DSApps 2023 @ TAU: Final Project Exploratory Data Analysis

Snacks!

Yoni Slutzky Itay Ofer

2023-08-22

Welcome

Welcome to the exploratory data analysis chapter of our final project. In this chapter, we will try to find some insights from our data, before any actual modelling and formal analysis.

Data preparation

To begin, we will first read the data and clean it up a bit. We will then continue to explore it, trying to find insightful attributes found in the data that may also assist us in the second chapter of the project.

```
food_train <- read_csv("./data/food_train.csv") %>%
  mutate(ingredients = ifelse(is.na(ingredients), "", ingredients))
nutrients <- read_csv("./data/nutrients.csv")
food_nutrients <- read_csv("./data/food_nutrients.csv")
```

We'll combine the datasets into a unified dataframe:

```
# Renaming the categories with shorter names:
food_train <- food_train %>% mutate(category = case_when(
              category == "cookies_biscuits" ~ "cookie",
              category == "cakes_cupcakes_snack_cakes" ~ "cake",
              category == "chips_pretzels_snacks" ~ "savory",
              category == "popcorn_peanuts_seeds_related_snacks" ~ "seeds",
              TRUE ~ category
              ))

# Used full names of nutrients and matched the nutrients found in each product
nutrients$fullname <- paste0(nutrients$name, " in ", nutrients$unit_name)
food_nutrients <- left_join(food_nutrients, select(nutrients, nutrient_id, fullname),
                            by="nutrient_id")
food_nutrients <- pivot_wider(food_nutrients, id_cols="idx", names_from="fullname",
                              values_from="amount")
food_nutrients <- food_nutrients %>%
  mutate(across(everything(), ~ifelse(is.na(.), 0, .)))
```

An important note is that by unifying the two tables, many nutrients (more than 75%) were dropped since they do not appear in any product.

```
# Unified all of the data with the nutrient features
food_train <- left_join(food_train, food_nutrients, by="idx")
```

First Look at The Nutrients

We'll begin our analysis by first examining the top 12 nutrients in across all categories, in terms of average frequency:

```
non_relevant <- c("idx", "brand", "description","ingredients", "serving_size",
                  "serving_size_unit", "household_serving_fulltext")
column_averages <- food_train %>%
  select(-non_relevant) %>%
  summarise(across(where(is.numeric), mean))

top_12_columns <- column_averages %>%
  pivot_longer(everything(), names_to = "column", values_to = "average") %>%
  arrange(desc(average)) %>%
  slice_head(n = 12) %>%
  pull(column)

top_12_data <- food_train %>%
  select(-non_relevant) %>%
  select(category, all_of(top_12_columns))

top_12_data_long <- pivot_longer(top_12_data, -category, names_to = "column")

averages <- top_12_data_long %>%
  group_by(column, category) %>%
  summarise(average = mean(value, na.rm = TRUE))
```

```
ggplot(averages, aes(x = column, y = average, fill = category)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Nutrients", y = "Average Values", fill = "Categories") +
  theme_minimal() +
  facet_wrap(~column, scales = "free", ncol = 3) +
  ggtitle("Average Values of Each Category in Top 12 Nutrients") +
  theme(strip.text = element_blank())
```

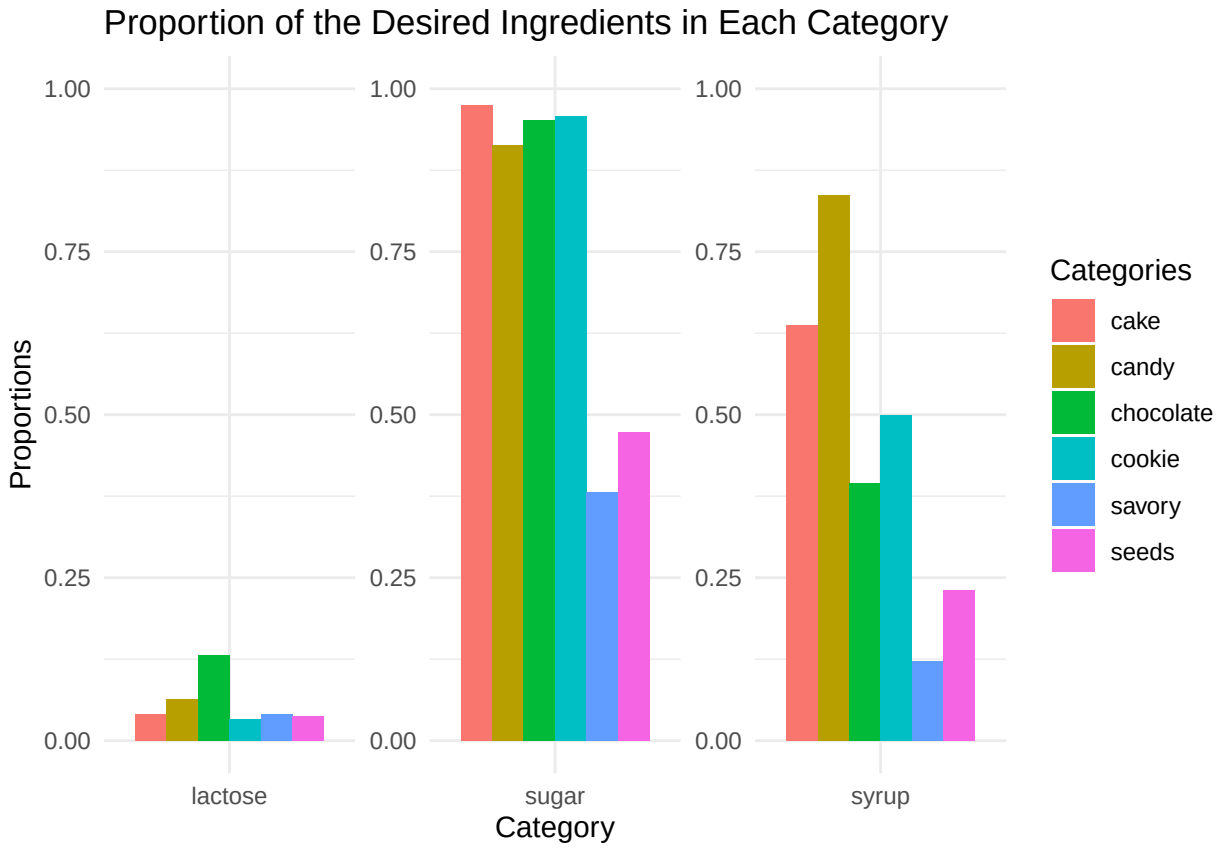
We can clearly see that the top nutrients are calories, Sodium, Carbohydrates, fats and more non-friendly nutrients.
Interestingly, savory snacks contain relatively a lot of Vitamin A and Potassium, which may point out to them being preferable for individuals who require these nutrients. They also contain relatively little sugar and little fatty acids when compared to the other categories. On the other hand, they contain a lot of Sodium - which should be carefully consumed.
Another interesting fact we can see is that seed related items contain relatively large amounts of calories and fats, but also a large amount of protein. Thus, this type of snack may benefit those who want to gain weight or replenish energy after a workout.

Helping Efraim

Meet Efraim, a snack addict who also (tragically) suffers from Diabetes. Efraim wants to know which type of snacks he's better off without. We'll help Efraim by continuing our analysis and expanding it not only to the nutrients themselves, but also to the ingredients of the snacks.\ We will first examine the probabilities for the presence of ingredients which are harmful for individuals suffering from Diabetes:

```
ingredient_presence <- function(ingredient) {
  props <- food_train %>%
  mutate(presence = ifelse(str_detect(ingredients, ingredient), 1, 0)) %>%
  group_by(category) %>%
  summarise(prop = mean(presence)) %>%
  return (props)
}

proportion_plots <- function(ingredients){
  probs <- data.frame()
  for (ingredient in ingredients) {
```

