Missing Data - Assignment 1

Aga, Nisse, Ruben

2024-02-22

Contents

1 Introduction

2 Methodology?

2.1 Data

Description of the dataset source and variables selection.

2.2

3 Loading data

We first specify our dependencies and read the data from the `data.rds` file.

```r
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)
```

```r
source <- readRDS("../data/data.rds") %>%
  as_tibble()
```

We then create a sub-selection of variables that are of interest to our model.

```r
data <- source %>%
  select(
    id,
    drink_regularly,
    sex,
    age,
    ethnicity,
    education,
    marital,
    household_income,
    dep1,
    dep2,
    dep3,
    dep4,
    dep5,
    dep6,
    dep7,
    dep8,
    dep9
  )
```

3.1 Variables description

Role	Variable	Name	Type	Characteristics	Target
Outcome	Drink regularly	drink_regularly	Categorical	Binary, yes and no	m/f, age 20-150
Predictor	Sex	sex	Categorical	Binary, male and female	m/f, age 0-150
Predictor	Age	age	Numeric	Discrete	m/f, age 0-150
Predictor	Ethnicity	ethnicity	Categorical	Nominal, 5 categories	m/f, age 0-150
Predictor	Education	marital	Categorical	Nominal, 5 categories	m/f, age 20-150
Predictor	Marital	marital	Categorical	Nominal, 5 categories	m/f, age 20-150
Predictor	Household income	household_income	Categorical	Nominal, 12 categories	m/f, age 0-150
Predictor	No interest in activity	dep1	Categorical	Ordinal, 1-3 scale	m/f, age 18-150
Predictor	Feeling depressed	dep2	Categorical	Ordinal, 1-3 scale	m/f, age 18-150
Predictor	Sleeping issues	dep3	Categorical	Ordinal, 1-3 scale	m/f, age 18-150
Predictor	Feeling tired	dep4	Categorical	Ordinal, 1-3 scale	m/f, age 18-150
Predictor	Eating issues	dep5	Categorical	Ordinal, 1-3 scale	m/f, age 18-150
Predictor	Feeling bad about yourself	dep6	Categorical	Ordinal, 1-3 scale	m/f, age 18-150
Predictor	Concentrating issues	dep7	Categorical	Ordinal, 1-3 scale	m/f, age 18-150
Predictor	Moving and speaking issues	dep8	Categorical	Ordinal, 1-3 scale	m/f, age 18-150
Predictor	Suicidial thoughts	dep9	Categorical	Ordinal, 1-3 scale	m/f, age 18-150

The table above lists the variables used in our subset selection, which will be utilised for the model in question. The predictor variables [*dep*1...*dep*9] are sourced from the same Depression Screener, where respondents of age 18 to 150 were ought to assign a number (1 to 3) regarding their mental and physical state within the last 2 weeks. The demographic variables - that being `sex`, `age`, `ethnicity`, `education` and `household_income` - were taken from the same screening component as well. The following should be noted, regarding these demographic variables:

- The variable `age` is topcoded at the value `80` for the respondents who were older than 80 years.
- The variable `education` was targeted at respondents of age `20` to `150`, thus excluding younger participants. This is due to the fact that this question includes responses such as `AA degree` and `College Graduate`.
- Similarly, the variable `marital` was also targeted at respondents of age `20` to `150`.
- The variable `household_income` is ordinal, rather than continuous.

As for the remaining demographic variables, namely `sex`, `age`, `ethnicity` and `household_income`, these are retrieved from target age `0` to `150`.

Finally, the `drink_regularly` variable was obtained from a an Alcohol Use questionnaire targeted at ages 20 and up.

4 EDA

4.1 Descriptive statistics

```
summary(data)
```

```
##        id        drink_regularly     sex           age
##  Min.   :41531   yes :307        male  :254   Min.   :20.00
##  1st Qu.:43912   no  :139        female:271   1st Qu.:33.00
##  Median :46357   NA's: 79                     Median :45.00
##  Mean   :46470                                Mean   :44.99
```

```
##  3rd Qu.:48934                                3rd Qu.:57.00
##  Max.   :51610                                Max.   :69.00
##
##               ethnicity              education                   marital
##  mexican_american  : 95   no_high_school  : 58   married            :279
##  other_hispanic    : 61   some_high_school:101   widowed            : 19
##  non-hispanic_white:220   high_school_grad:123   divorced           : 67
##  non-hispanic_black:124   some_college    :155   separated          : 14
##  other             : 25   college_grad    : 88   never_married      :102
##                                                  living_with_partner: 44
##
##     household_income      dep1             dep2             dep3
##  100000+    : 76     Min.   :0.0000   Min.   :0.0000   Min.   :0.000
##  25000:34999: 59     1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000
##  20000:24999: 52     Median :0.0000   Median :0.0000   Median :0.000
##  35000:44999: 51     Mean   :0.4095   Mean   :0.2817   Mean   :0.533
##  75000:99999: 49     3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.000
##  10000:14999: 45     Max.   :3.0000   Max.   :3.0000   Max.   :3.000
##  (Other)    :193                      NA's   :131      NA's   :131
##       dep4             dep5             dep6             dep7
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.7562   Mean   :0.3096   Mean   :0.2005   Mean   :0.3238
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :3.0000   Max.   :3.0000   Max.   :3.0000   Max.   :3.0000
##                   NA's   :131      NA's   :131
##       dep8            dep9
##  Min.   :0.000   Min.   :0.00000
##  1st Qu.:0.000   1st Qu.:0.00000
##  Median :0.000   Median :0.00000
##  Mean   :0.203   Mean   :0.06682
##  3rd Qu.:0.000   3rd Qu.:0.00000
##  Max.   :3.000   Max.   :3.00000
##  NA's   :52      NA's   :76
```

