Indian Agriculture Data Mining

CS685: Final Project Report [Group 18]

Sunil Dhaka
17817735
sunild@iitk.ac.in

Samyak Jain
180661
samyakj@iitk.ac.in

Sanjay Kumar
180669
sankumar@iitk.ac.in

Gaurav Kumar
170270
krgaurav@iitk.ac.in

Uday Kiran Damodara
150775
damodara@iitk.ac.in

Contents

Chapter 1

Abstract

India's agriculture history dates back to the Indus Valley Civilization. Agriculture sector has a major impact on the Indian economy as it contributes 18% to the country's GDP and 50% of the Indian workforce. Agriculture is demographically the broadest economic sector and plays a significant role in the overall socio-economic fabric of India.

India ranks second in worldwide farm output. India ranks first with highest net cropped area followed by the US and China. India is among the top three global producers of many crops, including wheat, rice, pulse, cotton, peanuts, fruits and vegetables. India has shown a steady average national average increase in the mass produced per hectare for some agricultural items, over the last 60 years. These gains have come mainly from India's Green revolution, improving infrastructure and modernization. Despite these recent accomplishments, agriculture has the potential for major productivity and total output gains, because crop yields in India are still just 30% to 60% of the best sustainable crop yield over other developed countries.

Chapter 2

Problem Statement

We aim to analyze Indian crop production data-set and extract key insights from it. Also, plotting the production vs other variables to easily visualize the insights obtained from the data-set.

Our problem statement can be divided into multiple tasks as follows-

- Data-set structure and in depth variable analysis (categorical and numeric)
- State wise, zone wise, crop wise, year wise, season wise, crop-category wise plots of production
- Find states dominating each category of crop
- Find most frequent crop and its geographical analysis
- Production area wise analysis of states
- Analysis of Northern and Southern parts of India

Chapter 3

Introduction and Motivation

In this project we analysed the data and presented before you the various insights and tools that we were able to extract and develop from the data. These tools and key insights can not only help us understand agriculture sector inefficiencies better and help them, but also help in improving the efficiency of the sectors by better managing their resources, targeting maximum throughput. It will have lots of major and minor facts which will help in charting a next successful revolution after 1965.

Chapter 4

Dataset used

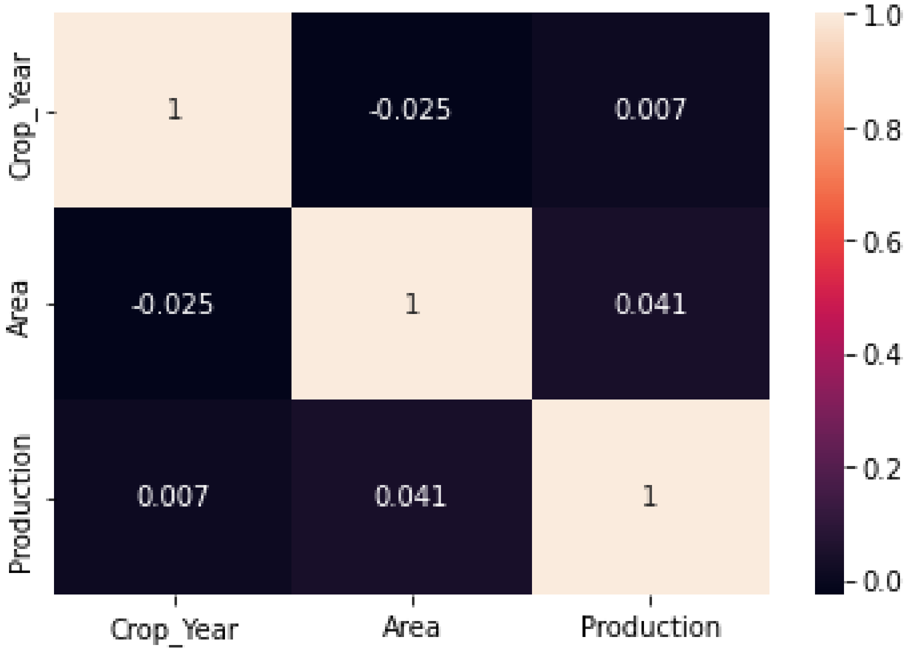
- The data-set used is available on the data.gov.in website.
- **Downloading the Dataset**
  The Dataset was downloaded using data.gov API in .csv format. For more info go to **data.gov**. How we have obtained data is given in below.
  what we are doing to get our datasets?
  - first, one needs an api-key to request full dataset. It is a simple sign-up process.
  - we include personal api-key in a 'login.py' file as a variable
  - import login and then access to use it
  - we request this particular dataset in particular format
  - as there are around 2.5 lakh data points we set upper-limit for api-url to 10 lakh to get whole dataset(trial and error). it might differ for other datasets
  - then we save requested content into a csv file
  - to use dataset we read it from local stored datasets folder rather than requesting it again and again.
- Total size of the downloaded data-set is 14.6 MB

Chapter 5

Methodology

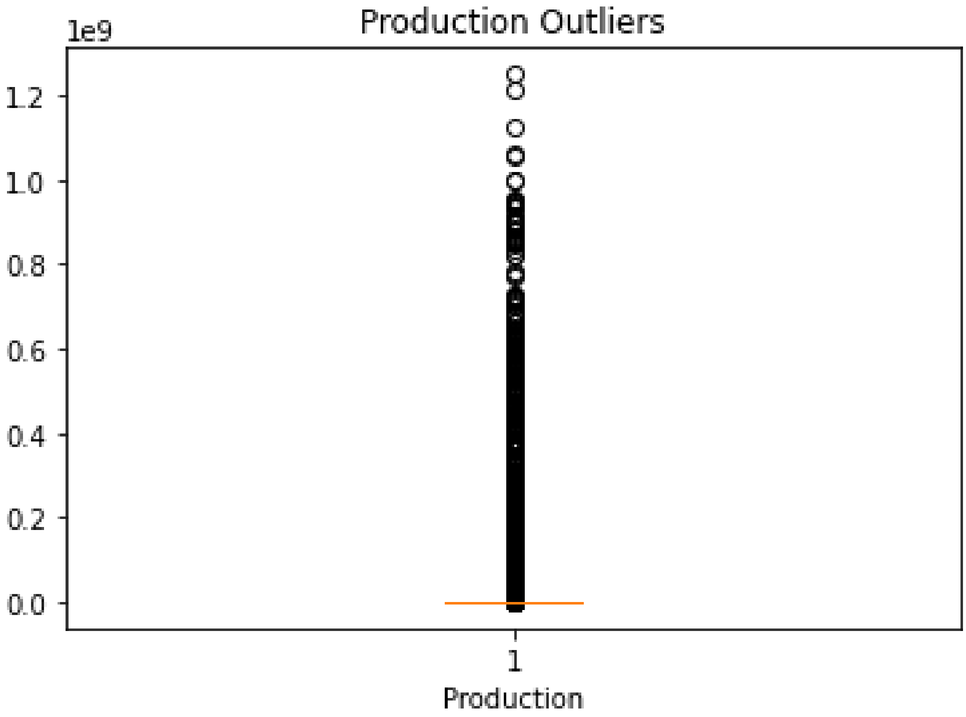
5.1 Data-set structure and variable analysis

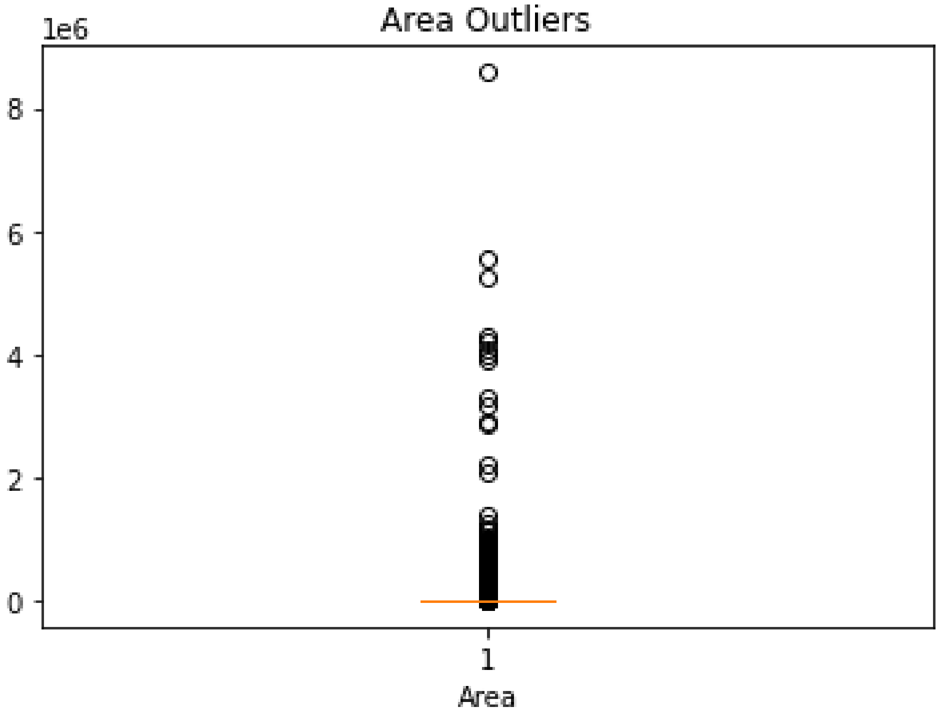
- Dataset consists of 246091 rows and 7 columns
- Out of 7 variables, we have
  - 4 categorical variables
    - State_Name
    - District_Name
    - Season
    - Crop type
  - 3 continuous variables
    - Area (float)
    - Production (float)
    - Crop Year (int)
- Missing points is Production col. 3730 NULL values (1.32%) are present in Production column. Others columns do not have Null values.
- Before with NULL values(in 'Production' col) we had 246091 data-points, now after dropping them we have 242361 sample size.
- Variable correlation



