Session 18 : RDD'S CONTD. & INTRODUCTION TO DATAFRAMES

Assignment 3

**Problem Statement**

1) Considering age groups of < 20 , 20-35, 35 > ,Which age group spends the most

amount of money travelling.

2) What is the amount spent by each age-group, every year in travelling?

Use the dataset given below:

<https://drive.google.com/drive/folders/0B_P3pWagdIrrVThBaUdVSUtzbms>

**Solution:-**

Starting Spark shell

Command - spark-shell



Loading data to dataframe:-

**Scala>** val holidaysRDD = sc.textFile("S18\_Dataset\_Holidays.txt");

**Scala>** val holidaysDF= holidaysRDD.map(lines=>lines.split(",")).map(arrays => (arrays(0),arrays(1),arrays(2),arrays(3),arrays(4),arrays(5))).toDF("Person\_ID","Source","Destination","Mode","Distance","Year");

**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","Fare");

**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");



All the datasets have been loaded in to temporary tables.

These dataframes would be used to find solution to problem statements

Listing out dataframes and column names:-

holidaysDF -> "Person\_ID","Source","Destination","Mode","Distance","Year"

transportDF -> "Transport\_Name","Fare"

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

**Starting with problem statements now:-**

**1) Considering age groups of < 20 , 20-35, 35 > ,Which age group spends the most**

**amount of money travelling.**

**scala>** val ageGrp = udf((age: String) => { if(age.toInt < 20) { "<20"; } else { if(age.toInt > 35) { ">35"; } else { "20-35"; }}});

**scala>** val userDFGrp = userDF.withColumn("AgeGrp", ageGrp($"Age"));

**scala>** val holidayWithPriceDF = holidaysDF.as("d1").join(transportDF.as("d2"), $"d1.Mode" === $"d2.Transport\_Name").select($"d1.\*", $"d2.Fare");

**scala>** val userHoliday = holidayWithPriceDF.as("d1").join(userDFGrp.as("d2"), $"d1.Person\_ID" === $"d2.Person\_ID").select($"d1.\*", $"d2.\*");

**scala>** val userGrpExpenditure = userHoliday.groupBy("AgeGrp").agg(sum("Fare")).orderBy($"sum(Fare)".desc);

**scala>** userGrpExpenditure.show(1);



**Age group 20-35 has spent most amounts i.e. 2210.0 on air travelling**

**2) What is the amount spent by each age-group, every year in travelling?**

**scala>** val userGrpYearlyExpenditure = userHoliday.groupBy("AgeGrp","Year").agg(sum("Fare")).orderBy($"AgeGrp", $"Year", $"sum(Fare)".desc);

userGrpYearlyExpenditure: org.apache.spark.sql.DataFrame = [AgeGrp: string, Year: string, sum(Fare): double]

**scala>** userGrpYearlyExpenditure.show;



***Above is a list of amount spent by each age-group, every year in travelling***