The “**kimchi**” data set contains “**Date, Region, Total Volume, Total Boxes, Small Boxes, Large Boxes, XLarge Boxes, and Price**”. The task is to predict the “**Price**” from this dataset.

The result

I have analyzed the dataset and tried to predict the “Price”. Those procedures “analysis to prediction” have briefly described below:

Overview of Dataset

Form the data I have found out some missing values and different types of features.

Table 1. Overview of Dataset

Column Name	Number of Data	Missing No. of Data	Type of Data
Date	648	0	datetime64[ns]
Region	648	0	object
Total Volume	647	1	float64
Total Boxes	648	0	float64
Small Boxes	648	0	float64
Large Boxes	648	0	float64
XLarge Boxes	648	0	float64
Price	644	4	float64

Data Analysis

From statistical values (mean, std, min, 25%, 50%, 75%, max), I have seen some unpleasant result. Ex. for “price”, the **min** value is **1.01**, the **max** value is **3003.00**, and the **mean** value is **6.533540**. So, some noise can predict inside the dataset from the statistical overview. It could be more clear if we consider the percentile value for “Price”. Now, I have to prepare the dataset for finding out the correlation of features between them and ready to fit for the machine learning model.

Preparing Dataset:

1. Dropping the missing values. Because the number missing value is tiny. If the number of missing values is large then probably I would fill it by using different types of methods, ex. “forward fill method”

2. In the “Region” column there are so repeated values. I have found out the unique values and filled them with integer values.

Finding the Correlations between features:

1. From figure 1(a) it is visible that the feature correlations between them are too poor.
2. I have missed one feature to prepare the dataset which is the “Date” feature. I have converted the date to an integer. Then, tried to find out the correlation. But the result had not improved as is visible from figure 1(b).
3. To better understand I have plotted the graphical view of correlation as is visible in figure 2.

