# REQUIRED

You are provided with Craft Brewery data to perform basic data analysis using Spark. The data includes three files:

* Beers.csv
* Breweries.csv
* Dictionary.doc

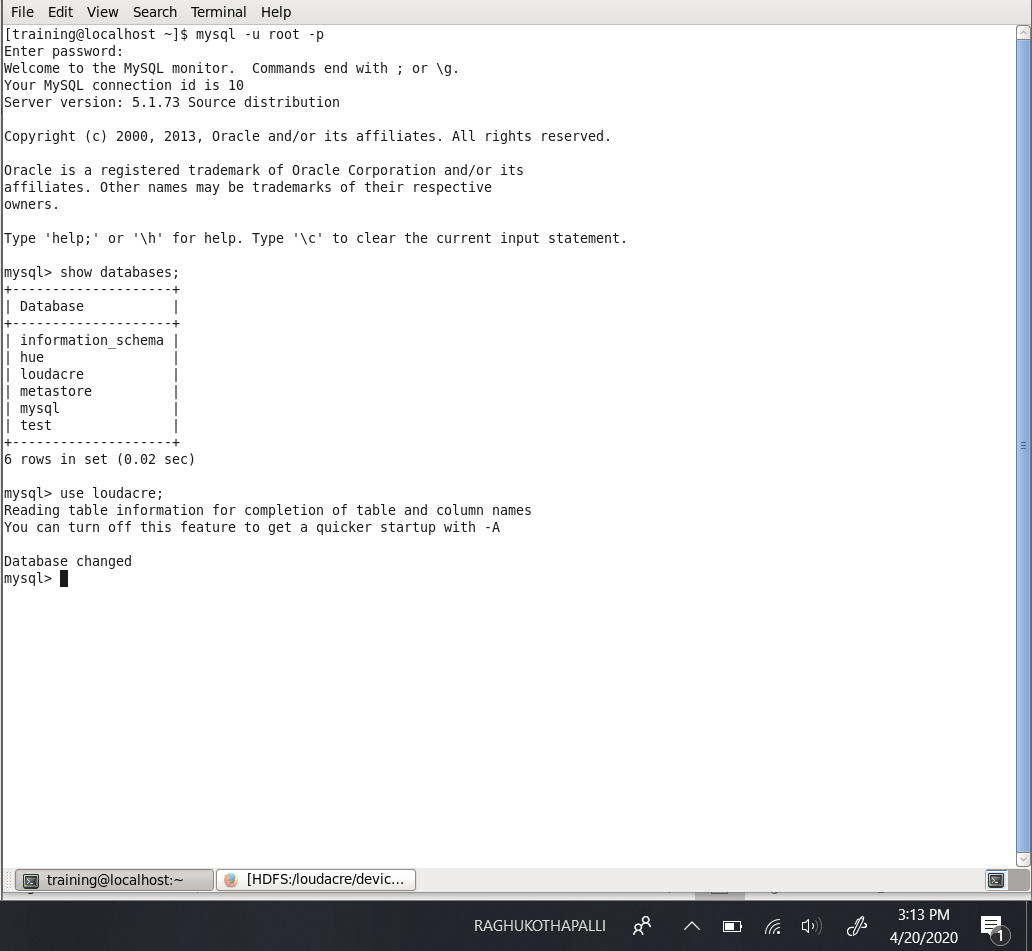
1. Import the brewery.csv into Sqoop. Move the brewery.csv into Spark from Sqoop (intermediate movement to another component in Hadoop is fine)

Answer

Going step by step firstly, to import the file named “brewery.csv” into Sqoop we need to follow the steps

**Using a database**

use the existing database “loudacre”



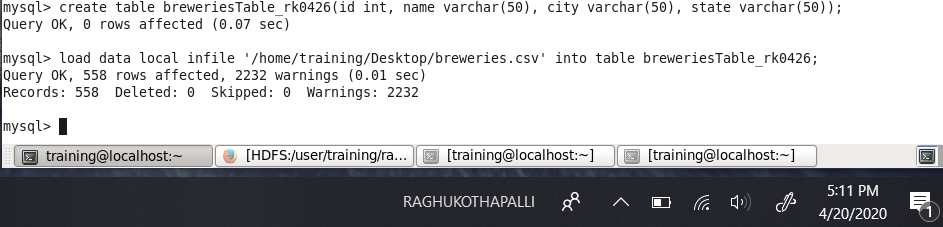
**Creating a table in the database selected**

create table breweriesTable\_rk0426 (id int, name varchar(50), city varchar (50), state varchar(20));

**Insert data into the table created**

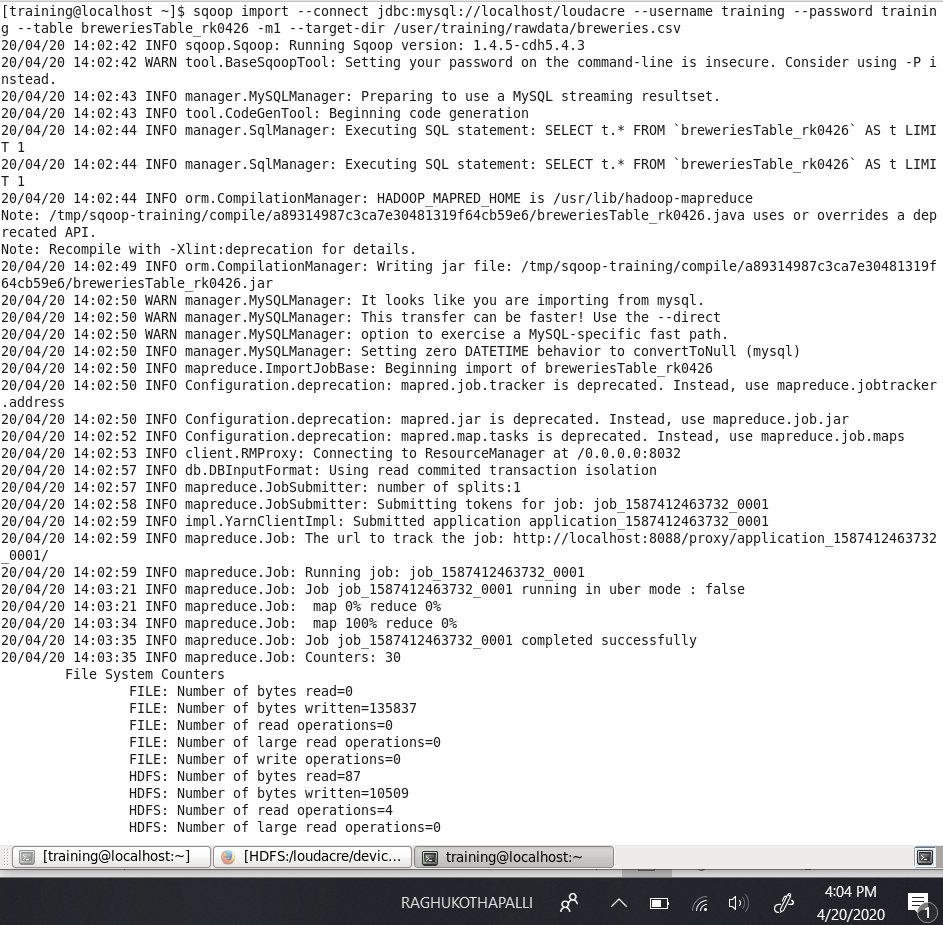
So, while inserting the data into the table we will be removing the name of the column because without the schema it will be easy to load into the Hadoop system.

