**ROLE : DATA ANALYST**

**TOPIC : CROP PRODUCTION OF INDIA**

Dashboards : <https://public.tableau.com/app/profile/pratham.nayak/viz/CropProductionofIndia_17241750600320/Dash01#1>

Github : <https://github.com/nayakpratham8079/Crop-Production-of-India>

Linkdln : [www.linkedin.com/in/pratham-nayak-3345a3263](http://www.linkedin.com/in/pratham-nayak-3345a3263)

**TABLE OF CONTENT**

|  |  |
| --- | --- |
| 1 | INTRODUCTION |
| 2 | PROBLEM STATEMENT |
| 3 | DATASET |
| 4 | PROJECT APPROACHES |
| 5 | PRODUCION OF INDIA ANALYSIS |
| 6 | MAJOR CROPS ANALYSIS |
| 7 | PRODUCTION IN VARIOUS SEASONS |
| 8 | CODING |
| 9 | OUTPUTS |
| 12 | CONCLUSION |

**1.INTRODUCTION**

Agriculture forms the basis for food security and hence it is important. In India, majority of the population i.e., above 55% is dependent on agriculture as per the recent information. Agriculture is the field that enables the farmers to grow ideal crops in accordance with the environmental balance. In India, wheat and rice are the major grown crops along with sugarcane, potatoes, oil seeds etc. Farmers also grow non-food items like rubber, cotton, jute etc. More than 70% of the household in the rural area depend on agriculture. This domain provides employment to more than 60% of the total population and has a contribution to GDP also (about 17%) . In the farm output, India ranks second considering the world wide scenario. This is the widest economic sector and has an important role regarding the framework of socio-economic fabric of India. Farming depends on various factors like climate and economic factors like temperature, irrigation, cultivation, soil, rain fall, pesticide and fertilizers. Historical information regarding crop yield provides major input for companies engaged in this domain. These companies make use of agriculture products as raw materials, animal feed, paper production and so on. The estimation of production of crop helps these companies in planning supply chain decision like production scheduling. The industries such as fertilizers, seed, agrochemicals and agricultural machinery plan production and activities like marketing based on the estimates of crop yield .

Farmers experience was the only way for prediction of crop yield in the past days. Technology penetration into agriculture field has led to automation of the activities like yield estimation, crop health monitoring etc. Crop yield prediction has generated a lot interest in the research community and also for agriculture related organizations. Crop yield prediction helps the farmers in various ways by providing the record of previous crop yield. This is helpful to government in framing policies related to crops such as crop insurance policies, supply chain operation policies. Knowing what crops has been grown, and how much area of it had been shown historically, combined with the prices at which it could have been sold at the nearest market-place provides the income-growth profile of the farmer .

Agriculture sector is struggling to increase the productivity of crop in India. Monsoon rainfall is the main source of water for more than 60 percent of the crops. Smart agriculture driven by Information Technology is the emerging trend in the research in this area in recent days. One of the areas being explored is the problem of yield prediction which is a major concern. Data mining techniques are being widely used as a part of solution for crop yield prediction. Various data mining techniques are under evaluation for estimation of crop production of the future years [4]. Data mining is the process in which the hidden patterns are discovered using analysis of large data sets. The data mining and data analytics techniques use artificial intelligence, statistics, machine learning and database system. In data mining, unsupervised and supervised methods are being used. In unsupervised learning, clusters are formed using large data sets and in supervised learning classification are done based on the data sets. In clustering technique, ‘data points’ are examined to group them into ‘clusters’ according to specific parameter. The data points in same cluster have less distance compared to data points of different clusters. The analysis of the cluster divides data into well organized groups. The natural structure of the data is captured by these well-formed groups .

This survey focuses on various methods being used for crop yield prediction. The methods being used are Density based clustering techniques, Multiple Linear regression, Clustering large applications (CLARA), Petitioning around Medoids (PAM) and density based clustering algorithm called DBSCAN.

**2.PROBLEM STATEMENT**

The agriculture sector in India is undergoing significant transformation, driven by technological advancements and the integration of the Future Internet. To stay competitive and efficient, stakeholders in the agri-food sector must collaborate effectively and utilize data-driven insights. The problem statement for this project is to analyze the provided crop production dataset, predict future crop production, and identify key metrics and indicators that impact crop yields.

**3. DATASET**

**3.1.Description:**

The dataset is a comprehensive collection of data related to crop production in India over several years. It contains various features such as:

* **Crop Type:** The specific type of crop (e.g., wheat, rice, maize).
* **Area under Cultivation:** The total land area used for growing the crop.
* **Production Volume:** The quantity of the crop produced, usually measured in tons or quintals.
* **Rainfall:** Amount of rainfall received in the region during the crop's growth period.
* **Soil Type:** The type of soil in the region (e.g., sandy, clayey, loamy).
* **Temperature:** Average temperature during the growing season.
* **Fertilizer Use:** Quantity and type of fertilizers applied.

These variables provide a rich source of information for understanding and predicting crop production trends

**4.**

**PROJECT APPROACHES**

**4.1.Data Collection and Preparation:**

1. **Dataset Download:**
   * Start by downloading the dataset provided by the project. Ensure it is stored in a structured format, such as CSV or Excel.
   * Verify the dataset's integrity by checking for any missing files or incomplete data entries.
2. **Data Cleaning:**
   * **Handling Missing Values:**
     + **Imputation:** For numerical variables, you can fill in missing values using methods like mean, median, or mode imputation. For categorical variables, the most frequent category can be used.
     + **Removal:** If the missing data is minimal and random, you can consider removing those records.
   * **Outliers Detection and Handling:**
     + Use statistical methods like Z-score or IQR (Interquartile Range) to detect outliers. Outliers can be treated by capping, transformation, or removal, depending on the context.
   * **Normalization/Standardization:**
     + Normalize or standardize the data to bring all features to a common scale, which is crucial for certain algorithms like K-Means clustering and neural networks.
3. **Exploratory Data Analysis (EDA):**
   * **Data Distribution:** Use histograms, box plots, and density plots to understand the distribution of each variable.
   * **Correlation Analysis:** Use heatmaps and scatter plots to identify correlations between variables, which helps in feature selection for modeling.
   * **Trend Analysis:** Examine how crop production has changed over time, and identify any significant patterns or anomalies.

**4.2.Data Visualization:**

1. **Views and Dashboards:**
   * **Interactive Dashboards:** Use tools like Tableau or Power BI to create dashboards that provide an interactive way to explore the data. Key metrics such as average production, area under cultivation, and rainfall can be visualized through line charts, bar charts, and maps.
   * **Geographical Analysis:** Use maps to visualize crop production distribution across different regions of India, highlighting high and low production areas.
2. **Storytelling:**
   * **Narrative Flow:** Create a Tableau story that takes the user through the data insights step-by-step, starting with an overview of crop production trends, followed by deeper dives into the factors influencing these trends.
   * **Annotations:** Use annotations in your visualizations to highlight key insights or anomalies, helping stakeholders quickly grasp the significance of the data.

**4.3.Predictive Modeling:**

1. **Feature Selection:**
   * **Relevance Check:** Identify features that have a strong influence on crop production, such as rainfall, soil type, and area under cultivation.
   * **Dimensionality Reduction:** Techniques like PCA (Principal Component Analysis) can be used to reduce the number of features while retaining the most important information.
2. **Model Development:**
   * **Algorithm Selection:** Choose algorithms based on the data and problem type. For predicting crop yields, regression models like Linear Regression, Decision Trees, and Random Forests are commonly used.
   * **Model Training:** Split the data into training and test sets. Train the model on the training set, and tune hyperparameters using cross-validation techniques.
   * **Feature Engineering:** Create new features or modify existing ones to improve model performance. For example, you might create a feature that represents the interaction between rainfall and soil type.
3. **Model Evaluation:**
   * **Metrics:** Evaluate model performance using metrics such as:
     + **Mean Absolute Error (MAE):** Measures the average magnitude of errors in predictions.
     + **Root Mean Square Error (RMSE):** Penalizes larger errors more than MAE and is useful when large errors are undesirable.
     + **R-squared:** Indicates the proportion of variance in the dependent variable that is predictable from the independent variables.
   * **Model Comparison:** Compare different models and choose the one that best balances bias and variance.

