Cereals & Ratings

A Statistical Overview

Data Description & Exploration

Data Source:
Cereal dataset with multiple different cereals with different ratings based on survey from the consumers.
16 Variables & 80 Entries

Variables:

Rating Calories Protein Carbs
Sugar Sodium Fiber Fat

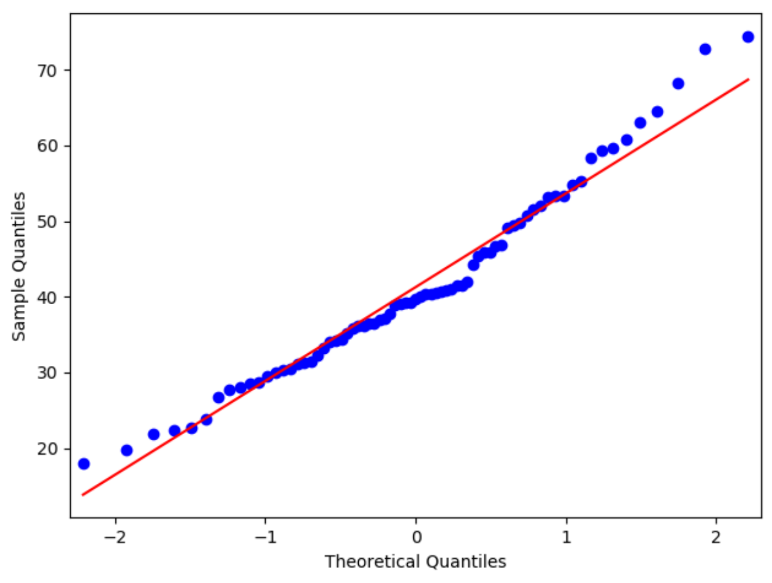
Standardize everything related to weight to **100 Grams**

Methodology:
- Data pre-processing
- Dimension Reduction
    - PCA
- Clustering
    - K-Means
- Linear Regression

Objective:
- Analysis of the ratings of the cereals
- Similarity analysis of the cereals
- Predict the ratings of new possible cereals

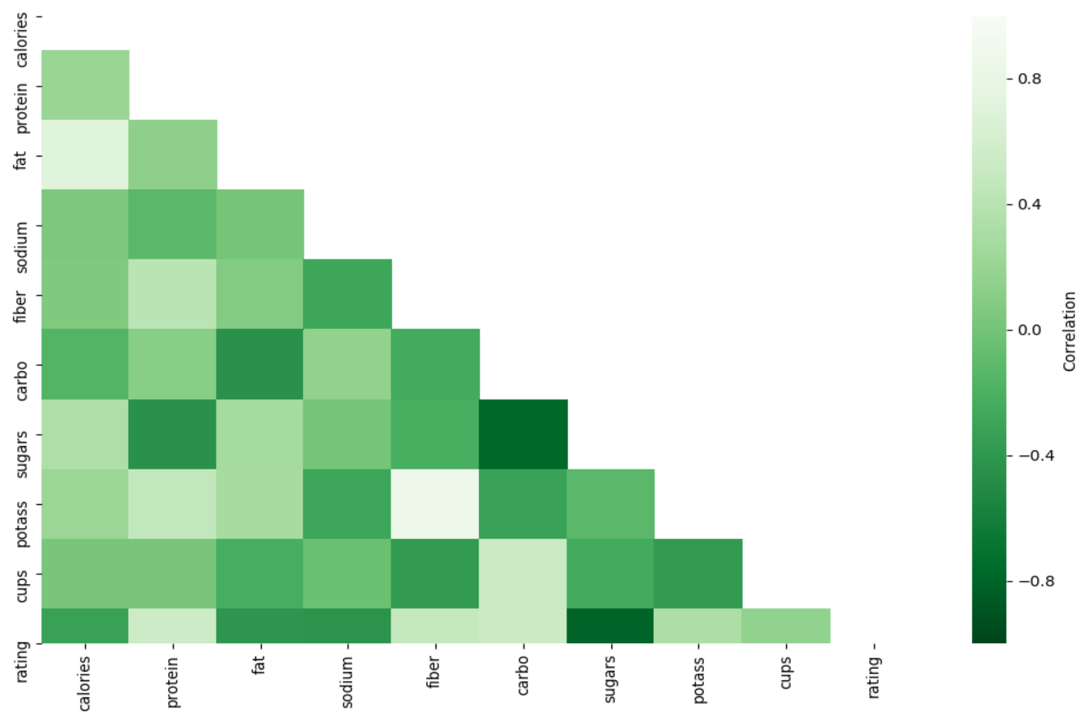
Normality Assumption:
In order to use standard techniques (parametric statistical methods), the dataset has to follow the normal distribution

