SciKit Learn

TECH TALENT ACADEMY

WOMEN IN DATA ACADEMY

Session Content

- What is machine learning
- What is Sci-kit Learn
- Features

TECH TALENT
ACADEMY

Machine Learning

It makes decisions based in what it has learned from the data

Does not require a human programmer to tell it what to do with the inputted data

Includes:

concept learning; function learning (or predictive modelling); clustering and looking for predictive patterns

Learns through available data observed through experiences or instructions

WOMEN IN DATA ACADEMY

The Steps for Machine Learning

- Load dataset
- Explore dataset
- Process data
- Visualise data
- Pick algorithms
- Train and test sets
- Standardise data (scale)
- Compile and fit
- Learning from the model and evaluations

What is Scikit - learn?

- Developed in 2007 (a Google summer project) but released to the public 2010
- Provides a range of supervised and unsupervised learning algorithms
- Built on SciPy (Scientific Python)
- Includes : NumPy, Pandas, SciPy, Matplotlib (2D/3D plotting), Ipython (interactive console) & Sympy (symbolic maths)
- The extension of SciPy is called SciKits
- Scikit provides an extension for learning algorithms

TECH TALENT
ACADEMY

Session Content

LOADING DATASETS

EXPLORING THE DATA

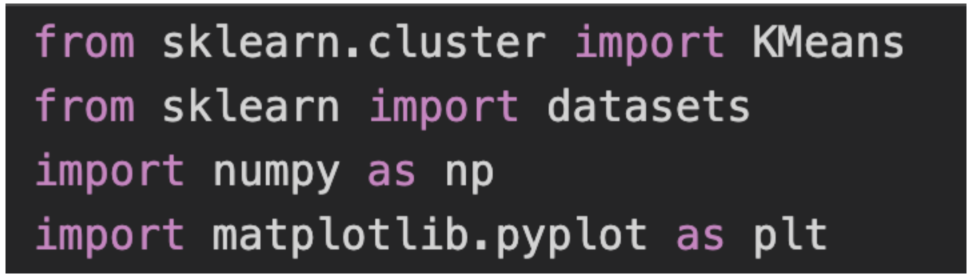
WOMEN IN DATA
ACADEMY

TECH TALENT
ACADEMY

Importing Required Libraries

First we need to declare the libraries we will be using.

1. SK Learn cluster for K-Means and datasets to generate the data.
2. NumPy.
3. Matplotlib for data visualisation.



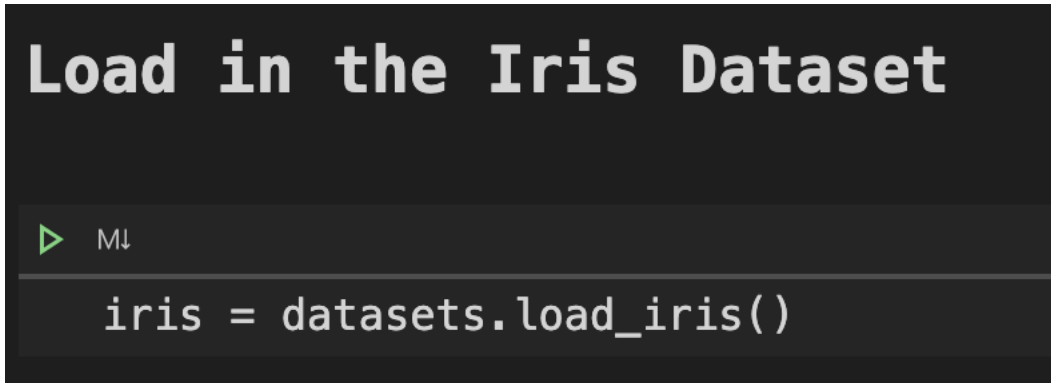
```
from sklearn.cluster import KMeans
from sklearn import datasets
import numpy as np
import matplotlib.pyplot as plt
```

WOMEN IN DATA
ACADEMY

Loading the Data set

Datasets is a module that is part of the library and allows you to load and investigate data.

- Step one in data science is to load your data.
- Data can be collated by you, or alternative systems and resources.
- We are going to be using data sets that are already available in the python library called **iris.**



