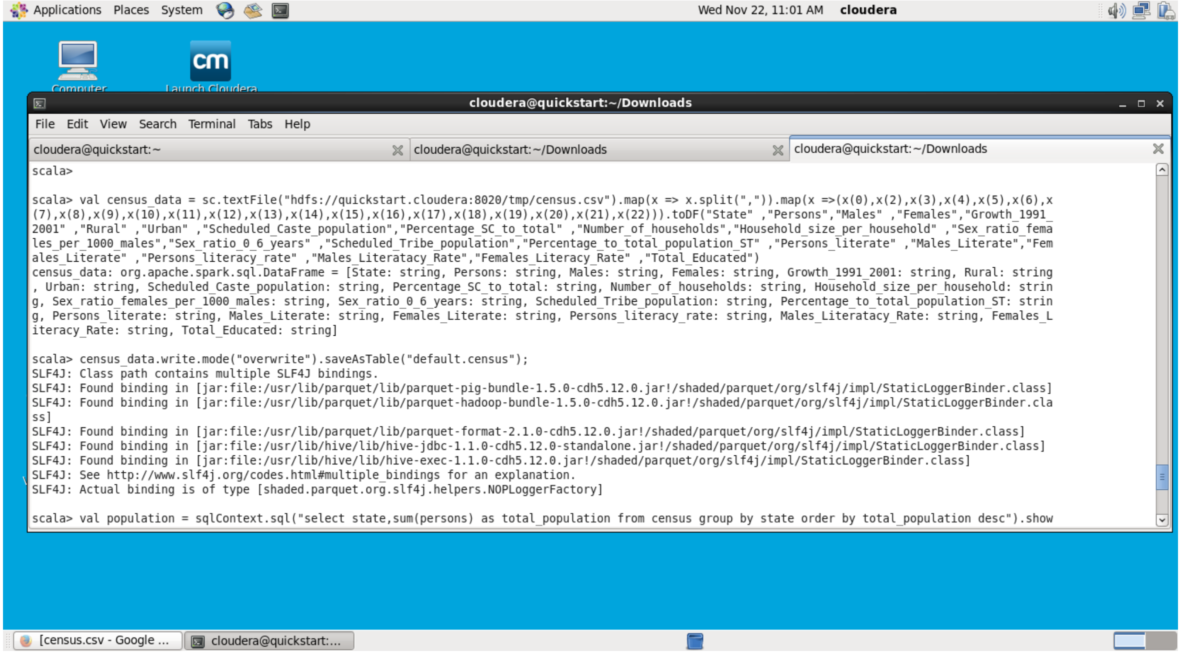
Census data analysis

1. Copy data to HDFS directory using Hadoop dfs -put command
2. Read csv file and create data from using csv data



Applications Places System
Wed Nov 22, 11:01 AM cloudera
Computer
Launch Cloudera
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
[census.csv - Google ...
cloudera@quickstart:...

