STA6714 Assignment Homework – Chapter 4

Mithun Mohanraj

Problem 4.1

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

a. **Which variables are quantitative/numerical? Which are ordinal? Which are nominal?**

The following variables are numerical/quantitative:

Calories,protein,fat,sodium,fiber,carbo,sugars,potass,vitamins,weight,cups,rating.

The following is ordinal:

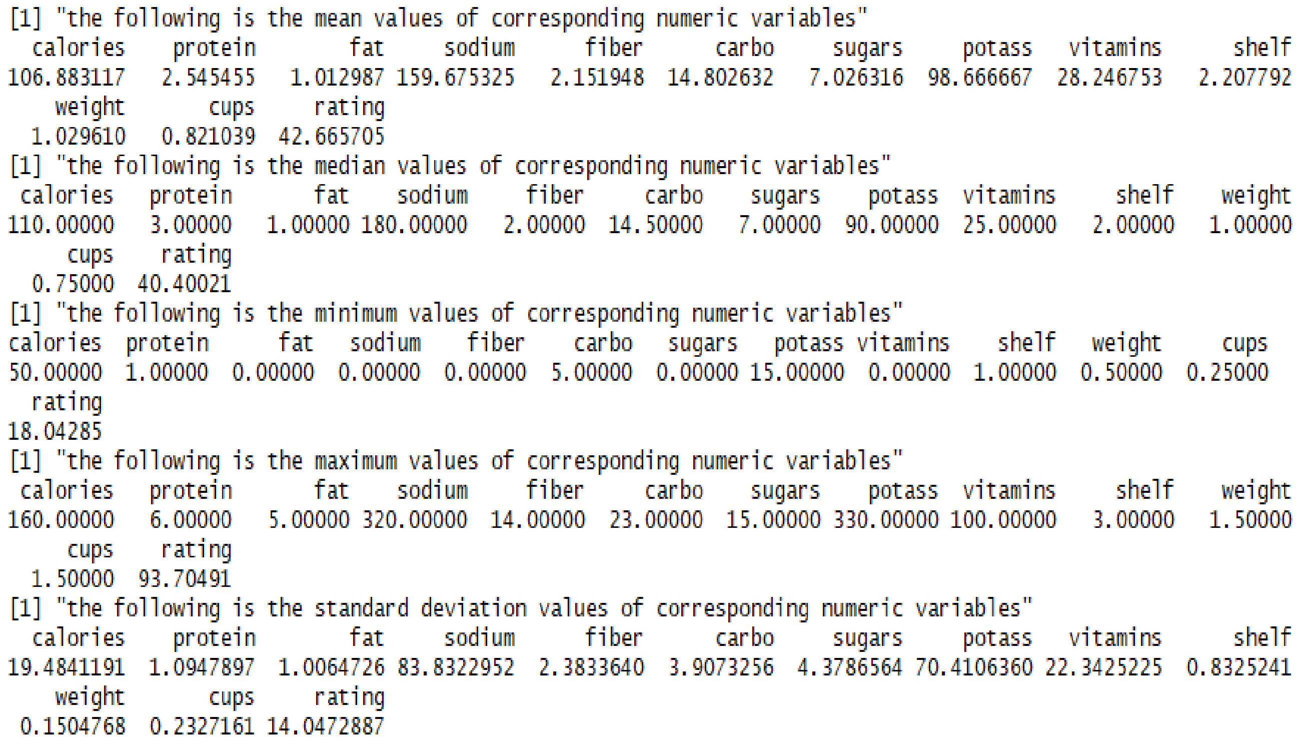
Shelf

The following are nominal:

Mfr ,type

b. Compute the mean, median, min, max, and standard deviation for each of the quantitative variables. This can be done through R's sapply() function (e.g., sapply(data, mean, na.rm = TRUE)).

The following image shows the result obtained by using sapply function:



```
[1] "the following is the mean values of corresponding numeric variables"
  calories    protein        fat     sodium      fiber      carbo     sugars     potass   vitamins      shelf 
106.883117   2.545455   1.012987 159.675325   2.151948  14.802632   7.026316  98.666667  28.246753   2.207792 
    weight       cups     rating 
  1.029610   0.821039  42.665705 
[1] "the following is the median values of corresponding numeric variables"
 calories   protein       fat    sodium     fiber     carbo    sugars    potass  vitamins     shelf    weight 
110.00000   3.00000   1.00000 180.00000   2.00000  14.50000   7.00000  90.00000  25.00000   2.00000   1.00000 
     cups    rating 
  0.75000  40.40021 
[1] "the following is the minimum values of corresponding numeric variables"
calories  protein      fat   sodium    fiber    carbo   sugars   potass vitamins    shelf   weight     cups 
50.00000  1.00000  0.00000  0.00000  0.00000  5.00000  0.00000 15.00000  0.00000  1.00000  0.50000  0.25000 
  rating 
18.04285 
[1] "the following is the maximum values of corresponding numeric variables"
 calories   protein       fat    sodium     fiber     carbo    sugars    potass  vitamins     shelf    weight 
160.00000   6.00000   5.00000 320.00000  14.00000  23.00000  15.00000 330.00000 100.00000   3.00000   1.50000 
     cups    rating 
  1.50000  93.70491 
[1] "the following is the standard deviation values of corresponding numeric variables"
  calories    protein        fat     sodium      fiber      carbo     sugars     potass   vitamins      shelf 
19.4841191  1.0947897  1.0064726 83.8322952  2.3833640  3.9073256  4.3786564 70.4106360 22.3425225  0.8325241 
    weight       cups     rating 
 0.1504768  0.2327161 14.0472887 
```

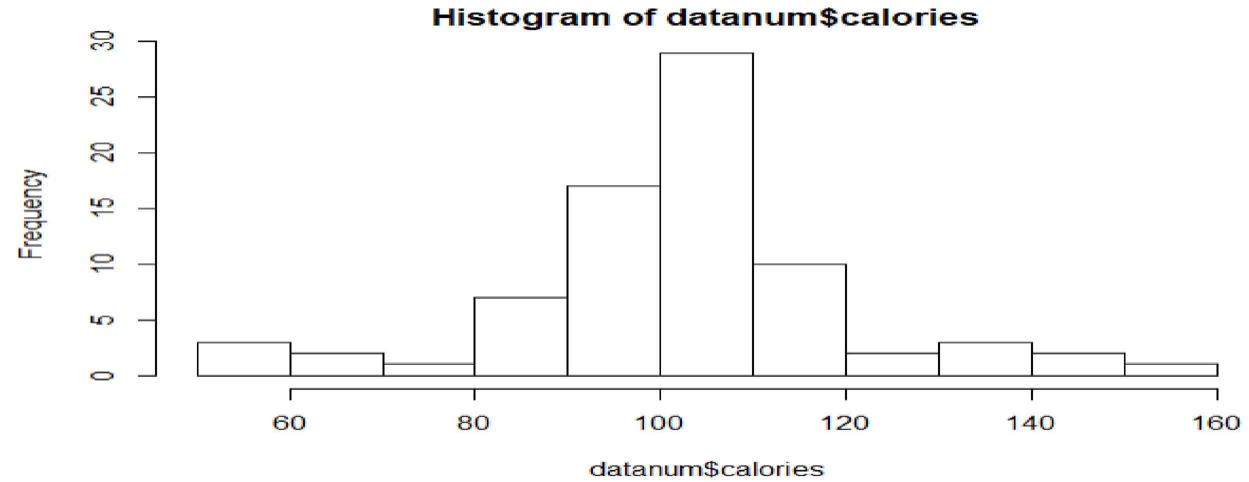
C,Use R to plot a histogram for each of the quantitative variables. Based on the histograms and summary statistics, answer the following questions:

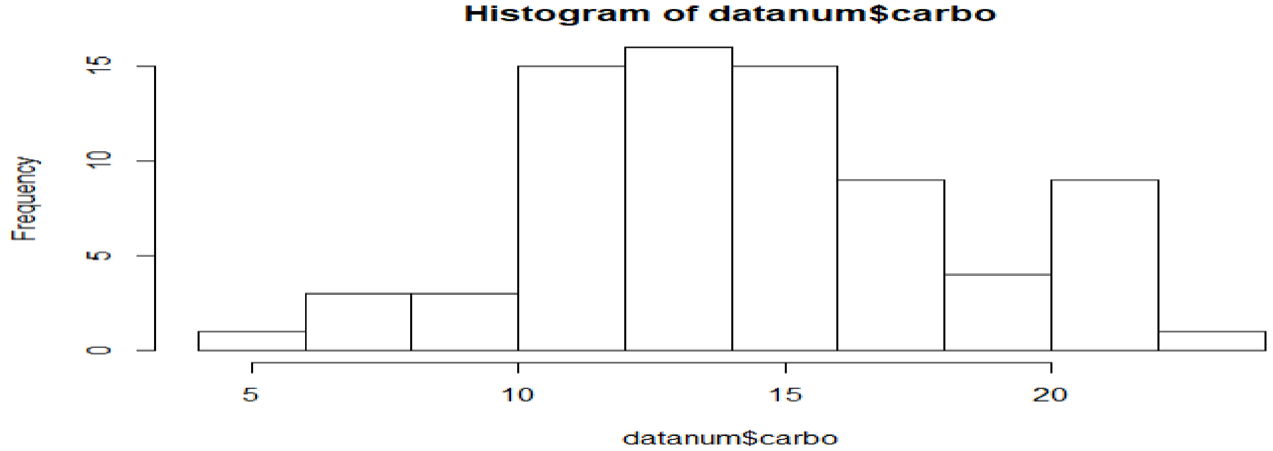
i. Which variables have the largest variability?

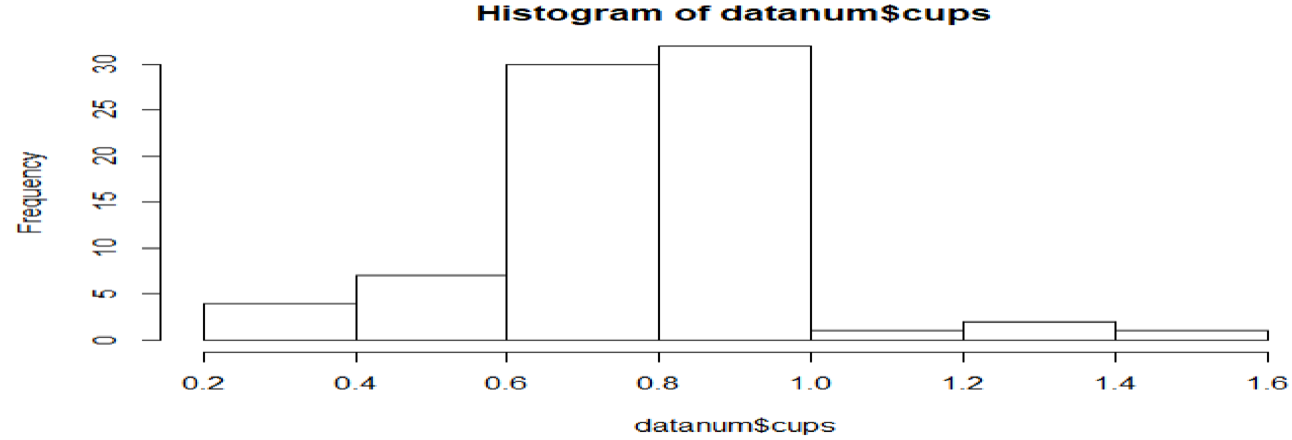
ii. Which variables seem skewed?

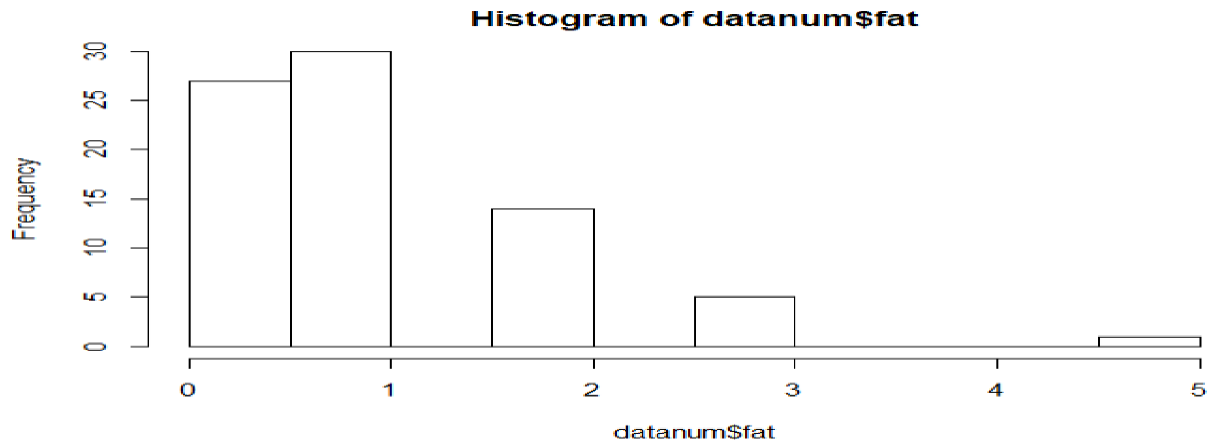
iii. Are there any values that seem extreme

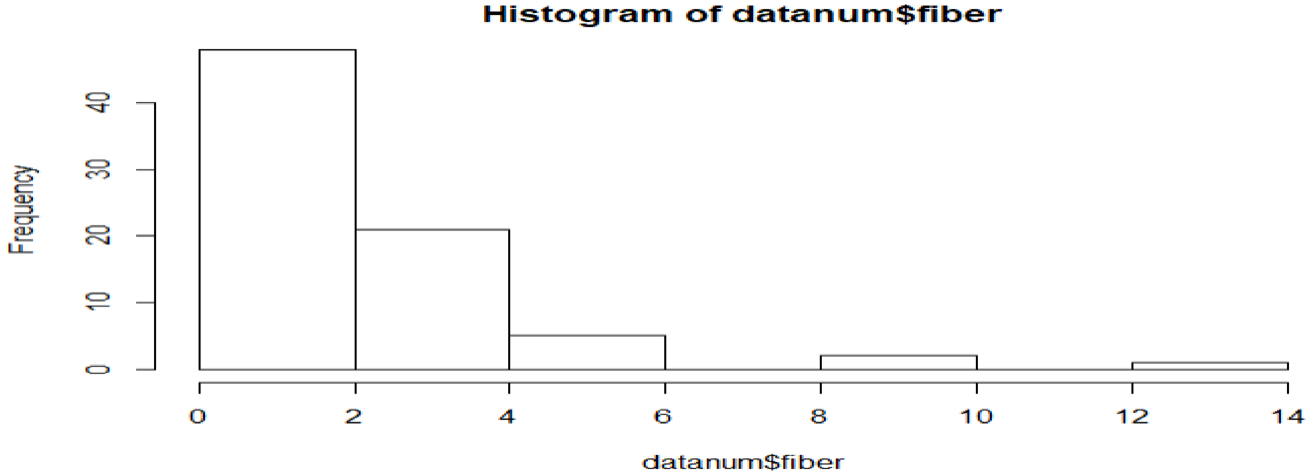
The following are the histogram plots for each of the quantitative variables:

