**STUDENTS EXAM SCORES**

Adhoksh Milind Sonawane Kaustubh Arun Pasalkar Sakshi Sanjay Dudhal Apoorv Vaishampayan

Each Member has contributed equally to this project.

# Part I : Introduction

**Background of the business/research problem:**

In order to create interventions and support systems that improve student outcomes, educators, legislators, and researchers must have a thorough understanding of the elements that affect students' academic attainment. By drilling further into the dataset and examining a range of characteristics, we may extract important information that helps educational institutions make decisions and allocate resources.

# Identifying Key Predictors:

By examining the information, we are able to pinpoint important factors that influence academic performance, including parental participation, socioeconomic status, student demographics, school resources, instructional strategies, and student engagement. Identifying the factors that most significantly affect academic achievement might help direct focused interventions and projects.

# Tailoring Interventions:

Equipped with knowledge from the dataset, teachers may customize interventions and support networks to meet the unique needs and difficulties that each student faces. For example, if data show that children from low-income households perform poorly academically, schools might implement targeted programs to provide these students with more assistance, resources, and mentoring.

# Addressing Achievement Gaps:

Analyzing the information might help identify achievement inequalities among various demographic groups, such as students of different racial or ethnic backgrounds, genders, or socioeconomic levels. By understanding the root causes of these discrepancies, teachers and lawmakers may develop strategies to decrease them and ensure that all children have equitable access to high-quality education.

# Improving Teaching Practices:

The dataset's insights may be used to build courses and improve teaching tactics. Teachers can include best practices into their lesson plans, for example, if data demonstrate that specific teaching strategies or instructional approaches lead to higher student achievement.

# Allocating Resources Wisely:

Data-driven decision-making may help educational institutions deploy resources more effectively. Institutions can maximize their expenditure by identifying the most crucial areas for additional resources, such as teacher professional development, infrastructure enhancements, or the creation of support programs for low-achieving students.

# Monitoring and evaluation:

By continuously monitoring and analyzing student outcomes using data analytics, the impact of interventions and initiatives may be assessed over time. Regularly assessing student performance data allows educators to spot patterns, evaluate progress over time, and make modifications as needed to ensure that interventions achieve their intended goals.

In conclusion, using the dataset's insights into factors influencing students' academic progress may influence evidence-based decision-making, enhance educational practices, and ultimately lead to greater student outcomes and success.

# Regression:

1. **Business/Research Question:**
   * What effects do socioeconomic and personal characteristics have on a student's academic performance?
   * Find the most impacting factors from gender, ethnicity, parental education, type of lunch, test-taking strategies, parental marital status, sports involvement, birth order, number of siblings, mode of transportation, amount of study time per week, and test results in math, reading, and writing for students.

# DV (Dependent Variable):

* + The **student's final score**, a continuous variable that reflects their overall academic achievement, is the dependent variable (DV).

# Potential IVs (Independent Variables):

* + Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMaritalStatus, PracticeSport, IsFirstChild, NrSiblings, TransportMeans, WklyStudyHours are potential independent variables.

# Importance in Business/Research Practice:

* + In order to create successful plans for student assistance and development, educational institutions must have a thorough understanding of the variables affecting students' academic performance. It supports equal educational results, efficient resource allocation, and intervention customization.

**- If any, hypotheses**: Specific hypotheses for regression analysis might be developed based on theoretical or previous research. We can propose the following hypothesis.

- Hypothesis 1: Students that obtain test preparation outperform their peers in terms of exam scores.

- Hypothesis 2: Student academic achievement in all disciplines is favorably correlated with parental education level.

# Classification:

1. **Business/Research Question:**
   * Is it possible to forecast a student's ultimate score over average (35 marks) based on their socioeconomic and personal characteristics?
   * Based on their demographic and background data, this question seeks to divide students into two groups: those with final scores above average and those with scores equal to or below average.

# DV (Dependent Variable):

* + As a binary outcome variable, the dependent variable (DV) is whether the student's **final score** is higher than average (35 marks).

# Potential IVs (Independent Variables):

* + Gender, EthnicGroup, ParentEduc, LunchType, TestPrep, ParentMaritalStatus, PracticeSport, IsFirstChild, NrSiblings, TransportMeans, WklyStudyHours are potential independent variables.

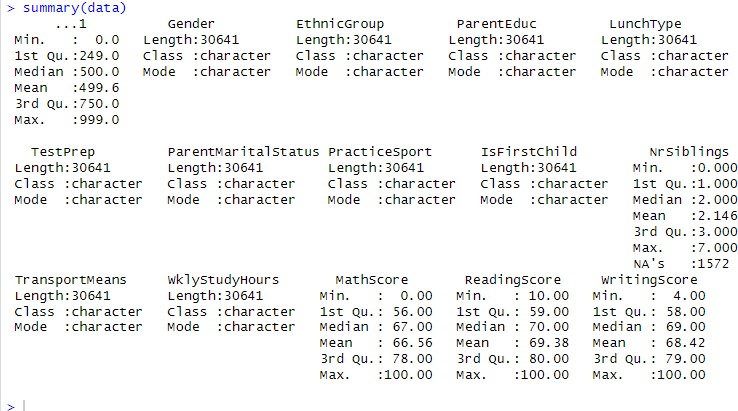
# Importance in Business/Research Practice:

* + Education professionals may better identify at-risk pupils and conduct targeted interventions to enhance their results by using predictive analytics to predict kids' academic performance categories (above or below average) based on their personal and socioeconomic aspects.

# If any, hypotheses:

* Hypothesis 1: Students who frequently participate in sports have a higher chance of scoring above average in comparison to their non-sports-playing peers.
* Hypothesis 2: Students from higher-income homes are more likely than those from lower-income families to score above average, as shown by the sort of lunch they eat.

# Part II : Data Preparation and EDA

Data Source : <https://www.kaggle.com/datasets/desalegngeb/students-exam-scores/data> Summary of the dataset:

**No. of Observations**: In the dataset, there are **6127** observations.

Every observation corresponds to a distinct public-school student, and the associated data captures the student's socioeconomic background, personal characteristics, and exam results in reading, writing, and math. In the dataset, each row represents a single student, and the columns stand for various characteristics or factors connected to that student.

# Variables:- Variable description and Variable type

1. Gender: The gender of the individual (Variable type: Categorical)
2. EthnicGroup: The ethnic group of the individual (Variable type: Categorical)
3. ParentEduc: The education level of the parent (Variable type: Categorical)
4. LunchType: The type of lunch received (Variable type: Categorical)
5. TestPrep: Indicates if test preparation was completed (Variable type: Categorical)
6. ParentMaritalStatus: The marital status of the parents (Variable type: Categorical)
7. PracticeSport: Indicates if the individual practices a sport (Variable type: Categorical)
8. IsFirstChild: Indicates if the individual is the first child (Variable type: Categorical)
9. NrSiblings: The number of siblings (Variable type: Float)
10. TransportMeans: The means of transport used (Variable type: Categorical)
11. WklyStudyHours: The weekly study hours (Variable type: Categorical)
12. MathScore: The score in mathematics (Variable type: Integer)
13. ReadingScore: The score in reading (Variable type: Integer)
14. WritingScore: The score in writing (Variable type: Integer)
15. Final Score: The final aggregated score (Variable type: Integer)

# EDA: Univariate Analysis

* + Results: Frequency counts for categorical variables and descriptive statistics such as the mean, median, quartiles, and range for numerical variables.
  + Remarks on Data Distribution: There is a little right skew in the NrSiblings distribution, meaning that more people have fewer siblings. The skewness of academic scores (MathScore, ReadingScore, WritingScore) is not statistically significant.
  + Missing Values: A number of variables have missing values; the largest number is 637 for TransportMeans, followed by EthnicGroup (379), TestPrep (382), and ParentEduc (413).
  + Extreme Outliers: The presence of severe outliers is suggested by the skewness of Above Average 35 Marks.

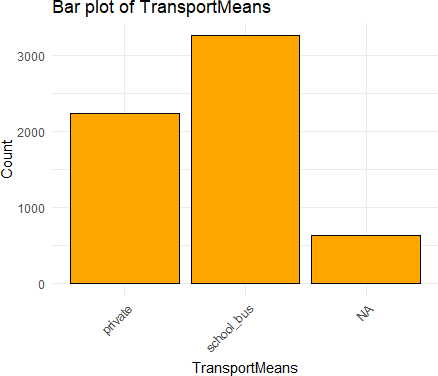
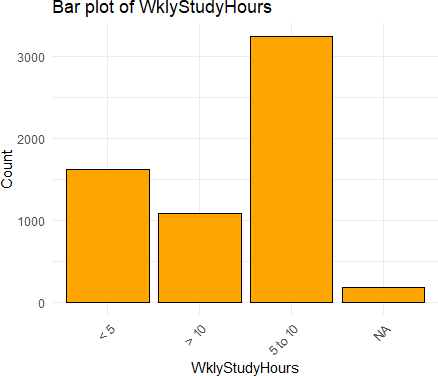
# EDA: Bivariate Analysis

* + Results: To investigate correlations between pairs of variables, bivariate analysis usually produces correlation matrices, scatter plots, or cross-tabulations.
  + Remarks on partnerships: Based on the produced outputs, these would assess the degree of positive, negative, or no correlation between the variables as well as the potential for one variable to predict another.

