

US Hospital Case Study

Given the hospital dataset do the following 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

SOLUTION

The source code for the spark program and the output is given below. The program does the following

1. Load the file into spark using spark.read method. We tell spark that it is a CSV file with the header and ask it to infer schema
2. We create a temporary view called "hospital" over the dataframe created by reading the CSV file
3. We use multiple SQL queries to answer desired questions. Below is the details of SQL queries for each questions

What is the average amount of AverageCoveredCharges per state

```
SELECT ProviderState,AVG(AverageCoveredCharges) FROM hospital GROUP BY ProviderState
```

find out the AverageTotalPayments charges per state

```
SELECT ProviderState,AVG(AverageTotalPayments) FROM hospital GROUP BY ProviderState
```

find out the AverageMedicarePayments charges per state.

```
SELECT ProviderState,AVG(AverageMedicarePayments) FROM hospital GROUP BY ProviderState
```

Find out the total number of Discharges per state and for each disease

Sort the output in descending order of totalDischarges

```
SELECT ProviderState, DRGDefinition AS Disease, SUM(TotalDischarges) AS Total  
FROM hospital GROUP BY ProviderState,DRGDefinition ORDER BY Total DESC
```

```
package demo
```

```
import org.apache.spark.sql.SparkSession  
import org.apache.spark.sql.functions.udf
```

```
object SparkSqlAssignment {
```

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

```
// create spark session and context
```

```
  val spark = SparkSession.builder().master(master="local")  
    .appName(name="spark sql example")  
    .config("spark.some.config.option", "some-value").getOrCreate()  
  spark.sparkContext.setLogLevel("WARN")
```

```

// read file in csv format
// tell spark that the file has headers and infer schema
val df = spark.read.format("csv").option("header","true")
    .option("inferSchema","true").option("mode","failfast")
    .load("/Users/mmisra/Desktop/acad/assignments/assignment_25.1/inpatientCharges.csv")
println("data frame schema is below")

// print the schema to make sure it is reading and inferring data correctly
df.printSchema()

// create a temp view to run our queries
val tv = df.createOrReplaceTempView("hospital")

println("Ans1. What is the average amount of AverageCoveredCharges per state")
val b = spark.sql("SELECT ProviderState,AVG(AverageCoveredCharges) FROM hospital GROUP BY
ProviderState")
b.show(false)

println("Ans2. find out the AverageTotalPayments charges per state")
val c = spark.sql("SELECT ProviderState,AVG(AverageTotalPayments) FROM hospital GROUP BY
ProviderState")
c.show(false)

println("Ans3. find out the AverageMedicarePayments charges per state")
val d = spark.sql("SELECT ProviderState,AVG(AverageMedicarePayments) FROM hospital GROUP BY
ProviderState")
d.show(false)

println("Ans4. Find out the total number of Discharges per state and for each disease")
val e = spark.sql("SELECT ProviderState, DRGDefinition AS Disease, SUM(TotalDischarges) AS
Total " +
                  " FROM hospital GROUP BY ProviderState,DRGDefinition ORDER BY
Total DESC ")
e.show(false)
}

```

Program Output

```

.....
18/10/01 11:09:05 INFO BlockManagerMasterEndpoint: Registering block manager 192.168.56.1:60827
with 1957.8 MB RAM, BlockManagerId(driver, 192.168.56.1, 60827, None)
18/10/01 11:09:05 INFO BlockManagerMaster: Registered BlockManager BlockManagerId(driver,
192.168.56.1, 60827, None)
18/10/01 11:09:05 INFO BlockManager: Initialized BlockManager: BlockManagerId(driver,
192.168.56.1, 60827, None)
data frame schema is below
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)

```

Ans1. What is the average amount of AverageCoveredCharges per state

```

+-----+-----+
|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

Ans2. find out the AverageTotalPayments charges per state

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

Ans3. find out the AverageMedicarePayments charges per state

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

Ans4.Find out the total number of Discharges per state and for each disease in descending sorted order of total

ProviderState	Disease	Total
CA	871 - SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC	34284
TX	470 - MAJOR JOINT REPLACEMENT OR REATTACHMENT OF LOWER EXTREMITY W/O MCC	30095
FL	470 - MAJOR JOINT REPLACEMENT OR REATTACHMENT OF LOWER EXTREMITY W/O MCC	29985
CA	470 - MAJOR JOINT REPLACEMENT OR REATTACHMENT OF LOWER EXTREMITY W/O MCC	29731
TX	871 - SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC	23144
NY	871 - SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC	21970
FL	392 - ESOPHAGITIS, GASTROENT & MISC DIGEST DISORDERS W/O MCC	21298
IL	470 - MAJOR JOINT REPLACEMENT OR REATTACHMENT OF LOWER EXTREMITY W/O MCC	20095
NY	470 - MAJOR JOINT REPLACEMENT OR REATTACHMENT OF LOWER EXTREMITY W/O MCC	19371
FL	871 - SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC	18660
TX	690 - KIDNEY & URINARY TRACT INFECTIONS W/O MCC	17384
NY	392 - ESOPHAGITIS, GASTROENT & MISC DIGEST DISORDERS W/O MCC	17337
MI	470 - MAJOR JOINT REPLACEMENT OR REATTACHMENT OF LOWER EXTREMITY W/O MCC	16847
PA	470 - MAJOR JOINT REPLACEMENT OR REATTACHMENT OF LOWER EXTREMITY W/O MCC	16712
FL	292 - HEART FAILURE & SHOCK W CC	16639
FL	690 - KIDNEY & URINARY TRACT INFECTIONS W/O MCC	16405
OH	470 - MAJOR JOINT REPLACEMENT OR REATTACHMENT OF LOWER EXTREMITY W/O MCC	16062
NC	470 - MAJOR JOINT REPLACEMENT OR REATTACHMENT OF LOWER EXTREMITY W/O MCC	15820
IL	871 - SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC	15610
MI	871 - SEPTICEMIA OR SEVERE SEPSIS W/O MV 96+ HOURS W MCC	15548

only showing top 20 rows

Process finished with exit code 0