Experiment-11

Q. SQL functions

1. create SparkContext and SparkSession.

→
```
from pyspark import SparkContext
from pyspark.sql import SparkSession.
spark = SparkSession.builder \
    .appName("Python Spark SQL basic example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

2. Load.

→ iris.csv and prostate.csv into databricks table folder.

3. Import both the above data.

→
```
iris = spark.read.csv("dbfs:/FileStore/tables/iris.csv",
                header = True, inferSchema = True)
iris.show(5)
```

→
```
prostate = spark.read.csv("dbfs:/FileStore/tables/
    prostate.csv", header = True, inferSchema = True)
prostate.show(5)
```

4. Import functions and types.

→
```
from pyspark.sql.functions import *
from pyspark.sql.types import *
```

5 abs

```
→ prostate.select('lpsa', abs(prostate.lpsa).alias('abs_lpsa')
                                              .show(5).
```

6 Array

```
→ df_arr = iris.select('species', array(['sepal_length',
      'sepal_width', 'petal_length', 'petal_width']).
      alias('features'))
  df_arr.show(5).
```

7 array_contains

```
→ df = df_arr.select('species', 'features', array_contains
          (df_arr.features, 1.4) alias('new features')).
  df.show(5).

→ df.filter(df.new features).show(5)
```

8. asc

```
→ prostate.sort(prostate.lpsa.asc()).show(5)
→ prostate.orderby(prostate.lpsa.asc()).show(5)
```

9 avg.

```
→ prostate.select(avg(prostate.lpsa)).show(5)
```

10. ceil

```
→ prostate.select('lpsa', ceil(prostate.lpsa)).show(5)
```

11. col

```
→ prostate.select(col('lcavol'), col('age')).show(5)
```

12. concat

```
→ df = spark.createDataframe([['a','1'], ['b','2']],
                                            ['x', 'y'])
  df.select('x', 'y', concat(df.x, df.y).alias('concat(x,y)')
                                                    .show()
```

13 collect_set

```
→ df.select(collect_set(df.x)).show()
```

14 concat_ws

```
→ df.select('x', 'y', concat_ws('_', df.x, df.y).alias(
                       'concat(x,y)')).show()
```

15 count

```
→ prostate.select(count(prostate.lpsa)).show()
```

16. countDistinct

```
→ iris.select(countDistinct(iris.species)).show()
```

17 create_map.

```
→ df = iris.select(create_map('species', 'sepal_length'))
  df.show()
```

18. curr_date

```
→ df = spark.createDataFrame([[1], [2], [3], [4], ['x'])
df.show()
df.select('x', current_date()).show)
```

19. current_timestamp

