Assignment 21

```
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions._
object Assignment_21 {

//declaring case class Sports
  case class Sports
(f_name:String,l_name:String,sports:String,medal_type:String,age:Int,year:Int,country:String
)

//creating spark session object
  def main(args: Array[String]): Unit = {
    println("hey scala")
    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark SQL Assignment 21")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()

    println("Spark Session Object created")


//loading of data from the text file
    val data =
spark.sparkContext.textFile("E:\\Avani\\Acadgild\\Assignment_21\\Sports_data.txt");
    val header=data.first()
    val data1 = data.filter(x => x != header)
     val num=println("Sports_Data->>"+data1.count())
    println("removed header")


//convert RDD into DataFrame, 's'
    import spark.implicits._
    val s =data1.map(x=>x.split(",")).map(x =>
Sports(x(0),x(1),x(2),x(3),x(4).trim.toInt,x(5).trim.toInt,x(6)))
      .toDF()
    s.show(25)
    println("Sports data")


//creating a temporary table
    s.registerTempTable("sport")
    println("temp table created")
```

Task 1

Using spark-sql, Find:

a. What are the total number of gold medal winners every year?

```
val gold=spark.sql("SELECT year,count(*) from sport where medal_type='gold' group by
year")
gold.show
println("Task1.1 output")
```

b. How many silver medals have been won by USA in each sport?

```
val silver=spark.sql("SELECT sports,count(*) from sport where medal_type='silver' and
country='USA' group by sports")

 silver.show

println("Task1.2 output")
```

Task 2

Using udfs on dataframe

a. Change firstname, lastname columns into

Mr.first_two_letters_of_firstname<space>lastname

for example - michael, phelps becomes Mr.mi phelps

```
//creating a nameudf method

def nameudf = ((f_name:String,
l_name:String)=>"Mr.".concat(f_name.substring(0,2)).concat("").concat(l_name:String))

//registering the method as a udf and storing it into variable, fullname
val fullname= spark.sqlContext.udf.register("fullname",nameudf)

//sql query to fetch the desired output
val name = spark.sql("SELECT fullname(f_name, l_name) FROM sport")
name.show(25)
println("Task 2.1 output")
```

b. Add a new column called ranking using udfs on dataframe, where :