Load in the Iris Dataset

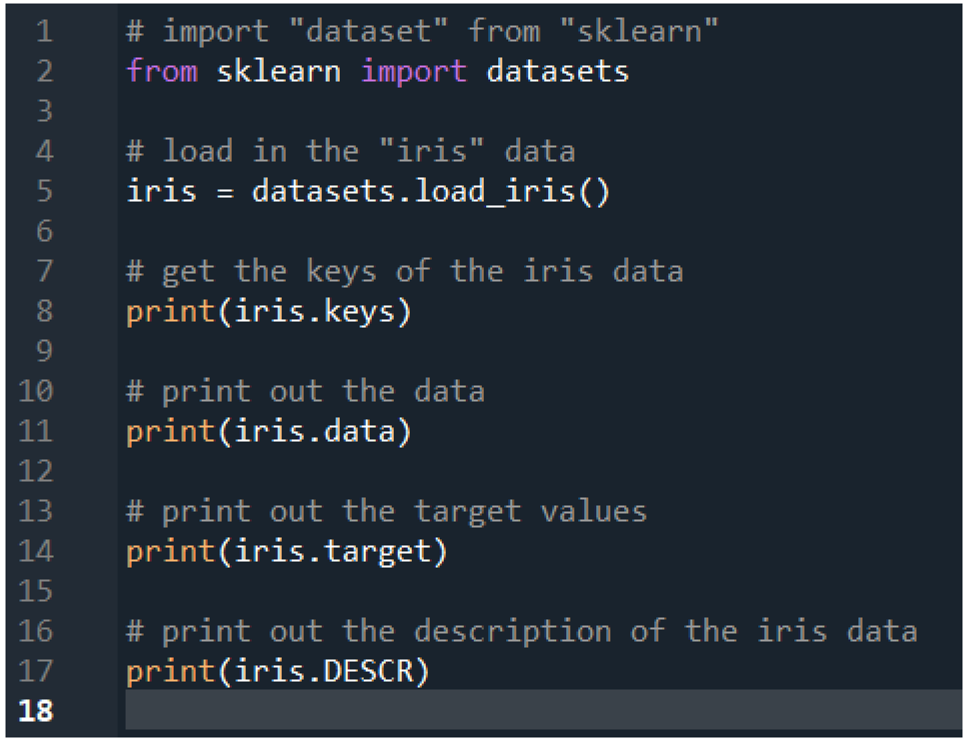
```
iris = datasets.load_iris()
```

TECH TALENT ACADEMY

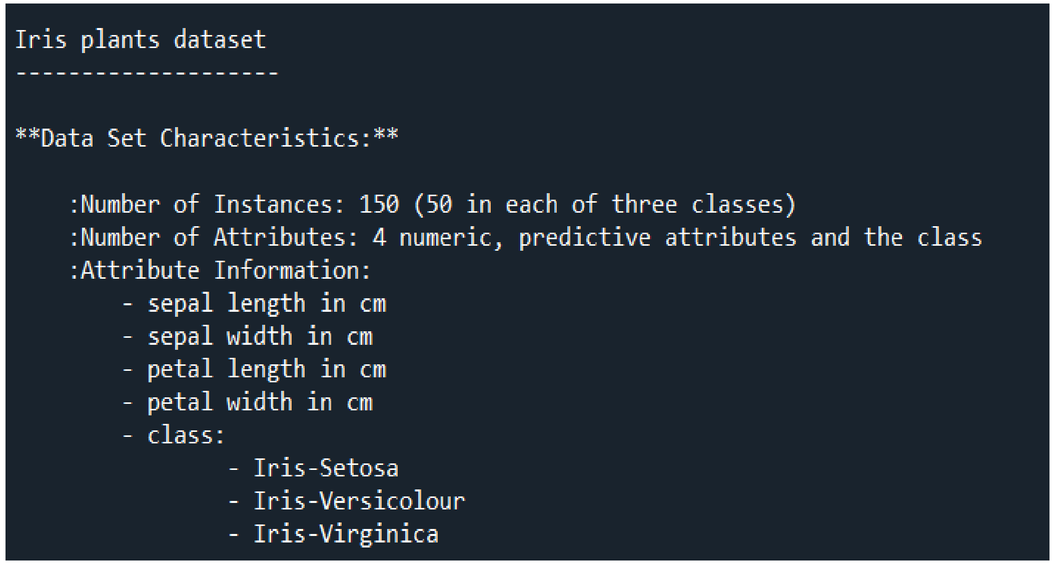
Explore the Data

- Its important to explore what you have, so it is a good idea to perform an exploratory data analysis (EDA) on the data.
- Start by collating the basic information.

data – actual data/numbers in dataset
target – access the values of labels/headers
DESCR – access description



```python
# import "dataset" from "sklearn"
from sklearn import datasets

# load in the "iris" data
iris = datasets.load_iris()

# get the keys of the iris data
print(iris.keys)

# print out the data
print(iris.data)

# print out the target values
print(iris.target)

# print out the description of the iris data
print(iris.DESCR)

```



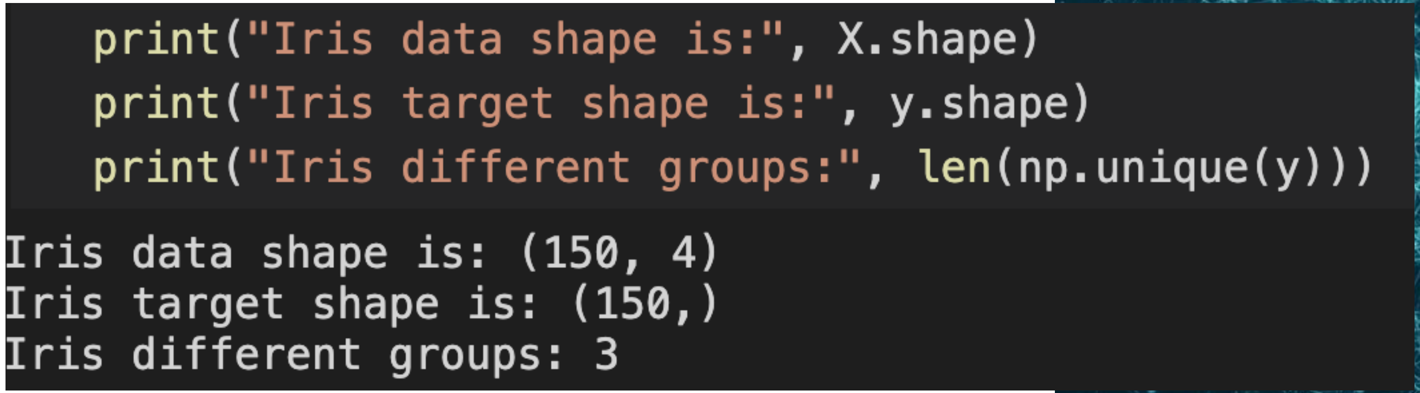
```
Iris plants dataset
--------------------

**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
```

