BS 100A Final Project

1. Statement of Research Question

The health outcome under consideration is Alcohol Use. The dataset was randomly collected from a survey that inquired about the frequency of alcohol consumption in a year, participants provided values on a continuous scale from 0 to 365 days. In Figure 1, we constructed a plot of the distribution of Alcohol Use. The distribution exhibits a normal shape, without any significant outliers. Interestingly, upon conducting a summary analysis of the dataset, it was revealed that the minimum recorded alcohol consumption is 23.65, while the maximum reaches 279.92 drinks. Before performing a bivariate plot, we anticipated a negative correlation between alcohol use and health literacy, indicating that higher alcohol use is associated with lower health literacy and vice-versa. Figure 2, the scatterplot, illustrates a negative correlation between the Health Literacy Score and Alcohol Use. Further reinforcing the idea, Figure 3 presents a Box Plot of Alcohol Use categorized by Health Literacy Score ranges in which lower health literacy scores correspond to higher levels of Alcohol Use. Therefore, both the Scatter Plot and Box Plot confirmed our expected relationship.

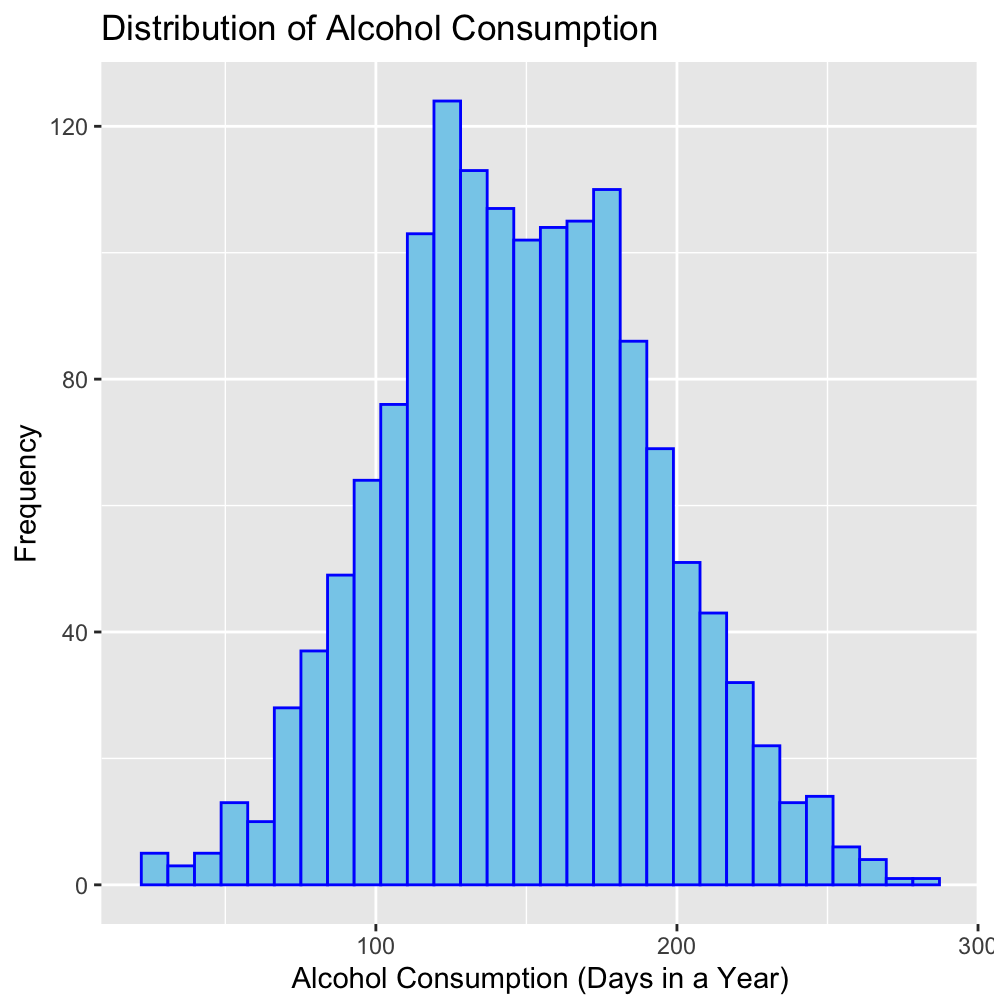


Figure 1. Alcohol Use Histogram

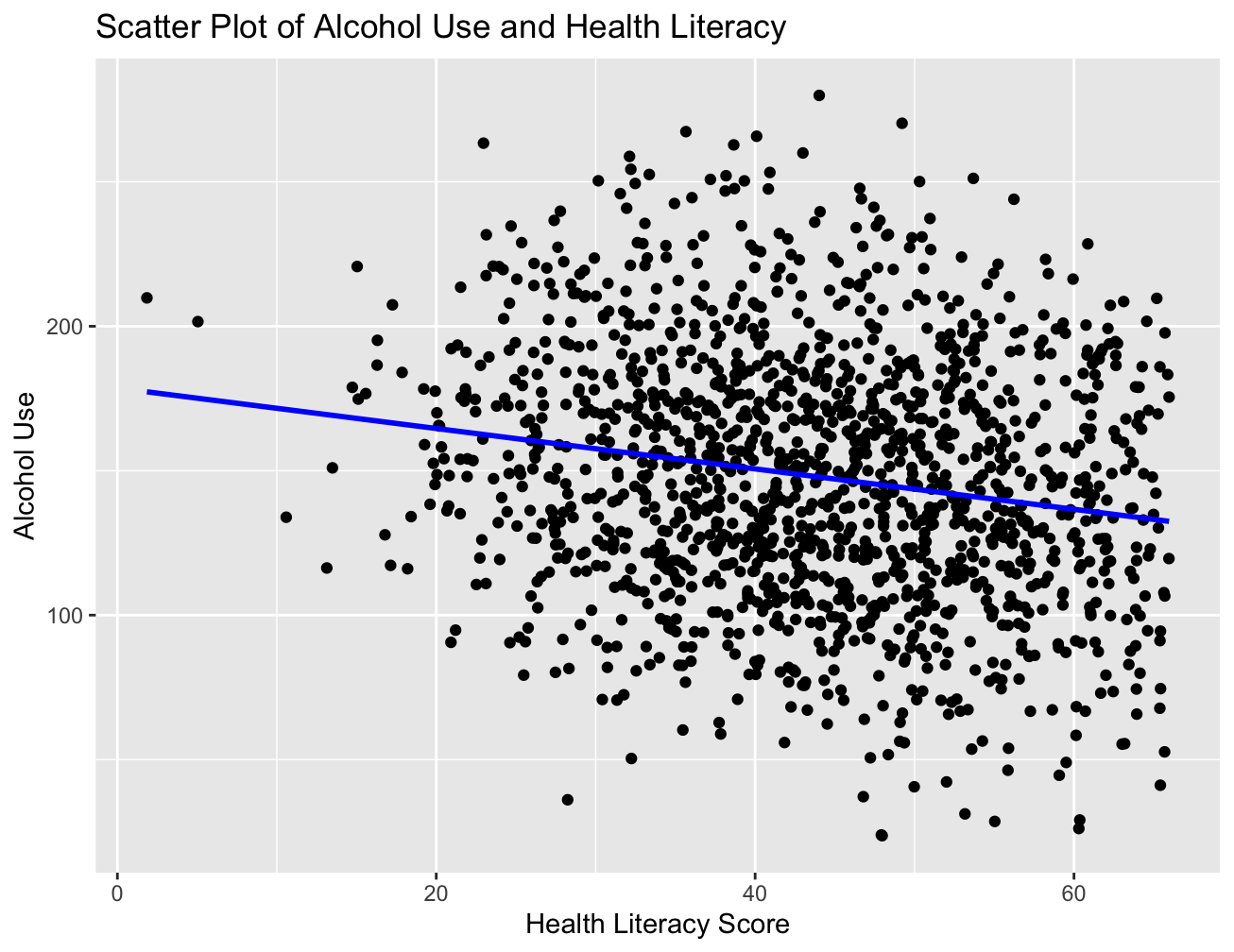
​​

Figure 2. Alcohol Use and Health Literacy Scatter Plot

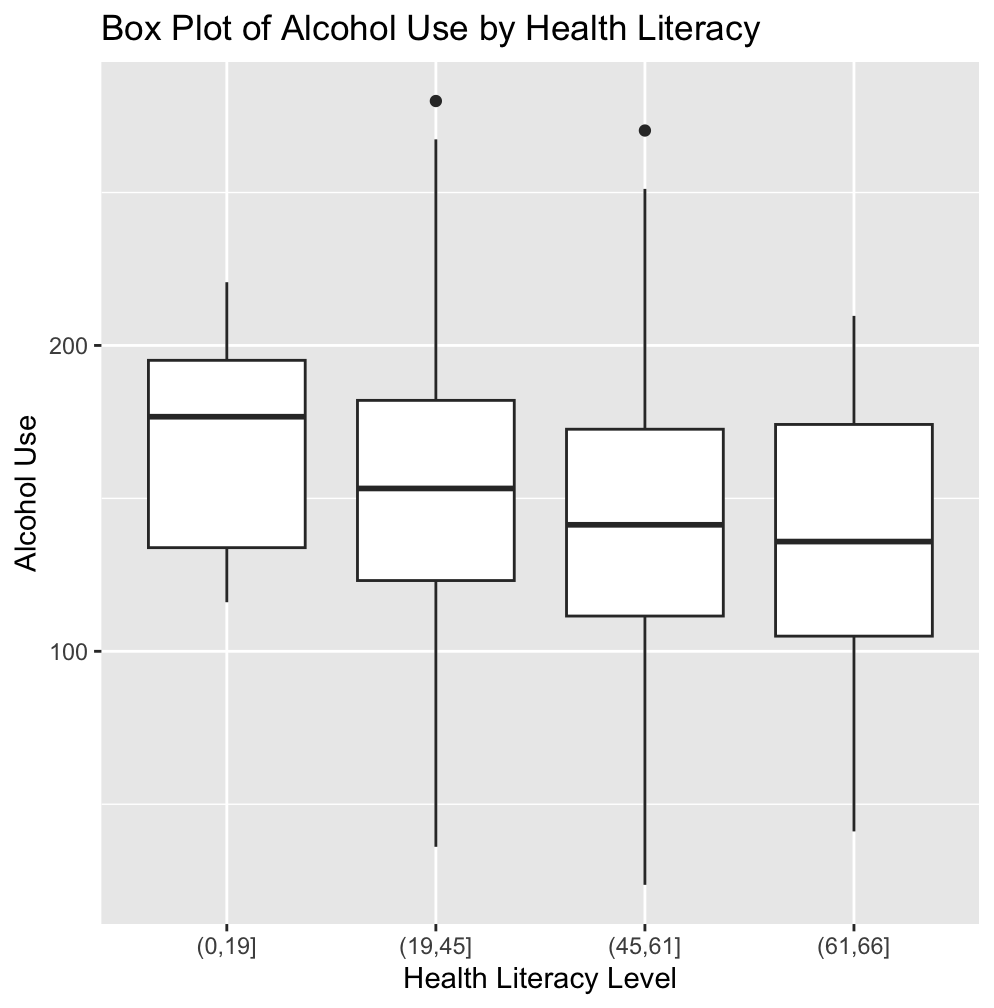


Figure 3. Alcohol Use by Health Literacy Box Plot

1. Description of Data
2. Health Literacy (hlth\_lit) is the primary focus of this report and it’s a continuous variable on the dataset. The Health Literacy scores within the 0 and 66 range are organized into distinct subcategories based on their scores which reflect their proficiency levels in reading materials. Figure 4 showcases a negative skew in the Health Literacy histogram with the majority of the population scoring over 30. However, there are some outliers that approach the score of 0. Figure 5 is a Scatter Plot between Alcohol Use and Health Literacy and shows a negative correlation.

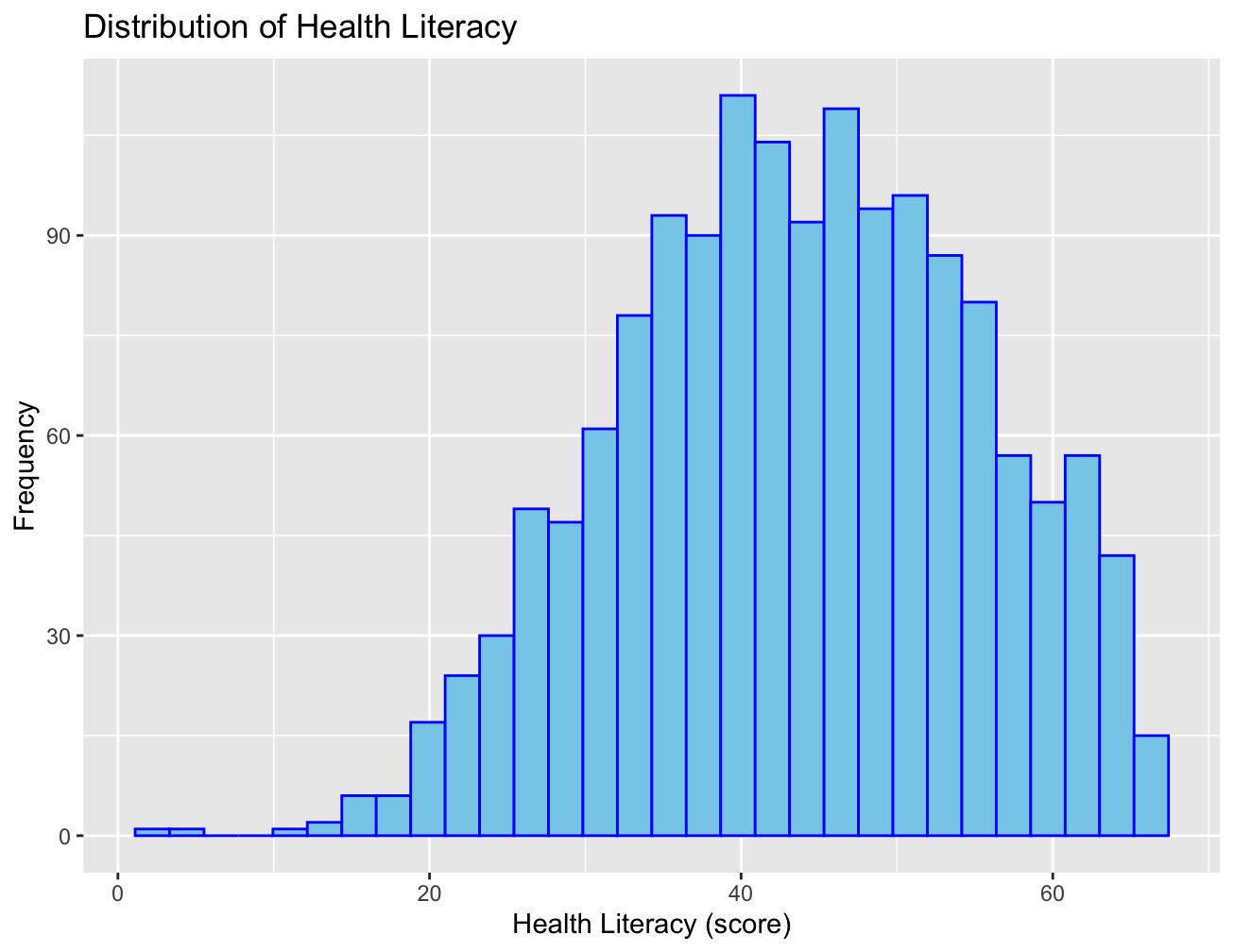


Figure 4. Health Literacy Histogram

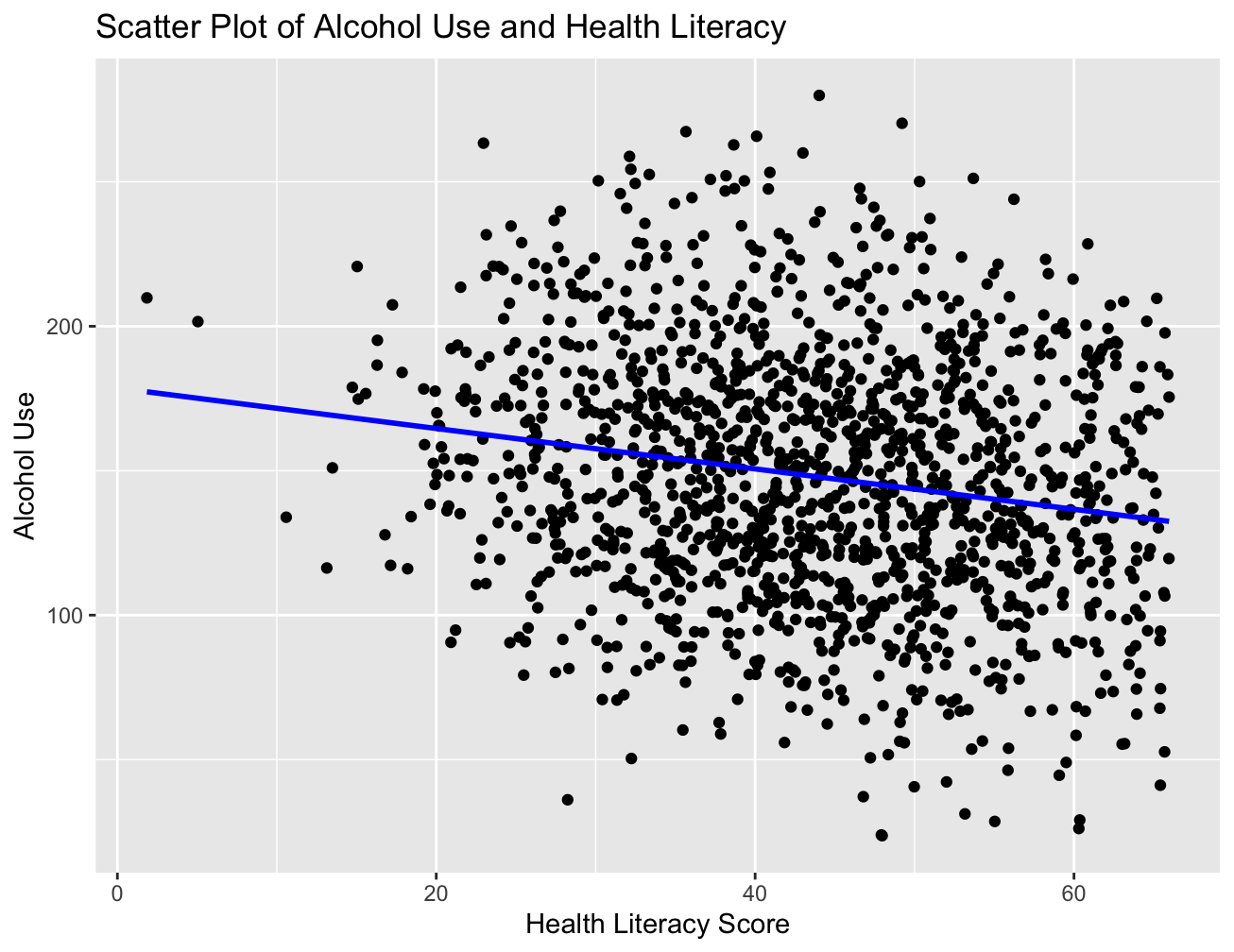


Figure 5. Alcohol Use and Health Literacy Scatter Plot

1. Sex is a categorical nominal variable, where male is coded as 0 and female as 1. In Figure 5, the distribution reveals that slightly over half of the population is female which can influence the study outcome. Figure 6 further illustrates that males tend to exhibit higher alcohol consumption compared to females. An interesting observation is that the lower alcohol use outliers are present in the female box plot, impacting the interpretation of the box plot's interquartile range.

