Missing Data - Assignment 1

Aga, Nisse, Ruben | Group 9

2024-02-29

Contents

1 Introduction

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: **To what extent the occurrence of regular alcohol consumption (12 or more in a year) can be predicted by the variables: depression level, age, sex, ethnicity, 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 referred 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 Variables Description section.

2.2 Variables Description

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. Multiple signs of depression were measured this way, which can later be combined to indicate an overall level of depression.

The demographic variables - that being `sex`, `age`, `ethnicity`, `education` and `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.

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

- 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 `income` is ordinal, rather than continuous.

As for the remaining demographic variables, namely `sex`, `age`, `ethnicity` and `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 Software

All of the data cleaning, processing and modeling was performed in R version 4.3.2 (R Core Team 2023). The following packages were used: tidyverse was used for data manipulation (Wickham et al. 2019), the mice (van Buuren and Groothuis-Oudshoorn 2011) and ggmice (Oberman 2024) packaged were used for missing data visualizations and list-wise deletion and mean imputation, the kableExtra pacckage was used for table styling (Zhu 2021), naniar package was used to make the visualization of missing data percentages (Figure 7) (Tierney and Cook 2023), Hmisc was utilized for mode imputation and vtable helped with the summary statistic table in the Appendix (Huntington-Klein 2023) and x was used to produce the summary tables for the two final models (Sjoberg et al. 2021).

2.4 Data processing

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 in the outcome variable and some of the depression questions. No outliers were found. Additionally, the missingness seems to be more prevalent among younger individuals, which gave a first indication of lack of MCAR mechanism. The important findings of this step are presented in 'EDA Results' section.

Following the EDA, the missing data problem was addressed. The extend, distribution and patterns of missingness were investigated. Testing was done to verify the missing data mechanism, including performing

a global test - Little MCAR test (Little 1986) and testing of the dependencies between missing values in one variable and observed values of other variables using t-test for continuous variables, Chi-squared test and Fisher test for categorical variables (Soetewey, n.d.). The Fisher test with simulated pValues was used to verify outcomes for combinations of variables that included an expected frequency of below 5, thus not meeting the assumptions of Chi-squared test (Soetewey, n.d.). When the assumptions of MCAR was not meet, MAR was assumed.

Two solutions to the missing data problem were implemented: list-wise deletion and mean imputation. These where chosen based on the convenience of implementation. For list-wise deletion all cases with one or more missing values were excluded from the modeling. The mean imputation method was implemented in the following way: For the missing depression questions the mean of the observed values within the variable was computed and then the missing values were replaces by the mean value. This approach was taken, as in case of ordinal variables on Likert scale treating them as continuous is an appropriate (Wu, Jia, and Enders 2015). For 'drink_regularly, the outcome variable, instead of a mean, the mode was computed. This was motivated by the need for a binary outcome variable in the logistic regression that was planned to answer the research question. Since the EDA found there was significantly more 'yes' values than 'no' values in 'drink_regularly', the missing values were replaces by 'yes'. Thus, two complete datasets with different missing data treatments were created.

Lastly, following the missing data treatments, the values for the depression questions were summed up to create an overall depression score for each individual. A sum was used, as opposed to other methods of aggregation (sucha as mean), as it preserves a convenient interpretation of a unit increase in a depression score. This overall score was used for modeling as opposed to the individual questions.

2.5 Modelling methodology

Including all of the variables in the two complete data sets is motivated by theoretical findings since other researchers found a relationship between them and drinking habits. Therefore, verifying the significance of these predictors and their relative importance in the presence of other variables was important. Thus, a decision was made to create two logistical models including all of the variables, instead of a top down or bottom up approach to model building. As opposed to removing the non-significant predictors as in the other approaches, it was decided to keep them and therefore have more complex models. The two logistics models were then compared and the impact of the two different missing data treatments was evaluated. For easier interpretation the model coefficients were exponentiated. Occurence of regular drinking was considered the "successful" outcome.

3 EDA Results (Nisse)

3.1 Descriptive statistics

Table 7 from the appendix shows summary statistics for each of the variables within the dataset; including mean values, standard deviations, IQR statistics and data range values. For categorical variables, a list of possible categories and their respective proportions is provided to replace the continuous summary statistics. Lastly, the column `N` shows the amount of cases with present data - that being non-NA 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 should therefore be cast as a factor. That said - and as mentioned in Section **??** - keeping the levels of `dep` as a numerical continuous datatype is beneficial for our missing data problem.

The dataset contains a total of 525 observations, each with 17 variables.

Our dichotomous 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 (`N`) of 446 suggests that 79 cases contain values outside of the set of possible binary values - most likely being missing values. Table 2 further confirms this.

Table 2: Drink regularly value distributions

drink_regularly	n
yes	307
no	139
NA	79

Predictor `sex` 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 1 suggests a uniform distribution of the age variable. Like `sex`, `age` has 525 cases with non-NA values, hence the variable does not contain missing data values.

