Time Series: CommoditIES INterim report

Monthly Trend Analysis

**Moderator:**

**Mr Srikar Muppidi**

**Report Submitted by:**

* Kevin Rajkumar
* Sukalp Shivam
* Manas Ghosh
* Gaurav Garhewal

**INTRODUCTION**

**About the Project:**

The retail price of key commodities is basically taken from Open Government Data Platform of India where prices of every commodities is varying with the different regions of India.

**Objective:**

The objective is to study the monthly trend of most important commodities and then forecast for the next 12 months in each region of India.

**Background:**

The dataset contains retail prices for key commodities recorded around 75 key market centres in India. The granularity of the dataset is at day level. The country is divided into 5 regions i.e. (North, South, East, West, North-East) and each region has different a centre. The dataset contains the Prices/rates of commodities of different regions with respect to time which is from 1997 to 2015.The dataset has 9 commodities i.e. (Onion, Rice, Tea Loose, Tur/Arhar Dal, Sugar, Salt Pack, Milk, Tomato, Sunflower oil) and for each commodity we have prices per kg. The dataset is talking about the trend of each commodity which is increasing and decreasing according to the data. So, we are basically analysing the trend of each commodity with respect to time and will forecast the trend for next 12 months.

**Visualization/Forecast:**

Since, this is essentially a time series related dataset, python time series libraries based on statsmodels.api will be used. So, we are using time series concept to get the result by using different algorithm and then comparing the accuracies. Hence, we are analysing for the reason for why the prices get high for each commodity in different region. Hence, we are trying to do a detailed analysis for the change in prices and the trend of each commodity in different region and the essential factors affecting those.

**Trend Analysis:**

Trend analysis is a technique used in [technical analysis](https://www.investopedia.com/terms/t/technicalanalysis.asp) that attempts to predict the future stock price movements based on recently observed trend data. In the given dataset of commodities, we are checking the trend in the region-wise price of various commodities over the period of 19 years. We will create a Time index in the dataset and use that to plot a timeseries line which will provide us the visualization of the trend over the years in term of price of that particular commodity. By looking at the movement of price in past observations we are going to predict for future points in time. The important thing to look in the trend analysis is the slope, it is giving to the timeseries line. Based on that information and seasonality (if present) we will be able to predict future points.

**Introduction to Agriculture Policy:**

Price policy plays a pioneer role in the economic development of a country. It is an important instrument for providing incentives to farmers for motivating them to go in for production-oriented investment and technology.

In a developing country like India where majority of the population devotes 2/3 of its expenditure on food alone and where majority of the population is engaged in agricultural sector, prices affect both income and consumption of the cultivators.

**Factors that affecting Commodity prices:**

1. **Production related –** Commodities are capital-intensive products i.e. they are influenced by natural factors like weather conditions, crop diseases, size of land cultivated and factors related to production like labour patterns, development in the tools and technologies used. Other than these there are factors like the economic and political environment which manifest itself in the form of trade constraints, subsidies, taxes to mention a few. Altogether these factors affect the cost of producing the commodity and the demand for it in a market where there is more than one participant.
2. **Global Inflation:** The rise in essential food items in India is Primarily because of the increasing commodity prices abroad, increase in fuel prices and fertilizers, which in turn affect the local produce by increasing input costs.
3. **Less space for cultivation:** With an increase in population, there is also an increase in the demand for vegetables and pulses. Increase in population has also led to increased urbanisation. Manufacturing, energy and service industries are all competing for land, water and human resources. With less availability of land, the prices for agricultural land are rising, leading to increased costs of agriculture produce.
4. **Costs involved in storage –** There are two types of costs involved in storing commodities. One is the financial cost and the other is the cost of physical storage and they both need to be factored in when computing the forward prices.
5. **Seasonality –** Some examples of such factors include weather related patterns, operational risk, climatic conditions and politics.
6. **Increased cost of transportation:** With increase in fuel prices, the transportation charges also increase, leading to a rise in prices of all commodities, and vegetables are no exception.
7. **Many mediators:** In India’s trading community, the end product reaches the consumer after passing through various mediators or middlemen. Each mediator tries to get profits by increasing the original cost and the end price becomes very higher than the actual prices.
8. **Supply chain mismanagement**: There has been mismanagement in the supply chain of vegetables and pulses from the farmers to the consumers. According to reports, the difference in wholesale and retail prices is anywhere between 40% and 60% and this margin is more within cities where there are the wholesale markets.

