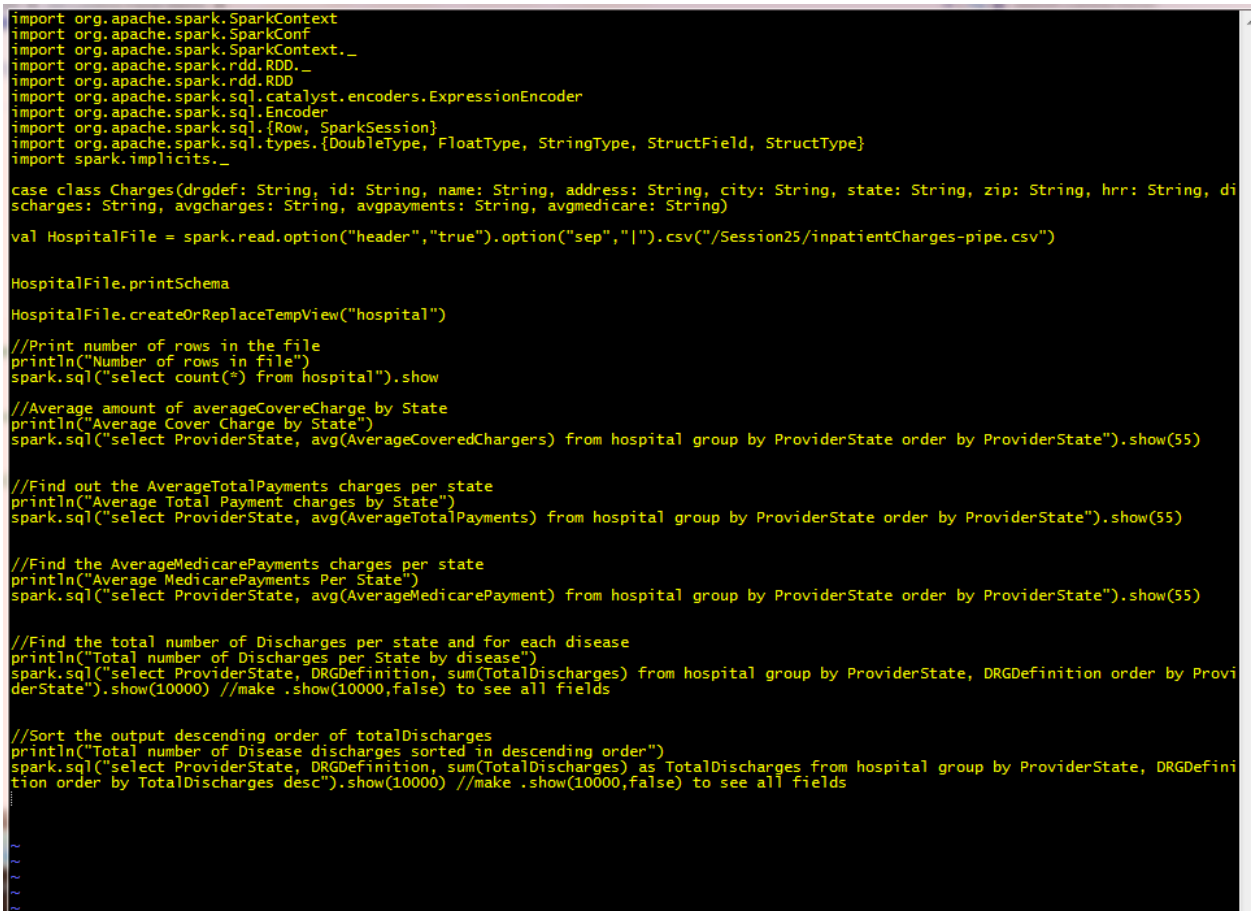
1. Load file into Spark
2. What is the average amount of AverageCoveredCharges per state?
3. Find the AverageTotalPayments charges per state.
4. Find the AverageMedicarePayments charges per state
5. Find the Total number of Discharges per state for each disease
6. Sort the output in descending order of TotalDischarges

Script:



```
import org.apache.spark.SparkContext
import org.apache.spark.SparkConf
import org.apache.spark.SparkContext._
import org.apache.spark.rdd.RDD._
import org.apache.spark.rdd.RDD
import org.apache.spark.sql.catalyst.encoders.ExpressionEncoder
import org.apache.spark.sql.Encoder
import org.apache.spark.sql.{Row, SparkSession}
import org.apache.spark.sql.types.{DoubleType, FloatType, StringType, StructField, StructType}
import spark.implicits._

case class Charges(drgdef: String, id: String, name: String, address: String, city: String, state: String, zip: String, hrr: String, di
scharges: String, avgcharges: String, avgpayments: String, avgmedicare: String)

val HospitalFile = spark.read.option("header","true").option("sep","|").csv("/Session25/inpatientCharges-pipe.csv")


HospitalFile.printSchema

HospitalFile.createOrReplaceTempView("hospital")

//Print number of rows in the file
println("Number of rows in file")
spark.sql("select count(*) from hospital").show

//Average amount of averageCovereCharge by State
println("Average Cover Charge by State")
spark.sql("select ProviderState, avg(AverageCoveredChargers) from hospital group by ProviderState order by ProviderState").show(55)


//Find out the AverageTotalPayments charges per state
println("Average Total Payment charges by State")
spark.sql("select ProviderState, avg(AverageTotalPayments) from hospital group by ProviderState order by ProviderState").show(55)


//Find the AverageMedicarePayments charges per state
println("Average MedicarePayments Per State")
spark.sql("select ProviderState, avg(AverageMedicarePayment) from hospital group by ProviderState order by ProviderState").show(55)


//Find the total number of Discharges per state and for each disease
println("Total number of Discharges per State by disease")
spark.sql("select ProviderState, DRGDefinition, sum(TotalDischarges) from hospital group by ProviderState, DRGDefinition order by Provi
derState").show(10000) //make .show(10000,false) to see all fields


//Sort the output descending order of totalDischarges
println("Total number of Disease discharges sorted in descending order")
spark.sql("select ProviderState, DRGDefinition, sum(TotalDischarges) as TotalDischarges from hospital group by ProviderState, DRGDefini
tion order by TotalDischarges desc").show(10000) //make .show(10000,false) to see all fields
```

Output of the script addresses all questions. For Brevity, script was modified to print only the default first 20 rows for the disease related queries.

```
tony@tony-PC ~
$ ssh acadgild@192.168.156.230
acadgild@192.168.156.230's password:
Last login: Tue Jan  8 09:26:57 2019 from 192.168.156.201
[acadgild@localhost ~]$ cd /myCode
-bash: cd: /myCode: No such file or directory
[acadgild@localhost ~]$ ls
```

```
[acadgild@localhost Session25]$ spark-shell -i Assignment.scala
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLeve
l(newLevel).
19/01/08 09:33:23 WARN util.NativeCodeLoader: Unable to load native-hadoop libra
ry for your platform... using builtin-java classes where applicable
19/01/08 09:33:23 WARN util.Utils: Your hostname, localhost.localdomain resolves
to a loopback address: 127.0.0.1; using 192.168.156.230 instead (on interface e
th16)
19/01/08 09:33:23 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to ano
ther address
19/01/08 09:33:25 WARN util.Utils: Service 'SparkUI' could not bind on port 4040
. Attempting port 4041.
Spark context Web UI available at http://192.168.156.230:4041
Spark context available as 'sc' (master = local[*], app id = local-1546920205580
).
Spark session available as 'spark'.
Loading Assignment.scala...
import org.apache.spark.SparkContext
import org.apache.spark.SparkConf
import org.apache.spark.SparkContext._
import org.apache.spark.rdd.RDD._
import org.apache.spark.rdd.RDD
import org.apache.spark.sql.catalyst.encoders.ExpressionEncoder
import org.apache.spark.sql.Encoder
import org.apache.spark.sql.{Row, SparkSession}
import org.apache.spark.sql.types.{DoubleType, FloatType, StringType, StructFiel
d, StructType}
import spark.implicits._
defined class Charges
HospitalFile: org.apache.spark.sql.DataFrame = [DRGDefinition: string, ProviderID:
string ... 10 more fields]
root
 |-- DRGDefinition: string (nullable = true)
 |-- ProviderID: string (nullable = true)
 |-- ProviderName: string (nullable = true)
 |-- ProviderStreetAddress: string (nullable = true)
 |-- ProviderCity: string (nullable = true)
 |-- ProviderState: string (nullable = true)
 |-- ProviderZipCode: string (nullable = true)
 |-- HospitalReferralRegionDescription: string (nullable = true)
 |-- TotalDischarges: string (nullable = true)
 |-- AverageCoveredChargers: string (nullable = true)
 |-- AverageTotalPayments: string (nullable = true)
 |-- AverageMedicarePayment: string (nullable = true)

Number of rows in file
+--------+
|count(1)|
+--------+
|  163065|
+--------+

Average Cover Charge by State
+-------------+----------------------------------------+
|ProviderState|avg(CAST(AverageCoveredChargers AS DOUBLE))|
+-------------+----------------------------------------+
|           AK|                       40348.74333333336|
|           AL|                      31316.462074277857|
|           AR|                       26174.52624576683|
|           AZ|                      41200.063019992995|
|           CA|                         67508.616535517|
|           CO|                       41095.13611111104|
|           CT|                        31318.4101143709|
|           DC|                       40116.66365800864|
|           DE|                      27071.699644670043|
|           FL|                       46016.23358673223|
|           GA|                       31096.93284218991|
|           HI|                      32174.748076923064|
|           IA|                      24168.742041522488|
|           ID|                      25565.547041742288|
|           IL|                       36061.84987861936|
```

IN	28144.7125446009
KS	31580.253663003667
KY	24523.80716940223
LA	33085.372791542846
MA	20534.00671264962
MD	13377.803789789768
ME	20394.957567567573
MI	24124.247209817277
MN	27894.36182060388
MO	31184.622902192626
MS	30292.785203319454
MT	22670.015237154144
NC	25140.952162269463
ND	21636.883459715627
NE	31736.427824858758
NH	27059.020801944105
NJ	66125.68627434729
NM	30011.406499454773
NV	61047.11541597337
NY	31435.685542601852
OH	28344.21854677697
OK	29587.57526587295
OR	27390.111870669723
PA	39633.9597629422
RI	29942.701122448976
SC	35862.49456269756
SD	29609.991543209864
TN	29279.931835412586
TX	41480.1934035738
UT	25092.80687158469
VA	29222.000487072903
VT	20074.958333333325
WA	34714.234074873886
WI	26149.325331686607
WV	19191.508634361224
WY	28700.59862348178
+-------------+--+

Average Total Payment charges by State

ProviderState	avg(CAST(AverageTotalPayments AS DOUBLE))
AK	14572.391731601727
AL	7568.232148555701
AR	8019.248805031454
AZ	10154.528211153991
CA	12629.668472137122
CO	9502.685550264529
CT	11365.450671307795
DC	12998.029415584406
DE	10360.072411167508
FL	8826.99043567904
GA	8925.793915056342
HI	12775.739524886882
IA	8312.571707035755
ID	9827.180090744107
IL	9790.67609685165
IN	8756.082532863868
KS	8455.476422466427
KY	8278.58884484363
LA	8638.66257680871
MA	10279.981535658479
MD	12608.947663663681
ME	8679.994977477474
MI	9754.420405978948
MN	9948.236962699833
MO	8724.631271249062
MS	8229.16482572614
MT	9252.802766798422
NC	9089.435711168418
ND	9827.63803317535
NE	9331.682523540492

NH	9289.661822600248
NJ	10678.98864691253
NM	9619.84109051253
NV	10291.718028286188
NY	11795.492051645204
OH	8808.127651169372
OK	8353.641035714261
OR	10436.192863741335
PA	9100.04321758069
RI	10509.566853741484
SC	9132.420758693366
SD	10141.688004115227
TN	8153.950854126662
TX	9243.97957265677
UT	9749.907076502734
VA	8887.75217682364
VT	11766.304481481482
WA	10543.151652267818
WI	9270.705617501746
WV	7968.4802454373785
WY	11398.485910931167
+-------------+--+

Average MedicarePayments Per State

ProviderState	avg(CAST(AverageMedicarePayment AS DOUBLE))
AK	12958.96943722943
AL	6418.007119669867
AR	6919.720832123854
AZ	8825.717239565045
CA	11494.381677893474
CO	8150.931391534383
CT	10104.592943809059
DC	11811.967705627709
DE	8959.67327411167
FL	7667.478694755685
GA	7667.581549919492
HI	10967.475045248866
IA	7148.119994232981
ID	8461.977513611617
IL	8385.28267290425
IN	7478.296323943639
KS	7224.643162393148
KY	7185.227810467647
LA	7387.704625041281
MA	9241.719323269102
MD	11480.121828828853
ME	7590.500810810822
MI	8662.157756043543
MN	8619.214982238007
MO	7589.245870904166
MS	7124.175751037359
MT	7981.088063241104
NC	7998.649702439984
ND	8752.048293838863
NE	7992.6272504707995
NH	8124.506852976913
NJ	9586.940055946912
NM	8300.111439476545
NV	8747.602828618963
NY	10620.73639790799
OH	7661.033712207648
OK	7207.706055555556
OR	9035.259961508847
PA	7919.184856483817
RI	9317.939115646255
SC	7876.33152441167
SD	8641.162078189302
TN	6946.489304222644
TX	7970.430797370104
UT	7829.6286885245845

```
|           VA|                          7538.847006001846|
|           VT|                         10546.969962962965|
|           WA|                          9076.509287257013|
|           WI|                          8002.597911079731|
|           WV|                          6900.852548772811|
|           WY|                          9539.392024291496|
+-------------+-----------------------------------------+

Total number of Discharges per State by disease
+-------------+--------------------+-----------------------------------+
|ProviderState|       DRGDefinition|sum(CAST(TotalDischarges AS DOUBLE))|
+-------------+--------------------+-----------------------------------+
|           AK|309 - CARDIAC ARR...|                              101.0|
|           AK|246 - PERC CARDIO...|                               37.0|
|           AK|286 - CIRCULATORY...|                               18.0|
|           AK| 638 - DIABETES W CC|                               32.0|
|           AK|689 - KIDNEY & UR...|                               12.0|
|           AK|247 - PERC CARDIO...|                              131.0|
|           AK|811 - RED BLOOD C...|                               13.0|
|           AK|389 - G.I. OBSTRU...|                               14.0|
|           AK|039 - EXTRACRANIA...|                               23.0|
|           AK|195 - SIMPLE PNEU...|                              126.0|
|           AK|291 - HEART FAILU...|                              111.0|
|           AK|690 - KIDNEY & UR...|                              188.0|
|           AK|065 - INTRACRANIA...|                              152.0|
|           AK|481 - HIP & FEMUR...|                               57.0|
|           AK|287 - CIRCULATORY...|                              143.0|
|           AK|282 - ACUTE MYOCA...|                               12.0|
|           AK|251 - PERC CARDIO...|                               63.0|
|           AK|303 - ATHEROSCLER...|                               12.0|
|           AK|491 - BACK & NECK...|                               84.0|
|           AK|300 - PERIPHERAL ...|                               12.0|
+-------------+--------------------+-----------------------------------+
only showing top 20 rows

Total number of Disease discharges sorted in descending order
+-------------+--------------------+---------------+
|ProviderState|       DRGDefinition|TotalDischarges|
+-------------+--------------------+---------------+
|           CA|871 - SEPTICEMIA ...|        34284.0|
|           TX|470 - MAJOR JOINT...|        30095.0|
|           FL|470 - MAJOR JOINT...|        29985.0|
|           CA|470 - MAJOR JOINT...|        29731.0|
|           TX|871 - SEPTICEMIA ...|        23144.0|
|           NY|871 - SEPTICEMIA ...|        21970.0|
|           FL|392 - ESOPHAGITIS...|        21298.0|
|           IL|470 - MAJOR JOINT...|        20095.0|
|           NY|470 - MAJOR JOINT...|        19371.0|
|           FL|871 - SEPTICEMIA ...|        18660.0|
|           TX|690 - KIDNEY & UR...|        17384.0|
|           NY|392 - ESOPHAGITIS...|        17337.0|
|           MI|470 - MAJOR JOINT...|        16847.0|
|           PA|470 - MAJOR JOINT...|        16712.0|
|           FL|292 - HEART FAILU...|        16639.0|
|           FL|690 - KIDNEY & UR...|        16405.0|
|           OH|470 - MAJOR JOINT...|        16062.0|
|           NC|470 - MAJOR JOINT...|        15820.0|
|           IL|871 - SEPTICEMIA ...|        15610.0|
|           MI|871 - SEPTICEMIA ...|        15548.0|
+-------------+--------------------+---------------+
only showing top 20 rows


Welcome to
      ____              __
     / __/__  ___ _____/ /__
    _\ \/ _ \/ _ `/ __/  '_/
   /___/ .__/\_,_/_/ /_/\_\   version 2.2.1
      /_/

Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_151)
Type in expressions to have them evaluated.
```

```
Type :help for more information.
scala>
```