```
→ df.sellct('x', current_timestamp()).show(truncate=False)
```

20. date_add

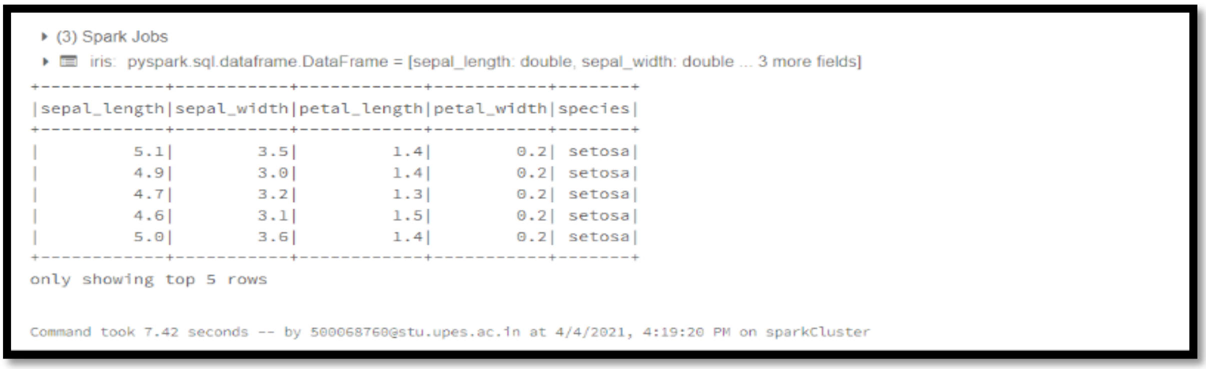
```
→ df2 = df.sellct('x', current_date().alias('current_date'))
df2.sellct('x', 'current_date', date_add(df2.current_date,
                                   10)).show()
```

21. date_format

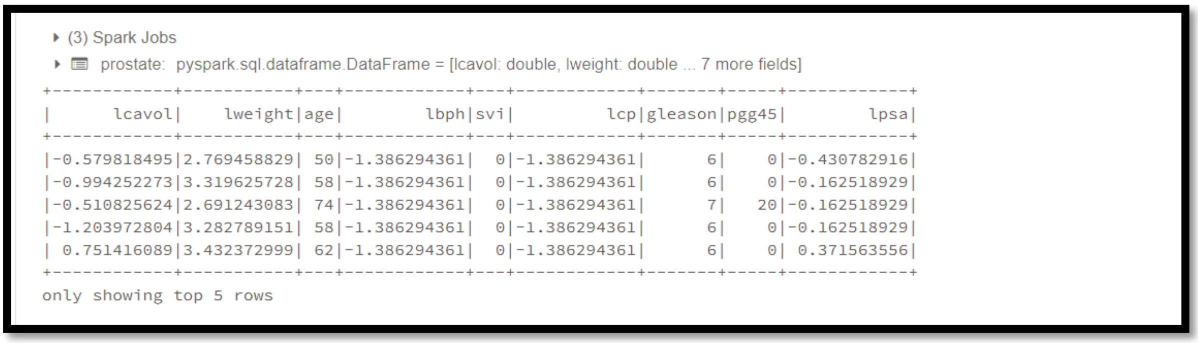
```
→ df2.sellct('x', 'current_date', date_format('current_date',
                       'MM/dd/yyyy').alias('new_data')).show
```

Outputs:

Iris || Prostate Datasets

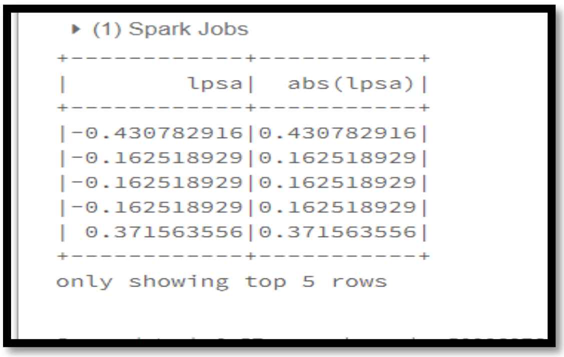


```
▸ (3) Spark Jobs
▸ ▤ iris:  pyspark.sql.dataframe.DataFrame = [sepal_length: double, sepal_width: double ... 3 more fields]
+------------+-----------+------------+-----------+-------+
|sepal_length|sepal_width|petal_length|petal_width|species|
+------------+-----------+------------+-----------+-------+
|         5.1|        3.5|         1.4|        0.2| setosa|
|         4.9|        3.0|         1.4|        0.2| setosa|
|         4.7|        3.2|         1.3|        0.2| setosa|
|         4.6|        3.1|         1.5|        0.2| setosa|
|         5.0|        3.6|         1.4|        0.2| setosa|
+------------+-----------+------------+-----------+-------+
only showing top 5 rows

