**Problem 1 – Cereal Data Factor Analysis**

Please note that this problem has been attempted in two ways. There are variables in the data called Soggy and Boring. These are negative expressions and we have assumed that when users would have rated the products / brands on these two parameters, they would have considered something like 1= “Not Soggy” & 5 = “Very Soggy”. Similarly for Boring.

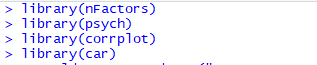
Hence, we have :

1. Tried to arrive at a conclusion using the raw data and making no changes to it
2. Tried to arrive at a conclusion by considering the data where we inversed the ratings in all rows for both Soggy and Boring.

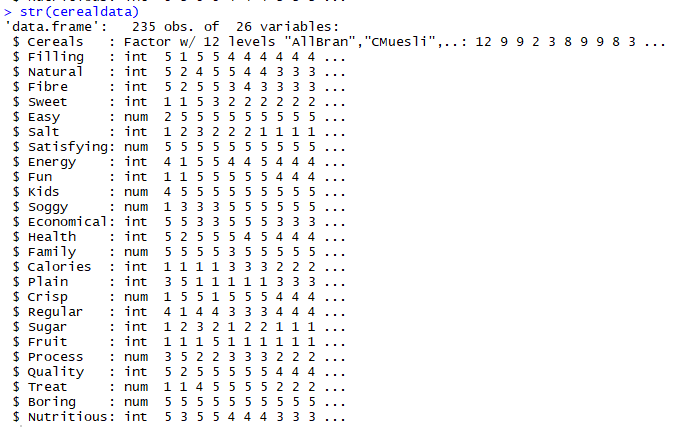
The results were more or less same except when we had to make further interpretations in the end. Please read on.

We started with exploring the data:

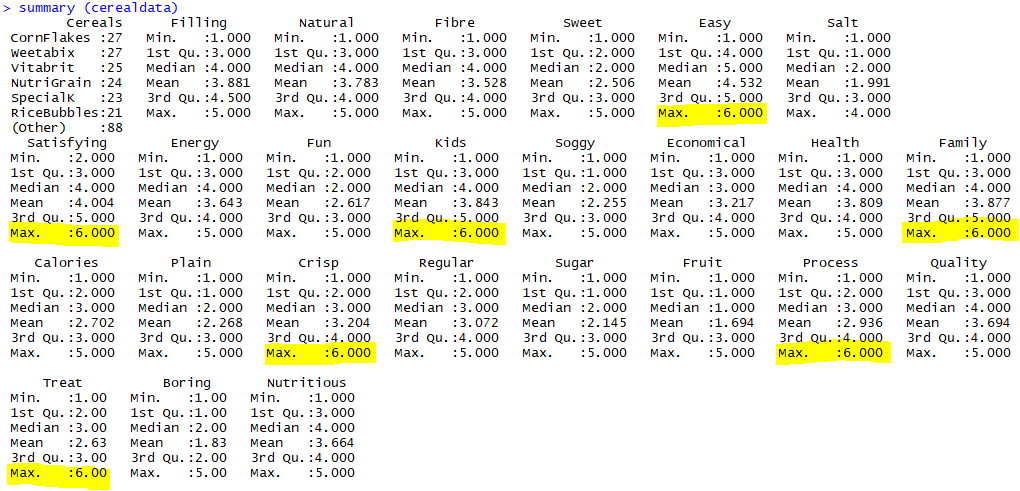
1. We load all required libraries upfront:



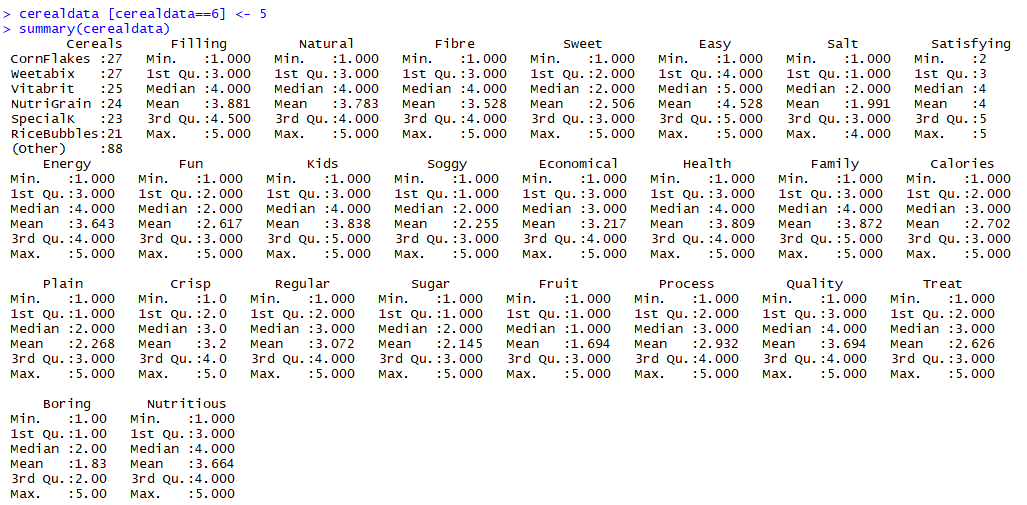
1. Reading the data file and taking a look at the structure:



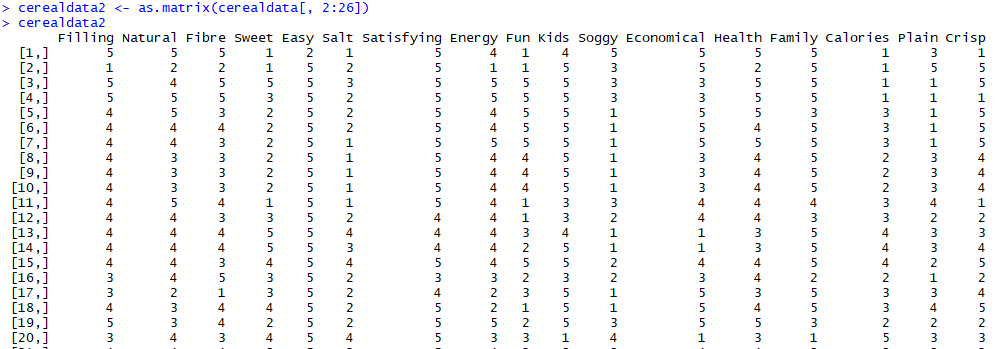
1. And the summary:



1. Interestingly the summary shows some variables having erroneous entries of value 6 as max value. Since we know from the problem statement that likert scale used was of 1-5, we will replace these entries of 6 with 5 & then take a look at the summary again:

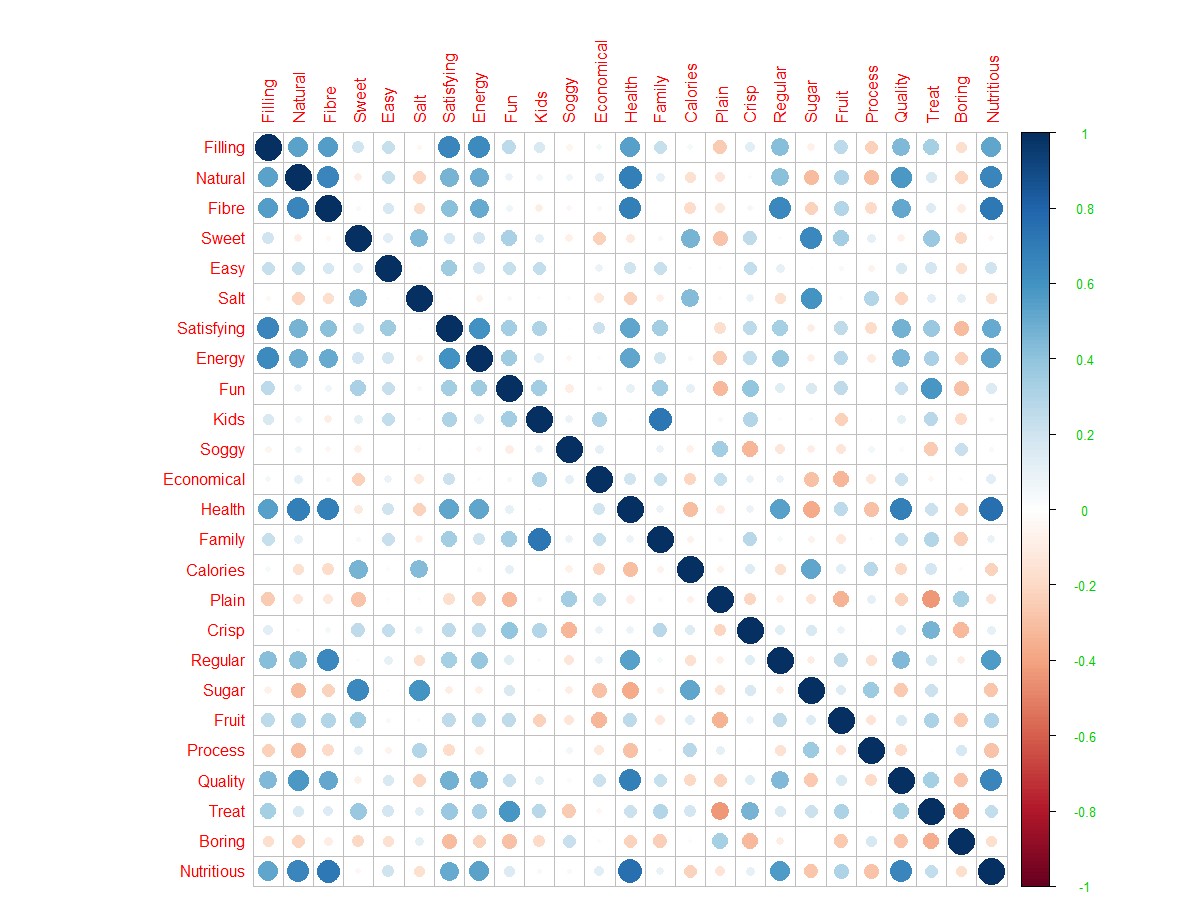


1. Now that the data looks good, we create a matrix out of the dataframe in order to use in the model:



1. Now let’s take a look at the correlation of these variables with each other visually





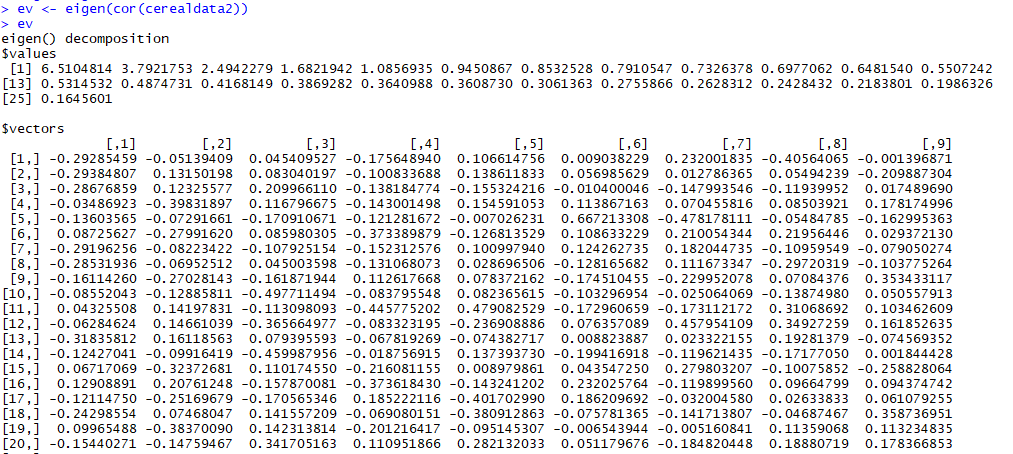
The correlation plot gives us some idea about the possible variable combinations we can have. As an example, some certain combinations of variables with strong correlation with each other that are evidently visible are:

Combination 1: Filling, Natural, Fibre, Satisfying, Energy, Health, Regular, Quality, Nutritious

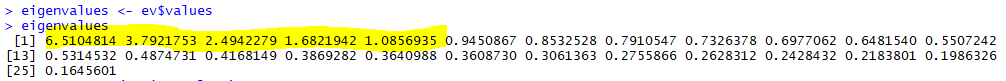
Combination 2: Family, Kids

And so on. Nevertheless, let’s delve further.

