Nutrient analysis of pizzas

Author: Romano Francesco

Abstract

The purpose of this report is to analyze, using the statistical method, the amount of nutrients in products sold by fast food pizza chain brands in the USA. The dataset (found on from kaggle.com) consists of three hundreds samples of ten different brands. The variables include not only calories, carbs, fats and proteins but also ash, moisture and sodium.

The goal is to create consumer-friendly graphs to help consumers select the most suitable product according to its own preferences and health. Since in the USA the obesity rate is at 39,6% (according to Wikipedia, 2016), the best way to help consumers is to create accessible and simplified data for them. In order to do so, it is quintessential to look at the analysed data from a critical perspective and apply the right statistical method. Deducing information from data is not an easy task, but with the right approach we can create useful metadata.

Statistical Method

The approach utilized for this study was, initially, an exploratory analysis to understand and explain the data. Then we used the correlation matrix to analyze the relations between nutrients.

Dataset

Here we show a quick sample of the data utilized:

```
dati = read.csv("C:/Users/COMPUTER/Desktop/reportstat/pizza.csv", header = TRUE)
head(dati)
```

```
  brand    id  mois  prot   fat  ash sodium carb  cal
1     A 14069 27.82 21.43 44.87 5.11   1.77 0.77 4.93
2     A 14053 28.49 21.26 43.89 5.34   1.79 1.02 4.84
3     A 14025 28.35 19.99 45.78 5.08   1.63 0.80 4.95
```

```
4     A 14016 30.55 20.15 43.13 4.79   1.61 1.38 4.74
5     A 14005 30.49 21.28 41.65 4.82   1.64 1.76 4.67
6     A 14075 31.14 20.23 42.31 4.92   1.65 1.40 4.67
```

Here is a summary of the most peculiar recurring variables. All the amounts are per 100 grams of sample:
Brand: From A to J, the variable anonymously represents the food chain brand which is selling the samples analyzed.
Moisture: The amount of water in the product. Moisture is also a key factor in food packaging because the more water is used, the more suitable the environment for bacteria. However, this also means that the product can be filled with chemical additives.
Ash: Ash refers to any inorganic material, like minerals, present in food. It is called ash because it is a remaining in the product when, by heating, water and organic material (fat, protein) are removed. Ash can include both compounds of essential minerals (calcium and potassium), and toxic materials, like mercury. Generally, any natural food will be less than 5% ash in content, while some processed foods can have ash content of more than 10%.
Sodium: Sodium is one of the most common minerals found in the human blood. It is fundamental for the organism because it regulates fluids' passage between cells and it is the main element concerning nervous system impulses. There are many foods with a substantial amount of sodium such as cold cuts and cheeses, very commonly found on many of the most ordered pizzas. Nonetheless, it is very important to adopt a low-sodium diet: an excessive amount of sodium in our body can cause diseases to the circulatory system,, leading to cardiovascular disease . The needed dose of sodium goes from 0,6 and 3,5 grams per day.

The other variables represent the amount of nutrients in the samples analyzed.

