Input data:

```
import sparkSession.implicits._

val holidaysData = sparkContext.textFile( path = "F:\\PDF Architect\\S20_Dataset_Holidays.txt")
val holidaysDF = holidaysData.map(_.split( regex = ",")).map(x => holidayClass(userId = x(0).toInt, source = x(1),
            destination = x(2),transport_mode = x(3),distance = x(4).toInt, year =  x(5).toInt )).toDF()
holidaysDF.createOrReplaceTempView( viewName = "Holiday_Data")
holidaysDF.show()

val userDetails = sparkContext.textFile( path = "F:\\PDF Architect\\S20_Dataset_User_details.txt")
val userDetailsDF = userDetails.map(_.split( regex = ",")).map(x => user_details(id = x(0).toInt, name = x(1), age = x(2).toInt)).toDF()
userDetailsDF.createOrReplaceTempView( viewName = "User_Details")
userDetailsDF.show()

val transportData = sparkContext.textFile( path = "F:\\PDF Architect\\S20_Dataset_Transport.txt")
val transportDataDF = transportData.map(_.split( regex = ",")).map(x => transportClass(transport_mode = x(0), cost_per_unit = x(1).toInt )).toDF()
transportDataDF.createOrReplaceTempView( viewName = "Transport_Data")
transportDataDF.show()
```

Class20Task

```
18/08/02 22:40:06 INFO DAGScheduler: ResultStage 0 (show at Assignment20.scala:19) finished in 0.428 s
18/08/02 22:40:06 INFO DAGScheduler: Job 0 finished: show at Assignment20.scala:19, took 0.528873 s
+------+------+-----------+--------------+--------+----+
|userId|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|
|     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|
|     7|   CHN|        RUS|      airplane|     200|1990|
|     8|   AUS|        CHN|      airplane|     200|1990|
|     9|   IND|        AUS|      airplane|     200|1991|
|    10|   RUS|        CHN|      airplane|     200|1992|
+------+------+-----------+--------------+--------+----+
only showing top 20 rows

18/08/02 22:40:06 INFO MemoryStore: Block broadcast_2 stored as values in memory (estimated size 216.7 KB, free 349.9 MB)
```

mpilation completed successfully in 8 s 253 ms (a minute ago)

```
18/08/02 22:40:08 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
18/08/02 22:40:08 INFO DAGScheduler: ResultStage 1 (show at Assignment20.scala:24) finished in 0.055 s
18/08/02 22:40:08 INFO DAGScheduler: Job 1 finished: show at Assignment20.scala:24, took 0.065120 s
+---+------+---+
| 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|
+---+------+---+

18/08/02 22:40:08 INFO MemoryStore: Block broadcast_4 stored as values in memory (estimated size 216.7 KB, free 349.7 MB)
18/08/02 22:40:08 INFO MemoryStore: Block broadcast_4_piece0 stored as bytes in memory (estimated size 20.5 KB, free 349.7 MB)
18/08/02 22:40:08 INFO BlockManagerInfo: Added broadcast_4_piece0 in memory on 192.168.1.8:55852 (size: 20.5 KB, free: 350.3 MB)
```

```
Class20Task
18/08/02 22:40:09 INFO TaskSetManager: Finished task 0.0 in stage 2.0 (TID 2) in 23 ms on localhost (executor driver) (1/1)
18/08/02 22:40:09 INFO TaskSchedulerImpl: Removed TaskSet 2.0, whose tasks have all completed, from pool
+--------------+-------------+
18/08/02 22:40:09 INFO DAGScheduler: ResultStage 2 (show at Assignment20.scala:29) finished in 0.044 s
|transport_mode|cost_per_unit|
18/08/02 22:40:09 INFO DAGScheduler: Job 2 finished: show at Assignment20.scala:29, took 0.048275 s
+--------------+-------------+
|      airplane|          170|
18/08/02 22:40:09 INFO SparkContext: Invoking stop() from shutdown hook
|           car|          140|
|         train|          120|
|          ship|          200|
+--------------+-------------+

18/08/02 22:40:09 INFO SparkUI: Stopped Spark web UI at http://192.168.1.8:4040
18/08/02 22:40:09 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/08/02 22:40:09 INFO MemoryStore: MemoryStore cleared
```