1. We now derive the eigen vector / factor loadings of these variables and the eigen values

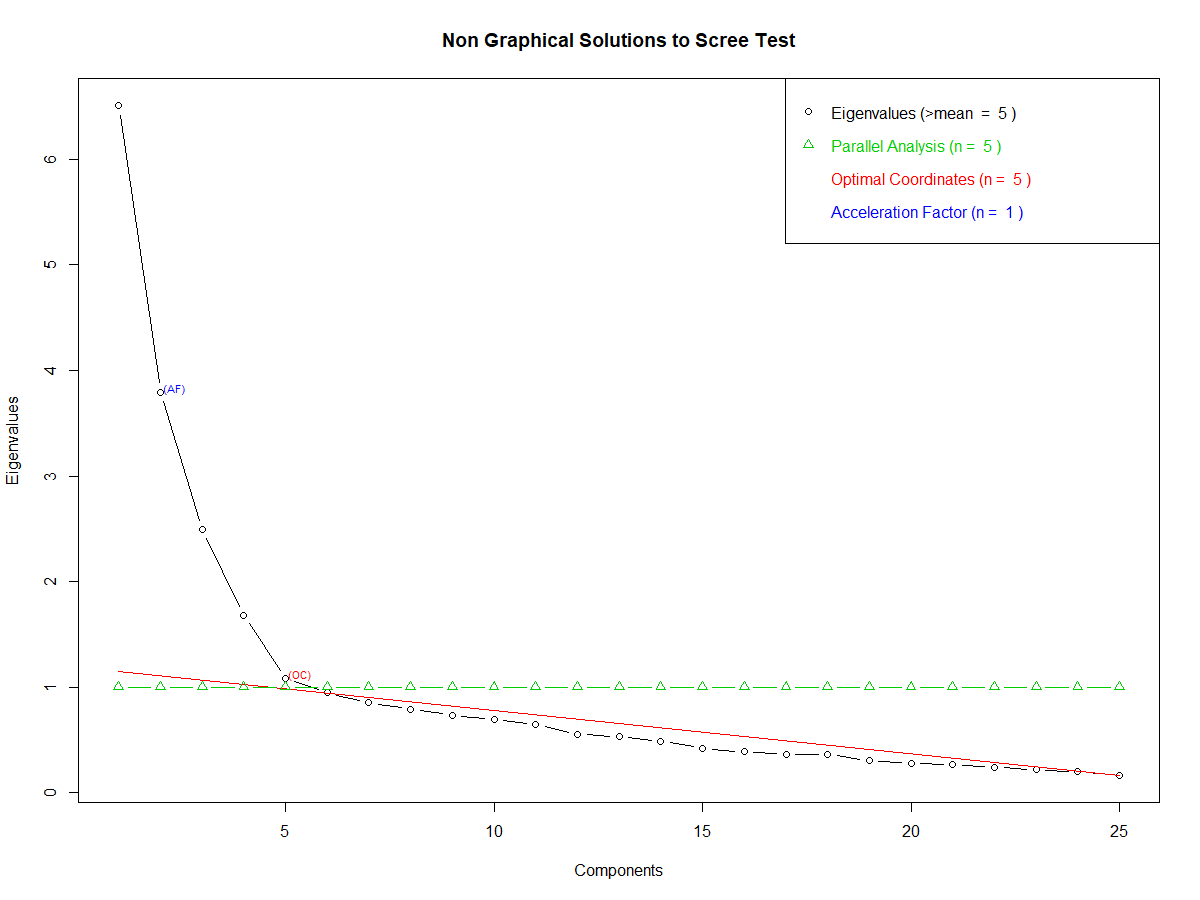


1. The eigen values give us an indication that we could have 5 factors:



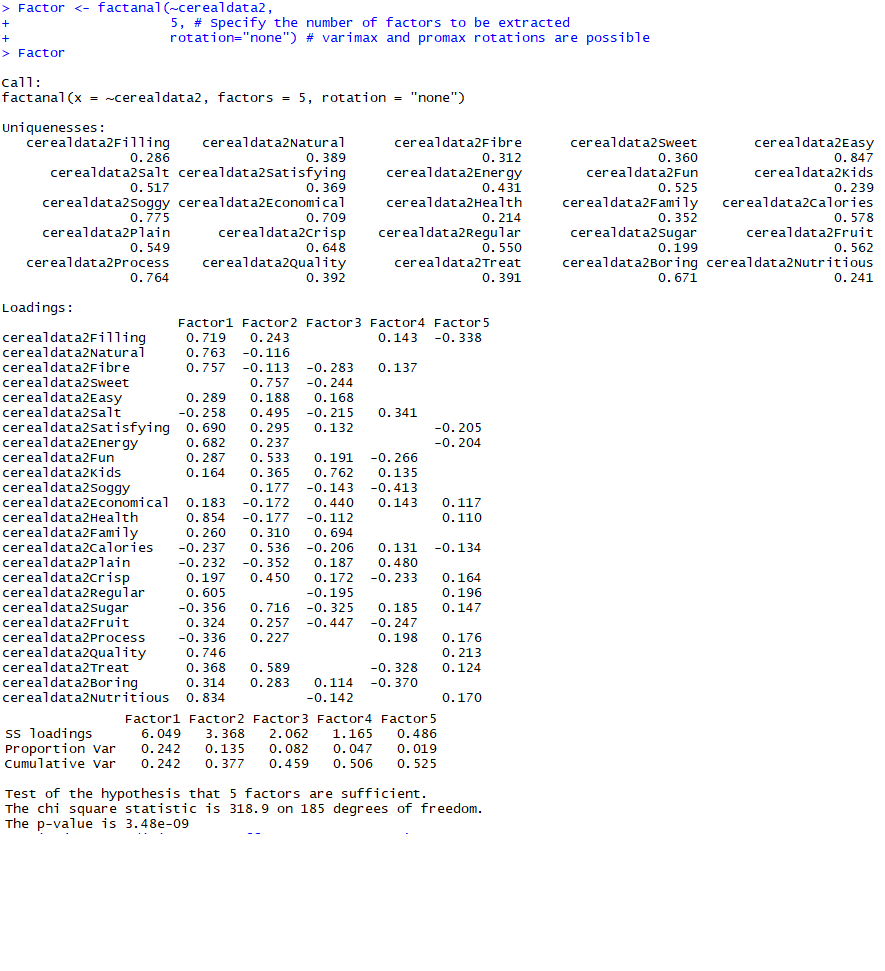
1. We will explore using the Scree Plot as well:





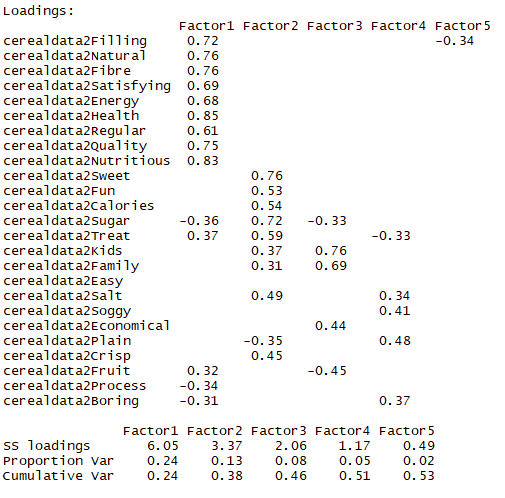
As per the scree plot we can follow the Kaiser rule (where eigen values > 1.0) and consider 5 factors that we can use for the factor analysis.

1. We now perform the factor analysis model with 5 factors and no rotation



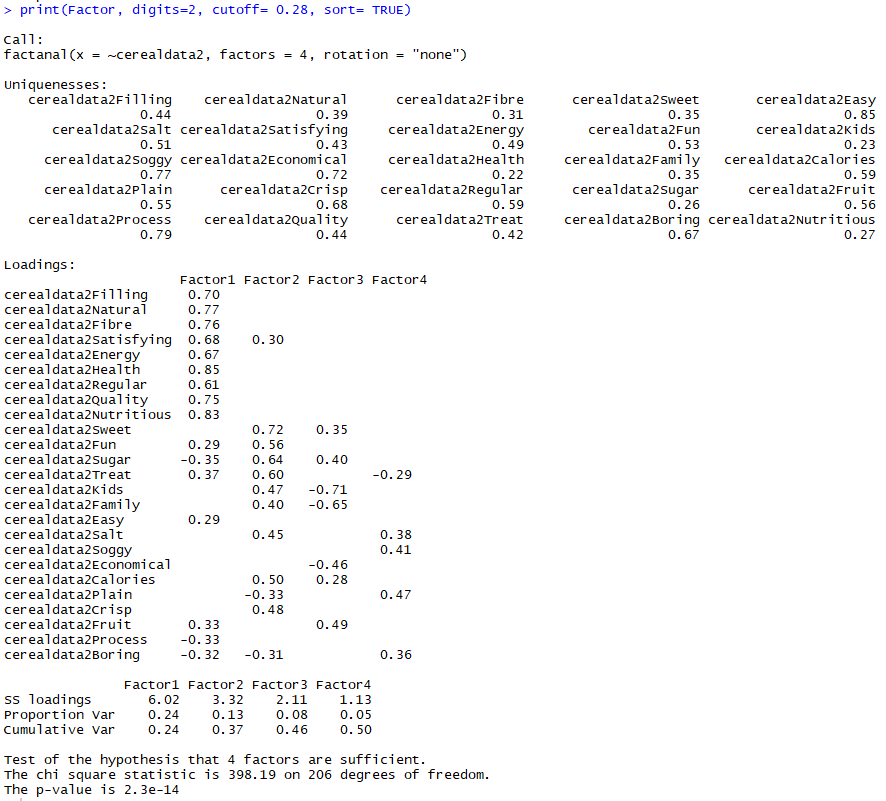
1. Let’s have a better view of these loadings with a random cut-off value of 0.30.



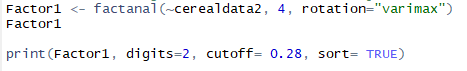


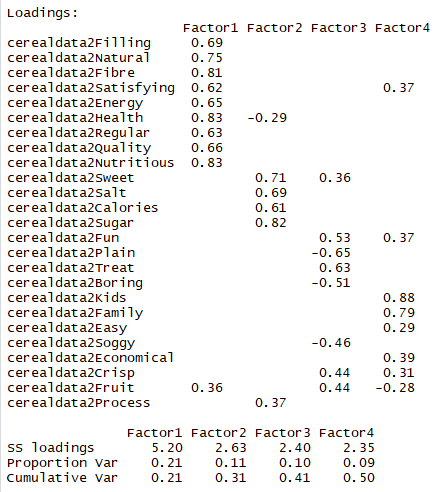
1. We can observe a couple of things:
   1. The 5th factor has almost no loadings. None of the variables is able to explain its variance on this particular 5th factor (except one which has a larger loading in Factor 1)
   2. Some variables, like “Easy” does not have an explanation, which indicates that we may want to revisit our cut off value.
2. Let’s perform the factor analysis with 4 factors and keep the cut off a bit lesser. We need to ensure that we capture maximum communality of the variables across the factors.





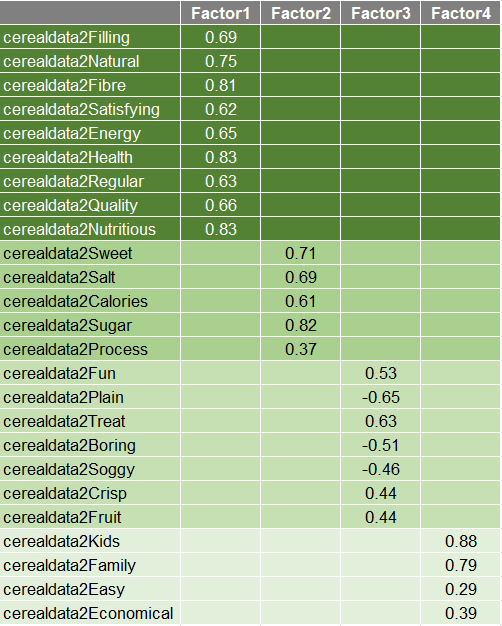
1. We can get a fair idea from above that our initial assumption based on the Corr plot was more or less right. However, the overall picture still looks a bit distorted. If we were to start combining the variables into each factor, we may not be sure where to include, for example, “Process” or “Easy” etc. We would still want to maximize the communality of the variables across the factors. This prompts us to try out one rotation of the factors. We could use Varimax Rotation. It will help maximize the sum of the variance of the squared loadings such that the result is a small number of important variables are highlighted, which makes it easier to interpret your results.





