# Missing Data - Assignment 1

Aga, Nisse, Ruben

2024-02-21

## Contents

## Introduction

## Methodology?

### Data

Description of the dataset source and variables selection. ##

## Load

```
library(tidyverse)
library(fastDummies)
library(kableExtra)
library(gridExtra, exclude="combine")
library(lubridate)
library(car)
library(ICC)
library(caret)
library(pROC)
library(naniar)
library(ggmice)
library(mice)
```

```
data <- readRDS("../data/data.rds") %>%
  select("drink_regularly", "sex", "age", "ethnicity", "education", "marital", "household_income",  "dep
  as_tibble()

#Adding the depression score from the individual depression items, removing the items
data1 <- mutate(data, dep_score = dep1 + dep2 + dep3 + dep4 + dep5 + dep6) %>%
  select("drink_regularly", "sex", "age", "ethnicity", "education", "marital", "household_income",  "dep
```

# EDA

```
summary(data)
```

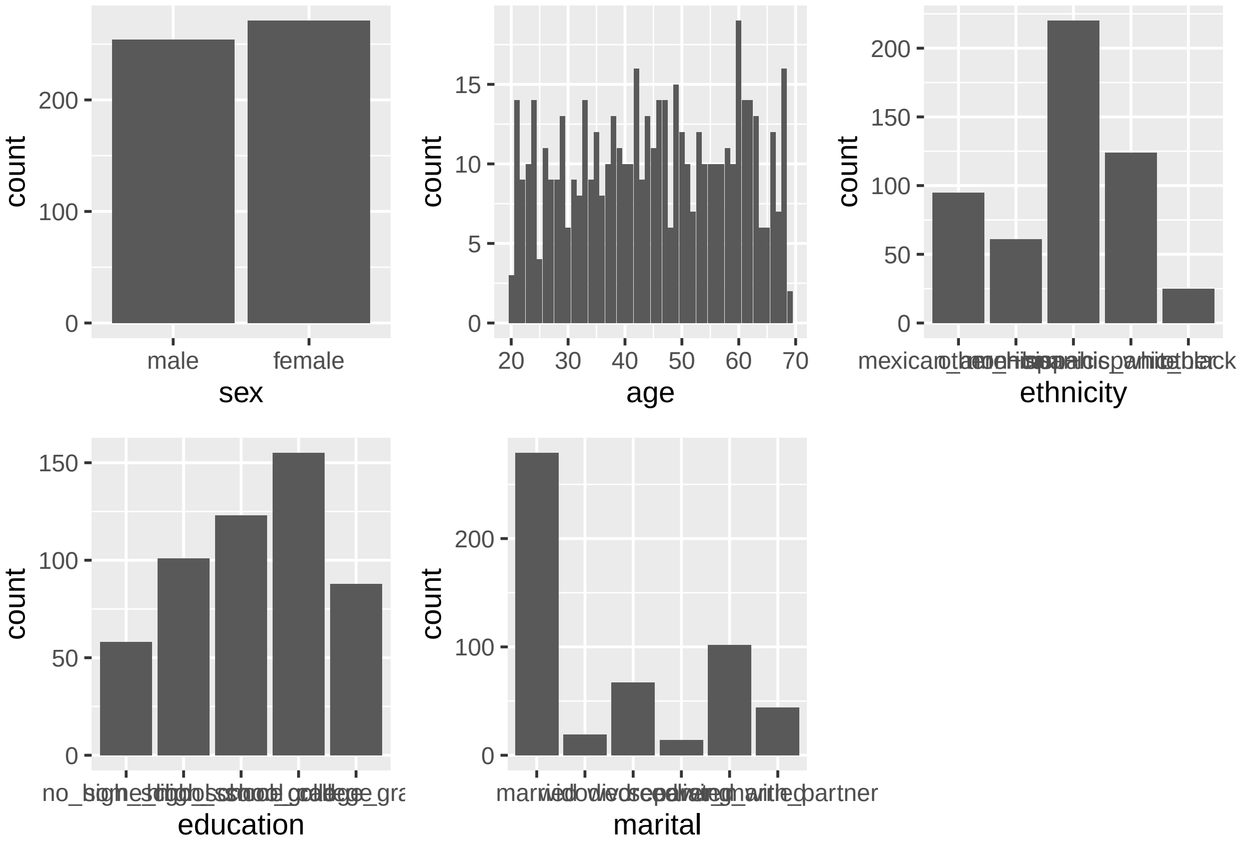
```
##  drink_regularly     sex           age                     ethnicity
##  yes :307        male  :254   Min.   :20.00   mexican_american  : 95
##  no  :139        female:271   1st Qu.:33.00   other_hispanic    : 61
##  NA’s: 79                     Median :45.00   non-hispanic_white:220
##                               Mean   :44.99   non-hispanic_black:124
##                               3rd Qu.:57.00   other             : 25
##                               Max.   :69.00
##
##             education                    marital       household_income
##  no_high_school  : 58   married            :279   100000+    : 76
##  some_high_school:101   widowed            : 19   25000:34999: 59
##  high_school_grad:123   divorced           : 67   20000:24999: 52
##  some_college    :155   separated          : 14   35000:44999: 51
##  college_grad    : 88   never_married      :102   75000:99999: 49
##                         living_with_partner: 44   10000:14999: 45
##                                                   (Other)    :193
##       dep1             dep2             dep3            dep4
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.000   Median :1.0000
##  Mean   :0.4095   Mean   :0.2817   Mean   :0.533   Mean   :0.7562
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :3.0000   Max.   :3.0000   Max.   :3.000   Max.   :3.0000
##                   NA’s   :131      NA’s   :131
##       dep5             dep6             dep7
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.3096   Mean   :0.2005   Mean   :0.3238
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :3.0000   Max.   :3.0000   Max.   :3.0000
##  NA’s   :131      NA’s   :131
```

```
str(data)
```

```
## tibble [525 x 14] (S3: tbl_df/tbl/data.frame)
##  $ drink_regularly : Factor w/ 2 levels "yes","no": 2 2 1 1 NA NA 1 1 2 2 ...
##  $ sex             : Factor w/ 2 levels "male","female": 2 2 1 1 2 2 1 2 2 1 ...
##  $ age             : int [1:525] 45 60 50 39 35 24 60 68 47 41 ...
##  $ ethnicity       : Factor w/ 5 levels "mexican_american",..: 1 2 3 3 3 1 3 3 4 4 ...
##  $ education       : Factor w/ 5 levels "no_high_school",..: 2 1 3 4 4 3 5 2 5 3 ...
##  $ marital         : Factor w/ 6 levels "married","widowed",..: 1 2 3 6 1 5 3 1 5 5 ...
##  $ household_income: Factor w/ 12 levels "0:4999","5000:9999",..: 7 1 4 11 5 5 10 3 10 6 ...
##  $ dep1            : int [1:525] 1 1 0 1 1 0 0 0 0 0 ...
##  $ dep2            : int [1:525] 1 NA NA 0 NA 1 NA 0 0 1 ...
##  $ dep3            : int [1:525] 1 NA NA 1 NA 0 NA 0 0 1 ...
##  $ dep4            : int [1:525] 1 1 0 1 3 1 0 1 0 1 ...
##  $ dep5            : int [1:525] 1 NA NA 0 NA 0 NA 0 0 0 ...
##  $ dep6            : int [1:525] 1 NA NA 1 NA 0 NA 0 0 0 ...
##  $ dep7            : int [1:525] 1 1 1 0 3 0 0 1 0 0 ...
```

```
grid.arrange(ncol = 3,
    ggplot(data, aes(sex)) + geom_histogram(stat = 'count'),
    ggplot(data, aes(age)) + geom_histogram(stat = 'count'),
    ggplot(data, aes(ethnicity)) + geom_histogram(stat = 'count'),
    ggplot(data, aes(education)) + geom_histogram(stat = 'count'),
    ggplot(data, aes(marital)) + geom_histogram(stat = 'count')
)
```

```
## Warning in geom_histogram(stat = "count"): Ignoring unknown parameters: ‘binwidth‘, ‘bins‘, and ‘pad
## Ignoring unknown parameters: ‘binwidth‘, ‘bins‘, and ‘pad‘
## Ignoring unknown parameters: ‘binwidth‘, ‘bins‘, and ‘pad‘
## Ignoring unknown parameters: ‘binwidth‘, ‘bins‘, and ‘pad‘
## Ignoring unknown parameters: ‘binwidth‘, ‘bins‘, and ‘pad‘
```

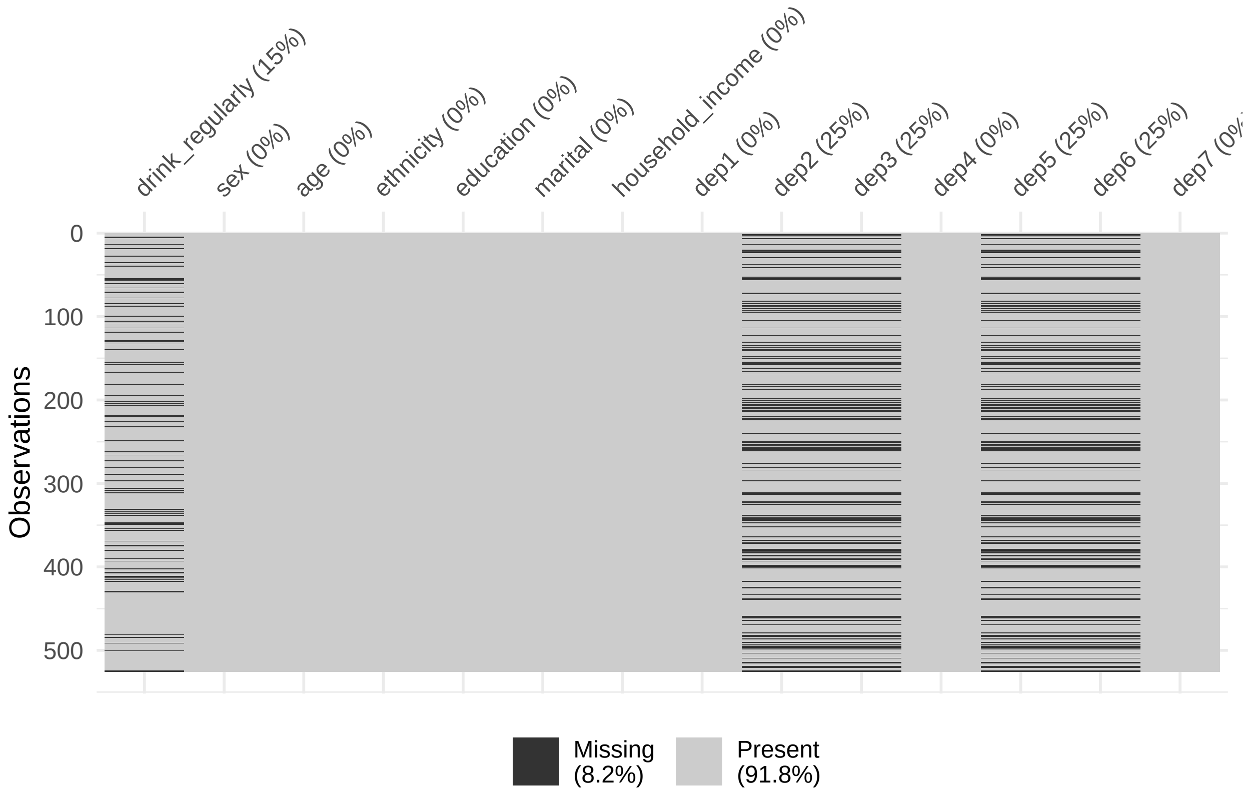


## Missing data and response Patterns

Firstly, we investigate the overall distribution of missing data in our dataset:

```r
# Creates a graph displaying the % of data missing in each variable

vis_miss(data)
```

