EDA Final Project

Dov Tuch

Welcome

Hello Giora, This EDA can be summed up into 2 parts. The first one is just general EDA with no story or purpose behind it. The second part is a market analysis I conducted with the data at hand, with a mission and a little bit of sarcasm :) Hope you enjoy (the second part).

First part: preparing the data set and general EDA.

I will make the 3 data sets to one single tibble. Where the columns are the nutritional values, and the first 10 ingredients in each snack.

```
fst_10ings = str_replace_all(food_train$ingredients, "[().*+-]", "") %>%
str_split_fixed( ',',n = 11) %>% as_tibble() %>%
  select(-V11)
```

```
# removing zero variance columns
zv_cols = nearZeroVar(food_nut, names = T)

food_nut2 = food_nut %>%
  select(-all_of(zv_cols))
food_joined = food_train %>%
  left_join(food_nut2, by = 'idx')
food_complete = bind_cols(food_joined,fst_10ings )
```

```
## New names:
## * `...1` -> `...9`
```

```
food_train %>%
  distinct(serving_size_unit)
```

```
## # A tibble: 2 x 1
##   serving_size_unit
##   <chr>
## 1 g
## 2 ml
```

```
food_complete %>%
  filter(serving_size_unit == 'ml') %>%
  nrow()
```

```
## [1] 8
```

Because we have only 8 observations that are in ml units (out of ~ 32K), I will drop the *serving_size_unit* column and those observations. The data set that I will be working with will look like where v_i is the i-th appearing ingredient in the ingredients column:

```
## Rows: 31,743
```

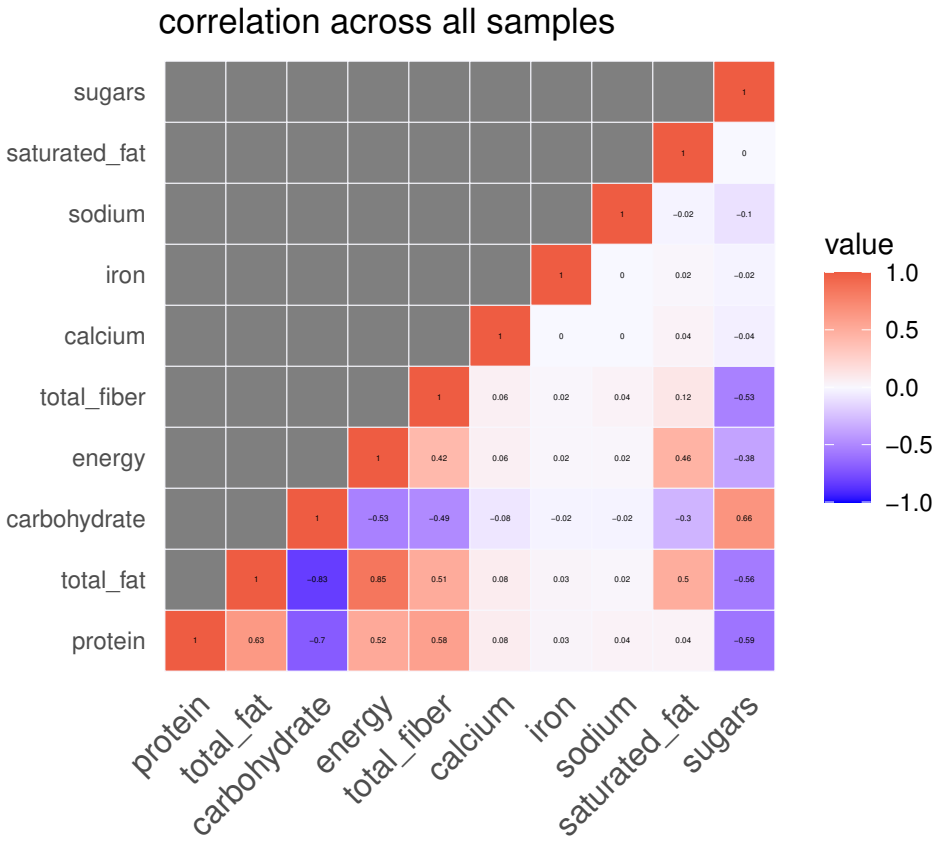
```
## Columns: 26
## $ idx                          <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9"~
## $ brand                        <chr> "brix chocolate", "target stores", "target ~
## $ description                  <chr> "milk chocolate", "frosted sugar cookies", ~
## $ serving_size                 <dbl> 28.0, 38.0, 30.0, 40.0, 40.0, 36.0, 28.0, 2~
## $ household_serving_fulltext <chr> "1 onz", "1 cookie", "2 cookies", "5 pieces~
## $ category                     <chr> "chocolate", "cookies_biscuits", "cookies_b~
## $ protein                      <dbl> 7.14, 2.63, 3.33, 5.00, 7.50, 5.56, 3.57, 7~
## $ total_fat                    <dbl> 35.71, 15.79, 15.00, 22.50, 42.50, 13.89, 2~
## $ carbohydrate                 <dbl> 53.57, 68.42, 70.00, 67.50, 47.50, 66.67, 5~
## $ energy                       <dbl> 536, 421, 433, 475, 600, 417, 429, 423, 297~
## $ total_fiber                  <dbl> 3.6, 0.0, 0.0, 2.5, 5.0, 2.8, 0.0, 10.6, 1.~
## $ calcium                      <dbl> 143, 0, 0, 100, 100, 0, 0, 70, 31, 0, 0, 0,~
## $ iron                         <dbl> 5.14, 1.89, 1.20, 1.80, 1.80, 3.00, 2.57, 2~
## $ sodium                       <dbl> 89, 276, 317, 312, 88, 278, 232, 317, 719, ~
## $ saturated_fat                <dbl> 25.00, 6.58, 6.67, 17.50, 17.50, 5.56, 8.93~
## $ sugars                       <dbl> 42.86, 42.11, 43.33, 47.50, 40.00, 33.33, 3~
## $ v1                           <chr> "sugar", "sugar", "sugar", "sugar", "semisw~
## $ v2                           <chr> " cocoa butter", " enriched bleached wheat ~
## $ v3                           <chr> " whole milk", " niacin", " niacin", " part~
## $ v4                           <chr> " chocolate liquor", " reduced iron", " red~
## $ v5                           <chr> " chocolate liquor processed with alkali", ~
## $ v6                           <chr> " soy lecithin emulsifier", " riboflavin", ~
## $ v7                           <chr> " vanilla", " folic acid", " folic acid", "~
## $ v8                           <chr> " and natural flavors", " margarine palm oi~
## $ v9                           <chr> NA, " water", " water", " skim milk", " haz~
## $ v10                          <chr> NA, " soybean oil", " soybean oil", " whey ~
```

