# BS 100A Final Project

Li Zhang 206305918

Researchers are concerned that in rural communities poor health literacy may be associated with poorer health outcomes. With this in mind, mailing lists for rural communities in the Los Angeles are obtained and 1,500 individuals are randomly selected to participate. Miraculously, all 1,500 individuals respond to the short survey! The following variables are included:

- **hlth_lit:** Health literacy (primary predictor of interest)– 66-item word recognition test. Scores and grade equivalents:
  - **Score of 0 to 18.99:** 3rd grade and below, will not be able to read most low-literacy materials
  - **Score of 19 to 44.99:** 4th to 6th grade, will need low-literacy materials
  - **Score of 45 to 60.99:** 7th to 8th grade, will struggle with most patient education materials
  - **Score of 61 to 66:** High school, able to read most materials
- **sex**: Sex – Male = 0, Female = 1
- **pol**: Living above the poverty line – No = 0, Yes = 1
- **daily_fol**: Daily Total folate intake – The recommended daily amount of folate for adults is 400 micrograms (mcg)
- **ins**: Insurance status – public insurance = 0, private insurance = 1, uninsured = 2
- **educ**: Highest level of education
  - Elementary school education = 0
  - High school graduate = 1
  - Some college = 2
  - College degree = 3
  - Graduate degree = 4
- **alc**: Alcohol use (dependent health outcome) – In the past 12 months, how many days did you drink any type of alcoholic beverage? (range: 0 – 365)
- **bmi**: BMI (dependent health outcome) – body mass index calculated by weight(kg)/height(m)$^2$. Weight status categories:
  - Underweight: <18.5
    Normal weight: 18.5–24.9
    Overweight: 25–29.9
    Obesity: ≥ 30
- **phq9**: PHQ9 (dependent health outcome) –a depression metric composed on 9 items, each ranging from 0 – 3. The score is formed by summing these nine items together (range: 0 – 27)
- **smoke**: Smoking (dependent health outcome) – Number of cigarette packs smoked in past 12 months

I. Statement of Research Question

a. Description of the health outcome

The health outcome is alcohol. It is continuous data indicating 'in the past 12 months, how many days did you drink any type of alcoholic beverage?' (range: 0 – 365)

b. Distribution of the health outcome variable

Figure 1 shows that it is slightly right-skewed. As can be seen in Figure 2, outliers where some individuals reported consuming alcohol on almost 300 days in the past year. This extreme frequency raises concerns about potential data accuracy issues or the presence of individuals with exceptionally high alcohol consumption patterns.

