Multivariate Analysis Project

Open Food Facts Dataset

Cristina Capdevilla Choy

Roberto Ruiz Wilson

Ana Giraldo Botero

1. **Description of dataset and problem**

The selected dataset gathers information related to the composition, characteristics and ingredients of food. The dataset was created in the aim of providing insights on nutrition facts related to food products from all around the world, the set used for this project was obtained from an already built site that seeks to deliver transparency to the consumer when purchasing processed food. In the Food Fact site ([world.openfoodfacts.org/data](https://world.openfoodfacts.org/data)), you can currently find applications to search particular facts on products from different countries and do advanced search based on some classification already done by the creators. We, on a different approach, will be focusing in the nutrition feature of 399,892 products of a selected subset, delivering a model to predict the rating/score of this particular product; so at the end the problem we will try to solve is to establish, accurately, whether a product is nutritional enough or not, according to the different features related to it.

Responsible variable:

* nutriscore\_score: a score given to each product, based on how balanced its macronutrients are. Values a to d

Supplementary variables: 11 features

* countries\_en: country of origin/commercialization of the product. 19 categorical values.
* nutriscore\_grade: categorical/factor. An assigned letter from “a” to “e”.
* additives\_n: number of additives found in nutrition value labels. Numerical value 1 to 24.
* nova\_group: it is a kind of grade of nutrition given to the foods, it is kept as a feature for extra info when grading the nutrition score. Numerical value from 1 to 4.
* pnns\_group\_2: categorical/factor correspond to the type of product, in an already defined list of groups of 6 values:
  + Biscuits and cakes,
  + Cereals,
  + Sweets,
  + Dressing and sauces,
  + Cheese,
  + Other.
* fat\_100g: fat supply of the product for a 100g portion.
* carbohydrates\_100g: carbohydrate supply of the product for a 100g portion.
* sugars\_100g: sugar supply of the product for a 100g portion.
* fiber\_100g: fiber supply of the product for a 100g portion.
* proteins\_100g: protein supply of the product for a 100g portion.
* salt\_100g: salt addition per 100g of the product.

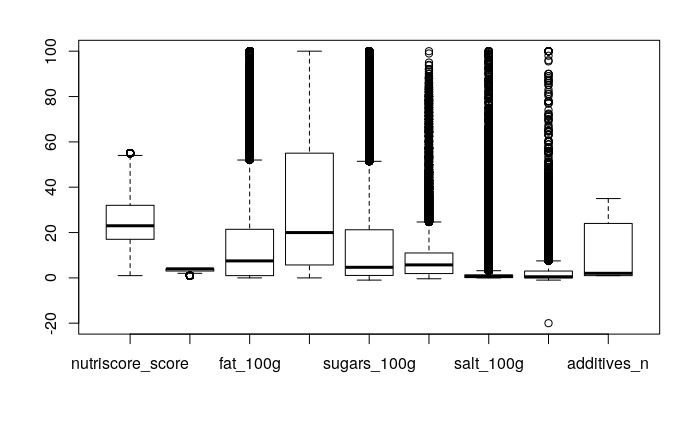
The feature selection is based on the main factors that, in general terms, define the nutritional value of any food: its chemical composition that is divided in the 3 main macronutrients: carbohydrates (fiber and sugar are types of it), proteins and fat. Some of the other features are included in order to analyze behavior of already created categories, classifications and groups.

1. **Data pre-process**

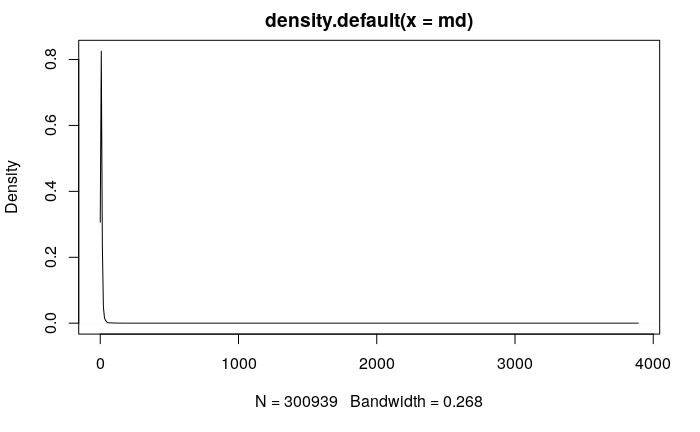
2.1. Data cleaning:

