Don Bosco Institute of Technology, Mumbai 400070
Department of Information Technology

Name: Shazmeen Shaikh

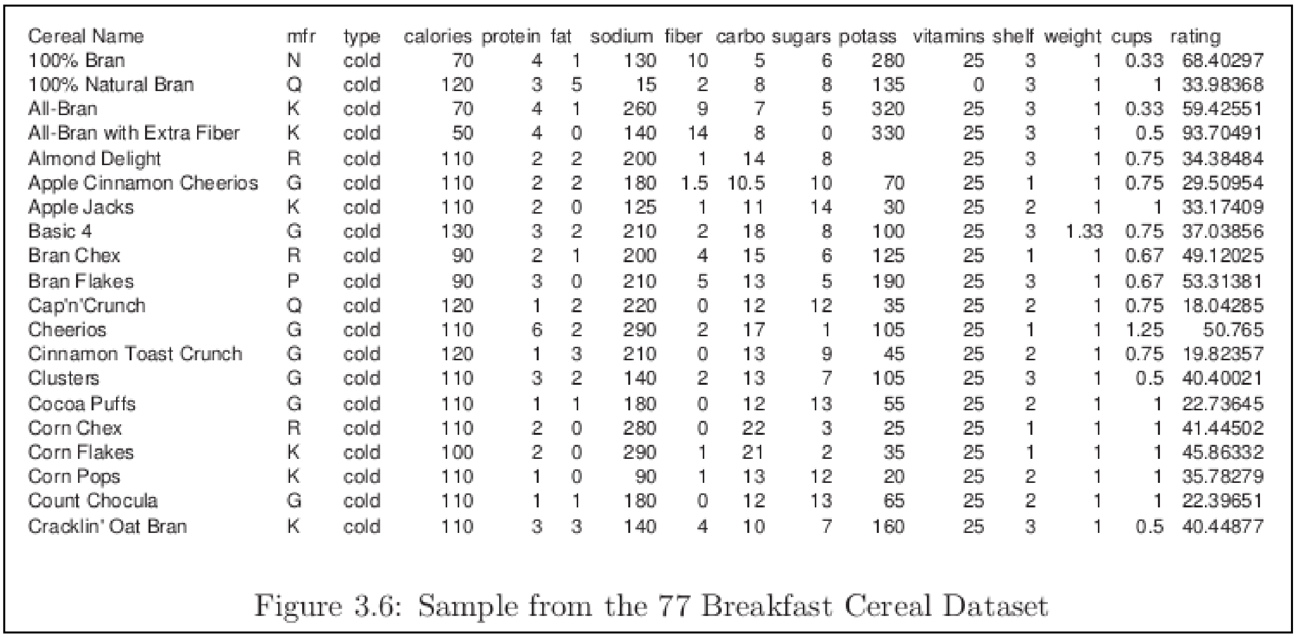
Roll No: 49

Date: 18/02/2024

Experiment No.: 2

Title: Tutorial-2 Solving exercises in Data Exploration

Dataset:



Cereal Name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
100% Bran	N	cold	70	4	1	130	10	5	6	280	25	3	1	0.33	68.40297
100% Natural Bran	Q	cold	120	3	5	15	2	8	8	135	0	3	1	1	33.98368
All-Bran	K	cold	70	4	1	260	9	7	5	320	25	3	1	0.33	59.42551
All-Bran with Extra Fiber	K	cold	50	4	0	140	14	8	0	330	25	3	1	0.5	93.70491
Almond Delight	R	cold	110	2	2	200	1	14	8		25	3	1	0.75	34.38484
Apple Cinnamon Cheerios	G	cold	110	2	2	180	1.5	10.5	10	70	25	1	1	0.75	29.50954
Apple Jacks	K	cold	110	2	0	125	1	11	14	30	25	2	1	1	33.17409
Basic 4	G	cold	130	3	2	210	2	18	8	100	25	3	1.33	0.75	37.03856
Bran Chex	R	cold	90	2	1	200	4	15	6	125	25	1	1	0.67	49.12025
Bran Flakes	P	cold	90	3	0	210	5	13	5	190	25	3	1	0.67	53.31381
Cap'n'Crunch	Q	cold	120	1	2	220	0	12	12	35	25	2	1	0.75	18.04285
Cheerios	G	cold	110	6	2	290	2	17	1	105	25	1	1	1.25	50.765
Cinnamon Toast Crunch	G	cold	120	1	3	210	0	13	9	45	25	2	1	0.75	19.82357
Clusters	G	cold	110	3	2	140	2	13	7	105	25	3	1	0.5	40.40021
Cocoa Puffs	G	cold	110	1	1	180	0	12	13	55	25	2	1	1	22.73645
Corn Chex	R	cold	110	2	0	280	0	22	3	25	25	1	1	1	41.44502
Corn Flakes	K	cold	100	2	0	290	1	21	2	35	25	1	1	1	45.86332
Corn Pops	K	cold	110	1	0	90	1	13	12	20	25	2	1	1	35.78279
Count Chocula	G	cold	110	1	1	180	0	12	13	65	25	2	1	1	22.39651
Cracklin' Oat Bran	K	cold	110	3	3	140	4	10	7	160	25	3	1	0.5	40.44877

Figure 3.6: Sample from the 77 Breakfast Cereal Dataset

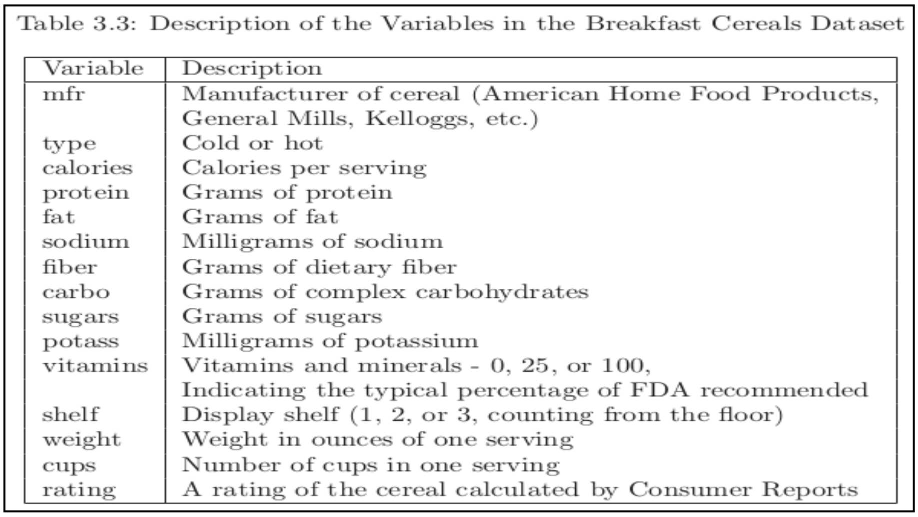


Table 3.3: Description of the Variables in the Breakfast Cereals Dataset

Variable	Description
mfr	Manufacturer of cereal (American Home Food Products, General Mills, Kelloggs, etc.)
type	Cold or hot
calories	Calories per serving
protein	Grams of protein
fat	Grams of fat
sodium	Milligrams of sodium
fiber	Grams of dietary fiber
carbo	Grams of complex carbohydrates
sugars	Grams of sugars
potass	Milligrams of potassium
vitamins	Vitamins and minerals - 0, 25, or 100, Indicating the typical percentage of FDA recommended
shelf	Display shelf (1, 2, or 3, counting from the floor)
weight	Weight in ounces of one serving
cups	Number of cups in one serving
rating	A rating of the cereal calculated by Consumer Reports

Use the data for the breakfast cereals example of section 3.7 of [1] to explore and summarize the data as follows:

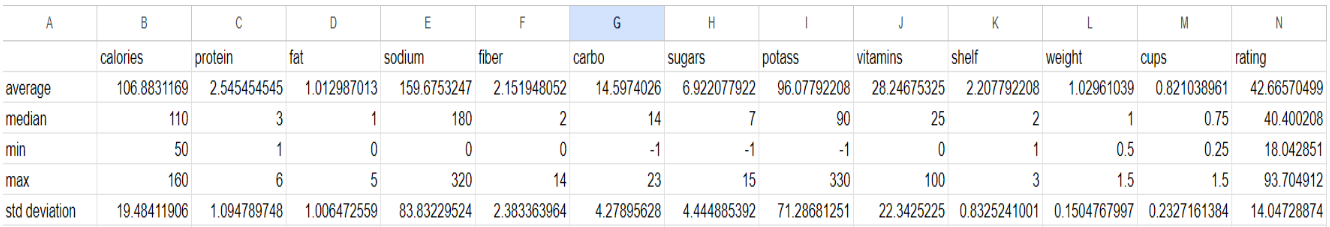
1. Which variables are quantitative/numeric? Which are ordinal? Which are nominal?

Quantitative/Numeric Variables: Calories, protein, fat, sodium, fiber, carbo, sugars, potass, vitamins, shelf, weight, cups, and rating are quantitative/numeric.

Ordinal: The "shelf" variable, which represents the display shelf (1, 2, or 3, counting from the floor), is ordinal.

Nominal: The "name," "mfr," and "type" variables are nominal.

2. Create a table with the average, median, min, max, and standard deviation for each of the quantitative variables.



A	B	C	D	E	F	G	H	I	J	K	L	M	N
	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
average	106.8831169	2.545454545	1.012987013	159.6753247	2.151948052	14.5974026	6.922077922	96.07792208	28.24675325	2.207792208	1.02961039	0.821038961	42.66570499
median	110	3	1	180	2	14	7	90	25	2	1	0.75	40.400208
min	50	1	0	0	0	-1	-1	-1	0	1	0.5	0.25	18.042851
max	160	6	5	320	14	23	15	330	100	3	1.5	1.5	93.704912
std deviation	19.48411906	1.094789748	1.006472559	83.83229524	2.383363964	4.27895628	4.444885392	71.28681251	22.3425225	0.8325241001	0.1504767997	0.2327161384	14.04728874

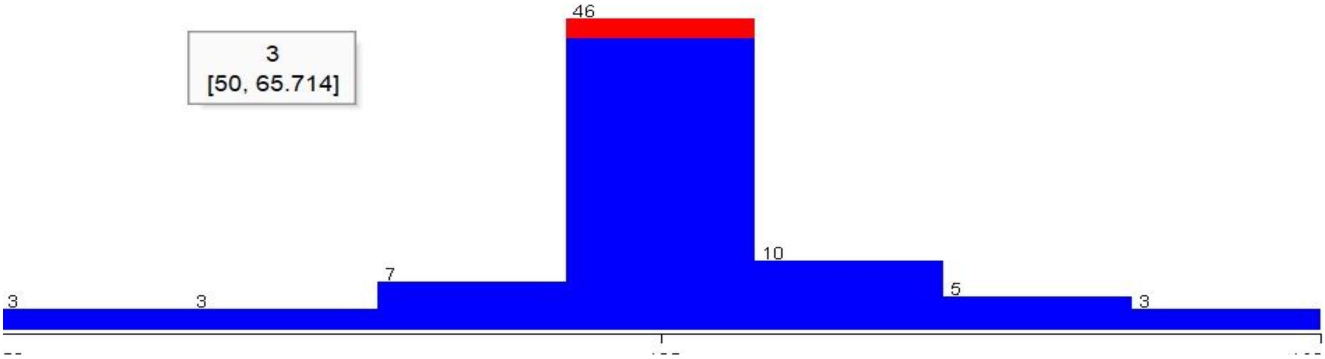
3. Use XLMiner/WEKA to plot a histogram for each of the quantitative variables. Based on the histograms and summary statistics, answer the following questions:

(a) Which variables have the largest variability?
Sodium, potass, vitamins, calories, rating

(b) Which variables seem skewed?
Fat, fibre, potass, rating

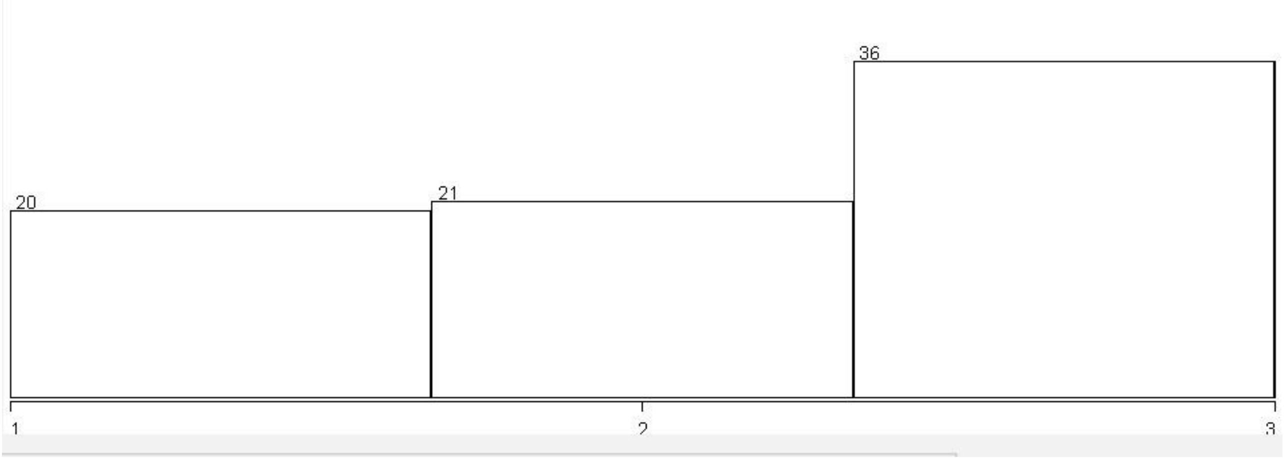
(c) Are there any values that seem extreme?
Calories, vitamins, weight

4. Use XLMiner/WEKA to plot a side-by-side boxplot comparing the calories in hot vs. cold cereals. What does this plot show us?



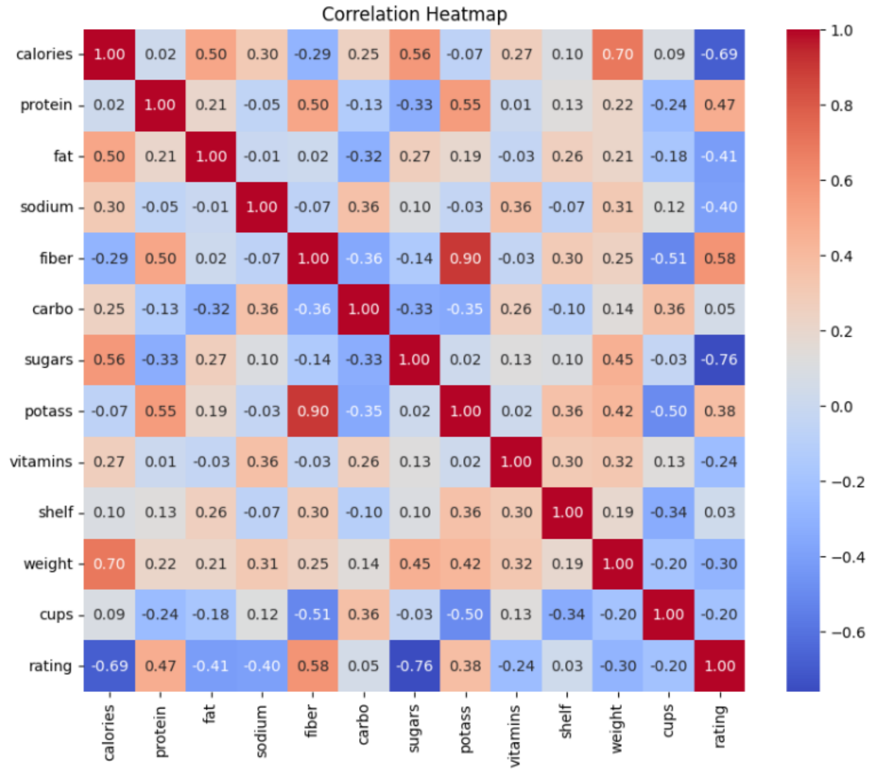
This indicates that the median value of hot cereals is higher than that of cold cereals. It also shows that cold cereals are more distributed as compared to hot cereals.

