Week 5 Assignment – Statistics & Analytical Techniques

Name: Anju M M

Dataset: Student Scores (student-scores.csv) – From Kaggle

Date: November 29, 2025

Dataset Description

- **Id -** Unique identifier assigned to each student
- **first_name -** The first name of a student.
- **last_name -** The last name of a student.
- **Email -** The email address of a student
- **Gender -** The gender of a student.
- **part_time_job -** This indicates whether a student is engaged in a part-time job
- **absence_days -** The total count of days the student was not present in class due to various reasons.
- **extracurricular_activities -** This captures whether a student participates in extracurricular activities. It could include clubs, sports, arts, or other activities outside of regular academic coursework.
- **weekly_self_study_hours -** This represents the number of hours a student spends on self-study each week. It indicates the amount of time the student dedicates to independent learning and studying outside of class.
- **career_aspiration -** This column records the student's career aspirations or goals for the future. It provides insight into the profession or field the student aims to pursue after completing their education.
- **math_score -** The score obtained by a student in the subject of mathematics (0 – 100).
- **history_score -** History score (0 – 100)
- **physics_score -** Physics score (0 – 100)
- **chemistry_score -** Chemistry score (0 – 100)
- **biology_score -** Biology score (0 – 100)
- **english_score -** English score (0 – 100)
- **geography_score -** Geography score (0 – 100)

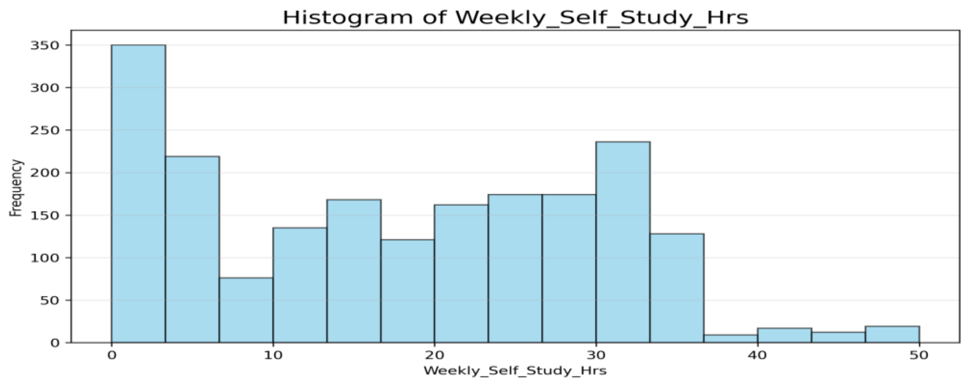
Basic Information

- **Number of rows:** 2,000
- **Number of columns:** 17
- **Total cells:** 34,000
- **Missing values:** 0 (No missing values in any column – dataset is 100% complete)
- **Duplicate rows:** 0

1. Descriptive Statistics

Statistic	Value
Mean	17.7555
Median	18.0
Mode	3
Range	50
Standard Deviation	12.129603595818578

Plot Histogram of any numerical data



The majority of students study 0–5 hours per week – that first bar is by far the tallest (over 350 students).

There is a second, smaller peak around 25–30 hours, and a long right tail with very few students studying more than 35–40 hours per week.

Overall, the distribution is heavily right-skewed: most students spend relatively little time on self-study, while only a small group invests a large number of hours each week.

2. Probability Analysis (Categorical column: gender)

Gender	**Probability**
Male	0.499
Female	0.501

Three Probability Questions

1. What is the probability a randomly selected student is female? → **0.501**
2. What is the probability a student has a part-time job? → **0.1580**
3. What is the probability that a student's career aspiration is Doctor?→ **0.0595**

Theoretical vs. Experimental Probability

- **Theoretical Probability** of a student being female is approximately 0.5 (assuming an equal gender ratio in the general population)
- **Experimental Probability** (calculated from the data) is 0.501
- The experimental probability is **very close** to the theoretical probability

