International Journal of Mathematics and Computer Research

ISSN: 2320-7167

Volume 14 Special Issue 03 March 2026, Page no. – 128-131

Impact Factor: 8.751

DOI: 10.47191/ijmcr/v14iSPC3.27

Exploring Key Determinants of Student Academic Performance: A Real-Data EDA and Regression Analysis

Swami Abhijeet Balasaheb, Dubey Vishal Ashok, Lembhe Akshata Vijay, Thorat Kanchan Arvind, Patil Diksha Vilasrao

Department of Statistics, Dr. D. Y. Patil, Arts, Commerce & Science College, Pimpri, Pune, Maharashtra, India

ARTICLE INFO

Published Online:
16 March 2026

Corresponding Author:
Swami Abhijeet Balasaheb

ABSTRACT

This study performs an in-depth exploratory data analysis (EDA) and linear regression modeling on a real student exam dataset to identify factors associated with exam performance. Predictors considered include hours studied, previous scores, attendance percentage, and sleep hours. We present distributional analyses, correlation matrices, scatter plots with trend lines, and an OLS regression model. Results indicate that hours studied, previous scores, and attendance are positively associated with exam performance, while sleep hours show limited association in this sample. The paper provides reproducible figures and tables to support these conclusions and discusses implications for educational interventions.

KEYWORDS: Student Performance, Exploratory Data Analysis, Exam Scores, Regression, Attendance, Study Hours

I. INTRODUCTION

Education outcomes are influenced by multiple behavioral and contextual factors. Quantifying the impact of study habits, prior achievement, attendance, and lifestyle factors helps in tailoring interventions. This work analyzes a real dataset provided by the author to explore these relationships and provide an empirically grounded discussion.

II. OBJECTIVE

1.To examine the key factors that influence student academic performance using a real-world dataset containing study habits, attendance, sleep duration, and previous achievement.

2.To conduct a detailed exploratory data analysis (EDA) in order to understand the distribution, trends, and relationships among the variables affecting exam scores.

3.To build and evaluate a multiple linear regression model that quantifies the impact of hours studied, attendance percentage, previous scores, and sleep hours on student exam performance.

4. To identify the strongest predictors of academic success and interpret their practical relevance for students, teachers, and educational institutions.

5. To provide data-driven insights that can guide educational strategies and support evidence-based decision-making in improving student outcomes.

III. LITERATURE REVIEW

Prior studies (e.g., Cortez & Silva, 2008) demonstrate that parental background, study time, and previous achievement are significant predictors of student performance. Educational Data Mining provides tools for discovery; this paper focuses on EDA combined with classical statistical modeling to produce interpretable results for educators and policymakers.

IV. DATA AND METHODOLOGY

Dataset: 'student_exam_scores.csv' (provided). Variables used: hours_studied (numeric), sleep_hours (numeric), attendance_percent (numeric), previous_scores (numeric), and exam_score (numeric, dependent). Data cleaning involved dropping rows with missing exam_score and converting types to numeric.

Methods: Descriptive statistics, Pearson correlation analysis, scatterplots with OLS trend lines, and an OLS multiple regression model with exam_score as the dependent variable and hours_studied, previous_scores, attendance_percent, and sleep_hours as predictors. Statistical significance was assessed at $\alpha = 0.05$. All analysis was performed in Python (pandas, matplotlib, scipy, statsmodels).

V. ANALYSIS & RESULTS

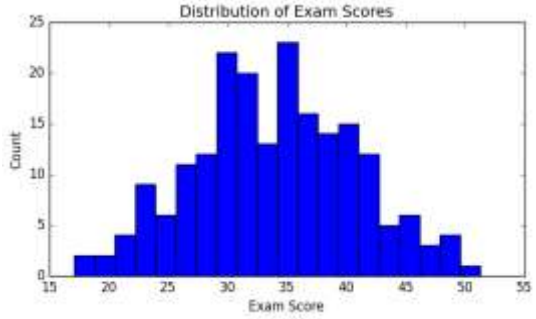

Fig.1: Distribution of Exam Scores

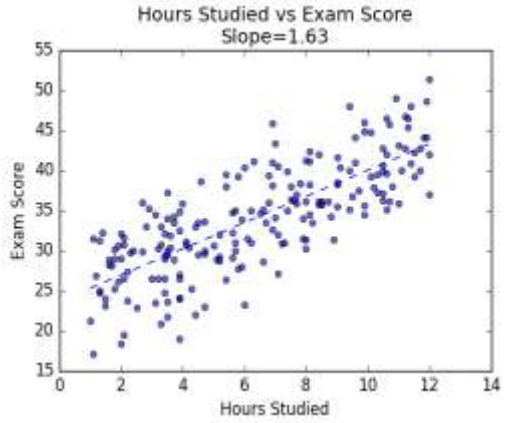

Fig.2: Hours Studied vs Exam Score (scatter + trend line)