gold medalist, with age >= 32 are ranked as pro

gold medalists, with age <= 31 are ranked amateur

silver medalist, with age >= 32 are ranked as expert

silver medalists, with age <= 31 are ranked rookie

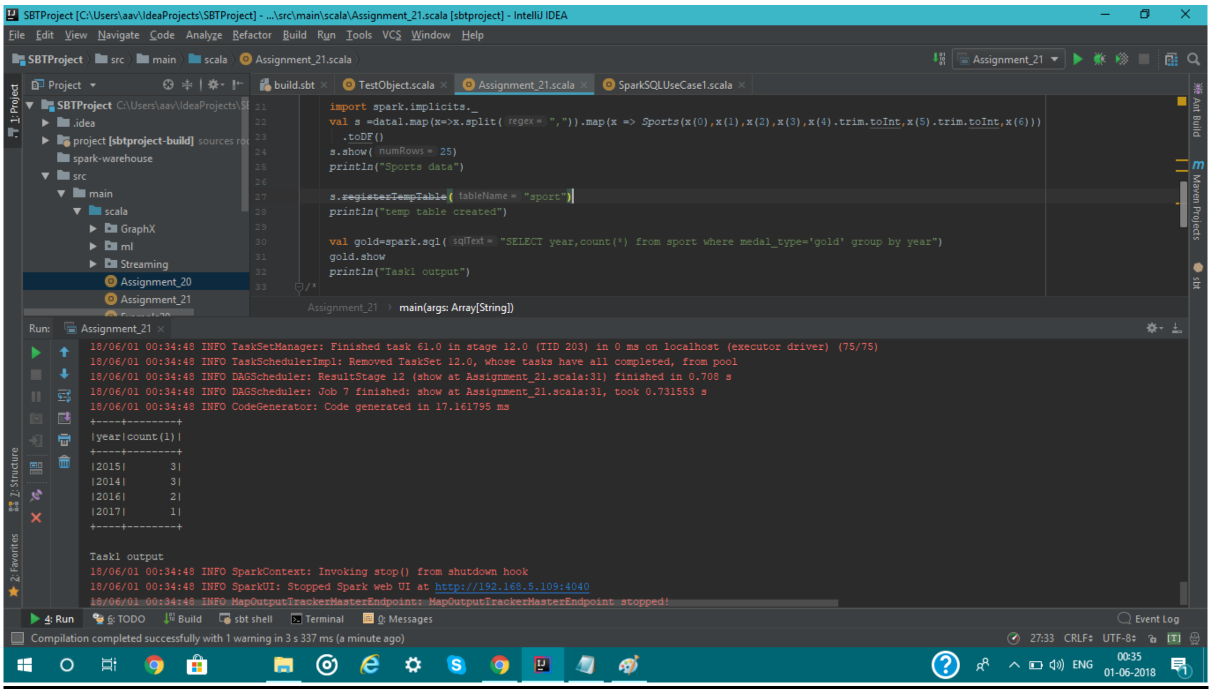
```
//declaring a Ranking udf along with the mentioned conditions
def Ranking = udf((medal_type: String, age: Int) => (medal_type,age) match
  {
    case (medal_type,age) if medal_type == "gold" && age >= 32 => "Pro"
    case (medal_type,age) if medal_type == "gold" && age <= 32 => "amateur"
    case (medal_type,age) if medal_type == "silver" && age >= 32 => "expert"
    case (medal_type,age) if medal_type == "silver" && age <= 32 => "rookie"
  })
//adding it (along with the parameters) as a column to the 's' dataframe, declared above
  val r = s.withColumn("Rank", Ranking(s("medal_type"),s("age")))
r.show(25)
  println("Task 2.2 output")
```

Complete code

```
import org.apache.spark.sql.SparkSession
import org.apache.spark.sql.functions._
object Assignment_21 {
  case class Sports
(f_name:String,l_name:String,sports:String,medal_type:String,age:Int,year:Int,country:String
)
  def main(args: Array[String]): Unit = {
    println("hey scala")
    val spark = SparkSession
      .builder()
      .master("local")
      .appName("Spark SQL Assignment 21")
      .config("spark.some.config.option", "some-value")
      .getOrCreate()

    println("Spark Session Object created")

    val data =
spark.sparkContext.textFile("E:\\Avani\\Acadgild\\Assignment_21\\Sports_data.txt");
    val header=data.first()
    val data1 = data.filter(x => x != header)
     val num=println("Sports_Data->>"+data1.count())
    println("removed header")

    import spark.implicits._
    val s =data1.map(x=>x.split(",")).map(x =>
Sports(x(0),x(1),x(2),x(3),x(4).trim.toInt,x(5).trim.toInt,x(6)))
      .toDF()
    s.show(25)
    println("Sports data")

    s.registerTempTable("sport")
    println("temp table created")

    val gold=spark.sql("SELECT year,count(*) from sport where medal_type='gold' group by
year")
    gold.show
    println("Task1.1 output")


        val silver=spark.sql("SELECT sports,count(*) from sport where medal_type='silver' and
country='USA' group by sports")
```

```
        silver.show
        println("Task1.2 output")

    def nameudf = ((f_name:String,
l_name:String)=>"Mr.".concat(f_name.substring(0,2)).concat("").concat(l_name:String))
    val fullname= spark.sqlContext.udf.register("fullname",nameudf)
    val name = spark.sql("SELECT fullname(f_name, l_name) FROM sport")
    name.show(25)
    println("Task 2.1 output")


    def Ranking = udf((medal_type: String, age: Int) => (medal_type,age) match
    {
      case (medal_type,age) if medal_type == "gold" && age >= 32 => "Pro"
      case (medal_type,age) if medal_type == "gold" && age <= 32 => "amateur"
      case (medal_type,age) if medal_type == "silver" && age >= 32 => "expert"
      case (medal_type,age) if medal_type == "silver" && age <= 32 => "rookie"
    })
  // val RANK= spark.sqlContext.udf.register("RANK",Ranking)
    val r = s.withColumn("Rank", Ranking(s("medal_type"),s("age")))
r.show(25)
    println("Task 2.2 output")

  }

}
```

