BS 100A Final Project

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

You will be responsible for compiling the following short report. I expect that these reports (with plots and tables) will be at least 10 pages. This should be double spaced, Times New Roman, 12 point. While there is no page requirement, please make sure that you answer each question to the best of your ability (while keeping answers succinct). All tests may be done at the $\alpha = 0.05$ level. You will submit one report (including all relevant plots & tables) and one .R file.

You have randomly been placed in a group with three other students and each assigned one of the four health outcomes. While the work you present must be your own, you may look for guidance (both in interpretation and coding) from your groupmates. I recommend you sit together for the last days of class, as there will be time to work on the project!

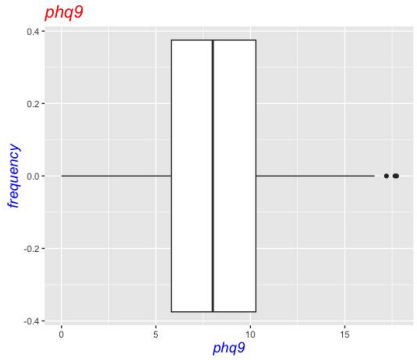
Your paper MUST contain the following sections (I – IV) with headings (and answer the corresponding questions as paraphs in these sections). Figures (plots) and tables should be included in the text so that they are easily referenced and must be labeled (e.g. "In Fig. 4, we see that…"). Grading of the report will be based on the quality of responses, readability, and statistical decision making. In addition, all tables and figures must be thoughtfully constructed and easy to read or points will be deducted. All code used in this analysis must be cleaned and submitted as a .R file. All work MUST be reproducible.

I. Statement of Research Question

a. Consider your health outcome. Describe it (Is it continuous? Categorical? Ordinal? Nominal?).

phq is continuous

b. Construct a plot of the distribution of this variable and describe it. Is it normal/skewed? Are there any outliers that you are concerned about?

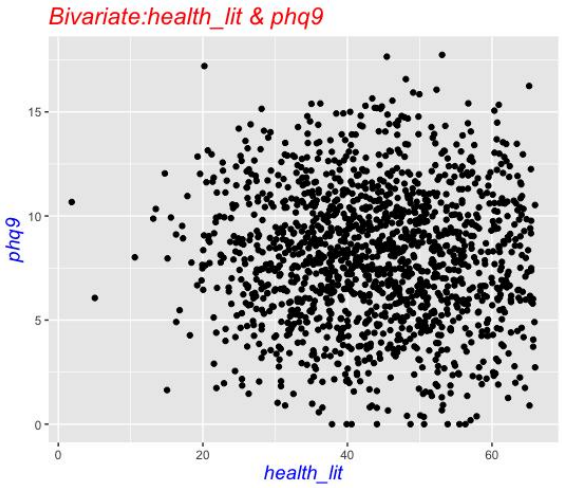


phq9 distribution concentrate in 7-10, outlier are 17 and 18

c. How do you expect health literacy to be associated with this outcome?

health literacy is positive associated with this outcome. In general, improvements in health literacy may lead to better health outcomes, but the specific associations depend on the specific research area and question.

d. Construct a bivariate plot comparing health literacy to your health outcome and discuss.



It can be observed that there seems to be some positive correlation between health literacy and health outcomes. In addition, a horizontal line be added which represents the average of the health outcomes.

e. Create a binary variable of high health literacy, defined as 1 if health literacy >= 45 and 0 if health literacy < 45.

df$high_health_literacy <- ifelse(df$health_lit >= 45, 1, 0)

II. Description of Data

a. For each variable non-health outcome variable, write a short paragraph with the following:

   i. 1-2 sentence description of the measure (can paraphrase from the description above). Consider how you would describe the variable. Is it continuous? Categorical? Ordinal? Nominal? For scale data, is there an implied range to the data?

- Health literacy (hlth_lit) is a primary predictor of interest and is measured using a 66-item word recognition test. The scores are categorized into four grade equivalents, ranging from 3rd grade and below to high school level. This variable is ordinal in nature, as the scores represent different levels of literacy.

- Sex (sex) is a categorical variable with two levels: male (0) and female (1). It is a nominal variable, as there is no implied order or ranking between the two categories.

- Living above the poverty line (pol) is a binary variable indicating whether an individual is above (1) or below (0) the poverty line. It is a nominal variable.

