

INT 375: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

Production of Crops In India

Submitted by: G somsunil

Registration No: 12322522

Programme and Section: B.TECH(CSE), K23GF

Course Code: INT375

Under the Guidance of

Mrs. Aashima Ma'am

Discipline of CSE/IT

Lovely School of Computer Science Engineering

Lovely Professional University, Phagwara

CERTIFICATE

This is to certify that G Somsunil bearing Registration no. 12322522 has completed INT375 project titled, **“Production of Crops in India”** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Signature and Name of the Supervisor

Designation of the Supervisor

School of Engineering(CSE)

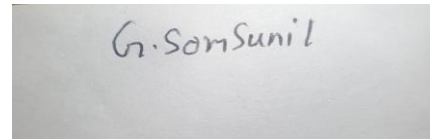
Lovely Professional University

Phagwara, Punjab.

Date: 12April, 2025

DECLARATION

I, G Somsunil student of B.Tech under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

A rectangular box containing a handwritten signature in blue ink that reads "G. SomSunil".

Date: 12 April, 2025

Signature

Registration No. 12322522

Student's Name: G. SomSunil

⇒ **Table of Content**

1. Introduction
2. Source of dataset
3. EDA process
4. Analysis on dataset (for each analysis)
 - i. Introduction
 - ii. General Description
 - iii. Specific Requirements, functions and formulas
 - iv. Analysis results
 - v. Visualization
5. Conclusion
6. References

1. Introduction:

Agriculture forms the backbone of the Indian economy, employing a vast segment of the population and contributing significantly towards the country's GDP. Out of the various aspects of agriculture, crop cultivation is an essential aspect in food security, employment generation, and export. Based on the geography, climate, and soil variance in India, the nation produces a vast number of crops every year.

This research study is centered on investigating the production trends of major crops in India across various years. Employing data analysis methods in Python, the research strives to detect significant patterns, detect high-yielding crops, and emphasize regional production capabilities. Employing robust libraries such as NumPy, Pandas, Matplotlib, and Seaborn, the research visualizes the data using heatmaps, fairplots, and pair plots. Furthermore, geospatial libraries such as Folium are employed to present production data geographically across Indian states and thus provide a

spatial perspective to agricultural productivity.

The main aim of this initiative is to enable data-driven conclusions that can potentially improve agricultural planning, guide policy, and maximize resource allocation. Identification of trends in crop production is important not only for farmers and the government but also for agri-tech companies and researchers engaged in improving productivity and sustainability in Indian agriculture.

2. Source of Dataset

The dataset used in this project is titled "**Open Government Data (OGD)** ", published by the **Central Government of India**

- **State Name**
- **Crop Year**
- **Seasons**
- **Area**

3. EDA Process (Exploratory Data Analysis)

The EDA stage is critical in detecting trends and patterns in the data set. This project used the crop production data set with the aid of Python libraries such as Pandas, NumPy, Seaborn, and Matplotlib.

1Data Overview

The dataset contains attributes like State_Name, District_Name, Crop_Year, Season, Crop, Area, and Production.

Initial inspection revealed missing values and inconsistencies, particularly in Production and Area columns.

```
[1]: import pandas as pd
import numpy as np
import seaborn as sn
import matplotlib.pyplot as plt
```

```
[7]: df = pd.read_csv("C:\\Users\\LENOVO\\Downloads\\apy (1).csv")
df
```

```
[7]:
```

	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254.0	2000.0
1	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2.0	1.0
2	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102.0	321.0
3	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176.0	641.0
4	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720.0	165.0
...
246086	West Bengal	PURULIA	2014	Summer	Rice	306.0	801.0
246087	West Bengal	PURULIA	2014	Summer	Sesamum	627.0	463.0
246088	West Bengal	PURULIA	2014	Whole Year	Sugarcane	324.0	16250.0
246089	West Bengal	PURULIA	2014	Winter	Rice	279151.0	597899.0
246090	West Bengal	PURULIA	2014	Winter	Sesamum	175.0	88.0