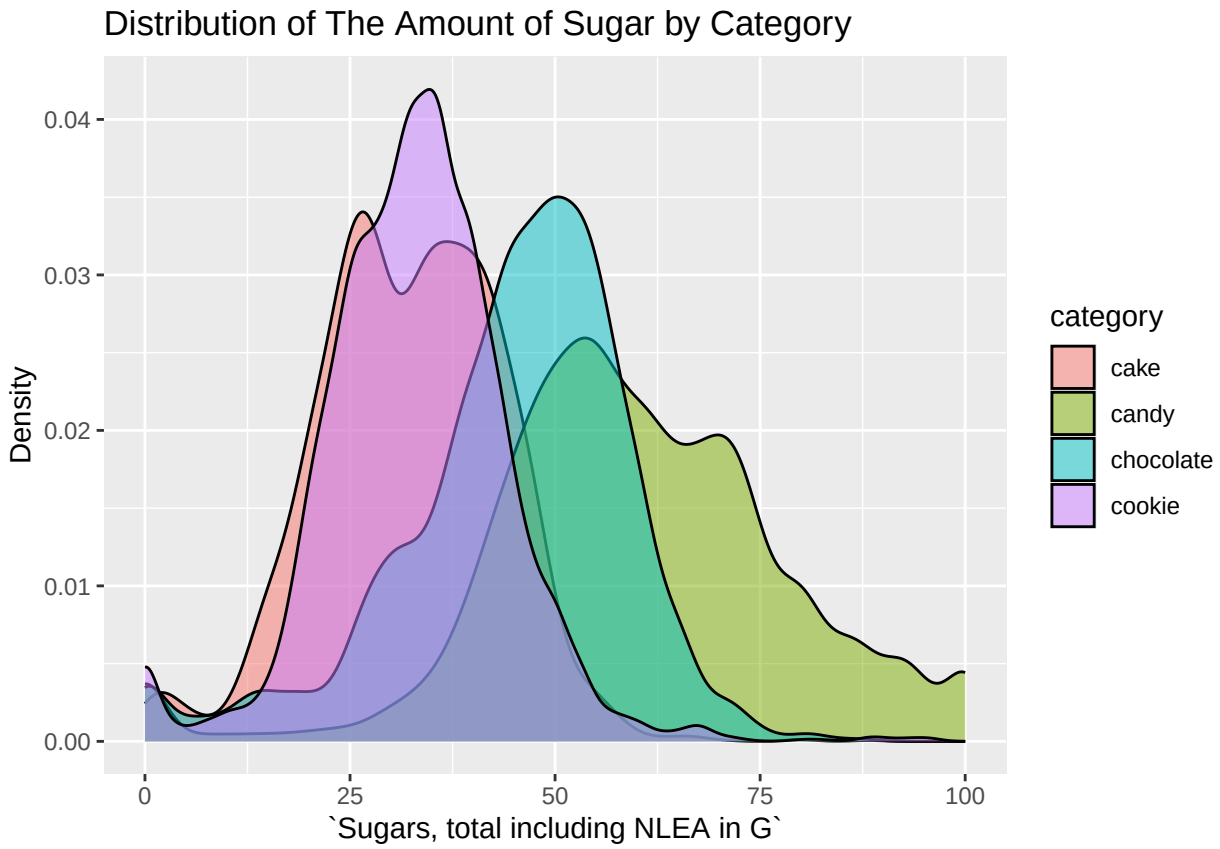
```
    probs <- rbind(probs, cbind(ingredient=ingredient,
                                ingredient_presence(ingredient)))
  }

  ggplot(probs, aes(x = ingredient, y = prop, fill = category)) +
    geom_bar(stat = "identity", position = "dodge") +
    labs(x = "Category", y = "Proportions", fill = "Categories") +
    coord_cartesian(ylim = c(0, 1)) +
    theme_minimal() +
    facet_wrap(~ingredient, scales = "free", ncol = 3) +
    ggtitle("Proportion of the Desired Ingredients in Each Category") +
    theme(strip.text = element_blank())
}

ingredients <- c("sugar", "syrup", "lactose")
proportion_plots(ingredients)
```



Not surprisingly, savory items and seeds are found at the bottom in terms of presence of non-diabetes friendly ingredients, while sweet items are drastically higher. An important note is that the ingredients we checked can be easily changed for Efraim's preference (if another ingredient is to be avoided).
We'll continue our analysis by examining the distribution of the sugar nutrient in each of the sweet categories (If Efraim has a sweet-tooth):

```
ggplot(food_train %>% select(`Sugars, total including NLEA in G`, category) %>%
         filter(!(category %in% c("savory", "seeds"))),
       aes(x = `Sugars, total including NLEA in G`, fill = category)) +
  geom_density(alpha = 0.5) +
```

```
  labs(x = "`Sugars, total including NLEA in G`", y = "Density") +
  ggtitle("Distribution of The Amount of Sugar by Category")
```



From the plot above we can conclude that when in a rush for something sweet, Efraim should resort to cookies and cakes instead of chocolate and candies. This also resonates with the former plot, in which cookies and cakes reach relatively lower results in the syrup and lactose presence.

Healthy Snacks - Is That So?!

