Modeling Student Success: How Lifestyle and Demographics Affect Academic Performance

AAI 500-02 – Final Project

Team 2: Paul Ancalima, Joseph Edwards, & Erika Gallegos

Introduction

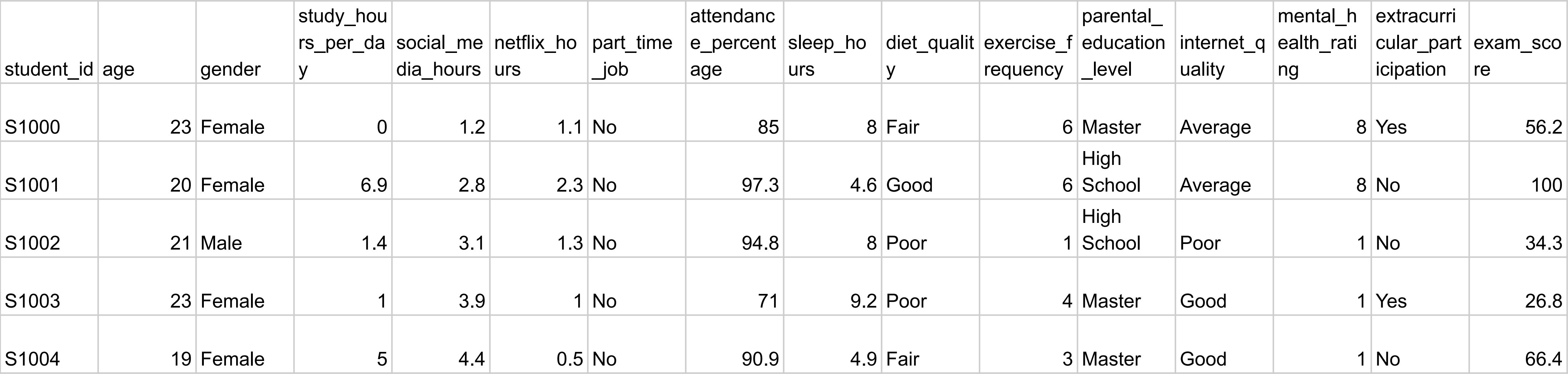
- Data Overview
  - How we decided on cleaning and preparing the data.
- Exploratory Data Analysis
  - How we understand relationships between our variables
- Regression
  - Where we determine which factors matter most
- Goals:
  - What patterns of association exist among students’ habits and background factors?
  - How do student habits and demographic variables affect exam scores?
  - What factors are most influential in determining whether a student achieves a passing versus failing exam grade?

Data Overview

- 1,000 student records
- 14 variables capturing habits, lifestyle, and demographics
    - Examples: study hours, social media, sleep, mental health, parental education
- Missing values (parental_education_level)
- Target variable: Exam Score (0-100)

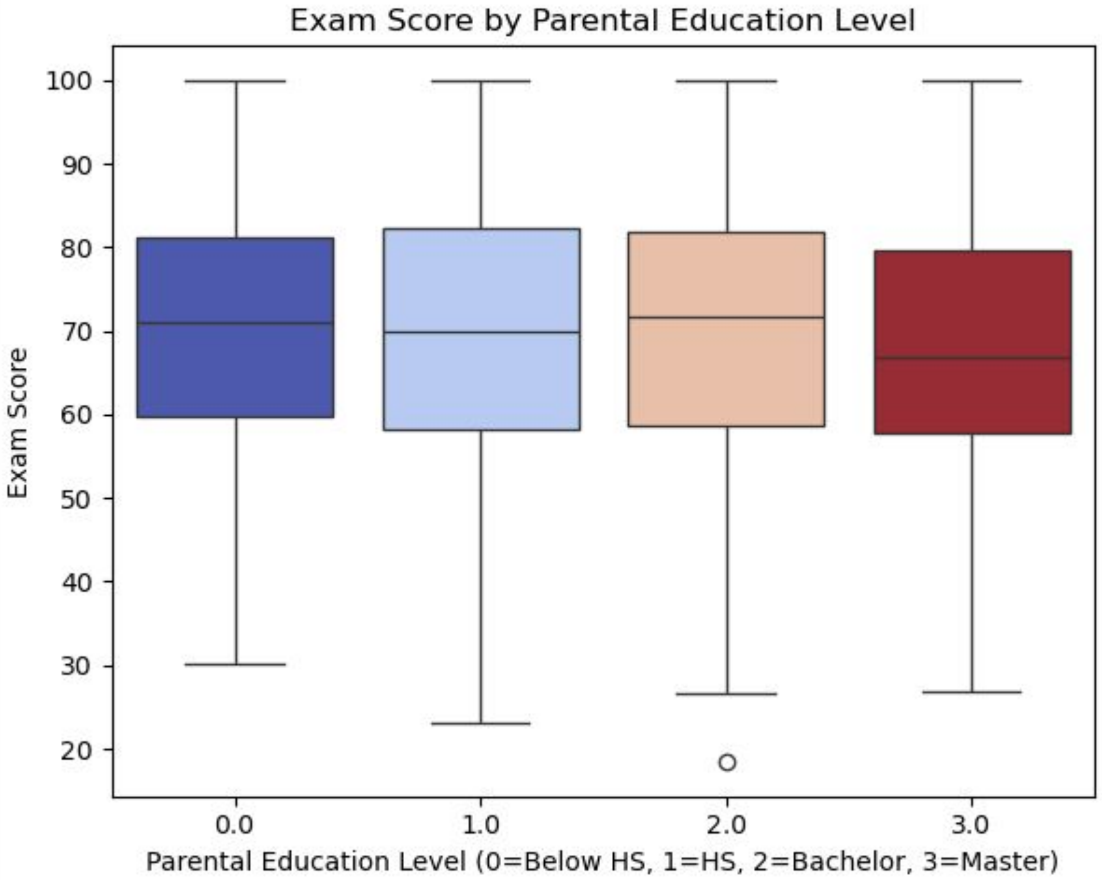
Dataset Overview

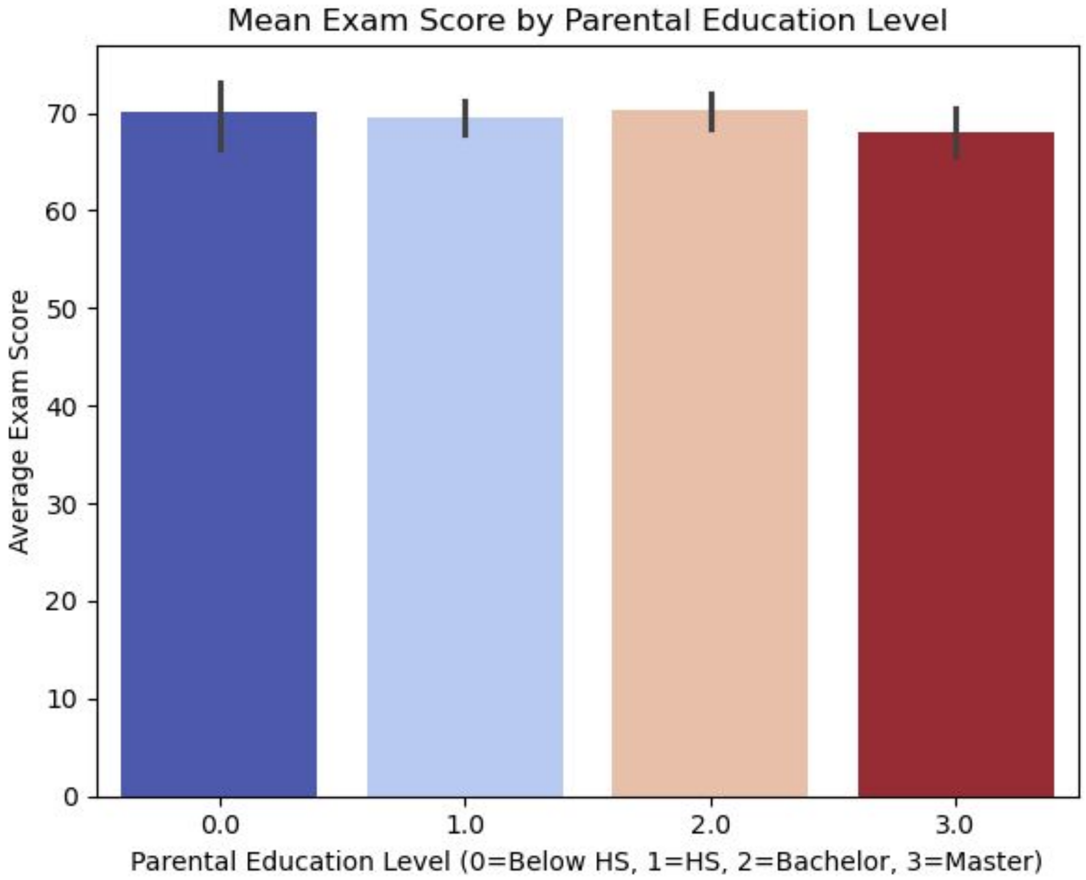
Student Habits and Exam Performance



student_id	age	gender	study_hours_per_day	social_media_hours	netflix_hours	part_time_job	attendance_percentage	sleep_hours	diet_quality	exercise_frequency	parental_education_level	internet_quality	mental_health_rating	extracurricular_participation	exam_score
S1000	23	Female	0	1.2	1.1	No	85	8	Fair	6	Master	Average	8	Yes	56.2
S1001	20	Female	6.9	2.8	2.3	No	97.3	4.6	Good	6	High School	Average	8	No	100
S1002	21	Male	1.4	3.1	1.3	No	94.8	8	Poor	1	High School	Poor	1	No	34.3
S1003	23	Female	1	3.9	1	No	71	9.2	Poor	4	Master	Good	1	Yes	26.8
S1004	19	Female	5	4.4	0.5	No	90.9	4.9	Fair	3	Master	Good	1	No	66.4

Data Visualization


Exam Score by Parental Education Level


Mean Exam Score by Parental Education Level

Data Cleaning

- Checked for missing values
- Dropped Student_ID column
- Found 91 missing values in parental_education_level
  - Tested their impact using a t-test
    - Showed no significant effect(t=0.238, p=0.795)
- Retained the rows for completeness and potential relevance in multivariate modeling
- Encoded categorical variables for modeling

Exploratory Data Analysis

Approaches

- Correlation Analysis: Relationship analysis
- Effect Size Quantification: How important that variables effect is
- ANOVA: Significance Across Groups
- Multiple Linear Regression: Shows each variables influence when controlling for other variables

Key Takeaways

- Highest predictors of exam scores were study time and mental health consistently across tests.
- Other variables deemed to have little effect on exam scores.
- Checking assumptions can give us higher confidence in results.

Assumptions Checked

ANOVA/T-Tests

- Normality
- Sample Size: n = 1000
- Equal Variances
- Where Variance Unequal, Compensated by Large n

Chi-Squared

- Independence
- Sample Size: n = 1000
- Minimum Frequencies Exceeded

RQ1: What patterns of association exist among students’ habits and background factors?

Attributes

- Study Hours per Day
- Mental Health Ratings
- Exercise Frequency
- Diet Quality
- Netflix Hours

Tests

- T-Tests: 2 Group Comparisons
- ANOVA: 3+ Group Comparisons
- Chi-Squared: Independence of Groups

T-Tests

2-Group Comparisons With Exam Scores

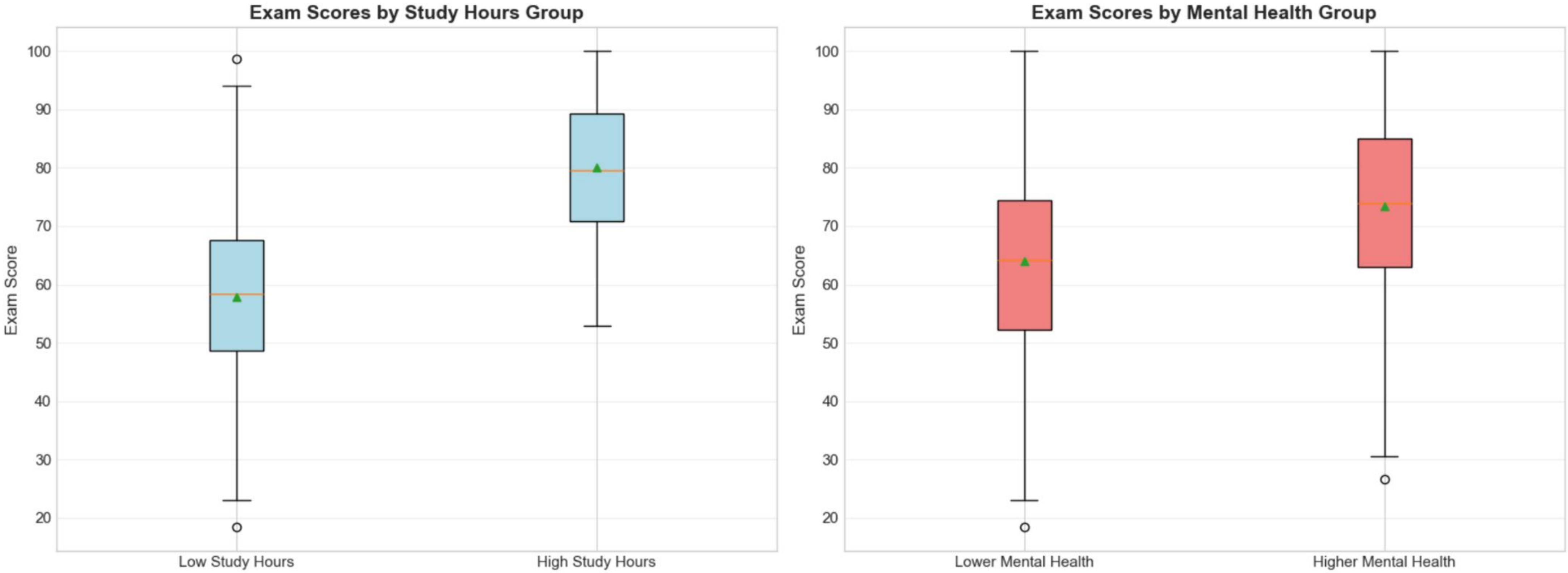
- **Low v High Study Hour Groups**
- **Hypothesis**:
  - H_0: Mean exam scores are equal between low and high study hour groups
  - H_1: Mean exam scores differ between low and high study hour groups
- **Results: Reject H_0**
  - T- Stat=27.71, P-Value=8.4759e-126, Effect Size=1.76, Significant at Alpha=0.05
- **Interpretation**: There is a large positive effect on exam score by higher study hours by a mean difference of 22.3 points.

T-Tests

2-Group Comparisons With Exam Scores

- **Low v High Mental Health Groups**
- **Hypothesis**:
  - H_0: Mean exam scores are equal between low and high mental health groups
  - H_1: Mean exam scores differ between low and high mental health groups
- **Results: Reject H_0**
  - T- Stat=9.05, P-Value=7.1393e-19, Effect Size=0.58, Significant at Alpha=0.05
- **Interpretation**: There is a moderate positive effect on exam score by higher mental health by a mean difference of 9.45 points.

Visualization of T-Tests



ANOVA Tests

3+ Group Comparison

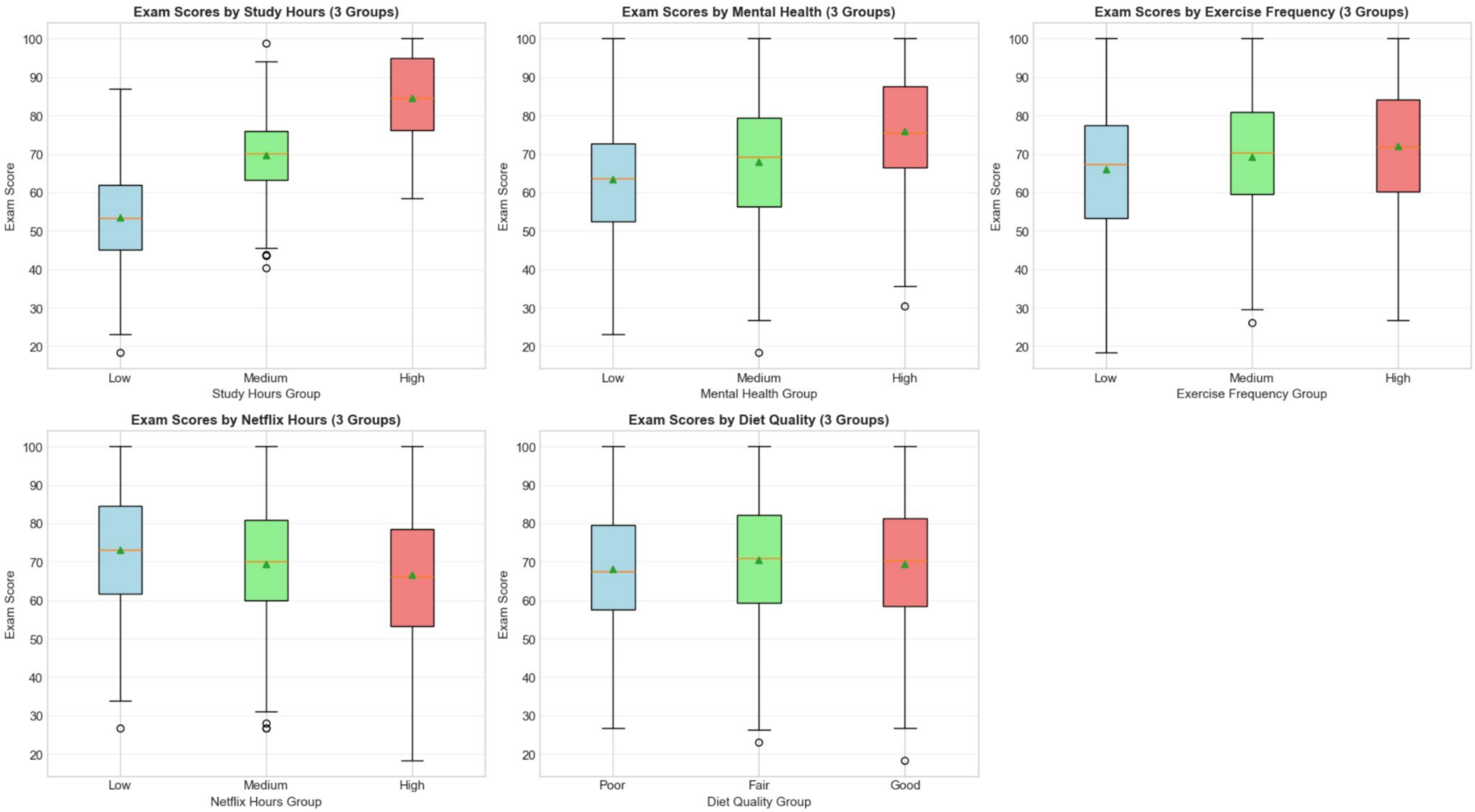
- **Mental Health Rating:** Low, Medium, and High groups
- **Hypothesis**:
  - H_0: Mean exam scores are equal between low, medium and high study hour groups
  - H_1: At least one group mean differs between low, medium and high study hour groups
- **Results: Reject H_0**
  - T- Stat=614.67, P-Value=1.1905e-174, Effect Size=0.55, Significant at Alpha=0.05
- **Interpretation**: There is a significant difference in means. Higher mental health groups outperform other groups.