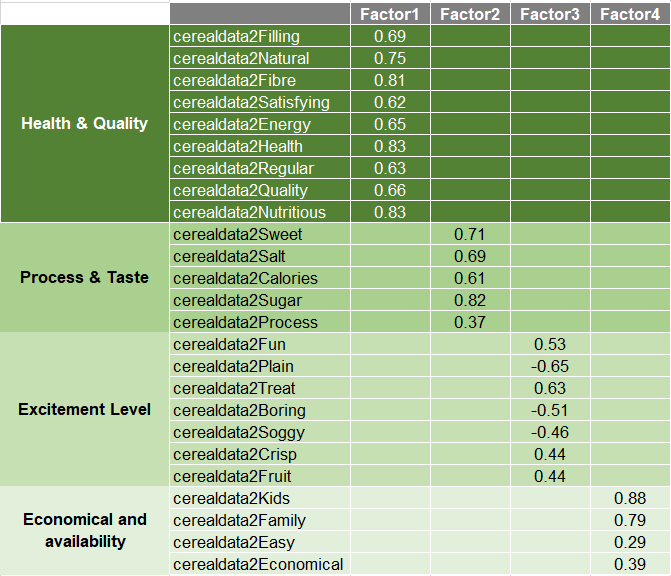
As we see the variance summarized hasn’t changed significantly, i.e., Factor 1 is still high followed by the rest. We tried to capture maximum communality though variances are not very significantly high but this looks still comprehensible.

1. Based on these learnings we can sort variables and factors as below:

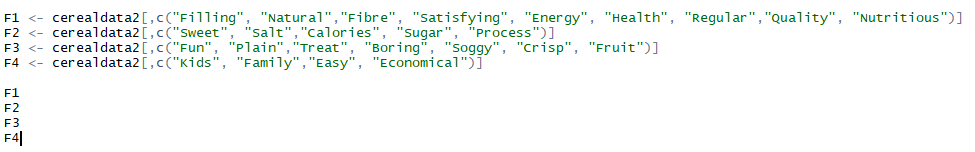


We could see that some variables in Factor 3 such as Plain, Boring and Soggy have a negative value – and rightly so. The customers have chosen in the order of 5 to 1 to indicate their expression range of, for example, “Very Soggy” to “Not Soggy” OR “Very Boring” to “Not Boring”. This is just an assumption. And finally when we analyse the data, the correlation that these variables have goes in a different direction as opposed to all the other variables (i.e., negative).

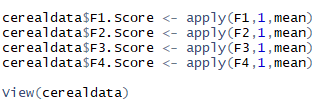
1. We could now characterize the overall consideration behaviours of the 12 selected brands based on the factors and according to the nature of the variables within each factor. Here’s a nomenclature proposition for the characterization of these roups:

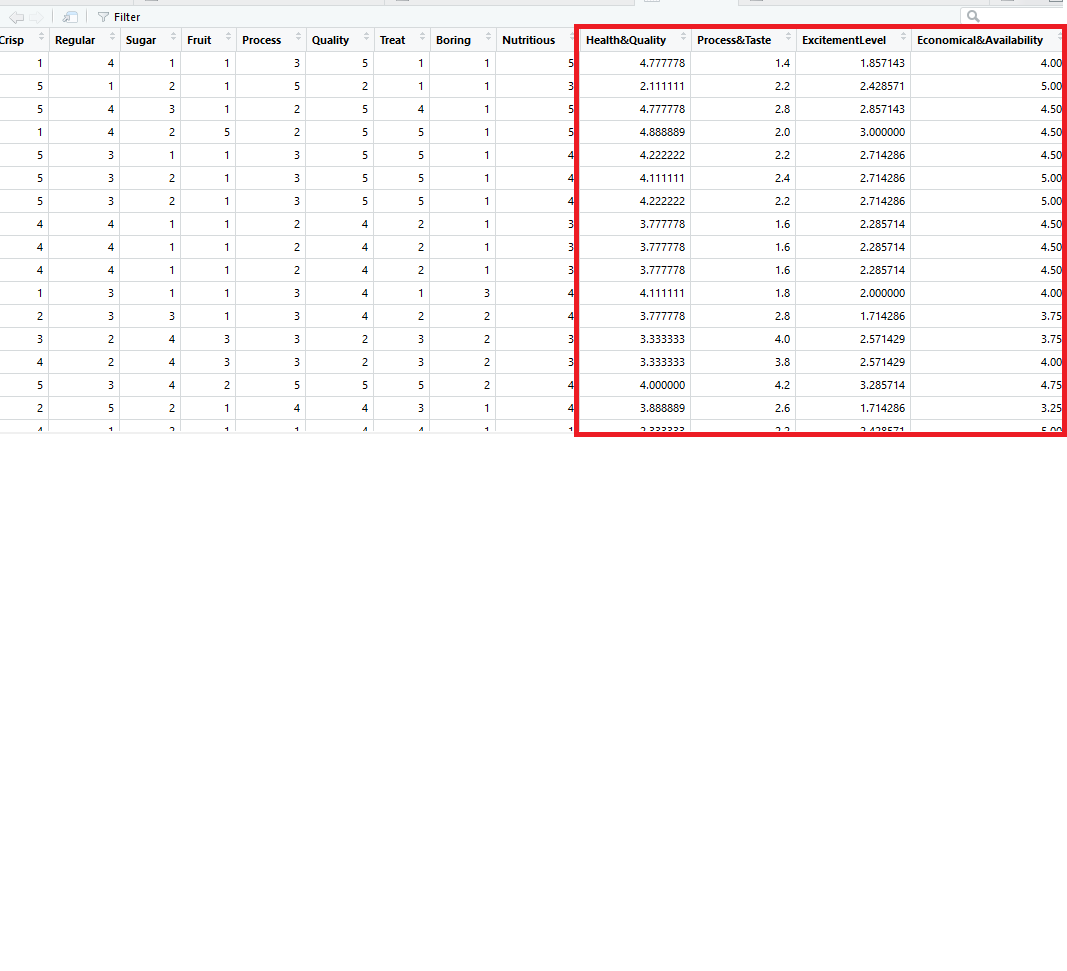


1. We can further factor groups as per the factor loadings:

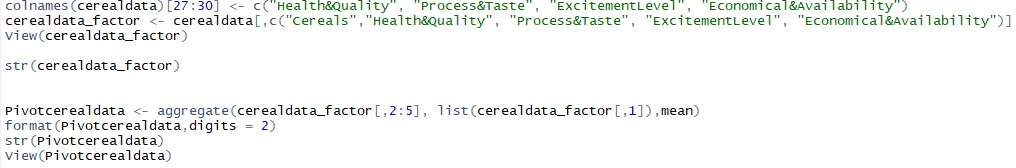


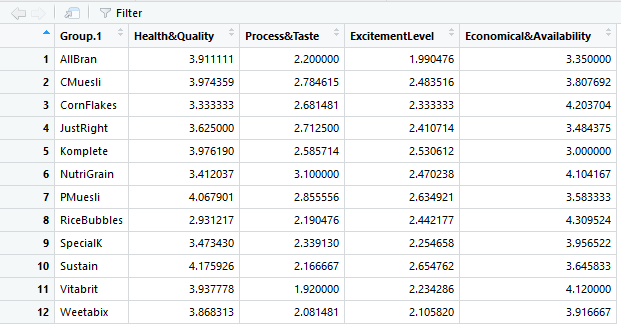
1. And then replacing the scores with the mean values:





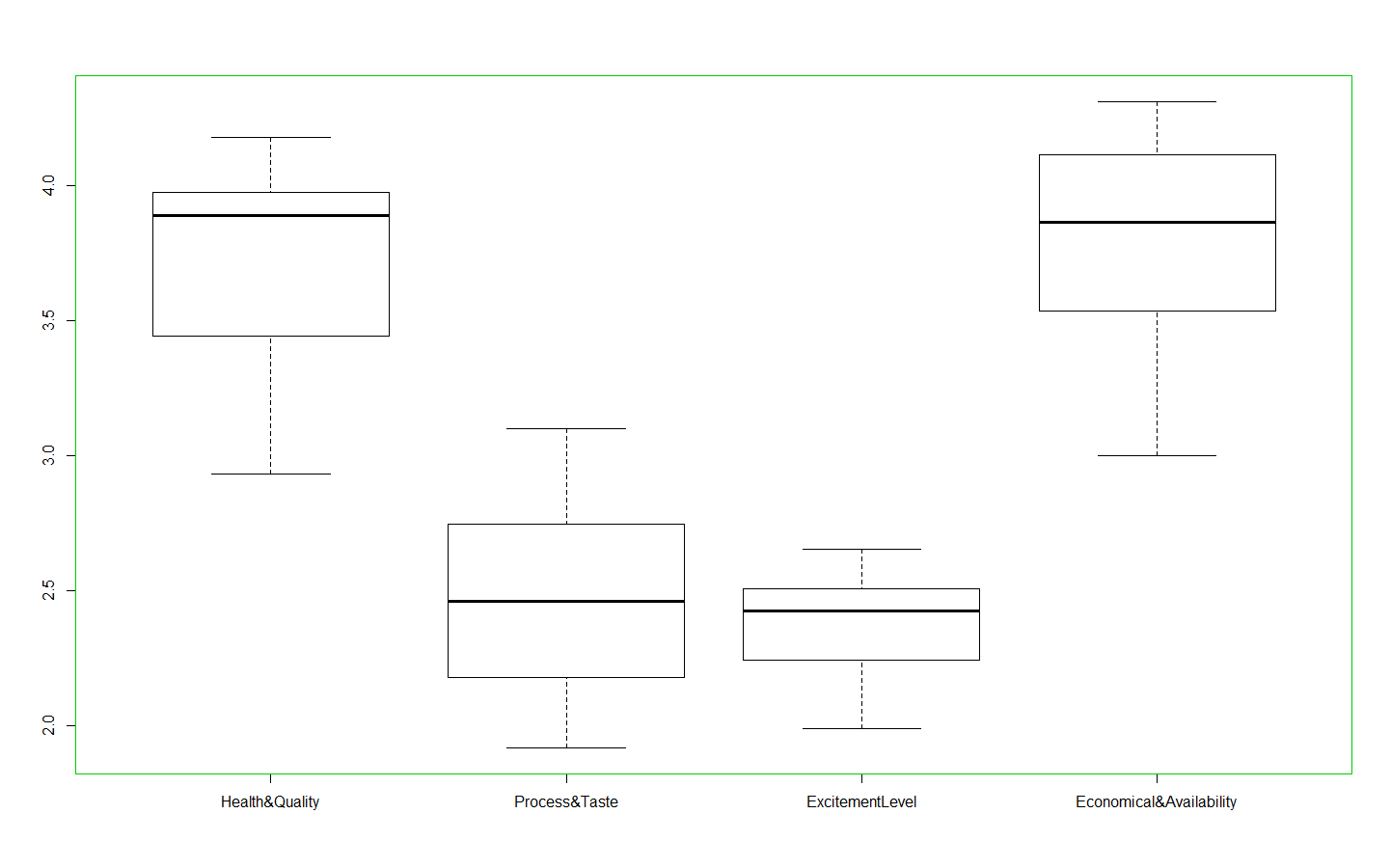
1. And finally considering only the last 4 columns above, i.e., the respective grouping. We also sub-set into a new data frame in order to aggregate and pivot the data based on the respective brands and the groups:





1. Further interpretation:

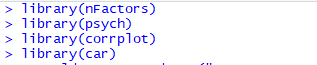
Further interpretation shows that the consumers surveyed have ranked these brands and products high on the health & quality and the economical & availability parameters while quite low on process & taste as well as Excitement Level. We may want to keep in mind that Boring and Soggy are variables capturing negative impression of users and they have influenced the mean of the factor “Excitement level” to be so low.



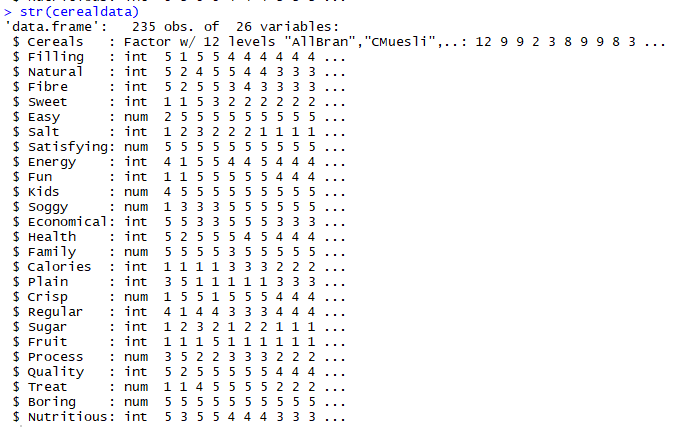
We have also re-done the analysis separately in the following pages by inversing the scores of Soggy and Boring to align with other scores.

We started with exploring the data:

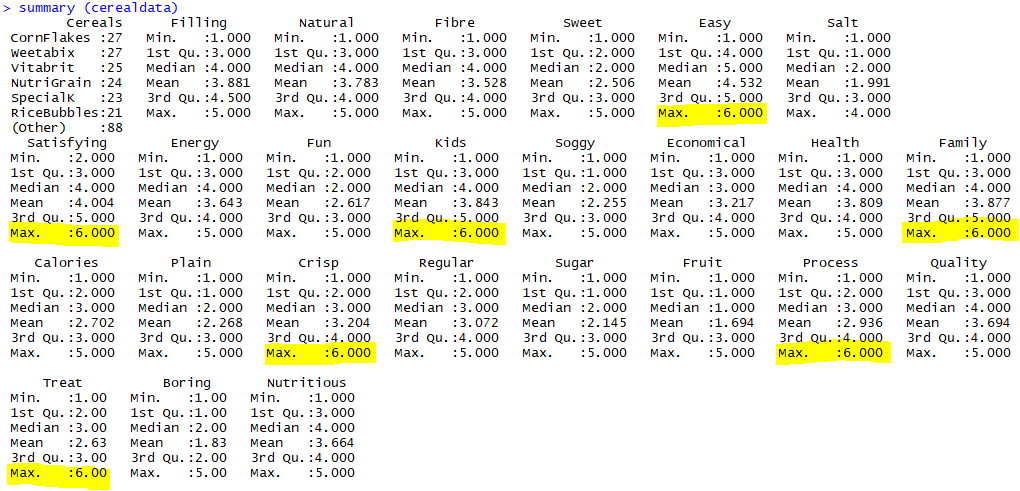
1. We load all required libraries upfront:



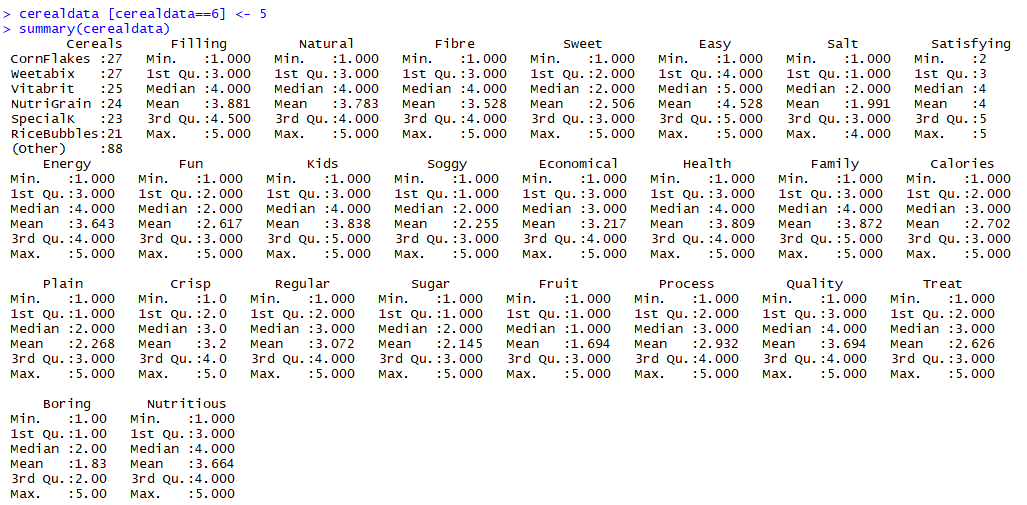
1. Reading the data file and taking a look at the structure:



1. And the summary:



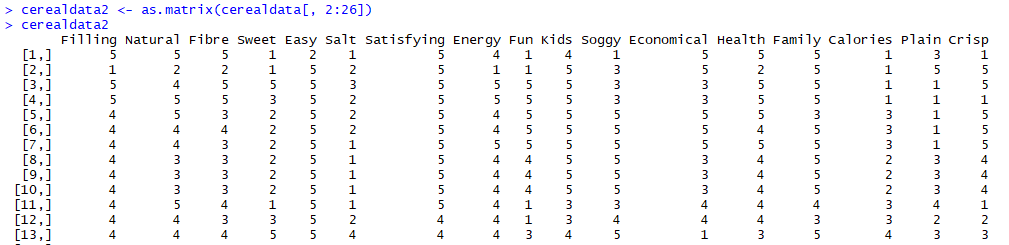
1. We will replace these entries of 6 with 5 & then take a look at the summary again:



1. We assume that some variables such as Plain, Boring and Soggy have a negative value – and rightly so. The customers must have chosen in the order of 5 to 1 to indicate their expression of, for example, “Very Soggy” to “Not Soggy” OR “Very Boring” to “Not Boring”. This is just an assumption. Hence, we inverse the scores of these variables to align with the other scores.

