# Business Analytics Project

## Final

15146302 Kang, Euiheyon
15146304 Ku, Seunghyo
15146312 Park, Gangmin

# CONTENTS

# 1. Introduction

# 1. Introduction

**Kimchi:**

1. Traditional Korean food
2. Third most popular food in Korea

“ Predict the **price** of **cabbages** after **one week** ”

# 1. Introduction

## Precedent Research

- **Study 1:** Predict wholesale prices of onions
  - **Data**: Cabbage wholesale prices, production of onion, cultivation areas of onion.
  - **Model**: ARDI(Autoregressive and distributed lags) model

- **Study 2:** Cabbage wholesale price forecasting using unstructured data
  - **Data**: Cabbage prices, frequency of agricultural weather keywords with crop growth in documents containing cabbage.
  - **Model**: Autoregressive model

# 1. Introduction

## Precedent Research

- **Study 1:** Predict wholesale prices of onions
  - **Different**: Using total 7 models
    (Gradient Boosting Regression, KNN Regression, Support Vector Regression, Linear Support Vector Regression, Linear Regression, Random Forest Regression, Decision tree Regression)

- **Study 2:** Cabbage wholesale price forecasting using unstructured data
  - **Different**: Using structured agricultural meteorological data

# 1. Introduction

**- Topic**

Predict cabbage prices after a week

- **Method**

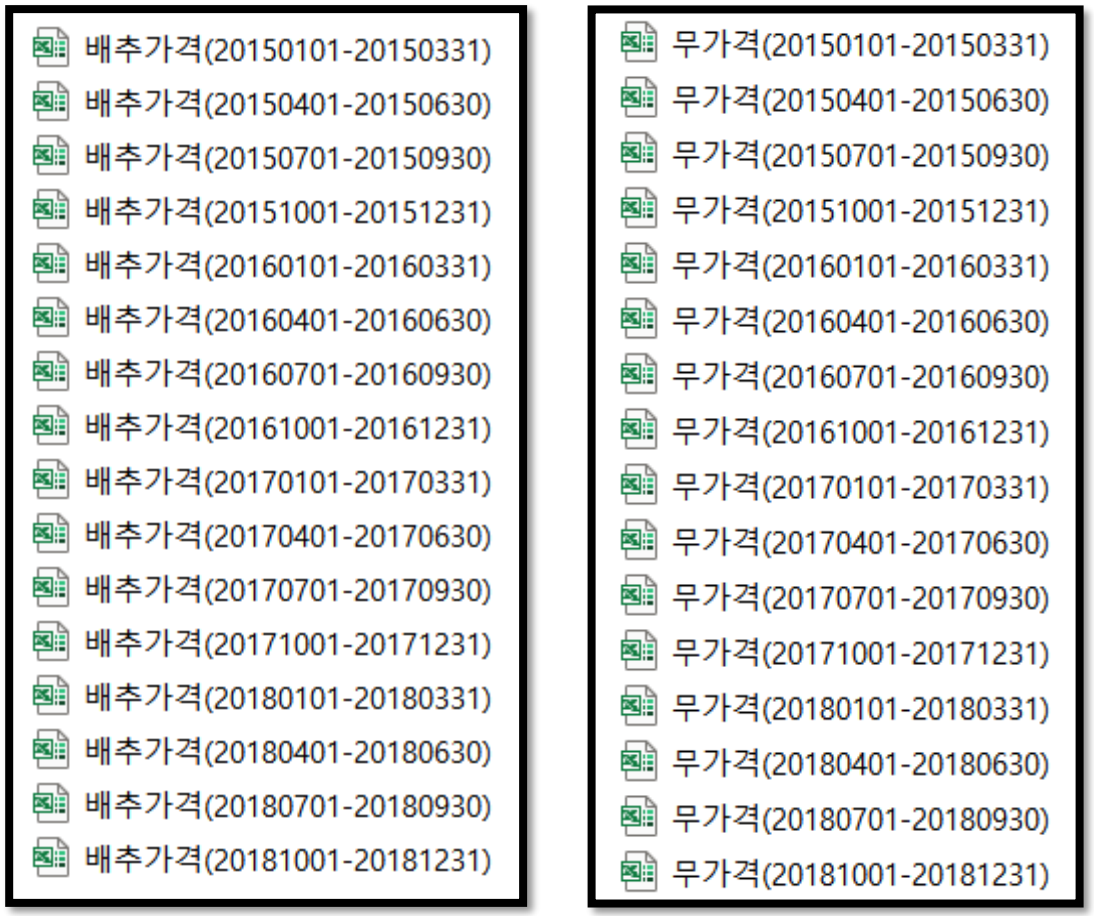
Use supervised learning with factors that affect the harvest of cabbages

- **Target:** Wholesale prices of cabbage (**Numeric**)
- **Column:** Time series data of weekly average data (**Numeric**)
    - Prices of cabbage, radish (2015-2018)
    - Weather information (temperature, precipitation, dew points.. Etc.)
    - Import information (import volume, import price)

# 2. Data

# 2.1 Data Collection

www.kamis.or.kr




**Price of cabbage, radish**

(2015~2018)

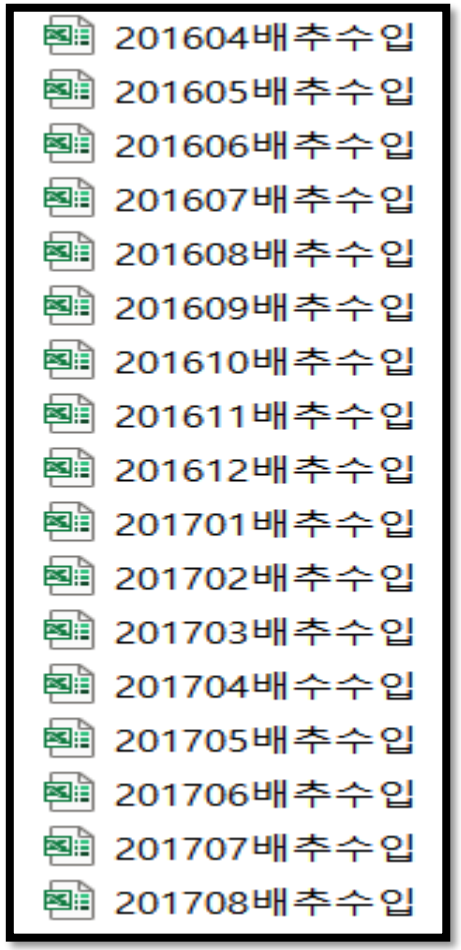
## Agricultural Transaction data

This data set consists of the day-to-day, average transaction price by region, and the normal year price from 2015 to 2018.