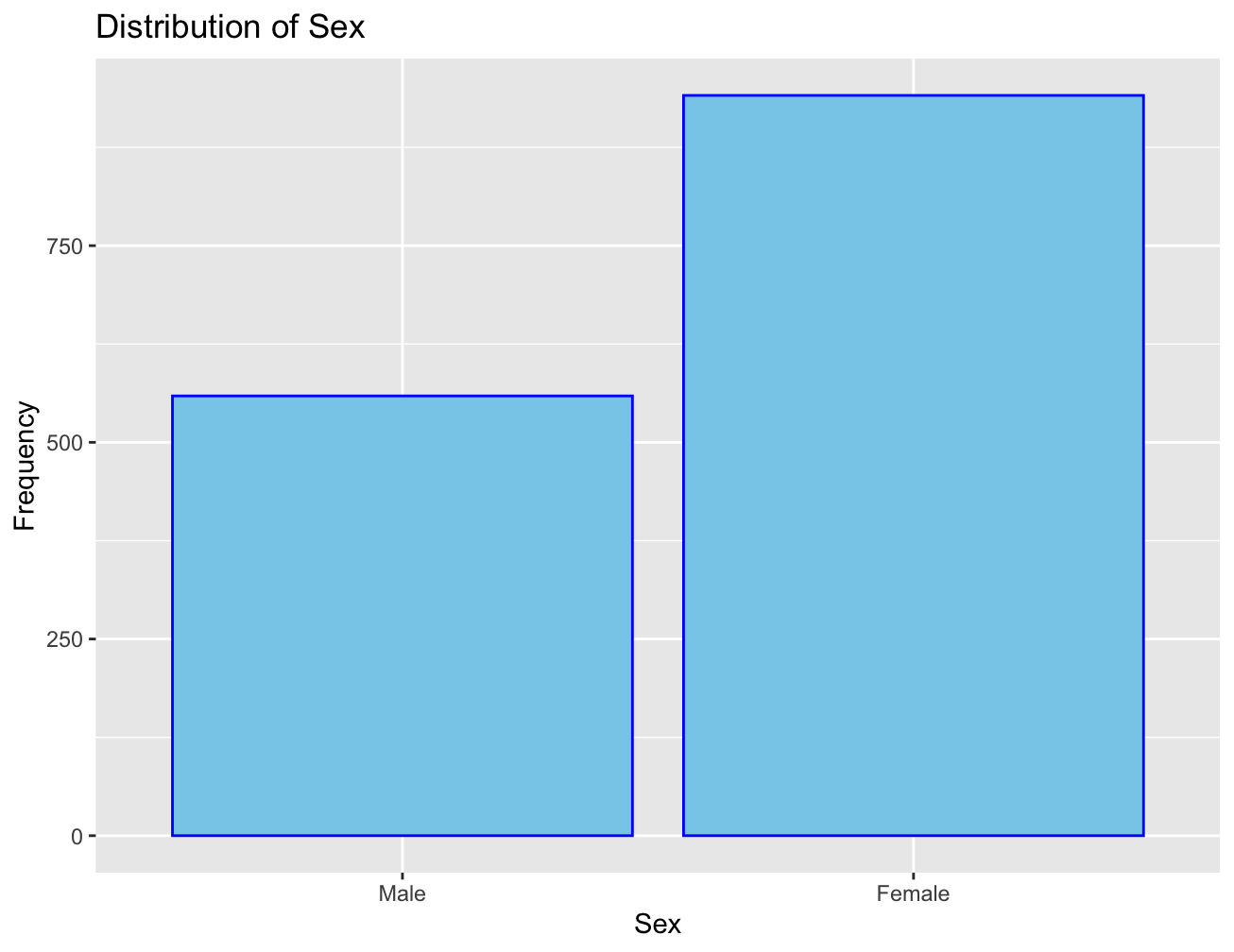


Figure 6. Distribution of Sex Bar Plot

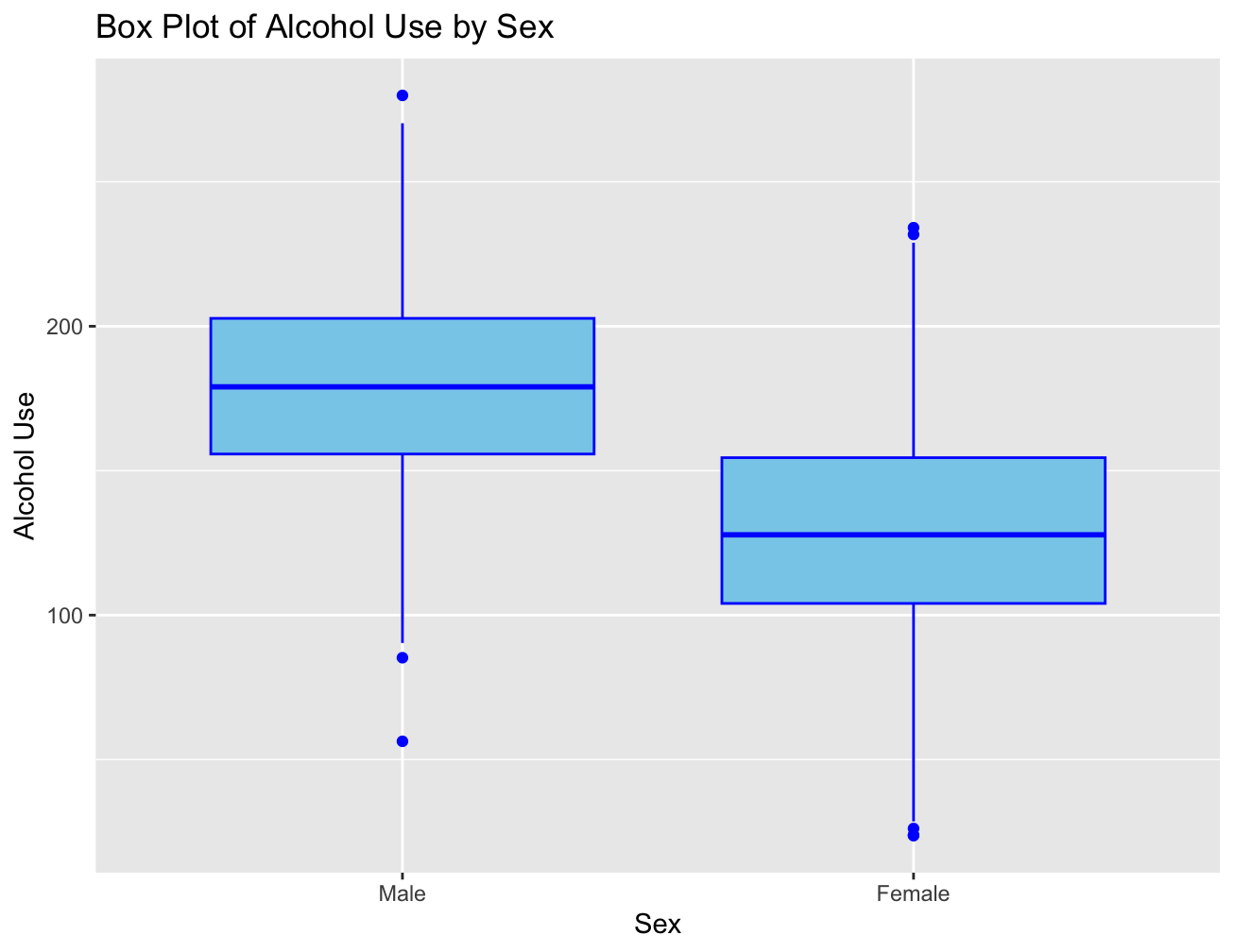


Figure 7. Alcohol Use and Sex Bar Plot

1. Living above the poverty line (pol) is a categorical ordinal variable, with 1 indicating 'yes' and 0 indicating 'no.' It is considered ordinal because there is an implied order of financial status based on a threshold. In Figure 8, the distribution reveals that roughly one-fourth of the population falls below the poverty line. In Figure 9, the Bar Plot of Alcohol Consumption with pol indicates a similar distribution of alcohol consumption for both above and below poverty line population.