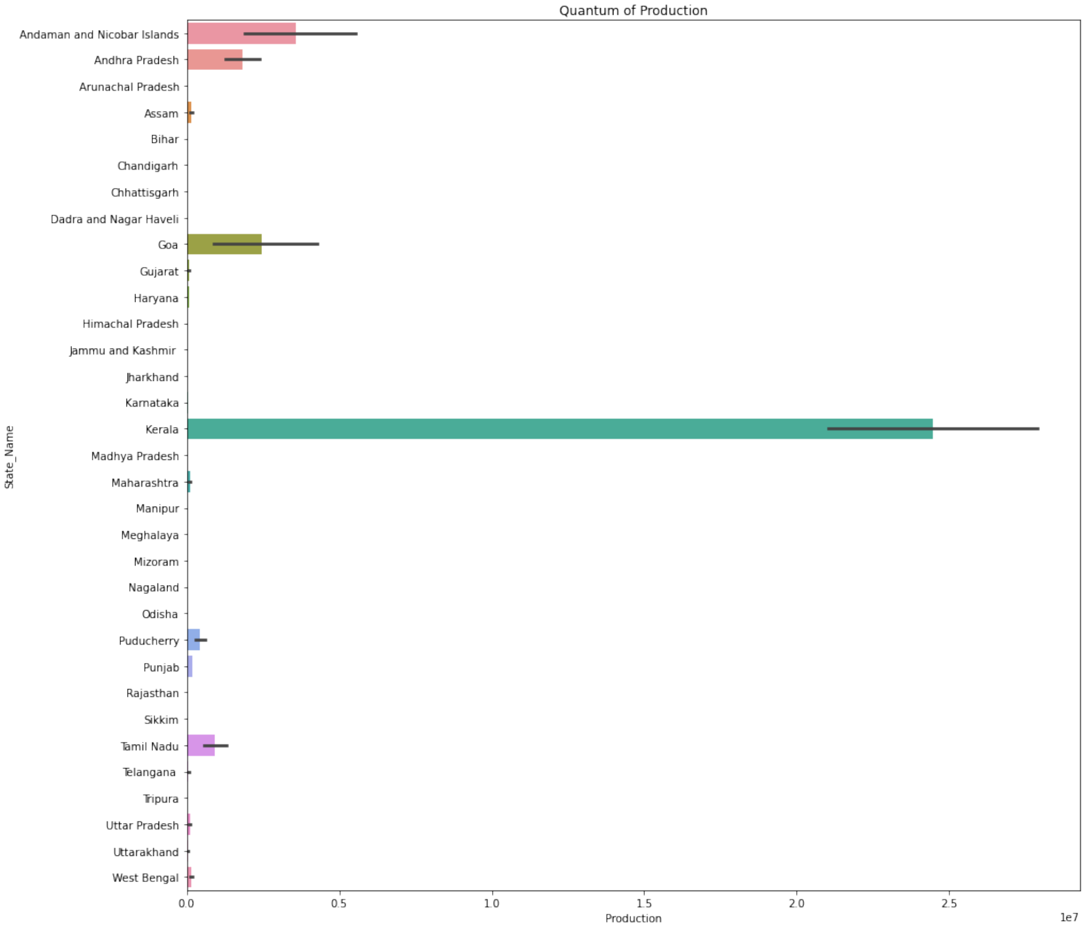
There is no variable showing high correlation with any other variable in the dataset.

- There are total 33 states in this dataset.
- On District front, we have have more data coming from Tumkur, Belgaum, Hassan, Bellary and Bijapur from Karantaka state.
- Our Data-set has data for 19 years from 1997 to 2015.
- Data-set consists of six different seasons i.e. Kharif, Annual, Autumn, Rabi, Summer and Winter crops.
- Data-set consist of 124 different crop varieties
- Production values range from 0(min) to 1250800000(max)
- Area and Production columns are highly skewed with lot of outliers. This is beacause most of Indian farmers cultivate on small farm lands unlike other developed countries where there are cultivation companies that do the farming; there are rarely farms that big in India. And as Production does follow area hence skewness is also present in it.





- Kerala is top state when we look at the quantum of Production for last 19 years. Area (size) of particular state also gives top rank to some states. It is not quite bias-free but it is useful.



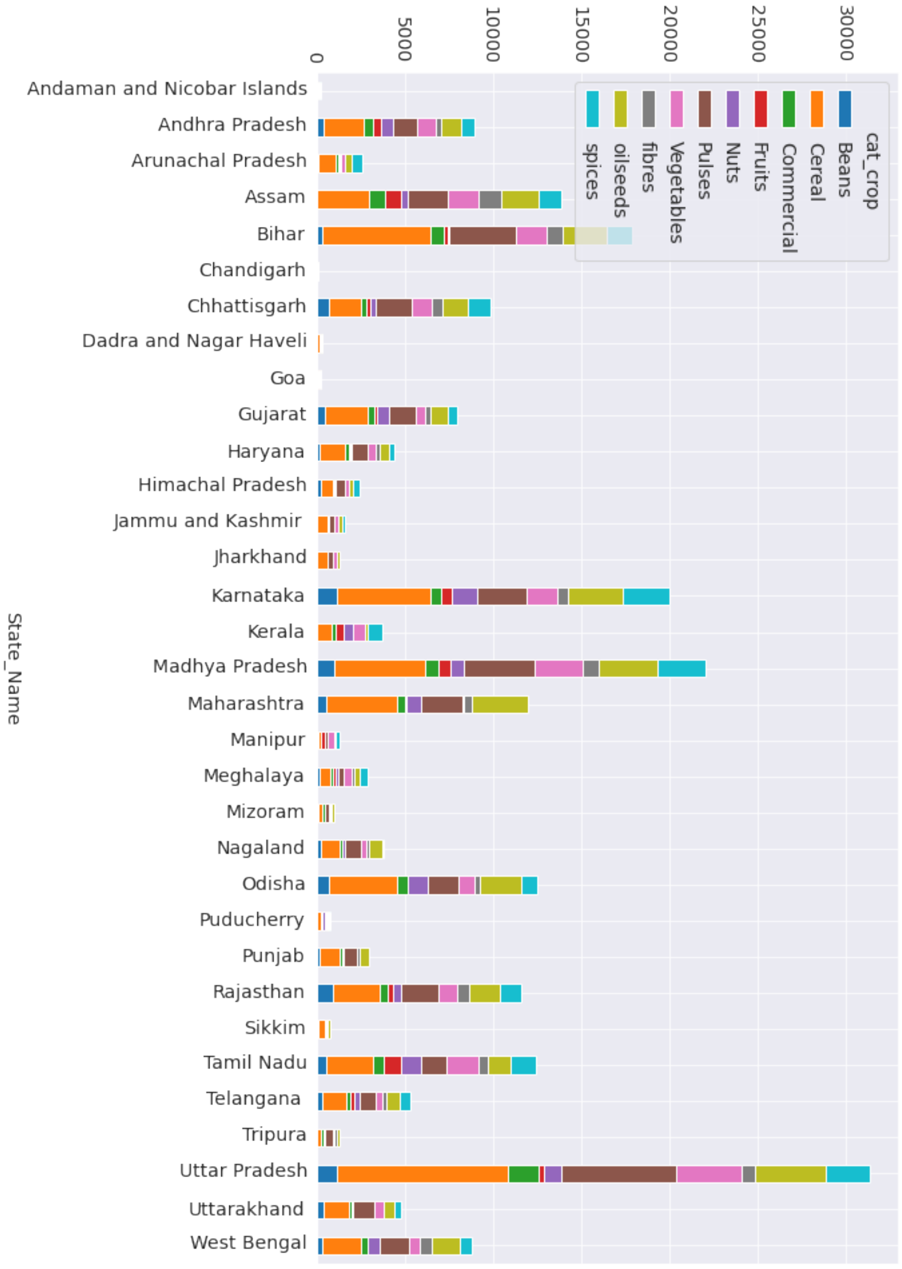
5.2 Creating Additional variable

- Different zones (Union Territory, South Zone, NE Zone, East Zone, North Zone, Central Zone and West Zone)
- Different categories (Cereal, Pulses,Fruits,Beans,Vegetables, Spices, fibres, Nuts, Natural Polymer,Coffee, Tea, Total food grain, Pulses, Oil seeds, Paddy, Commercial, Sugarcane, forage plants and Others)
- Note that we have considered UTs separately from states, to make comparison interpretable and fair.

5.3 Find states dominating in each category of crop

- Table using State Name and Crop categories

- Construct bar plot



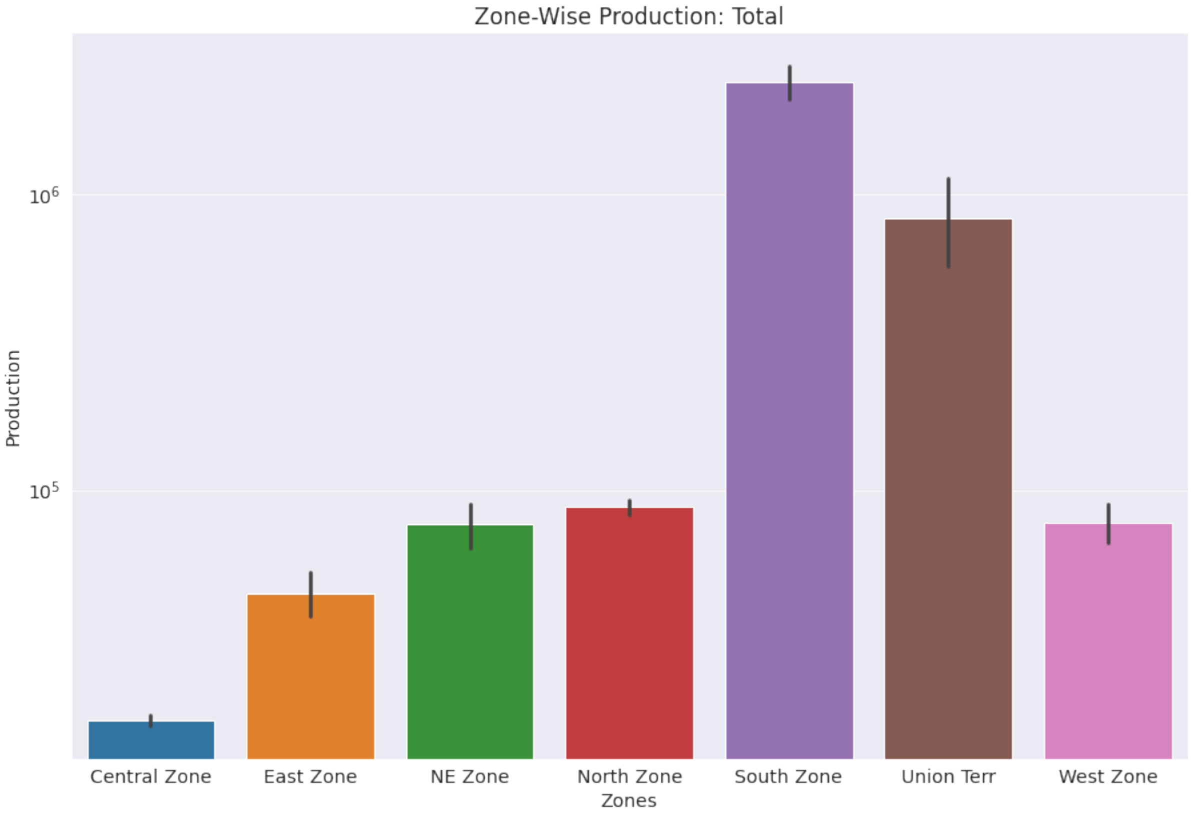
5.4 Most frequent crop and its geography

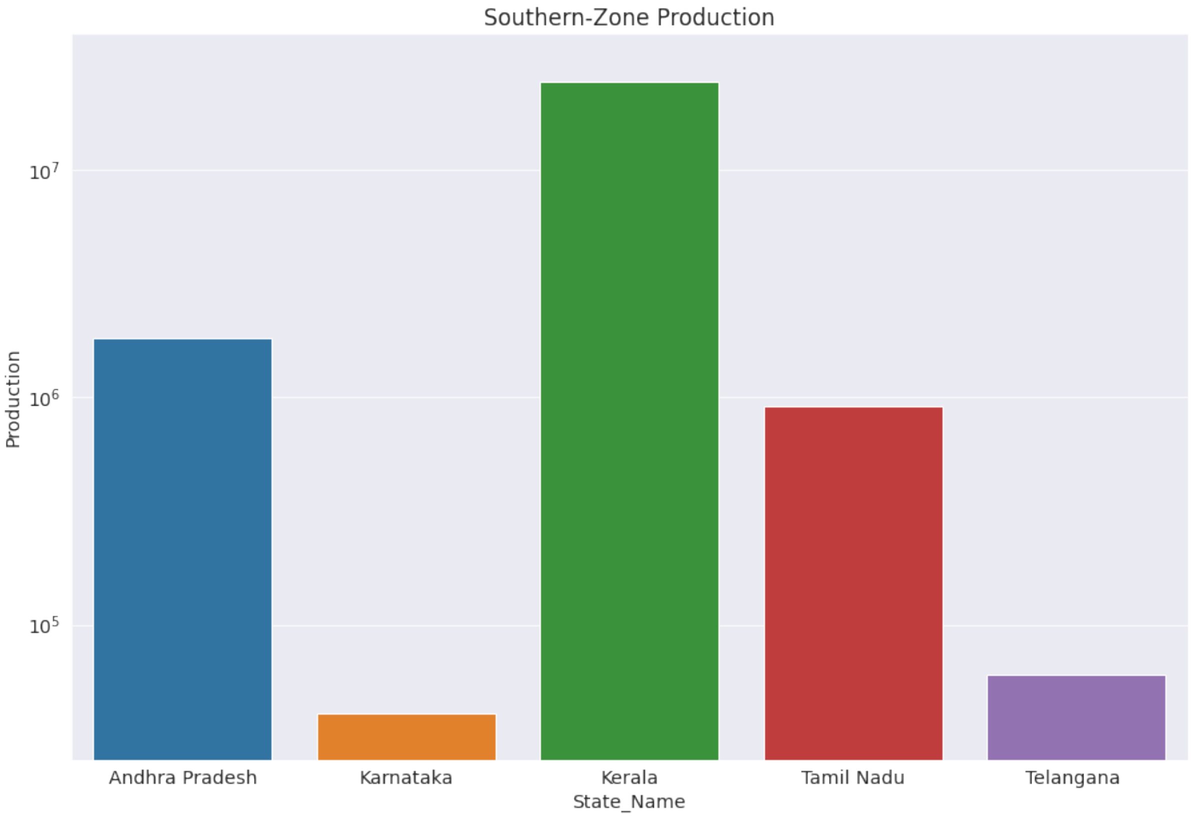
- Rice is grown heavily when we look the frequency of crops in India

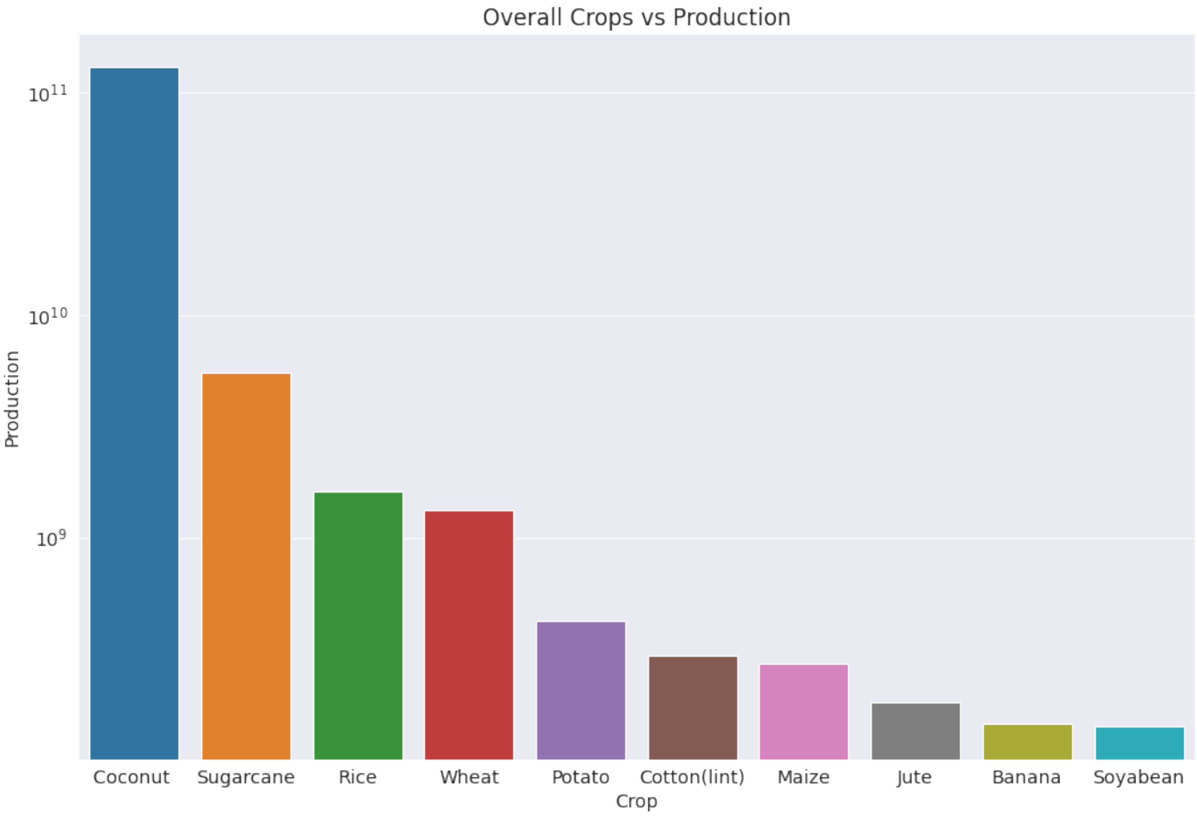
- Winter season is most importent for rice
- Statewise Punjab dominates in rice production
- District wise its BARDHAMAN(2.13
- Yearwise 2014 is the year when production reached the peak production
- Correlation between Area and Production shows high production is directly proportional to Area under cultivation.

5.5 Visualization on variables

5.5.1 Zone wise

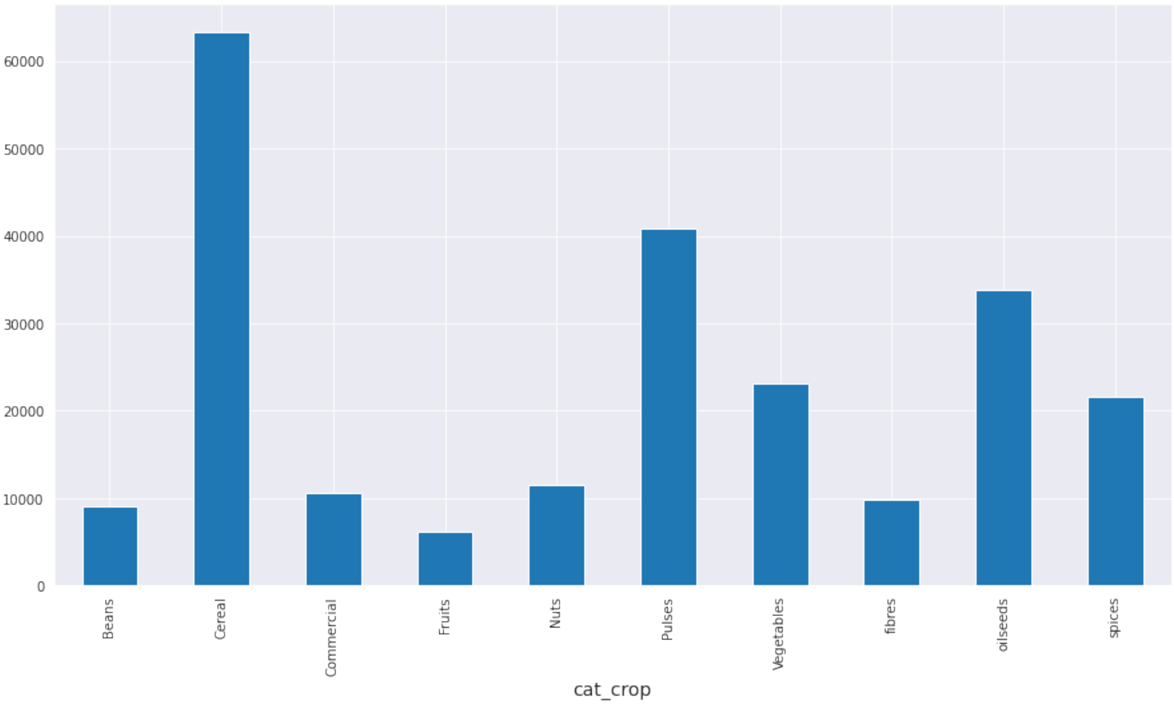


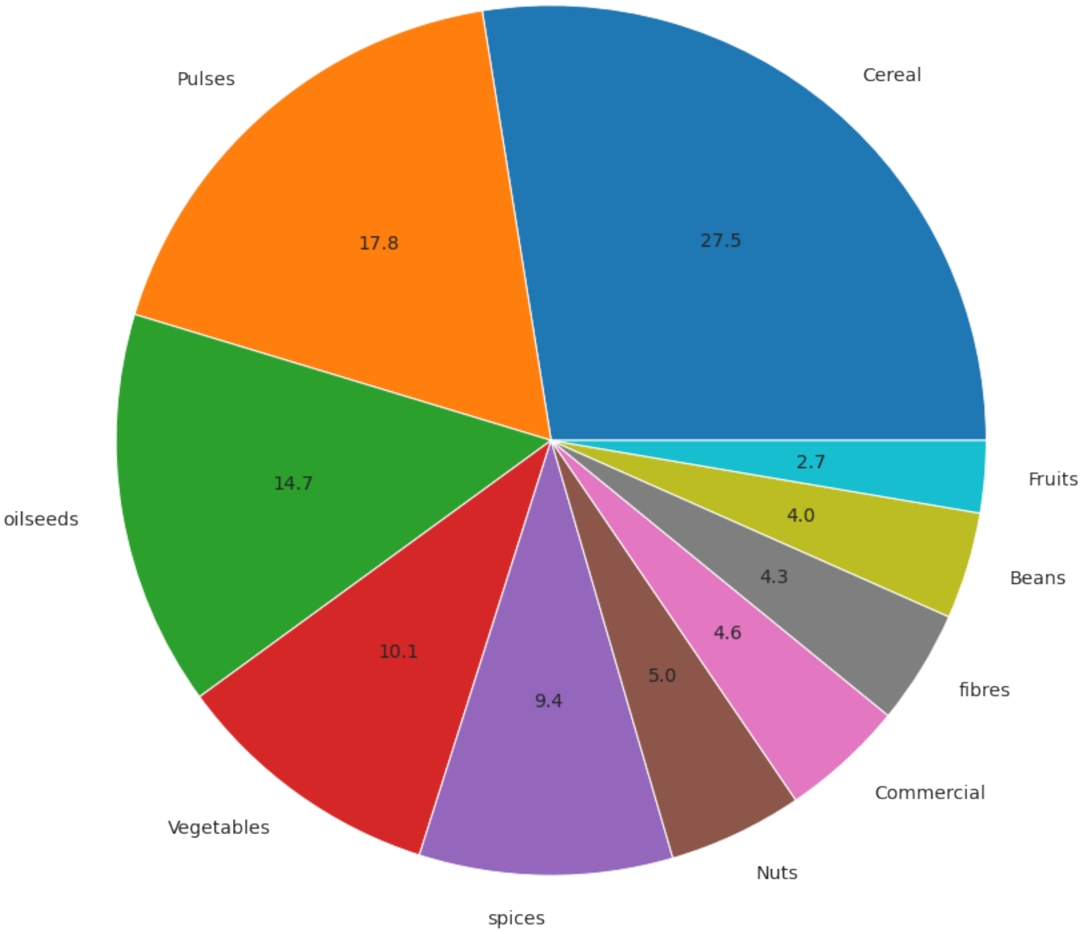

Southern-Zone Production
Production
10^7
10^6
10^5
Andhra Pradesh
Karnataka
Kerala
Tamil Nadu
Telangana
State_Name