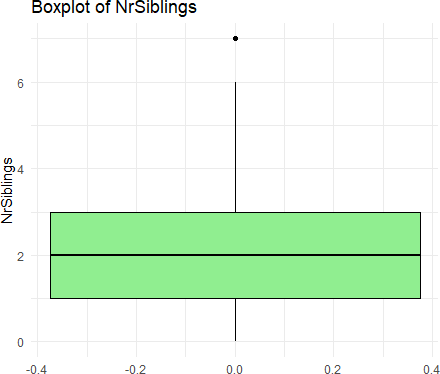
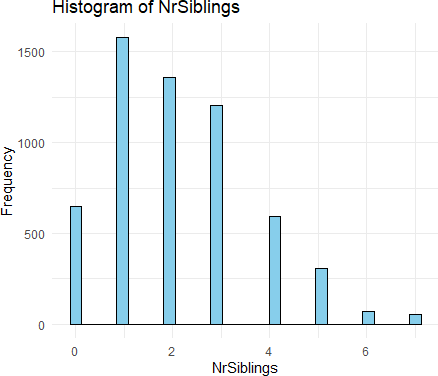
# Data Cleaning/Preparation

* + Extreme outliers should be properly managed and looked at as they may have an impact on the analysis's findings, especially in the Above Average 35 Marks variable.
  + If a variable is not normally distributed, it might need to be modified, especially when modeling is involved.
  + If the bivariate analysis shows that there are substantial interactions between variables, then new variables, such interaction terms, may be established.

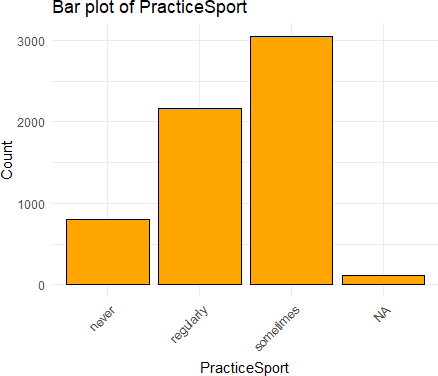
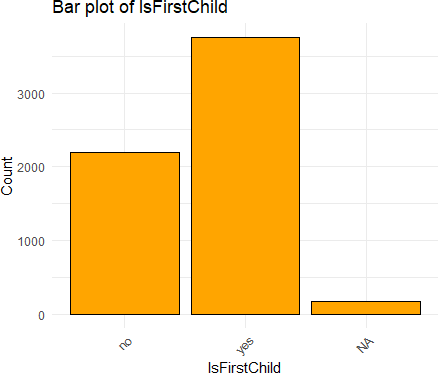
# Graphical representation:



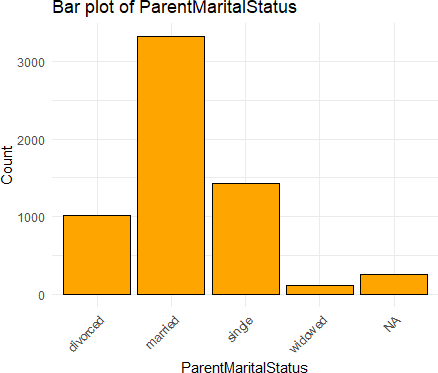
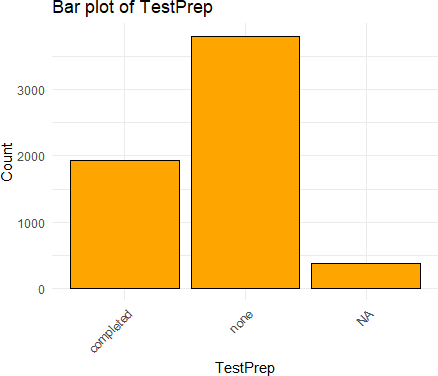
The x-axis of the WklyStudyHours bar plot displays four categories: '< 5', '5-10', '> 10', and 'N/A', which probably refers to not applicable or not accessible.Three categories are included in the TransportMeans bar plot: "Public," "SchoolBus," and "N/A."With a count of almost 2500, "SchoolBus" has the highest, indicating that it is the most often used kind of transportation.The data distribution in these plots indicates that most people have particular information about their study schedules and modes of transportation; most utilize school buses and spend five to ten hours a week studying. The 'N/A' categories in both graphs indicate that there are missing or irrelevant data in the dataset.