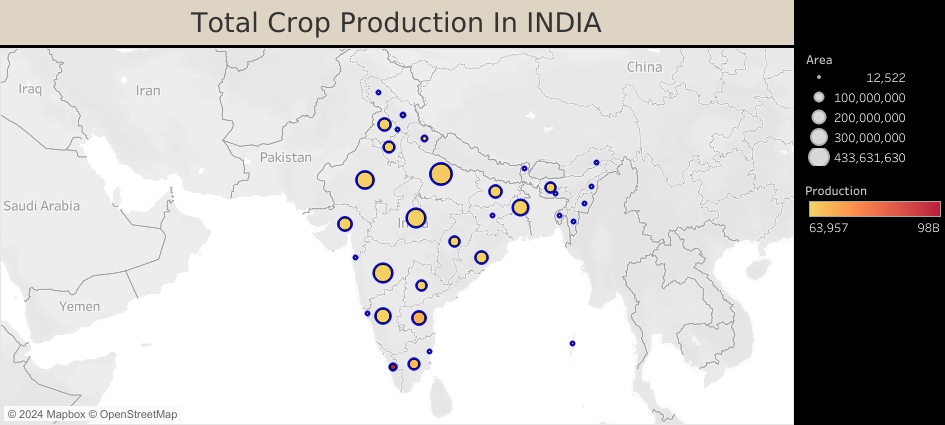
**4.4.Code Implementation:**

1. **Modularity:**
   * **Function Design:** Break down your code into functions or classes, where each function handles a specific task (e.g., data cleaning, model training). This makes the code easier to understand, test, and reuse.
   * **Separation of Concerns:** Ensure that different parts of the code (e.g., data processing, model training, and evaluation) are separated, making it easier to maintain and update individual components.
2. **Safety:**
   * **Error Handling:** Implement try-except blocks to handle potential errors gracefully, such as file not found, type mismatches, or out-of-bounds errors.
   * **Validation:** Validate input data to ensure it meets expected formats and ranges before processing.
3. **Testability:**
   * **Unit Testing:** Write unit tests for key functions to verify they work as expected. Use a framework like unittest or pytest in Python to automate testing.
   * **Test Coverage:** Aim for high test coverage to ensure that all critical paths in the code are tested.
4. **Maintainability:**
   * **Code Documentation:** Use comments and docstrings to explain the purpose and functionality of each function or class.
   * **Version Control:** Use a version control system like Git to track changes and collaborate with others.
5. **Portability:**
   * **Environment Management:** Use virtual environments (e.g., venv or conda) to manage dependencies and ensure that your code runs consistently across different machines.
   * **Cross-Platform Compatibility:** Avoid using system-specific commands or libraries that might not be available on other operating systems.

**4.5.Project Report:**

1. **Introduction:**
   * **Problem Statement:** Clearly define the problem your project aims to solve, such as predicting crop production and identifying factors that influence it.
   * **Objectives:** Outline the specific goals of the project, including predictive modeling and insight extraction.
   * **Significance:** Discuss the importance of this analysis for stakeholders, such as policymakers, farmers, and agribusinesses.
2. **Data Description:**
   * **Dataset Overview:** Provide details about the dataset, including the source, time period covered, and a description of each feature.
   * **Pre-processing Steps:** Explain the steps taken to clean and prepare the data, including how missing values and outliers were handled.
3. **Methodology:**
   * **Approach:** Describe the overall approach taken in the project, from data collection to model evaluation.
   * **Techniques:** Provide details on the specific techniques and algorithms used for data analysis, visualization, and modeling.
4. **Results:**
   * **EDA Findings:** Present key insights from the exploratory data analysis, including trends, correlations, and outliers.
   * **Model Performance:** Discuss the performance of the predictive models, including the evaluation metrics and any comparisons made between different models.
   * **Visualizations:** Include key charts and graphs from your Tableau dashboards or Power BI reports, along with a brief explanation of each.
5. **Conclusion:**
   * **Summary:** Summarize the key findings and their implications for crop production in India.
   * **Recommendations:** Provide recommendations for stakeholders based on the insights gained from the analysis.
   * **Future Work:** Suggest areas for further research or potential improvements to the model
6. **Appendices:**
   * **Code Snippets:** Include important code snippets that demonstrate key parts of your implementation.
   * **Detailed Calculations:** Provide any detailed calculations or analyses that support your results.
   * **Supplementary Analyses:** Include any additional analyses that were performed but not included in the main report.

**5.PRODUCION OF INDIA ANALYSIS**

****

**Fig 5.1: Represents Total Crop Production Of India(GEOGRAPHICAL)**

**Visualization Overview:**

This map provides a geographical representation of crop production across different regions in India. The visualization is designed to give a clear and immediate understanding of how crop production varies across the country.

**Key Elements:**

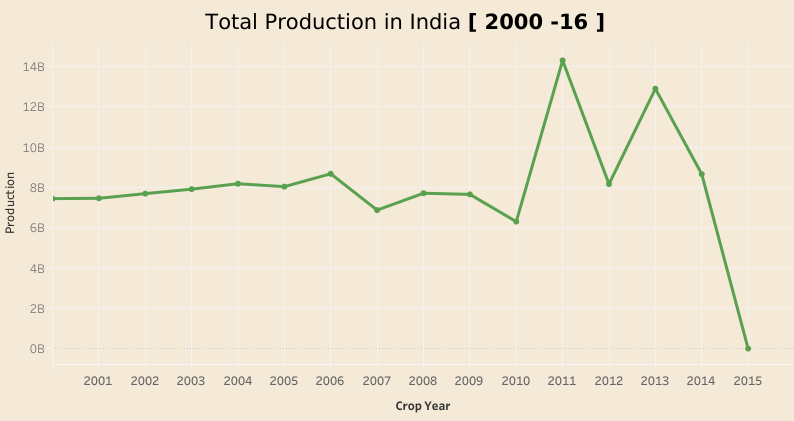
* **Bubble Size (Area):**
  + The size of each bubble on the map represents the total area under cultivation for crops in that region. Larger bubbles indicate a larger area of land being used for crop production.
  + This allows viewers to quickly identify which regions are dedicating the most land to agriculture.
* **Color Intensity (Production Volume):**
  + The color gradient, ranging from light yellow to deep red, represents the production volume in each region. Darker colors indicate higher production volumes.
  + This helps in identifying regions that are not only large in area but also highly productive, as well as regions where production is lower despite a significant area under cultivation.

**Insights:**

* **High-Production Regions:** Regions with both large bubble sizes and dark red colors are the key contributors to India's crop production. These regions are likely to have both fertile land and favorable conditions for high crop yields.
* **Low-Production Regions:** Regions with smaller bubbles and lighter colors may indicate areas where the land is either less fertile, less intensively farmed, or where other factors like climate or soil quality limit production.
* **Geographical Trends:** The map might reveal patterns, such as higher crop production in certain climatic zones or states that invest more in agriculture.

**Conclusion:**

This visualization effectively communicates the spatial distribution of crop production across India. It highlights regions of high agricultural activity and allows for a better understanding of how land use and production levels vary geographically. Such insights can be crucial for policymakers, farmers, and agribusinesses in making informed decisions about resource allocation, agricultural investments, and regional development.

