Case Study 5
Hospital Analysis

Dataset Description :

1. **DRG Definition:** The code and description identifying the MS-DRG. MS-DRGs are a classification system that groups similar clinical conditions (diagnoses) and procedures furnished by the hospital during their stay.

2. **Provider Id:** The CMS Certification Number (CCN) assigned to the Medicare-certified hospital facility.
3. **Provider Name:** The name of the provider.
4. **Provider Street Address:** The provider's street address.
5. **Provider City:** The city where the provider is located.
6. **Provider State:** The state where the provider is located.

7. **Provider Zip Code:** The provider's zip code.
8. **Provider HRR:** The Hospital Referral Region (HRR) where the provider is located.
9. **Total Discharges:** The number of discharges billed by the provider for inpatient hospital services.
10. **Average Covered Charges:** The provider's average charge for services covered by Medicare for all discharges in the MS-DRG. These will vary from hospital to hospital because of the differences in hospital charge structures.

11. **Average Total Payments**: The average total payments to all providers for the MS-DRG including the MSDRG amount,teaching, disproportionate share, capital, and outlier payments for all cases. Also included in the average total payments are co-payment and deductible amounts that the patient is responsible for and any additional payments by third parties for coordination of benefits.

12. **Average Medicare Payments:** The average amount that Medicare pays to the provider for Medicare's share of the MS-DRG. Average Medicare payment amounts include the MS-DRG amount, teaching, disproportionate share,capital, and outlier payments for all cases. Medicare payments DO NOT include beneficiary co-payments and deductible amounts nor any additional payments from third parties for coordination of benefits.

Case Study 5
Hospital Analysis

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

In below program, we have created StructType object to provide schema for this csv file. StructType object defines the schema of Spark DataFrame.
It contains a list of StructField objects that define the name, type, and nullable flag for each column in a DataFrame. Then we have printed the output of this structtype.

Scala Code :

```
import org.apache.spark.sql.SparkSession

import org.apache.spark.sql.types._

object Case_Study_5_Hospital_Ananlysis {


val HospitalSchema = new StructType(Array(new StructField("DRGDefinition",
StringType,false),
    new StructField("ProviderId", LongType, false),
    new StructField("ProviderName", StringType, false),
    new StructField("ProviderStreetAddress", StringType, false),
    new StructField("ProviderCity", StringType, false),
    new StructField("ProviderState", StringType, false),
    new StructField("ProviderZipCode", LongType, false),
    new StructField("HospitalReferralRegionDescription", StringType, false),
    new StructField("TotalDischarges", LongType, false),
    new StructField("AverageCoveredCharges", DoubleType, false),
    new StructField("AverageTotalPayments", DoubleType, false),
    new StructField("AverageMedicarePayments", DoubleType, false)))

  HospitalSchema.printTreeString()
```

Output :

```
root
 |-- DRGDefinition: string (nullable = false)
 |-- ProviderId: long (nullable = false)
 |-- ProviderName: string (nullable = false)
 |-- ProviderStreetAddress: string (nullable = false)
 |-- ProviderCity: string (nullable = false)
 |-- ProviderState: string (nullable = false)
 |-- ProviderZipCode: long (nullable = false)
 |-- HospitalReferralRegionDescription: string (nullable = false)
 |-- TotalDischarges: long (nullable = false)
 |-- AverageCoveredCharges: double (nullable = false)
 |-- AverageTotalPayments: double (nullable = false)
 |-- AverageMedicarePayments: double (nullable = false)
```

Case Study 5
Hospital Analysis

In below program, we have created Spark object by using SparkSession.

Scala Code :

```
def main(args : Array[String]) : Unit = {


  val spark = SparkSession
    .builder()
    .master("local")
    .appName("Case Study 5 Hospital Analysis")
    .config("spark.some.config.option", "some-value")
    .getOrCreate()

  println("Spark object created")
```

Output :

Spark object created

Then we have loaded data from csv file and converted it to DataFrame.
We have taken the count of rows present in that csv file and created a temporary view as **Patient_charges**.

Scala Code :

```
import spark.implicits._

    // Below statement will suppress all warnings
    spark.sparkContext.setLogLevel("WARN")


    val patientCharges =     spark.read.format("csv")
      .option("header", "true")
      .schema(HospitalSchema)
      .load("C:\\AcadGild Hadoop\\Assignments\\inpatientCharges.csv").toDF()


    println("Hospital_data_analysis data-->"+patientCharges.count())

    patientCharges.createOrReplaceTempView("patient_charges")
```

Output :

Hospital_data_analysis data-->163065

Case Study 5
Hospital Analysis

Here we have used **sql** transformation to create sql query and taken average of AverageCoveredCharges by using group by clause for ProviderState column from **patient_charges** view and printed the result from this query.

Scala Code :

```
println("Below is the average amount of AverageCoveredCharges per state")

val averageChargesPerState = spark.sql("select
cast(avg(AverageCoveredCharges) as DECIMAL(12,2)) as
AverageofAverageCoveredChargesPerState,ProviderState from patient_charges group
by ProviderState")

averageChargesPerState.show()
```

Output :

Below is the average amount of AverageCoveredCharges per state

AverageofAverageCoveredChargesPerState	ProviderState
41200.06	AZ
35862.49	SC
33085.37	LA
27894.36	MN
66125.69	NJ
40116.66	DC
27390.11	OR
29222.00	VA
29942.70	RI
24523.81	KY
28700.60	WY
27059.02	NH
24124.25	MI
61047.12	NV
26149.33	WI
25565.55	ID
67508.62	CA
31318.41	CT
31736.43	NE
22670.02	MT

only showing top 20 rows

Case Study 5
Hospital Analysis

- **Find out the AverageTotalPayments charges per state.**

Here we have used **sql** transformation to create sql query and taken sum of AverageTotalPayments by using group by clause for ProviderState column from **patient_charges** view and printed the result from this query.

Scala Code :

```
println("Below is the AverageTotalPayments charges per state")

val averagePaymentsPerState = spark.sql("select
cast(sum(AverageTotalPayments) as DECIMAL(14,2)) as
AverageTotalPaymentsPerState,ProviderState from patient_charges group by
ProviderState")

averagePaymentsPerState.show()
```

Output :

```
Below is the AverageTotalPayments charges per state
+--------------------------+-------------+
|AverageTotalPaymentsPerState|ProviderState|
+--------------------------+-------------+
|              28950559.93|           AZ|
|              26000001.90|           SC|
|              26149231.62|           LA|
|              22403429.64|           MN|
|              51536799.21|           NJ|
|               6005089.59|           DC|
|              13556614.53|           OR|
|              38501742.43|           VA|
|               6179625.31|           RI|
|              26731563.38|           KY|
|               2815426.02|           WY|
|               7645391.68|           NH|
|              52859204.18|           MI|
|              12370645.07|           NV|
|              26273179.72|           WI|
|               5414776.23|           ID|
|             164993988.92|           CA|
|              22855921.30|           CT|
|               9910246.84|           NE|
|               4681918.20|           MT|
+--------------------------+-------------+
only showing top 20 rows
```

Case Study 5
Hospital Analysis

- **Find out the AverageMedicarePayments charges per state.**

Below we have used **sql** transformation to create sql query and taken sum of AverageMedicarePayments by using group by clause for ProviderState column from **patient_charges** view and printed the result from this query.

Scala Code :

```
println("Below is the AverageMedicarePayments charges per state")

val averageMedicarePaymentsPerState = spark.sql("select
cast(sum(AverageMedicarePayments) as DECIMAL(14,2))
AverageMedicarePaymentsPerState,ProviderState from patient_charges group by
ProviderState")

averageMedicarePaymentsPerState.show()
```

Output :

```
Below is the AverageMedicarePayments charges per state
+-----------------------------+-------------+
|AverageMedicarePaymentsPerState|ProviderState|
+-----------------------------+-------------+
|                  25162119.85|           AZ|
|                  22423915.85|           SC|
|                  22362581.90|           LA|
|                  19410472.14|           MN|
|                  46266572.71|           NJ|
|                   5457129.08|           DC|
|                  11736802.69|           OR|
|                  32658285.23|           VA|
|                   5478948.20|           RI|
|                  23201100.60|           KY|
|                   2356229.83|           WY|
|                   6686469.14|           NH|
|                  46940232.88|           MI|
|                  10514618.60|           NV|
|                  22679362.48|           WI|
|                   4662549.61|           ID|
|                 150162602.24|           CA|
|                  20320336.41|           CT|
|                   8488170.14|           NE|
|                   4038430.56|           MT|
+-----------------------------+-------------+
only showing top 20 rows
```

Case Study 5
Hospital Analysis

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

Below we have used **sql** transformation to create sql query and taken sum of TotalDischarges by using group by clause for ProviderState and DRGDefinition columns from **patient_charges** view. Then we have sorted this output in the descending order of totalDischarges column and printed the result from this query.

