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)

### 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,  
# contraceptive\_decision\_maker = CP5, #R  
 current\_contraceptive\_use = CP2,  
# age\_first\_sexual\_intercourse = SB1, #R  
# currently\_married = MA1,  
# partner\_age = MA2, #R  
 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

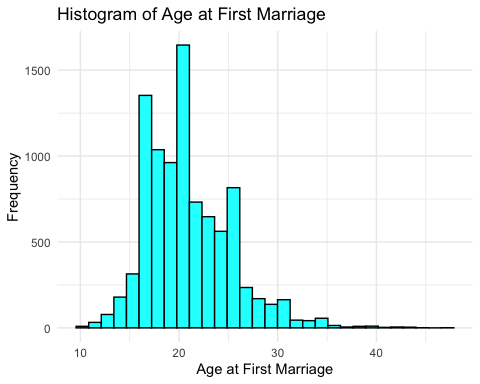
# Exporting female\_df to a CSV file in the current working directory  
#write.csv(female\_df, "female\_df.csv", row.names = FALSE)

### Distributions of Data

# Histogram for 'Age at First Marriage'  
afm\_hist <- ggplot(female\_df, aes(x = age\_first\_marriage)) +  
 geom\_histogram(fill = "cyan", 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()`).



# Saving the histogram  
#ggsave("histogram\_age\_first\_marriage.png", plot = afm\_hist, width = 8, height = 6, dpi = 300)

# # Plotting a boxplot for 'Age at First Marriage'  
# ggplot(female\_df, aes(y = age\_first\_marriage)) +  
# geom\_boxplot(fill = "cyan") +  
# labs(title = "Boxplot of Age at First Marriage", y = "Age at First Marriage") +  
# theme\_minimal()

# # Bar plot for 'used\_internet'  
# ggplot(female\_df, aes(x = used\_internet)) +  
# geom\_bar(fill = "yellow") +  
# theme\_minimal() +  
# ggtitle("Bar Plot of Used Internet") +  
# xlab("Used Internet") +  
# ylab("Count")

### 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 = "lightpink", color = "black", bins = 30) +  
 theme\_light() +  
 ggtitle("Age at First Marriage 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 = "plum", color = "black", bins = 30) +  
 theme\_light() +  
 ggtitle("Age at First Marriage 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))

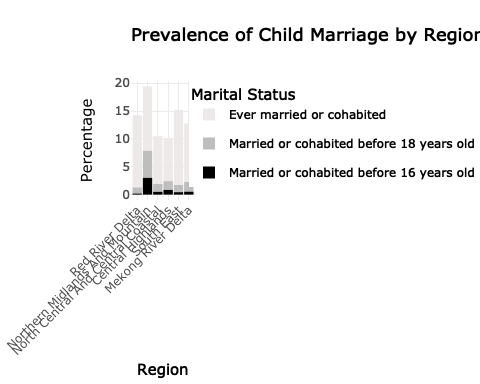
# Create a binary variable for child marriage under 16  
# The new variable "child\_marriage\_u16" will have a value of 1 if the marriage occurred before age 16, and 0 otherwise  
imputed\_df <- imputed\_df %>%  
 mutate(child\_marriage\_u16 = ifelse(age\_first\_marriage < 16, 1, 0))

# 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

# 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  
  
  
# Creating the plot with ggplot2, directly labeling regions  
p <- ggplot(counts\_df, aes(x = factor(region))) +  
 geom\_bar(aes(y = ever\_married, fill = "Ever married or cohabited"), stat = "identity") +  
 geom\_bar(aes(y = married\_u18, fill = "Married or cohabited before 18 years old"), stat = "identity") +  
 geom\_bar(aes(y = married\_u16, fill = "Married or cohabited before 16 years old"), stat = "identity") +  
 scale\_fill\_manual(values = c("Ever married or cohabited" = "snow2",   
 "Married or cohabited before 18 years old" = "grey",   
 "Married or cohabited before 16 years old" = "black"),  
 name = "Marital Status") +  
 labs(x = "Region", y = "Percentage", title = "Prevalence of Child Marriage by Regions Among Females") +  
 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)

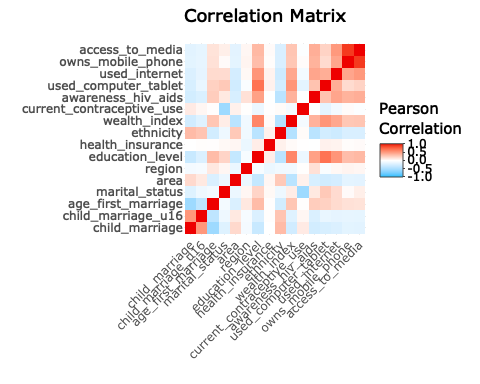


### Preparing for Regression Analysis

#### 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", "red2"),  
 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

# # Mapping for 'area'  
# area\_labels <- c("Urban", "Rural")  
# names(area\_labels) <- c("1", "2")  
#   
# # Mapping for 'region'  
# region\_labels <- c("Red River Delta", "Northern Midlands And Mountain", "North Central And Central Coastal", "Central Highlands", "South East", "Mekong River Delta")  
# names(region\_labels) <- c("1", "2", "3", "4", "5", "6")  
#   
# # Mapping for 'education\_level'  
# education\_level\_labels <- c("None or Pre-Primary", "Primary", "Lower Primary", "Upper Primary", "Vocational High School", "University/College/Higher")  
# names(education\_level\_labels) <- c("0", "1", "2", "3", "4", "5")   
#   
# # Mapping for 'ethnicity'  
# ethnicity\_labels <- c("Kinh and Hoa", "Tay, Thai, Muong, Nung", "Khmer", "Mong") # Add or modify as per your dataset  
# names(ethnicity\_labels) <- c("1", "2", "3", "4")   
#   
# # Mapping for 'wealth\_index'  
# wealth\_index\_labels <- c("Poorest", "Poor", "Middle", "Rich", "Richest")  
# names(wealth\_index\_labels) <- c("1", "2", "3", "4", "5")

# 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

cat("AIC (base):", AIC(baseline\_model), "\nBIC (base):", BIC(baseline\_model), "\n")

## AIC (base): 8523.675   
## BIC (base): 8633.655

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

## AIC (w/ interaction terms): 8288.662   
## BIC (w/ interaction terms): 8486.627

### 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>