Diet, Labels, and Lifestyle: A NHANES Analysis

2025-03-02

Table of contents

1 Setup and Data Ingest **3**
- 1.1 Initial Setup and Loading Package . . . 3
- 1.2 Loading the Raw Data . . . 3
- 1.3 Contents of the Raw Tibbles . . . 4
- 1.4 Merging the Data . . . 5
- 1.5 Checking the Merge . . . 5
- 1.6 Selecting the Variables . . . 5

2 Cleaning the Data **6**
- 2.1 The Variables . . . 6
  - 2.1.1 Identification and Population Variables . . . 6
  - 2.1.2 Quantitative Variables . . . 6
  - 2.1.3 Binary Variables . . . 6
  - 2.1.4 Multi-Categorical Variables . . . 6
- 2.2 Labeling Factors . . . 6
- 2.3 Filtering for Population of Interest . . . 7
- 2.4 Checking the Quantitative Variables . . . 7
- 2.5 Checking Binary Variables . . . 7
  - 2.5.1 The `RIDSTATR` Variable . . . 8
  - 2.5.2 The `RIAGENDR` Variable . . . 8
  - 2.5.3 The `ALQ151` Variable . . . 9
- 2.6 Checking Multi-Categorical Variables . . . 9
  - 2.6.1 The `DBQ750` Variable . . . 9
  - 2.6.2 The `DMDEDUC2` Variable . . . 10
- 2.7 Creating the Analytic Tibble . . . 11
- 2.8 Listing of Missing Values . . . 11

3 Codebook and Data Description **12**
- 3.1 Codebook . . . 12
- 3.2 Analytic Tibble . . . 12
- 3.3 Data Summary . . . 13

4 Imputation **14**

5 Analysis 1: Difference in Protein Consumption Among Men and Women **20**
5.1 The Question . . . 20
5.2 Describing The Data . . . 21
5.2.1 Graphical and Numerical Summaries . . . 21
5.2.2 BoxCox for Transformation . . . 24
5.2.3 Transformation of the Data . . . 25
5.2.4 Variance and Normality Checks . . . 26
5.3 Main Analysis . . . 27
5.4 Conclusions . . . 28

6 Analysis 2: Comparing 5 Population Means with Independent Samples **29**
6.1 The Question . . . 29
6.2 Describing The Data . . . 29
6.2.1 Numerical Summary . . . 29
6.2.2 Graphical Summaries . . . 30
6.2.3 Exploring a Transformation . . . 32
6.2.4 Transforming the Data . . . 32
6.2.5 Variance Check . . . 34
6.3 Main Analysis . . . 35
6.3.1 Analysis of Variance . . . 35
6.4 Conclusions . . . 37

7 Analysis D: Drinking Habitis between Genders **37**
7.1 The Question . . . 37
7.1.1 The 2x2 Table . . . 38
7.1.2 Checking Assumptions . . . 38
7.2 Main Analysis . . . 39
7.3 Conclusions . . . 40

8 Analysis 4: **41**
8.1 The Question . . . 41
8.2 Describing The Data . . . 41
8.2.1 The 5x5 Table . . . 41
8.3 Main Analysis . . . 42
8.3.1 Creating the Table for Analysis . . . 42
8.3.2 The Pearson χ^2 Test . . . 43
8.3.3 The Cochran Conditions . . . 43
8.3.4 The Association Plot and Table for the 5x5 Table . . . 44
8.4 Conclusions . . . 46

9 Session Information **46**

1 Setup and Data Ingest

1.1 Initial Setup and Loading Package

```
knitr::opts_chunk$set(warning = FALSE)
```

```
library(haven)
library(mice)
library(Hmisc)
library(MKinfer)
library(janitor)
library(tinytex)
library(naniar)
library(Epi)
library(epiR)
library(FSA)
library(sessioninfo)
library(car)
library(mosaic)
suppressPackageStartupMessages(library(here))
library(gt)
library(glue)
library(patchwork)
library(broom)
library(rstatix)
library(tidyverse)


theme_set(theme_bw())
knitr::opts_chunk$set(comment = NA, warning = FALSE)
options(dplyr.summarise.inform = FALSE)
```

1.2 Loading the Raw Data

```
# Load and Save P_DEMO (Demographics Data)
demo_raw <- read_xpt(here("data", "raw", "P_DEMO.XPT"))
saveRDS(demo_raw, here("data", "processed", "P_DEMO.Rds"))
demo_raw <- readRDS(here("data", "processed", "P_DEMO.Rds"))

# Load and Save P_DR2TOT (Protein Consumption Data)
protein_raw <- read_xpt(here("data", "raw", "P_DR2TOT.XPT"))
saveRDS(protein_raw, here("data", "processed", "P_DR2TOT.Rds"))
protein_raw <- readRDS(here("data", "processed", "P_DR2TOT.Rds"))

# Load and Save P_CBQPFA (Nutrition Label Use Data)
```

```r
nut_raw <- read_xpt(here("data", "raw", "P_CBQPFA.XPT"))
saveRDS(nut_raw, here("data", "processed", "P_CBQPFA.Rds"))
nut_raw <- readRDS(here("data", "processed", "P_CBQPFA.Rds"))

# Load and Save P_BMX (BMI Data)
bmi_raw <- read_xpt(here("data", "raw", "P_BMX.XPT"))
saveRDS(bmi_raw, here("data", "processed", "P_BMX.Rds"))
bmi_raw <- readRDS(here("data", "processed", "P_BMX.Rds"))

# Load and Save P_ALQ (Alcohol Consumption Data)
alc_raw <- read_xpt(here("data", "raw", "P_ALQ.XPT"))
saveRDS(alc_raw, here("data", "processed", "P_ALQ.Rds"))
alc_raw <- readRDS(here("data", "processed", "P_ALQ.Rds"))
```

1.3 Contents of the Raw Tibbles

Five tibbles were successfully loaded, each containing a subset of variables that will be used in the final analysis. Below, we verify the dimensions of each dataset to ensure they contain the expected number of observations and variables.

```r
dim(demo_raw)
```

```
[1] 15560    29
```

The `demo_raw` contains 15560 rows and 29 variables.

```r
dim(protein_raw)
```

```
[1] 14300    85
```

The dataset `protein_raw` contains 14300 rows and 85 variables.

```r
dim(nut_raw)
```

```
[1] 8552   66
```

The dataset `nut_raw` contains 8552 rows and 66 variables.

```r
dim(bmi_raw)
```

```
[1] 14300    22
```

The dataset `bmi_raw` contains 14300 rows and 22 variables.

```r
dim(alc_raw)
```

```
[1] 8965   10
```

The dataset `alc_raw` contains 8965 rows and 10 variables.

1.4 Merging the Data

To ensure all variables are stored within a single tibble, the data sets will be merged by their common column, `SEQN` which is a unique numerical identifier for each subject.

```
comdata <- left_join(demo_raw, protein_raw, by = "SEQN")
comdata2 <- left_join(comdata, nut_raw, by = "SEQN")
comdata3 <- left_join(comdata2, bmi_raw, by = "SEQN")
comdata4 <- left_join(comdata3, alc_raw, by = "SEQN")

dim(comdata4)
```

```
[1] 15560   208
```

Previously, the data sets contained a total of 212 variables (29 + 85 + 66 + 22 + 10). After merging, the final dataset contains 208 variables. This reduction occurs because `SEQN` is present in all five data sets and is used as the key to merge them. Since each merge retains only one instance of `SEQN`, a total of four duplicate columns are removed, leading to (212 - 4) = 208 columns in the final merged dataset.

1.5 Checking the Merge

To verify that each row represents a unique individual, we will check whether the number of unique `SEQN` identifiers matches the total number of rows in the merged dataset.

```
identical(n_distinct(comdata4$SEQN), comdata4 %>% nrow())
```

```
[1] TRUE
```

If the output returns TRUE, this confirms that each row correctly corresponds to one unique `SEQN` value, ensuring there are no duplicate subject entries.

1.6 Selecting the Variables

The merged dataset, `comdata4`, contains over 200 variables, but not all will be used in the final analysis. We will retain only those variables that are essential for analysis, subject identification, and further data cleaning.

```
NH_data <- comdata4 %>% select(SEQN, RIDAGEYR, RIDSTATR, RIAGENDR, DR2TPROT, DBQ750, BMXBMI,
```

We have selected the variables:

- `SEQN` - Unique identifier for each subject.
- `RIDAGEYR` and `RIDSTATR` - Used to define the population of interest.
- `RIAGENDR`, `DR2TPROT`, `DBQ750`, `BMXBMI`, `DMDEDUC2`, `ALQ151` - Variables for analysis

2 Cleaning the Data

2.1 The Variables

There are nine variables that will be used.

2.1.1 Identification and Population Variables

1. `SEQN` - The unique identification number assigned to each subject.

2.1.2 Quantitative Variables

2. `RIDAGEYR` - The subjects age, in years
3. `DR2TPROT` – The subject's total protein consumption, measured in grams.
4. `BMXBMI` - The subjects body mass index, in kg/m^2

2.1.3 Binary Variables

5. `RIDSTATR` – The subject's examination and interview status: (Interview Only, Interview and examination Complete)
6. `RIAGENDR` - The subjects gender (Male, female)
7. `ALQ151` – Response to the question: "*Have you ever in a point in your life had 4 or 5 drinks every day?*" (Yes, No)

2.1.4 Multi-Categorical Variables

8. `DBQ750` – How often the subject reads nutrition labels before purchasing food (Always, Most of the time, Sometimes, Rarely, Never)
9. `DMDEDUC2` - The subject's highest level of education completed ("Less Than 9th Grade, 9-11th Grade, High School/GED or Equivalent, Some College or AA degree, College Graduate or above)

2.2 Labeling Factors

All categorical variables (`RIAGENDR`, `RIDSTATR`, `DMDEDUC2`, `DBQ750`, `ALQ151`) will be converted from numeric values to factors for better interpret ability. Additionally, any categorical variable with values 7 or 9 (indicating "I don't know" or refusal to answer) will be treated as missing values.

```
NH_data <- NH_data %>% mutate(SEQN = as.character(SEQN),
                  RIAGENDR = factor(RIAGENDR, levels = c(1,2), labels = c("Male", "Female")),
                  RIDSTATR = factor(RIDSTATR, levels = c(1, 2), labels = c("Missing Examinati
                  DMDEDUC2 = factor(DMDEDUC2, levels = c(1,2,3,4,5), labels = c("Less Than 9t
                  DBQ750 = factor(DBQ750, levels = c(1,2,3,4,5), labels = c("Always", "Most o
                  ALQ151 = factor(ALQ151, levels = c(1,2), labels = c("Yes", "No")))
```

`SEQN` was converted to a character type to maintain it as a unique identifier rather than a numeric value.

2.3 Filtering for Population of Interest

The dataset `NH_data` will be filtered to include only subjects who meet the following criteria:

- Age between 21 and 79 years (`RIDAGEYR`)
- Completed both the NHANES examination and interview (`RIDSTATR`)

```
NH_data <- NH_data %>% filter(RIDAGEYR >= 21 & RIDAGEYR != 80, RIDSTATR == "Completed Examin
nrow(NH_data)
```

```
[1] 7853
```

After applying these filters, the `NH_data` tibble now contains 7,853 observations.

2.4 Checking the Quantitative Variables

The `NH_data` dataset contains three quantitative variables:

- `RIDAGEYR` – Age of the subject (in years).
- `DR2TPROT` – Total protein consumption (in grams).
- `BMXBMI` – Body mass index (BMI) (in kg/m^2).

To ensure these variables have plausible values and to assess the number of missing values, we generate summary statistics:

```
df_stats(~ RIDAGEYR + DR2TPROT + BMXBMI,  data = NH_data) %>% gt() %>%
  tab_options(table.width = pct(75)) %>%
  cols_align(align = "center") %>% fmt_number(decimals = 2, columns = c(2:8))
```

response	min	Q1	median	Q3	max	mean	sd	n	missing
RIDAGEYR	21.00	35.00	50.00	63.00	79.00	49.32	16.01	7853	0
DR2TPROT	0.00	48.95	69.45	96.32	417.30	76.64	40.82	6178	1675
BMXBMI	14.60	25.00	28.90	34.00	92.30	30.24	7.68	7736	117

The `RIDAHEYR` stat correctly displays that our data contains only adults between the ages of 21 and 79 years of age with no missing values. `DR2TPROT` and `BMXBMI` both display a plausible range of values. There are 1675 missing values for `DR2TPROT` and 118 missing values for `BMXBMI`. These missing values will be considered when conducting further analysis.