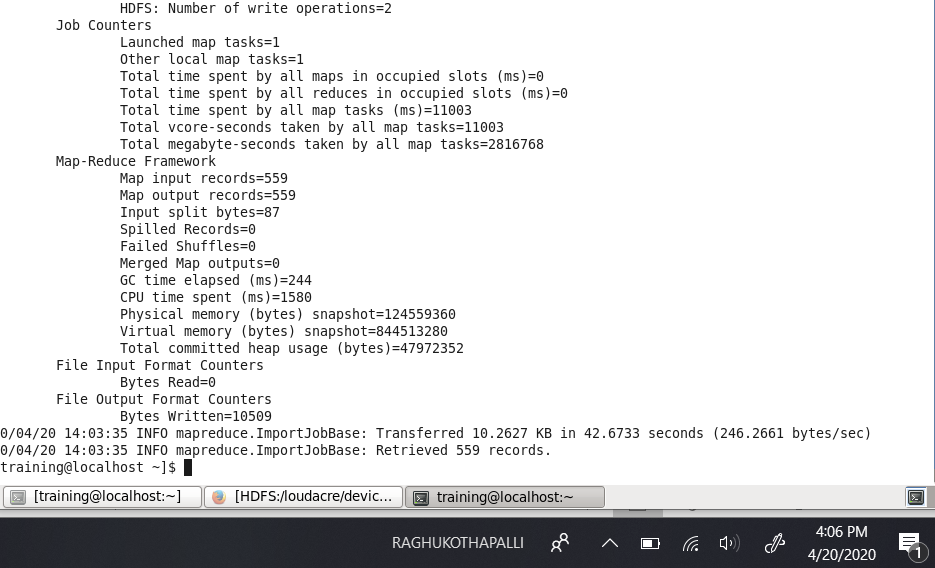
Loading a data from a local file named “breweries.csv” into the table “breweriesTable\_rk0426”



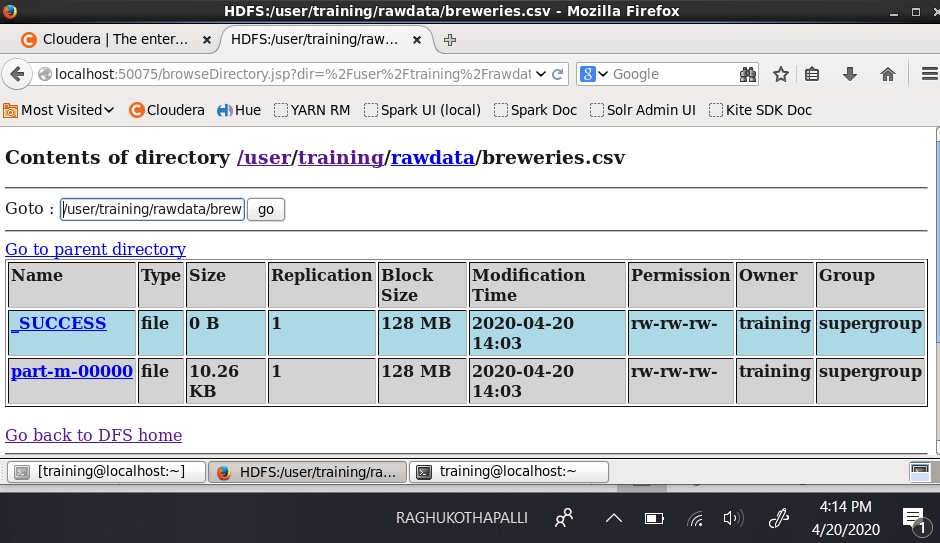
**Importing into the Sqoop**

So, now the data is in MySQL database then importing to Sqoop.

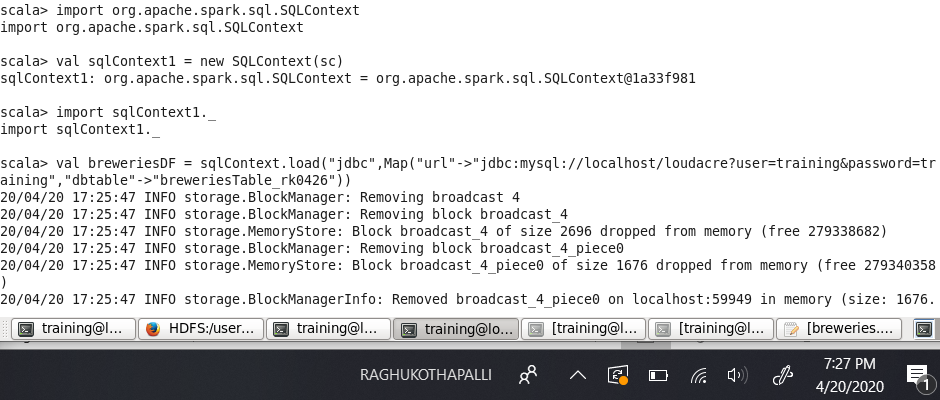




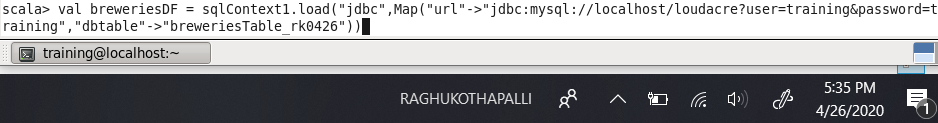
* Viewing the file in the in the Firefox Hadoop directory, as a proof that file is imported.



