hw3_p7

Haoyu_Zhao,2016012390

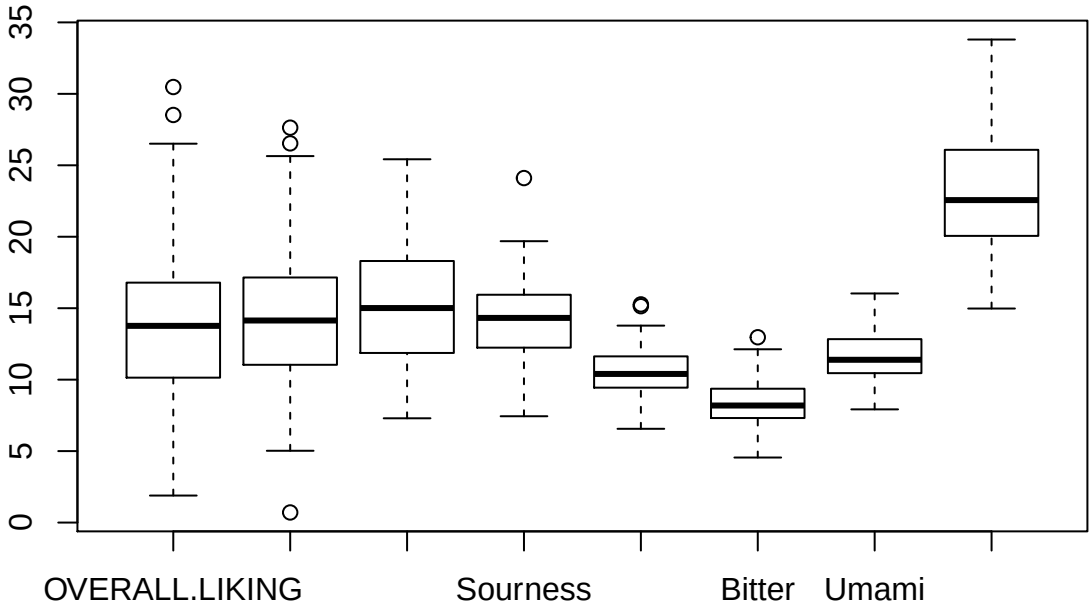
April 18, 2017

the code for the first sub-problem

Sub-problem 1

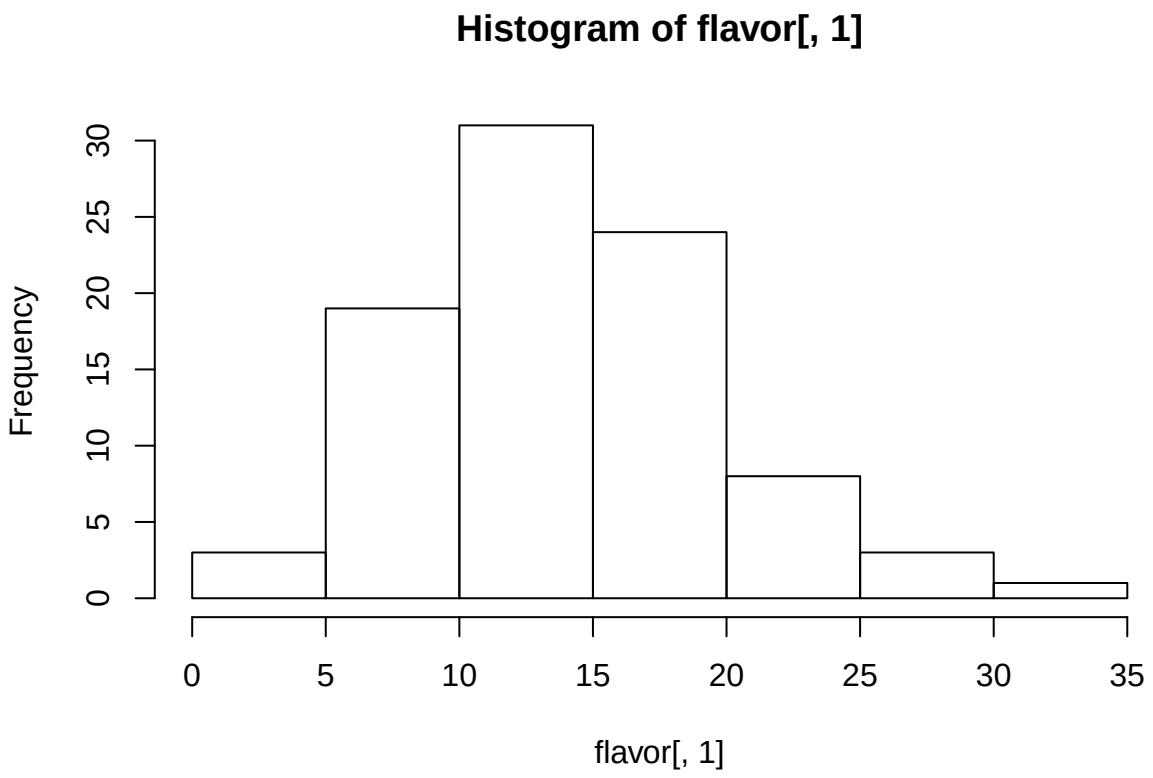
The boxplot of the flavor data

```
flavor <- read.csv("flavor.csv", sep=";")
flavor <- flavor[,2:9]
boxplot(flavor)
```