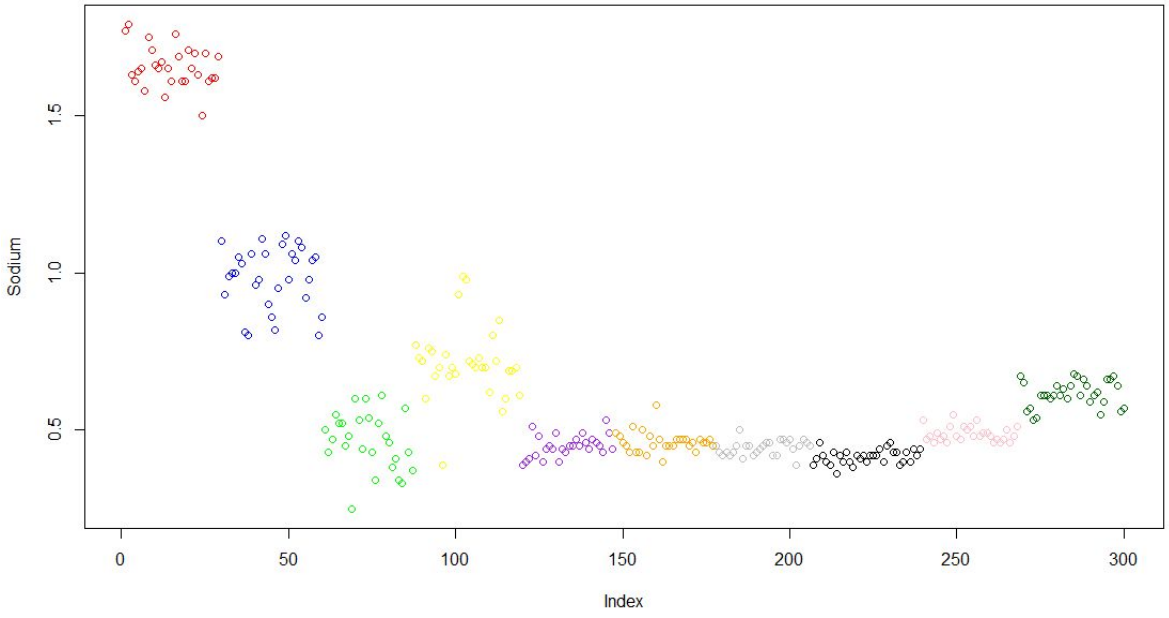
Analyzing the data

As stated before, obesity rate in the US is very high, mostly, this is caused by bad and unhealthy food habits and by a lack of knowledge about nutrients in

food consumed. Speaking of nutrients, the most dangerous are sodium, ashes and to an extent fats.

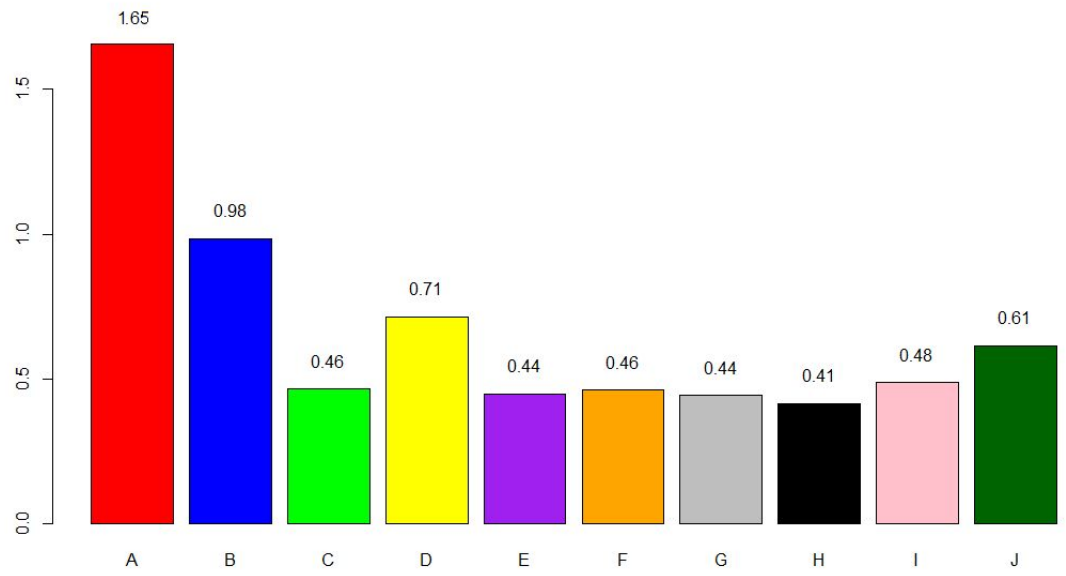
It is now important to have a look at the amount of sodium in these pizzas. We create a plot for the "*sodium*" variable with all the samples analyzed, each color represents a different brand:

```
plot(dati$sodium, col=c("red","blue","green","yellow","purple","orange",
                "gray", "black", "pink", "darkgreen")[dati$brand])
```



