Rendu 3

title: "Nutrition score - Sweets - Team D12"
author: "Simon Zeru & Toni Rey"
date: "2024-04-05"
output:
html_document:
toc: true
toc_depth: '3'
df_print: paged
pdf_document:
toc: true
toc_depth: 3
number_sections: true
latex_engine: pdflatex

```
library(dplyr)
library(ggplot2)
library(readr)
library(kableExtra)

sweets <- read_delim("zerus.csv", delim = "\\t")
```

Introduction

What is the Nutriscore ?

The Nutri-Score, introduced in France in 2017 and now adopted in several European countries, is a nutritional labeling system that ranks foods from A to E. Its aim is to help consumers make healthier food choices on a daily basis. To achieve this, it takes into account various nutritional elements such as calories, sugars, saturated fats, fibers, proteins, as well as the presence of fruits, vegetables, legumes, and nuts.

What are the products that we will study ?

The aim of our project is to find out whether the Nutriscore on sweets is reliable, using a filtered database to perform data analysis with visualization in order to draw relevant conclusions. (To meet expectations).

Our goal is to educate American consumers about their nutrition on sweets, by addressing them in a popularized way using a report written in English so that they can be aware of the reliability of this indicator.

We can measure the achievement of the objective (the success indicator) by public (customer) satisfaction.

We have one week to complete this project.

How many observations do we have ?

We dispose of 19843 observations of sweets in our database.

Statistical desciption of the different variables on the database

It contains 30 variables:

Qualitative

Nominal

- code, url, product_name, brands_tags, stores, owner, food_groups, labels_tags, countries, countries_tag, organic, vegan, vegetarian, palm_oil

Ordinal

- level_fat, level_saturated_fat, level_sugars, level_salt, nutriscore_score, nutriscore_grade

Quantitative

Continuous

- fat_100g, saturated_fat_100g, sugars_100g, proteins_100g, carbohydrates_100g, energy_100g, salt_100g, sodium_100g

Discrete

- Quantity

Distribution of sweets and Nutriscore (A to E)

```
# Define a custom color palette for the Nutriscore grades
nutriscore_colors <- c('a' = '#1AAE13',    # Green
                       'b' = '#77D700',    # Light green
                       'c' = '#FFF200',    # Yellow
                       'd' = '#FF8200',    # Orange
                       'e' = '#E40A0A'    # Red
                       )

# Create a bar plot with the custom color palette
ggplot(sweets, aes(x = nutriscore_grade)) +
  geom_bar(fill = nutriscore_colors, color = 'black') +
  labs(title = 'Distribution of Nutriscore Grades for Sweet
s',
       x = 'Nutriscore Grade',
       y = 'Frequency') +
  theme_minimal()
```

The bar plot indicates a right-skewed distribution, with the majority of products falling in the lower grades (D and E). Specifically, there is a significant concentration of products graded D and E, comprising over 10,000 observations for grade D and more than 8,000 observations for grade E. This concentration suggests that a substantial portion of sweets in the dataset are classified as less healthy options according to the Nutriscore system.

Conversely, there are relatively fewer products graded A, with only a small number of observations. Grades B and C fall in between, with grade C having approximately 3 to 4 times more observations than grade B. This indicates a gradual decrease in the frequency of products as we move towards healthier Nutriscore grades.

Overall, the distribution highlights the prevalence of less healthy sweets, as evidenced by the higher frequency of grades D and E compared to grades A, B, and C. This information can be valuable for consumers looking to make informed choices about the nutritional quality of sweets based on their Nutriscore classification.

Compositions

We separated the sweets in two distinct categories :

- Sweets graded "Healthy" with A or B Nutriscore
- Sweets graded "Unhealthy" with C, D or E Nutriscore

And we made grading colors also for future diagrams

```
# Filter the dataset for sweets graded 'a' or 'b'
healthy_sweets <- subset(sweets, nutriscore_grade %in% c('a', 'b'))

# Filter the dataset for sweets graded 'c', 'd' and 'e'
unhealthy_sweets <- subset(sweets, nutriscore_grade %in% c('c', 'd', 'e'))

# Define colors
grading_colors <- c("High" = "#E40A0A", "Medium" = "#FF8200", "Low" = "#77D700")
```

Proteins and Nutriscore

```
summary(sweets$proteins_100g)
```

On average, sweets contain 3.38 g of proteins per 100g. That's low but concerning, considering that sweets shouldn't have proteins at all. Proteins could come from pork fat added. Other ingredients, such as nuts, dairy products, or protein fortification, could also contribute to the protein content in sweets.

There are likely some extreme outliers in the data, given the large difference between the maximum value (73.330 grams per 100 grams).

These outliers may represent specialized or unique types of sweets with unusually high protein content.

Distributions

Proteins and "Healthy" Sweets

```
ggplot(healthy_sweets, aes(x = proteins_100g)) +
  geom_density(alpha = 0.5, fill = "skyblue") +
  labs(title = 'Distribution of Proteins in Healthy Sweet
s',
       x = 'Proteins for 100g',
       y = 'Density') +
  theme_minimal()
```

The majority of sweets fall below 2% of proteins. We can observe few spikes between 5g to 10g. There is outliers aroud 30g and 60g.

Basically, there is a very low density of healthy candies below the 10-gram mark. This shows that Nutriscore tends to value less proteins in candies.

Proteins and "Unhealthy" Sweets

```
ggplot(unhealthy_sweets, aes(x = proteins_100g)) +
  geom_density(alpha = 0.5, fill = "skyblue") +
  labs(title = 'Distribution of Proteins in Unhealthy Sweet
s',
       x = 'Proteins for 100g',
       y = 'Density') +
  theme_minimal()
```

We recreated the same graph, but this time for candies categorized as "Unhealthy." A significant difference in density is evident in this second graph. Candies in the "Unhealthy" category exhibit higher density within the range of 0.5 to 10 grams.
This observation can be explained by the fact that the "Unhealthy" category comprises candies with a higher protein content.

Fat and Nutriscore

```
summary(sweets$fat_100g)
```

The summary statistics reveal that the fat content in sweets varies considerably, with a minimum value of 0g per 100g and a maximum value of 71.43g per 100g. The median fat content is 5g per 100g, indicating that half of the sweets have a fat content below this value, while the other half have a fat content above it. The mean fat content is higher at 13.77g per 100g, suggesting that the distribution may be skewed towards higher fat content.

Distributions

Healthy Sweets

```
# Convert level_fat to factor with corresponding labels
healthy_sweets$level_fat <- factor(healthy_sweets$level_fa
t,
                                   levels = c("l", "m",
"h"),
                                   labels = c("Low", "Mediu
m", "High"))

# Create the plot
ggplot(healthy_sweets, aes(x = level_fat, fill = level_fa
t)) +
  geom_bar(show.legend = FALSE) +
  scale_fill_manual(values = grading_colors) +
  labs(title = "Distribution of Healthy Sweets (Graded A or
B) by Fat Levels",
       x = "Fat Level",
       y = "Frequency") +
  theme_minimal()
```

The graph illustrates that the majority of sweets graded A or B by the Nutriscore system have a low fat content, with approximately 225 instances falling into this category. Additionally, there are just under 50 instances classified as having a medium fat content, and very few instances categorized as high fat content.

It suggests that the Nutriscore system effectively distinguishes between sweets based on their fat levels. Products with higher Nutriscore grades ('a' or 'b') tend to have lower fat content, which aligns with the goals of the Nutriscore system to guide consumers towards healthier food choices. The prevalence of low fat content among sweets with higher Nutriscore grades indicates that these products are generally healthier options within the sweets category.

Unhealthy Sweets

```
# Convert level_fat to factor with corresponding labels
unhealthy_sweets$level_fat <- factor(unhealthy_sweets$level
_fat,
                                     levels = c("l", "m",
"h"),
                                     labels = c("Low", "Mediu
m", "High"))

# Create the plot
ggplot(unhealthy_sweets, aes(x = level_fat, fill = level_fa
t)) +
  geom_bar() +
  scale_fill_manual(values = grading_colors) +
  labs(title = "Distribution of Unhealthy Sweets (Graded C,
D or E) by Fat Levels",
       x = "Fat Level",
       y = "Frequency") +
  theme_minimal()
```

This shows a strange result :

The ratio of low fat candies categorized as "unhealthy" was expected to be the lowest, but almost 10000 candies are low-fat, putting them ahead of medium and high fat sweets.

It could be explained by other factors like sugar or energy levels...

It challenges the assumption that less fat means healthier option for Nutriscore.

Consumers may be misled by this label, overlooking potential health risks posed by high sugar content or other unhealthy additives.

Conversely, a notable number of sweets boast high fat levels, potentially contributing to calorie-dense diets and unhealthy eating habits if consumed excessively, making a great point for Nutriscore.

Boxplots

To verify our assumptions, we've generated a boxplot showing the relationship between fat levels and energy in candies.

Healthy Sweets

```
ggplot(healthy_sweets, aes(x = level_fat, y = energy_100g, fill = level_fat)) +
  geom_boxplot(show.legend = FALSE) +
  scale_fill_manual(values = grading_colors) +
  labs(title = "Boxplot of Energy by Fat Levels in Healthy Sweets (Graded A or B)",
       x = "Fat Level",
       y = "Energy per 100g") +
  theme_minimal()
```

High Fat, High Energy:

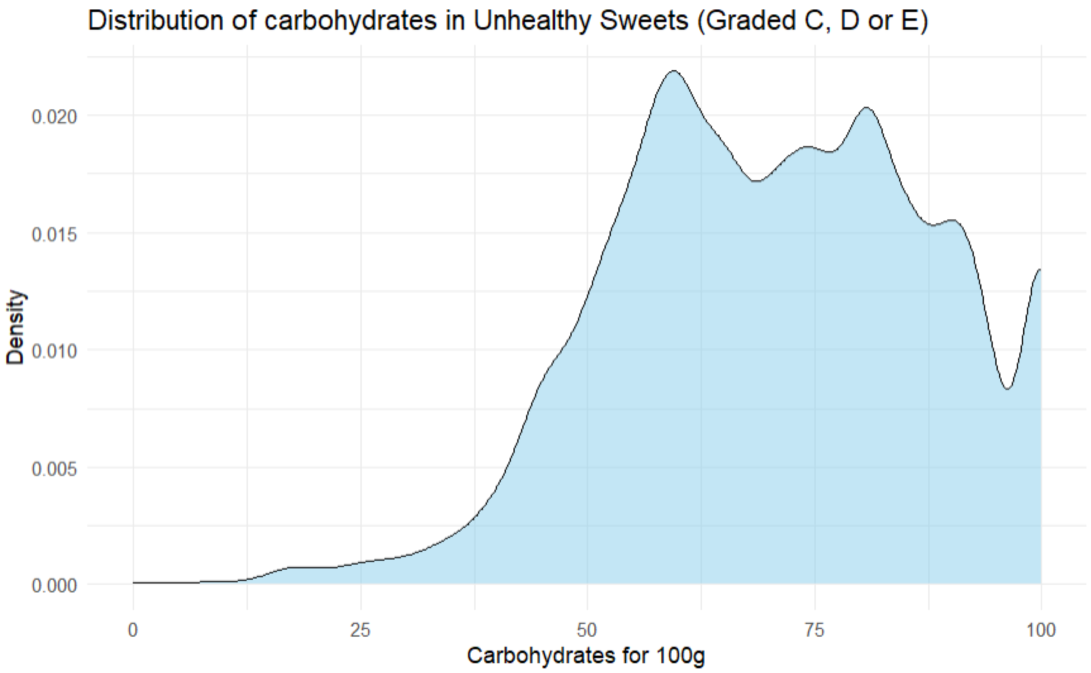
The sweets with high fat levels have a wider range and higher median energy content compared to those with low and medium fat levels. This suggests that high-fat sweets are more energy-dense, which is expected as fats have more than double the calories per gram compared to proteins and carbohydrates.

Furthermore, healthy candies with high fat levels in average are around 4 times higher in energy levels, showing us how much fat plays a role on calories.

Energy Content and Healthiness:

While these sweets are graded as healthy (A or B) on Nutriscore, it's important to note that they can still be quite energy-dense, especially those with high fat levels. Therefore, portion control is crucial even when consuming healthy sweets.

Unhealthy Sweets



```
ggplot(unhealthy_sweets, aes(x = level_fat, y = energy_100g, fill = level_fat)) +
  geom_boxplot(show.legend = FALSE) +
  scale_fill_manual(values = grading_colors) +
  labs(title = "Boxplot of Energy by Fat Levels in Unhealthy Sweets (Graded C, D or E)",
       x = "Fat Level",
       y = "Energy per 100g") +
  theme_minimal()
```

Comparing it with the previous graph, a striking contrast emerges in energy levels.

Low-fat candies labeled as "unhealthy" have a first quartile (25%) energy content of approximately 1200 kJ per 100g, significantly higher than their "healthy" counterparts.

Conversely, a considerable number of extreme low-energy values are observed, indicating a broad spectrum of energy content among sweets.

Both medium and high fat levels follows a logical sequel, confirming the relationship mentioned earlier.

Interestingly, this finding challenges our earlier assertion that low-fat candies were classified as "unhealthy" solely due to potential high energy content. Instead, it reveals a nuanced relationship between fat and energy levels in sweets.

Conclusion

This comparison underscores the importance of comprehensive nutritional understanding and not relying solely on single factors such as fat content when evaluating the healthiness of food products. It's always a good idea to look at the bigger picture, considering aspects like sugar content, artificial additives, and overall calorie content. It also highlights the need for transparency and availability of complete nutritional information for all food products. It's interesting to note that while high fat content generally leads to higher energy content, it doesn't necessarily mean the sweet is unhealthy. The overall nutritional profile needs to be considered.

Carbohydrates and Nutriscore

```
summary(sweets$carbohydrates_100g)
```

The summary statistics for the distribution of carbohydrates in sweets reveal that the carbohydrate content per 100g varies widely.

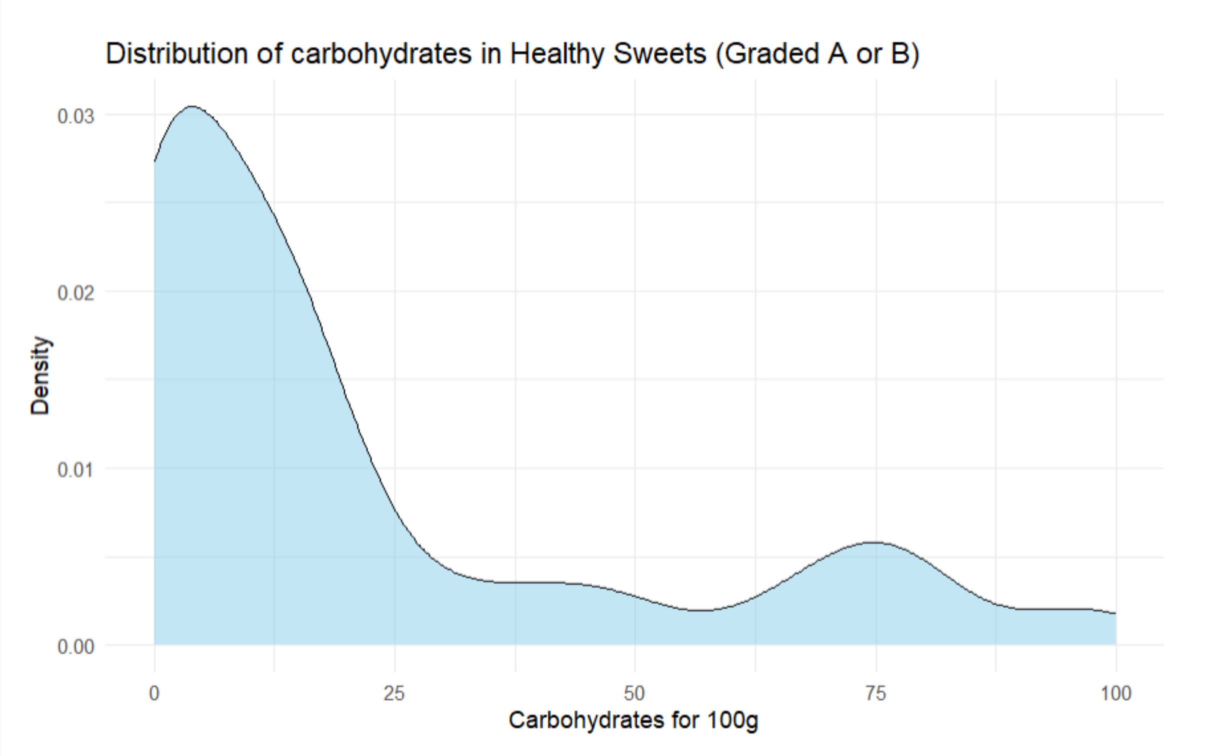
The data suggests that the majority of sweets contain carbohydrates ranging from approximately 57.5g to 83.33g per 100g serving, with a median value of 70.21g.

This distribution is indicative of a typical pattern observed in carbohydrate content, with a significant portion of sweets falling within this range.

However, it's worth noting that there are instances of extreme values, particularly at both ends of the spectrum, with some sweets exhibiting very low (close to 0g) or very high (close to 100g) carbohydrate content. These outliers contribute to the variability observed in the distribution, highlighting the diversity in carbohydrate content among sweets.

Distributions

Carbohydrates and "Healthy" Sweets



```
ggplot(healthy_sweets, aes(x = carbohydrates_100g)) +
  geom_density(alpha = 0.5, fill = "skyblue") +
  labs(title = 'Distribution of carbohydrates in Healthy Sweets (Graded A or B)',
       x = 'Carbohydrates for 100g',
       y = 'Density') +
  theme_minimal()
```

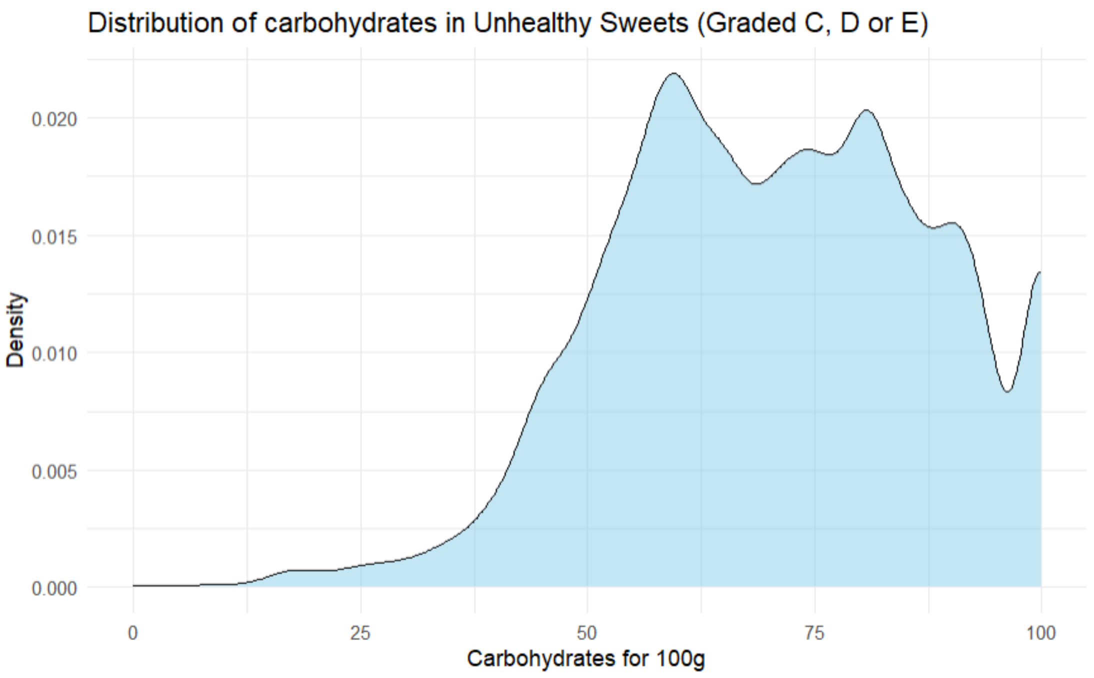
We created a density graph representing the distribution of carbohydrates in healthy candies. On the x-axis, we have carbohydrates per 100g, while on the y-axis, we have density. We observe a strong concentration of healthy candies between 0 and 30 grams of carbohydrates. This trend can be explained by the fact that the Nutri-Score is based on the quantity of carbohydrates in foods to determine their categories. Therefore, it is logical to find more healthy candies in the range of 0 to 25 grams of carbohydrates rather than above.

The observation of a slight increase in sugar content from 60g to 75g among sweets categorized as "healthy" is indeed intriguing. This anomaly prompts several potential explanations.

One possibility is that the Nutriscore system, while generally effective, may occasionally misclassify certain products due to its reliance on predefined nutritional thresholds. It's conceivable that some sweets with marginally higher sugar content could still receive favorable Nutriscore ratings if they meet other criteria for balanced nutrition.

Alternatively, these outliers could be attributed to variations in product formulations or labeling inaccuracies rather than a systematic issue with the Nutriscore grading itself. It's not uncommon for food products to deviate slightly from expected nutritional profiles due to manufacturing inconsistencies or inadvertent errors in labeling.

Carbohydrates and "Unhealthy" Sweets



```
ggplot(unhealthy_sweets, aes(x = carbohydrates_100g)) +
  geom_density(alpha = 0.5, fill = "skyblue") +
  labs(title = 'Distribution of carbohydrates in Unhealthy 
Sweets (Graded C, D or E)',
       x = 'Carbohydrates for 100g',
       y = 'Density') +
  theme_minimal()
```

We reproduced the same graph as mentioned earlier, but this time focusing on "Unhealthy" graded candies to support our hypothesis.
We observe a high density of candies with a carbohydrates content of more than 30 grams. This observation confirms our hypothesis that the Nutriscore partly relies on the amount of carbohydrates present in candies.