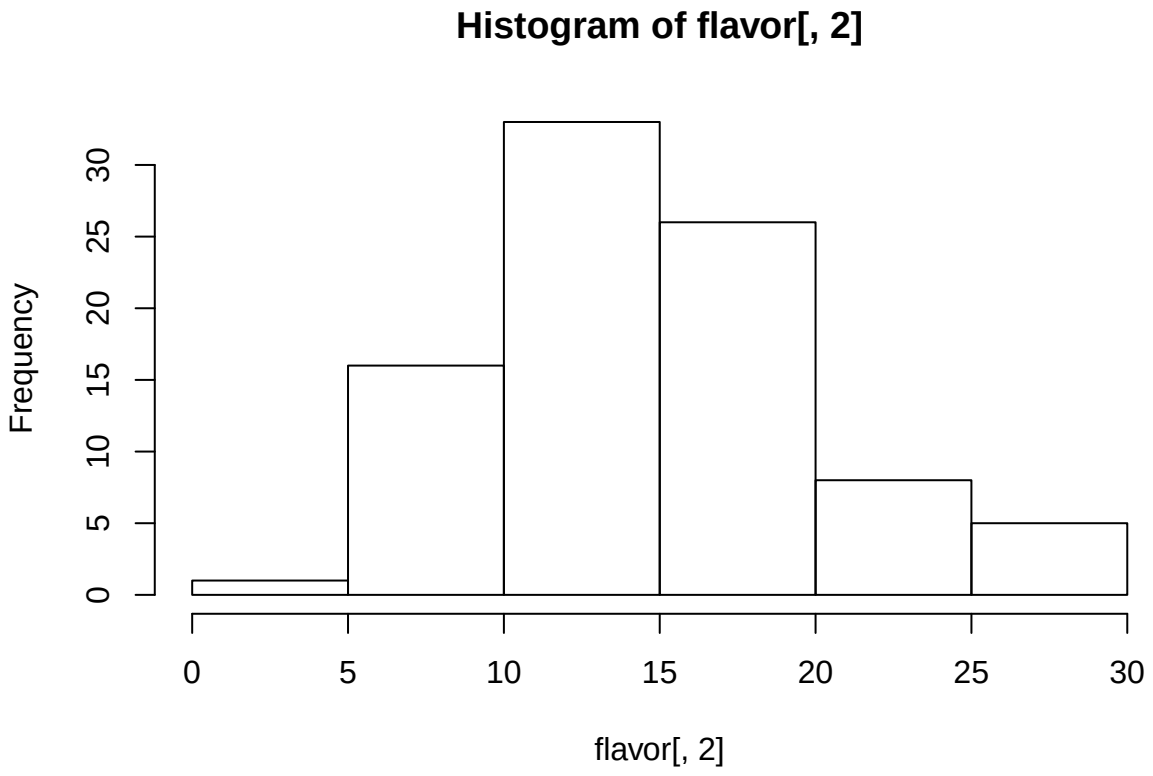


The histograms of the each of the elements

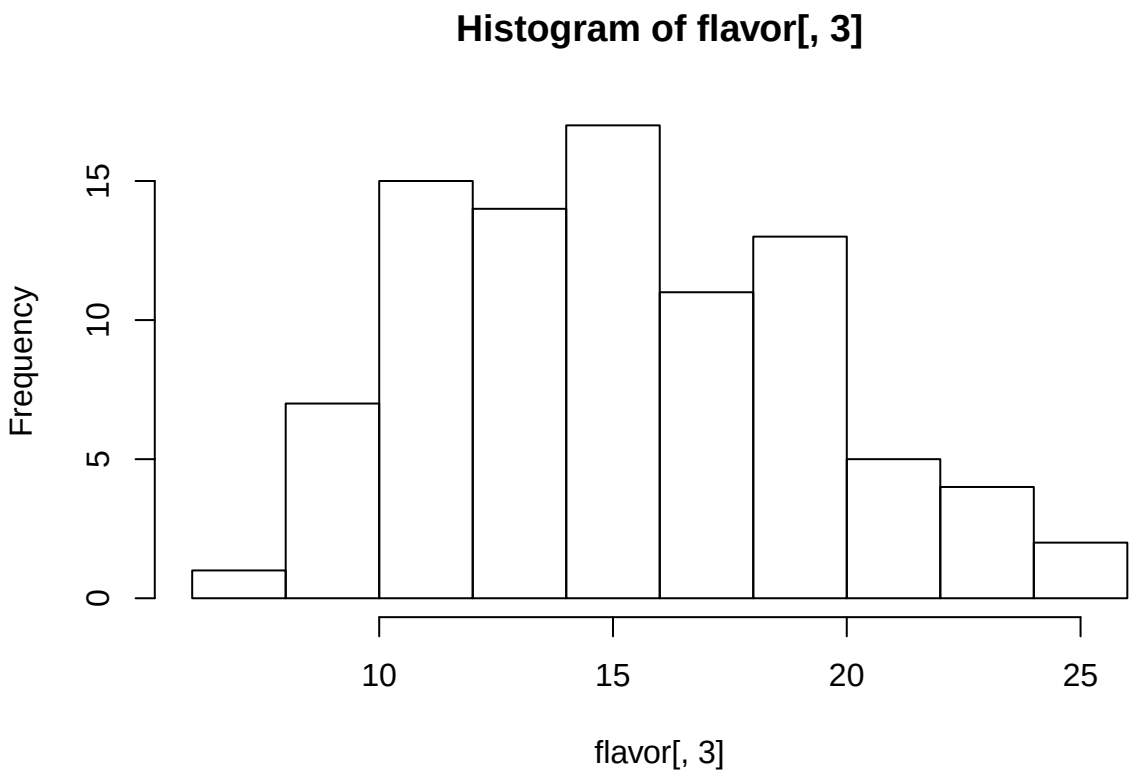
```
hist(flavor[,1])
```



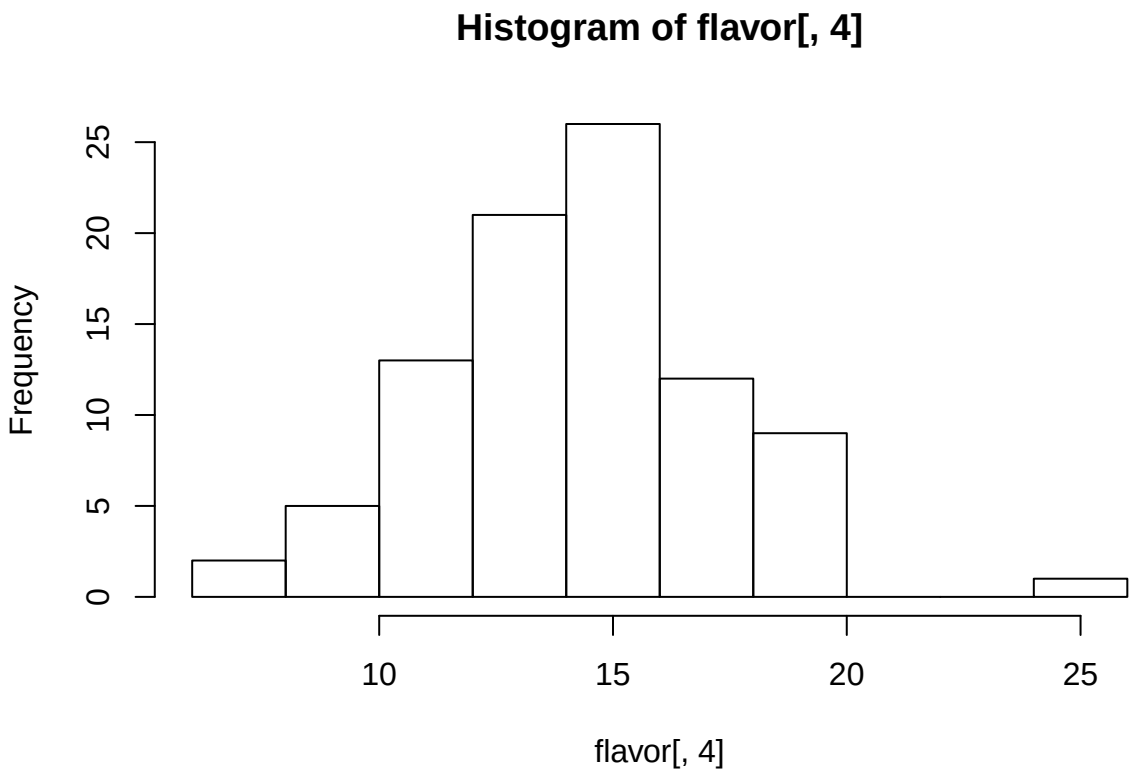

```
hist(flavor[,2])
```



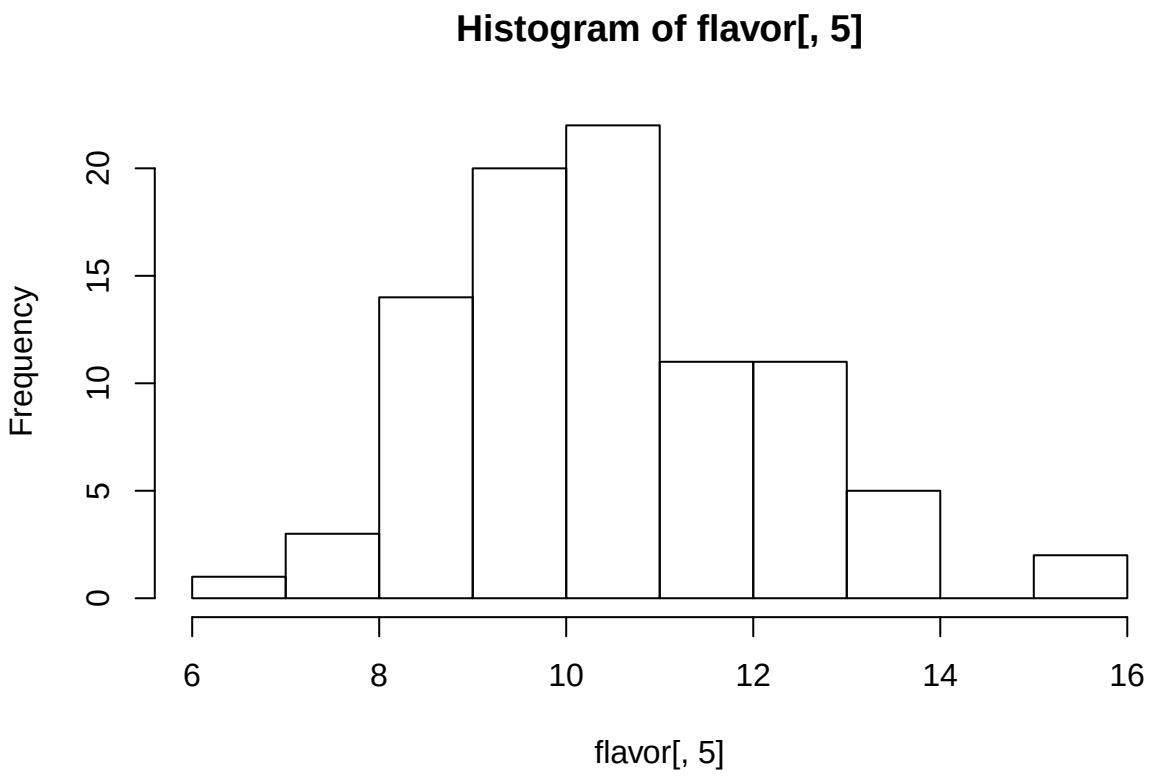