- Daily total folate intake (daily_fol) is a continuous variable representing the amount of folate consumed by an individual on a daily basis. There is an implied range of 0 to 365 micrograms.

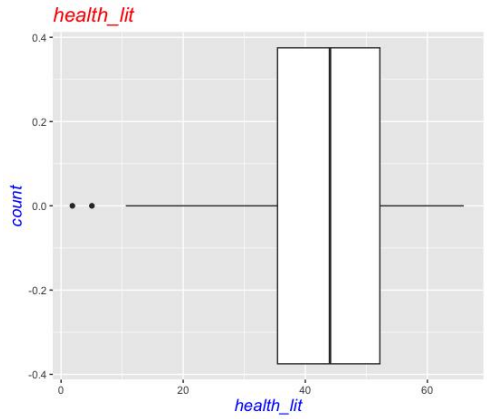
- Insurance status (ins) is a categorical variable with three levels: public insurance (0), private insurance (1), and uninsured (2). It is a nominal variable.

- Highest level of education (educ) is a categorical variable representing the educational attainment of an individual. It has five levels: elementary school education (0), high school graduate (1), some college (2), college degree (3), and graduate degree (4). It is an ordinal variable, as there is a logical order to the levels.

- Alcohol use (alc) is a continuous variable indicating the number of days an individual consumed any type of alcoholic beverage in the past 12 months. The range of this variable is from 0 to 365.

ii. Plot of the variable. Description of the distribution. Is it normal/skewed? Are there any outliers that you are concerned about?

hlth_lit:

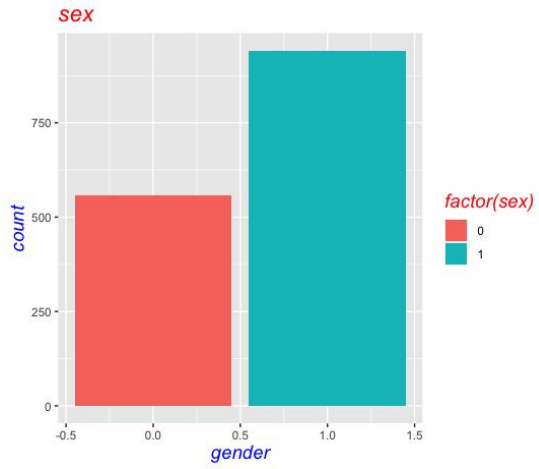


Most scores are concentrated between 30 and 50.

The median of the health_lit is 42.

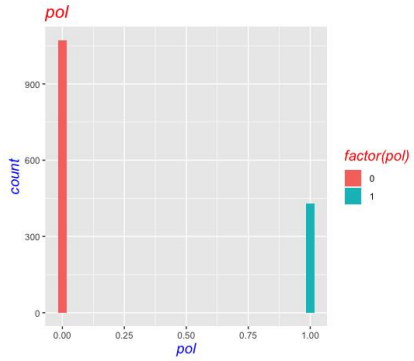
Outliers (outliers) in the box diagram are 2 and 5.

sex:



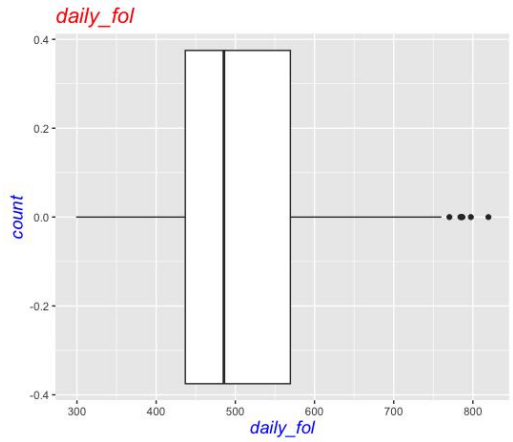
there are more females, no outliers

pol:



there are more people living below the poverty line,no outliers

daily_pol:

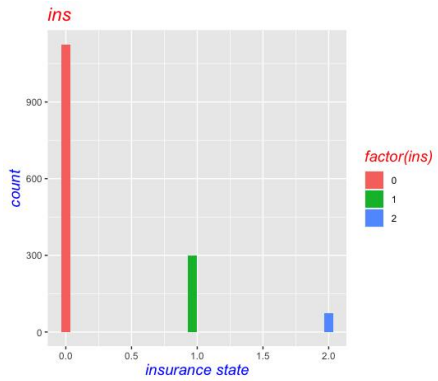


Daily total folate intake is concentrated between 450 and 550.