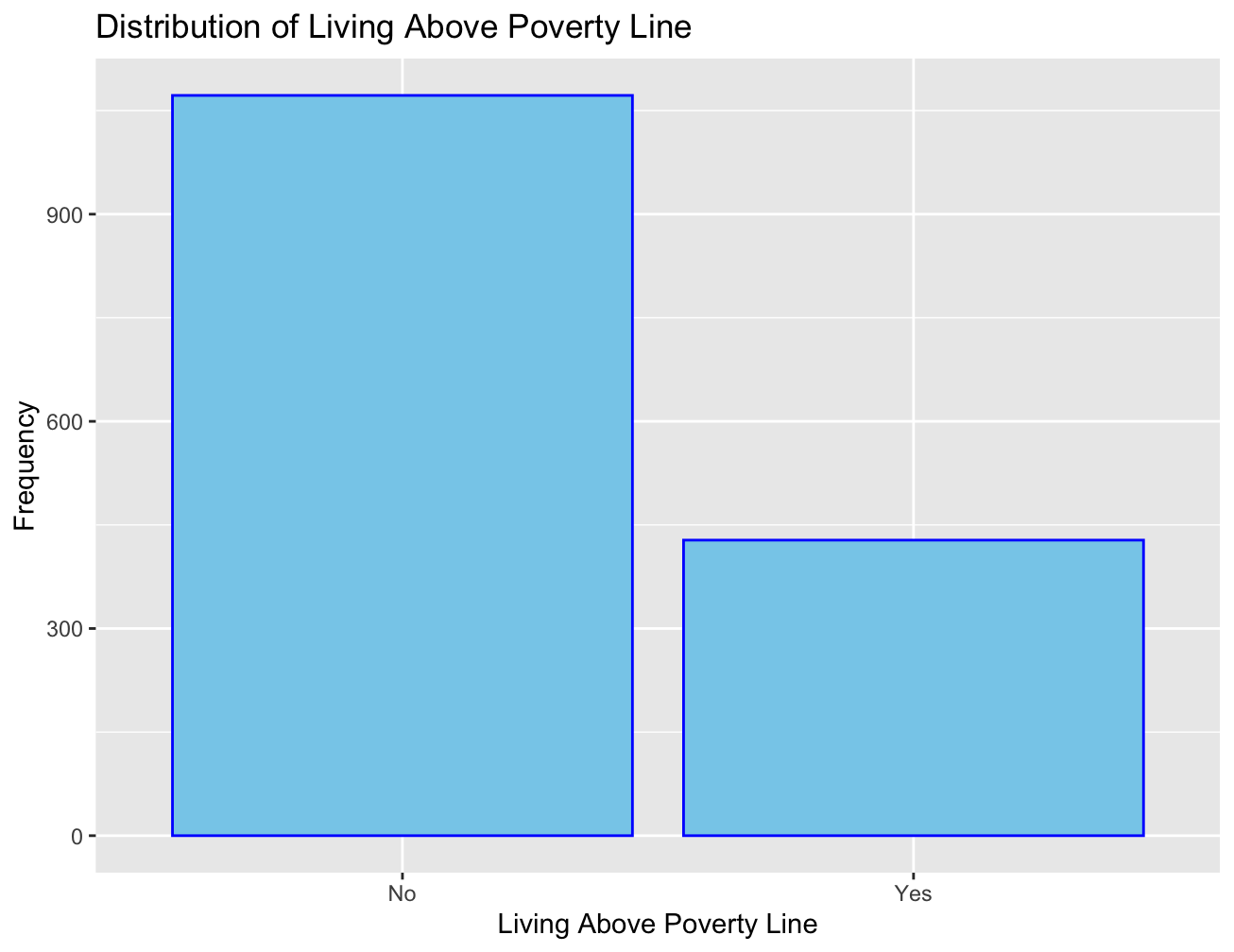


Figure 8. Distribution of Living Above the Poverty Line

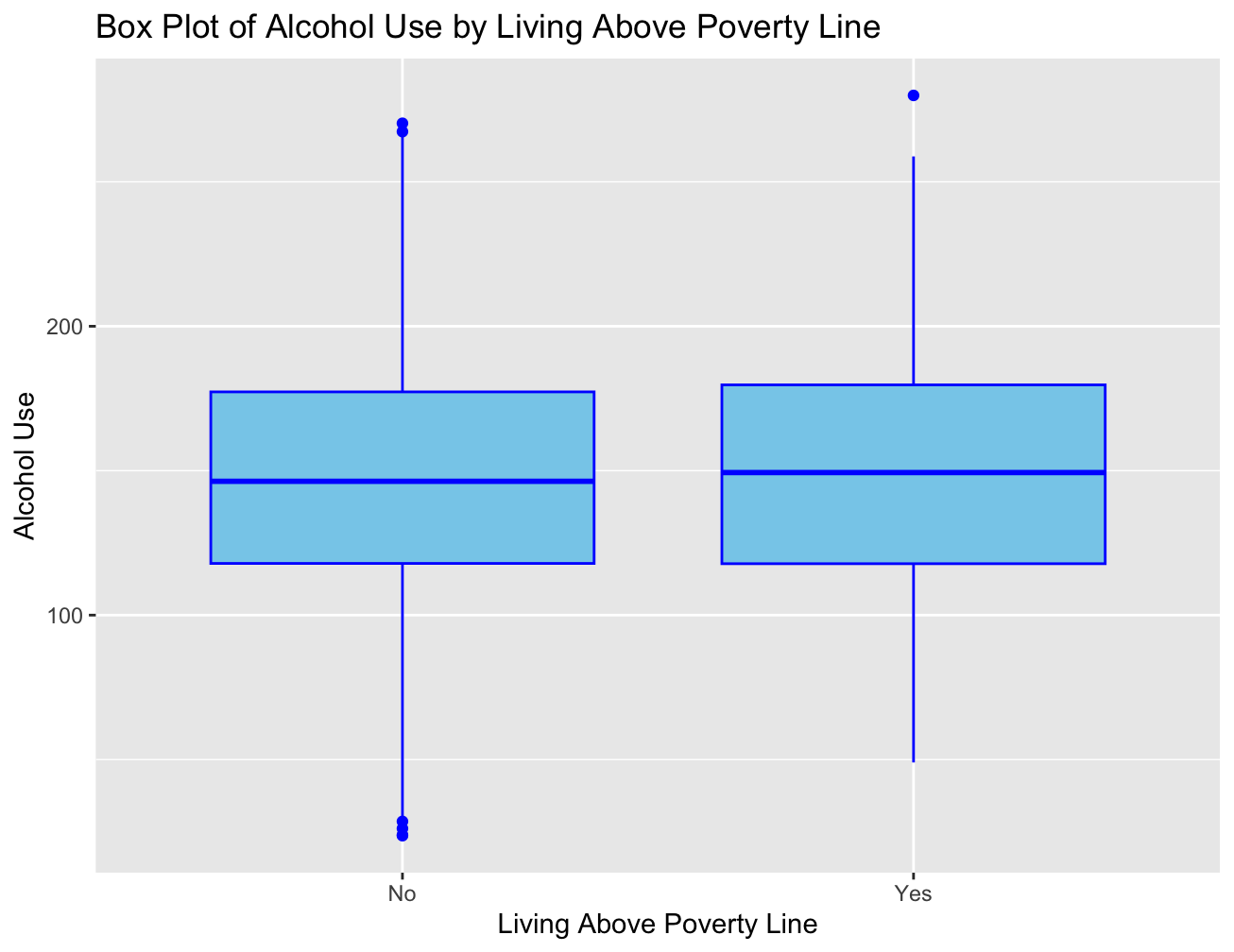


Figure 9. Alcohol Use and Living Above the Poverty Line Bar Plot

​​

1. Daily Total Folate Intake (daily\_fol) is a continuous variable. The recommended daily amount of folate for adults is 400 micrograms, and the positively skewed distribution in Figure 10 indicates that most of the population are above this threshold. In Figure 11, the scatter plot demonstrates a positive correlation, suggesting that as alcohol consumption increases, so does folate intake.

