{most code are in python, some parts in R}

# 1 Overall View

From the description and the training data, we can see the topic is about time series prediction.

We read train.csv as train and test.csv as test. First, we are going to explore something lying in the features.

In

1. **print**(train.columns)

Out

1. Index(['date', 'country', 'store', 'product', 'num\_sold'], dtype='object')

In these features, **country**, **store** and **product** are discrete variable and **num\_sold** is number of sold product on specific **date**:

1. country    ['Finland' 'Norway' 'Sweden']
2. store    ['KaggleMart' 'KaggleRama']
3. product    ['Kaggle Mug' 'Kaggle Hat' 'Kaggle Sticker']
4. sold max : 2884
5. sold min : 70

# 2 EDA

{ALL PLOTS AND CALCULATIONS ARE PRODUCED BY R CODE}

{For viewers’ convenience, we put the conclusion both here and the end of EDA part}

1 There is a common peak value for all 3 products: around new year.

2 Sales condition of three products are influenced equally by different warehouse.

3 Sales condition of three products have same distribution in different countries.

4 Three products have different sailing periodicity (New year peak doesn’t take into consideration here). For ‘Kaggle Mug’ and ‘Kaggle Hat’, the trend looks like sine function and for ‘Kaggle Sticker’, there is no obvious periodicity.

5 All the products witness an annual growth.

6 weekdays and weekends will influence sales.

## Inside view of data

### general level

1 We calculate the mean value of sales on weekdays and weekends, find that sales condition is better during weekends (449.4846 on weekends and 362.7057 on weekdays).

图表

描述已自动生成

图表, 散点图

描述已自动生成

图表

描述已自动生成

From the plot, we can see:

1 strong periodicity.

2 peak values around the new year.

Now, we will have a deeper looking into different groups(**country,warehouse,product**).

### product-specific view

#### Kaggle Mug

图表, 散点图

描述已自动生成

图表

描述已自动生成

图表

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表

描述已自动生成

These plots are about sailing condition of **‘Kaggle Mug’** in different countries and warehouses.

We can see that:

1 minimum sale happens every summer (June and July).

2 Two warehouses share similar trend but have different value.

3 small growth every year.

图表, 散点图

描述已自动生成图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

The y-axis is the ratio of **‘num\_sold’** in two warehouses in the same country. So, there is a obvious difference between two warehouses.

#### Kaggle Hat

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

From the plot above, we find:

1 the minimum sale can always be seen during the end of September.

2 Just like ‘Kaggle Mug’, different warehouse will have different impact on sale condition, and even the ratio is approximately same.

3 We can see small growth every year.

#### Kaggle Sticker

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成

手机屏幕截图

中度可信度描述已自动生成

From the plot above, we can see:

1 There is no strong periodicity.

2 Different warehouse impact on sales.

3 Obvious annual growth.

## Conclusion

1 There is a common peak value for all 3 products: around new year.

2 Sales condition of three products are influenced equally by different warehouse.

3 Sales condition of three products have same distribution in different countries.

4 Three products have different sailing periodicity (New year peak doesn’t take into consideration here). For ‘Kaggle Mug’ and ‘Kaggle Hat’, the trend looks like sine function and for ‘Kaggle Sticker’, there is no obvious periodicity.

5 All the products witness an annual growth.

6 weekdays and weekends will influence sales.

## Feature Engineering:

1 Periodicity. We can see strong periodicity on sale condition of Kaggle Mug and Kaggle Hat. To solve this situation, we add 20 sine and cosine functions into the feature to simulate the periodicity.

2 Annual growth. All three products witness an annual growth. According to the simple calculation (comparing their mean and median value), we set this parameter to each year [2015, 2016, 2017, 2018,2019] = [1,1.02,1.03,1.04,1.05].

3 Weekdays and weekends. We set dummy variables to make it binary.

4 New Year’s Day. We set dummy variables to make it binary.

# 3 Training

Using model ‘xgboost’ and package ‘optuna’. The loss function is Symmetric Mean Absolute Percentage Error (SMAPE).

Here is the two slides of prediction and true value.

Points: prediction Lines: true value Kaggle results (5.06)

图片包含 图表

描述已自动生成

图片包含 图形用户界面

描述已自动生成

# 4 Newly update

1/11/2022

Since we are going to explore sale condition of 3 products. So, we can divide the train and test into three parts and use three different models to fit the data. At the end, we make a combination of these results.

In the new data, we can have less features and less sparse features.

Kaggle results: 5.02