

Weikun Zhuang – SEC01 (NUID 001537998)

Big Data System Engineering with Scala

Fall 2022

Assignment No. #7



-List of Tasks Implemented

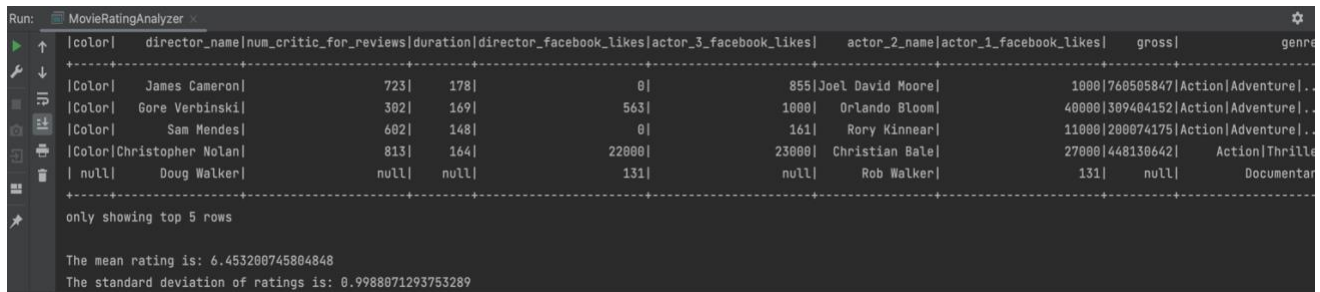
- Implemented a program to read in CSV file into Spark, and calculate the mean and standard deviation of the rating column (i.e. imdb_score)
- Implemented unit tests to test the functions

-Code

Please see the folder [assignment-movie-rating] in the repo.

-Unit tests

The returned mean and std are:



Run: MovieRatingAnalyzer

color	director_name	num_critic_for_reviews	duration	director_facebook_likes	actor_3_facebook_likes	actor_2_name	actor_1_facebook_likes	gross	genres
Color	James Cameron	723	178	0	855	Joel David Moore	1000	760505847	Action Adventure ...
Color	Gore Verbinski	302	169	563	1000	Orlando Bloom	40000	309404152	Action Adventure ...
Color	Sam Mendes	602	148	0	161	Rory Kinnear	11000	200074175	Action Adventure ...
Color	Christopher Nolan	813	164	22000	23000	Christian Bale	27000	448130642	Action Thriller ...
null	Doug Walker	null	null	131	null	Rob Walker	131	null	Documentary ...

only showing top 5 rows

The mean rating is: 6.453200745804848
The standard deviation of ratings is: 0.9988071293753289

