**Problem Statement**

Go through below blog and reiterate the same at your end.

https://docs.google.com/document/d/1csLBlMiEXs\_hXWV2Z8VpBlrj\_R6RoDQLlZUnA0uBT

Ck/edit

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

**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

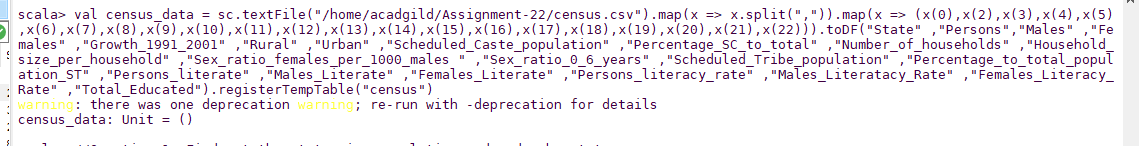
**Here is what we are taking**

"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"

Code:

val census\_data = sc.textFile("/home/acadgild/Assignment-22/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")

Screen-shot

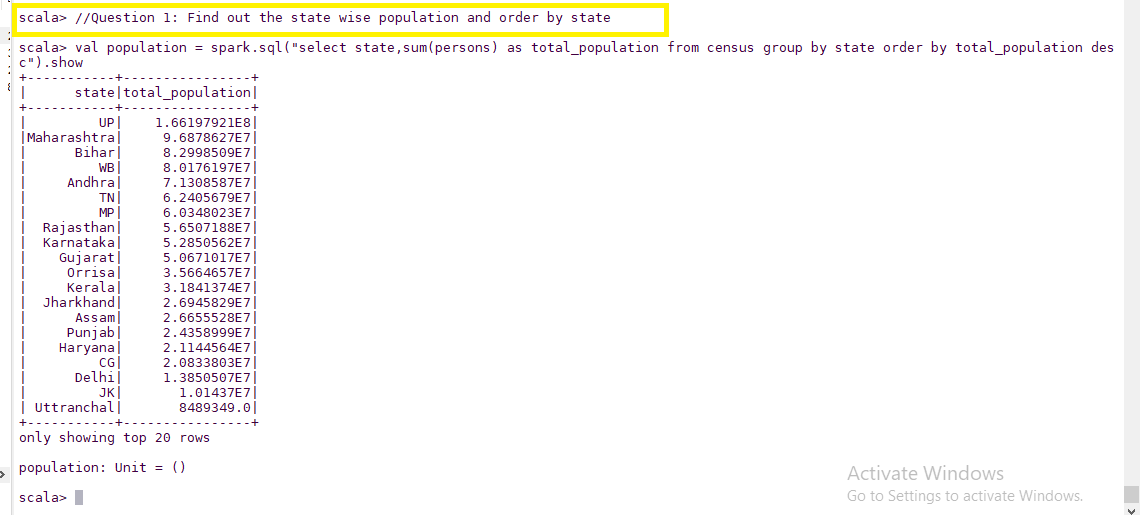


Question 1:

Find out the state wise population and order by state

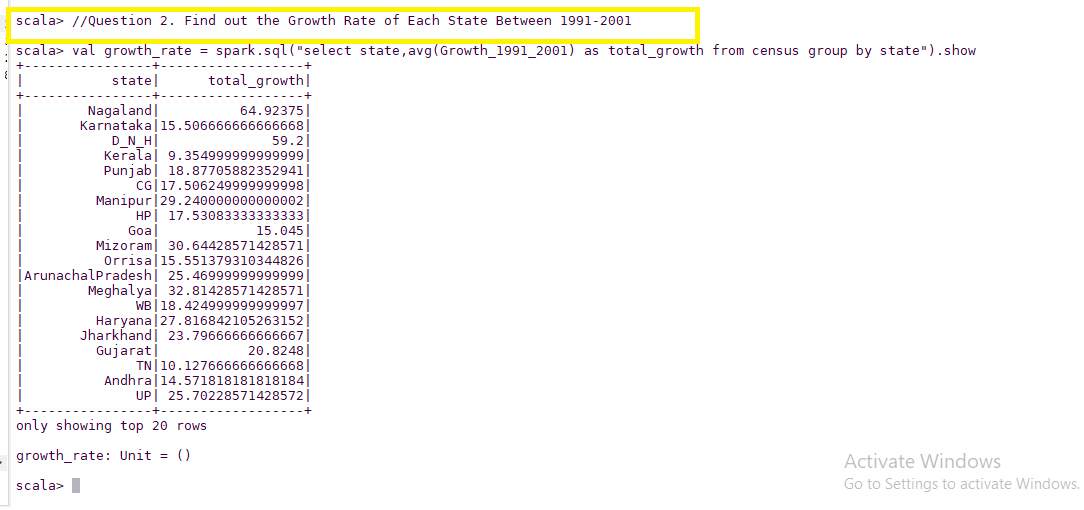
Code : val population = spark.sql("select state,sum(persons) as total\_population from census group by state order by total\_population desc").show

