

```
In [1]: 1 from pyspark.sql import SparkSession
        2 spark = SparkSession.builder.appName('Filter_operations').getOrCreate()
        3 spark
```

**Out[1]: SparkSession - in-memory
SparkContext**

[Spark UI \(http://DESKTOP-5NN8I6U:4042\)](http://DESKTOP-5NN8I6U:4042)

Version

v3.5.1

Master

local[*]

AppName

Filter_operations

```
In [10]: 1 df=spark.read.csv(r'C:\Users\Neha\Downloads\NARESH IT DATA SCIENCE\MAY\
        2
        3
```

```
In [11]: 1 df
```

Out[11]: DataFrame[Name: string, age: int, Experience: int, Salary: int]

```
In [12]: 1 df.show()
```

```
+-----+---+-----+-----+
|   Name|age|Experience|Salary|
+-----+---+-----+-----+
|   jack| 31|         10| 30000|
|   alex| 30|          8| 25000|
|caroline| 29|          4| 20000|
|   paul| 24|          3| 20000|
|  sandra| 21|          1| 15000|
|casandra| 23|          2| 18000|
+-----+---+-----+-----+
```

```
In [13]: 1 df.printSchema()
```

```
root
 |-- Name: string (nullable = true)
 |-- age: integer (nullable = true)
 |-- Experience: integer (nullable = true)
 |-- Salary: integer (nullable = true)
```

```
In [15]: 1 df.columns
```

Out[15]: ['Name', 'age', 'Experience', 'Salary']

Vector Assembler:

```
In [20]: 1 from pyspark.ml.feature import VectorAssembler
2 feature_assembler = VectorAssembler(inputCols=['age', 'Experience'],
3                                     outputCol = 'Independent Features')
```

```
In [21]: 1 feature_assembler
```

```
Out[21]: VectorAssembler_ed2086a260f0
```

```
In [22]: 1 output= feature_assembler.transform(df)
```

```
In [23]: 1 output.show()
```

```
+-----+-----+-----+-----+-----+
|   Name|age|Experience|Salary|Independent Features|
+-----+-----+-----+-----+-----+
|   jack| 31|      10| 30000|      [31.0,10.0]|
|   alex| 30|       8| 25000|      [30.0,8.0]|
|caroline| 29|       4| 20000|      [29.0,4.0]|
|   paul| 24|       3| 20000|      [24.0,3.0]|
|  sandra| 21|       1| 15000|      [21.0,1.0]|
|casandra| 23|       2| 18000|      [23.0,2.0]|
+-----+-----+-----+-----+-----+
```

```
In [24]: 1 output.columns
```

```
Out[24]: ['Name', 'age', 'Experience', 'Salary', 'Independent Features']
```

```
In [25]: 1 finalised_data= output.select ('Independent Features', 'Salary')
2
```

```
In [26]: 1 finalised_data.show()
```

```
+-----+-----+
|Independent Features|Salary|
+-----+-----+
|      [31.0,10.0]| 30000|
|      [30.0,8.0]| 25000|
|      [29.0,4.0]| 20000|
|      [24.0,3.0]| 20000|
|      [21.0,1.0]| 15000|
|      [23.0,2.0]| 18000|
+-----+-----+
```

```
In [ ]: 1
```

Linear Regression Model:

```
In [72]: 1 from pyspark.ml.regression import LinearRegression
2 train_data, test_data = finalised_data.randomSplit([0.75,0.25])
3 regressor = LinearRegression(featuresCol='Independent Features', labelCol='Salary')
4 regressor = regressor.fit(train_data)
```

```
In [73]: 1 regressor
```

```
Out[73]: LinearRegressionModel: uid=LinearRegression_2d0ec61160d5, numFeatures=2
```

```
In [74]: 1 pred_results = regressor.evaluate(test_data)
```

```
In [75]: 1 pred_results.predictions.show()
```

```
+-----+-----+-----+
|Independent Features|Salary|      prediction|
+-----+-----+-----+
|      [21.0,1.0]    |15000| 19285.71428571376|
|      [23.0,2.0]    |18000| 18571.428571428507|
|      [29.0,4.0]    |20000| 12857.142857144696|
+-----+-----+-----+
```

Executors

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(1)	0	638.2 KiB / 366.3 MiB	0.0 B	8	0	0	58	58	25 min (4 s)	2.6 MiB	0.0 B	0.0 B	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(1)	0	638.2 KiB / 366.3 MiB	0.0 B	8	0	0	58	58	25 min (4 s)	2.6 MiB	0.0 B	0.0 B	0

Executors

Show 20 entries

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Thread Dump	Add Time	Remove Time
driver	DESKTOP-5NN8IGU50483	Active	0	638.2 KiB / 366.3 MiB	0.0 B	8	0	0	58	58	25 min (4 s)	2.6 MiB	0.0 B	0.0 B	Thread Dump	2024-05-04 21:48:30	-

```
In [ ]: 1
```