MEMORANDUM

To: Yiwen Chiu

From: Cameron An and Lucas Kantorowski

Date: May 1, 2025

Re: Sustainable Literacy Statistical Analysis Recommendation

The purpose of this memo is to describe the statistical methods and findings from an analysis of your sustainable literacy data. I hope that this information helps you address your questions:

> *"Do environmental factors have an impact on one's sustainability knowledge score?"*
>
> *and,*
>
> *"Do educational factors have an impact on one's sustainability knowledge score?"*

This memo is organized into four sections.

- The first section, "**Background and Data**," includes a description of our understanding of your data and of your main statistical questions, along with a broad discussion about the design of the survey and the generalizability of the results. (page 2)

- The second section, "**Statistical Methods**," describes an analysis approach for your consideration. (pages 2-3)

- The third section, "**Results**," describes the results from an analysis for your consideration. (pages 3-8)

- The last section, "**Summary of Key Findings**," summarizes and discusses the key findings based on the statistical analyses. (page 8-9)

If you have any additional questions about this work following our consulting meeting today, feel free to contact us at **czan@calpoly.edu** or **lkantoro@calpoly.edu.**

I: Background and Data

From the initial consulting meeting, it is our understanding that you wanted to determine the effects of environmental and educational background on knowledge of sustainability literacy to improve curriculum in general education classes. It is also our understanding that a previous study was conducted with a similar aim, but due to small sample size and qualitative variables, the analytical power of the study was weak. This new survey was conducted using new questions with the goal of better analytical power. Some of the variables you have provided us include sustainability literacy knowledge score, year of college, and census data based on the home zip code of the student, including fire frequency and PM 2.5 concentration.

At our initial consulting meeting, you emphasized sustainability knowledge score (SKS) to be the **most important** variable of interest, as you wanted to remap the curriculum of general education classes to more effectively increase knowledge. After careful consideration, we decided to **focus on SKS** for our analysis.

For your **sampling methods**, it is our understanding that data collection was a mix of random sampling and voluntary methods. Large classes with students from a variety of majors were randomly selected to participate in an effort to obtain a representative sample of the population of students. Social media posts, e-mails, and fliers were made in an effort to have additional students volunteer to take the survey. Due to use of voluntary methods, our **ability to generalize** to all Cal Poly students is reduced.

II. Statistical Methods

We propose that you consider exploring your research question using **Multiple Linear Regression**, which is a statistical model that explores the linear relationship between a **response variable** and certain **explanatory variables**. This model is particularly useful for seeing how variables are related to each other and the extent to which they impact other variables. Here is the general form of a multiple linear regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

Note: The equation above only utilizes two explanatory variables (x_1 and x_2) but there could be more. There is also another term in this equation called the **interaction** term ($x_1 x_2$). The interaction term explains how one of the explanatory variables impacts the other explanatory variable's impact on the response variable. For example, consider the case where x_1 = fire frequency and x_2 = pm 2.5 concentration. The interaction term explains how different levels of fire frequency affect the pm 2.5 concentration impact on the response variable. For our analysis, **we did NOT consider interaction effects** due to the potential of overfitting and to simplify the nature of our model. However, we included this information to provide full transparency and for your consideration for future analyses.

For our data, we used SKS as the response variable as we wanted to explore the impact of certain explanatory variables on one's knowledge about sustainability. We conducted **two separate analyses** using multiple linear regression.

Our first analysis investigated the environmental impacts on SKS. The variables we considered as environmental factors were:

- Fire frequency
- Percent greenspace by area

- PM 2.5 concentration
- Poverty
- Zip code area

We decided to consider these specific variables in our analysis because we reasoned that each contributed to or shaped one's environment growing up. Using R, we selected the variables that were deemed "significant" and included those variables in the final environmental model equation. We considered significant variables as those with p-values of **less than 0.05**. Finally, for the selected variables, we assigned each of them a coefficient based on the R output. These coefficients explain the extent to which the selected variables affect the response variable. For example, in the case where we considered x_1 = fire frequency and x_2 = pm 2.5 concentration, a β_2 value of 0.5 would indicate that for each 1 $\mu g/m^3$ increase of pm 2.5 concentration, the SKS increases by 0.5 after adjusting for all other explanatory variables in the model.