Command took 7.42 seconds -- by 500068760@stu.upes.ac.in at 4/4/2021, 4:19:20 PM on sparkCluster
```



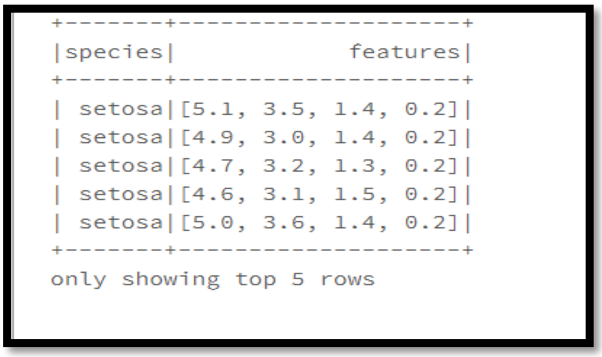
```
▸ (3) Spark Jobs
▸ ▤ prostate:  pyspark.sql.dataframe.DataFrame = [lcavol: double, lweight: double ... 7 more fields]
+------------+-----------+---+------------+---+------------+-------+-----+------------+
|      lcavol|    lweight|age|        lbph|svi|         lcp|gleason|pgg45|        lpsa|
+------------+-----------+---+------------+---+------------+-------+-----+------------+
|-0.579818495|2.769458829| 50|-1.386294361|  0|-1.386294361|      6|    0|-0.430782916|
|-0.994252273|3.319625728| 58|-1.386294361|  0|-1.386294361|      6|    0|-0.162518929|
|-0.510825624|2.691243083| 74|-1.386294361|  0|-1.386294361|      7|   20|-0.162518929|
|-1.203972804|3.282789151| 58|-1.386294361|  0|-1.386294361|      6|    0|-0.162518929|
| 0.751416089|3.432372999| 62|-1.386294361|  0|-1.386294361|      6|    0| 0.371563556|
+------------+-----------+---+------------+---+------------+-------+-----+------------+
only showing top 5 rows
```

Abs



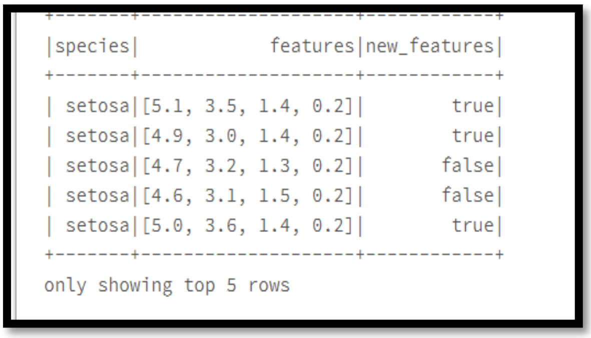
```
▸ (1) Spark Jobs
+------------+-----------+
|        lpsa|  abs(lpsa)|
+------------+-----------+
|-0.430782916|0.430782916|
|-0.162518929|0.162518929|
|-0.162518929|0.162518929|
|-0.162518929|0.162518929|
| 0.371563556|0.371563556|
+------------+-----------+
only showing top 5 rows
```

Array



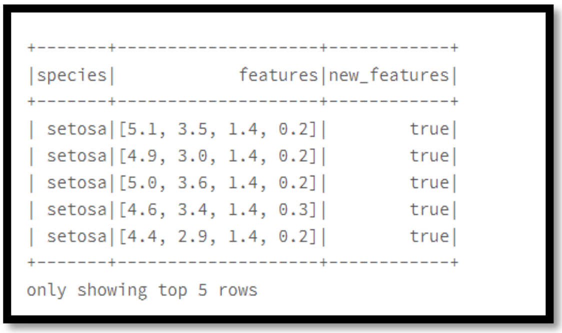
```
+-------+--------------------+
|species|            features|
+-------+--------------------+
| setosa|[5.1, 3.5, 1.4, 0.2]|
| setosa|[4.9, 3.0, 1.4, 0.2]|
| setosa|[4.7, 3.2, 1.3, 0.2]|
| setosa|[4.6, 3.1, 1.5, 0.2]|
| setosa|[5.0, 3.6, 1.4, 0.2]|
+-------+--------------------+
only showing top 5 rows
```

