

Assignment 20

```
import org.apache.spark.sql.SparkSession

object Assignment_20 {

  //Create a case class globally to be used inside the main method
  case class Transport(mode: String, amount: Int)
  case class User(id:Int, name:String,age :Int)
  case class Holiday (t_id:Int,Source:String,Destination: String,mode: String, distance:Int,
year:Int)

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

    //Create a spark session object
    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark SQL basic example")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()

    println("Spark Session Object created")

    import spark.implicits._          //to convert RDD into DataFrame
    val hl = spark.sparkContext
      .textFile("E:\\Avani\\Acadgild\\Assignment_20\\Data\\Holidays.txt")
      .map(_._split(","))
      .map(attributes => Holiday(attributes(0).trim.toInt,
attributes(1),attributes(2),attributes(3),attributes(4).trim.toInt,attributes(5).trim.toInt))
      .toDF()

    hl.show()

    println("Holiday data")
  }
}
```

```

val trans = spark.sparkContext
    .textFile("E:\\Avani\\Acadgild\\Assignment_20\\Data\\Transport.txt")
    .map(_ .split(", "))
    .map(attributes => Transport(attributes(0), attributes(1).trim.toInt))
    .toDF()
trans.show()

println("Transport data")

```

```

val user = spark.sparkContext
    .textFile("E:\\Avani\\Acadgild\\Assignment_20\\Data\\User.txt")
    .map(_ .split(", "))
    .map(attributes => User(attributes(0).trim.toInt, attributes(1), attributes(2).trim.toInt))
    .toDF()
user.show()

println("User data")

//Creating temporary tables
hl.registerTempTable("holiday")
user.registerTempTable("people")
trans.registerTempTable("transport")

```

a. What is the distribution of the total number of air-travellers per year?

```

val total = spark.sql("SELECT year,COUNT(*) from holiday group by year")
total.show()
println("Task1 output")

```

b. What is the total air distance covered by each user per year?

```

val dist = spark.sql("SELECT t_id,year,sum(distance) from holiday group by t_id,year
order by t_id,year")

```

```
dist.show()
println("Task2 output")
```

c. Which user has travelled the largest distance till date?

```
val max_d = spark.sql("SELECT t_id,max(d) from (SELECT t_id, sum(distance) d from
holiday group by t_id) group by t_id")
max_d.show(1)
println("Task3 output")
```

d. What is the most preferred destination for all users?

```
val ct=spark.sql("SELECT Destination from holiday group by 1 having count
(Destination)= (select max(c) from (SELECT Destination,count(Destination) c from
holiday group by Destination))")

ct.show()

println("Task4 output")
```

e. Which route is generating the most revenue per year?

```
val rev = spark.sql("SELECT h.source, h.destination, sum(h.distance*t.amount) revenue
from holiday h join transport t ON h.mode=t.mode group by h.source,h.destination")

val x= rev.sort($"revenue".desc).first

println("most revenue generated route-->" + x)
```

f. What is the total amount spent by every user on air-travel per year?

```
val amt = spark.sql("SELECT h.t_id, h.year,sum(t.amount) from holiday h join transport t
ON h.mode=t.mode group by h.t_id,h.year")

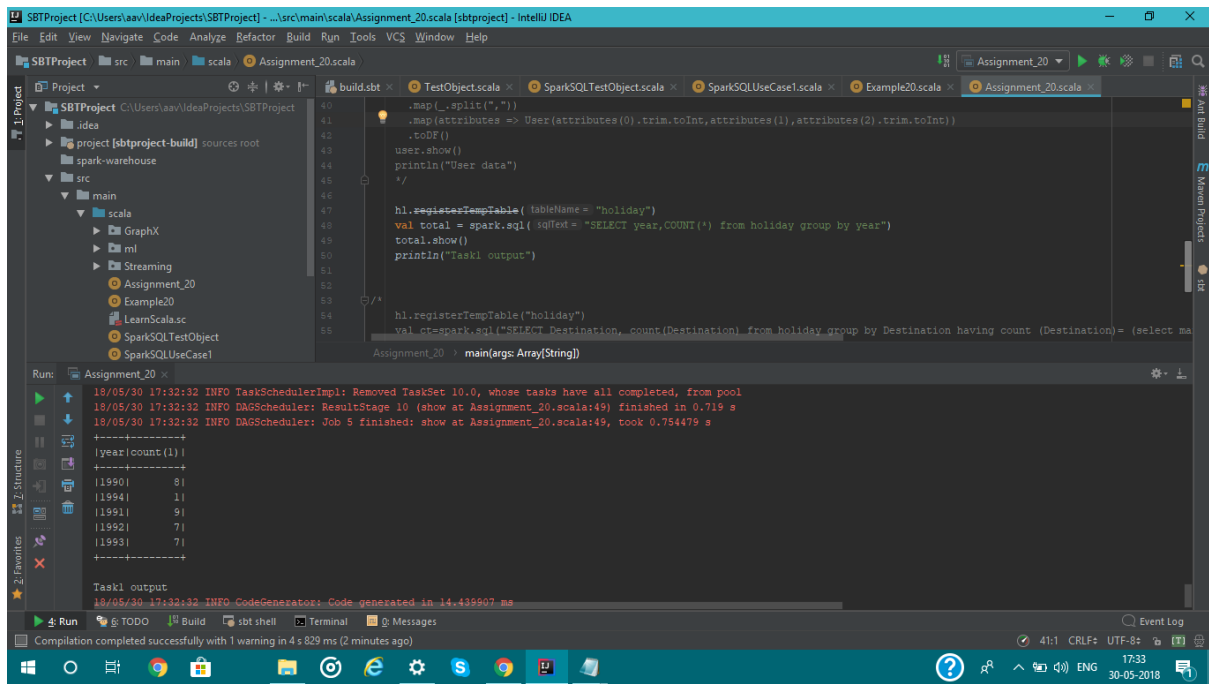
amt.show(32)

println("Task6")
```