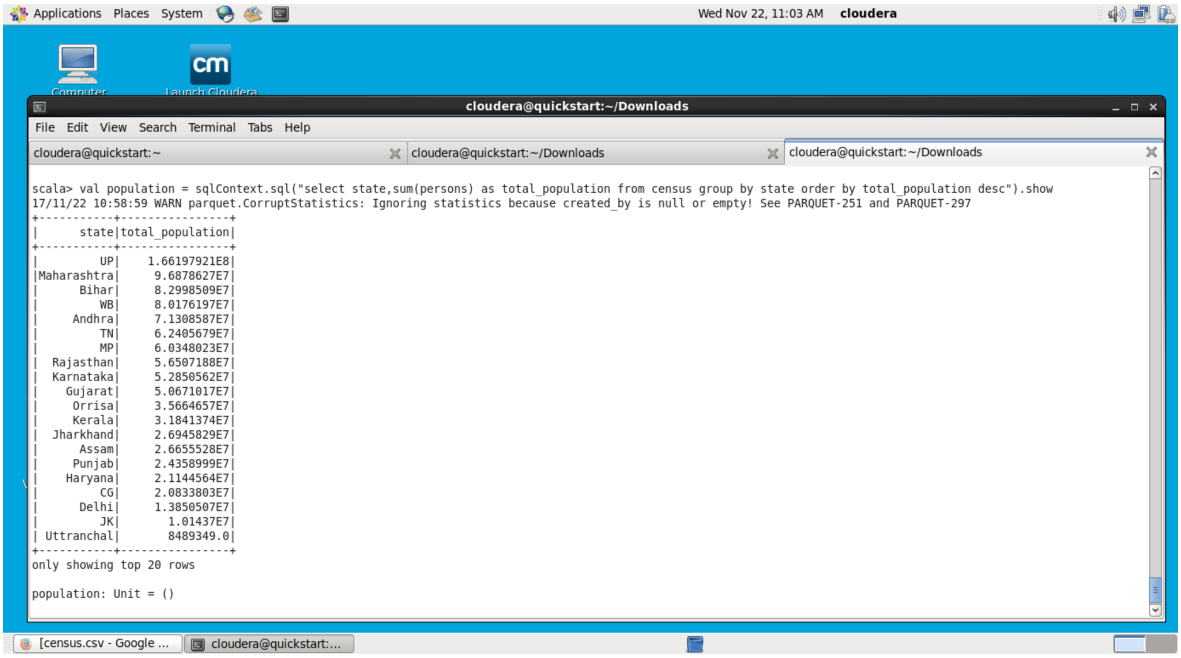
```
scala>

scala> val census_data = sc.textFile("hdfs://quickstart.cloudera:8020/tmp/census.csv").map(x => x.split(",")).map(x =>(x(0),x(2),x(3),x(4),x(5),x(6),x
(7),x(8),x(9),x(10),x(11),x(12),x(13),x(14),x(15),x(16),x(17),x(18),x(19),x(20),x(21),x(22))).toDF("State" ,"Persons","Males" ,"Females","Growth_1991_
2001" ,"Rural" ,"Urban" ,"Scheduled_Caste_population","Percentage_SC_to_total" ,"Number_of_households","Household_size_per_household" ,"Sex_ratio_fema
les_per_1000_males","Sex_ratio_0_6_years" ,"Scheduled_Tribe_population","Percentage_to_total_population_ST" ,"Persons_literate" ,"Males_Literate","Fem
ales_Literate" ,"Persons_literacy_rate" ,"Males_Literatacy_Rate","Females_Literacy_Rate" ,"Total_Educated")
census_data: org.apache.spark.sql.DataFrame = [State: string, Persons: string, Males: string, Females: string, Growth_1991_2001: string, Rural: string
, Urban: string, Scheduled_Caste_population: string, Percentage_SC_to_total: string, Number_of_households: string, Household_size_per_household: strin
g, Sex_ratio_females_per_1000_males: string, Sex_ratio_0_6_years: string, Scheduled_Tribe_population: string, Percentage_to_total_population_ST: strin
g, Persons_literate: string, Males_Literate: string, Females_Literate: string, Persons_literacy_rate: string, Males_Literatacy_Rate: string, Females_L
iteracy_Rate: string, Total_Educated: string]

scala> census_data.write.mode("overwrite").saveAsTable("default.census");
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/lib/parquet/lib/parquet-pig-bundle-1.5.0-cdh5.12.0.jar!/shaded/parquet/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/parquet/lib/parquet-hadoop-bundle-1.5.0-cdh5.12.0.jar!/shaded/parquet/org/slf4j/impl/StaticLoggerBinder.cla
ss]
SLF4J: Found binding in [jar:file:/usr/lib/parquet/lib/parquet-format-2.1.0-cdh5.12.0.jar!/shaded/parquet/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/hive-jdbc-1.1.0-cdh5.12.0-standalone.jar!/shaded/parquet/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/lib/hive/lib/hive-exec-1.1.0-cdh5.12.0.jar!/shaded/parquet/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [shaded.parquet.org.slf4j.helpers.NOPLoggerFactory]

scala> val population = sqlContext.sql("select state,sum(persons) as total_population from census group by state order by total_population desc").show
```

