**OPIM 5604 B15 – Predictive Modeling Assignment 3 Meghana Kasula (Net ID=mek15120)**

*“The work contained and presented here is my work and my work alone.”*

**4.1 Breakfast Cereals. Use the data for the breakfast cereals example in Section 4.8 (Cereals.jmp) to explore and summarize the data as follows:**

**a. Which variables are continuous/numerical? Which are ordinal? Which are**

**nominal?**

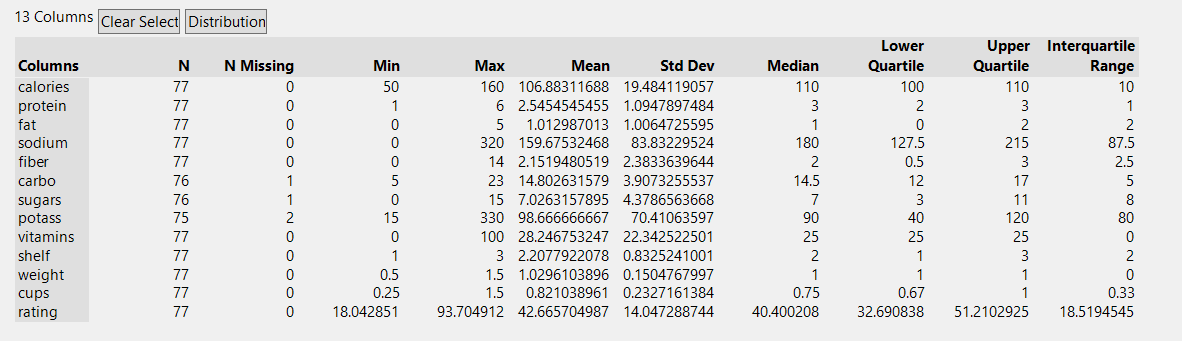
Continuous variables-calories protein, fat, sodium, fiber, carbo, sugars, potass, vitamins, shelf, weight, cups, rating.

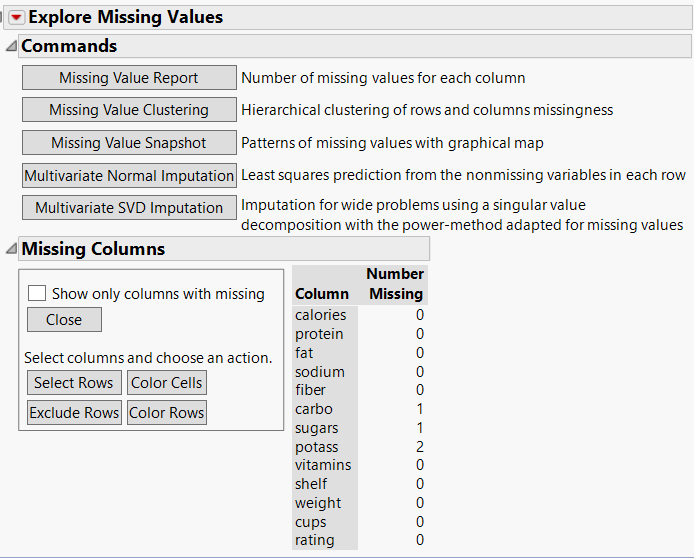
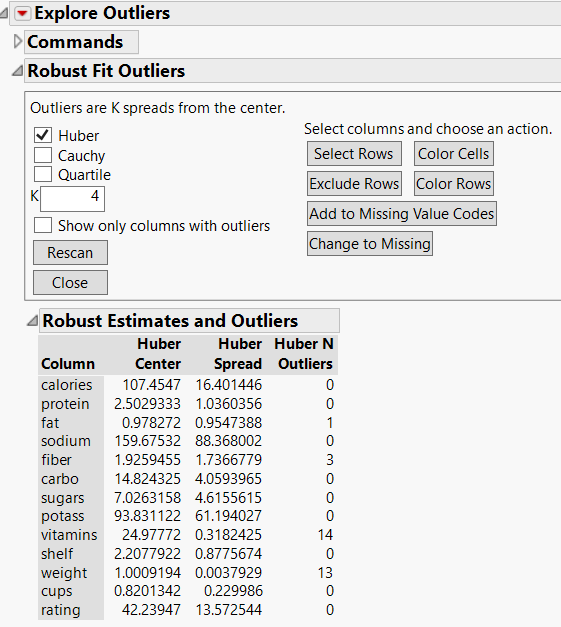
Nominal variables - name, mfr and type.

