

Bangladesh University of Business and Technology

Department of CSE

Assignment

Course Title: Data Mining **Course code**: CSE-476

Submitted By

Ishmoth Ura Nuri

ID: 18192103237

Intake: 41 Section: 03

Submitted To

Khan Md. Hasib

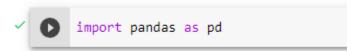
Lecturer

Computer Science & Engineering

Submission Date: 22.02.23

 Apply data preprocessing steps (such as: Viewing your data, Handling duplicates, Column cleanup, DataFrame slicing, selecting, extracting) in the following dataset https://www.kaggle.com/datasets/selinraja/irish-data

1. Import Library



2. Upload the dataset & Viewing the data

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

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).

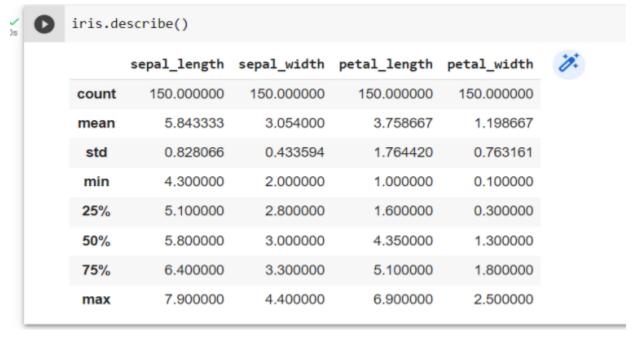
[3] iris = pd.read_csv("/content/drive/MyDrive/Colab Notebooks/Iris_Data.csv")
iris
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

3. View the top 10 rows of the dataset.

os ir	is.head(10)				
	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
5	5.4	3.9	1.7	0.4	Iris-setosa
6	4.6	3.4	1.4	0.3	Iris-setosa
7	5.0	3.4	1.5	0.2	Iris-setosa
8	4.4	2.9	1.4	0.2	Iris-setosa
9	4.9	3.1	1.5	0.1	Iris-setosa

4. Showing the description of the whole dataset.



5. Showing the info of the dataset.

```
iris.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 5 columns):
   Column Non-Null Count Dtype
 0 sepal_length 150 non-null
                               float64
 1 sepal_width 150 non-null float64
 2 petal_length 150 non-null float64
   petal width 150 non-null
                              float64
    species
            150 non-null
                               object
dtypes: float64(4), object(1)
memory usage: 6.0+ KB
```

6. Dropping the duplicate data

0	dienla	ay(iris.drop_d	dunlicates())			
	итэртс	ay(1113.u10p_0	upiicaces())			
		sepal_length	sepal_width	petal_length	petal_width	species
	0	5.1	3.5	1.4	0.2	Iris-setosa
	1	4.9	3.0	1.4	0.2	Iris-setosa
	2	4.7	3.2	1.3	0.2	Iris-setosa
	3	4.6	3.1	1.5	0.2	Iris-setosa
	4	5.0	3.6	1.4	0.2	Iris-setosa
	145	6.7	3.0	5.2	2.3	Iris-virginica
	146	6.3	2.5	5.0	1.9	Iris-virginica
	147	6.5	3.0	5.2	2.0	Iris-virginica
	148	6.2	3.4	5.4	2.3	Iris-virginica
	149	5.9	3.0	5.1	1.8	Iris-virginica

7. Column cleanup

```
[25] for x in iris.index:
           if iris.loc[x, "sepal_length"] > 5:
             iris.loc[x, "sepal_length"] = 5
           iris.head(10)
             sepal_length sepal_width petal_length petal_width
                                                                         species
         0
                       5.0
                                     3.5
                                                     1.4
                                                                   0.2 Iris-setosa
          1
                       4.9
                                     3.0
                                                     1.4
                                                                   0.2 Iris-setosa
         2
                       4.7
                                     3.2
                                                     1.3
                                                                   0.2 Iris-setosa
         3
                       4.6
                                     3.1
                                                     1.5
                                                                   0.2 Iris-setosa
                       5.0
                                     3.6
                                                     1.4
                                                                   0.2 Iris-setosa
          5
                       5.0
                                     3.9
                                                     1.7
                                                                   0.4 Iris-setosa
                                     3.4
                                                     1.4
                       4.6
                                                                   0.3 Iris-setosa
         7
                       5.0
                                     3.4
                                                     1.5
                                                                   0.2 Iris-setosa
                       4.4
                                     2.9
                                                     1.4
                                                                   0.2 Iris-setosa
          9
                       4.9
                                     3.1
                                                     1.5
                                                                   0.1 Iris-setosa
```

8. Showing the unique data of a specific column.

```
print("Species")
print(iris['species'].unique())

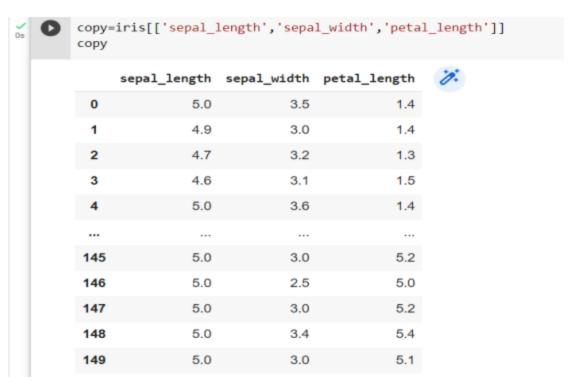
Species
['Iris-setosa' 'Iris-versicolor' 'Iris-virginica']
```

9. Showing the data frame slicing.

os D	<pre>iris1=iris.iloc[0:7] iris1</pre>					
		sepal_length	sepal_width	petal_length	petal_width	species
	0	5.0	3.5	1.4	0.2	Iris-setosa
	1	4.9	3.0	1.4	0.2	Iris-setosa
	2	4.7	3.2	1.3	0.2	Iris-setosa
	3	4.6	3.1	1.5	0.2	Iris-setosa
	4	5.0	3.6	1.4	0.2	Iris-setosa
	5	5.0	3.9	1.7	0.4	Iris-setosa
	6	4.6	3.4	1.4	0.3	Iris-setosa

10. Showing the data frame slicing.

	iris2=iris.loc[:,'sepal_length':'petal_width'] iris2					
	sepal_length	sepal_width	petal_length	petal_width		
0	5.0	3.5	1.4	0.2		
1	4.9	3.0	1.4	0.2		
2	4.7	3.2	1.3	0.2		
3	4.6	3.1	1.5	0.2		
4	5.0	3.6	1.4	0.2		
145	5.0	3.0	5.2	2.3		
146	5.0	2.5	5.0	1.9		
147	5.0	3.0	5.2	2.0		
148	5.0	3.4	5.4	2.3		
149	5.0	3.0	5.1	1.8		



11. Showing the data frame extracting.