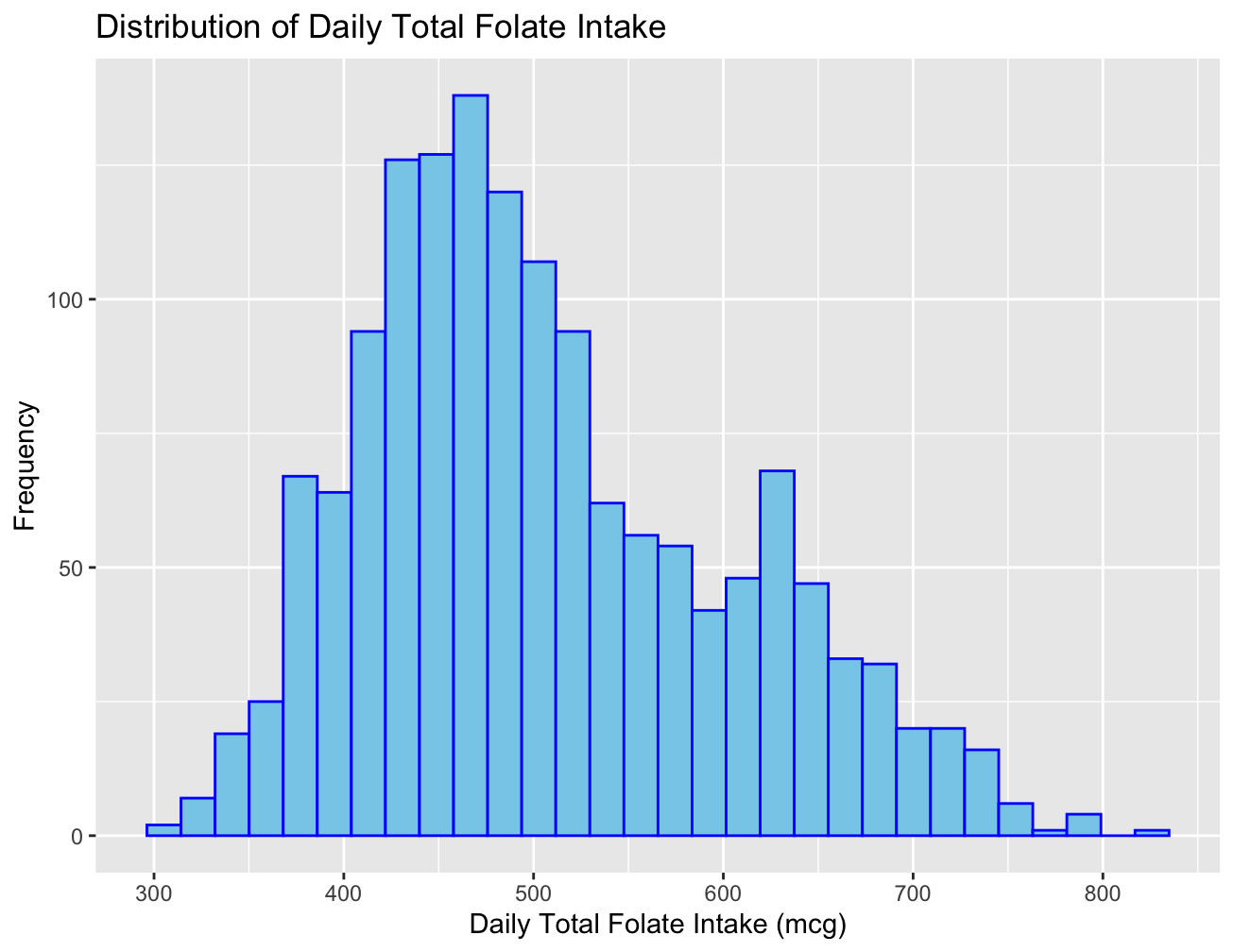


Figure 10. Distribution of Daily Total Folate Intake

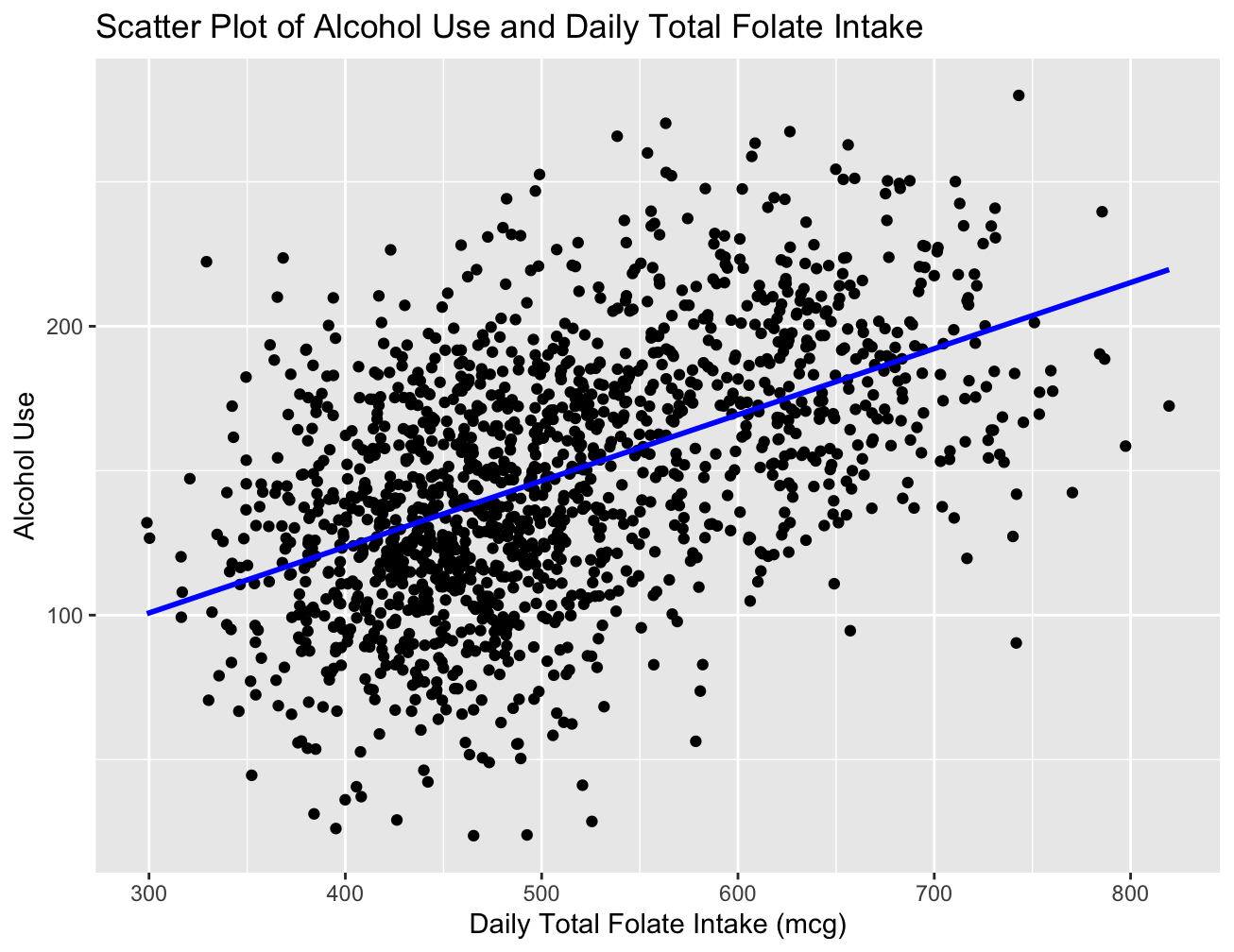


Figure 11. Alcohol Use and Daily Total Folate Intake

1. Insurance status is a categorical nominal variable as there is no inherent order of the categories of public insurance(0), private insurance(1) or being uninsured(2). The distribution of insurance seems heavily skewed towards public insurance, with a smaller proportion being uninsured, as depicted in Figure 12. In the box plot, Figure 13, all three subpopulations exhibit similar levels of alcohol use.

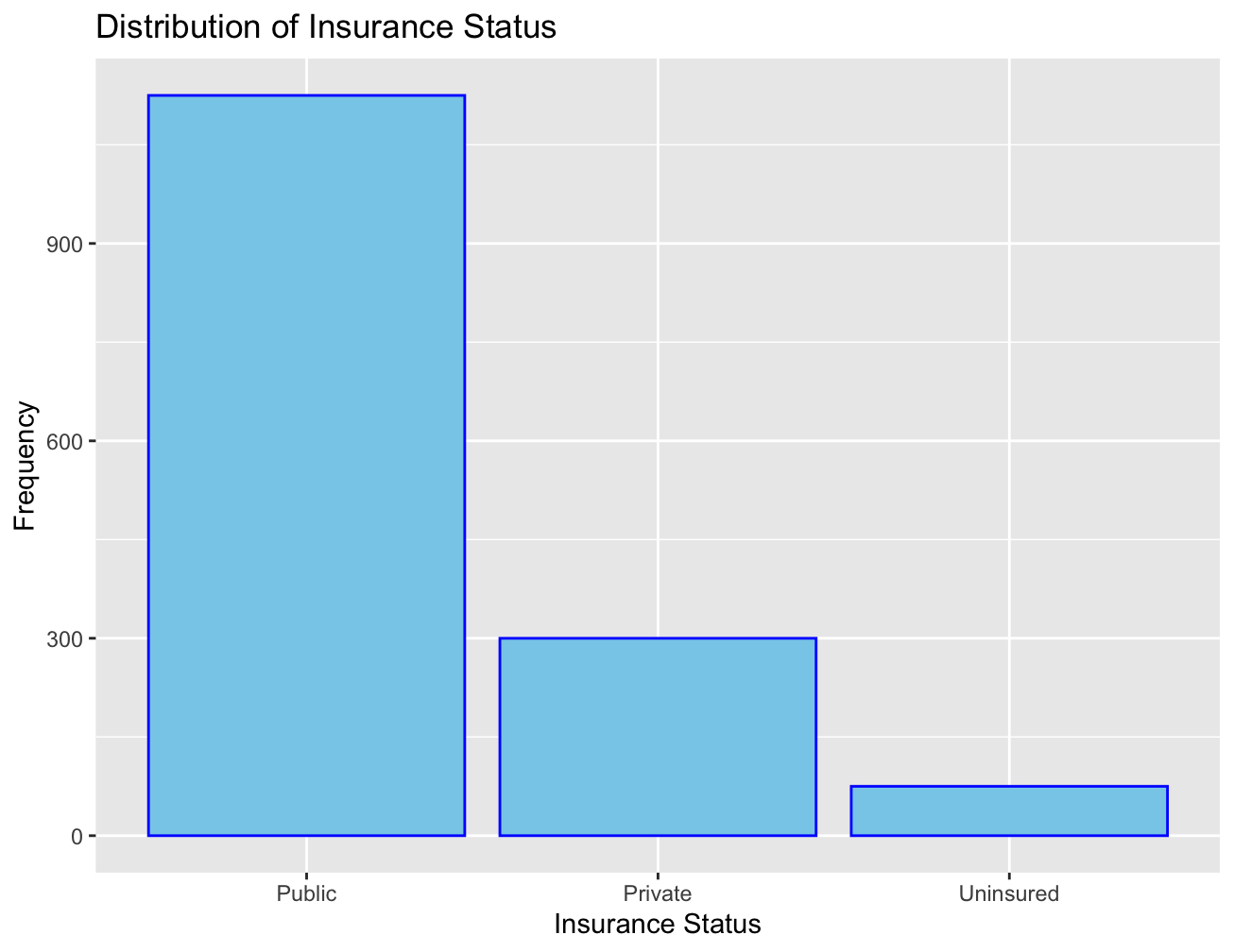


Figure 12. Distribution of Insurance Status

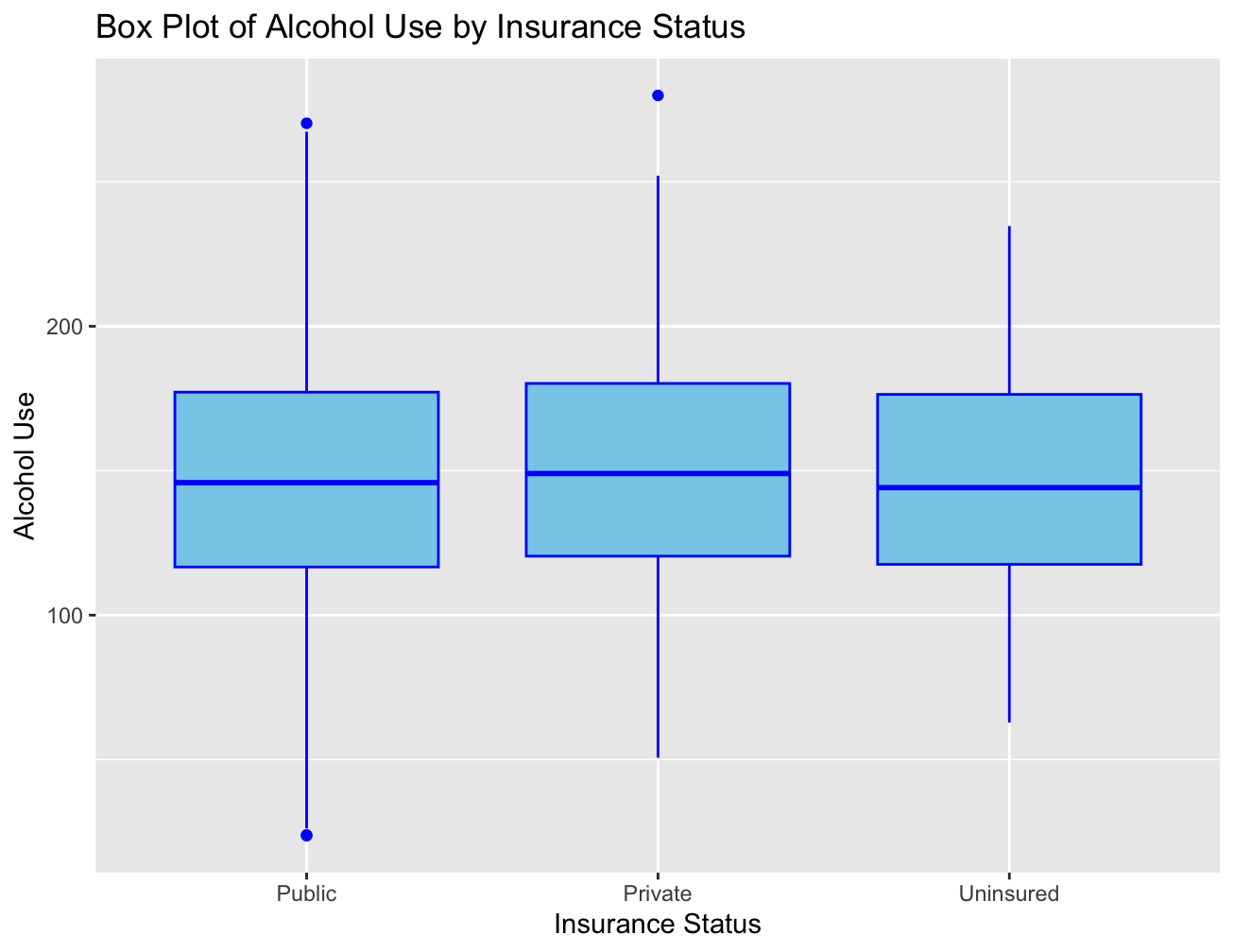


Figure 13. Alcohol Use and Insurance Bar Plot

1. Education is a categorical ordinal variable with a meaningful order which reflects progress to the next educational level. Categories include completion of elementary(0), high school(1), some college(2), college degree(3) and graduate degree(4). Figure 14 illustrates that the majority of the population falls within elementary, high school and some college categories, with very few individuals holding graduate degrees. The box plot comparing alcohol use and education level reveals similar alcohol use except for those with graduate degrees who exhibit higher alcohol use which is likely due to the smaller population of graduate degree holders.