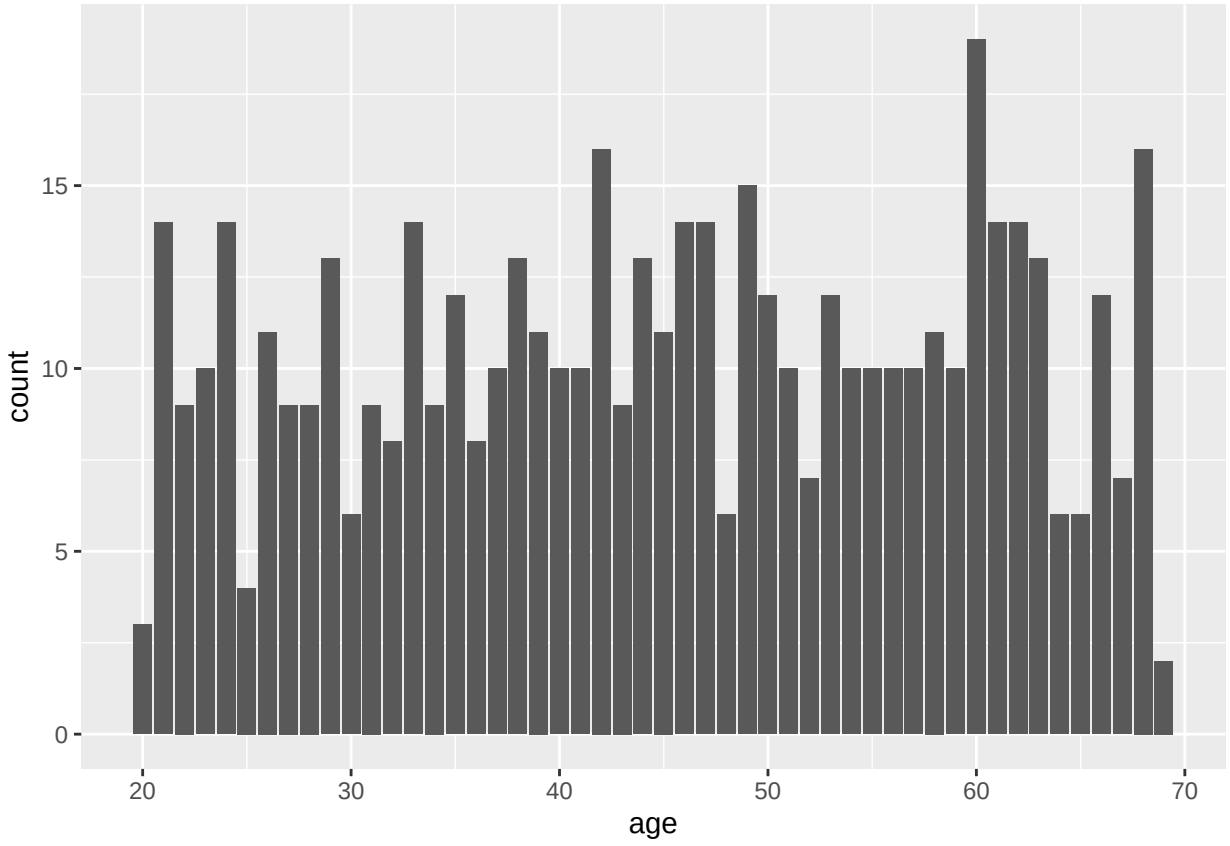
```r
n_rows <- n_distinct(data$id)
```

Notes:
- note: age < 20 is missing from data!!
- 525 unique rows / cases.