**MISSING VALUE TREATMENT**

In Statistics, **missing data**, or **missing values**, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be drawn from the data.

The concept of missing values is important to understand in order to successfully manage data.  If the missing values are not handled properly, we may end up drawing an inaccurate inference about the data. Due to improper handling, the result obtained us will differ from ones where the missing values are present. Time series data are become one of some data that most likely to have missing values in it.

**Need to fill null values in time series data**

Time Series is a data that observed in regular interval of time. Time series form can be found as a continuous data or discrete because observed in a regular interval of time such as daily, monthly and yearly etc. Some examples of continuous and discrete time series are sinusoidal signal, which is continuous, while daily stock prices, temperature data are discrete data. Data measurements are conducted several times with different conditions and missing data occurs due to several problems which is also known as “missingness mechanism”

1. **Missing completely at random (MCAR)**
2. **Missing at random (MAR):** a variable is missing at random if the probability of missingness is depending only on available information.
3. **Not missing at random (NMAR):** the missingness probability is depending on the variable itself

**Handling missing data**

Handling of Missing values is not a new topic of studies in data mining or data analysis since there are so many methods, approach, and techniques are proposed by the previous researchers, from the simplest one to the complicated one and their own advantage and drawbacks. Some basic methods are proposed such as ignoring, deleting, zero or mean or mode estimation methods. These methods above have the simplicity but only effective for low percentage of missing values, but in bigger percentage of missing values, the result will affect the result of overall analysis and can also result in biased predictions. The main ‬disadvantage of discarding incomplete observation is the loss of efficiency and biased estimation result especially when the ‬data missingness is systematic. The quality of data mining or analysis is influenced by the quality of the data. Therefore, the data that contain missing values should be estimated to provide complete case of data to get expected result from the data.‬ Below are some ways where we can impute missing values to a time series data accordingly‬‬

Mean, median, mode, random sample imputation

Data without trend and seasonality

Missing Values in Time Series

Data with trend and without seasonality

Imputation

Linear interpolation

Data with trend and seasonality

Seasonal adjustment + interpolation

**Data without trend and seasonality:**

One of the easiest method to impute or estimating missing values to get a complete sample is replacing each of the missing value with mean or median or mode of the observed data for the variables which is also known as unconditional mean/median/mode imputation.

**Data with trend and without seasonality:**

Interpolation is the process of finding a value between two points on a line or curve. The last valid value before the missing value and the first value after the missing value is used for interpolation.

Line drawn with the equation: (Y - Y0 / X - X0) = (Y1 - Y0)/(X1 – X0)

Point1: (X0, Y0)

Point2: (X1, Y1)

