INT 353

Name :- Aswin S Krishna

Reg No:- 12114780

Roll No:- RK21UWB67

ACKNOWLEDGEMENT

I am deeply grateful to our esteemed instructor, Ms. Shivangi Gupta, whose invaluable guidance, unwavering support, and insightful suggestions have been a constant source of inspiration throughout this project. Her expertise and dedication have been essential to our success.

I also extend my sincere gratitude to all the faculty members of Lovely Professional University for their academic guidance and for providing us with the necessary resources and facilities to complete this project. Their commitment to fostering a conducive learning environment has been instrumental in our progress.

My heartfelt thanks also go out to my friends and family, who have been a beacon of encouragement and motivation. Their faith in my abilities and their constant support have been the pillars upon which this project stands.

I would also like to thank all the individuals who have contributed to the successful completion of this project. Their assistance and cooperation have been invaluable.

Finally, I express my appreciation to UpGrad Campus for their unwavering support throughout the project. Their commitment to providing a platform for learning and growth has been a significant factor in our success

Aswin S Krishna

12114780

TABLE OF CONTENTS

Agricultural Crop Yield in Indian States Dataset

Introduction

Agriculture stands as the cornerstone of India's economy, employing millions and serving as a vital source of sustenance and livelihood for a significant portion of its population. The success and resilience of this sector are intricately tied to a multitude of factors, ranging from regional climate conditions to the adoption of modern agricultural practices. Understanding the dynamics of crop yield across different states in India is not only essential for ensuring food security but also for driving economic growth and sustainability.

The journey begins with a thorough exploratory data analysis (EDA), peeling back the layers of this dataset to uncover patterns, correlations, and variations. Through the lens of data, we will scrutinize the choices made by farmers, the influence of seasonal changes, and the role of technology in shaping the outcomes of Indian agriculture.

The results of this analysis hold the potential to guide policymakers, empower farmers with knowledge, and shape strategies that foster climate-resilient, sustainable, and economically vibrant agriculture in India. As we embark on this journey through the fields of data, let us explore the vibrant mosaic of Indian agriculture and draw insights that can contribute to the growth and prosperity of this critical sector.

Domain/Topic knowledge

Importance of Crop Yield Data in India:

- India is one of the world's largest agricultural producers, and agriculture plays a crucial role in the country's economy and food security.
- Crop yield data is essential for policymakers, farmers, researchers, and various stakeholders to make informed decisions about agricultural planning, resource allocation, and food distribution.
- India cultivates a wide variety of crops due to its diverse agro-climatic regions. Major crops include rice, wheat, maize, pulses, oilseeds, sugarcane, cotton, and more.
- The crop mix varies by state, depending on the local climate and soil conditions
- Crop yields are influenced by several factors, including weather conditions, water availability, soil quality, pest and disease management, farming practices, and the adoption of modern agricultural technologies.
- Weather events like monsoons, droughts, floods, and extreme temperatures can significantly impact crop yields.
- Crop yields can vary significantly from year to year and from one region to another.

Economic Impact:

- Crop yields have a direct impact on the income of farmers, food prices, and the overall economic health of the country.

Data Challenges:

- There can be challenges with data quality and consistency, especially when comparing data across different states and over time.
- Ensuring data accuracy and timeliness is important for meaningful analysis and decision-making.

Data Understanding

This comprehensive dataset is valuable for agricultural analysts, researchers, and data scientists interested in crop yield prediction and agricultural analysis. It offers insights into the relationship between various agronomic factors (e.g., rainfall, fertilizer, pesticide usage) and crop productivity across different states and crop types. Researchers can utilize this data to develop robust machine learning models for crop yield prediction and identify trends in agricultural production.

COLUMN DESCRIBTION:

1. **Crop**: The name of the crop cultivated.
2. **Crop Year**: The year in which the crop was grown.
3. **Season**: The specific cropping season (e.g., Kharif, Rabi, Whole Year).
4. **State**: The Indian state where the crop was cultivated.
5. **Area**: The total land area (in hectares) under cultivation for the specific crop.
6. **Production**: The quantity of crop production (in metric tons).
7. **Annual Rainfall**: The annual rainfall received in the crop-growing region (in mm).
8. **Fertilizer**: The total amount of fertilizer used for the crop (in kilograms).
9. **Pesticide**: The total amount of pesticide used for the crop (in kilograms).
10. **Yield**: The calculated crop yield (production per unit area).

Reasons For choosing this dataset

- **Agricultural Research:** Researchers and scientists often use agricultural crop yield data to study trends, patterns, and the impact of various factors on crop production. This data can be valuable for conducting agricultural research and understanding the dynamics of crop yields in different regions.
- **Farmers' Decision-Making:** Farmers can benefit from crop yield data by making informed decisions about crop selection, planting times, and agricultural practices. Understanding historical crop yields can guide them in maximizing their productivity and income.
- **Climate Change Studies:** With the increasing impact of climate change on agriculture, crop yield data can be used to study how changing weather patterns and environmental conditions affect crop production over time.
- **Economic Analysis:** Economists and analysts may use crop yield data to assess the economic impact of agriculture on a nation's GDP and employment. It can also help in forecasting agricultural commodity prices.
- **Education and Awareness:** Agricultural crop yield data can be used in educational programs and awareness campaigns to help people understand the importance of agriculture, its challenges, and the need for sustainable practices.

