BS 100A Final Project

Zongzhe Lin Uid: 206328707

Section I. Statement of Research Question:

a. Smoking as a Health Outcome: A Continuous Variable

In our study, smoking is defined as a continuous variable, quantified by the number of cigarette packs smoked in the past 12 months. This approach allows for a precise measurement of smoking intensity, ranging from occasional to heavy usage. By assessing smoking in this manner, we can effectively analyze its relationship with other indicators and socio-demographic variables in the dataset.

b. Figure 1: Histogram of Annual Cigarette Consumption

Figure 1 illustrates the annual cigarette consumption among the surveyed 1,500 individuals. The frequency of cigarette pack usage is displayed on the y-axis, while the x-axis indicates the number of packs. Notably, the data exhibits a moderate right skew, with a median near 53 packs, suggesting that cigarette use is prevalent among the sample population. This finding deviates slightly from my initial expectations of lower consumption. Also, a pronounced peak around 120 packs indicates the presence of heavier smokers and potential existence of outliers.

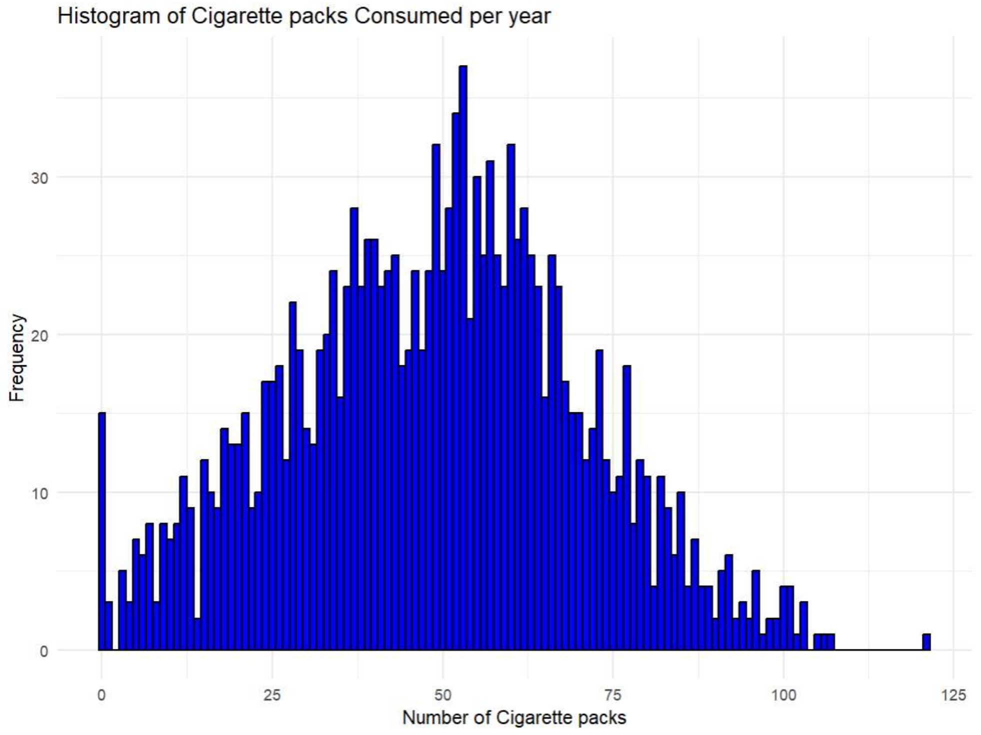

Fig. 1: Histogram of annual cigarette pack consumption among the surveyed individuals

c. Hypothesis on Health Literacy's Influence on Smoking

I propose that there exists a significant inverse relationship between health literacy and smoking frequency. Higher health literacy levels are anticipated to correlate with a lower number of cigarette packs consumed, as informed individuals are more likely to comprehend the adverse health effects of smoking and, therefore, may exhibit reduced smoking prevalence.

d. Figure 2: Bivariate Plot Analysis - Health Literacy vs. Smoking

Figure 2 showcases a scatter plot that reveals a negative trend between health literacy scores and the number of cigarette packs consumed annually. The trend line slopes downward, indicating that participants with higher health literacy generally smoke less. Despite this clear trend, there is notable scatter among the data points, especially in the middle range of health literacy scores, suggesting that other factors may also influence smoking behavior. This visual trend supports the hypothesis that increased health literacy is associated with reduced smoking prevalence.

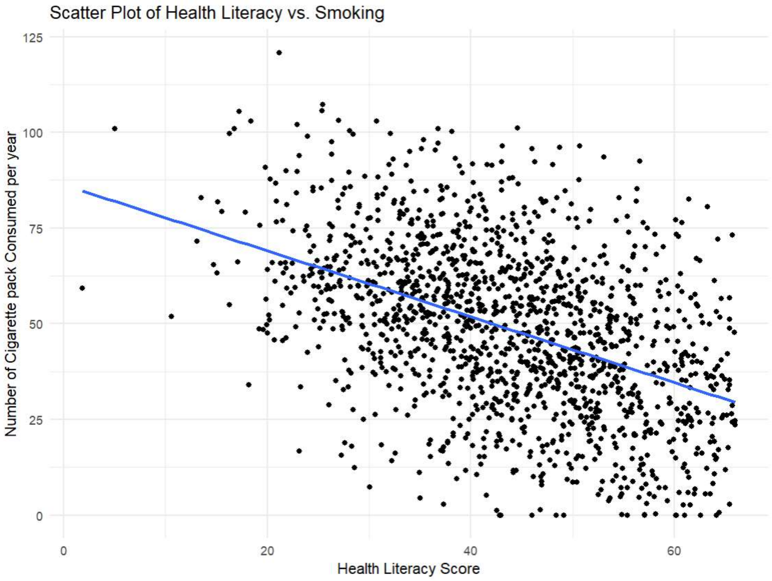

Fig. 2: Correlation between health literacy and annual cigarette pack consumption.

e. Construction of a Binary Health Literacy Variable

A binary variable, named 'high_health_literacy,' is created using R. This variable assigns a value of 1 to individuals with health literacy scores of 45 or above, indicating high health literacy, and a value of 0 to those with scores below 45.

Section II. Description of Data:

Our dataset encompasses six key non-health outcome variables, each serving as a potential predictor in our analysis. These variables are:

- **(hlth_lit)**: **Health Literacy** is a quantifiable measure of an individual's ability to recognize and comprehend health-related terms. Represented as an ordinal variable, it is distributed on a scale from 0 to 66, mirroring traditional educational grade levels. The distribution, shown in Figure 3a, appears to be multimodal. However, a significant portion of the study population has low literacy scores. The presence of a few scores near zero may represent outliers, potentially signaling individuals with very limited literacy skills.

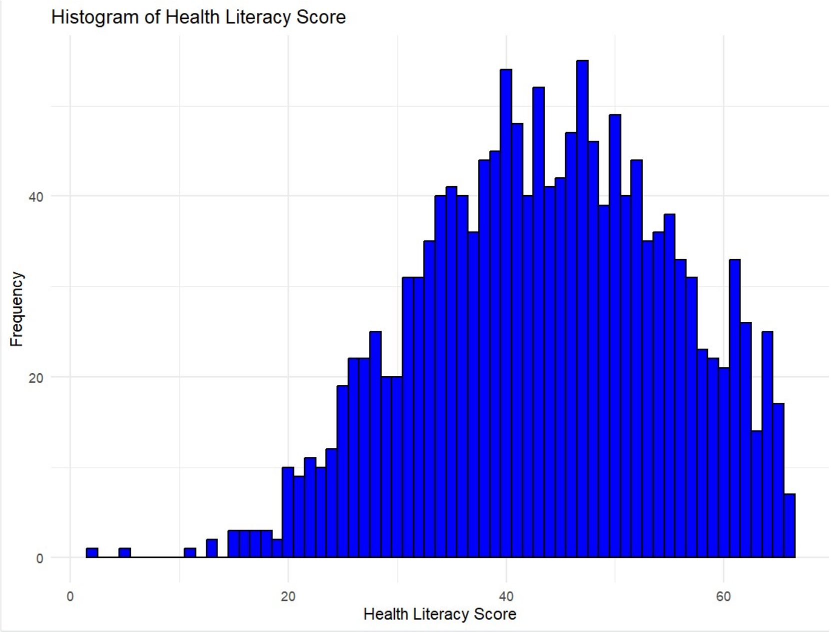

Fig. 3a: Distribution of health literacy scores with a left-skew and possible outliers.