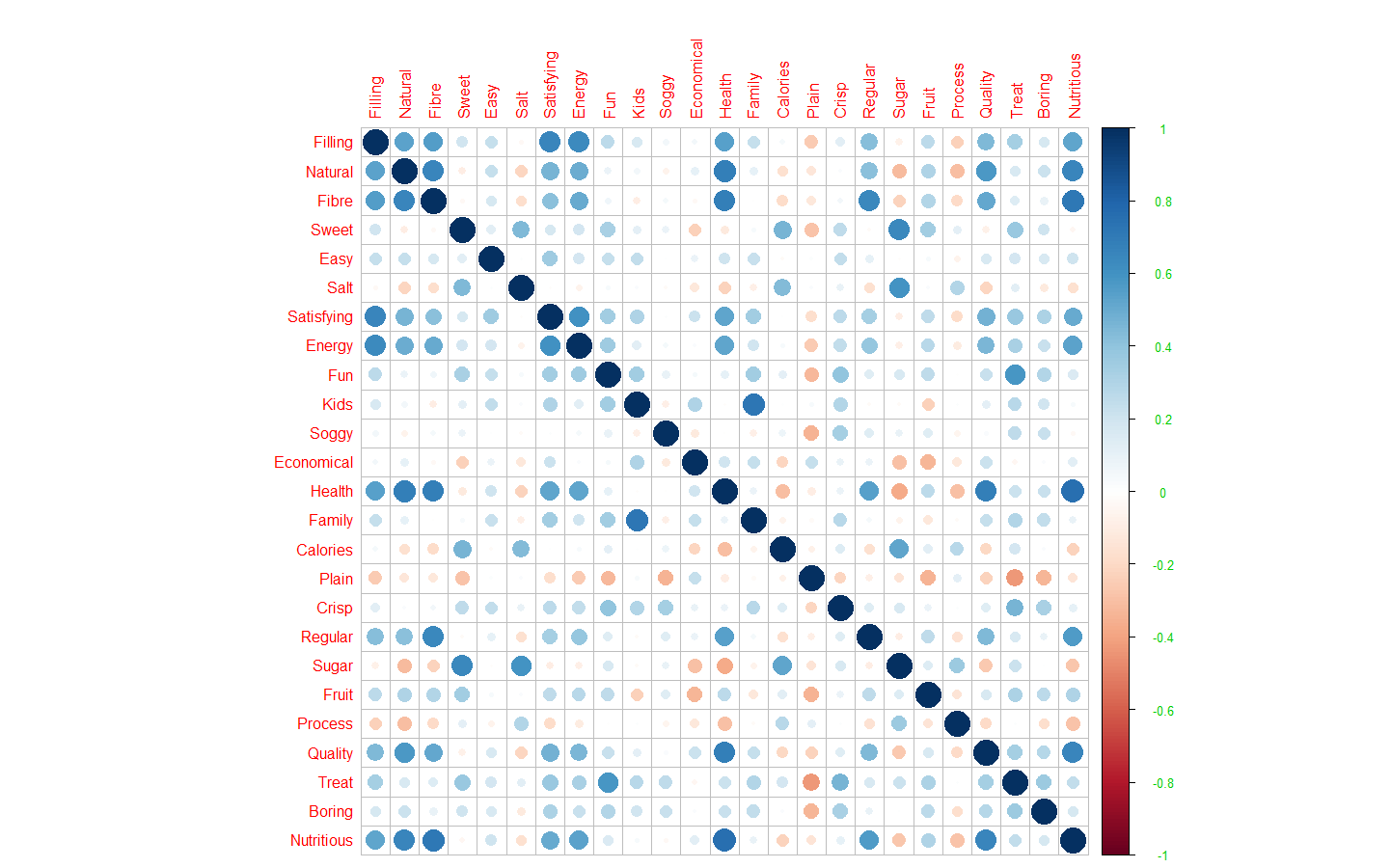


1. Now that the data looks good, we create a matrix out of the data-frame in order to use in the model:



1. Now let’s take a look at the correlation of these variables with each other visually





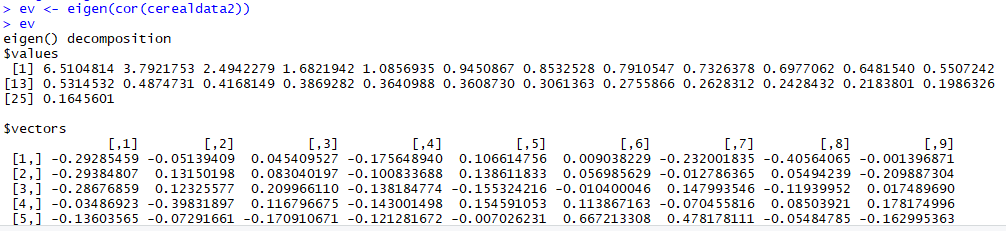
The correlation plot gives us some idea about the possible variable combinations we can have. As an example, some certain combinations of variables with strong correlation with each other that are evidently visible are:

Combination 1: Filling, Natural, Fibre, Satisfying, Energy, Health, Regular, Quality, Nutritious

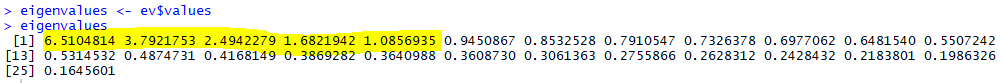
Combination 2: Family, Kids

And so on. Nevertheless, let’s delve further.