* Normal Year price excluding maximum and minimum values in the data for the five-year period based on this year's data

# 2.1 Data Collection

Kati.net



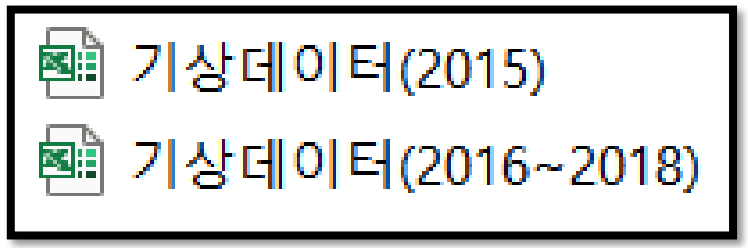

**Volume and Price of imported cabbage**
(2015 – 2018)

## Imported Agricultural Product data

This data set consists of monthly cabbage import volume and import price.

# 2.1 Data Collection

기상청 날씨누리 Weather.go.kr




**Weather data**
(2015 – 2018)

## Weather data

| | |
|---|---|
| Average Temperature | Lowest Temperature |
| Highest Temperature | Precipitation |
| Wind Speed | Dew Point |
| Relative Humidity | Vapor Pressure |
| Spot Atmospheric Pressure | Seal-level Pressure |
| Duration of Bright Sunshine | Solar Radiation Quantity |
| Maximum Depth of Snow Cover | Maximum Depth of New Snowfall |
| Amount of Cumulus | Average Temperature of Ground |
| Total Low Evaporation | Total Large Evaporation |
| Amount of Middle and Lower Cloud | |

# 2.2 Data Processing

| 구분 | | 01월 02일 | 01월 05일 | 01월 06일 | 01월 07일 |
|---|---|---|---|---|---|
| 평균 | | 2,800 | 2,800 | 2,800 | 2,800 |
| 평년 | | 5,080 | 7,462 | 7,502 | 6,124 |
| 서울 | 가을 | 3,000 | 3,000 | 3,000 | - |
| | 월동 | - | - | - | 3,000 |
| 부산 | 가을 | 3,000 | 3,000 | 3,000 | - |
| | 월동 | - | - | - | 3,000 |
| 대구 | 가을 | 3,000 | 3,000 | 3,000 | - |
| | 월동 | - | - | - | 3,000 |
| 광주 | 가을 | 2,000 | 2,000 | 2,000 | - |
| | 월동 | - | - | - | 2,000 |
| 대전 | 가을 | 3,000 | 3,000 | 3,000 | - |
| | 월동 | - | - | - | 3,000 |

| Location | Date | Average temperature | Lowest temperature | Highest temperature |
|---|---|---|---|---|
| 108 | 2015-01-01 | -7.7 | -9.8 | -4.3 |
| 133 | 2015-01-01 | -5.6 | -8.3 | -2.4 |
| 143 | 2015-01-01 | -3.9 | -5.9 | -0.5 |
| 156 | 2015-01-01 | -3.7 | -5.8 | -1.4 |
| 159 | 2015-01-01 | -2.2 | -4.9 | 1.2 |

**108 – Seoul**
**133 – Daejeon**
**143 – Daegu**
**156 – Gwangju**
**159 - Busan**

## Make total average data from whole regions

| Date | Average temperature | Lowest temperature | Highest temperature |
|---|---|---|---|
| 2015-01-01 | -4.62 | -6.94 | -1.48 |

```
for i in range(len(list_weather)):
    list_weather[i] = list_weather[i].reset_index(drop=True)

length = len(list_weather[0])
avg_weather = list_weather[0].copy()
columns = list_weather[0].columns
columns = ['Location',  'Average temperature', 'Lowest temperature',
       'Highest temperature', 'Precipitation', 'Wind speed', 'Dew point',
       'Relative humidity', 'Vapor pressure', 'Spot atmospheric pressure',
       'Sea-level pressure', 'Duration of bright sunshine',
       'Solar radiation quantity', 'Maximum depth of snow cover',
       'Maximum depth of new snowfall', 'Amount of cumulus',
       'Amount of middle and lower cloud', 'Average temperature of ground',
       'Total large evaporation', 'Toal low evaporation']

for i in columns:
        avg_weather[i] = list_weather[0][i].add(list_weather[1][i])
        avg_weather[i] = avg_weather[i].add(list_weather[2][i])
        avg_weather[i] = avg_weather[i].add(list_weather[3][i])
        avg_weather[i] = avg_weather[i].add(list_weather[4][i])
        avg_weather[i] = avg_weather[i].div(len(list_weather))

avg_weather = avg_weather.drop(['Location'],axis=1)
```