**Now moving the brewery.csv into Spark from Sqoop**



**Loading data to spark using Scala.**



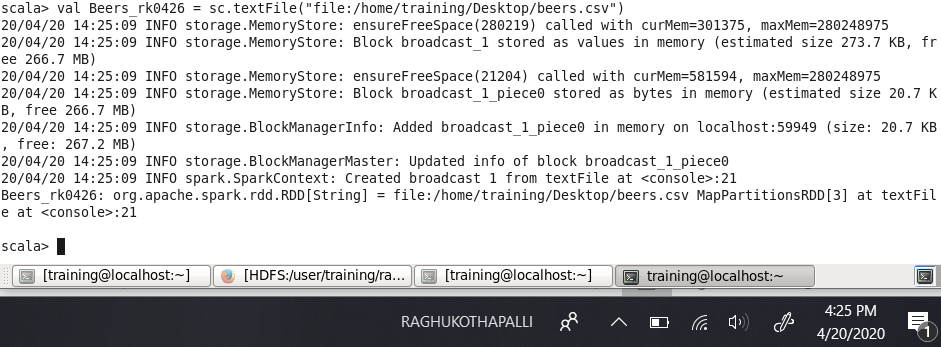
1. Create RDD for Beers.csv

Answer

**Creating RDD for Beers.csv**

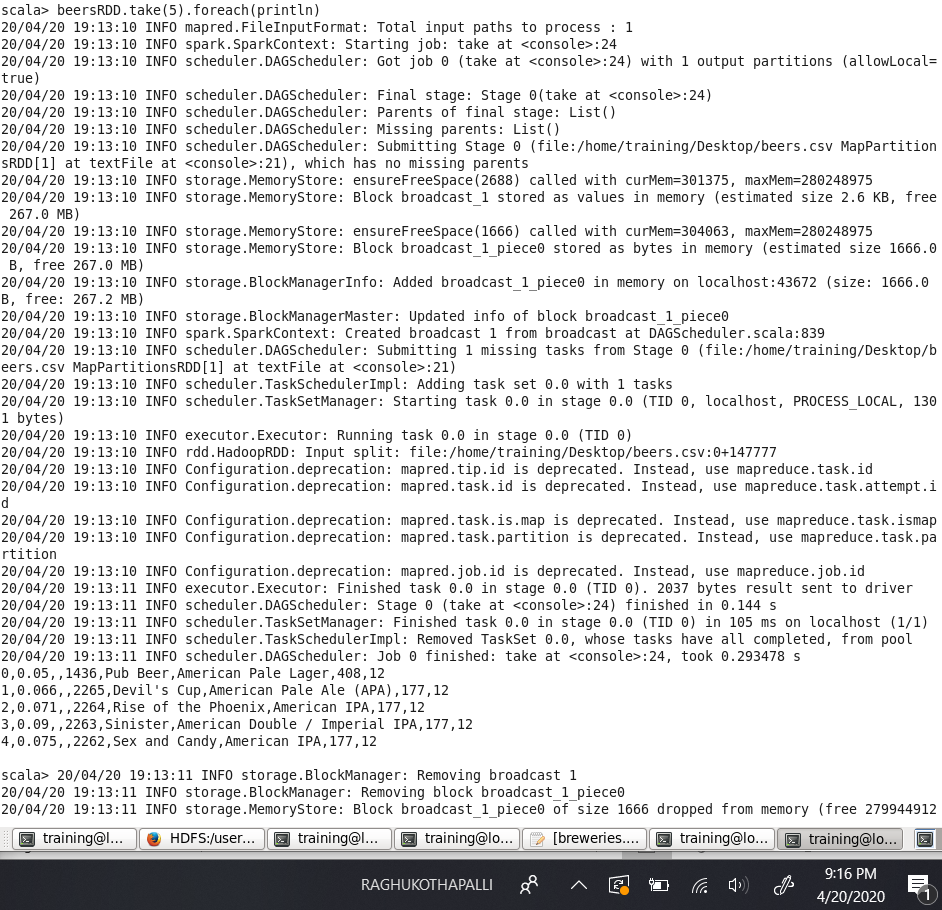
* Invoking the shell

Using “spark-shell”



**Creating a data frame for SQL**

Reference: <https://docs.databricks.com/spark/latest/dataframes-datasets/introduction-to-dataframes-scala.html>



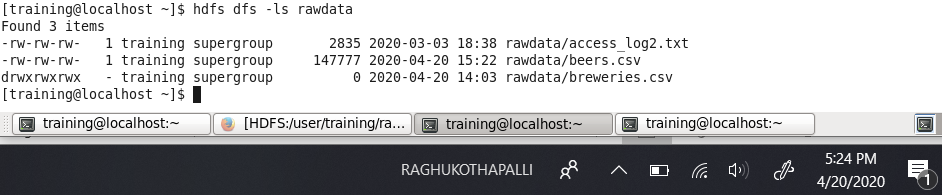
* Copy the beers.csv file to the desktop in VM then load the beers.csv file to HDFS rawdata directory using the following command

hdfs dfs -put Desktop/beers.csv rawdata;



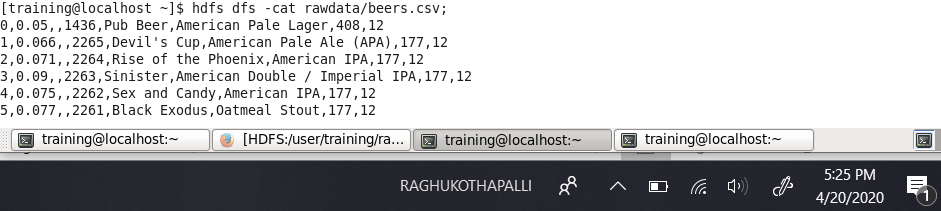
* **Verify the file is loaded**

hdfs dfs -ls rawdata



* **Now opening the beers.csv file**

hdfs dfs –cat rawdata/beers.csv



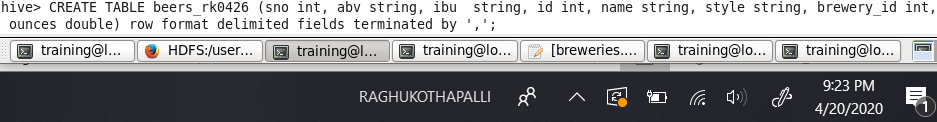
**Inserting the data into the table in hive**

