Starbucks

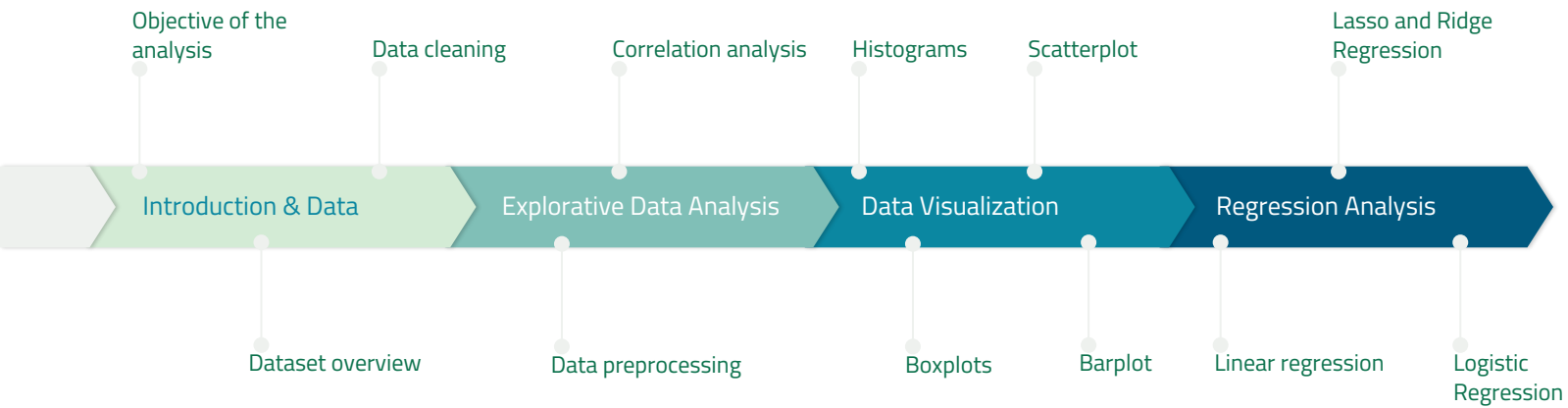
Statistical Learning Project

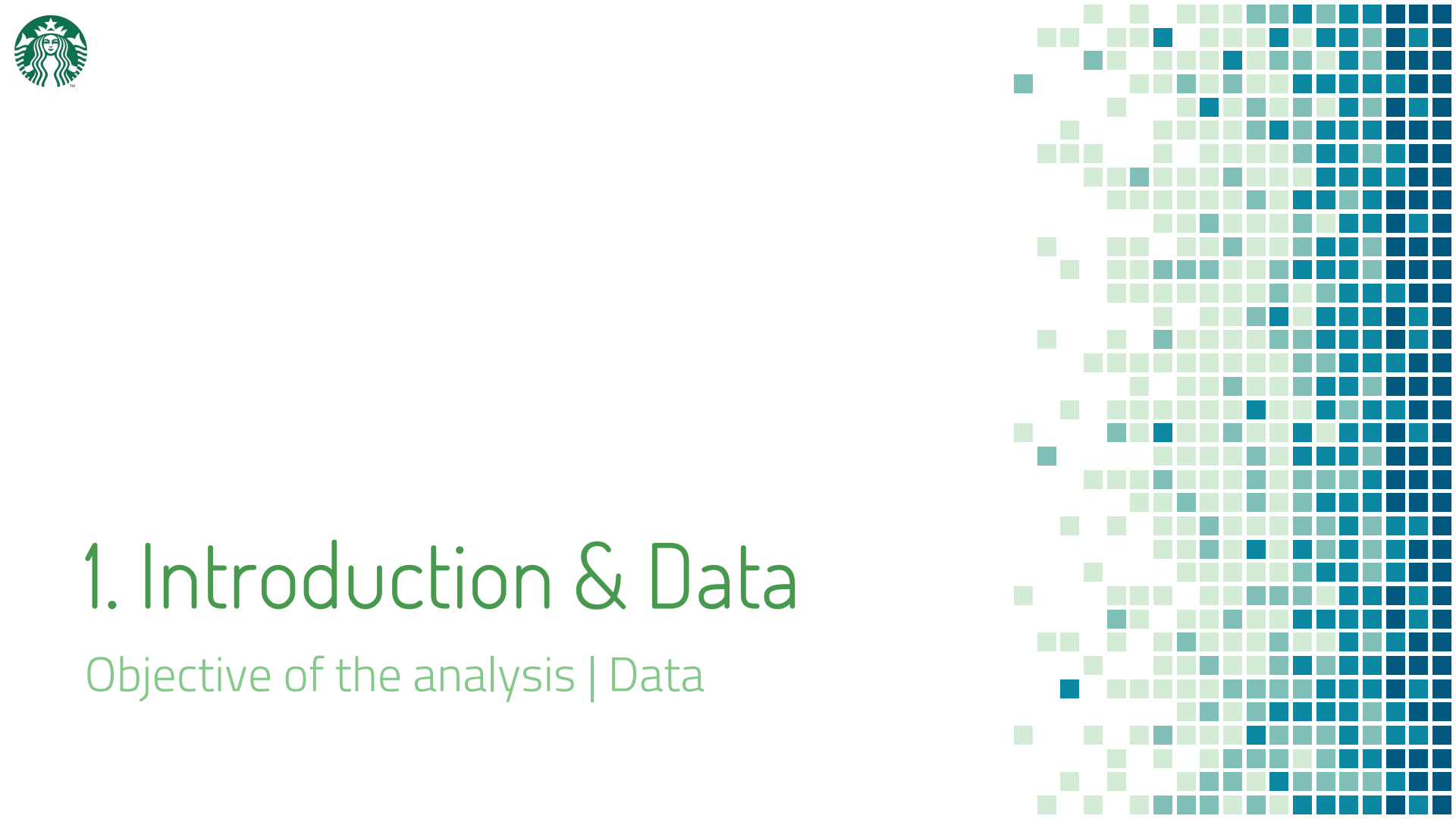
Alberto Calabrese
Eleonora Mesaglio
Greta d'Amore Grelli