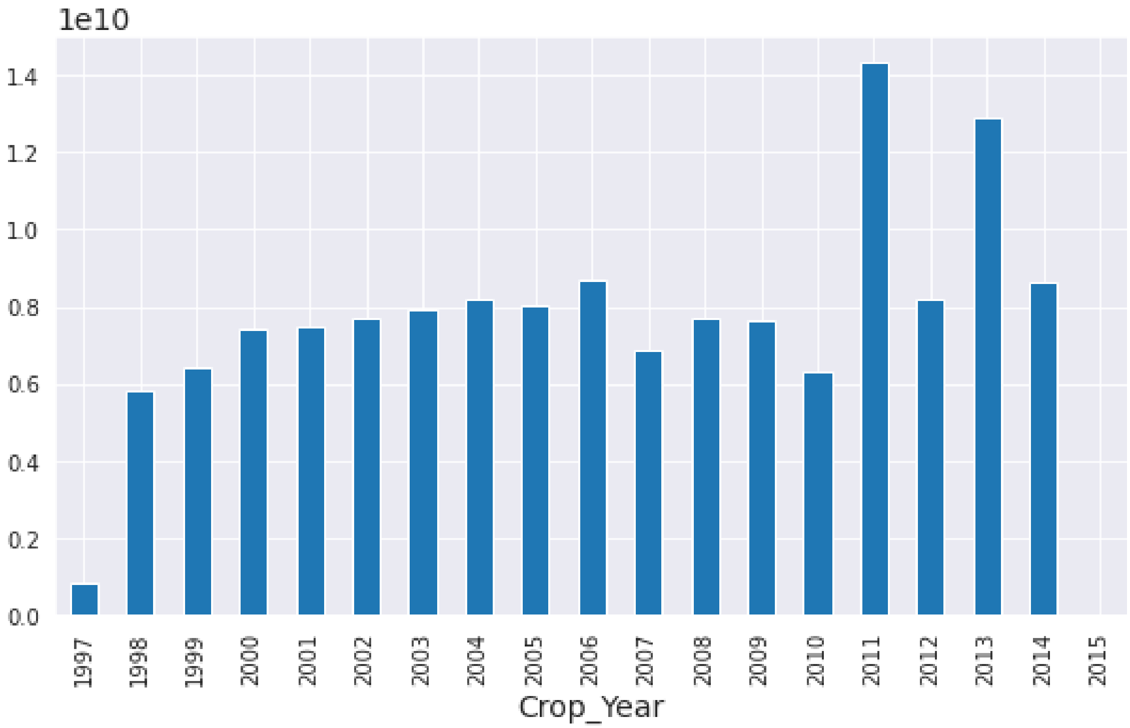
5.5.2 Crop wise

5.5.3 Crop Category wise

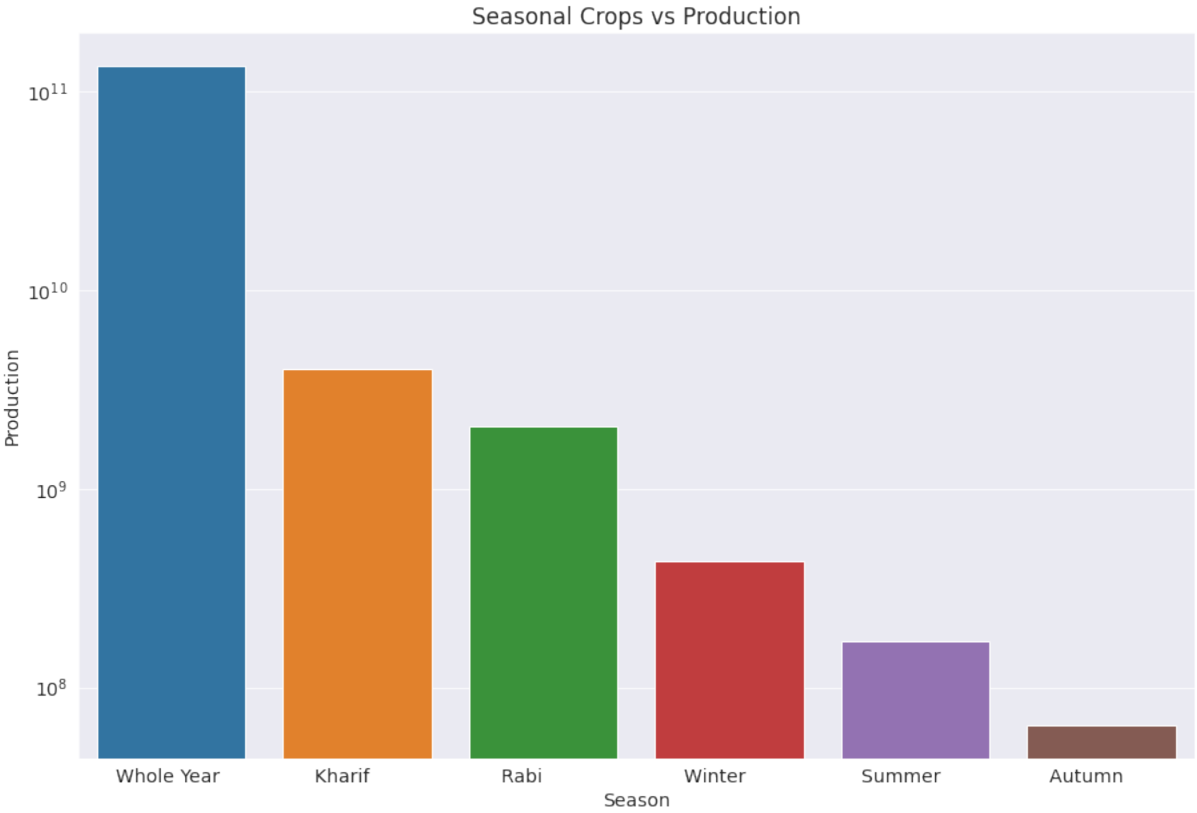



Cereal
27.5
Pulses
17.8
oilseeds
14.7
Vegetables
10.1
spices
9.4
Nuts
5.0
Commercial
4.6
fibres
4.3
Beans
4.0
Fruits
2.7

