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: **Can the occurrence of regular alcohol consumption (12 or more in a year) 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 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 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. 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 testing of the dependencies between missing values in one variable and observed values of other variables and performing the Little (1988) MCAR Test. Since 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 (source). 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.

2.4 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, we 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.

3 EDA Results (Nisse)

3.1 Descriptive statistics

Table 8 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.

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
no	139
yes	307
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.

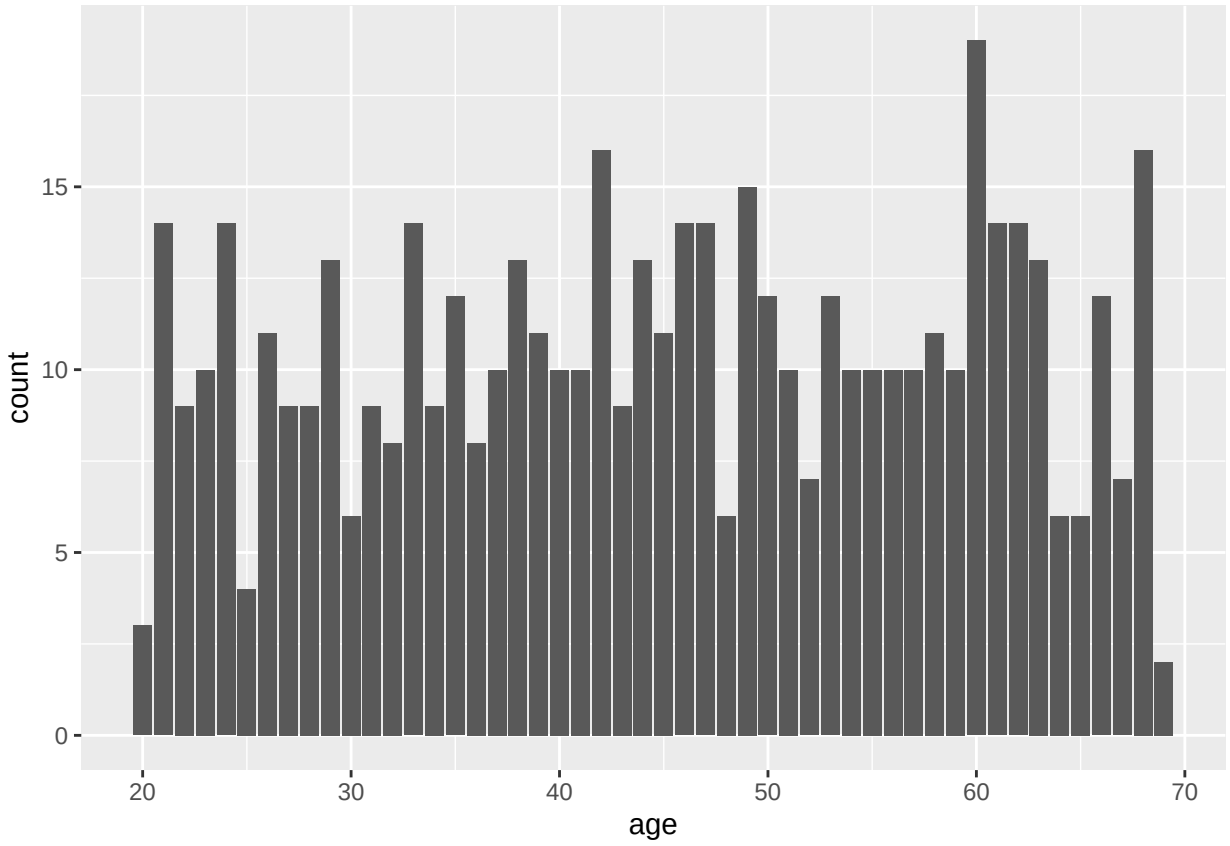


Figure 1: Distribution of age

Predictor `ethnicity` has 5 categories, with `non-hispanic_white` being the most prevalant 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. Prior to combining the `marital` categories, the category value `married` exceeded the the average frequency of other categories (that being 49.2) by a large margin - as can be seen in Figure 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.

