Case Study 4

```
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions._
object CaseStudy4 {

  case class Hospital(DRGDefinition: String, ProviderId: Int, ProviderName: String, street:
String, city: String, state: String, zip: Int, HReferra: String, TotalDischarges: Int,
avg_cover_charge: Double, avg_tot_pay: Double, avg_medicare: Double)

  def main(args: Array[String]): Unit = {
    println("hey scala")
    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Case study 4")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()
    println("Spark Session Object created")

  val hosp = spark.sqlContext.read.csv("E:\\Avani\\Acadgild\\Case Study
4\\inpatientCharges.csv")
  val hosDF= hosp.toDF()
    hosDF.show(truncate = false)

hosDF.registerTempTable("medical")
```

Objective 1

- Load file into spark

```
    val hosp = spark.sqlContext.read.csv("E:\\Avani\\Acadgild\\Case Study
4\\inpatientCharges.csv")
  val hosDF= hosp.toDF()
    hosDF.show(truncate = false)

hosDF.registerTempTable("medical")
```

Objective2

- What is the average amount of **AverageCoveredCharges** per state?

```
val avg_cover=spark.sql("SELECT (_c5),AVG(_c9) FROM MEDICAL GROUP BY (_c5)")
avg_cover.show()
println("Average covered charges")
```

- Find out the **AverageTotalPayments** charges per state

```
val avg_payment=spark.sql("SELECT (_c5),sum(_c10) FROM MEDICAL GROUP BY (_c5) order by (_c5)")
avg_payment.show()
println("Average Total payment charges")
```

- Find out the AverageMedicarePayments charges per state

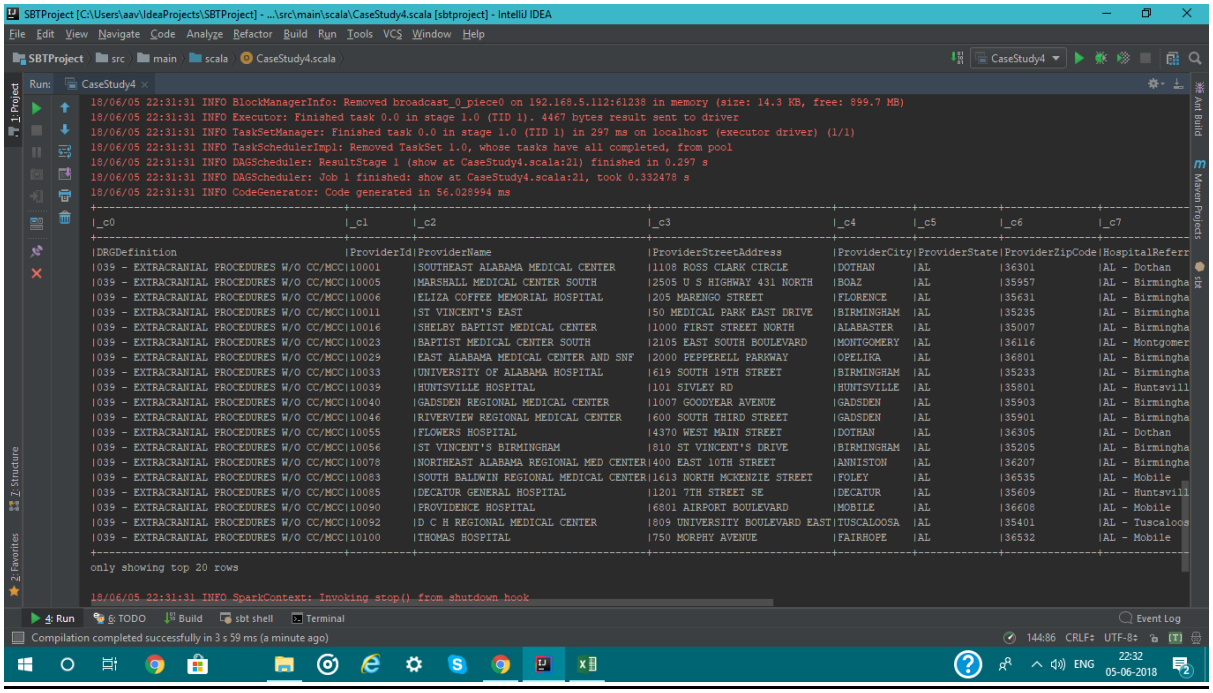
```
val avg_med=spark.sql("SELECT (_c5),sum(_c11) FROM MEDICAL GROUP BY (_c5) order by (_c5)")
avg_med.show()
println("Average Medicare payment charges")
```

Objective3

- Find out the total number of **Discharges** per state and for each disease
- Sort the output in descending order of **totalDischarges**