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

```
val y = spark.sql("select IF((p.age) <20, 'Kid', IF((p.age) between 20 and 35, 'Adult', 'Old'))  
age_1, count(age) a from people p join holiday h ON h.t_id = p.id group by p.age")  
val age_group= y.sort($"a".desc).first()  
println(age_group)
```

Screenshot



SBTProject [C:\Users\aa\IdeaProjects\SBTProject] - \src\main\scala\Assignment_20.scala [sbtpj] - IntelliJ IDEA

```
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
```

SBTProject src main scala Assignment_20.scala

Project Structure: SBTProject, .idea, project [sbtpj-build] sources root, spark-warehouse, src, main, scala, GraphX, ml, Streaming, Assignment_20, Example20, LearnScala.sc, SparkSQLTestObject, SparkSQLUseCase1

```
40 .map(_.split(","))
41 .map(attributes => User(attributes(0).trim.toInt, attributes(1), attributes(2).trim.toInt))
42 .toDF()
43 user.show()
44 println("User data")
45 */
46
47 hl.registerTempTable("holiday")
48 val total = spark.sql("SELECT year, COUNT(*) from holiday group by year")
49 total.show()
50 println("Task1 output")
51
52
53
54 hl.registerTempTable("holiday")
55 val ct=spark.sql("SELECT Destination, count(Destination) from holiday group by Destination having count (Destination) = (select ma
56
57 Assignment_20 main(args: Array[String])
```

Run: Assignment_20

```
18/05/30 17:32:32 INFO TaskSchedulerImpl: Removed TaskSet 10.0, whose tasks have all completed, from pool
18/05/30 17:32:32 INFO DAGScheduler: ResultStage 10 (show at Assignment_20.scala:49) finished in 0.719 s
18/05/30 17:32:32 INFO DAGScheduler: Job 5 finished: show at Assignment_20.scala:49, took 0.754479 s
```