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

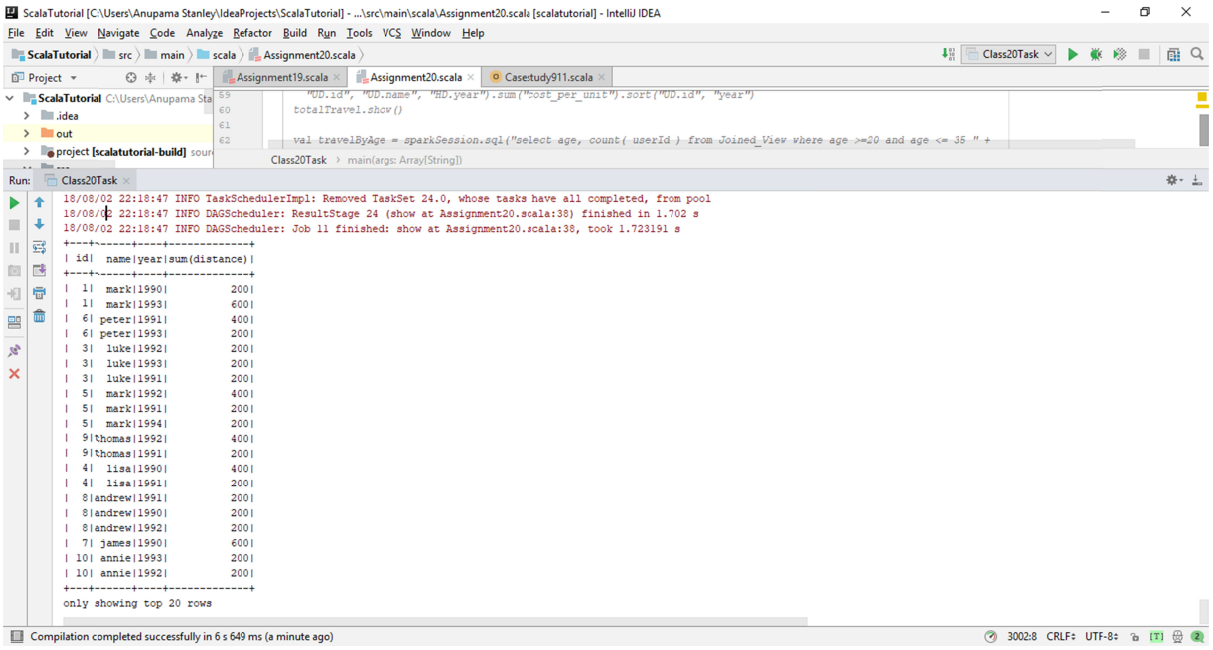
```
val totAirTravellers = sparkSession.sql( sqlText = "select year, count(transport_mode) from Holiday_Data where transport_mode = 'airplane' group by year ")
totAirTravellers.show()
```

```
18/08/02 22:18:43 INFO TaskSchedulerImpl: Removed TaskSet 12.0, whose tasks have all completed, from pool
18/08/02 22:18:43 INFO DAGScheduler: ResultStage 12 (show at Assignment20.scala:32) finished in 0.590 s
18/08/02 22:18:43 INFO DAGScheduler: Job 7 finished: show at Assignment20.scala:32, took 0.601899 s
+----+--------------------+
|year|count(transport_mode)|
+----+--------------------+
|1990|                   8|
|1994|                   1|
|1991|                   9|
|1992|                   7|
|1993|                   7|
+----+--------------------+

18/08/02 22:18:44 INFO CodeGenerator: Code generated in 14.015972 ms
18/08/02 22:18:44 INFO CodeGenerator: Code generated in 113.714354 ms
18/08/02 22:18:44 INFO CodeGenerator: Code generated in 18.039769 ms
```

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

```
val totDistancePerYear = sparkSession.sql( sqlText = "select id, name, year, sum( distance ) from Joined_View group by year,id, name" )
totDistancePerYear.show()
```



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

```
val joinedDF = holidaysDF.join(userDetailsDF, holidaysDF("userId") === userDetailsDF("id"), joinType = "inner")
joinedDF.createOrReplaceTempView( viewName = "Joined_View")
```

```
val maxTravelUser = sparkSession.sql( sqlText = "select name, year, sum( distance ) from Joined_View " +
                    "group by id, year, name order by sum( distance ) desc").take(1).mkString(",")
println(maxTravelUser)
```

```
18/08/02 22:09:46 INFO TaskSetManager: Finished task 174.0 in stage 27.0 (TID 532) in 17 ms on localhost (e
18/08/02 22:09:46 INFO TaskSchedulerImpl: Removed TaskSet 27.0, whose tasks have all completed, from pool
18/08/02 22:09:46 INFO DAGScheduler: ResultStage 27 (take at Assignment20.scala:41) finished in 2.436 s
18/08/02 22:09:46 INFO DAGScheduler: Job 12 finished: take at Assignment20.scala:41, took 2.784729 s
[mark,1993,600]
18/08/02 22:09:46 INFO CodeGenerator: Code generated in 10.383991 ms
18/08/02 22:09:46 INFO SparkContext: Invoking stop() from shutdown hook
18/08/02 22:09:46 INFO SparkUI: Stopped Spark web UI at http://192.168.1.8:4040
18/08/02 22:09:46 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
```

