ANALYSIS and METHODOLOGY REPORT

SocialCops Data Science Internship Technical Challenge

R&A | DS Task | APMC/Mandi

Objective:

1. Test and filter outliers.
2. Understand price fluctuations accounting the seasonal effect
    1. Detect seasonality type (multiplicative or additive) for each cluster of APMC and commodities
    2. De-seasonalise prices for each commodity and APMC according to the detected seasonality type
3. Compare prices in APMC/Mandi with MSP(Minimum Support Price)- raw and deseasonalized
4. Flag set of APMC/mandis and commodities with highest price fluctuation across different commodities in each relevant season, and year.

Name : PRANJAL BIYANI

E-mail : pranjalbiyani1996@gmail.com

Mobile :8336060863

BASIC OUTLINE :

- Exploratory Data Analysis and Data Cleansing
- Q-1 Outlier Detection,Testing and Filtering
- Data Preparation for Time-Series Analysis(Seasonality)
- Q-2.1 Seasonality Type Detection
- Q-2.2 Deseasonalizing Price
- Q-3 Comparison of Various Prices
- Q-4 Assessing Price Fluctuations

Data and Python Files

- Cleansed APMC/Mandi and MSP data
- Outlier Removed APMC/Mandi and MSP data
- Seasonality Analysis Data
- Seasonality Type Detected Mandi data
- Deseasonalized Prices Mandi data
- 7 Jupyter (.ipynb) notebook files for each section of above outline

METHODOLOGY

My aim was to understand the trend in the prices of commodities in various APMC/Mandi's and use the results and Analysis to help Maharashtra Government take informed decisions to regulate or control, quantity and price, and also to detect anomalies and corruption in APMC's with unusual pricing or explicit hoarding

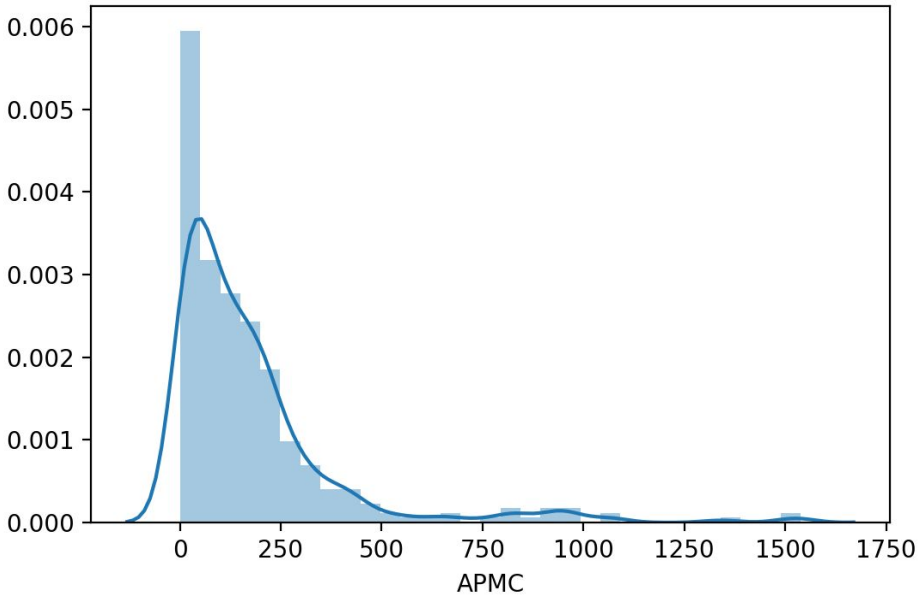
The technique used to conquer the above challenges-

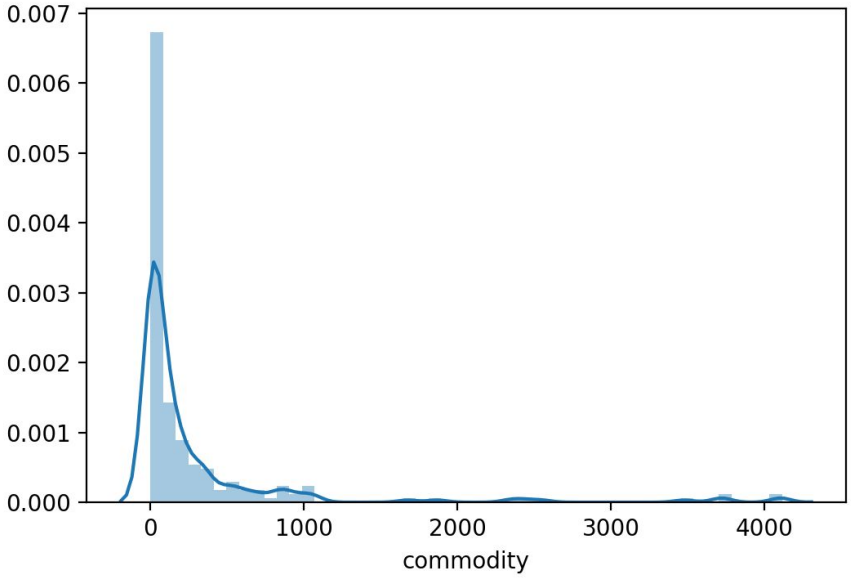
- Cleaning the data ,which had many issues from naming errors to missing values to limited or high degree Skeweness
- Removing the outliers, from removing abnormal prices due to incorrect data to illogical data values, using standard outlier filtering processes and visualising to confirm
- Analysis of Seasonality in the Prices and removing it to understand general trend in the price movements
- Comparing Prices (MSP,RAW and Deseasonalized) to observe APMC, as well as commodity pricing trends and come up with standard solutions to fix any anomalies if exist
- Finally, observing Price fluctuations across years, seasons for the government to focus on specific APMC,commodity pairs , by taking decisions based on price fluctuations

