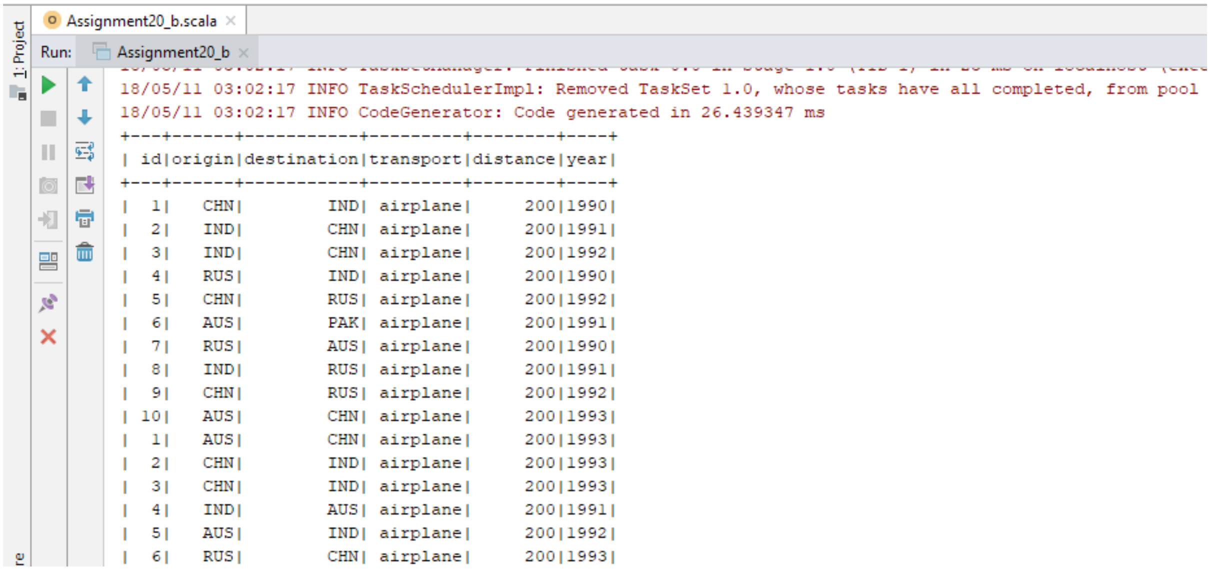
ASSIGNMENT 20.1

I have created this assignment in **IntelliJIDEA** application for scala. To solve the all seven problems, I created two scala file with objects **Assignment20** and **MaximumDistanceCoverByAgeGroupPerYear**.

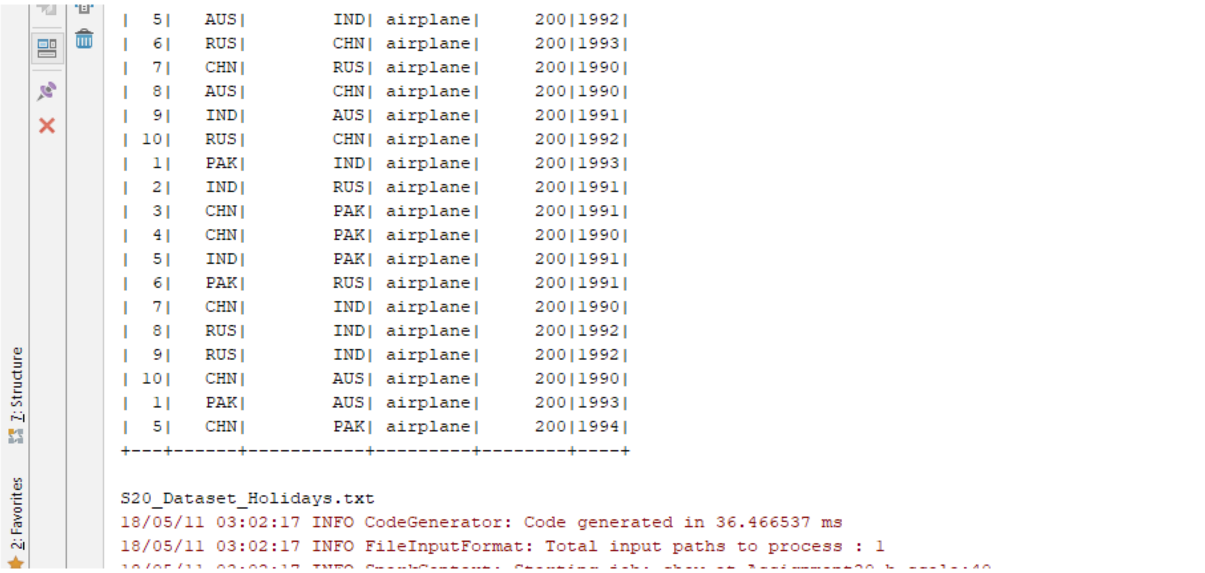
In **Assignment20.scala** codes for problem 1 to 6 are available and in **MaximumDistanceCoverByAgeGroupPerYear.scala** code for problem number7 is available with descriptions.

