Assignment 20 Spark SQL 1

Task 1

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

```
val obj1 = spark.sql("select count(id),year from Holiday where transport_mode = 'airplane'
group by(year) ")
```

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

```
val obj2 = spark.sql("select id,SUM(distance) a,year from Holiday where transport_mode =
'airplane' group by id,year order by year,id")
```

3) Which user has travelled the largest distance till date

```
val obj3 = spark.sql("select id,SUM(distance) AS TotalDistance from Holiday group by id
order by TotalDistance DESC ")
obj3.registerTempTable("temp_table")
println("Task 3 : Most distance")
val obj31 = spark.sql("select id,TotalDistance from temp_table where TotalDistance in
(select
```

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

```
val obj4 = spark.sql("select id,count(destination) as Total_Count from Holiday group by id
order by Total_Count DESC")
```

5)Which route is generating the most revenue per year

```
val revenue_per_year = Holidaydata.join(Transportdata,Holidaydata("transport_mode")===
Transportdata("transport_mode")).

groupBy("year","source","destination").sum("cost_per_unit").sort(desc("sum(cost_per_unit)"
)).show(10)
```

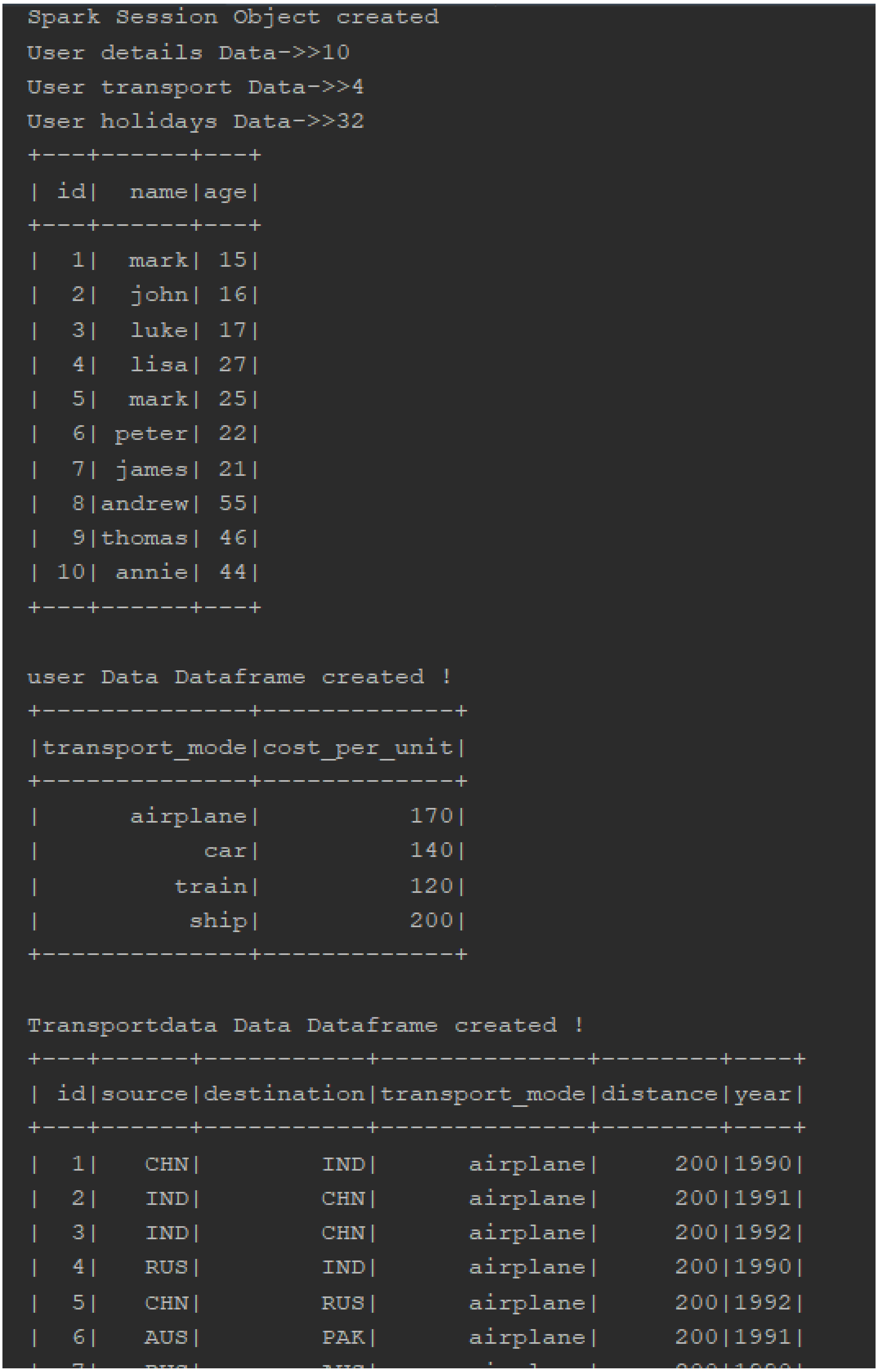
6) What is the total amount spent by every user on air-travel per year

```
val amount_spent_per_year =
Holidaydata.join(Transportdata,Holidaydata("transport_mode")===
Transportdata("transport_mode")).
  groupBy("id","year").sum("cost_per_unit").orderBy("id","year").show()
```

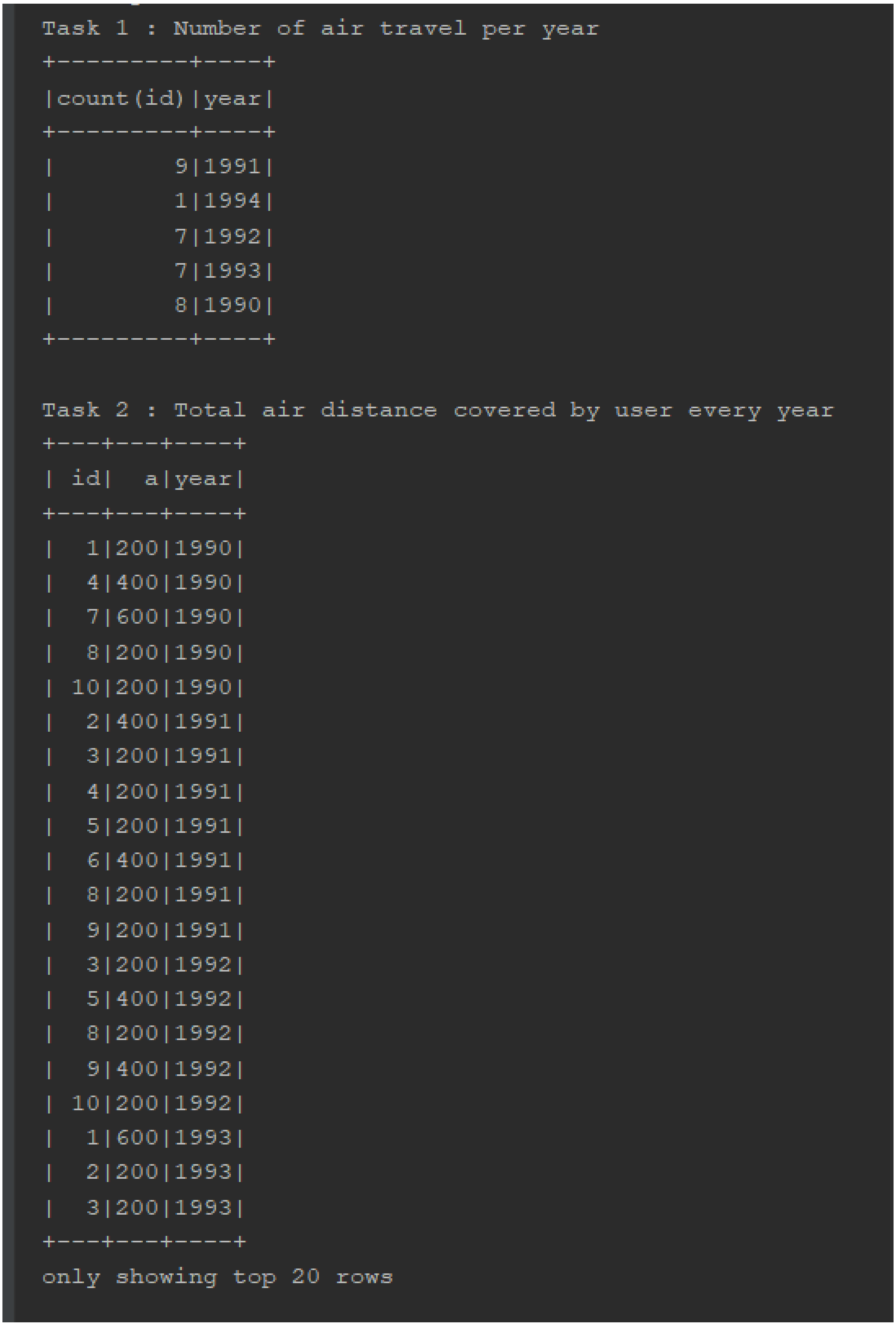
Source Code

```
package SQL

import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions._

object Assignment20 {

  case class User(id:Int , name:String , age:Int)
  case class Transport(transport_mode: String, cost_per_unit: Int)
  case class Holidays(id: Int, source: String, destination: String, transport_mode: String,
distance: BigInt, year: Long)
  case class temp_table(id:Int, TotalDistance:Int)

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

    println("hey scala")

    //Let us create a spark session object

    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark SQL Use Case 1 ")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()

    println("Spark Session Object created")


    //Set the log level as warning
    spark.sparkContext.setLogLevel("WARN")

    val data = spark.sparkContext.textFile("/Users/Vidya
Sagar/S20_Dataset_User_details.txt")
    println("User details Data->>" +data.count())

    val data1 = spark.sparkContext.textFile("/Users/Vidya Sagar/S20_Dataset_Transport.txt")
    println("User transport Data->>" +data1.count())

    val data2 = spark.sparkContext.textFile("/Users/Vidya Sagar/S20_Dataset_Holidays.txt")
    println("User holidays Data->>" +data2.count())

    //For implicit conversions like converting RDDs and sequences  to DataFrames
    import spark.implicits._


    val userdata = data.map(x => x.split(",")).map(x => User(x(0).toInt, x(1), x(2).toInt)).toDF
    userdata.show()

    println("user Data Dataframe created !")

    userdata.registerTempTable("User")

    val Transportdata = data1.map(x => x.split(",")).map(x => Transport(x(0),
x(1).toInt)).toDF

    Transportdata.show()

    println("Transportdata Data Dataframe created !")

    Transportdata.registerTempTable("Transport")

    val Holidaydata = data2.map(x => x.split(",")).map(x => Holidays(x(0).toInt, x(1),
x(2),x(3),x(4).toInt,x(5).toLong)).toDF

    Holidaydata.show()

    println("Holidaydata Data Dataframe created !")

    Holidaydata.registerTempTable("Holiday")

    println("Task 1 : Number of air travel per year")
    val obj1 = spark.sql("select count(id),year from Holiday where transport_mode = 'airplane'
group by(year) ")
    obj1.show()

    println("Task 2 : Total air distance covered by user every year")
    val obj2 = spark.sql("select id,SUM(distance) a,year from Holiday where transport_mode
= 'airplane' group by id,year order by year,id")
    obj2.show()

    println("Creating a temp table")
    val obj3 = spark.sql("select id,SUM(distance) AS TotalDistance from Holiday group by id
order by TotalDistance DESC ")
    obj3.show()

    obj3.registerTempTable("temp_table")
    println("Task 3 : Most distance")
    val obj31 = spark.sql("select id,TotalDistance from temp_table where TotalDistance in
(select MAX(TotalDistance) from temp_table) ")
    obj31.show

    println("Task 4 : preferred destinations")
    val obj4 = spark.sql("select id,count(destination) as Total_Count from Holiday group by id
order by Total_Count DESC")
    obj4.show()
    println("Task 5 : Route generating most revenue per year")
    val revenue_per_year =
Holidaydata.join(Transportdata,Holidaydata("transport_mode")===
Transportdata("transport_mode")).

groupBy("year","source","destination").sum("cost_per_unit").sort(desc("sum(cost_per_unit)"
)).show(10)

    println("Task 6 : total amount spent by every user on air travel per year")
    val amount_spent_per_year =
Holidaydata.join(Transportdata,Holidaydata("transport_mode")===
Transportdata("transport_mode")).
      groupBy("id","year").sum("cost_per_unit").orderBy("id","year").show()


   // val obj5 = spark.sql("select id,source,destination as Total_Count from Holiday group by
id order by Total_Count DESC")
    //obj5.show()


    // val obj6 = spark.sql("select Holiday.id, Transport.cost_per_unit,Holiday.year from
Holiday JOIN Transport where Holiday.id =Transport.id group by id,year order by year,id")

   // obj6.show()
  }
}
```

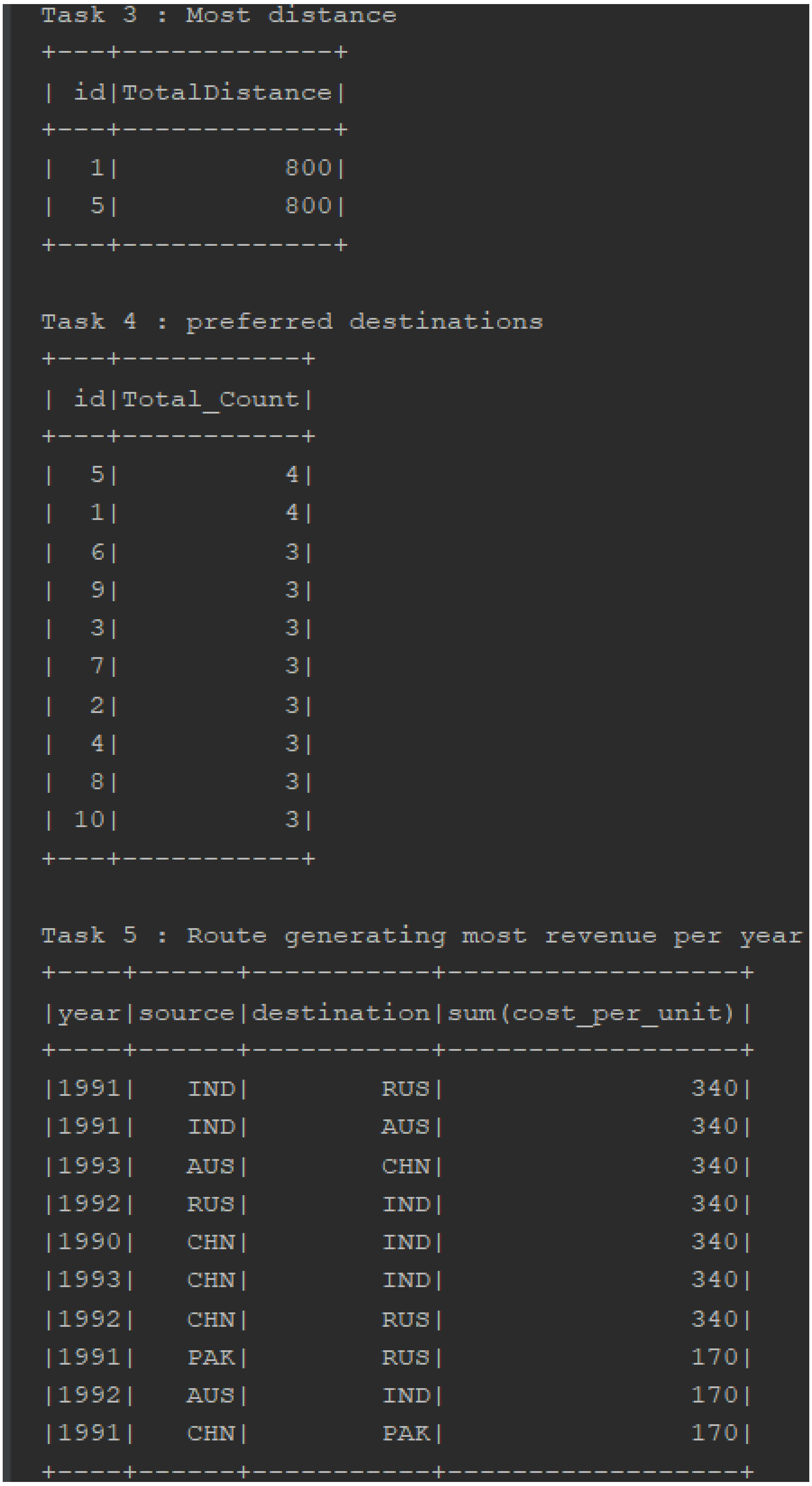
OUT PUT Screen shots



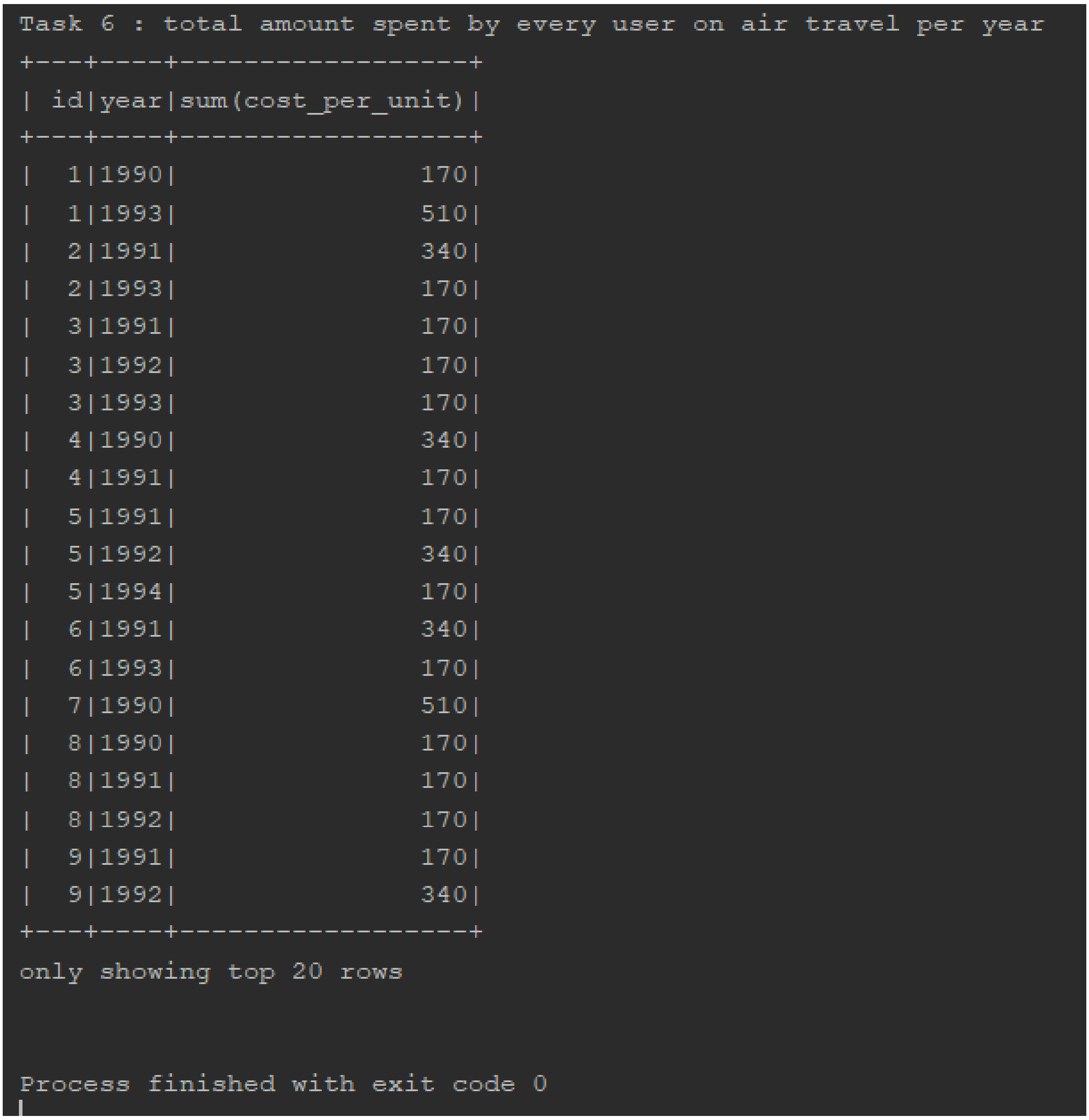
```
Spark Session Object created
User details Data->>10
User transport Data->>4
User holidays Data->>32
+---+------+---+
| id|  name|age|
+---+------+---+
|  1|  mark| 15|
|  2|  john| 16|
|  3|  luke| 17|
|  4|  lisa| 27|
|  5|  mark| 25|
|  6| peter| 22|
|  7| james| 21|
|  8|andrew| 55|
|  9|thomas| 46|
| 10| annie| 44|
+---+------+---+

user Data Dataframe created !
+--------------+-------------+
|transport_mode|cost_per_unit|
+--------------+-------------+
|      airplane|          170|
|           car|          140|
|         train|          120|
|          ship|          200|
+--------------+-------------+

Transportdata Data Dataframe created !
+---+------+-----------+--------------+--------+----+
| id|source|destination|transport_mode|distance|year|
+---+------+-----------+--------------+--------+----+
|  1|   CHN|        IND|      airplane|     200|1990|
|  2|   IND|        CHN|      airplane|     200|1991|
|  3|   IND|        CHN|      airplane|     200|1992|
|  4|   RUS|        IND|      airplane|     200|1990|
|  5|   CHN|        RUS|      airplane|     200|1992|
|  6|   AUS|        PAK|      airplane|     200|1991|
```



```
Task 1 : Number of air travel per year
+---------+----+
|count(id)|year|
+---------+----+
|        9|1991|
|        1|1994|
|        7|1992|
|        7|1993|
|        8|1990|
+---------+----+

Task 2 : Total air distance covered by user every year
+---+---+----+
| id|  a|year|
+---+---+----+
|  1|200|1990|
|  4|400|1990|
|  7|600|1990|
|  8|200|1990|
| 10|200|1990|
|  2|400|1991|
|  3|200|1991|
|  4|200|1991|
|  5|200|1991|
|  6|400|1991|
|  8|200|1991|
|  9|200|1991|
|  3|200|1992|
|  5|400|1992|
|  8|200|1992|
|  9|400|1992|
| 10|200|1992|
|  1|600|1993|
|  2|200|1993|
|  3|200|1993|
+---+---+----+
only showing top 20 rows
```



```
Task 3 : Most distance
+---+-------------+
| id|TotalDistance|
+---+-------------+
|  1|          800|
|  5|          800|
+---+-------------+

Task 4 : preferred destinations
+---+-----------+
| id|Total_Count|
+---+-----------+
|  5|          4|
|  1|          4|
|  6|          3|
|  9|          3|
|  3|          3|
|  7|          3|
|  2|          3|
|  4|          3|
|  8|          3|
| 10|          3|
+---+-----------+

Task 5 : Route generating most revenue per year
+----+------+-----------+-----------------+
|year|source|destination|sum(cost_per_unit)|
+----+------+-----------+-----------------+
|1991|   IND|        RUS|              340|
|1991|   IND|        AUS|              340|
|1993|   AUS|        CHN|              340|
|1992|   RUS|        IND|              340|
|1990|   CHN|        IND|              340|
|1993|   CHN|        IND|              340|
|1992|   CHN|        RUS|              340|
|1991|   PAK|        RUS|              170|
|1992|   AUS|        IND|              170|
|1991|   CHN|        PAK|              170|
+----+------+-----------+-----------------+
```



```
Task 6 : total amount spent by every user on air travel per year
+---+----+------------------+
| id|year|sum(cost_per_unit)|
+---+----+------------------+
|  1|1990|               170|
|  1|1993|               510|
|  2|1991|               340|
|  2|1993|               170|
|  3|1991|               170|
|  3|1992|               170|
|  3|1993|               170|
|  4|1990|               340|
|  4|1991|               170|
|  5|1991|               170|
|  5|1992|               340|
|  5|1994|               170|
|  6|1991|               340|
|  6|1993|               170|
|  7|1990|               510|
|  8|1990|               170|
|  8|1991|               170|
|  8|1992|               170|
|  9|1991|               170|
|  9|1992|               340|
+---+----+------------------+
only showing top 20 rows


Process finished with exit code 0
```