Final project code

Ruisi Geng

2023-12-07

Load data

```
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(pvclust)
library(cluster)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
# read data
data1 <- read.csv("~/Downloads/data_1.csv", header=TRUE)
data1<- data1[,-1]

# Data source and structure description
cat("Data Structure Description:\n")
```

```
## Data Structure Description:
```

```
str(data1)
```

```
## 'data.frame':    14728 obs. of  15 variables:
##  $ M2ID       : int  10005 10005 10005 10005 10005 10005 10005 10005 10015 10015 ...
##  $ B1SPWBU2   : num  48 48 48 48 48 48 48 48 38 38 ...
##  $ B1STINC1   : int  0 0 0 0 0 0 0 0 126250 126250 ...
##  $ B1PAGE_M2.x: int  80 80 80 80 80 80 80 80 53 53 ...
##  $ B1PGENDER  : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ B2DN_STR   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ race       : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ marital    : int  0 0 0 0 0 0 0 0 1 1 ...
##  $ B1SQ2      : int  10 10 10 10 10 10 10 10 8 8 ...
##  $ B1SQ1      : int  10 10 10 10 10 10 10 10 7 7 ...
##  $ B1SQ3      : int  10 10 10 10 10 10 10 10 9 9 ...
```

```
##  $ B2DNEGAV   : num  0 0 0 0.07 0 0 0 0 0.14 0 ...
##  $ meanNA     : num  0.00875 0.00875 0.00875 0.00875 0.00875 ...
##  $ sdNA       : num  0.0247 0.0247 0.0247 0.0247 0.0247 ...
##  $ gender     : int  0 0 0 0 0 0 0 0 0 0 ...
```

EDA

```
# Load necessary libraries for visualization
library(ggplot2)

# EDA
cat("Exploratory Data Analysis with Descriptive Statistics and Visualizations:\n")
```

The descriptive statistical analysis aims to understand sample characteristics

```
## Exploratory Data Analysis with Descriptive Statistics and Visualizations:
```

```
# Age Distribution
cat("Age Distribution:\n")
```

```
## Age Distribution:
```

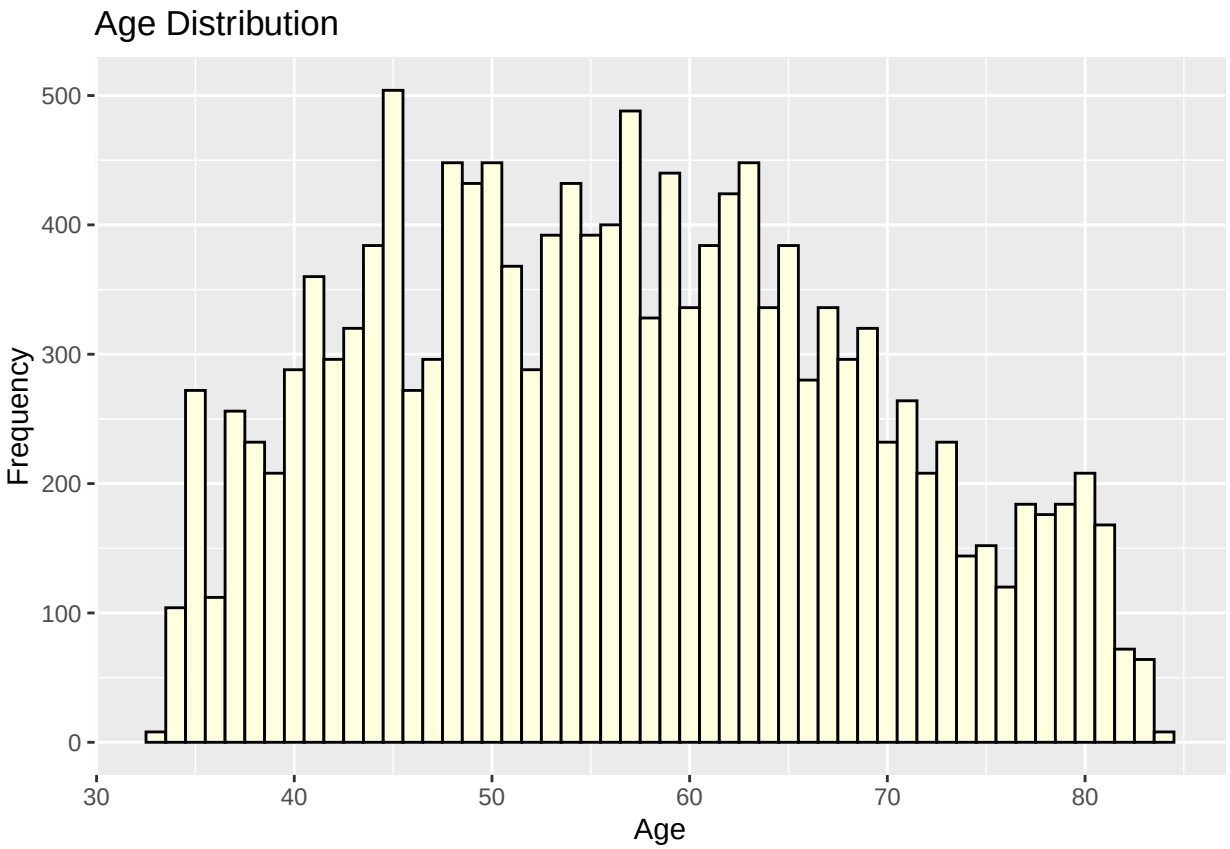
```
summary(data1$B1PAGE_M2.x, na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   33.00   47.00   56.00   56.51   65.00   84.00
```

```
sd(data1$B1PAGE_M2.x, na.rm = TRUE)
```

```
## [1] 12.2322
```

```
ggplot(data1, aes(x = B1PAGE_M2.x)) +
  geom_histogram(binwidth = 1, fill = "lightyellow", color = "black") +
  labs(title = "Age Distribution", x = "Age", y = "Frequency")
```



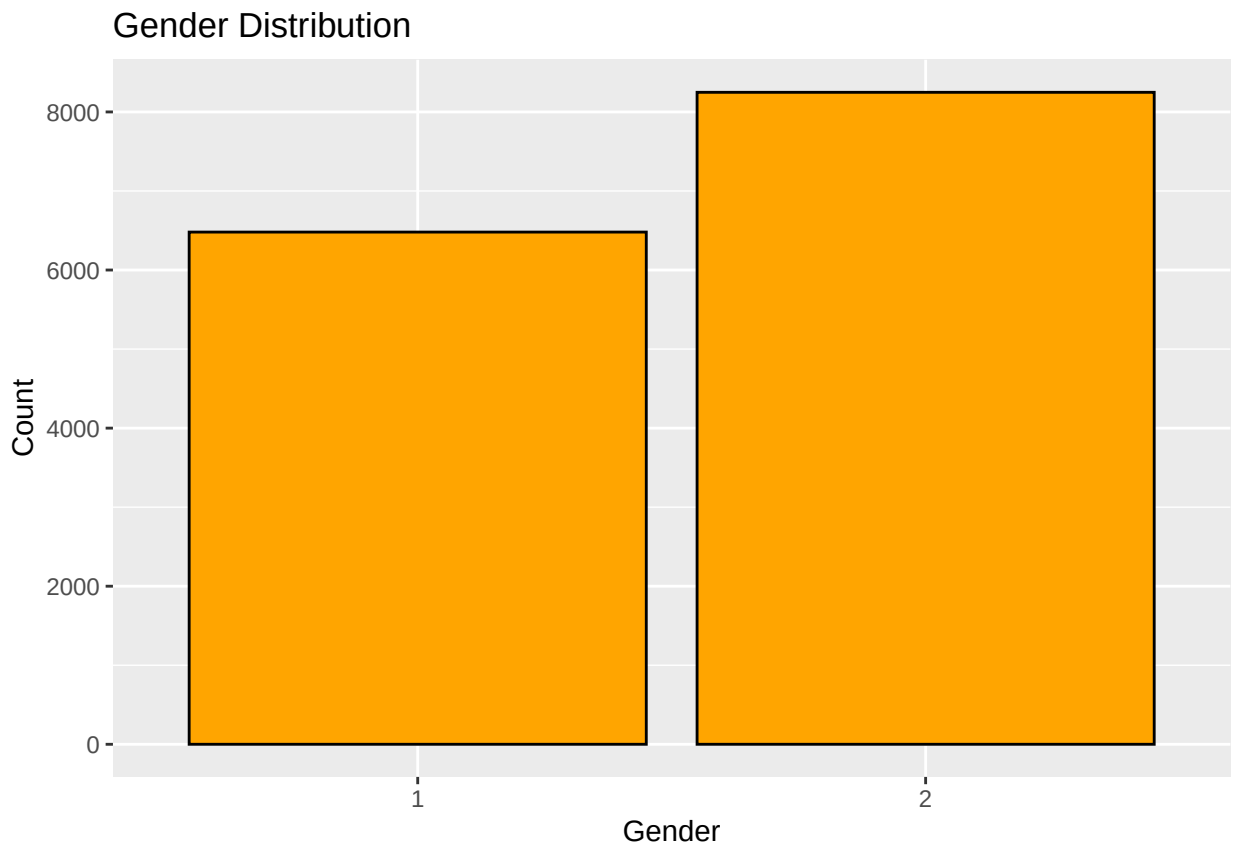

```
# Gender Distribution
cat("Gender Distribution:\n")
```

```
## Gender Distribution:
```

```
summary(data1$B1PGENDER, na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    1.00    2.00    1.56    2.00    2.00
```

```
ggplot(data1, aes(x = factor(B1PGENDER))) +
  geom_bar(fill = "orange", color = "black") +
  labs(title = "Gender Distribution", x = "Gender", y = "Count")
```




```
# Income Distribution
cat("Income Distribution:\n")
```

```
## Income Distribution:
```

```
summary(data1$B1STINC1, na.rm = TRUE)
```

