Session 20

Assignment 1

|  |  |
| --- | --- |
| **Prepared For:** | AcadGild |
|  |  |
| **Document Approval:** | **AcadGild** |
|  |  |
|  |  |
|  |  |
|  |  |
| **Project Title:** | Session 20 – Assignment 1 |
|  |  |
| **Prepared By:** | Duncan Burgess |
|  |  |
|  | dburgess@duncb.com |
|  |  |
| **Primary Engineer:** | Duncan Burgess |
|  |  |
| **Document Reference:** | **Session 20 – Assignment 1** |
|  |  |
| **Start Date:** | 20/10/2017 |
|  |  |
|  |  |



# 

# Contents

[Contents 2](#_Toc496382553)

[Change History 3](#_Toc496382554)

[1. Problem Statement 4](#_Toc496382555)

[2. Dataset 4](#_Toc496382556)

[3. Solution 5](#_Toc496382557)

[4. Results 6](#_Toc496382558)

# Change History

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Document Revision** | **Date** | **Authored By** | **Authorised By** | **Sections Affected** | **Reason for Change** |
| Rev 01 | 21/10/2017 | Duncan Burgess |  | All | Initial release. |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

# Problem Statement

**Follow the blog and iterate the same results.**

* Find out the state wise population and order by state

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

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

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

* Find the literacy rate of each state

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

* 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

* 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

# Dataset

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"

# Solution

I have set the environment and followed the blog as follows.

**Code written**

**package** com.duncb.spark

**import** org.apache.spark.\_

**import** org.apache.spark.SparkContext.\_

**import** org.apache.log4j.\_

**import** org.apache.spark.sql.\_

**import** org.apache.spark.sql.functions.\_

**object** deploy1 {

**def** main(args: Array[*String*]) {

// Create a SparkContext using every core of the local machine

**val** sc = **new** SparkContext("local[\*]", "deploy1")

**val** sqlContext = **new** org.apache.spark.sql.SQLContext(sc)

**import** sqlContext.implicits.\_

**val** spark = SparkSession

.builder

.appName("SparkSQL")

.master("local[1]")

.config("spark.sql.warehouse.dir", "file:///C:/temp") // Necessary to work around a Windows bug in Spark 2.0.0; omit if you're not on Windows.

.getOrCreate()

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

**val** census\_data = sc.textFile("file:///N:/Datasets/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")

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

println("The state wise population and order by state")

println("----------------------------------------------------------------------------------")

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

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

println("----------------------------------------------------------------------------------")

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

println("The literacy rate of each state")

println("----------------------------------------------------------------------------------")

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

println("The States with More Female Population")

println("----------------------------------------------------------------------------------")

**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

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

println("----------------------------------------------------------------------------------")

}

}

# Results

+-----------+----------------+

| 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

The state wise population and order by state

----------------------------------------------------------------------------------

+----------------+------------------+

| 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

The Growth Rate of Each State Between 1991-2001

----------------------------------------------------------------------------------

+----------------+------------------------------------------+

| 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|

+----------------+------------------------------------------+

only showing top 20 rows

The literacy rate of each state

----------------------------------------------------------------------------------

+----------------+-----------------------------------------------------------+

| 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|

+----------------+-----------------------------------------------------------+

only showing top 20 rows

The States with More Female Population

----------------------------------------------------------------------------------

+----------------+--------------------+

| 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|

+----------------+--------------------+

only showing top 20 rows

The Percentage of Population in Every State