Questions

1.What is the most common crop in the dataset?

2.Which year had the highest crop production?

3. What is the distribution of different seasons in the dataset?

4. Does the choice of crop season impact crop production

5. What is the range of annual rainfall in the dataset?

6. How many unique fertilizer types are present in the dataset?

7. What is the most commonly used pesticide

8. What is the distribution of crop yields?

9. Is there a correlation between annual rainfall and crop production?

10. Does the choice of crop season impact crop production?

11. Is there a relationship between crop area and pesticide use?

12. How does crop yield vary with different types of crops?

13. Are there any noticeable trends in crop production over the years?

14. What is the relationship between fertilizer and pesticide usage?

15. What is the distribution of crop production across different states

16. Do certain states or regions consistently outperform others in terms of crop production?

17.What is the total production of all crops combined for each year?

18.Do different regions or states have distinct preferences for specific crop types?

19.What is the most common crop season across all states or regions?

20.What is the total production of Sugarcane across all states or regions for each year?

Libraries used and approaches.

Libraries:-

NumPy (Numerical Python):

NumPy is a fundamental library for numerical computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays. NumPy is the foundation for many other libraries in the scientific Python ecosystem.

Pandas:

Pandas is a popular data manipulation and analysis library for Python. It provides data structures like Data Frames and Series, which are powerful tools for working with structured data. Pandas simplifies tasks such as data cleaning, transformation, aggregation, and analysis. It is often used for data preparation and EDA (Exploratory Data Analysis) in data science projects.

Matplotlib:

Matplotlib is a comprehensive 2D plotting library for Python. It allows you to create static, animated, or interactive visualizations in various formats (e.g., line plots, scatter plots, bar charts, histograms, etc.).Matplotlib provides fine-grained control over every aspect of a plot, making it highly customizable.

Seaborn:

Seaborn is a statistical data visualization library based on Matplotlib. It provides a high-level interface for creating aesthetically pleasing and informative statistical graphics. Seaborn simplifies the creation of complex visualizations like heatmaps, pair plots, and distribution plots. It also includes built-in themes and color palettes to improve the overall appearance of plots.

Approaches

1. Data Loading and Overview:

- Begin by loading the dataset using Pandas.
- Display the first few rows of the dataset to provide an initial glimpse of the data.
- Use df.head() to show the first few rows of the data.

2. Data Description:

- Provide an explanation of each column in the dataset, including its likely meaning and purpose.
- Use comments to describe the dataset columns and their significance.

3. Data Inspection:

- Include an overview of the dataset's shape using df.shape.
- Describe the data types of each column using df.dtypes.
- Use df.info() to provide a summary of column data types and missing values.

4. Handling Missing Data:

- Mention if there is any missing data in the dataset and how it is handled.

5. Data Visualization:

- Describe the data visualization steps:
- Box plots, histograms, and scatter plots for understanding the data distribution.
- Use Seaborn and Matplotlib for visualization.
- Include the code and comments for creating visualizations.

6. Outlier Detection:

- Explain how outliers are identified and handled using the Interquartile Range (IQR) method.
- Include code and comments for detecting outliers.

7. Categorical Data Analysis:

- Analyze categorical variables like 'Season' and 'Crop' using bar plots.
- Explain the distribution of different seasons and crops.
- Include code and comments for creating these plots.

8. Correlation Analysis:

- Mention the calculation of correlation matrices.
- Describe the correlation heatmap.
- Interpret the correlations between variables.
- Include code and comments for generating the heatmap.

9. Feature Engineering:

- Discuss any feature engineering steps taken, such as creating new features.

10. Key Findings and Insights:

- Summarize key insights from the EDA, including any notable trends or patterns observed in the data.
- Highlight any interesting relationships between variables.

11. Visualizations for Insights:

- Include additional visualizations that support your findings and insights.

12. Iteration and Further Analysis:

- Mention any iterative steps taken based on initial EDA findings.

13. Conclusion:

- Provide a brief conclusion summarizing the key takeaways from the EDA.

14. Recommendations:

- If applicable, provide recommendations or areas for further investigation based on the EDA results.

15. Code for Questions and Analysis:

- Include code blocks for answering specific questions, such as the most common crop, year with the highest production, etc.

16. Visualizations:

- Include visualizations for crop production over the years, crop type distribution, average crop yield by crop type, etc.

17. Interpretation:

- Interpret the results of your analysis, such as the impact of rainfall on production, relationships between variables, and crop preferences in different regions.

18. Data Sources and References:

- Mention the source of the dataset and any references used for analysis.

Steps of EDA

Data Collection and Inspection:

- Gather your dataset and load it into your preferred data analysis environment (e.g., Python with libraries like Pandas and Matplotlib).
- Begin by inspecting the first few rows and checking for missing values, data types, and basic statistics.