ANOVA Tests

3+ Group Comparison

- **Study Hours Per Day:** Low, Medium, and High groups
- **Hypothesis**:
  - H_0: Mean exam scores are equal between low, medium and high study hour groups
  - H_1: At least one group mean differs between low, medium and high study hour groups
- **Results: Reject H_0**
  - T- Stat=614.67, P-Value=1.1905e-174, Effect Size=0.55, Significant at Alpha=0.05
- **Interpretation**: There is a significant difference in means. High study hour groups outperform other groups.

Visualization of ANOVA Tests



Chi-Squared Tests

Exam Scores Converted to Low, Medium, and High Categories

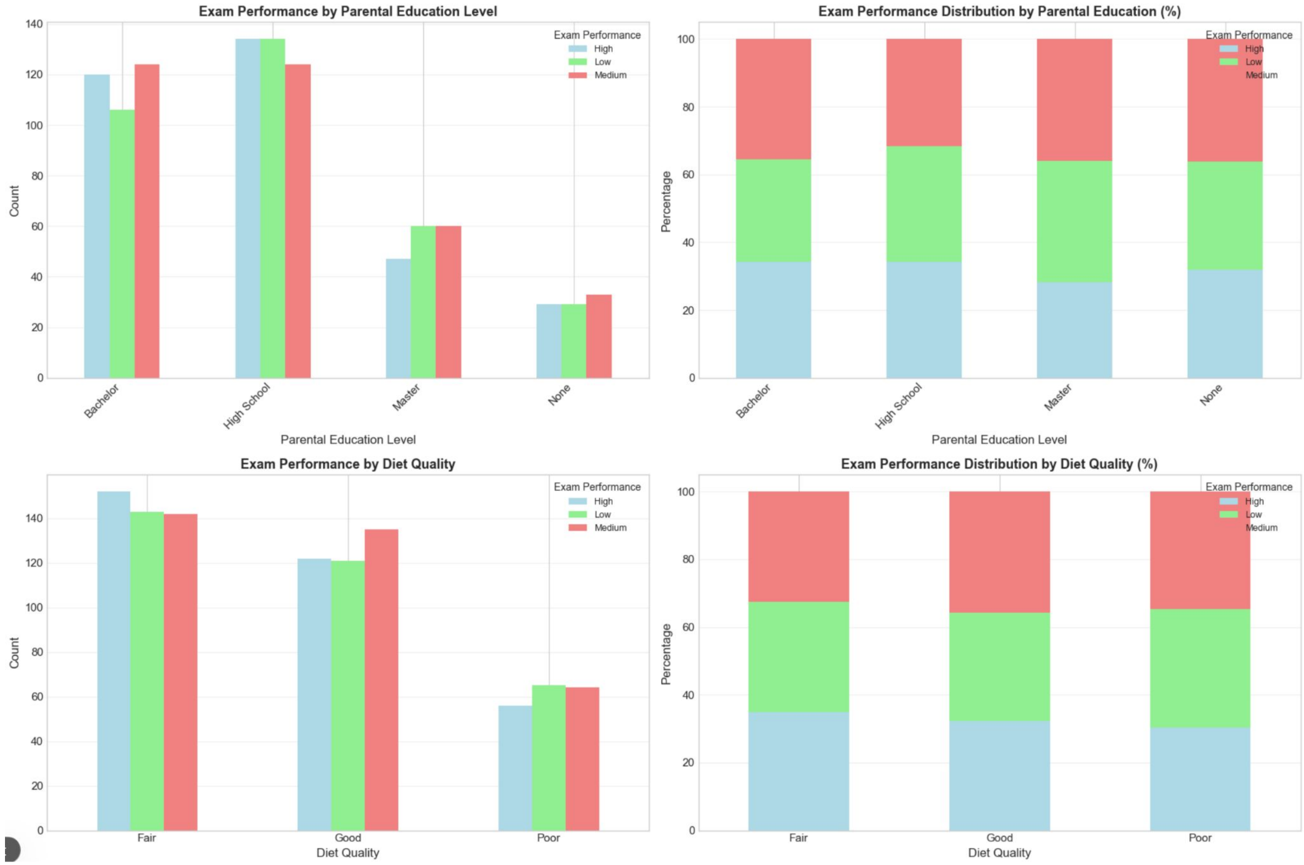
- **Parental Education Levels:** None, High School, Bachelors, Masters
- **Hypothesis**:
  - H_0: Exam Performance is independent of parental education level.
  - H_1: Exam Performance is dependent on parental education level.
- **Results: Fail to Reject H_0**
  - $\chi^2=4.16$, DOF=6, P-Value=0.66, Effect Size=0.05, NOT Significant at Alpha=0.05
- **Interpretation**: Exam performance is not significantly associated with parental education level.

