Interview Preparation II

Assignment

# Hospital CASE STUDY

## Dataset Description

- **DRG Definition**: The code and description identifying the MS-DRG. MS-DRGs are a classification system that groups similar clinical conditions (diagnoses) and procedures furnished by the hospital during their stay.
- **Provider Id**: The CMS Certification Number (CCN) assigned to the Medicare-certified hospital facility.
- **Provider Name**: The name of the provider.
- **Provider Street Address**: The provider's street address.
- **Provider City**: The city where the provider is located.
- **Provider State**: The state where the provider is located.
- **Provider Zip Code**: The provider's zip code.
- **Provider HRR**: The Hospital Referral Region (HRR) where the provider is located.
- **Total Discharges**: The number of discharges billed by the provider for inpatient hospital services.
- **Average Covered Charges**: The provider's average charge for services covered by Medicare for all discharges in the MS-DRG. These will vary from hospital to hospital because of the differences in hospital charge structures.
- **Average Total Payments**: The average total payments to all providers for the MS-DRG including the MSDRG amount, teaching, disproportionate share, capital, and outlier payments for all cases. Also included in the average total payments are co-payment and deductible amounts that the patient is responsible for and any additional payments by third parties for coordination of benefits.
- **Average Medicare Payments**: The average amount that Medicare pays to the provider for Medicare's share of the MS-DRG. Average Medicare payment amounts include the MS-DRG amount, teaching, disproportionate share, capital, and outlier payments for all cases. Medicare payments DO NOT include beneficiary c

## Load file into spark

//StructType objects define the schema of Spark DataFrames. StructType objects contain a list of StructField objects that define the name, type, and nullable flag for each column in a DataFrame.

```
package Usecase

import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.types._

object HospitalData_Anlysis {
```

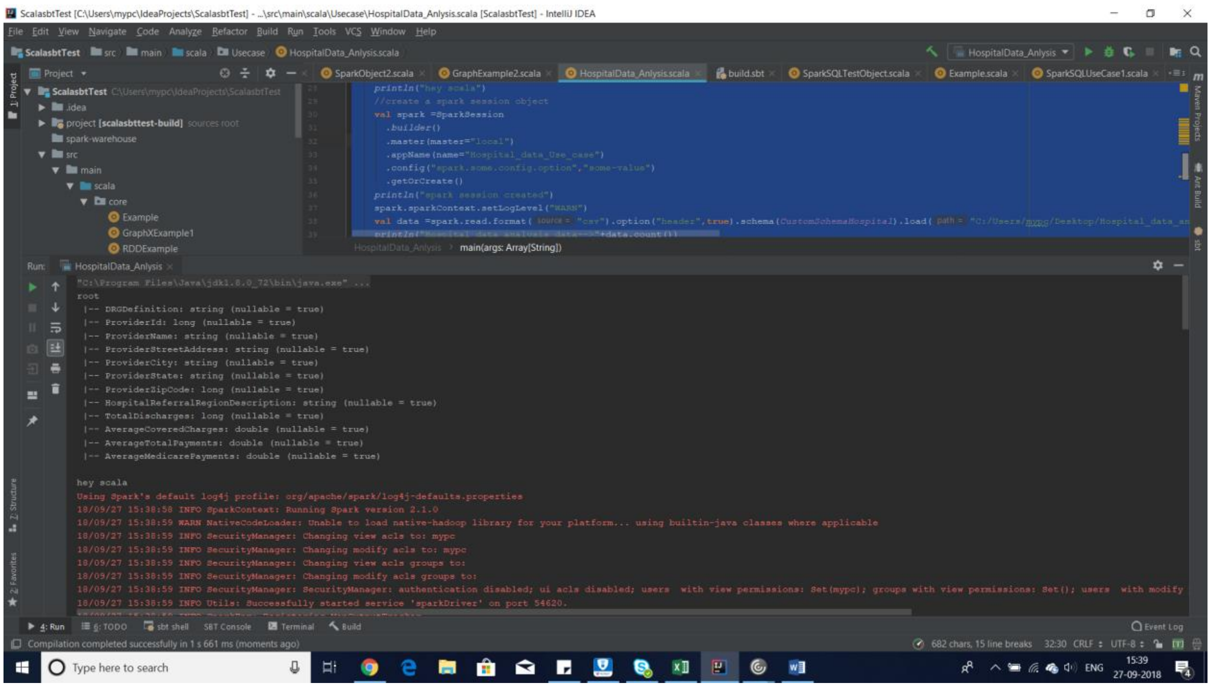
**//StructType objects define the schema of Spark DataFrames. StructType objects contain a list of StructField objects that define the name, type, and nullable flag for each column in a DataFrame.**

```
  val CustomSchemaHospital = new StructType(Array(
    StructField("DRGDefinition", StringType,true),
    StructField("ProviderId", LongType,true),
    StructField("ProviderName", StringType,true),
    StructField("ProviderStreetAddress", StringType,true),
    StructField("ProviderCity", StringType,true),
    StructField("ProviderState", StringType,true),
    StructField("ProviderZipCode", LongType,true),
    StructField("HospitalReferralRegionDescription", StringType,true),
    StructField("TotalDischarges", LongType,true),
    StructField("AverageCoveredCharges", DoubleType,true),
    StructField("AverageTotalPayments", DoubleType,true),
    StructField("AverageMedicarePayments", DoubleType,true))

CustomSchemaHospital.printTreeString()
```

```
 def main(args :Array[String]): Unit ={

  println("hey scala")

  //create a spark session object

  val spark =SparkSession

    .builder()

    .master(master="local")

    .appName(name="Hospital_data_Use_case")

    .config("spark.some.config.option","some-value")

    .getOrCreate()

  println("spark session created")

  spark.sparkContext.setLogLevel("WARN")

  val data
=spark.read.format("csv").option("header",true).schema(CustomSchemaHospital).load("C:/Users/mypc/Desktop/Hospital_data_analysis1.csv").toDF()

  println("Hospital_data_analysis data-->"+data.count())

  data.createOrReplaceTempView("hospital_data")
```

To see the contents inside the DataFrame, type the following:

```
data.show()
```



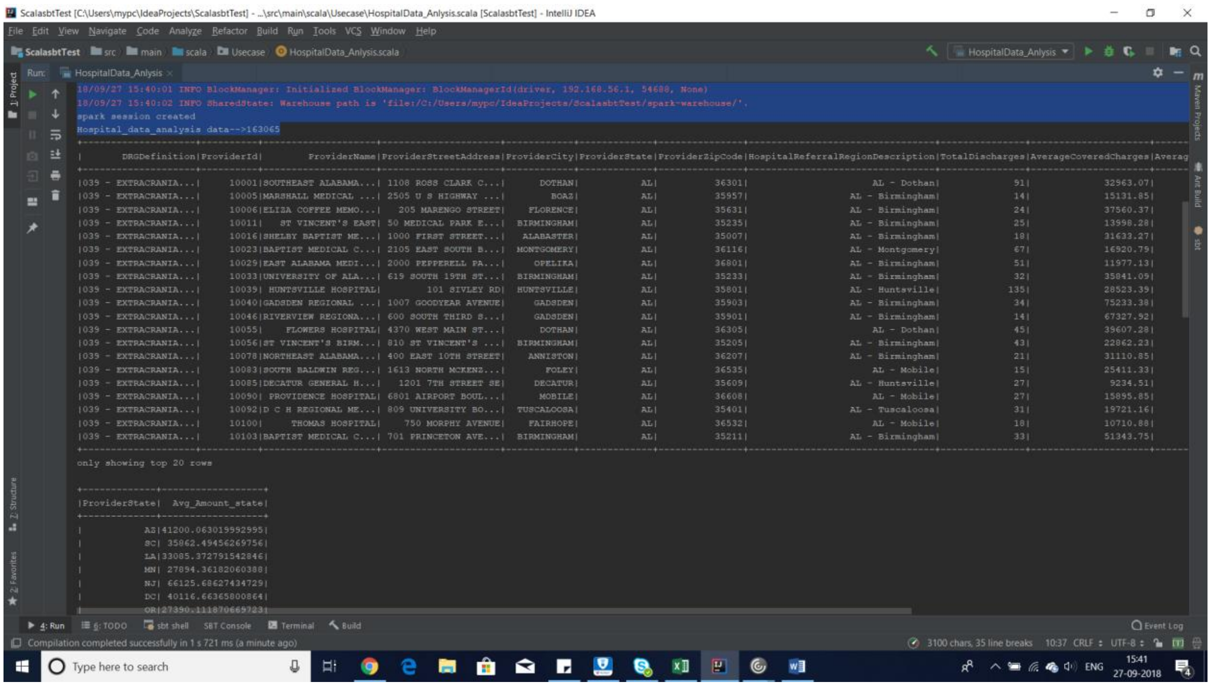
ScalasbtTest [C:\Users\mypc\IdeaProjects\ScalasbtTest] - ...\src\main\scala\Usecase\HospitalData_Anlysis.scala [ScalasbtTest] - IntelliJ IDEA

```
    println("hey scala")
    //create a spark session object
    val spark =SparkSession
      .builder()
      .master(master="local")
      .appName(name="Hospital_data_Use_case")
      .config("spark.some.config.option","some-value")
      .getOrCreate()
    println("spark session created")
    spark.sparkContext.setLogLevel("WARN")
    val data =spark.read.format( source = "csv").option("header",true).schema(CustomSchemaHospital).load( path = "C:/Users/mypc/Desktop/Hospital_data_an
    println("Hospital data analysis data-->"+data.count())
```

```
"C:\Program Files\Java\jdk1.8.0_72\bin\java.exe" ...
root
 |-- DRGDefinition: string (nullable = true)
 |-- ProviderId: long (nullable = true)
 |-- ProviderName: string (nullable = true)
 |-- ProviderStreetAddress: string (nullable = true)
 |-- ProviderCity: string (nullable = true)
 |-- ProviderState: string (nullable = true)
 |-- ProviderZipCode: long (nullable = true)
 |-- HospitalReferralRegionDescription: string (nullable = true)
 |-- TotalDischarges: long (nullable = true)
 |-- AverageCoveredCharges: double (nullable = true)
 |-- AverageTotalPayments: double (nullable = true)
 |-- AverageMedicarePayments: double (nullable = true)

hey scala
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
18/09/27 15:38:58 INFO SparkContext: Running Spark version 2.1.0
18/09/27 15:38:59 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
18/09/27 15:38:59 INFO SecurityManager: Changing view acls to: mypc
18/09/27 15:38:59 INFO SecurityManager: Changing modify acls to: mypc
18/09/27 15:38:59 INFO SecurityManager: Changing view acls groups to:
18/09/27 15:38:59 INFO SecurityManager: Changing modify acls groups to:
18/09/27 15:38:59 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(mypc); groups with view permissions: Set(); users  with modify
18/09/27 15:38:59 INFO Utils: Successfully started service 'sparkDriver' on port 54620.
```

