Data Science Internship at Data Glacier

Project: Retail Forecasting

Project Report

Group name: Red Tail Force

Group members: Gordon Poon, Pok Hei Tang (Keith), Joseph Xu

Email: gordontxpoon@gmail.com, zx1054@nyu.edu, keithtang0901@gmail.com

Country: United Kingdom, China

College: UCL, NYU, Durham University

Specialisation: Data Science

Table of Contents:

1. Project description

Dataset was provided by a large beverage company in Australia. They sell their products through various super-markets and also engage into heavy promotions throughout the year. Their demand is also influenced by various factors like holiday, seasonality. They needed a forecast of each of the products at item level every week in weekly buckets.

2. Business Understanding

Determine Business Objectives:

1. Forecast the item level of 6 products at each week

Assess Situation (Assumptions):

1. Relationship exists between sales and holidays
2. Relationship exists between sales and promotion
3. Relationship exists between sales and Covid
4. Relationship exists between sales and Google Mobility
5. Relationship exists between sales and discount
6. Relationship exists between sales of one product and another

Determine Data Science Goals:

1. Build 4-5 multivariate forecasting model
2. Demonstrate best in class forecast accuracy

3. Write a code in such a way in order to run the model in least time
4. Demonstrate explainability in the form of contribution of each variables

Project Plan:

1. Week 7: Business understanding
2. Week 8: Data Understanding
3. Week 9: Data Cleaning and preparation
4. Week 10: EDA
5. Week 11: EDA presentation and proposed modelling technique
6. Week 12: Model selection and model building

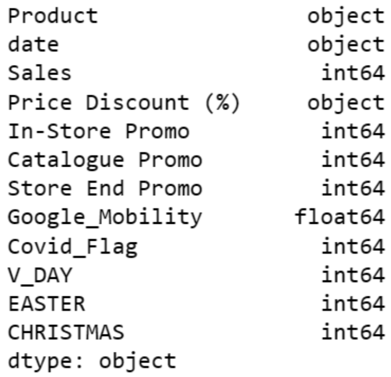
3. Data Understanding

Describe data:

Total number of observations	1218
Total number of features	12
Base format of the file	.xlsx
Size of the data	74 kB

Explore data and check data quality:

a. Check Data types:

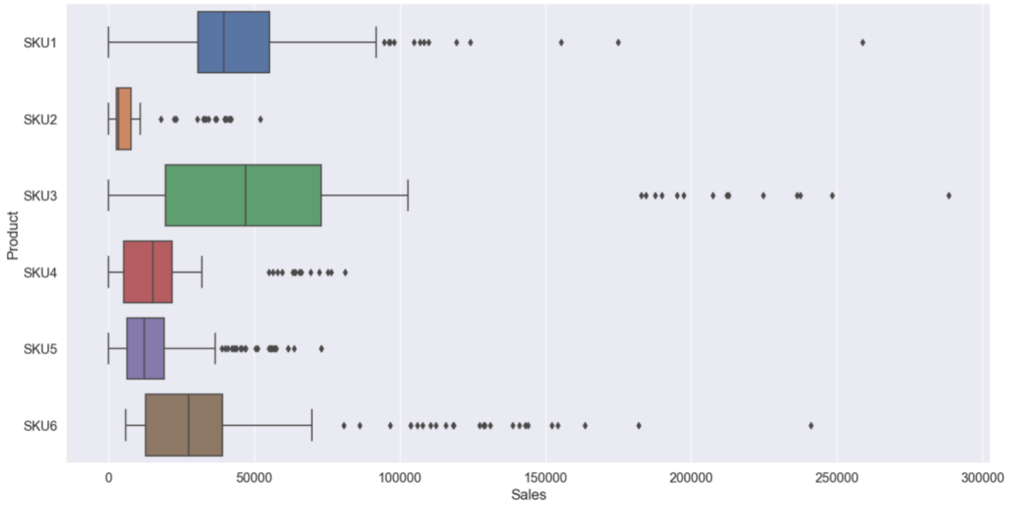


```
Product               object
date                  object
Sales                  int64
Price Discount (%)    object
In-Store Promo         int64
Catalogue Promo        int64
Store End Promo        int64
Google_Mobility      float64
Covid_Flag             int64
V_DAY                  int64
EASTER                 int64
CHRISTMAS              int64
dtype: object
```

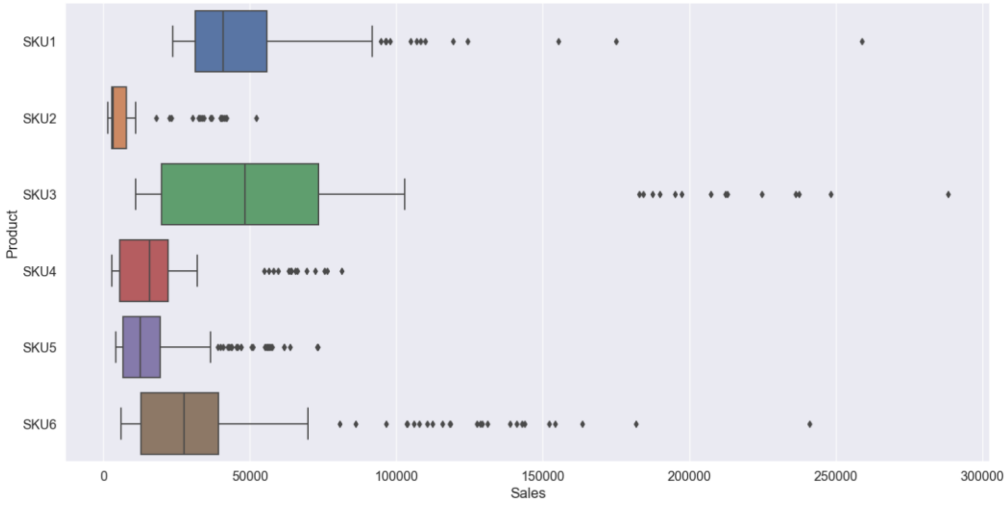
The data types of ‘date’ and ‘Price Discount (%)’ should be modified.

b. Check Missing values:
There is no missing values in the dataset

c. Check Outliers



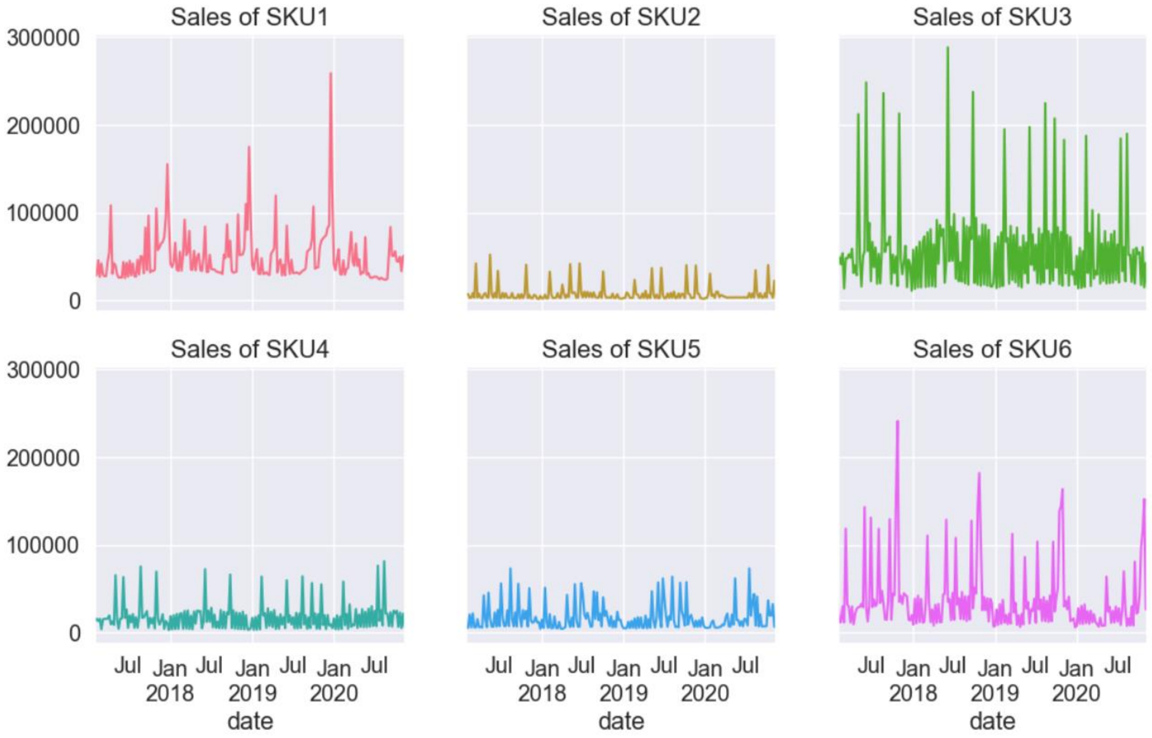
There are weeks with zero Sales in SKU1-5. It is believed that the testing weeks were mixed into the dataset.



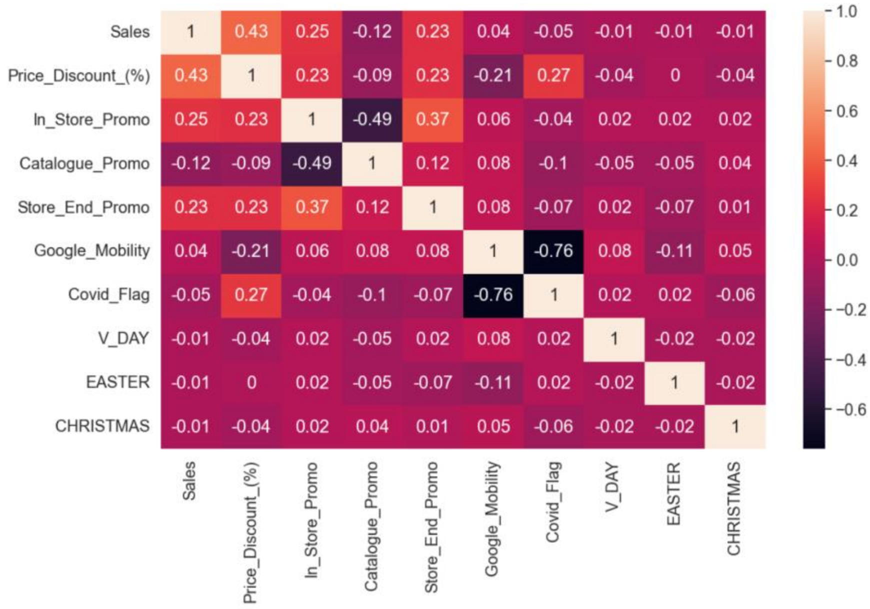
After modifying the data, there are still some outliers in the Sales feature. But since we do not have enough information on the components for the sales, it is not appropriate to treat it as an outlier.