With this plot we can already gather some useful informations from the data. But let's go even further and study the average amount of sodium for each brand:

```
barplotmedie = barplot(height = mediesodio$x, names = mediesodio$Group.1,
             col=c("red", "blue", "green", "yellow","purple", "orange",
"gray", "black", "pink", "darkgreen"))
text(barplotmedie, mediesodio$x + 0.1, paste(lbl))
```

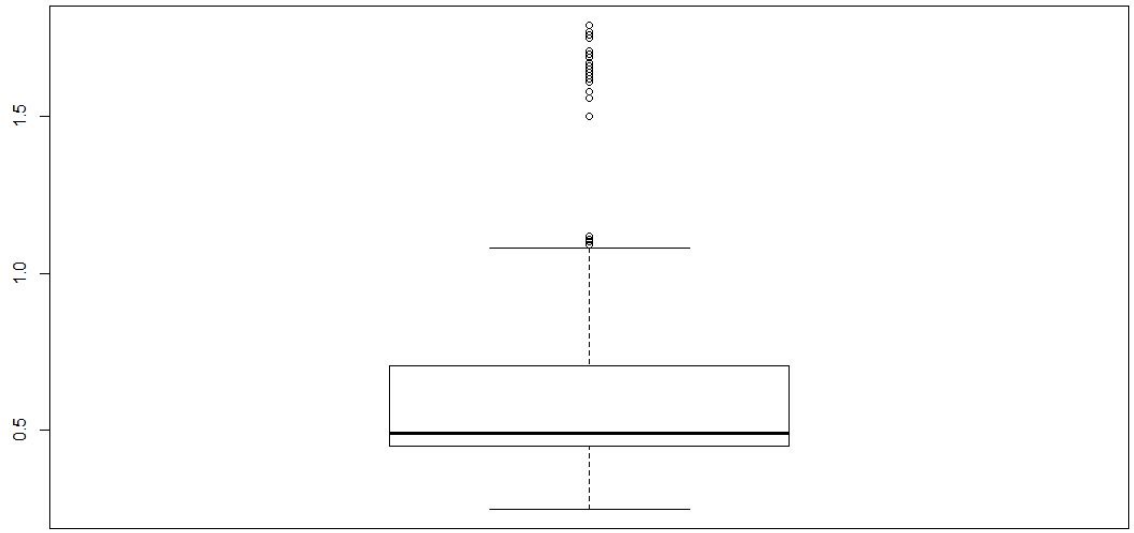


Each bar represents the mean of sodium amount for each brand, every brand has its own color. From a quick glance we can check that most of the brands have an average amount of sodium of 0.45, while four brands exceed that amount, some of them with a substantial gap.

The **WHO**(“*World Health Organization”)* recommend a daily assumption of sodium of 2 grams. Considering that most of the food we eat has a notable amount of sodium in it, the challenge of the WHO is to sensibilize people to a better consciousness on the topic. From the barplot we have two brands, red and blue, whom average amount of sodium is very substantial. For the red brand, 1.65 average amount of sodium per 100 grams of pizza is surprisingly high.