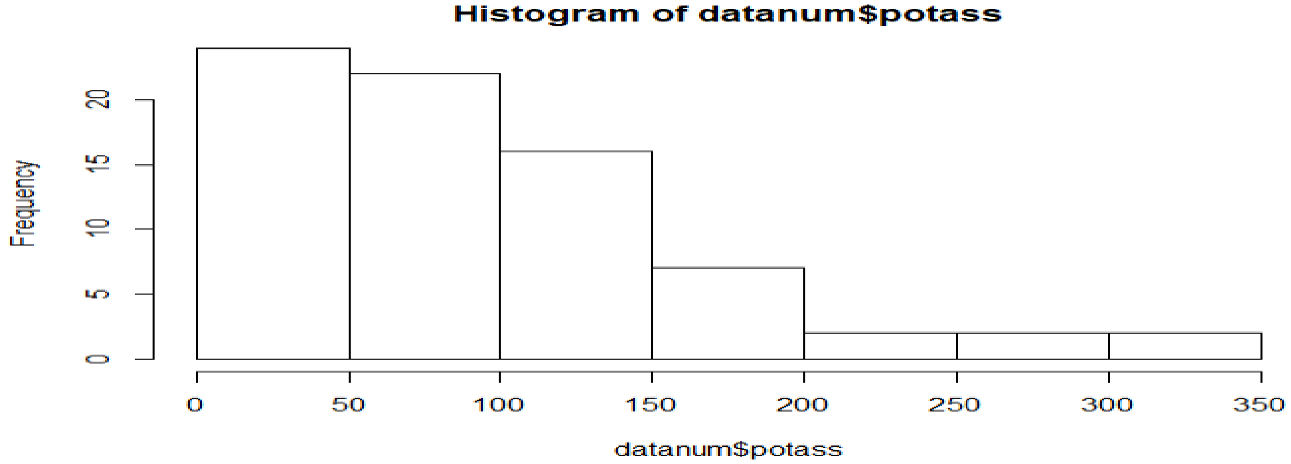


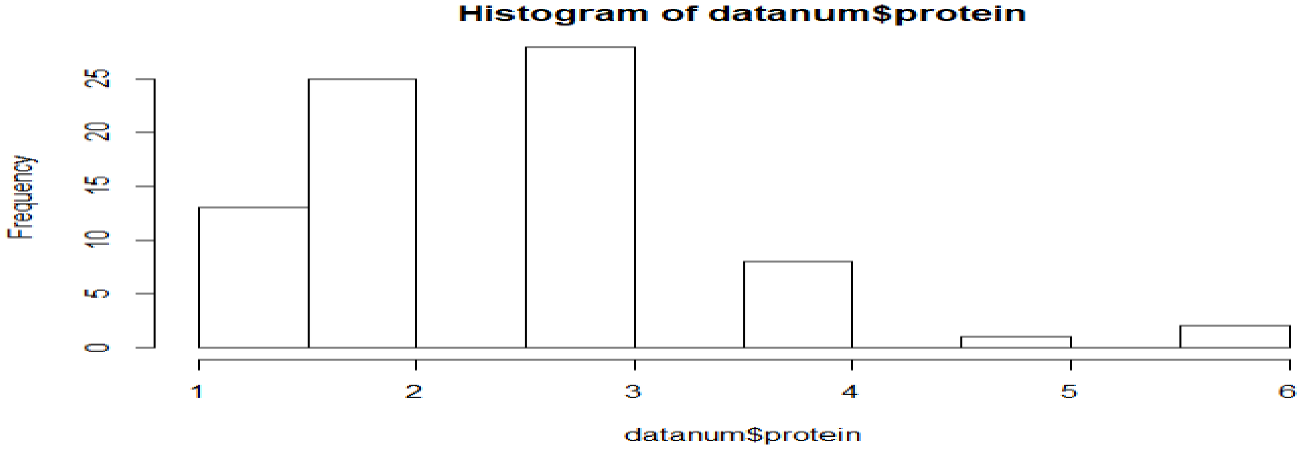


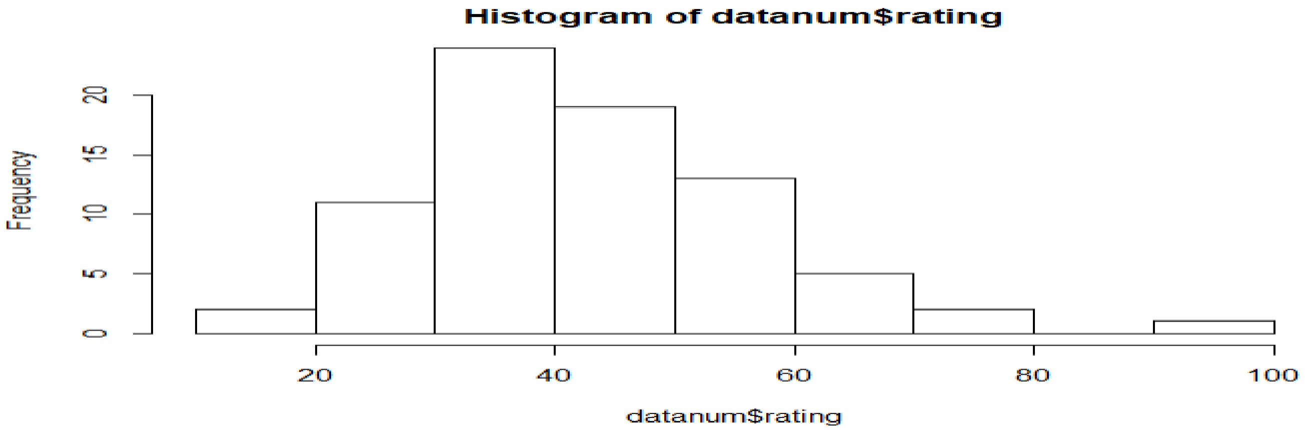


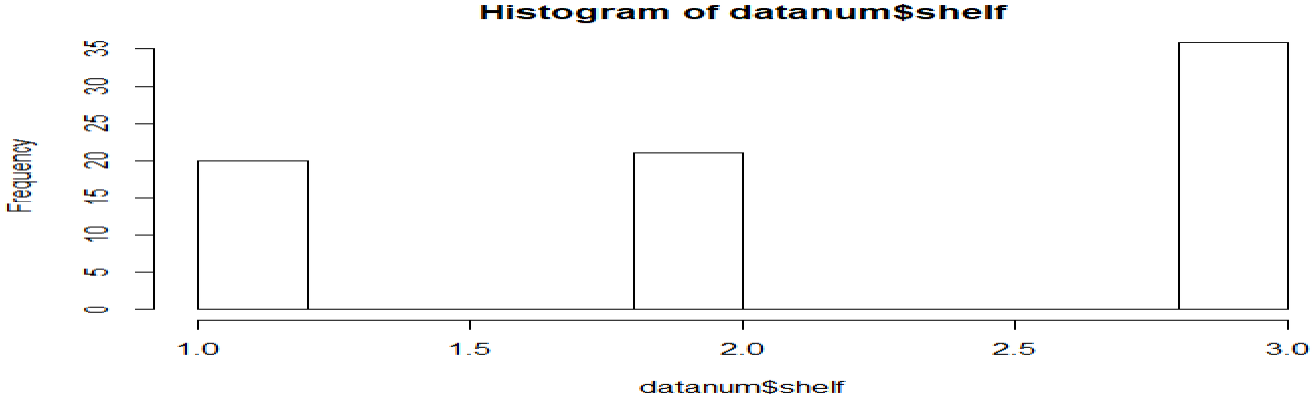

Histogram of datanum$fat
Frequency
0
5
10
15
20
25
30
0
1
2
3
4
5
datanum$fat

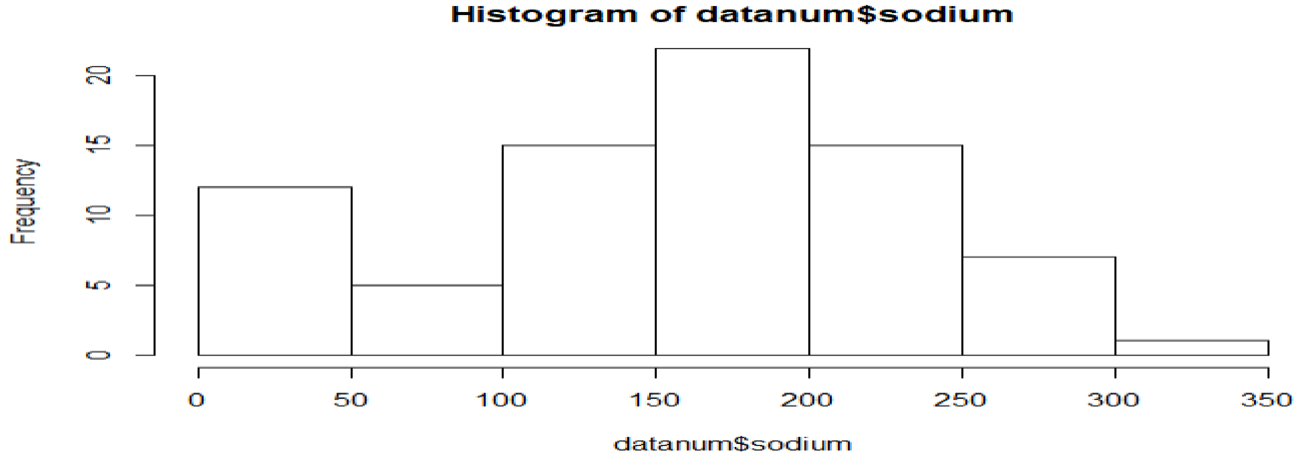

Histogram of datanum$fiber
Frequency
0
10
20
30
40
0
2
4
6
8
10
12
14
datanum$fiber

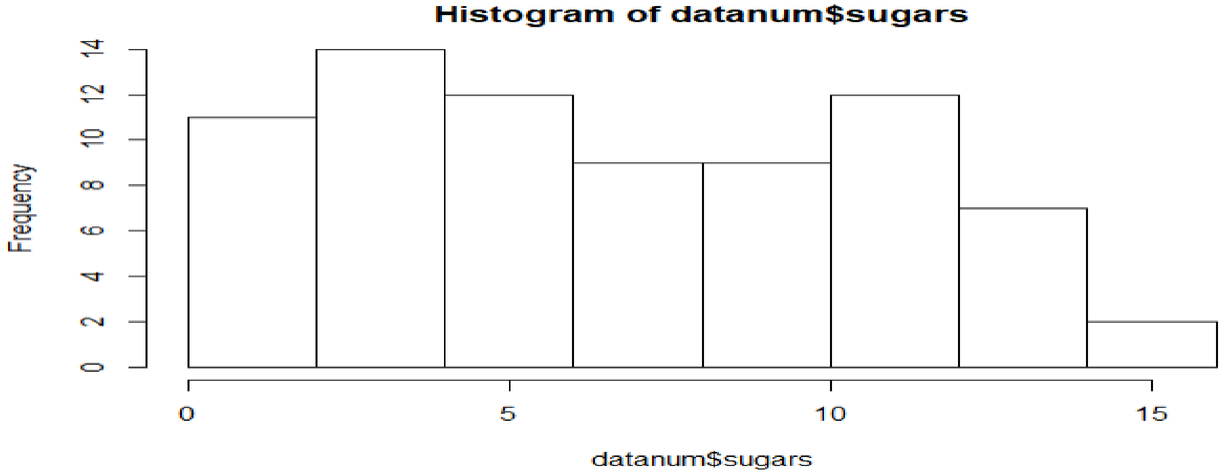

Histogram of datanum$potass
Frequency
0
5
10
15
20
0
50
100
150
200
250
300
350
datanum$potass

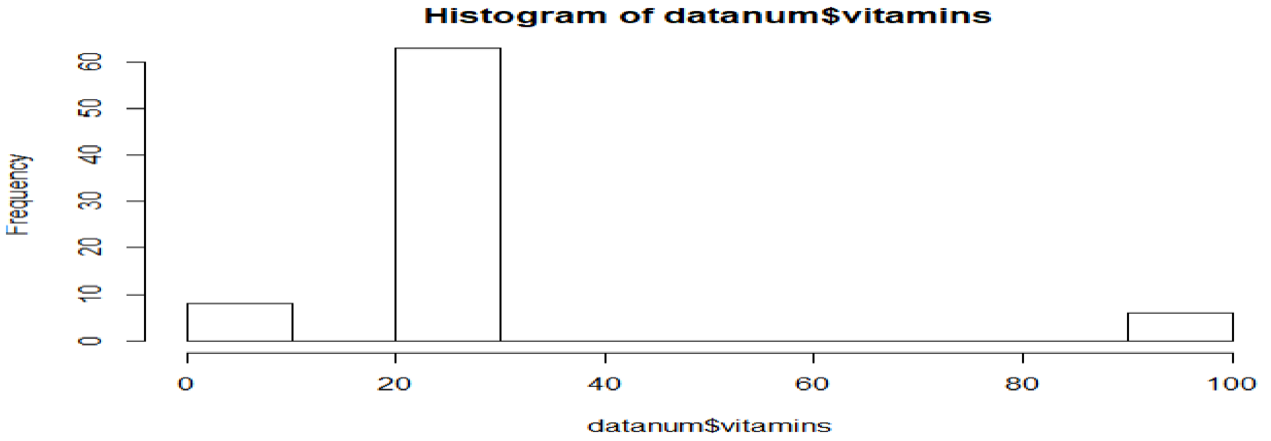

Histogram of datanum$protein
Frequency
0
5
10
15
20
25
1
2
3
4
5
6
datanum$protein

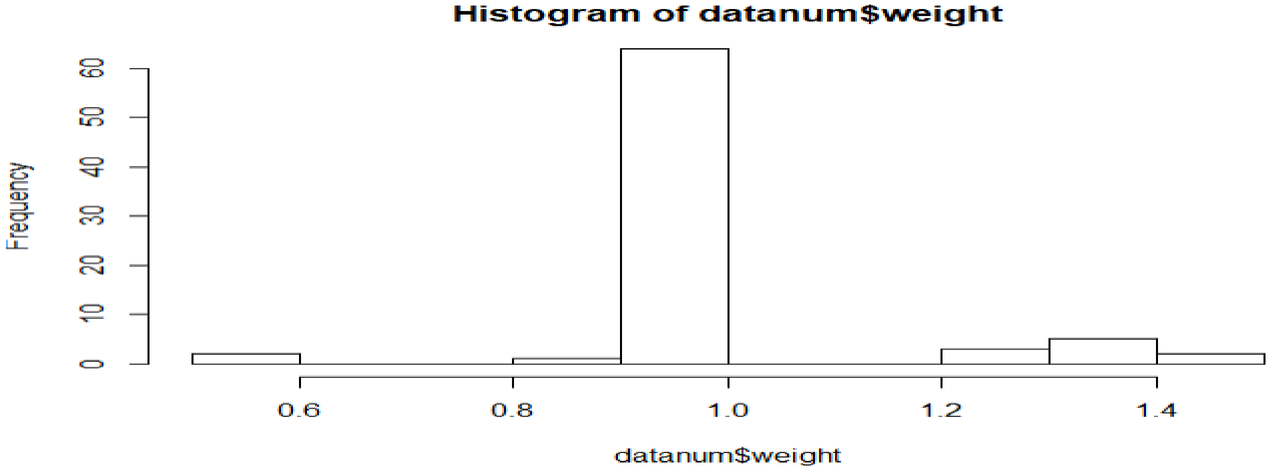

Histogram of datanum$rating
Frequency
0
5
10
15
20
20
40
60
80
100
datanum$rating


Histogram of datanum$shelf
Frequency
0
5
10
15
20
25
30
35
1.0
1.5
2.0
2.5
3.0
datanum$shelf


Histogram of datanum$sodium
Frequency
0
5
10
15
20
0
50
100
150
200
250
300
350
datanum$sodium


Histogram of datanum$sugars
Frequency
0
2
4
6
8
10
12
14
0
5
10
15
datanum$sugars





The following variables have the largest variability:

Sodium has the highest variance of 7027.854

Potassium has a variance of 4957.658

Vitamins has a variance of 499.1883

Calories has a variance of 379.6309

The following variables seems skewed:

Fiber,fat,potassium,vitamins

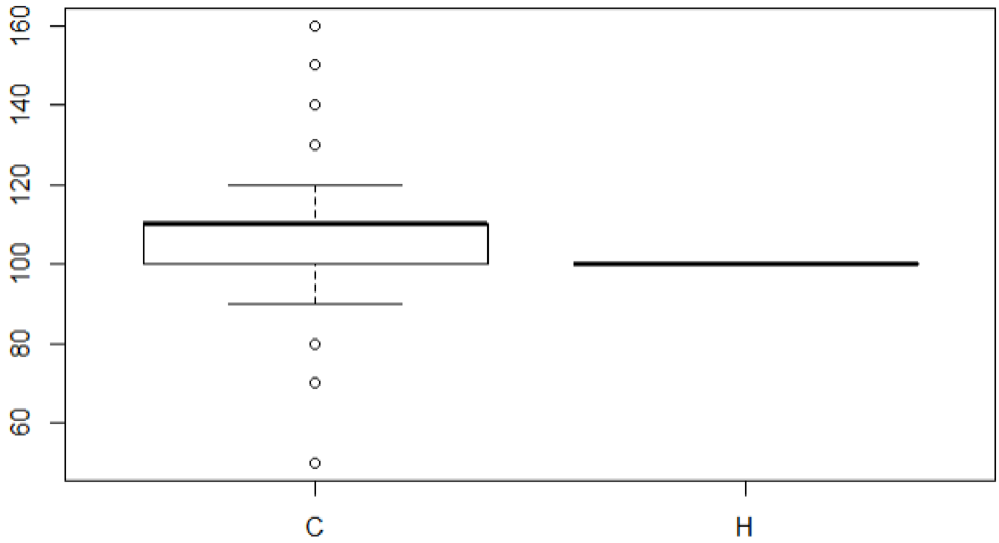
The following variables have extreme values as we can see above in their histogram plots:

vitamins,ratings,fat,fibre

d,Use R to plot a side-by-side boxplot comparing the calories in hot vs. cold cereals.