# 2.2 Data Processing

## Processing Missing Values

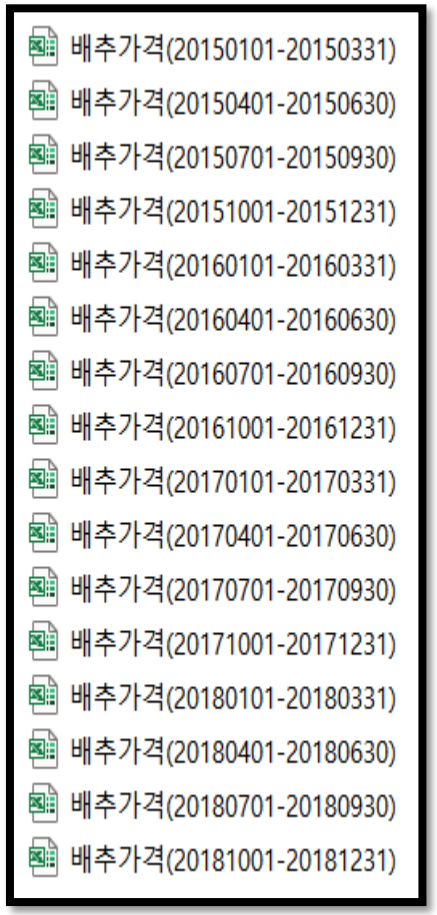
| | |
|---|---|
| Location | 0 |
| Date | 0 |
| Average temperature | 0 |
| Lowest temperature | 0 |
| Highest temperature | 0 |
| Precipitation | 868 |
| Wind speed | 0 |
| Dew point | 0 |
| Relative humidity | 1 |
| Vapor pressure | 1 |
| Spot atmospheric pressure | 0 |
| Sea-level pressure | 0 |
| Duration of bright sunshine | 0 |
| Solar radiation quantity | 0 |
| Maximum depth of snow cover | 1403 |
| Maximum depth of new snowfall | 1396 |
| Amount of cumulus | 0 |
| Amount of middle and lower cloud | 15 |
| Average temperature of ground | 0 |
| Total large evaporation | 242 |
| Toal low evaporation | 0 |

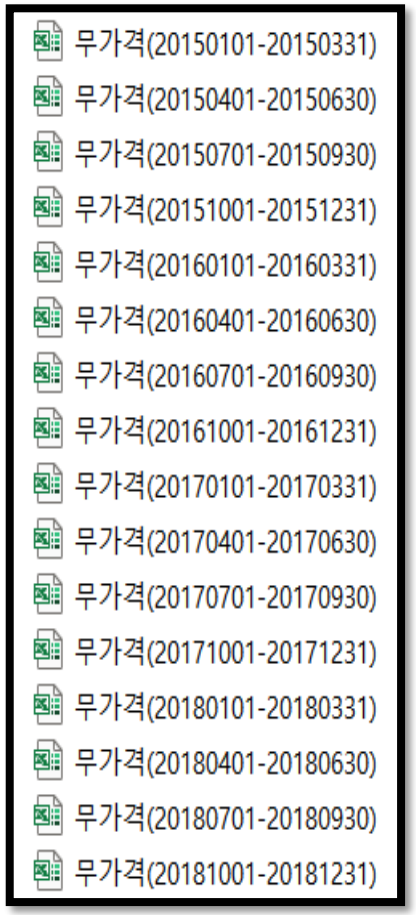
**Type 1** - **No events occurred**
- 'Maximum depth of snow cover', 'Maximum depth of new snowfall' 'Total large evaporation' 'Precipitation'
- Fill the missing values with zero

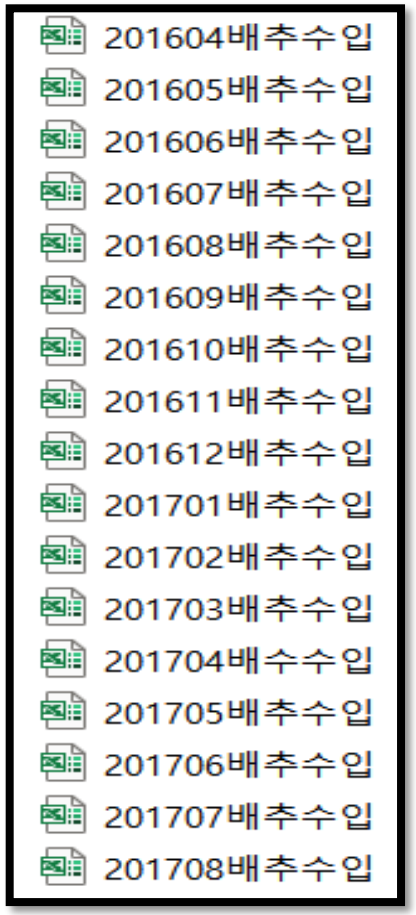
**Type 2** - **Not recorded**
- 'Amount of middle and lower cloud'
- Fill the missing values with the average of the other year's measurements

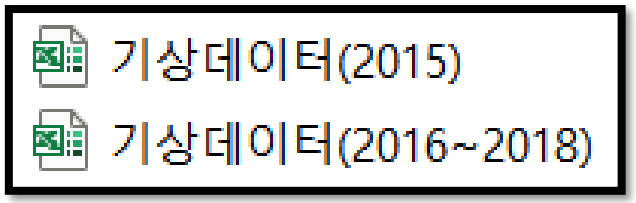
**Type 3** - **Not recorded(Don't have patterns)**
- 'Relative humidity' , 'Vapor pressure'
- Fill the missing value by referring to the data from the previous one

# 2.2 Data Processing



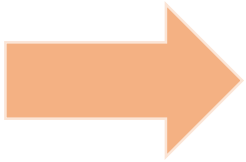




**Merging data**

Common key -> Date

# 2.2 Data Processing

| | | |
|---|---|---|
| cabbage | DataFrame | (981, 3) |
| radish | DataFrame | (979, 3) |
| cabbage_import | DataFrame | (48, 3) |
| avg_weather | DataFrame | (1461, 20) |



| | | |
|---|---|---|
| Date | 981 non-null | datetime64[ns] |
| Cabbage_Average | 981 non-null | object |
| Cabbage_Normalyear | 981 non-null | object |
| Radish_Average | 964 non-null | object |
| Radish_Normalyear | 964 non-null | object |
| Average temperature | 981 non-null | float64 |
| Lowest temperature | 981 non-null | float64 |
| Highest temperature | 981 non-null | float64 |
| Precipitation | 981 non-null | float64 |
| Wind speed | 981 non-null | float64 |
| Dew point | 981 non-null | float64 |
| Relative humidity | 981 non-null | float64 |
| Vapor pressure | 981 non-null | float64 |
| Spot atmospheric pressure | 981 non-null | float64 |
| Sea-level pressure | 981 non-null | float64 |
| Duration of bright sunshine | 981 non-null | float64 |
| Solar radiation quantity | 981 non-null | float64 |
| Maximum depth of snow cover | 981 non-null | float64 |
| Maximum depth of new snowfall | 981 non-null | float64 |
| Amount of cumulus | 981 non-null | float64 |
| Amount of middle and lower cloud | 981 non-null | float64 |
| Average temperature of ground | 981 non-null | float64 |
| Total large evaporation | 981 non-null | float64 |
| Toal low evaporation | 981 non-null | float64 |
| Import weight | 981 non-null | object |
| Amount of income | 981 non-null | object |

