# How to calculate the mean of a dataframe column and find the top 10%

I am very new to Scala and Spark, and am working on some self-made exercises using baseball statistics. I am using a case class create a RDD and assign a schema to the data, and am then turning it into a DataFrame so I can use SparkSQL to select groups of players via their stats that meet certain criteria.

Once I have the subset of players I am interested in looking at further, I would like to find the mean of a column; eg Batting Average or RBIs. From there I would like to break all the players into percentile groups based on their average performance compared to all players; the top 10%, bottom 10%, 40-50%

I've been able to use the DataFrame.describe() function to return a summary of a desired column (mean, stddev, count, min, and max) all as strings though. Is there a better way to get just the mean and stddev as Doubles, and what is the best way of breaking the players into groups of 10-percentiles?

So far my thoughts are to find the values that bookend the percentile ranges and writing a function that groups players via comparators, but that feels like it is bordering on reinventing the wheel.

I have the following imports currently:

```
import org.apache.spark.rdd.RDD
import org.apache.spark.sql.SQLContext
import org.apache.spark.{SparkConf, SparkContext}
import org.joda.time.format.DateTimeFormat
```

apache-spark scala

edited Jul 26 '15 at 21:49
Marmite Bomber
563 1 2 9

asked Jul 22 '15 at 14:16
the3rdNotch
28 1 6

Have you checked the scaladoc? It has an example for average and max: `.agg(avg(people("salary")), max(people("age")))` . With sorting you can probably find (using `skip` and `take` ) the percentiles, but there might be faster options. – Gábor Bakos Jul 22 '15 at 16:47

I had seen this previously in the scaladocs. When I try to use them like the example I receive and error `not found: value avg` and `not found: value max` – the3rdNotch Jul 22 '15 at 18:46

What are your imports? It might be easier to help if there is an example and you describe what were the problem. – Gábor Bakos Jul 22 '15 at 18:48

`import org.apache.spark.rdd.RDD import org.apache.spark.sql.SQLContext import org.apache.spark.{SparkConf, SparkContext} import org.joda.time.format.DateTimeFormat` – the3rdNotch Jul 22 '15 at 19:02

The following test might help start using DataFrame functions. It seems you have to import the `org.apache.spark.sql.functions._` too. (BTW.: I think the additional information is better added to the question itself and it is enough to add a comment after edit.) – Gábor Bakos Jul 22 '15 at 19:11

## 1 Answer

This is the import you need, and how to get the mean for a column named "RBIs":

```
import org.apache.spark.sql.functions._
df.select(avg($"RBIs")).show()
```

For the standard deviation, see scala - Calculate the standard deviation of grouped data in a Spark DataFrame - Stack Overflow

For grouping by percentiles, I suggest defining a new column via a user-defined function (UDF), and using groupBy on that column. See

- Spark SQL and DataFrames - Spark 1.5.1 Documentation - udf registration

edited Oct 21 '15 at 15:18 answered Oct 21 '15 at 15:00

nealmcb
**208** 2 6