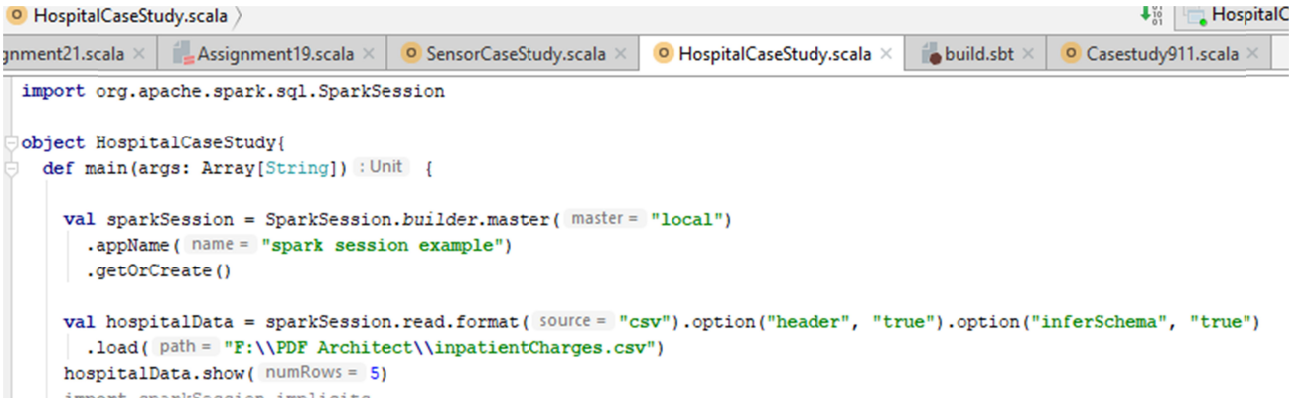
Hospital Data analysis in the US

Objective 1

Load file into spark




```
18/08/05 19:20:19 INFO ContextCleaner: Cleaned accumulator 70
18/08/05 19:20:19 INFO ContextCleaner: Cleaned accumulator 72
+--------------------+---------+--------------------+--------------------+------------+------------+--------------+--------------------+---------------+--------------------+
|       DRGDefinition|ProviderId|        ProviderName|ProviderStreetAddress|ProviderCity|ProviderState|ProviderZipCode|HospitalReferralRegionDescription|TotalDischarges|AverageCoveredCha|
+--------------------+---------+--------------------+--------------------+------------+------------+--------------+--------------------+---------------+--------------------+
|039 - EXTRACRANIA...|     10001|SOUTHEAST ALABAMA...| 1108 ROSS CLARK C...|      DOTHAN|          AL|         36301|          AL - Dothan|             91|              3294|
|039 - EXTRACRANIA...|     10005|MARSHALL MEDICAL ...| 2505 U S HIGHWAY ...|        BOAZ|          AL|         35957|      AL - Birmingham|             14|              1513|
|039 - EXTRACRANIA...|     10006|ELIZA COFFEE MEMO...|   205 MARENGO STREET|    FLORENCE|          AL|         35631|      AL - Birmingham|             24|              3754|
|039 - EXTRACRANIA...|     10011|   ST VINCENT'S EAST| 50 MEDICAL PARK E...|  BIRMINGHAM|          AL|         35235|      AL - Birmingham|             25|              1399|
|039 - EXTRACRANIA...|     10016|SHELBY BAPTIST ME...| 1000 FIRST STREET...|   ALABASTER|          AL|         35007|      AL - Birmingham|             18|              3163|
+--------------------+---------+--------------------+--------------------+------------+------------+--------------+--------------------+---------------+--------------------+
only showing top 5 rows

18/08/05 19:20:19 INFO SparkContext: Invoking stop() from shutdown hook
18/08/05 19:20:19 INFO SparkUI: Stopped Spark web UI at http://192.168.1.8:4040
```

Objective 2

- What is the average amount of AverageCoveredCharges per state

```
18/08/05 19:04:00 INFO Executor: Running task 91.0 in stage 18.0 (TID 242)
18/08/05 19:04:00 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/08/05 19:04:00 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 1 ms
+-------------+-------------------------+
|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

18/08/05 19:04:00 INFO Executor: Finished task 91.0 in stage 18.0 (TID 242). 2705 bytes result sent to driver
18/08/05 19:04:00 INFO TaskSetManager: Finished task 91.0 in stage 18.0 (TID 242) in 11 ms on localhost (executor driver) (94/94)
18/08/05 19:04:00 INFO TaskSchedulerImpl: Removed TaskSet 18.0, whose tasks have all completed, from pool
```

- Find out the AverageTotalPayments charges per state

```
18/08/05 19:13:46 INFO TaskSetManager: Finished task 91.0 in stage 18.0 (TID 242) in 34 ms on localhost (executor driver) (94/94)
18/08/05 19:13:46 INFO TaskSchedulerImpl: Removed TaskSet 18.0, whose tasks have all completed, from pool
+-------------+-------------------------+
|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

18/08/05 19:13:46 INFO DAGScheduler: ResultStage 18 (show at HospitalCaseStudy.scala:16) finished in 0.782 s
18/08/05 19:13:46 INFO DAGScheduler: Job 10 finished: show at HospitalCaseStudy.scala:16, took 0.792572 s
```

- Find out the AverageMedicarePayments charges per state.

```
18/08/05 19:18:10 INFO Executor: Running task 91.0 in stage 26.0 (TID 362)
18/08/05 19:18:10 INFO ShuffleBlockFetcherIterator: Getting 1 non-empty blocks out of 1 blocks
18/08/05 19:18:10 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
+-------------+---------------------------+
|ProviderState|sum(AverageMedicarePayments)|
+-------------+---------------------------+
|           AZ|        2.5162119849999946E7|
|           SC|        2.2423915850000024E7|
|           LA|        2.2362581899999958E7|
|           MN|        1.9410472139999993E7|
|           NJ|          4.62665727099998E7|
|           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

18/08/05 19:18:10 INFO Executor: Finished task 91.0 in stage 26.0 (TID 362). 2701 bytes result sent to driver
```

Objective 3

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

```
val hospitalData = sparkSession.read.format( source = "csv").option("header", "true").option("inferSchema", "true")
  .load( path = "F:\\PDF Architect\\inpatientCharges.csv")
hospitalData.show( numRows = 5)
import sparkSession.implicits._

val totalDischarge = hospitalData.groupBy( cols = $"ProviderState",$"DRGDefinition").sum( colNames = "TotalDischarges").show()
```

```
18/08/05 19:26:10 INFO DAGScheduler: ResultStage 4 (show at HospitalCaseStudy.scala:15) finished in 0.096 s
18/08/05 19:26:10 INFO DAGScheduler: Job 3 finished: show at HospitalCaseStudy.scala:15, took 10.462331 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

18/08/05 19:26:10 INFO SparkContext: Invoking stop() from shutdown hook
18/08/05 19:26:10 INFO SparkUI: Stopped Spark web UI at http://192.168.1.8:4040
```

- Sort the output in descending order of totalDischarges

```
val hospitalData = sparkSession.read.format( source = "csv").option("header", "true").option("inferSchema", "true")
  .load( path = "F:\\PDF Architect\\inpatientCharges.csv")
hospitalData.show( numRows = 5)
import sparkSession.implicits._

val totalDischarge = hospitalData.groupBy( cols = $"ProviderState",$"DRGDefinition")
    .sum( colNames = "TotalDischarges").withColumnRenamed( existingName = "sum(TotalDischarges)",  newName = "Sum")
totalDischarge.orderBy(($"Sum").desc ).show()
```

```
18/08/05 20:02:12 INFO DAGScheduler: ResultStage 8 (show at HospitalCaseStudy.scala:19) finished in 1.761 s
18/08/05 20:02:12 INFO DAGScheduler: Job 5 finished: show at HospitalCaseStudy.scala:19, took 3.861252 s
+-------------+--------------------+-----+
|ProviderState|      DRGDefinition|  Sum|
+-------------+--------------------+-----+
|           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

18/08/05 20:02:12 INFO SparkContext: Invoking stop() from shutdown hook
18/08/05 20:02:12 INFO SparkUI: Stopped Spark web UI at http://192.168.1.8:4040
```