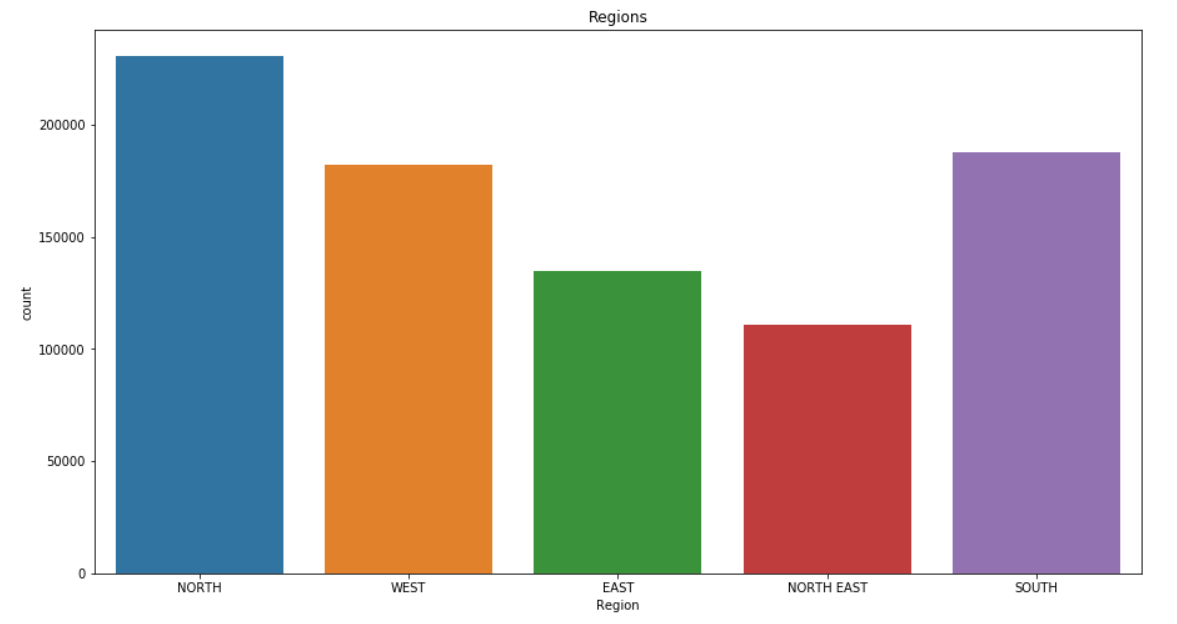
**Data with trend and seasonality:**

Seasonal adjustment is a statistical method to removes the seasonal component of the time series that exhibits a seasonal pattern. It is usually done when wanting to analyse the trend, and cyclical deviations from trend, of a time series independently of the seasonal components. And then interpolation is done.

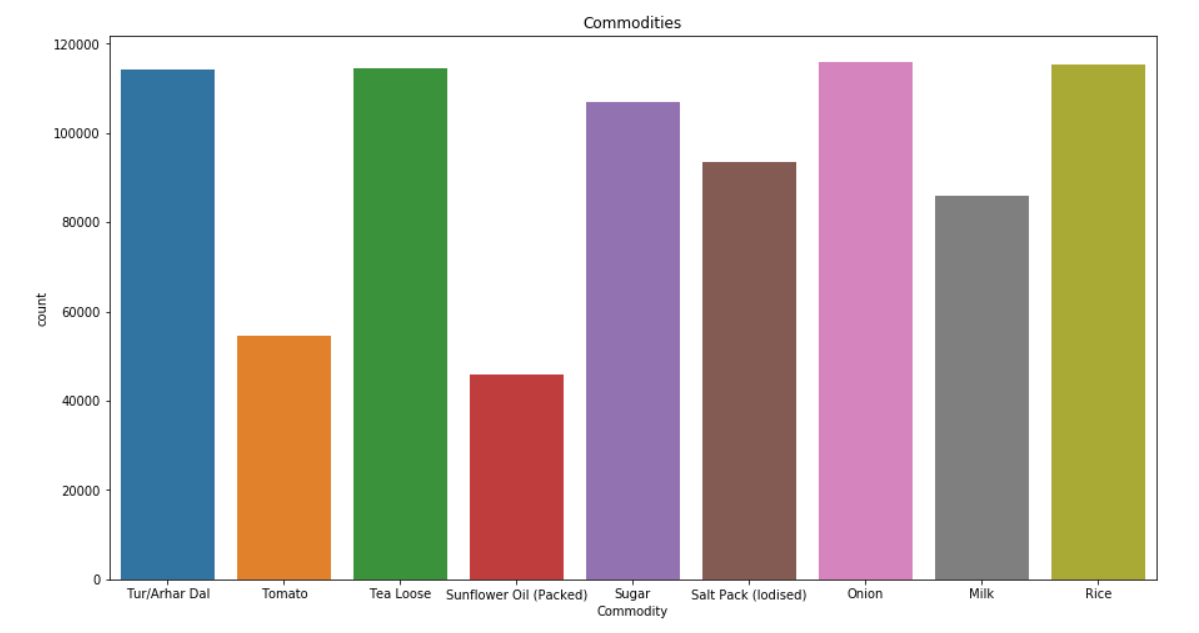
**EXPLORATORY DATA ANALYSIS**

The objective is to do region-wise analysis of the top commodities.