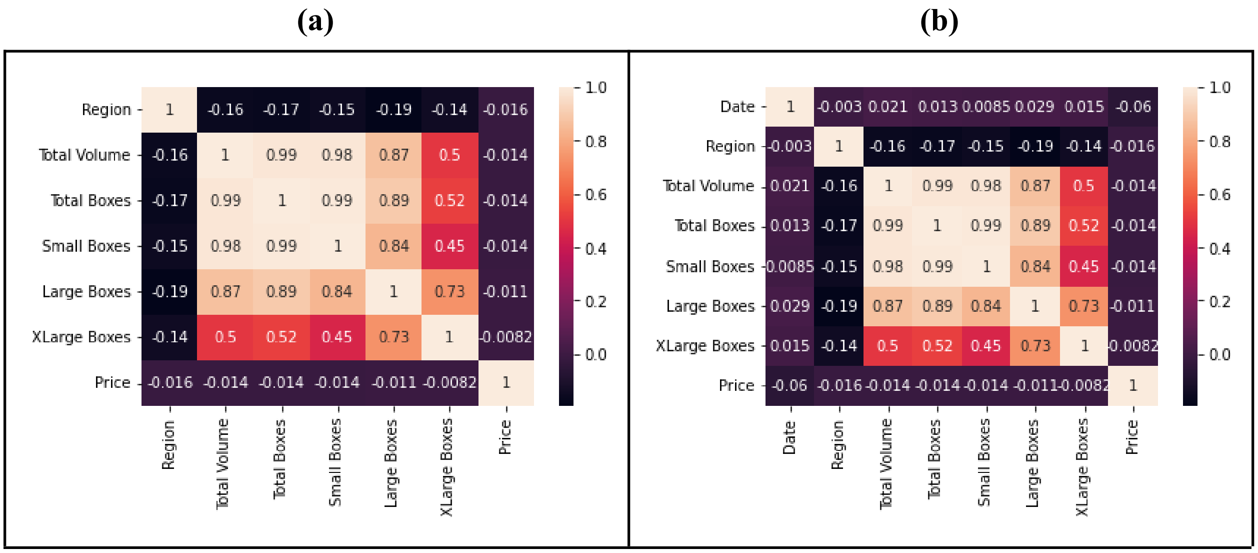

Figure 1. Correlation Heatmap

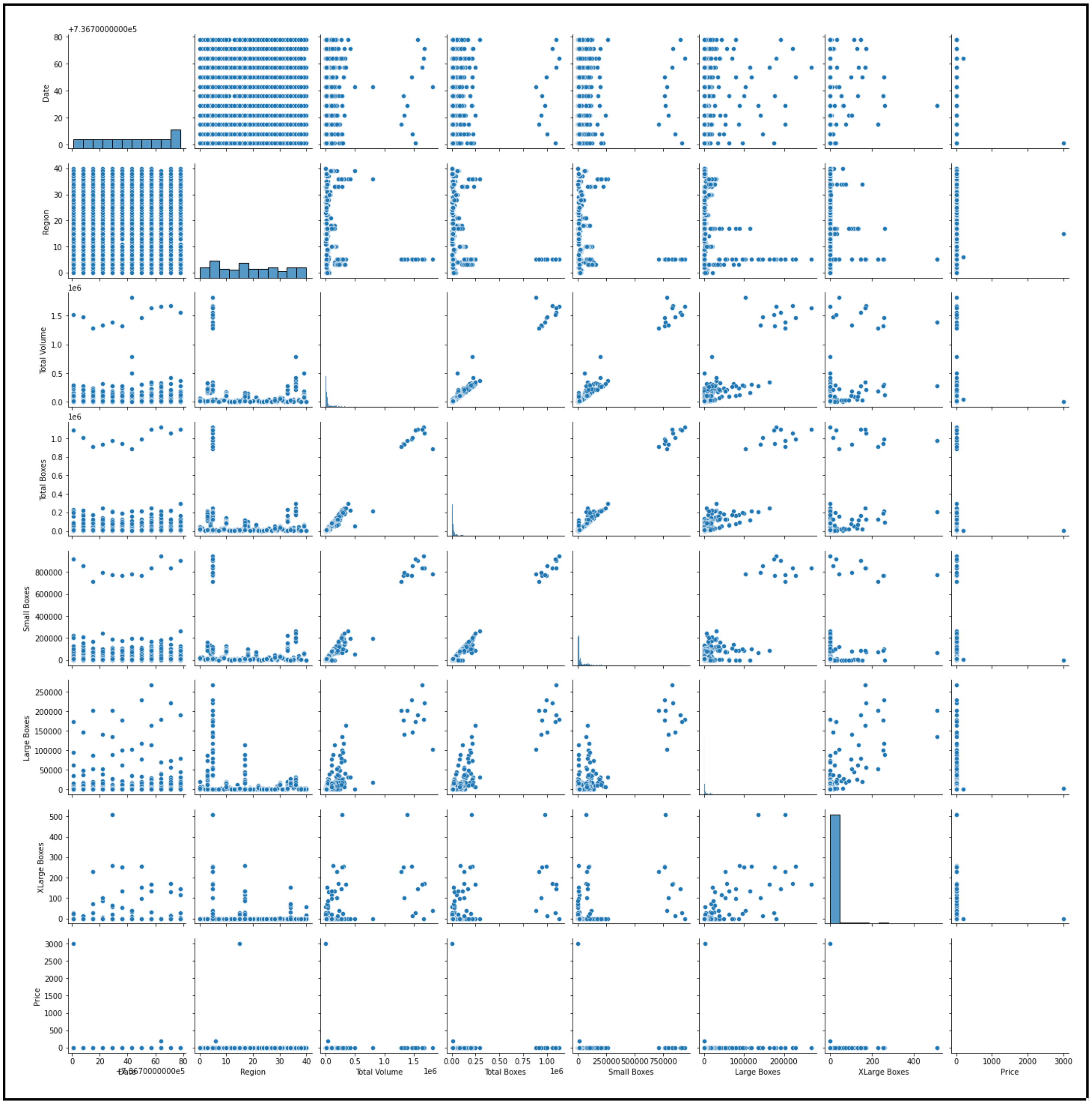

Figure 2. Correlation Pairplot

Model Selection

Now it seems like a classic regression problem where "**Price**" will be "**Y**" and **other features** will consider as "**X**". As the dataset is too small I would like to select **XgBoost** as this algorithm are proved to work well in this type of dataset.

1. Split the data set into 67% (training set) and 33% (testing set)
2. Got the accuracy of this regressor is 65.17% and rmse 0.1299

Model Evaluation

As I said earlier that our dataset's features are not well correlated. So, the accuracy shouldn't be well enough as expected. As per evaluation, I have applied the average **MAE** across the three repeats of 10-fold cross-validation. In this case, the model achieved an **MAE** of about **0.105**. This is not good enough but regarding our dataset it is okay. As I have shown an example that our model can predict nearly. And if we look around at scattered plots (figure 3(a)) the predictions are more or less in a line as it means that with this dataset it is well predicted (not well enough) but has some outliers. And distribution plots also show (figure 3(b)) that the overall price prediction the more or less well predicted.

