# CASE STUDY - 4: HOSPITAL DATA ANALYSIS

## 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 providerâ€™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 providerâ€™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 providerâ€™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 Medicareâ€™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 co-payments and deductible amounts nor any additional payments from third parties for coordination of benefits.

**TASKS:**

**1) Load file into spark.**

**2) What is the average amount of AverageCoveredCharges per state.**

**3) Find out the AverageTotalPayments charges per state.**

**4) Find out the AverageMedicarePayments charges per state.**

**5) Find out the total number of Discharges per state and for each disease.**

**6) Sort the output in descending order of totalDischarges.**

**EXPLANATION:** FOR BUILDING A SPARK SQL APPLICATION WE NEED TO SET THE MASTER AND CREATE A SPARK SESSION. USING BELOW COMMAND WE CAN CREATE A SPARK SESSION FOR SPARK SQL

**CODE: val session = org.apache.spark.sql.SparkSession.builder.master("local").appName("Spark CSV Reader").getOrCreate;**

**EXPLANATION TASK 1:** AFTER THIS WE LOAD THE CSV FILE INTO SPARK. THE DATASET HERE IS DOWNLOADED AND SAVED IN LOCAL DIRECTORY **PATH:///home/acadgild/inpatientCharges.csv.** NOW USING THE BELOW CODE WE LOAD THE CSV FILE INTO SPARK.

**CODE: val df = session.read.format("com.databricks.spark.csv").option("header", "true").option("inferSchema", "true").load("file:///home/acadgild/inpatientCharges.csv")**

HERE WE USED "**inferSchema**" AS AN OPTION SO IT WILL AUTOMATICALLY INFER THE DATA TYPES OF THE COLUMNS. WE CAN CHECK THE DATA TYPES FOR THE SCHEMA USING BELOW CODE:

**CODE: df.printSchema()**

**EXPLANATION:** USING THE BELOW CODE WE CAN CHECK THE LOADED DATA TYPES INTO THE VARIABLE "df".

**CODE: df.show()**

**EXPLANATION:** NOW TO DO THE NORMAL QUERY OPERATION WE NEED TO CONVERT THIS INTO A TEMPORARY TABLE. SO WE USE BELOW CODE.

**CODE: df.registerTempTable("hospital_charges")**

SO HERE CREATED A TEMPORARY TABLE OR VIEW NAMED **"hospitall_charges".**

**EXPLANATION TASK 2:** HERE WE CALCULATE THE AVERAGE OF "AverageCoveredCharges" AND GROUP IT IN THE ORDER OF "ProviderState". USING THE BELOW CODE WE GET THE RESULT.

**CODE: df.groupBy("ProviderState").avg("AverageCoveredCharges").show**

**EXPLANATION TASK 3:** HERE WE CALCULATE THE AVERAGE OF "AverageTotalPayments"

AND GROUP IT IN THE ORDER OF "ProviderState". USING THE BELOW CODE WE GET THE RESULT.

**CODE:**
```
df.groupBy("ProviderState").avg("AverageTotalPayments").show
```

**EXPLANATION TASK 4:** HERE WE CALCULATE THE AVERAGE OF "AverageMedicalPayments" AND GROUP IT IN THE ORDER OF "ProviderState". USING THE BELOW CODE WE GET THE RESULT.

**CODE:**
```
df.groupBy("ProviderState").avg("AverageMedicarePayments").show
```

**EXPLANATION TASK 5:** HERE WE CALCULATE THE SUM OF "TotalDischarges" AND GROUP IT IN THE ORDER OF "ProviderState" AND "DRGDefinition". USING THE BELOW CODE WE GET THE RESULT.

**CODE:**
```
df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").
show
```

**EXPLANATION TASK 6:** HERE WE CALCULATE THE SUM OF "TotalDischarges" THEN WE SORT SUM OF "TotalDischarges" IN THE DESCENDING ORDER. THIS CAN BE ACHIEVED IN TWO POSSIBLE WAY. BY USING "sort" AND "order by". THEN WE GROUP IT IN THE ORDER OF "ProviderState" AND "DRGDefinition". USING THE BELOW CODE WE GET THE RESULT.

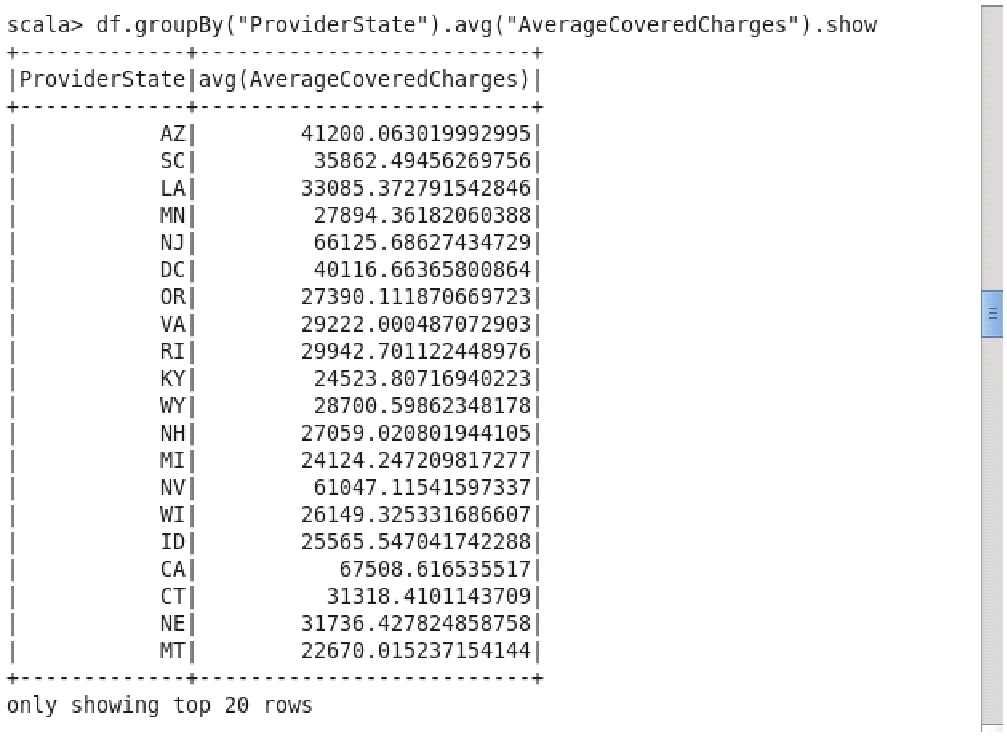
**CODE:**
```
df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").
sort(desc(sum                    ("TotalDischarges").toString)).show
```

**CODE:**
```
df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").
orderBy(desc(sum("TotalDischarges").toString)).show
```

**SOLUTION REPORT:**
```
scala> val session =
org.apache.spark.sql.SparkSession.builder.master("local").appName("Spar
k CSV Reader").getOrCreate;
Mon Mar 11 15:27:36 IST 2019 WARN: Establishing SSL connection without
server's identity verification is not recommended. According to MySQL
5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established
by default if explicit option isn't set. For compliance with existing
applications not using SSL the verifyServerCertificate property is set to
'false'. You need either to explicitly disable SSL by setting useSSL=false,
or set useSSL=true and provide truststore for server certificate
verification.
Mon Mar 11 15:27:38 IST 2019 WARN: Establishing SSL connection without
server's identity verification is not recommended. According to MySQL
5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established
by default if explicit option isn't set. For compliance with existing
applications not using SSL the verifyServerCertificate property is set to
'false'. You need either to explicitly disable SSL by setting useSSL=false,
or set useSSL=true and provide truststore for server certificate
verification.
```