3. Find out the state wise population and order by state

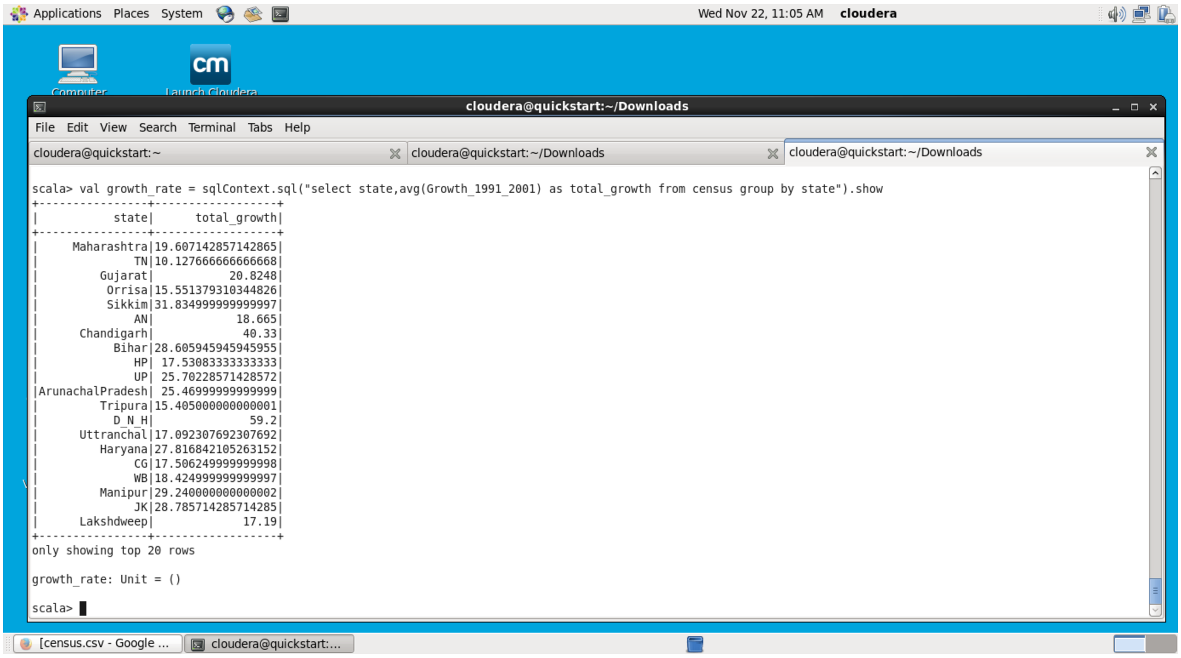


Applications Places System
Wed Nov 22, 11:03 AM cloudera
Computer
Launch Cloudera
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
[census.csv - Google ...
cloudera@quickstart:...

```
scala> val population = sqlContext.sql("select state,sum(persons) as total_population from census group by state order by total_population desc").show
17/11/22 10:58:59 WARN parquet.CorruptStatistics: Ignoring statistics because created_by is null or empty! See PARQUET-251 and PARQUET-297
+-----------+----------------+
|      state|total_population|
+-----------+----------------+
|         UP|    1.66197921E8|
|Maharashtra|     9.6878627E7|
|      Bihar|     8.2998509E7|
|         WB|     8.0176197E7|
|     Andhra|     7.1308587E7|
|         TN|     6.2405679E7|
|         MP|     6.0348023E7|
|  Rajasthan|     5.6507188E7|
|  Karnataka|     5.2850562E7|
|    Gujarat|     5.0671017E7|
|     Orrisa|     3.5664657E7|
|     Kerala|     3.1841374E7|
|  Jharkhand|     2.6945829E7|
|      Assam|     2.6655528E7|
|     Punjab|     2.4358999E7|
|    Haryana|     2.1144564E7|
|         CG|     2.0833803E7|
|      Delhi|     1.3850507E7|
|         JK|       1.01437E7|
| Uttranchal|       8489349.0|
+-----------+----------------+
only showing top 20 rows

population: Unit = ()
```

