Homework 2

Breakfast Cereals. Use the data *cereal.xlsx* to explore and summarize the data as follows: (note that a few records contain missing values; since there are just a few, a simple solution is to remove them first.)

a. Which variables are quantitative/numerical? Which are ordinal? Which are nominal? (10 point)

b. Create a table with the average, median, min, max, and standard deviation for each of the quantitative variables. This can be done through Excel's functions or Excel's *Data → Data Analysis → Descriptive Statistics* menu. (5 point)

c. Use RapidMiner to plot a histogram for each of the quantitative variables (5 pt). Based on the histograms and summary statistics, answer the following questions:

i. Which variables have the largest variability? (5 pt)

ii. Which variables seem skewed? (5 pt)

iii. Are there any values that seem extreme? (5 pt)

d. Plot a sidebyside boxplot comparing the calories in hot vs. cold cereals. What does this plot show us? (5 pt)

e. Plot a sidebyside 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 of shelf height? (5 pt)

f. Compute the correlation table for the quantitative variable

i. Which pair of variables is most strongly correlated? (2.5 pt)

ii. How can we reduce the number of variables based on these correlations? (2.5 pt)

Hari Vengadesh Elangeswaran

ANSWERS:

a. ***Which variables are quantitative/numerical? Which are ordinal? Which are nominal? (10 point)***

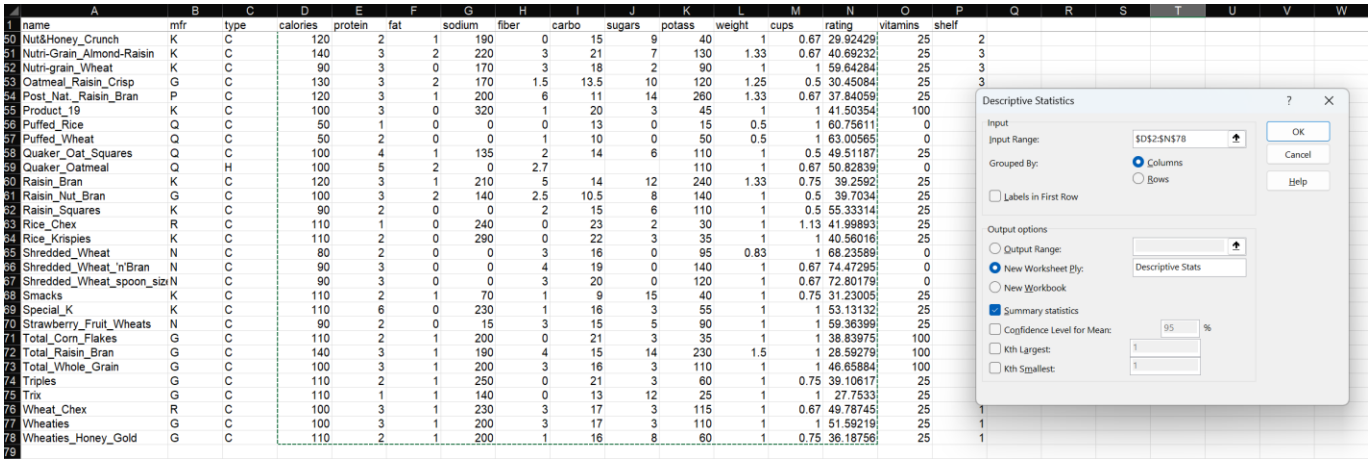
Order	Variable	Type	Description
1	Name	Nominal	Name of cereal
2	mfr	Nominal	Manufacturer (A = American Home Food Products, G = General Mills, etc.)
3	type	Nominal	Type of cereal (cold or hot)
4	calories	Quantitative	Calories per serving
5	protein	Quantitative	Grams of protein
6	fat	Quantitative	Grams of fat
7	sodium	Quantitative	Milligrams of sodium
8	fiber	Quantitative	Grams of dietary fiber
9	carbo	Quantitative	Grams of complex carbohydrates
10	sugars	Quantitative	Grams of sugars
11	potass	Quantitative	Milligrams of potassium
12	vitamins	Ordinal	Vitamins and minerals (0, 25, or 100% of FDA recommendation)
13	shelf	Ordinal	Display shelf position (1 = bottom, 2 = middle, 3 = top)
14	weight	Quantitative	Weight in ounces of one serving
15	cups	Quantitative	Number of cups in one serving
16	rating	Quantitative	Consumer Reports rating of the cereal

---

b. ***Create a table with the average, median, min, max, and standard deviation for each of the quantitative variables. This can be done through Excel's functions or Excel's Data → Data Analysis → Descriptive Statistics menu. (5 point)***

Data Analysis wasn't accessible at first and so I had to do the following to make it available:
File → Options → Add-ins → Manage: Excel Add-ins, and check Analysis ToolPak

To use Descriptive Analysis, I had to move the quantitative variable next each other as the data needed to be contiguous for the Data Analysis tool to work. This created a new sheet with a summary statistics of all the quantitative rows.



name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	weight	cups	rating	vitamins	shelf
Nut&Honey_Crunch	K	C	120	2	1	190	0	15	9	40	1	0.67	29.92429	25	2
Nutri-Grain_Almond-Raisin	K	C	140	3	2	220	3	21	7	130	1.33	0.67	40.69232	25	3
Nutri-grain_Wheat	K	C	90	3	0	170	3	18	2	90	1	1	59.64284	25	3
Oatmeal_Raisin_Crisp	G	C	130	3	2	170	1.5	13.5	10	120	1.25	0.5	30.45084	25	3
Post_Nat._Raisin_Bran	P	C	120	3	1	200	6	11	14	260	1.33	0.67	37.84059	25	
Product_19	K	C	100	3	0	320	1	20	3	45	1	1	41.50354	100	
Puffed_Rice	Q	C	50	1	0	0	0	13	0	15	0.5	1	60.75611	0	
Puffed_Wheat	Q	C	50	2	0	0	1	10	0	50	0.5	1	63.00565	0	
Quaker_Oat_Squares	Q	C	100	4	1	135	2	14	6	110	1	0.5	49.51187	25	
Quaker_Oatmeal	Q	H	100	5	2	0	2.7			110	1	0.67	50.82839	0	
Raisin_Bran	K	C	120	3	1	210	5	14	12	240	1.33	0.75	39.2592	25	
Raisin_Nut_Bran	G	C	100	3	2	140	2.5	10.5	8	140	1	0.5	39.7034	25	
Raisin_Squares	K	C	90	2	0	0	2	15	6	110	1	0.5	55.33314	25	
Rice_Chex	R	C	110	1	0	240	0	23	2	30	1	1.13	41.99893	25	
Rice_Krispies	K	C	110	2	0	290	0	22	3	35	1	1	40.56016	25	
Shredded_Wheat	N	C	80	2	0	0	3	16	0	95	0.83	1	68.23589	0	
Shredded_Wheat_'n'Bran	N	C	90	3	0	0	4	19	0	140	1	0.67	74.47295	0	
Shredded_Wheat_spoon_size	N	C	90	3	0	0	3	20	0	120	1	0.67	72.80179	0	
Smacks	K	C	110	2	1	70	1	9	15	40	1	0.75	31.23005	25	
Special_K	K	C	110	6	0	230	1	16	3	55	1	1	53.13132	25	
Strawberry_Fruit_Wheats	N	C	90	2	0	15	3	15	5	90	1	1	59.36399	25	
Total_Corn_Flakes	G	C	110	2	1	200	0	21	3	35	1	1	38.83975	100	
Total_Raisin_Bran	G	C	140	3	1	190	4	15	14	230	1.5	1	28.59279	100	
Total_Whole_Grain	G	C	100	3	1	200	3	16	3	110	1	1	46.65884	100	
Triples	G	C	110	2	1	250	0	21	3	60	1	0.75	39.10617	25	
Trix	G	C	110	1	1	140	0	13	12	25	1	1	27.7533	25	
Wheat_Chex	R	C	100	3	1	230	3	17	3	115	1	0.67	49.78745	25	1
Wheaties	G	C	100	3	1	200	3	17	3	110	1	1	51.59219	25	1
Wheaties_Honey_Gold	G	C	110	2	1	200	1	16	8	60	1	0.75	36.18756	25	1

calories		protein		fat		sodium		fiber		carbo		sugars		potass		weight		cups		rating	
Mean	107.027	Mean	2.513514	Mean	1	Mean	162.3649	Mean	2.175676	Mean	14.72973	Mean	7.108108	Mean	98.51351	Mean	1.030811	Mean	0.821622	Mean	42.37179
Standard E	2.306806	Standard E	0.125059	Standard E	0.117041	Standard E	9.621792	Standard E	0.281714	Standard E	0.452398	Standard E	0.506736	Standard E	8.239479	Standard E	0.017834	Standard E	0.027401	Standard E	1.631386
Median	110	Median	2.5	Median	1	Median	180	Median	2	Median	14.5	Median	7	Median	90	Median	1	Median	0.75	Median	40.25309
Mode	110	Mode	3	Mode	1	Mode	0	Mode	0	Mode	15	Mode	3	Mode	35	Mode	1	Mode	1	Mode	#N/A
Standard	19.84389	Standard	1.075802	Standard	1.006826	Standard	82.76979	Standard	2.423391	Standard	3.891675	Standard	4.359111	Standard	70.87868	Standard	0.153416	Standard	0.235715	Standard	14.03371
Sample Va	393.7801	Sample Va	1.157349	Sample Va	1.013699	Sample Va	6850.838	Sample Va	5.872825	Sample Va	15.14513	Sample Va	19.00185	Sample Va	5023.787	Sample Va	0.023536	Sample Va	0.055562	Sample Va	196.9451
Kurtosis	2.217752	Kurtosis	1.380291	Kurtosis	2.285592	Kurtosis	-0.21758	Kurtosis	8.272757	Kurtosis	-0.30027	Kurtosis	-1.16827	Kurtosis	1.915156	Kurtosis	5.123173	Kurtosis	0.313974	Kurtosis	1.520231
Skewness	-0.46066	Skewness	0.744335	Skewness	1.241525	Skewness	-0.61164	Skewness	2.383501	Skewness	0.118335	Skewness	0.047257	Skewness	1.398722	Skewness	0.280912	Skewness	-0.11498	Skewness	0.962866
Range	110	Range	5	Range	5	Range	320	Range	14	Range	18	Range	15	Range	315	Range	1	Range	1.25	Range	75.66206
Minimum	50	Minimum	1	Minimum	0	Minimum	0	Minimum	0	Minimum	5	Minimum	0	Minimum	15	Minimum	0.5	Minimum	0.25	Minimum	18.04285
Maximum	160	Maximum	6	Maximum	5	Maximum	320	Maximum	14	Maximum	23	Maximum	15	Maximum	330	Maximum	1.5	Maximum	1.5	Maximum	93.70491
Sum	7920	Sum	186	Sum	74	Sum	12015	Sum	161	Sum	1090	Sum	526	Sum	7290	Sum	76.28	Sum	60.8	Sum	3135.512
Count	74	Count	74	Count	74	Count	74	Count	74	Count	74	Count	74	Count	74	Count	74	Count	74	Count	74

After removing the other statistical values, we have the answer for the question below.

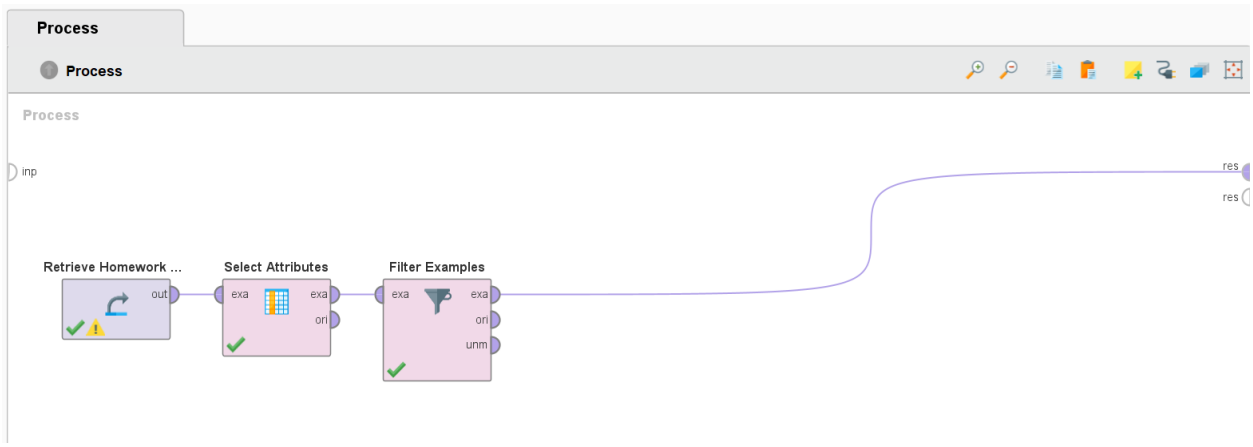
calories		***protein***		***fat***		***sodium***	
Mean	107.027	Mean	2.513514	Mean	1	Mean	162.3649
Median	110	Median	2.5	Median	1	Median	180
Standard Deviation	19.84389	Standard Deviation	1.075802	Standard Deviation	1.006826	Standard Deviation	82.76979
Minimum	50	Minimum	1	Minimum	0	Minimum	0
Maximum	160	Maximum	6	Maximum	5	Maximum	320

fiber		***carbo***		***sugars***		***potass***	
Mean	2.175676	Mean	14.72973	Mean	7.108108	Mean	98.51351
Median	2	Median	14.5	Median	7	Median	90
Standard Deviation	2.423391	Standard Deviation	3.891675	Standard Deviation	4.359111	Standard Deviation	70.87868
Minimum	0	Minimum	5	Minimum	0	Minimum	15
Maximum	14	Maximum	23	Maximum	15	Maximum	330

weight		***cups***		***rating***	
Mean	1.030811	Mean	0.821622	Mean	42.37179
Median	1	Median	0.75	Median	40.25309
Standard Deviation	0.153416	Standard Deviation	0.235715	Standard Deviation	14.03371
Minimum	0.5	Minimum	0.25	Minimum	18.04285
Maximum	1.5	Maximum	1.5	Maximum	93.70491

Hari Vengadesh Elangeswaran

c. *Use RapidMiner to plot a histogram for each of the quantitative variables (5 pt). Based on the histograms and summary statistics, answer the following questions:*



calories	Integer	0	Min 50	Max 160	Average 107.027	Deviation 19.844	Open visualizations
protein	Integer	0	Min 1	Max 6	Average 2.514	Deviation 1.076	Open visualizations
fat	Integer	0	Min 0	Max 5	Average 1	Deviation 1.007	Open visualizations
sodium	Integer	0	Min 0	Max 320	Average 162.365	Deviation 82.770	Open visualizations
fiber	Integer	0	Min 0	Max 14	Average 2.203	Deviation 2.421	Open visualizations
carbo	Integer	0	Min 5	Max 23	Average 14.757	Deviation 3.871	Open visualizations

i. Which variables have the largest variability? (5 pt)

Variability is represented by the standard deviation. The larger the standard deviation, the more variability exists in the data. From the data we can observe high standard deviation from the following variables.

- Sodium: Standard deviation of 82.77
- Potass: Standard deviation of 70.88
- Rating: Standard deviation of 14.03
- Sugars: Standard deviation of 4.36

So we can say that Sodium, Potassium, and Rating have the largest variability.

ii. Which variables seem skewed? (5 pt)

For skewness, we compare the mean and median of each variable. If the mean is significantly higher or lower than the median, this suggests skewness.

We can also observe this with the histograms as in a normal distribution, both tails are symmetrical. In left (negative) skew, the left tail is longer, in right (positive) skew, the right tail is longer, and the mean is greater than the mode. This is marked in the histograms to compare the skew and highlighted with **red** boxes.

So we can say **Sodium** is showing *slight negative skew***,** and **Rating and Potass** showing *positive skew*.

iii. Are there any values that seem extreme? (5 pt)

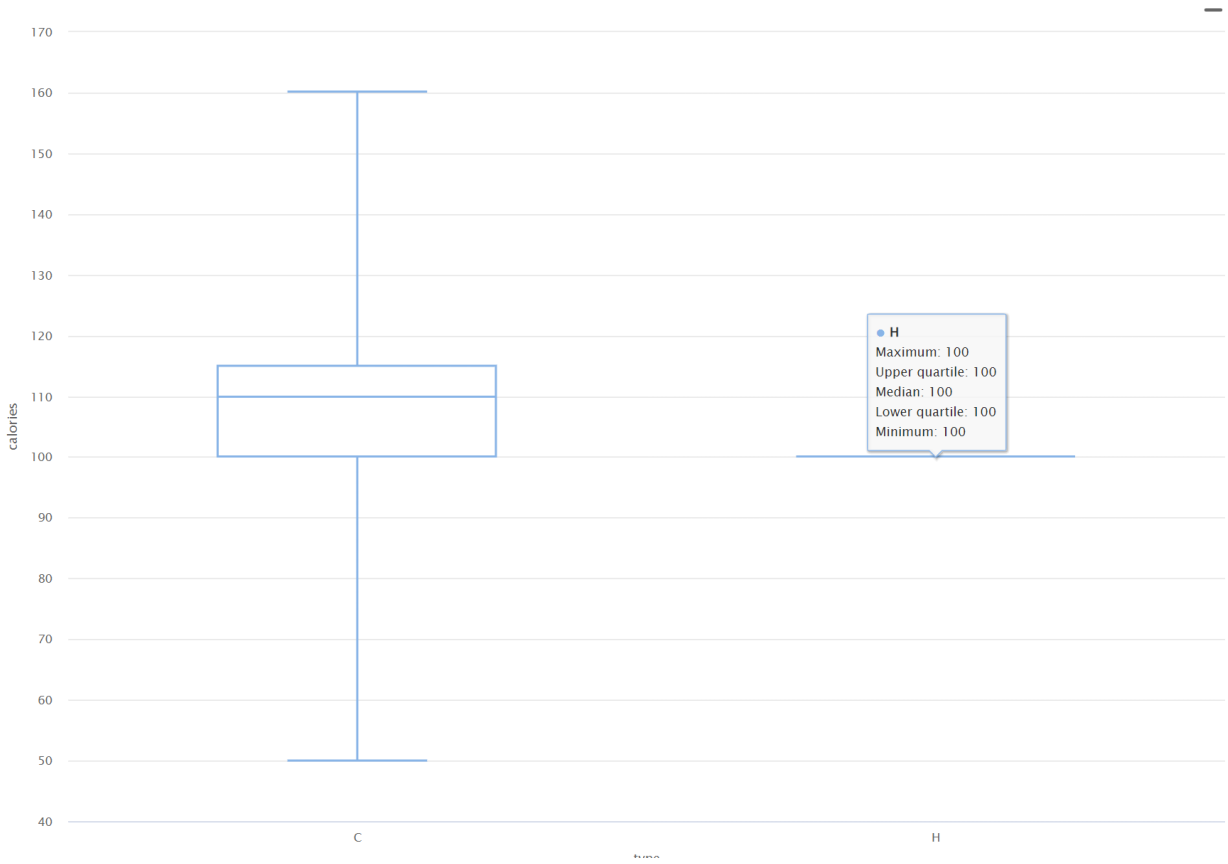
We may be able to identify extreme values with maximum and minimum. Using this we can see that Sodium, Potassium and Rating have maximums that are larger than the mean and having high standard deviation can suggest that they have extreme values.

But by looking at the histograms we can see that isolated bars in Fat, Fiber, Weight and Rating which suggests outliers as the rest of the bars are crammed together.

Using the above logic we can say **Fat, Fiber, Weight** and **Rating** have values that seem extreme.

---

d. *Plot a sidebyside boxplot comparing the calories in hot vs. cold cereals. What does this plot show us? (5 pt)*



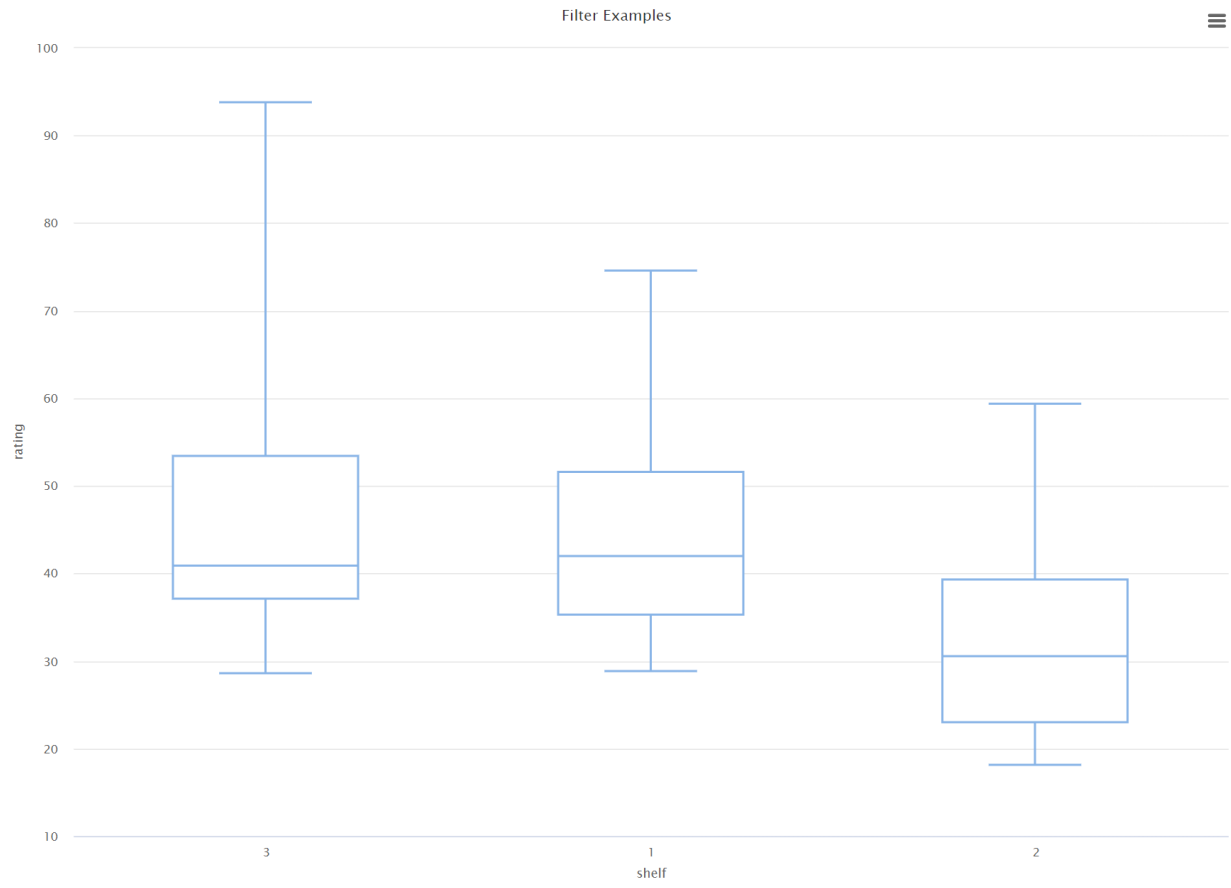
For cold cereals the lower whisker extends to 50 and the upper whisker extends to 160, with a median of 110, indicating the full range of data. The interquartile range, that is the middle 50% of values lie between 100 and 115. There's some variation but no extreme outliers.

For hot cereal values we can observe that all calorie values are exactly 100 with no variability at all. This tells us that there isn't a good distribution for hot cereals or hot cereals tend to be more consistent.

But since we observed the data before comparing the boxplot, we know this is because only 1 hot cereal entry exits, and the observations made above are due to that.

---

e. ***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 of shelf height? (5 pt)***



Shelf 3 has a wide range of ratings, from approximately 30 to 90, with a median around 45. The IQR (middle 50% of data) spans from around 40 to 60, indicating a diverse spread of consumer ratings.

Shelf 2 ratings are slightly more concentrated, with a range from around 30 to 75, and a median around 40. It has a narrower IQR compared to Shelf 3.

Shelf 2 has the smallest range, with ratings from approximately 30 to 70 and a median just above 35. The IQR is tighter compared to Shelves 1 and 3, indicating more consistency in ratings for products placed on this shelf.

The differences in ranges, medians, and spread suggests that removing any of the three categories may result in a loss of important information. So, it's mostly beneficial to keep all three shelf heights If we were to predict consumer rating from shelf height.

---

f. *Compute the correlation table for the quantitative variable*

Attribut...	calories	protein	fat	sodium	fiber	carbo	sugars	potass	weight	cups	rating
calories	1	0.034	0.507	0.296	-0.292	0.274	0.569	-0.071	0.696	0.089	-0.694
protein	0.034	1	0.202	0.012	0.517	-0.035	-0.287	0.579	0.231	-0.242	0.467
fat	0.507	0.202	1	0.001	0.022	-0.281	0.287	0.200	0.222	-0.158	-0.405
sodium	0.296	0.012	0.001	1	-0.068	0.332	0.037	-0.039	0.313	0.120	-0.383
fiber	-0.292	0.517	0.022	-0.068	1	-0.391	-0.145	0.913	0.249	-0.524	0.596
carbo	0.274	-0.035	-0.281	0.332	-0.391	1	-0.451	-0.366	0.147	0.354	0.051
sugars	0.569	-0.287	0.287	0.037	-0.145	-0.451	1	0.001	0.461	-0.032	-0.756
potass	-0.071	0.579	0.200	-0.039	0.913	-0.366	0.001	1	0.421	-0.502	0.416
weight	0.696	0.231	0.222	0.313	0.249	0.147	0.461	0.421	1	-0.202	-0.300
cups	0.089	-0.242	-0.158	0.120	-0.524	0.354	-0.032	-0.502	-0.202	1	-0.223
rating	-0.694	0.467	-0.405	-0.383	0.596	0.051	-0.756	0.416	-0.300	-0.223	1

i. Which pair of variables is most strongly correlated? (2.5 pt)

The pair of variables that are most strongly correlated are potass and fiber.

ii. How can we reduce the number of variables based on these correlations? (2.5 pt)

Attribut...	rating ↓	calories	protein	fat	sodium	fiber	carbo	sugars	potass	weight	cups
rating	1	-0.694	0.467	-0.405	-0.383	0.596	0.051	-0.756	0.416	-0.300	-0.223
fiber	0.596	-0.292	0.517	0.022	-0.068	1	-0.391	-0.145	0.913	0.249	-0.524
protein	0.467	0.034	1	0.202	0.012	0.517	-0.035	-0.287	0.579	0.231	-0.242
potass	0.416	-0.071	0.579	0.200	-0.039	0.913	-0.366	0.001	1	0.421	-0.502
carbo	0.051	0.274	-0.035	-0.281	0.332	-0.391	1	-0.451	-0.366	0.147	0.354
cups	-0.223	0.089	-0.242	-0.158	0.120	-0.524	0.354	-0.032	-0.502	-0.202	1
weight	-0.300	0.696	0.231	0.222	0.313	0.249	0.147	0.461	0.421	1	-0.202
sodium	-0.383	0.296	0.012	0.001	1	-0.068	0.332	0.037	-0.039	0.313	0.120
fat	-0.405	0.507	0.202	1	0.001	0.022	-0.281	0.287	0.200	0.222	-0.158
calories	-0.694	1	0.034	0.507	0.296	-0.292	0.274	0.569	-0.071	0.696	0.089
sugars	-0.756	0.569	-0.287	0.287	0.037	-0.145	-0.451	1	0.001	0.461	-0.032

After examining the correlation matrix, we can choose only the attributes with above 0.3 and below -0.3 correlation to the label, so all attributes with correlation between **-0.3** and **+0.3** can be rejected. Then to prevent any multicollinearity we can discard one of the variables of a pair of variables which have a correlation of above **+0.8** or below **-0.8**.

We can see in the image above that using the strategy above reduces the number of variables.
