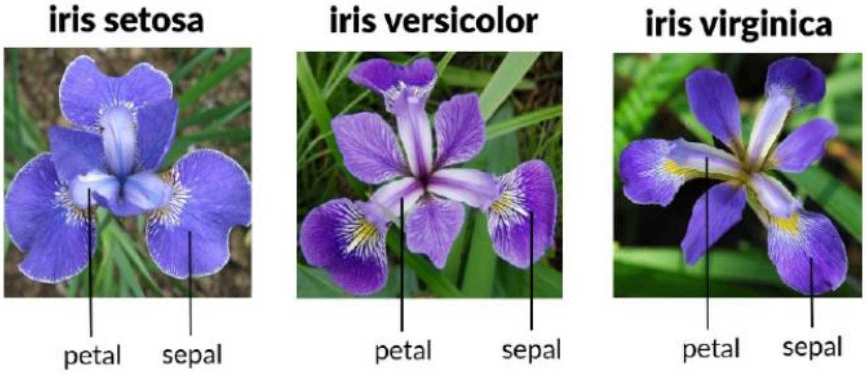
EDA for the Iris Flower Dataset



ANKIT RAJ
16/09/2023

INTRODUCTION

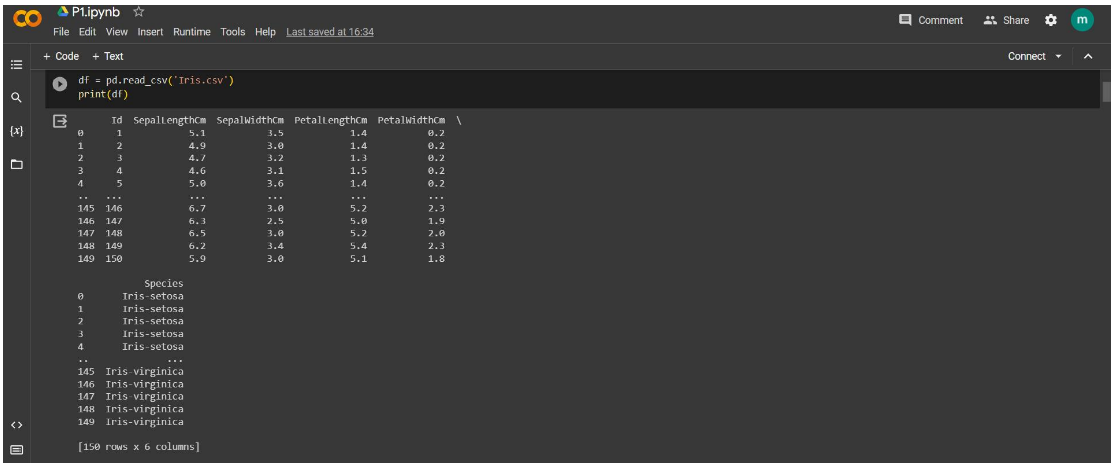
Every machine learning project begins by understanding what the data and drawing the objectives. While applying machine learning algorithms to your data set, you are understanding, building and analyzing the data as to get the end result.

Following are the steps involved in creating a well-defined ML project:
1] Understand and define the problem
2] Prepare the data
3] Explore and Analyse the data

1.1 Problem statement This data set consists of the physical parameters of three species of flower - Versicolor, Setosa and Virginica. The numeric parameters which the dataset contains are Sepal width, Sepal length, Petal width and Petal length. In this data we will be predicting the classes of the flowers based on these parameters. The data consists of continuous numeric values which describe the dimensions of the respective features.We will be training the model based on these features.

1.2 Data Prepare

It has been created Ronald Fisher in 1936. It contains the petal length, petal width,sepal length and sepal width of 150 iris flowers from 3 different species. Variables present in given dataset are SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm, Species.