Find out the Growth Rate of Each State Between 1991-2001

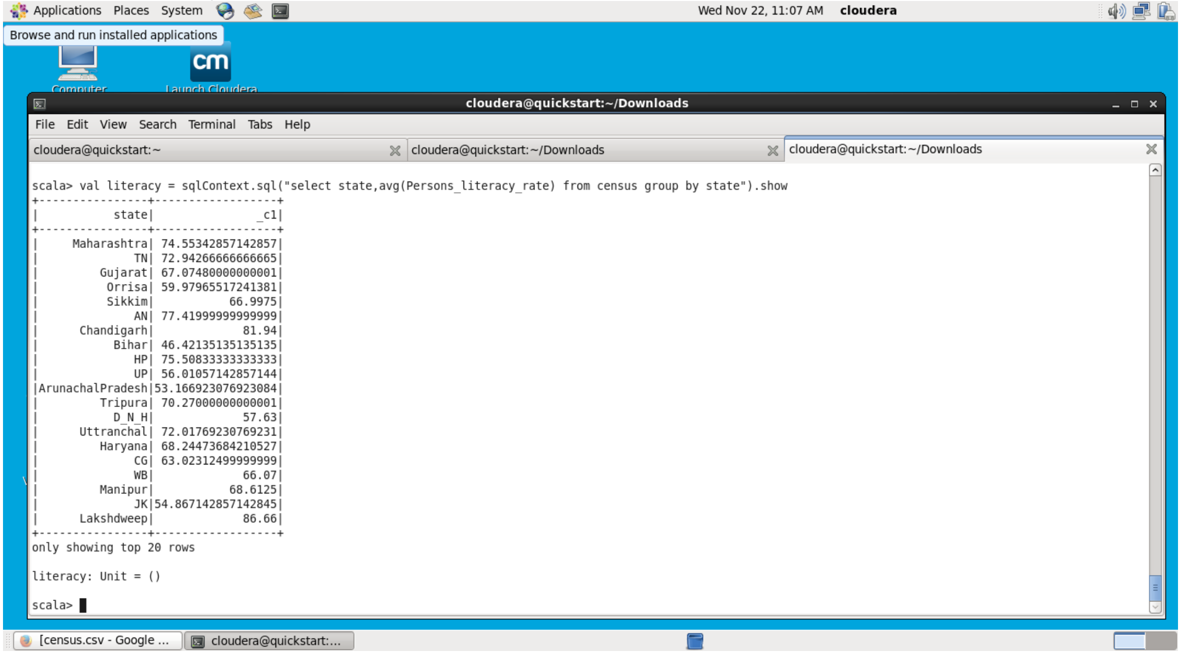


```
scala> val growth_rate = sqlContext.sql("select state,avg(Growth_1991_2001) as total_growth from census group by state").show
+----------------+-----------------+
|           state|     total_growth|
+----------------+-----------------+
|     Maharashtra|19.607142857142865|
|              TN|10.127666666666668|
|         Gujarat|          20.8248|
|          Orrisa|15.551379310344826|
|          Sikkim|31.834999999999997|
|              AN|           18.665|
|      Chandigarh|            40.33|
|           Bihar|28.605945945945955|
|              HP| 17.530833333333333|
|              UP| 25.70228571428572|
|ArunachalPradesh| 25.469999999999999|
|         Tripura|15.405000000000001|
|           D_N_H|             59.2|
|      Uttranchal|17.092307692307692|
|         Haryana|27.816842105263152|
|              CG|17.506249999999998|
|              WB|18.424999999999997|
|         Manipur|29.240000000000002|
|              JK|28.785714285714285|
|      Lakshdweep|            17.19|
+----------------+-----------------+
only showing top 20 rows

growth_rate: Unit = ()

scala> 
```

