ISSN: 0474-9030

*Our Heritage*

Vol-68-Issue-15

January-2020

UGC Care Listed Journal

# Market Analysis of Agricultural Products using Data Science tools

Mr. Rahul Borate[1], Mr. Rahul Navale[2], Dr. Milind Godase[3]
*[1]Assistant, [2,3]Professor*
*MCA Dept.,*
*Sinhgad Institute of Management, Pune*

***Abstract:***
*Agriculture is the backbone of Indian economy. Today data science plays an important role in farming domain. In farming sector where farmers and agribusinesses have to make numerous decisions every day and elaborate complexities involves the various factors influencing them. An essential issue for farming planning purpose is the accurate time for the various crops production involved in the planning. Data science tools are essential toward for accomplishing realistic and effective solutions for this difficulty. Unpredictability in market, input levels, combinations and product prices has made it all the more applicable for farmers to use information and get help to make significant farming decisions. This paper focuses on the analysis of the agricultural product market data and finding best possible parameters to take full advantage of the crop production using data science tools such as Linear Regression, density plot and pie-chart. Taking out the large amount of existing crop, market data and analyzing new, non-experimental data optimizes the production and makes farming more flexible according to market trend.*

**Background**

As India is predominately Agriculture County. According to FAO (Food and Agriculture Organization) of the United Nations, India is the world's largest producer of milk, pulses and jute, and ranks as the second largest producer of rice, wheat, sugarcane, groundnut, vegetables, fruit and cotton. It is also one of the leading producers of spices, fish, poultry, plantation crops and livestock. Worth $ 2.1 trillion, India is the world's third largest economy after the US and China [1]. In India, agriculture is unique business of crop production. Crop production is always depends on climate conditions, irrigation availability, rainy conditions. Even if in a particular year due to favorable climatic conditions, sufficient irrigation and satisfactory rainy conditions the crop production is high, but there is always doubt that framer's crop production will get good amount of market price. Hence in this paper we are looking at a solution that finds best period to cultivate a particular crop by the farmer in a year with the help of certain data science tools. This kind of investigation would help farmer to have better planning for his crop production which would results in best market place and price. This paper focuses on the analysis of the pomegranate production and market data to find best possible parameters to take full advantage of the crop production and market trends using data science tools such as Linear Regression, density plot and pie-chart. For this analysis we have considered the 13 districts of Maharashtra which used to cultivate and produce pomegranate in huge amount.

**Literature Survey**

There are different forecasting methodologies developed and evaluated by the researchers all over the world in the field of agriculture. Some of such studies are:Researchers like Ramesh and Vishnu Vardhan are analysed the agriculture data for the years 1965–2009 in the district East Godavari of Andhra Pradesh, India. Rain fall data is clustered into 4 clusters by adopting the K means clustering method. Multiple linear regressions (MLR) is the method used to model the linear relationship between a dependent variable and one or more independent variables. The dependent variable is rainfall and independent variables are year, area of sowing, production. Purpose of this work is to find suitable data models that

achieve high accuracy and a high generality in terms of yield prediction capabilities [3].Bangladesh offers several varieties of rice which has different cropping season [4]. For this a prior study of climate (effect on temperature and rainfall) in Bangladesh and its effect on agricultural production of rice has been done. Then this study was being taken into regression analysis with temperature and rainfall. Temperature puts an adverse consequence on the crop production. The data has been taken from the "Bangladesh Agricultural Research Council (BARC)" for past 20 years with 7 attributes: "rainfall", "max and min temperature", "sunlight", "speed of wind", "humidity" and "cloud-coverage".In Pre-processing, the whole dataset was divided in 3 month duration phases (March to June, July to October, November to February). For this duration, the average for every attribute has been taken and associated with it. This pre-processing has been done for each kind of rice variety. In clustering, the different pre-processed table has been analyzed to find the sharable group of region based on similar weather attribute. Soil characteristics are studied and analyzed using data mining techniques. As an example, the k-means clustering is used for clustering soils in combination with GPS based technologies [5]. Authors like Alberto Gonzalez-Sanchez, Juan Frausto-Solis and Waldo Ojeda-Bustamante have done extensive study on predictive ability of machine learning techniques such as multiple linear regressions, regression trees, artificial neural network, support vector regression and k-nearest neighbor for crop yield production [6]. Wheat yield prediction using machine learning and advanced sensing techniques has done by Pantazi, DimitriosMoshou, Thomas Alexandridis and Abdul Mounem-Mouazen [7]. The aim of their work is to predict within field variation in wheat yield, based on on-line multi-layer soil data, and satellite imagery crop growth characteristics. Supervised self-organizing maps capable of handling existent information from different soil and crop sensors by utilizing an unsupervised learning algorithm were used.The software tool 'Crop Advisor' has been developed by S. Veenadhari, B. Misra and CD Singh [8] is an user friendly web page for predicting the influence of climatic parameters on the crop yields. C4.5 algorithm is used to find out the most influencing climatic parameter on the crop yields of selected crops in selected districts of Madhya Pradesh.

