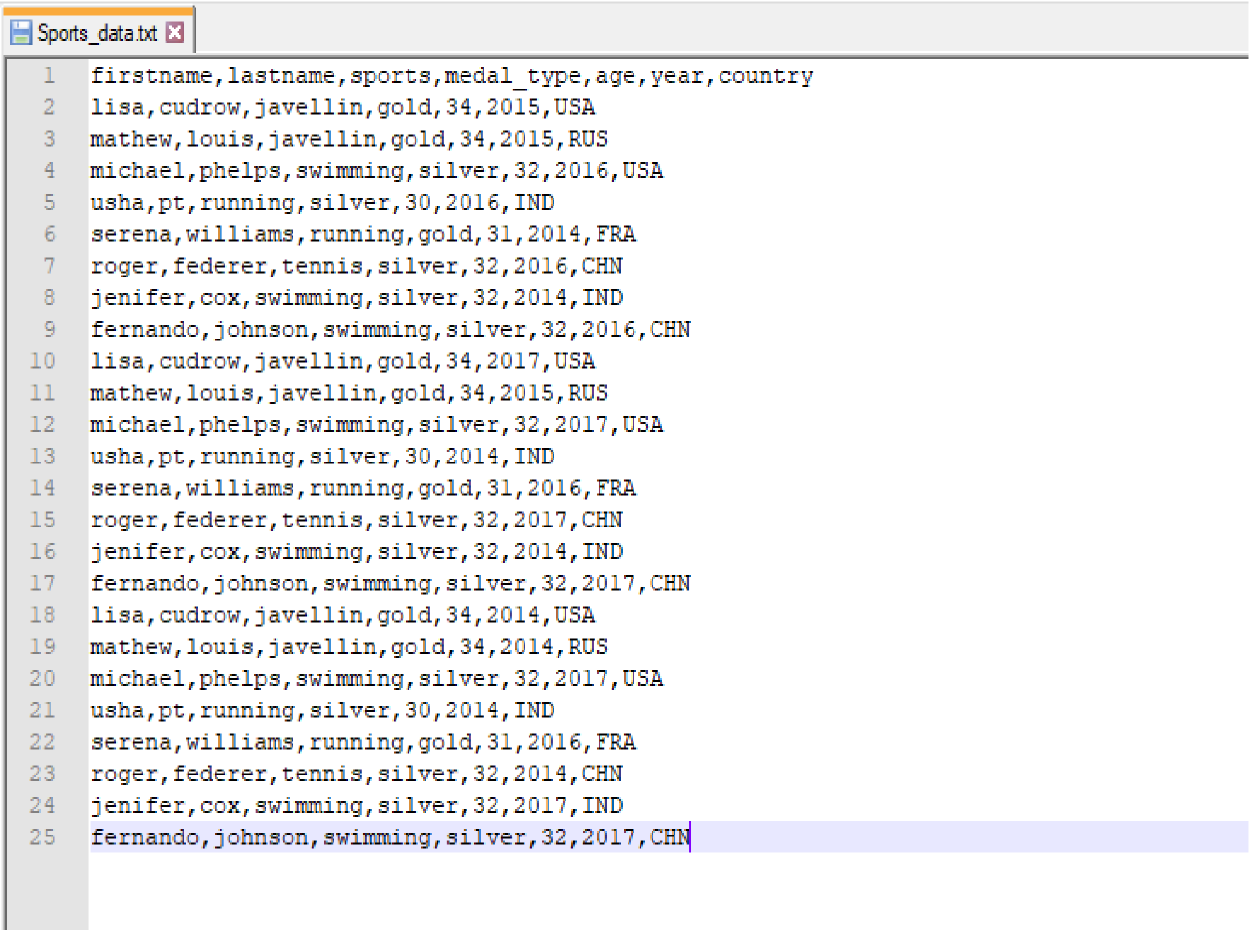
ASSIGNMENT 19.1

Input Dataset:

We have sports related data with us:

Sports_data

firstname, lastname, sports, medal_type, age, year, country



```
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
serena,williams,running,gold,31,2016,FRA
roger,federer,tennis,silver,32,2014,CHN
jenifer,cox,swimming,silver,32,2017,IND
fernando,johnson,swimming,silver,32,2017,CHN
```

Problem Statement:

Using spark-SQL, find:

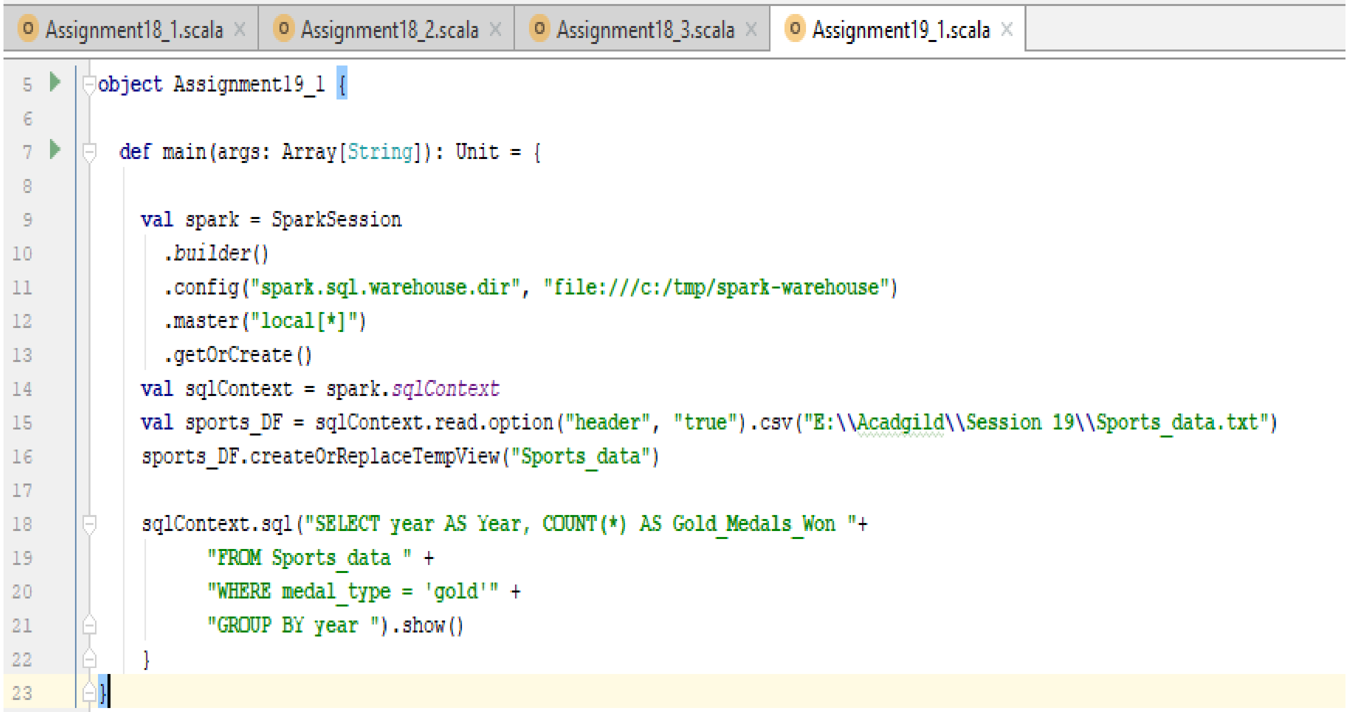
1. What are the total number of gold medal winners every year?

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

Solution:

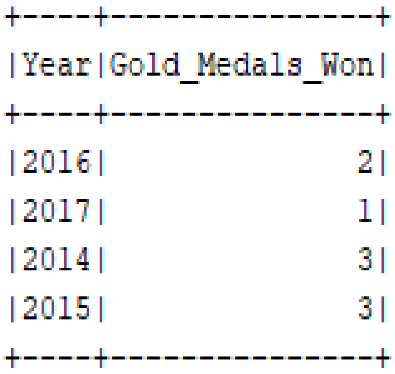
1. Here is the Spark code snippet to find the total number of gold medal winners every year:

```
// import required Spark packages

import org.apache.spark.sql.SparkSession

object Assignment18_1 {
  def main(args: Array[String]): Unit = {
// create a SparkSession object that can be used to create various contexts of Spark such as sqlContext
    val spark = SparkSession
      .builder()
      .config("spark.sql.warehouse.dir", "file:///c:/tmp/spark-warehouse")
      .master("local[*]")
      .getOrCreate()

// initialize sqlContext

    val sqlContext = spark.sqlContext

// load input data file – Sports_data.txt

    val sports_DF = sqlContext.read.option("header", "true").csv("E:\\Acadgild\\Session 19\\Sports_data.txt")

//create a temporary view – Sports_data

    sports_DF.createOrReplaceTempView("Sports_data")

// SQL query to find the total number of gold medal winners every year
    sqlContext.sql("SELECT year AS Year, COUNT(*) AS Gold_Medals_Won "+
          "FROM Sports_data " +
          "WHERE medal_type = 'gold' " +
          "GROUP BY year ").show()
  }
}
```



```
object Assignment19_1 {

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

    val spark = SparkSession
      .builder()
      .config("spark.sql.warehouse.dir", "file:///c:/tmp/spark-warehouse")
      .master("local[*]")
      .getOrCreate()
    val sqlContext = spark.sqlContext
    val sports_DF = sqlContext.read.option("header", "true").csv("E:\\Acadgild\\Session 19\\Sports_data.txt")
    sports_DF.createOrReplaceTempView("Sports_data")

    sqlContext.sql("SELECT year AS Year, COUNT(*) AS Gold_Medals_Won "+
          "FROM Sports_data " +
          "WHERE medal_type = 'gold'" +
          "GROUP BY year ").show()
    }
}
```

Output:



```
+----+---------------+
|Year|Gold_Medals_Won|
+----+---------------+
|2016|              2|
|2017|              1|
|2014|              3|
|2015|              3|
+----+---------------+
```

