Assignment19.2

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.

```
scala> val rdd=sc.textFile("Sports_data.txt")
rdd: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[1] at textFile at
<console>:27

scala> val head=rdd.first()
head: String = firstname,lastname,sports,medal_type,age,year,country

scala> val filter_set=rdd.filter(x=>x!=head)
filter_set: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[2] at filter at
<console>:31

scala> val sports=filter_set.map(x=>x.split(",")).map(arrays
=>(arrays(0),arrays(1),arrays(2),arrays(3),arrays(4),arrays(5),arrays(6))).toDF("fir
stname","lastname","sports","medal_type","age","year","country")
sports: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string,
sports: string, medal_type: string, age: string, year: string, country: string]

scala> sports.registerTempTable("sportstable")
```

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

probelem1

scala> val name=udf((fname:String,lname:String)=>{"Mr."+fname.slice(0,2)+" "+lname+""})
name: org.apache.spark.sql.UserDefinedFunction = UserDefinedFunction(<function2>,StringType,List(StringType, StringType))

scala> val first=sports.withColumn("Fullname",name($"firstname",$"lastname"))
first: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string, sports: string, medal_type: string, age: string, year: string, country: string, Fullname: string]

scala> val result=first.drop("firstname").drop("lastname")
result: org.apache.spark.sql.DataFrame = [sports: string, medal_type: string, age:

string, year: string, country: string, Fullname: string]