4.2 Distributions

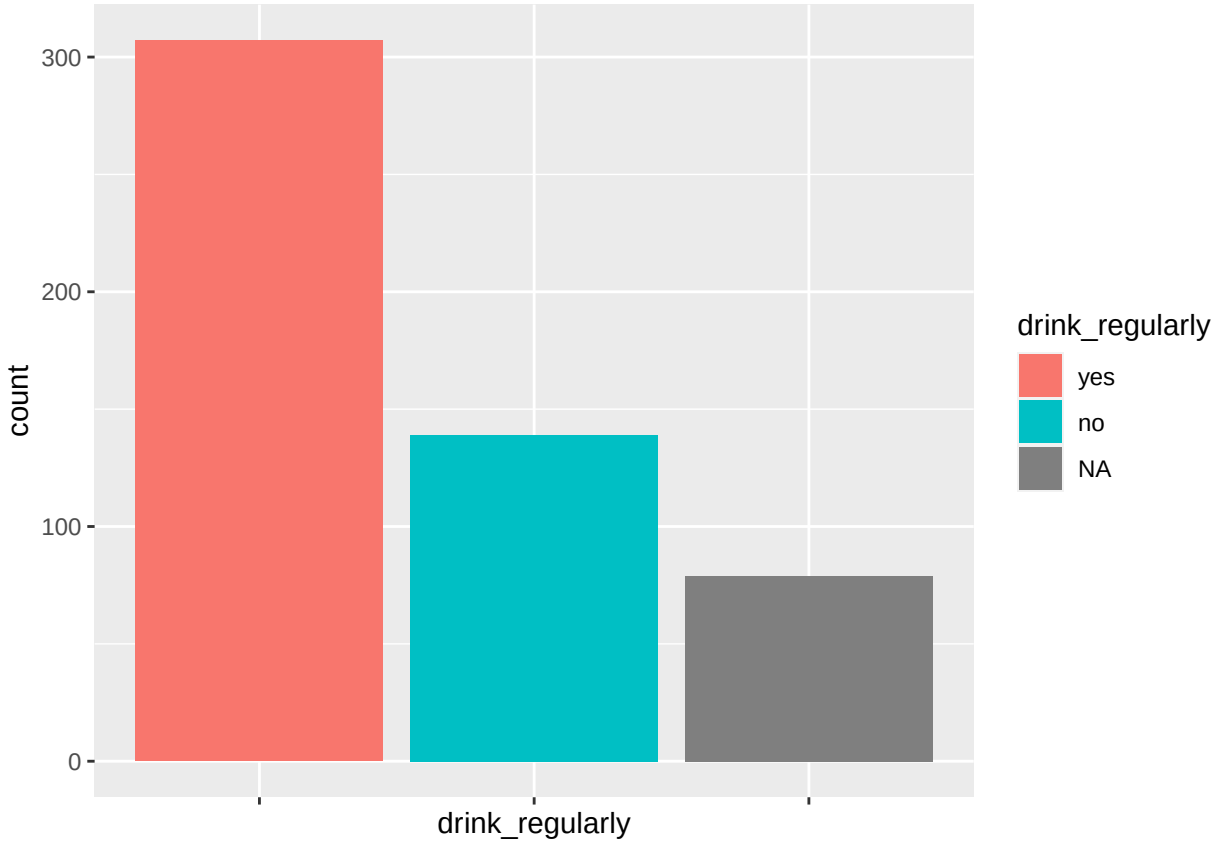
```r
# Continuous
ggplot(data, aes(age)) + geom_histogram(stat = 'count')
```

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



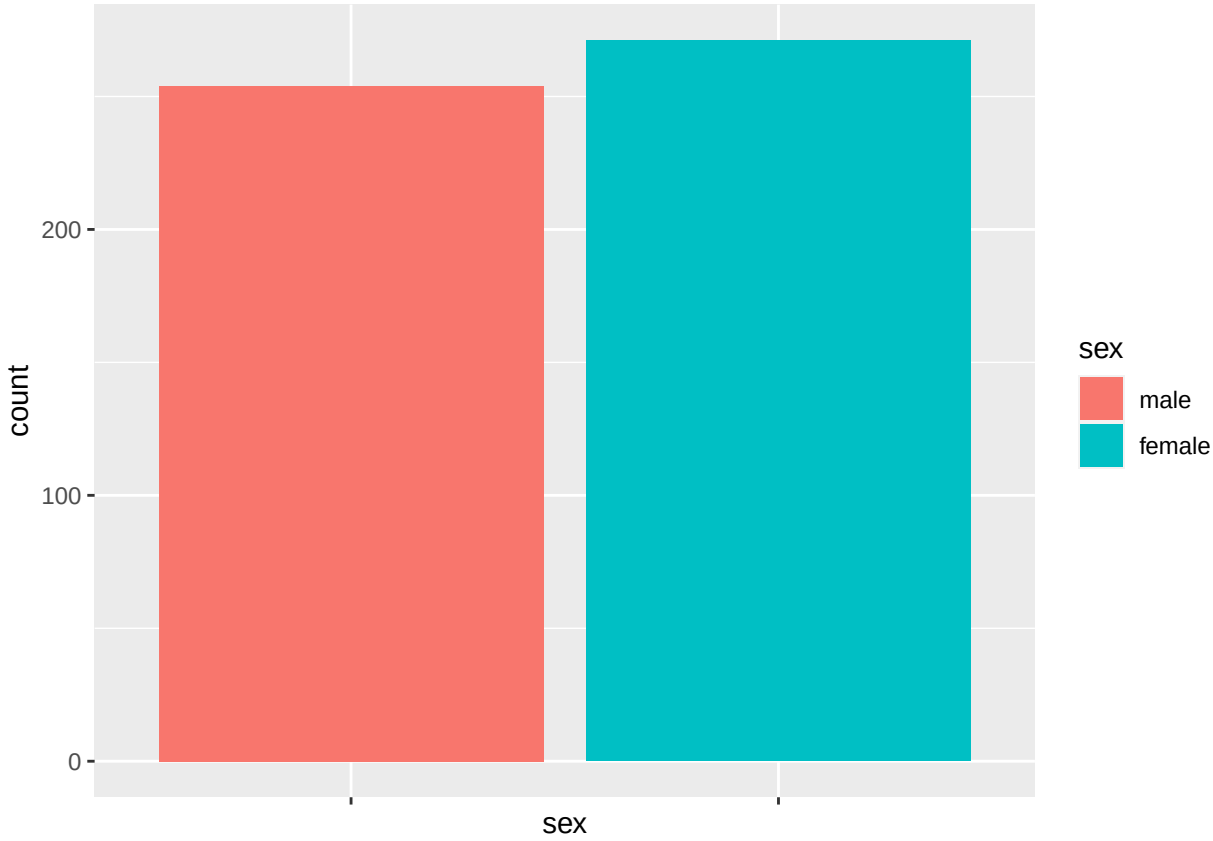

```
# Categorical
categorical_dist <- function(plot) {
  plot +
    geom_histogram(stat = 'count') +
     theme(axis.text.x = element_blank())
}

ggplot(data, aes(drink_regularly, fill = drink_regularly)) %>% categorical_dist()
```

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



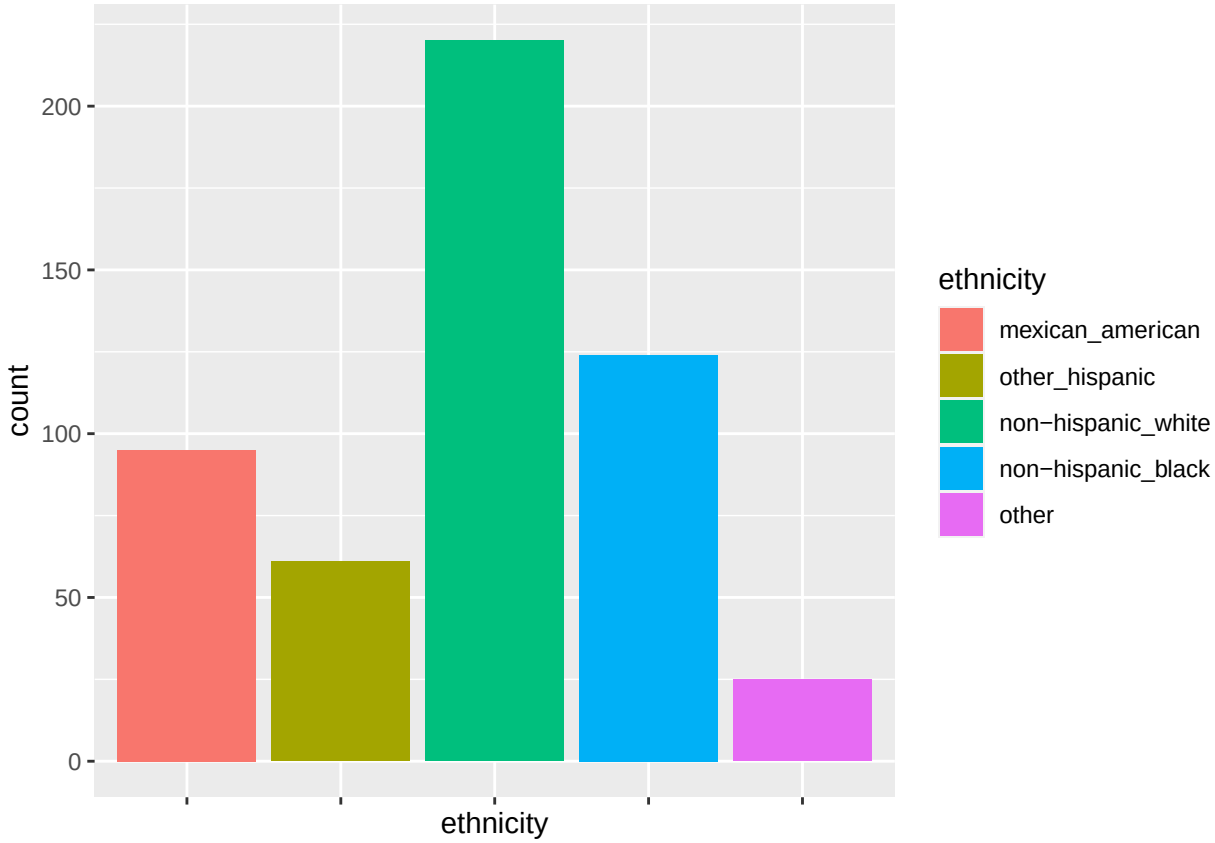

```
ggplot(data, aes(sex, fill = sex)) %>% categorical_dist()
```

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



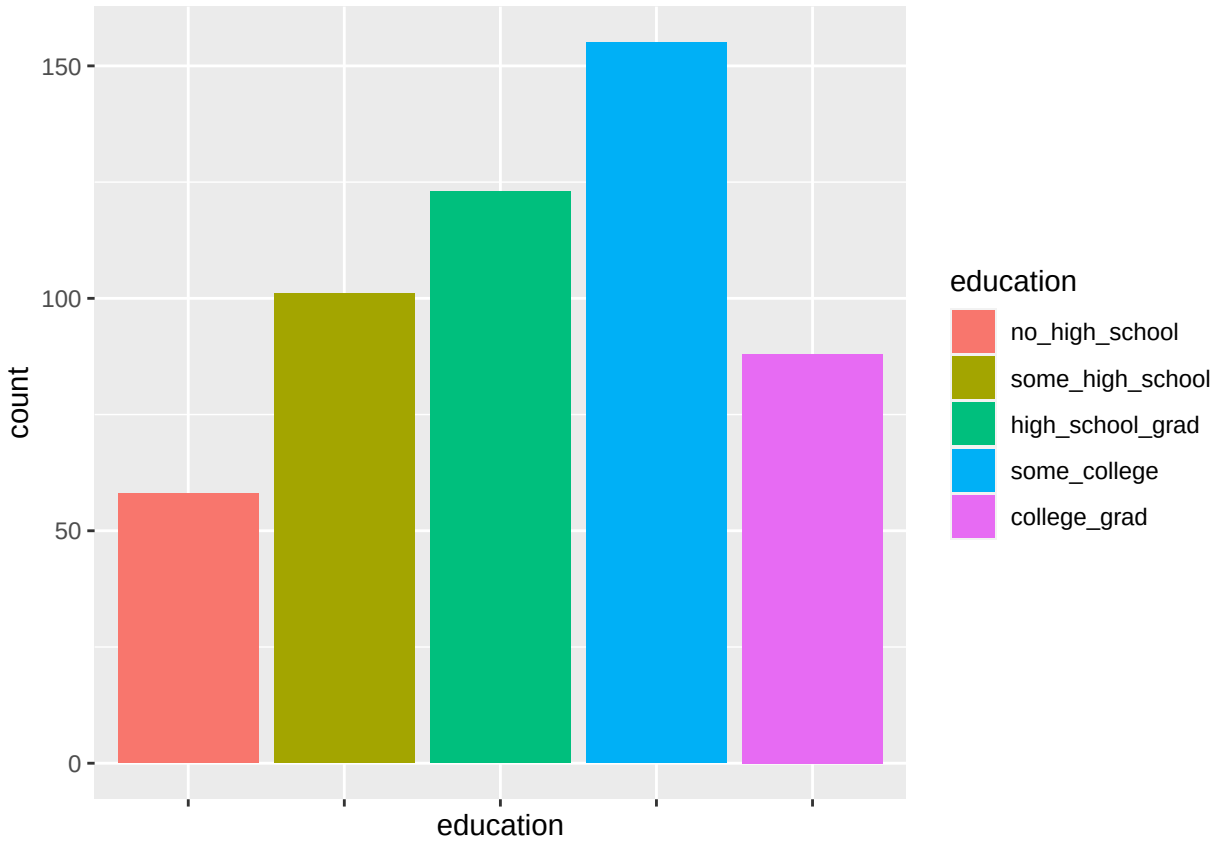

```
ggplot(data, aes(ethnicity, fill = ethnicity)) %>% categorical_dist()
```

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



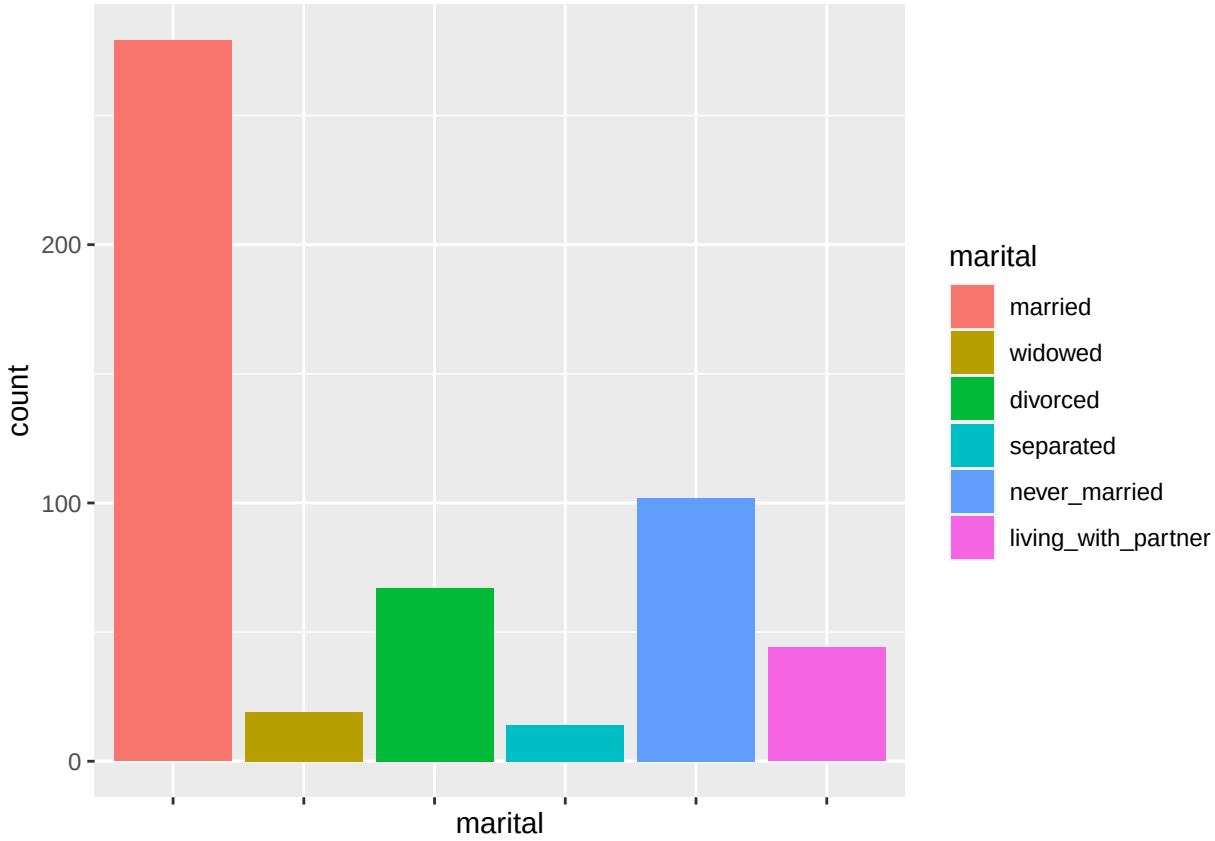

```
ggplot(data, aes(education, fill = education)) %>% categorical_dist()
```