Chi-Squared Tests

Exam Scores Converted to Low, Medium, and High Categories

- **Diet Quality:** Poor, Fair, Good
- **Hypothesis**:
  - H_0: Exam Performance is independent of diet quality.
  - H_1: Exam Performance is dependent on diet quality.
- **Results: Fail to Reject H_0**
  - χ^2=1.91, DOF=4, P-Value=0.75, Effect Size=0.03, NOT Significant at Alpha=0.05
- **Interpretation**: Exam performance is not significantly associated with diet quality.

Visualization of Chi-Squared Tests



Exam Performance by Parental Education Level

Exam Performance Distribution by Parental Education (%)

Exam Performance by Diet Quality

Exam Performance Distribution by Diet Quality (%)

RQ2: Predicting Exam Score

- How do student habits and demographic variables affect exam scores?
  - *Method:* Linear Regression
  - *Dependent Variable:* Exam Score (continuous)
  - *Train:* 80% split
  - *Test:* R^2, Adjusted-R^2, Residuals (linearity, homoscedasticity, normality)

$$\boxed{E(Exam\ Score) = \beta_0 + \beta_1(x_1) + \cdots \beta_n(x_n)}$$

RQ2: Predicting Exam Score

Variable	Coeff.	Std Error	t-statistic	p-value	95% CI [Lower, Upper]	
Intercept	8.24	2.95	2.79	0.005	2.45	14.02
Social Media Hours	-2.65	0.17	-16.02	< 0.001	-2.97	-2.32
Netflix Hours	-2.36	0.18	-13.03	< 0.001	-2.71	-2.00
Study Hours	9.50	0.13	72.71	< 0.001	9.25	9.76
Attendance Percentage	0.14	0.02	6.86	< 0.001	0.10	0.18
Sleep Hours	1.92	0.16	12.01	< 0.001	1.61	2.24
Exercise Frequency	1.49	0.09	15.76	< 0.001	1.30	1.67
Mental Health Rating	1.95	0.07	28.09	< 0.001	1.81	2.08
Age	-0.03	0.09	-0.34	0.732 (ns)	-0.20	0.14
Gender - Female	-0.26	0.40	-0.66	0.511 (ns)	-1.04	0.52
Gender - Other	0.41	0.96	0.43	0.669 (ns)	-1.48	2.30
Parent Edu. - Bachelor's	0.05	0.43	0.11	0.911 (ns)	-0.80	0.90
Parent Edu. - Master's	-0.21	0.55	-0.39	0.698 (ns)	-1.30	0.87
Part-Time Job - Yes	0.29	0.47	0.63	0.530 (ns)	-0.62	1.21
Extracurricular Partic. - Yes	-0.31	0.42	-0.75	0.452 (ns)	-1.13	0.50
Diet Quality - Fair	0.37	0.55	0.68	0.497 (ns)	-0.70	1.45
Diet Quality - Good	-0.24	0.56	-0.44	0.662 (ns)	-1.33	0.85
Internet Quality - Average	0.18	0.58	0.31	0.758 (ns)	-0.95	1.31
Internet Quality - Good	-0.54	0.57	-0.95	0.342 (ns)	-1.65	0.58

