```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

```
from google.colab import files
uploaded = files.upload()
```

Choose Files P7.csv

- **P7.csv**(text/csv) - 15316741 bytes, last modified: 7/8/2024 - 100% done

```
Saving P7.csv to P7 (1).csv
```

```
df = pd.read_csv('P7.csv')
df.head()
```

	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

```
df.tail()
```

	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
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

```
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
```

```
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
```

```
df.isnull().sum()
```

```
                  0
   State_Name     0
District_Name     0
    Crop_Year     0
       Season     0
         Crop     0
         Area     0
   Production  3730

dtype: int64
```

```
df.fillna(0, inplace=True)
df.isnull().sum()
```

```
               0
   State_Name  0
District_Name  0
    Crop_Year  0
       Season  0
         Crop  0
         Area  0
   Production  0

dtype: int64
```

```
df.nunique()
```

```
                   0
   State_Name     33
District_Name    646
    Crop_Year     19
       Season      6
         Crop    124
         Area  38442
   Production  51627

dtype: int64
```

```
df['State_Name'].unique()
```

```
array(['Andaman and Nicobar Islands', 'Andhra Pradesh',
       'Arunachal Pradesh', 'Assam', 'Bihar', 'Chandigarh',
       'Chhattisgarh', 'Dadra and Nagar Haveli', 'Goa', 'Gujarat',
       'Haryana', 'Himachal Pradesh', 'Jammu and Kashmir ', 'Jharkhand',
       'Karnataka', 'Kerala', 'Madhya Pradesh', 'Maharashtra', 'Manipur',
       'Meghalaya', 'Mizoram', 'Nagaland', 'Odisha', 'Puducherry',
       'Punjab', 'Rajasthan', 'Sikkim', 'Tamil Nadu', 'Telangana ',
       'Tripura', 'Uttar Pradesh', 'Uttarakhand', 'West Bengal'],
      dtype=object)
```

```
df['Crop_Year'].unique()
```

```
array([2000, 2001, 2002, 2003, 2004, 2005, 2006, 2010, 1997, 1998, 1999,
       2007, 2008, 2009, 2011, 2012, 2013, 2014, 2015])
```

```
df['Season'].unique()
```

```
array(['Kharif     ', 'Whole Year ', 'Autumn     ', 'Rabi       ',
       'Summer     ', 'Winter     '], dtype=object)
```

```
df['Crop'].unique()
```

```
array(['Arecanut', 'Other Kharif pulses', 'Rice', 'Banana', 'Cashewnut',
       'Coconut ', 'Dry ginger', 'Sugarcane', 'Sweet potato', 'Tapioca',
       'Black pepper', 'Dry chillies', 'other oilseeds', 'Turmeric',
       'Maize', 'Moong(Green Gram)', 'Urad', 'Arhar/Tur', 'Groundnut',
       'Sunflower', 'Bajra', 'Castor seed', 'Cotton(lint)', 'Horse-gram',
       'Jowar', 'Korra', 'Ragi', 'Tobacco', 'Gram', 'Wheat', 'Masoor',
       'Sesamum', 'Linseed', 'Safflower', 'Onion', 'other misc. pulses',
       'Samai', 'Small millets', 'Coriander', 'Potato',
       'Other  Rabi pulses', 'Soyabean', 'Beans & Mutter(Vegetable)',
       'Bhindi', 'Brinjal', 'Citrus Fruit', 'Cucumber', 'Grapes', 'Mango',
       'Orange', 'other fibres', 'Other Fresh Fruits', 'Other Vegetables',
       'Papaya', 'Pome Fruit', 'Tomato', 'Rapeseed &Mustard', 'Mesta',
       'Cowpea(Lobia)', 'Lemon', 'Pome Granet', 'Sapota', 'Cabbage',
       'Peas  (vegetable)', 'Niger seed', 'Bottle Gourd', 'Sannhamp',
       'Varagu', 'Garlic', 'Ginger', 'Oilseeds total', 'Pulses total',
       'Jute', 'Peas & beans (Pulses)', 'Blackgram', 'Paddy', 'Pineapple',
       'Barley', 'Khesari', 'Guar seed', 'Moth',
       'Other Cereals & Millets', 'Cond-spcs other', 'Turnip', 'Carrot',
       'Redish', 'Arcanut (Processed)', 'Atcanut (Raw)',
       'Cashewnut Processed', 'Cashewnut Raw', 'Cardamom', 'Rubber',
       'Bitter Gourd', 'Drum Stick', 'Jack Fruit', 'Snak Guard',
       'Pump Kin', 'Tea', 'Coffee', 'Cauliflower', 'Other Citrus Fruit',
       'Water Melon', 'Total foodgrain', 'Kapas', 'Colocosia', 'Lentil',
       'Bean', 'Jobster', 'Perilla', 'Rajmash Kholar',
       'Ricebean (nagadal)', 'Ash Gourd', 'Beet Root', 'Lab-Lab',
       'Ribed Guard', 'Yam', 'Apple', 'Peach', 'Pear', 'Plums', 'Litchi',
       'Ber', 'Other Dry Fruit', 'Jute & mesta'], dtype=object)
```