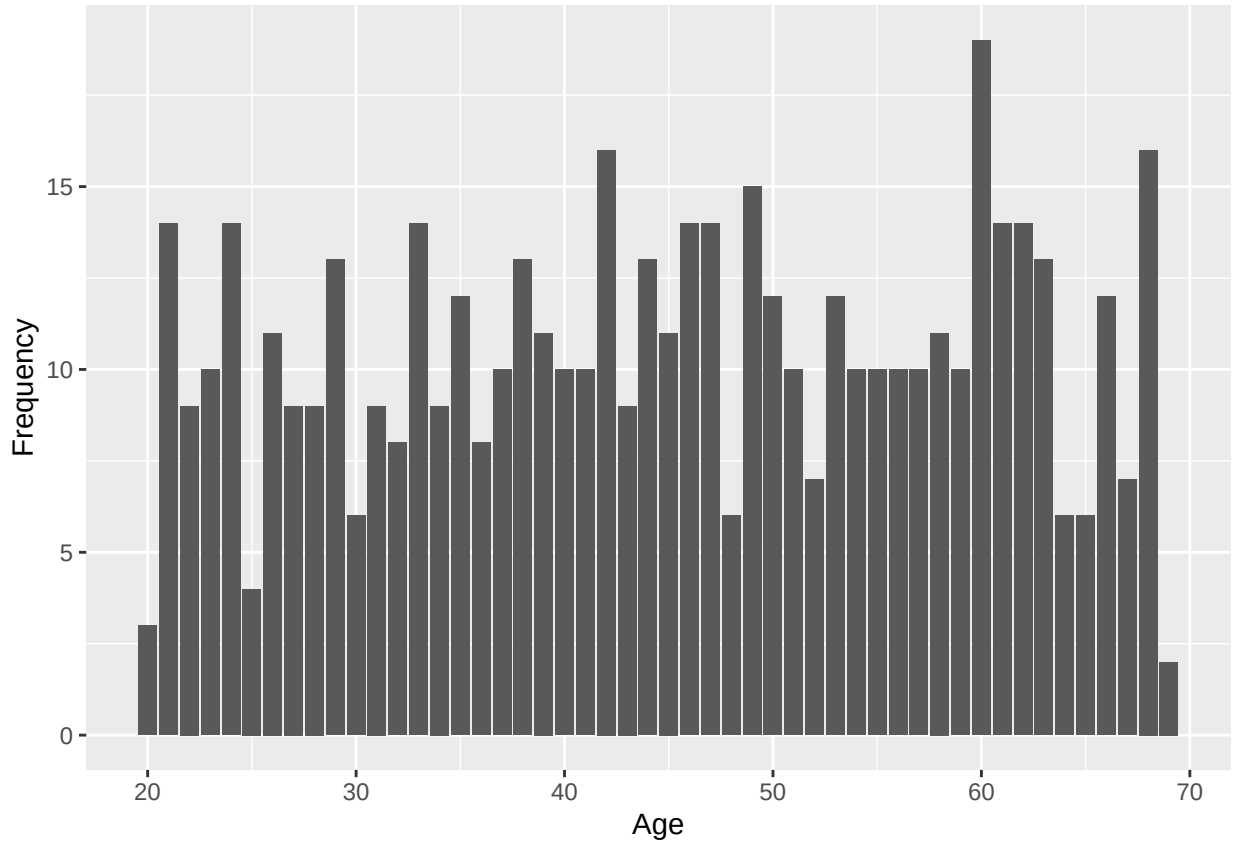


Figure 1: Distribution of age

Predictor `ethnicity` has 5 categories, with `non-hispanic_white` being the most prevalent category with 220 cases (42% of total), all the while `other` is the most infrequent category with 25 cases (5%). A total

of 525 cases contain `ethnicity` data, once again suggesting no presence of missing data values within this variable.

Predictor `education` is similar to `ethnicity`, having 5 categories. That said, the frequencies of said categories seem to be less out of proportion, with `some_college` being the most frequent category with 155 (30%) cases. Whilst the data was retrieved from respondents of ages 20 and higher, we do not observe any missing data values within the variable. This can be further justified by the fact that the minimum `age` within our dataset is 20, therefore foregoing possible issues with younger participants being unable to provide data for this specific variable.

The predictor variable `marital` contains 6 categories. It should be noted that the category value `married` exceeds the average frequency of other categories (that being 49.2) by a large margin - as can be seen in Figure 2. Specifically, 279 cases (53%) were `married`, with the other 5 categories made up the remaining 47% of the cases. `never_married` was the most frequent among those other 5 categories with 102 (19%) cases, whilst `separated` was the most infrequent category with only 14 (3%) cases. Furthermore, the meaning of `never_married` and `living_with_partners` is ambiguous in the sense that a person specifically living with their (non-married) partner can fall into both these categories.

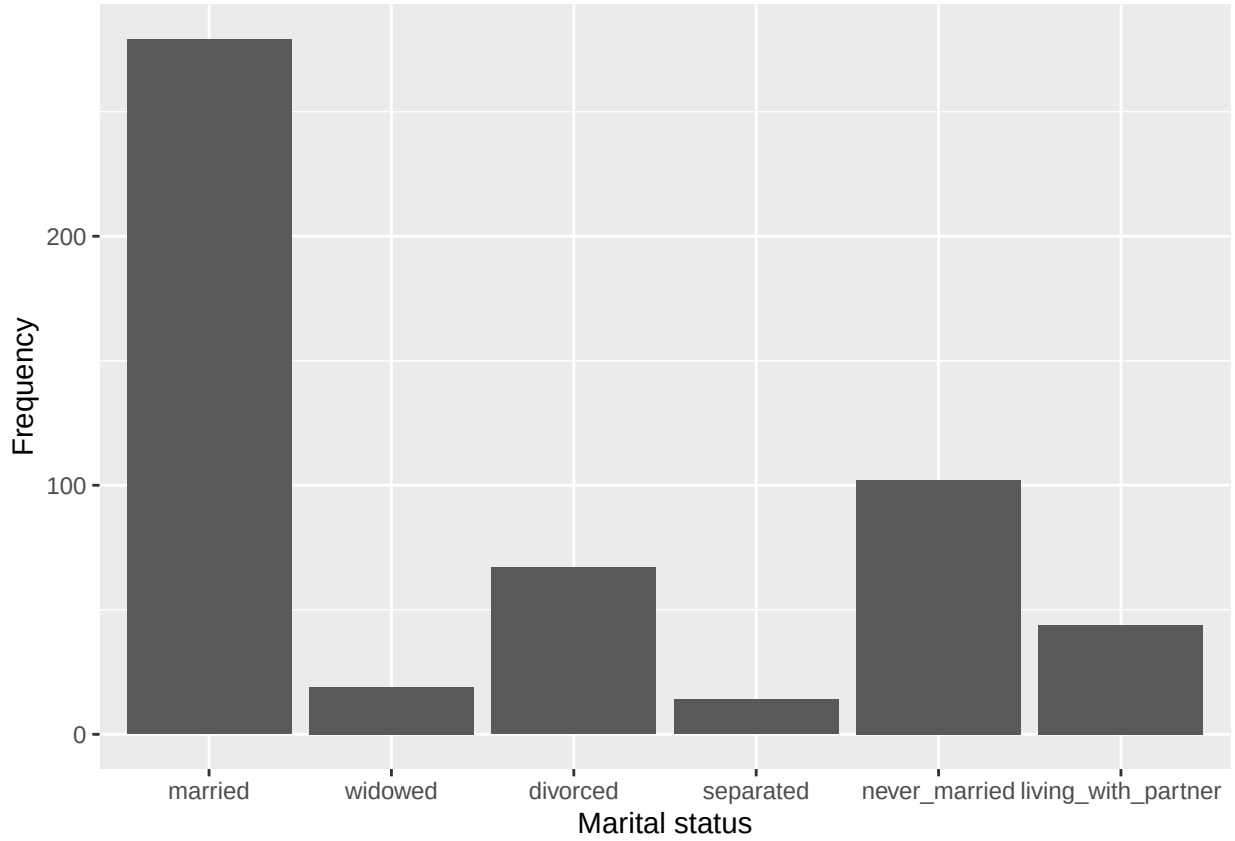


Figure 2: Distribution of marital status categories

Predictor `income` is a ordinal variable formed by 12 levels, ranging from income values 0 to 100000+. This data, like `age`, is topcoded at the value 100000. This also explains the highest income category - that being 100000+ - having 76 cases (9%), as can be seen in Figure 3. If we ignore the latter case, the most frequent category is instead `25000:34999`, which coincidentally is the mid-range category of our income data.
525 cases contain `income` data, therefore the variable does not contain missing data values.