1. EXPLORATORY DATA ANALYSIS

- Challenges-
    - Both the datasets had different naming styles for columns as well as data
        - Ex: **Commodity : commodity, Sugar-cane :Sugarcane , e.t.c**
    - MSP data had 10 missing values for commodity msprice
    - MSP data and APMC data differ on alphabetical grounds (Capital-Small)
    - APMC data started at 2014 and MSP started from 2012(2 years missing data)
    - Significant commodities in APMC data didn't have a msprice value in MSP data (204,32)
    - Max_price and Min_price columns in APMC dta had zero values, Illogical
    - Highly Skewed distribution of data for both APMC's and Commodities
        - Few commodities with high frequencies, still understandable
        - Few APMC's with high data, infact with just 1 commodity, Unusual
- Solutions -
    - All the column names were changed to small letters and observed naming errors were removed
    - Missing values were replaced with mean values from the past history of prices
    - All the values in commodity column for both datasets were converted to small letters
    - Further analysis where msprice was needed, only those commodities were considered that had existence in both datasets
    - Max_price and Min_price with zero values were removed, since they were small compared to the entire data
- **SUGGESTIONS -**
    - Data collection must be done uniformly addressing the skewness in the data
    - The two different sets of datasets collected must be aligned to have common grounds for analysis and predictions

Plots for Skewness in Distribution of Data for APMC’s and Commodities





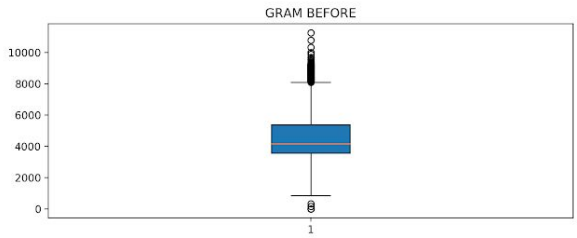
2. Outlier Detection, Testing and Filtering

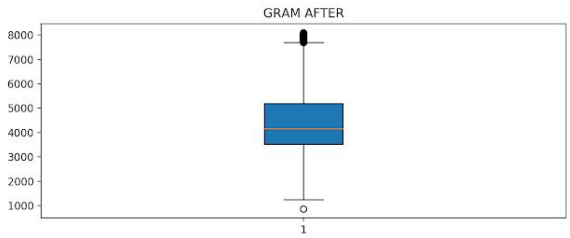
Two methods were used -

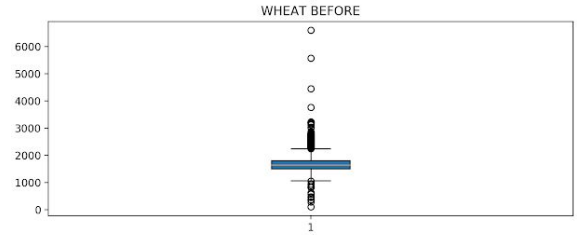
- Outlier Replacement - Replace values with maximum or minimum of the

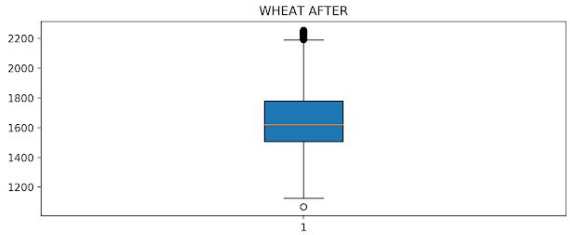
  Standard : (q1-1.5*IQR , q3+1.5*IQR) range
- Outlier Removal - Remove values not lying in the Range

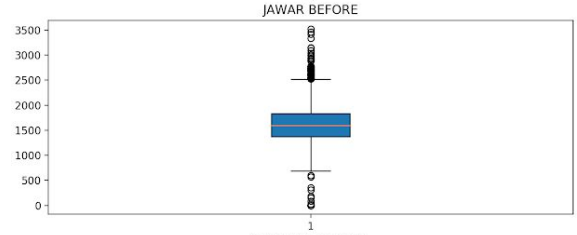
- Challenges -
  - Data had to be eliminated keeping in mind the loss factor to be <4~5%
  - Unusual data where min_price>max_price had to be removed
  - Naming errors in MSP data, ex: Soybean:Soyabean had to be fixed
  - Data had to be grouped appropriately to successfully find and remove outliers
- Solutions -
  - With loss < 4%, APMC data outliers were removed
  - With loss >14% ,MSP data outliers were replaced
  - Naming anomalies were fixed
  - Data was plotted to observe the comparison before and after removal of outliers
  - Unusually priced data was removed
- Observations-
  - APMC data had many outliers, can be because of corruption of faulty data collection
  - MSP data was fine , need not be fixed or changed