Sugar and Nutriscore

```
# Summary statistics for sugar content in sweets
summary(sweets$sugars_100g)
```

On average, sweets contain 54.82 g of sugar per 100g. That's more than half of the composition, which is quite concerning. Perhaps we should all reconsider consuming sweets...

The majority of sweets (50%) have a sugar content below the median value of 53.81 grams per 100g.

However, there is a noticeable right-skew in the distribution, with the mean (54.82 grams per 100g) being slightly higher than the median, indicating that some sweets have a high sugar content, pulling the mean upwards.

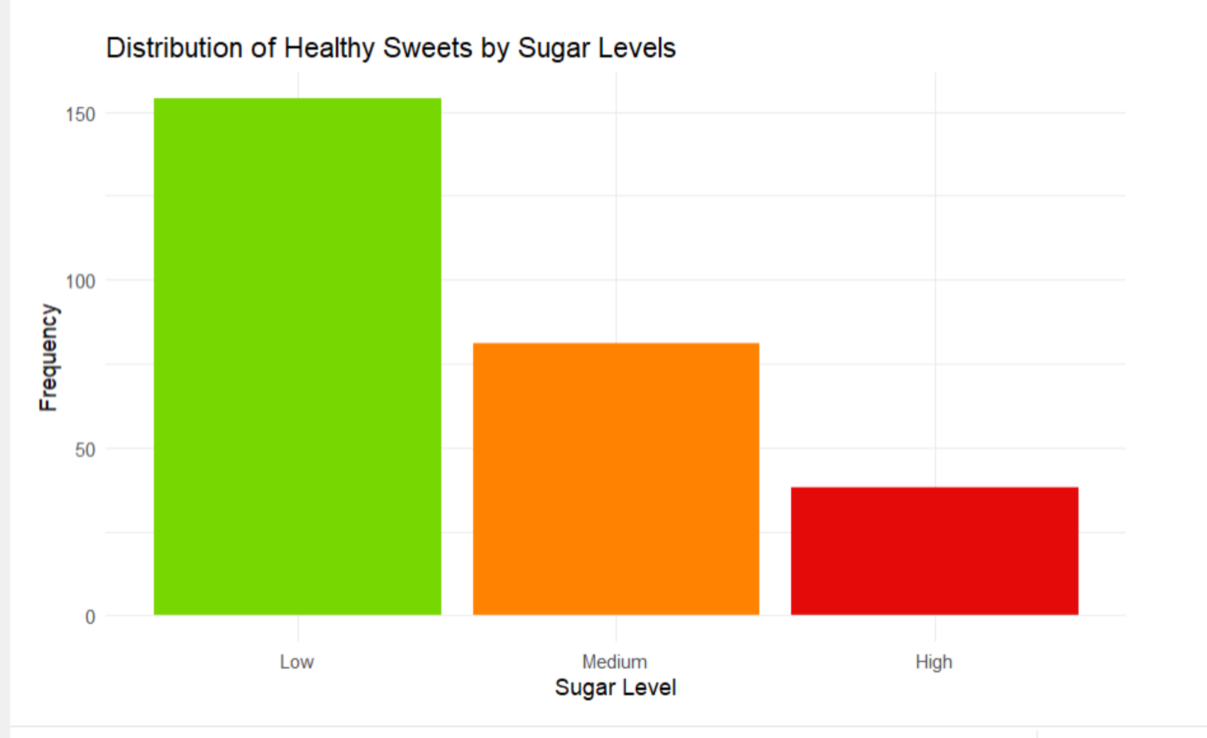
The maximum observed sugar content of 100.00 grams per 100g highlights the presence of sweets with very high sugar levels, which could pose health concerns if consumed excessively.

Theses results could give us confidence in the Nutriscore system's ability to accurately assess sugar content in food products relative to other nutritional components.

Let's decompose on healthy and unhealthy sweets.

Distributions

Healthy Sweets and Sugar Level (High, Medium or Low)



