Assignment 22.1:

Problem Statement:

Here we are going to work on Census Data.

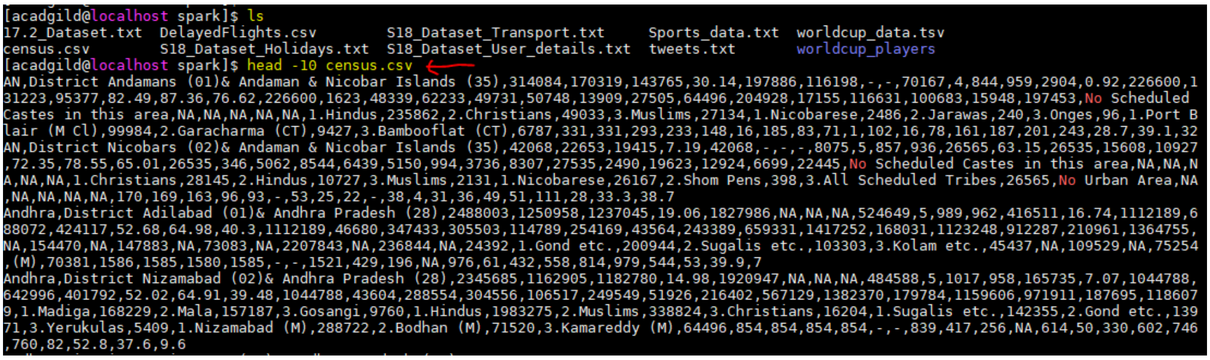
Here is the total dataset description

State String,District String,Persons String,Males int,Females int,Growth_1991_2001 int,Rural int,Urban int,Scheduled_Caste_population int,Percentage_SC_to_total int,Number_of_households int,Household_size_per_household int,Sex_ratio_females_per_1000_males int ,Sex_ratio_0_6_years int,Scheduled_Tribe_population int,Percentage_to_total_population_ST int,Persons_literate int,Males_Literate int,Females_Literate int,Persons_literacy_rate int,Males_Literatacy_Rate int,Females_Literacy_Rate int,Total_Educated int,Data_without_level int,Below_Primary int,Primary int,Middle int,Matric_Higher_Secondary_Diploma int,Graduate_and_Above int,X0_4_years int,X5_14_years int,X15_59_years int,X60_years_and_above_Incl_ANS int,Total_workers int,Main_workers int,Marginal_workers int,Non_workers int,SC_1_Name String,SC_1_Population int,SC_2_Name String,SC_2_Population int,SC_3_Name String,SC_3_Population int,Religeon_1_Name String,Religeon_1_Population int,Religeon_2_Name String,Religeon_2_Population int,Religeon_3_Name String,Religeon_3_Population int,ST_1_Name String,ST_1_Population int,ST_2_Name String,ST_2_Population int,ST_3_Name String,ST_3_Population int,Imp_Town_1_Name String,Imp_Town_1_Population int,Imp_Town_2_Name String,Imp_Town_2_Population int,Imp_Town_3_Name String,Imp_Town_3_Population int,Total_Inhabited_Villages int,Drinking_water_facilities int,Safe_Drinking_water int,Electricity_Power_Supply int,Electricity_domestic int,Electricity_Agriculture int,Primary_school int,Middle_schools int,Secondary_Sr_Secondary_schools int,College int,Medical_facility int,Primary_Health_Centre int,Primary_Health_Sub_Centre int,Post_telegraph_and_telephone_facility int,Bus_services int,Paved_approach_road int,Mud_approach_road int,Permanent_House int,Semi_permanent_House int,Temporary_House int

Due to the limitation of 22 elements for a map function, we are taking only 22 columns from the data set.