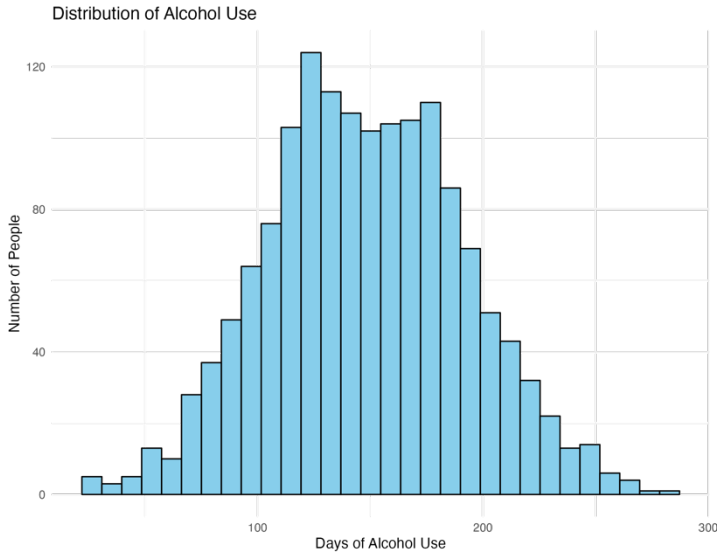


Figure 1: Distribution of Alcohol Use

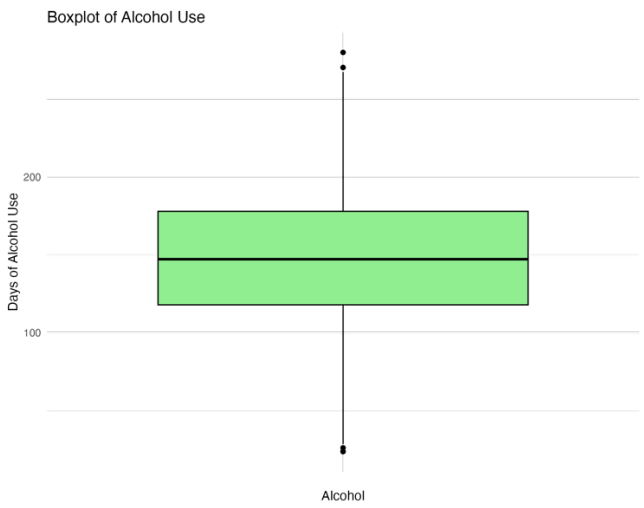

Figure 2: Boxplot of Alcohol Use

c. Expected association

Concerns arise regarding the potential influence of health literacy on individuals' drinking behaviors, as alcohol use is identified as a dependent health outcome. In rural communities, where poor health literacy may be correlated with adverse health outcomes, it is reasonable to anticipate that individuals with lower health literacy might manifest higher rates of alcohol use. Conversely, those with higher health literacy may demonstrate a greater understanding of the health risks linked to alcohol consumption, potentially resulting in more responsible drinking behaviors and, consequently, improved health outcomes.

d. The bivariate plot

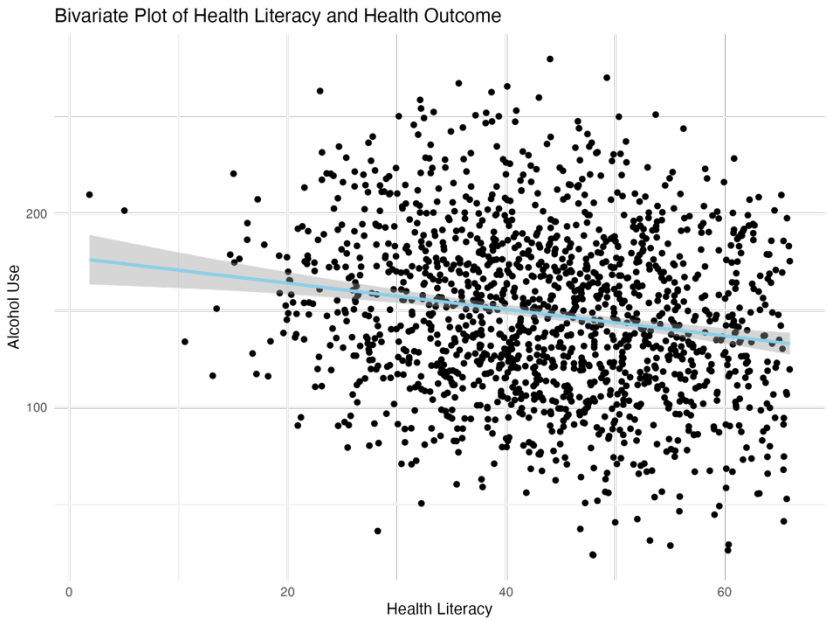

Figure 3: Bivariate Plot of Health Literacy and Alcohol Use

The Pearson correlation coefficient between health literacy (health_lit) and health outcome (alc) in the dataset is approximately -0.183. This negative correlation suggests a weak inverse relationship between health literacy and alcohol consumption. While the correlation is not strong, the negative sign implies that, on average, individuals with higher health literacy tend to have slightly lower levels of alcohol use (Figure 3). It's essential to interpret correlation cautiously, considering that correlation does not imply causation, and other factors may contribute to the observed relationship. Further analysis and contextual understanding are needed to draw meaningful conclusions about the association between health literacy and alcohol use in this dataset.

## II. Description of Data

### a. Description of non-health outcome variable

1. Health_lit: Health literacy is a continuous variable measured using a 66-item word recognition test, and scores(range:0-66) are categorized into four groups. Figure 4 displays a left-skewed distribution, indicating that the majority of participants have higher health literacy levels (40-50). There are some individuals with lower literacy scores, but they are relatively rare. Extreme low values, 5.048691 and 1.851406, are outliers that are of concern (Figure 5). The bivariate plot (Figure 3) suggests a negative correlation between the health literacy and health outcome.