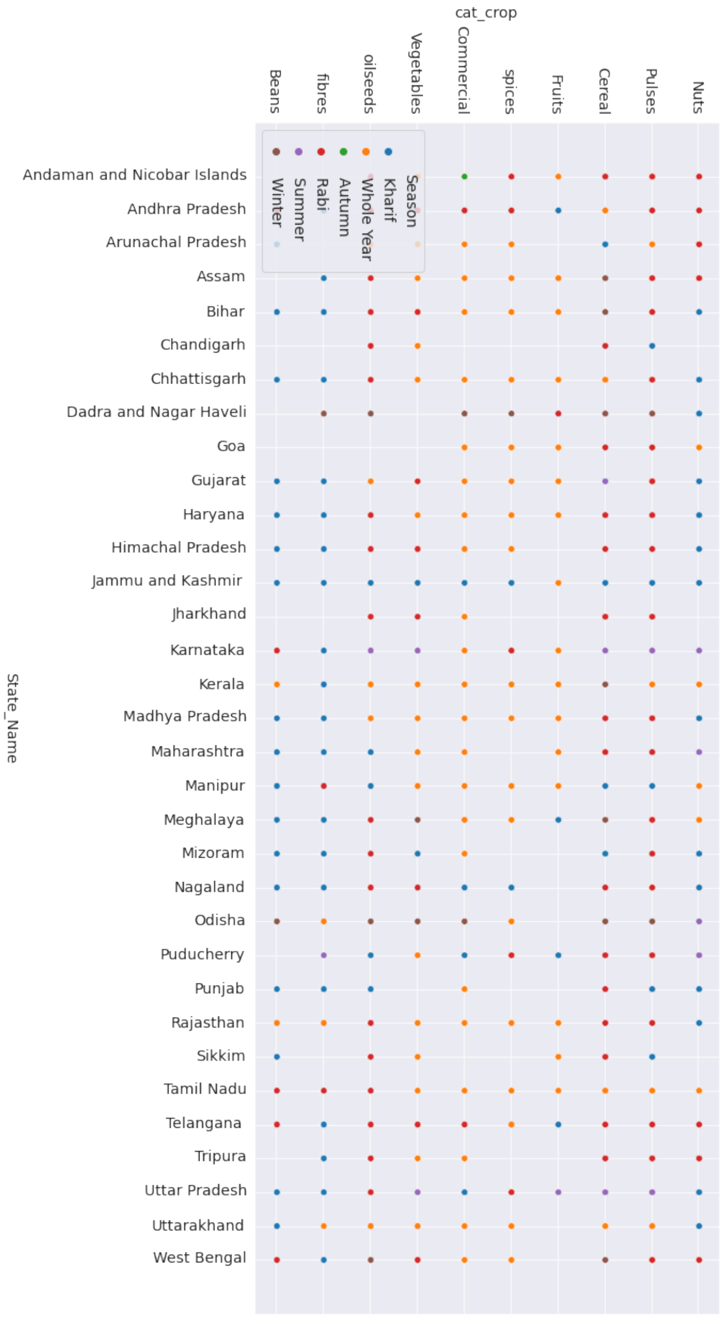
5.5.4 Year wise



5.5.5 Season wise

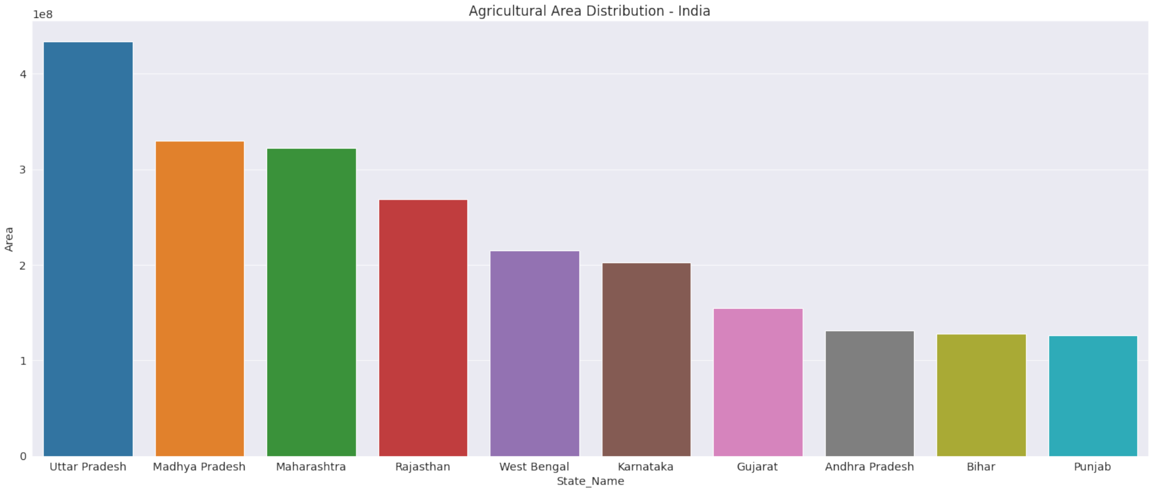


5.5.6 State vs Crop category vs Season

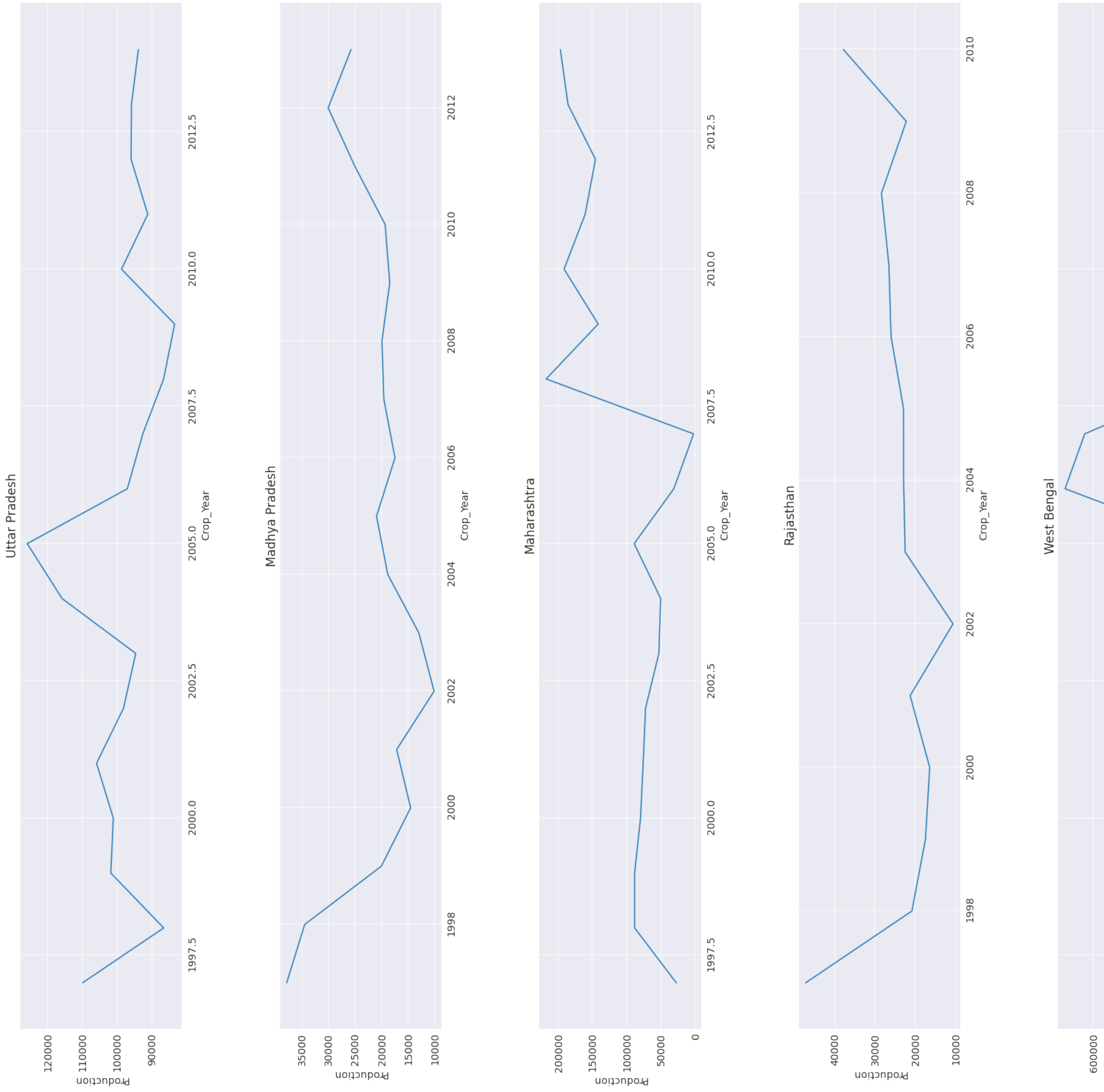


5.6 Production area wise analysis of states

- Find the state where Area under crop production is high and get top 10 states

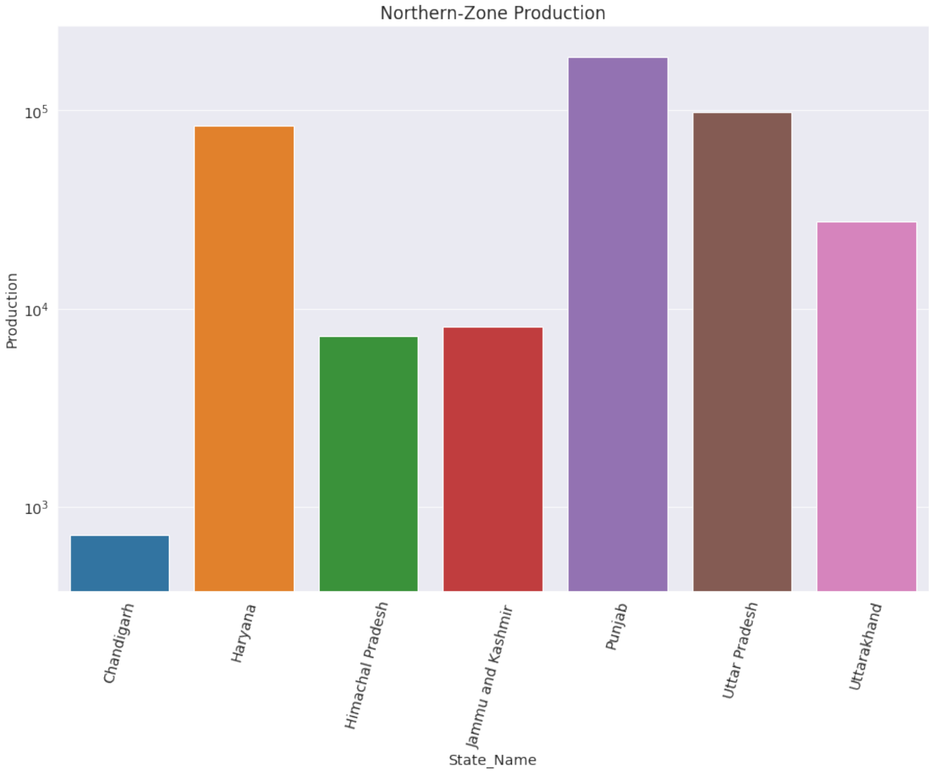


- Make a subset of the data with the above conditional data
- Look the cultivation status of these states year-wise

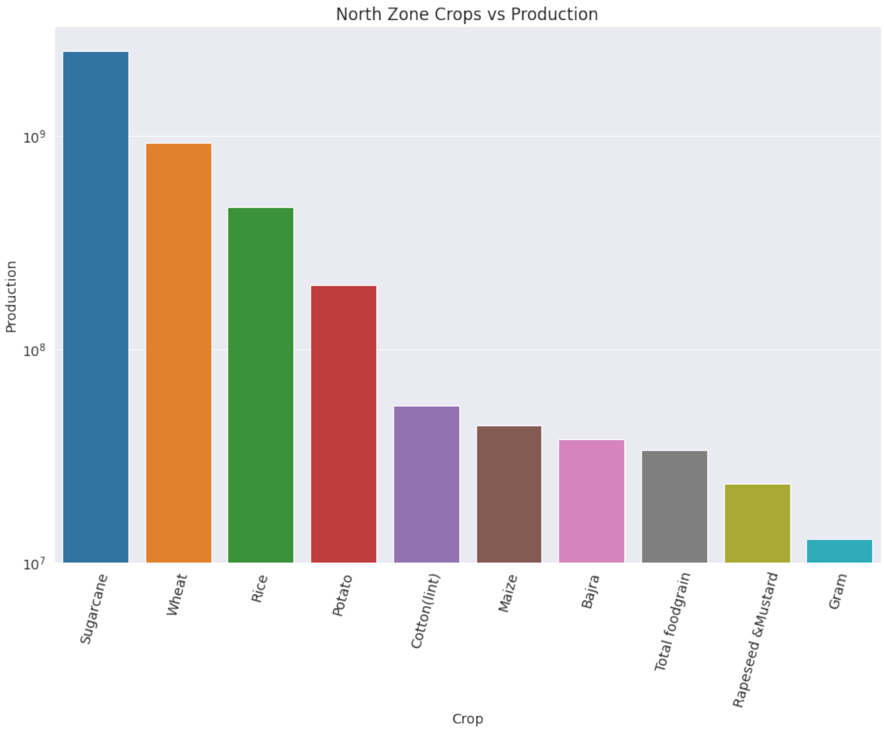


5.7 Analysis of Northern parts of India

- Extract North zone data as a subset



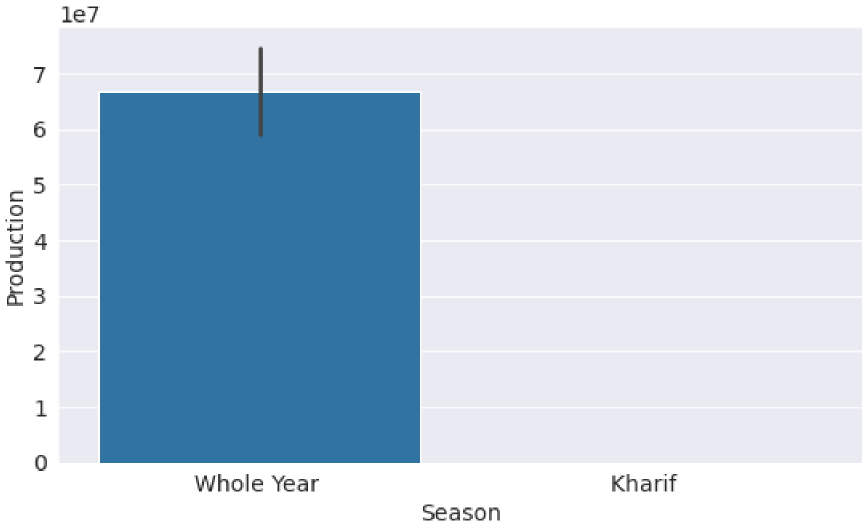
- Explore the top states of North which show high crop production stats