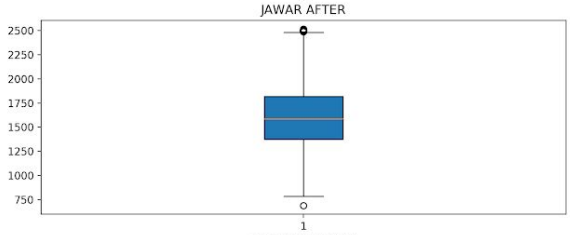

GRAM BEFORE
10000
8000
6000
4000
2000
0
1

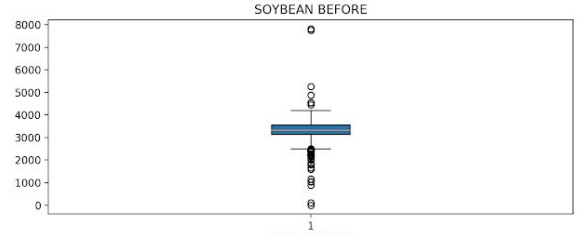

GRAM AFTER
8000
7000
6000
5000
4000
3000
2000
1000
1

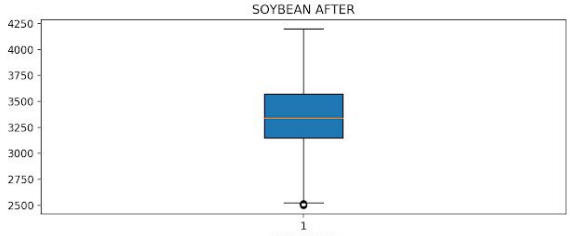

WHEAT BEFORE
6000
5000
4000
3000
2000
1000
0
1

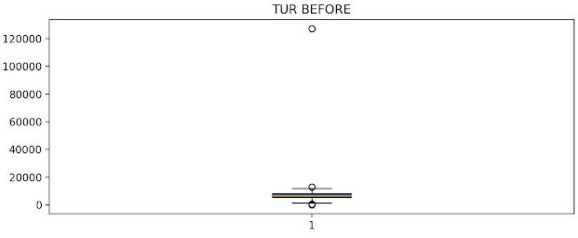

WHEAT AFTER
2200
2000
1800
1600
1400
1200
1

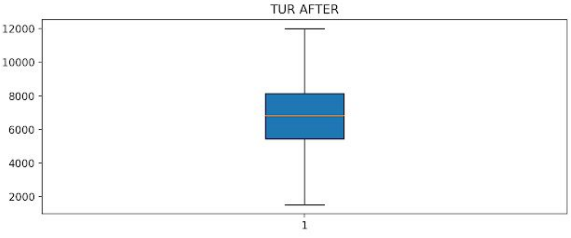

JAWAR BEFORE
3500
3000
2500
2000
1500
1000
500
0
1

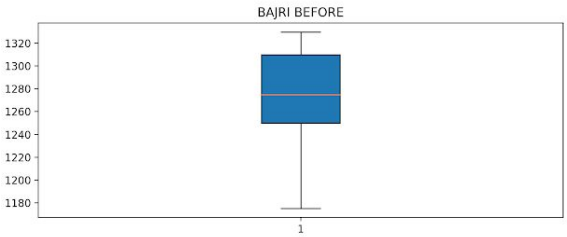

JAWAR AFTER
2500
2250
2000
1750
1500
1250
1000
750
1

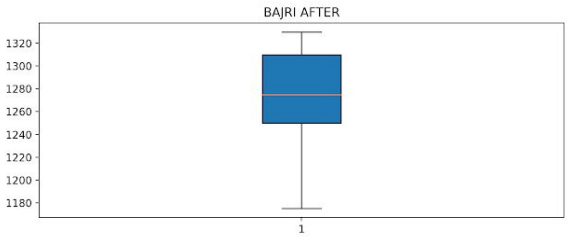

SOYBEAN BEFORE
8000
7000
6000
5000
4000
3000
2000
1000
0
1

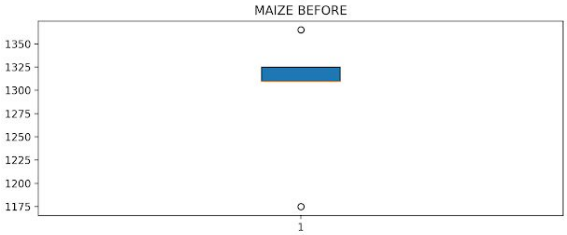

SOYBEAN AFTER
4250
4000
3750
3500
3250
3000
2750
2500
1

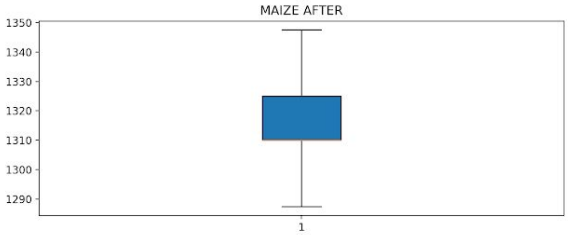

TUR BEFORE
120000
100000
80000
60000
40000
20000
0
1

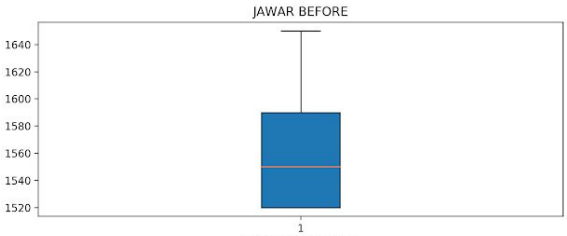

TUR AFTER
12000
10000
8000
6000
4000
2000
1

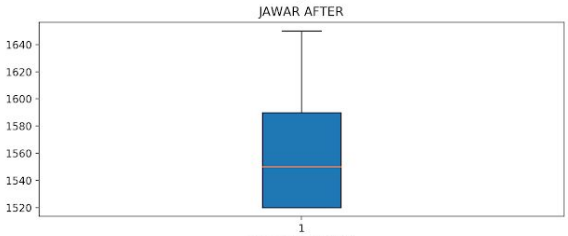

BAJRI BEFORE
1320
1300
1280
1260
1240
1220
1200
1180
1

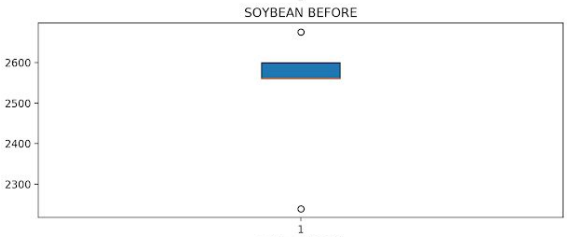

BAJRI AFTER
1320
1300
1280
1260
1240
1220
1200
1180
1

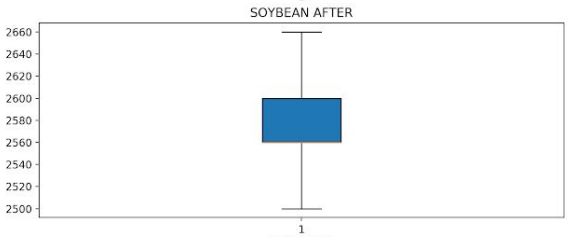

MAIZE BEFORE
1350
1325
1300
1275
1250
1225
1200
1175
1

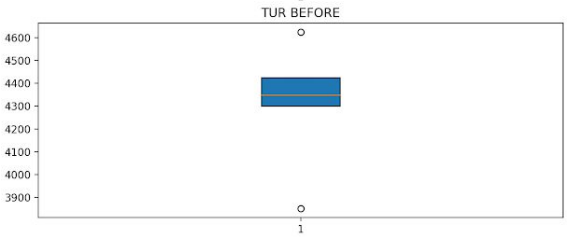

MAIZE AFTER
1350
1340
1330
1320
1310
1300
1290
1

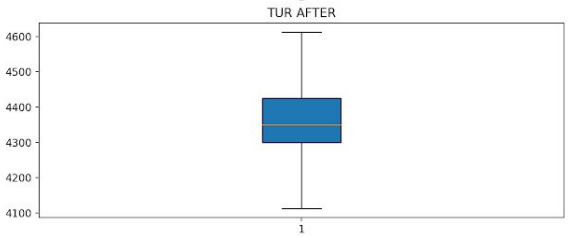

JAWAR BEFORE
1640
1620
1600
1580
1560
1540
1520
1


JAWAR AFTER
1640
1620
1600
1580
1560
1540
1520
1


SOYBEAN BEFORE
2600
2500
2400
2300
1


SOYBEAN AFTER
2660
2640
2620
2600
2580
2560
2540
2520
