

Exploratory Data Analysis

School of Computer Science and Engineering,

VIT Bhopal University

DSN3099: EPICS

Dr. Kanchan Lata Kashyap

Assistant Professor

**Acknowledgment**

**To be changed-**

*We sincerely thank Dr Kanchan Lata (Assistant Professor, VIT Bhopal University) for assisting with case selection.*

*We appreciate the assistance of the staff of the School of Computer Science And Engineering of VIT Bhopal University.*

*Special thanks go to Mari Beer (ABC Editing Company) for useful comments on and for editing an earlier draft of this manuscript.*

**Abstract**

Exploratory data analysis (EDA) plays a crucial role in understanding complex datasets and extracting meaningful insights. In this research paper, we present an EDA of crop production in India, aiming to uncover patterns, trends, and potential factors influencing agricultural outcomes.

Using a comprehensive dataset spanning multiple years, we examine various dimensions of crop production, including crop types, yield, geographical distribution, and socio-economic factors. Through descriptive statistics, data visualization techniques, and hypothesis testing, we gain valuable insights into the dynamics of crop production in India.

Our analysis reveals significant variations in crop yields across different regions and crop types. We identify key factors, such as rainfall, temperature, soil fertility, and access to irrigation, that impact crop productivity. Additionally, we explore the relationship between agricultural practices, such as fertilizer usage and crop yield, and investigate the influence of socio-economic factors, such as landholding size and education level of farmers.

Furthermore, we utilize time series analysis to examine temporal trends in crop production, identifying seasonal patterns and long-term changes. We investigate the impact of government policies, such as subsidies and crop insurance, on agricultural outcomes.

Overall, our exploratory data analysis provides a comprehensive overview of crop production in India, highlighting the complex interplay of natural, economic, and social factors. These findings can serve as a basis for policymakers, agricultural researchers, and stakeholders to make informed decisions and design effective interventions aimed at improving crop productivity, ensuring food security, and promoting sustainable agriculture in India.

**Introduction**

The agricultural sector plays a vital role in India's economy, providing livelihoods for a significant portion of the population and contributing to the country's food security. Understanding the patterns and trends in crop production is crucial for agricultural planning, policy-making, and sustainable development. Exploratory Data Analysis (EDA) is a powerful technique that allows researchers to gain insights, identify patterns, and uncover relationships in datasets. This research paper aims to conduct an EDA of crop production data in India to provide a comprehensive understanding of the agricultural landscape.

**Objective:**

The primary objective of this research paper is to explore and analyse the crop production data in India using EDA techniques. The specific goals include:

a) Examining the temporal trends and variations in crop production over a specific time period.

b) Identifying the dominant crops and their spatial distribution across different states or regions.

c) Analysing the relationship between crop production and key factors such as rainfall, temperature, and agricultural practices.

d) Identifying any significant changes or shifts in crop patterns over time. e) Providing actionable insights for policymakers and stakeholders to enhance agricultural productivity and sustainability.

**Implications and Significance:**

The findings from this research paper can have significant implications for agricultural policymakers, farmers, and researchers. The insights gained can inform decisions related to crop selection, resource allocation, infrastructure development, and climate-resilient agricultural practices. By understanding the patterns and dynamics of crop production in India, stakeholders can work towards enhancing food security, promoting sustainable agriculture, and improving the livelihoods of farmers.

**Data Collection and Methodology**

**Data Sources**

The research paper utilizes publicly available datasets from reliable sources such as the Ministry of Agriculture and Farmers Welfare, Government of India. The datasets include historical records of crop production, agricultural inputs, climatic variables, and geographical information.

**Methodology**

The EDA process involves several steps, including data collection, data cleaning, data pre-processing, and exploratory analysis. Descriptive statistics, data visualization techniques (e.g., bar charts, line plots, heatmaps), and statistical measures will be employed to analyse the data. Correlation analysis, clustering, and spatial analysis techniques may also be applied to uncover hidden patterns and relationships.

**Variables considered.**

Out of seven variables, we have four categorical variables (State\_Name, District\_Name, Season and Crop type) and three continuous variables in float (Area and Production) and integer (Crop\_Year) format.

**Data Preparation and Cleaning**

Data preparation or Data pre-processing is very crucial step in a Data Science project pipeline. This process includes, accessing different variables and verifying the accuracy of data collected. Working on missing data (either deleting it or imputing with appropriate measures). Outliers are also datapoints which must be flagged and investigated if they are conducive for including in the analysis.