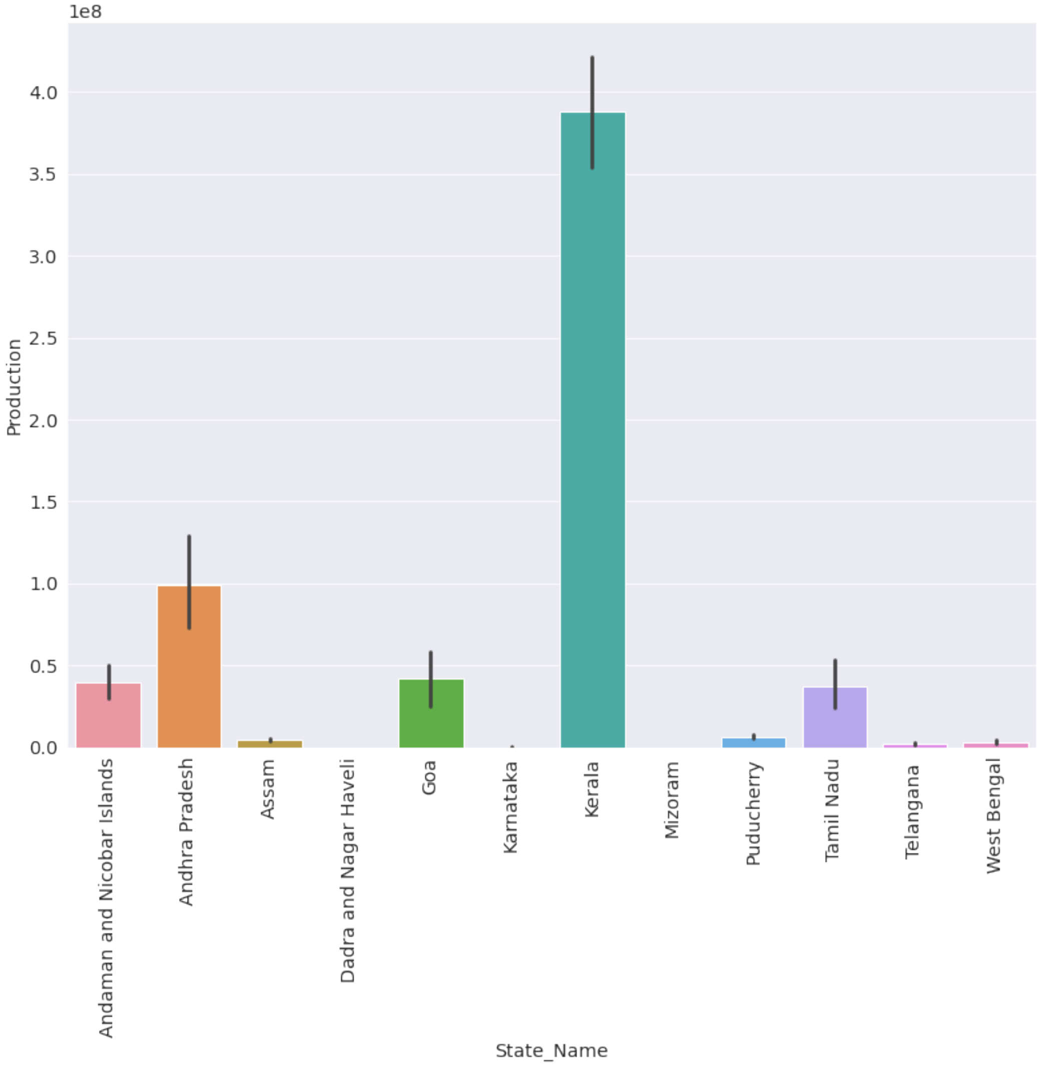
- Explore the top crops grown in North India

5.8 Analysis of South India

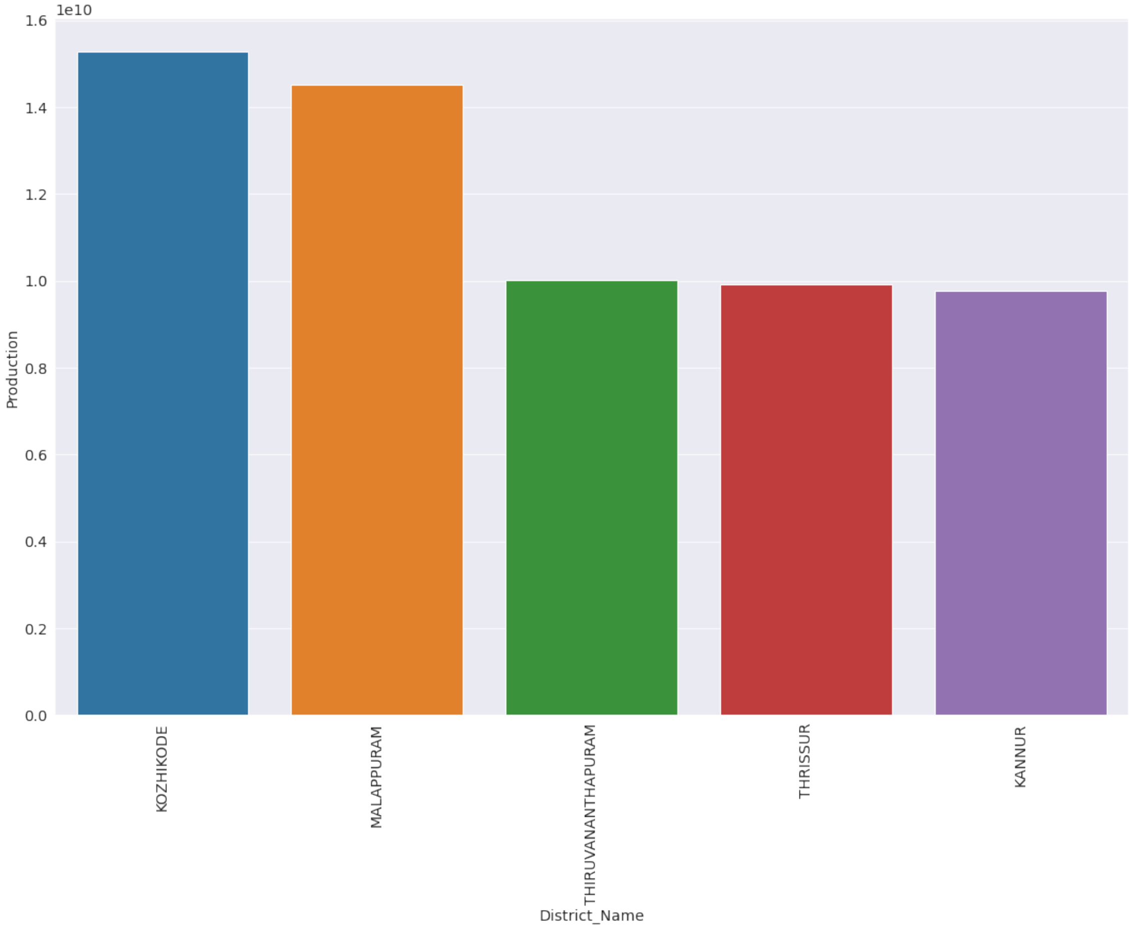
- Draw a subset of data for Coconut Production in India
- Find the conducive season for coconut production



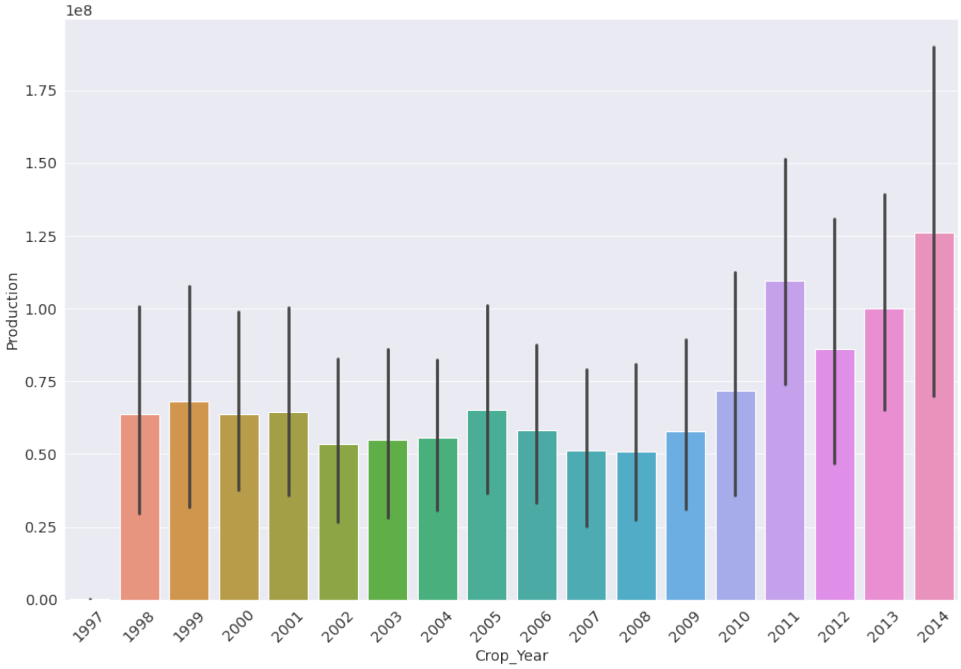
- Explore the states involved in coconut production