2500
1


TUR BEFORE
4600
4500
4400
4300
4200
4100
4000
3900
1


TUR AFTER
4600
4500
4400
4300
4200
4100
1

3. Data Preparation for Seasonality Analysis

- Challenges-
    - Date column’s datatype needed conversion to work with Time-Series Analysis
    - Minimum frequency, ex: 12 , required to use in time-series analysis, which was not available for all (APMC,Commodity) pair
- Solutions -
    - Basic pandas functionality solved data type issue
    - Dataset reduced from ~60k to ~22k size, after filtering for minimum frequency of (APMC,Commodity) pair data
    - Only, this data was exported for further Analysis
- Suggestions-
    - Level or Set a standard for data collection that enables Data Collection over a particular time frequency throughout the year for proper analysis

4. Seasonality Type Detection

- METHOD -
    - Use seasonal_decompose from the statools library of Python
    - Use it’s residuals and eventually their Auto-Correlation-Function values to detect Type of Seasonality (Multiplicative or Additive)
    - Collect unique pairs of APMC and Commodity and give them as an input to a function which uses the original dataset to get all the records for that pair and uses seasonal_decompose and acf to detect type
    - Also, create another function that is hard coded for the above functionality(in process)
    - Check for Seasonality by plotting Auto-correlation grams
    - Export the Dataset for Deseasonalizing Prices

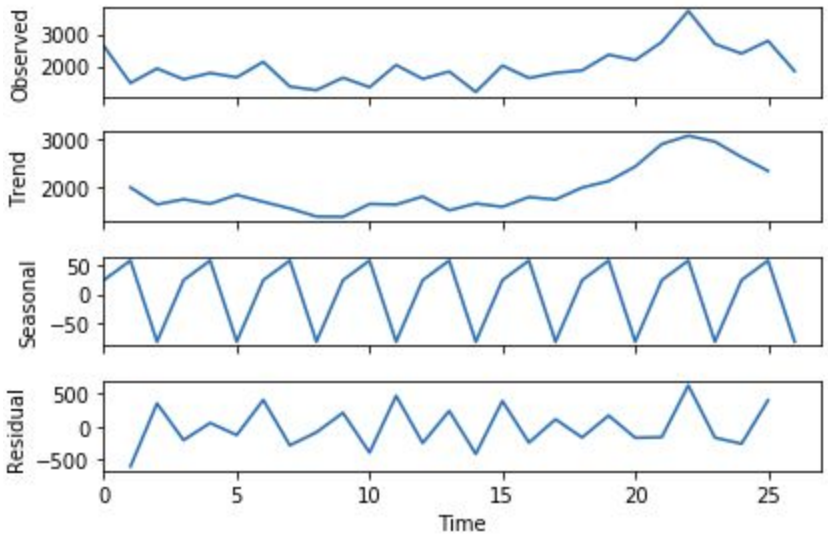
5. Deseasonalize Prices based on detected Seasonality Type