scala> first.show

```
+---------+--------+--------+----------+---+----+-------+-------------+
|firstname|lastname|  sports|medal_type|age|year|country|     Fullname|
+---------+--------+--------+----------+---+----+-------+-------------+
|     lisa|  cudrow|javellin|      gold| 34|2015|    USA|  Mr.li cudrow|
|   mathew|   louis|javellin|      gold| 34|2015|    RUS|   Mr.ma louis|
|  michael|  phelps|swimming|    silver| 32|2016|    USA|  Mr.mi phelps|
|     usha|      pt| running|    silver| 30|2016|    IND|      Mr.us pt|
|   serena|williams| running|      gold| 31|2014|    FRA|Mr.se williams|
|    roger| federer|  tennis|    silver| 32|2016|    CHN| Mr.ro federer|
|  jenifer|     cox|swimming|    silver| 32|2014|    IND|     Mr.je cox|
| fernando| johnson|swimming|    silver| 32|2016|    CHN| Mr.fe johnson|
|     lisa|  cudrow|javellin|      gold| 34|2017|    USA|  Mr.li cudrow|
|   mathew|   louis|javellin|      gold| 34|2015|    RUS|   Mr.ma louis|
|  michael|  phelps|swimming|    silver| 32|2017|    USA|  Mr.mi phelps|
|     usha|      pt| running|    silver| 30|2014|    IND|      Mr.us pt|
|   serena|williams| running|      gold| 31|2016|    FRA|Mr.se williams|
|    roger| federer|  tennis|    silver| 32|2017|    CHN| Mr.ro federer|
|  jenifer|     cox|swimming|    silver| 32|2014|    IND|     Mr.je cox|
| fernando| johnson|swimming|    silver| 32|2017|    CHN| Mr.fe johnson|
|     lisa|  cudrow|javellin|      gold| 34|2014|    USA|  Mr.li cudrow|
|   mathew|   louis|javellin|      gold| 34|2014|    RUS|   Mr.ma louis|
|  michael|  phelps|swimming|    silver| 32|2017|    USA|  Mr.mi phelps|
|     usha|      pt| running|    silver| 30|2014|    IND|      Mr.us pt|
+---------+--------+--------+----------+---+----+-------+-------------+
only showing top 20 rows
```

```
scala> first.show
+---------+--------+--------+----------+---+----+-------+-------------+
|firstname|lastname|  sports|medal_type|age|year|country|     Fullname|
+---------+--------+--------+----------+---+----+-------+-------------+
|     lisa|  cudrow|javellin|      gold| 34|2015|    USA|  Mr.li cudrow|
|   mathew|   louis|javellin|      gold| 34|2015|    RUS|   Mr.ma louis|
|  michael|  phelps|swimming|    silver| 32|2016|    USA|  Mr.mi phelps|
|     usha|      pt| running|    silver| 30|2016|    IND|      Mr.us pt|
|   serena|williams| running|      gold| 31|2014|    FRA|Mr.se williams|
|    roger| federer|  tennis|    silver| 32|2016|    CHN| Mr.ro federer|
|  jenifer|     cox|swimming|    silver| 32|2014|    IND|     Mr.je cox|
| fernando| johnson|swimming|    silver| 32|2016|    CHN| Mr.fe johnson|
|     lisa|  cudrow|javellin|      gold| 34|2017|    USA|  Mr.li cudrow|
|   mathew|   louis|javellin|      gold| 34|2015|    RUS|   Mr.ma louis|
|  michael|  phelps|swimming|    silver| 32|2017|    USA|  Mr.mi phelps|
|     usha|      pt| running|    silver| 30|2014|    IND|      Mr.us pt|
|   serena|williams| running|      gold| 31|2016|    FRA|Mr.se williams|
|    roger| federer|  tennis|    silver| 32|2017|    CHN| Mr.ro federer|
|  jenifer|     cox|swimming|    silver| 32|2014|    IND|     Mr.je cox|
| fernando| johnson|swimming|    silver| 32|2017|    CHN| Mr.fe johnson|
|     lisa|  cudrow|javellin|      gold| 34|2014|    USA|  Mr.li cudrow|
|   mathew|   louis|javellin|      gold| 34|2014|    RUS|   Mr.ma louis|
|  michael|  phelps|swimming|    silver| 32|2017|    USA|  Mr.mi phelps|
|     usha|      pt| running|    silver| 30|2014|    IND|      Mr.us pt|
+---------+--------+--------+----------+---+----+-------+-------------+
```

```
scala> result.show
+--------+----------+---+----+-------+-------------+
|  sports|medal_type|age|year|country|     Fullname|
+--------+----------+---+----+-------+-------------+
|javellin|      gold| 34|2015|    USA|  Mr.li cudrow|
|javellin|      gold| 34|2015|    RUS|   Mr.ma louis|
|swimming|    silver| 32|2016|    USA|  Mr.mi phelps|
| running|    silver| 30|2016|    IND|      Mr.us pt|
| running|      gold| 31|2014|    FRA|Mr.se williams|
|  tennis|    silver| 32|2016|    CHN| Mr.ro federer|
|swimming|    silver| 32|2014|    IND|     Mr.je cox|
|swimming|    silver| 32|2016|    CHN| Mr.fe johnson|
|javellin|      gold| 34|2017|    USA|  Mr.li cudrow|
|javellin|      gold| 34|2015|    RUS|   Mr.ma louis|
|swimming|    silver| 32|2017|    USA|  Mr.mi phelps|
| running|    silver| 30|2014|    IND|      Mr.us pt|
| running|      gold| 31|2016|    FRA|Mr.se williams|
|  tennis|    silver| 32|2017|    CHN| Mr.ro federer|
|swimming|    silver| 32|2014|    IND|     Mr.je cox|
|swimming|    silver| 32|2017|    CHN| Mr.fe johnson|
|javellin|      gold| 34|2014|    USA|  Mr.li cudrow|
|javellin|      gold| 34|2014|    RUS|   Mr.ma louis|
|swimming|    silver| 32|2017|    USA|  Mr.mi phelps|
| running|    silver| 30|2014|    IND|      Mr.us pt|
+--------+----------+---+----+-------+-------------+
```

