Missing Data - Assignment 1

Aga, Nisse, Ruben

2024-02-26

Contents

1 Introduction (Ruben) **1**

2 Methodology (Aga) **1**

2.1 Dataset 1

2.2 Data variables 2

2.3 Data proccessing methodology 2

2.4 Model methodology 3

3 EDA Results (Nisse) **3**

3.1 Descriptive statistics 3

3.2 Distributions 6

3.3 Outliers 7

3.4 Relations 8

4 Missing data problem (Aga) **8**

4.1 Missing data and response Patterns 8

4.2 Comparison of the two diffrent models in terms of missing data treatment !!! (Ruben) 12

4.3 Conclusion in terms of answering RQ (Nisse) 12

1 Introduction (Ruben)

1.0.1 RQ (Ruben)

2 Methodology (Aga)

2.1 Dataset

The dataset used is a subset of the data collected in the National Health and Nutrition Examination Survey (NHANES). The survey is a part of annual program that investigates the health and nutrition of a representative sample of people in the United States. The data we used contains information about 525 individuals that has been collected for the NHANES 2007-2008. This is a subset of the 12,946 individuals in that years' survey sample, out of which 78.4% was interviewed and 75.4% was examined in mobile examination centers.

The used dataset contains a wide range of variables related to the health of the individuals. We further subset the data by only including variables relevant to the study (demographics, alcohol use and answers to depression screener questions). The selected variables are further described in Variable Description section.

2.2 Data variables

Table 1: Variable descriptions

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 status	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

Table 1 lists the variables used in our subset selection, which will be utilised for the model in question. The predictor variables [$dep1...dep9$] 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 an Alcohol Use questionnaire targeted at ages 20 and up.

2.3 Data proccessing methodology

- what we investigated and why

2.4 Model methodology

3 EDA Results (Nisse)

3.1 Descriptive statistics

Table 2 shows summary statistics for each of the variables within the dataset; including mean values, standard deviations, IQR statistics and data range values. In general, it is worthy to note that all variables are interpreted as a factor, excluding `age` and the multiple depression levels of `dep`. Table 1 suggests that `dep` should in fact be categorical ordinal and hence also be cast as a factor. That said - and as mentioned in Section 2 - keeping the levels of `dep` as a numerical continuous datatype is beneficial for our missing data problem.

Our outcome variable `drink_regularly` has 307 cases of "yes" and 107 cases of "no", having a outcome balance of 69% and 31% respectively. This outcome ratio could be considered imbalanced, which could affect the accuracy of our logistic regression model. Additionally, the total amount of value entries of 446 suggests that 79 cases contain values outside of the set of possible binary values - possibly being missing values.

The variable `marital` contains 6 categories. Prior to combining the `marital` categories, the category value `married` exceeded the the average frequency of other categories (that being 49.2) by a large margin - as can be seen in Figure 1.

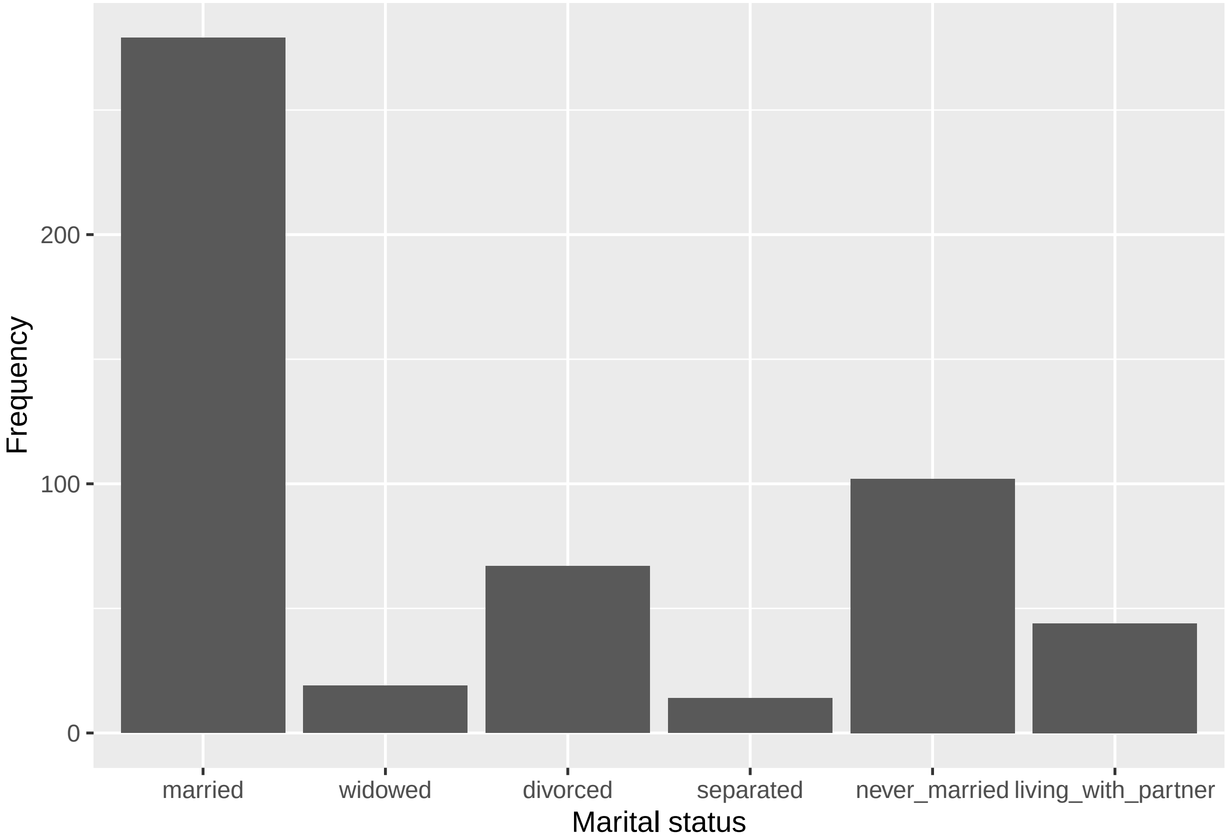


Figure 1: Distribution of marital status categories before pre-processing

Figure 2 shows the frequencies per `marital` category after the pre-processing step of combining the other values into one.

Table 2: Summary statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
id	525	46470	2898	41531	43912	48934	51610
drink_regularly	446						
... yes	307	69%					
... no	139	31%					
sex	525						
... male	254	48%					
... female	271	52%					
age	525	45	14	20	33	57	69
ethnicity	525						
... mexican_american	95	18%					
... other_hispanic	61	12%					
... non-hispanic_white	220	42%					
... non-hispanic_black	124	24%					
... other	25	5%					
education	525						
... no_high_school	58	11%					
... some_high_school	101	19%					
... high_school_grad	123	23%					
... some_college	155	30%					
... college_grad	88	17%					
marital	525						
... married	279	53%					
... widowed	19	4%					
... divorced	67	13%					
... separated	14	3%					
... never_married	102	19%					
... living_with_partner	44	8%					
household_income	525						
... 0:4999	13	2%					
... 5000:9999	24	5%					
... 10000:14999	45	9%					
... 15000:19999	40	8%					
... 20000:24999	52	10%					
... 25000:34999	59	11%					
... 35000:44999	51	10%					
... 45000:54999	44	8%					
... 55000:64999	35	7%					
... 65000:74999	37	7%					
... 75000:99999	49	9%					
... 100000+	76	14%					
dep1	525	0.41	0.79	0	0	1	3
dep2	394	0.28	0.58	0	0	0	3
dep3	394	0.53	0.86	0	0	1	3
dep4	525	0.76	0.9	0	0	1	3
dep5	394	0.31	0.7	0	0	0	3
dep6	394	0.2	0.56	0	0	0	3
dep7	525	0.32	0.71	0	0	0	3
dep8	473	0.2	0.59	0	0	0	3
dep9	449	0.067	0.37	0	0	0	3

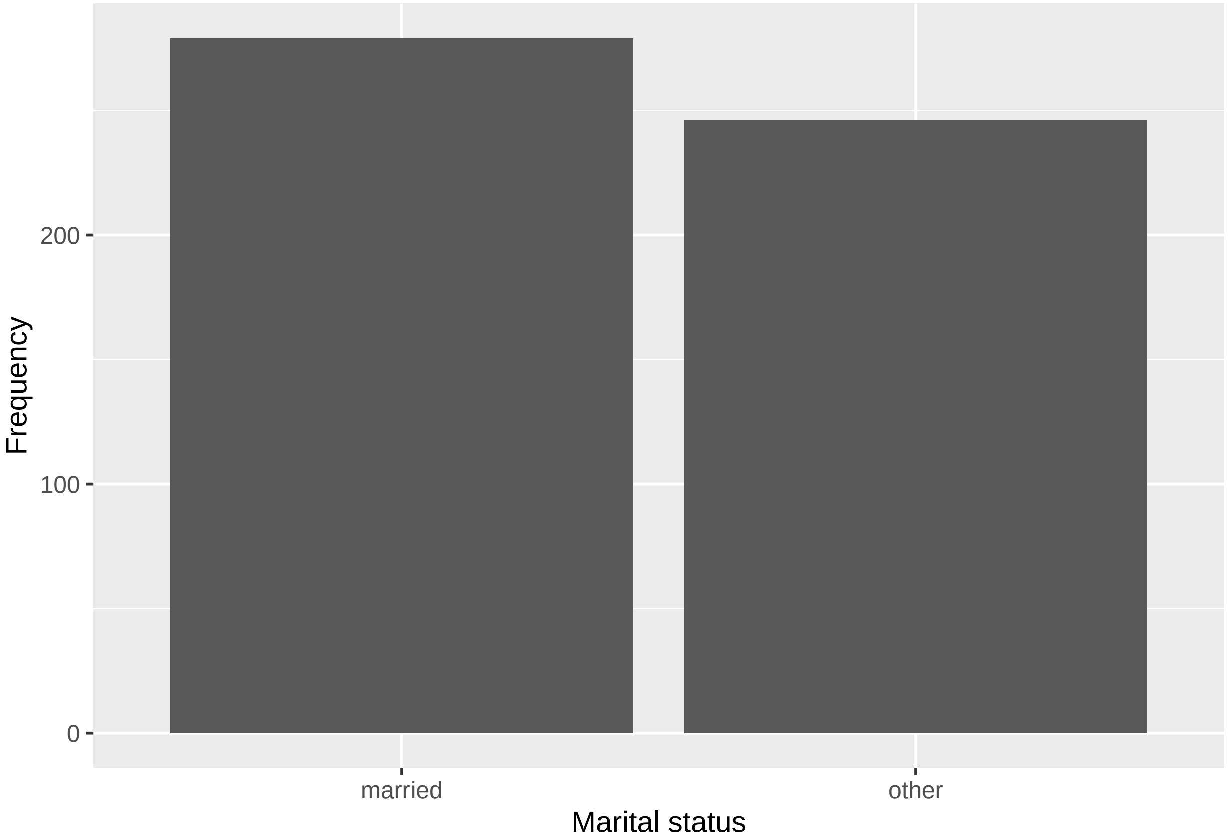


Figure 2: Distribution of marital status categories after pre-processing

Predictor `age` is a dichotomous variable with a balanced frequency distribution across the values “male” and “female”, that being 48% and 52% respectively. 525 entries contain one of these values, suggesting that the variable does not contain missing data.

The predictor `age` is the only continuous variable present within the sub-selected dataset. Although the survey was targeted at respondents of age 0-150 for most variables - the documentation even mentioning the topcoded entries for age 80+ - the dataset seems to only contain cases of people between the age of 20 and 69. Moreover, Figure 3 suggests a uniform distribution of the age variable.

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

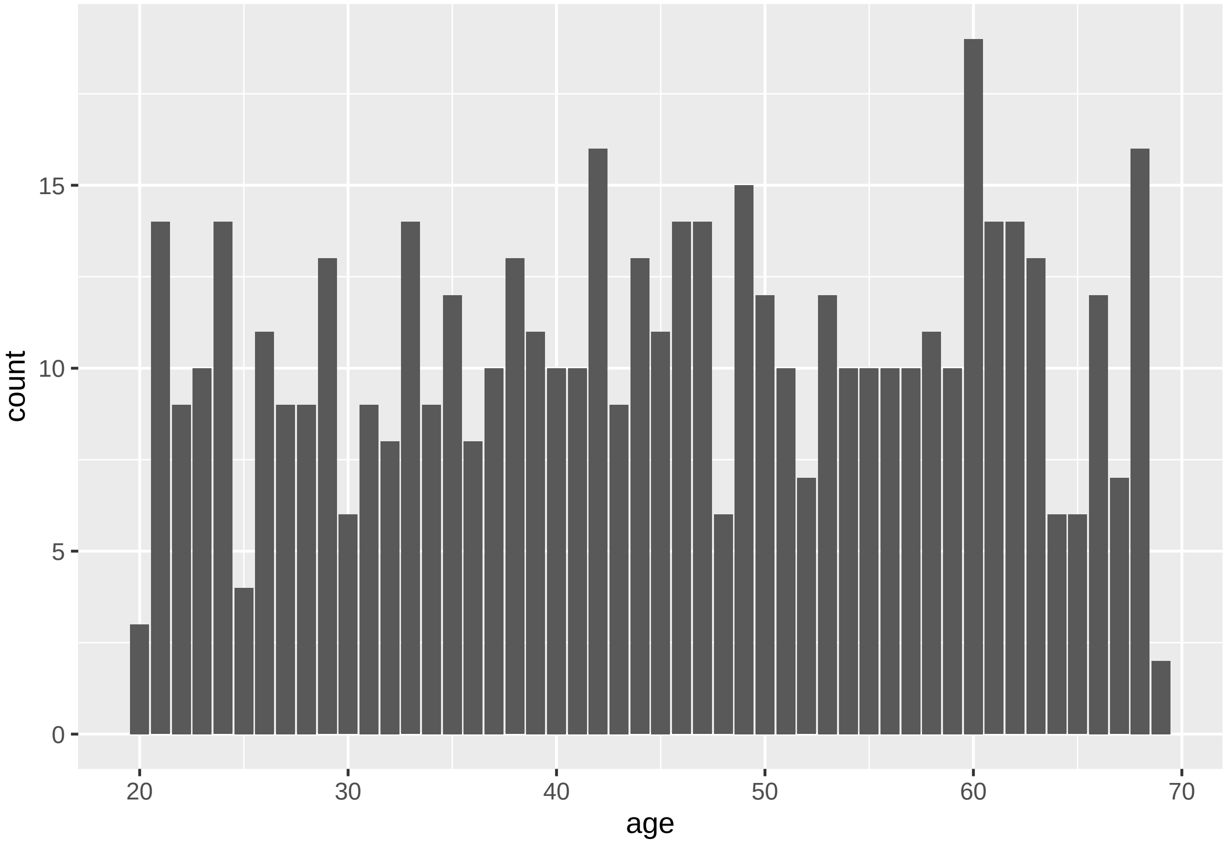


Figure 3: Distribution of age

3.2 Distributions

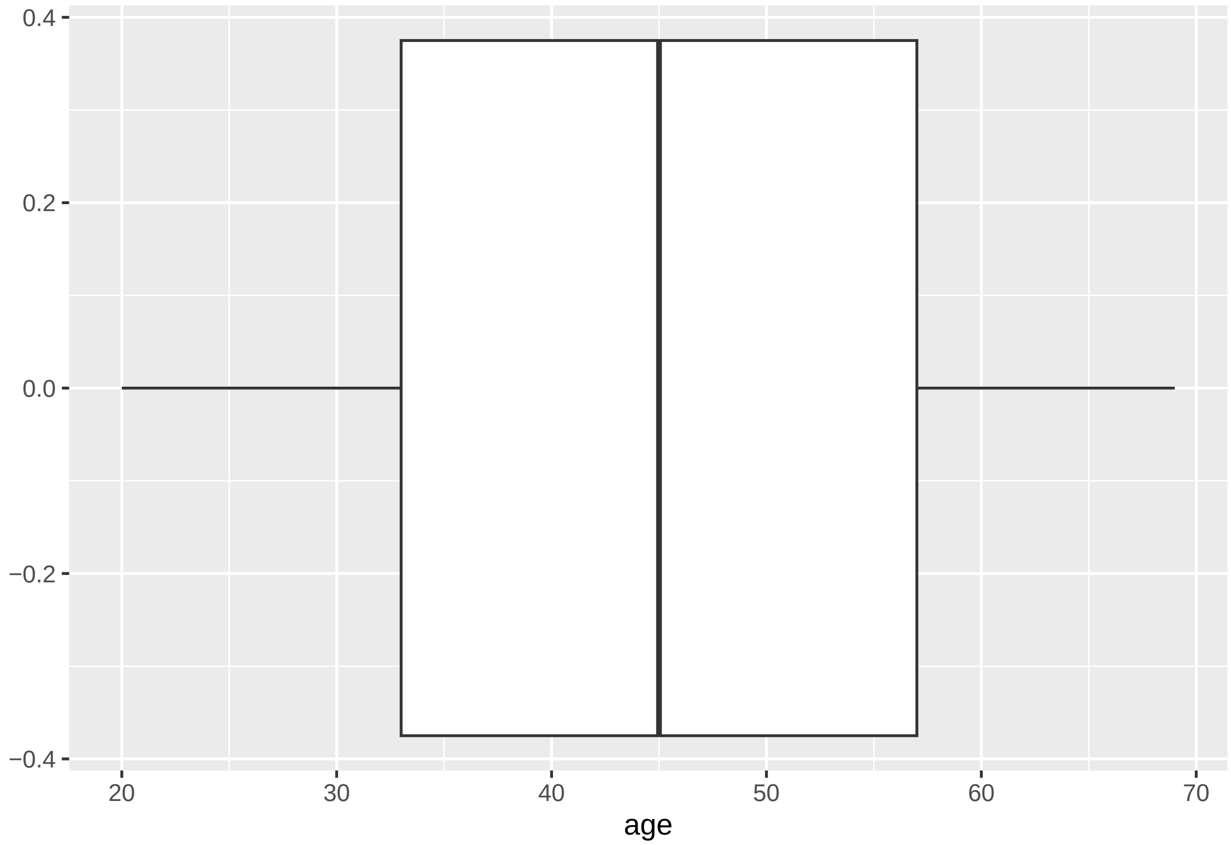
```
# Continuous
ggplot(data, aes(age)) + geom_histogram(stat = 'count')

# 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()
ggplot(data, aes(sex, fill = sex)) %>% categorical_dist()
