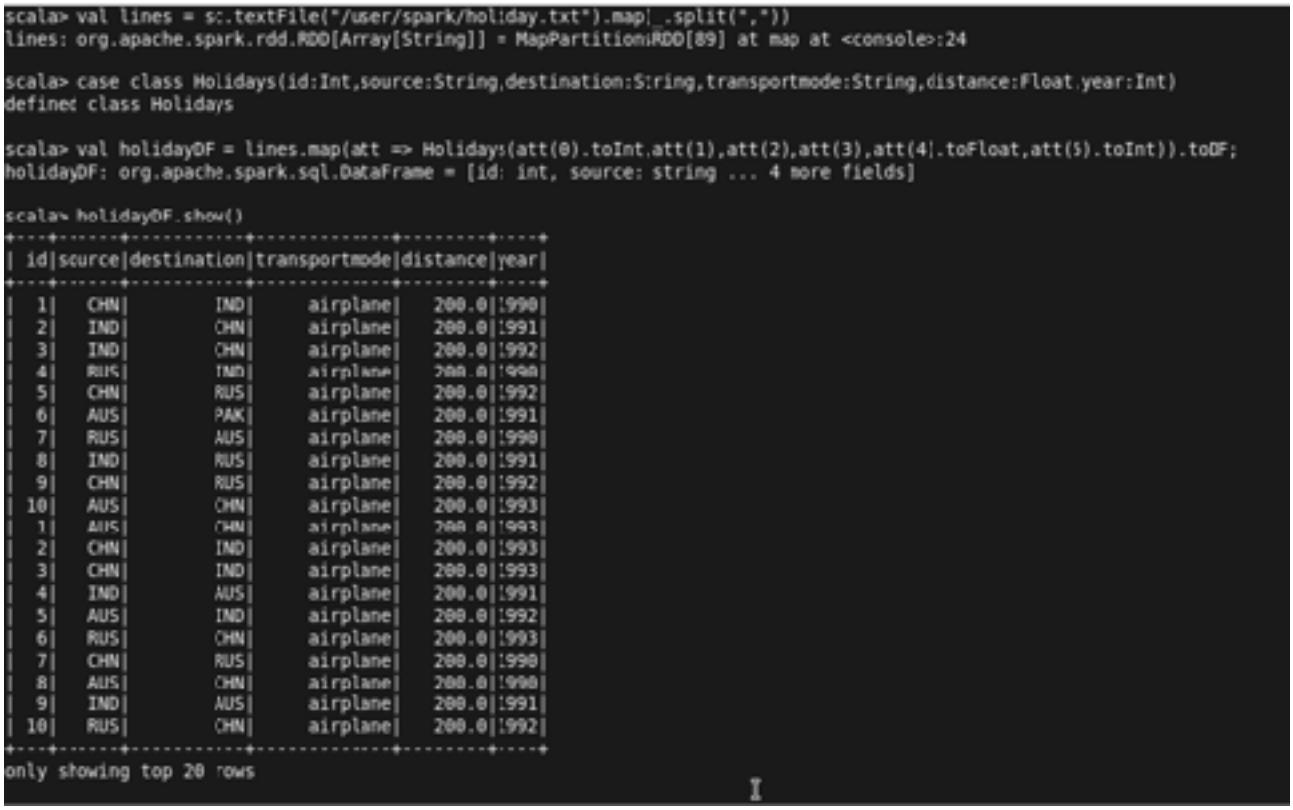
Task 1

1) What is the distribution of the total number of air-travelers per year

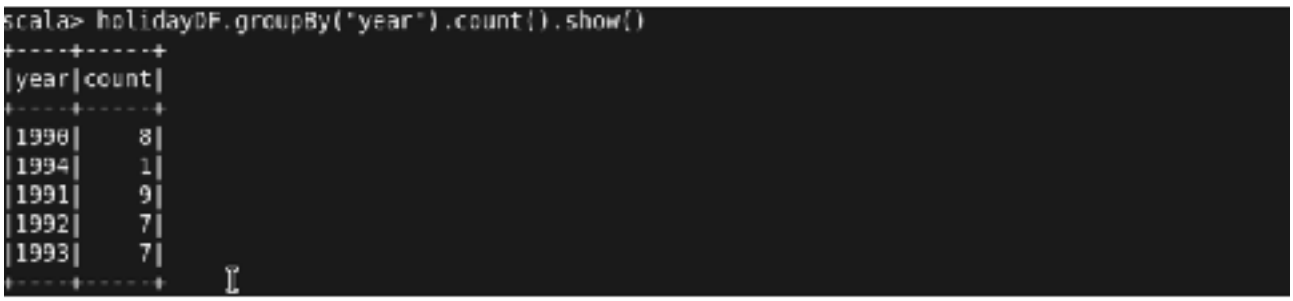


```
scala> val lines = sc.textFile("/user/spark/holiday.txt").map(_.split(","))
lines: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[89] at map at <console>:24

scala> case class Holidays(id:Int,source:String,destination:String,transportmode:String,distance:Float,year:Int)
defined class Holidays

scala> val holidayDF = lines.map(att => Holidays(att(0).toInt,att(1),att(2),att(3),att(4).toFloat,att(5).toInt)).toDF;
holidayDF: org.apache.spark.sql.DataFrame = [id: int, source: string ... 4 more fields]

scala> holidayDF.show()
+---+------+-----------+-------------+--------+----+
| id|source|destination|transportmode|distance|year|
+---+------+-----------+-------------+--------+----+
|  1|   CHN|        IND|     airplane|   200.0|1990|
|  2|   IND|        CHN|     airplane|   200.0|1991|
|  3|   IND|        CHN|     airplane|   200.0|1992|
|  4|   RUS|        IND|     airplane|   200.0|1990|
|  5|   CHN|        RUS|     airplane|   200.0|1992|
|  6|   AUS|        PAK|     airplane|   200.0|1991|
|  7|   RUS|        AUS|     airplane|   200.0|1990|
|  8|   IND|        RUS|     airplane|   200.0|1991|
|  9|   CHN|        RUS|     airplane|   200.0|1992|
| 10|   AUS|        CHN|     airplane|   200.0|1993|
|  1|   AUS|        CHN|     airplane|   200.0|1993|
|  2|   CHN|        IND|     airplane|   200.0|1993|
|  3|   CHN|        IND|     airplane|   200.0|1993|
|  4|   IND|        AUS|     airplane|   200.0|1991|
|  5|   AUS|        IND|     airplane|   200.0|1992|
|  6|   RUS|        CHN|     airplane|   200.0|1993|
|  7|   CHN|        RUS|     airplane|   200.0|1990|
|  8|   AUS|        CHN|     airplane|   200.0|1990|
|  9|   IND|        AUS|     airplane|   200.0|1991|
| 10|   RUS|        CHN|     airplane|   200.0|1992|
+---+------+-----------+-------------+--------+----+
only showing top 20 rows
```



```
scala> holidayDF.groupBy("year").count().show()
+----+-----+
|year|count|
+----+-----+
|1990|    8|
|1994|    1|