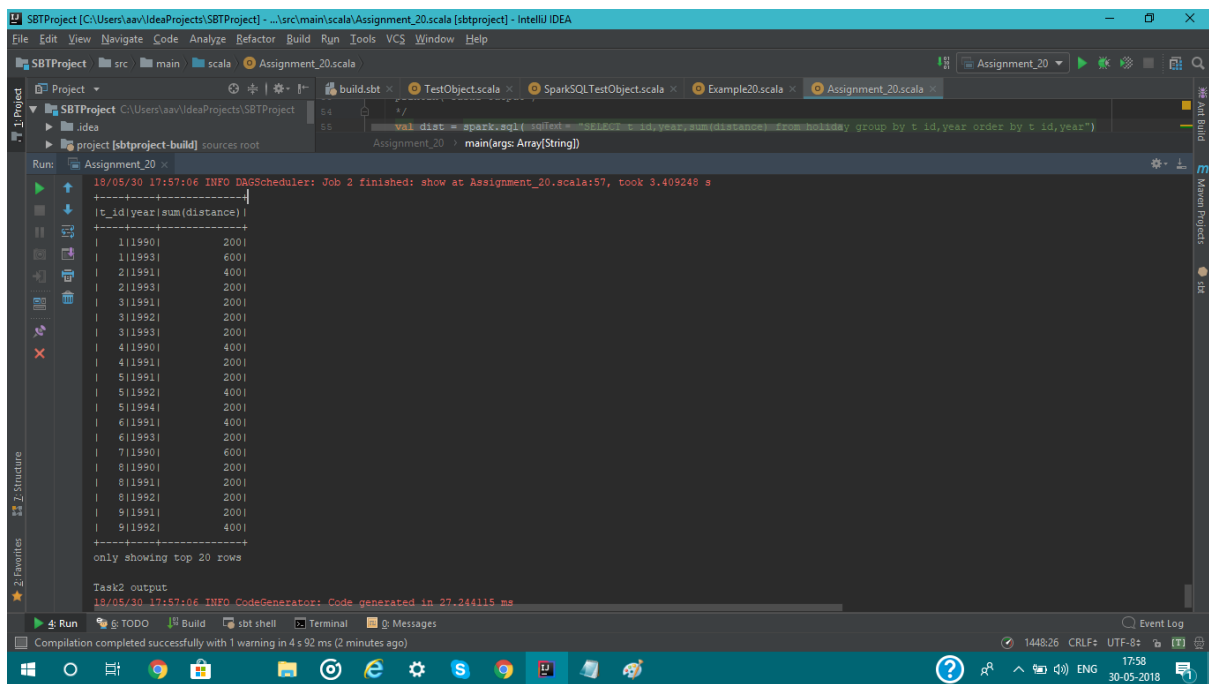
year	count(1)
119901	81
119941	11
119911	91
119921	71
119931	71

Task1 output

```
18/05/30 17:32:32 INFO CodeGenerator: Code generated in 14.439507 ms
```

Compilation completed successfully with 1 warning in 4 s 829 ms (2 minutes ago)

a.



SBTProject [C:\Users\aa\IdeaProjects\SBTProject] - \src\main\scala\Assignment_20.scala [sbtpj] - IntelliJ IDEA

```
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
```

SBTProject src main scala Assignment_20.scala

Project Structure: SBTProject, .idea, project [sbtpj-build] sources root, spark-warehouse, src, main, scala, GraphX, ml, Streaming, Assignment_20, Example20, LearnScala.sc, SparkSQLTestObject, SparkSQLUseCase1

```
54
55 val dist = spark.sql("select t.id,year,sum(distance) from holiday group by t.id,year order by t.id,year")
56
57 Assignment_20 main(args: Array[String])
```

Run: Assignment_20

```
18/05/30 17:57:06 INFO DAGScheduler: Job 2 finished: show at Assignment_20.scala:57, took 3.409248 s
```

t_id	year	sum(distance)
1	119901	2001
1	119931	6001
2	119911	4001
2	119931	2001
3	119911	2001
3	119921	2001
3	119931	2001
4	119901	4001
4	119911	2001
5	119911	2001
5	119921	4001
5	119941	2001
6	119911	4001
6	119931	2001
7	119901	6001
8	119901	2001
8	119911	2001
8	119921	2001
9	119911	2001
9	119921	4001

Task2 output

```
18/05/30 17:57:06 INFO CodeGenerator: Code generated in 27.244115 ms
```

Compilation completed successfully with 1 warning in 4 s 92 ms (2 minutes ago)

b.

```
1 //val total = spark.sql("SELECT year,COUNT(*) from holiday group by year")
2 total.show()
3 println("Task1 output")
4
5 val dist = spark.sql("SELECT t_id,year,sum(distance) from holiday group by t_id,year order by t_id,year")
6 dist.show()
7 println("Task2 output")
8
9 //
10 val max_d = spark.sql( sqlText = "SELECT t_id,max(d) from (SELECT t_id, sum(distance) d from holiday group by t_id) group by t_id")
11 max_d.show( numRows = 1)
12 println("Task3 output")
13
14 //
15 val ct=spark.sql("SELECT Destination from holiday group by 1 having count (Destination)= (select max(c) from (SELECT Destination,count(Destination) c from holiday group by Desti
16 ct.show()
17 println("Task4 output")
```

Run: Assignment_20 → main(args: Array[String])

```
18/05/30 20:32:47 INFO DAGScheduler: Job 3 finished: show at Assignment_20.scala:60, took 1.536005 s
18/05/30 20:32:47 INFO BlockManagerInfo: Removed broadcast_7_piece0 on 192.168.5.112:50587 in memory (size: 16.5 KB, free: 899.7 MB)
18/05/30 20:32:47 INFO CodeGenerator: Code generated in 15.142509 ms
18/05/30 20:32:47 INFO SparkContext: Invoking stop() from shutdown hook
+-----+
|t_id|max(d)|
+-----+
| 11 | 800 |
+-----+
only showing top 1 row

Task3 output
18/05/30 20:32:47 INFO SparkUI: Stopped Spark web UI at http://192.168.5.112:4040
18/05/30 20:32:47 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/30 20:32:47 INFO MemoryStore: MemoryStore cleared
18/05/30 20:32:47 INFO BlockManager: BlockManager stopped
```

c.

```
1 //
2 //val max_d = spark.sql("SELECT t_id from holiday group by t_id having max(distance) = (SELECT t_id, max(d) from (SELECT t_id, sum(dist
3 max_d.show()
4 println("Task3 output")
5
6 //
7 val ct=spark.sql( sqlText = "SELECT Destination from holiday group by 1 having count (Destination)= (select max(c) from (SELECT Destinati
8 ct.show()
9 println("Task4 output")
```

Run: Assignment_20 → main(args: Array[String])

```
18/05/30 19:31:51 INFO DAGScheduler: Job 7 finished: show at Assignment_20.scala:64, took 0.603383 s
18/05/30 19:31:51 INFO CodeGenerator: Code generated in 16.550279 ms
+-----+
|Destination|
+-----+
| IND |
+-----+

Task4 output
18/05/30 19:31:51 INFO SparkContext: Invoking stop() from shutdown hook
18/05/30 19:31:51 INFO SparkUI: Stopped Spark web UI at http://192.168.5.112:4040
```

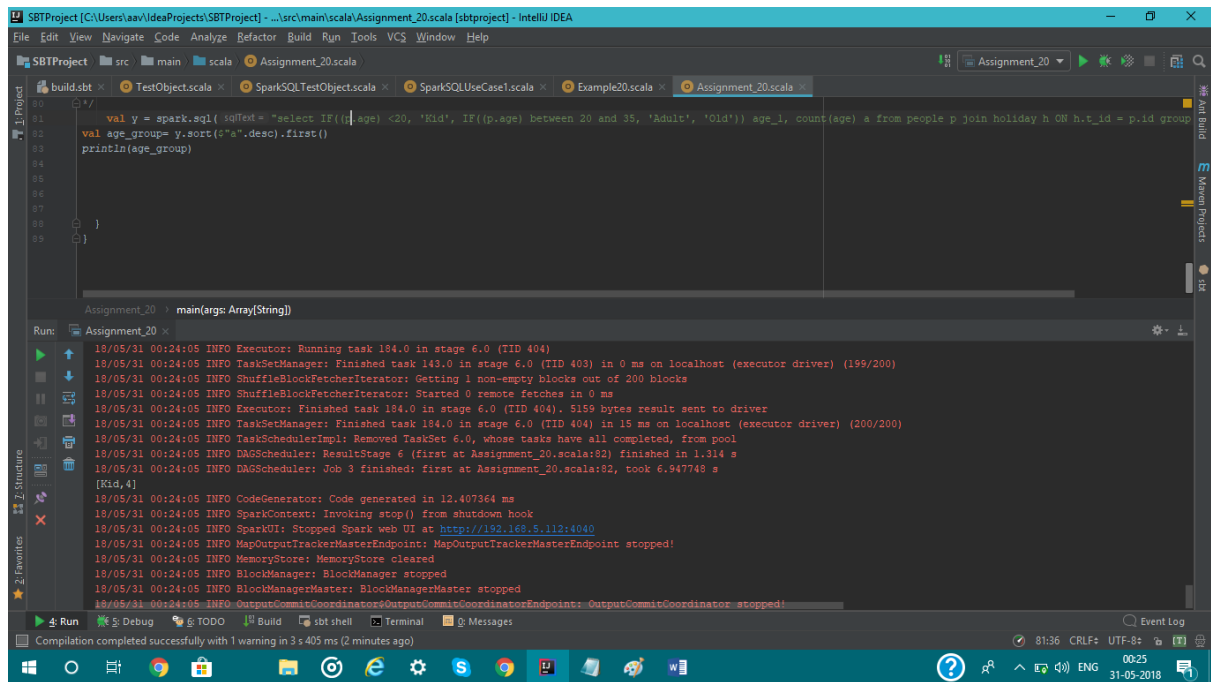
d.

```
1  SBTProject [C:\Users\aa\IdeaProjects\SBTProject] - \src\main\scala\Assignment_20.scala [sbtproject] - IntelliJ IDEA
2  File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
3  SBTProject src main scala Assignment_20.scala
4  build.sbt TestObject.scala SparkSQLTestObject.scala SparkSQLUseCase1.scala Example20.scala Assignment_20.scala
5  64 ct.show()
6  65 println("Task4 output")
6  66
6  67 val rev = spark.sql("SELECT h.source, h.destination, sum(h.distance*t.amount) revenue from holiday h join transport t ON h.mode=t.mode group by h.source,h.destination")
6  68 val x= rev.sort($"revenue".desc).first
6  69 println("most revenue generated route-->" + x)
6  70
6  71
6  72
7  Assignment_20 main(args: Array[String])
8  Run: Assignment_20 x
9  18/05/30 22:36:18 INFO ShuffleBlockFetcherIterator: Started 0 remote fetches in 0 ms
9  18/05/30 22:36:18 INFO Executor: Finished task 195.0 in stage 3.0 (TID 401). 5333 bytes result sent to driver
9  18/05/30 22:36:18 INFO TaskSetManager: Finished task 195.0 in stage 3.0 (TID 401) in 15 ms on localhost (executor driver) (200/200)
9  18/05/30 22:36:18 INFO TaskSchedulerImpl: Removed TaskSet 3.0, whose tasks have all completed, from pool
9  18/05/30 22:36:18 INFO DAGScheduler: ResultStage 3 (first at Assignment_20.scala:69) finished in 1.262 s
9  18/05/30 22:36:18 INFO DAGScheduler: Job 0 finished: first at Assignment_20.scala:69, took 6.926237 s
9  18/05/30 22:36:18 INFO CodeGenerator: Code generated in 16.330502 ms
9  most revenue generated route-->[CHN,IND,136000]
9  18/05/30 22:36:18 INFO SparkContext: Invoking stop() from shutdown hook
9  18/05/30 22:36:18 INFO SparkUI: Stopped Spark web UI at http://159.254.29.30:4040
9  18/05/30 22:36:18 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
9  18/05/30 22:36:18 INFO MemoryStore: MemoryStore cleared
9  18/05/30 22:36:18 INFO BlockManager: BlockManager stopped
9  18/05/30 22:36:18 INFO BlockManagerMaster: BlockManagerMaster stopped
9  18/05/30 22:36:18 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
9  18/05/30 22:36:18 INFO SparkContext: Successfully stopped SparkContext
9  18/05/30 22:36:18 INFO ShutdownHookManager: Shutdown hook called
9  18/05/30 22:36:18 INFO ShutdownHookManager: Deleting directory C:\Users\aa\AppData\Local\Temp\spark-0568e07d-c4e4-47a2-b152-122e38c999d6
9  Process finished with exit code 0
10 Run Debug TODO Build sbt shell Terminal Messages
11 Compilation completed successfully with 1 warning in 3 s 579 ms (4 minutes ago)
12 68.5 CRLF: UTF-8: 22:40 30-05-2018
```

