Data Expiration and Dimension Reduction
Cereals Dataset

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

Answer:

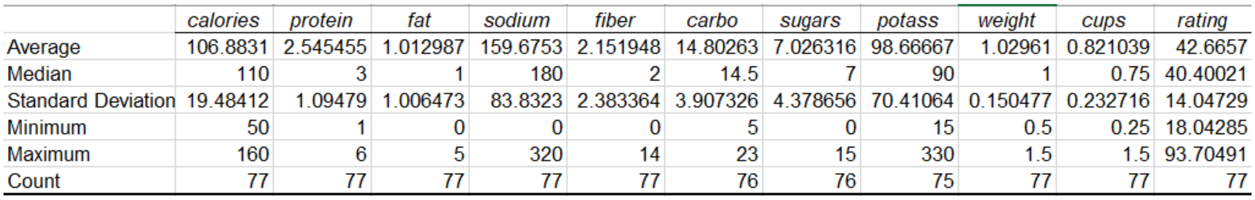
Numerical: The following variables are numerical: calories, protein, fat, sodium, fiber, carbo, sugars, potass, weight, cups, rating.

Ordinal: vitamins, type

Nominal: Name, mfr, shelf

2. Create a table with the average, median, min, max, standard deviation, count blank(the number of records with missing values) for each of the quantitative variables. This can be done through 1) Excel's functions or 2)Excel's Data à Data Analysis à Descriptive Statics menu and then use a excel function for counting missing values.

Answer:



	calories	*protein*	*fat*	*sodium*	*fiber*	*carbo*	*sugars*	*potass*	*weight*	*cups*	*rating*
Average	106.8831	2.545455	1.012987	159.6753	2.151948	14.80263	7.026316	98.66667	1.02961	0.821039	42.6657
Median	110	3	1	180	2	14.5	7	90	1	0.75	40.40021
Standard Deviation	19.48412	1.09479	1.006473	83.8323	2.383364	3.907326	4.378656	70.41064	0.150477	0.232716	14.04729
Minimum	50	1	0	0	0	5	0	15	0.5	0.25	18.04285
Maximum	160	6	5	320	14	23	15	330	1.5	1.5	93.70491
Count	77	77	77	77	77	76	76	75	77	77	77

The above table shows the average, median, min, max, standard deviation and number of missing records in each of the following quantitative records.

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

a. Which variable have the largest variability?

b. Which variables seem skewed?

c. Are there any values that seem extreme?

Answer:

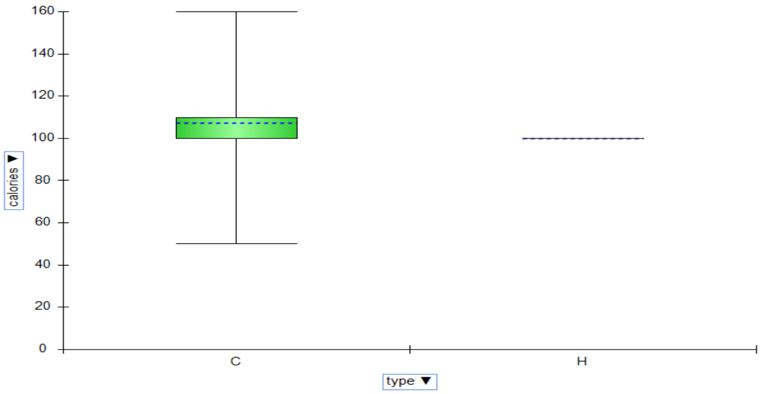
a) The variable with largest variability is sodium

b) Potassium, rating, protein, fat are skewed

c) Protein, sodium, fibre, potass, vitamins, weight and rating have few extreme values.

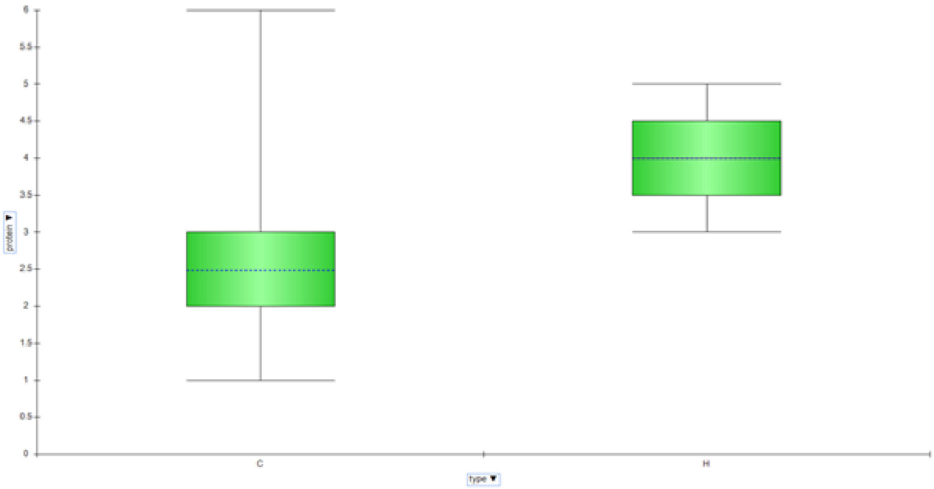
4. Use XLMiner to plot a side-by-side boxplot comparing the calories in hot vs cold cereals. What does this plot show us?

Answer: We plotted boxplots for the calorie values of the cold and hot cereals. The Y-axis represents the calorie value and the X-axis represents the type. After analyzing the boxplot for the cold cereal, we realized that the calorie range for the cold cereals is 110, lowest being 50, highest 110. The 1st quartile is 100 and the 3rd quartile is 110, which means that the calorie count in the interquartile section is from 100 – 110, the calories of the cold cereal isn’t spread out. The data for hot cereals was limited to only 3, nothing notable could be concluded.



5. Use XLMiner to plot a side-by-side boxplot comparing the protein in hot vs cold cereals. What does this plot show us?

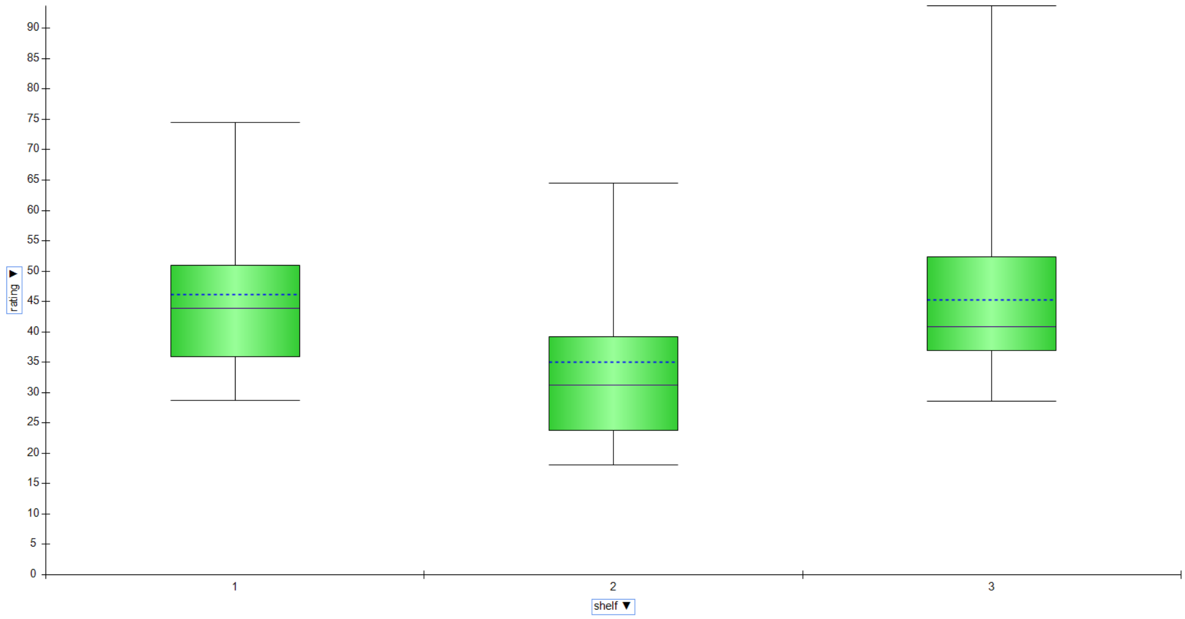
The Y axis of the box plots is Protein and the X axis is Type of the cereal. The range of cold cereals is 5, while that for hot cereals it is 3. 1&6 and 3&5 as minimum and maximum respectively. The 1st and the 3rd quartile for cold and hot cereals are 2,3 and 3.5,4.5 respectively, which tells that the interquartile section for both the type of cereals have a range of 1. Here, we can conclude that the values in the interquartile section are clustered for both the cereals.



6. Use XLMiner 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 of shelf height?

Answer:

The boxplots with Y axis as consumer rating and X axis as shelf are plotted. The mean rating for first shelf is 46.14, the highest and, second is 39.97 and third is 45.22. Also, the median for the first shelf is the highest, being 43.93, second is 31.23, and third is 40.80. The median and mean between the histogram and third is not much different but the mean and median for the second shelf is substantially low. Hence, statistically, we can say that the second shelf can discarded.



7. Compute the correlation table for the quantitative variable (use Excel's Data à Data Analysis à Correlation menu). In addition, use XLMiner to generate a matrix plot for these variables.

a. Which pair of variables is most strongly correlated?

Answer:

The variables potassium and fibre have strong positive correlation. As the values of potassium increases the values of fibre also increases. Therefore, according to the definition of the positive correlation large values of potassium are correlated to large values of fibre. The variables ratings and calories have strong negative correlation. As the values of calories increases the ratings values decreases. According to the definition of negative correlation large values of calories are correlated with small values with ratings.

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

Answer:

We can reduce the number of variables can be reduced by using the process of dimensional reduction. In order to reduce the number of variables we use their correlation analysis. The variables which are strongly correlated, we can predict the value of one variable based on the value of other variable, so these variables can be combined. For example, we have potassium and fibre have strong positive correlation, so we can combine these variables because if the value of potassium increases then fibre also increases.

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

Answer:

Normalization is getting all the data in a specific range. We do normalization to get better relations and establish a robust relationship, it tells us what variable's movement has an effect on other variables. When the relationship between two datasets is non-linear, we transform into a linear relationship. This helps us to find the correlation between these variables and we can get the relationship between them, But if the normalization involves only linear transformation, the correlation will not change.

8. Remove all records with missing numerical measurements from the dataset by creating a new worksheet. You can use the "Missing Data Handling utility in XLMiner".

Answer:

Used the Missing Data Handling utility in XLMiner to remove all the missing data in the dataset.
Once we select the dataset and remove these a new sheet with the data that has no missing values is created. The excel file attached has this dataset in the 'imputation' sheet.

9. Conduct a principal components analysis on the cleaned data and comment on the results. Should the data be normalized? Discuss what characterizes the components you consider key. Use the "principal components utility in XLMiner".

Answer:

Principal Component Analysis (PCA) is a very useful method for dimensionality reduction if the number of variables is very large. To conduct the PCA we have to use either correlation or covariance matrix. In any of the above mentioned cases we first normalize the data and then create the matrix based on the selected case. We normalize the data because principal component analysis is variance maximizing activity. From the principal component table we get to know the components which are more dependent on variables. The variance table shows the value of variance for each principal component. From the PCA table we get to know that first component is dependent on rating. Second, third and fourth components are dependent on weight, carbohydrates and fat respectively. Similarly, other components represent the variables with highest magnitude values in the table