|1991|    9|
|1992|    7|
|1993|    7|
+----+-----+
```

2) What is the total air distance covered by each user per year

```
scala> val userDetail = sc.textFile("/user/spark/userdetail.txt").map(_.split(","))
userDetail: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[102] at map at <console>:24

scala> case class UserDetail(id:Int, name:String, age:Int)
defined class UserDetail

scala> val userDetailDF = userDetail.map(attr => UserDetail(attr(0).toInt, attr(1), attr(2).toInt)).toDF;
userDetailDF: org.apache.spark.sql.DataFrame = [id: int, name: string ... 1 more field]

scala> userDetailDF.show
+---+------+---+
| id|  name|age|
+---+------+---+
|  1|  mark| 15|
|  2|  john| 16|
|  3|  luke| 17|
|  4|  lisa| 27|
|  5|  mark| 25|
|  6| peter| 22|
|  7| james| 21|
|  8|andrew| 55|
|  9|thomas| 46|
| 10| annie| 44|
+---+------+---+
```

```
scala> val user = holidayDF.join(userDetailDF, "id")
user: org.apache.spark.sql.DataFrame = [id: int, source: string ... 6 more fields]

scala> user.show
+---+------+-----------+-------------+--------+----+------+---+
| id|source|destination|transportmode|distance|year|  name|age|
+---+------+-----------+-------------+--------+----+------+---+
|  1|   CHN|        IND|     airplane|   200.0|1998|  mark| 15|
|  1|   AUS|        CHN|     airplane|   200.0|1993|  mark| 15|
|  1|   PAK|        IND|     airplane|   200.0|1993|  mark| 15|
|  1|   PAK|        AUS|     airplane|   200.0|1993|  mark| 15|
|  6|   AUS|        PAK|     airplane|   200.0|1991| peter| 22|
|  6|   RUS|        CHN|     airplane|   200.0|1993| peter| 22|
|  6|   PAK|        RUS|     airplane|   200.0|1991| peter| 22|
|  3|   IND|        CHN|     airplane|   200.0|1992|  luke| 17|
|  3|   CHN|        IND|     airplane|   200.0|1993|  luke| 17|
|  3|   CHN|        PAK|     airplane|   200.0|1991|  luke| 17|
|  5|   CHN|        RUS|     airplane|   200.0|1992|  mark| 25|
|  5|   AUS|        IND|     airplane|   200.0|1992|  mark| 25|
|  5|   IND|        PAK|     airplane|   200.0|1991|  mark| 25|
|  5|   CHN|        PAK|     airplane|   200.0|1994|  mark| 25|