Correlations between the nutrients across all samples and by categories.

```
# helper functions
# Get lower diagonal of the correlation matrix
get_lower_diag<-function(cormat){
    cormat[lower.tri(cormat)] <- NA
    return(cormat)
  }
reorder_cormat <- function(cormat){
# Use correlation between variables as distance
dd <- as.dist((1-cormat)/2)
hc <- hclust(dd)
cormat <-cormat[hc$order, hc$order]
}

# correltaion function heatmap
corr_mat = function(category_data){
  var_name = paste0(category_data)
  cor_nut = food_final %>%
  filter(category == var_name) %>%
  select(protein:sugars) %>%
  cor() %>%
  round(2)
  titleplot = str_replace_all(var_name, '_', " ")

  lower_diag= get_lower_diag(cor_nut)
```

```
  melted_cormat= reshape2::melt(lower_diag)
  heatmap_fun = ggplot(data = melted_cormat, aes(Var2, Var1, fill = value))+
   geom_tile(color = "ghostwhite")+
   scale_fill_gradient2(low = "blue1", high = "tomato2", mid = "ghostwhite",
     midpoint = 0, limit = c(-1,1), space = "Lab",
     ) +
    ggtitle(titleplot)+
    theme_minimal()+
   # theme(axis.text.x = element_text(angle = 45, vjust = 1,
   #    size = 12, hjust = 1))+
   coord_fixed()+
    geom_text(aes(Var2, Var1, label = value), color = "black", size = 2)+
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    axis.text.x = element_text(angle = 45, vjust = 1,
    size = 12, hjust = 1))

  return(heatmap_fun)

}
```

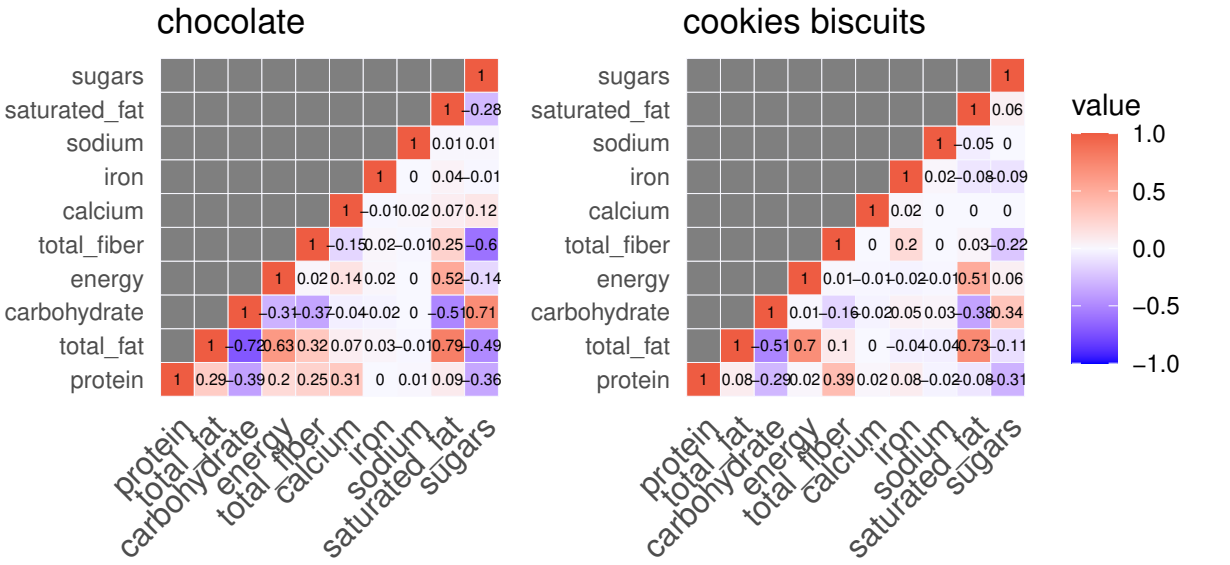
```
## Warning: Removed 45 rows containing missing values (geom_text).
```



