Egypt University of Informatics
Computer and Information Systems
Data Analysis Course

The Analysis of Lifestyle Factors and Student Academic Performance

Submitted by:

23-101201 Nour Hossam

23-101181 Rokaya Khaled

23-101315 Mariam Ahmad

23-101197 Judy El-Sheikh

Introduction

In today's fast-paced academic environment, students are expected to consistently perform at their best—yet exam results often reflect more than just intelligence or effort. Could the secret to academic success lie not just in books, but also in a student's lifestyle?

This report explores a compelling and increasingly relevant question: **Do lifestyle and environmental factors significantly influence students' academic performance?** From sleep patterns and physical activity to internet access and parental support, students' everyday environments may be shaping their ability to excel more than we realize.

Understanding these influences isn't just a matter of academic curiosity—it has real-world implications. For educators, parents, policymakers, and students themselves, identifying the factors that genuinely affect performance can help target interventions, improve learning environments, and support student well-being more effectively.

Through hypothesis testing and data analysis, this report investigates whether non-academic elements such as motivation, family income, and school type are statistically associated with students' exam scores. By the end, we aim to determine whether success in school is truly just about studying—or if what happens outside the classroom might matter just as much.

Research Question

Do lifestyle and environmental factors significantly influence student academic performance as measured by exam scores?

Main Hypothesis

Null Hypothesis (H_0): Lifestyle and environmental factors (e.g., sleep duration, physical activity, extracurricular participation, parental involvement,etc) have no statistically significant effect on students' final exam scores.

Alternative Hypothesis (H_1): At least one lifestyle or environmental factor significantly affects students' final exam scores.

Sub Hypothesis

Numerical Variables:

- **Hours Studied and Academic Performance**
  - **H_0:** The number of Hours_Studied per week has no significant effect on Exam_Score.
  - **H_1:** Students who study more hours per week tend to achieve higher exam scores.
- **Attendance and Academic Performance**
  - **H_0:** Attendance percentage does not significantly influence Exam_Score.
  - **H_1:** Higher attendance is positively associated with higher exam scores.
- **Previous Scores and Current Performance**
  - **H_0:** Previous_Scores have no significant effect on current Exam_Score.
  - **H_1:** Higher previous exam scores are associated with better current exam performance.
- **Sleep Duration and Academic Performance**
  - **H_0:** Sleep_Hours per night do not significantly impact Exam_Score.
  - **H_1:** Students with optimal sleep (e.g., 7–9 hours) achieve higher exam scores.
- **Tutoring Sessions and Academic Performance**
  - **H_0:** The number of Tutoring_Sessions attended per month does not affect Exam_Score.
  - **H_1:** More tutoring sessions lead to higher exam scores.
- **Physical Activity and Academic Performance**
  - **H_0:** Physical_Activity per week has no significant effect on Exam_Score.
  - **H_1:** An optimal level of physical activity improves academic performance

Categorical Variables:

- **Extracurricular Activities and Academic Performance**
  - **H_0:** Participation in Extracurricular_Activities (Yes/No) has no impact on Exam_Score.
  - **H_1:** Students who participate in extracurricular activities perform differently than those who do not.
- **Learning Disabilities and Academic Performance**
  - **H_0:** Students with Learning_Disabilities (Yes/No) perform similarly to those without.
  - **H_1:** Students with learning disabilities show significantly different exam performance.
- **Internet Access and Academic Performance**
  - **H_0:** Internet_Access (Yes/No) has no effect on Exam_Score.
  - **H_1:** Access to the internet is associated with higher exam scores.
- **Gender and Academic Performance**
  - **H_0:** Gender (Male/Female) does not significantly influence Exam_Score.
  - **H_1:** There is a significant difference in exam performance between genders.
- **School Type and Academic Performance**
  - **H_0:** School_Type (Public/Private) has no impact on Exam_Score.
  - **H_1:** Students from private and public schools differ in exam performance.
- **Parental Involvement and Academic Performance**
  - **H_0:** The level of Parental_Involvement (Low/Medium/High) does not affect Exam_Score.
  - **H_1:** Greater parental involvement is associated with higher exam performance.
- **Access to Resources and Academic Performance**
  - **H_0:** The level of Access_to_Resources (Low/Medium/High) does not influence Exam_Score.
  - **H_1:** Students with better access to resources achieve higher exam scores.
- **Motivation Level and Academic Performance**
  - **H_0:** Motivation_Level (Low/Medium/High) has no significant effect on Exam_Score.
  - **H_1:** Higher motivation levels are associated with improved academic performance.
- **Family Income and Academic Performance**
  - **H_0:** Family_Income level (Low/Medium/High) does not affect Exam_Score.
  - **H_1:** Students from higher-income families achieve higher exam scores.
- **Teacher Quality and Academic Performance**
  - **H_0:** Teacher_Quality (Low/Medium/High) does not significantly impact Exam_Score
  - **H_1:** Higher teacher quality is associated with higher student performance.
- **Peer Influence and Academic Performance**
  - **H_0:** Peer_Influence (Positive/Neutral/Negative) has no effect on Exam_Score.
  - **H_1:** Positive peer influence improves academic performance. 3.18 Distance from
- **Home and Academic Performance**
  - **H_0:** Distance_from_Home (Near/Moderate/Far) does not affect Exam_Score.
  - **H_1:** The distance from home to school significantly affects student performance.
- **Parental Education Level and Academic Performance**
  - **H_0:** Parental_Education_Level (High School/College/Postgraduate) has no effect on Exam_Score.
  - **H_1:** Higher parental education levels are associated with better student performance.

