Sai Sumana Puppala – SEC01 (NUID 002925158)

Big Data System Engineering with Scala  
Fall 2022   
Assignment No. 7



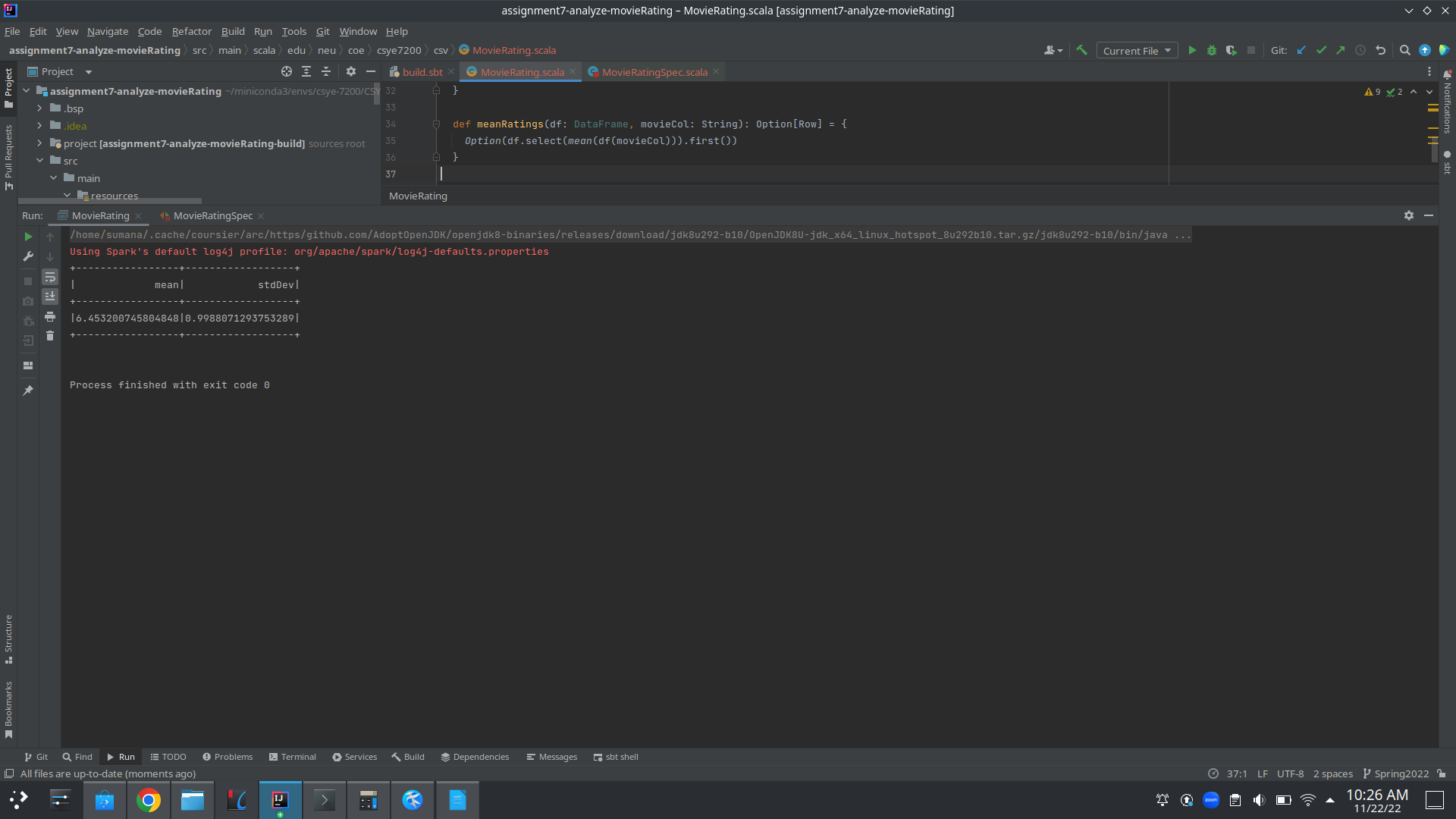
**-List of Tasks Implemented**

Analyzed movies\_metadata.csv to get the mean and standard deviation of the imdb score ratings using spark.

* MovieRating.scala: Created MovieRating case class and its companion object with the apply method that returns an instance of MovieRating.

case class MovieRating(resource: String)

object MovieRating extends App{  
 Logger.*getLogger*(("org")).setLevel(Level.*OFF*)  
 def apply(resource: String): MovieRating = new MovieRating(resource)



* + Defined meanRatings, stdDevRatings and processDF to return the mean, standard deviation of movie rating. ProcessesDF is an additional method that returns a dataframe of the movie ratings mean and standard deviation.
* MovieRatingSpec.scala: Wrote test cases to check for the following use cases:
  + Check if path of the csv file is parsed correctly.
  + Check if the csv file has been fetched correctly.
  + Check for correct movie rating statistics.