246091 rows × 7 columns

```
[3]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 246091 entries, 0 to 246090
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 
dtypes: float64(2), int64(1), object(4)
memory usage: 13.1+ MB
```

```
[15]: df.duplicated()
```

```
[15]: 0      False
1      False
2      False
3      False
4      False
...
246086 False
246087 False
246088 False
246089 False
246090 False
Length: 246091, dtype: bool
```



```
[19]: df.describe()
```

	Crop_Year	Area	Production
count	246091.000000	2.460910e+05	2.423610e+05
mean	2005.643018	1.200282e+04	5.825034e+05
std	4.952164	5.052340e+04	1.706581e+07
min	1997.000000	4.000000e-02	0.000000e+00
25%	2002.000000	8.000000e+01	8.800000e+01
50%	2006.000000	5.820000e+02	7.290000e+02
75%	2010.000000	4.392000e+03	7.023000e+03
max	2015.000000	8.580100e+06	1.250800e+09

4. Analysis on Dataset

1) The line chart displays the trend of total crop production (in green) and total cultivated area (in blue) in India from 1997 to 2015. A few key observations can be made:

Rising Production Trend:

From 1997 onwards, there is a significant increase in total production, peaking around 2011–2013. This could be due to improvements in agricultural techniques, irrigation, fertilizers, or high-yield crop varieties.

Stable Cultivation Area:

The total area under cultivation shows a relatively flat trend with only minor fluctuations. This suggests that the increase in production was

achieved without a significant expansion in land use, indicating better productivity per hectare.

Sharp Decline in 2015:

A noticeable drop in production is seen in 2015. This could be due to unfavorable weather conditions such as drought or floods, or possibly data gaps or errors in the dataset.

Yield Efficiency Growth:

Since the cultivated area remained nearly constant while production increased, it implies that **crop yield (production per unit area)** improved over time. This reflects positively on agricultural efficiency.

Code used

```
yearly_trend = df.groupby('Crop_Year')[['Area',  
'Production']].sum().reset_index()
```

```
plt.figure(figsize=(14, 6))
```

```
sn.lineplot(x='Crop_Year', y='Production', data=yearly_trend,  
label='Production', color='green')
```

```
sn.lineplot(x='Crop_Year', y='Area', data=yearly_trend, label='Area',  
color='blue')
```

```
plt.title("Total Production and Area Over the Years")
```

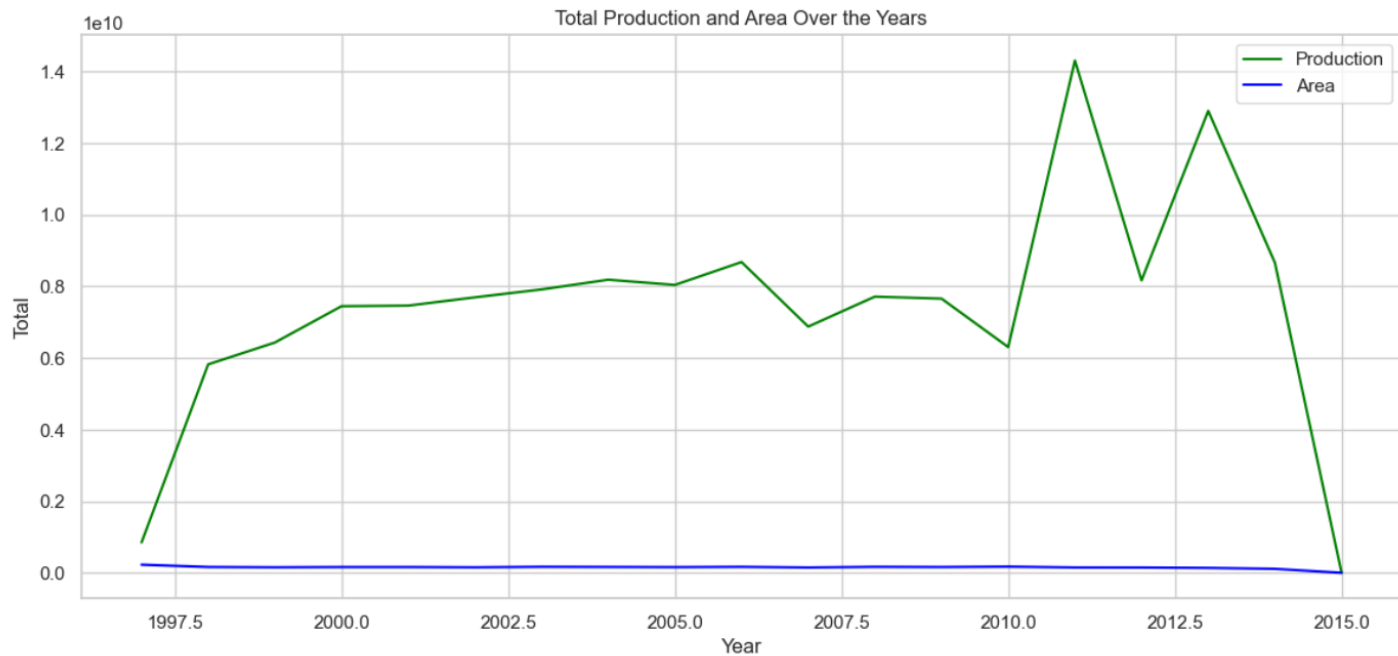
```
plt.xlabel("Year")
```

```
plt.ylabel("Total")
```

```
plt.legend()
```

```
plt.grid(True)
```

```
plt.show()
```



