BHARATIYA VIDYA BHAVAN'S
SARDAR PATEL INSTITUTE OF TECHNOLOGY
Bhavan's Campus, Munshi Nagar, Andheri (West), Mumbai – 400058-India

Branch:	BE ETRX
Batch:	A
Course:	Minors – Computer Science
Subject:	Machine Learning
Student Name:	Dhananjay Joshi
Experiment No.	01
UID:	2019110021

Aim:
Import the dataset and perform EDA such as number of data samples, number of features, number of classes, number of data samples per class, removing missing values, conversion to numbers, explore dimensionality, type the mean or average value, and using seaborn library to plot different graphs. Consider one of the datasets given below.

1. NASA: If you're interested in space and earth science, see what you can find among the tens of thousands of public datasets made available by NASA.

Software Used:
Google Colab Notebook

Dataset Description: Twentieth Century Crop Statistics 1900-2017

> The Twentieth Century Crop Statistics, 1900-2017 data set consists of national or subnational maize and wheat production, yield, and harvested area statistics for all available years for the period 1900-2017. It combines a new digitization of crop statistics from Italy, Spain, Indonesia, China, Mexico, Uruguay, Chile, Sweden, and Morocco with existing, publicly available, digitized data sets from India, Australia, the United States, Canada, Southern Brazil, Argentina, England, Austria, Belgium, Croatia, Czech Republic, Finland, Germany, Spain, Portugal, France, the Netherlands, and South Africa. All Units are converted to hectares (ha) for Units of harvested areas, tonnes for Units of production, and tonnes/ha for Units of yield. A ratio of 1/36.744 is used to convert wheat bushels to tonnes, and a value of 1/39.368 is used to convert maize bushels to tonnes. In all cases, the Harvest_year reported in the data set is the harvest year for the crop.

Analysis:

1. Going through the first ten samples:

```
df = pd.read_excel('/content/drive/MyDrive/foodcropstats.xlsx', s
heet_name="CropStats")
df.rename(columns = {'Unnamed: 0':'SrNo'}, inplace = True)
print("First ten samples in the dataset = ")
df.head(10)
```

SrNo	Harvest_year	nation	subnation	crop	hectares (ha)	production (tonnes)	year	yield(tonnes/ha)	admin2	notes
0	1902	Austria	NaN	wheat	NaN	NaN	1902	1.31	NaN	NaN
1	1903	Austria	NaN	wheat	NaN	NaN	1903	1.47	NaN	NaN
2	1904	Austria	NaN	wheat	NaN	NaN	1904	1.27	NaN	NaN
3	1905	Austria	NaN	wheat	NaN	NaN	1905	1.33	NaN	NaN
4	1906	Austria	NaN	wheat	NaN	NaN	1906	1.28	NaN	NaN
5	1907	Austria	NaN	wheat	NaN	NaN	1907	1.37	NaN	NaN
6	1908	Austria	NaN	wheat	NaN	NaN	1908	1.36	NaN	NaN
7	1909	Austria	NaN	wheat	NaN	NaN	1909	1.35	NaN	NaN
8	1910	Austria	NaN	wheat	NaN	NaN	1910	1.18	NaN	NaN
9	1911	Austria	NaN	wheat	NaN	NaN	1911	1.37	NaN	NaN

We can see that the data of yield i.e., production per hectare of wheat in Austria is given. We note that the columns "admin2" and "notes" are not needed to analyse the statistics of food crops. Therefore, we need to drop the two columns.

2. After dropping the two columns:

```
df2 = df.drop(['admin2', 'notes'], axis=1)
```

SrNo	Harvest_year	nation	subnation	crop	hectares (ha)	production (tonnes)	year	yield(tonnes/ha)
0	1902	Austria	NaN	wheat	NaN	NaN	1902	1.310000
1	1903	Austria	NaN	wheat	NaN	NaN	1903	1.470000
2	1904	Austria	NaN	wheat	NaN	NaN	1904	1.270000
3	1905	Austria	NaN	wheat	NaN	NaN	1905	1.330000
4	1906	Austria	NaN	wheat	NaN	NaN	1906	1.280000
...	...	...	...	...	...	...	...	...
36702	2013	China	zhejiang	wheat	75520.0	278300.0	2013	3.685117
36703	2014	China	zhejiang	wheat	82120.0	309500.0	2014	3.768875
36704	2015	China	zhejiang	wheat	89800.0	351300.0	2015	3.912027
36705	2016	China	zhejiang	wheat	76590.0	253900.0	2016	3.315054
36706	2017	China	zhejiang	wheat	103670.0	419200.0	2017	4.043600

```
Total samples in the original dataset =  403777
Total features in the original dataset =  11
Therefore shape of the original dataset =  (36707, 11)
Dimensions of the original dataset =  2


Total samples in the new dataset =  330363
Total features in the new dataset =  9
Therefore shape of the new dataset =  (36707, 9)
Dimensions of the new dataset =  2
```