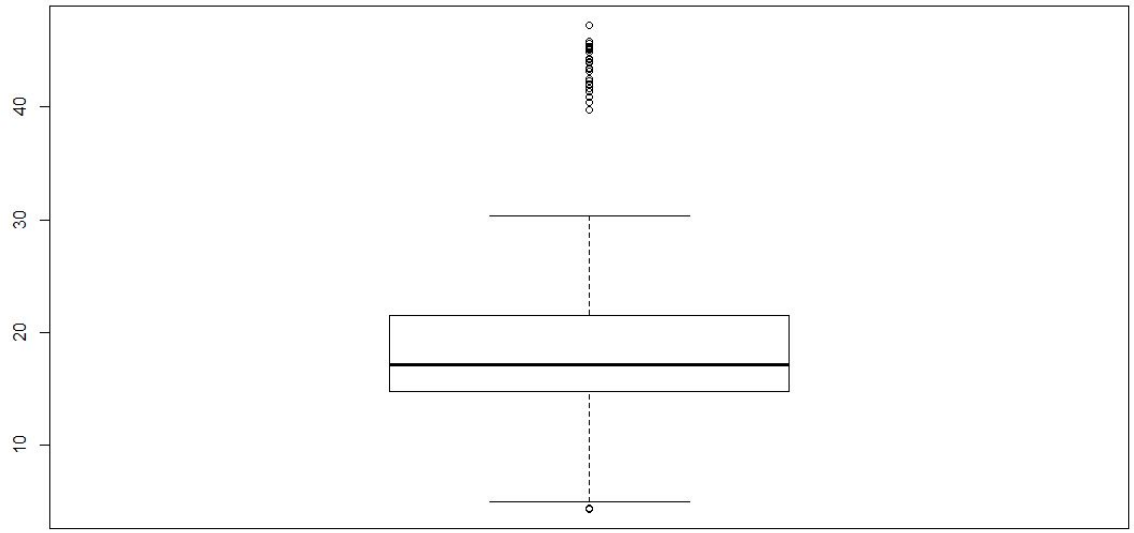
To aid our study of this variable we create the boxplot of the observations:

```
boxplot(dati$sodium)
```



The boxplot is a very useful tool that allow us to gather as much informations as possible from very little data. From the boxplot we can check not only if there are any anomalous values but we can also check (among many other things) the symmetry of the whole variable. In this case the boxplot highlights a clear positive asymmetry .
As we can see the samples belonging to the red brand counts as anomalous values or *outliers*.
It is up to us to speculate whether these values come from a different analysis or if there is really a brand selling products with high amount of Sodium.

We run in the same issue by analyzing the boxplot for the variable *fat*:



As we can see the amount of fat inside products sold by the red brand goes over 40 per 100 grams of sample.

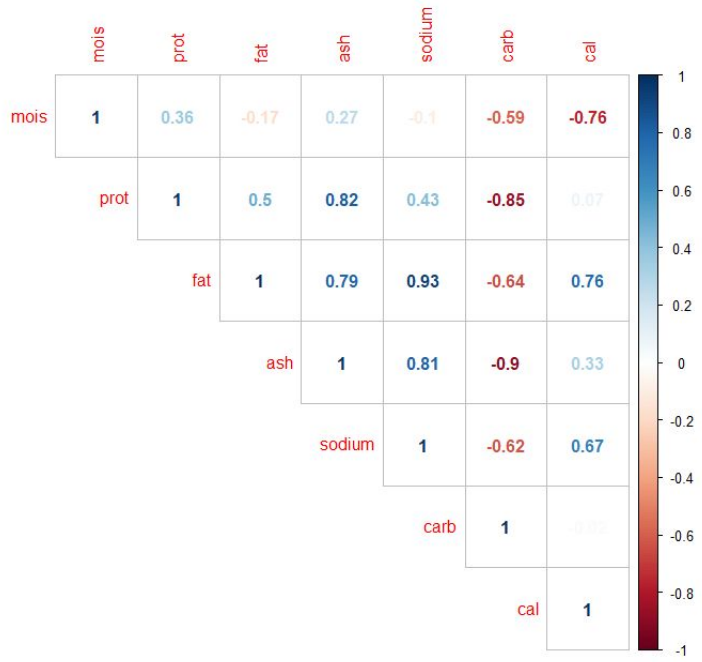
We compared this data to the average values of nutrients found in popular brands of frozen pizzas sold in italian supermarkets. While the average amount of fat inside american fast food pizza is **20.22** (per 100 grams), we found that frozen pizza sold in Italy has the average of **6.02** fat.

Looking at data with critical eyes:

Now that we explored our dataset a bit we can go further with our analysis and study the relations between some of our variables. This kind of approach is important because it lets us discover more properties regarding our available data.

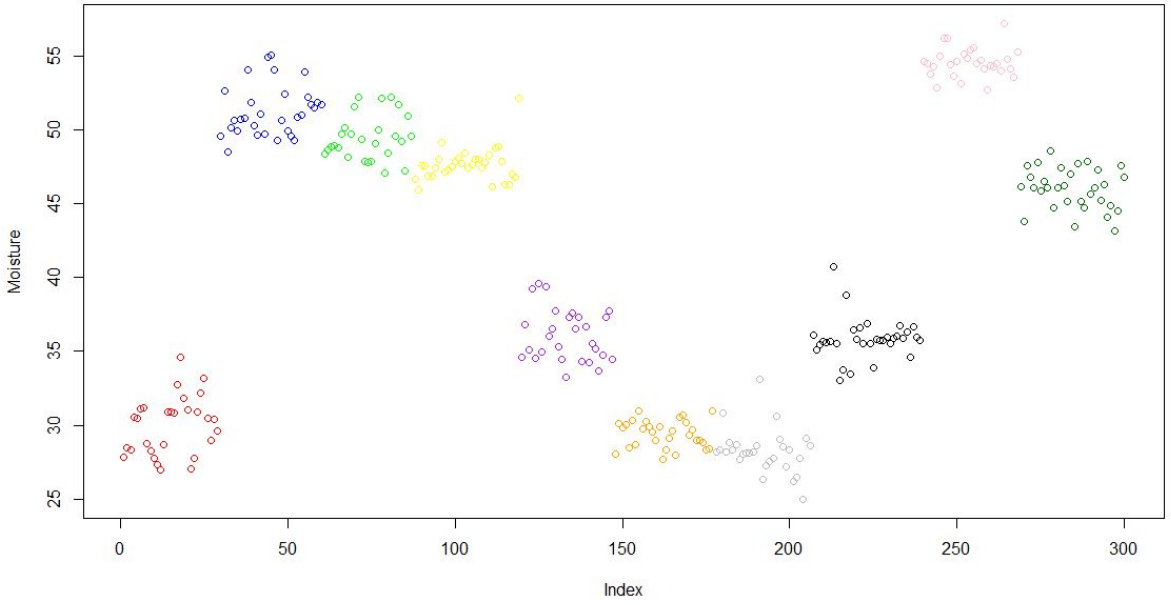
First thing first, we create, via R software, a correlation matrix. This tool helps us, understand if each variable has a relation with another one.

```
M = cor(x = matr)
corrplot(M, method = "number", type = "upper")
```

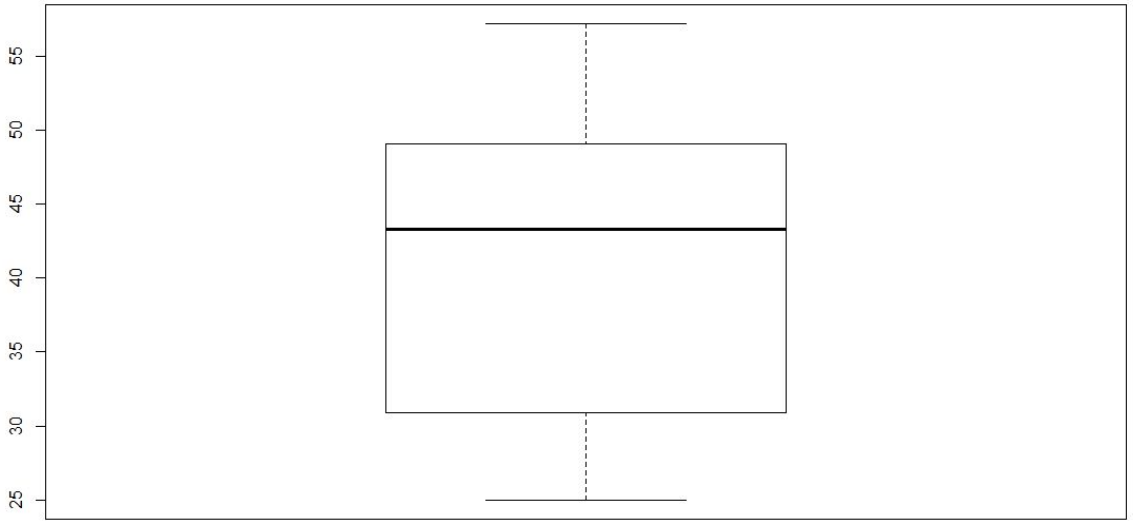


Let's look at the matrix and found the most peculiar relations:
First we have *sodium* and *fats*, we already analyzed these variables. It is clear that the fatter the ingredients in a pizza (cheese or salami) are, the more sodium they have.
Second, we look at the relations of *proteins* and *fats* with another peculiar variable: *ash*.
As already described ashes are what's left of inorganic matter after being burned, after a field research we've come to a possible explanation:
We initially thought this was related to wood piles burned in the ovens, but this was immediately discarded as, being fast food, the use of wood oven would not be that cost effective (and also not eco-friendly!). Finally, we can speculate that the reason is to be seen in the type of refined flour used, but we have no reliable means to check for the kind of flour used for each sample of each brand.
Lastly, we have *moisture* and *calories*, we will focus the next part of our study in the explanation of the possible connection between these two variables.

First of all let's look at the distribution of *moisture* for each brand:



This time the red brand is showing lower values. Is it a good thing?
Let's look at the boxplot for the entire variable:



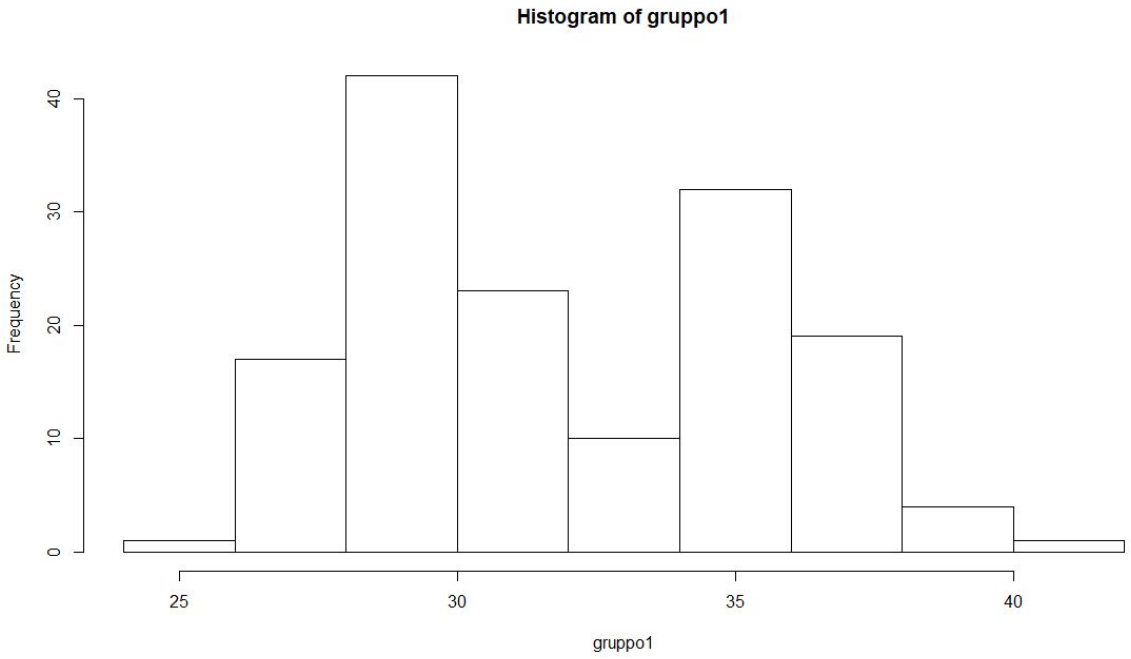
Surprisingly, there is no outlier this time. Although by looking at the boxplot we notice the negative asymmetry of the distribution.