Merge whole data through ‘Date’

# 2.3 EDA

## Merged Data Set

| X | DataFrame | (981, 26) |
|---|---|---|

**2015 - 248 EA**
**2016 - 246 EA**
**2017 - 243 EA**
**2018 - 244 EA**

## Columns

'Date' , 'Cabbage_Average' , 'Cabbage_Normalyear' , 'Radish_Average' , 'Radish_Normalyear'

'Lowest temperature' , 'Highest temperature' , 'Precipitation' , 'Wind speed' , 'Dew point'

'Vapor pressure' , 'Spot atmospheric pressure' , 'Sea-level pressure' , 'Duration of bright sunshine'

'Solar radiation quantity' , 'Maximum depth of snow cover' , 'Maximum depth of new snowfall'

'Amount of cumulus' , 'Amount of middle and lower cloud' , 'Average temperature of ground'

'Total large evaporation' , 'Total low evaporation' , 'Amount of income' , 'Relative humidity'

'Average temperature' , 'Import weight'

# 2.3 EDA

**Data type**

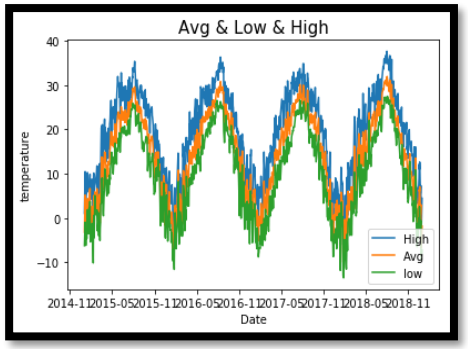
| | | |
|---|---|---|
| Date | 981 non-null | datetime64[ns] |
| Cabbage_Average | 981 non-null | float64 |
| Cabbage_Normalyear | 981 non-null | float64 |
| Radish_Average | 981 non-null | float64 |
| Radish_Normalyear | 981 non-null | float64 |
| Average temperature | 981 non-null | float64 |
| Lowest temperature | 981 non-null | float64 |
| Highest temperature | 981 non-null | float64 |
| Precipitation | 981 non-null | float64 |
| Wind speed | 981 non-null | float64 |
| Dew point | 981 non-null | float64 |
| Relative humidity | 981 non-null | float64 |
| Vapor pressure | 981 non-null | float64 |
| Spot atmospheric pressure | 981 non-null | float64 |
| Sea-level pressure | 981 non-null | float64 |
| Duration of bright sunshine | 981 non-null | float64 |
| Solar radiation quantity | 981 non-null | float64 |
| Maximum depth of snow cover | 981 non-null | float64 |
| Maximum depth of new snowfall | 981 non-null | float64 |
| Amount of cumulus | 981 non-null | float64 |
| Amount of middle and lower cloud | 981 non-null | float64 |
| Average temperature of ground | 981 non-null | float64 |
| Total large evaporation | 981 non-null | float64 |
| Toal low evaporation | 981 non-null | float64 |
| Import weight | 981 non-null | float64 |
| Amount of income | 981 non-null | float64 |

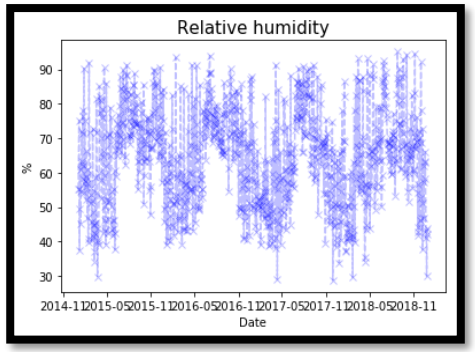


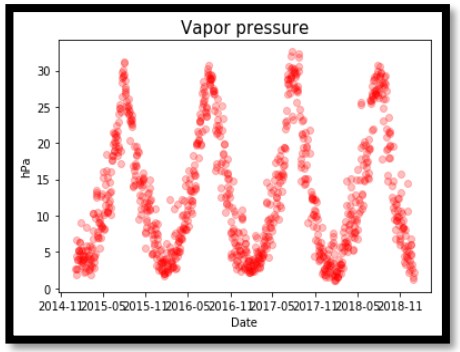
'Time Series Value'

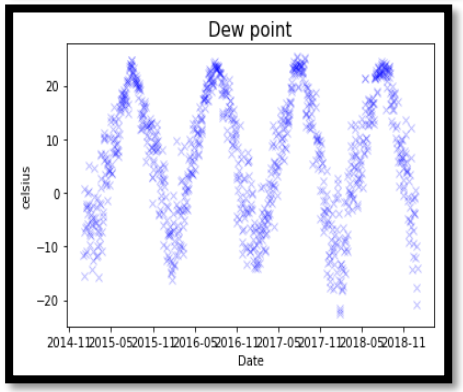
'Numeric Value'