* Launch hive

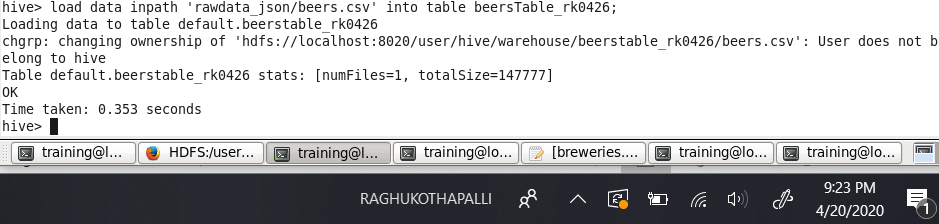
$DEV1/scripts/training\_setup\_dev1.sh

hive

* Creating table to insert the data in hive

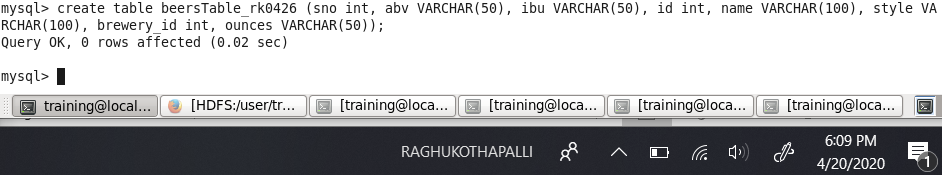


* **Insert the data into the created table**

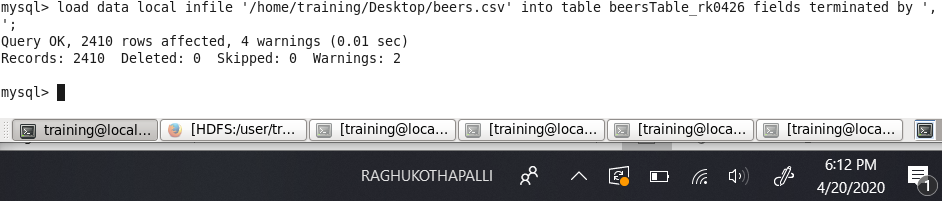


**Creating table in MySQL**

* Use loudacre
* Create a table named “beersTable\_rk0426”

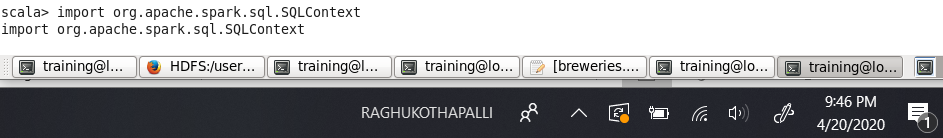


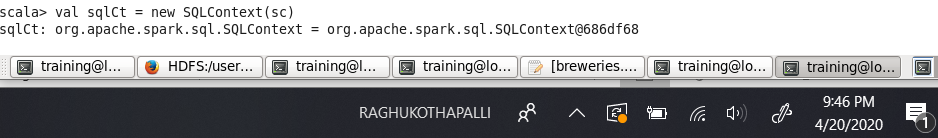
**Importing the data in to the created table**



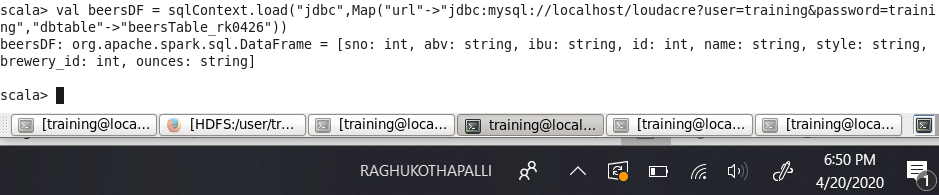
**Importing and Creating SQL context**

Reference: <https://spark.apache.org/docs/1.6.1/sql-programming-guide.html#sql>





**Loading data to Spark using Scala.**



Reference: <http://spark.apache.org/docs/1.6.2/api/R/registerTempTable.html>

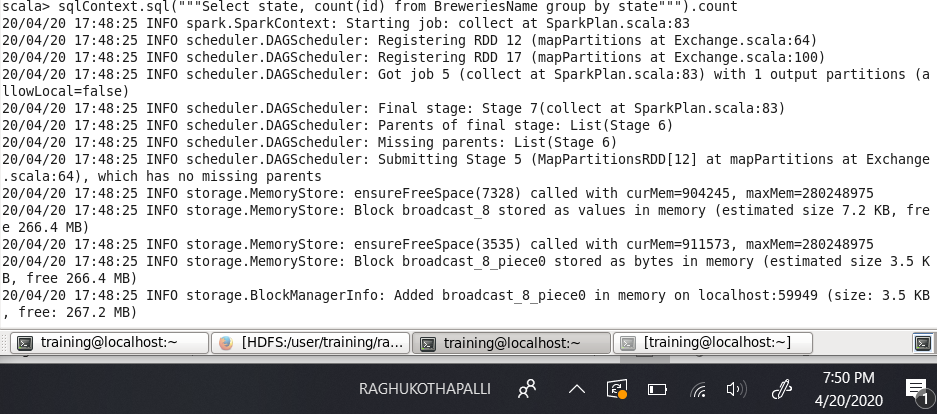
1. For the following questions use Spark SQL
   1. Determine the number of breweries in each state

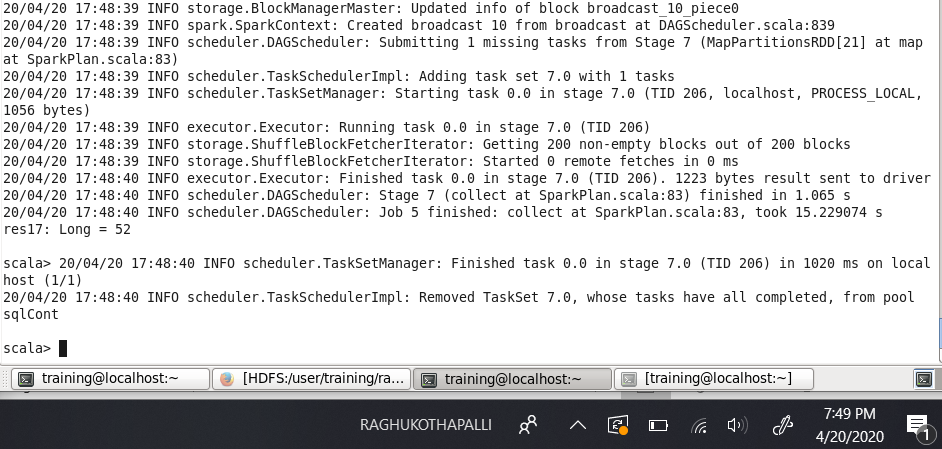
Answer

We determine the number of breweries in each state by the following command in scala.