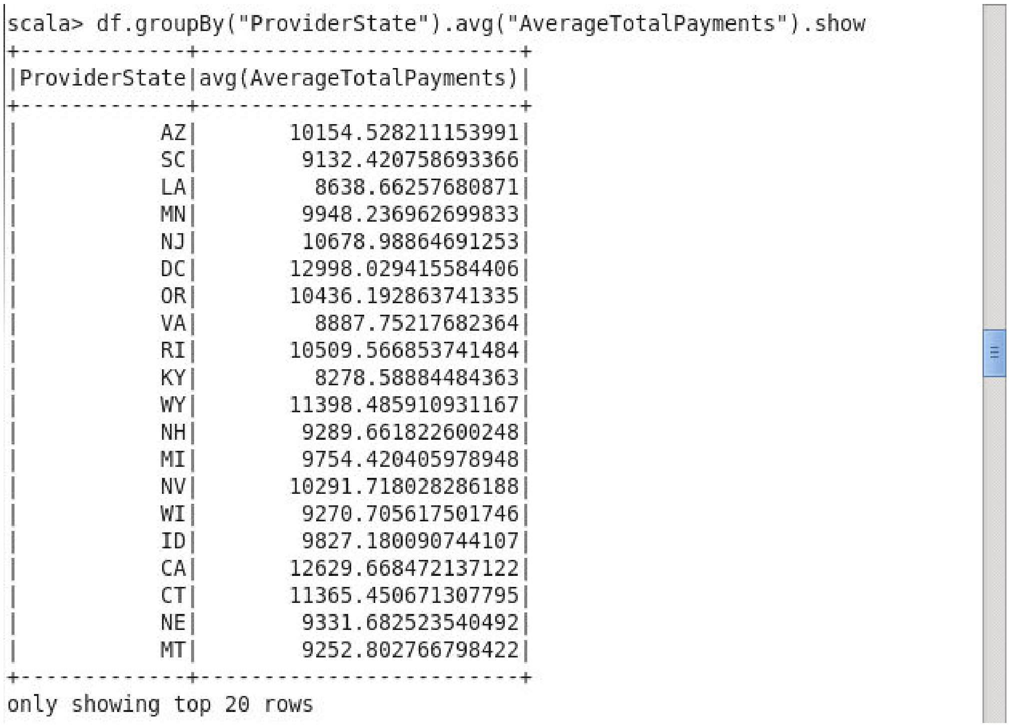
```
Mon Mar 11 15:27:39 IST 2019 WARN: Establishing SSL connection without
server's identity verification is not recommended. According to MySQL
5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established
by default if explicit option isn't set. For compliance with existing
applications not using SSL the verifyServerCertificate property is set to
'false'. You need either to explicitly disable SSL by setting useSSL=false,
or set useSSL=true and provide truststore for server certificate
verification.
Mon Mar 11 15:27:39 IST 2019 WARN: Establishing SSL connection without
server's identity verification is not recommended. According to MySQL
5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established
by default if explicit option isn't set. For compliance with existing
applications not using SSL the verifyServerCertificate property is set to
'false'. You need either to explicitly disable SSL by setting useSSL=false,
or set useSSL=true and provide truststore for server certificate
verification.
Mon Mar 11 15:27:40 IST 2019 WARN: Establishing SSL connection without
server's identity verification is not recommended. According to MySQL
5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established
by default if explicit option isn't set. For compliance with existing
applications not using SSL the verifyServerCertificate property is set to
'false'. You need either to explicitly disable SSL by setting useSSL=false,
or set useSSL=true and provide truststore for server certificate
verification.
Mon Mar 11 15:27:40 IST 2019 WARN: Establishing SSL connection without
server's identity verification is not recommended. According to MySQL
5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established
by default if explicit option isn't set. For compliance with existing
applications not using SSL the verifyServerCertificate property is set to
'false'. You need either to explicitly disable SSL by setting useSSL=false,
or set useSSL=true and provide truststore for server certificate
verification.
Mon Mar 11 15:27:40 IST 2019 WARN: Establishing SSL connection without
server's identity verification is not recommended. According to MySQL
5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established
by default if explicit option isn't set. For compliance with existing
applications not using SSL the verifyServerCertificate property is set to
'false'. You need either to explicitly disable SSL by setting useSSL=false,
or set useSSL=true and provide truststore for server certificate
verification.
Mon Mar 11 15:27:40 IST 2019 WARN: Establishing SSL connection without
server's identity verification is not recommended. According to MySQL
5.5.45+, 5.6.26+ and 5.7.6+ requirements SSL connection must be established
by default if explicit option isn't set. For compliance with existing
applications not using SSL the verifyServerCertificate property is set to
'false'. You need either to explicitly disable SSL by setting useSSL=false,
or set useSSL=true and provide truststore for server certificate
verification.
19/03/11 15:27:51 WARN metastore.ObjectStore: Failed to get database
global_temp, returning NoSuchObjectException
19/03/11 15:27:52 WARN sql.SparkSession$Builder: Using an existing
SparkSession; some configuration may not take effect.
session: org.apache.spark.sql.SparkSession =
org.apache.spark.sql.SparkSession@44a7661d

scala> val df =
session.read.format("com.databricks.spark.csv").option("header",
"true").option("inferSchema",
"true").load("file:///home/acadgild/inpatientCharges.csv")
df: org.apache.spark.sql.DataFrame = [DRGDefinition: string, ProviderId:
int ... 10 more fields]

scala> df.printSchema()
root
 |-- DRGDefinition: string (nullable = true)
 |-- ProviderId: integer (nullable = true)
 |-- ProviderName: string (nullable = true)
 |-- ProviderStreetAddress: string (nullable = true)
 |-- ProviderCity: string (nullable = true)
 |-- ProviderState: string (nullable = true)
 |-- ProviderZipCode: integer (nullable = true)
 |-- HospitalReferralRegionDescription: string (nullable = true)
 |-- TotalDischarges: integer (nullable = true)
 |-- AverageCoveredCharges: double (nullable = true)
 |-- AverageTotalPayments: double (nullable = true)
 |-- AverageMedicarePayments: double (nullable = true)

scala> df.show()
+--------------------+----------+--------------------+-----------------
----+------------+-------------+---------------+-----------------------
----------+---------------+---------------------+--------------------+-
----------------------+
|       DRGDefinition|ProviderId|
ProviderName|ProviderStreetAddress|ProviderCity|ProviderState|ProviderZ
ipCode|HospitalReferralRegionDescription|TotalDischarges|AverageCovered
Charges|AverageTotalPayments|AverageMedicarePayments|
+--------------------+----------+--------------------+-----------------
----+------------+-------------+---------------+-----------------------
----------+---------------+---------------------+--------------------+-
