|  |  |
| --- | --- |
| **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](https://nasa.github.io/data-nasa-gov-frontpage/#_blank):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

### Analysis:

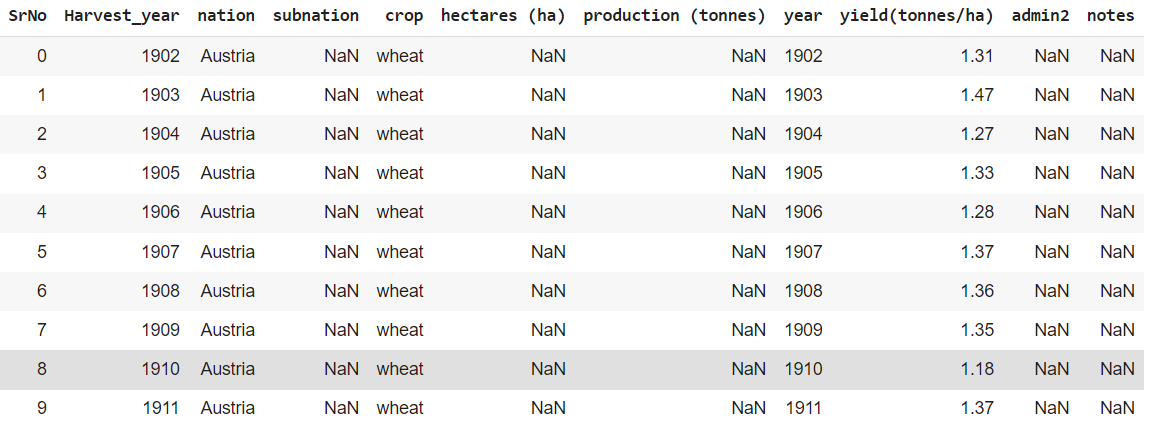
### Going through the first ten samples:

df = pd.read\_excel('/content/drive/MyDrive/foodcropstats.xlsx', sheet\_name="CropStats")

df.rename(columns = {'Unnamed: 0':'SrNo'}, inplace = True)

print("First ten samples in the dataset = ")

df.head(10)



### 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.

### After dropping the two columns:

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

### 

### 

### 

### 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)

### 

### 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.

### 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.

### Secondly, the highest yield is of maize at the value of 0.675 tonnes/hectare.

### 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:

### Based on the exploratory data analysis of Crop Statistics from year 1900-2017 we are able to derive inferences from the graphical plots.

### We are also able to clean the data using the python libraries such as pandas and numpy.

### We can now use this data to be used with Machine Learning Algorithms for further analysis.