2.5 Checking Binary Variables

The `NH_data` dataset contains three binary variables:

- `RIDSTATR` – Examination and interview completion status.
- `RIAGENDR` – Gender of the subject.

- `ALQ151` – Response to the question: "Have you ever, at any point in your life, had 4 or 5 drinks every day?"

We will verify that these variables contain correct values and assess the number of missing values.

2.5.1 The `RIDSTATR` Variable

The `RIDSTATR` has two levels:

- "Missing Examination"
- "Completed Examination and Interview".

Since we filtered the dataset to include only subjects who completed both, we will verify that this filtering was applied correctly.

```
NH_data %>% tabyl(RIDSTATR) %>% gt() %>%
  tab_options(table.width = pct(75)) %>%
  cols_align(align = "center")
```

RIDSTATR	n	percent
Missing Examination	0	0
Completed Examination and Interview	7853	1

There are no subjects with a *"Missing Examination"* and the number of subjects who completed the examination and interview corresponds to the number of rows in the `NH_data` tibble (7853 observations). This confirms that our filtering was applied correctly.

2.5.2 The `RIAGENDR` Variable

The `RIAGENDR` variable represents the gender of each subject and has two levels:

- Male
- Female

We will check for any missing values in this variable.

```
NH_data %>% tabyl(RIAGENDR) %>% gt() %>%
  tab_options(table.width = pct(75)) %>%
  cols_align(align = "center") %>% fmt_number(decimals = 2, columns = 3)
```

RIAGENDR	n	percent
Male	3794	0.48
Female	4059	0.52

The dataset contains 3,794 males and 4,059 females. There are no missing values for gender, confirming that every subject has their gender recorded.

2.5.3 The `ALQ151` Variable

The ALQ151 variable represents responses to the question: *"Have you ever, at any point in your life, had 4 or 5 drinks every day?"*

It has two levels:

- *"Yes"*
- *"No"*

```
NH_data %>% tabyl(ALQ151) %>% gt() %>%
  tab_options(table.width = pct(75)) %>%
  cols_align(align = "center") %>% fmt_number(decimals = 2, columns = c(3:4))
```

ALQ151	n	percent	valid_percent
Yes	1061	0.14	0.16
No	5633	0.72	0.84
NA	1159	0.15	NA

1,061 subjects responded *"Yes"*. 5,633 subjects responded *"No"*. A total of 6,694 subjects provided a response. There are 1,159 missing values for this variable.

2.6 Checking Multi-Categorical Variables

The `NH_data` dataset contains two multi-categorical variables:

- *DBQ750* – Frequency of reading nutrition labels before purchasing food.
- *DMDEDUC2* – Highest level of education completed.

We will check the number of observations for each level of these variables and assess the number of missing values.

2.6.1 The `DBQ750` Variable

The *DBQ750* variable represents how often a subject reads nutrition labels before purchasing food. It has five levels:

- *Always*
- *Most of the time*
- *Sometimes*
- *Rarely*
- *Never*

We will check the distribution of responses and assess the number of missing values.

```
NH_data %>% tabyl(DBQ750) %>% gt() %>%
  tab_options(table.width = pct(75)) %>%
  cols_align(align = "center") %>% fmt_number(decimals = 2, columns = c(3:4))
```

DBQ750	n	percent	valid_percent
Always	885	0.11	0.15
Most of the time	1451	0.18	0.25
Sometimes	2056	0.26	0.36
Rarely	766	0.10	0.13
Never	625	0.08	0.11
NA	2070	0.26	NA

5,783 subjects provided a response for `DBQ750`. 2,070 subjects have missing values for this variable.

2.6.2 The `DMDEDUC2` Variable

The DMDEDUC2 variable represents the highest level of education completed by each subject. It has five levels:

- *Less Than 9th Grade*
- *9-11th Grade*
- *High School/GED or Equivalent*
- *Some College or AA Degree*
- *College Graduate or Above*

We will check the distribution of responses and assess the number of missing values.

```
NH_data %>% tabyl(DMDEDUC2) %>% gt() %>%
  tab_options(table.width = pct(75)) %>%
  cols_align(align = "center") %>% fmt_number(decimals = 2, columns = c(3:4))
```

DMDEDUC2	n	percent	valid_percent
Less Than 9th Grade	602	0.08	0.08
9-11th Grade	843	0.11	0.11
High School/GED or Equivalent	1868	0.24	0.24
Some College or AA degree	2569	0.33	0.33
College Graduate or above	1963	0.25	0.25
NA	8	0.00	NA

7,845 subjects provided a response for DMDEDUC2.The level with the fewest responses is "Less Than 9th Grade" with 602 subjects. There are 8 missing values for this variable.

2.7 Creating the Analytic Tibble

The `NH_data` tibble will be updated to retain only the variables necessary for the four upcoming analyses, along with the unique subject identifier (`SEQN`).

```
NH_data <- NH_data %>% select(SEQN, RIAGENDR:ALQ151)
dim(NH_data)
```

```
[1] 7853    7
```

The updated `NH_data` tibble now contains:

- 7 variables (columns)
- 7,853 observations (rows)

This tidy dataset ensures that only the relevant variables are retained for further analysis.

2.8 Listing of Missing Values

Before proceeding with the analysis, we will check which variables contain missing values and determine the number of missing observations for each.

```
miss_var_summary(NH_data) %>% gt() %>%
  tab_options(table.width = pct(75)) %>%
  cols_align(align = "center") %>% fmt_number(decimals = 2, columns = 3)
```

variable	n_miss	pct_miss
DBQ750	2070	26.4
DR2TPROT	1675	21.3
ALQ151	1159	14.8
BMXBMI	117	1.49
DMDEDUC2	8	0.102
SEQN	0	0
RIAGENDR	0	0

Overall there are five variables with missing values:

- `DBQ750` - 2070 missing values
- `DR2TPROT` - 1675 missing values
- `ALQ151` - 1159 missing values
- `BMXBMI` - 117 missing values

- `DMDEDUC2` - 8 missing values

3 Codebook and Data Description

3.1 Codebook

The following codebook provides information about the seven variables in the tidy dataset `NH_data`. The "Type" column indicates whether a variable is character, quantitative, or categorical. For categorical variables, the number of levels is specified. The number of missing values is shown in brackets for variables that contain missing data. The "Description and Levels" column explains what each variable represents and, if applicable, lists its levels.

Variable	Type	Description and Levels
`SEQN`	Character	Respondent identification number
`RIAGENDR`	Categorical- 2	Male, female: The subjects gender
`DR2TPROT`	Quantitative	Total protein consumption over twenty-four hours in grams [**1675 NA**]
`DBQ750`	Categorical-5	Always, Most of the time, Sometimes, Rarely, Never: How often do you read the nutrition label before purchasing a food item? [**2070 NA**]
`BMXBMIL`	Quantitative	The subjects body mass index, in kg/m^2 [**117 NA**]
`DMDEDUC2`	Categorical-5	Less Than 9th Grade, 9-11th Grade, High School/GED or Equivalent, Some College or AA degree, College Graduate or above: The subjects highest level of education [**8 NA**]
`ALQ151`	Categorical- 2	Yes, no: Have you ever had a period in your life where you consumed 4 or 5 drinks almost every day? [**1159 NA**]

3.2 Analytic Tibble

The following check shows that the `NH_data` tibble is in a tibble data structure:

```
NH_data
```

```
# A tibble: 7,853 x 7
   SEQN   RIAGENDR DR2TPROT DBQ750           BMXBMI DMDEDUC2             ALQ151
```

```
   <chr>  <fct>       <dbl> <fct>             <dbl> <fct>                 <fct>
 1 109266 Female       62.9 Most of the time   37.8 College Graduate or ~ No
 2 109271 Male        140.  Sometimes          29.7 9-11th Grade          Yes
 3 109273 Male         81.0 Sometimes          21.9 Some College or AA d~ No
 4 109274 Male         47.9 Always             30.2 Some College or AA d~ No
 5 109282 Male         94.7 Sometimes          26.6 College Graduate or ~ No
 6 109284 Female      141.  Sometimes          39.1 9-11th Grade          <NA>
 7 109286 Female      105.  Rarely             28.9 College Graduate or ~ <NA>
 8 109290 Female      102.  Most of the time   28.1 College Graduate or ~ No
 9 109291 Female       21.8 Sometimes          31.3 College Graduate or ~ <NA>
10 109292 Male         NA   <NA>               30.5 High School/GED or E~ No
# i 7,843 more rows
```

Further confirmatory check

```
is_tibble(NH_data)
```

```
[1] TRUE
```

3.3 Data Summary

The following is a summary of the `NH_data` tibble, filtered to include only the specific population of interest and the relevant variables for analysis.

```
NH_data %>% describe()
```

.

```
 7  Variables      7853  Observations
--------------------------------------------------------------------------------
SEQN
       n  missing distinct
    7853        0     7853

lowest : 109266 109271 109273 109274 109282, highest: 124815 124817 124818 124821 124822
--------------------------------------------------------------------------------
RIAGENDR
       n  missing distinct
    7853        0        2

Value        Male Female
Frequency    3794   4059
Proportion  0.483  0.517
--------------------------------------------------------------------------------
DR2TPROT : Protein (gm)
       n  missing distinct     Info     Mean      Gmd      .05      .10
    6178     1675     5603        1    76.64    43.16    24.71    33.52
     .25      .50      .75      .90      .95
```

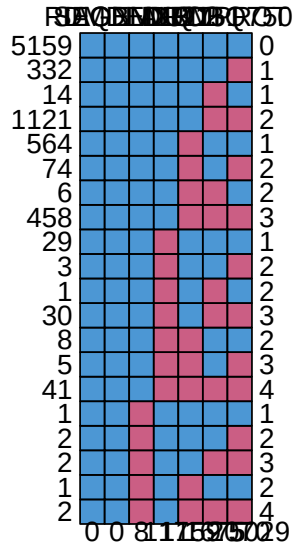
```
   48.95    69.45    96.32   126.93   150.76

lowest : 0      0.58   1.06   1.13   1.27  , highest: 322.3  329.56 345.53 352.25 417.3
---------------------------------------------------------------------------
DBQ750
       n  missing distinct
    5783     2070        5

Value                Always Most of the time        Sometimes           Rarely
Frequency               885             1451             2056              766
Proportion            0.153            0.251            0.356            0.132

Value                 Never
Frequency               625
Proportion            0.108
---------------------------------------------------------------------------
BMXBMI : Body Mass Index (kg/m**2)
       n  missing distinct     Info     Mean      Gmd      .05      .10
    7736      117      447        1    30.24    8.191    20.40    21.90
     .25      .50      .75      .90      .95
   25.00    28.90    34.00    40.10    44.42

lowest : 14.6 14.8 14.9 15   15.1, highest: 80.6 82   84.4 86.2 92.3
---------------------------------------------------------------------------
DMDEDUC2
       n  missing distinct
    7845        8        5

Less Than 9th Grade (602, 0.077), 9-11th Grade (843, 0.107), High School/GED or
Equivalent (1868, 0.238), Some College or AA degree (2569, 0.327), College
Graduate or above (1963, 0.250)
---------------------------------------------------------------------------
ALQ151
       n  missing distinct
    6694     1159        2

Value        Yes    No
Frequency   1061  5633
Proportion 0.159 0.841
---------------------------------------------------------------------------
```

4 Imputation

To address missing values in the dataset, we applied multiple imputation using the `mice` package. This approach was chosen to minimize bias and preserve sta-

tistical power, as complete-case analysis could lead to a substantial loss of data. Analyzing the patterns of missingness suggested that data were primarily Missing at Random (MAR), meaning the probability of missingness was related to observed variables rather than being completely random. Given this, imputation was performed separately for numerical and categorical variables: Predictive Mean Matching (PMM) was used for continuous variables (`DR2TPROT`, `BMXBMI`), logistic regression was applied to binary variables (`ALQ151`), and multinomial logistic regression was used for multi-category variables (`DBQ750`, `DMDEDUC2`). Five imputed data sets were generated, and a completed dataset was extracted for analysis. Post-imputation checks confirmed that the imputed values aligned with observed distributions, ensuring that the process did not introduce artificial distortions.

```
md.pattern(NH_data)
```



