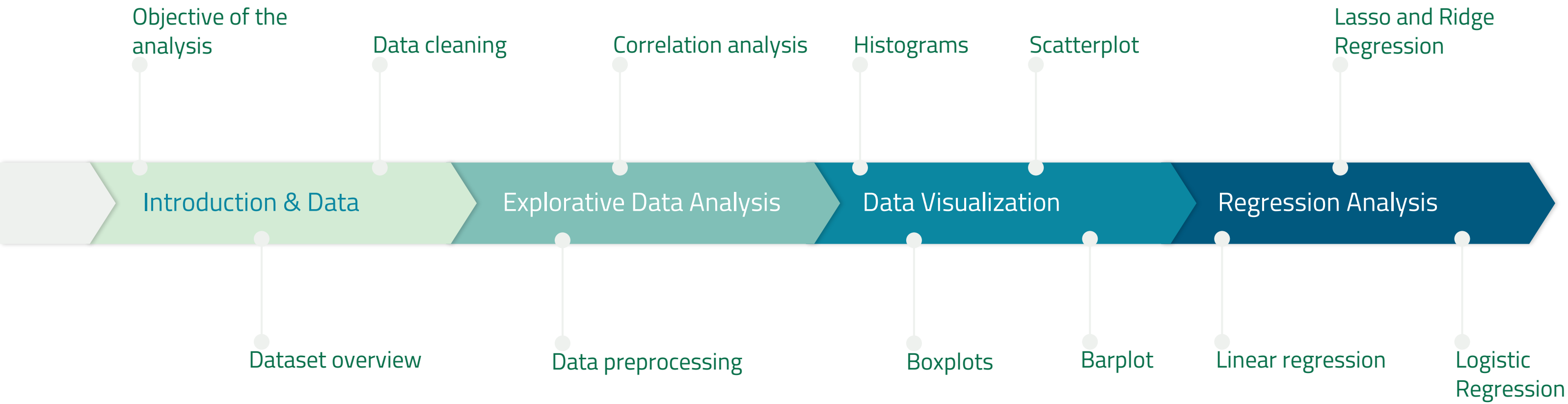
Starbucks™

Statistical Learning Project

Alberto Calabrese
Eleonora Mesaglio
Greta d'Amore Grelli

Content



Objective of the analysis

Introduction & Data

Dataset overview

Data cleaning

Data preprocessing

Correlation analysis

Explorative Data Analysis

Histograms

Data Visualization

Boxplots

Scatterplot

Barplot

Linear regression

Lasso and Ridge Regression

Regression Analysis

Logistic Regression

1. Introduction & Data

Objective of the analysis | Data

Objective of the analysis

In this project, we conduct a thorough analysis of the *Starbucks™ Beverage Components* dataset, which contains information about the ingredients of Starbucks™ beverages. Our goal is to gain a comprehensive understanding of the data and build models for accurate predictions.

The process involves several key steps:

1. Data Cleaning: We handle missing values and ensure the data is correctly formatted.
2. Exploratory Data Analysis (EDA): Using visual and quantitative methods, we explore the data structure and the relationships between variables.
3. Regression Analysis: We analyze the relationship between dependent and independent variables, focusing on predicting and understanding the factors influencing the Calories variable.

Dataset

The dataset we will analyze is the *Starbucks Beverage Components* from **Kaggle**, accessible at https://www.kaggle.com/datasets/henryshan/starbucks

This dataset provides a comprehensive guide to the nutritional content of beverages available on the Starbucks menu. It consists of 242 samples described by 18 variables, including the beverage name, categorization, preparation method, total caloric content, and various constituents of the beverages.

2. Explorative Data Analysis

Data preprocessing | Correlation Analysis

Problems with the data & data preprocessing

Data Cleaning

We transformed our row data into numeric values and we renamed the columns to ensure a easily comprehension of the variables.

NA's

We found some NA's value in the Caffein column, we replaced this values with the median of the variable to keep the distribution invariant.