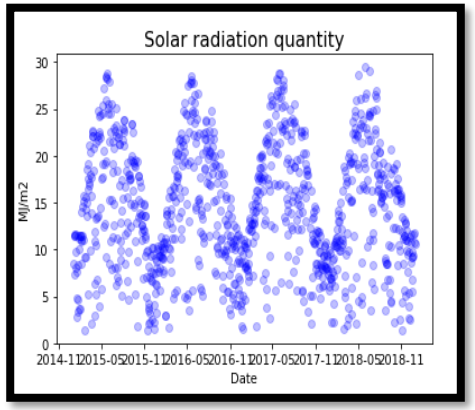
# 2.3 EDA

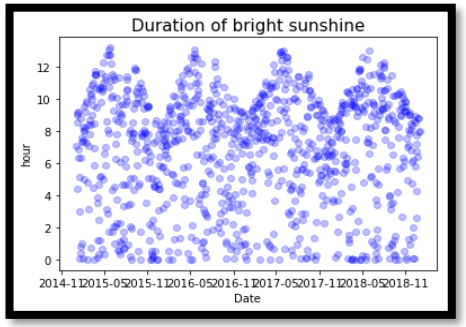


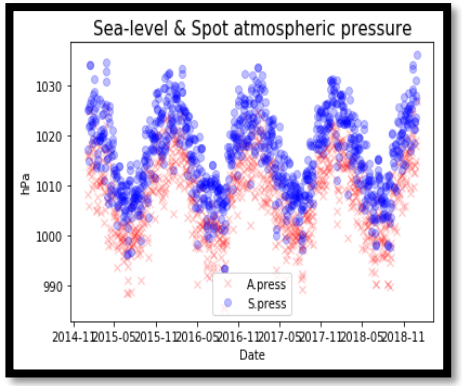



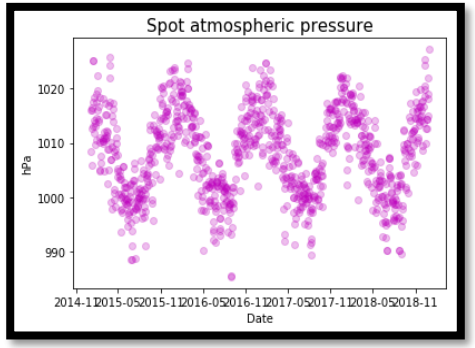



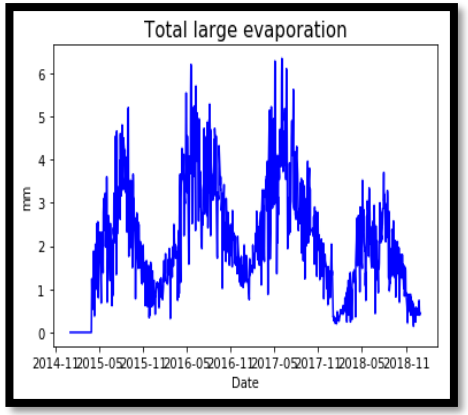



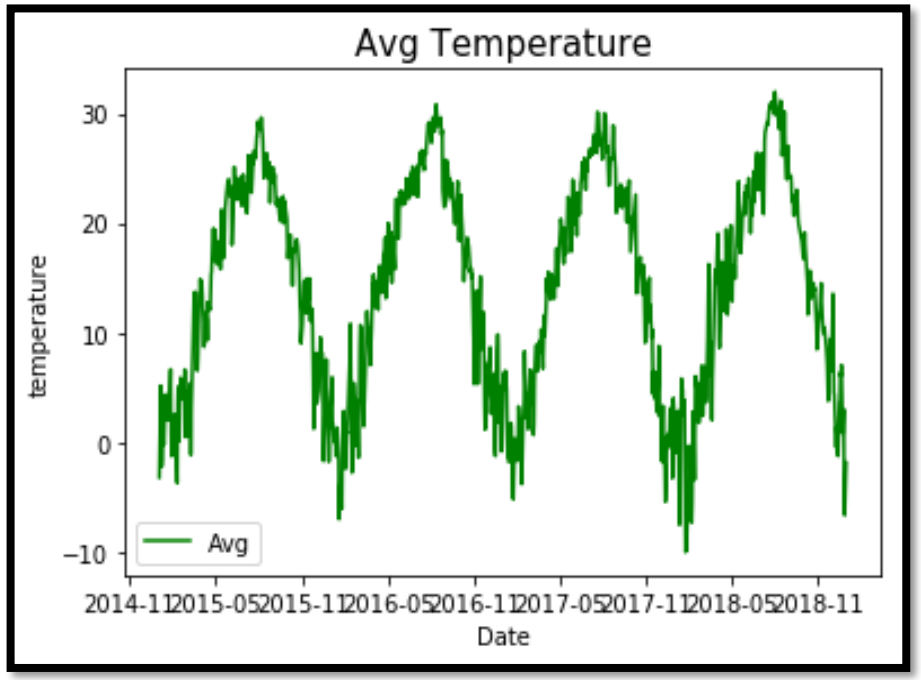



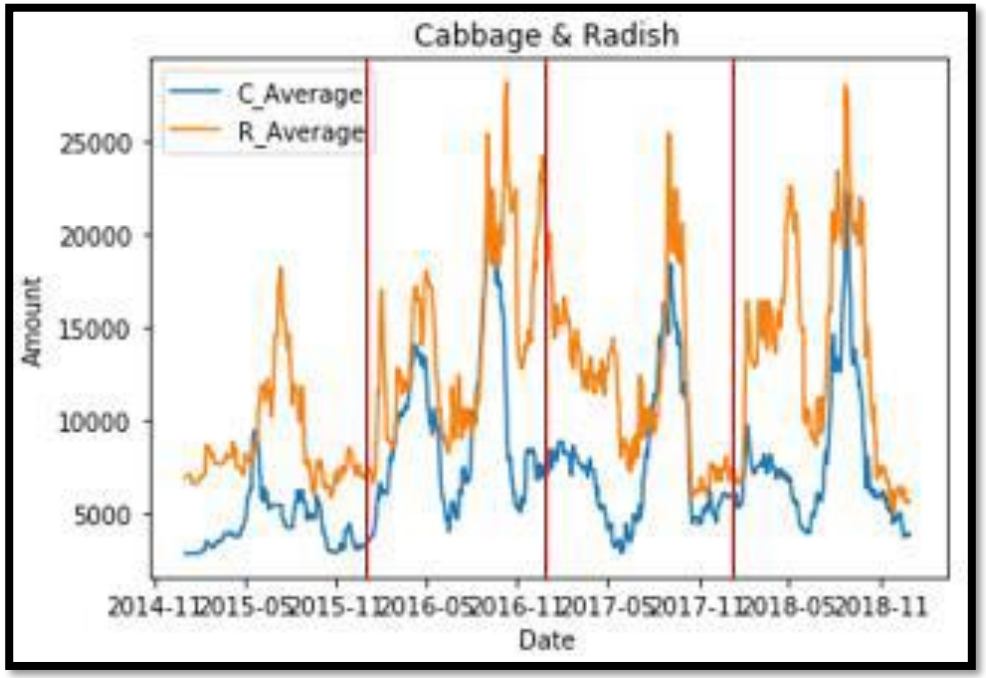




Most weather data have been found to be highly correlated with average temperature data

# 2.3 EDA




Similar graph shape and seasonal characteristics every year

# 3. Model

# 3.1 Model Training

From **Sea-level pressure** to **Radish Normal year**,

VIF Factor is more than 10

## Multicollinearity Test

```
In [8]: vif
Out[8]:
      VIF Factor                          features
0   56608.181859                Sea-level pressure_W
1   54500.013261          Spot atmospheric pressure_W
2   38012.539262                Average temperature_W
3   28952.010078                           Dew point_W
4    8042.601407                Highest temperature_W
5    7920.673867                 Lowest temperature_W
6    1541.038305                 Relative humidity_W
7    1091.312274      Average temperature of ground_W
8     261.679801        Duration of bright sunshine_W
9     253.284464             Total large evaporation_W
10    211.292730              Total low evaporation_W
11    111.145343                    Vapor pressure_W
12    109.044230                 Amount of cumulus_W
13    105.258127  Amount of middle and lower cloud_W
14     98.750587          Solar radiation quantity_W
15     37.429969                     Import price_W
16     36.478917                    Import weight_W
17     25.151146                      Wind speed_W
18     14.778542                Radish Normalyear_W
19      9.569020               Cabbage_Normalyear_W
20      9.065209                    Cabbage_Price_W
21      8.262303                     Radish_Price_W
22      5.833757        Maximum depth of snow cover_W
23      5.289138      Maximum depth of new snowfall_W
24      4.642909                      Precipitation_W
```

# 3.1 Model Training

## Multicollinearity Test

```
In [10]: vif
Out[10]:
      VIF Factor                          features
0   48063.591381                Sea-level pressure_W
1   45826.389579          Spot atmospheric pressure_W
2   26554.342402                       Dew point_W
3   14348.627680             Average temperature_W
4    4757.213022              Lowest temperature_W
5    1543.887022              Relative humidity_W
6    1043.442711   Average temperature of ground_W
7     257.114050     Duration of bright sunshine_W
8     175.882920          Total large evaporation_W
9     162.328363           Total low evaporation_W
10    104.809545                 Vapor pressure_W
11    102.633627                Amount of cumulus_W
12     99.309199         Solar radiation quantity_W
13     72.458978  Amount of middle and lower cloud_W
14     36.378661                  Import price_W
15     35.237003                 Import weight_W
16     25.049877                    Wind speed_W
17     14.480439              Radish_Normalyear_W
18      9.932128           Cabbage_Normalyear_W
19      8.530113                Cabbage_Price_W
20      8.029454                 Radish_Price_W
21      5.725926     Maximum depth of snow cover_W
22      5.075288  Maximum depth of new snowfall_W
23      4.568344                 Precipitation_W
```

Remove **Highest Temperature** from variable sets

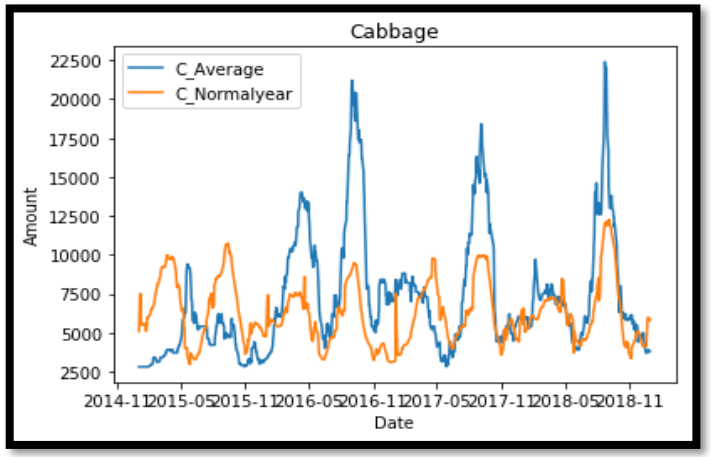
```
In [12]: vif
Out[12]:
      VIF Factor                          features
0   46127.276419                Sea-level pressure_W
1   44077.058682          Spot atmospheric pressure_W
2   23449.078326             Average temperature_W
3    7476.805882             Highest temperature_W
4    7218.550163              Lowest temperature_W
5    1070.303525                       Dew point_W
6    1032.642719   Average temperature of ground_W
7     247.626138     Duration of bright sunshine_W
8     177.727521          Total large evaporation_W
9     163.740135           Total low evaporation_W
10    104.849669                Amount of cumulus_W
11     97.784590                 Vapor pressure_W
12     95.589953         Solar radiation quantity_W
13     72.099589  Amount of middle and lower cloud_W
14     35.369003                  Import price_W
15     34.417530                 Import weight_W
16     25.675541                    Wind speed_W
17     14.464391              Radish_Normalyear_W
18      9.987378           Cabbage_Normalyear_W
19      8.530088                Cabbage_Price_W
20      8.020822                 Radish_Price_W
21      5.714485     Maximum depth of snow cover_W
22      5.074014  Maximum depth of new snowfall_W
23      3.862030                 Precipitation_W
```

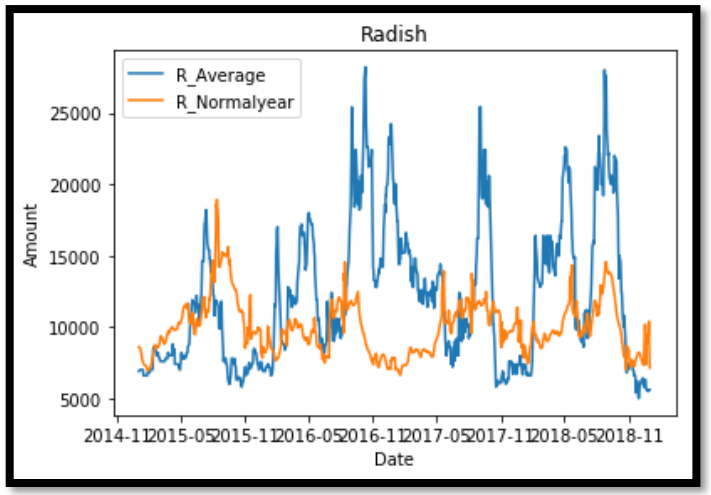
Remove **Highest Relative Humidity** from variable sets

**No Big difference**

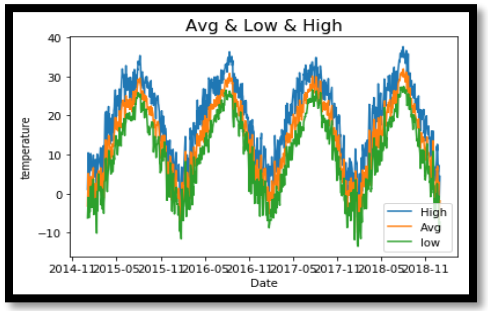
# 3.1 Model Training

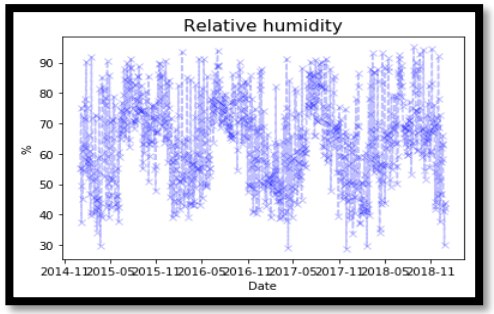
**Because from the EDA...**

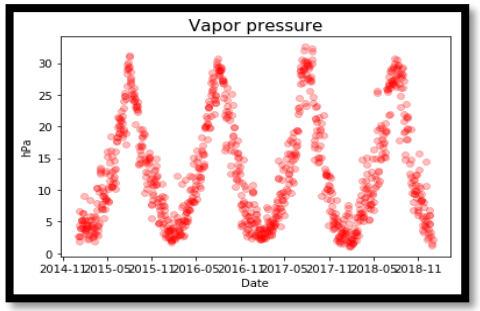


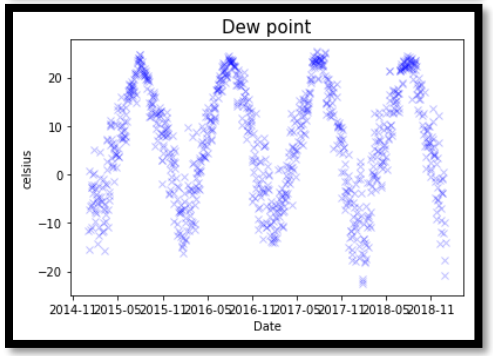



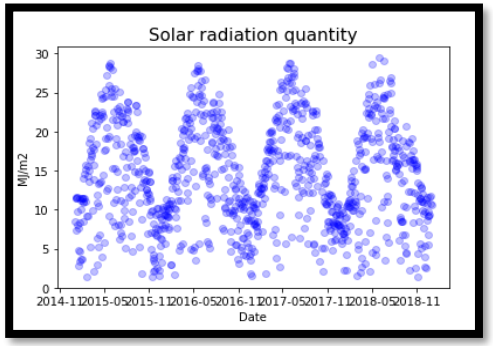

Similar Shape!