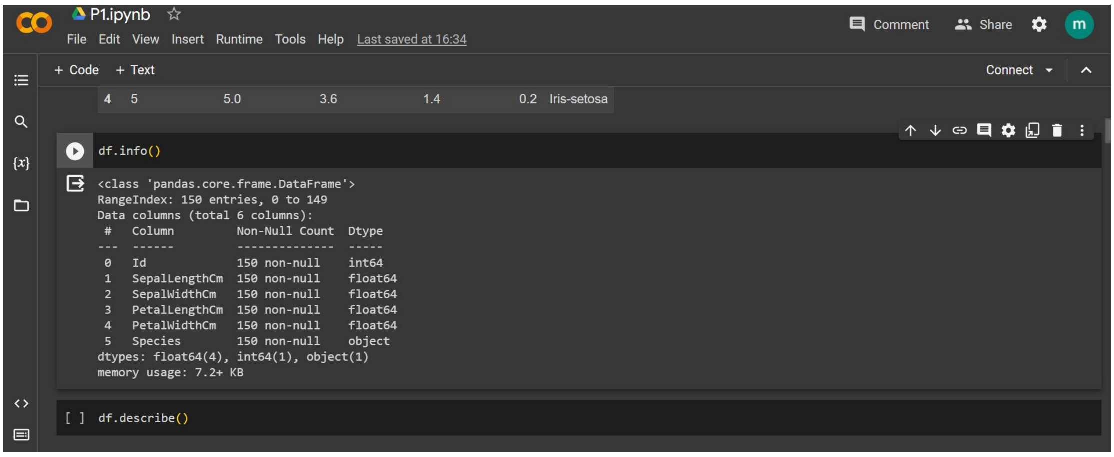

P1.ipynb
File Edit View Insert Runtime Tools Help Last saved at 16:34
Comment
Share
m
+ Code + Text
Connect

```
df = pd.read_csv('Iris.csv')
print(df)
```

```
      Id  SepalLengthCm  SepalWidthCm  PetalLengthCm  PetalWidthCm  \
0      1            5.1           3.5            1.4           0.2
1      2            4.9           3.0            1.4           0.2
2      3            4.7           3.2            1.3           0.2
3      4            4.6           3.1            1.5           0.2
4      5            5.0           3.6            1.4           0.2
..   ...            ...           ...            ...           ...
145  146            6.7           3.0            5.2           2.3
146  147            6.3           2.5            5.0           1.9
147  148            6.5           3.0            5.2           2.0
148  149            6.2           3.4            5.4           2.3
149  150            5.9           3.0            5.1           1.8

            Species
0       Iris-setosa
1       Iris-setosa
2       Iris-setosa
3       Iris-setosa
4       Iris-setosa
..              ...
145  Iris-virginica
146  Iris-virginica
147  Iris-virginica
148  Iris-virginica
149  Iris-virginica

[150 rows x 6 columns]
```


P1.ipynb
File Edit View Insert Runtime Tools Help Last saved at 16:34
Comment
Share
+ Code + Text
Connect

```
4   5   5.0   3.6   1.4   0.2   Iris-setosa
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 150 entries, 0 to 149
Data columns (total 6 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   Id             150 non-null    int64  
 1   SepalLengthCm  150 non-null    float64
 2   SepalWidthCm   150 non-null    float64
 3   PetalLengthCm  150 non-null    float64
 4   PetalWidthCm   150 non-null    float64
 5   Species        150 non-null    object 
dtypes: float64(4), int64(1), object(1)
memory usage: 7.2+ KB
```

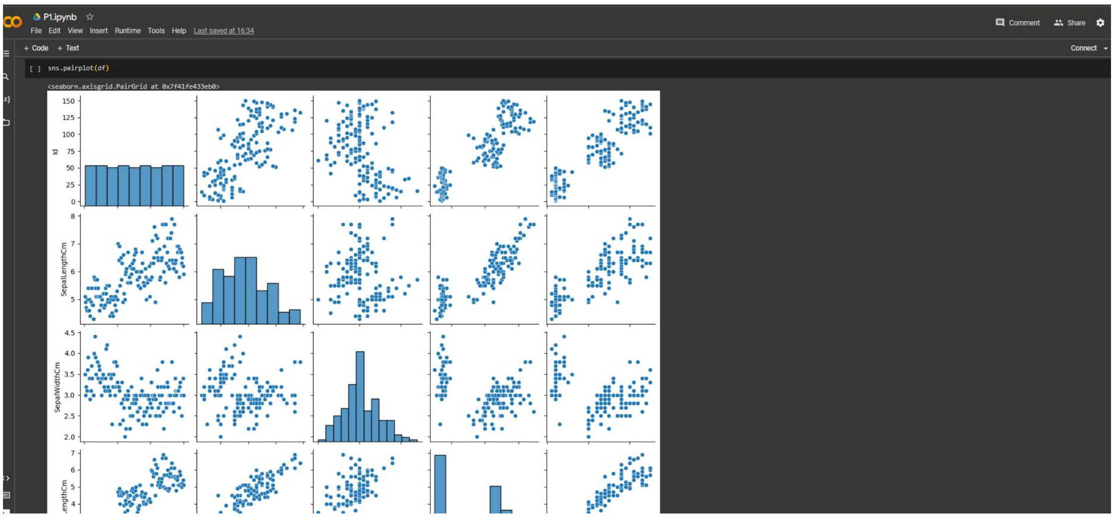
```
[ ]  df.describe()
```

Now, View the info of the data frame that contains details like the count of non-null variables and the column's datatype along with the column names. It will also show the memory usage.

2. Methodology

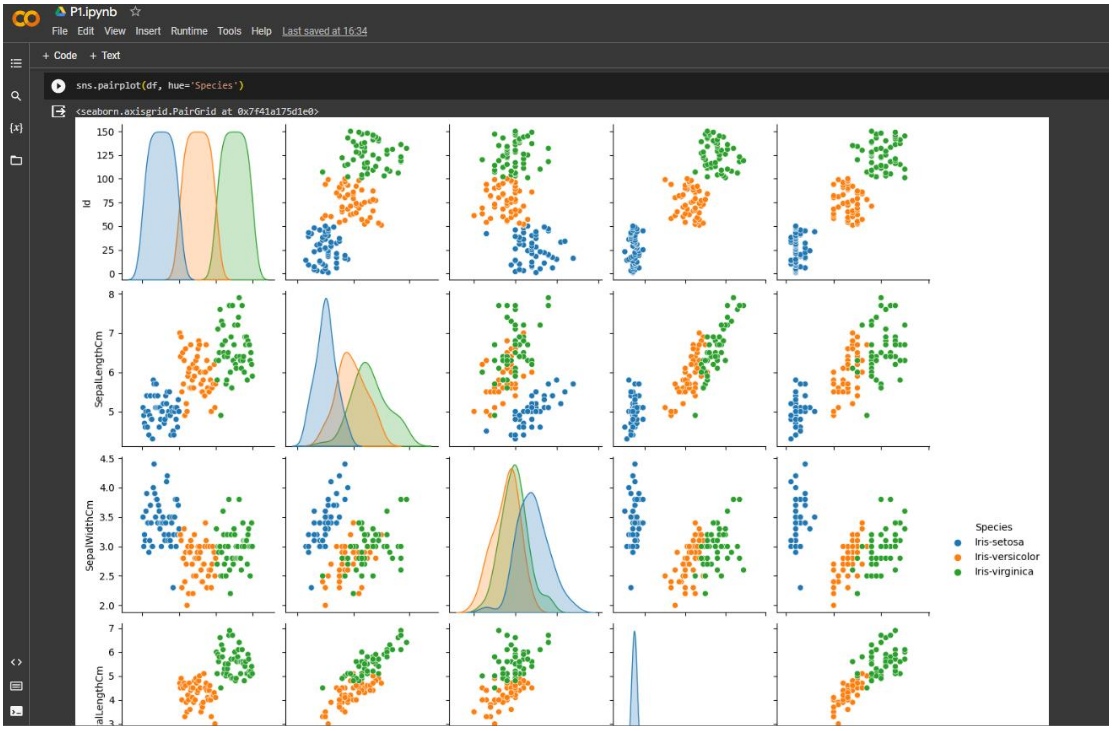
2.1 Pre Processing Any predictive modeling requires that we look at the data before we start modeling. However, in data mining terms looking at data refers to so much more than just looking. Looking at data refers to exploring the data, cleaning the data as well as visualizing the data through graphs and plots. This is often called as Exploratory Data Analysis.

2.1.1 Exploratory Data Analysis In exploring the data we have,If there are any missing values, then modify them before using the dataset. Formodifying you can use the fillna() method. It will fill null values.



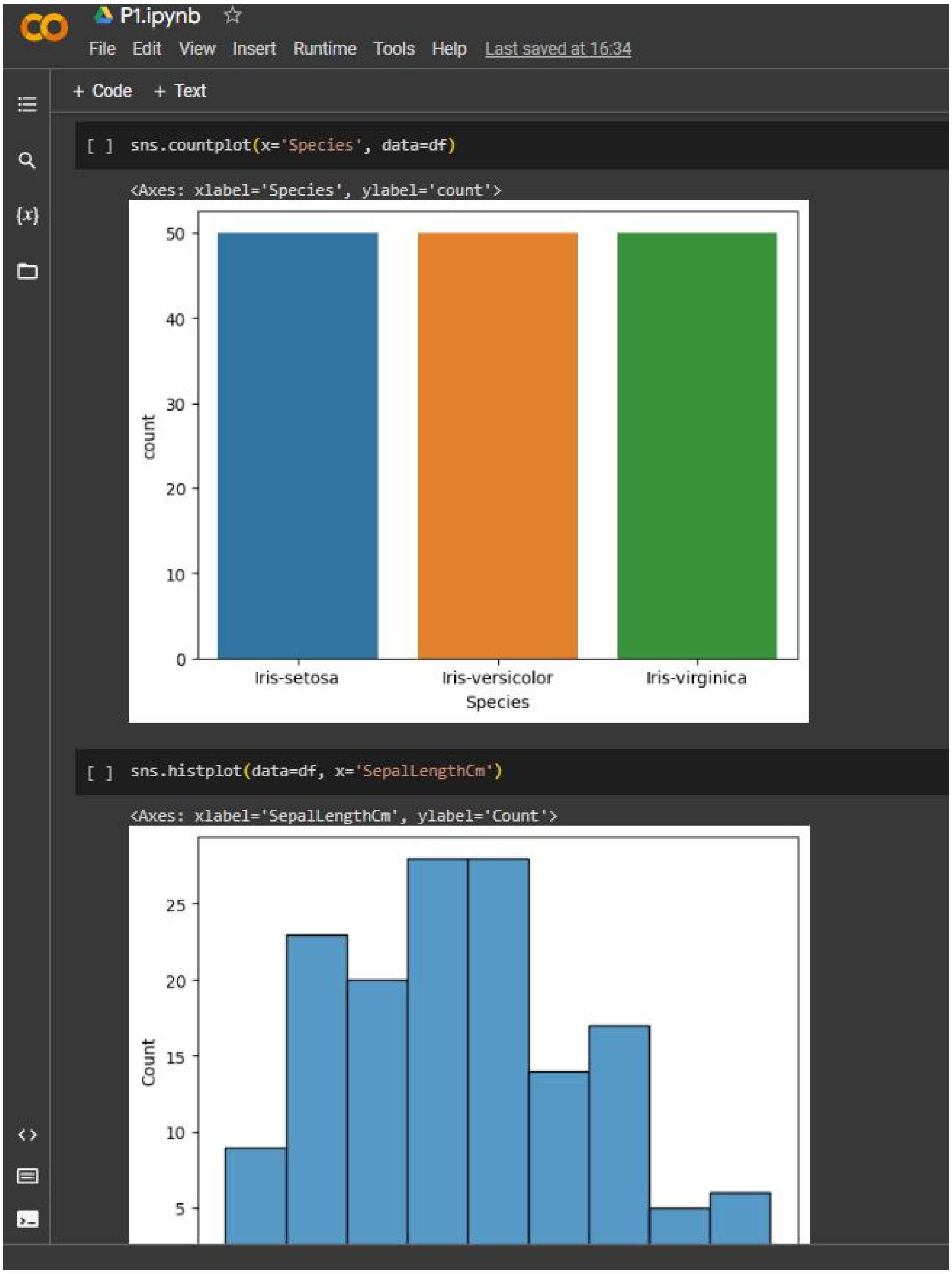
```
[ ]  sns.pairplot(df)
```

```
<seaborn.axisgrid.PairGrid at 0x7f41fe433eb0>
```



```
sns.pairplot(df, hue='Species')
```

```
<seaborn.axisgrid.PairGrid at 0x7f41a175d1e0>
```



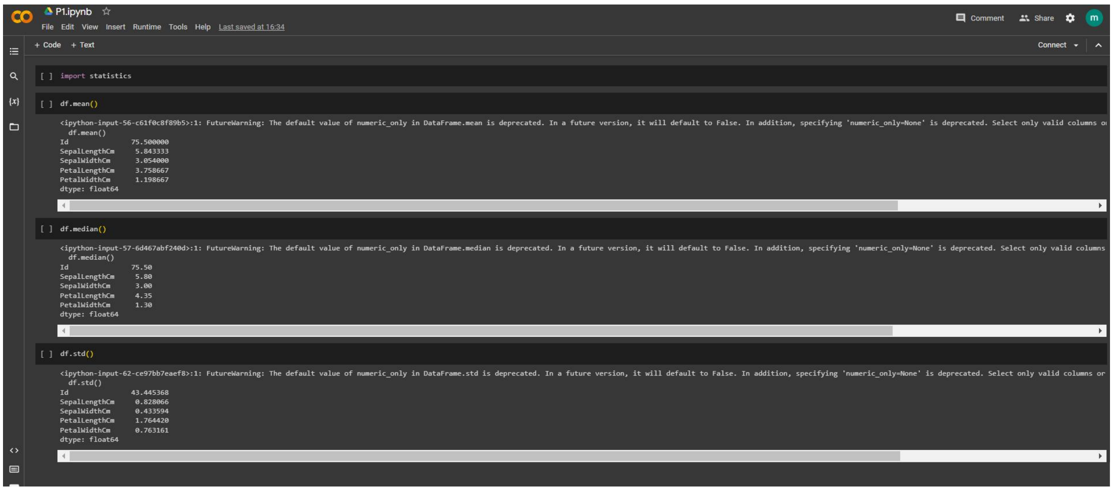
P1.ipynb
File Edit View Insert Runtime Tools Help Last saved at 16:34
+ Code + Text

```
[ ] sns.countplot(x='Species', data=df)
```

```
<Axes: xlabel='Species', ylabel='count'>
```

```
[ ] sns.histplot(data=df, x='SepalLengthCm')
```

```
<Axes: xlabel='SepalLengthCm', ylabel='Count'>
```



P1.ipynb
File Edit View Insert Runtime Tools Help Last saved at 16:34
Comment
Share
+ Code + Text
Connect

```
[ ] import statistics
```

```
[ ] df.mean()
```

```
<ipython-input-56-c61f0c8f89b5>:1: FutureWarning: The default value of numeric_only in DataFrame.mean is deprecated. In a future version, it will default to False. In addition, specifying 'numeric_only=None' is deprecated. Select only valid columns o
  df.mean()
Id               75.500000
SepalLengthCm     5.843333
SepalWidthCm      3.054000
PetalLengthCm     3.758667
PetalWidthCm      1.198667
dtype: float64
```

```
[ ] df.median()
```

```
<ipython-input-57-6d467abf240d>:1: FutureWarning: The default value of numeric_only in DataFrame.median is deprecated. In a future version, it will default to False. In addition, specifying 'numeric_only=None' is deprecated. Select only valid columns
  df.median()
Id               75.50
SepalLengthCm     5.80
SepalWidthCm      3.00
PetalLengthCm     4.35
PetalWidthCm      1.30
dtype: float64
```

```
[ ] df.std()
```

```
<ipython-input-62-ce97bb7eaef8>:1: FutureWarning: The default value of numeric_only in DataFrame.std is deprecated. In a future version, it will default to False. In addition, specifying 'numeric_only=None' is deprecated. Select only valid columns or
  df.std()
Id               43.445368
SepalLengthCm     0.828066
SepalWidthCm      0.433594
PetalLengthCm     1.764420
PetalWidthCm      0.763161
dtype: float64
```

Conclusion

Flower classification is a very important, simple, and basic project for any machine learning student. Every machine learning student should be thorough with the iris flowers dataset.