Reference: <https://www.tutorialspoint.com/spark_sql/spark_sql_dataframes.htm>

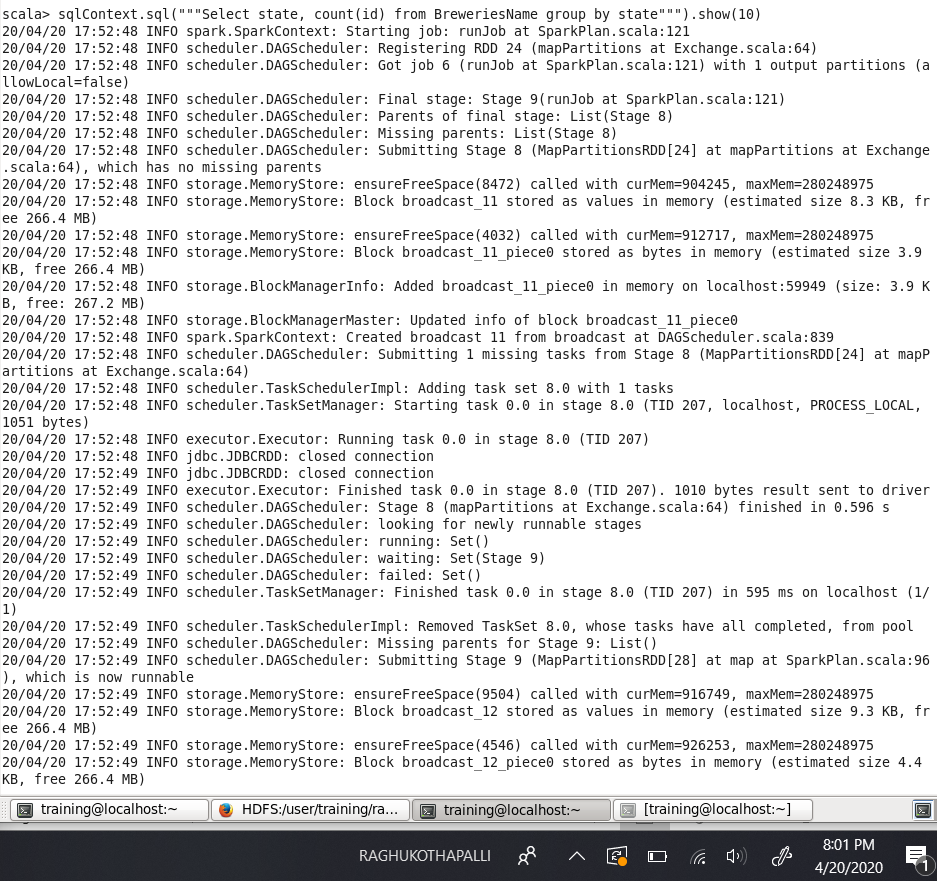
.count() command is used to figure out the number of records in the data set.

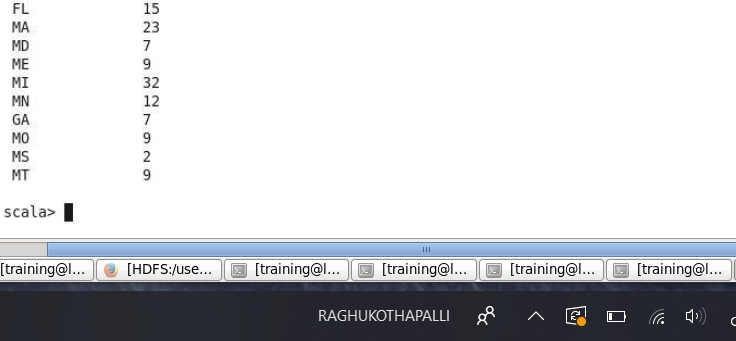




The result is 52 records. Now by using the following command .show() which helps in displaying the content of the data frame to the stdout.

As there are many records I have used show(10) which gives us the first 10 records of the data frame.





These are the top 10 breweries in each state irrespective of the order.

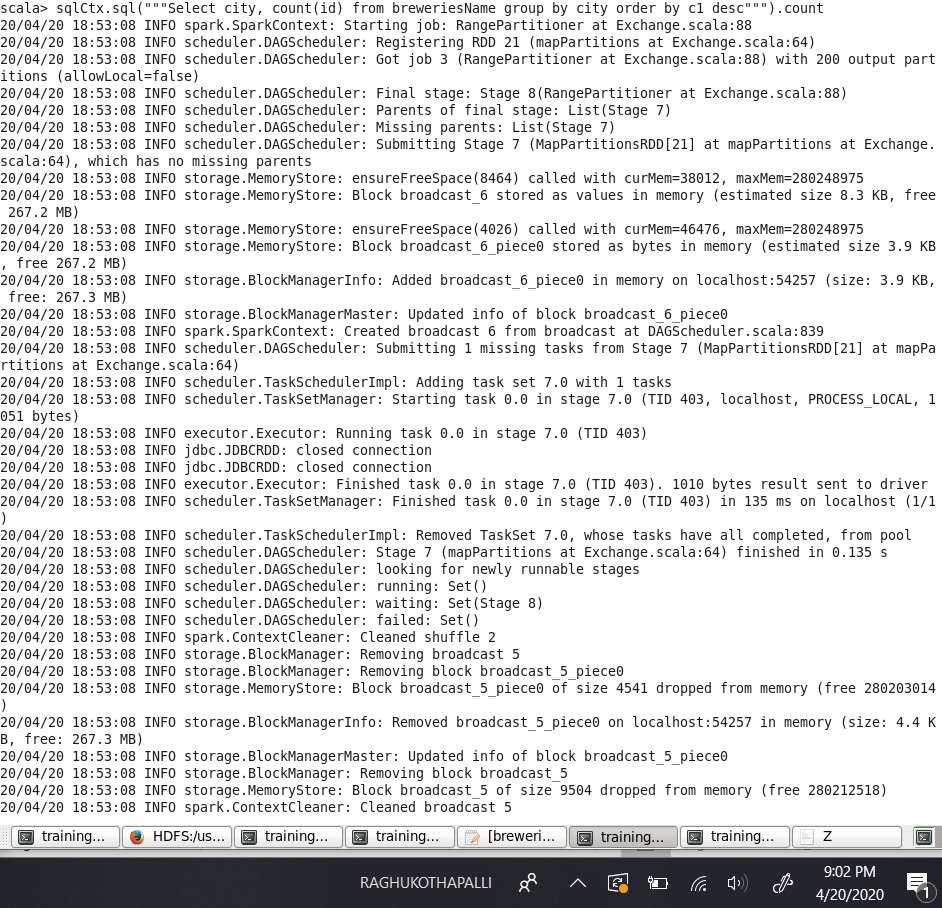
* 1. Determine the cities with most breweries