Find the literacy rate of each state

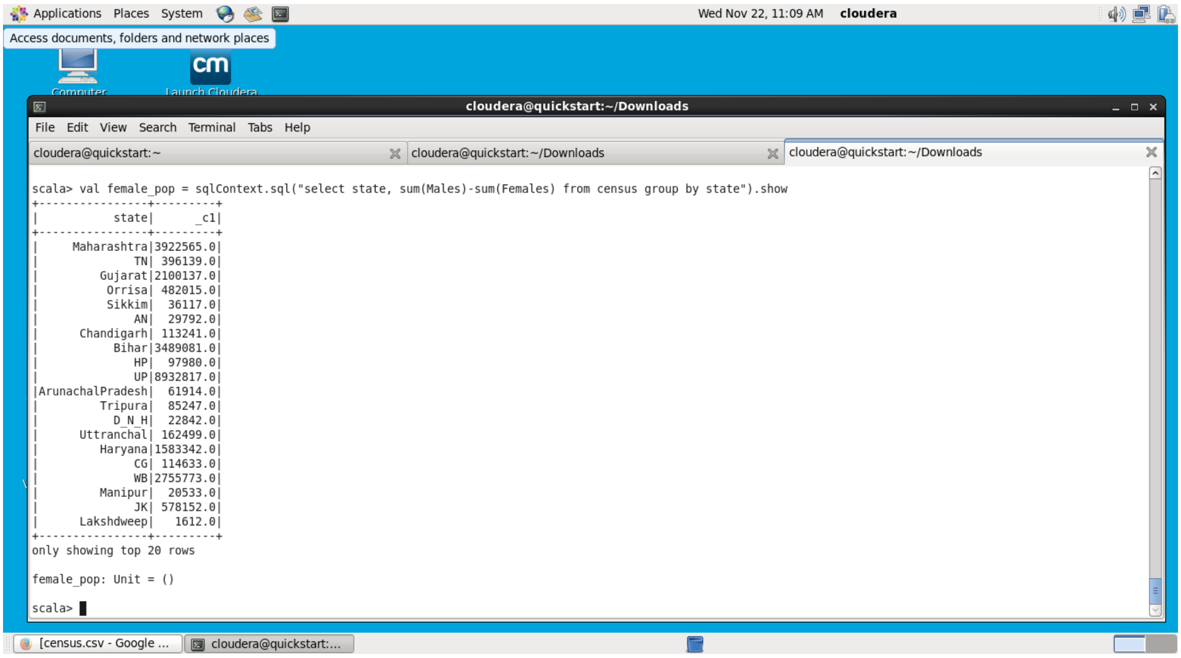


```
scala> val literacy = sqlContext.sql("select state,avg(Persons_literacy_rate) from census group by state").show
+----------------+-----------------+
|           state|              _c1|
+----------------+-----------------+
|     Maharashtra| 74.55342857142857|
|              TN| 72.94266666666665|
|         Gujarat| 67.07480000000001|
|          Orrisa| 59.97965517241381|
|          Sikkim|          66.9975|
|              AN| 77.41999999999999|
|      Chandigarh|            81.94|
|           Bihar| 46.42135135135135|
|              HP| 75.50833333333333|
|              UP| 56.01057142857144|
|ArunachalPradesh|53.166923076923084|
|         Tripura| 70.27000000000001|
|           D_N_H|            57.63|
|      Uttranchal| 72.01769230769231|
|         Haryana| 68.24473684210527|
|              CG| 63.02312499999999|
|              WB|            66.07|
|         Manipur|          68.6125|
|              JK|54.867142857142845|
|      Lakshdweep|            86.66|
+----------------+-----------------+
only showing top 20 rows

literacy: Unit = ()

scala> 
```

Find out the States with More Female Population

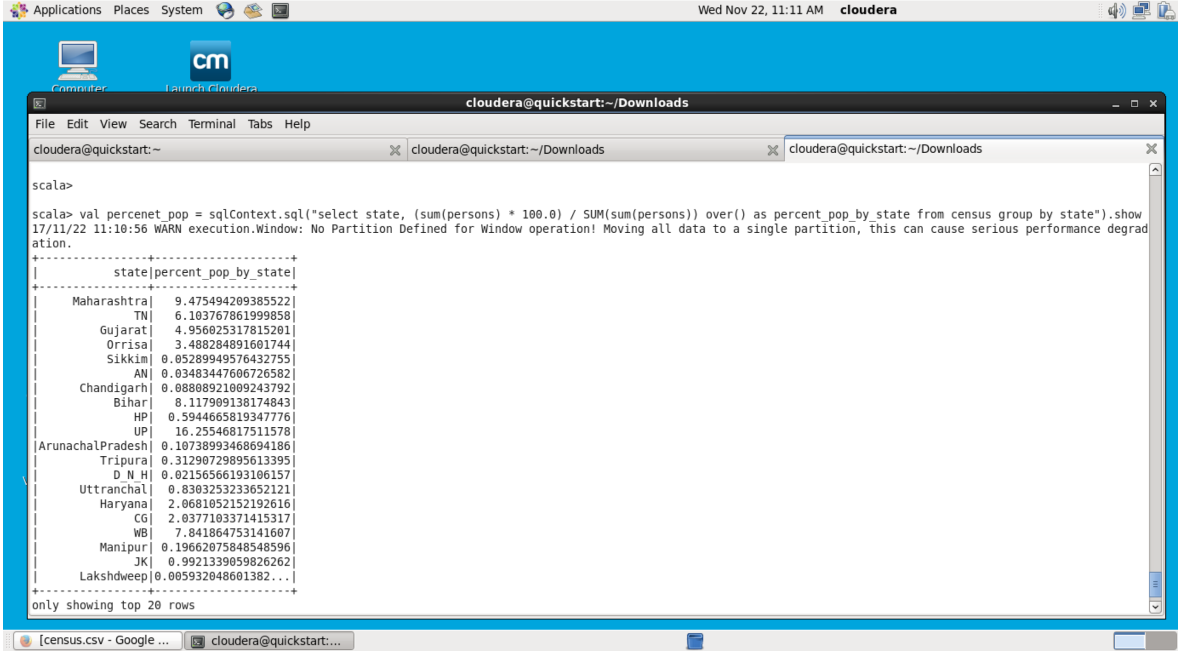

Applications Places System
Wed Nov 22, 11:09 AM cloudera
Access documents, folders and network places
Computer
Launch Cloudera
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
[census.csv - Google ...
cloudera@quickstart:...

```
scala> val female_pop = sqlContext.sql("select state, sum(Males)-sum(Females) from census group by state").show
+----------------+---------+
|           state|      _c1|
+----------------+---------+
|     Maharashtra|3922565.0|
|              TN| 396139.0|
|         Gujarat|2100137.0|
|          Orrisa| 482015.0|
|          Sikkim|  36117.0|
|              AN|  29792.0|
|      Chandigarh| 113241.0|
|           Bihar|3489081.0|
|              HP|  97980.0|
|              UP|8932817.0|
|ArunachalPradesh|  61914.0|
|         Tripura|  85247.0|
|           D_N_H|  22842.0|
|      Uttranchal| 162499.0|
|         Haryana|1583342.0|
|              CG| 114633.0|
|              WB|2755773.0|
|         Manipur|  20533.0|
|              JK| 578152.0|
|      Lakshdweep|   1612.0|
+----------------+---------+
only showing top 20 rows

female_pop: Unit = ()

scala> 
```

Find out the Percentage of Population in Every State


Applications Places System
Wed Nov 22, 11:11 AM cloudera
Computer
Launch Cloudera
cloudera@quickstart:~/Downloads
File Edit View Search Terminal Tabs Help
cloudera@quickstart:~
cloudera@quickstart:~/Downloads
cloudera@quickstart:~/Downloads
[census.csv - Google ...
cloudera@quickstart:...

```
scala>

scala> val percenet_pop = sqlContext.sql("select state, (sum(persons) * 100.0) / SUM(sum(persons)) over() as percent_pop_by_state from census group by state").show
17/11/22 11:10:56 WARN execution.Window: No Partition Defined for Window operation! Moving all data to a single partition, this can cause serious performance degrad
ation.
+----------------+--------------------+
|           state|percent_pop_by_state|
+----------------+--------------------+
|     Maharashtra|   9.475494209385522|
|              TN|   6.103767861999858|
|         Gujarat|   4.956025317815201|
|          Orrisa|   3.488284891601744|
|          Sikkim| 0.05289949576432755|
|              AN| 0.03483447606726582|
|      Chandigarh| 0.08808921009243792|
|           Bihar|   8.117909138174843|
|              HP|  0.5944665819347776|
|              UP|   16.25546817511578|
|ArunachalPradesh| 0.10738993468694186|
|         Tripura| 0.31290729895613395|
|           D_N_H| 0.02156566193106157|
|      Uttranchal|  0.8303253233652121|
|         Haryana|  2.0681052152192616|
|              CG|  2.0377103371415317|
|              WB|   7.841864753141607|
|         Manipur| 0.19662075848548596|
|              JK|  0.9921339059826262|
|      Lakshdweep|0.005932048601382...|
+----------------+--------------------+
only showing top 20 rows
```