Our second analysis investigated the educational impacts on SKS. The variables we considered as educational factors were:

- Linguistic isolation
- Some college (% of pop)
- College degree (associates +, % of pop)

Similarly to our first analysis, we selected the variables that were deemed "significant" in R and included those variables in the final educational model.

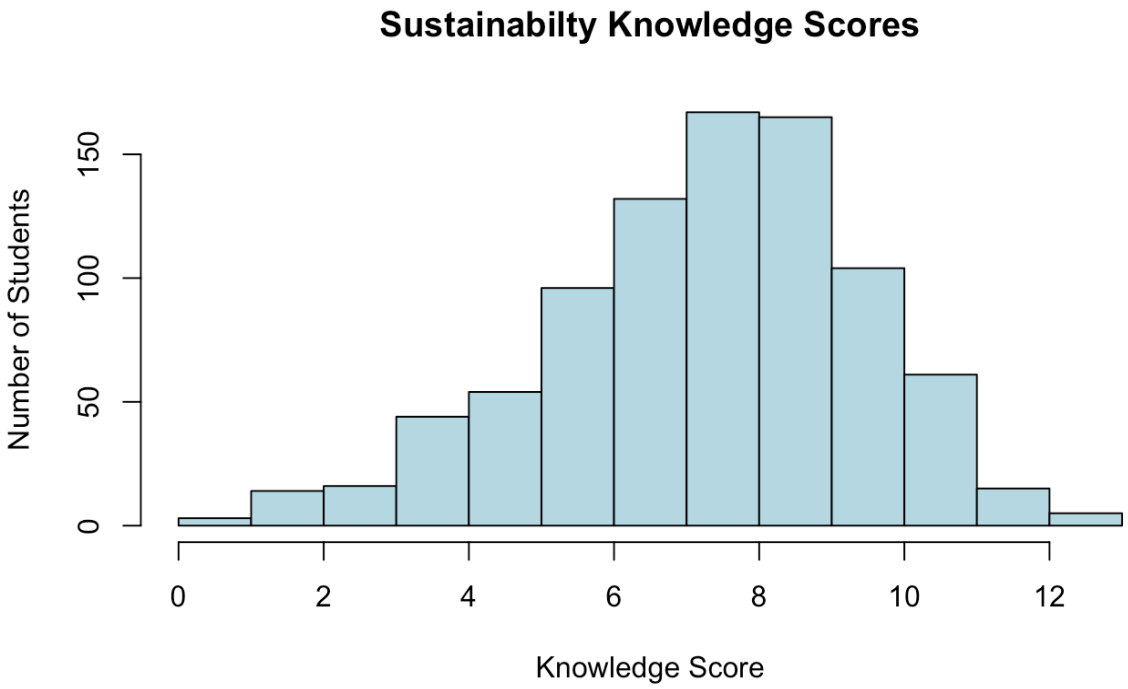
In both models, we realized it was important to adjust for other variables before concluding that certain variables had statistically significant impacts on SKS. This is why we decided to add the year of college to both the models. To intuitively understand the impact time spent at college had on SKS, we decided to change year to a quantitative variable, and treated all people who answered '5+' as fifth year students.

III. Results

We have included our findings from both our analyses below. The first analysis focuses on environmental factors and their effects on knowledge score. The second analysis focuses on educational factors and their effects on knowledge score.

Figure 1 below displays a histogram displaying the overall SKS values for all survey respondents. We included this graph to help with comparisons when using our multiple linear regression equations.

Figure 1: Histogram of Sustainability Knowledge Scores



Mean: 7.764 SD: 2.195

Analysis #1: (Environment Impact on SKS)

The first analysis uses fire frequency, percent greenspace by area, PM 2.5 concentration, poverty, zip code area, and year to indicate how these environmental factors affect students' sustainability knowledge score. *Figure 2* below shows that the intercept value, year, PM 2.5 concentration, and zip code area are significant predictors of SKS as indicated by the small p-values (below 0.05).

Figure 2: Table of Linear Regression Model of Environmental Factors on SKS

Variable	Estimate	p-value	Significance
Intercept	7.920	< 0.001	***
Year	0.2678	< 0.001	***
Fire Frequency	0.0208	0.17	
Percent Greenspace by Area	0.0019	0.71	
PM 2.5 Concentration	-0.0903	0.04	*
Poverty	-0.0065	0.37	
Zip Code Area	-0.0020	0.04	*

Therefore, a model we could consider for the environmental impact is:

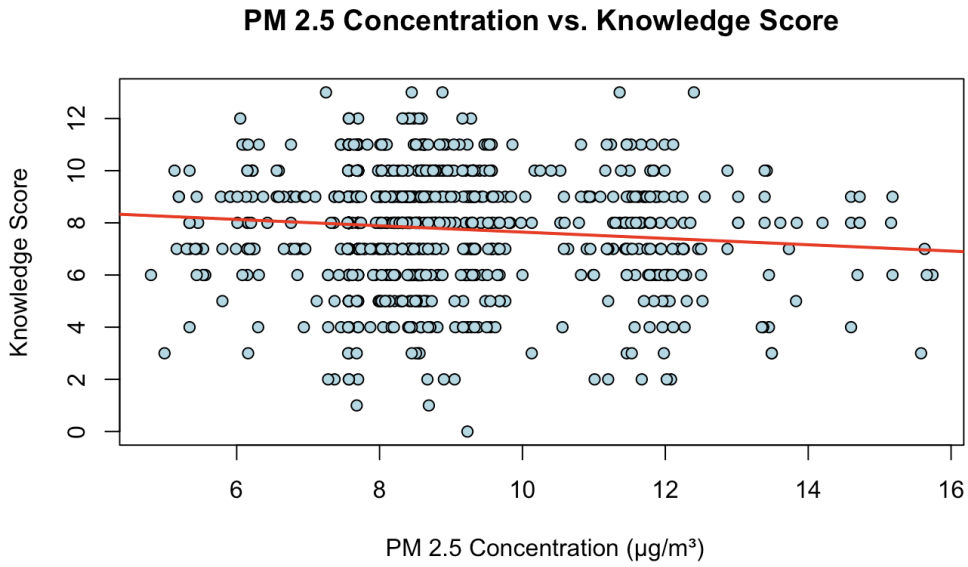
$$\widehat{SKS} = 7.920 + 0.2678x_1 - 0.0903x_2 - 0.0020x_3$$

where x_1= year, x_2= PM 2.5 concentration, and x_3= zip code area.

The year variable was considered to be a significant predictor of SKS and also yielded a slope estimate value of 0.2678. We can interpret the slope estimate value by stating that for each additional year a student stays in college, their expected SKS increases by 0.2678 after adjusting for all other predictors in the model.

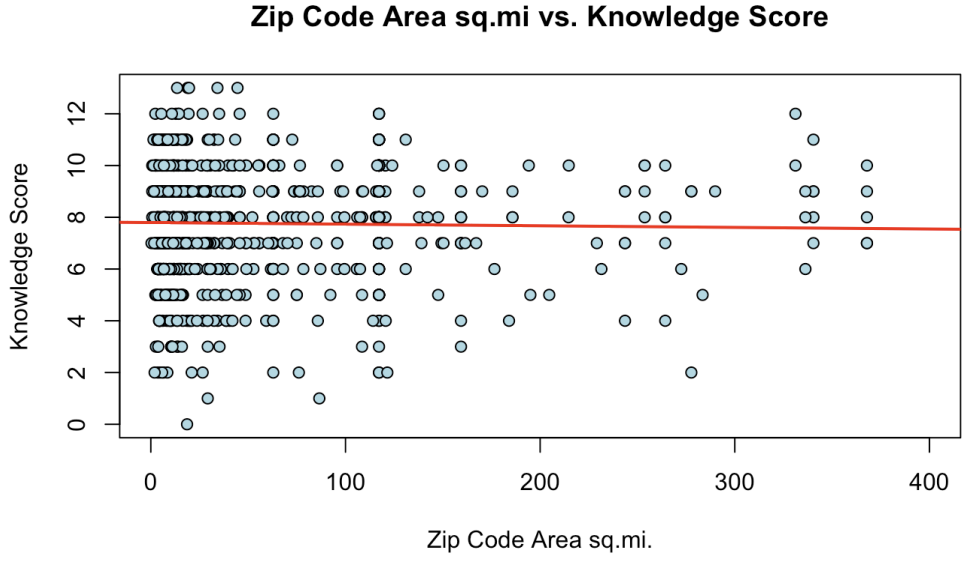
Similarly, the PM 2.5 concentration variable was considered significant and yielded a negative slope estimate of -0.0903. We can interpret this by stating for every 1 $\mu g/m^3$ increase in PM 2.5 concentration, the expected SKS decreases by 0.0903 after adjusting for all other predictors in the model. *Figure 3* below displays a scatterplot of PM 2.5 concentration vs. knowledge score. The **least squares line** (shown in red) shows a **slight negative correlation** between PM 2.5 concentration and SKS, which aligns with our findings in the equation above.

Figure 3: Knowledge Score vs PM 2.5 Concentration



Lastly, we can apply the same method to the zip code area variable. Since the slope estimate value is equal to -0.0020, we can interpret this by stating that for every 1 square mile increase in a student's zip code area, their expected SKS decreases by -0.0020. *Figure 4* below displays a scatterplot of zip code area vs. knowledge score. The least squares line (shown in red) shows a **slight negative correlation** between zip code area and SKS, which aligns with our findings above.

Figure 4: Knowledge Score Versus Zip Code Square Mileage With Model Regression Line



Analysis #2: (Education Impact on SKS)

This analysis focuses on educational background factors that contribute to the variation in SKS. These variables were taken from census data based on a student's home zip code. Variables include percent of population who graduated college, percent of population who have some college education, and percent of limited English speaking households. We decided to include a student's year in college as another predictor variable due to the large effect it has on SKS. Even after adjusting for the student's year in college, we found some significant predictors of SKS, shown in *Figure 5*.

Figure 5: Table of Linear Model of Education Background Factors on SKS

Variable	Estimate	p-value	Significance
Intercept	4.5212	< 0.001	***
Year	0.2696	< 0.001	***
Percent of Population With a College Degree from Home Zip Code	2.6566	0.02	*
Percent of Population With Some College Education from Home Zip Code	4.4415	0.16	
Linguistic Isolation	0.0134	0.5396	

Therefore, a model we could consider for the educational impact is:

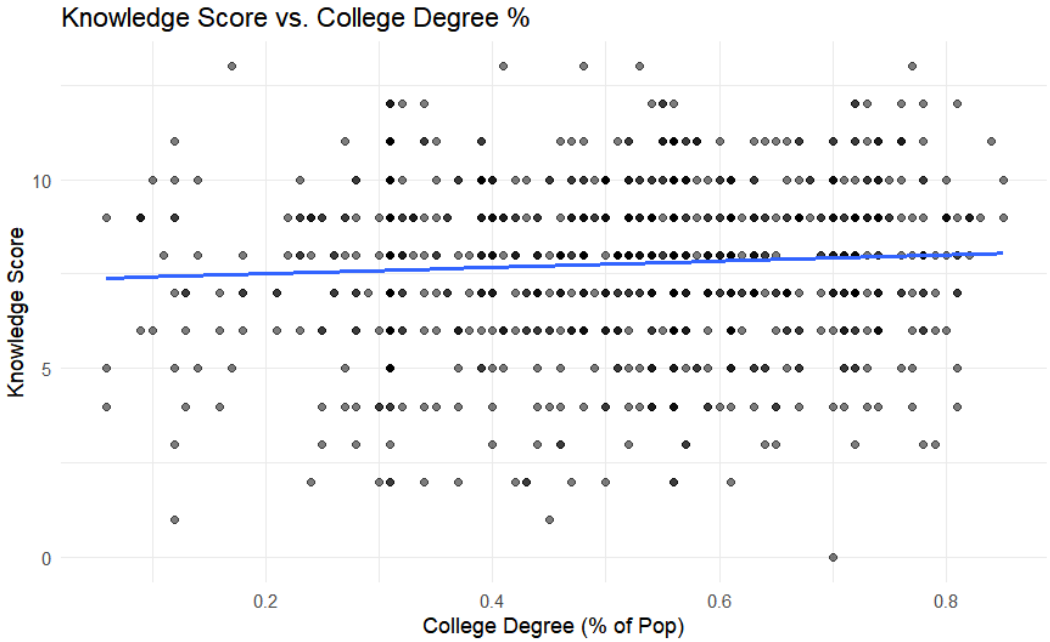
$$\widehat{SKS} = 4.5212 + 0.2696x_1 + 2.6566x_2$$

where x_1= year, x_2= Percent of Population with a college degree (within student's zip code).

For this analysis, the year variable was considered to be a significant predictor of SKS and also yielded a slope estimate value of 0.2696. We can interpret the slope estimate value by stating that for each additional year a student stays in college, their expected SKS increases by 0.2696 after adjusting for all other predictors in the model.

Similarly, the percent of population with a college degree in the respondent's zip code was also a significant predictor variable with a slope estimate of 2.6566. We can interpret this value by stating that for every 1 percentage point increase in college degree completion within a student's zip code, the expected SKS increases by 2.6566. *Figure 6* below displays a scatterplot of zip code area vs. knowledge score. The least squares line (shown in blue) shows a **slight positive correlation** between zip code area and SKS, which aligns with our findings above.

Figure 6: Knowledge Score vs. College Degree %



IV. Summary of Key Findings

Below in *Figure 7*, we have provided the key findings from our two analyses on environmental effects on SKS and educational effects on SKS. We have included:

- The p-value for each significant explanatory variable for each analysis.
- An estimate of how much a 1 unit increase of a variable affects the predicted SKS.

Our analysis on environmental impacts shows that a student's year in college, the PM 2.5 concentration in their zip code, and their zip code area (sq. mi.) have a significant impact on their SKS. Due to its extremely low p-value, the 'year' variable was deemed as the most significant predictor of SKS, which demonstrates that the longer a student is in college, the higher their SKS tends to be.

Our analysis on educational impacts shows that a student's year in college and the percentage of population with a college degree from their home zip code are significant predictors of SKS. Since the slope estimate of the percentage of population with a college degree from a student's home zip code is positive and the variable is significant, we can claim that students with a higher college degree percentage in their zip code tend to have higher SKS values.

Given that the data comes from both random sampling and voluntary responses, we cannot generalize our findings to the entire student body at Cal Poly. Voluntary responses may introduce the risk of self-selection bias, where individuals who choose to respond may differ systematically from those who do not. That being said, any conclusions drawn should be interpreted with caution, especially when making inferences about groups that may be underrepresented or missing entirely from the sample.

Figure 7: Summary of Key Results

Analysis	**p-value**	**Slope estimate**
Analysis #1 (Environment)	Year: 0.001 PM 2.5 conc: 0.04 Zip Code Area: 0.04	0.2678 -0.0903 -0.0020
Analysis #2 (Education)	Year: .001 Percent of Population With a College Degree from Home Zip Code: .02	0.2696 4.4415