4. Exploratory Data Analysis

4.1 Sales against time



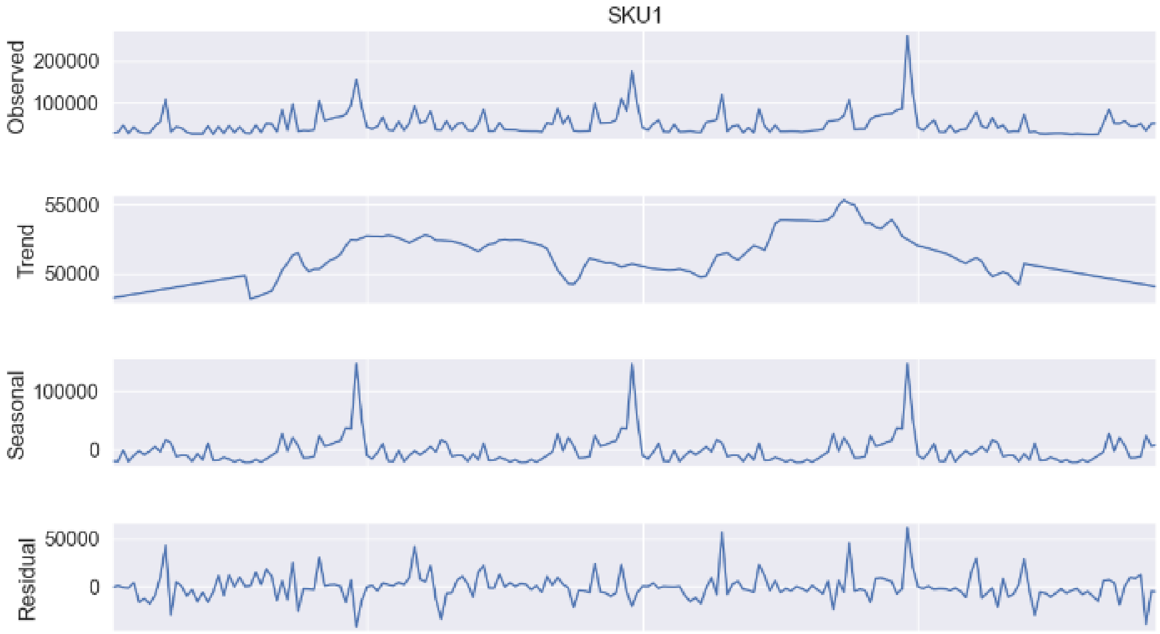
4.2 Correlation of features

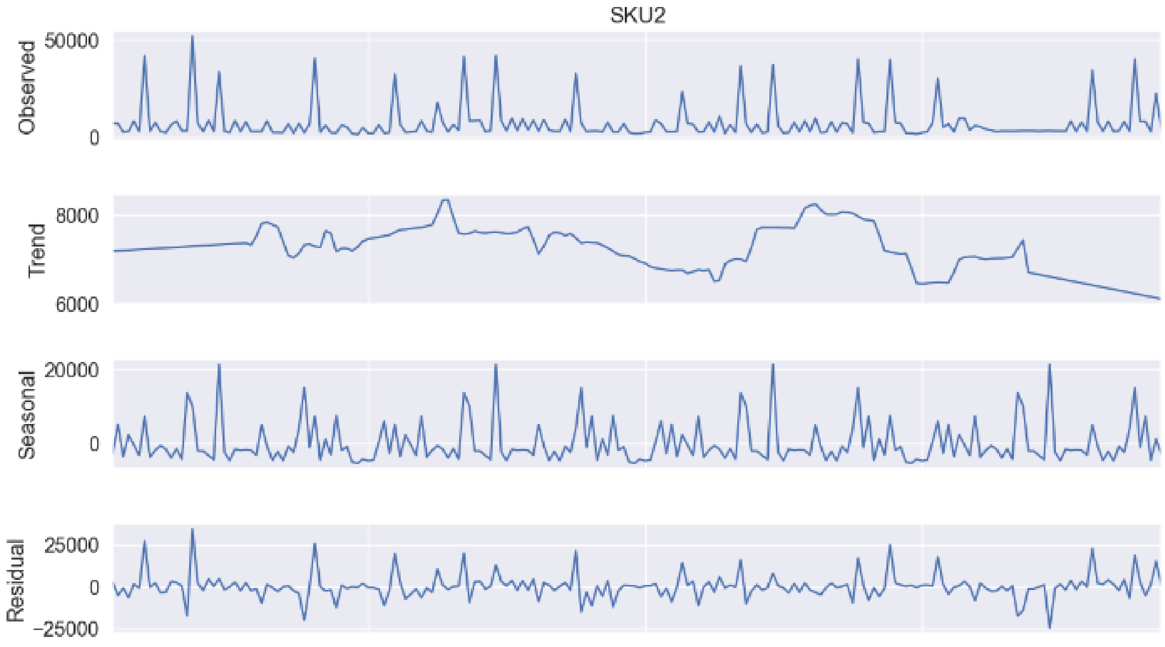


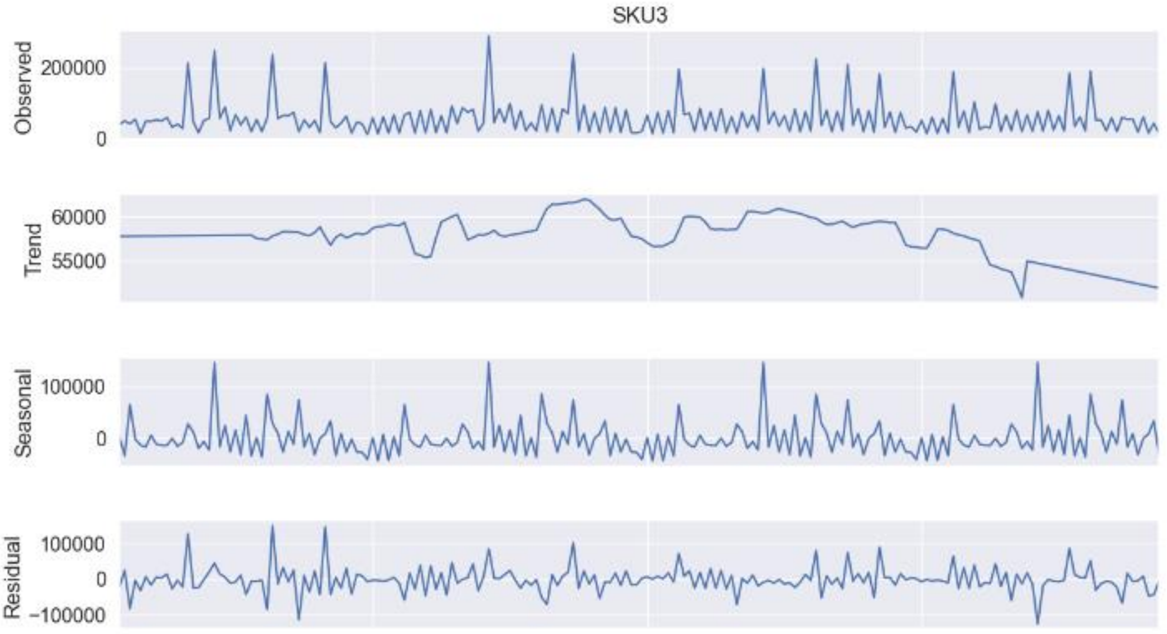
	Sales	Price_Discount_(%)	In_Store_Promo	Catalogue_Promo	Store_End_Promo	Google_Mobility	Covid_Flag	V_DAY	EASTER	CHRISTMAS
Sales	1	0.43	0.25	-0.12	0.23	0.04	-0.05	-0.01	-0.01	-0.01
Price_Discount_(%)	0.43	1	0.23	-0.09	0.23	-0.21	0.27	-0.04	0	-0.04
In_Store_Promo	0.25	0.23	1	-0.49	0.37	0.06	-0.04	0.02	0.02	0.02
Catalogue_Promo	-0.12	-0.09	-0.49	1	0.12	0.08	-0.1	-0.05	-0.05	0.04
Store_End_Promo	0.23	0.23	0.37	0.12	1	0.08	-0.07	0.02	-0.07	0.01
Google_Mobility	0.04	-0.21	0.06	0.08	0.08	1	-0.76	0.08	-0.11	0.05
Covid_Flag	-0.05	0.27	-0.04	-0.1	-0.07	-0.76	1	0.02	0.02	-0.06
V_DAY	-0.01	-0.04	0.02	-0.05	0.02	0.08	0.02	1	-0.02	-0.02
EASTER	-0.01	0	0.02	-0.05	-0.07	-0.11	0.02	-0.02	1	-0.02
CHRISTMAS	-0.01	-0.04	0.02	0.04	0.01	0.05	-0.06	-0.02	-0.02	1

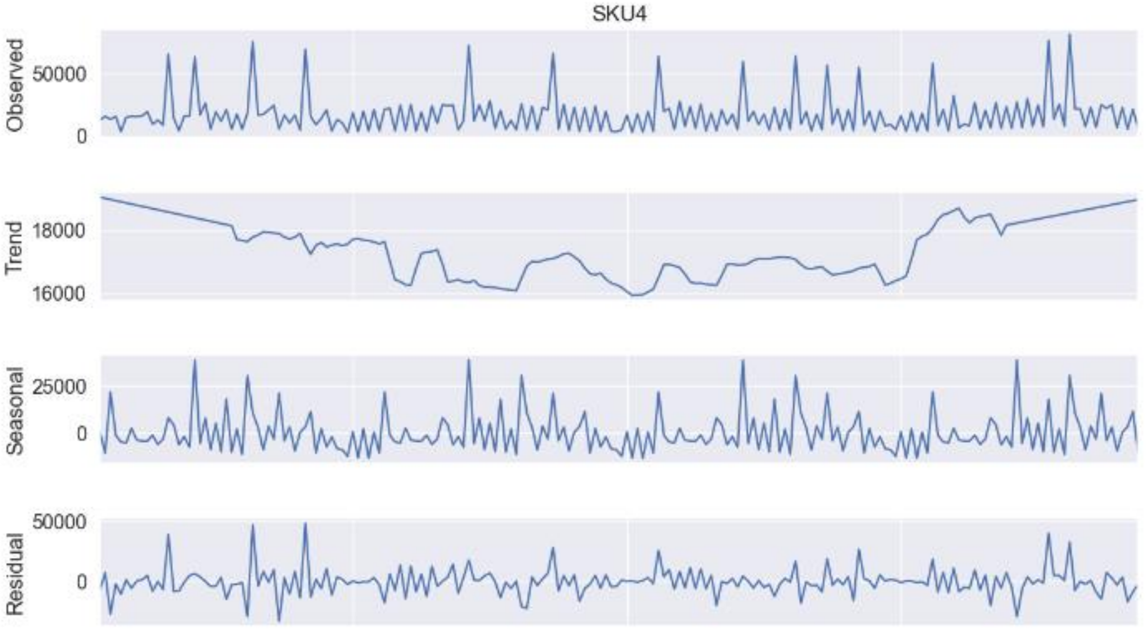
From the heatmap, we see that **Google_Mobility** has an inverse correlation with **Covid_Flag**. **In_Store_Promo** is slightly correlated with **Catalogue_Promo** and also for **Sales** and **Price_Discount_(%).**

