CASE STUDY 4

Objective1: Load file into Spark

Sol:

Step1 -> Create RDD and load file into it

//open spark-shell and create rdd

➔ val baseRDD = sc.textFile("/DatasetCS4/inpatientCharges.csv")

//remove header

➔ val header = baseRDD.first()
➔ val rdd1 = baseRDD.filter(row => row != header)

//define case class for the schema

➔ case class patient(DRGDef:String, ProviderId:String, ProviderName:String, ProviderStAddr:String, ProviderCity:String, ProviderState:String, ProviderZip:String, HospitalReferralRegionDes:String, TotalDischarges:String, AverageCoveredCharges:String, AverageTotalPayments:String, AverageMedicarePayments:String)

```
scala> val baseRDD = sc.textFile("/DatasetCS4/inpatientCharges.csv")
baseRDD: org.apache.spark.rdd.RDD[String] = /DatasetCS4/inpatientCharges.csv Map
PartitionsRDD[1] at textFile at <console>:24

scala> val header = baseRDD.first()
header: String = DRGDefinition,ProviderId,ProviderName,ProviderStreetAddress,Pro
viderCity,ProviderState,ProviderZipCode,HospitalReferralRegionDescription,TotalD
ischarges,AverageCoveredCharges,AverageTotalPayments,AverageMedicarePayments

scala> val rdd1 = baseRDD.filter(row => row != header)
rdd1: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at filter at <conso
le>:28

scala> case class patient(DRGDef: String, ProviderId: Int, ProviderName: String,
 ProviderStAddr:String, ProviderCity: String, ProviderState: String, ProviderZip
: Int, HospitalReferralRegionDes: String, TotalDischarges: Int, AverageCoveredCh
arges: Double,
     | AverageTotalPayments: Double, AverageMedicarePayments: Double)
defined class patient
```

//convert to dataframe

➔ val hospDF = rdd1.map(x=>x.split(",")).filter(x=>x.length>=12).map(x=> patient(x(0),x(1).toInt,x(2), x(3), x(4), x(5), x(6).toInt, x(7), x(8).toInt, x(9).toDouble, x(10).toDouble, x(11).toDouble)).toDF

➔ hospDF.show

```
scala> val hospDF = rdd1.map(x=>x.split(",")).filter(x=>x.length>=12).map(x=> patient(x(0),x(1).toInt,x(2), x(3), x(4), x(5), x(6).toI
nt, x(7), x(8).toInt, x(9).toDouble, x(10).toDouble, x(11).toDouble)).toDF
hospDF: org.apache.spark.sql.DataFrame = [DRGDef: string, ProviderId: int ... 10 more fields]

scala> hospDF.show
+--------------------+----------+--------------------+--------------------+------------+-------------+-----------+--------------------
-----+--------------+--------------------+--------------------+----------------------+
|             DRGDef|ProviderId|        ProviderName|      ProviderStAddr|ProviderCity|ProviderState|ProviderZip|HospitalReferralRegi
onDes|TotalDischarges|AverageCoveredCharges|AverageTotalPayments|AverageMedicarePayments|
+--------------------+----------+--------------------+--------------------+------------+-------------+-----------+--------------------
-----+--------------+--------------------+--------------------+----------------------+
|039 - EXTRACRANIA...|     10001|SOUTHEAST ALABAMA...|1108 ROSS CLARK C...|      DOTHAN|           AL|      36301|              AL - D
othan|            91|            32963.07|             5777.24|               4763.73|
|039 - EXTRACRANIA...|     10005|MARSHALL MEDICAL ...|2505 U S HIGHWAY ...|        BOAZ|           AL|      35957|          AL - Birmi
ngham|            14|            15131.85|             5787.57|               4976.71|
|039 - EXTRACRANIA...|     10006|ELIZA COFFEE MEMO...|  205 MARENGO STREET|    FLORENCE|           AL|      35631|          AL - Birmi
ngham|            24|            37560.37|             5434.95|               4453.79|
|039 - EXTRACRANIA...|     10011|   ST VINCENT'S EAST|50 MEDICAL PARK E...|  BIRMINGHAM|           AL|      35235|          AL - Birmi
ngham|            25|            13998.28|             5417.56|               4129.16|
|039 - EXTRACRANIA...|     10016|SHELBY BAPTIST ME...|1000 FIRST STREET...|   ALABASTER|           AL|      35007|          AL - Birmi
ngham|            18|            31633.27|             5658.33|               4851.44|
|039 - EXTRACRANIA...|     10023|BAPTIST MEDICAL C...|2105 EAST SOUTH B...|  MONTGOMERY|           AL|      36116|          AL - Montg
omery|            67|            16920.79|              6653.8|               5374.14|
|039 - EXTRACRANIA...|     10029|EAST ALABAMA MEDI...|2000 PEPPERELL PA...|     OPELIKA|           AL|      36801|          AL - Birmi
ngham|            51|            11977.13|             5834.74|               4761.41|
|039 - EXTRACRANIA...|     10033|UNIVERSITY OF ALA...|619 SOUTH 19TH ST...|  BIRMINGHAM|           AL|      35233|          AL - Birmi
ngham|            32|            35841.09|             8031.12|                5858.5|
|039 - EXTRACRANIA...|     10039| HUNTSVILLE HOSPITAL|      101 SIVLEY RD|  HUNTSVILLE|           AL|      35801|          AL - Hunts
ville|           135|            28523.39|             6113.38|                5228.4|
|039 - EXTRACRANIA...|     10040|GADSDEN REGIONAL ...|1007 GOODYEAR AVENUE|     GADSDEN|           AL|      35903|          AL - Birmi
ngham|            34|            75233.38|             5541.05|               4386.94|
|039 - EXTRACRANIA...|     10046|RIVERVIEW REGIONA...|600 SOUTH THIRD S...|     GADSDEN|           AL|      35901|          AL - Birmi
ngham|            14|            67327.92|             5461.57|               4493.57|
|039 - EXTRACRANIA...|     10055|    FLOWERS HOSPITAL|4370 WEST MAIN ST...|      DOTHAN|           AL|      36305|              AL - D
othan|            45|            39607.28|             5356.28|                4408.2|
```