Compilation completed successfully in 1 s 661 ms (moments ago)

682 chars, 15 line breaks 32:30 CRLF UTF-8



```
18/09/27 15:40:01 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver, 192.168.56.1, 54688, None)
18/09/27 15:40:02 INFO SharedState: Warehouse path is 'file:/C:/Users/mypc/IdeaProjects/ScalasbtTest/spark-warehouse/'.
spark session created
Hospital_data_analysis data-->163065
+--------------------+----------+--------------------+---------------------+------------+-------------+---------------+---------------------------------+---------------+---------------------+------
|       DRGDefinition|ProviderId|        ProviderName|ProviderStreetAddress|ProviderCity|ProviderState|ProviderZipCode|HospitalReferralRegionDescription|TotalDischarges|AverageCoveredCharges|Averag
+--------------------+----------+--------------------+---------------------+------------+-------------+---------------+---------------------------------+---------------+---------------------+------
|039 - EXTRACRANIA...|     10001|SOUTHEAST ALABAMA...| 1108 ROSS CLARK C...|      DOTHAN|           AL|          36301|                      AL - Dothan|             91|             32963.07|
|039 - EXTRACRANIA...|     10005|MARSHALL MEDICAL ...| 2505 U S HIGHWAY ...|        BOAZ|           AL|          35957|                  AL - Birmingham|             14|             15131.85|
|039 - EXTRACRANIA...|     10006|ELIZA COFFEE MEMO...|   205 MARENGO STREET|    FLORENCE|           AL|          35631|                  AL - Birmingham|             24|             37560.37|
|039 - EXTRACRANIA...|     10011|   ST VINCENT'S EAST| 50 MEDICAL PARK E...|  BIRMINGHAM|           AL|          35235|                  AL - Birmingham|             25|             13998.28|
|039 - EXTRACRANIA...|     10016|SHELBY BAPTIST ME...| 1000 FIRST STREET...|   ALABASTER|           AL|          35007|                  AL - Birmingham|             18|             31633.27|
|039 - EXTRACRANIA...|     10023|BAPTIST MEDICAL C...| 2105 EAST SOUTH B...|  MONTGOMERY|           AL|          36116|                  AL - Montgomery|             67|             16920.79|
|039 - EXTRACRANIA...|     10029|EAST ALABAMA MEDI...| 2000 PEPPERELL PA...|     OPELIKA|           AL|          36801|                  AL - Birmingham|             51|             11977.13|
|039 - EXTRACRANIA...|     10033|UNIVERSITY OF ALA...| 619 SOUTH 19TH ST...|  BIRMINGHAM|           AL|          35233|                  AL - Birmingham|             32|             35841.09|
|039 - EXTRACRANIA...|     10039| HUNTSVILLE HOSPITAL|       101 SIVLEY RD|  HUNTSVILLE|           AL|          35801|                  AL - Huntsville|            135|             28523.39|
|039 - EXTRACRANIA...|     10040|GADSDEN REGIONAL ...| 1007 GOODYEAR AVENUE|     GADSDEN|           AL|          35903|                  AL - Birmingham|             34|             75233.38|
|039 - EXTRACRANIA...|     10046|RIVERVIEW REGIONA...| 600 SOUTH THIRD S...|     GADSDEN|           AL|          35901|                  AL - Birmingham|             14|             67327.92|
|039 - EXTRACRANIA...|     10055|    FLOWERS HOSPITAL| 4370 WEST MAIN ST...|      DOTHAN|           AL|          36305|                      AL - Dothan|             45|             39607.28|
|039 - EXTRACRANIA...|     10056|ST VINCENT'S BIRM...| 810 ST VINCENT'S ...|  BIRMINGHAM|           AL|          35205|                  AL - Birmingham|             43|             22862.23|
|039 - EXTRACRANIA...|     10078|NORTHEAST ALABAMA...| 400 EAST 10TH STREET|    ANNISTON|           AL|          36207|                  AL - Birmingham|             21|             31110.85|
|039 - EXTRACRANIA...|     10083|SOUTH BALDWIN REG...| 1613 NORTH MCKENZ...|       FOLEY|           AL|          36535|                      AL - Mobile|             15|             25411.33|
|039 - EXTRACRANIA...|     10085|DECATUR GENERAL H...|   1201 7TH STREET SE|     DECATUR|           AL|          35609|                  AL - Huntsville|             27|              9234.51|
|039 - EXTRACRANIA...|     10090| PROVIDENCE HOSPITAL| 6801 AIRPORT BOUL...|      MOBILE|           AL|          36608|                      AL - Mobile|             27|             15895.85|
|039 - EXTRACRANIA...|     10092|D C H REGIONAL ME...| 809 UNIVERSITY BO...|  TUSCALOOSA|           AL|          35401|                  AL - Tuscaloosa|             31|             19721.16|
|039 - EXTRACRANIA...|     10100|     THOMAS HOSPITAL|   750 MORPHY AVENUE|    FAIRHOPE|           AL|          36532|                      AL - Mobile|             18|             10710.88|
|039 - EXTRACRANIA...|     10103|BAPTIST MEDICAL C...| 701 PRINCETON AVE...|  BIRMINGHAM|           AL|          35211|                  AL - Birmingham|             33|             51343.75|
+--------------------+----------+--------------------+---------------------+------------+-------------+---------------+---------------------------------+---------------+---------------------+------
only showing top 20 rows

+-------------+------------------+
|ProviderState|  Avg_Amount_state|
+-------------+------------------+
|           AZ|41200.063019992995|
|           SC| 35862.49456269756|
|           LA|33085.372791542846|
|           MN| 27894.36182060388|
|           NJ| 66125.68627434729|
|           DC| 40116.66365800864|
|           OR|27390.111870669723|
```

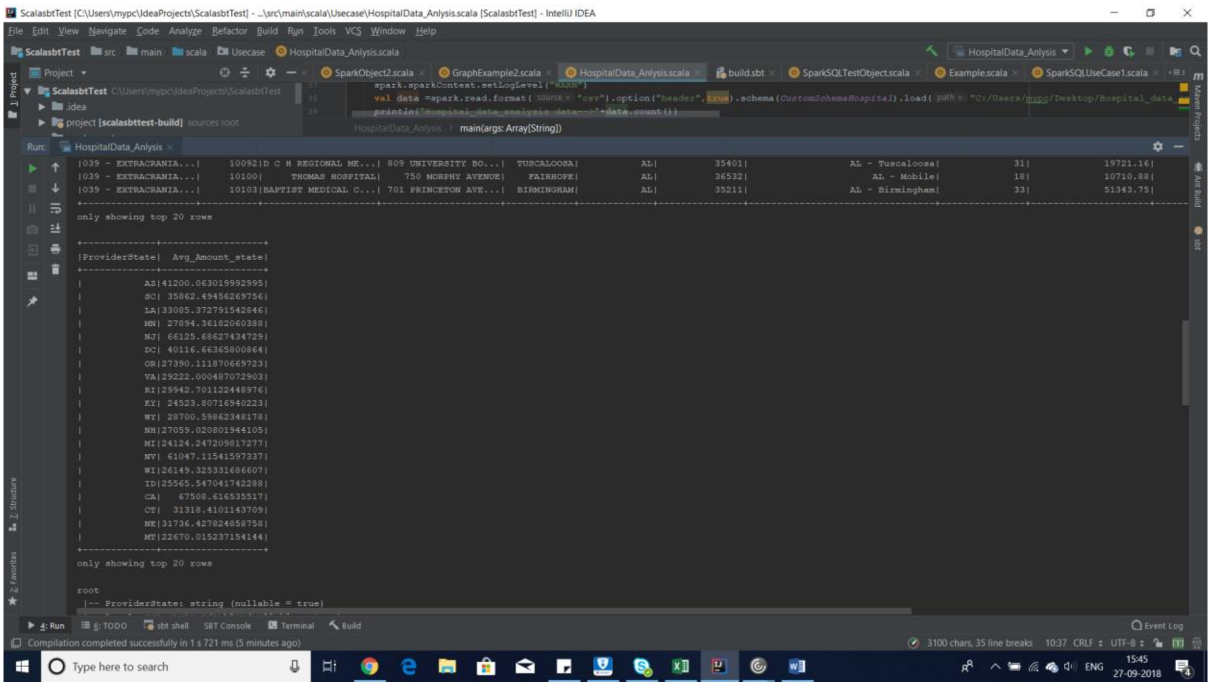
We will save the data in a table by registering it in a temp table as shown below.

data.createOrReplaceTempView("hospital_data")

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

**//Objective -1   What is the average amount of AverageCoveredCharges per state???**

**val a2= spark.sql("select ProviderState , avg(AverageCoveredCharges) as Avg_Amount_state from hospital_data group by ProviderState")**



```
spark.sparkContext.setLogLevel("WARN")
val data =spark.read.format( source = "csv").option("header",true).schema(CustomSchemaHospital).load( path = "C:/Users/mypc/Desktop/Hospital_data
println("Hospital_data_analysis data-->"+data.count())
```

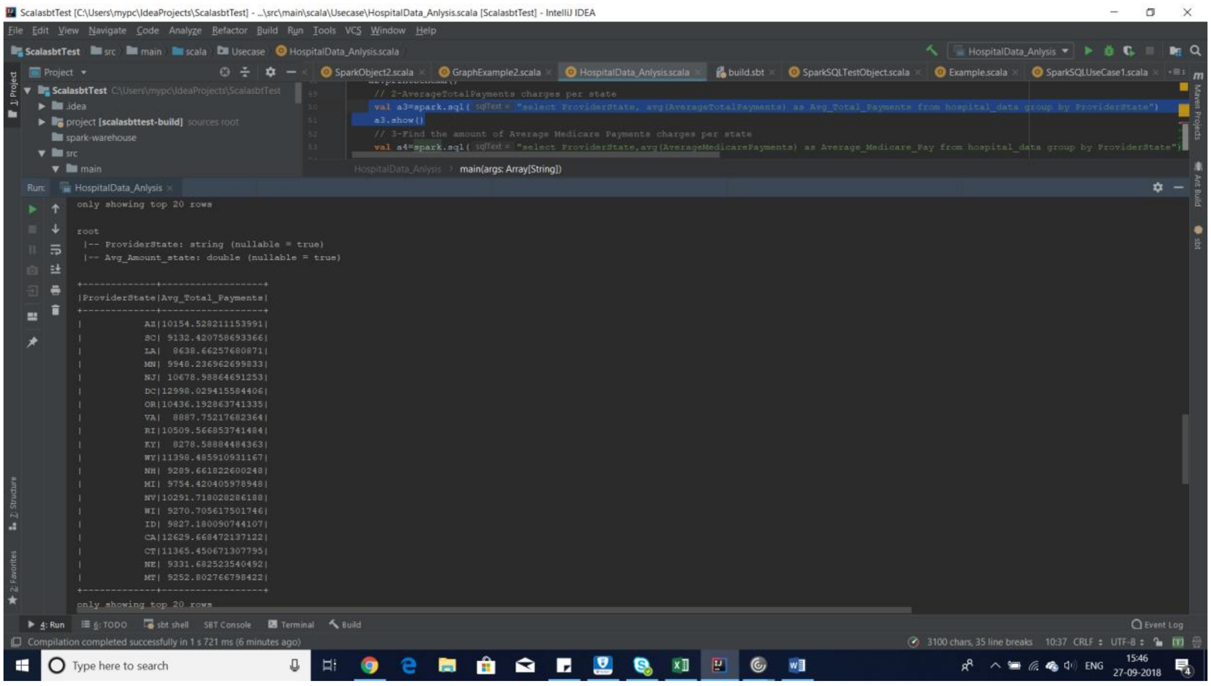
```
|039 - EXTRACRANIA...|     10092|D C H REGIONAL ME...| 809 UNIVERSITY BO...|  TUSCALOOSA|          AL|       35401|            AL - Tuscaloosa|        31|        19721.16|
|039 - EXTRACRANIA...|     10100|     THOMAS HOSPITAL|     750 MORPHY AVENUE|   FAIRHOPE|          AL|       36532|               AL - Mobile|        18|        10710.88|
|039 - EXTRACRANIA...|     10103|BAPTIST MEDICAL C...| 701 PRINCETON AVE...|  BIRMINGHAM|          AL|       35211|           AL - Birmingham|        33|        51343.75|
+--------------------+----------+--------------------+--------------------+------------+------------+------------+--------------------------+----------+----------------+
only showing top 20 rows

+-------------+------------------+
|ProviderState|  Avg_Amount_state|
+-------------+------------------+
|           AZ|41200.063019992995|
|           SC| 35862.49456269756|
|           LA|33085.372791542846|
|           MN| 27894.36182060388|
|           NJ| 66125.68627434729|
|           DC| 40116.66365800864|
|           OR|27390.111870669723|
|           VA|29222.000487072903|
|           RI|29942.701122448976|
|           KY| 24523.80716940223|
|           WY| 28700.59862348178|
|           NH|27059.020801944105|
|           MI|24124.247209817277|
|           NV| 61047.11541597337|
|           WI|26149.325331686607|
|           ID|25565.547041742288|
|           CA|   67508.616535517|
|           CT|  31318.4101143709|
|           NE|31736.427824858758|
|           MT|22670.015237154144|
+-------------+------------------+
only showing top 20 rows

root
 |-- ProviderState: string (nullable = true)
```

Compilation completed successfully in 1 s 721 ms (5 minutes ago)

## //find out the AverageTotalPayments charges per state

## val a3=spark.sql("select ProviderState, avg(AverageTotalPayments) as Avg_Total_Payments from hospital_data group by ProviderState")

## a3.show()



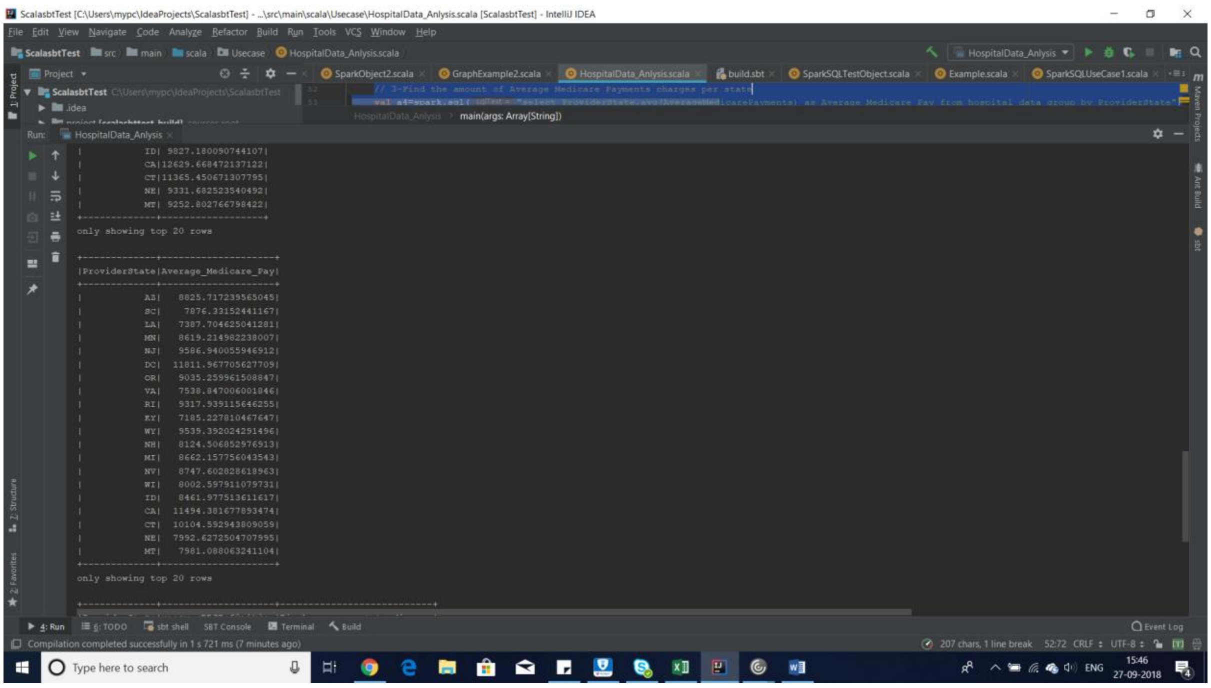
```
// 2-AverageTotalPayments charges per state
val a3=spark.sql( sqlText = "select ProviderState, avg(AverageTotalPayments) as Avg_Total_Payments from hospital_data group by ProviderState")
a3.show()
// 3-Find the amount of Average Medicare Payments charges per state
val a4=spark.sql( sqlText = "select ProviderState,avg(AverageMedicarePayments) as Average_Medicare_Pay from hospital_data group by ProviderState")
```

```
only showing top 20 rows

root
 |-- ProviderState: string (nullable = true)
 |-- Avg_Amount_state: double (nullable = true)

+-------------+------------------+
|ProviderState|Avg_Total_Payments|
+-------------+------------------+
|           AZ|10154.528211153991|
|           SC| 9132.420758693366|
|           LA|  8638.66257680871|
|           MN| 9948.236962699833|
|           NJ| 10678.98864691253|
|           DC|12998.029415584406|
|           OR|10436.192863741335|
|           VA|  8887.75217682364|
|           RI|10509.566853741484|
|           KY|  8278.58884484363|
|           WY|11398.485910931167|
|           NH| 9289.661822600248|
|           MI| 9754.420405978948|
|           NV|10291.718028286188|
|           WI| 9270.705617501746|
|           ID| 9827.180090744107|
|           CA|12629.668472137122|
|           CT|11365.450671307795|
|           NE| 9331.682523540492|
|           MT| 9252.802766798422|
+-------------+------------------+
only showing top 20 rows
```