**b. Calculate the following summary statistics: mean, median, min, max, and**

**standard deviation for each of the continuous variables, and the count for each**

**categorical variable. This can be done using *Cols* > *Columns Viewer*.**



****

**i. Is there any evidence of extreme values?**

Yes, if we explore outliers through col>model utilities > explore outliers.

We can see that vitamins, fat and weight has outliers which has been calculated using Huber spread and from Huber center.

**ii. Which, if any, of the variables is missing values?**

If we explore outliers through col>model utilities > explore missing values.

We can see that potass , sugar and carbo has missing values.

**c. Use *Analyze* > *Distribution* to plot a histogram for each of the continuous**

**variables and create summary statistics. Based on the histograms and summary**

**statistics, answer the following questions:**

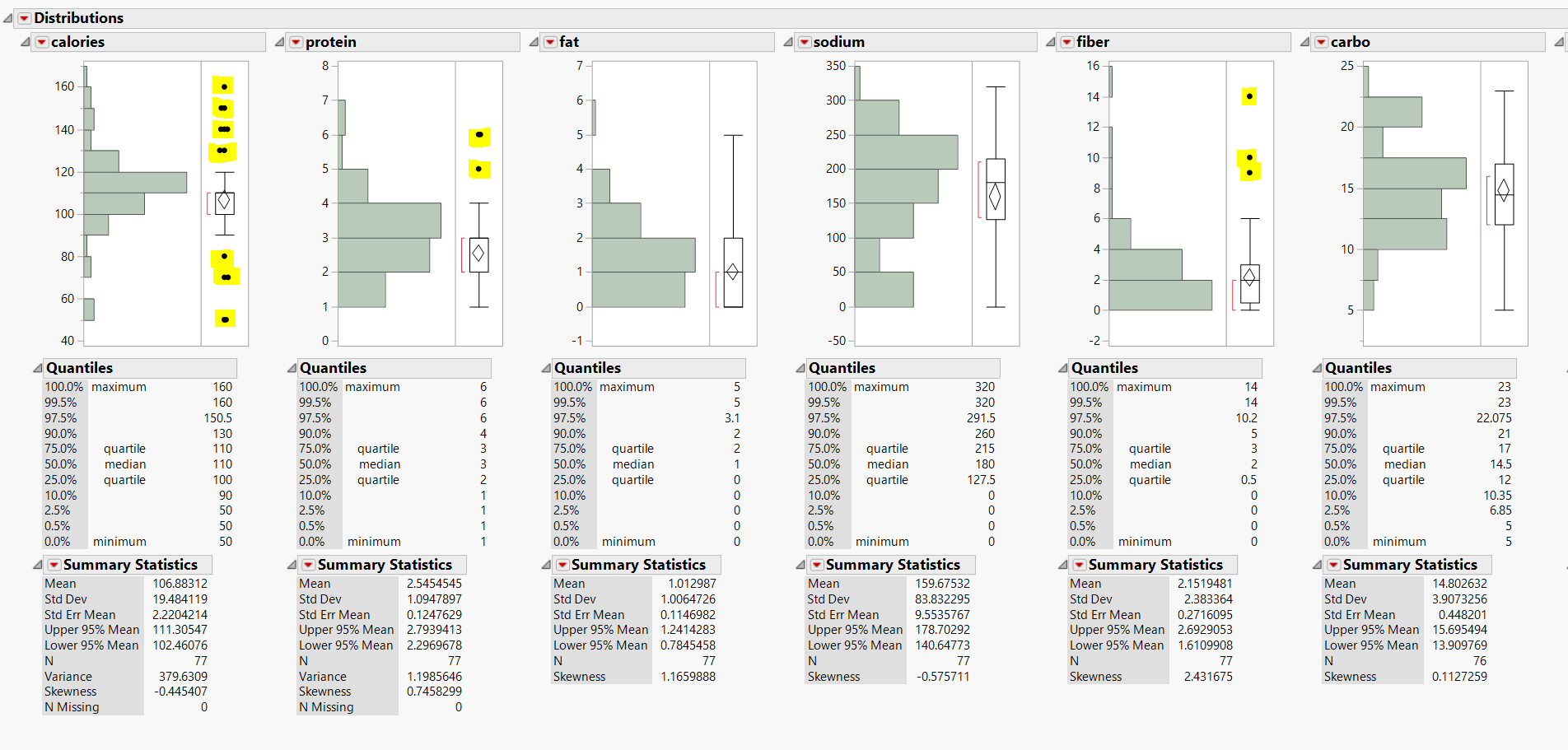
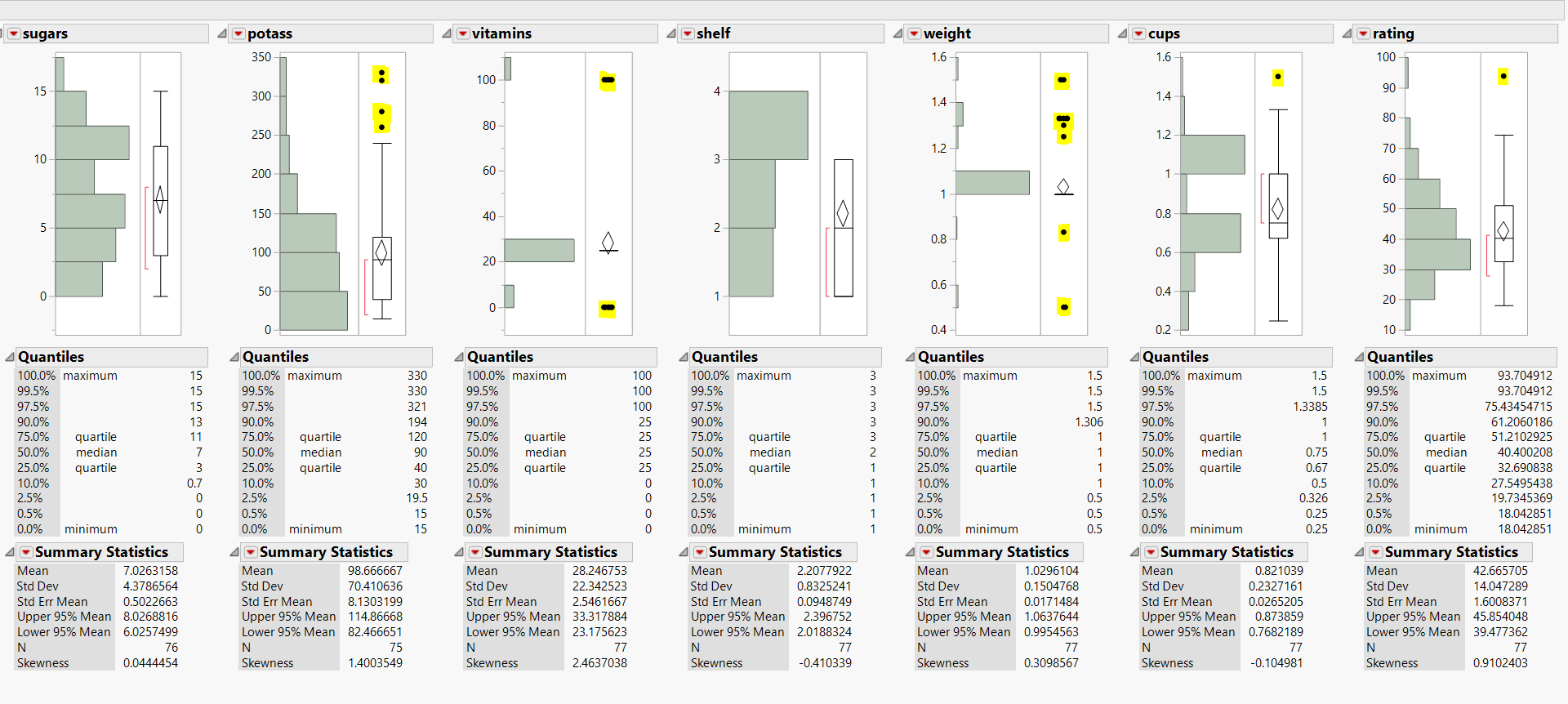
**i. Which variables have the largest variability?**