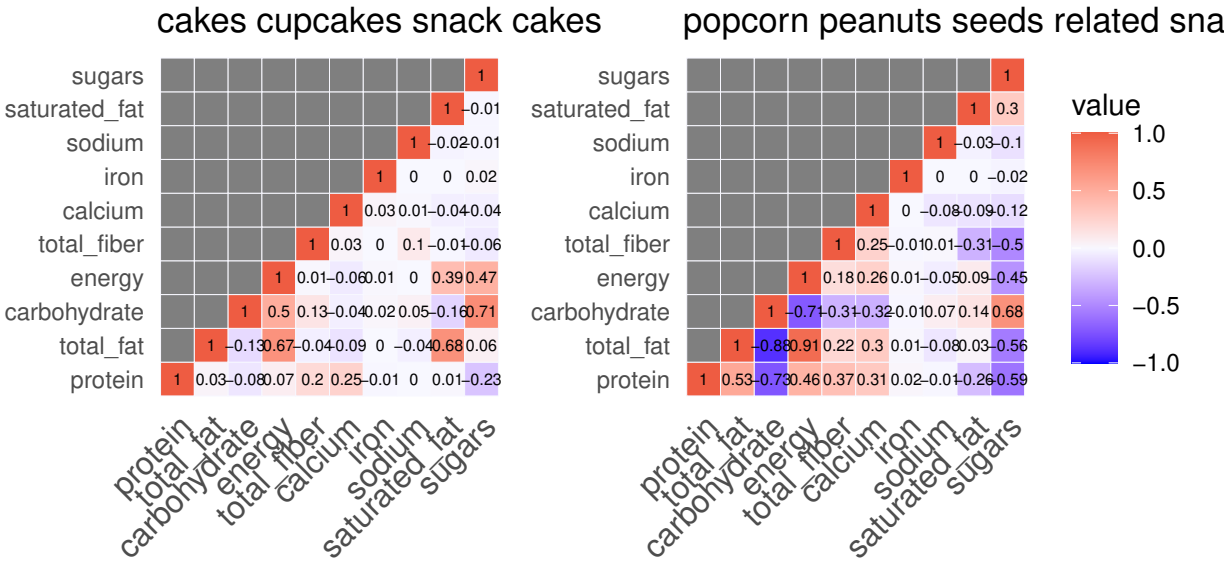
For all the snacks. total fat and protein are strongly inversely correlated. With carbohydrates. While energy and total fat are highly correlated.

```
## Warning: Removed 45 rows containing missing values (geom_text).
## Removed 45 rows containing missing values (geom_text).
```

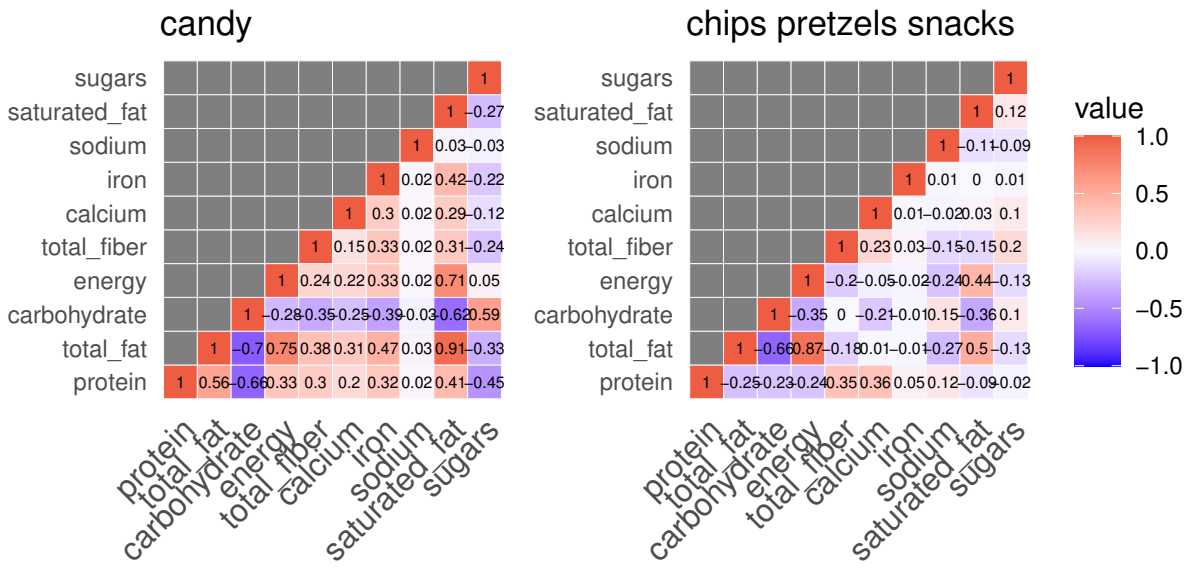
The corelation of nutrients by category




```
## Warning: Removed 45 rows containing missing values (geom_text).
## Removed 45 rows containing missing values (geom_text).
```




```
## Warning: Removed 45 rows containing missing values (geom_text).
## Removed 45 rows containing missing values (geom_text).
```