----------------------+
|039 - EXTRACRANIA...|     10001|SOUTHEAST ALABAMA...| 1108 ROSS CLARK
C...|      DOTHAN|           AL|          36301|
AL - Dothan|             91|             32963.07|            5777.24|
4763.73|
|039 - EXTRACRANIA...|     10005|MARSHALL MEDICAL ...| 2505 U S HIGHWAY
...|        BOAZ|           AL|          35957|                  AL -
Birmingham|             14|             15131.85|            5787.57|
4976.71|
|039 - EXTRACRANIA...|     10006|ELIZA COFFEE MEMO...|   205 MARENGO
STREET|    FLORENCE|           AL|          35631|                  AL
- Birmingham|             24|             37560.37|
5434.95|                4453.79|
|039 - EXTRACRANIA...|     10011|   ST VINCENT'S EAST| 50 MEDICAL PARK
E...|  BIRMINGHAM|           AL|          35235|                  AL -
Birmingham|             25|             13998.28|            5417.56|
4129.16|
|039 - EXTRACRANIA...|     10016|SHELBY BAPTIST ME...| 1000 FIRST
STREET...|   ALABASTER|           AL|          35007|
AL - Birmingham|            18|             31633.27|
5658.33|                 4851.44|
|039 - EXTRACRANIA...|     10023|BAPTIST MEDICAL C...| 2105 EAST SOUTH
B...|  MONTGOMERY|           AL|          36116|                 AL -
Montgomery|              67|             16920.79|              6653.8|
5374.14|
|039 - EXTRACRANIA...|     10029|EAST ALABAMA MEDI...| 2000 PEPPERELL
PA...|     OPELIKA|           AL|          36801|                 AL -
Birmingham|              51|             11977.13|             5834.74|
4761.41|
|039 - EXTRACRANIA...|     10033|UNIVERSITY OF ALA...| 619 SOUTH 19TH
ST...|  BIRMINGHAM|           AL|          35233|                 AL -
Birmingham|              32|             35841.09|             8031.12|
5858.5|
|039 - EXTRACRANIA...|     10039| HUNTSVILLE HOSPITAL|        101 SIVLEY
RD|  HUNTSVILLE|           AL|          35801|                 AL -
Huntsville|             135|             28523.39|             6113.38|
5228.4|
|039 - EXTRACRANIA...|     10040|GADSDEN REGIONAL ...| 1007 GOODYEAR
AVENUE|     GADSDEN|           AL|          35903|                 AL
- Birmingham|              34|             75233.38|
5541.05|                 4386.94|
|039 - EXTRACRANIA...|     10046|RIVERVIEW REGIONA...| 600 SOUTH THIRD
S...|     GADSDEN|           AL|          35901|                 AL -
Birmingham|              14|             67327.92|             5461.57|
4493.57|
|039 - EXTRACRANIA...|     10055|    FLOWERS HOSPITAL| 4370 WEST MAIN
ST...|      DOTHAN|           AL|          36305|
AL - Dothan|              45|             39607.28|             5356.28|
4408.2|
|039 - EXTRACRANIA...|     10056|ST VINCENT'S BIRM...| 810 ST VINCENT'S
...|  BIRMINGHAM|           AL|          35205|                 AL -
Birmingham|              43|             22862.23|             5374.65|
4186.02|
|039 - EXTRACRANIA...|     10078|NORTHEAST ALABAMA...| 400 EAST 10TH
STREET|    ANNISTON|           AL|          36207|                 AL
- Birmingham|              21|             31110.85|
5366.23|                 4376.23|
|039 - EXTRACRANIA...|     10083|SOUTH BALDWIN REG...| 1613 NORTH
MCKENZ...|       FOLEY|           AL|          36535|
AL - Mobile|              15|             25411.33|             5282.93|
4383.73|
|039 - EXTRACRANIA...|     10085|DECATUR GENERAL H...|   1201 7TH STREET
SE|     DECATUR|           AL|          35609|                 AL -
Huntsville|              27|              9234.51|             5676.55|
4509.11|
|039 - EXTRACRANIA...|     10090| PROVIDENCE HOSPITAL| 6801 AIRPORT
BOUL...|      MOBILE|           AL|          36608|
AL - Mobile|              27|             15895.85|             5930.11|
3972.85|
|039 - EXTRACRANIA...|     10092|D C H REGIONAL ME...| 809 UNIVERSITY
```

```
BO...|  TUSCALOOSA|          AL|          35401|                   AL -
Tuscaloosa|            31|             19721.16|            6192.54|
5179.38|
|039 - EXTRACRANIA...|     10100|     THOMAS HOSPITAL|    750 MORPHY
AVENUE|    FAIRHOPE|          AL|          36532|
AL - Mobile|            18|             10710.88|             4968.0|
3898.88|
|039 - EXTRACRANIA...|     10103|BAPTIST MEDICAL C...| 701 PRINCETON
AVE...|  BIRMINGHAM|          AL|          35211|                   AL
- Birmingham|            33|             51343.75|
5996.0|              4962.45|
+--------------------+----------+--------------------+-----------------
----+------------+-------------+---------------+----------------------
----------+---------------+---------------------+--------------------+-
----------------------+
only showing top 20 rows

scala> df.registerTempTable("hospital_charges")
warning: there was one deprecation warning; re-run with -deprecation for
details

scala> df.groupBy("ProviderState").avg("AverageCoveredCharges").show
+-------------+--------------------------+
|ProviderState|avg(AverageCoveredCharges)|
+-------------+--------------------------+
|           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
```