2)Analysis: Distribution of Production by Season

The box plot compares the distribution of crop production across different seasons — Kharif, Rabi, Autumn, Summer, Winter, and Whole Year. Here's what we can observe:

Winter Season Dominates:

The Winter season has the highest median production and also a wide range, indicating both high output and variability. Some extreme outliers suggest that a few crops or states perform exceptionally well in winter.

Consistent Performance in Autumn and Rabi:

These seasons show relatively moderate and consistent production levels. Autumn has slightly higher variability, while Rabi shows a tighter range around the median.

Summer and Kharif Seasons Have Lower Output:

Both Kharif and Summer seasons have lower median production. Summer shows the least variability, suggesting consistent but low production across the country.

Whole Year Category is Moderate:

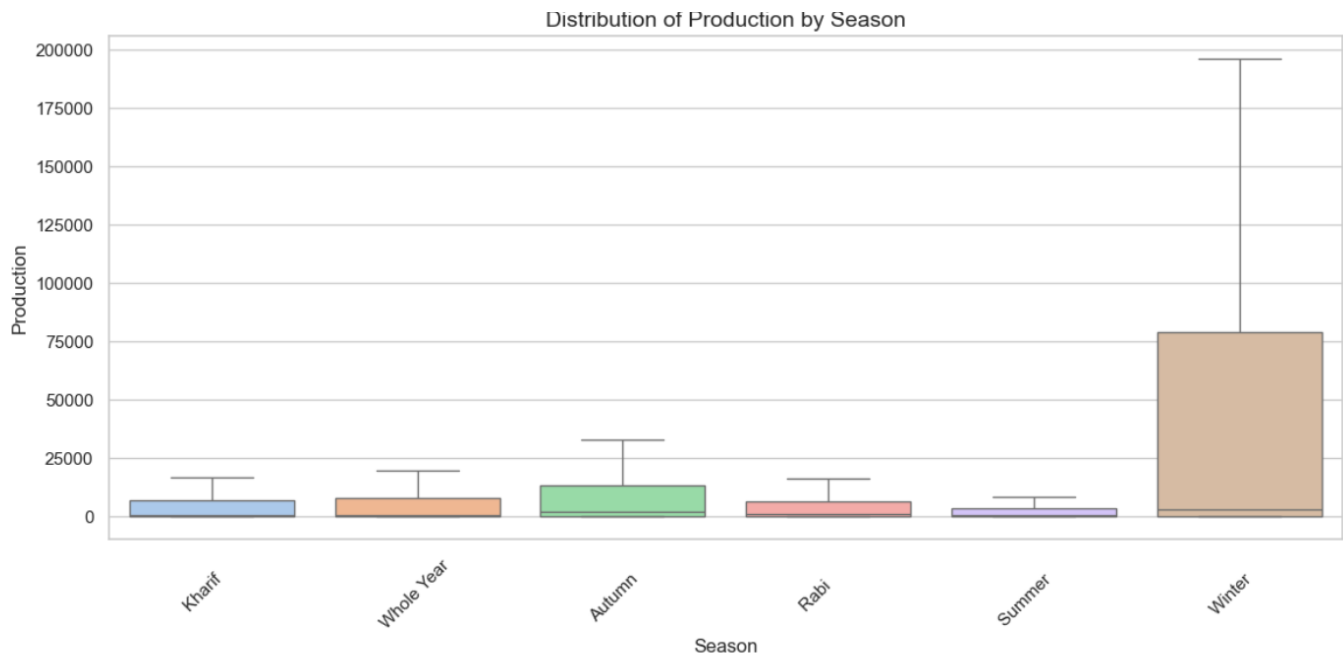
The "Whole Year" category likely includes perennial or multi-season crops. It has a moderate median with some outliers, but overall the range is narrower than that of Winter or Autumn.

