Census data analysis

This assignment deals with the analysis of Census data.

Below is the total dataset description-

State String,District String,Persons String,Males int,Females int,Growth_1991_2001int,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

Now since we can take only 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"

Below is the code where we are reading the file and mapping the columns-

- val census_data = sc.textFile("/home/acadgild/Assignment-22/census.csv")
- val census_map = census_data.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")

  Below are the screenshot for same with sample output-

```
scala> val census_data = sc.textFile("/home/acadgild/Assignment-22/census.csv")
census_data: org.apache.spark.rdd.RDD[String] = /home/acadgild/Assignment-22/census.csv MapPartitionsRDD[89] at textFile at <console>:55

scala> census_data.take(5).foreach(println)
AN,District Andamans (01)& Andaman & Nicobar Islands (35),314084,170319,143765,30.14,197886,116198,-,-,70167,4,844,959,2904,0.92,226600,131223,95377,82.49,87.36,76.62,226600,1623,48339,62233,49731,50748,13909,27505,64496,204928,17155,116631,100683,15948,197453,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 Blair (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,NA,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,524649,5,989,962,416511,16.74,1112189,688072,424117,52.68,64.98,40.3,1112189,46680,347433,305503,114789,254169,43564,243389,659331,1417252,168031,1123248,912287,210961,1364755,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,484588,5,1017,958,165735,7.07,1044788,642996,401792,52.02,64.91,39.48,1044788,43604,288554,304556,106517,249549,51926,216402,567129,1382370,179784,1159606,971911,187695,1186079,1.Madiga,168229,2.Mala,157187,3.Gosangi,9760,1.Hindus,1983275,2.Muslims,338824,3.Christians,16204,1.Sugalis etc.,142355,2.Gond etc.,13971,3.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
Andhra,District Karimnagar (03)& Andhra Pradesh (28),3491822,1747968,1743854,14.47,2813010,NA,NA,NA,813797,4,998,961,90636,2.6,1661089,1013328,647761,54.9,67.09,42.75,1661089,57595,445208,478883,189226,409110,81001,302570,796148,2077569,315535,1711559,1458954,252605,1780263,1.Madiga,396594,2.Mala,193030,3.Mala Sale etc.,25595,1.Hindus,3251834,2.Muslims,213811,3.Christians,20576,1.Sugalis etc.,51157,2.Gond etc.,13275,3.Yerukulas,13215,1.Ramagundam (M),236600,2.Karimnagar (M),205653,3.Jagtial (M),85521,1047,1047,1046,1047,-,-,1039,668,391,NA,863,70,475,795,912,823,218,55.8,36.4,7.8
```

```
scala> val census_map = census_data.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_map: Unit = ()
```

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

Below is the code used for same-

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

Below is the screenshot with 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|
+-----------+----------------+
```

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

Below is the code used for same-

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

Below is the screenshot with 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|
+----------------+------------------+
```

3. **Find the literacy rate of each state**

Below is the code used for same-

- val literacy = spark.sql("select state,avg(Persons_literacy_rate) from census group by state").show

Below is the screenshot with output-

```
scala> val literacy = spark.sql("select state,avg(Persons_literacy_rate) from census group by state").show
+----------------+----------------------------------------+
|           state|avg(CAST(Persons_literacy_rate AS DOUBLE))|
+----------------+----------------------------------------+
|        Nagaland|                                68.52875|
|       Karnataka|                   65.72666666666666|
|           D_N_H|                                57.63|
|          Kerala|                   90.52285714285713|
|          Punjab|                   68.61176470588235|
|              CG|                   63.02312499999999|
|         Manipur|                   68.61250000000001|
|              HP|                   75.50833333333333|
|             Goa|                   81.78999999999999|
|         Mizoram|                   85.55375000000001|
|          Orrisa|                   59.97965517241381|
|ArunachalPradesh|                  53.166923076923084|
|        Meghalya|                  60.722857142857144|
|              WB|                                66.07|
|         Haryana|                   68.24473684210527|
|       Jharkhand|                   50.51166666666667|
|         Gujarat|                   67.07480000000001|
|              TN|                   72.94266666666665|
|          Andhra|                   59.29363636363637|
|              UP|                   56.01057142857144|
+----------------+----------------------------------------+
```

4. Find out the States with More Female Population

Below is the code used for same-

- val female_pop = spark.sql("select state, sum(Males)-sum(Females) from census group by state").show

Below is the screenshot with output-

```
scala> val female_pop = spark.sql("select state, sum(Males)-sum(Females) from census group by state").show
+----------------+-------------------------------------------------------------+
|           state|(sum(CAST(Males AS DOUBLE)) - sum(CAST(Females AS DOUBLE)))|
+----------------+-------------------------------------------------------------+
|        Nagaland|                                                     104246.0|
|       Karnataka|                                                     947274.0|
|           D_N_H|                                                      22842.0|
|          Kerala|                                                    -904146.0|
|          Punjab|                                                    1611091.0|
|              CG|                                                     114633.0|
|         Manipur|                                                      20533.0|
|              HP|                                                      97980.0|
|             Goa|                                                      26828.0|
|         Mizoram|                                                      29645.0|
|          Orrisa|                                                     482015.0|
|ArunachalPradesh|                                                      61914.0|
|        Meghalya|                                                      33352.0|
|              WB|                                                    2755773.0|
|         Haryana|                                                    1583342.0|
|       Jharkhand|                                                     824245.0|
|         Gujarat|                                                    2100137.0|
|              TN|                                                     396139.0|
|          Andhra|                                                     826959.0|
|              UP|                                                    8932817.0|
+----------------+-------------------------------------------------------------+
```

5. Find out the Percentage of Population in Every State

Below is the code used for same-

- val percenet_pop = spark.sql("select state, (sum(persons) * 100.0) / SUM(sum(persons)) over() as percent_pop_by_state from census group by state").show

Below is the screenshot with output-

```
scala> val percenet_pop = spark.sql("select state, (sum(persons) * 100.0) / SUM(sum(persons)) over() as percent_pop_by_state from census group by state").show
18/01/11 00:48:19 WARN WindowExec: No Partition Defined for Window operation! Moving all data to a single partition, this can cause serious performance degrada
+----------------+--------------------+
|           state|percent_pop_by_state|
+----------------+--------------------+
|        Nagaland| 0.19464122457545488|
|       Karnataka|   5.169202018044398|
|           D_N_H| 0.02156566193106157|
|          Kerala|  3.1143376439044568|
|          Punjab|  2.3825023239741796|
|              CG|  2.0377103371415317|
|         Manipur| 0.19662075848548596|
|              HP|  0.5944665819347776|
|             Goa| 0.13181256512000492|
|         Mizoram| 0.08690945130876308|
|          Orrisa|   3.488284891601744|
|ArunachalPradesh| 0.10738993468694186|
|        Meghalya| 0.22679908989209513|
|              WB|   7.841864753141607|
|         Haryana|  2.0681052152192616|
|       Jharkhand|  2.6355147111714583|
|         Gujarat|   4.956025317815201|
|              TN|   6.103767861999858|
|          Andhra|   6.974542519042551|
|              UP|   16.25546817511578|
+----------------+--------------------+
```