Missing Data - Assignment A

Aga Kubica, Nisse Hermsen, Ruben Custers | Group 9

2024-03-13

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 (GBD 2016 Alcohol Collaborators 2018). 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 can the occurrence of regular alcohol consumption (12 or more in a year) 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, whilst household income will be positively correlated (Garnett et al. 2022; Moore et al. 2005). In addition, expectation entails that male (vs female) and white (vs other ethnicities) will be positively correlated with the regular occurrence of alcohol consumption (Moore et al. 2005). Regarding marital status, it is expected that the married status will be positively associated with the regular occurrence of alcohol consumption compared to other marital statuses. That said, the latter was based on a study where the factor marital status consisted of just married and other (Moore et al. 2005), while in this study more categories were considered.

2 Methodology

2.1 Dataset

The dataset used was a subset of the data collected in the National Health and Nutrition Examination Survey (NHANES). The survey was a part of an annual program that investigated the health and nutrition of a representative sample of people in the United States. The data used contained information about 525 individuals that were collected for the NHANES 2007-2008 survey. This was a subset of the 12,946 individuals in that years' survey sample, out of which 78.4% were interviewed whilst 75.4% participants were examined in mobile examination centers. The NHANES survey was further subdivided into themed sections - such as the Alcohol Use questionnaire - with each of these sections having separate documentations that will be referred to later on.

The "full" dataset contained a wide range of variables related to the health of the individuals. This "full" dataset excluding the 'id' variable was used for the multiple imputation and stochastic regression missing data treatments. For the remaining data processing, analysis and modeling the data was further subsetted to only include 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).

2.2 Variables Description

Table 1 lists the variables used in the subset selection, which were utilized for the model in question. The table for the "full" dataset can be found in the Appendix [number]. The predictor variables [$dep1...dep9$] were sourced from the same Depression Screener, where respondents of age 18 to 150 were ought to assign

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

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 be combined to create an overall depression score.

The demographic variables - that being `sex`, `age`, `ethnicity`, `education` and `income` - were taken from the same screening component as well. The original authors of the dataset topcoded the variable `age` was at the value `80` for the respondents who were older than 80 years. Similarly, `education` was targeted at respondents of age `20` to `150`, thus excluding younger participants. This was due to the fact that this question included responses such as `AA degree` and `College Graduate`. The variable `marital` was also targeted at respondents of age `20` to `150`. Lastly, the variable `income` was ordinal, rather than continuous. As for the remaining demographic variables, namely `sex`, `age`, `ethnicity` and `income`, these were 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* package was used for table styling (Zhu 2021), *naniar* package was used to make the visualization of missing data percentages (Figure 8) (Tierney and Cook 2023), *Hmisc* was utilized for mode imputation, *vtable* helped with the summary statistic table in the Appendix (Huntington-Klein 2023) and *gtsummary* 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, an Exploratory Data Analysis was performed. The distributions of the variables were investigated and the presence of missing data was found in the outcome variable and some of the depression questions. No outliers were found during the EDA. Additionally, the missingness seemed to be more prevalent among younger individuals, which gave a first indication of a lacking MCAR mechanism. The important findings of this step are presented in EDA Results (Section 3).

Following the EDA, the missing data problem was addressed. The extend, distributions and patterns of missingness were investigated. Testing was done to verify the missing data mechanism. This included performing a global test - Little MCAR test (Little 1986) and testing of the dependencies between missing values in one variable against observed values of other variables. The latter was done using t-tests for continuous variables, whilst categorical variables were checked via Chi-squared test and Fisher test (Soetewey, n.d.). The Fisher test with simulated p-values was used to verify outcomes for combinations of variables that included an expected frequency lower than 5, thus not meeting the assumptions of Chi-squared test (Soetewey, n.d.). When the assumptions of MCAR was not met, MAR was assumed.

Four solutions to the missing data problem were implemented: list-wise deletion, mean imputation, stochastic regression imputation and multiple imputation.

For list-wise deletion, all cases with one or more missing values were excluded from the modeling. Thus, the dateset was reduced by half to 263 observation.

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 replaced by said mean value. This approach was taken, since treating ordinal variables on a Likert scale as continuous variables instead was appropriate (Wu, Jia, and Enders 2015). As found in the EDA (Section 3.1), all of the depression item distributions are heavily skewed towards 0. Because of that, '0.28', '0.53', '0.31', '0.20', '0.20' and '0.067' were imputed for 'dep2', 'dep3', 'dep5', 'dep6', 'dep8' and 'dep9' respectively. In turn, these low means even further skewed the distribution towards lower values. For `drink_regularly` - the outcome variable - the mode was computed rather than the mean. This was motivated by the need for a binary outcome variable in the logistic regression model that was utilized to answer the research question. The EDA suggested significantly more "yes" than "no" values in `drink_regularly`, hence the missing values were replaces by "yes". The 'drink_regularly' variable went from '397' 'yes' answers to '476', with '139' 'no' answers remaining unchanged. Therefore, after this data treatment the outcome ratio of the logistic regression was even more imbalanced, which could negatively affect the accuracy of the model.

Stochastic regression imputation

Multiple imputation

Following each missing data treatment, the values for the depression questions were summed up to create an overall depression score for each individual case. A sum was used - as opposed to other methods of aggregation (e.g. mean) - as it preserved a convenient interpretation of a unit increase in a depression score. This overall score was used for modeling as opposed to the individual depression levels.

To conclude, the different missing data treatments created three complete datasets and a single MIDS object that were then used for statistical analysis.

2.5 Modelling methodology

Including all of the variables in the two complete data sets was 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 following the aforementioned model-building methods, it was decided to keep all the predictor variables and therefore have a more complex model. Following model creation, the two logistics models were compared, where the impact of the two different missing data treatments was evaluated. For easier interpretation the model coefficients were exponentiated. Occurrence of regular drinking was considered the "successful" outcome.

3 EDA Results

3.1 Descriptive statistics

Table 8 in Appendix A 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 was 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 was observed that all variables were interpreted as a factor, excluding `age` and the multiple depression levels of `dep`. Table 1 suggested that `dep` should in fact be categorical ordinal and should therefore be cast as a factor. That said - and as mentioned in Section 2.4 - keeping the levels of `dep` as a numerical continuous datatype was beneficial for the missing data problem.

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

The dichotomous outcome variable `drink_regularly` had 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 the logistic regression model. Additionally, the total amount of value entries (`N`) of 446 suggested that 79 cases contained values outside of the set of possible binary values - most likely being missing values. Table 2 further confirmed this.

Table 2: Drink regularly value distributions

drink_regularly	n
no	139
yes	307
NA	79

Predictor `sex` was a dichotomous variable with a balanced frequency distribution across the values "male" and "female", that being 48% and 52% respectively. 525 entries contained one of these values, suggesting that the variable did not contain missing data.

The predictor `age` was 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 seemed to only contain cases of people between the age of 20 and 69. Moreover, Figure 1 suggested a uniform distribution of the age variable. Like `sex`, `age` had 525 cases of non-NA values, hence the variable did not contain missing data values.

Predictor `ethnicity` had 5 categories, with `non-hispanic_white` being the most prevalent category with 220 cases (42% of total), all the while `other` was the most infrequent category with 25 cases (5%). A total of 525 cases contained `ethnicity` data, once again suggesting no presence of missing data values within this variable.

Predictor `education` was similar to `ethnicity`, having 5 categories. That said, the frequencies of said categories seemed 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, no missing data values were observed within the variable. This could be further justified by the fact that the minimum `age` within the used dataset was 20, therefore foregoing possible issues with younger participants being unable to provide data for this specific variable.

The predictor variable `marital` contained 6 categories. It should be noted that the category value `married` exceeded 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 making 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

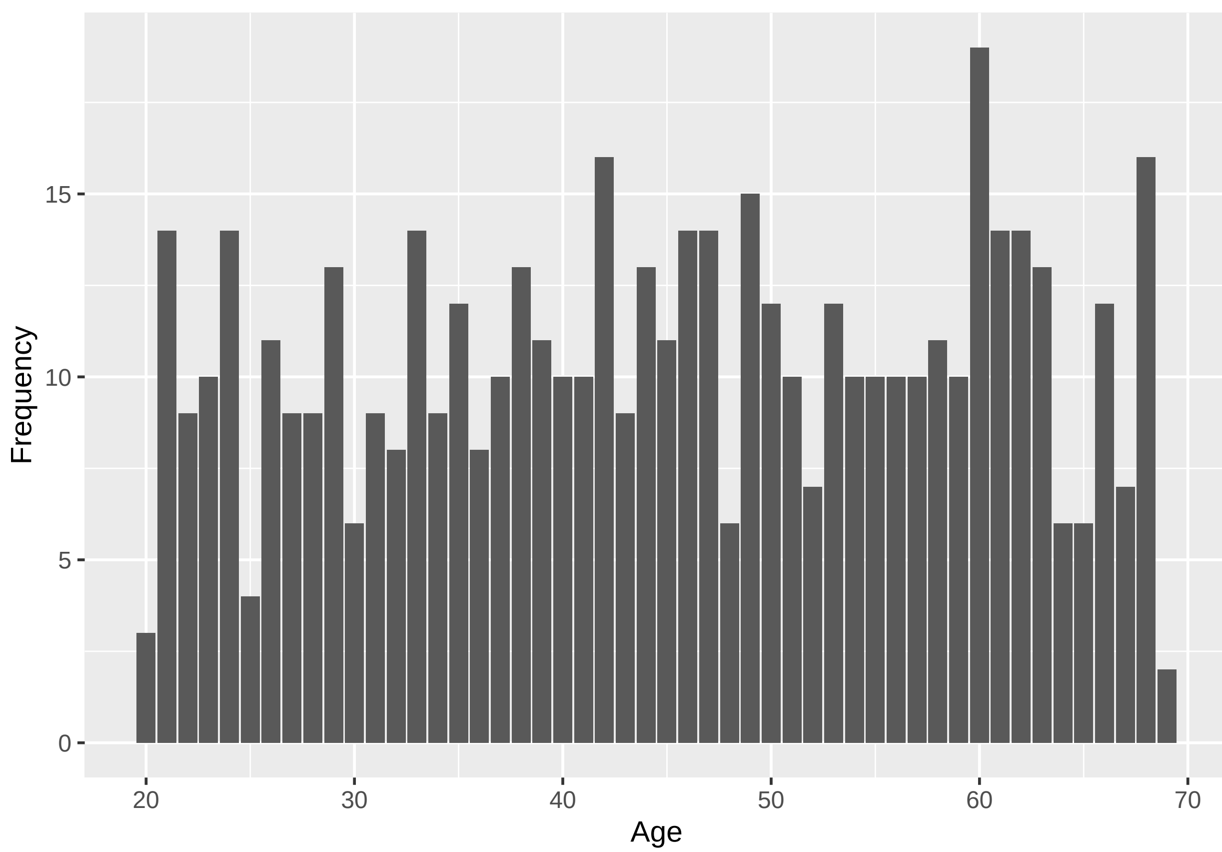


Figure 1: Distribution of age

of `never_married` and `living_with_partners` was ambiguous in the sense that a person specifically living with their (non-married) partner could fall into both these categories.

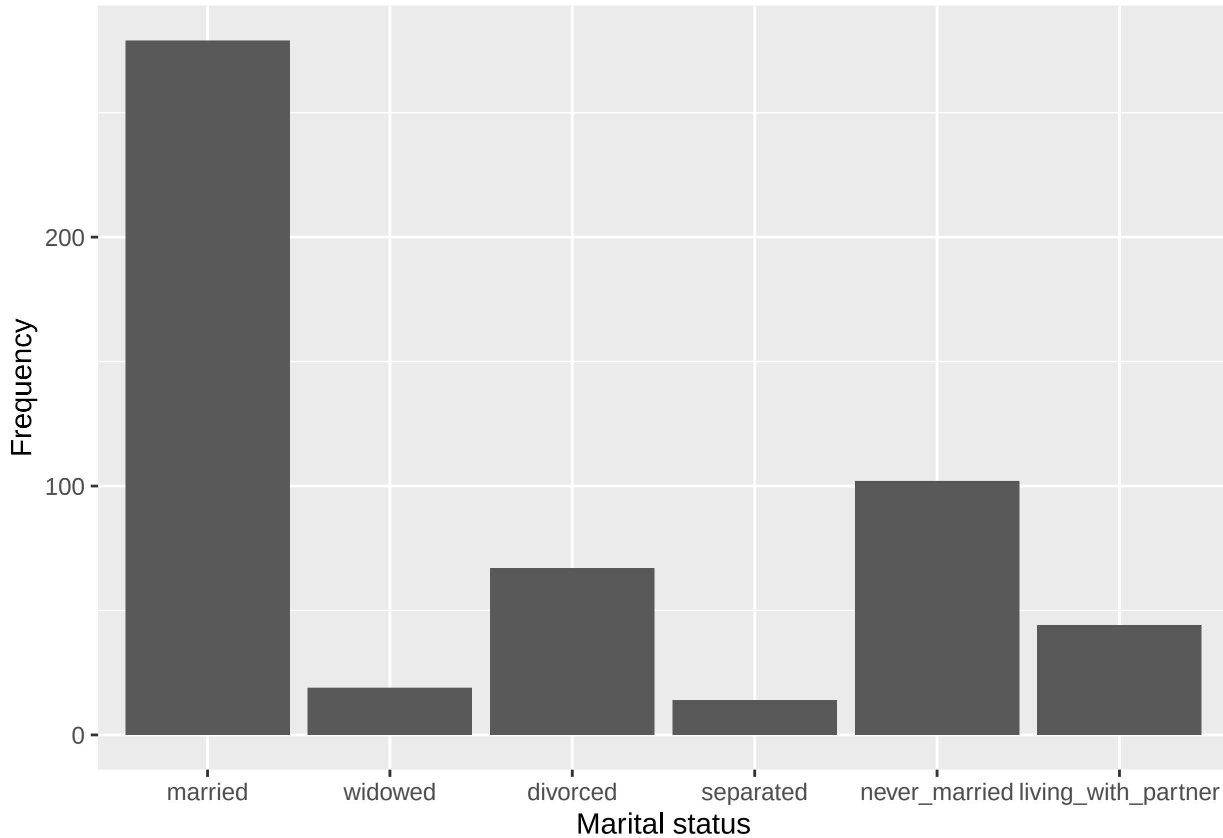


Figure 2: Distribution of marital status categories

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

The multiple levels of depression all had ranging scores between 0 and 3, with increments of 1 specifically. A higher score indicated an agreement with the signs of depression described in the survey. Only `dep1` and `dep7` had complete data data, whilst the other levels had varying amounts of missing data. Missing data could possibly be caused by the respondents being reluctant to answers these questions. It was also observed that `dep2`, `dep3`, `dep5`, `dep6` had the same amount of missing data values, hence it was possible that they had a shared cause of missing data. Lastly, `dep9` only had 18 cases of scores above 0, whilst also having the lowest mean value out of all depression levels. `dep9` described the presence of suicidal thoughts, hence it made sense that this was the lowest average score out of all levels. On the contrary, `dep4` (feeling tired) had the highest average score. Overall, the score distribution were heavily skewed towards the 0 values.

3.2 Outliers

For the only continuous predictor `age`, a boxplot was utilised to find potential outliers . Figure 4 shows the distribution of `age` in said boxplot, revealing no possible outliers. This was to be expected, since `age` was

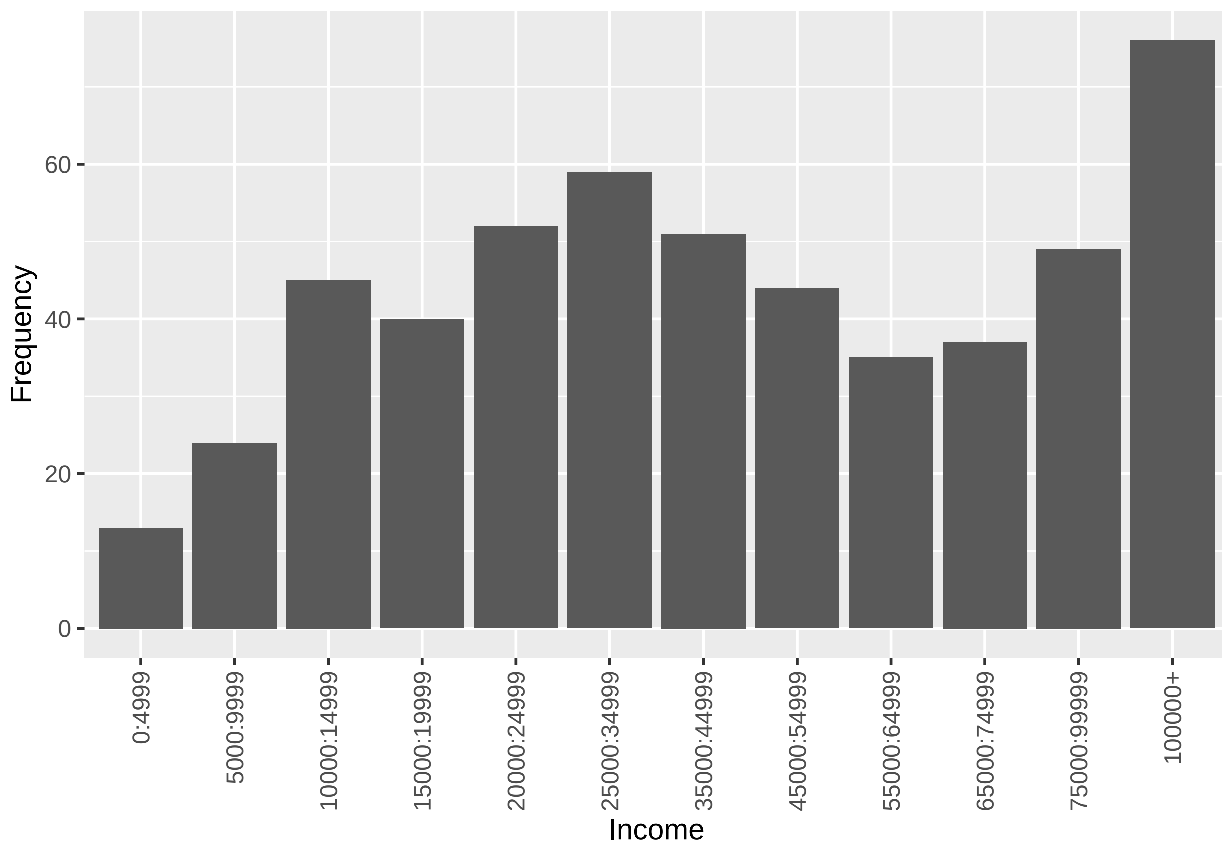


Figure 3: Distribution of income

uniformly distributed as was mentioned in Section 3.1.

