Machine Learning Project Name-

Indian Agriculture Analysis

Name - Tejaswi

Entry No -2018UCS0090

Course Code - CSP774

Indian Institute of Technology Jammu
Department of Computer Science and Engineering

Machine Learning (CSL774)

Contents

1 Abstract

In India, agriculture is considered to be the largest economic sector as per population. It plays a vital role in the development of the nation and contributing to its economy. Various crops are grown here including staples such as rice and wheat among the most important ones. Other food crops that grow here include pulses, potatoes and other vegetables. Cash crops such as sugarcane, oil-seeds, cotton, coffee, tea, rubber, and jute are also grown here. Despite the fact that agriculture is a part of such a major portion of the Indian economy and employs a large section of society, it is highly inefficient, unscientific, and incapable of meeting the high food demands in such a hugely populated country. Despite advancements in this area, these problems still persist in most of the areas. These problems can be solved by proper analysis of the agricultural scenario and extracting information to provide suggestions regarding effective ways of growing crops and making choices in the type of crops.

2 Introduction

In agriculture sector, the farmers and agro based industries have to take several decisions every day and there are various factors that influence them. Some of the factors on which agriculture depends are soil, climate, cultivation, irrigation, fertilizers, temperature, rainfall, harvesting and use of pesticides. Mining the large amount of existing crop, soil and climatic data and analyzing the environmental conditions can make it possible for farmers to use this information and get help to make critical farming decisions. This optimizes the production and makes agriculture more resilient to climatic change. Historical crop yield information is also important for supply chain operation of companies engaged in industries. These industries use agricultural products as raw material, livestock, food, animal feed, chemical, poultry, fertilizer, pesticides, seed and paper. An accurate estimate of crop production and risk helps these companies in planning supply chain decision like production scheduling. Business such as seed, fertilizer, agrochemical and agricultural machinery industries plan production and marketing activities based on crop production estimates.

2.1 Objectives(Problem Statement):

This project uses several machine learning techniques to extract information from agriculture data and to give suggestions regarding crops and make future predictions so that agriculture can be carried out in a planned manner. The objectives of the project include-

1. Finding trends in crops in terms of production, area, etc. over the years and studying the reasons behind the changing trends.
2. Finding how different factors that affect production are related to each other.
3. Prediction of suicide rate of the farmers.
4. Study of crops that do not follow the general trends and show an abnormal trend such as reduction in production.
5. Finding similar crops and similar states based on various factors.

6. Predictions of crops that might be rarely produced and the main crops that might be preferred by the farmers.

3 Data Collection

We have used agriculture data from:

1. Crop Production Statistics
2. Data gov crops related data

The data bases we have used are:

No.	File name	Name of DB in program	Details or Comments
1	apy	Crop Production Statistics (crop prod)	Main DB that has crop production info from 2000 to 2014.For different district of each State it includes, what are various crops produced, their area of production and total production, and what type of crop is it.(Kharif, Rabi)
2	Crops price	Crop prices (crop price)	Prices of some crops year wise change till 2013.For different commodities it has data for its price in rupees per quintal.
3	area cult	Crop cultivation area (crop cult)	Area of land a crop is produced on year by year for major crops from 2000 to 2009.
4	culti cost	Crop cultivation cost (culti cost)	State wise cost of cultivation of crops per hectare and per quintal. Three variant of cost are there(actual paid out cost plus imputed value of family labour(A2+FL),comprehensive cost including imputed rent and interest on owned land and capital(C2) and cost per quintal)
5	Mean _ Temperatures	Mean Temperatures(temperature)	Data of mean temperature from 2000-2012 for whole year and over interval of two months. This is used to determine effect of temperature on various crops
6	rainfall cleaned	Rainfall Statistics(rainfall)	State wise rainfall statics from year 2000-2015 annually and monthly in millimeter per square meter(area)
7	Avg annual Growth Rate _Major Crops	Crops Growth rate(growth)	Growth rate of various crops from 1997 to 2012 over a interval of five years. Growth rate represent increase in size, mass or number of crops over a period of time. It is

			used in analysis of preference of one crop over other.
8	suicides 10 14	Suicide Statics(suicides)	Data of no. of total cases of suicides in various states from year 2010-2014.While analysis it is taken into account to predict responsible factors.
9	India Export	Exports(exports)	Data regarding the amount of export of various materials and its price from 2003 to 2015.
10	data set	Data(data)	Combined data of various states from 2000 to 2014. This is combined representation of all data in one table. All other tables mentioned above are combine using the common features and merged.

4 Approach

4.1 Data Cleaning

The data needed to be cleaned in the beginning. The challenges faced while cleaning the data are-

1. The databases obtained composed of data of different years, which were not same across databases.
2. The names of some crops were not present in all the databases.
3. The database also contained a lot of missing data.
4. The data was of varying formats.
5. The naming conventions of crops and states were not the same across databases.
6. The units of measurements were different in different databases.

The databases were modified to contain data in a proper format and the missing values were replaced by the mean values of various years. Then the data was ready for any further processing.

4.2 Data Integration

The data from different tables were merged so that it can be analyzed. The tables were also unstacked when required, for proper understanding.

4.3 Data Analysis

The data were visualized and several plots were made for statistical analysis. Various machine learning algorithms were also used for finding patterns and making predictions. They are described in detail in the following sections.