```
df['State_Name'].value_counts()
```

	count
State_Name	
Uttar Pradesh	33306
Madhya Pradesh	22943
Karnataka	21122
Bihar	18885
Assam	14628
Odisha	13575
Tamil Nadu	13547
Maharashtra	12628
Rajasthan	12514
Chhattisgarh	10709
Andhra Pradesh	9628
West Bengal	9613
Gujarat	8436
Haryana	5875
Telangana	5649
Uttarakhand	4896
Kerala	4261
Nagaland	3906
Punjab	3173
Meghalaya	2867
Arunachal Pradesh	2546
Himachal Pradesh	2494
Jammu and Kashmir	1634
Tripura	1412
Manipur	1267
Jharkhand	1266
Mizoram	957
Puducherry	876
Sikkim	714
Dadra and Nagar Haveli	263
Goa	208
Andaman and Nicobar Islands	203
Chandigarh	90

dtype: int64

```
df['Crop_Year'].value_counts().sort_index()
```

```
           count
Crop_Year
     1997   8899
     1998  11533
     1999  12515
     2000  13658
     2001  13361
     2002  16671
     2003  17287
     2004  14117
     2005  13799
     2006  14328
     2007  14526
     2008  14550
     2009  14116
     2010  14065
     2011  14071
     2012  13410
     2013  13650
     2014  10973
     2015    562

dtype: int64
```

```
df['Season'].value_counts()
```

```
              count
    Season
    Kharif    95951
      Rabi    66987
Whole Year    57305
    Summer    14841
    Winter     6058
    Autumn     4949

dtype: int64
```

```
df['Crop'].value_counts()
```

Crop	count
Rice	15104
Maize	13947
Moong(Green Gram)	10318
Urad	9850
Sesamum	9046
...	...
Litchi	6
Coffee	6
Apple	4
Peach	4
Other Dry Fruit	1

124 rows × 1 columns

dtype: int64

```
df.head()
```

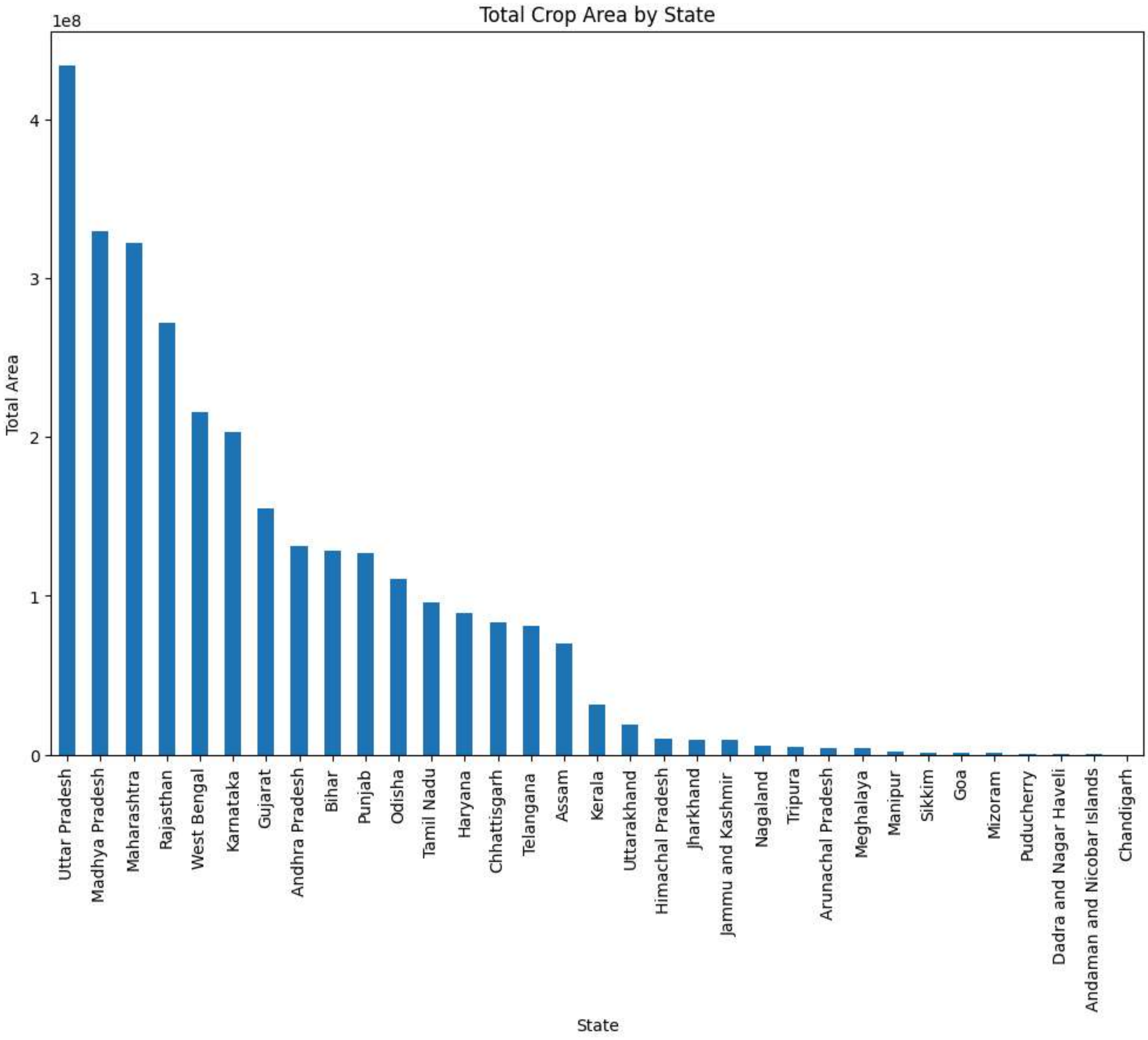
	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254.0	2000.0
1	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2.0	1.0
2	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102.0	321.0
3	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Banana	176.0	641.0
4	Andaman and Nicobar Islands	NICOBARS	2000	Whole Year	Cashewnut	720.0	165.0

```
state_area = df.groupby('State_Name')['Area'].sum().sort_values(ascending=False)
state_area
```

	Area
State_Name	
Uttar Pradesh	4.336316e+08
Madhya Pradesh	3.298131e+08
Maharashtra	3.222062e+08
Rajasthan	2.720249e+08
West Bengal	2.154052e+08
Karnataka	2.029101e+08
Gujarat	1.549440e+08
Andhra Pradesh	1.315458e+08
Bihar	1.282720e+08
Punjab	1.267256e+08
Odisha	1.105336e+08
Tamil Nadu	9.589787e+07
Haryana	8.959731e+07
Chhattisgarh	8.303966e+07
Telangana	8.136062e+07
Assam	7.037876e+07
Kerala	3.190807e+07
Uttarakhand	1.879318e+07
Himachal Pradesh	1.000388e+07
Jharkhand	9.391046e+06
Jammu and Kashmir	9.264623e+06
Nagaland	6.070974e+06
Tripura	4.641609e+06
Arunachal Pradesh	4.364346e+06
Meghalaya	4.035028e+06
Manipur	2.007264e+06
Sikkim	1.524479e+06
Goa	1.205680e+06
Mizoram	9.937352e+05
Puducherry	5.487420e+05
Dadra and Nagar Haveli	3.965150e+05
Andaman and Nicobar Islands	3.378961e+05
Chandigarh	1.252200e+04

dtype: float64

```
# Plotting the area values by state
plt.figure(figsize=(12, 8))
state_area.plot(kind='bar')
plt.title('Total Crop Area by State')
plt.xlabel('State')
plt.ylabel('Total Area')
plt.xticks(rotation=90)
plt.show()
```



From above we can conclude that :

1)Uttar Pradesh is the state with highest agricultural land.

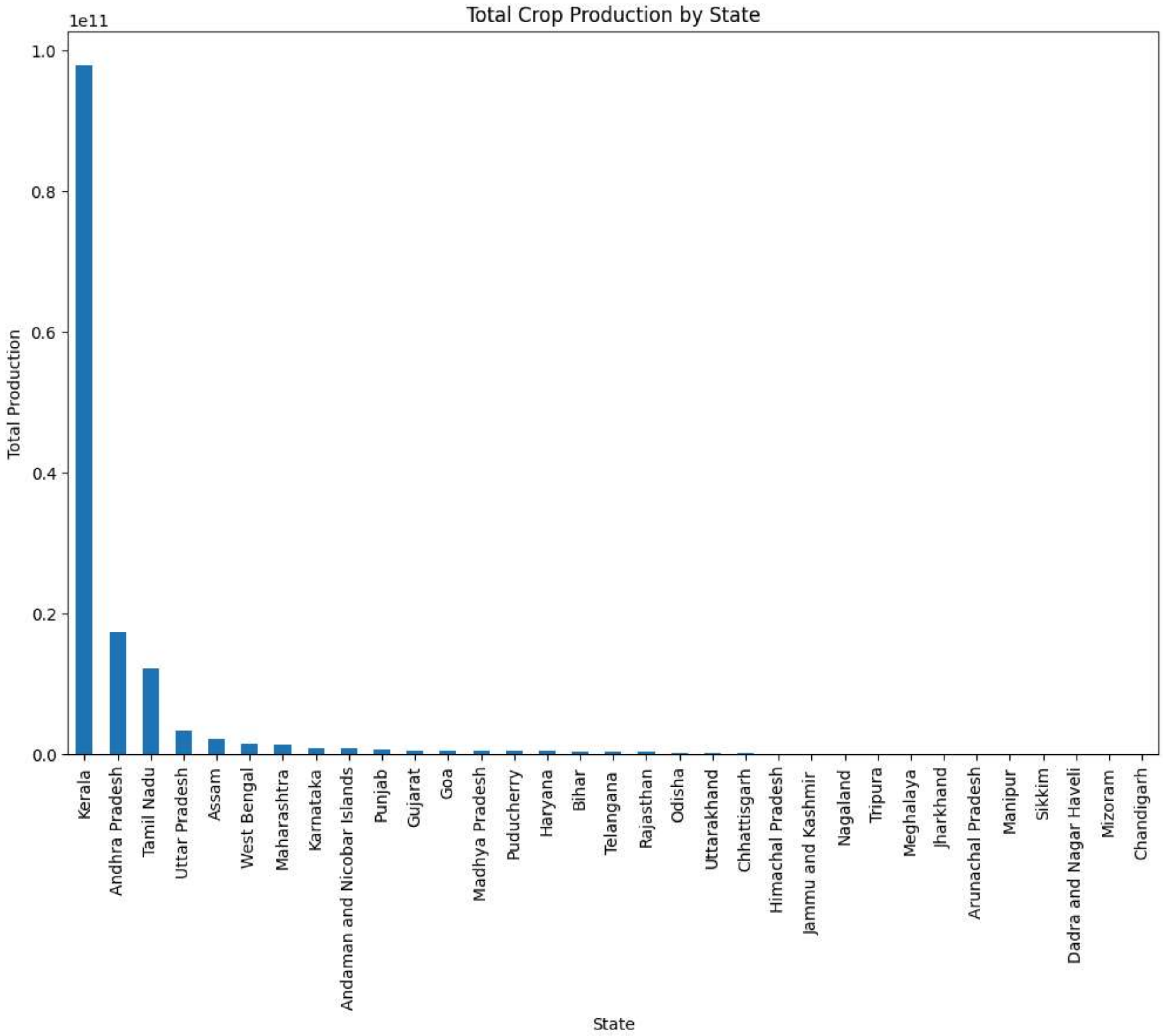
2)Chandigarh is the state with lowest agricultural land.

```
##Statewise Crop Production
state_production = df.groupby('State_Name')['Production'].sum().sort_values(ascending=False)
state_production
```

	Production
State_Name	
Kerala	9.788005e+10
Andhra Pradesh	1.732459e+10
Tamil Nadu	1.207644e+10
Uttar Pradesh	3.234493e+09
Assam	2.111752e+09
West Bengal	1.397904e+09
Maharashtra	1.263641e+09
Karnataka	8.634298e+08
Andaman and Nicobar Islands	7.182232e+08
Punjab	5.863850e+08
Gujarat	5.242913e+08
Goa	5.057558e+08
Madhya Pradesh	4.488407e+08
Puducherry	3.847245e+08
Haryana	3.812739e+08
Bihar	3.664836e+08
Telangana	3.351479e+08
Rajasthan	2.813203e+08
Odisha	1.609041e+08
Uttarakhand	1.321774e+08
Chhattisgarh	1.009519e+08
Himachal Pradesh	1.780517e+07
Jammu and Kashmir	1.329102e+07
Nagaland	1.276595e+07
Tripura	1.252292e+07
Meghalaya	1.211250e+07
Jharkhand	1.077774e+07
Arunachal Pradesh	6.823913e+06
Manipur	5.230917e+06
Sikkim	2.435735e+06
Dadra and Nagar Haveli	1.847871e+06
Mizoram	1.661540e+06
Chandigarh	6.395650e+04

dtype: float64

```
plt.figure(figsize=(12, 8))
state_production.plot(kind='bar')
plt.title('Total Crop Production by State')
plt.xlabel('State')
plt.ylabel('Total Production')
plt.xticks(rotation=90)
plt.show()
```



From above we can conclude that :

1) Kerala has highest production although not in top among agricultural area.

2) Chandigarh has lowest production but in sink with its rank in agricultural area.

```
df.groupby('Crop')['Area'].sum().sort_values(ascending=False)
```

```
                      Area
           Crop
           Rice  7.471253e+08
          Wheat  4.707136e+08
   Cotton(lint)  1.565681e+08
          Bajra  1.411408e+08
          Jowar  1.377159e+08
            ...           ...
            Ber  1.180000e+02
          Peach  4.200000e+01
         Litchi  2.500000e+01
          Apple  9.000000e+00
Other Dry Fruit  7.000000e+00
124 rows × 1 columns

dtype: float64
```

From above we can conclude that :

1)Rice crop has seen highest plantation.

2) Other Dry Fruits has lowest plantation according to area.

```
##Cropwise production
df.groupby('Crop')['Production'].sum().sort_values(ascending=False)
```

```
                    Production
             Crop
          Coconut  1.299816e+11
        Sugarcane  5.535682e+09
             Rice  1.605470e+09
            Wheat  1.332826e+09
           Potato  4.248263e+08
              ...           ...
Other Citrus Fruit  0.000000e+00
         Cucumber  0.000000e+00
           Litchi  0.000000e+00
          Lab-Lab  0.000000e+00
            Apple  0.000000e+00
124 rows × 1 columns

dtype: float64
```

From above we can conclude that :

1)Coconunt has highest production among all the crops.

2) There are many crops which has been planted but the production is almost zero.

3) The rice which has seen highest plantation is not the top production.

```
df.head()
```

```
                   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
```

```
df.groupby('Crop_Year')['Area'].sum()
```

```
                   Area
Crop_Year
     1997  2.317150e+08
     1998  1.669881e+08
     1999  1.586661e+08
     2000  1.652975e+08
     2001  1.652956e+08
     2002  1.577690e+08
     2003  1.720881e+08
     2004  1.678784e+08
     2005  1.631364e+08
     2006  1.706991e+08
     2007  1.527242e+08
     2008  1.712321e+08
     2009  1.656947e+08
     2010  1.766192e+08
     2011  1.536292e+08
     2012  1.524698e+08
     2013  1.415249e+08
     2014  1.157575e+08
     2015  4.601298e+06

dtype: float64
```

From above we can conclude that :

1) In 1997 the agriculture area is on top.

2) As the years are passing we can see a general trend that the crop area is reducing.

```
df.groupby('Crop_Year')['Production'].sum()
```

```
             Production
Crop_Year
   1997    8.512329e+08
   1998    5.825321e+09
   1999    6.434666e+09
   2000    7.449709e+09
   2001    7.465541e+09
   2002    7.696955e+09
   2003    7.917974e+09
   2004    8.189462e+09
   2005    8.043757e+09
   2006    8.681913e+09
   2007    6.879442e+09
   2008    7.717018e+09
   2009    7.660494e+09
   2010    6.307609e+09
   2011    1.430890e+10
   2012    8.171055e+09
   2013    1.290359e+10
   2014    8.664541e+09
   2015    6.935065e+06

dtype: float64
```

From above we can conclude that :

1) Year 2011 has highest production among all the years.

2) Also we can see that though the agriculture area is reducing the production is increasing year after year.

```
df.groupby('Season')['Area'].sum().sort_values(ascending=False)
```

```
                    Area
   Season
   Kharif     1.404845e+09
    Rabi      9.479874e+08
Whole Year    2.573005e+08
   Winter     2.195979e+08
   Summer     7.598406e+07
   Autumn     4.807113e+07

dtype: float64
```

From above we can conclude that :

1) In Kharif season more area is under plantation.

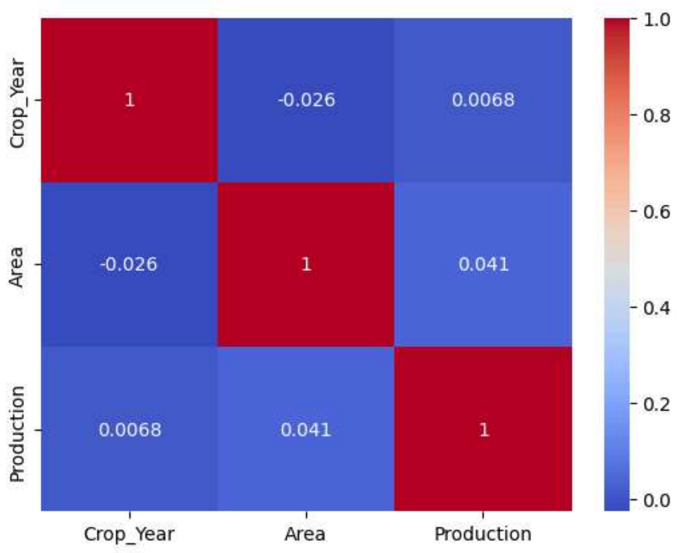
```
df.groupby('Season')['Production'].sum().sort_values(ascending=False)
```

```
              Production
      Season
  Whole Year  1.344248e+11
      Kharif  4.029970e+09
        Rabi  2.051688e+09
      Winter  4.345498e+08
      Summer  1.706579e+08
      Autumn  6.441377e+07

dtype: float64
```

```
# Select only numeric columns for correlation
numeric_df = df.select_dtypes(include=['number'])

# Compute correlation matrix
correlations = numeric_df.corr()
sns.heatmap(correlations, annot=True, cmap='coolwarm')
plt.show()
```



```
X = df.drop(columns=['Production'])
y = df['Production']
```

```
X.head()
```

	State_Name	District_Name	Crop_Year	Season	Crop	Area
0	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Arecanut	1254.0
1	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Other Kharif pulses	2.0
2	Andaman and Nicobar Islands	NICOBARS	2000	Kharif	Rice	102.0

```
y.head()
```

	Production
0	2000.0
1	1.0
2	321.0
3	641.0
4	165.0

dtype: float64

```
X.shape, y.shape
```

```
((246091, 6), (246091,))
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
```

Could not connect to the reCAPTCHA service. Please check your internet connection and reload to get a reCAPTCHA challenge.