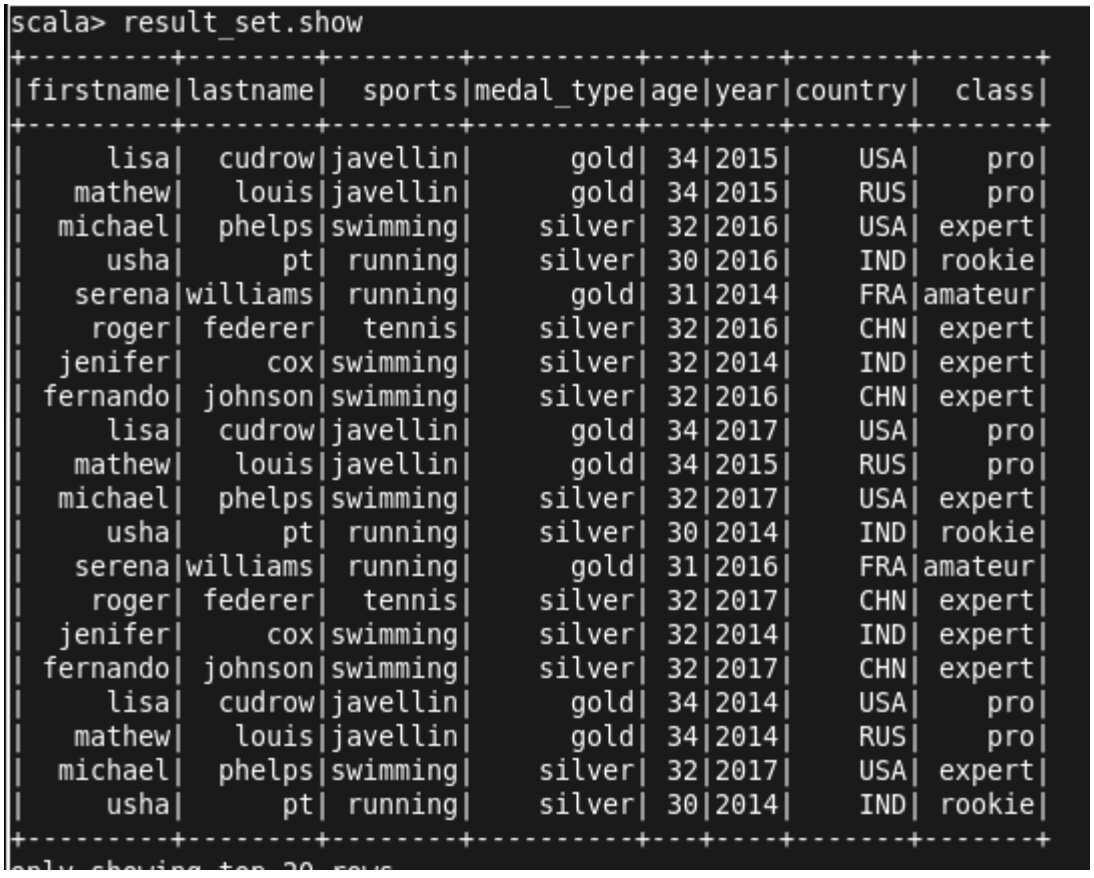
2. Add a new column called ranking using udfs on dataframe, where:
Gold medalist, with age >= 32 are ranked as pro
Gold medalists, with age <= 31 are ranked amateur
Silver medalist, with age >= 32 are ranked as expert
Silver medallists, with age <= 31 are ranked rookie.
Add new column into the created UDF.

```
scala> val
player=udf((medal:String,age:String)=>{if(age.toInt<=31&&medal=="silver")
{"rookie";}else{if(age.toInt>=32&&medal=="silver")
{"expert";}else{if(age.toInt<=31&&medal=="gold")
{"amateur";}else{if(age.toInt>=32&&medal=="gold"){"pro";}else{"could not
determine class";}}}}})
player: org.apache.spark.sql.UserDefinedFunction =
UserDefinedFunction(<function2>,StringType,List(StringType, StringType))

scala> val result_set=sports.withColumn("class",player($"medal_type",$"age"))
result_set: org.apache.spark.sql.DataFrame = [firstname: string, lastname: string,
sports: string, medal_type: string, age: string, year: string, country: string, class:
string]

scala> result_set.show
+---------+--------+--------+----------+---+----+-------+-------+
|firstname|lastname|  sports|medal_type|age|year|country|  class|
+---------+--------+--------+----------+---+----+-------+-------+
|     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



```
scala> result_set.show
+---------+--------+--------+----------+---+----+-------+-------+
|firstname|lastname|  sports|medal_type|age|year|country|  class|
+---------+--------+--------+----------+---+----+-------+-------+
|     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
```