Student-Math-Data-Analysis-Report

January 6, 2025

Author: Bosley Boka

1 Report Overview

This report illustrates the data findings from cleaning, merging, and analyzing the data for the Student Math Data Project. For additional details, code, and processes used to obtain this reporting, see Student-Math-Data-Analysis-Workbook that was provided with this report.

2 Data Observation Overview

2.1 Header Notes

- While the Usage & SIS files have a clearly labeled student_id column, the Assessment file's student_id column is labeled "id".
- File headers have inconsistent formats- spacing vs underscores, mixed casing, includes slash character.
- The instructions require "lessons completed", whereas the column header is "lesson completed".
- The merged file column "score" was changed to "math_score" for clarity.

2.2 Value Notes

- Benchmark level columns have inconsistent value of "Level4", i.e. no spacing.
- Duplicate student_id values exist in Assessment data, but this is expected since it includes both Math and ELA scores.
  - No duplicates exist after removing ELA subject rows.
- Assessment data column total_minutes has
  - 4 high value outliers
  - 3 values between 0 and 1, showing that 3 students completed between 14-23 lessons in under 1 minute. This may indicate a bug in the product software or in the data collection process.

```
     student_id  lesson completed benchmark_1_level benchmark_2_level  \
16      1254138                12            level4           level 3
24      1254393                16           level 2           level 2
69      1254280                 4           level 1           level 2
154     1254252                 9           level 3           level 3
171     1254367                13            level4           level 3
```

```
179     1254255                 14           level 2           level 3
261     1254284                 23           level 3           level 2

    benchmark_3_level benchmark_4_level  total_minutes  issue_column
16            level 1           level 3     255.231657  total_minutes
24            level 1           level 1       0.167026  total_minutes
69            level 1           level 1     261.391669  total_minutes
154           level 2            level4     247.638391  total_minutes
171           level 2           level 2     280.517175  total_minutes
179            level4           level 1       0.865787  total_minutes
261            level4           level 3       0.856146  total_minutes
```

2.3 Missing Data

- 47 NaN values present in SIS data column race_ethnicity.
- 114 NaN values present in SIS data column Free/Reduced Price Lunch.
- 8 rows in Usage data don’t exist in SIS data. Not included in final CSV (see below)

```
     student_id  lessons_completed benchmark_1_level benchmark_2_level  \
40      1254070                  7            level4           level 3
46      1254069                 23            level4           level 2
77      1254065                  5            level4           level 3
144     1254072                 24           level 1           level 3
200     1254068                 18           level 3           level 3
221     1254066                 16           level 1           level 1
263     1254067                  4           level 2            level4
309     1254071                 14           level 2           level 3

    benchmark_3_level benchmark_4_level  total_minutes
40            level 1            level4     134.195032
46            level 3           level 1      59.247430
77            level 1            level4     124.325568
144           level 1           level 3     129.205541
200            level4           level 1       5.879212
221            level4           level 3      12.492581
263           level 2           level 2      63.419238
309            level4           level 2     228.975415
```

