In [2]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

Reading the given dataset

In [3]:

```
crop_ds=pd.read_csv(r"G:\2. Python for data science\Datasets\apy_2.csv")
```

Checking for Missing values

In [4]:

```
crop_ds.isna().sum()
```

Out[4]:

```
Unnamed: 0          0
State_Name          0
District_Name       0
Crop_Year           0
Season              0
Crop                0
Area                0
Production       1148
dtype: int64
```

Dropping the rows with missing values

In [5]:

```
crop_ds.dropna(inplace=True)
```

In [6]:

```
crop_ds.isna().sum()
```

Out[6]:

```
Unnamed: 0       0
State_Name       0
District_Name    0
Crop_Year        0
Season           0
Crop             0
Area             0
Production       0
dtype: int64
```

In [7]:

```
crop_ds.head()
```

Out[7]:

	Unnamed: 0	State_Name	District_Name	Crop_Year	Season	Crop	Area	Production
0	0	Karnataka	UDUPI	2005	Rabi	Horse-gram	1122.0	836.0
1	1	Madhya Pradesh	GWALIOR	2003	Whole Year	Brinjal	194.0	0.0
2	2	Andhra Pradesh	CHITTOOR	2010	Rabi	Sesamum	334.0	118.0
3	3	Andhra Pradesh	KRISHNA	2014	Rabi	Tomato	538.0	7289.0
4	4	Uttar Pradesh	SULTANPUR	2011	Rabi	Coriander	59.0	33.0

In [8]:

```
crop_ds.tail()
```

Out[8]:

	Unnamed: 0	State_Name	District_Name	Crop_Year	Season	Crop	Area	Produ
73822	73822	Uttar Pradesh	SULTANPUR	2001	Rabi	Rapeseed &Mustard	3727.0	3
73823	73823	Chhattisgarh	DURG	2014	Rabi	Wheat	6364.0	7
73824	73824	Uttar Pradesh	RAE BARELI	2010	Summer	Moong(Green Gram)	489.0	
73825	73825	Assam	KAMRUP	1998	Rabi	Wheat	6431.0	6
73826	73826	Andhra Pradesh	GUNTUR	2005	Rabi	Urad	100125.0	78

Question no. 1

Which of these crops are produced during the Summer season? (select option with all that apply)

a. Arecanut, Arhar/Tur, Bajra, Castor seed

b. Paddy, Maize, Moong (Green Gram), Onion, Sunflower

c. Banana, Coriander, Gram, Rapeseed & Mustard

d. Rice, Sugarcane, Paddy, Tomato

In [9]:

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

Out[9]:

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

In [10]:

```
crop_ds[crop_ds['Season']=='Summer     ']['Crop'].unique()
```

Out[10]:

```
array(['Groundnut', 'Maize', 'Rice', 'Sesamum', 'Wheat', 'Onion', 'Urad',
       'Peas & beans (Pulses)', 'Sunflower', 'Ragi', 'Bajra',
       'Moong(Green Gram)', 'Cotton(lint)', 'Jowar', 'Cowpea(Lobia)',
       'Tobacco', 'Horse-gram', 'Dry chillies', 'Paddy', 'Turmeric',
       'Arhar/Tur', 'Banana', 'Potato', 'Dry ginger', 'Brinjal',
       'Sugarcane', 'Other  Rabi pulses', 'Small millets',
       'Total foodgrain'], dtype=object)
```

Answer: Option b

Question no. 2

During which year did Haryana have the highest crop production?

a. 2013

b. 2011

c. 1997

d. 2008

In [11]:

```
q2=crop_ds[['State_Name','Crop_Year','Production']].groupby(['State_Name','Crop_Yea
```

In [12]:

```
q2.loc['Haryana'].sort_values(by='Production', ascending=False)
```

```
Out[12]:            Production
           Crop_Year
                2008  9647100.0
                2005  9328300.0
                2011  9021300.0
                2003  8437562.0
                2012  7996416.0
                2002  7837700.0
                2004  7523400.0
                2007  7323500.0
                1999  7188901.0
                2001  7009900.0
                1998  6653600.0
                2006  6566600.0
                2000  5872800.0
                2010  5548900.0
                2009  5214480.0
                1997  3376800.0
```

Answer: Option d

Question no. 3

The maximum and minimum area for production were in the years?

a. 1997 and 2014

b. 1998 and 2015

c. 1997 and 2015

d. 1999 and 2005

```
In [13]: q3=crop_ds[['Crop_Year','Area']].groupby(['Crop_Year']).sum()
```

```
In [14]: q3.sort_values(by='Area',ascending=False)
```

Out[14]:

