In [1]:

```python
import findspark
findspark.init()
```

In [2]:

```python
from pyspark.shell import spark
```

```
Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /__ / .__/\_,_/_/ /_/\_\   version 3.2.2
      /_/

Using Python version 3.10.0 (tags/v3.10.0:b494f59, Oct  4 2021 19:00:18)
Spark context Web UI available at http://LAPTOP-SV7GEA9R:4040
Spark context available as 'sc' (master = local[*], app id = local-1669214396095).
SparkSession available as 'spark'.
```

In [3]:

```python
from pyspark.sql import SparkSession
```

In [4]:

```python
spark = SparkSession.builder.getOrCreate()
```

In [5]:

```python
df = spark.read.csv('C:/Users/Raghavendra K/Downloads/IRIS2.csv', inferSchema=True, header=True)
```

In [6]:

```python
df.printSchema()
df.show(5)
```

```
root
 |-- sepal_length: double (nullable = true)
 |-- sepal_width: double (nullable = true)
 |-- petal_length: double (nullable = true)
 |-- petal_width: double (nullable = true)
 |-- species: string (nullable = true)

+------------+-----------+------------+-----------+-----------+
|sepal_length|sepal_width|petal_length|petal_width|    species|
+------------+-----------+------------+-----------+-----------+
|         5.1|        3.5|         1.4|        0.2|Iris-setosa|
|         4.9|        3.0|         1.4|        0.2|Iris-setosa|
|         4.7|        3.2|         1.3|        0.2|Iris-setosa|
|         4.6|        3.1|         1.5|        0.2|Iris-setosa|
|         5.0|        3.6|         1.4|        0.2|Iris-setosa|
+------------+-----------+------------+-----------+-----------+
only showing top 5 rows
```

In [7]:

```python
df.count()
```

Out[7]:

```
150
```

In [8]:

```python
len(df.columns)
```

Out[8]:

```
5
```

In [9]:

```python
df.describe().show()
```

```
+-------+------------------+-------------------+------------------+------------------+--------------+
|summary|      sepal_length|        sepal_width|      petal_length|       petal_width|       species|
+-------+------------------+-------------------+------------------+------------------+--------------+
|  count|               150|                150|               150|               150|           150|
|   mean| 5.843333333333335| 3.0540000000000007|3.7586666666666693|1.1986666666666672|          null|
| stddev|0.8280661279778637|0.43359431136217375| 1.764420419952262|0.7631607417008414|          null|
|    min|               4.3|                2.0|               1.0|               0.1|   Iris-setosa|
|    max|               7.9|                4.4|               6.9|               2.5|Iris-virginica|
+-------+------------------+-------------------+------------------+------------------+--------------+
```

In [10]:

```python
df.head(5)
```

Out[10]:

```
[Row(sepal_length=5.1, sepal_width=3.5, petal_length=1.4, petal_width=0.2, species='Iris-setosa'),
 Row(sepal_length=4.9, sepal_width=3.0, petal_length=1.4, petal_width=0.2, species='Iris-setosa'),
 Row(sepal_length=4.7, sepal_width=3.2, petal_length=1.3, petal_width=0.2, species='Iris-setosa'),
 Row(sepal_length=4.6, sepal_width=3.1, petal_length=1.5, petal_width=0.2, species='Iris-setosa'),
 Row(sepal_length=5.0, sepal_width=3.6, petal_length=1.4, petal_width=0.2, species='Iris-setosa')]
```

In [11]:

```python
df.groupby("species").count().show()
```

```
+---------------+-----+
|        species|count|
+---------------+-----+
| Iris-virginica|   50|
|    Iris-setosa|   50|
|Iris-versicolor|   50|
+---------------+-----+
```

