

The Protective Power of Friendship: Peer Support and Depression Across Socioeconomic Groups

Introduction

Adolescent mental health remains a critical public health concern, with rising rates of depression reported among young people across socioeconomic backgrounds. While the role of economic disadvantage in shaping mental health outcomes is well documented, growing attention is being paid to the social support structures that may help buffer these effects. Among these, peer relationships and friendships are emerging as potential protective factors during adolescence.

This project investigates whether close peer support can mitigate the risk of depression for adolescents from lower-income households. Using data from Wave III of the National Longitudinal Study of Adolescent to Adult Health (Add Health), the analysis explores how varying levels of peer interaction relate to reported diagnoses of depression, and how this relationship may differ across socioeconomic groups.

Prior research provides the foundation for this investigation. Eisenberg et al. (2007) highlight that students with stronger social connections exhibit significantly lower levels of depressive symptoms, suggesting that perceived support from friends can foster emotional resilience during stressful periods. Their work emphasizes the need for integrating social context into mental health assessments among youth.

Goodman et al. (2003) demonstrate a clear gradient between household income and the prevalence of emotional disorders in adolescents, with those from lower-income families being disproportionately affected. This underscores the relevance of socioeconomic status as both a risk factor and a context for interpreting adolescent health outcomes.

Hammack et al. (2018) explore how peer relationships can serve as a critical buffer for youth facing systemic or interpersonal adversity. Their findings show that emotionally supportive friendships significantly reduce psychological distress, especially among marginalized or high-risk populations.

While these studies underscore the role of social support, they often overlook whether its protective effect differs across socioeconomic groups. This analysis addresses that gap by examining if peer support buffers against depression specifically for adolescents from lower-income households. By focusing on interaction effects and subgroup analyses, it offers insights into how friendship shapes mental health trajectories across different socioeconomic backgrounds.

Descriptive Statistics

This analysis focuses on Diagnosed Depression, a binary variable indicating whether a respondent reported a formal diagnosis of depression. This measure offers a concrete health outcome with relevance for both clinical and public health perspectives.

A total of 4,288 observations from Wave III of the National Longitudinal Study of Adolescent to Adult Health (Add Health) were included, covering adolescents aged 18 to 26. Depression diagnosis was chosen due to its critical importance in mental health research, particularly during young adulthood which is a period marked by significant social and emotional development.

A total of 17 explanatory variables were selected based on theoretical relevance and consistency with prior literature. These include:

- Demographic characteristics:
 - *Gender* (male/female)
 - *Race/ethnicity* (White, Other)
- Socioeconomic indicators (both parental and respondent-based):
 - *Parental income* at Wave I (categorised ranges)
 - *Respondent's perceived household income at Wave III*
 - *Education support from family and education aspiration*
- Peer support metrics:
 - *Time spent with friends* (Never, Sometimes, Most of the time)
 - *Perceived influence of friends versus family on decisions*
 - *Closeness to best male and female friends*
- Family dynamics:
 - *Closeness to mother and father*
 - *Whether parents discussed important matters*
 - *Household composition (living with both parents)*
- Mental health and behaviour controls:
 - *History of being diagnosed with a mental health issue*
 - *Past depressive symptoms (e.g., feeling sad in the past week)*

Each variable was recoded into interpretable categories and reviewed for missingness. Observations with missing values on the dependent variable or critical covariates were excluded from the regression sample.

Figure 1 below illustrates the association between closeness to friends and diagnosed depression, disaggregated by gender and welfare status. Adolescents who reported being less emotionally close to their friends exhibited higher rates of depression, particularly among females and low-income groups. The trend suggests that peer closeness may serve as a protective factor, though its strength appears to vary by both gender and socioeconomic background.

Depression Diagnosis by Closeness to Friends, Income, and Gender

Higher closeness is associated with lower depression rates across genders and income groups

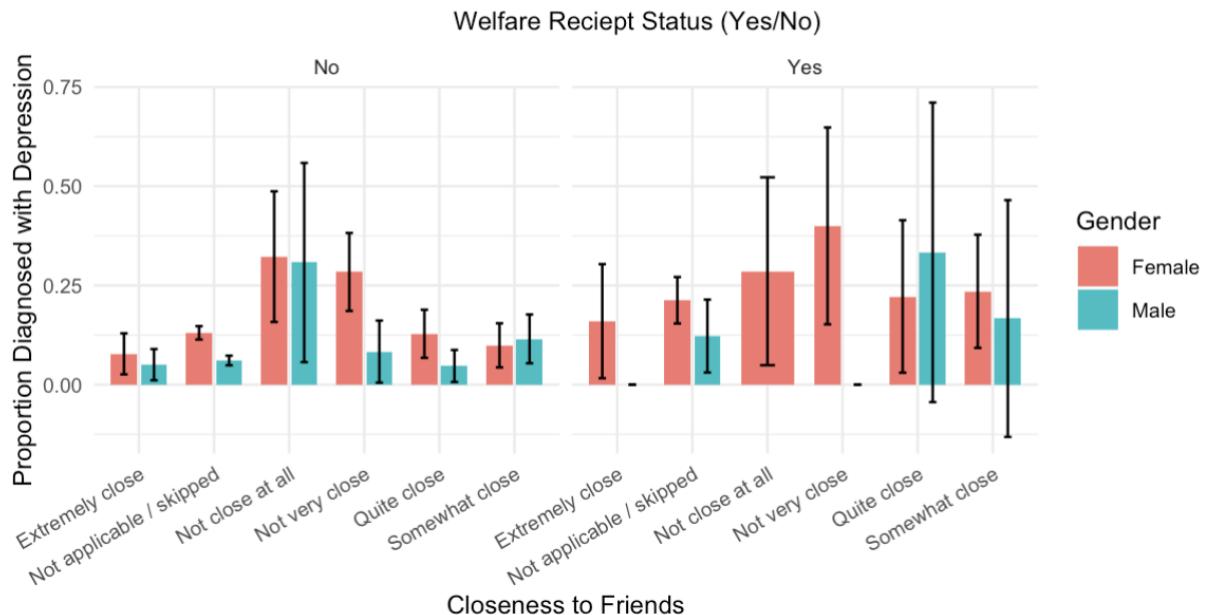


Figure 1. Depression Diagnosis by Closeness to Friends, Gender, and Welfare Status

Additional visualisations, including depression rates by household income, gender, and family closeness, are provided in the appendix. Together, these descriptive statistics offer preliminary evidence that social and economic factors jointly shape adolescent mental health outcomes.

Regression Analysis

To assess the association between social support and adolescent depression, a series of logistic regression models were estimated using Wave III of the Add Health dataset. The outcome variable is a binary indicator denoting whether the adolescent has ever been diagnosed with depression by a medical professional. The explanatory variables include peer and parental support indicators, along with demographic and socioeconomic controls.

Baseline Model - The baseline model incorporates only socio-demographic and behavioural predictors. Findings indicate that male adolescents and those identifying as White exhibit significantly lower odds of reporting depression. Not being economically well off and distant from family is strongly associated with higher depression rates. However, income-related variables do not show consistent effects, underscoring the need to incorporate social interaction indicators.

Full Model with Peer and Family Support - The full specification introduces peer-related variables (e.g., time spent with peers, whose opinion matters more) and parent-child closeness. The results suggest that weaker peer engagement and lack of emotional closeness with friends and family are associated with higher odds of depression. For instance, adolescents who “never or rarely” spent time with school peers were significantly more likely to report a depression diagnosis.

Stepwise Model - A stepwise selection procedure was used to identify a more parsimonious model while retaining key explanatory power. This model achieves a lower AIC (2728.8) compared to the full model (2752.2), suggesting improved model fit. Importantly, peer and parental support variables remain significant, reinforcing the hypothesis that social connectedness is a more consistent predictor of adolescent mental health than economic status.

Model	AIC	Key Predictors	Insights
Baseline	2878.2	Gender, race, substance use	Income effects limited
Full	2752.2	Peer/family support, closeness	Strong support for core hypothesis
Stepwise	2728.8	Peer interaction, family closeness	Best-fitting, interpretable model

Table 1. Comparison of regression models based on AIC and key predictors

Based on the model comparison table, the stepwise model demonstrated the best overall fit. This reinforces that peer interaction and family closeness, together with socioeconomic variables, are key predictors of adolescent mental health.

Additional Models

1. Gender-Specific Models - Separate regressions were run for males and females to test for gender heterogeneity. In both groups, lower peer engagement and weak emotional closeness were linked to higher depression rates.
2. Peer Support \times Low-Income Interaction - An interaction model assessed whether the effect of peer interaction varies by socioeconomic status. While lower interaction was linked to higher depression, the interaction term with low-income status was not significant, suggesting that peer support has a similarly protective effect across income groups.

Model	Key Finding
Gender-stratified	Stronger peer effects for males; peer support significant in both
Peer \times Income	No differential effect by income; support matters across SES

Table 2. Summary of additional models exploring gender differences and income interactions

These findings strengthen the case for peer-based interventions across diverse adolescent populations.

Limitations

This analysis provides relevant insights into how peer and social factors relate to adolescent depression, though a few considerations should be kept in mind. The cross-sectional design limits the ability to assess changes over time. Some variables are self-reported, which may affect measurement precision. While important controls were included, there may be other unobserved influences, such as school environment or personality traits, that could also play a role. Lastly, the definitions used for income and closeness provide broad patterns but may not reflect every nuance of adolescents' lived experiences.

Use of Weights

The Add Health dataset is derived from a complex survey design that includes clustering, stratification, and unequal probabilities of selection. In this analysis, weights were not applied. As a result, findings should be interpreted as sample-specific associations rather than population-level inferences.

While unweighted models offer valid insights into the relationships between depression and peer support within this sample, applying weights could improve generalizability. Given the project's focus on key factors that affect mental health rather than estimation at the population level, the decision to omit weights was deemed appropriate.

Conclusion

This project investigated the relationship between peer and family support and adolescent depression using Wave III of the Add Health dataset. The findings consistently point to the protective role of social connectedness particularly strong friendships and frequent peer interactions in reducing the likelihood of depressive symptoms. These associations remain robust across different model specifications and subgroups, providing evidence for the relevance of social support in adolescent mental health.

While the study does not make causal claims, the insights are highly informative. They reveal clear patterns that can inform school based interventions, youth mental health initiatives, and future research. By focusing on a adolescent sample, this analysis contributes to a growing body of evidence for stronger social integration and support systems during adolescence. Future work could build on this foundation using longitudinal or causal techniques, but the current results already underline the importance of considering peer dynamics in mental health policy and program design.

END

Appendix

Appendix A1: Variable Definitions

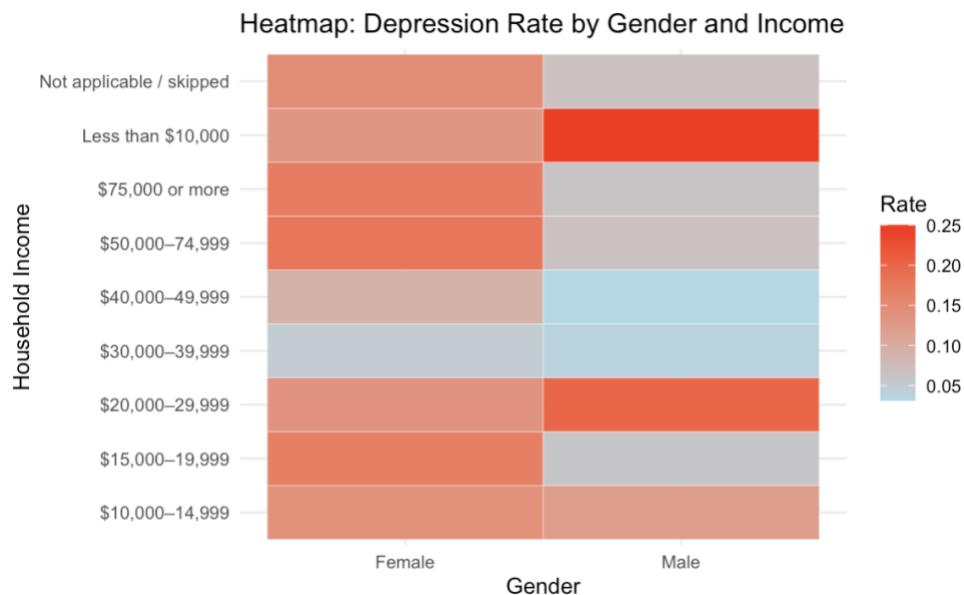
Variable Name	Description	Type	Recoding	Measurement Unit / Categories
Diagnosed Depression	Self-reported diagnosis by a healthcare professional	Binary	Yes = 1, No = 0	Yes/No
Gender	Respondent's biological sex	Categorical	Male/Female	Nominal
Race/Ethnicity	White vs. Other	Binary	White = 1, Other = 0	Nominal
Parental Income (Wave I)	Parental income reported at Wave I	Categorical	Grouped into 7 income brackets	Nominal
Household Income (Wave III)	Respondent's perception of household income	Categorical	Grouped into 7 income brackets	Nominal
Education Support	Parental support for education	Categorical	Levels of family education support	Ordinal
Education Aspiration	Respondent's level of educational aspiration	Categorical	Scale of aspirations	Ordinal
Time with Peers	Frequency of spending time with school peers	Categorical	Never, Sometimes, Most/all of the time	Ordinal
Influence on Decisions	Whether friends or family are more influential in decisions	Categorical	Friends vs. Family	Nominal
Closeness to Best Friend (Male/Female)	Emotional closeness with best friends	Categorical	5-point closeness scale	Ordinal
Closeness to Mother	Emotional closeness with mother	Categorical	5-point closeness scale	Ordinal
Closeness to Father	Emotional closeness with father	Categorical	5-point closeness scale	Ordinal
Parental Discussion of Important Matters	Did parents discuss important matters with respondent?	Binary	Yes/No	Yes/No
Living with Both Parents	Household composition including both biological parents	Binary	Yes/No	Yes/No
ADD/ADHD Diagnosis	Diagnosed with ADD/ADHD (Wave III)	Binary	Yes = 1, No = 0	Yes/No
Recent Depressive Feelings	Felt sad or low during the past week	Binary	Yes = 1, No = 0	Yes/No

Appendix A2: Summary Statistics

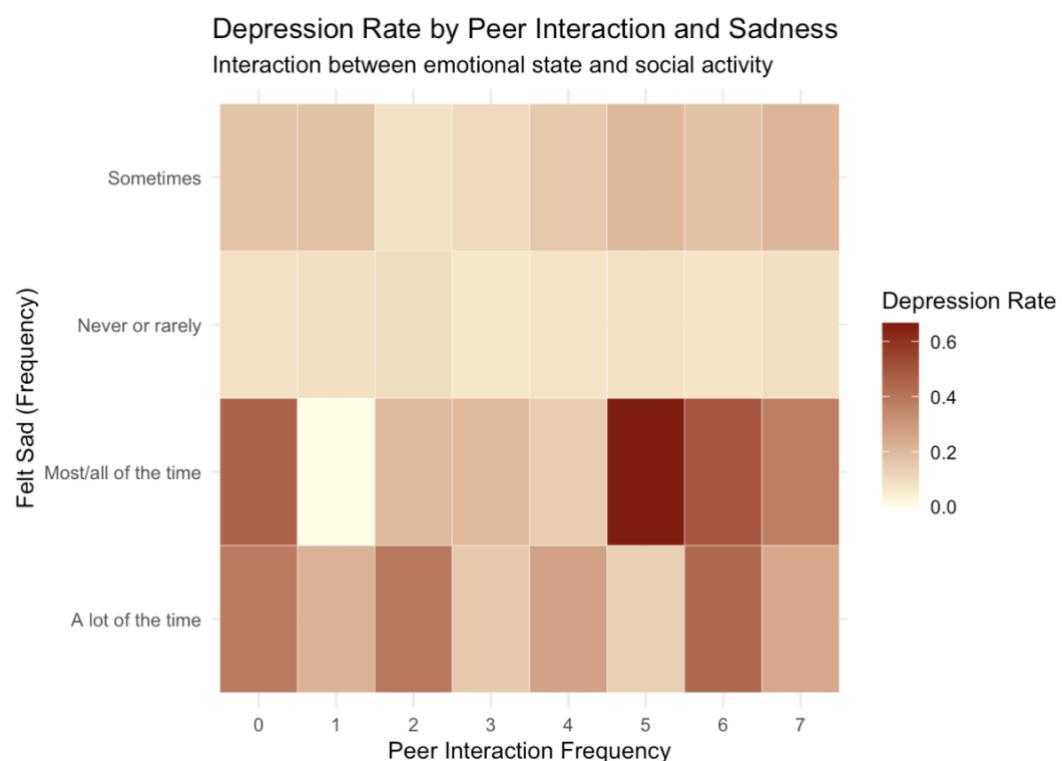
Variable	Category/Stat	Value	Variable	Category/Stat	Value
Diagnosed Depression	No Depression	3813	Closeness to Best Male Friend	Not applicable / skipped	3929
Diagnosed Depression	Diagnosed Depression	475	Closeness to Best Male Friend	Extremely close	161
Gender	Female	2314	Closeness to Best Male Friend	Quite close	95
Gender	Male	1974	Closeness to Best Male Friend	Somewhat close	63
Race (White)	White	3033	Closeness to Best Male Friend	Not very close	33
Race (White)	Not White	1255	Closeness to Best Male Friend	Not close at all	7
Parental Income (Wave I)	Not applicable / skipped	3661	Closeness to Best Female Friend	Not applicable / skipped	3321
Parental Income (Wave I)	Less than \$10,000	317	Closeness to Best Female Friend	Somewhat close	259
Parental Income (Wave I)	\$10,000-\$14,999	108	Closeness to Best Female Friend	Extremely close	251
Parental Income (Wave I)	\$20,000-\$29,999	80	Closeness to Best Female Friend	Quite close	249
Parental Income (Wave I)	\$15,000-\$19,999	75	Closeness to Best Female Friend	Not very close	149
Parental Income (Wave I)	\$30,000-\$39,999	28	Closeness to Best Female Friend	Not close at all	59
Parental Income (Wave I)	\$40,000-\$49,999	11	Parent Discussed Important Matters	Not applicable / skipped	4110
Parental Income (Wave I)	\$75,000 or more	4	Parent Discussed Important Matters	No	127
Parental Income (Wave I)	\$50,000-\$74,999	4	Parent Discussed Important Matters	Yes	51
Self-Reported Household Income	Not applicable / skipped	3809	Lived with Both Parents	No	4080
Self-Reported Household Income	\$75,000 or more	143	Lived with Both Parents	Yes	208
Self-Reported Household Income	\$50,000-\$74,999	101	Influence of Friends vs Family	Not Asked	3310
Self-Reported Household Income	\$40,000-\$49,999	65	Influence of Friends vs Family	Family more influential	591
Self-Reported Household Income	\$30,000-\$39,999	46	Influence of Friends vs Family	Friends more influential	387
Self-Reported Household Income	\$20,000-\$29,999	42	Diagnosed with ADHD	Yes	3200
Self-Reported Household Income	Less than \$10,000	31	Diagnosed with ADHD	No	1088
Self-Reported Household Income	\$15,000-\$19,999	29	Felt Sad (Past Week)	Mean	4.39
Self-Reported Household Income	\$10,000-\$14,999	22	Felt Sad (Past Week)	Std Dev	2.37
Time Spent with Friends	Never or rarely	3271	Felt Sad (Past Week)	Missing	0
Time Spent with Friends	Sometimes	776			
Time Spent with Friends	A lot of the time	158			
Time Spent with Friends	Most/all of the time	83			

Appendix A3: Additional Graphs & Visuals

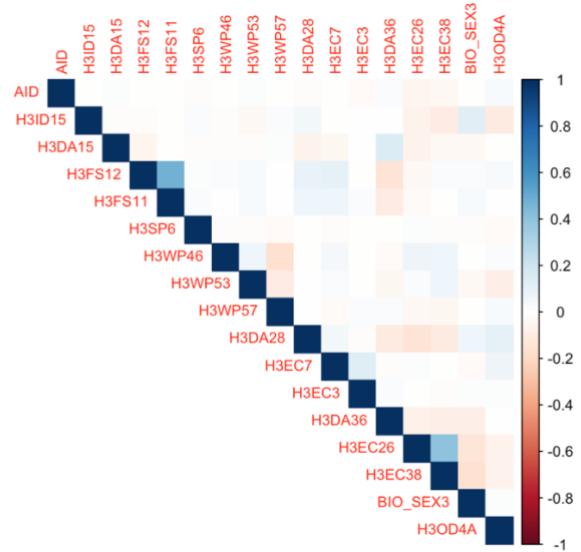
1. Heatmap – Depression Rate by Gender & Income



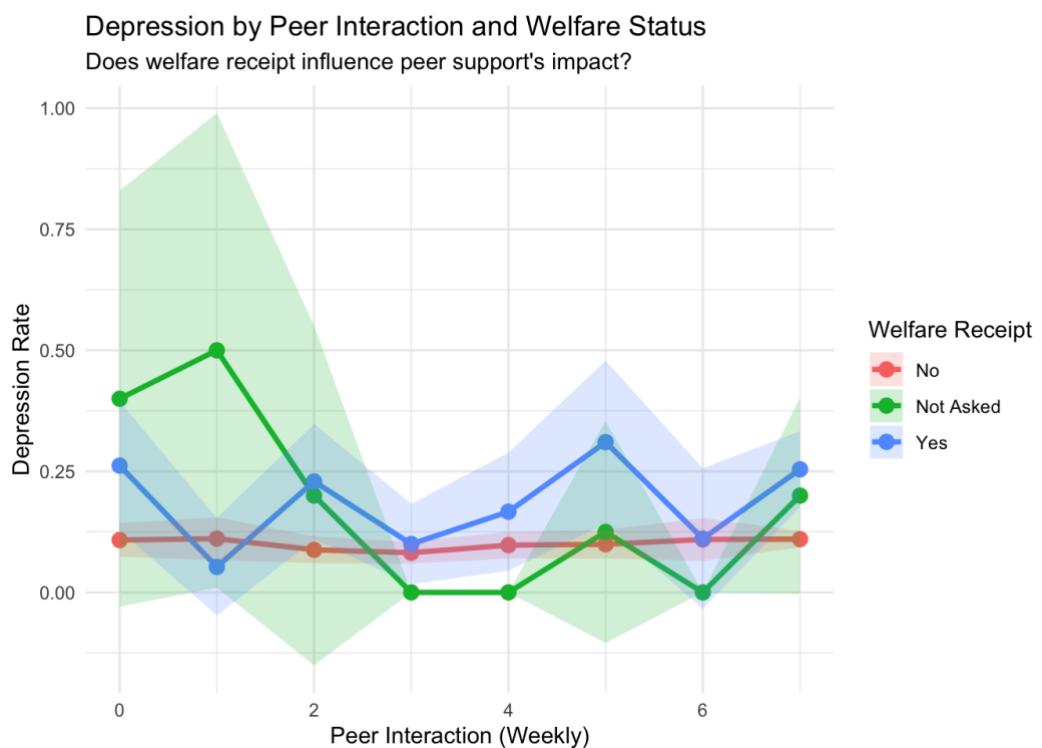
2. Depression Rate by Peer Interaction and Sadness



