

Spark by {Examples} (https://sparkbyexamples.com/)

PySpark Tutorial

[PySpark Tutorial For Beginners \(https://sparkbyexamples.com/pyspark-tutorial/\)](https://sparkbyexamples.com/pyspark-tutorial/)

[PySpark – Features \(https://sparkbyexamples.com/pyspark-tutorial/#features\)](https://sparkbyexamples.com/pyspark-tutorial/#features)

[PySpark – Advantages \(https://sparkbyexamples.com/pyspark-tutorial/#advantages\)](https://sparkbyexamples.com/pyspark-tutorial/#advantages)

[PySpark – Modules & Packages \(https://sparkbyexamples.com/pyspark-tutorial/#modules-packages\)](https://sparkbyexamples.com/pyspark-tutorial/#modules-packages)

[PySpark – Cluster Managers \(https://sparkbyexamples.com/pyspark-tutorial/#cluster-manager\)](https://sparkbyexamples.com/pyspark-tutorial/#cluster-manager)

[PySpark – Install on Windows \(https://sparkbyexamples.com/pyspark-tutorial/#pyspark-installation\)](https://sparkbyexamples.com/pyspark-tutorial/#pyspark-installation)

[PySpark – Web/Application UI \(https://sparkbyexamples.com/spark/spark-web-ui-understanding/\)](https://sparkbyexamples.com/spark/spark-web-ui-understanding/)

[PySpark – SparkSession \(https://sparkbyexamples.com/pyspark/pyspark-what-is-sparksession/\)](https://sparkbyexamples.com/pyspark/pyspark-what-is-sparksession/)

[PySpark – RDD \(https://sparkbyexamples.com/pyspark-rdd\)](https://sparkbyexamples.com/pyspark-rdd)

[PySpark – Parallelize \(https://sparkbyexamples.com/pyspark/pyspark-parallelize-create-rdd/\)](https://sparkbyexamples.com/pyspark/pyspark-parallelize-create-rdd/)

[PySpark – repartition\(\) vs coalesce\(\) \(https://sparkbyexamples.com/pyspark/pyspark-repartition-vs-coalesce/\)](https://sparkbyexamples.com/pyspark/pyspark-repartition-vs-coalesce/)

[PySpark – Broadcast Variables \(https://sparkbyexamples.com/pyspark/pyspark-broadcast-variables/\)](https://sparkbyexamples.com/pyspark/pyspark-broadcast-variables/)

[PySpark \(https://sparkbyexamples.com/pyspark-tutorial/\)](https://sparkbyexamples.com/pyspark-tutorial/)

[Hive \(https://sparkbyexamples.com/apache-hive-tutorial/\)](https://sparkbyexamples.com/apache-hive-tutorial/)

[HBase \(https://sparkbyexamples.com/apache-hbase-tutorial/\)](https://sparkbyexamples.com/apache-hbase-tutorial/)

[Kafka \(https://sparkbyexamples.com/apache-kafka-tutorials-with-examples/\)](https://sparkbyexamples.com/apache-kafka-tutorials-with-examples/)

[PySpark \(https://sparkbyexamples.com/category/pyspark/\)](https://sparkbyexamples.com/category/pyspark/)

[PySpark provides built-in standard](https://sparkbyexamples.com/pyspark/)

questions/

[More \(https://sparkbyexamples.com/\)](https://sparkbyexamples.com/)

Setu Lean Lite

Checkout the latest products

Setu Nutrition
Setu Nutrition

PySpark Aggregate Functions with Examples

PySpark provides built-in standard Aggregate functions defines in DataFrame API, these come in handy when we need to make aggregate operations on DataFrame columns. Aggregate functions operate on a group of rows and calculate a single return value for every group.

Data Science Course Bangalore Certified

Datamites - Data Science Courses in I

6-month/400 learning hours, 120-hour training, capstone & client projects

 WEBSITE

 [

All these aggregate functions accept input as, Column type or column name in a string and several other arguments based on the function and return Column type.



[PySpark – Accumulator
\(https://sparkbyexamples.com/pyspark/pyspark-accumulator-with-example/\)](https://sparkbyexamples.com/pyspark/pyspark-accumulator-with-example/)

PySpark DataFrame

[PySpark – Create a DataFrame
\(https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/\)](https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/)

[PySpark – Create an empty DataFrame
\(https://sparkbyexamples.com/pyspark/pyspark-create-an-empty-dataframe/\)](https://sparkbyexamples.com/pyspark/pyspark-create-an-empty-dataframe/)

[PySpark – Convert RDD to DataFrame
\(https://sparkbyexamples.com/pyspark/convert-pyspark-rdd-to-dataframe/\)](https://sparkbyexamples.com/pyspark/convert-pyspark-rdd-to-dataframe/)

[PySpark – Convert DataFrame to Pandas
\(https://sparkbyexamples.com/pyspark/convert-pyspark-dataframe-to-pandas/\)](https://sparkbyexamples.com/pyspark/convert-pyspark-dataframe-to-pandas/)

[PySpark – show\(\)
\(https://sparkbyexamples.com/pyspark/pyspark-show-display-dataframe-contents-in-table/\)](https://sparkbyexamples.com/pyspark/pyspark-show-display-dataframe-contents-in-table/)

[PySpark – StructType & StructField
\(https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/\)](https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/)

[PySpark – Row Class
\(https://sparkbyexamples.com/pyspark/pyspark-row-using-rdd-dataframe/\)](https://sparkbyexamples.com/pyspark/pyspark-row-using-rdd-dataframe/)

[PySpark – Column Class
\(https://sparkbyexamples.com/pyspark/pyspark-column-functions/\)](https://sparkbyexamples.com/pyspark/pyspark-column-functions/)

[PySpark – select\(\)
\(https://sparkbyexamples.com/pyspark/select-columns-from-pyspark-dataframe/\)](https://sparkbyexamples.com/pyspark/select-columns-from-pyspark-dataframe/)

[PySpark – collect\(\)
\(https://sparkbyexamples.com/pyspark/pyspark-collect/\)](https://sparkbyexamples.com/pyspark/pyspark-collect/)

[PySpark – withColumn\(\)
\(https://sparkbyexamples.com/pyspark/pyspark-withcolumn/\)](https://sparkbyexamples.com/pyspark/pyspark-withcolumn/)

When possible try to leverage standard library as they are little bit more compile-time safety, handles null and perform better when compared to UDF's. If your application is critical on performance try to avoid using custom UDF at all costs as these are not guarantee on performance.

PySpark Aggregate

Functions

PySpark SQL Aggregate functions are grouped as “agg_funcs” in Pyspark. Below is a list of functions defined under this group. Click on each link to learn with example.

- [approx_count_distinct](#)
- [avg](#)
- [collect_list](#)
- [collect_set](#)
- [countDistinct](#)
- [count](#)
- [grouping](#)
- [first](#)
- [last](#)
- [kurtosis](#)
- [max](#)
- [min](#)
- [mean](#)
- [skewness](#)
- [stddev](#)
- [stddev_samp](#)
- [stddev_pop](#)
- [sum](#)
- [sumDistinct](#)
- [variance](#), [var_samp](#), [var_pop](#)

PySpark Aggregate

Functions Examples

First, let's [create a DataFrame \(https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/\)](https://sparkbyexamples.com/pyspark/different-ways-to-create-dataframe-in-pyspark/) to work with PySpark aggregate functions. All examples provided here are also available at [PySpark Examples GitHub \(https://github.com/spark-](https://github.com/spark-)

[PySpark – withColumnRenamed\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-rename-dataframe-column/>).

[PySpark – where\(\) & filter\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-where-filter/>).

[PySpark – drop\(\) & dropDuplicates\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-distinct-to-drop-duplicates/>).

[PySpark – orderBy\(\) and sort\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-orderby-and-sort-explained/>).

[PySpark – groupBy\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-groupby-explained-with-example/>).

[PySpark – join\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-join-explained-with-examples/>).

[PySpark – union\(\) & unionAll\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-union-and-unionall/>).

[PySpark – unionByName\(\).](#)
(<https://sparkbyexamples.com/spark/spark-merge-two-dataframes-with-different-columns/>).

[PySpark – UDF \(User Defined Function\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-udf-user-defined-function/>).

[PySpark – map\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-map-transformation/>).

[PySpark – flatMap\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-flatmap-transformation/>).

[pyspark – foreach\(\).](#)
(<https://sparkbyexamples.com/pyspark/pyspark-loop-iterate-through-rows-in-dataframe/#use-foreach-loop-through-dataframe>).

[PySpark – sample\(\) vs sampleBy\(\).](#)

[examples/pyspark-examples/blob/master/pyspark-aggregate.py](#))project.

Data Science Course Bangalore Certified

Datamites - Data Science Courses in I

6-month/400 learning hours, 120-hour training, capstone & client projects



```
simpleData = [("James", "Sales", 3000),
              ("Michael", "Sales", 4600),
              ("Robert", "Sales", 4100),
              ("Maria", "Finance", 3000),
              ("James", "Sales", 3000),
              ("Scott", "Finance", 3300),
              ("Jen", "Finance", 3900),
              ("Jeff", "Marketing", 3000),
              ("Kumar", "Marketing", 2000),
              ("Saif", "Sales", 4100)]

schema = ["employee_name", "department", "salary"]
df = spark.createDataFrame(simpleData, schema)
df.printSchema()
df.show(truncate=False)
```

Yields below output.

```
+-----+-----+-----+
|employee_name|department|salary|
+-----+-----+-----+
|James|Sales|3000|
|Michael|Sales|4600|
|Robert|Sales|4100|
|Maria|Finance|3000|
|James|Sales|3000|
|Scott|Finance|3300|
|Jen|Finance|3900|
|Jeff|Marketing|3000|
|Kumar|Marketing|2000|
|Saif|Sales|4100|
+-----+-----+-----+
```

Now let's see how to aggregate data in PySpark.

[\(https://sparkbyexamples.com/pyspark/pyspark-sampling-example/\)](https://sparkbyexamples.com/pyspark/pyspark-sampling-example/).

[PySpark – fillna\(\) & fill\(\). \(https://sparkbyexamples.com/pyspark/pyspark-fillna-fill-replace-null-values/\)](https://sparkbyexamples.com/pyspark/pyspark-fillna-fill-replace-null-values/).

[PySpark – pivot\(\)_\(Row to Column\). \(https://sparkbyexamples.com/pyspark/pyspark-pivot-and-unpivot-dataframe/\)](https://sparkbyexamples.com/pyspark/pyspark-pivot-and-unpivot-dataframe/).

[PySpark – partitionBy\(\). \(https://sparkbyexamples.com/pyspark/pyspark-partitionby-example/\)](https://sparkbyexamples.com/pyspark/pyspark-partitionby-example/).

[PySpark – ArrayType Column \(Array\). \(https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/\)](https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/).

[PySpark – MapType \(Map/Dict\). \(https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/\)](https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/).

PySpark SQL Functions

[PySpark – Aggregate Functions \(https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/\)](https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/).

[PySpark – Window Functions \(https://sparkbyexamples.com/pyspark/pyspark-window-functions/\)](https://sparkbyexamples.com/pyspark/pyspark-window-functions/).

[PySpark – Date and Timestamp Functions \(https://sparkbyexamples.com/pyspark/pyspark-sql-date-and-timestamp-functions/\)](https://sparkbyexamples.com/pyspark/pyspark-sql-date-and-timestamp-functions/).

[PySpark – JSON Functions \(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/\)](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/).

PySpark Datasources

[PySpark – Read & Write CSV File \(https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/\)](https://sparkbyexamples.com/pyspark/pyspark-read-csv-file-into-dataframe/).

[PySpark – Read & Write Parquet File \(https://sparkbyexamples.com/pyspark/pyspark-read-write-parquet-file/\)](https://sparkbyexamples.com/pyspark/pyspark-read-write-parquet-file/).

approx_count_distinct

Aggregate Function

In PySpark
approx_count_distinct() function returns the count of distinct items in a group.

```
//approx_count_distinct()
print("approx_count_distinct: "
      str(df.select(approx_count_distinct('col')).collect())

//Prints approx_count_distinct: 3
```

avg (average) Aggregate Function

avg() function returns the average of values in the input column.

```
//avg
print("avg: " + str(df.select(avg('col')).collect())

//Prints avg: 3400.0
```

collect_list Aggregate Function

collect_list() function returns all values from an input column with duplicates.

Time Tracking Software

Best employee time utilization. View application usage manage cost.

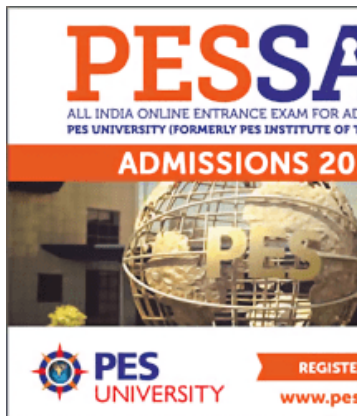
ProHance Analytics

Open



[pyspark/pyspark-read-and-write-parquet-file/](#)

[PySpark – Read & Write JSON file](#)
(<https://sparkbyexamples.com/pyspark/pyspark-read-json-file-into-dataframe/>)



PySpark Built-In Functions

[PySpark – when\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-when-otherwise/>)

[PySpark – expr\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-sql-expr-expression-function/>)

[PySpark – lit\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/>)

[PySpark – split\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-convert-string-to-array-column/>)

[PySpark – concat_ws\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-convert-array-column-to-string-column/>)

[Pyspark – substring\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-substring-from-a-column/>)

[PySpark – translate\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/#translate-replace-character-by-character>)

[PySpark – regexp_replace\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-replace-column->

```
//collect_list
df.select(collect_list("salary")

+-----+
|collect_list(salary)|
+-----+
|[3000, 4600, 4100, 3000, 3000,
```

collect_set Aggregate Function

collect_set() function returns all values from an input column with duplicate values eliminated.

```
//collect_set
df.select(collect_set("salary")

+-----+
|collect_set(salary)|
+-----+
|[4600, 3000, 3900, 4100, 3300,
```

countDistinct Aggregate Function

countDistinct() function returns the number of distinct elements in a columns

```
//countDistinct
df2 = df.select(countDistinct("c
df2.show(truncate=False)
print("Distinct Count of Depart
```

count function

count() function returns number of elements in a column.

[values/#regex_replace-replace-string-columns](#)).

[PySpark – overlay\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-replace-column-values/#overlay-function>)

[PySpark – to_timestamp\(\)](#)
(https://sparkbyexamples.com/spark/pyspark-to_timestamp-convert-string-to-timestamp-type/)

[PySpark – to_date\(\)](#)
(https://sparkbyexamples.com/pyspark/pyspark-to_date-convert-timestamp-to-date/)

[PySpark – date_format\(\)](#)
(https://sparkbyexamples.com/pyspark/pyspark-date_format-convert-date-to-string-format/)

[PySpark – datediff\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-difference-between-two-dates-days-months-years/#datediff>)

[PySpark – months_between\(\)](#)
([https://sparkbyexamples.com/pyspark/pyspark-difference-between-two-dates-days-months-years/#months_between\(\)](https://sparkbyexamples.com/pyspark/pyspark-difference-between-two-dates-days-months-years/#months_between()))

[PySpark – explode\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-explode-nested-array-into-rows/>)

[PySpark – array_contains\(\)](#)
(https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/#array_contains)

[PySpark – array\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-arraytype-column-with-examples/#array>)

[PySpark – collect_list\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/#collect-list>)

[PySpark – collect_set\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/#collect-set>)

[PySpark – create_map\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-convert-dataframe-columns-to-maptype-dict/>)

```
print("count: "+str(df.select(c  
  
Prints county: 10
```

grouping function

`grouping()` Indicates whether a given input column is aggregated or not. returns 1 for aggregated or 0 for not aggregated in the result. If you try grouping directly on the salary column you will get below error.

```
Exception in thread "main" org.  
// grouping() can only be used
```

first function

`first()` function returns the first element in a column when `ignoreNulls` is set to true, it returns the first non-null element.

```
//first  
df.select(first("salary")).show  
  
+-----+  
|first(salary, false)|  
+-----+  
|3000                |  
+-----+
```

last function

`last()` function returns the last element in a column. when `ignoreNulls` is set to true, it returns the last non-null element.

[PySpark – map_keys\(\)](#)
(https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/#map_keys)

[PySpark – map_values\(\)](#)
(https://sparkbyexamples.com/pyspark/pyspark-maptype-dict-examples/#map_values)

[PySpark – struct\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-structtype-and-structfield/#update-struct-function>)

[PySpark – countDistinct\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-count-distinct-from-dataframe/>)

[PySpark – sum\(\).avg\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-dataframe-groupby-and-sort-by-descending-order/>)

[PySpark – row_number\(\)](#)
(https://sparkbyexamples.com/pyspark/pyspark-window-functions/#row_number)

[PySpark – rank\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-window-functions/#rank>)

[PySpark – dense_rank\(\)](#)
(https://sparkbyexamples.com/pyspark/pyspark-window-functions/#dense_rank)

[PySpark – percent_rank\(\)](#)
(https://sparkbyexamples.com/pyspark/pyspark-window-functions/#percent_rank)

[PySpark – typedLit\(\)](#)
(<https://sparkbyexamples.com/pyspark/pyspark-lit-add-literal-constant/#typedlit>)

[PySpark – from_json\(\)](#)
(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#from_json)

[PySpark – to_json\(\)](#)
(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#to_json)

[PySpark – json_tuple\(\)](#)
(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#json_tuple)

```
//last
df.select(last("salary")).show()
```

```
+-----+
|last(salary, false)|
+-----+
|4100           |
+-----+
```

kurtosis function

kurtosis() function returns the kurtosis of the values in a group.

```
df.select(kurtosis("salary")).show()
```

```
+-----+
|kurtosis(salary) |
+-----+
|-0.64678030303032 |
+-----+
```

max function

max() function returns the maximum value in a column.

```
df.select(max("salary")).show()
```

```
+-----+
|max(salary) |
+-----+
|4600           |
+-----+
```

min function

min() function

[PySpark – get_json_object\(\)
\(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#get_json_object\)](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#get_json_object)

[PySpark – schema_of_json\(\)
\(https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#schema_of_json\)](https://sparkbyexamples.com/pyspark/pyspark-json-functions-with-examples/#schema_of_json)

```
df.select(min("salary")).show(t

+-----+
|min(salary)|
+-----+
| 2000      |
+-----+
```

mean function

mean() function returns the average of the values in a column. Alias for Avg

```
df.select(mean("salary")).show(

+-----+
|avg(salary)|
+-----+
| 3400.0     |
+-----+
```

skewness function

skewness() function returns the skewness of the values in a group.

```
df.select(skewness("salary")).s

+-----+
|skewness(salary)|
+-----+
|-0.12041791181069571|
+-----+
```

stddev(), stddev_samp() and stddev_pop()

stddev() alias for stddev_samp.

stddev_samp() function returns the sample standard deviation of values in a column.

`stddev_pop()` function returns the population standard deviation of the values in a column.

```
df.select(stddev("salary"), stddev_pop("salary")).show(truncate=False)
```

stddev_samp(salary)	stddev_pop(salary)
765.9416862050705	765.9416862050705

sum function

`sum()` function Returns the sum of all values in a column.

```
df.select(sum("salary")).show(truncate=False)
```

sum(salary)
34000

sumDistinct function

`sumDistinct()` function returns the sum of all distinct values in a column.

```
df.select(sumDistinct("salary")).show(truncate=False)
```

sum(DISTINCT salary)
20900

variance(), var_samp(), var_pop()

`variance()` alias for `var_samp`

var_samp() function returns the unbiased variance of the values in a column.

var_pop() function returns the population variance of the values in a column.

```
df.select(variance("salary"),var_pop("salary"))\
    .show(truncate=False)
```

var_samp(salary)	var_pop(salary)
586666.6666666666	586666.6666666666

Source code of PySpark
Aggregate examples

```

import pyspark
from pyspark.sql import SparkSession
from pyspark.sql.functions import col
from pyspark.sql.functions import countDistinct
from pyspark.sql.functions import sum
from pyspark.sql.functions import sumDistinct
from pyspark.sql.functions import variance

spark = SparkSession.builder.appName("Salary Analysis").getOrCreate()

simpleData = [("James", "Sales", 4600),
              ("Michael", "Sales", 4600),
              ("Robert", "Sales", 4100),
              ("Maria", "Finance", 3000),
              ("James", "Sales", 3000),
              ("Scott", "Finance", 3300),
              ("Jen", "Finance", 3900),
              ("Jeff", "Marketing", 3000),
              ("Kumar", "Marketing", 2000),
              ("Saif", "Sales", 4100)]

schema = ["employee_name", "department", "salary"]

df = spark.createDataFrame(data=simpleData, schema=schema)
df.printSchema()
df.show(truncate=False)

print("approx_count_distinct: " + str(df.select(approx_count_distinct("salary")).show(truncate=False)))

print("avg: " + str(df.select(avg("salary")).show(truncate=False)))

print(df.select(collect_list("salary")).show(truncate=False))

print(df.select(collect_set("salary")).show(truncate=False))

df2 = df.select(countDistinct("salary"))
df2.show(truncate=False)
print("Distinct Count of Department Salary")

print("count: " + str(df.select(count("salary")).show(truncate=False)))
print(df.select(first("salary")).show(truncate=False))
print(df.select(last("salary")).show(truncate=False))
print(df.select(kurtosis("salary")).show(truncate=False))
print(df.select(max("salary")).show(truncate=False))
print(df.select(min("salary")).show(truncate=False))
print(df.select(mean("salary")).show(truncate=False))
print(df.select(skewness("salary")).show(truncate=False))
print(df.select(stddev("salary"), stddev_pop("salary")).show(truncate=False))
print(df.select(sum("salary")).show(truncate=False))
print(df.select(sumDistinct("salary")).show(truncate=False))
print(df.select(variance("salary"), variance_samp("salary")).show(truncate=False))

```

Conclusion

In this article, I've consolidated and listed all PySpark Aggregate functions with scala examples and also learned the benefits of using PySpark SQL functions.

Happy Learning !!

Share this:



(<https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/?share=facebook&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/?share=reddit&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/?share=pinterest&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/?share=tumblr&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/?share=pocket&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/?share=linkedin&nb=1>)



(<https://sparkbyexamples.com/pyspark/pyspark-aggregate-functions/?share=twitter&nb=1>)

TAGS: [AGGREGATE FUNCTIONS](#)
(<https://sparkbyexamples.com/tag/aggregate-functions/>)



[NNK](#)

(<https://Sparkbyexamples.Com/Author/Admin/>)

(<https://sparkbyexamples.com/author/admin/>)

SparkByExamples.com is a Big Data and Spark examples community page, all examples are simple and easy to understand and well tested in our development environment [Read more ...](#)
(<https://sparkbyexamples.com/about-sparkbyexamples/>)

> THIS POST HAS ONE COMMENT



Tw

30 JAN 2021 [REPLY](#)



Thank you for all your efforts in putting PySpark operations together. It is very helpful.

Leave a Reply

Categories

Apache Hadoop
(<https://sparkbyexamples.com/category/hadoop/>)

Apache Spark
(<https://sparkbyexamples.com/category/spark/>)

Apache Spark Streaming
(<https://sparkbyexamples.com/category/spark/apache-spark-streaming/>)

Apache Kafka
(<https://sparkbyexamples.com/category/kafka/>)

Recent Posts

Spark regexp_replace() – Replace String Value
(https://sparkbyexamples.com/spark/spark-regexp_replace-replace-string-value/)

How to Run a PySpark Script from Python?
(<https://sparkbyexamples.com/pyspark/run-pyspark-script-from-python-subprocess/>)

Spark SQL like() Using Wildcard Example
(<https://sparkbyexamples.com/spark/spark-sql-like-using-wildcard-example/>)

About SparkByExamples.Com

SparkByExamples.com is a Big Data and Spark examples community page, all examples are simple and easy to understand, and well tested in our development environment Read more ..
(<https://sparkbyexamples.com/about-sparkbyexamples/>)

Follow Us



Apache HBase
(<https://sparkbyexamples.com/category/hbase/>)

Apache Cassandra
(<https://sparkbyexamples.com/category/cassandra/>)

Snowflake Database
(<https://sparkbyexamples.com/category/snowflake/>)

H2O Sparkling Water
(<https://sparkbyexamples.com/category/h2o-sparkling-water/>)

PySpark
(<https://sparkbyexamples.com/category/pyspark/>)

Spark isin() & IS NOT IN Operator Example
(<https://sparkbyexamples.com/spark/spark-isin-is-not-in-operator-example/>)

Spark – Get Size/Length of Array & Map Column
(<https://sparkbyexamples.com/spark/spark-get-size-length-of-array-map-column/>)

Spark Using Length/Size Of a DataFrame Column
(<https://sparkbyexamples.com/spark/spark-using-length-size-of-a-dataframe-column/>)

Spark rlike() Working with Regex Matching Examples
(<https://sparkbyexamples.com/spark/spark-rlike-regex-matching-examples/>)

Spark Check String Column Has Numeric Values
(<https://sparkbyexamples.com/spark/spark-check-string-column-has-numeric-values/>)

Spark Check Column Data Type is Integer or String
(<https://sparkbyexamples.com/spark/spark-check-column-data-type-is-integer-or-string/>)

(<https://www.facebook.com/sparkbyexamples/>)

(<https://www.linkedin.com/company/sparkbyexamples/>)

(<https://twitter.com/sparkbyexamples>)

(<https://www.youtube.com/channel/UC8193l83R3v2W361W5DncA>)

(<https://github.com/sparkbyexamples>)

(<https://medium.com/@sparkbyexamples>)

(<https://www.b860a.com/examples/>)

(<https://www.scribd.com/document/81931931/Spark-Examples>)

