

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.....	3
2. Business understanding.....	3

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

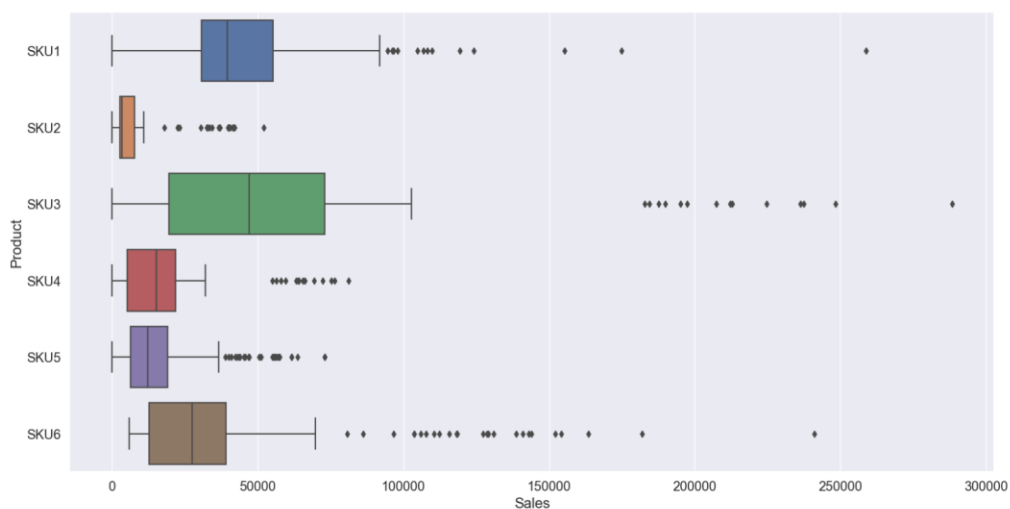
```

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

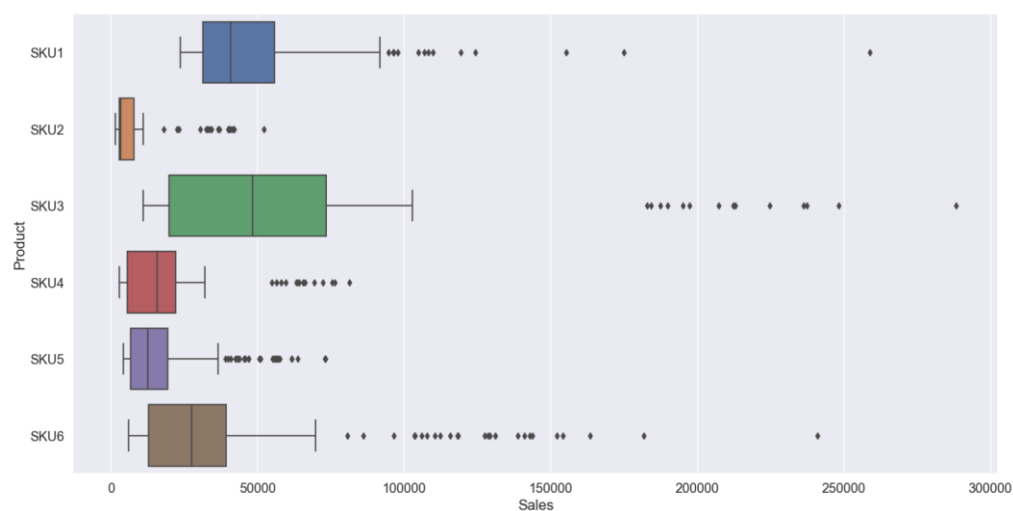
b. Check Missing values:

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.



4. Exploratory Data Analysis

Figure 1 displays six time-series plots showing sales of SKUs 1 through 6 from July 2018 to July 2020. The y-axis for all plots ranges from 0 to 300,000. The x-axis for all plots is labeled 'date' and shows months from Jul 2018 to Jul 2020. The plots are arranged in a 2x3 grid. The top row shows Sales of SKU1 (red), Sales of SKU2 (orange), and Sales of SKU3 (green). The bottom row shows Sales of SKU4 (teal), Sales of SKU5 (blue), and Sales of SKU6 (purple). SKU1 shows a significant spike in early 2020, reaching nearly 260,000. SKU2 shows low, periodic sales. SKU3 shows high, frequent sales. SKU4 shows low, periodic sales. SKU5 shows low, periodic sales. SKU6 shows moderate sales with a peak of approximately 240,000 in early 2019.

	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_ (%)**.