Screen-shot



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

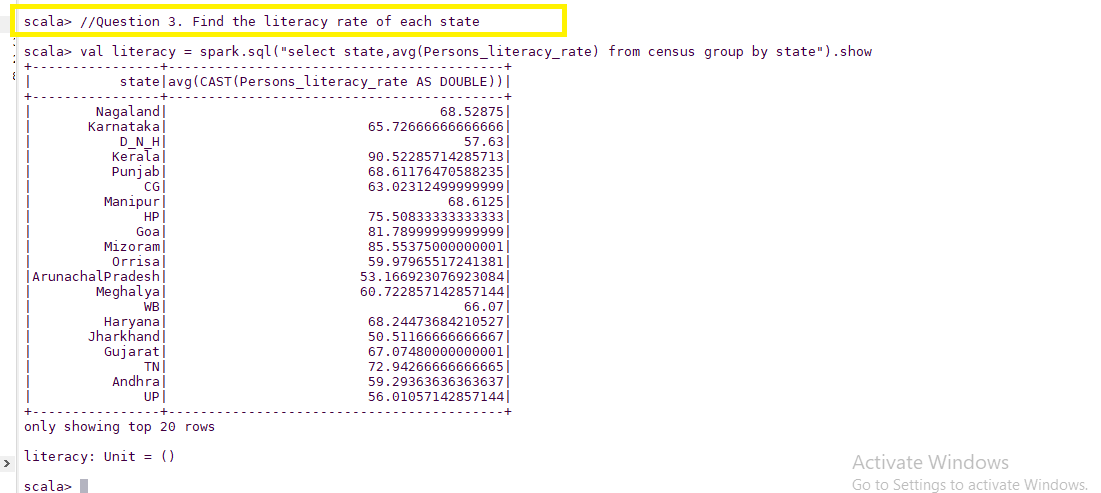
Code 2: val growth\_rate = spark.sql("select state,avg(Growth\_1991\_2001) as total\_growth from census group by state").show



Question 3:

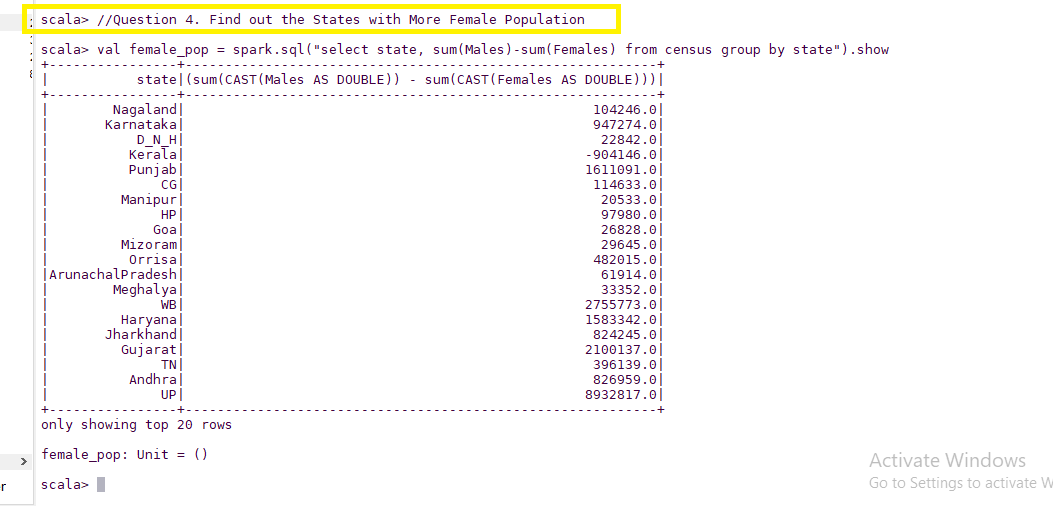
Find the literacy rate of each state

val literacy = spark.sql("select state,avg(Persons\_literacy\_rate) from census group by state").show



Question 4: Find out the States with More Female Population

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



Question 5:

Find out the Percentage of Population in Every State

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

