Olasunkanmi Olayinka – SEC01 (NUID 001512266)

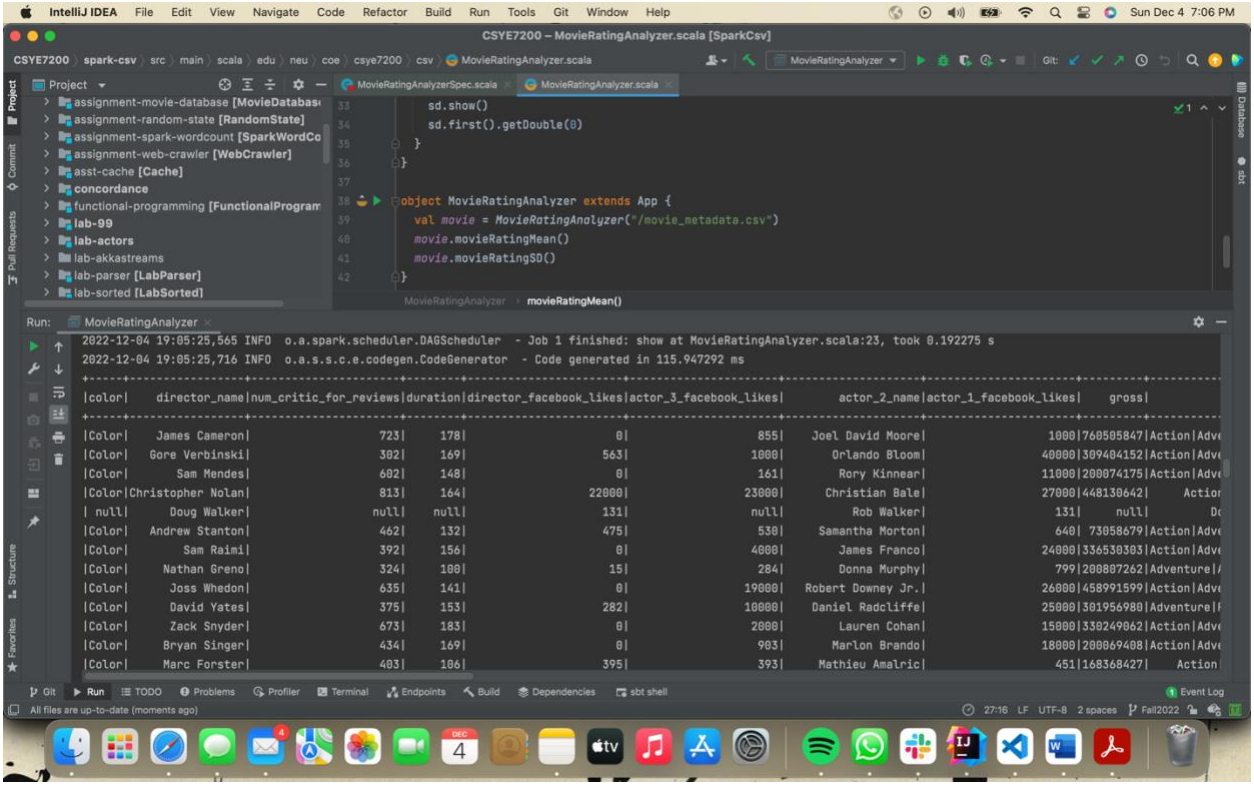
Big Data System Engineering with Scala
Fall 2022
Assignment No. 07

-List of Tasks Implemented

- Analyze a movie rating dataset stored in a CSV file
- Read the file using spark
- Calculate the mean and standard deviation for all movies
- Write a test case to ensure the programs works

