Session 18 Assignment 1

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

```
[bigdata@localhost bin]$ ./spark-shell -i /home/bigdata/deepak/docs/Acadgild/Session18Assignment1/Session18Assignment1_1.scala
2018-05-17 13:26:07 WARN  Utils:66 - Your hostname, localhost.localdomain resolves to a loopback address: 127.0.0.1; using 192.168.0.6 instead (on interface en
p0s3)
2018-05-17 13:26:07 WARN  Utils:66 - Set SPARK_LOCAL_IP if you need to bind to another address
```

```
Loading /home/bigdata/deepak/docs/Acadgild/Session18Assignment1/Session18Assignment1_1.scala...
import spark.implicits._
2018-05-17 13:29:58 WARN  ObjectStore:568 - Failed to get database global_temp, returning NoSuchObjectException
dataset_holiday: org.apache.spark.sql.DataFrame = [_c0: string, _c1: string ... 4 more fields]
################################################################################
      Problem 1 - Total air travellers per year
################################################################################
+----+-----+
| _c5|count|
+----+-----+
|1992|    7|
|1994|    1|
|1993|    7|
|1990|    8|
|1991|    9|
+----+-----+
```

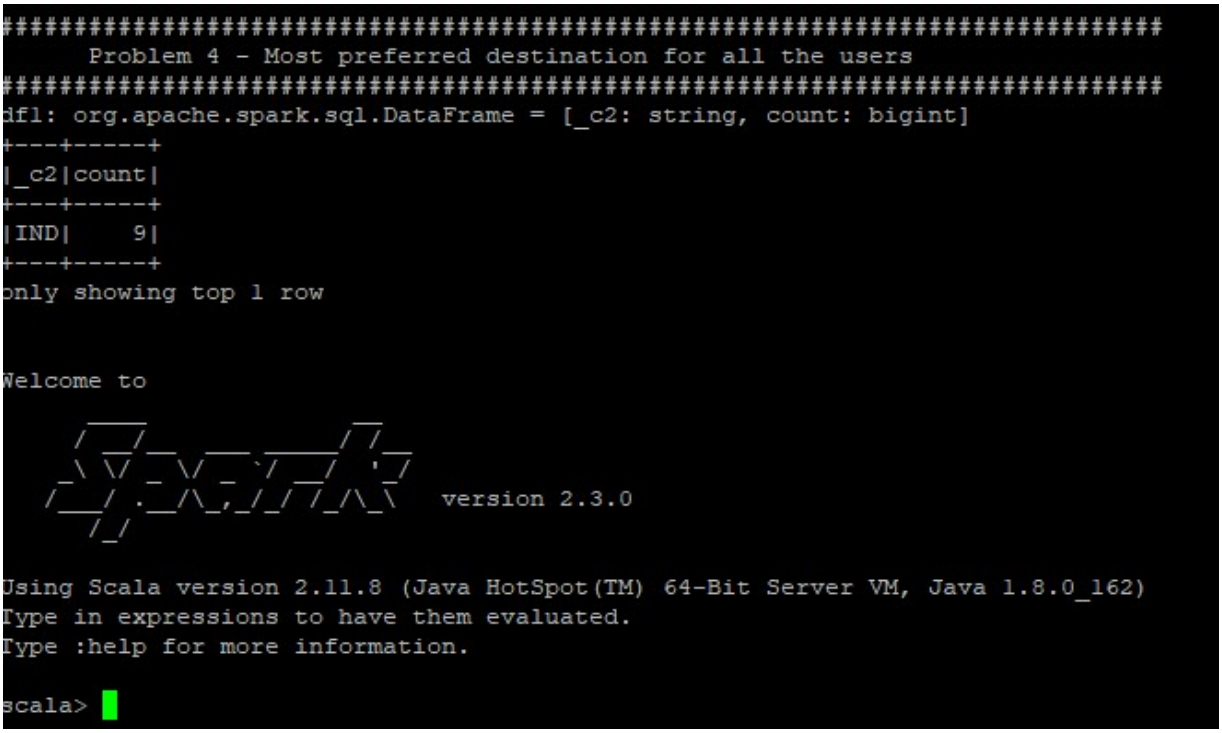
2) What is the total air distance covered by each user per year

3) Which user has travelled the largest distance till date

```
################################################################################
      Problem 2 - Total air distance covered by each user per year
################################################################################
+---+----+--------+
|_c0| _c5|sum(_c4)|
+---+----+--------+
|  1|1990|   200.0|
| 10|1992|   200.0|
|  7|1990|   600.0|
|  6|1993|   200.0|
| 10|1990|   200.0|
|  8|1990|   200.0|
|  8|1991|   200.0|
|  3|1991|   200.0|
|  1|1993|   600.0|
|  9|1992|   400.0|
|  2|1991|   400.0|
|  3|1993|   200.0|
|  4|1990|   400.0|
|  2|1993|   200.0|
|  4|1991|   200.0|
| 10|1993|   200.0|
|  9|1991|   200.0|
|  8|1992|   200.0|
|  3|1992|   200.0|
|  5|1992|   400.0|
+---+----+--------+
only showing top 20 rows

################################################################################
      Problem 3 - User which has travelled largest distance till date
################################################################################
df1: org.apache.spark.sql.DataFrame = [_c0: string, sum(_c4): double]
+---+--------+
|_c0|sum(_c4)|
+---+--------+
|  5|   800.0|
+---+--------+
only showing top 1 row
```

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


##
Problem 4 - Most preferred destination for all the users
##
df1: org.apache.spark.sql.DataFrame = [_c2: string, count: bigint]
+---+-----+
|_c2|count|
+---+-----+
|IND| 9|
+---+-----+
only showing top 1 row
Welcome to
version 2.3.0
Using Scala version 2.11.8 (Java HotSpot(TM) 64-Bit Server VM, Java 1.8.0_162)
Type in expressions to have them evaluated.
Type :help for more information.
scala>