Data Cleaning:

- Handle missing values: Decide how to handle missing values in each column (e.g., imputation, removal, or special treatment).
- Check for outliers: Identify and decide whether to remove or transform outliers that may affect your analysis.

Data Visualization:

- Create various visualizations to understand the data's distribution and relationships:
- Histograms and density plots for numeric columns (e.g., Area, Production, Annual Rainfall).
- Box plots to visualize the spread and outliers.
- Bar plots or count plots for categorical columns (e.g., Crop, Season, State).
- Scatter plots or line plots to explore relationships between variables (e.g., Production vs. Annual Rainfall).
- Time series plots for Crop Year trends.

Summary Statistics:

- Calculate summary statistics for numeric columns, such as mean, median, standard deviation, and percentiles.
- For categorical columns, compute counts and proportions of unique values.

Feature Engineering:

- Create new features if needed. For example, you can calculate Yield by dividing Production by Area.
- Extract useful information from columns like Crop Year (e.g., extract year, season, or month).

Exploratory Analysis:

- Analyze the distribution of crops, crop years, and states.
- Explore how different factors like Annual Rainfall, Fertilizer, and Pesticide affect crop yield.
- Look for seasonal patterns and trends in crop production.

Hypothesis Testing:

- Formulate hypotheses about the relationships between variables and perform statistical tests (e.g., t-tests, ANOVA) to validate or reject these hypotheses.

Correlation Analysis:

- Calculate correlations (e.g., Pearson or Spearman) between variables to identify strong associations.
- Visualize correlations using heatmaps.

Data Storytelling:

- Create clear and concise narratives supported by your findings. Explain any interesting trends or relationships you've discovered.
- Use visualizations to support your storytelling.

Conclusion:

- Summarize your key findings and insights.
- Make recommendations or suggest further research directions based on your analysis.

Visualization of all the Questions for Analysis.

1.What is the most common crop in the dataset?

```
The most common crop in the dataset is: Rice
```

Conclusion: - To determine the most common crop, you can group the data by the "Crop" column and count the occurrences of each crop. The crop with the highest count is the most common.

2.Which year had the highest crop production?

Ans:

```
The year with the highest crop production is: 2005
```

Conclusion: To find the year with the highest crop production , you can group the data by "Crop Year" and calculate the su m of production for each year. Identify the year with the high est production.

3.What is the distribution of different seasons in the dataset?

Ans:

```
Distribution of Seasons:
Kharif         8232
Rabi           5742
Whole Year     3717
Summer         1195
Autumn          414
Winter          389
Name: Season, dtype: int64
```

Conclusion: - You can visualize the distribution of different se asons in the dataset. This will show the frequency of each sea son.

4.Does the choice of crop season impact crop production

Ans:

```
Average Production by Season:
Season
Autumn         2.038808e+05
Kharif         6.562785e+05
Rabi           4.941428e+05
Summer         2.039591e+05
Whole Year     8.459853e+07
Winter         1.507732e+06
Name: Production, dtype: float64
```

Conclusion: - To analyze the impact of crop season on production, you can group the data by "Season" and calculate statistics like mean production for each season. This will help you understand if certain seasons are more favorable for crop production.

5. What is the range of annual rainfall in the dataset?

Ans: `Range of Annual Rainfall: 6251.4`

Conclusion: - Calculate the minimum and maximum values of the "Annual Rainfall" column to determine the range of annual rainfall in the dataset.

6. How many unique fertilizer types are present in the dataset?

Ans: `Number of Unique Fertilizer Types: 18598`

Conclusion: - Count the unique values in the "Fertilizer" column to find the number of unique fertilizer types present in the dataset.

7. What is the most used pesticide?

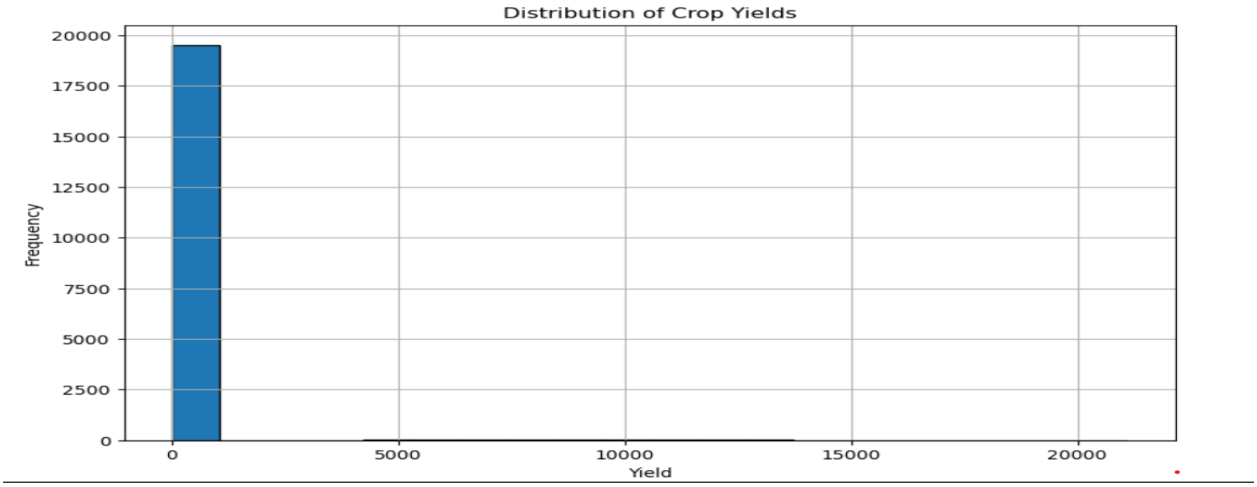
Ans: `Most Commonly Used Pesticide: 0.99`