Variability & Outliers:

The presence of significant outliers, especially in the Winter season, highlights the non-uniformity in crop production — possibly due to regional or crop-specific differences.

Code used

```
df_clean = df.dropna(subset=['Production'])  
plt.figure(figsize=(12, 6))  
sn.boxplot(  
    data=df_clean,  
    x="Season",  
    y="Production",  
    showfliers=False,  
    palette="pastel"  
)  
plt.title("Distribution of Production by Season", fontsize=14)  
plt.xticks(rotation=45)  
plt.xlabel("Season", fontsize=12)  
plt.ylabel("Production", fontsize=12)  
plt.tight_layout()  
plt.show()
```



3) Analysis: Area vs Production for Top 5 Crops

This scatter plot visualizes the relationship between the area cultivated (in hectares) and the production output (in tonnes) for India's top five crops — **Rice, Maize, Moong (Green Gram), Urad, and Sesamum**.

Key Observations:

Rice Leads in Both Area and Production:

Rice (in red) clearly dominates the plot with the **highest number of data points**, the **widest spread**, and the **highest production values**, even with moderate area usage. This underlines rice's central role in Indian agriculture.

Maize Shows Strong Correlation:

Maize (in blue) follows a more **linear trend**, showing that as the area increases, production increases proportionally. It suggests maize has relatively consistent yield efficiency.

Pulses (Moong, Urad, Sesamum) Show Limited Scale:

The remaining crops — Moong (green), Urad (purple), and Sesamum (orange) — are clustered at the lower end of both axes, indicating **lower cultivated area and production**. These are likely grown in limited regions or under specific conditions.

High Density & Overlap:

The plot is heavily dense in the bottom-left quadrant, reflecting that **most records** fall under **lower production and area** ranges. Only a few data points extend into very high area/production values (outliers).

Efficiency Pointers:

Rice and maize seem to have better scalability and productivity per hectare compared to the other crops, which could influence future agricultural policy and crop planning.

Code used

```
top_crops = df_clean['Crop'].value_counts().head(5).index
```

```
df_top_crops = df_clean[df_clean['Crop'].isin(top_crops)]
```

```
plt.figure(figsize=(10, 6))
```

```
sn.scatterplot(data=df_top_crops, x='Area', y='Production', hue='Crop', alpha=0.6,  
palette='Set1')
```

```
plt.title("Area vs Production (Top 5 Crops)")
```

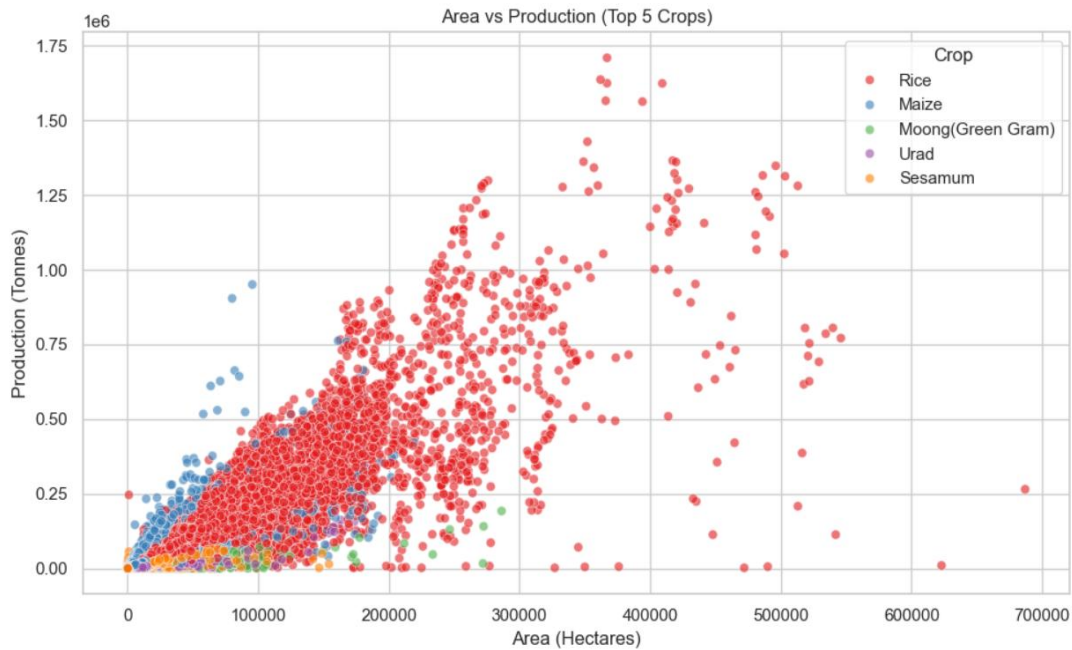
```
plt.xlabel("Area (Hectares)")
```

```
plt.ylabel("Production (Tonnes)")
```

```
plt.legend(title='Crop')
```

```
plt.tight_layout()
```

```
plt.show()
```



4)Top 15 Crops by Number of Records: Insights

This horizontal bar chart presents the 15 most frequently recorded crops in the dataset, providing insight into **data availability and crop cultivation trends**.

Key Highlights:

Rice Dominates:

With over **14,000 records**, rice tops the list, reflecting its massive cultivation scale and consistent tracking across regions and years.

Maize & Pulses Follow:

Maize, Moong (Green Gram), and Urad are next, each with **10,000+ records**, indicating their importance in Indian agriculture, especially in rain-fed areas.

Oilseeds & Legumes Well Represented:

Crops like **Sesamum, Groundnut, and Rapeseed & Mustard** have significant representation, pointing to their widespread cultivation and economic importance.

Staples & Commercial Crops in Middle Tier:

Sugarcane, Wheat, Gram, Jowar, and Onion hold moderate positions, showing a balanced level of data presence and possibly region-specific cultivation.

Potato & Dry Chillies at the Bottom of the Top 15:

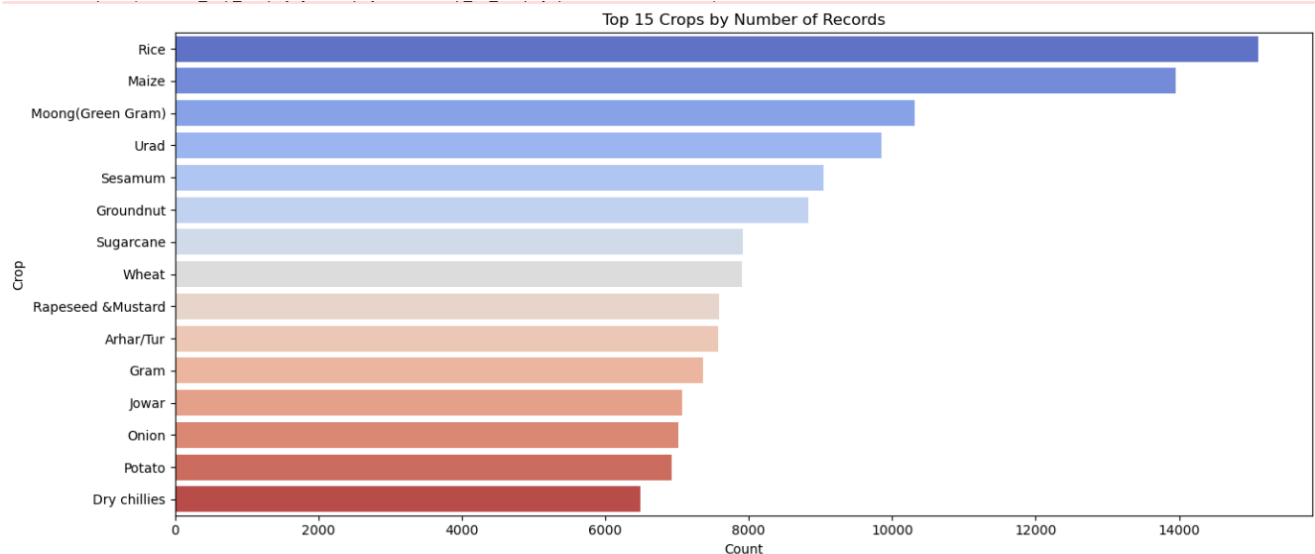
Although they are essential crops, their lower record count could be due to regional limitations or lesser data collection compared to cereals and pulses.

Code used

```
top_15_crops = df['Crop'].value_counts().head(15).index
```

```
df_top_crops = df[df['Crop'].isin(top_15_crops)]
```

```
plt.figure(figsize=(14, 6))
sn.countplot(data=df_top_crops, y='Crop', order=top_15_crops, palette='coolwarm')
plt.title("Top 15 Crops by Number of Records")
plt.xlabel("Count")
plt.ylabel("Crop")
plt.tight_layout()
plt.show()
```



5) Analysis of Top 10 Crops by Total Production

The bar chart illustrates the top ten crops based on their total production. The analysis reveals a significant disparity in production volumes across different crops.

1. Coconut - The Leading Crop

Coconut stands out remarkably as the highest-produced crop, with a total production of **129,981,629,216 units**. This figure is **more than 23 times greater** than that of the second-most produced crop, Sugarcane. The exceptionally high number may be attributed to the unit of measurement (e.g., number of coconuts) or its vast cultivation in tropical regions.

2. Sugarcane

Sugarcane ranks second with a production of **5,535,681,525 units**. It plays a vital role in sugar and ethanol industries, and is widely cultivated in countries like India and Brazil.

3–4. Rice and Wheat

Rice (**1.6 billion units**) and Wheat (**1.3 billion units**) are essential staple foods for a large portion of the global population. Their production reflects their importance in daily diets and food security.

5–7. Potato, Cotton (Lint), and Maize

These crops fall in the mid-tier range:

- **Potato:** 424 million units
 - **Cotton (lint):** 297 million units
 - **Maize:** 273 million units
- Each of these has multiple uses — from food to textile and industrial applications.

8–10. Jute, Banana, and Soybean

Jute (**181 million**), Banana (**146 million**), and Soybean (**141 million**) complete the list. Despite lower production volumes, these crops are economically significant. For example, jute is used in packaging, while soybeans are crucial in oil and protein production.

Key Insights

- **Skewed Distribution:** Coconut heavily skews the distribution, showing a possible anomaly or a unique measurement scale.
- **Staple Crops:** Rice, wheat, and maize maintain high production due to their demand as food staples.
- **Diversification:** The list includes both food (e.g., potato, banana) and non-food crops (e.g., cotton, jute), showing agricultural diversity.

Code used

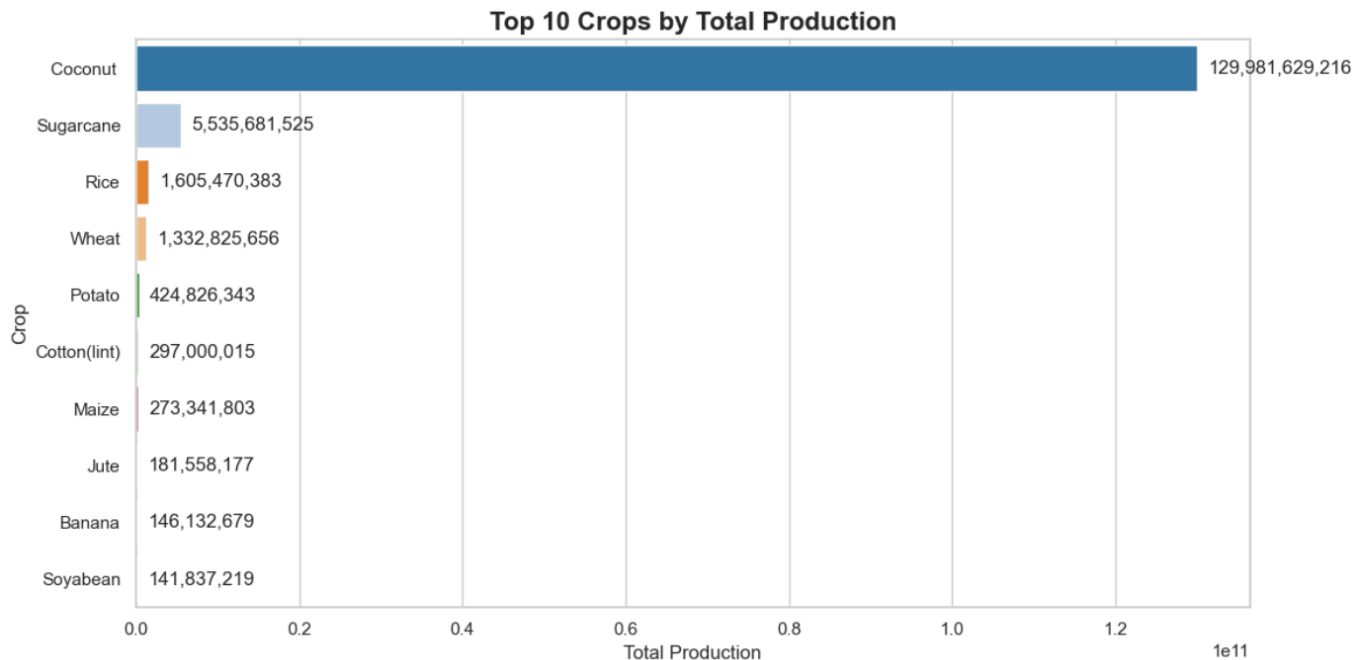
```
top_crops =  
df_clean.groupby('Crop')['Production'].sum().sort_values(ascending=False).head(10)  
  
plt.figure(figsize=(12, 6))  
  
sn.barplot(  
    x=top_crops.values,  
    y=top_crops.index,
```

```

palette=sn.color_palette("tab20", n_colors=10) # vibrant color palette
)
plt.title("Top 10 Crops by Total Production", fontsize=16, fontweight='bold')
plt.xlabel("Total Production", fontsize=12)
plt.ylabel("Crop", fontsize=12)
for index, value in enumerate(top_crops.values):
    plt.text(value + max(top_crops.values)*0.01, index, f'{int(value):,}', va='center')

plt.tight_layout()
plt.show()

```



6) Analysis of Correlation between Area and Production

The heatmap illustrates the correlation between the area under cultivation and total production of crops. The values range from -1 to +1, where:

- +1 indicates a perfect positive correlation,

- 0 indicates no correlation, and
- -1 indicates a perfect negative correlation.

Key Findings:

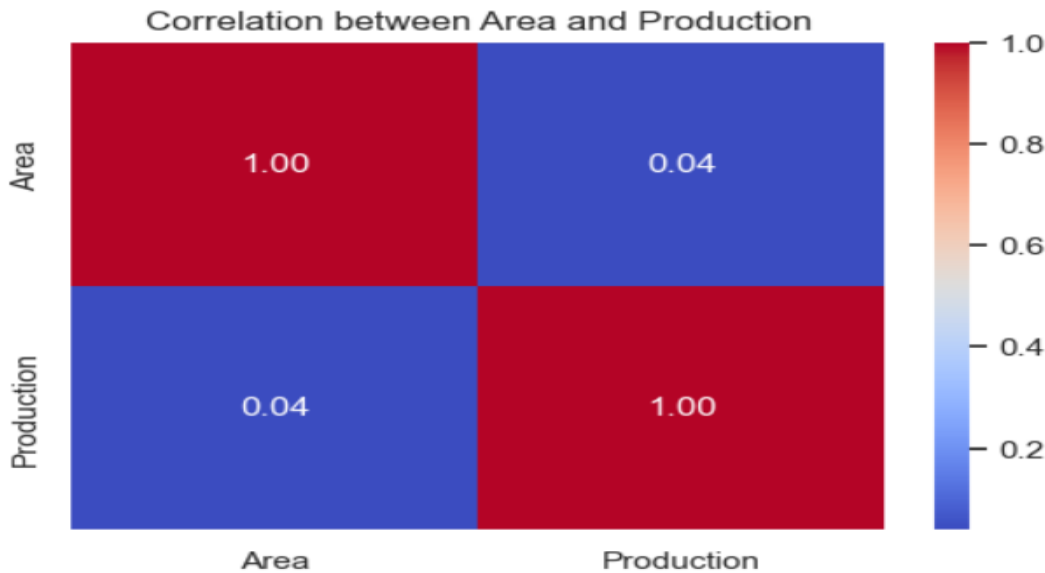
- **Correlation Value: 0.04**
 - The correlation between Area and Production is 0.04, which is very close to zero.
 - This indicates that there is almost no linear relationship between the amount of land used for cultivation and the total production output.
 - In other words, increasing the area does not necessarily increase production proportionally.

Interpretation:

- **Low Correlation Could Mean:**
 - Some crops may be high-yielding, producing large quantities from small areas (e.g., Coconut).
 - Other crops might require more space but produce less in terms of quantity.
 - Factors such as crop type, farming techniques, irrigation, soil fertility, and technology can influence production more than just land area.

Code used

```
plt.figure(figsize=(6, 4))  
correlation = df[['Area', 'Production']].corr()  
sn.heatmap(correlation, annot=True, cmap='coolwarm', fmt='.2f')  
plt.title("Correlation between Area and Production")  
plt.show()
```



5. Conclusion

The analysis of crop production data reveals that **Coconut** dominates total production by a significant margin, far exceeding other crops. However, a heatmap of correlation between area and production shows a **very weak correlation (0.04)**. This implies that **increasing the cultivation area does not necessarily lead to higher production**. Instead, crop productivity is influenced more by **factors such as crop type, yield efficiency, and farming practices**. Therefore, to improve agricultural output, focus should be placed on enhancing **productivity per unit area** rather than merely expanding cultivated land.

6. References:

- <https://www.data.gov.in/> Open Data Portal
- Python Libraries: NumPy, Pandas, Matplotlib, Seaborn
- Documentation:
 - ⇒ Pandas: <https://pandas.pydata.org/docs/>
 - ⇒ Matplotlib: <https://matplotlib.org/stable/index.html>

⇒ Seaborn: <https://seaborn.pydata.org/>

⇒ Dataset: [NCHS - Leading Causes of Death: United States - Catalog](#)

LinkedIn Activity:

<https://www.linkedin.com/in/sunil-guthula/>