-Code

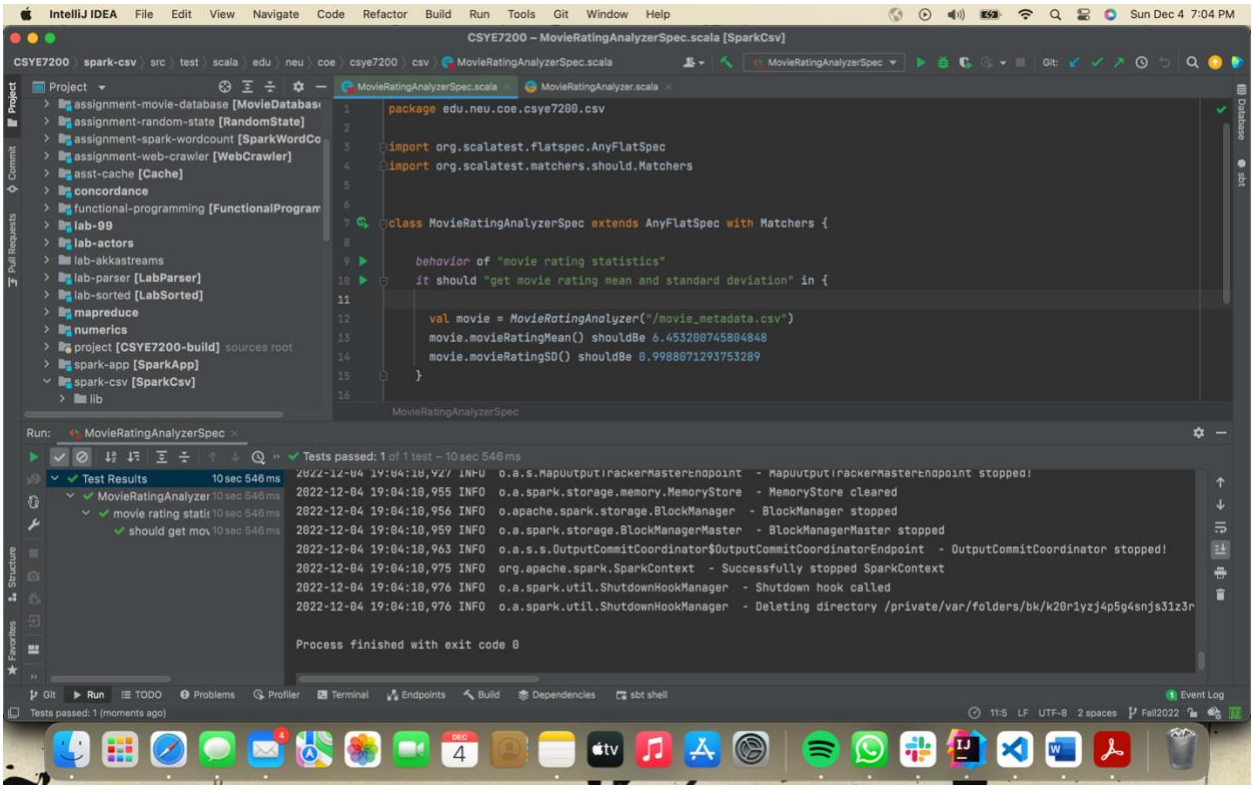
```
case class MovieRatingAnalyzer(resource: String) {

  val spark: SparkSession = SparkSession
          .builder()
          .appName( name = "MovieRating")
          .master( master = "local[*]")
          .getOrCreate()

  spark.sparkContext.setLogLevel("ERROR")

  def apply(resource: String): MovieRatingAnalyzer = new MovieRatingAnalyzer(resource)

  val df: DataFrame = spark.read.format( source = "csv").option("header", "true")
    .load(getClass.getResource(resource).getPath)
  df.show()

  def movieRatingMean(): Double = {
    val mean = df.select(functions.avg( columnName = "imdb_score"))
    mean.show()
    mean.first().getDouble(0)
  }

  def movieRatingSD(): Double = {
    val sd = df.select(functions.stddev( columnName = "imdb_score"))
    sd.show()
    sd.first().getDouble(0)
  }
}

```




```
2022-12-04 19:05:26,529 INFO  o.a.s.s.c.e.codegen.CodeGenerator  - Code generated in 20.165542 ms
+------------------+
|  avg(imdb_score)|
+------------------+
|6.453200745804848|
+------------------+
```

```
2022-12-04 19:05:27,150 INFO  o.a.spark.scheduler.DAGScheduler  - Job 4 finished: show at MovieRatingAnalyzer.scala:33, took 0.095737 s
+----------------------+
|stddev_samp(imdb_score)|
+----------------------+
|     0.998807129375328|
+----------------------+
```

-Unit tests



- Result

Mean = 6.453200745804848

Standard Deviation = 0.9988071293753289