5 Data Plots

5.1 Crop Prices

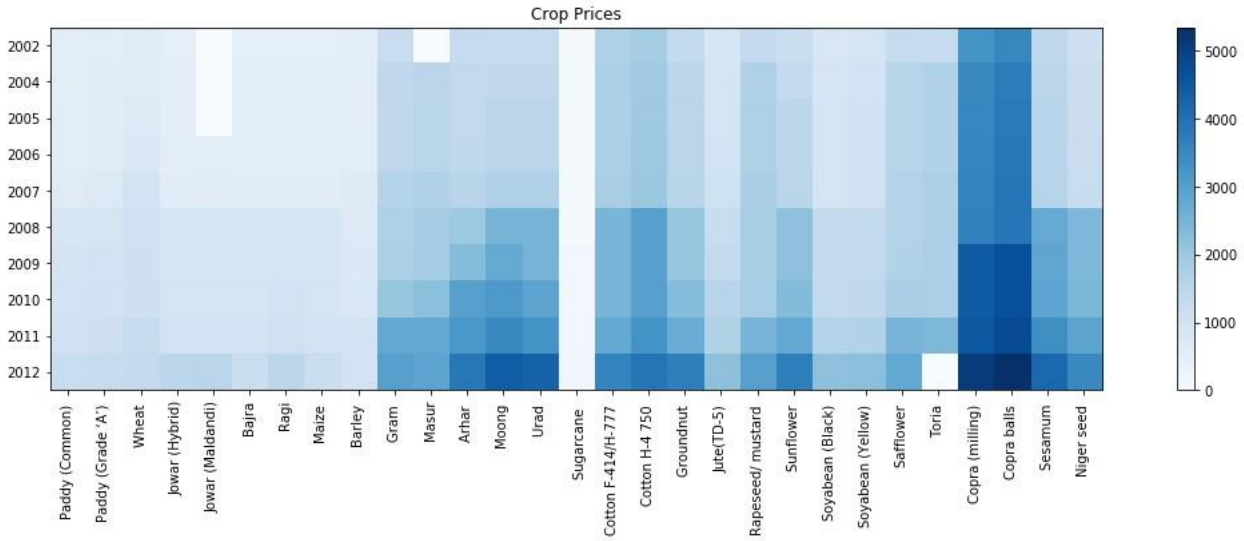

Figure 1: Crop price of various crops in different year in Rs/quintal

5.2 Cultivation Area

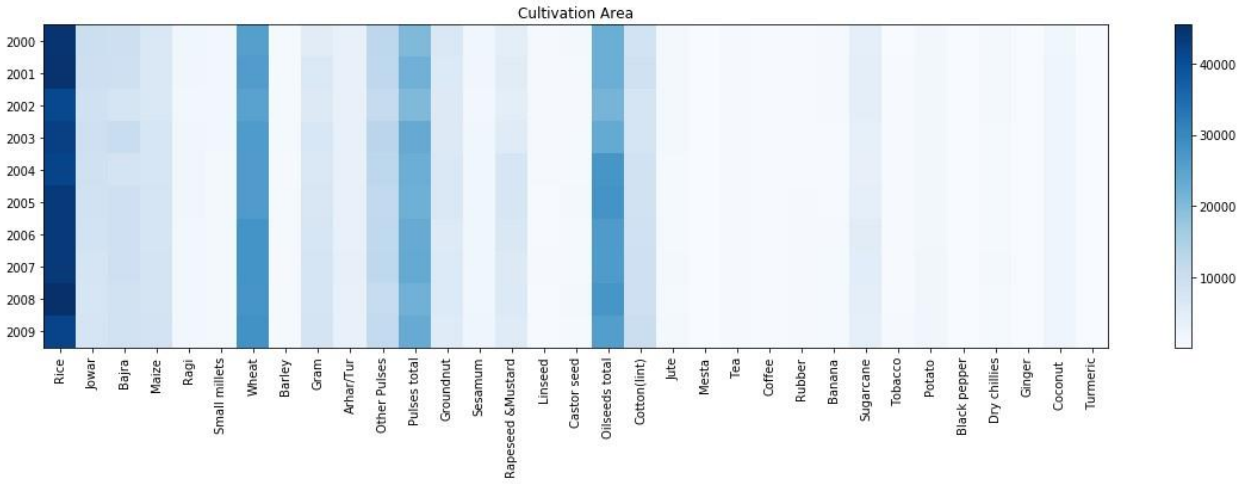

Figure 2: Cultivation area of various crops in hectares

5.3 Cultivation Cost A2+FL and C2

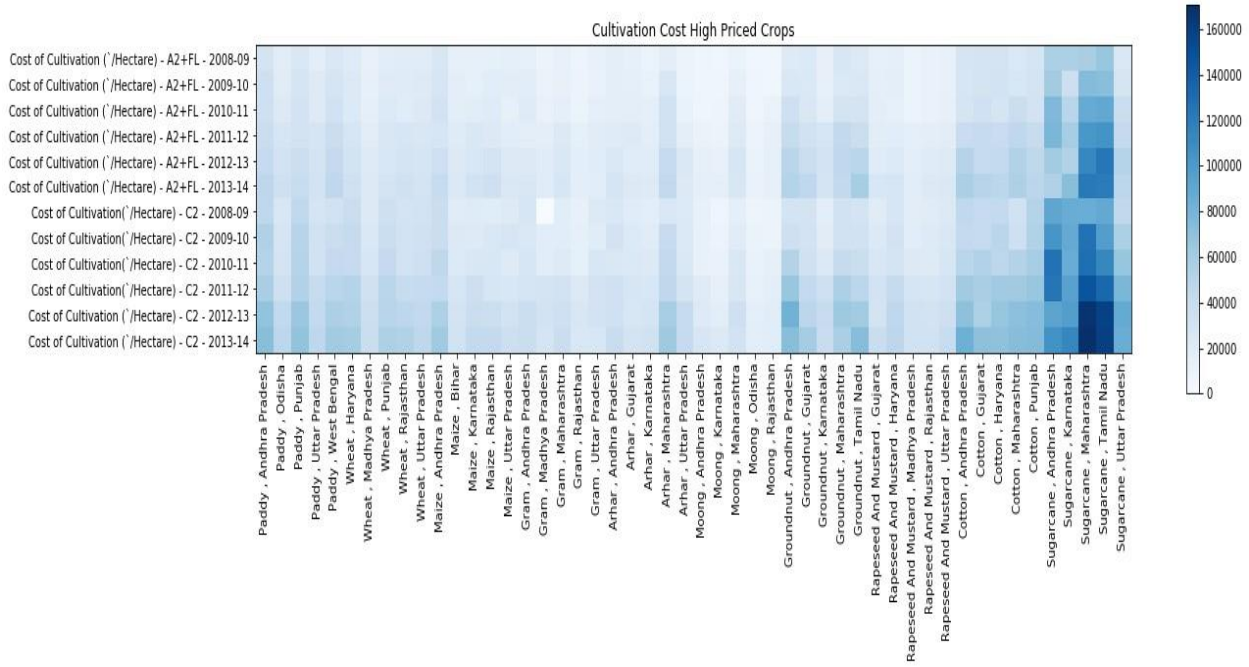

Figure 3: Cultivation cost by area of major crops in respective states in Rs/hectare

5.4 Cultivation Cost by Quintal

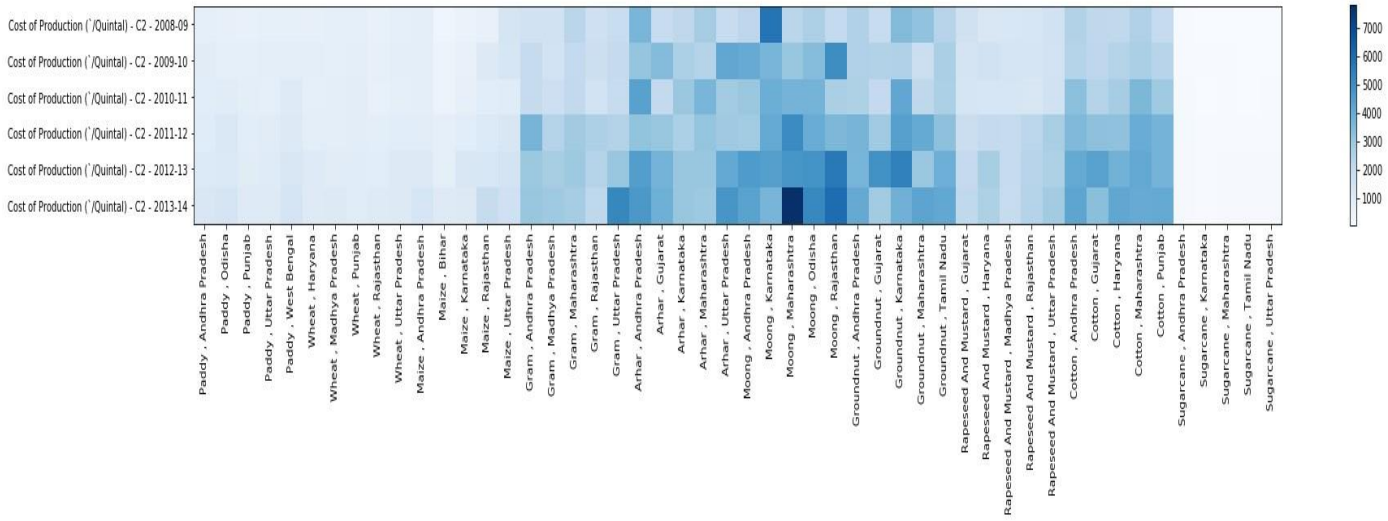

Figure 4: Cultivation cost by quintal of major crops in respective states

5.5 Farmer Suicides

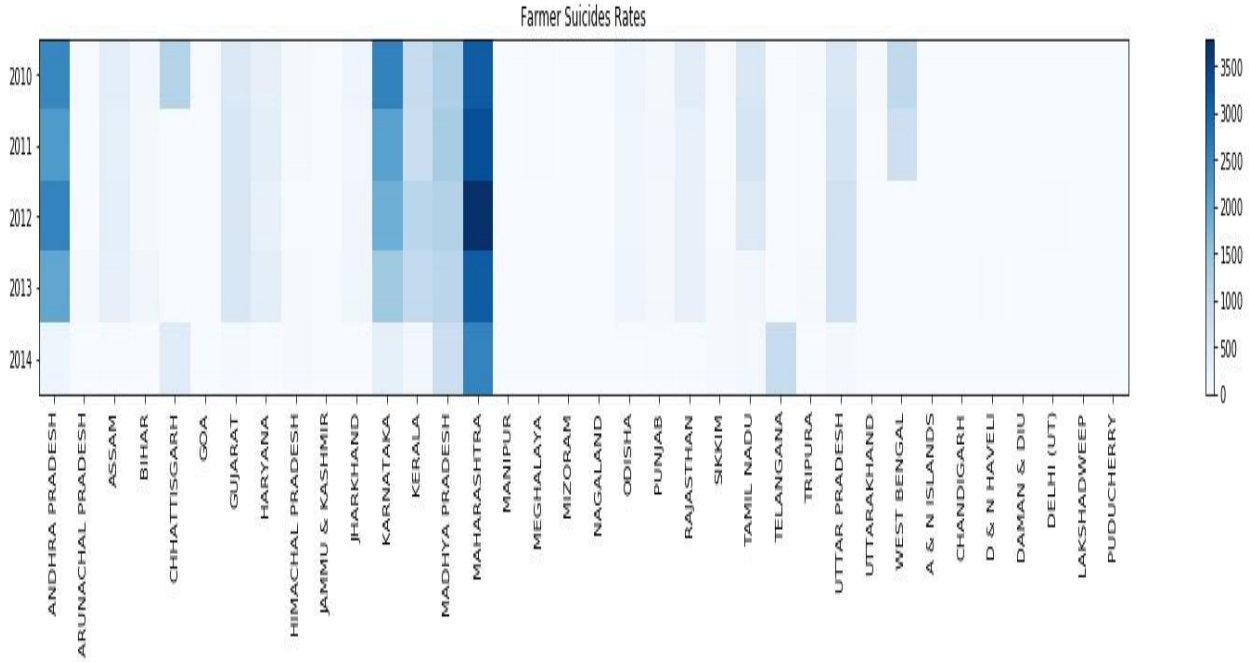

Figure 5: Farmer suicide rate

5.6 Growth Rate of Major Crops

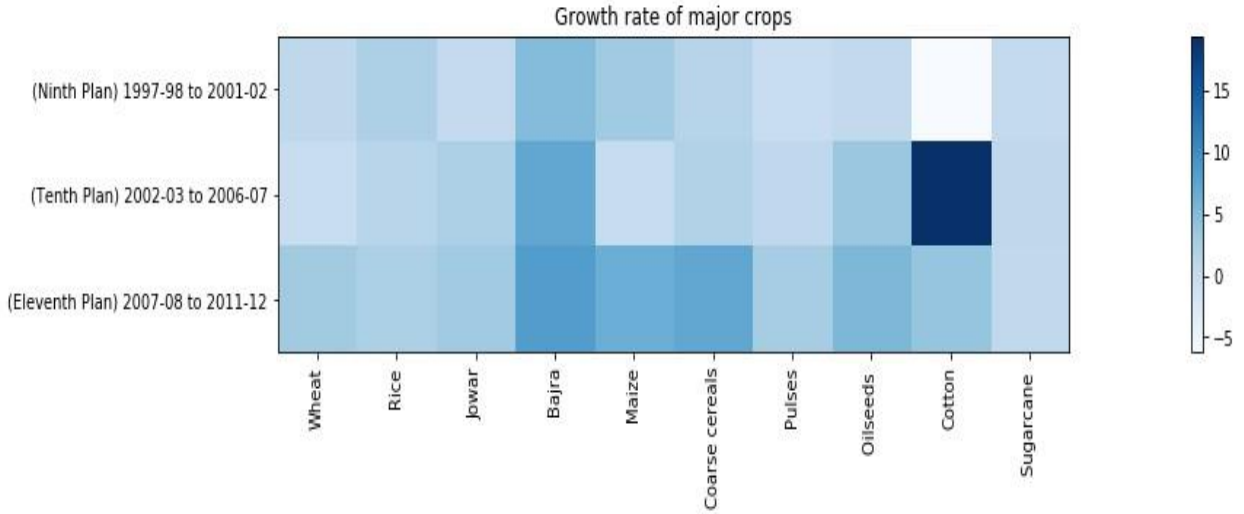

Figure 6: Growth rate of major crops

5.7 Annual Rainfall

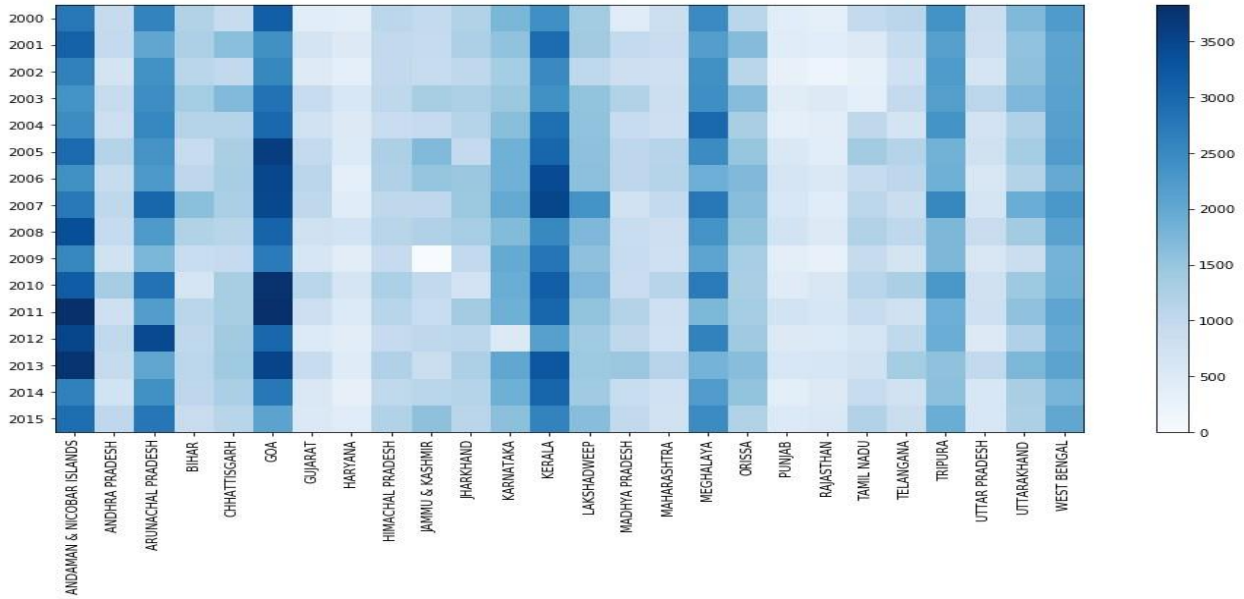

Figure 7: Annual rainfall

5.8 Temperature Variations

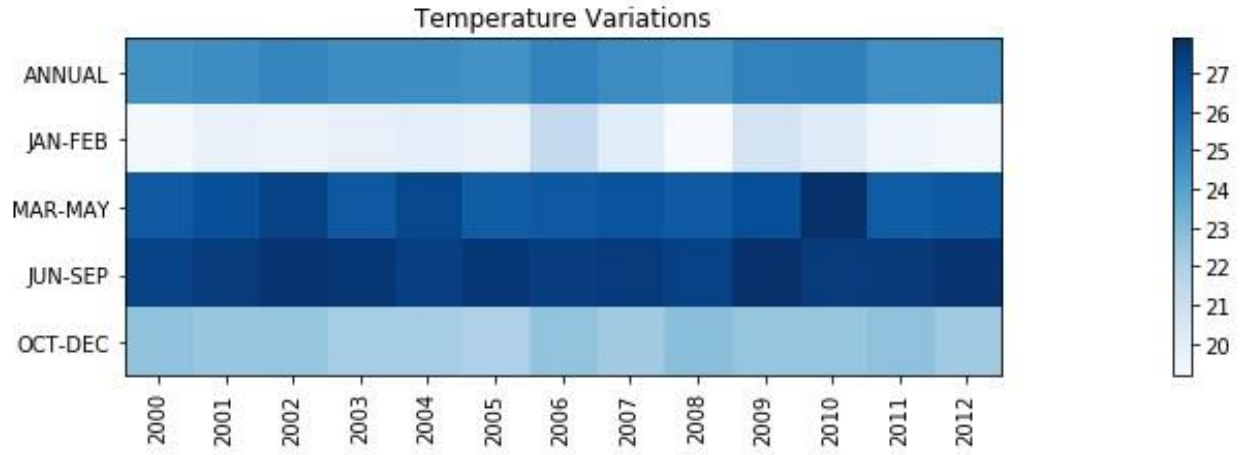

Figure 8: Temperature Variations of various year in Centigrade

6 Observations

6.1) In the chart above, we see the area under cultivation of particular crops from 2002 to

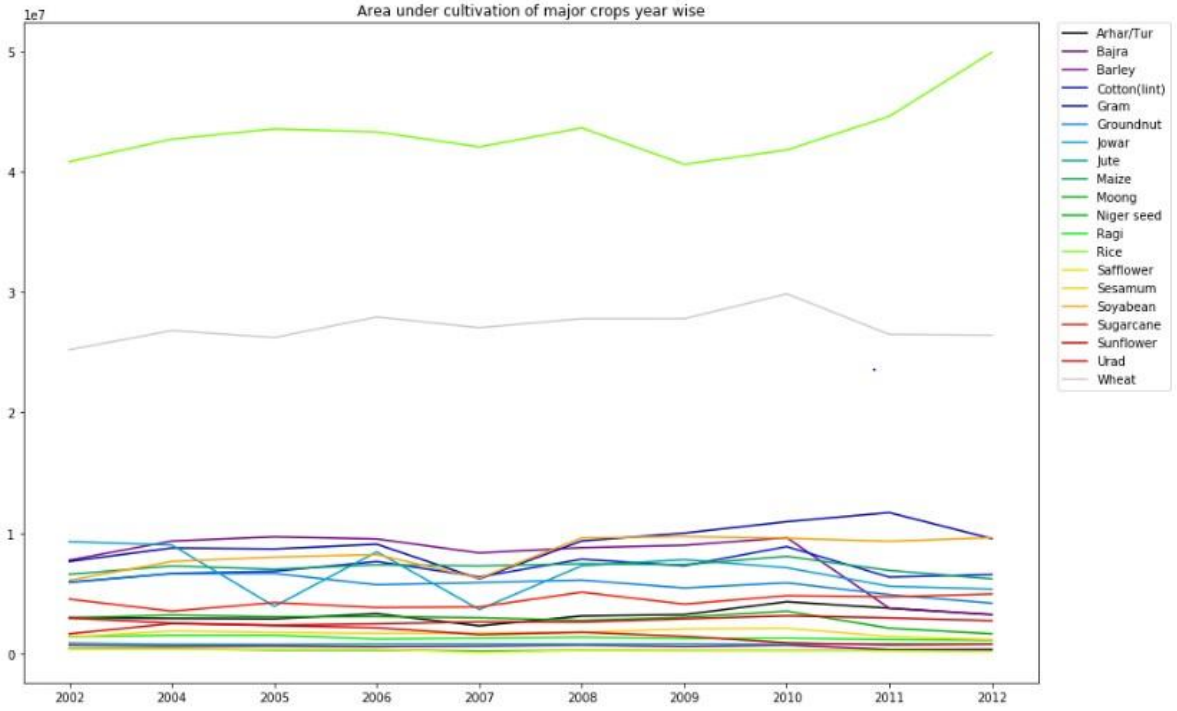

Figure 9: Area under Cultivation of Major Crops in various years in hectares

2012.Overall, we see that the area under cultivation of Rice and Wheat takes the maximum proportions over all years covered . As of 2012-13, wheat and rice accounted for 75% of the food grains production in the country. Area under rice has increased over the years from about 40 million in 2002 to 50 million in 2012. Area under wheat has remained constant over the years with very small variations in short term. The country's increasing requirement for these food grains is attributed to its population increase over the decade.

6.2)

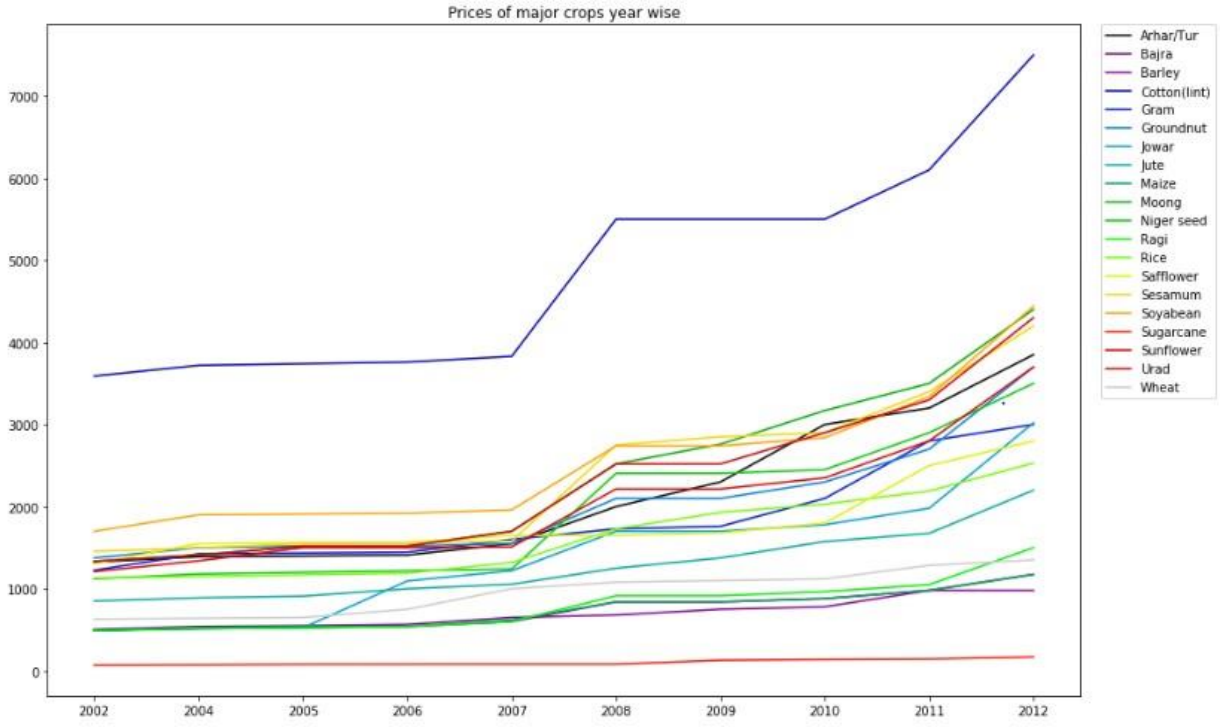

Figure 10: Price per quintal of Major Crops by year in various years

Over the years, the price of all major crops have increased slightly considering the increased demand due to the population (except some rare crops) with a sudden increase in 2008 due to 2007-08 food crisis. Cotton prices are high due to its high cost of cultivation.