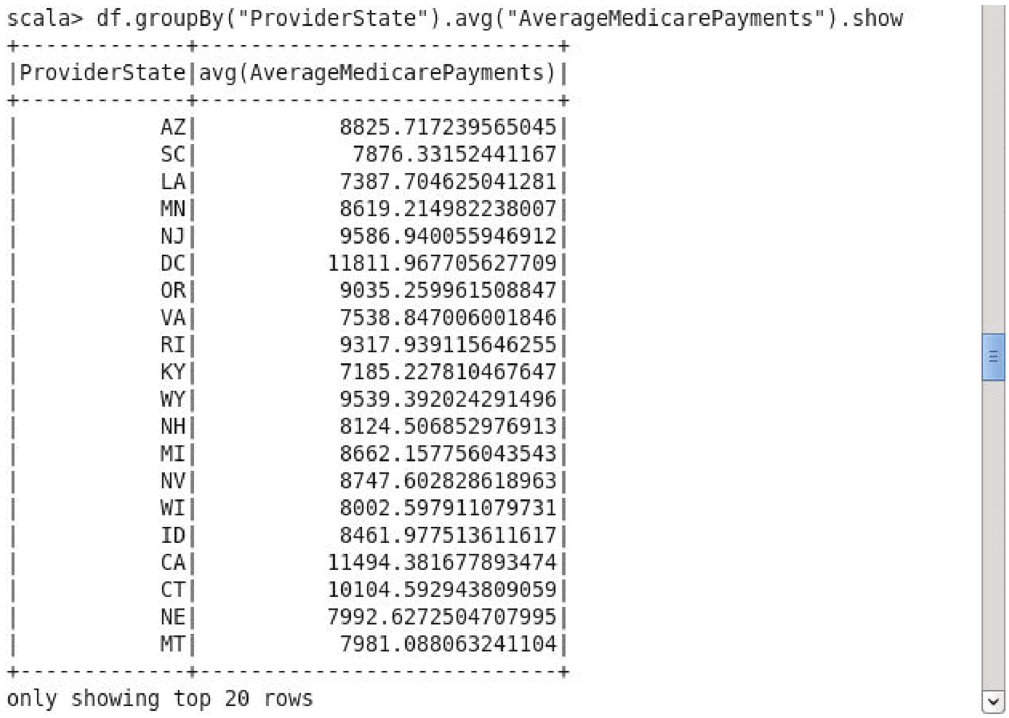
```
scala> df.groupBy("ProviderState").avg("AverageCoveredCharges").show
+-------------+--------------------------+
|ProviderState|avg(AverageCoveredCharges)|
+-------------+--------------------------+
|           AZ|        41200.063019992995|
|           SC|         35862.49456269756|
|           LA|        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
```

```
scala> df.groupBy("ProviderState").avg("AverageTotalPayments").show
+-------------+-------------------------+
|ProviderState|avg(AverageTotalPayments)|
+-------------+-------------------------+
|           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
```