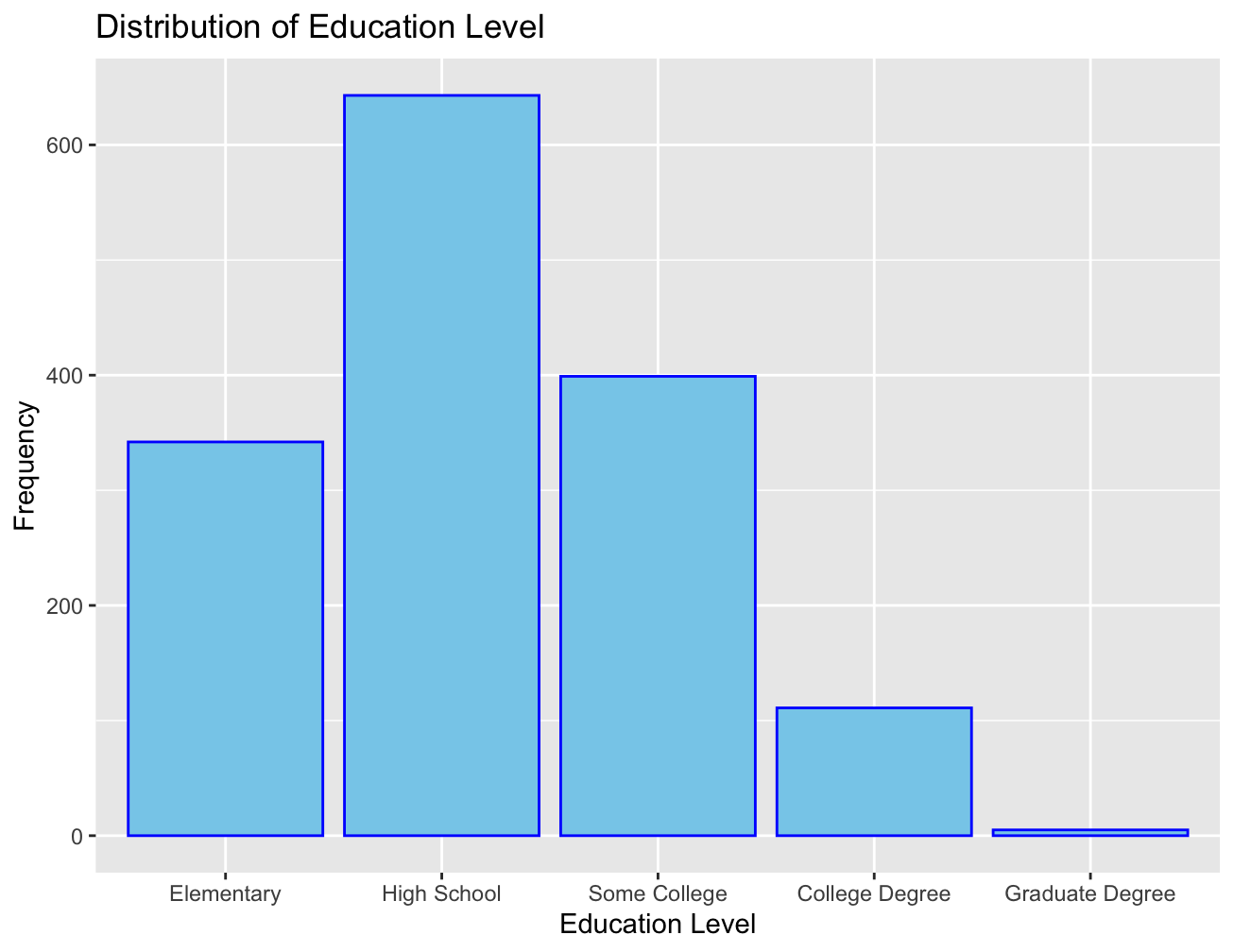


Figure 14. Distribution of Education Level

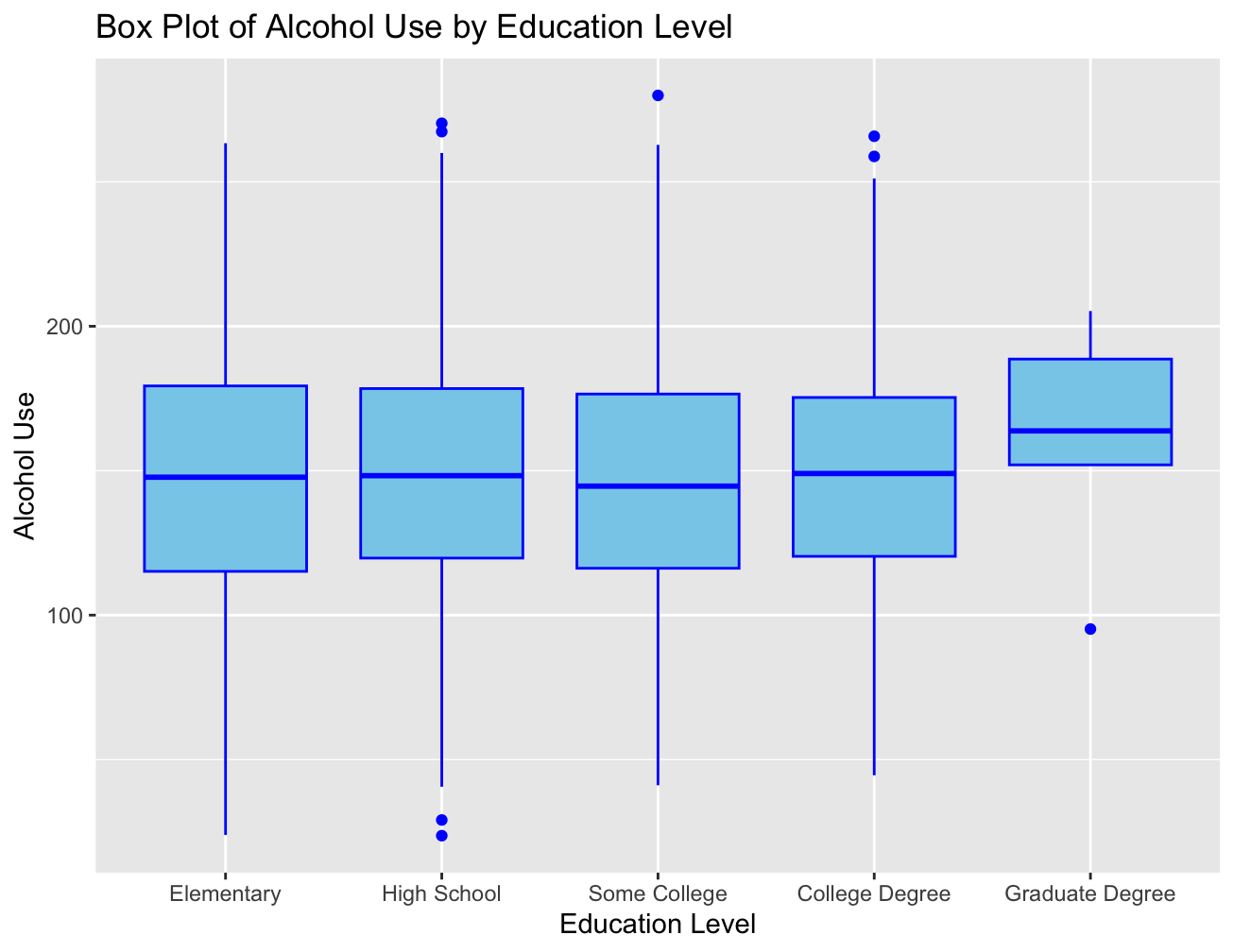


Figure 15. Alcohol Use and Education Level Bar Plot

1. Results

|  | Total  (N= )  Percent or  Mean (SD) | High Health Literacy  (N= )  Percent or  Mean (SD) | Low Health Literacy  (N= )  Percent or  Mean (SD) | p-value |
| --- | --- | --- | --- | --- |
| **Characteristics** |  |  |  |  |
| **Sex** |  |  |  | 0.905 |
| Male | 559(37.3%) | 260(37.0%) | 299(37.5%) |  |
| Female | 941(62.7%) | 442(63.0%) | 499(62.5%) |  |
| **Below the Poverty Line** |  |  |  | 0.582 |
| No | 1072(71.5%) | 507(72.2%) | 565(70.8%) |  |
| Yes | 428(28.5%) | 195(27.9%) | 233(29.2%) |  |
| **Daily Folate** | 506.8(96.1) | 505.6(94.3) | 507.8(97.8) | 0.650 |
| **Insurance** |  |  |  | 0.873 |
| Public | 1125(75.0%) | 527(75.1%) | 598(74.9%) |  |
| Private | 300(20.0%) | 142(20.2%) | 158(19.8%) |  |
| Uninsured | 75(5.0%) | 33(4.7%) | 42(5.3%) |  |
| **Education** |  |  |  | 0.147 |
| Elementary school | 342(22.8%) | 143(20.4%) | 199(24.9%) |  |
| High School | 643(42.9%) | 310(44.2%) | 333(41.7%) |  |
| Some College | 399(26.6%) | 189(26.9%) | 210(26.3%) |  |
| College degree | 111(7.4%) | 56(8.0%) | 55(6.9%) |  |
| Graduate Degree | 5(0.3%) | 4(0.6%) | 1(0.1%) |  |
| **Health Outcome**  **(Alcohol)** | 148.5(43.6) | 141.2(43.9) | 154.1(42.5) | 9.3\*10^-9 |

The overall purpose of Table 1 is to explore the relationship between health literacy and the different variables by comparing the high and low literacy groups with the variables and obtaining a p-value. The null hypothesis for each variable row states that there is no association between the variable and health literacy while the alternative hypothesis states that there is association. The variable “Sex” yielded a high p-value following the chi-square test of independence when testing the association between sex and high health literacy. The assumptions for the test are independence between observations, random sampling and a large sample size. The test resulted in a p-value of 0.905, exceeding the alpha value of 0.10, thus failing to reject the null hypothesis. For the “Below the Poverty Line”, “Insurance” and “Education” variables additional chi-square tests of independence were made with the same assumptions. The resulting p-values were 0.582, 0.873 and 0.147 respectively. Therefore, the variables also failed to reject the null hypothesis as they exceeded the alpha value of 0.10. For the variable “Daily Folate”, a paired t-test was performed with assumptions of normal distribution, random sampling, and independence. The resulting p-value of 0.650 led to the failure to reject the null hypothesis. For the given Health Outcome variable, “Alcohol Use”, we conducted another paired t-test with the same assumptions, yielding a small p-value of 9.312 x 10^-9. Therefore, we can reject the null hypothesis, indicating a strong association between alcohol use and health literacy.

| Table 2. Linear Regression Model Predicting Health Outcome (N = 1500 ) | |
| --- | --- |
|  |  |
| **Coefficients** | B (SE) |
| Intercept | 50.76050(1.025447)\*\*\* |
