Vijay's Assignment – Spark Assignment1

Spark SQL was used to implement this usecase.

Generic code:

```
package vksp1

import org.apache.spark.SparkConf
import org.apache.spark.SparkContext
import org.apache.spark.rdd.RDD.rddToPairRDDFunctions
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions.udf
import org.apache.spark.sql.functions._

import org.apache.spark.sql.types._
object vkassign1 {

    def main(args: Array[String]) {

    val conf = new SparkConf()
      .setAppName("vkassign1")
      .setMaster("local")

    val sc = new SparkContext(conf)

    val spark = SparkSession
      .builder()
      .appName("Spark Assignment1")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()

    import spark.implicits._
```

Task 1:

Load file into spark

Using read.format function and csv option the given csv file was loaded into the DF inpaDF and the header record was seperated out to form the actual schema. This DF was registered as the temp table inpatient

```
val inpaDF = spark.read.format("csv")
    .option("header", "true")
    .option("inferschema", "true")
    .load("C:\\Users\\VIJAYLAKSHMANAN\\spark\\inpatientCharges.csv")


    inpaDF.registerTempTable("inpatient")
```

Task 2:

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

Code:

```
val query1DF = spark.sql("select ProviderState, avg(AverageCoveredCharges) from inpatient group by ProviderState")
    query1DF.show()
```

Output:

```
19/03/11 18:12:10 INFO DAGScheduler: Job 5 finished: show at vkassign1.scala:43, took 1.088519 s
+-------------+-------------------------+
|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
```

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

Code:

```
val query2DF = spark.sql("select ProviderState, sum(AverageTotalPayments) from inpatient group by ProviderState")
      query2DF.show()
```

Output:

```
19/03/11 18:45:01 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
+-------------+-------------------------+
|ProviderState|sum(AverageTotalPayments)|
+-------------+-------------------------+
|           AZ|     2.8950559930000026E7|
|           SC|     2.6000001900000013E7|
|           LA|     2.6149231619999968E7|
|           MN|     2.2403429640000023E7|
|           NJ|     5.1536799209999874E7|
|           DC|        6005089.589999995|
|           OR|     1.3556614529999994E7|
|           VA|      3.850174243000001E7|
|           RI|        6179625.309999993|
|           KY|     2.6731563380000085E7|
|           WY|        2815426.019999998|
|           NH|        7645391.680000004|
|           MI|      5.285920417999992E7|
|           NV|     1.2370645069999998E7|
|           WI|     2.6273179719999947E7|
|           ID|        5414776.230000002|
|           CA|     1.6499398891999936E8|
|           CT|     2.2855921299999975E7|
|           NE|        9910246.840000004|
|           MT|        4681918.200000002|
+-------------+-------------------------+
only showing top 20 rows
```

➢ find out the **AverageMedicarePayments** charges per state.

Code:

```
val query3DF = spark.sql("select ProviderState, sum(AverageMedicarePayments) from
inpatient group by ProviderState")
          query3DF.show()
```

Output:

```
19/03/11 18:46:31 INFO DAGScheduler: Job 5 finished: show at vkassign1.scala:47, took 1.209041 s
+-------------+---------------------------+
|ProviderState|sum(AverageMedicarePayments)|
+-------------+---------------------------+
|           AZ|        2.5162119849999946E7|
|           SC|        2.2423915850000024E7|
|           LA|        2.2362581899999958E7|
|           MN|        1.9410472139999993E7|
|           NJ|          4.626657270999998E7|
|           DC|           5457129.080000001|
|           OR|        1.1736802689999992E7|
|           VA|        3.2658285229999997E7|
|           RI|           5478948.199999998|
|           KY|         2.320110060000003E7|
|           WY|          2356229.8299999996|
|           NH|                6686469.14|
|           MI|         4.694023287999996E7|
|           NV|        1.0514618599999994E7|
|           WI|        2.2679362479999956E7|
|           ID|           4662549.610000001|
|           CA|        1.5016260224000034E8|
|           CT|         2.032033641000002E7|
|           NE|           8488170.13999999|
|           MT|          4038430.559999998|
+-------------+---------------------------+
only showing top 20 rows
```

Task 3:

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

Code:

```
val query4DF = spark.sql("select ProviderState, DRGDefinition,sum(TotalDischarges)
from inpatient group by ProviderState,DRGDefinition")
          query4DF.show()
```

Output:

```
19/03/11 18:48:47 INFO DAGScheduler: Job 2 finished: show at vkassign1.scala:50, took 2.684953 s
+-------------+--------------------+--------------------+
|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**

Code:

```
val query5DF = spark.sql("select ProviderState, DRGDefinition,sum(TotalDischarges)
as tot from inpatient group by ProviderState,DRGDefinition order by tot desc ")
    query5DF.show()
```

Output:

```
19/03/11 18:52:56 INFO CodeGenerator: Code generated in 16.6327 ms
+-------------+--------------------+-----+
|ProviderState|      DRGDefinition|  tot|
+-------------+--------------------+-----+
|           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
```