Assignment21:

Task 1 Using spark-sql, Find:

1. What are the total number of gold medal winners every year

Query used:

```
//To get the number of Gold Medal winners every year.
val GoldMedals = spark.sql("select year, MedalType, count(MedalType) from MedalsList
group by year, MedalType having MedalType = 'gold'")
GoldMedals.show()
```

Output:

```
+----+---------+----------------+
|year|MedalType|count(MedalType)|
+----+---------+----------------+
|2016|     gold|               2|
|2014|     gold|               3|
|2015|     gold|               3|
|2017|     gold|               1|
+----+---------+----------------+
```

2. How many silver medals have been won by USA in each sport Task

Query used:

```
//To get the number of Gold Medal winners every year.
val SilverMedals = spark.sql("select Sports, country, MedalType, count(MedalType) from
MedalsList group by Sports, country, MedalType having MedalType = 'silver' and country
= 'USA'")
SilverMedals.show()
```

Output:

```
+--------+-------+---------+----------------+
|  Sports|country|MedalType|count(MedalType)|
+--------+-------+---------+----------------+
|swimming|    USA|   silver|               3|
+--------+-------+---------+----------------+
```

Task 2:

Using udfs on dataframe 1. Change firstname, lastname columns into Mr.first_two_letters_of_firstnamelastname for example - michael, phelps becomes Mr.mi phelps

Query used:

```
//udf to add a new column containing MR.+ 2charactersof first name + last name
def changefname = org.apache.spark.sql.functions.udf((value: String, value2: String)
=> {
 val newsubstrwith2char = value.substring(0, 2)
 var newsubstrMr: String = "Mr."
 var newsubstr1: String = ""
 newsubstr1 = newsubstrMr + newsubstrwith2char + " " + value2
 newsubstr1
})
val newname = data1.withColumn("NewName", changefname($"fname",$"lname"))
newname.show(20)
val droppedfield = newname.drop("fname","lname")
val finaloutput =
droppedfield.select("NewName","Sports","MedalType","Age","Year","country")
finaloutput.show(20)
```

Output:

```
16/05/28 22:30:37 INFO SparkContext: Invoking stop() from shutdown hoo
|       NewName|  Sports| MedalType|Age|Year|country|
+--------------+--------+----------+---+----+-------+
|Mr.f lastname|  sports|medal_type|age|year|country|
|  Mr.li cudrow|javellin|      gold| 34|2015|    USA|
|   Mr.ma louis|javellin|      gold| 34|2015|    RUS|
|  Mr.mi phelps|swimming|    silver| 32|2016|    USA|
|      Mr.us pt| running|    silver| 30|2016|    IND|
|Mr.se williams| running|      gold| 31|2014|    FRA|
| Mr.ro federer|  tennis|    silver| 32|2016|    CHN|
|     Mr.je cox|swimming|    silver| 32|2014|    IND|
| Mr.fe johnson|swimming|    silver| 32|2016|    CHN|
|  Mr.li cudrow|javellin|      gold| 34|2017|    USA|
|   Mr.ma louis|javellin|      gold| 34|2015|    RUS|
|  Mr.mi phelps|swimming|    silver| 32|2017|    USA|
|      Mr.us pt| running|    silver| 30|2014|    IND|
|Mr.se williams| running|      gold| 31|2016|    FRA|
| Mr.ro federer|  tennis|    silver| 32|2017|    CHN|
|     Mr.je cox|swimming|    silver| 32|2014|    IND|
| Mr.fe johnson|swimming|    silver| 32|2017|    CHN|
|  Mr.li cudrow|javellin|      gold| 34|2014|    USA|
|   Mr.ma louis|javellin|      gold| 34|2014|    RUS|
|  Mr.mi phelps|swimming|    silver| 32|2017|    USA|
+--------------+--------+----------+---+----+-------+
only showing top 20 rows
```

2. Add a new column called ranking using udfs on dataframe, where : gold medalist, with age >= 32 are ranked as pro gold medalists, with age <= 31 are ranked amateur silver medalist, with age >= 32 are ranked as expert silver medalists, with age <= 31 are ranked rookie

Query used:

```
//udf to Add a new column called ranking using udfs on dataframe
def NewRankColumn = org.apache.spark.sql.functions.udf((medaltype: String, age: Long)
=> {
 var newsubstr:String = ""
 if(medaltype == "gold" && age >= 32)
   newsubstr +=  "PROF GOLD MEDALIST"
 if(medaltype == "gold" && age <= 31)
   newsubstr += "AMATEUR SILVER MEDALIST"
 if(medaltype == "silver" && age >= 32)
   newsubstr += "EXPERT SILVER MEDALIST"
```

```
 if(medaltype == "silver" && age <= 32)
    newsubstr += "Rookie"
 newsubstr
})
```

Output:

```
|   fname|   lname|  Sports|MedalType|Age|Year|country|             Ranking|
+--------+--------+--------+---------+---+----+-------+--------------------+
|    lisa|  cudrow|javelin|     gold| 34|2015|    USA|  PROF GOLD MEDALIST|
|  mathew|   louis|javelin|     gold| 34|2015|    RUS|  PROF GOLD MEDALIST|
| michael|  phelps|swimming|   silver| 32|2016|    USA|EXPERT SILVER MED...|
|    usha|      pt| running|   silver| 30|2016|    IND|              Rookie|
|  serena|williams| running|     gold| 31|2014|    FRA|AMATEUR SILVER ME...|
|   roger| federer|  tennis|   silver| 32|2016|    CHN|EXPERT SILVER MED...|
| jenifer|     cox|swimming|   silver| 32|2014|    IND|EXPERT SILVER MED...|
|fernando| johnson|swimming|   silver| 32|2016|    CHN|EXPERT SILVER MED...|
|    lisa|  cudrow|javelin|     gold| 34|2017|    USA|  PROF GOLD MEDALIST|
|  mathew|   louis|javelin|     gold| 34|2015|    RUS|  PROF GOLD MEDALIST|
| michael|  phelps|swimming|   silver| 32|2017|    USA|EXPERT SILVER MED...|
|    usha|      pt| running|   silver| 30|2014|    IND|              Rookie|
|  serena|williams| running|     gold| 31|2016|    FRA|AMATEUR SILVER ME...|
|   roger| federer|  tennis|   silver| 32|2017|    CHN|EXPERT SILVER MED...|
| jenifer|     cox|swimming|   silver| 32|2014|    IND|EXPERT SILVER MED...|
+--------+--------+--------+---------+---+----+-------+--------------------+
only showing top 15 rows
```