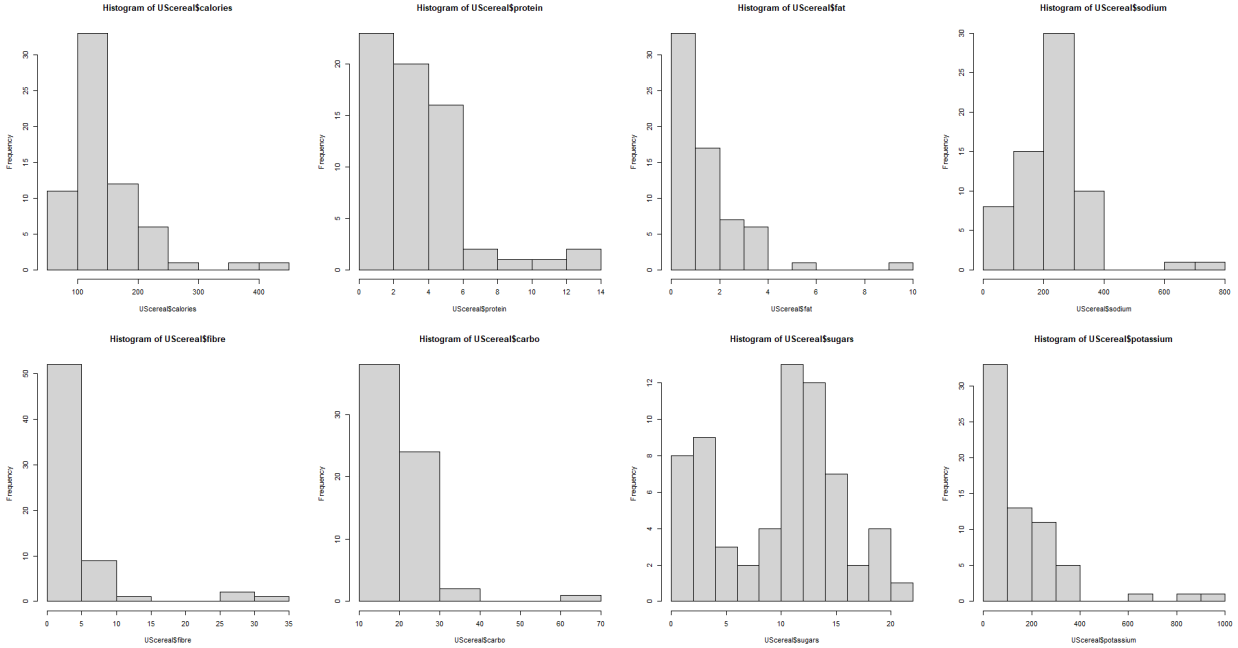
Data frame `UScereal` distributed with the MASS library (install MASS then type `library(mass)`) describes 65 commonly available breakfast cereals in the US, based on information available on the mandatory food label on the package. The measurements are normalized to a serving size of one American cup (see help page `?UScereal`). I will use the term *numerical nutrition variables* to mean all variables except for `mfr`, `shelf` and `vitamins` (variables 2:8, 10). This data set is an exercise in the classic V&R3, where the authors asked, (Q1) Is there any way to discriminate among the major manufacturers by cereal characteristics, or do they each have a balanced portfolio of cereals? (Q2) Are there interpretable clusters of cereals? (Q3) Can you describe why cereals are displayed on high, low or middle shelves? I add: (Q4) How to visualize the data and show clusters or manufacturers? Keep the following in mind as you search for clusters:

- This problem addresses a core topic in marketing management, **market segmentation** (as opposed to **customer segmentation** discussed earlier, where the customers in some market segment are further partitioned into smaller groups). Customers have heterogeneous wants and needs. Large manufacturers like General Mills identify clusters (segments) based on these wants and needs and then create ("target") a brand for each segment.[1] Smaller manufacturers may not be able to compete across segments, and instead adopt a **niche strategy**, where they focus on small segment with unique needs that large manufactures may ignore. Your clusters should identify the mainstream segments as well as niches. Visualizations should show the strategies of different manufactures.
- Should you standardize and/or transform the data prior to clustering and/or using PCA/FA? Should you cluster on the raw numerical variables, PCs or factors? Are there interpretable clusters of cereals?

After checking the histogram of each numerical variable, we found that the sugars variable is roughly normally distributed, but the rest of the variables have a skewed distribution.



Therefore, we will perform log-transformation for these variables: calories, protein, fat, sodium, fiber, carbo, and potassium. We use log(x+1) for some variables to avoid performing log-transformation on value 0.

```
#log-transformation
UScereal$log_calories = log(UScereal$calories)
UScereal$log_protein = log(UScereal$protein)
UScereal$log_fat = log(UScereal$fat + 1)
UScereal$log_sodium = log(UScereal$sodium + 1)
UScereal$log_fibre = log(UScereal$fibre + 1)
UScereal$log_carbo = log(UScereal$carbo)
UScereal$log_potassium = log(UScereal$potassium)
```

The units of these numerical variables in the dataset are incommensurate, for example, protein is in “gram” and sodium is in “milligrams”. Therefore, we definitely need to standardize the data prior to clustering.

calories

number of calories in one portion.

protein

grams of protein in one portion.

fat

grams of fat in one portion.

sodium

milligrams of sodium in one portion.

fibre

grams of dietary fibre in one portion.

carbo

grams of complex carbohydrates in one portion.

sugars

The summary for standardized variables:

```
> #Standardization
> ZCereal =data.frame(lapply(UScereal[,c(8,12,13,14,15,16,17,18)], scale, scale=T)) #log-transformed variables
> summary(ZCereal)
     sugars         log_calories      log_protein          log_fat          log_sodium         log_fibre          log_carbo         log_potassium
 Min.   :-1.7224   Min.   :-2.8721   Min.   :-1.98239   Min.   :-1.17029   Min.   :-3.31102   Min.   :-1.20898   Min.   :-1.69069   Min.   :-1.90964
 1st Qu.:-1.0369   1st Qu.:-0.6678   1st Qu.:-0.55808   1st Qu.:-1.17029   1st Qu.: 0.09795   1st Qu.:-1.20898   1st Qu.:-0.65232   1st Qu.:-0.79507
 Median : 0.3340   Median :-0.1091   Median : 0.03222   Median :-0.01258   Median : 0.26355   Median :-0.01902   Median :-0.01116   Median :-0.02015
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.00000
 3rd Qu.: 0.6768   3rd Qu.: 0.6952   3rd Qu.: 0.61526   3rd Qu.: 0.66464   3rd Qu.: 0.40932   3rd Qu.: 0.63310   3rd Qu.: 0.52181   3rd Qu.: 0.81494
 Max.   : 1.8585   Max.   : 3.2081   Max.   : 2.06510   Max.   : 2.69067   Max.   : 1.06330   Max.   : 2.52108   Max.   : 3.77901   Max.   : 2.31984
```

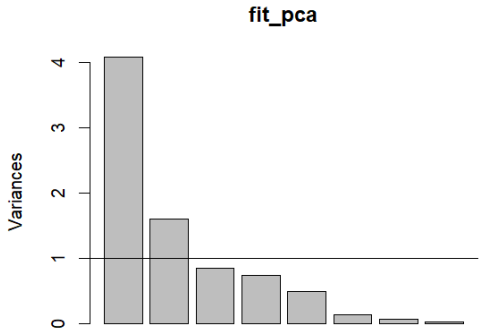
We decide to cluster on the PCs or factors. By doing that, we could reduce the dimension and make the final results easy to interpret.

- I suggest examining both PCs and varimax rotated PCs. Which do you prefer? Plotting components in a two-dimensional scatterplot is a form of a perceptual map. Empty regions may reveal opportunities to launch new brands (or cereals that nobody wants to eat!).

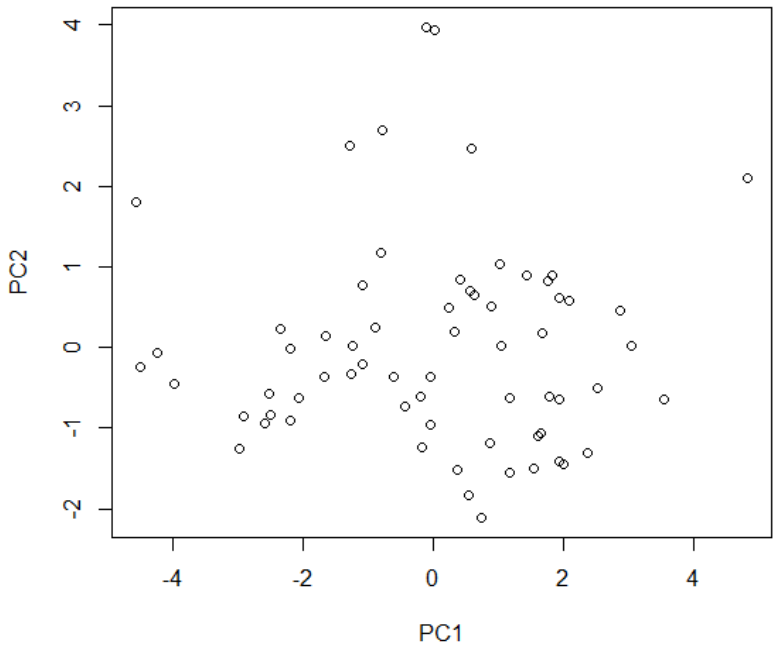
Firstly, we examine PCs.

```
> #PC
> fit_pca <- prcomp(ZCereal)
> summary(fit_pca)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     2.0203 1.2666 0.9228 0.86055 0.69831 0.37469 0.26315 0.15712
Proportion of Variance 0.5102 0.2005 0.1064 0.09257 0.06096 0.01755 0.00866 0.00309
Cumulative Proportion  0.5102 0.7107 0.8172 0.90975 0.97071 0.98826 0.99691 1.00000
> fit_pca$rotation ##hard to interpret
                    PC1        PC2         PC3           PC4          PC5         PC6         PC7         PC8
sugars        -0.2292649 -0.5361194  0.31595249 -0.4499738163  0.415033298 -0.07809532 -0.04819401 -0.4207576
log_calories  -0.4413273 -0.1337862 -0.24817961 -0.3652908087  0.139113216 -0.01767775  0.17914290  0.7352168
log_protein   -0.4344383  0.2492395  0.07026645  0.1929388549 -0.027861785 -0.83345878 -0.02631915 -0.1042006
log_fat       -0.2893372 -0.4470491 -0.01315259  0.0006530075 -0.830211143  0.05641137  0.09282946 -0.1233190
log_sodium    -0.1378265 -0.4868964 -0.42008174  0.6782141682  0.323185402  0.04157928 -0.01648002 -0.0323020
log_fibre     -0.4099164  0.2513937  0.36119399  0.2523682629  0.116782176  0.36887737  0.64028705 -0.1222407
log_carbo     -0.2999787  0.3315640 -0.67314600 -0.2831645415  0.016403882  0.19347933 -0.03645662 -0.4772190
log_potassium -0.4507775  0.1500080  0.27095248  0.1517036779 -0.003476412  0.34715881 -0.73804266  0.1123820
> plot(fit_pca$x)
> screeplot(fit_pca)
> abline(h=1) ##Choose 2 PC
```

The Kaiser criterion suggests that 2 PCs are enough and these 2 PCs can explain 71.07% of the total variation for this dataset.



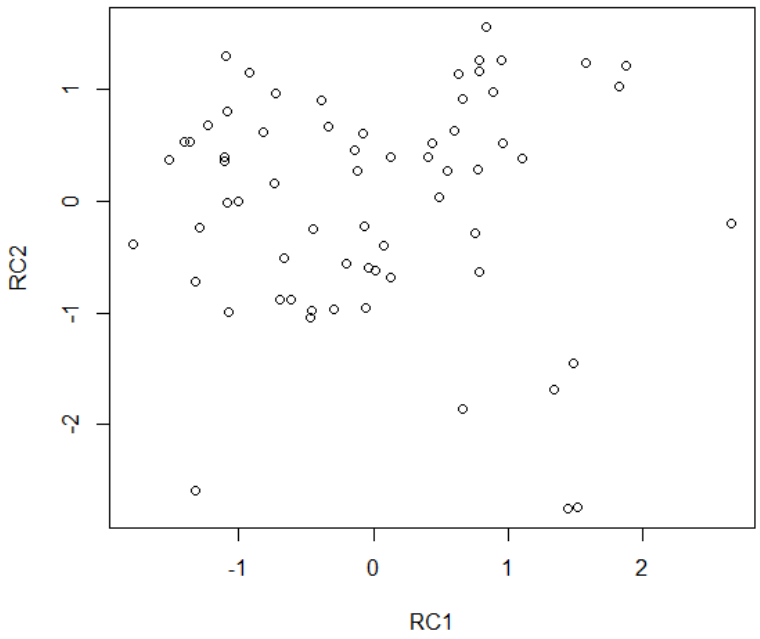
The PC1 vs PC2 plot is listed below. Roughly 4-6 clusters can be observed in the plot.



Secondly, we examine varimax rotated PCs.

```
> #varimax rotated PC
> library(psych)
> fit_pca_varimax <- principal(ZCereal, nfactor=2)
> print(fit_pca_varimax$loadings, cutoff=0.3, digits=3, sort=TRUE)

Loadings:
              RC1    RC2
log_calories   0.708  0.568
log_protein    0.923
log_fibre      0.881
log_carbo      0.732
log_potassium  0.893
sugars                0.817
log_fat               0.774
log_sodium            0.675

                 RC1   RC2
SS loadings    3.537 2.149
Proportion Var 0.442 0.269
Cumulative Var 0.442 0.711
> plot(fit_pca_varimax$scores)
```

The PC1 vs PC2 plot is listed below. Roughly 4-6 clusters can be observed in the plot.



We decided to use varimax rotated PCs for clustering, because the two rotated factors are easy to interpret. RC1 consists of calories, protein, fiber, carbohydrates, and potassium. This factor can be labeled as the “Healthy Nutrients” factor. RC2 consists of sugars, fat, and sodium. This factor can be labeled as the “Unhealthy Nutrients” factor.

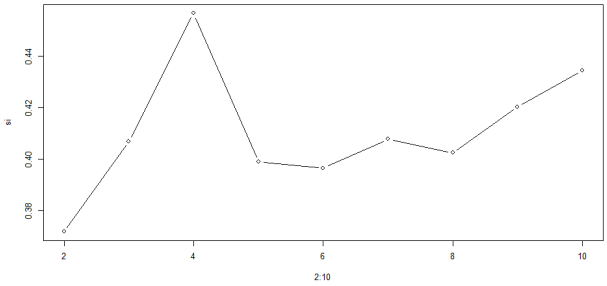
```
> print(fit_pca_varimax$loadings, cutoff=0.3, digits=3, sort=TRUE)

Loadings:
              RC1   RC2
log_calories  0.708 0.568
log_protein   0.923
log_fibre     0.881
log_carbo     0.732
log_potassium 0.893
sugars              0.817
log_fat             0.774
log_sodium          0.675

                 RC1   RC2
SS loadings    3.537 2.149
Proportion Var 0.442 0.269
Cumulative Var 0.442 0.711
```

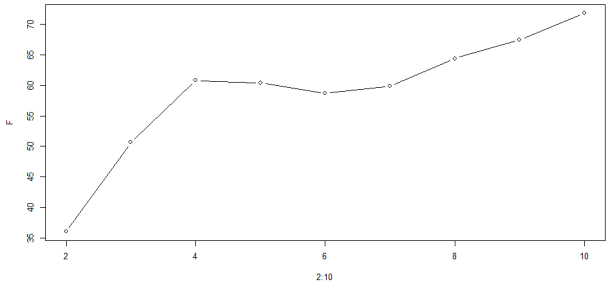
In the later stage, we will also analyze the 3-dimensional results for clustering in order to find some extra information that we might omit using 2-dimensional results. On top of the previous two factors, the 3rd factor represents “High Energy”. We see that the third factor has high values of carbohydrates, calories, some protein, and low values of sugar. Therefore, this factor can be thought of as representing high energy or cereal for fueling athletic activities.

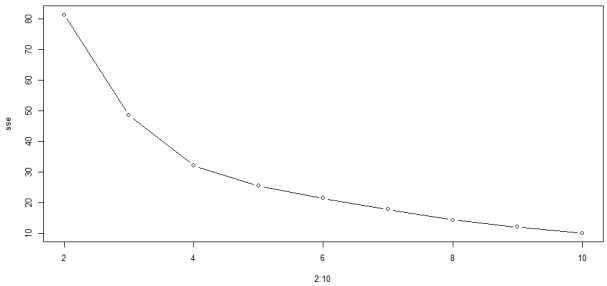
```
> #3 factors
> fit_pca_varimax_3 <- principal(Zcereal, nfactor=3)
> print(fit_pca_varimax_3$loadings, cutoff=0.3, digits=3, sort=TRUE)

Loadings:
              RC1    RC2    RC3
log_protein    0.861          0.344
log_fibre      0.942
log_potassium  0.926
sugars                0.764 -0.326
log_calories   0.566  0.597  0.446
log_fat               0.768
log_sodium            0.726
log_carbo      0.367          0.891

                 RC1   RC2   RC3
SS loadings    3.112 2.122 1.303
Proportion Var 0.389 0.265 0.163
Cumulative Var 0.389 0.654 0.817
```

- Which clustering methods do you suggest? Why?
- I think that unsupervised (PCA/cluster) methods are the way to go with Q1, Q2 and Q4, but one could make a case to use supervised for Q3 (or parts of Q1).

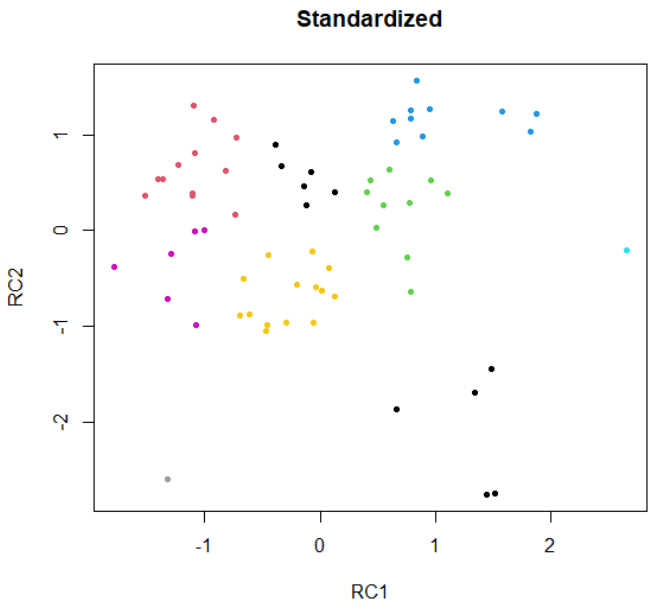
We first examine SSE, F, and Silhouette statistics and plots in the k-means' setting. The highest Silhouette value and a spike in F can be observed when K is 4. However, our purpose is to identify market segments especially the niche market segments. Clustering the market into 4 groups will not serve this purpose. We need to choose a bigger K in order to understand the market niches and unlock the potential of underserved customer blocs.



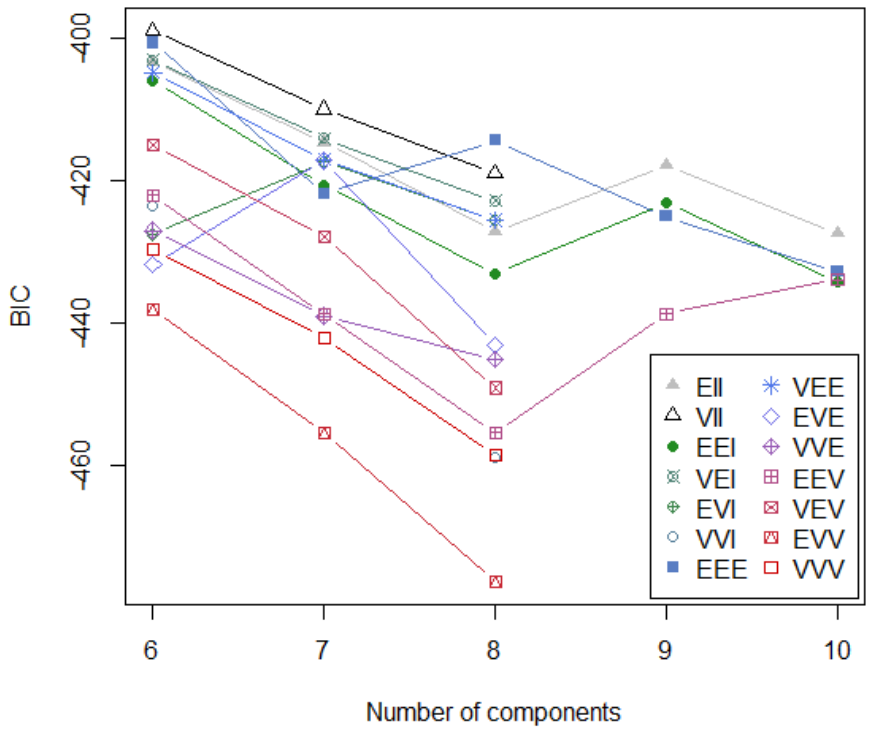




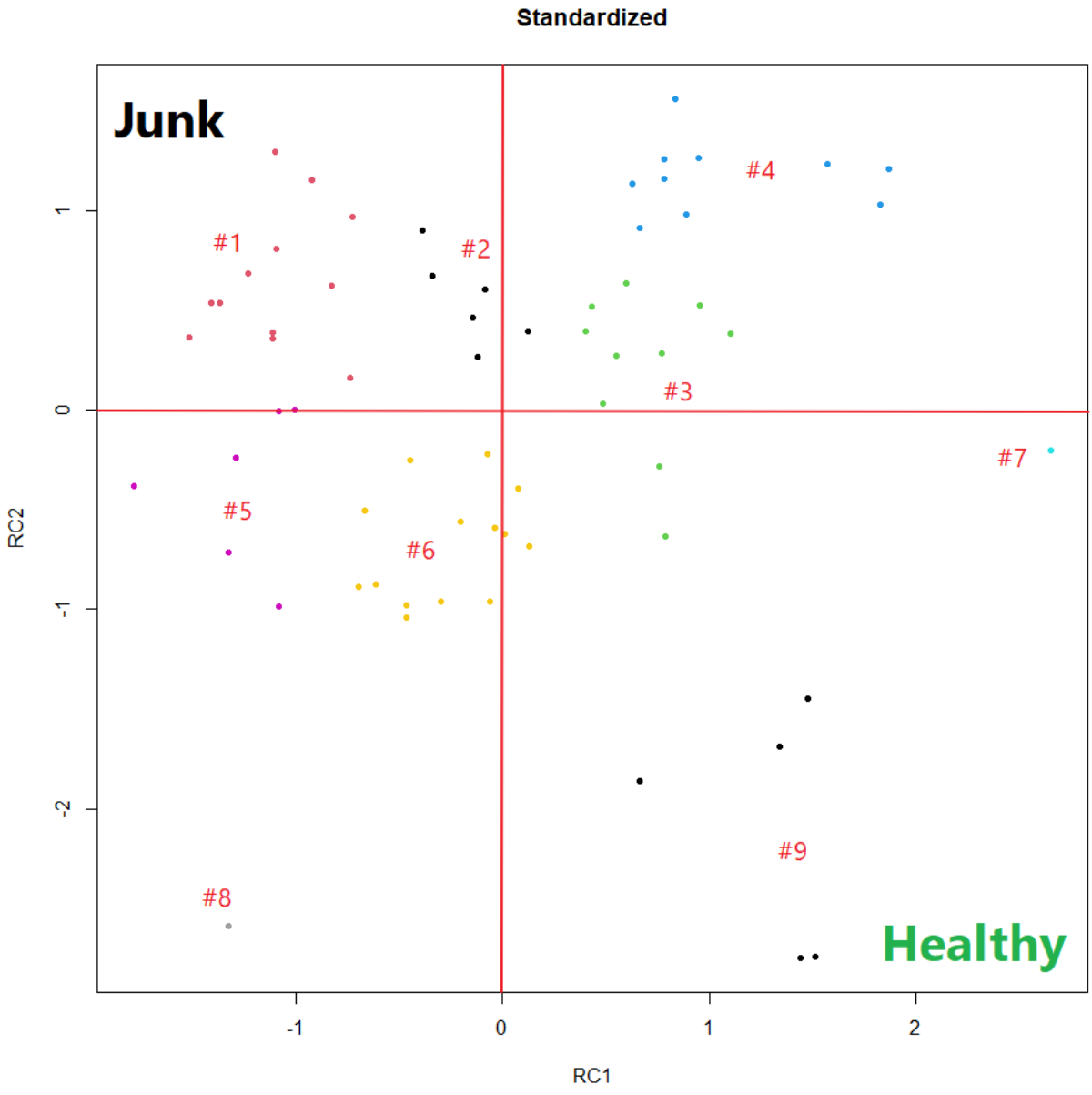
We try to cluster the market into 6 to 10 groups. After examining the plots, we find that the 9-cluster solution is ideal, because it identifies the outliers as well as the clusters.



Then we try to use Gaussian Mixtures to cluster the data into 6 to 10 groups. The BIC graph suggests that the 9-cluster solution with the EII model has a relatively high BIC value. EII produces round clusters of equal “volume”, which is equivalent to the k-means method. As the result of that, we suggest using the k-means method and cluster the data into 9 groups.



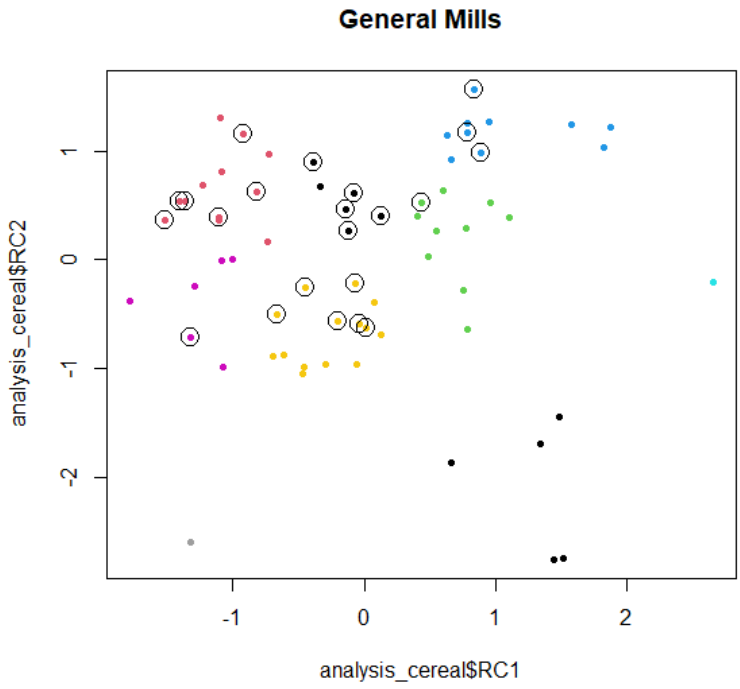
These 9 clusters are renamed as below:
#1: [**Junk++**] Low Healthy Nutrients, High Unhealthy Nutrients
#2: [**Junk**] Average Healthy Nutrients, High Unhealthy Nutrients
#3: [**Mixed-Lean-Healthy**] High Healthy Nutrients, Average Unhealthy Nutrients
#4: [**Mixed-Lean-Healthy++**] High Healthy Nutrients, High Unhealthy Nutrients
#5: [**Non Nutrient++**] Low Healthy Nutrients, Low Unhealthy Nutrients
#6: [**Non Nutrient**] Average Healthy Nutrients, Low Unhealthy Nutrients
#7: [**Healthy**] Extremely High Healthy Nutrients, Average Unhealthy Nutrients
#8: [**Low Value**] Extremely Low Healthy Nutrients, Extremely Low Unhealthy Nutrients
#9: [**Healthy++**] High Healthy Nutrients, Extremely Low Unhealthy Nutrients



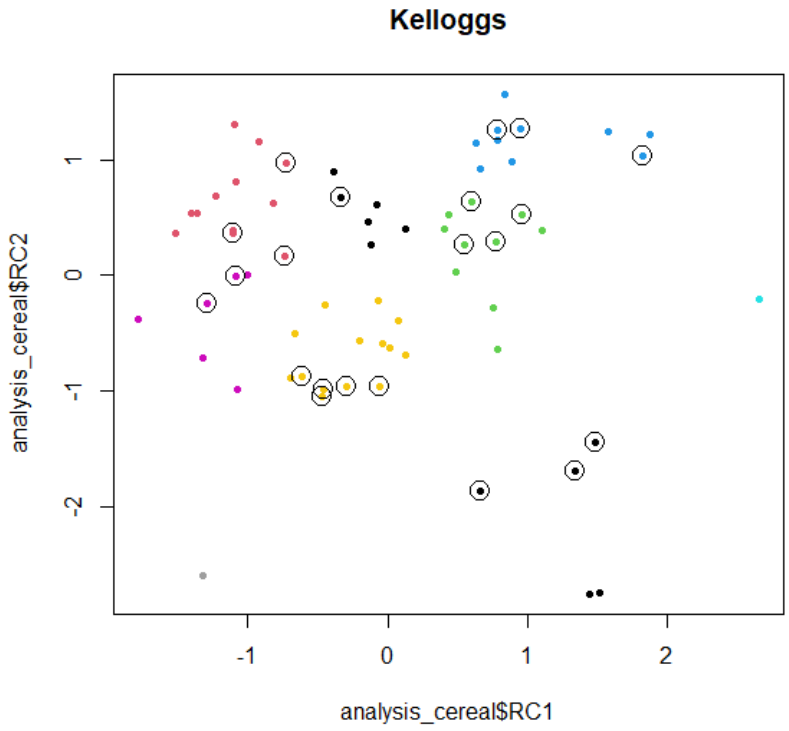
(Q1) Is there any way to discriminate among the major manufacturers by cereal characteristics, or do they each have a balanced portfolio of cereals?

These manufacturers have different portfolios of cereals, which means that they have different market strategies. We will examine these manufacturers individually.

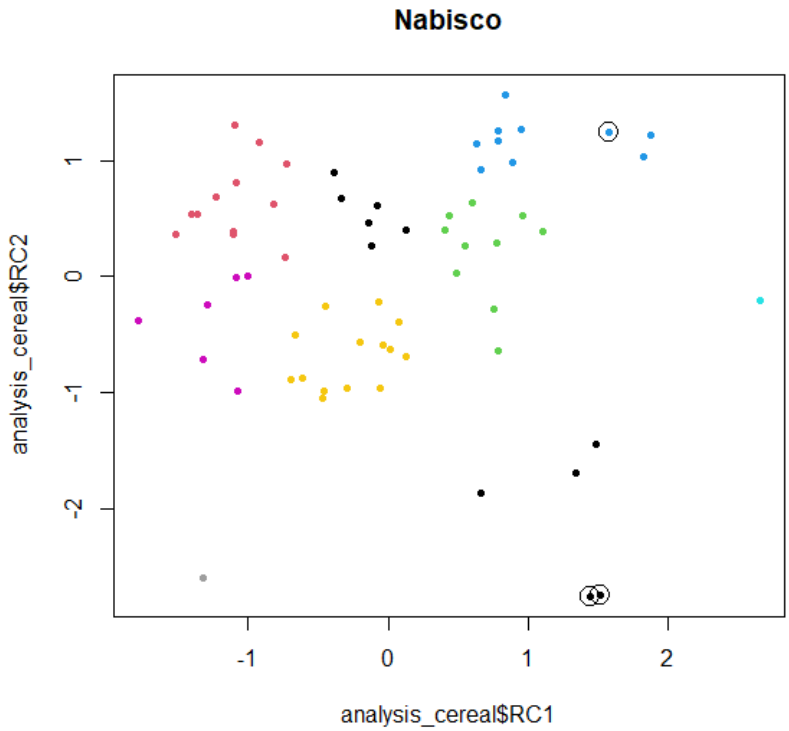
General Mills: It has a balanced portfolio of cereals. As General Mills is a major cereal manufacturer, it actually makes sense to launch various product lines to meet the public's needs.



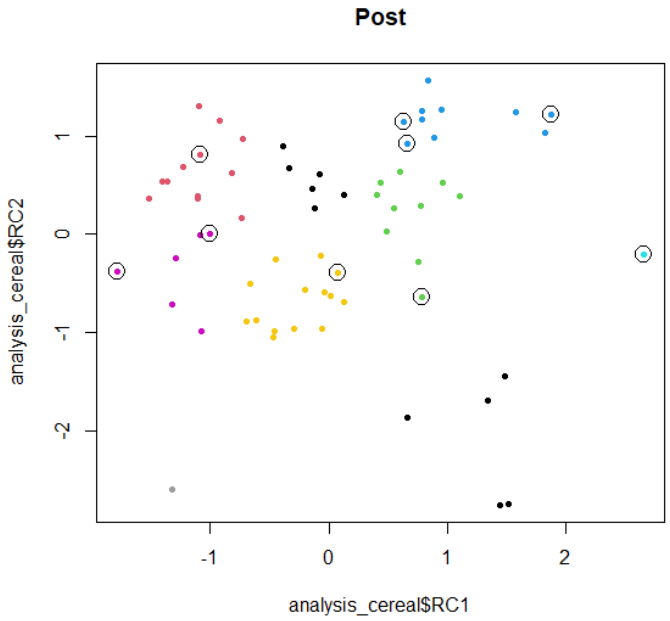
Kelloggs: It has a balanced portfolio of cereals. Compared with General Mills, Kellogg's market positioning is more distinct. For example, there are 3 products landing in extremely healthy territory. It seems like Kellogg's is trying to target those customers with a healthy lifestyle.



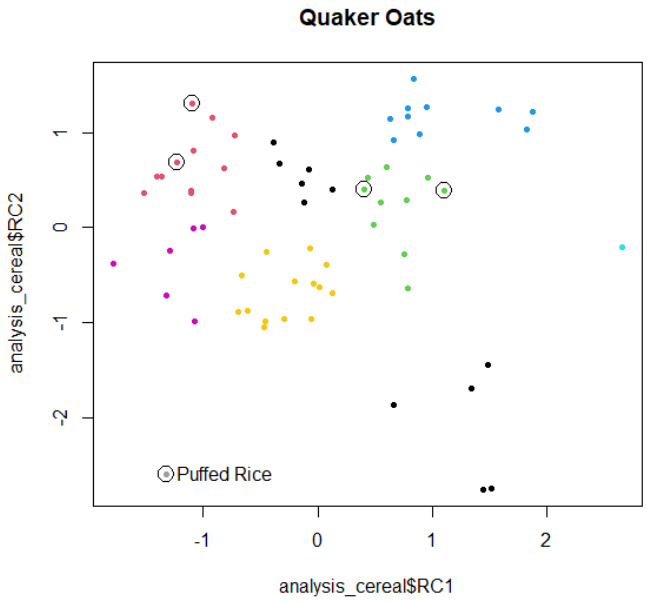
Nabisco: Nabisco, on the other hand, focuses on the niche markets. Its cereals are branded as “super healthy” with an aim to win some of those health-conscious customers.



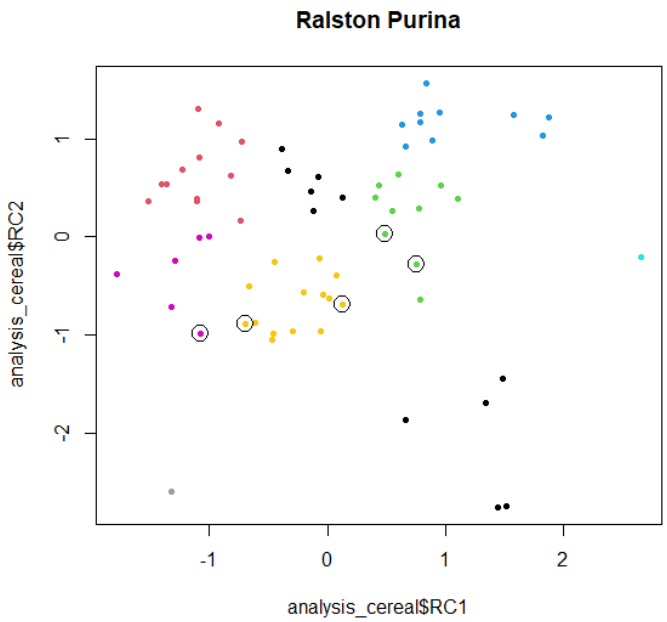
Post: Post also targets the niche markets. But unlike Nabisco only focusing on those health-conscious customers, Post focuses on various different niche markets, ranging from healthy to unhealthy.



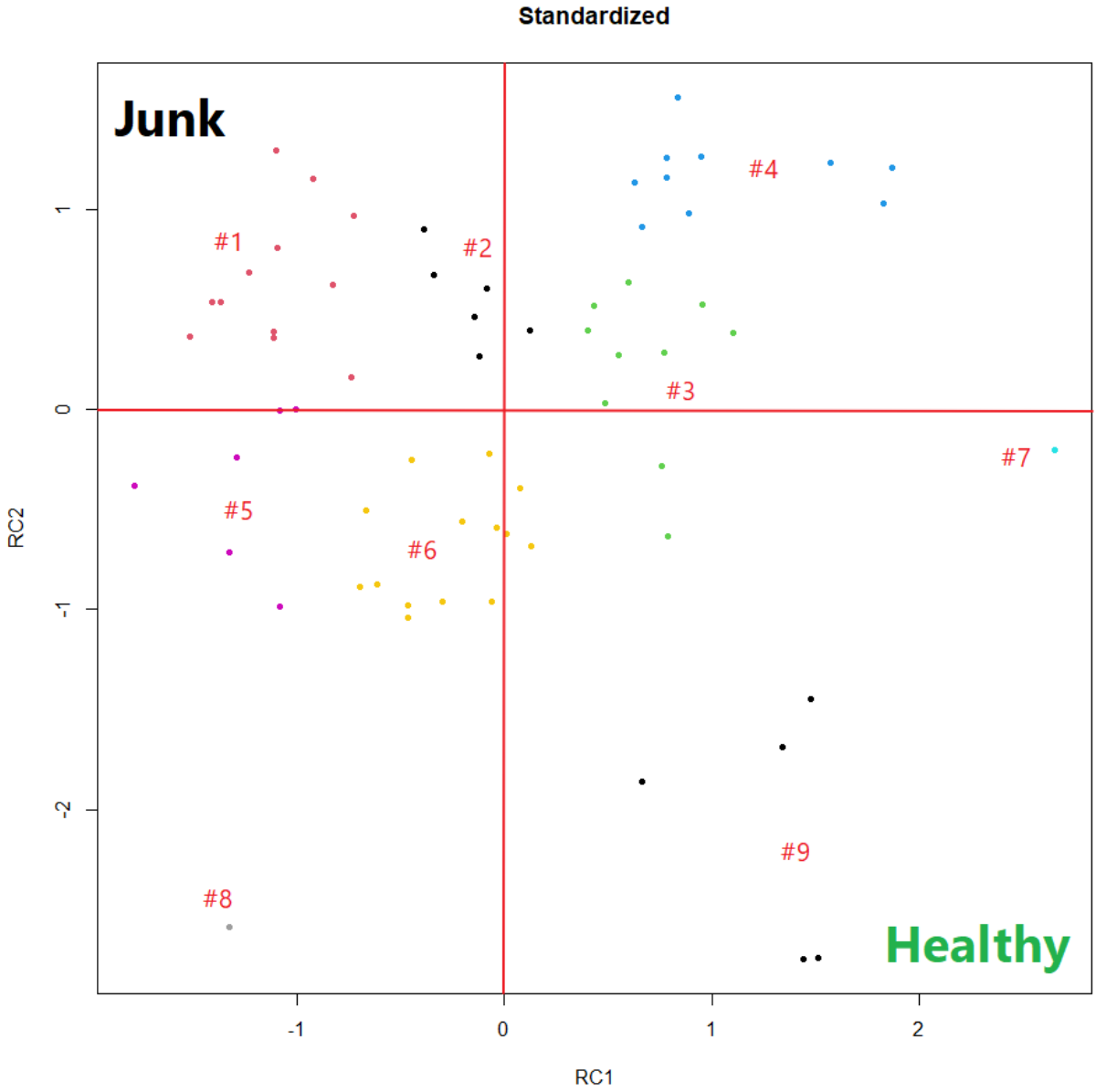
Quaker Oats: Quaker Oats' cereals mainly have two groups of products, classified as Junk++ and Mixed-Lean-Healthy in our categories. It also has one extreme product, Puffed Rice, which is low in both healthy and unhealthy nutrients.



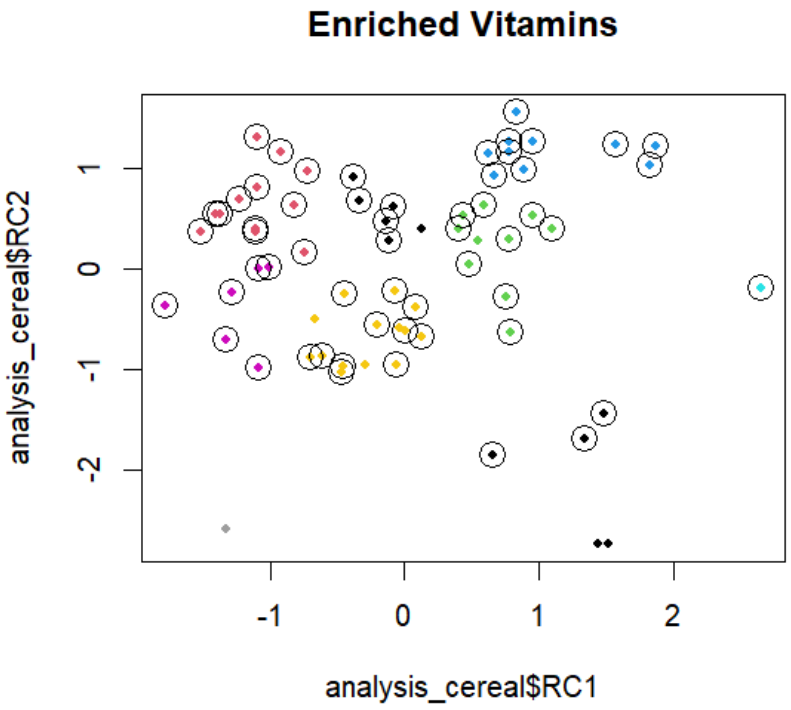
Ralston Purina: Ralston Purina's cereals are overall with little nutrients, landing in the Non Nutrient and Mixture area.

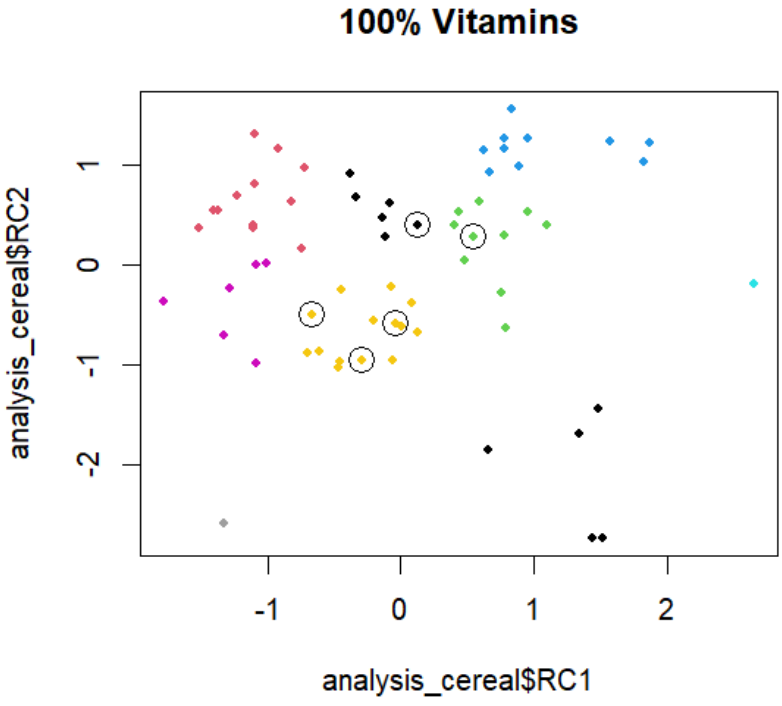


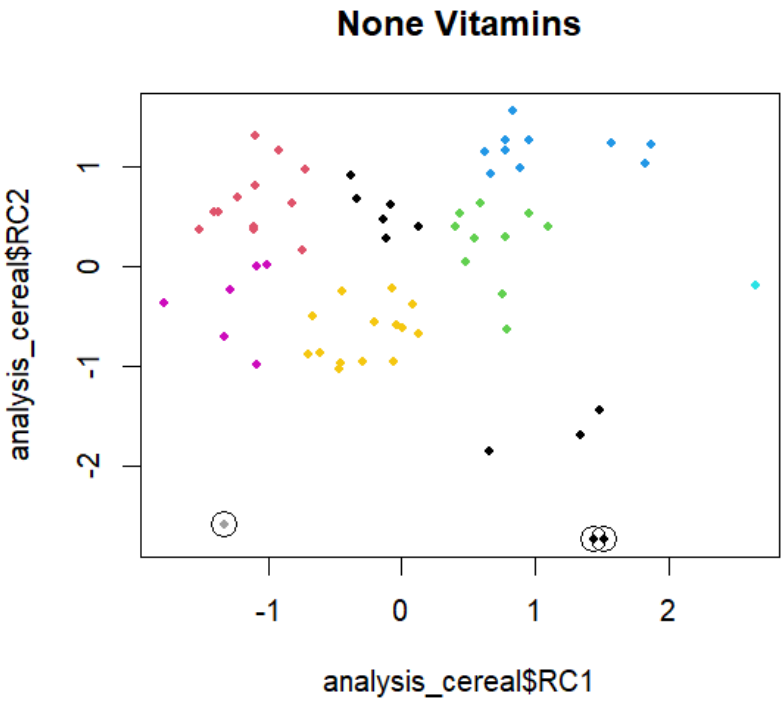
(Q2) Are there interpretable clusters of cereals?
Yes. We identify 9 clusters shown below:
#1: [**Junk++**] Low Healthy Nutrients, High Unhealthy Nutrients
#2: [**Junk**] Average Healthy Nutrients, High Unhealthy Nutrients
#3: [**Mixed-Lean-Healthy**] High Healthy Nutrients, Average Unhealthy Nutrients
#4: [**Mixed-Lean-Healthy++**] High Healthy Nutrients, High Unhealthy Nutrients
#5: [**Non Nutrient++**] Low Healthy Nutrients, Low Unhealthy Nutrients
#6: [**Non Nutrient**] Average Healthy Nutrients, Low Unhealthy Nutrients
#7: [**Healthy**] Extremely High Healthy Nutrients, Average Unhealthy Nutrients
#8: [**Low Value**] Extremely Low Healthy Nutrients, Extremely Low Unhealthy Nutrients
#9: [**Healthy++**] High Healthy Nutrients, Extremely Low Unhealthy Nutrients



We also compare our clusterings with the original vitamins labels. Only the None vitamins plot can be interpreted meaningfully. The 100% and enriched vitamins are not correlated with RC1 and RC2. None-vitamin products have lower unhealthy nutrients.


Enriched Vitamins
analysis_cereal$RC2
1
0
-1
-2
-1
0
1
2
analysis_cereal$RC1


100% Vitamins
analysis_cereal$RC2
1
0
-1
-2
-1
0
1
2
analysis_cereal$RC1

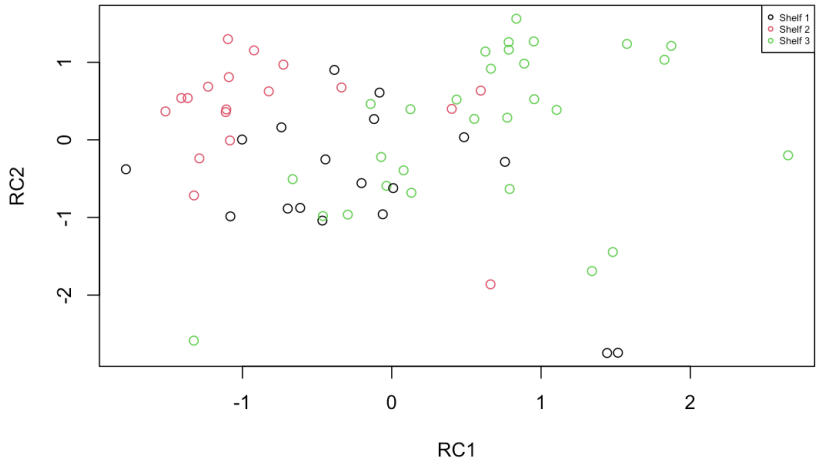


(Q3) Can you describe why cereals are displayed on high, low or middle shelves?
After fitting two factors, manufacturers, vitamins, and clusters into the logistic regression model to determine the value of shelf, we found that only the two factors can be used in the model. The final logistic regression result is shown below:

```
> glm1 <- glm(factor(shelf) ~ RC1+RC2, data = analysis_cereal, family = 'binomial')
> summary(glm1)

Call:
glm(formula = factor(shelf) ~ RC1 + RC2, family = "binomial",
    data = analysis_cereal)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9672  -0.9005   0.5405   0.7869   1.9197

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1339     0.3219   3.523 0.000427 ***
RC1           0.3926     0.3289   1.194 0.232578
RC2           0.8828     0.3247   2.719 0.006550 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 76.703  on 64  degrees of freedom
Residual deviance: 66.984  on 62  degrees of freedom
AIC: 72.984

Number of Fisher Scoring iterations: 4
```

We can see that RC2 has a small P-value, which provides strong evidence against the null hypothesis. Higher the RC2 (Unhealthy nutrients) values means highly likely this cereal will be

placed in a higher shelf level. RC1 (Healthy nutrients) does not have any effect on determining the shelf level.



If we color the cereals according to which shelf they are placed at, we can see that:

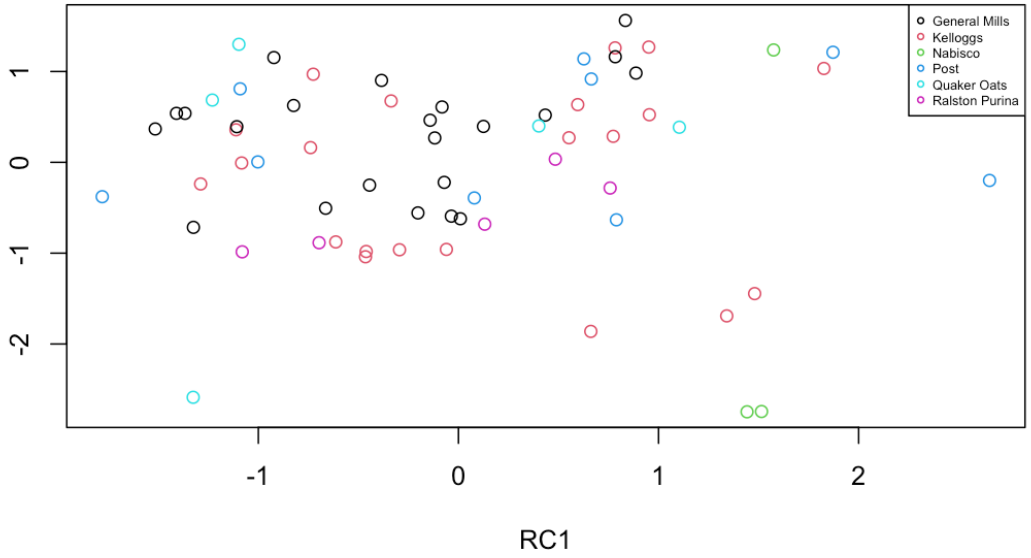
- Cereals in the lower shelf generally have average “Unhealthy Nutrients” and average “Healthy Nutrients”
- Cereals in the middle shelf generally have high “Unhealthy Nutrients” and low “Healthy Nutrients”
- Cereals in the higher shelf generally have high “Unhealthy Nutrients” and high “Healthy Nutrients”

It seems like the unhealthy cereals are the most popular ones among customers, because these cereals are placed in the middle shelf (easy to see and grab).

(Q4) How to visualize the data and show clusters or manufacturers?

We can draw the plot for both RC1 and RC2 and color these points according to the 9 clusters. Then we can point out the manufacturers individually. Check (Q1) for details.

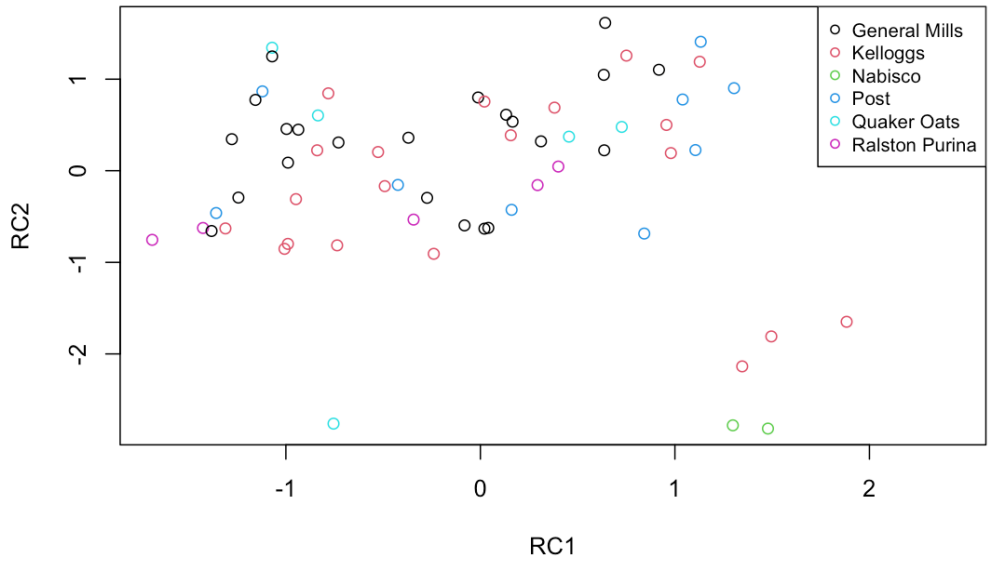
2 Factor Varimax

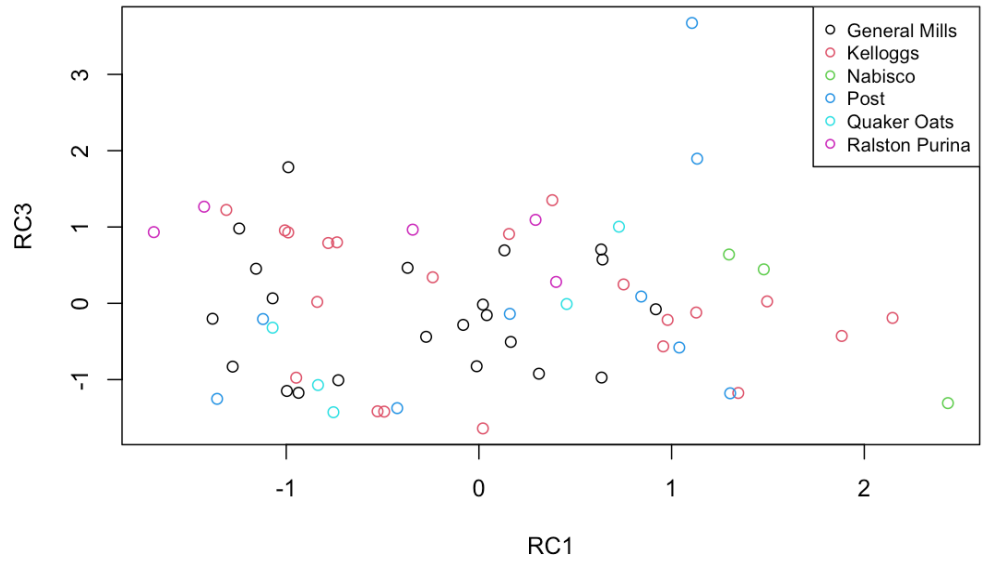


Here, we can observe the distinctions amongst manufacturers as described above, in Q3. For example, we can see General Mills and Kellogs' balanced portfolios. We see that Nabisco fills a very small niche of healthy cereals. Post offers a wide variety of cereals, ranging from healthy to unhealthy. Quaker Oats offers mainly junky cereals with one outlier, Puffed Rice, that is neither healthy or unhealthy and offers little nutritional value at all. Ralston Purina has a somewhat balanced but slightly leaning unhealthy, small portfolio of cereals.

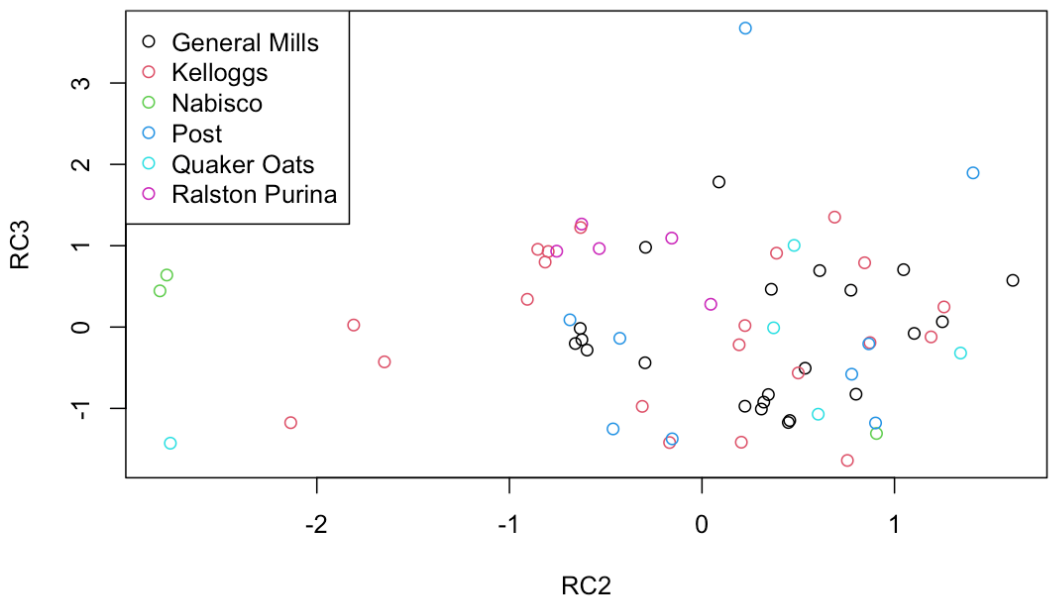
3 Factor Varimax

Investigating three factors can help gain more insight into niches of the cereal market that manufacturers might be missing out on. Below, we see the RC1 vs. RC2 plot, which is fairly in line with the same plot for only two factors. However, we can go deeper by analyzing each of RC1 and RC2 against RC3, the energy factor.





Above, we can observe that there are only two cereals that are both highly energizing and somewhat healthy, with RC3 around 2 and above and RC1 at 1. Both of these cereals are manufactured by Post, Grape Nuts (RC3 = 3.674) and Great Grains Pecan (RC3 = 1.895). This finding confirms and expands upon the finding that Post targets niche markets, specifically high value cereals. Few other manufacturers are venturing into the extremely high energy market, with most cereals having average levels of energizing nutrients, falling between 1 and -1 on RC3. One notable exception is General Mills, with its slightly unhealthy but energizing Triples cereal.



Similarly, in the above plot, we also see that Post has the two highest energy cereals. Grape Nuts (RC = 0.226) is healthier than Great Grains Pecan (RC2 = 1.408), as indicated by its average RC2 score and higher RC1 score. Therefore, Grape Nuts is a high energy, healthy cereal. Kelloggs, General Mills, and Quaker Oats have very average portfolios in terms of energy. Ralston Purina seems to have a portfolio consisting of high energy cereals with a range of healthiness.

In conclusion, investigating three factors in tandem with our main, two factor analysis reveals some untapped market potential. Aside from the Post cereals, high energy and healthy cereals is a customer segment that is underserved and presents a potential opportunity for manufacturers. Especially given the rise of healthy eating for fitness goals, a cereal targeted at athletes seems like a great option in an industry crowded by unhealthy, low value cereals.