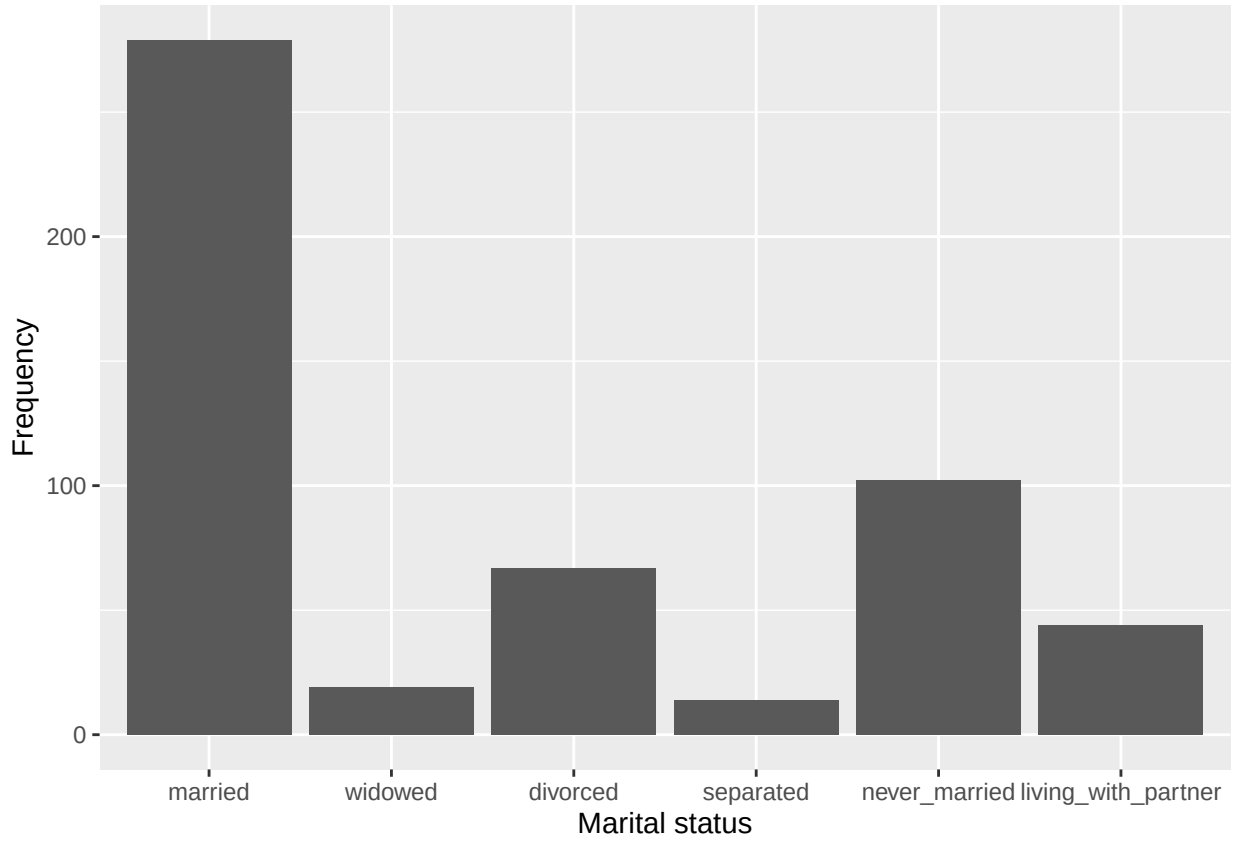


Figure 2: Distribution of marital status categories before pre-processing

3.2 Correlations

```
## Warning in chisq.test(data$marital, data$drink_regularly): Chi-squared
## approximation may be incorrect
```

Table 3: Contigency table of marital vs. drink regularly

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

```
##
##  Pearson's Chi-squared test
##
## data:  data$marital and data$drink_regularly
## X-squared = 10.218, df = 5, p-value = 0.06928
```

```
##                      X^2 df P(> X^2)
## Likelihood Ratio  9.8952  5 0.078259
## Pearson          10.2183  5 0.069280
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.15
## Cramer's V        : 0.151
```

TODO (from pract):

- Variable with most NA's
- Summary of non-NA vars
-

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 3 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.

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 4 , 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:

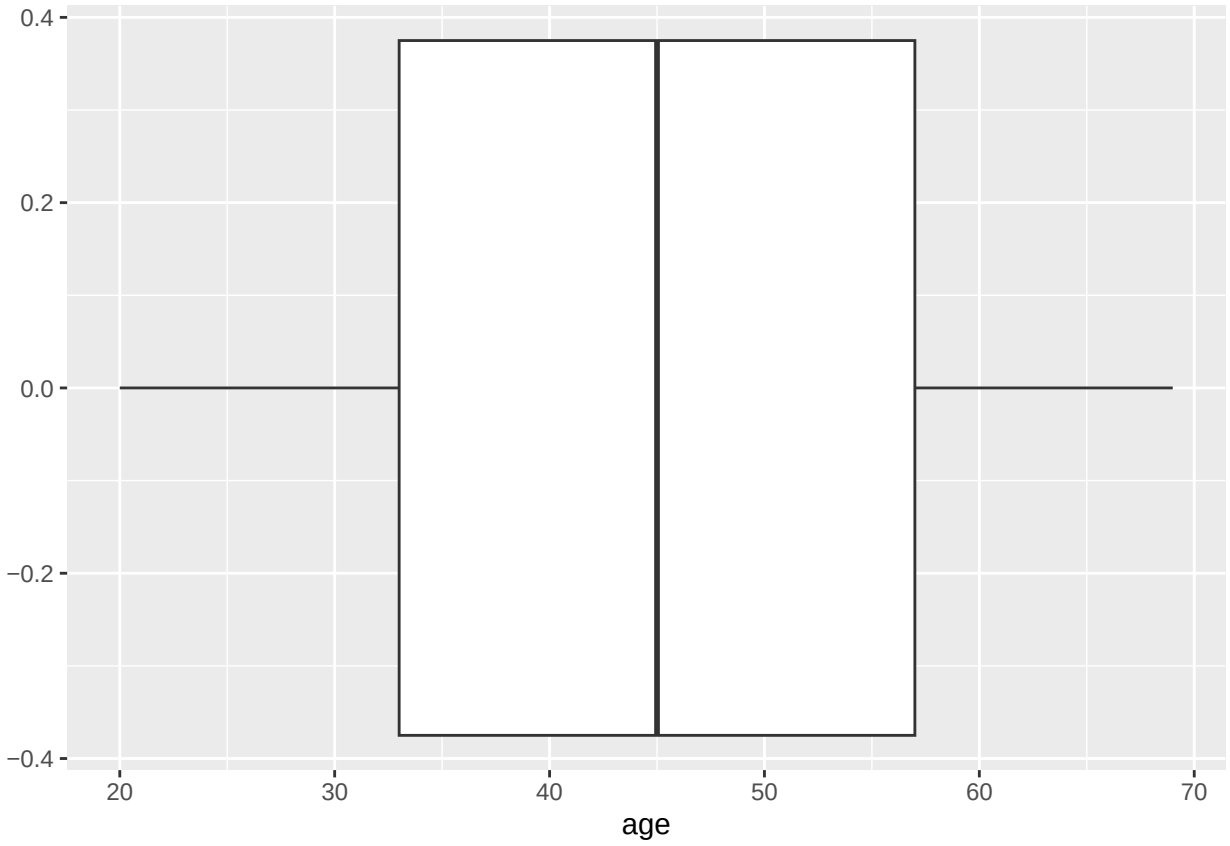


Figure 3: Boxplot of age

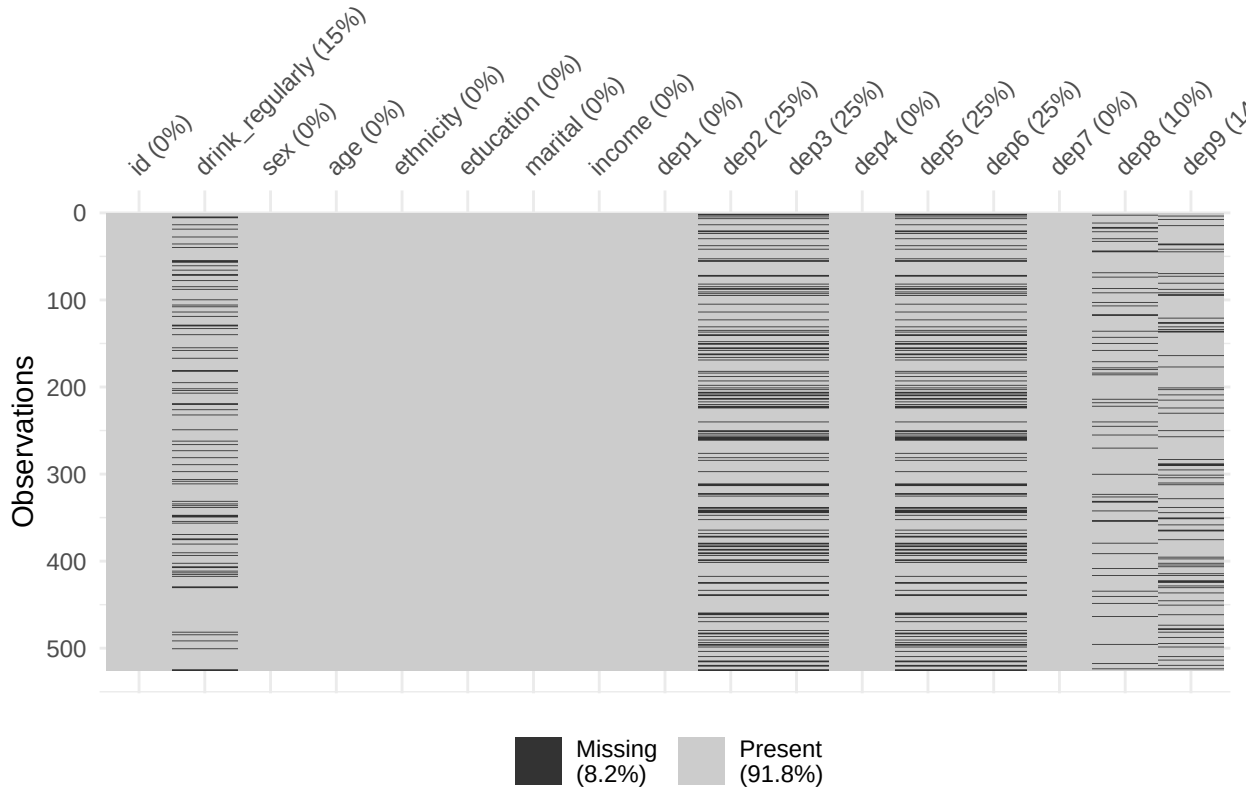


Figure 4: Distribution of missing data in each variable

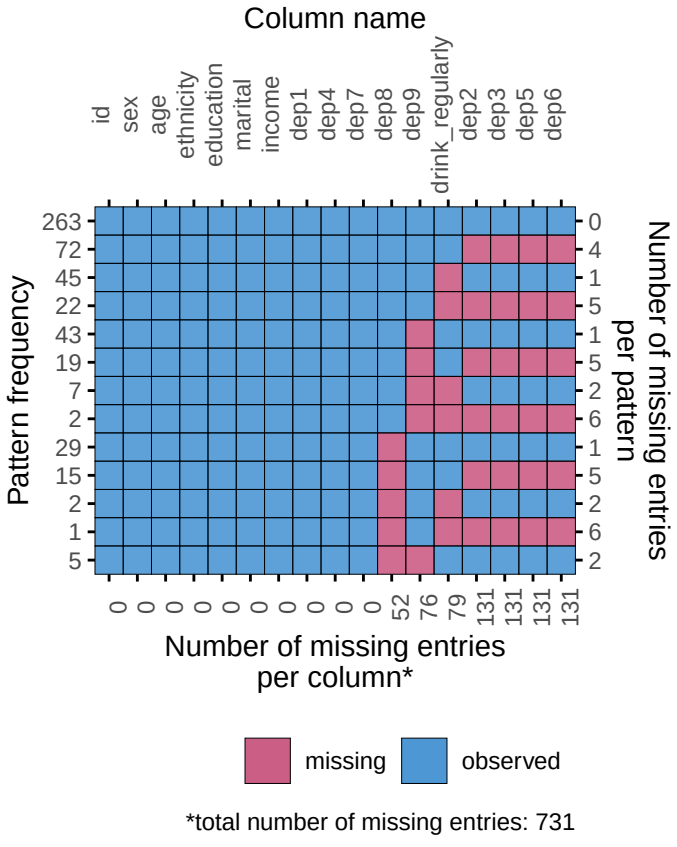


Figure 5: Response patterns and their frequency

Figure 5 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.

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 the table below 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	drink_regularly	dep2/3/5/6	dep8	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	-

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   no :161         male  :254   Min.   :20.00
##  1st Qu.:43912   yes:364         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.   :-0.3542   Min.   :-0.1686
##  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 : 0.1224
##  35000:44999: 51   Mean   :0.4095   Mean   : 0.3510   Mean   : 0.6821
##  75000:99999: 49   3rd Qu.:1.0000   3rd Qu.: 0.7030   3rd Qu.: 1.0000
##  10000:14999: 45   Max.   :3.0000   Max.   : 3.0000   Max.   : 3.1378
##  (Other)    :193
##       dep4             dep5              dep6              dep7
##  Min.   :0.0000   Min.   :-0.3069   Min.   :-0.1745   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   : 0.3590   Mean   : 0.3281   Mean   :0.3238
##  3rd Qu.:1.0000   3rd Qu.: 0.6110   3rd Qu.: 0.4757   3rd Qu.:0.0000
##  Max.   :3.0000   Max.   : 3.0000   Max.   : 3.1091   Max.   :3.0000
##
##       dep8              dep9
##  Min.   :-0.2806   Min.   :-0.38236
##  1st Qu.: 0.0000   1st Qu.: 0.00000
##  Median : 0.0000   Median : 0.00000
##  Mean   : 0.2058   Mean   : 0.06482
##  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.

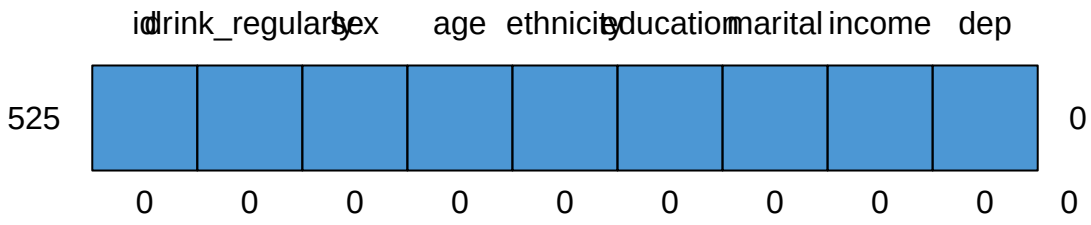
TODO:

- 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.

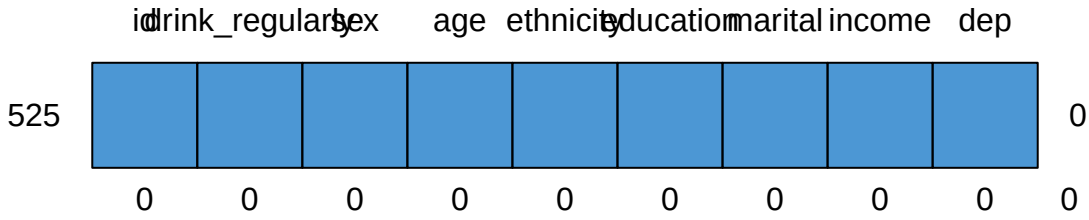
```
##
##  iter imp variable
##   1   1  dep2  dep3  dep5  dep6  dep8  dep9

##  /\     /\
## {  `---'  }
## {  O   O  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   `-----'
```



```
##     id drink_regularly sex age ethnicity education marital income dep
## 525  1               1   1   1         1         1       1      1   1 0
##      0               0   0   0         0         0       0      0   0 0
```

```
##  /\     /\
## {  `---'  }
## {  0   0  }
## ==>  V <==  No need for mice. This data set is completely observed.
##  \  \|/  /
##   `-----'
```




```
##     id drink_regularly sex age ethnicity education marital income dep
## 525  1               1   1   1         1         1       1      1   1 0
##      0               0   0   0         0         0       0      0   0 0
```

Table 5: Mean imputation 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			

Table 5: Mean imputation model *(continued)*

Characteristic	OR	SE	p-value
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 6: 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

Table 6: Mean imputation model *(continued)*

Characteristic	OR	SE	p-value
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

Table 7: Model fit statistics

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

4.4 Interpretation of the two models.

Two logistic models were created, one model with the dataset resulting from using listwise deletion as the method used to treat the missing data and the other model with the dataset resulting from using mean imputation as the method used to treat the missing data. In the interpretation section below only the significant predictors are mentioned as most are non-significant and including them would add an unnecessary amount of text. They are still included in the created models, see table 5 and 6.

Looking at the listwise deletion model, only sex ($p < 0.001$) and the marital status divorced ($p < 0.001$) are significant predictors, all the other predictors and 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 compared to males ($0.16 < 1$ so it less likely). The odds ratio of people with the marital status divorced compared to people with the marital status 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 ($p < 0.001$) is also a significant predictor next to sex ($p < 0.001$) and marital status divorced ($p < 0.004$), the intercept ($p < 0.001$) is 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 is 20.5[3.40, 136]. Looking at the model fit statistics, it seems like the listwise deletion models seems to fit better with a lower AIC, BIC and D value as show in table 7.

4.5 Comparison of the two models in terms the missing data and resulting model

The answer to the research question changes depending on the used method to treat missing data as can be seen in the interpretation section above. The models differ in: the number of significant predictors, the odds ratios, standard errors (these are lower in the mean imputation model, see tables 5 and 6) and the fitness statistics. Both methods used to treat missing data only work correctly if the missingness is MCAR but as the missingness is not MCAR as is explained in section 4, both methods caused bias in the results. In the case of logistic regression (as this study does) listwise deletion could still give unbiased results even if the missingness isn't MCAR if the missing values are either in the predictor variables or in the outcome variable. This isn't the case in the data used in this study, both predictor variables as the outcome variables contain missing values.

When listwise deletion is used on a dataset where the missingness isn't MCAR it will result in bias in the regression coefficients and standard errors that are too large. As the models in this study are logistic regression models the regression coefficients are 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 41% of the cases are deleted. Which 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 has 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 when the missingness wasn't MCAR the created models are biased in multiple ways and thus aren't valid.

4.6 Conclusion in terms of answering RQ (Nisse)

5 References

A Appendix

Table 8: Summary statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
id	525	46470	2898	41531	43912	48934	51610
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