Exercise 1 (Mandatory)

Submit two in-class exercises (four .scala files).

Exercise 2 (Mandatory)

Develop a script (from scratch or expand on given examples) and do the following steps:

1. Read insurance.csv file (uploaded in slack channel week6)
2. Print the size
3. Print sex and count of sex (use group by in sql)
4. Filter smoker=yes and print again the sex,count of sex
5. Group by region and sum the charges (in each region), then print rows by descending order (with respect to sum)

Exercise 3 (Mandatory)

Take the following script and fill up the missing codes:

```
package dataset

import org.apache.spark.sql.SparkSession

object CaseClass {

 case class Number(i: Int, english: String, french: String)

 def main(args: Array[String]) {
   val spark =
     SparkSession.builder()
       .appName("Dataset-CaseClass")
       .master("local[4]")
       .getOrCreate()

   import spark.implicits._

   val numbers = Seq(
     Number(1, "one", "un"),
     Number(2, "two", "deux"),
     Number(3, "three", "trois"))

   val numberDS=numbers.toDS()
```

```
  println("Dataset Types")
  // your code goes here

  println("filter dataset where i>1")
   // your code goes here

 println("select the number with English column and display")
   // your code goes here

 println("select the number with English column and filter for i>1")
   // your code goes here

  println("sparkSession dataset")
  val anotherDS=spark.createDataset(numbers)
  println("Spark Dataset Types")
   // your code goes here

 }
}
```

Exercise 4 (Optional)

Take the script, load data and do the steps. You can write additional exploitation queries to better understand the data.

https://github.com/apache/spark/blob/master/examples/src/main/scala/org/apache/spark/examples/sql/SparkSQLExample.scala