Below are the screen shots for the input dataset and output obtained by the code for each problem in IntelliJIDEA application.

Below screen, shot shows the input data-set S20_Dataset_Holidays.txt created in textfile RDD



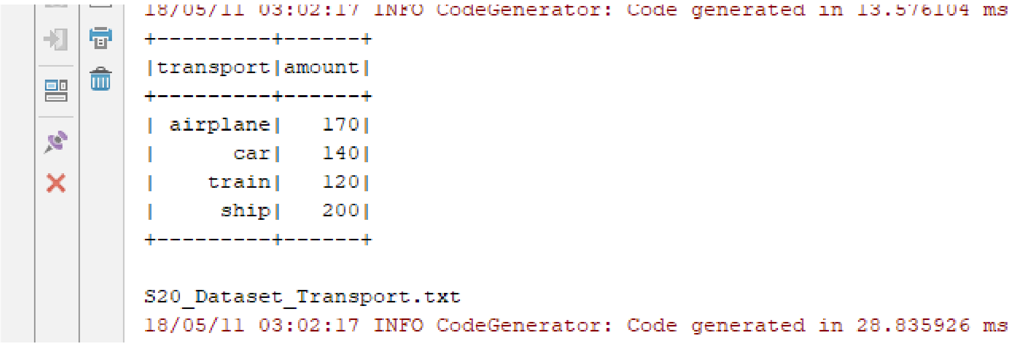
```
18/05/11 03:02:17 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/05/11 03:02:17 INFO CodeGenerator: Code generated in 26.439347 ms
+---+------+-----------+---------+--------+----+
| id|origin|destination|transport|distance|year|
+---+------+-----------+---------+--------+----+
|  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|
|  7|   RUS|        AUS| airplane|     200|1990|
|  8|   IND|        RUS| airplane|     200|1991|
|  9|   CHN|        RUS| airplane|     200|1992|
| 10|   AUS|        CHN| airplane|     200|1993|
|  1|   AUS|        CHN| airplane|     200|1993|
|  2|   CHN|        IND| airplane|     200|1993|
|  3|   CHN|        IND| airplane|     200|1993|
|  4|   IND|        AUS| airplane|     200|1991|
|  5|   AUS|        IND| airplane|     200|1992|
|  6|   RUS|        CHN| airplane|     200|1993|
```



```
|  5|   AUS|        IND| airplane|     200|1992|