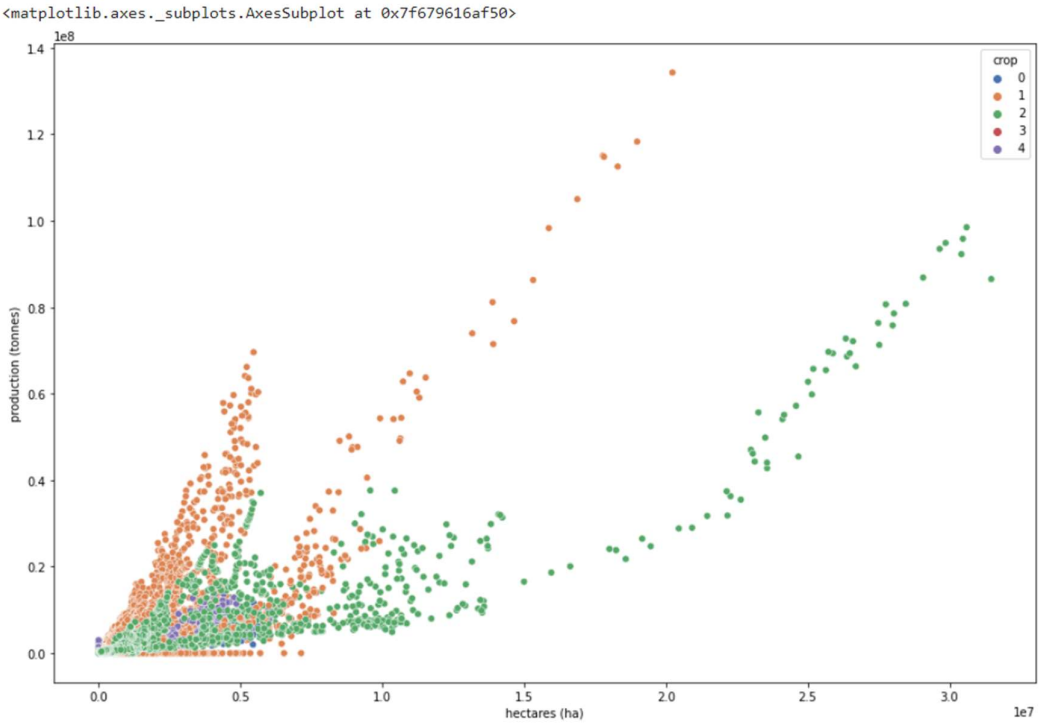
3. After assigning numeric values to the different classes of crops

```
df2['production (tonnes)'] = df2['production (tonnes)'].fillna(0)
df2['subnation'] = df2['subnation'].fillna(0)
df2['hectares (ha)'] = df2['hectares (ha)'].fillna(0)
df2['crop'].replace(['cereals', 'maize', 'wheat', 'spring wheat',
 'winter wheat'], [0, 1, 2, 3, 4], inplace=True)
```

SrNo	Harvest_year	nation	subnation	crop	hectares (ha)	production (tonnes)	year	yield(tonnes/ha)
0	1902	Austria	0	2	0.0	0.0	1902	1.310000
1	1903	Austria	0	2	0.0	0.0	1903	1.470000
2	1904	Austria	0	2	0.0	0.0	1904	1.270000
3	1905	Austria	0	2	0.0	0.0	1905	1.330000
4	1906	Austria	0	2	0.0	0.0	1906	1.280000
...	...	...	...	...	...	...	...	...
36702	2013	China	zhejiang	2	75520.0	278300.0	2013	3.685117
36703	2014	China	zhejiang	2	82120.0	309500.0	2014	3.768875
36704	2015	China	zhejiang	2	89800.0	351300.0	2015	3.912027
36705	2016	China	zhejiang	2	76590.0	253900.0	2016	3.315054
36706	2017	China	zhejiang	2	103670.0	419200.0	2017	4.043600