6.3)

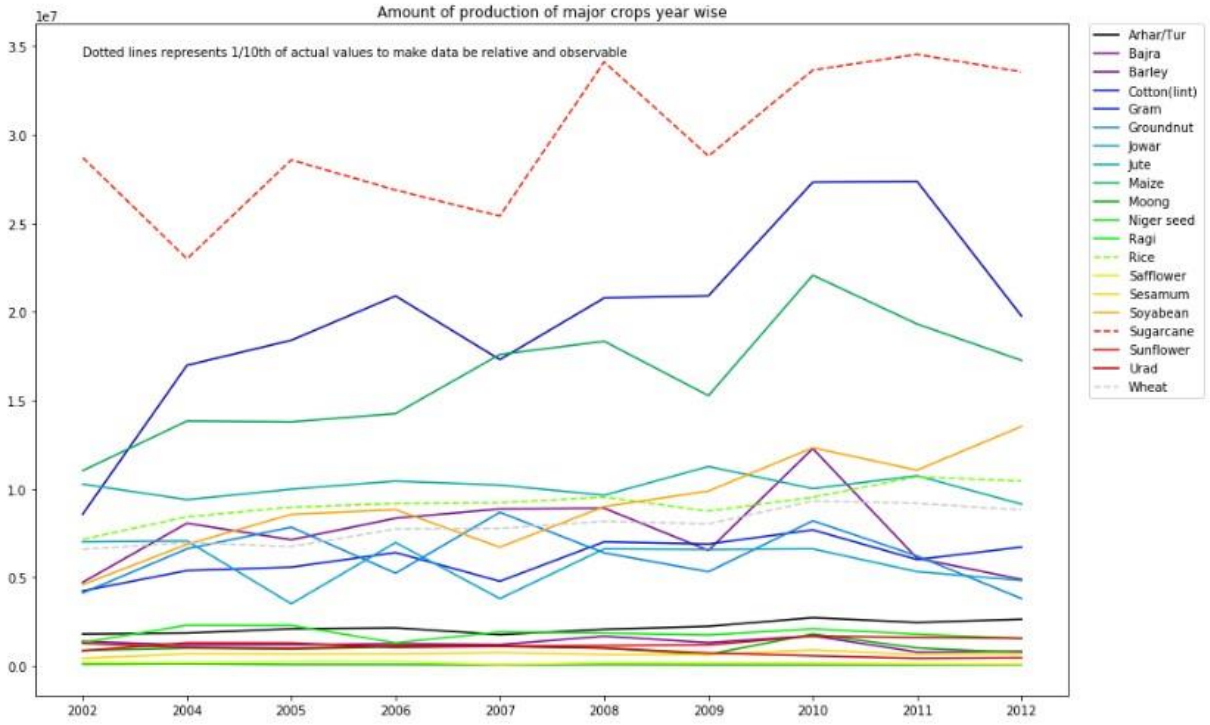

Figure 11: Production of Major Crops in various years

This graph shows the production of different crops over the years. The highly produced crop in India is Sugarcane. India is the second largest producer of this popular cash crop. It has one of the longest growing periods. Other crops which are highly produced over the country includes rice , wheat, cotton etc.

6.4)

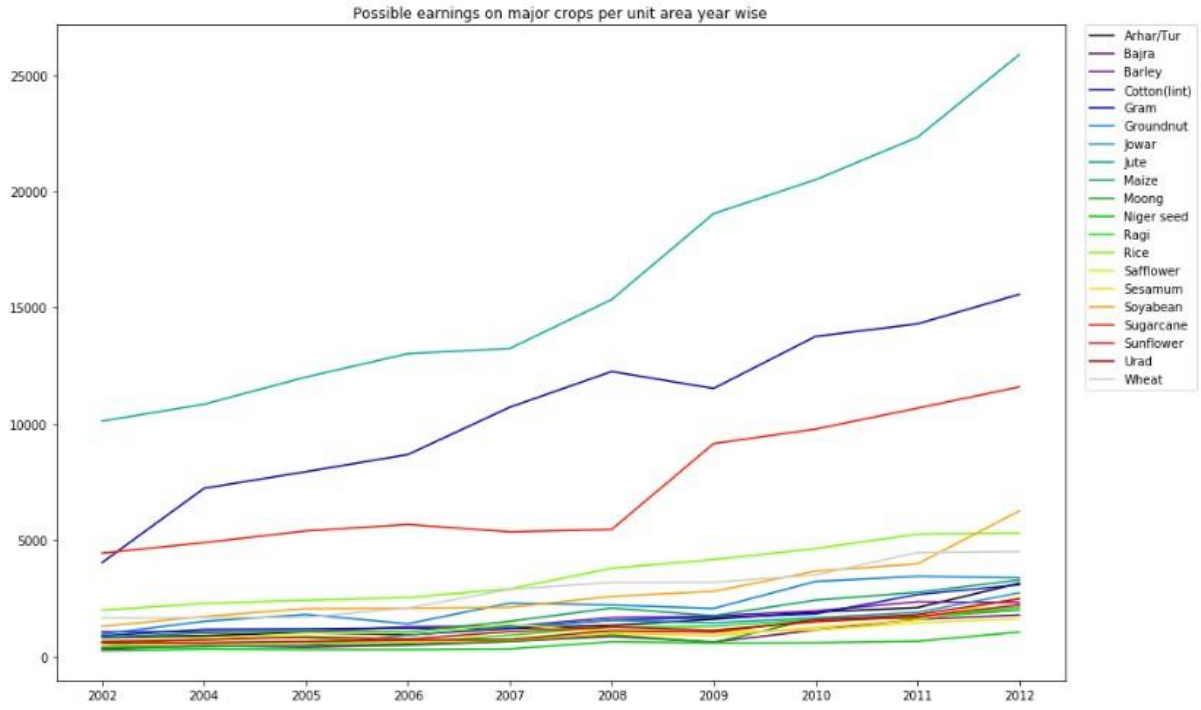

Figure 12: Possible earnings on major crops per unit area in various years

By analyzing the above graph, we find that most of the possible earnings per unit area is contributed by Groundnut, Gram and Sugarcane. The earnings have increased over the years because of the increase in price of the crops.

6.5) In these plots, influence of rainfall on the area under cultivation for different states is shown. Most of the agricultural area in India is still depending on monsoon rainfall. Rainfall can have direct or indirect impact on the area.
For example, lack of rainfall can cause short of irrigation water supplies leading to lesser area under cultivation.
As shown in these plots, steepest relation between area and rainfall can be seen in the states like Bihar, Haryana, Madhya Pradesh, Rajasthan and Telangana.

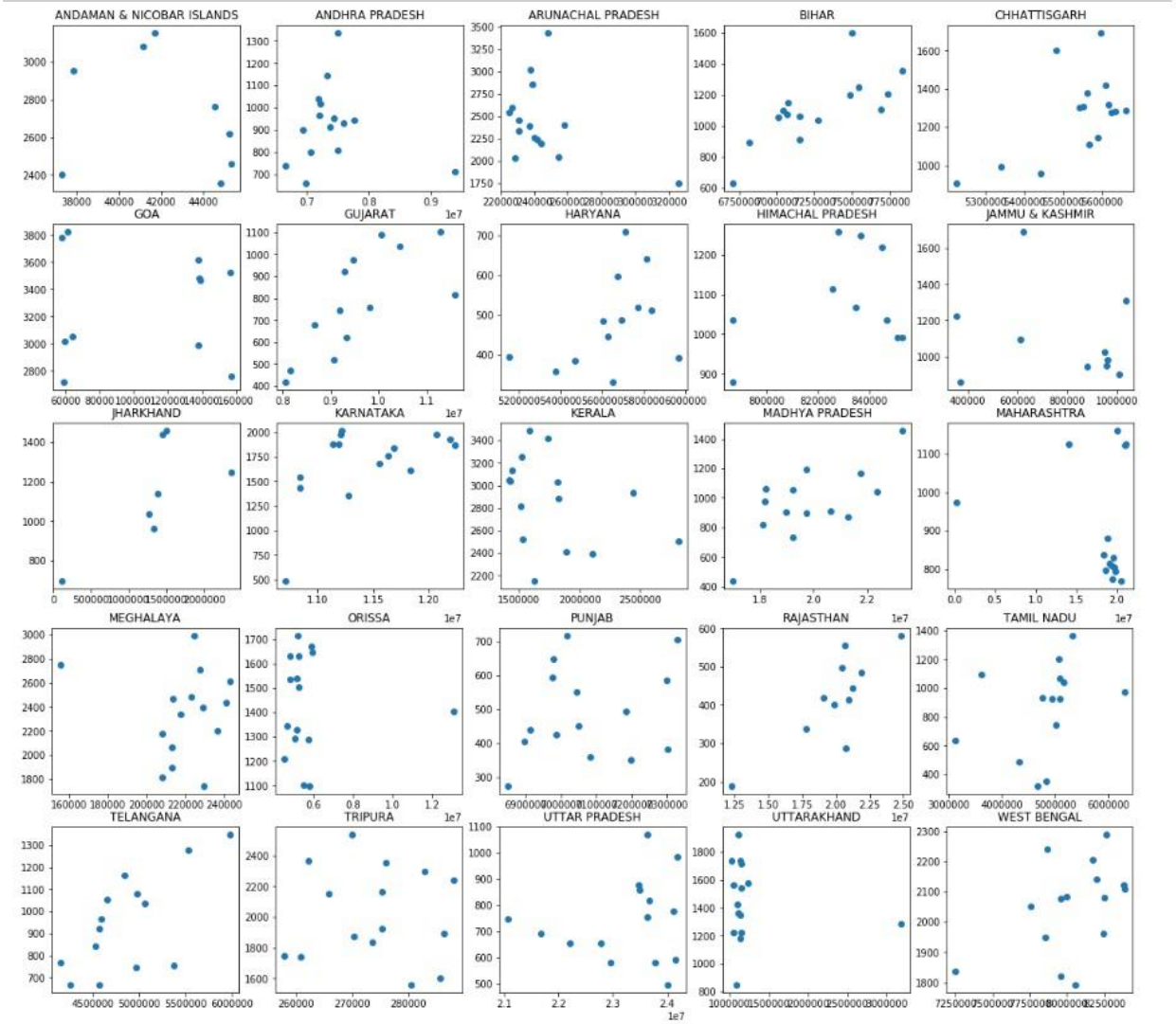

Figure 13: Relationship between Avg Rainfall and Area in different states

6.6)