Quantile-Quantile Plot (QQ-Plot):
A QQ-Plot is a graphical method to verify a normal distribution. If the datapoints (represents the quantiles of the dataset) are close to line (represents normal distribution), the data is normally distributed.



Interpretation QQ-Plot:
The datapoints are close to the line. Hence, we can assume normal distribution and are able to apply standard parametric statistical methods.

Correlation:
The correlation heatmap shows
1. Strong negative correlation between **sugar** and the **rating**
2. Strong negative correlation between **sugar** and **carbohydrates**
3. Strong positive correlation between **fiber** and **potassium**
4. Average correlation among all other variables

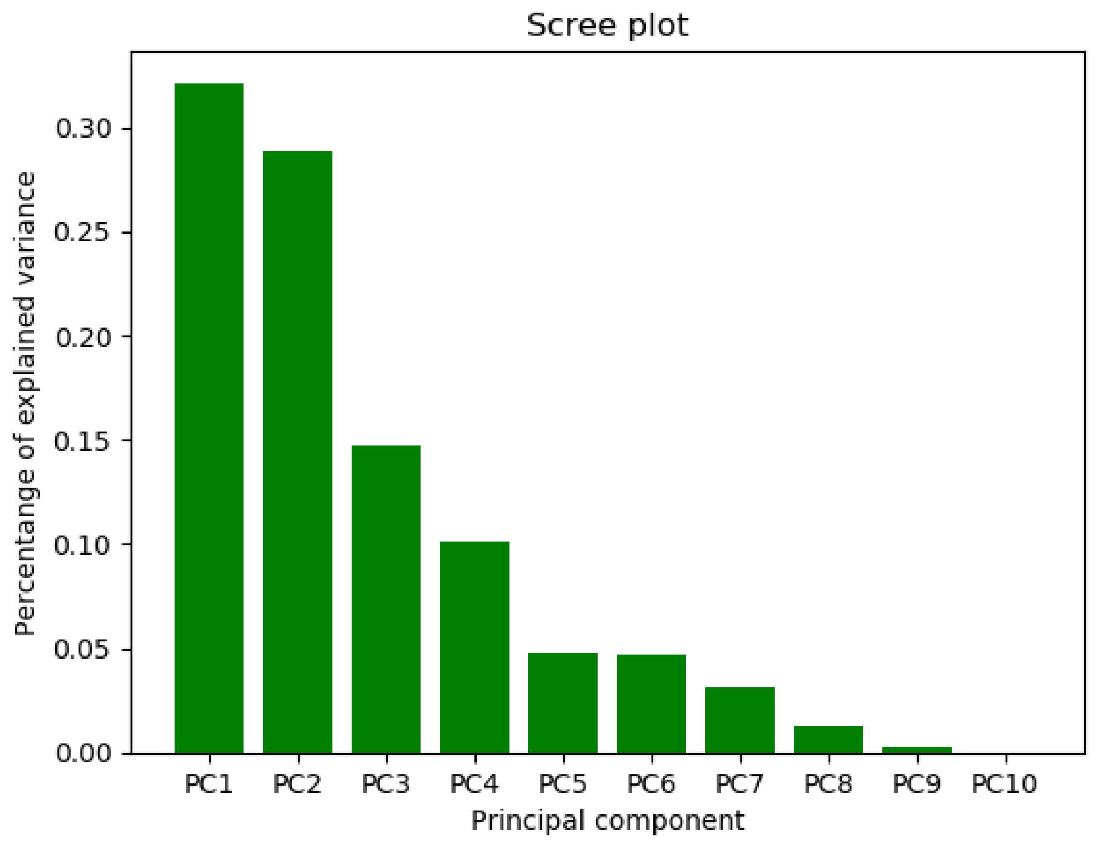


Principal Component Analysis (PCA)

Objectives of PCA:
- Build new variables (linear combinations of existing variables)
- Reduce dimensionality

Interpretation Scree Plot:
- PC 1 explains 32 % of the variation
- PC1 – PC4 explains 86 % of the variation which can be considered as adequate

Further use of PCA:
- Computed PC's can be used for the following clustering analysis or a linear regression

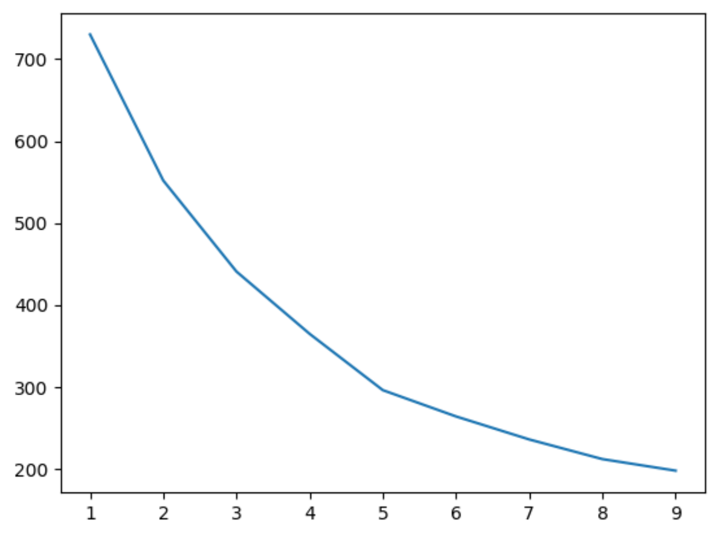