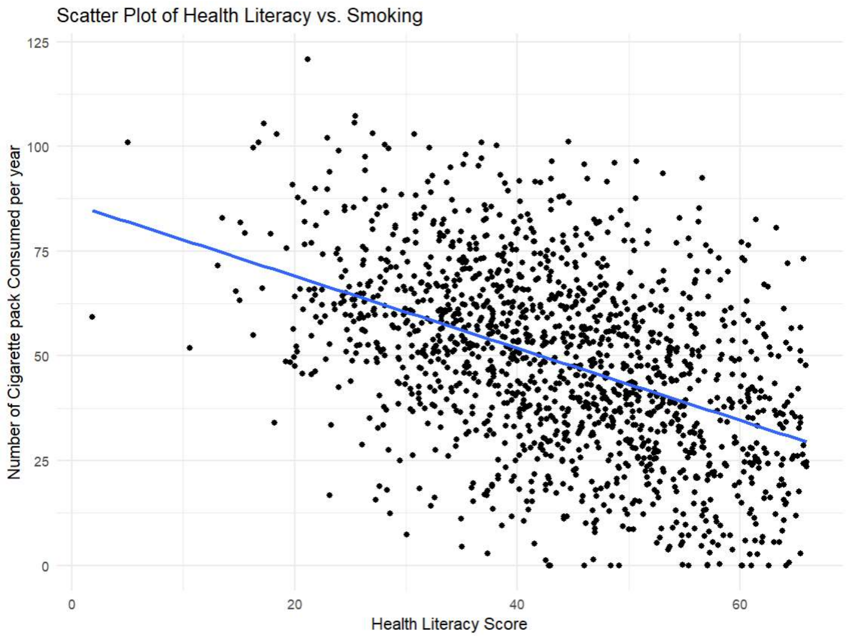

Fig. 3b: Health literacy scores against yearly cigarette consumption,

- **(sex): Sex** is a demographic variable. It categorizes respondents into two groups: Male (0) and Female (1). Figure 4a reveals a greater number of female respondents compared to males, indicating a gender imbalance in our sample. The bar chart does not exhibit skewness, as it displays categorical data, but it does show a few outliers for both genders. Moving to smoking habits, Figure 4b, a box plot, provides a comparative view of cigarette consumption by gender. It is evident that male respondents tend to smoke more, as indicated by the higher median and larger interquartile range in cigarette use. This visual comparison underscores a gender disparity in smoking prevalence within our dataset.