WOMEN IN DATA ACADEMY

TECH TALENT
ACADEMY

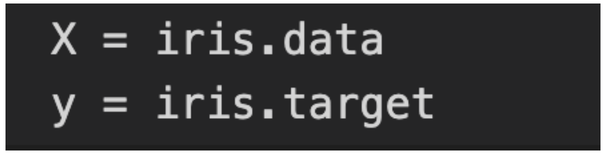
Assigning Values

- In Data Science, you will see the values X and y used a lot!
- X represents the data
- While y represents the target.
- Let's reassign our values.



```
print("Iris data shape is:", X.shape)
print("Iris target shape is:", y.shape)
print("Iris different groups:", len(np.unique(y)))
```

```
Iris data shape is: (150, 4)
Iris target shape is: (150,)
Iris different groups: 3
```



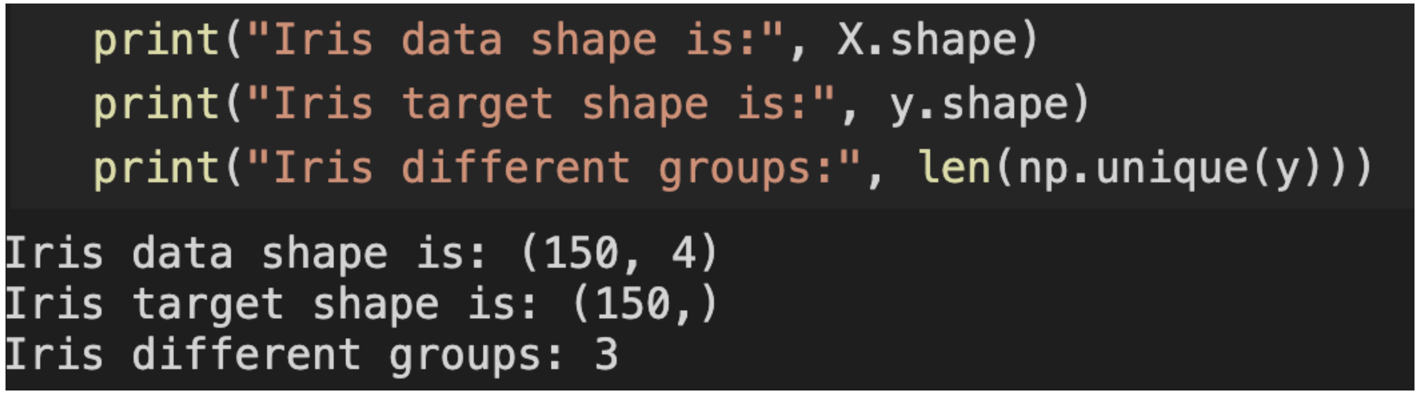
```
X = iris.data
y = iris.target
```

WOMEN IN DATA
ACADEMY

Check the Data Shape

iris.data :There are 150 samples and 4 features.

iris.targets are also 150



```
    print("Iris data shape is:", X.shape)
    print("Iris target shape is:", y.shape)
    print("Iris different groups:", len(np.unique(y)))

Iris data shape is: (150, 4)
Iris target shape is: (150,)
Iris different groups: 3
```