	Area
Crop_Year	
1997	68245111.00
2004	53445367.24
1999	51841049.00
2006	51393712.02
2010	50813109.78
2009	50772638.00
2000	50530198.00
2008	49604477.00
1998	49317287.00
2003	48954140.97
2002	48328531.15
2011	45835655.41
2005	45296672.88
2007	44971304.38
2012	44369268.00
2001	43603438.77
2013	41895685.00
2014	34188857.84
2015	1313314.00

Answer: Option c

Question no. 4

Which state in India had the second lowest crop production? (overall, for all years)

a. Meghalaya

b. Chandigarh

c. Mizoram

d. Manipur

In [15]:

```
q4=crop_ds[['State_Name','Production']].groupby(['State_Name']).sum()
```

In [16]:

```
q4.sort_values(by='Production',ascending=True)
```

```
Out[16]:                               Production
                        State_Name
                        Chandigarh    1.580450e+04
                           Mizoram    4.579428e+05
                            Sikkim    5.507250e+05
            Dadra and Nagar Haveli    6.247060e+05
                           Manipur    1.658617e+06
                 Arunachal Pradesh    2.035912e+06
                         Jharkhand    3.319141e+06
                         Meghalaya    3.639914e+06
                          Nagaland    3.925012e+06
                 Jammu and Kashmir    4.018134e+06
                           Tripura    4.278173e+06
                  Himachal Pradesh    5.198802e+06
                      Chhattisgarh    3.132986e+07
                       Uttarakhand    4.017398e+07
                            Odisha    4.563306e+07
                        Puducherry    7.326832e+07
                         Rajasthan    8.671658e+07
                           Haryana    1.145473e+08
                             Bihar    1.151581e+08
                         Telangana    1.202722e+08
                    Madhya Pradesh    1.337226e+08
                               Goa    1.421826e+08
                           Gujarat    1.538659e+08
                            Punjab    1.724376e+08
       Andaman and Nicobar Islands    2.032759e+08
                         Karnataka    2.637879e+08
                       Maharashtra    3.930627e+08
                             Assam    5.201106e+08
                       West Bengal    5.327023e+08
                     Uttar Pradesh    1.069989e+09
                        Tamil Nadu    1.787126e+09
                    Andhra Pradesh    3.141848e+09
```

	Production
State_Name	
Kerala	2.299779e+10

Answer: Option c

Question no. 5

What were the top three produced crops in the year 2012?

a. Wheat, Potato, Rice

b. Coconut, Potato, Sugarcane

c. Coconut, Sugarcane, Rice

d. Rice, Sugarcane, Maize

```
In [17]: q5=crop_ds[['Crop_Year','Crop','Production']].groupby(['Crop_Year','Crop']).sum()
```

```
In [18]: q5.loc[2012].sort_values(by='Production',ascending=False)
```

Out[18]:

	Production
Crop	
Coconut	1.208299e+09
Sugarcane	1.143474e+08
Rice	3.371318e+07
Wheat	2.448045e+07
Potato	1.006466e+07
...	...
Pome Granet	8.720000e+02
Blackgram	7.000000e+01
Grapes	1.800000e+01
Cardamom	1.200000e+01
other oilseeds	1.000000e+00

69 rows × 1 columns

Answer: Option c

Question no. 6

What is the standard deviation for Area of production?

a. 52957.44 (approx.)

b. 12167.42 (approx.)

c. 49177.60 (approx.)

d. 48848.27 (approx.)

```
In [19]: crop_ds['Area'].std()
```

```
Out[19]: 49177.60312712377
```

Answer: Option c

Question no. 7

Which is the crop that gave the highest production to the state of Andhra Pradesh?

a. Sugarcane

b. Wheat

c. Banana

d. Coconut

```
In [20]: q7=crop_ds[['State_Name','Crop','Production']].groupby(['State_Name','Crop']).sum()
```

```
In [21]: q7.loc['Andhra Pradesh'].sort_values(by="Production",ascending=False)
```

Out[21]:

	Production
Crop	
Coconut	2.979218e+09
Sugarcane	7.585185e+07
Rice	4.599650e+07
Groundnut	5.430417e+06
Maize	4.853811e+06
...	...
Cucumber	0.000000e+00
Bottle Gourd	0.000000e+00
Other Vegetables	0.000000e+00
Peas (vegetable)	0.000000e+00
other fibres	0.000000e+00

67 rows × 1 columns

Answer: Option d

Question no. 8

Which of the following statements is true? (Select all that applies)

a. The overall production during the Kharif season is 2,029,970,000 (approx.)

b. The overall production during the Summer season is 51,992,900 (approx.)

c. The overall production during the Autumn season is 14,413,770 (approx.)

d. The overall production during the Kharif season is 1,282,056,700 (approx.)

```
In [22]: crop_ds[crop_ds['Season']=='Kharif     ']['Production'].sum()
```

```
Out[22]: 1282056680.69
```

```
In [23]: crop_ds[crop_ds['Season']=='Summer     ']['Production'].sum()
```

```
Out[23]: 51992876.699999996
```

```
In [24]: crop_ds[crop_ds['Season']=='Autumn     ']['Production'].sum()
```

```
Out[24]: 18896594.060000002
```

Answer: Options b & d

Question no. 9

Which state has the lowest area of production?

a. Puducherry

b. Chandigarh

c. Kerala

d. Goa

```
In [25]: q9=crop_ds[['State_Name','Area']].groupby(['State_Name']).sum()
```

```
In [26]: q9.sort_values(by='Area',ascending=True)
```

Out[26]:

	Area
State_Name	
Chandigarh	2.791000e+03
Andaman and Nicobar Islands	8.531894e+04
Dadra and Nagar Haveli	1.182820e+05
Puducherry	1.206340e+05
Goa	2.702040e+05
Mizoram	2.711313e+05
Sikkim	3.922050e+05
Manipur	6.121800e+05
Meghalaya	1.270237e+06
Arunachal Pradesh	1.283609e+06
Tripura	1.605874e+06
Nagaland	1.800631e+06
Jammu and Kashmir	2.744164e+06
Jharkhand	2.748473e+06
Himachal Pradesh	2.897344e+06
Uttarakhand	5.714422e+06
Kerala	8.871677e+06
Assam	2.065623e+07
Telangana	2.225371e+07
Chhattisgarh	2.547540e+07
Haryana	2.756965e+07
Tamil Nadu	2.922639e+07
Odisha	3.132268e+07
Punjab	3.692255e+07
Andhra Pradesh	3.864198e+07
Bihar	3.881541e+07
Gujarat	4.835700e+07
Karnataka	5.735193e+07
West Bengal	6.637353e+07
Rajasthan	8.258156e+07
Maharashtra	9.356024e+07
Madhya Pradesh	9.968280e+07

	Area
State_Name	
Uttar Pradesh	1.251196e+08

Answer: Option b

Question no. 10

What is the mean for the area of production?

a. 17065.81 (approx.)

b. 12035.39 (approx.)

c. 11868.49 (approx.)

d. 58250.34 (approx.)

```
In [27]: crop_ds['Area'].mean()
```

```
Out[27]: 12035.385977242393
```

Answer: Option b

Question no. 11

What is the correlation coefficient between Area and Production?

a. 37.686

b. 0.37686

c. 3.7686

d. 0.037686

```
In [28]: crop_ds.corr()
```

```
C:\Users\DELL DESKTOP\AppData\Local\Temp\ipykernel_12996\3772326663.py:1: FutureWa
rning: The default value of numeric_only in DataFrame.corr is deprecated. In a fut
ure version, it will default to False. Select only valid columns or specify the va
lue of numeric_only to silence this warning.
  crop_ds.corr()
```

```
Out[28]:
              Unnamed: 0  Crop_Year      Area  Production
 Unnamed: 0     1.000000  -0.007143  0.002912    0.002058
  Crop_Year    -0.007143   1.000000 -0.025927    0.005928
       Area     0.002912  -0.025927  1.000000    0.037686
 Production     0.002058   0.005928  0.037686    1.000000
```

Answer: Option d

Question no. 12

The crops that had the highest production (in the correct order) were?

a. Coconut, Sugarcane, Cucumber, Potato, Rice

b. Gram, Jute, Soya bean, Maize, Cotton

c. Coconut, Sugarcane, Rice, Wheat, Potato

d. Sugarcane, Wheat, Soya bean, Potato, Coconut

```
In [29]: q12=crop_ds[['Crop','Production']].groupby(['Crop']).sum()
```

```
In [30]: q12.sort_values(by='Production',ascending=False)
```

```
Out[30]:
                      Production
                Crop
             Coconut  2.870917e+10
           Sugarcane  1.769219e+09
                Rice  4.744665e+08
               Wheat  3.849682e+08
              Potato  1.250747e+08
                 ...           ...
                 Ber  0.000000e+00
            Cucumber  0.000000e+00
            Pump Kin  0.000000e+00
  Other Citrus Fruit  0.000000e+00
               Apple  0.000000e+00

122 rows × 1 columns
```

Answer: Option c

Question no. 13

Which is the only crop that has the highest production during autumn, summer, and winter?

a. Rice

b. Maize

c. Paddy

d. Jute

```
In [31]: q13=crop_ds[['Season','Crop','Production']].groupby(['Season','Crop']).sum()
```

```
In [32]: q13.loc['Autumn     '].sort_values(by='Production',ascending=False)
```

```
Out[32]:
                          Production
                     Crop
                     Rice  14820963.70
                    Maize   3089100.78
                    Paddy    521140.00
                     Jute    155210.90
                     Ragi    121504.38
                Groundnut     97968.50
                     Urad     39608.00
        Moong(Green Gram)     24799.50
             Dry chillies     10513.00
                  Sesamum      5302.10
                   Banana      3980.00
                Arhar/Tur      3937.00
                Sugarcane      1291.20
            Small millets       595.00
               Dry ginger       360.00
                  Tapioca       130.00
                   Potato       100.00
     Peas & beans (Pulses)       50.00
                    Onion        20.00
                 Turmeric        20.00
```

```
In [34]: q13.loc['Summer     '].sort_values(by='Production',ascending=False)
```

```
Out[34]:
                       Production
                 Crop
                 Rice  38709052.2
                Maize   4840564.0
            Groundnut   2083123.3
                Bajra   1779222.0
                Paddy   1638436.0
    Moong(Green Gram)    735738.3
              Sesamum    618864.8
                Onion    596379.0
               Banana    272080.0
            Sunflower    170886.0
               Potato    168904.0
                 Urad    107210.8
                 Ragi    103252.8
                Jowar     59373.0
          Dry chillies    54145.0
                Wheat     17009.9
          Cotton(lint)    10622.0
 Peas & beans (Pulses)    10066.0
              Tobacco      5318.0
            Sugarcane      4570.0
            Arhar/Tur      3198.0
        Cowpea(Lobia)      2511.0
           Dry ginger       860.0
       Total foodgrain      617.0
              Brinjal       552.0
           Horse-gram       287.6
             Turmeric        20.0
         Small millets       11.0
     Other Rabi pulses        3.0
```

```
In [35]: q13.loc['Winter     '].sort_values(by='Production',ascending=False)
```

Out[35]:

	Production
Crop	
Rice	1.113293e+08
Potato	8.277893e+06
Sugarcane	4.839051e+06
Paddy	4.622613e+06
Horse-gram	8.732610e+04
Ragi	8.064870e+04
Urad	7.479070e+04
Moong(Green Gram)	6.656960e+04
Sesamum	3.270780e+04
Arhar/Tur	1.260100e+04
Groundnut	1.229040e+04
Rapeseed &Mustard	9.776100e+03
Banana	8.770000e+03
Maize	3.129800e+03
Wheat	1.652000e+03
Dry chillies	7.000000e+02
Gram	5.740000e+02
Niger seed	3.015000e+02
Peas & beans (Pulses)	1.100000e+02
Sannhamp	7.300000e+01
Dry ginger	4.000000e+01
Sweet potato	3.000000e+01
Onion	2.000000e+01

Answer: Option a

Prepare the dataset further by following the steps given below:

```
●Ensure the datatypes of the columns are appropriate
●Drop all the variables except “Area” and “Production”
●Split the data into the train (70%) and test (30%) sets, and set
the random state for the train-test split instance as 42
```

Build a linear regression model using the training dataset by having “Area” as the independent variable and “Production” as the dependent variable. Using the model that has been built, answer the following question

Question no. 14

The Root mean square value of the Linear regression model is

a. 13850999.74575 (approx)

b. 1001531.33109 (approx)

c. 13524820.12533 (approx)

d. 14599645.26554 (approx)

```
In [36]: ml=crop_ds.drop(['Unnamed: 0','State_Name','District_Name','Crop_Year','Season','Cr
```

```
In [37]: ml.head()
```

Out[37]:

	Area	Production
0	1122.0	836.0
1	194.0	0.0
2	334.0	118.0
3	538.0	7289.0
4	59.0	33.0

```
In [38]: x=pd.DataFrame(ml['Area'])
```

```
In [39]: y=pd.DataFrame(ml['Production'])
```

```
In [40]: from sklearn.model_selection import train_test_split
```

```
In [41]: x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.3,random_state=42)
```

```
In [42]: from sklearn.linear_model import LinearRegression
```

```
In [43]: model=LinearRegression()
```

```
In [44]: model.fit(x_train,y_train)
```

```
Out[44]: ▾ LinearRegression
         LinearRegression()
```

In [45]:

```
prediction=model.predict(x_test)
```

In [47]:

```
from sklearn.metrics import mean_squared_error
import numpy as np
```

In [48]:

```
np.sqrt(mean_squared_error(y_test,prediction))
```

Out[48]:

```
13850999.745759705
```

Answer: Option a

Question no. 15

The MAE of the Linear regression model is

a. 18529629.51147 (approx)

b. 827676.37303 (approx)

c. 13524820.12533 (approx)

d. 112599645.26554 (approx)

In [49]:

```
from sklearn.metrics import mean_absolute_error
```

In [50]:

```
mean_absolute_error(y_test,prediction)
```

Out[50]:

```
827676.3730329004
```

Answer: Option b

----Thank You -----

In [ ]:

```

```