ETL Process - NYS Exam Dataset Cleanup (Math & ELA)

1. Check for null values

2. Check for duplicate values

3. Validate quantity bias - every year approximately the same amount students logged is roughly the same.

4. State assumption on the analysis:
    - The test is the same in difficulty every year
    - Pandemic and teaching format have some effect on the students' performances.

5. Split DBN number into 3 categories:
    - District
    - Borough
    - School

6. Map split borough code column to the full name of the borough:
    - M = Manhattan
    - X = Bronx
    - K = Brooklyn
    - Q = Queens
    - R = Staten Island

7. Confirm the intended format of our data is accurate
    - E.g., year read in as an integer - without commas

8. Drop "All Grades" from the grade column

9. Drop "Category" column

10. Drop the "Unnamed: 0" column

11. Drop the "Number Tested" column

12. Merging two tables (horizontally), into 2 sets of columns.

13. Generate primary keys for the following dimension tables:
    - location_ID - begin at 100
    - level_ID - begin at 200
    - year_ID - begin at 300

14. Insert primary keys for dimension tables as foreign keys, and add the grades, for the fact table:
    - location_ID
    - level_ID
    - year_ID
    - grades
        i. (2) separate columns on fact table for ELA & Math scores