This step will clean the available data in such a way that it can be further used in the project pipeline without any hassle (reducing any biasness).

import seaborn as sns

import matplotlib.pyplot as plt

import pandas as pd

import numpy as np

%matplotlib inline

df=pd.read\_csv("./crop-production-in-india/crop\_production.csv")

df.info()

Data columns (total 7 columns):

# Column Non-Null Count Dtype

0 State\_Name 246091 non-null object

1 District\_Name 246091 non-null object

2 Crop\_Year 246091 non-null int64

3 Season 246091 non-null object

4 Crop 246091 non-null object

5 Area 246091 non-null float64

6 Production 242361 non-null float64

Out of seven variables, we have four categorical variables(State\_Name,District\_Name, Season and Crop type) and three continuous variables in float(Area and Production) and integer(Crop\_Year) format.

**Checking for Missing Data**

df.isnull().sum()

State\_Name 0

District\_Name 0

Crop\_Year 0

Season 0

Crop 0

Area 0

Production 3730

Checking for missing values showed Production variable showing huge number of missing values amounting to 3730 values and no missing values for other variables. Next step would be dropping these samples as we have sufficiently big dataset. Missing value accounts for only 1.5% of total sample size.

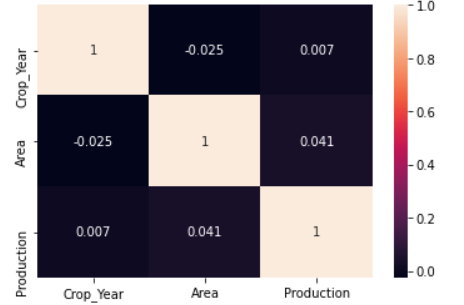
df.dropna(subset=["Production"],axis=0,inplace=True)

We get no missing values now.

**Checking for Correlation between variables**

plt.tick\_params(labelsize=10)

sns.heatmap(df.corr(),annot=True);



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

After Performing UniVariate Analysis , we get to know that

* On District front, we have more data coming from Tumkur, Belgaum, Hassan, Bellary and Bijapur from Karantaka state.
* Our Dataset has data for 19 years from 1997 to 2015. Top years are 2003, 2002, 2007–08 and 2006.
* Dataset shows data for 124 different crop varieties. On doing df.value\_counts(), we can show top crops are Rice(15082),Maize(13787) and Moong(Green Gram, 10106).
* We have proportion of area under cultivation for different states and districts. Production depends on the proportion of area allocated for cultivation.

And after performing BiVariate Analysis, We get to know that Kerala is top state when we look at the quantum of Production for last 19.years.

**Descriptive Statistics**

**Overview of Crop Production in Madhya Pradesh:**

Crop production in India is a critical component of the country's agricultural sector, contributing significantly to its economy and food security. Conducting an exploratory data analysis (EDA) of crop production in India can provide valuable insights into the patterns, trends, and factors influencing agricultural productivity. Here is an overview of crop production in India in reference to EDA:

**Crop Diversity**: India is known for its diverse range of crops due to its varied agro-climatic conditions. An EDA can explore the different types of crops cultivated across various regions of the country. It can identify the dominant crops in specific areas and assess the changes in crop choices over time.

**Yield Analysis**: Analysing crop yields is crucial to understand productivity levels and variations. EDA can help identify the highest yielding crops and regions, as well as assess the factors contributing to yield fluctuations. It can also uncover any temporal patterns or seasonality in crop yields.

**Spatial Distribution**: EDA can examine the spatial distribution of crop production in India. This involves analysing the cultivation patterns across different states, districts, or agro-ecological zones. It can reveal regional disparities, identify areas with high agricultural potential, and explore the impact of geographical factors on crop productivity.

**Crop Rotation and Succession**: EDA can investigate crop rotation and succession patterns in different regions. It can explore the sequence of crops planted and assess the benefits of specific crop rotations, such as pest control, soil fertility maintenance, and risk diversification.

**Input Usage**: Analysing the usage of inputs like fertilizers, pesticides, and irrigation can provide insights into their impact on crop production. EDA can assess the levels and variations in input usage across different crops and regions, helping to identify efficient and sustainable agricultural practices.

By conducting an EDA of crop production in India using relevant variables, researchers can gain valuable insights into the dynamics of agricultural systems, identify areas for improvement, and inform evidence-based policies and interventions to enhance food security and agricultural productivity in the country.

**Diversity of crops grown across the country:**

Rice-Wheat: Madhya Pradesh, West Bengal, UP, Bihar, Punjab, Haryana.

Rice-Rice: Irrigated and Humid coastal system of Tamil Nadu, Orissa, Kerala, Andhra Pradesh and Karnataka.

Rice-Pulses: Bihar, Chhattisgarh and Orissa.

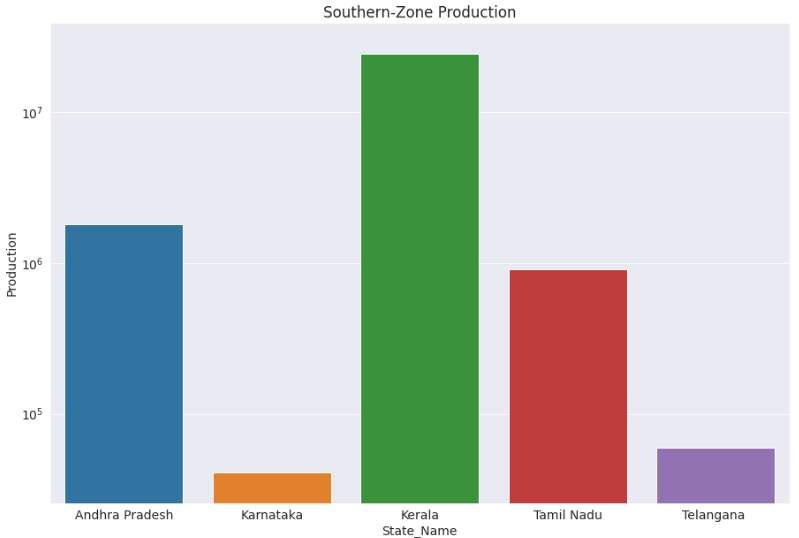
Maize-Wheat: MP, UP, Rajasthan and Bihar.

Sugarcane-Wheat: Punjab, Haryana and UP makes up for 68% of the area under sugarcane. The other states which cover the crops are MP and Karnataka.

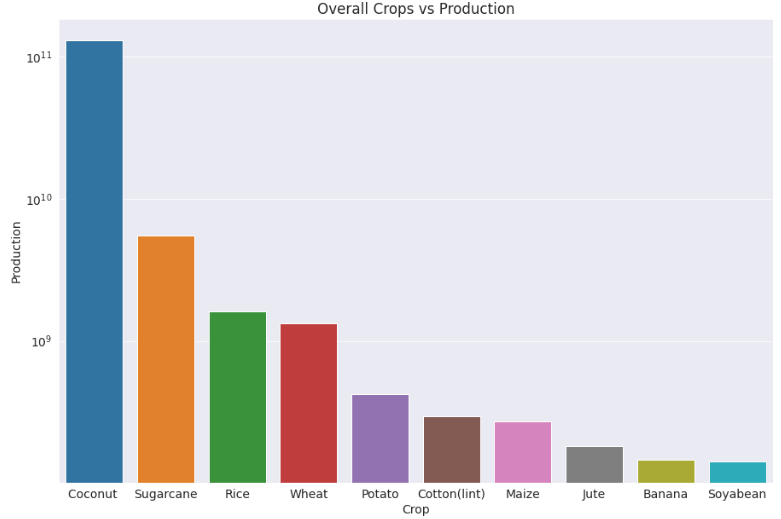
Cotton-Wheat: Haryana, Punjab, Karnataka, Andhra Pradesh, Tamil Nadu.

Soya bean-Wheat: MP, Maharashtra, and Rajasthan.