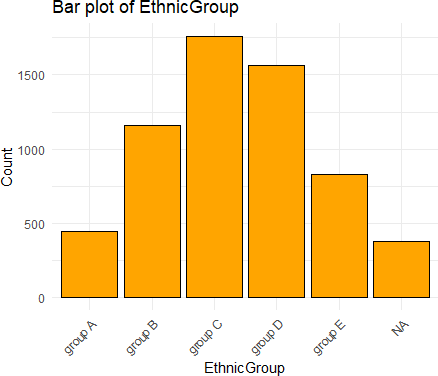
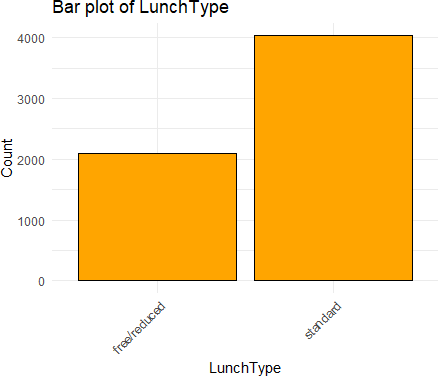
According to both graphs, there is a right-skewed distribution in the number of siblings, with the most frequent number being one and the rarest numbers being more. Six siblings is not the usual number for this group, according to the box plot's outlier. This is supported by the histogram, which demonstrates a low prevalence of people with six siblings. According to the research, most people have between one and three siblings; bigger families are less typical.



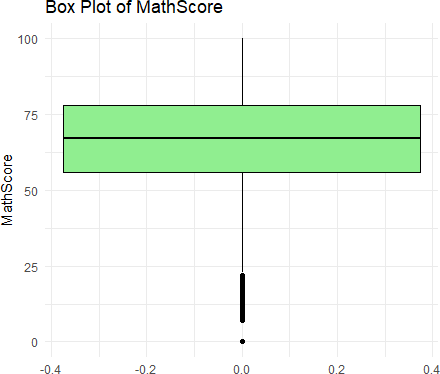
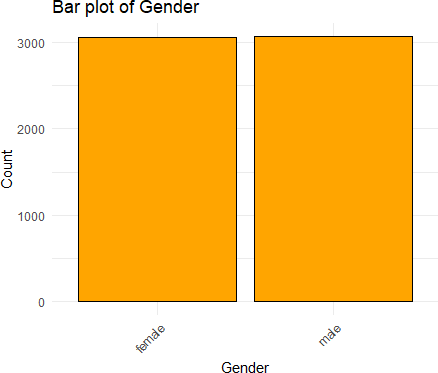
The distribution of these two variables shows that the majority of the people in the sample are first-generation children and that they participate in sports to some extent, with full engagement being less prevalent than sporadic involvement. For these two variables, the 'N/A' counts in both graphs are low, indicating strong data completeness.

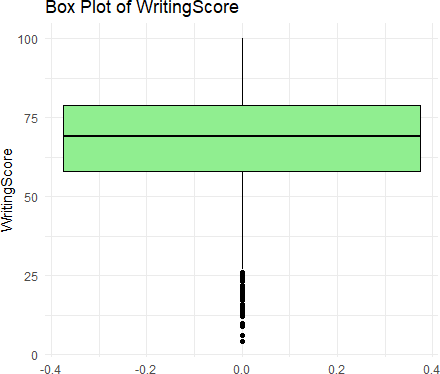
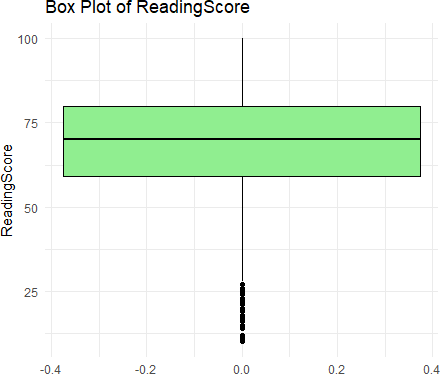
The majority of people have married parents, according to the distribution of "ParentMaritalStatus," with single, divorced, and widowed statuses being less prevalent. According to the statistics, test preparation is not consistently finished by the people in the dataset when it comes to TestPrep. The absence of data does not significantly affect either variable.



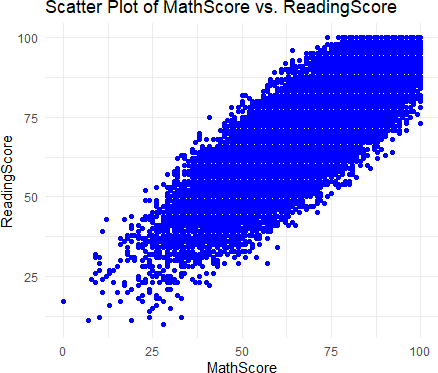
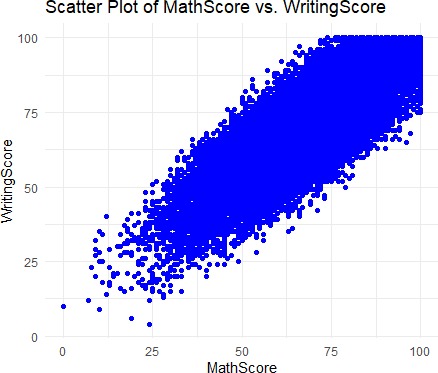
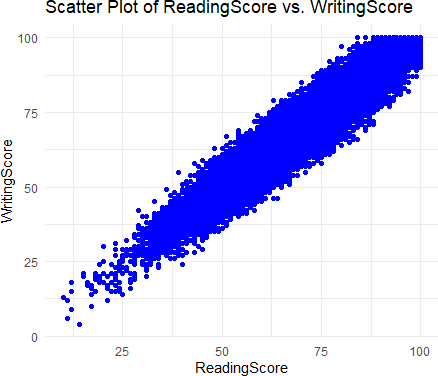
These distributions provide information on the demographics and lunch preferences of the dataset's participants. Standard lunches are more popular than free or reduced lunches, and a variety of ethnic groups are represented, with 'Group C' being the most prominent. The EthnicGroup variable has a 'N/A' category, which means that some of the data is either missing or not recorded.



A gender-balanced dataset is suggested by the gender plot, which is helpful for analysis that compares results across genders. Many students fall into a reasonable range, according to the MathScore box plot, but there are outliers at the lower end, with some kids scoring far lower than the others. There aren't any exceptionally high scores that qualify as anomalies. This distribution may indicate a left-skewed distribution in which a smaller percentage of children struggle with arithmetic and the majority of students achieve at a medium to high level.

The reading and writing score distributions are rather comparable in terms of median and IQR, as both charts show. Lower-end outliers are present in both disciplines, which implies that some students are doing below what their peers are doing. While there are high achievers scoring at the top of the scale, their scores are not abnormally high in relation to the remainder of the distribution, as indicated by the absence of upper outliers. The distributions for writing and reading are comparable, which may indicate that there is a relationship between the two abilities among the pupils.



The distribution of points in all three plots indicates that students are likely to score higher in the other subjects if they do well in the first one, albeit the strength of the connection may change significantly between the subject pairs. These scatter plots show rather consistent variation throughout the score ranges, with no clear outliers or distinguishable clusters.

Multiple illustrations suggest a relationship between students' academic achievement in different topics. However, regression analysis or correlation coefficients might be used to further examine how much progress in one subject predicts better in another.

# Part III Analysis and Findings

From the outputs of both the stepwise regression and the Ridge regression, we calculated the best predictor terms based on their statistical significance, size of coefficient estimates, and consistency across both models are as follows:

# Best Predictor Terms WritingScore:

Stepwise Regression Coefficient: 0.838752 (t-value: 108.808, P-value < 2e-16) Ridge Regression Coefficient: 0.83677186

Interpretation: With the greatest coefficient and t-value of any predictor, this one clearly has a considerable positive impact on ReadingScore. Reading scores rise in tandem with writing skills, indicating a tightly related skill set between the two.

# MathScore:

Stepwise Regression Coefficient: 0.134491 (t-value: 17.030, P-value < 2e-16) Ridge Regression Coefficient: 0.13601205

Interpretation: Like WritingScore, MathScore significantly improves ReadingScore as well. This association highlights how reading comprehension and analytical skills are improved by numerical aptitude.

**Correlation Matrix**

A computer screen shot of a computer program

Description automatically generated

With a statistically significant (p-value < 2.2e-16) Pearson correlation coefficient of 0.8229, the analysis shows a high positive association between Math and Reading scores. Given that gains in Reading scores are likely to result in higher Math results as well, this shows a strong linear link between these two variables. The correlation's 95% confidence interval, which spans from 0.8127 to 0.8326, shows how accurate the estimation was. The correlation matrix also reveals a substantial association between each of the three evaluated subjects—math, reading, and writing. The strong correlations between reading and writing (0.9539) and math and writing (0.8108) highlight the substantial performance overlap in these areas. This suggests that pupils who excel in one area also typically excel in other subjects, pointing to a shared underlying factor influencing success across these academic areas.

# TestPrepnone:

Stepwise Regression Coefficient: 1.974909 (t-value: 13.369, P-value < 2e-16) Ridge Regression Coefficient: 1.95936533

Interpretation: The fact that pupils who did not prepare for the test performed better may indicate that they were already proficient readers and did not require more study.

# LunchTypestandard:

Stepwise Regression Coefficient: -1.041463 (t-value: -6.919, P-value 5.33e-12) Ridge Regression Coefficient: -1.04128356

Interpretation: This coefficient, which is negative but significant, shows that students who ate a regular lunch—which may indicate a higher socioeconomic status—surprisingly performed worse. More socio-economic research may be interested in this surprising discovery.

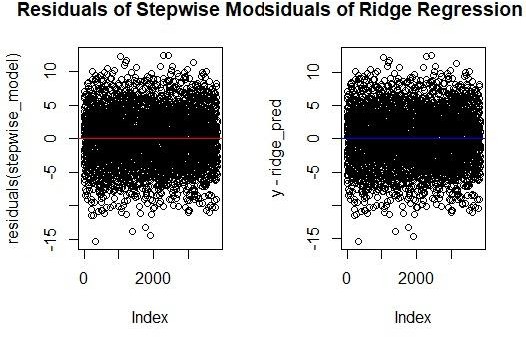
# EthnicGroupgroup D:

Stepwise Regression Coefficient: -1.990819 (t-value: -7.065, P-value 1.90e-12) Ridge Regression Coefficient: -1.22259030

Interpretation:When compared to the baseline group, this predictor exhibits a substantial negative influence, indicating discrepancies that may have their origins in different sociocultural or educational contexts.

# Conclusion

The best predictors—WritingScore, MathScore, TestPrepnone, LunchTypestandard, and EthnicGroupgroup D—are crucial for understanding and predicting ReadingScore.



A graph of a log line

Description automatically generated

# Linear Regression Model:

* Selected Predictor Terms: The model has selected NrSiblings, EthnicGroup (with several levels), TestPrep, WklyStudyHours (with several levels), LunchType, MathScore, and WritingScore as significant predictors for ReadingScore.
* Coefficient Estimates: The coefficients for significant variables include:
  + TestPrepnone: 1.975198
  + LunchTypestandard: -0.871765
  + MathScore: 0.125239
  + WritingScore: 0.842740
  + EthnicGroupgroup D: -1.744027
* Cross-Validated Test Errors: The RMSE for the Linear Regression model is 4.1175.
* Residual Diagnosis: Based on the residuals summary, the Linear Regression model seems to perform well with no indication of systematic errors.

# Random Forest Model:

* Selected Predictor Terms: Random Forest uses all provided predictors but gives a measure of importance for each variable which is not detailed in the provided output.
* Coefficient Estimates: Random Forest does not provide coefficient estimates in the same way Linear Regression does, but it gives a variable importance plot (not provided here) which shows the impact of each predictor on the model accuracy.
* Cross-Validated Test Errors: The RMSE for the Random Forest model is 4.4074.
* Residual Diagnosis: There is no detailed output provided for the residuals of the Random Forest model, but the RMSE indicates the average prediction error.

# Optimal Regression Model:

When comparing the two models, the Linear Regression has a slightly lower RMSE compared to the Random Forest model, indicating better average prediction accuracy. Additionally, the Linear Regression model has a higher interpretability due to the clear coefficient estimates associated with each predictor.

Given these points, the Linear Regression model appears to be the optimal choice for predicting ReadingScore based on the given dataset and within the scope of this analysis.

# Comments on Findings Based on the Optimal Model:

* Academic Skills: The positive impact of WritingScore and MathScore on ReadingScore suggests that students who perform well in writing and math also tend to perform well in reading. This underscores the interrelated nature of academic skills.
* Socioeconomic Factors: The negative coefficient for LunchTypestandard indicates that students who receive a standard lunch, which might be a proxy for higher socioeconomic status, do not necessarily perform better in reading. This counterintuitive result warrants further investigation.
* Cultural Factors: The significant negative coefficient for EthnicGroupgroup D suggests that students from this group, on average, score lower in reading compared to the baseline group. This might reflect educational disparities that need to be addressed.
* Educational Practices: The fact that TestPrepnone has a positive coefficient indicates that students who did not engage in test preparation have higher reading scores. This could suggest these students already possess strong reading skills or that test preparation is not as effective as intended.

This information is really useful for teachers and the people who make education policies. It points out a few key things they should pay attention to:

* Kids need to build a wide range of skills. Being good at just one subject isn't enough – they should be good at many, like reading, writing, and math.
* Schools might need to look again at how they help kids from different backgrounds, especially when these backgrounds might mean they don't have as many resources or support at home.
* Every group of kids is different, and some groups might face unique challenges, especially if they're from various ethnic backgrounds. Schools need to make sure they're meeting these specific needs.
* It's also a good idea to check if the programs designed to help kids get ready for tests are helping them learn better. If students who skip these programs are doing better, we need to figure out why.

In simpler terms, these insights can help schools tailor their teaching and support to help every student do their best, no matter where they come from or what they need.

# Classification Problem Analysis

1. **Proper List of Predictor Terms**
   * Logistic Regression (glmnet): Uses all predictors available in the dataset.
   * LDA: Also uses all predictors available in the dataset.
   * KNN: Uses all predictors as well, as KNN does not inherently perform feature selection.

# Selected Predictor Terms and Coefficients

* + Logistic Regression (glmnet): The specific coefficients for each predictor are not detailed in your output, but glmnet typically selects variables based on the penalty applied during training. The mix between L1 and L2 regularization is controlled by the alpha parameter. The optimal lambda selected is from cross-validation but not specified in your output.

# Cross-validated Test Error, Specificity, and Sensitivity

* + Logistic Regression (glmnet):
    - ROC: Approx. 0.9997 at best parameters.
    - Sensitivity: Up to 0.62 at lower lambdas, decreases as lambda increases.
    - Specificity: 1 across most models.
  + LDA:
    - ROC: 0.9973781
    - Sensitivity: 0.3733333
    - Specificity: 1
  + KNN:
    - ROC: Highest at 0.7446258 for k=7.
    - Sensitivity: Very low (max 0.02 for k=3).
    - Specificity: Essentially perfect at nearly 1.0.

These test errors represent the proportion of the test dataset for which each model predicted the wrong category (Pass/Fail).

* + Logistic Regression (glmnet):

Accuracy: 0.9839704

Test Error = 1−0.9839704=0.01602961−0.9839704=0.0160296

* + K-Nearest Neighbors (KNN):

Accuracy: 0.9839704

Test Error = 1−0.9839704=0.01602961−0.9839704=0.0160296

* + Linear Discriminant Analysis (LDA):

Accuracy: 0.9901356

Test Error = 1−0.9901356=0.00986441−0.9901356=0.0098644

The frequency with which each model mispredicts whether students will pass or fail based on their results is shown by the test error rates computed for three distinct models. The accuracy of the K-Nearest Neighbors (KNN) and Logistic Regression (glmnet) models is 98.39704%, with a test error of 1.60296%. This indicates that these models forecast incorrectly around 1.6% of the time. Conversely, the Linear Discriminant Analysis (LDA) model exhibits a lower test error of 0.98644% due to its greater accuracy of 99.01356%. This smaller test error indicates that, in comparison to the other two models, the LDA model is more successful in generalizing and producing accurate predictions on fresh, unexplored data.

# Optimal K for KNN

* + K: The optimal number of neighbors is 7, as this configuration provided the highest ROC value.

# Comparison of Models and Determination of the Optimal Model

* + The LDA model shows a strong balance between ROC and specificity, although sensitivity is relatively low. This might suggest that while the model is excellent at predicting true negatives (specificity of 1), it struggles with true positives (lower sensitivity).
  + The Logistic Regression model shows very high ROC and perfect specificity but also suffers from a similar issue with sensitivity, especially at higher lambda values where regularization is stronger.
  + The KNN model, while having the highest specificity, suffers from very low sensitivity and a lower ROC compared to the other models.

# Optimal Model:

Considering all metrics, the LDA model might be considered the best among the three due to its high ROC and perfect specificity while maintaining a better sensitivity score than the KNN model. Logistic Regression also performs well but the exact balance of lambda and alpha would need further refinement to improve sensitivity without losing much on specificity.

# Business/Research Insights:

* + The analysis shows that LDA, despite its simplicity compared to the regularized logistic regression model, provides robust predictions for the binary classification of passing or failing based on the provided predictors.
  + Sensitivity remains a challenge across models; thus, efforts should be directed towards improving the identification of true positives without sacrificing the high specificity achieved.
  + Business or research decisions should consider the trade-off between catching as many positive cases as possible (sensitivity) and maintaining the accuracy of the negative predictions (specificity), especially in scenarios where false negatives may carry higher risks or costs than false positive.

# Part IV: Conclusions and Recommendations

**Conclusions**

The comprehensive analysis performed using various statistical models and techniques has provided deep insights into factors influencing student performance, particularly regarding scoring above 35 marks. Key findings from this study include:

1. High Predictive Accuracy: The Linear Discriminant Analysis (LDA) model demonstrated the highest predictive accuracy, suggesting that the selected predictors are significantly influential in determining student performance.
2. Importance of Predictor Variables: Key variables such as study hours, participation in test preparation programs, and socioeconomic factors like lunch type were identified as critical in influencing students' scores. These variables correlate strongly with performance, indicating areas where interventions could be most effective.
3. Model Comparisons and Effectiveness: LDA outperformed other models, indicating its effectiveness in handling the categorical nature of the data and the underlying relationships between predictors and student performance.
4. Data Quality and Management: The handling of missing values and outliers has shown that proper data management directly impacts the quality of predictive outcomes, emphasizing the importance of clean and well-managed datasets for accurate analyses.

# Recommendations for Business/Research Decisions

Based on the conclusions derived from the analysis, the following recommendations are proposed to guide business or educational decisions:

1. Targeted Interventions: Based on model insights, targeted educational programs focusing on critical predictors like test preparation and study hours should be developed. These programs should aim to provide additional support where it is most needed, potentially lifting the lower-performing students to achieve above the pass threshold.
2. Policy Development: Educational policymakers should consider these findings to develop or adjust policies that ensure equitable access to resources that significantly impact

performance, such as nutritional programs (free or reduced-price lunches) and after-school tutoring or test preparation services.

1. Continuous Monitoring and Adjustment: Institutions should implement continuous monitoring of the implemented changes' effectiveness through ongoing data collection and analysis. This iterative process will help fine-tune interventions and ensure that resources are being used effectively to improve student outcomes.
2. Investment in Data Systems: To support the ongoing data analysis needs, investment in robust data collection and analysis systems is recommended. These systems will facilitate the detailed tracking of student progress over time, helping to identify trends and adjust strategies promptly.
3. Collaborations and Partnerships: Schools should consider partnerships with local businesses and community organizations to support and fund programs that address the identified key predictors of student success. These partnerships can provide both financial support and opportunities for students to engage in real-world applications of their learning, potentially enhancing motivation and performance.
4. Professional Development: Educators should receive professional development to better understand the use of data in identifying student needs and differentiating instruction accordingly. This training will help teachers more effectively target their instructional strategies to meet diverse learners' needs within their classrooms.

By implementing these recommendations, educational institutions and policymakers can better align their strategies and resources with the factors that most significantly impact student success, thereby improving educational outcomes in a measurable and data-informed manner.

**Key Findings**

* Generally, students perform well in exams, with notable strengths in reading and writing.
* There is a strong correlation between high math scores and high scores in reading and writing.
* Student performance in exams is influenced by gender and ethnic background.
* Females tend to excel in writing and reading, whereas males generally perform better in math.
* The educational level of parents impacts student exam scores, with higher parental education correlating with higher student scores.
* Students from ethnic group E typically perform well in all exams, regardless of parental education level, even though they are a minority group.
* Students who have completed a test preparation course generally outperform those who have not.
* Students dedicating more than 10 hours a week to studying usually achieve high scores in math

# Part V References

* 1. Desalegn, G. (2023). *Students Exam Scores*. Kaggle. Retrieved from <https://www.kaggle.com/datasets/desalegngeb/students-exam-scores/data>
  2. Wikipedia contributors. (n.d.). Generalized linear model. In *Wikipedia, The Free Encyclopedia*. Retrieved from <https://en.wikipedia.org/wiki/Generalized_linear_model>
  3. Wikipedia contributors. (n.d.). Statistical classification. In *Wikipedia, The Free Encyclopedia*. Retrieved April 25, 2024, from <https://en.wikipedia.org/wiki/Statistical_classification>
  4. UCLA: Institute for Digital Research and Education. (n.d.). Logistic regression. Retrieved from https://stats.idre.ucla.edu/stata/dae/logistic-regression/
  5. Brownlee, J. (2018). How to Implement k-Nearest Neighbors From Scratch in Python. *Machine Learning Mastery*. Retrieved from https://machinelearningmastery.com/tutorial-to-implement-k-nearest-neighbors-in-python-from-scratch/
  6. MathWorks. (n.d.). Understand LDA. Retrieved from https://www.mathworks.com/help/stats/understand-linear-discriminant-analysis.html