2. Here is the Spark code to find the total number of silver medals won by USA in each sport:

// import required Spark packages

import org.apache.spark.sql.SparkSession

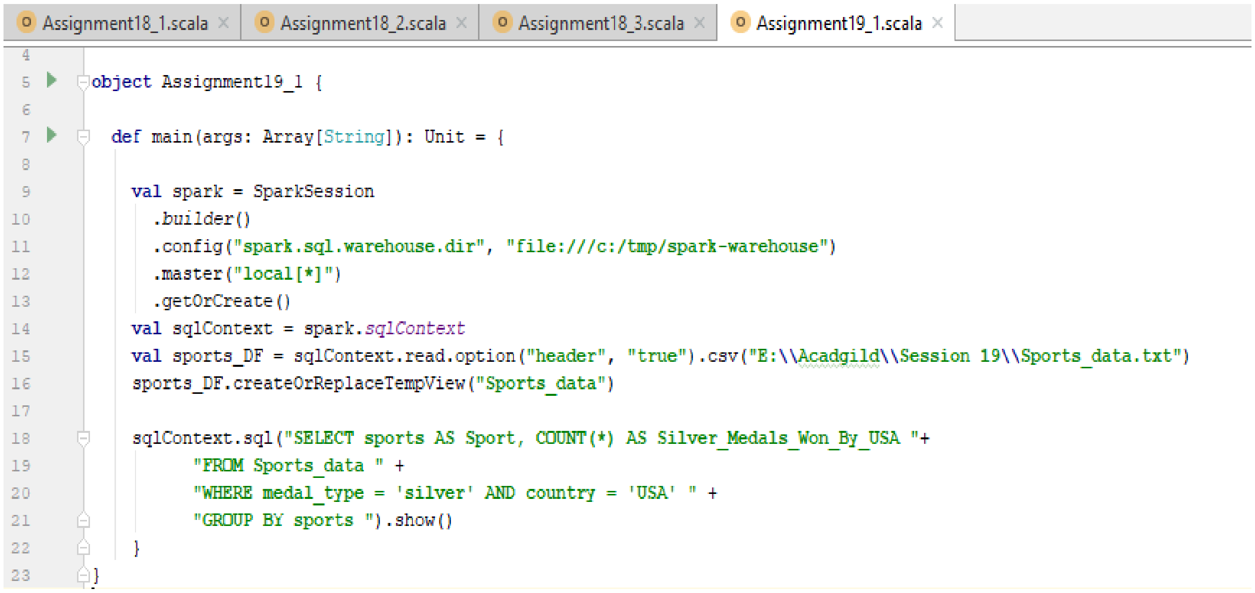
object Assignment18_1 {

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

// create a SparkSession object that can be used to create various contexts of Spark such as sqlContext

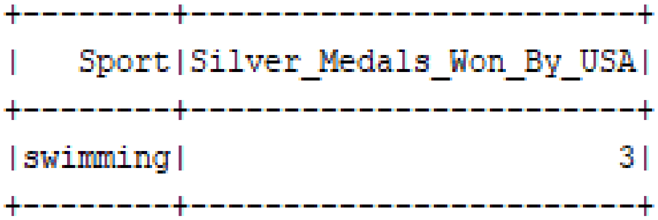
    val spark = SparkSession

```
      .builder()
      .config("spark.sql.warehouse.dir", "file:///c:/tmp/spark-warehouse")
      .master("local[*]")
      .getOrCreate()

// initialize sqlContext

    val sqlContext = spark.sqlContext

// load input data file – Sports_data.txt

    val sports_DF = sqlContext.read.option("header", "true").csv("E:\\Acadgild\\Session 19\\
Sports_data.txt")

//create a temporary view – Sports_data

    sports_DF.createOrReplaceTempView("Sports_data")

// SQL query to find the total number of gold medal winners every year
    sqlContext.sql("SELECT sports AS Sport, COUNT(*) AS Silver_Medals_Won_By_USA "+
          "FROM Sports_data " +
          "WHERE medal_type = 'silver' AND country = 'USA' " +
          "GROUP BY sports ").show()
  }
}
```



```

object Assignment19_1 {

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

    val spark = SparkSession
      .builder()
      .config("spark.sql.warehouse.dir", "file:///c:/tmp/spark-warehouse")
      .master("local[*]")
      .getOrCreate()
    val sqlContext = spark.sqlContext
    val sports_DF = sqlContext.read.option("header", "true").csv("E:\\Acadgild\\Session 19\\Sports_data.txt")
    sports_DF.createOrReplaceTempView("Sports_data")

    sqlContext.sql("SELECT sports AS Sport, COUNT(*) AS Silver_Medals_Won_By_USA "+
          "FROM Sports_data " +
          "WHERE medal_type = 'silver' AND country = 'USA' " +
          "GROUP BY sports ").show()
    }
}
```

Output:



```
+--------+------------------------+
|   Sport|Silver_Medals_Won_By_USA|
+--------+------------------------+
|swimming|                       3|
+--------+------------------------+
```

To summarize, USA has won **3** silver medals in swimming category.