Summary Statistics and Data Exploration

The dataset here comprimises 7,555 individuals aged between 25 and 64, of which only 18.2% reported visiting a doctor in January 2010. This is visualized in Figure 1 which shows that a substantial majority of respondents did not seek medical care during that month, furthermore, a key determinant appears to be private health insurance coverage, which 65.7% of individuals reported having. This is seen in figure 2 which shows that those with insurance were significantly more likely to visit a doctor than those without, highlighting the potential role of coverage in improving healthcare access.

When studying figure 3 we see a positive relationship between age and likelihood of doctor visits, with older individuals more likely to report having seen a doctor. Health status also plays a strong role, Figures 4 and 5 show that doctor visit rates increase substantially among those reporting poorer physical or mental health, peaking at over 40% for individuals in "poor" health. This aligns with expectations, as individuals in worse self-reported health typically require more medical attention. Exploring other demographics, we can see that Figure 6 shows some regional variation, with slightly higher doctor visit rates in the Northeast and Midwest compared to the South and West to add to this family size (Figure 7) reveals a non-linear pattern, with visit rates decreasing for larger families up to a certain size, possibly due to resource constraints, before spiking at extreme values, though these may reflect small sample sizes.

Income (Figure 8) also appears to matter in this case, doctor visit rates rise with income up to the $30–40 range before slightly leveling off, suggesting economic capacity contributes to healthcare-seeking behaviour. Race and gender disparities are evident in our data as well this is seen in Figure 9 which shows that white individuals are more likely to visit the doctor compared to other racial groups, and Figure 10 highlights that females have higher visit rates than males.

Overall, the descriptive analysis suggests that doctor visits are associated with multiple factors, especially insurance coverage, health status, age, and income, and indicates substantial variation across sociodemographic groups. These patterns will be explored in greater depth in my modelling section.

Model Proposition – Doctor Visits

To explain the likelihood that an individual visited a doctor in January 2010, I propose using a probit regression model, which as we learnt is suited for binary outcomes. In this case the dependent variable Doctorvisit equals 1 if the individual reported visiting a doctor, and 0 otherwise.

Model Specification

Let Y_i denote whether individual i visited a doctor. The model assumes there's an unobserved, continuous latent variable Y_i^* representing their tendency to seek care, modeled as:

$$Y_i^* = X_i'\beta + \varepsilon_i, \quad \text{where } \varepsilon_i \sim N(0,1)$$

The observed outcome is determined by whether this latent tendency exceeds zero:

$$Y_i = 1 \text{ if } Y_i^* > 0, \text{ else } 0$$

The probability that an individual visits a doctor is given by:

$$\Pr(Y_i = 1 \mid X_i) = \Phi(X_i'\beta)$$

Where Φ is the cumulative distribution function of the standard normal distribution, and X_i is a vector of individual characteristics.

The idea behind my proposed model is that each person has an internal tendency shaped by their circumstances to visit the doctor. That tendency isn't directly measurable, but we use observable characterestics such as the explanatory variables below to estimate it. If the tendency is high enough, we observe a doctor visit. The model allows us to see how things like insurance, income, and health status affect that probability.

Variable Selection and Reasoning

The variables I chose to include are informed by both the visual analysis in Question 1 and by economic and

behavioral logic as well as intuition, whether or not these variables are truly significant can be teted later on once the model is estimated. The explanatory variables chosen are as below:

- **Private Insurance**: Individuals with insurance were clearly more likely to visit the doctor. It likely lowers financial barriers, so I expect a strong positive effect.
- **Age**: Older individuals were more likely to visit a doctor, likely due to increased health needs. A positive effect is expected.
- **Gender (Male)**: Males were less likely to visit a doctor. I include a male dummy variable and expect a negative effect.
- **Income**: Higher-income individuals showed higher visit rates, likely due to better access and affordability. A positive effect is expected.
- **Education**: Education is often linked to health awareness. I include it with an expected positive impact.
- **Family Size**: Likely that higher family sizes means higher visits in terms of pediatrician visits etc. I expect a positive affect
- **Married**: May reflect social support and higher health motivation. Expected effect is positive.
- **MSA (Urban residence)**: Urban dwellers may have greater access to healthcare. Expected effect is positive.
- **Race**: Racial differences in visit rates warrant inclusion of a categorical Race variable (White, Black, Asian, Other).
- **Physical Health**: Individuals with poorer physical health were more likely to visit a doctor. I include dummies from "very good" to "poor", with "excellent" as the reference. A positive gradient is expected.
- **Mental Health**: Same trend as physical health; dummies included for all levels below "excellent". Expected effect is positive.

Variables Not Included

Some variables were excluded either due to weak visual evidence, low frequency, or lack of theoretical justification:
- **Region**: Small regional differences observed as seen in the graph in the appendix, but not enough to justify inclusion in the base model.
- **phpoor / mhpoor**: Very few observations recorded in our data set; could produce unstable estimates.
- **Workplace variable**s: Focus is on individual-level predictors, so variables like sick leave policy and employer size are omitted.

Estimating the Model and Interpreting Results

To examine the determinants of doctor visits, I estimated the probit model proposed in Question 2 using the set of explanatory variables that included demographic, economic, and health-related characteristics as seen above. The dependent variable is binary and equals 1 if the individual visited a doctor in January 2010, and 0 otherwise. The probit model is ideal for this setting as it accounts for the binary nature of the response while allowing for a non-linear relationship between predictors and the probability of visiting a doctor.

The overall model is statistically significant. A likelihood ratio test(see appendix table.4) comparing the full model to a null model yielded a chi-square value of 487.7 with 18 degrees of freedom ($p < 0.001$), indicating a strong improvement over the null. The McFadden pseudo R^2 was 0.068, a respectable value for health behavior models. Multiple variables were statistically significant at the 5% level: private insurance, age, male, education, income, family size, married status, race (Black and White), and all levels of deteriorating physical health. These results align well with theoretical expectations and the trends observed in the data visualizations. Interestigly, none of the mental health indicators (mhverygood, mhgood, mhfair) were statistically significant. This may suggest that self-reported mental health is not a primary driver of doctor visits in the short term which makes sense intuitivly, or that mental health needs may lead to non-physical or specialist services not captured by the dependent variable.

The AMEs as seen in Table.1 in the appendix provide insights into the real-world impact of each variable. Physical health is by far the most important predictor: being in 'poor' physical health increases the probability of visiting a doctor by nearly 25 percentage points compared to being in 'excellent' health. 'Fair' and 'good' health also show substantial effects. Gender and insurance are the next most influential factors, males are about 9.6 percentage points less likely to visit, while those with private insurance are 4.3 points more likely. Socioeconomic factors such as education and income have smaller but meaningful effects, while being married or identifying as Black or White is also associated with increased likelihood of visiting a doctor. Mental health, while included in the model, showed no practical or statistical significance.

Overall, the probit model confirms that doctor visits are strongly influenced by insurance status, physical health, and demographic factors like gender and age. These findings reinforce earlier visual observations and provide both statistically significant and practically meaningful evidence about the key drivers of health-seeking behavior.

Effect of Health Insurance and Other Explanatory Variables

This section investigates how health insurance and other personal characteristics influence the probability of visiting a doctor. Using the estimated probit model, we calculate average marginal effects (AMEs), which are used here to approximate what the question refers to as average partal effects (APEs). These AMEs represent the average change in predicted probability of a doctor visit in response to a one-unit change in each explanatory variable, holding other factors constant.

Having private health insurance was one of the most influential explanatory variables, as reported in Table 5 in the appendix, the average marginal effect of insurance is approximately 4.29 percentage points and this estimate is highly statistically significant. This indicates that, all else equal, an insured individual is around 4.3% more likely to have visited a doctor compared to someone without insurance. This is further supported by Appendix Figure 11, which visually shows that those with insurance have a much higher predicted probability of visiting a doctor than those without.

Among all variables in the model, self-reported physical health had the strongest influence on predicted probabilities. Compared to those in excellent health, individuals in poor physical health had a 24.98 percentage point higher probability of visiting the doctor. Those in fair and good health also had significantly higher probabilities of doctor visits, as shown in Table A1 and visually confirmed in Appendix Figure 13. The probability increases in a clear and consistent gradient from excellent to poor health.

Gender had a substantial effect: being male was associated with a 9.62 percentage point decrease in the likelihood of visiting a doctor. This gender gap is consistently visible across racial subgroups in Appendix Figure A2. Age also showed a positive but smakler effect, with each additional year of age increasing the probability by 0.17 percentage points. Education also had a meaningful effect as well with each additional year of education increasing the predicted probability by 1.47 percentage points. Appendix Figure 15 illustrates that individuals with a bachelor's degree or higher had a much higher probability of visiting the doctor than those with less than a high school education.

Race and marital status were also important and significant, Black and White individuals were significantly more likely to visit a doctor than those identifying with other racial categories, to add to this being married increased the probability by approximately 4.17 percentage points. These effects are both statistically significant (Table A1) and visually supported in Appendix Figure 12.

Income had a positive but smaller marginal effect, with a one-dollar increase in hourly income increasing the visit probability by 0.09 percentage points. However, Appendix Figure 16 reveals an important interaction, the benefit of having insurance is more pronounced at lower income levels. Among those earning under $20/hour, having insurance increases the likelihood of a doctor visit far more than among higher-income groups. This finding suggests that expanding insurance access to lower-income populations may be especially effective in promoting healthcare utilization. Appendix Figure 14 further explores the role of age by grouping individuals into decade-based categories. The predicted probability of visiting the doctor steadily increases from ages 25–34 to 55–64. This pattern reflects the growing health needs and risk perceptions associated with aging, reinforcing the importance of age as a determinant of health service utilization.

In summary, private insurance, poor physical health, gender, education, and age are all strong and significant predictors of doctor visits. The results show that health-seeking behavior is shaped by both access (e.g., insurance and income) and need (e.g., age and health status). Visual toolsuch as the plots in Appendix Figures 11–16 provide additional insight into how predicted probabilities differ across subgroups, making the effects more tangible.

Predicting the Outcome and Evaluating Model Performance

To evaluate the performance of the estimated probit model, we first created a binary classification variable using a 0.5 threshold on the predicted probabilities. Individuals with predicted probabilities greater than 0.5 were classified as likely to have visited a doctor (predicted = 1), and those below this threshold were predicted not to have visited (predicted = 0). The results for the overall model as well as individual variable prediction accuracies are in the appendix labeled "Accuracy measure Table.7"

The overall matrix reveals that the model correctly identifies most individuals who did not visit a doctor (True Negatives = 6158), but severely underpredicts doctor visits (True Positives = 24). The number of False Negatives is 1351, indicating that many actual doctor visits were not detected by the model. This highlights a significant imbalance showing that while overall accuracy is high, the model's sensitivity is very low.

The model's overall prediction accuracy is 81.83%. However, this high score is driven primarily by the model's ability to predict the majority class being those who did not visit a doctor. Given that over 80% of the sample did not visit the

doctor in January 2010, even a simplistic model that predicts no visits for everyone would achieve high accuracy. Thus, overall accuracy can be misleading without examining performance across relevant subgroups.

To better understand where the model performs well or poorly, I analyzed prediction accuracy across several key demographic and health-related groups. These group-level accuracies are summarized in Accuracy measure Table 6 in the appendix. Several patterns emerge with the most interesing ones being:

• **Insurance Status**: Prediction accuracy is higher for individuals without insurance (85.93%) than those with insurance (79.68%). This likely reflects the lower variance and overall lower doctor visit rate in the uninsured group, making their behavior easier to classify.

• **Gender**: The model predicts male outcomes more accurately (86.92%) than female outcomes (76.73%). This again reflects a general tendency to predict the majority (non-visiting) class which is more prevalent among men.

• **Physical Health**: Prediction accuracy declines as self-reported physical health worsens. For example, the model correctly predicts 87.32% of cases in excellent health, but only 62.35% for those in poor health. This is a critical weaknes as individuals in poor health are more likely to visit a doctor—and failing to identify them is problematic from a policy and clinical perspective.

• **Education**: Prediction accuracy declines with increasing education. Individuals with less than a high school education are correctly predicted 89.15% of the time, while the rate falls to 74.69% for those with a Bachelor's degree or higher. This suggests more educated individuals may have less predictable health behaviors potentially due to greater autonomy or access to varied care options as their degree of education increases

In conclusion, the model performs well in predicting the majority class (no doctor visit), achieving a high overall accuracy of 81.83%. However, it performs poorly in identifying individuals who actually visited a doctor, as evidenced by the low number of true positives and the high number of false negatives. The model's performance also varies substantially across subgroups. It is most accurate among individuals who are male, less educated, uninsured, in good health, or of younger age and is is least accurate among those who are female, in poor health, older, or highly educated. These differences highlight the importance of evaluating predictive models not only by overall metrics but also by subgroup performance, particularly in health researchw where disparities matter.

Alternative Model and Performance Comparison

To evaluate whether an alternative specification improves model performance, I estimated a new model using a logistic regression (logit) instead of the original probit model. Additionally, an interaction term between private insurance and gender was added to investigate whether insurance coverage influences male and female behavior differently.

The alternative model is specified as follows:

$$Doctorvisit \sim Privateins + age + male + educ + income + famsize + MSA + married + Race + phverygood + phgood + phfair + phpoor + mhverygood + mhgood + mhfair + Privateins:male$$

The model performance metrics for both specifications are summarized in the appendix under Table.9

Confusion matrices as found in the Appendix (Table.8 Logit V Probit measures) for both models highlight the same critical issue both models achieve high overall accuracy but perform poorly in identifying actual doctor visits. The confusion matrices reveal that both models are highly affective at predicting non-visits (specificity > 99%) but fail to capture the minority class actual doctor visits. The logit model improves slightly on sensitivity (2.11% vs. 1.74%) but not enough to justify adopting the model given other metrics.

Although the logit model included a meaningful interaction term and employed a different link function, it did not yield any substantial improvement. The AIC is slightly higher in the logit model, and its pseudo R^2 is marginally lower. Accuracy is nearly identical between the two models. This suggests that the change in functional form and specification did not improve fit or predictive capability.

In conclusion, the original probit model remains the preferred specification due to its slightly better AIC, higher pseudo R^2, and simpler functional form. The alternative model did not offer practical or statistical improvements in classification or fit.

Physical Health Measure Construction

In the MEPS dataset, physical health is reported using five mutually exclusive binary variables:

- phpoor, phfair, phgood, phverygood, and phexcellent.

These dummies indicate self-reported physical health levels ranging from "poor" to "excellent." Since these variables are **ordinal in nature**, it makes sense to consolidate them into a single **numerical measure** that preserves this ordering for easier analysis and modelling.

Methodology

To capture this ordering, I constructed a new variable, phscore, that assigns numeric values from 1 to 5 to each of the physical health categories as seen in the Appendix Table 10 (Dummy Variable Assignment), This transformation allows us to treat physical health as an ordinal response for subsequent modelling techniques (e.g., ordered logit or probit).

Proposed Model for Physical Health

The newly constructed variable phscore represents physical health on a **discrete ordinal scale** ranging from 1 (Poor) to 5 (Excellent). Each unit increase reflects a step up in perceived physical well-being. Since the categories are **ordered but not continuous**, treating phscore as a linear outcome (via OLS regression) would **violate the assumptions of homoscedastkicity and interval equality** between categories.

Therefore, we need a model that:

- Recognizes the **ordinal** nature of the outcome.
- Does **not assume equal distances** between physical health categories.
- Models the **cumulative probabilities** of being at or below a certain health level.

Model Choice: Ordered Probit Model

To analyze physical health, I propose using an Ordered Probit Model given the ordinal nature of the physical health variable. The model captures the ordered but not interval-scaled nature of the physical health status, which is categorized into five levels: Poor, Fair, Good, Very Good, and Excellent.

The underlying latent variable approach assumes that there exists an unobserved continuous variable, physical health status (PH^*), determined by individual characteristics, which can be represented mathematically as:

$$PH_i^* = \beta_0 + \beta_1 \text{Private Insurance}_i + \beta_2 \text{Age}_i + \beta_3 \text{Gender}_i + \beta_4 \text{Education}_i + \beta_5 \text{Income}_i + \beta_6 \text{Family Size}_i + \beta_7 \text{MSA}_i + \beta_8 \text{Married}_i + \gamma' \text{Race}_i + \epsilon_i$$

where:

- PH_i^* is the latent (unobserved) physical health status for individual ii.
- β_0 is the intercept term.
- $\beta_1, \beta_2, \ldots, \gamma'$ are the coefficients associated with each explanatory variable.
- ϵi represents the error term, assumed to follow a standard normal distribution.

The observed ordinal physical health categories are related to this latent variable by thresholds τ1<τ2<τ3<τ4 such that:

$$PHi = 1 \; if \; PHi * \leq \tau 1$$

$$2 \; if \; \tau 1 < PHi * \leq \tau 2$$

$$3 if \; \tau 2 < PHi * \leq \tau 3$$

$$4 \; if \; \tau 3 < PHi * \leq \tau 4$$

$$5 \; if \; PHi * > \tau 4$$

Intuition Behind the Choice:

The choice of the Ordered Probit Model is intuitive as it aligns perfectly with the ordinal nature of the dependent variable. Physical health categories, though ordered, do not have equal intervals, meaning that the distance between

Poor and Fair health may not necessarily equal the distance between Very Good and Excellent. Thus, using an Ordered Probit Model respects this structure and provides a more realistic and statistically sound framework for analysis. Also, this model allows me to estimate the impact of explanatory variables on the probability of falling into each health category clearly and intuitively.

Estimation Results and Interpretation of Ordered Probit Model

We estimate an ordered probit model to analyse the determinants of physical health. The dependent variable, physical health, is measured on an ordinal scale ranging from 1 (Poor) to 5 (Excellent). The results from our estimation are presented in Table 11 in the apendix.

Estimation Results

The model provides insights into the factors that significantly influence an individual's physical health:

- **Age** negatively impacts physical health significantly, with a coefficient of -0.0118, statistically significant at the 5% level (p-value < 0.001). This indicatethat as individuals grow older, their self-reported physical health tends to deteriorate.
- **Education (educ)** exhibits a strong positive relationship with physical health, with a coefficient of 0.0513, statistically significant at the 5% level. This demonstrates that higher educaton levels are associated with better perceived physical health, possibly due to increased health literacy and healthier lifestyle choices.
- **Income** positively influences physical health, with a coefficient of 0.0067, statistically significant at the 5% level. Higher income provides better access to healthcare services, nutritious food, and safer living conditions, thereby enhancing physical health.
- **Family size (famsize)** also positively impacts physical health (coefficient: 0.0217, p-value < 0.05). A larger family size may provide stronger social support networks, contributing positively to individual physical health.

Other variables such as gender (male), marital status (married), private insurance coverage (Privateins), metropolitan residence (MSA), and racial categories (RaceBlack, RaceOther, RaceWhite) do not show statistically significant effects in this model at the 5% level.

Average Partial Effects (APE)

Table 12 in the appendix summarizes the average partial effects of each significant variable:

- **Age:** An additional year in age decreases the probability of reporting better physical health by approximately 0.03%, statistically significant at the 5% level.
- **Education:** Each additional year of education increases the probability of higher physical health rating by about 0.15%, highly significant at the 5% level.
- **Income:** An increase in income significantyly increases the probability of reporting better physical health, although the effect is relatively small (0.02%).
- **Family Size:** A larger family increases the probability of reporting better physical health by about 0.06%, significant at the 5% level.

Practical Significance and Insights

The analysis here reveals that age and education are the most impactful determinants of physical health with age negatively affecting it and education having a robust positive effect. Income, while significant, has a smaller magnitude of impact compared to education. These findings suggest policy interventions focusing on educational programs and health awareness could substantially enhance population-level health outcomes.

Model Prediction and Evaluation of Physical Health

Overall Predictive Performance

The ordered probit model was employed to predict individuals' physical health status. To assess the model's performance, a confusion matrix was generated, comparing the model’s predictions against actual observed physical health categories (See **Table 13/14** in Appendix).

The overall predictive accuracy was approximately **35.9%**, indicating that the model correctly classified roughly a third of individuals into their actual health categories. Given that there are five categories of physical health, this

accuracy is better than random chance (20%) however indicates some challenges in accurately distingiushing among categories.

Confusion Matrix Analysis

Analysis of the confusion matrix (**Table 13/14** in Appendix) reveals that the model struggled significantly in identifying the two lower categories ("Poor" and "Fair"), failing entirely to predict these categories explicitly. Predictions concentrated heavily on the middle ("Good") and upper ("Very Good" and "Excellent") categories.

Specifically:

- Category "Good" (3) was frequently predicted for individuals whose actual health ranged broadly from "Poor" to "Excellent," highlighting difficulty in precisely classifying individuals at the extremes of health status.
- The model predominantly classified individuals into "Very Good" (4) and "Excellent" (5) categories, potentially reflecting the distribution of actual health status in the dataset.

Predictive Accuracy across Groups of Interest

To understand if predictive accuracy varied systematically across different population segments the analysis further explored accuracy within subgroups defined by **education level** and **age group**:

- **Education Level** (**Table 14** in Appendix):
    - Individuals with **Bachelor or higher degrees** had the highest predictive accuracy at **40.0%**, likely due to greater consistency or clarity in self-reported health assessments in better-educated individuals.
    - Those with **less than high school education** had significantly lower accuracy (**34.4%**). This disparity suggests that individuals with lower education levels may have less predictable health outcomes or potentially greater variability in health perceptions or reporting accuracy.
- **Age Group** (**Table 14** in Appendix):
    - Predictive accuracy increased slightly with age, from **36.0%** for ages 25–34 to **37.2%** for ages 55–64, reflecting potentially more stable health patterns in older individuals. Younger cohorts may have more diverse or volatile health status, compliating accurate predictions.

Interpretation and Practical Implications

Overall while the ordered probit model provides useful insights, its predictive limitations especially at the lower end of the physical health spectrum highlights the complexity inherent in health assessments. The observed variation across education and age groups suggests potential avenues for refining the model or targeted policy interventions.