The multiple levels of depression all have ranging scores between 0 and 3, with increments of 1 specifically. A higher score indicates agreement with the signs of depression described in the survey. Only `dep1` and

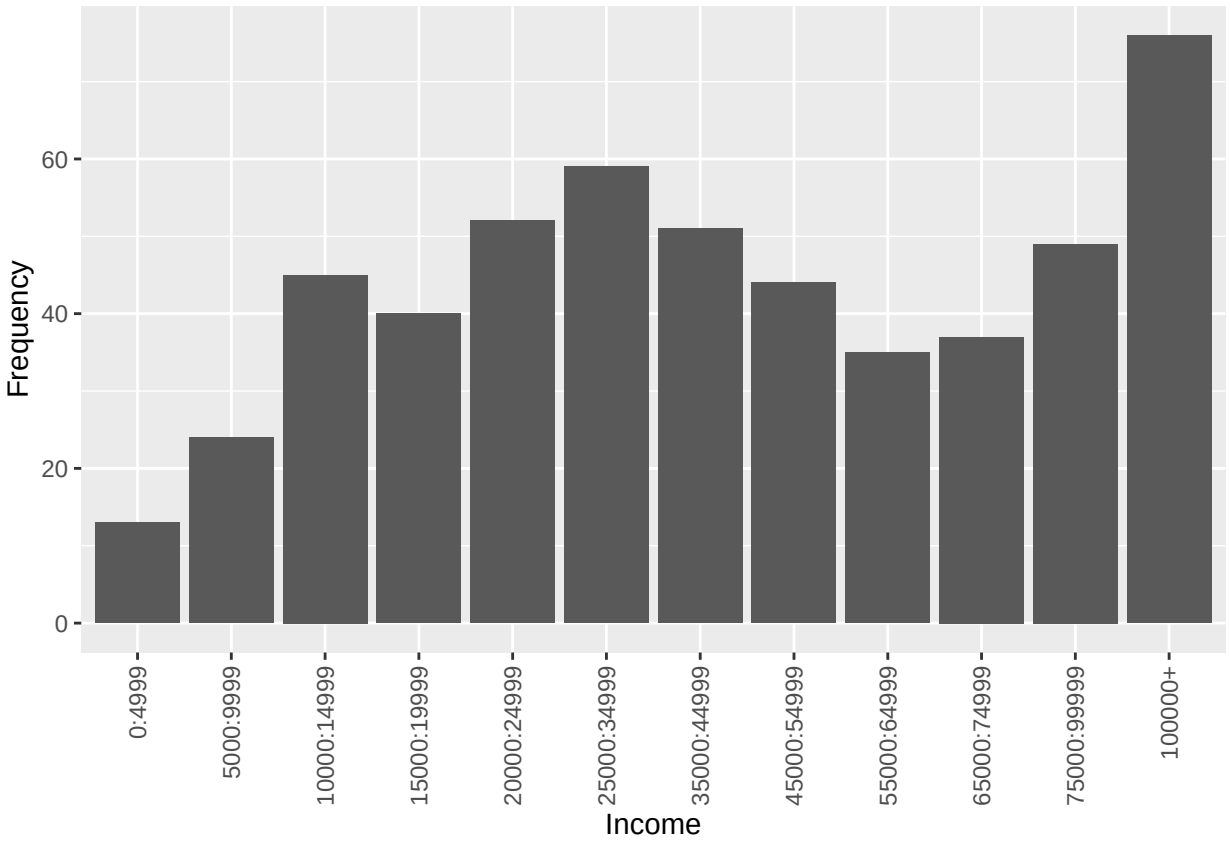


Figure 3: Distribution of income

dep7 have complete data data, whilst the other levels have varying amounts of missing data. Missing data could possibly be caused by the respondents being reluctant to answers these questions. We also observe that dep2, dep3, dep5, dep6 have the same amount of missing data value, hence it is plausible that they have a shared cause of missing data. Lastly, dep9 only has 18 cases of scores above 0, whilst also having the lowest mean value out of all depression levels. dep9 describes the presence of suicidal thoughts, hence it makes sense that this has the lowest score out of all levels. On the contrary, dep4 (feeling tired) has the highest average score.
Overall, the score distribution are heavily skewed towards the 0 values.

3.2 Response rates

As mentioned, some levels of of depression as well as the outcome variable drink_regularly contain missing data values. Figure 4 further confirms this, showing that dep2, dep3, dep5 and dep6 have an equal amount of cases with missingness, whilst also being the variables with the most amount of missing data. drink_regularly is the variable with the second most missingness, followed by dep9 and dep8. In total, 7 out of the 17 variables have NA values.