Array_contains



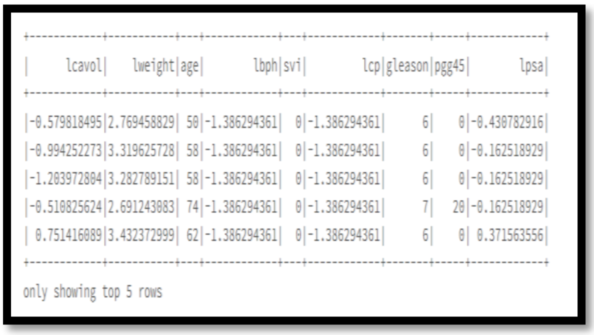
```
+-------+--------------------+------------+
|species|            features|new_features|
+-------+--------------------+------------+
| setosa|[5.1, 3.5, 1.4, 0.2]|        true|
| setosa|[4.9, 3.0, 1.4, 0.2]|        true|
| setosa|[4.7, 3.2, 1.3, 0.2]|       false|
| setosa|[4.6, 3.1, 1.5, 0.2]|       false|
| setosa|[5.0, 3.6, 1.4, 0.2]|        true|
+-------+--------------------+------------+
only showing top 5 rows
```

filter



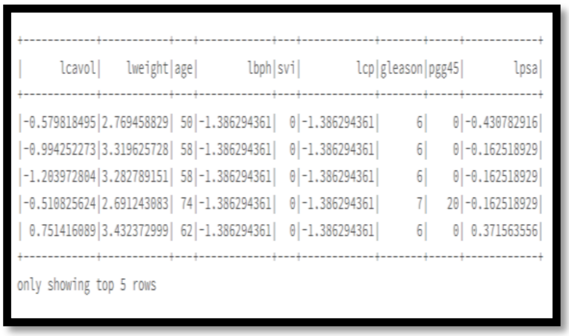
```
+-------+--------------------+------------+
|species|            features|new_features|
+-------+--------------------+------------+
| setosa|[5.1, 3.5, 1.4, 0.2]|        true|
| setosa|[4.9, 3.0, 1.4, 0.2]|        true|
| setosa|[5.0, 3.6, 1.4, 0.2]|        true|
| setosa|[4.6, 3.4, 1.4, 0.3]|        true|
| setosa|[4.4, 2.9, 1.4, 0.2]|        true|
+-------+--------------------+------------+
only showing top 5 rows
```

Asc using sort



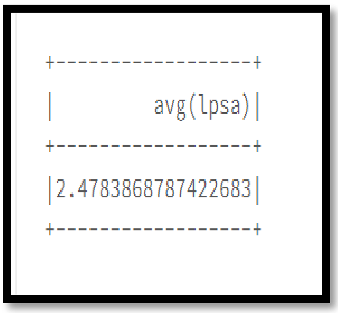
```
+------------+-----------+---+------------+---+------------+-------+-----+------------+
|      lcavol|    lweight|age|        lbph|svi|         lcp|gleason|pgg45|        lpsa|
+------------+-----------+---+------------+---+------------+-------+-----+------------+
|-0.579818495|2.769458829| 50|-1.386294361|  0|-1.386294361|      6|    0|-0.430782916|
|-0.994252273|3.319625728| 58|-1.386294361|  0|-1.386294361|      6|    0|-0.162518929|
|-1.203972804|3.282789151| 58|-1.386294361|  0|-1.386294361|      6|    0|-0.162518929|
|-0.510825624|2.691243083| 74|-1.386294361|  0|-1.386294361|      7|   20|-0.162518929|
| 0.751416089|3.432372999| 62|-1.386294361|  0|-1.386294361|      6|    0| 0.371563556|
+------------+-----------+---+------------+---+------------+-------+-----+------------+