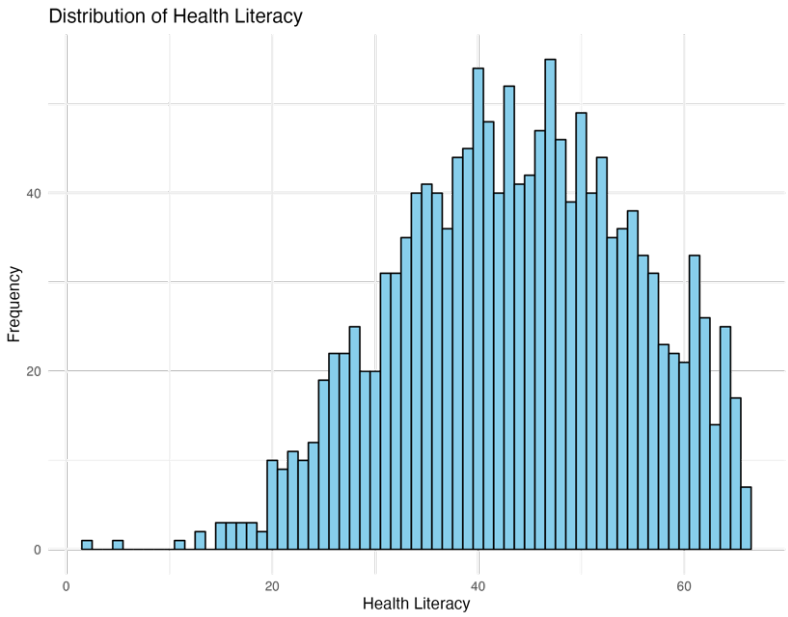

Figure 4: Distribution of Health Literacy

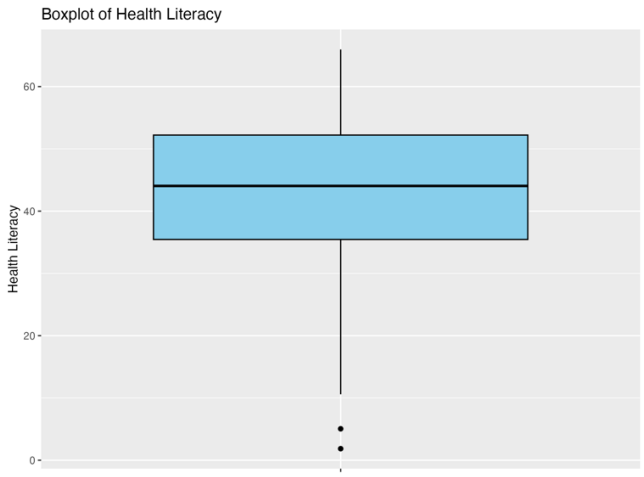


Figure 5: Boxplot of Health Literacy

2. Sex: a categorical variable representing gender, with 0 indicating male and 1 indicating female. The dataset includes 559 male and 941 female (Figure 6).
   Figure 7 illustrates an association between sex and alcohol use, with females generally exhibiting a tendency toward lower alcohol consumption. The dispersion levels between the two genders appear similar, suggesting that the variability in data is not significantly different between males and females.