The median of the health_lit is 500.

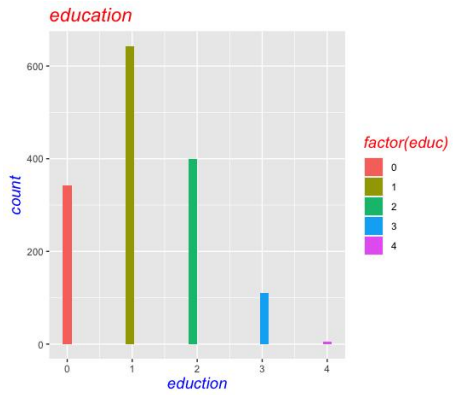
Outliers (outliers) in the box diagram are 750、770、800 and 830.

ins:



the number of public insurance is the most, private insurance is the middle, uninsured is the least

educ:

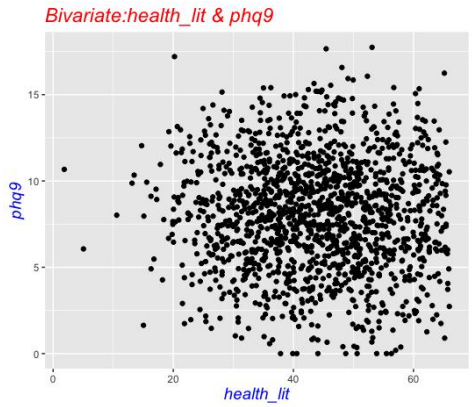


elementary school education (0), high school graduate (1), some college (2), college degree (3), and graduate degree (4)

the number of high shcool graduate is the most, graduare degree is the least.

iii. Bivariate plot of the variable with your chosen health outcome. Do these appear to be associated? Do you see any differences in the distribution of health outcome by these other variables?

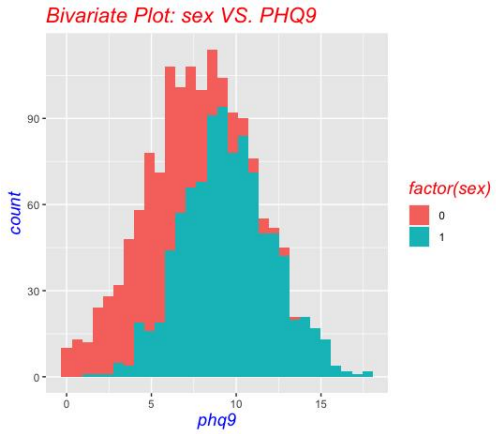
Bivariate:health_lit & phq9:



Associated: the points have a pattern,concentrate in 30-50 of health_lit and 5-10 of phq9, so there are an association between the variables.

there are noticeable differences in the distribution of phq9 at different levels of health_lit, in 30-50 of health_lit, there are more phq9, in 0-20 and 60-70 of health_lit, there are less phq9.

Bivariate:sex & phq9:

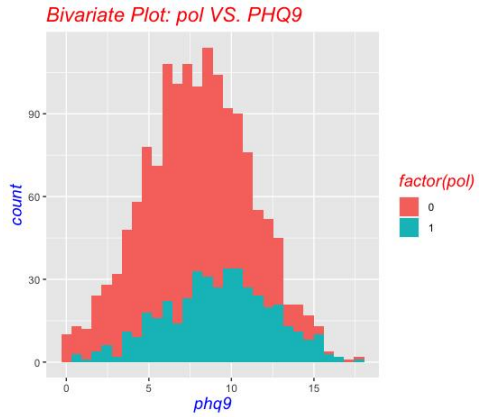


the distribution of male&phq9 is left skewed,

the distribution of female&phq9 is normal

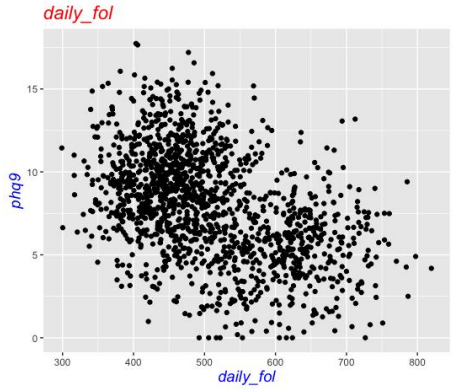
there are more phq9 in male

Bivariate:pol & phq9:



