Data Science Internship at Data Glacier

Project: Retail Forecasting

Week 9: Deliverables

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:

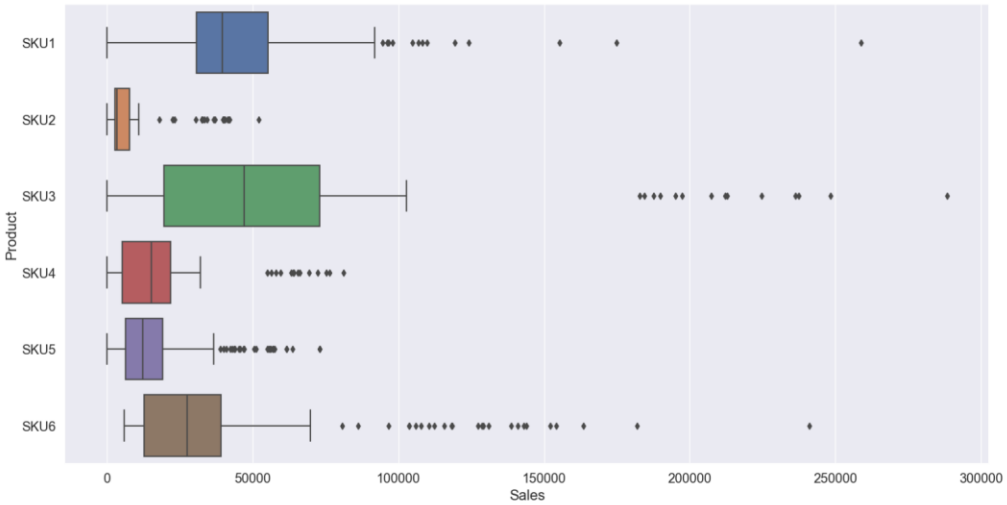
```
Product               object
date                  object
Sales                  int64
Price Discount (%)    object
In-Store Promo         int64
Catalogue Promo        int64
Store End Promo        int64
Google_Mobility      float64
Covid_Flag             int64
V_DAY                  int64
EASTER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

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

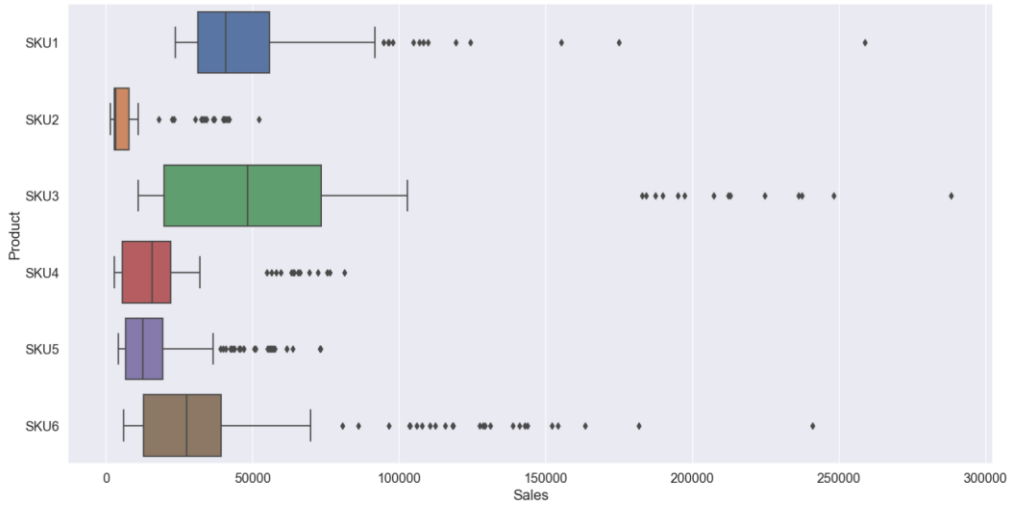
b. Check Missing values:

There is no missing values in the dataset

c. Check Outliers



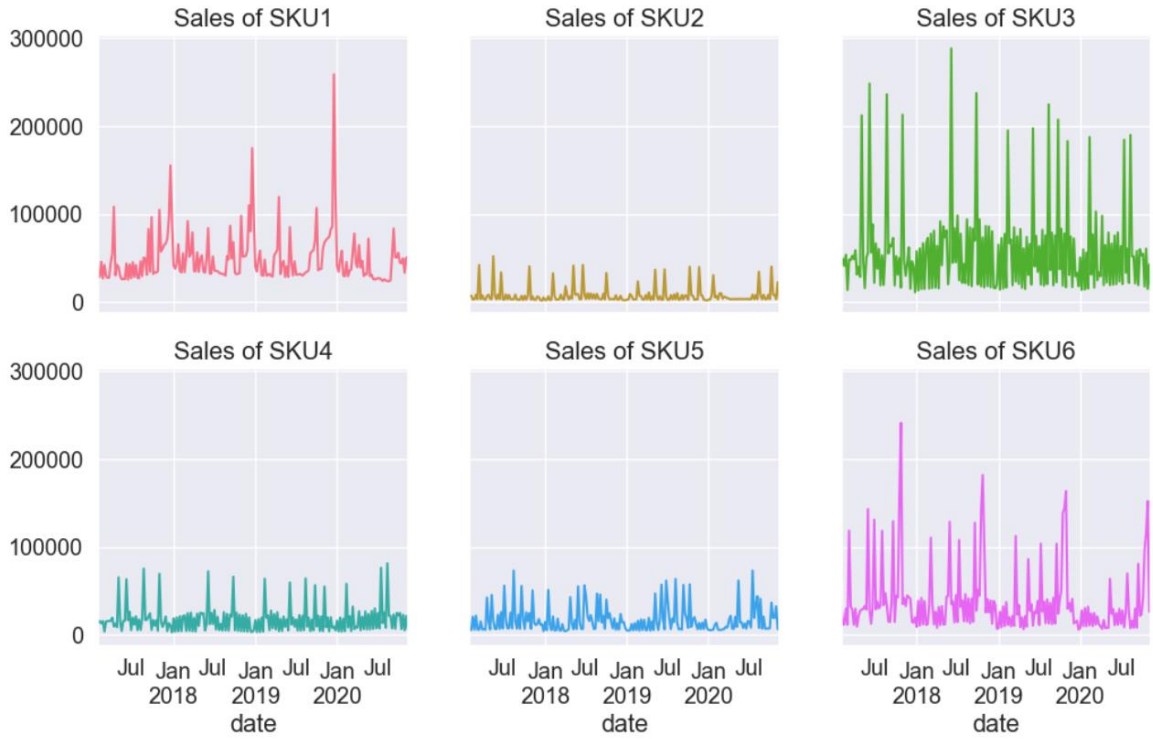
There are weeks with zero Sales in SKU1-5. It is believed that the testing weeks were mixed into the dataset.



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

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_(%).**