

```
In [1]: import pandas as pd
import numpy as np
import statistics as st
```

```
In [2]: 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]

```
In [3]: df.columns
df.head()
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
```

```
In [4]: df.dtypes
df.isnull().sum()
df.describe()
```

Out[4]:

| | Id | SepalLengthCm | SepalWidthCm | PetalLengthCm | PetalWidthCm |
|--------------|------------|---------------|--------------|---------------|--------------|
| count | 150.000000 | 150.000000 | 150.000000 | 150.000000 | 150.000000 |
| mean | 75.500000 | 5.843333 | 3.054000 | 3.758667 | 1.198667 |
| std | 43.445368 | 0.828066 | 0.433594 | 1.764420 | 0.763161 |
| min | 1.000000 | 4.300000 | 2.000000 | 1.000000 | 0.100000 |
| 25% | 38.250000 | 5.100000 | 2.800000 | 1.600000 | 0.300000 |
| 50% | 75.500000 | 5.800000 | 3.000000 | 4.350000 | 1.300000 |
| 75% | 112.750000 | 6.400000 | 3.300000 | 5.100000 | 1.800000 |
| max | 150.000000 | 7.900000 | 4.400000 | 6.900000 | 2.500000 |

```
In [5]: df.value_counts("Species")
```

Out[5]: Species
Iris-setosa 50
Iris-versicolor 50
Iris-virginica 50
dtype: int64

```
In [6]: from sklearn.datasets import load_iris  
iris = load_iris()#Load directly iris and get description  
#view data description and information  
print(iris.DESCR)
```

```
.. _iris_dataset:
```

```
Iris plants dataset
```

```
-----
```

```
**Data Set Characteristics:**
```

```
:Number of Instances: 150 (50 in each of three classes)
:Number of Attributes: 4 numeric, predictive attributes and the class
:Attribute Information:
  - sepal length in cm
  - sepal width in cm
  - petal length in cm
  - petal width in cm
  - class:
    - Iris-Setosa
    - Iris-Versicolour
    - Iris-Virginica
```

```
:Summary Statistics:
```

```
=====
      Min  Max  Mean  SD  Class Correlation
=====
sepal length:  4.3  7.9   5.84  0.83    0.7826
sepal width:   2.0  4.4   3.05  0.43   -0.4194
petal length:   1.0  6.9   3.76  1.76    0.9490 (high!)
petal width:   0.1  2.5   1.20  0.76    0.9565 (high!)
=====
```

```
:Missing Attribute Values: None
:Class Distribution: 33.3% for each of 3 classes.
:Creator: R.A. Fisher
:Donor: Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov)
:Date: July, 1988
```

The famous Iris database, first used by Sir R.A. Fisher. The dataset is taken from Fisher's paper. Note that it's the same as in R, but not as in the UCI Machine Learning Repository, which has two wrong data points.

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. (See Duda & Hart, for example.) The data set contains 3 classes of 50 instances each, where each class refers to a

type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

```
.. topic:: References
```

- Fisher, R.A. "The use of multiple measurements in taxonomic problems" Annual Eugenics, 7, Part II, 179-188 (1936); also in "Contributions to Mathematical Statistics" (John Wiley, NY, 1950).
- Duda, R.O., & Hart, P.E. (1973) Pattern Classification and Scene Analysis. (Q327.D83) John Wiley & Sons. ISBN 0-471-22361-1. See page 218.
- Dasarthy, B.V. (1980) "Nosing Around the Neighborhood: A New System

Structure and Classification Rule for Recognition in Partially Exposed Environments". IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-2, No. 1, 67-71.

- Gates, G.W. (1972) "The Reduced Nearest Neighbor Rule". IEEE Transactions on Information Theory, May 1972, 431-433.
- See also: 1988 MLC Proceedings, 54-64. Cheeseman et al's AUTOCLASS II conceptual clustering system finds 3 classes in the data.
- Many, many more ...

In []: