DSC 102: Systems for Scalable Analytics

Programming Assignment 2: Grading Scheme

Programming correctness (90)

For each task, we will run several tests on it. Your code must pass all the tests to be counted pass for the task.

Task No.	Task Description	Score (pass/fail)
1	Average and count of rating per product	20/0
2	Flattening schema, dealing with array and map type	10/0
3	Flattening schema, conducting self-joins	30/0
4	Typecasting and data imputation	10/0
5	Apply word2vec on string data	15/0
6	One-hot encoding and PCA on categorical data	15/0

Timeout: each submission (all six tasks put together) will be given at most 30 minutes to finish on a four-worker cluster. Your final score will be the total scores of tasks passed before timeout. We will deduct points if it takes longer as per the following.

- Between 30min and 60min: -10 points
- Over 60min: -30 points

If any task fails you might still get partial credits for the task.

Extra Credit (10)

Your code will also be timed and based on the run time of all six tasks together, teams in the top 10 percentile of runtimes will receive a 10-point bonus.