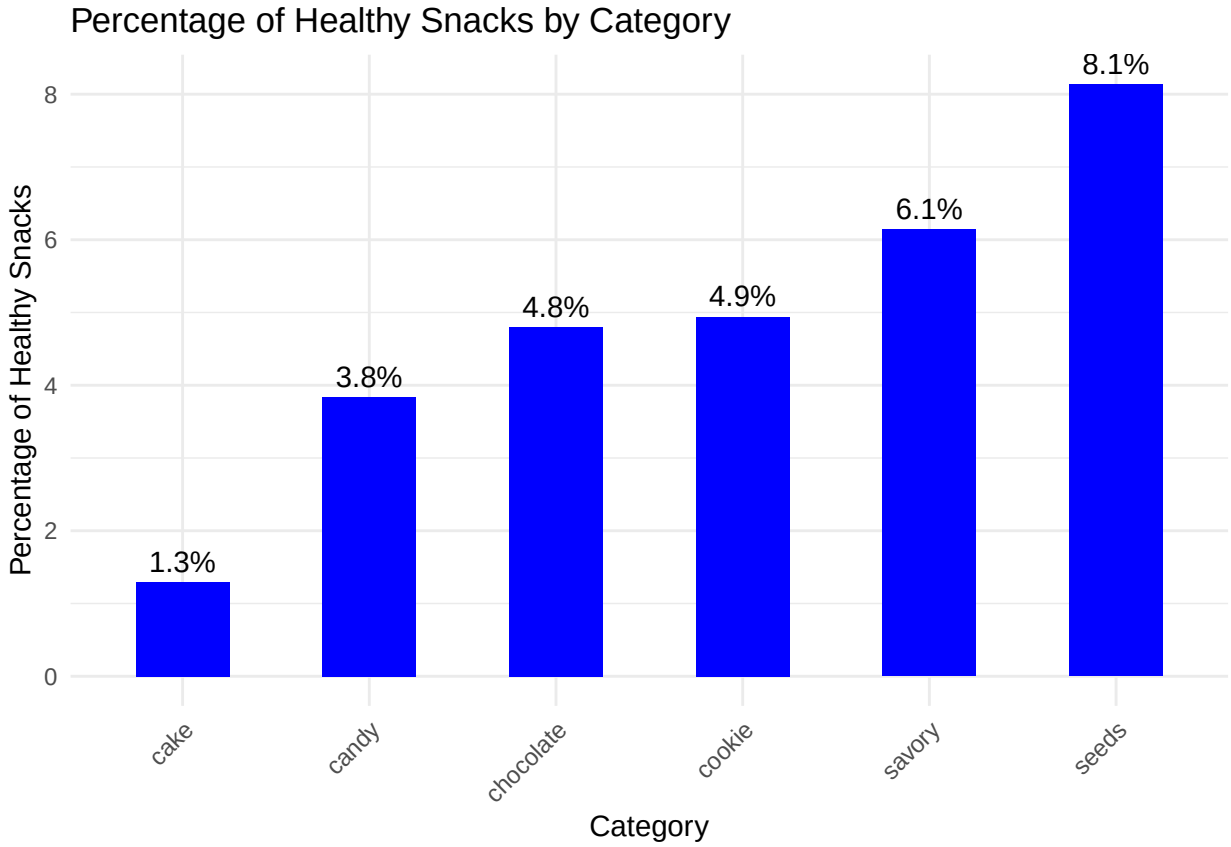
Introducing Osnat - a mother of three who is a health junkie and wants to make sure her children only consume healthy snacks. Her husband suggested her to only buy snacks which are branded as healthy. However, Osnat heard from a fellow gym member that these types of snacks might be worse than regular ones. She asked for our help to give her insight on which snacks to buy.

We will partition the snacks to "healthy" and "non-healthy" using some common-sense keywords (which can be modified to Osnat's liking) found in the descriptions, and then try to gain insights when comparing snacks from the same categories in the two groups.

```
commonsense <- c("free", " low", "natural", "calorie", "healthy", "organic")
food_train <- food_train %>%
  mutate(healthy = ifelse(grepl(paste(commonsense, collapse = "|"), description),
                          "Healthy", "Non-healthy"))
```

We'll first analyze the frequency of healthy snacks in each of the categories. This will help Osnat by pointing to which snack category she should first look into when buying snacks for her kids:

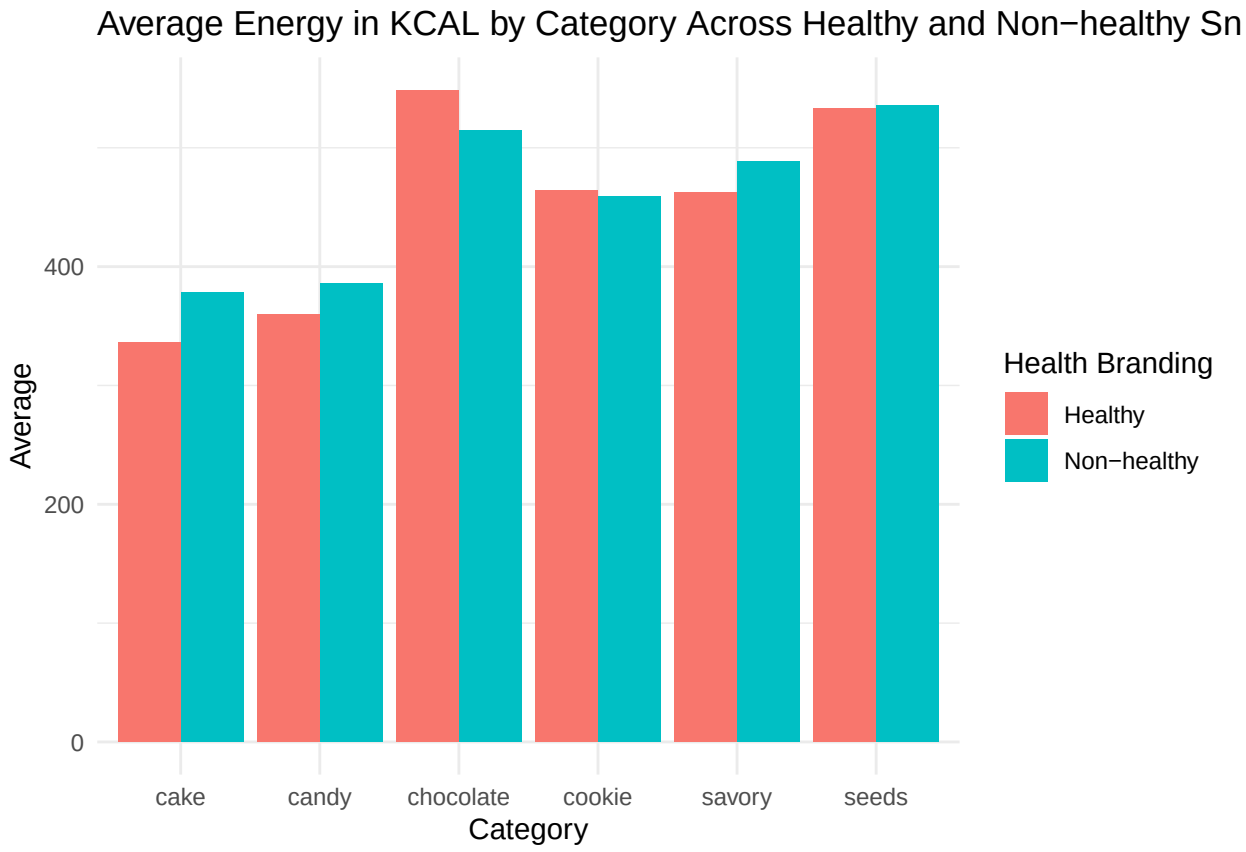
```r
percentage_data <- food_train %>%
  group_by(category) %>%
  summarize(healthy_percentage = sum(ifelse(healthy == "Healthy", 1, 0)) / n() * 100)

ggplot(percentage_data, aes(x = category, y = healthy_percentage)) +
  geom_bar(stat = "identity", fill = "blue", width = 0.5) +
  geom_text(aes(label = sprintf("%.1f%%", healthy_percentage)), vjust = -0.5) +
  labs(x = "Category", y = "Percentage of Healthy Snacks",
       title = "Percentage of Healthy Snacks by Category") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

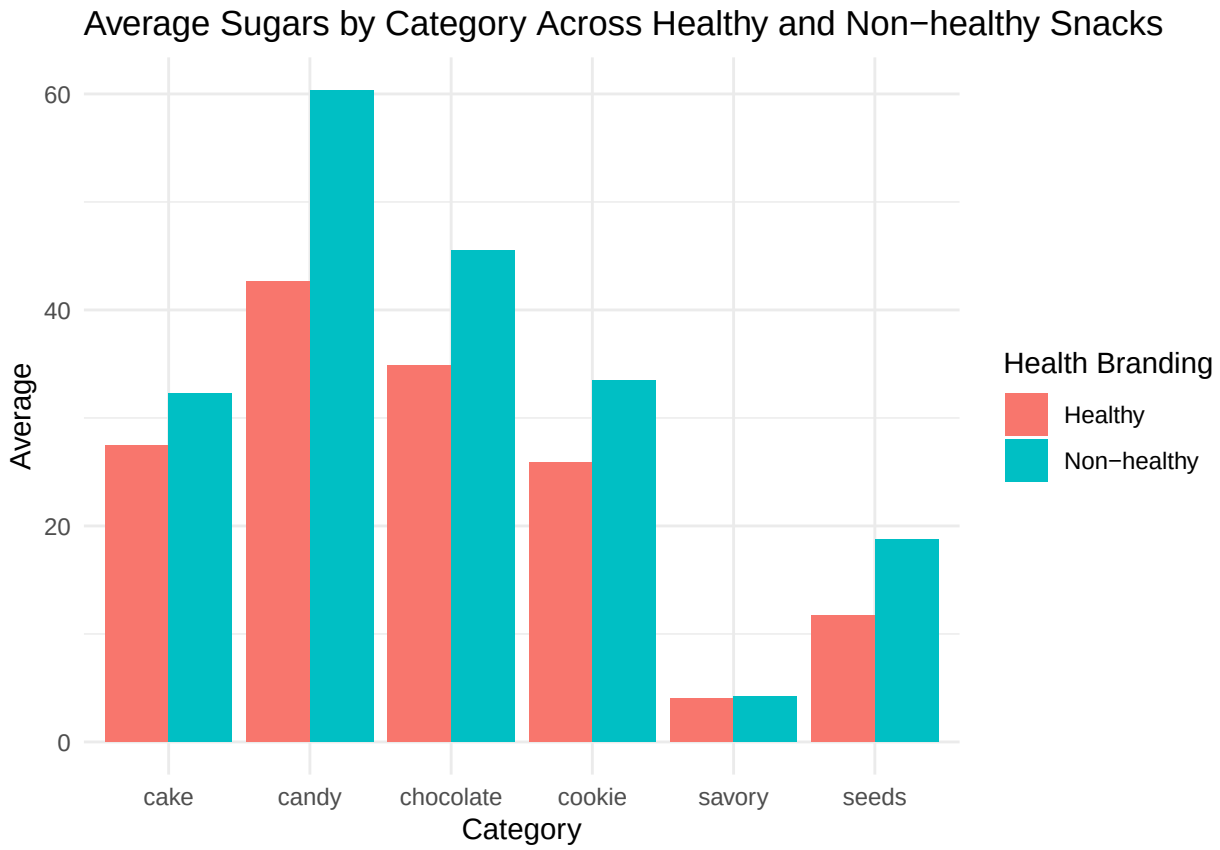


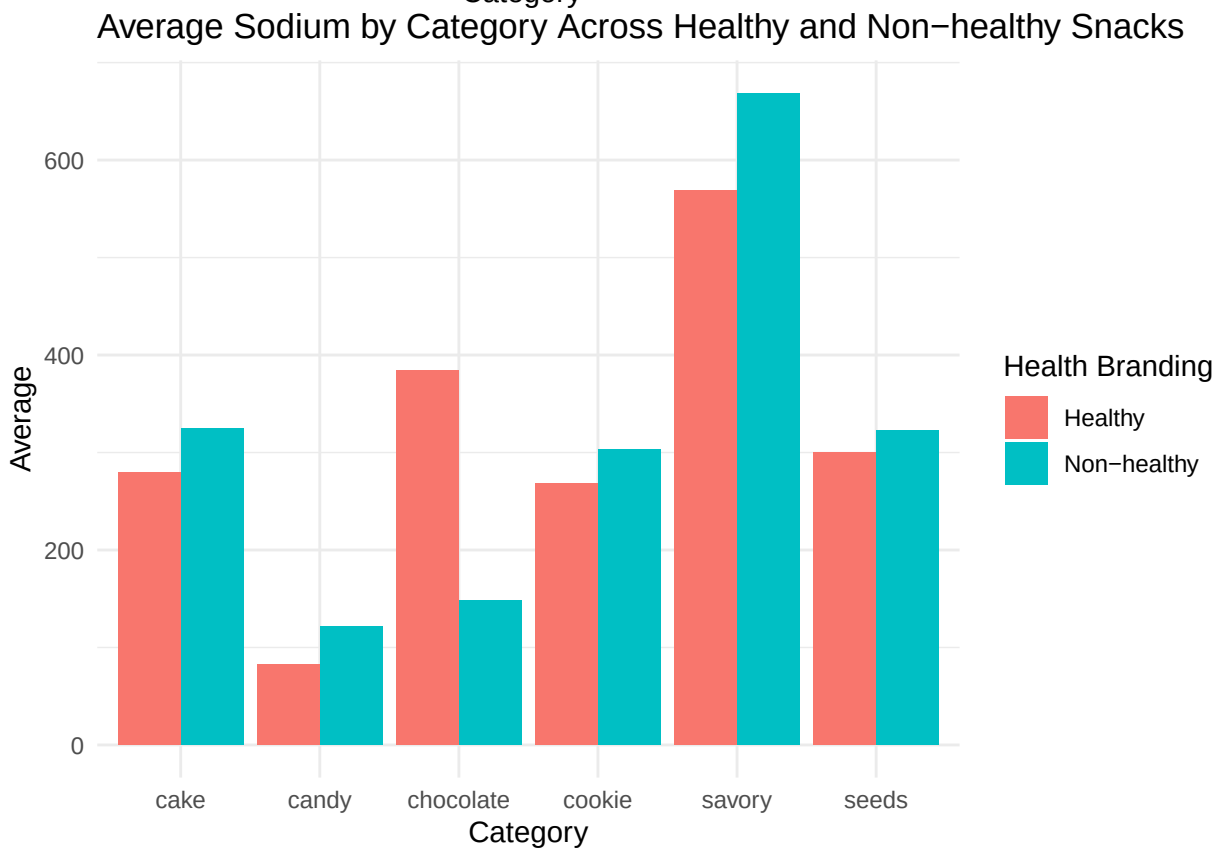
The main conclusion is that most snacks are unhealthy (as expected). Among the snacks categories, the best ones in terms of health considerations are seeds and savory snacks. If Osnat's children insist on getting sweet snacks, the preferable ones are cookies followed by chocolates.
We'll continue to examine some of the nutrients in each category, comparing the two groups:

```r
nutrient_compare <- function(nutrient){
  averages_df <- food_train %>%
    group_by(category, healthy) %>%
    summarise(average = mean(.data[[nutrient]])) %>%
    ungroup()

  plot <- ggplot(averages_df, aes(x = category, y = average, fill = factor(healthy))) +
    geom_bar(stat = "identity", position = "dodge") +
    labs(title = paste("Average", strsplit(nutrient, ",")[[1]],
                       "by Category Across Healthy and Non-healthy Snacks"),
```

```
         x = "Category",
         y = "Average",
         fill = "Health Branding") +
    theme_minimal()
  return (plot)
}

nutrients <- c("Energy in KCAL", "Sugars, total including NLEA in G",
               "Sodium, Na in MG")
for (nutrient in nutrients) {
  print(nutrient_compare(nutrient))
}
```







Inspecting the plots above, we can see the following trends:

1. The average amount of calories is relatively the same in each category, with non-healthy snacks having more calories (as expected), except for the chocolate category where healthy snacks have a small increase relatively to non-healthy ones.
2. The average amount of sugars is quite different between healthy and non-healthy snacks, where healthy snacks have quite an advantage, especially in the candy and chocolate categories.
3. The average amount of Sodium is relatively the same in each category, with non-healthy snacks having more sodium (as expected), except for the chocolate category, where surprisingly healthy snacks have a large increase relatively to non-healthy ones.
The conclusion is that healthy branded chocolate snacks are worse than non-healthy branded ones (especially in terms of Sodium).
Non-healthy snacks from other categories do seem to be worse nutritionally when compared to healthy ones.
The overall conclusion is that Osnat should get seeds and savory items, and if a sweet item is a must, then she should first consider cookies.

Final note

After some thought we decided not to use the image data in this chapter. This is due to the fact that after performing a shallow processing of the images we couldn't reach any meaningful insights, and we preferred to reserve the deeper processing (namely, through the usage of deep learning tools) for the modelling chapter. In the modelling chapter, we deploy a finetuned computer vision model to assist us in utilizing the image data.