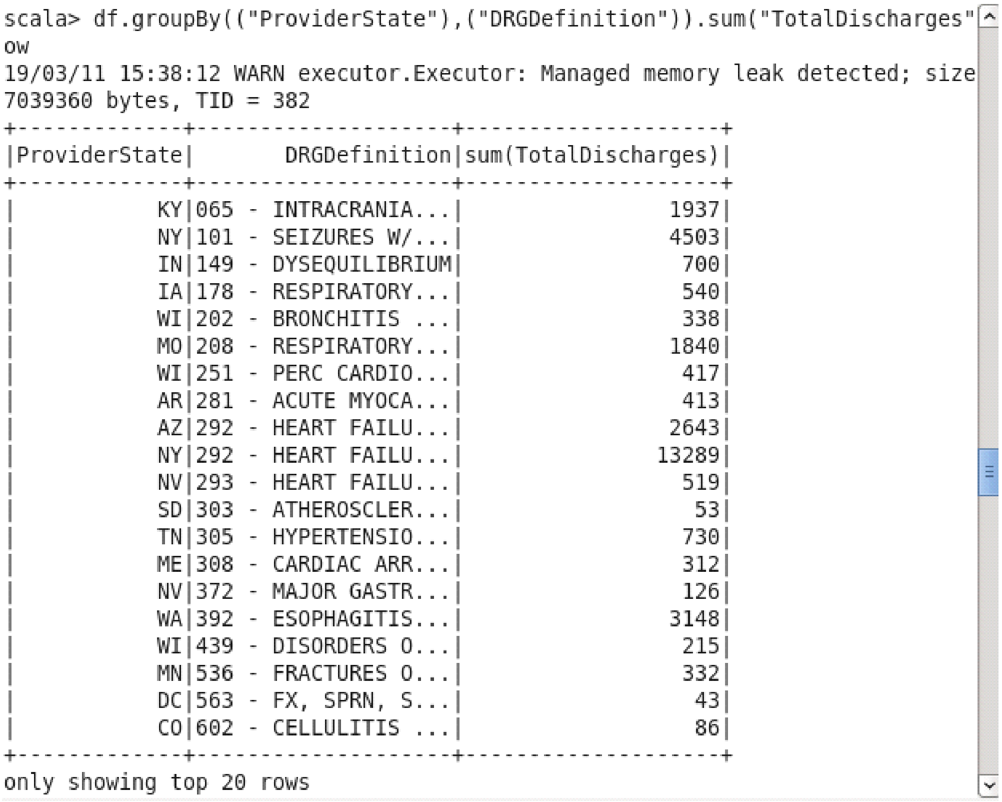
```
scala> df.groupBy("ProviderState").avg("AverageTotalPayments").show
+-------------+-------------------------+
|ProviderState|avg(AverageTotalPayments)|
+-------------+-------------------------+
|           AZ|       10154.528211153991|
|           SC|        9132.420758693366|
|           LA|         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
```

```
scala> df.groupBy("ProviderState").avg("AverageMedicarePayments").show
+-------------+----------------------------+
|ProviderState|avg(AverageMedicarePayments)|
+-------------+----------------------------+
|           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|           9539.392024291496|
|           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
```