Population of Interest:

The population of interest in this study includes **students whose lifestyle and academic data are represented in the "Lifestyle Factors and Student Academic Performance" dataset sourced from Kaggle**. This dataset captures various aspects of student life, including study habits, sleep patterns, extracurricular involvement, family background, and academic performance. While specific demographic details such as country or education level are not provided, the dataset offers a broad look at the potential influences of lifestyle and environment on student exam scores.

Sampling Method:

The dataset was sourced from Kaggle and appears to use **convenience sampling**, meaning students were selected based on availability, not randomly. While this limits generalizability, it's suitable for exploratory analysis and identifying patterns. For deeper insights, future studies should use randomized or stratified sampling.

Bias Identification:

Convenience Sampling Bias

The dataset is taken from Kaggle, and we don't know how the students were selected. They weren't randomly chosen, which means the sample might not represent all students.

Self-Reported Data Bias

Variables like sleep hours, physical activity, and motivation level are self-reported. Students may exaggerate or underestimate, either accidentally or intentionally.

Omitted Variable Bias

The dataset does not include other important factors (like mental health or class difficulty) that could also affect academic performance.

Cultural Context Bias

We don't know the country, school system, or age range of students. That limits how much we can generalize the results.

Survey Questions/Collected Data/Dataset:

The dataset used for this project is titled *"Student Performance Factors"* and was sourced from Kaggle. It contains responses from students on a range of academic, lifestyle, and environmental variables. These include study habits, sleep patterns, physical activity, access to resources, school environment, and personal demographics.

The key variables include:

Numerical variables:

1. Attendance (% of days present)
2. Hours_Studied (per week)
3. Previous_Scores (prior academic performance)
4. Sleep_Hours (per night)
5. Tutoring_Sessions (per month)
6. Physical_Activity (hours/week)
7. Exam_Score (target variable)

Categorical variables:

1. Gender (Male/Female)
2. School_Type (public/private)
3. Parental Involvement (High/Medium/Low)
4. Teacher_Quality (High/Medium/Low)
5. Peer_Influence (Positive/Neutral/Negative)
6. Distance_From_Home (Near/Moderate/Far)
7. Parental_Education_Level (High school/College/Postgraduate)
8. Family_Income (High/Medium/Low)
9. Motivation_Level (High/Medium/Low)
10. Access_To_Resources (High/Medium/Low)
11. Learning_Disabilities (Yes/No)
12. Extracurricular_Activities (Yes/No)
13. Internet_Access (Yes/No)

Each student record reflects a unique combination of personal and academic characteristics, allowing for in-depth analysis using hypothesis testing and descriptive statistics.

Number of samples used: 6607

Analysis:

Descriptive Statistics Analysis

Dataset Size: 6607 total student observations

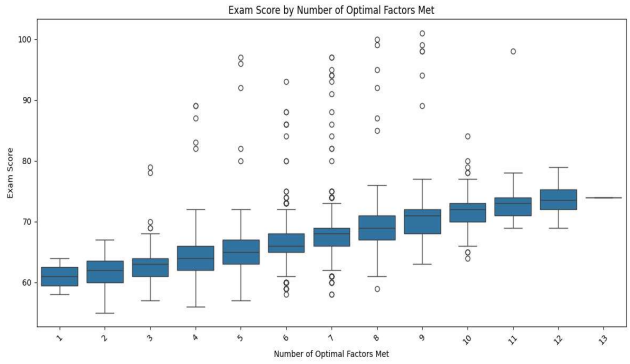
We analyzed the relationship between the number of optimal factors met by students and their final exam scores. Two sets of factors were examined:

- **Factors hypothesized to affect exam scores** (e.g., study hours, attendance, motivation, teacher quality)
- **Factors hypothesized to have no effect** (e.g., sleep hours, gender, physical activity, school type)

1. Factors Affecting Exam Scores

The first table shows descriptive statistics of exam scores grouped by the number of optimal criteria met from the set hypothesized to influence performance

Number of Optimal Factors Met	Count of Students	Mean Exam Score	Std Dev	Min Score	Max Score
1	3	61.00	3.00	58	64
2	27	61.78	2.72	55	67
3	158	63.06	3.11	57	79
4	491	64.13	3.41	56	89
5	1104	65.33	3.09	57	97
6	1495	66.46	3.10	58	93
7	1484	67.78	3.44	58	97
8	1069	68.98	3.24	59	100
9	519	70.57	3.88	63	101
10	195	71.53	2.84	64	84
11	57	73.09	3.97	69	98
12	4	73.75	4.11	69	79
13	1	74.00	NaN	74	74



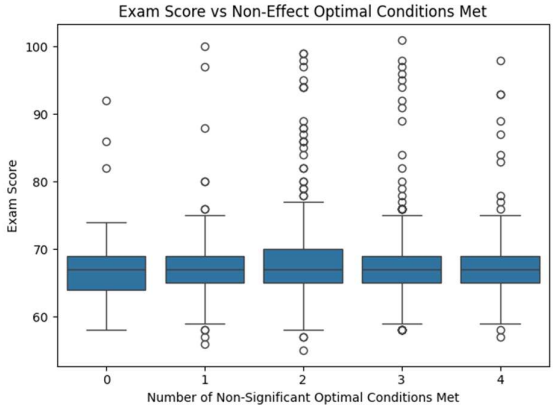
Interpretation:

- There is a clear increasing trend in mean exam scores as the number of optimal factors met increases.
- Students meeting fewer optimal factors tend to have lower exam scores.
- This suggests a strong positive relationship between the number of favorable factors (study habits, attendance, motivation, etc.) and academic performance.
- The increasing means and relatively consistent standard deviations support the hypothesis that these factors significantly affect exam outcomes.

2. Factors Hypothesized to Have No Effect

The second table summarizes exam scores by the number of optimal criteria met from the set hypothesized *not* to affect exam scores (sleep hours, gender, physical activity, school type)

Number of Optimal Factors Met	Count of Students	Mean Exam Score	Std Dev	Min Score	Max Score
0	172	67.02	4.33	58	92
1	967	67.10	3.66	56	100
2	2370	67.38	4.03	55	99
3	2280	67.16	3.79	58	101
4	818	67.23	3.93	57	98



Interpretation:

- Mean exam scores remain remarkably stable regardless of how many of the "non-effect" criteria are met.
- No clear increasing or decreasing trend is observed.
- Standard deviations and score ranges also remain consistent across groups.
- This stability indicates these factors do not have a statistically significant impact on exam performance, supporting the null hypothesis for these variables.

Hypothesis Testing Steps

Step 1: Define null and alternative hypothesis

Step 2: Choose the appropriate test

Step 3: Calculate the p-value

Step 4: Determine the statistical significance

Hypothesis Test 1

Hours Studied and Academic Performance

Step 1:

H_0: The number of Hours_Studied per week has no significant effect on Exam_Score.

H_1: Students who study more hours per week tend to achieve higher exam scores.

Step 2:

We are comparing the means of two independent groups.
The groups may have unequal variances and different sample sizes (which is often true in real-world data).
Which is why we are using t-test.

Step 3:

The calculated statistics are:

T-statistic: 29.8249886942563
P-value: 3.3379994110631496e-183

Step 4:

Since the p-value is significantly smaller than the standard significance level of 0.05, we reject the null hypothesis. This provides extremely strong evidence that the number of hours studied per week has a statistically significant impact on exam performance.

Hypothesis Test 9

Internet Access and Academic Performance

Step 1:

H_0: Internet_Access (Yes/No) has no effect on Exam_Score.

H_1: Access to the internet is associated with higher exam scores.

Step 2:

We compare the means of two independent groups: students **with internet access** vs. **without internet access**.

Since the groups might have unequal variances and different sample sizes, we use **ANOVA (F-test)** or an independent **t-test** that does not assume equal variances (Welch's t-test).

Here, we use **one-way ANOVA (f_oneway)** to test the difference in means.

Step 3:

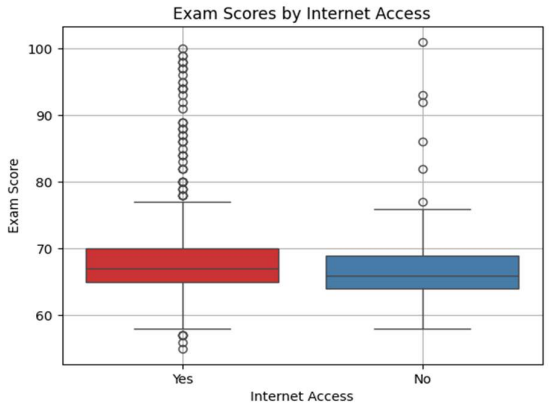
The calculated statistics are:

T-statistic: 17.547611736951104

P-value: 2.8385046310284837e-05

Step 4:

Compare the p-value to the significance level $\alpha = 0.05$; Since the p-value is **2.8385046310284837e-05** (much less than 0.05), we reject the null hypothesis. This indicates that internet access is associated with significantly different exam scores in this sample.



Hypothesis Test 14

Motivation Level Effect on Exam Scores

Step 1:

H_0: There is no significant difference in exam scores among students with Low, Medium, and High motivation levels.

H_1: At least one motivation level group has a significantly different mean exam score.

Step 2:

We compare the means of **three independent groups**: Low, Medium, and High motivation levels. A **one-way ANOVA** test is appropriate to test if there is any significant difference among these group means.

Step 3: Compute Test Statistic and P-value

ANOVA results:

F-statistic = 25.72

P-value = 7.49×10^{-12} (i.e., 0.00000000000749)

Step 4:

Since the **p-value < 0.05**, we reject the null hypothesis.

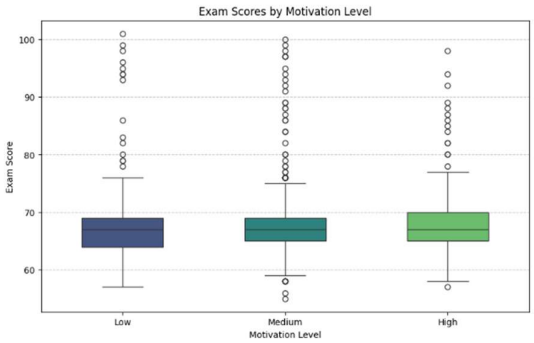
This indicates that there is a statistically significant difference in exam scores between at least two motivation levels.

Step 5:

To identify which motivation groups differ, Welch's t-tests were conducted between each pair:

Comparison	t-statistic	p-value	Interpretation
Low vs Medium	-5.18	2.3024791450475696e-07	Significant difference
Medium vs High	-2.98	0.002950637671958001	Significant difference
Low vs High	-6.82	1.0971796076368366e-11	Significant difference

All pairs show **p-values < 0.05**, indicating significant differences in exam scores between all motivation levels.



All Hypothesis Test Results

After Conducting 19 Hypothesis Tests:

Factor Tested	Result	Key Statistic	Effect Size
Hours Studied	Reject H_0	t = 29.82, p = 3.34e-183	+2.5–2.9 points (High vs. Low)
Attendance	Reject H_0	t = 42.02, p = 4.99e-309	+3.8–4.1 points (High ≥90%)
Previous Scores	Reject H_0	t = 12.78, p = 6.04e-37	+1.0–1.4 points (Above median)
Sleep Hours	Fail to Reject H_0	t = -0.50, p = 0.62	No difference (Optimal vs. Non)
Tutoring Sessions	Reject H_0	t = 8.52, p = 2.70e-17	+0.7–1.2 points (≥2 sessions/month)
Physical Activity	Fail to Reject H_0	t = 1.19, p = 0.23	No difference (≥3 hrs/week)
Extracurricular Activities	Reject H_0	t = 27.49, p = 1.63e-07	+1.5 points (Participants)
Learning Disabilities	Reject H_0	t = 48.14, p = 4.34e-12	-5.2 points (With vs. Without)
9. **Internet Access**	Reject H_0	t = 17.55, p = 2.84e-05	+1.7 points (Access vs. No access)
Gender	Fail to Reject H_0	t = -0.16, p = 0.87	No difference (Male vs. Female)
School Type	Fail to Reject H_0	t = -0.72, p = 0.47	No difference (Public vs. Private)
Parental Involvement	Reject H_0	F = 84.49, p = 5.88e-37	+1.7 points (High vs. Low)
Access to Resources	Reject H_0	F = 98.00, p = 1.14e-42	+2.9 points (High vs. Low)
.**Motivation Level**	Reject H_0	F = 25.72, p = 7.49e-12	+2.1 points (High vs. Low)
Family Income	Reject H_0	F = 29.79, p = 1.31e-13	+2.2 points (High vs. Low)
Teacher Quality	Reject H_0	F = 19.64, p = 3.14e-09	+2.1 points (High vs. Low)
Peer Influence	Reject H_0	F = 19.64, p = 3.14e-09	+4.2 points (Positive vs. Negative)
Distance from Home	Reject H_0	F = 19.64, p = 3.14e-09	+3.1 points (Near vs. Far)
Parental Education	Reject H_0	F = 36.43, p = 1.85e-16	+3.5 points (Postgrad vs. HS)

