Data Report

Rabeya Siddika Sajia - 23084716

Math Matters: Examining the New York State Students Performance Over the Time (2006-2023)

Mathematics is a fundamental part of life, influencing every aspect of daily activities and decision-making. From early childhood, math skills lay the foundation for critical thinking and problem-solving. This project analyzes 18 years of student performance data from New York State Math exams to uncover long-term trends and insights. By combining two datasets, 2006 to 2012 and 2013 to 2023, it will investigate long-term patterns in test outcomes between grade levels, demographic groupings, and geographic areas in New York.

I. Question

How the students' performance in New York City Math exams expanded in the past 18 years, and what factors are involved in the changes in different demographics and regions?

II. Data Sources

Data source 1: **Math Test Results 2006-2012**

- Metadata URL: https://catalog.data.gov/dataset/math-test-results-2006-2012
- Data URL: https://data.cityofnewyork.us/api/views/e5c5-ieuv/rows.csv
- Data Type: CSV File

Data source 2: **Math Test Results 2013-2023**

- Metadata URL: https://catalog.data.gov/dataset/math-test-results-2013-2023
- Data URL: https://data.cityofnewyork.us/api/views/74kb-55u9/rows.csv
- Data Type: CSV File

These datasets provide a complete look at New York State Math exam results over a long period, providing a strong foundation for examining trends in student performance. The datasets include information across grades, demographics, and regional data, enabling a detailed analysis of the factors influencing outcomes.

A. Data Structure

The data is highly structured and organized, adhering to a tabular format with defined rows and columns. Each column corresponds to specific attributes such as year, grade level, proficiency rates, demographic details, and regional classifications. This structure facilitates seamless integration and analysis using standard data processing tools.

B. Data Quality

Accuracy: The data contains real world students' performance in different parts of New York City

Completeness: Contains all necessary information to conduct the analysis.

Consistency: The data is mostly consistent in its formats and units.

Timeliness: Data reflects results from 2006 to 2023.

Relevancy: The datasets focus on students' math performance according to different demographic groupings and geographic areas.

C. License

Both datasets are sourced from the United States Government's open data platform, which explicitly states that the datasets are intended for public access and use. According to the terms and conditions, users are free to access, use, and analyze the datasets, but they should do so with the understanding that the providers are not responsible for the accuracy or completeness of the information. In adherence to these terms, I will ensure that the data is used responsibly, without engaging in any criminal activity, misuse, or actions that violate the stated guidelines.

Link: https://www.nyc.gov/home/terms-of-use.page

III. Data Pipeline

In this project Python is used to build the Data Pipeline. The traditional data pipeline architecture ETA (Extraction, Transformation and Loading) is used which extracts data from internet in CSV format, cleans the data by removing unnecessary columns and missing values, and finally saves the transformed data in a structured SQLite database for future analysis.

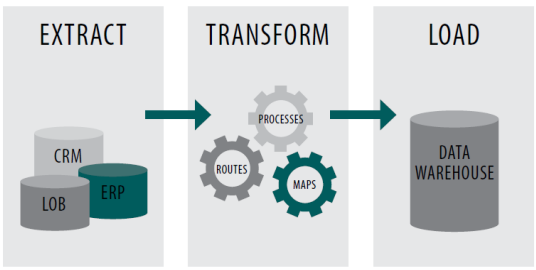

Figure 1. ETA Data Pipeline

A. Data Extraction:

Both datasets are extracted from the data.gov websites in CSV format. Then I used the function read_csv from pandas to store it in a dictionary to further process the data.

B. Cleaning and Transformation:

Since the datasets spanned two periods, 2006-2012 and 2013-2023, I combined them into a single dataset. During this process, I identified and removed entries with incorrect values to ensure data quality. Missing values were addressed wherever possible by filling in gaps with appropriate estimates. To account for variations in grading systems over the years, I adjusted raw scores and aligned demographic and regional categories across both datasets. Finally, the data became clean, consistent and ready for analysis

C. Data Loading:

The final cleaned and transformed data was stored into a SQLite database. Two tables from the two different sources were extracted and stored inside the database.

IV. Problems and Solution

Irregular Formatting: In some columns, such as "Geographic Subdivision," there were irregularities where both string and integer values were present. To address this problem, I removed rows containing integer values to maintain uniform formatting.

Duplicate Rows: The datasets contained some duplicate entries. To clean this up, I used Pandas' drop_duplicates function, effectively eliminating redundancy and ensuring each entry was unique.

Large Dataset Size: Both of my datasets are too large to handle in memory. That is why I implemented optimization techniques, such as chunking the data

during processing, to handle the datasets more efficiently.

A. Meta-Quality Measures

I ensured the accuracy of all transformations by cross-checking the results. Every modification made during the data cleaning and integration process was thoroughly recorded to preserve data extraction and ensure that the analysis could be done reliably.

IV. Results

The final clean and transformed outcome of the pipeline was stored in a SQLite Database called ‘math_results_2006_2023’. It is of high quality with no missing values. SQLite database ensures data integrity and consistency which makes it easy for handling large datasets and performing structured data analysis.

V. Limitations

Despite cleaning and validation efforts, some errors still may be present in the data, such as inaccuracies in reported test scores or demographic information.

Certain demographic or regional data points were missing and had to be estimated. This introduces some uncertainty in the analysis.

The datasets may not fully capture the diversity of all students, as some subgroups or regions might be underrepresented due to missing or incomplete data.

VI. Conclusion

Despite overall progress, challenges remain in addressing demographic disparities and regional inequalities. In conclusion I can say that the ETL data pipeline constructively managed to execute the full process correctly and stored the final dataset in a single table in the data repository.