SESSION 22 ASSIGNMENT 1

PROBLEM : Census data analysis

1. Find out the state wise population and order by state.
2. Find out the Growth Rate of Each State Between 1991-2001.
3. Find the literacy rate of each state.
4. Find out the States with More Female Population.
5. Find out the Percentage of Population in Every State.

SOLUTION :

import org.apache.log4j.{Level, Logger}

import org.apache.spark.sql.SparkSession

object Assignment\_22\_1 extends App {

Logger.getLogger("org").setLevel(Level.OFF)

Logger.getLogger("akka").setLevel(Level.OFF)

val spark = SparkSession.builder

.master("local")

.appName("example")

.config("spark.sql.warehouse.dir","C://ACADGILD")

.getOrCreate()

// A CSV dataset is pointed to by path.

// The path can be either a single text file or a directory storing text files.

val dataset\_1 = spark.sqlContext.read.format("csv").option("header", "false")

.option("inferSchema", true)

.load("C:/ACADGILD/Big Data/census.csv")

import spark.implicits.\_

val dataset\_2 = dataset\_1.map(x => (x(0).toString,x(2).toString.toInt,x(3).toString,

x(4).toString,x(5).toString,x(6).toString,x(7).toString,x(8).toString,x(9).toString,x(10).toString,

x(11).toString,x(12).toString,x(13).toString,x(14).toString,x(15).toString,x(16).toString,

x(17).toString,x(18).toString,x(19).toString,x(20).toString,x(21).toString,x(22).toString))

.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")

.createOrReplaceTempView("census")

//--------------------------------------------------------------------------------------------------------------------

//------------------------------------------ PROBLEM 1 ---------------------------------------------------------------

//--------------------------------------------------------------------------------------------------------------------

//Spark SQL is a Spark module for structured data processing.One use of Spark SQL is to execute SQL queries.

println("The state wise population is : ")

val population = spark.sql("select State,sum(Persons) as total\_population from census " +

" group by State order by sum(Persons) desc")

.show()

//--------------------------------------------------------------------------------------------------------------------

//------------------------------------------ PROBLEM 2 ---------------------------------------------------------------

//--------------------------------------------------------------------------------------------------------------------

println("The Growth Rate of Each State Between 1991-2001 is : ")

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

//--------------------------------------------------------------------------------------------------------------------

//------------------------------------------ PROBLEM 3 ---------------------------------------------------------------

//--------------------------------------------------------------------------------------------------------------------

//// .show - Displays the content of the DataFrame to stdout

println("The literacy rate of each state : ")

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

//--------------------------------------------------------------------------------------------------------------------

//------------------------------------------ PROBLEM 4 ---------------------------------------------------------------

//--------------------------------------------------------------------------------------------------------------------

println("The States with More Female Population : ")

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

//--------------------------------------------------------------------------------------------------------------------

//------------------------------------------ PROBLEM 5 ---------------------------------------------------------------

//--------------------------------------------------------------------------------------------------------------------

println("The Percentage of Population in Every State :")

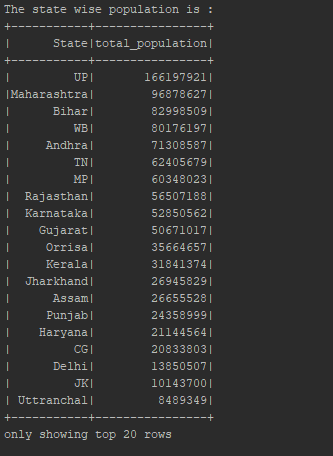
val percenet\_pop = spark.sql("select state, (sum(persons) \* 100.0) / SUM(sum(persons)) over()"

+ " as percent\_pop\_by\_state from census group by state").show

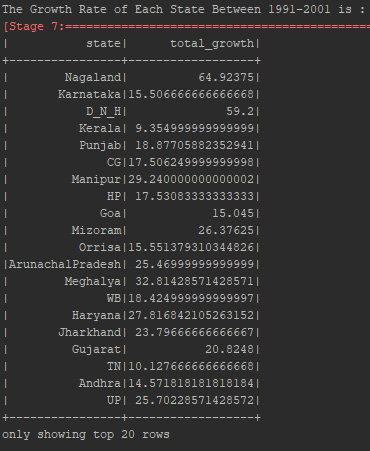
}

OUTPUT :

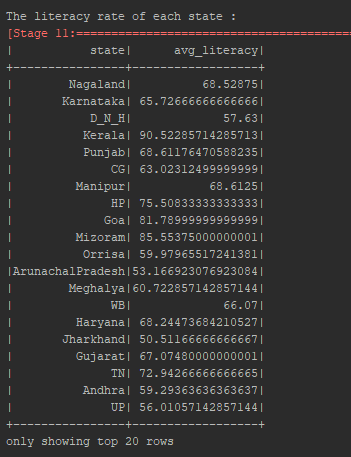
1.



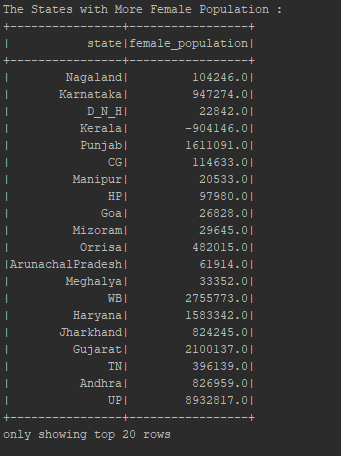
2.



3.



4.



5.