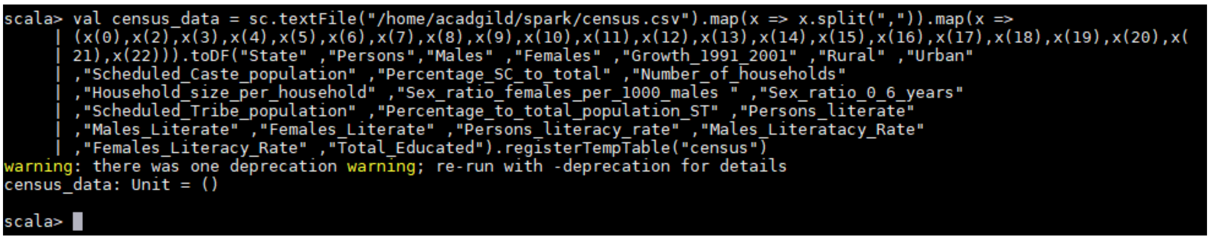
Here is what we are taking

"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_females_per_1000_males " ,"Sex_ratio_0_6_years" ,"Scheduled_Tribe_population" ,"Percentage_to_total_population_ST" ,"Persons_literate" ,"Males_Literate" ,"Females_Literate" ,"Persons_literacy_rate" ,"Males_Literatacy_Rate" ,"Females_Literacy_Rate" ,"Total_Educated"



```
[acadgild@localhost spark]$ ls
17.2_Dataset.txt  DelayedFlights.csv        S18_Dataset_Transport.txt     Sports_data.txt  worldcup_data.tsv
census.csv        S18_Dataset_Holidays.txt  S18_Dataset_User_details.txt  tweets.txt       worldcup_players
[acadgild@localhost spark]$ head -10 census.csv
AN,District Andamans (01)& Andaman & Nicobar Islands (35),314084,170319,143765,30.14,197886,116198,-,-,70167,4,844,959,2904,0.92,226600,1
31223,95377,82.49,87.36,76.62,226600,1623,48339,62233,49731,50748,13909,27505,64496,204928,17155,116631,100683,15948,197453,No Scheduled
Castes in this area,NA,NA,NA,NA,NA,1.Hindus,235862,2.Christians,49033,3.Muslims,27134,1.Nicobarese,2486,2.Jarawas,240,3.Onges,96,1.Port B
lair (M Cl),99984,2.Garacharma (CT),9427,3.Bambooflat (CT),6787,331,331,293,233,148,16,185,83,71,1,102,16,78,161,187,201,243,28.7,39.1,32
AN,District Nicobars (02)& Andaman & Nicobar Islands (35),42068,22653,19415,7.19,42068,-,-,-,8075,5,857,936,26565,63.15,26535,15608,10927
,72.35,78.55,65.01,26535,346,5062,8544,6439,5150,994,3736,8307,27535,2490,19623,12924,6699,22445,No Scheduled Castes in this area,NA,NA,N
A,NA,NA,1.Christians,28145,2.Hindus,10727,3.Muslims,2131,1.Nicobarese,26167,2.Shom Pens,398,3.All Scheduled Tribes,26565,No Urban Area,NA
,NA,NA,NA,NA,170,169,163,96,93,-,53,25,22,-,38,4,31,36,49,51,111,28,33.3,38.7
Andhra,District Adilabad (01)& Andhra Pradesh (28),2488003,1250958,1237045,19.06,1827986,NA,NA,NA,524649,5,989,962,416511,16.74,1112189,6
88072,424117,52.68,64.98,40.3,1112189,46680,347433,305503,114789,254169,43564,243389,659331,1417252,168031,1123248,912287,210961,1364755,
NA,154470,NA,147883,NA,73083,NA,2207843,NA,236844,NA,24392,1.Gond etc.,200944,2.Sugalis etc.,103303,3.Kolam etc.,45437,NA,109529,NA,75254
,(M),70381,1586,1585,1580,1585,-,-,1521,429,196,NA,976,61,432,558,814,979,544,53,39.9,7
Andhra,District Nizamabad (02)& Andhra Pradesh (28),2345685,1162905,1182780,14.98,1920947,NA,NA,NA,484588,5,1017,958,165735,7.07,1044788,
642996,401792,52.02,64.91,39.48,1044788,43604,288554,304556,106517,249549,51926,216402,567129,1382370,179784,1159606,971911,187695,118607
9,1.Madiga,168229,2.Mala,157187,3.Gosangi,9760,1.Hindus,1983275,2.Muslims,338824,3.Christians,16204,1.Sugalis etc.,142355,2.Gond etc.,139
71,3.Yerukulas,5409,1.Nizamabad (M),288722,2.Bodhan (M),71520,3.Kamareddy (M),64496,854,854,854,854,-,-,839,417,256,NA,614,50,330,602,746
,760,82,52.8,37.6,9.6
```