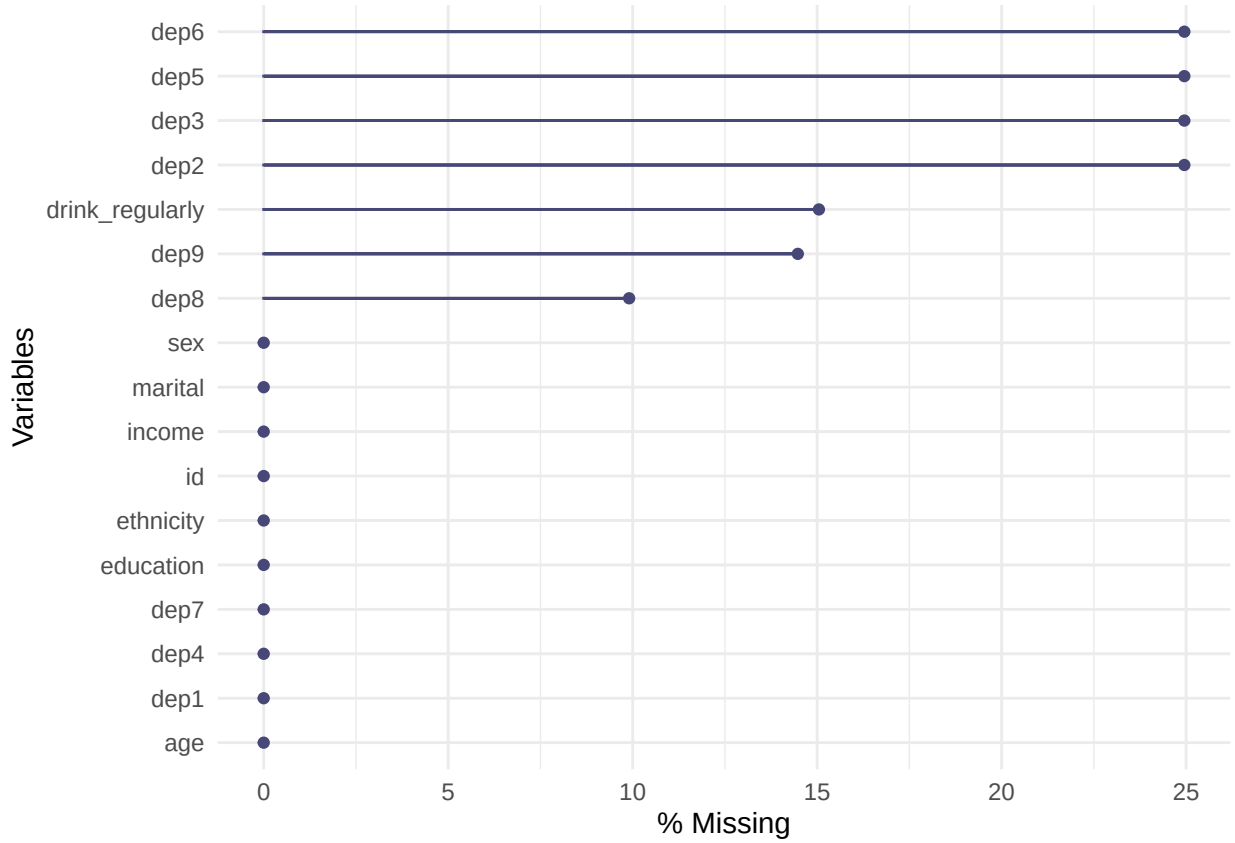


Figure 4: Percentages of missing data per variable

3.3 Outliers

For our only continuous predictor age, a boxplot shows potential outliers if datapoints fall outside of the interquartile range (whiskers of the plot). Figure 5 shows the distribution of age in said boxplot, revealing no possible outliers. This is to be expected, since age was uniformly distributed as mentioned in Section 3.1.