3. Correlation Analysis

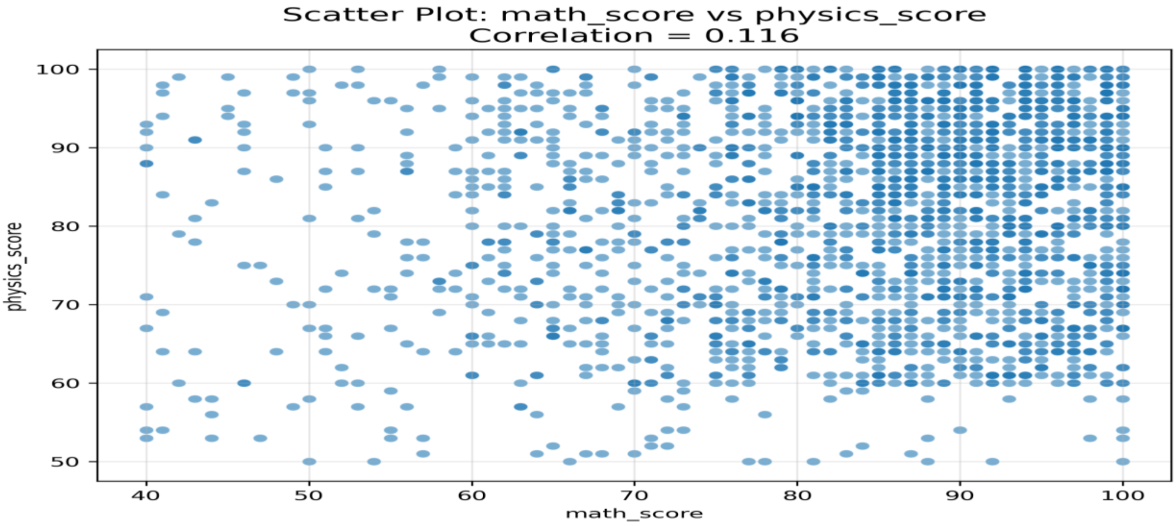
Variables Analyzed:

- Independent (x): math_score
- Dependent (y): physics_score

Correlation Coefficient (r): 0.116

R-squared: 0.0134 (only 1.34% of variation explained)

Scatter Plot: A scatter plot was created to visually represent this very weak positive relationship



Interpretation:

- **Strength:** Very weak (r = 0.116)
- **Direction:** Positive
- **Conclusion:** There is almost **no meaningful linear relationship** between math score and physics score. Students who score high in math do **not necessarily** score high (or low) in physics — the two subjects appear largely independent in this dataset.

Regression Analysis (Prediction)

Simple Linear Regression

Independent Variable (X): weekly_self_study_hours

Dependent Variable (y): math_score

Regression Equation:

math_score = 0.429 * weekly_self_study_hours + 75.83

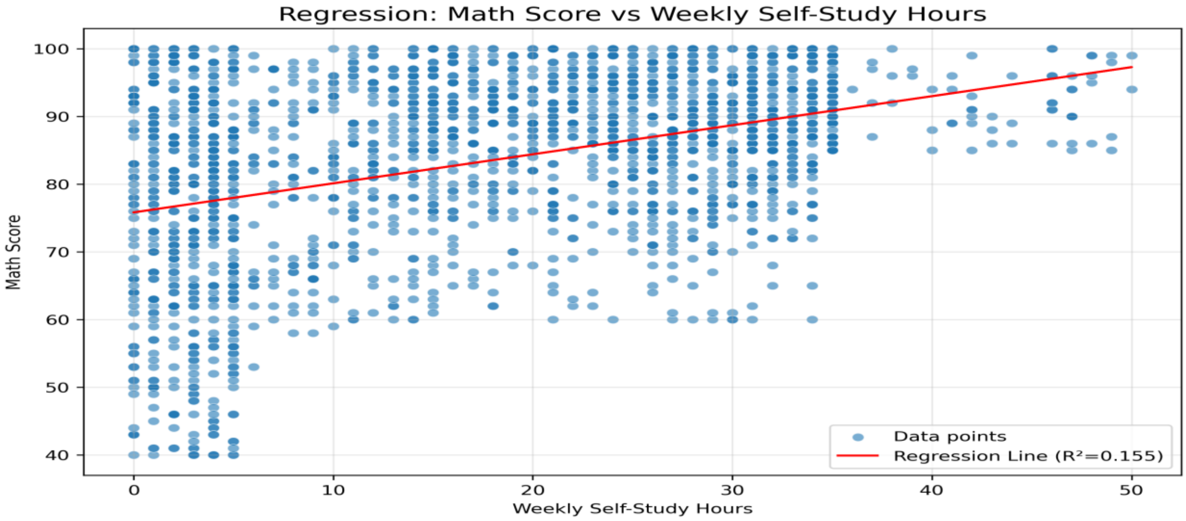
R-squared = 0.155

Prediction for a New Input:

- For a student who studies 30 hours/week:

  Expected Math Score = 0.429 * 30 + 75.83 ~ 88.71

Regression Plot: A scatter plot with the regression line was generated, where $R^2 = 0.155$



Task 5 – Hypothesis Testing

Test Chosen: Independent Two-Sample T-Test **Groups:** Male vs Female students **Variable:** math_score

Hypotheses

H₀: There is no difference in average math scores between males and females (µ_male = µ_female)

H₁: There is a difference in average math scores between males and females (µ_male ≠ µ_female)

Results

- T-statistic = 2.237
- p-value = 0.0254
- Significance level (α) = 0.05

Decision: p-value < 0.05 → **Reject H_0 Conclusion:** There is statistically significant evidence that male and female students have different average math scores.

Final Insights

1. **Polarized Study Habits:** While the average weekly self-study time is approx 17.76 hours , the **mode is only 3 hours** , and the standard deviation is high (approx 12.13). This suggests that a large group of students studies for very short periods, while a smaller group studies much longer, resulting in polarized study habits across the population.
2. **Study Time is a Weak Predictor:** The Simple Linear Regression showed that while **weekly self-study hours** has a positive relationship with math_score (slope approx 0.429), it is a **weak predictor**. The R-squared value (approx 0.155) means that study time explains only about 15.5% of the variation in math scores, indicating that other variables are more influential.
3. **Significant Gender Disparity in Math:** The T-Test concluded that there is a **statistically significant difference** in math scores between genders (P-value 0.02539 < 0.05). On average, male students scored higher (approx 84.11) than female students (approx 82.79).
4. **Low Inter-Subject Correlation:** The relationship between math_score and physics_score is a **very weak positive correlation** (approx 0.116). This implies that performance in one subject is not a reliable indicator of performance in the other.

Recommendations

Based on these findings, the following recommendations are suggested for improving student performance and future analysis:

- **Target Low Study Hours:** Focus intervention efforts on the large group of students who study for only 3 hours per week (the mode). Investigating the reasons for this low engagement could reveal barriers (e.g., part-time jobs, competing responsibilities) that the institution could address.
- **Investigate Gender-Specific Factors:** Given the significant difference in math_score based on gender, the institution should conduct a deeper analysis to understand the

underlying causes, such as curriculum bias, teaching methodologies, or differences in student confidence, and implement targeted programs to promote equity in scores.

- **Build a More Robust Predictive Model:** Since weekly_self_study_hours is a weak lone predictor of score R^2 approx 0.155), future analysis should incorporate other variables (e.g., absence_days, extracurricular_activities) into a **Multiple Linear Regression** model to create a more accurate and comprehensive forecasting tool.

Learning Summary – Week 5

This week I mastered the complete analytics workflow:

- Descriptive statistics for data summarization
- Correlation & regression for relationship and prediction modeling
- Hypothesis testing with p-values for evidence-based decision making