Scala Code :

```
   println("Below is the total number of Discharges per state and for each
disease")

    val DischargesPerStatePerDisease = spark.sql("select
ProviderState,DRGDefinition, sum(TotalDischarges) as DischargesPerStatePerDisease
from patient_charges group by ProviderState,DRGDefinition")

    DischargesPerStatePerDisease.show()
```

Output :

Below is the total number of Discharges per state and for each disease

ProviderState	DRGDefinition	DischargesPerStatePerDisease
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

Case Study 5
Hospital Analysis

- **Sort the output in descending order of totalDischarges.**

  Below we have sorted the result in the descending order of totalDischarges column.

Scala Code :

```
println("Below is the output sorted in the descending order of totalDischarges
")

val TotalDischargesDesc = spark.sql("select ProviderState,DRGDefinition,
sum(TotalDischarges) as DischargesPerStatePerDisease   from patient_charges group
by ProviderState,DRGDefinition order by DischargesPerStatePerDisease desc")

TotalDischargesDesc.show()
```

Output :

Below is the total output sorted in the descending order of totalDischarges

```
+-------------+--------------------+--------------------------+
|ProviderState|      DRGDefinition|DischargesPerStatePerDisease|
+-------------+--------------------+--------------------------+
|           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
```

Case Study 5
Hospital Analysis

Complete Scala Program:

```
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.types._

object Case_Study_5_Hospital_Ananlysis {


  val HospitalSchema = new StructType(Array(new StructField("DRGDefinition",
StringType,false),
    new StructField("ProviderId", LongType, false),
    new StructField("ProviderName", StringType, false),
    new StructField("ProviderStreetAddress", StringType, false),
    new StructField("ProviderCity", StringType, false),
    new StructField("ProviderState", StringType, false),
    new StructField("ProviderZipCode", LongType, false),
    new StructField("HospitalReferralRegionDescription", StringType, false),
    new StructField("TotalDischarges", LongType, false),
    new StructField("AverageCoveredCharges", DoubleType, false),
    new StructField("AverageTotalPayments", DoubleType, false),
    new StructField("AverageMedicarePayments", DoubleType, false)))

  HospitalSchema.printTreeString()

  def main(args : Array[String]) : Unit = {


    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Case Study 5 Hospital Analysis")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()

    println("Spark object created")

    import spark.implicits._

    // Below statement will suppress all warnings
    spark.sparkContext.setLogLevel("WARN")



    val patientCharges =     spark.read.format("csv")
      .option("header", "true")
      .schema(HospitalSchema)
      .load("C:\\AcadGild Hadoop\\Assignments\\inpatientCharges.csv").toDF()


    println("Hospital_data_analysis data-->"+patientCharges.count())

    patientCharges.createOrReplaceTempView("patient_charges")

    val patient_charges = spark.sql("select * from patient_charges ")


    println("Below is the average amount of AverageCoveredCharges per state")

    val averageChargesPerState = spark.sql("select  cast(avg(AverageCoveredCharges)
as DECIMAL(12,2)) as AverageofAverageCoveredChargesPerState,ProviderState from
patient_charges group by ProviderState")

  averageChargesPerState.show()


    println("Below is the AverageTotalPayments charges per state")
```

Case Study 5
Hospital Analysis

```
    val averagePaymentsPerState = spark.sql("select cast(sum(AverageTotalPayments)
as DECIMAL(14,2)) as AverageTotalPaymentsPerState,ProviderState from
patient_charges group by ProviderState")

   averagePaymentsPerState.show()


    println("Below is the AverageMedicarePayments charges per state")

    val averageMedicarePaymentsPerState = spark.sql("select
cast(sum(AverageMedicarePayments) as DECIMAL(14,2))
AverageMedicarePaymentsPerState,ProviderState from patient_charges group by
ProviderState")

   averageMedicarePaymentsPerState.show()


    println("Below is the total number of Discharges per state and for each
disease")

    val DischargesPerStatePerDisease = spark.sql("select
ProviderState,DRGDefinition, sum(TotalDischarges) as DischargesPerStatePerDisease
from patient_charges group by ProviderState,DRGDefinition")

    DischargesPerStatePerDisease.show()


    println("Below is the output sorted in the descending order of
totalDischarges")

    val TotalDischargesDesc = spark.sql("select ProviderState,DRGDefinition,
sum(TotalDischarges) as DischargesPerStatePerDisease   from patient_charges group
by ProviderState,DRGDefinition order by DischargesPerStatePerDisease desc")

    TotalDischargesDesc.show()

}
}
```

Case Study 5
Hospital Analysis

Complete Output:

```
root
 |-- DRGDefinition: string (nullable = false)
 |-- ProviderId: long (nullable = false)
 |-- ProviderName: string (nullable = false)
 |-- ProviderStreetAddress: string (nullable = false)
 |-- ProviderCity: string (nullable = false)
 |-- ProviderState: string (nullable = false)
 |-- ProviderZipCode: long (nullable = false)
 |-- HospitalReferralRegionDescription: string (nullable = false)
 |-- TotalDischarges: long (nullable = false)
 |-- AverageCoveredCharges: double (nullable = false)
 |-- AverageTotalPayments: double (nullable = false)
 |-- AverageMedicarePayments: double (nullable = false)

Spark object created
Hospital_data_analysis data-->163065
Below is the average amount of AverageCoveredCharges per state
+-----------------------------------+-------------+
|AverageofAverageCoveredChargesPerState|ProviderState|
+-----------------------------------+-------------+
|                           41200.06|           AZ|
|                           35862.49|           SC|
|                           33085.37|           LA|
|                           27894.36|           MN|
|                           66125.69|           NJ|
|                           40116.66|           DC|
|                           27390.11|           OR|
|                           29222.00|           VA|
|                           29942.70|           RI|
|                           24523.81|           KY|
|                           28700.60|           WY|
|                           27059.02|           NH|
|                           24124.25|           MI|
|                           61047.12|           NV|
|                           26149.33|           WI|
|                           25565.55|           ID|
|                           67508.62|           CA|
|                           31318.41|           CT|
|                           31736.43|           NE|
|                           22670.02|           MT|
+-----------------------------------+-------------+
only showing top 20 rows
```

Case Study 5
Hospital Analysis

Below is the AverageTotalPayments charges per state

```
+---------------------------+-------------+
|AverageTotalPaymentsPerState|ProviderState|
+---------------------------+-------------+
|                28950559.93|           AZ|
|                26000001.90|           SC|
|                26149231.62|           LA|
|                22403429.64|           MN|
|                51536799.21|           NJ|
|                 6005089.59|           DC|
|                13556614.53|           OR|
|                38501742.43|           VA|
|                 6179625.31|           RI|
|                26731563.38|           KY|
|                 2815426.02|           WY|
|                 7645391.68|           NH|
|                52859204.18|           MI|
|                12370645.07|           NV|
|                26273179.72|           WI|
|                 5414776.23|           ID|
|               164993988.92|           CA|
|                22855921.30|           CT|
|                 9910246.84|           NE|
|                 4681918.20|           MT|
+---------------------------+-------------+
only showing top 20 rows
```

Below is the AverageMedicarePayments charges per state

```
+-----------------------------+-------------+
|AverageMedicarePaymentsPerState|ProviderState|
+-----------------------------+-------------+
|                  25162119.85|           AZ|
|                  22423915.85|           SC|
|                  22362581.90|           LA|
|                  19410472.14|           MN|
|                  46266572.71|           NJ|
|                   5457129.08|           DC|
|                  11736802.69|           OR|
|                  32658285.23|           VA|
|                   5478948.20|           RI|
|                  23201100.60|           KY|
|                   2356229.83|           WY|
|                   6686469.14|           NH|
|                  46940232.88|           MI|
|                  10514618.60|           NV|
|                  22679362.48|           WI|
|                   4662549.61|           ID|
|                 150162602.24|           CA|
|                  20320336.41|           CT|
```

Case Study 5
Hospital Analysis

```
|              8488170.14|        NE|
|              4038430.56|        MT|
+------------------------------+----------+
only showing top 20 rows
```

Below is the total number of Discharges per state and for each disease

```
+-------------+--------------------+--------------------------+
|ProviderState|      DRGDefinition|DischargesPerStatePerDisease|
+-------------+--------------------+--------------------------+
|           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
```

Case Study 5
Hospital Analysis

Below is the output sorted in the descending order of totalDischarges

```
+-------------+--------------------+--------------------------+
|ProviderState|       DRGDefinition|DischargesPerStatePerDisease|
+-------------+--------------------+--------------------------+
|           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