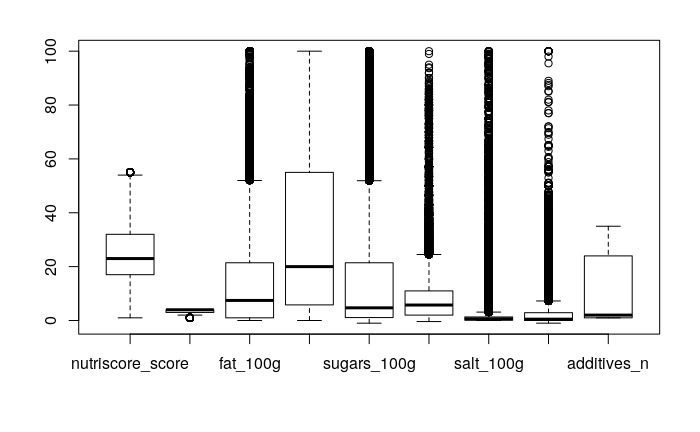
* Transformar " " o NA values no reportados como tal a NA, los valores nutricionales que eran NA cambie a cero, ya que cuando no estan es porqe su valor cero en el producto(como norma general), los valores nutricionales que eran mayor a 1000 estan divididos entre 1000 porqe esto es que estan puestos en mg, y los que quedan mayor a 100 estan quitados porqe no estan bien(los valores de nutri van sobre 100g) pero no se puede saber porqe estan mal
* quitar todas las lineas donde nutriscore es nulo, quitar todas las lineas donde el pais es nulo, dejar solo los productos que vengan de un solo pais y que ese pais tenga mas de
* a partir de los datos originales: quitar todas las lineas donde nutriscore es nulo, quitar todas las lineas donde el pais es nulo, dejar solo los productos que vengan de un solo pais y que ese pais tenga mas de 1,000 productos

2.2. Outliers



Multivariate outlier detection using the Mahalanobis distance can be used. Plot of the classical and the robust (based on the MCD) Mahalanobis distance is drawn.





el cuantil 0.975 es una distancia de 26.77667

1. **Visualization and latent factors detection**

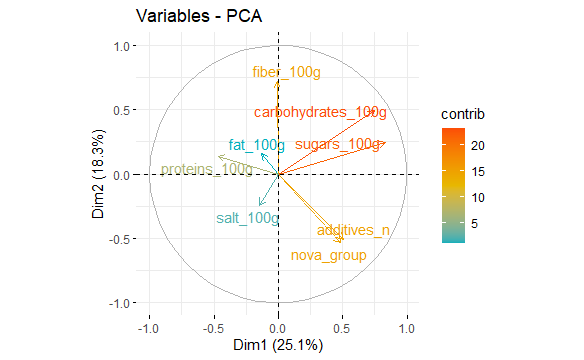
3.1. PCA Analysis

To perform a Principal Component Analysis all categorical variables/factors must be removed from the dataset, in that order, there are 8 numerical variables remaining as explanatory as we have also removed the responsible variable from the data set.

In the first exploratory analysis we have considered all variables to be equally weighted to perform a standardized PCA where the first principal component of our data frame X with X1,X2,X3,....X8 is the linear combination of the features:

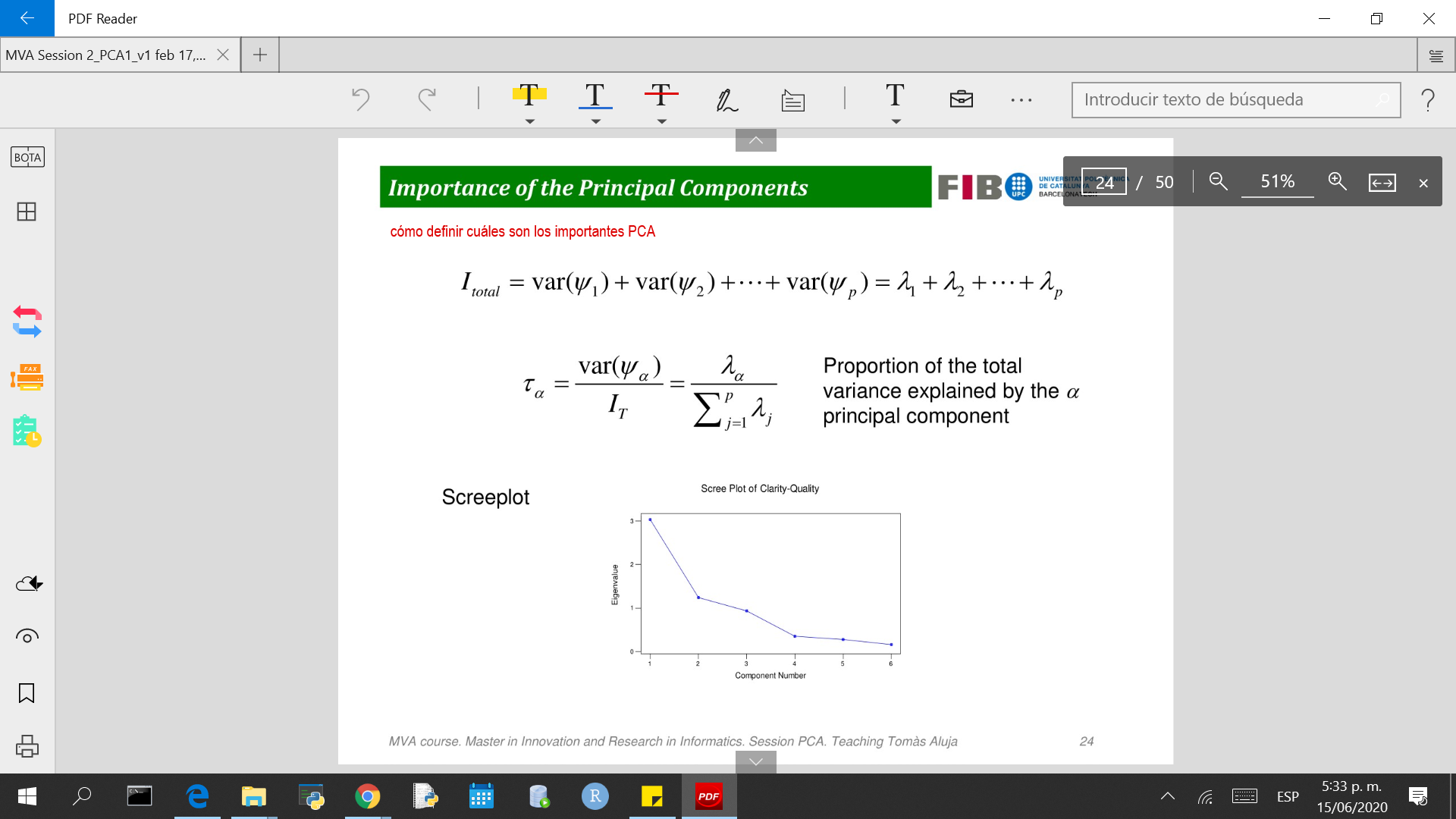
Z1 =ϕ11X1 + ϕ21X2 + … + ϕ81X8

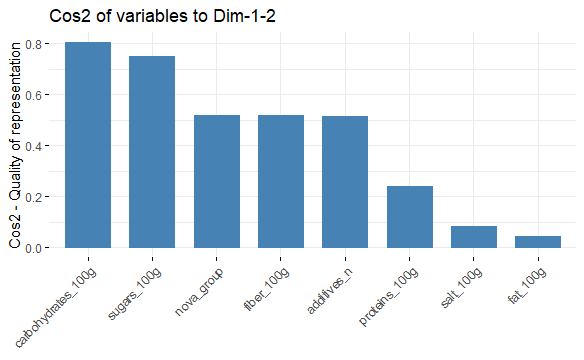
where the largest variance is captured in the first component ϕ1, an artificial value that is a combination of the original value and its variance, later on computed repeatedly until we get as many artificial variables as original values.