Objective2:

2.1 What is the average amount of AverageCoveredCharges per state

Sol: Step1: Register the dataframe as temp table

➔ hospDF.registerTempTable("hospDF")

Step2: write sql query to calculate average

➔ spark.sql("select ProviderState, avg(AverageCoveredCharges) as AvgAmt from hospDF group by ProviderState").show

```
acadgild@localhost:~
scala> hospDF.registerTempTable("hospDF")
warning: there was one deprecation warning; re-run with -deprecation for details

scala> spark.sql("select ProviderState, avg(AverageCoveredCharges) as AvgAmt from hospDF group by ProviderState").show
18/05/27 14:30:30 WARN executor.Executor: Managed memory leak detected; size = 17039360 bytes, TID = 34
+--------------------+------------------+
|       ProviderState|            AvgAmt|
+--------------------+------------------+
|            TOWANDA|              17.0|
|           SAN PABLO|              27.2|
|        PO BOX 1727"|              null|
|          CUMBERLAND| 54.57142857142857|
|             HANCOCK|              18.0|
|           PRINCETON|              51.1|
|           WATERTOWN|30.571428571428573|
|             EDMONDS|23.571428571428573|
|         MCMINNVILLE|              38.0|
|               BOAZ|              37.5|
|              BAXLEY|              14.0|
|               30002|              null|
|              140082|              null|
|              150089|              null|
|              330024|              null|
|   750 MORPHY AVENUE|              null|
|2500 ROCKY MOUNTA...|              null|
|          20 YORK ST|              null|
|    1000 MAR-WALT DR|              null|
|    14000 FIVAY ROAD|              null|
+--------------------+------------------+
only showing top 20 rows
```

2.2 find out the AverageTotalPayments charges per state

Sol: Step1: write query to calculate the sum

➔ spark.sql("select ProviderState, sum(AverageTotalPayments) as TotalCharges from hospDF group by ProviderState").show

```
acadgild@localhost:~

scala> spark.sql("select ProviderState, sum(AverageTotalPayments) as TotalCharges from hospDF group by ProviderState").show
18/05/27 14:57:34 WARN executor.Executor: Managed memory leak detected; size = 17039360 bytes, TID = 40
+--------------------+------------------+
|       ProviderState|      TotalCharges|
+--------------------+------------------+
|            TOWANDA|         186600.04|
|           SAN PABLO|310797.86999999994|
|        PO BOX 1727"|             254.0|
|          CUMBERLAND|          80783.02|
|             HANCOCK|          22929.57|
|           PRINCETON|331020.16000000003|
|           WATERTOWN|         152122.09|
|             EDMONDS|         260504.04|
|         MCMINNVILLE|          91759.72|
|               BOAZ|           34349.8|
|              BAXLEY|          10866.35|
|               30002|           85006.0|
|              140082|           60640.0|
|              150089|           47303.0|
|              330024|           10029.0|
|   750 MORPHY AVENUE|              60.0|
|2500 ROCKY MOUNTA...|              44.0|
|          20 YORK ST|             344.0|
|    1000 MAR-WALT DR|             107.0|
|    14000 FIVAY ROAD|              74.0|
+--------------------+------------------+
only showing top 20 rows
```

