Dataset exploration

Luigi Borzì

To do

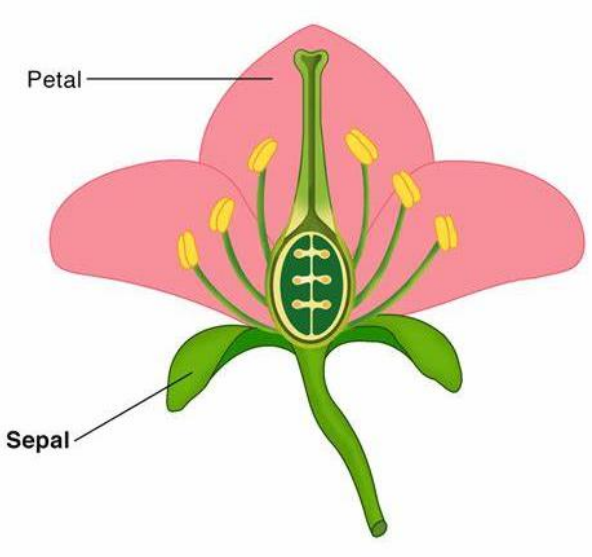
For each dataset, we will explore the data and visualize:

- Composition
- Distribution of variables
- Correlation between variables

Choose the appropriate method for each question.

1.Iris dataset

sepal_length	sepal_width	petal_length	petal_width	species
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
7	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
5.8	2.7	5.1	1.9	virginica



Number of variables: 4

Number of classes: 3

1.Iris dataset: questions

- How many different flowers in the dataset?
- What is the most represented class?
- What is the distribution of each variable?
- Are the distributions from different variables different?
- Is there any correlation between variables?
- Compare flower characteristics for each class
- Compare flower characteristics between classes

2.Cereal dataset

name	manufacturer	type (cold/hot)	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
100% Bran	N	C	70	4	1	130	10	5	6	280	25	3	1	0.33	68.40297
100% Natural Bran	Q	C	120	3	5	15	2	8	8	135	0	3	1	1	33.98368
All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1	0.33	59.42551
All-Bran with Extra Fiber	K	C	50	4	0	140	14	8	0	330	25	3	1	0.5	93.70491
Almond Delight	R	C	110	2	2	200	1	14	8	-1	25	3	1	0.75	34.38484

Number of variables: 15

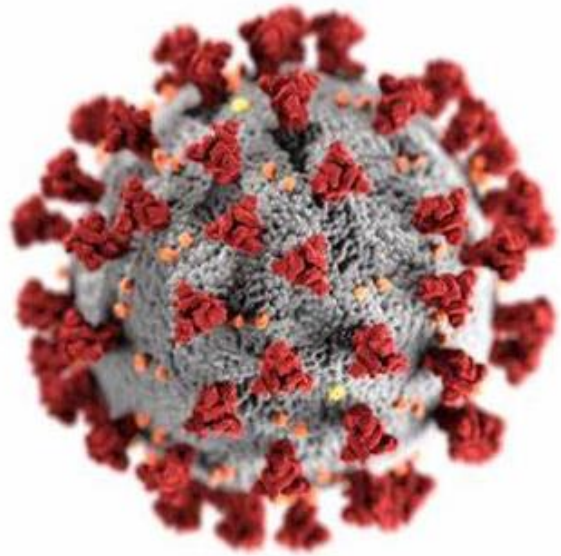
Response variable: rating

2.Cereal dataset: questions

- How many different cereals in the dataset?
- What is the distribution of each variable?
- Are the distributions from different variables different?
- Is there any correlation between variables?
- Is there any correlation between variables and response column?
- Compare different cereal characteristics
- What is the characteristic that most influence the rating?

3.Covid dataset

Country/ Region	Confirmed	Deaths	Recovered	Active	New cases	New deaths	New recovered	Deaths / 100 Cases	Recovered / 100 Cases	Deaths / 100 Recovered	Confirmed last week	1 week change	1 week % increase	WHO Region	Country/ Region
Afghanistan	36263	1269	25198	9796	106	10	18	3.5	69.49	5.04	35526	737	2.07	Eastern Mediterranean	Afghanistan
Albania	4880	144	2745	1991	117	6	63	2.95	56.25	5.25	4171	709	17	Europe	Albania
Algeria	27973	1163	18837	7973	616	8	749	4.16	67.34	6.17	23691	4282	18.07	Africa	Algeria
Andorra	907	52	803	52	10	0	0	5.73	88.53	6.48	884	23	2.6	Europe	Andorra
Angola	950	41	242	667	18	1	0	4.32	25.47	16.94	749	201	26.84	Africa	Angola



Number of variables: 16

Response variable: none

3.Covid dataset: questions

- What is the distribution of deaths?
- What is the correlation between recovered and deaths?
- What is the distribution of WHO regions?
- Is there any correlation between variables?
- Is there any correlation between variables?
- Compare the characteristics of different countries