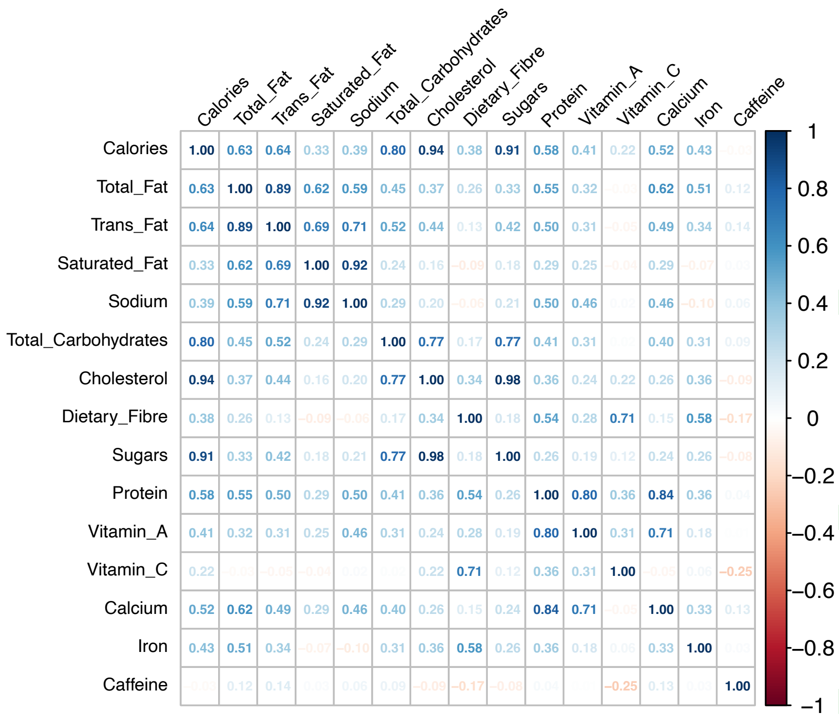
Multicollinearity

We noted that in our data we have a problem with multicollinearity that we tried to solve during regression analysis

Correlation Analysis

We calculated the correlation matrix for our dataset. This computation helps us in comprehending the interrelationships among the dataset's variables.

In the correlation matrix, a value near to 1 at the ij position indicates a strong positive correlation between the i-th and j-th variables. Conversely, a value close to −1 signifies a strong negative correlation. A value near 0 suggests that the two variables do not significantly influence each other.



	Calories	Total_Fat	Trans_Fat	Saturated_Fat	Sodium	Total_Carbohydrates	Cholesterol	Dietary_Fibre	Sugars	Protein	Vitamin_A	Vitamin_C	Calcium	Iron	Caffeine
Calories	1.00	0.63	0.64	0.33	0.39	0.80	0.94	0.38	0.91	0.58	0.41	0.22	0.52	0.43	−0.03
Total_Fat	0.63	1.00	0.89	0.62	0.59	0.45	0.37	0.26	0.33	0.55	0.32	−0.03	0.62	0.51	0.12
Trans_Fat	0.64	0.89	1.00	0.69	0.71	0.52	0.44	0.13	0.42	0.50	0.31	−0.05	0.49	0.34	0.14
Saturated_Fat	0.33	0.62	0.69	1.00	0.92	0.24	0.16	−0.09	0.18	0.29	0.25	−0.04	0.29	−0.07	0.03
Sodium	0.39	0.59	0.71	0.92	1.00	0.29	0.20	−0.06	0.21	0.50	0.46	0.02	0.46	−0.10	0.06
Total_Carbohydrates	0.80	0.45	0.52	0.24	0.29	1.00	0.77	0.17	0.77	0.41	0.31	0.02	0.40	0.31	0.09
Cholesterol	0.94	0.37	0.44	0.16	0.20	0.77	1.00	0.34	0.98	0.36	0.24	0.22	0.26	0.36	−0.09
Dietary_Fibre	0.38	0.26	0.13	−0.09	−0.06	0.17	0.34	1.00	0.18	0.54	0.28	0.71	0.15	0.58	−0.17
Sugars	0.91	0.33	0.42	0.18	0.21	0.77	0.98	0.18	1.00	0.26	0.19	0.12	0.24	0.26	−0.08
Protein	0.58	0.55	0.50	0.29	0.50	0.41	0.36	0.54	0.26	1.00	0.80	0.36	0.84	0.36	0.04
Vitamin_A	0.41	0.32	0.31	0.25	0.46	0.31	0.24	0.28	0.19	0.80	1.00	0.31	0.71	0.18	0.01
Vitamin_C	0.22	−0.03	−0.05	−0.04	0.02	0.02	0.22	0.71	0.12	0.36	0.31	1.00	−0.05	0.06	−0.25
Calcium	0.52	0.62	0.49	0.29	0.46	0.40	0.26	0.15	0.24	0.84	0.71	−0.05	1.00	0.33	0.13
Iron	0.43	0.51	0.34	−0.07	−0.10	0.31	0.36	0.58	0.26	0.36	0.18	0.06	0.33	1.00	0.03
Caffeine	−0.03	0.12	0.14	0.03	0.06	0.09	−0.09	−0.17	−0.08	0.04	0.01	−0.25	0.13	0.03	1.00

3. Data Visualization

Add something here

Data
Visualization

Histograms

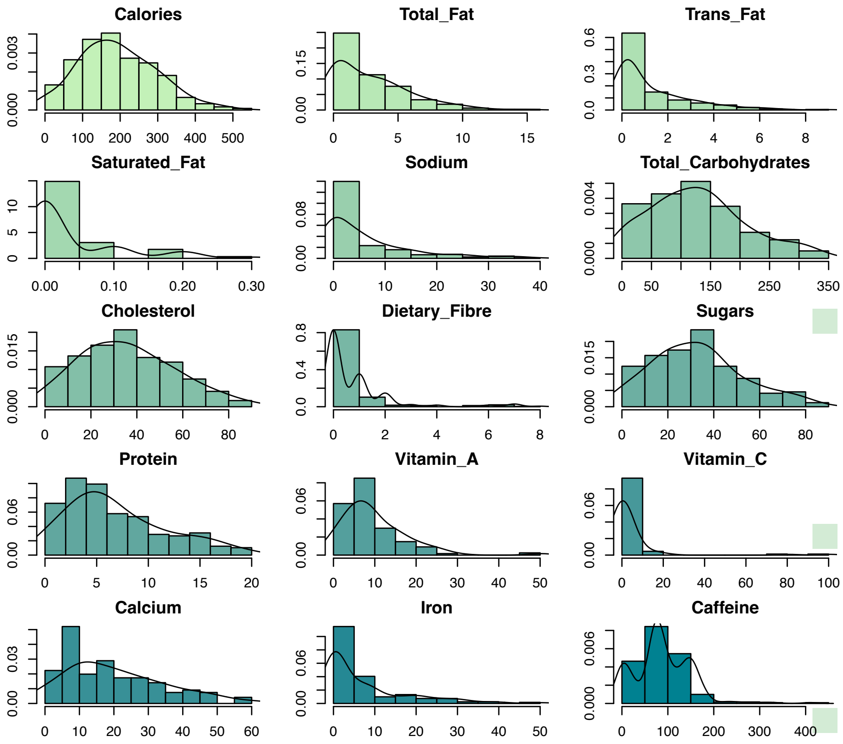
Histograms serve as a graphical interpretation of data distribution.

In a histogram, each bar corresponds to the counted frequency within each bin or interval.

We introduce these plots to see if our data is normally distributed, skewed, or has outlier values.

By looking at the graphs, we can notice that the variables "Calories", "Total_Carbohydrates", "Cholesterol", and "Sugars" exhibit distributions that are nearly normal.

Conversely, the distributions of the remaining variables display a noticeable skewness towards the left.



