Problem Statement

1. **Create a dataframe with 1 to 100 and save as parquet file.**

In order to proceed first we need to create a RDD for numbers between 1 to 100.

Below screenshot shows the same-

```
scala> val nums = sc.parallelize(1 to 100)
nums: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[0] at parallelize at <console>:24

scala> nums
res0: org.apache.spark.rdd.RDD[Int] = ParallelCollectionRDD[0] at parallelize at <console>:24

scala> nums.collect()
res1: Array[Int] = Array(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 3
8, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67, 68, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79
, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 95, 96, 97, 98, 99, 100)

scala>
```

Now we will create the dataframe with above RDD as shown below-

```
scala> val numsDF = nums.toDF()
numsDF: org.apache.spark.sql.DataFrame = [value: int]

scala> numsDF.show()
+-----+
|value|
+-----+
|    1|
|    2|
|    3|
|    4|
|    5|
|    6|
|    7|
|    8|
|    9|
|   10|
|   11|
|   12|
|   13|
|   14|
|   15|
|   16|
|   17|
|   18|
|   19|
|   20|
+-----+
only showing top 20 rows
```

Now below is the final screenshot which shows that we are writing a parquet file from above defined datframe and then reading it-

```
scala> numsDF.write.parquet("/home/acadgild/Assignment-19/nums.parquet")

scala> val numsRead = spark.read.parquet("/home/acadgild/Assignment-19/nums.parquet")
numsRead: org.apache.spark.sql.DataFrame = [value: int]

scala> numsRead.show()
+-----+
|value|
+-----+
|   51|
|   52|
|   53|
|   54|
|   55|
|   56|
|   57|
|   58|
|   59|
|   60|
|   61|
|   62|
|   63|
|   64|
|   65|
|   66|
|   67|
|   68|
|   69|
|   70|
+-----+
```