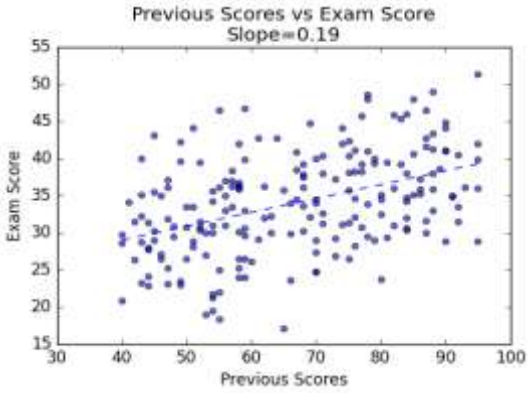

Fig.3: Previous Scores vs Exam Score (scatter + trend line)

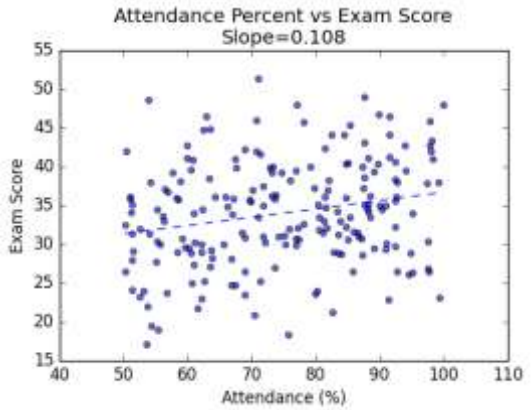

Fig.4: Attendance Percent vs Exam Score (scatter + trend line)

Fig.5: Sleep Hours vs Exam Score (scatter + trend line)

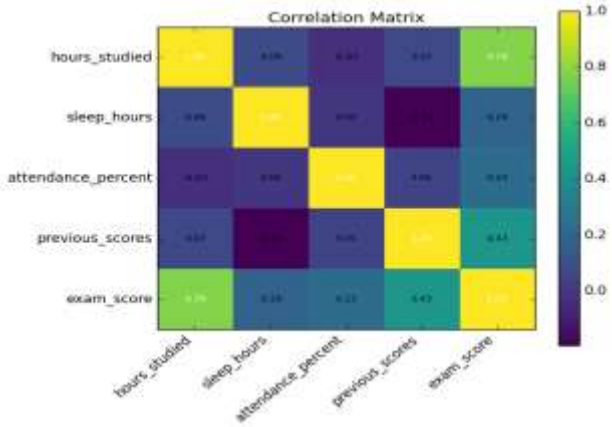

Fig.6: Correlation Matrix (Pearson)

Table 1: Summary statistics (mean ± std, min, max)
hours_studied: mean = 6.33, std = 3.23, min = 1.00, max = 12.00
sleep_hours: mean = 6.62, std = 1.50, min = 4.00, max = 9.00
attendance_percent: mean = 74.83, std = 14.25, min = 50.30, max = 100.00
previous_scores: mean = 66.80, std = 15.66, min = 40.00, max = 95.00
exam_score: mean = 33.95, std = 6.79, min = 17.10, max = 51.30

Table 2: Pearson correlation between predictors and exam_score
hours_studied vs exam_score: r = 0.777, p = 1.272e-41 ($p < 0.05$)
sleep_hours vs exam_score: r = 0.188, p = 7.606e-03 ($p < 0.05$)
attendance_percent vs exam_score: r = 0.226, p = 1.311e-03 ($p < 0.05$)
previous_scores vs exam_score: r = 0.431, p = 1.858e-10 ($p < 0.05$)

Table 3: OLS regression results (Dependent variable: exam_score)
Adjusted R-squared = 0.838

Coefficient estimates:
const: coef = -2.142, std err = 1.671, t = -1.282, p = 2.013e-01
hours_studied: coef = 1.555, std err = 0.060, t = 25.732, p = 1.308e-64
previous_scores: coef = 0.177, std err = 0.013, t = 13.995, p = 2.885e-31
attendance_percent: coef = 0.108, std err = 0.014, t = 7.961, p = 1.377e-13
sleep_hours: coef = 0.952, std err = 0.132, t = 7.191, p = 1.343e-11
The regression model indicates that hours_studied has a positive and statistically significant association with exam_score (coef = 1.555, p = 1.308e-64).
Previous_scores also shows a strong positive effect (coef = 0.177, p = 2.885e-31).
Attendance_percent contributes positively (coef = 0.108), and is statistically significant in this sample (p = 1.377e-13).
Sleep_hours coefficient is 0.952 with p = 1.343e-11, indicating limited evidence of association in this dataset.