|  9|   CHN|        RUS|     airplane|   200.0|1992|thomas| 46|
|  9|   IND|        AUS|     airplane|   200.0|1991|thomas| 46|
|  9|   RUS|        IND|     airplane|   200.0|1992|thomas| 46|
|  4|   RUS|        IND|     airplane|   200.0|1998|  lisa| 27|
|  4|   IND|        AUS|     airplane|   200.0|1991|  lisa| 27|
|  4|   CHN|        PAK|     airplane|   200.0|1998|  lisa| 27|
+---+------+-----------+-------------+--------+----+------+---+
only showing top 20 rows
```

```
scala> user.groupBy("year","name").agg(sum("distance")).show
+----+------+-------------+
|year|  name|sum(distance)|
+----+------+-------------+
|1991|  mark|        200.0|
|1990|andrew|        200.0|
|1991|andrew|        200.0|
|1991|  luke|        200.0|
|1993|  mark|        600.0|
|1991| peter|        400.0|
|1993|  luke|        200.0|
|1991|thomas|        200.0|
|1993|  john|        200.0|
|1991|  john|        400.0|
|1990| annie|        200.0|
|1994|  mark|        200.0|
|1990|  mark|        200.0|
|1990|  lisa|        400.0|
|1992|thomas|        400.0|
|1990| james|        600.0|
|1993| peter|        200.0|
|1991|  lisa|        200.0|
|1992|  luke|        200.0|
|1992| annie|        200.0|
+----+------+-------------+
only showing top 20 rows
```

3) Which user has travelled the largest distance till date

```
scala> user.groupBy("year","name").agg(max("distance")).sort(desc("year")).show
+----+------+-------------+
|year|  name|max(distance)|
+----+------+-------------+
|1994|  mark|        200.0|
|1993|  luke|        200.0|
|1993| annie|        200.0|
|1993| peter|        200.0|
|1993|  john|        200.0|
|1993|  mark|        200.0|
|1992|andrew|        200.0|
|1992| annie|        200.0|
|1992|thomas|        200.0|
|1992|  luke|        200.0|
|1992|  mark|        200.0|
|1991|  luke|        200.0|
|1991|thomas|        200.0|
|1991|  mark|        200.0|
|1991|andrew|        200.0|
|1991|  lisa|        200.0|
|1991|  john|        200.0|
|1991| peter|        200.0|
|1990| james|        200.0|
|1990|  mark|        200.0|
+----+------+-------------+
only showing top 20 rows
```

4) What is the most preferred destination for all users.