What is Starbucks ?

Starbucks is a global coffeehouse chain known for its specialty coffee drinks, teas, and pastries.
Founded in Seattle in 1971, it has since expanded worldwide, offering a variety of beverages and snacks in a cozy, café-style environment.
Starbucks is also noted for its customer-centric approach and ethically sourced coffee beans.

Content


Objective of the analysis
Data cleaning
Correlation analysis
Histograms
Scatterplot
Lasso and Ridge Regression
Introduction & Data
Explorative Data Analysis
Data Visualization
Regression Analysis
Dataset overview
Data preprocessing
Boxplots
Barplot
Linear regression
Logistic Regression



1. Introduction & Data

Objective of the analysis | Data

Objective of the analysis

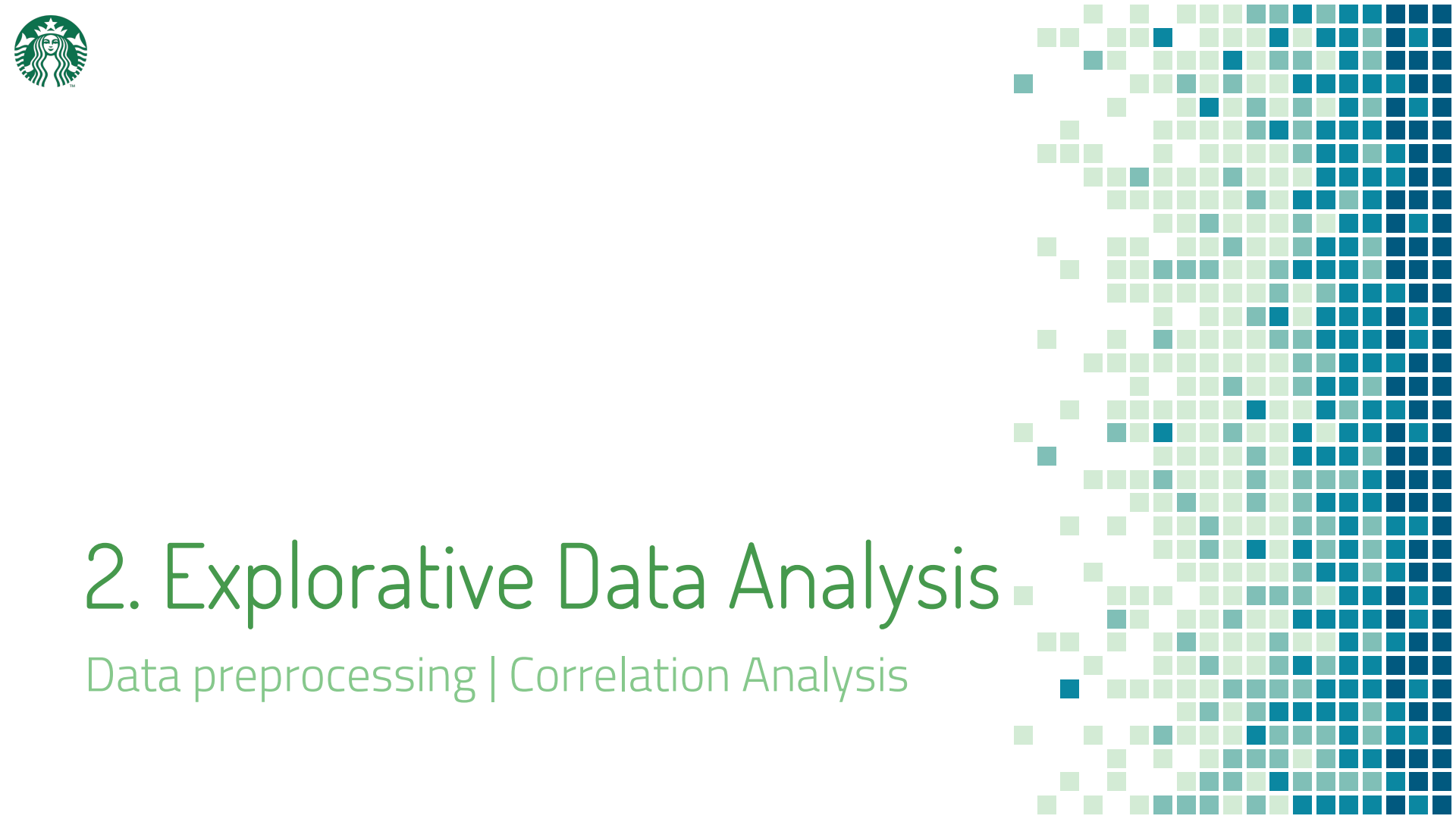
In this project, we conducted a thorough analysis of the ***Starbucks Beverage Components*** dataset, which contains information about the ingredients in *Starbucks* beverages. Our goal was to gain a comprehensive understanding of the data and build models for accurate predictions.

Dataset

The dataset we analyzed is the *Starbucks Beverage Components* dataset from **Kaggle**. This dataset provides a comprehensive guide to the nutritional content of beverages available on the *Starbucks*™ menu.

242
Samples

18
Features



2. Explorative Data Analysis

Data preprocessing | Correlation Analysis

Problems with the data & data preprocessing

Data Cleaning

We transformed our raw data into numeric values and renamed the columns to ensure easy comprehension of the variables.

NA's

We found some NA values in the Caffeine column and replaced these values with the median of the variable to maintain the distribution.

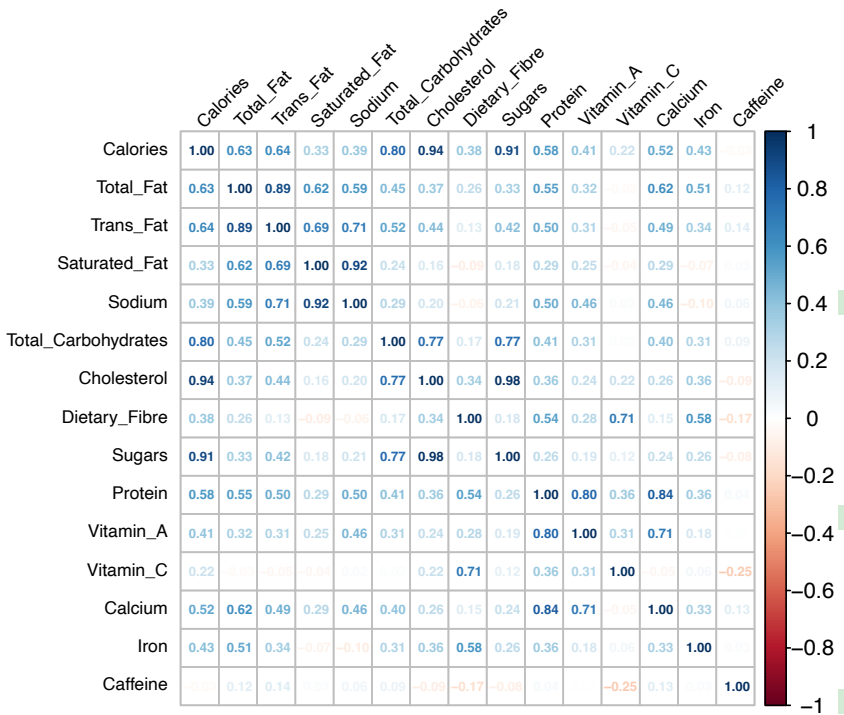
Multicollinearity

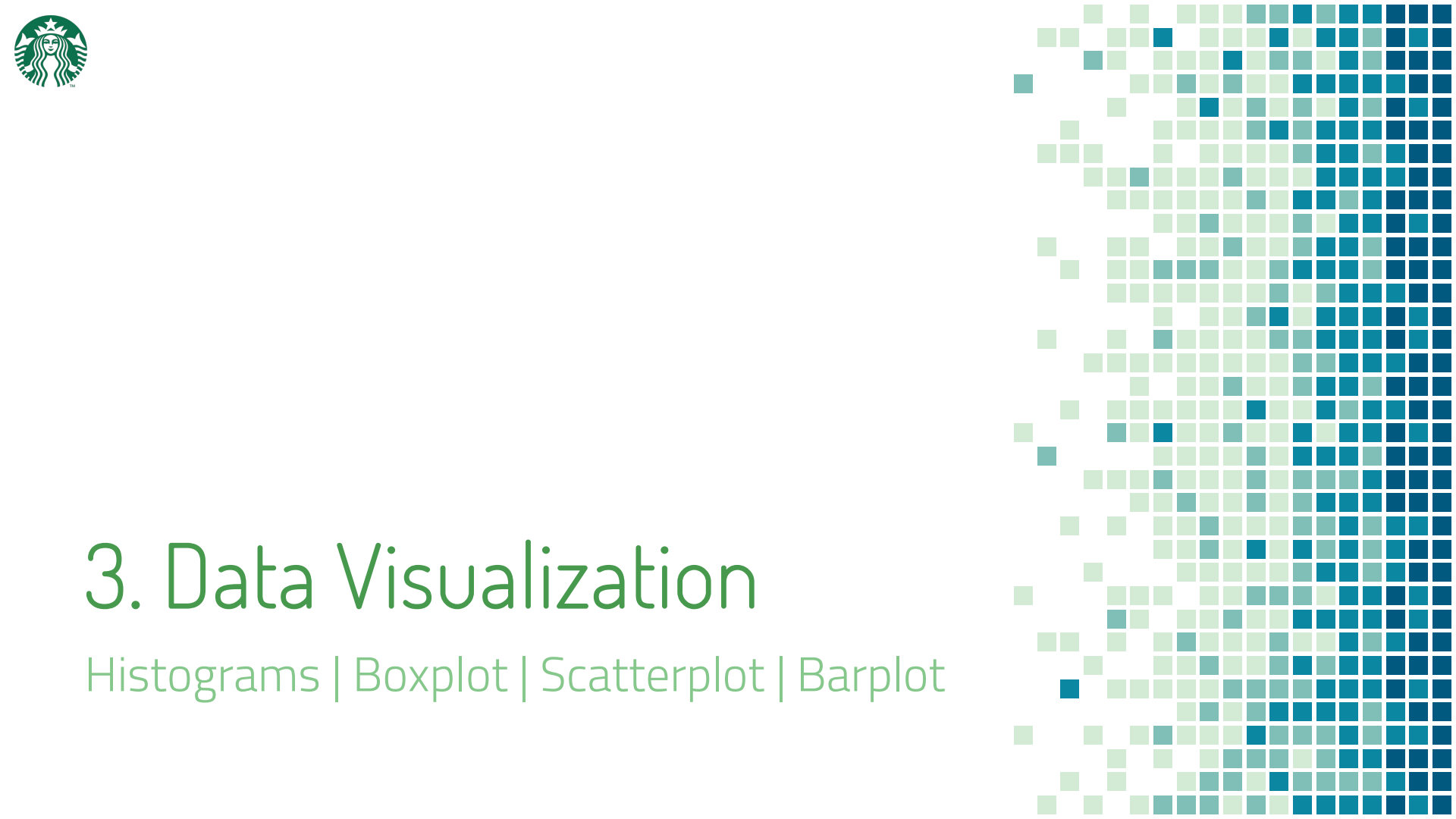
We noted a problem with multicollinearity in our data, which we addressed during the regression analysis.

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.

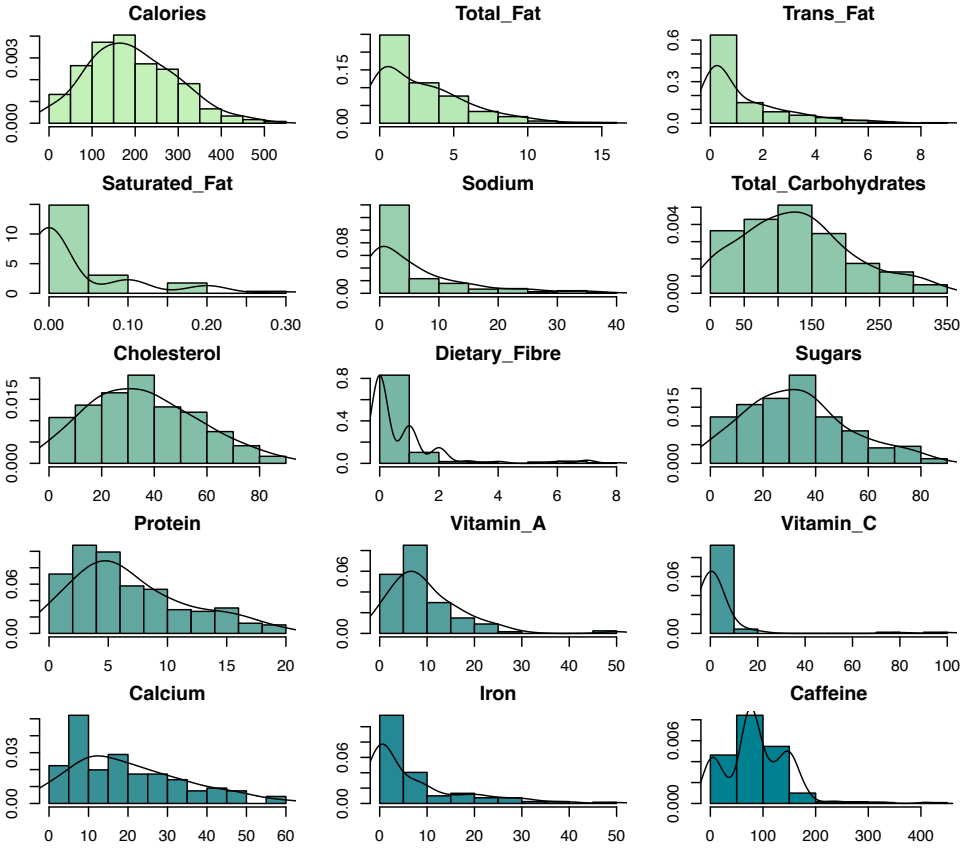




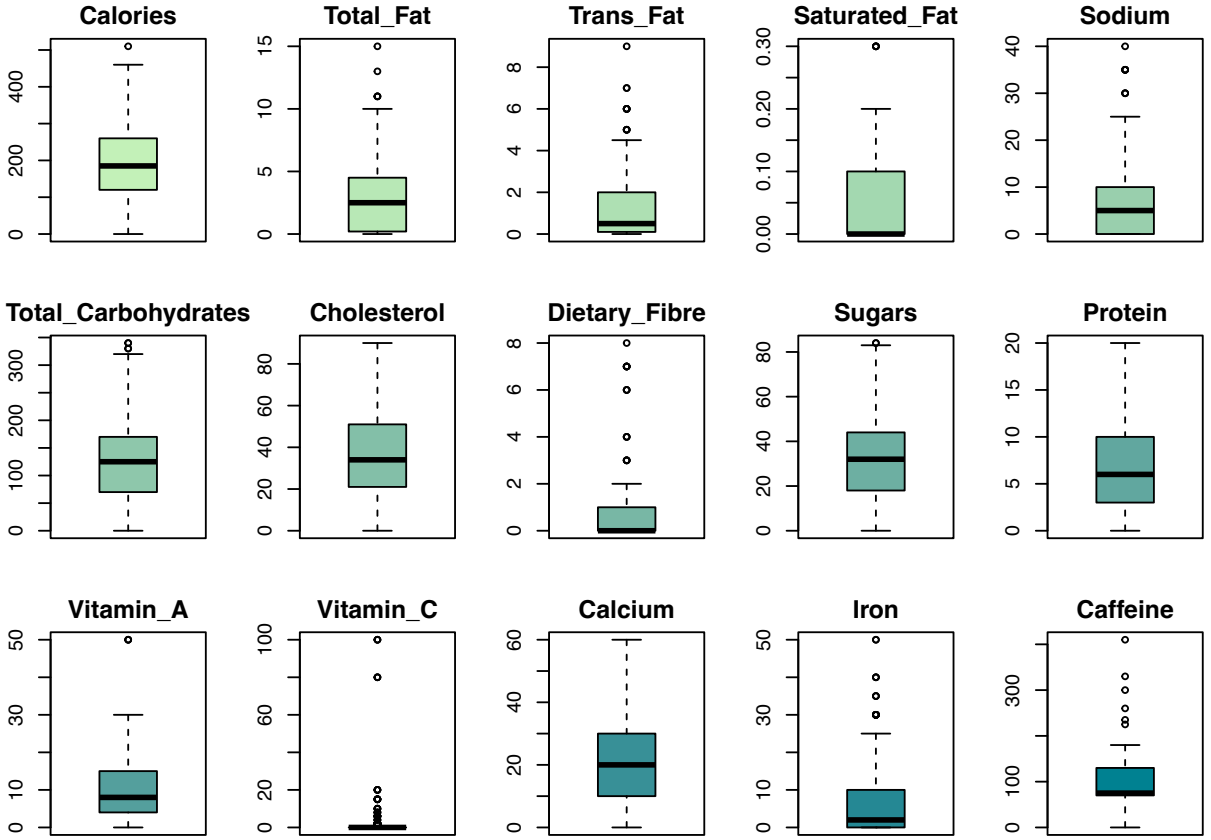
3. Data Visualization

Histograms | Boxplot | Scatterplot | Barplot

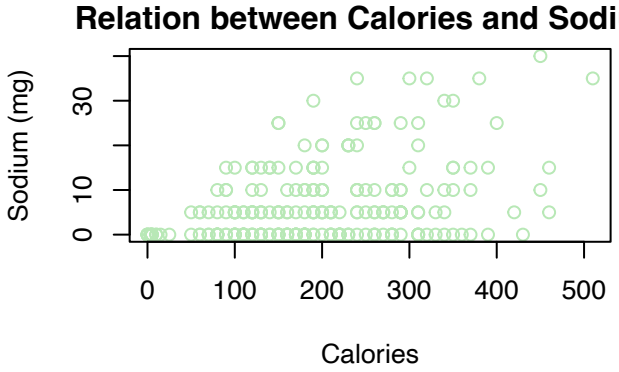
Histograms

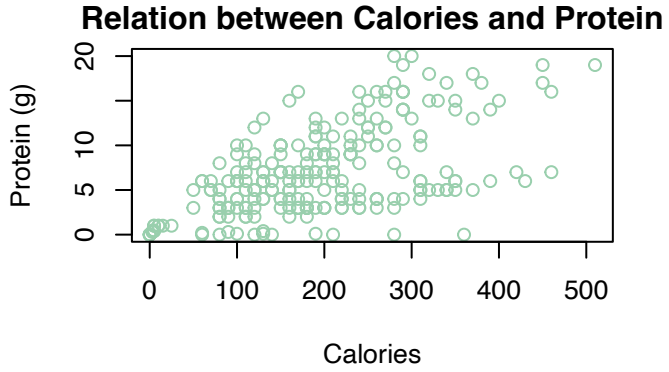

Calories
Total_Fat
Trans_Fat
Saturated_Fat
Sodium
Total_Carbohydrates
Cholesterol
Dietary_Fibre
Sugars
Protein
Vitamin_A
Vitamin_C
Calcium
Iron
Caffeine