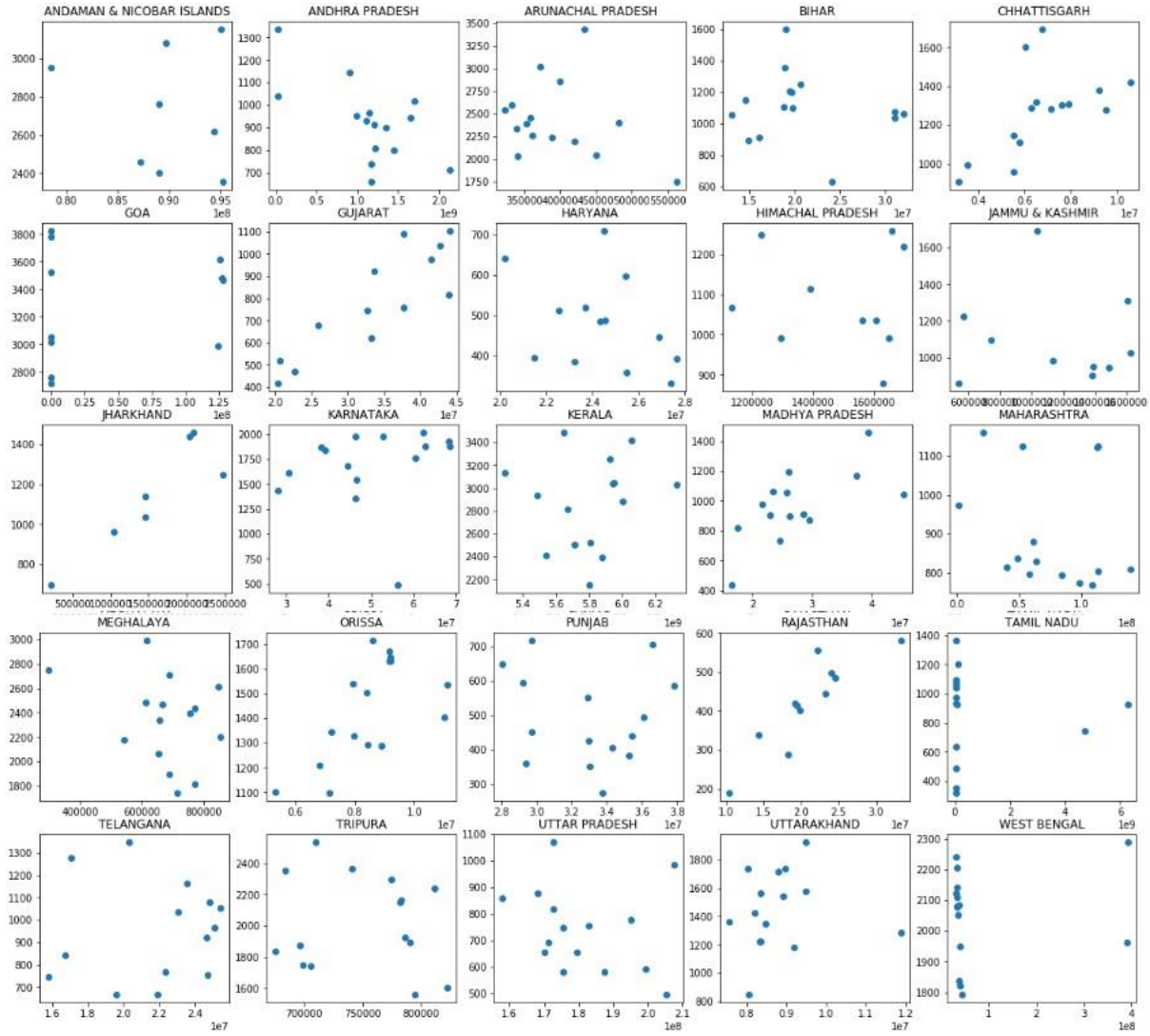

Figure 14: Relationship between Average Rainfall and Production in different states

In these plots, influence of rainfall on the production for different states is shown. Rainfall can affect the production of crops in a great way(sometimes in long term as well).For example, below normal rainfall can cause damaged crops and damaged soil quality. As shown from the plots, in states like Chhattisgarh, Gujarat, Jharkhand, Karnataka etc (where major crop is rice and sugarcane) increased rainfall is leading to increased production. In Haryana, Uttar Pradesh, Rajasthan(major crop is wheat) increased rainfall is showing to have negative impact on production

6.7)

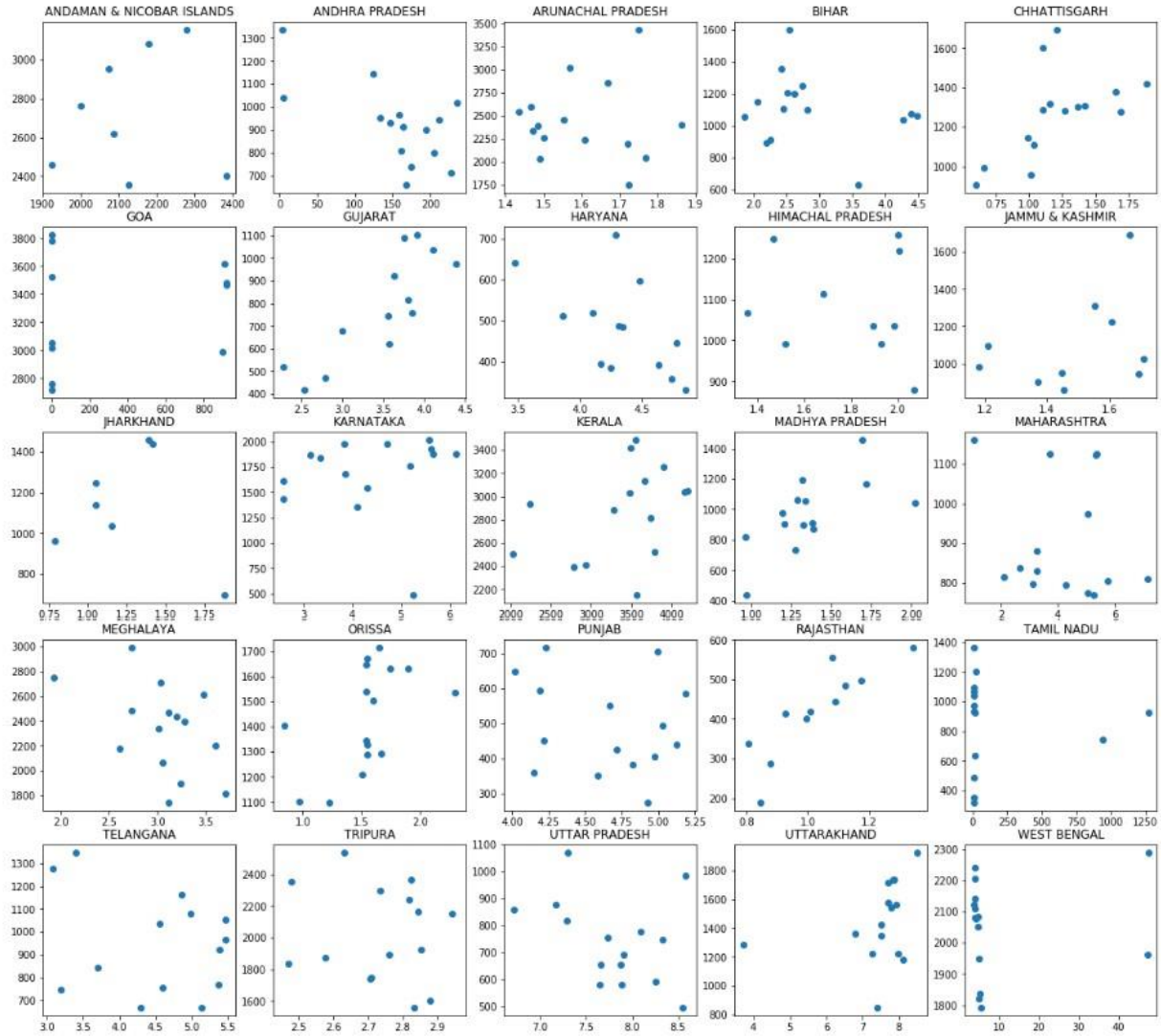

Figure 15: Relationship between Average Rainfall and Yield in different states

In these plots, influence of rainfall on the yield for different states is shown. While investigating the impacts of rainfall variability, it is more important to consider yield(Production/Area) than production. Focusing on yield could give results that could help to identify the extreme severe conditions like severe drought, therefore, it has more economic importance. From these plots, it can be observed that rainfall have positive impact on yield in states like Gujarat, Chhattisgarh, Rajasthan, Himachal Pradesh, Madhya Pradesh (major crop is rice and sugarcane).

6.8)