Screenshots



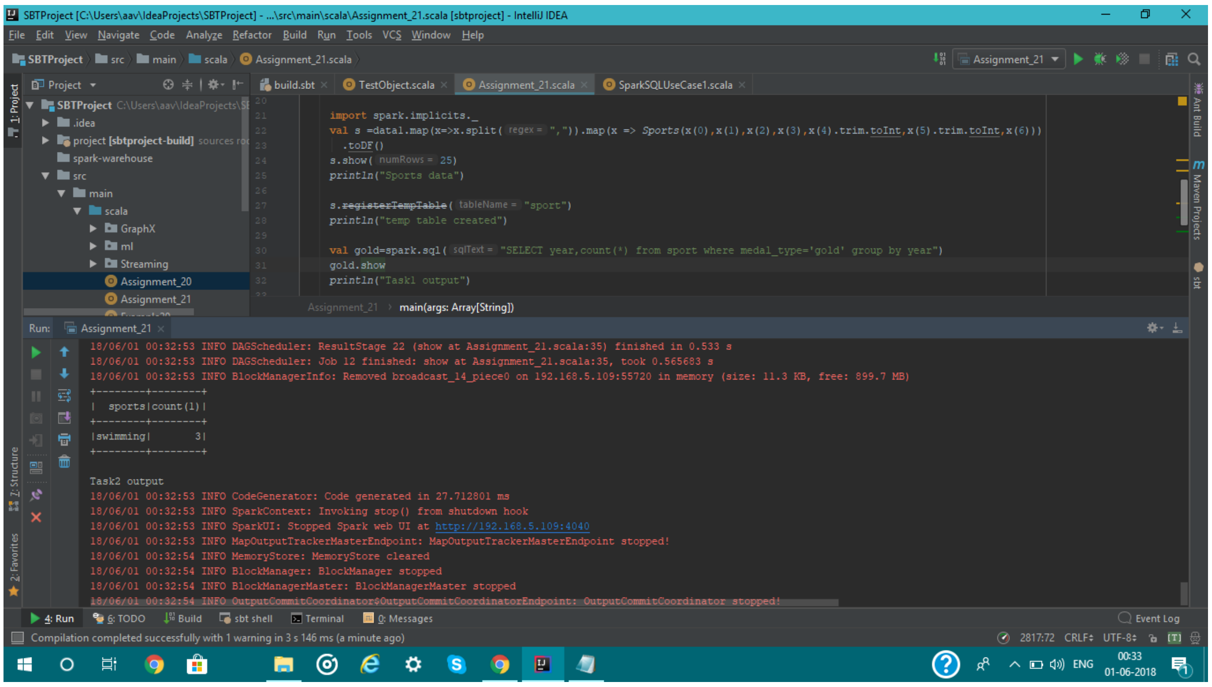
SBTProject [C:\Users\aav\IdeaProjects\SBTProject] - ...\src\main\scala\Assignment_21.scala [sbtproject] - IntelliJ IDEA

```scala
import spark.implicits._
val s =datal.map(x=>x.split(",")).map(x => Sports(x(0),x(1),x(2),x(3),x(4).trim.toInt,x(5).trim.toInt,x(6)))
  .toDF()
s.show(25)
println("Sports data")

s.registerTempTable("sport")
println("temp table created")

val gold=spark.sql("SELECT year,count(*) from sport where medal_type='gold' group by year")
gold.show
println("Taskl output")
/*
```

```
18/06/01 00:34:48 INFO TaskSetManager: Finished task 61.0 in stage 12.0 (TID 203) in 0 ms on localhost (executor driver) (75/75)
18/06/01 00:34:48 INFO TaskSchedulerImpl: Removed TaskSet 12.0, whose tasks have all completed, from pool
18/06/01 00:34:48 INFO DAGScheduler: ResultStage 12 (show at Assignment_21.scala:31) finished in 0.708 s
18/06/01 00:34:48 INFO DAGScheduler: Job 7 finished: show at Assignment_21.scala:31, took 0.731553 s
18/06/01 00:34:48 INFO CodeGenerator: Code generated in 17.161795 ms
+----+--------+
|year|count(1)|
+----+--------+
|2015|       3|
|2014|       3|
|2016|       2|
|2017|       1|
+----+--------+

Taskl output
18/06/01 00:34:48 INFO SparkContext: Invoking stop() from shutdown hook
18/06/01 00:34:48 INFO SparkUI: Stopped Spark web UI at http://192.168.5.109:4040
18/06/01 00:34:48 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
```

Compilation completed successfully with 1 warning in 3 s 337 ms (a minute ago)

1.a.



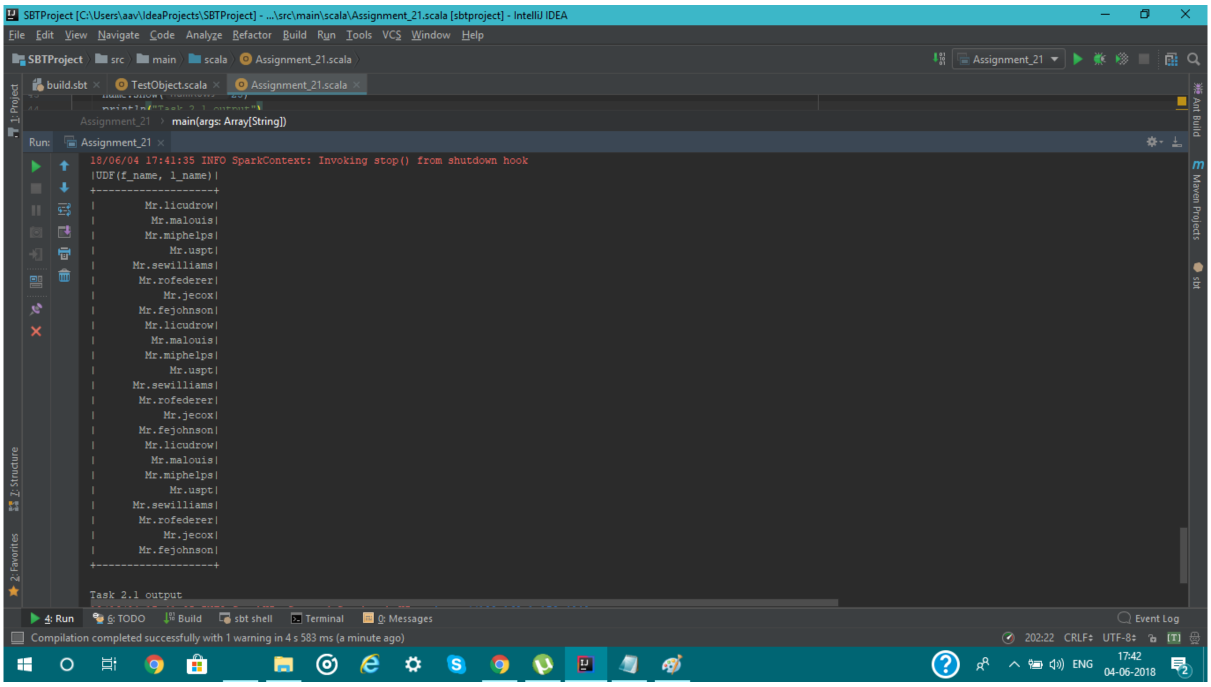
SBTProject [C:\Users\aav\IdeaProjects\SBTProject] - ...\src\main\scala\Assignment_21.scala [sbtproject] - IntelliJ IDEA

```scala
import spark.implicits._
val s =datal.map(x=>x.split(",")).map(x => Sports(x(0),x(1),x(2),x(3),x(4).trim.toInt,x(5).trim.toInt,x(6)))
  .toDF()
s.show(25)
println("Sports data")

s.registerTempTable("sport")
println("temp table created")

val gold=spark.sql("SELECT year,count(*) from sport where medal_type='gold' group by year")
gold.show
println("Taskl output")
```

```
18/06/01 00:32:53 INFO DAGScheduler: ResultStage 22 (show at Assignment_21.scala:35) finished in 0.533 s
18/06/01 00:32:53 INFO DAGScheduler: Job 12 finished: show at Assignment_21.scala:35, took 0.565683 s
18/06/01 00:32:53 INFO BlockManagerInfo: Removed broadcast_14_piece0 on 192.168.5.109:55720 in memory (size: 11.3 KB, free: 899.7 MB)
+--------+--------+
|  sports|count(1)|
+--------+--------+
|swimming|       3|
+--------+--------+

Task2 output
18/06/01 00:32:53 INFO CodeGenerator: Code generated in 27.712801 ms
18/06/01 00:32:53 INFO SparkContext: Invoking stop() from shutdown hook
18/06/01 00:32:53 INFO SparkUI: Stopped Spark web UI at http://192.168.5.109:4040
18/06/01 00:32:53 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/06/01 00:32:54 INFO MemoryStore: MemoryStore cleared
18/06/01 00:32:54 INFO BlockManager: BlockManager stopped
18/06/01 00:32:54 INFO BlockManagerMaster: BlockManagerMaster stopped
18/06/01 00:32:54 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
```