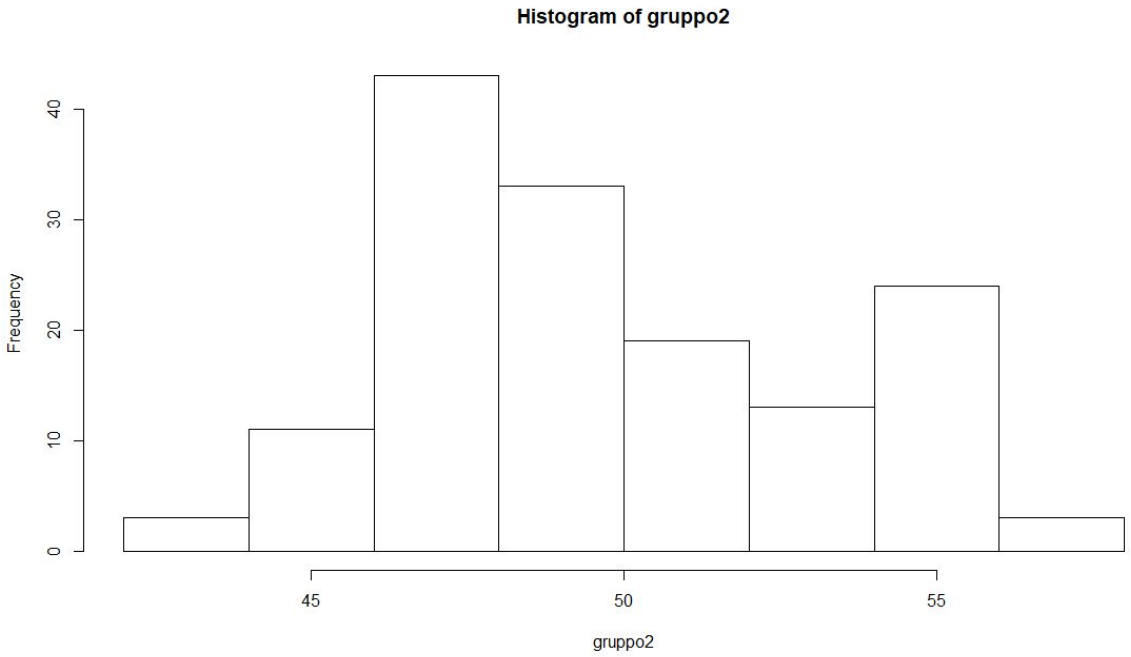
Why there is such a difference between groups of brands? Moisture indicates water presence in the product, among many ingredients, the one most capable of

holding water is of course flour. Even if, as stated before, we have no means of knowing what kind of flour was used, the gap can prove as key point at least to understand the strength of the flour. The term “strength” (labeled W in some books) referred to flour indicates the amount of water that kind of flour can hold.

First of all, we divide the distribution in two groups, one with higher moisture and the other with lower amounts, then we check for the normality of the distribution. Since our champion is too big we use the histogram as tool.

```
hist(gruppo1)
hist(gruppo2)
```





Both groups have no normal distribution, future tests should be conducted while aware of this condition.

Next, in order to prove that there is a true difference between means of each group, we conduct a **t-test**.

```
t.test(gruppo1, gruppo2)
```

This is our output:

```
     Welch Two Sample t-test

data:  gruppo1 and gruppo2
t = -43.841, df = 292.03, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.55932 -16.96457
sample estimates:
mean of x mean of y
 31.96443  49.72638
```

The tests indicates that there is a difference between the two means, the alternative hypothesis is valid.

With the t-test proving the difference between means, it is proven that there is a strong difference between the two groups. If we attribute this difference to the kind of flour used (strong vs weak flour) we have, thanks to our correlation matrix, positive results.

The more strength has a kind of flour, more water can absorb. But that's not all, higher strength equals higher amount of proteins inside the flour and less carbs. This proves not only the inverse relation between *moisture* and *calories* (and also *carb*) but also the small positive relation that moisture has with proteins.

The usefulness of this result is clear: more water contains a certain type of flour more the gluten it produces, so, while celiac must still to avoid these products, some people who chose to eat less gluten might choice brands where they use a weak flour.