	SEQN	RIAGENDR	DMDEDUC2	BMXBMI	ALQ151	DR2TPROT	DBQ750	
5159	1	1	1	1	1	1	1	0
332	1	1	1	1	1	1	0	1
14	1	1	1	1	1	0	1	1
1121	1	1	1	1	1	0	0	2
564	1	1	1	1	0	1	1	1
74	1	1	1	1	0	1	0	2
6	1	1	1	1	0	0	1	2
458	1	1	1	1	0	0	0	3
29	1	1	1	0	1	1	1	1
3	1	1	1	0	1	1	0	2
1	1	1	1	0	1	0	1	2

```
30      1         1         1       0       1         0       0    3
8       1         1         1       0       0         1       1    2
5       1         1         1       0       0         1       0    3
41      1         1         1       0       0         0       0    4
1       1         1         0       1       1         1       1    1
2       1         1         0       1       1         1       0    2
2       1         1         0       1       1         0       0    3
1       1         1         0       1       0         1       1    2
2       1         1         0       1       0         0       0    4
        0         0         8     117    1159      1675    2070 5029
```

Defining the imputation method for each variable.

```
impute_methods <- make.method(NH_data)

impute_methods["SEQN"] <- ""
impute_methods["DR2TPROT"] <- "pmm"
impute_methods["BMXBMI"] <- "pmm"
impute_methods["ALQ151"] <- "logreg"
impute_methods["DBQ750"] <- "polyreg"
impute_methods["DMDEDUC2"] <- "polyreg"
```

Multiple imputation will be down using 5 imputed data sets

```
imp_model <- mice(NH_data, method = impute_methods, m = 5, seed = 444)
```

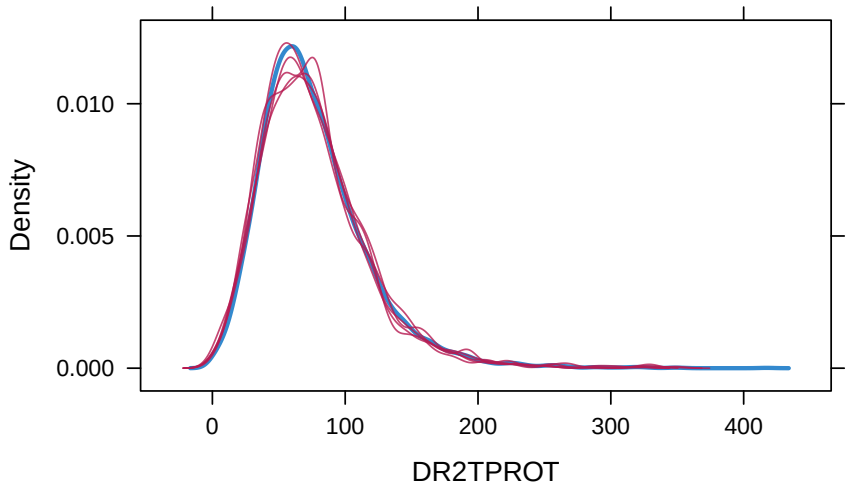
```
 iter imp variable
  1   1  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  1   2  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  1   3  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  1   4  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  1   5  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  2   1  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  2   2  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  2   3  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  2   4  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  2   5  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  3   1  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  3   2  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  3   3  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  3   4  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  3   5  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  4   1  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  4   2  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  4   3  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  4   4  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
```

```
  4   5  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  5   1  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  5   2  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  5   3  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  5   4  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
  5   5  DR2TPROT  DBQ750  BMXBMI  DMDEDUC2  ALQ151
```

```
summary(imp_model)
```

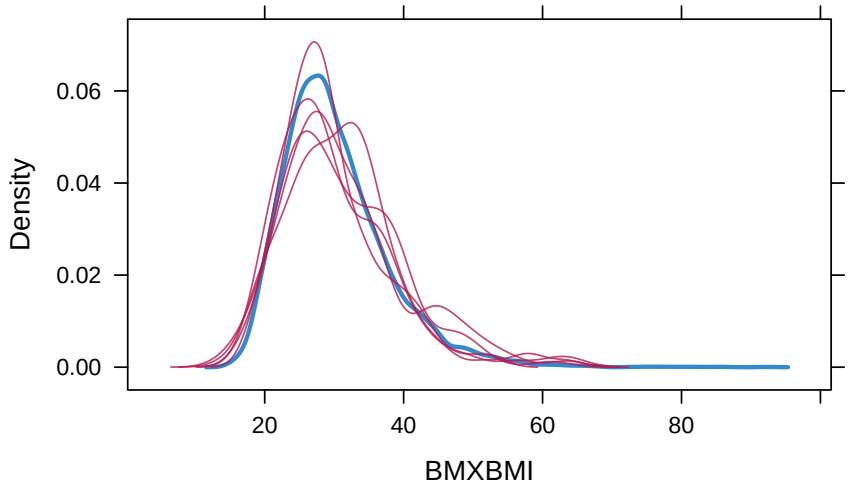
```
Class: mids
Number of multiple imputations:  5
Imputation methods:
     SEQN  RIAGENDR  DR2TPROT    DBQ750    BMXBMI  DMDEDUC2    ALQ151
       ""        ""     "pmm" "polyreg"     "pmm" "polyreg"  "logreg"
PredictorMatrix:
         SEQN RIAGENDR DR2TPROT DBQ750 BMXBMI DMDEDUC2 ALQ151
SEQN        0        1        1      1      1        1      1
RIAGENDR    0        0        1      1      1        1      1
DR2TPROT    0        1        0      1      1        1      1
DBQ750      0        1        1      0      1        1      1
BMXBMI      0        1        1      1      0        1      1
DMDEDUC2    0        1        1      1      1        0      1
Number of logged events:  1
  it im dep     meth  out
1  0  0     constant SEQN
```

```
densityplot(imp_model, ~DR2TPROT)
```



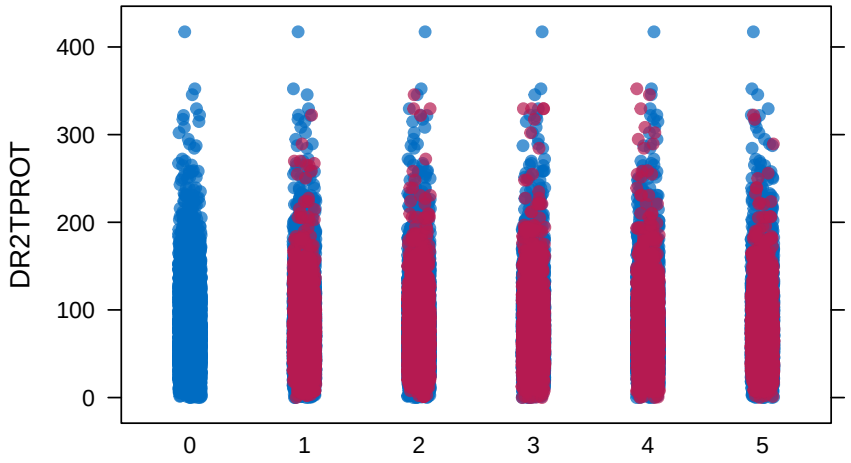
To assess the effectiveness of our multiple imputation process, we generated density plots comparing the distribution of observed and imputed values for `DR2TPROT` (total protein consumption). The plot shows that the imputed values (red lines) closely align with the observed data (blue line), indicating that the imputation method preserved the original data structure well. There are no significant deviations, suggesting that the imputation did not introduce bias.There are slight varriations that appear on the right tail of the distribution, but this occurs due to the selected of Predictive Mean Matching (PPM), which selected imputed values based on real observations. Overall, the imputed values maintain the integrity of the original dataset, supporting the choice of the imputation method.

```
densityplot(imp_model, ~BMXBMI)
```



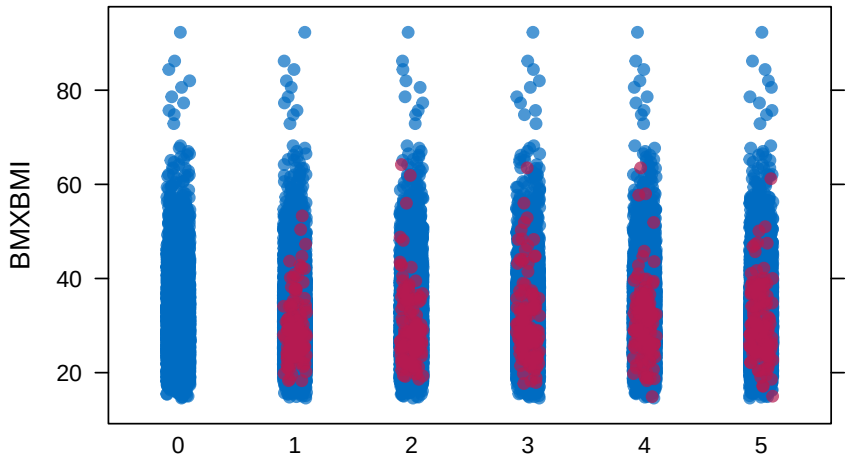
The density plot for `BMXBMI` (Body Mass Index) compares the distribution of observed values (blue line) with the imputed values (red lines) across multiple imputed data sets. The imputed values closely follow the observed distribution, indicating that the Predictive Mean Matching (PMM) method successfully captured the overall data structure. There is a slight increase in variability among the imputed data sets, particularly in the higher BMI range, but this is expected given the natural spread of BMI values and the small number of extreme cases. Importantly, no major deviations or unrealistic patterns are observed, confirming that the imputation process preserved the integrity of the original data while effectively handling missing values.

```
stripplot(imp_model, DR2TPROT, pch = 20, cex = 1.2)
```



The strip plot for `DR2TPROT` visualizes the distribution of observed values and imputed values (red) across the five data sets. The imputed values are well integrated with the observed data following a similar spread and range indicating the the PPM captured the natural variation of the predictor. There do not appear to be any apparent clusters or extreme deviation among the imputed points suggesting no bias was introdcued to the dataset. The presence of higher protein consumption outliers in both observed and imputed data further supports the validity of the imputation approach. Overall, this visualization confirms that the missing values were imputed in a manner consistent with the original data distribution.

```
stripplot(imp_model, BMXBMI, pch = 20, cex = 1.2)
```



The strip plot for `BMXBMI` shows a similar consistency as the `DR2TPROT` plot suggesting the chosen imputation method captured the structure of the data successfully. There are no apparent clusters or extreme deviations in the data suggesting no artificial bias.

The imputed dataset will be extracted.

```
NH_data_imp <- complete(imp_model)
```

Check for the any missing variables in the dataset

```
sum(is.na(NH_data_imp))
```

```
[1] 0
```

The imputated dataset will be saved

```
saveRDS(NH_data_imp, here("data", "processed", "NH_data_imputed.Rds"))
```

5 Analysis 1: Difference in Protein Consumption Among Men and Women

5.1 The Question

The mean `DR2TPROT` (total protein consumption) of the male population will be compared to that of the female population. This analysis will use independent samples, as there is no direct relationship between the two groups in `RIAGENDR` that would influence each others `DR2TPROT` values. Additionally, the number of

males and females in RIAGENDR is unequal, making the sample sizes unbalanced.

Our research question:

Is there a meaningful difference between total protein consumption between males and females?

5.2 Describing The Data

```
favstats(DR2TPROT ~ RIAGENDR, data = NH_data_imp) %>% gt() %>% fmt_number(decimals = 2, colu
  tab_options(table.width = pct(75)) %>%
  cols_align(align = "center")
```

RIAGENDR	min	Q1	median	Q3	max	mean	sd	n	missing
Male	0.00	56.22	80.27	108.41	352.25	86.58	43.91	3794	0
Female	0.00	42.65	60.98	82.56	417.30	66.22	34.56	4059	0

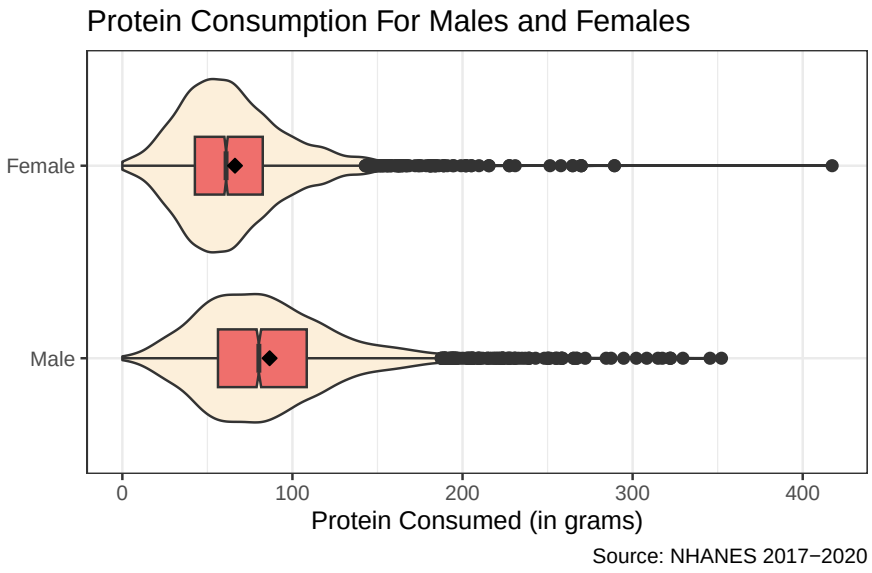
The numerical summary of total protein consumption (DR2TPROT) reveals notable differences between males and females. On average, males consume more protein (86.58g) than females (66.22g), with a higher median intake (80.27g vs. 60.98g) and a wider overall distribution. The interquartile range (IQR) for males (56.22g to 108.41g) is also higher compared to females (42.65g to 82.56g), suggesting greater variability in protein intake among men. The standard deviation further supports this, with males showing higher variability (43.91g) than females (34.56g). Both groups have a minimum reported intake of 0g. These results suggest a meaningful difference in protein consumption between genders, which will be further analyzed to determine statistical significance.

5.2.1 Graphical and Numerical Summaries

The boxplot and violin plot illustrate the distribution of total protein consumption (DR2TPROT) for males and females using complete case analysis. The violin plot shows the overall distribution shape, while the boxplot provides key summary statistics, including the median, interquartile range (IQR), and outliers. The notched boxplot visually represents confidence intervals around the median, helping to assess potential differences between groups.

```
NH_data_imp %>% ggplot(aes(x = RIAGENDR, y = DR2TPROT)) +
  geom_violin(fill = "#FCEFDA") +
  geom_boxplot(width = 0.3, notch = TRUE, fill = "#EF6F6C", outlier.size = 2) +
  coord_flip() +
  stat_summary(fun = "mean", geom = "point", shape = 23, size = 2, fill = "black") +
  labs(title = "Protein Consumption For Males and Females",
       x = "",
```

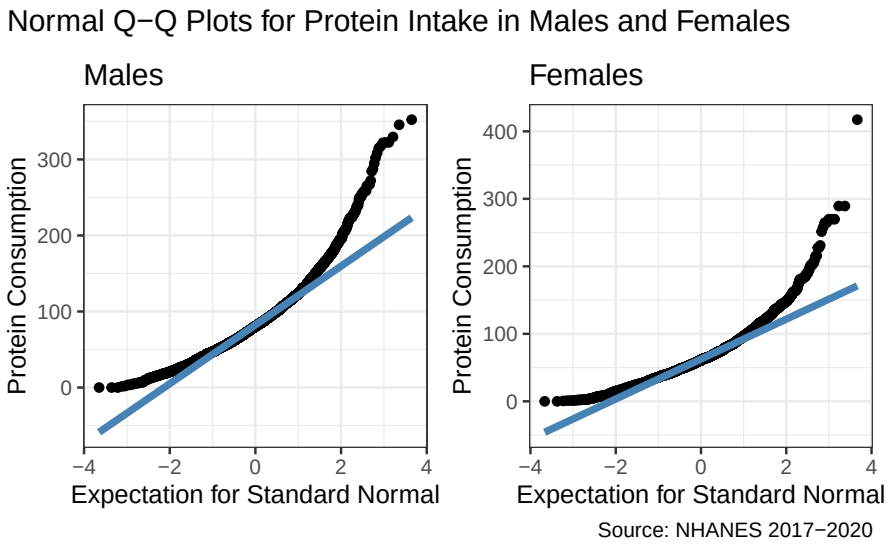
```
       y = "Protein Consumed (in grams)",
       caption = "Source: NHANES 2017-2020") -> Protein_Gender_Boxplot


ggsave(filename = here("figures", "Protein_Gender_Boxplot.png"), plot = Protein_Gender_Boxpl

Protein_Gender_Boxplot
```



The plot confirms that males generally consume more protein than females, as indicated by a higher median and mean protein intake. The spread of protein consumption is also greater for males, suggesting more variability in dietary habits. Both groups have outliers in the higher protein consumption range, which may reflect individuals with higher protein intake due to diet or lifestyle factors. This visualization reinforces the numerical summary findings and will be further examined through statistical testing.

```
# QQ plot for males
prot_m <- NH_data_imp %>%
  filter(RIAGENDR == "Male") %>%
  ggplot(aes(sample = DR2TPROT)) +
  geom_qq() +
  geom_qq_line(col = "steelblue", lwd = 1.5) +
  theme(aspect.ratio = 1) +
  labs(
    title = "Males",
    x = "Expectation for Standard Normal",
    y = "Protein Consumption")
```

```r
# QQ plot for females
prot_f <- NH_data_imp %>%
  filter(RIAGENDR == "Female") %>%
  ggplot(aes(sample = DR2TPROT)) +
  geom_qq() +
  geom_qq_line(col = "steelblue", lwd = 1.5) +
  theme(aspect.ratio = 1) +
  labs(
    title = "Females",
    x = "Expectation for Standard Normal",
    y = "Protein Consumption",
    caption = "Source: NHANES 2017-2020")

Protein_QQ_Plots <- prot_m + prot_f +  plot_annotation(title = "Normal Q-Q Plots for Protein

# Save the combined plot as a high-resolution PNG
ggsave(filename = here("figures", "Protein_QQ_Plots.png"), plot = Protein_QQ_Plots, width =

Protein_QQ_Plots
```



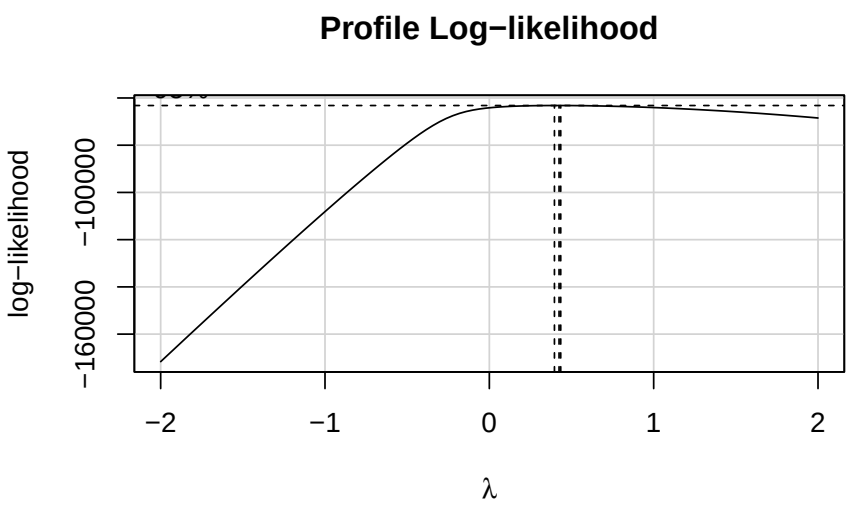
The Normal Q-Q plots for protein consumption (`DR2TPROT`) in males and females assess the assumption of normality by comparing the observed distribution to a theoretical normal distribution. In both groups, data points deviate significantly from the diagonal reference line, particularly at the tails, suggesting that protein intake is not normally distributed. The rightward curvature in the upper tails indicates the presence of positive skew, meaning some individuals

consume substantially more protein than average. Additionally, the lower tails deviate downward, reinforcing the presence of skewness and potential outliers at both extremes. Given this non-normal distribution, non-parametric statistical methods or transformations may be necessary for further analysis.

5.2.2 BoxCox for Transformation

A Box-Cox plot was used for `DR2TPROT` to address its non-normal distribution, as suggested by the Q-Q plots. Since Box-Cox transformations require strictly positive values, a small constant (0.01 grams) was added to all observations to accommodate subjects reporting 0 grams of protein intake.

```
boxCox(lm(DR2TPROT+0.01 ~ RIAGENDR, data = NH_data_imp))
```




```
powerTransform(lm(DR2TPROT+ 0.01 ~ RIAGENDR, data = NH_data_imp)) # 0.41
```

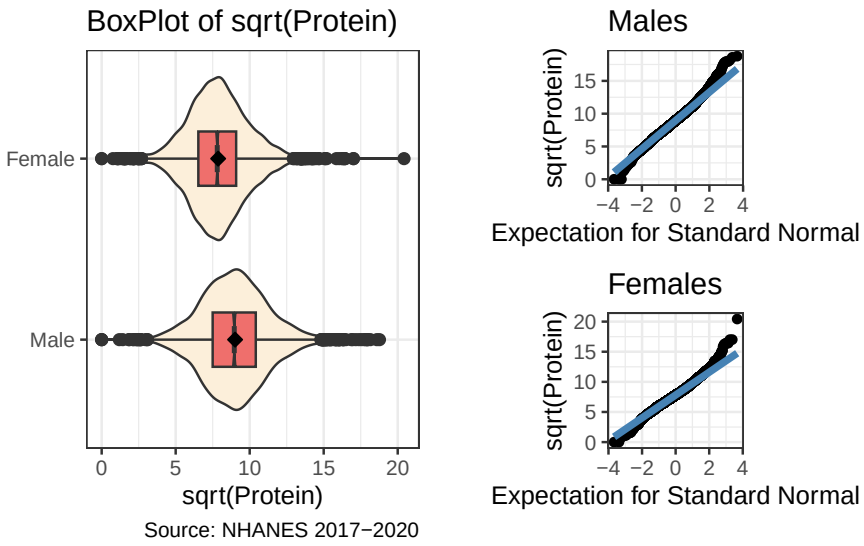
```
Estimated transformation parameter
       Y1
0.4143391
```

The Box-Cox plot suggests a transformation parameter close to 0.5, and the estimated transformation value from `powerTransform()` is 0.41, further confirming this recommendation. A power of 0.5 corresponds to a square root transformation, which will be applied to assess whether it improves normality. This transformation aims to reduce skewness and stabilize variance, making the data more suitable for parametric analyses.

5.2.3 Transformation of the Data

```
pp <- NH_data_imp %>%
  ggplot(aes(x = RIAGENDR, y = sqrt(DR2TPROT))) +
  geom_violin(fill = "#FCEFDA") +
  geom_boxplot(width = 0.3, notch = TRUE, fill = "#EF6F6C", outlier.size = 2) +
  coord_flip() +
  stat_summary(fun = "mean", geom = "point", shape = 23, size = 2, fill = "black") +
  labs(
    title = "BoxPlot of sqrt(Protein)",
    x = "",
    y = "sqrt(Protein)",
    caption = "Source: NHANES 2017-2020"
  )


prot_m2 <- NH_data_imp %>%
  filter(RIAGENDR == "Male") %>%
  ggplot(aes(sample = sqrt(DR2TPROT))) +
  geom_qq() +
  geom_qq_line(col = "steelblue", lwd = 1.5) +
  theme(aspect.ratio = 1) +
  labs(
    title = "Males",
    x = "Expectation for Standard Normal",
    y = "sqrt(Protein)"
  )


prot_f2 <- NH_data_imp %>%
  filter(RIAGENDR == "Female") %>%
  ggplot(aes(sample = sqrt(DR2TPROT))) +
  geom_qq() +
  geom_qq_line(col = "steelblue", lwd = 1.5) +
  theme(aspect.ratio = 1) +
  labs(
    title = "Females",
    x = "Expectation for Standard Normal",
    y = "sqrt(Protein)"
  )


sqrt_Protein_Plots <- pp + (prot_m2 / prot_f2)

ggsave(filename = here("figures", "sqrt_Protein_Plots.png"), plot = sqrt_Protein_Plots, widt
```

```
sqrt_Protein_Plots
```



The BoxPlot and Q-Q plots visualize the effects of applying a square root transformation to total protein consumption (DR2TPROT). The transformation was performed based on the Box-Cox analysis, which suggested a transformation parameter close to 0.5. The violin-boxplot shows that while the transformed data remains slightly skewed, the distribution appears more symmetric compared to the original scale. The Q-Q plots for males and females indicate improved normality, with data points aligning more closely with the diagonal reference line. However, slight deviations persist in the lower and upper tails, suggesting that while the transformation reduces skewness, perfect normality is not achieved. This transformation will enhance the validity of parametric statistical tests by stabilizing variance and making the distribution more suitable for further analysis.

5.2.4 Variance and Normality Checks

To determine whether the assumption of equal variances holds between males and females, we perform Levene's test on the square root-transformed DR2TPROT values.

```
leveneTest(sqrt(DR2TPROT) ~ RIAGENDR, data = NH_data_imp)
```

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    1  38.764 5.029e-10 ***
      7851
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To assess whether protein consumption for each group follows a normal distribution after transformation, we apply the Shapiro-Wilk test to the square root-transformed values.

```
# Ensure the variable is numeric
shapiro.test(NH_data_imp %>% filter(RIAGENDR == "Male") %>% pull(DR2TPROT) %>% sqrt())
```

```
    Shapiro-Wilk normality test

data:  NH_data_imp %>% filter(RIAGENDR == "Male") %>% pull(DR2TPROT) %>% sqrt()
W = 0.99197, p-value = 8.66e-14
```

```
shapiro.test(NH_data_imp %>% filter(RIAGENDR == "Female") %>% pull(DR2TPROT) %>% sqrt())
```

```
    Shapiro-Wilk normality test

data:  NH_data_imp %>% filter(RIAGENDR == "Female") %>% pull(DR2TPROT) %>% sqrt()
W = 0.98968, p-value < 2.2e-16
```

The results indicate that Levene's test is highly significant ($p < 0.001$), meaning the assumption of equal variances is violated, requiring the use of Welch's t-test instead of a pooled-variance t-test. Additionally, the Shapiro-Wilk test results for both males ($p = 8.66e\text{-}14$) and females ($p < 2.2e\text{-}16$) suggest strong deviations from normality, even after transformation. Given these findings, a non-parametric alternative (such as bootstrapping) remains the most robust approach for comparing protein consumption between genders.

5.3 Main Analysis

```
set.seed(444)

male_pro <- NH_data_imp %>% filter(RIAGENDR == "Male") %>% pull(DR2TPROT)
female_pro <- NH_data_imp %>% filter(RIAGENDR == "Female") %>% pull(DR2TPROT)

boot.t.test(x = male_pro, y = female_pro, alternative = 'two.sided', conf.level = 0.95, pari
```

```
    Bootstrap Welch Two Sample t-test

data:  male_pro and female_pro
number of bootstrap samples:  10000
bootstrap p-value < 1e-04
bootstrap difference of means (SE) = 20.33699 (0.8954913)
```

```
95 percent bootstrap percentile confidence interval:
 18.58277 22.08416

Results without bootstrap:
t = 22.72, df = 7200.9, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 18.59644 22.10851
sample estimates:
mean of x mean of y
 86.57570  66.22322
```

The bootstrapped Welch two-sample t-test was conducted to compare the mean protein consumption (`DR2TPROT`) between males and females. Using 10000 bootstrap resamples, the bootstrap p-value was < 0.0001, indicating a statistically significant difference in protein intake between genders. The bootstrap estimate of the mean difference was 20.34 grams (SE = 0.90), with a 95% confidence interval (18.58g, 22.08g), suggesting that males consistently consume more protein than females.

5.4 Conclusions

The results of the bootstrapped two-sample t-test provide strong evidence that males consume significantly more protein than females. With 10,000 bootstrap resamples, the estimated mean difference in protein intake was 20.34 grams (SE = 0.90), and the 95% confidence interval (18.58g, 22.08g) confirms that this difference is unlikely due to random variation. The bootstrap p-value (< 0.0001) further supports the statistical significance of this finding. Since bootstrapping does not rely on normality or equal variance assumptions, these results are robust and reliable, reinforcing the conclusion that males have a consistently higher protein intake compared to females.

One limitation of our analysis is that `DR2TPROT` only reflects total protein consumption for a single day, which may not accurately represent an individual's long-term dietary habits. Daily protein intake can vary due to lifestyle factors or temporary dietary choices, meaning our estimates may not fully capture habitual protein consumption patterns between gender. Additionally, we did not account for dietary reporting bias, where individuals may under report or over report their intake. Future research should consider analyzing multiple days of dietary intake or incorporating habitual dietary assessment methods to obtain a more accurate picture of protein consumption trends.

6 Analysis 2: Comparing 5 Population Means with Independent Samples

6.1 The Question

This analysis will use the variables `DBQ750` and `BMXBMI` to compare the mean BMI (`BMXBMI`) across different groups based on how frequently individuals read nutrition labels before purchasing food items. The five groups represent individuals who always, most of the time, sometimes, rarely, or never read nutrition labels. These groups are independent, meaning that an individual's frequency of reading nutrition labels has no effect on the BMI of other subjects. To compare the groups, an analysis of variance and 95% confidence intervals will be used.

Research Question: Is there an association between an individual's BMI and the frequency with which they read nutrition labels on food products before purchasing them?

6.2 Describing The Data

6.2.1 Numerical Summary

To begin, we examine the numerical summary of BMI (`BMXBMI`) across the five groups categorized by nutrition label reading frequency (`DBQ750`).

```
favstats(BMXBMI ~ DBQ750, data = NH_data_imp) %>% gt() %>% fmt_number(decimals = 2, columns
  tab_options(table.width = pct(75)) %>%
  cols_align(align = "center")
```

DBQ750	min	Q1	median	Q3	max	mean	sd	n	missi
Always	15.40	25.00	28.55	33.40	74.80	29.71	7.11	1218	0
Most of the time	14.60	25.20	28.90	34.00	80.60	30.39	7.70	1910	0
Sometimes	14.90	25.20	29.40	34.90	92.30	30.76	8.09	2771	0
Rarely	15.40	24.80	28.70	34.10	77.30	30.17	7.65	1037	0
Never	15.10	24.40	27.90	32.60	66.20	29.00	6.80	917	0

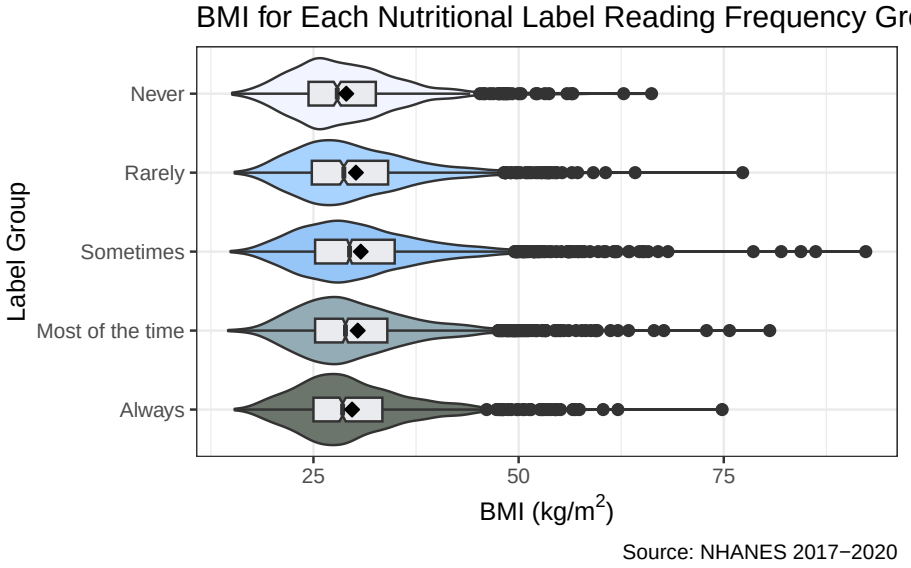
The mean BMI across groups is relatively similar, with the "Sometimes" group having the highest mean BMI (30.76) and the "Never" group having the lowest (29.00). This pattern is also reflected in the median values, where individuals who sometimes read nutrition labels have the highest median BMI (29.40), while those who never read them have the lowest (27.90). In all groups, the mean BMI is higher than the median, indicating that the distributions are right-skewed, with a subset of individuals having higher BMI values. The standard deviations range from 6.80 to 8.09, suggesting some variability in BMI within each group. Furthermore, the number of subjects in each category varies, with the "Sometimes" group having the largest sample size (n = 2771) and the "Never" group

having the smallest (n = 917). The unequal sample sizes further confirm that subjects are independent across groups, ensuring validity in comparing their BMI distributions.

6.2.2 Graphical Summaries

A Box plot and violin plot will be used to examine the distribution of BMI (`BMXBMI`) across the five groups categorized by nutrition label reading frequency (`DBQ750`) to compare their spread and central tendencies.

```
NH_data_imp %>% ggplot(aes(x = DBQ750, y = BMXBMI, fill = DBQ750)) +
  geom_violin() +
  geom_boxplot(width = 0.3, notch = TRUE, outlier.size = 2, fill = "#EAEBEE") +
  coord_flip() +
  stat_summary(fun = "mean", geom = "point", shape = 23, size = 2, fill = "black") +
  theme(legend.position = "none") +
  scale_fill_manual(values = c("#6C756B", "#93ACB5", "#96C5F7", "#A9D3FF", "#F2F4FF")) +
  labs(title = "BMI for Each Nutritional Label Reading Frequency Group",
       x = "Label Group",
       y = expression(paste("BMI (kg/m"^2, ")")),
       caption = "Source: NHANES 2017-2020") -> BMI_Label_Plot

ggsave(filename = here("figures", "BMI_Label_Plot.png"), plot = BMI_Label_Plot,width = 12, h

BMI_Label_Plot
```



The violin and boxplots reveal that each group has a considerable number of outliers at the higher end of BMI, which helps explain the higher means observed

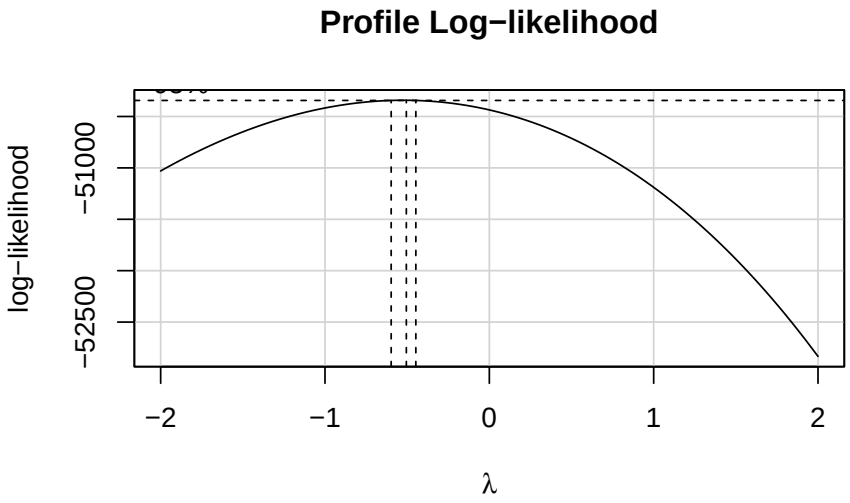
in the numerical summary. The "Sometimes" group has the widest distribution of BMI values, indicating greater variability compared to the other groups. In contrast, the "Never" group has the smallest range, suggesting more consistency in BMI values within this group. Despite these differences in distribution spread, the median BMI values are relatively close across all groups, and the notches in the boxplot indicate there is not a difference in the the median BMI for each group.

```
NH_data_imp %>% ggplot(aes(sample = BMXBMI)) +
  geom_qq() +
  geom_qq_line(col = "steelblue", lwd = 1.5) +
  theme(aspect.ratio = 1) +
  labs(title = "Q-Q Plot of Body Mass Index by Nutrition Label Reading Frequency",
       x = "Expectation for Standard Normal ",
       y = "BMI") + facet_wrap(~DBQ750) -> BMI_QQ

ggsave(filename = here("figures", "BMI_Label_Plot.png"), plot = BMI_QQ ,width = 12, height =

BMI_QQ
```

The Q-Q plots reveal that BMI (`BMXBMI`) is not normally distributed across any of the five nutrition label reading groups. In all cases, the data deviates from the reference line, particularly in the upper tail, indicating strong right-skewness. This suggests that a subset of individuals in each group have significantly higher BMI values than expected under normality. To account for this deviation from normality, non-parametric methods may be considered (e.g., Kruskal-Wallis or bootstrapping). Additionally, a transformation will be considered.

6.2.3 Exploring a Transformation

```
boxCox(lm(BMXBMI ~ DBQ750, data = NH_data_imp))
```




```
powerTransform(lm(BMXBMI ~ DBQ750, data = NH_data_imp))
```

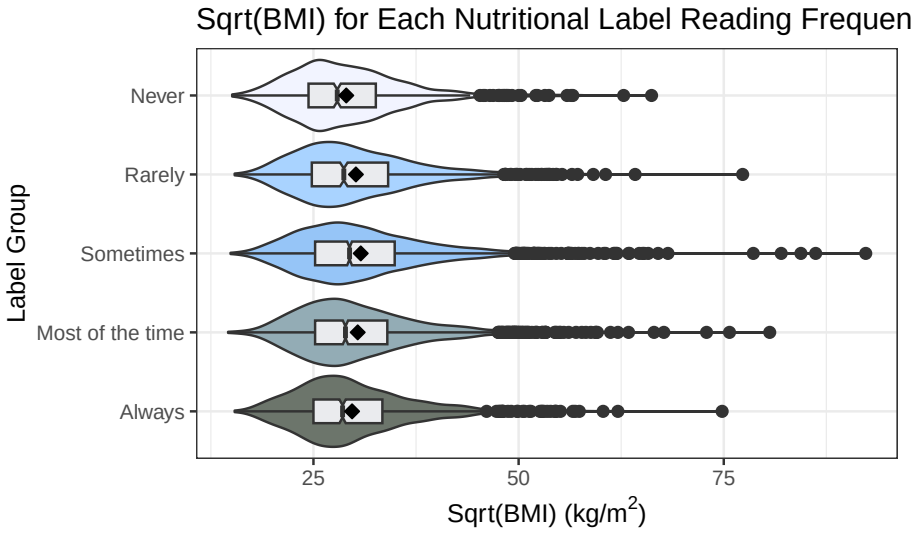
```
Estimated transformation parameter
        Y1
-0.5218437
```

The BoxCox plot and maximum likelihood ratio points to a transformation of -0.5 indicating an inverse square root transformation for our outcome variable BMI (`BMXBMI`).

6.2.4 Transforming the Data

```
NH_data_imp %>% ggplot(aes(x = DBQ750, y = BMXBMI, fill = DBQ750)) +
  geom_violin() +
  geom_boxplot(width = 0.3, notch = TRUE, outlier.size = 2, fill = "#EAEBEE") +
  coord_flip() +
  stat_summary(fun = "mean", geom = "point", shape = 23, size = 2, fill = "black") +
  theme(legend.position = "none") +
  scale_fill_manual(values = c("#6C756B", "#93ACB5", "#96C5F7", "#A9D3FF", "#F2F4FF")) +
  labs(title = "Sqrt(BMI) for Each Nutritional Label Reading Frequency Group",
       x = "Label Group",
       y = expression(paste("Sqrt(BMI) (kg/m"^2, ")")),
       caption = "Source: NHANES 2017-2020") -> Sqrt_BMI_Label_Plot
```

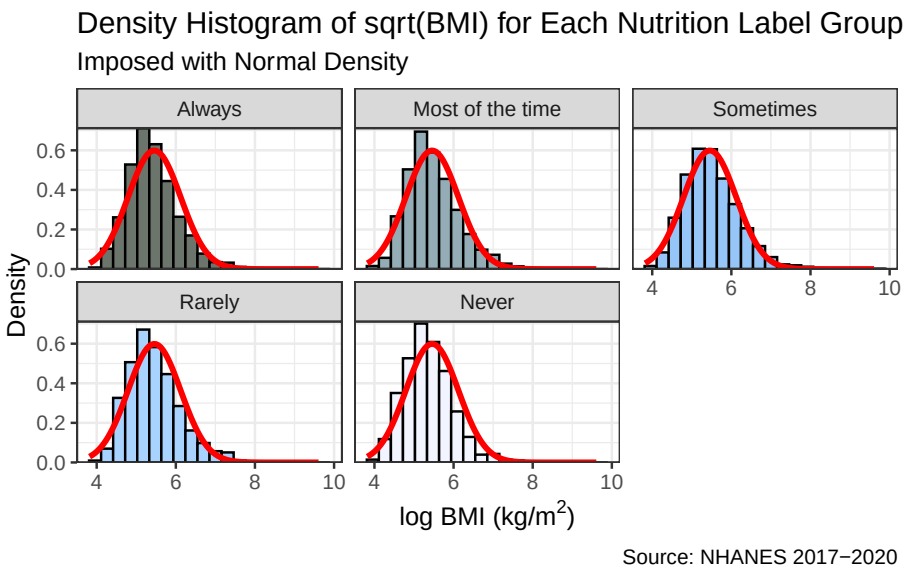
```
ggsave(filename = here("figures", "Sqrt_BMI_Label_Plot.png"), plot = Sqrt_BMI_Label_Plot ,wi

Sqrt_BMI_Label_Plot
```



The square root transformation was applied to BMI (`BMXBMI`) to address the previously observed right skewness in its distribution. Compared to the untransformed version, the spread of values is more compressed, and the right tail has been reduced, indicating a moderate improvement in normality. However, some degree of asymmetry and high-value outliers still persist, particularly in the "Sometimes" and "Rarely" groups. The general ranking of median BMI across groups remains unchanged, with the "Sometimes" group exhibiting the widest range, while the "Never" group has the narrowest distribution.

```
NH_data_imp %>% ggplot(aes(x = sqrt(BMXBMI), fill = DBQ750)) +
  geom_histogram(aes(y = after_stat(density)), col = "black", bins = 20) +
  scale_fill_manual(values = c("#6C756B", "#93ACB5", "#96C5F7", "#A9D3FF", "#F2F4FF")) +
  stat_function(fun = dnorm, args = list(mean = mean(sqrt(NH_data_imp$BMXBMI), na.rm = TRUE)
                                         sd = sd(sqrt(NH_data_imp$BMXBMI), na.rm = TRUE)), c
  scale_y_continuous(expand = c(0,0)) +
  guides(fill = "none") +
  labs(title = "Density Histogram of sqrt(BMI) for Each Nutrition Label Group",
       subtitle = 'Imposed with Normal Density',
       x = expression(paste("log BMI (kg/m"^2, ")")),
       y = "Density",
       caption = "Source: NHANES 2017-2020") + facet_wrap(~DBQ750) -> Sqrt_BMI_Density_Plot
```

```r
ggsave(filename = here("figures", "Sqrt_BMI_Density_Plot.png"), plot = Sqrt_BMI_Density_Plot

Sqrt_BMI_Density_Plot
```



The density histograms display the distribution of square root transformed BMI (`BMXBMI`) for each nutrition label reading frequency group (DBQ750), overlaid with a normal density curve (red). The transformation has improved symmetry, bringing the distributions closer to normality, but some deviations remain. Across all groups, the histograms generally align with the normal curve in the center, suggesting that the majority of individuals follow an approximately normal distribution. However, the tails show some departures from normality, particularly in the upper range, where mild right-skewness is still present. The "Sometimes" and "Rarely" groups exhibit more noticeable right tails, indicating a subset of individuals with higher BMI values. Additionally, the "Never" and "Always" groups appear to have slightly narrower distributions, suggesting lower variance in BMI within these groups. This could be a potential indication of dissimlar variance within each of the groups.

6.2.5 Variance Check

```r
variance_summary <- NH_data_imp %>% group_by(DBQ750) %>% summarise(Variance = var(sqrt(BMXBM


variance_summary %>% gt() %>% fmt_number(decimals = 2) %>%
  tab_options(table.width = pct(75)) %>%
  cols_align(align = "center")
```

DBQ750	Variance
Always	0.39
Most of the time	0.44
Sometimes	0.48
Rarely	0.44
Never	0.37

The table presents the calculated variance of sqrt(BMI) (`BMXBMI`) for each nutrition label reading frequency group (`DBQ750`). the values are now lower and more similar across groups, ranging from 0.37 ("Never" group) to 0.48 ("Sometimes" group"). The "Sometimes" group still has the highest variability, while the "Never" group has the lowest, mirroring the pattern seen in the untransformed data but with more compressed variance differences.

```
leveneTest(sqrt(BMXBMI) ~ DBQ750, data = NH_data_imp)
```

```
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    4  5.2399 0.0003263 ***
      7848
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The result of the Levene's test is statistically significant ($p < 0.001$), meaning there is strong evidence that the variances are not equal across groups. This violation of the homogeneity of variance assumption is important because parametric tests such as ANOVA assume equal variances. Given this result, adjustments or alternative methods may be necessary, such as Welch's ANOVA, which accounts for unequal variances, or non-parametric approaches like the Kruskal-Wallis test.

6.3 Main Analysis

6.3.1 Analysis of Variance

Given that Levene's test remained significant even after applying the square root transformation, we concluded that variance differences across groups persist. Additionally, previous normality checks (e.g., Q-Q plots and density histograms) confirmed that BMI (`BMXBMI`) is not normally distributed across nutrition label reading groups. Since ANOVA assumes both normality and equal variances, and Welch's ANOVA, while robust to variance differences, still assumes normality, we opted for the Kruskal-Wallis test. This non-parametric test does not require normality or homogeneity of variance, making it the best choice for analyzing

differences in median BMI across groups. To keep interpretation straightforward, we will apply the Kruskal-Wallis test to the untransformed BMI variable.

```
krus_bmi <- kruskal.test(BMXBMI ~ DBQ750, data = NH_data_imp)
krus_bmi
```

```
    Kruskal-Wallis rank sum test

data:  BMXBMI by DBQ750
Kruskal-Wallis chi-squared = 37.175, df = 4, p-value = 1.658e-07
```

The Kruskal-Wallis test was conducted to compare median BMI (`BMXBMI`) across the five nutrition label reading frequency groups (`DBQ750`). The test yielded a statistically significant result ($ ^2 = 37.175$, df $= 4$, $p < 0.0001$), indicating that at least one group differs in median BMI. However, the test does not specify which groups are significantly different from each other. To determine which specific group differences are driving this result, we will conduct a Dunn's post hoc test with a Bonferroni correction for multiple comparisons.

```
dunn_BMI <- dunnTest(BMXBMI ~ DBQ750, data = NH_data_imp, method = "bonferroni")
dunn_BMI
```

```
Dunn (1964) Kruskal-Wallis multiple comparison

  p-values adjusted with the Bonferroni method.
```

	Comparison	Z	P.unadj	P.adj
1	Always - Most of the time	-2.0501471	4.035008e-02	4.035008e-01
2	Always - Never	2.1706125	2.996047e-02	2.996047e-01
3	Most of the time - Never	4.2333553	2.302303e-05	2.302303e-04
4	Always - Rarely	-0.9820304	3.260849e-01	1.000000e+00
5	Most of the time - Rarely	0.8731954	3.825566e-01	1.000000e+00
6	Never - Rarely	-3.0089272	2.621719e-03	2.621719e-02
7	Always - Sometimes	-3.5075110	4.523196e-04	4.523196e-03
8	Most of the time - Sometimes	-1.5268574	1.267965e-01	1.000000e+00
9	Never - Sometimes	-5.6562053	1.547563e-08	1.547563e-07
10	Rarely - Sometimes	-2.1726011	2.981035e-02	2.981035e-01

Following the significant Kruskal-Wallis test, a Dunn's post hoc test with Bonferroni correction was conducted to determine which specific groups differ in median BMI (`BMXBMI`) based on nutrition label reading frequency. The results indicate that individuals who never read nutrition labels have a significantly different median BMI compared to those who sometimes read them ($p < 0.0001$). Additionally, those who always read nutrition labels also have significantly different median BMI compared to the sometimes group ($p = 0.0045$). Other comparisons, such as "Most of the time" vs. "Never" ($p = 0.023$) and "Rarely" vs. "Sometimes" ($p = 0.298$), were not statistically significant after adjusting for multiple comparisons. These findings suggest that BMI differences are more

pronounced when comparing individuals at the extremes of nutrition label reading habits, particularly between those who never or always check labels and those who fall into the sometimes category. However, the differences between other intermediate groups are less distinct.

6.4 Conclusions

Our analysis found detectable differences in BMI based on the frequency of reading nutrition labels.

Instead of ANOVA, we used the Kruskal-Wallis test, which confirmed a statistically significant difference in median BMI across the nutrition label reading groups ($\$^{\wedge 2} = 37.175$, $p < 0.0001$). To identify which groups were driving this difference, we conducted a Dunn's post hoc test with Bonferroni correction. The results revealed that significant differences were observed between the Never vs. Sometimes ($p < 0.0001$) and Always vs. Sometimes ($p = 0.0045$) groups. This suggests that individuals who sometimes read nutrition labels have a significantly different median BMI compared to those who never or always read them.

Interestingly, individuals who never read nutrition labels had a lower median BMI compared to those in the sometimes category, while those who always read labels also had a significantly different median BMI compared to the sometimes group. However, other comparisons, such as Never vs. Most of the Time ($p = 0.023$) and Rarely vs. Sometimes ($p = 0.298$), were not statistically significant. These findings indicate that BMI differences are more pronounced at the extremes of label-reading frequency rather than among intermediate groups.

Overall, the results suggest that nutrition label reading frequency is associated with differences in BMI, but the relationship may not be strictly linear. Further research is needed to explore the underlying factors driving these differences, such as dietary habits, health awareness, or socioeconomic influences.

7 Analysis D: Drinking Habitis between Genders

7.1 The Question

We are interested in the association between a subject's gender and whether they have ever experienced a period in their life where they consumed multiple alcoholic drinks daily. The variable `RIAGENDR` represents gender, with two levels: male and female. To assess heavier drinking habits, we will use the variable `ALQ151`, a categorical variable with two levels: "Yes" and "No", based on whether the respondent reported consuming four or five alcoholic drinks almost every day at some point in their life. To analyze this association, we will construct a 2x2 contingency table, apply a Bayesian augmentation, and calculate

90% confidence intervals for the proportional probabilities.

Our research question is there a difference in the proportion of males and females who report having consumed four or five alcoholic drinks almost every day at some point in their life?

7.1.1 The 2x2 Table

The 2x2 table with `RIAGENDR` and `ALQ151`:

```
NH_data_imp %>% tabyl(RIAGENDR, ALQ151) %>%
                adorn_totals(where = "row") %>%
                adorn_percentages(denominator = "row") %>%
                adorn_pct_formatting() %>%
                adorn_ns(position = "front") %>%
                adorn_title(placement = "combined") %>%
                gt() %>%
                tab_options(table.width = pct(75)) %>%
                cols_align(align = "center")
```

RIAGENDR/ALQ151	Yes	No
Male	862 (22.7%)	2,932 (77.3%)
Female	352 (8.7%)	3,707 (91.3%)
Total	1,214 (15.5%)	6,639 (84.5%)

The contingency table presents the distribution of self-reported heavy drinking behavior (`ALQ151`) by gender (`RIAGENDR`). Overall, 15.5% of respondents reported having consumed four or five alcoholic drinks almost every day at some point in their life, while 84.5% did not.

Among males, 22.7% (862 out of 3,794) reported engaging in heavy drinking at some point, compared to only 8.7% (352 out of 4,059) of females. This suggests that males are more than twice as likely as females to report past heavy drinking behavior. Conversely, the proportion of respondents who answered "No" is higher among females (91.3%) than males (77.3%), indicating that a greater percentage of women report never having engaged in heavy drinking.

These proportions suggest a potential association between gender and heavy drinking behavior, with males being more likely to report a history of frequent alcohol consumption. Further statistical analysis is needed to determine whether this observed difference is statistically significant.

7.1.2 Checking Assumptions

Before interpreting the results of our chi-square test, we must ensure that key assumptions are met:

- Independence of Observations ** Each subject contributes to only one cell in the contingency table, meaning observations are independent and do not overlap across categories.
- Expected Cell Counts ** Two-by-two analyses, including the calculation of relative risk (RR) and odds ratios (OR), require sufficiently large sample sizes in each cell to produce reliable estimates. ** All cells contain well above the commonly recommended threshold of 5 expected counts, ensuring the stability of our estimates.
- Sufficient Sample Size ** Since we used a Bayesian-augmented approach, exact confidence intervals were computed for RR, OR, and probability differences, mitigating concerns about small sample bias. ** With a total of 7,853 subjects, our analysis meets the requirements for accurate inference

Since these assumptions are satisfied, we can confidently interpret the estimated relative risk, odds ratios, and probability differences.

7.2 Main Analysis

A two by two table with a Bayesian augmentation will be used to assess if there is an association between `RIAGENDR` and `ALQ151`

```
twoby2(matrix(c(862, 352, 2932, 3707), nrow = 2, byrow = TRUE),
       conf.level = 0.95)
```

```
2 by 2 table analysis:
------------------------------------------------------
Outcome   : Col 1
Comparing : Row 1 vs. Row 2

      Col 1 Col 2    P(Col 1) 95% conf. interval
Row 1   862   352      0.7100    0.6839   0.7349
Row 2  2932  3707      0.4416    0.4297   0.4536

                                  95% conf. interval
            Relative Risk: 1.6078    1.5371   1.6818
        Sample Odds Ratio: 3.0962    2.7103   3.5369
Conditional MLE Odds Ratio: 3.0957    2.7057   3.5472
    Probability difference: 0.2684    0.2397   0.2960

            Exact P-value: 0.0000
       Asymptotic P-value: 0.0000
------------------------------------------------------
```

A Bayesian-augmented two-by-two table analysis was conducted to examine the association between gender (`RIAGENDR`) and self-reported past heavy drinking (`ALQ151`), defined as consuming four or five alcoholic drinks almost every day at some point in life.

The probability of a male reporting past heavy drinking is 0.710 (71.0%), with a 95% confidence interval of (0.6839, 0.7349). In contrast, the probability for females is 0.4416 (44.2%), with a 95% confidence interval of (0.4297, 0.4536). This suggests that males are more likely than females to report a history of heavy alcohol consumption.

The relative risk (RR) of reporting past heavy drinking for males compared to females is 1.6078, meaning that males are approximately 1.61 times more likely to report this behavior (95% CI: 1.5371, 1.6818). Additionally, the odds ratio (OR) is 3.0962, indicating that the odds of reporting heavy drinking are over three times higher for males than females (95% CI: 2.7103, 3.5369). The conditional MLE odds ratio, a more precise estimate, is 3.0957 (95% CI: 2.7057, 3.5472).

The absolute probability difference between males and females is 0.2684 (26.8%), with a 95% confidence interval of (0.2397, 0.2960). This represents the estimated difference in heavy drinking prevalence between the two groups. Finally, the exact p-value and asymptotic p-value are both < 0.0001, providing very strong statistical evidence that gender is significantly associated with self-reported past heavy drinking behavior. These findings highlight a notable gender disparity, with males being significantly more likely to report a history of frequent heavy drinking

7.3 Conclusions

The confidence intervals for both males and females do not overlap, indicating a substantial difference in the estimated probability of reporting past heavy drinking when comparing the two genders. The relative risk (RR) of having reported heavy drinking for males versus females was estimated to be 1.61, with a 95% confidence interval of (1.537, 1.682). This suggests that males are approximately 1.61 times more likely to report a history of frequent heavy drinking compared to females. The odds ratio (OR) was estimated to be 3.10, with a 95% confidence interval of (2.710, 3.537), reinforcing that males have significantly higher odds of reporting past heavy drinking behavior. The probability difference between males and females was estimated to be 0.268 (26.8%), with a 95% confidence interval of (0.240, 0.296), further demonstrating the gender disparity in reported drinking habits. The exact p-value was <0.0001, providing strong statistical evidence that gender and heavy drinking history are significantly associated.

A limitation of our analysis is that for the female subjects who reported a history of heavy drinking, the conditional probability was 0.4416 (44.2%), with a 95% confidence interval of (0.4297, 0.4536). While we utilized a Bayesian augmentation to provide more precise estimates for proportions and confidence intervals, the accuracy of this method depends on having sufficiently large sample sizes per group and probability estimates within a reasonable range (10%–90%). Given our large overall sample size (7,853 subjects), these concerns may have been mitigated, but further sampling with more balanced group probabilities would

strengthen the validity of this association.

8 Analysis 4:

8.1 The Question

We are interested in exploring whether an individual's highest level of educational attainment is associated with how frequently they read nutrition labels on food products before making a purchase. To investigate this, we will use the DMDEDUC2 and DBQ750 variables from the NHANES 2017-2020 dataset. DMDEDUC2 represents the highest level of education attained, categorized into five levels: Less than 9th grade, 9-11th grade, High school/GED or equivalent, Some college or an AA degree, and College graduate or above. DBQ750 measures how often a subject reads the nutrition label before purchasing a food item, with responses classified as Always, Most of the time, Sometimes, Rarely, or Never.

Our research question: Is there an association between an individual's level of education and how frequently they read nutrition labels on food products before purchasing them?

8.2 Describing The Data

8.2.1 The 5x5 Table

We will start by examining the total counts and percentages for each group.

```
NH_data_imp %>% tabyl(DMDEDUC2, DBQ750) %>%
  adorn_totals(where = "row") %>%
  adorn_percentages(denominator = "row") %>%
  adorn_pct_formatting() %>%
  adorn_ns(position = "front") %>%
  adorn_title(placement = "combined") %>%
  gt() %>% tab_options(table.width = pct(100)) %>% cols_align(align = "center")
```

DMDEDUC2/DBQ750	Always	Most of the time	Sometimes	Rar
Less Than 9th Grade	111 (18.4%)	61 (10.1%)	179 (29.7%)	97 (16
9-11th Grade	130 (15.4%)	115 (13.6%)	310 (36.7%)	118 (14
High School/GED or Equivalent	238 (12.7%)	350 (18.7%)	684 (36.6%)	309 (16
Some College or AA degree	391 (15.2%)	668 (26.0%)	926 (36.0%)	352 (13
College Graduate or above	348 (17.7%)	716 (36.4%)	672 (34.2%)	161 (8
Total	1,218 (15.5%)	1,910 (24.3%)	2,771 (35.3%)	1,037 (1

For this analysis, we aimed to ensure that each cell contained a minimum of 15 observations to meet assumptions for statistical tests. From the table, we

observe that all education levels have sufficient sample sizes across the five nutrition label reading frequency categories. The highest proportion of individuals who always read nutrition labels are those with a college degree or above (17.7%), while the highest proportion of individuals who never read nutrition labels are those with less than a 9th-grade education (25.6%).

Additionally, the sometimes category consistently has the largest number of individuals across all education levels, with proportions ranging from 29.7% (less than 9th grade) to 36.7% (9-11th grade). This suggests that the majority of individuals, regardless of education level, tend to read nutrition labels only sometimes rather than always or never. However, there appears to be a trend where higher education levels correspond to a greater likelihood of frequently reading nutrition labels.

The smallest observed count is 70 individuals in the college graduate group who reported never reading nutrition labels (3.6%), while the largest observed count is 926 individuals in the some college/AA degree group who reported reading nutrition labels sometimes (36.0%). These findings indicate a potential association between education level and frequency of reading nutrition labels, which will be further examined through statistical testing.

8.3 Main Analysis

8.3.1 Creating the Table for Analysis

A matrix will be created with the values in our 5 by 5 table for analysis:

```
edu_tab <- matrix(c(111,  61, 179,  97, 154,
                    130, 115, 310, 118, 171,
                    238, 350, 684, 309, 289,
                    391, 668, 926, 352, 233,
                    348, 716, 672, 161,  70),
                  ncol = 5, nrow = 5, byrow = TRUE)


rownames(edu_tab) <- c("Less Than 9th Grade", "9-11th Grade", "High School/GED or Equivalent
                       "Some College or AA degree", "College Graduate or above")
colnames(edu_tab) <- c("Always", "Most of the time", "Sometimes", "Rarely", "Never")


edu_tab
```

```
                            Always Most of the time Sometimes Rarely Never
Less Than 9th Grade            111               61       179     97   154
9-11th Grade                   130              115       310    118   171
High School/GED or Equivalent  238              350       684    309   289
Some College or AA degree      391              668       926    352   233
College Graduate or above      348              716       672    161    70
```

8.3.2 The Pearson χ^2 Test

We conducted a Pearson's Chi-Square test to assess whether there is an association between education level (`DMDEDUC2`) and nutrition label reading frequency (`DBQ750`). This test was chosen over Fisher's exact test, which is generally more appropriate for smaller sample sizes, whereas the Pearson Chi-Square test is more reliable for larger contingency tables like ours.

```
chisq.test(edu_tab)
```

```
    Pearson's Chi-squared test

data:  edu_tab
X-squared = 620.88, df = 16, p-value < 2.2e-16
```

The Chi-Square test statistic gives is a χ^2 620.88 with 16 degrees of freedom, yielding a p-value $<$ 2.2e-16. Since the p-value is extremely small, we have strong statistical evidence to reject the null hypothesis, which assumes that education level and nutrition label reading frequency are independent. Instead, these results suggest a significant association between an individual's education level and how frequently they read nutrition labels before purchasing food products. We will perform a post-hoc test to determine where the significant differences lie.

8.3.3 The Cochran Conditions

To validate the assumptions of the Pearson's Chi-Square test, we examine the expected counts for each cell in our 5x5 contingency table.

```
expcount <- NH_data_imp %>% tabyl(DMDEDUC2, DBQ750) %>% chisq.test()

expcount$expected %>% gt() %>% tab_options(table.width = pct(100)) %>% cols_align(align = "c
```

DMDEDUC2	Always	Most of the time	Sometimes	Rarely	Never
Less Than 9th Grade	93.37	146.42	212.42	79.49	70.30
9-11th Grade	130.90	205.28	297.81	111.45	98.55
High School/GED or Equivalent	290.04	454.82	659.85	246.94	218.36
Some College or AA degree	398.61	625.07	906.85	339.37	300.10
College Graduate or above	305.08	478.41	694.07	259.75	229.69

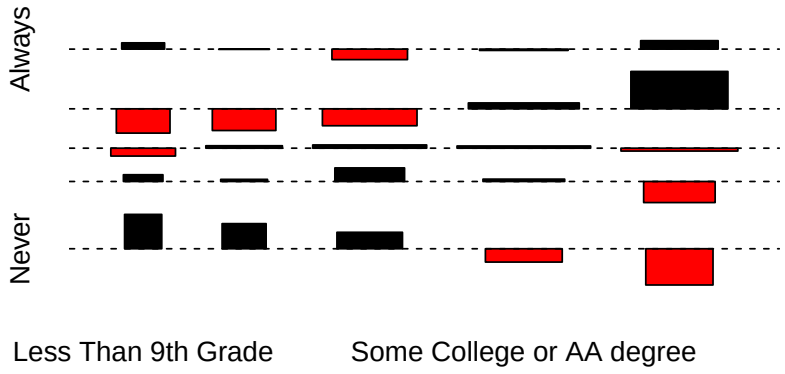
The Cochran conditions require that:

- No cells have an expected count of zero
- At least 80% of the expected counts are 5 or greater

From our table, we confirm that no cells have expected counts of zero and the smallest expected count in our table is 70.30, which is well above the threshold of 5. Since all conditions are met, we can confidently rely on the results of Pearson's Chi-Square test. This ensures that our test statistic is valid, and the association between education level and nutrition label reading frequency is not influenced by violations of assumption.

8.3.4 The Association Plot and Table for the 5x5 Table

```
assocplot(edu_tab)
```



The association plot visually represents the relationship between education level (`DMDEDUC2`) and nutrition label reading frequency (`DBQ750`), highlighting deviations from expected counts. Black bars indicate a higher-than-expected frequency, while red bars indicate a lower-than-expected frequency. The plot shows that individuals with some college or an AA degree are more likely to always read nutrition labels than expected, while those with less than a 9th-grade education are significantly less likely to do so. Similarly, individuals with lower education levels are more likely to rarely or never read nutrition labels, whereas those with higher education levels are less likely to fall into the "Rarely" and "Never" categories. The size of the bars indicates the strength of these deviations, with larger bars reflecting more substantial differences. Overall, the plot reinforces the significant association between education level and nutrition label reading behavior, supporting the conclusion that higher education levels are associated with more frequent nutrition label reading, while lower education levels are linked to less frequent use of nutrition labels.

The following table presents the percent difference between the observed and expected frequencies for the relationship between education level (DMDEDUC2) and nutrition label reading frequency (DBQ750). Positive values indicate that the observed frequency is higher than expected, whereas negative values indicate that the observed frequency is lower than expected.

```
obs <- matrix(c(62,28,109,51,70,
                79,82,207,85,115,
                168,249,488,219,206,
                311,525,713,282,181,
                265,566,539,128,53), ncol = 5, nrow = 5, byrow = TRUE)


exp <- matrix(c(48.99,80.26,113.81,42.35,34.60,
                86.95,142.47,202.01,75.16,61.41,
                203.61,333.59,473.01,176,143.79,
                308.01,504.65,715.56, 266.25,217.52,
                237.44,389.02,551.61,205.24,167.68), ncol = 5, nrow = 5, byrow = TRUE)

freq_table <- ((obs - exp)/exp) *100

freq_table <- as_tibble(freq_table)

rows_names <- c("Less Than 9th Grade", "9-11th Grade", "High School/GED or Equivalent", "Som

tbs_1 <- tibble(Education = rows_names,
        Always = freq_table[[1]],
       `Most of the Time` = freq_table[[2]],
       `Sometimes` = freq_table[[3]],
       `Rarely` = freq_table[[4]],
       `Never` = freq_table[[5]])

tbs_1 %>% gt() %>% tab_options(table.width = pct(100)) %>% cols_align(align = "center") %>%
```

	Always	Most of the Time	Sometimes	Rarely	Never
Less Than 9th Grade	26.56	-65.11	-4.23	20.43	102.31
9-11th Grade	-9.14	-42.44	2.47	13.09	87.27
High School/GED or Equivalent	-17.49	-25.36	3.17	24.43	43.26
Some College or AA degree	0.97	4.03	-0.36	5.92	-16.79
College Graduate or above	11.61	45.49	-2.29	-37.63	-68.39

From the table, we observe that college graduates and those with some college education have a higher observed frequency of reading nutrition labels before purchasing a food product and are less likely to read the label only sometimes,

rarely, or never. Conversely, individuals with a high school education or less are generally less likely to read nutrition labels frequently, except for those with less than a 9th-grade education, who show an unexpectedly higher frequency of reading labels always. These findings further support the association between education level and nutrition label reading behavior, indicating that higher education is linked to more frequent nutrition label usage.

8.4 Conclusions

The test statistic and p-value from the Pearson's chi-squared test provide sufficient evidence to conclude that educational attainment and the frequency of reading nutrition labels are significantly associated. From the association plot, we observed that college graduates had observed frequencies 17.7% and 36.4% higher than expected for the "always" and "most of the time" categories, respectively. Additionally, this group had lower-than-expected frequencies of -34.2%, -37.5%, and -68.3% for the "sometimes", "rarely", and "never" categories, respectively, suggesting that higher education is linked to more frequent nutrition label usage.

In contrast, individuals with less than a 9th-grade education had an observed frequency 25.6% higher than expected for the "never" group and 16.1% higher for the "rarely" group, indicating a lower likelihood of regularly reading nutrition labels. These findings provide strong evidence that education level is associated with nutrition label reading frequency, with higher educational attainment being linked to more frequent label use.

A limitation of this study is the imbalance in group sizes, where some educational categories had significantly more participants than others. To strengthen the reliability of our findings, future research could aim for a more balanced sample across education levels to improve the statistical power of the analysis and minimize potential biases.

9 Session Information

```
# rmarkdown::render("Report.Rmd", output_format = "pdf_document")
```

```
sessionInfo()
```

```
R version 4.4.1 (2024-06-14 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 26100)

Matrix products: default


locale:
```

```
[1] LC_COLLATE=English_United States.utf8
[2] LC_CTYPE=English_United States.utf8
[3] LC_MONETARY=English_United States.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.utf8

time zone: America/New_York
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] lubridate_1.9.3   forcats_1.0.0     stringr_1.5.1     purrr_1.0.2
 [5] readr_2.1.5       tidyr_1.3.1       tibble_3.2.1      tidyverse_2.0.0
 [9] rstatix_0.7.2     broom_1.0.6       patchwork_1.3.0   glue_1.7.0
[13] gt_0.11.0         here_1.0.1        mosaic_1.9.1      mosaicData_0.20.4
[17] ggformula_0.12.0  dplyr_1.1.4       Matrix_1.7-0      ggplot2_3.5.1
[21] lattice_0.22-6    car_3.1-2         carData_3.0-5     sessioninfo_1.2.2
[25] FSA_0.9.6         epiR_2.0.80       survival_3.7-0    Epi_2.55
[29] naniar_1.1.0      tinytex_0.55      janitor_2.2.0     MKinfer_1.2
[33] Hmisc_5.1-3       mice_3.16.0       haven_2.5.4

loaded via a namespace (and not attached):
  [1] rstudioapi_0.16.0       jsonlite_1.8.8          shape_1.4.6.1
  [4] dunn.test_1.3.6         magrittr_2.0.3          jomo_2.7-6
  [7] farver_2.1.2            nloptr_2.1.1            rmarkdown_2.29
 [10] ragg_1.3.3              vctrs_0.6.5             minqa_1.2.8
 [13] askpass_1.2.0           base64enc_0.1-3         htmltools_0.5.8.1
 [16] Formula_1.2-5           mitml_0.4-5             KernSmooth_2.23-24
 [19] htmlwidgets_1.6.4       plyr_1.8.9              zoo_1.8-12
 [22] uuid_1.2-1              lifecycle_1.0.4         cmprsk_2.2-12
 [25] iterators_1.0.14        pkgconfig_2.0.3         R6_2.5.1
 [28] fastmap_1.2.0           snakecase_0.11.1        digest_0.6.37
 [31] numDeriv_2016.8-1.1     colorspace_2.1-1        rprojroot_2.0.4
 [34] textshaping_0.4.0       labeling_0.4.3          fansi_1.0.6
 [37] timechange_0.3.0        abind_1.4-8             mgcv_1.9-1
 [40] compiler_4.4.1          proxy_0.4-27            withr_3.0.1
 [43] fontquiver_0.2.1        pander_0.6.5            htmlTable_2.4.3
 [46] backports_1.5.0         DBI_1.2.3               BiasedUrn_2.0.12
 [49] pan_1.9                 MASS_7.3-60.2           openssl_2.2.1
 [52] classInt_0.4-10         tools_4.4.1             units_0.8-5
 [55] foreign_0.8-86          zip_2.3.1               visdat_0.6.0
 [58] nnet_7.3-19             nlme_3.1-164            grid_4.4.1
 [61] sf_1.0-17               checkmate_2.3.2         cluster_2.1.6
 [64] generics_0.1.3          gtable_0.3.5            tzdb_0.4.0
```

```
 [67] labelled_2.13.0         class_7.3-22             data.table_1.16.0
 [70] hms_1.1.3               xml2_1.3.6               utf8_1.2.4
 [73] foreach_1.5.2           pillar_1.9.0             arrangements_1.1.9
 [76] mitools_2.4             splines_4.4.1            etm_1.1.1
 [79] gmp_0.7-5               tidyselect_1.2.1         fontLiberation_0.1.0
 [82] knitr_1.49              fontBitstreamVera_0.1.1  gridExtra_2.3
 [85] xfun_0.51               mosaicCore_0.9.4.0       stringi_1.8.4
 [88] yaml_2.3.10             boot_1.3-30              evaluate_1.0.0
 [91] codetools_0.2-20        officer_0.6.7            gdtools_0.4.1
 [94] cli_3.6.3               rpart_4.1.23             systemfonts_1.1.0
 [97] munsell_0.5.1           exactRankTests_0.8-35    Rcpp_1.0.13
[100] parallel_4.4.1          miceadds_3.17-44         lme4_1.1-35.5
[103] glmnet_4.1-8            MKdescr_0.8              scales_1.3.0
[106] e1071_1.7-16            ggridges_0.5.6           flextable_0.9.7
[109] rlang_1.1.4
```