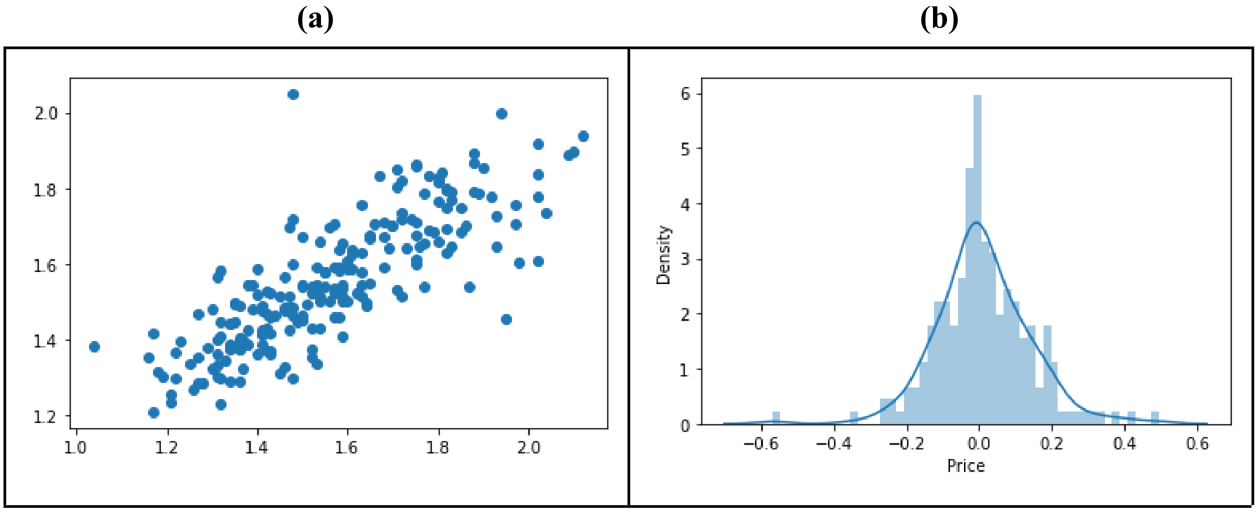
(a) **(b)**

Figure 3. Price Predictions Graphical View

Model Outliers Detection and Remove

As in prediction it is visible some outliers were found in our prediction. If it is possible to improve better accuracy by removing outliers. From the boxplot (figure 4) of the "Price" we can see some outliers is there and if we take min as 1% and max as 99.70% in quantile some outliers could be delected and it might be possible to improve our model. After detecting the outliers and removing the outliers accuracy is not significantly change but improved (accuracy 65.83% and rmse 0.1329) and if we look at our new plots, scattered plot (figure 5(a)) represents that it is precise than earlier and distribution plot (figure 5(b)) looks more good that earlier. Though our data is not well correlated this is well predicted I think. As I said form statistical calculation I found some unpleasant result from it. Lets try with another feature to detect outliers and removing those. It could make more improve our model. I have plotted "Total Volume" in boxplot as the distribution of this data is not well (from statistical result). And from the boxplot (figure 6) it also shows that too many outliers are there. But after removing those our accuracy should be improved. But it didn't happen as it shows on scattered plot (figure 7(a)) and distribution plot (figure 7(b)) but **rmse** value improved. Because the dataset has gone smaller. With this small data better accuracy findings might not be possible. So I didn't go through another feature.