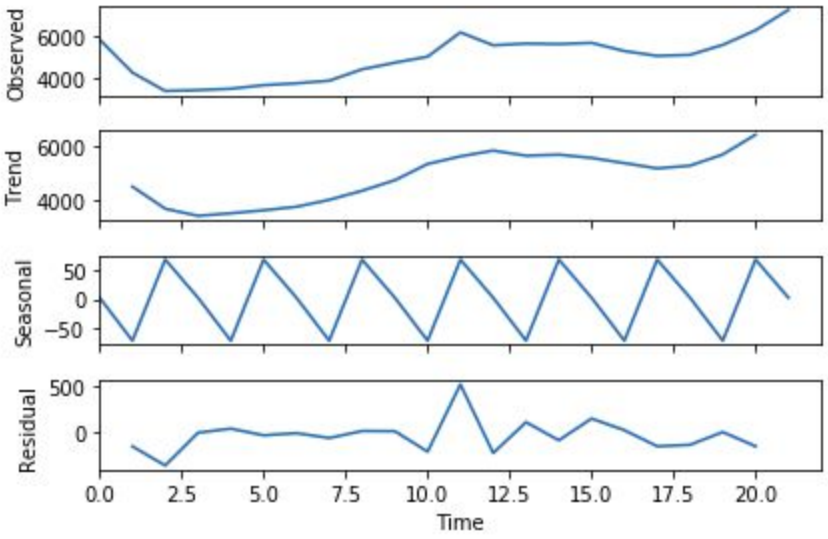
METHOD-

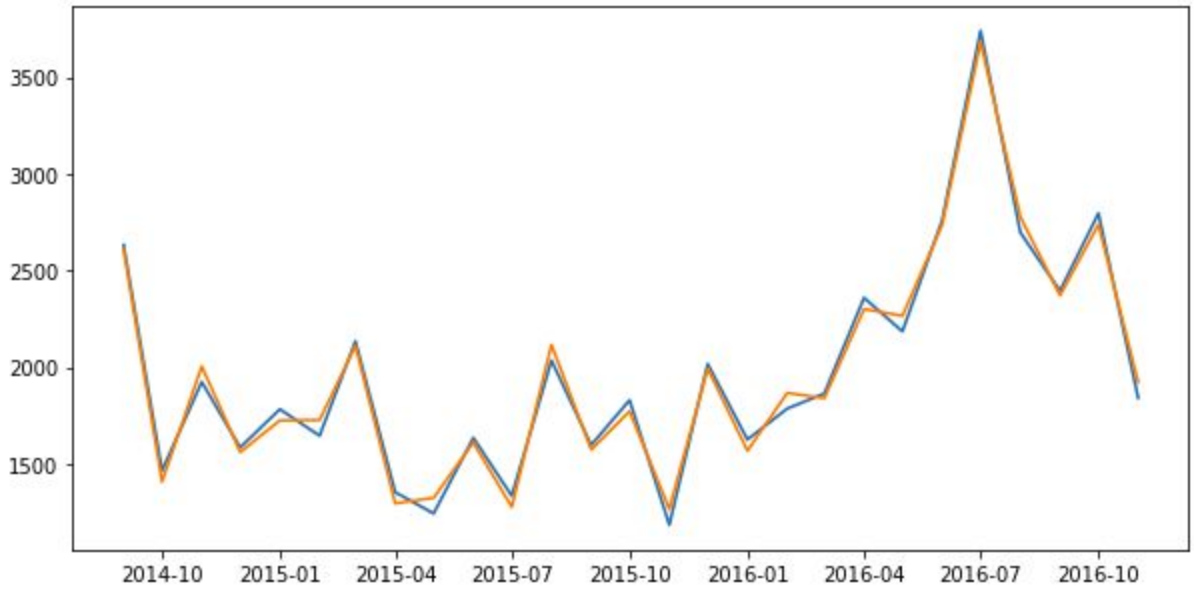
- Again, using seasonal_decompose to get the seasonal component of the time-series data for a particular pair of APMC,Commodity
- Removing the seasonal component from the price for all the pairs
- Substantiating and Statistically testing Stationarity (Deseasonalization) in the prices using the standard Augmented Dickey Fuller Test from the adfuller method, statools
- Plotting the result from the seasonal_decompose to effectively know if we actually have seasonality component, which did exist
- Plotting the Raw(modal_price) and Deseasonalized Price to observe smoothening of the line, indicating removal of Seasonality and Stationarity in Time-Series

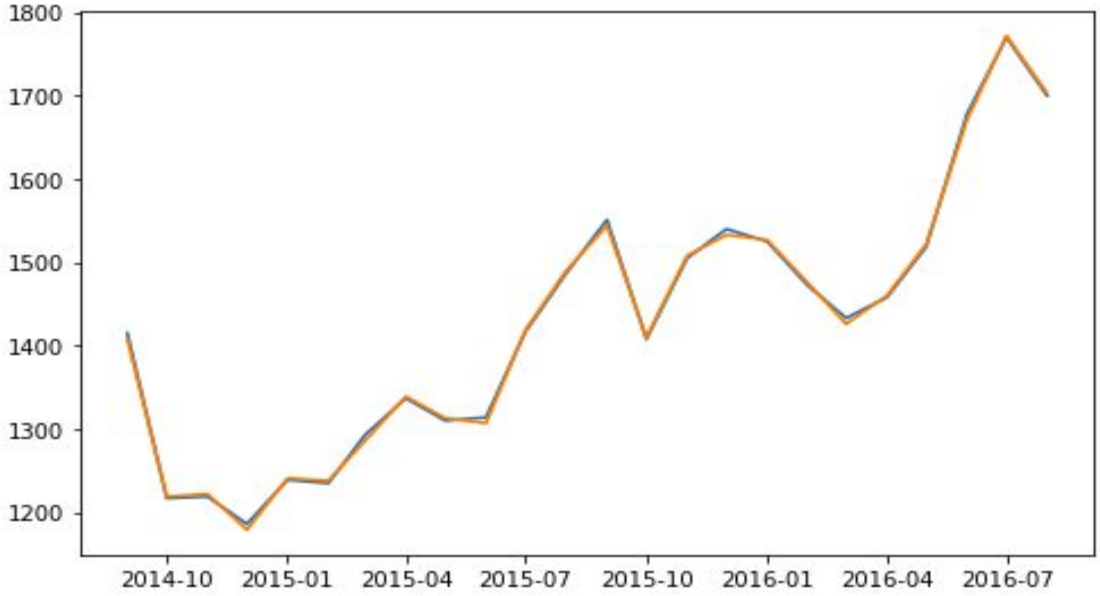
OBSERVATIONS -

- Some APMC,Commodity pairs have high seasonality ex: seasonal crops like tomatoes
- Some APMC,Commodity pairs have low seasonality ex: all-year crops like maize


Observed
3000
2000
Trend
3000
2000
Seasonal
50
0
−50
Residual
500
0
−500
0
5
10
15
20
25
Time


Observed
6000
4000
Trend
6000
4000
Seasonal
50
0
−50
Residual
500
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Time