e.

```
1  SBTProject [C:\Users\aa\IdeaProjects\SBTProject] - \src\main\scala\Assignment_20.scala [sbtproject] - IntelliJ IDEA
2  File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
3  SBTProject src main scala Assignment_20.scala
4  build.sbt TestObject.scala SparkSQLTestObject.scala SparkSQLUseCase1.scala Example20.scala Assignment_20.scala
5  76 rev_user.show()
6  Assignment_20 main(args: Array[String])
7  Run: Assignment_20 x
8  +-----+
8  |t_id|year|sum(amount)|
8  +-----+
8  | 3|1991|      170|
8  | 6|1993|      170|
8  | 3|1992|      170|
8  | 7|1990|      510|
8  |10|1993|      170|
8  | 6|1991|      340|
8  | 2|1991|      340|
8  | 4|1991|      170|
8  | 5|1991|      170|
8  | 5|1994|      170|
8  | 8|1991|      170|
8  |11|1990|      170|
8  | 5|1992|      340|
8  | 4|1990|      340|
8  | 3|1993|      170|
8  |10|1990|      170|
8  | 2|1993|      170|
8  |11|1993|      510|
8  | 9|1991|      170|
8  | 9|1992|      340|
8  | 8|1990|      170|
8  |10|1992|      170|
8  | 8|1992|      170|
8  +-----+
8  Task6 output
8  18/05/30 23:57:29 INFO SparkUI: Stopped Spark web UI at http://192.168.5.112:4040
9  Run Debug TODO Build sbt shell Terminal Messages
10 Compilation completed successfully with 1 warning in 3 s 901 ms (a minute ago)
11 3205:24 CRLF: UTF-8: 23:58 30-05-2018
```

f.



g.