Github Link:

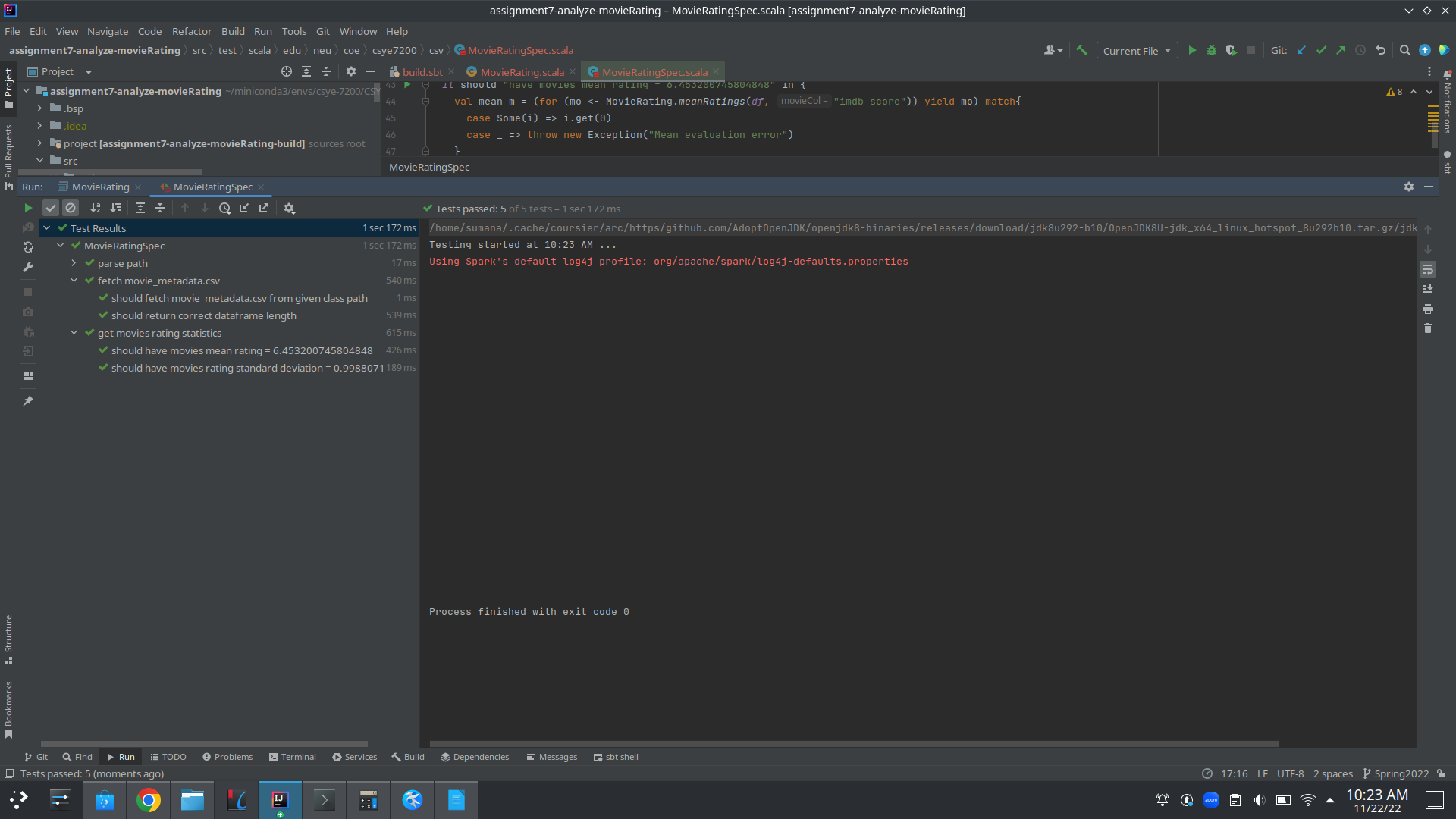
https://github.com/puppala-sumana/CSYE7200/tree/Spring2022/assignment7-analyze-movieRating

**-Code**

The three main methods:

def meanRatings(df: DataFrame, movieCol: String): Option[Row] = {  
 *Option*(df.select(*mean*(df(movieCol))).first())  
}  
  
def stdDevRatings(df: DataFrame, movieCol: String): Option[Row] = {  
 *Option*(df.select(*stddev*(df(movieCol))).first())  
}  
  
def processDF(df: DataFrame, movieCol: String): DataFrame = {  
 df.select(*mean*(movieCol).alias("mean"), *stddev*(movieCol).alias("stdDev"))  
}

**-Unit tests**

****