3500
3000
2500
2000
1500
2014-10
2015-01
2015-04
2015-07
2015-10
2016-01
2016-04
2016-07
2016-10


1800
1700
1600
1500
1400
1300
1200
2014-10
2015-01
2015-04
2015-07
2015-10
2016-01
2016-04
2016-07

6. Price Comparison (Raw,MSP,Deseasonalized)

Three types-

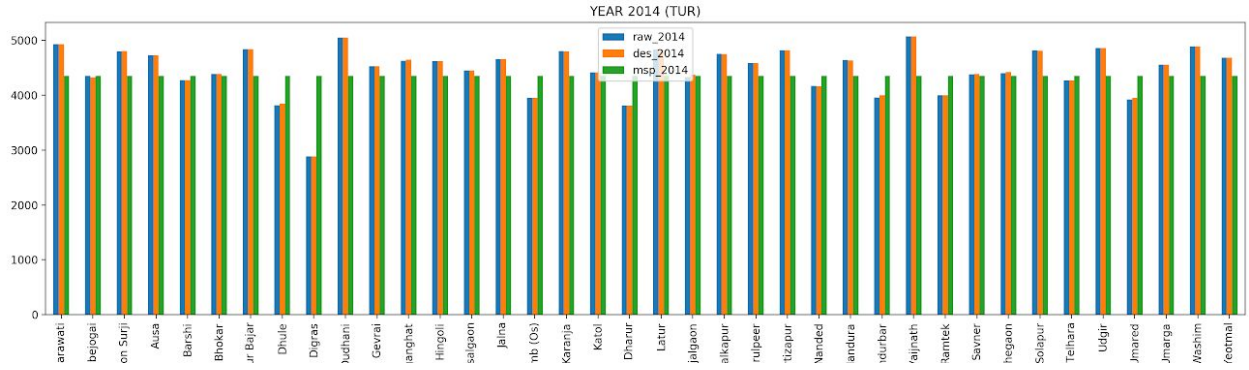
- Yearly average prices for different commodities of the same APMC
- Yearly average prices for different APMC's of particular commodity
- Time-Series comparison of a pair of APMC and commodity

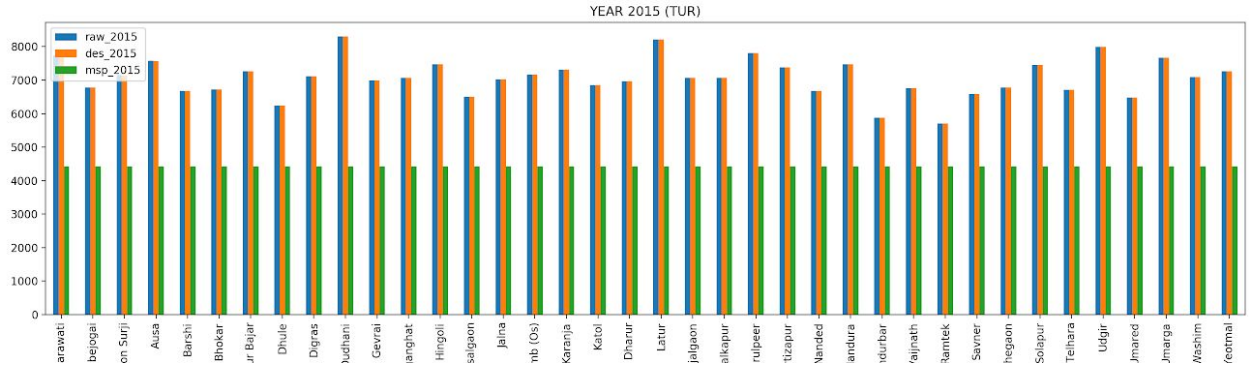
PROBLEMS with Data -

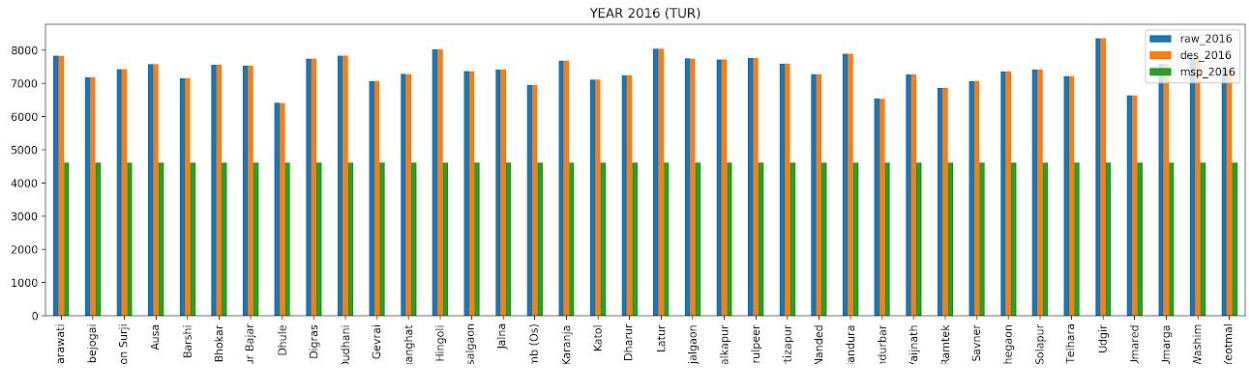
- We only have data for three years 2014,2015,2016
- We only have 17 common commodities between MSP and APMC data
- MSP data , didn't have all three years data