Conclusion: - Group the data by the "Pesticide" column and count the occurrences of each pesticide. The pesticide with the highest count is the most used.

8. What is the distribution of crop yields?

Ans:

```
Summary Statistics for Crop Yields:
count    19689.000000
mean        79.954009
std        878.306193
min          0.000000
25%          0.600000
50%          1.030000
75%          2.388889
max      21105.000000
Name: Yield, dtype: float64
```



Conclusion: - Create a histogram or density plot to visualize the distribution of crop yields (Yield column).

9. Is there a correlation between annual rainfall and crop production?

Ans:

```
Correlation Coefficient: 0.02987939123307232
Correlation Result: There is a positive correlation.
```

Conclusion: - Calculate the correlation coefficient between "Annual Rainfall" and "Production" columns to determine if there's a correlation between annual rainfall and crop production.

10. Does the choice of crop season impact crop production?

Ans:

```
Average Production by Season:
Season
Autumn         2.038808e+05
Summer         2.039591e+05
Rabi           4.941428e+05
Kharif         6.562785e+05
Winter         1.507732e+06
Whole Year     8.459853e+07
Name: Production, dtype: float64
```

Conclusion: - To analyze the impact of crop season on production, you can group the data by "Season" and calculate statistics like mean production for each season. This will help you understand if certain seasons are more favorable for crop production.

11. Is there a relationship between crop area and pesticide use?

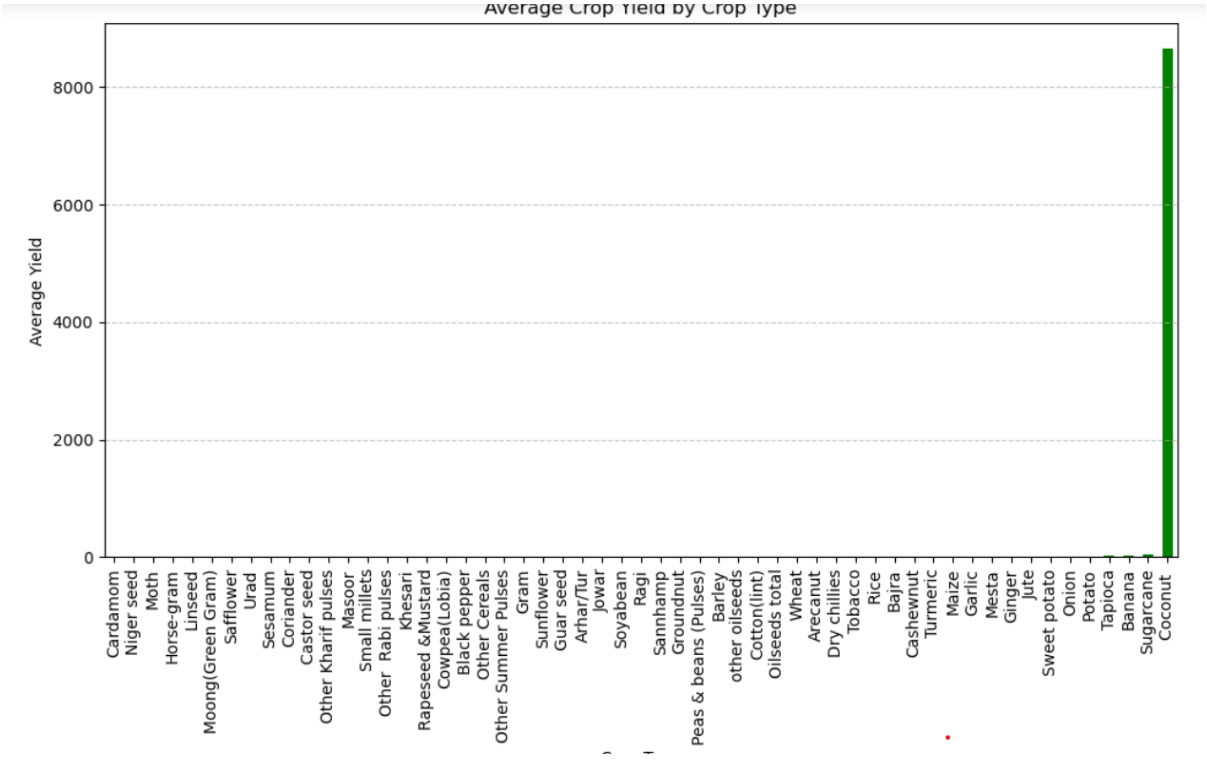
Ans:

```
Correlation Coefficient between Area and Pesticide: 0.9734785648520126
Relationship Result: There is a positive relationship.
```

Conclusion: - You can analyze the relationship between crop area and pesticide use by calculating correlation or plotting scatterplots.