We can calculate the variability using coefficient of variance. Here coefficient of variance is particularly important since we are comparing data of two different sets, like potass and fat etc. this can give us the clear idea about variability in all measures. It’s formula is-

Coefficient of variance = standard Deviation / Mean

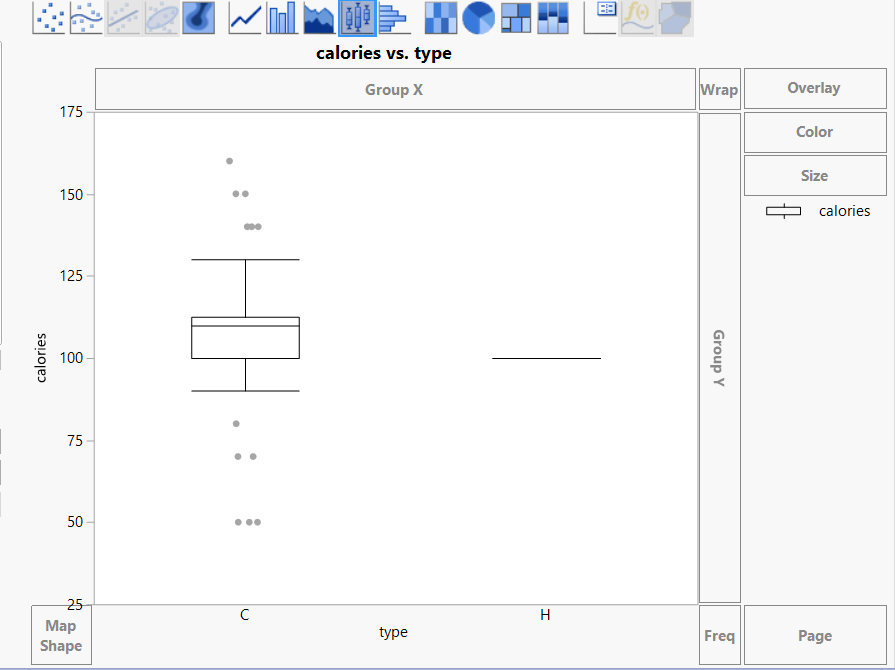
Hence, after conducting the formula for each of the measures, I get to know that the vitamins have the highest coefficient of variance. Which is 0.788.

**ii. Which variables seem skewed?**

**iii. Are there any values that seem extreme?**

Rule of thumb is that acceptable amount of skewness is between -1 to +1. Hence I have highlighted the values which are either lesser than -1 and more than 1. So, very skewed values are fat, fiber, potass and vitamins.