	Potassium & Carbo	High Sugar	High Calories	Low Sodium
calories	-0.492	-0.264	**0.759**	0.065
protein	-0.344	0.670	0.380	-0.046
fat	-0.645	-0.320	0.548	-0.025
sodium	0.275	-0.255	0.244	**-0.858**
fiber	-0.693	0.573	-0.240	-0.069
carbo	**0.713**	0.528	0.324	-0.083
sugars	-0.334	**-0.867**	-0.142	0.171
potass	**-0.806**	0.485	-0.086	-0.032
cups	0.588	0.144	0.470	0.473

Conclusion PC:
- PC1 (Potassium & Carbs) is strongly positive correlated to carbo (complex carbohydrates) and highly negative correlated to potassium, therefore the PC1 increases with increasing in carbo and decrease with increasing in potassium
- PC2 (High Sugar) has a strong negative correlation with sugar and a relatively high positive correlation with protein. PC2 mostly explains variable sugar
- PC3 (High Calories) has a positive correlation with calories and is primarily an explanation of calories and increases with increasing calories
- PC4 (Low Sodium) has a strong negative correlation with sodium and is primarily a measure of sodium. This PC decreases with increasing in sodium

Clustering

Objectives Clustering:
- Grouping cereals with similar components
- K-Means was the chosen method with 3 clusters based on the elbow graph below

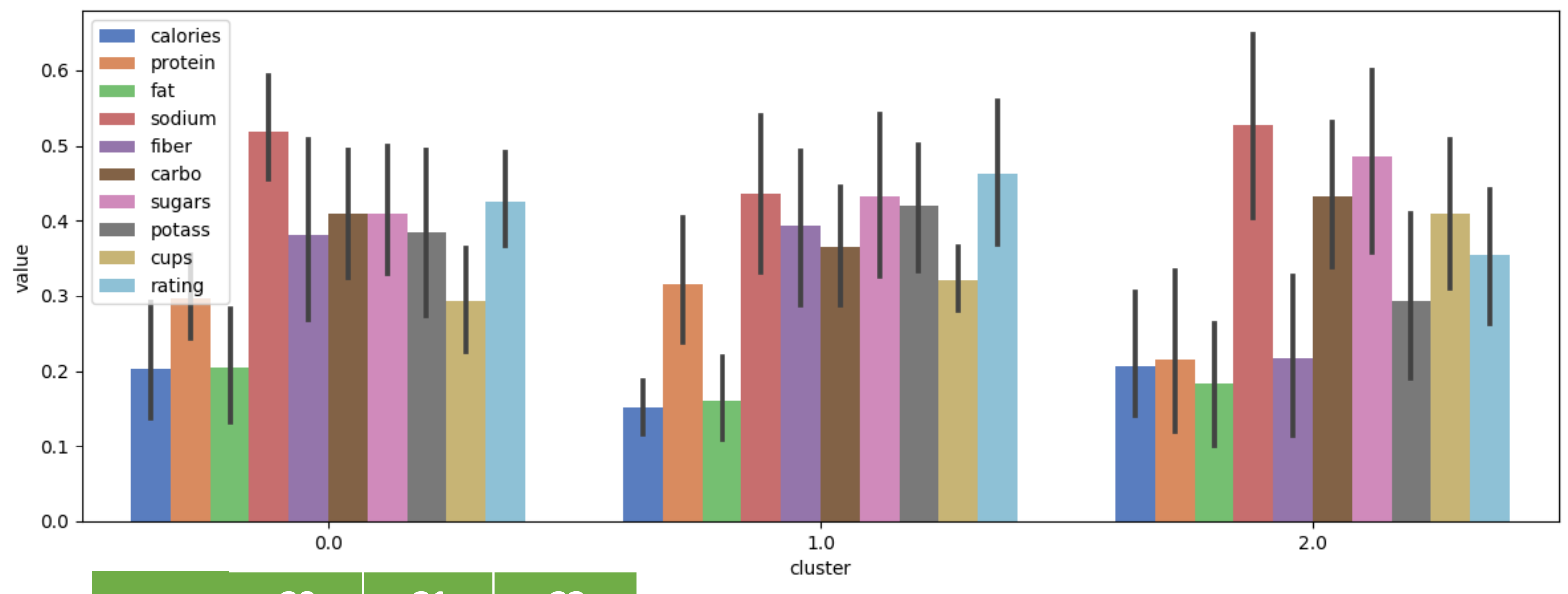


Interpretation Clustering:
On the side plot we can have a visual look of our final clusters according our variables
Cluster 0 - <u>Medium Rating, High Protein</u>
- This group are composed by the <u>least healthy cereals</u> with High Calories; High Fat and High Fiber

Cluster 1 - <u>High Rating, Low Calories</u>
- This group are composed by the <u>healthy cereals</u> with Low Calories; Low Fat and Low Sugar

Cluster 2 - <u>Low Rating, High Sugar</u>
- This group are composed by the <u>sweet cereals</u> with an Average Calorie and Average Fat but High Sugar

The calculation of the distance between clusters using Euclidean distances it's also presented on the right table.



	C0	C1	C2
C0	0	47.2	21.9
C1	47.2	0	47.8
C2	21.9	47.8	0

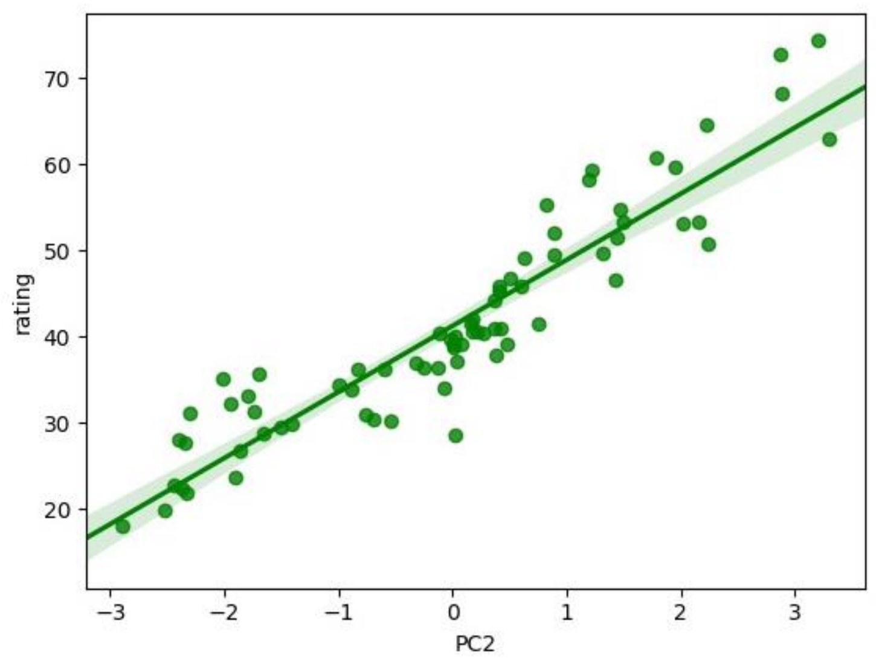
Conclusion Clustering:
After looking into the final clusters we realize that, unlike to our initial thinking, healthier cereals have a higher rating than sweeter cereals.

Linear Regression

Objectives Linear Regression:
- Predicting the rating of cereals
- Understanding the importance of each variable through the coefficients

Assumption of the Model:
- ✓ Linear in parameters
- ✓ No perfect collinearity
- ✓ Zero conditional zeros
- ✓ Homoscedastictiy
- ✓ No correlation of errors
- ✓ Normality

Simple Linear Regression with 1 Principal Component:
- Adjusted R-squared: 86%
- Intercept: 41.2
- Slope: 7.6
- Prob: $1.15e^{-32}$



Multiple Linear Regression with 2 Principal Components

$$y_i = \beta_0 + \beta_1\chi_{i1} + \beta_2\chi_{i2} + \epsilon_i$$
$$i = 1,2,\ldots,73$$

<u>R-squared</u>: 87%
<u>Significance</u>: PC1 and PC2 are highly significant on cereals rating.

Conclusion Linear Regression:
The rating is mostly driven by sugars. Since our PC2 explains the most variance of sugars it's the one which fits better the model for rating.
The addiction of PC1 improves the model score by adding the Potassium and Carbs components while keeping an easy interpretability.

Conclusion

Cereals are grown for their highly nutritious edible seeds which are often referred to as grains, but once they are processed and mix with other ingredients in order to be commercialized, they might become a dangerous to our health.

- **PCA analysis** we can see that the most import feature in cereals are values for Potassium and Carbs; Sugar; Calories and Sodium.
- **Clustering analysis** show us an overview of the differences and similarities between groups of cereals with emphasis on the rating variable made by the consumers of the different cereals. We can easily see that healthier cereals have higher ratings.
- **Linear Regression** proves the conclusions above as a higher amount of sugars will lead to a lower rating.

NOVA IMS Information Management School

Nova IMS – Data Science & Advanced Analytics
Statistics for Data Science
2019/20

Ana Cláudia Alferes	M20190932
Lennart Dangers	M20190251
Pedro Santos	M20190420