Case Study - V

Objective – 1

Load a CSV file directly into the Spark SQL context

val session = org.apache.spark.sql.SparkSession.builder.master("local").appName("Spark CSV Reader").getOrCreate;
val df = session.read.format("com.databricks.spark.csv").option("header", "true").option("inferSchema", "true").load("file:///home/kiran/Documents/datasets/inpatientCharges.csv")

check the schema of the DataFrame using inferschema

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)

To see the contents inside the DataFrame

df.show

Save the data in a table by registering it in a temp table

df.registerTempTable("hospital_charges")

Objective – 2

What is the average amount of AverageCoveredCharges per state

df.groupBy("ProviderState").avg("AverageCoveredCharges").show

output

```
scala> df.groupBy("ProviderState").avg("AverageCoveredCharges").show
+-------------+--------------------------+
|ProviderState|avg(AverageCoveredCharges)|
+-------------+--------------------------+
|           AZ|         41200.06301999297|
|           SC|         35862.49456269757|
|           LA|        33085.372791542744|
|           MN|         27894.36182060391|
|           NJ|         66125.68627434727|
|           DC|         40116.66365800866|
|           OR|         27390.11187066973|
|           VA|        29222.000487072957|
|           RI|        29942.701122448976|
|           KY|        24523.807169402284|
|           WY|         28700.59862348179|
|           NH|        27059.020801944105|
|           MI|        24124.247209817382|
|           NV|         61047.11541597339|
|           WI|        26149.325331686654|
|           ID|        25565.547041742295|
|           CA|         67508.61653551755|
|           CT|         31318.41011437097|
|           NE|        31736.427824858765|
|           MT|        22670.015237154144|
+-------------+--------------------------+
```

Find out the AverageTotalPayments charges per state

df.groupBy("ProviderState").avg("AverageTotalPayments").show

```
scala> df.groupBy("ProviderState").avg("AverageTotalPayments").show
+-------------+-------------------------+
|ProviderState|avg(AverageTotalPayments)|
+-------------+-------------------------+
|           AZ|       10154.528211153982|
|           SC|        9132.42075869336|
|           LA|        8638.662576808716|
|           MN|        9948.23696269982|
|           NJ|       10678.988646912556|
|           DC|       12998.029415584415|
|           OR|       10436.192863741338|
|           VA|        8887.752176823638|
|           RI|       10589.566853741495|
|           KY|         8278.5888448436|
|           WY|       11398.485910931175|
|           NH|        9289.661822600241|
|           MI|        9754.420405978964|
|           NV|       10291.718028286188|
|           WI|        9270.705617501762|
|           ID|        9827.180090744105|
|           CA|       12629.668472137168|
|           CT|       11365.450671307808|
|           NE|        9331.68252354049|
|           MT|        9252.802766798417|
+-------------+-------------------------+
```

Find out the AverageMedicarePayments charges per state.

df.groupBy("ProviderState").avg("AverageMedicarePayments").show

```
scala> df.groupBy("ProviderState").avg("AverageMedicarePayments").show
+-------------+----------------------------+
|ProviderState|avg(AverageMedicarePayments)|
+-------------+----------------------------+
|           AZ|          8825.717239565063|
|           SC|          7876.331524411663|
|           LA|          7387.704625041294|
|           MN|           8619.21498223801|
|           NJ|           9586.94005594695|
|           DC|         11811.967705627703|
|           OR|          9035.259961508855|
|           VA|          7538.847006001844|
|           RI|           9317.93911564626|
|           KY|          7185.227810467634|
|           WY|           9539.3920242915|
|           NH|          8124.506852976911|
|           MI|          8662.157756043538|
|           NV|           8747.60282861897|
|           WI|          8002.597911079743|
|           ID|          8461.977513611615|
|           CA|         11494.381677893432|
|           CT|          10104.59294380905|
|           NE|          7992.627250470812|
|           MT|          7981.088063241105|
+-------------+----------------------------+
```

Objective – 3

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

df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").show

```
scala> df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").show
16/10/04 15:38:19 WARN TaskMemoryManager: leak 4.3 MB memory from org.apache.spark.unsafe.map.BytesToBytesMap@33d17650
16/10/04 15:38:19 WARN TaskMemoryManager: leak a page: org.apache.spark.unsafe.memory.MemoryBlock@156193f3 in task 829
16/10/04 15:38:19 WARN TaskMemoryManager: leak a page: org.apache.spark.unsafe.memory.MemoryBlock@514dc5c1 in task 829
16/10/04 15:38:19 WARN Executor: Managed memory leak detected; size = 4456448 bytes, TID = 829
+-------------+--------------------+--------------------+
|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
```

Sort the output in descending order of totalDischarges

df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").sort(desc(sum("TotalDischarges").toString)).show

```
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
```

df.groupBy(("ProviderState"),("DRGDefinition")).sum("TotalDischarges").orderBy(desc(sum("TotalDischarges").toString)).show

```
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|
```