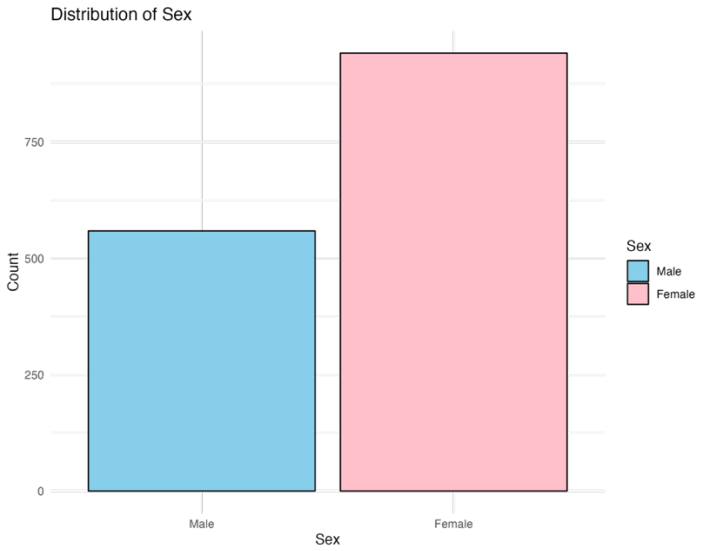


Figure 6: Distribution of Sex

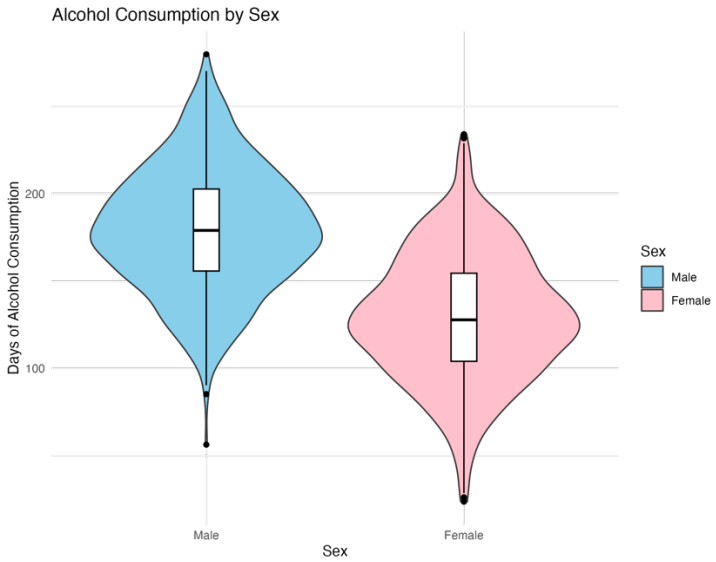


Figure 7: Alcohol Consumption by Sex

3. Pol: Living above the poverty line is a binary categorical variable (0 for No, 1 for Yes) indicating the economic status of participants. The dataset includes 1072 participants living under the poverty line and 428 who above (Figure 8).
   Figure 9 suggests that there is no significant difference between different financial status. It is also not easy to tell the association between the financial status and alcohol use.

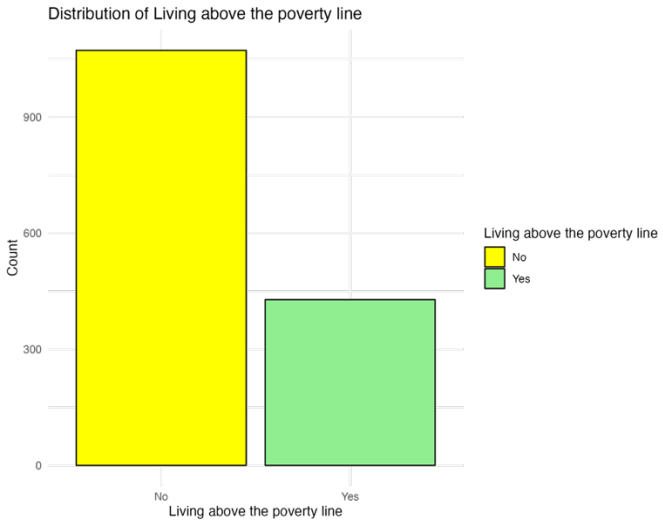


Figure 8: Distribution of Living above the Poverty Line


Figure 9: Alcohol Use by Financial Status

4. Daily_fol: Daily total folate intake is a continuous variable representing the amount of folate consumed daily. Figure 10 shows that outliers are around 800 micrograms (mcg), which is nearly twice the recommended daily amount for adults. Additionally, most participants exceed this amount. The distribution appears right-skewed (Figure 11). As can be seen in Figure 12, it appears that daily total folate intake has a positive correlation with alcohol use.