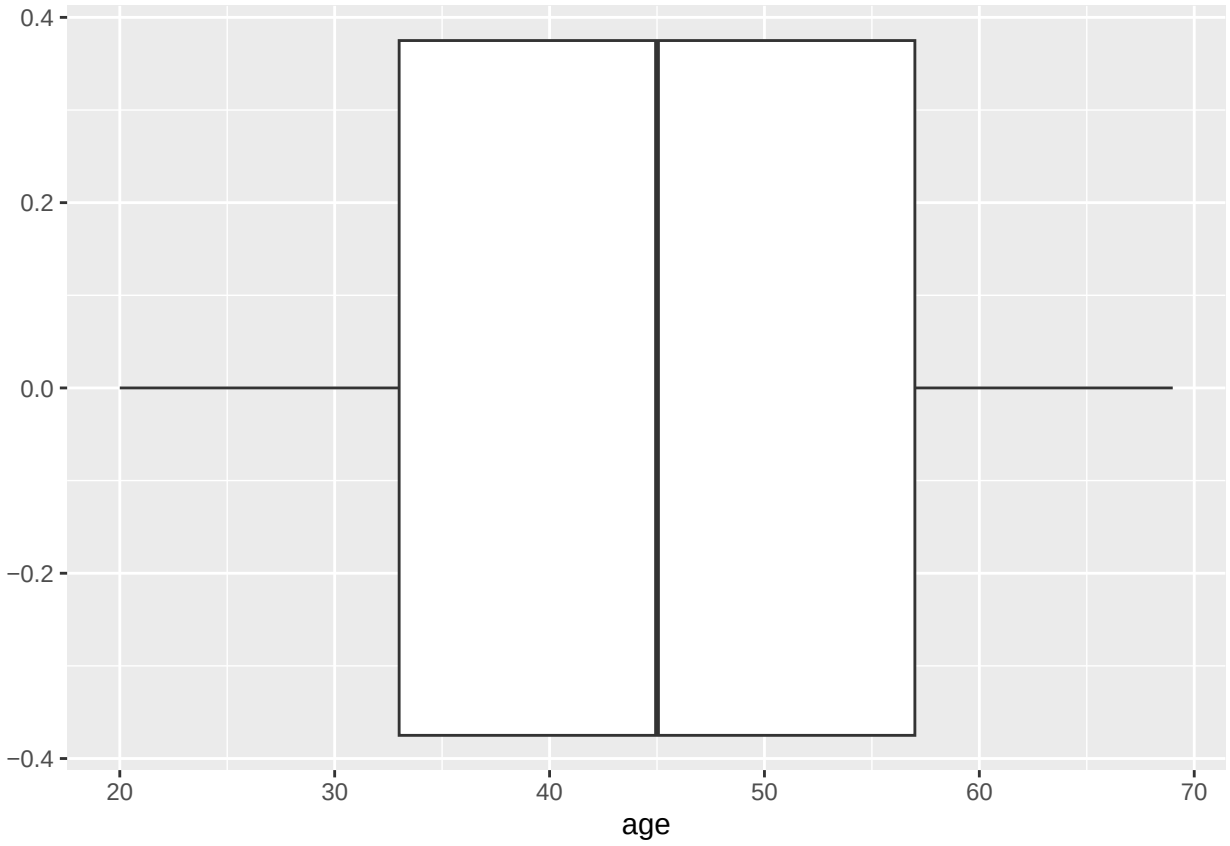


Figure 5: Boxplot of age

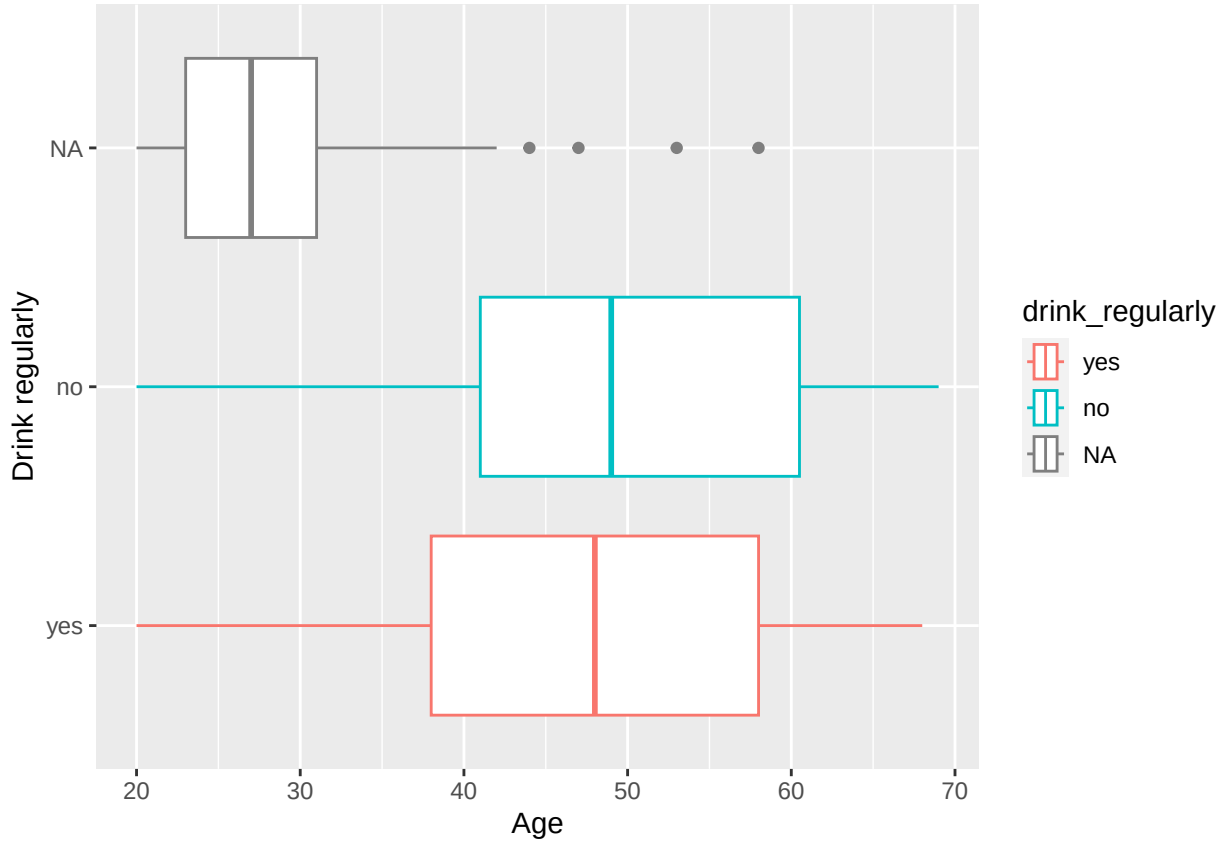
Looking at the distributions of the various categorical variables, the `marital` showed the aforementioned imbalance in frequency distribution, with the category `married` being overly dominant within the data. We considered combining the other `marital` categories into one, as their infrequency could allude to being possible outliers. Table 3 shows the interrelations between the `martial` variable and our outcome `drink_regularly`. From this, we concluded that combining the remaining 5 categories - that being `widowed`, `divorced`, `separated`, `never_married` and `living_with_partner` - was not feasible, since these relations seem to differ per category. For example, `widowed` has more "no" cases compared to "yes", whilst the relation is reversed for `separated`.

Table 3: Contingency table of marital status vs. drinking regularly

	yes	no
married	175	81
widowed	8	11
divorced	48	14
separated	8	4
never_married	45	23
living_with_partner	23	6

3.4 Correlations

Figure **??** shows the distributions of `age` over the outcome variable `drink_regularly` in a boxplot. Not considering the NA values, `age` does not seem to be strongly correlated with `drink_regularly`, with "no" cases having only slightly higher ages than its counterpart.
It should however be noted that the NA values of `drink_regularly` are primarily present in younger participants, therefore possibly attenuating the correlation effect between these two variables. For example, a large amount of missing values belonging to "yes" could shift the distribution of these cases towards lower `age` values, hence increasing the difference in distributions for each `drink_regularly` outcome.



Contingency tables and frequency distribution plots (including NA values) were used to observe the correlation between our categorical predictors and the outcome variable. From this, we observed that `sex` seemed to be highly correlated with `drink_regularly` - 172 cases of `male` drank regularly, whilst only 39 didn't. Compared to the 135 cases of `female` of "yes" and 100 cases of "no", this ratio seemed to differ greatly per `sex` category. The missingness of `drink_regularly` was evenly distributed across both `male` and `female` cases.

Likewise, `marital` showed a similarly strong correlation effect. The category `divorced` had a significantly higher ratio of respondents who drank regularly, compared to cases of `widowed`, where respondents were more likely to not drink regularly as can be seen in Figure 6. We also noted that the missing data of `drink_regularly` was not evenly distributed across the `marital` categories. Rather, the missingness was more prevalent in the `never_married` and `living_with_partner` cases. This possibly ties back into the fact that more data was missing amongst younger respondents, which can be direcly seen in Figure **??** where `never_married` and `living_with_partner` have cases of a younger age as well.

`ethnicity` showed a weaker correlation effect with the outcome variable. Only `non-hispanic_black` and `other` showed different frequency distributions compared to the other categories. Missing data of `drink_regularly` was present in all categories of `ethnicity`, though `other_hispanic` and `other` had close to none compared to the remaining categories.

The variable `education` was considered to have the weakest correlation with `drink_regularly`, having similar success ratios across all categories all the while missing data of `drink_regularly` was also evenly spread.

`income` was considered to be correlated with the outcome variable. Performing a similar analysis, we observed that respondents with an income up until 24999 were less likely to drink regularly than cases with a higher income. Akin to `education`, missing data of `drink_regularly` seemed to be evenly spread across the `income` values.

Finally, all levels of depression showed a effect of correlation with `drink_regularly`, where higher scores tended to decrease the ratio of "yes" to "no" cases, that is to say that respondents were less likely to drink

regularly if signs of depression were present.