What does this plot show us?

```
boxplot(datanum$calories ~ train$type)
```



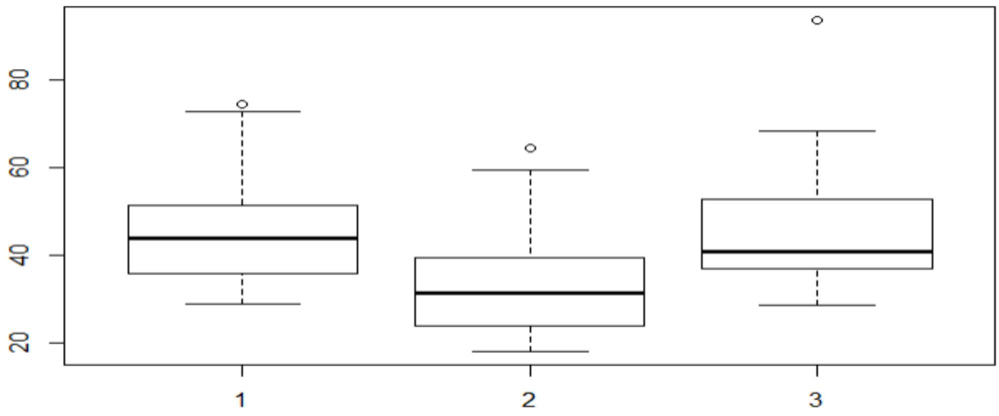
This box plot shows that the number of hot cereals type is very low(just 3) and all the three equals 100 calories

The number of cold cereals are very large(centered around 110),has few outliers towards both maximum and minimum value of calories.

e. Use R 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?

```
boxplot(datanum$rating ~ datanum$shelf)
```



To predict the consumer rating it doesn’t appear that we need to keep all three categories because shelf1 and shelf3 looks similar and can be combined.

f. Compute the correlation table for the quantitative variable (function cor()). In addition,

generate a matrix plot for these variables (function plot(data)).

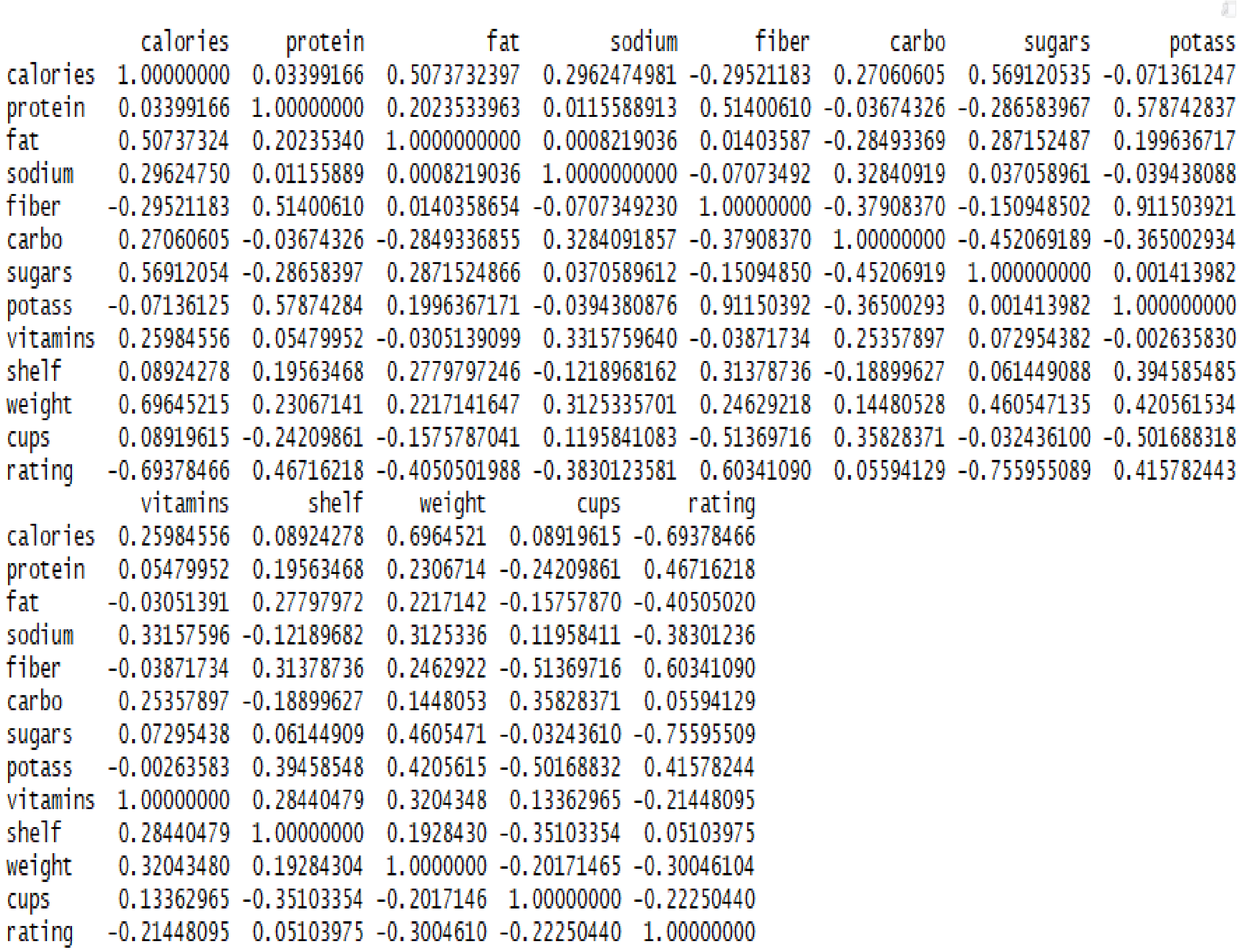
i. Which pair of variables is most strongly correlated?

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

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

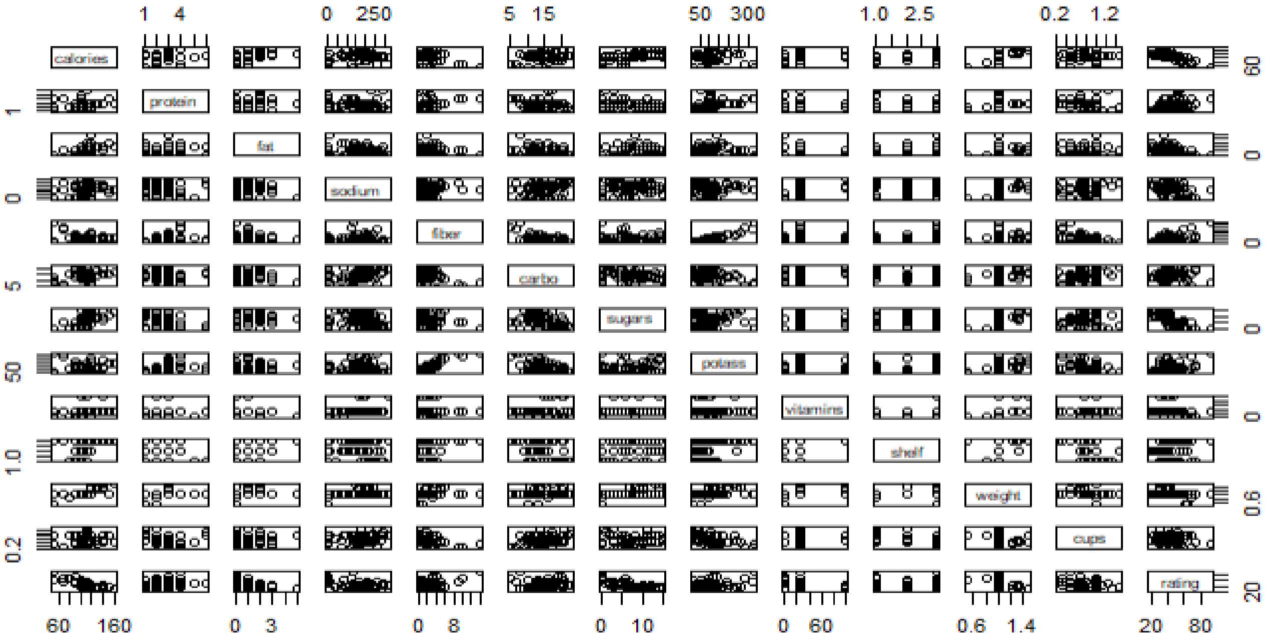
```
cor(datanum,use="complete.obs")
```

The following image shows the correlation table for the quantitative variables:



```
            calories     protein           fat        sodium       fiber       carbo       sugars       potass
calories  1.00000000  0.03399166  0.5073732397  0.2962474981 -0.29521183  0.27060605  0.569120535 -0.071361247
protein   0.03399166  1.00000000  0.2023533963  0.0115588913  0.51400610 -0.03674326 -0.286583967  0.578742837
fat       0.50737324  0.20235340  1.0000000000  0.0008219036  0.01403587 -0.28493369  0.287152487  0.199636717
sodium    0.29624750  0.01155889  0.0008219036  1.0000000000 -0.07073492  0.32840919  0.037058961 -0.039438088
fiber    -0.29521183  0.51400610  0.0140358654 -0.0707349230  1.00000000 -0.37908370 -0.150948502  0.911503921
carbo     0.27060605 -0.03674326 -0.2849336855  0.3284091857 -0.37908370  1.00000000 -0.452069189 -0.365002934
sugars    0.56912054 -0.28658397  0.2871524866  0.0370589612 -0.15094850 -0.45206919  1.000000000  0.001413982
potass   -0.07136125  0.57874284  0.1996367171 -0.0394380876  0.91150392 -0.36500293  0.001413982  1.000000000
vitamins  0.25984556  0.05479952 -0.0305139099  0.3315759640 -0.03871734  0.25357897  0.072954382 -0.002635830
shelf     0.08924278  0.19563468  0.2779797246 -0.1218968162  0.31378736 -0.18899627  0.061449088  0.394585485
weight    0.69645215  0.23067141  0.2217141647  0.3125335701  0.24629218  0.14480528  0.460547135  0.420561534
cups      0.08919615 -0.24209861 -0.1575787041  0.1195841083 -0.51369716  0.35828371 -0.032436100 -0.501688318
rating   -0.69378466  0.46716218 -0.4050501988 -0.3830123581  0.60341090  0.05594129 -0.755955089  0.415782443
            vitamins       shelf    weight        cups      rating
calories  0.25984556  0.08924278 0.6964521  0.08919615 -0.69378466
protein   0.05479952  0.19563468 0.2306714 -0.24209861  0.46716218
fat      -0.03051391  0.27797972 0.2217142 -0.15757870 -0.40505020
sodium    0.33157596 -0.12189682 0.3125336  0.11958411 -0.38301236
fiber    -0.03871734  0.31378736 0.2462922 -0.51369716  0.60341090
carbo     0.25357897 -0.18899627 0.1448053  0.35828371  0.05594129
sugars    0.07295438  0.06144909 0.4605471 -0.03243610 -0.75595509
potass   -0.00263583  0.39458548 0.4205615 -0.50168832  0.41578244
vitamins  1.00000000  0.28440479 0.3204348  0.13362965 -0.21448095
shelf     0.28440479  1.00000000 0.1928430 -0.35103354  0.05103975
weight    0.32043480  0.19284304 1.0000000 -0.20171465 -0.30046104
cups      0.13362965 -0.35103354 -0.2017146 1.00000000 -0.22250440
rating   -0.21448095  0.05103975 -0.3004610 -0.22250440  1.00000000
```

The following plot is the matrix plot obtained for all variables using plot():



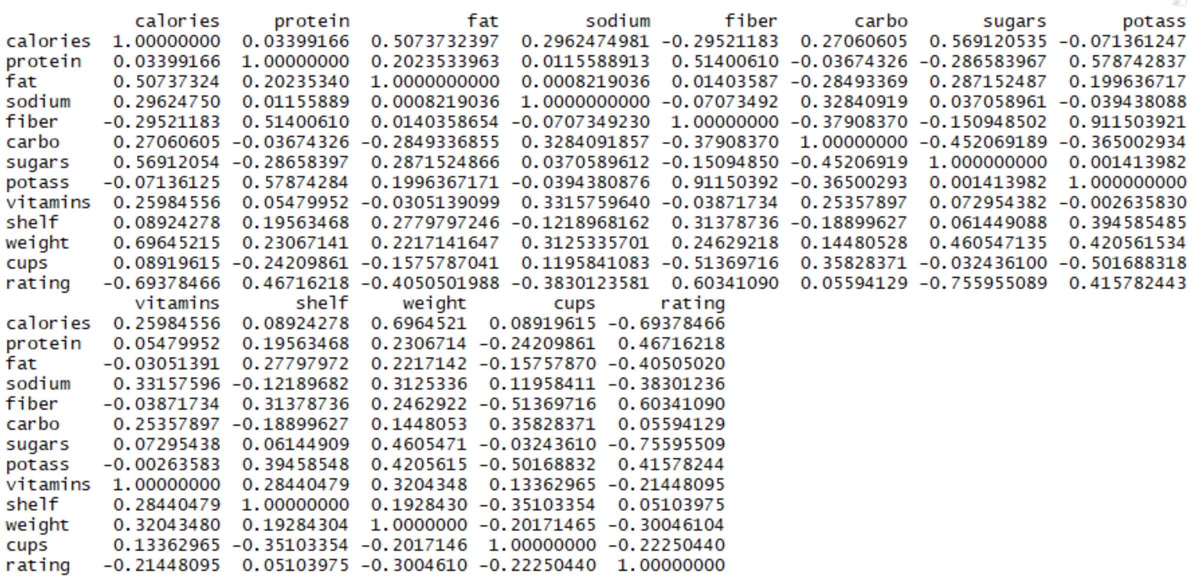
The following pair of variables are strongly correlated:

Potassium and fiber have a correlation of 0.911

We can reduce the number of variables or dimensions by:

Using principle component analysis by combining the variables that are highly correlated.

The following image is the correlation table for the variables after normalizing:



```
          calories     protein          fat        sodium      fiber      carbo       sugars       potass
calories  1.00000000  0.03399166  0.5073732397  0.2962474981 -0.29521183  0.27060605  0.569120535 -0.071361247
protein   0.03399166  1.00000000  0.2023533963  0.0115588913  0.51400610 -0.03674326 -0.286583967  0.578742837
fat       0.50737324  0.20235340  1.0000000000  0.0008219036  0.01403587 -0.28493369  0.287152487  0.199636717
sodium    0.29624750  0.01155889  0.0008219036  1.0000000000 -0.07073492  0.32840919  0.037058961 -0.039438088
fiber    -0.29521183  0.51400610  0.0140358654 -0.0707349230  1.00000000 -0.37908370 -0.150948502  0.911503921
carbo     0.27060605 -0.03674326 -0.2849336855  0.3284091857 -0.37908370  1.00000000 -0.452069189 -0.365002934
sugars    0.56912054 -0.28658397  0.2871524866  0.0370589612 -0.15094850 -0.45206919  1.000000000  0.001413982
potass   -0.07136125  0.57874284  0.1996367171 -0.0394380876  0.91150392 -0.36500293  0.001413982  1.000000000
vitamins  0.25984556  0.05479952 -0.0305139099  0.3315759640 -0.03871734  0.25357897  0.072954382 -0.002635830
shelf     0.08924278  0.19563468  0.2779797246 -0.1218968162  0.31378736 -0.18899627  0.061449088  0.394585485
weight    0.69645215  0.23067141  0.2217141647  0.3125335701  0.24629218  0.14480528  0.460547135  0.420561534
cups      0.08919615 -0.24209861 -0.1575787041  0.1195841083 -0.51369716  0.35828371 -0.032436100 -0.501688318
rating   -0.69378466  0.46716218 -0.4050501988 -0.3830123581  0.60341090  0.05594129 -0.755955089  0.415782443
          vitamins       shelf      weight       cups      rating
calories  0.25984556  0.08924278  0.6964521  0.08919615 -0.69378466
protein   0.05479952  0.19563468  0.2306714 -0.24209861  0.46716218
fat      -0.03051391  0.27797972  0.2217142 -0.15757870 -0.40505020
sodium    0.33157596 -0.12189682  0.3125336  0.11958411 -0.38301236
fiber    -0.03871734  0.31378736  0.2462922 -0.51369716  0.60341090
carbo     0.25357897 -0.18899627  0.1448053  0.35828371  0.05594129
sugars    0.07295438  0.06144909  0.4605471 -0.03243610 -0.75595509
potass   -0.00263583  0.39458548  0.4205615 -0.50168832  0.41578244
vitamins  1.00000000  0.28440479  0.3204348  0.13362965 -0.21448095
shelf     0.28440479  1.00000000  0.1928430 -0.35103354  0.05103975
weight    0.32043480  0.19284304  1.0000000 -0.20171465 -0.30046104
cups      0.13362965 -0.35103354 -0.2017146  1.00000000 -0.22250440
rating   -0.21448095  0.05103975 -0.3004610 -0.22250440  1.00000000
```

The correlation table remains same even after the normalization of data.

g. Consider the first PC of the analysis of the 13 numerical variables in Table 4.11.

Describe briefly what this PC represents.

first principal component measures the balance between 2 quantities:

1, calories and cups(positives)

2,protein,potassium,fiber and ratings(negatives)

The weights or co-efficient values in the first pc shows that the cereal is high in calories and amount per bowl ,low in protein and potassium. consequently gets the low consumer rating.

4.2 University Rankings. The dataset on American college and university rankings (available from www.dataminingbook.com) contains information on 1302 American colleges and universities offering an undergraduate program. For each university, there are 17 measurements that include continuous measurements (such as tuition and graduation rate) and categorical measurements (such as location by state and whether it is a private or a public school).

a. Remove all categorical variables. Then remove all records with missing numerical measurements from the dataset.

The dataset contains 1302 observations of 20 variables. Of these variables, two variables(State and Public..1…Private..2) are categorical variables and the college name column are both removed.

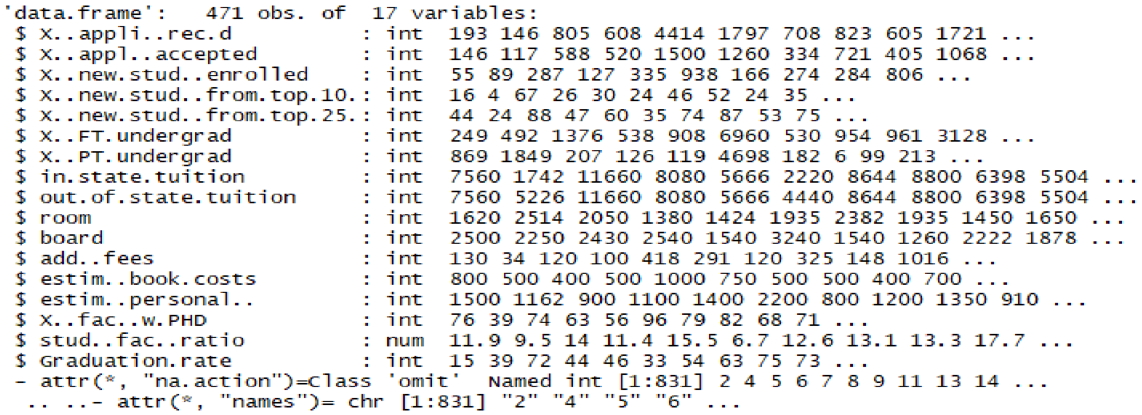
Then any observation with missing values are also removed.

The following image shows the cleaned data of university datasets:

It contains 471 observations of 17 variables.

dat_clean<-na.omit(within(dat, rm(College.Name, State, Public..1...Private..2.)))

str(dat_clean)



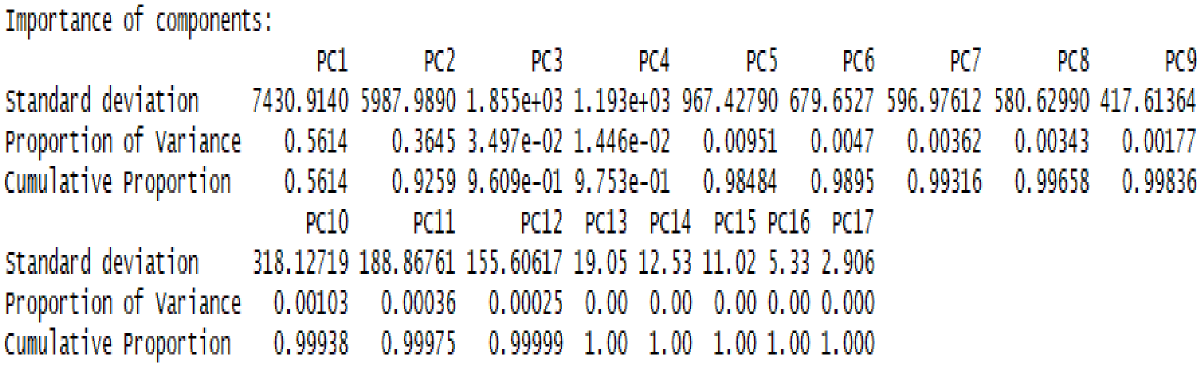
```
'data.frame':   471 obs. of  17 variables:
 $ X..appli..rec.d           : int  193 146 805 608 4414 1797 708 823 605 1721 ...
 $ X..appl..accepted         : int  146 117 588 520 1500 1260 334 721 405 1068 ...
 $ X..new.stud..enrolled     : int  55 89 287 127 335 938 166 274 284 806 ...
 $ X..new.stud..from.top.10.: int  16 4 67 26 30 24 46 52 24 35 ...
 $ X..new.stud..from.top.25.: int  44 24 88 47 60 35 74 87 53 75 ...
 $ X..FT.undergrad           : int  249 492 1376 538 908 6960 530 954 961 3128 ...
 $ X..PT.undergrad           : int  869 1849 207 126 119 4698 182 6 99 213 ...
 $ in.state.tuition          : int  7560 1742 11660 8080 5666 2220 8644 8800 6398 5504 ...
 $ out.of.state.tuition      : int  7560 5226 11660 8080 5666 4440 8644 8800 6398 5504 ...
 $ room                      : int  1620 2514 2050 1380 1424 1935 2382 1935 1450 1650 ...
 $ board                     : int  2500 2250 2430 2540 1540 3240 1540 1260 2222 1878 ...
 $ add..fees                 : int  130 34 120 100 418 291 120 325 148 1016 ...
 $ estim..book.costs         : int  800 500 400 500 1000 750 500 500 400 700 ...
 $ estim..personal..         : int  1500 1162 900 1100 1400 2200 800 1200 1350 910 ...
 $ X..fac..w.PHD             : int  76 39 74 63 56 96 79 82 68 71 ...
 $ stud..fac..ratio          : num  11.9 9.5 14 11.4 15.5 6.7 12.6 13.1 13.3 17.7 ...
 $ Graduation.rate           : int  15 39 72 44 46 33 54 63 75 73 ...
 - attr(*, "na.action")=Class 'omit'  Named int [1:831] 2 4 5 6 7 8 9 11 13 14 ...
  .. ..- attr(*, "names")= chr [1:831] "2" "4" "5" "6" ...
```

b,**Conduct a principal components analysis on the cleaned data and comment on the results. Should the data be normalized? Discuss what characterizes the components you consider key**.

The following image shows the result of applying pca on cleaned university dataset:

pca_results<- prcomp(dat_clean)

summary(pca_results)



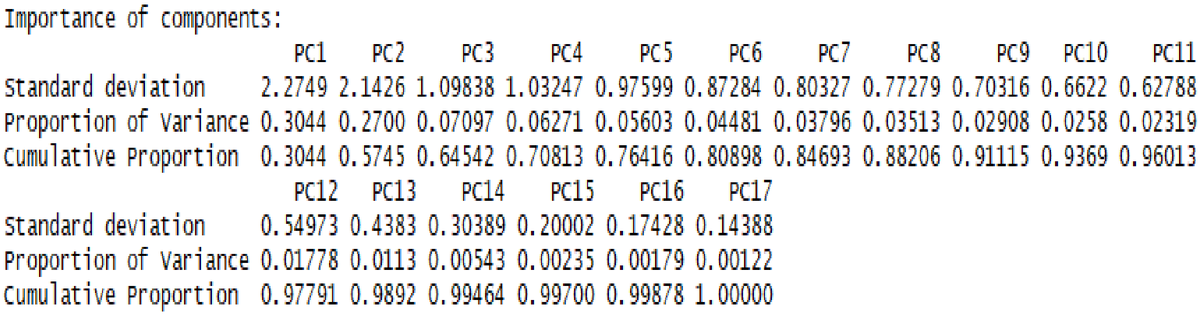
```
Importance of components:
                             PC1       PC2       PC3       PC4       PC5      PC6       PC7       PC8       PC9
Standard deviation     7430.9140 5987.9890 1.855e+03 1.193e+03 967.42790 679.6527 596.97612 580.62990 417.61364
Proportion of Variance    0.5614    0.3645 3.497e-02 1.446e-02   0.00951   0.0047   0.00362   0.00343   0.00177
Cumulative Proportion     0.5614    0.9259 9.609e-01 9.753e-01   0.98484   0.9895   0.99316   0.99658   0.99836
                            PC10      PC11      PC12  PC13  PC14  PC15 PC16  PC17
Standard deviation     318.12719 188.86761 155.60617 19.05 12.53 11.02 5.33 2.906
Proportion of Variance   0.00103   0.00036   0.00025  0.00  0.00  0.00 0.00 0.000
Cumulative Proportion    0.99938   0.99975   0.99999  1.00  1.00  1.00 1.00 1.000
```

The above photo shows that the pc1 and pc2 shows almost all the variance of the data. This is because the data is not normalized. since we have variables of different units, we have to normalize the data.

The following image shows the result of applying pca on normalized university dataset:

pca_results<- prcomp(dat_clean,scale. = T)

summary(pca_results)

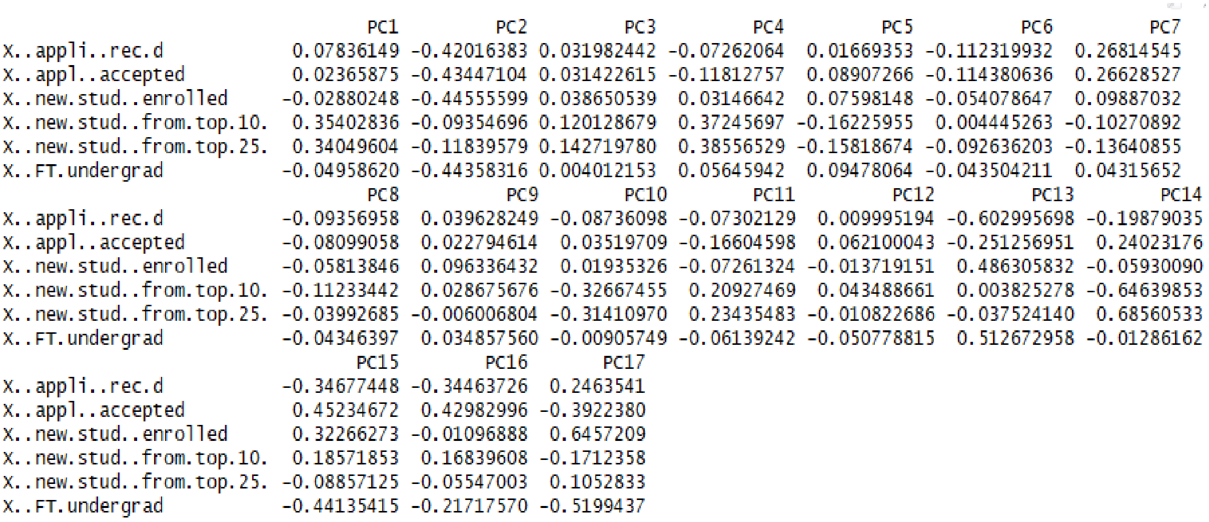


```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9   PC10    PC11
Standard deviation     2.2749 2.1426 1.09838 1.03247 0.97599 0.87284 0.80327 0.77279 0.70316 0.6622 0.62788
Proportion of Variance 0.3044 0.2700 0.07097 0.06271 0.05603 0.04481 0.03796 0.03513 0.02908 0.0258 0.02319
Cumulative Proportion  0.3044 0.5745 0.64542 0.70813 0.76416 0.80898 0.84693 0.88206 0.91115 0.9369 0.96013
                          PC12   PC13    PC14    PC15    PC16    PC17
Standard deviation     0.54973 0.4383 0.30389 0.20002 0.17428 0.14388
Proportion of Variance 0.01778 0.0113 0.00543 0.00235 0.00179 0.00122
Cumulative Proportion  0.97791 0.9892 0.99464 0.99700 0.99878 1.00000
```

As we normalize the data, the proportion of variance explained by each pc is accurate. The first 9 pc's explains the 90% of variance in the data,so we can use 9 variables(pc's) instead of 17 variables which proves the dimensionality reduction property of PCA.

The following image shows the weights or coefficients of the pc's:

head(pca_results$rotation)



```
                                   PC1         PC2         PC3         PC4         PC5          PC6         PC7
X..appli..rec.d             0.07836149 -0.42016383 0.031982442 -0.07262064  0.01669353 -0.112319932  0.26814545
X..appl..accepted           0.02365875 -0.43447104 0.031422615 -0.11812757  0.08907266 -0.114380636  0.26628527
X..new.stud..enrolled      -0.02880248 -0.44555599 0.038650539  0.03146642  0.07598148 -0.054078647  0.09887032
X..new.stud..from.top.10.   0.35402836 -0.09354696 0.120128679  0.37245697 -0.16225955  0.004445263 -0.10270892
X..new.stud..from.top.25.   0.34049604 -0.11839579 0.142719780  0.38556529 -0.15818674 -0.092636203 -0.13640855
X..FT.undergrad            -0.04958620 -0.44358316 0.004012153  0.05645942  0.09478064 -0.043504211  0.04315652
                                   PC8          PC9        PC10        PC11         PC12         PC13        PC14
X..appli..rec.d            -0.09356958  0.039628249 -0.08736098 -0.07302129  0.009995194 -0.602995698 -0.19879035
X..appl..accepted          -0.08099058  0.022794614  0.03519709 -0.16604598  0.062100043 -0.251256951  0.24023176
X..new.stud..enrolled      -0.05813846  0.096336432  0.01935326 -0.07261324 -0.013719151  0.486305832 -0.05930090
X..new.stud..from.top.10.  -0.11233442  0.028675676 -0.32667455  0.20927469  0.043488661  0.003825278 -0.64639853
X..new.stud..from.top.25.  -0.03992685 -0.006006804 -0.31410970  0.23435483 -0.010822686 -0.037524140  0.68560533
X..FT.undergrad            -0.04346397  0.034857560 -0.00905749 -0.06139242 -0.050778815  0.512672958 -0.01286162
                                  PC15        PC16       PC17
X..appli..rec.d            -0.34677448 -0.34463726  0.2463541
X..appl..accepted           0.45234672  0.42982996 -0.3922380
X..new.stud..enrolled       0.32266273 -0.01096888  0.6457209
X..new.stud..from.top.10.   0.18571853  0.16839608 -0.1712358
X..new.stud..from.top.25.  -0.08857125 -0.05547003  0.1052833
X..FT.undergrad            -0.44135415 -0.21717570 -0.5199437
```

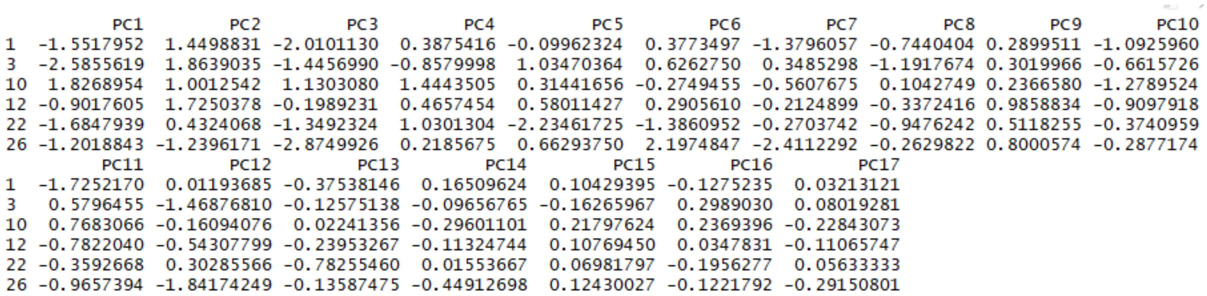
In the above table, we can see that each pc weights represent a relationship between the pc and the corresponding variable .Each pc measures the balance between two components :

1,variables with large positive weight

2,variables with large negative weight

The following image shows the new dimensions of the original observation:

head(pca_results$x)



```
          PC1        PC2        PC3        PC4         PC5        PC6        PC7        PC8       PC9       PC10
1  -1.5517952  1.4498831 -2.0101130  0.3875416 -0.09962324  0.3773497 -1.3796057 -0.7440404 0.2899511 -1.0925960
3  -2.5855619  1.8639035 -1.4456990 -0.8579998  1.03470364  0.6262750  0.3485298 -1.1917674 0.3019966 -0.6615726
10  1.8268954  1.0012542  1.1303080  1.4443505  0.31441656 -0.2749455 -0.5607675  0.1042749 0.2366580 -1.2789524
12 -0.9017605  1.7250378 -0.1989231  0.4657454  0.58011427  0.2905610 -0.2124899 -0.3372416 0.9858834 -0.9097918
22 -1.6847939  0.4324068 -1.3492324  1.0301304 -2.23461725 -1.3860952 -0.2703742 -0.9476242 0.5118255 -0.3740959
26 -1.2018843 -1.2396171 -2.8749926  0.2185675  0.66293750  2.1974847 -2.4112292 -0.2629822 0.8000574 -0.2877174
         PC11        PC12        PC13        PC14        PC15       PC16        PC17
1  -1.7252170  0.01193685 -0.37538146  0.16509624  0.10429395 -0.1275235  0.03213121
3   0.5796455 -1.46876810 -0.12575138 -0.09656765 -0.16265967  0.2989030  0.08019281
10  0.7683066 -0.16094076  0.02241356 -0.29601101  0.21797624  0.2369396 -0.22843073
12 -0.7822040 -0.54307799 -0.23953267 -0.11324744  0.10769450  0.0347831 -0.11065747
22 -0.3592668  0.30285566 -0.78255460  0.01553667  0.06981797 -0.1956277  0.05633333
26 -0.9657394 -1.84174249 -0.13587475 -0.44912698  0.12430027 -0.1221792 -0.29150801
```