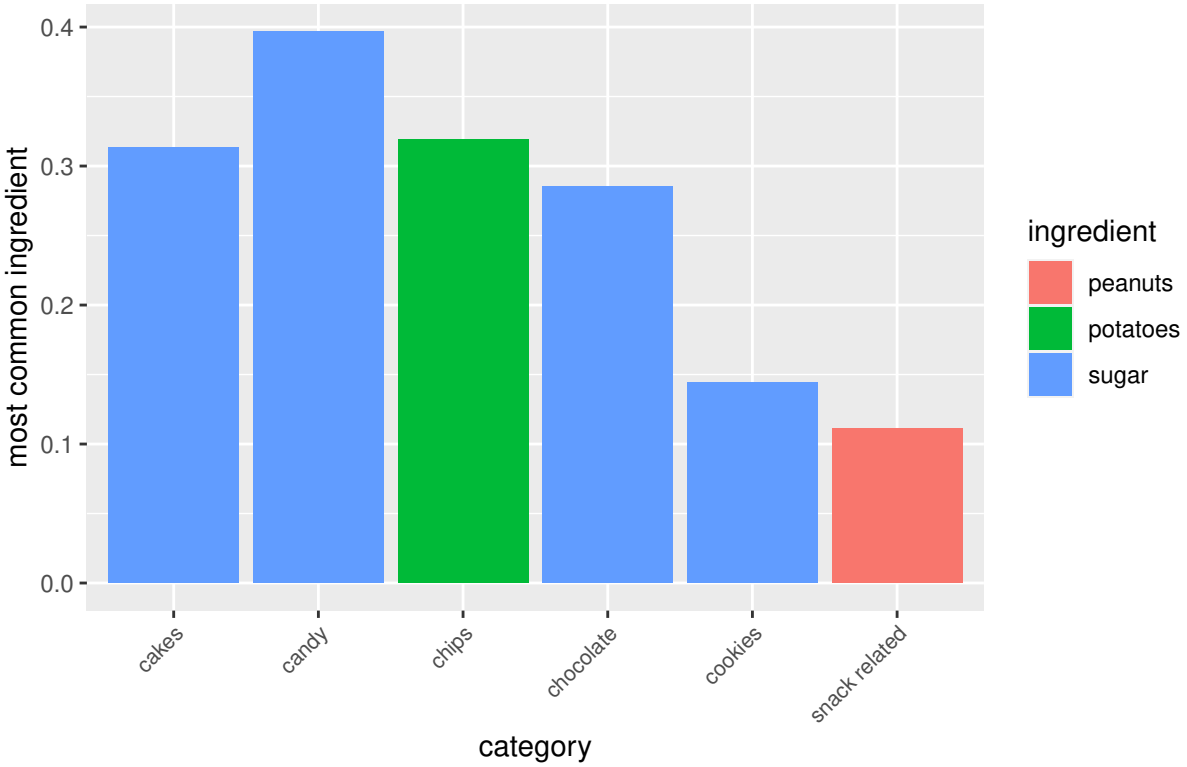


When we look at the nutrient correlation by category we find different results. Candies total fat is very strongly correlated with saturated fat, but for peanuts There is no correlation between the two nutritional values. In the cakes category carbohydrates are strongly correlated with sugars but not with chips. This realtionships can be used later as predictors for each category.

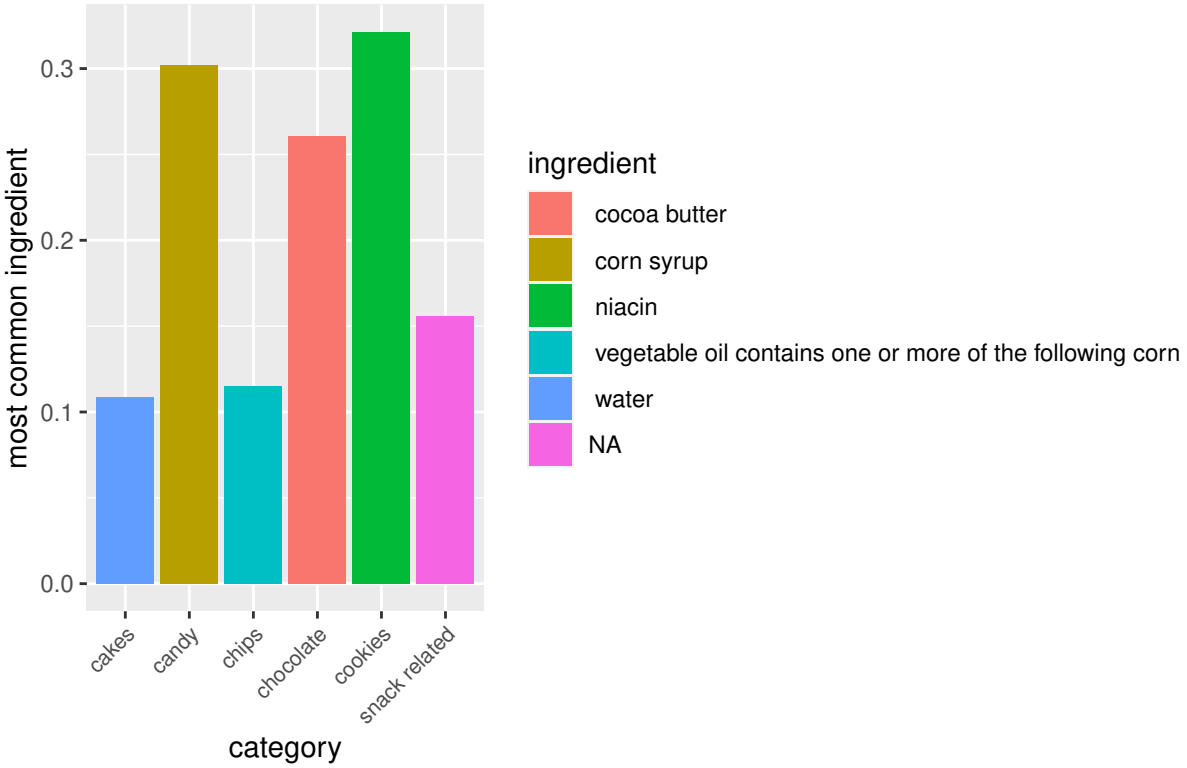
```
table_0 = table_0 %>%
  mutate(category =
  case_when(category == "cakes_cupcakes_snack_cakes" ~ 'cakes',
            category =="chips_pretzels_snacks" ~ 'chips',
            category == 'cookies_biscuits' ~ 'cookies',
            category == 'popcorn_peanuts_seeds_related_snacks' ~ 'snack related',
            TRUE ~ as.character(category))
  )
```

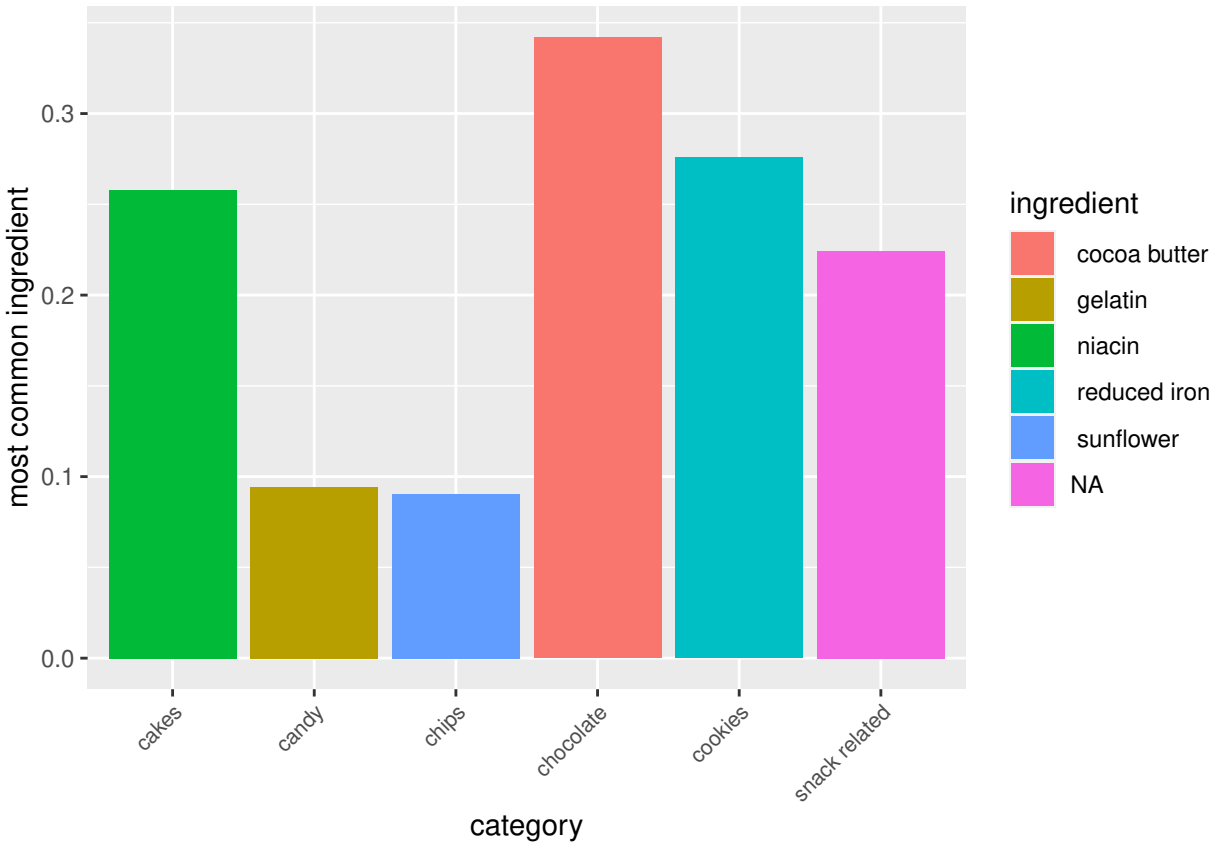
Let's look at the most occurring ingredients by category.