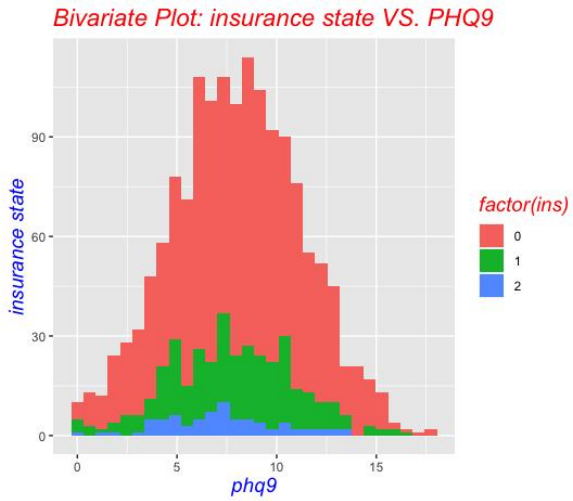
it is normal distribution, there are more people below the poverty line when phq9 is 5-10, they are asoociated, both of them have the most number in 9 of phq9 and decrease on two side.

Bivariate: daily_fol & phq9:



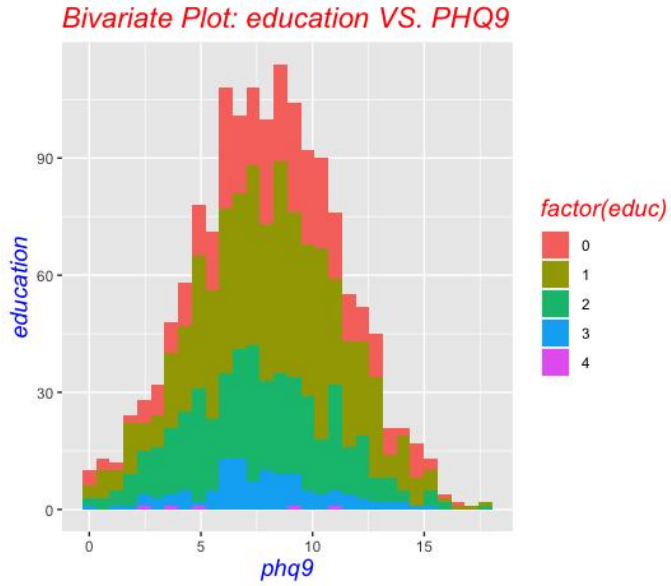
daily_fol and phq9 have a negative associated, more daily_fol, less phq9

Bivariate: ins & phq9:



the distribution are normal, the number of public insurance is the most, private insurance is the middle, uninsured is the least. All of insurance state are highest when phq9 is 8, and decrease both side.

Bivariate: educ & phq9:



elementary school education (0), high school graduate (1), some college (2), college degree (3), and graduate degree (4)

the distribution are normal, the number of elementary shcool education is the most, graduare degree is the least. All of number of education are highest when phq9 is 8, and decrease both side.

The distribution of health outcome by these other variables are similar, All of number of are highest when phq9 is 8, and decrease both side. there are also some difference in daily_fol, it is negative associated.

III. **Results**

– This section will have 2 tables with discussions following each one

a. Table 1 – Univariate and Bivariate Summaries (Follow the example given below)
    i. Using the binary category you constructed in Section I, create a table which provides numerical summaries of all covariates: first overall, and then by your binary health literacy group.
    ii. Then, test differences by the binary outcome, making sure to use the appropriate statistical test. Report the corresponding p-value in the table.
    iii. In a paragraph, describe the total sample. Then for each statistical test performed, please briefly describe the test (what type of test, assumptions, and conclusion). You do not need to calculate the test-statistic or p-value by hand.

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

	Total (N= 1500) Percent or Mean (SD)	High Health Literacy (N= 702) Percent or Mean (SD)	Low Health Literacy (N= 798) Percent or Mean (SD)	p-value
Characteristics				
Sex				
Male	(N=559)37.27%	(N=260)37.04%	(N=299)37.47%	0.863
Female	(N=941)62.73%	(N=442)62.96%	(N=499)62.53%	0.863
Below the Poverty Line				
No	(N=428)28.53%	(N=195)27.78%	(N=233)29.20%	0.5433
Yes	(N=1072)71.47%	(N=507)72.22%	(N=565)70.80%	0.5433
Daily Folate	506.79 (96.13)	505.60(94.30)	507.85(97.77)	0.6509
Insurance				
Public	(N=1125)75%	(N=527)75.07%	(N=598)74.94%	0.8731
Private	(N=300)20%	(N=142)20.23%	(N=158)19.80%	0.8731
Uninsured	(N=75)5%	(N=33)4.7%	(N=42)5.26%	0.8731
Education				
Elementary school	(N=342)22.80%	(N=143)20.37%	(N=199)24.94%	0.1474
High School	(N=643)42.87%	(N=310)44.16%	(N=333)41.73%	0.1474

	Some College	(N=399)26.60%	(N=189)26.92%	(N=210)26.32%	0.1474
	College degree	(N=111)7.4%	(N=56)7.98%	(N=55)6.89%	0.1474
	Graduate	(N=5)0.33%	(N=4)0.57%	(N=1)0.13%	0.1474
Degree					
	Health Outcome	7.99(3.25)	7.93(3.34)	8.04(3.18)	0.5263

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

The total sample size is 1,500 individuals which are randomly selected to participate. However, the table presents characteristics of the sample divided into two groups: high health literacy and low health literacy. The characteristics include sex, below the poverty line, daily folate intake, insurance status, education level, and health outcome.

For the statistical tests performed, the type of test, assumptions, and conclusion are as follows: Chi-square is used for categorical plus categorical, and two-sample t test is used for continuous plus categoricl variable, categorical variable use percent, continuous variable use mean or sd.

1. Sex: The test performed is a chi-square test to compare the distribution of sex between the high health literacy group and the low health literacy group. The assumptions for this test include having independent observations and a sufficient sample size. The null hypothesis is that there is no association between health literacy and sex. The p-value for this test is 0.863, indicating that there is no significant difference in the distribution of sex between the two groups.

2. Below the Poverty Line: The test performed is also a chi-square test to compare the distribution of being below the poverty line between the high health literacy group and the low health literacy group. The assumptions for this test are the same as the previous test. The null hypothesis is that there is no association between health literacy and being below the poverty line. The p-value for this test is 0.5433, indicating that there is no significant difference in the distribution of being below the poverty line between the two groups.

3. Daily Folate: The third statistical test performed is a t-test to compare the mean daily folate intake between the high health literacy group and the low health literacy group. The assumptions for this test include having independent observations, normally distributed data, and equal variances between the two groups. The null hypothesis is that there is no difference in the mean daily folate intake between the two groups. The p-value for this test is 0.6509, indicating that there is no significant difference in the mean daily folate intake between the two groups.

4.Insurance: The fourth statistical test performed is another chi-square test to compare the distribution of insurance type between the high health literacy group and the low health literacy group. The assumptions for this test are the same as the previous chi-square tests. The null hypothesis is that there is no association between health literacy and insurance type. The p-value for this test is 0.8731, indicating that there is no significant difference in the distribution of insurance type between the two groups.

5. Education: The fifth statistical test performed is a chi-square test to compare the distribution of education level between the high health literacy group and the low health literacy group. The

assumptions for this test are the same as the previous chi-square tests. The null hypothesis is that there is no association between health literacy and education level. The p-value for this test is 0.1474, indicating that there is no significant difference in the distribution of education level between the two groups.

6. Health Outcome: The final statistical test performed is a t-test to compare the mean health outcome score between the high health literacy group and the low health literacy group. The assumptions for this test include having independent observations, normally distributed data, and equal variances between the two groups. The null hypothesis is that there is no difference in the mean health outcome score between the two groups. The p-value for this test is 0.5263, indicating that there is no significant difference in the mean health outcome score between the two groups.

b. Table 2 – Results of SLR

   i. Present the results of your simple linear regression corresponding to your research question in Section I. (e.g. predict your health outcome using health literacy)

The results of the simple linear regression model predicting the health outcome using health literacy are as follows:

- Intercept: The intercept coefficient is 8.3928 with a standard error of 0.3325. This means that when health literacy is zero, the predicted health outcome is 8.3928.

