Missing Data - Assignment 1

Aga, Nisse, Ruben

2024-02-27

Contents

1 Introduction (Ruben) **1**

2 Methodology (Aga) **2**

2.1 Dataset 2

2.2 Variables description 2

3 EDA Results (Nisse) **3**

3.1 Descriptive statistics 3

3.2 Distributions 4

3.3 Outliers 11

3.4 Relations 12

4 Missing data problem (Aga) **12**

4.1 Missing data and response Patterns 12

4.2 Comparison of the two diffrent models in terms of missing data treatment !!! (Ruben) 16

4.3 Conclusion in terms of answering RQ (Nisse) 16

1 Introduction (Ruben)

1.0.1 RQ (Ruben)

Alcohol consumption led to 2.8 million deaths in 2016 and accounts for almost 10% of global deaths in people aged 15-49 years. Higher levels of alcohol consumption leads to a higher risk of mortality and the only level of alcohol consumption that minimizes this risk is zero alcohol consumption (study). This means that drinking any alcohol at all increases the risk of mortality. All of this makes it vital to know and understand the factors behind alcohol consumption.

Moore et al. 2005 found that age, sex, ethnicity, marital status, education level and household income were all either negatively or positively associated with alcohol consumption. In Garnett et all. 2022 it was found that the level of depression was negatively associated with alcohol consumption. Considering that the only safe level of alcohol consumption is zero, it is important to understand what factors predict regular consumption of alcohol instead of only looking at the amount of alcohol consumption as in the two studies mentioned above.

The research question of this study is whether the occurrence of regular alcohol consumption (12 or more in a year) can be predicted by the variables: depression level, age, sex, ethnicity (white vs other), marital status and household income. It is expected that age and depression level will be negatively correlated with the occurrence of alcohol consumption while household income will be positively correlated (sources). It is also expected that being male (vs female) and white (vs other ethnicities) will be positively correlated with the regular occurrence of alcohol consumption (source). Regarding marital status it is expected that the married status will be positively associated with the regular occurrence of alcohol consumption compared to others but that is based on a study where the factor marital status consisted of just married and other (source), while in this study more categories are considered.

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 NHANES is further subdivided into themed sections, such as the Alcohol Use questionnaire, that have separate documentation that will be refered to later.

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 Variables description

The table below lists the variables used in our subset selection, which will be utilized 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. These questions are intended as components of an overall Depression Score that will be created by summing up the individual values.
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. It is likely due to the fact that it intends to register the "final" level of education and participats would not be likely to accomplish `AA degree` or `College Graduate` earlier.
- 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.

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

2.2.1 Data proccessing methodology

After arriving at the subset dataset, as Exploratory Data Analysis was performed. The distributions of the variables were investigated and presence of missing data was found. The important findings of this step are presented in 'EDA Results' section. It is important to note, that after the uneven distribution in 'marital' categories was discovered and the relationship of the categories with other variables was investigated, it was decided to combine the not-married categories into one. Thus, a less complex, binary variable was created.

Following the EDA, the missing data problem was addressed. The extend, distribution and patterns of missingness were investigated.

