DSC 102: Systems for Scalable Analytics

Programming Assignment 1: Grading Scheme

For evaluation we will run your code on a smaller hidden test dataset. The below runtime thresholds are for this hidden dataset. If your runtime on the public dataset is close to ours, you can expect similar results on the hidden dataset.

Accuracy (80)

Your solutions rounded to 2 decimal points must match exactly as our solutions, otherwise you will receive 0 points. We will tolerate a margin of 1.5% only for the median value. There is no partial marking.

Task No.	Task Description	Points
1	% Missing values for all columns in *reviews* table	6
2	% Missing values for all columns in *products* table	6
3	Pearson correlation between rating and price	14
4	Descriptive stats for price	12
5	Number of products for each super-category	14
6	Check for dangling reference of product ids from *reviews* table to *products* table	14
7	Check for dangling reference of product ids in the *related* column to *asin* column of *products* table	14

Runtime (20)

We will run your functions thrice and take the average for obtaining the runtime measurement. If your accuracy points are greater than 40, the runtime grading is based on the below tables. If your accuracy points fall below 40 then you will get partial credit based on manual inspection by TAs.

Absolute 5 node runtime	Points
Under 5 mins	5
Between 5 mins to 10 mins	3
Between 10 mins to 30 mins	2
Anything above 30 mins	0

Runtime speedup with 5 nodes over 2 nodes	Points
Greater than 3.0x	5
Between 2.25x to 3.0x	3
Between 1x to 2.25x	2
Under 1x	0

Extra Credit (+15)

If you receive full accuracy points, and you implement a parallelized solution for Q6 (more details in problem statement), and the runtime is less than 3 minutes and speedup greater than 2.5 on our hidden dataset you will receive an additional 15 points.