Name - Sarvesh Karanjkar

PRN - 20210812002

BDA LAB 04

```
pip install pyspark
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Collecting pyspark
  Downloading pyspark-3.3.2.tar.gz (281.4 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 281.4/281.4 MB 4.0 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Collecting py4j==0.10.9.5
  Downloading py4j-0.10.9.5-py2.py3-none-any.whl (199 kB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 199.7/199.7 KB 20.3 MB/s eta 0:00:00
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.3.2-py2.py3-none-any.whl size=281824025 sha256=3ba5c5078e74170cc27e881455b25301a346
  Stored in directory: /root/.cache/pip/wheels/6c/e3/9b/0525ce8a69478916513509d43693511463c6468db0de237c86
Successfully built pyspark
Installing collected packages: py4j, pyspark
Successfully installed py4j-0.10.9.5 pyspark-3.3.2
```

```
import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()

df = spark.read.csv("Data Set - Copy.csv")
df.show()
```

```
+---+-------------+------------+-------------+------------+-----------+
|_c0|          _c1|         _c2|          _c3|         _c4|        _c5|
+---+-------------+------------+-------------+------------+-----------+
| Id|SepalLengthCm|SepalWidthCm|PetalLengthCm|PetalWidthCm|    Species|
|  1|          5.1|         3.5|          1.4|         0.2|Iris-setosa|
|  2|          4.9|         3.0|          1.4|         0.2|Iris-setosa|
|  3|          4.7|         3.2|          1.3|         0.2|Iris-setosa|
|  4|          4.6|         3.1|          1.5|         0.2|Iris-setosa|
|  5|          5.0|         3.6|          1.4|         0.2|Iris-setosa|
|  6|          5.4|         3.9|          1.7|         0.4|Iris-setosa|
|  7|          4.6|         3.4|          1.4|         0.3|Iris-setosa|
|  8|          5.0|         3.4|          1.5|         0.2|Iris-setosa|
|  9|          4.4|         2.9|          1.4|         0.2|Iris-setosa|
| 10|          4.9|         3.1|          1.5|         0.1|Iris-setosa|
| 11|          5.4|         3.7|          1.5|         0.2|Iris-setosa|
| 12|          4.8|         3.4|          1.6|         0.2|Iris-setosa|
| 13|          4.8|         3.0|          1.4|         0.1|Iris-setosa|
| 14|          4.3|         3.0|          1.1|         0.1|Iris-setosa|
| 15|          5.8|         4.0|          1.2|         0.2|Iris-setosa|
| 16|          5.7|         4.4|          1.5|         0.4|Iris-setosa|
| 17|          5.4|         3.9|          1.3|         0.4|Iris-setosa|
| 18|          5.1|         3.5|          1.4|         0.3|Iris-setosa|
| 19|          5.7|         3.8|          1.7|         0.3|Iris-setosa|
+---+-------------+------------+-------------+------------+-----------+
only showing top 20 rows
```

```
#Checked datatypes
df.printSchema()
```

```
root
 |-- _c0: string (nullable = true)
 |-- _c1: string (nullable = true)
 |-- _c2: string (nullable = true)
 |-- _c3: string (nullable = true)
 |-- _c4: string (nullable = true)
 |-- _c5: string (nullable = true)
```

```
df.columns
```

```
['_c0', '_c1', '_c2', '_c3', '_c4', '_c5']
```

```
#Add column
df.withColumn("NEW_COL_01",df[0]).show()
```

```
+---+-------------+------------+-------------+------------+-----------+----------+
|_c0|          _c1|         _c2|          _c3|         _c4|        _c5|NEW_COL_01|
+---+-------------+------------+-------------+------------+-----------+----------+
| Id|SepalLengthCm|SepalWidthCm|PetalLengthCm|PetalWidthCm|    Species|        Id|
```

```
|  1|          5.1|          3.5|          1.4|          0.2|Iris-setosa|         1|
|  2|          4.9|          3.0|          1.4|          0.2|Iris-setosa|         2|
|  3|          4.7|          3.2|          1.3|          0.2|Iris-setosa|         3|
|  4|          4.6|          3.1|          1.5|          0.2|Iris-setosa|         4|
|  5|          5.0|          3.6|          1.4|          0.2|Iris-setosa|         5|
|  6|          5.4|          3.9|          1.7|          0.4|Iris-setosa|         6|
|  7|          4.6|          3.4|          1.4|          0.3|Iris-setosa|         7|
|  8|          5.0|          3.4|          1.5|          0.2|Iris-setosa|         8|
|  9|          4.4|          2.9|          1.4|          0.2|Iris-setosa|         9|
| 10|          4.9|          3.1|          1.5|          0.1|Iris-setosa|        10|
| 11|          5.4|          3.7|          1.5|          0.2|Iris-setosa|        11|
| 12|          4.8|          3.4|          1.6|          0.2|Iris-setosa|        12|
| 13|          4.8|          3.0|          1.4|          0.1|Iris-setosa|        13|
| 14|          4.3|          3.0|          1.1|          0.1|Iris-setosa|        14|
| 15|          5.8|          4.0|          1.2|          0.2|Iris-setosa|        15|
| 16|          5.7|          4.4|          1.5|          0.4|Iris-setosa|        16|
| 17|          5.4|          3.9|          1.3|          0.4|Iris-setosa|        17|
| 18|          5.1|          3.5|          1.4|          0.3|Iris-setosa|        18|
| 19|          5.7|          3.8|          1.7|          0.3|Iris-setosa|        19|
+---+-------------+-------------+-------------+-------------+-----------+----------+
only showing top 20 rows
```

```
#COLUMN RENAMED
df.withColumnRenamed("NEW_COL_01","RENAMED_COL_01").show()
```

```
+---+-------------+-------------+-------------+-------------+-----------+
|_c0|          _c1|          _c2|          _c3|          _c4|        _c5|
+---+-------------+-------------+-------------+-------------+-----------+
| Id|SepalLengthCm|SepalWidthCm|PetalLengthCm|PetalWidthCm|    Species|
|  1|          5.1|          3.5|          1.4|          0.2|Iris-setosa|
|  2|          4.9|          3.0|          1.4|          0.2|Iris-setosa|
|  3|          4.7|          3.2|          1.3|          0.2|Iris-setosa|
|  4|          4.6|          3.1|          1.5|          0.2|Iris-setosa|
|  5|          5.0|          3.6|          1.4|          0.2|Iris-setosa|
|  6|          5.4|          3.9|          1.7|          0.4|Iris-setosa|
|  7|          4.6|          3.4|          1.4|          0.3|Iris-setosa|
|  8|          5.0|          3.4|          1.5|          0.2|Iris-setosa|
|  9|          4.4|          2.9|          1.4|          0.2|Iris-setosa|
| 10|          4.9|          3.1|          1.5|          0.1|Iris-setosa|
| 11|          5.4|          3.7|          1.5|          0.2|Iris-setosa|
| 12|          4.8|          3.4|          1.6|          0.2|Iris-setosa|
| 13|          4.8|          3.0|          1.4|          0.1|Iris-setosa|
| 14|          4.3|          3.0|          1.1|          0.1|Iris-setosa|
| 15|          5.8|          4.0|          1.2|          0.2|Iris-setosa|
| 16|          5.7|          4.4|          1.5|          0.4|Iris-setosa|
| 17|          5.4|          3.9|          1.3|          0.4|Iris-setosa|
| 18|          5.1|          3.5|          1.4|          0.3|Iris-setosa|
| 19|          5.7|          3.8|          1.7|          0.3|Iris-setosa|
+---+-------------+-------------+-------------+-------------+-----------+
only showing top 20 rows
```

conclusion - Performed basic operations on dataframe using pyspark module.

Double-click (or enter) to edit

0s completed at 9:44 AM