S2.04 - The Link Between Nutriscore and Nutritional Values

Leo Ducruet and Lynn Hayot, Group B1

What is our project

The aim of this project is to produce for our customer a visual and statistical support with relevant data which enables to establish a link between the nutriscore grade and the nutritional values of a french product in the "en:one-dish-meal" category.

The nutriscore is a nutrition label and nutritional rating system, it assigns products a rating letter from A (best) to E (worst) and an a color associated to the letter, depending on the nutritional values provided by the product.

To respond to the request, we will use the database hayotj.csv which groups data on french products in the "en:one-dish-meals" category without duplication. This database contains 22,216 products, it's an extract from the http://world-en.openfoodfacts.org/ database which contains about 3 millions products.

Useful variables

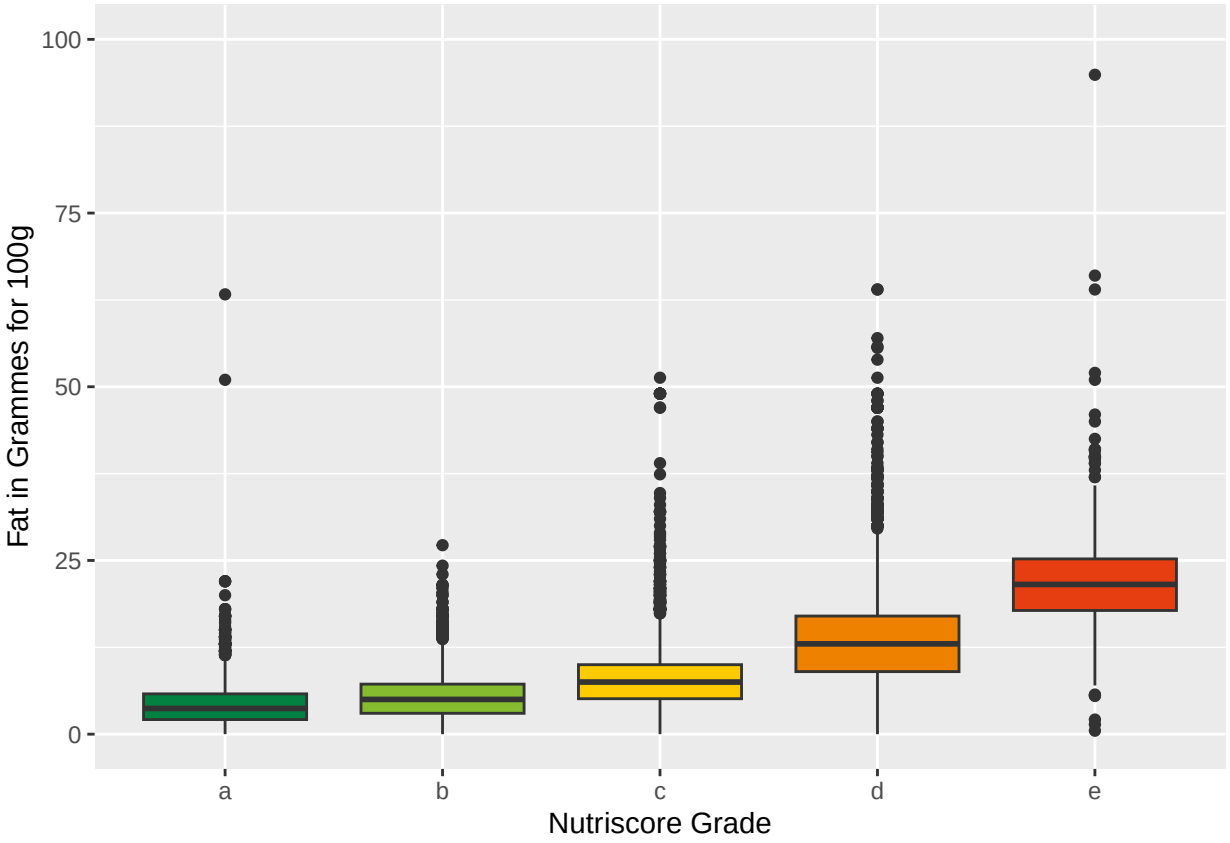
We have a few variables that will be helpful to respond to the subject :

The code and the url of a product are nominal qualitative variables that distinguish the product from the others. The product_name is the name of the product on the market. The nutriscore_score and nutriscore_grade are ordinal qualitative variables that rank product quality, the nutriscore_score is expressed as a number, instead of the nutriscore_grade expressed as a letter (a, b, c, d, e). The quantitative variables like salt_100g, carbohydrates_100g, sugars_100g, fat_100g or proteins_100g give us the nutritional values per 100g of the product. The energy_100g expresses the amount of energy given by the product in kj, the computed_energy_100g is about the same value but calculated manually. Finally, the ordinal qualitative variables level_fat, level_salt, level_saturated_fat and level_sugars which indicate the level of each category.

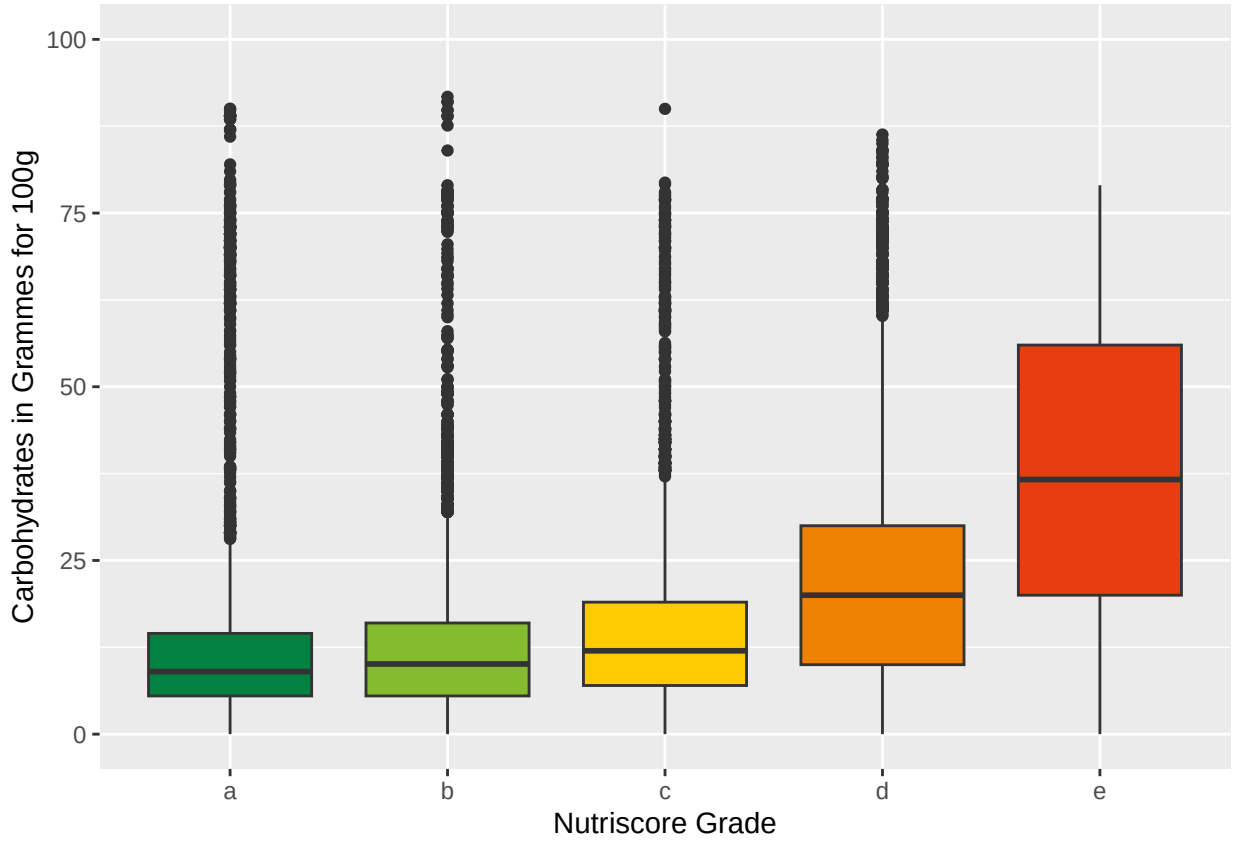
Values removed

Here are some graphs displaying every quantitative variables we have before filtering them.

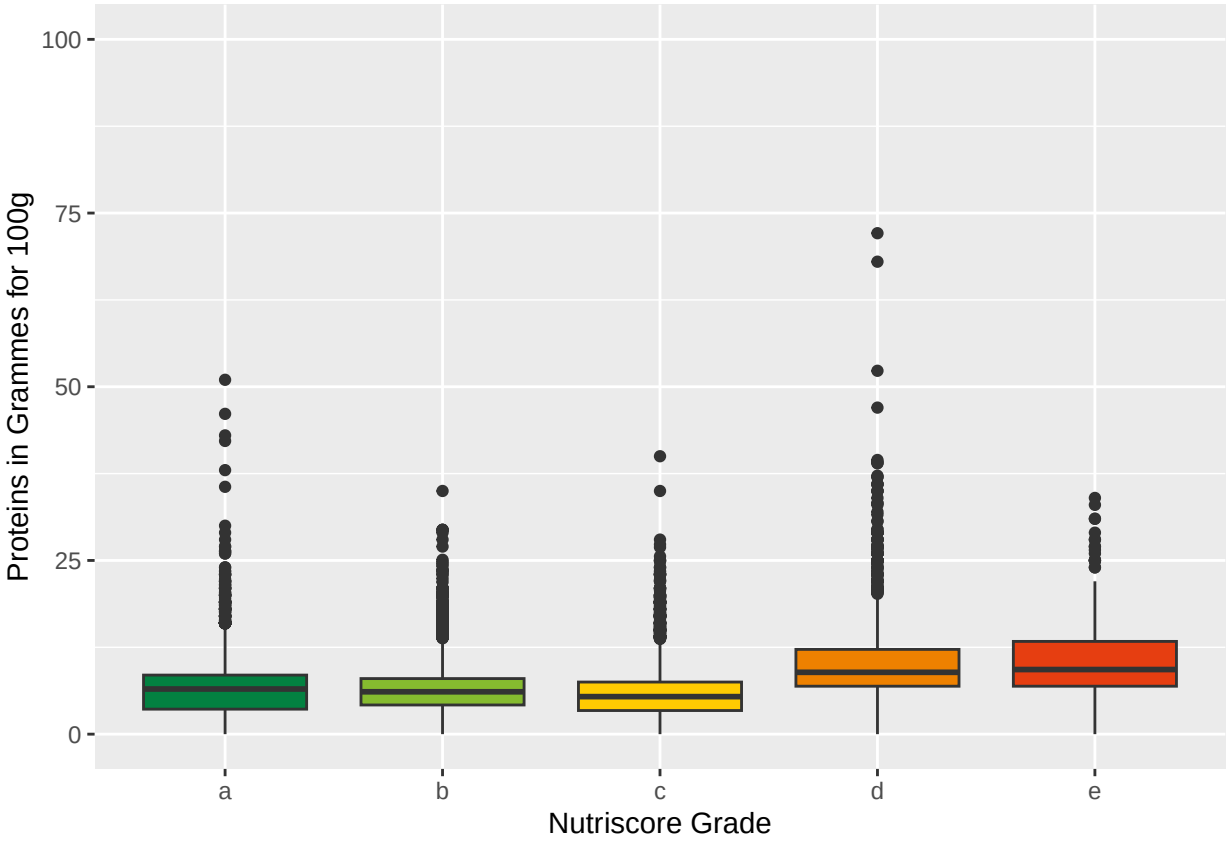
```
nutriscore_colors <- c('a' = '#038141', 'b' = '#85bb2f', 'c' = '#fecb02', 'd' = '#ee8100', 'e' = '#e63e

ggplot(hayotj)+
  geom_boxplot(aes(x=nutriscore_grade, y=fat_100g), fill = nutriscore_colors)+
  labs(x = "Nutriscore Grade", y = "Fat in Grammes for 100g") +
  ylim(0, 100)
```



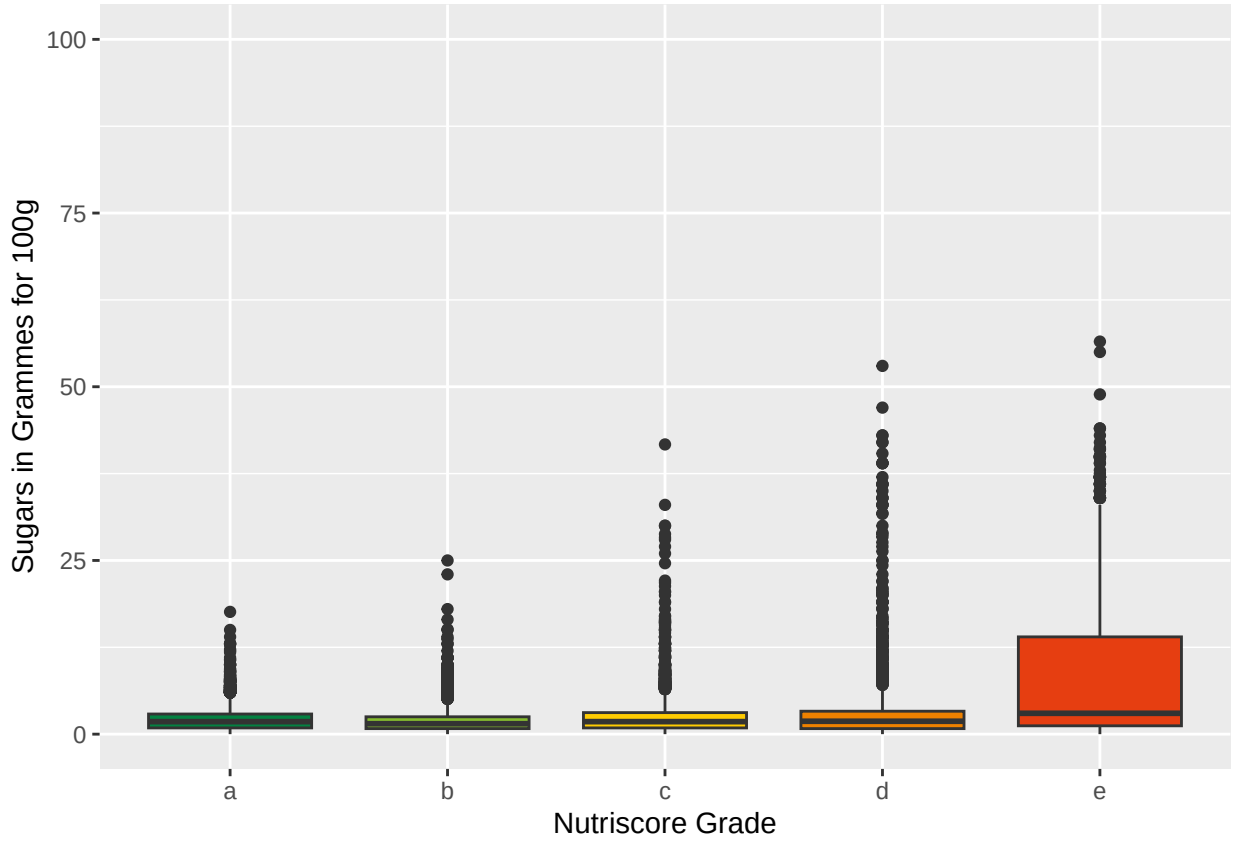

```
ggplot(hayotj)+
  geom_boxplot(aes(x=nutriscore_grade, y=carbohydrates_100g), fill = nutriscore_colors)+
  labs(x = "Nutriscore Grade", y = "Carbohydrates in Grammes for 100g") +
  ylim(0, 100)
```



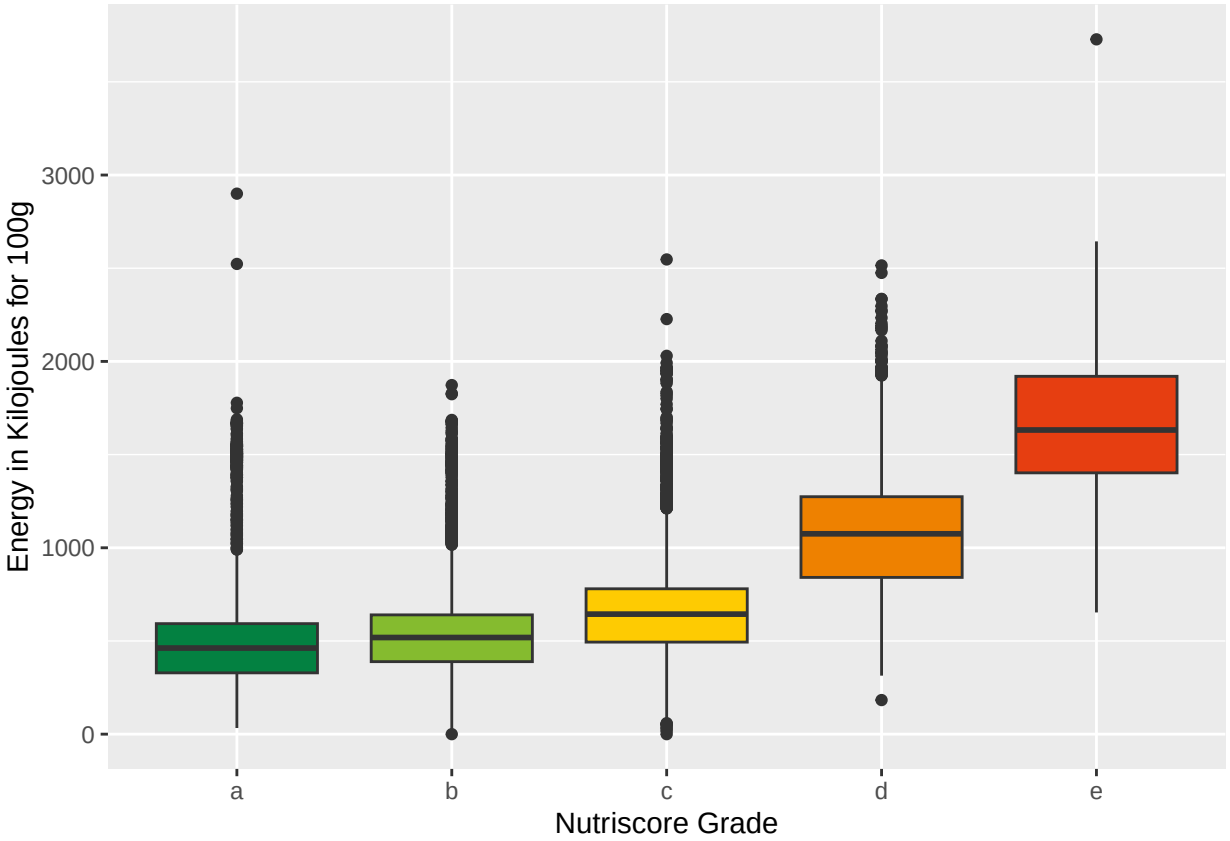

```
ggplot(hayotj) +
  geom_boxplot(aes(x=nutriscore_grade, y=proteins_100g), fill = nutriscore_colors)+
  labs(x = "Nutriscore Grade", y = "Proteins in Grammes for 100g") +
  ylim(0, 100)
```




```
ggplot(hayotj) +
  geom_boxplot(aes(x = nutriscore_grade, y = sugars_100g), fill = nutriscore_colors) +
  labs(x = "Nutriscore Grade", y = "Sugars in Grammes for 100g") +
  ylim(0, 100)
```




```
ggplot(hayotj)+
  geom_boxplot(aes(x=nutriscore_grade, y=energy_100g), fill = nutriscore_colors)+
  labs(x = "Nutriscore Grade", y = "Energy in Kilojoules for 100g")
```



We decided to exclude some abnormally high values because they were too far away from the other values and they were not significant.

First for the proteins_100g category, we excluded 2 products which contained more than 60% of proteins because it was products specifically made to contain a lot of proteins, which is not really useful to study.

For fat_100g category, we excluded a specific product (Lard Cream) which contained nearly 95% fat, to compare, the second product with the most rate of fat contains 66%, so we though that this product wasn't interesting for our study.

For energy_100g category, we removed the products that had more than 2400kJ because there is only 9 products superior to this amount and they are high above the average so it is not very meaningful and also those with 0kJ because it is not possible.

And for the categories sugars_100g and carbohydrates_100g, we kept all the values because there wasn't any abnormal values.

Visualisation of the importants variables

```
hayotj %>%
  filter(proteins_100g < 60 & fat_100g < 90 & energy_100g < 2400 & energy_100g > 0) %>%
  select(product_name, nutriscore_grade, fat_100g, sugars_100g, energy_100g, carbohydrates_100g, level_
  slice_head(n = 10)
```

```
##                          product_name nutriscore_grade fat_100g sugars_100g
```

```
## 1                      0sso-buco de veau                b      2.8         3.6
## 2                        100G Minestrone                d      3.4        15.6
## 3   10 Accras au cabillaud et à la morue                c     11.0         1.4
## 4  10 Crêpes Champignons Jambon Emmental                c      6.5         3.5
## 5                     10 Crêpes Chocolat                d     22.0        39.0
## 6       10 crêpes demi-lune aux fromages                b      8.7         3.7
## 7                     10 crêpes emmental                b      4.3         2.6
## 8                     10 crêpes emmental                b      4.8         3.7
## 9      10 Crêpes gourmandes campagnardes                b      5.0         3.2
## 10  10 crêpes gourmandes jambon emmental                b      7.1         3.7
##    energy_100g carbohydrates_100g level_fat level_sugars
## 1          534                9.0         l            l
## 2         1397               57.7         m            h
## 3          977               21.0         m            l
## 4          674               20.0         m            l
## 5         1946               59.0         h            h
## 6          761               18.4         m            l
## 7          607               20.0         m            l
## 8          609               17.0         m            l
## 9          619               17.0         m            l
## 10         708               17.0         m            l
```

This table shows us the 10 first rows of the database with the useful variables for our study. We can notice that the worst nutriscore_grade here is d, attributed to the "100G Minestrone" and the "10 Crêpes Chocolat". Their common points are their high fat and sugar content and the large amount of energy that they provide. First, we are going to look at the fat content of the products and its impact on the nutriscore grade.

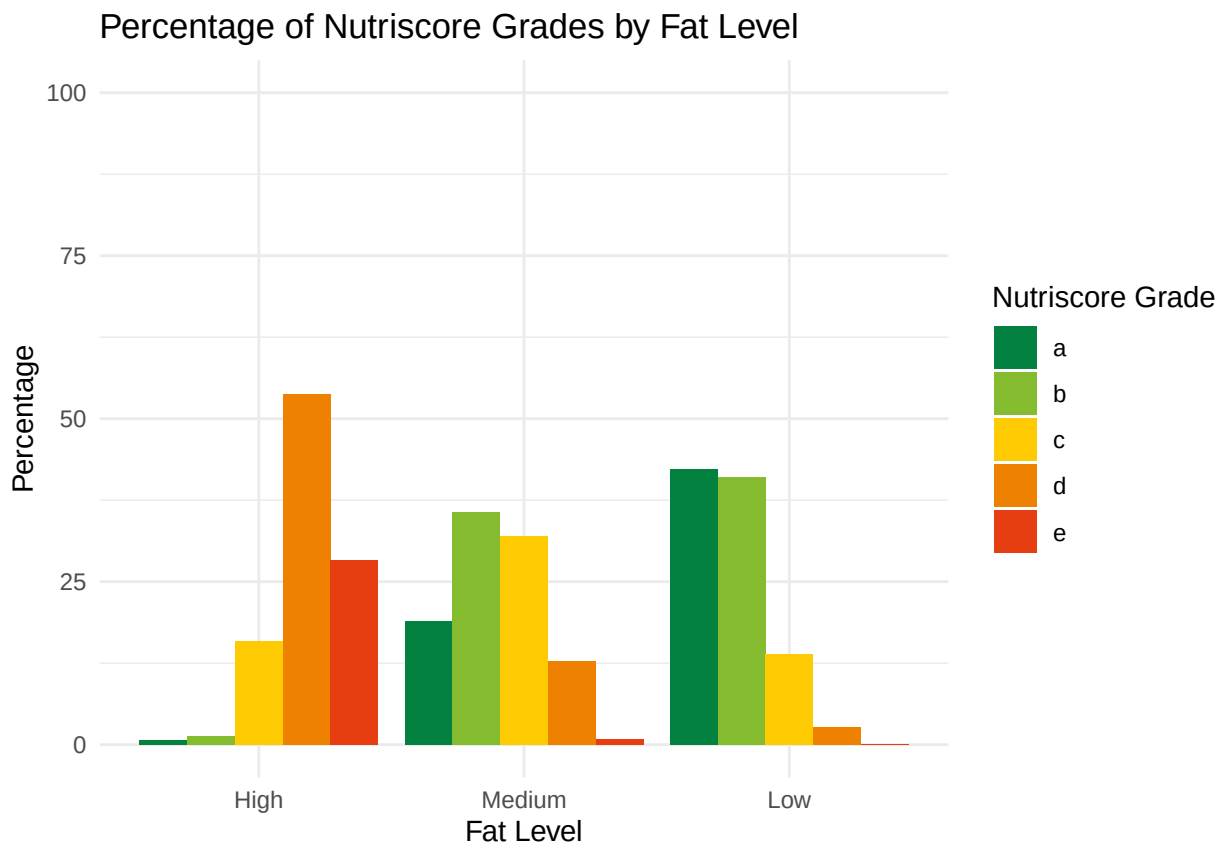
Impact of the Fat Level on the Nutriscore

```
hayotj$level_fat <- factor(hayotj$level_fat, levels = c('h', 'm', 'l'))

percentage_data <- hayotj %>%
  filter(proteins_100g < 60 & fat_100g < 90 & energy_100g < 2400 & energy_100g > 0) %>%
  group_by(level_fat, nutriscore_grade) %>%
  summarise(count = n()) %>%
  mutate(percentage = count / sum(count) * 100)
```

```
## `summarise()` has grouped output by 'level_fat'. You can override using the
## `.groups` argument.
```

```
nutriscore_colors <- c('a' = '#038141', 'b' = '#85bb2f', 'c' = '#fecb02', 'd' = '#ee8100', 'e' = '#e63e

ggplot(percentage_data, aes(x = level_fat, y = percentage, fill = nutriscore_grade)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Percentage of Nutriscore Grades by Fat Level",
       x = "Fat Level",
       y = "Percentage",
       fill = "Nutriscore Grade") +
  scale_fill_manual(values = nutriscore_colors) +  # You can choose any palette you prefer
```

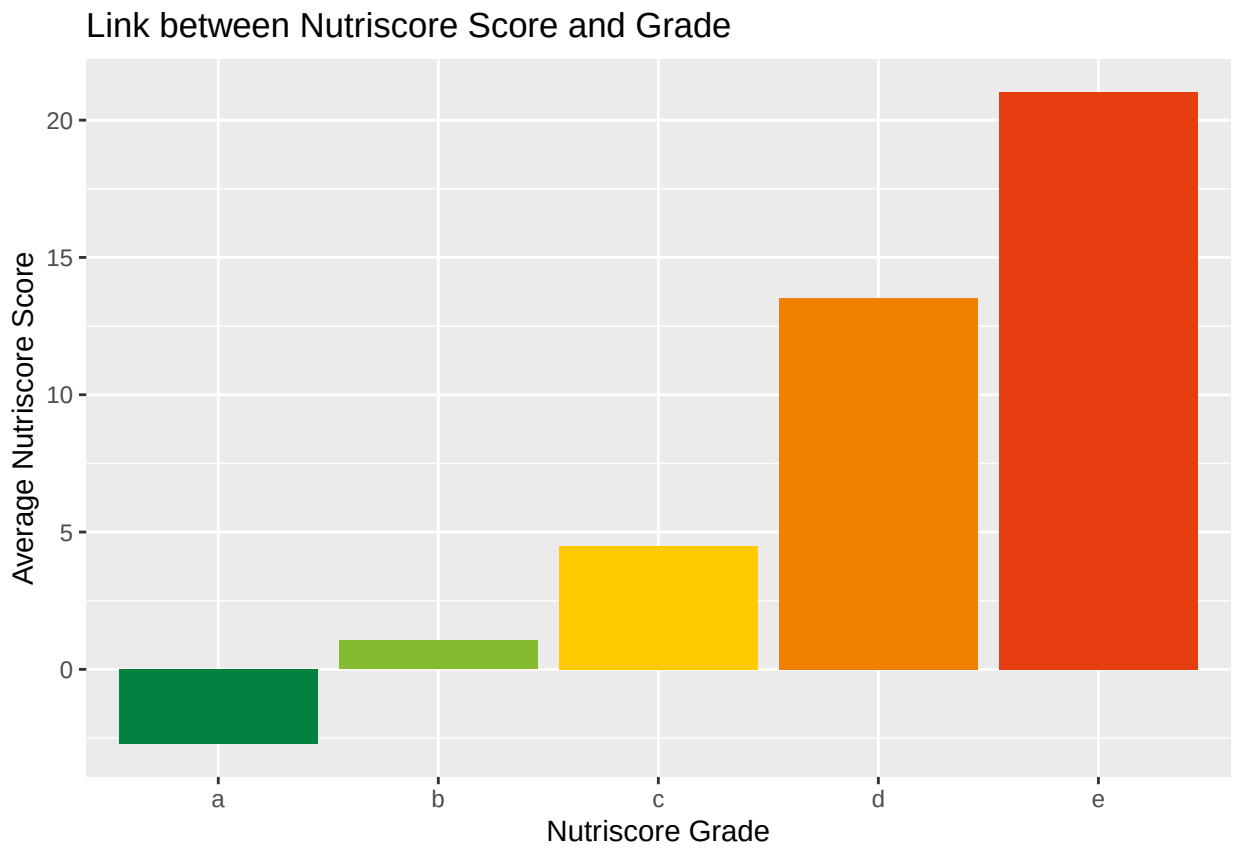
```
  scale_x_discrete(labels = c('High', 'Medium', 'Low')) + # Renaming factor levels
  theme_minimal() +
  ylim(0, 100)
```



This bar chart shows us that more than half of the product with a high fat level have a poor nutriscore rating (d or e), whereas the majority of product with a low fat level have a good nutriscore rating (a or b). The products with a medium fat level have a wider range of grades. So we can conclude that fat content have a major impact on nutriscore scores : the higher the fat content, lower the nutriscore score. Now we are going to take a look at the carbohydrate content of a product and more specifically its sugar content. Before that, we need to establish a link between the product's nutriscore score and its grade.

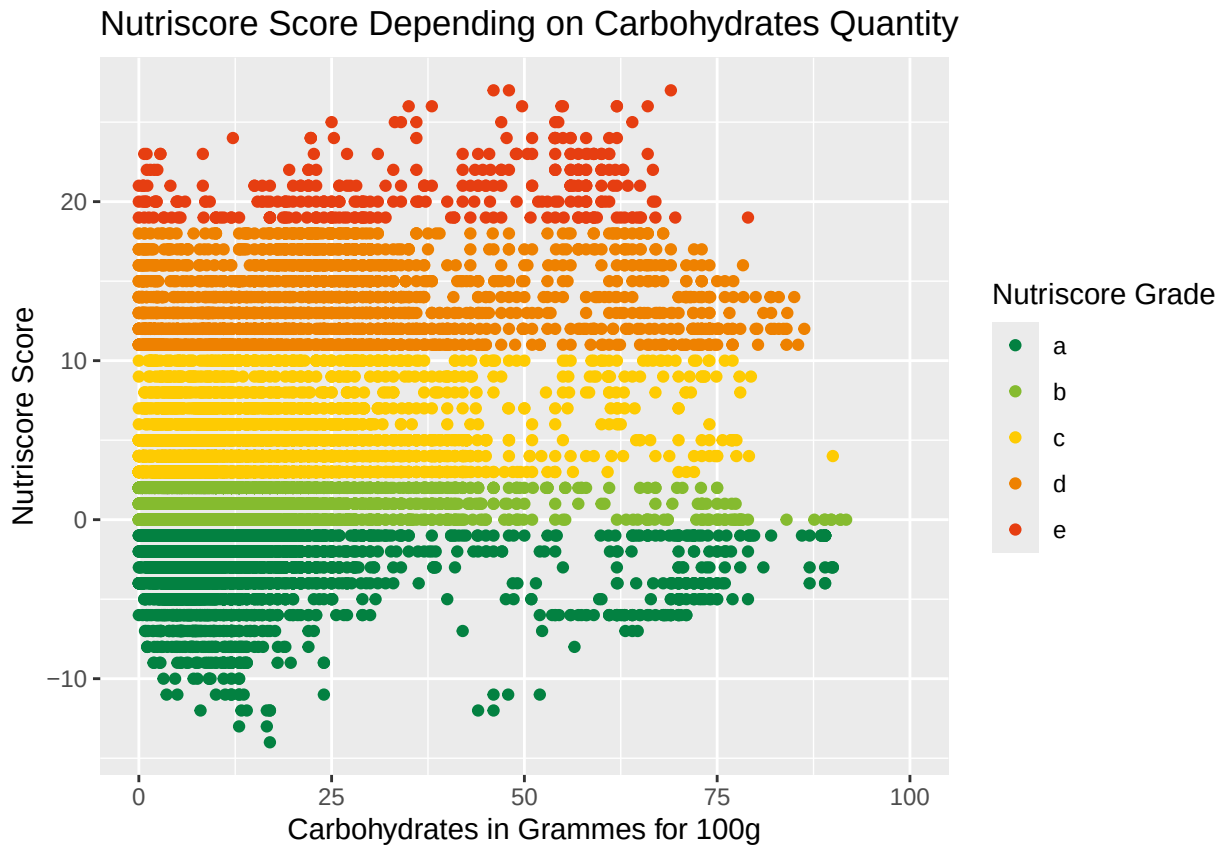
Are Nutriscore Score and Grade related ?

```
average_score <- hayotj%>%
  filter(proteins_100g < 60 & fat_100g < 90 & energy_100g < 2400 & energy_100g > 0) %>%
  group_by(nutriscore_grade)%>%
  summarise(avg_score = mean(nutriscore_score))

ggplot(average_score, aes(x= nutriscore_grade, y=avg_score ))+
  geom_bar(stat="identity", fill= c('a' = '#038141', 'b' = '#85bb2f', 'c' = '#fecb02', 'd' = '#ee8100',
  labs(x = "Nutriscore Grade", y = "Average Nutriscore Score", title = "Link between Nutriscore Score a
```

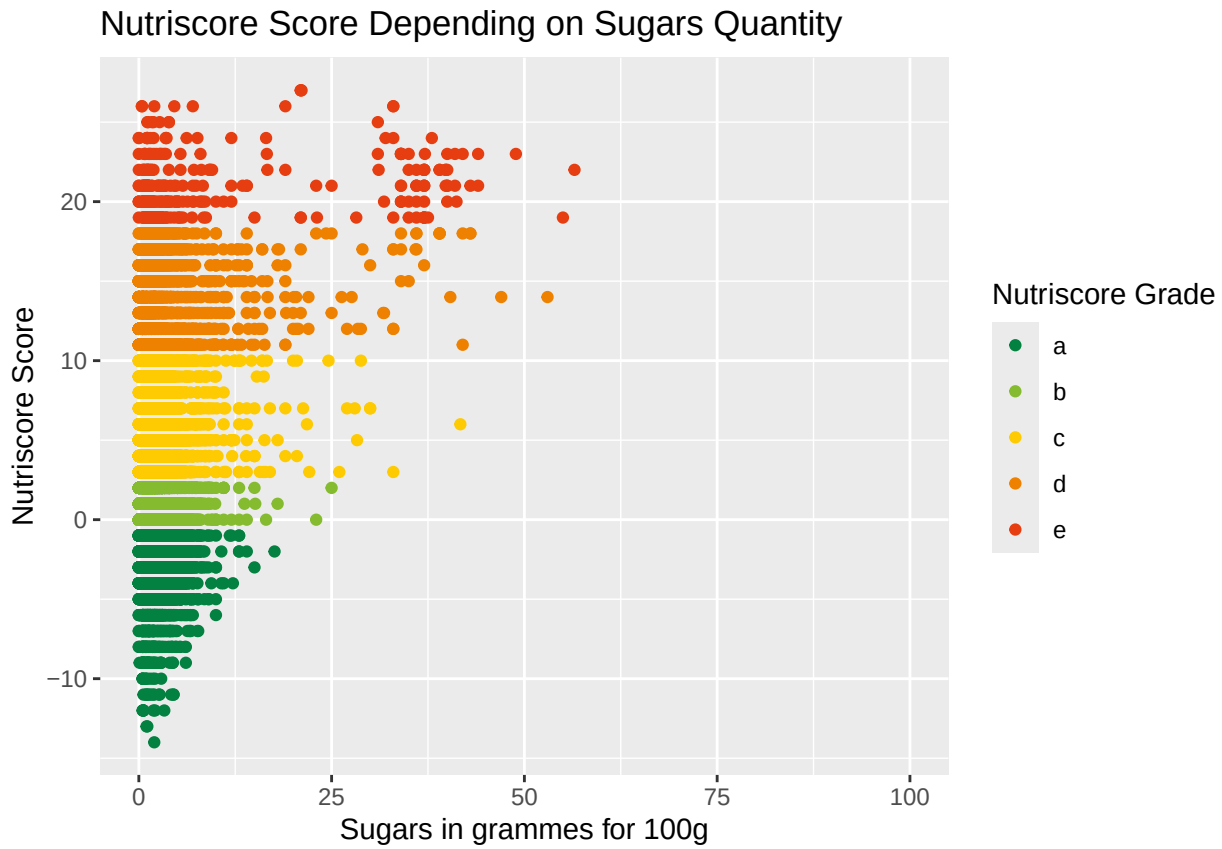


The bar chart shows us the different nutriscore grades on the X-axis by the average of the nutriscore score on the Y-axis. The average of products with nutriscore grade E is the highest, with a score of 23. In contrast, the lowest average nutriscore score is attributed to the nutriscore grade A with a score of -6. A clear link can therefore be established between nutriscore score and nutriscore grade: the lower the nutriscore score, the better the grade. We can now look at the carbohydrate and sugar content of products. # Impact of the Carbohydrates and Sugars Level on the Nutriscore

```
hayotj %>%
filter(proteins_100g < 60 & fat_100g < 90 & energy_100g < 2400 & energy_100g > 0) %>%
  ggplot(aes(x= carbohydrates_100g, y=nutriscore_score, color=nutriscore_grade))+
  geom_point()+
  scale_color_manual(values = nutriscore_colors) +
  labs(x = "Carbohydrates in Grammes for 100g", y = "Nutriscore Score", color = "Nutriscore Grade", tit
  xlim(0,100)
```



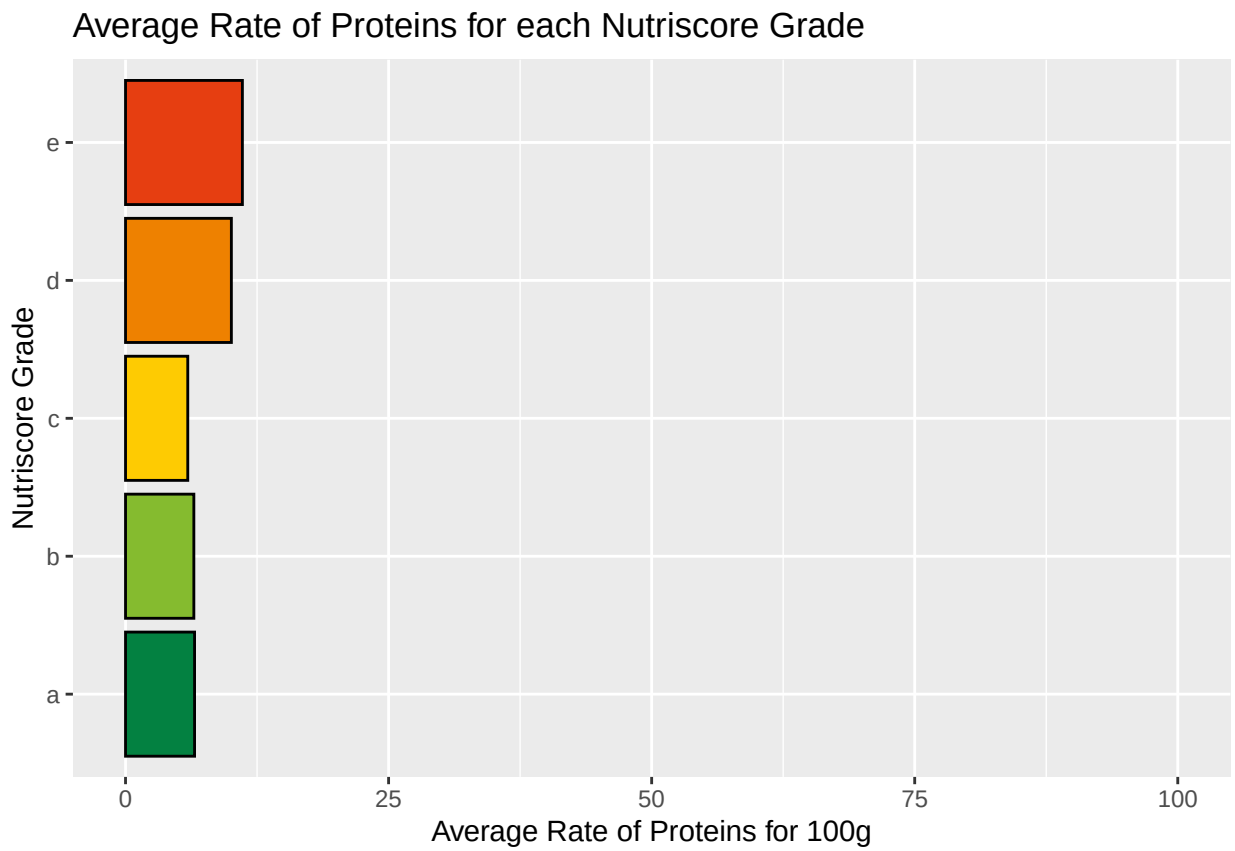

```
nutriscore_colors <- c('a' = '#038141', 'b' = '#85bb2f', 'c' = '#fecb02', 'd' = '#ee8100', 'e' = '#e63e

hayotj %>%
filter(proteins_100g < 60 & fat_100g < 90 & energy_100g < 2400 & energy_100g > 0) %>%
  ggplot(aes(x= sugars_100g, y=nutriscore_score, color=nutriscore_grade))+
  geom_point()+
  scale_color_manual(values = nutriscore_colors)+
  labs(x = "Sugars in grammes for 100g", y = "Nutriscore Score", color = "Nutriscore Grade", title = "Nu
  xlim(0,100)
```



If we look at the carbohydrates graph, we can't see major differences between each grade, but we can still notice that products with a very high quantity of carbohydrates are located in the better grades (a and b) even if it is not very significant. But if we go deeper and look at the quantity of sugars, we can clearly see that products in a and b grades have low quantities of sugars and, on the other hand, products with c, d or e grades tend to have a higher quantity of sugars. So carbohydrates in general doesn't have a meaningful impact, but sugars have a pretty decent impact on the grade of the product. Now we have to look at the third nutrional value of a product : the proteins. # Impact of the Proteins Rate on the Nutriscore Grade

```
nutriscore_colors = c('a' = '#038141', 'b' = '#85bb2f', 'c' = '#fecb02', 'd' = '#ee8100', 'e' = '#e63e1

average_score <- hayotj%>%
  filter(proteins_100g < 60 & fat_100g < 90 & energy_100g < 2400 & energy_100g > 0) %>%
  group_by(nutriscore_grade)%>%
  summarise(avg_proteins = mean(proteins_100g))

ggplot(average_score, aes(x= avg_proteins, y=nutriscore_grade))+
  geom_histogram(stat="identity", fill = nutriscore_colors, color='black') +
  labs(x = "Average Rate of Proteins for 100g", y = "Nutriscore Grade", title = "Average Rate of Protei
  xlim(0, 100)
```

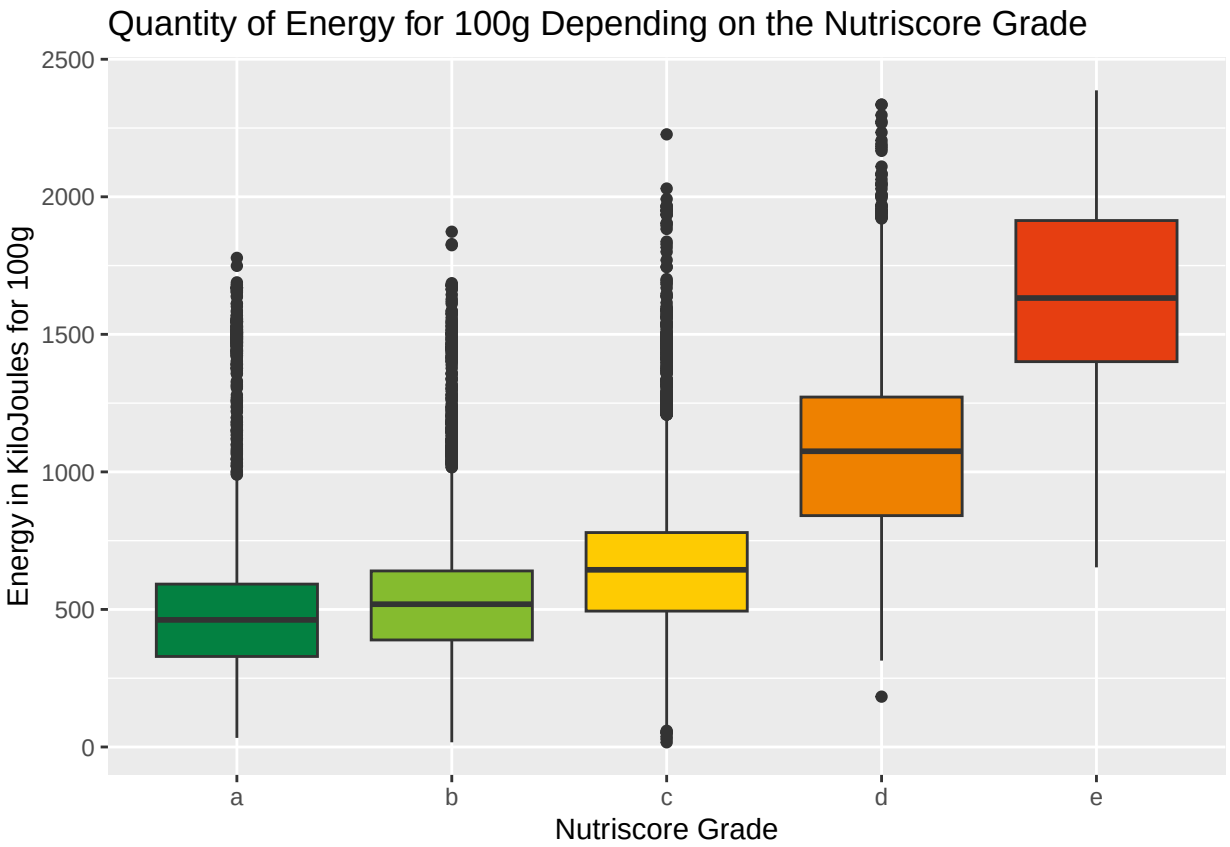
```
## Warning in geom_histogram(stat = "identity", fill = nutriscore_colors, color =
## "black"): Ignoring unknown parameters: `binwidth`, `bins`, and `pad`
```



This histogram shows the average protein percentage, on the X-axis, as a function of nutriscore grade, on the Y-axis. We can see that the percentage of proteins of the nutriscore grades D and E are almost the same. On the other hand, we have a similar protein percentages for grades A, B and C. The protein percentage of products with nutriscore grade E is only 5% higher than the protein percentage of products with grade A. So we can conclude that the protein content of a product has no real impact on its nutriscore grade. At the beginning, we noticed that an other value differed with the nutriscore grade : let's take a look at the energy variable.

Impact of the Quantity of Energy on the Nutriscore Grade

```
nutriscore_colors = c('a' = '#038141', 'b' = '#85bb2f', 'c' = '#fecb02', 'd' = '#ee8100', 'e' = '#e63e1

hayotj %>%
  filter(proteins_100g < 60 & fat_100g < 90 & energy_100g < 2400 & energy_100g > 0) %>%
  ggplot() +
  geom_boxplot(aes(x = nutriscore_grade, y = energy_100g), fill = nutriscore_colors) +
  labs(x = "Nutriscore Grade", y = "Energy in KiloJoules for 100g", title = "Quantity of Energy for 100
```


Quantity of Energy for 100g Depending on the Nutriscore Grade
Energy in KiloJoules for 100g
2500
2000
1500
1000
500
0
a
b
c
d
e
Nutriscore Grade