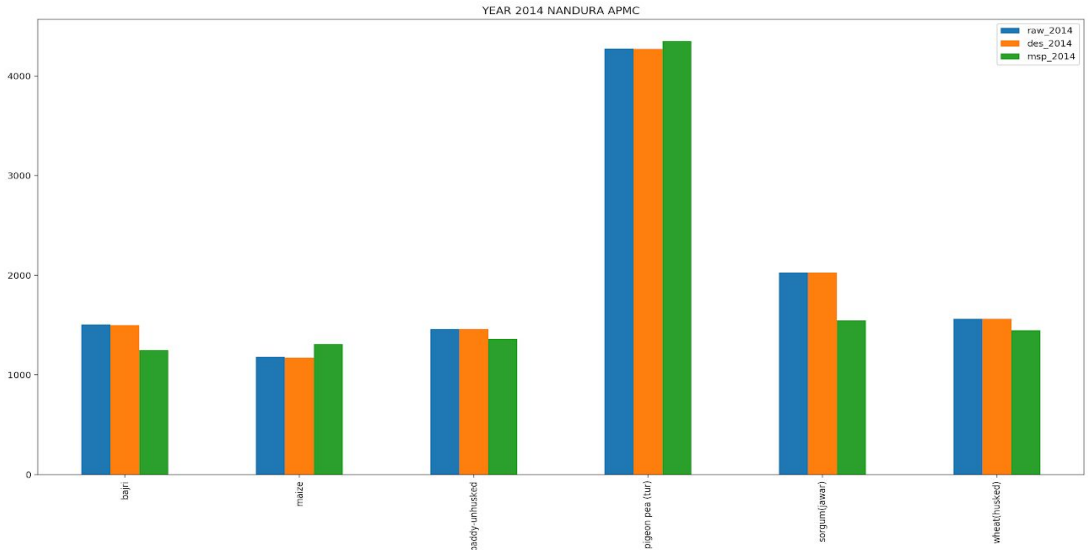
OBSERVATIONS -

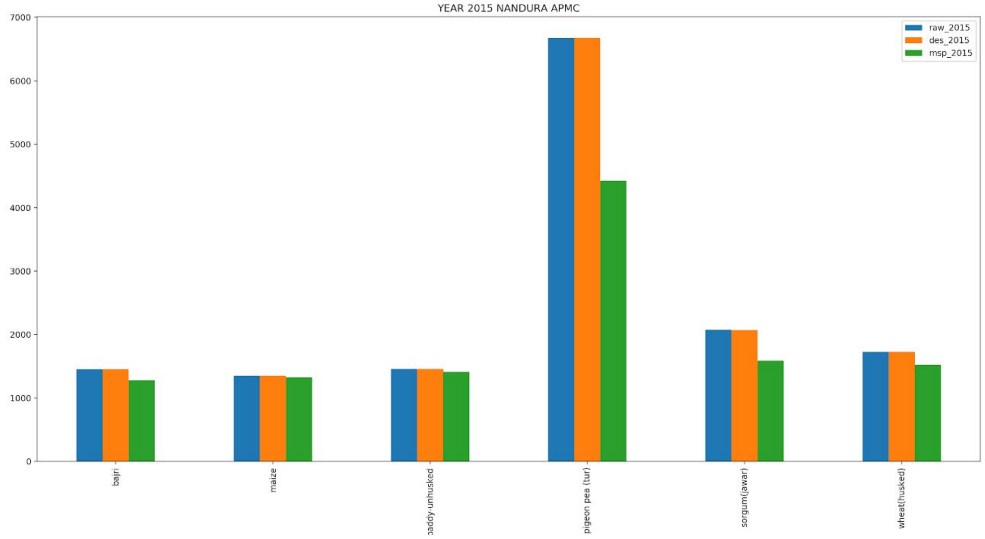
- Some APMC's have commodity prices below MSP
- Some APMC's have price much higher than MSP
- 2014 had many pricing anomalies, 2015 and 2016 had few comparatively
- Commodities with high MSP's have higher variations and differences b/w raw and msp
- Some commodities fluctuate above and below MSP, ex: Paddy
- Some commodity prices always stay above MSP

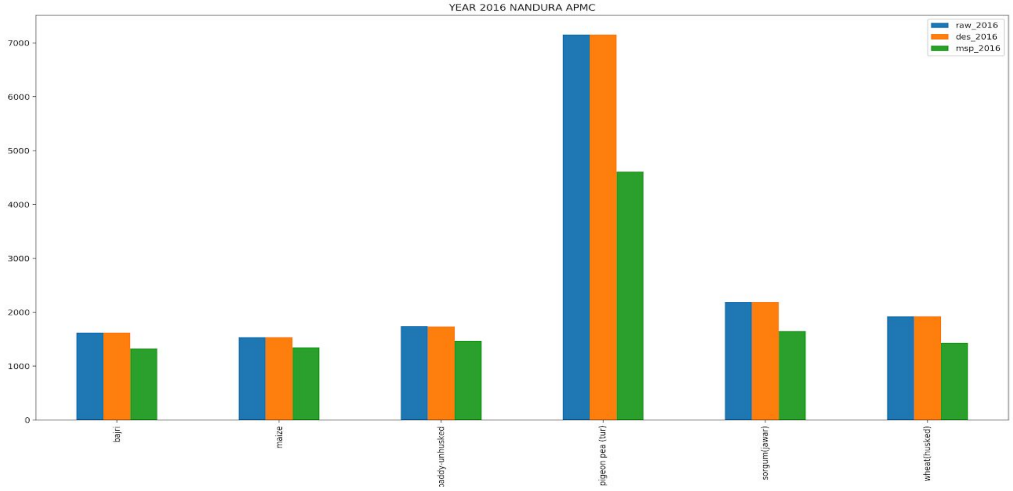

YEAR 2014 (TUR)
raw_2014
des_2014
msp_2014
0
1000
2000
3000
4000
5000
arawati
bejogai
on Surji
Ausa
Barshi
Bhokar
ur Bajar
Dhule
Digras
udhani
Gevrai
anghat
Hingoli
salgaon
Jalna
nb (Os)
Karanja
Katol
Dharur
Latur
jalgaon
alkapur
rulpeer
tizapur
Nanded
andura
ndurbar
aijnath
Ramtek
Savner
hegaon
Solapur
Telhara
Udgir
Jmared
Jmarga
Washim
Yeotmal