ggplot(data, aes(ethnicity, fill = ethnicity)) %>% categorical_dist()
ggplot(data, aes(education, fill = education)) %>% categorical_dist()
ggplot(data, aes(marital, fill = marital)) %>% categorical_dist()
ggplot(data, aes(household_income, fill = household_income)) %>% categorical_dist()

# 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.

3.3 Outliers

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

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



No outliers using IQR.

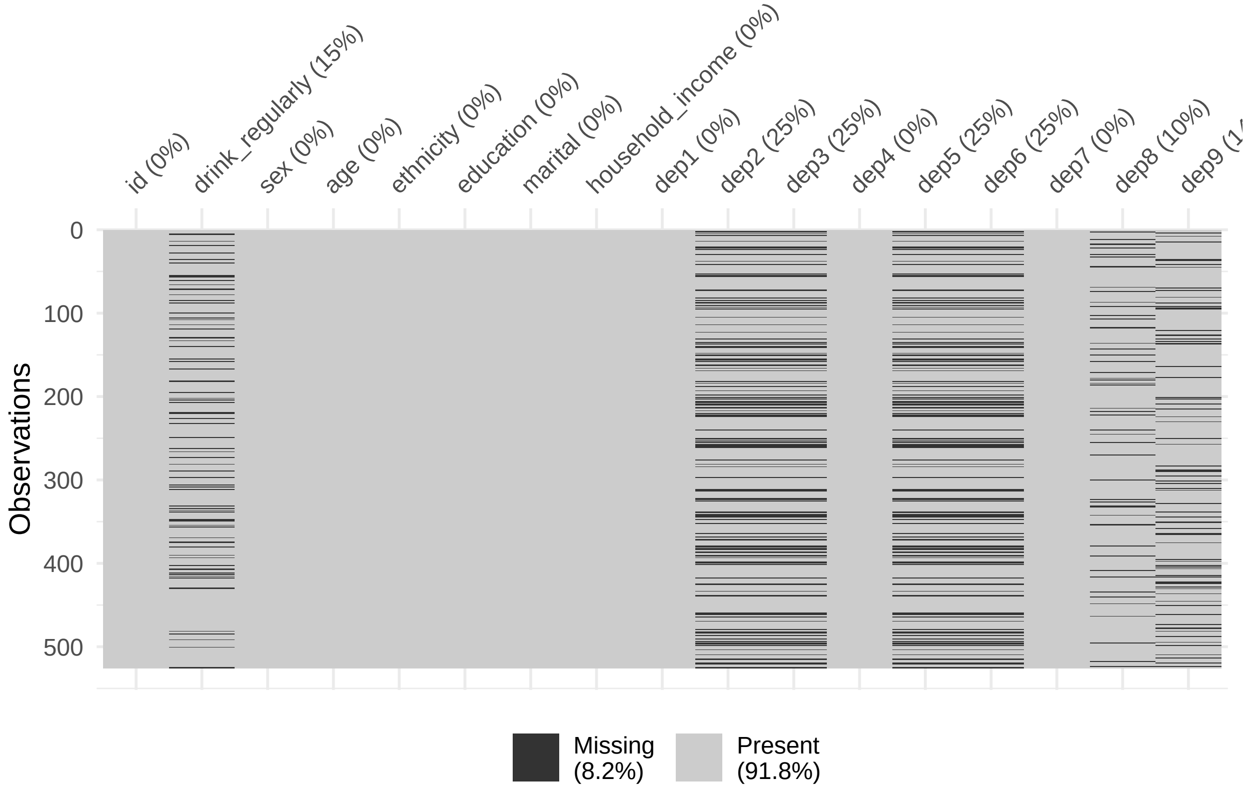
3.4 Relations

4 Missing data problem (Aga)

4.1 Missing data and response Patterns

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

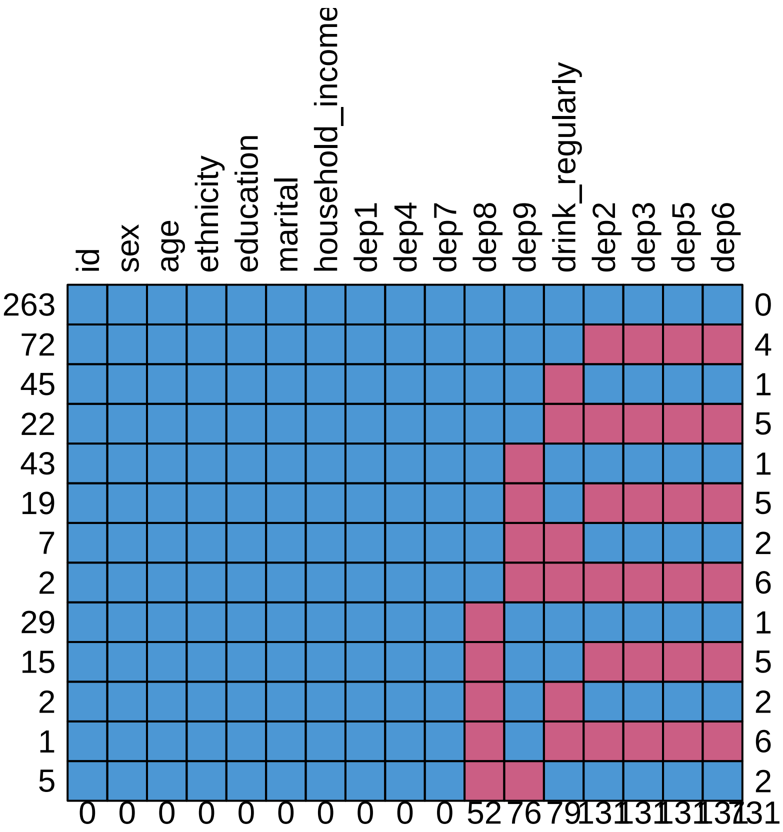
```
# Creates a graph displaying the % of data missing in each variable

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    1   0
## 72     1               1    0    0    0    0   4
## 45     1               0    1    1    1    1   1
## 22     1               0    0    0    0    0   5
## 43     0               1    1    1    1    1   1
## 19     0               1    0    0    0    0   5
## 7      0               0    1    1    1    1   2
## 2      0               0    0    0    0    0   6
## 29     1               1    1    1    1    1   1
## 15     1               1    0    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
```

4.1.0.1 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)
```

```
## # A tibble: 1 x 4
##   statistic    df p.value missing.patterns
##       <dbl> <dbl>   <dbl>            <int>
## 1      465.   164       0               13
```

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

- what's the missing data mechnism?

4.1.1 Result models with deletion and imputation (Nisse)

- formula
- table with coefficients and pval (make sure to exponential the coefficients for easier interpretation)
- Interpretation of model result

```
miceOut <- mice(data, defaultMethod = c("norm.predict", "logreg", "polyreg", "polr"), m = 1, maxit = 1)
```

```
##
##  iter imp variable
##   1   1  drink_regularly  dep2  dep3  dep5  dep6  dep8  dep9
```

```
## Warning: Number of logged events: 1
```

```
reg_imp_data <- complete(miceOut)
summary(reg_imp_data)
```

```
##        id        drink_regularly     sex           age
##  Min.   :41531   yes:364         male  :254   Min.   :20.00
##  1st Qu.:43912   no :161         female:271   1st Qu.:33.00
##  Median :46357                                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   Length:525
##  other_hispanic    : 61   some_high_school:101   Class :character
##  non-hispanic_white:220   high_school_grad:123   Mode  :character
##  non-hispanic_black:124   some_college    :155
##  other             : 25   college_grad    : 88
##
##
##    household_income      dep1             dep2              dep3
##  100000+    : 76     Min.   :0.0000   Min.   :-0.1803   Min.   :-0.20715
##  25000:34999: 59     1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.00000
##  20000:24999: 52     Median :0.0000   Median : 0.0000   Median : 0.06551
##  35000:44999: 51     Mean   :0.4095   Mean   : 0.3526   Mean   : 0.67666
##  75000:99999: 49     3rd Qu.:1.0000   3rd Qu.: 0.6856   3rd Qu.: 1.00000
##  10000:14999: 45     Max.   :3.0000   Max.   : 3.0000   Max.   : 3.00000
##  (Other)    :193
##       dep4             dep5              dep6              dep7
##  Min.   :0.0000   Min.   :-0.1840   Min.   :-0.1752   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   : 0.3522   Mean   : 0.3282   Mean   :0.3238
##  3rd Qu.:1.0000   3rd Qu.: 0.5922   3rd Qu.: 0.4620   3rd Qu.:0.0000
##  Max.   :3.0000   Max.   : 3.0000   Max.   : 3.4502   Max.   :3.0000
##
##       dep8              dep9
##  Min.   :-0.2302   Min.   :-0.35994
##  1st Qu.: 0.0000   1st Qu.: 0.00000
##  Median : 0.0000   Median : 0.00000
##  Mean   : 0.2046   Mean   : 0.06599
##  3rd Qu.: 0.0000   3rd Qu.: 0.00000
##  Max.   : 3.0000   Max.   : 3.00000
##
```

4.2 Comparison of the two diffrent models in terms of missing data treatment !!! (Ruben)

4.3 Conclusion in terms of answering RQ (Nisse)