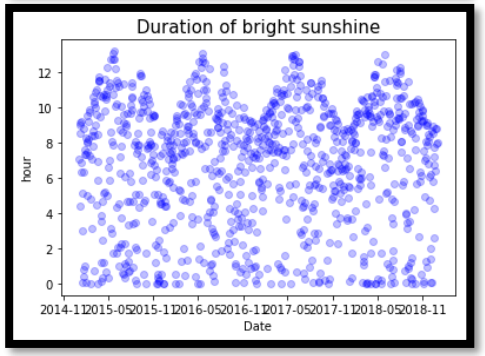


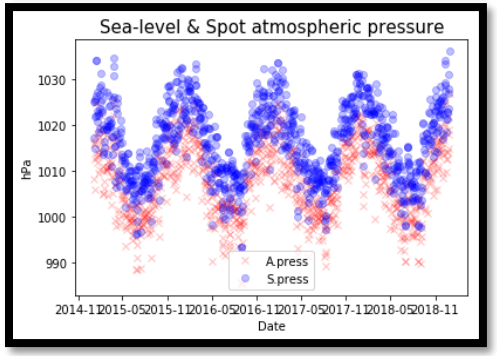



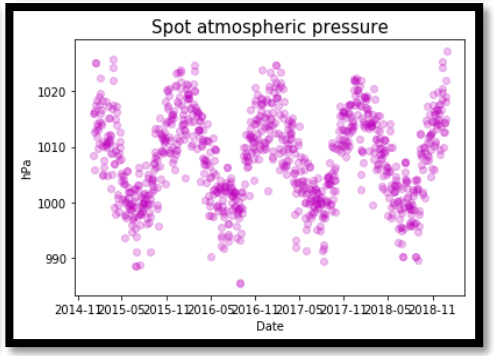



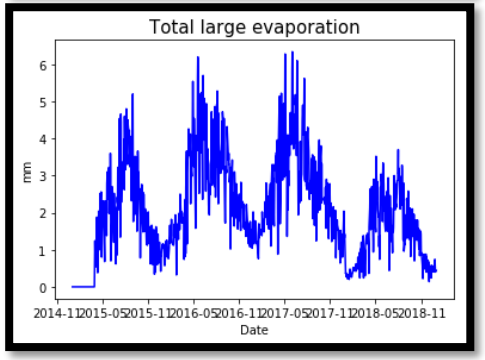



# 3.1 Model Training

| | VIF Factor | feature |
|---|---|---|
| 0 | 48084.703408 | Sea-level pressure_I |
| 1 | 45841.794674 | Spot atmospheric pressure_I |
| 2 | 36819.492206 | Average temperature_I |
| 3 | 26730.655590 | Dew point_I |
| 4 | 7551.560260 | Highest temperature_I |
| 5 | 7321.667346 | Lowest temperature_I |
| 6 | 1559.323067 | Relative humidity_I |
| 7 | 1043.492891 | Average temperature of ground_I |
| 8 | 258.881311 | Duration of bright sunshine_I |
| 9 | 177.727575 | Total large evaporation_I |
| 10 | 163.865282 | Total low evaporation_I |
| 11 | 109.182668 | Vapor pressure_I |
| 12 | 104.904881 | Amount of cumulus_I |
| 13 | 99.553236 | Solar radiation quantity_I |
| 14 | 72.459196 | Amount of middle and lower cloud_I |
| 15 | 36.622435 | Import price_I |
| 16 | 35.608547 | Import weight_I |
| 17 | 25.681921 | Wind speed_I |
| 18 | 14.610881 | Radish Normalyear I |

Remove all variables with more than 10 multicollinearity

| | | |
|---|---|---|
| 19 | 9.989151 | Cabbage_Normalyear_W |
| 20 | 8.530164 | Cabbage_Price_W |
| 21 | 8.049111 | Radish_Price_W |
| 22 | 5.727877 | Maximum depth of snow cover_W |
| 23 | 5.098014 | Maximum depth of new snowfall_W |
| 24 | 4.571857 | Precipitation_W |

| VIF Factor | feature |
|---|---|
| 48084.703408 | Sea-level pressure_I |
| 45841.794674 | Spot atmospheric pressure_I |
| 36819.492206 | Average temperature_I |
| 26730.655590 | Dew point_I |
| 7551.560260 | Highest temperature_I |
| 7321.667346 | Lowest temperature_I |
| 1559.323067 | Relative humidity_I |
| 1043.492891 | Average temperature of ground_I |
| 258.881311 | Duration of bright sunshine_I |
| 177.727575 | Total large evaporation_I |
| 163.865282 | Total low evaporation_I |
| 109.182668 | Vapor pressure_I |
| 104.904881 | Amount of cumulus_I |
| 99.553236 | Solar radiation quantity_I |
| 72.459196 | Amount of middle and lower cloud_I |
| 36.622435 | Import price_I |
| 35.608547 | Import weight_I |
| 25.681921 | Wind speed_I |
| 14.610881 | Radish Normalyear I |

Add one by one

| | |
|---|---|
| 9.989151 | Cabbage_Normalyear_W |
| 8.530164 | Cabbage_Price_W |
| 8.049111 | Radish_Price_W |
| 5.727877 | Maximum depth of snow cover_W |
| 5.098014 | Maximum depth of new snowfall_W |
| 4.571857 | Precipitation_W |

# 3.1 Model Training

**Pseudo algorithm for add variable one by one**

```
variableList = all Independece variable;
remainList = new remainList();
for(variable : variableList):
        if(variable.VIF > 10):
                variableList.remove(variable);
                remainList.add(variable);
for(variable : remainList)
        if(variableList.add(variable).VIF_all <10 && variable == minimum VIF increasing):
                        variableList.add(variable)
```

# 3.1 Model Training

**Pseudo algorithm for add variable one by one**

| | VIF Factor | features |
|---|---|---|
| 0 | 9.159643 | Relative humidity_W |
| 1 | 8.347921 | Radish_Normalyear_W |
| 2 | 8.171331 | Wind speed_W |
| 3 | 6.948952 | Cabbage_Normalyear_W |
| 4 | 6.051729 | Cabbage_Price_W |
| 5 | 6.012943 | Radish_Price_W |
| 6 | 4.835797 | Maximum depth of snow cover_W |
| 7 | 4.555513 | Maximum depth of new snowfall_W |
| 8 | 4.536822 | Spot atmospheric pressure_W |
| 9 | 2.676105 | Precipitation_W |
| 10 | 1.568088 | Import price_W |

11 variables remain

# 3.2 Model Validation

**Separate train set and validation set**

**Validation**

**2018**

**Train**

**2017 2016 2015**

# 3.2 Model Validation

**Model for numeric variables**

Gradient Boosting Regression

KNN Regression

Support Vector Regression

Linear Support Vector Regression

Linear Regression

Random Forest Regression

Decision Tree Regression

# 3.2 Model Validation

**Feature selection using Backward elimination method**

**Best score**

| method | score |
|---|---|
| GradientBoos… | 0.626506 |
| SupportVecto… | 0.57413 |
| Linear_SV_Re… | 0.573937 |
| Linear Regression | 0.567522 |
| KNeighbors_R… | 0.564965 |
| RandomForest… | 0.483727 |
| DecisionTree… | 0.210532 |

Original variable set's score

**Best score**

```
<bound method Series.max of 5    0.640714
1    0.617634
2    0.595854
3    0.569943
0    0.565908
4    0.467732
6    0.214656
```

Only **Radish normal year** is removed

# 3.2 Model Validation

| | Method | Score |
|---|---|---|
| **1** | Gradient Boosting Regression | 0.64004 |
| **2** | KNN Regression | 0.617634 |
| **3** | Support Vector Regression | 0.595854 |
| **4** | Linear Support Vector Regression | 0.571686 |
| **5** | Linear Regression | 0.565908 |
| **6** | Random Forest Regression | 0.503584 |
| **7** | Decision Tree Regression | 0.213532 |

Best R_square is about 64%

# 4. Conclusion

# 4. Conclusion


- Agricultural Transaction data
- Imported Agricultural Product data
- Weather data

- Move total average data
- Processing Missing values
- Merging distributed data
- Merging whole data through date

- Find relationship between the data

- Remove high VIF factors

- Separate train set and validation set
- Using several models
- Feature selection
- Best score: Gradient Boosting Regression(0.64004)

# Thank you