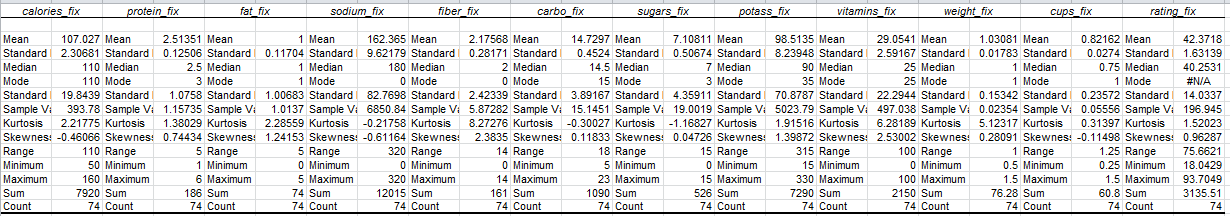
Assignment #3 PCA by Joshua Troup

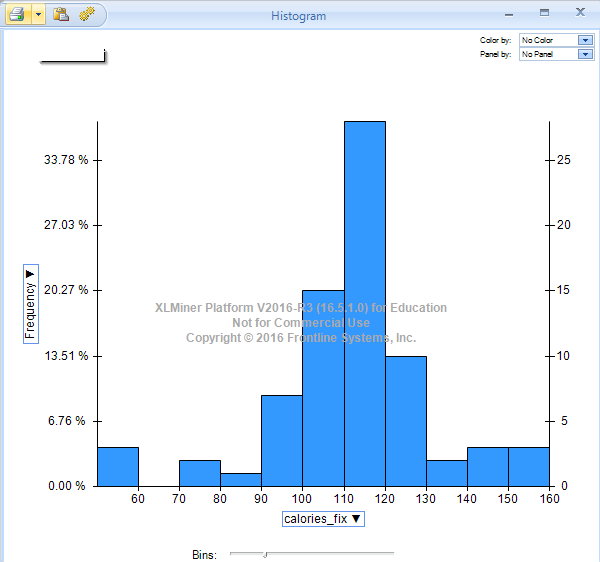
1. Which variables are quantitative/numerical? Which are ordinal? Which are nominal? (create a Table for easy display of various types)

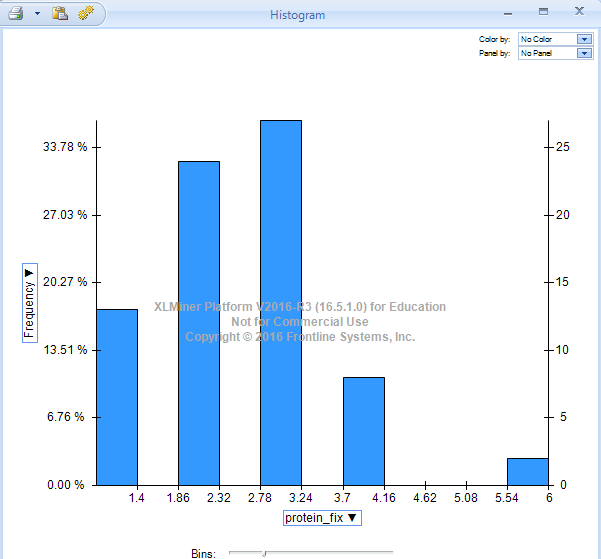
|  |  |  |
| --- | --- | --- |
| **Quantitative** | **Ordinal** | **Nominal** |
| Calories Per Serving | Displaying Shelfs | Manufacture Type |
| Grams of Protein |  | Type of Cereal |
| Grams of Fat |  | Name of Cereal |
| Milligrams of Sodium |  |  |
| Milligrams of Potassium |  |  |
| Grams of Sugar |  |  |
| Grams of Complex Carbs |  |  |
| Rating of Cereals |  |  |
| Grams of Dietary Fiber |  |  |
| Vitamins & Minerals |  |  |
| Weight in Ounces |  |  |
| Cups in One Serving |  |  |

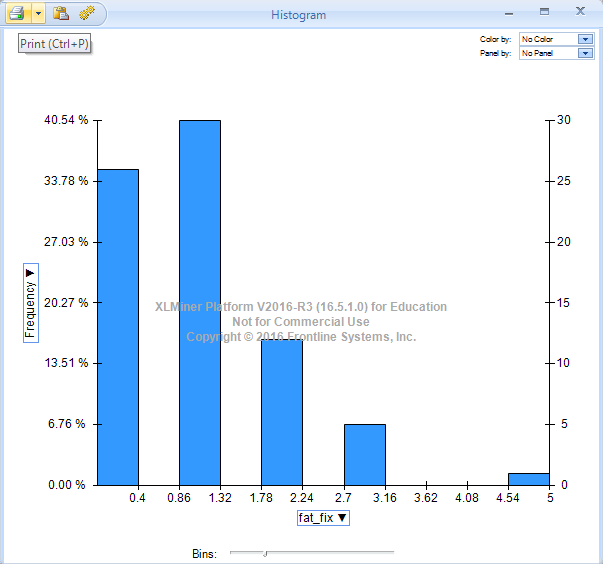
1. Create a table with the average, median, min, max, and standard deviation for each of the quantitative variables (descriptive statistics).