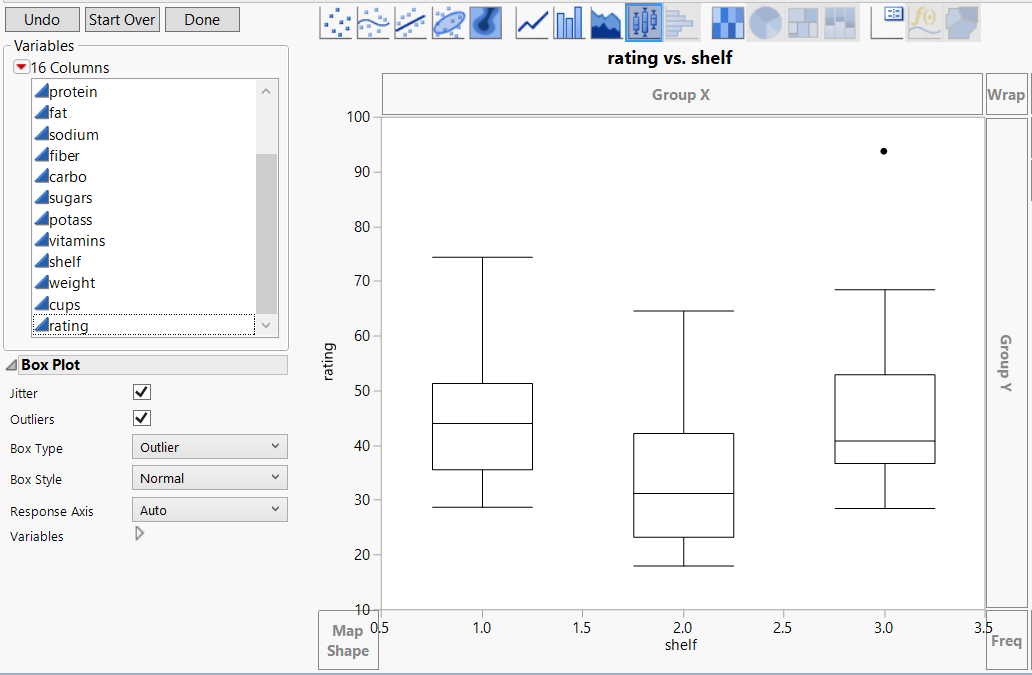
Yes, there are extreme values! I have highlighted all the extreme values in Yellow. Those are the outliers.

**d. Use the *Graph Builder* to plot a side-by-side boxplot comparing the calories in**

**hot vs. cold cereals. What does this plot show us?**

The plot shows us that there is a large range in Cold cereals ranging from 50 -160. It also includes outliers.

Hot cereals have only one value of calories, which is 100. So the mean, median and mode has the same value.

**e. Use the *Graph Builder* to plot a side-by-side boxplot of consumer rating as a**

**function of the shelf height (the variable *shelf*). If we were to predict consumer**

**rating from shelf height, does it appear that we need to keep all three categories**

**of shelf height?**

As we look at the box plot on the right, we can observe that the plot at shelf 1 and shelf 3 are similar looking. Moreover, the plot at shelf 3 looks like a subset of plot 1. Hence, we can conclude combining the shelf 3 and 1 would not make much difference. We can effectively reduce the numbers of shelves to two.

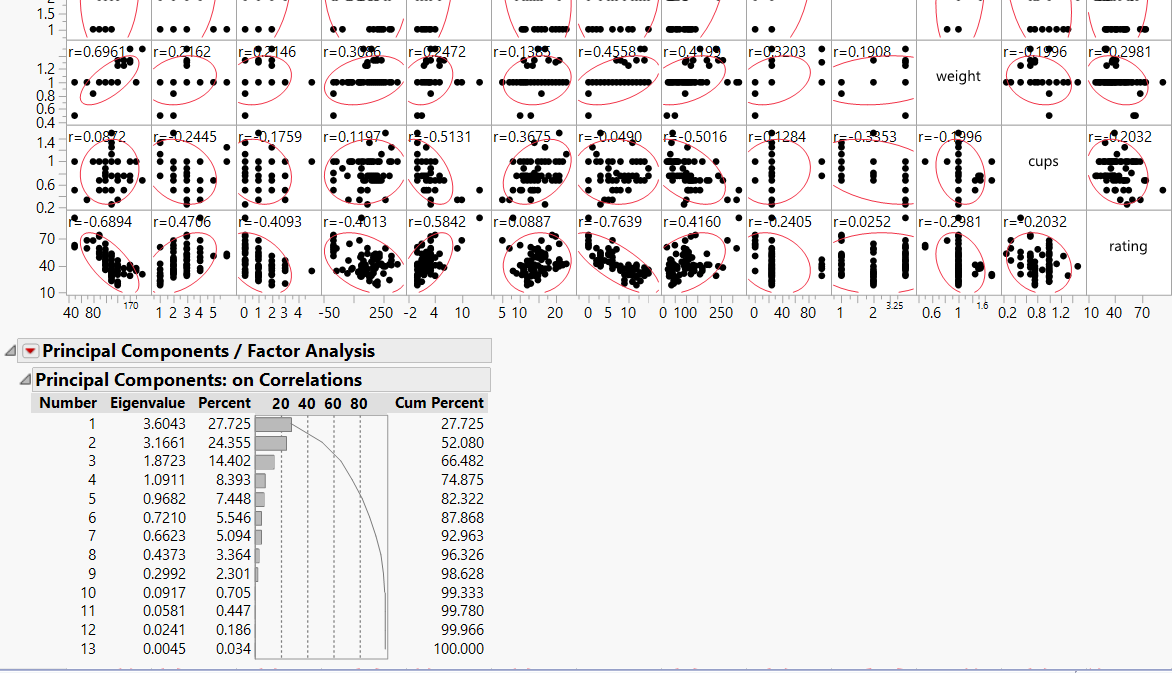
**f. Compute the correlation table and generate a scatterplot matrix for the**