Decrease exam score

Increase exam score

Key Takeaway: Each additional hour studying per day associated with 9.50-point increase in exam score

Adj R^2 = 0.899

RQ3: Predicting Pass/Fail

- What factors are most influential in determining whether a student achieves a passing versus failing exam grade?
  - *Method:* Binary Logistic Regression
  - *Dependent Variable:* Pass (1), Fail (0)
  - *Train:* 80% split
  - *Test:* Accuracy, Precision, Recall, Specificity, F1 Score, AUC

$$\boxed{log\left(\frac{P(Pass)}{1-P(Pass)}\right) = \beta_0 + \beta_1(x_1) + \cdots \beta_n(x_n)}$$

RQ3: Predicting Pass/Fail

Variable	Coeff.	Std Error	z-statistic	p-value	Odds Ratio
Intercept	-23.77	2.91	-8.17	< 0.001	0.00
Social Media Hours	-1.00	0.16	-6.35	< 0.001	0.37
Netflix Hours	-0.95	0.17	-5.67	< 0.001	0.39
Study Hours	3.69	0.34	10.81	< 0.001	40.16
Attendance Percentage	0.07	0.02	4.20	< 0.001	1.07
Sleep Hours	0.79	0.13	6.02	< 0.001	2.20
Exercise Frequency	0.60	0.09	6.52	< 0.001	1.83
Mental Health Rating	0.82	0.09	9.57	< 0.001	2.27
Age	-0.08	0.07	-1.24	0.215 (ns)	0.92
Gender - Female	0.10	0.31	0.32	0.747 (ns)	1.11
Gender - Other	-1.01	0.68	-1.49	0.137 (ns)	0.37
Parent Edu. - Bachelor’s	0.19	0.33	0.57	0.571 (ns)	1.21
Parent Edu. - Master’s	0.11	0.44	0.24	0.807 (ns)	1.11
Part-Time Job - Yes	0.07	0.36	0.18	0.855 (ns)	1.07
Extracurricular Partic. - Yes	-0.16	0.33	-0.49	0.624 (ns)	0.85
Diet Quality - Fair	0.36	0.42	0.86	0.390 (ns)	1.44
Diet Quality - Good	-0.05	0.42	-0.11	0.914 (ns)	0.96
Internet Quality - Average	-0.56	0.44	-1.28	0.201 (ns)	0.57
Internet Quality - Good	-0.34	0.43	-0.79	0.431 (ns)	0.71

Key Takeaway: Each additional hour studying per day increased odds of passing by 40.16

Predictions on Test Data:

- Accuracy = 0.868
- Precision = 0.884
- Recall = 0.866
- Specificity = 0.871
- F1 Score = 0.875
- AUC = 0.959

Collaborative Efforts

Paul Ancalima

- Data Cleaning
- Exploratory Data Analysis

Joseph Edwards

- Exploratory Data Analysis
- Predictor Relationships
- Paper Conclusions

Erika Gallegos

- Linear Regression
- Logistic Regression
- Paper Intro & Conclusions