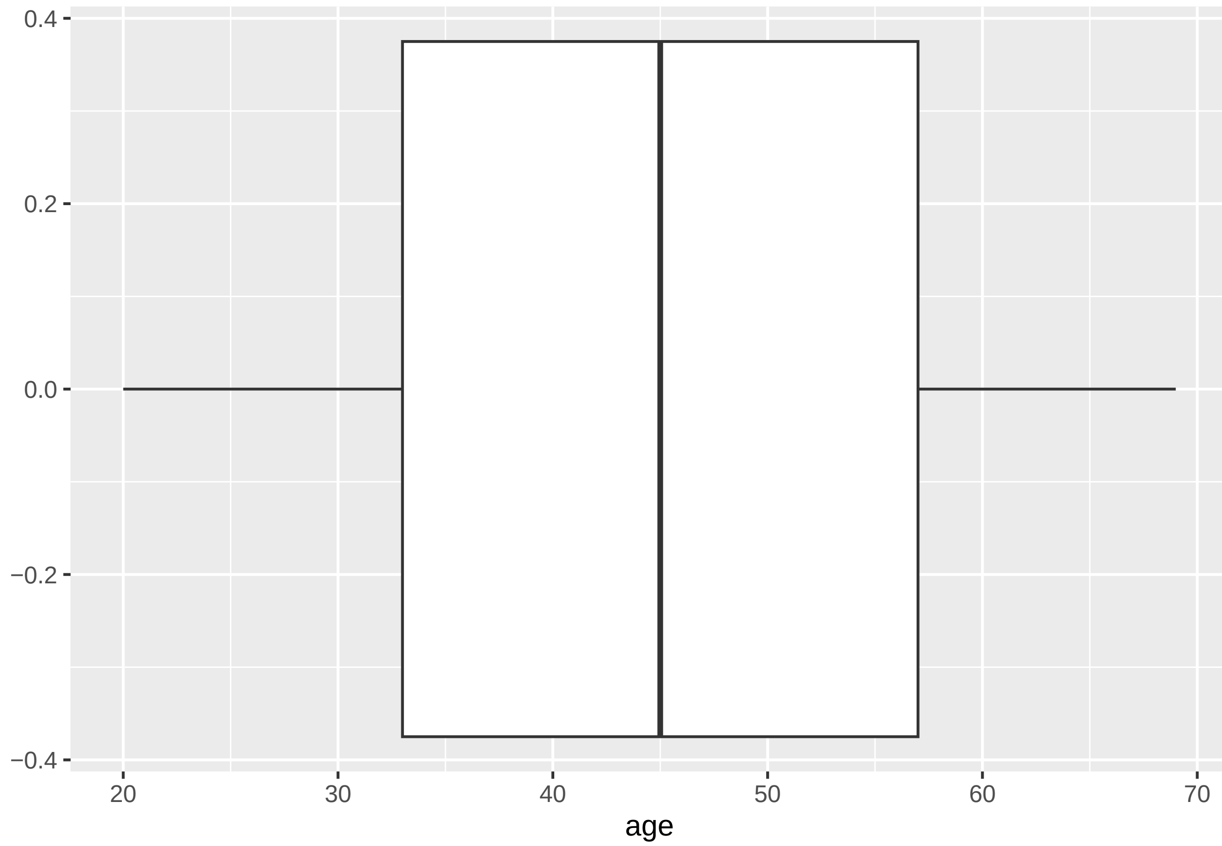


Figure 4: Boxplot of age

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. It was 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 the outcome `drink_regularly`. From this, it was concluded that combining the remaining 5 categories - that being `widowed`, `divorced`, `separated`, `never_married` and `living_with_partner` - was not feasible, since these relations seemed to differ per category. For example, `widowed` had more "no" cases compared to "yes", whilst the relation was reversed for `separated`.

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

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

3.3 Correlations

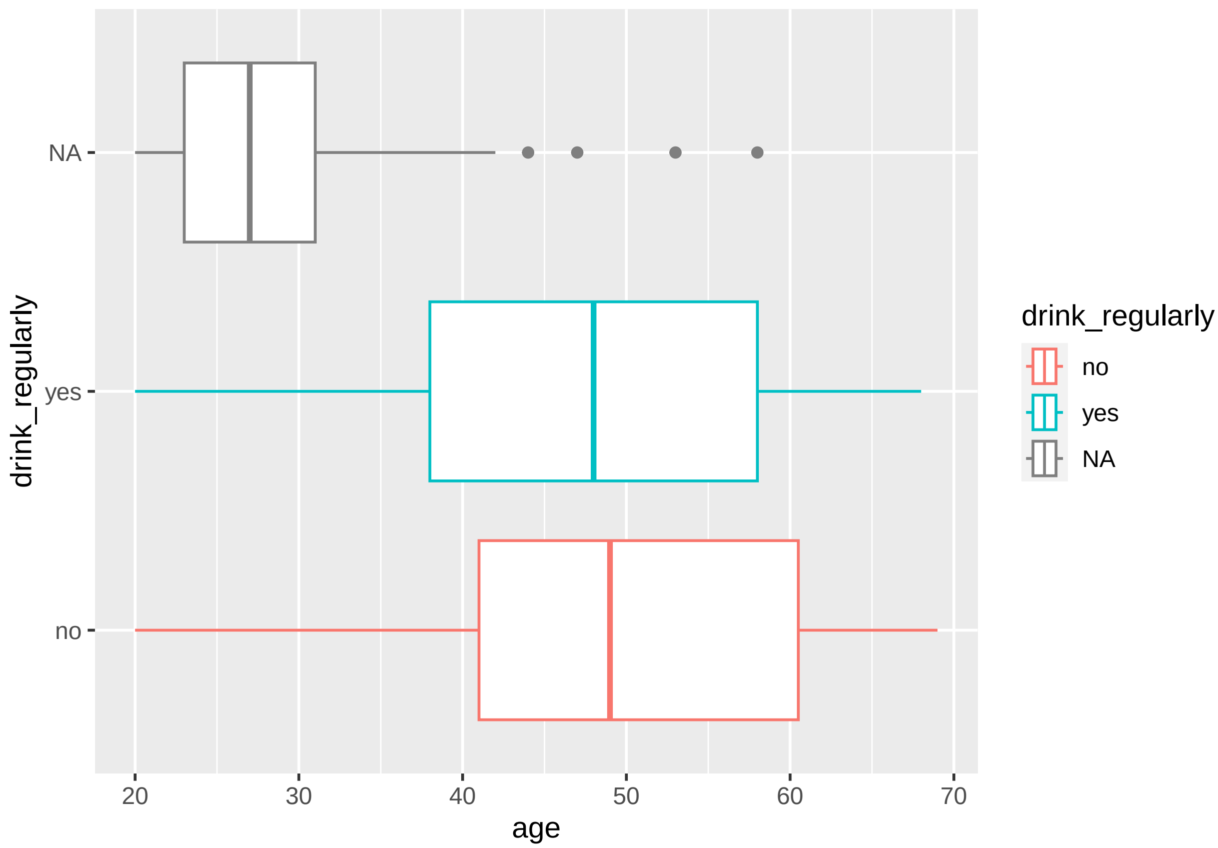


Figure 5: Correlations of age vs. drink regularly

Figure 5 shows the distributions of `age` over the outcome variable `drink_regularly` in a boxplot. Not considering the NA values, `age` did 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` were 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 the categorical predictors and the outcome variable. From these, it was 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. It was 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 7 where `never_married` and `living_with_partner` have more cases of a younger age as well.

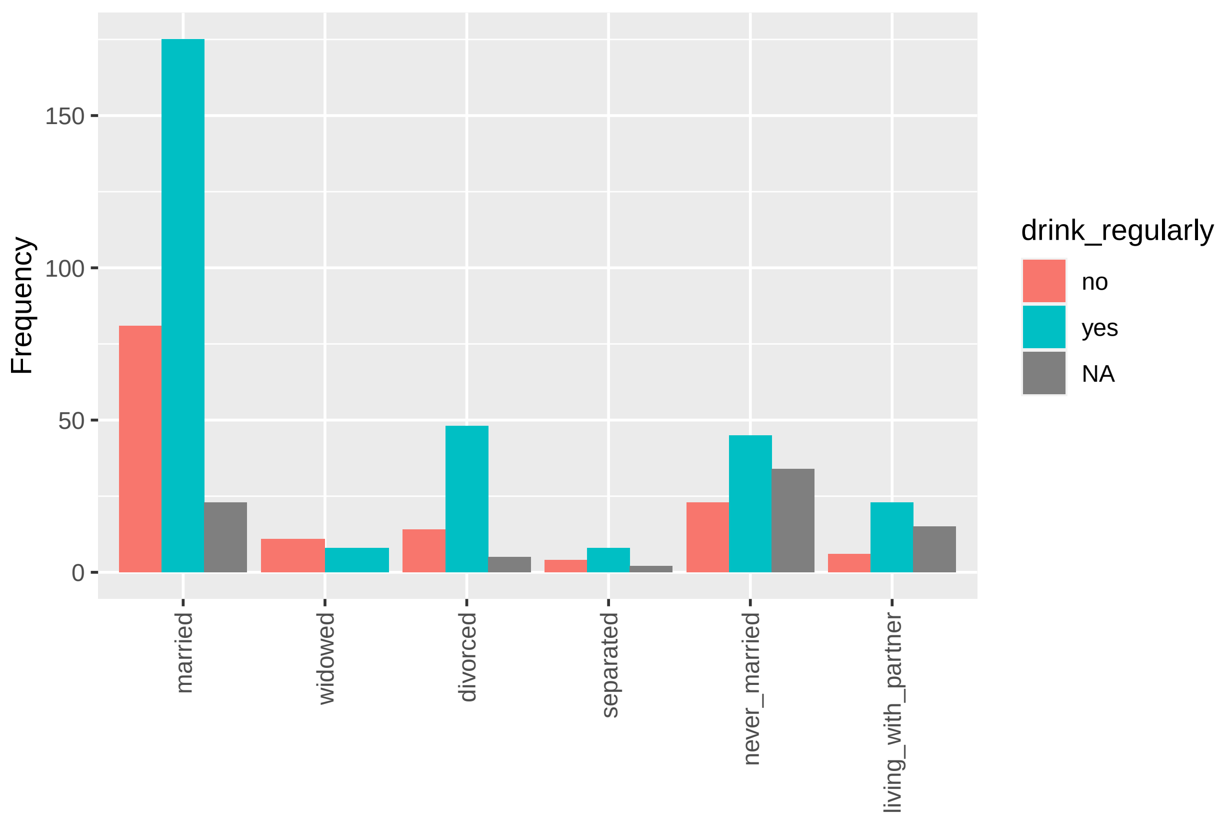


Figure 6: Correlations of marital vs. drink regularly

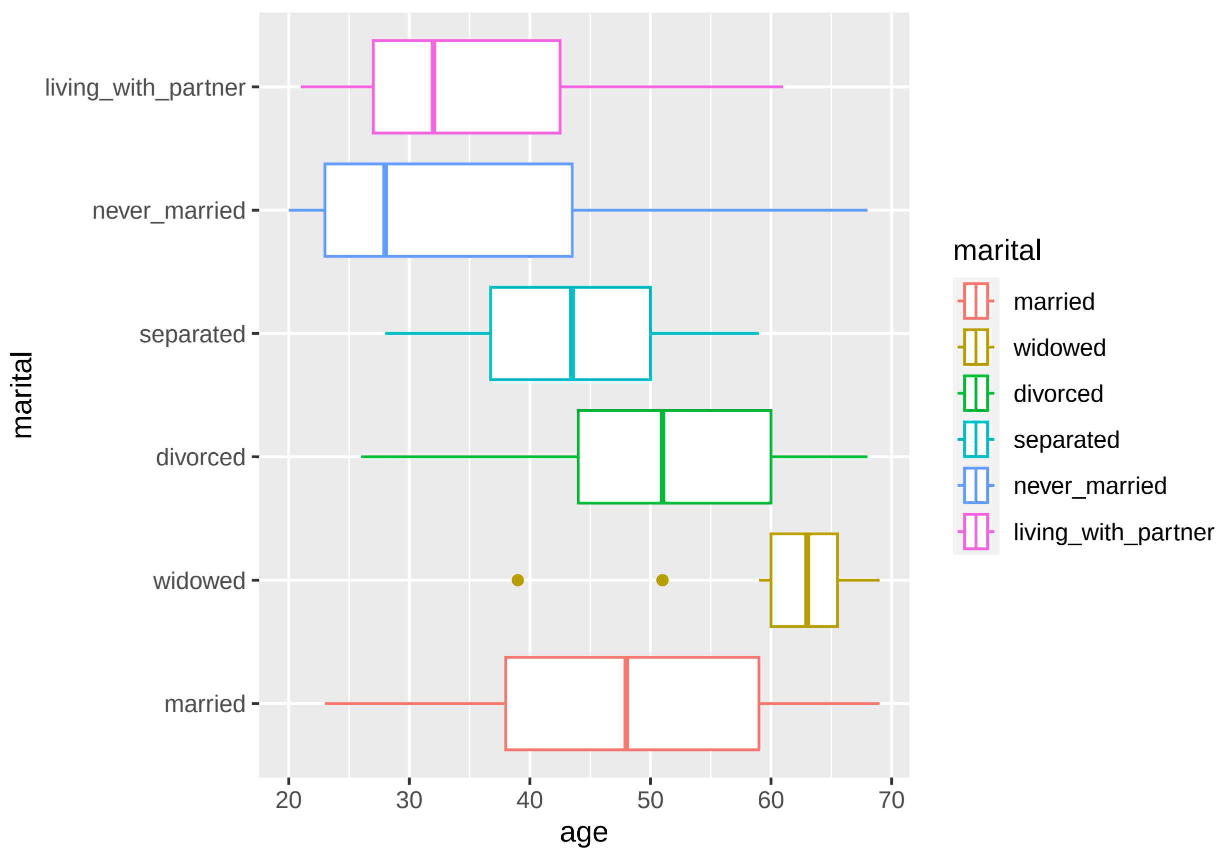


Figure 7: Age distributions over marital statuses

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, it was 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 an 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.

4 Missing data problem

4.1 Missing data and response patterns

Firstly, the overall distribution of missingness within the dataset was investigated.

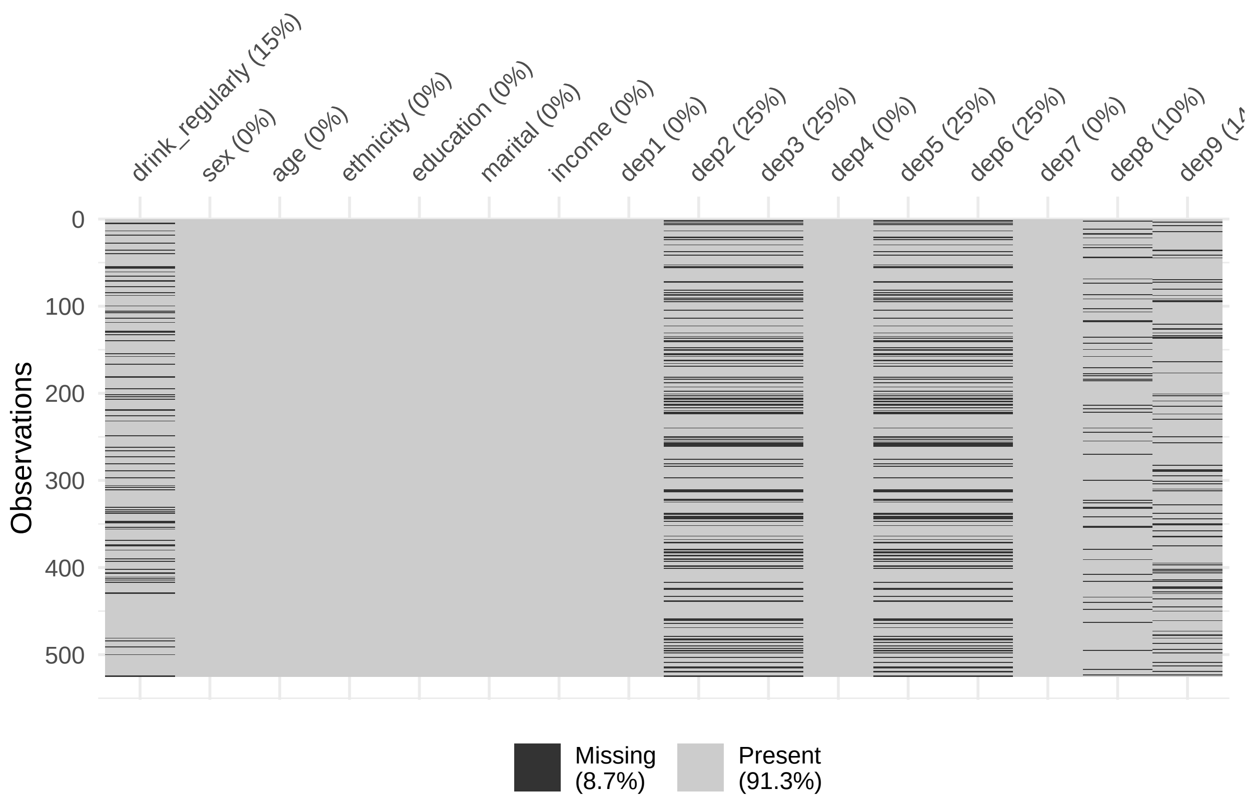


Figure 8: Distribution of missing data in each variable

As can be seen on Figure 8, 8.2% of the data was missing. The missing values occurred in the outcome variable drink_regularly and in the responses to questions dep2, dep3, dep5, dep6, dep8 and dep9 that created the depression score variable. 15% of responses were missing for the predictor variable. 25% of the responses were missing for the individual depression questions 2, 3, 5 and 6, whilst 10% were missing for question 8. Finally, 14% of data was missing for question 9.

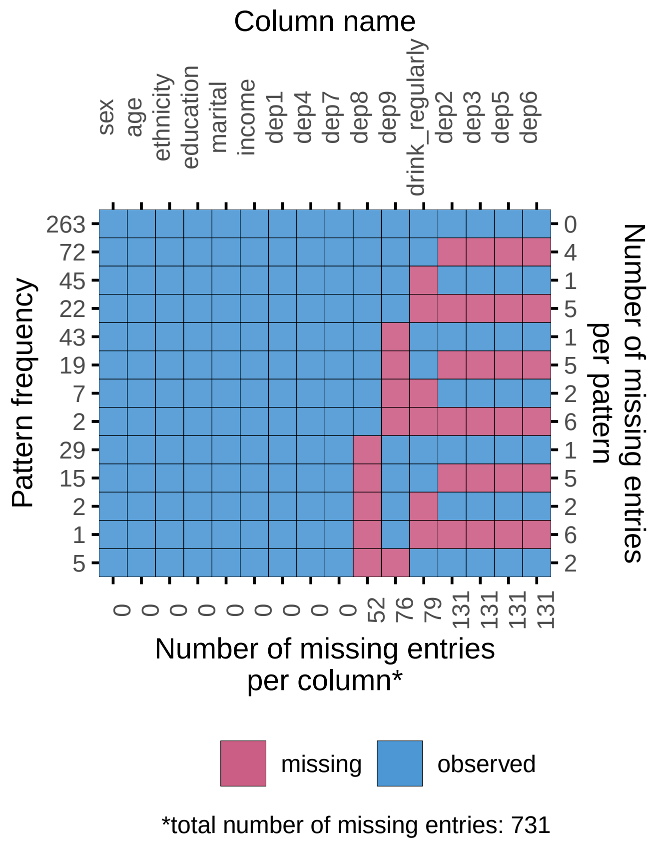


Figure 9: Response patterns and their frequency

The missing data patterns were further investigated by looking at the response patterns. Figure 9 reveals that there were 13 distinct response patterns in the dataset, with the missingness *not* being monotone. The most frequent pattern was no missing entries, with 263 cases. It is important to note that the depression questions 2, 3, 5 and 6 were always either all present or all missing. It is very probable that the reason for the item non-response regarding these depression levels was the same, since there were no cases with partial missingness in these 4 variables.

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

4.2 Missing data mechanism

Missing completely at random (MCAR) missingness mechanism is often an important assumption for statistical analysis, including this one. To gain some inside whether the data was MCAR or not, we deployed a variety of tests. If missing values of a variable was MCAR, said variable should not have a significant dependency with other (observed) variables.

Firstly, Little MCAR test was performed to verify if the missing data was MCAR at a global level, thus for all of the instances of missingess. The test was significant ($\chi^2(164) = 465.18, p < 0.01$), therefore the data was assumed to *not* be 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 were 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 suggests 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 was impossible to test the dependency of missing values and observed values within the same variable, thus these values were missing from the table.