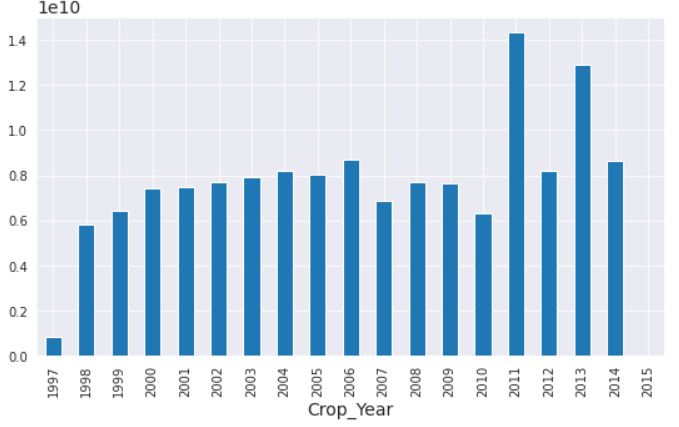
**Data Visualisation**

**Zonal distribution of crops:**

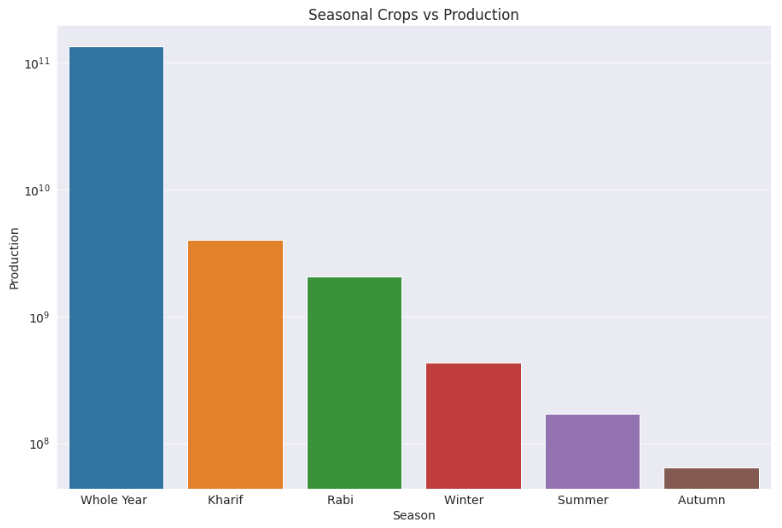
**Crop wise Production status:**



**Year wise Production Status:**

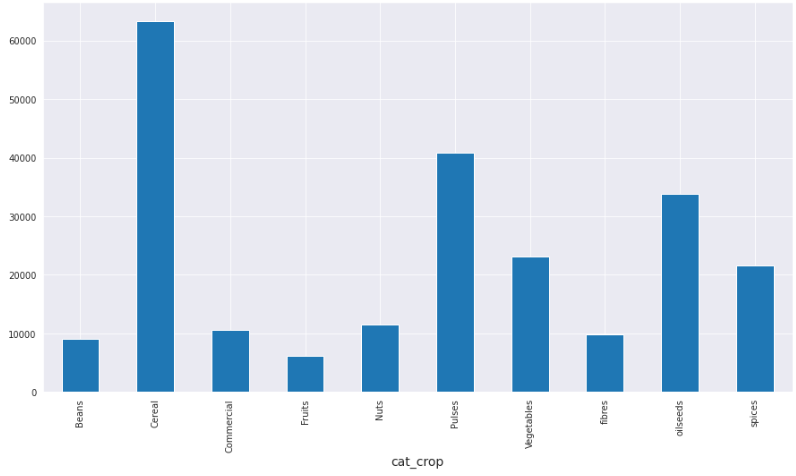


**Season wise Production Status:**



Top crop categories which show high production values are Whole Year (Annual growing plants), Kharif and Rabi crops. It clearly shows these crops heavily dependent on seasonal monsoons.

**Crop wise Production plot describing production values for all crop types.**



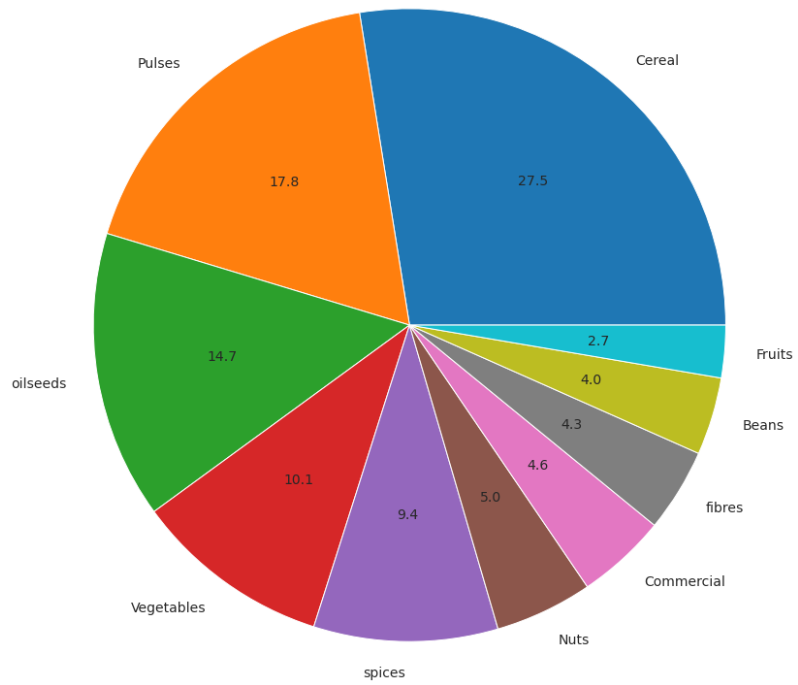
Top crop categories are Cereal, Pulses and Oilseeds.