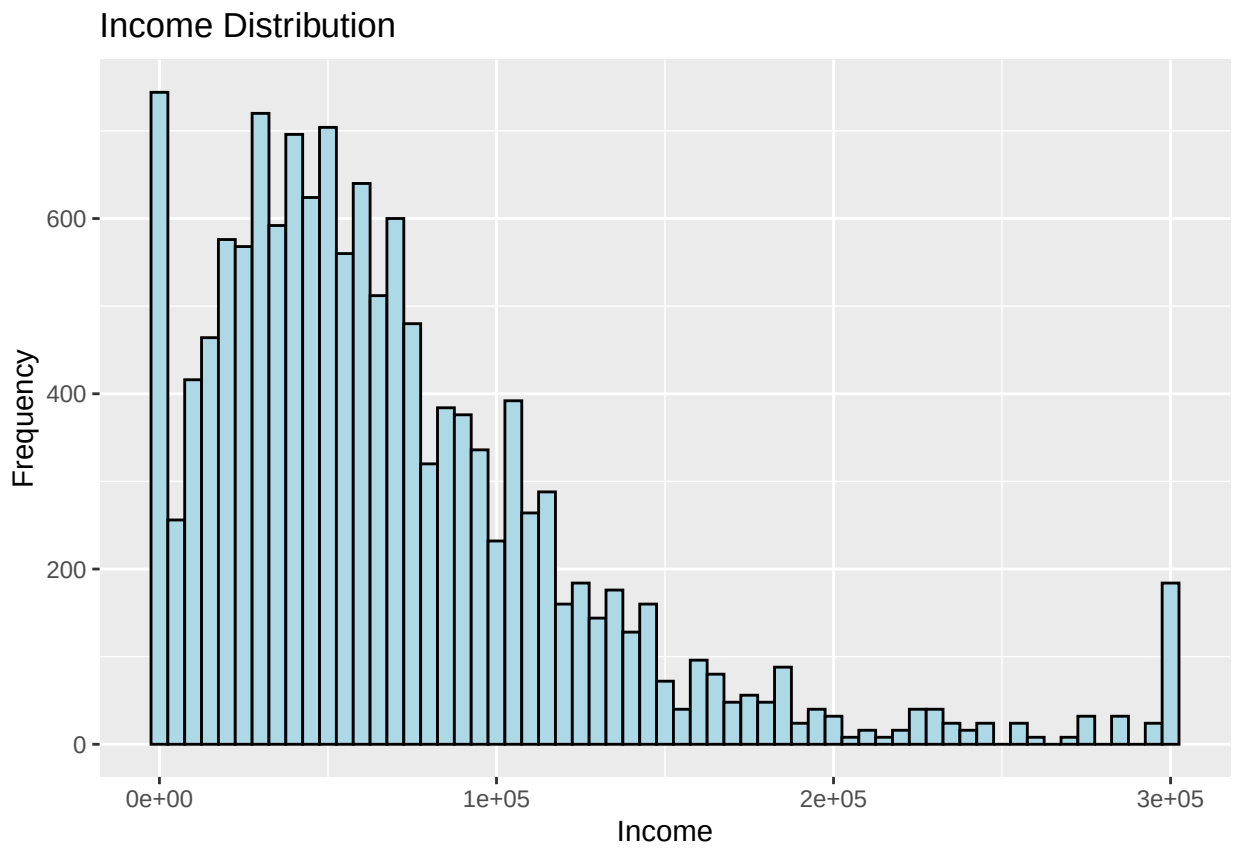
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0   30178   57500   70508   93750  300000     904
```

```
sd(data1$B1STINC1, na.rm = TRUE)
```

```
## [1] 57837.37
```

```
ggplot(data1, aes(x = B1STINC1)) +
  geom_histogram(binwidth = 5000, fill = "lightblue", color = "black") +
  labs(title = "Income Distribution", x = "Income", y = "Frequency")
```