```
scala> user.groupBy("name","destination").agg(count("destination")).orderBy("destination").show
+------+-----------+------------------+
|  name|destination|count(destination)|
+------+-----------+------------------+
|  lisa|        AUS|                 1|
|thomas|        AUS|                 1|
|  mark|        AUS|                 1|
| annie|        AUS|                 1|
| james|        AUS|                 1|
| annie|        CHN|                 2|
|andrew|        CHN|                 1|
|  mark|        CHN|                 1|
|  john|        CHN|                 1|
| peter|        CHN|                 1|
|  luke|        CHN|                 1|
|  john|        IND|                 1|
|thomas|        IND|                 1|
|  mark|        IND|                 3|
|andrew|        IND|                 1|
| james|        IND|                 1|
|  lisa|        IND|                 1|
|  luke|        IND|                 1|
|  lisa|        PAK|                 1|
|  luke|        PAK|                 1|
+------+-----------+------------------+
only showing top 20 rows
```

5)Which route is generating the most revenue per year

```
scala> val transport = sc.textFile("/user/spark/transport.txt").map( .split(','))
transport: org.apache.spark.rdd.RDD[Array[String]] = MapPartitionsRDD[371] at map at <console>:24

scala> case class TransportDetail(transportmode:String, cost:Int)
defined class TransportDetail

scala> val transportDF = transport.map(attr => TransportDetail(attr(0), attr(1).toInt)).toDF;
transportDF: org.apache.spark.sql.DataFrame = [transportmode: string, cost: int]

scala> val userInfo = user.join(transportDF, "transportmode").toDF
userInfo: org.apache.spark.sql.DataFrame = [transportmode: string, id: int ... 7 more fields]

scala> userInfo.show
+-------------+---+------+-----------+--------+----+------+---+----+
|transportmode| id|source|destination|distance|year|  name|age|cost|
+-------------+---+------+-----------+--------+----+------+---+----+
|     airplane|  1|   CHN|        IND|   200.0|1990|  mark| 15| 170|
|     airplane|  1|   AUS|        CHN|   200.0|1993|  mark| 15| 170|
|     airplane|  1|   PAK|        IND|   200.0|1993|  mark| 15| 170|
|     airplane|  1|   PAK|        AUS|   200.0|1993|  mark| 15| 170|
|     airplane|  6|   AUS|        PAK|   200.0|1991| peter| 22| 170|
|     airplane|  6|   RUS|        CHN|   200.0|1993| peter| 22| 170|
|     airplane|  6|   PAK|        RUS|   200.0|1991| peter| 22| 170|
|     airplane|  3|   IND|        CHN|   200.0|1992|  luke| 17| 170|
|     airplane|  3|   CHN|        IND|   200.0|1993|  luke| 17| 170|
|     airplane|  3|   CHN|        PAK|   200.0|1991|  luke| 17| 170|
|     airplane|  5|   CHN|        RUS|   200.0|1992|  mark| 25| 170|
|     airplane|  5|   AUS|        IND|   200.0|1992|  mark| 25| 170|
|     airplane|  5|   IND|        PAK|   200.0|1991|  mark| 25| 170|
|     airplane|  5|   CHN|        PAK|   200.0|1994|  mark| 25| 170|
|     airplane|  9|   CHN|        RUS|   200.0|1992|thomas| 46| 170|
|     airplane|  9|   IND|        AUS|   200.0|1991|thomas| 46| 170|
|     airplane|  9|   RUS|        IND|   200.0|1992|thomas| 46| 170|
|     airplane|  4|   RUS|        IND|   200.0|1990|  lisa| 27| 170|