|  6|   RUS|        CHN| airplane|     200|1993|
|  7|   CHN|        RUS| airplane|     200|1990|
|  8|   AUS|        CHN| airplane|     200|1990|
|  9|   IND|        AUS| airplane|     200|1991|
| 10|   RUS|        CHN| airplane|     200|1992|
|  1|   PAK|        IND| airplane|     200|1993|
|  2|   IND|        RUS| airplane|     200|1991|
|  3|   CHN|        PAK| airplane|     200|1991|
|  4|   CHN|        PAK| airplane|     200|1990|
|  5|   IND|        PAK| airplane|     200|1991|
|  6|   PAK|        RUS| airplane|     200|1991|
|  7|   CHN|        IND| airplane|     200|1990|
|  8|   RUS|        IND| airplane|     200|1992|
|  9|   RUS|        IND| airplane|     200|1992|
| 10|   CHN|        AUS| airplane|     200|1990|
|  1|   PAK|        AUS| airplane|     200|1993|
|  5|   CHN|        PAK| airplane|     200|1994|
+---+------+-----------+---------+--------+----+

S20_Dataset_Holidays.txt
18/05/11 03:02:17 INFO CodeGenerator: Code generated in 36.466537 ms
18/05/11 03:02:17 INFO FileInputFormat: Total input paths to process : 1
```

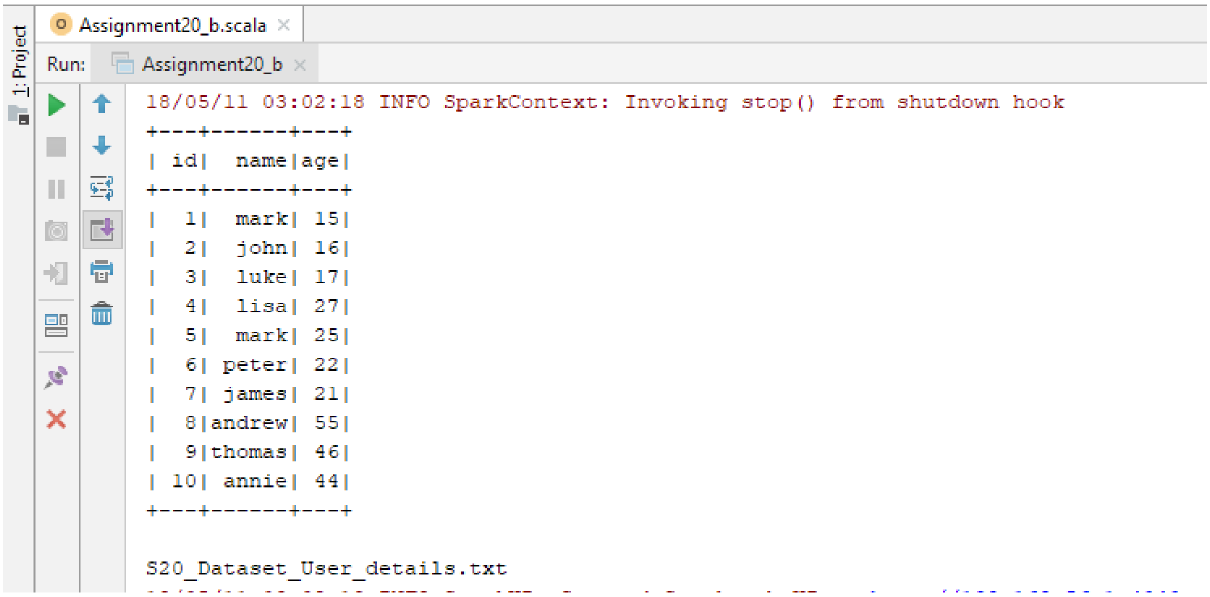
Below screen, shot shows the input data-set S20_Dataset_Transport.txt created in textfile RDD.



```
18/05/11 03:02:17 INFO CodeGenerator: Code generated in 13.576104 ms
+---------+------+
|transport|amount|
+---------+------+
| airplane|   170|
|      car|   140|
|    train|   120|
|     ship|   200|
+---------+------+

S20_Dataset_Transport.txt
18/05/11 03:02:17 INFO CodeGenerator: Code generated in 28.835926 ms
```

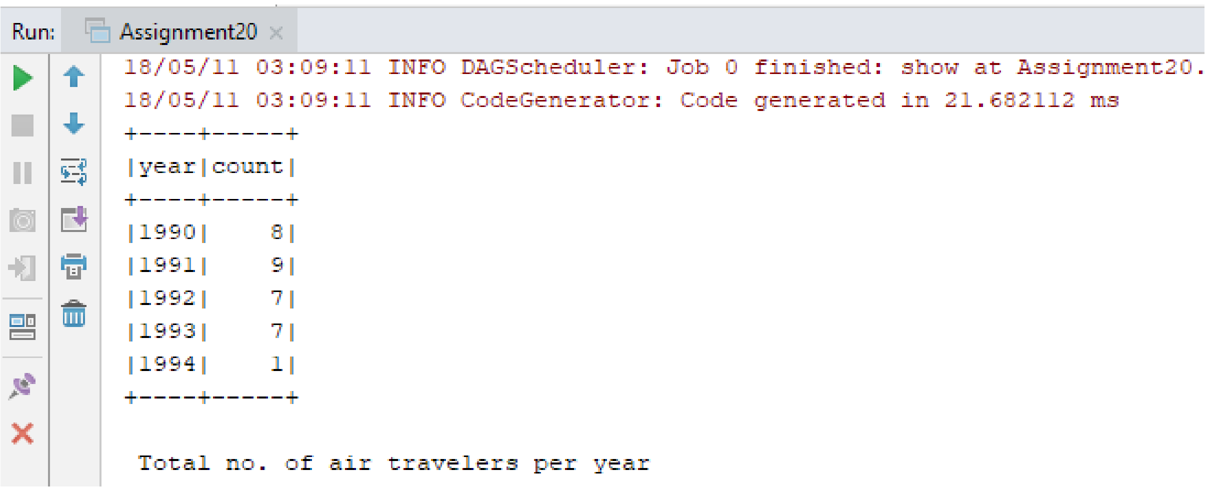
Below screen, shot shows the input data-set S20_Dataset_User_details.txt created in textfile RDD.



```
18/05/11 03:02:18 INFO SparkContext: Invoking stop() from shutdown hook
+---+------+---+
| 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|