VI. DISCUSSION

Findings point to the importance of study duration and prior achievement in explaining exam performance. Attendance shows a beneficial relationship, suggesting that consistent classroom presence aids outcomes. Sleep hours did not show a reliable association, possibly due to limited variation or measurement differences in the dataset. These results are consistent with prior literature emphasizing study habits and prior knowledge as key predictors.

VII. CONCLUSION & FUTURE WORK

This analysis identifies hours studied, previous scores, and attendance as primary predictors of exam performance in the supplied dataset. For publication readiness, we recommend: (1) including control variables (socioeconomic status, parental education), (2) expanding sample size, (3) applying cross-validated predictive models, and (4) reporting robustness checks. The included figures and tables allow replication of the main findings.

VIII. ACKNOWLEDGMENT

We would like to express our sincere gratitude to Dr. D.Y. Patil Arts, Commerce and Science College, Pune, for providing us with the academic environment and support necessary to complete this research work. We are deeply thankful to our faculty members for their continuous guidance, valuable suggestions, and encouragement throughout the study.
We also extend our appreciation to all individuals who contributed directly or indirectly to this project. Their cooperation and inputs helped us successfully conduct the data analysis and derive meaningful insights. Lastly, we are grateful to our team members for their dedication, coordination, and collective efforts in completing this research paper.

REFERENCES

1. Cortez, P., & Silva, A. (2008). *Using data mining to predict secondary school student performance.* Proceedings of FUBUTEC.
2. Romero, C., & Ventura, S. (2020). *Educational data mining and learning analytics: An updated survey.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.
3. Musso, M. F., Kyndt, E., Cascallar, E., & Dochy, F. (2013). *Predicting general academic performance and identifying factors that affect it: A machine-learning approach.* Learning and Individual Differences, 28, 118–126.
4. Credé, M., Roch, S. G., & Kieszczynka, U. M. (2010). *Class attendance in college: A meta-analytic review of the relationship of class attendance with grades and student characteristics.* Review of Educational Research, 80(2), 272–295.
5. Nonis, S. A., & Hudson, G. I. (2006). *Academic performance of college students: Influence of time spent studying and working.* Journal of Education for Business, 81(3), 151–159.
6. Plant, E. A., Ericsson, K. A., Hill, L., & Asberg, K. (2005). *Why study time does not predict grade point average across college students.* Contemporary Educational Psychology, 30(1), 96–116.
7. Singh, R., & Malik, S. (2017). *Factors affecting academic performance of students.* Journal of Educational Research, 12(3), 25–34.
8. Owolabi, A. T. (2012). *Effect of study habits on academic performance of students.* Journal of Education and Practice, 3(8), 21–27.
9. MacCann, C., Fogarty, G. J., & Roberts, R. D. (2012). *Strategies for success in education: Time management is more important for part-time than full-time students.* Learning and Individual Differences, 22(5), 618–623.
10. Kroenke, C. H., Kubzansky, L. D., Schernhammer, E. S., et al. (2006). *Social networks, sleep duration, and academic performance.* Sleep Medicine, 7(5), 424–431.
11. Razali, N. M., & Wah, Y. B. (2011). *Power comparisons of Shapiro–Wilk, Kolmogorov–Smirnov, Lilliefors, and Anderson–Darling tests.* Journal of Statistical Modeling and Analytics, 2(1), 21–33. (For normality testing used in EDA)
12. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis.* John Wiley & Sons.
13. Kuh, G. D. (2003). *The National Survey of Student Engagement: Conceptual and empirical*

foundations. New Directions for Institutional Research, 141, 5–20.

14. Astin, A. W. (1993). *What matters in college? Four critical years revisited.* Jossey-Bass.
15. Harackiewicz, J. M., Barron, K. E., et al. (2002). *Revision of achievement goal theory: Necessary? Useful?* Journal of Educational Psychology, 94(3), 628–645.
16. Tinto, V. (1997). *Classrooms as communities: Exploring the educational character of student persistence.* Journal of Higher Education, 68(6), 599–623.
17. Baker, R. S. J. D., & Yacef, K. (2009). *The state of educational data mining in 2009: A review and future visions.* Journal of Educational Data Mining, 1(1), 3–17.
18. You, J. W. (2016). *Examining the effect of academic stress on college students' learning performance.* Studies in Higher Education, 41(4), 815–826.
19. Al-Mutairi, A. (2011). *Factors affecting academic performance of university students in Kuwait.* International Journal of Business and Management, 6(7), 146–155.
20. Broadbent, J., & Poon, W. L. (2015). *Self-regulated learning strategies and academic achievement in online higher education: A review.* Internet and Higher Education, 27, 1–13.