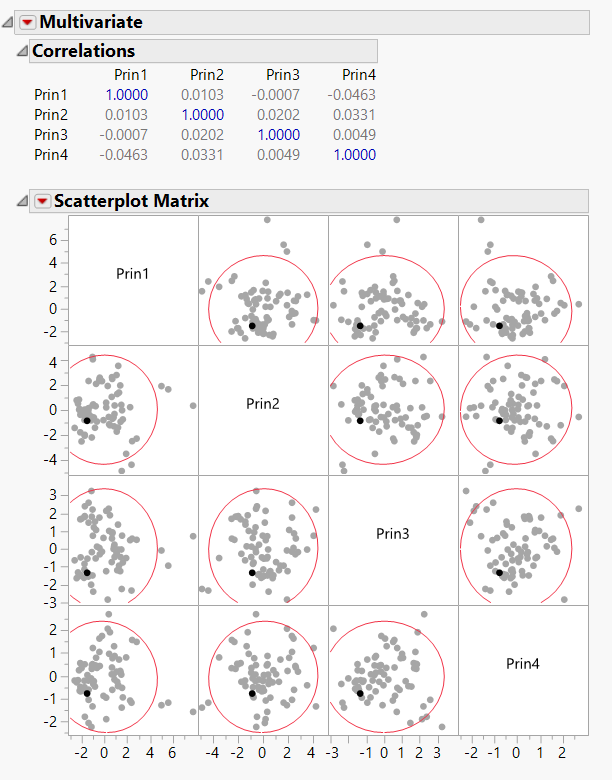
**continuous variables (use *Analyze* > *Multivariate Methods* > *Multivariate*).**

**i. Which pair of variables is most strongly correlated?**

As we analyze the correlations matrix above, the highest correlation seems to be between the pair fiber and potass. They are highlighted in blue.

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



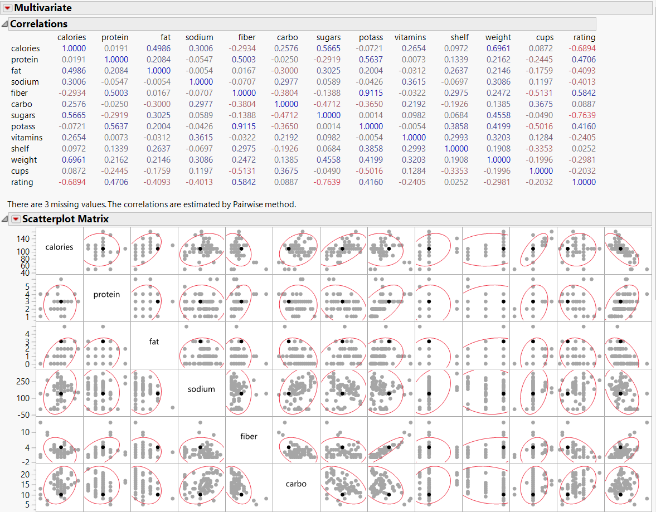


As we select on the red triangle beside principal components and save principal components “4”, we get 4 new columns based on the correlations where all the measures are reduced.