+---+------+---+

S20_Dataset_User_details.txt
```

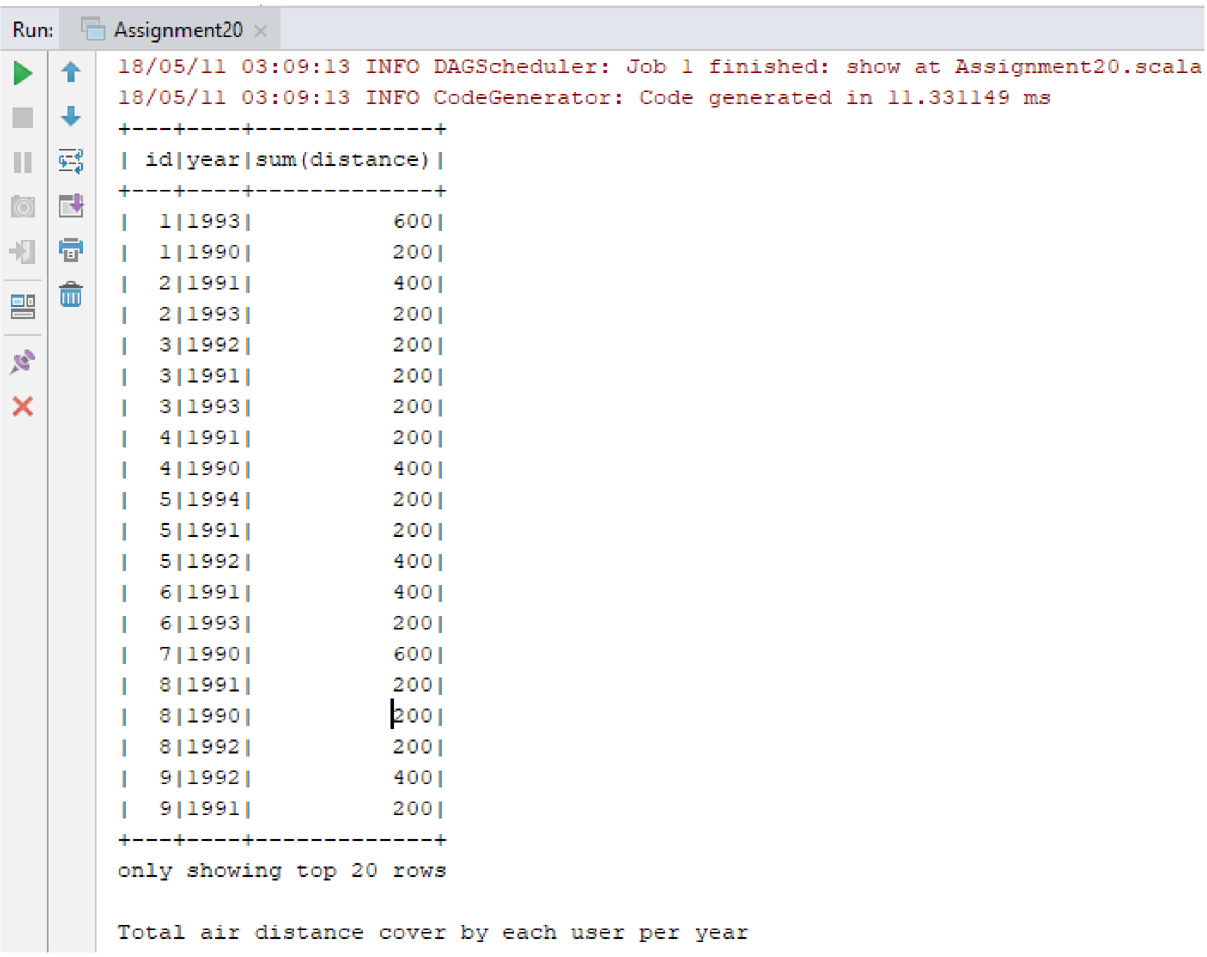
1) What is the distribution of the total number of air-travellers per year.



```
18/05/11 03:09:11 INFO DAGScheduler: Job 0 finished: show at Assignment20.
18/05/11 03:09:11 INFO CodeGenerator: Code generated in 21.682112 ms
+----+-----+
|year|count|
+----+-----+
|1990|    8|
|1991|    9|
|1992|    7|
|1993|    7|
|1994|    1|
+----+-----+

 Total no. of air travelers per year
```

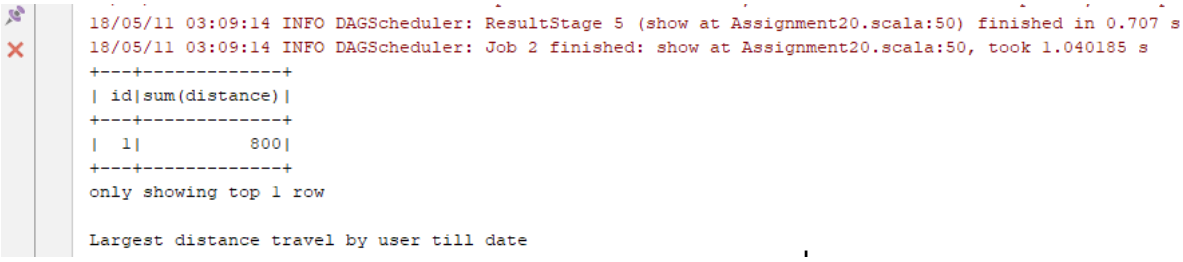
2) What is the total air distance covered by each user per year.