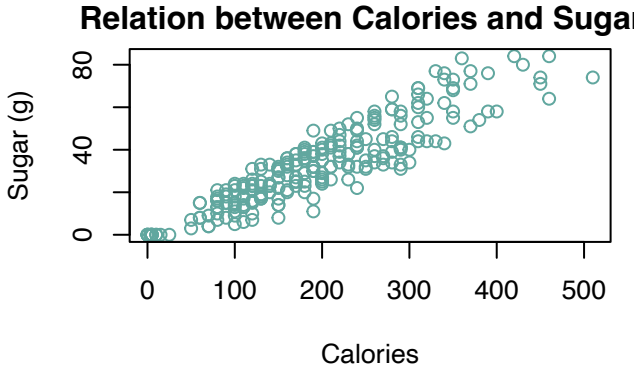
Boxplot

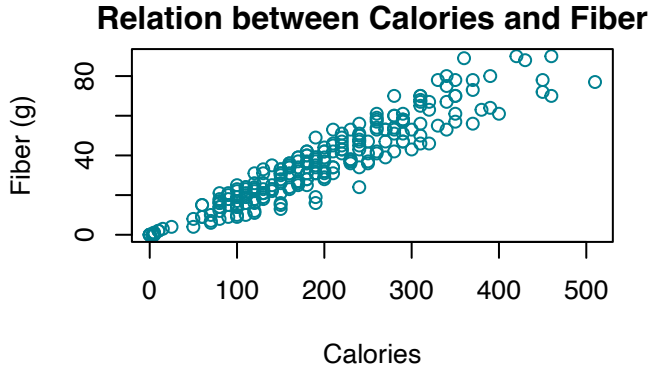

Calories
0
200
400
Total_Fat
0
5
10
15
Trans_Fat
0
2
4
6
8
Saturated_Fat
0.00
0.10
0.20
0.30
Sodium
0
10
20
30
40
Total_Carbohydrates
0
100
200
300
Cholesterol
0
20
40
60
80
Dietary_Fibre
0
2
4
6
8
Sugars
0
20
40
60
80
Protein
0
5
10
15
20
Vitamin_A
0
10
30
50
Vitamin_C
0
20
60
100
Calcium
0
20
40
60
Iron
0
10
30
50
Caffeine
0
100
300