Session Content

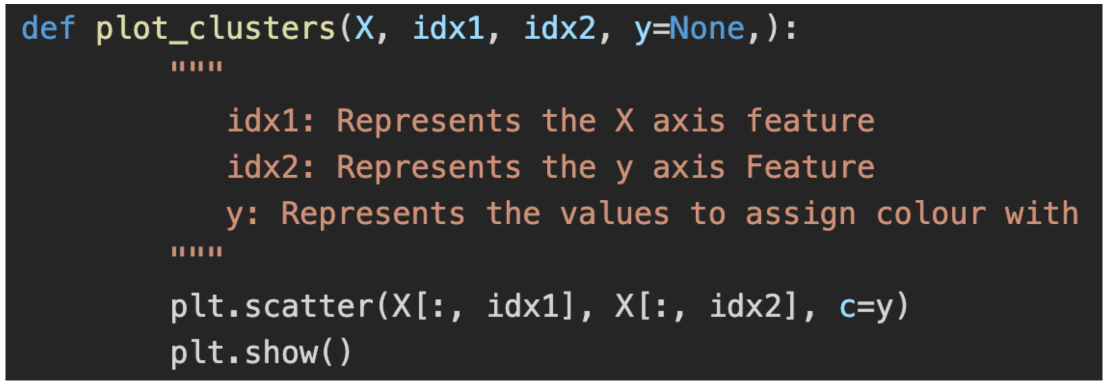
REUSABLE CODE

MATPLOTLIB DATA VISUALISATION

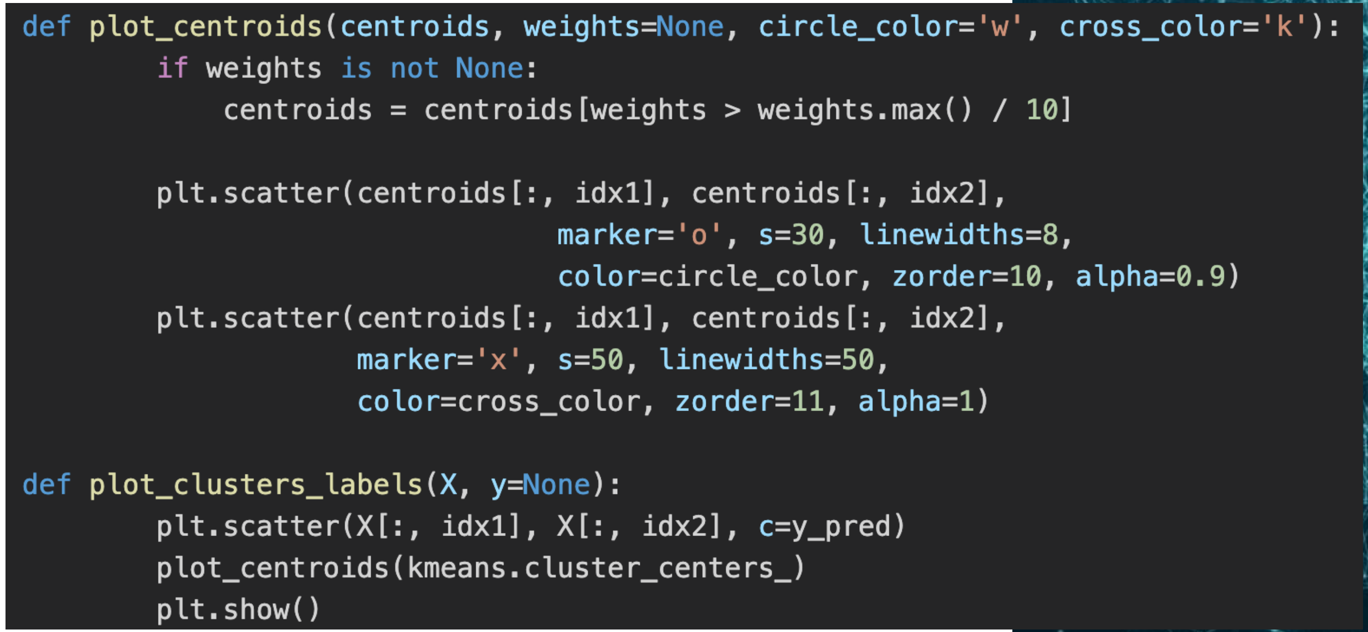
SCATTERPLOT

Reusable Code

- To make exploring the data less repetitive, we can create reusable functions.
- This will be useful for the next stage, but it's not essential.



```python
def plot_clusters(X, idx1, idx2, y=None,):
        """
           idx1: Represents the X axis feature
           idx2: Represents the y axis Feature
           y: Represents the values to assign colour with
        """
        plt.scatter(X[:, idx1], X[:, idx2], c=y)
        plt.show()
```



```python
def plot_centroids(centroids, weights=None, circle_color='w', cross_color='k'):
        if weights is not None:
            centroids = centroids[weights > weights.max() / 10]

        plt.scatter(centroids[:, idx1], centroids[:, idx2],
                                marker='o', s=30, linewidths=8,
                                color=circle_color, zorder=10, alpha=0.9)
        plt.scatter(centroids[:, idx1], centroids[:, idx2],
                    marker='x', s=50, linewidths=50,
                    color=cross_color, zorder=11, alpha=1)

def plot_clusters_labels(X, y=None):
        plt.scatter(X[:, idx1], X[:, idx2], c=y_pred)
        plot_centroids(kmeans.cluster_centers_)
        plt.show()
```