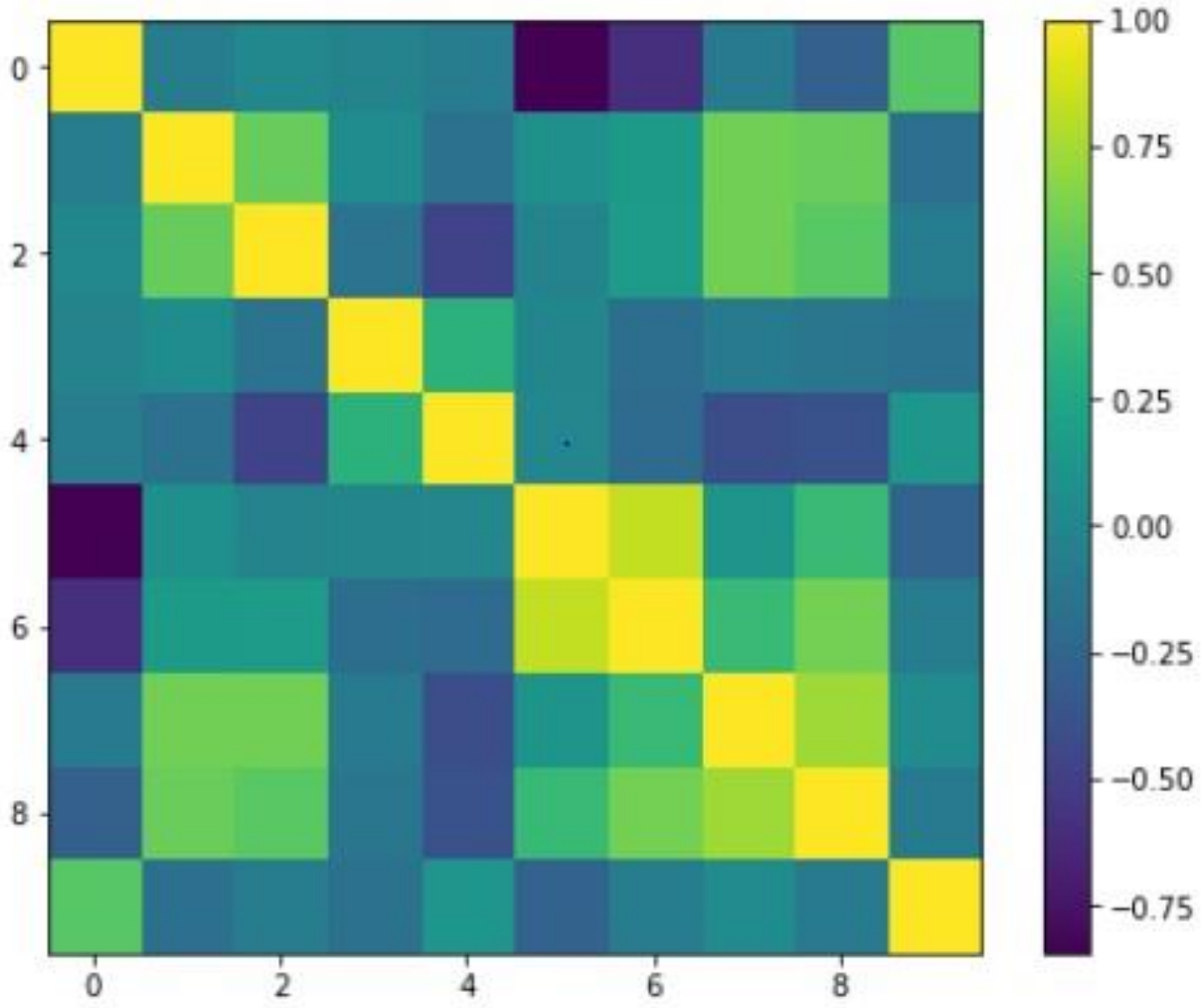

Figure 16: Correlation between different features

This plot helps us to visualize the correlation between different features of the table created after merging all data related to Area, Production, Rain, Suicides etc. As shown from this plot, all features are sufficiently independent.

6.9)

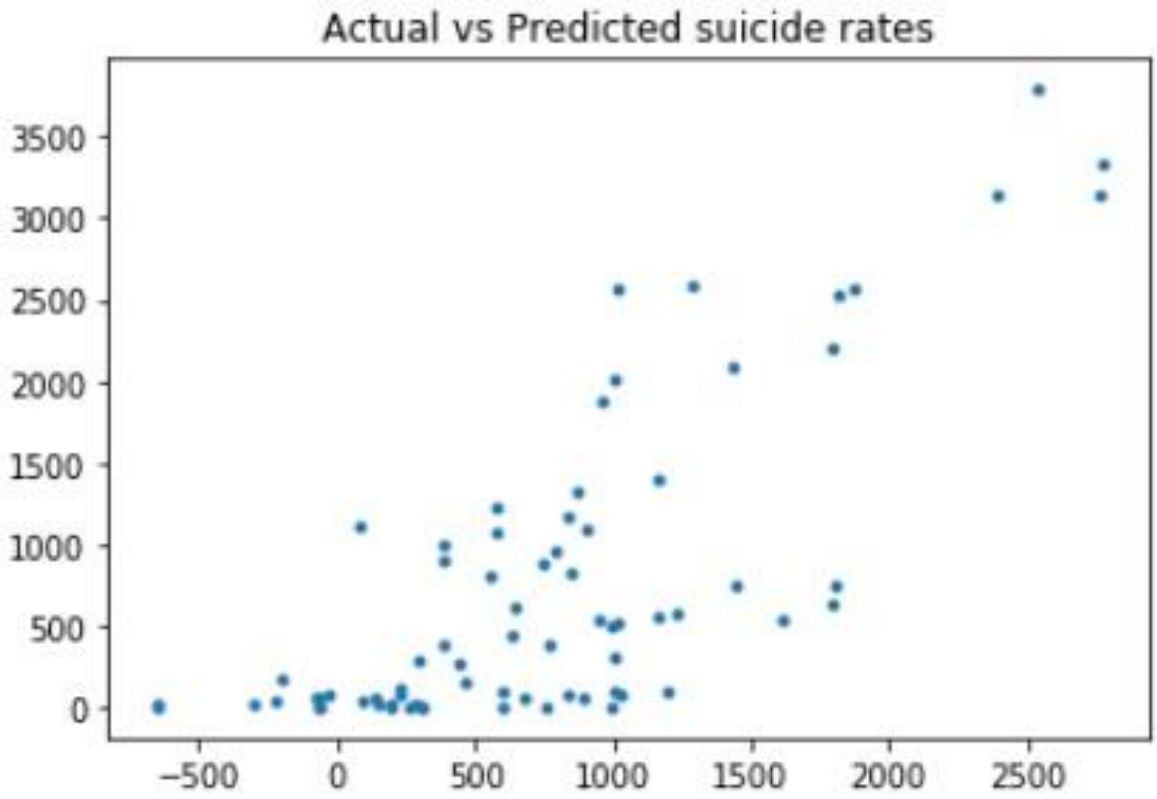

Figure 17: Relationship between actual and predicted suicide rates with Linear Regression

Using Linear Regression to predict the suicides by remaining features like rainfall, production, area under cultivation and others, gives us the accuracy of 62.07%.

6.10) Using Classification and Regression Trees, accuracy of 83% could be achieved.

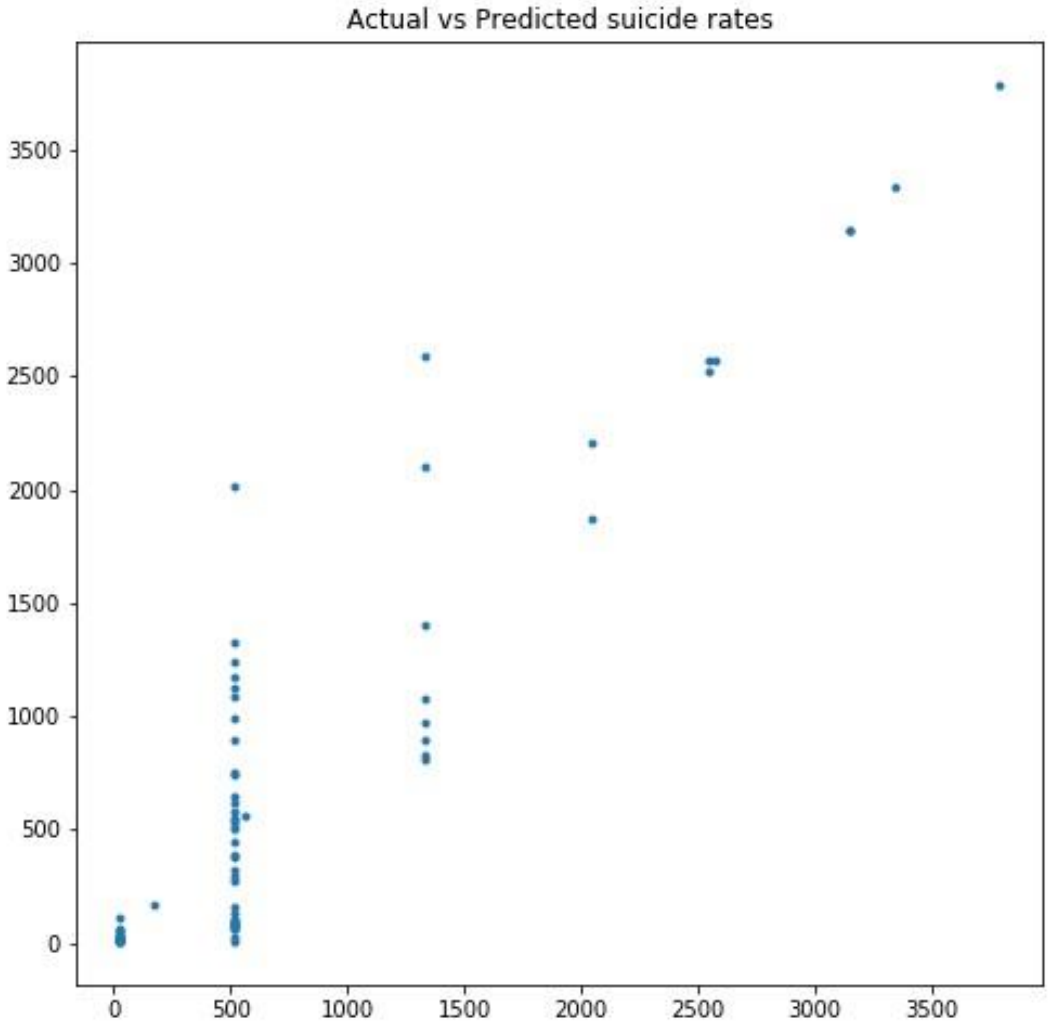

Figure 18: Relationship between actual and predicted suicide rates with Classification and Regression Trees below

Given is the graphical visualisation of Decision Tree modelled

6.11)

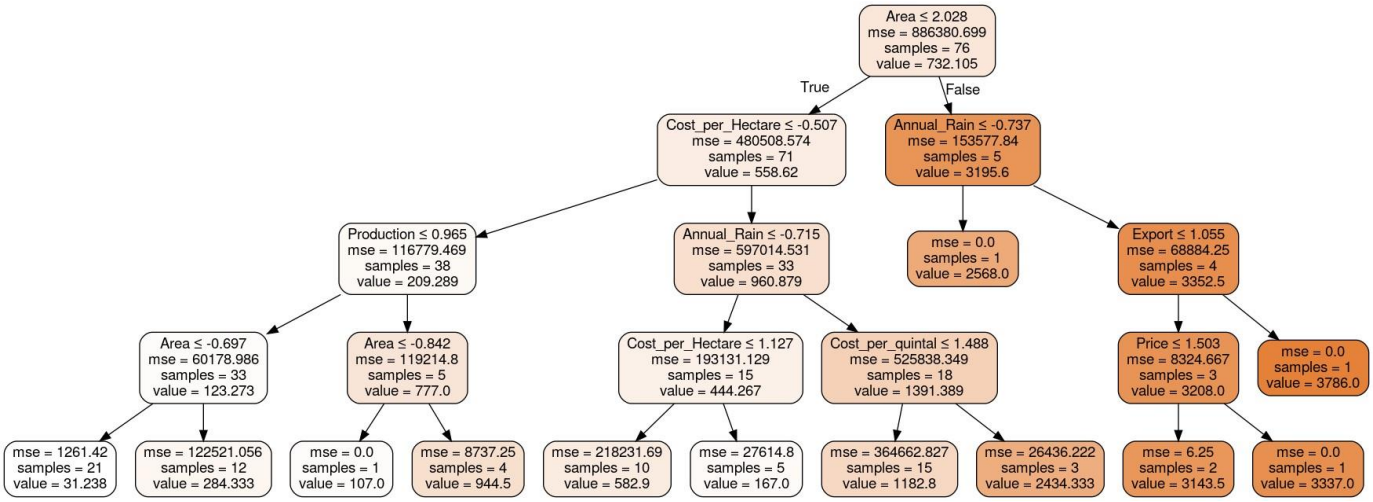

Figure 19: Decision Tree for Suicide Prediction

The decision making in decision tree regressor can be seen here for prediction of number of suicides. The depth of tree is limited to 4. As mentioned the accuracy we achieved is 83%.