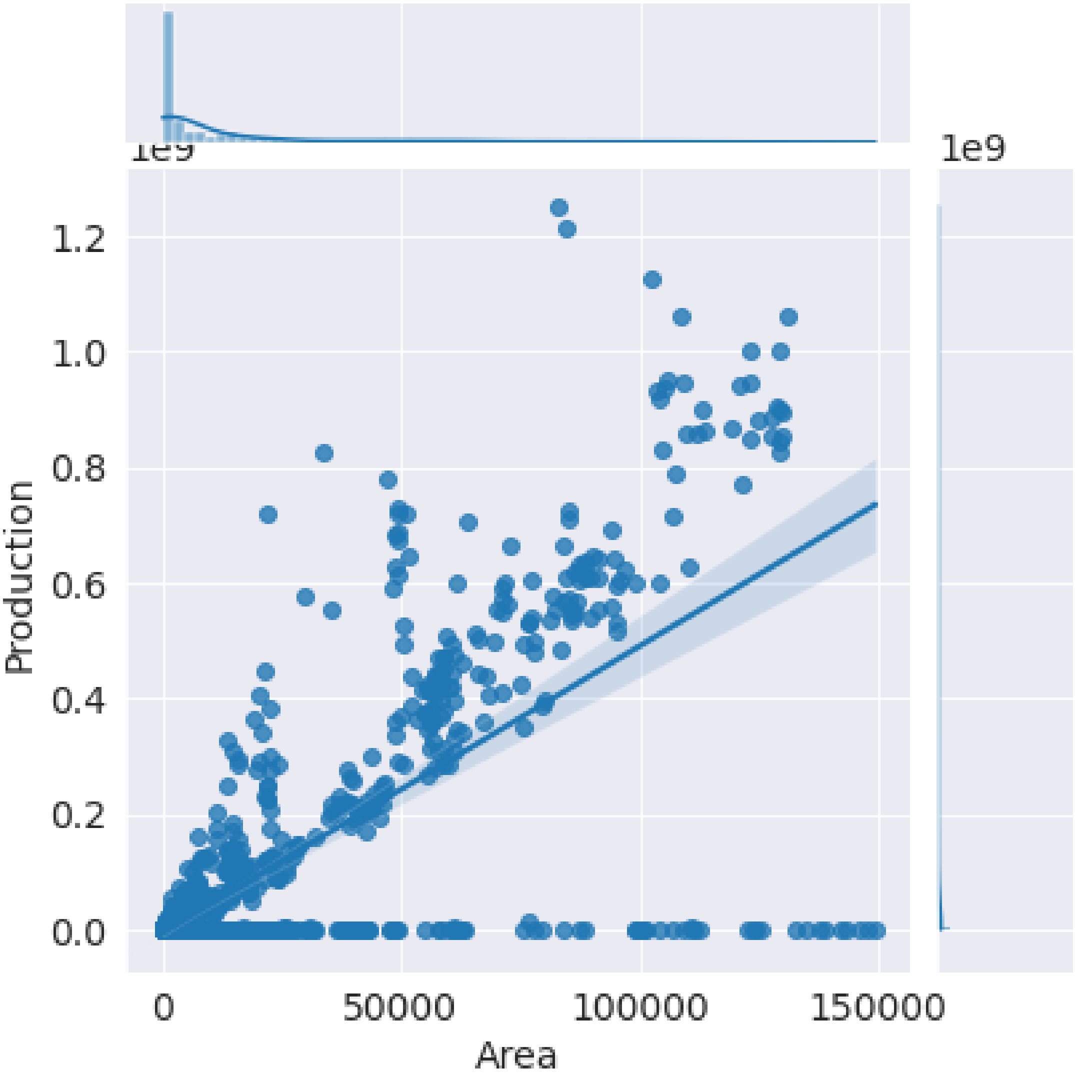
- Explore the southern district involved in coconut production



- Find the cultivation status of coconut year-wise



- Find the Area under cultivation in South India


1e9
1e9
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Production
0
50000
100000
150000
Area

Chapter 6

Spatial Analysis Results

6.1

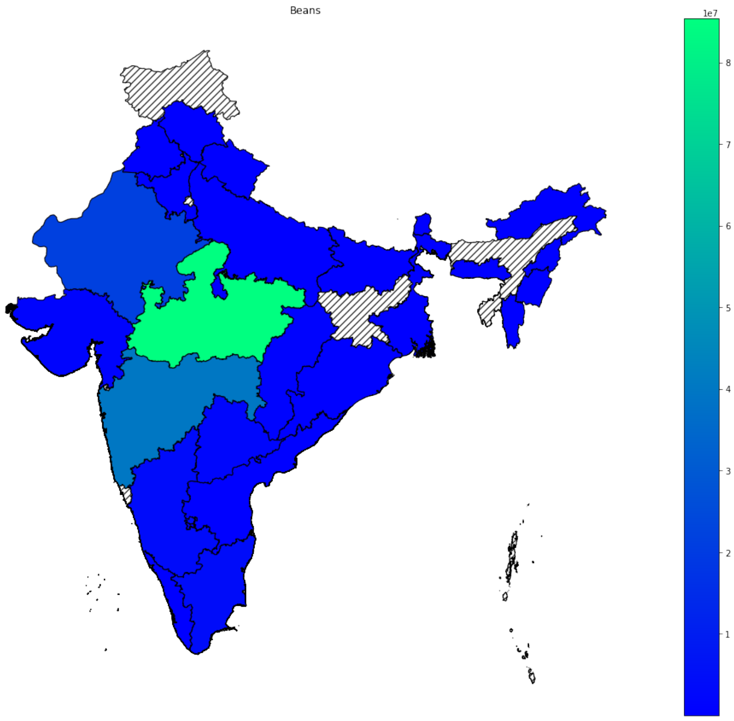
Observations: - In oilseeds Rajasthan and WB are top producers over 19 years. - Gujarat and Karnatka are two top states in the production of Nuts. - In fibers WB, Gujrat, and Maharastra are top producers over 19 years. - KT, AP, TG are some top states in the production of Spices. - In beans MP and MH are top producers over 19 years. There is no beans production data for JH and Assam. - TamilNadu tops in the production of Fruits. - In Creals UP, PB and HR are some top producers over 19 years. - In Pulses MP, UP, and MH are some top producers over 19 years. - UP and Wb are two top states in the production of Vegetables. - UttarPradesh tops in the production of commercial crops.

6.2 Top Category Crops Analysis

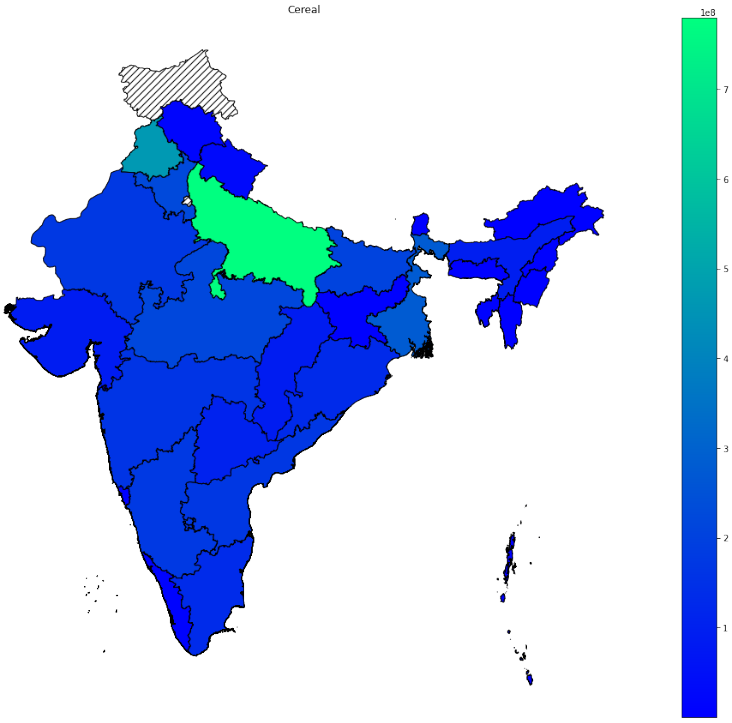
Observatoins: - States in North and East specifically Uttar Pradesh, West Bengal, Punjab and Andra Pradesh are top producers of Rice in India over 19 years(1997-2015) - States in North specifically Uttar Pradesh, Punjab and Hariyana are top producers of Wheat in India over 19 years(1997-2015) - There are only four states that produce Paddy and Assam is the top producer of Paddy in India over 19 years(1997-2015)

6.3 Zonal Analysis

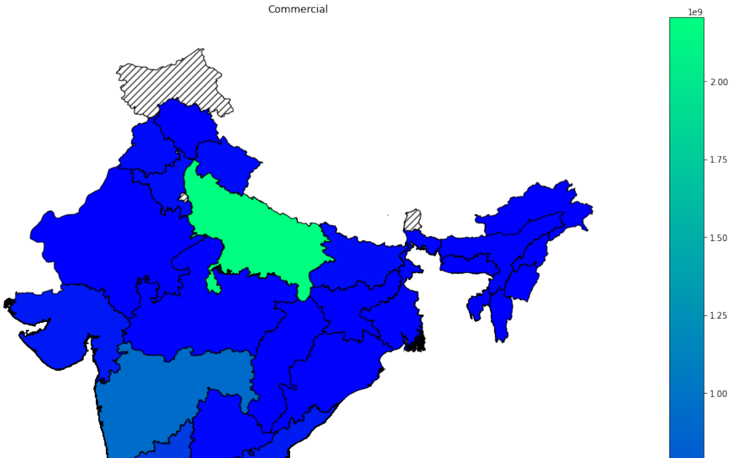
Observations: - Top producers of crops in each region: - 'North Zone': Uttar Pradesh - 'South Zone': Kerela - 'East Zone': West bangal - 'West Zone': Maharastra - 'Central Zone': Madhay Pradesh - 'NE Zone': Assam - 'Union Terr': Andman and Nicobars - Most of the result are not new to our intutions. All states that are in top rank has at least one major river[that flows throughout the year] passing through their cultivation zones.