The proprtion of the top most used ingredient by category



The proprtion of the top 2nd most used ingredient by category





We can get a lot of predicting power with those values, when potatoes are the first ingredient we can have a good guess that it's from the chips category. It's not a big surprise that peanuts and seeds have mostly one ingredient, that's why the 2nd and 3rd most common "ingredient" is NA.

Second Part: Market Analysis!

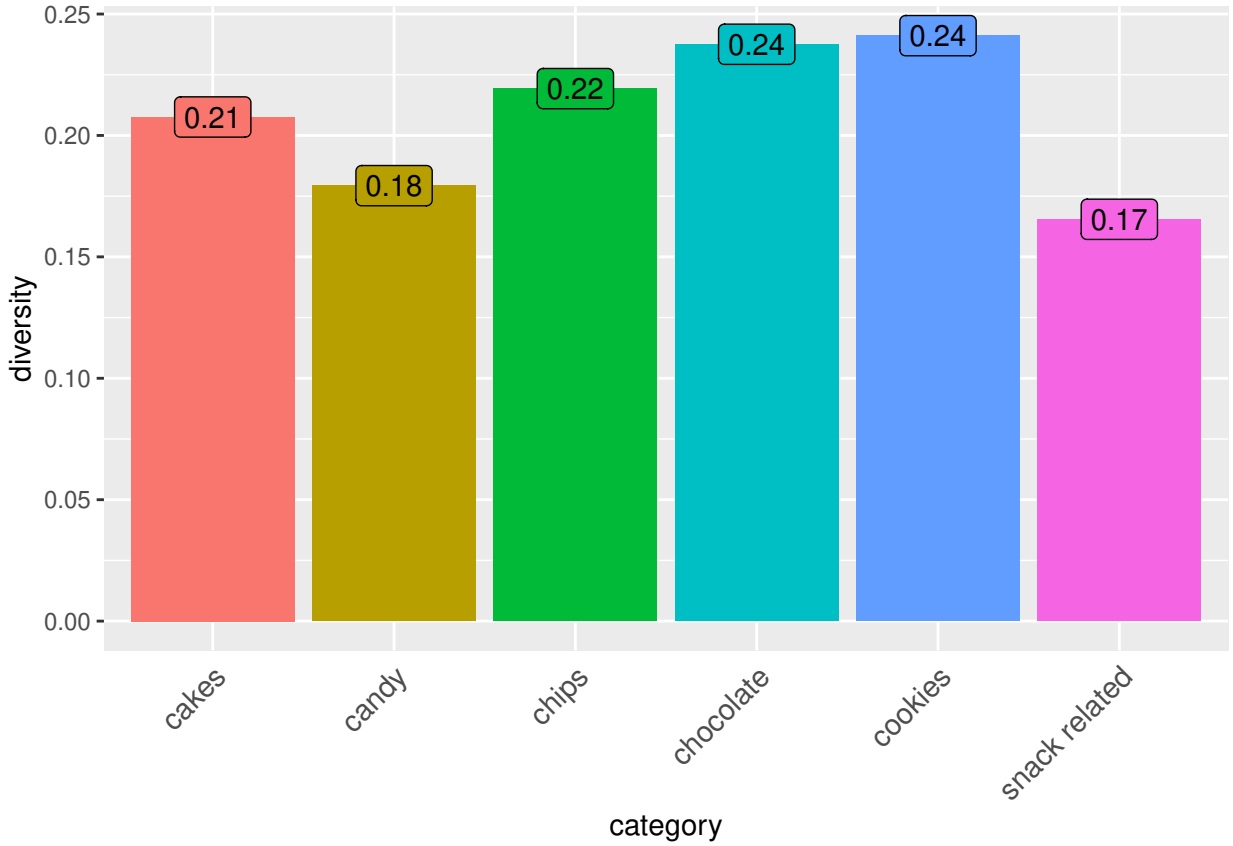
During this EDA my mother started to nag me that I don't do anything with my Econ major, just statistics all day. So to make mummy proud I decided to go from general EDA to a market research!

First question, where is the competition?

To answer this we need to understand how many brands are in each category? To measure this we will look at the ratio of n_brands/n_products in the category. We will call this measure as the *Diversity*. so if the *Diversity* is high there are many brands in this market and the highest value it can get is 1. When there is low market diversity it means that there are less players in the market and so the diversity will be close to 0.

```
## # A tibble: 6 x 3
##   category                                n num_brands
##   <chr>                               <int>      <int>
## 1 cakes_cupcakes_snack_cakes           3786        786
## 2 candy                                7584       1360
## 3 chips_pretzels_snacks                3680        807
## 4 chocolate                            3772        896
## 5 cookies_biscuits                     5284       1274
## 6 popcorn_peanuts_seeds_related_snacks 7645       1265
```

```
table1 %>%
  ggplot(mapping = aes(x = category, y = diversity, fill = category))+
  geom_col()+
  theme(
        axis.ticks.x=element_blank(),
        legend.position  = "none"
        )+
  theme(
    axis.text.x = element_text(angle = 45, vjust = 1,
    size = 11, hjust = 1)
   )+
  #geom_text(aes( label = paste0("n = ", total_snacks)), nudge_y = +0.02)+
  geom_label(aes(label = round(diversity, 2)))
```



We can see that popcorn peanuts and snacks have the lowest diversity. But it also the most common category in the training sample. As always, this breaks some economical models assumptions. As we saw before a large parcentege of those products have only one ingredient (nuts or seeds). It's a generic product so we expect it to have more sellers. Another possible explanation is that most of these sellers operate on a small business scale, So They don't need an FDA registration to sell their nuts.