5. Use XLMiner/WEKA to plot a side-by-side boxplot of consumer rating as a function of the shelf
height. If we were to predict consumer rating from shelf height, does it appear that we need to keep all three categories (1,2,3) of shelf height?



This suggests that there is a positive correlation between shelf height and consumer rating, with higher shelves generally receiving higher ratings. The analysis of the provided data indicates that shelf height does have an impact on consumer ratings. Therefore, it may be necessary to keep all three categories of shelf height (1, 2, and 3) when predicting consumer ratings.

6. Compute the correlation table for the quantitative variable (use Libre's office’s Data > Statistics > Correlation menu). In addition, use XLMiner/WEKA to generate a matrix plot for these variables.



	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
calories	1.00	0.02	0.50	0.30	-0.29	0.25	0.56	-0.07	0.27	0.10	0.70	0.09	-0.69
protein	0.02	1.00	0.21	-0.05	0.50	-0.13	-0.33	0.55	0.01	0.13	0.22	-0.24	0.47
fat	0.50	0.21	1.00	-0.01	0.02	-0.32	0.27	0.19	-0.03	0.26	0.21	-0.18	-0.41
sodium	0.30	-0.05	-0.01	1.00	-0.07	0.36	0.10	-0.03	0.36	-0.07	0.31	0.12	-0.40
fiber	-0.29	0.50	0.02	-0.07	1.00	-0.36	-0.14	0.90	-0.03	0.30	0.25	-0.51	0.58
carbo	0.25	-0.13	-0.32	0.36	-0.36	1.00	-0.33	-0.35	0.26	-0.10	0.14	0.36	0.05
sugars	0.56	-0.33	0.27	0.10	-0.14	-0.33	1.00	0.02	0.13	0.10	0.45	-0.03	-0.76
potass	-0.07	0.55	0.19	-0.03	0.90	-0.35	0.02	1.00	0.02	0.36	0.42	-0.50	0.38
vitamins	0.27	0.01	-0.03	0.36	-0.03	0.26	0.13	0.02	1.00	0.30	0.32	0.13	-0.24
shelf	0.10	0.13	0.26	-0.07	0.30	-0.10	0.10	0.36	0.30	1.00	0.19	-0.34	0.03
weight	0.70	0.22	0.21	0.31	0.25	0.14	0.45	0.42	0.32	0.19	1.00	-0.20	-0.30
cups	0.09	-0.24	-0.18	0.12	-0.51	0.36	-0.03	-0.50	0.13	-0.34	-0.20	1.00	-0.20
rating	-0.69	0.47	-0.41	-0.40	0.58	0.05	-0.76	0.38	-0.24	0.03	-0.30	-0.20	1.00

(a) Which pair of variables is most strongly correlated?

Fibre and Potass is the most strongly correlated pair.

(b) How can we reduce the number of variables based on these correlations?

To reduce the number of variables based on correlations, we can identify and remove highly correlated variables (with correlation coefficients close to 1 or -1) as they provide redundant information, retaining only one variable from each highly correlated pair to simplify the dataset while preserving essential information.

(c) How would the correlations change if we normalized the data first?

Normalizing the data before calculating correlations would standardize the scales of the variables, potentially altering the correlation values. This normalization process could lead to changes in the magnitude and direction of correlations, particularly if variables originally had vastly different scales or units. However, the relative strength of relationships between variables is likely to remain similar, with the overall patterns and associations between variables preserved.

References:

1) G. Shmueli, N.R. Patel, P.C. Bruce, "Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner", 2nd Edition, Wiley India.
2) http://www.wekaleamstudios.co.uk/posts/summarising-data-using-box-and-whisker-plots/
3) https://colab.research.google.com/
4) https://docs.google.com/spreadsheets/u/0/