```
## Warning: Removed 904 rows containing non-finite values (`stat_bin()`).
```



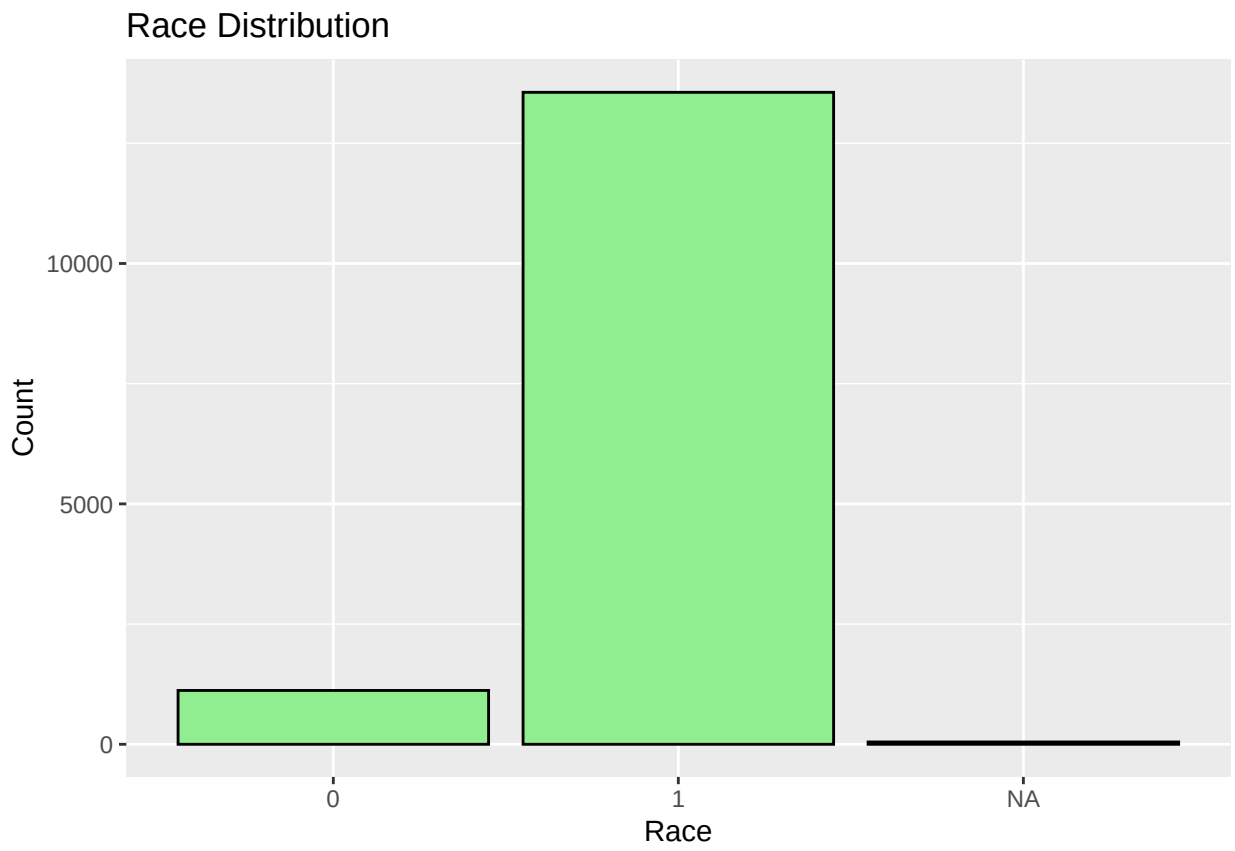

```
# Race Distribution
cat("Race Distribution:\n")
```

```
## Race Distribution:
```

```
summary(data1$race, na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  1.0000  1.0000  0.9237  1.0000  1.0000      48
```

```
ggplot(data1, aes(x = factor(race))) +
  geom_bar(fill = "lightgreen", color = "black") +
  labs(title = "Race Distribution", x = "Race", y = "Count")
```



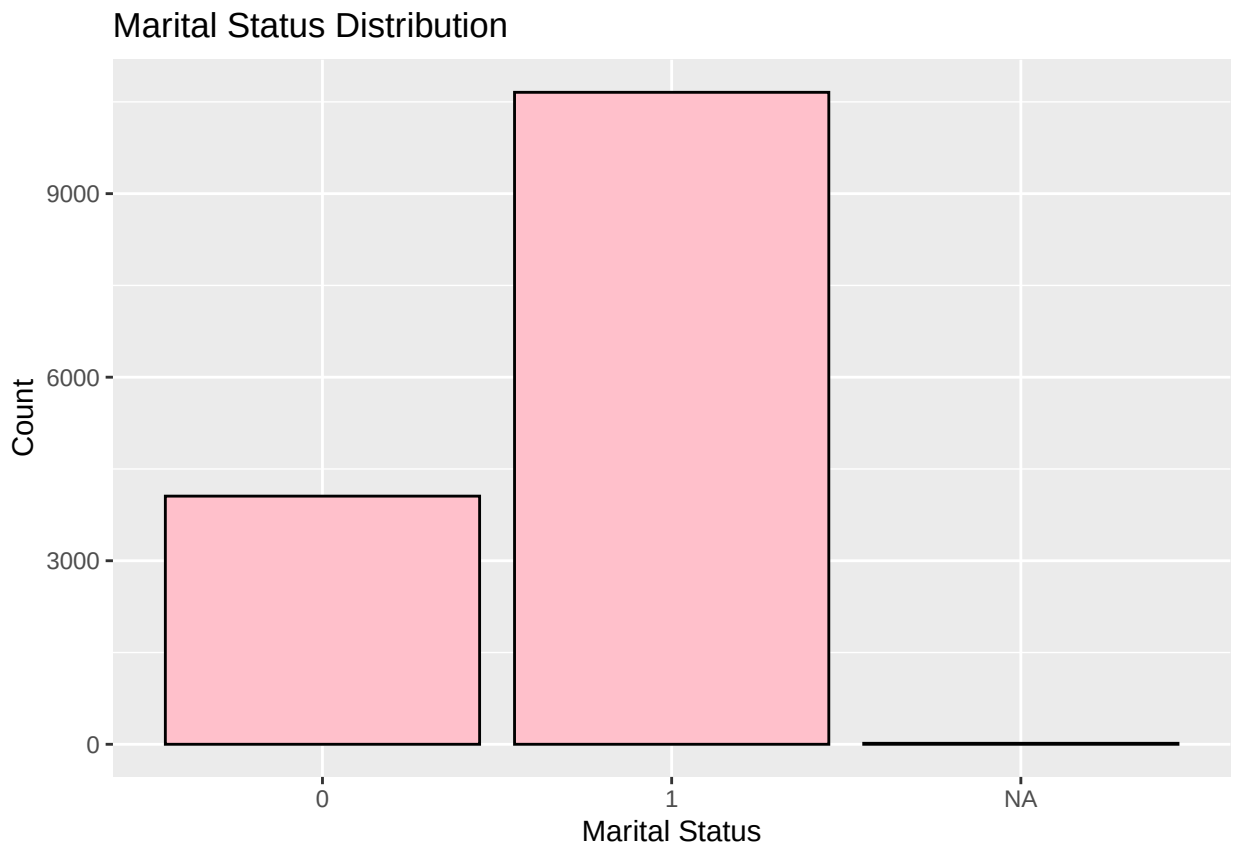