In [12]:

```python
df.groupby("sepal_width").count().show()
```

```
+-----------+-----+
|sepal_width|count|
+-----------+-----+
|        2.4|    3|
|        3.5|    6|
|        2.9|   10|
|        3.7|    3|
|        2.3|    4|
|        3.4|   12|
|        2.5|    8|
|        3.1|   12|
|        2.7|    9|
|        4.1|    1|
|        2.2|    3|
|        2.8|   14|
|        4.0|    1|
|        3.9|    2|
|        3.8|    6|
|        4.2|    1|
|        3.2|   13|
|        3.0|   26|
|        2.0|    1|
|        2.6|    5|
+-----------+-----+
only showing top 20 rows
```

In [16]:

```python
from pyspark.ml.linalg import Vector
from pyspark.ml.feature import VectorAssembler
```

In [17]:

```python
df.columns
```

Out[17]:

```
['sepal_length', 'sepal_width', 'petal_length', 'petal_width', 'species']
```

In [18]:

```python
input_cols =['sepal_length', 'sepal_width', 'petal_length', 'petal_width']
```

In [19]:

```python
vec_assembler = VectorAssembler(inputCols = input_cols,
                                outputCol = "features")
```

In [20]:

```python
final_data = vec_assembler.transform(df)
```

In [21]:

```python
final_data.show()
```

```
+------------+-----------+------------+-----------+-----------+-----------------+
|sepal_length|sepal_width|petal_length|petal_width|    species|         features|
+------------+-----------+------------+-----------+-----------+-----------------+
|         5.1|        3.5|         1.4|        0.2|Iris-setosa|[5.1,3.5,1.4,0.2]|
|         4.9|        3.0|         1.4|        0.2|Iris-setosa|[4.9,3.0,1.4,0.2]|
|         4.7|        3.2|         1.3|        0.2|Iris-setosa|[4.7,3.2,1.3,0.2]|
|         4.6|        3.1|         1.5|        0.2|Iris-setosa|[4.6,3.1,1.5,0.2]|
|         5.0|        3.6|         1.4|        0.2|Iris-setosa|[5.0,3.6,1.4,0.2]|
|         5.4|        3.9|         1.7|        0.4|Iris-setosa|[5.4,3.9,1.7,0.4]|
|         4.6|        3.4|         1.4|        0.3|Iris-setosa|[4.6,3.4,1.4,0.3]|
|         5.0|        3.4|         1.5|        0.2|Iris-setosa|[5.0,3.4,1.5,0.2]|
|         4.4|        2.9|         1.4|        0.2|Iris-setosa|[4.4,2.9,1.4,0.2]|
|         4.9|        3.1|         1.5|        0.1|Iris-setosa|[4.9,3.1,1.5,0.1]|
|         5.4|        3.7|         1.5|        0.2|Iris-setosa|[5.4,3.7,1.5,0.2]|
|         4.8|        3.4|         1.6|        0.2|Iris-setosa|[4.8,3.4,1.6,0.2]|
|         4.8|        3.0|         1.4|        0.1|Iris-setosa|[4.8,3.0,1.4,0.1]|
|         4.3|        3.0|         1.1|        0.1|Iris-setosa|[4.3,3.0,1.1,0.1]|
|         5.8|        4.0|         1.2|        0.2|Iris-setosa|[5.8,4.0,1.2,0.2]|
|         5.7|        4.4|         1.5|        0.4|Iris-setosa|[5.7,4.4,1.5,0.4]|
|         5.4|        3.9|         1.3|        0.4|Iris-setosa|[5.4,3.9,1.3,0.4]|
|         5.1|        3.5|         1.4|        0.3|Iris-setosa|[5.1,3.5,1.4,0.3]|
|         5.7|        3.8|         1.7|        0.3|Iris-setosa|[5.7,3.8,1.7,0.3]|
|         5.1|        3.8|         1.5|        0.3|Iris-setosa|[5.1,3.8,1.5,0.3]|
+------------+-----------+------------+-----------+-----------+-----------------+
only showing top 20 rows
```

In [22]:

```python
from pyspark.ml.clustering import KMeans
from pyspark.ml.evaluation import ClusteringEvaluator
```