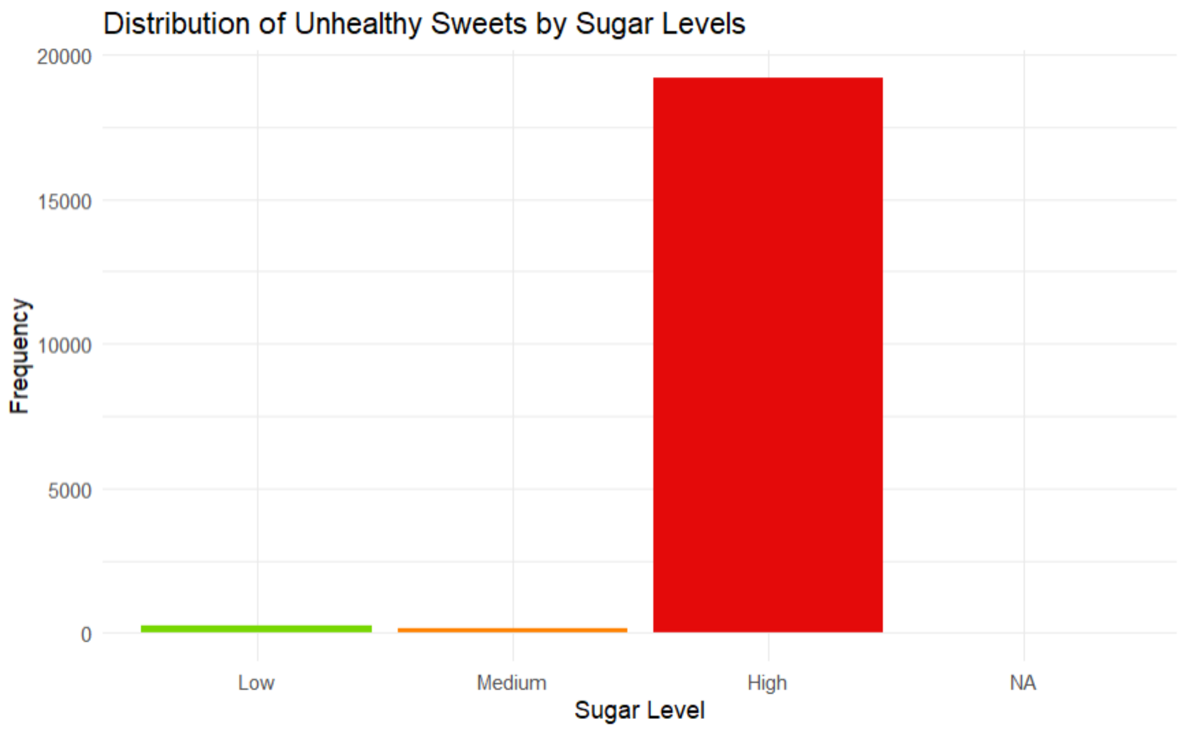
```
# Convert level_sugars to factor with corresponding labels
healthy_sweets$level_sugars <- factor(healthy_sweets$level_sugars,
                                      levels = c("l", "m", "h"),
                                      labels = c("Low", "Medium", "High"))

# Create the plot
ggplot(healthy_sweets, aes(x = level_sugars, fill = level_sugars)) +
  geom_bar(show.legend = FALSE) +
  scale_fill_manual(values = grading_colors) +
  labs(title = "Distribution of Healthy Sweets by Sugar Levels",
       x = "Sugar Level",
       y = "Frequency") +
  theme_minimal()
```

The graph indicates that the majority of sweets graded A or B according to the Nutriscore system have a low sugar content, with over 150 instances falling into this category. Additionally, there are over 75 instances classified as having a medium sugar content and around 40 instances categorized as high sugar content.

This distribution suggests that the Nutriscore system is effective in differentiating between sweets based on their sugar levels. Products with higher Nutriscore grades ('a' or 'b') tend to have lower sugar content, aligning with the goals of the Nutriscore system to guide consumers towards healthier food choices. The prevalence of low sugar content among sweets with higher Nutriscore grades reinforces the idea that these products are generally healthier options within the sweets category.

Unhealthy Sweets and Sugar Level (High, Medium or Low)



```
# Convert level_sugars to factor with corresponding labels
unhealthy_sweets$level_sugars <- factor(unhealthy_sweets$le
vel_sugars,
                                      levels = c("l", "m", "h"),
                                      labels = c("Low", "Medium", "High"))

# Create the plot
ggplot(unhealthy_sweets, aes(x = level_sugars, fill = level_sugars)) +
  geom_bar(show.legend = FALSE) +
  scale_fill_manual(values = grading_colors) +
  labs(title = "Distribution of Unhealthy Sweets by Sugar Levels",
       x = "Sugar Level",
       y = "Frequency") +
  theme_minimal()
```

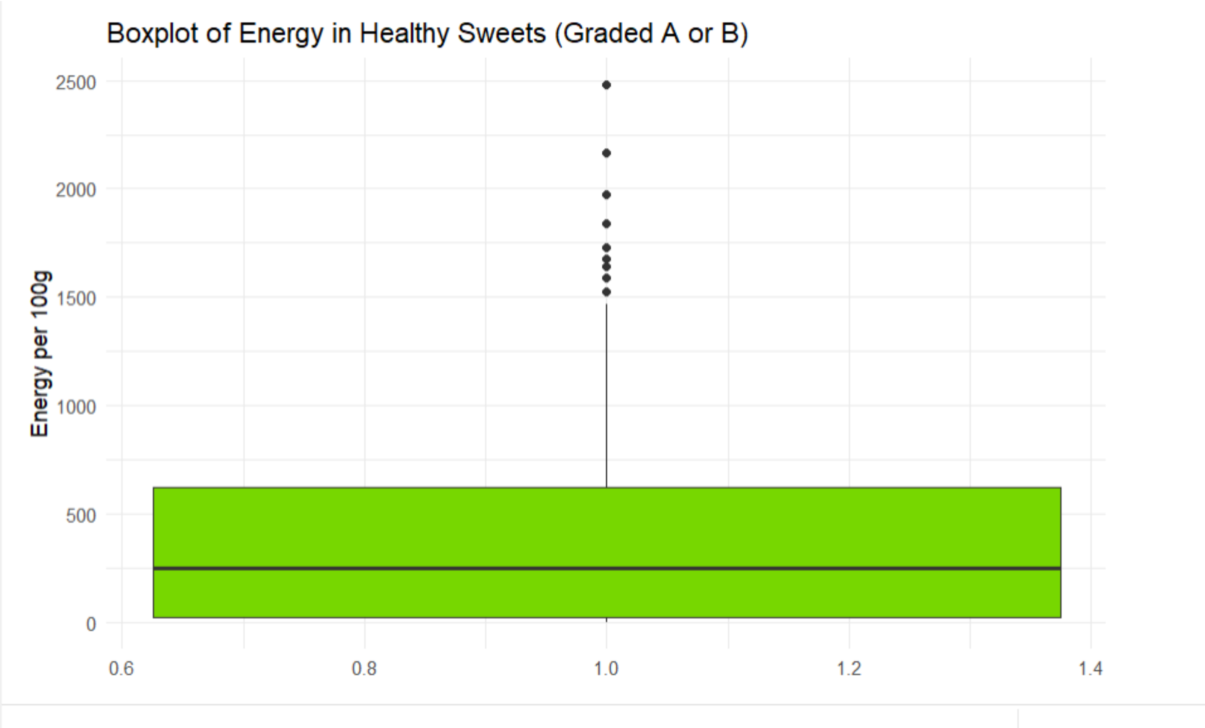
We reproduced the same graph, but this time focusing on candies from the Unhealthy category.
One striking observation is that the frequency of candies with low to moderate sugar levels is almost zero, while the frequency of candies with high sugar levels is very high. Almost all "unhealthy" sweets are high in sugar.
This observation can be explained by the fact that sugar content plays an important role in the Nutriscore. Therefore, it is logical to find candies with a high sugar concentration in the Unhealthy category.