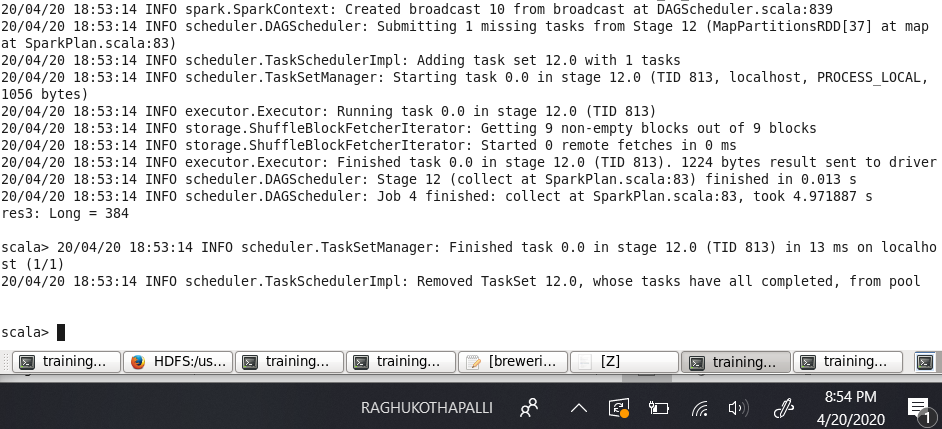
Answer

We determine the number of breweries in each state by the following command in scala.

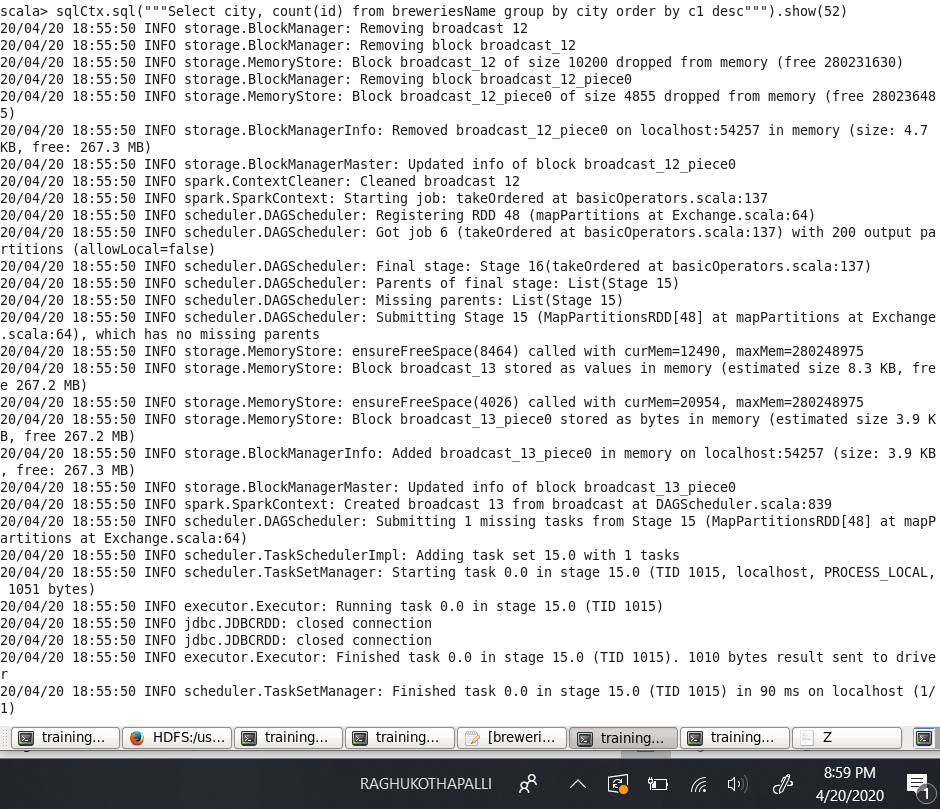
Reference: <https://www.tutorialspoint.com/spark_sql/spark_sql_dataframes.htm>

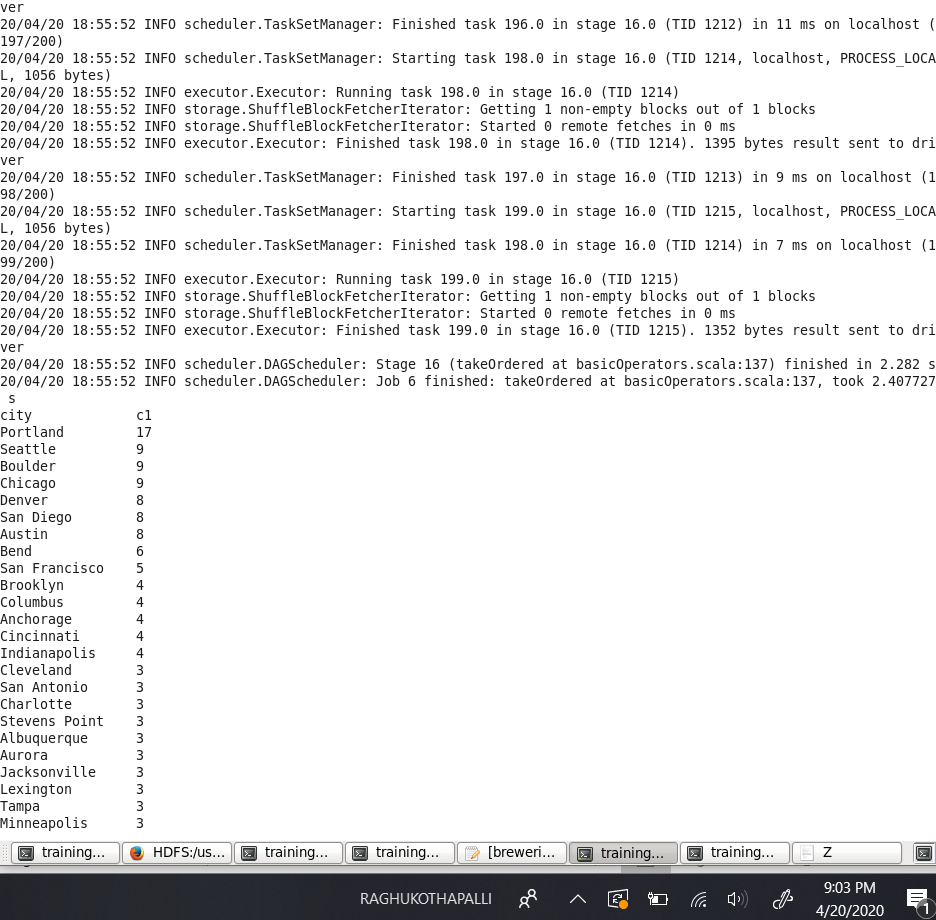
.count() command is used to figure out the number of records in the data set.

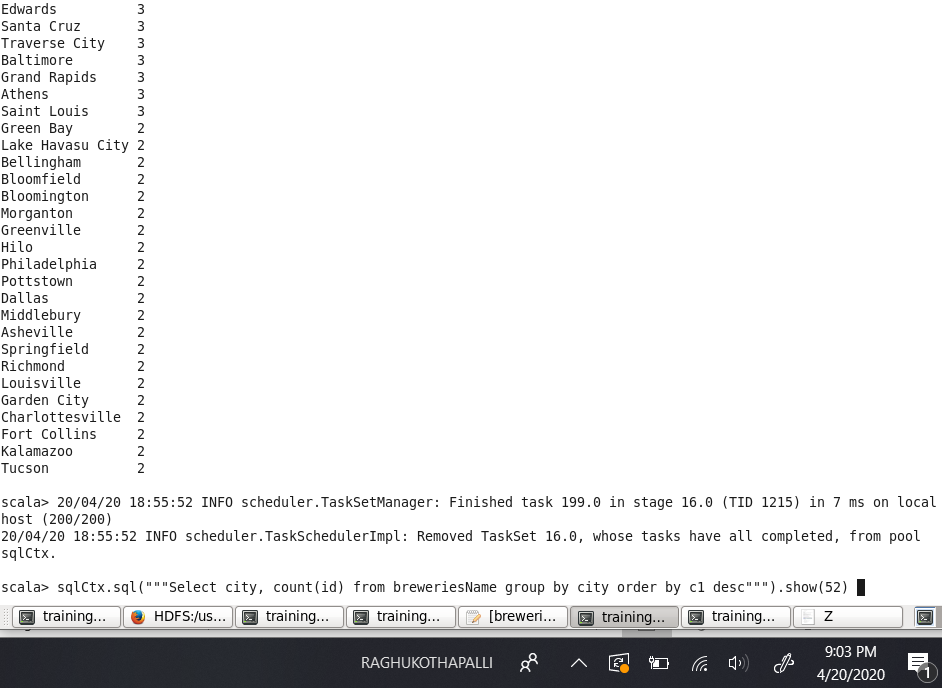




There are 384 records in the data frame. The value 384 is stored in the res3.

Now by using the following command .show() which helps in displaying the content of the data frame to the stdout.

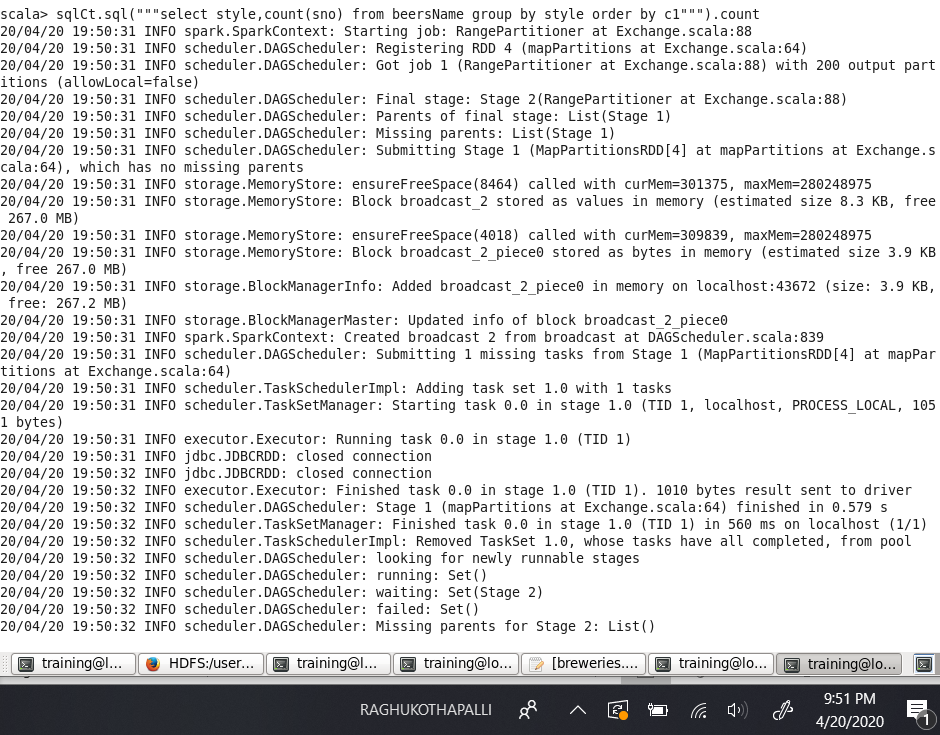


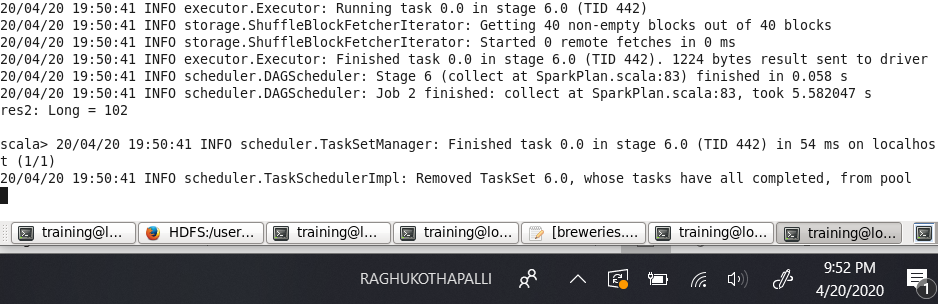


By this we get to see that Portland is the city with the greatest number of breweries.

* 1. Determine the most brewed beer style

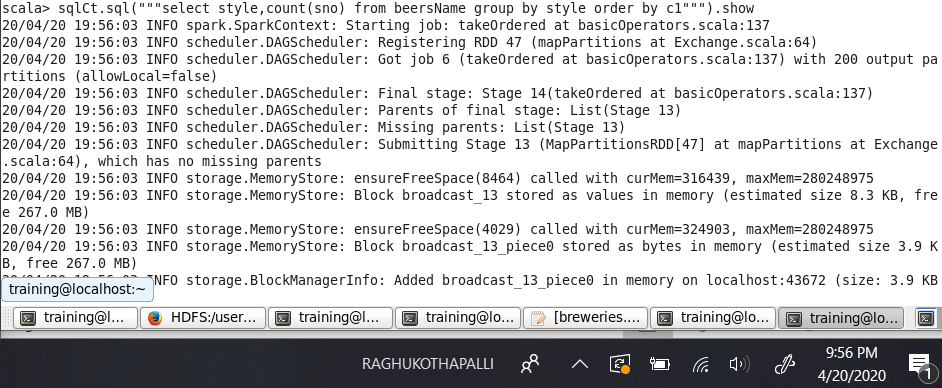
Answer

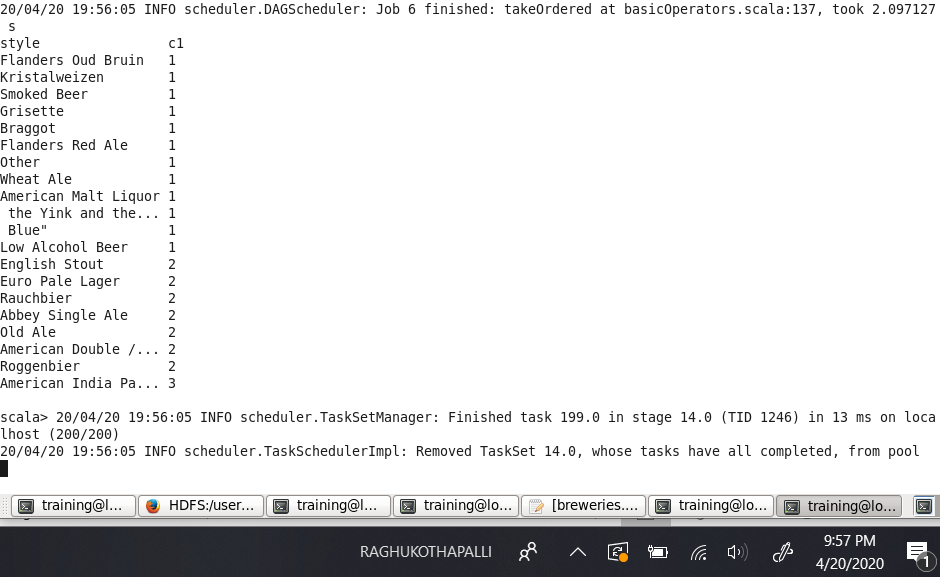
Now we use the following scala command to get the count of the number of records of the most brewed style.



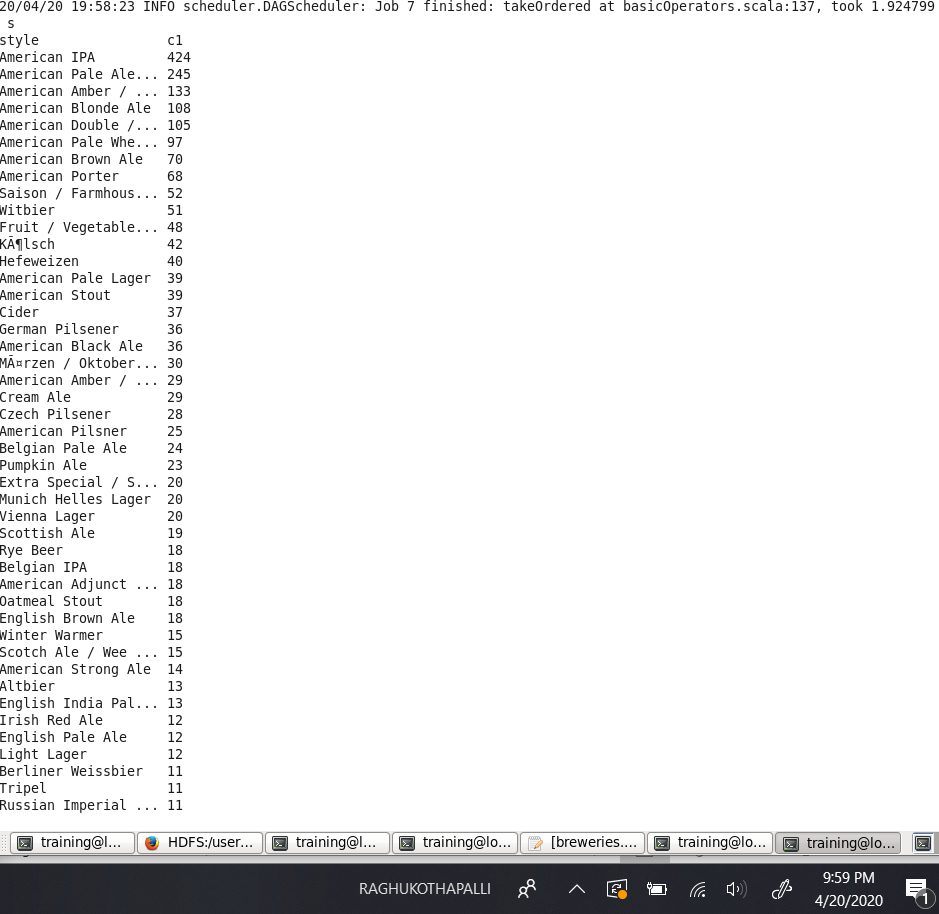
The result is 102. Indicates the number of records in the dataframe with the respective sql statement.

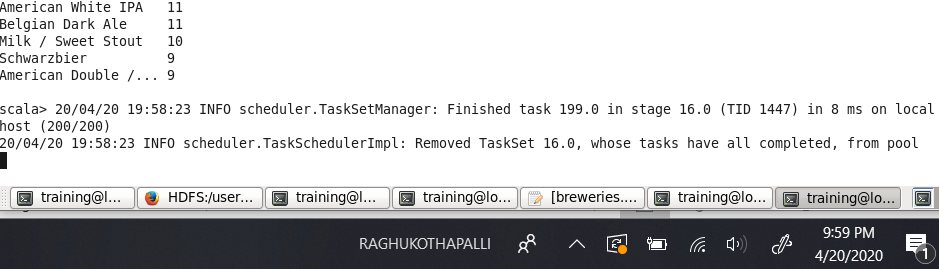
Show(20) records regardless of whether it is descending or ascending.





For show(50) used desc to get the most brewed beer style. To obtain the top 50 most brewed. Most brewed we use the descending order in order to obtain them.





By the above results we get to know that American IPA is the most brewed beer style.