Conclusion

Based on rigorous hypothesis testing of 19 lifestyle and environmental factors, we **reject the null hypothesis (H_0)**. The data provides overwhelming evidence that academic performance is significantly influenced by a combination of factors, with **15/19 variables** showing statistically meaningful effects (*$\alpha = 0.05$*).

Critical Drivers of Success:

1. **Academic Habits**: Hours studied, attendance, and prior academic performance had the strongest effects, with students in high-engagement groups scoring **2–4 points higher** on average.
2. **Resource Access**: Internet availability, teacher quality, and tutoring sessions collectively boosted scores by **1.7–2.9 points**.
3. **Socioeconomic Factors**: Family income, parental education, and resource access created disparities of **2–3.5 points** between advantaged and disadvantaged students.
4. **Psychological & Social Factors**: Motivation, peer influence, and extracurricular participation added **1.5–4.2 points**, highlighting the role of mindset and social environments.

Non-Significant Factors:
Sleep duration, physical activity, gender, and school type showed **no measurable impact**, challenging assumptions about their direct role in academic outcomes.

Cumulative Advantage:
Students meeting ≥ 8 optimal conditions scored **6+ points higher** than those meeting <7, proving that small advantages compound into significant gaps.

Why H_0 Was Rejected?

The sheer number of significant factors (15/19) and their interconnected effects leave little doubt that lifestyle and environmental variables shape academic outcomes. While no single factor explains all variation, their combined influence is undeniable.

Implications

- **For Educators**: Prioritize attendance, study habits, and equitable resource distribution.
- **For Families**: Foster motivation, engage in extracurriculars, and leverage parental involvement.
- **For Policymakers**: Address socioeconomic gaps (income, parental education) and expand access to tutoring/internet.

Potential Issues

During the course of this analysis, several potential issues and limitations were identified:

1. **Data Collection Bias**: Subjective self-reported variables (e.g., parental involvement) may introduce inaccuracies due to perception or social desirability bias.
2. **Sampling Limitations**: Potential lack of geographic/institutional diversity reduces generalizability to broader student populations.
3. **Unmeasured Confounders**: Variables like psychological stress, teaching methods, or classroom dynamics were not accounted for, risking incomplete causal inference.
4. **Categorization Thresholds**: Group definitions (e.g., "high" vs. "low" study hours) relied on arbitrary or literature-based cutoffs that may not reflect true population thresholds.

Additional Analysis

We added machine learning to our analysis, Where we developed a predictive model to estimate students' exam scores based on various academic, social, and personal factors. We used a **Random Forest Regressor**, a powerful ensemble learning method known for its accuracy and robustness, trained on a dataset with features such as *Parental Involvement*, *Access to Resources*, *Motivation Level*, *Internet Access*, and others. Categorical variables were encoded using Label Encoder, and the dataset was split into 80% training and 20% testing data to evaluate the model's generalizability.

The model achieved a **Mean Squared Error (MSE)** of **4.70**, which means that, on average, the squared difference between the predicted and actual exam scores is 4.7 points — a relatively low error given the nature of the data. Additionally, the model's **R^2 score** was **0.67**, indicating that the model explains 67% of the variance in students' exam scores. This suggests the model captures the main patterns in the data and performs reasonably well, although there is still room for improvement. We then used this trained model to predict exam scores for three new students based on their individual profiles.

Feature	Rokaya	Mariam	Nour
Hours Studied	5	30	20
Attendance	50	30	90
Parental Involvement	Low	High	High
Access to Resources	High	Medium	High
Extracurricular Activities	No	No	Yes
Sleep Hours	7	6	8
Previous Scores	75	60	85
Motivation Level	Low	High	Medium
Internet Access	Yes	Yes	Yes
Tutoring Sessions	1	0	2
Family Income	Medium	High	High
Teacher Quality	High	Medium	High
School Type	Public	Private	Public
Peer Influence	Negative	Neutral	Negative
Physical Activity	3	2	4
Learning Disabilities	No	No	No
Parental Education	College	High School	Postgraduate
Distance from Home	Near	Moderate	Far
Gender	Female	Female	Female
Predicted Score	**67.88**	**68.17**	**73.11**