| Health Literacy | -0.04785(0.006644) \*\*\* |

The regression analysis, summarized in Table 2, provides insights into the relationship between Alcohol Use and Health Literacy. The intercept of 50.76050 represents the estimated Health Literacy score when Alcohol Use is 0 and acts as a reference. The coefficient of Alcohol Use, -0.04785 indicates that for each coefficient increase in Alcohol Use, the Health Literacy score is expected to decrease by approximately 0.04785. However, the R-squared value of 0.0334678 suggests that only about 3.35% of the variability in Health Literacy can be explained by the Alcohol Use variable. This small R-squared implies that there are many other factors that are likely to influence Health Literacy. The results from Table 2 which presents the regression analysis, do not entirely align with the associations observed in Table 1 and the scatter plot from the first section. In Table 1, a highly significant p-value of 9 x 10^-9 was obtained, indicating a strong association between Alcohol Use and Health Literacy. The scatter plot in the first section further illustrated a negative correlation, suggesting that higher alcohol use is linked to lower health literacy. However, the regression results in Table 2 reveal a weak correlation. This implies that while statistically significant, the impact of Alcohol Use on Health Literacy scores may not be as pronounced as initially believed by the p-value and Scatter Plot in section 1.

1. Conclusion

In conclusion, there were some significant findings of the relationship between alcohol use and health literacy found using Table 1 & 2 and graphing tools. The effects of alcohol use on health literacy were first highlighted by the Scatter Plot and Box Plot of Health Literacy and Alcohol Use in the Statement of the Research Question. The Scatter Plot highlighted the negative relationship while providing a general look of the data and the Box Plot showed that participants with Score of 0 to 18.99 had a higher alcohol use than participants with a Score of 61 to 66. Then, Table 1 showed a correlation with a statistically significant small p-value of 9 x 10^-9. The regression analysis in Table 2 revealed a small R squared value,-0.04785, suggesting that only a small percentage of the variability is explained by Alcohol Use. For future studies, variables like socioeconomic status, education level, sex and daily folate are relevant to include for variability in the health result. For example, education and socioeconomic can be related to decision making, sex can suggest that one gender is more prone to consuming alcohol and folate can be inhibited by large alcohol consumption. On the other hand, insurance can be related to socioeconomic status, thus another variable could be added instead. Variables like mental health scores, diet, exercise frequency could contribute to a more comprehensive understanding of health. These future variables are directly related to the current health of the individual which can often provide a better understanding of a person and show a better relationship to our health outcome, alcohol.