6.12)

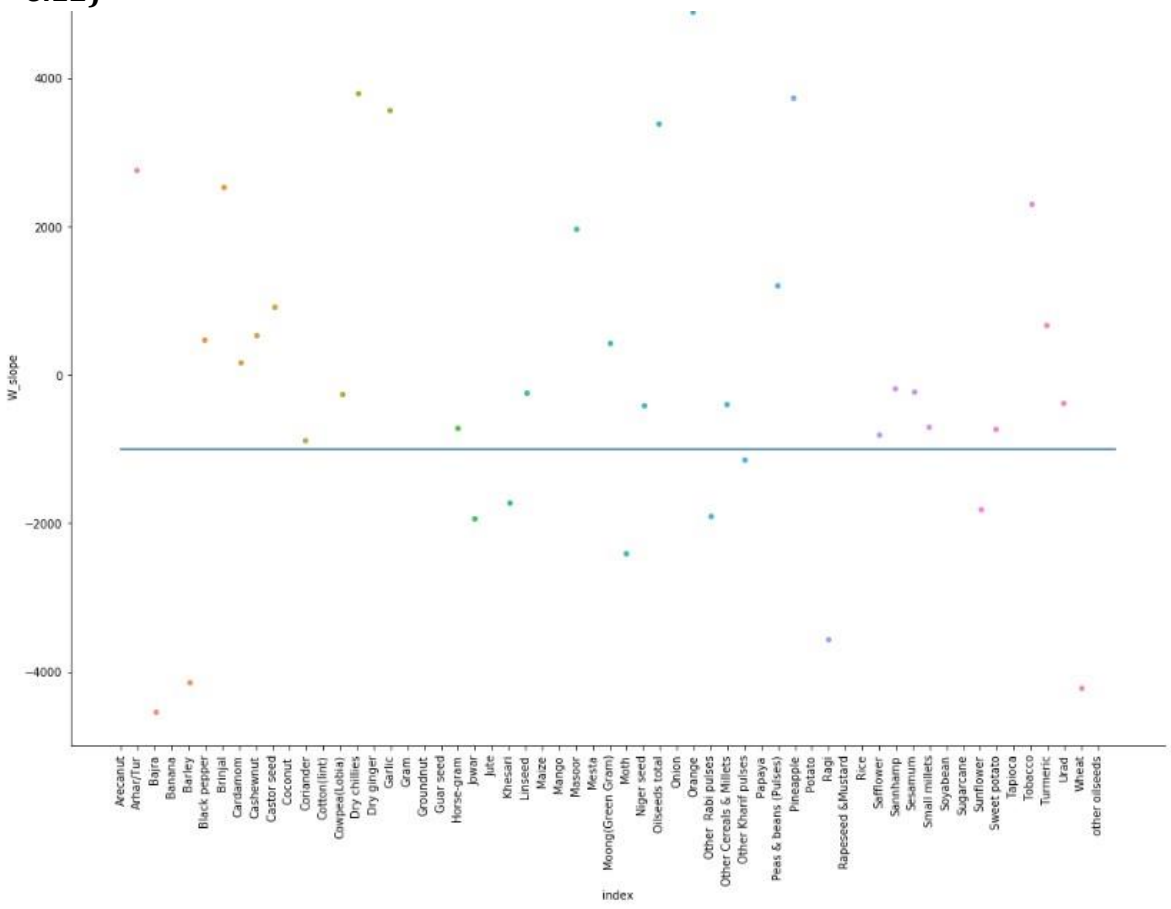

Figure 20: Plot for crops that have more reduction in production over the years

From the above graph, it can be concluded that for some of the crops like Bajra,Groundnut,Jowar etc production is reducing over the years as the slope of the production vs year is negative.
We have considered threshold as -1000 for the slope.

6.13) We are clustering the crops in this figure on the basis of data available except

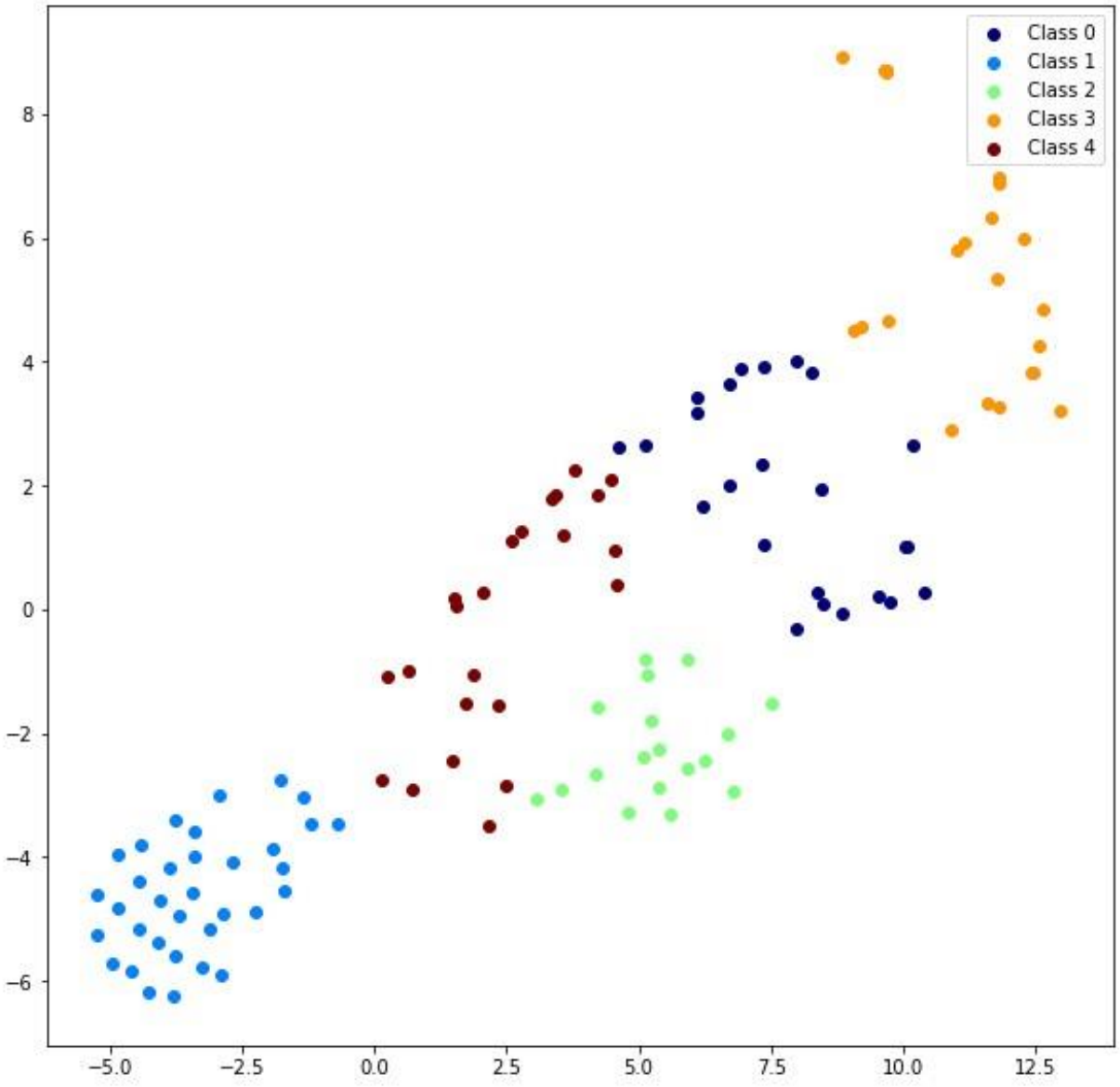

Figure 21: Clustering of similar crops

for export. This data was converted to 2 dimesions using tSNE as keeps the neighbours in original space closer in lower dimensions as well. The cluster number was chosen by plotting the data first. K-means clustering is used for this as even number of points in each cluster can be observed.

Cluster 1 :

Other Citrus Fruit, Perilla, Ricebean (nagadal), Varagu, Water Melon, Blackgram, Ber, Sannhamp, Cauliflower, Carrot, Snak Guard, Other Dry Fruit, Apple, Peas (vegetable), Pear, Cucumber, otherfibres, Bean, Ribed Guard, Turnip, other misc. pulses, Bitter Gourd, Bottle Gourd, Yam, Ash Gourd, Jobster, Lab-Lab, Plums, Pump Kin, Beet Root, Peach, Redish, Lentil, Litchi and Kapas.

Cluster 2 :

Jute, Soyabean, Atcanut (Raw), Tapioca, Jowar, Guar seed, Rubber, Bajra, Groundnut, Banana, Total foodgrain, Potato, Gram, Cotton(lint), Sugarcane, Wheat, Rapeseed & Mustard, Maize and Rice.

Cluster 3 :

Cardamom, Other Cereals & Millets, Safflower, Other Kharif pulses, Oilseeds total, Linseed, Cashewnut, Rajmash Kholar, Sesamum, Samai, Black pepper, Peas & beans (Pulses), Niger seed, Cowpea(Lobia), Cashewnut Processed, Jack Fruit, Korra, Coriander, Pulses total, Other Rabi pulses, Small millets, Horse-gram and Drum Stick.

Cluster 4 :

Sapota, Papaya, other oilseeds, Cashewnut Raw, Urad, Barley, Pineapple, Moong, Arhar/Tur, Arecanut, Orange, Mango, Arcanut (Processed), Dry chillies, Ragi, Coffee, Khesari, Onion, Masoor, Lemon, Tomato, Sunflower, Moth, Castor seed and Grapes.

Cluster 5 :

Mesta, Sweet potato, Dry ginger, Garlic, Pome Granet, Other Vegetables, Turmeric, Other Fresh Fruits, Tobacco, Brinjal, Pome Fruit, Beans & Mutter(Vegetable), Bhindi, Citrus Fruit, Tea, Cabbage, Colocosia, Ginger and Jute & mesta.