2.2.2 Model methodology

3 EDA Results (Nisse)

3.1 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
```

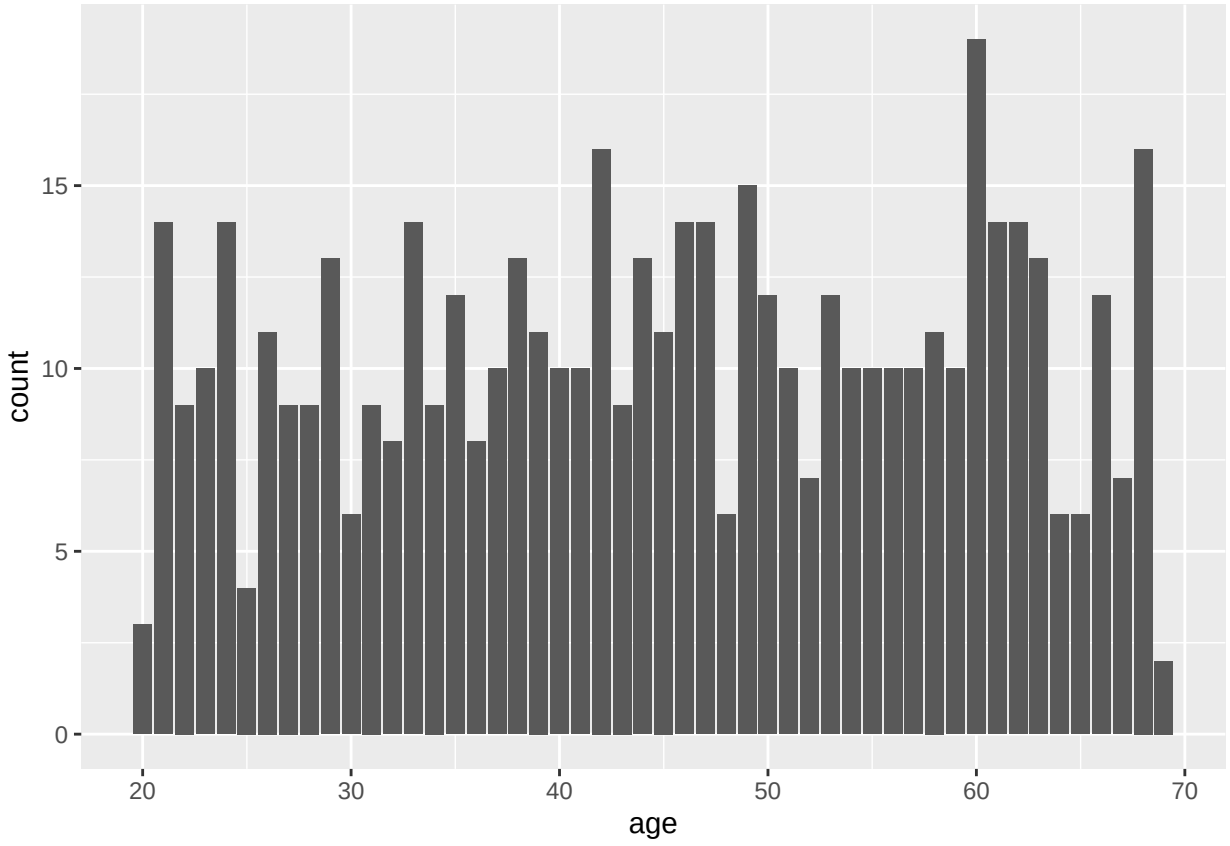
```r
n_rows <- n_distinct(data$id)
```

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