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



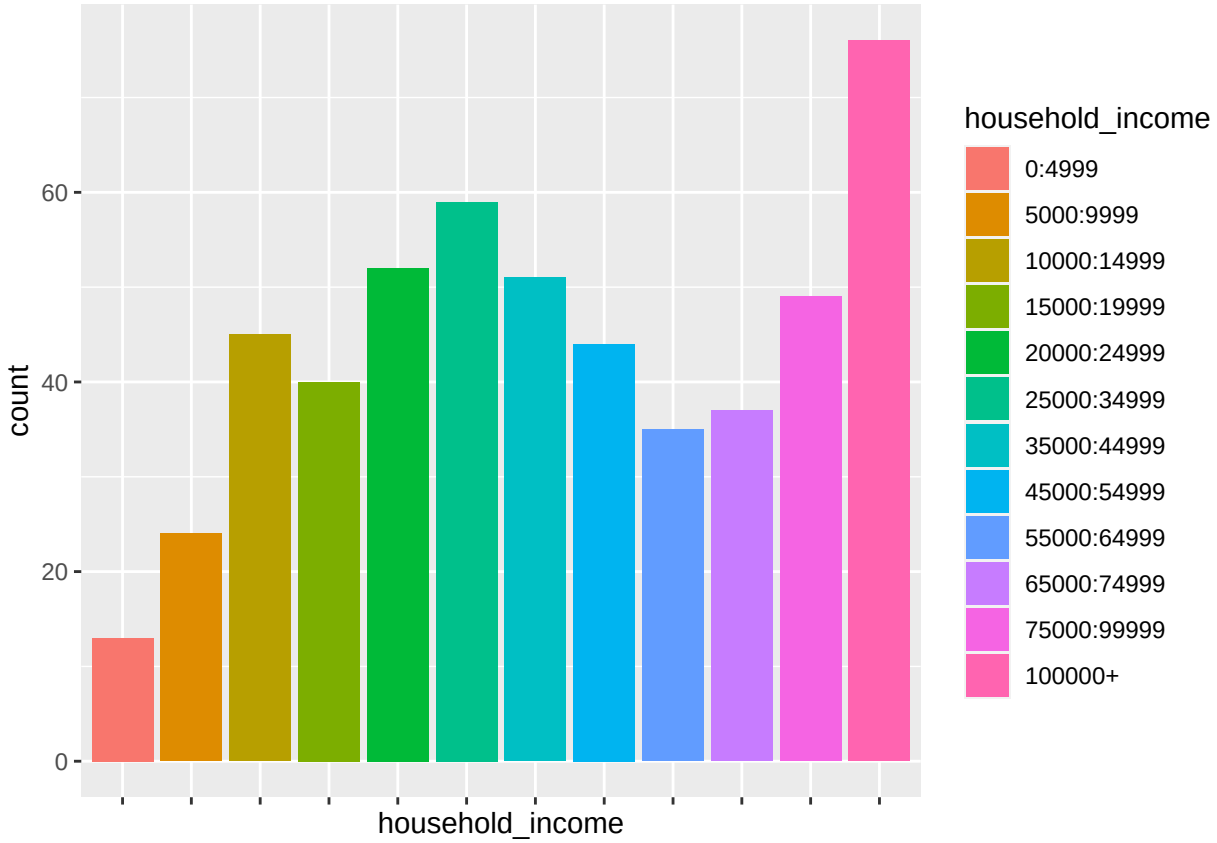

```
ggplot(data, aes(marital, fill = marital)) %>% categorical_dist()
```

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




```
ggplot(data, aes(household_income, fill = household_income)) %>% categorical_dist()
```

