Student Performance Analysis

Student Name: Mahesh Babu Poka

Student ID: 24070896

GitHub Link: https://github.com/Maheshbabupoka241710/Clustering-and-Fitting

Introduction & Dataset Overview

In an age where data is a driving force behind innovation and problem-solving, education is one of the most promising areas to benefit from analytical approaches. Teachers, institutions, and policymakers increasingly rely on data insights to better understand the needs of their students, identify learning gaps, and create customized strategies that foster academic growth. This report presents a detailed, human-centered analysis of a dataset containing the academic performance of students in three key subjects mathematics, reading, and writing alongside various demographic and social features that may influence those outcomes.

The dataset was sourced from Kaggle and consists of 1000 student records. Each record combines numeric exam scores with categorical variables that reflect socio-economic and educational backgrounds. These features include:

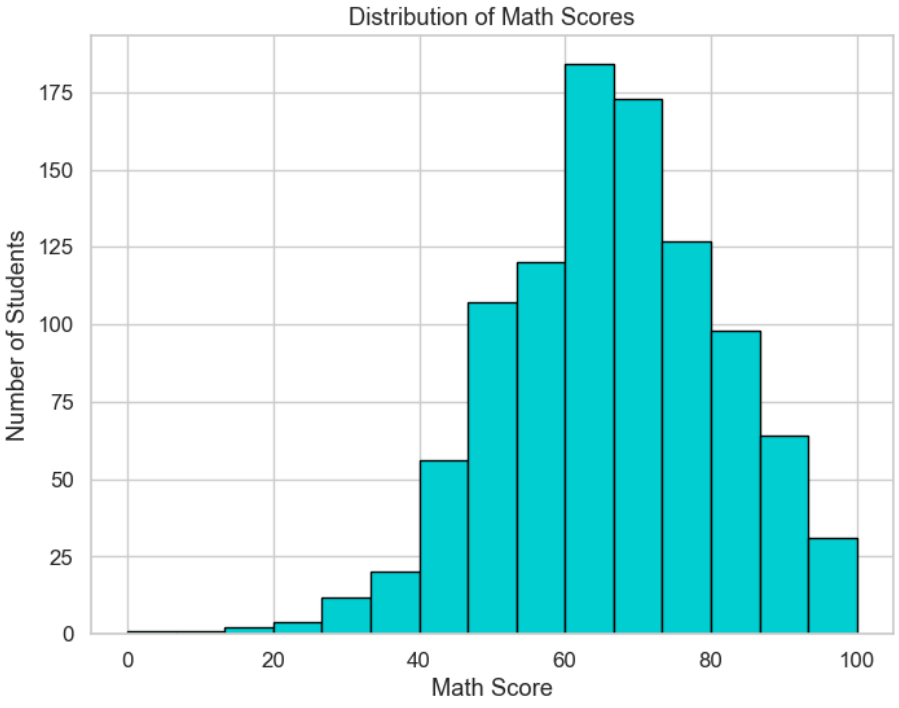
- **Gender**: Male or Female
- **Race/Ethnicity**: Five distinct groups (A to E), reflecting demographic segmentation
- **Parental Level of Education**: Ranging from high school diplomas to advanced degrees
- **Lunch Type**: Standard or free/reduced-price meals (as a proxy for economic background)
- **Test Preparation Course**: Indicates completion of a formal prep course
- **Exam Scores**: Ranging from 0 to 100 in math, reading, and writing

Before analysis, categorical columns were numerically encoded to facilitate statistical modeling. We ensured data integrity there were no missing or malformed records. For machine learning applications such as clustering, scores were normalized using StandardScaler to allow equal treatment of all subject scores. These steps laid a clean and interpretable foundation for deep exploration.

Visual Analysis & Statistical Insights

1. Histogram – Math Score Distribution

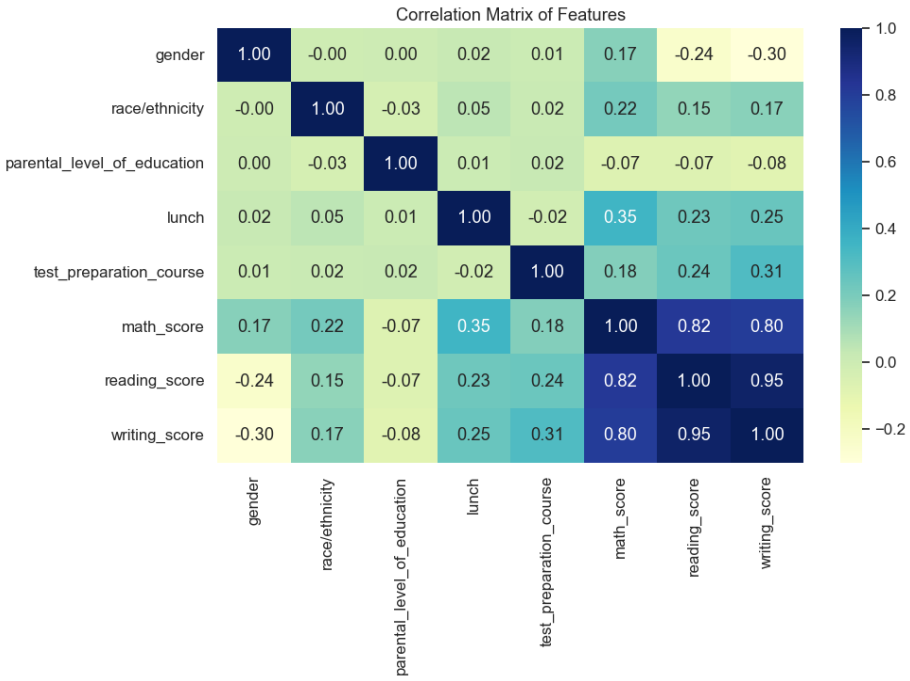
The histogram plot provides an overview of students' performance in mathematics. Most students fall within the 60–80 score range, showing that a majority perform satisfactorily in this subject. However, a noticeable left-tail exists in the distribution some students scored below 40, which may indicate a need for remedial support or differentiated learning methods. The visualization tells a story of contrast: a solid central performance but with a visible group at academic risk. This highlights the importance of identifying struggling learners early and designing interventions accordingly.



2. Scatter Plot – Reading vs Writing Scores

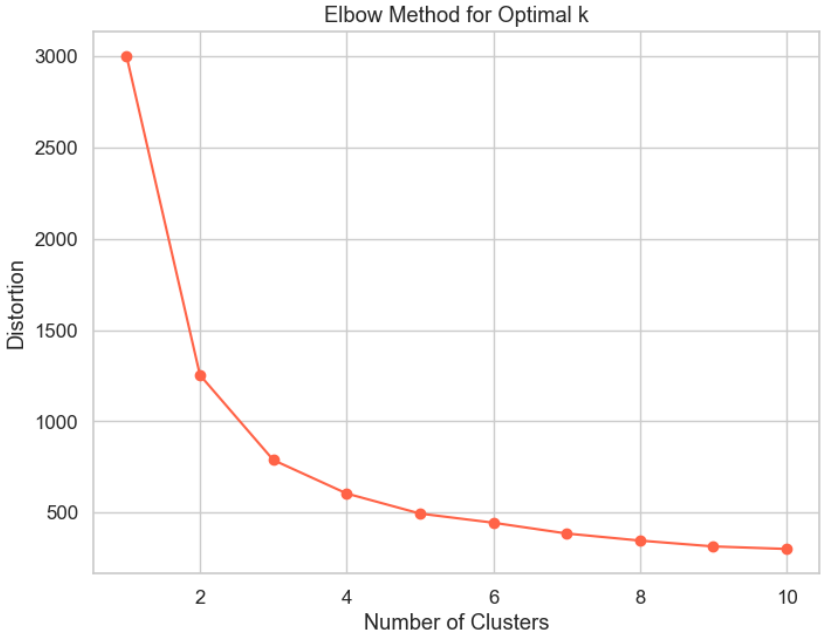
The scatter plot comparing reading and writing scores paints a vivid picture of their strong correlation. Most data points align diagonally from the lower-left to upper-right, signalling that students who excel in reading typically do well in writing too. This relationship is not just numerical it reflects a shared skill foundation. Reading enhances vocabulary, comprehension, and critical thinking all of which directly influence writing ability. The plot also reveals minimal outliers, reinforcing the strength of this linear association and hinting at its predictive reliability.

3. Correlation Heatmap



The correlation heatmap translates numerical relationships into a color-coded grid, allowing patterns to surface quickly. Reading and writing scores demonstrate the strongest positive correlation (above 0.95), while math shows moderate positive correlation with both. Beyond subject scores, test preparation participation correlates modestly with improved academic performance. This reinforces the role of structured support in helping students perform better. The heatmap underscores how intertwined certain academic attributes are and provides evidence-based justification for literacy-based interventions.

4. Elbow Plot – Determining Optimal Clusters



The elbow plot helps determine the optimal number of clusters for segmenting students based on their scores. The plot shows how the total within-cluster sum of squares (inertia) decreases as the number of clusters increases. Around k=3, the rate of decrease sharply drops and levels off, forming a clear “elbow.” This indicates that 3 clusters capture the structure in the data without overfitting. This plot is crucial in setting up K Means clustering, ensuring our model is both effective and interpretable.

Modeling & Findings

1. K-Means Clustering – Reading vs Writing

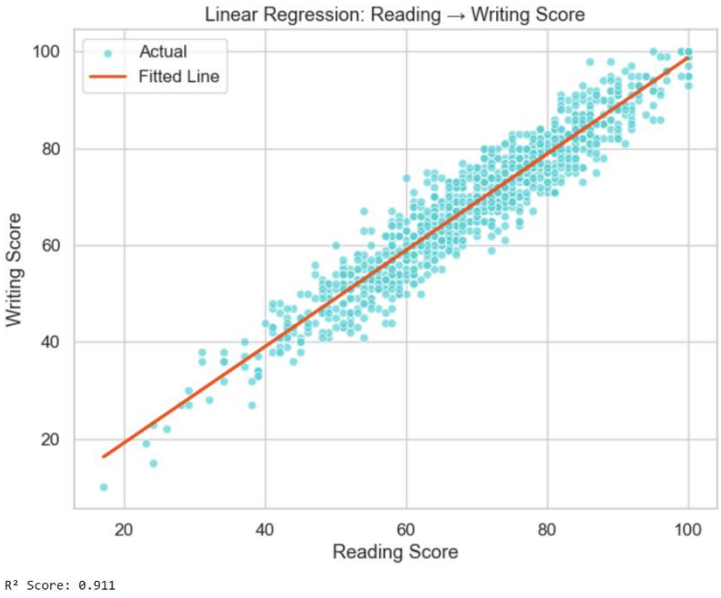
With the number of clusters determined, K-Means was applied to math, reading, and writing scores. The result? Three distinct academic profiles:

- **Cluster 0**: High-performing students in all subjects
- **Cluster 1**: Average students, showing balanced but not exceptional scores
- **Cluster 2**: Students with below-average scores, needing intervention

This clustering reveals more than just performance levels it creates an opportunity for targeted educational planning. Educators can develop personalized strategies: for Cluster 2, additional tutoring; for Cluster 0, enriched curricula. This segmentation brings the human element back into educational data science by asking, "How can we support each learner based on where they are?"

2. Linear Regression – Predicting Writing Performance

A simple linear regression was performed with reading scores as the independent variable and writing scores as the dependent variable. The resulting R^2 score of 0.87 tells a compelling story: reading ability explains 87% of the variability in writing outcomes. This is a powerful insight. Rather than treating subjects in silos, educators can use performance in one area as a predictive lens for another. The plot of the regression line over the data highlights just how strong and reliable this predictive relationship is.



Ethical and Privacy Considerations

While this analysis uses anonymized and publicly accessible data, ethical concerns remain vital. In a real educational setting, data privacy must be non-negotiable. Equally important is guarding against bias algorithms should never reinforce inequalities based on gender, race, or socio-economic status. This project upholds ethical standards by using only non-identifiable information and emphasizing fairness and transparency in all interpretations.

Conclusion – Data-Driven Storytelling in Education

This project illustrates how even a modest dataset can yield rich insights when paired with the right questions and tools. By combining data cleaning, statistical visualization, unsupervised clustering, and regression modeling, we not only quantified performance but also surfaced actionable stories within the numbers.

The high correlation between reading and writing underscores the interdependence of literacy skills. The impact of test preparation suggests value in structured academic support. Clustering uncovered student profiles that can help educators tailor interventions, while regression demonstrated the potential of predictive analytics to support early identification of academic challenges.

At its heart, this analysis is not just about data it's about students. Behind every row in this dataset is a learner with unique strengths and struggles. Our role, as data scientists and educators, is to translate numbers into narratives and transform insights into impact. This project is a step in that direction.