2.4 Additional Observations

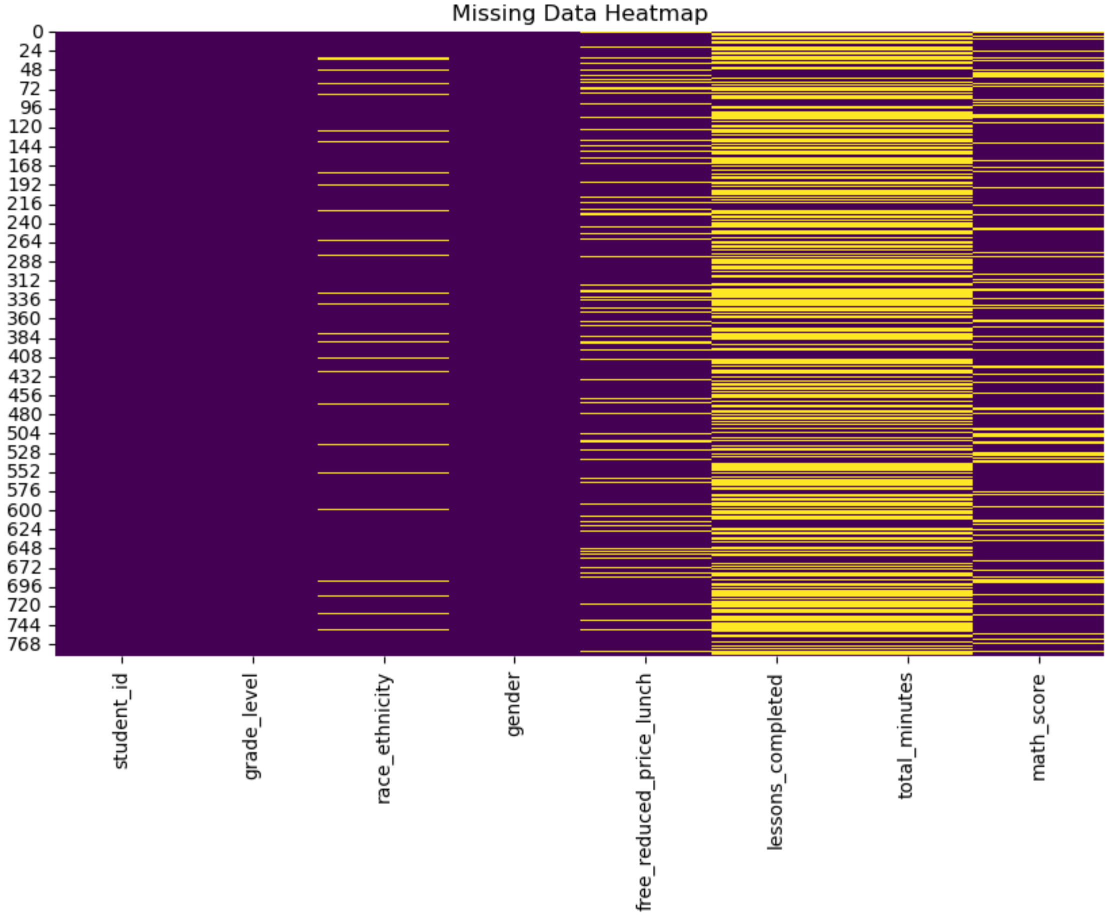
- IDs in Assessment but not in SIS: 0
- IDs in Usage but not in SIS: 8
- IDs in SIS but not in Assessment: 143
- IDs in SIS but not in Usage: 442
- IDs that exist in all files: 231

3 Data Analysis of Merged File

3.1 Analyze Missing Data

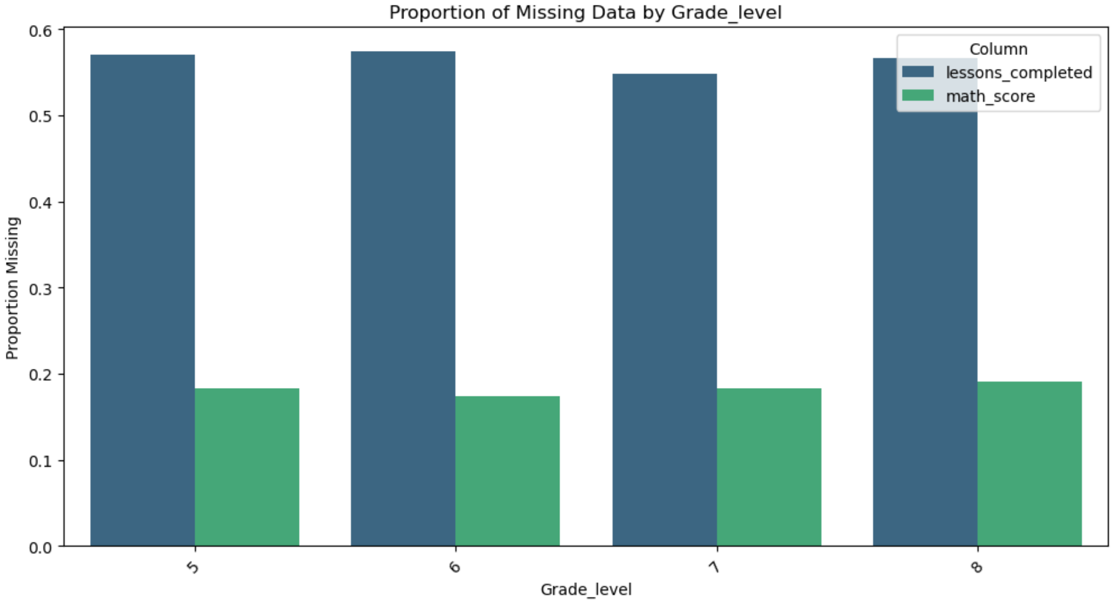
- Most Null/NaN values are from the Usage File, as is shown by the table and heatmap below. Columns lessons_completed and total_minutes are from Usage File.