**Interesting facts:**

South zone: Top producing state Kerala shows an abundance of whole year seasonal crops.

North Zone: Top producing state Uttar Pradesh shows abundance of Kharif, Rabi and Summer crops.

**Different proportion of Crop Categories for India**



**Discussion**

Exploratory data analysis (EDA) of crop production in India can provide valuable insights that have implications for agricultural practices and policy recommendations. Here are some key implications and recommendations that can be derived from EDA:

**Improved Crop Selection**: This EDA reveals the dominant crops in different regions of India. This information can guide farmers and policymakers in selecting suitable crops based on agro-climatic conditions, market demand, and resource availability. Recommendations can be made to promote crop diversification, encouraging the cultivation of crops that are better suited to specific regions.

**Enhanced Yield Management**: This paper identifies regions or periods with high or low crop yields. This information can guide efforts to enhance yield management practices in areas with lower yields. It can also identify regions with high yields as potential models for best practices. Recommendations can be made to promote the adoption of improved agricultural techniques, such as precision farming, better irrigation methods, and optimized fertilizer use.

**Resource Allocation and Infrastructure Development**: Highlighting regions where crop production is concentrated or lacking. This information can guide resource allocation decisions, such as investment in irrigation infrastructure, rural electrification, and transportation networks. Policy recommendations can be made to ensure equitable access to resources, enhance market linkages, and improve post-harvest infrastructure.

**Data-Driven Decision Making**: EDA highlights the importance of data-driven decision making in agriculture. Recommendations can be made for the collection, analysis, and dissemination of accurate and timely data on crop production, yield, input usage, and market trends. This can empower policymakers, researchers, and farmers to make informed decisions and implement evidence-based interventions.

These implications and policy recommendations derived from EDA can contribute to the development of effective agricultural practices and policies in India. By incorporating data-driven insights, policymakers can support sustainable and resilient crop production, improve farmer livelihoods, and enhance food security in the country.

**Some Important Questions asked and explored further.**

**Q1**: **Which State dominates in crop production with different categories of crops?**

Uttar Pradesh is topping in producing more crop categories than any other Indian state and the stats are: Beans (1112), Cereal (9719), Commercial (1741), Fruits (269), Nuts (958), Pulses (6549), Vegetables (3734), Fibres (724), oil seeds (4028) and spices (2529).

**Q2**: **Which Crop is seen in high frequency and when and where is it grown in India?**

* Rice is grown heavily when we look the frequency of crops in India.
* Rice needs Winter for it mature.
* State wise Punjab dominates in rice production.
* Districts wise its BARDHAMAN (2.13%), MEDINIPUR WEST (1.8%) and WEST GODAVARI (1.73%) which contributes to total rice production.
* Year wise 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.

**Q3: Which states ranks high in area wise crop production in India? Substantiate with facts and Figures?**

Top cultivating states based on the Cultivation area are Uttar Pradesh (4.33e+08), Madhya Pradesh (3.29e+08) and Maharashtra (3.22e+08)

Year wise Statues of these States:

* + **Uttar Pradesh**: High Production was seen in 2005 and after that it has been reducing gradually.
  + **Madhya Pradesh**: 1998 showed a high production and then there was gradual reduction, but it picked up and 2012 also showed a peak in Production.
  + **Maharashtra**: Production went down drastically in 2006 and again the levels went up and hit a high peak after 2007.
  + **Rajasthan**: the production hit an all-time low in the year 2002 and then picked up by 2010.
  + **West Bengal**: the production hit a peak around 2006 but it has hit a low after 2007 and never recovered back.