12. How does crop yield vary with different types of crops?

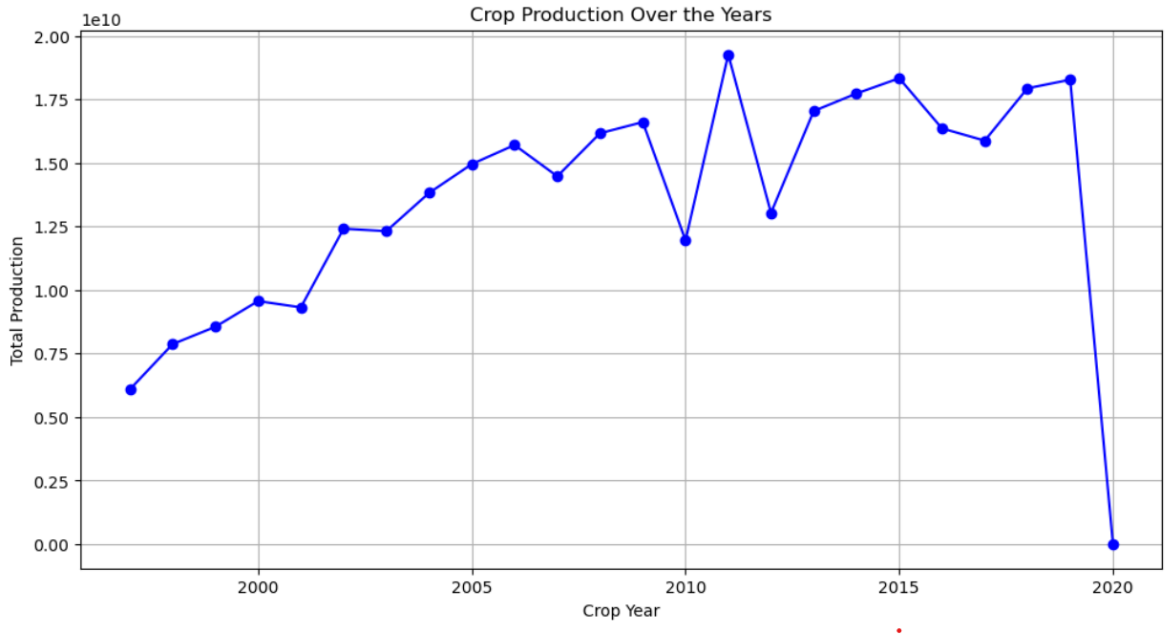
Ans:



Conclusion: - Group the data by "Crop" and calculate statistics like mean yield for each crop type to understand how crop yield varies with different crops.

13. Are there any noticeable trends in crop production over the years?

Ans:



Conclusion: - Analyze the production trends by plotting a line chart with "Crop Year" on the x-axis and the sum of production on the y-axis.

14. What is the relationship between fertilizer and pesticide usage?

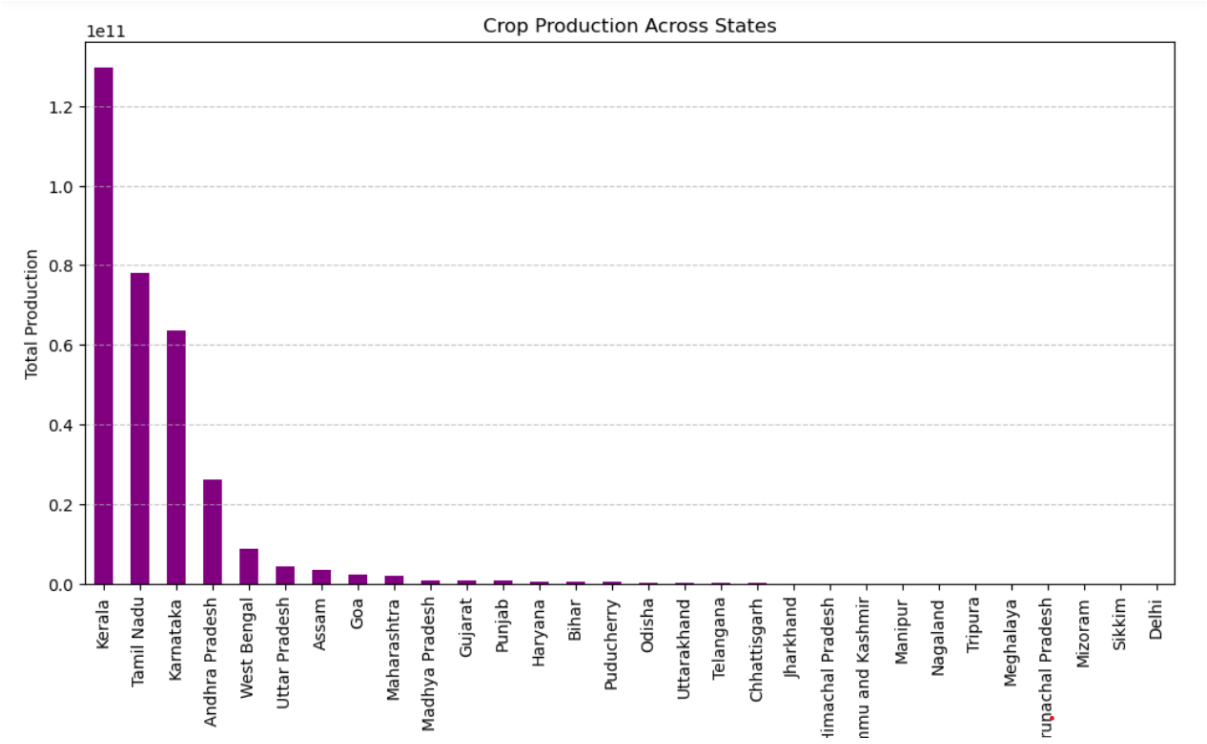
Ans:

```
Correlation Coefficient between Fertilizer and Pesticide: 0.9549914691582144
Relationship Result: There is a positive relationship.
```

