Child Marriage Phenomenon In Vietnam

Holly Duong

2024-02-29

# load packages  
library(readr)  
library(dplyr)  
library(ggplot2)  
library(gridExtra)  
library(corrplot)  
library(plotly)  
library(reshape2)  
library(car)  
library(kableExtra)  
library(broom)  
library(knitr)  
library(pROC)  
library(ggpattern)  
library(tidyr)  
library(ResourceSelection)  
library(sf)

### Setting up Data

# import dataset  
female <- read\_csv("/Users/hollyduong/Desktop/DA 401/ChildMarriageInVietnam/Data/wm.csv")

# get a glimpse of dataset  
head(female)

## # A tibble: 6 × 458  
## HH1 HH2 LN WM1 WM2 WM3 WMINT WM4 WM5 WM6D WM6M WM6Y WM8  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 2 3 1 2 3 92 91 92 19 11 2020 2  
## 2 1 4 2 1 4 2 92 91 92 18 11 2020 1  
## 3 1 9 4 1 9 4 92 91 92 18 11 2020 1  
## 4 1 10 4 1 10 4 92 91 92 18 11 2020 2  
## 5 1 11 4 1 11 4 92 91 92 19 11 2020 2  
## 6 1 11 5 1 11 5 92 91 92 18 11 2020 2  
## # ℹ 445 more variables: WM9 <dbl>, WM17 <dbl>, WM7H <dbl>, WM7M <dbl>,  
## # WM10H <dbl>, WM10M <dbl>, WM11 <dbl>, WM12 <dbl>, WM13 <dbl>, WM14 <dbl>,  
## # WM15 <dbl>, WM22 <dbl>, WM23 <dbl>, WM24 <dbl>, WMHINT <dbl>, WMFIN <dbl>,  
## # WB3M <dbl>, WB3Y <dbl>, WB4 <dbl>, WB5 <dbl>, WB6A <dbl>, WB6B <dbl>,  
## # WB7 <dbl>, WB9 <lgl>, WB10A <lgl>, WB10B <lgl>, WB11 <lgl>, WB12A <lgl>,  
## # WB12B <lgl>, WB14 <dbl>, WB15 <dbl>, WB16 <dbl>, WB17 <dbl>, WB18 <dbl>,  
## # WB19A <chr>, WB19B <chr>, WB19C <chr>, WB19D <chr>, WB19E <chr>, …

# Select the specified columns to create a new dataframe  
female\_df <- select(female, WAGEM, MSTATUS, HH6, HH7, welevel, insurance, ethnicity, windex5, CP2, HA1, MT4, MT9, MT11)  
  
# View the first few rows of the new dataframe  
summary(female\_df)

## WAGEM MSTATUS HH6 HH7 welevel   
## Min. :10.00 Min. :1.000 Min. :1.000 Min. :1.000 Min. :0.00   
## 1st Qu.:18.00 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:2.000 1st Qu.:1.00   
## Median :20.00 Median :1.000 Median :2.000 Median :3.000 Median :2.00   
## Mean :20.92 Mean :1.413 Mean :1.683 Mean :3.404 Mean :2.46   
## 3rd Qu.:23.00 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:5.000 3rd Qu.:3.00   
## Max. :47.00 Max. :9.000 Max. :2.000 Max. :6.000 Max. :9.00   
## NA's :2026 NA's :11 NA's :11 NA's :11 NA's :524   
## insurance ethnicity windex5 CP2   
## Min. :1.000 Min. :1.000 Min. :0.000 Min. :1.000   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000   
## Median :1.000 Median :1.000 Median :2.000 Median :1.000   
## Mean :1.135 Mean :2.035 Mean :2.494 Mean :1.431   
## 3rd Qu.:1.000 3rd Qu.:3.000 3rd Qu.:4.000 3rd Qu.:2.000   
## Max. :9.000 Max. :6.000 Max. :5.000 Max. :9.000   
## NA's :524 NA's :864   
## HA1 MT4 MT9 MT11   
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.00   
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.00   
## Median :1.000 Median :2.000 Median :1.000 Median :1.00   
## Mean :1.215 Mean :1.664 Mean :1.365 Mean :1.12   
## 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:2.000 3rd Qu.:1.00   
## Max. :9.000 Max. :9.000 Max. :9.000 Max. :9.00   
## NA's :524 NA's :524 NA's :2496 NA's :524

# Rename the columns  
female\_df <- female\_df %>%  
 rename(  
 age\_first\_marriage = WAGEM,  
 marital\_status = MSTATUS,  
 area = HH6,  
 region = HH7,  
 education\_level = welevel,  
 health\_insurance = insurance,  
 ethnicity = ethnicity,  
 wealth\_index = windex5,  
 current\_contraceptive\_use = CP2,  
 awareness\_hiv\_aids = HA1,  
 used\_computer\_tablet = MT4,  
 used\_internet = MT9,  
 owns\_mobile\_phone = MT11  
)

# Summarize missing values by column  
summarize\_missing <- sapply(female\_df, function(x) sum(is.na(x)))  
print(summarize\_missing)

## age\_first\_marriage marital\_status area   
## 2026 11 11   
## region education\_level health\_insurance   
## 11 524 524   
## ethnicity wealth\_index current\_contraceptive\_use   
## 0 0 864   
## awareness\_hiv\_aids used\_computer\_tablet used\_internet   
## 524 524 2496   
## owns\_mobile\_phone   
## 524

# Recode the value 9 to NA for specified variables  
female\_df <- female\_df %>%  
 mutate(  
 current\_contraceptive\_use = na\_if(current\_contraceptive\_use, 9),  
 used\_internet = na\_if(used\_internet, 9),  
 health\_insurance = na\_if(health\_insurance, 9),  
 education\_level = na\_if(education\_level, 9),  
 awareness\_hiv\_aids = na\_if(awareness\_hiv\_aids, 9),  
 used\_computer\_tablet = na\_if(used\_computer\_tablet, 9),  
 owns\_mobile\_phone = na\_if(owns\_mobile\_phone, 9),  
 marital\_status = na\_if(marital\_status, 9)  
 )  
  
# Recode the value 6 to NA for 'ethnicity'  
female\_df$ethnicity <- na\_if(female\_df$ethnicity, 6)  
  
# Recode the value 0 to NA for 'wealth\_index'  
female\_df$wealth\_index <- na\_if(female\_df$wealth\_index, 0)

# Recoding variables from 1 (Yes) and 2 (No) to 1 (Yes) and 0 (No)  
female\_df$health\_insurance <- ifelse(female\_df$health\_insurance == 2, 0, female\_df$health\_insurance)  
female\_df$current\_contraceptive\_use <- ifelse(female\_df$current\_contraceptive\_use == 2, 0, female\_df$current\_contraceptive\_use)  
female\_df$awareness\_hiv\_aids <- ifelse(female\_df$awareness\_hiv\_aids == 2, 0, female\_df$awareness\_hiv\_aids)  
female\_df$used\_computer\_tablet <- ifelse(female\_df$used\_computer\_tablet == 2, 0, female\_df$used\_computer\_tablet)  
female\_df$owns\_mobile\_phone <- ifelse(female\_df$owns\_mobile\_phone == 2, 0, female\_df$owns\_mobile\_phone)  
female\_df$used\_internet <- ifelse(female\_df$used\_internet == 2, 0, female\_df$used\_internet)

# View the changes to ensure NA substitution has been correctly applied  
summary(female\_df)

## age\_first\_marriage marital\_status area region   
## Min. :10.00 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:18.00 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:2.000   
## Median :20.00 Median :1.000 Median :2.000 Median :3.000   
## Mean :20.92 Mean :1.408 Mean :1.683 Mean :3.404   
## 3rd Qu.:23.00 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:5.000   
## Max. :47.00 Max. :3.000 Max. :2.000 Max. :6.000   
## NA's :2026 NA's :18 NA's :11 NA's :11   
## education\_level health\_insurance ethnicity wealth\_index   
## Min. :0.00 Min. :0.0000 Min. :1.000 Min. :1.000   
## 1st Qu.:1.00 1st Qu.:1.0000 1st Qu.:1.000 1st Qu.:1.000   
## Median :2.00 Median :1.0000 Median :1.000 Median :2.000   
## Mean :2.46 Mean :0.8659 Mean :1.586 Mean :2.615   
## 3rd Qu.:3.00 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:4.000   
## Max. :5.00 Max. :1.0000 Max. :4.000 Max. :5.000   
## NA's :525 NA's :525 NA's :1148 NA's :524   
## current\_contraceptive\_use awareness\_hiv\_aids used\_computer\_tablet  
## Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:0.0000   
## Median :1.0000 Median :1.0000 Median :0.0000   
## Mean :0.5842 Mean :0.7975 Mean :0.3412   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000   
## NA's :885 NA's :541 NA's :532   
## used\_internet owns\_mobile\_phone  
## Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:1.0000   
## Median :1.0000 Median :1.0000   
## Mean :0.6394 Mean :0.8831   
## 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000   
## NA's :2501 NA's :529

### Creating New Variable Access to Media

# Combine individual variables for access to the internet, phone, and computer into a single variable  
# This new variable "access\_to\_media" will have a value of 1 if the respondent has access to any of these media sources, and 0 if not  
# This provides a more comprehensive measure of media access  
female\_df <- female\_df %>%  
 mutate(access\_to\_media = ifelse(used\_computer\_tablet == 1 | used\_internet == 1 | owns\_mobile\_phone == 1, 1, 0))  
  
# In later analysis, use access\_to\_media instead of the 3 separate variables

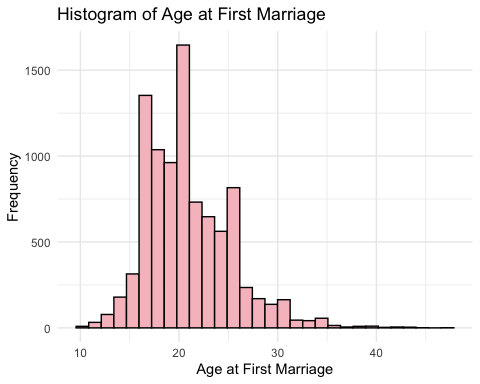
# Exporting female\_df to a CSV file in the current working directory  
#write.csv(female\_df, "female\_df.csv", row.names = FALSE)

### Distributions of Data

# Histogram for 'Age at First Marriage'  
afm\_hist <- ggplot(female\_df, aes(x = age\_first\_marriage)) +  
 geom\_histogram(fill = "#F7C0C8", color = "black") +  
 theme\_minimal() +  
 ggtitle("Histogram of Age at First Marriage") +  
 xlab("Age at First Marriage") +  
 ylab("Frequency")  
  
print(afm\_hist)

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 2026 rows containing non-finite outside the scale range  
## (`stat\_bin()`).