1. We now derive the eigen vector / factor loadings of these variables and the eigen values

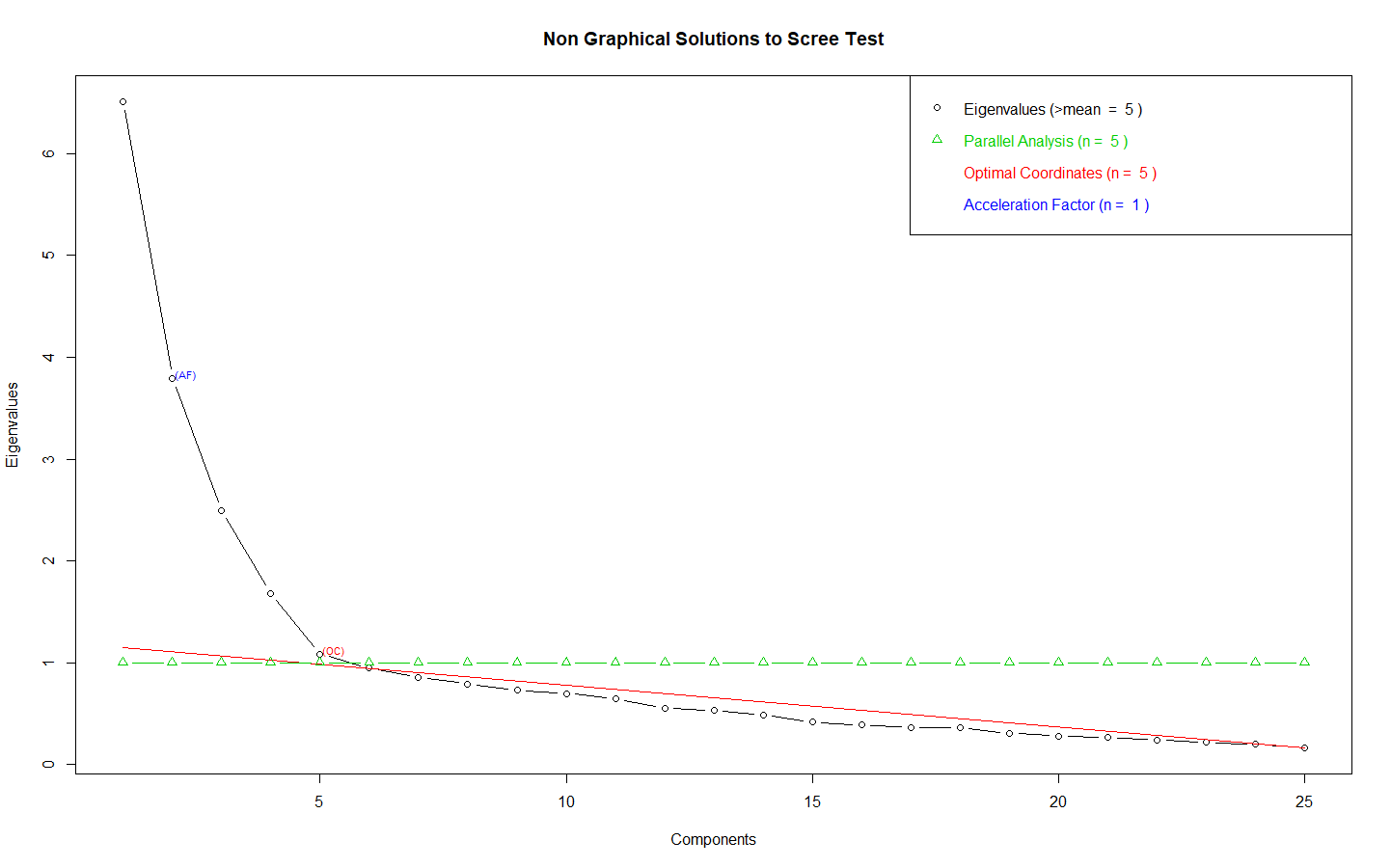


1. The eigen values give us an indication that we could have 5 factors:



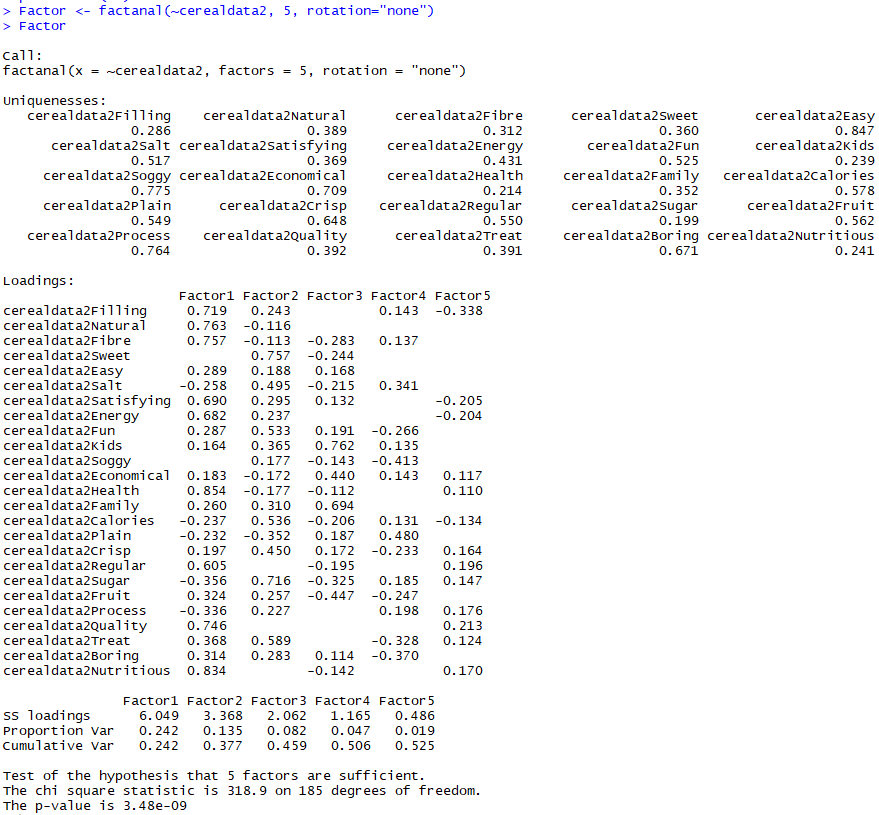
1. We will explore using the Scree Plot as well:





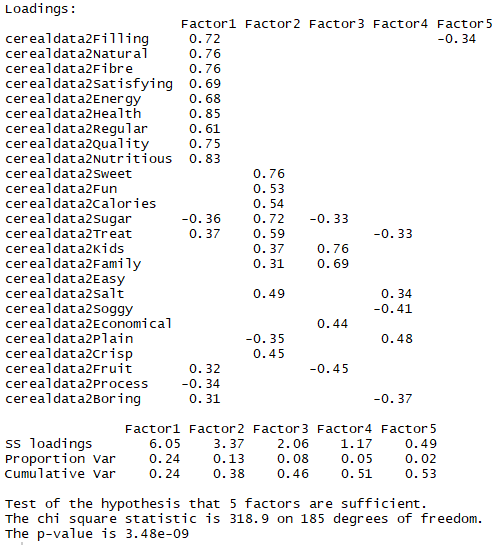
As per the scree plot we can follow the Kaiser rule (where eigen values > 1.0) and consider 5 factors that we can use for the factor analysis.

1. We now perform the factor analysis model with 5 factors and no rotation



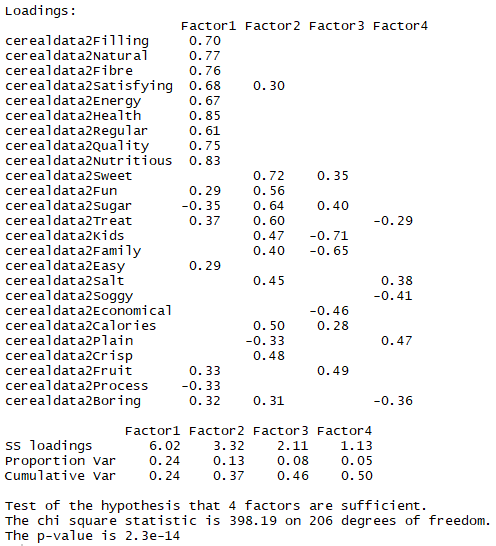
1. Let’s have a better view of these loadings with a random cut-off value of 0.30



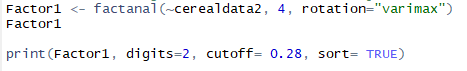


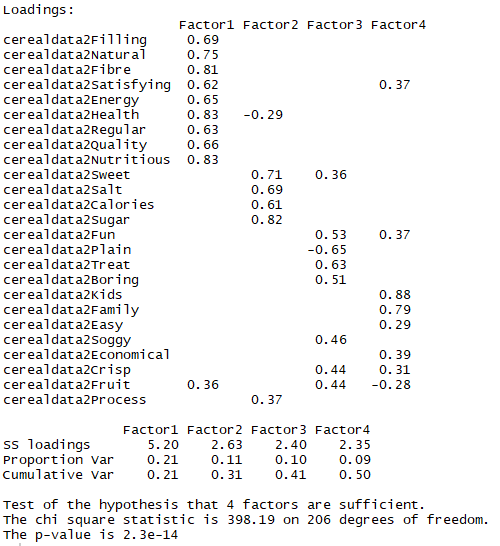
1. We can observe a couple of things:
   1. The 5th factor has almost no loadings. None of the variables is able to explain its variance on this particular 5th factor (except one which has a larger loading in Factor 1)
   2. Some variables, like “Easy” does not have an explanation, which indicates that we may want to revisit our cut off value.
2. Let’s perform the factor analysis with 4 factors and keep the cut off a bit lesser. We need to ensure that we capture maximum communality of the variables across the factors.





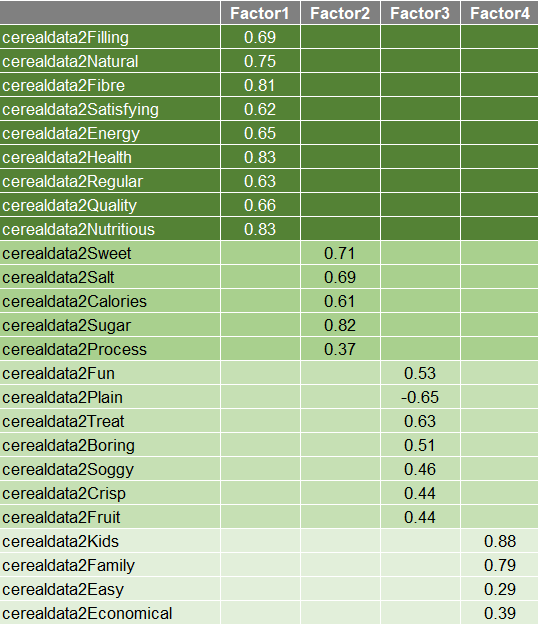
1. We can get a fair idea from above that our initial assumption based on the Corr plot was more or less right. However, the overall picture still looks a bit distorted. If we were to start combining the variables into each factor, we may not be sure where to include, for example, “Process” or “Easy” etc. We would still want to maximize the communality of the variables across the factors. This prompts us to try out one rotation of the factors. We could use Varimax Rotation. It will help maximize the sum of the variance of the squared loadings such that the result is a small number of important variables are highlighted, which makes it easier to interpret the results.