Part B – Analysis of Mental Health

Proposing a Measure for Mental Health

To effectively capture the mental health infomation encoded in the five dummy variables (excellent, very good, good, fair, poor), I propose an **ordinal mental health index**, ranging from 1 (Poor) to 5 (Excellent). This approach ensures we retain the ordered nature of the health states allowing intuitive interpretation and allowing me to do subsequent modelling with ordinal regression techniques. Doing so in R and summarizing it allows us to see how populated each index is, this is seen in Table 15 in the appendix.

Proposed Model for Mental Health (Ordered Probit Model)

Intuition and Explanation of the Ordered Probit Model:

The Ordered Probit Model assumes there is an underlying, latent (unobserved) continuous measure of mental health, that determines the observed ordinal outcomes. We can formally represent this as:

$$MHi* = Xi\beta + ui, ui \mid Xi \sim N(0,1)$$

Where:

- MH_i^* is the latent mental health level for individual iii.
- X_i is a vector of explanatory variables (e.g., age, education, family size, income, marital status, private insurance, race).

- β represents the vector of parameters to estimate.
- u_i is the error term, assumed normally distributed with mean zero and unit variance.

While MH_i^* is unobserved, the ordered responses (Poor, Fair, Good, Very Good, Excellent) are observed according to thresholds (cut points), as follows:

$$MHi = \{1(Poor)ifMHi * \leq \alpha 1$$

$$2(Fair)if\alpha 1 < MHi * \leq \alpha 2$$

$$3(Good)if\alpha 2 < MHi * \leq \alpha 3$$

$$4(Very\ Good)if\alpha 3 < MHi * \leq \alpha 4$$

$$5(Excellent)ifMHi * > \alpha 4$$

The threshold parameters (α1,α2,α3,α4) are estimated from the data along with the regression parameters (β).

Explanatory Variables (Included in Model):

Based on intuitionan and previous findings, the following variables will be included in my model:

- **Age:** Older age may negatively impact mental health due to cumulative stress or declining health.
- **Education:** Higher educational qualifications likely positively impacts mental health through improved access to resources and coping mechanisms.
- **Family size:** Larger families may offer social support, potentially positively impacting mental health.
- **Income:** Higher income levels are likely to positively impact mental health by reducing financial stress and increasing access to healthcare.
- **Marital status (Married):** Married individuals my have better mental health due to emotional support.
- **Gender (Male):** Gender may impact mental health differently, with potential gender-specific stressors.
- **Race (Black, White, Other):** Differences in mental health may exist across racial groups due to various socio-economic or cultural factors.
- **Private Insurance:** Having insurance can positively impact mental health through improved access to mental healthcare.

Model Estimation and Interpretation

As seen in **Table 16** in the appendix, The ordered probit model estimation reveals several key predictors of self-reported mental health. Education emerges as a highly signifiant factor, with a coefficient of 0.0479 and an average partial effect (AME) of 0.0004 ($p < 0.001$). This indicates that each additional year of education increases the likelihood of reporting better mental health, likely due to increased health literacy, access to resources, and improved problem-solving skills, to add to this Family size also demonstrates a significant positive impact (coefficient: 0.0470, AME: 0.0003, $p < 0.001$), suggesting that individuals with larger families may benefit from greater social and emotional support netorks. Income shows a small but statistically significant positive relationship with mental health (coefficient: 0.0068, AME: 0.0001, $p = 0.002$), reinforcing the idea that financial stability and access to quality care contribute to better mental well-being. Additionally, RaceWhite shows a small positive and significant effect (AME: 0.0011, $p = 0.039$), indicating that White individuals in the sample are marginally more likely to report better mental health, possibly due to systemitc advantages or differential access to support structures.

In contrast variables such as male ($p = 0.076$), married ($p = 0.707$), Private Insurance ($p = 0.636$), and RaceBlack ($p = 0.136$) are not statistically significant at the 5% level. While RaceOther has a negative coefficient (−0.1349) and AME of 0.0027, its p-value of 0.013 suggests it is significant and indicates worse mental health outcomes for this group—likely reflecting underlying socio-cultural or systemic disparities.

In summary, education, family size, income, and race (White and Other) significantly influence mental health at the 5% level. Compared to physical health, whee age and education were key drivers, mental health appears more sensitive to socioeconomic and social support factors, underscoring the hard to predict nature of mental wllbeing.

Further analysis using visual statistics

Mental Health and Socioeconomic Characteristics

Figures 17 and 18 in the Appendix provide a visual representation of how mental health varies across education and income groups, respectively.

Figure 17: Mental Health Distribution by Education Level
This stacked bar chart shows the distribution of mental health scores across different education levels. We observe a clear positive gradient: as education level increases, the proportion of individuals reporting better mental health (scores of 4 = "Very Good" and 5 = "Excellent") also increases. For instance, individuals with a Bachelor's degree or higher show the highest consentration of excellent mental health ratings, whereas those with less than a high school education exhibit higher proportions of lower mental health scores. This visual evidence supports the regression finding that **education has a significant and positive impact on mental health** (AME = -0.0004, $p < 0.01$), suggesting that education enhances individuals' coping mechanisms and access to mental health resources.

Figure 18: Average Mental Health Score by Income Group
This bar chart illustrates that average mental health scores increase with income. Individuals in the lowest income group (0–10) report the lowest average mental health scores (just below 4), while those earning $40 or more per hour report scores closer to 4.5. This trend reflects the significant and positive effect of **income** on mental health found in the model (AME = 0.0004, $p = 0.0024$). Higher income levels likely reduce stress related to financial insecurity and improve access to healthcare and well-being services.

Together, these figures emphasize the **strong influence of socioeconomic factors on mental health**, reinforcing the statistical conclusions from the ordered probit model and demonstrating the value of targeted social and educational interventions to promote mental well-being.

Overall Model Accuracy

The confusion matrix (Appendix **Table.17**) reveals the distribution of predicted vs. actual mental health scores. The majority of predictions fall within categories 4 and 5 (Very Good and Excellent), which is consistent with the distribution of responses in the sample. The model here achieves an **overall accuracy of 43.5%**, while not perfect, is relatively reasonable given the complexity and ordinal nature of the outcome variable. This is a notable improvement over random guessing in a five-category model (baseline of 20%).

Accuracy Across Demographic Groups

To explore the model's performance acoss different population segments, prediction accuracy was calculated for each of the following:

- **Education Levels** (Appendix **Table.17**): Accuracy improves with educational attainment. Those with a *Bachlor's degree or higher* have the highest prediction accuracy (0.541), while individuals with less than a high school education show lower accuracy (0.348). This suggests the model performs better at predicting mental health among more educated individuals, possibly due to more consistent patterns in this group.
- **Income Groups** (Appendix **Table.17**): A similar trend is observed with income. The lowest income group (0–10) has the lowest accuracy (0.367), while the highest group (40+) achieves 0.557 accuracy. This supports the idea that socioeconomic indicators are strong predictors of mental health and are better captured by the model for higher-income individuals.
- **Age Groups** (Appendix **Table.17**): Prediction accuracy is highest among younger individuals (ages 25–34: 0.483), and decreases slightly for older groups (ages 55–64: 0.399). This may reflect greater variability in mental health patterns in older age groups which as a result makes prediction more difficult.

The model demonstrates moderate predictive power overall and performs significantly better for groups with higher education and income levels. This likely reflects more systematic patterns in mental health reporting within these groups, which the model can detect more effectively. However, the reduced accuracy among lower-income and less-educated groups may point to missing predictors that could be incorporated into future models.

Relationship Between Mental and Physical Health

To explore the interacion between mental and physical health, we use both graphical tools and statistical summaries based on the constructed ordinal measures for each dimension (ranging from 1 = Poor to 5 = Excellent).

Descriptive Graphical Analysis

Figure 19 shows the **stacked proportional bar chart** of physical health scores within each level of mental health. It reveals that as mental health improves, so does physical health. Among individuals with the lowest mental health score of 1 (Poor), nearly **70%** report a physical health score of 1 or 2, indicating poor physical condition. Conversely, among those with a mental health score of 5 (Excellent), over **50%** report a physical health score of 5, with a further **30%** reporting a score of 4. This visual gradient highlights a strong positive association.

Figure 20 presents **boxplots** of mental health across levels of physical health. The median mental health score rises steadily from **2 (Poor)** for those in the lowest physical health category to nearly **5 (Excellent)** in the highest. To add to this, the variability of mental health narrows significantly at higher physical health levels suggesting that excellent physical health is more consistently associated with excellent mental health.

Figure 21 displays a **heatmap** of joint proportions. The diagonal dominance (e.g., 65% of individuals with physical health score 3 also have mental health score 3) reinforces the aligned nature of mental and pbsical health. Notably, for those with the top mental health rating (5), **52%** simultaneously report excellent physical health. For those with mental health score 4, **62%** report physical health score 4, reflecting alignment in well-being.

Statistical Summary

To quantify this association, we calculate the **Pearson correlation coefficient** between the two scores (both treated as ordered numeric scales). The correlation value is:

$$r = 0.566$$

This indicates a **moderate-to-strong positive linear association** between mental and physical health, statistically confirming the trend observed in the plots.

Interpretation

These results demonstrate that **individuals with better mental health are substantially more likely to report better physical health**, and vice versa. The strength of the relationship is consistent across the visualizations and supported by the correlation metric. For example:

- Over **80%** of respondents with excellent mental health (score 5) reported physical health scores of 4 or 5.
- In contrast, nearly **75%** of individuals with poor or fair mental health (scores 1–2) fall into the lower physical health categories (1–3).
- The narrowing interquartile ranges in Figure 20 indicate that higher physical health is not only associated with better mental health on average but also less variability in mental well-being.

Modelling the Relationship Between Mental and Physical Health

Given the strong observed association between mental and physical health, an appropriate next step would be to jointly model the two outcomes allowing for correlation between their unobserved determinants. Both health outcomes are ordinal and measured on a five-point scale. Therefore, a natural modelling approach would be the **Bivariate Ordered Probit Model** (Greene, 2018). In this framework, two latent variables are defined, one for mental health and one for physical health, each modelled as a function of explanatory variables and random disturbances. The random errors are assumed to follow a bivariate normal distribution with a correlaton coefficient, capturing the fact that unobserved factors (such as personaliy traits, resilience, or social support) may simulteneously affect both health outcomes. The bivariate ordered probit model estimates the joint probability of observing different combinations of mental and physical health outcomes, accounting for the ordinal nature of the data and potential simultaneity. This approach would allow for a more nuanced understanding of how covariates influence each dimension separately while recognizing the possibility of interdependence between the two outcomes. We could also use something called the **Seemingly Unrelated Ordered Probit Models (SUOP)** could be employed if the primary interest lies in estimating two separate equations while allowing correlated errors (Cappellari and Jenkins, 2003). This method treats mental and physical health as distinct but related outcomes, linked through their error erms rather than explicit simultaneity.

Overall, the most direct and suitable approach would be the **Bivariate Ordered Probit Model**, due to its ability to respect the ordered nature of the data and capture correlation between unobserved influences on mental and physical health.

References

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley.

