group29_project

Ismail Sarah and Steffe Colin

15 December, 2020

1 Introduction

1.1 Overview and motivation

Nowadays, cancer has become the second leading cause of death in the world, a disease that can affect anyone directly or indirectly, through family members or close friends.
Both having either already lost a loved one very quickly and suddenly to the advanced stage of cancer disease or learned that one of them was diagnosed with cancer. We therefore feel intrigued by this disease which remains until today a mystery in terms of medical care, treatment and recommendations to stabilize and ideally reduce the growth of cancer cells in the body.
No one is spared and it can affect people with a very healthy lifestyle just as much as others who do not necessarily care about their health. This is why we have noted with our different experiences and various conversations on the subject that professionals in the field do not necessarily have the same recommendations in terms of food advice and diets, or even contradictory in certain situations.

So, we decided to look at the relationship between - diet and cancer. Knowing that the increase of the disease in the human body is due to several factors including bad eating habits, we ask ourselves the question: what are really these bad eating habits? is meat one of them? Some diets encourage people to eat more meat, such as the super protein diet while other diets strongly discourage meat consumption. Yet in 2015, WHO had classified red meat as a carcinogen.

So, our objectives are to see if there is any relation between the different foods, that is to say vegetables, fruits, meats and cancer. We will also take into account all other variables that are not related to diet such as smoking cigarettes, age, gender, physical activity, alcohol consumption and the economic status of our respondents. Ideally, we would like to bring a more scientific and informative approach to the topic as there is a lot of confusion on this topic due to the wide availability information on the Internet and trends in different regimes.

1.2 Research questions

1. Does eating meat have a relationship with cancer ?
   Our hypothesis is that meat consumption may increase the risk of cancer.
2. Does eating dairy products have a relationship with cancer ?
   Our hypothesis is that dairy products consumption may increase the risk of cancer.
3. Does eating vegetable have a relationship with cancer ?
   Our hypothesis is that vegetable consumption may decrease the risk of cancer.
4. Does eating fruit have a relationship with cancer ?
   Our hypothesis is that fruit consumption may decrease the risk of cancer.
5. Does having an healthy has a effect on cancer ?
   Our hypothesis is that having a healthy diet may decrease the risk of cancer.

At the beginning we wanted to mainly focus on meat, but finally we have decided to broaden our scope. Moreover these questions might evolve once we will be working on the analysis part.

1.3 Related work

WHO report says eating processed meat and red meat are carcinogenic : https://www.hsph.harvard.edu/nutritionsource/2015/11/03/report-says-eating-processed-meat-is-carcinogenic-understanding-the-findings/ (https://www.hsph.harvard.edu/nutritionsource/2015/11/03/report-says-eating-processed-meat-is-carcinogenic-understanding-the-findings/)

Health Concerns About Dairy. From Washington Physician committee responsible medecine:
https://www.pcrm.org/good-nutrition/nutrition-information/health-concerns-about-dairy# (https://www.pcrm.org/good-nutrition/nutrition-information/health-concerns-about-dairy#):~:text=Milk%20and%20other%20dairy%20products,%2C%20ovarian%2C%20and%20prostate%20cancers.

Does having a healthy diet reduce my risk of cancer? From Cancer research UK:
https://www.cancerresearchuk.org/about-cancer/causes-of-cancer/diet-and-cancer/does-having-a-healthy-diet-reduce-my-risk-of-cancer (https://www.cancerresearchuk.org/about-cancer/causes-of-cancer/diet-and-cancer/does-having-a-healthy-diet-reduce-my-risk-of-cancer)

Diet and Physical Activity: What's the Cancer Connection? From American Cancer Society:
https://www.cancer.org/cancer/cancer-causes/diet-physical-activity/diet-and-physical-activity.html (https://www.cancer.org/cancer/cancer-causes/diet-physical-activity/diet-and-physical-activity.html)

2 Data

2.1 Sources

Our data comes from the National Health and Nutrition Examination Survey (https://www.cdc.gov/nchs/nhanes/ (https://www.cdc.gov/nchs/nhanes/)). Each year this institution asks a large amount of questions to a wide representative panel of respondents. Those questions are related to demographic, social economics, health and nutrition. We chose the year 2005 because it is the year that has the

richest nutritional database in terms of variety and it is also the one that we found the most relevant and the most qualitative compared to other years. There is a very strong relationship between all our datasets as the survey participants are labeled. Therefore, many of them responded to every surveys.

For our study we will use 7 datasets :

- Nutrition dataset -> https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/FFQRAW_D.htm (https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/FFQRAW_D.htm)
- Diet Behavior & Nutrition dataset -> https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/DBQ_D.htm#DBQ700 (https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/DBQ_D.htm#DBQ700)
- Medical condition dataset -> https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/MCQ_D.htm#MCQ220 (https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/MCQ_D.htm#MCQ220)
- alcohol dataset -> https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/ALQ_D.htm (https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/ALQ_D.htm)
- Demography dataset -> https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/DEMO_D.htm (https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/DEMO_D.htm)
- Physical activity dataset ->https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/PAQ_D.htm (https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/PAQ_D.htm)
- Smoking - Cigarette Use dataset -> https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/SMQ_D.htm (https://wwwn.cdc.gov/Nchs/Nhanes/2005-2006/SMQ_D.htm)

2.2 Data description

2.2.1 Nutrition dataset

Here is a preview of our first dataset.

```
#> Rows: 6,013
#> Columns: 15
#> $ SEQN     <dbl> 31129, 31131, 31132, 31133, 31134, 31139, 31141...
#> $ WTS_FFQ  <dbl> 36621, 21244, 50102, 6008, 47303, 4176, 5791, 6...
#> $ DRDINT   <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,...
#> $ FFQ_MISS <dbl> 1, 0, 1, 1, 0, 1, 0, 1, 0, 0, 3, 0, 0, 5, 1, 1,...
#> $ FFQ0001  <dbl> 1, 1, 1, 1, 4, 1, 2, 1, 1, 3, 1, 1, 2, 2, 3, 5,...
#> $ FFQ0002  <dbl> 5, 2, 1, 3, 4, 1, 1, 3, 4, 5, 3, 5, 2, 8, 2, 8,...
#> $ FFQ0003  <dbl> 5, 7, 1, 4, 2, 1, 3, 3, 1, 5, 5, 5, 1, 1, 2, 6,...
#> $ FFQ0004  <dbl> 5, 2, 1, 5, 1, 1, 3, 3, 2, 4, 4, 4, 1, 1, 2, 3,...
#> $ FFQ0005  <dbl> 6, 2, 1, 1, 2, 1, 3, 2, 1, 4, 5, 8, 1, 2, 2, 6,...
#> $ FFQ0006  <dbl> 8, 8, 1, 2, 3, 7, 5, 7, 3, 8, 3, 6, 1, 1, 1, 88...
#> $ FFQ0006A <dbl> 1, 1, 88, 1, 3, 5, 1, 2, 88, 1, 1, 1, 88, 88, 8...
#> $ FFQ0007  <dbl> 6, 1, 1, 88, 2, 7, 9, 8, 2, 8, 6, 3, 2, 1, 6, 2...
#> $ FFQ0007A <dbl> 1, 88, 88, 2, 3, 1, 4, 2, 4, 1, 1, 1, 1, 88, 1,...
#> $ FFQ0008  <dbl> 1, 1, 1, 3, 1, 1, 1, 3, 2, 1, 1, 1, 1, 1, 1, 8,...
#> $ FFQ0009  <dbl> 1, 1, 2, 1, 1, 2, 1, 1, 88, 1, 2, 1, 1, 2, 2, 1...
```

This first dataset about nutrition has 6013 observations and 225 variables, it is important to highlight that most of our variables are categorical numeric variables ranged from 1 to 11 :
-1=never
-2=1-6 times per year
-3=7-11 times per year
-4=1 time per month
-5=2-3 times per month
-6=1 time per week
-7=2 times per week
-8=3-4 times per week
-9=5-6 times per week
-10=1 time per day
-11=2 or more times per day
Other values are either blank or error and will be transformed to NA during cleaning process.

In this dataset we will use many variables that are of interest for our analysis, including :
- `SEQN` - Respondent Sequence Number, it will be our reference to merge our different datasets.

- `FFQ0069` - Q.69 Roast beef sandwiches eaten?
- `FFQ0070` - Q.70 Did you eat cold cuts?
- `FFQ0071` - Q.71 Did you eat luncheon ham?
- `FFQ0072` - Q.72 Did you eat other cold cuts?
- `FFQ0074` - Q.74 Did you eat GROUND chicken?
- `FFQ0075` - Q.75 Did you eat beef hamburgers?
- `FFQ0076` - Q.76 Ground beef mixtures eaten?
- `FFQ0077` - Q.77 Did you eat hot dogs?
- `FFQ0078` - Q.78 Other beef mixtures eaten?
- `FFQ0079` - Q.79 Roast beef eaten at other times?
- `FFQ0080` - Q.80 Did you eat steak?

- FFQ0081 - Q.81 Did you eat spareribs?
- FFQ0082 - Q.82 Did you eat roast turkey?
- FFQ0083 - Q.83 Did you eat chicken in mixtures?
- FFQ0084 - Q.84 Did you eat baked chicken?
- FFQ0085 - Q.85 Did you eat baked ham?
- FFQ0086 - Q.86 Did you eat pork?
- FFQ0088 - Q.88 Did you eat liver?
- FFQ0089 - Q.89 Did you eat bacon?
- FFQ0090 - Q.90 Did you eat sausage?

- FFQ0007 - Q.7 How often drink milk as a beverage?
- FFQ0108 - Q.108 Did you eat yogurt?
- FFQ0109 - Q.109 Did you eat cottage cheese?
- FFQ0110 - Q.110 Did you eat cheese?
- FFQ0111 - Q.111 Did you eat frozen yogurt?
- FFQ0112 - Q.112 Did you eat ice cream?
- FFQ0137 - Q.137 Did you eat cream cheese?
- FFQ0138 - Q.138 Did you eat sour cream?

- FFQ0028 - Q.28 Did you eat cooked greens? (such as spinach, turnip, collard, mustard, chard, or kale)
- FFQ0029 - Q.29 Did you eat raw greens? (such as spinach, turnip, collard, mustard, chard, or kale)
- FFQ0030 - Q.30 Did you eat coleslaw?
- FFQ0031 - Q.31 Did you eat sauerkraut?
- FFQ0032 - Q.32 Did you eat carrots?
- FFQ0033 - Q.33 Did you eat string beans?
- FFQ0034 - Q.34 Did you eat peas?
- FFQ0035 - Q.35 Did you eat corn?
- FFQ0036 - Q.36 Did you eat broccoli?
- FFQ0037 - Q.37 Did you eat cauliflower?
- FFQ0038 - Q.38 Did you eat mixed veggies?
- FFQ0039 - Q.39 Did you eat onions?
- FFQ0040 - Q.40 Did you eat peppers?
- FFQ0041 - Q.41 Did you eat cucumbers?
- FFQ0042 - Q.42 Fresh tomatoes eaten?
- FFQ0043 - Q.43 Did you eat summer squash?
- FFQ0044 - Q.44 Did you eat lettuce salads?

- FFQ0015 - Q.15 Did you eat applesauce?
- FFQ0016 - Q.16 Did you eat apples?
- FFQ0017 - Q.17 Did you eat pears?
- FFQ0018 - Q.18 Did you eat bananas?
- FFQ0019 - Q.19 Did you eat pineapple?
- FFQ0020 - Q.20 Did you eat dried fruit?
- FFQ0021 - Q.21 Did you eat peaches?
- FFQ0022 - Q.22 Did you eat grapes?
- FFQ0023 - Q.23 Did you eat melons?
- FFQ0024 - Q.24 Fresh strawberries eaten?
- FFQ0025 - Q.25 Did you eat oranges?
- FFQ0026 - Q.26 Did you eat grapefruit?
- FFQ0027 - Q.27 Did you eat other kinds of fruit?

Here we can observe the proportion of missing values for the selected variables.

```
#>         missing %
#> SEQN          0 0
#> FFQ0069       0 0
#> FFQ0070       0 0
#> FFQ0071       0 0
#> FFQ0072       0 0
#> FFQ0074       0 0
#> FFQ0075       0 0
#> FFQ0076       0 0
#> FFQ0077       0 0
#> FFQ0078       0 0
#> FFQ0079       0 0
#> FFQ0080       0 0
#> FFQ0081       0 0
#> FFQ0082       0 0
#> FFQ0083       0 0
```

There is no NA in this dataset.

2.2.2 Diet Behavior & Nutrition dataset

Here is a preview of our second dataset.

```
#> Rows: 10,348
#> Columns: 15
#> $ SEQN    <dbl> 31127, 31128, 31129, 31130, 31131, 31132, 31133,...
#> $ DBQ010  <dbl> 1, NA, NA, NA, NA, NA, NA, NA, 1, NA, NA, 1, NA,...
#> $ DBD020  <dbl> 121, NA, NA, NA, NA, NA, NA, NA, 152, NA, NA, 91...
#> $ DBD030  <dbl> 121, NA, NA, NA, NA, NA, NA, NA, 182, NA, NA, 15...
#> $ DBD040  <dbl> 121, NA, NA, NA, NA, NA, NA, NA, 152, NA, NA, 91...
#> $ DBD050  <dbl> 304, NA, NA, NA, NA, NA, NA, NA, 0, NA, NA, 365,...
#> $ DBD060  <dbl> 304, NA, NA, NA, NA, NA, NA, NA, 0, NA, NA, 365,...
#> $ DBD072A <dbl> 10, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 10, ...
#> $ DBD072B <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ DBD072C <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ DBD072D <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ DBD072U <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ DBD080  <dbl> 212, NA, NA, NA, NA, NA, NA, NA, 182, NA, NA, 27...
#> $ DBQ700  <dbl> NA, NA, NA, 3, 3, 2, 5, 3, NA, 3, NA, NA, 3, NA,...
#> $ DBQ197  <dbl> NA, 3, 3, 3, 1, 3, 0, 2, NA, 0, 3, 3, 3, 3, 3, 3...
```

This second dataset has 10348 observations and 53 variables. for our analysis we are only going to use two variables :
- `SEQN` - Respondent sequence number. It will be our reference to merge our different datasets.

- `DBQ700` - How healthy is the diet, this variable take a values : 1= Excellent, 2=Very good, 3=Good, 4= fair, 5= poor. all other values are either “refused” or “don’t know” and will be cleaned.

Here we can observe the proportion of missing values for the selected variables.

```
#>        missing  %
#> DBQ700    4209 41
#> SEQN         0  0
```

2.2.3 Medical dataset

Here is a preview of our third dataset.

```
#> Rows: 9,822
#> Columns: 15
#> $ SEQN    <dbl> 31128, 31129, 31130, 31131, 31132, 31133, 31134,...
#> $ MCQ010  <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, ...
#> $ MCQ025  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 1, NA, N...
#> $ MCQ035  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 2, NA, N...
#> $ MCQ040  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ MCQ050  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ MCQ053  <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ...
#> $ MCQ080  <dbl> NA, NA, 2, 2, 2, 2, 1, 1, NA, NA, 1, NA, 2, NA, ...
#> $ MCQ092  <dbl> 2, 2, 9, 1, 2, 2, 2, 2, 2, NA, 2, 2, 2, 2, 2, 2,...
#> $ MCD093  <dbl> NA, NA, NA, 3, NA, NA, NA, NA, NA, NA, NA, NA, N...
#> $ MCQ140  <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, ...
#> $ MCQ149  <dbl> 2, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 2, NA...
#> $ MCQ150G <dbl> 1, 1, NA, NA, NA, 1, NA, NA, 1, NA, 2, 1, 1, 1, ...
#> $ MCQ150Q <dbl> 2, 1, NA, NA, NA, 12, NA, NA, 3, NA, NA, 3, 0, 8...
#> $ MCQ160A <dbl> NA, NA, 2, 2, 2, NA, 2, 2, NA, NA, NA, NA, NA, N...
```

This third dataset is about medical condition has 9822 observations and 89 variables. For our analysis we are going to use :
- `SEQN` - Respondent sequence number, which will be our reference to merge our different datasets.

- `MCQ220` - Ever told you had cancer or malignancy ? This variable is the one that we will use the most in our analysis.
- `MCQ230A` - What kind of cancer
- `MCQ230B` - What kind of cancer
- `MCQ230C` - What kind of cancer
The variable : what kind of cancer exists 4 times(A,B,C,D) in case respondents had multiple cancers, but we will not use the 4th as nobody ever got a 4th cancer.

Here we can observe the proportion of missing values for the selected variables.

```
#>         missing   %
#> MCQ230C    9818 100
#> MCQ230B    9778 100
#> MCQ230A    9413  96
#> MCQ220     4847  49
#> SEQN          0   0
```

Only half of the respondents answered if they had cancer or not, and only 4% provided the type of cancer they got.

2.2.4 Alcohol dataset

Here is a preview of our forth dataset.

```
#> Rows: 4,773
#> Columns: 9
#> $ SEQN    <dbl> 31130, 31131, 31132, 31134, 31144, 31149, 31150,...
#> $ ALQ101  <dbl> NA, 2, 1, 1, 1, 2, 1, 2, 2, 1, 1, 2, NA, 1, 1, 2...
#> $ ALQ110  <dbl> NA, 1, NA, NA, NA, 2, NA, 1, 2, NA, NA, 2, NA, N...
#> $ ALQ120Q <dbl> NA, 0, 4, 2, 2, NA, 7, 0, NA, 0, 3, NA, NA, 4, 3...
#> $ ALQ120U <dbl> NA, NA, 1, 1, 2, NA, 1, NA, NA, NA, 1, NA, NA, 1...
#> $ ALQ130  <dbl> NA, NA, 1, 2, 2, NA, 3, NA, NA, NA, 3, NA, NA, 2...
#> $ ALQ140Q <dbl> NA, NA, 0, 0, 0, NA, 0, NA, NA, NA, 2, NA, NA, 4...
#> $ ALQ140U <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 3, NA, N...
#> $ ALQ150  <dbl> NA, 2, 2, 2, 2, NA, 1, 2, NA, 1, 2, NA, NA, 2, 2...
```

This 4th dataset is about Alcohol consumption, it has 4773observations and 9 variables. For our analysis we are going to use :
- `SEQN` - Respondent sequence number, it will be our reference to merge our different datasets.

- `ALQ130` - Avg # alcoholic drinks/day -past 12 mos

Here we can observe the proportion of missing values for the selected variables.

```
#>        missing  %
#> ALQ130    1953 41
#> SEQN         0  0
```

There is 41% of missing values about alcohol consumption.

2.2.5 Demography dataset

Here is a preview of our fifth dataset.

```
#> Rows: 10,348
#> Columns: 15
#> $ SEQN     <dbl> 31127, 31128, 31129, 31130, 31131, 31132, 31133...
#> $ SDDSRVYR <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4,...
#> $ RIDSTATR <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2,...
#> $ RIDEXMON <dbl> 2, 1, 2, 2, 2, 2, 2, 2, 2, NA, 1, 1, 2, 1, 2, 1...
#> $ RIAGENDR <dbl> 1, 2, 1, 2, 2, 1, 2, 1, 1, 2, 2, 1, 2, 2, 1, 1,...
#> $ RIDAGEYR <dbl> 0, 11, 15, 85, 44, 70, 16, 73, 0, 41, 14, 3, 18...
#> $ RIDAGEMN <dbl> 11, 132, 189, NA, 535, 842, 193, 882, 10, 493, ...
#> $ RIDAGEEX <dbl> 12, 132, 190, NA, 536, 843, 194, 883, 11, NA, 1...
#> $ RIDRETH1 <dbl> 3, 4, 4, 3, 4, 3, 4, 3, 5, 4, 4, 1, 2, 3, 1, 1,...
#> $ DMQMILIT <dbl> NA, NA, NA, 2, 2, 1, NA, 1, NA, 2, NA, NA, 2, N...
#> $ DMDBORN  <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
#> $ DMDCITZN <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,...
#> $ DMDYRSUS <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,...
#> $ DMDEDUC3 <dbl> NA, 4, 10, NA, NA, NA, 9, NA, NA, NA, 6, NA, 13...
#> $ DMDEDUC2 <dbl> NA, NA, NA, 4, 4, 5, NA, 3, NA, 4, NA, NA, NA, ...
```

This 5th dataset is about demography it has 10348 observations and 43 variables. For our analysis we are going to use :
- `SEQN` - Respondent sequence number, which will be our reference to merge our different datasets.

- `RIAGENDR` - Gender
- `RIDAGEYR` - Age at Screening Adjudicated - Recode
- `INDHHINC` - Annual Household Income

Here we can observe the proportion of missing values for the selected variables.

```
#>          missing %
#> INDHHINC     364 4
#> SEQN           0 0
#> RIAGENDR       0 0
#> RIDAGEYR       0 0
```

There is only 4% of missing value about income, and no missing values about gender and age.

2.2.6 Physical activity dataset

Here is a preview of our sixth dataset.

```
#> Rows: 9,424
#> Columns: 20
#> $ SEQN    <dbl> 31128, 31129, 31130, 31131, 31132, 31133, 31134,...
#> $ PAD020  <dbl> NA, 2, 2, 1, 2, 2, 1, 2, 2, NA, 1, NA, 2, NA, 1,...
#> $ PAQ050Q <dbl> NA, NA, NA, 10, NA, NA, 3, NA, NA, NA, 1, NA, NA...
#> $ PAQ050U <dbl> NA, NA, NA, 2, NA, NA, 2, NA, NA, NA, 1, NA, NA,...
#> $ PAD080  <dbl> NA, NA, NA, 20, NA, NA, 30, NA, NA, NA, 120, NA,...
#> $ PAQ100  <dbl> NA, NA, 2, 1, 1, 2, 1, 1, NA, NA, 1, NA, 1, NA, ...
#> $ PAD120  <dbl> NA, NA, NA, 9, 9, NA, 30, 9, NA, NA, 4, NA, 7, N...
#> $ PAD160  <dbl> NA, NA, NA, 60, 60, NA, 120, 180, NA, NA, 25, NA...
#> $ PAQ180  <dbl> NA, NA, 2, 1, 2, 3, 3, 1, NA, NA, 2, NA, 1, NA, ...
#> $ PAD200  <dbl> NA, 1, 2, 2, 2, 2, 2, 1, 1, NA, 1, NA, 1, NA, 1,...
#> $ PAD320  <dbl> NA, 1, 2, 2, 1, 2, 1, 1, 1, NA, 1, NA, 1, NA, 2,...
#> $ PAD440  <dbl> NA, 1, 2, 2, 2, 2, 2, 2, 2, NA, 1, NA, 1, NA, 2,...
#> $ PAD460  <dbl> NA, 30, NA, NA, NA, NA, NA, NA, NA, NA, 30, NA, ...
#> $ PAQ500  <dbl> NA, 1, 3, 2, 3, 3, 3, 2, 3, NA, 1, NA, 3, NA, 3,...
#> $ PAQ520  <dbl> NA, 3, 1, 3, 1, 3, 1, 2, 1, NA, 1, NA, 3, NA, 1,...
#> $ PAQ540  <dbl> NA, NA, 2, 2, 3, NA, 2, 2, NA, NA, NA, NA, NA, N...
#> $ PAQ560  <dbl> 3, NA, NA, NA, NA, NA, NA, NA, NA, 5, NA, 7, NA,...
#> $ PAD590  <dbl> 4, 3, 2, 2, 2, 5, 1, 1, 1, 0, 2, 2, 0, 5, 2, 3, ...
#> $ PAD600  <dbl> 0, 2, 6, 0, 0, 5, 0, 3, 0, 6, 6, 0, 3, 1, 4, 6, ...
#> $ PAAQUEX <dbl> 1, 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, ...
```

This 6th dataset is about physical activity it has 9424 observations and 20 variables. For our analysis we are going to use :
- `SEQN` - Respondent sequence number, which will be our reference to merge our different datasets.

- `PAQ180` - Avg level of physical activity each day

Here we can observe the proportion of missing values for the selected variables.

```
#>        missing  %
#> PAQ180    3291 35
#> SEQN         0  0
```

There is 35% missing values about the physical activity variable.

2.2.7 Smoking - Cigarette Use dataset

Here is a preview of our seventh dataset.

```
#> Rows: 7,186
#> Columns: 15
#> $ SEQN    <dbl> 31129, 31130, 31131, 31132, 31133, 31134, 31136,...
#> $ SMQ020  <dbl> NA, 2, 2, 2, NA, 2, 2, NA, NA, NA, NA, 2, NA, 2,...
#> $ SMD030  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ SMQ040  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ SMQ050Q <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ SMQ050U <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ SMD055  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ SMD057  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ SMD070  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ SMD075  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ SMQ077  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 4, NA, N...
#> $ SMD641  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 2, NA, N...
#> $ SMD650  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 2, NA, N...
#> $ SMD093  <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
#> $ SMDUPCA <chr> "", "", "", "", "", "", "", "", "", "", "", "", ...
```

This 7th dataset is about smoking it has 7186 observations and 39 variables. For our analysis we are going to use :
- `SEQN` - Respondent sequence number , which will be our reference to merge our different datasets.

- `SMD070` - # cigarettes smoked per day now

Here we can observe the proportion of missing values for the selected variables.

```
#>        missing  %
#> SMD070    6298 88
#> SEQN         0  0
```

there is 88% of missing values about the average cigarettes smoked per day.

2.3 Missing values in our data set

About all the NAs we decide to deal with them and filter them everytime we'll use a particular data set in our other parts. Because as we will merge 7 datasets in the end, if we remove all the NAs right at the start we will loose the majority of observations. Otherwise it would mean that we will have only the respondents who answered every questionnaire and every question, and this would not be good since for example only smokers would answer the questionnaire about smoking.

3 Exploratory data analysis

3.1 Food EDA

In this section, we'll explore all of the food variables that may be linked to cancer.

this is a reminder that all the food variables that we are going to explore are categorical as well as numeric variables and they all refer to a certain frequency of consumption.

-1=never
-2=1-6 times per year
-3=7-11 times per year
-4=1 time per month
-5=2-3 times per month
-6=1 time per week
-7=2 times per week
-8=3-4 times per week
-9=5-6 times per week
-10=1 time per day
-11=2 or more times per day
-88=Blank
-99=Error

The frequencies of consumption of all our food variables will be reduced from 11 to 4 frequencies.

-1->Never
-2,3->Low
-4,5,6-> Moderate
-7,8,9,10,11->High
-88,99->NA

3.1.1 Meat

We first start with the meat variable, we create an array (Meat_vars) connecting our respondent number and meat variables.

Meat variables

SEQN	FFQ0069	FFQ0070	FFQ0071	FFQ0072	FFQ0074	FFQ0075	FFQ0076	FFQ0077	FFQ0078	FFQ0079	FFQ0080	FFQ0081
31129	5	5	5	8	5	5	5	5	2	2	3	2
31131	1	6	7	7	1	5	3	5	2	4	2	3
31132	1	1	3	4	1	5	2	2	2	1	5	5
31133	1	5	5	4	7	7	7	5	5	5	5	5
31134	4	5	4	3	1	7	7	5	5	5	4	3
31139	3	2	1	2	1	3	3	1	3	2	2	1

We merge all our meat variables to one variable `meat_cons` .

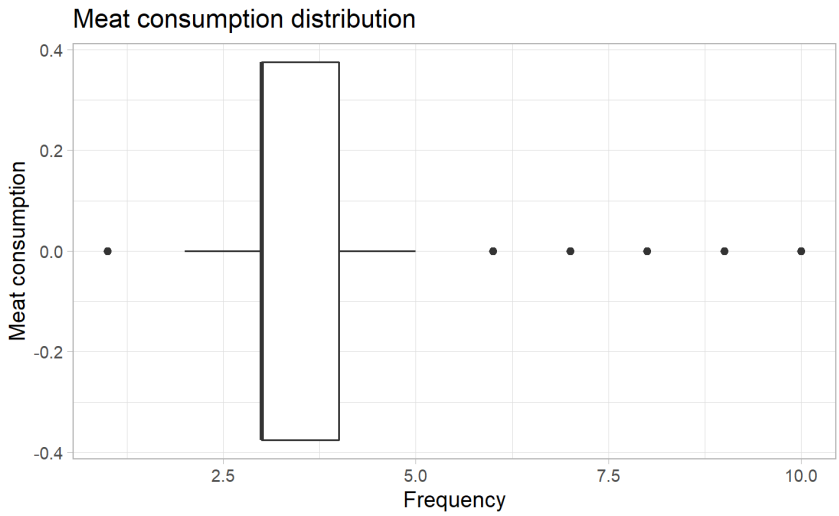
Meat variable

SEQN	meat_cons
31129	4
31131	4
31132	3

SEQN	meat_cons
31133	5
31134	5
31139	2

In order to understand the distribution of meat consumption we make a summary and a boxplot.

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>    1.00    3.00    3.00    3.43    4.00   10.00
```

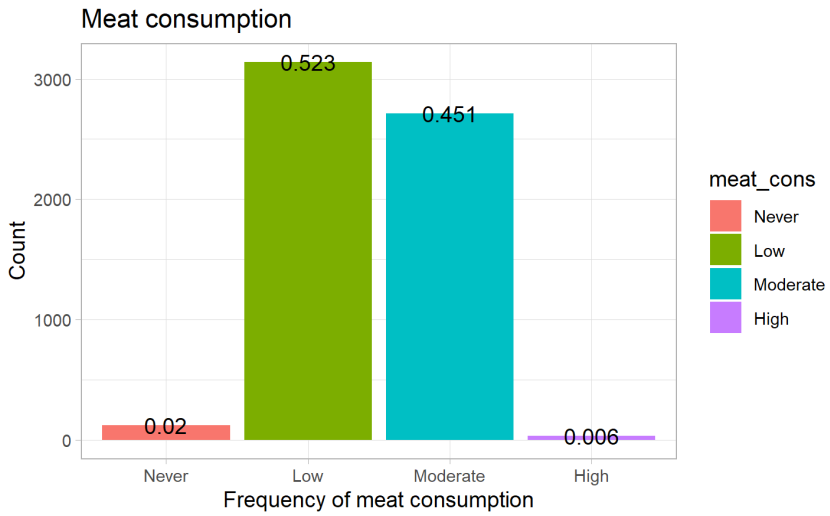


With our method we can see that most of the respondents eat meat 7-11 times per year/1 time per month, with a minimum of 1 (vegetarian) and a max of 10 meaning the person eats meat 1 time per day.

We change the value of the `meat_cons` variable from 1-11 to “never”, “low”, “moderate” and “high”, so it is more understandable. And we display a table to count the number of respondent and the proportion per frequency of meat consumption

Meat consumption count and proportion

meat_cons	count	proportion
Never	119	0.020
Low	3140	0.523
Moderate	2711	0.451
High	35	0.006



We can observe that the main frequency of meat consumption of our respondents is in majority low(52.2%) and moderate(45.1%). There is only 6% with a high consumption and 2% of vegetarians.

3.1.2 Diary Products

We create an array (dairy_vars) connecting our respondent number and all the variables of dairy products.

Dairy product variables

SEQN	FFQ0007	FFQ0108	FFQ0109	FFQ0110	FFQ0111	FFQ0112	FFQ0137	FFQ0138
31129	6	1	1	3	7	3	1	1
31131	1	3	1	8	7	8	3	3
31132	1	1	1	5	1	5	1	1
31133	NA	5	1	7	1	5	1	5
31134	2	1	4	8	1	8	1	3
31139	7	1	1	1	NA	4	1	1

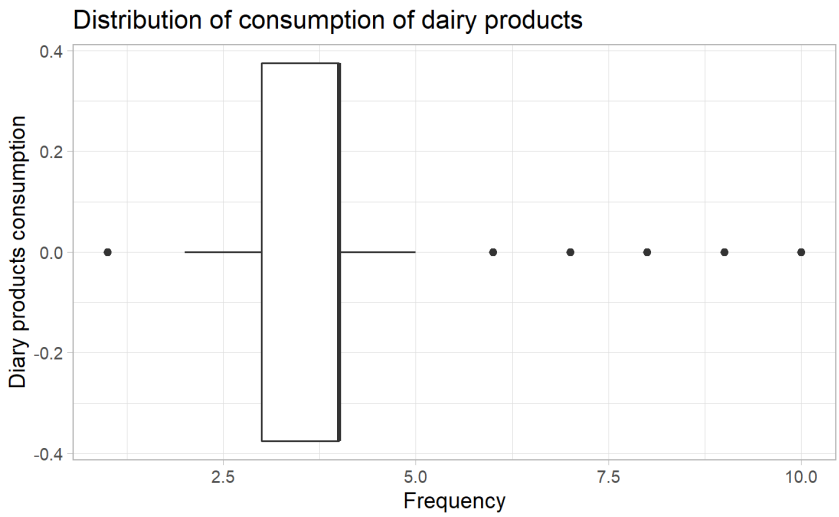
We merge all our dairy product variables to one variable `dairy_cons` .

Diary products variable

SEQN	dairy_cons
31129	3
31131	4
31132	2
31133	4
31134	4
31139	2

In order to understand the distribution of dairy products variable, we make a summary and a boxplot.

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>    1.00    3.00    4.00    3.74    4.00   10.00
```

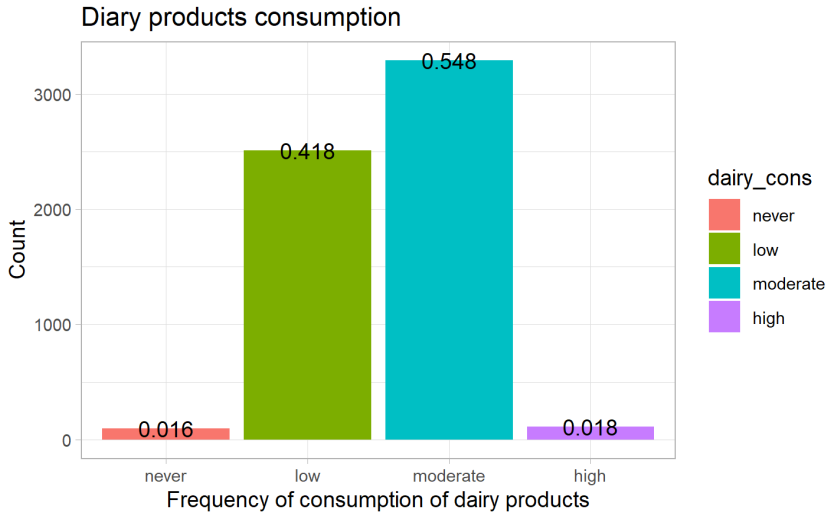


we can see that most of the respondents eat dairy products around 1 time per month with a minimum of 1 (never), and a max of 10, meaning eating dairy products 1 time per day.

We change the value of variable `dairy_cons` from 1-11 to “never”, “low”, “moderate” and “high”, and we display a table to count the number and proportion of respondents according to the frequency of consumption of dairy products.

Count and proportion of consumption of dairy products

dairy_cons	count	proportion
never	94	0.016
low	2509	0.418
moderate	3294	0.548
high	111	0.018



We can observe that the main frequency of dairy productions consumption of our respondents is in majority moderate (56%) and low (41.17%). There is only 1.8% with a high consumption and 1.6% never.

3.1.3 Vegetables

Here we explore our vegetable variable, we create an array (vege_var) connecting our respondent number and our vegetable variables.

Vegetable variables

SEQN	FFQ0028	FFQ0029	FFQ0030	FFQ0031	FFQ0032	FFQ0033	FFQ0034	FFQ0035	FFQ0036	FFQ0037	FFQ0038	FFQ0039	FFQ0
31129	7	2	2	3	2	8	8	1	7	5	8	5	
31131	8	5	2	2	3	7	3	1	9	3	3	8	
31132	3	2	2	1	8	8	5	1	2	2	1	2	
31133	2	1	1	1	1	3	1	1	3	1	1	5	
31134	1	1	3	3	7	7	3	1	3	3	3	7	
31139	3	1	1	4	3	2	3	2	2	1	2	7	

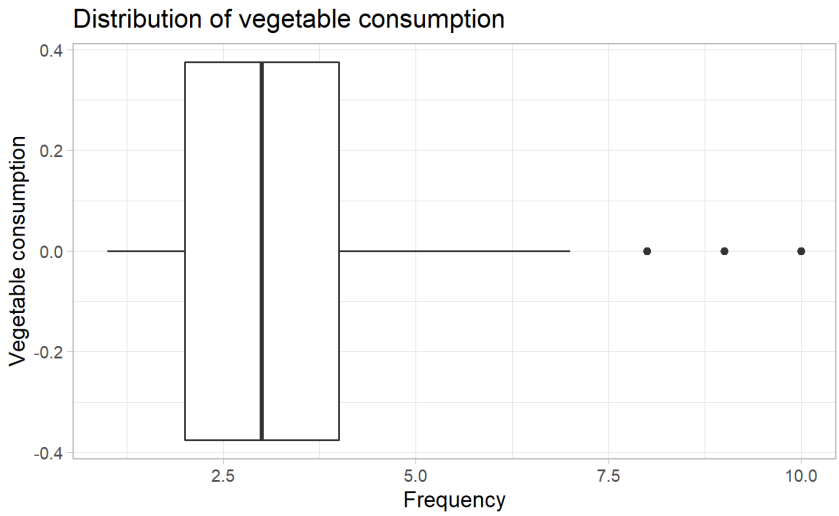
We merge all our vegetable variables to one variable `vege_cons` .

Vegetable variable

SEQN	vege_cons
31129	5
31131	4
31132	3
31133	2
31134	3
31139	3

In order to understand the distribution of vegetable variable, we make a summary and a boxplot.

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>     1.0     2.0     3.0     3.3     4.0    10.0
```

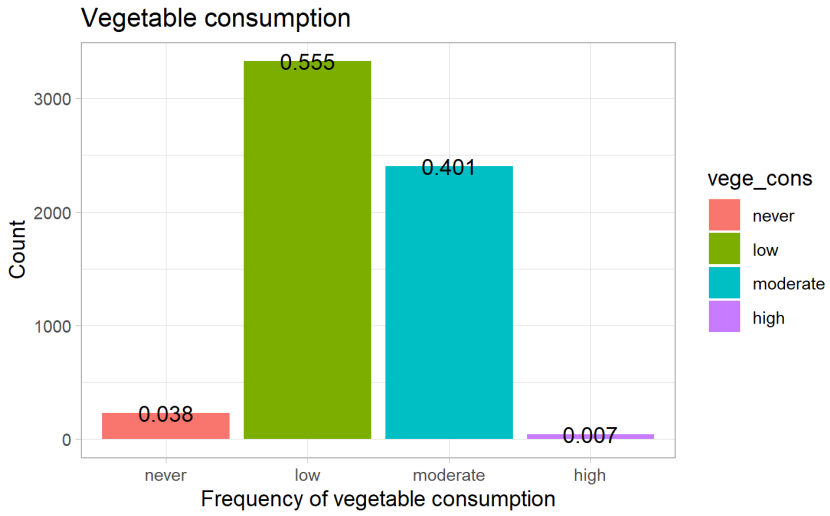


we can see that most of the respondents eat vegetables around 7-11 times per year with a minimum of 1 (never) and a max of 10, meaning eating vegetables products 1 time per day.

We change the value of variable `vege_cons` from 1 to 10 to “never”,“loW”,“moderate”,“high”, and display a table to count the number and proportion of respondents according to the frequency of vegetable consumption.

Count and proportion of vegetable consumption

vege_cons	count	proportion
never	227	0.038
low	3330	0.555
moderate	2405	0.401
high	41	0.007



We can observe that the main frequency of vegetable consumption of our respondents is in majority low(55.4%) and moderate(40.5%). There is only 1% with a high consumption and 3.8% who never eat fruit.

3.1.4 Fruits

Here we explore our fruits variable, we create an array (fruit_var) connecting our respondent number and our vegetable variable.

fruit variables

SEQN	FFQ0015	FFQ0016	FFQ0017	FFQ0018	FFQ0019	FFQ0020	FFQ0021	FFQ0022	FFQ0023	FFQ0024	FFQ0025	FFQ0026	FFQ(
31129	5	7	1	1	5	1	1	9	2	2	2	2	
31131	8	5	1	8	5	7	1	8	1	1	1	2	
31132	1	6	1	6	2	2	1	3	1	1	2	2	
31133	3	2	1	2	1	1	1	3	1	1	1	2	
31134	1	2	2	8	2	5	1	7	1	1	2	1	
31139	1	8	1	10	1	1	1	1	1	1	2	2	

We merge all our fruit variables to one variable `vege_cons` .

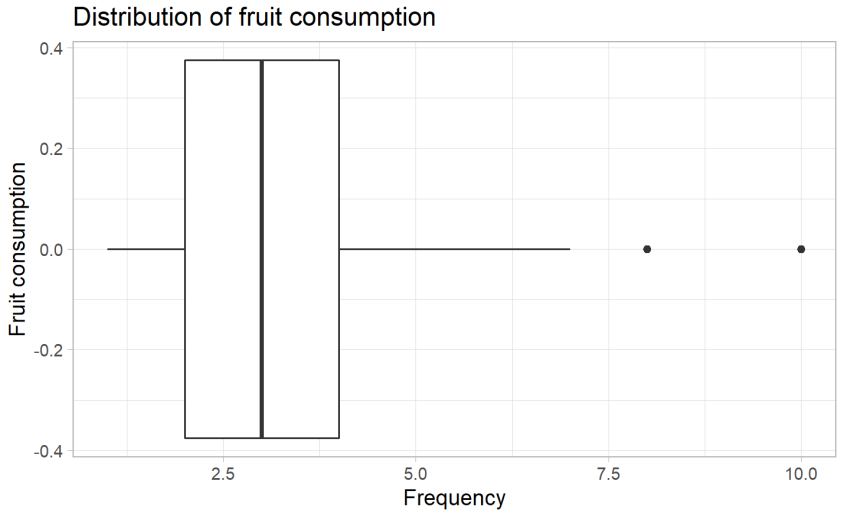
Fruit variable

SEQN	fruit_cons
31129	3
31131	4

SEQN	fruit_cons
31132	2
31133	2
31134	3
31139	2

In order to understand the distribution of fruit variable, we make a summary and a boxplot

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>    1.00    2.00    3.00    2.96    4.00   10.00
```

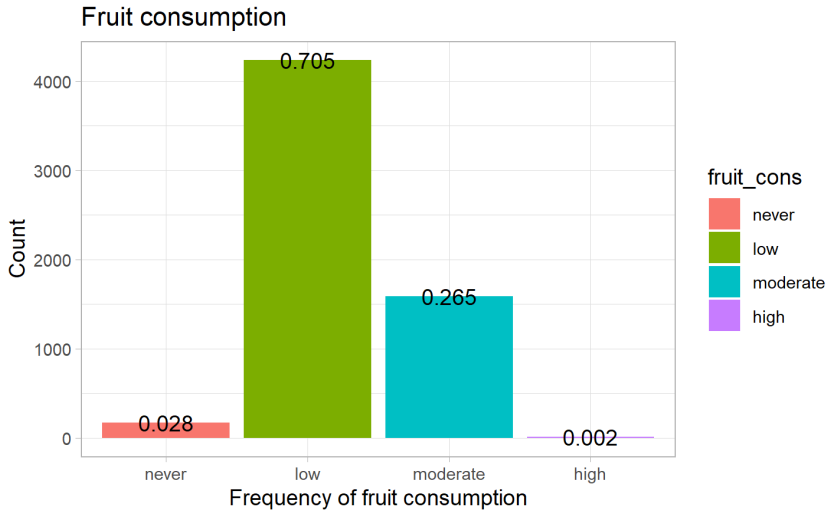


With our method we can see that most of the respondents eat vegetables around 7-11 times per year with a minimum of 1 (never) and a max of 10, meaning eating vegetables products 1 time per day.

We change the value of variable `fruit_cons` from 1 to “never”, “low”, “moderate” and “high” and we display a table to count the number and proportion of respondents according to the frequency of fruit consumption.

Count and proportion of fruit consumption

fruit_cons	count	proportion
never	171	0.028
low	4238	0.705
moderate	1591	0.265
high	10	0.002



We can observe that the main frequency of fruit consumption of our respondents is in majority low(70.5%), then moderate(26.6%). There is 0% with a high consumption and 2.8% who never eat fruit.

3.1.5 Diet

Here we explore the type of diet, we create an array (Dietary_vars) connecting our respondent number and our vegetable variable.

Diteray variables

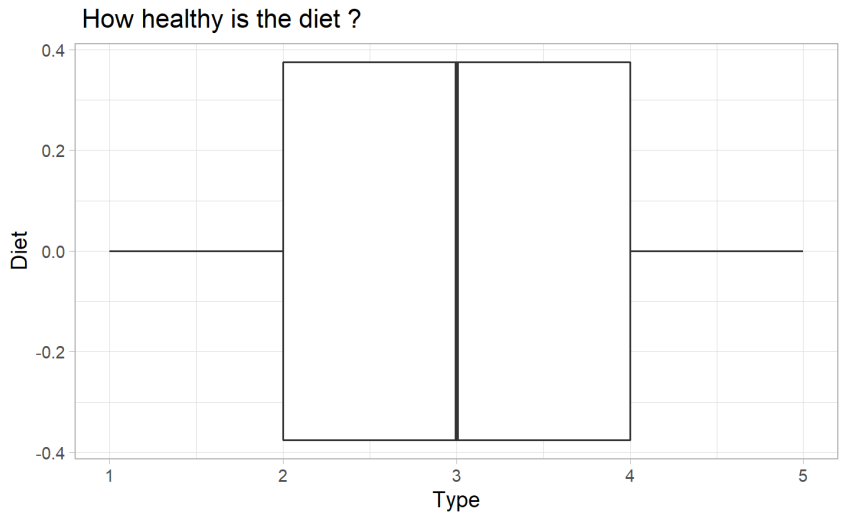
	SEQN	diet
	31130	3
	31131	3
	31132	2
	31133	5
	31134	3
	31136	3

There are different categories defining how healthy the diet is :

1 = Excellent
2 = Very good
3 = Good
4 = Fair
5 = Poor
7 = Refused
9 = Don't know

In order to understand the distribution of this `diet` variable, we make a summary and a boxplot.

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
#>       1       2       3       3       4       5    4222
```



The different categories are now changed to this :
1-> Excellent
2-> Very good
3->Good
4->Fair
5->Poor
7,9->NA

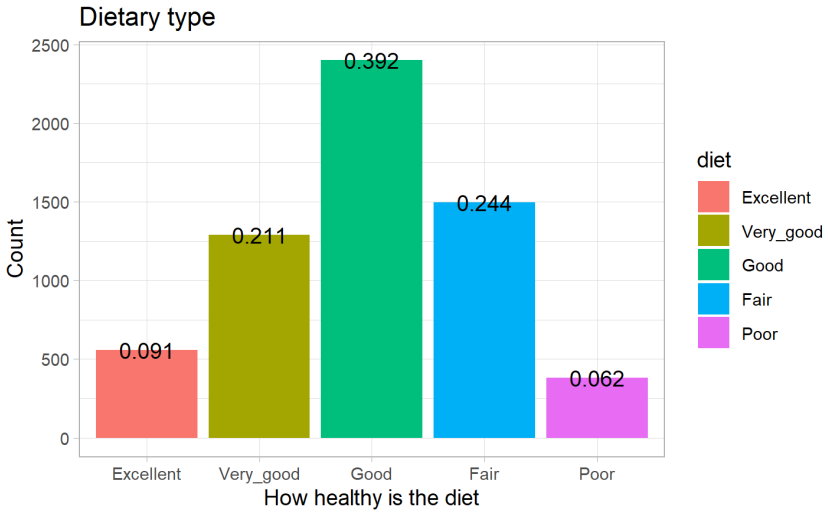
Dietary variables

	SEQN	diet
	31130	Good
	31131	Good
	31132	Very_good
	31133	Poor
	31134	Good
	31136	Good

We display a table to count the number and proportion of respondents.

Diet - Count and proportion

diet	count	proportion
Excellent	559	0.091
Very_good	1290	0.211
Good	2401	0.392
Fair	1496	0.244
Poor	380	0.062



We can observe that the distribution is very similar to a normal distribution.

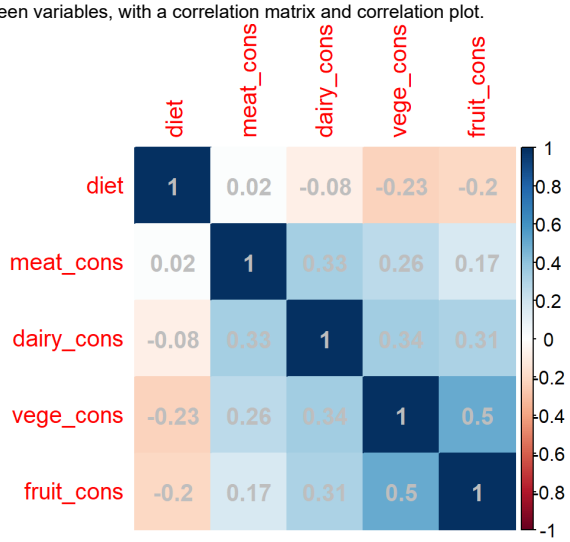
3.1.6 correlations

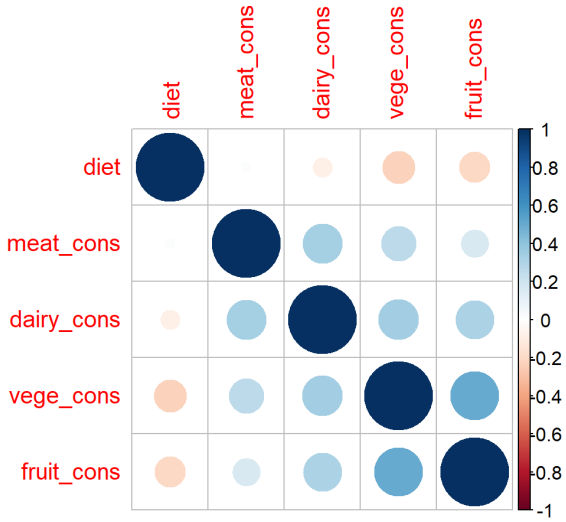
We create a table by joining all the previous food variables by their SEQN number to see if there is a correlation between them.

Variables table

SEQN	meat_cons	dairy_cons	vege_cons	fruit_cons	diet
31129	4	3	5	3	NA
31131	4	4	4	4	3
31132	3	2	3	2	2
31133	5	4	2	2	5
31134	5	4	3	3	3
31139	2	2	3	2	3

We then take a look at our correlations between variables, with a correlation matrix and correlation plot.





We can observe healthy diet is correlated with vegetable (-0,23) and fruit (-0.2) so people who eat more vegetable and fruits tend to state more that they eat healthy. Also can observe that all the different consumption are positively correlated and particularly vegetable and and fruit consumption(0.5).

3.2 Non Food variables EDA

In this section, we will explore all the variables that could be linked to cancer but that are not classified as foods.

3.2.1 Age

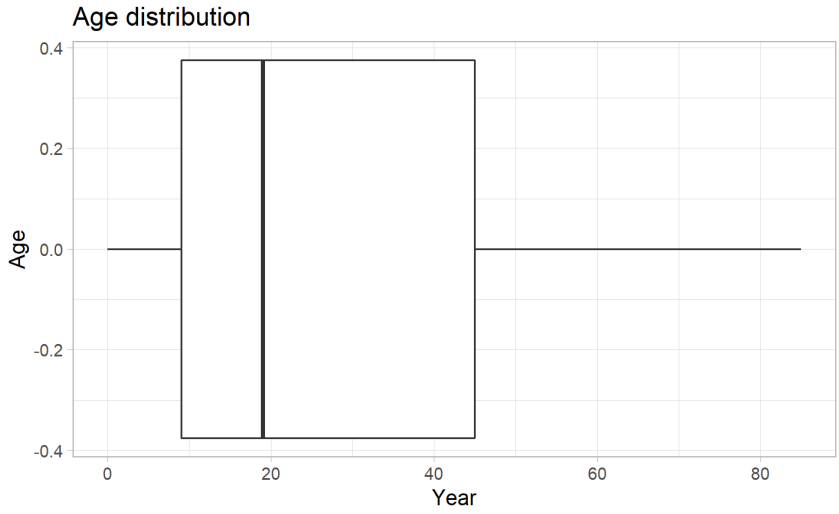
We start first by the variable age, we create a table (age_var) with the variables `SEQN` and `RIDAGEYR` and then we rename `RIDAGEYR` to `age`.

Age variable

SEQN	age
31127	0
31128	11
31129	15
31130	85
31131	44
31132	70

In order to understand the distribution of `age` we make a boxplot and a summary.

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>       0       9      19      28      45      85
```

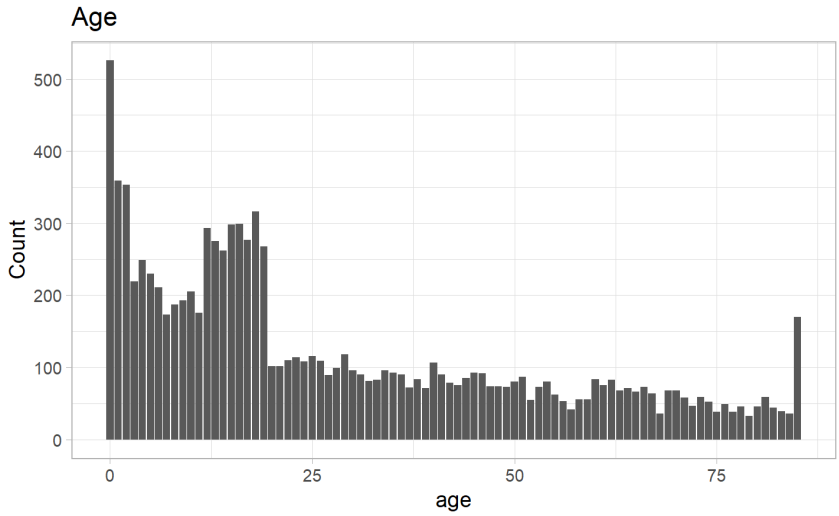


We note that our respondent pool is rather young with 75% under 45 and half of our observation is under 19. In total, we have an average of 28 years. It is important to mention that people over 85 will always tick 85 as this is the maximum age.

We display a table to count the number of respondent and the proportion per age.

Age count and proportion

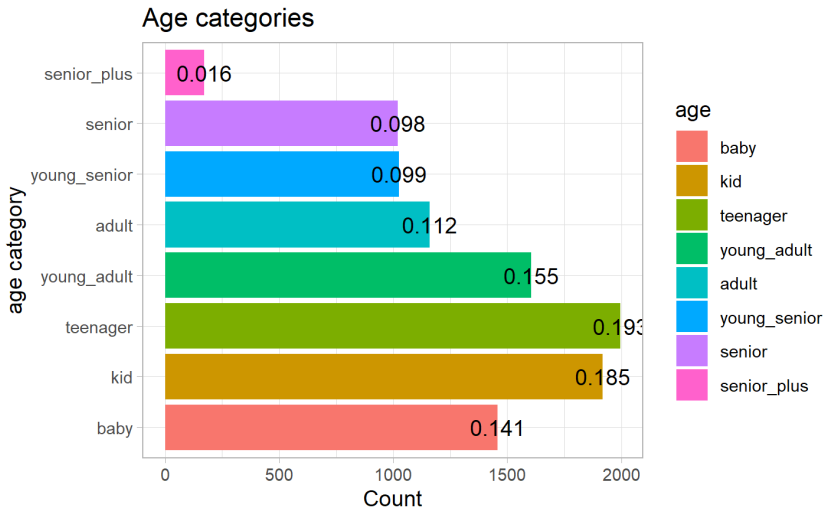
age	count	proportion
0	526	0.051
1	359	0.035
2	353	0.034
18	316	0.031
16	299	0.029
15	298	0.029
12	293	0.028
17	277	0.027



We regroup our data by categories of age.

Age count and proportion by categories of age

age	count	proportion
baby	1457	0.141
kid	1917	0.185
teenager	1995	0.193
young_adult	1606	0.155
adult	1159	0.112
young_senior	1025	0.099
senior	1019	0.098
senior_plus	170	0.016



These plots and tables confirm our previous observation that our respondent pool is rather young.

3.2.2 Gender

We create a table(gender_var) with the variables `SEQN` and `RIAGENDR` and then we rename `RIAGENDR` to `gender`.

Gender variable

SEQN	gender
31127	1
31128	2
31129	1
31130	2
31131	2
31132	1

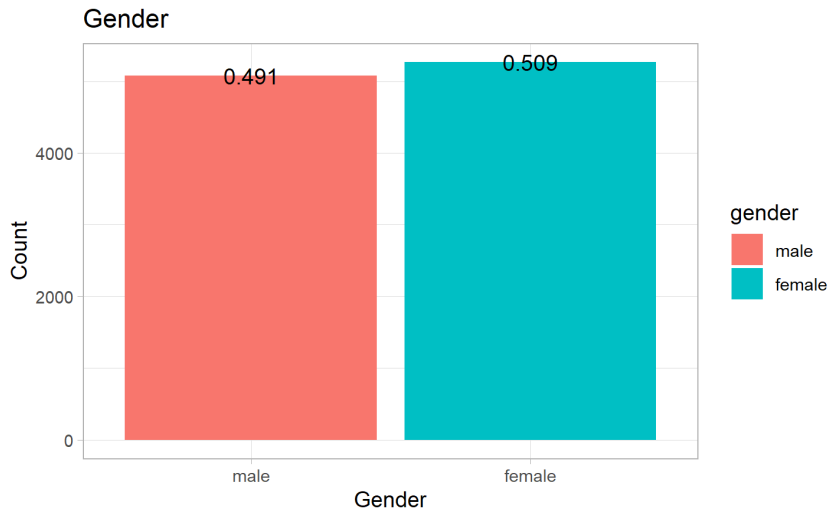
In order to understand the variable we make a summary statistics.

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>    1.00    1.00    2.00    1.51    2.00    2.00
```

We display a table to count the number of respondents and the proportion by `gender`.

Gender count and proportion

gender	count	proportion
male	5080	0.491
female	5268	0.509



We can observe that the gender is fairly well balanced in general but there are 188 more women.

3.2.3 Income

We create a table(income_var) with the variables `SEQN` and `INDHHINC`, and we rename `INDHHINC` to `income`.

income variable

SEQN	income
31127	4
31128	8
31129	10
31130	4
31131	11
31132	11

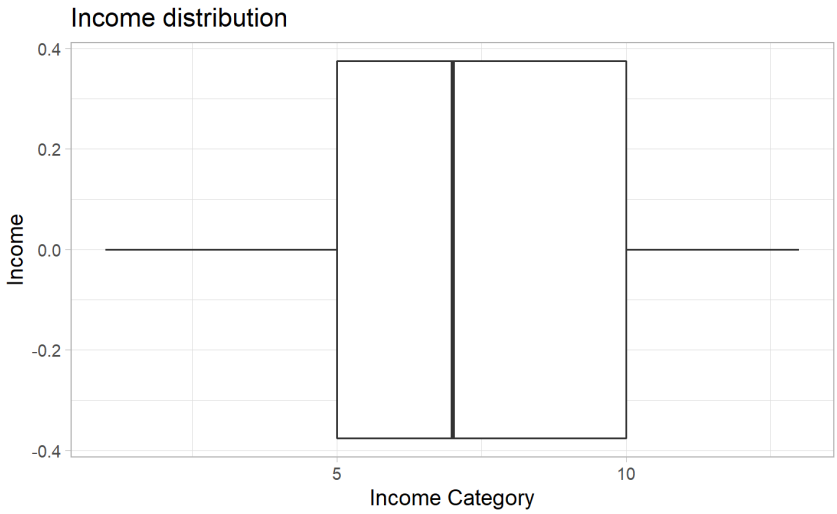
It is important to know that in this variable, `income` is a numeric categorical variable, which refers to certain intervals:

-1 = $ 0 to $ 4,999
-2 = $ 5,000 to $ 9,999
-3 = $10,000 to $14,999
-4 = $15,000 to $19,999
-5 = $20,000 to $24,999
-6 = $25,000 to $34,999
-7 = $35,000 to $44,999
-8 = $45,000 to $54,999
-9 = $55,000 to $64,999
-10 =$65,000 to $74,999

-11 =$75,000 and Over
-12 =Over $20,000
-13 =Under $20,000

In order to understand the distribution of `income` , we make a summary and a boxplot.

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
#>       1       5       7       7      10      13     364
```

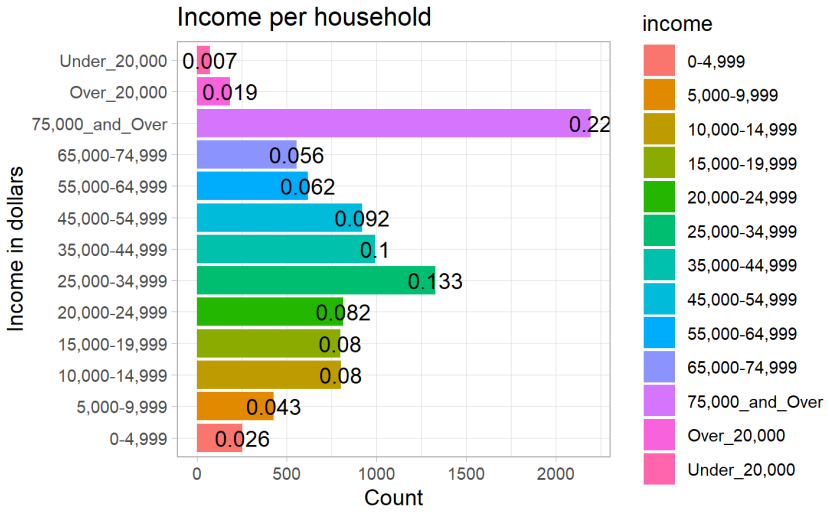


We can observe that our respondent pool has a median income and that the mean equal to 7, that is to say $ 35,000 to $ 44,999. 50% of our pool earn between $ 20,000 and $ 74,999 and 25% earn less than that and the other 25% earn more.

We display a table to count the number of respondent and the proportion by `income` .

Income count and proportion

income	count	proportion
0-4,999	255	0.026
5,000-9,999	428	0.043
10,000-14,999	803	0.080
15,000-19,999	801	0.080
20,000-24,999	818	0.082
25,000-34,999	1331	0.133
35,000-44,999	995	0.100
45,000-54,999	932	0.093



We can observe that in our pool of respondent, household earning more 75,000$ dollars are over represented comparing to other categories.

3.2.4 physical activity

We create a table(physical_var) with our respondent number variable and then we rename `PAQ180` to `avg_physical_activity` .

Physical activity variable

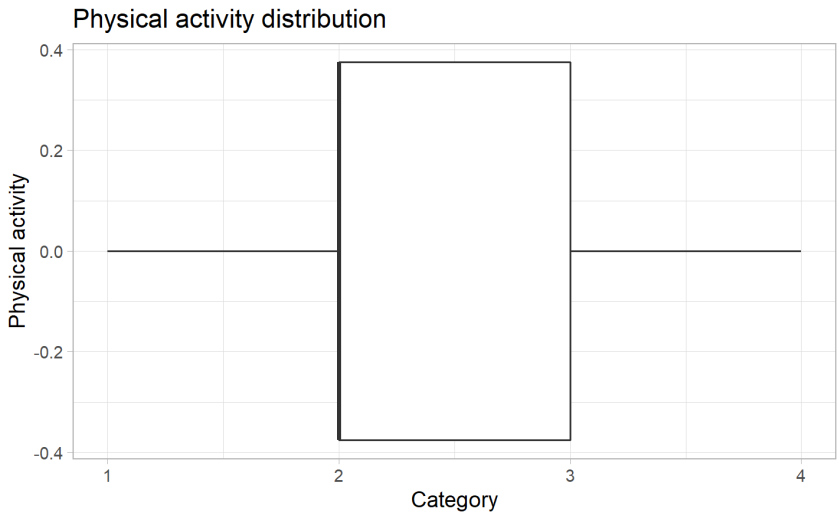
SEQN	avg_physical_activity
31130	2
31131	1
31132	2
31133	3
31134	3
31136	1

It is important to mention that this variable is categorical and numeric, which refers to different levels of activity:

-1 {you sit/he/she sits} during the day and {do/does} not walk about very much.
-2 {you stand or walk/he/she stands or walks} about a lot during the day, but {do/does}not have to carry or lift things very often
-3 {you/he/she} lift(s) light load or {have/has} to climb stairs or hills often.
-4 {you/he/she} {do/does} heavy work or {carry/carries} heavy loads.

In order to understand the distribution of physical activity, we make a summary and a boxplot.

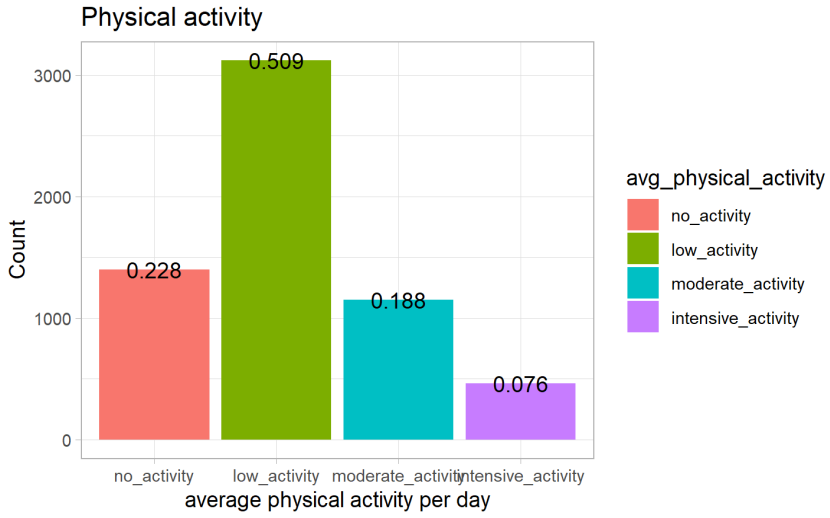
```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
#>       1       2       2       2       3       4    3291
```



We display a table to count the number of respondents and the proportion by physical activity.

Physical activity count and proportion

avg_physical_activity	count	proportion
no_activity	1399	0.228
low_activity	3120	0.509
moderate_activity	1150	0.188
intensive_activity	464	0.076



It is found that 22.8% have practically no physical activity, 50% have a light activity and 25% have a moderate or high physical activity.

3.2.5 Alcohol

we create a table(Alcohol_var) with our variable respondent number and `ALQ130` . Next, we rename `ALQ130` to `avg_alcohol` .

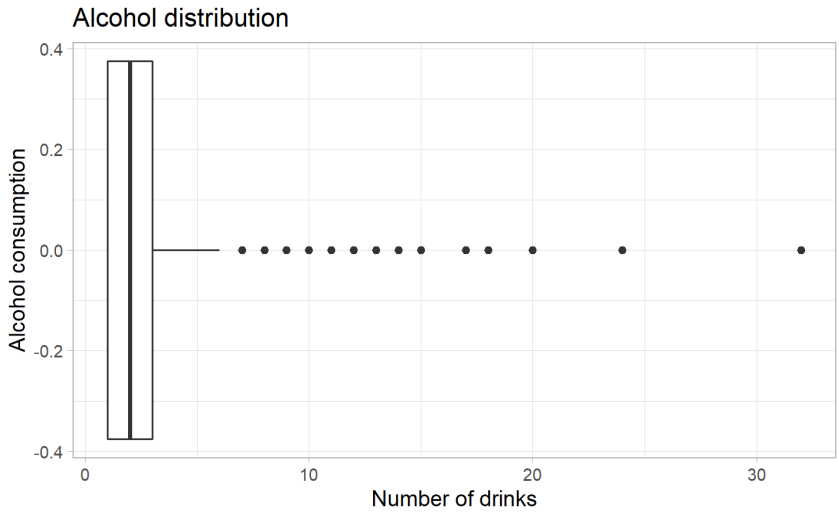
Average consumption variable

SEQN	avg_alcohol
31132	1
31134	2

SEQN	avg_alcohol
31144	2
31150	3
31154	3
31158	2

In order to understand the distribution of the average alcohol consumption, we make a summary and a boxplot.

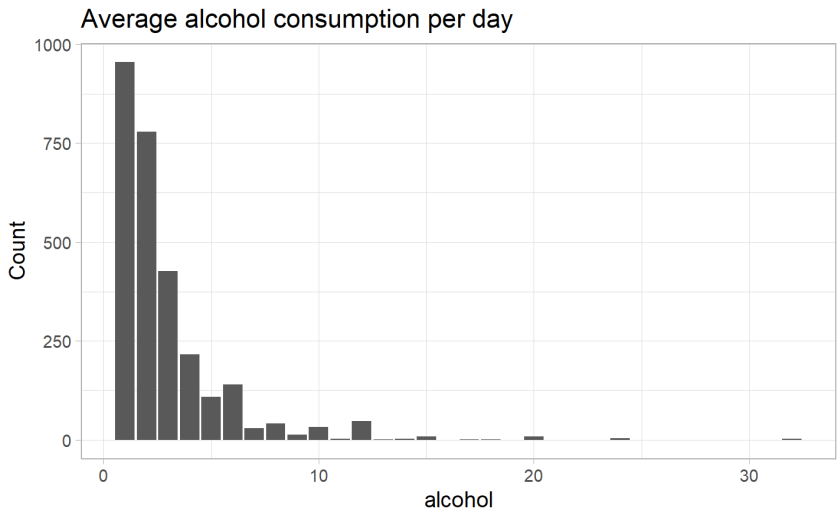
```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
#>       1       1       2       3       3      32    1953
```



We display a table to count the number of respondent and the proportion by average alcohol drinks per day.

Average alcohol consumption count and proportion

avg_alcohol	count	proportion
1	955	0.339
2	779	0.276
3	426	0.151
4	216	0.077
5	109	0.039
6	140	0.050
7	30	0.011
8	43	0.015



It can be seen that most respondents only drink 1 to 3 glasses per day but there is an impressive outlier with 32 glasses per day on average.

3.2.6 smoking

We create a table(Smoking_var) with our variable respondent number and `SMD070` . Then we rename `SMD070` to `cigarets_per_day` .

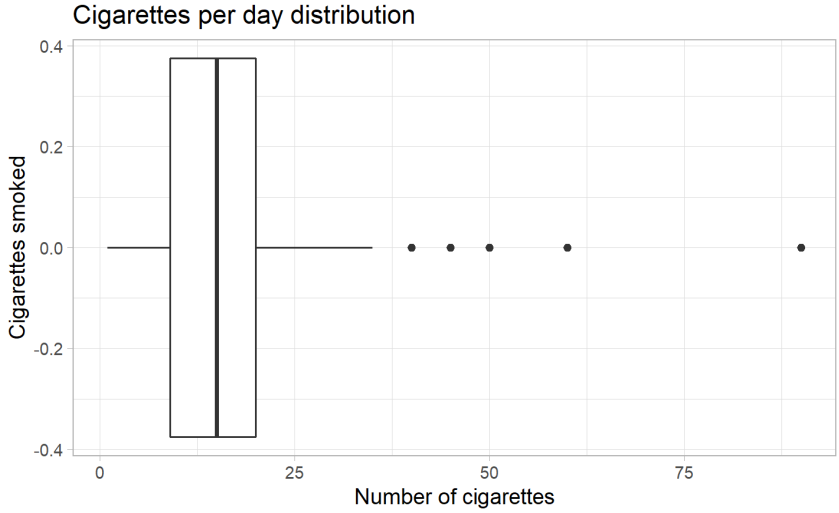
Cigarettes per day variable

SEQN	cigarets_per_day
31154	15
31158	20
31167	20
31186	1
31210	10
31253	15

It is important to note that it is not possible for us to know if a respondent does not smoke because this information was not requested. We therefore have no precise way of knowing whether a respondent's noted NA is due to the fact that he is a non-smoker or if he did not respond.

In order to understand the distribution of `cigarets_per_day` we make a summary and a boxplot

```
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
#>       1       9      15      16      20      90    6298
```

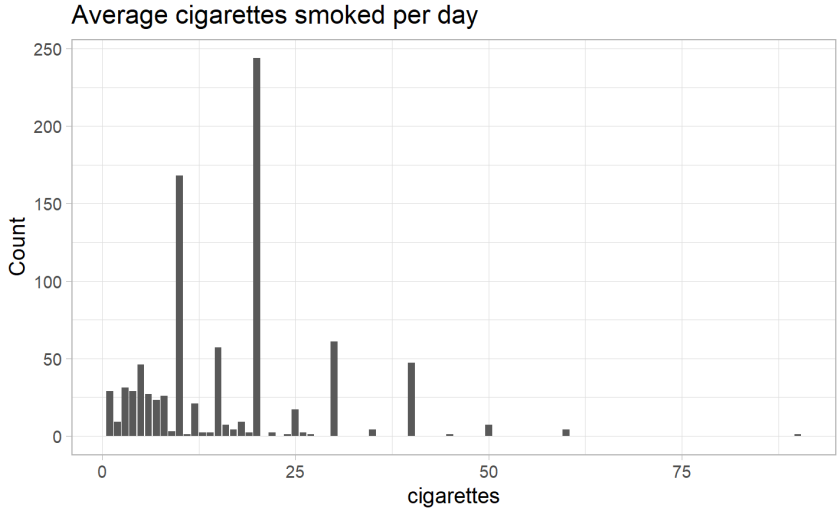


It can be seen that smokers smoke an average of 15 cigarettes and with the interquantile interval, we observe that 50% of smokers smoke 9 to 20 cigarettes per day. There is also an impressive maximum of 90 cigarettes smoked per day.

We display a table to count the number of respondents and the proportion of cigarettes smoked per day

Average cigarettes smoked per day count and proportion

cigarets_per_day	count	proportion
1	29	0.033
2	9	0.010
3	31	0.035
4	29	0.033
5	46	0.052
6	27	0.030
7	23	0.026
8	26	0.030



Here we can clearly see that the respondents approximated their number of cigarettes smoked because most of them answer either 5-10-15-20-30-40. This reminds us that our data is based on declarative observations and is not precise.

3.2.7 Correlations

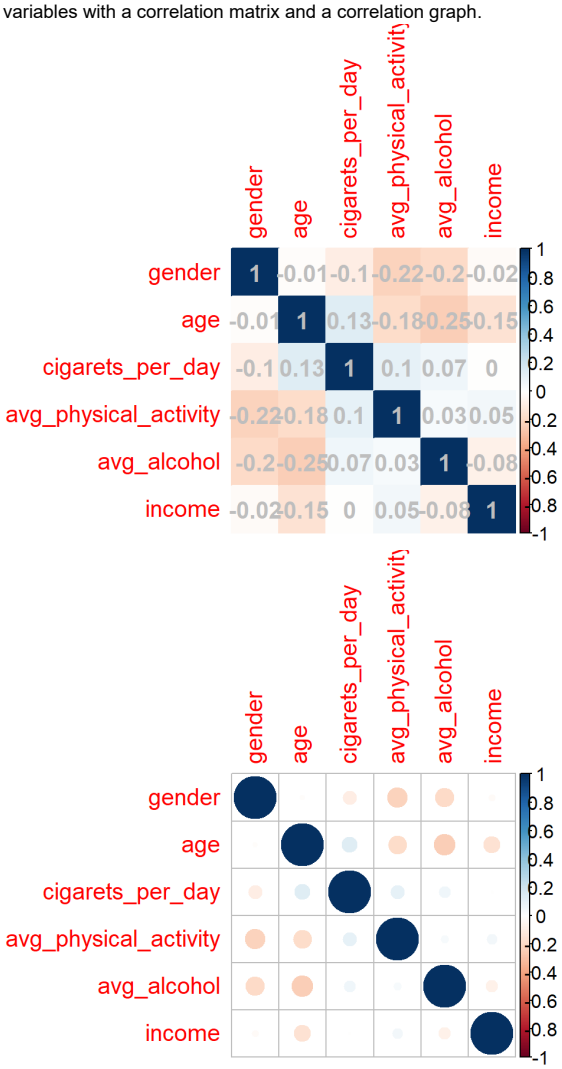
We create a table by joining all the previous non-food related variables by their SEQN number to examine their correlations.

Other variables table

SEQN	age	gender	income	avg_physical_activity	avg_alcohol	cigarets_per_day
31130	85	2	4	2	NA	NA
31131	44	2	11	1	NA	NA
31132	70	1	11	2	1	NA
31134	73	1	12	3	2	NA
31144	21	1	3	2	2	NA

SEQN	age	gender	income	avg_physical_activity	avg_alcohol	cigarets_per_day
31149	85	2	1	1	NA	NA

We then examine our correlations between variables with a correlation matrix and a correlation graph.



There are no strong correlations between the variables. Our highest correlations are :
- Age and average alcohol consumption -24.67%
- Gender and physical activity -22.30%, Gender and alcohol consumption -20%

So this suggests that women do less physical activity, and drink less alcohol, and older people drink less alcohol, these correlations are not strong, but have to be considered. Other correlations are weak.

3.3 Cancer EDA

In this section, we will explore our variables on cancer. Later, we will regroup all our variables seen previously and we will explore the relationships with cancer and finally we will model it, in part 4.

3.3.1 Had cancer

We create a table (cancer_var) with our variable respondent number and `MCQ220` which we will rename `got_cancer`, this variable got_cancer is the one that we will try to find relationship and to model in our analysis.

Ever had cancer variable

SEQN	got_cancer
31130	0
31131	0

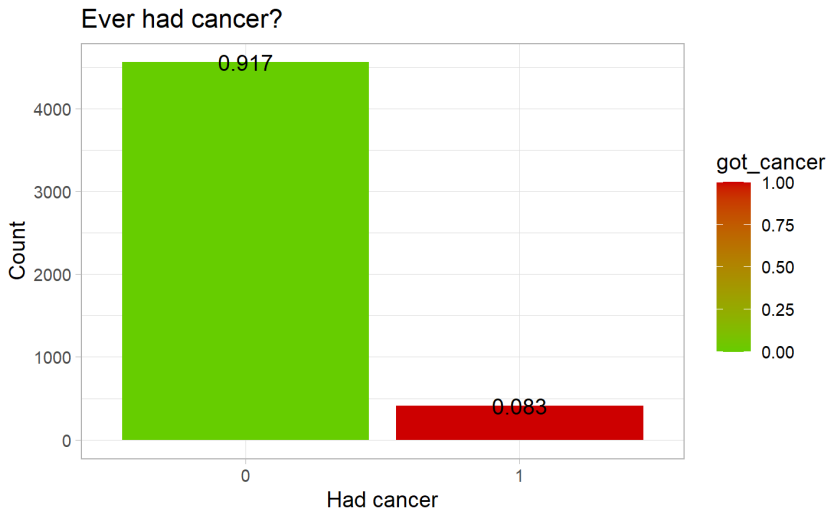
SEQN	got_cancer
31132	0
31134	0
31136	0
31144	0

Our variable got_cancer can only take two values, either 0 if no cancer or 1 if ever had cancer.

We display a table to count the number of respondents and the proportion who ever had cancer or not.

Ever had cancer count and proportion

got_cancer	count	proportion
0	4561	0.917
1	414	0.083



We can observe that 8.3% of our respondents have already had cancer once in their life. You might think this number is high, but in fact it is quite low because we know the lifetime probability of getting cancer is around 40% from the american cancer of society : https://www.cancer.org/cancer/cancer-basics/lifetime-probability-of-developing-or-dying-from-cancer.html (https://www.cancer.org/cancer/cancer-basics/lifetime-probability-of-developing-or-dying-from-cancer.html)
This small result could be due to several reasons, one of which may be because our respondent pool is rather young, as we observed above.

3.3.2 Cancer types

We create a table(cancer_type_var) with our variables respondent number `SEQN` , `MCQ230A` , `MCQ230B` , `MCQ230C` that we will rename to `cancer_typeQ1` , `cancer_typeQ2` and `cancer_typeQ3` .

Type of Cancer

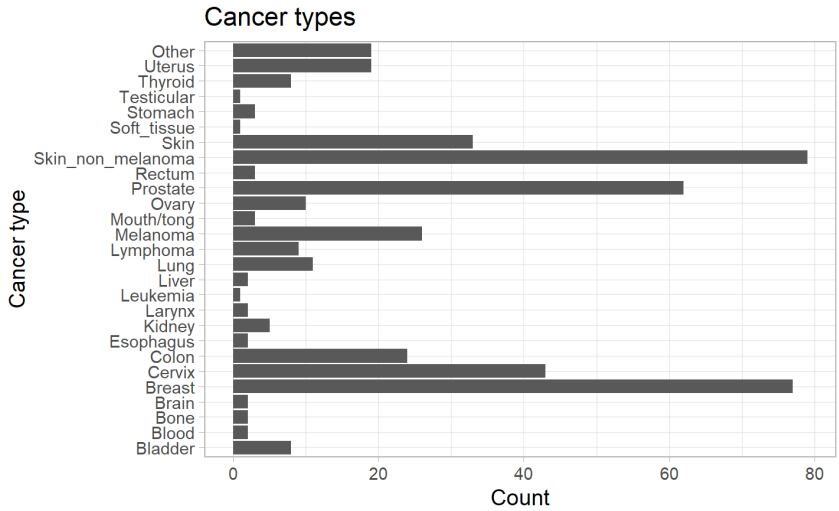
SEQN	cancer_typeQ1	cancer_typeQ2	cancer_typeQ3
31149	16	NA	NA
31150	16	NA	NA
31208	39	NA	NA
31214	14	NA	NA

SEQN	cancer_typeQ1	cancer_typeQ2	cancer_typeQ3
31233	32	NA	NA
31243	30	NA	NA

Now that some respondents have had cancer several times, we need to change the way our data is displayed and for that we use the “pivot longer” function to have only one variable with the type of cancer.

Type of cancer count and proportion

type	count	proportion
Skin_non_melanoma	79	0.173
Breast	77	0.168
Prostate	62	0.136
Cervix	43	0.094
Skin	33	0.072
Melanoma	26	0.057
Colon	24	0.053
Uterus	19	0.042



It is observed that the most frequent cancers are breast, non-melanoma skin, prostate and cervical cancers.

3.4 Relation between food variables and others

In this section, we will study the correlations between the variables related to food and our other variables (age, income, smoking, etc.)

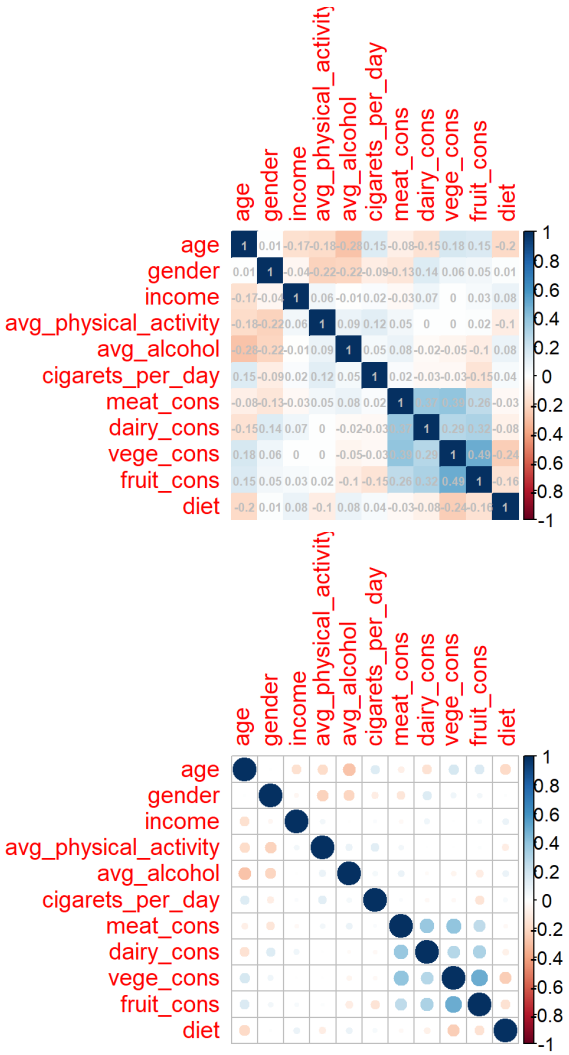
3.4.1 correlations

We first create a table: join_food_other_var, where we will merge all our food and non-food variables. Then we will explore if there are strong correlations between them.

Food variables and others

SEQN	age	gender	income	avg_physical_activity	avg_alcohol	cigarets_per_day	meat_cons	dairy_cons	vege_cons	fruit_cons	diet
31131	44	2	11	1	NA	NA	4	4	4	4	3

SEQN	age	gender	income	avg_physical_activity	avg_alcohol	cigarets_per_day	meat_cons	dairy_cons	vege_cons	fruit_cons	diet
31132	70	1	11	2	1	NA	3	2	3	2	2
31134	73	1	12	3	2	NA	5	4	3	3	3
31144	21	1	3	2	2	NA	5	7	6	5	1
31150	79	1	3	4	3	NA	3	4	6	4	4
31151	59	2	7	1	NA	NA	2	3	3	2	4



There are no strong correlations between the food variables and the other variables, only age and gender seem to have small correlations. we can observe that there is a correlation of -0.2 between age and diet, which means that older people would eat healthier.
We also see that age is negatively correlated with the consumption of dairy products (-0.15) and positively correlated with the consumption of vegetables (+0.18) as well as with the consumption of fruits (+0.15), which would mean that the lower the age, the higher the consumption of dairy products is, and the smaller fruits and vegetables consumption is.
Regarding gender, we can observe a negative correlation with the consumption of meat (-0.13) and a positive correlation with the consumption of dairy products (+0.14), which means that women tend to eat a little less meat and a little more dairy than men, according to our data.
On the other hand, there is no correlation between gender and diet, consumption of vegetables and consumption of fruits. Smoking is also a little negatively correlated with fruit consumption (-0.15).

3.4.2 Most correlated variables investigation

Based on the correlation, we study the relationships between our variables age, gender, meat_cons, dairy_cons, vege_cons, fruit_cons.

Age, Gender and Food variables

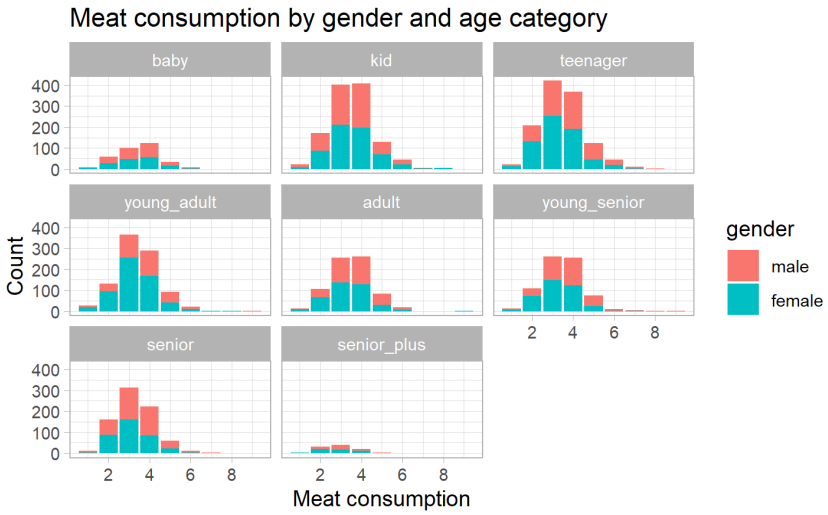
SEQN	age	gender	meat_cons	dairy_cons	vege_cons	fruit_cons
31129	teenager	male	4	3	5	3
31131	adult	female	4	4	4	4
31132	senior	male	3	2	3	2
31133	teenager	female	5	4	2	2
31134	senior	male	5	4	3	3
31139	teenager	female	2	2	3	2

We make a a table of the average consumption by `age` and by `gender`.

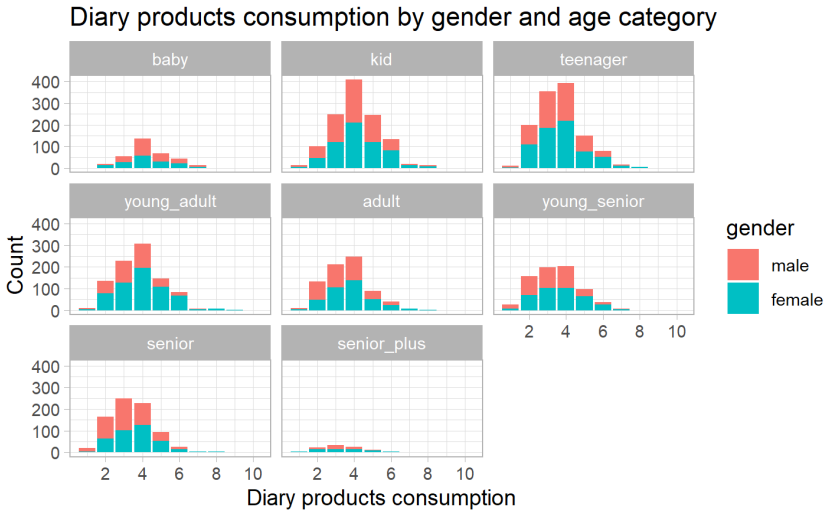
Average consumption of meat, dairy products, vegetables, fruits by age and gender categories

age	gender	avg_meat_cons	avg_dairy_cons	avg_vege_cons	avg_fruit_cons
baby	male	4	5	3	4
baby	female	4	5	4	4
kid	male	4	5	3	4
kid	female	4	5	4	4
teenager	male	4	4	3	3
teenager	female	4	4	3	3
young_adult	male	4	4	4	3
young_adult	female	4	4	4	3

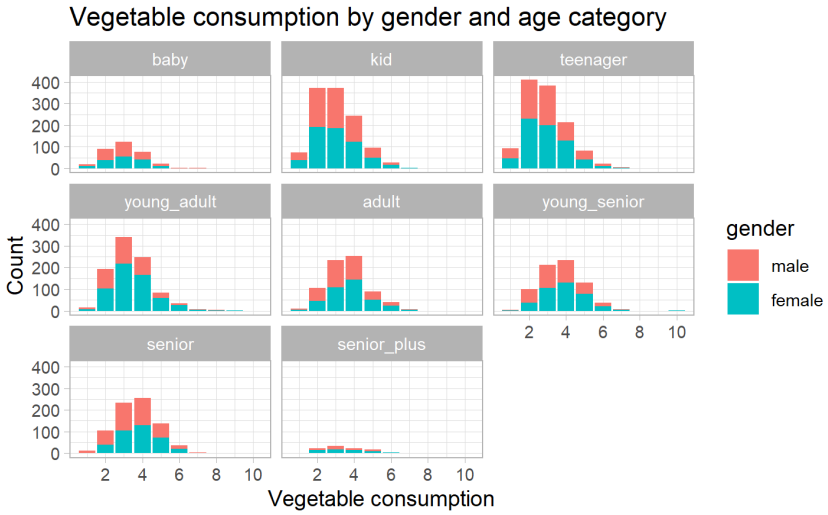
This table shows us that the consumption of dairy products is more present in babies and children after this period the consumption of milk dairy products decreases. Another thing that we could point out would be that babies, children and the elderly eat a little more fruit than adolescents, young people and adults. We don't see the effect of gender.



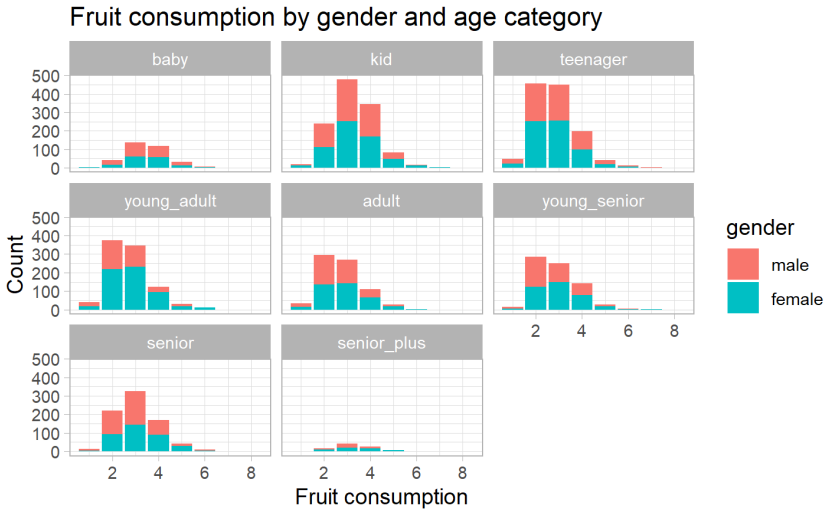
We can observe that the consumption of meat increases during growth i.e. from kid to teenager but this relationship stops from teenager and takes the opposite turn, the consumption of meat decreases later from teenager to senior plus.
It can also be noted that the consumption of men is higher than that of women, as we observe a higher proportion in the high frequencies and a smaller in low frequencies.



Regarding now the consumption of dairy products, we can see that babies, kids, and teenagers consume more often dairy products. In addition, the majority of them consume these products once a month in general and women who consume more dairy products, as we can observe high proportion in higher frequencies.



With age vegetables consumption increase once adult, and women eat more vegetables at every age category.



We don't see any an effect of age, but we can see that women eat more fruit, as the proportion of female increase in the higher frequencies.

3.5 Most important observations recap

About food variables:
we combined and merged the variables into a single variable based on its type, then grouped them into frequency categories. This gave us observations on the frequency of food consumption that were mostly low or moderate. There are no correlations that we could consider important or strong, however there is a slight correlation that we can note is that between the consumption of vegetables and fruits (corr = 0.5).

About other variables that could be linked to cancer:
The pool of respondents is rather young because half of the observations concern people under the age of 19. However, the average is 28 years old. The proportion of male and female sex is balanced. When it comes to income, people earning over $ 75,000 are over-represented. The majority of our respondents do not engage in any physical activity and not even from time to time. Cigarette smoke per day is approximate, we observed that many respondents rounded their consumption to 10, 20, 30, 40 which reminds us that when people take a survey, nothing is really sure or precise. There is no strong correlation between these variables, the highest correlation is a negative correlation, between age and average alcohol consumption of -0.25.

About cancer variables:
8.3% of respondents have had cancer at least once in their life. The most frequent cancers are that of the breast, that of non-melanoma, that of the prostate and finally that of the cervix. The relationship between cancer and other variables will be explored in depth in the analysis section.

About the relationship between food variables and other variables:
There are no strong correlations, only age and sex are slightly correlated with food variables.

4 Analysis

4.1 Approach and Method

We will begin our analysis on the different variables that can influence cancer, by examining the direct relationship between cancer and each of our non-food variables, then we will examine the relationship between cancer and each of our food variables which will be examined further in the detail and depth compared to other variables.

For the non-food variables, we will establish a proportion barplot then we will look at the p-value of the coefficient and so if the latter is significant then it will be integrated into a generalized linear model. For food variables, we will do the same thing but in addition, we will model their individual effects on cancer and we will illustrate the regression.

Finally, we will select all variables significant at the 5% level and train a multivariate model with those selected variables, then we will compare this model to a full model including all variables. We chose the GLM regression with the binary method, because the variable we want to predict “got_cancer” is binary as it takes the value 0 if the respondent has never had cancer and the value of 1 if the respondent has already had cancer. cancer.

At the start of our project, we only wanted to take into account the variables related to our research questions but finally we decided to widen our scope by including more variables that could be confounding, and influence or impact our results.

For the analysis, we create a table where we group all our data so all our explanatory variables, this table will be useful to explain `got_cancer` .

cancer, food and other variables

SEQN	got_cancer	age	gender	income	avg_physical_activity	avg_alcohol	cigarets_per_day	meat_cons	dairy_cons	vege_cons	fruit
31131	0	44	2	11	1	NA	NA	4	4	4	
31132	0	70	1	11	2	1	NA	3	2	3	
31134	0	73	1	12	3	2	NA	5	4	3	
31144	0	21	1	3	2	2	NA	5	7	6	
31150	1	79	1	3	4	3	NA	3	4	6	
31151	0	59	2	7	1	NA	NA	2	3	3	

4.2 Cancer and non food variables

In this section, we will briefly study the relationship with `got_cancer` and the variables which are not related to our research questions but which may have an impact on our interpretation. To study them, we will observe the proportion of cancer as well as the individual significance to determine if these variables would be useful for modeling `got_cancer` .

4.2.1 Cancer and age

We create a table including our `got_cancer` and `age` variables.

Cancer and age variables

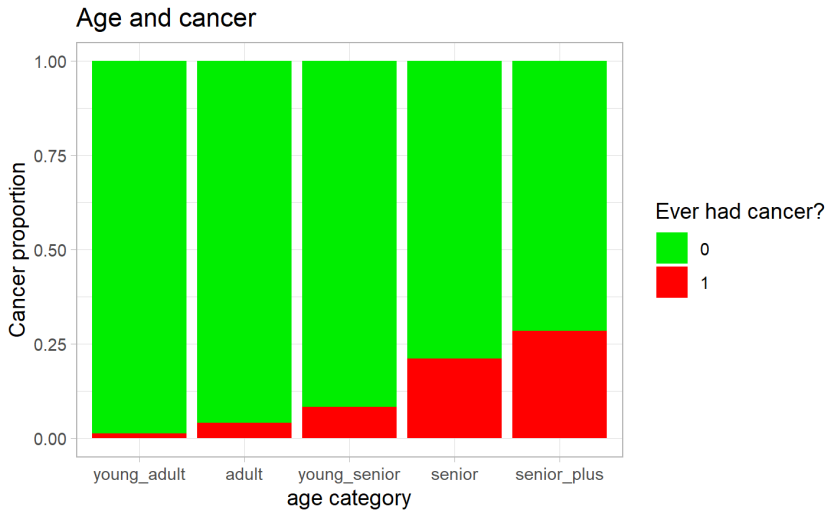
	SEQN	age	got_cancer
	31130	senior_plus	0

SEQN	age	got_cancer
31131	adult	0
31132	senior	0
31134	senior	0
31136	adult	0
31144	young_adult	0

We display a table and a barplot to count the number of cancers and the proportion by age categories.

Cancer by age categories

age	got_cancer	count	proportion
baby	NA	0	NaN
kid	NA	0	NaN
teenager	NA	0	NaN
young_adult	0	1587	0.988
young_adult	1	19	0.012
adult	0	1112	0.959
adult	1	47	0.041
young_senior	0	938	0.917



We can observe that we have no data about babies, kid and teenagers. We can clearly see that the older the respondents, the greater the proportion of them who have had cancer.

We would like to see the effect of the coefficient and if its p-value is significant.

```
#> [1] "coefficient"
#> [1] 0.0635
#> [1] "p-value"
#> [1] 0.00000000000000000000000000000000000000000000000000775
```

We can see that the coefficient has a small positive effect on cancer, but it is significant because its value is close to 0, this variable will be useful for our partial multivariate mode.

4.2.2 Cancer and gender

We create a table including our `got_cancer` and `gender` variables.

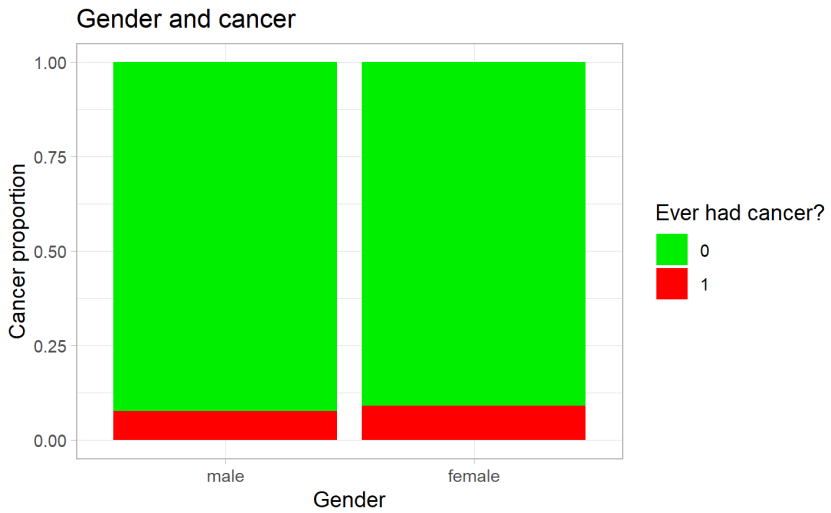
Cancer and gender

	SEQN	gender	got_cancer
	31130	female	0
	31131	female	0
	31132	male	0
	31134	male	0
	31136	female	0
	31144	male	0

We display a table and a barplot to count the number of cancers and the proportion by gender.

Cancer by gender

gender	got_cancer	count	proportion
male	0	2204	0.924
male	1	181	0.076
female	0	2357	0.910
female	1	233	0.090



There is a small difference between men (7.6%) and women (9%), it is necessary to know if this difference of 1.4% is significant or not.

We would like to see the effect of the coefficient and if its p-value is significant.

```
#> [1] "coefficient"
#> [1] 0.0599
#> [1] "p-value"
#> [1] 0.625
```

It can be seen that the coefficient has a small positive effect on cancer, but it is not at all significant because the p-value is very large (0.607). This variable will therefore only be taken into account in the complete model.

4.2.3 Cancer and income

We create a table including `got_cancer` and `income` variables.

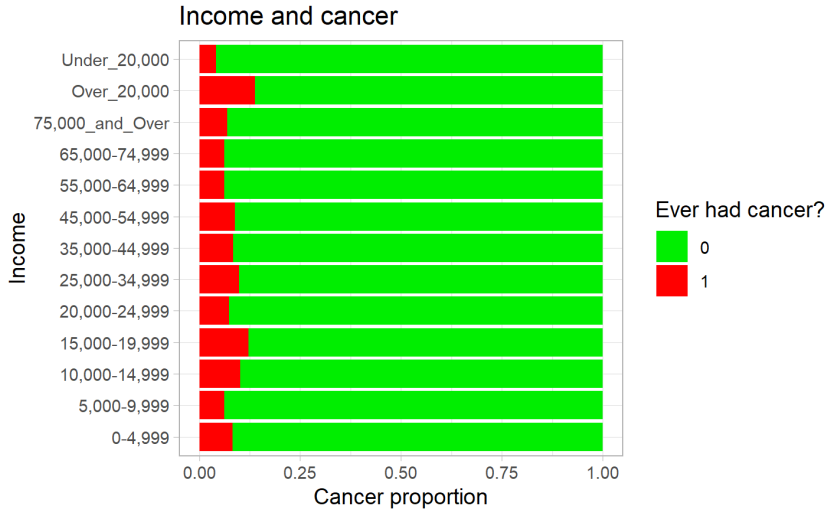
Cancer and income

	SEQN	income	got_cancer
	31130	15,000-19,999	0
	31131	75,000_and_Over	0
	31132	75,000_and_Over	0
	31134	Over_20,000	0
	31136	35,000-44,999	0
	31144	10,000-14,999	0

We display a table and a barplot to count the number of cancers and the proportion by income categories.

Cancer by income categories

income	got_cancer	count	proportion
0-4,999	0	78	0.918
0-4,999	1	7	0.082
5,000-9,999	0	198	0.938
5,000-9,999	1	13	0.062
10,000-14,999	0	311	0.899
10,000-14,999	1	35	0.101
15,000-19,999	0	309	0.878
[illegible]	1	[illegible]	[illegible]



We don’t see any trend in the relationship between income and cancer, our data does not suggest that income is related to cancer.

We would like to see the effect of the coefficient and if its p-value is significant.

```
#> [1] "Coefficient"
#> [1] -0.0347
#> [1] "p-value"
#> [1] 0.0936
```

We can see that the coefficient has a small negative effect on cancer, but it is only significant at the 0.10 level. So this variable will only be used in the full model.

4.2.4 Cancer and physical activity

We create a table including `got_cancer` and `avg_physical_activity` variables.

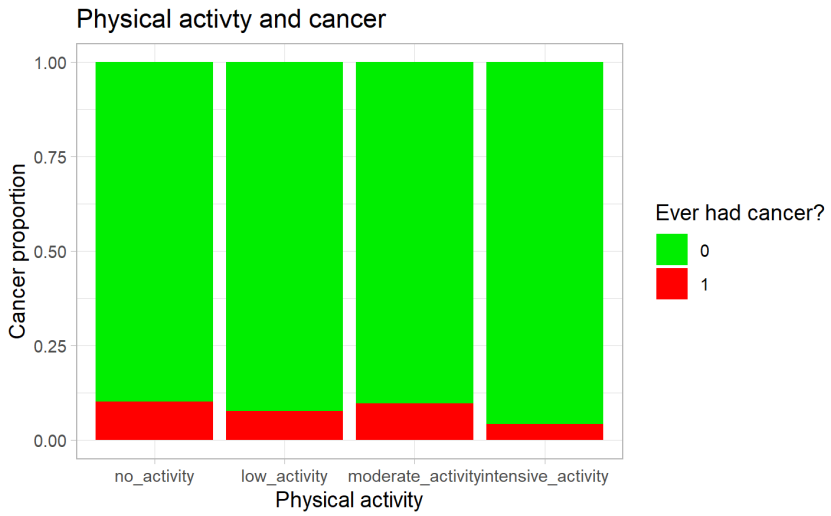
Cancer and physical activity

	SEQN	avg_physical_activity	got_cancer
	31130	low_activity	0
	31131	no_activity	0
	31132	low_activity	0
	31134	moderate_activity	0
	31136	no_activity	0
	31144	low_activity	0

We display a table and a barplot to count the number of cancer and the proportion by income categories.

Cancer by physical activity

avg_physical_activity	got_cancer	count	proportion
no_activity	0	1067	0.898
no_activity	1	121	0.102
low_activity	0	2324	0.924
low_activity	1	192	0.076
moderate_activity	0	794	0.904
moderate_activity	1	84	0.096
intensive_activity	0	371	0.959
intensive_activity	1	16	0.041



We do not see any particular trend, respondents with intensive activity have a little less cancer 4.1% while the others have around 9%.

We would like to see the effect of the coefficient and if its p-value is significant.

```
#> [1] "coefficient"
#> [1] -0.0639
#> [1] "p-value"
#> [1] 0.398
```

We can see that the coefficient has a small negative effect on cancer but it is not significant since the p-value is very large. This variable will only be used in the full model.

4.2.5 Cancer and alcohol

We create a table including `got_cancer` and `avg_alcohol` variables.

Cancer and Average alcohol consumption

SEQN	avg_alcohol	got_cancer
31130	NA	0
31131	NA	0
31132	1	0
31134	2	0
31144	2	0
31149	NA	1

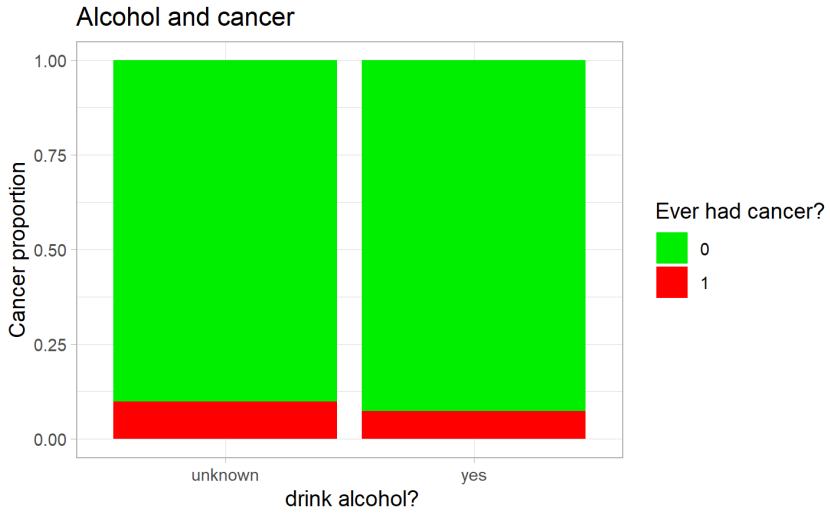
About this variable `avg_alcohol` , it's a bit special because we only have the information on the number of drinks, but there is no answer if the respondents do not drink, so the undetermined proportion (NA), maybe people who didn't answer because they don't drink alcohol at all, or it could just be people who didn't answer the question.

We are now investigating whether people who did not respond about their alcohol consumption have a different proportion of cancer, but these results need to be taken with a lot of hindsight.

declared alcohol drinkers and cancer

drink	got_cancer	count	proportion
unknown	0	1761	0.903
unknown	1	190	0.097

drink	got_cancer	count	proportion
yes	0	2613	0.927
yes	1	205	0.073

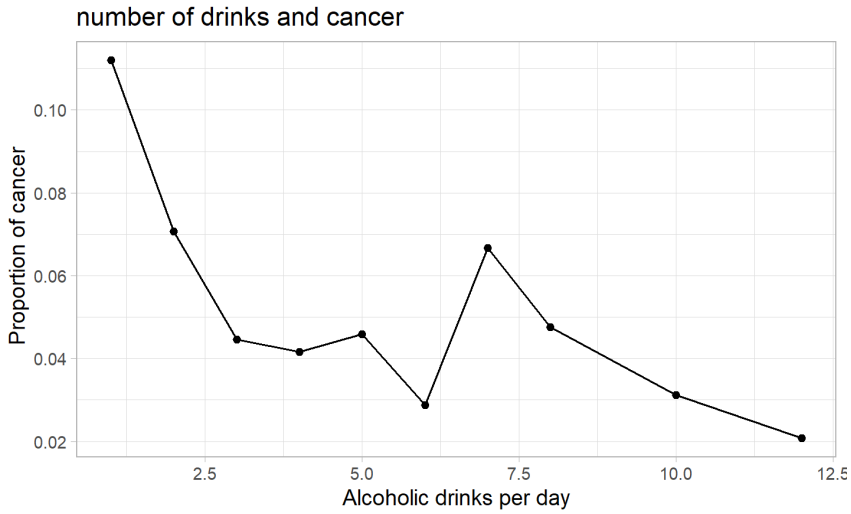


We observe that people who say they drink have less cancer (7.3%) than those who did not answer the question (9.7%).

Next we want to look at the amount of drinks per day and cancer. We make a table with proportion in number of drinks and a scatterplot.

Alcohol quantity and cancer

avg_alcohol	got_cancer	count	proportion
1	0	848	0.888
1	1	107	0.112
2	0	723	0.929
2	1	55	0.071
3	0	407	0.955
3	1	19	0.045
4	0	207	0.958
4	1	9	0.042



We can observe a negative relationship between cancer and alcohol, which is quite surprising. But don't forget that NA has a double meaning in this alcohol variable, which does not necessarily give us a very realistic result.

We would like to see the effect of the coefficient and if its p-value is significant.

```
#> [1] "coefficient"
#> [1] -0.31
#> [1] "p-value"
#> [1] 0.00000255
```

We see that the coefficient has a negative effect on cancer and is very significant at its p-value close to 0. This variable will still be used in the partial model.

4.2.6 Cancer and smoking

We create a table including cancer and smoking variables.

Cancer and Cigarettes per day

SEQN	cigarets_per_day	got_cancer
31154	15	0
31158	20	0
31167	20	0
31186	1	0
31210	10	0
31253	15	0

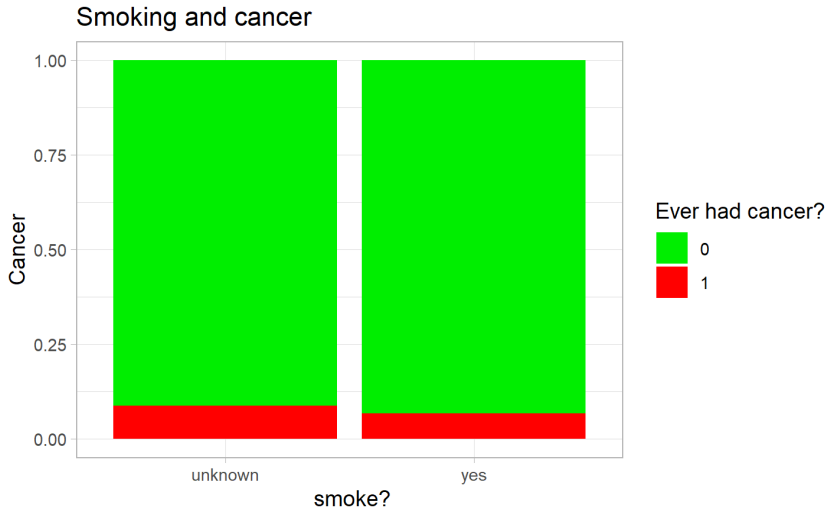
About this variable, it is the same as for alcohol because we only have the information on the number of cigarettes but there is no answer for the respondents who do not smoke, therefore the undetermined proportion (NA) , maybe people who didn't respond because they don't smoke at all, or it could just be people who didn't respond.

We are now investigating whether people who did not respond about their cigarette consumption have a different proportion of cancer, but these results need to be taken with a lot of hindsight.

Smoking and cancer

smoke	got_cancer	count	proportion
unknown	0	3732	0.913
unknown	1	355	0.087

smoke	got_cancer	count	proportion
yes	0	829	0.934
yes	1	59	0.066

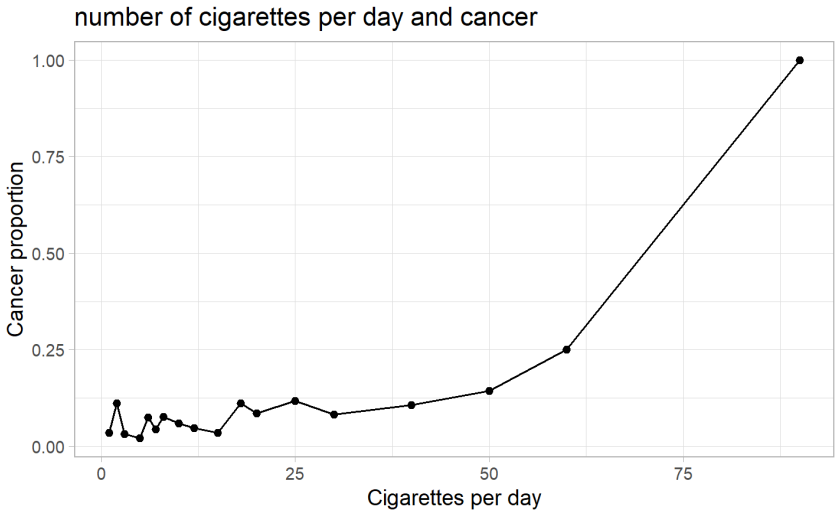


We observe that people who report smoking have a lower proportion of cancer (6.6%) than those who did not answer the question (8.7%).

Next we want to look at the number of cigarettes per day and cancer. we make a table with proportion in number of cigarettes and a scatterplot.

Cirattes per day and cancer

cigarets_per_day	got_cancer	count	proportion
1	0	28	0.966
1	1	1	0.034
2	0	8	0.889
2	1	1	0.111
3	0	30	0.968
3	1	1	0.032
4	0	29	1.000
5	0	45	0.978



We can see a positive trend but we have to take into account that there is an outlier of 90 cigarettes smoked per day and who had cancer, so it makes it look like the relationship is strong, but this it's not the case.

We would like to see the effect of the coefficient and if its p-value is significant.

```
#> [1] "coefficient"
#> [1] 0.0349
#> [1] "p-value"
#> [1] 0.00597
```

We can see that the coefficient has a positive effect on cancer, and the p-value is significant at a level of 0.01. This variable will be used in the partial model.

4.3 Cancer and food variables relations

In this section, we will explore in more depth the relationship with `got_cancer` and the variables related to our research questions. To study them, we will observe the proportion of cancer with these variables then we will model `got_cancer` , illustrate the relationship between each variable individually with got_cancer and finally interpret the results.

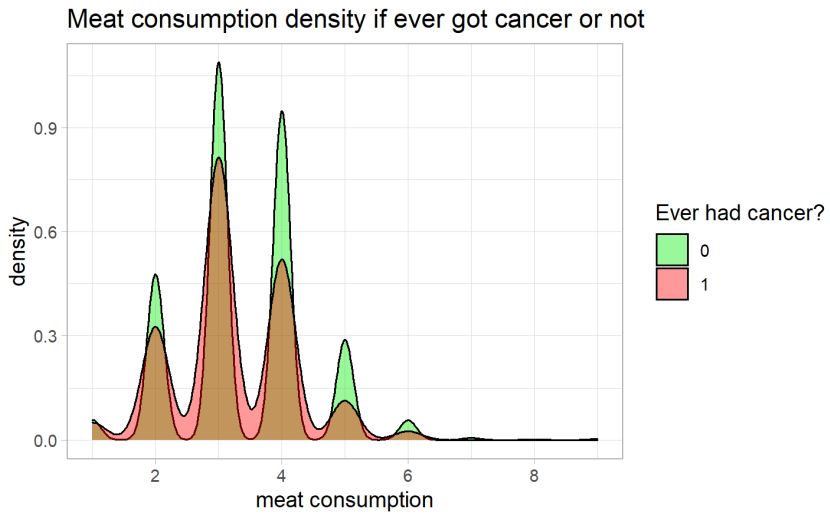
4.3.1 Cancer and Meat

We create a table including our `got_cancer` and our `meat_cons` variables.

Cancer and meat consumption variables

SEQN	meat_cons	got_cancer
31131	Moderate	0
31132	Low	0
31134	Moderate	0
31144	Moderate	0
31150	Low	1
31151	Low	0

Here is an illustration of the meat consumption density depending whether the respondents ever had cancer or not.

