ASSIGNMENT 18.3

**Input Datasets:**

We have an airline data with us:

**user details:**

user\_id, name, age

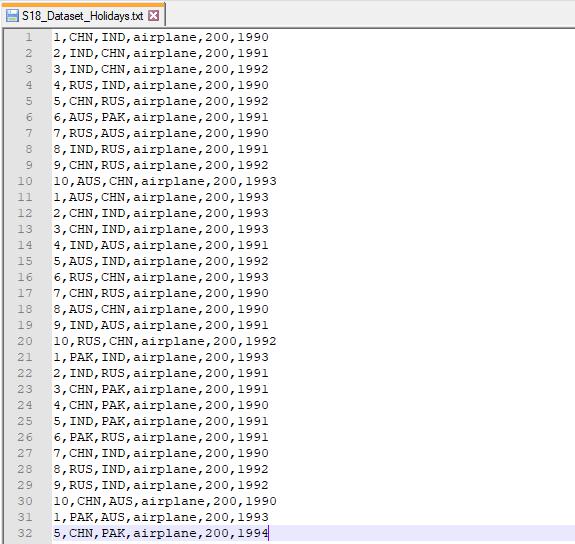
**holidays:**

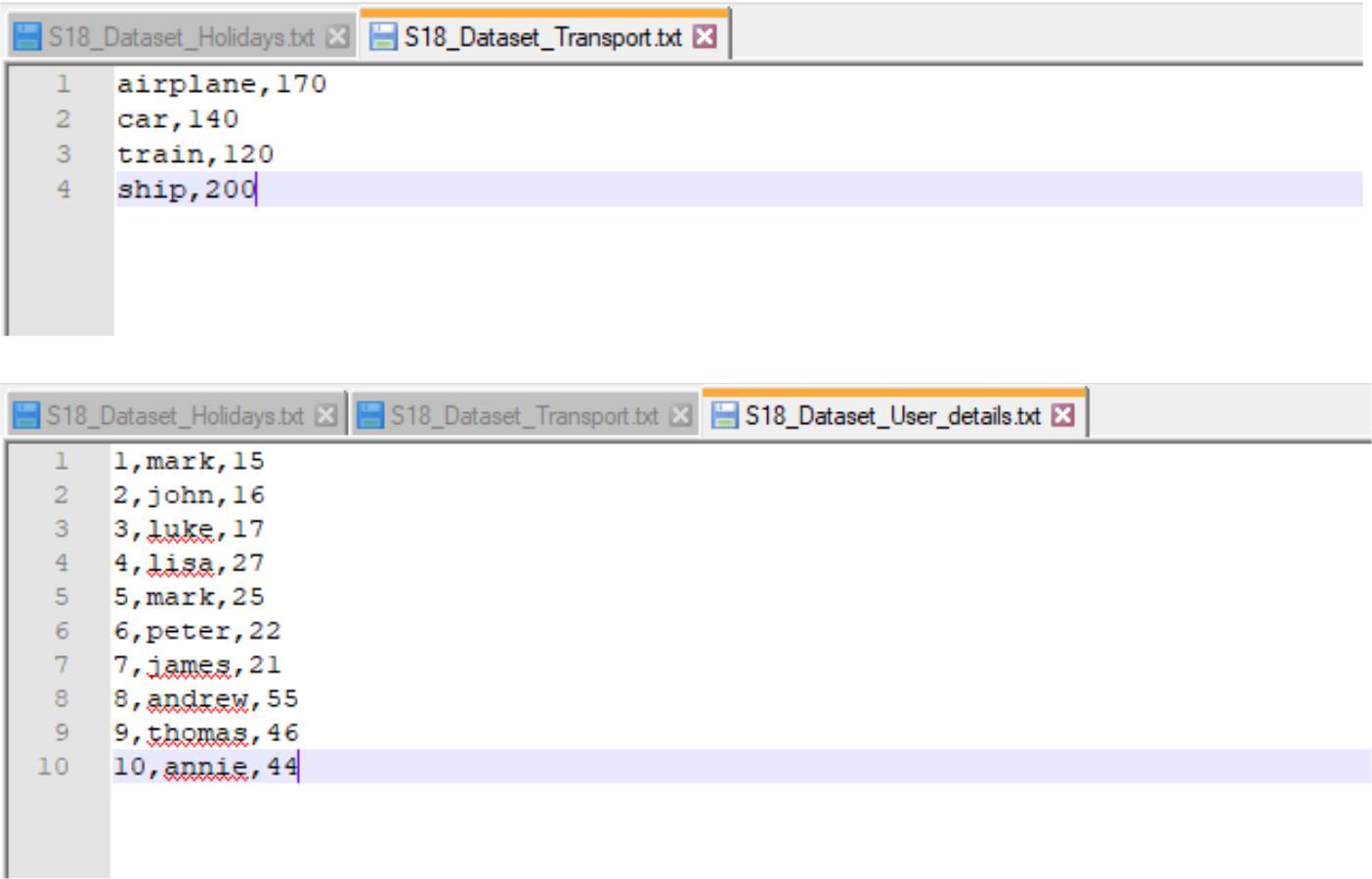
user\_id, src, dest, travel\_mode, distance, year\_of\_travel

**transport:**

travel\_mode, cost\_per\_unit

Here are the datasets which we will be using for this assignment in all problems. It has been kept in local file system:





**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?

**Solution:**

1. Here is the Spark code snippet to find the age group that spends most amount of money travelling:

// import required Spark packages

import org.apache.spark.sql.SparkSession

import org.apache.spark.sql.types.{IntegerType, StringType}

object Assignment18\_2 {

def main(args: Array[String]): Unit = {

val spark = SparkSession // create a SparkSession object that can be used to

.builder() // create various contexts of Spark such as sqlContext

.config("spark.sql.warehouse.dir", "file:///c:/tmp/spark-warehouse")

.master("local[\*]")

.getOrCreate()

val sqlContext = spark.sqlContext // initialize sqlContext

val input\_df1 = sqlContext.read.csv ("E:\\Acadgild\\Session 18\\S18\_Datasets \\S18\_Dataset

\_Transport.txt") // load input data file – transport.txt val transportDF = input\_df1.select( // define schema for input data loaded input\_df1("\_c0").cast(StringType).as("travel\_mode"), //assign column names to the data frame input\_df1("\_c1").cast(IntegerType).as("cost\_per\_unit")) transportDF.createOrReplaceTempView("transport") // create a temporary view - transport

val input\_df2 = sqlContext.read.csv("E:\\Acadgild\\Session 18\\S18\_Datasets\\S18\_Dataset\_

Holidays.txt")

val holidaysDF = input\_df2.select(

input\_df2("\_c0").cast(IntegerType).as("user\_id"),

// load input data file – holidays.txt

// define schema for input data loaded

//assign column names to the data frame

input\_df2("\_c1").cast(StringType).as("src"),

input\_df2("\_c2").cast(StringType).as("dest"),

input\_df2("\_c3").cast(StringType).as("travel\_mode"),

input\_df2("\_c4").cast(IntegerType).as("distance"),

input\_df2("\_c5").cast(IntegerType).as("year\_of\_travel"))

holidaysDF.createOrReplaceTempView("holidays") // create a temporary view - holidays

val input\_df3 = sqlContext.read.csv("E:\\Acadgild\\Session 18\\S18\_Datasets\\S18\_Dataset\_

User\_details.txt")

val usersDF = input\_df3.select(

input\_df3("\_c0").cast(IntegerType).as("user\_id"),

// load input data file – user\_details.txt // define schema for input data loaded //assign column names to the data frame

input\_df3("\_c1").cast(StringType).as("name"),

input\_df3("\_c2").cast(IntegerType).as("age"))

usersDF.createOrReplaceTempView("users")

// create a temporary view – users

* SQL query to find the age group that spends most amount of money travelling sqlContext.sql(""SELECT z.age, SUM(x.cost\_per\_unit) as total\_amount " +

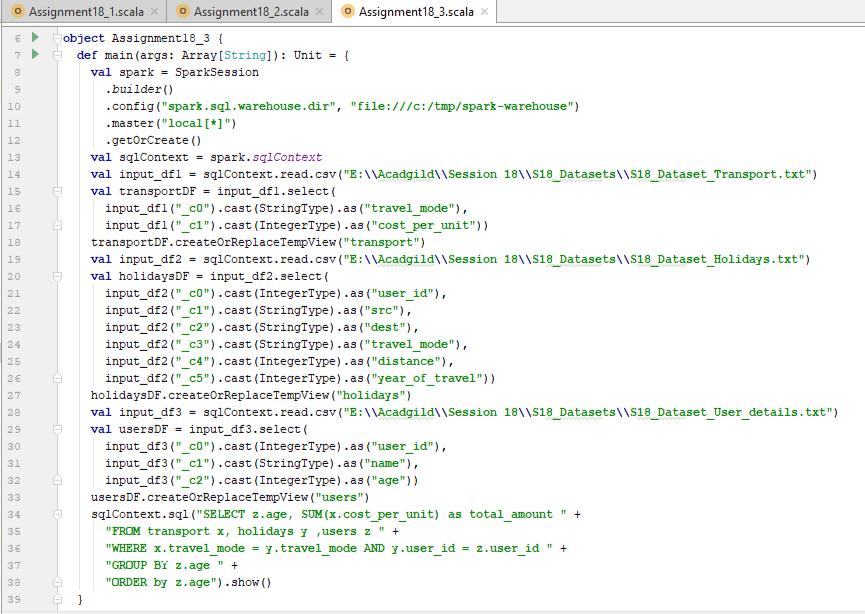
"FROM transport x, holidays y ,users z " +

"WHERE x.travel\_mode = y.travel\_mode AND y.user\_id = z.user\_id " + "GROUP BY z.age " +

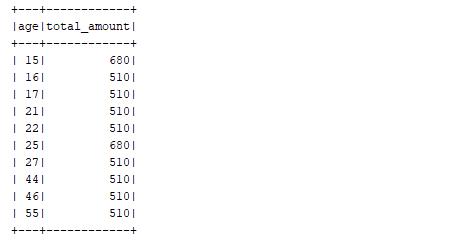
"ORDER by z.age").show() // print the result

}

}



**Output:**



The total amount spent by the age group below 15 is **1700,** the age group 20 – 35 spends **2210** and

the age group above 35 spends **1530.** So we can deduce that **the age group 20** **–** **35 spends the**

**most on travelling.**