Cappellari, L., & Jenkins, S. P. (2003). Multivariate probit regression using simulated maximum likelihood. *The Stata Journal, 3*(3), 278–294.

Greene, W. H. (2018). *Econometric analysis* (8th ed.). Pearson.

Nelsen, R. B. (2006). *An introduction to copulas* (2nd ed.). Springer.

Appendix

Appendix A: Visual Summaries of Doctor Visit Patterns

Table 1. Summary Statistics

Variable	Mean	SD	Min	Max
Doctorvisit	0.182	0.3859	0	1
Privateins	0.6567	0.4749	0	1
Age	42.5906	10.5737	25	64
Male	0.4999	0.5	0	1
Educ	13.4332	2.8921	0	17

Famsize	2.9317	1.5952	1	14
Income	20.0944	11.9902	0.4	73.08
MSA	0.8683	0.3382	0	1
Married	0.7824	0.4126	0	1
White	0.7392	0.4391	0	1
Black	0.17	0.3756	0	1
Asian	0.0809	0.2727	0	1
Northeast	0.1407	0.3477	0	1
Midwest	0.2255	0.418	0	1
South	0.3693	0.4826	0	1
West	0.2645	0.4411	0	1
PH Excellent	0.2621	0.4398	0	1
PH Very Good	0.3464	0.4759	0	1
PH Good	0.2968	0.4569	0	1
PH Fair	0.0835	0.2767	0	1
PH Poor	0.0113	0.1055	0	1
MH Excellent	0.4328	0.4955	0	1
MH Very Good	0.3026	0.4594	0	1
MH Good	0.2287	0.42	0	1
MH Fair	0.0328	0.1782	0	1
MH Poor	0.003	0.0551	0	1

Table.2 Model estimation

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.660794   0.174463 -15.251  < 2e-16 ***
Privateins   0.174613   0.040349   4.328 1.51e-05 ***
age          0.007090   0.001829   3.876 0.000106 ***
male        -0.391532   0.036060 -10.858  < 2e-16 ***
educ         0.059838   0.007888   7.586 3.29e-14 ***
income       0.003813   0.001678   2.272 0.023100 *
famsize     -0.041622   0.012856  -3.238 0.001205 **
MSA          0.071363   0.052733   1.353 0.175959
married      0.169585   0.051016   3.324 0.000887 ***
RaceBlack    0.167396   0.079721   2.100 0.035749 *
RaceOther    0.300758   0.194321   1.548 0.121686
RaceWhite    0.320501   0.069632   4.603 4.17e-06 ***
phverygood   0.231401   0.052230   4.430 9.41e-06 ***
phgood       0.401168   0.056740   7.070 1.55e-12 ***
phfair       0.691365   0.074948   9.225  < 2e-16 ***
phpoor       1.016500   0.149801   6.786 1.16e-11 ***
mhverygood  -0.035731   0.045867  -0.779 0.435975
mhgood      -0.079477   0.052770  -1.506 0.132041
mhfair       0.034093   0.101365   0.336 0.736613
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table.3

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7168.3  on 7554  degrees of freedom
Residual deviance: 6680.6  on 7536  degrees of freedom
AIC: 6718.6

Number of Fisher Scoring iterations: 5

>
> # Pseudo R-squared
> pR2(probit_model)
fitting null model for pseudo-r2
          llh       llhNull            G2      McFadden          r2ML          r2CU
-3.340323e+03 -3.584173e+03  4.876997e+02  6.803518e-02  6.251379e-02  1.020129e-01
```

Table.4 LM measures

```
Likelihood ratio test

Model 1: Doctorvisit ~ Privateins + age + male + educ + income + famsize +
    MSA + married + Race + phverygood + phgood + phfair + phpoor +
    mhverygood + mhgood + mhfair
Model 2: Doctorvisit ~ 1
  #Df  LogLik  Df Chisq Pr(>Chisq)
1  19 -3340.3
2   1 -3584.2 -18 487.7  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1. Distribution of Doctor Visits