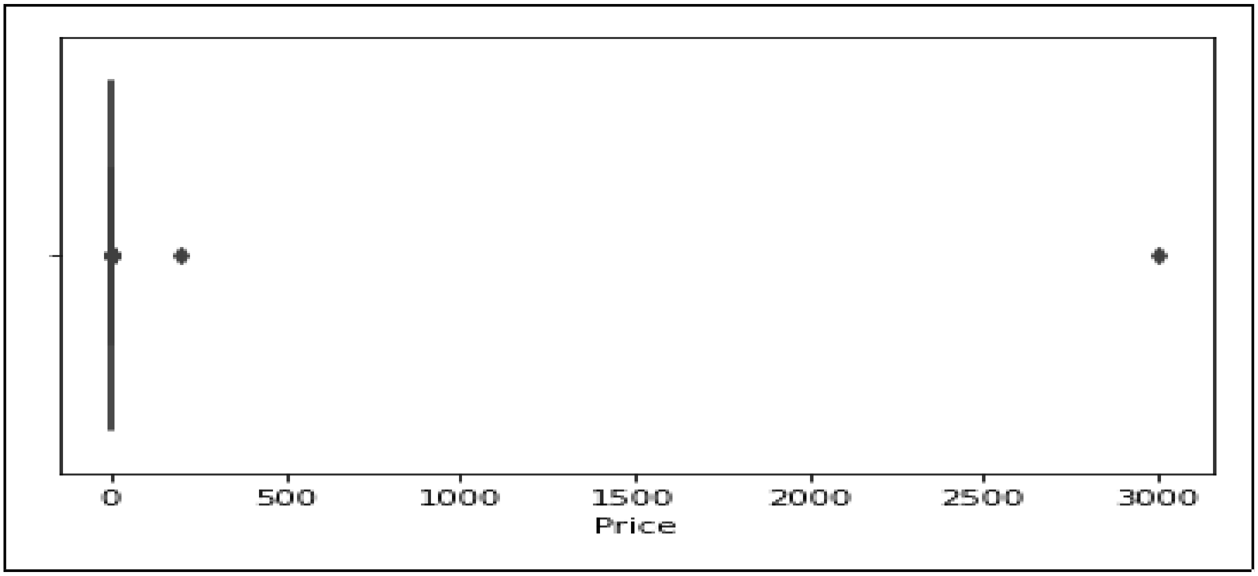

Figure 4. "Price" Feature Boxplot Graphical View

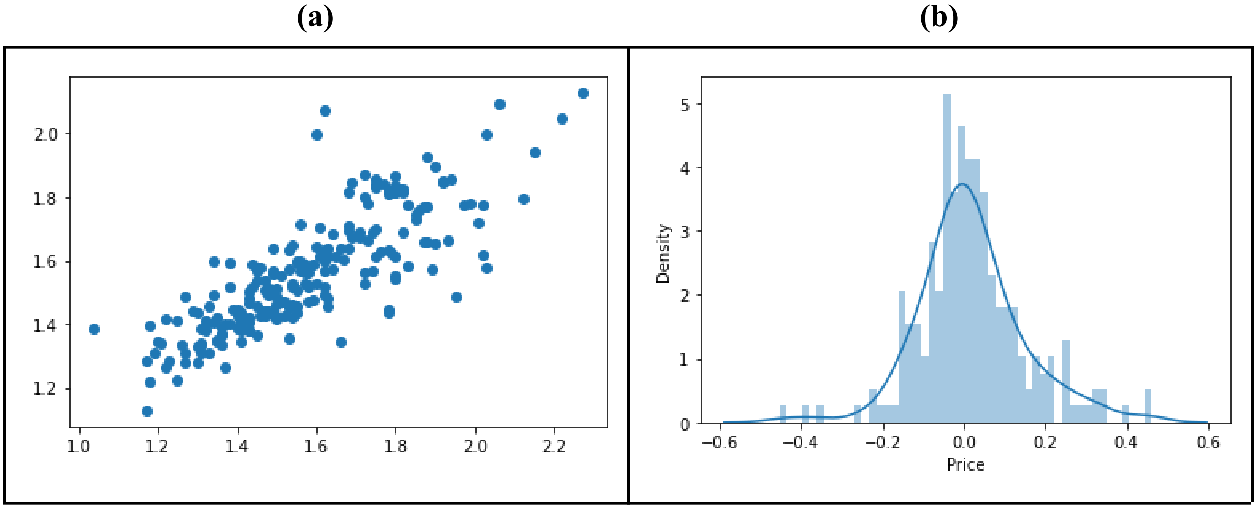

Figure 5. Price Predictions Graphical View

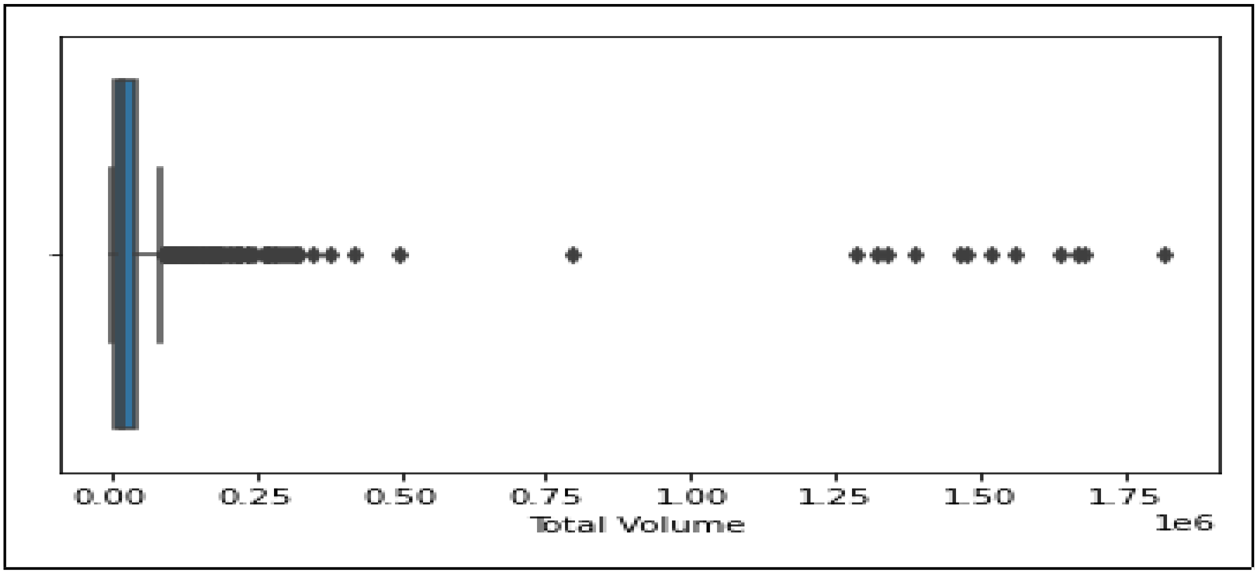

Figure 6. "Price" Feature Boxplot Graphical View

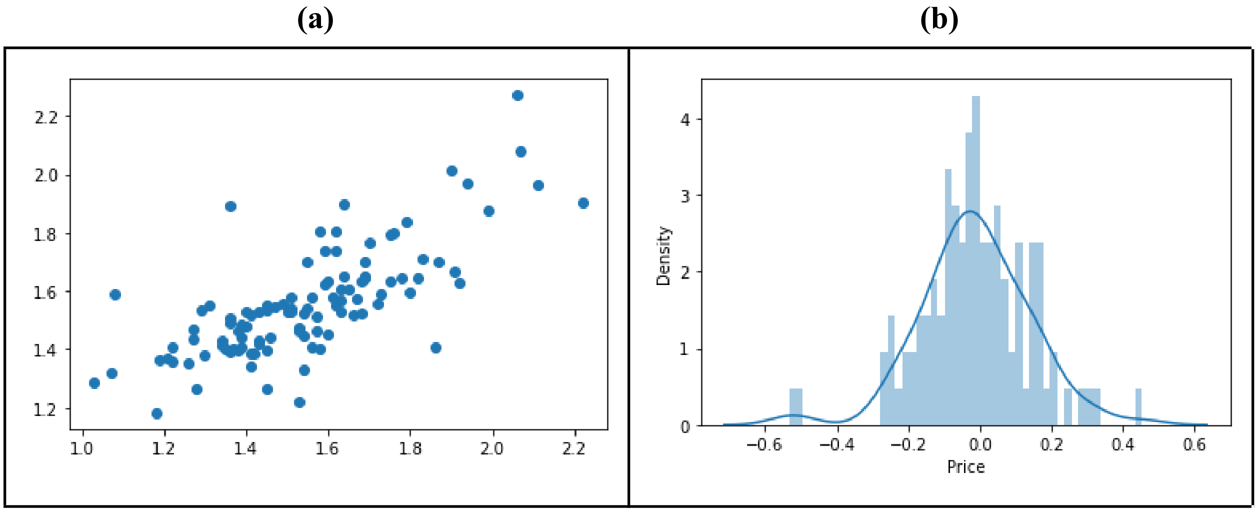

Figure 7. Price Predictions Graphical View

<u>Findings</u>

1. The dataset's features are not well correlated.
2. Too many outliers in feature but after removing it the dataset has gone to smaller which will not fit for machine learning algorithms to predict.
3. With this dataset accuracy score is 65.17% and rmse is 0.1299 (before preprocessing) and accuracy score is 65.83% and rmse is 0.1329 (after preprocessing) which represents slightly improved.