This is a great point for the Nutriscore on preventing consumers about high sugar levels.

Energy and Nutriscore

Boxplots

Boxplot of Energy in Healthy Sweets

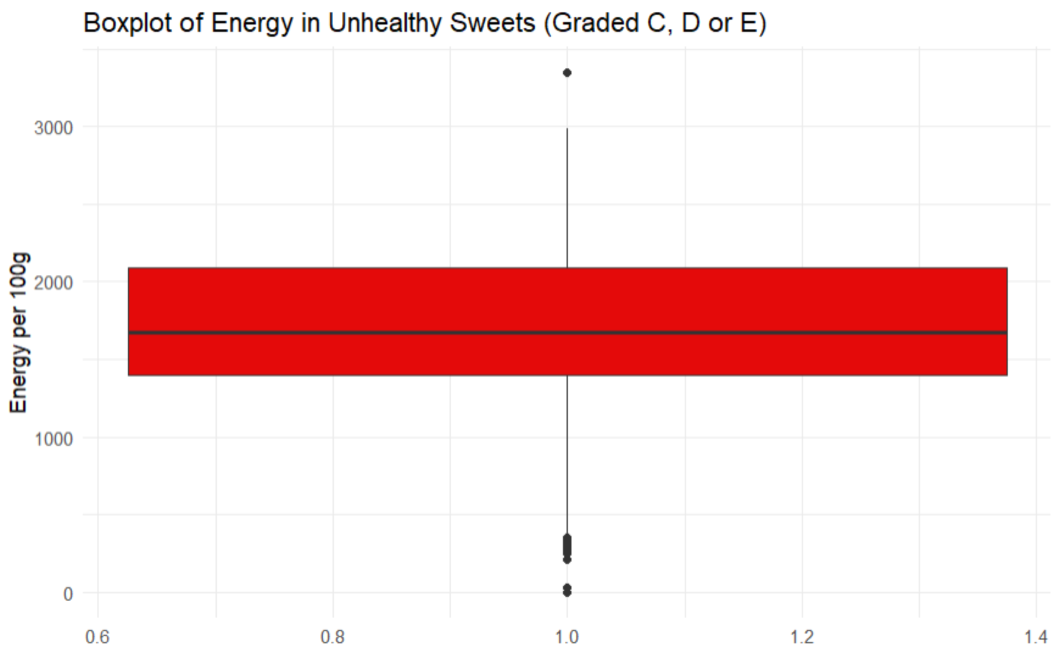


```
ggplot(healthy_sweets, aes(x = 1, y = energy_100g, fill =
'Energy')) +
  geom_boxplot(show.legend = FALSE) +
  labs(title = "Boxplot of Energy in Healthy Sweets (Graded
A or B)",
       x = NULL,
       y = "Energy per 100g") +
  scale_fill_manual(values = '#77D700') +
  theme_minimal()
```

We designed a box plot representing the distribution of candies according to their energy. Through this representation, we have access to measures such as the median, the mean, standard deviations, and extreme values.
Examining the graph, we notice that most of the healthy candies have their energy concentrated between 0 and 600 kJ, with a median at 260 kJ which is lower than the mean. This observation highlights the importance of considering extreme values, as they can influence the results, leading to a median lower than the mean.
Furthermore, we note that all candies considered healthy remain below 1000 kJ. This finding underscores the importance of low energy content for candies categorized as healthy.

Despite the majority of healthy candies falling within the lower energy range, there may be instances where candies with higher energy levels are still classified as healthy. This discrepancy could indicate potential misclassifications within the Nutriscore system, where certain candies may not accurately reflect their nutritional quality based solely on energy content.

Let's compare it with unhealthy sweets...



```r
ggplot(unhealthy_sweets, aes(x = 1, y = energy_100g, fill = 'Energy')) +
  geom_boxplot(show.legend = FALSE) +
  labs(title = "Boxplot of Energy in Unhealthy Sweets (Graded C, D or E)",
       x = NULL,
       y = "Energy per 100g") +
  scale_fill_manual(values = '#E40A0A') +
  theme_minimal()
```

We reproduced the same graph, but this time focusing on candies from the Unhealthy category.

At first glance, we notice that the majority of values are significantly higher than for the Healthy candy category. This time, the range of values is between 1400 and 2500, with a median of 1750, which is almost equal to the mean.

This observation suggests firstly that extreme values have less impact than the first time. Moreover, it highlights the fact that candies in the Unhealthy category contain much more energy than healthy candies, which supports our theory about the importance of the energy content of candies in their classification.

Overall, the consistent observation that all candies categorized as healthy remain below 1000 kJ, and the notable disparity in energy levels between candies from the Healthy and Unhealthy categories, further reinforces the Nutriscore system's adherence to established nutritional guidelines. This consistent trend underscores the system's effectiveness in guiding consumers towards healthier choices based on energy content, while also highlighting potential areas for improvement or refinement to ensure its continued accuracy and relevance in promoting healthier dietary habits.

Oher factors

Are sweets vegan ?

```
# Calculate the percentage of sweets that are classified as vegan
percentage_vegan <- mean(sweets$vegan == 't', na.rm = TRUE) * 100

# Print the result
print(paste("Percentage of sweets classified as vegan:", round(percentage_vegan, 2), "%"))
```

Based on the analysis, it appears that none of the sweets in the dataset are classified as vegan. This suggests that all the products included in the dataset contain ingredients derived from animal sources. The absence of vegan sweets may indicate a lack of plant-based alternatives in the dataset or a prevalence of animal-derived ingredients commonly used in sweet products.

Because the Nutriscore, designed to assess the overall nutritional quality of food products, typically rewards items with higher fruit and vegetable content due to their positive health attributes, it makes sense that sweets on average would be badly graded based on this parameter.

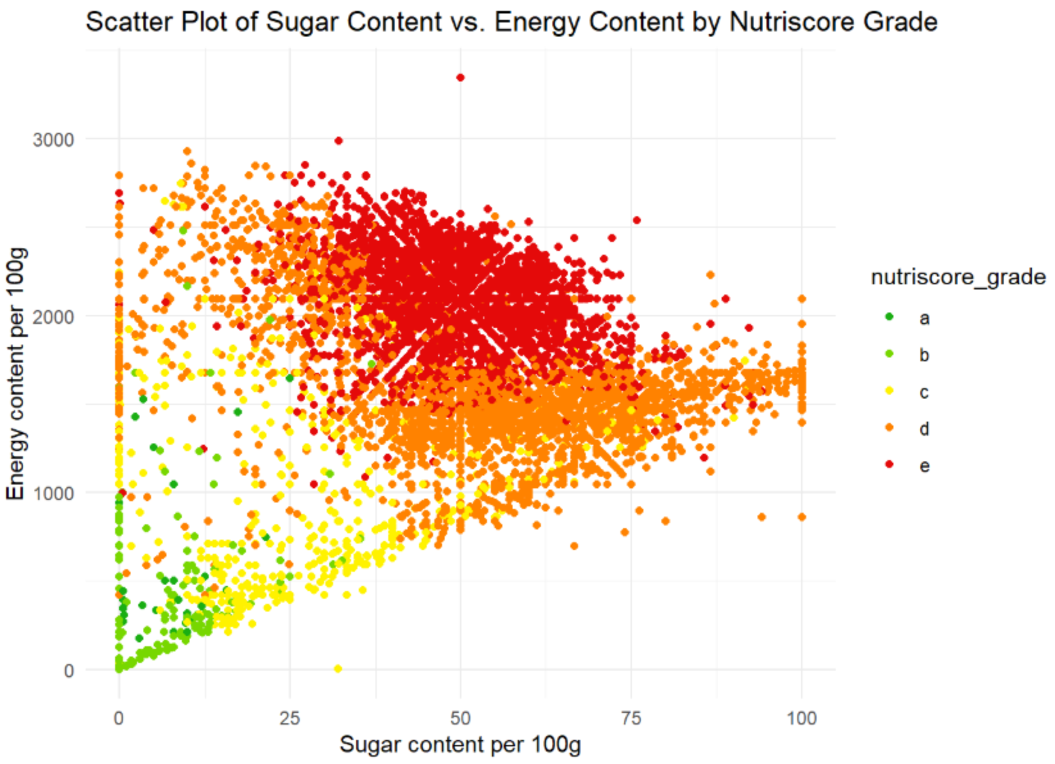
Does Sweets Contain Palm Oil ?

```
# Calculate the percentage of sweets that are classified as vegan
percentage_palm <- mean(sweets$palm_oil == 't', na.rm = TRUE) * 100

# Print the result
print(paste("Percentage of sweets containing palm oil :", round(percentage_palm, 2), "%"))
```

None of the Sweets contain palm oil. No further analysis needed.

Energy Content vs Sugar Content



Conclusion

Based on our analysis, we can make several relevant observations about the Nutriscore system:

Correlation with Healthiness: Our analysis reveals a correlation between the Nutriscore grading and various nutritional aspects of sweets. For example, sweets graded as "healthy" tend to have lower levels of fat, sugars, and energy compared to those graded as "unhealthy." This suggests that the Nutriscore system is effective in categorizing sweets based on their nutritional quality.

Limitations: Despite its effectiveness in categorizing sweets, the Nutriscore system has its limitations. For instance, we observed instances where sweets categorized as "healthy" had unexpected characteristics, such as higher energy levels. This indicates that while the Nutriscore system provides a quick assessment of nutritional quality, it may not capture all nuances of a product's composition.

Consumer Awareness: Our analysis underscores the importance of consumer awareness and education regarding nutritional labeling systems like Nutriscore. Consumers should understand that while Nutriscore provides valuable information, it should be used in conjunction with a broader understanding of nutrition and dietary choices.

Our analysis highlights the need for continuous refinement and improvement of the Nutriscore system on Sweets. By incorporating feedback from nutritional experts and considering emerging research in nutrition science, the Nutriscore system can evolve to provide more accurate and comprehensive assessments of food products.

Overall, while the Nutriscore system on sweets offers a valuable tool for consumers to make informed food choices, it is essential to recognize its limitations and continue striving for improvements to enhance its effectiveness.