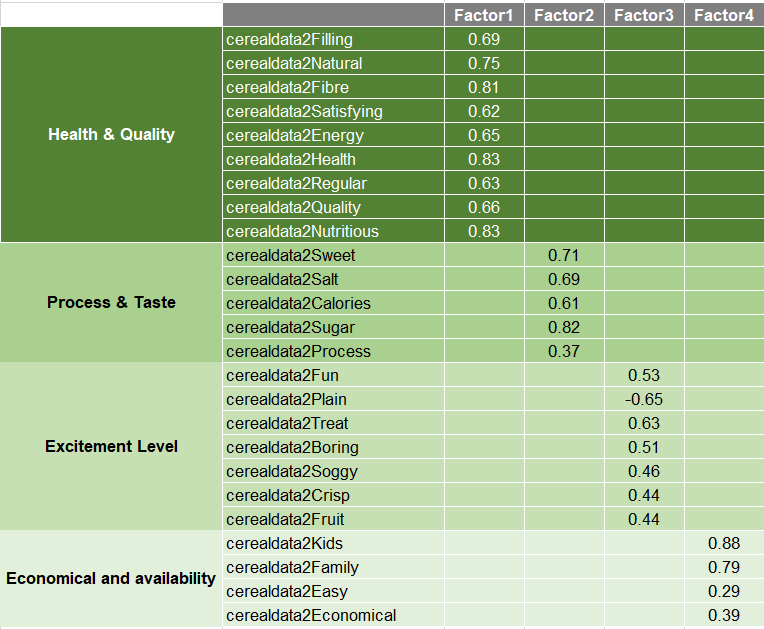
As we see the variance summarized hasn’t changed significantly, i.e., Factor 1 is still high followed by the rest. We tried to capture maximum communality though variances are not very significantly high but this looks still comprehensible.

1. Based on these learnings we can sort variables and factors as below:

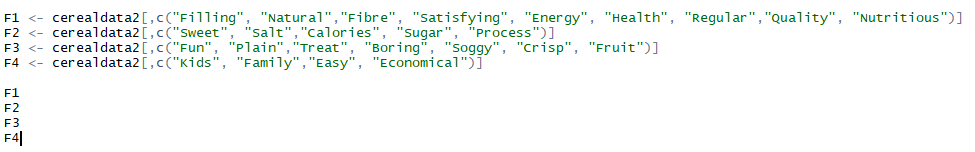


We could see that some variables in Factor 3 such as Plain, has a negative value. The customers must have chosen in the order of 5 to 1 to indicate their expression range of, for example, “Very Plain” to “Not Plain”. This is just an assumption.

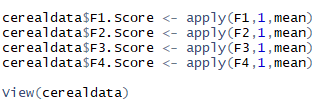
1. We could now characterize the overall consideration behaviours of the 12 selected brands based on the factors and according to the nature of the variables within each factor. Here’s a nomenclature proposition for the characterization of these roups:

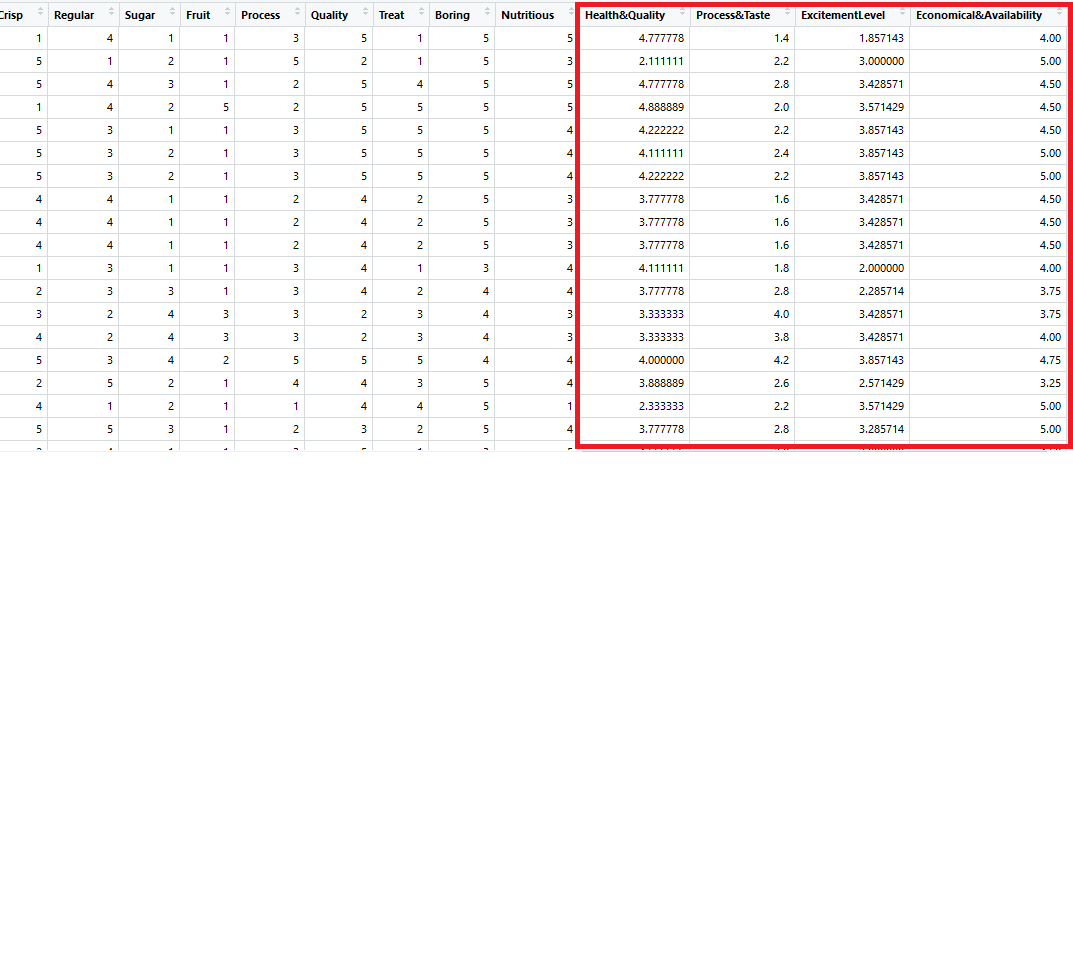


1. We can further factor groups as per the factor loadings:

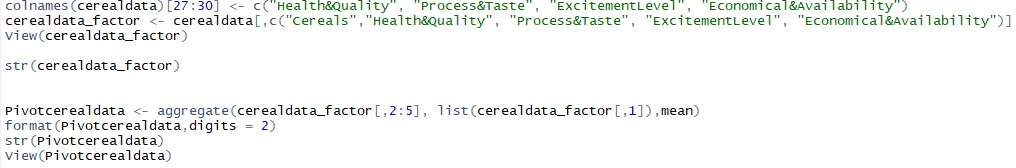


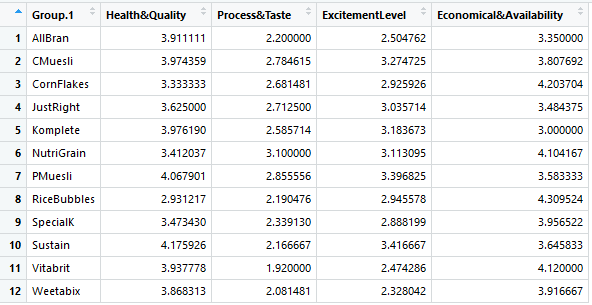
1. And then replacing the scores with the mean values:





1. And finally considering only the last 4 columns above, i.e., the respective grouping. We also sub-set into a new data frame in order to aggregate and pivot the data based on the respective brands and the groups:





1. Further interpretation:

Further interpretation shows that the consumers surveyed have ranked these brands and products high on “Economical and Availability” (i.e., price, availability of the product, suitability for family and kids); followed by “Health & Quality” (good nutrition, healthy, fibrous and so on) and then followed by the “Excitement Level” and “Process & Taste”. The companies may want to focus on improving the process & taste in general that has such a low mean rating of ~2.5. They may want to experiment with the excitement factor a bit more – change existing flavours, add new flavours or adopt other strategies.