**Region-wise Commodities**:



**Count of Commodities across all regions**:



There are five regions to do analysis and 9 commodities, we are selecting 5 top commodities on the basis of above count information and popularity of those commodities in the market. So, we will do analysis on following commodities in this project:

1. Onion 115991 rows
2. Rice 115370 rows
3. Tea Loose 114495 rows
4. Tur/Arhar Dal 114224 rows
5. Sugar 106953 rows

We are working on the commodities dataset, where we are trying to look at the pattern in the price throughout multiple years on which we are doing a monthly trend analysis. So before plotting the graph, we need to do certain feature engineering, which will involve following steps:

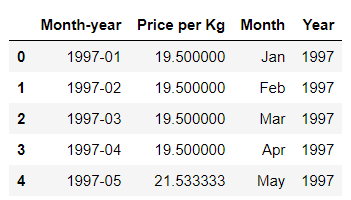
* Load the dataset into a variable
* Extract date info from the date feature of the dataset
* Use this date, to create new feature called Time index
* Pick region and commodity to do analysis
* Create new data frame for the analysis containing only required region and commodity
* Extract month-year from this Time index
* Use group by function to level the data frame to the monthly level
* Create month and year features in the data frame by extracting them from time index for further analysis

As, we have 5 regions and 5 commodities, we are creating 25 different models for prediction. However, EDA part will be same for all of them, so we will select one region and one commodity to visualize the trend in this report.

Region- North

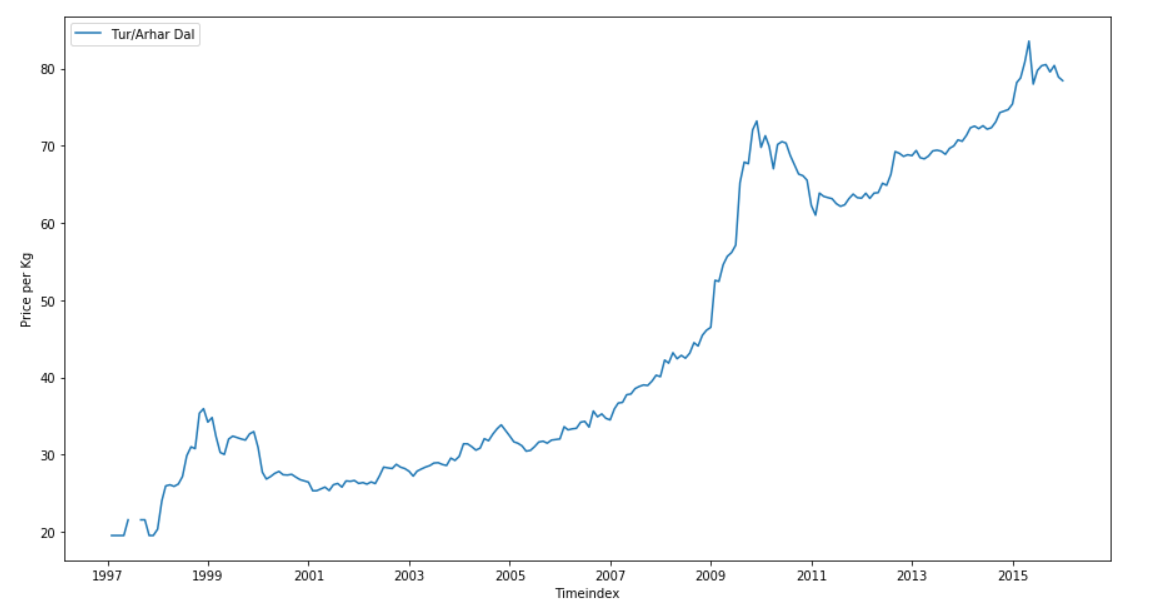
Commodity- Tur/Arhar Dal

After doing feature engineering steps, our new data frame will look like:

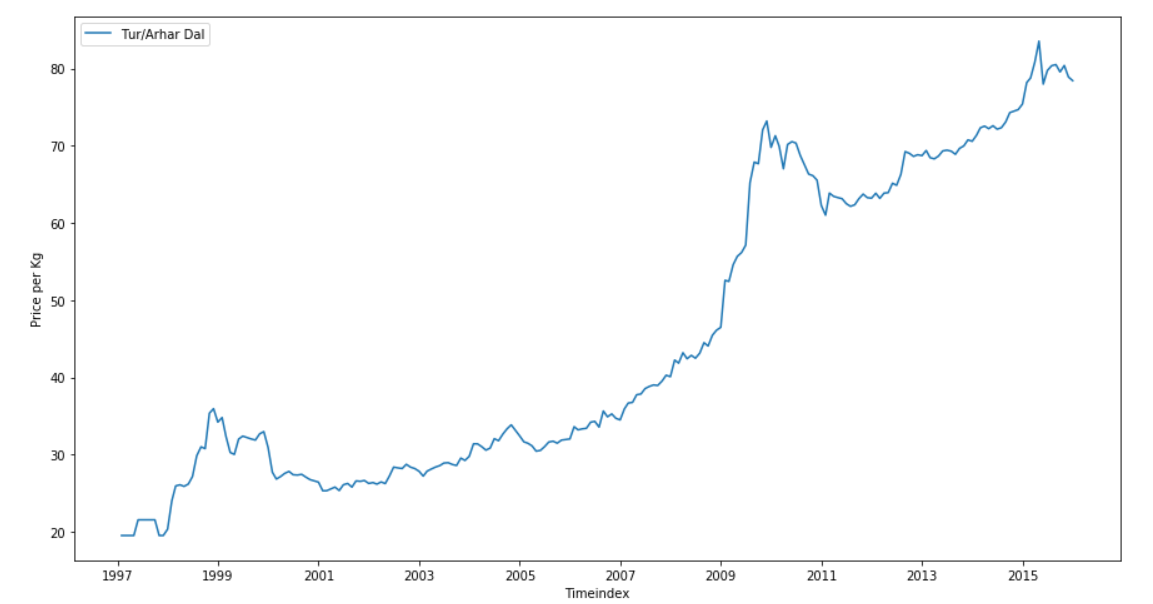


Here we need to look for the missing months in the dataset. The best way to check for missing months at this level is by sorting the dataset by Month-year column and then check for first and last value. Calculate the number of months that are supposed to be between first and last value and check it with the number of rows in this dataset. In this example the first value is 1997-01 and last value is 2015-12. The total number of months that are supposed to be between these values are 12\*19=228 months. But in this data frame only 226 values are present. So, we need to find where these values are missing and insert a row there and in place of price per kg, show null so that we can interpolate it later.

This thing can be done by setting Month-year as index and reindex it with complete month-year series. This will give null in place of missing rows. After this we will plot the timeseries:



As you can clearly see that there is a null value present in year 1997, we can also see that there is a trend in this dataset but no seasonality, so we will go with linear interpolation method of filling the missing value. The output will be:



These are the datapoints that we will use to create models and forecast for next 12 months in each region for top five commodities.

**MODEL TECHNIQUES**

A time series is a sequence of numerical data points in successive order. In investing, a time series tracks the movement of the chosen data points, such as a security’s price, over a specified period of time with data points recorded at regular intervals. There is no minimum or maximum amount of time that must be included, allowing the data to be gathered in a way that provides the information being sought by the investor or analyst examining the activity. A time series can be taken on any variable that changes over time. In investing, it is common to use a time series to track the price of a security over time.