Scatterplot

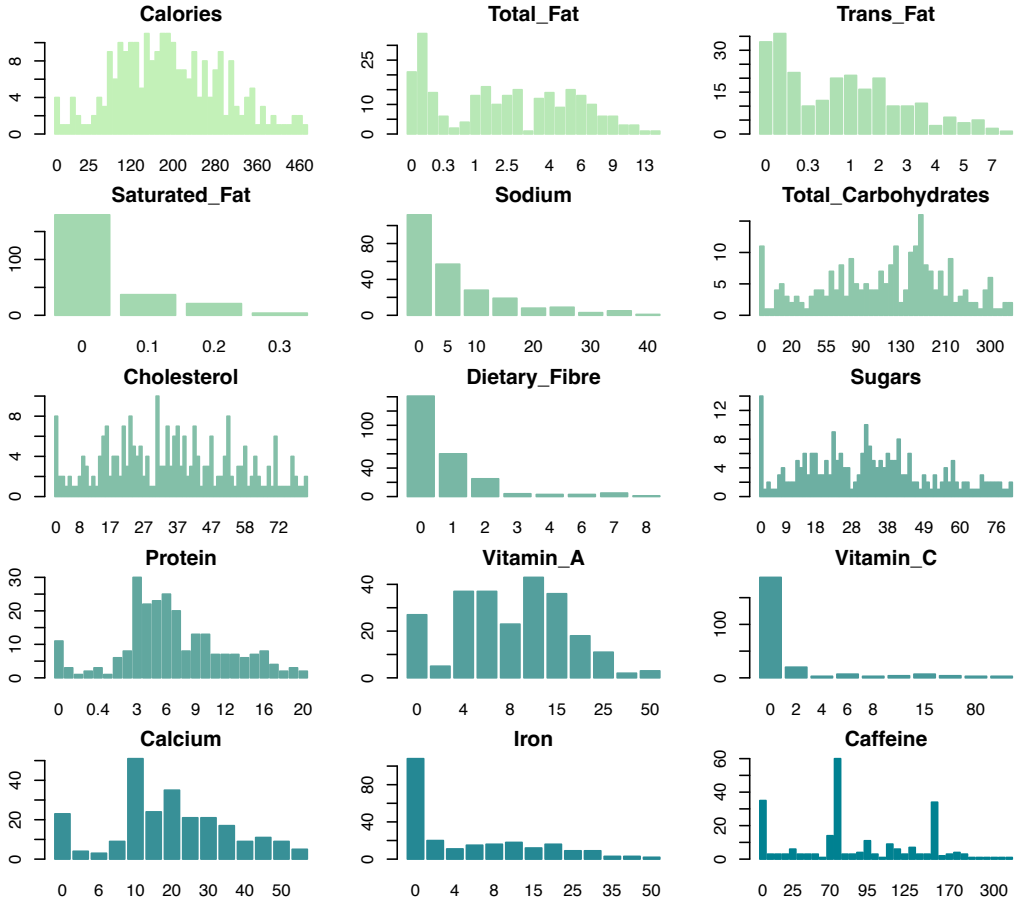








Barplot

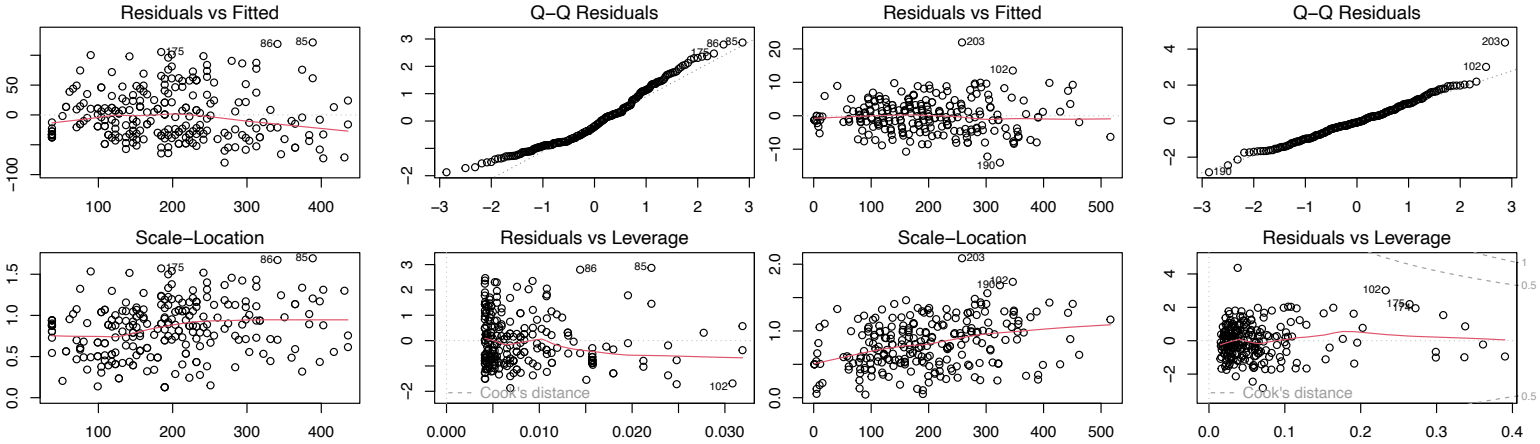

Calories
0 4 8
0 25 120 200 280 360 460
Total_Fat
0 10 25
0 0.3 1 2.5 4 6 9 13
Trans_Fat
0 15 30
0 0.3 1 2 3 4 5 7
Saturated_Fat
0 100
0 0.1 0.2 0.3
Sodium
0 40 80
0 5 10 20 30 40
Total_Carbohydrates
0 5 10
0 20 55 90 130 210 300
Cholesterol
0 4 8
0 8 17 27 37 47 58 72
Dietary_Fibre
0 40 100
0 1 2 3 4 6 7 8
Sugars
0 4 8 12
0 9 18 28 38 49 60 76
Protein
0 10 20 30
0 0.4 3 6 9 12 16 20
Vitamin_A
0 20 40
0 4 8 15 25 50
Vitamin_C
0 100
0 2 4 6 8 15 80
Calcium
0 20 40
0 6 10 20 30 40 50
Iron
0 40 80
0 4 8 15 25 35 50
Caffeine
0 20 40 60
0 25 70 95 125 170 300