ws × 9 columns

4. Using Scatterplot to plot production v/s hectares graph – yield as the slope

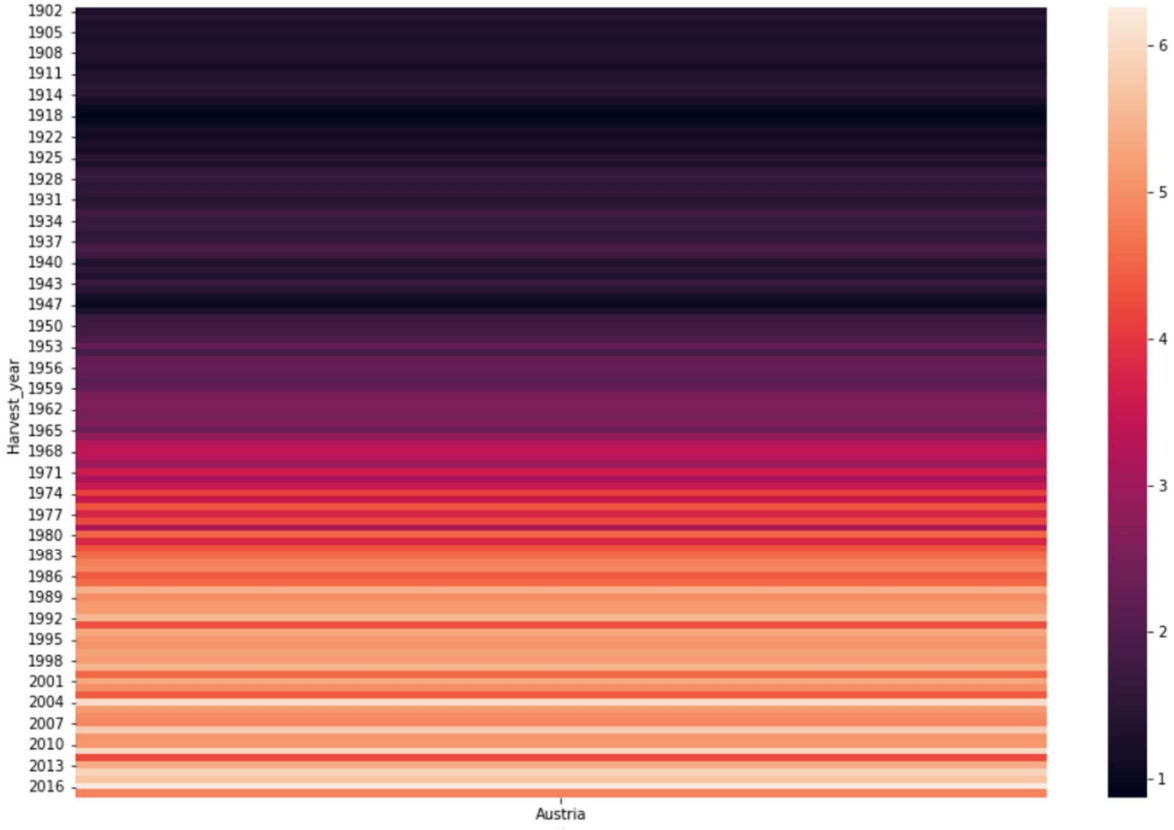


From the scatterplot, we can see the relation of each crop based on the area in hectares and the production in tonnes.

a. First of all, we can infer that the production of maize – 1 and wheat – 2 is lower (0.1 – 0.2 tonnes) when the area used by the nation is between 0 – 0.7 hectares.
b. Secondly, the highest yield is of maize at the value of 0.675 tonnes/hectare.

5. Using a heatmap to understand the yield of wheat in Austria from 1900-2017

```
df_aus = df_aus.dropna()
df_aus = df_aus.reset_index(drop=True)
df2_hmap = df_aus.pivot("Harvest_year","nation","yield(tonnes/ha)")
plt.figure(figsize=(15,10))
ax = sns.heatmap(df2_hmap)
```



We can differentiate between the yield of wheat for Austria for the harvest years and can infer that the range of yield is higher in the years 1948-1965 where the yield is between 2-3 tonnes/ha.

Conclusion:

1. Based on the exploratory data analysis of Crop Statistics from year 1900-2017 we are able to derive inferences from the graphical plots.
2. We are also able to clean the data using the python libraries such as pandas and numpy.
3. We can now use this data to be used with Machine Learning Algorithms for further analysis.