val census_data = sc.textFile("/home/acadgild/spark/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_females_per_1000_males " ,"Sex_ratio_0_6_years" ,"Scheduled_Tribe_population" ,"Percentage_to_total_population_ST" ,"Persons_literate" ,"Males_Literate" ,"Females_Literate" ,"Persons_literacy_rate" ,"Males_Literatacy_Rate" ,"Females_Literacy_Rate" ,"Total_Educated").registerTempTable("census")



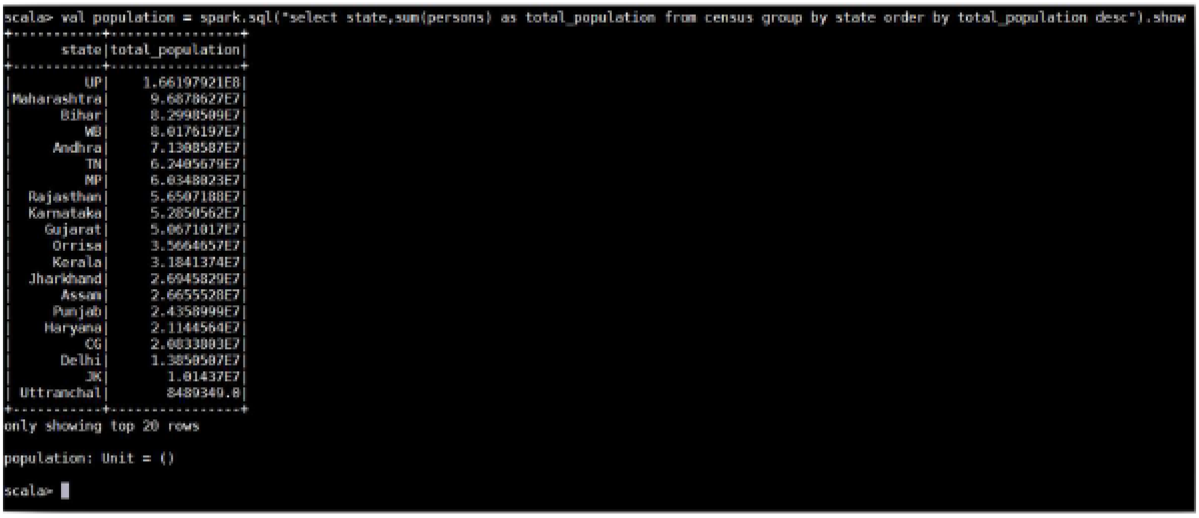
```
scala> val census_data = sc.textFile("/home/acadgild/spark/census.csv").map(x => x.split(",")).map(x =>
     | (x(0),x(2),x(3),x(4),x(5),x(6),x(7),x(8),x(9),x(10),x(11),x(12),x(13),x(14),x(15),x(16),x(17),x(18),x(19),x(20),x(
     | 21),x(22))).toDF("State" ,"Persons","Males" ,"Females" ,"Growth_1991_2001" ,"Rural" ,"Urban"
     | ,"Scheduled_Caste_population" ,"Percentage_SC_to_total" ,"Number_of_households"
     | ,"Household_size_per_household" ,"Sex_ratio_females_per_1000_males " ,"Sex_ratio_0_6_years"
     | ,"Scheduled_Tribe_population" ,"Percentage_to_total_population_ST" ,"Persons_literate"
     | ,"Males_Literate" ,"Females_Literate" ,"Persons_literacy_rate" ,"Males_Literatacy_Rate"
     | ,"Females_Literacy_Rate" ,"Total_Educated").registerTempTable("census")
warning: there was one deprecation warning; re-run with -deprecation for details
census_data: Unit = ()

scala> 
```