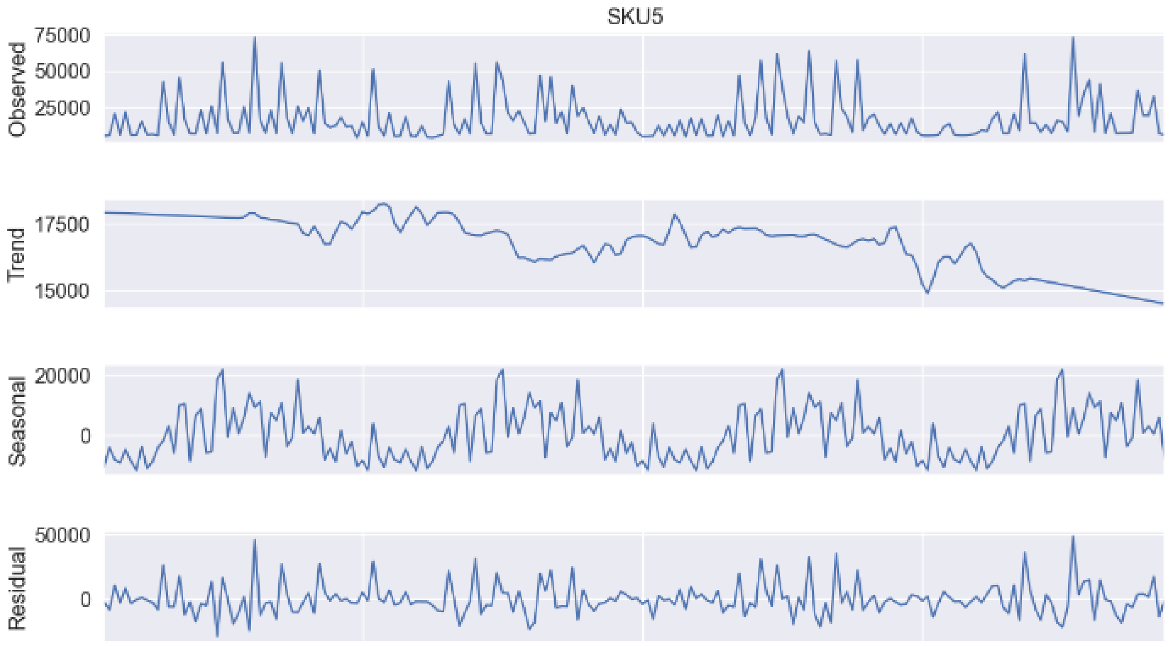
4.3 Trend and Seasonality

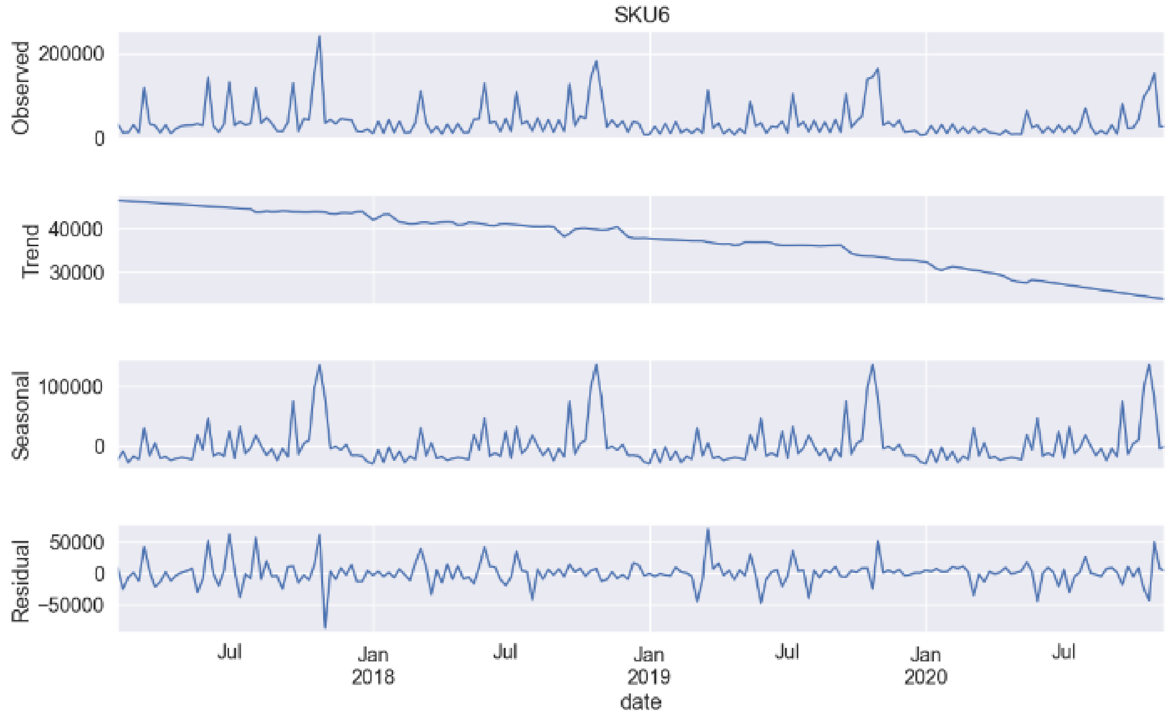




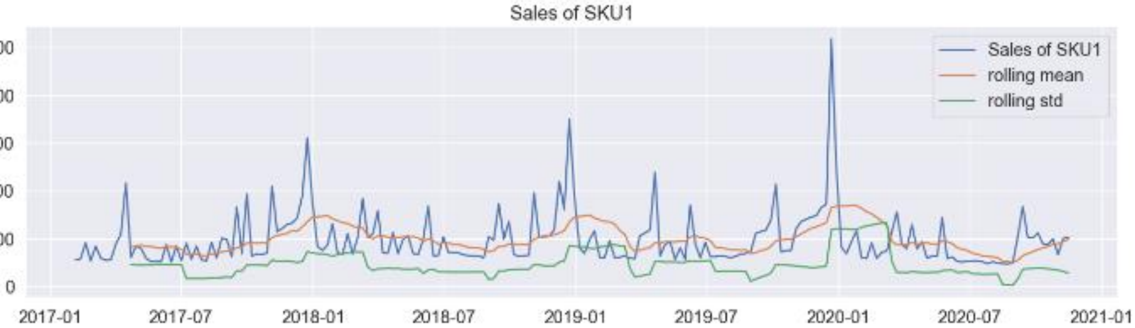

SKU3
Observed
200000
0
Trend
60000
55000
Seasonal
100000
0
Residual
100000
0
−100000