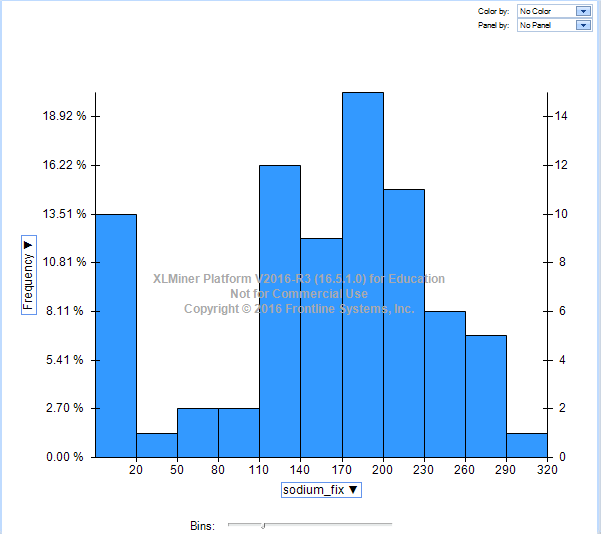


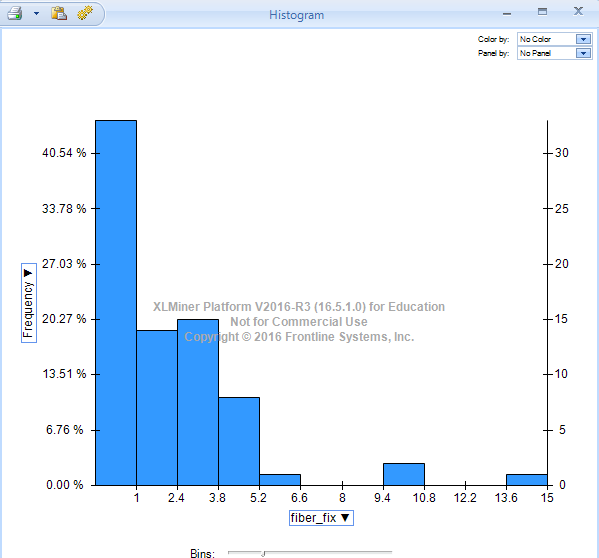
1. Use XLMiner to plot a histogram for each of the quantitative variables. Based on the histograms and summary statistics, answer the following questions:

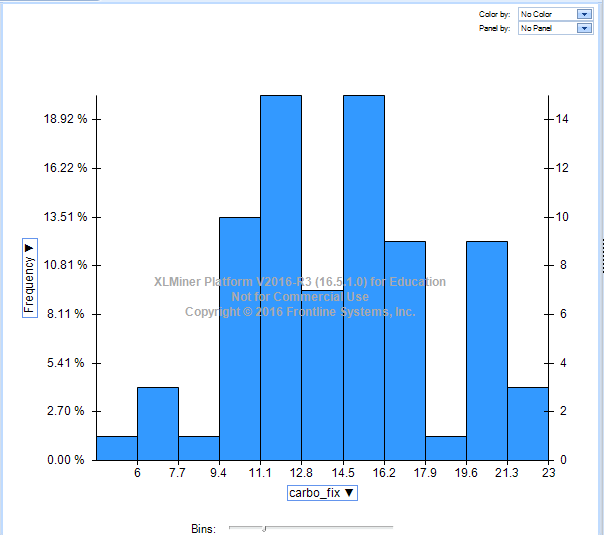


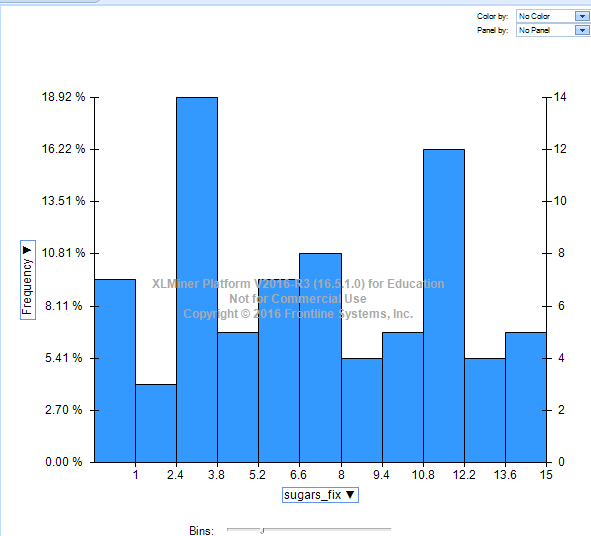


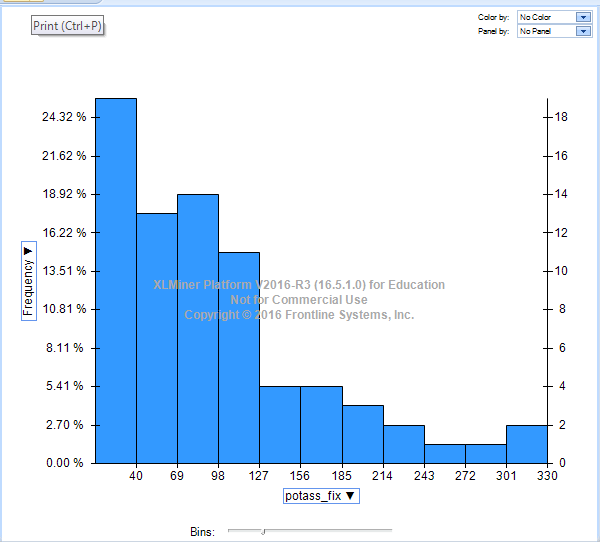


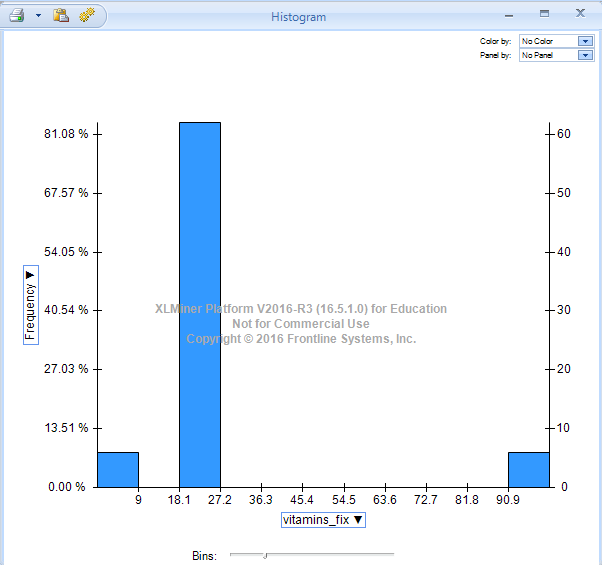


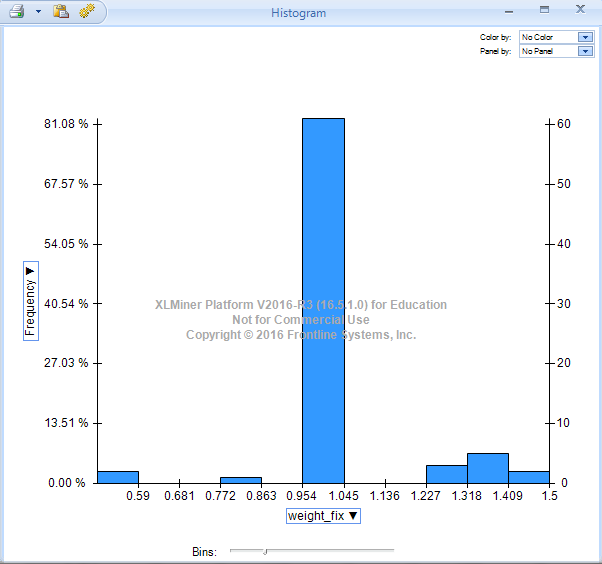


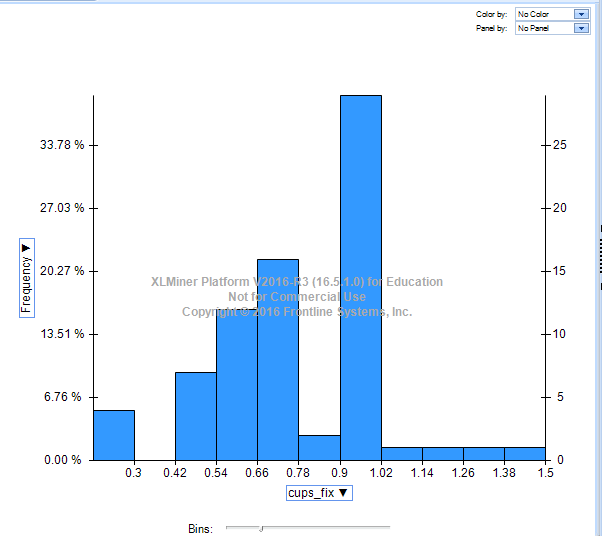


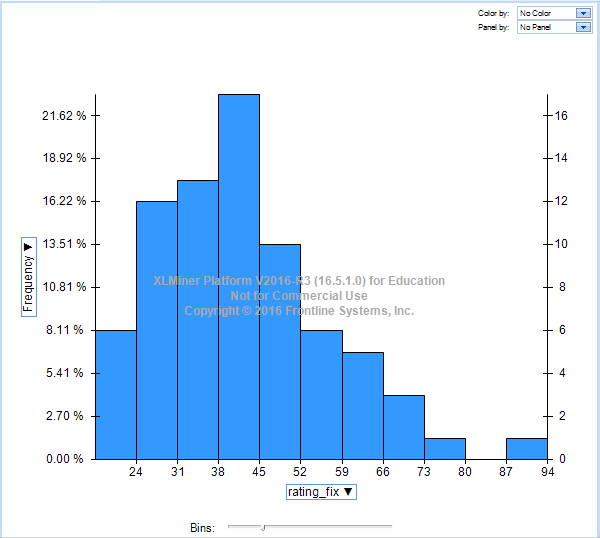












a. Which variables have the largest variability?

**Sodium and Potassium have the largest variability**

**Sodium-Sample Variance=6850 Potassium-Sample Variance=5023**

b. Which variables seem skewed? (use the descriptive statistics and histograms to clearly sate positive or negative skew)

**Potassium and Fiber seem to have positive skews with the longer tails to the right of the frequency distribution shown in the histogram. Calories seems to be negatively skewed with the longer tail to the left of the frequency distribution shown in the histogram.**

c. Are there any values that seem extreme? (list the variables and the extreme values for each one of them)

**Protein-6**

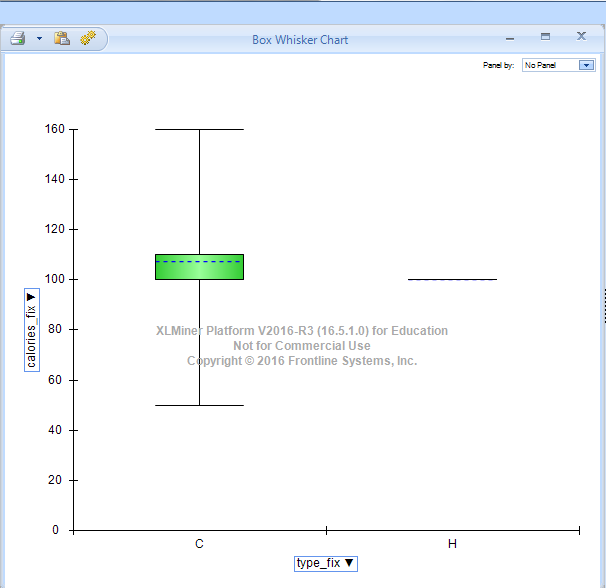
**Fat-5**

**Sodium-15**

**Fiber-14,10**

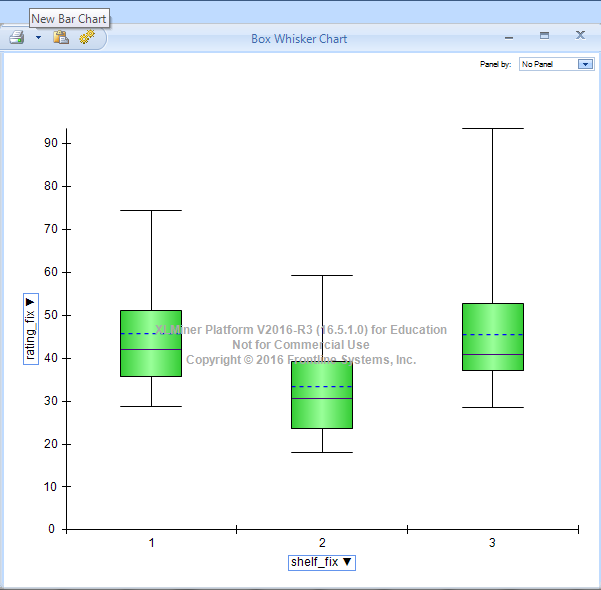
**Vitamins-0, 100**

d. Use XLMiner to plot a side-by-side boxplot comparing the calories in hot vs. cold cereals. What does this plot show us? (use the calories values to distinguish them)



**Cold cereals vary in calories ranging from approximately 50-160. The median is 110 while the mean is 107. Hot cereals all have 100 calories since there is only one hot cereal in the data set.**

e. 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?



**Shelf 1 has a range of ratings from 28-74.**

**Median-41**

**Mean-45**

**Shelf 2 has a range of ratings from 18-59.**

**Median-30**

**Mean-33**

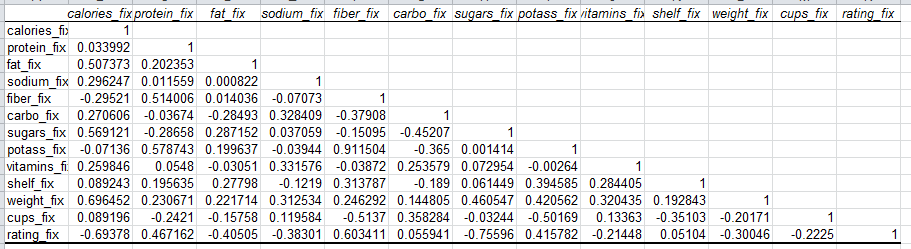
**Shelf 3 has a range of ratings from 28-93.**

**Median-40**

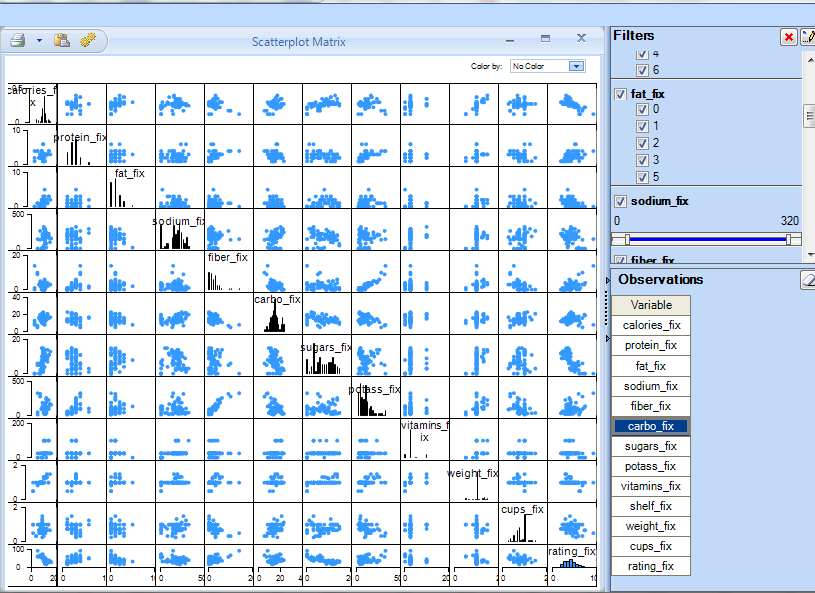
**Mean-45**

**Shelf 1 & 3 seem to be predictors of consumer ratings with similar values and higher values than shelf 2.**

f. Compute the correlation table for the quantitative variable (use Excel’s Data → Data Analysis → Correlation menu).

****

In addition, use XLMiner to generate a matrix plot for these variables in the correlation table generated.

****

Which pair of variables is most strongly correlated?

**Potassium and Fiber seem to be the most correlated based on the matrix plot above.**

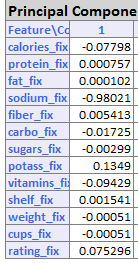
ii. How can we reduce the number of variables based on these correlations?

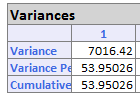
**The pairs of variables with strong correlation contain corresponding information therefore the data is redundant and should be removed to avoid multicollinearity**

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

**The correlations would not change since they’re not affected by the normalization of data.**

iv. Now, run the PCA algorithm (under Transformation) with default settings and the output may look as shown below.

****

****

Consider the first column which is the Principal Component. It shows the weights carried by different variables. Describe briefly what this column represents in terms of the measure of each variable. Which is the dominant variable?

**The first columns shows how each variable contributes to the data. If principal component 1 was included then 53% of the information would be accounted for. The dominant variable is Sodium with 98%.**