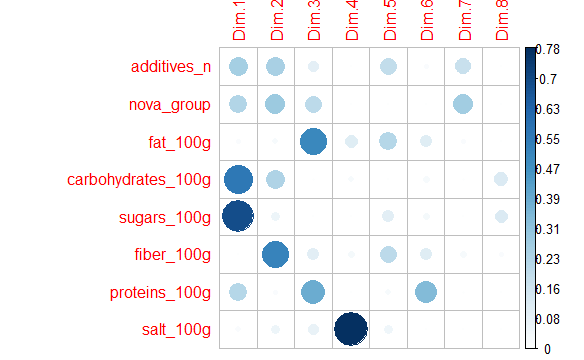


Each principal component vector defines a direction in feature space and keeps a portion of the total inertia of the variables also known as the eigenvalue, it is our objective to retain as much as possible of it in the least number of variables, performing a reduction of dimensions.

The total inertia, variance of the cloud of points, is given by the sum of the eigenvalues:





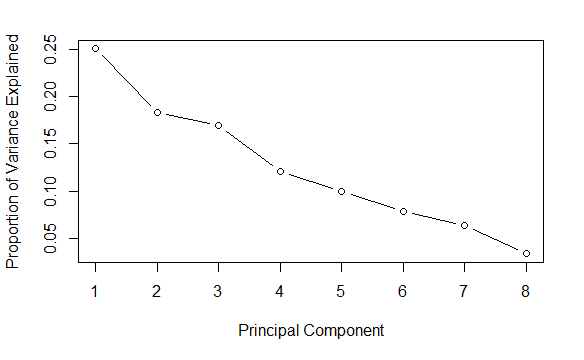


The most well represented features are carbohydrates along sugars, interpreted from the plot; these two dimensions are closely related; explaining in a very similar way the same information about food. This can be due to the fact that both one and the other are usually taken as the same macronutrient in food, and in many cases the carbohydrate supply in one food is determined by the amount of sugar contained in it. Sugar is the simplest type of carbohydrate in its composition and it is frequently the most used one in food processing for being cheap and easy to get. Also inferred from the plot we could say that these two factors are redundant, as one can be explained by the other and we won’t lose significant information in the analysis when choosing only one of them for further analysis.

Nova group and additives are also very close in relation, as it is a general thought that additives are determinant in the nutrition grade given to a food ELABOATE

The least well represented feature is salt, it is also poorly correlated with carbohydrates, indicating that they measure food from opposite dimensions. This is very true when analyzing a food value as salt is by no mean a macronutrient but a mineral (also known as micronutrient), so it is less significant when grading a food, and its contribution to the nutrition score is low in proportion.

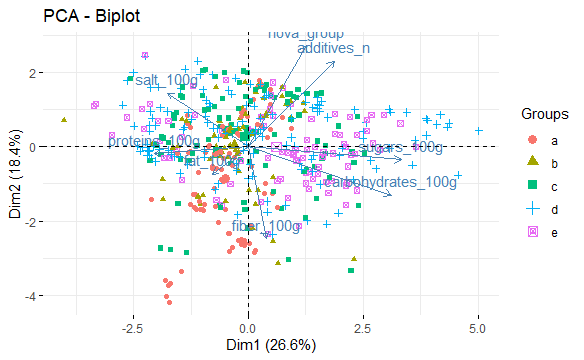
Fiber dimension is very close to ,



According to last elbow rule, the first 4 PCA are significant, explaining a total of 73% of the correlation of variables.

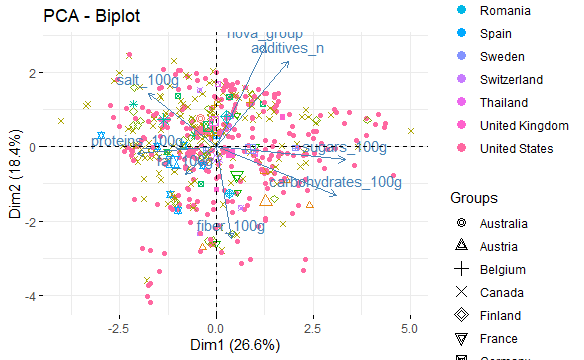
In the aim of analysing further the supplementary variables through a visualization of them in the same factor plane we have sampled a 500 row dataset and plot 2 of the categorical variables and interpret some of the results.

Nutriscore grade



High nutrition score grades are

country



1. Clustering
2. Train and test sampling
3. Modeling
4. Conclusions