**Q4: Discuss the top crops grown in Northern parts of India?**

* Production wise top states of North zone are Punjab(5.86e+08), Uttar Pradesh (3.23e+09), and Haryana (3.81e+08)
* Top crops of these states are Sugarcane, Wheat, and Rice.

**Q5**: **Top crop of South India is Coconut, Find out that status of Coconut production in India?**

* Coconut cultivation is yearlong and does not get restricted to any seasons.
* Top states involved in coconut production are Kerala, Andhra Pradesh, and Tamil Nadu
* Top districts featuring in coconut production is KOZHIKODE (11.75%), MALAPPURAM (11.16%) and THIRUVANANTHAPURAM (7.7%)
* Year wise coconut cultivation is strong and its increasing healthy.
* High coconut cultivation is directly proportional to area under cultivation.

**Inferences and Conclusion**

We started with 246091 samples with 7 columns. Production Variable had 3730 (1.5% of total sample size) missing values which was dropped and working dataset has 242361 sample size. Also checked for multicollinearity of variables using heatmap.

**Uni-Variate Analysis:**

* **State\_Name**: 33 Names including Union territories. Top states contributing to dataset are Uttar Pradesh, Madhya Pradesh, and Karnataka.
* **District\_Name**: 646 districts are represented in the dataset and Top contributors are Tumkur, Belgaum, and Bijapur from Karnataka
* **Crop\_Year**: Dataset represents data for 19 years from 1997 to 2015 and maximum data from 2003, 2002 & 2007.
* **Season**: we see six seasons with maximum data from Kharif, Rabi and Whole year
* **Crop**: we data for 124 different crops with top occurrence from Rice, Maize and Moong (Green Gram)
* **Area**: Huge margin area used for production from 1 to 8580100-unit area. Distribution is highly right skewed due to lot of outliers.
* **Production**: Production value ranges from 0 to 1.25e+09 and Distribution is highly right skewed due to lot of outliers.

**Bi-Variate Analysis:**

State\_Name Vs Production: Top states are Kerala, Andhra Pradesh, and Tamil Nadu

**New Variables created:**

* **Zones**: A new variable was created based on 33 State\_Name. States were divided bucketed into North Zone, South Zone, East Zone, West Zone, Central Zone, NE Zone, and Union Terr. Dataset has more data coming from South, North and East zones.
* **Cat\_crop**: 124 crops were divided into Cereal, Pulses, oilseeds, Vegetables, spices, Nuts, Commercial, fibres, Beans, Fruits. Dataset shows top categories are Cereal, Pulses, and oilseeds.

**Visualization done on:**

* **Zonal distribution of crops**: Production wise top zone is South zone with Kerala topping its high crop production.
* **Crop wise Production status**: Production wise top crop type are Coconut, Sugarcane and Rice.
* **Year wise Production Status**: Total Production hit peak in the year 2011 and 2013.
* **Season wise Production Status**: Top crop categories which shows high production values are Whole Year (Annual growing plants), Kharif and Rabi crops. It clearly shows these crops heavily dependent on seasonal monsoons.
* **Crop Category wise Production Status**: Top crop categories are Cereal, Pulses and Oilseeds
* State versus crop Category versus Season Status: Top producing state Kerala shows a abundance of whole year seasonal crops and Top producing state Uttar Pradesh shows abundance of Kharif, Rabi and Summer crops.
* **Different Proportion of crop Categories Status**: Top Crop Categories are Cereal (27.5), Pulses (17.8) and oil seeds (14.7) and these crops contribute towards 60% of total crop production.

**Future Work**

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 over the years, if the production has gone up (Good case scenario) and if production is gone down can we investigate the causation of this trend.



**References**

<https://www.currentscience.ac.in/Volumes/123/03/0488.pdf>

<https://agricoop.nic.in/Documents/annual-report-2020-21.pdf>

<https://agricoop.gov.in/Documents/agristatglance2018.pdf>

<https://www.indiastat.com/data/agriculture/commercial-crops>

<https://www.insightsonindia.com/agriculture/major-crops-and-cropping-patterns-in-various-parts-of-the-country/major-crops-grown-in-india/#:~:text=Rice%2DRice%3A%20Irrigated%20and%20Humid,%2C%20Rajasthan%2C%20MP%20and%20Bihar>.