**There are two main goals of Time series Analysis:**

1. Identifying the nature of the phenomenon represented by the sequence of observations.
2. Forecasting (predicting future values of the time series variables)

Both of these goals require that the pattern of observed time series data is identified and more or less formally described. Once the pattern is established, we can interpret and integrate it with other data (i.e., use it in our theory of the investigated phenomenon, e.g., seasonal commodity prices).

Time series forecasting uses information regarding historical values and associated patterns to predict future activity. Most often, this relates to trend analysis, cyclical fluctuation analysis and issues of seasonality. As with all forecasting methods, success is not guaranteed. The Dataset is having different Commodities where prices are varying with the regions and centres. The time series forecasting methods can be classified as:

In Decomposition technique, we decompose the timeseries into trend, seasonality and irregular components and in Regression technique we make use of auto-regression and moving average

**TIMESERIES DECOMPOSITION:**

In this technique, we are dividing the series into three components:

* Trend Component: a long-term monotonic change of the average level of the time series.
* Seasonal Component: fluctuations in time series that recur during specific time periods.
* Residual Component that represents all the influences on the time series that are not explained by the other two components.

There are three models that we have in decomposition technique:

**Simple Exponential Smoothing (SES):**

SES is a good for forecasting data with *no clear trend or seasonal pattern*. Forecasts are calculated using weighted averages, which means the largest weights are associated with most recent observations, while the smallest weights are associated with the oldest observations:

**Holt’s Method**:

Holt extended simple exponential smoothing to allow the *forecasting of data with trends.* Holt’s method involves a forecast equation and two smoothing equations (one for the level and one for the trend):

**Holt-Winters’ Method:**

Holt-Winters’ Method is suitable for *data with trends and seasonality* which includes a seasonality smoothing parameter γ. Holt-Winters’ method includes a trend component and a seasonal component. There are two variations to this method:

* Additive method: the seasonal variations are roughly constant through the series.
* Multiplicative method: the seasonal variations are changing proportionally to the level of the series.

**TIMESERIES REGRESSION:**

Time series regression is a statistical method for predicting a future response based on the response history (known as autoregressive dynamics) and the transfer of dynamics from relevant predictors. Time series regression can help us understand and predict the behaviour of dynamic systems from experimental or observational data. Time series regression is commonly used for modelling and forecasting of economic, financial, and biological systems. There are two models of Timeseries regression we will be exploring in this project:

**Autoregressive Integrated Moving Average (ARIMA):**

An ARIMA model is a class of statistical models for analyzing and forecasting time series data. It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skilful time series forecasts.

ARIMA is an acronym that stands for Autoregressive Integrated Moving Average. It is a generalization of the simpler Autoregressive Moving Average and adds the notion of integration. This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:

* **AR**: *Autoregression*. A model that uses the dependent relationship between an observation and some number of lagged observations.
* **I**: *Integrated*. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
* **MA**: *Moving Average*. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.

Each of these components are explicitly specified in the model as a parameter. A standard notation is used of ARIMA(p,d,q) where the parameters are substituted with integer values to quickly indicate the specific ARIMA model being used.

The parameters of the ARIMA model are defined as follows:

* **p**: The number of lag observations included in the model, also called the lag order.
* **d**: The number of times that the raw observations are differenced, also called the degree of differencing.
* **q**: The size of the moving average window, also called the order of moving average.

**Seasonal Autoregressive Integrated Moving-Average (SARIMA):**

SARIMA method models the next step in the sequence as a linear function of the differenced observations, errors, differenced seasonal observations, and seasonal errors at prior time steps. It combines the ARIMA model with the ability to perform the same autoregression, differencing and moving average modelling at the seasonal level.

The notation for the model involves specifying the order for the AR(p), I(d), and MA(q) models as parameters to an ARIMA function and AR(P), I(D), MA(Q) and m parameters at the seasonal level, time steps in each season. A SARIMA model can be used to develop AR, MA, ARMA and ARIMA models. The method is suitable for univariate time series with trend and/or seasonal components.