Compilation completed successfully with 1 warning in 3 s 146 ms (a minute ago)

1.b.

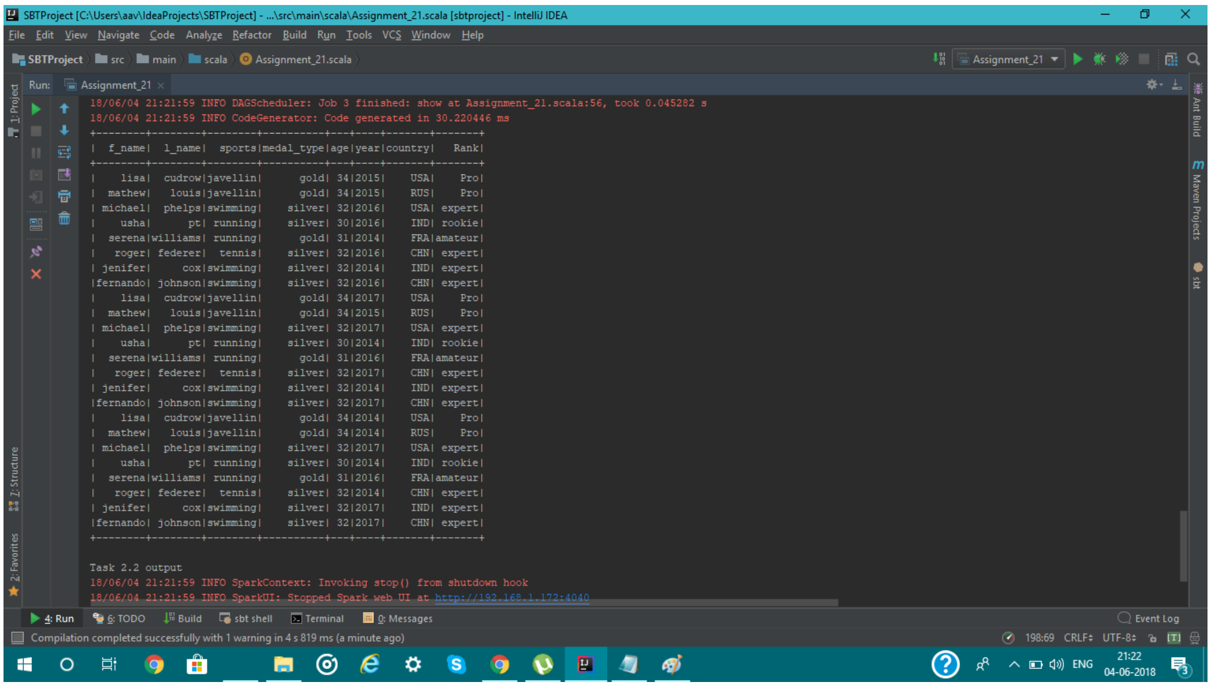


SBTProject [C:\Users\aav\IdeaProjects\SBTProject] - ...\src\main\scala\Assignment_21.scala [sbtproject] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
SBTProject src main scala Assignment_21.scala
build.sbt TestObject.scala Assignment_21.scala
Assignment_21 main(args: Array[String])
Run: Assignment_21
4: Run 6: TODO Build sbt shell Terminal 0: Messages Event Log
202:22 CRLF UTF-8
17:42 04-06-2018

```
18/06/04 17:41:35 INFO SparkContext: Invoking stop() from shutdown hook
|UDF(f_name, l_name)|
+-------------------+
|        Mr.licudrow|
|         Mr.malouis|
|        Mr.miphelps|
|            Mr.uspt|
|      Mr.sewilliams|
|       Mr.rofederer|
|           Mr.jecox|
|       Mr.fejohnson|
|        Mr.licudrow|
|         Mr.malouis|
|        Mr.miphelps|
|            Mr.uspt|
|      Mr.sewilliams|
|       Mr.rofederer|
|           Mr.jecox|
|       Mr.fejohnson|
|        Mr.licudrow|
|         Mr.malouis|
|        Mr.miphelps|
|            Mr.uspt|
|      Mr.sewilliams|
|       Mr.rofederer|
|           Mr.jecox|
|       Mr.fejohnson|
+-------------------+

Task 2.1 output
```

Compilation completed successfully with 1 warning in 4 s 583 ms (a minute ago)

2.a.



SBTProject [C:\Users\aav\IdeaProjects\SBTProject] - ...\src\main\scala\Assignment_21.scala [sbtproject] - IntelliJ IDEA
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
SBTProject src main scala Assignment_21.scala
Run: Assignment_21
4: Run 6: TODO Build sbt shell Terminal 0: Messages Event Log
198:69 CRLF UTF-8
21:22 04-06-2018

```
18/06/04 21:21:59 INFO DAGScheduler: Job 3 finished: show at Assignment_21.scala:56, took 0.045282 s
18/06/04 21:21:59 INFO CodeGenerator: Code generated in 30.220446 ms
+--------+--------+--------+----------+---+----+-------+-------+
|  f_name|  l_name|  sports|medal_type|age|year|country|   Rank|