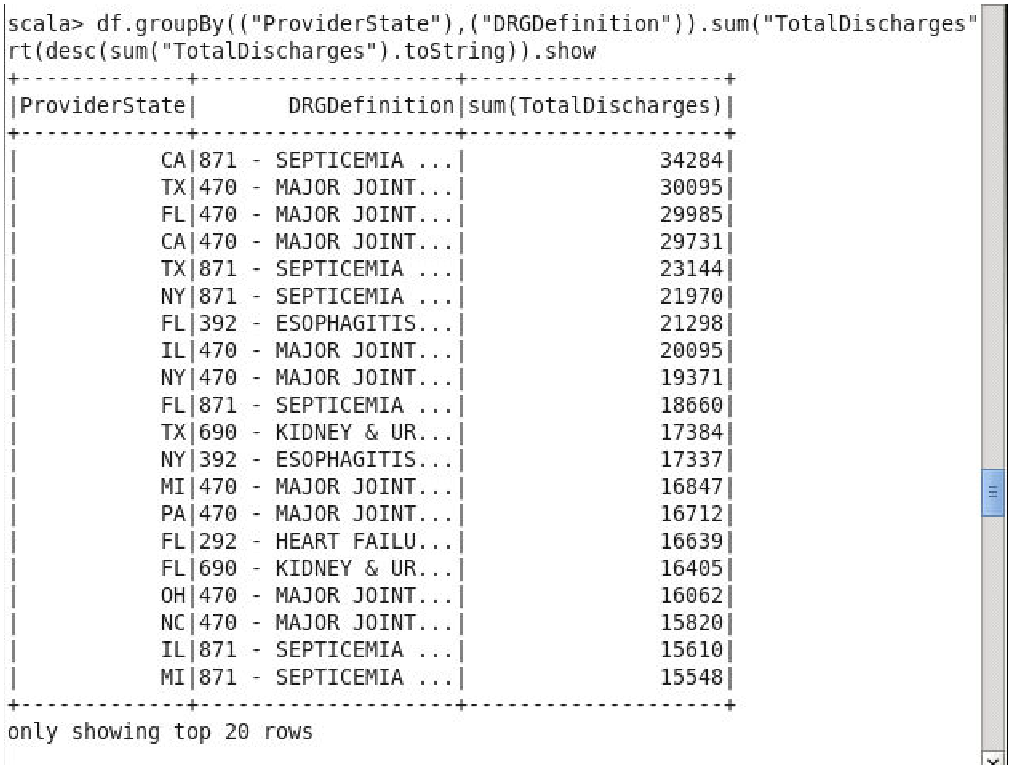
```
scala> df.groupBy("ProviderState").avg("AverageMedicarePayments").show
+-------------+----------------------------+
|ProviderState|avg(AverageMedicarePayments)|
+-------------+----------------------------+
|           AZ|            8825.71723956504|
|           SC|            7876.33152441167|
|           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|           9539.392024291496|
|           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
```

```
scala>
df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").
show
19/03/11 15:38:12 WARN executor.Executor: Managed memory leak detected;
size = 17039360 bytes, TID = 382
+-------------+--------------------+--------------------+
|ProviderState|       DRGDefinition|sum(TotalDischarges)|
+-------------+--------------------+--------------------+
|           KY|065 - INTRACRANIA...|                1937|
|           NY|101 - SEIZURES W/...|                4503|
|           IN|149 - DYSEQUILIBRIUM|                 700|
|           IA|178 - RESPIRATORY...|                 540|
|           WI|202 - BRONCHITIS ...|                 338|
|           MO|208 - RESPIRATORY...|                1840|
|           WI|251 - PERC CARDIO...|                 417|
|           AR|281 - ACUTE MYOCA...|                 413|
|           AZ|292 - HEART FAILU...|                2643|
|           NY|292 - HEART FAILU...|               13289|
|           NV|293 - HEART FAILU...|                 519|
|           SD|303 - ATHEROSCLER...|                  53|
|           TN|305 - HYPERTENSIO...|                 730|
|           ME|308 - CARDIAC ARR...|                 312|
|           NV|372 - MAJOR GASTR...|                 126|
|           WA|392 - ESOPHAGITIS...|                3148|
|           WI|439 - DISORDERS O...|                 215|
|           MN|536 - FRACTURES O...|                 332|
|           DC|563 - FX, SPRN, S...|                  43|
|           CO|602 - CELLULITIS ...|                  86|
+-------------+--------------------+--------------------+
only showing top 20 rows
```