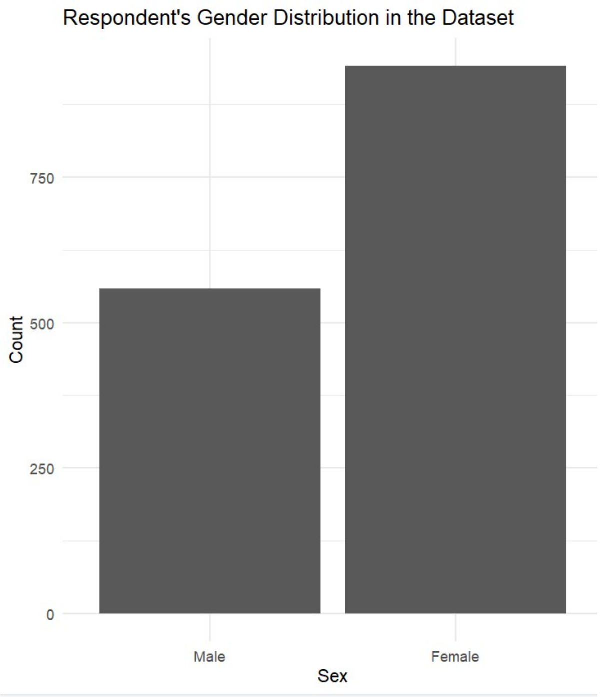

Fig. 4a: Bar chart showing the distribution of respondents by gender

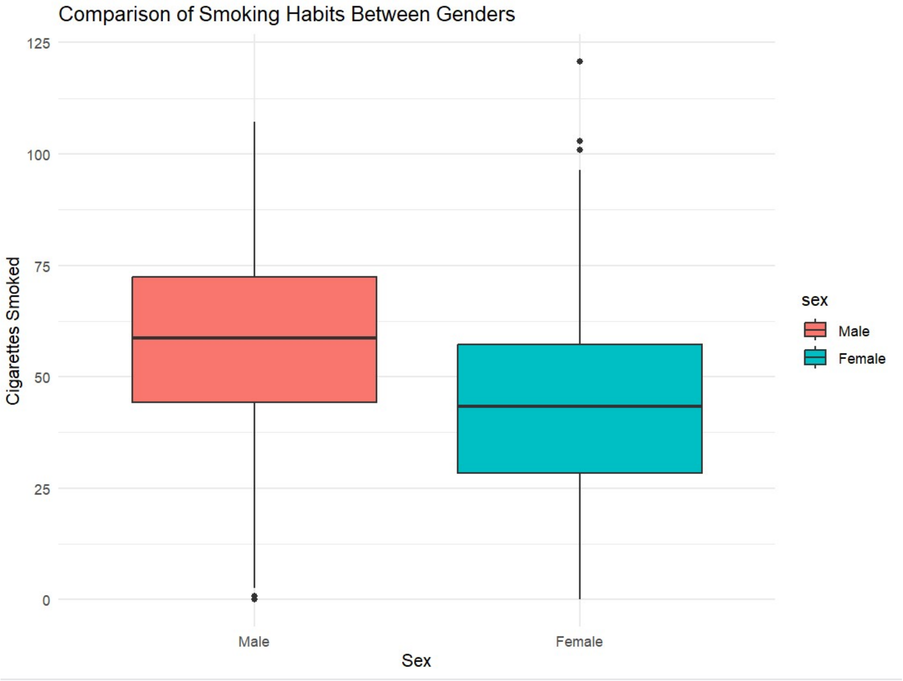

Fig. 4b: Box plot comparing cigarette consumption between genders

- **(Pol): Living above the poverty line** is a binary categorical variable. It categorizes respondents based on their economic status in relation to the poverty line. Figure 5a shows a disproportionately larger number of individuals residing below the poverty line compared to those above it, highlighting economic disparities within the sample. When examining smoking habits in relation to economic status (Figure 5b), we observe that the distribution of cigarette consumption is quite similar across both groups. Interestingly, individuals above the poverty line exhibit a marginally higher median in cigarette usage, a finding that counters typical expectations. While there are outliers present, they do not appear significant enough to skew the overall comparison. This suggests that economic status, as defined by the poverty line, may not be a predominant factor in determining the levels of cigarette consumption among the respondents.

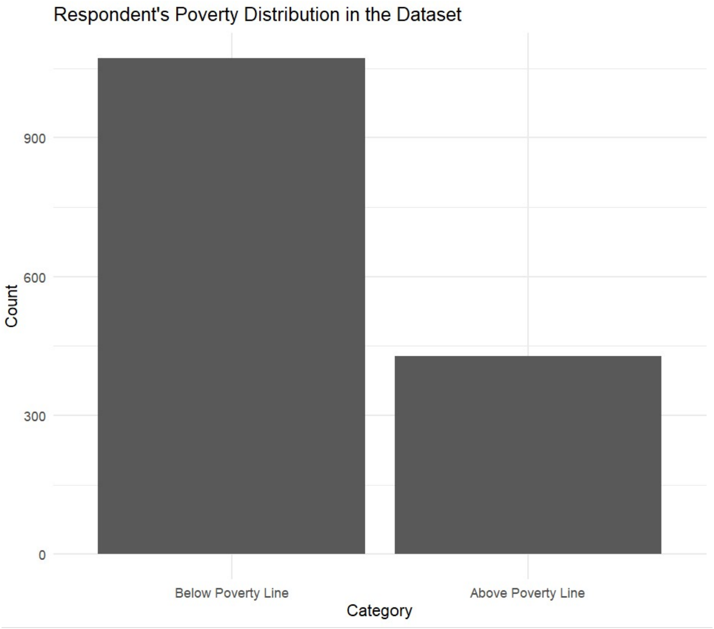

Fig. 5a: Bar chart depicting the distribution of respondents by poverty line status

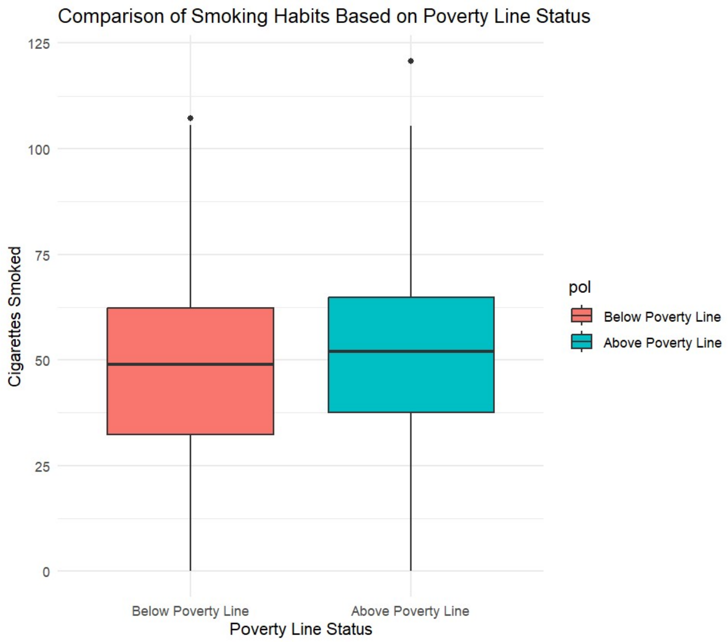

Fig. 5b: Box plot comparison of smoking habits based on poverty line status

- **(Daily_fol): Daily Total folate intake** is a continuous variable related to the daily intake of folate. it can also be considered scale data with an implied range of recommended daily amount of folate for adults, which is 400 micrograms(mcg).

Figure 6a's histogram suggests a distribution that is right-skewed, with a concentration around 450 micrograms, slightly above the recommended daily intake. This skewness indicates that while most individuals meet the recommended levels, there is a subset with significantly higher intakes. In the scatter plot of Figure 6b, an upward trend is noticeable, where higher daily folate intake is associated with an increase in cigarette consumption. This counterintuitive trend might suggest a complex relationship between nutritional intake and smoking habits, possibly influenced by other lifestyle factors not captured by the folate variable alone.

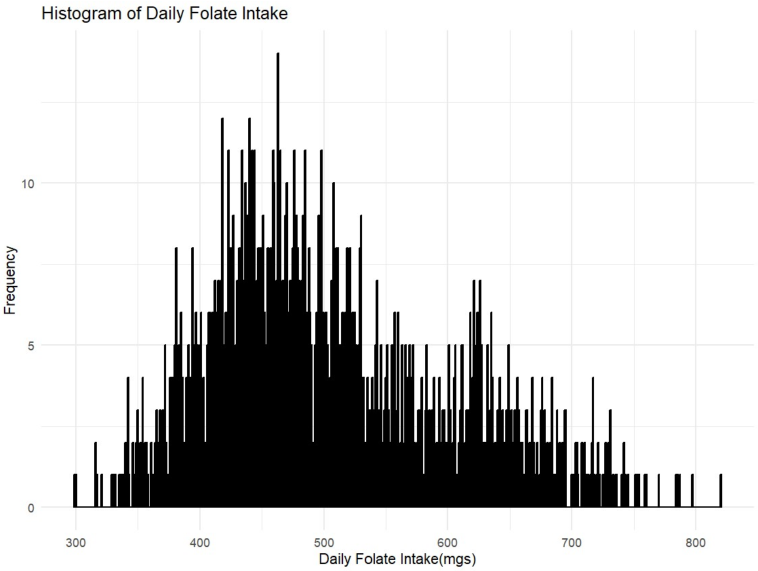

Fig. 6a: Histogram of daily folate intake

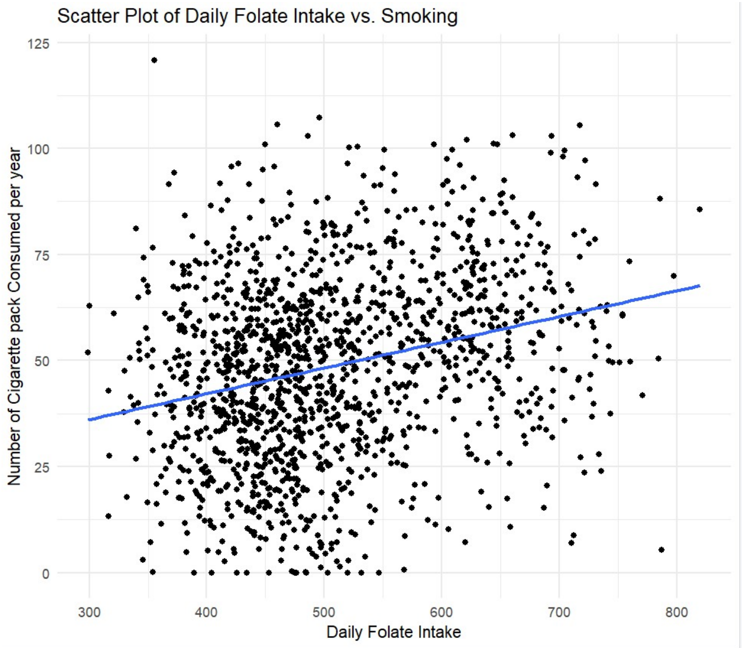

Fig. 6b: Scatter plot between daily folate intake and the pack of cigarettes smoked annually.

- **Ins: Insurance status** is a categorical variable. The 'ins' variable categorizes respondents based on their insurance status: public insurance (0), private insurance (1), and uninsured (2). As illustrated in Figure 7a, our dataset contains a significantly higher number of individuals with public insurance compared to those with private insurance or without any insurance. From the box plot in Figure 7b, we notice a couple of outliers within the public insurance group. Despite these outliers, the median cigarette consumption across the insurance categories is quite similar. However, it's noteworthy that the uninsured group shows a higher mean number of cigarettes smoked. This could reflect socioeconomic factors affecting health behavior, where uninsured individuals may engage in higher-risk health behaviors such as smoking.

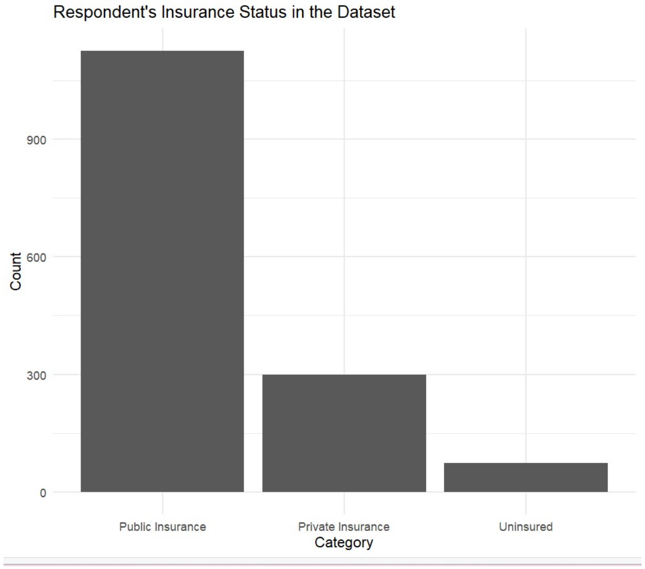

Fig. 7a: Bar chart illustrating different types of insurance among the respondents

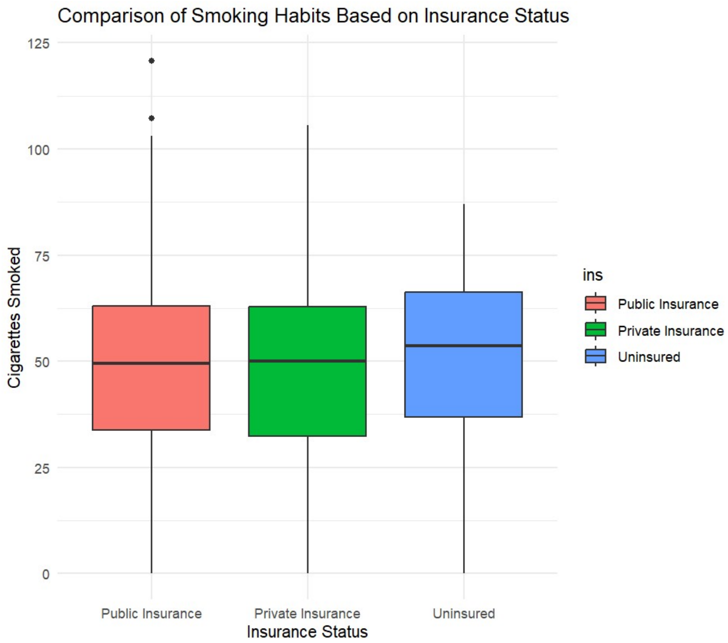

Fig. 7b: Box plot demonstrating smoking habits across different insurance statuses

- **Educ: Education Level** is an ordinal variable. The data presented in Fig. 8a and Fig. 8b elucidate the relationship between education level and smoking habits among the survey respondents. Fig. 8a reveals a predominant concentration of individuals whose highest educational attainment is a high school diploma. A lesser proportion of the surveyed population has achieved education beyond the high school level, indicating that the majority of respondents fall below a college degree in educational attainment. Contrastingly, Fig. 8b offers compelling visual evidence of an inverse relationship between education level and cigarette consumption. It becomes apparent that individuals with higher educational credentials, such as college and graduate degrees, tend to smoke cigarettes less frequently. This trend is indicative of a potential correlation where increased education correlates with healthier lifestyle choices, in this case, reflected by reduced smoking habits.

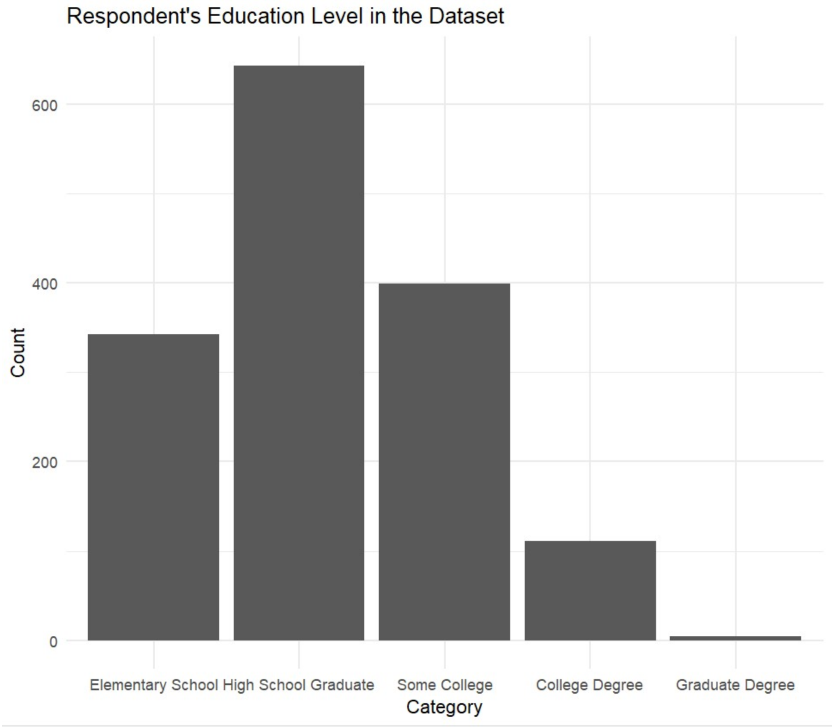

Fig. 8a: Bar Chart of respondents' highest completed education level

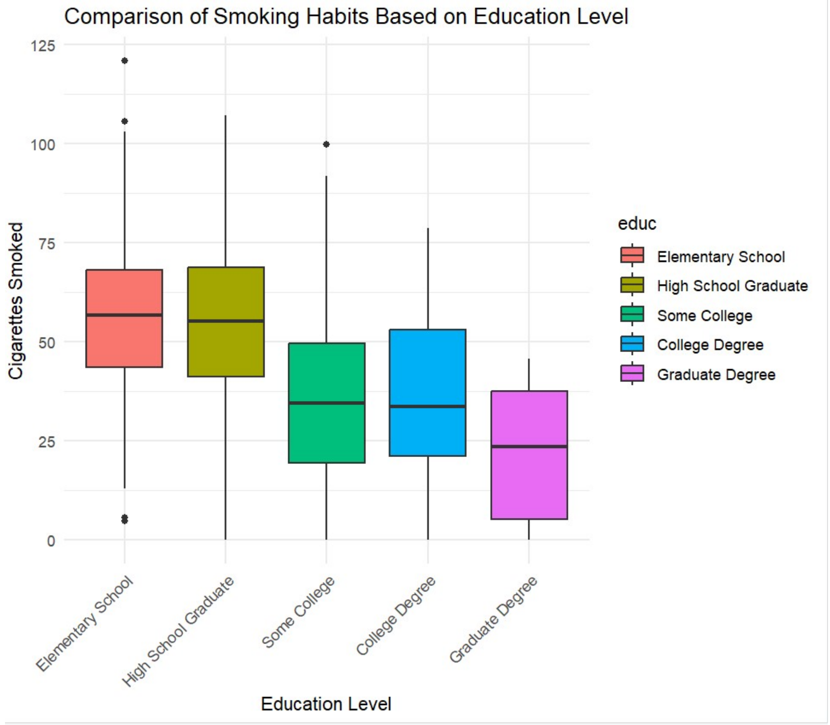

Fig. 8b: Box Plot of Smoking by Education Level

Section III. Results:

Table 1. Characteristics of the Sample in Total and by Health Literacy
Sex, Pol, Folate, Insurance, and Edu used chi-square, the rest used two-proportion z test.
For the two continuous variables, Daily Folate and Health Outcome(Smoke), we use t test.

	Total (N=1500) Percent or Mean(SD)	High Health Literacy (N=702) Percent or Mean (SD)	LowHealth Literacy (N= 798) Percent or Mean (SD)	p-value
Characteristics				
Sex	1500(100%)	702(46.8%)	798(53.2%)	0.863
Male	559(37.27%)	260(37%)	299(37.5%)	0.023*
Female	941(62.7%)	442(63%)	499(62.5%)	0.0098**
Above the Poverty Line	1500(100%)	702(46.8%)	798(53.2%)	0.543
No	1072(71.5%)	507(72.2%)	565(70.8%)	0.014*
Yes	428(28.5%)	195(27.8%)	233(29.2%)	0.011*
Daily Folate - Mean (SD)	506.79(96.13)	505.60(94.30)	507.85(97.77)	0.65
Insurance	1500(100%)	702(46.8%)	798(53.2%)	0.873
Public	1125(75%)	527(75.1%)	598(74.9%)	0.0031**
Private	300(20%)	142(20.2%)	158(19.8%)	0.221
Uninsured	75(5%)	33(4.7%)	42(5.26%)	0.191
Education	1500(100%)	702(46.8%)	798(53.2%)	0.147
Elementary school	342(22.8%)	143(20.4%)	199(24.9%)	2.6e-0.5
High School	643(42.9%)	310(44.2%)	333(41.7%)	0.22
Some College	399(26.6%)	189(26.9%)	210(26.3%)	0.157
College degree	111(7.4%)	56(7.98%)	55(6.89%)	1
Graduate Degree	5(0.333%)	4(0.57%)	1(0.125%)	0.206
Health Outcome - Mean (SD)	48.68(21.71)	39.98(20.12)	56.34(20.12)	2e-16***

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

The sample comprises 1500 respondents, with 702 classified as having high health literacy and 798 with low health literacy. To examine the characteristics of this sample, three statistical tests were employed:

1. The **Chi-Square Test** was utilized for categorical variables including Sex, Above the Poverty Line, Insurance, and Education. The assumptions for this test include independent samples, a sufficiently large sample size, random sampling, and categorical data. A p-value below the conventional significance level of 0.05 indicates a statistically significant association between the variables. Conversely, a non-significant result implies that the observed association could be attributable to chance rather than a true effect.
2. The **Two-Proportion Z-Test** was applied to sub-categories of the Poverty Line, Insurance, and Education since the Chi-Square Test is not appropriate for 1x2 tables. The assumptions of this test are similar to the Chi-Square Test, with the addition of requiring a large sample size that justifies a normal approximation. A p-value less than 0.05 suggests a significant difference between the two proportions. A non-significant result indicates that the difference between sample proportions reflect random sampling variability.
3. The **T-Test** was conducted for two continuous variables: Daily Folate and the health outcome variable "Smoke". This test assumes independent samples, normal distribution of data, homoscedasticity, random sampling, and a continuous scale of measurement. A p-value below 0.05 signifies a statistically significant difference in means between the two groups. A non-significant result points to the possibility that any observed difference in sample means is due to random variation rather than a genuine difference in the population.

Table 2.
Linear Regression Model Predicting Health Outcome **Smoke** (N = 1500)
Multiple R-squared: 0.205

Coefficients	B (SE)
Intercept	86.31(1.98)
Health Literacy	-0.86(0.044)

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

Result Explanation:

The line in Figure 2 is my simple linear regression. We can see that the line is going downward as the health literacy score increases. It represents a general phenomenon that people with higher health literacy score tend to smoke lesser cigarette.

Table 2 displays the results of my linear regression model with data from 1500 individuals regarding the impact of health literacy score on health outcome “smoke”. The Intercept is 86.31, meaning that without health literacy, our model will predict in total 86.31 packs of cigarette each individual consumed per year. The Health Literacy coefficient is -0.86, meaning that with every single point of Health Literacy Score, individual’s cigarette consumptions per year will go down by 0.86. This negative relationship align with my previous statement that higher health literacy is associated with lower smoking rates. The R^2 value of this model is 0.205, indicating that health literacy accounts for 20.5% of the variance in the cigarette consumption. This score suggests that health literacy has a significant influence. We can see from Table 1 that there is a big difference in smoking behavior based on health literacy, and the trend is the same as our linear regression suggests in Table 2. Thus Table 2 supports the difference in Table 1.

Section IV Conclusion:

a. The table demonstrates a significant inverse relationship between health literacy and smoking. Specifically, a unit increase in health literacy corresponds to a 0.86 unit decrease in smoking. This statistically significant finding suggests that higher health literacy is associated with lower smoking levels.

b. Figure 2 stands out as the pivotal plot, showcasing a linear trend between health literacy and smoking through a scatter plot with a regression line. It corroborates our hypothesis and statistical findings, highlighting the direct relationship between increased health literacy and reduced smoking.

c. Variables like Poverty Line and Insurance exhibit notable differences across health literacy groups. These factors could be influential, potentially explaining a portion of the variability in cigarette consumption."

d. To further explore smoking behavior, incorporating variables such as alcohol consumption, drug usage, stress levels, mental health status, income, and employment status would be beneficial.