This leads as to the second question..

Question2: Why there is less competition in the popcorn market?

Let's dig further into popcorn category, I will look at the brands and look if there are any significant players.

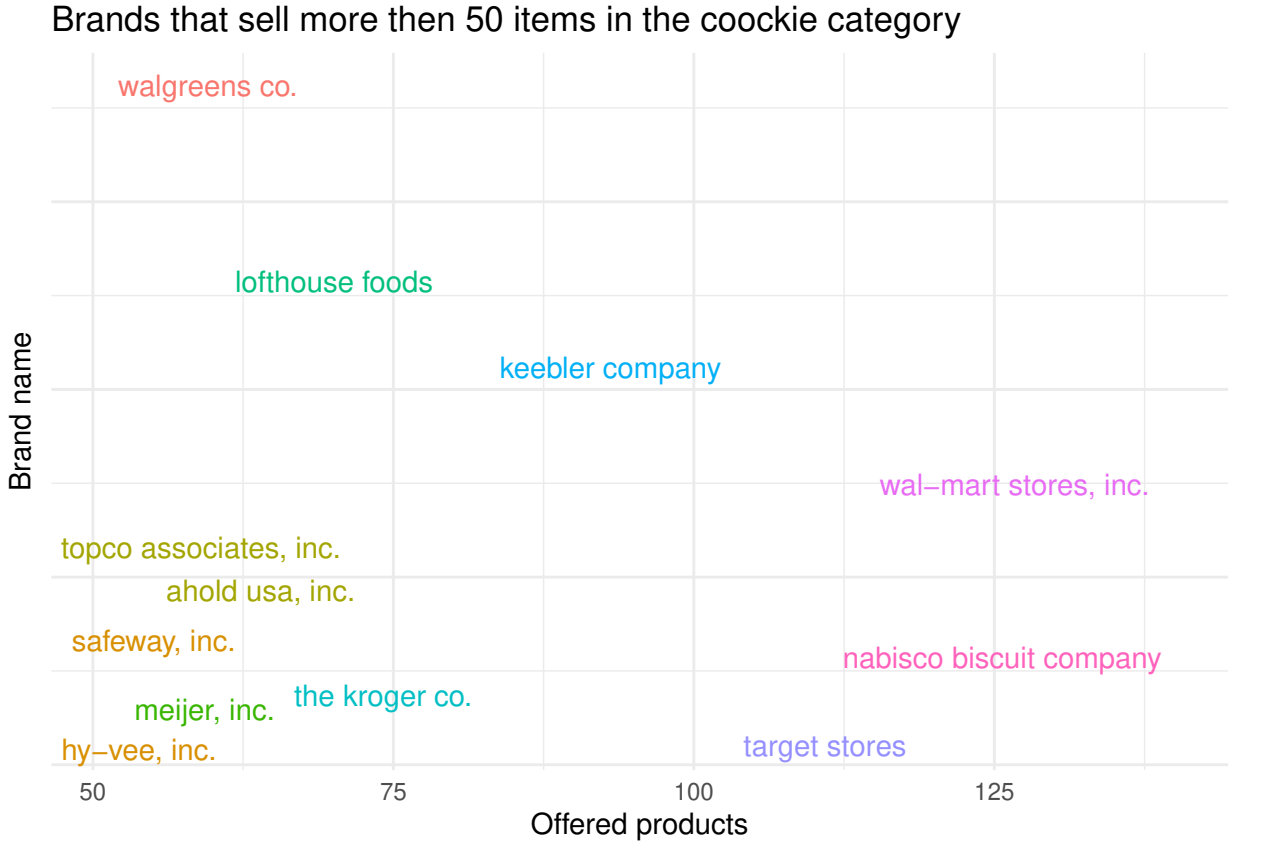
```
set.seed(12323)
food_train %>%
  filter(category == "popcorn_peanuts_seeds_related_snacks" ) %>%
  count(brand, sort = TRUE) %>%
  mutate(y_random = c(runif(1265, min = -1, max = 1))) %>%
  filter(n > 50) %>%
  ggplot(aes(x = n, y = y_random, label = brand, color = as.factor(n))) +
  ggtitle('Brands that sell more then 50 items in the seed category')+
  geom_text_repel(max.overlaps = 1265) +
  theme_minimal() +
  theme(legend.position = "None",
        axis.ticks.y = element_blank(),
        axis.text.y = element_blank()) +
  labs(x = "Offered products",
       y = "Brand name")
```

Looking at the biggest sellers here (meijer and target) we can understand that the biggest players are not food focused companies but big retailers. with much less products pharmacies like cvs and walgreens are also big players. only nabisco is a company that focuses mainly on packaged foods.

