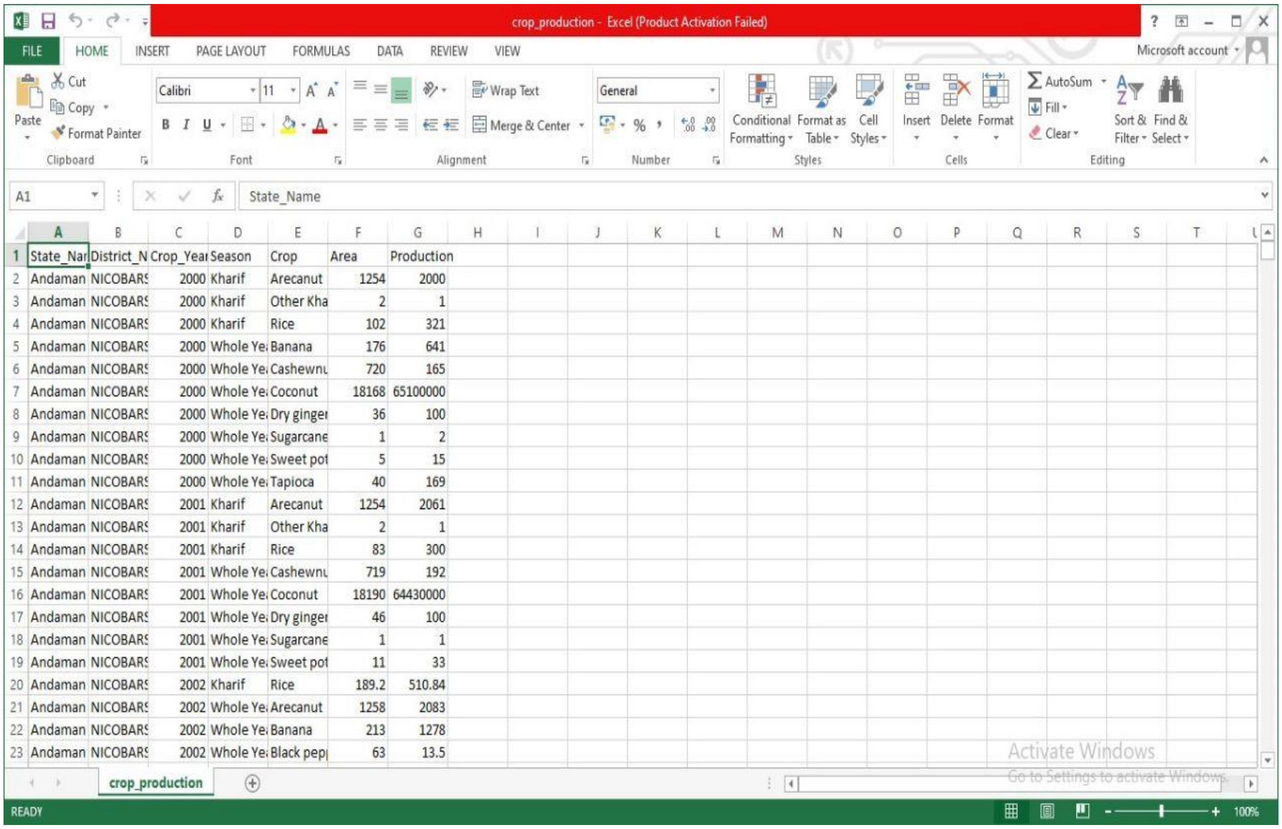
WORKING WITH THE DATASET

Date	28 OCTOBER 2022
Team ID	PNT2022TMID25946
Project Name	ESTIMATION OF CROP YIELD PREDICTION USING DATA ANALYTICS

LOADING AND UNDERSTANDING THE DATASET:

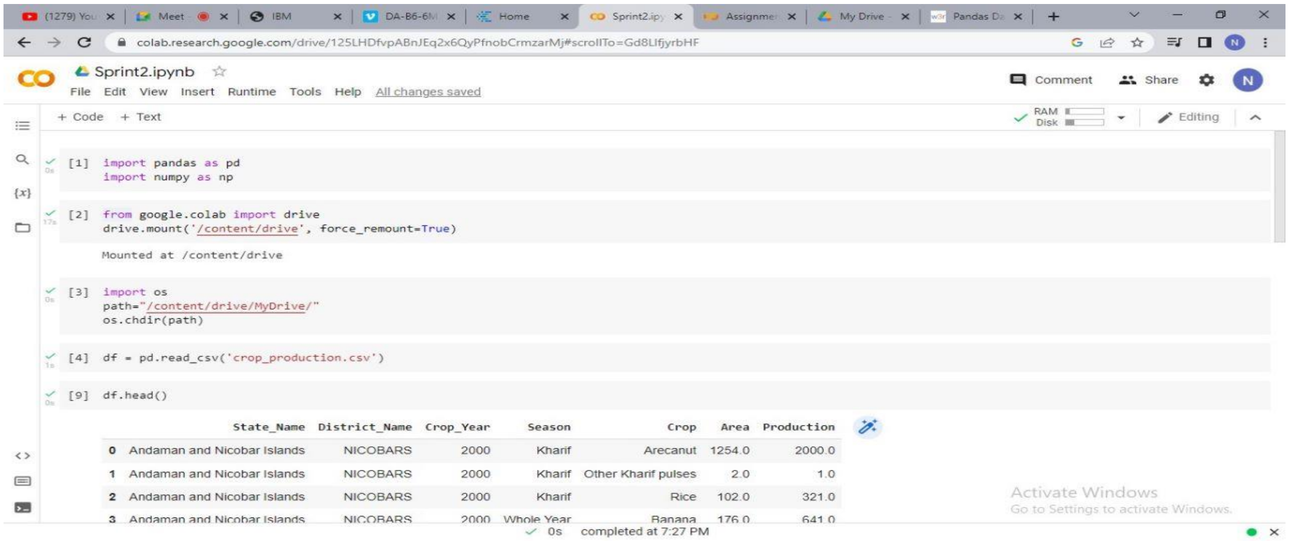
The dataset is downloaded from the Kaggle website with reference to the hyperlink provided in the project flow.



	A	B	C	D	E	F	G
1	State_Nar	District_N	Crop_Year	Season	Crop	Area	Production
2	Andaman	NICOBARS	2000	Kharif	Arecanut	1254	2000
3	Andaman	NICOBARS	2000	Kharif	Other Kha	2	1
4	Andaman	NICOBARS	2000	Kharif	Rice	102	321
5	Andaman	NICOBARS	2000	Whole Ye	Banana	176	641
6	Andaman	NICOBARS	2000	Whole Ye	Cashewnu	720	165
7	Andaman	NICOBARS	2000	Whole Ye	Coconut	18168	65100000
8	Andaman	NICOBARS	2000	Whole Ye	Dry ginger	36	100
9	Andaman	NICOBARS	2000	Whole Ye	Sugarcane	1	2
10	Andaman	NICOBARS	2000	Whole Ye	Sweet pot	5	15
11	Andaman	NICOBARS	2000	Whole Ye	Tapioca	40	169
12	Andaman	NICOBARS	2001	Kharif	Arecanut	1254	2061
13	Andaman	NICOBARS	2001	Kharif	Other Kha	2	1
14	Andaman	NICOBARS	2001	Kharif	Rice	83	300
15	Andaman	NICOBARS	2001	Whole Ye	Cashewnu	719	192
16	Andaman	NICOBARS	2001	Whole Ye	Coconut	18190	64430000
17	Andaman	NICOBARS	2001	Whole Ye	Dry ginger	46	100
18	Andaman	NICOBARS	2001	Whole Ye	Sugarcane	1	1
19	Andaman	NICOBARS	2001	Whole Ye	Sweet pot	11	33
20	Andaman	NICOBARS	2002	Kharif	Rice	189.2	510.84
21	Andaman	NICOBARS	2002	Whole Ye	Arecanut	1258	2083
22	Andaman	NICOBARS	2002	Whole Ye	Banana	213	1278
23	Andaman	NICOBARS	2002	Whole Ye	Black pep	63	13.5

DESCRIPTION OF THE DATASET:

The given dataset was analyzed to get the knowledge about that dataset. It is done by writing a python code in Google Colab platform.


colab.research.google.com/drive/125LHDfvpABnJEq2x6QyPfnobCrmzarMj#scrollTo=Gd8LlfjyrbHF
Sprint2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Comment
Share
RAM Disk
Editing
Activate Windows
Go to Settings to activate Windows.
0s completed at 7:27 PM

```
[1] import pandas as pd
    import numpy as np
```

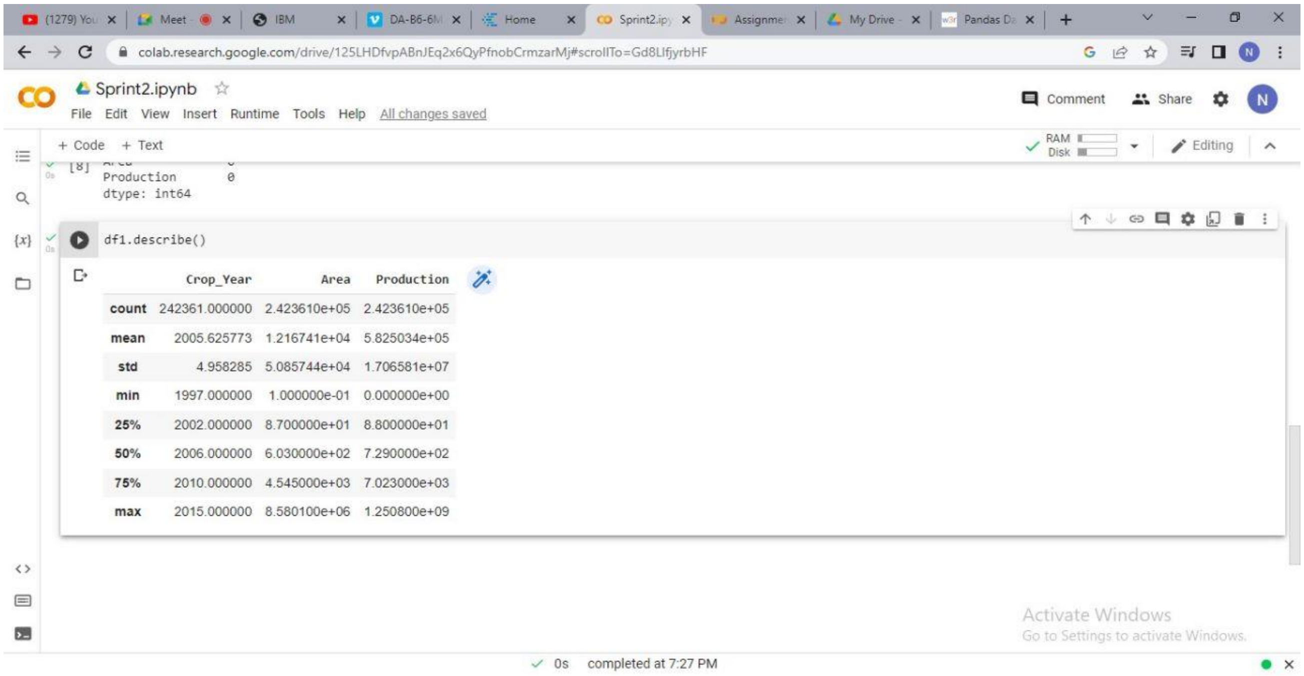
```
[2] from google.colab import drive
    drive.mount('/content/drive', force_remount=True)

    Mounted at /content/drive
```

```
[3] import os
    path="/content/drive/MyDrive/"
    os.chdir(path)
```

```
[4] df = pd.read_csv('crop_production.csv')
```

```
[9] 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


colab.research.google.com/drive/125LHDfvpABnJEq2x6QyPfnobCrmzarMj#scrollTo=Gd8LlfjyrbHF
Sprint2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
+ Code + Text
Comment
Share
RAM Disk
Editing
Activate Windows
Go to Settings to activate Windows.
0s completed at 7:27 PM

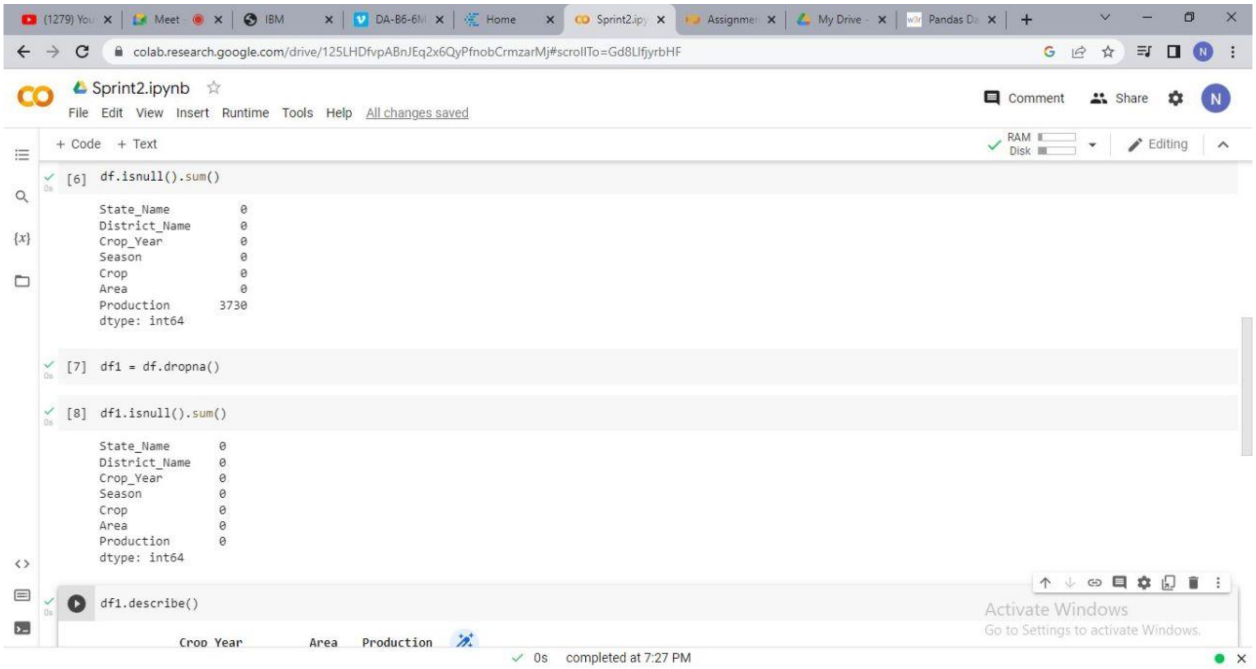
```
[8] Production       0
    dtype: int64
```

```
df1.describe()
```

	Crop_Year	Area	Production
count	242361.000000	2.423610e+05	2.423610e+05
mean	2005.625773	1.216741e+04	5.825034e+05
std	4.958285	5.085744e+04	1.706581e+07
min	1997.000000	1.000000e-01	0.000000e+00
25%	2002.000000	8.700000e+01	8.800000e+01
50%	2006.000000	6.030000e+02	7.290000e+02
75%	2010.000000	4.545000e+03	7.023000e+03
max	2015.000000	8.580100e+06	1.250800e+09

CLEANING THE DATASET:

The dataset may contain null values, so it must cleaned before using and also cleaning process is important to perform accurate visualization.



colab.research.google.com/drive/125LHDfvpABnJEq2x6QyPfnobCrmzarMj#scrollTo=Gd8LlfjyrbHF
Sprint2.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
RAM Disk
Editing
Activate Windows
Go to Settings to activate Windows.

```
[6] df.isnull().sum()

    State_Name         0
    District_Name      0
    Crop_Year          0
    Season             0
    Crop               0
    Area               0
    Production      3730
    dtype: int64

[7] df1 = df.dropna()

[8] df1.isnull().sum()

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

    df1.describe()

             Crop_Year          Area    Production
```

✓ 0s completed at 7:27 PM