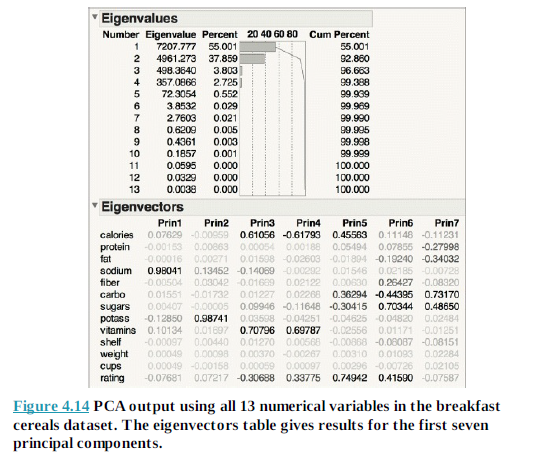
The second picture represents the correlation matrix of those reduced variables. This is how we can reduce number of variables using correlations.

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

****Correlation matrix is done only after the data is normalized. Hence, there would not be any difference if we differently normalize the data apply the correlation matrix.

The above one is the correlation matrix before standardizing attributes. The beside one is after. We can observe no change at all.

**g. Consider the first column on the left under *Eigenvectors* in Figure 4.14. Describe**

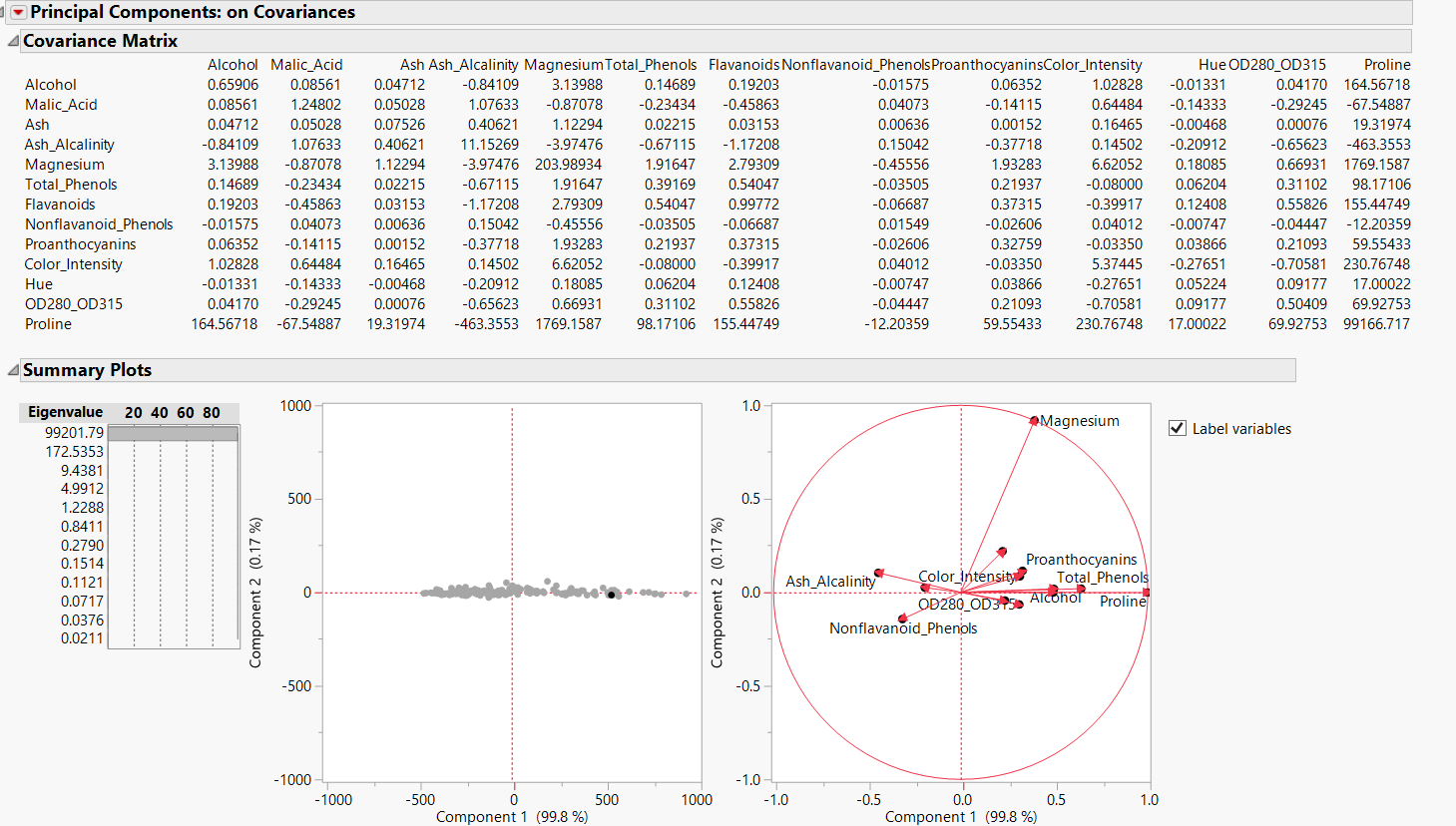
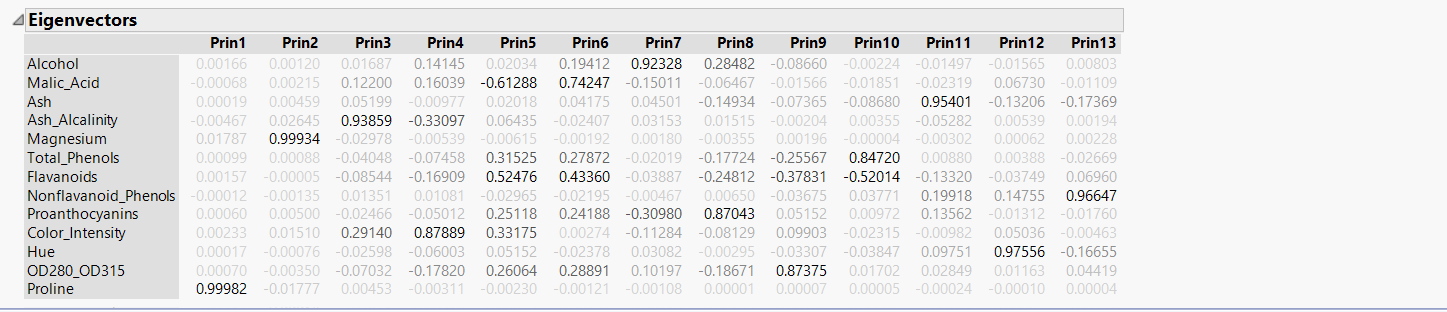
**briefly what this column represents.**

Here, in first column, the Eigen vector Prin 1, it is evident that the first principal component is dominated by Sodium content. It has the highest value of 0.98041 at 55.001%.

**4.2 Chemical Features of Wine. Figure 4.17 shows the PCA output for analyses conducted on normalized (correlations) and non normalized data (covariances).Each variable represent a chemical characteristics of wine, and each case is a different wine.**

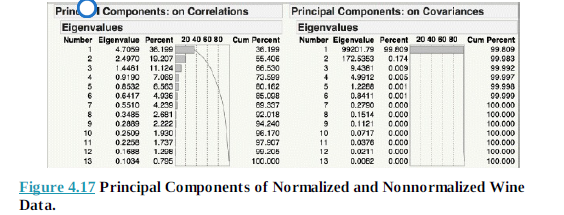
**a. The data are in the file Wine.jmp. Consider the variances in the columns labeled *Eigenvalue* for the principal components analysis conducted using covariance. Why is the variance for the first principal component so much larger than the others?**

Here, instead of the correlations we will take covariance. We will have to do the following. Analyze> Multivariate Methods> Principal component > on covariance.



Under the Principal component 1, Proline has highest variance. It is because it has the higher scale as compared to others.

**b. Comment on the use of correlations versus covariances. Would the results change dramatically if PCA (in this example) were conducted on the correlations instead?**



As we can notice, the covariance is not entirely distributed. The first component itself is taking 99.808%. Since, proline is measures in different weight which is more than other measures or variable present. Where as in the correlation chart , the distribution is further more than covariance, and also the variance in weights are further considered and the data is normalized too.