3.2 Distributions

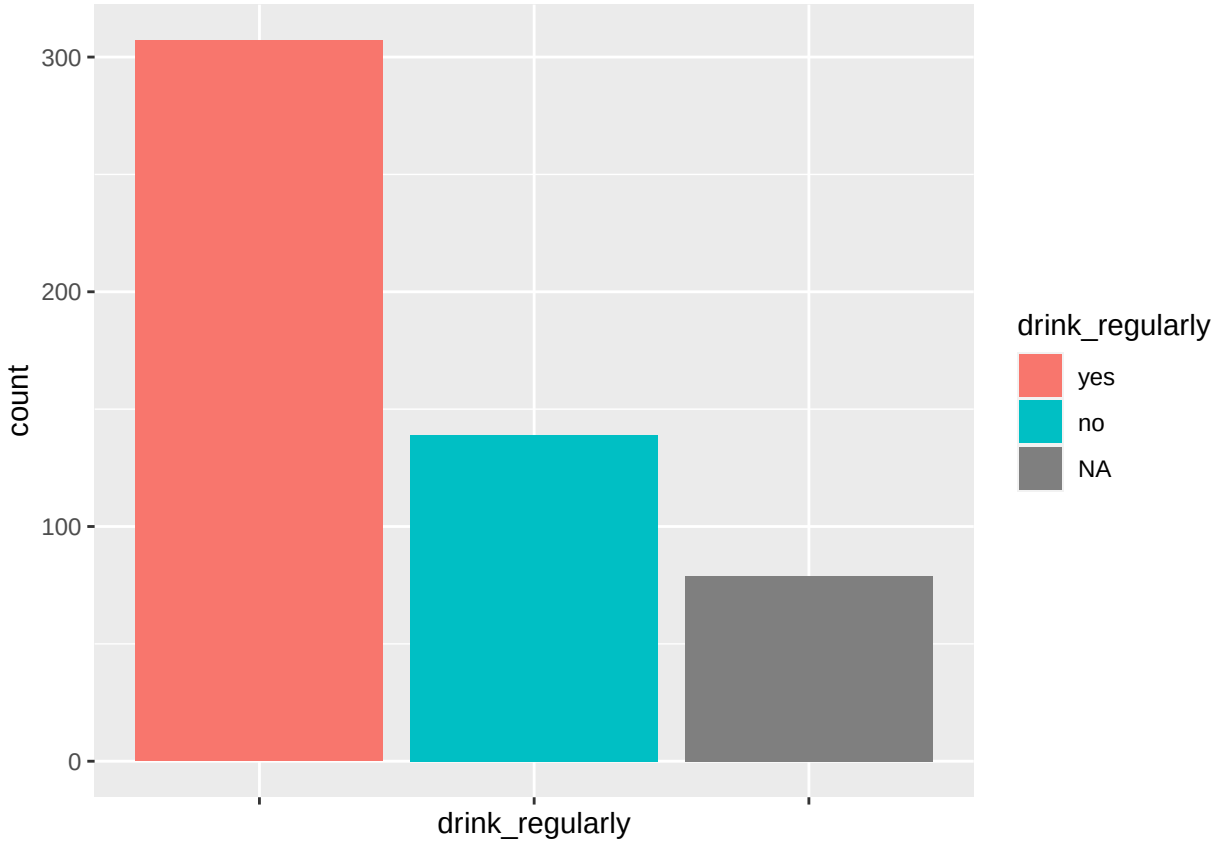
```r
# Continuous
ggplot(data, aes(age)) + geom_histogram(stat = 'count')
```

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



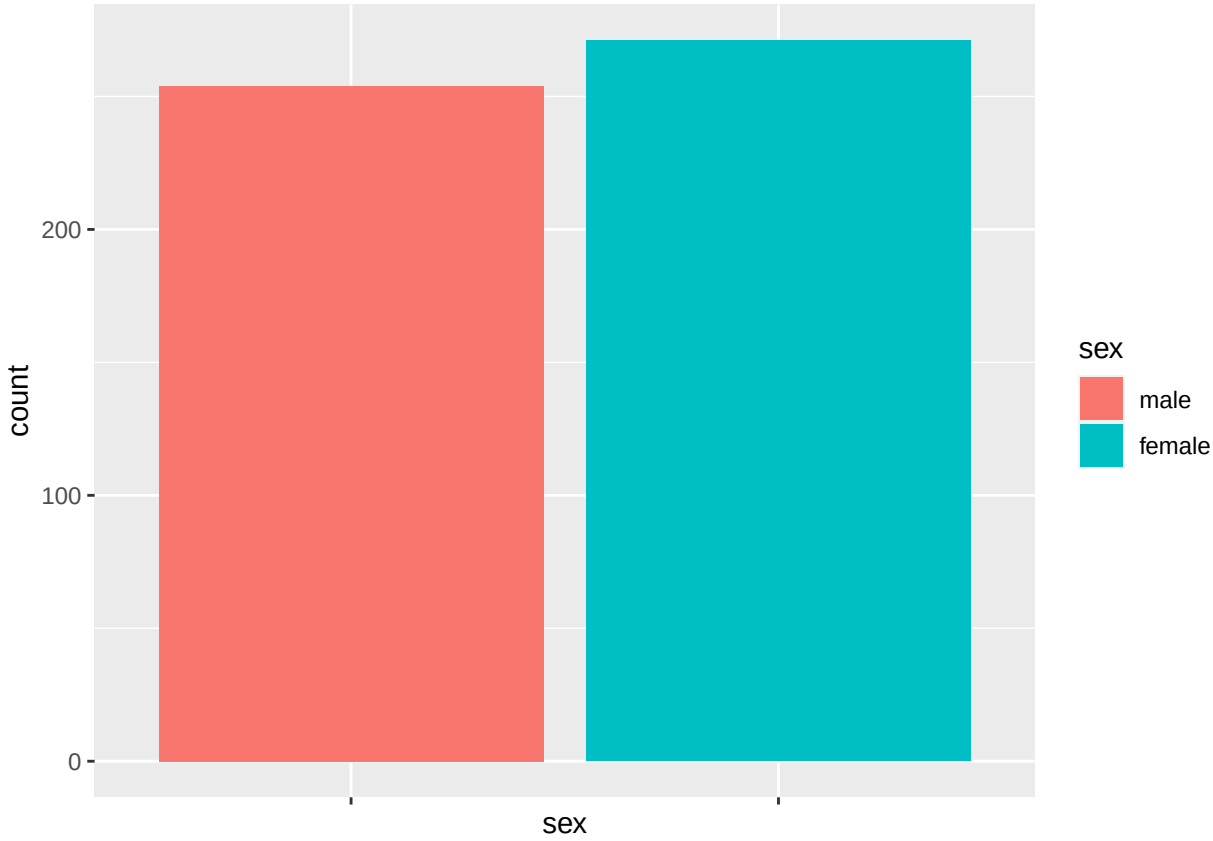

```
# Categorical
categorical_dist <- function(plot) {
  plot +
    geom_histogram(stat = 'count') +
     theme(axis.text.x = element_blank())
}

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

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



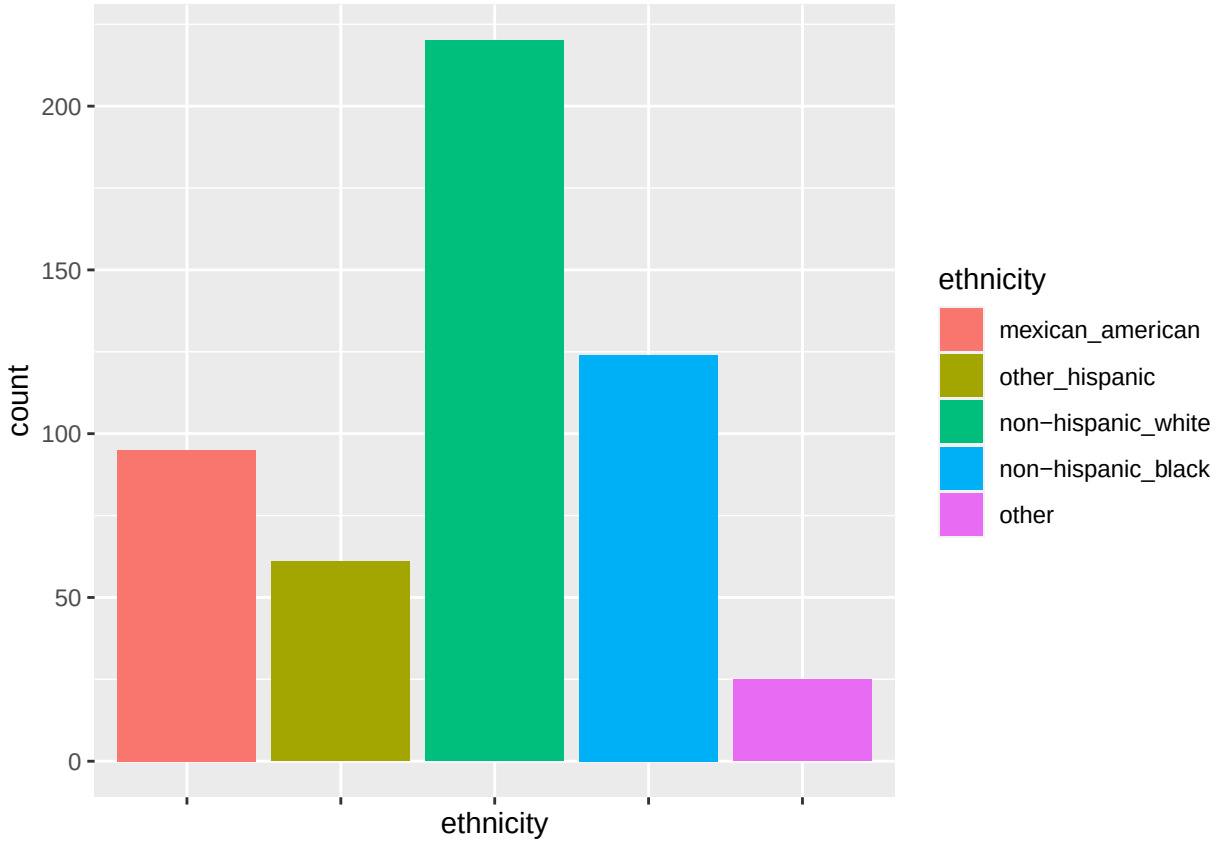

```
ggplot(data, aes(sex, fill = sex)) %>% categorical_dist()
```

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



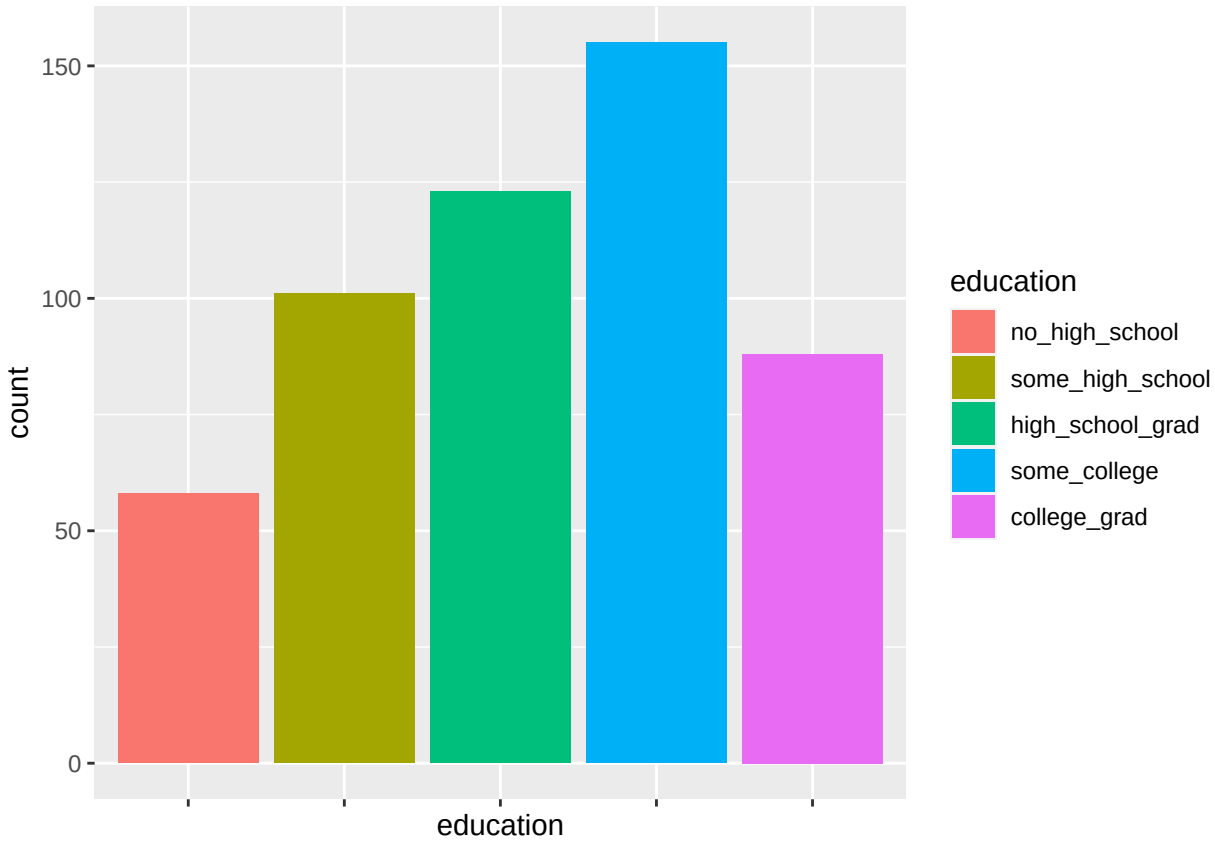

```
ggplot(data, aes(ethnicity, fill = ethnicity)) %>% categorical_dist()
```

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



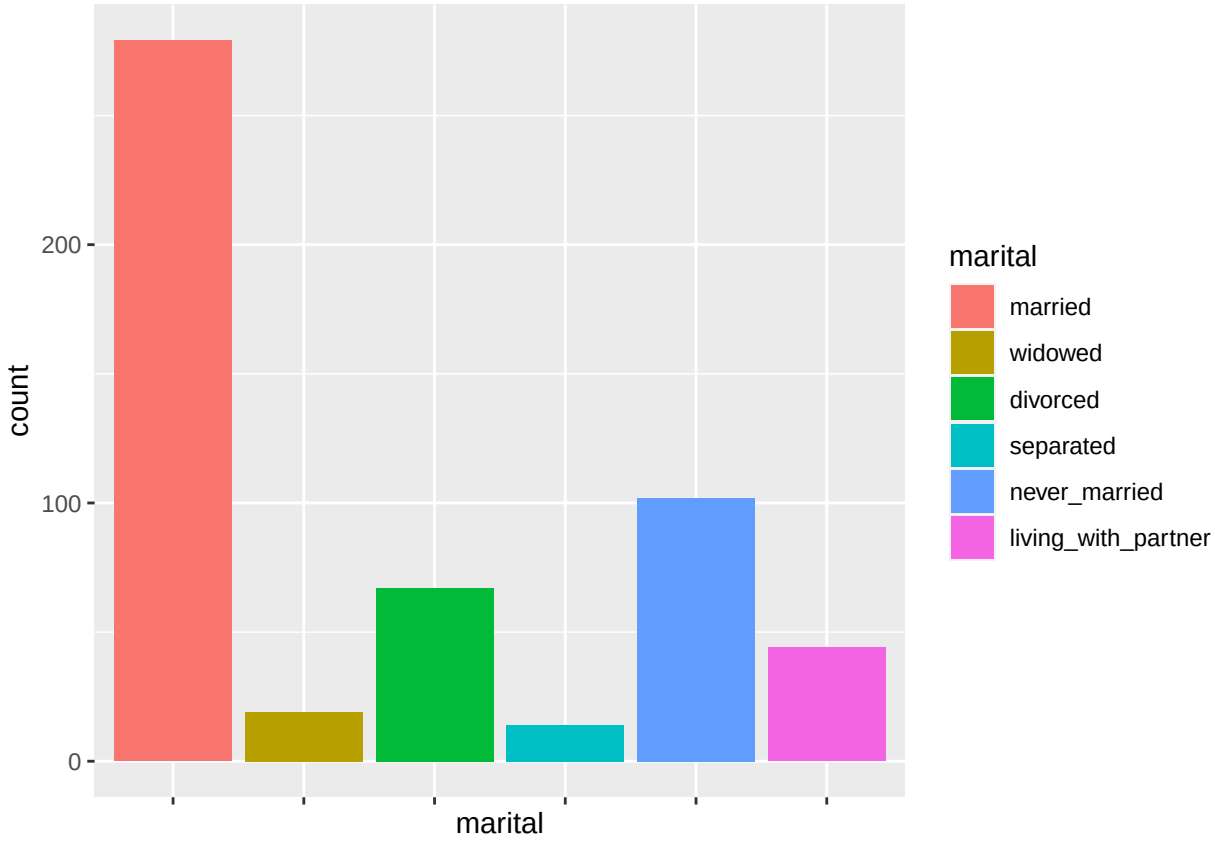

```
ggplot(data, aes(education, fill = education)) %>% categorical_dist()
```

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



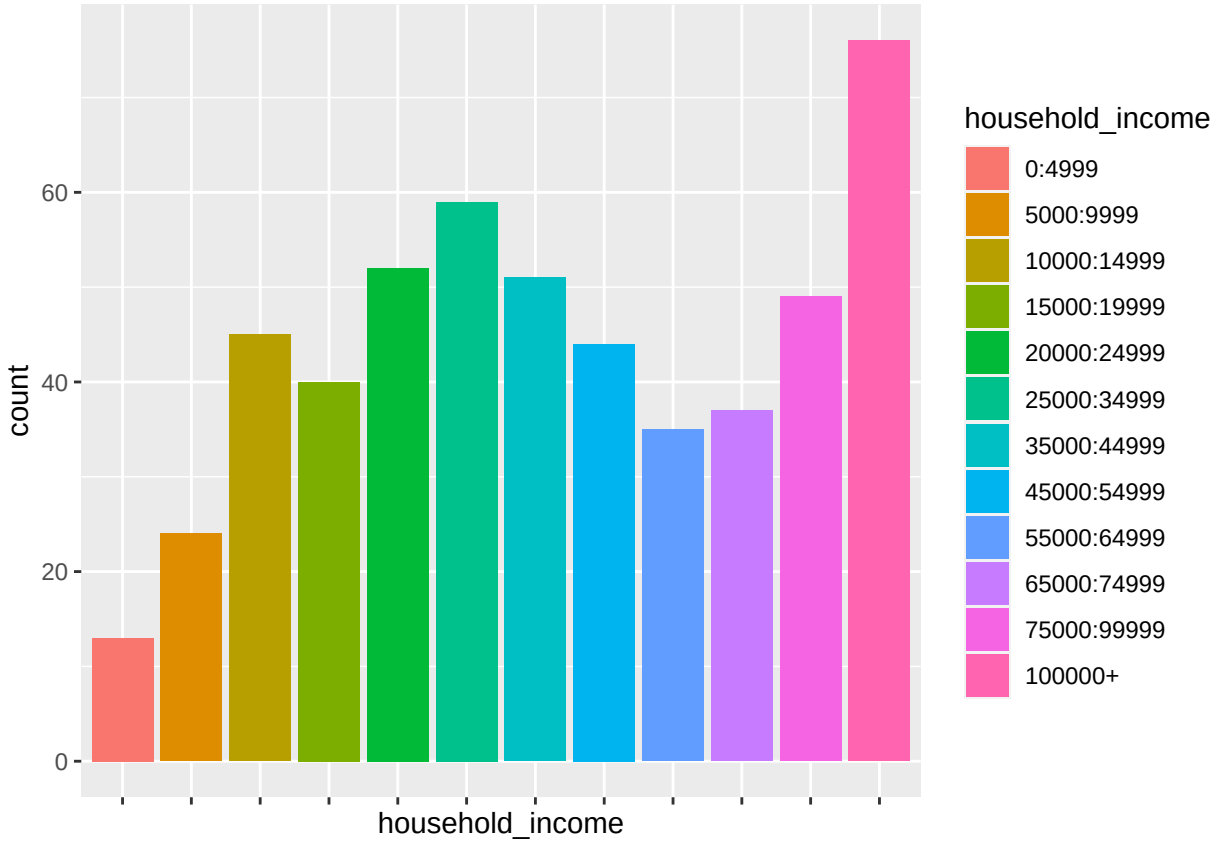

```
ggplot(data, aes(marital, fill = marital)) %>% categorical_dist()
```

```
## 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`
```



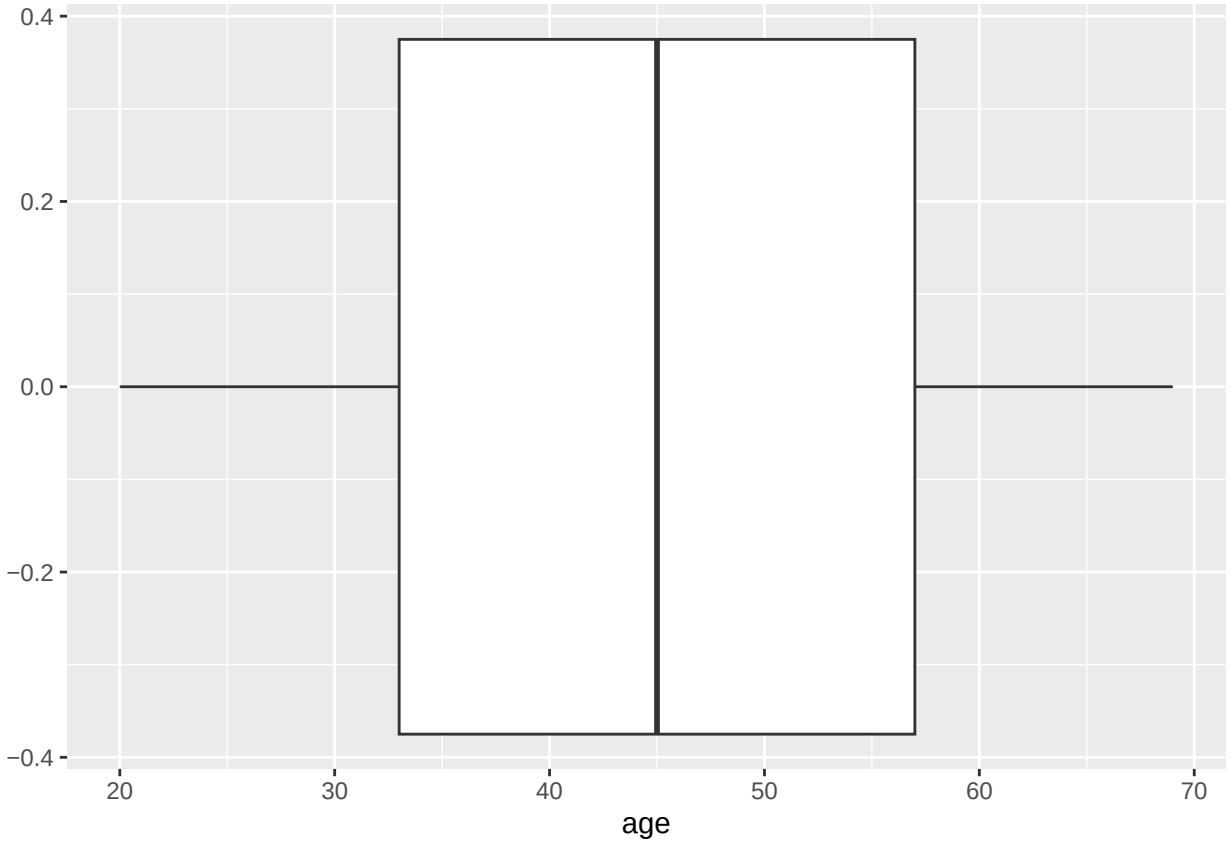

```
# TODO depression data
```

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.

3.3 Outliers

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

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



No outliers using IQR.

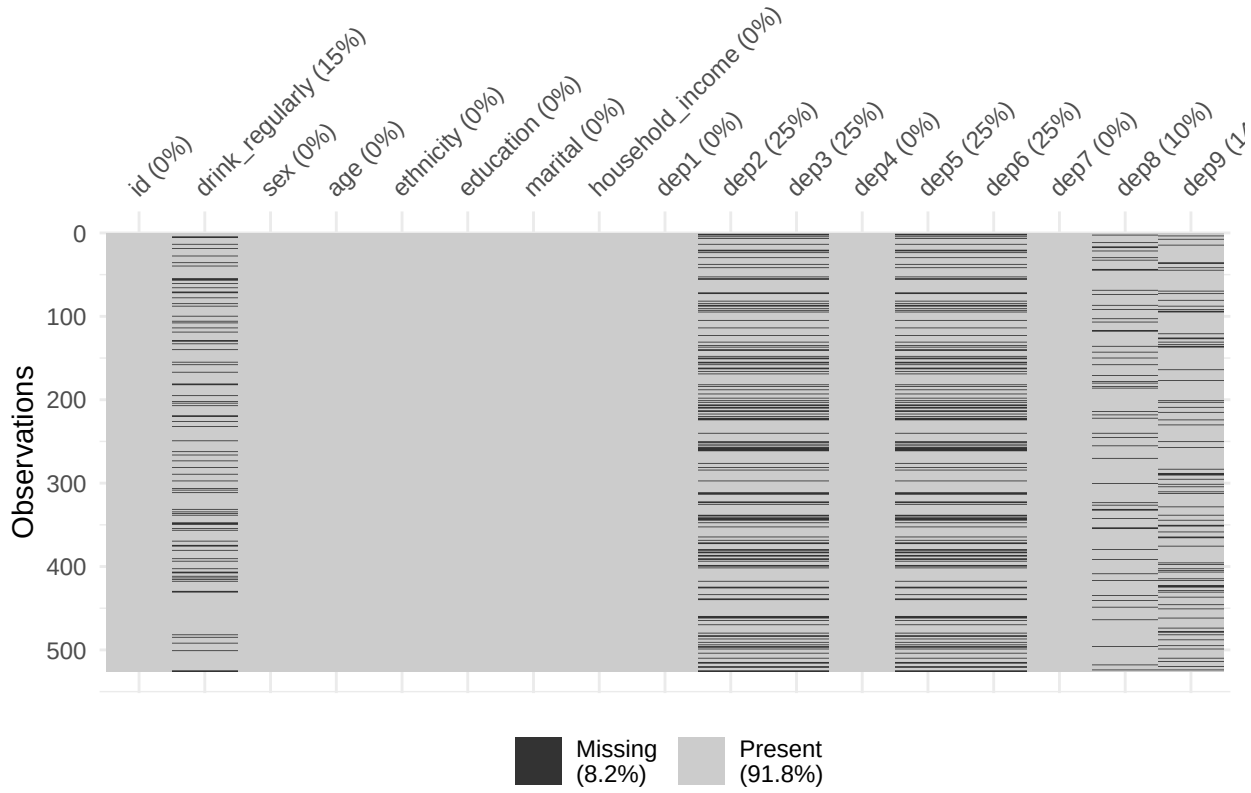
3.4 Relations

4 Missing data problem (Aga)

4.1 Missing data and response Patterns

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

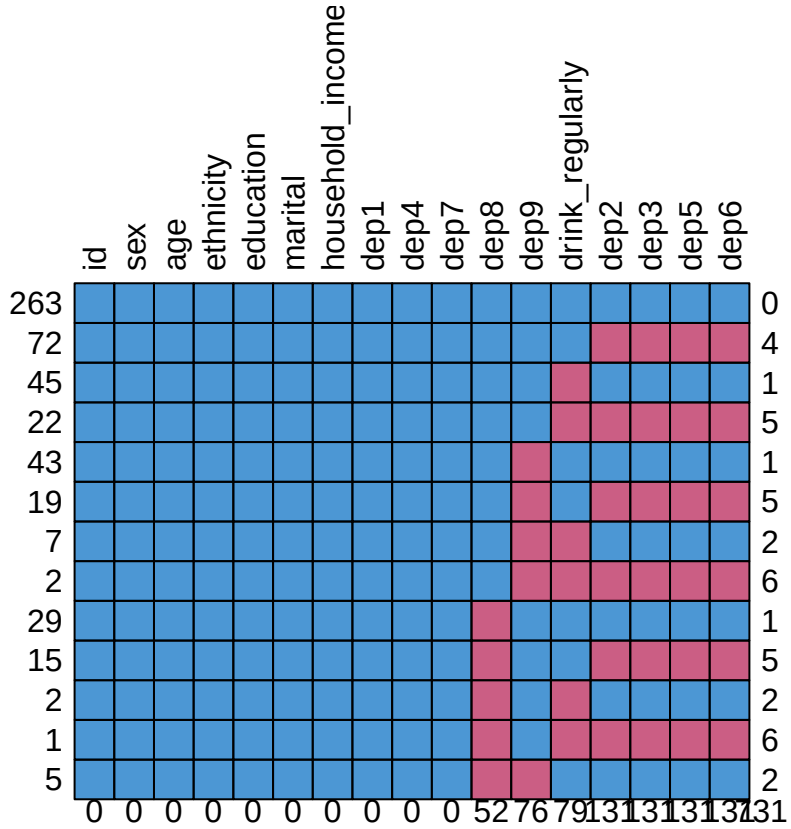
```
# Creates a graph displaying the % of data missing in each variable

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    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
```

4.1.0.1 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)
```

statistic	df	p.value	missing.patterns
471.1203	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
```

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

```
##        id        drink_regularly     sex           age
##  Min.   :41531   yes:363         male  :254   Min.   :20.00
##  1st Qu.:43912   no :162         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   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.   :-0.3354   Min.   :-0.09603
##  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 : 0.12476
##  35000:44999: 51     Mean   :0.4095   Mean   : 0.3531   Mean   : 0.69126
##  75000:99999: 49     3rd Qu.:1.0000   3rd Qu.: 0.7310   3rd Qu.: 1.00000
##  10000:14999: 45     Max.   :3.0000   Max.   : 3.0000   Max.   : 3.24036
##  (Other)    :193
##       dep4             dep5              dep6              dep7
##  Min.   :0.0000   Min.   :-0.5062   Min.   :-0.1410   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   : 0.3507   Mean   : 0.3230   Mean   :0.3238
##  3rd Qu.:1.0000   3rd Qu.: 0.5616   3rd Qu.: 0.5031   3rd Qu.:0.0000
##  Max.   :3.0000   Max.   : 3.0000   Max.   : 3.0000   Max.   :3.0000
##
##       dep8              dep9
##  Min.   :-0.2060   Min.   :-0.31357
##  1st Qu.: 0.0000   1st Qu.: 0.00000
##  Median : 0.0000   Median : 0.00000
##  Mean   : 0.2069   Mean   : 0.06586
##  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)