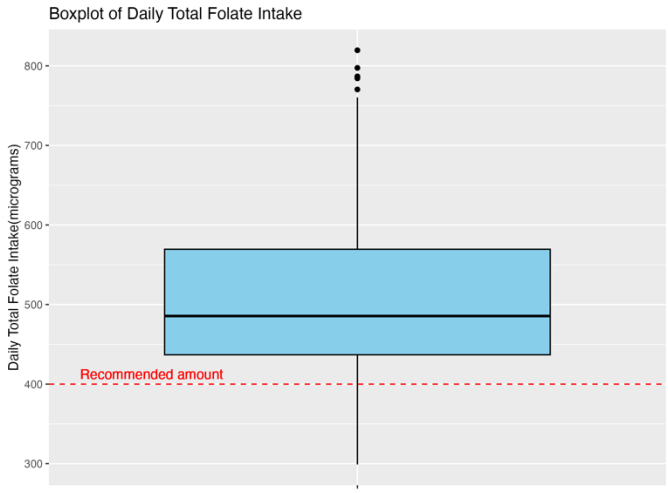


Figure 10: Boxplot of Daily Total Folate Intake

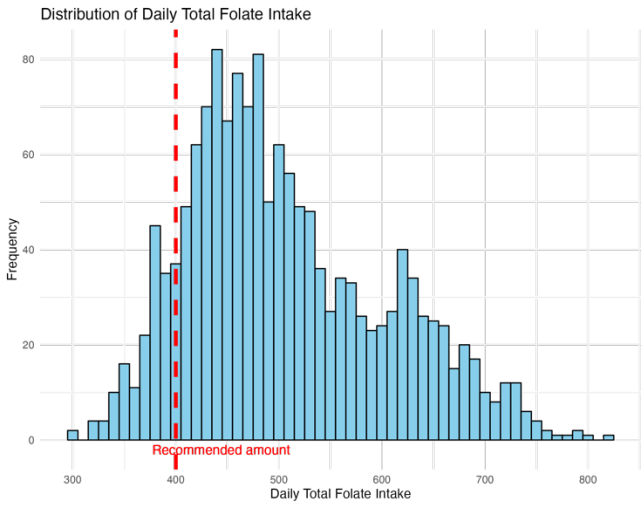


Figure 11: Distribution of Daily Total Folate Intake

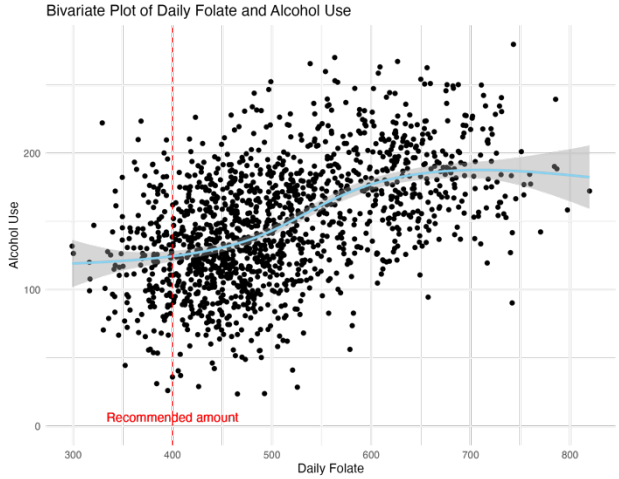


Figure 12: Bivariate of Daily Total Folate Intake and Alcohol Use

5. Ins: Insurance status is a categorical variable (0 for public insurance, 1 for private insurance, 2 for uninsured) indicating participants' insurance coverage (Figure 13). It is surprising that there seems no association between insurance status and alcohol use. Additionally, participants with different insurance status seem to have similar amount of alcohol use (Figure 14).

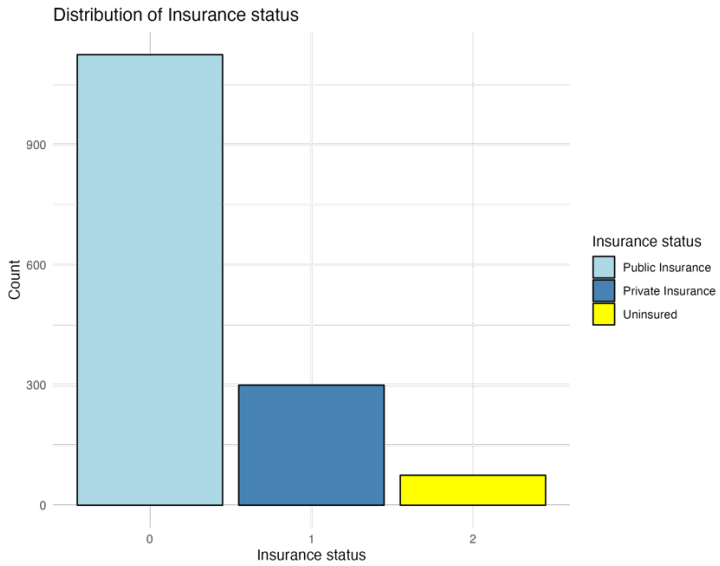

Figure 13: Distribution of Insurance Status

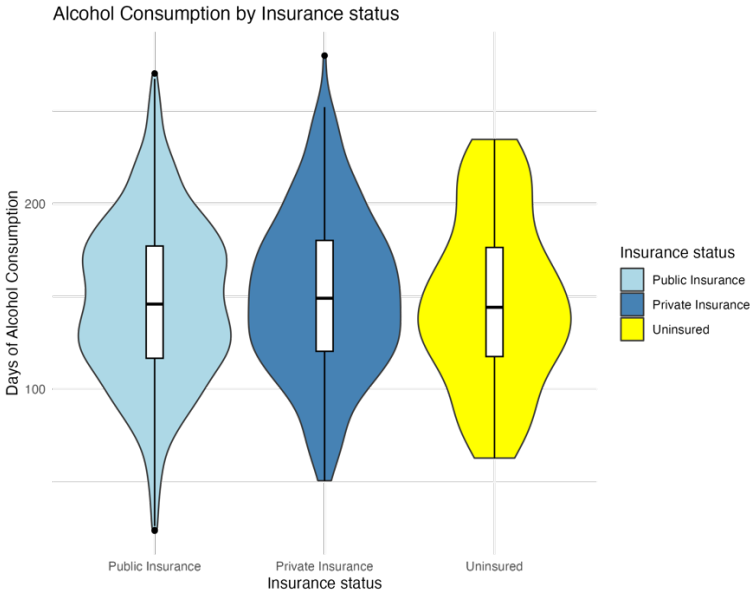

Figure 14: Alcohol Use by Insurance Status

6. Educ: Education is an ordinal variable indicating the highest level of education attained by participants, ranging from elementary school (0) to graduate degree (4). The distribution is shown in Figure 15.
   As shown in Figure 16, there is a lack of significant association between education level and alcohol use. Individuals with graduate degrees seem to exhibit higher alcohol consumption, while others demonstrate similar patterns. However, it is important to note that the sample size of participants with graduate degrees is very small, potentially introducing bias.

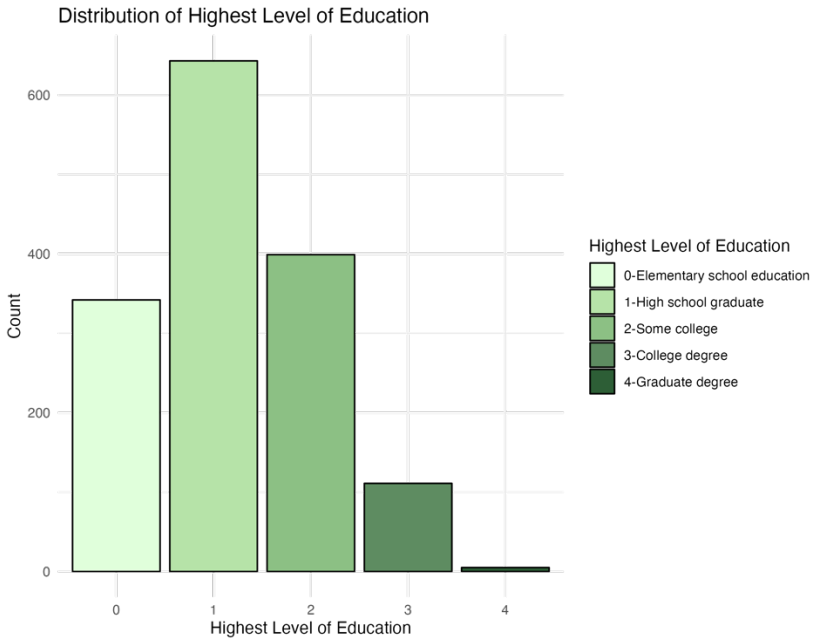


Figure 15: Distribution of Highest Level of Education

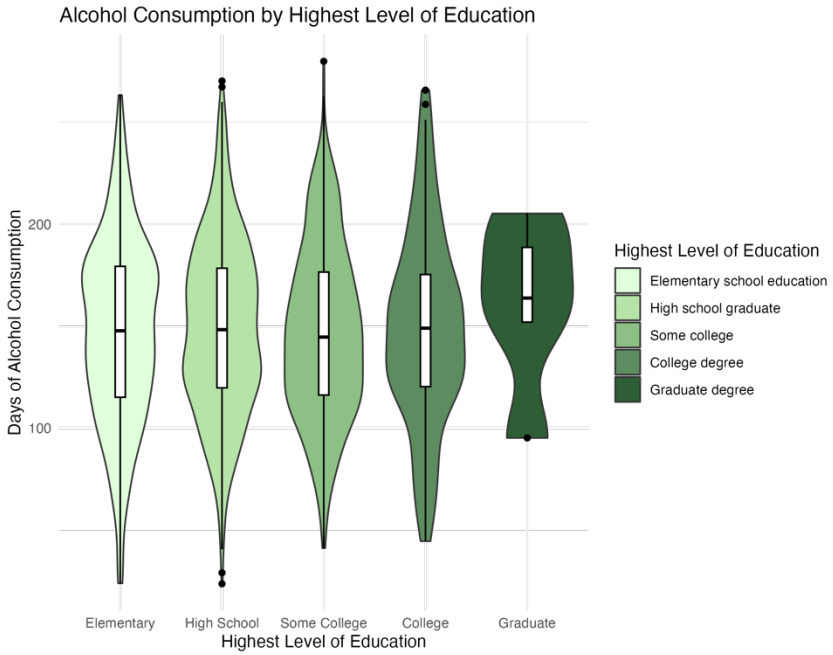


Figure 16: Alcohol Use by Highest Level of Education

## III. Results

### a. Table 1 – Univariate and Bivariate Summaries

As can be seen in table 1, the total sample comprises 1,500 individuals residing in rural communities in Los Angeles, with male participants constituting nearly twice the number of females. The population below the poverty line exceeds twice that of those above it. Mean daily folate intake is 506.79 micrograms, surpassing the recommended 400 micrograms. Regarding insurance coverage, three-fourths of participants have public insurance, while the remaining individuals have private insurance, with very few being uninsured. In terms of education levels, almost half of the individuals are high school graduates, followed by those with some college education. Elementary school graduates are slightly fewer than those with some college degree and only a small fraction possesses a college degree or higher. Over the past year, the average number of days of alcohol use was 148.05, with a standard deviation of 43.59.

Table 1. Characteristics of the Sample in Total and by Health Literacy

| | Total (N= 1500) Count (%) or Mean (SD) | High Health Literacy (N=702) Count (%) or Mean (SD) | Low Health Literacy (N= 798) Count (%) or Mean (SD) | p-value |
|---|---|---|---|---|
| **Characteristics** | | | | |
| Sex | | | | 0.905 |
| Male | 539(37.27%) | 260(37.04%) | 299(37.47%) | |
| Female | 941(62.73%) | 442(62.96%) | 499(62.53%) | |
| Above the Poverty Line | | | | 0.582 |
| No | 1072(71.47%) | 507(72.22%) | 565(70.80%) | |
| Yes | 428(28.53%) | 233(28.53%) | 233(29.20%) | |
| Daily Folate | 506.79(96.13) | 505.60(94.30) | 507.85(97.77) | 0.650 |
| Insurance | | | | 0.873 |
| Public | 1125(75.00%) | 527(75.07%) | 598(74.94%) | |
| Private | 300(20.00%) | 142(20.23%) | 158(19.80%) | |
| Uninsured | 75(5.00%) | 33(4.70%) | 42(5.26%) | |
| Education | | | | 0.151 |
| Elementary school | 342(22.80%) | 143(20.37%) | 199(24.94%) | |
| High School | 643(42.87%) | 310(44.16%) | 333(41.73%) | |
| Some College | 399(26.60%) | 189(26.92%) | 189(26.32%) | |
| College degree | 111(7.40%) | 56(7.98%) | 56(6.89%) | |
| Graduate Degree | 5(0.33%) | 4(0.57%) | 4(0.13%) | |
| Health Outcome | 148.05(43.59) | 141.18(43.86) | 154.10(42.46) | *** |

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

1) Sex: Chi-squared test
   Assumptions: Each subject only contributes to one cell. Observations are Independent. All expected values are at least 5.
   Conclusion: There is not enough evidence to suggest a significant difference in gender between the high health literacy and low health literacy groups at alpha=0.05. (P=0.905>0.05. We cannot reject the null hypothesis.)