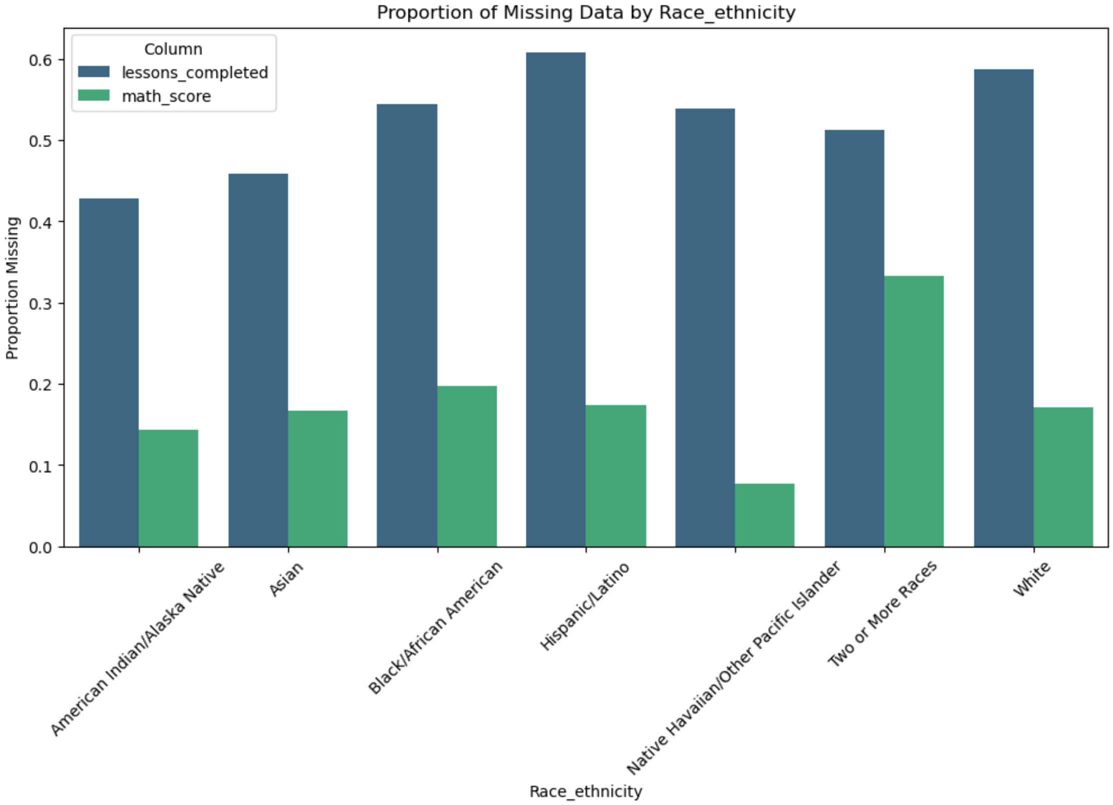
```
                   Column  Null Count  Null Percentage
0              student_id           0         0.000000
1             grade_level           0         0.000000
2          race_ethnicity          47         6.002554
3                  gender           0         0.000000
4  free_reduced_price_lunch       114        14.559387
5       lessons_completed         442        56.449553
6           total_minutes         442        56.449553
7              math_score         143        18.263091
```

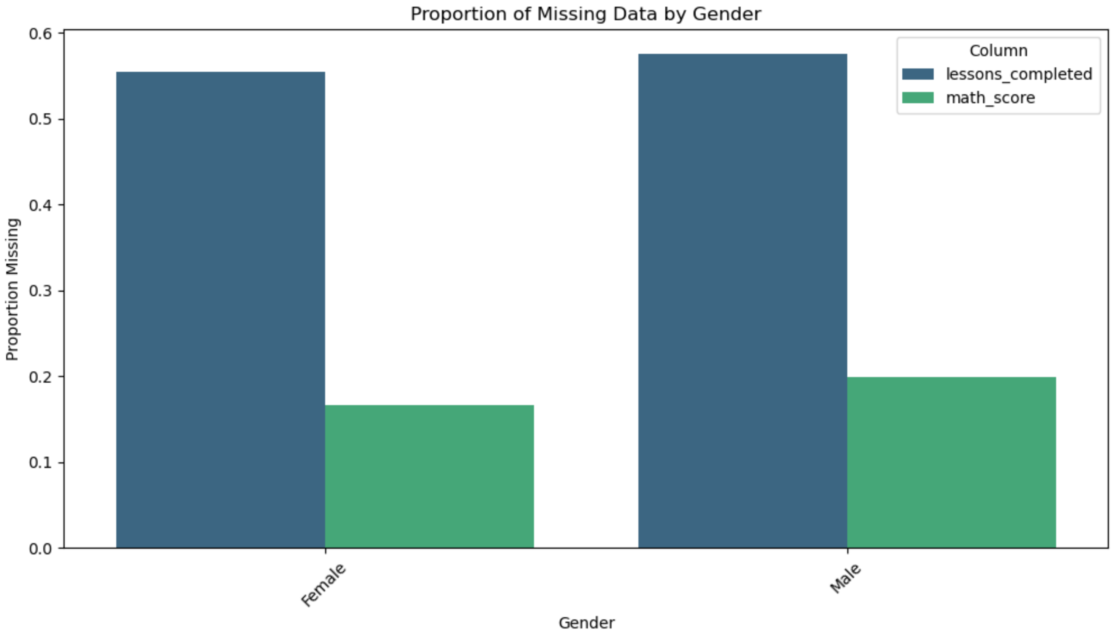


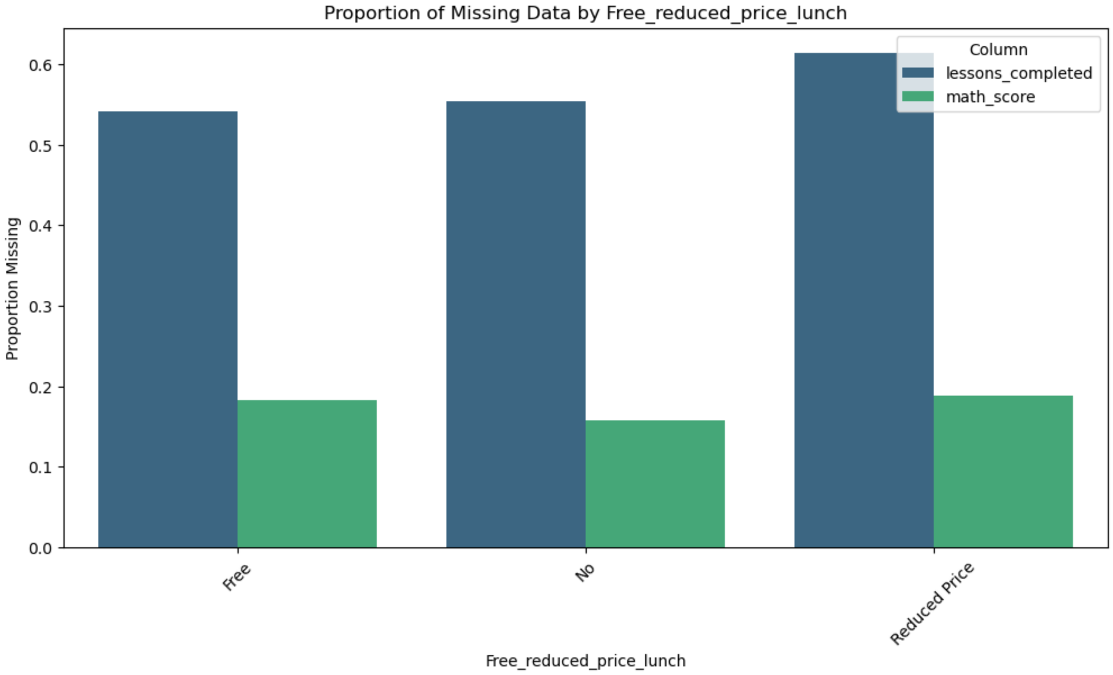
- The demographics with the highest missing data with respect to proportion:
  - Hispanic/Latino missing Usage Data (i.e. lessons_completed and total_minutes).
  - Two or More Races missing Assessment Data (i.e. math_score).

– All other demographics are normal.





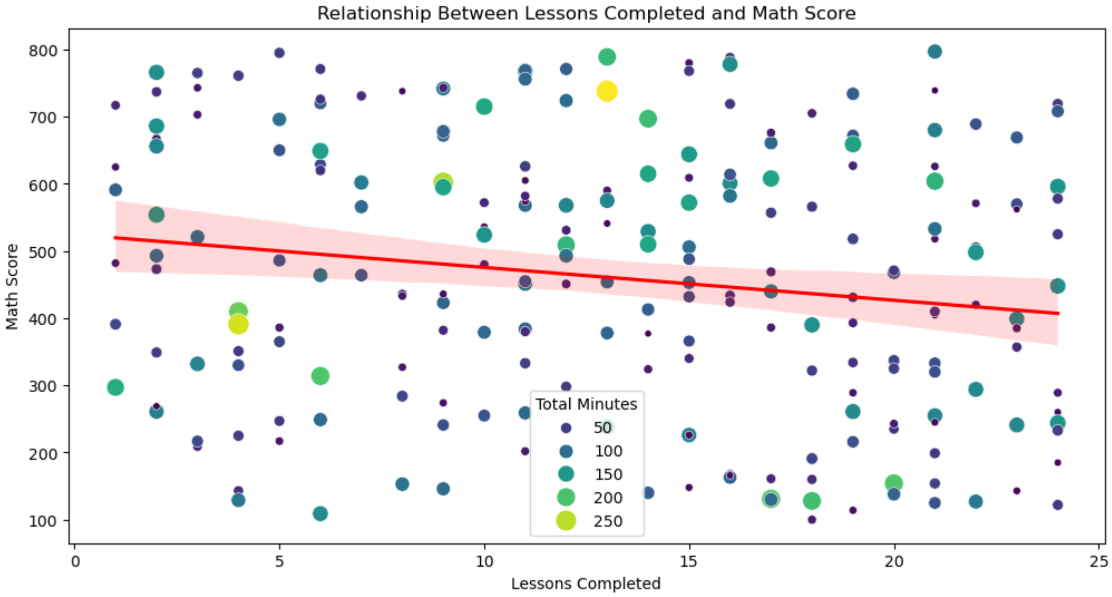




3.2 Analyze Math Data

- There doesn't seem to be any correlation between the math product usage and higher assessment scores. If anything, scores lower as the product is used, as is shown with the illustration below.

```
Correlation with Score for lessons_completed: -0.16664187473992423
Correlation with Score for total_minutes: 0.1072420936688609
```



Relationship Between Lessons Completed and Math Score