4. Regression Analysis

Linear Regression | Logistic Regression

Linear Regression

Simple and Multiple

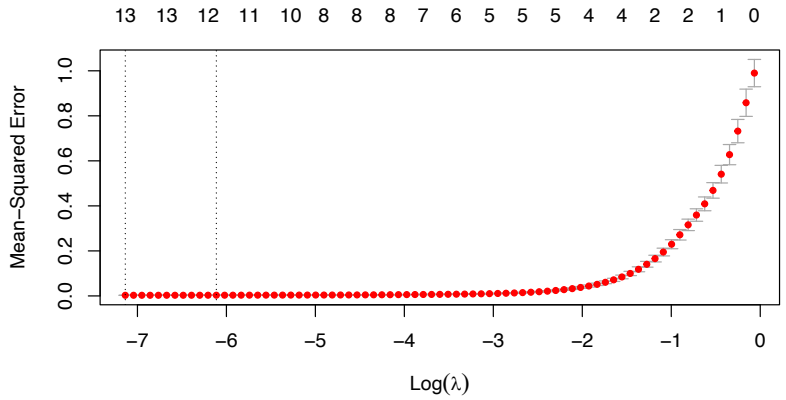


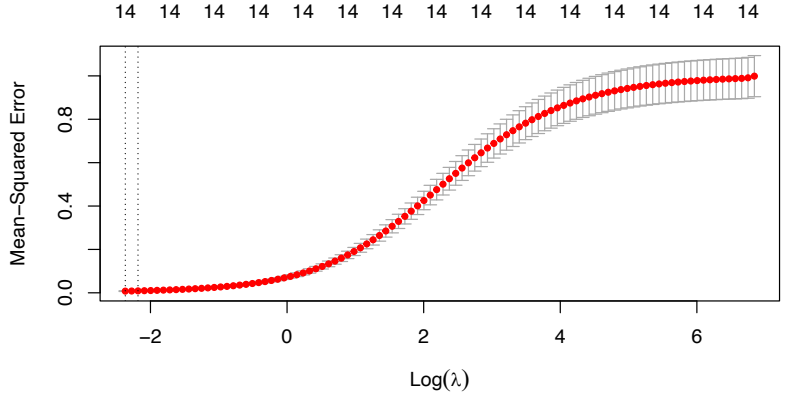
	AIC	BIC	R^2	Adjusted R^2
Simple linear regression	**2509**	**2519**	**0.827**	**0.826**

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

Lasso and Ridge Regression

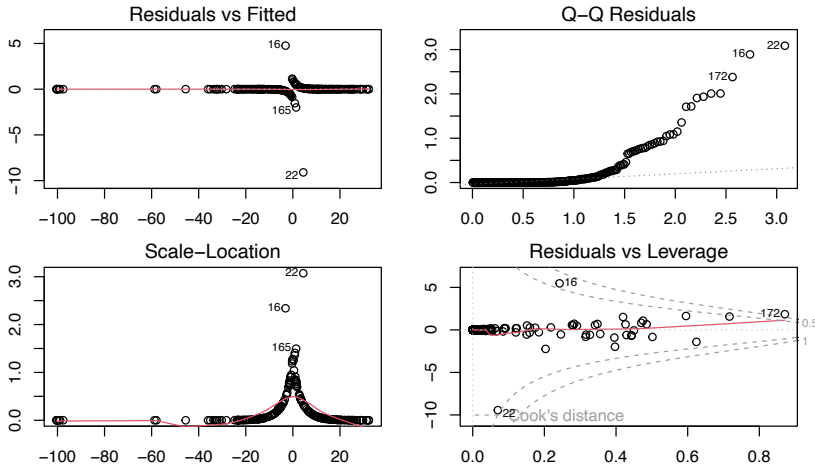
	R^2	MSE
Lasso Regression	**0.9975**	**0.0024**
Ridge Regression	**0.9941**	**0.0066**





Logistic Regression

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

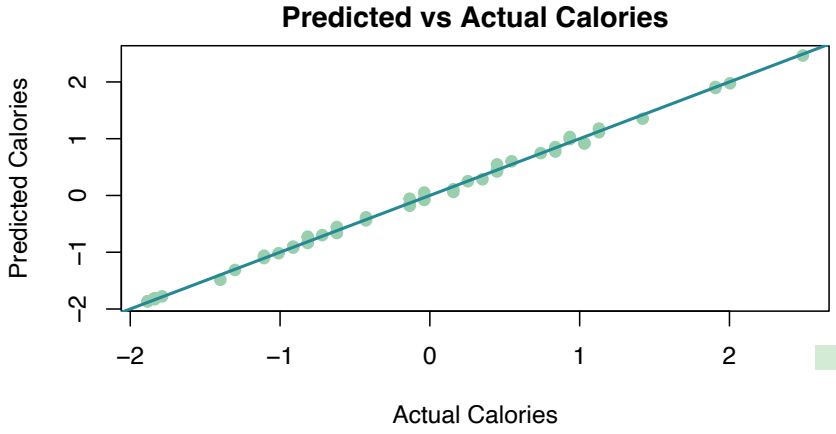


	AIC	BIC	R^2	Residual Deviance	Null Deviance
Multiple linear regression	**69.42**	**121.75**	**0.88**	**39.42**	**335.48**

Cross Validation

Lasso Regression Model

We decided to split the data, allocating 80% of the examples for training and 20% for testing. We then evaluated the model using the testing set and calculated the mean squared error and the root mean squared error to assess its 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