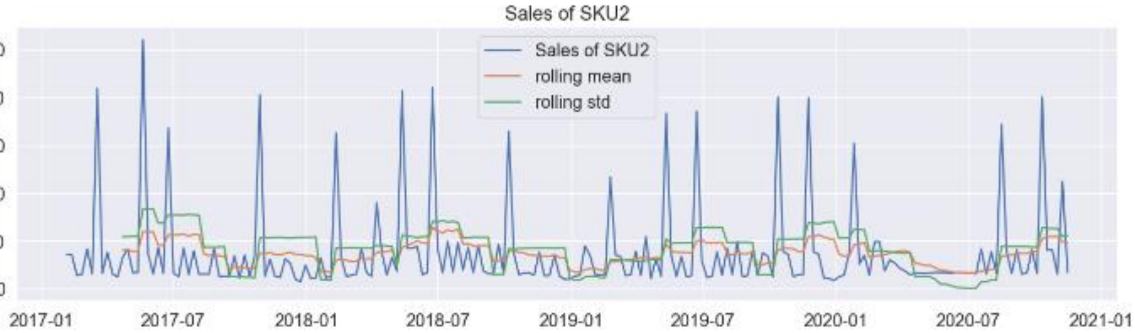

SKU4
Observed
50000
0
Trend
18000
16000
Seasonal
25000
0
Residual
50000
0

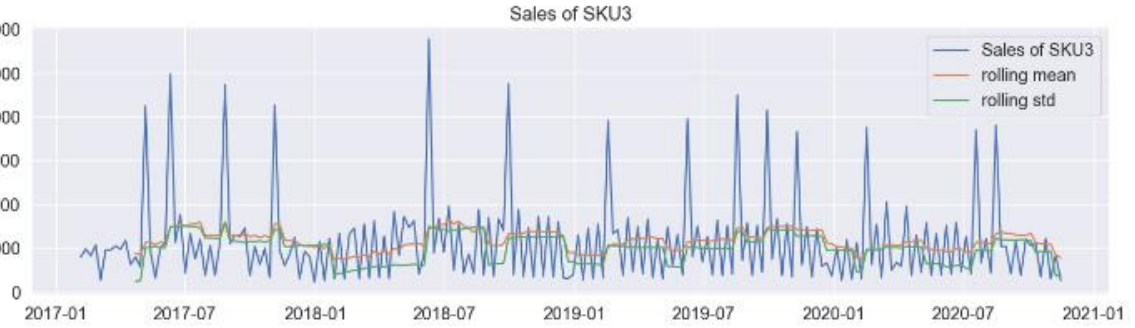


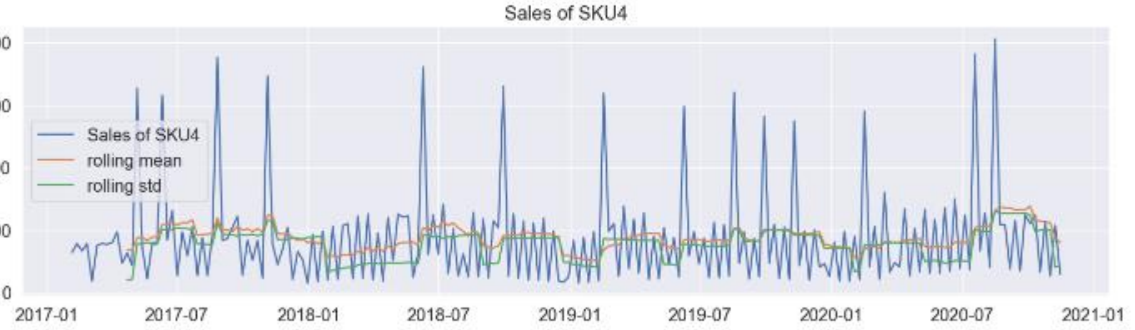


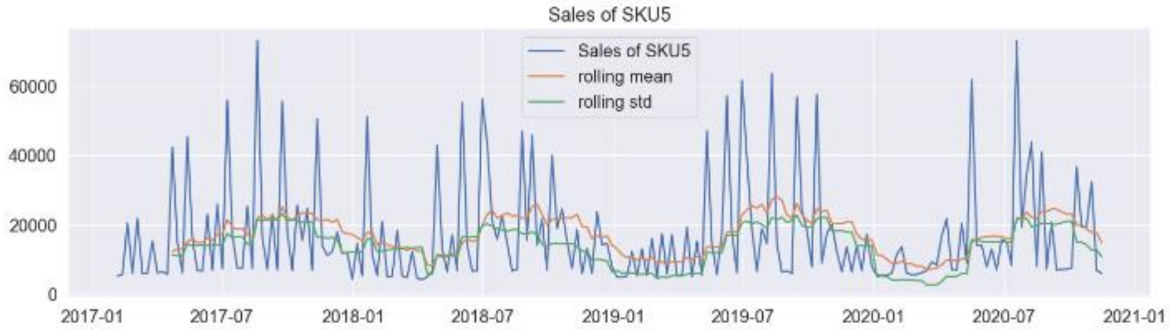
4.4 Stationarity

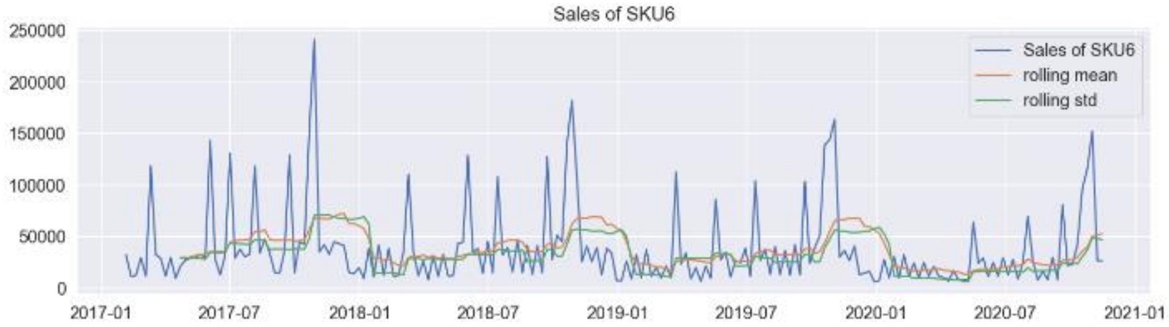

Sales of SKU1
Sales of SKU1
rolling mean
rolling std
250000
200000
150000
100000
50000
0
2017-01
2017-07
2018-01
2018-07
2019-01
2019-07
2020-01
2020-07
2021-01