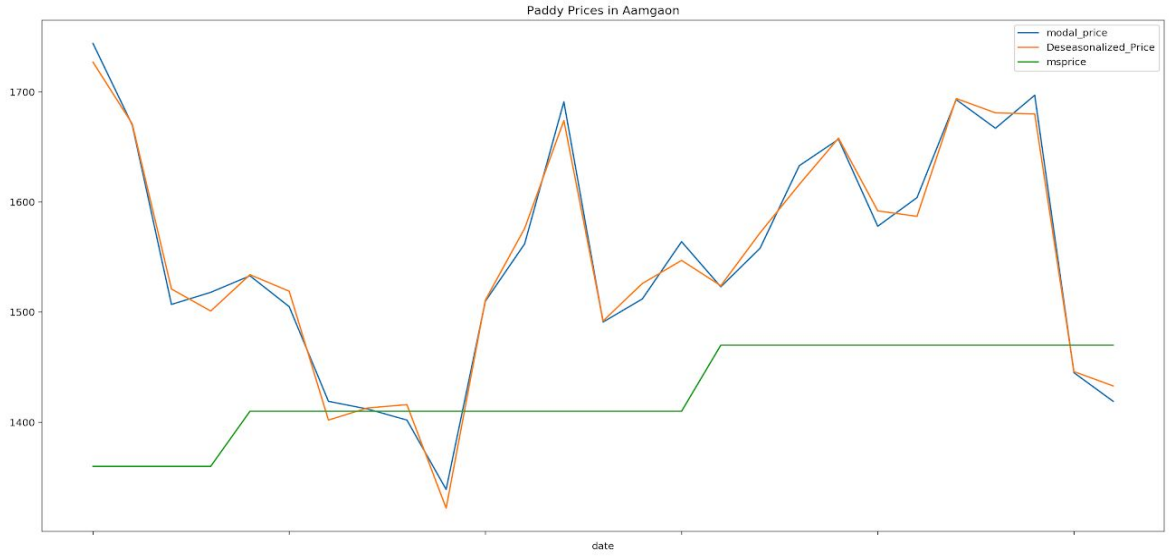

YEAR 2015 (TUR)
raw_2015
des_2015
msp_2015
0
1000
2000
3000
4000
5000
6000
7000
8000
arawati
bejogai
on Surji
Ausa
Barshi
Bhokar
ur Bajar
Dhule
Digras
udhani
Gevrai
anghat
Hingoli
salgaon
Jalna
nb (Os)
Karanja
Katol
Dharur
Latur
jalgaon
alkapur
rulpeer
tizapur
Nanded
andura
ndurbar
aijnath
Ramtek
Savner
hegaon
Solapur
Telhara
Udgir
Jmared
Jmarga
Washim
Yeotmal


YEAR 2016 (TUR)
raw_2016
des_2016
msp_2016
0
1000
2000
3000
4000
5000
6000
7000
8000
arawati
bejogai
on Surji
Ausa
Barshi
Bhokar
ur Bajar
Dhule
Digras
udhani
Gevrai
anghat
Hingoli
salgaon
Jalna
nb (Os)
Karanja
Katol
Dharur
Latur
jalgaon
alkapur
rulpeer
tizapur
Nanded
andura
ndurbar
aijnath
Ramtek
Savner
hegaon
Solapur
Telhara
Udgir
Jmared
Jmarga
Washim
Yeotmal


YEAR 2014 NANDURA APMC
raw_2014
des_2014
msp_2014
4000
3000
2000
1000
0
bajri
maize
paddy-unhusked
pigeon pea (tur)
sorgum(jawar)
wheat(husked)


YEAR 2015 NANDURA APMC
raw_2015
des_2015
msp_2015
7000
6000
5000
4000
3000
2000
1000
0
bajri
maize
paddy-unhusked
pigeon pea (tur)
sorgum(jawar)
wheat(husked)


YEAR 2016 NANDURA APMC
raw_2016
des_2016
msp_2016
7000
6000
5000
4000
3000
2000
1000
0
bajri
maize
paddy-unhusked
pigeon pea (tur)
sorgum(jawar)
wheat(husked)


Paddy Prices in Aamgaon
modal_price
Deseasonalized_Price
msprice
1700
1600
1500
1400
date


Soybean Prices in Akhadabalapur
modal_price
Deseasonalized_Price
msprice
3600
3400
3200
3000
2800
2600
date

7. Highest Price Fluctuations (Yearly , Seasonally)

METHOD -

- Accounting for Fluctuations Annually for 2014,2015,2016
- Fluctuations quarterly (Seasonal assuming Quarter)
- Fluctuations by Crop Type (Kharif,Rabi,Others assuming seasonality in Production)

By evaluating Fluctuation based on the difference between the maximum and minimum price yearly,quarterly or by Type, we find APMC,Commodity pairs with highest Fluctuations by the above methods constraints

SUGGESTIONS -

- For yearly fluctuation we can intersect pairs of all years to get perfect and fixed pairs
- For quarterly pairs, we can be cautious with specific pairs ahead of time every year
- For Type method pairs, we can ensure regularity of produce and supply according to the cropping season type

End Notes -

- Major issues with the Analysis were-
    - Lack of Structured data
    - Improper Alignment of one with the other
    - Matching percentage was very low for a diverse analysis

- Scope of Improvement
    - We could use Natural Language Processing to reduce naming errors by using similarity detection algorithms
    - We could increase the size of data by using discounted models for price by developing a discounting model from the time-series
    - We could also setup a digital unit to record data in a standard time-frame

- Personal Thoughts
    - A more structured query would have been easy to answer with respect to the specific needs of the Government
    - Questions were a little confusing to properly adapt, define and answer

THANKYOU

-Pranjal Biyani