+--------+--------+--------+----------+---+----+-------+-------+
|    lisa|  cudrow|javellin|      gold| 34|2015|    USA|    Pro|
|  mathew|   louis|javellin|      gold| 34|2015|    RUS|    Pro|
| michael|  phelps|swimming|    silver| 32|2016|    USA| expert|
|    usha|      pt| running|    silver| 30|2016|    IND| rookie|
|  serena|williams| running|      gold| 31|2014|    FRA|amateur|
|   roger| federer|  tennis|    silver| 32|2016|    CHN| expert|
| jenifer|     cox|swimming|    silver| 32|2014|    IND| expert|
|fernando| johnson|swimming|    silver| 32|2016|    CHN| expert|
|    lisa|  cudrow|javellin|      gold| 34|2017|    USA|    Pro|
|  mathew|   louis|javellin|      gold| 34|2015|    RUS|    Pro|
| michael|  phelps|swimming|    silver| 32|2017|    USA| expert|
|    usha|      pt| running|    silver| 30|2014|    IND| rookie|
|  serena|williams| running|      gold| 31|2016|    FRA|amateur|
|   roger| federer|  tennis|    silver| 32|2017|    CHN| expert|
| jenifer|     cox|swimming|    silver| 32|2014|    IND| expert|
|fernando| johnson|swimming|    silver| 32|2017|    CHN| expert|
|    lisa|  cudrow|javellin|      gold| 34|2014|    USA|    Pro|
|  mathew|   louis|javellin|      gold| 34|2014|    RUS|    Pro|
| michael|  phelps|swimming|    silver| 32|2017|    USA| expert|
|    usha|      pt| running|    silver| 30|2014|    IND| rookie|
|  serena|williams| running|      gold| 31|2016|    FRA|amateur|
|   roger| federer|  tennis|    silver| 32|2014|    CHN| expert|
| jenifer|     cox|swimming|    silver| 32|2017|    IND| expert|
|fernando| johnson|swimming|    silver| 32|2017|    CHN| expert|
+--------+--------+--------+----------+---+----+-------+-------+

Task 2.2 output
18/06/04 21:21:59 INFO SparkContext: Invoking stop() from shutdown hook
18/06/04 21:21:59 INFO SparkUI: Stopped Spark web UI at http://192.168.1.172:4040
```

Compilation completed successfully with 1 warning in 4 s 819 ms (a minute ago)

2.b.