2.3 Find out the AverageMedicarePayments charges per state

Sol: Step1: write query to calculate medicareCharges

➔ spark.sql("select ProviderState, sum(AverageMedicarePayments) as MedicareCharges from hospDF group by ProviderState").show

```
acadgild@localhost:~

scala> spark.sql("select ProviderState, sum(AverageMedicarePayments) as MedicareCharges from hospDF group by ProviderState").show
18/05/27 14:58:21 WARN executor.Executor: Managed memory leak detected; size = 17039360 bytes, TID = 42
+--------------------+------------------+
|       ProviderState|   MedicareCharges|
+--------------------+------------------+
|             TOWANDA| 85933.35999999999|
|           SAN PABLO| 55995.32000000001|
|        PO BOX 1727"|146123.02000000002|
|          CUMBERLAND|          76058.12|
|             HANCOCK|          16149.57|
|           PRINCETON| 69283.06000000001|
|           WATERTOWN|          65885.64|
|             EDMONDS| 59794.41999999999|
|         MCMINNVILLE|          25478.34|
|                BOAZ|12612.869999999999|
|              BAXLEY|           4328.57|
|               30002|              null|
|              140082|              null|
|              150089|              null|
|              330024|              null|
|   750 MORPHY AVENUE|          29557.27|
|2500 ROCKY MOUNTA...|          26560.64|
|          20 YORK ST|          58816.36|
|    1000 MAR-WALT DR|           92185.2|
|    14000 FIVAY ROAD|          65990.82|
+--------------------+------------------+
only showing top 20 rows
```

Objective3:

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

Sol: command used:

➔ spark.sql("select DRGDef, ProviderState, sum(TotalDischarges) as SumDischarges from hospDF group by ProviderState, DRGDef").show

```
acadgild@localhost:~

scala> spark.sql("select DRGDef, ProviderState, sum(TotalDischarges) as SumDischarges from hospDF group by ProviderState, DRGDef").sho
w
18/05/27 15:24:12 WARN executor.Executor: Managed memory leak detected; size = 17039360 bytes, TID = 255
+--------------------+-------------+-------------+
|              DRGDef|ProviderState|SumDischarges|
+--------------------+-------------+-------------+
|039 - EXTRACRANIA...|        TULSA|         null|
|057 - DEGENERATIV...|      SAGINAW|         null|
|065 - INTRACRANIA...|      CONYERS|         null|
|065 - INTRACRANIA...|           KY|       1747.0|
|066 - INTRACRANIA...|    GREENBRAE|         null|
|101 - SEIZURES W/...|           NY|       4305.0|
|101 - SEIZURES W/...| JOHNSON CITY|         null|
|101 - SEIZURES W/...|    LYNCHBURG|         null|
|149 - DYSEQUILIBRIUM|           IN|        700.0|
|176 - PULMONARY E...|   BINGHAMTON|         null|
|178 - RESPIRATORY...|           IA|        540.0|
|190 - CHRONIC OBS...|   FITZGERALD|         null|
|192 - CHRONIC OBS...|      ARDMORE|         null|
|193 - SIMPLE PNEU...|SAN FRANCISCO|         null|
|193 - SIMPLE PNEU...|  PAINTSVILLE|         null|
|193 - SIMPLE PNEU...|    LAFAYETTE|         null|
|202 - BRONCHITIS ...|   FALL RIVER|         null|
|202 - BRONCHITIS ...|   GREENSBORO|         null|
|202 - BRONCHITIS ...|           WI|        324.0|
|208 - RESPIRATORY...|           MO|       1754.0|
+--------------------+-------------+-------------+
only showing top 20 rows
```

3.2 Sort the output in descending order of totalDischarges

Sol: Command used:

➔ spark.sql("select DRGDef, ProviderState, sum(TotalDischarges) as SumDischarges from hospDF group by ProviderState, DRGDef order by SumDischarges desc").show