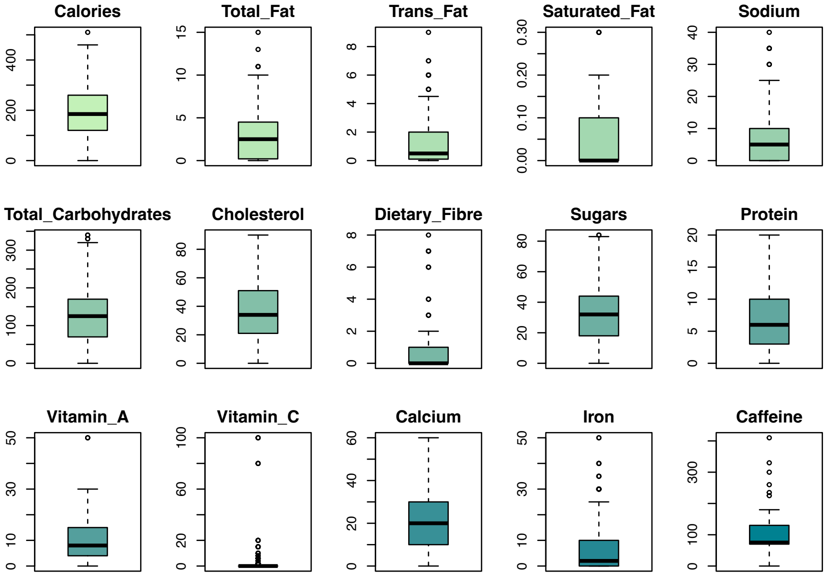
Boxplot

Data
Visualization

Boxplots are a type of graphical representation used to display the distribution of a dataset. They provide a visual summary of the data, enabling us to quickly identify key statistical measures such as median, quartiles and outliers. This visualization also helps us to determine the spread and variability of the data.

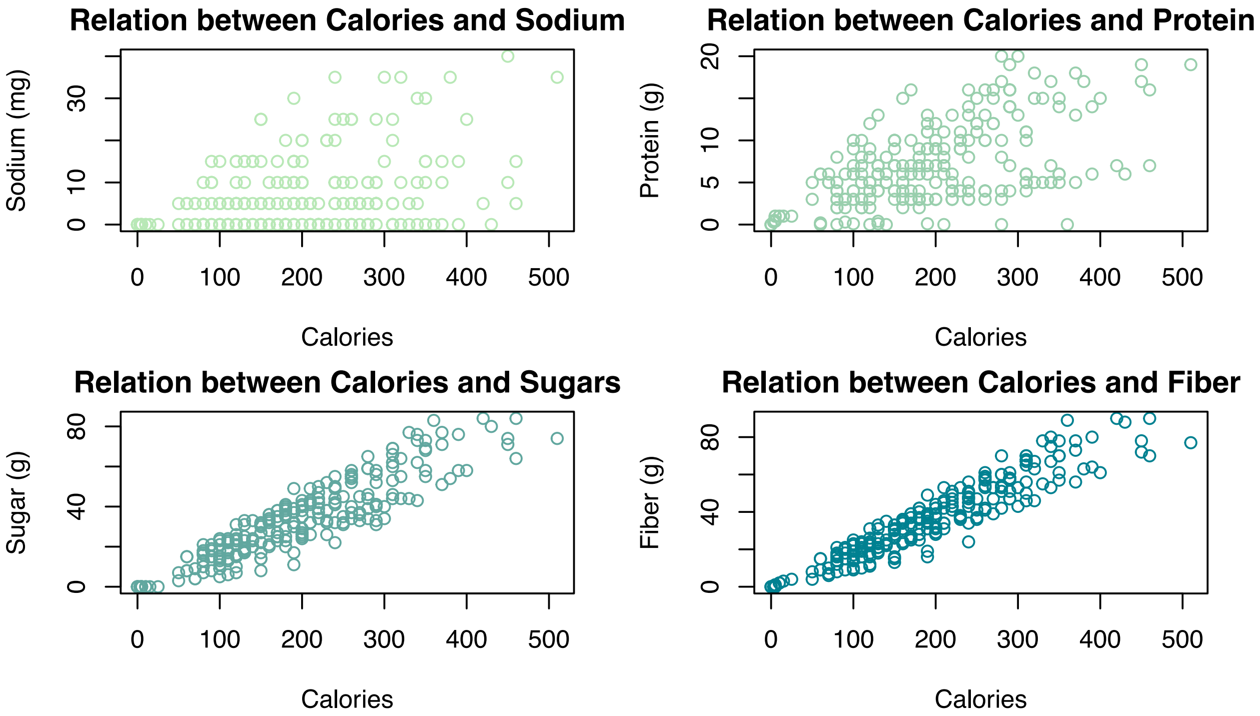
As we observed earlier, the majority of the graphs exhibits a skewness towards zero, with the exceptions being “Calories”, “Total_Carbohydrates”, “Cholesterol”, and “Sugars”.

Another aspect that has not been previously highlighted is the presence of outliers. These are notably evident in “Dietary_Fiber”, “Vitamin_C”, and



Scatterplot

Data Visualization

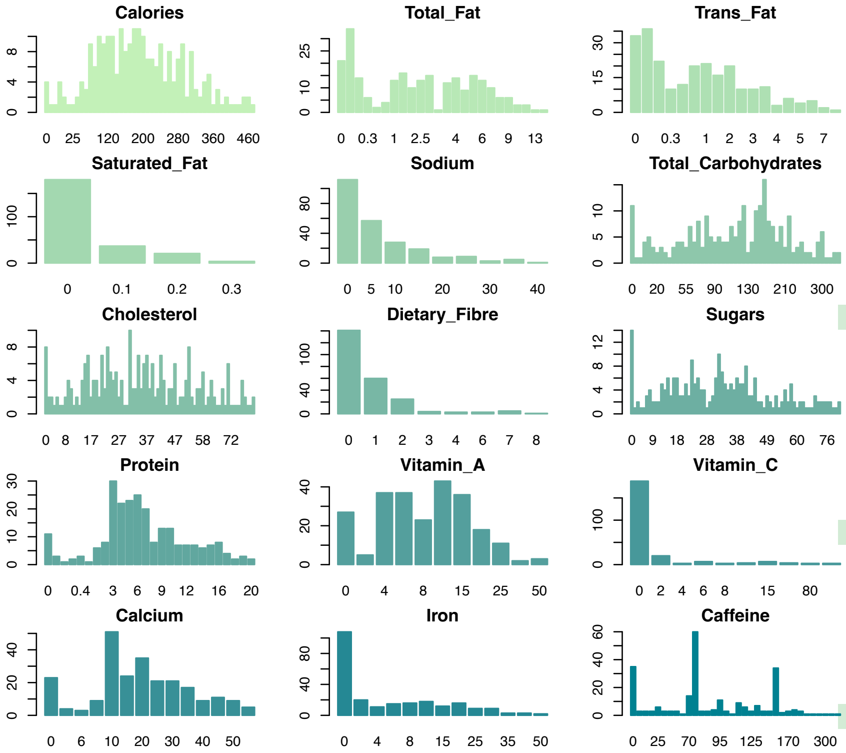


Barplot

The primary use of bar plots is to make comparisons between the amounts of different categories.

Indeed, each bar corresponds to a category and the height of the bar represents the frequency or proportion of that category.

These graphs are commonly used for categorical data, or numerical data that has been binned into categories.



Looking to the plot, we can notice that variables such as "Saturated_Fat",

"Dietary_Fibre", "Vitamin_C", and "Iron" are typically either absent or present in small quantities in the beverages.

In particular, the frequency of these

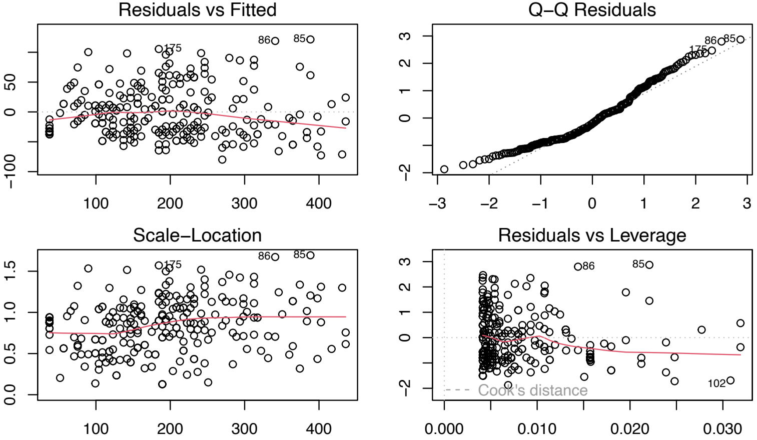
4. Regression Analysis

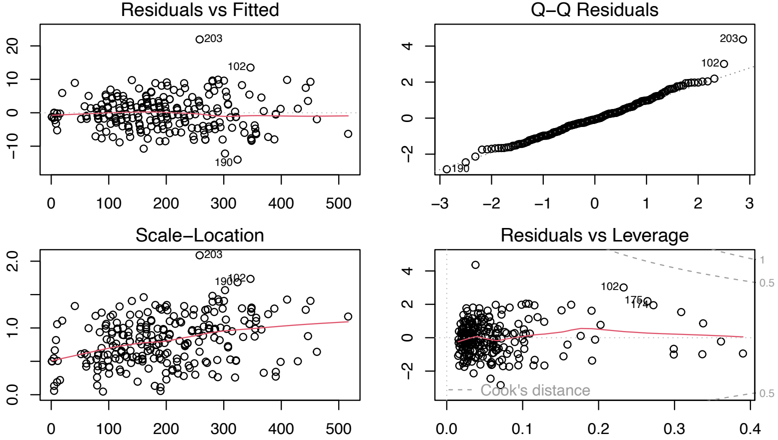
Linear Regression | Logistic Regression

Regression Analysis

Linear Regression

Simple and Multiple





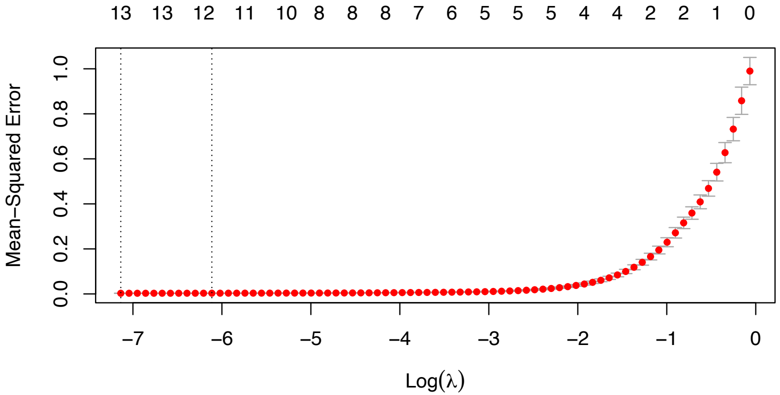
	AIC	BIC	R^2	Adjusted R^2
Simple linear regression	**2509**	**2519**	**0.827**	**0.826**

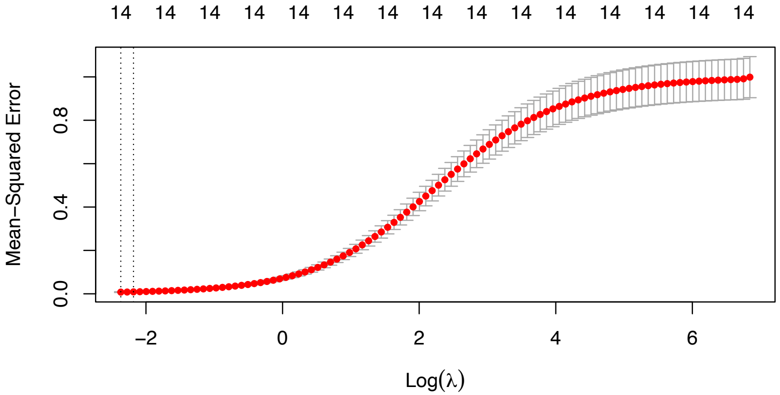
	AIC	BIC	R^2	Adjusted R^2
Multiple linear regression	**1494**	**1550**	**0.997**	**0.997**