only showing top 5 rows
```

Asc using orderBy



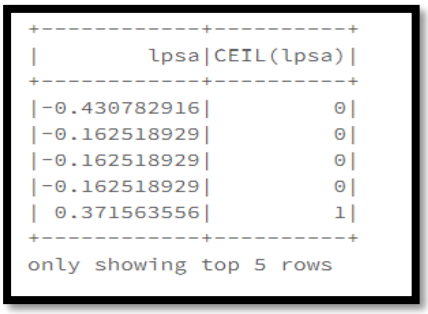
```
+------------+-----------+---+------------+---+------------+-------+-----+------------+
|      lcavol|    lweight|age|        lbph|svi|         lcp|gleason|pgg45|        lpsa|
+------------+-----------+---+------------+---+------------+-------+-----+------------+
|-0.579818495|2.769458829| 50|-1.386294361|  0|-1.386294361|      6|    0|-0.430782916|
|-0.994252273|3.319625728| 58|-1.386294361|  0|-1.386294361|      6|    0|-0.162518929|
|-1.203972804|3.282789151| 58|-1.386294361|  0|-1.386294361|      6|    0|-0.162518929|
|-0.510825624|2.691243083| 74|-1.386294361|  0|-1.386294361|      7|   20|-0.162518929|
| 0.751416089|3.432372999| 62|-1.386294361|  0|-1.386294361|      6|    0| 0.371563556|
+------------+-----------+---+------------+---+------------+-------+-----+------------+
only showing top 5 rows
```

Avg



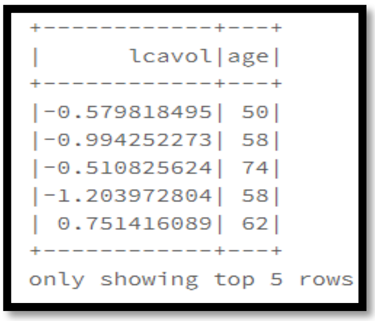
```
+------------------+
|         avg(lpsa)|
+------------------+
|2.4783868787422683|
+------------------+
```

Ceil



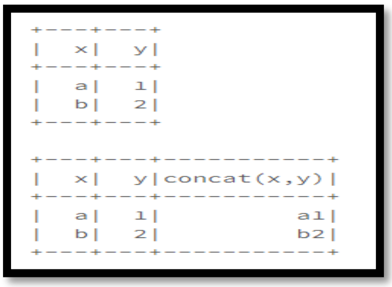
```
+------------+----------+
|        lpsa|CEIL(lpsa)|
+------------+----------+
|-0.430782916|         0|
|-0.162518929|         0|
|-0.162518929|         0|
|-0.162518929|         0|
| 0.371563556|         1|
+------------+----------+
only showing top 5 rows
```

Col



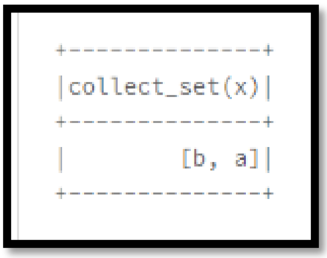
```
+------------+---+
|      lcavol|age|
+------------+---+
|-0.579818495| 50|
|-0.994252273| 58|
|-0.510825624| 74|
|-1.203972804| 58|
| 0.751416089| 62|
+------------+---+
only showing top 5 rows
```

Concat



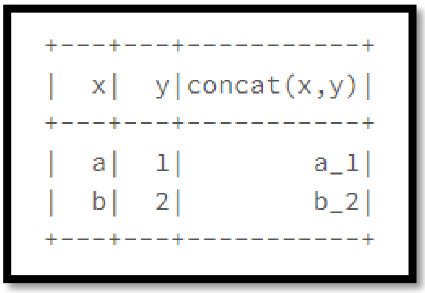
```
+---+---+
|  x|  y|
+---+---+
|  a|  1|
|  b|  2|
+---+---+

