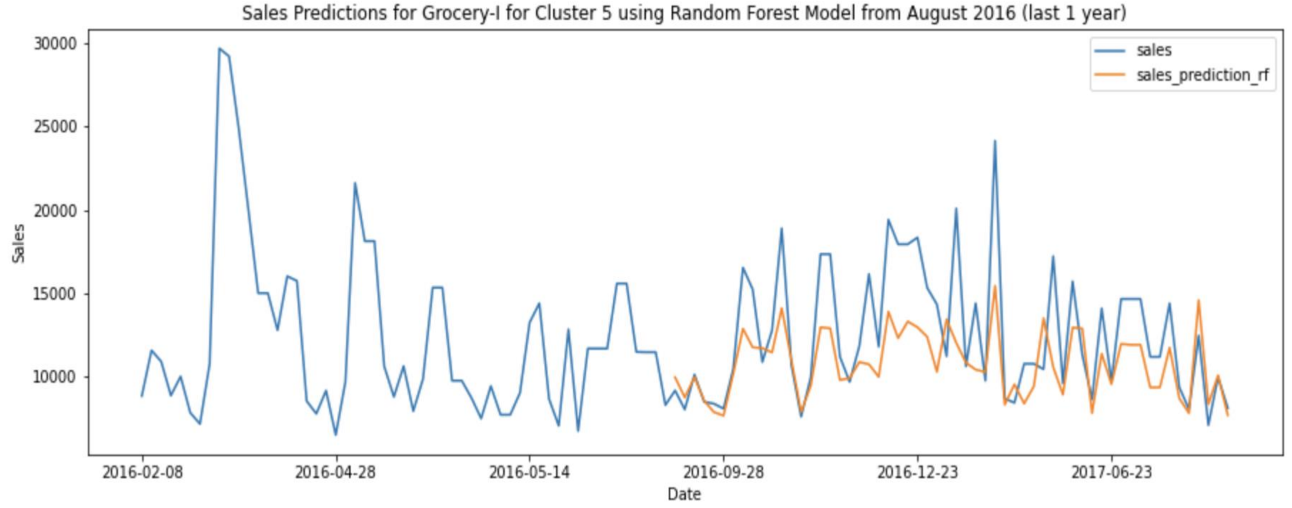
RandomForest

```
[64] subset_sales_rf[["store_nbr","day_of_week","Holiday","Work Day","sales","sales_prediction_rf","Error_perc"]].tail()
```

date	store_nbr	day_of_week	Holiday	Work Day	sales	sales_prediction_rf	Error_perc
2017-07-25	44	2	0	0	8047.0	7814.0	-2.90
2017-08-05	44	6	1	0	12463.0	14588.0	17.05
2017-08-10	44	4	1	0	7097.0	8351.0	17.67
2017-08-11	44	5	0	0	9979.0	10079.0	1.00
2017-08-15	44	2	1	0	8123.0	7695.0	-5.27

```
plt.figure()
subset_sales_rf[['sales','sales_prediction_rf']].plot(figsize=(15, 5))
plt.xlabel("Date")
plt.ylabel("Sales")
plt.title("Sales Predictions for Grocery-I for Cluster 5 using Random Forest Model from August 2016 (last 1 year)")
plt.show()
```

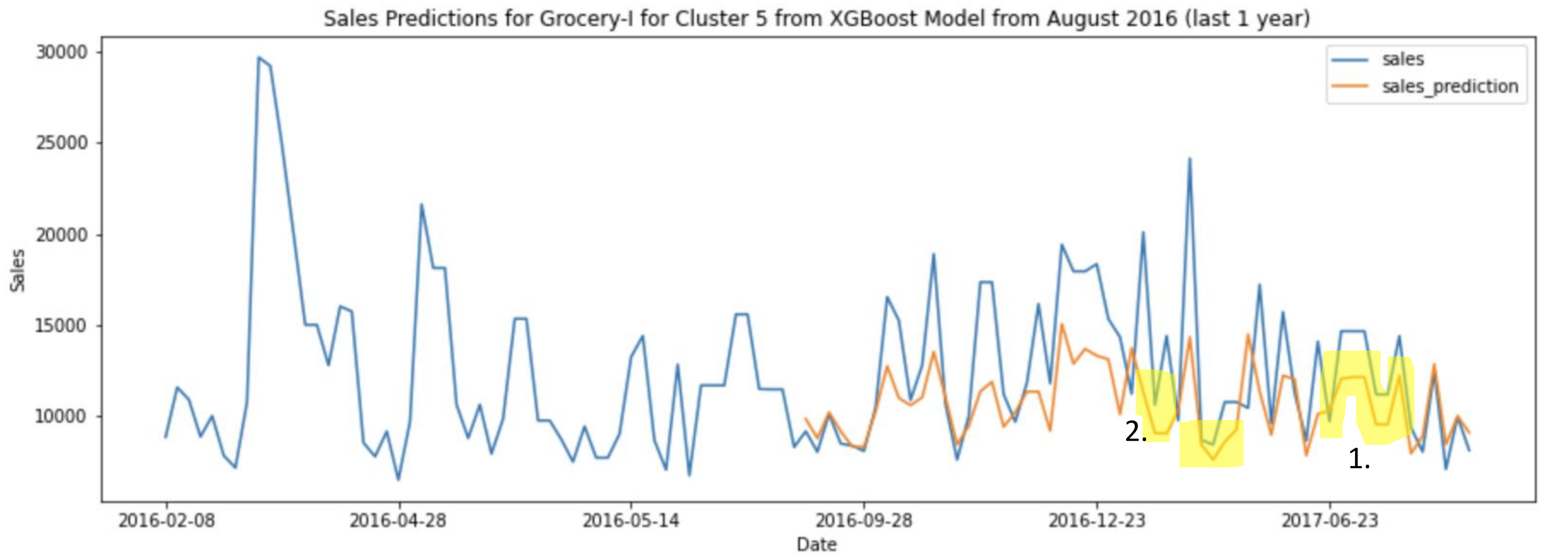
<Figure size 432x288 with 0 Axes>

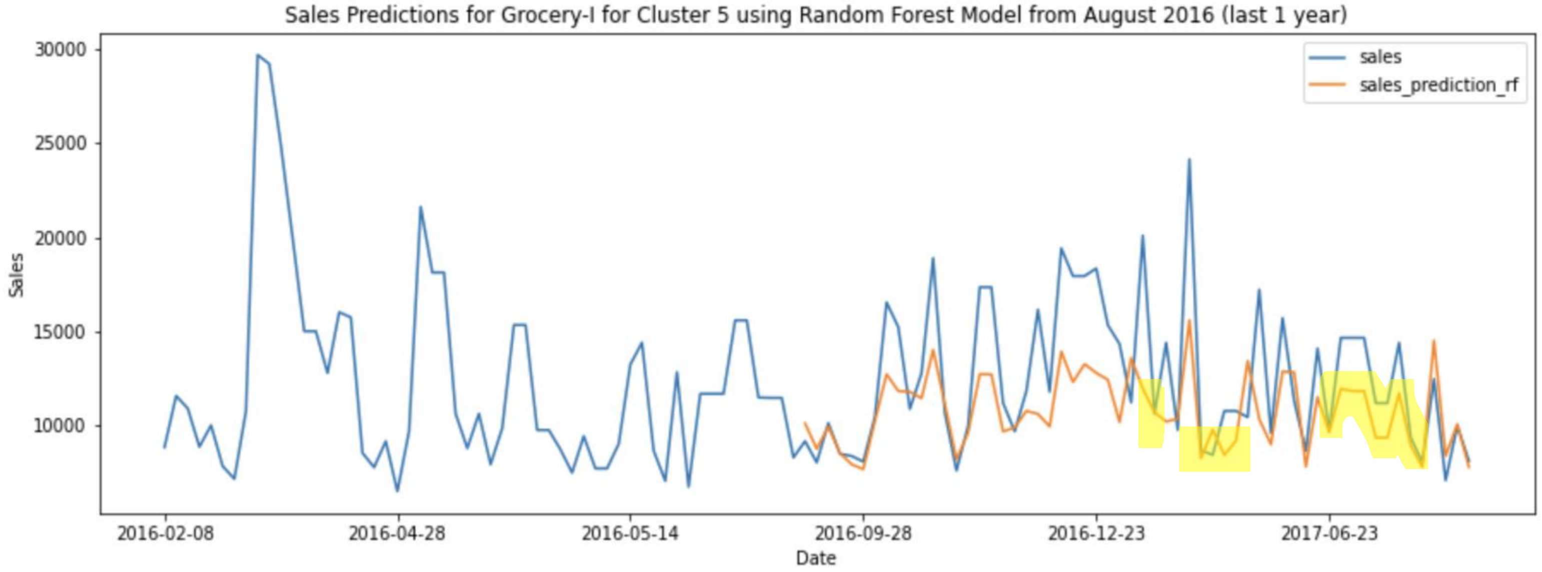



```
[59] mape(y_test, y_pred)
```

26.80456359567298

Methodology - ML Models





XGBOOST (XGB) vs Random Forest (RF):

Quantitative:
MAPE : 26.40 (XGB) vs 26.80 (RF) are similar!
As both are tree-based ensemble methods.

Qualitative:

1. Both the models identified some pattens in the data. With more engineered features the predict power of the models might increase.
2. Random forest predictions show lesser extreme variations as compared to XGB predictions

Results

Prediction Objective	Model	Model Type	MAPE	Training and Test Data	Usecase
Yearly per unit sales for next n months	Holt-winters Exponential Smoothing	Statistical	7.05	Univariate	Annual Budgets and Plans in Company Review
Monthly per unit sales for next n months	ARIMA	Statistical	2.25, 3.1 and 7.42	Univariate	Supply chain and demand forecasting planning
Monthly per unit sales for next n months	SARIMAX	Statistical	67.34 and 24.12	Univariate and Multivariate	Supply chain and demand forecasting planning
Daily Per unit sales per store for next 1 year	XGBoost	Machine Learning (Decision tree-based ensemble (Boosting))	26.40	Multivariate	Resource allocation and cashflow management planning
Daily Per unit sales per store for next 1 year	Random Forest	Machine Learning (Decision tree-based ensemble ML mode (Bagging))	26.80	Multivariate	Resource allocation and cashflow management planning

LVX VERITAS VIRTVS

Conclusion and Future Works

- **Conclusion**: We analyzed the sales and transactions data at different cross sections along with exogenous factors and evaluated specific modelling techniques for different use cases of time series forecasting of sales for a retail store.
- **Future Works**:
  - Expansion of Univariate analysis to multivariate analysis
  - Feature engineering and hybrid models
  - UI for better user experience
  - Build a ML pipeline using MLOps

LVX VERITAS VIRTVS

Thank you

Open to Questions