```
# Marital Status
cat("Marital Status Distribution:\n")
```

```
## Marital Status Distribution:
```

```
summary(data1$marital, na.rm = TRUE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  1.0000  0.7243  1.0000  1.0000      16
```

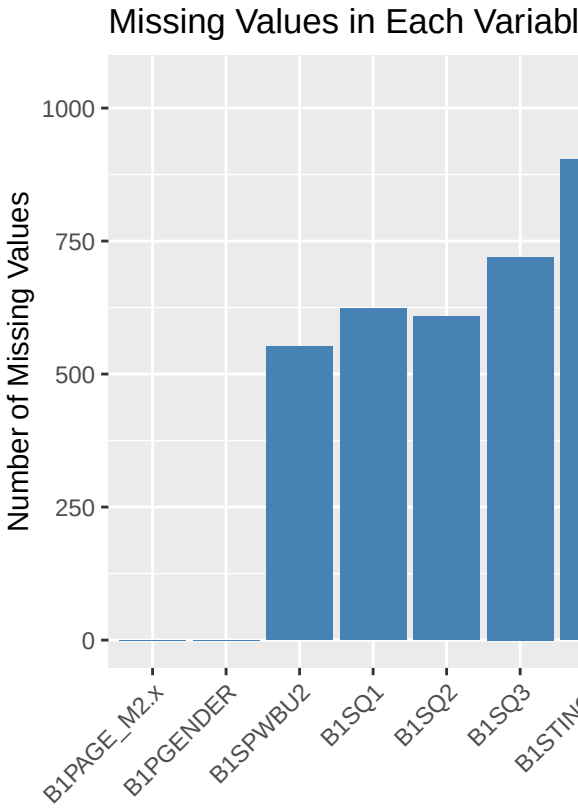
```
ggplot(data1, aes(x = factor(marital))) +
  geom_bar(fill = "pink", color = "black") +
  labs(title = "Marital Status Distribution", x = "Marital Status", y = "Count")
```



Missing value Handling

```
# Create a data frame with the number of missing values for each variable
na_counts <- data1 %>% summarise_all(~sum(is.na(.))) %>% gather(key = "Variable", value = "NA_Count")

# Draw a bar chart of missing values
ggplot(na_counts, aes(x = Variable, y = NA_Count)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Missing Values in Each Variable", x = "Variable", y = "Number of Missing Values")
```



First, visually display the missing values, and then process them.

```
# Function to calculate the mode
get_mode <- function(v) {
  uniqv <- unique(na.omit(v))
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Missing value handling

# Use median to fill missing values for continuous variables
data1$B1SPWBU2[is.na(data1$B1SPWBU2)] <- median(data1$B1SPWBU2, na.rm = TRUE)
data1$B1STINC1[is.na(data1$B1STINC1)] <- median(data1$B1STINC1, na.rm = TRUE)
data1$B2DNEGAV[is.na(data1$B2DNEGAV)] <- median(data1$B2DNEGAV, na.rm = TRUE)

# Use mode to fill missing values for categorical variables
data1$B2DN_STR[is.na(data1$B2DN_STR)] <- get_mode(data1$B2DN_STR)
data1$race[is.na(data1$race)] <- get_mode(data1$race)
data1$marital[is.na(data1$marital)] <- get_mode(data1$marital)