6.14) In this we tried to cluster the states with similarities over all data in different years

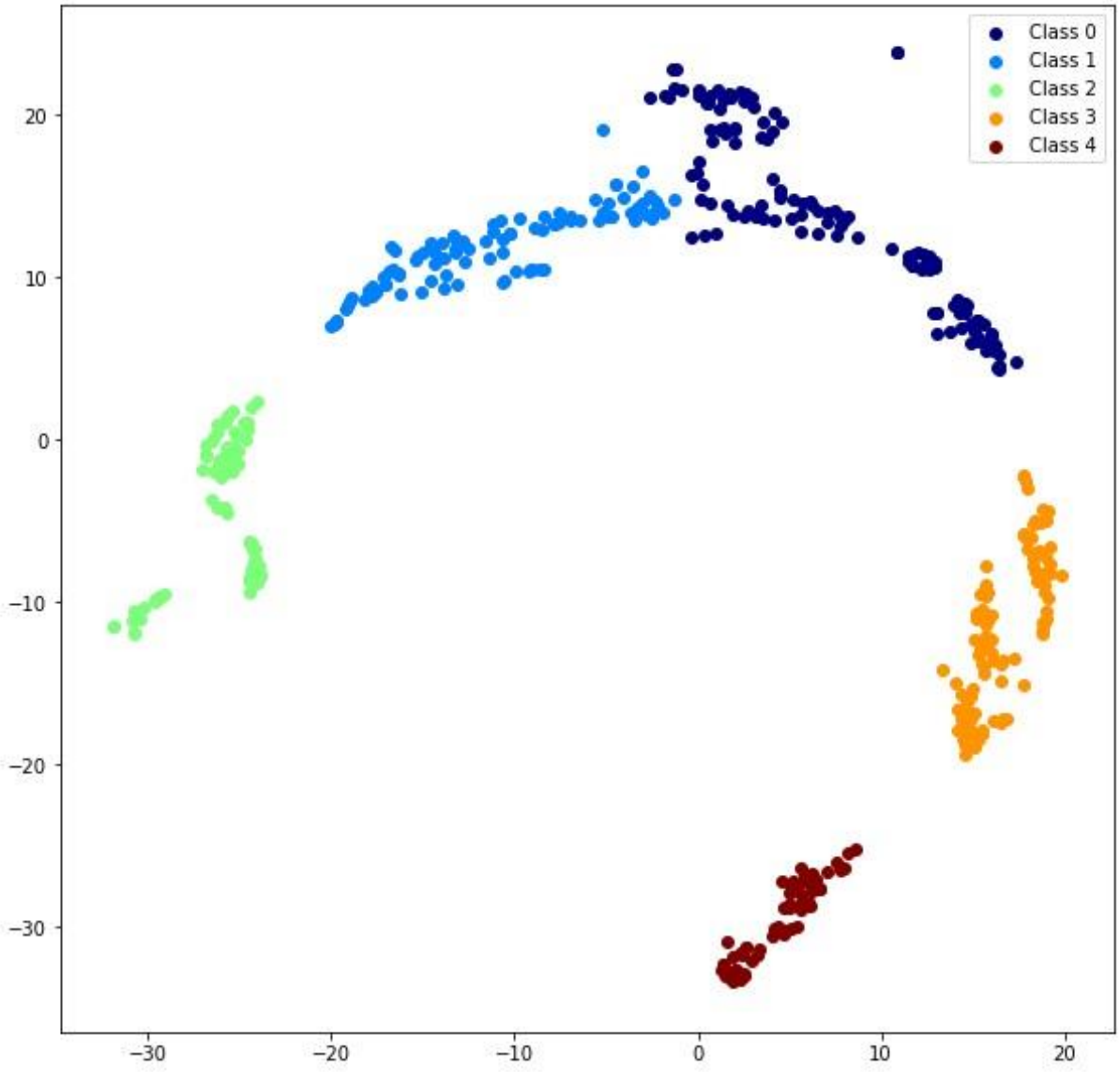

Figure 22: Clustering of similar states

i.e. state and year as class. The data was converted to 2 dimensions using tSNE due to reasons mentioned earlier. The clustering used in this was agglomerative clustering as cluster can be seen as uneven and separated. Agglomerative clustering captures this keeping the clusters uneven. We used 5 cluster as it can be visually oberved.

Cluster 1 :

UTTARAKHAND, JAMMU & KASHMIR, BIHAR, GUJARAT, ANDHRA PRADESH, ASSAM, CHHATTISGARH, PUDUCHERRY, HIMACHAL PRADESH, TAMIL NADU, MADHYA PRADESH, MAHARASHTRA, JHARKHAND, RAJASTHAN, TELANGANA, KARNATAKA, HARYANA and ORISSA.

Cluster 2 :

PUNJAB, BIHAR, GUJARAT, ANDHRA PRADESH, PUDUCHERRY, TAMIL NADU, MADHYA PRADESH, MAHARASHTRA, WEST BENGAL, KARNATAKA, HARYANA and ORISSA.

Cluster 3 :

GOA, UTTAR PRADESH, ANDHRA PRADESH, ASSAM, KERALA, TAMIL NADU, MAHARASHTRA, ANDAMAN & NICOBAR ISLANDS and WEST BENGAL.

Cluster 4 :

MEGHALAYA, GOA, MANIPUR, JAMMU & KASHMIR, PUDUCHERRY, HIMACHAL PRADESH, NAGALAND, MAHARASHTRA, JHARKHAND, SIKKIM, ARUNACHAL PRADESH and TRIPURA.

Cluster 5 :

MEGHALAYA, MANIPUR, CHANDIGARH, DADRA AND NAGAR HAVELI, MIZORAM, NAGALAND and SIKKIM.

7 Results:

7.1)

	index	W_slope
2	Bajra	-4537.162480
4	Barley	-4141.496114
18	Groundnut	-12034.886469
19	Guar seed	-12780.917262
21	Jowar	-1927.821737
22	Jute	-54720.777614
23	Khesari	-1717.234226
28	Mesta	-5094.770567
30	Moth	-2411.072500
35	Other Rabi pulses	-1908.905869
37	Other Kharif pulses	-1140.953895
42	Ragi	-3564.839019
43	Rapeseed &Mustard	-6685.204140
51	Sunflower	-1812.156509
53	Tapioca	-11008.462271
57	Wheat	-4216.058051

Figure 23: Crop with reduced production over Years

The above table shows the crops that have negative overall slope or in other words have seen decrease in production over the years. This same can be observed in figure 21. The threshold we used is -1000.

7.2) Finding crops that have reduction in production but their price is increasing

	index	production var	price var
0	Barley	-129.610556	51.277056
1	Jute	-43011.146430	125.248918
2	Niger seed	-144.009443	249.534632
3	Safflower	-1031.251935	122.445887
4	Sunflower	-4511.467222	223.906926

Figure 24: Crop with reduced production and Increased price

In this we tried to find the crops that have reduction in production of the crop but there is still increase in price of the crop. This shows that the production has been decreasing but the demand for the same crops is not as can be observed by the positive value of slope.

7.3) Finding the crops which has lower increase in production but are increasing in price

	index	production var	price var
0	Arhar/Tur	5237.572915	260.551948
1	Groundnut	1758.235163	199.339827
2	Jowar	3648.323679	234.956710
3	Jute	-43011.146430	125.248918
4	Moong	941.797058	308.993506
5	Niger seed	-144.009443	249.534632
6	Safflower	-1031.251935	122.445887
7	Sesamum	1494.291172	279.404762
8	Sunflower	-4511.467222	223.906926
9	Urad	2670.533797	284.469697

Figure 25: Crops with lower increase in production but are increasing in price

The above crops are those that have lower increase in production but has increase higher increase in price. This shows that the increase in production of that crop is not as much as demand. These crops will be more profitable to produce. Threshold used for production is -10000 and for price is 100.

7.4) Finding crops that have lower rate of increase in cost per hectare than price

	index	cost_per_hec var	price var	grow_ratio
5	Moong	571.901298	308.993506	0.540292
0	Arhar/Tur	1155.433382	260.551948	0.225501
2	Gram	865.233108	169.123377	0.195466
1	Cotton(lint)	2477.784265	379.372294	0.153109
3	Groundnut	1623.278151	199.339827	0.122801
6	Rice	1253.114850	150.562771	0.120151
4	Maize	868.141613	68.593074	0.079011
8	Wheat	1122.494264	81.904762	0.072967
7	Sugarcane	3281.720563	10.367338	0.003159

Figure 26: Crops with lower rate of increase in cost per hectare than price

We tried to find out the crops that has lower rate of increase in cost per hectare but has increase in price of that crop more. This shows that these crops can give more returns.

8 Conclusion

The agriculture of a nation depends on several factors and its proper study is immensely useful. In this project, we have tried to get and organize agricultural data in a way in which it can be used for analysis. Visualization helps understanding the data better and hence, the data sets are visualized. The datasets are merged and studied regarding various factors that affect crop production. Simple statistical inferences help us to learn the changing pattern over the years, which motivates us to find the reasons behind these changing patterns. The factors that affect crop production are not always independent and lack of one of the factors can therefore affect crop production hugely. Using these data, something as significant as suicide rates can also be predicted with a good accuracy, which helps in better planning and taking preventive measures and formulating insurance policies. We have used linear regression and decision trees for prediction. The accuracy with decision trees is 83%. Decision trees makes the prediction more intuitive to understand, highlighting how each of the factors affect. Finding crops that have special changing patterns over the years like sudden decrease in production can help us understand the reasons behind them in a more specific way. We can also use these data to recommend new crops that can be grown in places which has the suitable weather and economy conditions. If similar analytical study is done and followed, it can prevent wastage of land and increase in production, which will help us to meet the demands of the people and also boost our economy.

9 Future Work

The study can be extended to large data sets with large number of attributes. The data can be taken more locally to study the various regions within the states. The above analysis is done by taking crop production, cultivation cost,crop yield, area under cultivation, farmer's suicide rate, production growth rate, temperature and rainfall as features. More features can also affect the production of crops. For example, We have taken features like rainfall, temperature data to show effects on crops production, price etc. Other features like Soil data (soil type, PH value),Weather data (winds, humidity) , fertilizers used,can also be taken into account for much better results. We can also improve the prediction and clustering by using some other algorithms. Apart from these, predictions of various other factors can be made, like crop production in the following year. If we will be able to predict the factors affecting the crop production, then we can plan the cultivation in a better way.

10 References

1. Crop Production Statistics
2. Data gov crops related data