FairStorage

Demand Forecasting

Ahmed Fedi Lassoued
Daniel Lim
Tristan Koh
Vasu Namdeo
Zen Goh

01

Executive Summary

02

EDA Insights

03

Best Approach

04

The Alternatives

05

Limitations

06

Business Insights

01

Executive Summary

Executive Summary

TASK

Ensure that enough stock is available to meet forecasted demand

ISSUE

Data is profuse, difficult to parse out real granular trends

SOLUTION

Use Computational Methods to aid in Time Series Forecasting

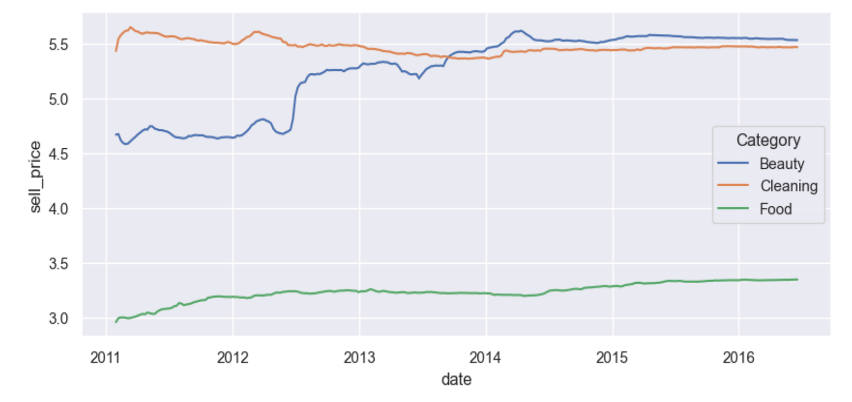
02

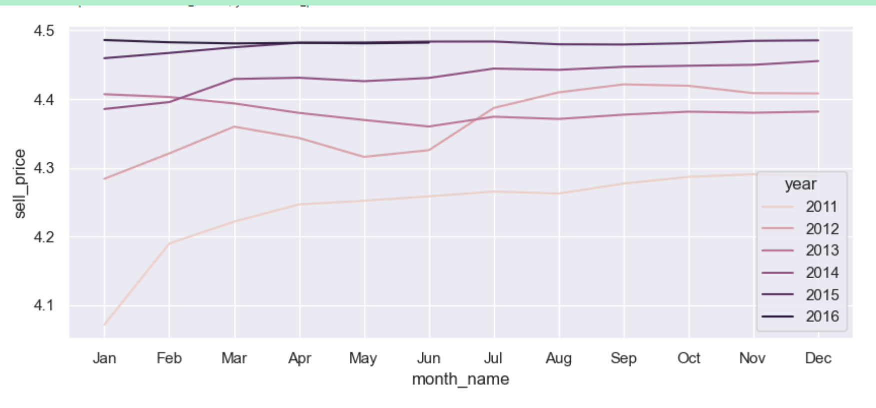
Exploratory Data Analysis

Structure of Data

30490 Total records
3049 Items
1919 Days
10 Different stores
3 Categories of items (beauty, cleaning, food)
7 Subcategories of items (beauty1-2, cleaning1-2, food1-3)
3 Different regions (east, central west)

Prices – Initial Observations

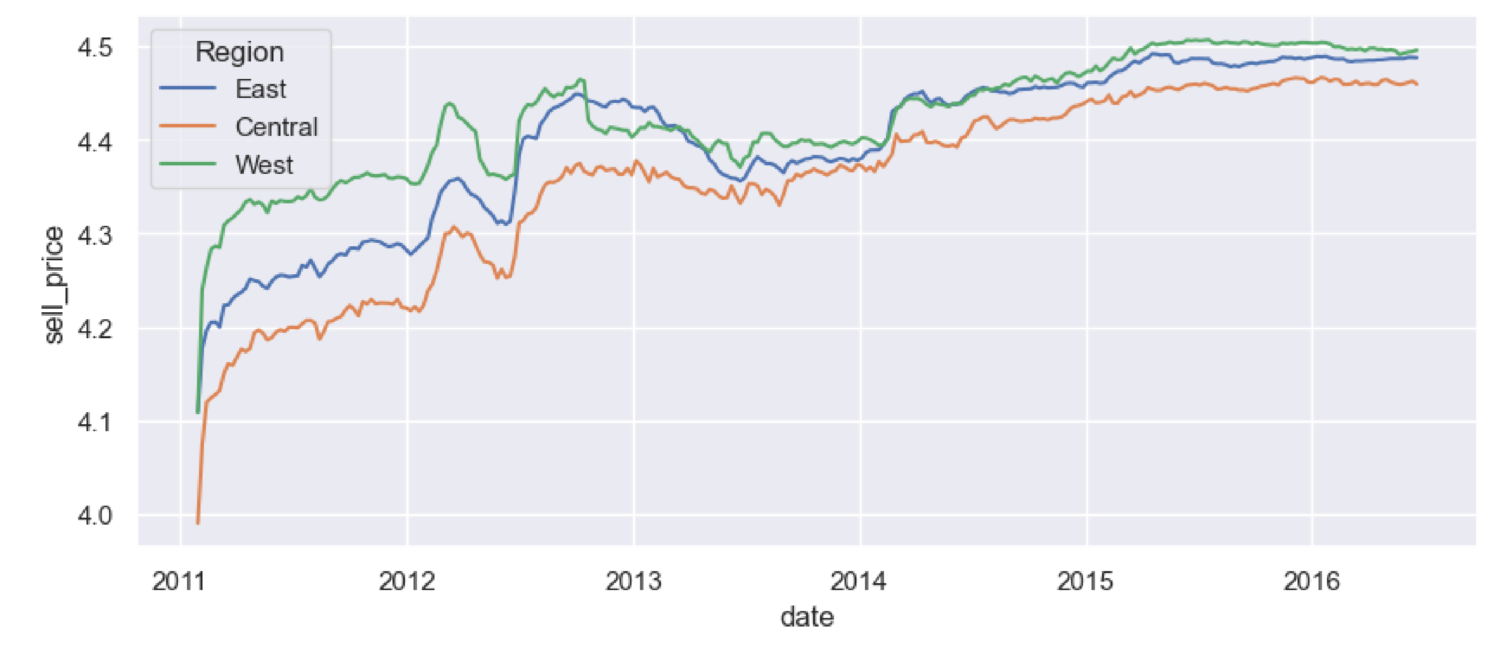




Gradual increase of prices for beauty and food
Slight decrease for prices of cleaning products