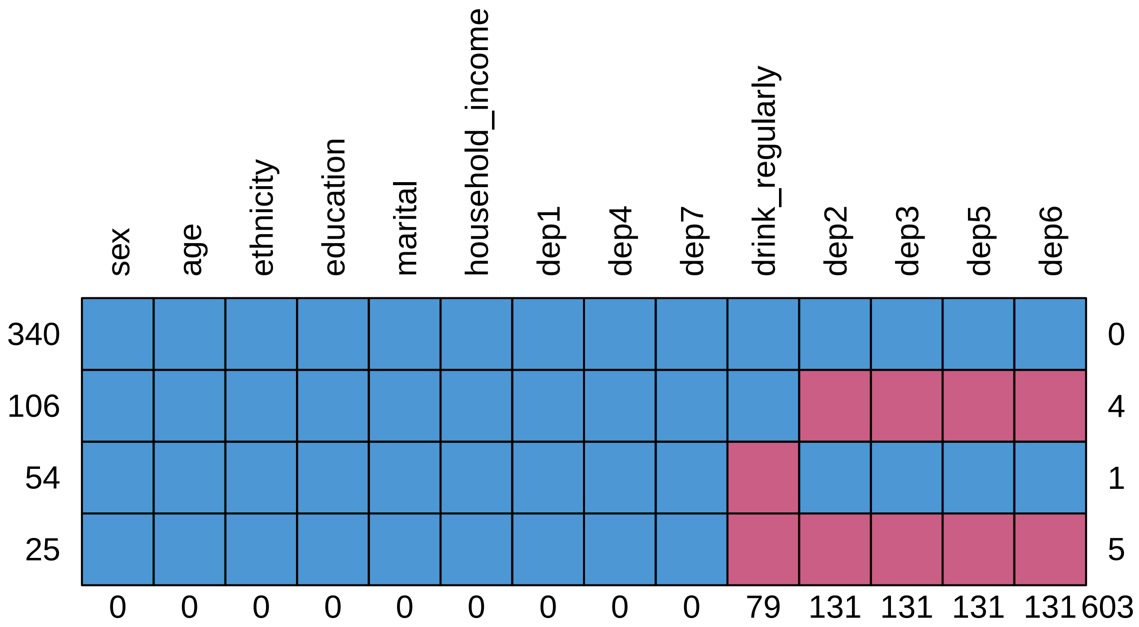


As can be seen on the graph above, 8.2% of the data is missing. The missing values occur in the outcome variable ‘drink_regularly’ and in the responses to questions ‘dep2’, ‘dep3’, ‘dep5’ and ’dep6’that create the depression score variable. 15% of responses are missing for the predictor variable and 25% of the responses are missing for the individual depression questions.

We further investigate the missing data patterns by looking at the response patters:

```r
#Creates a graph with all of the response patterns in the dataset and their frequency

md.pattern(data, rotate = TRUE)
```



```
##     sex age ethnicity education marital household_income dep1 dep4 dep7
## 340   1   1         1         1       1                1    1    1    1
## 106   1   1         1         1       1                1    1    1    1
## 54    1   1         1         1       1                1    1    1    1
## 25    1   1         1         1       1                1    1    1    1
##       0   0         0         0       0                0    0    0    0
##     drink_regularly dep2 dep3 dep5 dep6
## 340               1    1    1    1    1   0
## 106               1    0    0    0    0   4
## 54                0    1    1    1    1   1
## 25                0    0    0    0    0   5
##                  79  131  131  131  131 603
```

This figure reveals that there are four distinct response patterns in the dataset. The most frequent one is no missing entries, with 340 cases. Alternatively, either all four depression entries are missing (106 cases), the predictor variable is missing (54 cases) or both (25 cases). It is very probable that the reason for item non-response for the depression items is the same, since there are no cases of only some of them missing. Since the depression items are missing in this pattern, 25% of the overall depression score will be missing.

```
# Creating vectors that indicate if a value is missing in a given variable. Since the pattern in depres

mdrink <- is.na(data$drink_regularly)
```

```
mdep <- is.na(data$dep2)

# Testing dependency between missing value in var1 and values of var2. Null hypothesis: no dependency.

out1 <- t.test(age ~ mdrink, data = data)
out1$statistic
```

### Testing dependency of missing values

```
##        t
## 19.31658
```

```
out1$p.value
```

```
## [1] 3.099076e-45
```

```
# Should this be on data1 or data?
mcar_test(data1)
```

| statistic | df | p.value | missing.patterns |
|---:|---:|---:|---:|
| 171.3685 | 20 | 0 | 4 |

Thus, the missing values are definitelly not missing at random.