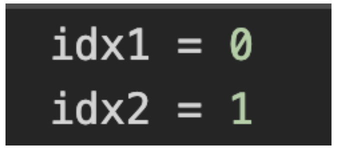
TECH TALENT
ACADEMY

MatplotLib Data Visualization

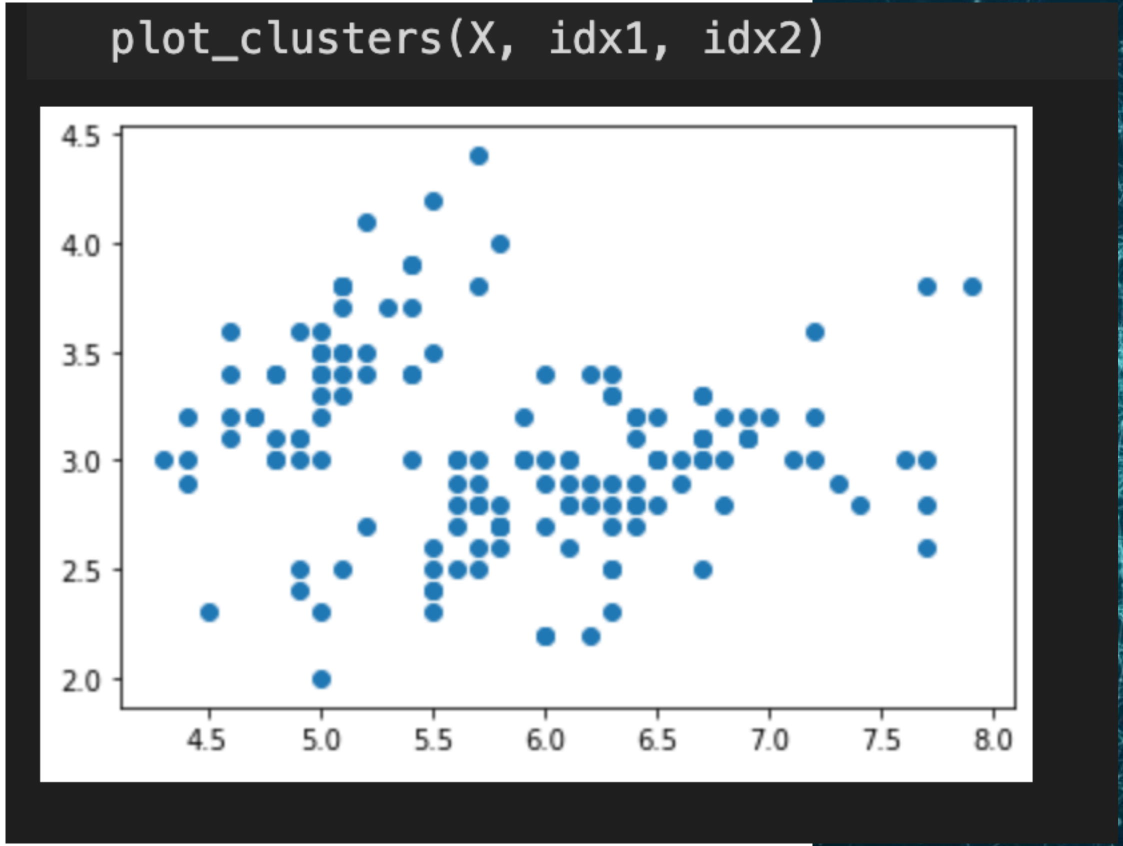
WOMEN IN DATA
ACADEMY

Feature Selection

- How many features does the dataset have?
- We will assign the features we want to variables
- We will then pass them to the plotting function



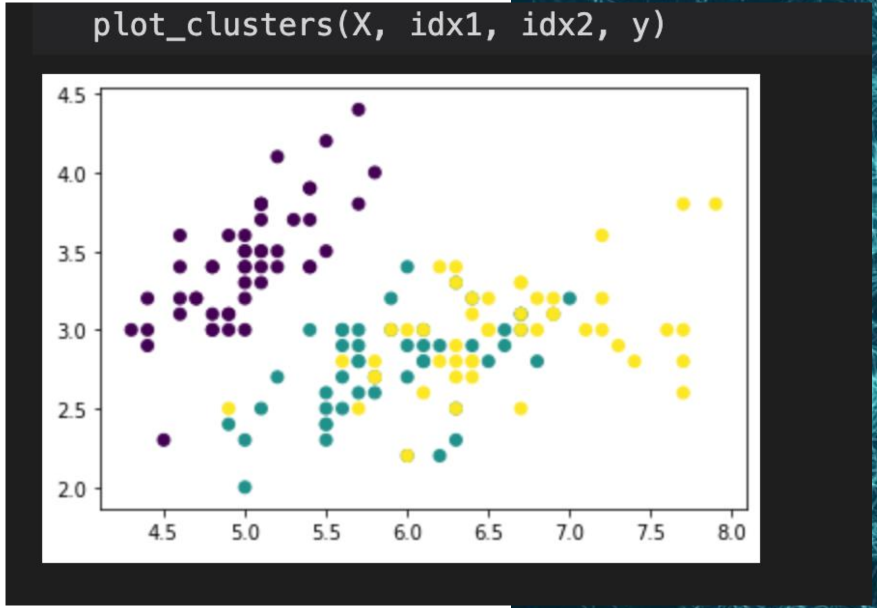
```
idx1 = 0
idx2 = 1
```



```
plot_clusters(X, idx1, idx2)
```

Using labels

- For clustering we would not usually have labels, but as we are learning and the dataset has labels, we can use these to see the general data distribution.
- We will replot the data but will pass through the y variable
- This is referred to as the “Ground Truth”
- Try using a combination of features to see what ones match well together.



```
plot_clusters(X, idx1, idx2, y)
```

K-Means Algorithm

The simplest and most widely used algorithm

Used to solve clustering problems

A simplistic route to classify data through a certain number of clusters that are configured prior to running the algorithm – this number is called **K**

The k-means will find the nearest cluster centre for each data point and assign the data point closest to that cluster

When all data has been assigned to clusters the cluster centre will be recomputed, the process is then repeated until most data points stick to the same cluster

Session Content

CLUSTERING THE IRIS DATA

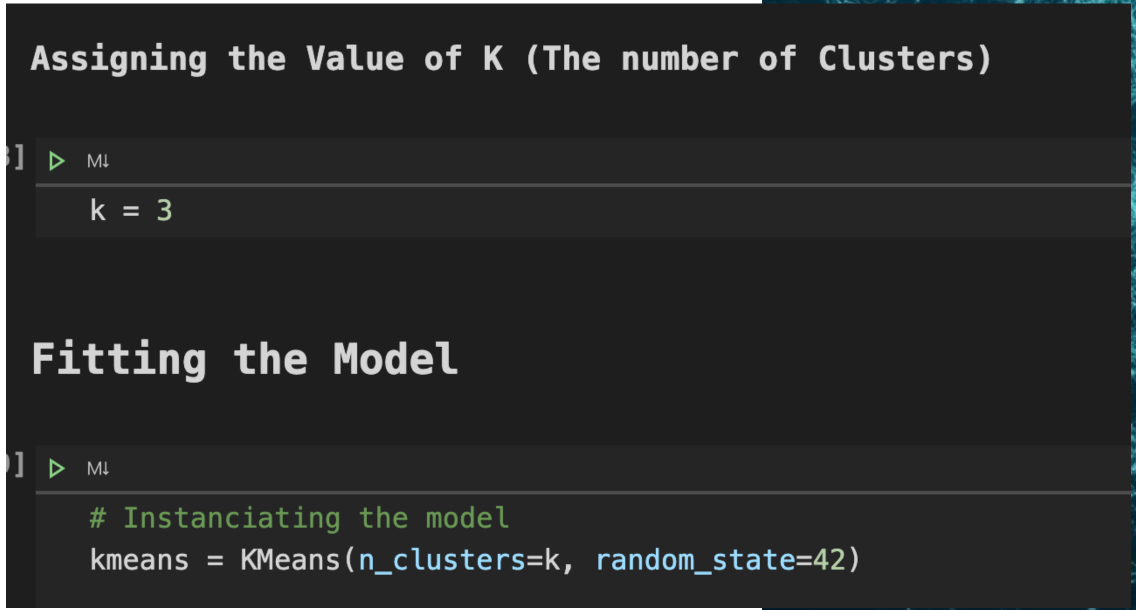
PREDICT THE LABELS

VISUALISATION

Clustering the iris data

- Find the clusters of your training set (K-Means)
- K-Means required the user to tell it how many clusters there are within the dataset.
- These are referred to as k (clusters).
- We know that there are three, but in the real world its up to you to decide, based on your data exploration and the hypothesis you make exploring.

Set the n_clusters argument to 3, this indicates the number of clusters or groups you want your data to form and number of centroids to generate.



Assigning the Value of K (The number of Clusters)

```
k = 3
```

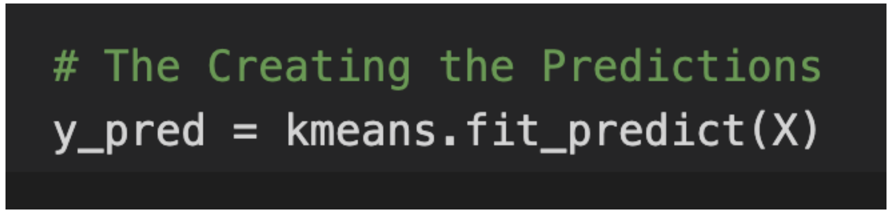
Fitting the Model

```
# Instanciating the model
kmeans = KMeans(n_clusters=k, random_state=42)
```

TECH TALENT
ACADEMY

Fitting the Model

- Now that we have created our K-means model.
- Time to get it to cluster the data into class.

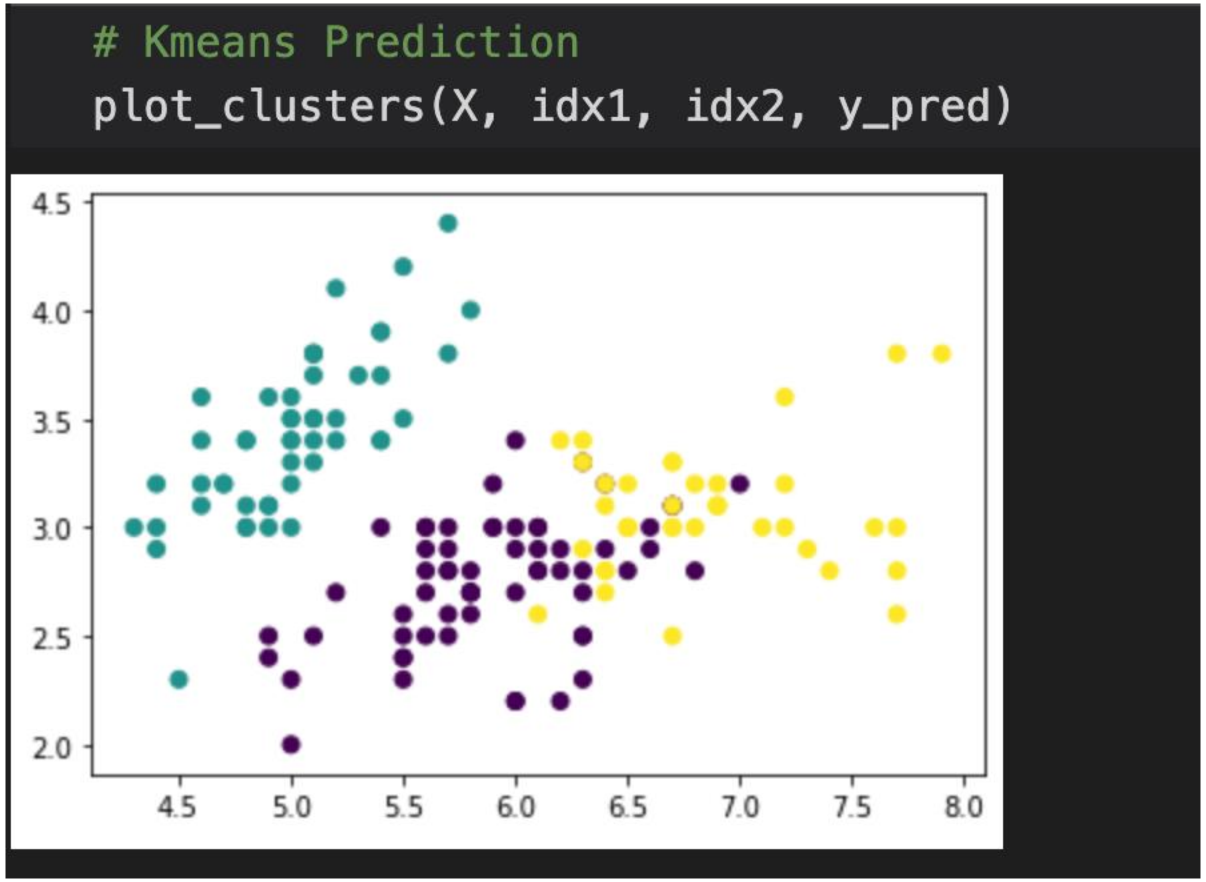


```
# The Creating the Predictions
y_pred = kmeans.fit_predict(X)
```