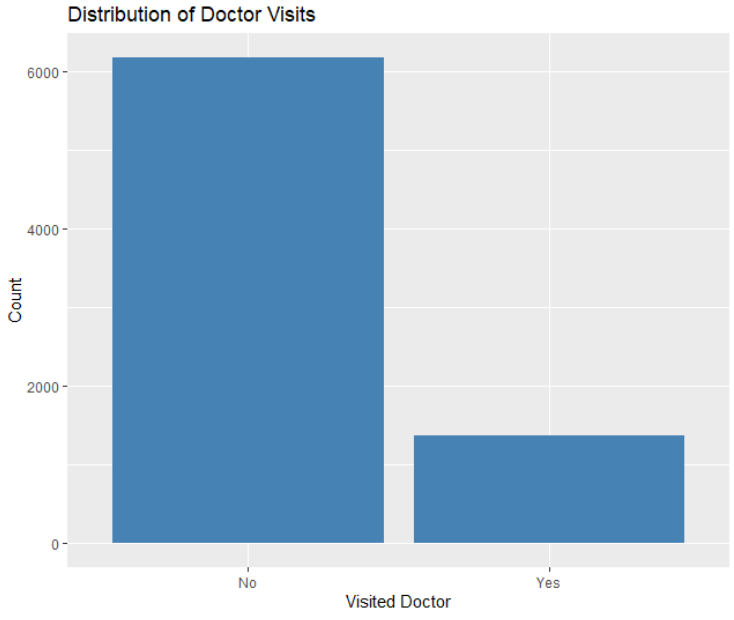

Figure 2. Doctor Visit Rate by Health Insurance

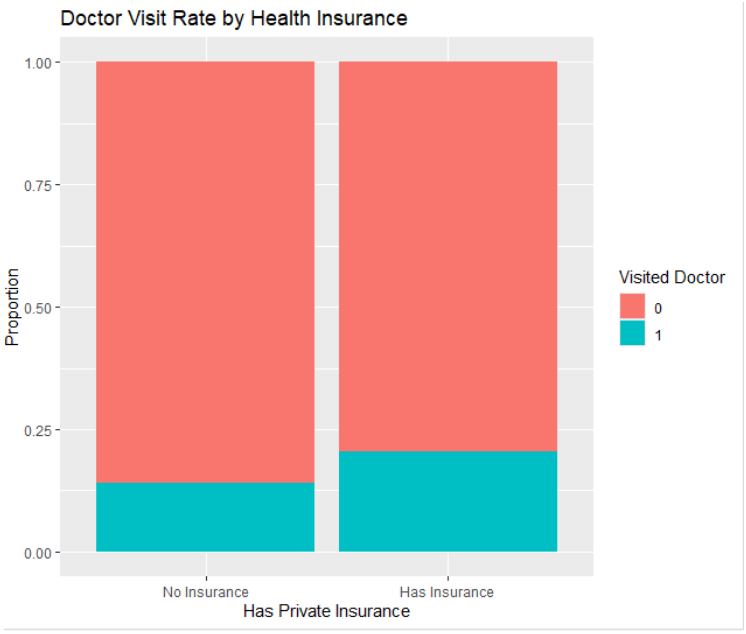

Figure 3. Age Distribution by Doctor Visit

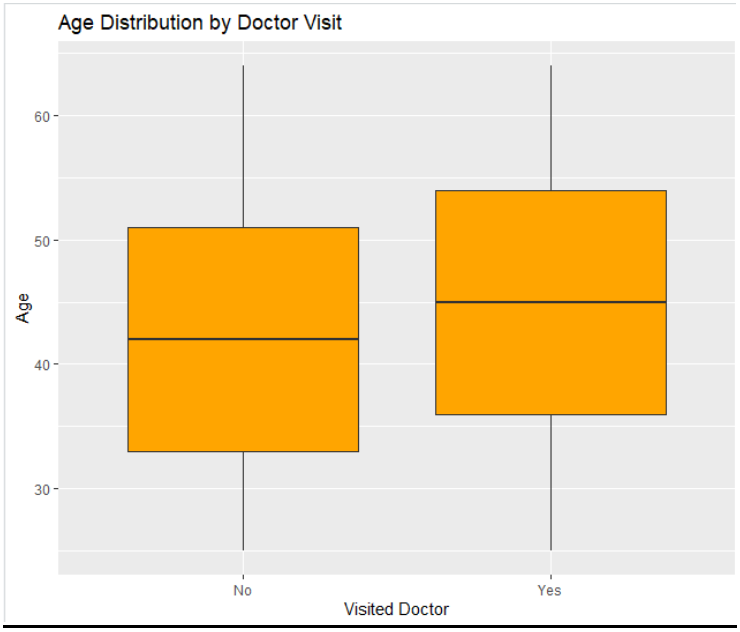

Figure 4. Doctor Visits by Physical Health Status

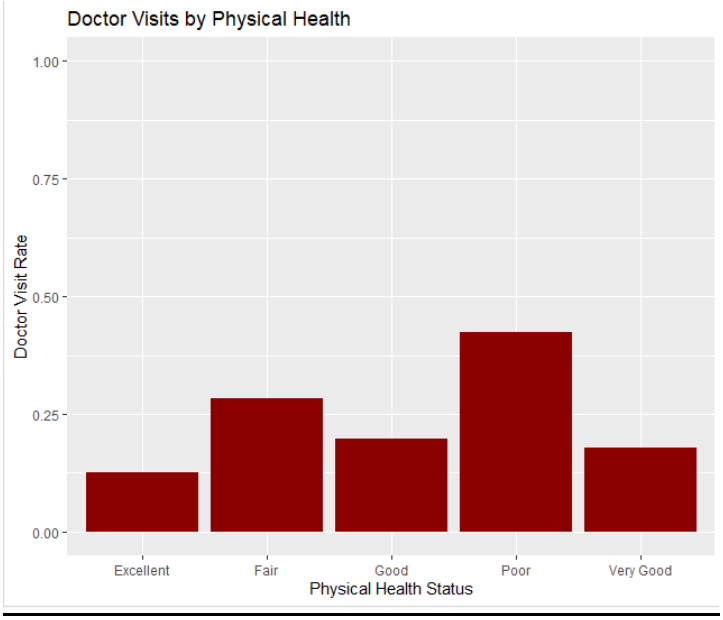

Figure 5. Doctor Visits by Mental Health Status

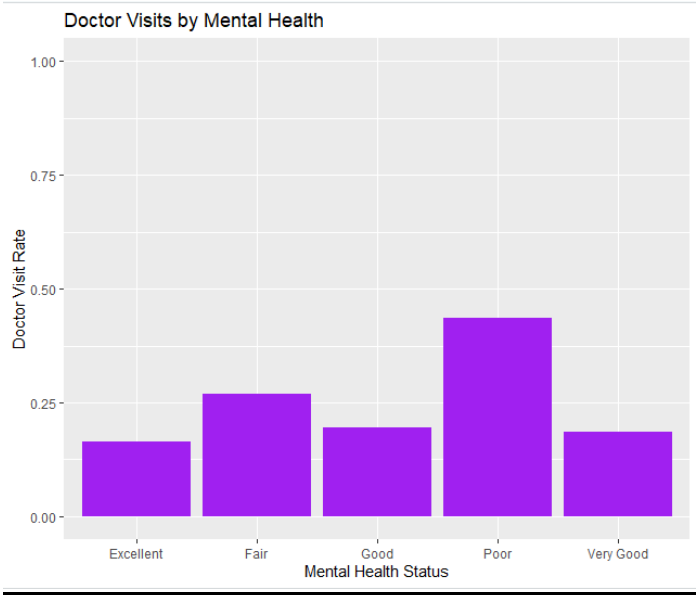

Figure 6. Doctor Visits by Region

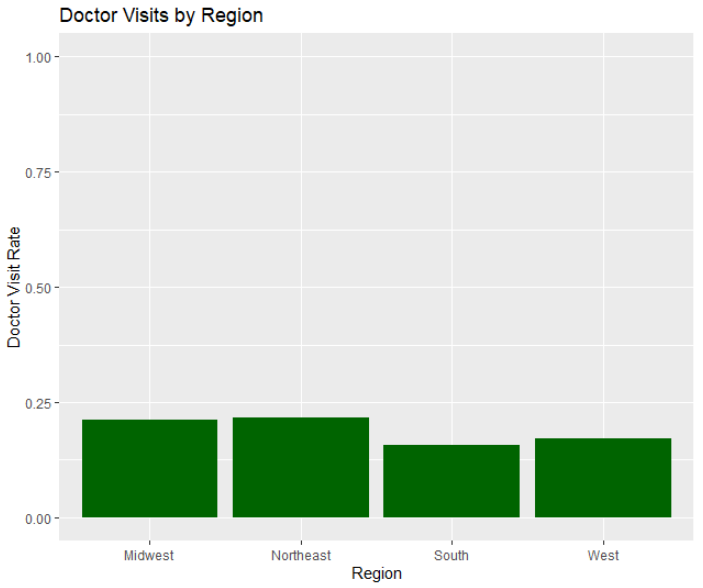

Figure 7. Doctor Visits by Family Size

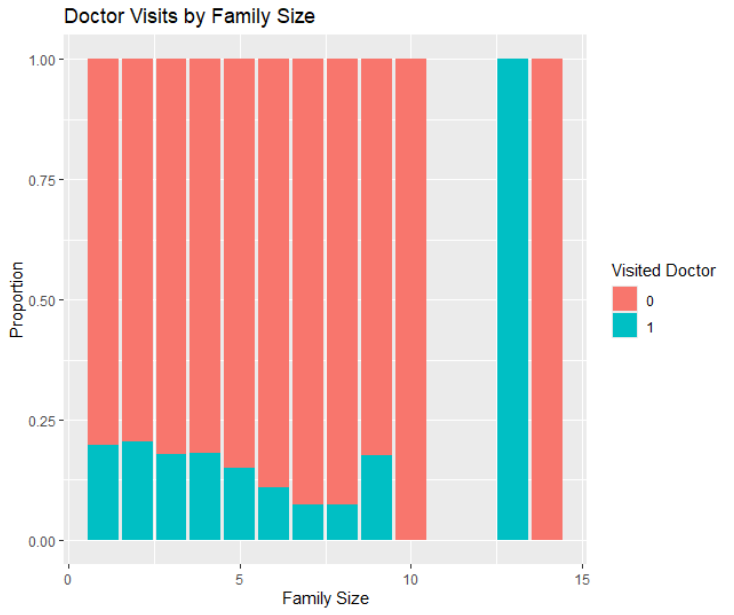

Figure 8. Doctor Visits by Income Group

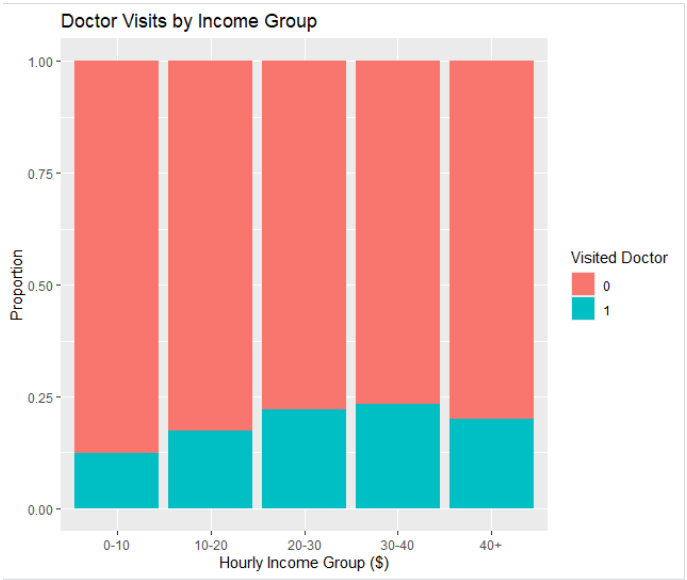

Figure 9. Doctor Visits by Race

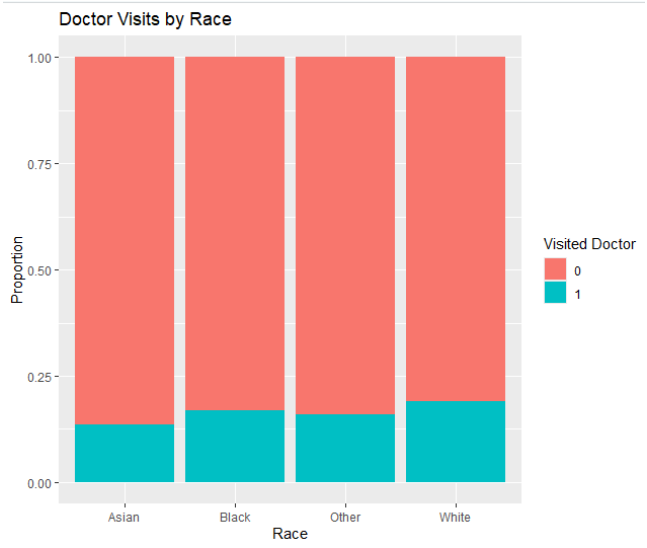

Figure 10. Doctor Visits by Gender

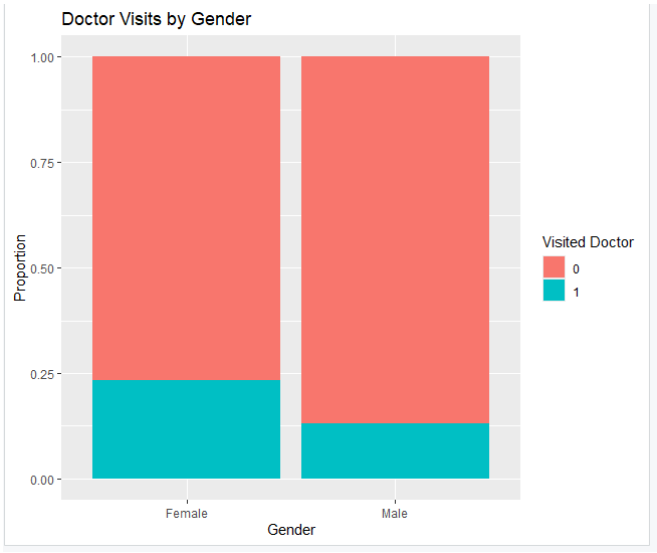

Table.5 AME of explanatory variables

Variable	AME	Std. Error	z-value	p-value	95% CI (Lower)	95% CI (Upper)
Private Insurance	0.0429	0.0099	4.34	<0.001	0.0235	0.0623
Age	0.0017	0.0004	3.88	<0.001	0.0009	0.0026
Male	-0.0962	0.0087	-11.02	<0.001	-0.1134	-0.0791
Education (years)	0.0147	0.0019	7.63	<0.001	0.0109	0.0185
Income (hourly)	0.0009	0.0004	2.27	0.023	0.0001	0.0017
Family Size	-0.0102	0.0032	-3.24	0.001	-0.0164	-0.004
Married	0.0417	0.0125	3.33	<0.001	0.0171	0.0662
Race: Black	0.0347	0.016	2.16	0.03	0.0033	0.0662
Race: White	0.0716	0.0137	5.23	<0.001	0.0448	0.0984
Physical: Very Good	0.0569	0.0128	4.44	<0.001	0.0318	0.082
Physical: Good	0.0986	0.0139	7.11	<0.001	0.0714	0.1258
Physical: Fair	0.1699	0.0182	9.33	<0.001	0.1342	0.2056
Physical: Poor	0.2498	0.0366	6.83	<0.001	0.1782	0.3215