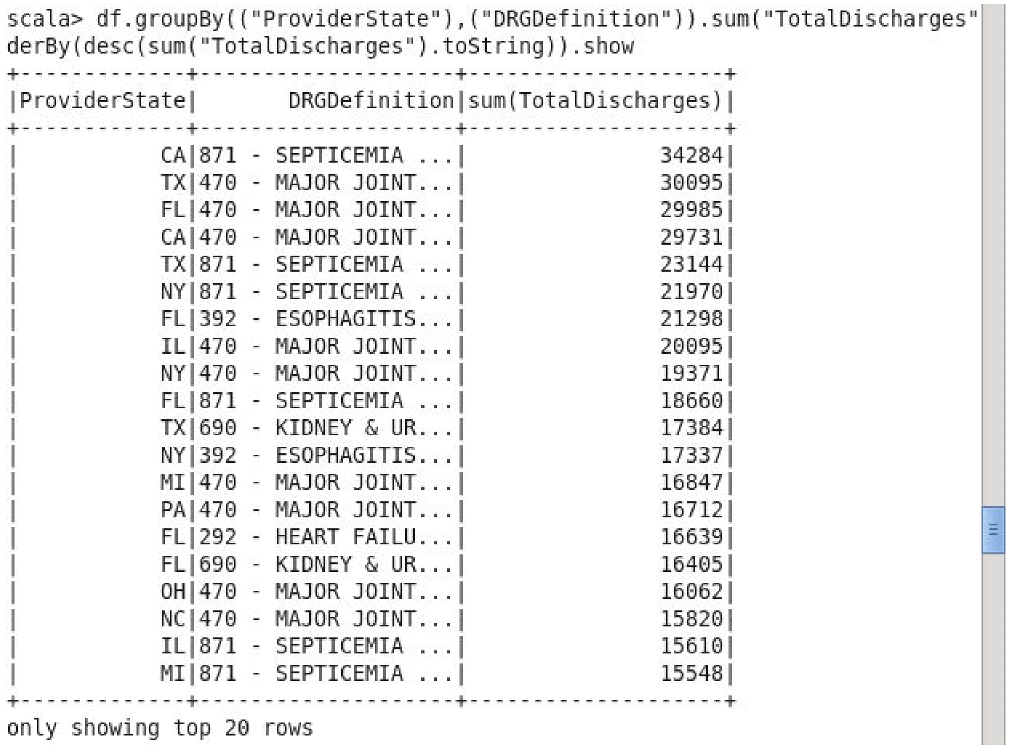
```
scala> df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges"
ow
19/03/11 15:38:12 WARN executor.Executor: Managed memory leak detected; size
7039360 bytes, TID = 382
+-------------+-------------------+--------------------+
|ProviderState|      DRGDefinition|sum(TotalDischarges)|
+-------------+-------------------+--------------------+
|           KY|065 - INTRACRANIA...|                1937|
|           NY|101 - SEIZURES W/...|                4503|
|           IN|149 - DYSEQUILIBRIUM|                 700|
|           IA|178 - RESPIRATORY...|                 540|
|           WI|202 - BRONCHITIS ...|                 338|
|           MO|208 - RESPIRATORY...|                1840|
|           WI|251 - PERC CARDIO...|                 417|
|           AR|281 - ACUTE MYOCA...|                 413|
|           AZ|292 - HEART FAILU...|                2643|
|           NY|292 - HEART FAILU...|               13289|
|           NV|293 - HEART FAILU...|                 519|
|           SD|303 - ATHEROSCLER...|                  53|
|           TN|305 - HYPERTENSIO...|                 730|
|           ME|308 - CARDIAC ARR...|                 312|
|           NV|372 - MAJOR GASTR...|                 126|
|           WA|392 - ESOPHAGITIS...|                3148|
|           WI|439 - DISORDERS O...|                 215|
|           MN|536 - FRACTURES O...|                 332|
|           DC|563 - FX, SPRN, S...|                  43|
|           CO|602 - CELLULITIS ...|                  86|
+-------------+-------------------+--------------------+
only showing top 20 rows
```

```
scala>
df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").
sort(desc(sum("TotalDischarges").toString)).show
+-------------+--------------------+--------------------+
|ProviderState|       DRGDefinition|sum(TotalDischarges)|
+-------------+--------------------+--------------------+
|           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
```



```
scala> df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges"
rt(desc(sum("TotalDischarges").toString)).show
+-------------+--------------------+--------------------+
|ProviderState|       DRGDefinition|sum(TotalDischarges)|
+-------------+--------------------+--------------------+
|           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
```

```
scala>
df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").
orderBy(desc(sum("TotalDischarges").toString)).show
+-------------+--------------------+--------------------+
|ProviderState|       DRGDefinition|sum(TotalDischarges)|
+-------------+--------------------+--------------------+
|           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
```



```
scala> df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges"
derBy(desc(sum("TotalDischarges").toString)).show
+-------------+--------------------+--------------------+
|ProviderState|       DRGDefinition|sum(TotalDischarges)|
+-------------+--------------------+--------------------+
|           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
```