Assuming an alpha of 0.05, it was observed that missing values of `drink_regularly` were dependent on `age`, `dep2` and `dep9`. Similarly, `dep2`, `dep3`, `dep5` and `dep6` seemed to be dependent on `age` and the remaining depression questions. For missingness in `dep8` and `dep9`, there were no significant differences across groups in any of the numerical variables. These tests suggested that at least `drink_regularly`, `dep2`, `dep3`, `dep5` and `dep6` were not MCAR, thus MAR was 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 was not appropriate for categorical variables, we performed the Chi-squared test for said variables. The null hypothesis of the tests entailed no significant relationships between the categorical variables. The test compared 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` were significant. Thus, it further supported the non-MCAR behaviour of missing data in these items.

However, the Chi-squared test had an requirement that the smallest expected frequencies had to be higher than 5. This requirement was 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 was not met for these combination, the p-values might not have been correctly estimated. Therefore, the Fischer test - which has the same hypothesis, but not the same assumption - was 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` were significant. Therefore, it was concluded the dependence of missingness in these variables on multiple other observed variables and rejected MCAR in their case.

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

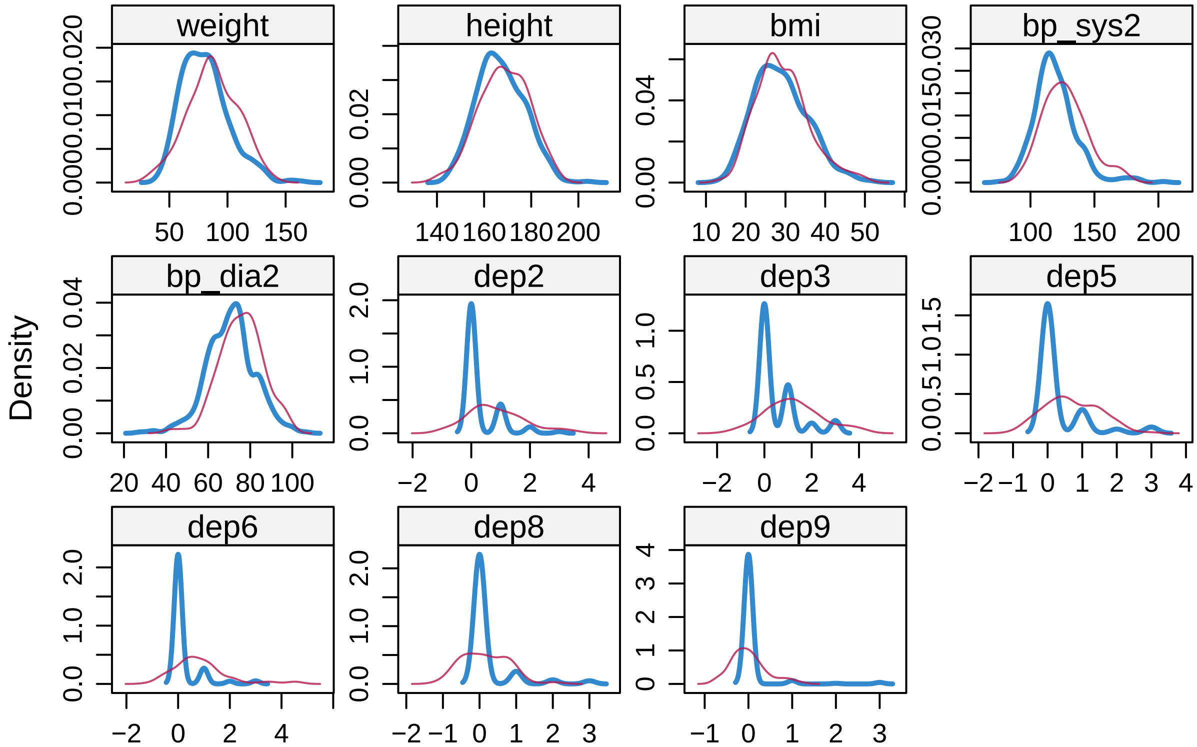
5 Modelling

5.1 Interpretation of the two models

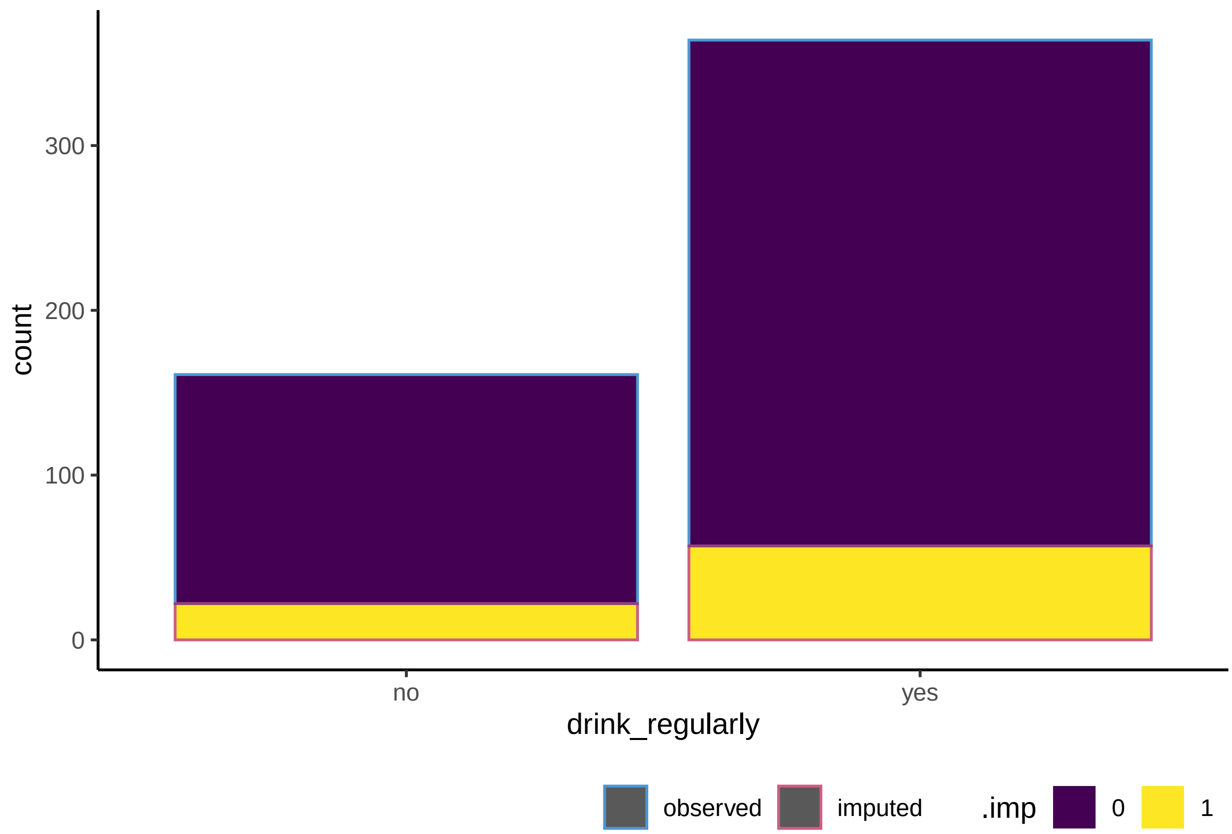
TODO: determine derived variable imputation (in our case the sum of all the depression levels).