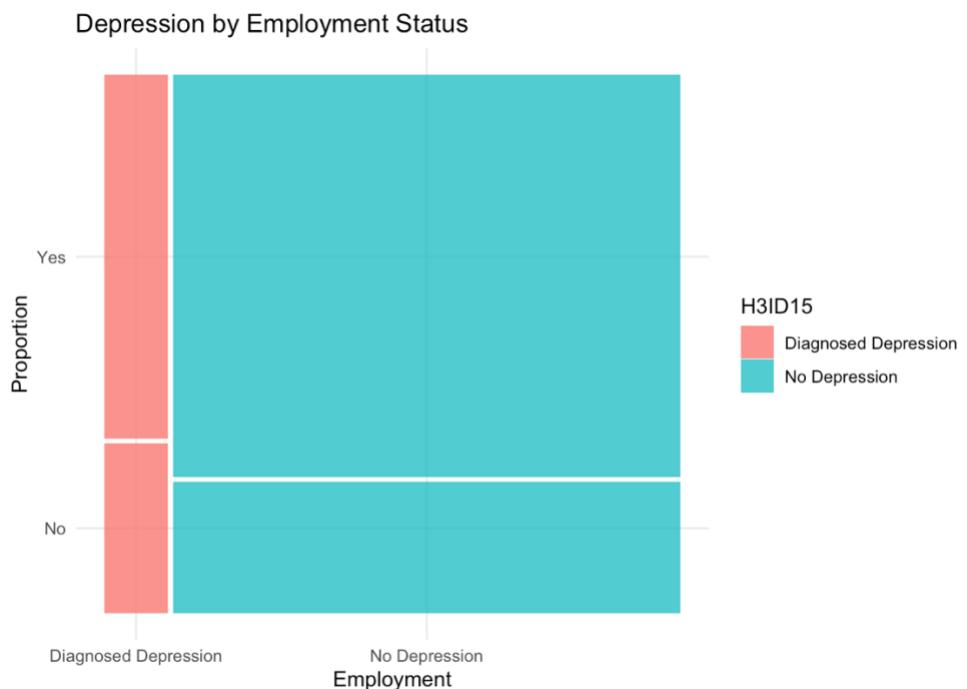
3. Correlation Plot



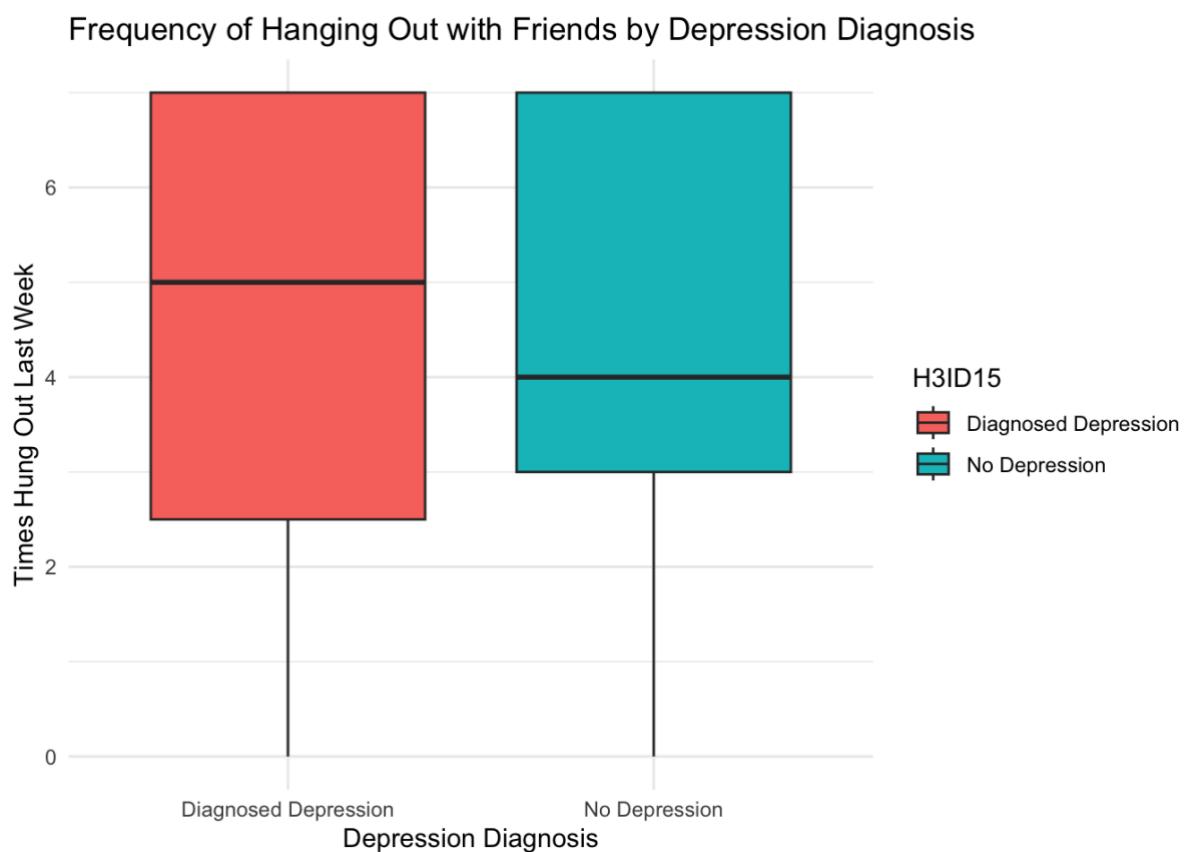
4. Depression by Peer Interaction and Welfare Status



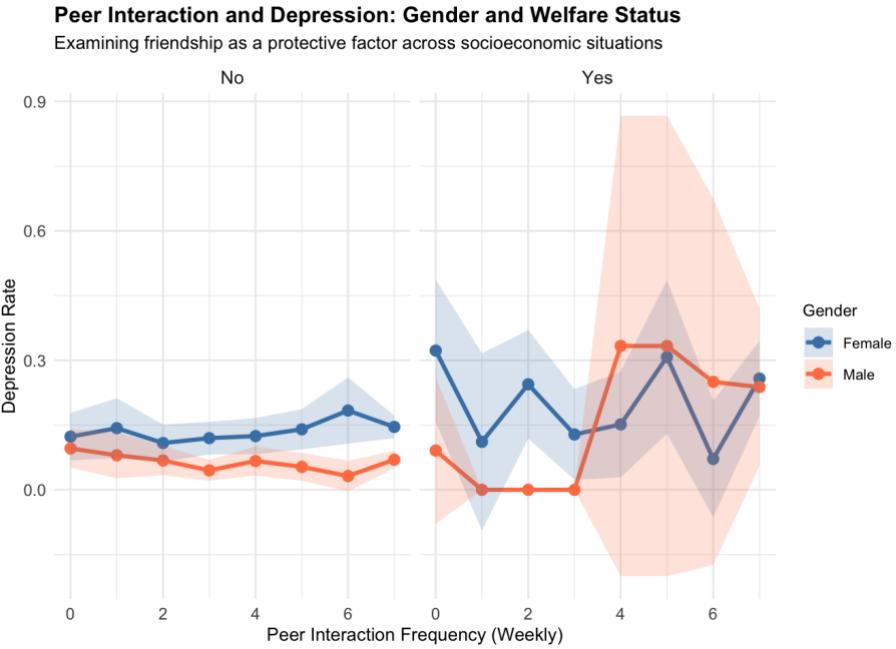
5. Depression by Employment Status:



6. Frequency of Hanging Out with Friends by Depression Diagnosis



7. Peer Interaction and Depression: Gender and Welfare Status



Appendix A4: Regression Output Tables

1. Baseline Regression Model Output

```
##
## Call:
## glm(formula = H3ID15 ~ H3EC3 + H3EC7 + H3DA28 + H3DA36 + BIO_SEX3 +
##      H3D04A, family = binomial, data = data)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                2.0059    0.7097   2.826 0.004709 **
## H3EC3$15,000-19,999       -0.4479    0.4879  -0.918 0.358636
## H3EC3$20,000-29,999       -0.5133    0.4750  -1.081 0.279819
## H3EC3$30,000-39,999       -0.6891    0.6397  -1.077 0.281335
## H3EC3$40,000-49,999        11.9528   260.8805   0.046 0.963456
## H3EC3$50,000-74,999       -1.5048    1.2388  -1.215 0.224472
## H3EC3$75,000 or more      -1.4505    1.2306  -1.179 0.238535
## H3EC3Less than $10,000     -0.2761    0.3675  -0.751 0.452530
## H3EC3Not applicable / skipped -0.1923   0.3301  -0.582 0.560277
## H3EC7$15,000-19,999       0.2259    0.8906   0.254 0.799760
## H3EC7$20,000-29,999       -0.2564    0.7654  -0.335 0.737633
## H3EC7$30,000-39,999        1.1271    0.9677   1.165 0.244129
## H3EC7$40,000-49,999        0.9564    0.8256   1.158 0.246708
## H3EC7$50,000-74,999        0.2023    0.7078   0.286 0.775018
## H3EC7$75,000 or more       0.2734    0.6947   0.394 0.693910
## H3EC7Less than $10,000      -0.4803    0.7946  -0.604 0.545521
## H3EC7Not applicable / skipped 0.3610    0.6421   0.562 0.573950
## H3DA28Yes                  0.3974    0.1093   3.637 0.000276 ***
## H3DA36Yes                  0.1107    0.1040   1.064 0.287140
## BIO_SEX3Male                0.8652    0.1087   7.959 1.73e-15 ***
## H3D04AWhite                 -0.9325   0.1322  -7.056 1.72e-12 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2985.6 on 4287 degrees of freedom
## Residual deviance: 2836.2 on 4267 degrees of freedom
## AIC: 2878.2
##
## Number of Fisher Scoring iterations: 13
```

2. Full Regression Model Output:

```

## 
## Call:
## glm(formula = H3ID15 ~ H3FS12 + H3FS11 + H3SP6 + H3WP46 + H3WP53 +
##      H3WP57 + H3DA15 + H3DA28 + H3DA36 + H3EC7 + H3EC3 + H3EC26 +
##      H3EC38 + BIO_SEX3 + H3OD4A, family = binomial, data = data)
## 

## Coefficients: (1 not defined because of singularities)
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  1.85679  0.87326  2.126  0.03348 *
## H3FS12Friends more influential -0.42142  0.22121 -1.905  0.05677 .
## H3FS12Not Asked              -0.21419  0.21730 -0.986  0.32428
## H3FS11All                     0.98530  0.76752  1.284  0.19923
## H3FS11Most                   0.60785  0.30961  1.963  0.04961 *
## H3FS11None                   -0.32705  0.41821 -0.782  0.43419
## H3FS11Not Asked              NA       NA       NA       NA
## H3FS11One                     0.55973  0.75849  0.738  0.46054
## H3FS11Some                    -0.10794  0.27011 -0.400  0.68943
## H3SP6Most/all of the time   -0.33855  0.32061 -1.056  0.29098
## H3SP6Never or rarely        1.27747  0.20774  6.149  7.78e-10 ***
## H3SP6Sometimes                0.50124  0.22018  2.276  0.02282 *
## H3WP46Not applicable / skipped -0.01928  0.31095 -0.062  0.95056
## H3WP46Not close at all     -1.26003  0.89413 -1.409  0.15877
## H3WP46Not very close       -0.68161  0.50380 -1.353  0.17608
## H3WP46Quite close           0.90025  0.48230  1.867  0.06196 .
## H3WP46Somewhat close       -0.11284  0.44370 -0.254  0.79925
## H3WP53Not applicable / skipped -0.40227  0.26763 -1.503  0.13282
## H3WP53Not close at all     -1.57468  0.39576 -3.979  6.92e-05 ***
## H3WP53Not very close       -1.16842  0.33242 -3.515  0.00044 ***
## H3WP53Quite close          -0.39348  0.33433 -1.177  0.23924
## H3WP53Somewhat close       -0.51437  0.32426 -1.586  0.11267
## H3WP57Not applicable / skipped -0.55520  0.34646  1.602  0.10905
## H3WP57Yes                   0.09226  0.50458  0.183  0.85492
## H3DA15                      -0.03106  0.02222 -1.398  0.16216
## H3DA28Yes                   0.33311  0.11547  2.885  0.00392 **
## H3DA36Yes                   -0.03152  0.11085 -0.284  0.77614
## H3EC7$15,000–19,999         0.41120  0.92216  0.446  0.65566
## H3EC7$20,000–29,999         -0.42385  0.79055 -0.536  0.59185
## H3EC7$30,000–39,999         1.12181  0.99544  1.127  0.25976
## H3EC7$40,000–49,999         0.75476  0.84960  0.888  0.37434
## H3EC7$50,000–74,999         0.10759  0.73420  0.147  0.88349
## H3EC7$75,000 or more        0.11557  0.71820  0.161  0.87216
## H3EC7Less than $10,000      -0.46176  0.82989 -0.556  0.57793
## H3EC7Not applicable / skipped -0.29100  0.66489  0.438  0.66163
## H3EC3$15,000–19,999         -0.48037  0.51578 -0.931  0.35168
## H3EC3$20,000–29,999         -0.46785  0.49984 -0.936  0.34927
## H3EC3$30,000–39,999         -0.74911  0.65527 -1.143  0.25295
## H3EC3$40,000–49,999         11.89368 253.80613  0.047  0.96262
## H3EC3$50,000–74,999         -1.67275  1.30285 -1.284  0.19917
## H3EC3$75,000 or more        -2.04968  1.24262 -1.649  0.09905 .
## H3EC3Less than $10,000       -0.40141  0.38516 -1.042  0.29733
## H3EC3Not applicable / skipped -0.30255  0.34766 -0.870  0.38417
## H3EC26Yes                   -0.22082  0.22370 -0.987  0.32360
## H3EC38Not Asked              -0.86712  0.40539 -2.139  0.03244 *
## H3EC38Yes                   -0.45899  0.17169 -2.673  0.00751 **
## BIO_SEX3Male                 0.68594  0.11358  6.040  1.55e-09 ***
## H3OD4AWhite                  -1.13867  0.14088 -8.083  6.33e-16 ***
## ---

## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

## 

## Null deviance: 2985.6 on 4287 degrees of freedom
## Residual deviance: 2658.2 on 4241 degrees of freedom
## AIC: 2752.2

## 
## Number of Fisher Scoring iterations: 13

```

3. Stepwise Regression Output:

```
##  
## Call:  
## glm(formula = H3ID15 ~ H3FS12 + H3SP6 + H3WP46 + H3WP53 + H3DA28 +  
##      H3EC38 + BIO_SEX3 + H3D4A, family = binomial, data = data)  
##  
## Coefficients:  
##                                     Estimate Std. Error z value Pr(>|z|)  
## (Intercept)                 1.9588    0.4219   4.643 3.44e-06 ***  
## H3FS12Friends more influential -0.4218    0.2160  -1.953 0.050795 .  
## H3FS12Not Asked              -0.3043    0.1633  -1.864 0.062362 .  
## H3SP6Most/all of the time     -0.3105    0.3181  -0.976 0.329062  
## H3SP6Never or rarely          1.3151    0.2050   6.414 1.42e-10 ***  
## H3SP6Sometimes                0.5252    0.2176   2.413 0.015806 *  
## H3WP46Not applicable / skipped 0.1898    0.2577   0.737 0.461390  
## H3WP46Not close at all       -1.2005    0.8492  -1.414 0.157442  
## H3WP46Not very close          -0.7283    0.4878  -1.493 0.135433  
## H3WP46Quite close             0.8826    0.4791   1.842 0.065452 .  
## H3WP46Somewhat close          0.2108    0.4364  -0.483 0.628979  
## H3WP53Not applicable / skipped -0.3043   0.2583  -1.178 0.238730  
## H3WP53Not close at all        -1.5129   0.3902  -3.877 0.000106 ***  
## H3WP53Not very close          -1.0974   0.3280  -3.346 0.000819 ***  
## H3WP53Quite close              0.3333   0.3309  -1.007 0.313905  
## H3WP53Somewhat close           0.4753   0.3199  -1.486 0.137345  
## H3DA28Yes                      0.3396   0.1132   3.000 0.002703 **  
## H3EC38Not Asked                -0.8125   0.3972  -2.046 0.040776 *  
## H3EC38Yes                      -0.5263   0.1550  -3.395 0.000685 ***  
## BIO_SEX3Male                   0.7007   0.1117   6.274 3.52e-10 ***  
## H3D4AWhite                     -1.1196   0.1386  -8.079 6.56e-16 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for binomial family taken to be 1)  
##  
## Null deviance: 2985.6  on 4287  degrees of freedom  
## Residual deviance: 2686.8  on 4267  degrees of freedom  
## AIC: 2728.8  
##  
## Number of Fisher Scoring iterations: 5
```

4. Regression By Income Output:

```
##  
## Call:  
## glm(formula = H3ID15 ~ H3SP6 * LowIncome + H3FS12 + H3DA28 +  
##      H3EC38 + BIO_SEX3 + H3D4A, family = binomial, data = data)  
##  
## Coefficients:  
##                                     Estimate Std. Error z value Pr(>|z|)  
## (Intercept)                 1.7352    0.2748   6.315 2.70e-10 ***  
## H3SP6Most/all of the time     -0.1725    0.3396  -0.508 0.611444  
## H3SP6Never or rarely          1.4026    0.2131   6.580 4.70e-11 ***  
## H3SP6Sometimes                0.5135    0.2260   2.272 0.023062 *  
## LowIncome                    0.2792    0.6420   0.435 0.663594  
## H3FS12Friends more influential -0.4787    0.2144  -2.233 0.025558 *  
## H3FS12Not Asked               -0.3610    0.1620  -2.229 0.025811 *  
## H3DA28Yes                      0.3135    0.1123   2.791 0.005253 **  
## H3EC38Not Asked                -0.8029    0.3980  -2.018 0.043639 *  
## H3EC38Yes                      -0.5698    0.1514  -3.763 0.000168 ***  
## BIO_SEX3Male                   0.7397    0.1112   6.655 2.84e-11 ***  
## H3D4AWhite                     -1.0695    0.1358  -7.876 3.39e-15 ***  
## H3SP6Most/all of the time:LowIncome -1.0275    0.9238  -1.112 0.266049  
## H3SP6Never or rarely:LowIncome    -0.5000    0.6669  -0.750 0.453444  
## H3SP6Sometimes:LowIncome         0.1528    0.7231   0.211 0.832662  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for binomial family taken to be 1)  
##  
## Null deviance: 2985.6  on 4287  degrees of freedom  
## Residual deviance: 2720.9  on 4273  degrees of freedom  
## AIC: 2750.9  
##  
## Number of Fisher Scoring iterations: 5
```

5. Regression By Gender (Male) Output:

```
##  
## Call:  
## glm(formula = H3ID15 ~ H3FS12 + H3SP6 + H3WP53 + H3DA28 + H3EC38 +  
##      H3D4A, family = binomial, data = male_data)  
##  
## Coefficients:  
##                                     Estimate Std. Error z value Pr(>|z|)  
## (Intercept)                 2.60630   0.64052  4.069 4.72e-05 ***  
## H3FS12Friends more influential -0.31600   0.42404 -0.745 0.456146  
## H3FS12Not Asked              -0.41325   0.32011 -1.291 0.196707  
## H3SP6Most/all of the time    -0.40417   0.57430 -0.704 0.481580  
## H3SP6Never or rarely         1.37105   0.36677  3.738 0.000185 ***  
## H3SP6Sometimes               0.66992   0.39725  1.686 0.091718 .  
## H3WP53Not applicable / skipped -0.02455   0.43814 -0.056 0.955324  
## H3WP53Not close at all       -1.98211   0.75217 -2.635 0.008409 **  
## H3WP53Not very close        -0.25378   0.68556 -0.370 0.711255  
## H3WP53Quite close            -0.03774   0.58267 -0.065 0.948354  
## H3WP53Somewhat close         -0.70892   0.52784 -1.343 0.179252  
## H3DA28Yes                   0.49182   0.20899  2.353 0.018609 *  
## H3EC38Not Asked              -1.66290   0.58713 -2.832 0.004622 **  
## H3EC38Yes                   -0.66702   0.39926 -1.671 0.094793 .  
## H3D4AWhite                  -1.21154   0.26566 -4.560 5.10e-06 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for binomial family taken to be 1)  
##  
## Null deviance: 984.84  on 1973  degrees of freedom  
## Residual deviance: 913.10  on 1959  degrees of freedom  
## AIC: 943.1  
##  
## Number of Fisher Scoring iterations: 6
```

6. Regression By Gender (Female) Output:

```
##  
## Call:  
## glm(formula = H3ID15 ~ H3FS12 + H3SP6 + H3WP53 + H3DA28 + H3EC38 +  
##      H3D4A, family = binomial, data = female_data)  
##  
## Coefficients:  
##                                     Estimate Std. Error z value Pr(>|z|)  
## (Intercept)                 2.1784   0.4314  5.050 4.42e-07 ***  
## H3FS12Friends more influential -0.4834   0.2510 -1.926 0.054073 .  
## H3FS12Not Asked              -0.2702   0.1895 -1.426 0.153979  
## H3SP6Most/all of the time    -0.3469   0.3790 -0.915 0.359957  
## H3SP6Never or rarely         1.2780   0.2463  5.188 2.12e-07 ***  
## H3SP6Sometimes               0.4509   0.2593  1.739 0.082062 .  
## H3WP53Not applicable / skipped -0.3334   0.3102 -1.075 0.282524  
## H3WP53Not close at all       -1.3622   0.4491 -3.033 0.002421 **  
## H3WP53Not very close         -1.3113   0.3827 -3.426 0.000612 ***  
## H3WP53Quite close            -0.3728   0.3955 -0.943 0.345869  
## H3WP53Somewhat close          -0.1962   0.3949 -0.497 0.619268  
## H3DA28Yes                   0.2822   0.1338  2.109 0.034926 *  
## H3EC38Not Asked              -0.4148   0.5222 -0.794 0.426996  
## H3EC38Yes                   -0.5239   0.1666 -3.145 0.001659 **  
## H3D4AWhite                  -1.0926   0.1622 -6.738 1.61e-11 ***  
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## (Dispersion parameter for binomial family taken to be 1)  
##  
## Null deviance: 1931.5  on 2313  degrees of freedom  
## Residual deviance: 1773.7  on 2299  degrees of freedom  
## AIC: 1803.7  
##  
## Number of Fisher Scoring iterations: 5
```

Appendix A5: Final Variable Selection From Add-Health Data

Category	Variable Code	Label	Source
Target Variable	H3ID15	Have you ever been diagnosed with depression?	DS0008
Friends & Family	H3DA15	Times hung out with friends or talked > 5 min in past 7 days	DS0008
Friends & Family	H3FS12	Friends influenced more than family or vice versa (past 5 years)	DS0008
Friends & Family	H3FS11	Still friends with high school friends?	DS0008
Friends & Family	H3SP6	Could not shake the blues, even with help (past 7 days)	DS0008
Friends & Family	H3WP46	Closeness to biological mother	DS0008
Friends & Family	H3WP53	Closeness to biological father	DS0008
Friends & Family	H3WP57	Do biological parents live together?	DS0008
Socioeconomic Factors	H3DA28	Do you currently have a job?	DS0008
Socioeconomic Factors	H3EC7	Total household income before taxes (self-reported)	DS0008
Socioeconomic Factors	H3EC3	Total personal income before taxes (self-reported)	DS0008
Socioeconomic Factors	H3DA36	Currently enrolled in school/training?	DS0008
Socioeconomic Factors	H3EC26	Receiving AFDC/public assistance/welfare?	DS0008
Socioeconomic Factors	H3EC38	Ever received other public assistance/welfare?	DS0008
Demographics	BIO_SEX3	Respondent's Gender	DS0008
Demographics	H3OD4A	Race: White	DS0008

Healthcare Analytics Individual Assignment

R Code

Part 0 - Data Import

Importing Libraries

```
library(haven)
library(ggplot2)
```

Extraction of Dataset and Storing it

```
# Define the directory where your files are stored
data_dir <- "/Users/harsh/Downloads/ICPSR_21600"

# Initialize a list to store the datasets
data_list <- list()

# Loop through DS0001 to DS0042
for (i in 1:42) {

  # Format the dataset number with leading zeros (e.g., DS0001)
  ds_num <- sprintf("DS%04d", i)

  # Build the correct file name (note the difference in pattern)
  if (i < 10) {
    file_name <- paste0("21600-000", i, "-Data.dta")
  } else {
    file_name <- paste0("21600-00", i, "-Data.dta")
  }

  # Build the full file path
  file_path <- file.path(data_dir, ds_num, file_name)

  # Print progress (optional)
  cat("Reading:", file_path, "\n")

  # Check if the file exists to avoid crashing
  if (file.exists(file_path)) {
    data_list[[ds_num]] <- read_dta(file_path)
  } else {
    cat("File not found:", file_path, "\n")
  }
}
```

```

## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0001/21600-0001-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0002/21600-0002-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0003/21600-0003-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0004/21600-0004-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0005/21600-0005-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0006/21600-0006-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0007/21600-0007-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0008/21600-0008-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0009/21600-0009-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0010/21600-0010-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0011/21600-0011-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0012/21600-0012-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0013/21600-0013-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0014/21600-0014-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0015/21600-0015-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0016/21600-0016-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0017/21600-0017-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0018/21600-0018-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0019/21600-0019-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0020/21600-0020-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0021/21600-0021-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0022/21600-0022-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0023/21600-0023-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0024/21600-0024-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0025/21600-0025-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0026/21600-0026-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0027/21600-0027-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0028/21600-0028-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0029/21600-0029-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0030/21600-0030-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0031/21600-0031-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0032/21600-0032-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0033/21600-0033-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0034/21600-0034-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0035/21600-0035-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0036/21600-0036-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0037/21600-0037-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0038/21600-0038-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0039/21600-0039-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0040/21600-0040-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0041/21600-0041-Data.dta
## Reading: /Users/harsh/Downloads/ICPSR_21600/DS0042/21600-0042-Data.dta

```

```

# Check the loaded datasets
names(data_list)

```

```

## [1] "DS0001" "DS0002" "DS0003" "DS0004" "DS0005" "DS0006" "DS0007" "DS0008"
## [9] "DS0009" "DS0010" "DS0011" "DS0012" "DS0013" "DS0014" "DS0015" "DS0016"
## [17] "DS0017" "DS0018" "DS0019" "DS0020" "DS0021" "DS0022" "DS0023" "DS0024"
## [25] "DS0025" "DS0026" "DS0027" "DS0028" "DS0029" "DS0030" "DS0031" "DS0032"
## [33] "DS0033" "DS0034" "DS0035" "DS0036" "DS0037" "DS0038" "DS0039" "DS0040"
## [41] "DS0041" "DS0042"

```

List of Target Variables

```
vars <- c(
  "AID", "H3ID15", # ID + target
  "H3DA15", "H3FS12", "H3FS11", "H3SP6", "H3WP46", "H3WP53", "H3WP57", # Friends & Family
  "H3DA28", "H3EC7", "H3EC3", "H3DA36", "H3EC26", "H3EC38", # Socioeconomic
  "BI0_SEX3", "H30D4A" # Demographics
)
```

Saving the Subset of Dataset

```
wave3_data <- read_dta("/Users/harsh/Downloads/ICPSR_21600/DS0008/21600-0008-Data.dta")

data_subset <- wave3_data[ , vars]

str(data_subset)
```

```

## tibble [4,882 x 17] (S3: tbl_df/tbl/data.frame)
## $ AID      : chr [1:4882] "57100270" "57101310" "57103869" "57104676" ...
## ..- attr(*, "label")= chr "RESPONDENT IDENTIFIER"
## ..- attr(*, "format.stata")= chr "%9s"
## $ H3ID15   : dbl+lbl [1:4882] 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, ...
## ..@ label    : chr "S11Q15 EVER BEEN DX WITH DEPRESSION-W3"
## ..@ format.stata: chr "%18.0f"
## ..@ labels   : Named num [1:4] 0 1 8 9
## ... - attr(*, "names")= chr [1:4] "(0) No" "(1) Yes" "(8) Don't know" "(9) Not
applicable"
## $ H3DA15   : dbl+lbl [1:4882] 0, 7, 6, 0, 7, 3, 1, 7, 0, 2, 4, 4, 6, 0, 2, 5, 2,
0, ...
## ..@ label    : chr "S33Q15 HANG OUT WITH FRIENDS - W3"
## ..@ format.stata: chr "%19.0f"
## ..@ labels   : Named num [1:10] 0 1 2 3 4 5 6 7 96 98
## ... - attr(*, "names")= chr [1:10] "(0) Not at all" "(1) 1 time" "(2) 2 times"
"(3) 3 times" ...
## $ H3FS12   : dbl+lbl [1:4882] 5, 5, 5, 1, 1, 7, 5, 5, 5, 2, 7, 5, 5, 5, 5, 2,
5, ...
## ..@ label    : chr "S6Q12 FRIEND OR FAMILY INFLUENCE MORE-W3"
## ..@ format.stata: chr "%41.0f"
## ..@ labels   : Named num [1:6] 1 2 5 7 8 9
## ... - attr(*, "names")= chr [1:6] "(1) Your friends have influenced you more"
"(2) Your family has influenced you more" "(5) Question not asked of this respondent"
"(7) Legitimate skip" ...
## $ H3FS11   : dbl+lbl [1:4882] 95, 95, 95, 5, 3, 97, 95, 95, 95, 0, 97, 95, 95,
9...
## ..@ label    : chr "S6Q11 STILL FRIENDS W/HS FRIENDS-W3"
## ..@ format.stata: chr "%42.0f"
## ..@ labels   : Named num [1:9] 0 1 2 3 4 5 95 97 98
## ... - attr(*, "names")= chr [1:9] "(0) None" "(1) One" "(2) A few" "(3) Some"
...
## $ H3SP6    : dbl+lbl [1:4882] 1, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 3, 0, 0,
2, ...
## ..@ label    : chr "S12Q1 PAST 7 DAYS SHAKE OFF BLUES-W3"
## ..@ format.stata: chr "%36.0f"
## ..@ labels   : Named num [1:7] 0 1 2 3 6 8 9
## ... - attr(*, "names")= chr [1:7] "(0) Never/rarely" "(1) Sometimes" "(2) A lo
t of the time" "(3) Most of the time/all of the time" ...
## $ H3WP46   : dbl+lbl [1:4882] 97, 97, 97, 97, 97, 97, 97, 97, 97, 97, 97, 97,
97, ...
## ..@ label    : chr "S3Q46 CLOSE TO BIO MOM-W3"
## ..@ format.stata: chr "%20.0f"
## ..@ labels   : Named num [1:6] 1 2 3 4 5 97
## ... - attr(*, "names")= chr [1:6] "(1) Extremely close" "(2) Quite close" "(3)
Somewhat close" "(4) Not very close" ...
## $ H3WP53   : dbl+lbl [1:4882] 1, 97, 1, 97, 2, 97, 97, 97, 97, 5, 97, 97, 97,
...
## ..@ label    : chr "S3Q53 CLOSE TO BIO DAD-W3"
## ..@ format.stata: chr "%20.0f"
## ..@ labels   : Named num [1:7] 1 2 3 4 5 97 99
## ... - attr(*, "names")= chr [1:7] "(1) Extremely close" "(2) Quite close" "(3)
Somewhat close" "(4) Not very close" ...
## $ H3WP57   : dbl+lbl [1:4882] 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7,

```

```

7, ...
##   ..@ label      : chr "S3Q57 BIO DAD/BIO MOM LIVE TOGETHER-W3"
##   ..@ format.stata: chr "%19.0f"
##   ..@ labels     : Named num [1:5] 0 1 7 8 9
##   ... -- attr(*, "names")= chr [1:5] "(0) No" "(1) Yes" "(7) Legitimate skip" "
##   (8) Don't know" ...
## $ H3DA28 : dbl+lbl [1:4882] 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1,
0, ...
##   ..@ label      : chr "S33Q28 CURRENTLY HAVE JOB-W3"
##   ..@ format.stata: chr "%21.0f"
##   ..@ labels     : Named num [1:5] 0 1 6 8 9
##   ... -- attr(*, "names")= chr [1:5] "(0) No (skip to Q.36)" "(1) Yes" "(6) Refused"
##   "(8) Don't know" ...
## $ H3EC7  : dbl+lbl [1:4882] 97, 97, 97, 97, 97, 8, 97, 97, 97, 97, 97, 97, 97,
9...
##   ..@ label      : chr "S15Q7 BEST GUESS HOUSEHOLD INC B TAX-W3"
##   ..@ format.stata: chr "%21.0f"
##   ..@ labels     : Named num [1:12] 1 2 3 4 5 6 7 8 96 97 ...
##   ... -- attr(*, "names")= chr [1:12] "(1) Less than $10,000" "(2) $10,000-$14,99
9" "(3) $15,000-$19,999" "(4) $20,000-$29,000" ...
## $ H3EC3  : dbl+lbl [1:4882] 97, 97, 97, 97, 97, 1, 97, 97, 97, 98, 97, 97,
9...
##   ..@ label      : chr "S15Q3 BEST GUESS INCOME BEFORE TAX-W3"
##   ..@ format.stata: chr "%21.0f"
##   ..@ labels     : Named num [1:12] 1 2 3 4 5 6 7 8 96 97 ...
##   ... -- attr(*, "names")= chr [1:12] "(1) Less than $10,000" "(2) $10,000-$14,99
9" "(3) $15,000-$19,999" "(4) $20,000-$29,000" ...
## $ H3DA36 : dbl+lbl [1:4882] 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0,
0, ...
##   ..@ label      : chr "S33Q36 ENROLLED SCHOOL/VOC TRAIN-W3"
##   ..@ format.stata: chr "%21.0f"
##   ..@ labels     : Named num [1:5] 0 1 6 8 9
##   ... -- attr(*, "names")= chr [1:5] "(0) No (skip to Q.43)" "(1) Yes" "(6) Refused"
##   "(8) Don't know" ...
## $ H3EC26 : dbl+lbl [1:4882] 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
0, ...
##   ..@ label      : chr "S15Q26 CURRENTLY GETTING AFDC-W3"
##   ..@ format.stata: chr "%19.0f"
##   ..@ labels     : Named num [1:6] 0 1 6 7 8 9
##   ... -- attr(*, "names")= chr [1:6] "(0) No" "(1) Yes" "(6) Refused" "(7) Legiti
mate skip" ...
## $ H3EC38 : dbl+lbl [1:4882] 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, ...
##   ..@ label      : chr "S15Q38 RECEIVED OTH PUB ASSISTANCE-W3"
##   ..@ format.stata: chr "%41.0f"
##   ..@ labels     : Named num [1:7] 0 1 5 6 7 8 9
##   ... -- attr(*, "names")= chr [1:7] "(0) No (skip to Q.53)" "(1) Yes" "(5) Quest
ion not asked of this respondent" "(6) Refused" ...
## $ BIO_SEX3: dbl+lbl [1:4882] 2, 2, 1, 1, 1, 1, 1, 1, 2, 2, 2, 1, 1, 2, 2, 2,
2, ...
##   ..@ label      : chr "BIOLOGICAL SEX-W3"
##   ..@ format.stata: chr "%10.0f"
##   ..@ labels     : Named num [1:2] 1 2
##   ... -- attr(*, "names")= chr [1:2] "(1) Male" "(2) Female"
## $ H30D4A : dbl+lbl [1:4882] 0, 0, 0, 0, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,
1, ...

```

```

## ..@ label      : chr "S1Q4A RACE-WHITE-W3"
## ..@ format.stata: chr "%18.0f"
## ..@ labels     : Named num [1:5] 0 1 6 8 9
## ...- attr(*, "names")= chr [1:5] "(0) Not marked" "(1) Marked" "(6) Refused"
## "(8) Don't know" ...
## - attr(*, "label")= chr "National Longitudinal Study of Adolescent to Adult Health (Add Health), 1994-2000"

```

```
summary(data_subset)
```

```

##      AID          H3ID15        H3DA15        H3FS12
## Length:4882    Min.   :0.0000   Min.   : 0.0   Min.   :1.000
## Class  :character 1st Qu.:0.0000  1st Qu.: 3.0   1st Qu.:5.000
## Mode   :character Median :0.0000  Median : 5.0   Median :5.000
##                   Mean   :0.1247  Mean   : 4.8   Mean   :4.404
##                   3rd Qu.:0.0000  3rd Qu.: 7.0   3rd Qu.:5.000
##                   Max.   :9.0000  Max.   :98.0   Max.   :9.000
##
##      H3FS11        H3SP6         H3WP46        H3WP53
## Min.   : 0.00  Min.   :0.0000  Min.   : 1.00  Min.   : 1.00
## 1st Qu.:95.00  1st Qu.:0.0000  1st Qu.:97.00  1st Qu.:97.00
## Median :95.00  Median :0.0000  Median :97.00  Median :97.00
## Mean   :74.88  Mean   :0.3343  Mean   :89.43  Mean   :75.92
## 3rd Qu.:95.00  3rd Qu.:0.0000  3rd Qu.:97.00  3rd Qu.:97.00
## Max.   :98.00  Max.   :9.0000  Max.   :97.00  Max.   :99.00
##
##      H3WP57        H3DA28        H3EC7         H3EC3
## Min.   :0.000  Min.   :0.0000  Min.   : 1.00  Min.   : 1.00
## 1st Qu.:7.000  1st Qu.:0.0000  1st Qu.:97.00  1st Qu.:97.00
## Median :7.000  Median :1.0000  Median :97.00  Median :97.00
## Mean   :6.736  Mean   :0.7524  Mean   :87.05  Mean   :82.46
## 3rd Qu.:7.000  3rd Qu.:1.0000  3rd Qu.:97.00  3rd Qu.:97.00
## Max.   :9.000  Max.   :9.0000  Max.   :98.00  Max.   :99.00
##                   NA's   :5       NA's   :5
##
##      H3DA36        H3EC26        H3EC38        BI0_SEX3
## Min.   :0.0000  Min.   :0.0000  Min.   :0.0000  Min.   :1.000
## 1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:0.0000  1st Qu.:1.000
## Median :0.0000  Median :0.0000  Median :0.0000  Median :2.000
## Mean   :0.3863  Mean   :0.0609  Mean   :0.1833  Mean   :1.539
## 3rd Qu.:1.0000  3rd Qu.:0.0000  3rd Qu.:0.0000  3rd Qu.:2.000
## Max.   :9.0000  Max.   :8.0000  Max.   :9.0000  Max.   :2.000
##                   NA's   :5       NA's   :5
##
##      H3OD4A
## Min.   :0.0000
## 1st Qu.:0.0000
## Median :1.0000
## Mean   :0.7552
## 3rd Qu.:1.0000
## Max.   :9.0000
##

```

```
colSums(is.na(data_subset))
```

```

##      AID    H3ID15    H3DA15    H3FS12    H3FS11    H3SP6    H3WP46    H3WP53
##      0        0        0        0        0        0        0        0        0
##  H3WP57    H3DA28    H3EC7     H3EC3    H3DA36    H3EC26    H3EC38  BIO_SEX3
##      0        0        5        5        0        5        5        0
##  H30D4A
##      0

```

```

cleaned_data <- na.omit(data_subset)

write.csv(cleaned_data, "cleaned_addhealth_data.csv", row.names = FALSE)

```

Dataset Preview

```
head(cleaned_data)
```

```

## # A tibble: 6 × 17
##   AID    H3ID15    H3DA15    H3FS12    H3FS11    H3SP6    H3WP46    H3WP53    H3WP57
##   <chr>  <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl> <dbl+lbl>
## 1 57100270 0 [(0) No] 0 [(0)... 5 [(5)... 95 [(95... 1 [(1)... 97 [(97... 1 [(1)... 7 [(7)...
## 2 57101310 0 [(0) No] 7 [(7)... 5 [(5)... 95 [(95... 1 [(1)... 97 [(97... 97 [(97... 7 [(7)...
## 3 57103869 0 [(0) No] 6 [(6)... 5 [(5)... 95 [(95... 0 [(0)... 97 [(97... 1 [(1)... 7 [(7)...
## 4 57104676 0 [(0) No] 0 [(0)... 1 [(1)... 5 [(5)... 0 [(0)... 97 [(97... 97 [(97... 7 [(7)...
## 5 57109625 0 [(0) No] 7 [(7)... 1 [(1)... 3 [(3)... 1 [(1)... 97 [(97... 2 [(2)... 7 [(7)...
## 6 57111071 0 [(0) No] 3 [(3)... 7 [(7)... 97 [(97... 0 [(0)... 97 [(97... 97 [(97... 7 [(7)...
## # i 8 more variables: H3DA28 <dbl+lbl>, H3EC7 <dbl+lbl>, H3EC3 <dbl+lbl>,
## #   H3DA36 <dbl+lbl>, H3EC26 <dbl+lbl>, H3EC38 <dbl+lbl>, BIO_SEX3 <dbl+lbl>,
## #   H30D4A <dbl+lbl>

```

Part 1 - Dataset Preparation

Re-labelling

A) Target Variable

```

# Recode H3ID15 – Have you ever been diagnosed with depression?
# 0 = No, 1 = Yes, all others = NA

cleaned_data$H3ID15_recode <- ifelse(cleaned_data$H3ID15 %in% c(0, 1),
                                         cleaned_data$H3ID15,
                                         NA)

# Convert to labeled factor
cleaned_data$H3ID15_recode <- factor(cleaned_data$H3ID15_recode,
                                         levels = c(0, 1),
                                         labels = c("No Depression", "Diagnosed Depression"))

```

B) Independent Variable

B1) Friends & Family

"Yes",
"Not applicable / skipped"))

B2) Socioeconomic Factors


```

        "$40,000–49,999",
        "$50,000–74,999",
        "$75,000 or more",
        "Not applicable / skipped"),
ordered = TRUE)

#4
# Recode H3DA36 – Enrolled in school/job training/vocational education
# 0 = No, 1 = Yes, others = NA

cleaned_data$H3DA36_recode <- ifelse(cleaned_data$H3DA36 %in% c(0, 1),
                                      cleaned_data$H3DA36,
                                      NA)

# Convert to labeled factor
cleaned_data$H3DA36_recode <- factor(cleaned_data$H3DA36_recode,
                                       levels = c(0, 1),
                                       labels = c("No", "Yes"))

#5
# Recode H3EC26 – Currently receiving AFDC/public assistance/welfare?
# 0 = No, 1 = Yes, others = NA

cleaned_data$H3EC26_recode <- ifelse(cleaned_data$H3EC26 %in% c(0, 1),
                                       cleaned_data$H3EC26,
                                       NA)

# Convert to labeled factor
cleaned_data$H3EC26_recode <- factor(cleaned_data$H3EC26_recode,
                                       levels = c(0, 1),
                                       labels = c("No", "Yes"))

#6
# Recode H3EC38 – Ever received public assistance (excluding food stamps)?
# 0 = No, 1 = Yes, 5 = Not Asked, others = NA

cleaned_data$H3EC38_recode <- ifelse(cleaned_data$H3EC38 %in% c(0, 1, 5),
                                       cleaned_data$H3EC38,
                                       NA)

# Convert to labeled factor
cleaned_data$H3EC38_recode <- factor(cleaned_data$H3EC38_recode,
                                       levels = c(0, 1, 5),
                                       labels = c("No", "Yes", "Not Asked"))

```

B3) Demographics

```
#1
# Recode BI0_SEX3 – Respondent's gender
# 1 = Male, 2 = Female

cleaned_data$BI0_SEX3_recode <- factor(cleaned_data$BI0_SEX3,
                                         levels = c(1, 2),
                                         labels = c("Male", "Female"))

#2
# Recode H30D4A – Race: White
# 0 = Not White, 1 = White, others = NA

cleaned_data$H30D4A_recode <- ifelse(cleaned_data$H30D4A %in% c(0, 1),
                                       cleaned_data$H30D4A,
                                       NA)

# Convert to labeled factor
cleaned_data$H30D4A_recode <- factor(cleaned_data$H30D4A_recode,
                                       levels = c(0, 1),
                                       labels = c("Not White", "White"))
```

Saving the Recoded Variables as a separate dataset

```

# Create a new data frame with only recoded variables and AID for reference
recoded_data <- cleaned_data[, c(
  "AID",
  # Target
  "H3ID15_recode",
  # Friends & Family
  "H3DA15_recode",
  "H3FS12_recode",
  "H3FS11_recode",
  "H3SP6_recode",
  "H3WP46_recode",
  "H3WP53_recode",
  "H3WP57_recode",
  # SES
  "H3DA28_recode",
  "H3EC7_recode",
  "H3EC3_recode",
  "H3DA36_recode",
  "H3EC26_recode",
  "H3EC38_recode",
  # Demographics
  "BI0_SEX3_recode",
  "H30D4A_recode"
)]

```

Create a complete-case dataset (drop rows with any NA)

```

complete_data <- na.omit(recoded_data)

```

Save to CSV

```

write.csv(complete_data, "complete_addhealth_data.csv", row.names = FALSE)

```

Finding NA in each Variable

```

# Count NA values in each variable of your recoded dataset
colSums(is.na(complete_data))

```

	AID	H3ID15_recode	H3DA15_recode	H3FS12_recode	H3FS11_recode
##	0	0	0	0	0
##	H3SP6_recode	H3WP46_recode	H3WP53_recode	H3WP57_recode	H3DA28_recode
##	0	0	0	0	0
##	H3EC7_recode	H3EC3_recode	H3DA36_recode	H3EC26_recode	H3EC38_recode
##	0	0	0	0	0
##	BI0_SEX3_recode	H30D4A_recode			
##	0	0			

Part 2 - Visualisation

```
# Load libraries
library(tidyverse)

## — Attaching core tidyverse packages —————— tidyverse 2.0.0 —
## ✓ dplyr    1.1.4    ✓ readr    2.1.5
## ✓ forcats  1.0.0    ✓ stringr  1.5.1
## ✓ lubridate 1.9.4    ✓ tibble   3.2.1
## ✓ purrr   1.0.4    ✓ tidyrr   1.3.1
## — Conflicts —————— tidyverse_conflicts() —
## ✘ dplyr::filter() masks stats::filter()
## ✘ dplyr::lag()   masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
# Step 1: Read and clean the dataset
data <- read.csv("complete_addhealth_data.csv", stringsAsFactors = TRUE)
names(data) <- gsub("_recode$", "", names(data))

# Step 2: Make sure key variables are treated as factors
categorical_vars <- c("H3ID15", "BI0_SEX3", "H3FS11", "H3FS12", "H3SP6",
                      "H3WP46", "H3WP53", "H3WP57", "H3DA28", "H3EC7", "H3EC3",
                      "H3DA36", "H3EC26", "H3EC38", "H3OD4A")

data[categorical_vars] <- lapply(data[categorical_vars], as.factor)

# Step 3: Summary stats for numeric variable
summary(data$H3DA15)
```

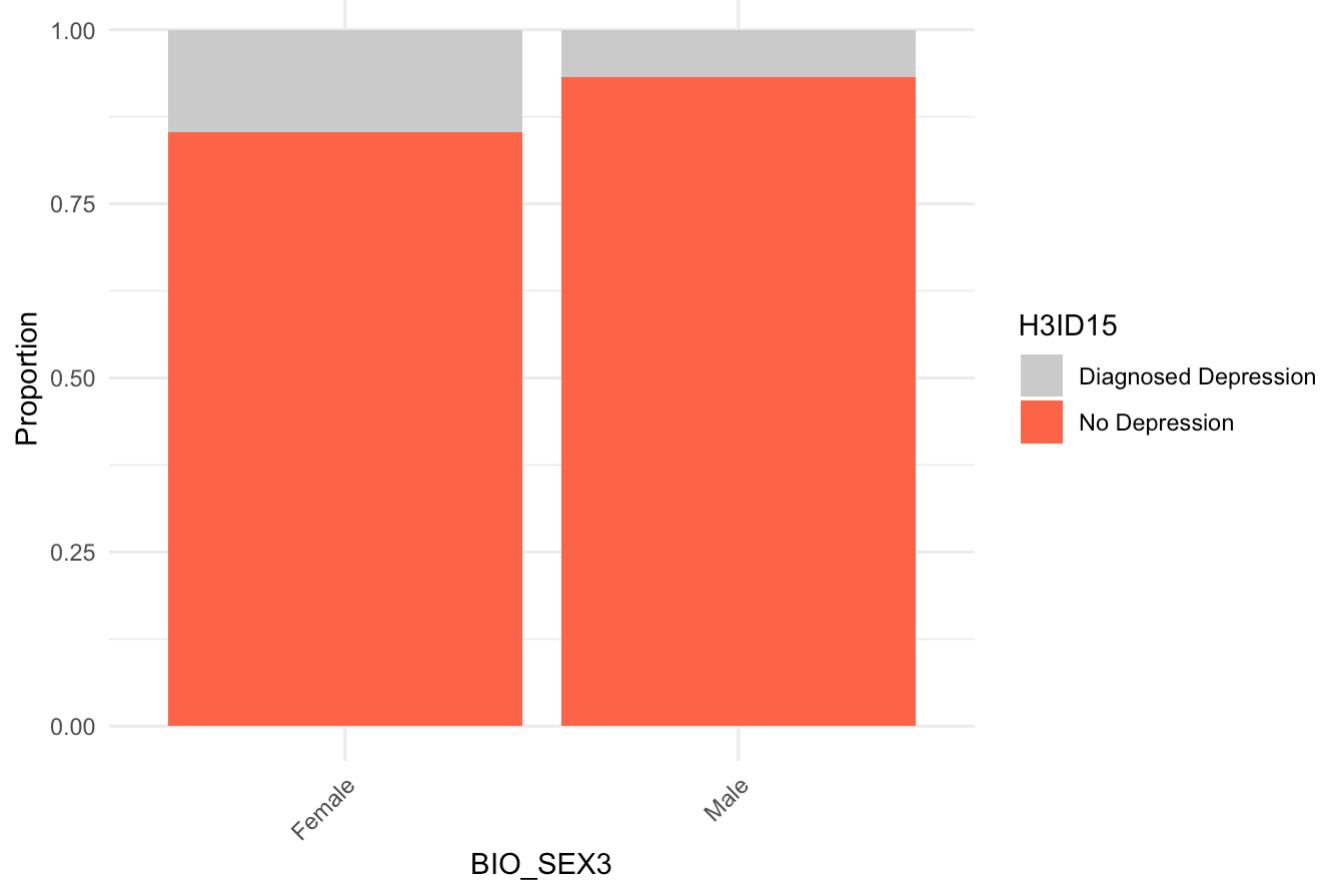
```
##      Min. 1st Qu. Median     Mean 3rd Qu.    Max.
##      0.000  3.000  4.000  4.389  7.000  7.000
```

```
# Step 4: Bar plots – Bivariate with Depression
plot_vars <- categorical_vars[categorical_vars != "H3ID15"] # exclude target

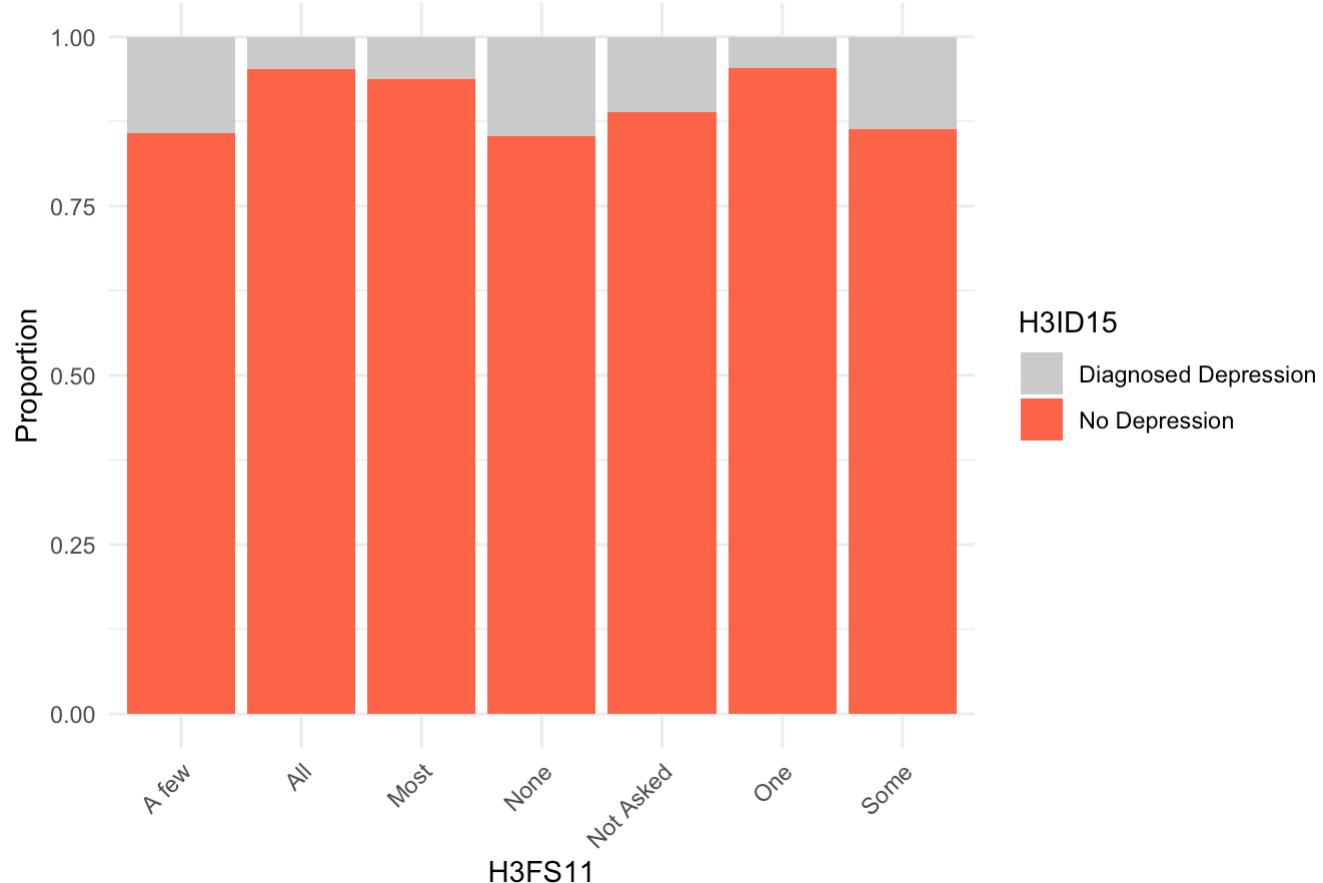
for (var in plot_vars) {
  print(
    ggplot(data, aes_string(x = var, fill = "H3ID15")) +
      geom_bar(position = "fill") +
      labs(title = paste("Depression by", var),
           x = var, y = "Proportion") +
      scale_fill_manual(values = c("gray80", "tomato")) +
      theme_minimal() +
      theme(axis.text.x = element_text(angle = 45, hjust = 1))
  )
}
```

```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.  
## i Please use tidy evaluation idioms with `aes()`.  
## i See also `vignette("ggplot2-in-packages")` for more information.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was  
## generated.
```

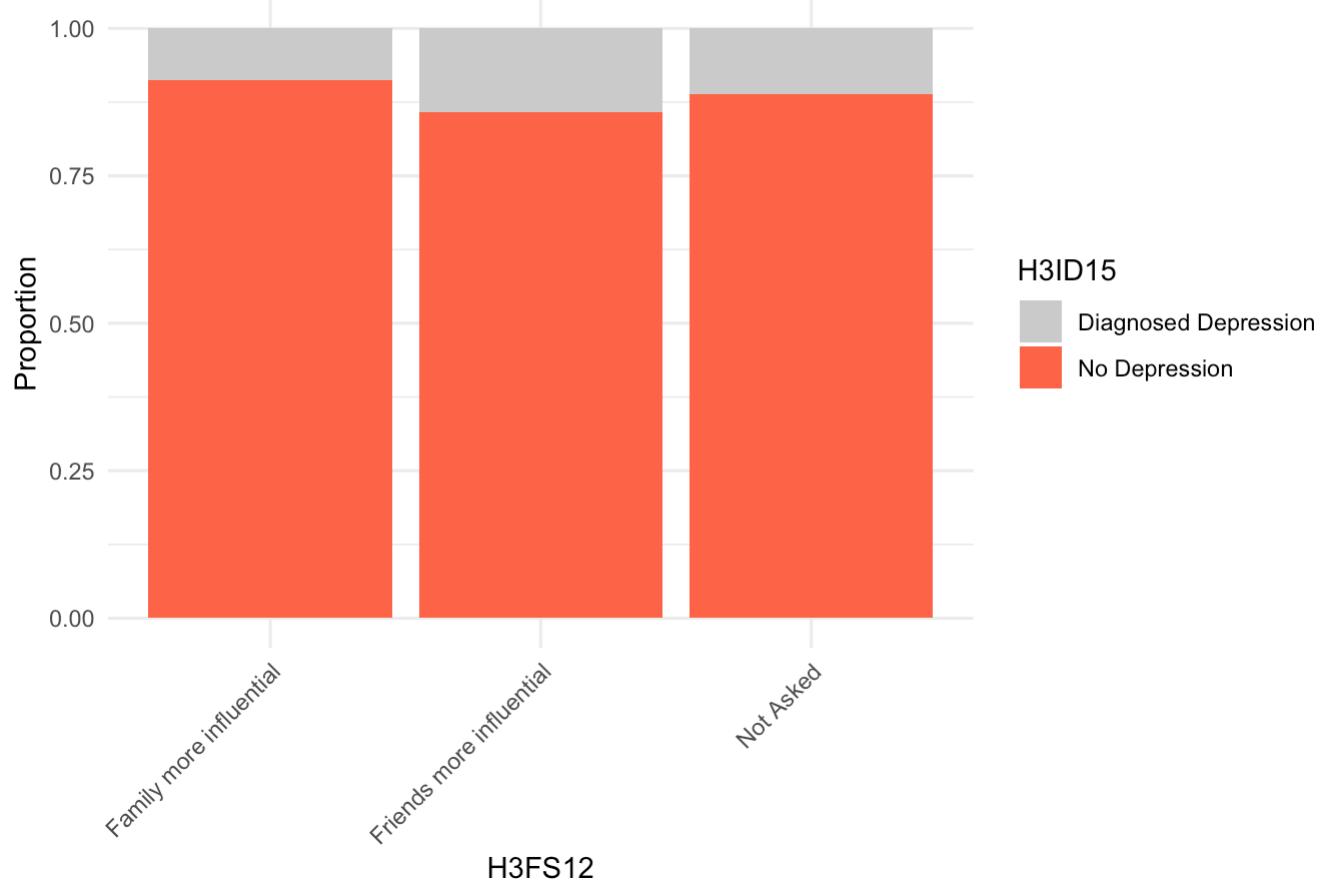
Depression by BIO_SEX3



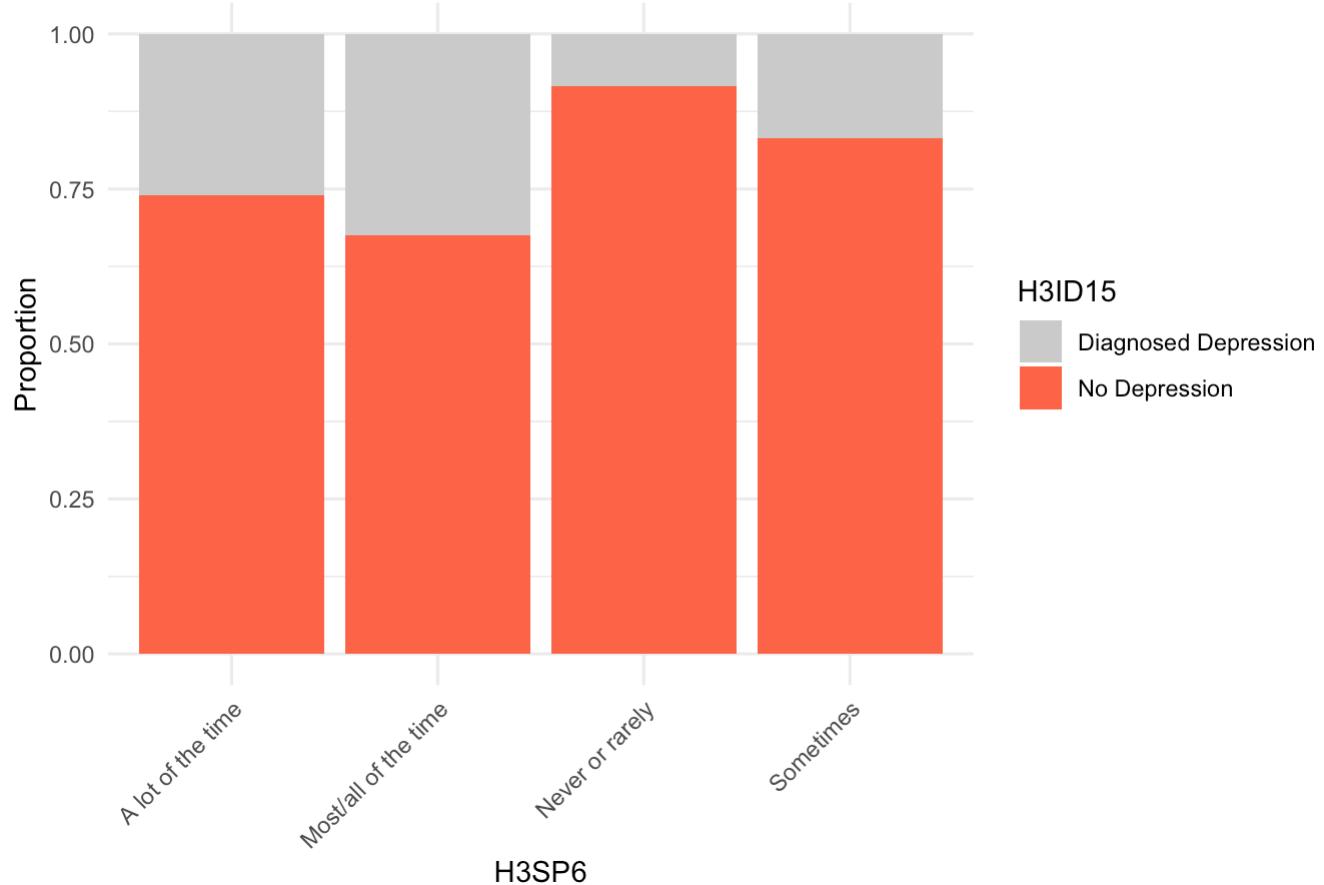
Depression by H3FS11



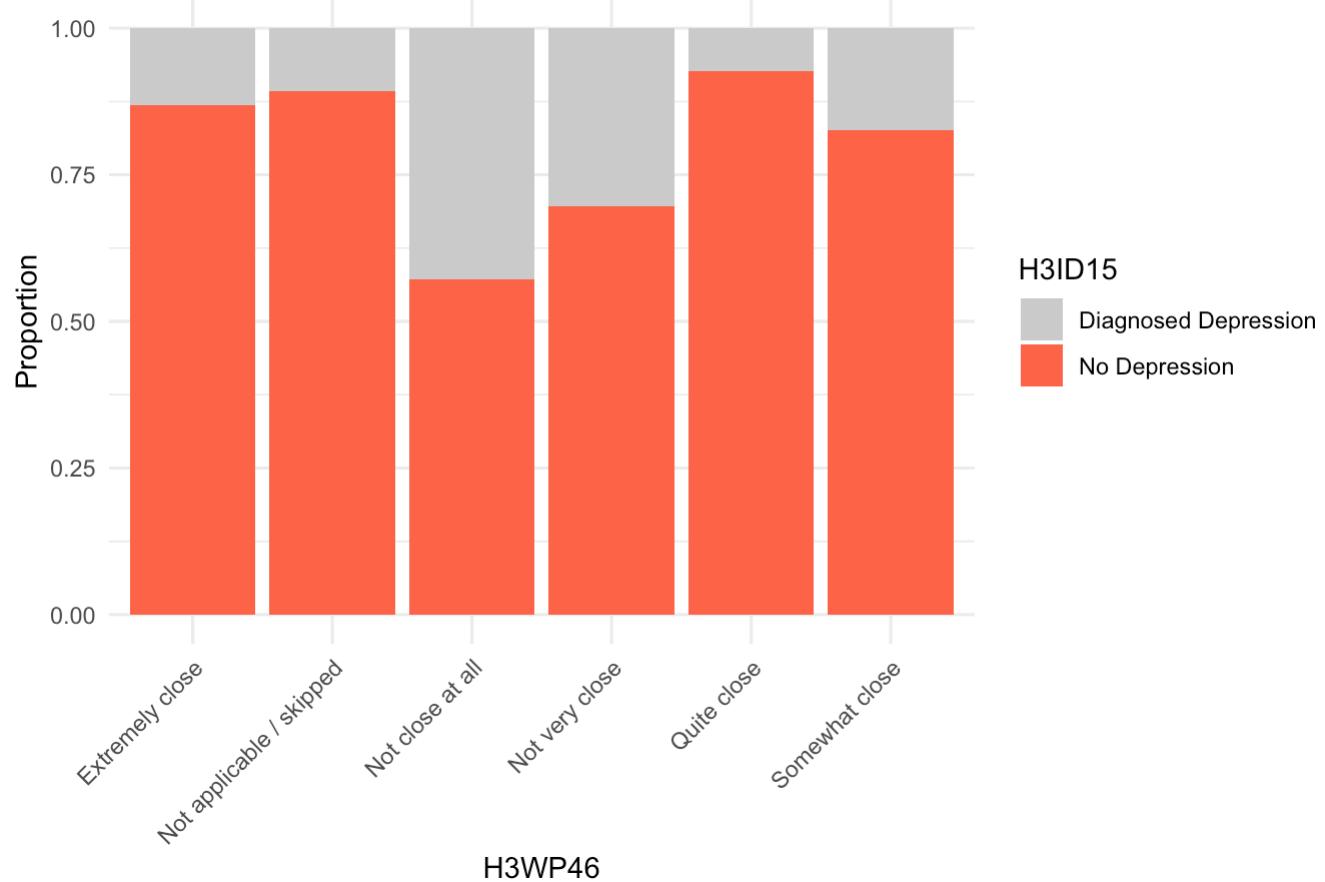
Depression by H3FS12



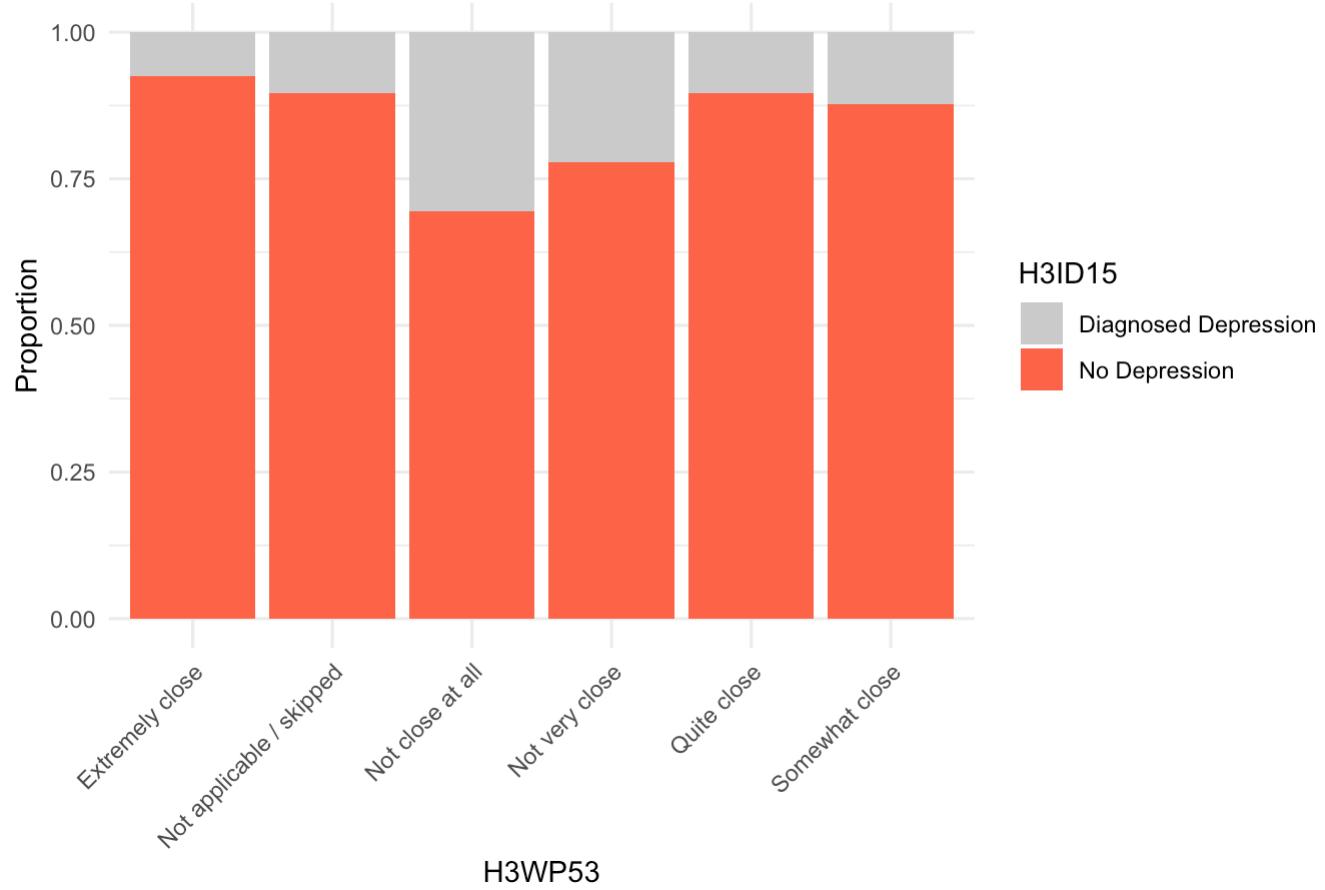
Depression by H3SP6



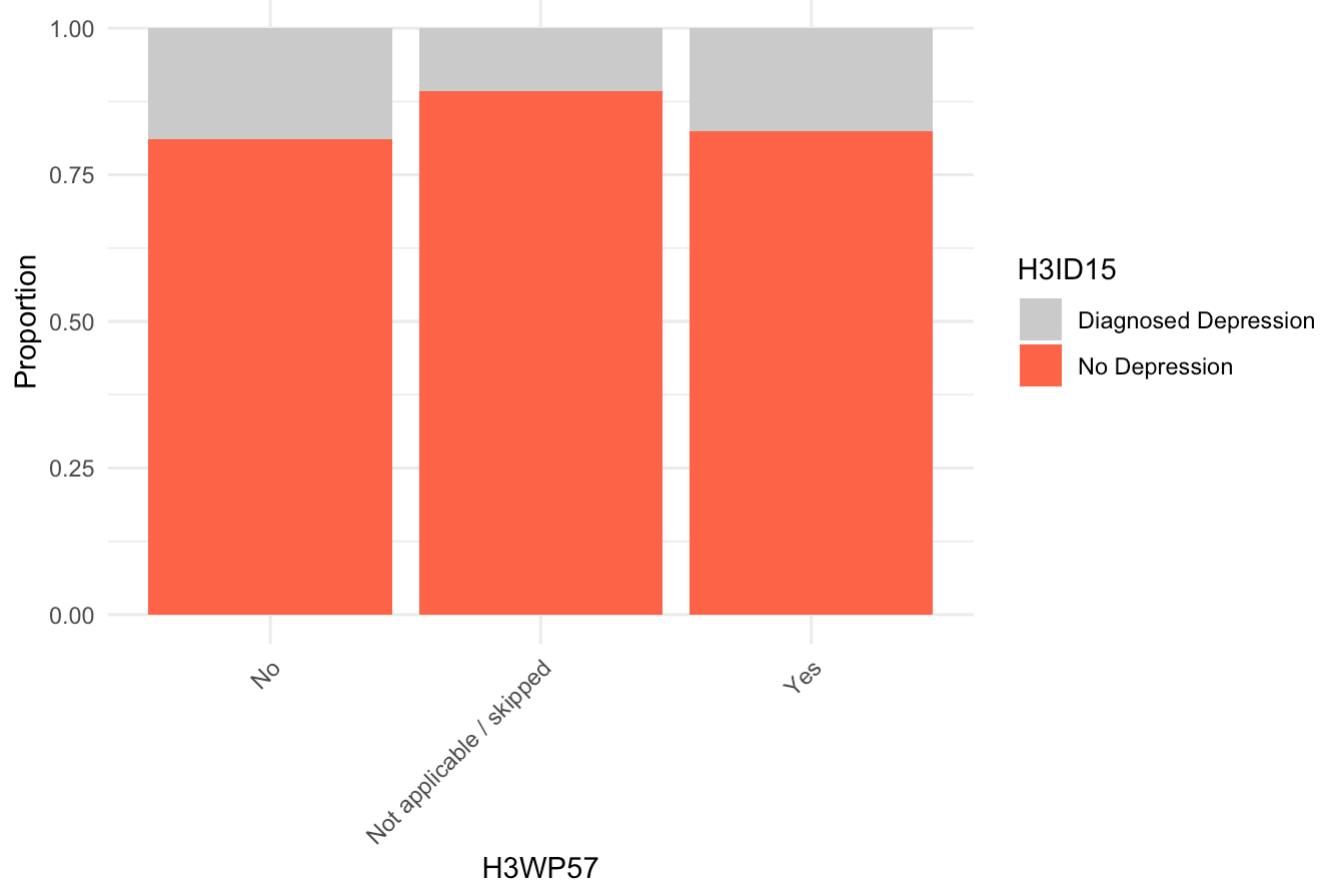
Depression by H3WP46



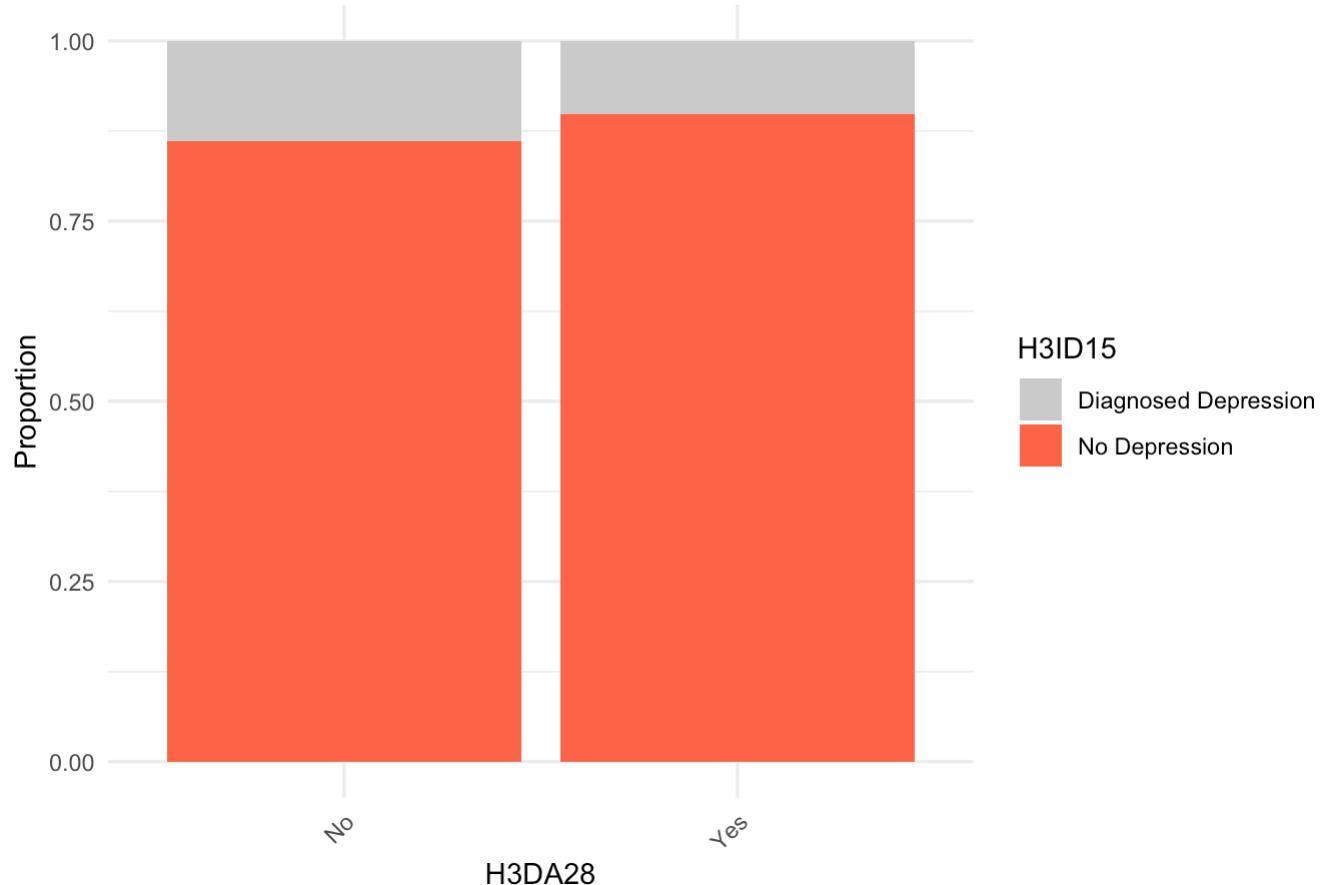
Depression by H3WP53



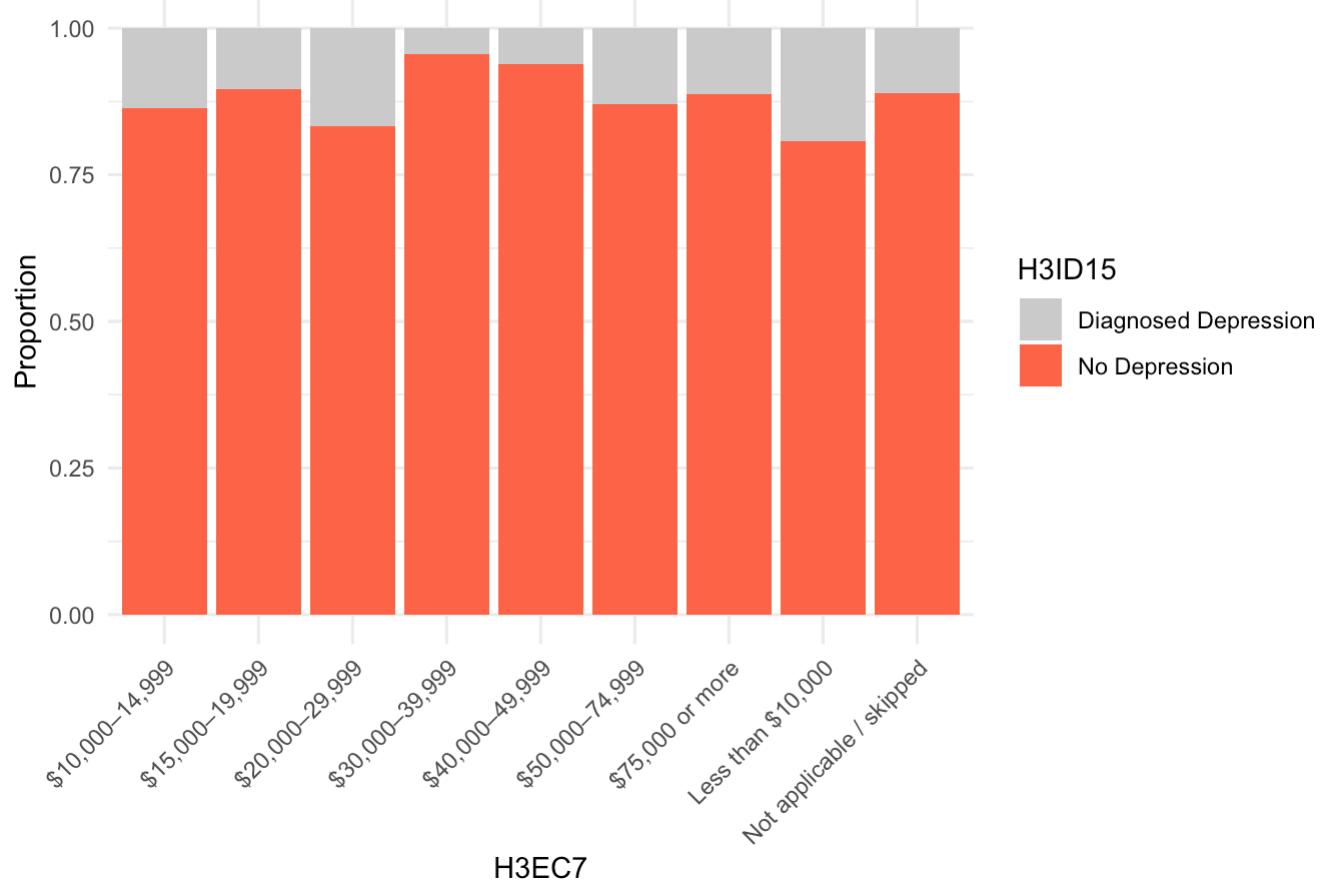
Depression by H3WP57



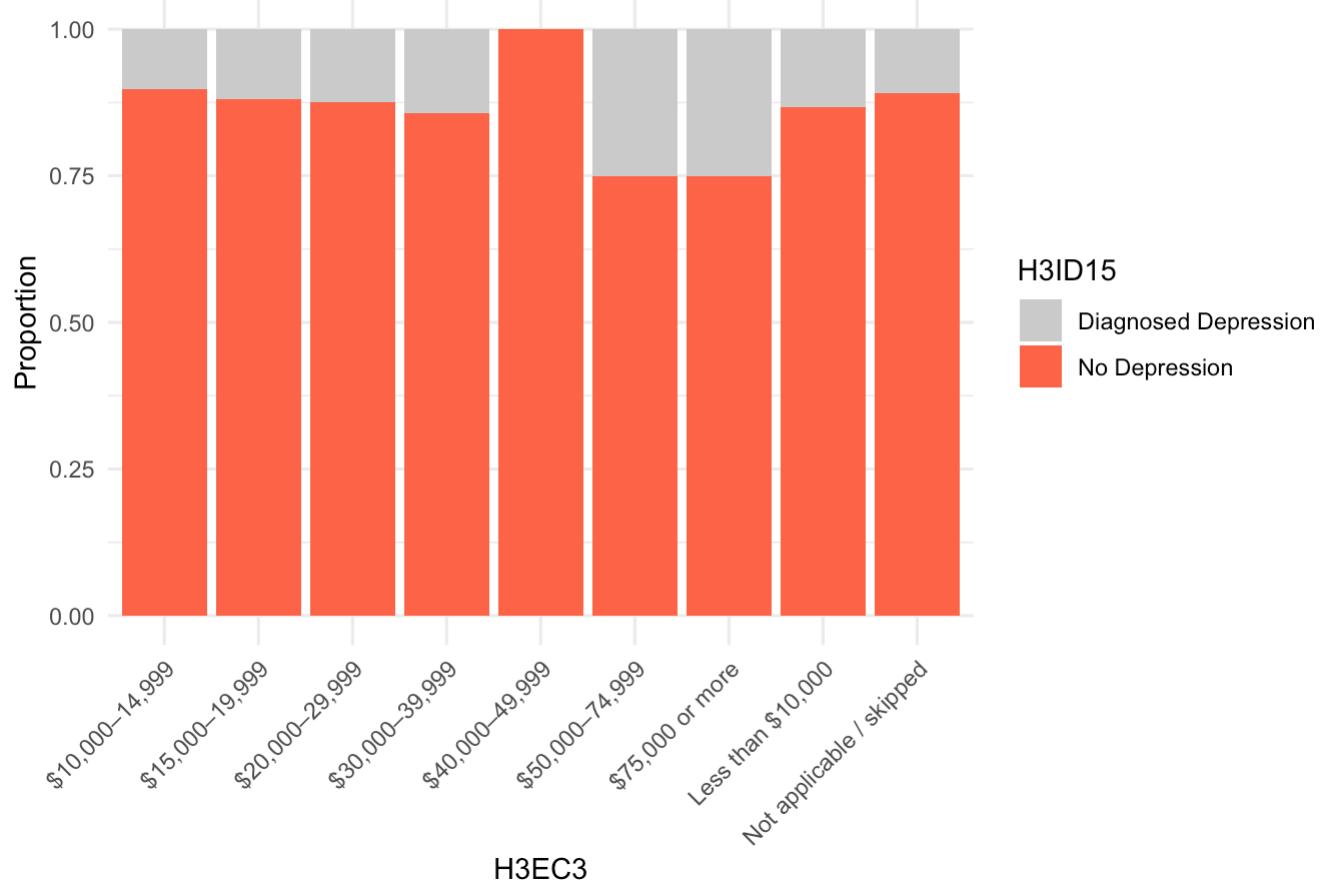
Depression by H3DA28



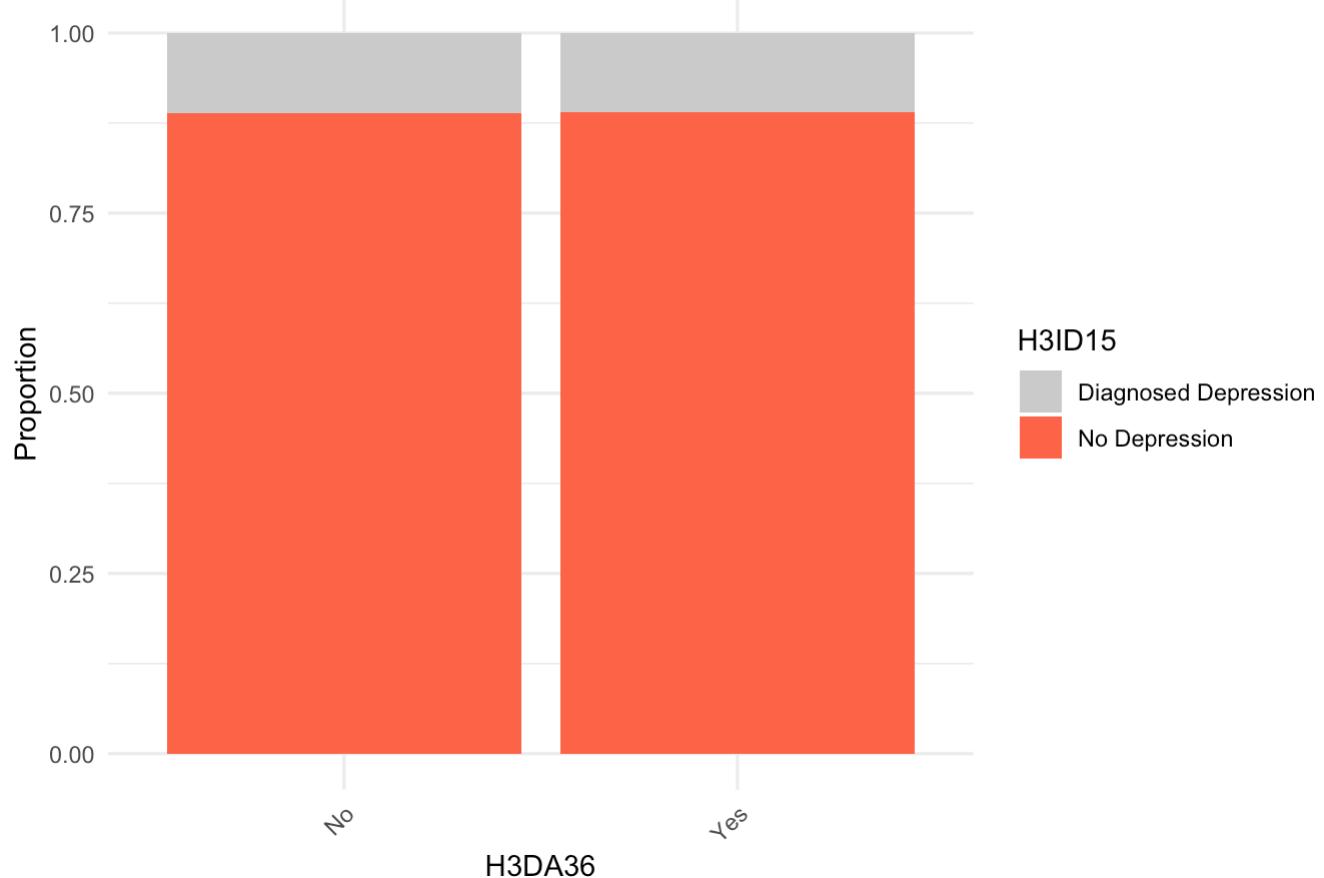
Depression by H3EC7



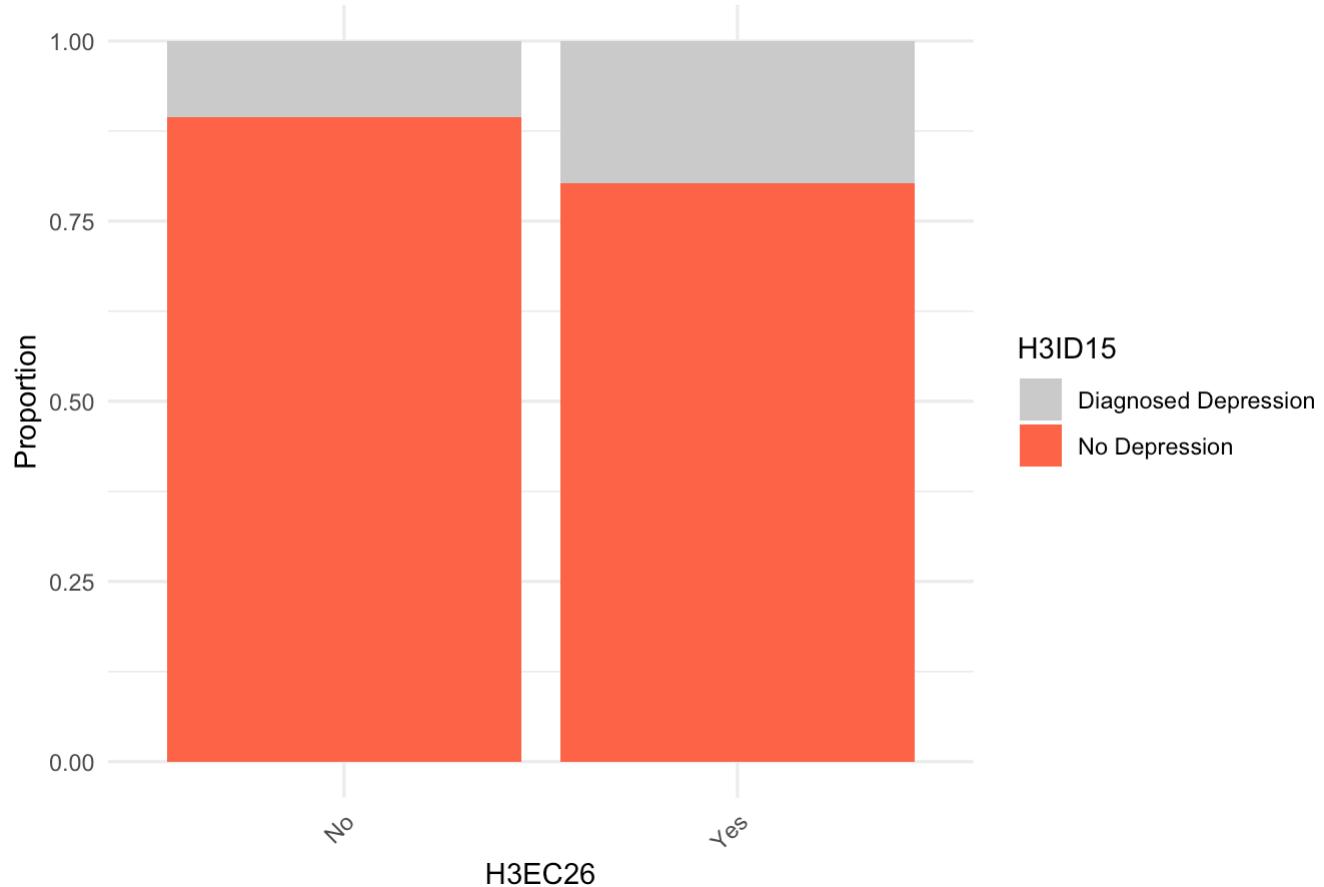
Depression by H3EC3



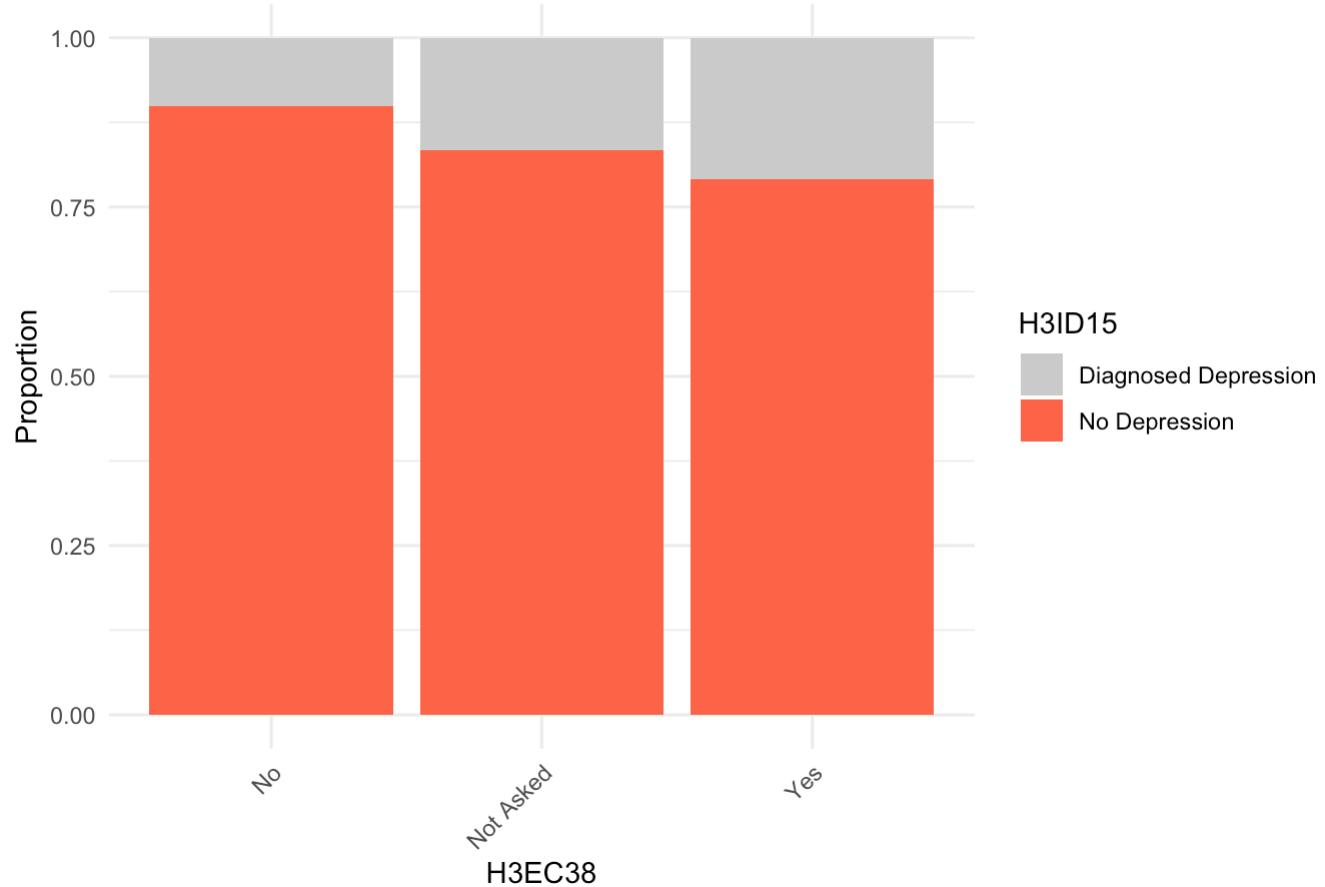
Depression by H3DA36



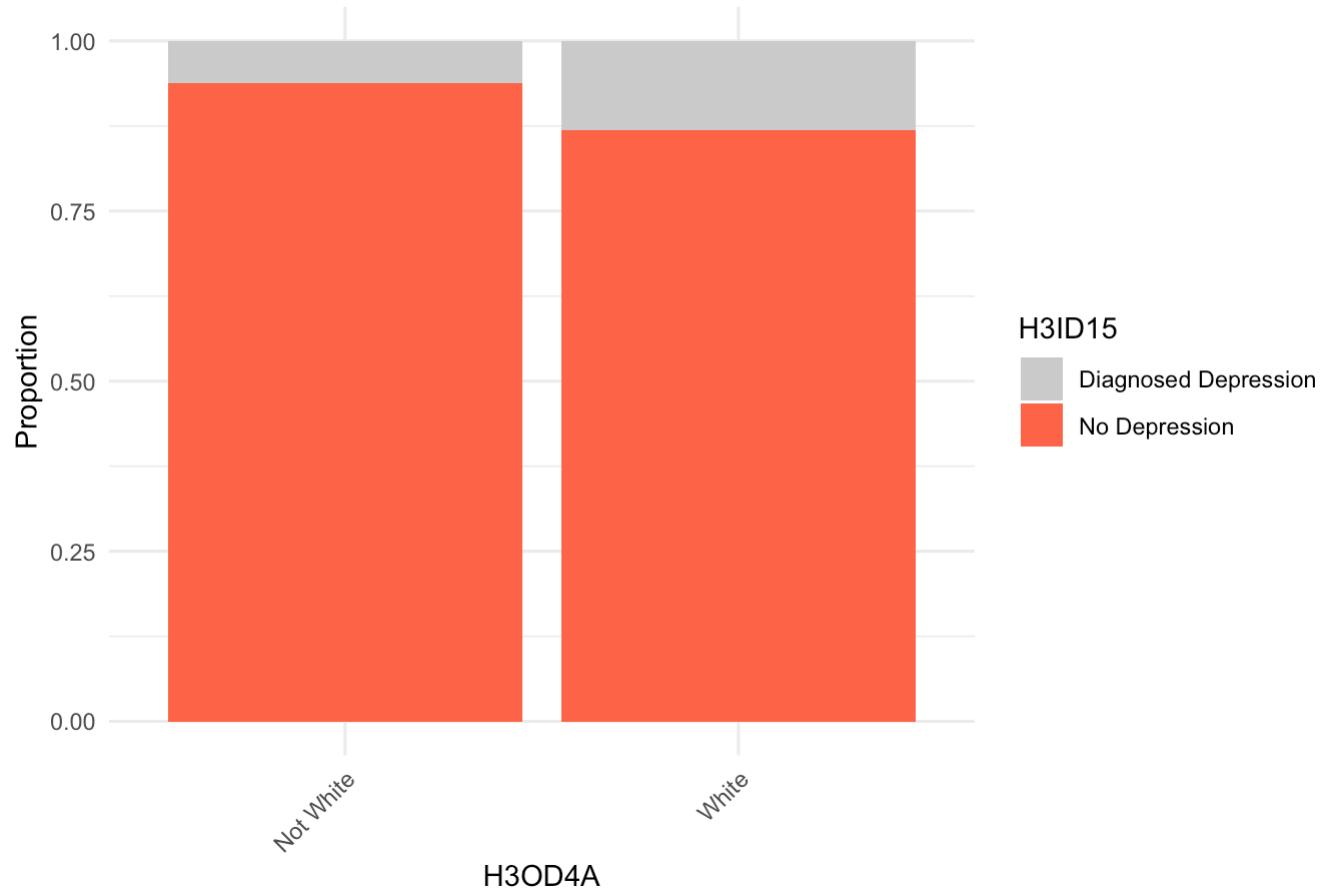
Depression by H3EC26



Depression by H3EC38

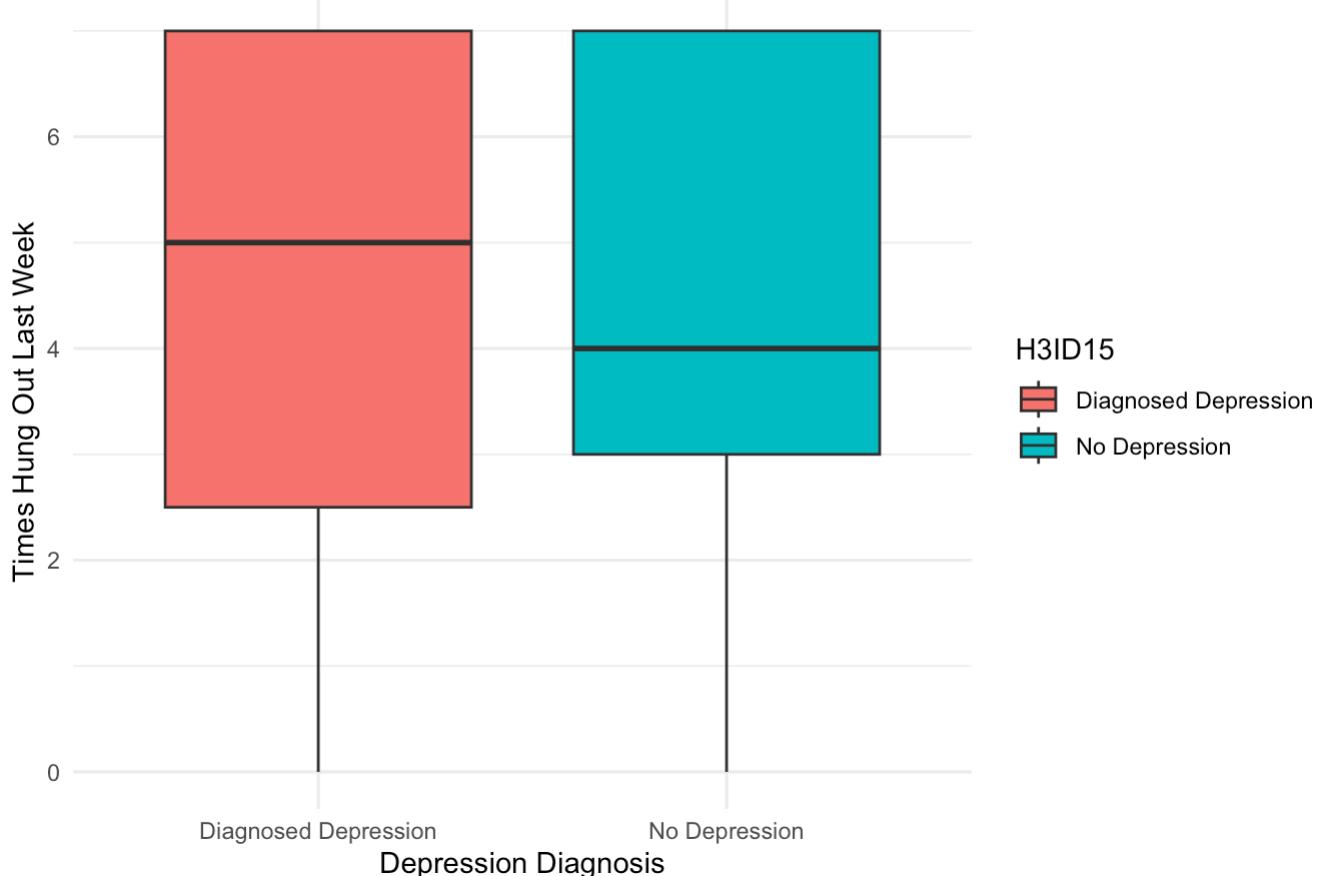


Depression by H3OD4A



```
# Step 5: Boxplot – Numeric Predictor vs Target
ggplot(data, aes(x = H3ID15, y = H3DA15, fill = H3ID15)) +
  geom_boxplot() +
  labs(title = "Frequency of Hanging Out with Friends by Depression Diagnosis",
       x = "Depression Diagnosis", y = "Times Hung Out Last Week") +
  theme_minimal()
```

Frequency of Hanging Out with Friends by Depression Diagnosis



```
# Step 6: Cross-tabs
```

```
# Peer support vs Depression
print("Depression by Peer Interaction:")
```

```
## [1] "Depression by Peer Interaction:"
```

```
print(prop.table(table(data$H3DA15, data$H3ID15), margin = 1))
```

```
##
##      Diagnosed Depression No Depression
##  0          0.13068182    0.86931818
##  1          0.11312217    0.88687783
##  2          0.10389610    0.89610390
##  3          0.08253968    0.91746032
##  4          0.10101010    0.89898990
##  5          0.11389522    0.88610478
##  6          0.10849057    0.89150943
##  7          0.12254570    0.87745430
```

```
# Income vs Depression
print("Depression by Household Income:")
```

```
## [1] "Depression by Household Income:"
```

```
print(prop.table(table(data$H3EC7, data$H3ID15), margin = 1))
```

```
##
##                                     Diagnosed Depression No Depression
## $10,000–14,999                      0.13636364   0.86363636
## $15,000–19,999                      0.10344828   0.89655172
## $20,000–29,999                      0.16666667   0.83333333
## $30,000–39,999                      0.04347826   0.95652174
## $40,000–49,999                      0.06153846   0.93846154
## $50,000–74,999                      0.12871287   0.87128713
## $75,000 or more                     0.11188811   0.88811189
## Less than $10,000                    0.19354839   0.80645161
## Not applicable / skipped            0.11052770   0.88947230
```

```
# Gender vs Depression
print("Depression by Gender:")
```

```
## [1] "Depression by Gender:"
```

```
print(prop.table(table(data$BI0_SEX3, data$H3ID15), margin = 1))
```

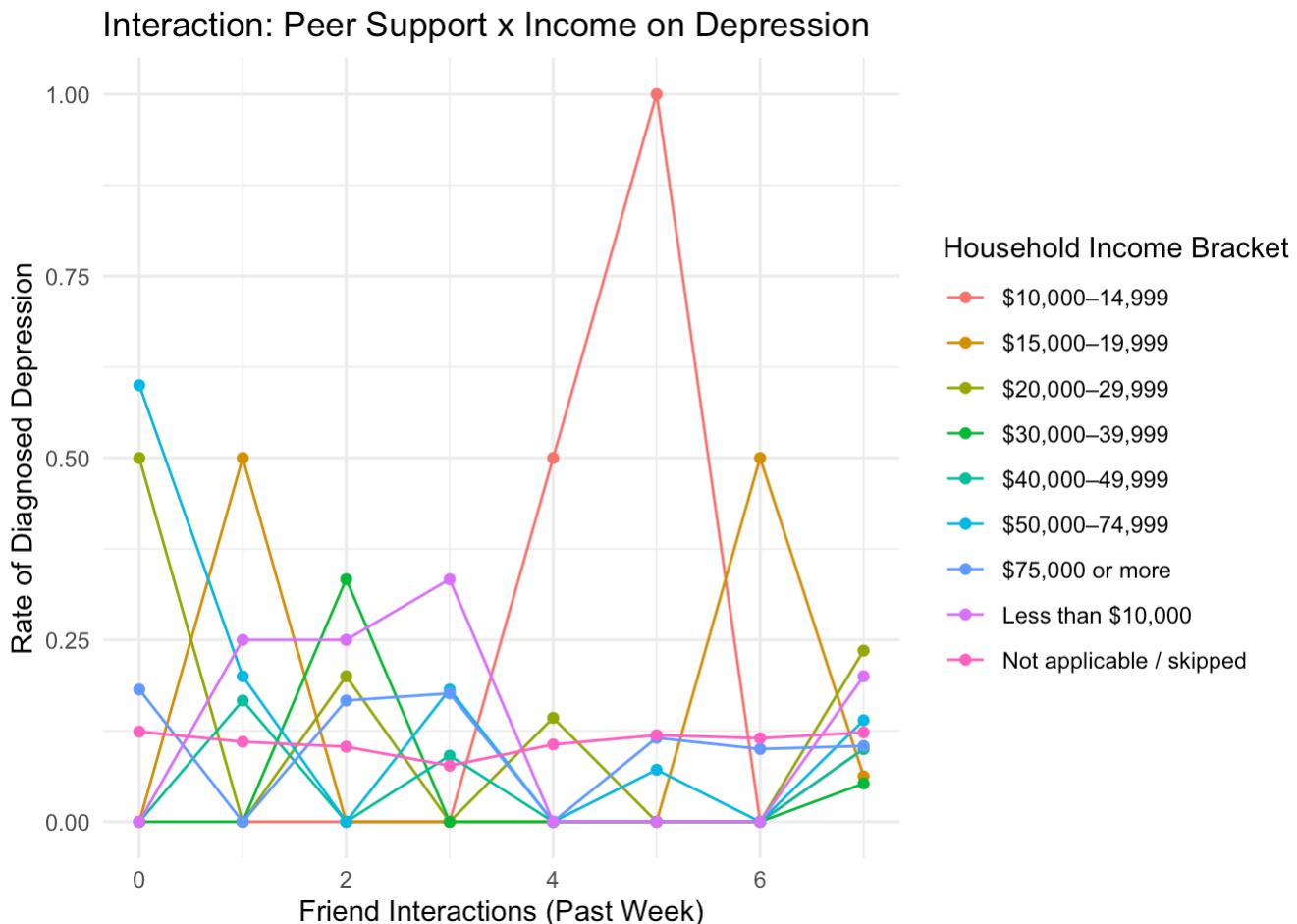
```
##
##                                     Diagnosed Depression No Depression
## Female                           0.14693172   0.85306828
## Male                            0.06838906   0.93161094
```

```
library(dplyr)
```

```
# Collapse to average depression rate
interaction_summary <- data %>%
  group_by(H3DA15, H3EC7) %>%
  summarise(depression_rate = mean(H3ID15 == "Diagnosed Depression"))
```

```
## `summarise()` has grouped output by 'H3DA15'. You can override using the
## `.` argument.
```

```
# Plot
ggplot(interaction_summary, aes(x = H3DA15, y = depression_rate, color = H3EC7)) +
  geom_line() +
  geom_point() +
  labs(title = "Interaction: Peer Support x Income on Depression",
       x = "Friend Interactions (Past Week)",
       y = "Rate of Diagnosed Depression",
       color = "Household Income Bracket") +
  theme_minimal()
```



```
library(ggmosaic)

ggplot(data = data) +
  geom_mosaic(aes(weight = 1, x = product(H3DA28, H3ID15), fill = H3ID15)) +
  labs(title = "Depression by Employment Status",
       x = "Employment", y = "Proportion") +
  theme_minimal()
```

```
## Warning: The `scale_name` argument of `continuous_scale()` is deprecated as of ggplot2
## 3.5.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```

## Warning: The `trans` argument of `continuous_scale()` is deprecated as of ggplot2
3.5.0.
## i Please use the `transform` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

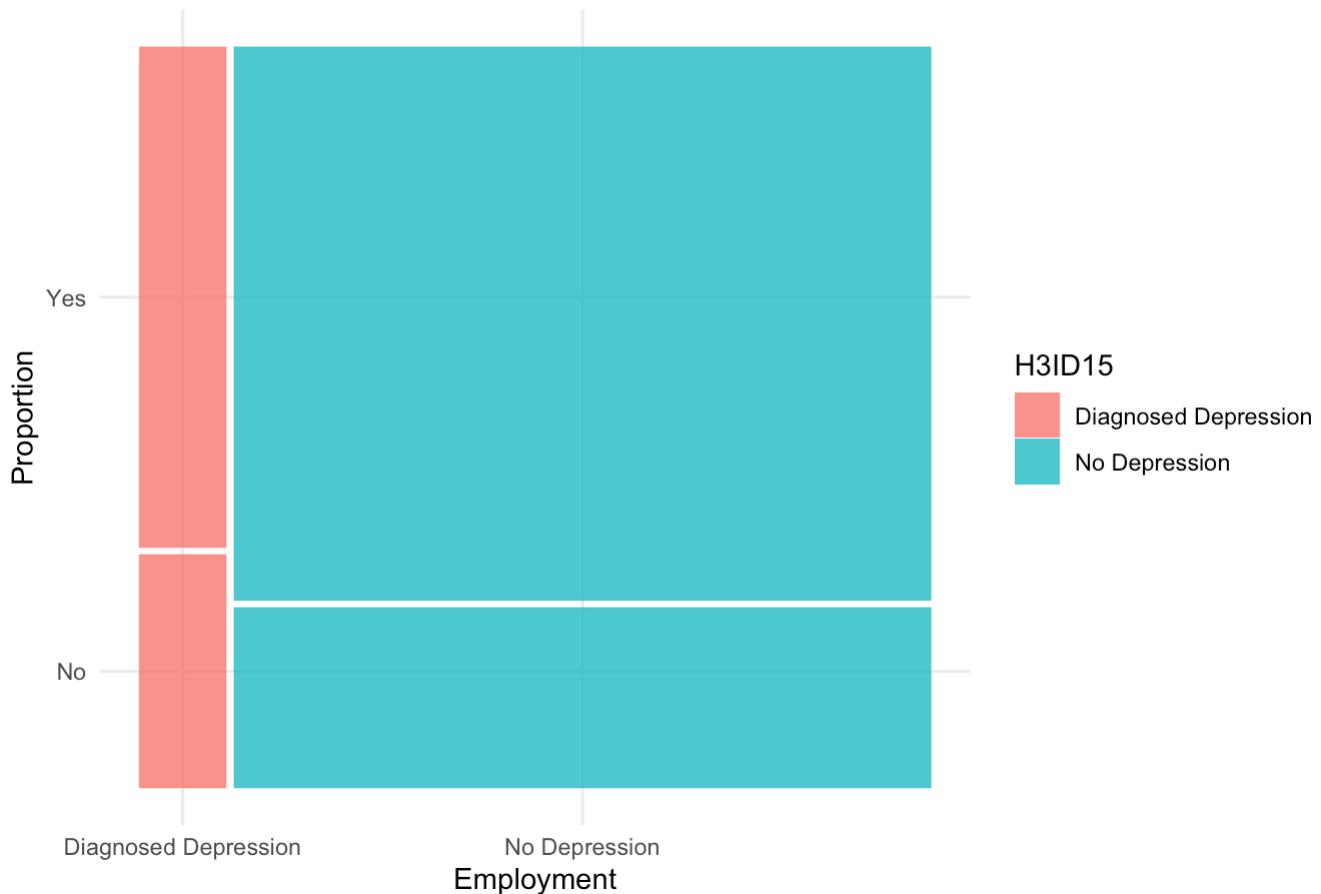
```

```

## Warning: `unite_()` was deprecated in tidyverse 1.2.0.
## i Please use `unite()` instead.
## i The deprecated feature was likely used in the ggmosaic package.
## Please report the issue at <https://github.com/haleyjeppson/ggmosaic>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

```

Depression by Employment Status



```

heat_data <- data %>%
  group_by(BIO_SEX3, H3EC7) %>%
  summarise(depression_rate = mean(H3ID15 == "Diagnosed Depression"))

```

```

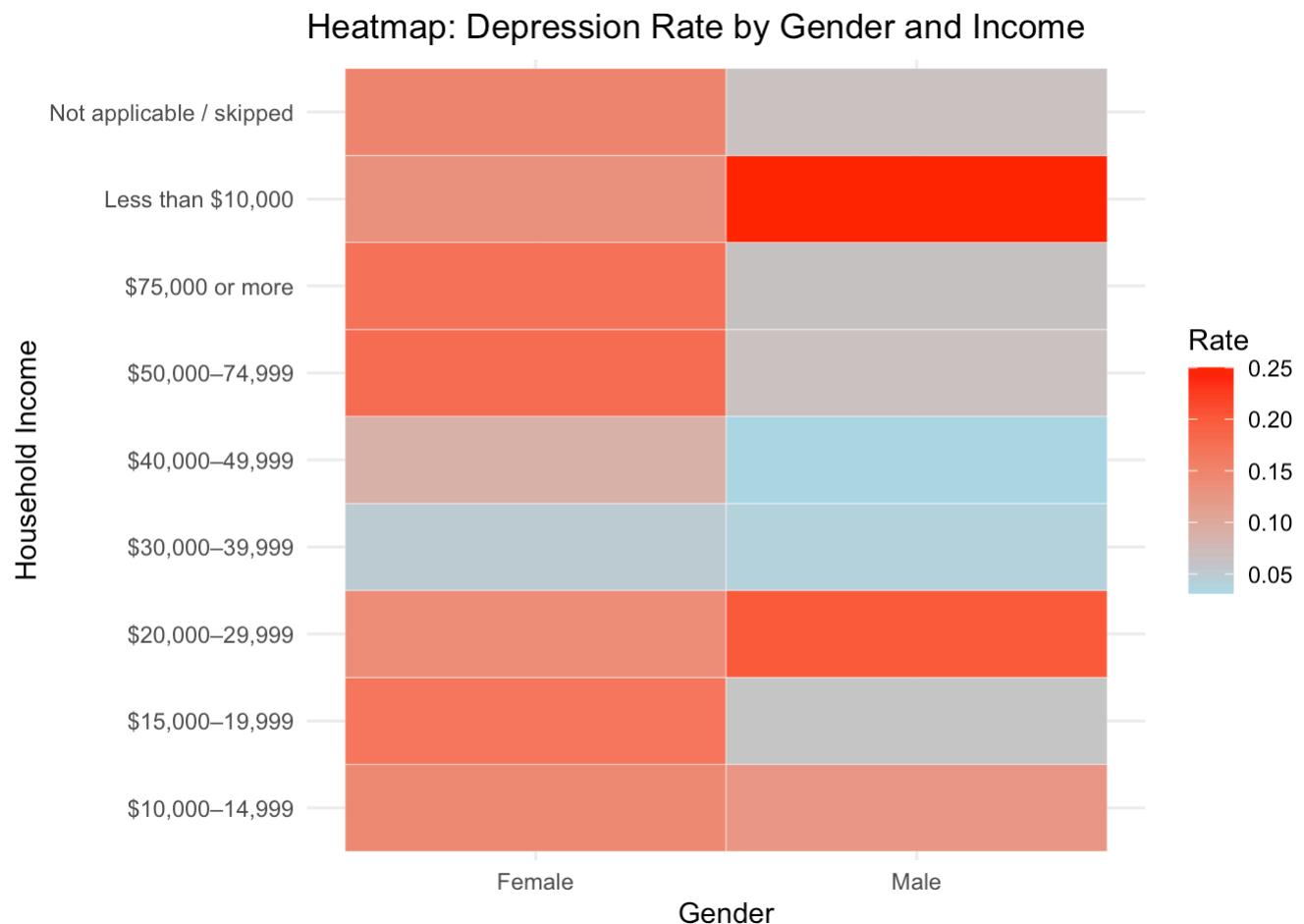
## `summarise()` has grouped output by 'BIO_SEX3'. You can override using the
## `.`groups` argument.

```

```

ggplot(heat_data, aes(x = BI0_SEX3, y = H3EC7, fill = depression_rate)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightblue", high = "red") +
  labs(title = "Heatmap: Depression Rate by Gender and Income",
       x = "Gender", y = "Household Income", fill = "Rate") +
  theme_minimal()

```



```

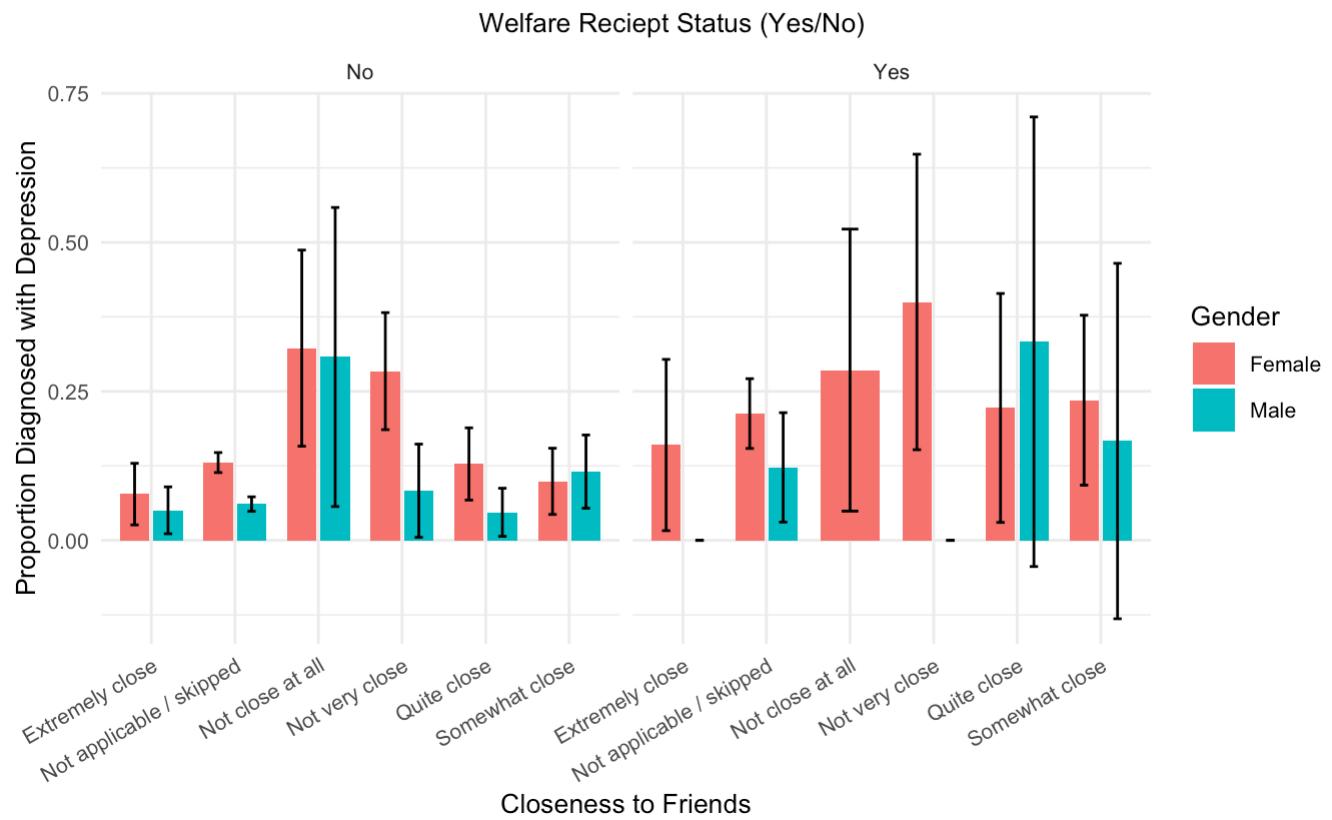
# Filter out non-responses and prepare data
plot_df <- data %>%
  filter(H3WP53 != "Not Asked", H3EC38 != "Not Asked") %>%
  group_by(H3WP53, H3EC38, BI0_SEX3) %>%
  summarise(
    DepressionRate = mean(H3ID15 == "Diagnosed Depression", na.rm = TRUE),
    N = n(),
    SE = sqrt((DepressionRate * (1 - DepressionRate))/N),
    .groups = "drop"
  )

# Plot
ggplot(plot_df, aes(x = H3WP53, y = DepressionRate, fill = BI0_SEX3)) +
  geom_col(position = position_dodge(width = 0.8), width = 0.7) +
  geom_errorbar(
    aes(ymin = DepressionRate - 1.96 * SE, ymax = DepressionRate + 1.96 * SE),
    width = 0.2, position = position_dodge(width = 0.8)
  ) +
  facet_wrap(~H3EC38) +
  labs(
    title = "Depression Diagnosis by Closeness to Friends, Income, and Gender ",
    subtitle = "Higher closeness is associated with lower depression rates across genders and income groups \n\n",
    x = "Closeness to Friends",
    y = "Proportion Diagnosed with Depression",
    fill = "Gender"
  ) +
  theme_minimal(base_size = 10) +
  theme(axis.text.x = element_text(angle = 30, hjust = 1))

```

Depression Diagnosis by Closeness to Friends, Income, and Gender

Higher closeness is associated with lower depression rates across genders and income groups



```

library(dplyr)

plot_data <- data %>%
  group_by(H3DA15, BI0_SEX3) %>%
  summarise(DepressionRate = mean(H3ID15 == "Diagnosed Depression"),
            N = n(),
            SE = sqrt((DepressionRate * (1 - DepressionRate)) / N),
            .groups = 'drop')

# Add confidence intervals
plot_data <- plot_data %>%
  mutate(lower = DepressionRate - 1.96 * SE,
        upper = DepressionRate + 1.96 * SE)

library(ggplot2)

ggplot(plot_data, aes(x = H3DA15, y = DepressionRate, color = BI0_SEX3, group = BI0_SEX3)) +
  geom_line(size = 1.2) +
  geom_point(size = 3) +
  geom_ribbon(aes(ymin = lower, ymax = upper, fill = BI0_SEX3), alpha = 0.2, color = NA) +
  scale_color_manual(values = c("steelblue", "coral")) +
  scale_fill_manual(values = c("steelblue", "coral")) +
  labs(
    title = "Depression Rates by Peer Interaction and Gender",
    subtitle = "Exploring how peer interactions relate to depression for males and females",
    x = "Peer Interaction (Times Hung Out per Week)",
    y = "Depression Rate",
    color = "Gender",
    fill = "Gender"
  ) +
  theme_minimal(base_size = 15)

```

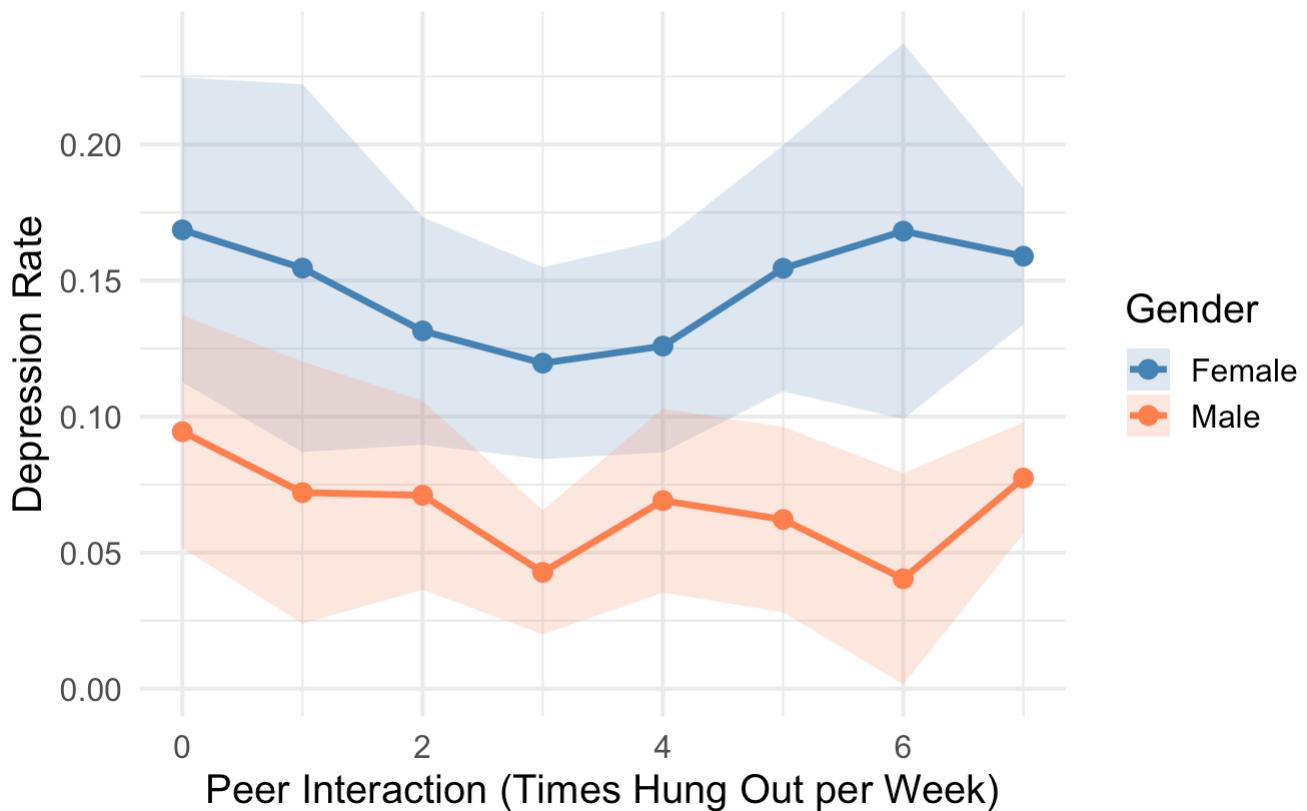
```

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

```

Depression Rates by Peer Interaction and Gender

Exploring how peer interactions relate to depression for males and females.



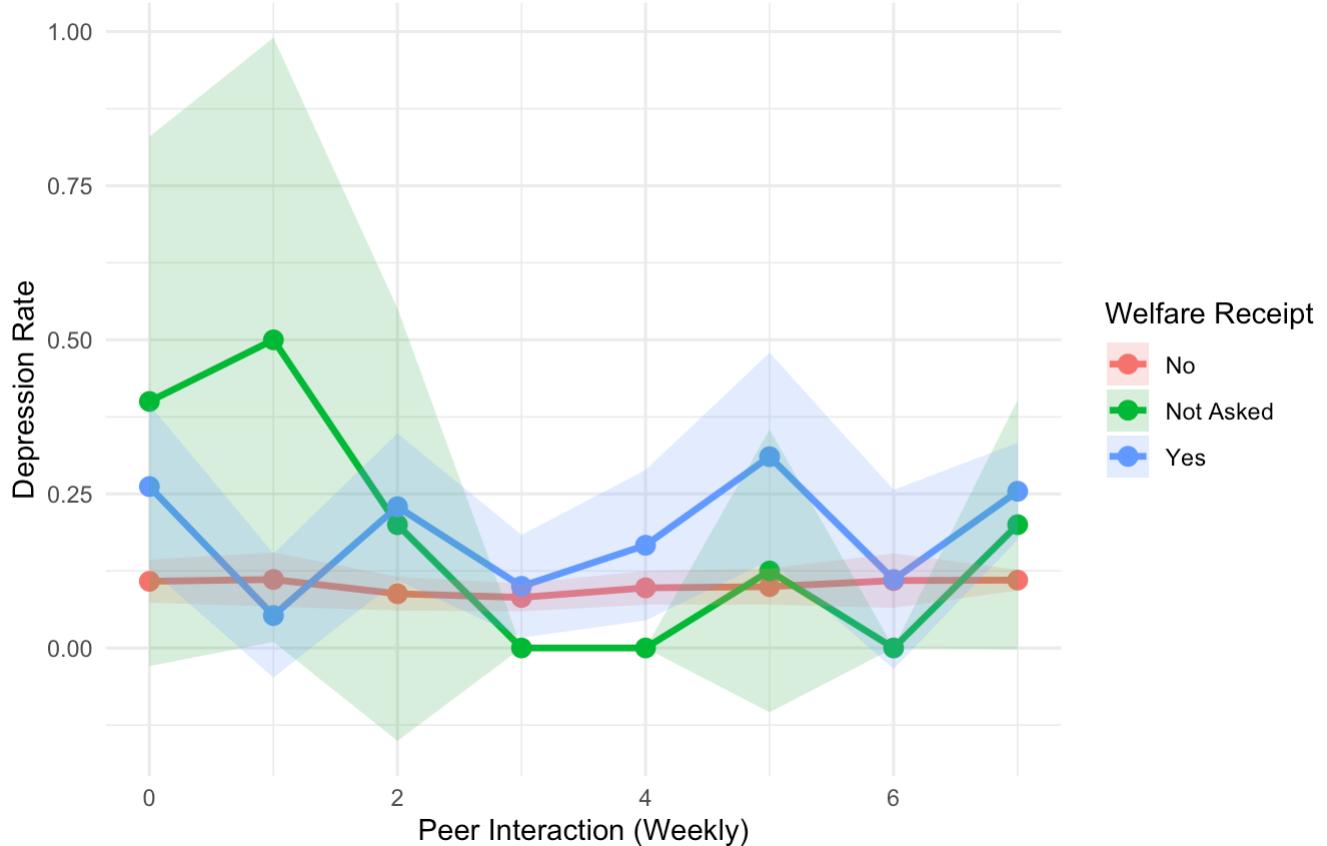
```
library(dplyr)
library(ggplot2)

plot_data_welfare <- data %>%
  group_by(H3DA15, H3EC38) %>%
  summarise(DepressionRate = mean(H3ID15 == "Diagnosed Depression"),
            N = n(),
            SE = sqrt((DepressionRate * (1 - DepressionRate))/N),
            .groups = 'drop')

ggplot(plot_data_welfare, aes(x = H3DA15, y = DepressionRate, color = H3EC38, group =
H3EC38)) +
  geom_line(size = 1.2) +
  geom_point(size = 3) +
  geom_ribbon(aes(ymax = DepressionRate + 1.96*SE, ymin = DepressionRate - 1.96*SE, fill = H3EC38), alpha = 0.2, color = NA) +
  labs(title = "Depression by Peer Interaction and Welfare Status",
       subtitle = "Does welfare receipt influence peer support's impact?",
       x = "Peer Interaction (Weekly)",
       y = "Depression Rate",
       color = "Welfare Receipt",
       fill = "Welfare Receipt") +
  theme_minimal()
```

Depression by Peer Interaction and Welfare Status

Does welfare receipt influence peer support's impact?

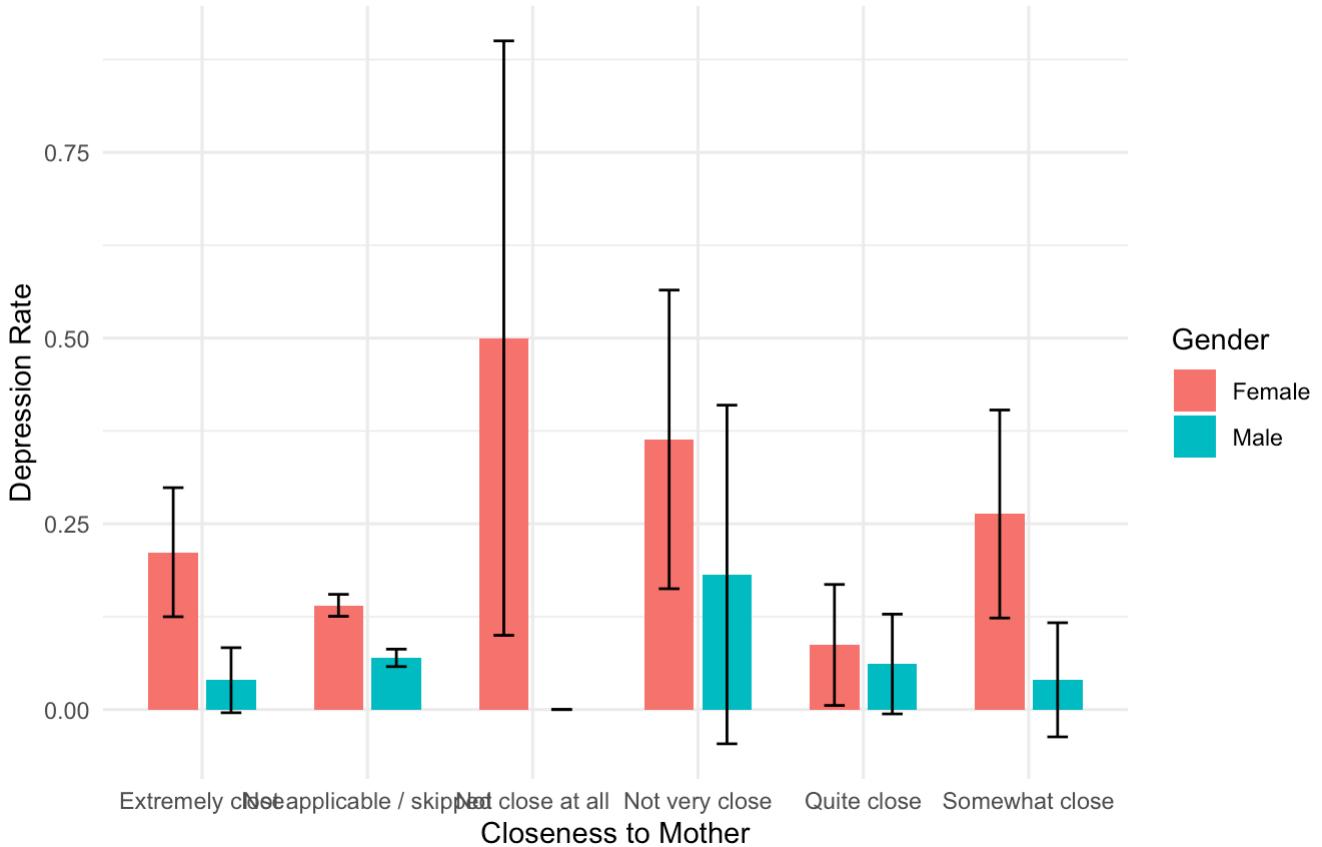


```
plot_data_parents <- data %>%
  group_by(H3WP46, BI0_SEX3) %>% # Using mother's closeness here; you can replicate
  summarise(DepressionRate = mean(H3ID15 == "Diagnosed Depression"),
            N = n(),
            SE = sqrt((DepressionRate * (1 - DepressionRate))/N),
            .groups = 'drop')

ggplot(plot_data_parents, aes(x = H3WP46, y = DepressionRate, group = BI0_SEX3, fill =
= BI0_SEX3)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.7), width=0.6) +
  geom_errorbar(aes(ymin = DepressionRate - 1.96*SE, ymax = DepressionRate + 1.96*S
E),
                position = position_dodge(width = 0.7), width = 0.25) +
  labs(title = "Depression by Mother's Closeness and Gender",
       subtitle = "Does parental closeness affect depression differently by gender?",
       x = "Closeness to Mother",
       y = "Depression Rate",
       fill = "Gender") +
  theme_minimal()
```

Depression by Mother's Closeness and Gender

Does parental closeness affect depression differently by gender?



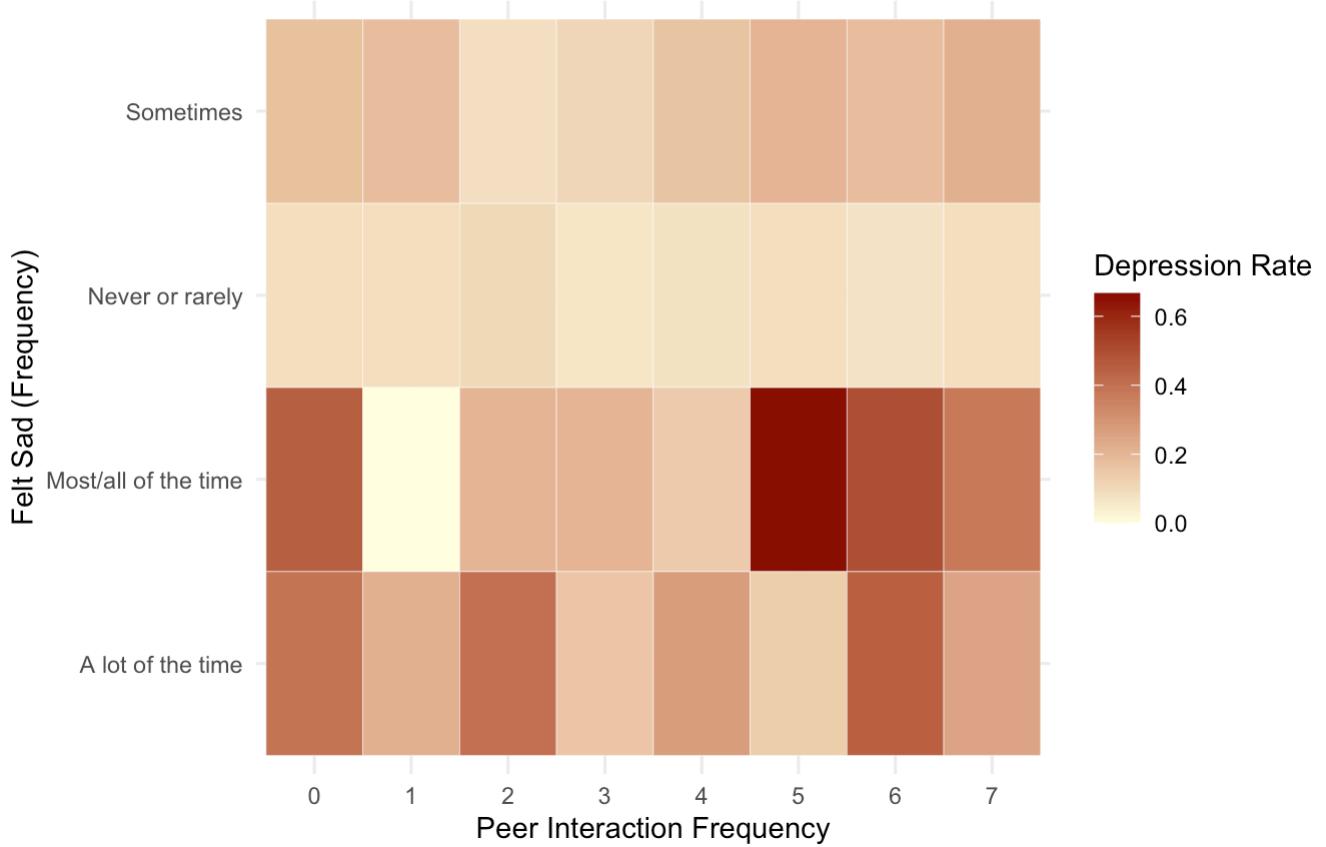
```
library(ggplot2)
library(dplyr)

heatmap_data <- data %>%
  group_by(H3DA15, H3SP6) %>%
  summarise(DepressionRate = mean(H3ID15 == "Diagnosed Depression"), .groups = 'drop')
```

```
ggplot(heatmap_data, aes(x = factor(H3DA15), y = H3SP6, fill = DepressionRate)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightyellow", high = "darkred") +
  labs(title = "Depression Rate by Peer Interaction and Sadness",
       subtitle = "Interaction between emotional state and social activity",
       x = "Peer Interaction Frequency",
       y = "Felt Sad (Frequency)",
       fill = "Depression Rate") +
  theme_minimal()
```

Depression Rate by Peer Interaction and Sadness

Interaction between emotional state and social activity



```

library(dplyr)
library(ggplot2)

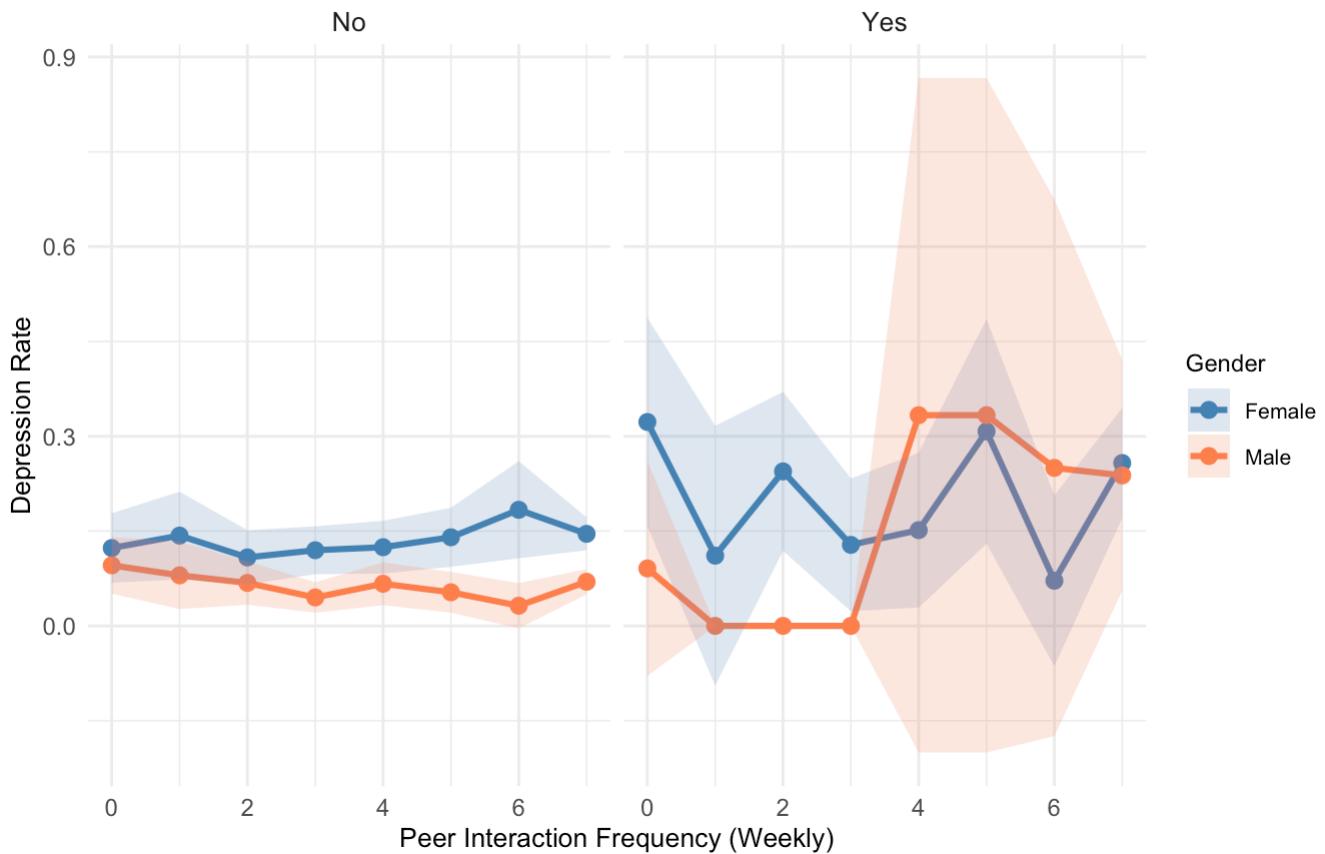
# Exclude 'Not Asked'
plot_main <- data %>%
  filter(H3EC38 != "Not Asked") %>%
  group_by(H3DA15, BI0_SEX3, H3EC38) %>%
  summarise(DepressionRate = mean(H3ID15 == "Diagnosed Depression"),
             N = n(),
             SE = sqrt((DepressionRate * (1 - DepressionRate))/N),
             .groups = 'drop')

# Plot with smaller text sizes
ggplot(plot_main, aes(x = H3DA15, y = DepressionRate, color = BI0_SEX3, group = BI0_SEX3)) +
  geom_line(size = 1.1) +
  geom_point(size = 2.5) +
  geom_ribbon(aes(ymin = DepressionRate - 1.96*SE, ymax = DepressionRate + 1.96*SE, fill = BI0_SEX3), alpha = 0.2, color = NA) +
  facet_wrap(~H3EC38) +
  scale_color_manual(values = c("steelblue", "coral")) +
  scale_fill_manual(values = c("steelblue", "coral")) +
  labs(
    title = "Peer Interaction and Depression: Gender and Welfare Status",
    subtitle = "Examining friendship as a protective factor across socioeconomic situations",
    x = "Peer Interaction Frequency (Weekly)",
    y = "Depression Rate",
    color = "Gender",
    fill = "Gender"
  ) +
  theme_minimal(base_size = 10) + # Reduce base text size
  theme(
    plot.title = element_text(size = 12, face = "bold"),
    plot.subtitle = element_text(size = 10),
    axis.title = element_text(size = 10),
    axis.text = element_text(size = 9),
    strip.text = element_text(size = 10),
    legend.title = element_text(size = 9),
    legend.text = element_text(size = 8)
  )

```

Peer Interaction and Depression: Gender and Welfare Status

Examining friendship as a protective factor across socioeconomic situations



###Part 3 - Regressions

Correlation Matrix:

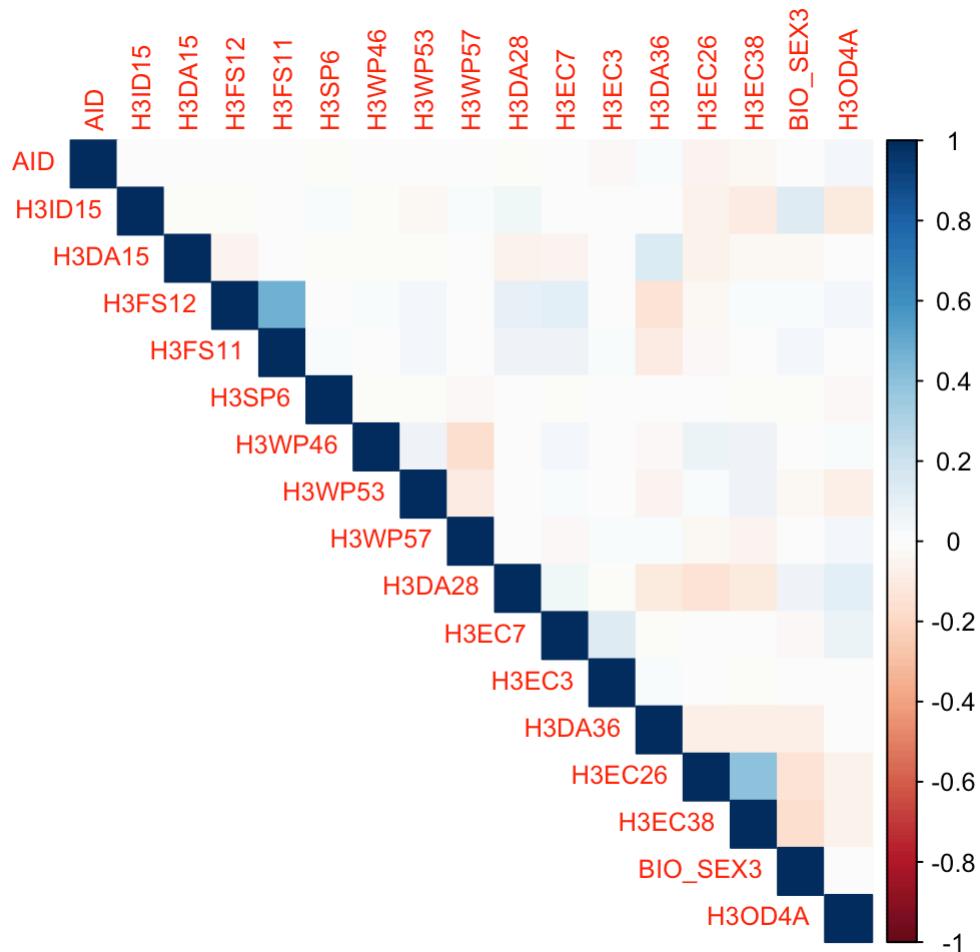
```
# Load libraries
library(tidyverse)
library(corrplot)

## corrplot 0.95 loaded

# Read and clean data
data <- read.csv("complete_addhealth_data.csv", stringsAsFactors = TRUE)
names(data) <- gsub("_recode$", "", names(data))

# Encode factors as numeric for correlation
data_corr <- data
for (col in names(data_corr)) {
  if (is.factor(data_corr[[col]]) || is.character(data_corr[[col]])) {
    data_corr[[col]] <- as.numeric(as.factor(data_corr[[col]]))
  }
}

# Plot correlation matrix
cor_matrix <- cor(data_corr, use = "complete.obs")
corrplot(cor_matrix, method = "color", type = "upper", tl.cex = 0.8)
```



Baseline Regression Model:

```

# Make sure variables are factors where needed
factor_vars <- c("H3ID15", "BIO_SEX3", "H3DA28", "H3DA36", "H3OD4A", "H3EC7", "H3EC3")
data[factor_vars] <- lapply(data[factor_vars], as.factor)

# Baseline model with only SES + demographic predictors
baseline_model <- glm(H3ID15 ~ H3EC3 + H3EC7 + H3DA28 + H3DA36 + BIO_SEX3 + H3OD4A,
                      data = data, family = binomial)

summary(baseline_model)

```

```

## 
## Call:
## glm(formula = H3ID15 ~ H3EC3 + H3EC7 + H3DA28 + H3DA36 + BI0_SEX3 +
##      H30D4A, family = binomial, data = data)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                2.0059    0.7097   2.826 0.004709 **
## H3EC3$15,000–19,999       -0.4479    0.4879  -0.918 0.358636
## H3EC3$20,000–29,999       -0.5133    0.4750  -1.081 0.279819
## H3EC3$30,000–39,999       -0.6891    0.6397  -1.077 0.281335
## H3EC3$40,000–49,999        11.9528   260.8805   0.046 0.963456
## H3EC3$50,000–74,999       -1.5048    1.2388  -1.215 0.224472
## H3EC3$75,000 or more      -1.4505    1.2306  -1.179 0.238535
## H3EC3Less than $10,000     -0.2761    0.3675  -0.751 0.452530
## H3EC3Not applicable / skipped -0.1923    0.3301  -0.582 0.560277
## H3EC7$15,000–19,999       0.2259    0.8906   0.254 0.799760
## H3EC7$20,000–29,999       -0.2564    0.7654  -0.335 0.737633
## H3EC7$30,000–39,999        1.1271    0.9677   1.165 0.244129
## H3EC7$40,000–49,999        0.9564    0.8256   1.158 0.246708
## H3EC7$50,000–74,999        0.2023    0.7078   0.286 0.775018
## H3EC7$75,000 or more       0.2734    0.6947   0.394 0.693910
## H3EC7Less than $10,000     -0.4803    0.7946  -0.604 0.545521
## H3EC7Not applicable / skipped 0.3610    0.6421   0.562 0.573950
## H3DA28Yes                  0.3974    0.1093   3.637 0.000276 ***
## H3DA36Yes                  0.1107    0.1040   1.064 0.287140
## BI0_SEX3Male                0.8652    0.1087   7.959 1.73e-15 ***
## H30D4AWhite                 -0.9325   0.1322  -7.056 1.72e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2985.6 on 4287 degrees of freedom
## Residual deviance: 2836.2 on 4267 degrees of freedom
## AIC: 2878.2
##
## Number of Fisher Scoring iterations: 13

```

Full Model

```
# Ensure factor encoding for relevant variables
factor_vars <- c("H3ID15", "BIO_SEX3", "H3DA28", "H3DA36", "H30D4A",
                 "H3EC7", "H3EC3", "H3FS12", "H3FS11", "H3SP6",
                 "H3WP46", "H3WP53", "H3WP57", "H3EC26", "H3EC38")
data[factor_vars] <- lapply(data[factor_vars], as.factor)

# Fit the full model
full_model <- glm(H3ID15 ~ H3FS12 + H3FS11 + H3SP6 + H3WP46 + H3WP53 + H3WP57 +
                    H3DA15 + H3DA28 + H3DA36 + H3EC7 + H3EC3 + H3EC26 + H3EC38 +
                    BIO_SEX3 + H30D4A,
                    data = data,
                    family = binomial)

# View summary
summary(full_model)
```

```

## 
## Call:
## glm(formula = H3ID15 ~ H3FS12 + H3FS11 + H3SP6 + H3WP46 + H3WP53 +
##      H3WP57 + H3DA15 + H3DA28 + H3DA36 + H3EC7 + H3EC3 + H3EC26 +
##      H3EC38 + BI0_SEX3 + H3OD4A, family = binomial, data = data)
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.85679   0.87326   2.126  0.03348 *
## H3FS12Friends more influential -0.42142   0.22121  -1.905  0.05677 .
## H3FS12Not Asked -0.21419   0.21730  -0.986  0.32428
## H3FS11All      0.98530   0.76752   1.284  0.19923
## H3FS11Most     0.60785   0.30961   1.963  0.04961 *
## H3FS11None     -0.32705   0.41821  -0.782  0.43419
## H3FS11Not Asked NA          NA        NA      NA
## H3FS110ne      0.55973   0.75849   0.738  0.46054
## H3FS11Some     -0.10794   0.27011  -0.400  0.68943
## H3SP6Most/all of the time -0.33855   0.32061  -1.056  0.29098
## H3SP6Never or rarely    1.27747   0.20774   6.149 7.78e-10 ***
## H3SP6Sometimes   0.50124   0.22018   2.276  0.02282 *
## H3WP46Not applicable / skipped -0.01928   0.31095  -0.062  0.95056
## H3WP46Not close at all    -1.26003   0.89413  -1.409  0.15877
## H3WP46Not very close     -0.68161   0.50380  -1.353  0.17608
## H3WP46Quite close       0.90025   0.48230   1.867  0.06196 .
## H3WP46Somewhat close    -0.11284   0.44370  -0.254  0.79925
## H3WP53Not applicable / skipped -0.40227   0.26763  -1.503  0.13282
## H3WP53Not close at all    -1.57468   0.39576  -3.979 6.92e-05 ***
## H3WP53Not very close     -1.16842   0.33242  -3.515  0.00044 ***
## H3WP53Quite close       -0.39348   0.33433  -1.177  0.23924
## H3WP53Somewhat close    -0.51437   0.32426  -1.586  0.11267
## H3WP57Not applicable / skipped 0.55520   0.34646   1.602  0.10905
## H3WP57Yes             0.09226   0.50458   0.183  0.85492
## H3DA15               -0.03106   0.02222  -1.398  0.16216
## H3DA28Yes            0.33311   0.11547   2.885  0.00392 **
## H3DA36Yes            -0.03152   0.11085  -0.284  0.77614
## H3EC7$15,000–19,999   0.41120   0.92216   0.446  0.65566
## H3EC7$20,000–29,999   -0.42385   0.79055  -0.536  0.59185
## H3EC7$30,000–39,999   1.12181   0.99544   1.127  0.25976
## H3EC7$40,000–49,999   0.75476   0.84960   0.888  0.37434
## H3EC7$50,000–74,999   0.10759   0.73420   0.147  0.88349
## H3EC7$75,000 or more   0.11557   0.71820   0.161  0.87216
## H3EC7Less than $10,000 -0.46176   0.82989  -0.556  0.57793
## H3EC7Not applicable / skipped 0.29100   0.66489   0.438  0.66163
## H3EC3$15,000–19,999   -0.48037   0.51578  -0.931  0.35168
## H3EC3$20,000–29,999   -0.46785   0.49984  -0.936  0.34927
## H3EC3$30,000–39,999   -0.74911   0.65527  -1.143  0.25295
## H3EC3$40,000–49,999   11.89368  253.80613  0.047  0.96262
## H3EC3$50,000–74,999   -1.67275   1.30285  -1.284  0.19917
## H3EC3$75,000 or more   -2.04968   1.24262  -1.649  0.09905 .
## H3EC3Less than $10,000 -0.40141   0.38516  -1.042  0.29733
## H3EC3Not applicable / skipped -0.30255   0.34766  -0.870  0.38417
## H3EC26Yes             -0.22082   0.22370  -0.987  0.32360
## H3EC38Not Asked      -0.86712   0.40539  -2.139  0.03244 *
## H3EC38Yes             -0.45899   0.17169  -2.673  0.00751 **
## BI0_SEX3Male          0.68594   0.11358   6.040 1.55e-09 ***

```

```
## H30D4AWhite           -1.13867    0.14088   -8.083 6.33e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2985.6  on 4287  degrees of freedom
## Residual deviance: 2658.2  on 4241  degrees of freedom
## AIC: 2752.2
##
## Number of Fisher Scoring iterations: 13
```

Step-Wise Regression Model

```
# Full model with all variables
full_model <- glm(H3ID15 ~ ., data = data, family = binomial)

# Stepwise model selection
step_model <- step(full_model, direction = "both")
```

```

## Start: AIC=2754.01
## H3ID15 ~ AID + H3DA15 + H3FS12 + H3FS11 + H3SP6 + H3WP46 + H3WP53 +
##      H3WP57 + H3DA28 + H3EC7 + H3EC3 + H3DA36 + H3EC26 + H3EC38 +
##      BI0_SEX3 + H3OD4A
##
##          Df Deviance    AIC
## - H3EC3     8  2664.7 2744.7
## - H3EC7     8  2665.5 2745.5
## - H3DA36    1  2658.1 2752.1
## - AID       1  2658.2 2752.2
## - H3WP57    2  2660.7 2752.7
## - H3EC26    1  2659.0 2753.0
## - H3FS11    5  2667.2 2753.2
## - H3DA15    1  2659.9 2753.9
## <none>        2658.0 2754.0
## - H3WP46    5  2668.1 2754.1
## - H3FS12    1  2661.7 2755.7
## - H3DA28    1  2666.1 2760.1
## - H3EC38    2  2668.4 2760.4
## - H3WP53    5  2684.4 2770.4
## - BI0_SEX3   1  2696.3 2790.3
## - H3OD4A    1  2735.8 2829.8
## - H3SP6     3  2743.3 2833.3
##
## Step: AIC=2744.71
## H3ID15 ~ AID + H3DA15 + H3FS12 + H3FS11 + H3SP6 + H3WP46 + H3WP53 +
##      H3WP57 + H3DA28 + H3EC7 + H3DA36 + H3EC26 + H3EC38 + BI0_SEX3 +
##      H3OD4A
##
##          Df Deviance    AIC
## - H3EC7     8  2672.5 2736.5
## - H3DA36    1  2664.7 2742.7
## - AID       1  2664.8 2742.8
## - H3WP57    2  2667.3 2743.3
## - H3EC26    1  2665.7 2743.7
## - H3FS11    5  2673.9 2743.9
## - H3WP46    5  2674.5 2744.5
## - H3DA15    1  2666.6 2744.6
## <none>        2664.7 2744.7
## - H3FS12    1  2668.5 2746.5
## - H3DA28    1  2672.4 2750.4
## - H3EC38    2  2674.9 2750.9
## + H3EC3     8  2658.0 2754.0
## - H3WP53    5  2691.1 2761.1
## - BI0_SEX3   1  2703.4 2781.4
## - H3OD4A    1  2741.4 2819.4
## - H3SP6     3  2750.8 2824.8
##
## Step: AIC=2736.54
## H3ID15 ~ AID + H3DA15 + H3FS12 + H3FS11 + H3SP6 + H3WP46 + H3WP53 +
##      H3WP57 + H3DA28 + H3DA36 + H3EC26 + H3EC38 + BI0_SEX3 + H3OD4A
##
##          Df Deviance    AIC
## - H3DA36    1  2672.6 2734.6
## - AID       1  2672.7 2734.7

```

```

## - H3WP57 2 2675.1 2735.1
## - H3EC26 1 2673.6 2735.6
## - H3FS11 5 2681.6 2735.6
## - H3WP46 5 2682.3 2736.3
## - H3DA15 1 2674.3 2736.3
## <none> 2672.5 2736.5
## - H3FS12 1 2676.2 2738.2
## - H3EC38 2 2682.6 2742.6
## - H3DA28 1 2680.7 2742.7
## + H3EC7 8 2664.7 2744.7
## + H3EC3 8 2665.5 2745.5
## - H3WP53 5 2698.8 2752.8
## - BI0_SEX3 1 2710.5 2772.5
## - H30D4A 1 2748.7 2810.7
## - H3SP6 3 2759.9 2817.9
##
## Step: AIC=2734.56
## H3ID15 ~ AID + H3DA15 + H3FS12 + H3FS11 + H3SP6 + H3WP46 + H3WP53 +
##      H3WP57 + H3DA28 + H3EC26 + H3EC38 + BI0_SEX3 + H30D4A
##
##          Df Deviance   AIC
## - AID      1 2672.7 2732.7
## - H3WP57  2 2675.1 2733.1
## - H3EC26  1 2673.6 2733.6
## - H3FS11  5 2681.6 2733.6
## - H3WP46  5 2682.3 2734.3
## - H3DA15  1 2674.4 2734.4
## <none>    2672.6 2734.6
## - H3FS12  1 2676.2 2736.2
## + H3DA36  1 2672.5 2736.5
## - H3EC38  2 2682.6 2740.6
## - H3DA28  1 2680.8 2740.8
## + H3EC7   8 2664.7 2742.7
## + H3EC3   8 2665.6 2743.6
## - H3WP53  5 2698.8 2750.8
## - BI0_SEX3 1 2711.0 2771.0
## - H30D4A  1 2748.7 2808.7
## - H3SP6   3 2760.1 2816.1
##
## Step: AIC=2732.7
## H3ID15 ~ H3DA15 + H3FS12 + H3FS11 + H3SP6 + H3WP46 + H3WP53 +
##      H3WP57 + H3DA28 + H3EC26 + H3EC38 + BI0_SEX3 + H30D4A
##
##          Df Deviance   AIC
## - H3WP57  2 2675.3 2731.3
## - H3EC26  1 2673.7 2731.7
## - H3FS11  5 2681.7 2731.7
## - H3WP46  5 2682.5 2732.5
## - H3DA15  1 2674.6 2732.6
## <none>    2672.7 2732.7
## - H3FS12  1 2676.3 2734.3
## + AID     1 2672.6 2734.6
## + H3DA36  1 2672.7 2734.7
## - H3EC38  2 2682.8 2738.8
## - H3DA28  1 2681.0 2739.0
## + H3EC7   8 2664.9 2740.9

```

```

## + H3EC3     8  2665.8 2741.8
## - H3WP53    5  2698.9 2748.9
## - BI0_SEX3   1  2711.1 2769.1
## - H30D4A    1  2749.0 2807.0
## - H3SP6     3  2760.2 2814.2
##
## Step: AIC=2731.27
## H3ID15 ~ H3DA15 + H3FS12 + H3FS11 + H3SP6 + H3WP46 + H3WP53 +
##      H3DA28 + H3EC26 + H3EC38 + BI0_SEX3 + H30D4A
##
##          Df Deviance   AIC
## - H3FS11    5  2684.2 2730.2
## - H3EC26    1  2676.3 2730.3
## - H3DA15    1  2677.1 2731.1
## <none>        2675.3 2731.3
## + H3WP57    2  2672.7 2732.7
## - H3WP46    5  2686.8 2732.8
## - H3FS12    1  2679.0 2733.0
## + AID       1  2675.1 2733.1
## + H3DA36    1  2675.2 2733.2
## - H3DA28    1  2683.5 2737.5
## - H3EC38    2  2685.7 2737.7
## + H3EC7     8  2667.5 2739.5
## + H3EC3     8  2668.4 2740.4
## - H3WP53    5  2702.1 2748.1
## - BI0_SEX3   1  2713.6 2767.6
## - H30D4A    1  2752.5 2806.5
## - H3SP6     3  2763.2 2813.2
##
## Step: AIC=2730.2
## H3ID15 ~ H3DA15 + H3FS12 + H3SP6 + H3WP46 + H3WP53 + H3DA28 +
##      H3EC26 + H3EC38 + BI0_SEX3 + H30D4A
##
##          Df Deviance   AIC
## - H3EC26    1  2685.2 2729.2
## - H3DA15    1  2686.0 2730.0
## <none>        2684.2 2730.2
## - H3FS12    2  2689.0 2731.0
## + H3FS11    5  2675.3 2731.3
## + H3WP57    2  2681.7 2731.7
## + AID       1  2684.0 2732.0
## - H3WP46    5  2696.1 2732.1
## + H3DA36    1  2684.2 2732.2
## - H3DA28    1  2692.1 2736.1
## - H3EC38    2  2694.9 2736.9
## + H3EC7     8  2676.6 2738.6
## + H3EC3     8  2677.3 2739.3
## - H3WP53    5  2711.4 2747.4
## - BI0_SEX3   1  2724.4 2768.4
## - H30D4A    1  2763.1 2807.1
## - H3SP6     3  2774.2 2814.2
##
## Step: AIC=2729.17
## H3ID15 ~ H3DA15 + H3FS12 + H3SP6 + H3WP46 + H3WP53 + H3DA28 +
##      H3EC38 + BI0_SEX3 + H30D4A
##

```

```

##          Df Deviance   AIC
## - H3DA15    1  2686.8 2728.8
## <none>           2685.2 2729.2
## - H3FS12    2  2689.8 2729.8
## + H3EC26    1  2684.2 2730.2
## + H3FS11    5  2676.3 2730.3
## + H3WP57    2  2682.7 2730.7
## + AID       1  2685.0 2731.0
## + H3DA36    1  2685.2 2731.2
## - H3WP46    5  2697.2 2731.2
## - H3DA28    1  2693.8 2735.8
## + H3EC7     8  2677.5 2737.5
## + H3EC3     8  2678.3 2738.3
## - H3EC38    2  2699.7 2739.7
## - H3WP53    5  2712.5 2746.5
## - BI0_SEX3   1  2726.1 2768.1
## - H30D4A    1  2763.3 2805.3
## - H3SP6     3  2775.7 2813.7
##
## Step:  AIC=2728.77
## H3ID15 ~ H3FS12 + H3SP6 + H3WP46 + H3WP53 + H3DA28 + H3EC38 +
##      BI0_SEX3 + H30D4A
##
##          Df Deviance   AIC
## <none>           2686.8 2728.8
## + H3DA15    1  2685.2 2729.2
## - H3FS12    2  2691.4 2729.4
## + H3FS11    5  2677.9 2729.9
## + H3EC26    1  2686.0 2730.0
## + H3WP57    2  2684.3 2730.3
## + AID       1  2686.6 2730.6
## - H3WP46    5  2698.7 2730.7
## + H3DA36    1  2686.7 2730.7
## - H3DA28    1  2695.5 2735.5
## + H3EC7     8  2679.2 2737.2
## + H3EC3     8  2679.9 2737.9
## - H3EC38    2  2700.7 2738.7
## - H3WP53    5  2713.6 2745.6
## - BI0_SEX3   1  2728.2 2768.2
## - H30D4A    1  2764.5 2804.5
## - H3SP6     3  2777.0 2813.0

```

```

# Summary of selected model
summary(step_model)

```

```

## 
## Call:
## glm(formula = H3ID15 ~ H3FS12 + H3SP6 + H3WP46 + H3WP53 + H3DA28 +
##      H3EC38 + BI0_SEX3 + H30D4A, family = binomial, data = data)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 1.9588    0.4219   4.643 3.44e-06 ***
## H3FS12Friends more influential -0.4218    0.2160  -1.953 0.050795 .
## H3FS12Not Asked             -0.3043    0.1633  -1.864 0.062362 .
## H3SP6Most/all of the time   -0.3105    0.3181  -0.976 0.329062
## H3SP6Never or rarely        1.3151    0.2050   6.414 1.42e-10 ***
## H3SP6Sometimes              0.5252    0.2176   2.413 0.015806 *
## H3WP46Not applicable / skipped 0.1898    0.2577   0.737 0.461390
## H3WP46Not close at all     -1.2005    0.8492  -1.414 0.157442
## H3WP46Not very close       -0.7283    0.4878  -1.493 0.135433
## H3WP46Quite close          0.8826    0.4791   1.842 0.065452 .
## H3WP46Somewhat close       -0.2108    0.4364  -0.483 0.628979
## H3WP53Not applicable / skipped -0.3043   0.2583  -1.178 0.238730
## H3WP53Not close at all     -1.5129    0.3902  -3.877 0.000106 ***
## H3WP53Not very close       -1.0974    0.3280  -3.346 0.000819 ***
## H3WP53Quite close          0.3333    0.3309  -1.007 0.313905
## H3WP53Somewhat close       -0.4753    0.3199  -1.486 0.137345
## H3DA28Yes                  0.3396    0.1132   3.000 0.002703 **
## H3EC38Not Asked            -0.8125    0.3972  -2.046 0.040776 *
## H3EC38Yes                  -0.5263    0.1550  -3.395 0.000685 ***
## BI0_SEX3Male                0.7007    0.1117   6.274 3.52e-10 ***
## H30D4AWhite                -1.1196    0.1386  -8.079 6.56e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2985.6 on 4287 degrees of freedom
## Residual deviance: 2686.8 on 4267 degrees of freedom
## AIC: 2728.8
##
## Number of Fisher Scoring iterations: 5

```

Regression By Gender

```

# Load necessary library
library(dplyr)

# Filter data into male and female subsets
male_data <- filter(data, BI0_SEX3 == "Male")
female_data <- filter(data, BI0_SEX3 == "Female")

# Run logistic regression for males
male_model <- glm(H3ID15 ~ H3FS12 + H3SP6 + H3WP53 +
                   H3DA28 + H3EC38 + H30D4A,
                   family = binomial, data = male_data)
summary(male_model)

```

```

## 
## Call:
## glm(formula = H3ID15 ~ H3FS12 + H3SP6 + H3WP53 + H3DA28 + H3EC38 +
##      H30D4A, family = binomial, data = male_data)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                2.60630   0.64052   4.069 4.72e-05 ***
## H3FS12Friends more influential -0.31600   0.42404  -0.745 0.456146
## H3FS12Not Asked            -0.41325   0.32011  -1.291 0.196707
## H3SP6Most/all of the time  -0.40417   0.57430  -0.704 0.481580
## H3SP6Never or rarely       1.37105   0.36677   3.738 0.000185 ***
## H3SP6Sometimes              0.66992   0.39725   1.686 0.091718 .
## H3WP53Not applicable / skipped -0.02455   0.43814  -0.056 0.955324
## H3WP53Not close at all     -1.98211   0.75217  -2.635 0.008409 **
## H3WP53Not very close       -0.25378   0.68556  -0.370 0.711255
## H3WP53Quite close          -0.03774   0.58267  -0.065 0.948354
## H3WP53Somewhat close       -0.70892   0.52784  -1.343 0.179252
## H3DA28Yes                  0.49182   0.20899   2.353 0.018609 *
## H3EC38Not Asked            -1.66290   0.58713  -2.832 0.004622 **
## H3EC38Yes                  -0.66702   0.39926  -1.671 0.094793 .
## H30D4AWhite                -1.21154   0.26566  -4.560 5.10e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 984.84  on 1973  degrees of freedom
## Residual deviance: 913.10  on 1959  degrees of freedom
## AIC: 943.1
##
## Number of Fisher Scoring iterations: 6

```

```

# Run logistic regression for females
female_model <- glm(H3ID15 ~ H3FS12 + H3SP6 + H3WP53 +
                      H3DA28 + H3EC38 + H30D4A,
                      family = binomial, data = female_data)
summary(female_model)

```

```

## 
## Call:
## glm(formula = H3ID15 ~ H3FS12 + H3SP6 + H3WP53 + H3DA28 + H3EC38 +
##      H3OD4A, family = binomial, data = female_data)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.1784    0.4314   5.050 4.42e-07 ***
## H3FS12Friends more influential -0.4834    0.2510  -1.926 0.054073 .
## H3FS12Not Asked             -0.2702    0.1895  -1.426 0.153979
## H3SP6Most/all of the time   -0.3469    0.3790  -0.915 0.359957
## H3SP6Never or rarely        1.2780    0.2463   5.188 2.12e-07 ***
## H3SP6Sometimes               0.4509    0.2593   1.739 0.082062 .
## H3WP53Not applicable / skipped -0.3334    0.3102  -1.075 0.282524
## H3WP53Not close at all       -1.3622    0.4491  -3.033 0.002421 **
## H3WP53Not very close         -1.3113    0.3827  -3.426 0.000612 ***
## H3WP53Quite close            -0.3728    0.3955  -0.943 0.345869
## H3WP53Somewhat close          -0.1962    0.3949  -0.497 0.619268
## H3DA28Yes                   0.2822    0.1338   2.109 0.034926 *
## H3EC38Not Asked              -0.4148    0.5222  -0.794 0.426996
## H3EC38Yes                   -0.5239    0.1666  -3.145 0.001659 **
## H3OD4AWhite                  -1.0926    0.1622  -6.738 1.61e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 1931.5 on 2313 degrees of freedom
## Residual deviance: 1773.7 on 2299 degrees of freedom
## AIC: 1803.7
##
## Number of Fisher Scoring iterations: 5

```

Regression by Income

```

# Create binary variable: 1 = Low income, 0 = Others
data$LowIncome <- ifelse(data$H3EC3 %in% c("Less than $10,000", "$10,000-14,999", "$1
5,000-19,999"), 1, 0)

interaction_model <- glm(H3ID15 ~ H3SP6 * LowIncome + H3FS12 + H3DA28 +
                           H3EC38 + BIO_SEX3 + H3OD4A,
                           family = binomial, data = data)

summary(interaction_model)

```

```

## 
## Call:
## glm(formula = H3ID15 ~ H3SP6 * LowIncome + H3FS12 + H3DA28 +
##      H3EC38 + BI0_SEX3 + H3OD4A, family = binomial, data = data)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                  1.7352   0.2748   6.315 2.70e-10 ***
## H3SP6Most/all of the time   -0.1725   0.3396  -0.508 0.611444
## H3SP6Never or rarely        1.4026   0.2131   6.580 4.70e-11 ***
## H3SP6Sometimes               0.5135   0.2260   2.272 0.023062 *
## LowIncome                    0.2792   0.6420   0.435 0.663594
## H3FS12Friends more influential -0.4787  0.2144  -2.233 0.025558 *
## H3FS12Not Asked              -0.3610  0.1620  -2.229 0.025811 *
## H3DA28Yes                   0.3135   0.1123   2.791 0.005253 **
## H3EC38Not Asked              -0.8029  0.3980  -2.018 0.043639 *
## H3EC38Yes                   -0.5698  0.1514  -3.763 0.000168 ***
## BI0_SEX3Male                 0.7397   0.1112   6.655 2.84e-11 ***
## H3OD4AWhite                  -1.0695  0.1358  -7.876 3.39e-15 ***
## H3SP6Most/all of the time:LowIncome -1.0275  0.9238  -1.112 0.266049
## H3SP6Never or rarely:LowIncome     -0.5000  0.6669  -0.750 0.453444
## H3SP6Sometimes:LowIncome         0.1528   0.7231   0.211 0.832662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 2985.6 on 4287 degrees of freedom
## Residual deviance: 2720.9 on 4273 degrees of freedom
## AIC: 2750.9
##
## Number of Fisher Scoring iterations: 5

```