Regression Analysis

Lasso and Ridge Regression

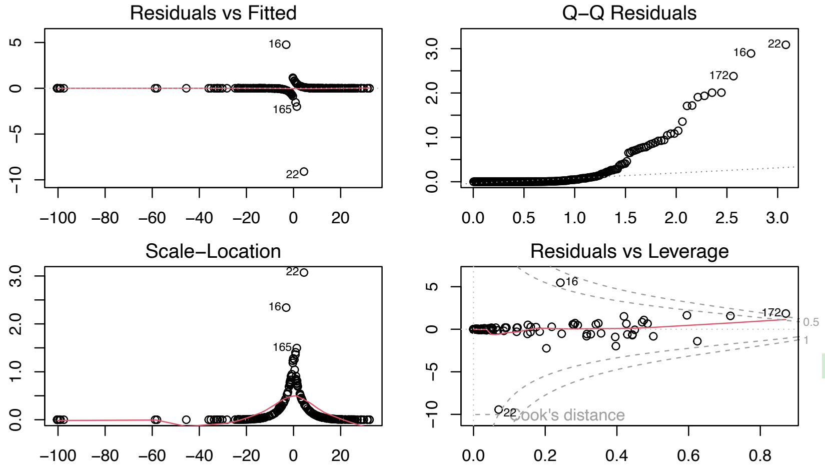
	R^2	MSE
Lasso Regression	**0.9975**	**0.0024**
Ridge Regression	**0.9941**	**0.0066**





Logistic Regression

	Accuracy	Precision	Recall	F1 Score
Cross Validation	**0.91**	**0.92**	**0.92**	**0.92**



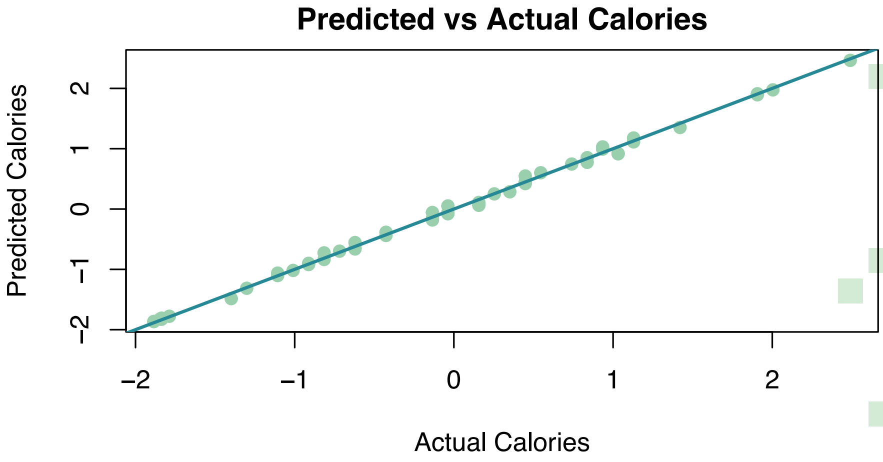
	AIC	BIC	R^2	Residual Deviance	Null Deviance
Multiple linear regression	**69.42**	**121.75**	**0.88**	**39.42**	**335.48**

Cross Validation

Lasso Regression Model

We decided to split the data with 80% of examples for training and 20% for testing.

We evaluate the model using the testing set. We make predictions using the testing set and calculate the mean squared error and the root mean squared error to assess the model's accuracy.



	Accuracy	MSE	R^2
Lasso regression model	**0.997**	**0.0026**	**0.997**

5. Conclusions

Conclusions | Potential implementations

Conclusions

In this project, we conduct a thorough analysis of the *Starbucks™ Beverage Components* dataset, which contains information about the ingredients of Starbucks™ beverages. Our goal is to gain a comprehensive understanding of the data and build models for accurate predictions.

The process involves several key steps:

1. **Data Cleaning:** We handle missing values and ensure the data is correctly formatted.
2. **Exploratory Data Analysis (EDA):** Using visual and quantitative methods, we explore the data structure and the relationships between variables.
3. **Regression Analysis:** We analyze the relationship between dependent and independent variables, focusing on predicting and understanding the factors influencing the Calories variable.

Potential Implementations

Here we can write the idea of propose our best model as solution for companies that want to create a new kind of beverage and thanks to our model can predict the amount of calories based on other variables.

This could be useful specially in US where there is an important obesity disease and a tool like this can really make the difference!

Write better of course. ☺

THANKS!

Any questions?

Alberto Calabrese Eleonora Mesaglio Greta d’Amore Grelli

Add the other images