### Handling Missing Data

# Before that, let's double check the count of missing values by column  
count\_missing <- sapply(female\_df, function(x) sum(is.na(x)))  
print(count\_missing)

## age\_first\_marriage marital\_status area   
## 2026 18 11   
## region education\_level health\_insurance   
## 11 525 525   
## ethnicity wealth\_index current\_contraceptive\_use   
## 1148 524 885   
## awareness\_hiv\_aids used\_computer\_tablet used\_internet   
## 541 532 2501   
## owns\_mobile\_phone access\_to\_media   
## 529 529

# Make a copy of female\_df for imputation  
imputed\_df <- female\_df

# Handle missing values in 'Age at First Marriage'  
  
# Calculate the median value for 'Age at First Marriage', excluding NA values  
median\_age\_first\_marriage <- median(imputed\_df$age\_first\_marriage, na.rm = TRUE)  
  
# Impute missing values in 'Age at First Marriage' with the median value  
imputed\_df$age\_first\_marriage[is.na(imputed\_df$age\_first\_marriage)] <- median\_age\_first\_marriage

# Define a function to calculate mode for categorical variables  
getMode <- function(v) {  
 # The mode is the value that appears most frequently in the data  
 uniqv <- unique(na.omit(v)) # Omit NA values and get unique values  
 uniqv[which.max(tabulate(match(v, uniqv)))] # Return the value with the highest frequency  
}

# For 'ethnicity', an ordinal variable with predefined categories, it makes sense to impute missing values with the mode.  
mode\_ethnicity <- getMode(imputed\_df$ethnicity)  
imputed\_df <- mutate(imputed\_df, ethnicity = ifelse(is.na(ethnicity), mode\_ethnicity, ethnicity))  
mode\_area <- getMode(imputed\_df$area)  
imputed\_df <- mutate(imputed\_df, area = ifelse(is.na(area), mode\_area, area))  
mode\_region <- getMode(imputed\_df$region)  
imputed\_df <- mutate(imputed\_df, region = ifelse(is.na(region), mode\_region, region))  
mode\_marital\_status <- getMode(imputed\_df$marital\_status)  
imputed\_df <- mutate(imputed\_df, marital\_status = ifelse(is.na(marital\_status), mode\_marital\_status, marital\_status))  
  
  
# Binary variables like 'current\_contraceptive\_use', 'health\_insurance','awareness\_hiv\_aids', 'used\_internet', 'used\_computer\_tablet', 'owns\_mobile\_phone', and 'access\_to\_media'  
# should be imputed with the mode since it represents the most frequent category (either 0 or 1).  
  
# Calculate the mode for each binary variable  
mode\_used\_internet <- getMode(imputed\_df$used\_internet)  
mode\_current\_contraceptive\_use <- getMode(imputed\_df$current\_contraceptive\_use)  
mode\_health\_insurance <- getMode(imputed\_df$health\_insurance)  
mode\_awareness\_hiv\_aids <- getMode(imputed\_df$awareness\_hiv\_aids)  
mode\_used\_computer\_tablet <- getMode(imputed\_df$used\_computer\_tablet)  
mode\_owns\_mobile\_phone <- getMode(imputed\_df$owns\_mobile\_phone)  
mode\_access\_to\_media <- getMode(imputed\_df$access\_to\_media)  
  
# Impute missing values for binary variables  
imputed\_df <- mutate(imputed\_df,  
 used\_internet = ifelse(is.na(used\_internet), mode\_used\_internet, used\_internet),  
 current\_contraceptive\_use = ifelse(is.na(current\_contraceptive\_use), mode\_current\_contraceptive\_use, current\_contraceptive\_use),  
 health\_insurance = ifelse(is.na(health\_insurance), mode\_health\_insurance, health\_insurance),  
 awareness\_hiv\_aids = ifelse(is.na(awareness\_hiv\_aids), mode\_awareness\_hiv\_aids, awareness\_hiv\_aids),  
 used\_computer\_tablet = ifelse(is.na(used\_computer\_tablet), mode\_used\_computer\_tablet, used\_computer\_tablet),  
 owns\_mobile\_phone = ifelse(is.na(owns\_mobile\_phone), mode\_owns\_mobile\_phone, owns\_mobile\_phone),  
 access\_to\_media = ifelse(is.na(access\_to\_media), mode\_access\_to\_media, access\_to\_media)  
)  
  
# 'education\_level' is an ordinal variable where the median could be a more suitable measure of central tendency than the mode.  
# However, given the categorical nature of the levels (e.g., "Primary", "Secondary"), using the mode may still be appropriate.  
mode\_education\_level <- getMode(imputed\_df$education\_level)  
imputed\_df <- mutate(imputed\_df, education\_level = ifelse(is.na(education\_level), mode\_education\_level, education\_level))  
  
# 'wealth\_index' is an ordinal variable where the median could be a more suitable measure of central tendency than the mode.  
# However, given the categorical nature of the levels (e.g., "Poorest", "Poor",...), using the mode may still be appropriate.  
mode\_wealth\_index <- getMode(imputed\_df$wealth\_index)  
imputed\_df <- mutate(imputed\_df, wealth\_index = ifelse(is.na(wealth\_index), mode\_wealth\_index, wealth\_index))

# Check the resulting dataset to confirm changes  
summary(imputed\_df)

## age\_first\_marriage marital\_status area region   
## Min. :10.00 Min. :1.000 Min. :1.000 Min. :1.000   
## 1st Qu.:18.00 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:2.000   
## Median :20.00 Median :1.000 Median :2.000 Median :3.000   
## Mean :20.76 Mean :1.408 Mean :1.684 Mean :3.403   
## 3rd Qu.:23.00 3rd Qu.:1.000 3rd Qu.:2.000 3rd Qu.:5.000   
## Max. :47.00 Max. :3.000 Max. :2.000 Max. :6.000   
## education\_level health\_insurance ethnicity wealth\_index   
## Min. :0.000 Min. :0.0000 Min. :1.000 Min. :1.00   
## 1st Qu.:1.000 1st Qu.:1.0000 1st Qu.:1.000 1st Qu.:1.00   
## Median :2.000 Median :1.0000 Median :1.000 Median :2.00   
## Mean :2.438 Mean :0.8721 Mean :1.527 Mean :2.54   
## 3rd Qu.:3.000 3rd Qu.:1.0000 3rd Qu.:2.000 3rd Qu.:4.00   
## Max. :5.000 Max. :1.0000 Max. :4.000 Max. :5.00   
## current\_contraceptive\_use awareness\_hiv\_aids used\_computer\_tablet  
## Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:0.0000   
## Median :1.0000 Median :1.0000 Median :0.0000   
## Mean :0.6168 Mean :0.8072 Mean :0.3251   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000   
## used\_internet owns\_mobile\_phone access\_to\_media   
## Min. :0.0000 Min. :0.0000 Min. :0.0000   
## 1st Qu.:0.0000 1st Qu.:1.0000 1st Qu.:1.0000   
## Median :1.0000 Median :1.0000 Median :1.0000   
## Mean :0.7192 Mean :0.8886 Mean :0.9071   
## 3rd Qu.:1.0000 3rd Qu.:1.0000 3rd Qu.:1.0000   
## Max. :1.0000 Max. :1.0000 Max. :1.0000

# After imputation, let's check for missing values  
count\_imputation <- sapply(imputed\_df, function(x) sum(is.na(x)))  
print(count\_imputation)

## age\_first\_marriage marital\_status area   
## 0 0 0   
## region education\_level health\_insurance   
## 0 0 0   
## ethnicity wealth\_index current\_contraceptive\_use   
## 0 0 0   
## awareness\_hiv\_aids used\_computer\_tablet used\_internet   
## 0 0 0   
## owns\_mobile\_phone access\_to\_media   
## 0 0

### Visualization Comparing Before vs. After Imputation

# Histogram for 'age\_first\_marriage' before imputation  
afm\_0 <- ggplot(female\_df, aes(x = age\_first\_marriage)) +   
 geom\_histogram(fill = "#F7C0C8", color = "black", bins = 30) +  
 theme\_light() +  
 ggtitle("Before Imputation") +  
 xlab("Age at First Marriage") +  
 ylab("Frequency")  
  
# Histogram for 'age\_first\_marriage' after imputation  
afm\_imputed <- ggplot(imputed\_df, aes(x = age\_first\_marriage)) +   
 geom\_histogram(fill = "#E83853", color = "black", bins = 30) +  
 theme\_light() +  
 ggtitle("After Imputation") +  
 xlab("Age at First Marriage") +  
 ylab("Frequency")  
  
# Arrange the two plots side by side  
grid.arrange(afm\_0, afm\_imputed, ncol = 2)

## Warning: Removed 2026 rows containing non-finite outside the scale range  
## (`stat\_bin()`).



# Arrange the two plots side by side and capture the layout as a grob  
combined\_plots <- arrangeGrob(afm\_0, afm\_imputed, ncol = 2)

## Warning: Removed 2026 rows containing non-finite outside the scale range  
## (`stat\_bin()`).

# Now, use ggsave to save the combined plot  
#ggsave("combined\_age\_first\_marriage.png", plot = combined\_plots, width = 10, height = 5)

### Creating Binary Variables for Child Marriage Under 18 and 16

# Convert "age at first marriage" into a binary variable to indicate child marriage  
# Child marriage is defined as marriage before the age of 18  
# The new binary variable "child\_marriage" will have a value of 1 if the marriage occurred before age 18, and 0 otherwise  
imputed\_df <- imputed\_df %>%  
 mutate(child\_marriage = ifelse(age\_first\_marriage < 18, 1, 0))

# Create a binary variable for child marriage under 16  
# The new variable "child\_marriage\_u16" will have a value of 1 if the marriage occurred before age 16, and 0 otherwise  
imputed\_df <- imputed\_df %>%  
 mutate(child\_marriage\_u16 = ifelse(age\_first\_marriage < 16, 1, 0))

# Move "child\_marriage" and "child\_marriage\_u16" to the front of the dataframe  
imputed\_df <- imputed\_df %>%  
 select(child\_marriage, child\_marriage\_u16, everything())

# Exporting female\_df to a CSV file in the current working directory  
#write.csv(imputed\_df, "imputed\_df.csv", row.names = FALSE)

### EDA

#### Vietnam Map with Regions and Their Percentages of Child Marriage (under 18)

# Aggregate the data by region to get the total number of child marriages under 18 per region  
region\_counts <- aggregate(child\_marriage ~ region, data = imputed\_df, FUN = sum)  
  
# Calculate the total number of child marriages under 18 in the dataset  
total\_child\_marriages <- sum(region\_counts$child\_marriage)  
  
# Calculate the percentage for each region  
region\_counts$married\_u18\_perc\_of\_total <- (region\_counts$child\_marriage / total\_child\_marriages) \* 100

# Mapping region numbers to names  
region\_names <- c("Red River Delta", "Northern Midlands And Mountain",   
 "North Central And Central Coastal", "Central Highlands",   
 "South East", "Mekong River Delta")  
names(region\_counts)[1] <- "region\_name"  
region\_counts$region\_name <- factor(region\_counts$region\_name, levels = 1:6, labels = region\_names)  
  
# Display the final data frame  
print(region\_counts)

## region\_name child\_marriage married\_u18\_perc\_of\_total  
## 1 Red River Delta 142 7.226463  
## 2 Northern Midlands And Mountain 888 45.190840  
## 3 North Central And Central Coastal 215 10.941476  
## 4 Central Highlands 271 13.791349  
## 5 South East 194 9.872774  
## 6 Mekong River Delta 255 12.977099

# Updated mapping including all provinces and cities in the Red River Delta  
province\_to\_region <- data.frame(  
 NAME\_1 = c(  
 'Bắc Ninh', 'Hà Nam', 'Hà Nội', 'Hải Dương', 'Hải Phòng', 'Hoà Bình', 'Hưng Yên', 'Nam Định', 'Ninh Bình', 'Thái Bình', 'Vĩnh Phúc', # Red River Delta 11  
   
 'Bắc Giang', 'Bắc Kạn', 'Cao Bằng', 'Hà Giang', 'Lạng Sơn', 'Lào Cai', 'Phú Thọ', 'Quảng Ninh', 'Thái Nguyên', 'Tuyên Quang', 'Yên Bái', 'Điện Biên', 'Lai Châu', 'Sơn La', # Northern Midlands And Mountain 14  
   
 'Bình Định', 'Bình Thuận', 'Khánh Hòa', 'Ninh Thuận', 'Phú Yên', 'Quảng Nam', 'Quảng Ngãi', 'Thừa Thiên Huế', 'Đà Nẵng', 'Hà Tĩnh', 'Nghệ An', 'Quảng Bình', 'Quảng Trị', 'Thanh Hóa', # North Central And Central Coastal 14  
   
 'Đắk Lắk', 'Đắk Nông', 'Gia Lai', 'Kon Tum', 'Lâm Đồng', # Central Highlands 5  
   
 'Bà Rịa - Vũng Tàu', 'Bình Dương', 'Bình Phước', 'Đồng Nai', 'Hồ Chí Minh', 'Tây Ninh', # South East 6  
   
 'An Giang', 'Bạc Liêu', 'Bến Tre', 'Cà Mau', 'Cần Thơ', 'Đồng Tháp', 'Hậu Giang', 'Kiên Giang', 'Long An', 'Sóc Trăng', 'Tiền Giang', 'Trà Vinh', 'Vĩnh Long' # Mekong River Delta 13  
 ),  
 Region = c(  
 rep('Red River Delta', 11),   
 rep('Northern Midlands And Mountain', 14),   
 rep('North Central And Central Coastal', 14),   
 rep('Central Highlands', 5),   
 rep('South East', 6),   
 rep('Mekong River Delta', 13)  
 )  
)

# Check the mapping  
#print(province\_to\_region)

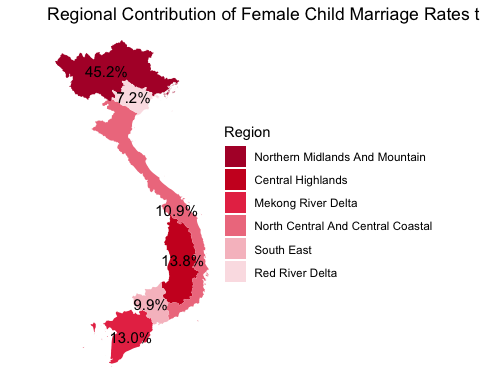
# Read shapefile  
vietnam\_shape <- st\_read('/Users/hollyduong/Desktop/DA 401/ChildMarriageInVietnam/gadm41\_VNM\_shp')

## Multiple layers are present in data source /Users/hollyduong/Desktop/DA 401/ChildMarriageInVietnam/gadm41\_VNM\_shp, reading layer `gadm41\_VNM\_2'.  
## Use `st\_layers' to list all layer names and their type in a data source.  
## Set the `layer' argument in `st\_read' to read a particular layer.  
## Reading layer `gadm41\_VNM\_2' from data source   
## `/Users/hollyduong/Desktop/DA 401/ChildMarriageInVietnam/gadm41\_VNM\_shp'   
## using driver `ESRI Shapefile'  
## Simple feature collection with 710 features and 13 fields  
## Geometry type: MULTIPOLYGON  
## Dimension: XY  
## Bounding box: xmin: 102.1446 ymin: 8.381355 xmax: 109.4692 ymax: 23.39269  
## Geodetic CRS: WGS 84

# Join the shapefile with the province-to-region mapping  
vietnam\_shape\_with\_region <- vietnam\_shape %>%  
 left\_join(province\_to\_region, by = "NAME\_1")  
  
# Aggregate the shapefile data by region  
vietnam\_regions <- vietnam\_shape\_with\_region %>%  
 group\_by(Region) %>%  
 summarise(geometry = st\_union(geometry), .groups = 'drop')

# Join the aggregated shapefile data with the child marriage data  
vietnam\_map\_data <- vietnam\_regions %>%  
 left\_join(region\_counts, by = c("Region" = "region\_name"))  
  
# Define colors with your specific choices  
colors\_ordered <- setNames(c("#B20033", "#CD0A25", "#E83853", "#EF7D8D", "#F7C0C8", "#FBE1E5"),  
 c("Northern Midlands And Mountain", "Central Highlands",   
 "Mekong River Delta", "North Central And Central Coastal",   
 "South East", "Red River Delta"))  
  
# Plot  
mapvn <- ggplot(data = vietnam\_map\_data) +  
 geom\_sf(aes(fill = factor(Region, levels = names(colors\_ordered))), color = NA) +  
 geom\_sf\_text(aes(label = sprintf("%.1f%%", married\_u18\_perc\_of\_total)), size = 4, hjust = 0.5, vjust = 0.5) +  
 scale\_fill\_manual(values = colors\_ordered, name = "Region") +  
 labs(title = "Regional Contribution of Female Child Marriage Rates to Total Rates in Vietnam") +  
 theme\_void() +  
 theme(legend.position = "right")  
  
print(mapvn)

## Warning in st\_point\_on\_surface.sfc(sf::st\_zm(x)): st\_point\_on\_surface may not  
## give correct results for longitude/latitude data

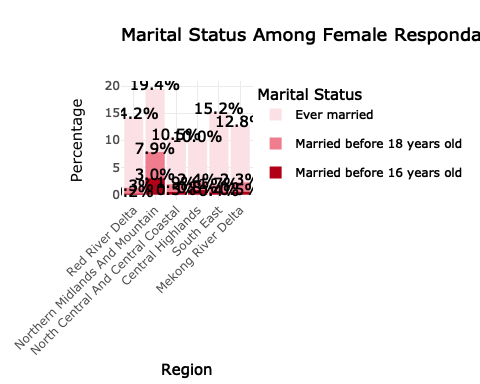


#ggsave("mapvn.png", plot = mapvn, width = 8, height = 6, dpi = 300)

#### Marital Status Among Female Respondants by Region in Vietnam

# Calculate total observations  
total\_obs <- nrow(imputed\_df)  
  
# Aggregate counts for each category by region without altering the original 'region' field  
counts\_df <- aggregate(cbind(ever\_married = imputed\_df$marital\_status %in% c(1, 2),   
 married\_u18 = imputed\_df$child\_marriage == 1,   
 married\_u16 = imputed\_df$child\_marriage\_u16 == 1) ~ region,   
 data = imputed\_df,   
 FUN = sum)  
  
# Convert counts to percentages  
counts\_df$ever\_married <- (counts\_df$ever\_married / total\_obs) \* 100  
counts\_df$married\_u18 <- (counts\_df$married\_u18 / total\_obs) \* 100  
counts\_df$married\_u16 <- (counts\_df$married\_u16 / total\_obs) \* 100

p <- ggplot(counts\_df, aes(x = factor(region))) +  
 geom\_bar(aes(y = ever\_married, fill = "Ever married"), stat = "identity") +  
 geom\_bar(aes(y = married\_u18, fill = "Married before 18 years old"), stat = "identity") +  
 geom\_bar(aes(y = married\_u16, fill = "Married before 16 years old"), stat = "identity") +  
 # Adding text labels for ever\_married  
 geom\_text(aes(y = ever\_married, label = sprintf("%.1f%%", ever\_married)),   
 position = position\_stack(vjust = 1.025),   
 size = 4, color = "black") +  
 # Adding text labels for married\_u18  
 geom\_text(aes(y = married\_u18, label = sprintf("%.1f%%", married\_u18)),   
 position = position\_stack(vjust = 1.05),   
 size = 4, color = "black") +  
 # Adding text labels for married\_u16  
 geom\_text(aes(y = married\_u16, label = sprintf("%.1f%%", married\_u16)),   
 position = position\_stack(vjust = 1.1),   
 size = 4, color = "black") +  
 scale\_fill\_manual(values = c("Ever married" = "#FBE1E5",   
 "Married before 18 years old" = "#EF7D8D",   
 "Married before 16 years old" = "#B20016"),  
 name = "Marital Status") +  
 labs(x = "Region", y = "Percentage", title = "Marital Status Among Female Respondants by Region") +  
 theme\_minimal() +  
 theme(plot.margin = margin(t = 10, r = 10, b = 10, l = 10, unit = "mm"),   
 axis.text.x = element\_text(angle = 45, hjust = 1),   
 legend.position = "right") +  
 scale\_x\_discrete(labels = c("1" = "Red River Delta", "2" = "Northern Midlands And Mountain",   
 "3" = "North Central And Central Coastal", "4" = "Central Highlands",   
 "5" = "South East", "6" = "Mekong River Delta"))  
  
# Convert ggplot object to plotly for interactive visualization  
ggplotly(p)

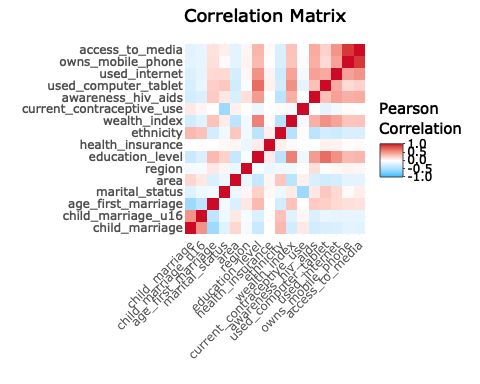


### Preparing for Regression Analysis

#### Correlation Matrix

# Calculate correlation matrix  
cor\_matrix <- cor(imputed\_df %>% select\_if(is.numeric), use = "complete.obs")  
  
# Melt the correlation matrix  
melted\_cor\_matrix <- melt(cor\_matrix)

# Generate an interactive heatmap  
corr\_matrix <- ggplot(melted\_cor\_matrix, aes(Var1, Var2, fill = value)) +  
 geom\_tile() +  
 scale\_fill\_gradientn(  
 colours = c("deepskyblue", "white", "#CD0A25"),  
 values = scales::rescale(c(-1, 0, 1)),  
 limits = c(-1, 1),  
 name="Pearson\nCorrelation"  
 ) +  
 theme\_minimal() +   
 theme(axis.text.x = element\_text(angle = 45, hjust = 1)) +  
 xlab("") +   
 ylab("") +  
 ggtitle("Correlation Matrix")   
  
# Convert ggplot object to plotly for interactivity  
ggplotly(corr\_matrix)



#### Converting Categorical and Binary Variables to Factors

# Convert nominal and ordinal variables to factors  
imputed\_df$area <- as.factor(imputed\_df$area)  
imputed\_df$region <- as.factor(imputed\_df$region)  
imputed\_df$education\_level <- factor(imputed\_df$education\_level, ordered = FALSE)  
imputed\_df$ethnicity <- as.factor(imputed\_df$ethnicity)  
imputed\_df$wealth\_index <- factor(imputed\_df$wealth\_index, ordered = FALSE)  
  
# Binary variables are already in the correct format and can be used as is

### Base Logistic Regression Model

# Baseline model for reference  
baseline\_model <- glm(child\_marriage ~ area + education\_level + wealth\_index + health\_insurance + current\_contraceptive\_use + awareness\_hiv\_aids + access\_to\_media, family = binomial(), data = imputed\_df)  
  
# Summarize the baseline model  
summary(baseline\_model)

##   
## Call:  
## glm(formula = child\_marriage ~ area + education\_level + wealth\_index +   
## health\_insurance + current\_contraceptive\_use + awareness\_hiv\_aids +   
## access\_to\_media, family = binomial(), data = imputed\_df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.73391 0.13169 -5.573 2.50e-08 \*\*\*  
## area2 0.45543 0.07999 5.693 1.25e-08 \*\*\*  
## education\_level1 -0.41350 0.08705 -4.750 2.03e-06 \*\*\*  
## education\_level2 -0.58693 0.08518 -6.890 5.57e-12 \*\*\*  
## education\_level3 -1.52807 0.11502 -13.286 < 2e-16 \*\*\*  
## education\_level4 -3.24812 0.51198 -6.344 2.24e-10 \*\*\*  
## education\_level5 -3.19482 0.25323 -12.616 < 2e-16 \*\*\*  
## wealth\_index2 -0.46183 0.07994 -5.777 7.61e-09 \*\*\*  
## wealth\_index3 -0.80552 0.09964 -8.085 6.24e-16 \*\*\*  
## wealth\_index4 -0.61389 0.10926 -5.618 1.93e-08 \*\*\*  
## wealth\_index5 -0.89341 0.14898 -5.997 2.01e-09 \*\*\*  
## health\_insurance 0.06907 0.08094 0.853 0.393   
## current\_contraceptive\_use 0.40123 0.06065 6.615 3.71e-11 \*\*\*  
## awareness\_hiv\_aids -0.47875 0.06960 -6.879 6.05e-12 \*\*\*  
## access\_to\_media -0.01461 0.08121 -0.180 0.857   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 10439.1 on 11293 degrees of freedom  
## Residual deviance: 8493.7 on 11279 degrees of freedom  
## AIC: 8523.7  
##   
## Number of Fisher Scoring iterations: 7

#### Accessing Base Model’s Multicollinearity

# Calculate VIF (A VIF value > 5 indicates high multicollinearity)  
base\_vif\_results <- vif(baseline\_model)  
print(base\_vif\_results)

## GVIF Df GVIF^(1/(2\*Df))  
## area 1.110732 1 1.053913  
## education\_level 1.707863 5 1.054983  
## wealth\_index 1.401192 4 1.043067  
## health\_insurance 1.021120 1 1.010505  
## current\_contraceptive\_use 1.034853 1 1.017277  
## awareness\_hiv\_aids 1.500511 1 1.224954  
## access\_to\_media 1.268704 1 1.126367

# Check if any VIF value is greater than a typical threshold, like 5 or 10.  
base\_high\_vif <- base\_vif\_results[base\_vif\_results > 5]  
print(base\_high\_vif)

## numeric(0)

### Adding Fixed Effects to Base Model

# Enhanced model with additional fixed effects  
enhanced\_model <- glm(child\_marriage ~ area + region + education\_level + ethnicity + wealth\_index + health\_insurance + current\_contraceptive\_use + awareness\_hiv\_aids + access\_to\_media,   
 family = binomial(),   
 data = imputed\_df)  
  
# Summarize the new model with FEs  
summary(enhanced\_model)

##   
## Call:  
## glm(formula = child\_marriage ~ area + region + education\_level +   
## ethnicity + wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media, family = binomial(),   
## data = imputed\_df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.54522 0.17906 -8.630 < 2e-16 \*\*\*  
## area2 0.28689 0.08488 3.380 0.000725 \*\*\*  
## region2 0.37758 0.12919 2.923 0.003469 \*\*   
## region3 0.18170 0.12745 1.426 0.153967   
## region4 0.34057 0.12599 2.703 0.006869 \*\*   
## region5 -0.05340 0.12612 -0.423 0.672028   
## region6 -0.05277 0.13336 -0.396 0.692301   
## education\_level1 -0.05925 0.09353 -0.634 0.526382   
## education\_level2 -0.22571 0.09227 -2.446 0.014438 \*   
## education\_level3 -1.17196 0.12172 -9.629 < 2e-16 \*\*\*  
## education\_level4 -2.96956 0.51385 -5.779 7.51e-09 \*\*\*  
## education\_level5 -2.89574 0.25599 -11.312 < 2e-16 \*\*\*  
## ethnicity2 0.07771 0.10592 0.734 0.463184   
## ethnicity3 0.27807 0.12830 2.167 0.030210 \*   
## ethnicity4 1.15504 0.10626 10.870 < 2e-16 \*\*\*  
## wealth\_index2 -0.12498 0.08523 -1.466 0.142533   
## wealth\_index3 -0.46330 0.10565 -4.385 1.16e-05 \*\*\*  
## wealth\_index4 -0.26442 0.11695 -2.261 0.023760 \*   
## wealth\_index5 -0.54064 0.16030 -3.373 0.000744 \*\*\*  
## health\_insurance -0.05157 0.08277 -0.623 0.533244   
## current\_contraceptive\_use 0.48480 0.06265 7.739 1.00e-14 \*\*\*  
## awareness\_hiv\_aids -0.19387 0.07592 -2.554 0.010661 \*   
## access\_to\_media -0.07880 0.08517 -0.925 0.354845   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 10439.1 on 11293 degrees of freedom  
## Residual deviance: 8240.2 on 11271 degrees of freedom  
## AIC: 8286.2  
##   
## Number of Fisher Scoring iterations: 7

#### Model with FEs’ Multicollinearity

# Calculate VIF (A VIF value > 5 indicates high multicollinearity)  
FE\_vif\_results <- vif(enhanced\_model)  
print(FE\_vif\_results)

## GVIF Df GVIF^(1/(2\*Df))  
## area 1.254248 1 1.119932  
## region 4.916245 5 1.172636  
## education\_level 1.956030 5 1.069394  
## ethnicity 4.235677 3 1.272000  
## wealth\_index 1.934925 4 1.086008  
## health\_insurance 1.051524 1 1.025438  
## current\_contraceptive\_use 1.049061 1 1.024237  
## awareness\_hiv\_aids 1.679129 1 1.295812  
## access\_to\_media 1.312005 1 1.145428

# Check if any VIF value is greater than a typical threshold, like 5.  
FE\_high\_vif <- FE\_vif\_results[FE\_vif\_results > 5]  
print(FE\_high\_vif)

## numeric(0)

No VIF is significant larger than 5. This suggests that the independent variables in my model with Fixed Effects do not suffer from severe multicollinearity.

#### Assessing AICs and BICs of Base Model vs Fixed Effects Model

# Model comparison using AIC  
aic\_base <- AIC(baseline\_model)  
aic\_enhanced <- AIC(enhanced\_model)  
cat("AIC - Base Model:", aic\_base, "\n")

## AIC - Base Model: 8523.675

cat("AIC - Enhanced Model:", aic\_enhanced, "\n")

## AIC - Enhanced Model: 8286.168

Additionally, the lower AIC for the model with two fixed effects compared to the baseline model suggests that adding these fixed effects improves the model’s overall fit. This aligns with my theoretical justification for including region and ethnicity as relevant factors in predicting child marriage.

# Model comparison using BIC  
bic\_base <- BIC(baseline\_model)  
bic\_enhanced <- BIC(enhanced\_model)  
cat("BIC - Base Model:", bic\_base, "\n")

## BIC - Base Model: 8633.655

cat("BIC - Enhanced Model:", bic\_enhanced, "\n")

## BIC - Enhanced Model: 8454.805

Lower BIC values suggest that, despite the added complexity (more parameters), the model with both region and ethnicity provides a better overall fit to my data when adjusted for the number of predictors. This is a strong indication that these variables are meaningful in explaining the variance in child marriage occurrences in my dataset.

### Model with Interaction Terms (Area:Wealth\_Index)

# Adding the interaction term between education level and wealth index  
model\_interaction <- update(enhanced\_model, . ~ . + area:wealth\_index)  
summary(model\_interaction)

##   
## Call:  
## glm(formula = child\_marriage ~ area + region + education\_level +   
## ethnicity + wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media + area:wealth\_index,   
## family = binomial(), data = imputed\_df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.78525 0.21262 -8.396 < 2e-16 \*\*\*  
## area2 0.54558 0.15257 3.576 0.000349 \*\*\*  
## region2 0.36523 0.13009 2.807 0.004994 \*\*   
## region3 0.16778 0.12857 1.305 0.191877   
## region4 0.32337 0.12701 2.546 0.010892 \*   
## region5 -0.06803 0.12726 -0.535 0.592956   
## region6 -0.06059 0.13410 -0.452 0.651371   
## education\_level1 -0.06117 0.09352 -0.654 0.513085   
## education\_level2 -0.21895 0.09231 -2.372 0.017701 \*   
## education\_level3 -1.17282 0.12176 -9.632 < 2e-16 \*\*\*  
## education\_level4 -2.96880 0.51416 -5.774 7.74e-09 \*\*\*  
## education\_level5 -2.89550 0.25820 -11.214 < 2e-16 \*\*\*  
## ethnicity2 0.06171 0.10624 0.581 0.561352   
## ethnicity3 0.27262 0.12824 2.126 0.033518 \*   
## ethnicity4 1.12852 0.10687 10.559 < 2e-16 \*\*\*  
## wealth\_index2 0.26715 0.19445 1.374 0.169473   
## wealth\_index3 -0.15465 0.21122 -0.732 0.464075   
## wealth\_index4 -0.04152 0.22195 -0.187 0.851622   
## wealth\_index5 -0.32837 0.26777 -1.226 0.220078   
## health\_insurance -0.04238 0.08289 -0.511 0.609164   
## current\_contraceptive\_use 0.49215 0.06274 7.844 4.35e-15 \*\*\*  
## awareness\_hiv\_aids -0.18696 0.07599 -2.460 0.013883 \*   
## access\_to\_media -0.07199 0.08522 -0.845 0.398196   
## area2:wealth\_index2 -0.47937 0.21523 -2.227 0.025928 \*   
## area2:wealth\_index3 -0.38304 0.24189 -1.584 0.113299   
## area2:wealth\_index4 -0.26224 0.25631 -1.023 0.306247   
## area2:wealth\_index5 -0.25049 0.31993 -0.783 0.433646   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 10439.1 on 11293 degrees of freedom  
## Residual deviance: 8234.7 on 11267 degrees of freedom  
## AIC: 8288.7  
##   
## Number of Fisher Scoring iterations: 7

#### Assessing AICs and BICs of Fixed Effects Model vs Fixed Effects Model w/ Interaction Terms

cat("AIC (enhanced):", AIC(enhanced\_model), "\nAIC (w/ interaction terms):", AIC(model\_interaction), "\n")

## AIC (enhanced): 8286.168   
## AIC (w/ interaction terms): 8288.662

cat("BIC (enhanced):", BIC(enhanced\_model), "\nBIC (w/ interaction terms):", BIC(model\_interaction), "\n")

## BIC (enhanced): 8454.805   
## BIC (w/ interaction terms): 8486.627

* The AIC and BIC results for my two models – the enhanced\_model and model\_interaction – indicate that the simpler model (without interaction terms) has a better trade-off between model fit and complexity.
* The AIC is lower for the enhanced\_model compared to the model\_interaction. This suggests that, despite possibly having a good fit, the addition of interaction terms to the model\_interaction does not improve the model’s performance enough to justify the added complexity. A lower AIC value is generally preferred, implying that the enhanced\_model is a better model choice from an information criterion standpoint.
* The BIC, like the AIC, is also lower for the enhanced\_model. BIC, which includes a stricter penalty for the number of parameters, reinforces the conclusion drawn from the AIC: the additional complexity introduced by interaction terms does not lead to a proportional improvement in the model fit. BIC being more stringent, particularly in models with larger sample sizes and more parameters, suggests a strong preference for the simpler enhanced\_model.
* Both AIC and BIC results point towards the enhanced\_model as a more optimal balance between goodness of fit and parsimony. This indicates that, while interaction terms add detail and complexity to the model, they might not be contributing significantly to explaining the variability in your dependent variable.

### Logistic Regression Models Comparison (Odd Ratios and 95% Confidence Intervals)

# Function to add significance asterisks  
add\_asterisks <- function(p\_value) {  
 if (is.na(p\_value)) {  
 return(NA)  
 } else if (p\_value < 0.001) {  
 return("\*\*\*")  
 } else if (p\_value < 0.01) {  
 return("\*\*")  
 } else if (p\_value < 0.05) {  
 return("\*")  
 } else {  
 return("")  
 }  
}

# Function to format confidence intervals as a string  
format\_ci <- function(lower, upper) {  
 paste0("(", round(lower, 2), ", ", round(upper, 2), ")")  
}

# Tidy the baseline model with confidence intervals  
tidy\_baseline <- tidy(baseline\_model, conf.int = TRUE, exponentiate = TRUE)  
  
# Tidy the enhanced model with confidence intervals  
tidy\_enhanced <- tidy(enhanced\_model, conf.int = TRUE, exponentiate = TRUE)  
  
# Tidy the interaction model with confidence intervals  
tidy\_interaction <- tidy(model\_interaction, conf.int = TRUE, exponentiate = TRUE)

# Apply the function to each model's p.value  
tidy\_baseline$asterisks <- sapply(tidy\_baseline$p.value, add\_asterisks)  
tidy\_enhanced$asterisks <- sapply(tidy\_enhanced$p.value, add\_asterisks)  
tidy\_interaction$asterisks <- sapply(tidy\_interaction$p.value, add\_asterisks)

# Create OR strings with asterisks and format CIs as a string  
tidy\_baseline <- tidy\_baseline %>%  
 mutate(  
 OR = ifelse(is.na(estimate), NA, paste0(round(estimate, 2), asterisks)),  
 CI = ifelse(is.na(conf.low) | is.na(conf.high), NA, format\_ci(conf.low, conf.high))  
 )  
  
tidy\_enhanced <- tidy\_enhanced %>%  
 mutate(  
 OR = ifelse(is.na(estimate), NA, paste0(round(estimate, 2), asterisks)),  
 CI = ifelse(is.na(conf.low) | is.na(conf.high), NA, format\_ci(conf.low, conf.high))  
 )  
  
tidy\_interaction <- tidy\_interaction %>%  
 mutate(  
 OR = ifelse(is.na(estimate), NA, paste0(round(estimate, 2), asterisks)),  
 CI = ifelse(is.na(conf.low) | is.na(conf.high), NA, format\_ci(conf.low, conf.high))  
 )

# Add a 'Model' column to each tidied dataframe  
tidy\_baseline <- tidy\_baseline %>% mutate(Model = "Baseline")  
tidy\_enhanced <- tidy\_enhanced %>% mutate(Model = "Enhanced")  
tidy\_interaction <- tidy\_interaction %>% mutate(Model = "Interaction")  
  
# Combine and pivot the dataframes  
combined\_results <- bind\_rows(  
 tidy\_baseline %>% select(term, OR, CI, Model),  
 tidy\_enhanced %>% select(term, OR, CI, Model),  
 tidy\_interaction %>% select(term, OR, CI, Model)  
) %>%  
 pivot\_wider(names\_from = Model, values\_from = c(OR, CI))

# Replace NAs with "—"  
combined\_results[is.na(combined\_results)] <- "—"

# Reordering columns to have OR and CI next to each other for each model  
combined\_results <- combined\_results %>%  
 select(term,   
 OR\_Baseline, CI\_Baseline,   
 OR\_Enhanced, CI\_Enhanced,   
 OR\_Interaction, CI\_Interaction)

# Print the final combined table  
print(combined\_results)

## # A tibble: 27 × 7  
## term OR\_Baseline CI\_Baseline OR\_Enhanced CI\_Enhanced OR\_Interaction  
## <chr> <chr> <chr> <chr> <chr> <chr>   
## 1 (Intercept) 0.48\*\*\* (0.37, 0.6… 0.21\*\*\* (0.15, 0.3) 0.17\*\*\*   
## 2 area2 1.58\*\*\* (1.35, 1.8… 1.33\*\*\* (1.13, 1.5… 1.73\*\*\*   
## 3 education\_lev… 0.66\*\*\* (0.56, 0.7… 0.94 (0.78, 1.1… 0.94   
## 4 education\_lev… 0.56\*\*\* (0.47, 0.6… 0.8\* (0.67, 0.9… 0.8\*   
## 5 education\_lev… 0.22\*\*\* (0.17, 0.2… 0.31\*\*\* (0.24, 0.3… 0.31\*\*\*   
## 6 education\_lev… 0.04\*\*\* (0.01, 0.0… 0.05\*\*\* (0.02, 0.1… 0.05\*\*\*   
## 7 education\_lev… 0.04\*\*\* (0.02, 0.0… 0.06\*\*\* (0.03, 0.0… 0.06\*\*\*   
## 8 wealth\_index2 0.63\*\*\* (0.54, 0.7… 0.88 (0.75, 1.0… 1.31   
## 9 wealth\_index3 0.45\*\*\* (0.37, 0.5… 0.63\*\*\* (0.51, 0.7… 0.86   
## 10 wealth\_index4 0.54\*\*\* (0.44, 0.6… 0.77\* (0.61, 0.9… 0.96   
## # ℹ 17 more rows  
## # ℹ 1 more variable: CI\_Interaction <chr>

### Models Validation

#### Hosmer-Lemeshow Test

# 1. Hosmer-Lemeshow Test for the Baseline Model  
hoslem.test(baseline\_model$y, fitted(baseline\_model), g=10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: baseline\_model$y, fitted(baseline\_model)  
## X-squared = 10.565, df = 8, p-value = 0.2276

# 2. Hosmer-Lemeshow Test for the Enhanced Model (Baseline + Fixed Effects)  
hoslem.test(enhanced\_model$y, fitted(enhanced\_model), g=10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: enhanced\_model$y, fitted(enhanced\_model)  
## X-squared = 10.538, df = 8, p-value = 0.2293

# 3. Hosmer-Lemeshow Test for the Interaction Model (Baseline + Fixed Effects + Interaction Terms)  
hoslem.test(model\_interaction$y, fitted(model\_interaction), g=10)

##   
## Hosmer and Lemeshow goodness of fit (GOF) test  
##   
## data: model\_interaction$y, fitted(model\_interaction)  
## X-squared = 10.272, df = 8, p-value = 0.2465

A large p-value (>0.05) indicates a good fit, meaning that there’s no significant difference between the observed and predicted values. Through each model, the p-value increases which suggests that our decision to include fixed effects and interaction terms are significant.

#### Likelihood Ratio Test (ANOVA)

# Baseline vs. Baseline + Fixed Effects  
anova(baseline\_model, enhanced\_model, test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: child\_marriage ~ area + education\_level + wealth\_index + health\_insurance +   
## current\_contraceptive\_use + awareness\_hiv\_aids + access\_to\_media  
## Model 2: child\_marriage ~ area + region + education\_level + ethnicity +   
## wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)   
## 1 11279 8493.7   
## 2 11271 8240.2 8 253.51 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

The Enhanced Model provides a significantly better fit to the data compared to the Baseline Model, as indicated by the large decrease in residual deviance and the very small p-value (< 2.2e-16).

# Baseline + Fixed Effects vs. Baseline + Fixed Effects + Interaction Terms  
anova(enhanced\_model, model\_interaction, test="Chisq")

## Analysis of Deviance Table  
##   
## Model 1: child\_marriage ~ area + region + education\_level + ethnicity +   
## wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media  
## Model 2: child\_marriage ~ area + region + education\_level + ethnicity +   
## wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media + area:wealth\_index  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 11271 8240.2   
## 2 11267 8234.7 4 5.506 0.2392

Adding the interaction terms between area and wealth\_index does not significantly improve the model fit compared to the Enhanced Model without interaction terms. This is indicated by the relatively high p-value and the smaller decrease in residual deviance.

#### ROC Curve and AUC

# For each model  
roc\_response\_baseline <- roc(imputed\_df$child\_marriage, fitted(baseline\_model))  
auc\_baseline <- auc(roc\_response\_baseline)  
  
roc\_response\_enhanced <- roc(imputed\_df$child\_marriage, fitted(enhanced\_model))  
auc\_enhanced <- auc(roc\_response\_enhanced)  
  
roc\_response\_interaction <- roc(imputed\_df$child\_marriage, fitted(model\_interaction))  
auc\_interaction <- auc(roc\_response\_interaction)  
  
# Compare AUC values  
print(paste("AUC Baseline Model:", auc\_baseline))

## [1] "AUC Baseline Model: 0.794541604239919"

print(paste("AUC Enhanced Model:", auc\_enhanced))

## [1] "AUC Enhanced Model: 0.809330968003957"

print(paste("AUC Interaction Model:", auc\_interaction))

## [1] "AUC Interaction Model: 0.80987822317723"

1. Baseline Model (AUC: 0.7945):

The AUC value is close to 0.8, which indicates that the Baseline Model has good discriminative ability. In other words, it is capable of distinguishing between cases and controls with a high degree of accuracy. An AUC of 0.5 represents a model with no discriminative ability (akin to random guessing), while an AUC of 1.0 represents perfect discrimination. So, my model is performing substantially better than random guessing.

1. Enhanced Model (AUC: 0.8093):

This model shows a slight improvement in AUC over the Baseline Model. The increase suggests that the additional variables (or adjustments) you made in the Enhanced Model contribute positively to its ability to differentiate between cases and controls. The difference in AUC between the Baseline and Enhanced models, while modest, is still meaningful, especially in practical, real-world contexts.

1. Interaction Model (AUC: 0.8099):

The Interaction Model shows a very slight improvement in AUC over the Enhanced Model. This indicates that adding interaction terms provides a marginal improvement in the model’s discriminatory power. However, the improvement is very minimal, which aligns with my earlier findings that the interaction terms did not significantly improve the model fit.

1. Overall:

All models demonstrate good ability to distinguish between cases and controls. An AUC greater than 0.7 is generally considered acceptable, and my models are around or above 0.8. The Enhanced and Interaction Models only show marginal improvements in AUC compared to the Baseline Model. This suggests that while the additional complexity (more variables and interaction terms) does contribute slightly to model performance, the gains are not substantial. Given the slight increases in AUC with added model complexity, consider the trade-offs. A simpler model might be preferable if it is easier to interpret and communicate, especially if the increase in predictive power is minimal. Personally, given this result, I think Enhanced Model might be a better-suited model overall.

### Gut Check: Redo Analysis with Actual Data (no imputation)

#### Recreate the dataframe with removed missing data

# No imputation -- Remove rows with any missing data from 'female\_df'  
og\_df <- na.omit(female\_df)  
head(og\_df)

## # A tibble: 6 × 14  
## age\_first\_marriage marital\_status area region education\_level  
## <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 17 1 1 1 2  
## 2 22 1 1 1 5  
## 3 26 1 1 1 5  
## 4 21 1 1 1 3  
## 5 18 1 1 1 1  
## 6 22 1 1 1 4  
## # ℹ 9 more variables: health\_insurance <dbl>, ethnicity <dbl>,  
## # wealth\_index <dbl>, current\_contraceptive\_use <dbl>,  
## # awareness\_hiv\_aids <dbl>, used\_computer\_tablet <dbl>, used\_internet <dbl>,  
## # owns\_mobile\_phone <dbl>, access\_to\_media <dbl>

#### Recreate necessary variables

# Convert "age at first marriage" into a binary variable to indicate child marriage  
# Child marriage is defined as marriage before the age of 18  
# The new binary variable "child\_marriage" will have a value of 1 if the marriage occurred before age 18, and 0 otherwise  
og\_df <- og\_df %>%  
 mutate(child\_marriage = ifelse(age\_first\_marriage < 18, 1, 0))

# Create a binary variable for child marriage under 16  
# The new variable "child\_marriage\_u16" will have a value of 1 if the marriage occurred before age 16, and 0 otherwise  
og\_df <- og\_df %>%  
 mutate(child\_marriage\_u16 = ifelse(age\_first\_marriage < 16, 1, 0))

# Move "child\_marriage" and "child\_marriage\_u16" to the front of the dataframe  
og\_df <- og\_df %>%  
 select(child\_marriage, child\_marriage\_u16, everything())

#### Converting Categorical and Binary Variables to Factors

# Convert nominal and ordinal variables to factors  
og\_df$area <- as.factor(og\_df$area)  
og\_df$region <- as.factor(og\_df$region)  
og\_df$education\_level <- factor(og\_df$education\_level, ordered = FALSE)  
og\_df$ethnicity <- as.factor(og\_df$ethnicity)  
og\_df$wealth\_index <- factor(og\_df$wealth\_index, ordered = FALSE)  
  
# Binary variables are already in the correct format and can be used as is

#### Recreate the regression models without imputed data

# Base model  
base\_model\_no\_impute <- glm(child\_marriage ~ area + education\_level + wealth\_index + health\_insurance + current\_contraceptive\_use + awareness\_hiv\_aids + access\_to\_media, family = binomial(), data = og\_df)  
  
# Summarize the base model  
summary(base\_model\_no\_impute)

##   
## Call:  
## glm(formula = child\_marriage ~ area + education\_level + wealth\_index +   
## health\_insurance + current\_contraceptive\_use + awareness\_hiv\_aids +   
## access\_to\_media, family = binomial(), data = og\_df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.12852 0.15839 -0.811 0.41714   
## area2 0.27195 0.09515 2.858 0.00426 \*\*   
## education\_level1 -0.44184 0.10273 -4.301 1.70e-05 \*\*\*  
## education\_level2 -0.51162 0.10336 -4.950 7.43e-07 \*\*\*  
## education\_level3 -1.27268 0.13835 -9.199 < 2e-16 \*\*\*  
## education\_level4 -3.39394 0.71961 -4.716 2.40e-06 \*\*\*  
## education\_level5 -3.17824 0.46414 -6.848 7.51e-12 \*\*\*  
## wealth\_index2 -0.61389 0.09307 -6.596 4.22e-11 \*\*\*  
## wealth\_index3 -0.92891 0.11199 -8.295 < 2e-16 \*\*\*  
## wealth\_index4 -0.75155 0.12357 -6.082 1.19e-09 \*\*\*  
## wealth\_index5 -1.02338 0.16941 -6.041 1.53e-09 \*\*\*  
## health\_insurance 0.10983 0.09266 1.185 0.23587   
## current\_contraceptive\_use -0.10949 0.07039 -1.555 0.11984   
## awareness\_hiv\_aids -0.46608 0.08253 -5.647 1.63e-08 \*\*\*  
## access\_to\_media 0.01062 0.10367 0.102 0.91843   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 6819.3 on 6443 degrees of freedom  
## Residual deviance: 5889.2 on 6429 degrees of freedom  
## AIC: 5919.2  
##   
## Number of Fisher Scoring iterations: 7

# Enhanced model with additional fixed effects  
enhanced\_model\_no\_impute <- glm(child\_marriage ~ area + region + education\_level + ethnicity + wealth\_index + health\_insurance + current\_contraceptive\_use + awareness\_hiv\_aids + access\_to\_media, family = binomial(), data = og\_df)  
  
# Summarize the new model with FEs  
summary(enhanced\_model\_no\_impute)

##   
## Call:  
## glm(formula = child\_marriage ~ area + region + education\_level +   
## ethnicity + wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media, family = binomial(),   
## data = og\_df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.70949 0.23279 -7.343 2.08e-13 \*\*\*  
## area2 0.17214 0.10044 1.714 0.08654 .   
## region2 0.10548 0.16065 0.657 0.51143   
## region3 0.03594 0.15601 0.230 0.81778   
## region4 -0.10432 0.17061 -0.611 0.54090   
## region5 0.06092 0.14434 0.422 0.67298   
## region6 0.10315 0.14996 0.688 0.49156   
## education\_level1 0.12519 0.11676 1.072 0.28362   
## education\_level2 0.07838 0.11878 0.660 0.50933   
## education\_level3 -0.71860 0.15086 -4.763 1.91e-06 \*\*\*  
## education\_level4 -2.82957 0.72295 -3.914 9.08e-05 \*\*\*  
## education\_level5 -2.61912 0.46855 -5.590 2.27e-08 \*\*\*  
## ethnicity2 0.55070 0.13875 3.969 7.21e-05 \*\*\*  
## ethnicity3 0.33886 0.14008 2.419 0.01556 \*   
## ethnicity4 1.90113 0.16173 11.755 < 2e-16 \*\*\*  
## wealth\_index2 -0.05571 0.10820 -0.515 0.60665   
## wealth\_index3 -0.34294 0.12995 -2.639 0.00831 \*\*   
## wealth\_index4 -0.14989 0.14419 -1.040 0.29857   
## wealth\_index5 -0.41552 0.19232 -2.161 0.03073 \*   
## health\_insurance -0.07933 0.09623 -0.824 0.40971   
## current\_contraceptive\_use -0.02792 0.07341 -0.380 0.70370   
## awareness\_hiv\_aids -0.08636 0.09460 -0.913 0.36132   
## access\_to\_media 0.17785 0.11026 1.613 0.10674   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 6819.3 on 6443 degrees of freedom  
## Residual deviance: 5658.7 on 6421 degrees of freedom  
## AIC: 5704.7  
##   
## Number of Fisher Scoring iterations: 7

#### Assessing AICs and BICs of Base Model vs Fixed Effects Model

# Model comparison using AIC  
aic\_base <- AIC(base\_model\_no\_impute)  
aic\_enhanced <- AIC(enhanced\_model\_no\_impute)  
cat("AIC - Base Model:", aic\_base, "\n")

## AIC - Base Model: 5919.176

cat("AIC - Enhanced Model:", aic\_enhanced, "\n")

## AIC - Enhanced Model: 5704.682

Additionally, the lower AIC for the model with two fixed effects compared to the baseline model suggests that adding these fixed effects improves the model’s overall fit. This aligns with my theoretical justification for including region and ethnicity as relevant factors in predicting child marriage.

# Model comparison using BIC  
bic\_base <- BIC(base\_model\_no\_impute)  
bic\_enhanced <- BIC(enhanced\_model\_no\_impute)  
cat("BIC - Base Model:", bic\_base, "\n")

## BIC - Base Model: 6020.739

cat("BIC - Enhanced Model:", bic\_enhanced, "\n")

## BIC - Enhanced Model: 5860.413

Lower BIC values suggest that, despite the added complexity (more parameters), the model with both region and ethnicity provides a better overall fit to my data when adjusted for the number of predictors. This is a strong indication that these variables are meaningful in explaining the variance in child marriage occurrences in my dataset.

### Model with Interaction Terms (Area:Wealth\_Index)

# Adding the interaction term between education level and wealth index  
interaction\_model\_no\_impute <- update(enhanced\_model\_no\_impute, . ~ . + area:wealth\_index)  
summary(interaction\_model\_no\_impute)

##   
## Call:  
## glm(formula = child\_marriage ~ area + region + education\_level +   
## ethnicity + wealth\_index + health\_insurance + current\_contraceptive\_use +   
## awareness\_hiv\_aids + access\_to\_media + area:wealth\_index,   
## family = binomial(), data = og\_df)  
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.71245 0.30284 -5.655 1.56e-08 \*\*\*  
## area2 0.15095 0.24236 0.623 0.53341   
## region2 0.12616 0.16152 0.781 0.43477   
## region3 0.05816 0.15716 0.370 0.71134   
## region4 -0.08816 0.17115 -0.515 0.60647   
## region5 0.08191 0.14672 0.558 0.57664   
## region6 0.12494 0.15100 0.827 0.40799   
## education\_level1 0.12367 0.11674 1.059 0.28943   
## education\_level2 0.07710 0.11879 0.649 0.51630   
## education\_level3 -0.71184 0.15091 -4.717 2.39e-06 \*\*\*  
## education\_level4 -2.79700 0.72341 -3.866 0.00011 \*\*\*  
## education\_level5 -2.58385 0.46919 -5.507 3.65e-08 \*\*\*  
## ethnicity2 0.55095 0.13897 3.965 7.35e-05 \*\*\*  
## ethnicity3 0.33370 0.14008 2.382 0.01721 \*   
## ethnicity4 1.90109 0.16307 11.658 < 2e-16 \*\*\*  
## wealth\_index2 0.02709 0.27554 0.098 0.92168   
## wealth\_index3 -0.35431 0.28892 -1.226 0.22007   
## wealth\_index4 -0.24012 0.30088 -0.798 0.42484   
## wealth\_index5 -0.66376 0.37279 -1.781 0.07499 .   
## health\_insurance -0.07528 0.09636 -0.781 0.43465   
## current\_contraceptive\_use -0.02755 0.07345 -0.375 0.70764   
## awareness\_hiv\_aids -0.08558 0.09457 -0.905 0.36551   
## access\_to\_media 0.17709 0.11027 1.606 0.10827   
## area2:wealth\_index2 -0.11570 0.29382 -0.394 0.69374   
## area2:wealth\_index3 0.01145 0.31297 0.037 0.97081   
## area2:wealth\_index4 0.13487 0.32848 0.411 0.68137   
## area2:wealth\_index5 0.37698 0.41566 0.907 0.36443   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 6819.3 on 6443 degrees of freedom  
## Residual deviance: 5656.6 on 6417 degrees of freedom  
## AIC: 5710.6  
##   
## Number of Fisher Scoring iterations: 7

cat("AIC (enhanced):", AIC(enhanced\_model\_no\_impute), "\nAIC (w/ interaction terms):", AIC(interaction\_model\_no\_impute), "\n")

## AIC (enhanced): 5704.682   
## AIC (w/ interaction terms): 5710.578

cat("BIC (enhanced):", BIC(enhanced\_model\_no\_impute), "\nBIC (w/ interaction terms):", BIC(interaction\_model\_no\_impute), "\n")

## BIC (enhanced): 5860.413   
## BIC (w/ interaction terms): 5893.392

* The AIC and BIC results for my two models – the enhanced\_model\_no\_impute and interaction\_model\_no\_impute – indicate that the simpler model (without interaction terms) has a better trade-off between model fit and complexity.
* The AIC is lower for the enhanced\_model\_no\_impute compared to the interaction\_model\_no\_impute. This suggests that, despite possibly having a good fit, the addition of interaction terms to the interaction\_model\_no\_impute does not improve the model’s performance enough to justify the added complexity. A lower AIC value is generally preferred, implying that the enhanced\_model\_no\_impute is a better model choice from an information criterion standpoint.
* The BIC, like the AIC, is also lower for the enhanced\_model\_no\_impute. BIC, which includes a stricter penalty for the number of parameters, reinforces the conclusion drawn from the AIC: the additional complexity introduced by interaction terms does not lead to a proportional improvement in the model fit. BIC being more stringent, particularly in models with larger sample sizes and more parameters, suggests a strong preference for the simpler enhanced\_model\_no\_impute.
* Both AIC and BIC results point towards the enhanced\_model\_no\_impute as a more optimal balance between goodness of fit and parsimony. This indicates that, while interaction terms add detail and complexity to the model, they might not be contributing significantly to explaining the variability in your dependent variable.

# VIFs check (A VIF value > 5 indicates high multicollinearity)  
# Base model  
base\_no\_impute\_vif <- vif(base\_model\_no\_impute)  
print(base\_no\_impute\_vif)

## GVIF Df GVIF^(1/(2\*Df))  
## area 1.161113 1 1.077549  
## education\_level 1.635562 5 1.050429  
## wealth\_index 1.521114 4 1.053829  
## health\_insurance 1.028226 1 1.014015  
## current\_contraceptive\_use 1.009489 1 1.004733  
## awareness\_hiv\_aids 1.471075 1 1.212879  
## access\_to\_media 1.212302 1 1.101046

# Enhanced model  
enhanced\_no\_impute\_vif <- vif(enhanced\_model\_no\_impute)  
print(enhanced\_no\_impute\_vif)

## GVIF Df GVIF^(1/(2\*Df))  
## area 1.295560 1 1.138227  
## region 5.816954 5 1.192531  
## education\_level 2.116453 5 1.077856  
## ethnicity 7.824091 3 1.408983  
## wealth\_index 2.687468 4 1.131535  
## health\_insurance 1.087063 1 1.042623  
## current\_contraceptive\_use 1.025473 1 1.012656  
## awareness\_hiv\_aids 1.772812 1 1.331470  
## access\_to\_media 1.256323 1 1.120858

# Interacton model  
interaction\_no\_impute\_vif <- vif(interaction\_model\_no\_impute)

## there are higher-order terms (interactions) in this model  
## consider setting type = 'predictor'; see ?vif

print(interaction\_no\_impute\_vif)

## GVIF Df GVIF^(1/(2\*Df))  
## area 7.510342 1 2.740500  
## region 6.056756 5 1.197358  
## education\_level 2.125445 5 1.078313  
## ethnicity 7.978570 3 1.413581  
## wealth\_index 497.956732 4 2.173446  
## health\_insurance 1.090010 1 1.044035  
## current\_contraceptive\_use 1.026554 1 1.013190  
## awareness\_hiv\_aids 1.771947 1 1.331145  
## access\_to\_media 1.256367 1 1.120878  
## area:wealth\_index 466.507434 4 2.155794

* Most VIF values are just slightly above 1, which indicates no concerning multicollinearity in this model.
* Except for wealth\_index and area:wealth\_index, their VIF values are significantly higher (around 2.17 and 2.16, respectively). These values suggest a considerable level of multicollinearity, which is often expected when including interaction terms in a model.
* High multicollinearity in models, especially those with interaction terms, doesn’t invalidate the model but does complicate the interpretation of specific coefficients.

#### Comparing Models With vs Without Imputed Data

# Extracting data from the models  
extract\_model\_data <- function(model) {  
 model\_summary <- summary(model)  
 coeffs <- model\_summary$coefficients  
 data.frame(  
 Term = rownames(coeffs),  
 Estimate = sprintf("%.3f", coeffs[, "Estimate"]),  
 pValue = ifelse(coeffs[, "Pr(>|z|)"] < 0.001,   
 format(coeffs[, "Pr(>|z|)"], scientific = TRUE),  
 sprintf("%.3f", coeffs[, "Pr(>|z|)"])),  
 Significance = sapply(coeffs[, "Pr(>|z|)"], add\_asterisks)  
 )  
}

# Applying the function to each model  
base\_impute\_data <- extract\_model\_data(baseline\_model)  
base\_no\_impute\_data <- extract\_model\_data(base\_model\_no\_impute)  
  
# Combine the data for comparison  
base\_comparison\_data <- merge(base\_impute\_data, base\_no\_impute\_data, by = "Term", suffixes = c("\_Impute", "\_NoImpute"), sort = FALSE)  
  
# View the comparison  
print(base\_comparison\_data)

## Term Estimate\_Impute pValue\_Impute Significance\_Impute  
## 1 (Intercept) -0.734 2.502722e-08 \*\*\*  
## 2 area2 0.455 1.245520e-08 \*\*\*  
## 3 education\_level1 -0.413 2.032253e-06 \*\*\*  
## 4 education\_level2 -0.587 5.570356e-12 \*\*\*  
## 5 education\_level3 -1.528 2.806068e-40 \*\*\*  
## 6 education\_level4 -3.248 2.235120e-10 \*\*\*  
## 7 education\_level5 -3.195 1.720431e-36 \*\*\*  
## 8 wealth\_index2 -0.462 7.607170e-09 \*\*\*  
## 9 wealth\_index3 -0.806 6.235617e-16 \*\*\*  
## 10 wealth\_index4 -0.614 1.926274e-08 \*\*\*  
## 11 wealth\_index5 -0.893 2.012553e-09 \*\*\*  
## 12 health\_insurance 0.069 0.393   
## 13 current\_contraceptive\_use 0.401 3.710837e-11 \*\*\*  
## 14 awareness\_hiv\_aids -0.479 6.047566e-12 \*\*\*  
## 15 access\_to\_media -0.015 0.857   
## Estimate\_NoImpute pValue\_NoImpute Significance\_NoImpute  
## 1 -0.129 0.417   
## 2 0.272 0.004 \*\*  
## 3 -0.442 1.701057e-05 \*\*\*  
## 4 -0.512 7.430972e-07 \*\*\*  
## 5 -1.273 3.606682e-20 \*\*\*  
## 6 -3.394 2.400883e-06 \*\*\*  
## 7 -3.178 7.512748e-12 \*\*\*  
## 8 -0.614 4.224206e-11 \*\*\*  
## 9 -0.929 1.089044e-16 \*\*\*  
## 10 -0.752 1.186365e-09 \*\*\*  
## 11 -1.023 1.532441e-09 \*\*\*  
## 12 0.110 0.236   
## 13 -0.109 0.120   
## 14 -0.466 1.630404e-08 \*\*\*  
## 15 0.011 0.918

* Education and Wealth Index variables are consistent, robust and significant predictors of the outcome across both models.
* For some variables like ‘current contraceptive use’, imputation significantly impacts its role in the model. It could indicate the missing data mechanism’s influence.
* Significance of Area: This variable’s significance suggests geographical variation is an important factor, but its influence is reduced in the non-imputed model.
* The overall patterns suggest that while some predictors are consistently influential, others are sensitive to how missing data is handled.

# Applying the function to each model  
enhanced\_impute\_data <- extract\_model\_data(enhanced\_model)  
enhanced\_no\_impute\_data <- extract\_model\_data(enhanced\_model\_no\_impute)  
  
# Combine the data for comparison  
enhanced\_comparison\_data <- merge(enhanced\_impute\_data, enhanced\_no\_impute\_data, by = "Term", suffixes = c("\_Impute", "\_NoImpute"), sort = FALSE)  
  
# View the comparison  
print(enhanced\_comparison\_data)

## Term Estimate\_Impute pValue\_Impute Significance\_Impute  
## 1 (Intercept) -1.545 6.147348e-18 \*\*\*  
## 2 area2 0.287 7.251287e-04 \*\*\*  
## 3 region2 0.378 0.003 \*\*  
## 4 region3 0.182 0.154   
## 5 region4 0.341 0.007 \*\*  
## 6 region5 -0.053 0.672   
## 7 region6 -0.053 0.692   
## 8 education\_level1 -0.059 0.526   
## 9 education\_level2 -0.226 0.014 \*  
## 10 education\_level3 -1.172 6.057140e-22 \*\*\*  
## 11 education\_level4 -2.970 7.510453e-09 \*\*\*  
## 12 education\_level5 -2.896 1.143797e-29 \*\*\*  
## 13 ethnicity2 0.078 0.463   
## 14 ethnicity3 0.278 0.030 \*  
## 15 ethnicity4 1.155 1.603993e-27 \*\*\*  
## 16 wealth\_index2 -0.125 0.143   
## 17 wealth\_index3 -0.463 1.158858e-05 \*\*\*  
## 18 wealth\_index4 -0.264 0.024 \*  
## 19 wealth\_index5 -0.541 7.444127e-04 \*\*\*  
## 20 health\_insurance -0.052 0.533   
## 21 current\_contraceptive\_use 0.485 1.004528e-14 \*\*\*  
## 22 awareness\_hiv\_aids -0.194 0.011 \*  
## 23 access\_to\_media -0.079 0.355   
## Estimate\_NoImpute pValue\_NoImpute Significance\_NoImpute  
## 1 -1.709 2.081962e-13 \*\*\*  
## 2 0.172 0.087   
## 3 0.105 0.511   
## 4 0.036 0.818   
## 5 -0.104 0.541   
## 6 0.061 0.673   
## 7 0.103 0.492   
## 8 0.125 0.284   
## 9 0.078 0.509   
## 10 -0.719 1.905736e-06 \*\*\*  
## 11 -2.830 9.081803e-05 \*\*\*  
## 12 -2.619 2.273156e-08 \*\*\*  
## 13 0.551 7.214439e-05 \*\*\*  
## 14 0.339 0.016 \*  
## 15 1.901 6.687210e-32 \*\*\*  
## 16 -0.056 0.607   
## 17 -0.343 0.008 \*\*  
## 18 -0.150 0.299   
## 19 -0.416 0.031 \*  
## 20 -0.079 0.410   
## 21 -0.028 0.704   
## 22 -0.086 0.361   
## 23 0.178 0.107

* Education levels and ethnicity (specific categories) are robust across both models.
* Certain variables like area, region, and wealth index are sensitive to imputation, suggesting that missing data in these areas could be informative.
* The loss of significance in many variables in the non-imputed model suggests that the missing data may not be random and could be related to the variables’ influence on the outcome.

# Applying the function to each model  
interaction\_impute\_data <- extract\_model\_data(model\_interaction)  
interaction\_no\_impute\_data <- extract\_model\_data(interaction\_model\_no\_impute)  
  
# Combine the data for comparison  
interaction\_comparison\_data <- merge(interaction\_impute\_data, interaction\_no\_impute\_data, by = "Term", suffixes = c("\_Impute", "\_NoImpute"), sort = FALSE)  
  
# View the comparison  
print(interaction\_comparison\_data)

## Term Estimate\_Impute pValue\_Impute Significance\_Impute  
## 1 (Intercept) -1.785 4.599871e-17 \*\*\*  
## 2 area2 0.546 3.488899e-04 \*\*\*  
## 3 region2 0.365 0.005 \*\*  
## 4 region3 0.168 0.192   
## 5 region4 0.323 0.011 \*  
## 6 region5 -0.068 0.593   
## 7 region6 -0.061 0.651   
## 8 education\_level1 -0.061 0.513   
## 9 education\_level2 -0.219 0.018 \*  
## 10 education\_level3 -1.173 5.833944e-22 \*\*\*  
## 11 education\_level4 -2.969 7.739192e-09 \*\*\*  
## 12 education\_level5 -2.896 3.473185e-29 \*\*\*  
## 13 ethnicity2 0.062 0.561   
## 14 ethnicity3 0.273 0.034 \*  
## 15 ethnicity4 1.129 4.598972e-26 \*\*\*  
## 16 wealth\_index2 0.267 0.169   
## 17 wealth\_index3 -0.155 0.464   
## 18 wealth\_index4 -0.042 0.852   
## 19 wealth\_index5 -0.328 0.220   
## 20 health\_insurance -0.042 0.609   
## 21 current\_contraceptive\_use 0.492 4.348509e-15 \*\*\*  
## 22 awareness\_hiv\_aids -0.187 0.014 \*  
## 23 access\_to\_media -0.072 0.398   
## 24 area2:wealth\_index2 -0.479 0.026 \*  
## 25 area2:wealth\_index3 -0.383 0.113   
## 26 area2:wealth\_index4 -0.262 0.306   
## 27 area2:wealth\_index5 -0.250 0.434   
## Estimate\_NoImpute pValue\_NoImpute Significance\_NoImpute  
## 1 -1.712 1.561885e-08 \*\*\*  
## 2 0.151 0.533   
## 3 0.126 0.435   
## 4 0.058 0.711   
## 5 -0.088 0.606   
## 6 0.082 0.577   
## 7 0.125 0.408   
## 8 0.124 0.289   
## 9 0.077 0.516   
## 10 -0.712 2.393561e-06 \*\*\*  
## 11 -2.797 1.104547e-04 \*\*\*  
## 12 -2.584 3.649218e-08 \*\*\*  
## 13 0.551 7.354197e-05 \*\*\*  
## 14 0.334 0.017 \*  
## 15 1.901 2.093819e-31 \*\*\*  
## 16 0.027 0.922   
## 17 -0.354 0.220   
## 18 -0.240 0.425   
## 19 -0.664 0.075   
## 20 -0.075 0.435   
## 21 -0.028 0.708   
## 22 -0.086 0.366   
## 23 0.177 0.108   
## 24 -0.116 0.694   
## 25 0.011 0.971   
## 26 0.135 0.681   
## 27 0.377 0.364

* The effects of imputation are most pronounced in geographical variables (area and region) and some interactions, indicating a potential non-random pattern in how missing data occurs in these variables.
* Certain variables like higher education levels and some ethnicity categories maintain their predictive power regardless of missing data imputation, indicating a stronger and more consistent relationship with the outcome.