Principal Component Analysis

```
library(ggplot2)
library(ggbiplot)
```

```
df <- read.table("cereals.txt", header=T )
```

```
head(df)
```

```
##                         name mfr type calories protein fat sodium fiber carbo
## 1                  100%_Bran   N    C       70       4   1    130  10.0   5.0
## 2          100%_Natural_Bran   Q    C      120       3   5     15   2.0   8.0
## 3                   All-Bran   K    C       70       4   1    260   9.0   7.0
## 4 All-Bran_with_Extra_Fiber   K    C       50       4   0    140  14.0   8.0
## 5             Almond_Delight   R    C      110       2   2    200   1.0  14.0
## 6   Apple_Cinnamon_Cheerios   G    C      110       2   2    180   1.5  10.5
##   sugars potass vitamins shelf weight cups   rating
## 1      6    280       25     3      1 0.33 68.40297
## 2      8    135        0     3      1 1.00 33.98368
## 3      5    320       25     3      1 0.33 59.42551
## 4      0    330       25     3      1 0.50 93.70491
## 5      8     -1       25     3      1 0.75 34.38484
## 6     10     70       25     1      1 0.75 29.50954
```

Let's take a look at the dataset

```
summary(df)
```

```
##      name               mfr                type              calories
##  Length:77          Length:77          Length:77          Min.   : 50.0
##  Class :character   Class :character   Class :character   1st Qu.:100.0
##  Mode  :character   Mode  :character   Mode  :character   Median :110.0
##                                                           Mean   :106.9
##                                                           3rd Qu.:110.0
##                                                           Max.   :160.0
##     protein           fat            sodium          fiber
##  Min.   :1.000   Min.   :0.000   Min.   :  0.0   Min.   : 0.000
##  1st Qu.:2.000   1st Qu.:0.000   1st Qu.:130.0   1st Qu.: 1.000
##  Median :3.000   Median :1.000   Median :180.0   Median : 2.000
##  Mean   :2.545   Mean   :1.013   Mean   :159.7   Mean   : 2.152
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:210.0   3rd Qu.: 3.000
##  Max.   :6.000   Max.   :5.000   Max.   :320.0   Max.   :14.000
##      carbo          sugars          potass          vitamins
##  Min.   :-1.0   Min.   :-1.000   Min.   : -1.00   Min.   :  0.00
```

```
##  1st Qu.:12.0   1st Qu.: 3.000   1st Qu.: 40.00   1st Qu.: 25.00
##  Median :14.0   Median : 7.000   Median : 90.00   Median : 25.00
##  Mean   :14.6   Mean   : 6.922   Mean   : 96.08   Mean   : 28.25
##  3rd Qu.:17.0   3rd Qu.:11.000   3rd Qu.:120.00   3rd Qu.: 25.00
##  Max.   :23.0   Max.   :15.000   Max.   :330.00   Max.   :100.00
##      shelf           weight          cups           rating
##  Min.   :1.000   Min.   :0.50   Min.   :0.250   Min.   :18.04
##  1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.670   1st Qu.:33.17
##  Median :2.000   Median :1.00   Median :0.750   Median :40.40
##  Mean   :2.208   Mean   :1.03   Mean   :0.821   Mean   :42.67
##  3rd Qu.:3.000   3rd Qu.:1.00   3rd Qu.:1.000   3rd Qu.:50.83
##  Max.   :3.000   Max.   :1.50   Max.   :1.500   Max.   :93.70
```

- We have no missing values.
- Because PCA works best with numerical data, I'll exclude all the categorical columns for now. the categorical columns are : name,mfr,type

```
no_cat_df = df[,-c(1:3)]

summary(no_cat_df)
```

```
##     calories        protein           fat            sodium
##  Min.   : 50.0   Min.   :1.000   Min.   :0.000   Min.   :  0.0
##  1st Qu.:100.0   1st Qu.:2.000   1st Qu.:0.000   1st Qu.:130.0
##  Median :110.0   Median :3.000   Median :1.000   Median :180.0
##  Mean   :106.9   Mean   :2.545   Mean   :1.013   Mean   :159.7
##  3rd Qu.:110.0   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:210.0
##  Max.   :160.0   Max.   :6.000   Max.   :5.000   Max.   :320.0
##      fiber            carbo          sugars          potass
##  Min.   : 0.000   Min.   :-1.0   Min.   :-1.000   Min.   : -1.00
##  1st Qu.: 1.000   1st Qu.:12.0   1st Qu.: 3.000   1st Qu.: 40.00
##  Median : 2.000   Median :14.0   Median : 7.000   Median : 90.00
##  Mean   : 2.152   Mean   :14.6   Mean   : 6.922   Mean   : 96.08
##  3rd Qu.: 3.000   3rd Qu.:17.0   3rd Qu.:11.000   3rd Qu.:120.00
##  Max.   :14.000   Max.   :23.0   Max.   :15.000   Max.   :330.00
##     vitamins          shelf           weight          cups
##  Min.   :  0.00   Min.   :1.000   Min.   :0.50   Min.   :0.250
##  1st Qu.: 25.00   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:0.670
##  Median : 25.00   Median :2.000   Median :1.00   Median :0.750
##  Mean   : 28.25   Mean   :2.208   Mean   :1.03   Mean   :0.821
##  3rd Qu.: 25.00   3rd Qu.:3.000   3rd Qu.:1.00   3rd Qu.:1.000
##  Max.   :100.00   Max.   :3.000   Max.   :1.50   Max.   :1.500
##      rating
##  Min.   :18.04
##  1st Qu.:33.17
##  Median :40.40
##  Mean   :42.67
##  3rd Qu.:50.83
##  Max.   :93.70
```

PCA

```
df.pca <- prcomp(no_cat_df, center = TRUE,)

summary(df.pca)
```

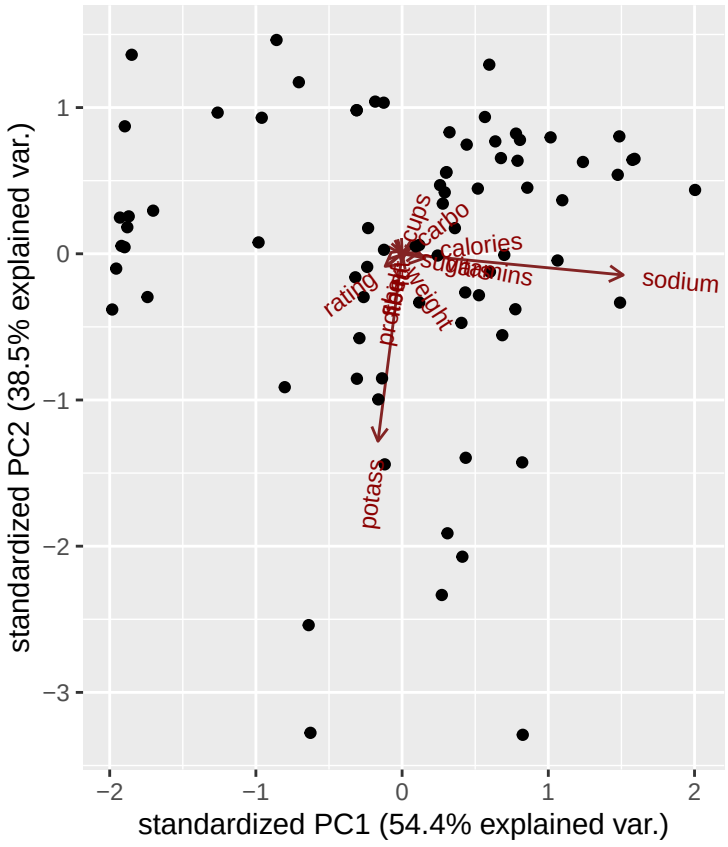
```
## Importance of components:
##                             PC1     PC2      PC3     PC4     PC5     PC6     PC7
## Standard deviation     84.8289 71.3721 22.37869 18.8655 8.62931 2.37580 2.08502
## Proportion of Variance  0.5438  0.3850  0.03785  0.0269 0.00563 0.00043 0.00033
## Cumulative Proportion   0.5438  0.9288  0.96661  0.9935 0.99914 0.99956 0.99989
##                            PC8     PC9    PC10   PC11    PC12     PC13
## Standard deviation     0.80551 0.69493 0.53223 0.1844 0.06684 5.26e-08
## Proportion of Variance 0.00005 0.00004 0.00002 0.0000 0.00000 0.00e+00
## Cumulative Proportion  0.99994 0.99998 1.00000 1.0000 1.00000 1.00e+00
```

```
df.pca$rotation
```

```
##                    PC1           PC2           PC3           PC4           PC5
## calories  0.0761921453  0.0100060433 -0.6119143204 -0.612823065  0.464049505
## protein  -0.0013978750 -0.0083327685  0.0003558018  0.001782938  0.055718533
## fat      -0.0001175350 -0.0025953318 -0.0157634654 -0.025981796 -0.016691341
## sodium    0.9830207900 -0.1121776549  0.1418722668 -0.003522443  0.015170455
## fiber    -0.0046056965 -0.0298561770  0.0193455258  0.020272965  0.017484643
## carbo     0.0196167313  0.0179111948 -0.0153790645  0.031108147  0.348460854
## sugars    0.0062502419 -0.0015278082 -0.1004208194 -0.112461025 -0.287017306
## potass   -0.1072760936 -0.9905116321 -0.0282744796 -0.043969463 -0.041434486
## vitamins  0.1011060939 -0.0221968424 -0.7026451242  0.702849413 -0.024219704
## shelf    -0.0008931428 -0.0040856560 -0.0122027951  0.005839518 -0.004883186
## weight    0.0005089675 -0.0009469476 -0.0036129977 -0.002659340  0.003546237
## cups      0.0004630786  0.0015556523 -0.0007262406  0.001027953  0.002259186
## rating   -0.0754052237 -0.0663275447  0.3159387583  0.337280622  0.758017288
##                    PC6           PC7           PC8           PC9           PC10
## calories  0.131449665  0.082059479 -0.011293097  0.0297126851  0.0436328489
## protein   0.228648039  0.041511354  0.467877635 -0.5326734198 -0.2431158187
## fat       0.174345223 -0.181299406 -0.309421436 -0.1981802917 -0.8351295525
## sodium    0.014092225  0.019892953 -0.000122729 -0.0039424456  0.0008386517
## fiber     0.090922096  0.259026544 -0.577270945  0.3768249749 -0.1618103786
## carbo    -0.834993279 -0.323977480 -0.022460749 -0.0505757062 -0.1692547786
## sugars   -0.419688201  0.796458048  0.052303354 -0.1808324967 -0.1738497577
## potass   -0.028964116 -0.042076871  0.015654281 -0.0007188526  0.0056799769
## vitamins  0.024394871  0.012348939  0.006334232  0.0082137949 -0.0062742306
## shelf    -0.008075171 -0.042104631 -0.585837090 -0.7043890473  0.3908283529
## weight   -0.006793883  0.013271236  0.006429771  0.0202563639  0.0343363977
## cups     -0.014562392 -0.007665034  0.071899033  0.0211719970 -0.0491759189
## rating    0.128335360  0.384637008  0.006195629 -0.0417057658 -0.0089215916
##                   PC11          PC12          PC13
## calories  0.0045681129 -0.0074150943 -4.193592e-02
## protein  -0.0321443663  0.0145062508  6.162939e-01
## fat      -0.0164228495  0.0370773738 -3.184690e-01
## sodium    0.0003046364 -0.0002509910 -1.026023e-02
```

```
## fiber     0.0290055364  0.0010964656  6.483602e-01
## carbo    -0.0215171793  0.0027536872  2.056936e-01
## sugars   -0.0075113676 -0.0043715898 -1.364878e-01
## potass    0.0002021444 -0.0008627762 -6.400483e-03
## vitamins -0.0016297703 -0.0007749528 -9.642514e-03
## shelf     0.0755062446  0.0109830422 -7.005538e-09
## weight   -0.0772785997  0.9960628633 -8.093091e-08
## cups      0.9927878610  0.0778211349  2.596195e-08
## rating    0.0034248384 -0.0035448206 -1.882863e-01
```

```
ggbiplot(df.pca)
```



Running PCA without scaling we can see the first principal component is dominated by the sodium content in the cereal. On the other hand, the second principal component seems to be measuring the amount of potassium. However, the problem with this output is that sodium and potassium are measured in milligrams while other nutrients are in grams.This gives the two nutrients a much larger variance than the other variables. The solution would be to normalize our dataset before running PCA, that way we give all variable equal importance in terms of variability.

PCA on scaled data

```
scaled_df.pca <- prcomp(no_cat_df, scale. = TRUE)

summary(scaled_df.pca)
```

```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     1.9000 1.7716 1.3479 1.03068 0.98540 0.84803 0.81930
## Proportion of Variance 0.2777 0.2414 0.1398 0.08171 0.07469 0.05532 0.05163
## Cumulative Proportion  0.2777 0.5191 0.6589 0.74057 0.81526 0.87058 0.92221
##                           PC8     PC9    PC10    PC11    PC12      PC13
## Standard deviation     0.6707 0.55076 0.36151 0.27035 0.23294 1.473e-08
## Proportion of Variance 0.0346 0.02333 0.01005 0.00562 0.00417 0.000e+00
## Cumulative Proportion  0.9568 0.98015 0.99020 0.99583 1.00000 1.000e+00
```

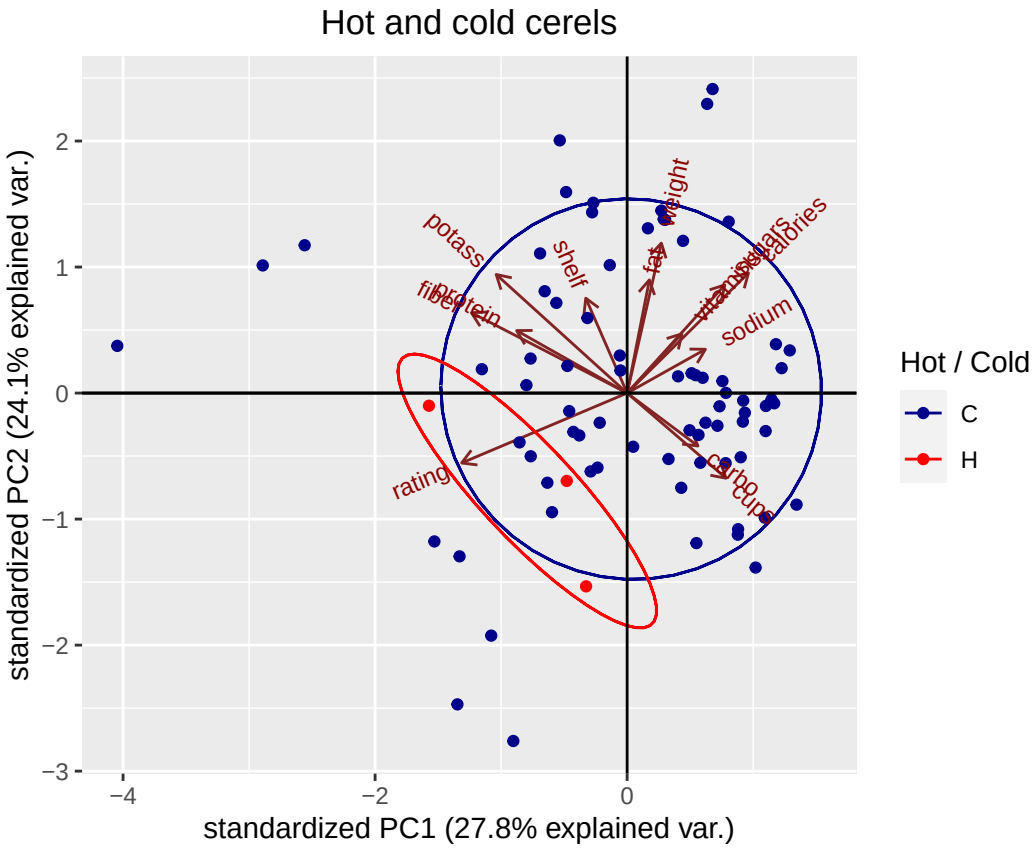
```
scaled_df.pca$rotation
```

```
##                  PC1        PC2         PC3          PC4         PC5
## calories  0.33325770  0.3540139 -0.07308253  0.310355769  0.10011713
## protein  -0.30350996  0.1840913 -0.21402139  0.487982044  0.18592601
## fat       0.06117068  0.3334252  0.27702228  0.364438411  0.44961181
## sodium    0.21498742  0.1289957 -0.39268086 -0.023246751 -0.30711733
## fiber    -0.42768885  0.2366169 -0.10160606 -0.093597811 -0.24394397
## carbo     0.19454080 -0.1558489 -0.55656859  0.111598513  0.12697220
## sugars    0.26823424  0.3185932  0.31783521 -0.203435176 -0.29985232
## potass   -0.35969408  0.3494167 -0.08431129 -0.007793219 -0.18777422
## vitamins  0.15038361  0.1745444 -0.40671499 -0.438602292  0.22243824
## shelf    -0.11360533  0.2795025 -0.04010916 -0.483363806  0.57904983
## weight    0.09400496  0.4417863 -0.24306975  0.119565685 -0.25454786
## cups      0.27161796 -0.2508459 -0.14300053  0.160753010  0.08103983
## rating   -0.45456749 -0.2059514 -0.20769280  0.058186320  0.04467135
##                  PC6         PC7         PC8         PC9         PC10
## calories -0.26460458  0.01498978 -0.02281052  0.01334435 -0.19957001
## protein   0.15170567 -0.22719127 -0.32251078  0.54173171 -0.15116767
## fat       0.31169592  0.12818669  0.14648175 -0.45797475  0.01304935
## sodium    0.62270867  0.38210484  0.17939260  0.23312260  0.04240406
## fiber     0.07240825 -0.08258368  0.24298629 -0.22599391 -0.09908523
## carbo    -0.35737318  0.27484150  0.19065854 -0.16580913 -0.41780268
## sugars   -0.14064192 -0.22842387  0.04051536  0.22755181 -0.47283223
## potass    0.04174594 -0.13336127  0.28351947 -0.18083652 -0.24633333
## vitamins  0.22240340 -0.38321224 -0.47879609 -0.32410564 -0.07167993
## shelf    -0.12424005  0.13697250  0.35305822  0.40528584  0.11986130
## weight   -0.38496596 -0.05973177 -0.02919860 -0.04510871  0.65471082
## cups      0.13621347 -0.67930809  0.55537769  0.05698475  0.13237275
## rating   -0.18951565 -0.05255890  0.03489036 -0.06805101 -0.03241640
##                 PC11         PC12          PC13
## calories  0.28043868  0.640419468 -2.288634e-01
## protein  -0.03161281 -0.154300238  1.889853e-01
## fat       0.05426156 -0.342665063 -8.977980e-02
## sodium    0.08437340 -0.028669463 -2.409229e-01
## fiber     0.59848142  0.070600095  4.328289e-01
## carbo    -0.09599574 -0.296433152  2.465290e-01
## sugars    0.11830639 -0.449901336 -1.699277e-01
## potass   -0.66867383  0.222146469 -1.278002e-01
## vitamins  0.01591208 -0.003684911 -6.034371e-02
## shelf     0.02296900  0.020642042 -1.633608e-09
## weight   -0.06087022 -0.266099498 -3.411092e-09
## cups      0.03401393  0.007559801  1.692284e-09
## rating    0.27592100 -0.178577824 -7.408338e-01
```

- The first two components account for 52% of Proportion of Variance. Reducing the number of components to just two would mean losing a lot of information.

PCA Plot

```
ggbiplot(scaled_df.pca ,ellipse = TRUE, groups = df$type)+labs(title='Hot and cold cerels',color="Hot /
scale_color_manual(values=c("blue4", "red"))+ theme(plot.title = element_text(hjust = 0.5))+ geom_hline
```



intepreting the graph

- we have very few outliers
- fat and weight have large positive loadings on component 2.
- sodium has the largest positive loading on component 1
- Some strong correlations are:
  - vitamins, sugars and calories
  - fat and weight -carbo and cups
  - fiber and protein
- Hot cereals are clustered

PLot with "groups" as the classifier :

```
ggbiplot(scaled_df.pca ,ellipse = TRUE, groups = df$mfr)
```