Assignment 21.1

Task 1
Using spark-sql, Find:
1. What are the total number of gold medal winners every year
2. How many silver medals have been won by USA in each sport

Task 2
Using udfs on dataframe
1. Change firstname, lastname columns into
Mr.first_two_letters_of_firstname<space>lastname
for example - michael, phelps becomes Mr.mi phelps
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

Solution

The output of the complete program is given below. Complete source code with comments has been pasted later in the document.

TASK1
Ans1. What are the total number of gold medal winners every year

```
+----+----------+
|year|gold_count|
+----+----------+
|2015|         3|
|2014|         3|
|2016|         2|
|2017|         1|
+----+----------+
```

Ans2. How many silver medals have been won by USA in each sport

```
+--------+------------+
|  sports|silver_count|
+--------+------------+
|swimming|           3|
+--------+------------+
```

Task2

```
+--------------+--------+----------+---+----+-------+
|         name|  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|
|     Mr.us pt| running|    silver| 30|2014|    IND|
+--------------+--------+----------+---+----+-------+
only showing top 20 rows

+---------+--------+--------+----------+---+----+-------+-------+
|firstname|lastname|  sports|medal_type|age|year|country|Ranking|
+---------+--------+--------+----------+---+----+-------+-------+
|     lisa|  cudrow|javellin|      gold| 34|2015|    USA|    pro|
|   mathew|   louis|javellin|      gold| 34|2015|    RUS|    pro|
|  michael|  phelps|swimming|    silver| 32|2016|    USA| expert|
|     usha|     pt| running|    silver| 30|2016|    IND| rookie|
|   serena|williams| running|      gold| 31|2014|    FRA|amateur|
|    roger|  federer|  tennis|    silver| 32|2016|    CHN| expert|
|  jenifer|     cox|swimming|    silver| 32|2014|    IND| expert|
| fernando|  johnson|swimming|    silver| 32|2016|    CHN| expert|
|     lisa|  cudrow|javellin|      gold| 34|2017|    USA|    pro|
|   mathew|   louis|javellin|      gold| 34|2015|    RUS|    pro|
|  michael|  phelps|swimming|    silver| 32|2017|    USA| expert|
|     usha|     pt| running|    silver| 30|2014|    IND| rookie|
|   serena|williams| running|      gold| 31|2016|    FRA|amateur|
|    roger|  federer|  tennis|    silver| 32|2017|    CHN| expert|
|  jenifer|     cox|swimming|    silver| 32|2014|    IND| expert|
| fernando|  johnson|swimming|    silver| 32|2017|    CHN| expert|
```

```
|    lisa| cudrow|javellin|     gold| 34|2014|    USA|    pro|
|  mathew|  louis|javellin|     gold| 34|2014|    RUS|    pro|
| michael| phelps|swimming|   silver| 32|2017|    USA| expert|
|    usha|     pt| running|   silver| 30|2014|    IND| rookie|
+--------+-------+--------+---------+---+----+-------+-------+
only showing top 20 rows
```

Source Code

```
package demo

import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions.udf


object SparkSqlAssignment {

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

    val spark = SparkSession.builder().master(master="local")
      .appName(name="spark sql example")
      .config("spark.some.config.option","some-value").getOrCreate()
    spark.sparkContext.setLogLevel("WARN")
    // use this to create dataframes
    import spark.implicits._
    // create dataframes by reading the data set file
    // Since the file has headers, we set it to true
    // we will as spark to infer the schema for us
    val df = spark.read.format("csv").option("header","true")
      .option("inferSchema","true").option("mode","failfast")
      .load("/Users/mmisra/Desktop/acad/assignments/assignment_21.1/Sports_data.txt")

    println("data frame schema is below")
    df.printSchema()

    // create a temporary view
    val tv = df.createOrReplaceTempView("olympic")

    val b =spark.sql("select year,COUNT(medal_type)as gold_count from olympic where
medal_type='gold' group by year")
    println("Ans1. What are the total number of gold medal winners every year")
    b.show()

    println("Ans2. How many silver medals have been won by USA in each sport")
    val c =spark.sql("select sports,COUNT(medal_type) as silver_count from olympic
where medal_type='silver' AND country='USA' group by sports")
    c.show()


    spark.sqlContext.udf.register("ChangeName",ChangePlayerName(_:String,_:String))
    val d =spark.sql("select ChangeName(firstname,lastname)as name, sports,
medal_type, age, year, country from olympic ")
    d.show()

    spark.sqlContext.udf.register("RankingUDF",Ranking(_:String,_:Int))
    val e =spark.sql("select *, RankingUDF(medal_type,age) as Ranking from olympic ")
    e.show()

  }
```

```
  // UDF for changing the name of the player
  def ChangePlayerName(first:String,last:String):String=
  {

    //1. Change firstname, lastname columns into
    //Mr.first_two_letters_of_firstname<space>lastname
    //for example - michael, phelps becomes Mr.mi phelps
    // register the UDF
    return "Mr."+first.charAt(0)+first.charAt(1) +" " + last
  }

  def Ranking(medal_type:String,age:Int):String=
  {

    //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

    if(medal_type=="gold")
      {
       if(age>=32)
         return "pro"
       else return "amateur"
      }

    if(medal_type=="silver") {
      if (age >= 32)
        return "expert"
      else return "rookie"
    }

    "Unknown"
  }

}
```