Agriculture Commodities, Prices & Seasons

In the problem, there are two type of datasets

1. Monthly_data_cmo
2. CMO_MSP_Mandi

Monthly_data_cmo:

Data value 62429 observation and 11 columns / variables.

There are 11 variables:

1. APMC : Name Agricultural produce market committee
2. Commodity: Types of commodity
3. Year : Years when data has recorded
4. Month: Month when data has recorded
5. arrivals_in_qtl : How much quantity has arrived
6. min_price: Min price of the crops when data has recorded
7. max_price: Max price of the crops when data has recorded
8. modal_price: Avg price of the crops when data has recorded
9. date: Date when data has recorded
10. district_name: District name of the APMC
11. state_name: State name of the APMC

Exploratory Analysis:

- From given below you can check the min, max and median value of all numeric variables.

```
In [16]: Monthly_data.describe()
```

Out[16]:

	Year	arrivals_in_qtl	min_price	max_price	modal_price
count	62429.000000	6.242900e+04	6.242900e+04	6.242900e+04	62429.000000
mean	2015.337503	6.043088e+03	2.945228e+03	3.688814e+03	3296.003989
std	0.690451	3.470331e+04	1.318396e+04	7.662962e+03	3607.792534
min	2014.000000	1.000000e+00	0.000000e+00	0.000000e+00	0.000000
25%	2015.000000	3.800000e+01	1.250000e+03	1.600000e+03	1450.000000
50%	2015.000000	2.110000e+02	1.976000e+03	2.797000e+03	2425.000000
75%	2016.000000	1.364000e+03	3.900000e+03	4.647000e+03	4257.000000
max	2016.000000	1.450254e+06	3.153038e+06	1.600090e+06	142344.000000

- In above table, the min value of min price, max price and modal price values are zero, which is not possible in real scenario. It's means it's errors. So we will remove those observation where values are zeros.
- Data don't have any missing values in all variables.
- After that we have identified the outliers, which are extreme values and make our result wrong.

- For removing outliers, I have used Density-Based Spatial Clustering of Applications with Noise (DBSCAN). We can't use boxplot here because box plot is use when data is normally distributed.

CMO_MSP_Mandi:

- **Data have 155 observations and 5 columns / variables.**
  1. **Commodity : Types of** commodity
  2. **Year :** Years when data has recorded
  3. **Type : Type of crops**
  4. **Msprice: Minimum Support Price (MSP)**
  5. **msp_filter: Unknown**

```
In [21]: CMO_MSP_Mandi.describe()
```

Out[21]:

	year	msprice	msp_filter
count	155.000000	145.000000	155.0
mean	2014.000000	2822.448276	1.0
std	1.418798	1441.725928	0.0
min	2012.000000	170.000000	1.0
25%	2013.000000	1470.000000	1.0
50%	2014.000000	2970.000000	1.0
75%	2015.000000	4000.000000	1.0
max	2016.000000	6240.000000	1.0

- **Above table is explaining about min and max value of numeric variables.**
- **Data have missing values:**

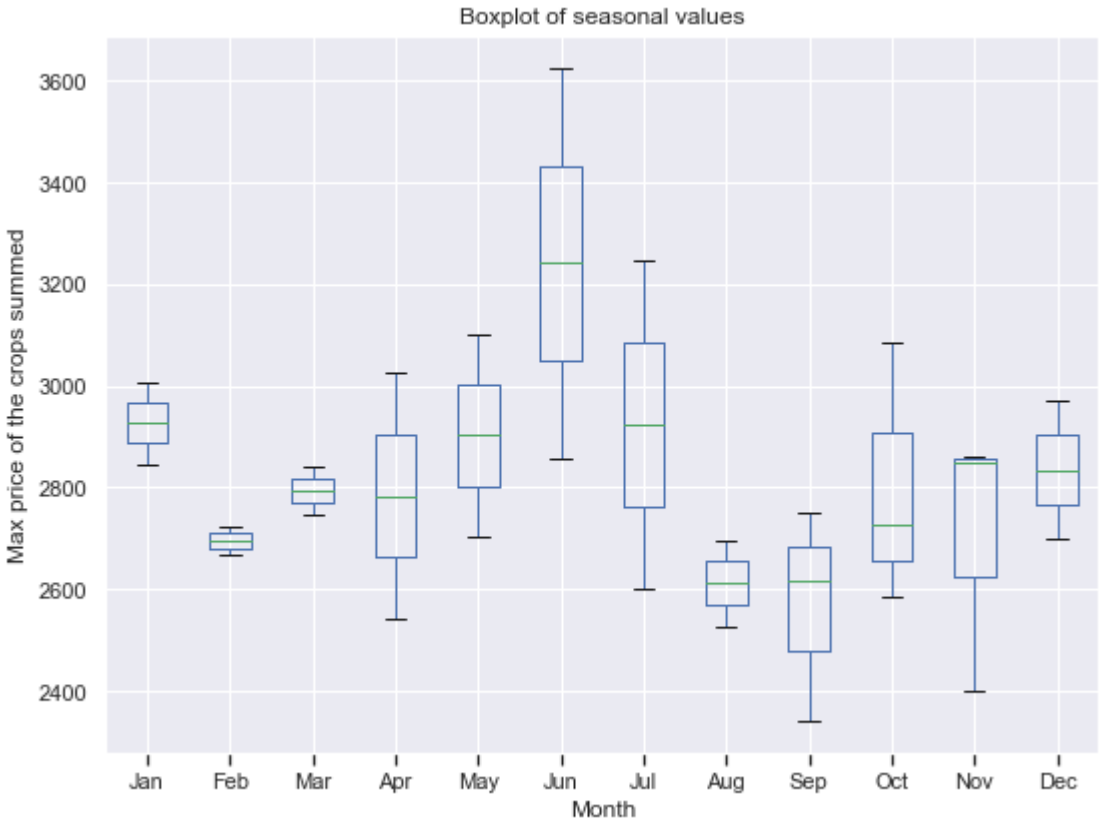
```
In [22]: CMO_MSP_Mandi.isnull().sum()

Out[22]: commodity      0
         year           0
         Type           0
         msprice       10
         msp_filter     0
         dtype: int64
```