In [23]:

```python
kmeans =KMeans (featuresCol = "features", k=3)
```

In [24]:

```python
model = kmeans.fit(final_data)
```

In [25]:

```python
model
```

Out[25]:

```
KMeansModel: uid=KMeans_ebdb9a2b0f0e, k=3, distanceMeasure=euclidean, numFeatures=4
```

In [26]:

```python
model.transform(final_data).groupby("prediction").count().show()
```

```
+----------+-----+
|prediction|count|
+----------+-----+
|         1|   39|
|         2|   61|
|         0|   50|
+----------+-----+
```

In [27]:

```python
prediction = model.transform(final_data)
```

In [28]:

```python
prediction.show()
```

```
+------------+-----------+------------+-----------+-----------+-----------------+----------+
|sepal_length|sepal_width|petal_length|petal_width|    species|         features|prediction|
+------------+-----------+------------+-----------+-----------+-----------------+----------+
|         5.1|        3.5|         1.4|        0.2|Iris-setosa|[5.1,3.5,1.4,0.2]|         0|
|         4.9|        3.0|         1.4|        0.2|Iris-setosa|[4.9,3.0,1.4,0.2]|         0|
|         4.7|        3.2|         1.3|        0.2|Iris-setosa|[4.7,3.2,1.3,0.2]|         0|
|         4.6|        3.1|         1.5|        0.2|Iris-setosa|[4.6,3.1,1.5,0.2]|         0|
|         5.0|        3.6|         1.4|        0.2|Iris-setosa|[5.0,3.6,1.4,0.2]|         0|
|         5.4|        3.9|         1.7|        0.4|Iris-setosa|[5.4,3.9,1.7,0.4]|         0|
|         4.6|        3.4|         1.4|        0.3|Iris-setosa|[4.6,3.4,1.4,0.3]|         0|
|         5.0|        3.4|         1.5|        0.2|Iris-setosa|[5.0,3.4,1.5,0.2]|         0|
|         4.4|        2.9|         1.4|        0.2|Iris-setosa|[4.4,2.9,1.4,0.2]|         0|
|         4.9|        3.1|         1.5|        0.1|Iris-setosa|[4.9,3.1,1.5,0.1]|         0|
|         5.4|        3.7|         1.5|        0.2|Iris-setosa|[5.4,3.7,1.5,0.2]|         0|
|         4.8|        3.4|         1.6|        0.2|Iris-setosa|[4.8,3.4,1.6,0.2]|         0|
|         4.8|        3.0|         1.4|        0.1|Iris-setosa|[4.8,3.0,1.4,0.1]|         0|
|         4.3|        3.0|         1.1|        0.1|Iris-setosa|[4.3,3.0,1.1,0.1]|         0|
|         5.8|        4.0|         1.2|        0.2|Iris-setosa|[5.8,4.0,1.2,0.2]|         0|
|         5.7|        4.4|         1.5|        0.4|Iris-setosa|[5.7,4.4,1.5,0.4]|         0|
|         5.4|        3.9|         1.3|        0.4|Iris-setosa|[5.4,3.9,1.3,0.4]|         0|
|         5.1|        3.5|         1.4|        0.3|Iris-setosa|[5.1,3.5,1.4,0.3]|         0|
|         5.7|        3.8|         1.7|        0.3|Iris-setosa|[5.7,3.8,1.7,0.3]|         0|
|         5.1|        3.8|         1.5|        0.3|Iris-setosa|[5.1,3.8,1.5,0.3]|         0|
+------------+-----------+------------+-----------+-----------+-----------------+----------+
only showing top 20 rows
```

In [29]:

```python
prediction.groupby("species", "prediction").count().show()
```

```
+---------------+----------+-----+
|        species|prediction|count|
+---------------+----------+-----+
|Iris-versicolor|         2|   47|
|    Iris-setosa|         0|   50|
| Iris-virginica|         1|   36|
| Iris-virginica|         2|   14|
|Iris-versicolor|         1|    3|
+---------------+----------+-----+
```

In [ ]: