Homework 2

Using Knime and/or Excel on the cereal.csv dataset...

- Explore whether there are missing values for any of the variables
- Determine whether there are any outliers among the "Sodium" values
- Normalize the variables “Calories", "Sodium" and "Potassium" (for example using Min-Max transformation)
- Analyze and interpret the correlations between each one of the variables and the variable "Rating"

Output Report:

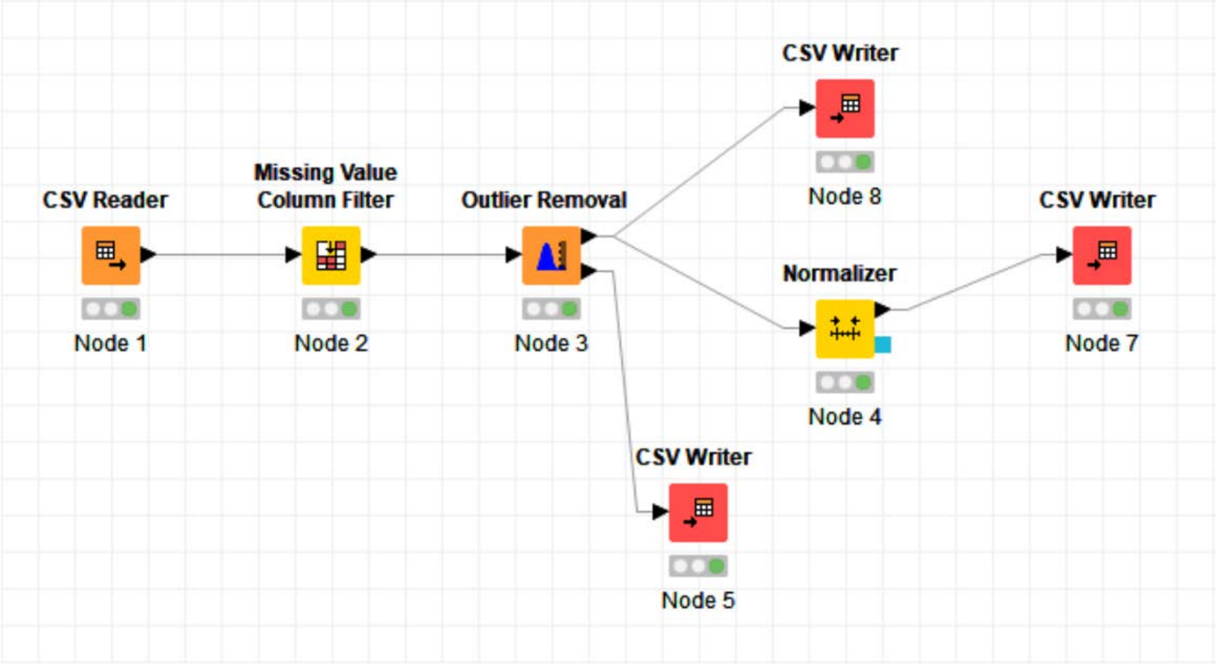
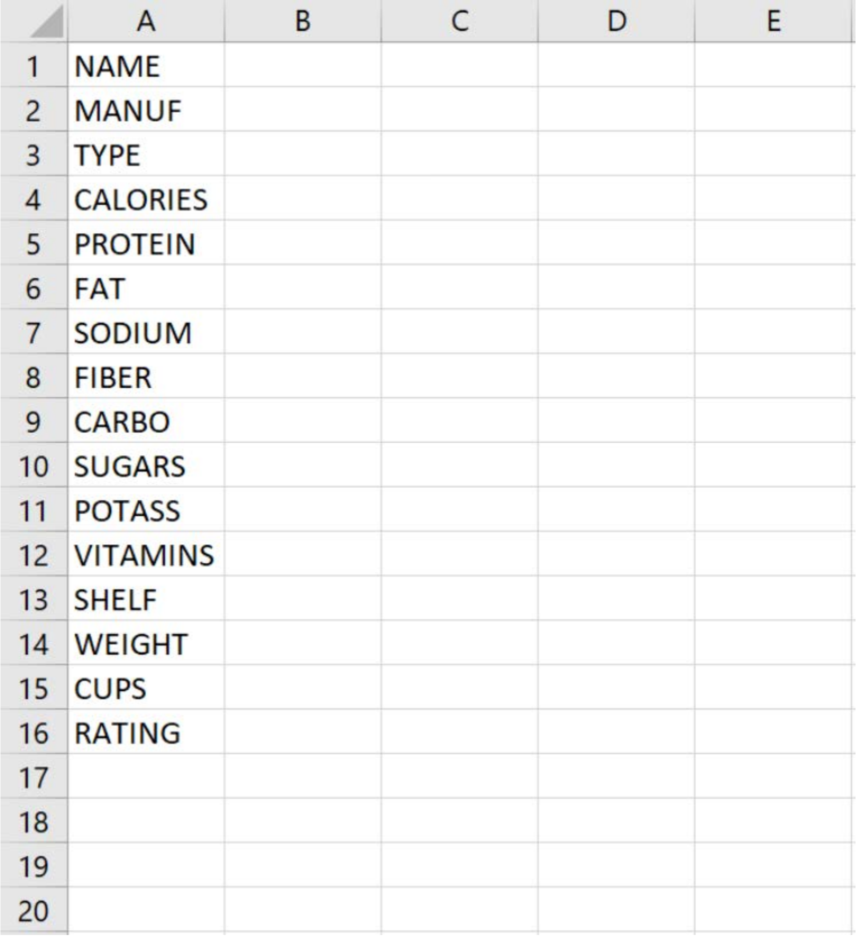


Figure 1 Project Work Board

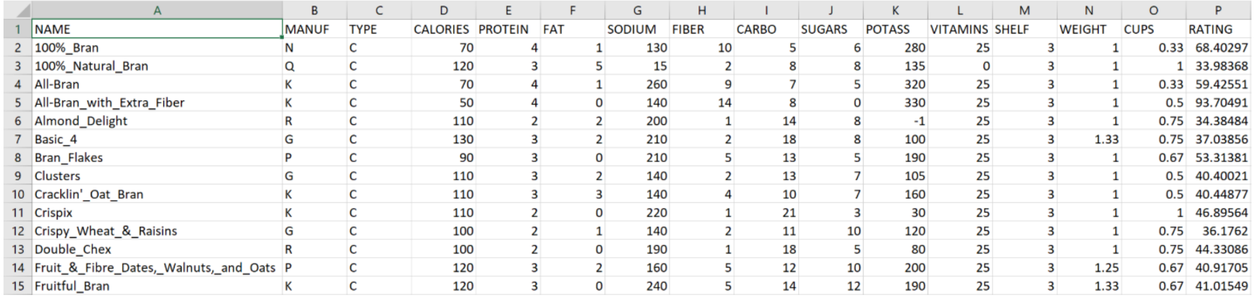
The figure 1 shows the project work board for this assignment. I imported the ‘cereal.csv’ file as an input to ‘CSV Reader’ and was passed through a ‘Missing Value Filter’ with a 90% threshold. After that the csv file was passed through an ‘Outlier Removal’ process for only the column of Sodium. The Outlier were recorded in a CSV file and are shown as in figure 2. There were no recorded outliers from ‘Sodium’.



	A	B	C	D	E
1	NAME				
2	MANUF				
3	TYPE				
4	CALORIES				
5	PROTEIN				
6	FAT				
7	SODIUM				
8	FIBER				
9	CARBO				
10	SUGARS				
11	POTASS				
12	VITAMINS				
13	SHELF				
14	WEIGHT				
15	CUPS				
16	RATING				
17					
18					
19					
20					

Figure 2 Outliers

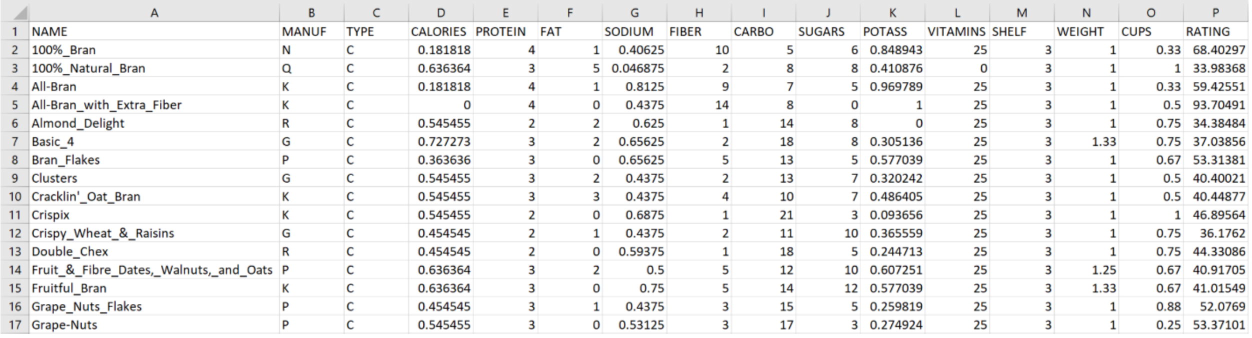
The filtered output from 'Outlier Remover' was also recorded and a part of it is shown in figure 3.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	NAME	MANUF	TYPE	CALORIES	PROTEIN	FAT	SODIUM	FIBER	CARBO	SUGARS	POTASS	VITAMINS	SHELF	WEIGHT	CUPS	RATING
2	100%_Bran	N	C	70	4	1	130	10	5	6	280	25	3	1	0.33	68.40297
3	100%_Natural_Bran	Q	C	120	3	5	15	2	8	8	135	0	3	1	1	33.98368
4	All-Bran	K	C	70	4	1	260	9	7	5	320	25	3	1	0.33	59.42551
5	All-Bran_with_Extra_Fiber	K	C	50	4	0	140	14	8	0	330	25	3	1	0.5	93.70491
6	Almond_Delight	R	C	110	2	2	200	1	14	8	-1	25	3	1	0.75	34.38484
7	Basic_4	G	C	130	3	2	210	2	18	8	100	25	3	1.33	0.75	37.03856
8	Bran_Flakes	P	C	90	3	0	210	5	13	5	190	25	3	1	0.67	53.31381
9	Clusters	G	C	110	3	2	140	2	13	7	105	25	3	1	0.5	40.40021
10	Cracklin'_Oat_Bran	K	C	110	3	3	140	4	10	7	160	25	3	1	0.5	40.44877
11	Crispix	K	C	110	2	0	220	1	21	3	30	25	3	1	1	46.89564
12	Crispy_Wheat_&_Raisins	G	C	100	2	1	140	2	11	10	120	25	3	1	0.75	36.1762
13	Double_Chex	R	C	100	2	0	190	1	18	5	80	25	3	1	0.75	44.33086
14	Fruit_&_Fibre_Dates,_Walnuts,_and_Oats	P	C	120	3	2	160	5	12	10	200	25	3	1.25	0.67	40.91705
15	Fruitful_Bran	K	C	120	3	0	240	5	14	12	190	25	3	1.33	0.67	41.01549

Figure 3 Part of Filtered Output

After filtering the output from through an outlier filter, it was passed to a 'Normalizer' for only the columns of 'Sodium', 'Calories' and 'Potass' using Min-Max Transformation. The normalized output were in the range of 0 to 1. A part of the output is shown in figure 3.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	NAME	MANUF	TYPE	CALORIES	PROTEIN	FAT	SODIUM	FIBER	CARBO	SUGARS	POTASS	VITAMINS	SHELF	WEIGHT	CUPS	RATING
2	100%_Bran	N	C	0.181818	4	1	0.40625	10	5	6	0.848943	25	3	1	0.33	68.40297
3	100%_Natural_Bran	Q	C	0.636364	3	5	0.046875	2	8	8	0.410876	0	3	1	1	33.98368
4	All-Bran	K	C	0.181818	4	1	0.8125	9	7	5	0.969789	25	3	1	0.33	59.42551
5	All-Bran_with_Extra_Fiber	K	C	0	4	0	0.4375	14	8	0	1	25	3	1	0.5	93.70491
6	Almond_Delight	R	C	0.545455	2	2	0.625	1	14	8	0	25	3	1	0.75	34.38484
7	Basic_4	G	C	0.727273	3	2	0.65625	2	18	8	0.305136	25	3	1.33	0.75	37.03856
8	Bran_Flakes	P	C	0.363636	3	0	0.65625	5	13	5	0.577039	25	3	1	0.67	53.31381
9	Clusters	G	C	0.545455	3	2	0.4375	2	13	7	0.320242	25	3	1	0.5	40.40021
10	Cracklin'_Oat_Bran	K	C	0.545455	3	3	0.4375	4	10	7	0.486405	25	3	1	0.5	40.44877
11	Crispix	K	C	0.545455	2	0	0.6875	1	21	3	0.093656	25	3	1	1	46.89564
12	Crispy_Wheat_&_Raisins	G	C	0.454545	2	1	0.4375	2	11	10	0.365559	25	3	1	0.75	36.1762
13	Double_Chex	R	C	0.454545	2	0	0.59375	1	18	5	0.244713	25	3	1	0.75	44.33086
14	Fruit_&_Fibre_Dates,_Walnuts,_and_Oats	P	C	0.636364	3	2	0.5	5	12	10	0.607251	25	3	1.25	0.67	40.91705
15	Fruitful_Bran	K	C	0.636364	3	0	0.75	5	14	12	0.577039	25	3	1.33	0.67	41.01549
16	Grape_Nuts_Flakes	P	C	0.454545	3	1	0.4375	3	15	5	0.259819	25	3	1	0.88	52.0769
17	Grape-Nuts	P	C	0.545455	3	0	0.53125	3	17	3	0.274924	25	3	1	0.25	53.37101

Figure 4 Normalized Output

To analyze the relation for ratings, 'Rank Correlation' node was added [shown in figure 5] and a correlation matrix was developed [shown in figure 6]. From the graph it is clear that the ratings are inversely proportional to 'calories' and 'sugar' and is directly proportional to 'protein' and 'fiber'. There are other relations as well, detailed in the figure 6.

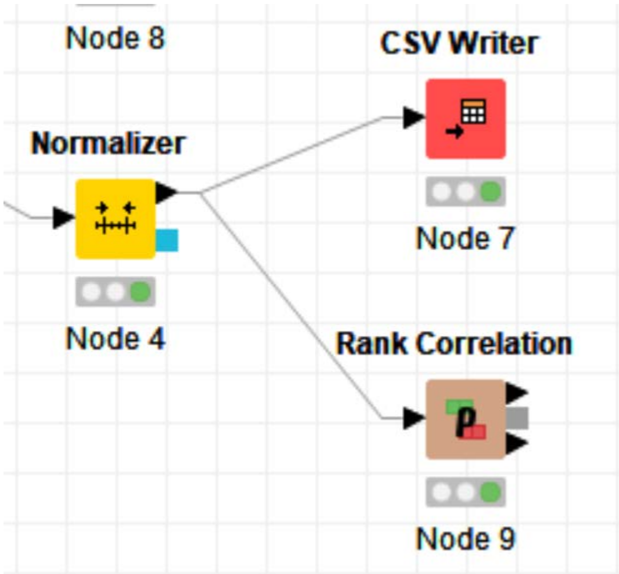


Figure 5 Rank Correlation Node

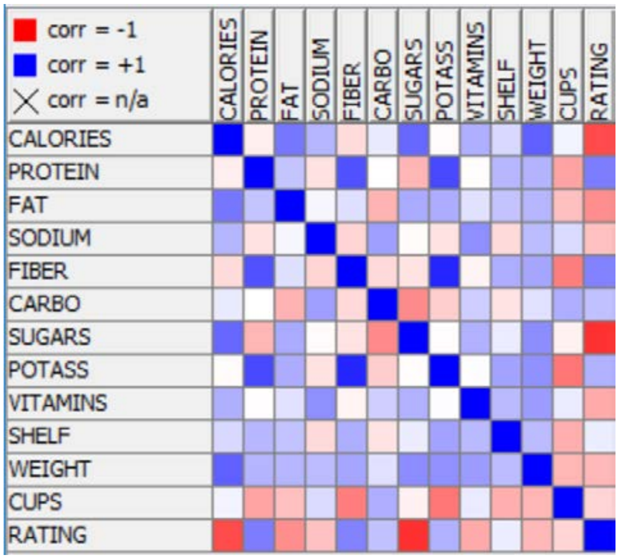


Figure 6 Correlation Matrix