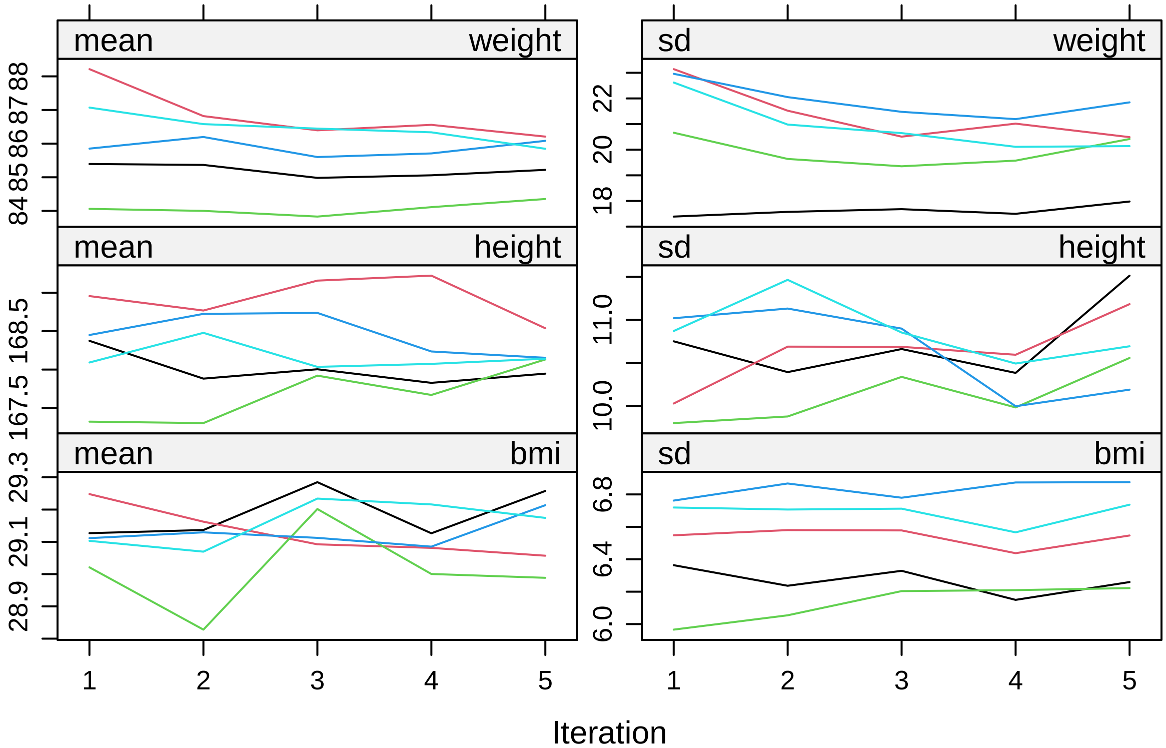
```
imp_single$method
```

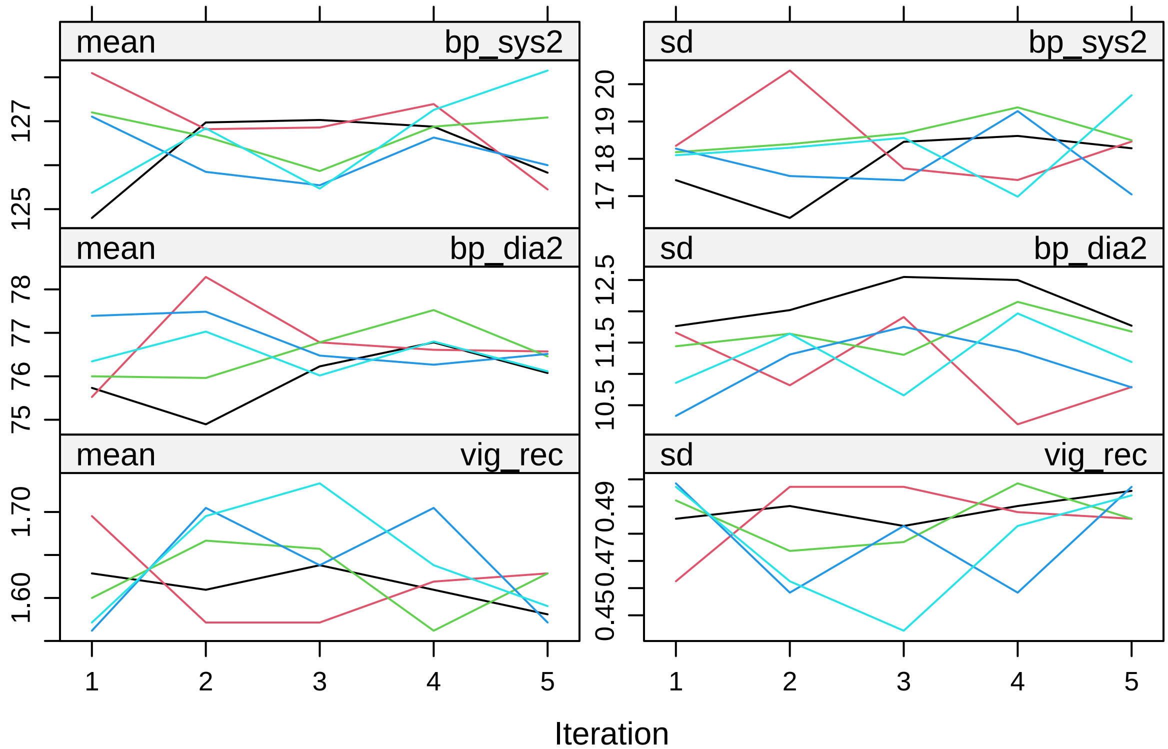
```
##               sex               age         ethnicity         education
##                ""                ""                ""                ""
##           marital    household_size            income            weight
##                ""                ""                ""         "norm.nob"
##            height               bmi             pulse           bp_sys1
##        "norm.nob"        "norm.nob"                ""                ""
##           bp_dia1           bp_sys2           bp_dia2          vig_work
##                ""        "norm.nob"        "norm.nob"                ""
##          mod_work        walk_cycle           vig_rec           mod_rec
##                ""                ""          "logreg"          "logreg"
##          time_sed        n_sex_year n_unsafe_sex_year   drink_regularly
##                ""                ""         "polyreg"          "logreg"
##     days_drinking              dep1              dep2              dep3
##                ""                ""        "norm.nob"        "norm.nob"
##              dep4              dep5              dep6              dep7
##                ""        "norm.nob"        "norm.nob"                ""
##              dep8              dep9
##        "norm.nob"        "norm.nob"
```

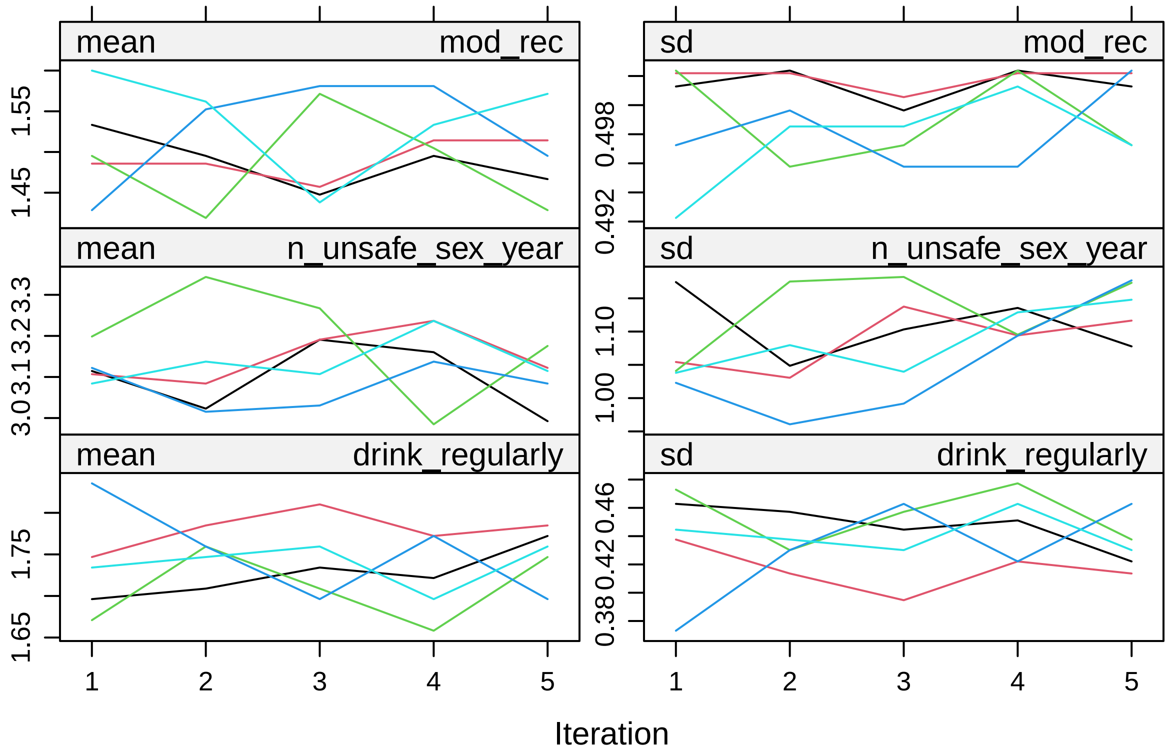


Density
weight
height
bmi
bp_sys2
bp_dia2
dep2
dep3
dep5
dep6
dep8
dep9

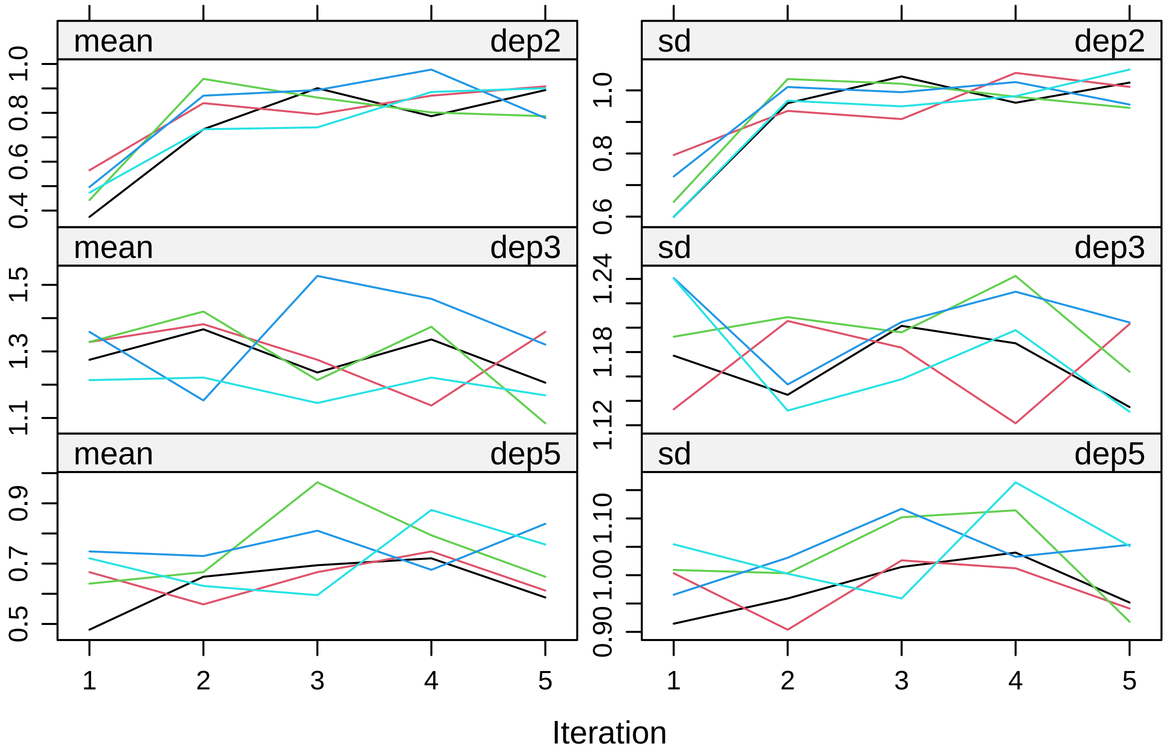

count
300
200
100
0
no
yes
drink_regularly
observed
imputed
.imp
0
1

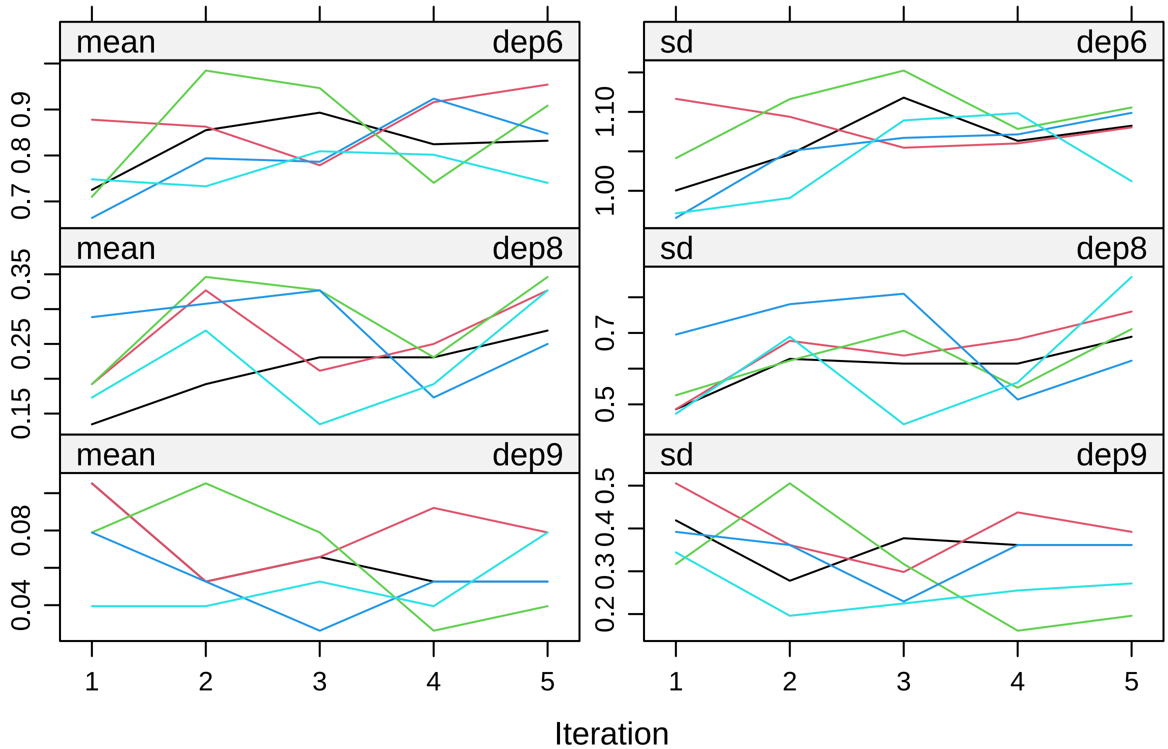

mean
weight
sd
weight
mean
height
sd
height
mean
bmi
sd
bmi
84 85 86 87 88
18
20
22
167.5
168.5
10.0
11.0
28.9
29.1
29.3
6.0
6.4
6.8
1
2
3
4
5
1
2
3
4
5
Iteration

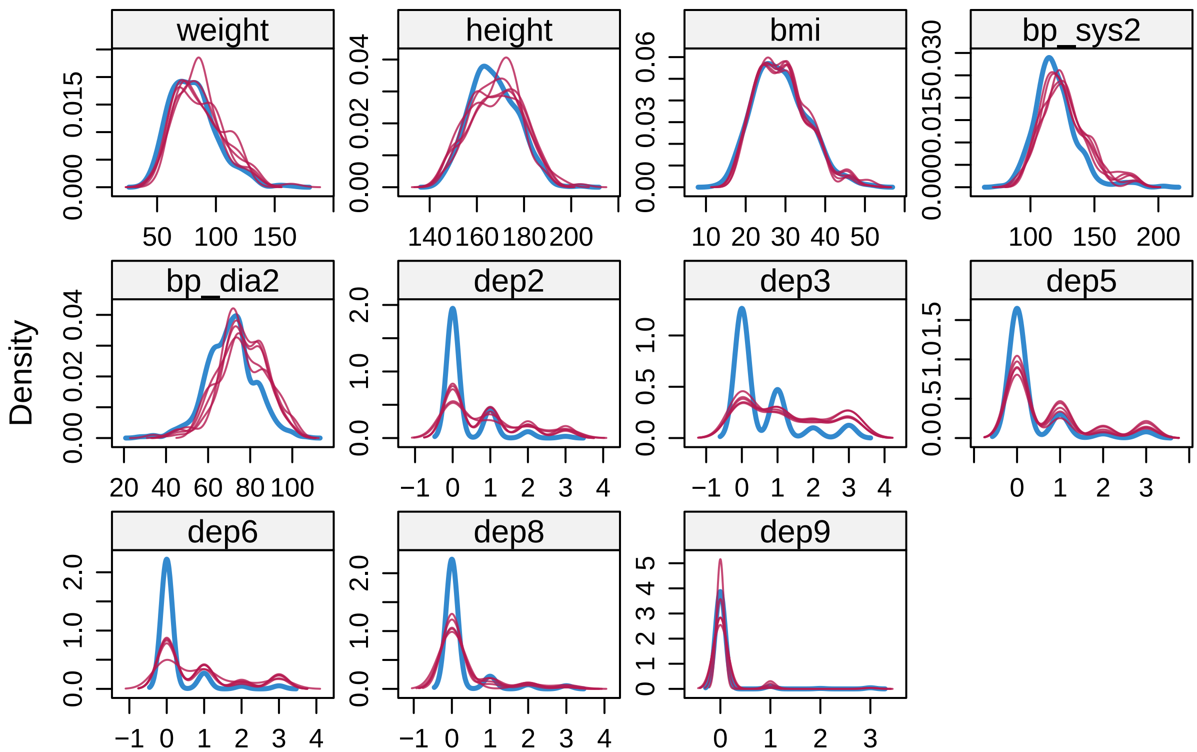

mean
bp_sys2
sd
bp_sys2
mean
bp_dia2
sd
bp_dia2
mean
vig_rec
sd
vig_rec
Iteration



mean
mod_rec
sd
mod_rec
mean
n_unsafe_sex_year
sd
n_unsafe_sex_year
mean
drink_regularly
sd
drink_regularly
1.45
1.55
0.492
0.498
3.0
3.1
3.2
3.3
1.00
1.10
1.65
1.75
0.38
0.42
0.46
1
2
3
4
5
Iteration


mean
dep2
sd
dep2
mean
dep3
sd
dep3
mean
dep5
sd
dep5
0.4 0.6 0.8 1.0
1.1 1.3 1.5
0.5 0.7 0.9
0.6 0.8 1.0
1.12 1.18 1.24
0.90 1.00 1.10
1 2 3 4 5
1 2 3 4 5
Iteration


mean
dep6
sd
dep6
mean
dep8
sd
dep8
mean
dep9
sd
dep9
0.7 0.8 0.9
1.00
1.10
0.15
0.25
0.35
0.5
0.7
0.04
0.08
0.2 0.3 0.4 0.5
1
2
3
4
5
Iteration


weight
height
bmi
bp_sys2
bp_dia2
dep2
dep3
dep5
dep6
dep8
dep9
Density

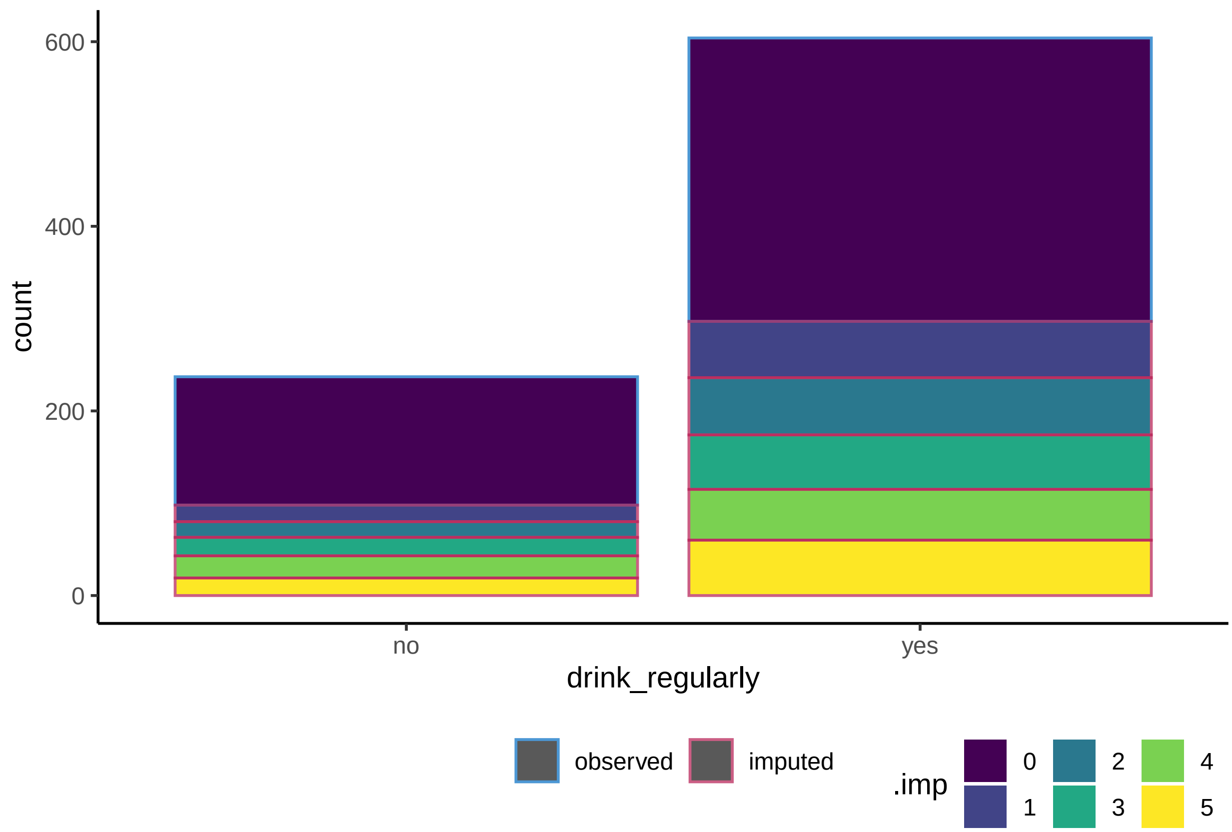


Table 7: Model fit statistics

model	AIC	BIC	Residual_Deviance
listwise deletion	342.6925	442.7129	286.6925
mean imputation	576.6865	696.0617	520.6865

Two logistic models were created, one for each of the imputed datasets that resulted from both the ad-hoc mean imputation and listwise deletion missing data methods. The interpretation below only includes significant predictors, since most of the model's predictors were non-significant. That said, they are still reported in the model summary tables, as can be seen Table 9 and Table 10 in Appendix A.