2) Above the Poverty Line: Chi-squared test
   Assumptions: Each subject only contributes to one cell. Observations are Independent. All expected values are at least 5.
   Conclusion: There is not enough evidence to suggest a significant difference in poverty status between the high health literacy and low health literacy groups at alpha=0.05. (P=0.582>0.05. We cannot reject the null hypothesis.)

3) Daily Folate: two sample t-test
   Assumptions: Random sampling. Independent observations. $n_A \geqslant 30$ & $n_B \geqslant 30$ or both populations are normal.
   Conclusion: There is not enough evidence to suggest a significant difference in daily folate between the high health literacy and low health literacy groups at alpha=0.05. (P=0.650>0.05. We cannot reject the null hypothesis.)

4) Insurance: Chi-squared test
   Assumptions: Each subject only contributes to one cell. Observations are Independent. All expected values are at least 5.
   Conclusion: There is not enough evidence to suggest a significant difference in insurance status between the high health literacy and low health literacy groups at alpha=0.05. (P=0.873>0.05. We cannot reject the null hypothesis.)

5) Education: Fisher's Exact Test
   Assumptions: Each subject only contributes to one cell. Random sampling. Observations are Independent. Applicability to the 2x2 Table. Small sample sizes.
   Conclusion: There is not enough evidence to suggest a significant difference in highest level of education between the high health literacy and low health literacy groups at alpha=0.05. (P=0.151>0.05. We cannot reject the null hypothesis.)

6) Health outcome(alcohol): two sample t-test
   Assumptions: Random sampling. Independent observations. $n_A \geqslant 30$ & $n_B \geqslant 30$ or both populations are normal.
   Conclusion: There is enough evidence to suggest a significant difference in alcohol use between the high health literacy and low health literacy groups at alpha=0.05. (P= 9.313e-09<0.05. We can reject the null hypothesis.)

b. Table 2 shows the results of the simple linear regression corresponding to the research question in Section I: to predict the health outcome using health literacy.

Table 2. Linear Regression Model Predicting Health Outcome (N = 1500)

| **Coefficients** | B (SE) |
|---|---|
| Intercept | 178.60 (4.38)*** |
| Health Literacy | -0.70(0.10)*** |

* p < 0.05, ** p < 0.01, *** p < 0.001

Intercept (178.60) represents the estimated health outcome (alcohol use) when the health literacy is zero.

Health Literacy (-0.70): This coefficient represents the estimated change in the health outcome (alcohol use) for a one-unit increase in health literacy, while holding other variables constant. If an individual's health literacy increases by one unit, the model predicts that their alcohol use will decrease by approximately 0.70 units.

R-squared (0.03347): The R-squared value measures the proportion of variance in the dependent variable (health outcome) that is explained by the independent variable (health literacy) in the model. In this case, the R-squared value is 0.03347, suggesting that approximately 3.35% of the variability in the health outcome can be explained by health literacy.

In summary, the results suggest that there is a statistically significant relationship between health literacy and the health outcome. However, the R-squared value is relatively low (3.35%), indicating that health literacy alone explains a small proportion of the variability in the health outcome. It supports the conclusion that there is enough evidence to suggest a significant difference in alcohol use between the high health literacy and low health literacy groups, as shown in table 1. It is also consistent with the figure 3, which indicates a negative correlation between them.

## IV. Conclusion

The negative coefficient for health literacy suggests that higher levels of health literacy are associated with lower health outcomes, and this relationship is highly significant. Figure 3 holds particular importance for the research question, indicating a correlation between health literacy and health outcomes. While other plots may suggest that health literacy alone explains only a small proportion of the variability in health outcomes, it's essential to approach the interpretation of these findings with caution. Further investigation and consideration of potential confounding variables may be necessary.

Based on Table 1, 'education' emerges as a critical factor, given the observed variations in health literacy across different education levels. Similarly, considering 'Above the Poverty Line,' despite lacking statistical significance, may provide valuable insights into the nuanced influence of economic status on health outcomes.

To comprehensively examine the health outcome (alcohol use), it is essential to collect information on several additional variables such as Support and Intervention History, Mental Health Status, and Genetic and Family History. Past or current

interventions, treatments, or support systems likely affect individuals' alcohol consumption. Assessing mental health indicators, such as stress levels, anxiety, and depression, would contribute to understanding their potential influence on physical health. Genetic predispositions and family history of alcohol-related issues may also help understand potential factors influencing alcohol consumption.