# For rating variables (assuming a 1-10 scale), use median or mode
data1$B1SQ2[is.na(data1$B1SQ2)] <- median(data1$B1SQ2, na.rm = TRUE)
data1$B1SQ1[is.na(data1$B1SQ1)] <- median(data1$B1SQ1, na.rm = TRUE)
data1$B1SQ3[is.na(data1$B1SQ3)] <- median(data1$B1SQ3, na.rm = TRUE)

# Print the updated data to check
sum(is.na(data1))
```

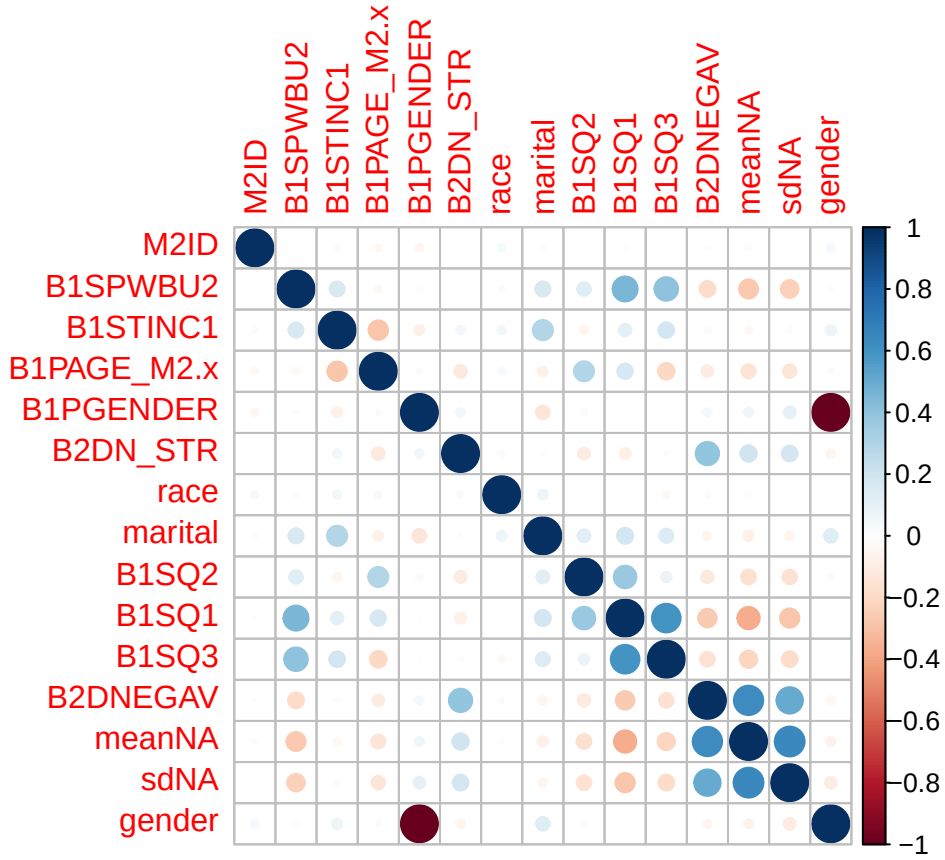
```
## [1] 0
```

More dimensions of data visualization

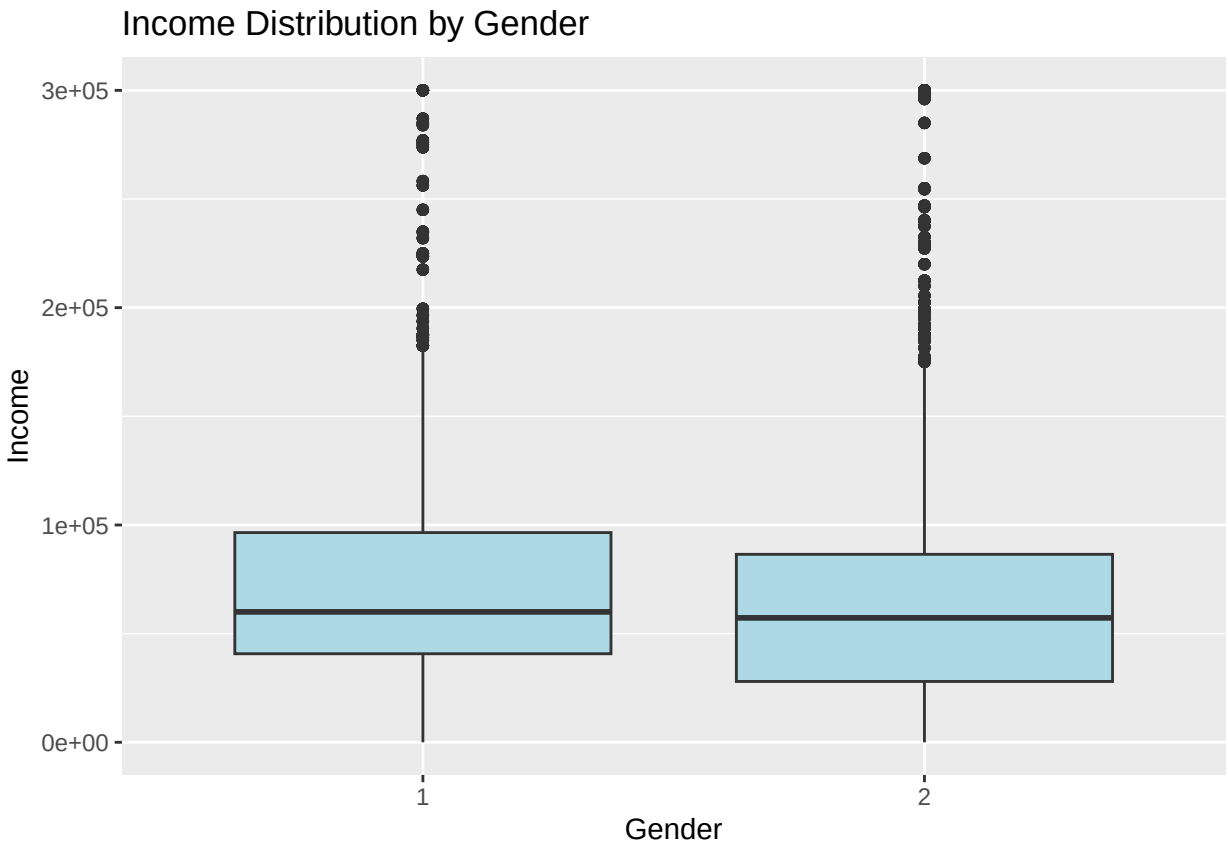
```
# correlation matrix - Correlation between variables
library(corrplot)
```

```
## corrplot 0.92 loaded
```

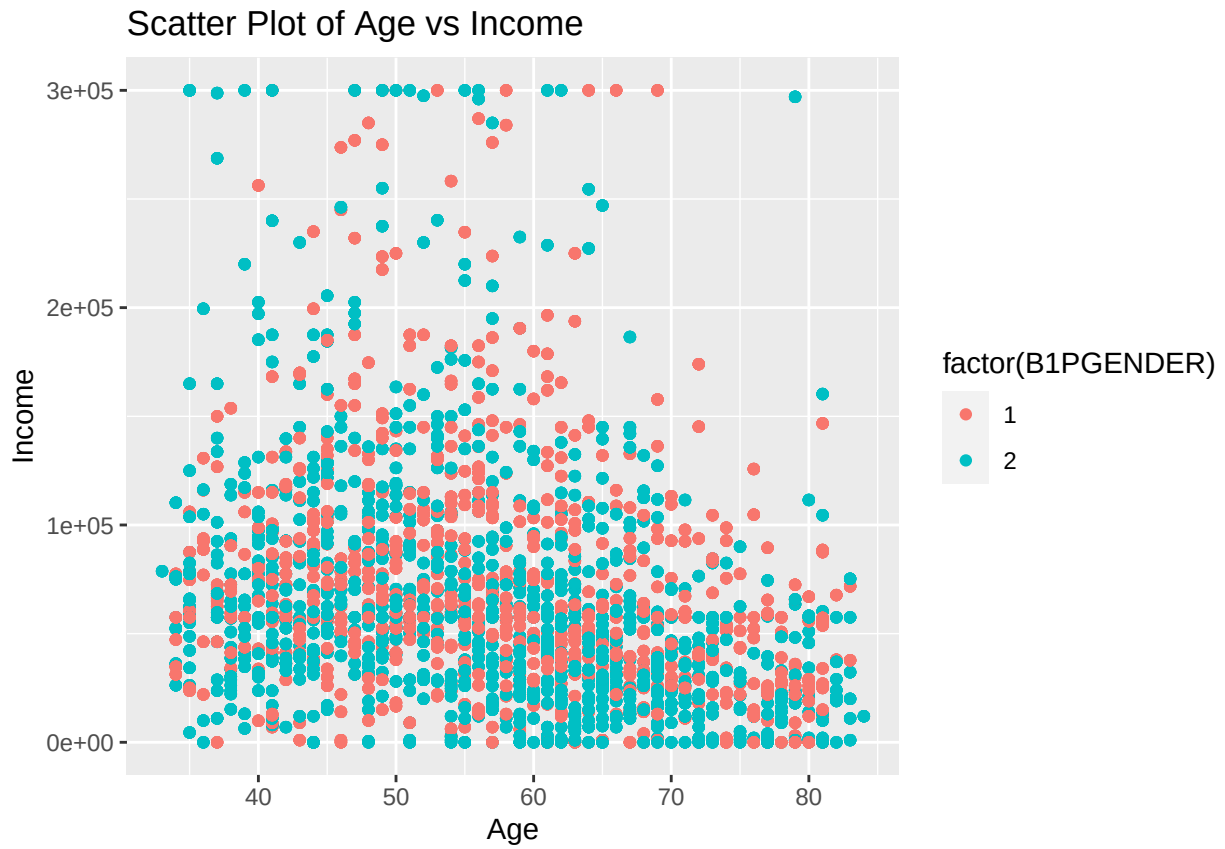
```
M <- cor(data1, use = "complete.obs")
corrplot(M, method = "circle")
```



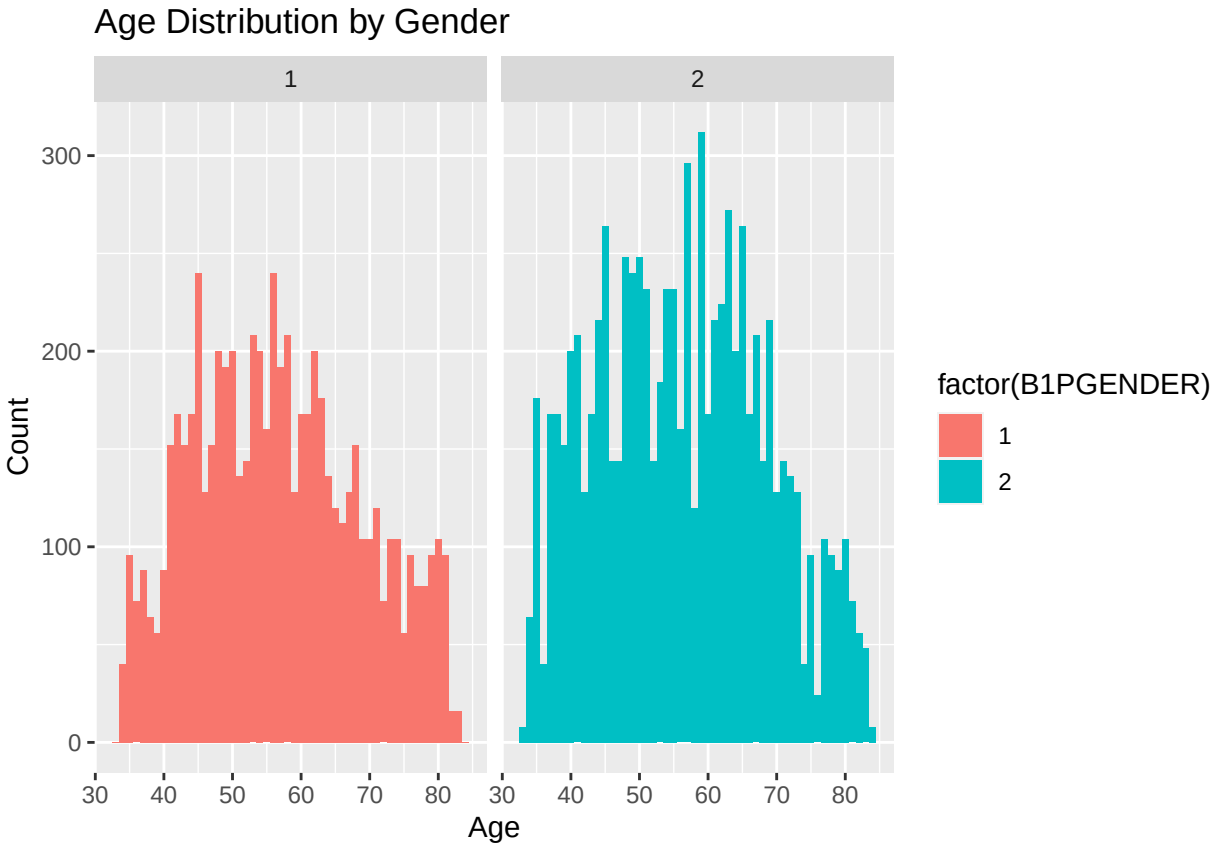

```
# Comparative analysis - Income distribution by sex
ggplot(data1, aes(x = factor(B1PGENDER), y = B1STINC1)) +
  geom_boxplot(fill = "lightblue") +
  labs(title = "Income Distribution by Gender", x = "Gender", y = "Income")
```




```
# Correlation analysis - Scatter plot of age and income
ggplot(data1, aes(x = B1PAGE_M2.x, y = B1STINC1)) +
  geom_point(aes(color = factor(B1PGENDER))) +
  labs(title = "Scatter Plot of Age vs Income", x = "Age", y = "Income")
```




```
# multivariate analysis - gender difference of income
ggplot(data1, aes(x = B1PAGE_M2.x, fill = factor(B1PGENDER))) +
  geom_histogram(binwidth = 1) +
  facet_wrap(~ B1PGENDER) +
  labs(title = "Age Distribution by Gender", x = "Age", y = "Count")
```


Age Distribution by Gender
1
2
300
200
100
0
Count
30
40
50
60
70
80
Age
factor(B1PGENDER)
1
2