So I saw that retail stores and pharmacies of the branded snacks in the seeds category. Let's see what is the situation in a category with more *diversity*. for example the cookies category.

```
set.seed(12323)
food_train %>%
  filter(category == "cookies_biscuits" ) %>%
  count(brand, sort = TRUE) %>%
  mutate(y_random = c(runif(1274, min = -1, max = 1))) %>%
```

```
  filter(n > 50) %>%
  ggplot(aes(x = n, y = y_random, label = brand, color = as.factor(n))) +
  ggtitle('Brands that sell more then 50 items in the coockie category')+
  geom_text_repel(max.overlaps = 1274) +
  theme_minimal() +
  theme(legend.position = "None",
        axis.ticks.y = element_blank(),
        axis.text.y = element_blank()) +
  labs(x = "Offered products",
       y = "Brand name")
```



In the cookie market only two out of the largest 5 selling brands are not food focused. Nabisco biscuit company lives up to it's name :). To support the effectiveness I gave of measuring the competition we can see here that there are less brands that sell more then 50 products. (12 in this category and 29 in the seed category) although there are slightly more brands in this category then the last one (1274 and 1265)/

the only category where the top 5 sellers are food focused companies is the chocolate category. To be in the chocolate business you really need to focus.

Question 3: how is nabisco different?

let's analyze nabisco in the cookie category,how are they're product differ from the the others?

I will conduct a paired t test for each nutritional value with the hypothesis that the means of nabisco are the same as the avarege cookie & biscuit company.

```
food_final %>%
  filter(category == "cookies_biscuits") %>%
```

```
  select(-c(idx, category)) %>%
  summarise(across(where(is.numeric), list(avg = mean, sd = sd))) %>%
  bind_rows(food_final %>%
  filter(category == "cookies_biscuits") %>%
  filter(brand == "nabisco biscuit company") %>%
  select(-c(idx, category)) %>%
  summarise(across(where(is.numeric), list(avg = mean, sd = sd))))
```

```
## # A tibble: 2 x 22
##   serving_size_avg serving_size_sd protein_avg protein_sd total_fat_avg
##              <dbl>           <dbl>       <dbl>      <dbl>         <dbl>
## 1             33.0            13.1        5.44       2.97          20.1
## 2             31.7             8.99       5.37       2.86          18.4
## # ... with 17 more variables: total_fat_sd <dbl>, carbohydrate_avg <dbl>,
## #   carbohydrate_sd <dbl>, energy_avg <dbl>, energy_sd <dbl>,
## #   total_fiber_avg <dbl>, total_fiber_sd <dbl>, calcium_avg <dbl>,
## #   calcium_sd <dbl>, iron_avg <dbl>, iron_sd <dbl>, sodium_avg <dbl>,
## #   sodium_sd <dbl>, saturated_fat_avg <dbl>, saturated_fat_sd <dbl>,
## #   sugars_avg <dbl>, sugars_sd <dbl>
```

```
cookie_avg = food_final %>%
  filter(category == "cookies_biscuits") %>%
  select(-c(idx, category)) %>%
  summarise(across(where(is.numeric), list(avg = mean))) %>%
  slice(1) %>%
  as.numeric()

cookie_sd =food_final %>%
  filter(category == "cookies_biscuits") %>%
  select(-c(idx, category)) %>%
  summarise(across(where(is.numeric), list(sd = sd))) %>%
  slice(1) %>%
  as.numeric()

nobosco_avgs = food_final %>%
  filter(category == "cookies_biscuits") %>%
  filter(brand == "nabisco biscuit company") %>%
  select(-c(idx, category)) %>%
  summarise(across(where(is.numeric), list(avg = mean))) %>%
  slice(1) %>%
  as.numeric()

nobosco_sd = food_final %>%
  filter(category == "cookies_biscuits") %>%
  filter(brand == "nabisco biscuit company") %>%
  select(-c(idx, category)) %>%
  summarise(across(where(is.numeric), list(sd = sd))) %>%
  slice(1) %>%
  as.numeric()
```

```
all_ttest = function(nb_avg, nb_sd, all_avg, all_sd,n1 ,n2){
  m = length(nb_avg)
  t = numeric(m)
  for (i in 1:m){
```

```
    sp = sqrt(((n1- 1)*(nb_sd[i])^2 + (n2-1)*(all_sd[i])^2)/(n1+n2 -2))
    t[i] = abs((nb_avg[i]-all_avg[i])/(sp*sqrt(1/n1+1/n2))) > qt(0.975,n1 + n2- 2)
  }
  return(t)
}



b = all_ttest(nobosco_avgs,nobosco_sd,cookie_avg,cookie_sd, 140,5284)
```

From this test I can search for nad reject the null hypothesis with 0.05 significance that carbohydrate, energy, calcium, sodium and saturated fats are different in the nobisco brand from the average cookie/biscuit product.

And That's It. Good luck with the rest of your life!

Thank you for all the mind blows I had in this semester in your course.

Statement:

Please add this statement to the pdf you are submitting, with your full name, ID number and signature:

I confirm that the work for the following project in the Applications of Data Science course was solely undertaken by myself and that no help was provided from other sources as those allowed.

Name: Dov Tuch, ID: 207049719, Date: 08/08/22,
Signature: