Steven Alnemri
Dr. Susanna Lange
DATA 11800: Introduction to Data Science I
12 February 2024

Midterm Report

Part A: Introducing the Dataset

The study provides web-based categorical survey data on American 12th grade students' values, behaviors, and lifestyle orientations. This is a longitudinal prospective trend study, meaning survey outcomes track changes in behavior each new year starting from the beginning of the study, but samples change each year. There are 1,400 unique variables in the full data set spread across 6 different versions of the survey, each sharing a common set of "core" questions. From the description citation and my code notebook, 9, 599 students in total participated (Form 1: 1,536, Form 2: 1,587, Form 3: 1,593, Form 4: 1,633, Form 5: 1,586, and Form 6: 1,664). Many of the characteristics of students and their households (e.g father education, race, sex) were collected using the core data consisting of 9,599 cases. Multistage sampling (three stages) was implemented, randomizing by geographic area, school, then students within the school. Primary sampling based on geography accounted for household count, ensuring higher representation from highly populated areas to reduce selection bias. However, there may be response bias due to questions about illegal behavior, causing students to lie, and non-response bias due to schools and students declining to participate; schools or students who choose not to participate may be different than those who do, making the sample not truly representative of the American 12th grade population.

Part B: Characteristics of sample

Surveyed students were asked to best describe their average grades in the current year with a grade value (A, A-, etc.). The results are graphed in **Figure 1**. The number of students increases as we move up grade categories. Students who chose not to answer were also included to account for non-response bias as students with poor grades might prefer not to share. The amount of missing values is significant when compared to low-acheiving students (C, C-, and D), indicating the possibility for more students earning lower grades.

Another characteristic surveyed was father education level (**Figure 2**). The common milestone accomplishments "Completed high school" and "Completed college" are the most represented categories. Most fathers have completed at least high school with relatively few having completed less. The "Don't Know, N/A" and "Not answered" categories are included to account for students who may not have a father figure to report data on. Non-response bias may cause students also omit answers when father education level is low.

Students were also asked to report the number of school days skipped (**Figure 3**) in the last four weeks. Since the range of days missed for each category (e.g "4-5 Days," "6-10 Days," etc.), a table was better suited to represent the data; it might be misleading to have inconsistent intervals. The 1,217 students who chose not to answer were included to account for non-response bias since those who skip more days may not want to report their behavior. A relatively large amount of students did not cut class, and the number of students decreased as we move to catefgories of more days skipped.

Figure 1

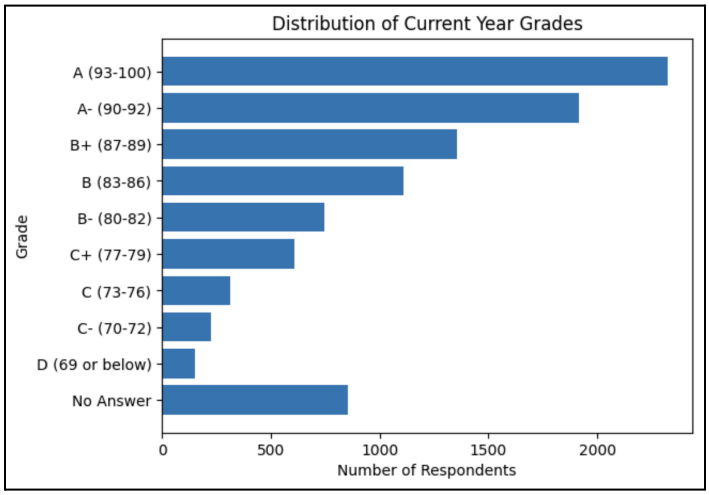
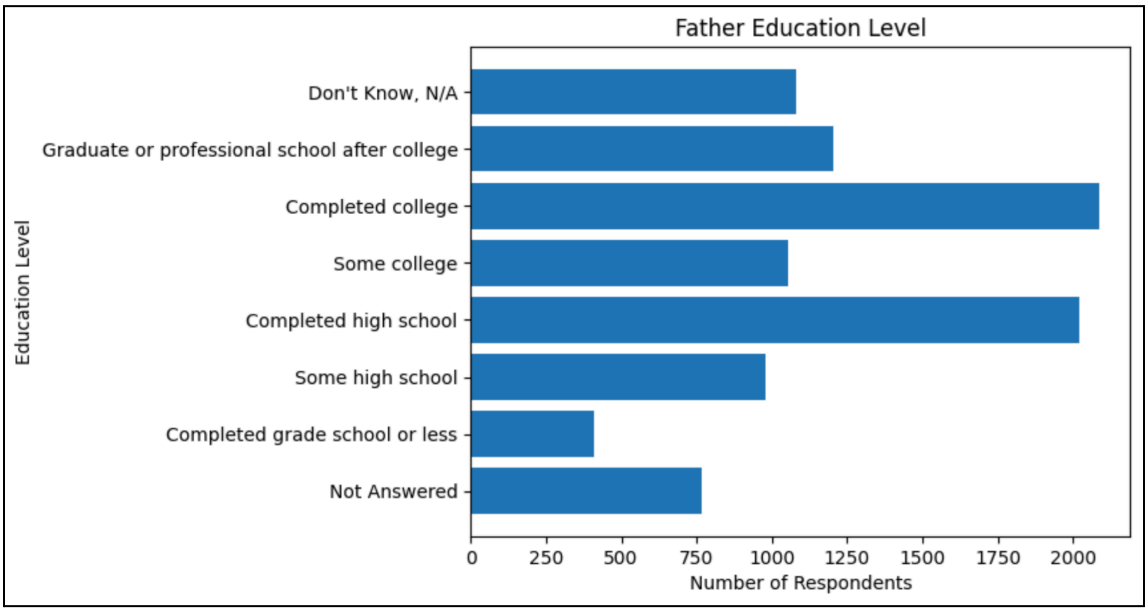


Figure 3

	School Days Skipped (Last 4 Weeks)	count
0	Not Answered	1217
1	None	6051
2	1 Day	947
3	2 days	521
4	3 Days	353
5	4-5 Days	249
6	6-10 Days	117
7	11 or More	144

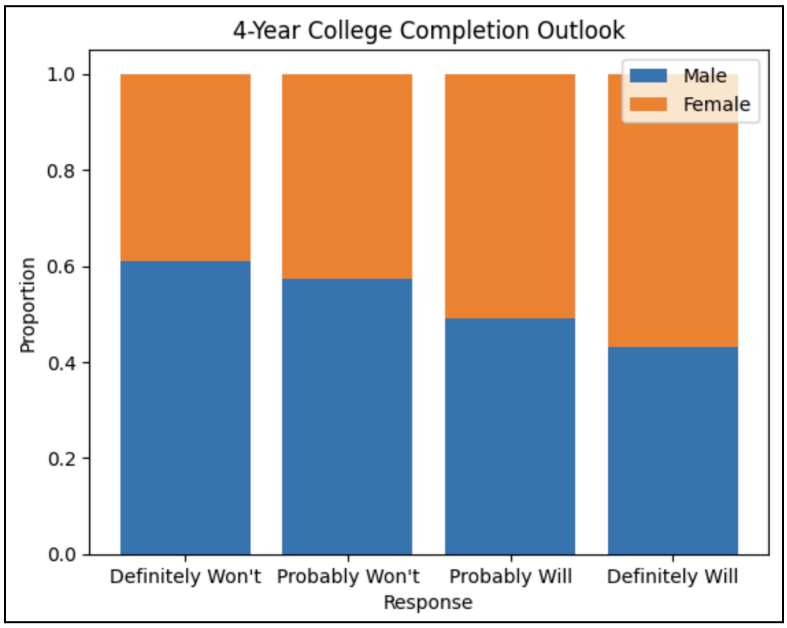
Figure 2



Parts C: Relationships Between Variables

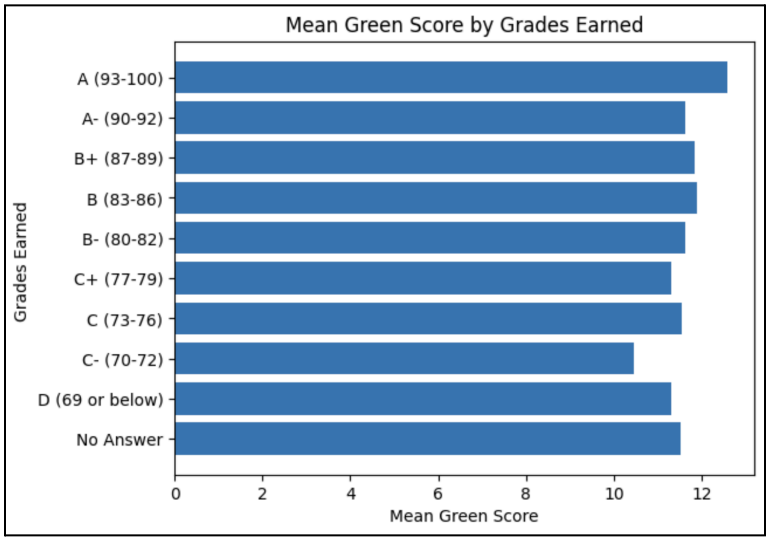
One survey question asked students if they expect to complete a four-year college degree (**Figure 4**). Students not identifying as male or female were not included because of small representation. The proportion of female students increased as we move to categories of higher certainty of college graduation. In general, female students have a greater outlook for college completion than male students. More behavioral problems might explain differences in college outlook (Goldin 2006).

Figure 4



Additionally, there were several questions related to environmental consciousness such as transit usage and diet. Using these responses, I created a new variable named “Green Score” that is a measure of how much a student cares about the environment overall, and compared mean scores across academic performance (**Figure 5**). Green scores range from 0 to 20 and were only calculated for those who responded to all environmental questions. There seems to be no significant correlation between grades and mean Green score.

Figure 5



Part D: Provide Context

The higher college outlook may imply a shift in attitude towards women's participation in the workforce. More women are seeking professional careers than before, and have even surpassed men in college participation. Shifting attitudes may have equalized participation between genders, but behavioral issues in young males may cause women to surpass them in college outlook (Goldin 2006). A potential collider is behavior as being male and having low college aspirations can both contribute to behavioral issues.

We might expect higher achieving students to be more knowledgable about coursework related to the environment and adjust their behavior benefit the environment. However, since Green Scores don't vary significantly between academic performance categories, this could imply a lack of environmental education in general, or a pattern of viewing environmental education ad purely academic (not something to implement in their lives). High performing students may care about the environment, but may not be willing to make actual changes (Nijhuis). One potential confounder is lack of environmental education as it can impact Green Score and grades of the students.

I'd be interested to see how Green score and college outlook can vary across neighborhoods with different median incomes. In the current study, there is nothing accounting for school funding or parents' income that might show differences in Green Score, college outlook, or other behaviors.

Part E: Conclusion

From this study I have learned that environmental consiousness does not vary across academic performance, but college outlook does change with gender. The former may be a result of a lack of environmental education/emphasis in schools. Through my analysis and description of the study, I've noticed its important to recognize biases when reporting stigmatized behavior as many students refused an answer for school days skipped. Additionally, it's important implement strategic sampling to represent a population accurately; for example, complete randomization may have caused over-representation of locations with low population.

Bibliography

Claudia Goldin & Lawrence F. Katz & Ilyana Kuziemko, 2006. "The Homecoming of American College Women: The Reversal of the College Gender Gap," Journal of Economic Perspectives, American Economic Association, vol. 20(4), pages 133-156, Fall.

Kloska, D. D., Miech, R. A., Johnston, L. D., Bachman, J. G., O'Malley, P. M., Schulenberg, J. E., and Patrick, M. E. (2023). Codebook for Monitoring the Future: 12th grade surveys, 2022. Ann Arbor: Institute for Social Research. The University of Michigan. Available from https://www.icpsr.umich.edu/web/NAHDAP/series/35.

OpenAI. "How to create stacked bar graph" ChatGPT, [16, February 2024].

Nijhuis, Michelle.
https://e360.yale.edu/features/green_failure_whats_wrong_with_environmental_education

Bibliography