Looking at the listwise deletion model, only sex ($p < 0.001$) and the marital category `divorced` ($p < 0.001$) were significant predictors, whilst all the other predictors (including the intercept) were insignificant. The odds ratio of females drinking regularly compared to males was 0.16[0.08, 0.31], meaning they were 0.16 times as likely to drink regularly compared to males. The odds ratio of `divorced` respondents compared to those who were `married` was 7.37[2.54, 25.0], meaning divorced people were 7.37 times as likely to drink regularly compared to married people.

In the mean imputation model, `age` ($p < 0.001$) was also a significant predictor. This was in addition to `sex` ($p < 0.001$) and marital status `divorced` ($p < 0.004$). Moreover, the `intercept` ($p < 0.001$) was also significant. For a unit increase in age the odds ratio of drinking regularly decreases 0.97[0.95, 0.99] times. The odds ratio of females drinking regularly compared to males was 0.24[0.14, 0.38], meaning they were 0.24 times as likely to drink compared to males. The odds ratio of people with the marital status `divorced` compared to people with the marital status `married` was 3.05[1.47, 6.69], meaning divorced people were 3.05 times as likely to drink regularly compared to married people. If all predictor variables were in their reference state, the baseline odds ratio was 20.5[3.40, 136]. Looking at the model fit statistics, the listwise deletion model performed better with a lower AIC, BIC and residual deviance value, as is shown in Table 7.

5.2 Results comparison ad-hoc methods

Following Section 5.1, the answer to the research question will differ depending on the ad-hoc method used to treat the missing data. Said difference is comprised of: the number of significant predictors, the odds ratios, standard errors and the fitness statistics. The SE of the mean imputation model was lower than the listwise deletion model, which can be see in Table 10 and 9.

It was concluded that the missing data treating methods caused biased results within each model, since these methods assume the missingness to be MCAR. This assumption does not hold, in case the data is assumed to be MAR - as was described in Section 4.2. In the specific case of logistic regression - which was utilised in this study - listwise deletion could still give unbiased results even if the missingness isn't MCAR, but only if the missing values are either in the predictor variables or in the outcome variable. From the EDA and missing data specification it was observed that both predictor *and* outcome variables contained missing data, hence the results were biased as well.

When listwise deletion is used on a dataset where the missingness isn't MCAR, it will result in a bias in the regression coefficients, with standard errors that are too large. As the models in this study were logistic regression models, the regression coefficients were transformed into odds ratios instead. Another negative of listwise deletion is that it is wasteful, a lot of data goes unused. This is also the case in this study, as 263 out of a total of 525 cases were deleted or in other words only half the cases were used. This might explain why the listwise deletion model has a lower D value; there is just a lot less data that might not fit in the model. Another consequence of removing so much data from the dataset is that it becomes more difficult to find effects, this is a possible explanation for why the listwise deletion model had less significant predictors.

Using mean imputation on a dataset where the missingness isn't MCAR will give biased results and underestimated standard errors. It is indeed the case that the standard errors in the mean imputation model are lower than those in the listwise deletion model. As listwise deletion usually overestimates the standard errors, the real values are probably somewhere in between.

So as both methods were used on data where the missingness wasn't MCAR, the created models were biased in multiple ways.

6 Conclusion

The research question was **To what extent can the occurrence of regular alcohol consumption (12 or more in a year) be predicted by the variables: depression level, age, sex, ethnicity, marital status and household income?** Here, it was expected that sex `male`, ethnicity `non-hispanic_white` and marital status `married` would exhibit a positive relation with drinking regularly, whilst age and depression would have a negative relation with said outcome variable.

The resulting listwise deletion model - using a logistic regression approach - only had `sex` and marital status `divorced` (in reference to `married`) as significant predictors. Using this ad-hoc method, no significant relation between the predictors `depression` and `ethnicity` and the outcome variable `drink_regularly` could be deduced. Additionally, `divorced` was positively related to drinking regularly, which goes against the expectation of categories other than `married` exhibiting a negative relation with the outcome. The model did suggest that `male` respondents are more likely to drink regularly than `female` participants, and can therefore be used to predict alcohol consumption.

As for the mean imputation model, the results similarly suggested non-significant predictors for `ethnicity` and `depression`. `age` was significant, as opposed to in the listwise deletion model - this was in addition to the other significant predictors, those being the married status `divorced` and `sex`. Once again, the model aligned with expectations of `male` respondents drinking more regularly. `age` also showed a negative relation with the outcome, albeit a small decrease per unit change of `age`. Similar to the previous model, `divorced` showed an unexected positive behaviour with the outcome variable.

Overall, it was concluded from both models that `depression` and `ethnicity` cannot be used to predict whether a case is drinking regularly. Marital status is a valid predictor, although its relation does not align with research previously done by others. Whether `age` can be used as a predictor, depends on the ad-hoc method used to treat missing data. Lastly, the coefficients of significant predictors in both models differed greatly. The coefficient of `divorced` in the listwise deletion model exceeded the mean imputation variant twofold.

7 References

Garnett, Claire, Sabrina Kastaun, Jamie Brown, and Daniel Kotz. 2022. "Alcohol Consumption and Associations with Sociodemographic and Health-Related Characteristics in Germany: A Population Survey." *Addictive Behaviors* 125 (February): 107159. https://doi.org/10.1016/j.addbeh.2021.107159.

GBD 2016 Alcohol Collaborators. 2018. "Alcohol Use and Burden for 195 Countries and Territories, 1990–2016: A Systematic Analysis for the Global Burden of Disease Study 2016." *Lancet* 392 (10152): 1015–35. https://doi.org/10.1016/s0140-6736(18)31310-2.

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.

Moore, Alison A., Robert Gould, David B. Reuben, Gail A. Greendale, M. Kallin Carter, Kefei Zhou, and Arun Karlamangla. 2005. "Longitudinal Patterns and Predictors of Alcohol Consumption in the United States." *American Journal of Public Health* 95 (3): 458–64. https://doi.org/10.2105/ajph.2003.019471.

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 8: Summary statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
drink_regularly	446						
... no	139	31%					
... yes	307	69%					
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

Table 9: Listwise deletion model

Characteristic	**OR**	**SE**	**p-value**
(Intercept)	6.71	1.28	0.14
sex			
male	—	—	
female	0.16***	0.350	<0.001
age	1.00	0.013	0.8
ethnicity			
mexican_american	—	—	
other_hispanic	1.49	0.594	0.5
non-hispanic_white	1.58	0.435	0.3
non-hispanic_black	0.55	0.485	0.2
other	1.04	0.792	>0.9
education			
no_high_school	—	—	
some_high_school	0.82	0.582	0.7
high_school_grad	1.30	0.589	0.7
some_college	1.05	0.604	>0.9
college_grad	1.66	0.647	0.4
marital			
married	—	—	
widowed	1.37	0.711	0.7
divorced	7.37***	0.578	<0.001
separated	1.98	1.14	0.5
never_married	1.34	0.435	0.5
living_with_partner	4.26	0.862	0.093
income			
0:4999	—	—	
5000:9999	0.17	1.24	0.15
10000:14999	0.43	1.07	0.4
15000:19999	0.43	1.05	0.4
20000:24999	0.40	1.04	0.4
25000:34999	0.65	1.04	0.7
35000:44999	0.79	1.07	0.8
45000:54999	0.40	1.12	0.4
55000:64999	1.06	1.13	>0.9
65000:74999	0.36	1.11	0.4
75000:99999	0.56	1.13	0.6
100000+	0.79	1.06	0.8
dep	1.03	0.051	0.6

[1] *p<0.05;* ***p<0.01;*** p<0.001

[2] OR = Odds Ratio, SE = Standard Error

Table 10: Mean imputation model

Characteristic	**OR**	**SE**	**p-value**
(Intercept)	20.5**	0.933	0.001
sex			
male	—	—	
female	0.24***	0.247	<0.001
age	0.97***	0.009	<0.001
ethnicity			
mexican_american	—	—	
other_hispanic	1.19	0.411	0.7
non-hispanic_white	1.42	0.338	0.3
non-hispanic_black	0.57	0.367	0.12
other	0.59	0.543	0.3
education			
no_high_school	—	—	
some_high_school	1.16	0.428	0.7
high_school_grad	1.48	0.422	0.4
some_college	1.22	0.420	0.6
college_grad	1.98	0.479	0.2
marital			
married	—	—	
widowed	0.85	0.547	0.8
divorced	3.05**	0.385	0.004
separated	1.36	0.678	0.7
never_married	1.22	0.335	0.5
living_with_partner	2.28	0.508	0.10
income			
0:4999	—	—	
5000:9999	0.64	0.825	0.6
10000:14999	0.92	0.756	>0.9
15000:19999	0.40	0.738	0.2
20000:24999	0.51	0.732	0.4
25000:34999	1.20	0.752	0.8
35000:44999	0.87	0.753	0.9
45000:54999	0.96	0.791	>0.9
55000:64999	0.93	0.785	>0.9
65000:74999	0.60	0.790	0.5
75000:99999	1.19	0.785	0.8
100000+	1.18	0.752	0.8
dep	1.02	0.035	0.6

[1] *p<0.05;* ***p<0.01;*** p<0.001

[2] OR = Odds Ratio, SE = Standard Error

B Appendix