Compilation completed successfully in 1 s 721 ms (6 minutes ago)

## // 3-Find the amount of Average Medicare Payments charges per state

val a4=spark.sql("select ProviderState,avg(AverageMedicarePayments) as Average_Medicare_Pay from hospital_data group by ProviderState")



```
// 3-Find the amount of Average Medicare Payments charges per state
val a4=spark.sql( sqlText = "select ProviderState,avg(AverageMedicarePayments) as Average_Medicare_Pay from hospital_data group by ProviderState"
```

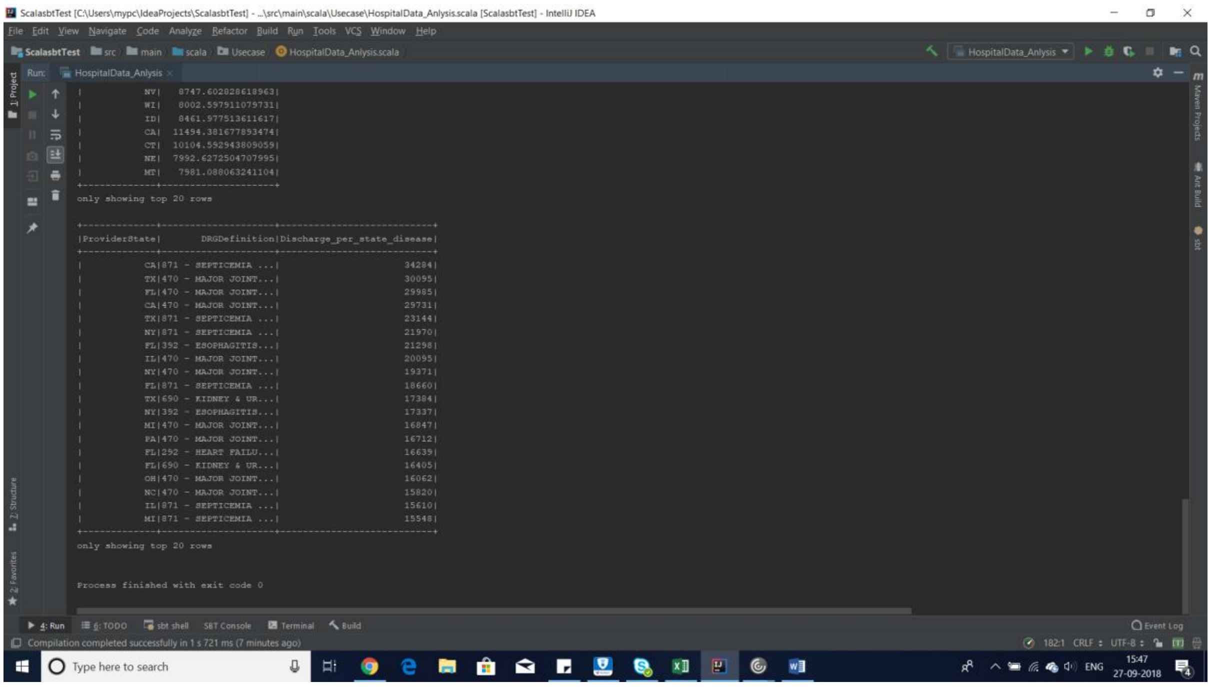
```
|           ID| 9827.180090744107|
|           CA|12629.668472137122|
|           CT|11365.450671307795|
|           NE| 9331.682523540492|
|           MT| 9252.802766798422|
+-------------+------------------+
only showing top 20 rows

+-------------+--------------------+
|ProviderState|Average_Medicare_Pay|
+-------------+--------------------+
|           AZ|   8825.717239565045|
|           SC|    7876.33152441167|
|           LA|   7387.704625041281|
|           MN|   8619.214982238007|
|           NJ|   9586.940055946912|
|           DC|  11811.967705627709|
|           OR|   9035.259961508847|
|           VA|   7538.847006001846|
|           RI|   9317.939115646255|
|           KY|   7185.227810467647|
|           WY|   9535.392024291496|
|           NH|   8124.506852976913|
|           MI|   8662.157756043543|
|           NV|   8747.602828618963|
|           WI|   8002.597911079731|
|           ID|   8461.977513611617|
|           CA|  11494.381677893474|
|           CT|  10104.592943809059|
|           NE|  7992.6272504707995|
|           MT|   7981.088063241104|
+-------------+--------------------+
only showing top 20 rows
```

