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

Name: Shruti Chaube
Roll no: 66
Date: 01 -03-21

Experiment No. : 1

Title : Tutorial-1 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?
Ans. **Quantitative / Numeric** :- Calories, protein, fat, sodium, fiber, carbo, sugars, potass, vitamins, shelf, weight, cups, rating.
Nominal :- Cereal name, mfr, type.

2. Create a table with the average, median, min, max, and standard deviation for each of the quantitative variables.
This can be done through Libre office's functions – Libre's office's Data > Statistics > Descriptive Statistics menu.

Ans.

	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	weight	cups	rating
Average	103	3	1	185	3	13	7	117	24	2	1	1	41
Median	110	2	1	190	2	13	8	100	25	2	1	1	39
Min	50	1	0	15	0	5	0	20	0	1	1	0	18
Max	130	6	5	290	14	22	14	330	25	3	1	1	94
Standard Deviation	19.762	1.318	1	69.19	3.8	4.36	4.097	98.35	5.5902	0.83	0.074	0.3	18

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:

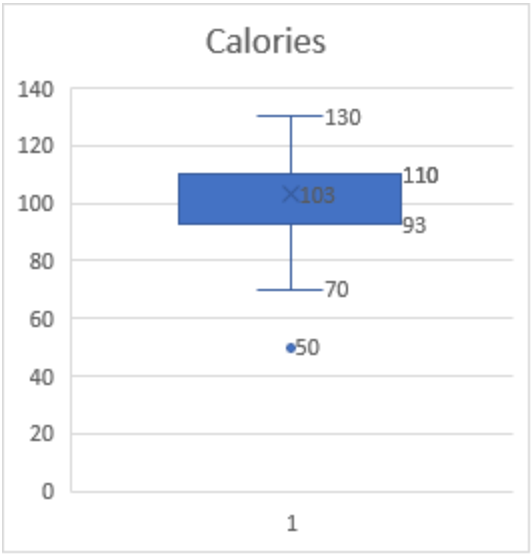
(a) Which variables have the largest variability?
Ans. Sugars, shelf, cups

(b) Which variables seem skewed?
Ans. Negatively skewed :- Calories, Sodium, Sugar, Vitamins, Self, Cups
Positively skewed :- Protein, Fat, Fiber, Carbo, Potass, weight, rating

(c) Are there any values that seem extreme?
Ans. Calories, Protien, sodium, Fiber, carbo, vitamins, weight, Rating have extreme values.

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?

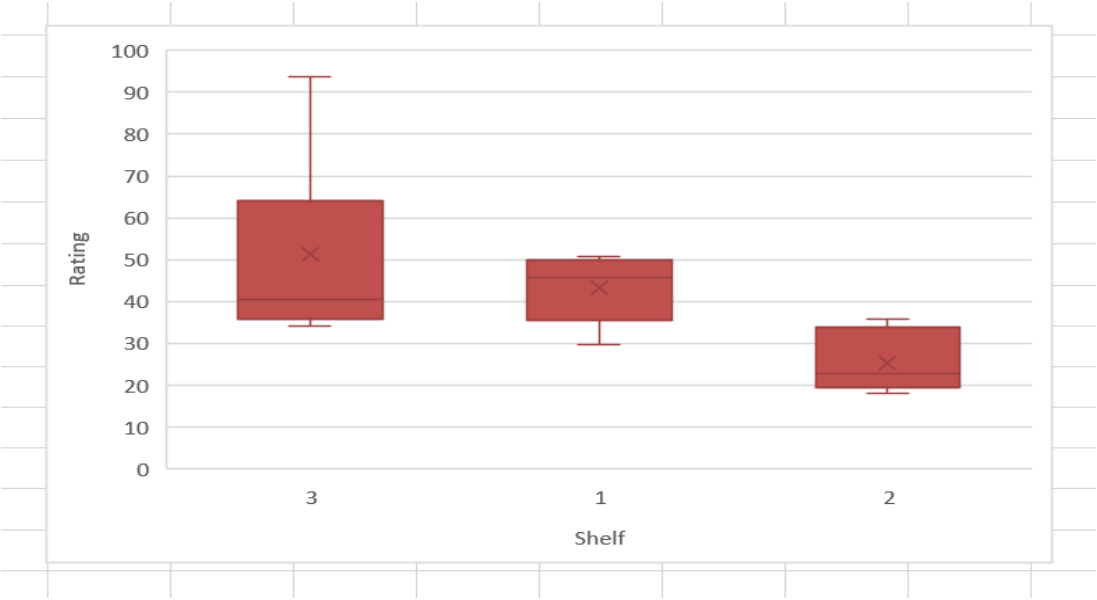
Ans. We don't have data for calories in hot cereals therefore we do not have a second box plot to compare.



This boxplot show us the median, 1st quartile (25th percentile), & 3rd quartile(75th percentile) as well as the outliers present in the data.

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?

Ans. Yes, if we were to predict consumer rating from shelf height, we need to keep all three categories (1,2,3) of shelf height.



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												
protein	-0.444594165	1											
fat	0.478380126	0.091185832	1										
sodium	-0.023867137	0.121219618	-0.353765009	1									
fiber	-0.912517609	0.606809106	-0.222564458	-0.135483193	1								
carbo	0.452597798	-0.176188273	-0.250627637	0.625861194	-0.56757441	1							
sugars	0.528478459	-0.706675009	0.140787555	-0.417857595	-0.54375523	-0.24148389	1						
potass	-0.729184354	0.365943245	-0.04770028	-0.198435464	0.824981135	-0.64031983	-0.23709907	1					
vitamins	-0.202481415	-0.089299953	-0.644916928	0.576658219	0.059862651	0.254884393	-0.03159654	-0.04341196	1				
shelf	-0.230061305	0.239566489	0.259519426	-0.486464138	0.471077321	-0.58465226	0.033905832	0.459016423	-0.225912998	1			
weight	0.321588129	0.089299953	0.107486155	0.086753891	-0.05986265	0.284553476	0.031596544	0.018790251	0.052631579	0.225912998	1		
cups	0.579303714	-0.237193679	-0.01890721	0.123129909	-0.69862903	0.562573798	0.203370387	-0.663655006	-0.208306993	-0.606018658	-0.023145221	1	
rating	-0.882001184	0.703093052	-0.374251008	0.01839087	0.916806771	-0.25263357	-0.75916594	0.634032973	0.09789949	0.274546503	-0.05804699	-0.482326093	1

(a) Which pair of variables is most strongly correlated?
Ans. Rating and Fiber has the highest correlation.

(b) How can we reduce the number of variables based on these correlations?
Ans. Removal of weight, vitamins, sugar, fat.

(c) How would the correlations change if we normalized the data first?
Ans. There is no change in correlation even if we normalize the data first.

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/