1. **Find out the state wise population and order by state**

Code:

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

Output:



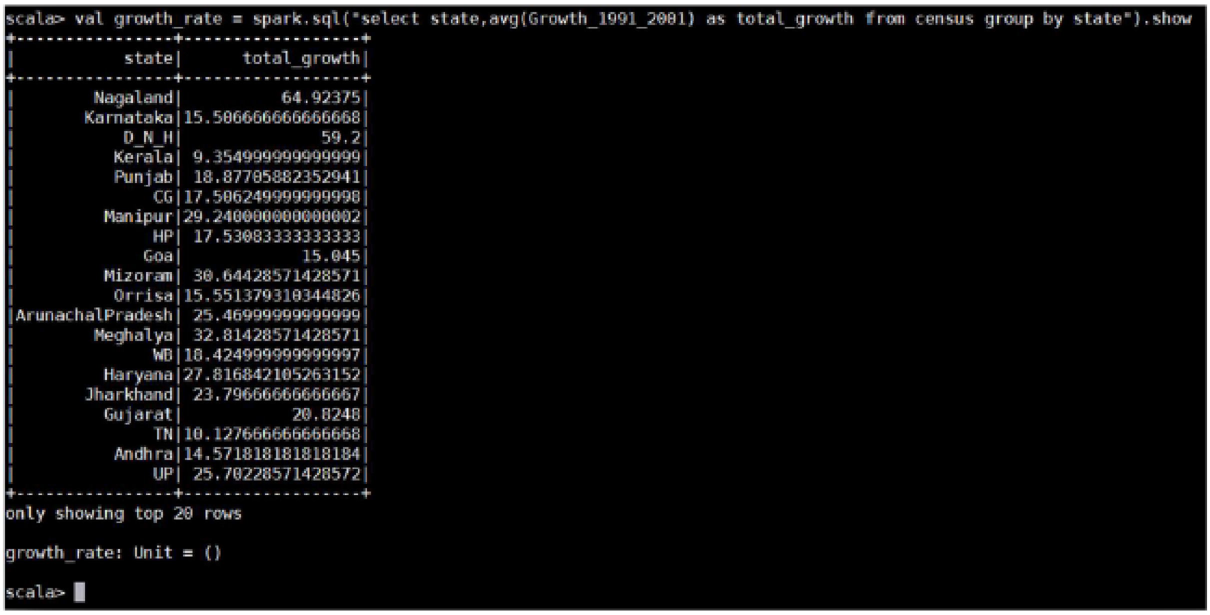
```
scala> val population = spark.sql("select state,sum(persons) as total_population from census group by state order by total_population desc").show
+-----------+----------------+
|      state|total_population|
+-----------+----------------+
|         UP|    1.66197921E8|
|Maharashtra|     9.6878627E7|
|      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 = ()

scala> █
```

2. Find out the growth rate of each state between 1991-2001

Code:

val growth_rate = spark.sql("select state,avg(Growth_1991_2001) as total_growth from census group by state").show

Output:



```
scala> val growth_rate = spark.sql("select state,avg(Growth_1991_2001) as total_growth from census group by state").show
+---------------+------------------+
|          state|      total_growth|
+---------------+------------------+
|       Nagaland|          64.92375|
|      Karnataka|15.506666666666668|
|          D_N_H|              59.2|
|         Kerala| 9.354999999999999|
|         Punjab| 18.87705882352941|
|             CG|17.506249999999998|
|        Manipur|29.240000000000002|
|             HP| 17.53083333333333|
|            Goa|            15.045|
|        Mizoram| 30.64428571428571|
|         Orrisa|15.551379310344826|
|ArunachalPradesh| 25.46999999999999|
|       Meghalya| 32.81428571428571|
|             WB|18.424999999999997|
|        Haryana|27.816842105263152|
|      Jharkhand| 23.79666666666667|
|        Gujarat|           20.8248|
|             TN|10.127666666666668|
|         Andhra|14.571818181818184|
|             UP| 25.70228571428572|
+---------------+------------------+
only showing top 20 rows

growth_rate: Unit = ()

scala> █
```