Compilation completed successfully in 1 s 721 ms (7 minutes ago)

**//4-Find out the total number of Discharges per state and for each disease.**

val a5= spark.sql("select ProviderState,DRGDefinition, sum(TotalDischarges)as Discharge_per_state_disease  from hospital_data group by ProviderState,DRGDefinition order by Discharge_per_state_disease desc ")

a5.show()


ScalasbtTest [C:\Users\mypc\IdeaProjects\ScalasbtTest] - ...\src\main\scala\Usecase\HospitalData_Anlysis.scala [ScalasbtTest] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
ScalasbtTest src main scala Usecase HospitalData_Anlysis.scala
Run: HospitalData_Anlysis
4: Run 6: TODO sbt shell SBT Console Terminal Build
Event Log
Compilation completed successfully in 1 s 721 ms (7 minutes ago)
182:1 CRLF UTF-8
Type here to search
ENG
15:47
27-09-2018

```
|          NV|   8747.602828618963|
|          WI|   8002.597911079731|
|          ID|   8461.977513611617|
|          CA|  11494.381677893474|
|          CT|  10104.592943809059|
|          NE|  7992.6272504707995|
|          MT|   7981.088063241104|
+------------+--------------------+
only showing top 20 rows

+------------+--------------------+--------------------------+
|ProviderState|        DRGDefinition|Discharge_per_state_disease|
+------------+--------------------+--------------------------+
|          CA|871 - SEPTICEMIA ...|                     34284|
|          TX|470 - MAJOR JOINT...|                     30095|
|          FL|470 - MAJOR JOINT...|                     29985|
|          CA|470 - MAJOR JOINT...|                     29731|
|          TX|871 - SEPTICEMIA ...|                     23144|
|          NY|871 - SEPTICEMIA ...|                     21970|
|          FL|392 - ESOPHAGITIS...|                     21298|
|          IL|470 - MAJOR JOINT...|                     20095|
|          NY|470 - MAJOR JOINT...|                     19371|
|          FL|871 - SEPTICEMIA ...|                     18660|
|          TX|690 - KIDNEY & UR...|                     17384|
|          NY|392 - ESOPHAGITIS...|                     17337|
|          MI|470 - MAJOR JOINT...|                     16847|
|          PA|470 - MAJOR JOINT...|                     16712|
|          FL|292 - HEART FAILU...|                     16639|
|          FL|690 - KIDNEY & UR...|                     16405|
|          OH|470 - MAJOR JOINT...|                     16062|
|          NC|470 - MAJOR JOINT...|                     15820|
|          IL|871 - SEPTICEMIA ...|                     15610|
|          MI|871 - SEPTICEMIA ...|                     15548|
+------------+--------------------+--------------------------+
only showing top 20 rows


Process finished with exit code 0
```