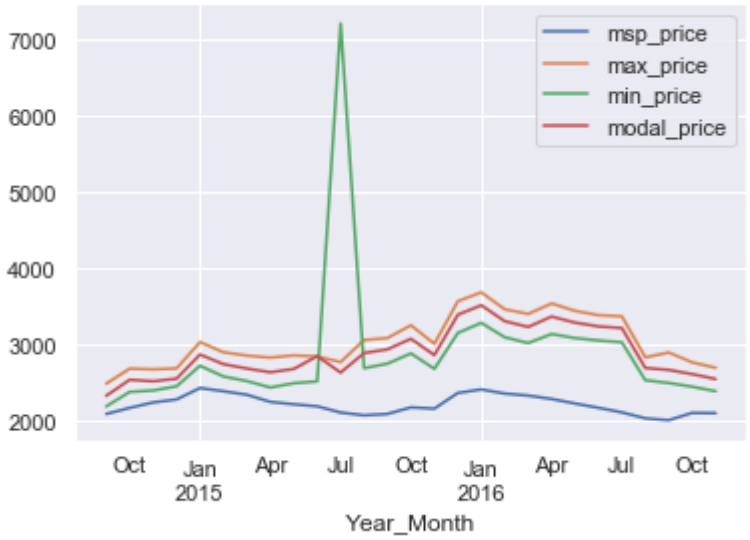
3. **These missing values will impute from** Monthly_data_cmo in to **CMO_MSP_Mandi by mean values according to commodity**.

- Now we have removed the outliers and missing values.
- We know, here time is dependent variable, so we will use time series not regression. Because data is depending on time.
- Now we will check the data are stationary or not?
- For checking the stationary, we will use **Dickey-Fuller test on data.**

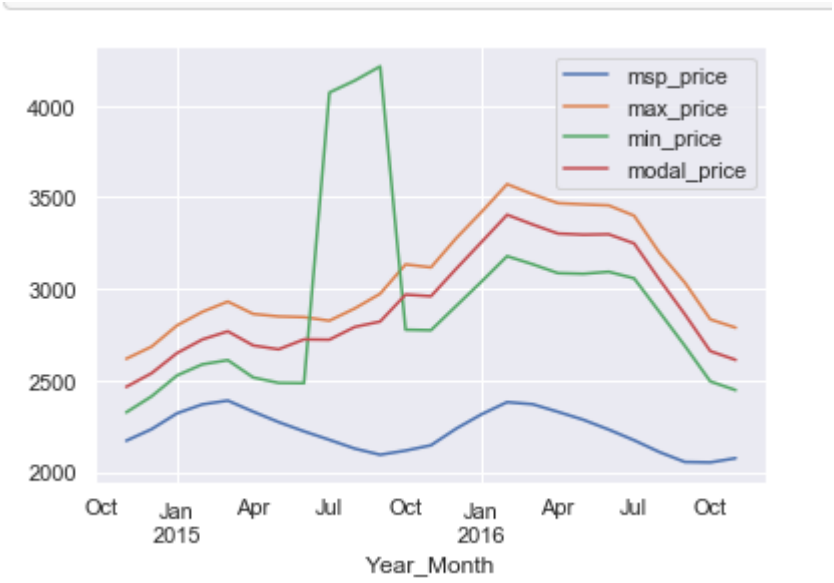
- **Then remove the trend and seasonality from the data.**
- **Below graph is telling** median price of the crops is high at June.



-
- **About the seasonality and trend, I have explained in the ipython notebook.**
- **Compare the APMC with MSP:**



- **Above graph is showing during July min price was very high due to some seasonality effect.**
- **Now we will remove seasonality, we take 3 months means and draw a (de-seasonalised) graph below:**



- **In Below graph , the brown curve above helps us to identify which APMCs have hight and low fluctuations in prices.**