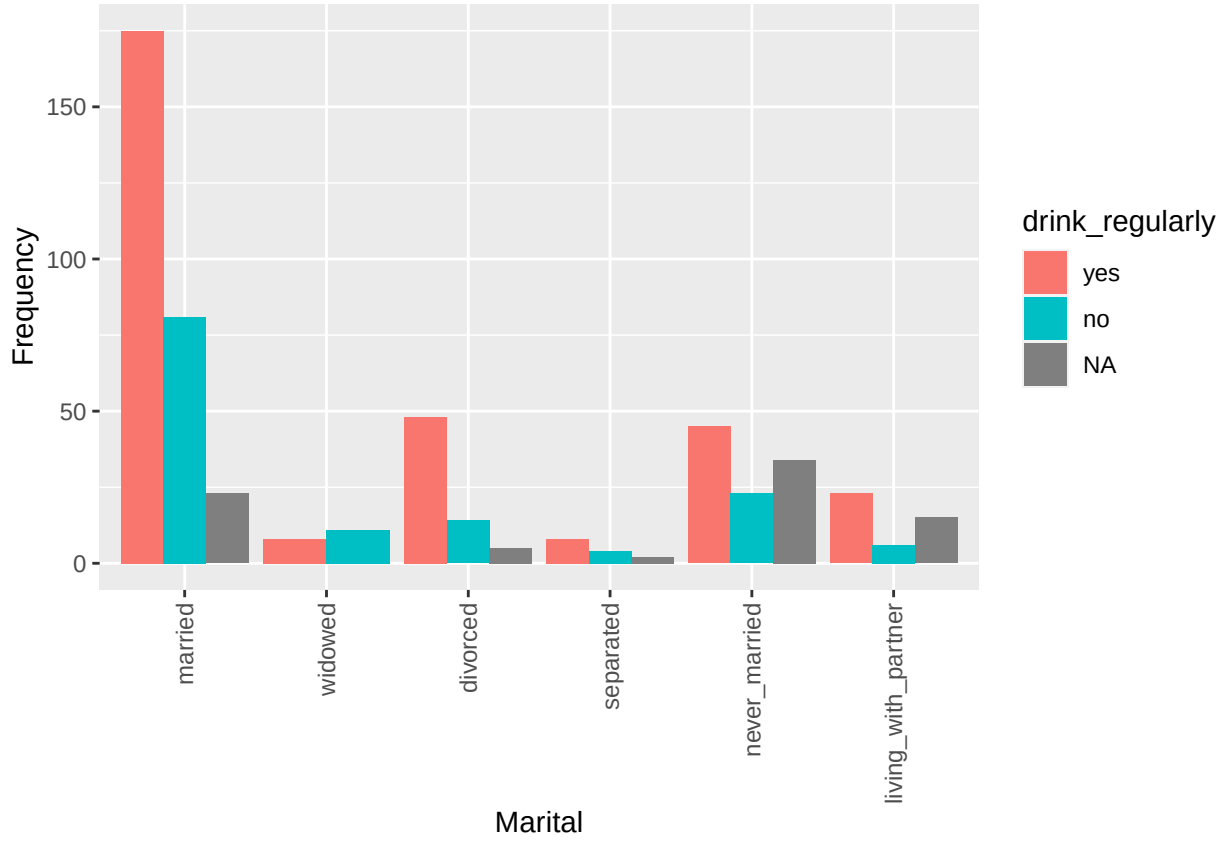


Figure 6: Correlations of marital vs. drink regularly

4 Missing data problem (Aga)

4.1 Missing data and response Patterns

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

As can be seen on Figure 7 , 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', 'dep6', 'dep8' and 'dep9' that create the depression score variable. 15% of responses are missing for the predictor variable. 25% of the responses are missing for the individual depression questions 2, 3, 5 and 6, 10% are missing for question 8 and 14% for question 9.

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

Figure 8 reveals that there are 13 distinct response patterns in the dataset and the missingness pattern is not monotone. The most frequent pattern is no missing entries, with 263 cases. It is important to note that the depression questions 2, 3, 5 and 6 are always either all present or all missing. 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.

Based on the missingness pattern of the depression items, 41% of the overall depression score includes at least one missing value.

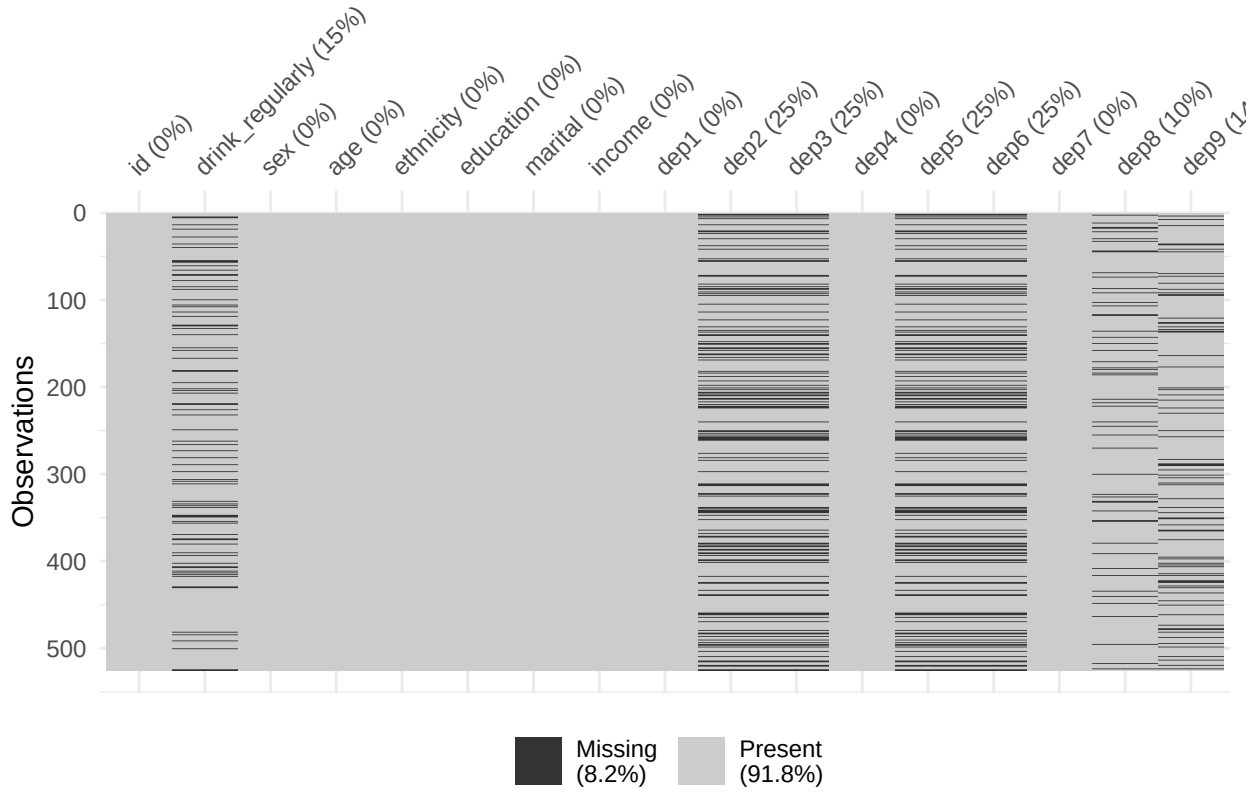


Figure 7: Distribution of missing data in each variable

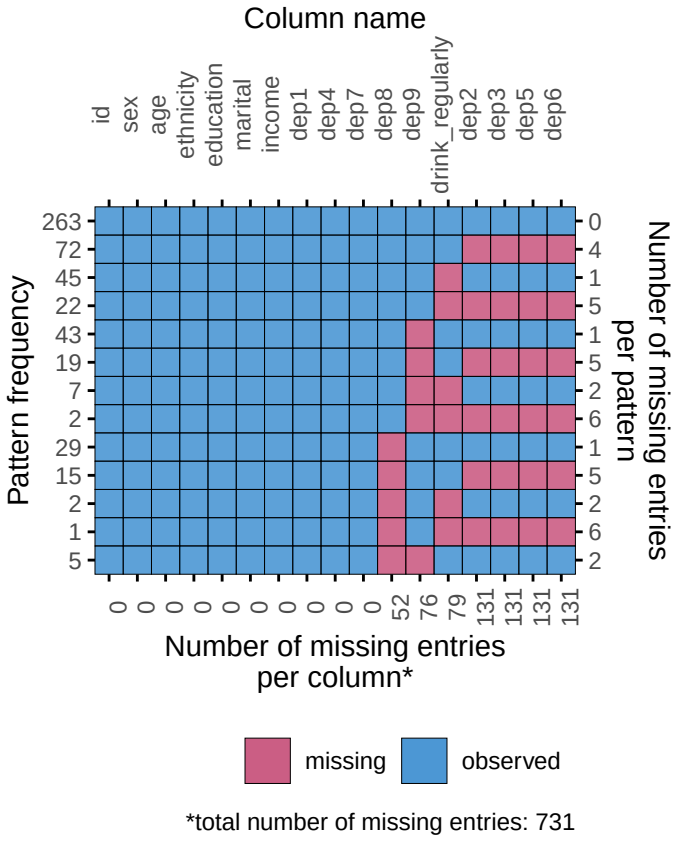


Figure 8: Response patterns and their frequency

4.2 Missing data mechanism

Missing completly at random (MCAR) missingneses mechanism is often an important assumption for statistical analysis, including this one. To gain some inside whether the data is MCAR or not, we deploy a variety of tests. If missing values of a variable are MCAR then there should be no significant dependency of them with other variables.

Firstly, Little MCAR test was preformed to verify if the missing data is MCAR at a global level, thus for all of the instances of missigness. The test was significant ($\chi^2(164) = 465.18, p < 0.01$), therefore the data may be assumed as not MCAR.

Since the data being MCAR can be questioned following the Little's test, a t-test was performed for all missing values vectors and numerical variables to check which variables are the likely culprit. The test compared the observed means of a given numerical variable within the group of individuals that had an observed value and the group with a missing value for another variable. Effectively testing if there is a significant difference in the continuous variable in the group that answered another question and the group that did not. Since the null hypothesis is no difference in means across the groups, a single significant value for a missingness of a given variable suggests that the missing values in that variable depend on the observed values of another variable. Therefore, it is likely that the missing values are not MCAR. Due to the identical pattern of responses in 'dep2', 'dep3', 'dep5' and 'dep6' it was sufficient to only test once for the dependency of their missing values. By the design of the test it is impossible to test the dependency of missing values and observed values within the same variable, thus these values are missing from the table.

Assuming an alpha of 0.05, it can be observed that missing values of 'drink_regularly' are dependent on 'age', 'dep2' and 'dep9'. Similarly, 'dep2', 'dep3', 'dep5' and 'dep6' seem to be dependent on 'age' and the remaining depression questions. For missingness in 'dep8' and 'dep9' there are no significant differences across groups in any of the numerical variables. These tests suggest that at least 'drink_regularly', 'dep2', 'dep3', 'dep5' and 'dep6'are not MCAR, thus MAR will be assumed for them. In Table 4 the exact t-statistic and the p-value are reported.

Table 4: Dependency t-test: mean comparison in numerical variables across missing values and observed values in the other variables

numerical variable	missing drink_regularly	missing dep2/3/5/6	missing dep8	missing dep9
age	t = 19.3, pVal = 3.1e-45	t = 1.38, pVal = 0.17	t = -1.16, pVal = 0.249	t = -0.884, pVal = 0.379
dep1	t = 1.39, pVal = 0.167	t = -6.72, pVal = 2.88e-10	t = 0.464, pVal = 0.644	t = -0.444, pVal = 0.658
dep2	t = 3.63, pVal = 0.000405	-	t = 0.332, pVal = 0.742	t = 1.22, pVal = 0.224
dep3	t = 0.797, pVal = 0.428	-	t = -0.137, pVal = 0.892	t = 0.0545, pVal = 0.957
dep4	t = -0.541, pVal = 0.59	t = -7.8, pVal = 6.11e-13	t = -0.68, pVal = 0.499	t = -1.28, pVal = 0.202
dep5	t = 2.01, pVal = 0.0466	-	t = -0.605, pVal = 0.549	t = -0.731, pVal = 0.467
dep6	t = 0.822, pVal = 0.414	-	t = 2.3, pVal = 0.0242	t = 0.00637, pVal = 0.995
dep7	t = -0.382, pVal = 0.703	t = -8.19, pVal = 1.24e-13	t = -0.239, pVal = 0.812	t = -0.0635, pVal = 0.949
dep8	t = -0.32, pVal = 0.75	t = -4.26, pVal = 3.86e-05	-	t = 0.595, pVal = 0.553
dep9	t = 2.48, pVal = 0.0135	t = -2.64, pVal = 0.00947	t = -0.295, pVal = 0.769	-

Table 5: Independency Chi-squared test

categorical variable	missing drink_regularly	missing dep2/3/5/6	missing dep8	missing dep9
drink_regularly	-	X^2 = 4.2, pVal = 0.0404	X^2 = 1.52, pVal = 0.218	X^2 = 0.464, pVal = 0.496
sex	X^2 = 1.09, pVal = 0.296	X^2 = 5.04, pVal = 0.0248	X^2 = 0.469, pVal = 0.494	X^2 = 0.317, pVal = 0.573
ethnicity	X^2 = 5.49, pVal = 0.241	X^2 = 1.64, pVal = 0.802	X^2 = 7.74, pVal = 0.102	X^2 = 5.8, pVal = 0.215
education	X^2 = 2.7, pVal = 0.609	X^2 = 5.72, pVal = 0.221	X^2 = 1.63, pVal = 0.803	X^2 = 5.13, pVal = 0.275
marital	X^2 = 55.7, pVal = 9.58e-11	X^2 = 22.2, pVal = 0.000477	X^2 = 5.08, pVal = 0.407	X^2 = 3.97, pVal = 0.553
income	X^2 = 2.94, pVal = 0.991	X^2 = 8.84, pVal = 0.636	X^2 = 9.4, pVal = 0.585	X^2 = 17.3, pVal = 0.0995

Since the t-test is not appropriate for categorical variables, we perform the Chi-squared test for them. The null hypothesis of the test is that there is no significant relationship between the categorical variables. The

test compares observed frequencies to expected frequencies if there was no relationship between the variables.

The outcomes of the test are presented in Table 5. However, only 'martial' with missingness of 'drink_regularly' and 'marital' with missingness of 'dep2', 'dep3', 'dep5' and 'dep6' are significant. Thus, it further supports the non-MCAR nature of missing data in these items.

However, the Chi-squared test has an requirement that the smallest expected frequencies have to be higher than 5. This requirement is not met for some combinations of missing values and categorical variables. Specifically: missing 'dep8/9' and 'ethnicity', 'marital', 'income'; missing 'dep2/3/5/6' and 'marital', 'income'; missing 'drink_regularly' and 'ethnicity', 'marital', 'income'. Since this assumption is not met for these combination, the pValues might not be correctly estimated. Therefore, the Fischer test, test with the same hypothesis but without the assumption, will be performed for these combinations.

Table 6: Fisher test

categorical variable	missing drink_regularly	missing dep2/3/5/6	missing dep8	missing dep9
ethnicity	pVal = 0.232	-	pVal = 0.0804	pVal = 0.21
marital	pVal = 1e-05	pVal = 0.00089	pVal = 0.399	pVal = 0.446
income	pVal = 0.987	pVal = 0.64	pVal = 0.561	pVal = 0.058

The outcomes of the FIsher test are presented in Table 6. Like in the case of Chi-squared test, only 'martial' with missingness of 'drink_regularly' and 'marital' with missingness of 'dep2', 'dep3', 'dep5' and 'dep6' are significant. Therefore, we can conclude the dependence of missingness in these variables on multiple other observed variables and reject MCAR in their case.

In all of the tests above 'dep8' and 'dep9' seem not to be dependent on any of the observed variables. Thus, perhaps MCAR could be assumed. However, since the depression scores will be collapsed into a single depression score ('dep') and other depression questions are MAR, the overall score is also assumed to be MAR.

4.2.1 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:367         male  :254   Min.   :20.00
##  1st Qu.:43912   no :158         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
##
##          income         dep1             dep2              dep3
##  100000+    : 76   Min.   :0.0000   Min.   :-0.3655   Min.   :-0.2842
##  25000:34999: 59   1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  20000:24999: 52   Median :0.0000   Median : 0.0000   Median : 0.1497
##  35000:44999: 51   Mean   :0.4095   Mean   : 0.3443   Mean   : 0.6760
##  75000:99999: 49   3rd Qu.:1.0000   3rd Qu.: 0.6794   3rd Qu.: 1.0000
##  10000:14999: 45   Max.   :3.0000   Max.   : 3.0000   Max.   : 3.0000
##  (Other)    :193
##       dep4             dep5              dep6              dep7
##  Min.   :0.0000   Min.   :-0.2874   Min.   :-0.1433   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   : 0.3481   Mean   : 0.3217   Mean   :0.3238
##  3rd Qu.:1.0000   3rd Qu.: 0.5391   3rd Qu.: 0.4741   3rd Qu.:0.0000
##  Max.   :3.0000   Max.   : 3.0000   Max.   : 3.0000   Max.   :3.0000
##
##       dep8              dep9
##  Min.   :-0.2293   Min.   :-0.33189
##  1st Qu.: 0.0000   1st Qu.: 0.00000
##  Median : 0.0000   Median : 0.00000
##  Mean   : 0.2062   Mean   : 0.06596
##  3rd Qu.: 0.0000   3rd Qu.: 0.00000
##  Max.   : 3.0000   Max.   : 3.00000
##
```

4.3 Model creation (maybe combine with next section?)

Mostly code for now, will add more later on.

TODO:

- Use mode imputation for depression and outcome.
- Mention n rows removal in case of listwise

Summary of code:
1. Create imputed data for each ad-hoc method. 2. combine deps levels for each dataset. 3. create models for analysis.

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

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

Depending on the chosen method used to treat the missing data, the answer on the research question changes. In the model fitted on the data where listwise-deletion was used as the missing data treatment only sex and marital status divorced were significant predictors (stats). Using mean imputation as the missing data treatment resulted in a fitted model where next to sex and marital status divorced, age and the intercept were also significant. The used method also changes the odds resulting from the model, in predictors that were significant in one but not the other (see intercept for example) but also in predictors that were significant in both (see age and divorced). As can be seen the estimated odds differ quite a lot. The standard errors in the listwise deletion model are also considerably smaller than those in the mean imputation model. All of the above shows that the used missing data treatment method can cause bias in the odds, standards errors and significant values.

The missing data treatment method also influences the look of the resulting model as a whole. The model fitted on the listwise deletion data looks better, the null deviance and residual deviance are both considerably lower compared to the model fitted on the data resulting from mean imputation (values). Looking at AIC also gives the impression that the liswise deletion method gives a betger model.

To summarize the chosen method to treat the missing data causes bias in the individual predictors and in the look of the fitted model as a whole. As a result the answer to the studied research question depends on which method is chosen.

4.5 Conclusion in terms of answering RQ (Nisse)

- go back to papers, reflect hypo
- stuffs

5 References

Huntington-Klein, Nick. 2023. *Vtable: Variable Table for Variable Documentation.* https://CRAN.R-project.org/package=vtable.

Little, Roderick J. 1986. "A Test of Missing Completely at Random for Multivariate Data with Missing Values." *Journal of the American Statistical Association* 83 (404): 1198–1202. https://doi.org/10.1080/01621459.1988.10478722.

Oberman, Hanne. 2024. *Ggmice: Visualizations for 'Mice' with 'Ggplot2'.* https://github.com/amices/ggmice.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sjoberg, Daniel D., Karissa Whiting, Michael Curry, Jessica A. Lavery, and Joseph Larmarange. 2021. "Reproducible Summary Tables with the Gtsummary Package." *The R Journal* 13: 570–80. https://doi.org/10.32614/RJ-2021-053.

Soetewey, Antoine. n.d. "Chi-Square Test of Independence in r." *Stats and R.* https://statsandr.com/blog/chi-square-test-of-independence-in-r/#chi-square-test-of-independence-in-r.

Tierney, Nicholas, and Dianne Cook. 2023. "Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations." *Journal of Statistical Software* 105 (7): 1–31. https://doi.org/10.18637/jss.v105.i07.

van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "mice: Multivariate Imputation by Chained Equations in r." *Journal of Statistical Software* 45 (3): 1–67. https://doi.org/10.18637/jss.v045.i03.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wu, Wei, Fan Jia, and Craig Enders. 2015. "A Comparison of Imputation Strategies for Ordinal Missing Data on Likert Scale Variables." *Multivariate Behavioral Research* 50 (5): 484–503. https://doi.org/10.1080/00273171.2015.1022644.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.

A Appendix

Table 7: 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%					
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