**Assignment 18.1:**

**Problem Statements:**

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

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

3) Which user has travelled the largest distance till date

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

**Solution:-**

Loading data to dataframe:-



scala> val transportRDD = sc.textFile("S18\_Dataset\_Transport.txt");

scala> val transportDF= transportRDD.map(lines=>lines.split(",")).map(arrays=>(arrays(0),arrays(1))).toDF("Transport\_Name","Transport\_ID");



scala> val userRDD = sc.textFile("S18\_Dataset\_User\_details.txt");

scala> val userDF= userRDD.map(lines=>lines.split(",")).map(arrays=>(arrays(0),arrays(1),arrays(2))).toDF("Person\_ID","Name","Age");



Creating temporary view out of all data frames.

scala> holidaysDF.registerTempTable("holidays");

scala> transportDF.registerTempTable("transport");

scala> userDF.registerTempTable("user");



All the datasets have been loaded in to temporary tables.

These temporary tables and dataframes would be used to find solution to problem statements

Listing out temporary tables and column names:-

holidays -> "Person\_ID","Source","Destination","Mode","Fare","Year"

transport -> "Transport\_Name","Transport\_ID"

user -> "Person\_ID","Name","Age"

Starting with problem statements now:-

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



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







Rest of output --->>>>>



**3) Which user has travelled the largest distance till date**



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