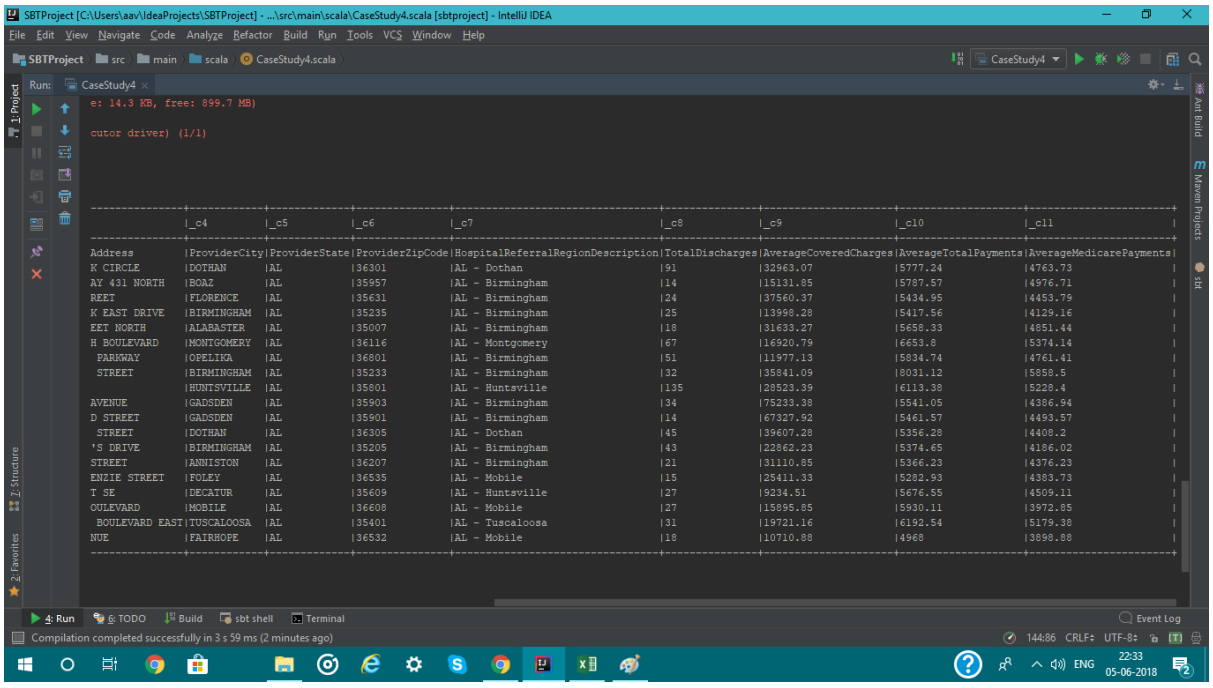
```
val tot_dis=spark.sql("SELECT (_c5), (_c0), sum(_c8) total_discharge FROM MEDICAL GROUP BY (_c5), (_c0)")
tot_dis.show(truncate = false)        //to display the complete content of a column
println("Total discharges")
val order= tot_dis.orderBy(desc("total_discharge"))        //sorting the above data in descending order
 order.show(truncate = false)
println("sorted data")
```

Screenshots



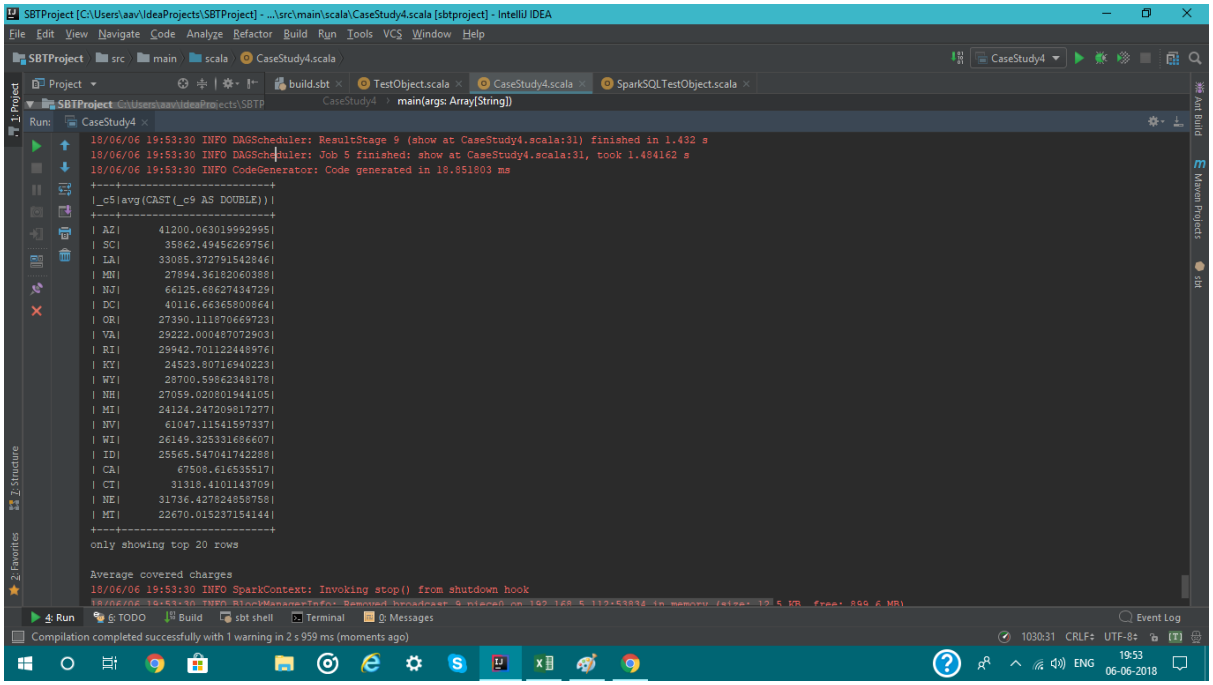
```
18/06/05 22:31:31 INFO BlockManagerInfo: Removed broadcast_0_piece0 on 192.168.5.112:61238 in memory (size: 14.3 KB, free: 899.7 MB)
18/06/05 22:31:31 INFO Executor: Finished task 0.0 in stage 1.0 (TID 1). 4467 bytes result sent to driver
18/06/05 22:31:31 INFO TaskSetManager: Finished task 0.0 in stage 1.0 (TID 1) in 297 ms on localhost (executor driver) (1/1)
18/06/05 22:31:31 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/06/05 22:31:31 INFO DAGScheduler: ResultStage 1 (show at CaseStudy4.scala:21) finished in 0.297 s
18/06/05 22:31:31 INFO DAGScheduler: Job 1 finished: show at CaseStudy4.scala:21, took 0.332478 s
18/06/05 22:31:31 INFO CodeGenerator: Code generated in 56.028994 ms
+------------------------------------+---------+--------------------------------+--------------------------+--------------+-------------+-------------+--------------------+
|_c0                                 |_c1      |_c2                             |_c3                       |_c4           |_c5          |_c6          |_c7                 |
+------------------------------------+---------+--------------------------------+--------------------------+--------------+-------------+-------------+--------------------+
|DRGDefinition                       |ProviderId|ProviderName                    |ProviderStreetAddress     |ProviderCity  |ProviderState|ProviderZipCode|HospitalReferr      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10001    |SOUTHEAST ALABAMA MEDICAL CENTER|1108 ROSS CLARK CIRCLE    |DOTHAN        |AL           |36301        |AL - Dothan         |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10005    |MARSHALL MEDICAL CENTER SOUTH   |2505 U S HIGHWAY 431 NORTH|BOAZ          |AL           |35957        |AL - Birmingha      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10006    |ELIZA COFFEE MEMORIAL HOSPITAL  |205 MARENGO STREET        |FLORENCE      |AL           |35631        |AL - Birmingha      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10011    |ST VINCENT'S EAST               |50 MEDICAL PARK EAST DRIVE|BIRMINGHAM    |AL           |35235        |AL - Birmingha      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10016    |SHELBY BAPTIST MEDICAL CENTER   |1000 FIRST STREET NORTH   |ALABASTER     |AL           |35007        |AL - Birmingha      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10023    |BAPTIST MEDICAL CENTER SOUTH    |2105 EAST SOUTH BOULEVARD |MONTGOMERY    |AL           |36116        |AL - Montgomer      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10029    |EAST ALABAMA MEDICAL CENTER AND SNF|2000 PEPPERELL PARKWAY|OPELIKA       |AL           |36801        |AL - Birmingha      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10033    |UNIVERSITY OF ALABAMA HOSPITAL  |619 SOUTH 19TH STREET     |BIRMINGHAM    |AL           |35233        |AL - Birmingha      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10039    |HUNTSVILLE HOSPITAL             |101 SIVLEY RD             |HUNTSVILLE    |AL           |35801        |AL - Huntsvill      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10040    |GADSDEN REGIONAL MEDICAL CENTER |1007 GOODYEAR AVENUE      |GADSDEN       |AL           |35903        |AL - Birmingha      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10046    |RIVERVIEW REGIONAL MEDICAL CENTER|600 SOUTH THIRD STREET   |GADSDEN       |AL           |35901        |AL - Birmingha      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10055    |FLOWERS HOSPITAL                |4370 WEST MAIN STREET     |DOTHAN        |AL           |36305        |AL - Dothan         |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10056    |ST VINCENT'S BIRMINGHAM         |810 ST VINCENT'S DRIVE    |BIRMINGHAM    |AL           |35205        |AL - Birmingha      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10078    |NORTHEAST ALABAMA REGIONAL MED CENTER|400 EAST 10TH STREET  |ANNISTON      |AL           |36207        |AL - Birmingha      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10083    |SOUTH BALDWIN REGIONAL MEDICAL CENTER|1613 NORTH MCKENZIE STREET|FOLEY     |AL           |36535        |AL - Mobile         |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10085    |DECATUR GENERAL HOSPITAL        |1201 7TH STREET SE        |DECATUR       |AL           |35609        |AL - Huntsvill      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10090    |PROVIDENCE HOSPITAL             |6801 AIRPORT BOULEVARD    |MOBILE        |AL           |36608        |AL - Mobile         |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10092    |D C H REGIONAL MEDICAL CENTER   |809 UNIVERSITY BOULEVARD EAST|TUSCALOOSA |AL           |35401        |AL - Tuscaloos      |
|039 - EXTRACRANIAL PROCEDURES W/O CC/MCC|10100    |THOMAS HOSPITAL                 |750 MORPHY AVENUE         |FAIRHOPE      |AL           |36532        |AL - Mobile         |
+------------------------------------+---------+--------------------------------+--------------------------+--------------+-------------+-------------+--------------------+
only showing top 20 rows

18/06/05 22:31:31 INFO SparkContext: Invoking stop() from shutdown hook
```

1.a

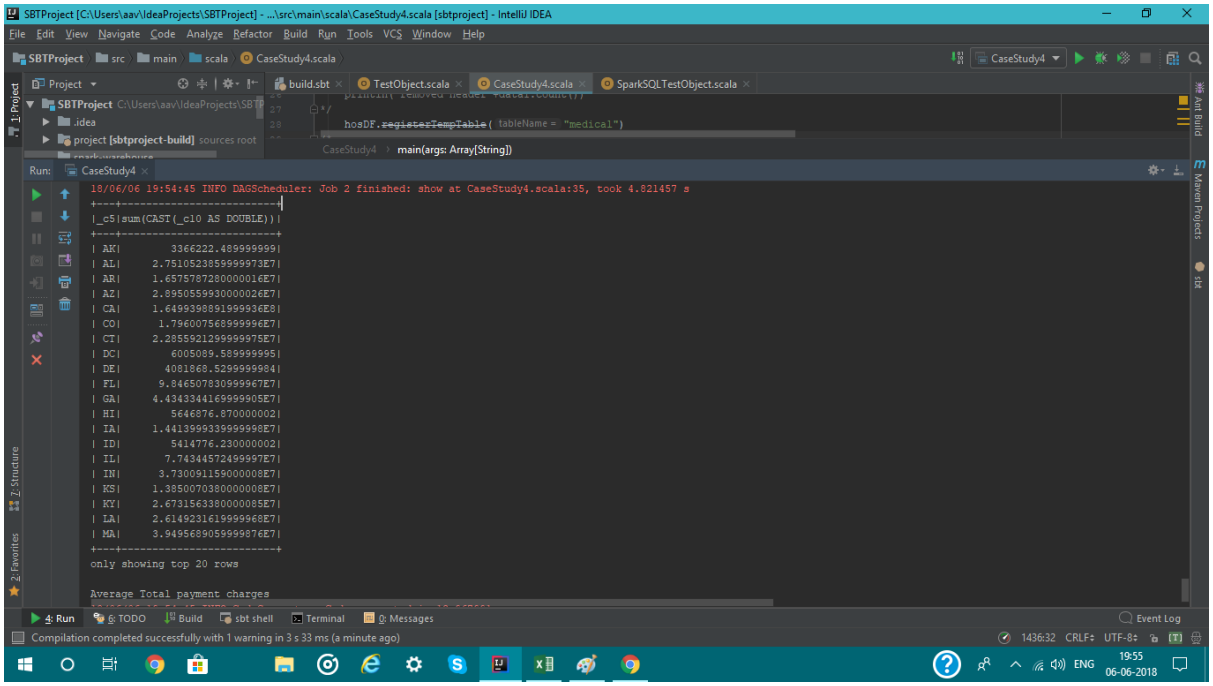


```
e: 14.3 KB, free: 899.7 MB)

cutor driver) (1/1)

+--------------+------------+-----------+-------------+--------------------------------+---------------+-----------------+-------------------+---------------------+
|              |_c4         |_c5        |_c6          |_c7                             |_c8            |_c9              |_c10               |_c11                 |
+--------------+------------+-----------+-------------+--------------------------------+---------------+-----------------+-------------------+---------------------+
|Address       |ProviderCity|ProviderState|ProviderZipCode|HospitalReferralRegionDescription|TotalDischarges|AverageCoveredCharges|AverageTotalPayments|AverageMedicarePayments|
|K CIRCLE      |DOTHAN      |AL         |36301        |AL - Dothan                     |91             |32963.07         |5777.24            |4763.73              |
|AY 431 NORTH  |BOAZ        |AL         |35957        |AL - Birmingham                 |14             |15131.85         |5787.57            |4976.71              |
|REET          |FLORENCE    |AL         |35631        |AL - Birmingham                 |24             |37560.37         |5434.95            |4453.79              |
|K EAST DRIVE  |BIRMINGHAM  |AL         |35235        |AL - Birmingham                 |25             |13998.28         |5417.56            |4129.16              |
|EET NORTH     |ALABASTER   |AL         |35007        |AL - Birmingham                 |18             |31633.27         |5658.33            |4851.44              |
|H BOULEVARD   |MONTGOMERY  |AL         |36116        |AL - Montgomery                 |67             |16920.79         |6653.8             |5374.14              |
| PARKWAY      |OPELIKA     |AL         |36801        |AL - Birmingham                 |51             |11977.13         |5834.74            |4761.41              |
| STREET       |BIRMINGHAM  |AL         |35233        |AL - Birmingham                 |32             |35841.09         |8031.12            |5858.5               |
|              |HUNTSVILLE  |AL         |35801        |AL - Huntsville                 |135            |28523.39         |6113.38            |5228.4               |
|AVENUE        |GADSDEN     |AL         |35903        |AL - Birmingham                 |34             |75233.38         |5541.05            |4386.94              |
|D STREET      |GADSDEN     |AL         |35901        |AL - Birmingham                 |14             |67327.92         |5461.57            |4493.57              |
| STREET       |DOTHAN      |AL         |36305        |AL - Dothan                     |45             |39607.28         |5356.28            |4408.2               |
|'S DRIVE      |BIRMINGHAM  |AL         |35205        |AL - Birmingham                 |43             |22862.23         |5374.65            |4186.02              |
|STREET        |ANNISTON    |AL         |36207        |AL - Birmingham                 |21             |31110.85         |5366.23            |4376.23              |
|ENZIE STREET  |FOLEY       |AL         |36535        |AL - Mobile                     |15             |25411.33         |5282.93            |4383.73              |
|T SE          |DECATUR     |AL         |35609        |AL - Huntsville                 |27             |9234.51          |5676.55            |4509.11              |
|OULEVARD      |MOBILE      |AL         |36608        |AL - Mobile                     |27             |15895.85         |5930.11            |3972.85              |
| BOULEVARD EAST|TUSCALOOSA |AL         |35401        |AL - Tuscaloosa                 |31             |19721.16         |6192.54            |5179.38              |
|NUE           |FAIRHOPE    |AL         |36532        |AL - Mobile                     |18             |10710.88         |4968               |3898.88              |
+--------------+------------+-----------+-------------+--------------------------------+---------------+-----------------+-------------------+---------------------+
```

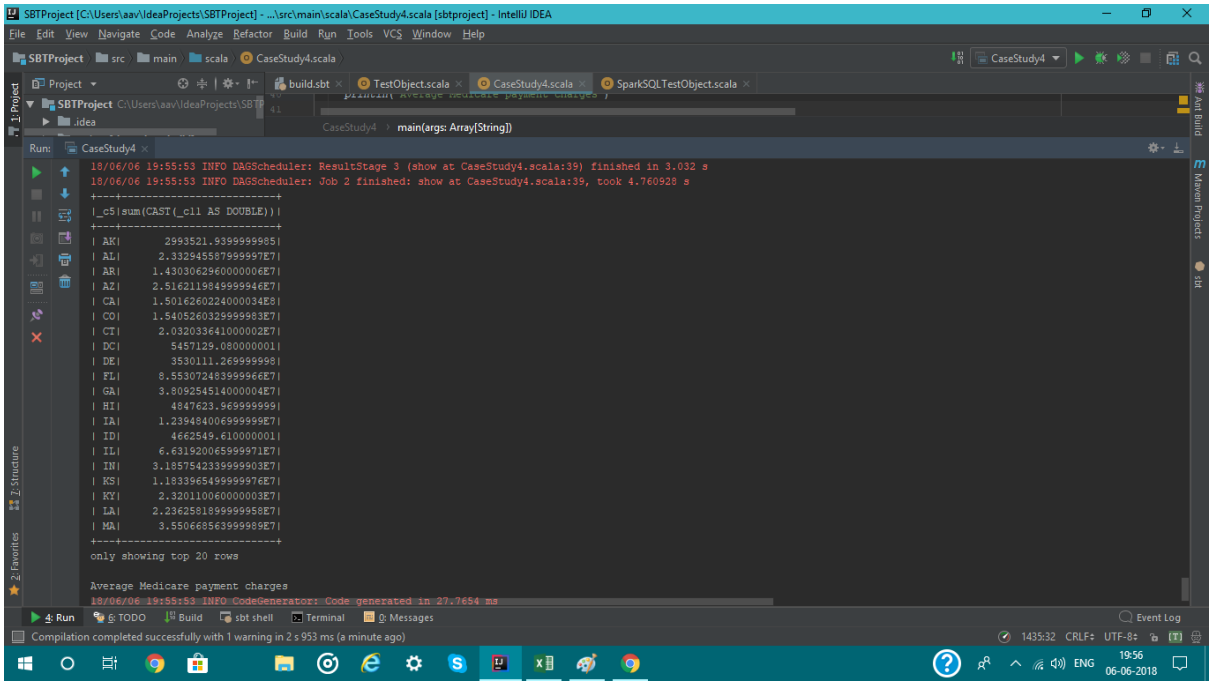
1.b(contd..)


SBTProject [C:\Users\aav\IdeaProjects\SBTProject] - ...\src\main\scala\CaseStudy4.scala [sbtproject] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
SBTProject src main scala CaseStudy4.scala
CaseStudy4
Project
build.sbt
TestObject.scala
CaseStudy4.scala
SparkSQLTestObject.scala
CaseStudy4 main(args: Array[String])
Run: CaseStudy4
18/06/06 19:53:30 INFO DAGScheduler: ResultStage 9 (show at CaseStudy4.scala:31) finished in 1.432 s
18/06/06 19:53:30 INFO DAGScheduler: Job 5 finished: show at CaseStudy4.scala:31, took 1.484162 s
18/06/06 19:53:30 INFO CodeGenerator: Code generated in 18.851803 ms
+---+------------------------+
|_c5|avg(CAST(_c9 AS DOUBLE))|
+---+------------------------+
AZ	41200.063019992995
SC	35862.494562697564
LA	33085.372791542846
MN	27894.361820603884
NJ	66125.686274347294
DC	40116.663658008644
OR	27390.111870669723
VA	29222.000487072903
RI	29942.701122448976
KY	24523.807169402234
WY	28700.598623481784
NH	27059.020801944105
MI	24124.247209817277
NV	61047.115415973374
WI	26149.325331686607
ID	25565.547041742288
CA	67508.61653551714
CT	31318.410114370904
NE	31736.427824858758
MT	22670.015237154144
+---+------------------------+
only showing top 20 rows
Average covered charges
18/06/06 19:53:30 INFO SparkContext: Invoking stop() from shutdown hook
4: Run 6: TODO Build sbt shell Terminal 0: Messages Event Log
Compilation completed successfully with 1 warning in 2 s 959 ms (moments ago)
1030:31 CRLF: UTF-8:
ENG 19:53 06-06-2018

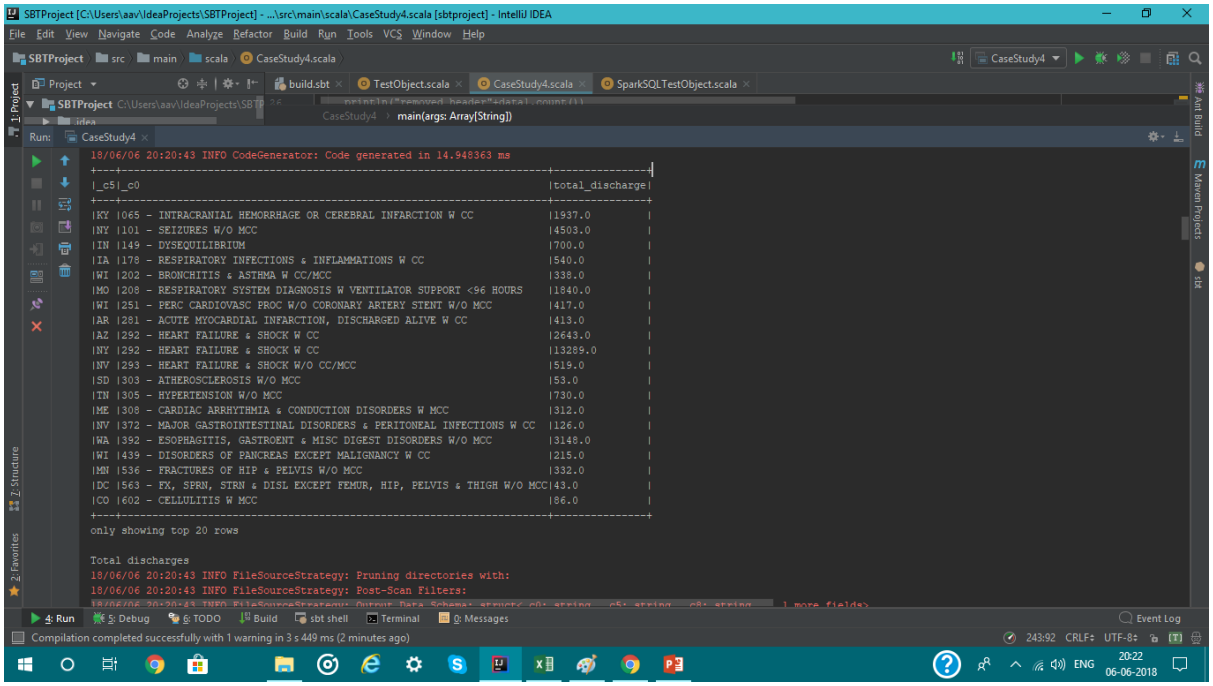
2.a


SBTProject [C:\Users\aav\IdeaProjects\SBTProject] - ...\src\main\scala\CaseStudy4.scala [sbtproject] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
SBTProject src main scala CaseStudy4.scala
CaseStudy4
Project
build.sbt
TestObject.scala
CaseStudy4.scala
SparkSQLTestObject.scala
hosDF.registerTempTable(tableName = "medical")
CaseStudy4 main(args: Array[String])
Run: CaseStudy4
18/06/06 19:54:45 INFO DAGScheduler: Job 2 finished: show at CaseStudy4.scala:35, took 4.821457 s
+---+--------------------------+
|_c5|sum(CAST(_c10 AS DOUBLE))|
+---+--------------------------+
AK	3366222.489999999
AL	2.7510523859999973E7
AR	1.6575787280000016E7
AZ	2.8950559930000026E7
CA	1.6499398891999936E8
CO	1.796007568999996E7
CT	2.2855921299999975E7
DC	6005089.589999995
DE	4081868.5299999984
FL	9.846507830999967E7
GA	4.4343344169999905E7
HI	5646876.870000002
IA	1.4413999339999998E7
ID	5414776.230000002
IL	7.7434457249999975E7
IN	3.7300911590000008E7
KS	1.3850070380000008E7
KY	2.6731563380000085E7
LA	2.6149231619999968E7
MA	3.9495689059999876E7
+---+--------------------------+
only showing top 20 rows
Average Total payment charges
4: Run 6: TODO Build sbt shell Terminal 0: Messages Event Log
Compilation completed successfully with 1 warning in 3 s 33 ms (a minute ago)
1436:32 CRLF: UTF-8:
ENG 19:55 06-06-2018

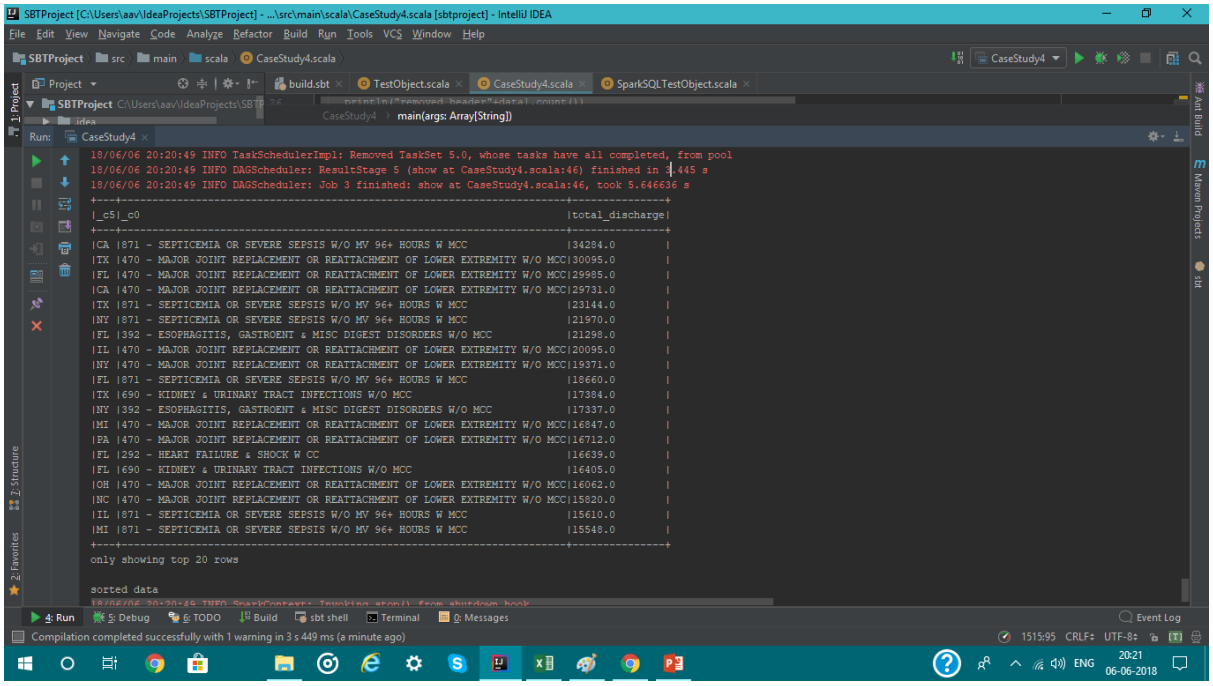
2.b


SBTProject [C:\Users\aav\IdeaProjects\SBTProject] - ...\src\main\scala\CaseStudy4.scala [sbtproject] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
build.sbt
TestObject.scala
CaseStudy4.scala
SparkSQLTestObject.scala
CaseStudy4 › main(args: Array[String])
Run: CaseStudy4
18/06/06 19:55:53 INFO DAGScheduler: ResultStage 3 (show at CaseStudy4.scala:39) finished in 3.032 s
18/06/06 19:55:53 INFO DAGScheduler: Job 2 finished: show at CaseStudy4.scala:39, took 4.760928 s
+---+--------------------------+
|_c5|sum(CAST(_c11 AS DOUBLE))|
+---+--------------------------+
AK	2993521.9399999985
AL	2.3329455879999997E7
AR	1.4303062960000006E7
AZ	2.5162119849999946E7
CA	1.5016260224000034E8
CO	1.5405260329999983E7
CT	2.0320336410000002E7
DC	5457129.080000001
DE	3530111.269999998
FL	8.553072483999966E7
GA	3.8092545140000047E7
HI	4847623.969999999
IA	1.2394840069999997E7
ID	4662549.610000001
IL	6.631920065999971E7
IN	3.1857542339999903E7
KS	1.1833965499999976E7
KY	2.3201100600000037E7
LA	2.2362581899999958E7
MA	3.5506685639999889E7
+---+--------------------------+
only showing top 20 rows
Average Medicare payment charges
18/06/06 19:55:53 INFO CodeGenerator: Code generated in 27.7654 ms
4: Run
6: TODO
Build
sbt shell
Terminal
0: Messages
Event Log
Compilation completed successfully with 1 warning in 2 s 953 ms (a minute ago)
1435:32 CRLF: UTF-8:
ENG
19:56
06-06-2018

2.c


SBTProject [C:\Users\aav\IdeaProjects\SBTProject] - ...\src\main\scala\CaseStudy4.scala [sbtproject] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
build.sbt
TestObject.scala
CaseStudy4.scala
SparkSQLTestObject.scala
CaseStudy4 › main(args: Array[String])
Run: CaseStudy4
18/06/06 20:20:43 INFO CodeGenerator: Code generated in 14.948363 ms
+---+--+---------------+
|_c5|_c0 |total_discharge|
+---+--+---------------+
KY	065 - INTRACRANIAL HEMORRHAGE OR CEREBRAL INFARCTION W CC	1937.0
NY	101 - SEIZURES W/O MCC	4503.0
IN	149 - DYSEQUILIBRIUM	700.0
IA	178 - RESPIRATORY INFECTIONS & INFLAMMATIONS W CC	540.0
WI	202 - BRONCHITIS & ASTHMA W CC/MCC	338.0
MO	208 - RESPIRATORY SYSTEM DIAGNOSIS W VENTILATOR SUPPORT <96 HOURS	1840.0
WI	251 - PERC CARDIOVASC PROC W/O CORONARY ARTERY STENT W/O MCC	417.0
AR	281 - ACUTE MYOCARDIAL INFARCTION, DISCHARGED ALIVE W CC	413.0
AZ	292 - HEART FAILURE & SHOCK W CC	2643.0
NY	292 - HEART FAILURE & SHOCK W CC	13289.0
NV	293 - HEART FAILURE & SHOCK W/O CC/MCC	519.0
SD	303 - ATHEROSCLEROSIS W/O MCC	53.0
TN	305 - HYPERTENSION W/O MCC	730.0
ME	308 - CARDIAC ARRHYTHMIA & CONDUCTION DISORDERS W MCC	312.0
NV	372 - MAJOR GASTROINTESTINAL DISORDERS & PERITONEAL INFECTIONS W CC	126.0
WA	392 - ESOPHAGITIS, GASTROENT & MISC DIGEST DISORDERS W/O MCC	3148.0
WI	439 - DISORDERS OF PANCREAS EXCEPT MALIGNANCY W CC	215.0
MN	536 - FRACTURES OF HIP & PELVIS W/O MCC	332.0
DC	563 - FX, SPRN, STRN & DISL EXCEPT FEMUR, HIP, PELVIS & THIGH W/O MCC	43.0
CO	602 - CELLULITIS W MCC	86.0
+---+--+---------------+
only showing top 20 rows
Total discharges
18/06/06 20:20:43 INFO FileSourceStrategy: Pruning directories with:
18/06/06 20:20:43 INFO FileSourceStrategy: Post-Scan Filters:
4: Run
5: Debug
6: TODO
Build
sbt shell
Terminal
0: Messages
Event Log
Compilation completed successfully with 1 warning in 3 s 449 ms (2 minutes ago)
243:92 CRLF: UTF-8:
ENG
20:22
06-06-2018

3.a



3.b

Complete code

```
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions._
object CaseStudy4 {

 case class Hospital(DRGDefinition: String, ProviderId: Int, ProviderName: String, street: String, city: String, state: String, zip: Int, HReferra: String, TotalDischarges: Int, avg_cover_charge: Double, avg_tot_pay: Double, avg_medicare: Double)

 def main(args: Array[String]): Unit = {
  println("hey scala")
  val spark = SparkSession
    .builder()
    .master("local")
    .appName("Case study 4")
    .config("spark.some.config.option", "some-value")
    .getOrCreate()
  println("Spark Session Object created")

 val hosp = spark.sqlContext.read.csv("E:\\Avani\\Acadgild\\Case Study
```

```
4\\inpatientCharges.csv")
  val hosDF= hosp.toDF()
    hosDF.show(truncate = false)
   val header=hosp.first()
    val data1 = hosp.filter(row => row != header)  //removing of header
    data1.count()
    println("Hospital Data->>"+hosp.show(truncate = false))
    println("removed header"+data1.count())

    hosDF.registerTempTable("medical")

    val avg_cover=spark.sql("SELECT (_c5),AVG(_c9) FROM MEDICAL GROUP BY (_c5)")
    avg_cover.show()
    println("Average covered charges")

    val avg_payment=spark.sql("SELECT (_c5),sum(_c10) FROM MEDICAL GROUP BY (_c5)
order by (_c5)")
    avg_payment.show()
    println("Average Total payment charges")

    val avg_med=spark.sql("SELECT (_c5),sum(_c11) FROM MEDICAL GROUP BY (_c5) order by
(_c5)")
    avg_med.show()
    println("Average Medicare payment charges")

    val tot_dis=spark.sql("SELECT (_c5), (_c0), sum(_c8) total_discharge FROM MEDICAL
GROUP BY (_c5), (_c0)")
    tot_dis.show(truncate = false)
    println("Total discharges")
    val order= tot_dis.orderBy(desc("total_discharge"))
      order.show(truncate = false)
    println("sorted data")

  }
}
```