WOMEN IN DATA
ACADEMY

Visualising the Results

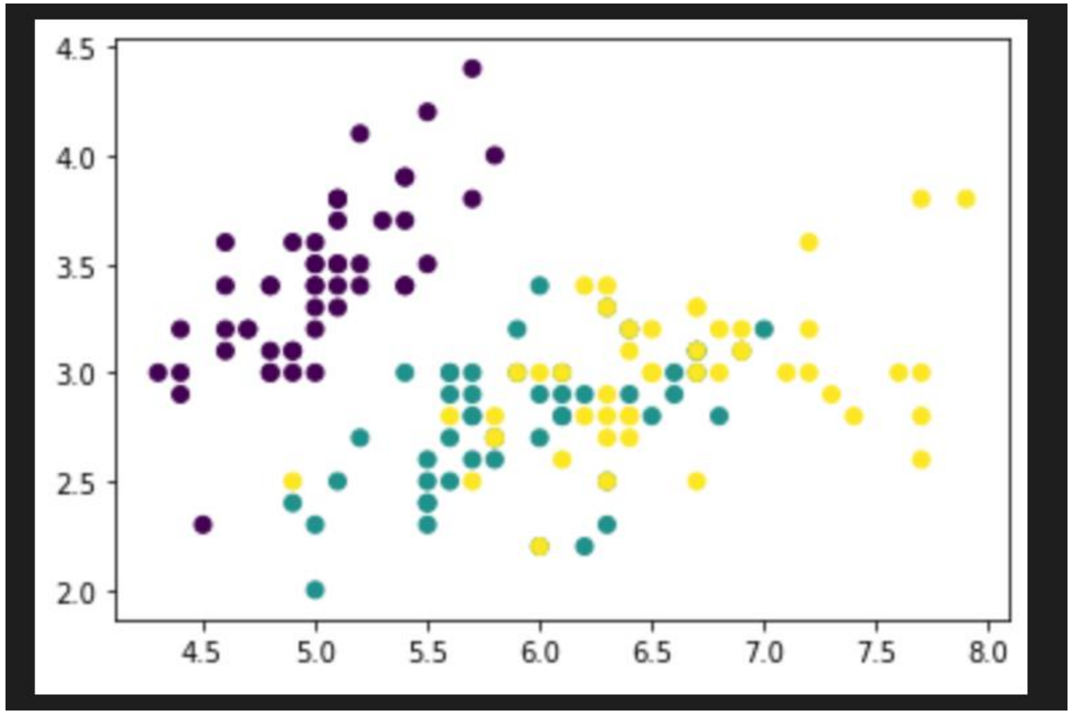
We will do the same as before, but we will use the y_pred for the prediction's classes.

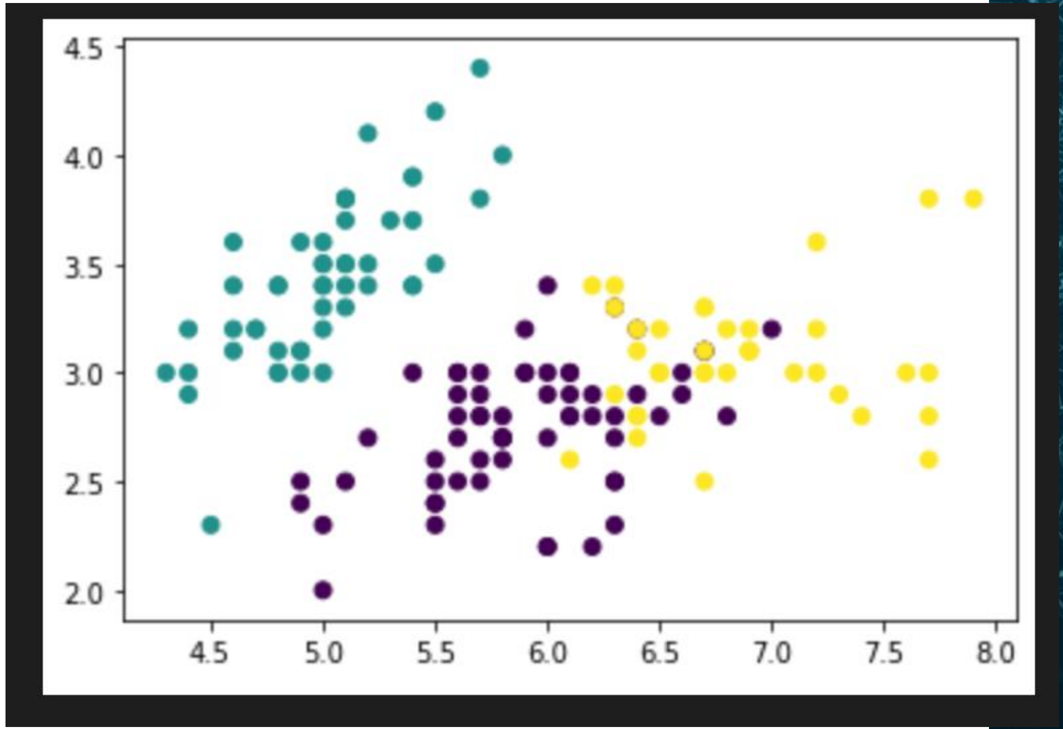


```
# Kmeans Prediction
plot_clusters(X, idx1, idx2, y_pred)
```

Evaluation of your Clustering Model

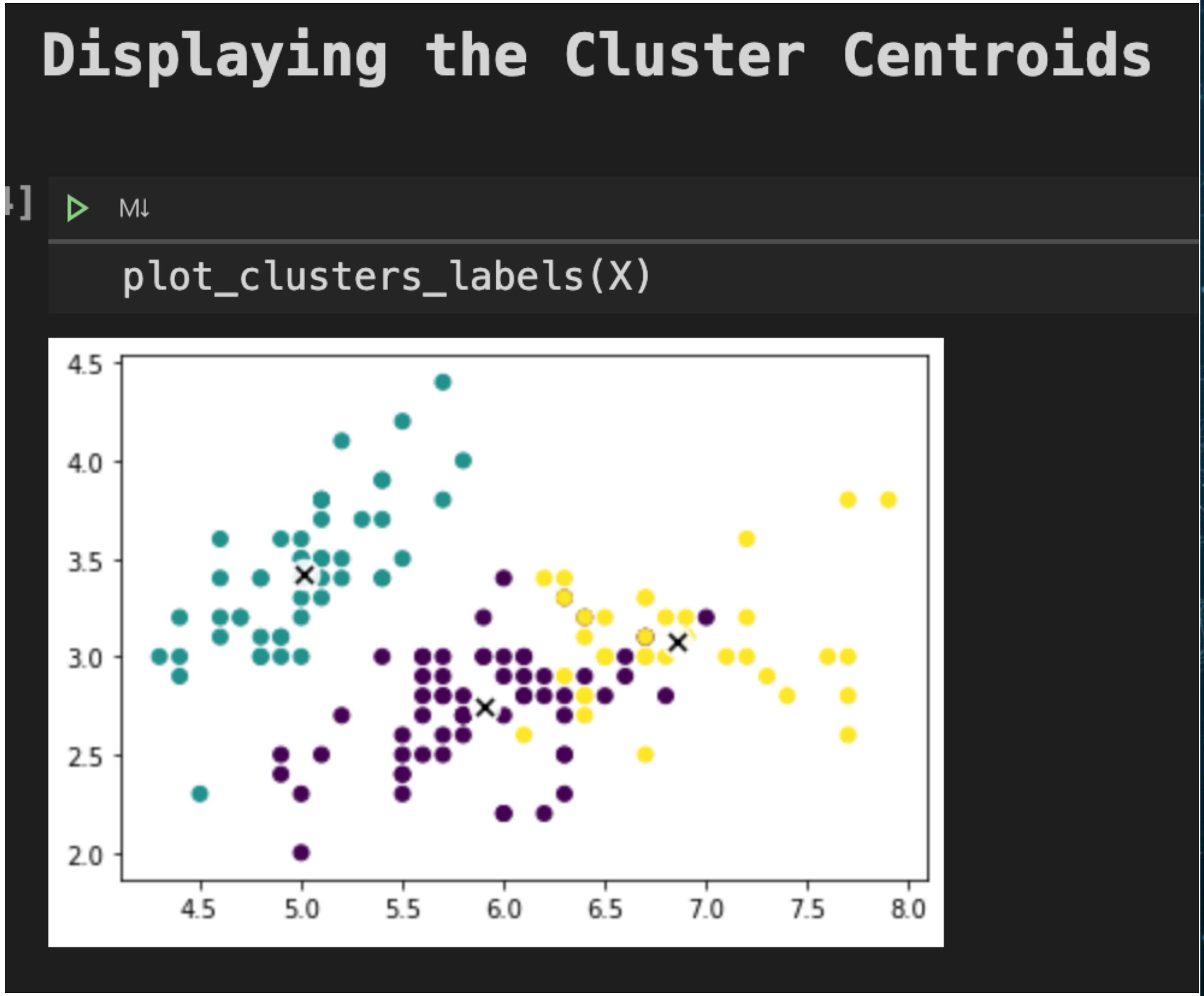
- Let's compare the Predicted clusters with the ground truth
- Notice any difference with the Ground Truth?





Seeing the Cluster Centres

- The centres of the clusters are referred to the centroids
- We can see where the data has been assigned to.



Displaying the Cluster Centroids

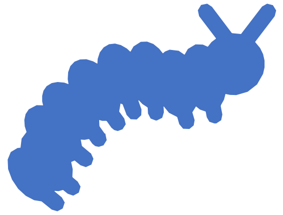
```
plot_clusters_labels(X)
```

Further Learning

- https://scikit-learn.org/stable/auto_examples/index.html
- https://towardsdatascience.com/5-powerful-scikit-learn-models-e9b12375320d

TECH TALENT
ACADEMY

Home Learning Task



https://stackabuse.com/linear-regression-in-python-with-scikit-learn/

From this website select follow and complete the Linear Regression Model.

WOMEN IN DATA
ACADEMY

WOMEN IN DATA
ACADEMY

TECH TALENT
ACADEMY