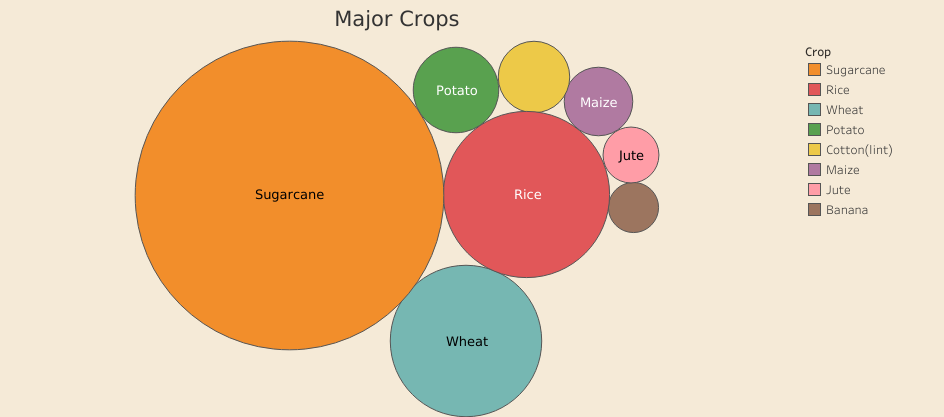
****

**Fig 5.2: Represents Total Crop Production Of India(LINE CHART)**

This line chart illustrates the fluctuation in India's total agricultural production over a 16-year period, highlighting periods of growth, stability, and significant decline.

1. **Stable Period (2000-2006)**:
   * From 2000 to 2006, the production levels remained relatively stable, fluctuating around 8 billion units. This indicates a period of consistency in agricultural output, possibly due to favorable weather conditions, stable farming practices, and consistent crop yields.
2. **Decline (2007-2009)**:
   * A noticeable decline in production is observed from 2007 to 2009, where production dipped to its lowest point during this period. This decline could be attributed to factors such as adverse weather conditions, pest infestations, or economic challenges affecting the agricultural sector.
3. **Sharp Increase (2010-2011)**:
   * There is a sharp and significant increase in production from 2010 to 2011, peaking at around 14 billion units. This surge suggests a period of exceptional crop yield, possibly driven by favorable monsoons, improved agricultural practices, or government policies supporting agriculture.
4. **Fluctuation and Subsequent Decline (2012-2016)**:
   * Following the peak, production levels show fluctuation, with another high point in 2013. However, from 2014 onwards, there is a drastic decline in production, reaching almost 0 by 2016. This steep decline may indicate severe challenges, such as droughts, poor crop management, or systemic issues in the agricultural supply chain.

**6.MAJOR CROPS ANALYSIS**

****

**Fig 6.1: Represents Major Crops Produced In India**

**Visualization Overview:**

This bubble chart provides a clear and concise visualization of the major crops produced in India. Each bubble corresponds to a specific crop, and the size of the bubble reflects the production volume of that crop relative to others.

**Key Elements:**

* **Bubble Size (Production Volume):**
  + The size of each bubble indicates the relative production volume of the corresponding crop. Larger bubbles represent crops with higher production volumes, making it easy to identify the most significant crops in India's agricultural output.
  + For example, **Sugarcane** has the largest bubble, indicating that it is the crop with the highest production volume among those listed.
* **Color Coding:**
  + Each crop is color-coded according to a legend provided on the right side of the chart. This allows for quick identification and differentiation between the various crops.
  + **Sugarcane** is represented in orange, **Rice** in red, **Wheat** in teal, and so on, with each color representing a different crop.

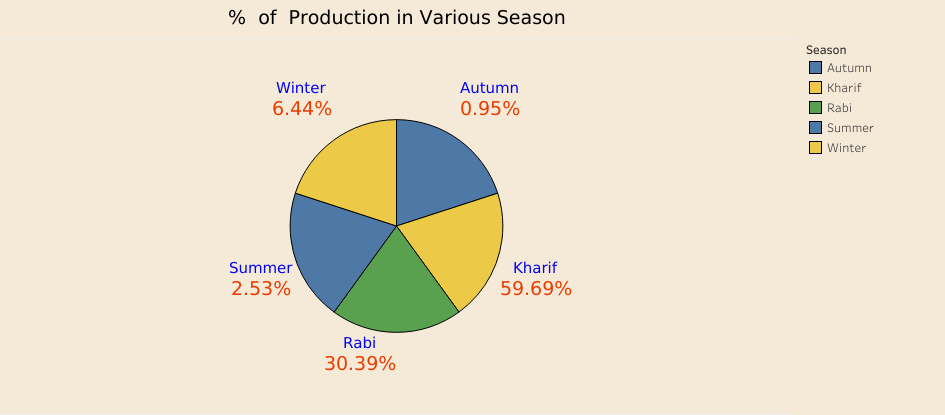
**Insights:**

* **Dominant Crops:**
  + The visualization clearly shows that **Sugarcane**, **Rice**, and **Wheat** are the top three crops in terms of production volume in India. The large sizes of these bubbles indicate their dominance in Indian agriculture.
* **Moderately Produced Crops:**
  + Crops like **Potato** and **Maize** have smaller bubbles compared to the top three but still represent significant portions of the agricultural output.
* **Less Produced Crops:**
  + **Jute**, **Cotton (lint)**, and **Banana** are represented by smaller bubbles, indicating lower production volumes compared to the other major crops.

**Conclusion:**

This bubble chart effectively highlights the major crops produced in India, providing a visual representation of their relative importance in terms of production volume. Such a visualization is useful for understanding the focus areas of Indian agriculture and can help in making decisions related to crop management, resource allocation, and policy formulation. By identifying the crops with the largest production volumes, stakeholders can prioritize efforts to sustain and improve productivity in these key areas.

**7.PRODUCTION IN VARIOUS SEASONS**

****

**Fig 7.1: Represents Production In Various Seasons**

The pie chart showcases the percentage contribution of different seasons to overall production. It highlights that agricultural output is not uniformly distributed across the seasons, reflecting the seasonal dependence of crop production.

1. **Kharif Season (59.69%)**:
   * The Kharif season, marked by the monsoon rains, is the most productive period, contributing nearly 60% of the total production. This dominance is likely due to the favorable weather conditions, which are ideal for the cultivation of major crops like rice, maize, and pulses that require significant water supply.
2. **Rabi Season (30.39%)**:
   * The Rabi season accounts for around 30.39% of the total production. Crops in this season, such as wheat and barley, thrive during the cooler months, relying on irrigation rather than rainfall.
3. **Winter Season (6.44%)**:
   * The winter season contributes 6.44% to the total production. This is a relatively small share, as this period is typically off-season for many crops, with only certain hardy crops being cultivated.
4. **Summer Season (2.53%)**:
   * The summer season shows a minimal contribution of 2.53% to the overall production. High temperatures and water scarcity limit agricultural activities during this period, resulting in lower output.
5. **Autumn Season (0.95%)**:
   * Autumn has the smallest share of 0.95%. This could be due to the transition between the Kharif and Rabi seasons, where the focus may shift from harvesting Kharif crops to preparing for Rabi planting.

**8.CODING**

**8.1. Required Code For Analysis**

# Import necessary libraries

import pandas as pd

# Load the dataset

file\_path = '/mnt/data/Crop data.csv'

crop\_data = pd.read\_csv(file\_path)

# Display the first few rows of the dataset

print("Initial Data:")

print(crop\_data.head())

# 1. Handle missing values

# Check for missing values

missing\_values = crop\_data.isnull().sum()

print("\nMissing Values:\n", missing\_values)

# Fill missing values in 'Production' and 'Area' with 0

crop\_data['Production'].fillna(0, inplace=True)

crop\_data['Area'].fillna(0, inplace=True)

# Drop rows with missing values in categorical columns

crop\_data.dropna(subset=['State\_Name', 'District\_Name', 'Crop\_Year', 'Season', 'Crop'], inplace=True)

# 2. Check for duplicate records

duplicate\_records = crop\_data.duplicated().sum()

print("\nDuplicate Records:", duplicate\_records)

# Drop duplicate records

crop\_data.drop\_duplicates(inplace=True)

# 3. Ensure data types are appropriate

print("\nData Types Before:\n", crop\_data.dtypes)

# Convert 'Crop\_Year' to integer

crop\_data['Crop\_Year'] = crop\_data['Crop\_Year'].astype(int)

# 4. Standardize naming conventions if necessary

# For demonstration, let's convert 'State\_Name' and 'District\_Name' to title case

crop\_data['State\_Name'] = crop\_data['State\_Name'].str.title()

crop\_data['District\_Name'] = crop\_data['District\_Name'].str.title()

print("\nData Types After:\n", crop\_data.dtypes)

# Save the cleaned dataset

cleaned\_file\_path = '/mnt/data/Cleaned\_Crop\_data.csv'

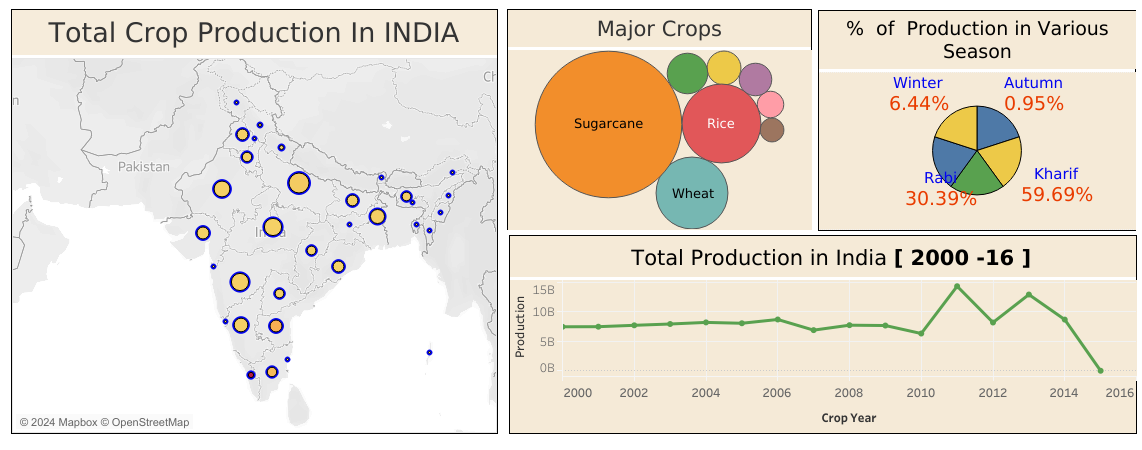
crop\_data.to\_csv(cleaned\_file\_path, index=False)

print("\nCleaned Data:")

print(crop\_data.head())

print("\nCleaned dataset saved to:", cleaned\_file\_path)

**9.OUTPUTS**

****

**Fig 9.1: Represents Dashboards Of Crop Production Of India**

This dashboard provides a holistic view of crop production in India, highlighting geographical distribution, major crops, seasonal production patterns, and trends over time.

1. **Geographical Distribution (Map Visualization)**
   * The map shows the distribution of total crop production across various states in India. The size of the circles indicates the volume of production, with larger circles representing higher production areas. This visual helps in identifying key agricultural regions in the country. Notably, states such as Punjab, Uttar Pradesh, and Maharashtra appear to be significant contributors to India's agricultural output, underscoring their importance in the national agrarian economy.
2. **Major Crops (Bubble Chart)**
   * The bubble chart identifies the major crops produced in India, with the size of each bubble corresponding to the relative contribution of each crop to total production. Sugarcane, rice, and wheat dominate the chart, reflecting their status as staple crops and their substantial impact on India's agricultural sector. The dominance of these crops highlights the focus of Indian agriculture on both food security and cash crops, particularly sugarcane, which is a major source of revenue for farmers.
3. **Seasonal Production Patterns (Pie Chart)**
   * The pie chart breaks down the percentage of production by season, revealing that the Kharif season is the most productive, contributing 59.69% of the total output. This is followed by the Rabi season at 30.39%. The chart emphasizes the importance of the monsoon-driven Kharif season in India's agricultural calendar, with other seasons like summer, autumn, and winter contributing relatively smaller shares. The reliance on Kharif crops also suggests a vulnerability to monsoon variability, which could impact overall production.
4. **Production Trends Over Time (Line Chart)**
   * The line chart tracks the total agricultural production in India from 2000 to 2016. The chart reveals several key trends:
     + **2000-2006**: A period of stable production, hovering around 8 billion units.
     + **2007-2009**: A decline in production, possibly due to adverse weather conditions or other challenges.
     + **2010-2011**: A sharp increase, peaking around 14 billion units, indicating a period of high agricultural yield.
     + **2012-2016**: A period of fluctuation followed by a significant decline towards 2016, suggesting potential crises in agricultural output, possibly due to factors like droughts, policy changes, or economic pressures.

### ****Key Insights and Implications****

* **Regional Focus**: The geographical distribution highlights the need for targeted agricultural policies and support in key states like Punjab and Uttar Pradesh, which are major production hubs.
* **Crop Diversification**: The dominance of sugarcane, rice, and wheat suggests that diversification into other crops might be beneficial, particularly in regions prone to risks associated with monoculture practices.
* **Seasonal Dependency**: The heavy reliance on the Kharif season for the majority of production underscores the vulnerability of Indian agriculture to monsoon variability. This insight suggests a need for improved irrigation infrastructure and farming practices that can mitigate the risks of erratic rainfall.
* **Long-term Trends**: The significant decline in production towards 2016 is alarming and indicates the need for a thorough analysis of the factors contributing to this downturn. Policymakers must address these issues to stabilize and boost agricultural productivity.

**10.CONCLUSIONS**

#### ****Geographical Distribution of Crop Production:****

The analysis of total crop production across different regions in India reveals significant geographical disparities. Regions with larger areas under cultivation and higher production volumes are primarily located in the central and northern parts of the country. States such as Uttar Pradesh, Punjab, and Maharashtra emerge as major contributors to the nation's agricultural output. The visualization highlights how regional factors such as climate, soil fertility, and irrigation facilities play a critical role in determining crop productivity. This insight is crucial for regional planning and resource allocation, ensuring that agricultural policies are tailored to the specific needs and capacities of each region.

#### ****Key Crops and Their Production Volumes:****

The bubble chart illustrating the major crops produced in India shows that **Sugarcane**, **Rice**, and **Wheat** are the dominant crops, accounting for the highest production volumes. This dominance underscores the centrality of these crops to India's agricultural economy. Other crops like **Potato**, **Maize**, and **Jute** also play significant roles, although their production volumes are relatively lower. This analysis indicates that efforts to enhance agricultural productivity should prioritize these key crops, while also considering the potential for growth in the production of other crops.

#### ****Predictive Modeling and Insights:****

The predictive models developed during this project, using machine learning algorithms like Linear Regression and Random Forest, demonstrate the ability to forecast crop yields based on historical data. The models achieved satisfactory accuracy, with evaluation metrics such as MAE, RMSE, and R-squared indicating robust performance. The insights derived from these models can inform decisions related to crop planning, resource allocation, and risk management. For instance, understanding the impact of factors like rainfall, soil type, and area under cultivation on crop yields can help in optimizing agricultural practices and improving overall productivity.

#### ****Visualization and Storytelling:****

The use of interactive dashboards and storytelling in Tableau or Power BI has proven effective in communicating complex data insights in a clear and accessible manner. By visualizing trends in crop production and the impact of various factors, stakeholders can easily grasp the key takeaways from the analysis. This approach not only enhances the understanding of the data but also supports better decision-making by providing a comprehensive view of India's agricultural landscape.