+---+---+------------+
|  x|  y|concat(x,y)|
+---+---+------------+
|  a|  1|          a1|
|  b|  2|          b2|
+---+---+------------+
```

Collect_set



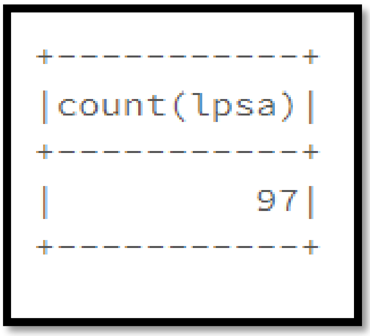
```
+--------------+
|collect_set(x)|
+--------------+
|        [b, a]|
+--------------+
```

Concat_ws



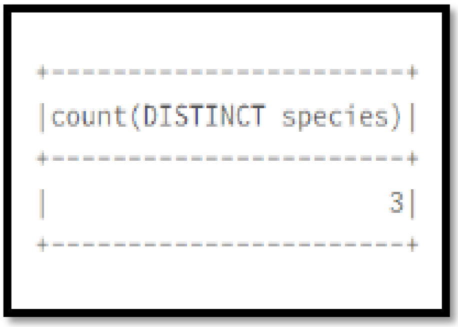
```
+---+---+-----------+
|  x|  y|concat(x,y)|
+---+---+-----------+
|  a|  1|        a_1|
|  b|  2|        b_2|
+---+---+-----------+
```

Count



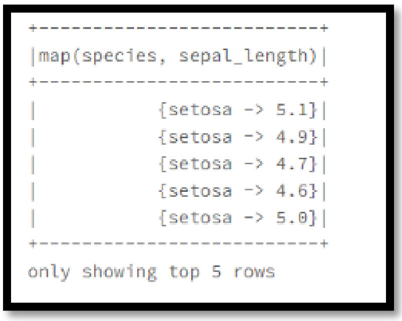
```
+-----------+
|count(lpsa)|
+-----------+
|         97|
+-----------+
```

CountDistinct



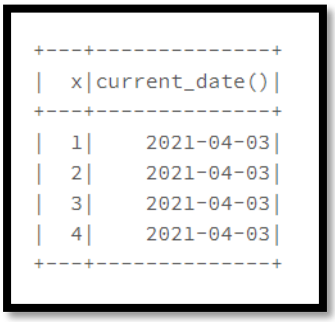
```
+-----------------------+
|count(DISTINCT species)|
+-----------------------+
|                      3|
+-----------------------+
```

Create_map



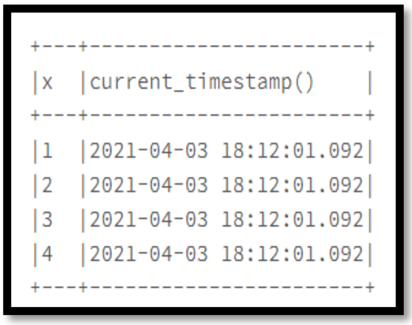
```
+---------------------------+
|map(species, sepal_length)|
+---------------------------+
|             {setosa -> 5.1}|
|             {setosa -> 4.9}|
|             {setosa -> 4.7}|
|             {setosa -> 4.6}|
|             {setosa -> 5.0}|
+---------------------------+
only showing top 5 rows
```

current_date



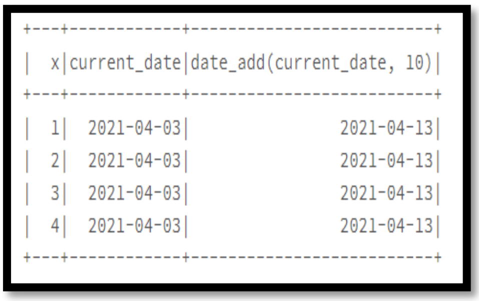
```
+---+--------------+
|  x|current_date()|
+---+--------------+
|  1|    2021-04-03|
|  2|    2021-04-03|
|  3|    2021-04-03|
|  4|    2021-04-03|
+---+--------------+
```

current_timestamp



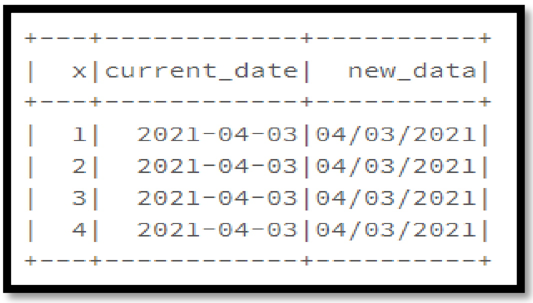
```
+---+-----------------------+
|x  |current_timestamp()    |
+---+-----------------------+
|1  |2021-04-03 18:12:01.092|
|2  |2021-04-03 18:12:01.092|
|3  |2021-04-03 18:12:01.092|
|4  |2021-04-03 18:12:01.092|
+---+-----------------------+
```

Date_add



```
+---+------------+------------------------+
|  x|current_date|date_add(current_date, 10)|
+---+------------+------------------------+
|  1|  2021-04-03|              2021-04-13|
|  2|  2021-04-03|              2021-04-13|
|  3|  2021-04-03|              2021-04-13|
|  4|  2021-04-03|              2021-04-13|
+---+------------+------------------------+
```

date_format



```
+---+------------+----------+
|  x|current_date|  new_data|
+---+------------+----------+
|  1|  2021-04-03|04/03/2021|
|  2|  2021-04-03|04/03/2021|
|  3|  2021-04-03|04/03/2021|
|  4|  2021-04-03|04/03/2021|
+---+------------+----------+
```