**Methods**

The aim of proposed work is to analyze the agriculture production and market data using data science tools. In proposed work, pomegranate production data has been collected from following sources:
[https://numerical.co.in/numerons/collection/58df68014e976264035a1f85] [9],
[http://nrcpomegranate.icar.gov.in/EPublications] [10],
Pomegranate market data of APMC Pune [http://www.puneapmc.org/rates.aspx] [11],
Input dataset consist of 3 year data with following parameters namely: year, State-Maharashtra (13 districts), District, crop (Pomegranate), area (in hectares), production (in tonnes), input (in kg) and price (RS/kg) required.

**Pie chart:**

R pie chart is created using the **pie()** function [12] which takes positive numbers as a vector input. The additional parameters are used to control appearance of pie charts in R are labels, color, title etc.

**Syntax R Pie chart**

The basic syntax for creating a pie chart using the R is:
**pie (x, labels, radius, main, col, clockwise)**
Following is the description of the parameters used:
**x** is a vector containing the numeric values used in the pie chart.
**labels** is used to give description to the slices.
**radius** indicates the radius of the circle of the pie chart.(value between -1 and +1).
**main** indicates the title of the chart.

**col** indicates the color palette.
**clockwise** is a logical value indicating if the slices are drawn clockwise or anti clockwise.
Since most of the data scientist collect is quantitative, data tables and charts are usually used to organize the information. Graphs are created from data tables. They allow the investigator to get a visual image of the observations, which simplifies interpretation and drawing conclusions

**Linear Regression:**
Linear regression is one of the most commonly used predictive modelling techniques [13]. The aim of linear regression is to find a mathematical equation for a continuous response variable Y as a function of one or more X variable(s). So you can use this regression model to predict the Y when only the X is known.
This mathematical equation can be generalized as follows:
$\mathbf{Y=\beta1+\beta2X+\epsilon}$
Where,
**β1** is the intercept **and β2X** is the slope.
Collectively, they are called regression coefficients and ϵ is the error term.
**Evaluation Methods**

Pomegranate cultivation in Maharashtra (area, production)

1.32 lakh hectare under pomegranate plantation, accounting for nearly two-thirds of the total area under pomegranate cultivation in the country, Approximate number of farmers engaged in pomegranate cultivation 2,00,000 [9].
Experimental Analysis of Production District wise:
For this analysis we considering the last three years data for pomegranate cultivation data and we calculate the average area and production based on it we calculate tonnes per hectare production.

| District | Production | Area | tonnes/hectare |
|---|---|---|---|
| Beed | 12345 | 2845 | 4.34 |
| Aurangabad | 31800 | 7300 | 4.36 |
| Buldana | 5479 | 842 | 6.51 |
| Jalna | 19100 | 2424 | 7.88 |
| Solapur | 169798 | 20033 | 8.48 |
| Pune | 108061 | 12010 | 9.00 |
| Osmanabad | 24790 | 2550 | 9.72 |
| Ahmednagar | 162096 | 16113 | 10.06 |
| Satara | 43618 | 3947 | 11.05 |
| Dhule | 108004 | 8308 | 13.00 |
| Latur | 7705 | 571 | 13.49 |
| Nashik | 679378 | 48527 | 14.00 |
| Sangli | 114841 | 7656 | 15.00 |

Table 1: Average pomegranate production tonnes/hectare

We divide the districts in three main clusters based on tonnes per hectare production capacity like

**Cluster 1:** Low production(0-5 tonne/hectare)
**Cluster 2 :** Moderate production (6-10tonnes per hectare)
**Cluster 3:** High production (11-15 tonnes per hectare)

| **Low Production Districts (0 to 5 tonnes/hectare)** | **Moderate Production Districts (6 to 10 tonnes/hectare)** | **High Production Districts ( 11to15 tonnes/hectare)** |
|---|---|---|
| Beed<br>Aurangabad | Buldana<br>Jalna<br>Solapur<br>Pune<br>Osmanabad | Ahmednagar<br>Satara<br>Dhule<br>Latur<br>Nashik<br>Sangli |

**Table 2: Cluster wise districts**

**Experimental Results**
**R codes to visualize the tonnes/ hectare production by using pie chart:**

```
> tonnes_per_hectare <- c(4.34,4.36,6.51,7.88,8.48,9.00,
        +  9.72,10.06,11.05,13.00,13.49,14.00,15.00 )
> labels <- c("Beed","Aurangabad","Buldana","Jalna",
        + "solapur","Pune","Osmanabad","Ahmednagar","Satara",
        + "Dhule","Latur","Nashik","Sangli")
> pie(tonnes_per_hectare, labels=tonnes_per_hectare ,
        + main = "Districtwise Tonnes Per Hectare Production Chart",
        + col = rainbow(length(x)),clockwise=TRUE)
> legend("topright", c("Beed","Aurangabad","Buldana",
        + "Jalna","solapur","Pune","Osmanabad","Ahmednagar","Satara",
        + "Dhule","Latur","Nashik","Sangli"), cex = 0.4,
        +    fill = rainbow(length(x)))
```

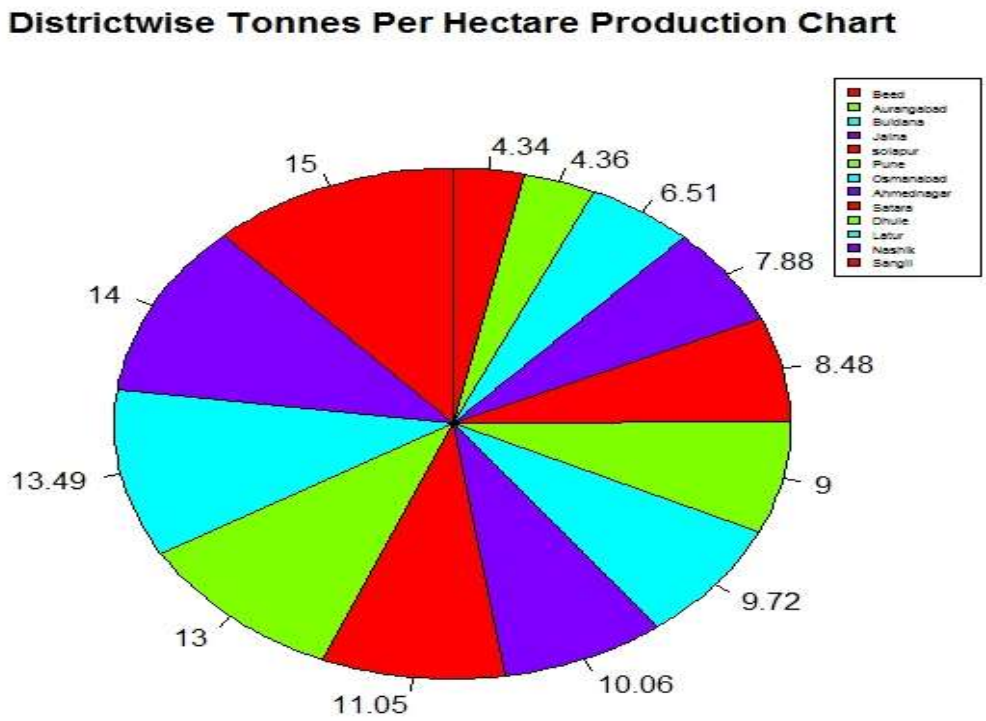
ISSN: 0474-9030

*Our Heritage*

Vol-68-Issue-15

January-2020

UGC Care Listed Journal




**Experimental Analysis of Input and Price by using linear Regression:**

For market analysis purpose we are taking data from apmc pune market data where we observe the input level and modal price on daily basis of pomegranate and draw some conclusion by using linear regression model of R language.

The aim of this eperiment is to build a simple regression model that we can use to predict Price (Mprice) by establishing a statistically significant linear relationship with Input level (Input).

We are building market_data dataset , that makes it convenient to demonstrate linear regression in a simple and easy to understand fashion. You can access this dataset simply by typing in market_data in R console. You will find that it consists of 50 observations(rows) and 2 variables (columns) – Input and Mprice. Lets print out the first six observations here.

**Scatter plot**: Visualize the linear relationship between the Input and Price

Scatter plots can help visualize any linear relationships between the Price (Mprice) variable and Input (Input) variables.

```
> market_data <- read.csv("apmc.csv", header = TRUE)
> head(market_data)
     Date  Input Mprice
1 1-Nov-19  80010   87.5
2 3-Nov-19  70000   87.5
3 4-Nov-19 140000   82.5
4 5-Nov-19 146000   82.5
5 6-Nov-19 152012   87.5
6 7-Nov-19 164000   60.0
> scatter.smooth(x=market_data$Input, y=market_data$Mprice, main="Input ~ Price")  # scatterplot
```

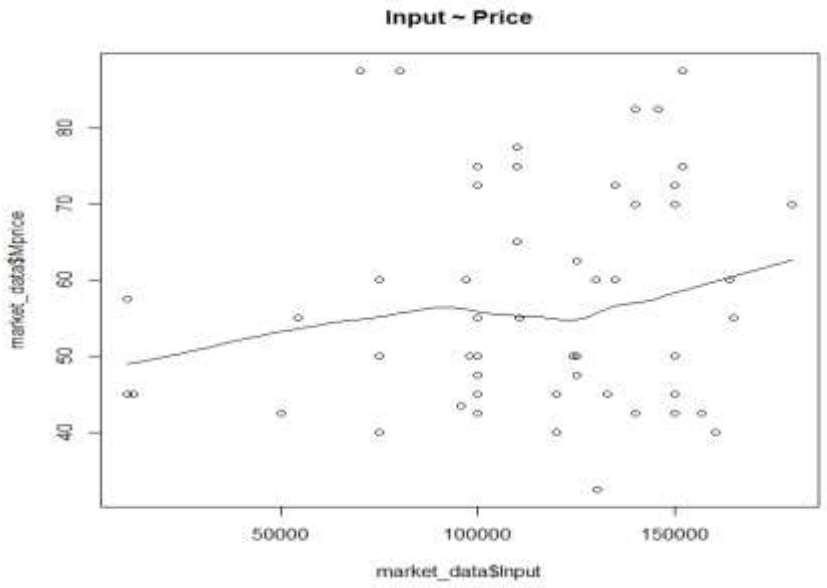
ISSN: 0474-9030

*Our Heritage*

Vol-68-Issue-15

January-2020

UGC Care Listed Journal




The scatter plot along with the smoothing line above suggests a linearly increasing relationship between the Input and Mprice variables. This is a good thing, because, one of the underlying assumptions in linear regression is that the relationship between the Price and Input variables is linear and additive.

**Density plot:**

To see the distribution of the price variable. Ideally, a close to normal distribution (a bell shaped curve), without being skewed to the left or right is preferred.

```
library(e1071)
par(mfrow=c(1, 2))
> plot(density(market_data$Input), main="Density Plot: Input", ylab="Frequency", sub=paste("Skewness:", round(e1071::skewness(market_data$Input), 2)))  # density plot for 'Input'
> polygon(density(market_data$Input), col="red")
> plot(density(market_data$Mprice), main="Density Plot: Price", ylab="Frequency", sub=paste("Skewness:", round(e1071::skewness(market_data$Mprice), 2)))  # density plot for 'Price'
> polygon(density(market_data$Mprice), col="red")
```

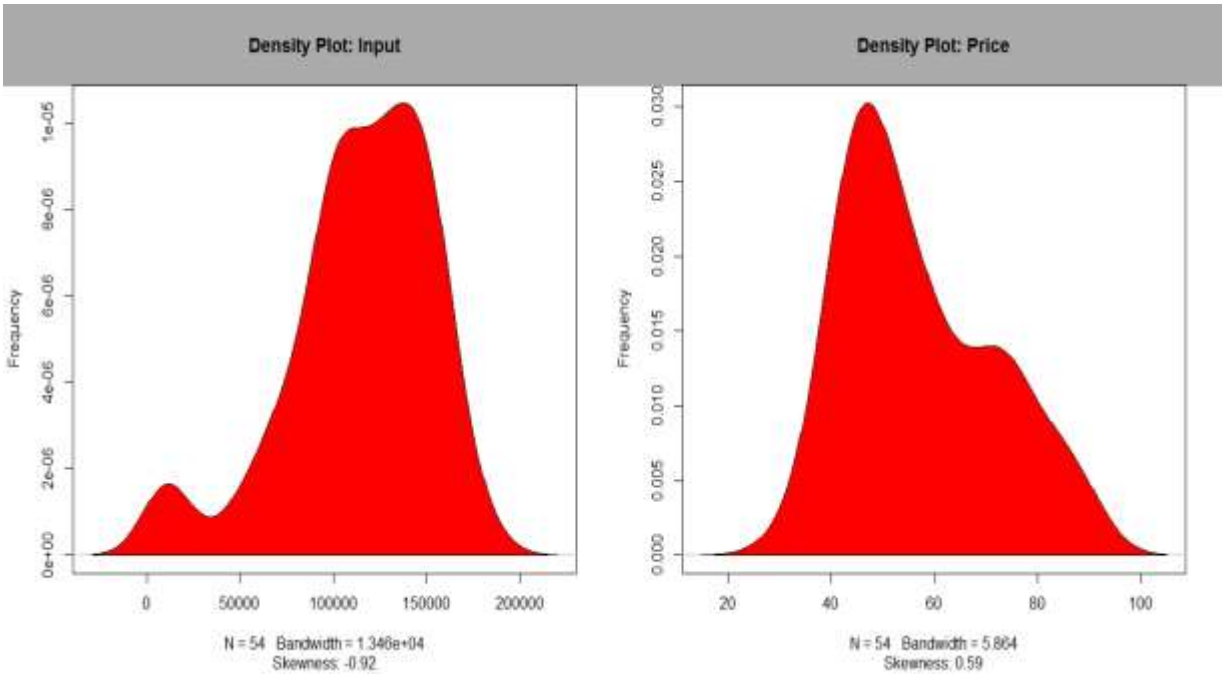
ISSN: 0474-9030

*Our Heritage*

Vol-68-Issue-15

January-2020

UGC Care Listed Journal




**Conclusion:**

From the secondary data collected we formed 3 clusters such as low (0-5 tonnes/hectare), moderate (6-10 tonnes/hectare) and high (11-15 tonnes/hectare) of pomegranate production in 13 districts of Maharashtra. Final distribution of 13 districts across 3 clusters is as follow

1) Beed and Auragangabad had the low pomegranate production
2) Buldana, Jalana, Solapur, Pune, Osmanabad are moderate pomegranate production districts.
3) Ahamadnagr, Satara, Sangli, Nashik, Dhule, Latur are high pomegranate production districts.

We have used Pune APMC Market data for analyzing market trend by using parameters such as input level and price. Result of experiment shows that as input level increases the price also goes on increasing , so it is controversial to market supply demand concept .in our further study we taking more number attribute to find out the input and price relation.

**References:**

1. "India at a glance", http://www.fao.org/india/fao-in-india/india-at-a-glance/en/
2. Majumdar J, Ankalaki S., "Comparison of clustering algorithms using quality metrics with invariant features extracted from plant leaves", International conference on computational science and engineering, 2016.
3. Ramesh D, Vishnu Vardhan B. Data mining techniques and applications to agricultural yield data. In: International journal of advanced research in computer and communication engineering. 2013; 2(9).
4. Motiur Rahman M, Haq N, Rahman RM. Application of data mining tools for rice yield prediction on clustered regions of Bangladesh. IEEE. 2014;2014:8–13.
5. Verheyen K, Adrianens M, Hermy S Deckers. High resolution continuous soil classification using morphological soil profile descriptions. Geoderma. 2001;101:31–48.

6. Gonzalez-Sanchez Alberto, Frausto-Solis Juan, Ojeda-Bustamante W. Predictive ability of machine learning methods for massive crop yield prediction. Span J Agric Res. 2014;12(2):313–28.
7. Pantazi XE, Moshou D, Alexandridis T, Mouazen AM. Wheat yield prediction using machine learning and advanced sensing techniques. Comput Electron Agric. 2016;121:57 – 65.
8. Veenadhari S, Misra B, Singh D. Machine learning approach for forecasting crop yield based on climatic parameters. In: Paper presented at international conference on computer communication and informatics (ICCCI-2014), Coimbatore, 2014.
9. https://numerical.co.in/numerons/collection/58df68014e976264035a1f85
10. http://nrcpomegranate.icar.gov.in/EPublications
11. http://www.puneapmc.org/rates.aspx
12. https://www.r-bloggers.com/the-ultimate-guide-to-partitioning-clustering/
13. https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/