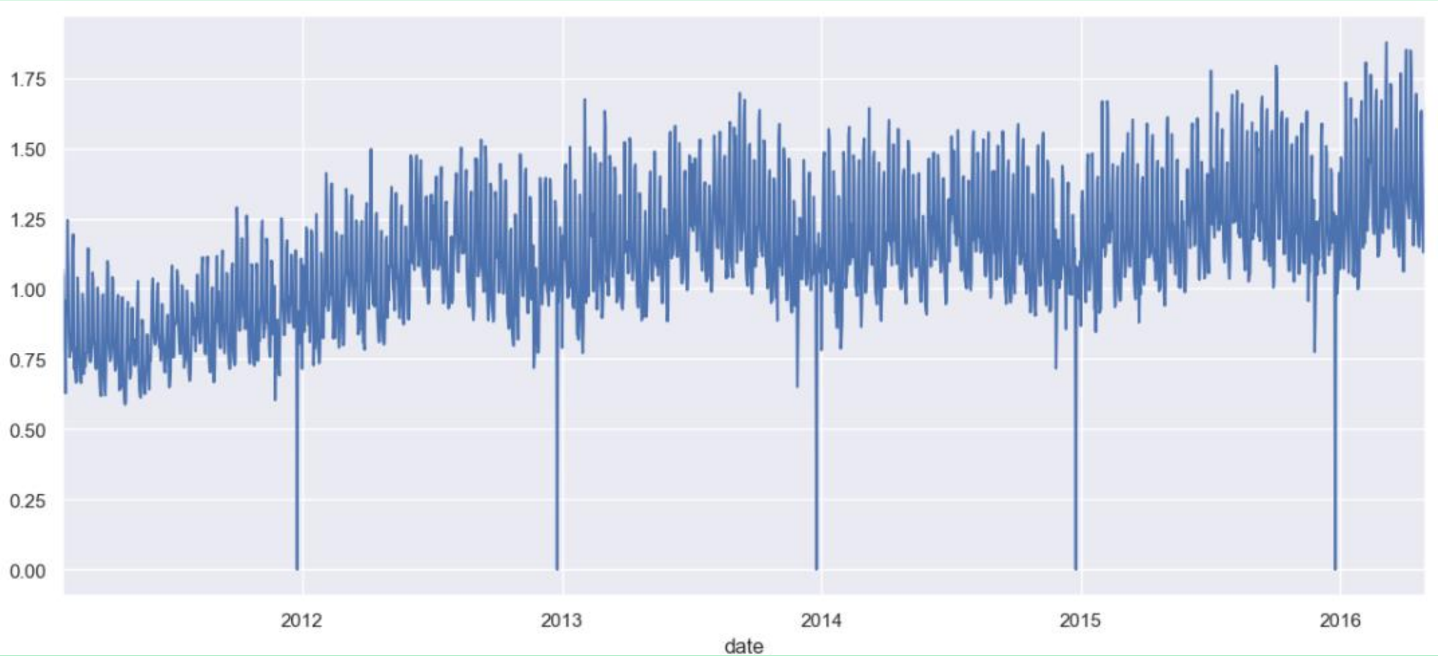
Increase of overall prices
Lack of any obvious trend across months

Prices – Increasing Prices (by Region)



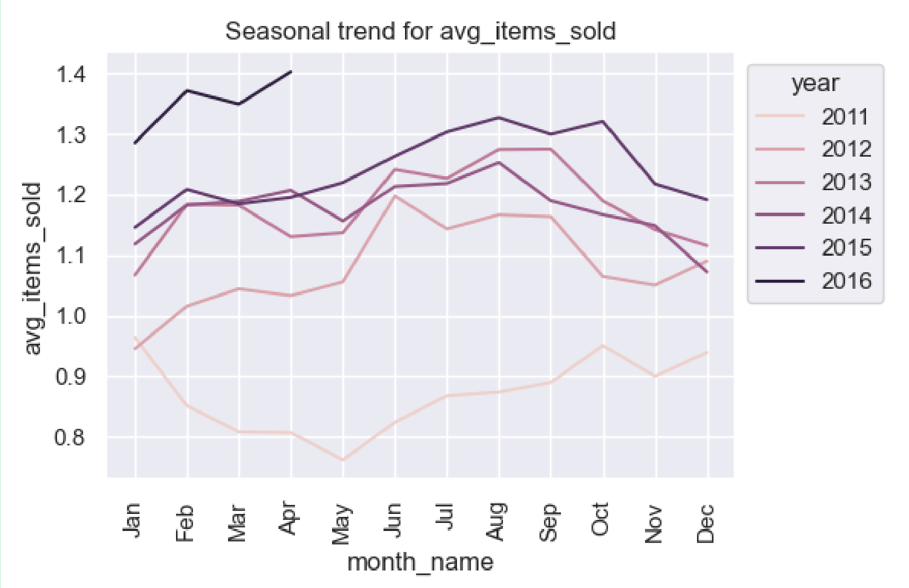
Difference between the regions becomes smaller in recent years

Orders – Increasing Demand



Note the outliers, occurring near the end of every calendar year

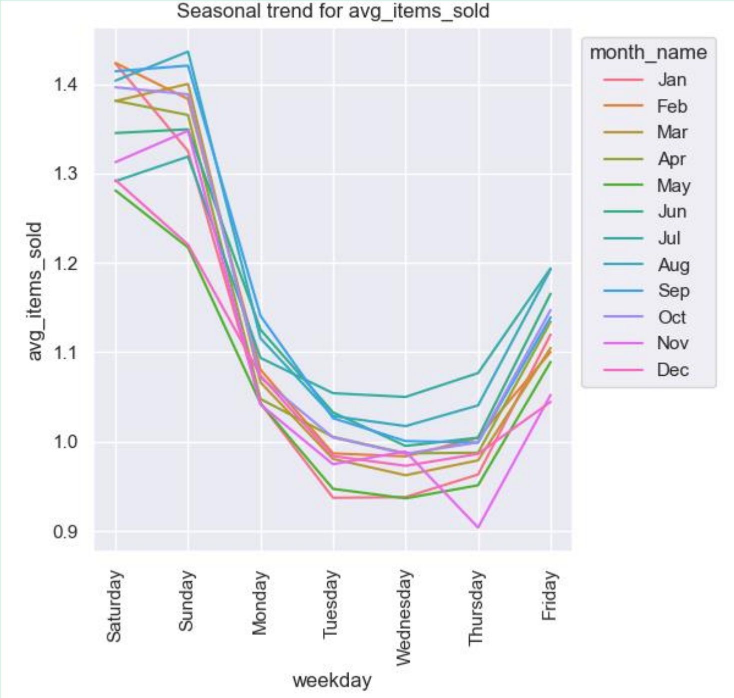
Orders – Single Peak Trend



2011 stands as the only exception, unlikely for trend to reoccur

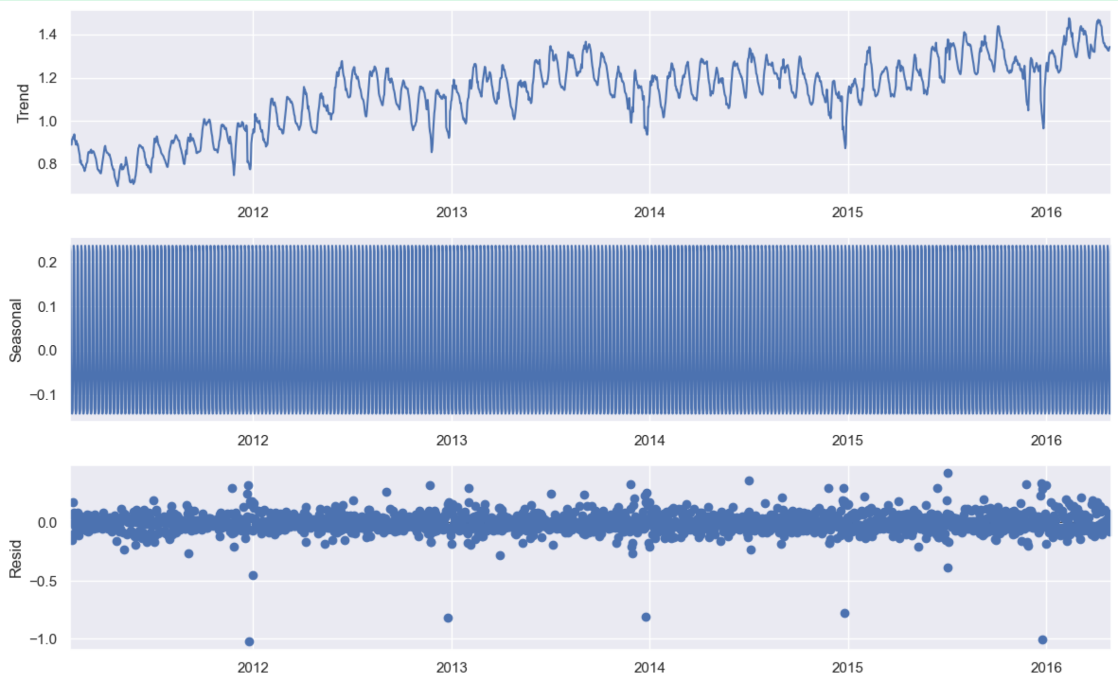
Volatility from Apr - Jun

Orders – Significant Seasonal Trend by Week



Sales are high on the weekends, decrease by a decreasing amount over the weekdays and begin to rise back up Thursday onwards

Orders – Decomposition Plot



Seasonality

Indicates strong cycles or seasonality in the data

Residuals

Variance does not change much over the years

Stationarity – ADF v KPSS

ADF

```
Results of Dickey-Fuller Test:
Test Statistic                  -1.602145
p-value                          0.482516
#Lags Used                      26.000000
Number of Observations Used   1892.000000
Critical Value (1%)             -3.433811
Critical Value (5%)             -2.863069
Critical Value (10%)            -2.567584
```

Non-stationary

p-value = 0.48 > 0.05
Not statistically significant

Cannot reject null hypothesis of non-stationary

KPSS

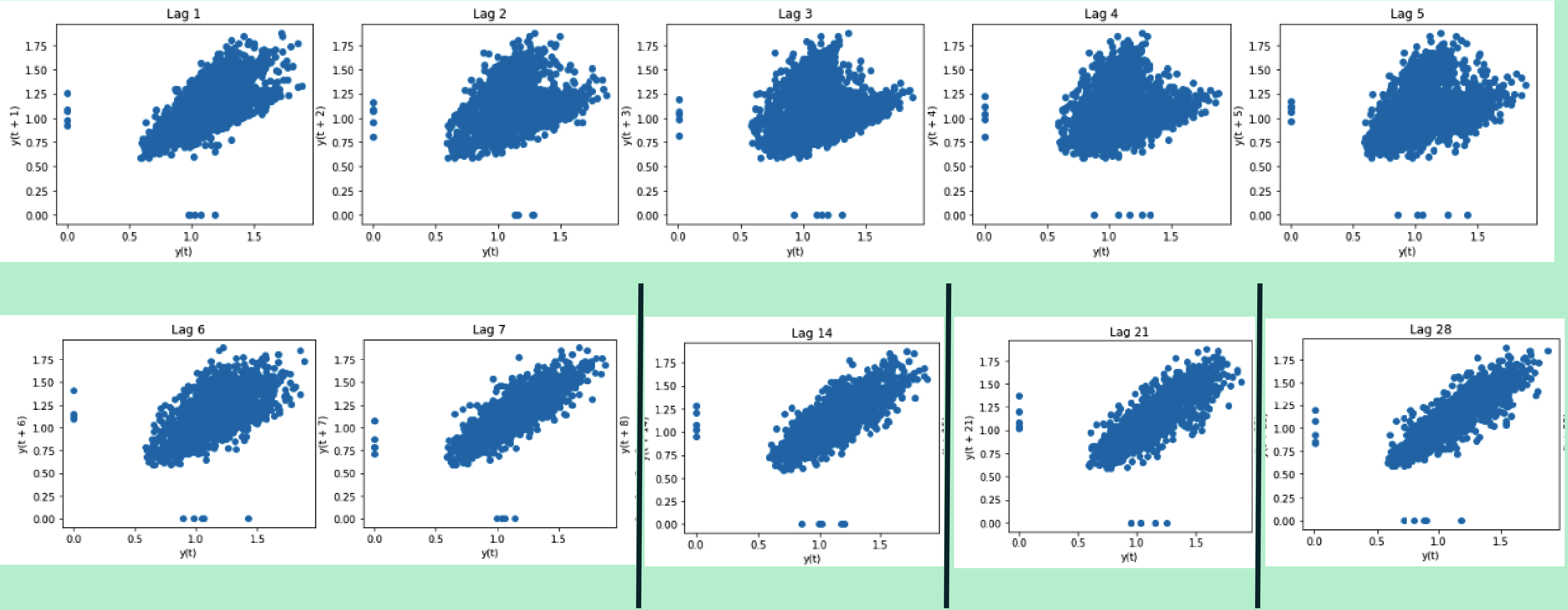
```
Results of KPSS Test:
Test Statistic            6.443358
p-value                   0.010000
Lags Used                20.000000
Critical Value (10%)      0.347000
Critical Value (5%)       0.463000
Critical Value (2.5%)     0.574000
Critical Value (1%)       0.739000
```

Non-stationary

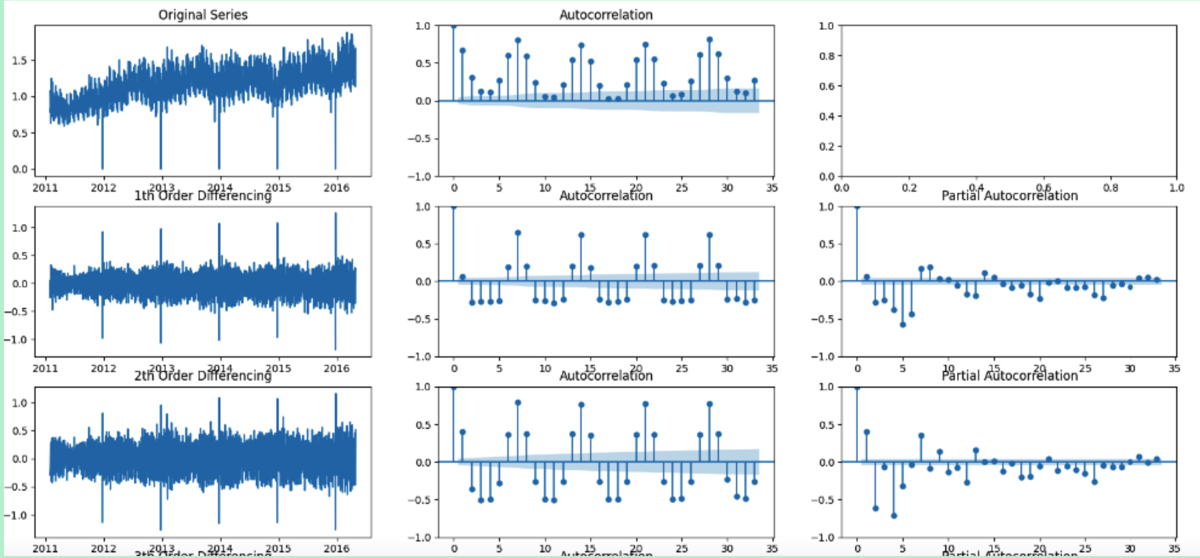
p-value = 0.01 < 0.05
Statistically significant

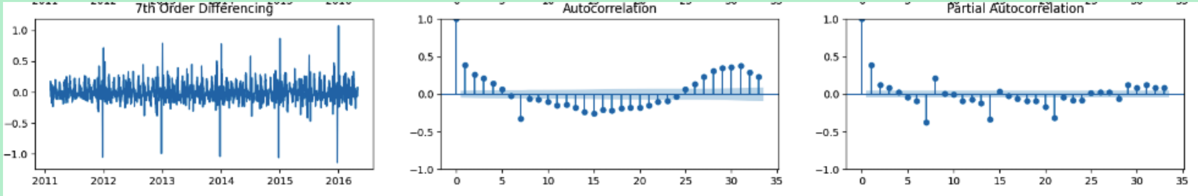
Reject null hypothesis of stationary

Stationarity – Lag Plots



Stationarity – ACF & PACF





7-Day Seasonality

Proceed with Differencing

Stationarity – Differenced ADF v KPSS

ADF

Results of Dickey-Fuller Test:	
Test Statistic	-20.958364
p-value	0.000000
#Lags Used	23.000000
Number of Observations Used	1888.000000
Critical Value (1%)	-3.433818
Critical Value (5%)	-2.863072
Critical Value (10%)	-2.567586

Stationary

p-value = 0.00 < 0.05
Statistically significant

Reject null hypothesis of non-stationary

KPSS

Results of KPSS Test:	
Test Statistic	0.008453
p-value	0.100000
Lags Used	19.000000
Critical Value (10%)	0.347000
Critical Value (5%)	0.463000
Critical Value (2.5%)	0.574000
Critical Value (1%)	0.739000

Stationary

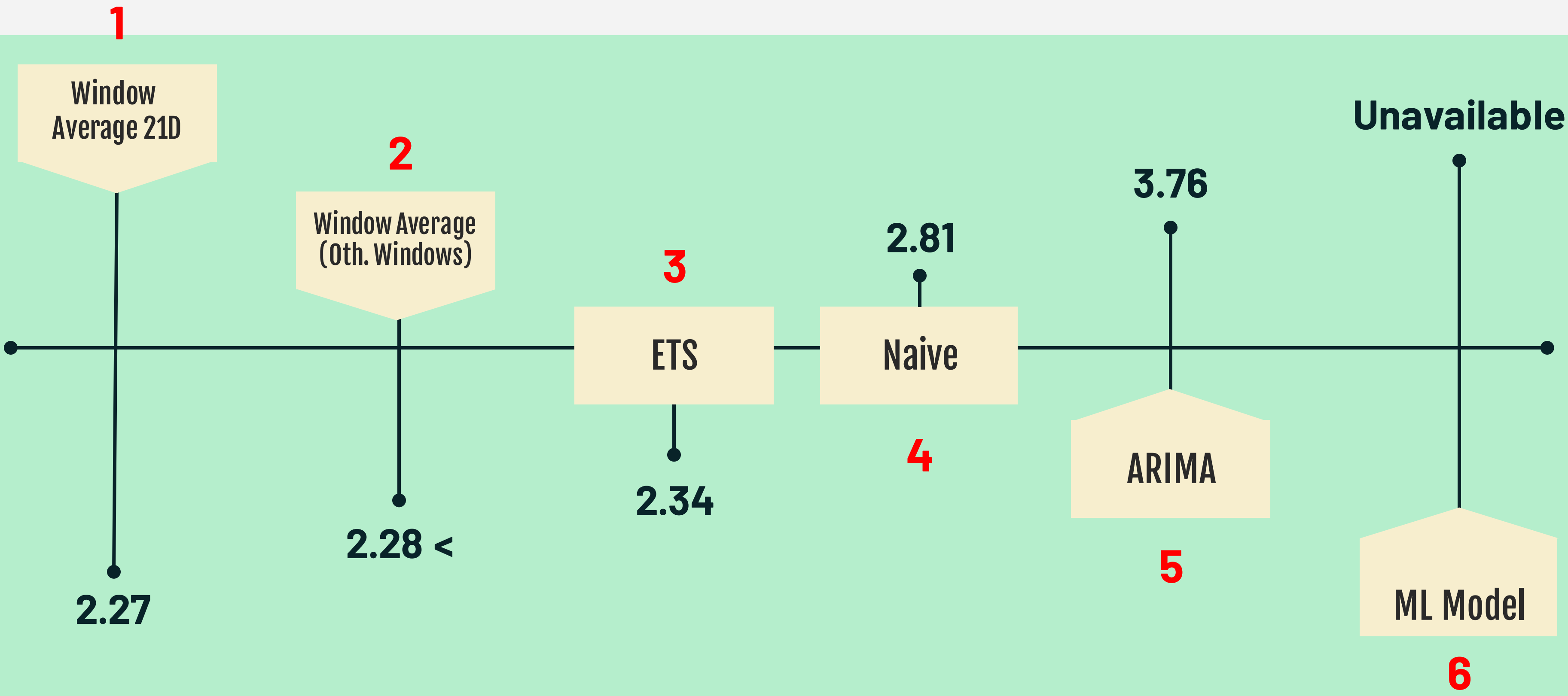
p-value = 0.10 > 0.05
Not statistically significant

Cannot reject null hypothesis of stationary

03

Best Approach

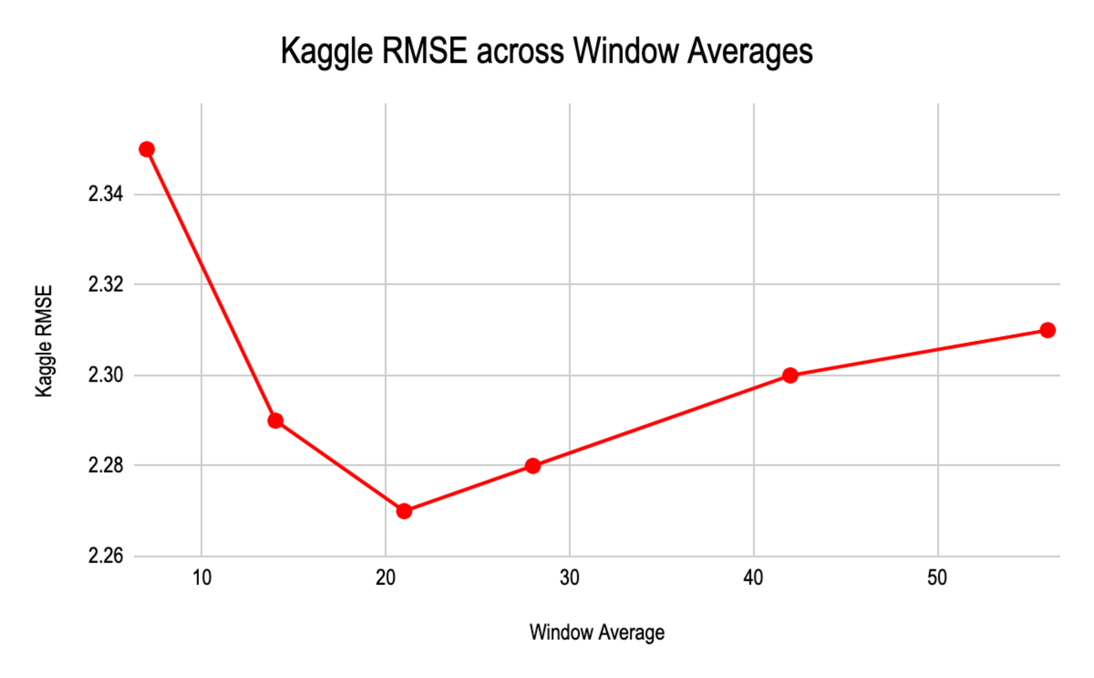
Modelling Overview



Selection of Window Average

	RMSE
Naive Model	4.61
Seasonal Naive Model	4.80
Window Average Model	4.49
Seasonal Window Average Model	4.70

Optimizing Lengths for Window Average



In order to expedite the fine tuning process, created our own 80-20 train test split by taking all days before 29 Jan 2015 as the training set and days after as the test set.

04

Alternative Modelling

Error Trend Seasonality Model

Assessed RMSE on Kaggle for 3 ETS methods

Best performing model

Holt-Winters' Seasonality model; ETS(A,A,A)

3rd Place in Score

(S)ARIMA

We used the "from pmdarima.arima import auto_arima"
which optimizes given an information criterion and provides the most optimal order of difference(d), and hyper-parameters,(p) and (q).

Since we detected seasonality, we attempted SARIMA as well

4th Place in Score

ARIMA

We first applied ARIMA on the average units sold
Then, we aggregated per category; found new & better predictions
Next, we repeated the last step for subcategories
Finally, we did the same for stores
With each new iteration, we found better and better result

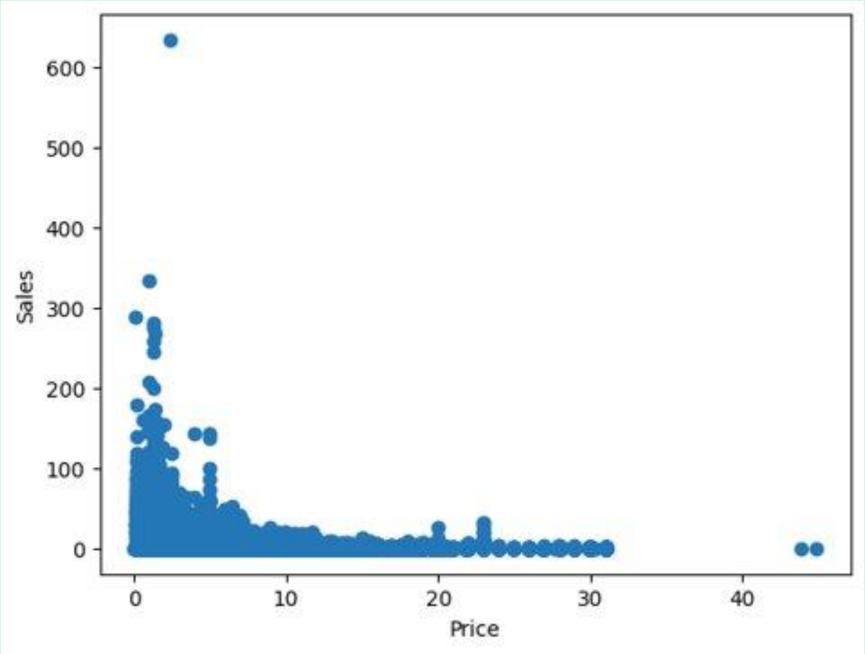
4th Place in Score

ML Approach – Overview

The technical process for ML predictions was to

1) Restructure train.csv to long form data
2) Select 1% of records randomly(462K datapoints)
2) Manipulate the data to introduce new time-based and price features
3) Fit an LGBM model

ML Approach – Negative Correlation of Price and Sales



Price should be included as a feature in our model

Corr = -0.15

ML Approach – Feature Engineering

Date features such as day of the year, week of the month etc.

Price lags up to 7 days

ML Approach – Fit and Result

Obtain LGBM Model by training on 1% of the data for computational efficiency

Testing on 20%: RMSE of 3.87
Forecasted last 21 days of test set: RMSE of 2.01

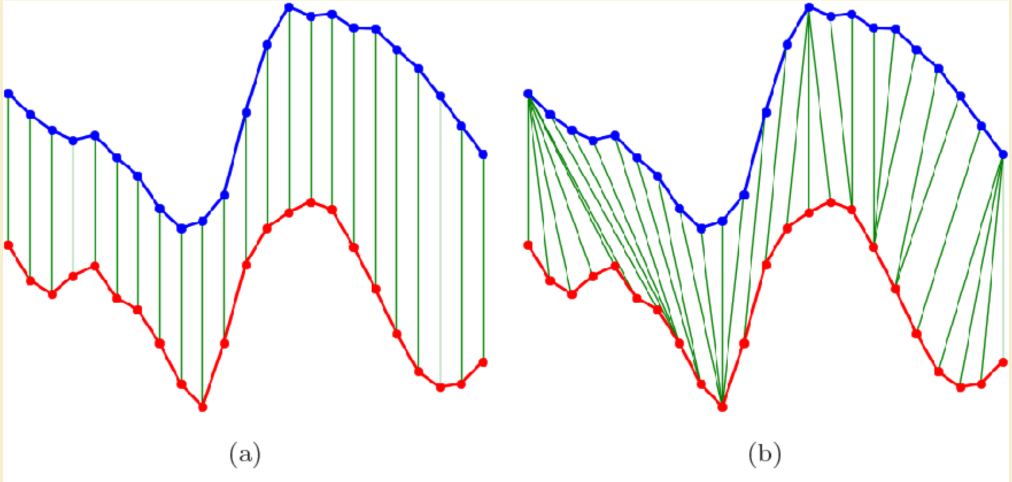
05

Limitations

Dynamic Time Wrapping

Why DTW?
Matching Time Series

Issues faced?
Volume of data



Clustering

Attempting Dynamic Time Wrapping Clustering

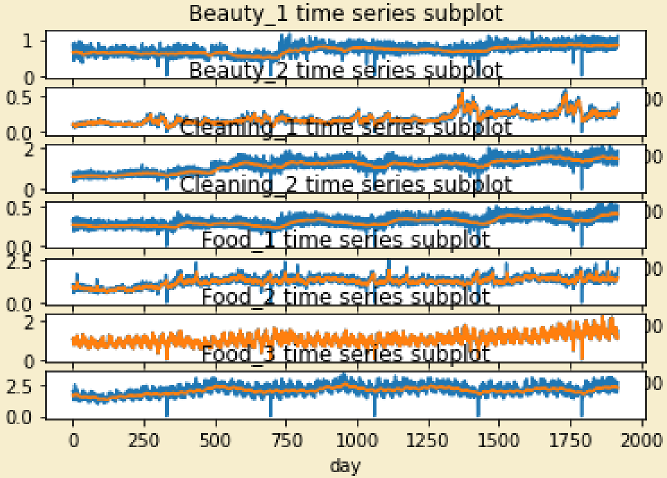
Clustering based on existing labelled categories

Next steps applying hierarchical clustering

Trying ETS performance on labels

Using ETS on subcategories performs well for some categories

Here the prediction is performing well for Beauty_2, Food_1, and Food_2
This is applying Beauty_1 on the rest



Computational Limitations

The absolute size of this data set makes it difficult to conduct comprehensive modelling techniques

ML Specific

Data could be subset by time instead of randomly selected

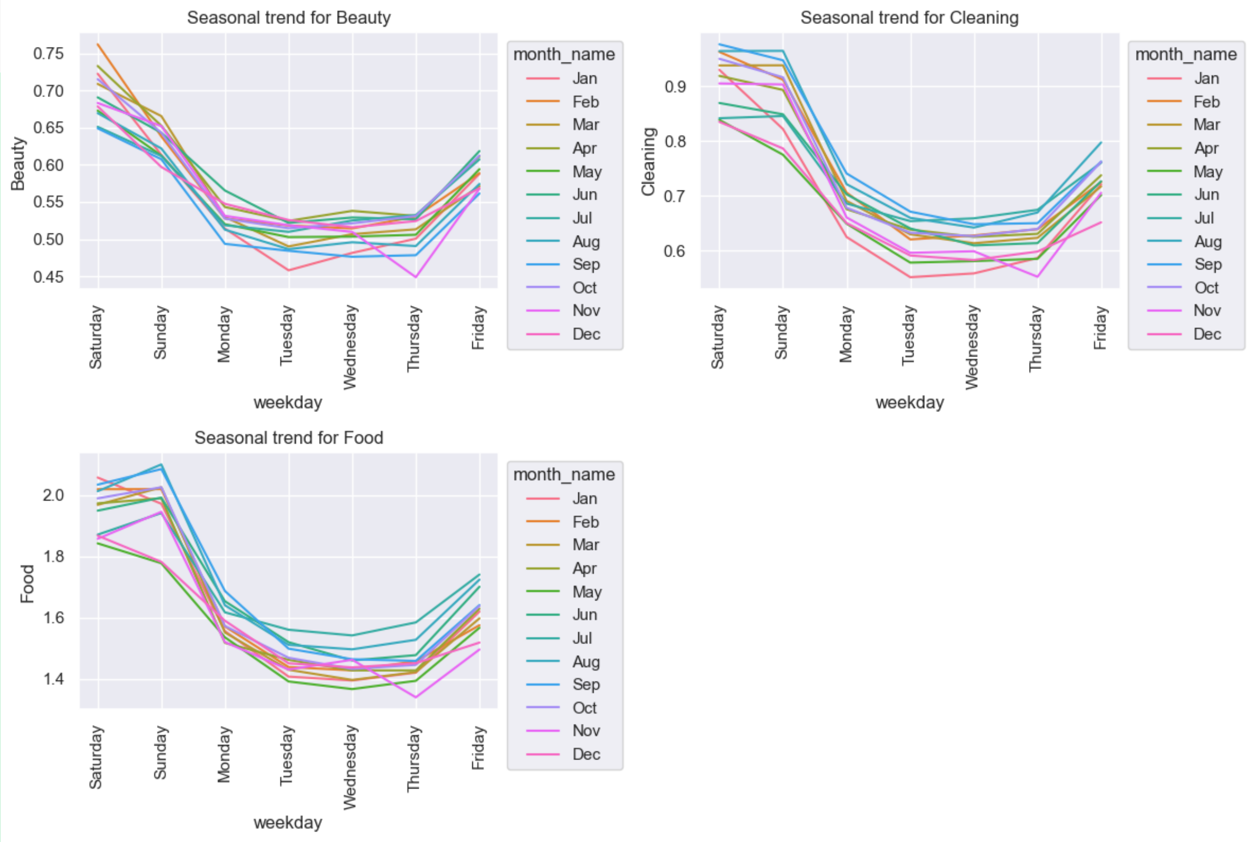
More feature engineering could be done: Price data, time lags and rolling window

Group by different subsets to train separate models (e.g. store, categories etc.)

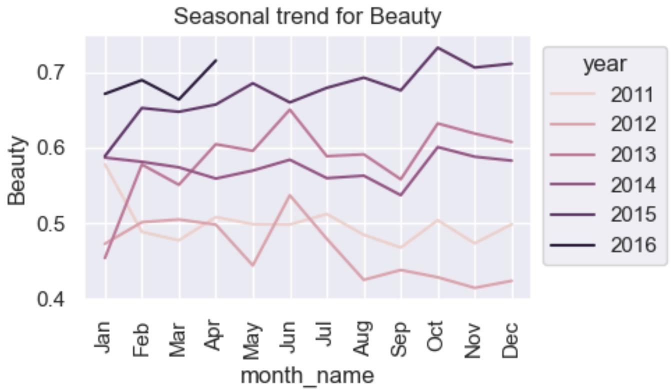
06

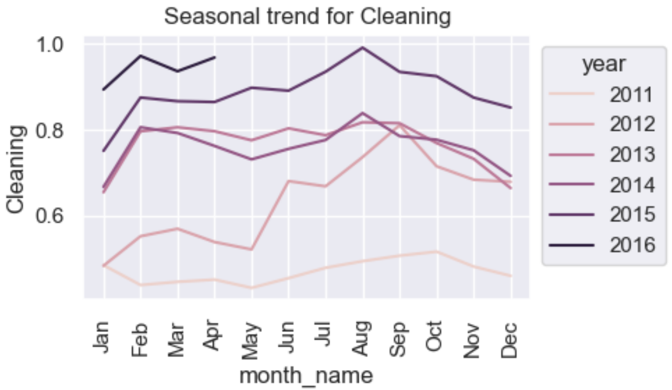
Business Insights

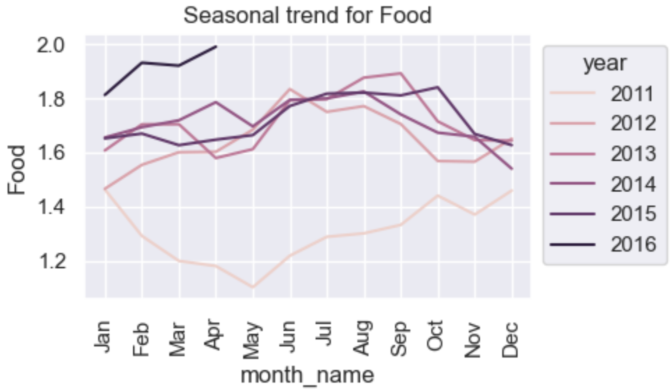
Weekly Trend By Category



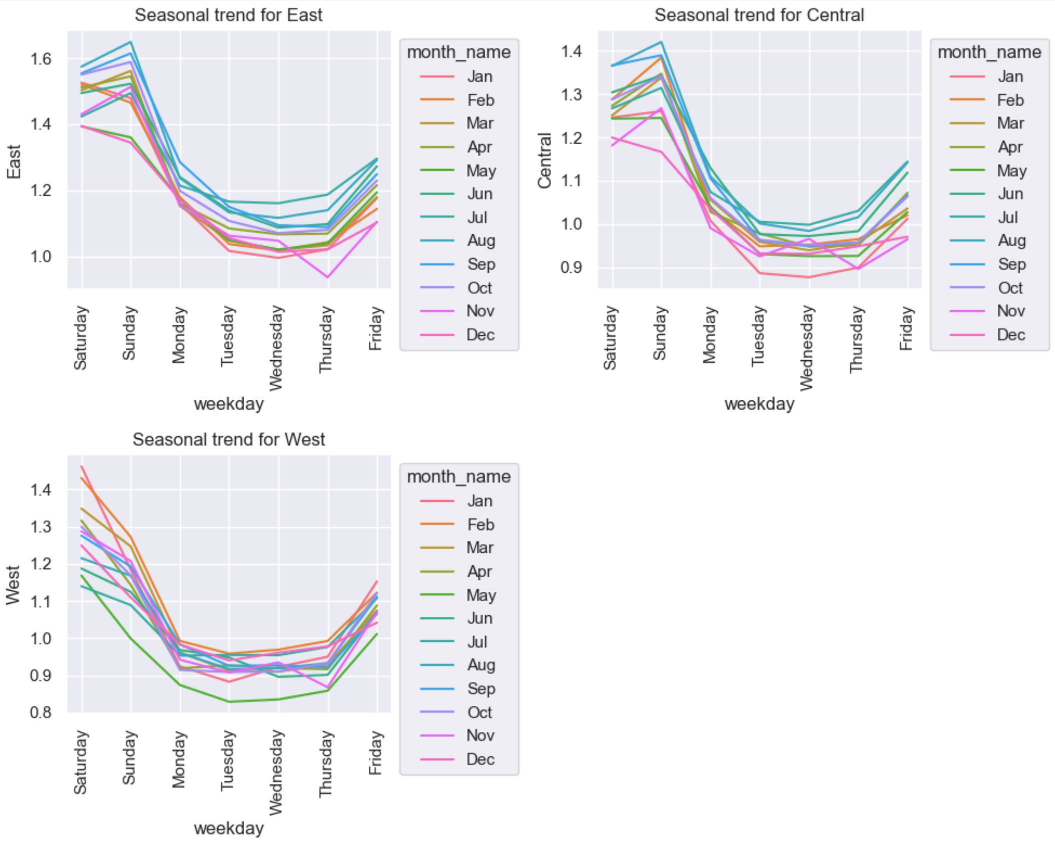
Monthly Trend By Category



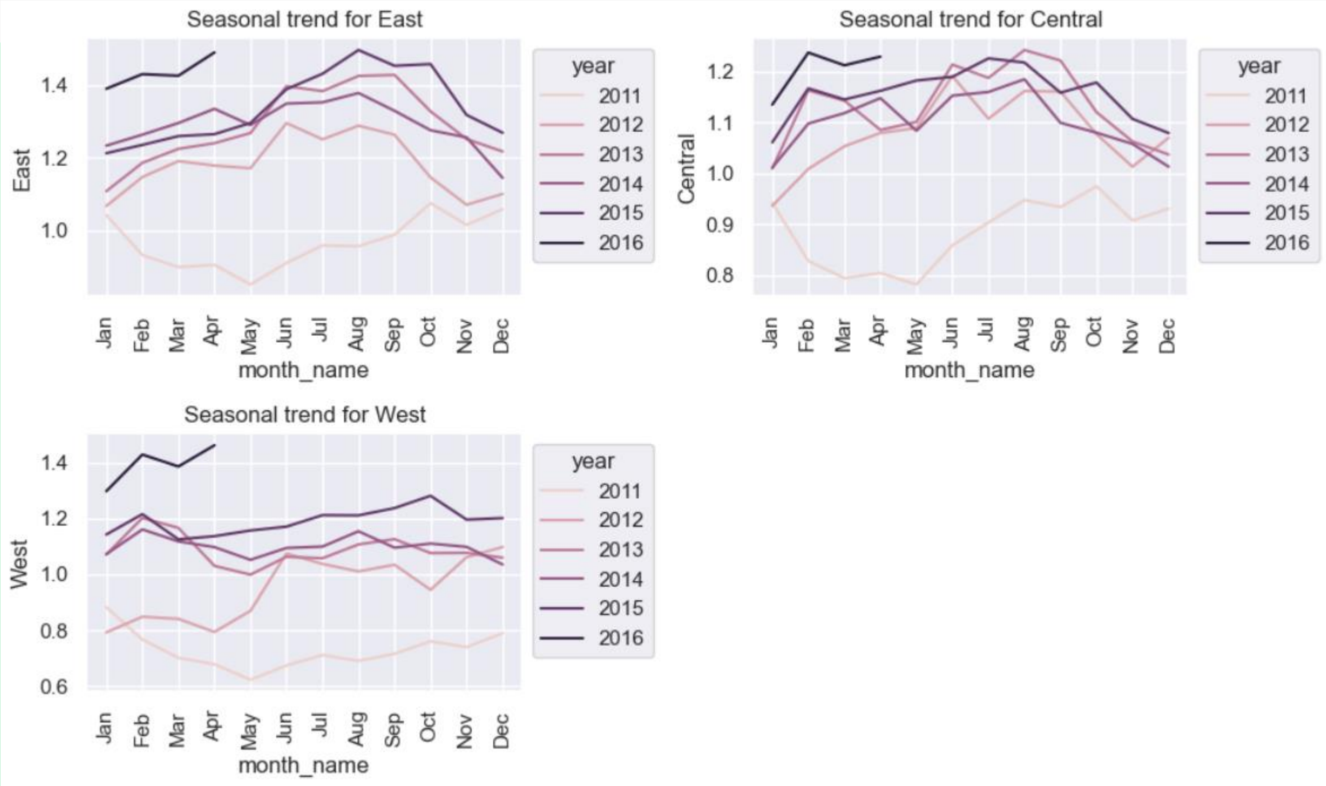




Weekly Trend per Region



Monthly Trend per Region



Insights from Best Method

The Window Average method forecasted a standard output for each item across the 21 days

Stores can either maintain sufficient stocks as predicted, or adjust their pricing strategies according to the forecasted demand to drive orders toward or away accordingly

Scalable and computationally efficient

Demand Forecast

	id	d_1920	d_1921	d_1922	d_1923	d_1924	d_1925	d_1926	d_1927	d_1928	...	d_1936	d_1937	d_1938	d_1939	d_1940
14	Beauty_1_015_East_1	2	2	2	2	2	2	2	2	2	...	2	2	2	2	2
3063	Beauty_1_015_East_2	8	8	8	8	8	8	8	8	8	...	8	8	8	8	8
6112	Beauty_1_015_East_3	6	6	6	6	6	6	6	6	6	...	6	6	6	6	6
9161	Beauty_1_015_East_4	3	3	3	3	3	3	3	3	3	...	3	3	3	3	3
12210	Beauty_1_015_Central_1	3	3	3	3	3	3	3	3	3	...	3	3	3	3	3
15259	Beauty_1_015_Central_2	3	3	3	3	3	3	3	3	3	...	3	3	3	3	3
18308	Beauty_1_015_Central_3	5	5	5	5	5	5	5	5	5	...	5	5	5	5	5
21357	Beauty_1_015_West_1	1	1	1	1	1	1	1	1	1	...	1	1	1	1	1
24406	Beauty_1_015_West_2	1	1	1	1	1	1	1	1	1	...	1	1	1	1	1
27455	Beauty_1_015_West_3	3	3	3	3	3	3	3	3	3	...	3	3	3	3	3

Utility of Demand Forecast

Remain competitive and attract more customers where needed

Improve store operations and customer experience ->
Increased customer loyalty

Providing the right products and services at the right time

Thank you!