Figure 11. Predicted Probability of Doctor Visit by Insurance Status

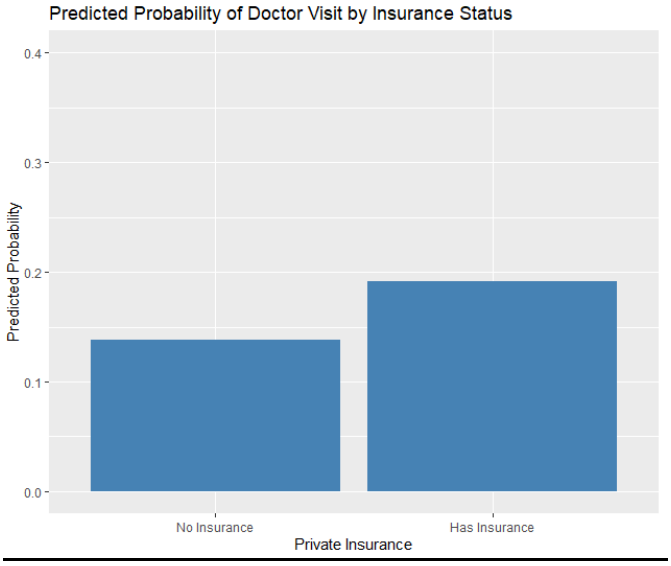

Figure 12. Predicted Probability of Doctor Visit by Race and Gender

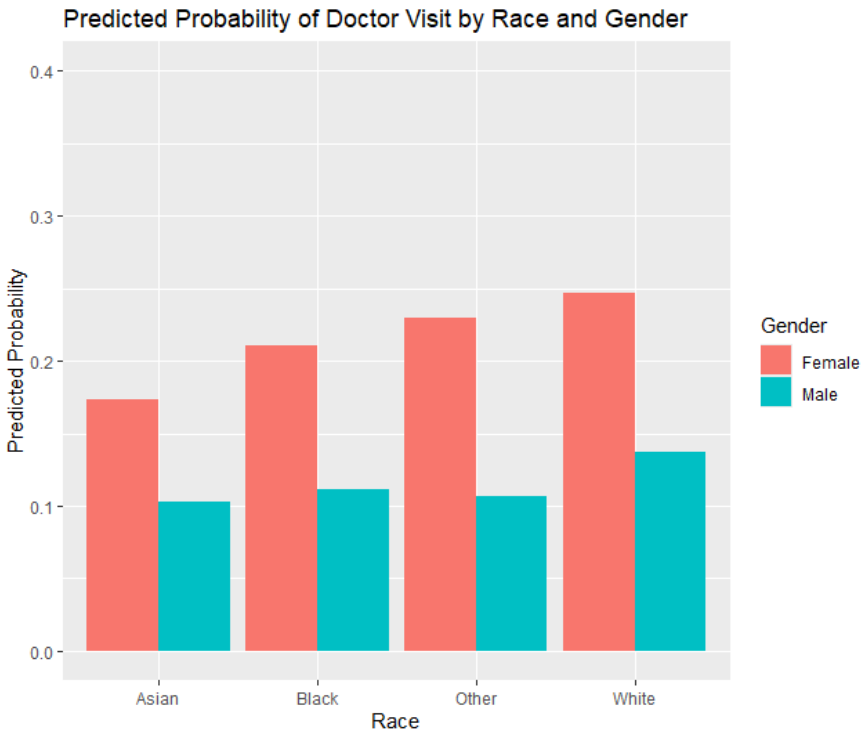

Figure 13. Predicted Probability of Doctor Visit by Physical Health Status

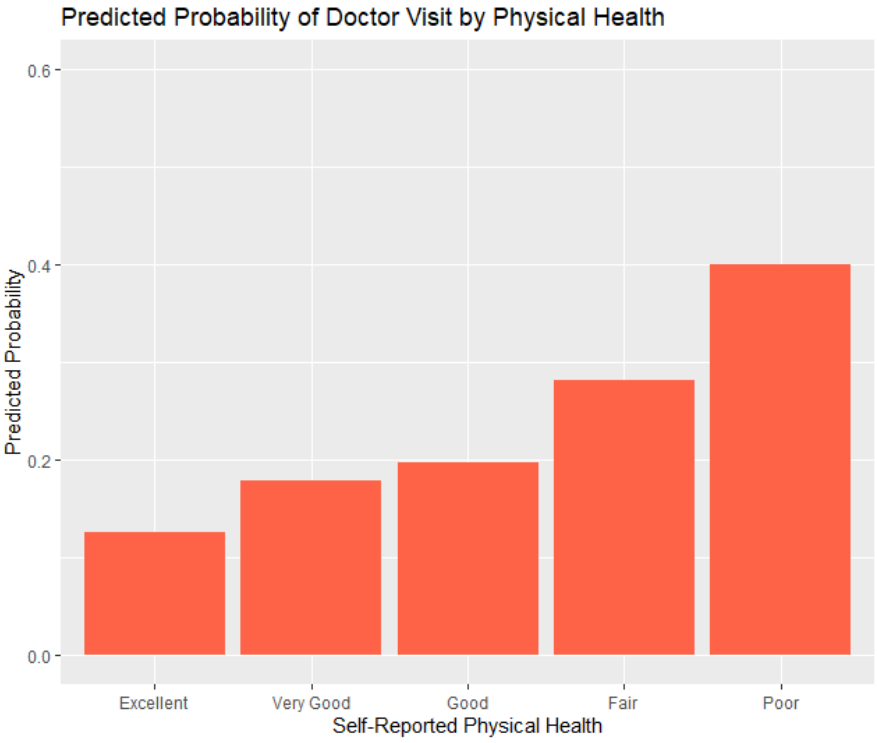

Figure 14. Predicted Probability of Doctor Visit by Age Group

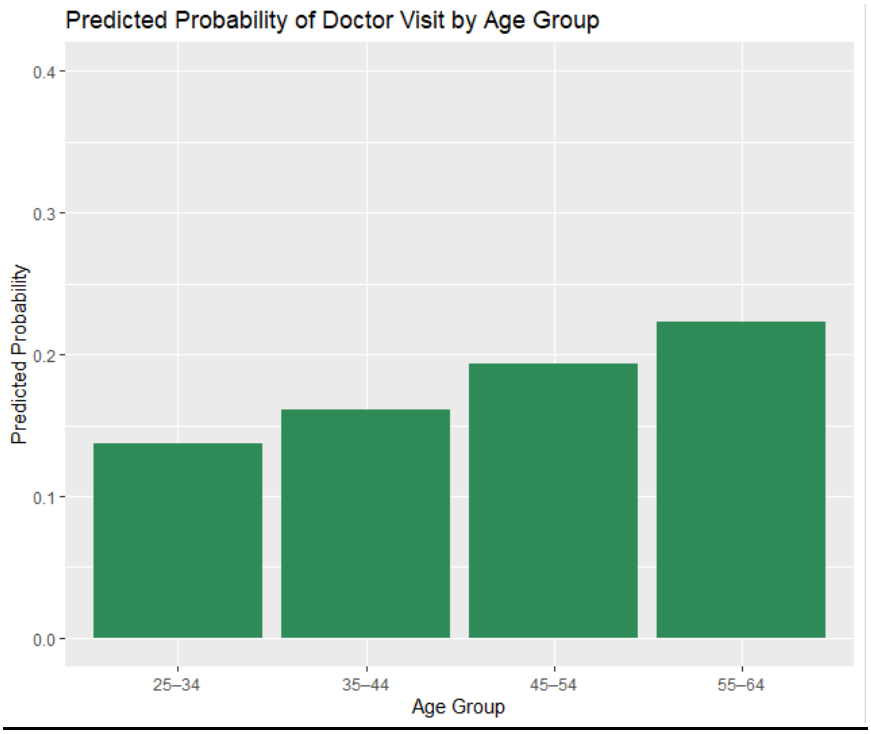

Figure 15. Predicted Probability of Doctor Visit by Education Level

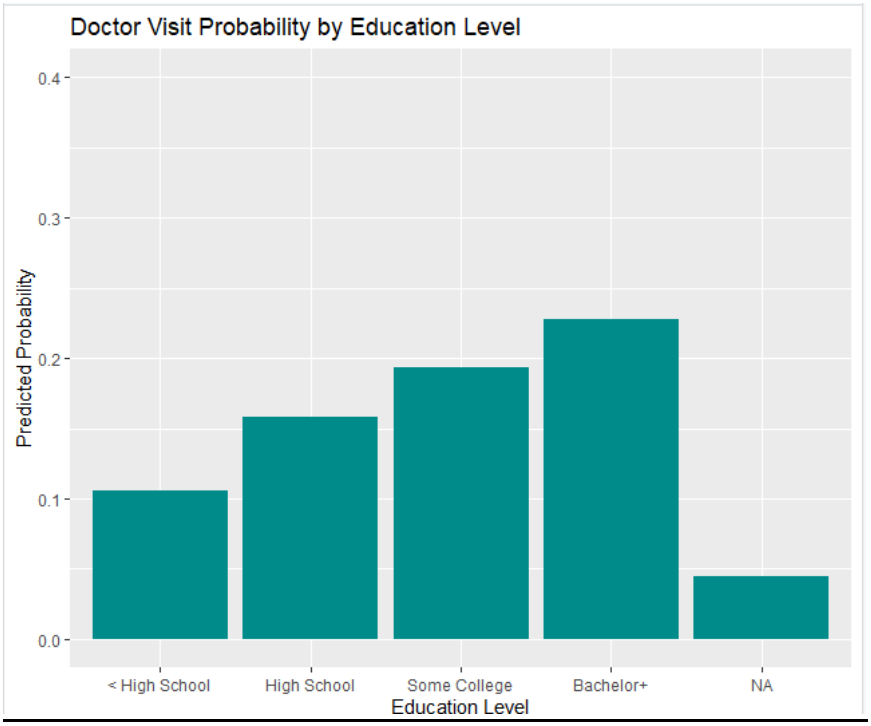

Figure 16. Predicted Probability of Doctor Visit by Income Group and Insurance Status

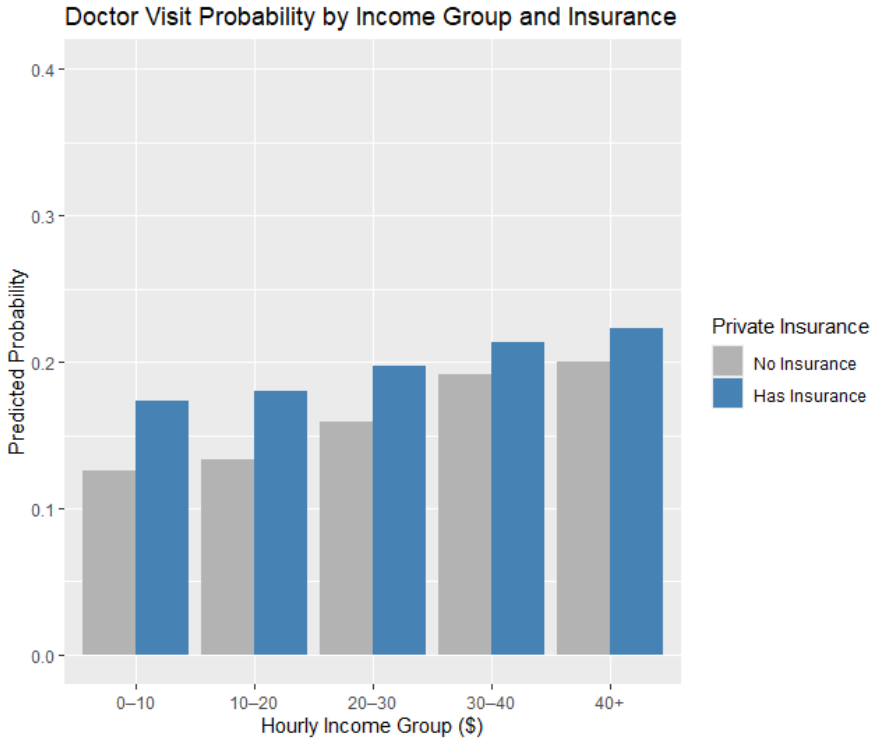

Table.6 Probit group accuracy measures

Group	Accuracy	Group Type
Overall	0.8183	<NA>
0	0.8593	Private Insurance
1	0.7968	Private Insurance
Female	0.7673	Gender
Male	0.8692	Gender
xcellent	0.8732	Physical Health
Fair	0.7132	Physical Health
Good	0.8029	Physical Health
Poor	0.6235	Physical Health
ery Good	0.8216	Physical Health
25-34	0.8547	Age Group
35-44	0.8278	Age Group
45-54	0.8178	Age Group
55-64	0.7424	Age Group
h School	0.8915	Education
h School	0.8294	Education
College	0.8056	Education
achelor+	0.7469	Education
Asian	0.8642	Race
Black	0.8287	Race
Other	0.8400	Race
White	0.8106	Race
0-10	0.8747	Income Group
10-20	0.8251	Income Group
20-30	0.7803	Income Group
30-40	0.7658	Income Group
40+	0.8010	Income Group

Table.7 Probit Matrix

	Actual	
Predicted	0	1
0	6158	1351
1	22	24

> |

Table.8 Logit V Probit Measures

```
Original Probit AIC: 6718.646
> cat("Alternative Logit AIC:", alt_aic, "\n")
Alternative Logit AIC: 6722.709
> cat("Alternative Model Accuracy:", round(alt_accuracy, 4), "\n")
Alternative Model Accuracy: 0.8181
> cat("Alternative McFadden R2:", round(mcfadden_r2, 4), "\n")
Alternative McFadden R2: 0.0677
> cat("Likelihood Ratio Test p-value:", round(p_value, 4), "\n")
Likelihood Ratio Test p-value: 1
> # Confusion matrix for the alternative logit model
> table(Predicted = meps$alt_pred_class, Actual = meps$Doctorvisit)
         Actual
Predicted    0    1
        0 6152 1346
        1   28   29
>
> # Optional: calculate sensitivity and specificity
> TP <- sum(meps$alt_pred_class == 1 & meps$Doctorvisit == 1)
> TN <- sum(meps$alt_pred_class == 0 & meps$Doctorvisit == 0)
> FP <- sum(meps$alt_pred_class == 1 & meps$Doctorvisit == 0)
> FN <- sum(meps$alt_pred_class == 0 & meps$Doctorvisit == 1)
>
> sensitivity <- TP / (TP + FN)  # True positive rate
> specificity <- TN / (TN + FP)  # True negative rate
> accuracy <- (TP + TN) / (TP + TN + FP + FN)
>
> cat("Sensitivity:", round(sensitivity, 4), "\n")
Sensitivity: 0.0211
> cat("Specificity:", round(specificity, 4), "\n")
Specificity: 0.9955
> cat("Accuracy:", round(accuracy, 4), "\n")
Accuracy: 0.8181
```

Table.9 Accuracy measures

Original Probit Model AIC: 6718.6 Alternative Logit Model AIC: 6722.7 Probit Accuracy: 81.83% Logit Accuracy: 81.81% Probit McFadden R^2: 0.0680 Logit McFadden R^2: 0.0677

Table.10 Dummy Variable Assignments

Physical Health Level	Dummy Variable	Score
Poor	phpoor	1
Fair	phfair	2
Good	phgood	3
Very Good	phverygood	4
Excellent	phexcellent	5

Table.11 Ordered Probit Model estimation and Thresholds

```
Coefficients:
               Value Std. Error  t value
Privateins  0.026178   0.027850  0.93996
age        -0.011788   0.001299 -9.07674
male        0.046925   0.025245  1.85880
educ        0.051321   0.004998 10.26873
income      0.006715   0.001234  5.43993
famsize     0.021669   0.008598  2.52013
married     0.002862   0.034671  0.08256
MSA         0.001499   0.036750  0.04080
RaceBlack  -0.096077   0.053738 -1.78789
RaceOther  -0.184053   0.130852 -1.40658
RaceWhite   0.026235   0.046718  0.56156

Intercepts:
    Value    Std. Error t value
1|2  -1.9232   0.1174   -16.3802
2|3  -0.9223   0.1116    -8.2651
3|4   0.1505   0.1112     1.3532
4|5   1.0883   0.1116     9.7516

Residual Deviance: 19812.80
AIC: 19842.80
```

```
                 Value  Std. Error      t value      p-value
Privateins  0.026177603 0.027849756   0.93995808 3.472391e-01
age        -0.011788430 0.001298752  -9.07673858 1.118769e-19
male        0.046924502 0.025244578   1.85879527 6.305616e-02
educ        0.051321074 0.004997802  10.26872787 9.748158e-25
income      0.006714685 0.001234334   5.43992541 5.330288e-08
famsize     0.021668640 0.008598227   2.52012876 1.173119e-02
married     0.002862416 0.034670595   0.08256033 9.342011e-01
MSA         0.001499399 0.036749511   0.04080051 9.674549e-01
RaceBlack  -0.096077086 0.053737673  -1.78789071 7.379364e-02
RaceOther  -0.184052882 0.130851755  -1.40657557 1.595533e-01
RaceWhite   0.026234857 0.046718047   0.56155723 5.744177e-01
1|2        -1.923223339 0.117411407 -16.38020858 2.648335e-60
2|3        -0.922332548 0.111593797  -8.26508798 1.395918e-16
3|4         0.150463905 0.111194724   1.35315687 1.760055e-01
4|5         1.088316404 0.111604278   9.75156526 1.816456e-22
>
```

Table.12 APE of ordered probit model

```
> summary(ape_ord_probit)
     factor      AME     SE        z      p   lower   upper
        age  0.0003 0.0001  3.1654 0.0015  0.0001  0.0005
      educ -0.0015 0.0003 -5.0859 0.0000 -0.0020 -0.0009
    famsize -0.0006 0.0002 -2.6114 0.0090 -0.0011 -0.0002
     income -0.0002 0.0001 -3.0384 0.0024 -0.0003 -0.0001
       male -0.0013 0.0007 -1.8546 0.0636 -0.0027  0.0001
    married -0.0001 0.0010 -0.0824 0.9343 -0.0020  0.0019
        MSA -0.0000 0.0010 -0.0409 0.9674 -0.0021  0.0020
 Privateins -0.0007 0.0008 -0.9213 0.3569 -0.0023  0.0008
  RaceBlack  0.0030 0.0020  1.4922 0.1356 -0.0009  0.0070
  RaceOther  0.0064 0.0057  1.1160 0.2644 -0.0048  0.0176
  RaceWhite -0.0007 0.0012 -0.5832 0.5598 -0.0031  0.0017
>
```

Table.13/14 Ordered Probit Accuracy Matrix

```
Overall Accuracy: 0.359
>
> # Confusion matrix
> confusion_matrix <- table(Predicted = meps$predicted_phscore, Actual = meps$phscore)
> print(confusion_matrix)
         Actual
Predicted    1    2    3    4    5
        3   35  255  618  501  318
        4   46  344 1468 1772 1340
        5    4   32  156  344  322
>
```

educ_group	Accuracy
<fct>	<dbl>
< High School	0.344
High School	0.336
Some College	0.370
Bachelor+	0.400
NA	0.464

	age_group	Accuracy
	<fct>	<dbl>
1	25-34	0.360
2	35-44	0.348
3	45-54	0.362
4	55-64	0.372

Table.15

```
>
> # Quick summary
> table(meps$mhscore)

     Poor      Fair      Good Very Good Excellent
       23       248      1728      2286      3270
>
```

Table.16

```
                Value  Std. Error     t value      p value
educ        0.04794674 0.005092461   9.4152391 4.720069e-21
famsize     0.04703693 0.008601298   5.4685849 4.536428e-08
income      0.00815420 0.001287046   6.3355946 2.364277e-10
married    -0.01241731 0.033100235  -0.3751427 7.075543e-01
male        0.07789642 0.026140831   2.9798755 2.883655e-03
Privateins -0.01354770 0.028789426  -0.4705789 6.379415e-01
RaceBlack  -0.09943402 0.056654450  -1.7550964 7.924282e-02
RaceOther  -0.31431989 0.134360024  -2.3393855 1.931549e-02
RaceWhite  -0.13987134 0.049359948  -2.8337012 4.601233e-03
1|2        -1.96828565 0.112455839 -17.5027430 1.365326e-68
2|3        -1.00484879 0.093490410 -10.7481482 6.046326e-27
3|4         0.19940939 0.091985061   2.1678453 3.017046e-02
4|5         1.01721728 0.092390421  11.0099864 3.420554e-28
>
> # Average Partial Effects
> library(margins)
> ape_mental <- margins(mental_model)
> summary(ape_mental)
     factor     AME     SE       z      p   lower   upper
       educ -0.0004 0.0001 -5.6851 0.0000 -0.0006 -0.0003
    famsize -0.0004 0.0001 -4.0413 0.0001 -0.0006 -0.0002
     income -0.0001 0.0000 -3.3858 0.0007 -0.0001 -0.0000
       male -0.0007 0.0003 -2.7623 0.0057 -0.0012 -0.0002
    married  0.0001 0.0003  0.3685 0.7125 -0.0005  0.0007
 Privateins  0.0001 0.0003  0.4637 0.6429 -0.0004  0.0006
  RaceBlack  0.0007 0.0005  1.4300 0.1527 -0.0003  0.0017
  RaceOther  0.0031 0.0021  1.5029 0.1329 -0.0010  0.0072
  RaceWhite  0.0011 0.0005  2.0637 0.0390  0.0001  0.0021
>
```

Figure.17 Mental Health distribution by education level

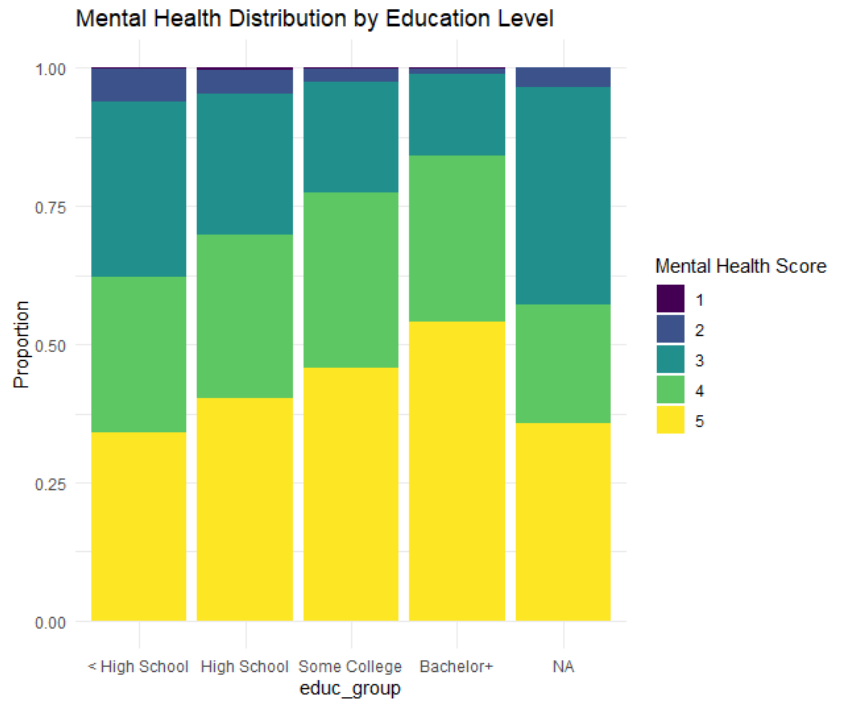

Figure.18 Average mental health score by income

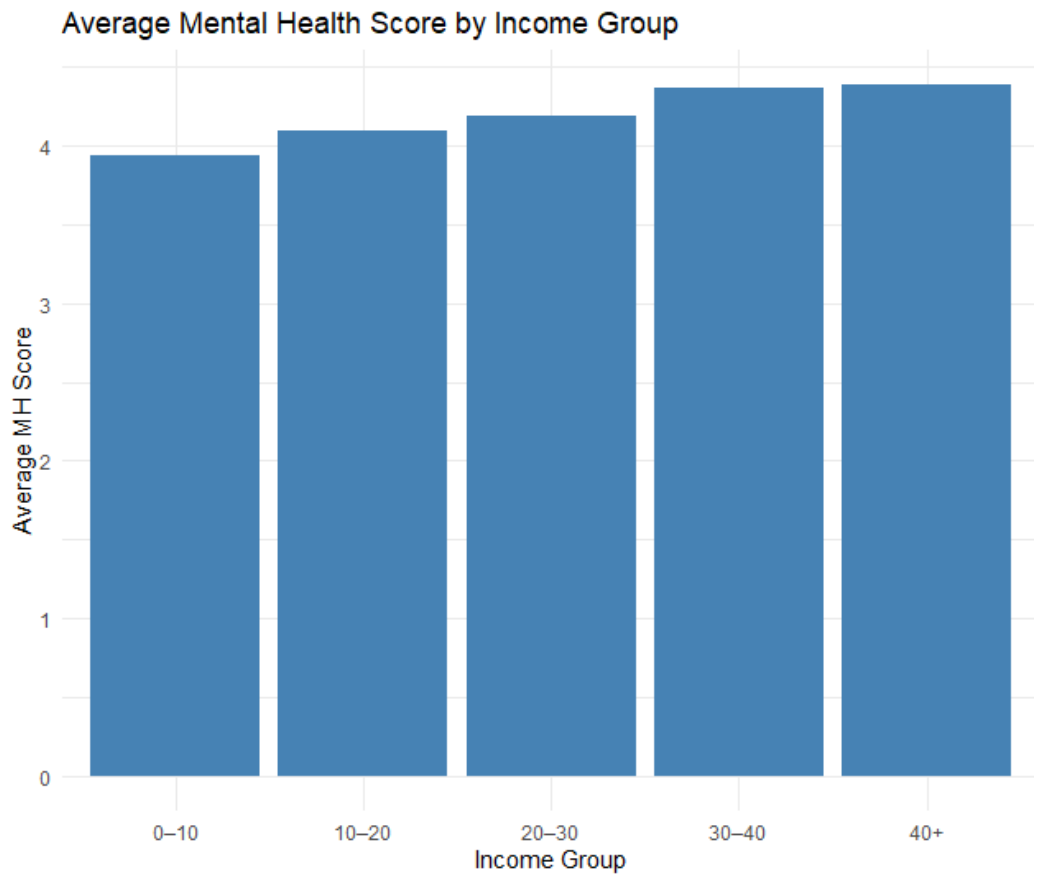

Table.17 Accuracy measures

```
> # Confusion matrix
> confusion_matrix <- table(Predicted = predicted_mhscore, Actual = actual_mhscore)
> print(confusion_matrix)
         Actual
Predicted    1    2    3    4    5
        3    1   32  130   93  105
        4    2   10   33   34   46
        5   20  206 1565 2159 3119
>
> # Overall accuracy
> accuracy <- mean(predicted_mhscore == actual_mhscore, na.rm = TRUE)
> cat("Overall Accuracy:", round(accuracy, 3), "\n")
Overall Accuracy: 0.435
```

```
> # View outputs
> print(acc_educ)
# A tibble: 5 x 2
  educ_group    Accuracy
  <fct>            <dbl>
1 < High School    0.348
2 High School      0.403
3 Some College     0.457
4 Bachelor+        0.541
5 NA               0.393
> print(acc_income)
# A tibble: 5 x 2
  income_group Accuracy
  <fct>           <dbl>
1 0-10            0.367
2 10-20           0.411
3 20-30           0.456
4 30-40           0.536
5 40+             0.557
> print(acc_age)
# A tibble: 4 x 2
  age_group Accuracy
  <fct>        <dbl>
1 25-34        0.483
2 35-44        0.450
3 45-54        0.393
4 55-64        0.399
>
```

Figure.19 Distribution of Physical Health by Mental Health Level

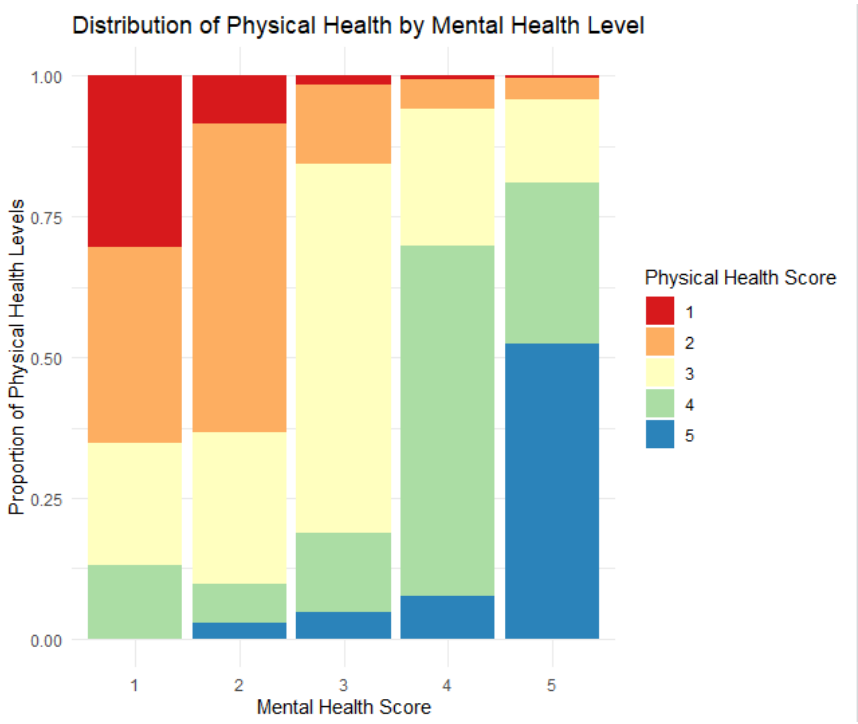

Figure.20 Mental health X Physical health box plot

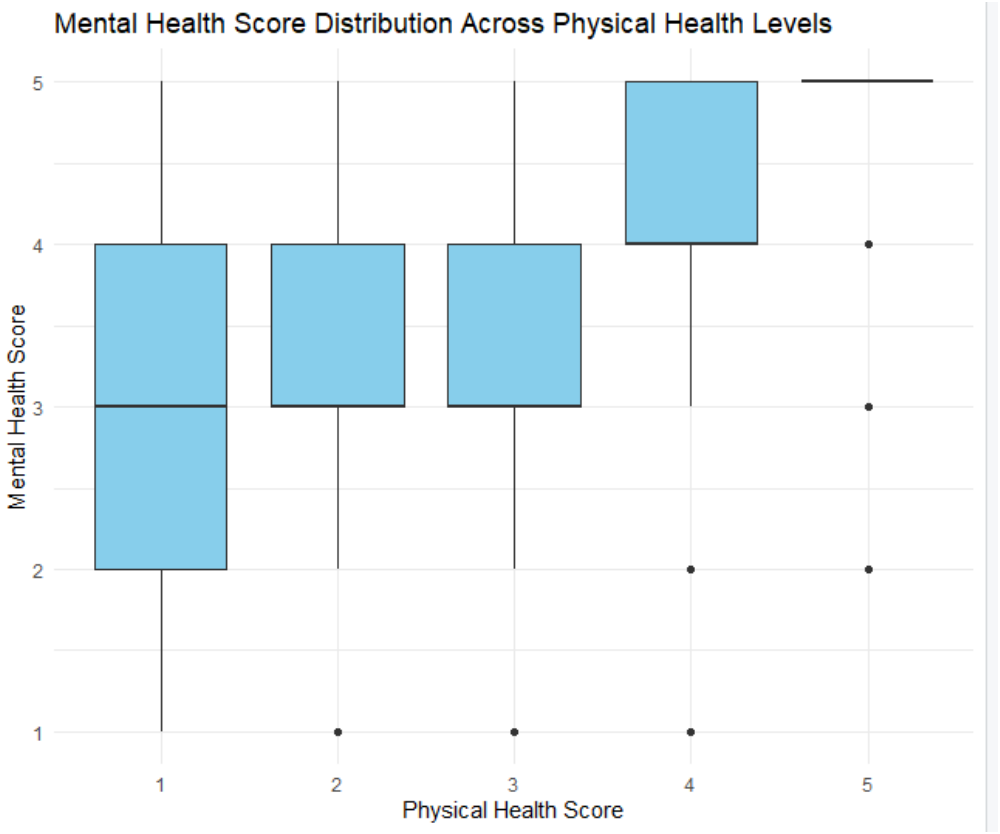

Figure 21. Heatmap

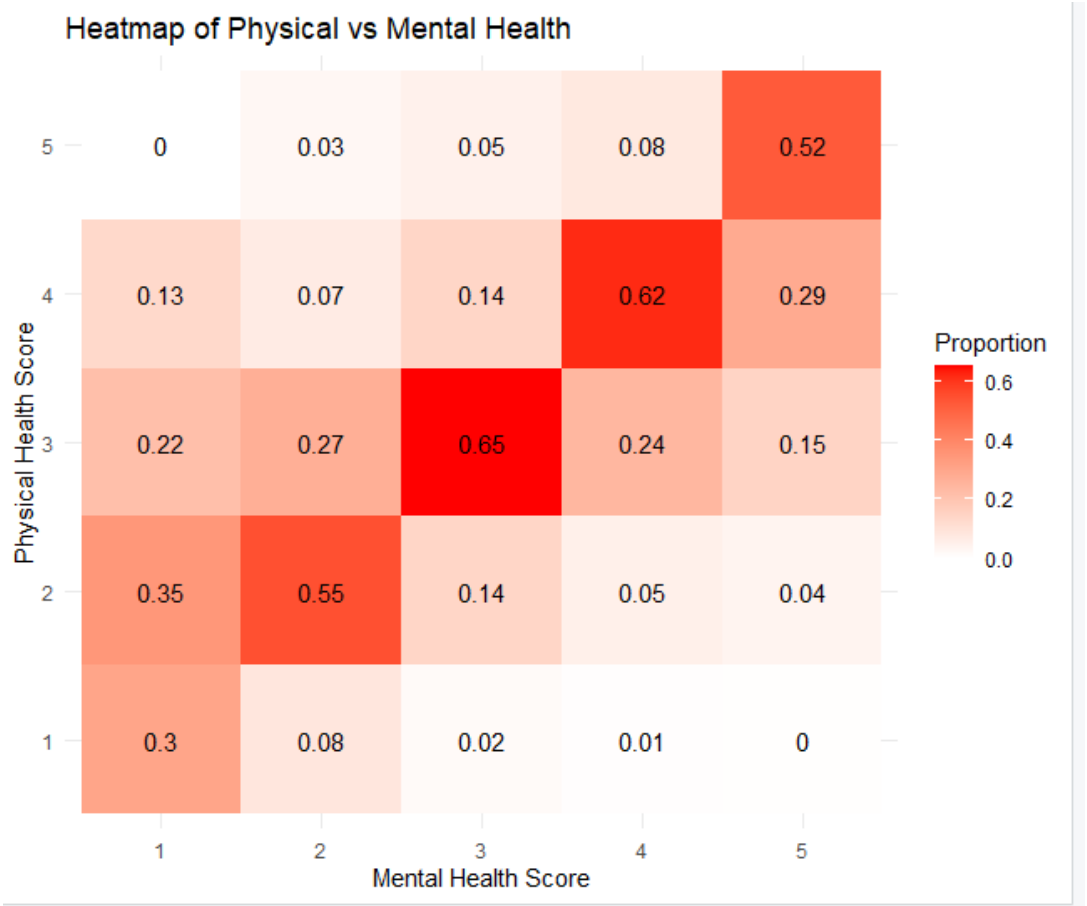

Table.18 Covariance

```
> cor(as.numeric(meps$mhscore), as.numeric(meps$phscore), use = "complete.obs")
[1] 0.5661878
```