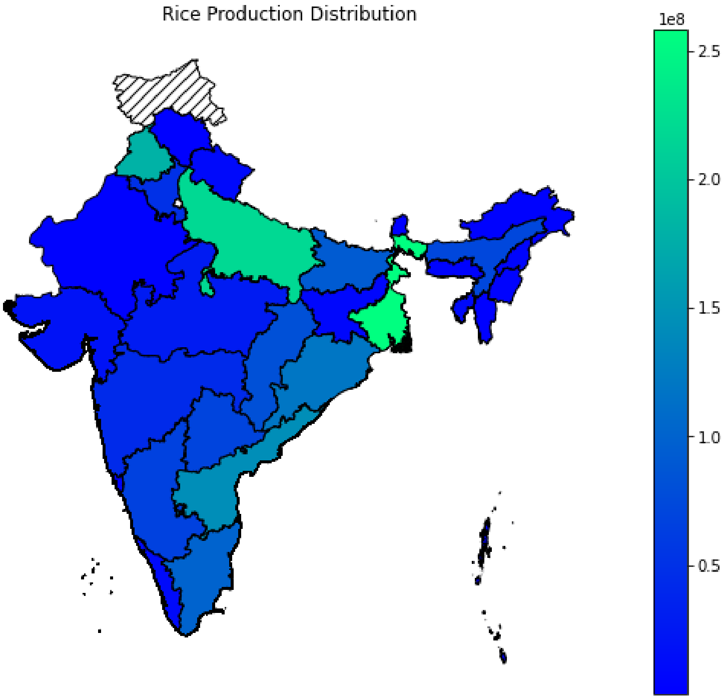

Beans
1e7
8
7
6
5
4
3
2
1

.5 .5

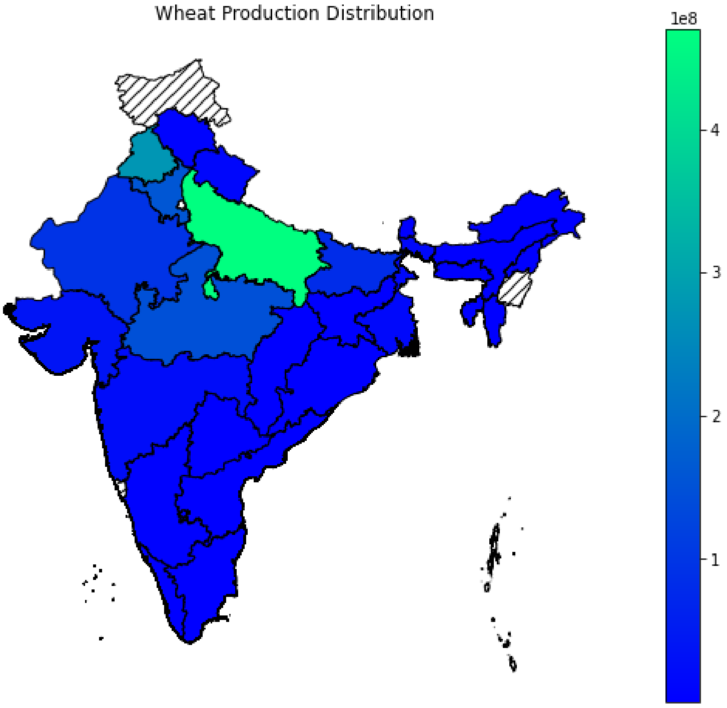

Cereal
1e8
7
6
5
4
3
2
1

.5

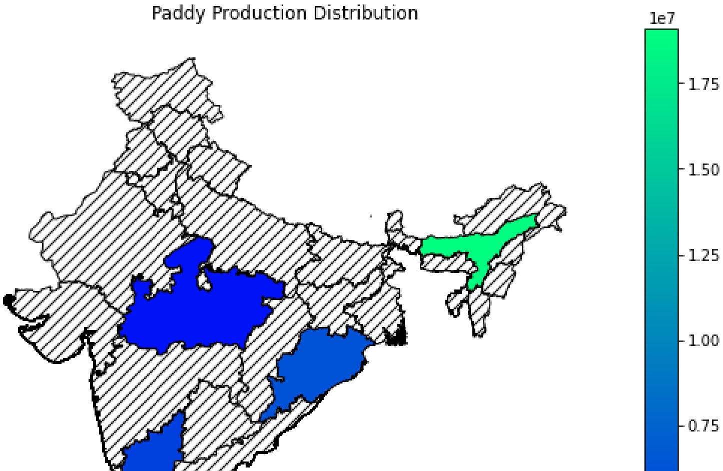

Commercial
1e9
2.00
1.75
1.50
1.25
1.00

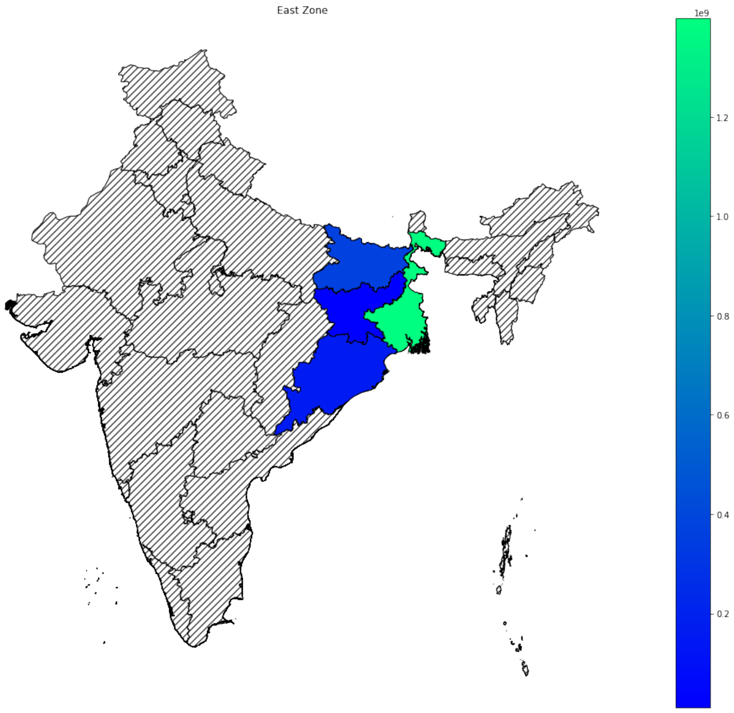

Rice Production Distribution
1e8
2.5
2.0
1.5
1.0
0.5

.5 .5

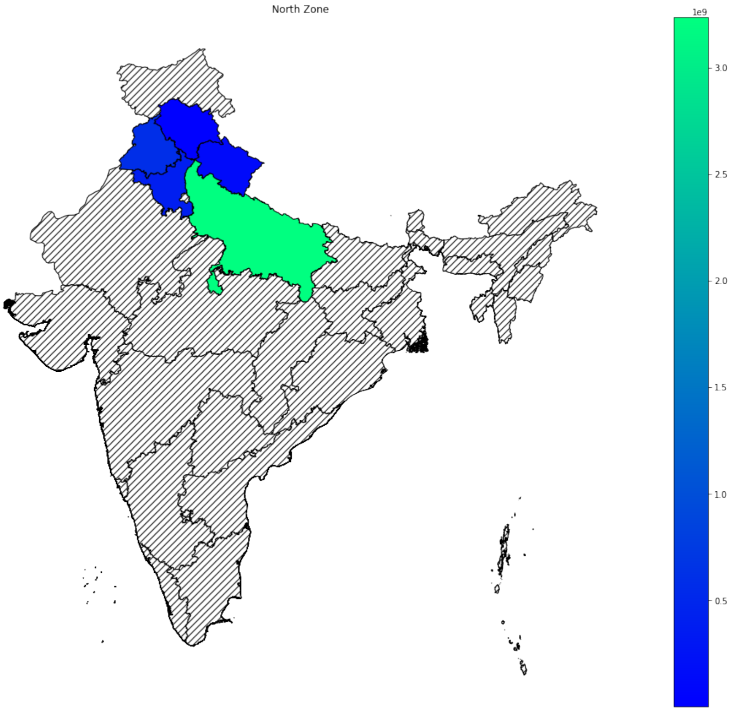

Wheat Production Distribution
1e8
4
3
2
1

.5

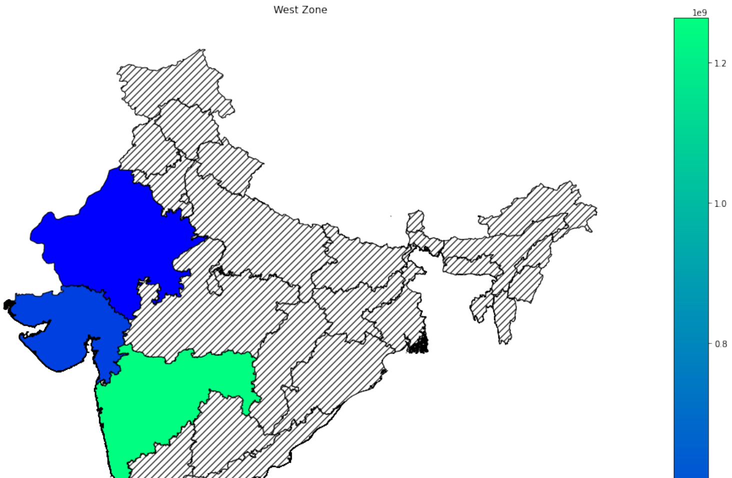

Paddy Production Distribution
1e7
1.75
1.50
1.25
1.00
0.75


East Zone
1e9
1.2
1.0
0.8
0.6
0.4
0.2

.5 .5

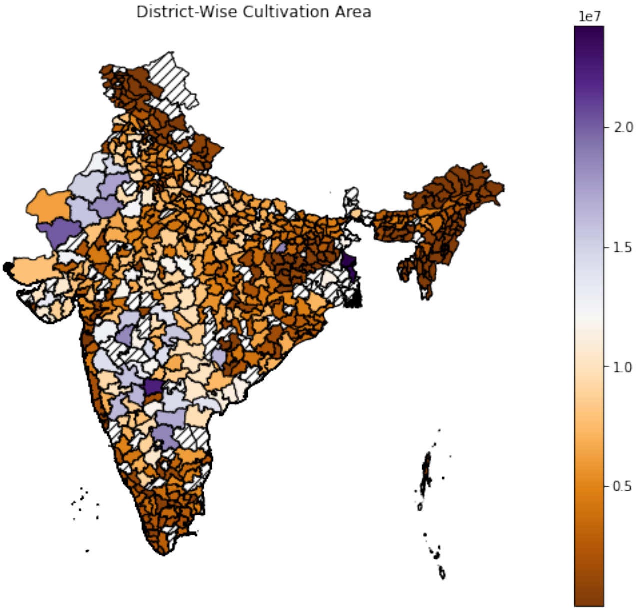

North Zone
1e9
3.0
2.5
2.0
1.5
1.0
0.5

.5


West Zone
1e9
1.2
1.0
0.8

6.4 District Wise Cultivation Area

Observations: - Quite clearly Thar Desert[in Rajasthan] districts have very less cultivation area - Some in between districts[mainly in central zone due to rocky terrain] that have low cultivation are districts that have water shortage like not having mansoon rain, river or any other source of water. - In North-East region cultivation area is huge due to heavy mansoon rains, major rivers flows through or near them; basically they have all the waters. - Southern area also has quite heavy rains in mansoon season and rivers are also there; so as expected they have higher cultivation area. - Indeed Ganga Plains have high cultivation areas, as expected.


District-Wise Cultivation Area
1e7
2.0
1.5
1.0
0.5

Chapter 7

Future Work

7.1 What next??

As always there is scope for improvement and here we include some of the tasks or points that can be explored further. This analysis is just a tip of iceberg, with nineteen year crop production data, a lot could be done and some of the ideas are:

- Instead of deleting missing data for Production(3730 data points), we could impute based on the area used for cultivation and state.
- Zone wise cultivation status and predict future production prediction using regression.
- Crop Categories and status of their cultivation obver the years, if the production has gone up (Good case scenerio) and if production is gone down (bad case scenerio)...can we look into the causation of this trend.
- Asking further important questions like, Kerela is low in area coverage compared to other southern states but still in production levels its high why?
- We can explore each zone separately on deeper level, meaning on district level.
- After predicting missing production values we can re-explore entire file with those points included. Although it shall not make much difference.
- Each crop category can be analysed separately to see which crop dominates a particular crop category.