```
18/05/11 03:09:13 INFO DAGScheduler: Job 1 finished: show at Assignment20.scala
18/05/11 03:09:13 INFO CodeGenerator: Code generated in 11.331149 ms
+---+----+-------------+
| id|year|sum(distance)|
+---+----+-------------+
|  1|1993|          600|
|  1|1990|          200|
|  2|1991|          400|
|  2|1993|          200|
|  3|1992|          200|
|  3|1991|          200|
|  3|1993|          200|
|  4|1991|          200|
|  4|1990|          400|
|  5|1994|          200|
|  5|1991|          200|
|  5|1992|          400|
|  6|1991|          400|
|  6|1993|          200|
|  7|1990|          600|
|  8|1991|          200|
|  8|1990|          200|
|  8|1992|          200|
|  9|1992|          400|
|  9|1991|          200|
+---+----+-------------+
only showing top 20 rows

Total air distance cover by each user per year
```

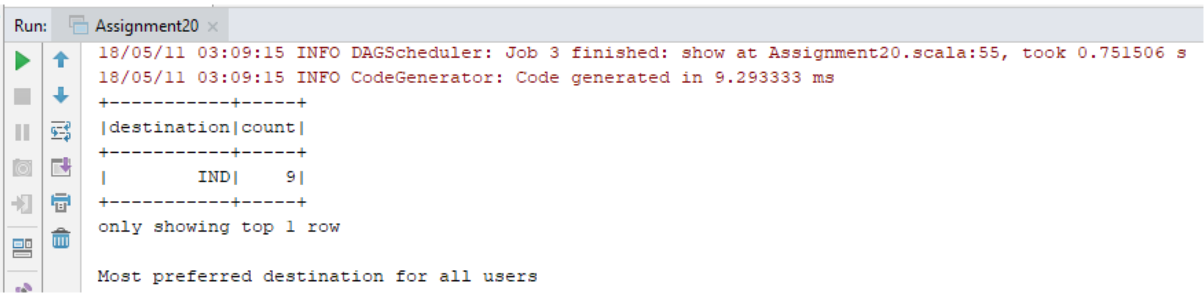
3) Which user has travelled the largest distance until date.



```
18/05/11 03:09:14 INFO DAGScheduler: ResultStage 5 (show at Assignment20.scala:50) finished in 0.707 s
18/05/11 03:09:14 INFO DAGScheduler: Job 2 finished: show at Assignment20.scala:50, took 1.040185 s
+---+-------------+
| id|sum(distance)|
+---+-------------+
|  1|          800|
+---+-------------+
only showing top 1 row

Largest distance travel by user till date
```

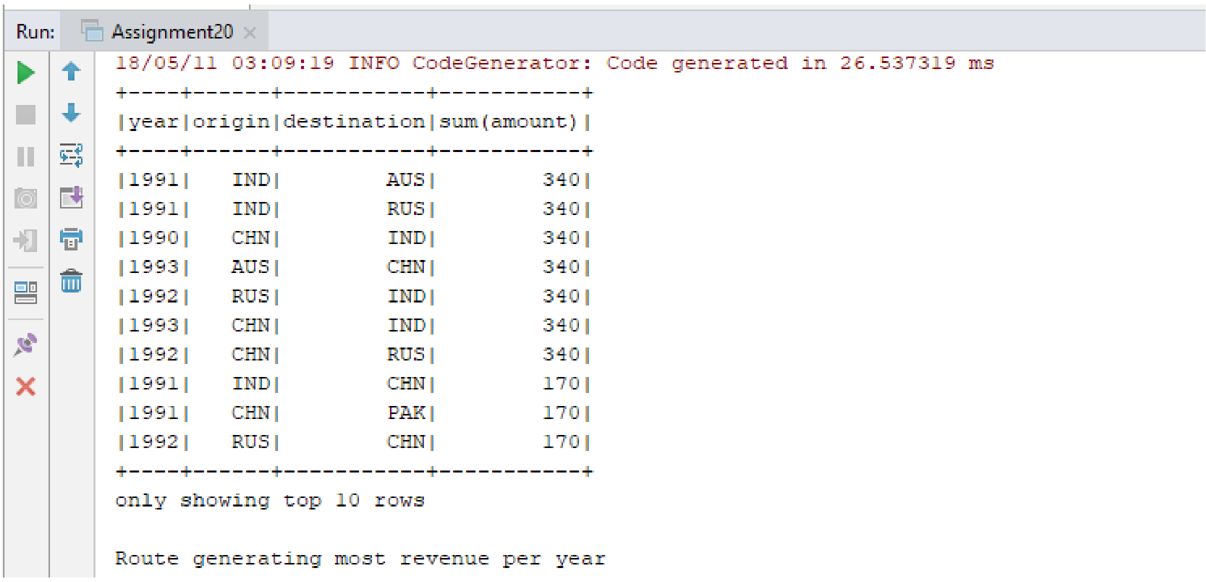
4) What is the most preferred destination for all users.