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



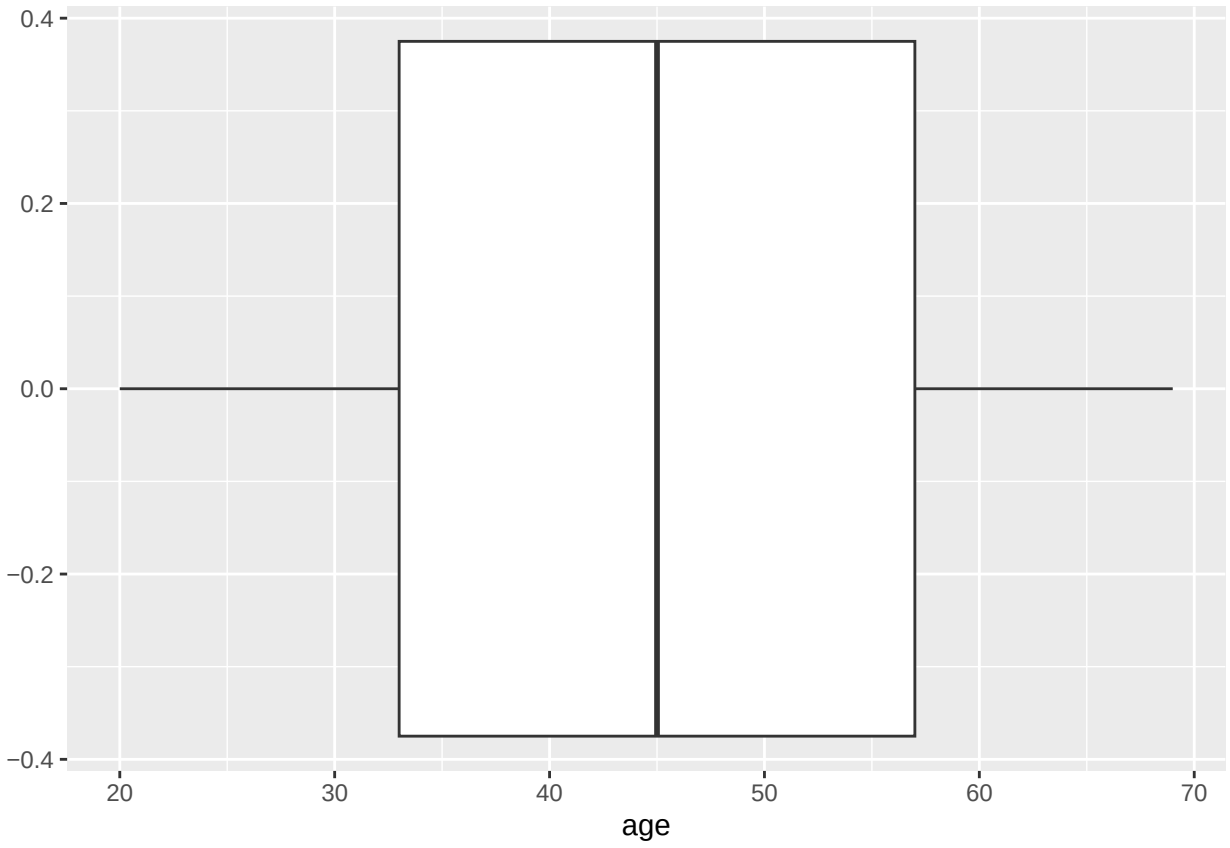

```
# TODO depression data
```

Notes:
- Age is not normally distributed, moreover might be unknowingly missing data < 20 and > 70?
- Missing data in outcome (and depression).
- Lots of married people compared to other marital statuses.

4.3 Outliers

Can only check continuous variables, hence only `age`.

```
ggplot(data, aes(age)) +
  geom_boxplot()
```



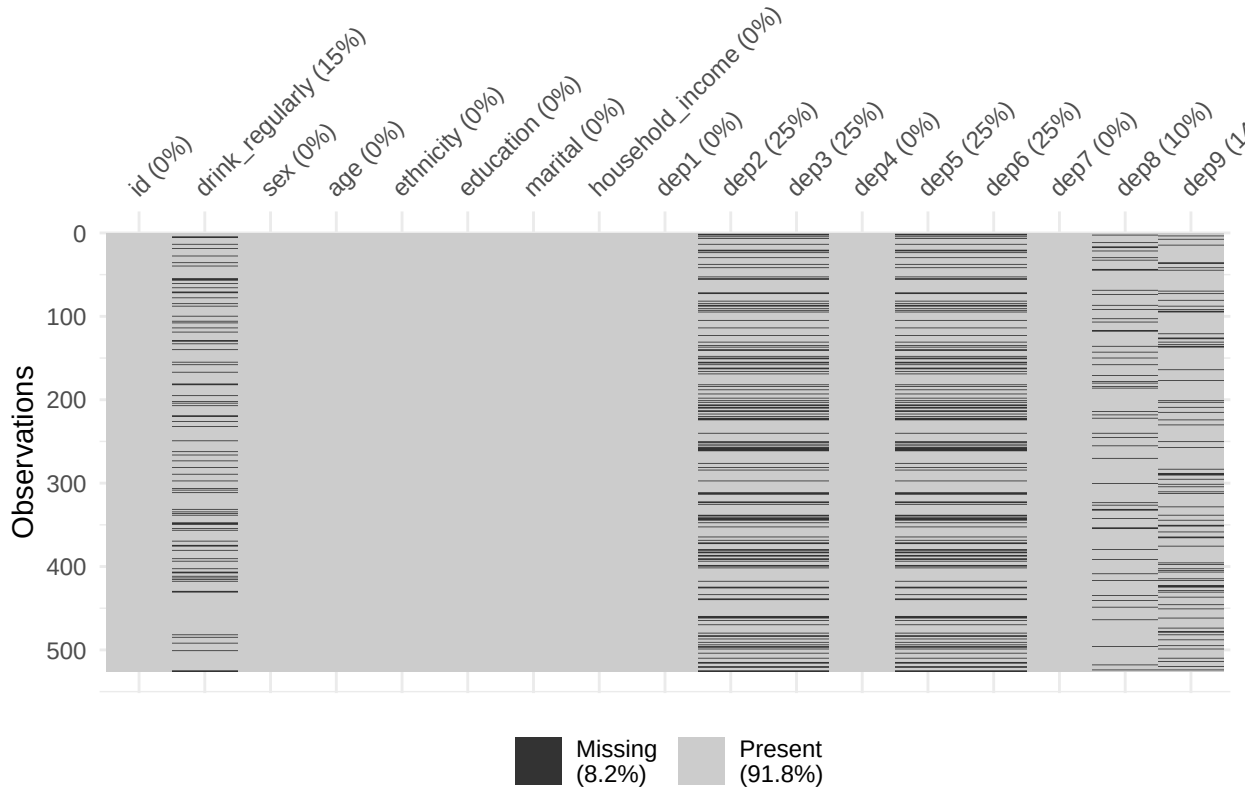
No outliers using IQR.

4.4 Relations

4.5 Missing data and response Patterns

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

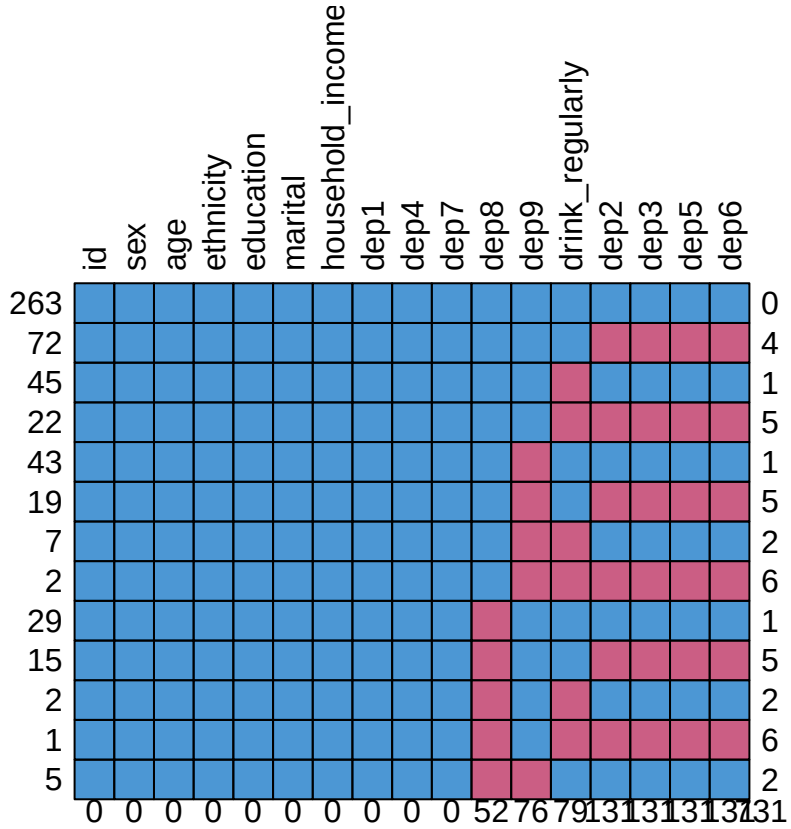
```
# Creates a graph displaying the % of data missing in each variable

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:

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

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




```
##     id sex age ethnicity education marital household_income dep1 dep4 dep7 dep8
## 263  1   1   1         1         1       1                 1    1    1    1    1
## 72   1   1   1         1         1       1                 1    1    1    1    1
## 45   1   1   1         1         1       1                 1    1    1    1    1
## 22   1   1   1         1         1       1                 1    1    1    1    1
## 43   1   1   1         1         1       1                 1    1    1    1    1
## 19   1   1   1         1         1       1                 1    1    1    1    1
## 7    1   1   1         1         1       1                 1    1    1    1    1
## 2    1   1   1         1         1       1                 1    1    1    1    1
## 29   1   1   1         1         1       1                 1    1    1    1    0
## 15   1   1   1         1         1       1                 1    1    1    1    0
## 2    1   1   1         1         1       1                 1    1    1    1    0
## 1    1   1   1         1         1       1                 1    1    1    1    0
## 5    1   1   1         1         1       1                 1    1    1    1    0
##      0   0   0         0         0       0                 0    0    0    0   52
##     dep9 drink_regularly dep2 dep3 dep5 dep6
## 263    1                1    1    1    1    0
## 72     1                1    0    0    0    4
## 45     1                0    1    1    1    1
## 22     1                0    0    0    0    5
## 43     0                1    1    1    1    1
## 19     0                1    0    0    0    5
## 7      0                0    1    1    1    2
## 2      0                0    0    0    0    6
## 29     1                1    1    1    1    1
## 15     1                1    0    0    0    5
```

```
## 2      1               0    1    1    1    1   2
## 1      1               0    0    0    0    0   6
## 5      0               1    1    1    1    1   2
##       76              79  131  131  131  131 731
```

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
```

4.5.0.1 Testing dependency of missing values

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

```
out1$p.value
```

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

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

statistic	df	p.value	missing.patterns
471.1203	164	0	13

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