|     airplane|  4|   IND|        AUS|   200.0|1991|  lisa| 27| 170|
|     airplane|  4|   CHN|        PAK|   200.0|1990|  lisa| 27| 170|
+-------------+---+------+-----------+--------+----+------+---+----+
```

```
scala> userInfo.groupBy("year","source","destination").agg(sum("cost")).show
+----+------+-----------+---------+
|year|source|destination|sum(cost)|
+----+------+-----------+---------+
|1993|   RUS|        CHN|      170|
|1991|   IND|        PAK|      170|
|1993|   PAK|        IND|      170|
|1990|   CHN|        IND|      340|
|1990|   CHN|        RUS|      170|
|1991|   IND|        CHN|      170|
|1992|   RUS|        CHN|      170|
|1994|   CHN|        PAK|      170|
|1991|   AUS|        PAK|      170|
|1992|   RUS|        IND|      340|
|1992|   AUS|        IND|      170|
|1991|   CHN|        PAK|      170|
|1990|   CHN|        AUS|      170|
|1993|   AUS|        CHN|      340|
|1990|   RUS|        AUS|      170|
|1993|   PAK|        AUS|      170|
|1990|   CHN|        PAK|      170|
|1991|   IND|        AUS|      340|
|1990|   AUS|        CHN|      170|
|1993|   CHN|        IND|      340|
+----+------+-----------+---------+
```

6) What is the total amount spent by every user on air-travel per year

```
scala> userInfo.groupBy("year").agg(sum("cost")).show
+----+---------+
|year|sum(cost)|
+----+---------+
|1990|     1360|
|1994|      170|
|1991|     1530|
|1992|     1190|
|1993|     1190|
+----+---------+
```

7) Considering age groups of < 20 , 20-35, 35 > ,Which age group is travelling the most every year.

```
scala> userInfo.filter($"age" < 20 || ($"age" <=20 && $"age" <=35) || $"age" > 35).show
+-------------+---+------+-----------+--------+----+--------+---+----+
|transportmode| id|source|destination|distance|year|    name|age|cost|
+-------------+---+------+-----------+--------+----+--------+---+----+
|     airplane|  1|   CHN|        IND|   200.0|1990|    mark| 15| 170|
|     airplane|  1|   AUS|        CHN|   200.0|1993|    mark| 15| 170|
|     airplane|  1|   PAK|        IND|   200.0|1993|    mark| 15| 170|
|     airplane|  1|   PAK|        AUS|   200.0|1993|    mark| 15| 170|
|     airplane|  3|   IND|        CHN|   200.0|1992|    luke| 17| 170|
|     airplane|  3|   CHN|        IND|   200.0|1993|    luke| 17| 170|
|     airplane|  3|   CHN|        PAK|   200.0|1991|    luke| 17| 170|
|     airplane|  9|   CHN|        RUS|   200.0|1992|  thomas| 46| 170|
|     airplane|  9|   IND|        AUS|   200.0|1991|  thomas| 46| 170|
|     airplane|  9|   RUS|        IND|   200.0|1992|  thomas| 46| 170|
|     airplane|  8|   IND|        RUS|   200.0|1991|  andrew| 55| 170|
|     airplane|  8|   AUS|        CHN|   200.0|1990|  andrew| 55| 170|
|     airplane|  8|   RUS|        IND|   200.0|1992|  andrew| 55| 170|
|     airplane| 10|   AUS|        CHN|   200.0|1993|   annie| 44| 170|
|     airplane| 10|   RUS|        CHN|   200.0|1992|   annie| 44| 170|
|     airplane| 10|   CHN|        AUS|   200.0|1990|   annie| 44| 170|
|     airplane|  2|   IND|        CHN|   200.0|1991|    john| 16| 170|
|     airplane|  2|   CHN|        IND|   200.0|1993|    john| 16| 170|
|     airplane|  2|   IND|        RUS|   200.0|1991|    john| 16| 170|
+-------------+---+------+-----------+--------+----+--------+---+----+
```