Here are the columns that are present in thedatasets:

Building.csv –BuildingID, BuildingMgr, BuildingAge,HVACproduct,Country

HVAC.csv –Date, Time, TargetTemp, ActualTemp, System, SystemAge,BuildingID

Working with SensorData

.Load HVAC.csv file into temporarytable

.Add a new column, tempchange -set to 1, if there is a change of greater than +/-5 betweenactual

and target temperature

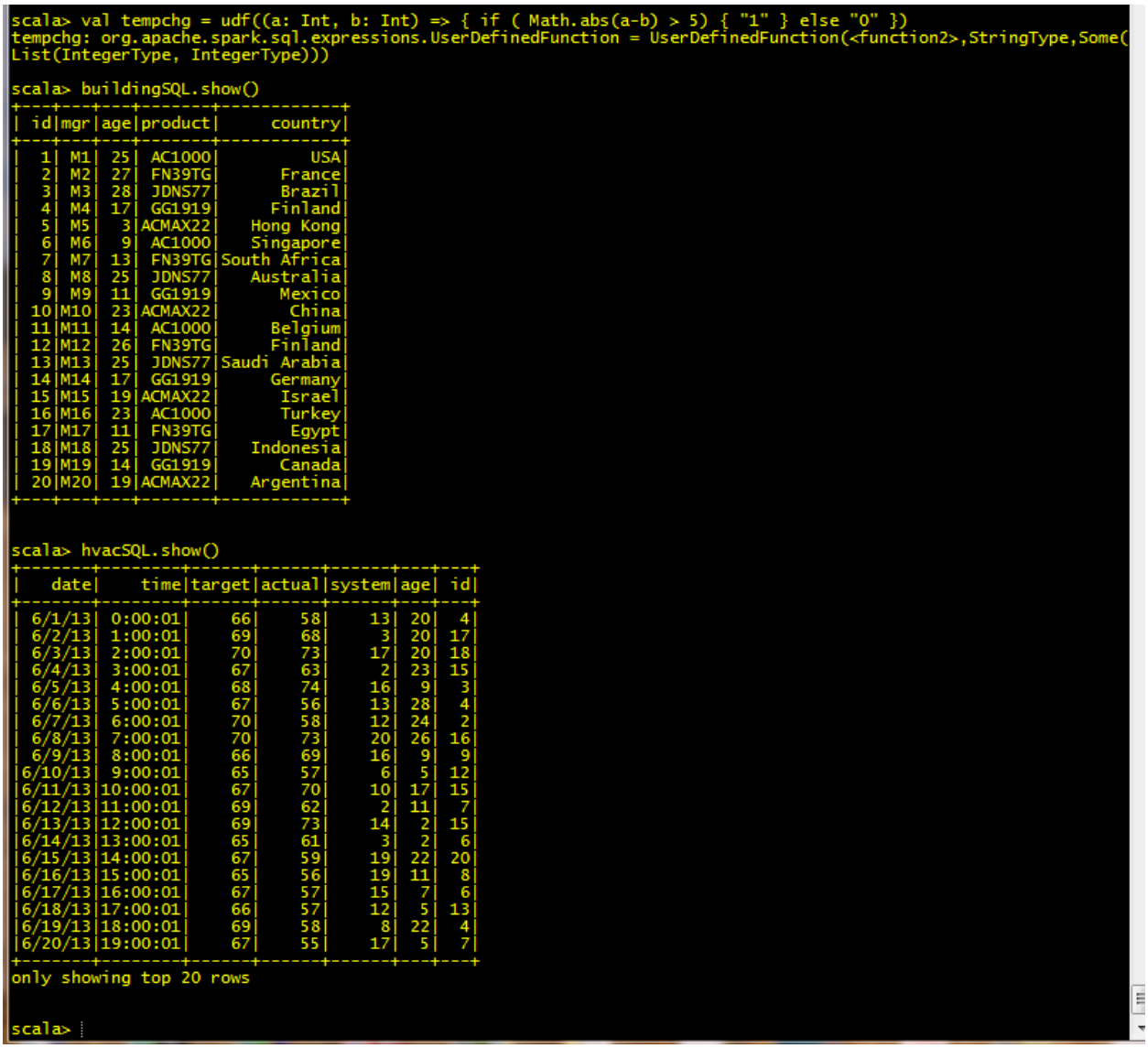
STARTER CODE:

```
import org.apache.spark.sql.functions.udf

import org.apache.spark.sql.catalyst.encoders.ExpressionEncoder

import org.apache.spark.sql.Encoder

import org.apache.spark.sql.{Row, SparkSession}

import org.apache.spark.sql.types.{DoubleType, StringType, StructField, StructType}

import spark.implicits._

case class Building(id: Int, mgr: String, age: Int, product: String, country: String)

case class Hvac(date: String, time: String, target: Int, actual: Int, system: Int, age: Int, id: Int)

val buildingDF = sc.textFile("/Session21/buildingNC.csv")

val buildingDFmap = buildingDF.map(_.split(","))

val buildingDFdata = buildingDFmap.map(attributes =>
Building(attributes(0).toInt,attributes(1),attributes(2).toInt,attributes(3),attributes(4).trim))

val buildingSQL = buildingDFdata.toDF()

val HVACDF = sc.textFile("/Session21/HVACNC.csv")

val HVACDFmap = HVACDF.map(_.split(","))

val HVACDFdata = HVACDFmap.map(attributes =>
Hvac(attributes(0),attributes(1),attributes(2).toInt,attributes(3).toInt,attributes(4).toInt,attributes(5).toInt,attributes(6).trim.toInt))

val hvacSQL = HVACDFdata.toDF()
```

```
scala>

scala> val hvacSQL = HVACDFdata.toDF()
hvacSQL: org.apache.spark.sql.DataFrame = [date: string, time: string ... 5 more fields]

scala> buildingSQL.show()
+---+---+---+-------+------------+
| id|mgr|age|product|     country|
+---+---+---+-------+------------+
|  1| M1| 25| AC1000|         USA|
|  2| M2| 27| FN39TG|      France|
|  3| M3| 28| JDNS77|      Brazil|
|  4| M4| 17| GG1919|     Finland|
|  5| M5|  3|ACMAX22|   Hong Kong|
|  6| M6|  9| AC1000|   Singapore|
|  7| M7| 13| FN39TG|South Africa|
|  8| M8| 25| JDNS77|   Australia|
|  9| M9| 11| GG1919|      Mexico|
| 10|M10| 23|ACMAX22|       China|
| 11|M11| 14| AC1000|     Belgium|
| 12|M12| 26| FN39TG|     Finland|
| 13|M13| 25| JDNS77|Saudi Arabia|
| 14|M14| 17| GG1919|     Germany|
| 15|M15| 19|ACMAX22|      Israel|
| 16|M16| 23| AC1000|      Turkey|
| 17|M17| 11| FN39TG|       Egypt|
| 18|M18| 25| JDNS77|   Indonesia|
| 19|M19| 14| GG1919|      Canada|
| 20|M20| 19|ACMAX22|   Argentina|
+---+---+---+-------+------------+


scala> havcSQL.show()
<console>:112: error: not found: value havcSQL
       havcSQL.show()
       ^

scala> hvacSQL.show()
+-------+--------+------+------+------+---+---+
|   date|    time|target|actual|system|age| id|
+-------+--------+------+------+------+---+---+
| 6/1/13| 0:00:01|    66|    58|    13| 20|  4|
| 6/2/13| 1:00:01|    69|    68|     3| 20| 17|
| 6/3/13| 2:00:01|    70|    73|    17| 20| 18|
| 6/4/13| 3:00:01|    67|    63|     2| 23| 15|
| 6/5/13| 4:00:01|    68|    74|    16|  9|  3|
| 6/6/13| 5:00:01|    67|    56|    13| 28|  4|
| 6/7/13| 6:00:01|    70|    58|    12| 24|  2|
| 6/8/13| 7:00:01|    70|    73|    20| 26| 16|
| 6/9/13| 8:00:01|    66|    69|    16|  9|  9|
|6/10/13| 9:00:01|    65|    57|     6|  5| 12|
|6/11/13|10:00:01|    67|    70|    10| 17| 15|
|6/12/13|11:00:01|    69|    62|     2| 11|  7|
|6/13/13|12:00:01|    69|    73|    14|  2| 15|
|6/14/13|13:00:01|    65|    61|     3|  2|  6|
|6/15/13|14:00:01|    67|    59|    19| 22| 20|
|6/16/13|15:00:01|    65|    56|    19| 11|  8|
|6/17/13|16:00:01|    67|    57|    15|  7|  6|
|6/18/13|17:00:01|    66|    57|    12|  5| 13|
|6/19/13|18:00:01|    69|    58|     8| 22|  4|
|6/20/13|19:00:01|    67|    55|    17|  5|  7|
+-------+--------+------+------+------+---+---+
only showing top 20 rows


scala>
```

Define UDF

val tempchg = udf((a: Int, b: Int) => { if (Math.abs(a-b) > 5) { "1" } else "0" })



```
scala> val tempchg = udf((a: Int, b: Int) => { if ( Math.abs(a-b) > 5) { "1" } else "0" })
tempchg: org.apache.spark.sql.expressions.UserDefinedFunction = UserDefinedFunction(<function2>,StringType,Some(
List(IntegerType, IntegerType)))

scala> buildingSQL.show()
+---+---+---+-------+------------+
| id|mgr|age|product|     country|
+---+---+---+-------+------------+
|  1| M1| 25| AC1000|         USA|
|  2| M2| 27| FN39TG|      France|
|  3| M3| 28| JDNS77|      Brazil|
|  4| M4| 17| GG1919|     Finland|
|  5| M5|  3|ACMAX22|   Hong Kong|
|  6| M6|  9| AC1000|   Singapore|
|  7| M7| 13| FN39TG|South Africa|
|  8| M8| 25| JDNS77|   Australia|
|  9| M9| 11| GG1919|      Mexico|
| 10|M10| 23|ACMAX22|       China|
| 11|M11| 14| AC1000|     Belgium|
| 12|M12| 26| FN39TG|     Finland|
| 13|M13| 25| JDNS77|Saudi Arabia|
| 14|M14| 17| GG1919|     Germany|
| 15|M15| 19|ACMAX22|      Israel|
| 16|M16| 23| AC1000|      Turkey|
| 17|M17| 11| FN39TG|       Egypt|
| 18|M18| 25| JDNS77|   Indonesia|
| 19|M19| 14| GG1919|      Canada|
| 20|M20| 19|ACMAX22|   Argentina|
+---+---+---+-------+------------+


scala> hvacSQL.show()
+-------+--------+------+------+------+---+---+
|   date|    time|target|actual|system|age| id|
+-------+--------+------+------+------+---+---+
| 6/1/13| 0:00:01|    66|    58|    13| 20|  4|
| 6/2/13| 1:00:01|    69|    68|     3| 20| 17|
| 6/3/13| 2:00:01|    70|    73|    17| 20| 18|
| 6/4/13| 3:00:01|    67|    63|     2| 23| 15|
| 6/5/13| 4:00:01|    68|    74|    16|  9|  3|
| 6/6/13| 5:00:01|    67|    56|    13| 28|  4|
| 6/7/13| 6:00:01|    70|    58|    12| 24|  2|
| 6/8/13| 7:00:01|    70|    73|    20| 26| 16|
| 6/9/13| 8:00:01|    66|    69|    16|  9|  9|
|6/10/13| 9:00:01|    65|    57|     6|  5| 12|
|6/11/13|10:00:01|    67|    70|    10| 17| 15|
|6/12/13|11:00:01|    69|    62|     2| 11|  7|
|6/13/13|12:00:01|    69|    73|    14|  2| 15|
|6/14/13|13:00:01|    65|    61|     3|  2|  6|
|6/15/13|14:00:01|    67|    59|    19| 22| 20|
|6/16/13|15:00:01|    65|    56|    19| 11|  8|
|6/17/13|16:00:01|    67|    57|    15|  7|  6|
|6/18/13|17:00:01|    66|    57|    12|  5| 13|
|6/19/13|18:00:01|    69|    58|     8| 22|  4|
|6/20/13|19:00:01|    67|    55|    17|  5|  7|
+-------+--------+------+------+------+---+---+
only showing top 20 rows

scala>
```

```
scala> val tempDiff = hvacSQL.withColumn("More than 5degrees", tempchg($"target",$"actual"))
tempDiff: org.apache.spark.sql.DataFrame = [date: string, time: string ... 6 more fields]

scala> tempDiff.show()
+-------+--------+------+------+------+---+---+------------------+
|   date|    time|target|actual|system|age| id|More than 5degrees|
+-------+--------+------+------+------+---+---+------------------+
| 6/1/13| 0:00:01|    66|    58|    13| 20|  4|                 1|
| 6/2/13| 1:00:01|    69|    68|     3| 20| 17|                 0|
| 6/3/13| 2:00:01|    70|    73|    17| 20| 18|                 0|
| 6/4/13| 3:00:01|    67|    63|     2| 23| 15|                 0|
| 6/5/13| 4:00:01|    68|    74|    16|  9|  3|                 1|
| 6/6/13| 5:00:01|    67|    56|    13| 28|  4|                 1|
| 6/7/13| 6:00:01|    70|    58|    12| 24|  2|                 1|
| 6/8/13| 7:00:01|    70|    73|    20| 26| 16|                 0|
| 6/9/13| 8:00:01|    66|    69|    16|  9|  9|                 0|
|6/10/13| 9:00:01|    65|    57|     6|  5| 12|                 1|
|6/11/13|10:00:01|    67|    70|    10| 17| 15|                 0|
|6/12/13|11:00:01|    69|    62|     2| 11|  7|                 1|
|6/13/13|12:00:01|    69|    73|    14|  2| 15|                 0|
|6/14/13|13:00:01|    65|    61|     3|  2|  6|                 0|
|6/15/13|14:00:01|    67|    59|    19| 22| 20|                 1|
|6/16/13|15:00:01|    65|    56|    19| 11|  8|                 1|
|6/17/13|16:00:01|    67|    57|    15|  7|  6|                 1|
|6/18/13|17:00:01|    66|    57|    12|  5| 13|                 1|
|6/19/13|18:00:01|    69|    58|     8| 22|  4|                 1|
|6/20/13|19:00:01|    67|    55|    17|  5|  7|                 1|
+-------+--------+------+------+------+---+---+------------------+
only showing top 20 rows

scala>
```

Load building.csv file into temporary table

```
scala> buildingSQL.show()
+---+---+---+-------+------------+
| id|mgr|age|product|     country|
+---+---+---+-------+------------+
|  1| M1| 25| AC1000|         USA|
|  2| M2| 27| FN39TG|      France|
|  3| M3| 28| JDNS77|      Brazil|
|  4| M4| 17| GG1919|     Finland|
|  5| M5|  3|ACMAX22|   Hong Kong|
|  6| M6|  9| AC1000|   Singapore|
|  7| M7| 13| FN39TG|South Africa|
|  8| M8| 25| JDNS77|   Australia|
|  9| M9| 11| GG1919|      Mexico|
| 10|M10| 23|ACMAX22|       China|
| 11|M11| 14| AC1000|     Belgium|
| 12|M12| 26| FN39TG|     Finland|
| 13|M13| 25| JDNS77|Saudi Arabia|
| 14|M14| 17| GG1919|     Germany|
| 15|M15| 19|ACMAX22|      Israel|
| 16|M16| 23| AC1000|      Turkey|
| 17|M17| 11| FN39TG|       Egypt|
| 18|M18| 25| JDNS77|   Indonesia|
| 19|M19| 14| GG1919|      Canada|
| 20|M20| 19|ACMAX22|   Argentina|
+---+---+---+-------+------------+
```

Figure out the number of times temperature has changed by 5 degrees or more for each country

Joining the two tables

```
scala> tempDiff.show()
+-------+--------+------+------+------+---+---+-----------+
|   date|    time|target|actual|system|age| id|fivedegrees|
+-------+--------+------+------+------+---+---+-----------+
| 6/1/13| 0:00:01|    66|    58|    13| 20|  4|          1|
| 6/2/13| 1:00:01|    69|    68|     3| 20| 17|          0|
| 6/3/13| 2:00:01|    70|    73|    17| 20| 18|          0|
| 6/4/13| 3:00:01|    67|    63|     2| 23| 15|          0|
| 6/5/13| 4:00:01|    68|    74|    16|  9|  3|          1|
| 6/6/13| 5:00:01|    67|    56|    13| 28|  4|          1|
| 6/7/13| 6:00:01|    70|    58|    12| 24|  2|          1|
| 6/8/13| 7:00:01|    70|    73|    20| 26| 16|          0|
| 6/9/13| 8:00:01|    66|    69|    16|  9|  9|          0|
|6/10/13| 9:00:01|    65|    57|     6|  5| 12|          1|
|6/11/13|10:00:01|    67|    70|    10| 17| 15|          0|
|6/12/13|11:00:01|    69|    62|     2| 11|  7|          1|
|6/13/13|12:00:01|    69|    73|    14|  2| 15|          0|
|6/14/13|13:00:01|    65|    61|     3|  2|  6|          0|
|6/15/13|14:00:01|    67|    59|    19| 22| 20|          1|
|6/16/13|15:00:01|    65|    56|    19| 11|  8|          1|
|6/17/13|16:00:01|    67|    57|    15|  7|  6|          1|
|6/18/13|17:00:01|    66|    57|    12|  5| 13|          1|
|6/19/13|18:00:01|    69|    58|     8| 22|  4|          1|
|6/20/13|19:00:01|    67|    55|    17|  5|  7|          1|
+-------+--------+------+------+------+---+---+-----------+
only showing top 20 rows


scala> buildingSQL.show()
+---+---+---+-------+------------+
| id|mgr|age|product|     country|
+---+---+---+-------+------------+
|  1| M1| 25| AC1000|         USA|
|  2| M2| 27| FN39TG|      France|
|  3| M3| 28| JDNS77|      Brazil|
|  4| M4| 17| GG1919|     Finland|
|  5| M5|  3|ACMAX22|   Hong Kong|
|  6| M6|  9| AC1000|   Singapore|
|  7| M7| 13| FN39TG|South Africa|
|  8| M8| 25| JDNS77|   Australia|
|  9| M9| 11| GG1919|      Mexico|
| 10|M10| 23|ACMAX22|       China|
| 11|M11| 14| AC1000|     Belgium|
| 12|M12| 26| FN39TG|     Finland|
| 13|M13| 25| JDNS77|Saudi Arabia|
| 14|M14| 17| GG1919|     Germany|
| 15|M15| 19|ACMAX22|      Israel|
| 16|M16| 23| AC1000|      Turkey|
| 17|M17| 11| FN39TG|       Egypt|
| 18|M18| 25| JDNS77|   Indonesia|
| 19|M19| 14| GG1919|      Canada|
| 20|M20| 19|ACMAX22|   Argentina|
+---+---+---+-------+------------+


scala>
```

Two tables joined

```
scala> val countrySQL = buildingSQL.join(tempDiff, "id")
countrySQL: org.apache.spark.sql.DataFrame = [id: int, mgr: string ... 10 more fields]

scala> countrySQL.show()
+---+---+---+-------+-------+-------+--------+------+------+------+---+-----------+
| id|mgr|age|product|country|   date|    time|target|actual|system|age|fivedegrees|
+---+---+---+-------+-------+-------+--------+------+------+------+---+-----------+
| 12|M12| 26| FN39TG|Finland|6/10/13| 9:00:01|    65|    57|     6|  5|          1|
| 12|M12| 26| FN39TG|Finland|6/18/13|23:13:19|    66|    75|     1| 13|          1|
| 12|M12| 26| FN39TG|Finland| 6/2/13|13:43:51|    65|    72|    20| 26|          1|
| 12|M12| 26| FN39TG|Finland|6/13/13| 0:13:20|    67|    77|     8| 19|          1|
| 12|M12| 26| FN39TG|Finland|6/16/13| 3:13:20|    67|    55|    11| 16|          1|
| 12|M12| 26| FN39TG|Finland|6/30/13|17:13:20|    65|    57|    17|  9|          1|
| 12|M12| 26| FN39TG|Finland| 6/1/13|18:13:20|    68|    65|     7| 21|          0|
| 12|M12| 26| FN39TG|Finland|6/25/13|18:33:07|    70|    66|    20| 20|          0|
| 12|M12| 26| FN39TG|Finland|6/17/13|16:00:01|    69|    68|    16|  4|          0|
| 12|M12| 26| FN39TG|Finland| 6/5/13|16:43:51|    69|    69|    19| 15|          0|
| 12|M12| 26| FN39TG|Finland|6/23/13|10:13:20|    65|    61|     1|  1|          0|
| 12|M12| 26| FN39TG|Finland|6/29/13|16:13:20|    67|    80|    12|  8|          1|
| 12|M12| 26| FN39TG|Finland| 6/4/13|21:13:20|    66|    72|     7|  1|          1|
| 12|M12| 26| FN39TG|Finland| 6/3/13| 2:00:01|    69|    72|     7| 21|          0|
| 12|M12| 26| FN39TG|Finland|6/16/13|15:00:01|    67|    77|     4| 22|          1|
| 12|M12| 26| FN39TG|Finland|6/22/13|21:00:01|    70|    77|    13| 12|          1|
| 12|M12| 26| FN39TG|Finland|6/26/13| 7:43:51|    65|    62|     6|  6|          0|
| 12|M12| 26| FN39TG|Finland|6/26/13|13:13:20|    65|    63|    20|  9|          0|
| 12|M12| 26| FN39TG|Finland|6/30/13|17:13:20|    66|    62|    14| 26|          0|
| 12|M12| 26| FN39TG|Finland|6/10/13| 3:33:07|    70|    78|     5|  9|          1|
+---+---+---+-------+-------+-------+--------+------+------+------+---+-----------+
only showing top 20 rows


scala> |
```

Filtering only five degrees

```
scala> val country5SQL = countrySQL.filter($"fivedegrees" === "1")
country5SQL: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [id: int, mgr: string ... 10 more fields]

scala> country5SQL.show()
+---+---+---+-------+-------+-------+--------+------+------+------+---+-----------+
| id|mgr|age|product|country|   date|    time|target|actual|system|age|fivedegrees|
+---+---+---+-------+-------+-------+--------+------+------+------+---+-----------+
| 12|M12| 26| FN39TG|Finland|6/10/13| 9:00:01|    65|    57|     6|  5|          1|
| 12|M12| 26| FN39TG|Finland|6/18/13|23:13:19|    66|    75|     1| 13|          1|
| 12|M12| 26| FN39TG|Finland| 6/2/13|13:43:51|    65|    72|    20| 26|          1|
| 12|M12| 26| FN39TG|Finland|6/13/13| 0:13:20|    67|    77|     8| 19|          1|
| 12|M12| 26| FN39TG|Finland|6/16/13| 3:13:20|    67|    55|    11| 16|          1|
| 12|M12| 26| FN39TG|Finland|6/30/13|17:13:20|    65|    57|    17|  9|          1|
| 12|M12| 26| FN39TG|Finland|6/29/13|16:13:20|    67|    80|    12|  8|          1|
| 12|M12| 26| FN39TG|Finland| 6/4/13|21:13:20|    66|    72|     7|  1|          1|
| 12|M12| 26| FN39TG|Finland|6/16/13|15:00:01|    67|    77|     4| 22|          1|
| 12|M12| 26| FN39TG|Finland|6/22/13|21:00:01|    70|    77|    13| 12|          1|
| 12|M12| 26| FN39TG|Finland|6/10/13| 3:33:07|    70|    78|     5|  9|          1|
| 12|M12| 26| FN39TG|Finland| 6/1/13| 0:00:01|    69|    58|    18|  8|          1|
| 12|M12| 26| FN39TG|Finland|6/27/13| 8:43:51|    69|    57|    19| 23|          1|
| 12|M12| 26| FN39TG|Finland| 6/6/13|23:13:20|    65|    57|     8|  5|          1|
| 12|M12| 26| FN39TG|Finland|6/22/13|21:45:56|    66|    60|     2|  5|          1|
| 12|M12| 26| FN39TG|Finland|6/19/13| 0:45:56|    65|    75|     9| 22|          1|
| 12|M12| 26| FN39TG|Finland|6/26/13|18:00:01|    67|    56|     1| 28|          1|
| 12|M12| 26| FN39TG|Finland| 6/6/13|10:43:51|    65|    78|    20| 24|          1|
| 12|M12| 26| FN39TG|Finland|6/11/13|15:43:51|    66|    80|    11|  6|          1|
| 12|M12| 26| FN39TG|Finland|6/12/13| 4:00:01|    69|    61|     6| 20|          1|
+---+---+---+-------+-------+-------+--------+------+------+------+---+-----------+
only showing top 20 rows
```

Group by country and count

```
scala> val country5Count = country5SQL.groupBy("country").count()
country5Count: org.apache.spark.sql.DataFrame = [country: string, count: bigint]

scala> country5Count.show()
+------------+-----+
|     country|count|
+------------+-----+
|   Singapore|  230|
|      Turkey|  243|
|     Germany|  196|
|      France|  251|
|   Argentina|  230|
|     Belgium|  199|
|     Finland|  473|
|       China|  241|
|   Hong Kong|  248|
|      Israel|  232|
|         USA|  213|
|      Mexico|  228|
|   Indonesia|  243|
|Saudi Arabia|  233|
|      Canada|  232|
|      Brazil|  226|
|   Australia|  225|
|       Egypt|  236|
|South Africa|  237|
+------------+-----+


scala>
```