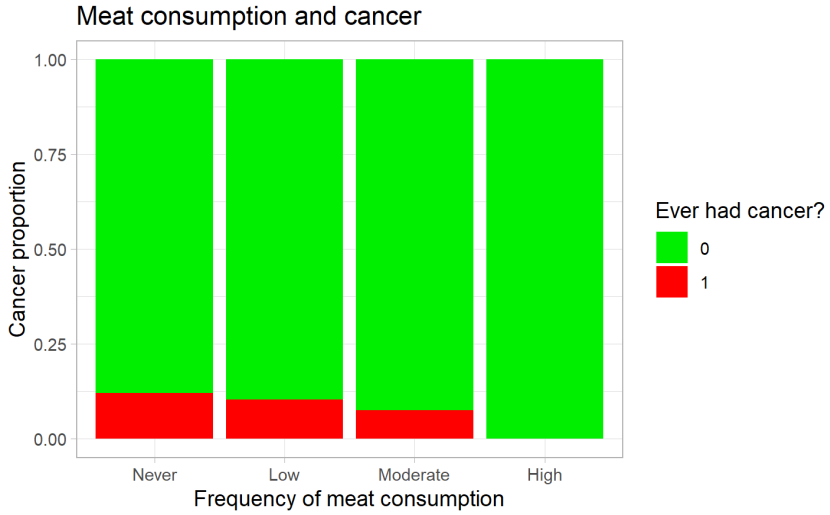


We can see that the greater the frequency is, the greater the gap in density is.

We display a table and a barplot to count the number of cancer and the proportion by frequency of meat consumption.

Cancer by meat consumption categories

meat_cons	got_cancer	count	proportion
Never	0	59	0.881
Never	1	8	0.119
Low	0	1592	0.897
Low	1	182	0.103
Moderate	0	1315	0.926
Moderate	1	105	0.074
High	0	12	1.000



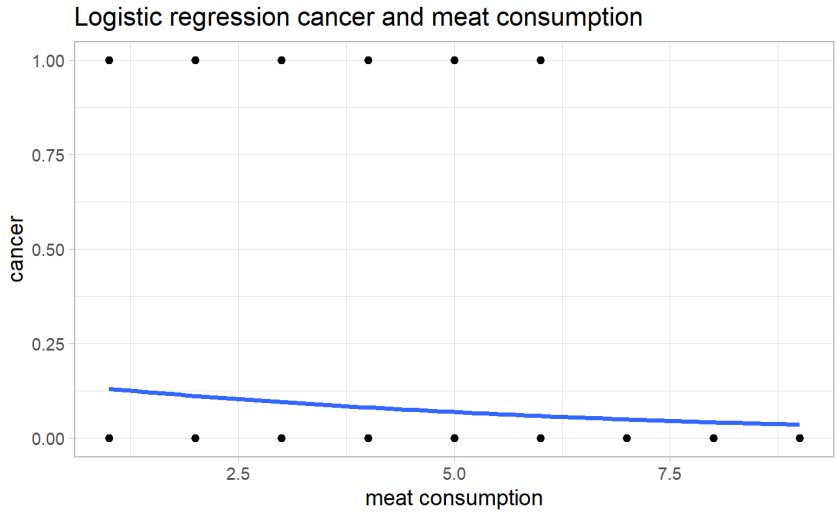
It can be seen that, surprisingly, respondents who never eat meat, which can therefore be considered vegetarians are the category which got more cancer in proportion, we can observe a trend but we must take into account that the high category have no cancer also because this category is small.

We create a GLM regression with meat explaining cancer and we compute the pseudo Rsquared.

```
#>
#> Call:
#> glm(formula = got_cancer ~ meat_cons, family = binomial(link = "logit"),
#>     data = join_food_other_cancer)
#>
#> Deviance Residuals:
#>    Min      1Q  Median      3Q     Max
#> -0.527  -0.446  -0.446  -0.410   2.386
#>
#> Coefficients:
#>             Estimate Std. Error z value            Pr(>|z|)
#> (Intercept)  -1.7266     0.2069   -8.34 <0.0000000000000002 ***
#> meat_cons    -0.1766     0.0612   -2.89              0.0039 **
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 1981.3  on 3266  degrees of freedom
#> Residual deviance: 1972.8  on 3265  degrees of freedom
#>   (3 observations deleted due to missingness)
#> AIC: 1977
#>
#> Number of Fisher Scoring iterations: 5
#> fitting null model for pseudo-r2
#>        llh    llhNull         G2   McFadden       r2ML       r2CU
#> -986.38955 -990.63354    8.48797    0.00428    0.00259    0.00571
```

It is seen that the meat consumption coefficient has a negative effect on cancer and a significant p-value at a level of 5%. However, McFadden's pseudo R^2 is 0.43%, which means that meat explains 0.43% of the variation in cancer. This variable will be included in the partial model because it is significant.

Here is an illustration of the GLM regression:.



4.3.2 Cancer and dairy products

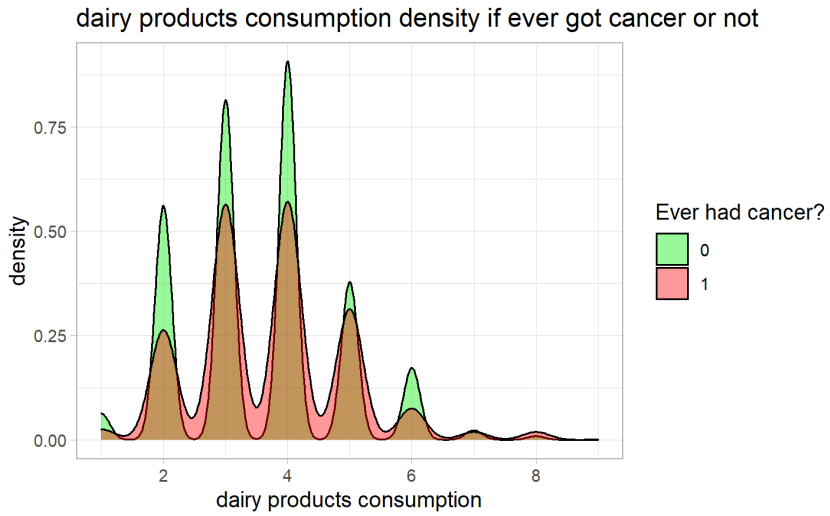
We create a table including our got_cancer and our dairy products variables.

Cancer and dairy products consumption variables

	SEQN	dairy_cons	got_cancer
	31131	moderate	0
	31132	low	0
	31134	moderate	0
	31144	high	0
	31150	moderate	1

SEQN	dairy_cons	got_cancer
31151	low	0

Here is an illustration of the dairy products consumption density depending whether the respondens ever had cancer or not.

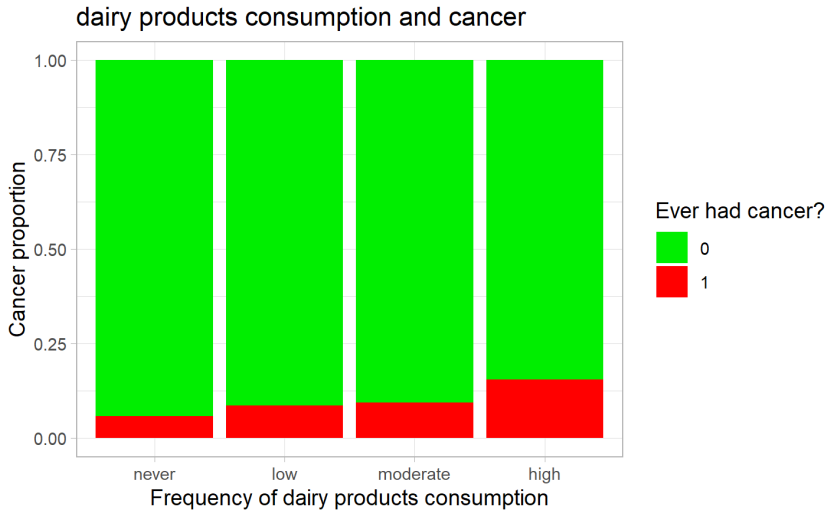


We can see that the greater the frequency is, the smaller the gap in density is between the two groups.

We display a table and a barplot to count the number of cancers and the proportion by frequency of consumption of dairy products.

Cancer by frequency of consumption of dairy products

dairy_cons	got_cancer	count	proportion
never	0	66	0.943
never	1	4	0.057
low	0	1400	0.914
low	1	132	0.086
moderate	0	1481	0.906
moderate	1	153	0.094
high	0	33	0.846
high	1	6	0.154



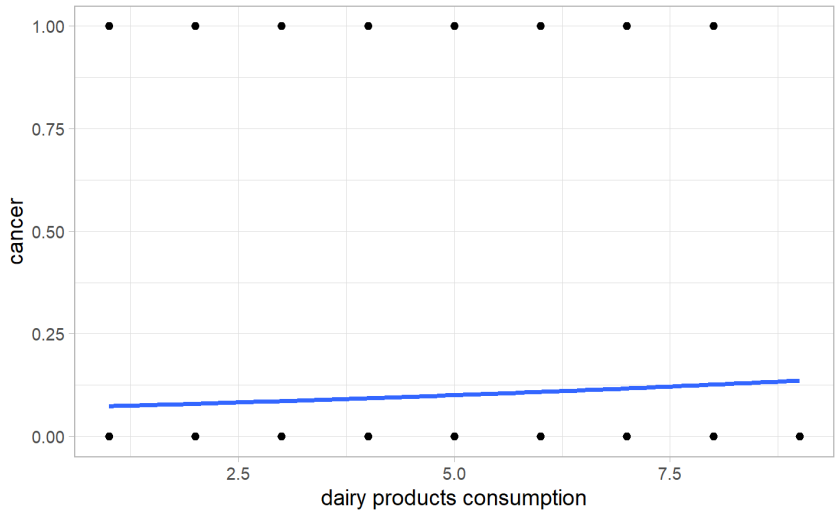
It can be seen here that the more the frequency of consumption of dairy products increases, the more respondents actually had cancer.

We create a GLM regression with dairy products consumption explaining cancer and we compute the pseudo Rsquared

```
#>
#> Call:
#> glm(formula = got_cancer ~ dairy_cons, family = binomial(link = "logit"),
#>     data = join_food_other_cancer)
#>
#> Deviance Residuals:
#>    Min      1Q  Median      3Q     Max
#> -0.540  -0.442  -0.424  -0.407   2.285
#>
#> Coefficients:
#>             Estimate Std. Error z value            Pr(>|z|)
#> (Intercept)  -2.6195     0.1887  -13.88 <0.0000000000000002 ***
#> dairy_cons    0.0854     0.0485    1.76               0.079 .
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 1981.3  on 3266  degrees of freedom
#> Residual deviance: 1978.2  on 3265  degrees of freedom
#>   (3 observations deleted due to missingness)
#> AIC: 1982
#>
#> Number of Fisher Scoring iterations: 5
#> fitting null model for pseudo-r2
#>         llh     llhNull          G2    McFadden        r2ML
#> -989.101825 -990.633537    3.063426    0.001546    0.000937
#>        r2CU
#>    0.002061
```

We can see that the dairy products consumption coefficient has a positive effect on cancer and a significant p-value at 10%, so it will not be included in the partial model. McFadden's pseudo R^2 is 0.16%, which means that dairy products explains 0.16% of the variation in cancer.

Here is an illustration of the GLM regression:



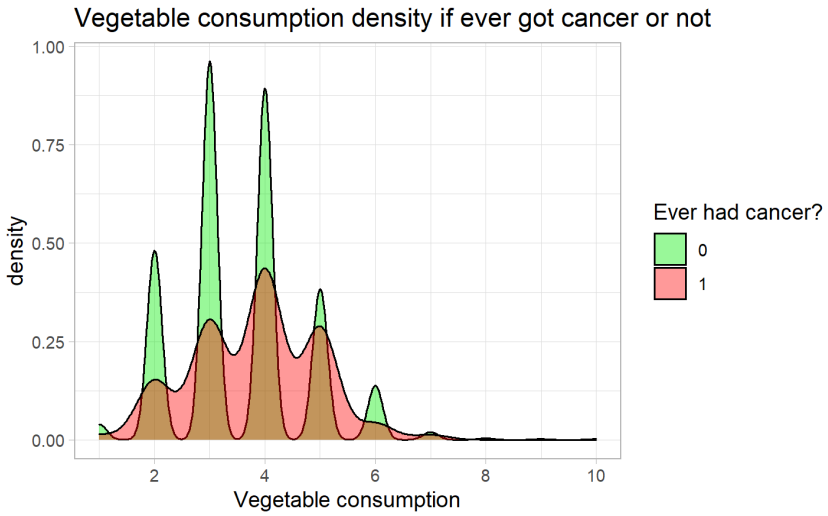
4.3.3 Cancer and Vegetables

We create a table including our got_cancer and our vegetables variables.

Cancer and vegetables consumption variables

SEQN	vege_cons	got_cancer
31131	moderate	0
31132	low	0
31134	low	0
31144	moderate	0
31150	moderate	1
31151	low	0

Here is an illustration of the vegetable consumption density depending whether respondents ever had cancer or not.

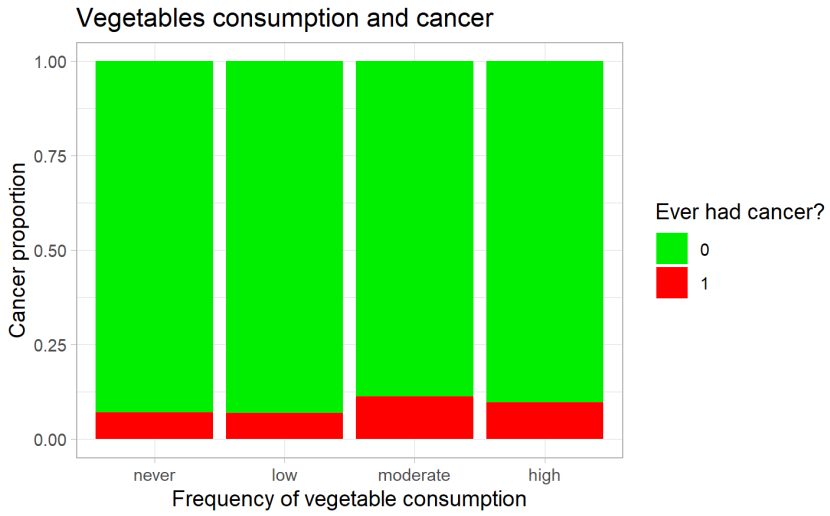


We display a table and a barplot to count the number of cancers and the proportion by frequency of consumption of vegetables.

Cancer by frequency of consumption of vegetables

vege_cons	got_cancer	count	proportion

vege_cons	got_cancer	count	proportion
never	0	40	0.930
never	1	3	0.070
low	0	1470	0.932
low	1	108	0.068
moderate	0	1438	0.888
moderate	1	181	0.112
high	0	28	0.903
high	1	3	0.097



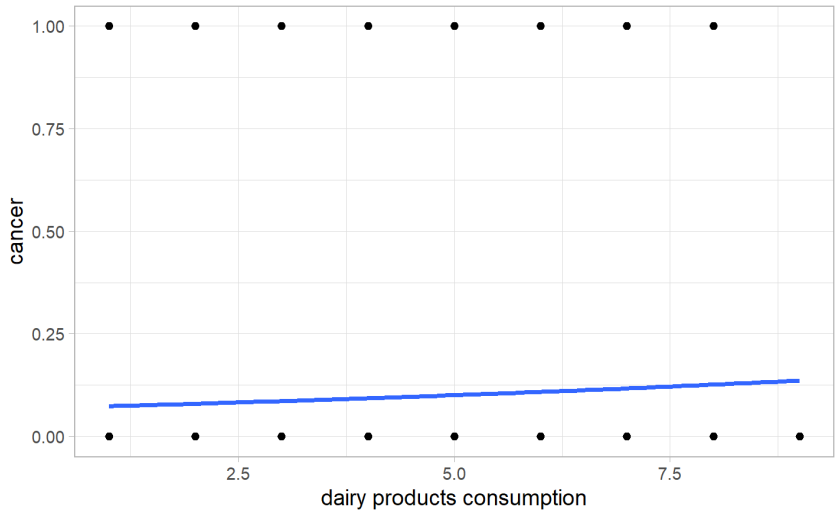
We do not necessarily see a difference or relationship, depending on the frequency of vegetable consumption and the fact of having had cancer.

We create a GLM regression with vegetable consumption explaining cancer and we compute the pseudo Rsquared

```
#>
#> Call:
#> glm(formula = got_cancer ~ vege_cons, family = binomial(link = "logit"),
#>     data = join_food_other_cancer)
#>
#> Deviance Residuals:
#>    Min      1Q  Median      3Q     Max
#> -0.731  -0.447  -0.410  -0.377   2.386
#>
#> Coefficients:
#>             Estimate Std. Error z value             Pr(>|z|)
#> (Intercept)  -2.9660     0.2009  -14.76 < 0.0000000000000002 ***
#> vege_cons     0.1781     0.0506    3.52              0.00043 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 1981.3  on 3266  degrees of freedom
#> Residual deviance: 1969.1  on 3265  degrees of freedom
#>   (3 observations deleted due to missingness)
#> AIC: 1973
#>
#> Number of Fisher Scoring iterations: 5
#> fitting null model for pseudo-r2
#>        llh    llhNull         G2   McFadden       r2ML       r2CU
#> -984.56768 -990.63354   12.13172    0.00612    0.00371    0.00815
```

It is seen that the coefficient of vegetable consumption has a positive effect on cancer and a significant p-value at a level of 0.1%, this variable will be included in the partial model. McFadden's pseudo R ^ 2 is 0.62%, which means that vegetable consumption explains 0.62% of the variation in cancer.

Here is an illustration of the GLM regression:



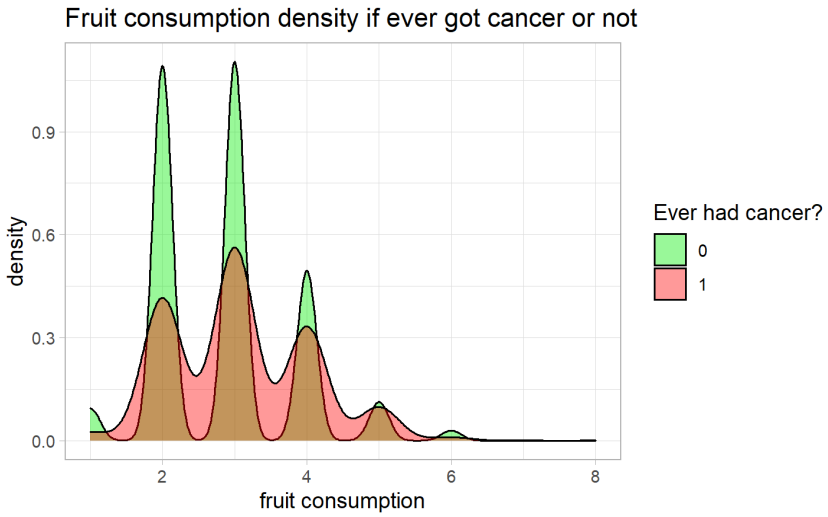
4.3.4 Cancer and Fruits

We create a table including our got_cancer and our fruits variables.

Cancer and fruits consumption variables

SEQN	fruit_cons	got_cancer
31131	moderate	0
31132	low	0
31134	low	0
31144	moderate	0
31150	moderate	1
31151	low	0

Here is an illustration of the fruit consumption density depending whether respondents ever had cancer or not.

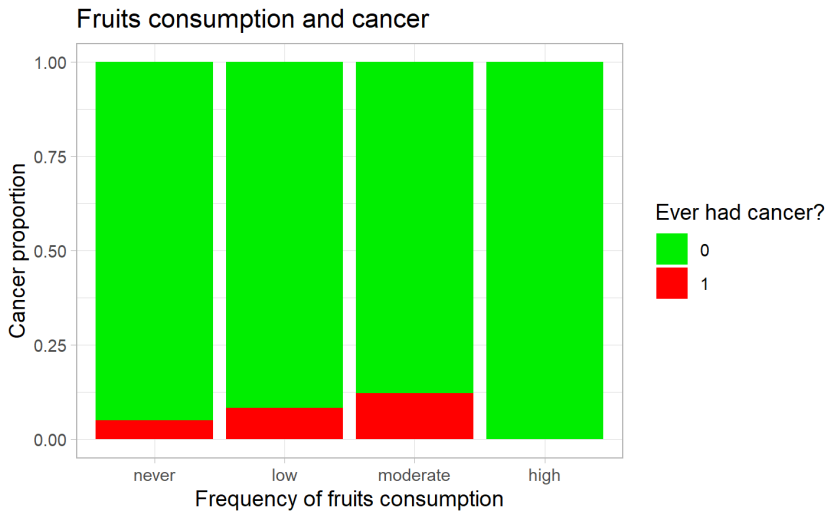


We can see that the greater the frequency is, the smaller the gap in density is between the two groups.

We display a table and a barplot to count the number of cancers and the proportion by frequency of consumption of fruits.

Cancer by frequency of consumption of fruits

fruit_cons	got_cancer	count	proportion
never	0	96	0.950
never	1	5	0.050
low	0	2235	0.918
low	1	200	0.082
moderate	0	647	0.878
moderate	1	90	0.122
high	0	3	1.000



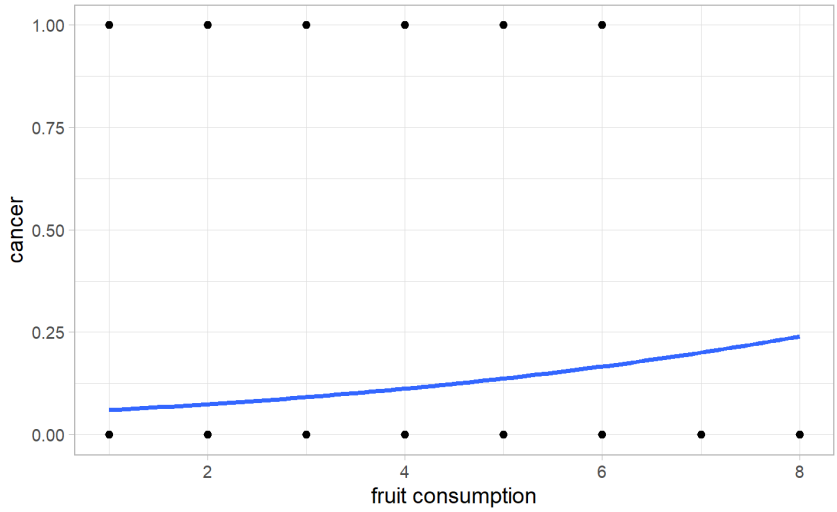
We can clearly see that the higher the frequency of fruit consumption among our respondents, the more they have cancer. Respondents with high consumption have no cancer probably because this category is very small.

We create a GLM regression with fruit consumption explaining cancer and we compute the pseudo Rsquared

```
#>
#> Call:
#> glm(formula = got_cancer ~ fruit_cons, family = binomial(link = "logit"),
#>     data = join_food_other_cancer)
#>
#> Deviance Residuals:
#>    Min      1Q  Median      3Q     Max
#> -0.739  -0.437  -0.437  -0.392   2.374
#>
#> Coefficients:
#>             Estimate Std. Error z value             Pr(>|z|)
#> (Intercept)  -2.9835     0.1938  -15.39 < 0.0000000000000002 ***
#> fruit_cons    0.2282     0.0605    3.77              0.00016 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 1981.3  on 3266  degrees of freedom
#> Residual deviance: 1967.5  on 3265  degrees of freedom
#>   (3 observations deleted due to missingness)
#> AIC: 1971
#>
#> Number of Fisher Scoring iterations: 5
#> fitting null model for pseudo-r2
#>        llh    llhNull         G2   McFadden       r2ML       r2CU
#> -983.74643 -990.63354   13.77422    0.00695    0.00421    0.00925
```

It can be seen that the fruit consumption coefficient has a positive effect on cancer and a significant p-value at 0.1%, this variable will be included in the partial model. McFadden's pseudo R^2 is 0.7%, which means that fruits explains 0.7% of the variation in cancer.

Here is an illustration of the GLM regression:



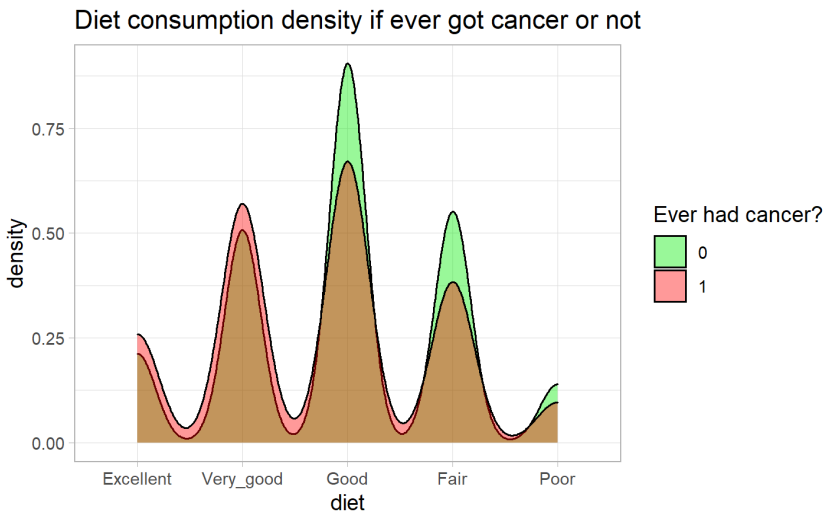
4.3.5 Cancer and diet type.

We create a table including our `got_cancer` and the type of `diet` variables.

Cancer and the type of diet followed variables

SEQN	diet	got_cancer
31130	Good	0
31131	Good	0
31132	Very_good	0
31134	Good	0
31136	Good	0
31144	Excellent	0

Here is an illustration of the diet density depending whether respondents ever had cancer or not.

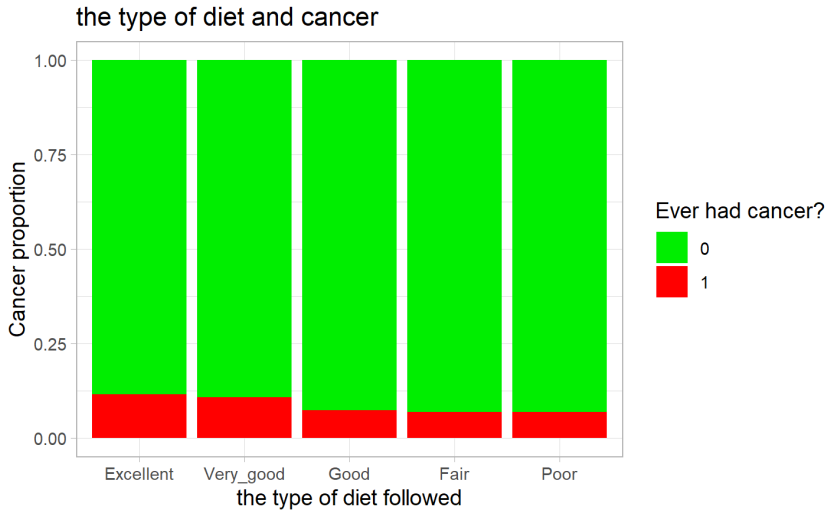


We can see that for excellent and very good diet the density is higher for the group who got cancer, then the density is relatively higher for the group who never got cancer.

We display a table and a barplot to count the number of cancers and the proportion by the type of diet followed.

Cancer by the type of diet followed

diet	got_cancer	count	proportion
Excellent	0	417	0.885
Excellent	1	54	0.115
Very_good	0	997	0.893
Very_good	1	119	0.107
Good	0	1779	0.927
Good	1	140	0.073
Fair	0	1084	0.931
Fair	1	80	0.069
Poor	0	274	0.932
Poor	1	20	0.068



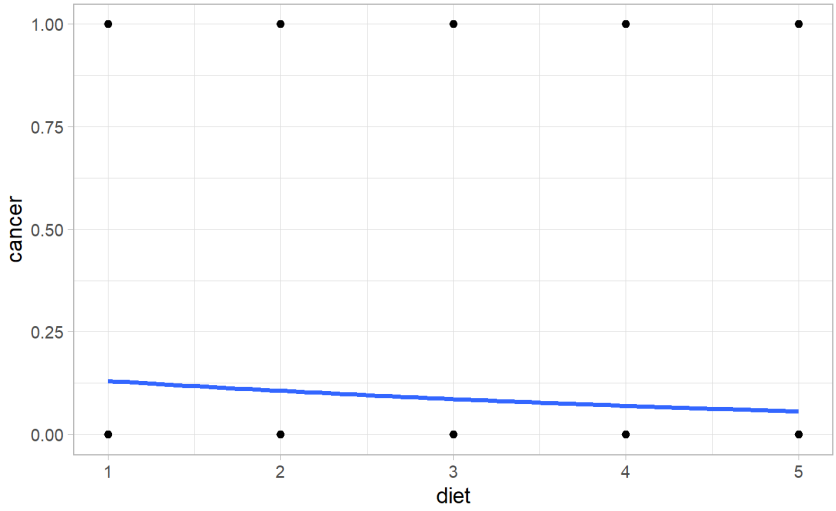
We can clearly see that the higher the frequency of fruit consumption among our respondents, the more they have cancer. This result which seems quite predictable to us because the fruits are basically sweet and we know that sugar (glucose) provides the necessary nourishment for every cell in our body and also for cancer cells.

We create a GLM regression with diet explaining cancer and we compute the pseudo Rsquared.

```
#>
#> Call:
#> glm(formula = got_cancer ~ diet, family = binomial(link = "logit"),
#>     data = join_food_other_cancer)
#>
#> Deviance Residuals:
#>    Min      1Q  Median      3Q     Max
#> -0.528  -0.473  -0.424  -0.380   2.401
#>
#> Coefficients:
#>             Estimate Std. Error z value             Pr(>|z|)
#> (Intercept)  -1.6692     0.1717   -9.72 < 0.0000000000000002 ***
#> diet         -0.2312     0.0602   -3.84              0.00012 ***
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 1980.3  on 3261  degrees of freedom
#> Residual deviance: 1965.4  on 3260  degrees of freedom
#>   (8 observations deleted due to missingness)
#> AIC: 1969
#>
#> Number of Fisher Scoring iterations: 5
#> fitting null model for pseudo-r2
#>        llh    llhNull         G2   McFadden       r2ML       r2CU
#> -982.70808 -990.15997   14.90377    0.00753    0.00456    0.01002
```

We can see that the weight of the diet has a negative effect on cancer. Here, we have to be careful because a healthy diet = 1 and a bad diet = 5. This therefore means that a healthier diet has a positive relationship with cancer and as the p-value is significant at 0.1%, this variable will be included in the partial model. McFadden's pseudo R ^ 2 is 0.75%, which means that the diet type explains 0.75% of the variation in cancer.

Here is an illustration of the GLM regression:



4.4 Multivariate cancer modelling

Now we will start by calculating the partial model with all the individually significant variables, then we will create a model with all the variables (significant and non-significant) and finally we will compare these two models.

4.4.1 Multivariate partial model

We create a partial model with all the significant variables at the 5% level, which we saw previously:

- Age
- Alcohol
- Smoking
- Meat consumption
- Vegetable consumption
- Fruit consumption
- Type of diet

Here is the GLM regression explaining cancer and with pseudo Rsquared

```
#>
#> Call:
#> glm(formula = got_cancer ~ age + avg_alcohol + cigarets_per_day +
#>     meat_cons + vege_cons + fruit_cons + diet, family = binomial(link = "logit"),
#>     data = join_food_other_cancer)
#>
#> Deviance Residuals:
#>    Min      1Q  Median      3Q     Max
#> -0.767  -0.349  -0.286  -0.221   2.634
#>
#> Coefficients:
#>                  Estimate Std. Error z value Pr(>|z|)
#> (Intercept)      -4.59715    1.68482   -2.73   0.0064 **
#> age               0.02363    0.01609    1.47   0.1420
#> avg_alcohol      -0.02913    0.09846   -0.30   0.7674
#> cigarets_per_day  0.00472    0.02289    0.21   0.8366
#> meat_cons        -0.36447    0.25843   -1.41   0.1585
#> vege_cons         0.29062    0.25981    1.12   0.2633
#> fruit_cons        0.13678    0.32000    0.43   0.6691
#> diet              0.14981    0.24160    0.62   0.5352
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 158.30  on 394  degrees of freedom
#> Residual deviance: 150.18  on 387  degrees of freedom
#>   (2875 observations deleted due to missingness)
#> AIC: 166.2
#>
#> Number of Fisher Scoring iterations: 6
#> fitting null model for pseudo-r2
#>      llh  llhNull       G2 McFadden     r2ML     r2CU
#> -75.0890 -79.1480   8.1178   0.0513   0.0203   0.0616
```

We notice that our partial model including only our individually significant variables, gives us as results that none of our variables is now significant, only the intercept is significant. We can also notice that the pseudo McFadden rsquared has a value of 0.0513 and therefore that only 5.13% of the variation in cancer is explained by this model. Therefore, this model is extremely poor at predicting cancer as a good model should ideally have a McFadden rsquared value greater than 70% or even 80%.

4.4.2 Multivariate full model

We build a complete model including all the variables we explored previously, including those that were not significant.

```
#>
#> Call:
#> glm(formula = got_cancer ~ age + gender + income + avg_physical_activity +
#>     avg_alcohol + cigarets_per_day + meat_cons + dairy_cons +
#>     vege_cons + fruit_cons + diet, family = binomial(link = "logit"),
#>     data = join_food_other_cancer)
#>
#> Deviance Residuals:
#>    Min      1Q  Median      3Q     Max
#> -0.797  -0.337  -0.224  -0.161   2.915
#>
#> Coefficients:
#>                       Estimate Std. Error z value Pr(>|z|)
#> (Intercept)           -8.16906    2.56406   -3.19   0.0014 **
#> age                    0.02535    0.01880    1.35   0.1775
#> gender                 1.40936    0.63000    2.24   0.0253 *
#> income                -0.02046    0.09039   -0.23   0.8209
#> avg_physical_activity  0.12066    0.29711    0.41   0.6847
#> avg_alcohol            0.04693    0.10699    0.44   0.6609
#> cigarets_per_day       0.00751    0.02574    0.29   0.7705
#> meat_cons             -0.42881    0.29311   -1.46   0.1435
#> dairy_cons             0.41119    0.24489    1.68   0.0931 .
#> vege_cons              0.19348    0.30446    0.64   0.5251
#> fruit_cons            -0.09215    0.37662   -0.24   0.8067
#> diet                   0.27474    0.26588    1.03   0.3015
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> (Dispersion parameter for binomial family taken to be 1)
#>
#>     Null deviance: 145.50  on 385  degrees of freedom
#> Residual deviance: 129.01  on 374  degrees of freedom
#>   (2884 observations deleted due to missingness)
#> AIC: 153
#>
#> Number of Fisher Scoring iterations: 6
#> fitting null model for pseudo-r2
#>      llh  llhNull       G2 McFadden     r2ML     r2CU
#> -64.5073 -72.7520  16.4894   0.1133   0.0418   0.1332
```

For the full model, we can see that gender is significant at a 5% level. This could be due to the fact that as we observed in part 3.3.2, breast and cervix are cancers that only women can get, are among the most common cancer types in our sample. The other significant variable at the 10% level is the consumption of dairy products. McFadden's pseudo rsquared is equal to 0.1133 which means that only 11.33% of the variation in cancer is explained by this model, which is unfortunately very poor.

4.4.3 Model comparison and interpretration

The full model is therefore better than the partial model with a McFadden pseudo-square of 11.33% compared to 5.13%. Even though it is preferable, 11.33% is still a very poor value for predicting cancer. There must be either an important variable explaining a part of the cancer than the one we have studied or else the cancer has a large number of variables with each of them having a small effect, which is the most probable as we included some variables known to cause some cancers.

In the end, there are two significant variables which we can be identified: - Gender at the 5% level - Consumption of dairy products at the 10% level (which could be considered too high)

However, one thing that is interesting to observe is the direction of the coefficient of each variable, with a model including all variables:

- `age` 0.025 is positive, as we would expect the greater the age is the greater the odds that you got cancer are.
- `gender` 1.409 is positive meaning the odds to get a cancer are greater for women.
- `income` -0.02 is negative meaning the odds to get a cancer are smaller the more income you get.
- `avg_physical_activity` 0.121 is positive, which is the opposite of what we would expect meaning more physical activity the greater the odds to have a cancer are.
- `avg_alcohol` 0.047 is positive which we would expect before starting this project, but which contradict what we observe in point 4.2.5, where it appeared surprisingly more as a negative relationship with cancer.
- `cigarettes_per_day` 0.008 is positive as we could expect.

- `meat_cons` -0.429 is negative meaning the odds to get cancer are smaller the more you eat meat, which contradict our hypothesis.
- `dairy_cons` 0.411 is positive meaning the odds to get cancer are greater the more you eat dairy products, which goes in the same direction than our hypothesis.
- `vege_cons` 0.193 is positive meaning the odds to get cancer are greater the more vegetables you eat, which contradict our hypothesis.
- `fruit_cons` -0.092 is negative, which goes in the same direction than our hypothesis, but contradict what we observed in point 4.3.4. Moreover we have to keep in mind that this variable is correlated (50%) with vegetable.
- `diet` 0.275 is positive in that case (for this variable value are ranked for 1 for healthy to 5 for poor) it means the less healthy you eat the greater the odds to get cancer are, which goes in the same direction than our hypothesis, but contradict what we observed in point 4.3.5.

4.5 Answering research questions

Finally, our variables are not good enough to predict cancer, because the best model we could make could only explain 11.33% of the variation in cancer, so the influence of our variables individually is extremely small and questionable as we have seen.
Our hypotheses were that eating healthy, meaning eating vegetables and fruits and therefore having a good diet would reduce the risk of cancer and that, on the contrary, animal products (meat and dairy products) would increase the risk of cancer.
However, these assumptions are not supported by our model, when it includes confounding variables, except for the dairy product which has an effect but weak and is the only food variable which is significant, and only at a level of 10%.

5 Conclusion

At the first sight when we investigated the relationship between food and cancer it looked like there were relationships, but it was no longer the case once we introduced potential confounding variables. In the end, only gender and potentially dairy product consumption have an effect. So, it was a good application of this principle for us, as at the very beginning of our project we were not thinking about that which would have lead us to very biased and unrealistic results.

Although we included many variables that are known to have a relationship with certain cancer, such as smoking, drinking, it is very far from enough to predict cancer, as our model showed us. All the variables in our study have only a very small or no effect, it is therefore very difficult to determine a particular cause. There could also be for sure much more variables that were not in the scope of this study (such as genetic or stress level, and more) that could influence cancer or influence the significance of other variables. Moreover, there was only cancer data about adults which might lower the age's overall effect on our results. Therefore remaining cautious, we cannot come out of this study with precisely identifying a clear relationship or effect between eating habits and cancer. But aside of that our data suggest that gender has an effect on cancer, meaning women got significantly more cancers than men in our sample of respondents.

For the future, investigating every particular cancer types compared to particular eating habits could be a potential interesting work. But to do that we would need other data sets with a wider pool of observations for each type of cancer, as our dataset had a very limited number of observations by type of cancer.