Shuyan Bian – SEC01 (NUID 002309442)

Big Data System Engineering with Scala
Spring 2024
Assignment No. 10

- GitHub Repo URL -

- List of Tasks Implemented

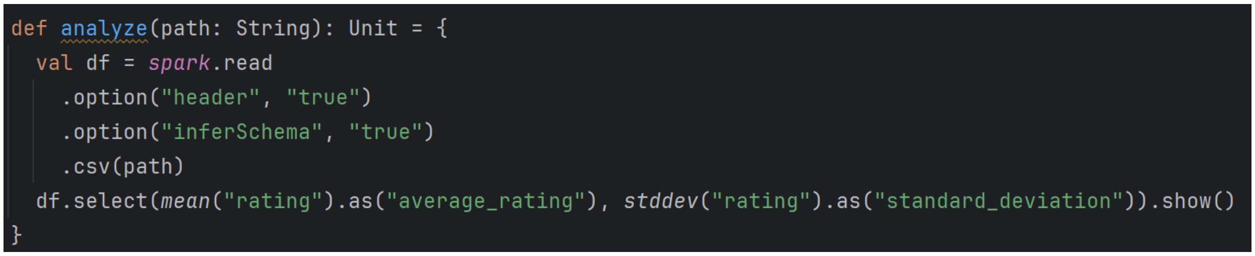
- Code

FileName.Scala (Either text or screenshot can be added below of your code snippet, don’t add the whole code just the part you completed)

Dataset source:

https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?resource=download&select=ratings.csv

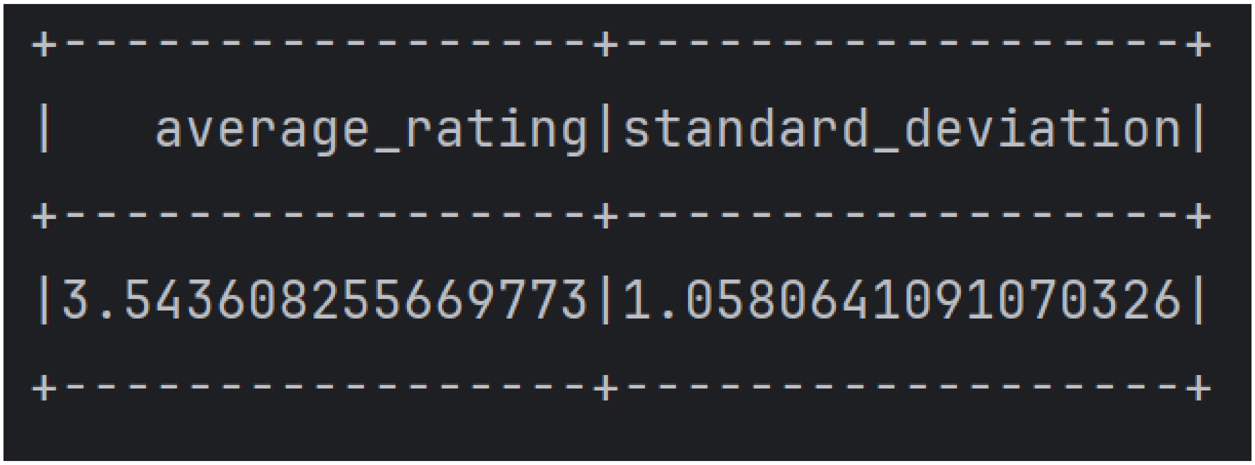
https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?resource=download&select=ratings_small.csv



```scala
def analyze(path: String): Unit = {
  val df = spark.read
    .option("header", "true")
    .option("inferSchema", "true")
    .csv(path)
  df.select(mean("rating").as("average_rating"), stddev("rating").as("standard_deviation")).show()
}
```

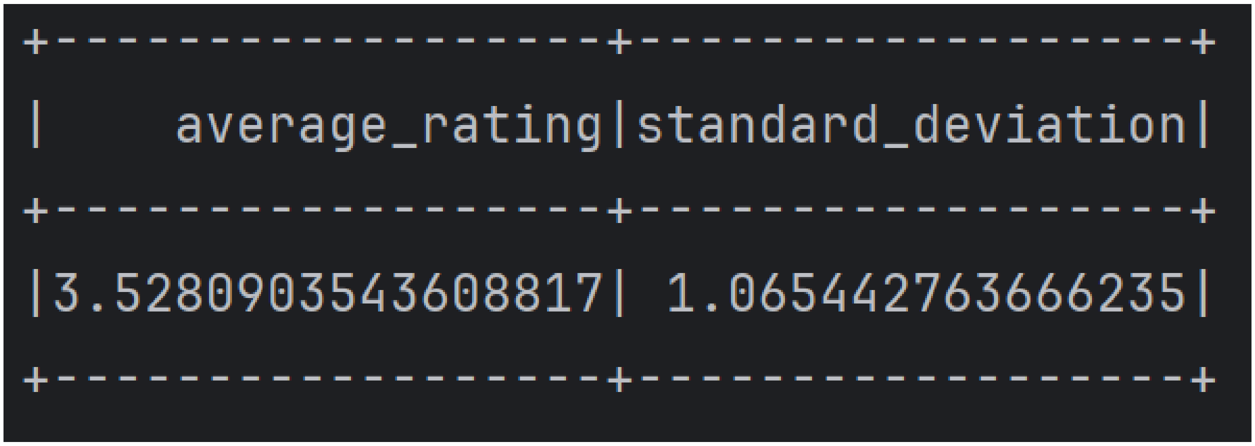
- Unit tests / Results

I use ratings_small as test cases.



```
+-----------------+------------------+
|   average_rating|standard_deviation|
+-----------------+------------------+
|3.543608255669773|1.0580641091070326|
+-----------------+------------------+
```

And the final result for ratings is as follows.



```
+------------------+------------------+
|    average_rating|standard_deviation|
+------------------+------------------+
|3.5280903543608817| 1.065442763666235|
+------------------+------------------+
```