- Health Literacy: The coefficient for health literacy is -0.0092 with a standard error of 0.0074. This means that for every one unit increase in health literacy, the predicted health outcome decreases by 0.0092 units.

formula: ^health outcome=8.3928 - 0.0092*health literacy

more details:

Residuals:

Min	1Q	Median	3Q	Max
-8.0449	-2.1875	-0.0169	2.3012	9.8335

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.392782	0.332502	25.241	<2e-16 ***
health_lit	-0.009187	0.007366	-1.247	0.213

Residual standard error: 3.252 on 1498 degrees of freedom

Multiple R-squared: 0.001037, Adjusted R-squared: 0.0003704

F-statistic: 1.555 on 1 and 1498 DF, p-value: 0.2125

ii. Provide a summary of Table 2. Specifically, interpret each coefficient. What is the R^2 value of this model and how do we interpret it? Does Table 2 support the differences you saw in Table 1? How about the plot you presented in table 1?

R^2 =0.010

Table 2. Linear Regression Model Predicting

Health Outcome (N =)

Coefficients	B (SE)
Intercept	8.3928(0.3325)
Health Literacy	-0.0092(0.0074)

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

- Intercept: The intercept coefficient is 8.3928 with a standard error of 0.3325. This means that when health literacy is zero, the predicted health outcome is 8.3928.

- Health Literacy: The coefficient for health literacy is -0.0092 with a standard error of 0.0074. This means that for every one unit increase in health literacy, the predicted health outcome decreases by 0.0092 units.

The R-squared value of this model is 0.001037, which indicates that only 0.1% of the variation in the health outcome can be explained by the variation in health literacy. This implies that health literacy alone is not a strong predictor of the health outcome.

Table 2 does not support the differences observed in Table 1. The p-value for the health literacy coefficient is 0.213, which is greater than the significance level of 0.05. This suggests that the coefficient is not statistically significant, meaning that there is no strong evidence to suggest a relationship between health literacy and the health outcome. However, in Table 2, the coefficient for health literacy suggests that higher health literacy is associated with a slightly lower health outcome.

The plot presented in Table 1 may have shown some differences, but the results of the regression analysis in Table 2 indicate that these differences are not statistically significant.

IV. Conclusion

a. Summarize your findings on the relationship between your health outcome and the health literacy measure.
b. Choose 1-3 plots that you created in the report that you believe are the most important to the explaining your research question. Explain how and why you chose these plots.
c. Based on Table 1, comment on which other variables may be important to include in a future regression analysis to account for additional variability in your health outcome.
d. What other variables would you like to be collected to examine your health outcome?

a. The relationship between health outcome and health literacy measure is not statistically significant (p-value = 0.5263). This suggests that there is no significant difference in health outcome between individuals with high health literacy and those with low health literacy, what is more, the R-squared value of this model is 0.001037, which indicates that only 0.1% of the variation in the health outcome can be explained by the variation in health literacy. This implies that health literacy alone is not a strong predictor of the health outcome.

b. One of the most important plots in explaining the research question is the boxplot of the distribution of health outcome (phq9). This plot provides an overview of the spread and central tendency of the health outcome variable, allowing for easy comparison between different health literacy groups.

Another important plot is the bivariate plot comparing health literacy to the health outcome. This plot helps visualize the relationship between the two variables and identify any potential patterns or trends.

These plots were chosen because they directly relate to the research question of examining the relationship between health literacy and the health outcome. They provide visual representations of the data, allowing for a better understanding and interpretation of the results.

c. Based on Table 1, other variables that may be important to include in a future regression analysis to account for additional variability in the health outcome could be age, socioeconomic status. These variables may have an impact on both health literacy and health outcome, and including them in the analysis could help to better understand the relationship between health literacy and health outcome.

d. Other variables that I would like to be collected to examine the health outcome could include information on lifestyle factors such as diet, exercise, smoking status, and alcohol consumption. Additionally, it would be helpful to have data on any chronic health conditions or medications that individuals in the sample may have, as these factors could also influence health outcome.