```
18/05/11 03:09:15 INFO DAGScheduler: Job 3 finished: show at Assignment20.scala:55, took 0.751506 s
18/05/11 03:09:15 INFO CodeGenerator: Code generated in 9.293333 ms
+-----------+-----+
|destination|count|
+-----------+-----+
|        IND|    9|
+-----------+-----+
only showing top 1 row

Most preferred destination for all users
```

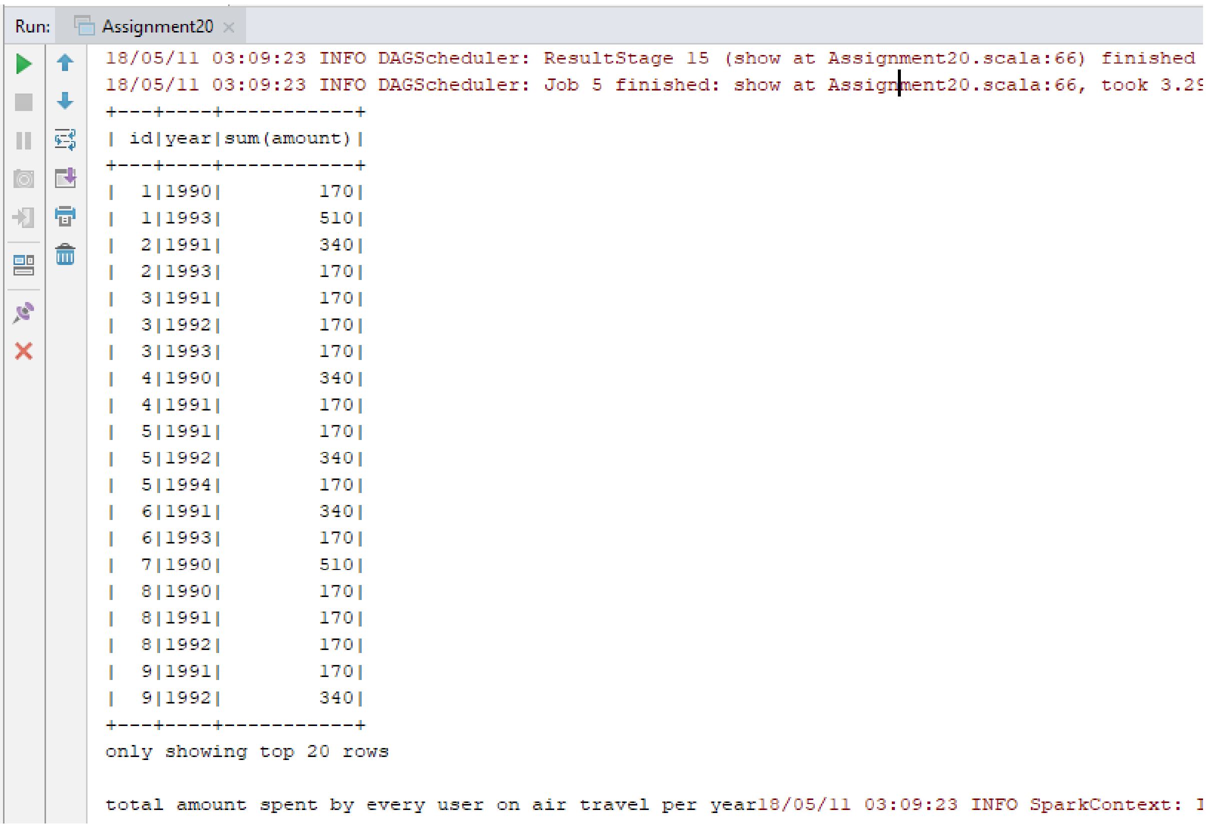
5) Which route is generating the most revenue per year.



```
18/05/11 03:09:19 INFO CodeGenerator: Code generated in 26.537319 ms
+----+------+-----------+-----------+
|year|origin|destination|sum(amount)|
+----+------+-----------+-----------+
|1991|   IND|        AUS|        340|
|1991|   IND|        RUS|        340|
|1990|   CHN|        IND|        340|
|1993|   AUS|        CHN|        340|
|1992|   RUS|        IND|        340|
|1993|   CHN|        IND|        340|
|1992|   CHN|        RUS|        340|
|1991|   IND|        CHN|        170|
|1991|   CHN|        PAK|        170|
|1992|   RUS|        CHN|        170|
+----+------+-----------+-----------+
only showing top 10 rows

Route generating most revenue per year
```

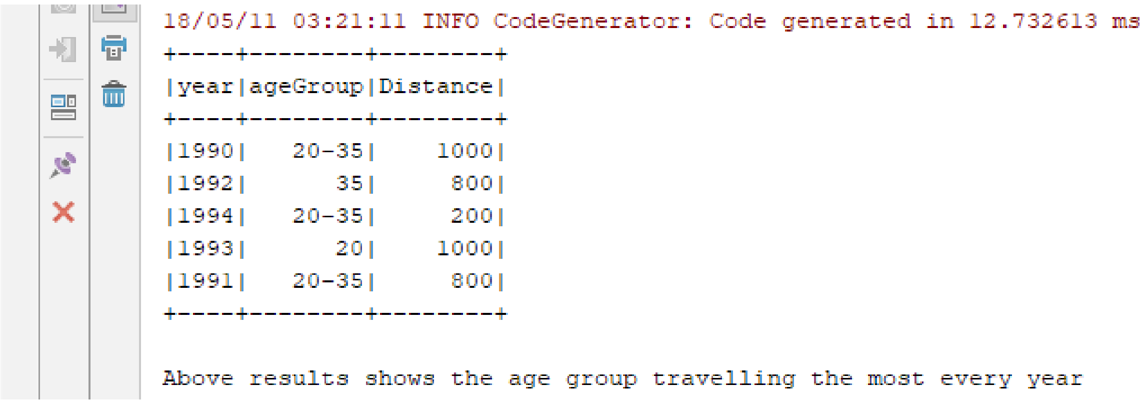
6) What is the total amount spent by every user on air-travel per year



```
18/05/11 03:09:23 INFO DAGScheduler: ResultStage 15 (show at Assignment20.scala:66) finished
18/05/11 03:09:23 INFO DAGScheduler: Job 5 finished: show at Assignment20.scala:66, took 3.2
+---+----+-----------+
| id|year|sum(amount)|
+---+----+-----------+
|  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

total amount spent by every user on air travel per year18/05/11 03:09:23 INFO SparkContext: I
```

7) Considering age groups of < 20, 20-35, 35 >, which age group is travelling the most every year.



```
18/05/11 03:21:11 INFO CodeGenerator: Code generated in 12.732613 ms
+----+--------+--------+
|year|ageGroup|Distance|
+----+--------+--------+
|1990|   20-35|    1000|
|1992|      35|     800|
|1994|   20-35|     200|
|1993|      20|    1000|
|1991|   20-35|     800|
+----+--------+--------+

Above results shows the age group travelling the most every year
```