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

```
val maxTravelUser = sparkSession.sql( sqlText = "select name, year, sum( distance ) from Joined_View " +
                    "group by id, year, name order by sum( distance ) desc").take(1).mkString(",")
println(maxTravelUser)
```

```
18/08/02 22:06:36 INFO Executor: Finished task 183.0 in stage 29.0 (TID 733). 3951 bytes result sent to driver
18/08/02 22:06:36 INFO TaskSetManager: Finished task 183.0 in stage 29.0 (TID 733) in 7 ms on localhost (executor
18/08/02 22:06:36 INFO TaskSchedulerImpl: Removed TaskSet 29.0, whose tasks have all completed, from pool
18/08/02 22:06:36 INFO DAGScheduler: ResultStage 29 (show at Assignment20.scala:46) finished in 1.002 s
18/08/02 22:06:36 INFO DAGScheduler: Job 13 finished: show at Assignment20.scala:46, took 1.188584 s
+-----------+------------------+
|destination|count(destination)|
+-----------+------------------+
|        IND|                 9|
+-----------+------------------+
only showing top 1 row

18/08/02 22:06:36 INFO CodeGenerator: Code generated in 15.909239 ms
18/08/02 22:06:36 INFO SparkContext: Invoking stop() from shutdown hook
```

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

```
47
48  val routesDF = holidaysDF.withColumn( colName = "Route", struct( colName = "destination", colNames = "source" )).toDF()
49  routesDF.createOrReplaceTempView( viewName = "Routes")
50  routesDF.show()
51
52  val maxRoute = sparkSession.sql( sqlText = "select Route, Sum( cost_per_unit ) from Routes JOIN Transport_Data on " +
53                  "Routes.transport_mode = Transport_Data.transport_mode group by Route order by Sum( cost_per_unit) desc")
54
55  maxRoute.show( numRows = 1)
56
```

```
18/08/02 20:06:06 INFO DAGScheduler: ResultStage 34 (show at Assignment20.scala:55) finished in 1.952 s
18/08/02 20:06:06 INFO DAGScheduler: Job 15 finished: show at Assignment20.scala:55, took 10.903583 s
18/08/02 20:06:06 INFO CodeGenerator: Code generated in 28.356746 ms
+----------+------------------+
|     Route|sum(cost_per_unit)|
+----------+------------------+
|[IND, CHN]|               680|
+----------+------------------+
only showing top 1 row

18/08/02 20:06:06 INFO SparkContext: Invoking stop() from shutdown hook
```

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

```
maxRoute.show(1) */

val totalTravel = holidaysDF.as( alias = "HD").join(userDetailsDF.as( alias = "UD"), joinExprs = $"HD.userId" === $"UD.id" ).
  join(transportDataDF.as( alias = "TD"),  joinExprs = $"HD.transport_mode" === $"TD.transport_mode").groupBy(
   col1 = "UD.id",  cols = "UD.name", "HD.year").sum( colNames = "cost_per_unit").sort( sortCol = "UD.id",  sortCols = "year")
totalTravel.show()
```

```
18/08/02 22:28:17 INFO TaskSchedulerImpl: Removed TaskSet 30.0, whose tasks have all completed, from pool
18/08/02 22:28:17 INFO DAGScheduler: ResultStage 30 (show at Assignment20.scala:60) finished in 1.460 s
18/08/02 22:28:17 INFO DAGScheduler: Job 12 finished: show at Assignment20.scala:60, took 14.689175 s
+---+------+----+------------------+
| id|  name|year|sum(cost_per_unit)|
+---+------+----+------------------+
|  1|  mark|1990|               170|
|  1|  mark|1993|               510|
|  2|  john|1991|               340|
|  2|  john|1993|               170|
|  3|  luke|1991|               170|
|  3|  luke|1992|               170|
|  3|  luke|1993|               170|
|  4|  lisa|1990|               340|
|  4|  lisa|1991|               170|
|  5|  mark|1991|               170|
|  5|  mark|1992|               340|
|  5|  mark|1994|               170|
|  6| peter|1991|               340|
|  6| peter|1993|               170|
|  7| james|1990|               510|
|  8|andrew|1990|               170|
|  8|andrew|1991|               170|
|  8|andrew|1992|               170|
|  9|thomas|1991|               170|
|  9|thomas|1992|               340|
+---+------+----+------------------+
only showing top 20 rows

18/08/02 22:28:17 INFO CodeGenerator: Code generated in 20.594713 ms
```

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

```
val travelByAge = sparkSession.sql( sqlText = "select age, count( userId ) from Joined_View where age >=20 and age <= 35 " +
                  "group by age, userId order by count( userId ) desc")

travelByAge.show()
}
```

```
18/08/02 20:01:24 INFO TaskSchedulerImpl: Removed TaskSet 4.
18/08/02 20:01:24 INFO DAGScheduler: ResultStage 43 (show at
18/08/02 20:01:24 INFO DAGScheduler: Job 17 finished: show
+---+-------------+
|age|count(userId)|
+---+-------------+
| 25|            4|
| 22|            3|
| 21|            3|
| 27|            3|
+---+-------------+
```