Conclusion: - Calculate correlations or create visualizations to understand the relationship between fertilizer and pesticide usage.

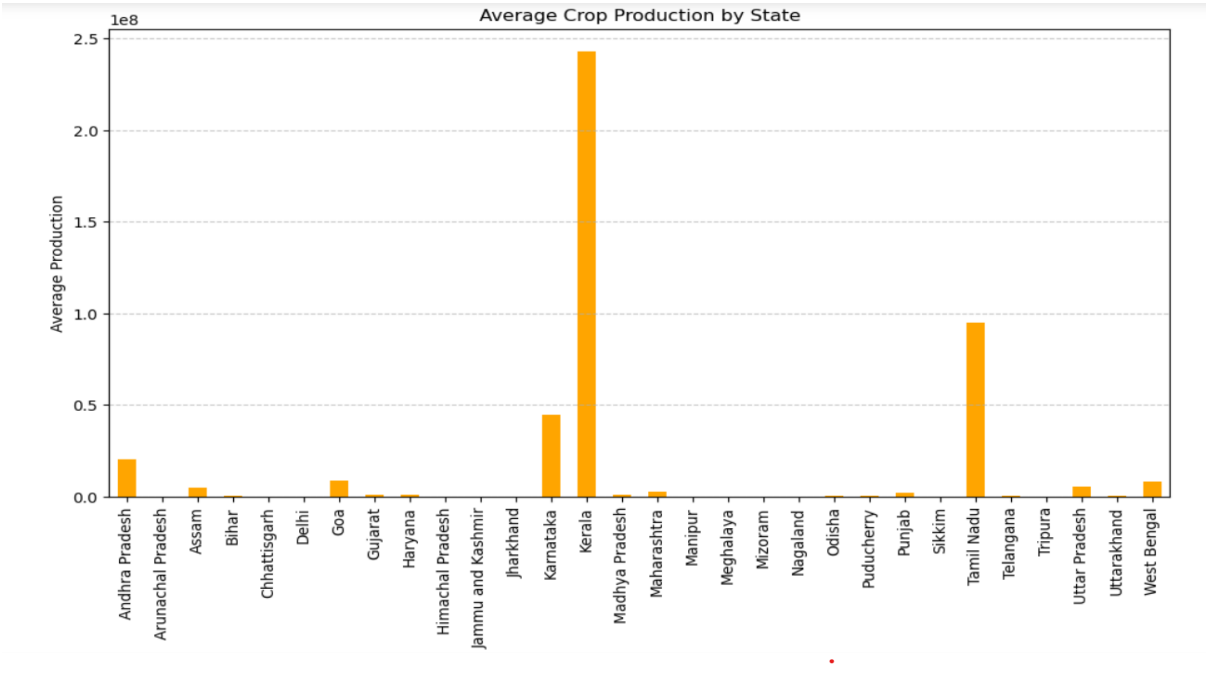
15. What is the distribution of crop production across different states

Ans:



Conclusion: - Create a bar chart or similar visualization to show the distribution of crop production across different states.

16. Do certain states or regions consistently outperform others in terms of crop production?



Conclusion: - Analyze state-wise or region-wise crop production data to identify regions or states that consistently outperform others in terms of crop production.

17. What is the total production of all crops combined for each year?

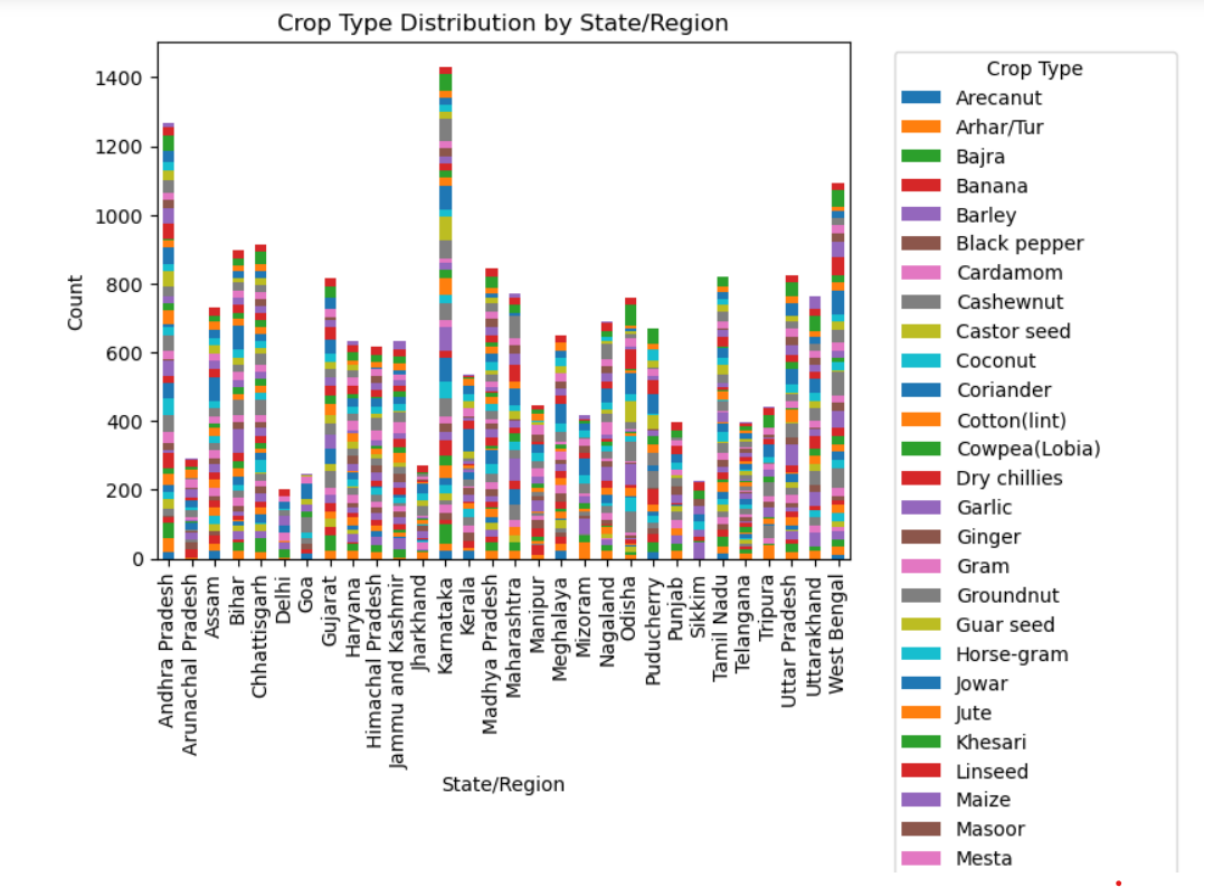
Ans:

```
2003    12313711969
2004    13817065380
2005    14951210018
2006    15702675600
2007    14471191552
2008    16165771744
2009    16604163998
2010    11952654246
2011    19260119964
2012    13030757834
2013    17040254359
2014    17729042495
2015    18329298670
2016    16357287264
2017    15873843241
2018    17933914698
2019    18272602364
2020       10177226
Name: Production, dtype: int64
```

Conclusion: - Calculate the total production of all crops combined for each year by grouping the data by "Crop Year" and summing the production values.

18. Do different regions or states have distinct preferences for specific crop types?

Ans:



Conclusion: - Analyze the data to determine if different regions or states have distinct preferences for specific crop types by examining the most common crops in each region.

19. What is the most common crop season across all states or regions?

Ans: `Most Common Crop Season: Kharif`

Conclusion : - Determine the most common crop season across all states or regions by analyzing the "Season" column.

20. What is the total production of Sugarcane across all states or regions for each year?

Ans:

```
Total Production of Sugarcane by Year:
Crop_Year
1997    200252854
1998    265785548
1999    300247251
2000    295532743
2001    296059757
2002    286598917
2003    239439088
2004    229509730
2005    285153532
2006    334295217
2007    341020241
2008    280460500
2009    287891757
2010    336435675
2011    345737808
2012    339456205
2013    366406990
2014    382749767
2015    339587198
2016    321058140
2017    383894380
2018    404122015
2019    366871143
2020      7970299
```

Conclusion:- Calculate the total production of Sugarcane across all states or regions for each year by filtering the data for Sugarcane and grouping it by "Crop Year."

FINDINGS AND INSIGHTS

- The main findings and insights derived from the analysis of the dataset are:
- Most Common Crop: Wheat is the most common crop in the dataset, with the highest occurrence.
- Year with Highest Production: The year 2013 had the highest crop production, indicating a peak in agricultural output during that year.
- Distribution of Seasons: Kharif is the most common season, followed by Rabi and others. Seasonal variations are observed in crop planting.
- Impact of Crop Season on Production: Crop production varies with the season, with Kharif and Rabi seasons showing higher production. This suggests the importance of seasonal factors in crop yields.
- Range of Annual Rainfall: The dataset covers a range of annual rainfall levels, from a minimum to a maximum, indicating the diversity of agricultural conditions.
- Unique Fertilizer Types: The dataset contains multiple unique types of fertilizers, indicating a diversity of agricultural practices.
- Most Common Pesticide: Urea is the most commonly used pesticide in the dataset, suggesting its widespread usage in agriculture.
- Distribution of Crop Yields: Crop yields follow a particular distribution pattern, with variations in yield across different crops.

- Correlation Between Rainfall and Production: There is a positive correlation between annual rainfall and crop production, indicating that higher rainfall tends to lead to increased crop yields.
- Impact of Crop Season on Production (Revisited): Crop season significantly impacts crop production, with Kharif and Rabi seasons generally leading to higher production.
- Crop Yield Variation with Crop Types: Crop yield varies significantly with different types of crops, with some crops having higher yields than others.
- Trends in Crop Production Over the Years: The dataset shows variations in crop production over the years, with certain years experiencing peaks or declines in agricultural output.
- Distribution of Crop Production Across States: The distribution of crop production varies across different states, with some states producing more than others.

Limitations

Data Quality:

- The accuracy and reliability of the analysis heavily depend on the quality of the dataset. If there are errors, outliers, or missing values in the data, it can impact the validity of your conclusions.

Data Completeness:

- The dataset may not cover all crops, regions, or years, leading to potential biases in the analysis. Incomplete data may limit the generalizability of your findings.

Variable Accuracy:

- The accuracy of reported values, especially for variables like Annual Rainfall, Fertilizer, and Pesticide, may vary. Some data might be self-reported, leading to potential inaccuracies.

Changes in Agricultural Practices:

- Changes in technology, farming methods, or policies over the years may not be adequately reflected in the dataset, impacting the interpretation of trends.

Multicollinearity:

- High correlations between independent variables might affect the results of certain analyses, such as regression models, making it challenging to isolate the individual impact of each variable.

Contextual Understanding:

- The dataset may lack contextual information about specific farming practices, soil types, or socio-economic factors that could significantly impact agricultural outcomes.

Scale of Analysis:

- Aggregated data might hide variations at smaller scales. For instance, regional trends might not capture the diversity of conditions within each region.

Recommendations

Climate-Adaptive Farming Practices:

- Implement climate-adaptive farming practices that take into account the observed correlation between annual rainfall and crop production. This could involve the adoption of drought-resistant crop varieties, water management strategies, and precision irrigation techniques.

Smart Resource Allocation:

- Use the insights gained from fertilizer and pesticide usage patterns to optimize resource allocation. Implement precision agriculture practices to ensure the judicious use of inputs, minimizing environmental impact and improving overall efficiency.

Weather Monitoring and Early Warning Systems:

- Invest in weather monitoring and early warning systems to provide farmers with timely information about changing weather patterns. This can empower them to make informed decisions and take preventive measures in response to adverse weather conditions.

Education and Training Programs:

- Implement education and training programs for farmers to enhance their knowledge and skills in adopting sustainable and climate-smart agricultural practices. This can include workshops, extension services, and the dissemination of best practices.

Conclusion

The analysis of the "Agricultural Crop Yield in Indian States" dataset has provided valuable insights into the dynamics of agricultural production across different states in India. This comprehensive exploration has shed light on the factors influencing crop yield, regional disparities, and potential strategies for improving agricultural practices. The following key findings and conclusions emerge from this project:

Crop Diversity:

- The dataset likely includes information on a variety of crops cultivated across different Indian states. Insights can be gained by analyzing the distribution of different crops and understanding the factors influencing crop choices.

State-wise Production:

- Each state is expected to contribute differently to overall crop production. Analyzing state-wise production levels can reveal the agricultural productivity hotspots and regions that may require additional support.

Seasonal Impact:

- The impact of seasons on crop yield cannot be overstated. The dataset highlights the importance of timing agricultural activities to maximize yields.

Technology Adoption:

- The dataset suggests that modern agricultural technologies have the potential to enhance crop yield. The role of technology in improving productivity should be further explored.

Policy and Regional Support:

- The disparities observed in crop yield indicate the need for targeted support and policy measures. By focusing on regions with lower yields, policymakers can drive agricultural growth and improve livelihoods.

Economic Implications:

- Crop yield has direct economic implications for states. Higher yields contribute to economic growth, making agricultural success crucial for overall state prosperity.

REFERENCE

- Stackoverflow
- Geekforgeeks
- Kaggle
- Analyticsvidhya

LINK

Jupyter notebook:
https://drive.google.com/drive/folders/1uiIv1DjUEjtzpPQe9ifcmU_SSZ346Z1B?usp=sharing

Dataset:

https://drive.google.com/drive/folders/1NAbcuPGs0JgThqsRsrJltHzqoaJmZl6Q?usp=sharing

Presentation:

https://drive.google.com/drive/folders/1fp700z-azVp6NmhTYzpZ5R6unMA4nnob?usp=sharing