Sales of SKU2
Sales of SKU2
rolling mean
rolling std
50000
40000
30000
20000
10000
0
2017-01
2017-07
2018-01
2018-07
2019-01
2019-07
2020-01
2020-07
2021-01

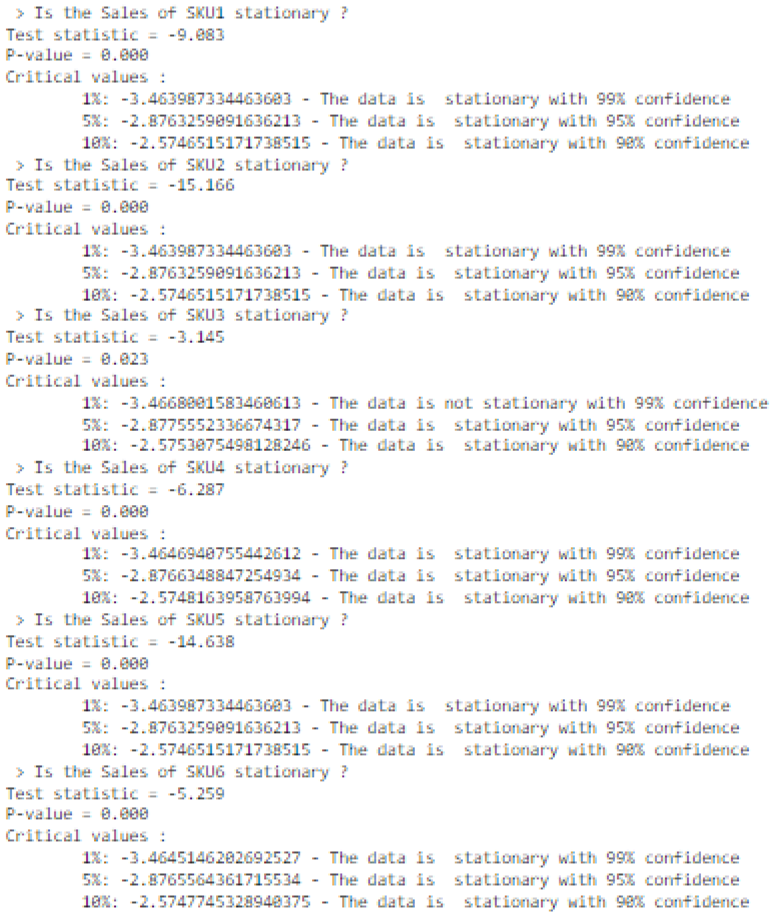

Sales of SKU3
Sales of SKU3
rolling mean
rolling std
300000
250000
200000
150000
100000
50000
0
2017-01
2017-07
2018-01
2018-07
2019-01
2019-07
2020-01
2020-07
2021-01


Sales of SKU4
Sales of SKU4
rolling mean
rolling std
80000
60000
40000
20000
0
2017-01
2017-07
2018-01
2018-07
2019-01
2019-07
2020-01
2020-07
2021-01



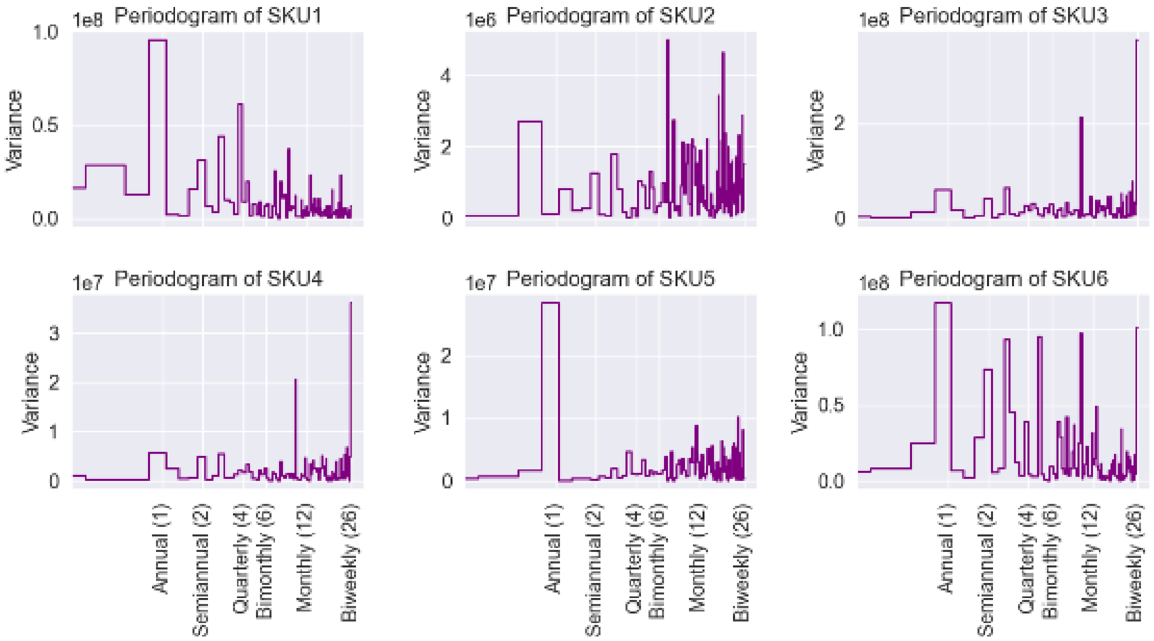


Augmented Dickey-Fuller Test:



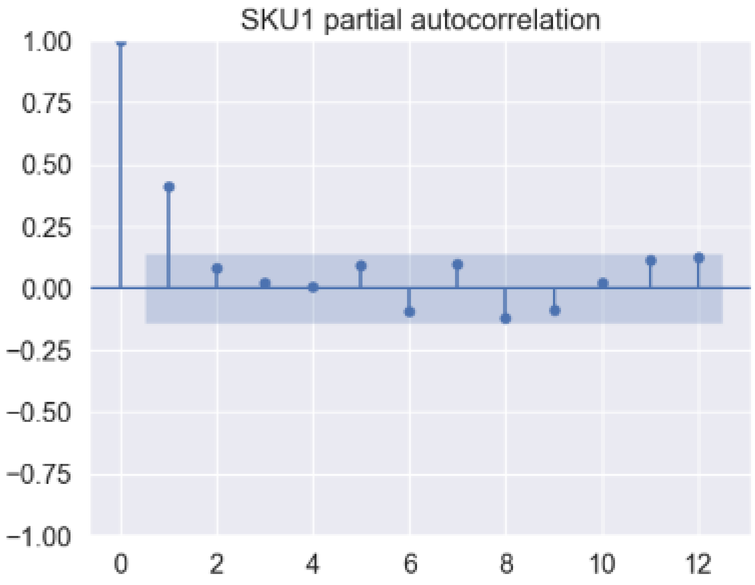
```
 > Is the Sales of SKU1 stationary ?
Test statistic = -9.083
P-value = 0.000
Critical values :
        1%: -3.463987334463603 - The data is  stationary with 99% confidence
        5%: -2.8763259091636213 - The data is  stationary with 95% confidence
        10%: -2.5746515171738515 - The data is  stationary with 90% confidence
 > Is the Sales of SKU2 stationary ?
Test statistic = -15.166
P-value = 0.000
Critical values :
        1%: -3.463987334463603 - The data is  stationary with 99% confidence
        5%: -2.8763259091636213 - The data is  stationary with 95% confidence
        10%: -2.5746515171738515 - The data is  stationary with 90% confidence
 > Is the Sales of SKU3 stationary ?
Test statistic = -3.145
P-value = 0.023
Critical values :
        1%: -3.4668001583460613 - The data is not stationary with 99% confidence
        5%: -2.8775552336674317 - The data is  stationary with 95% confidence
        10%: -2.5753075498128246 - The data is  stationary with 90% confidence
 > Is the Sales of SKU4 stationary ?
Test statistic = -6.287
P-value = 0.000
Critical values :
        1%: -3.4646940755442612 - The data is  stationary with 99% confidence
        5%: -2.8766348847254934 - The data is  stationary with 95% confidence
        10%: -2.5748163958763994 - The data is  stationary with 90% confidence
 > Is the Sales of SKU5 stationary ?
Test statistic = -14.638
P-value = 0.000
Critical values :
        1%: -3.463987334463603 - The data is  stationary with 99% confidence
        5%: -2.8763259091636213 - The data is  stationary with 95% confidence
        10%: -2.5746515171738515 - The data is  stationary with 90% confidence
 > Is the Sales of SKU6 stationary ?
Test statistic = -5.259
P-value = 0.000
Critical values :
        1%: -3.4645146202692527 - The data is  stationary with 99% confidence
        5%: -2.8765564361715534 - The data is  stationary with 95% confidence
        10%: -2.5747745328940375 - The data is  stationary with 90% confidence
```

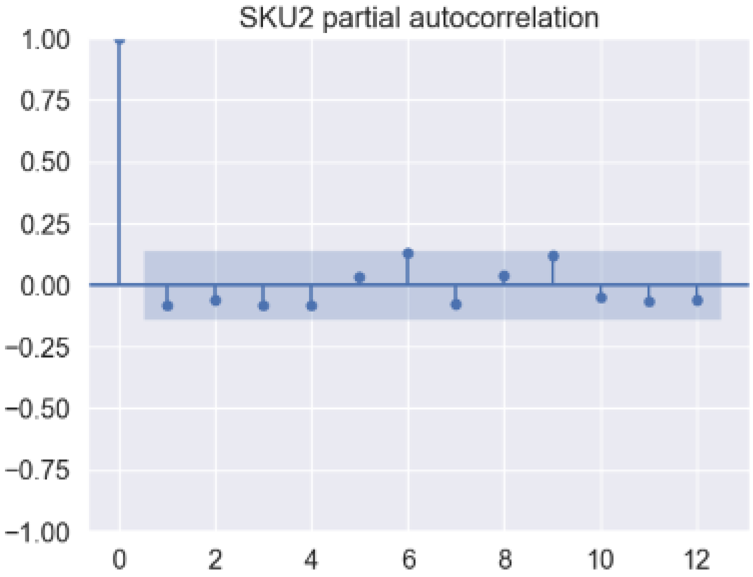
4.5 Periodogram

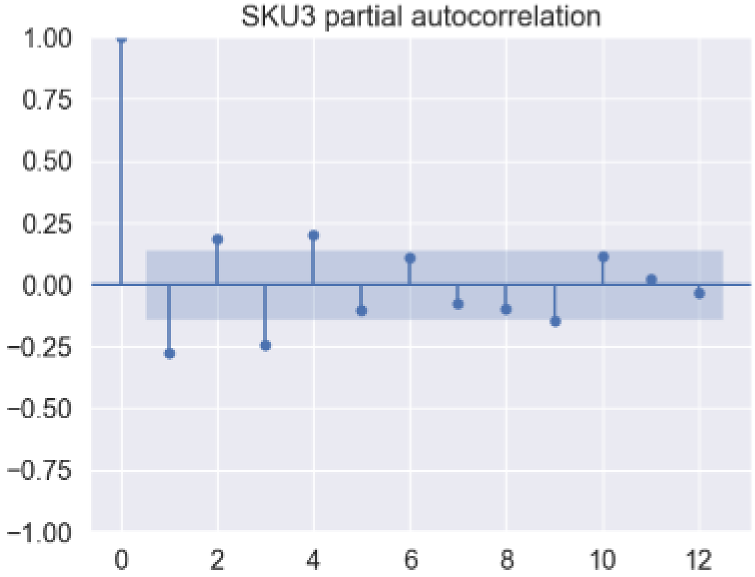


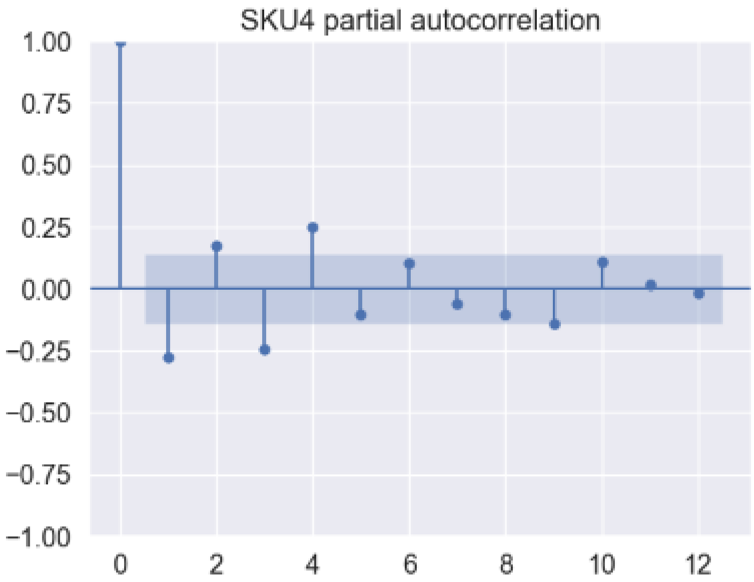
5. Feature engineering and transformation

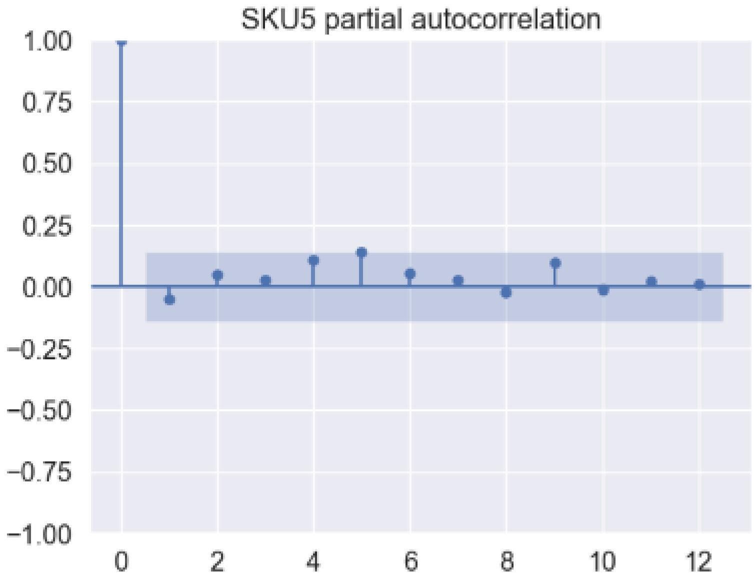
5.1 Partial Autocorrelaion

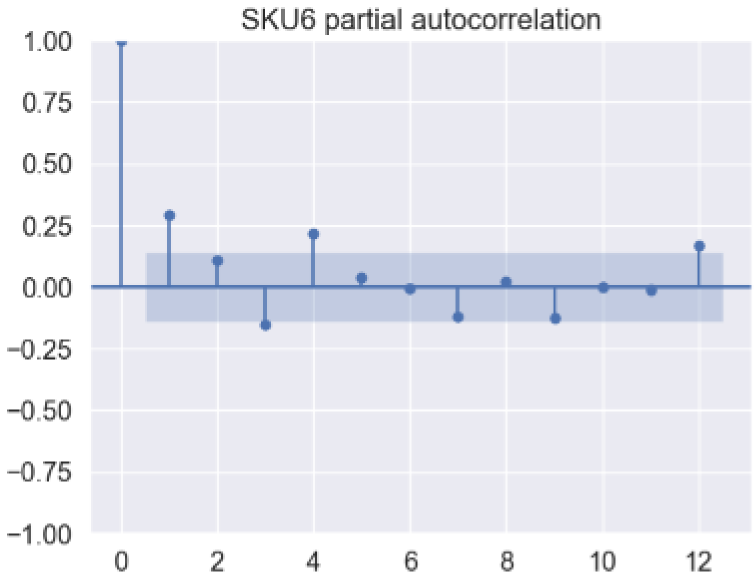



SKU2 partial autocorrelation


SKU3 partial autocorrelation


SKU4 partial autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
2
4
6
8
10
12

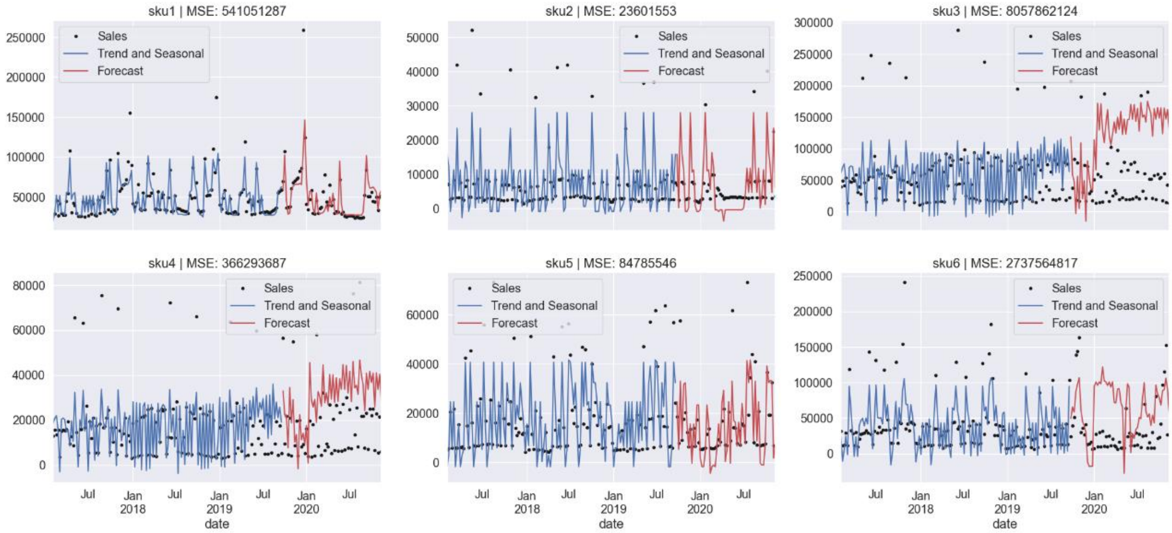

SKU5 partial autocorrelation
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
0
2
4
6
8
10
12



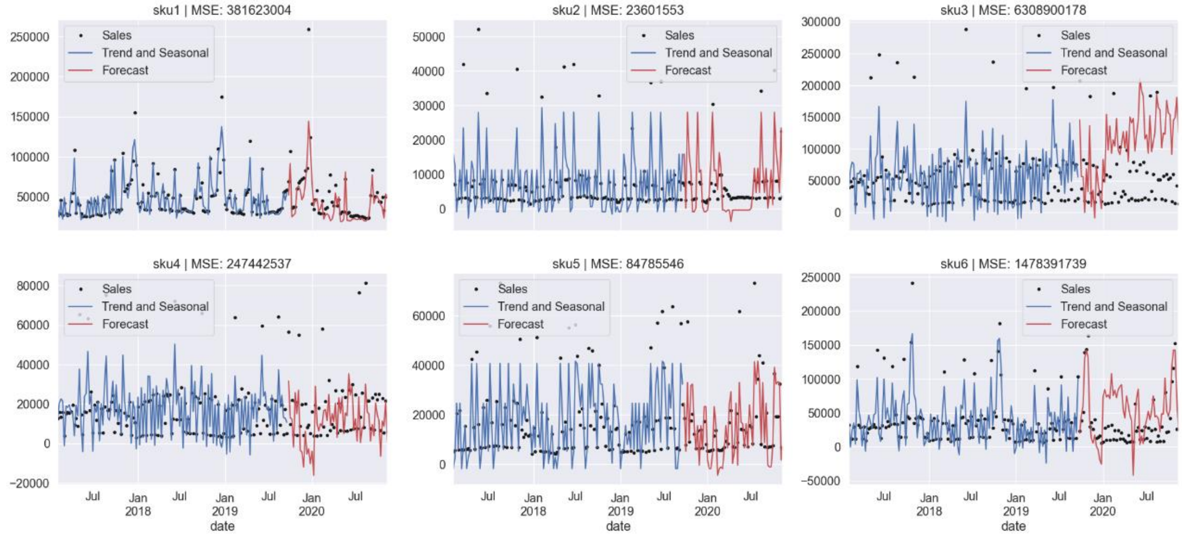
From the partial autocorrelation plot, we added lag 1 time series to SKU1 and SKU6 and lag 1 to lag 4 time series to SKU3 and SKU4.

6. Model training

We used linear regression to train the data and mean square error to evaluate the result



After tuning the hyper parameters, we improve the errors:


sku1 | MSE: 381623004
sku2 | MSE: 23601553
sku3 | MSE: 6308900178
sku4 | MSE: 247442537
sku5 | MSE: 84785546
sku6 | MSE: 1478391739
Sales
Trend and Seasonal
Forecast
250000
200000
150000
100000
50000
50000
40000
30000
20000
10000
0
300000
250000
200000
150000
100000
50000
0
80000
60000
40000
20000
0
−20000
60000
40000
20000
0
250000
200000
150000
100000
50000
0
−50000
Jul
Jan
2018
Jul
Jan
2019
Jul
Jan
2020
Jul
date