1. Dataset and Research Question

• What dataset are you working with? (Include brief description of dataset, variables, and sample size).

Student Performance Dataset. The dataset shows how well different genders/ethnicities scored on math, reading, and writing. The dataset variables include gender, race/ethnicity, parental level of education, lunch, and test preparation course. The sample size is n = 1000.

• Why did you choose this dataset? What real-world context or problem does it represent?

We chose this dataset because we wanted to research how social status can affect a student's academics. The dataset represents students' academic success based on their background and the financial well-being of their families.

• Formulate a clear research question that can be answered with the dataset. Why is this question relevant/important in a business/economic context? You can tie the research question to the regression you will run in your analysis.

How does a student's background, financial well-being, and parental education affect their academic performance? This question carries meaningful implications for modern hiring practices and workplace equity.

This is because many companies today still rely on academic performance and school prestige as indicators of talent. These outcomes are shaped by socioeconomic factors. Therefore, relying on them can reinforce bias. That is why understanding these influences helps businesses create inclusive hiring practices that focus on true ability rather than inherited advantages.

2. Data Classification and Description

• Choose 1-2 dependent variables (DV) and 2-3 independent variables (IV) from the dataset for your analysis. Go for at least one more IV than DV.

Dependent Variables: Math Scores, Reading Scores, Writing Scores,
Independent Variables: Parental Education, Lunch, Gender, Absences

• Classify your variables: Which are categorical (nominal, ordinal) and which are numerical (interval, ratio)?

Categorical Data: Parental Education (Ordinal), Lunch, Gender
Numerical Data: Math Scores, Reading Scores, Writing Scores, Absences

• Provide descriptive statistics for your main quantitative variables by using the Descriptive Statistics function in Excel. Also, generate the 5-point summary. Briefly provide the interpretation of these measures – these can include explaining if the data is skewed and why, what the average represents, how spread out the data is, etc.

- The math scores have a mean of 66.09 and a median of 66. Therefore, indicating that the average student scored in the mid-60s and that the distribution is fairly centered around that value.
- The five point summary for math scores is as follows:
- Minimum = 0, Q1 = 57, Median = 66, Q3 = 77, Maximum = 100
- For reading the five point summary scores is as follows:
- Minimum = 17, Q1 = 59, Median = 70, Q3 = 79, Maximum = 100
-

• Include one appropriate graph or frequency/relative frequency table for your DVs and IVs. For quantitative variables, choose a boxplot if going for a graph. For qualitative variables, you can opt for bar chart or pie chart as graph options.
• Bonus points: Identify any outliers and give a brief explanation of their meaning in the context of your data.
- Outlier Interpretation: The outliers in the math score data occur only on the lower end of the distribution. These low scores may represent students who did not complete the exam or struggle with the material. Moreover, these increase variability but reflect real differences in student circumstances.

3. Point Estimates and Confidence Intervals
• Treat your dataset as a sample from a larger population. What would the population for your dataset potentially be?

The population for our dataset would be all high school students in the broader education system who take standardized tests in math and reading. Our 1,000 students are treated as a sample from that larger group.

• Provide a point estimate of the population mean for your DV.

Point estimate:

1. Math Scores = 66.089

2. Reading Scores = 69.169

• Construct one confidence interval for your DV with a 95% confidence interval. Interpret your confidence interval in plain English that can be explained to a business decision-maker.

We are 95% confident that the DV sample mean is between (65.14919983,67.0288)
• Bonus points: Construct one more confidence interval for a proportion in your dataset.

3. Hypothesis Testing (One or Two Samples)

Research question for One T Test and Two T Test:

H0: Is the average overall academic performance of students significantly different from the benchmark proficiency score of 70?
H0: $\mu = 70$
H1: $\mu =/ 70$

In this study, academic performance was measured by combining both math and reading scores. When averaged together, the performance came out to be 67.629. Therefore, we are doing a one sample t-test to check whether the average academic performance is the same as a benchmark score of 70. We set up the null hypothesis as a true average academic performance being 70, while the alternative hypothesis being different from 70. We use this since the population's standard deviation is unknown.

Using Excel, the calculated t-value was -5.284955701. At the 5% significance level, the p-value is less than 0.05 which means we reject the null.

Therefore, since the sample mean is lower than the benchmark, the findings suggest that students are performing below the proficiency level. Suggesting that academic performance may be influenced by background factors such as parental education.

Two T Test

How does a student's background with both financial well-being and parental education affect their academic performance?

To connect this hypothesis, we will be focusing on parental education as one key dimension of any potential advantage.

Group 1 = students whose parents DO have a college degree
Group 2 = students whose parents DO NOT have at least one college degree

Let group 1's population mean be at least one college degree, and group 2's population mean be with no college degree. It will be the mean academic performance of both math and reading, which is roughly 67.629. This is the DV we will use for our one-sample hypothesis test.

Null Hypothesis (H0): μ - population mean of the overall academic performance (both math and reading / 2)

- Parental education does not affect academic performance
- H0 : μcollege = μno college
  Alternative Hypothesis (H1):
- Parental education does affect academic performance.
- H1: μcollege =/ μno college

5. Regression Analysis
• Select two quantitative variables – 1 IV and 1 DV. Write the regression equation youv will test: $y = \beta 0 + \beta 1 x + \epsilon$, by substituting your x and y values.

Writing Score, Absences

• Fit a simple linear regression model using Excel. You can use any test from what we covered in class. Report all the results: coefficient of determination, regression equation, now with the intercept and slope coefficients; significance test (p-value) of the slope.

Coefficient of Determination: 0.000183091
Regression Equation: y=67.70232265+0.024228546*x+ ϵ,
Intercept: 67.70232265
Slope: 0.024228546
P-Value: 0.669105452

• Interpret the slope and intercept in business terms. What do they mean in the context of your research question?

The intercept is the value of y (writing Score) when x (absences) is 0. This represents the baseline performance. It predicts that a student who has zero absences would achieve an average writing score of approximately 67.7. This gives educators or employers a benchmark for understanding what a "fully present" student achieves on average. The slope, however, represents the change in y for every 1-unit increase in x. For every, one additional absence, a student's writing score is predicted to increase by 0.024 points. In the context of the research question, this number is so close to zero that it indicates attendance is not the driver of performance for this sample. From a business perspective, looking at the slope would conclude that tracking attendance is irrelevant for predicting who has strong writing skills, and they should look for other indicators instead.

6. Statistical Thinking & Reflection

• Overall, were your results surprising, expected, or inconclusive? Why?

The results were surprising. We expected absences to impact writing scores negatively, but the analysis showed almost no relationship and a slightly positive slope. This contradicts the conventional wisdom that poor attendance automatically leads to lower academic performance.

• How does your analysis help answer the research question?

It refines our understanding by showing that not all variables predict success equally. The earlier tests confirmed that background factors like parental education do matter, and the

regression proves that behavioral metrics like attendance are not always strong indicators of performance in isolation.

• What are the limitations of your data, methods, or findings?

The standardized test scores do not measure soft skills or total intelligence. The regression ignores the confounding factors like a student can be absent frequently but still scores well due to private tutoring. The absences data does not distinguish between excused and unexcused absences.

• If you had access to more/better data, how would you extend the analysis?

We would replace the “Lunch” variable with actual household income data for accuracy.

• How could a business decision-maker use your findings?

Decision makers such as HR directors can use these findings to support skills-based hiring, which improves the earning potential of a company. Since attendance and GPA can be biased based on socioeconomic background or irrelevant to actual performance. Firms should focus on assessing a candidate’s actual skills rather than their background or traditional academic path.