```
hist(flavor[,3])
```



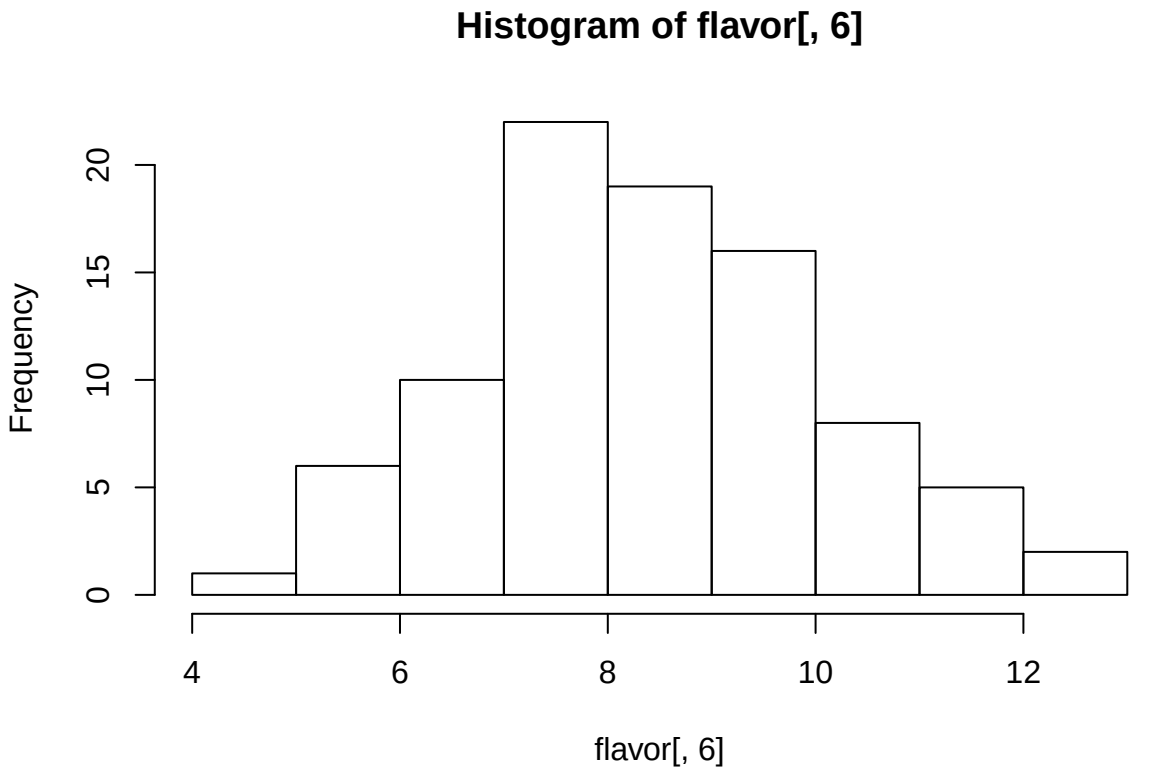

```
hist(flavor[,4])
```



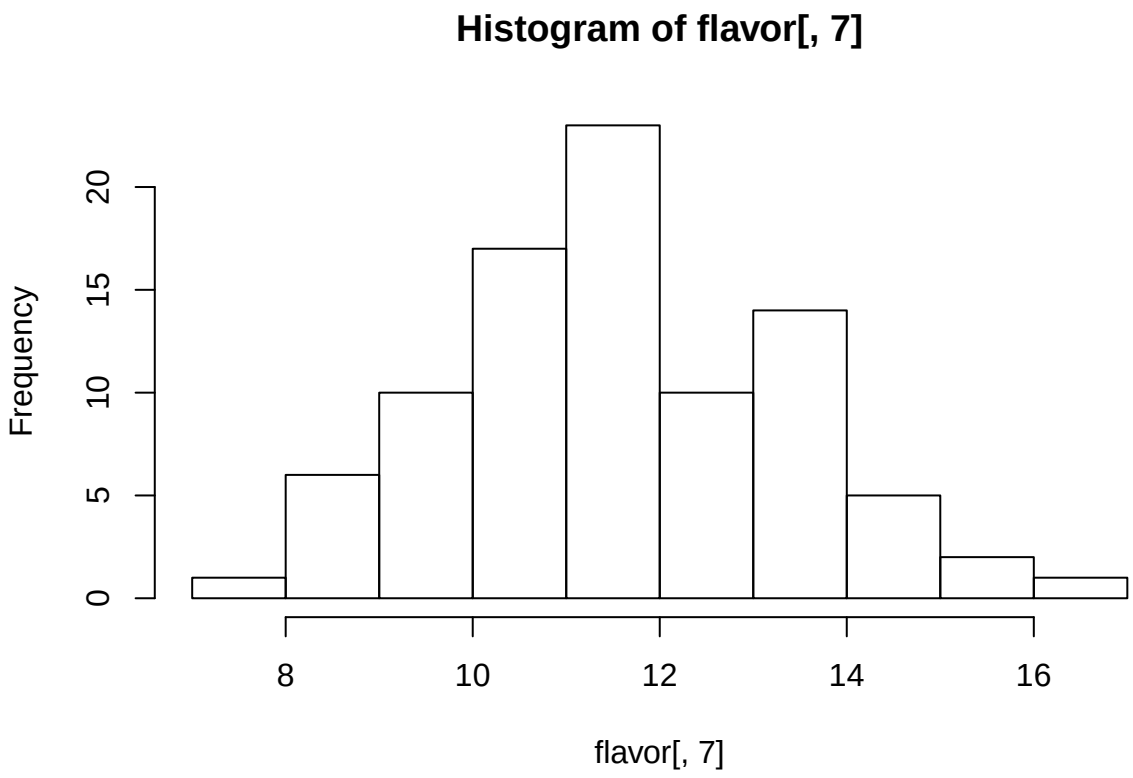

```
hist(flavor[,5])
```



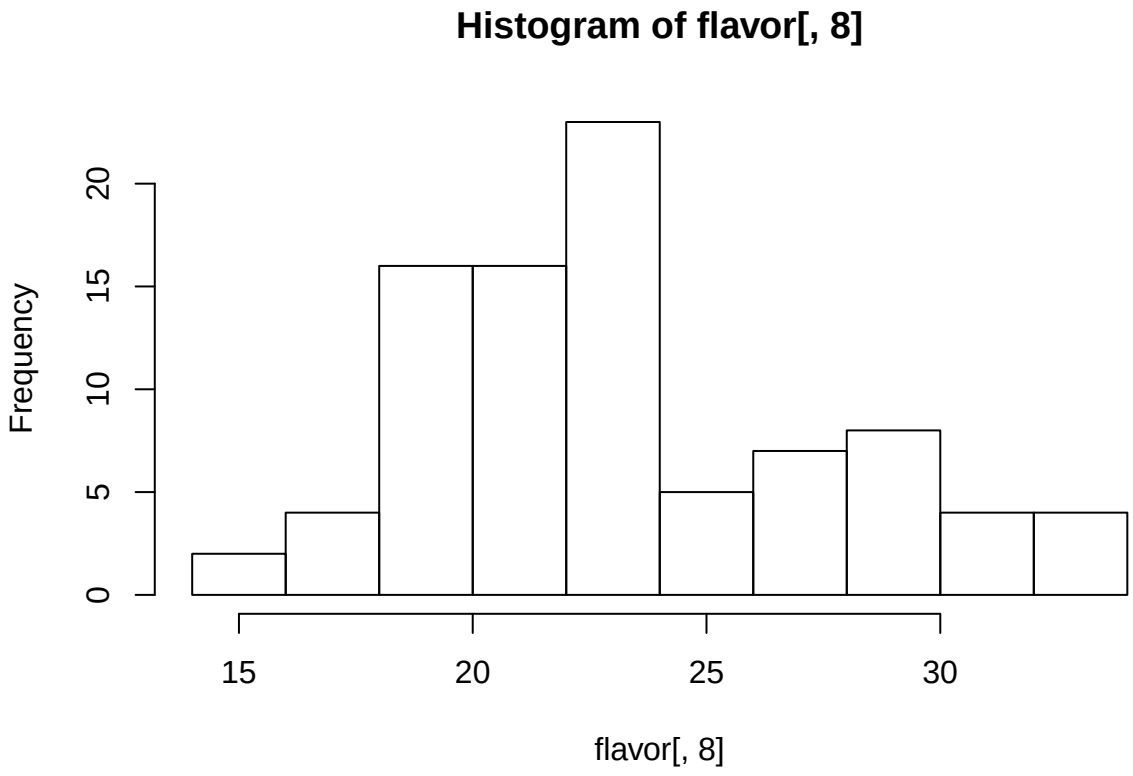

```
hist(flavor[,6])
```




```
hist(flavor[,7])
```

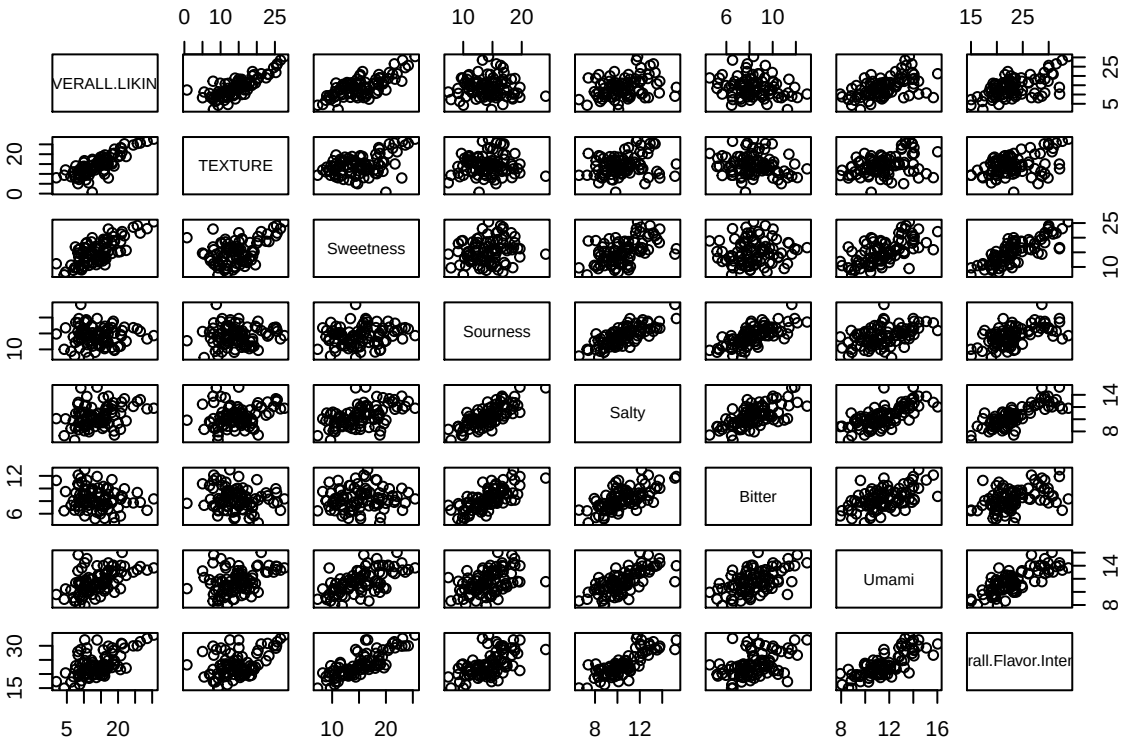



```
hist(flavor[,8])
```



The scatter plots of the variables

```
pairs(flavor)
```



The outlier detection can be done by the boxplot and the scatter plot and by the PC analysis mentioned below.

sub-problem 2

Skip the report of the fundamental analysis of the data for the chemical data.\ Just read the data from the .csv file and store it in a variable.

```
chemical <- read.csv("chemical.csv", sep=";")
chemical <- chemical[,2:69]
```

sub-problem 3

PCA analysis of the flavor data and the chemical data

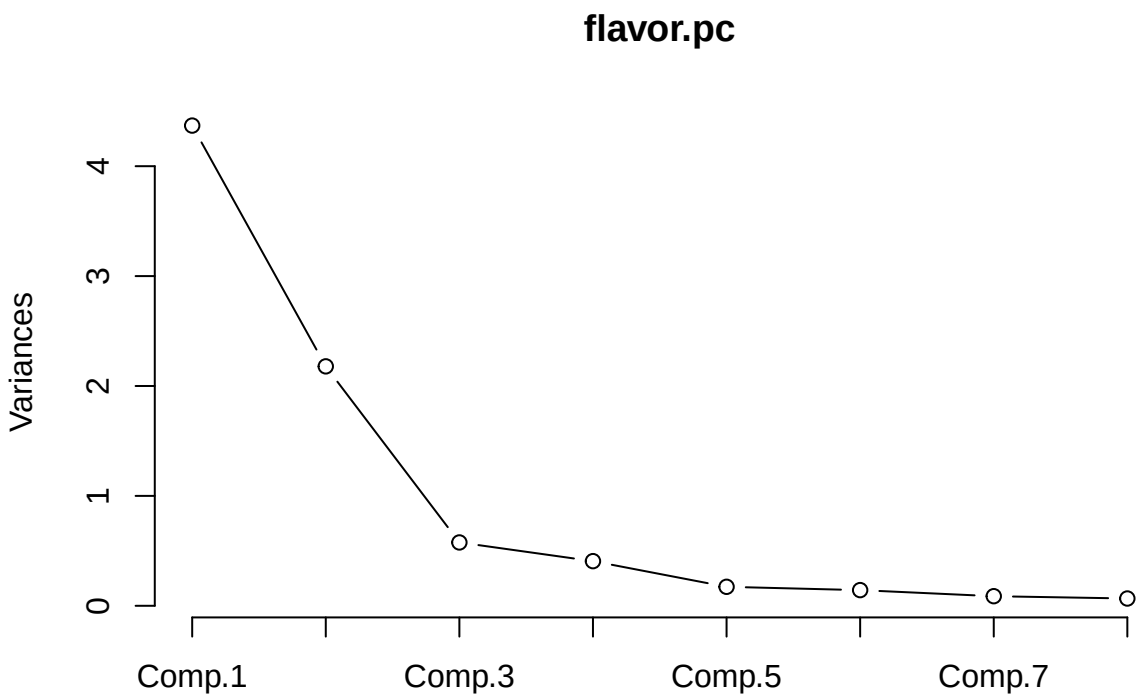
```
#use the build-in method to do the PC analysis for the flavor data
flavor.pc <- princomp(flavor,cor=TRUE)

#print the summary with the loadings
summary(flavor.pc, loadings = TRUE)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3    Comp.4     Comp.5
## Standard deviation     2.0903122 1.4760046 0.75926337 0.6374110 0.41582060
## Proportion of Variance 0.5461756 0.2723237 0.07206011 0.0507866 0.02161335
## Cumulative Proportion  0.5461756 0.8184993 0.89055944 0.9413460 0.96295938
##                            Comp.6     Comp.7      Comp.8
## Standard deviation     0.37765536 0.29548406 0.257663613
```

```
## Proportion of Variance 0.01782795 0.01091385 0.008298817
## Cumulative Proportion  0.98078733 0.99170118 1.000000000
##
## Loadings:
##                          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## OVERALL.LIKING           -0.308  0.493                -0.141        -0.464
## TEXTURE                  -0.249  0.438 -0.692 -0.158  0.278         0.202
## Sweetness                -0.372  0.265  0.511  0.324  0.426        -0.292
## Sourness                 -0.304 -0.416 -0.414  0.417 -0.120  0.526 -0.314
## Salty                    -0.416 -0.234         0.197 -0.381 -0.735
## Bitter                   -0.250 -0.515        -0.416  0.625 -0.188 -0.128
## Umami                    -0.413         0.230 -0.648 -0.411  0.339
## Overall.Flavor.Intensity -0.454         0.159  0.238         0.140  0.737
##                          Comp.8
## OVERALL.LIKING            0.638
## TEXTURE                  -0.351
## Sweetness                -0.394
## Sourness
## Salty                    -0.207
## Bitter                    0.235
## Umami                    -0.266
## Overall.Flavor.Intensity  0.381
```

```
#print the scree plot of the variance
screeplot(flavor.pc, type = "lines")
```



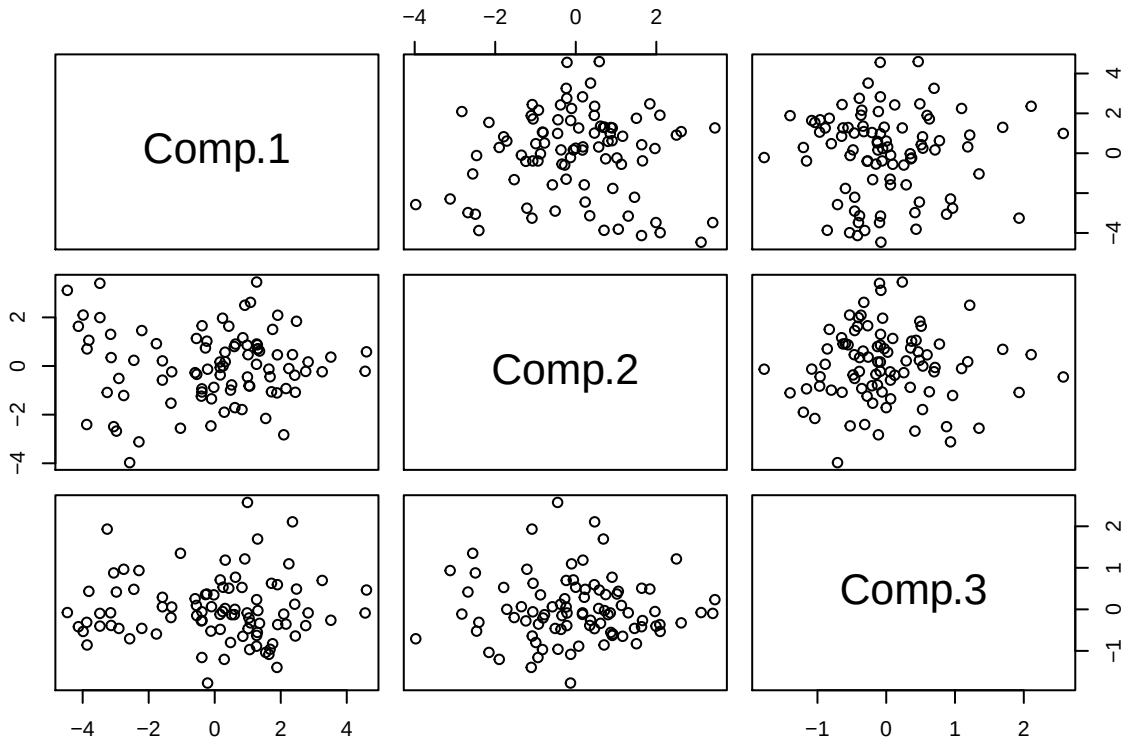

```
#from the scree plot we see that the first 3 component is
#really important compared to other components, so we just plot
#the scores and the scatter plot for the first 3 components.
flavor.pc$scores[,1:3]
```

```
##            Comp.1       Comp.2        Comp.3
##  [1,] -1.32538966 -1.528728585 -0.1972647759
```

```
##  [2,] -2.44940743  0.234074655  0.4820309228
##  [3,]  1.71898265 -1.059876154  0.6269681307
##  [4,] -0.21631151 -0.131237706 -1.7748067883
##  [5,] -3.05984277 -2.493204425  0.8771257511
##  [6,] -2.97868814 -2.676855319  0.4177913154
##  [7,]  2.75618749 -0.222871753 -0.3910462297
##  [8,] -2.57646950 -3.973533550 -0.7084703252
##  [9,]  1.26541606  0.071970445 -0.8873239367
## [10,]  1.89826872  0.460129505  0.5941129713
## [11,] -0.38318731  1.656846911 -0.2693533310
## [12,]  1.75368917  1.506156433 -0.8279850005
## [13,]  1.30766430  0.713579545 -0.0290877319
## [14,]  1.29090799  0.870908572 -0.5550974103
## [15,] -0.23772540  1.022239647  0.3696093756
## [16,]  1.35720536  0.614773304 -0.3367718950
## [17,] -0.55459286  1.133094108  0.0944933400
## [18,]  1.67711337 -0.441982216 -0.9615857253
## [19,] -0.54431783 -0.342204167 -0.1496507219
## [20,] -3.15395305  1.303265269 -0.0829330587
## [21,] -2.90317512 -0.511175685 -0.4595815399
## [22,]  0.16678373 -0.366532096 -0.4825045998
## [23,]  2.09098897 -2.828551034 -0.1139861169
## [24,] -1.30224418 -0.238537924  0.0565570269
## [25,] -0.59391825 -0.274330154  0.2557154792
## [26,] -0.40476669 -1.239759319 -0.2779796979
## [27,]  0.28873170 -1.900196562 -1.2048091071
## [28,]  2.44175459 -1.080704368 -0.6418017224
## [29,]  4.55710991 -0.217162651 -0.0874115887
## [30,]  1.64086828 -0.126632673 -1.0834165593
## [31,] -3.47723971  1.991533010 -0.4037809184
## [32,] -1.58464718  0.211355861  0.2906305337
## [33,]  2.83016829  0.168389757 -0.0852940929
## [34,]  1.90760631  2.089075739 -0.3668306784
## [35,]  0.42679344  1.637145489  0.5091486453
## [36,]  2.42414485 -0.378197542  0.1235664969
## [37,]  0.51263848 -0.775891596 -0.1279281898
## [38,]  3.25617062 -0.236861837  0.6961703826
## [39,]  3.51748763  0.366590066 -0.2624958880
## [40,]  0.16014115  0.166306159 -0.1208826822
## [41,] -0.09819471 -1.349984021  0.0639024103
## [42,] -0.28050502  0.742388689  0.3557500129
## [43,]  1.26930541  3.454738870  0.2339211465
## [44,] -4.46435716  3.112372302 -0.0769902825
## [45,] -3.81297736  1.056939236  0.4341298876
## [46,] -3.47905887  3.401452949 -0.0974930453
## [47,] -3.98575498  2.095895577 -0.5323268207
## [48,] -3.86676370  0.702750671 -0.8573286371
## [49,] -1.76656919  0.918927300 -0.5936639021
## [50,] -4.12973386  1.632993501 -0.4180343480
## [51,] -1.58498140 -0.580808889  0.0618764556
## [52,]  1.88516327 -1.105685283 -1.3979191286
## [53,] -2.29500163 -3.119442020  0.9347697969
## [54,] -1.03838817 -2.558985022  1.3494199873
## [55,]  4.59691133  0.583861582  0.4651080582
```

```
## [56,]  0.61056024 -1.711978876  0.0002011065
## [57,] -0.38876153 -0.938629982 -1.1589779501
## [58,]  1.01074999  0.464448025 -0.4646325981
## [59,]  0.98144668  0.856801704 -0.0849857419
## [60,]  0.23711198  1.967005889 -0.0522792701
## [61,]  2.15887361 -0.926235605 -0.3553945018
## [62,] -3.87750494 -2.406373729 -0.3124304626
## [63,]  0.47856671 -0.991677118 -0.7982990219
## [64,]  1.54455927 -2.156121156 -1.0375940904
## [65,]  1.08333059  2.617778266 -0.3254324219
## [66,]  0.91137508  2.499389203  1.2139377681
## [67,]  2.47611946  1.839479594  0.4888801732
## [68,]  1.29814961  0.685247263  1.6937445762
## [69,] -0.38733679 -1.067418410 -0.0597385602
## [70,]  0.31882585  0.177086672  1.1851096646
## [71,]  0.31363123  0.568583613  0.0208116568
## [72,]  0.59499220  0.801525402 -0.1272943650
## [73,] -2.20902031  1.458755698 -0.4586110364
## [74,]  0.85227014  1.163569239 -0.6486949954
## [75,]  1.27018819  0.906991201 -0.6277740599
## [76,] -2.75736679 -1.212965463  0.9652244549
## [77,]  2.24613435 -0.104386268  1.0957429170
## [78,]  0.16674305 -0.059601734  0.7094768165
## [79,]  1.03884878 -0.812282907 -0.2075078342
## [80,]  0.82639956 -1.790059592  0.5278761461
## [81,] -3.25709183 -1.087285487  1.9323320887
## [82,]  1.05923701 -0.826529270 -0.9683225943
## [83,]  0.62848464  0.901851250  0.7736433240
## [84,]  0.25259644 -0.001911541  0.5325168470
## [85,] -0.02655530 -0.876669964  0.3491844646
## [86,] -0.11498911 -2.462825805 -0.5241664843
## [87,]  0.99124789 -0.450713874  2.5749042278
## [88,]  2.35544952  0.468289931  2.1073829547
## [89,] -3.13730591  0.347041231 -0.3877948486
```

```
pairs(flavor.pc$scores[,1:3])
```



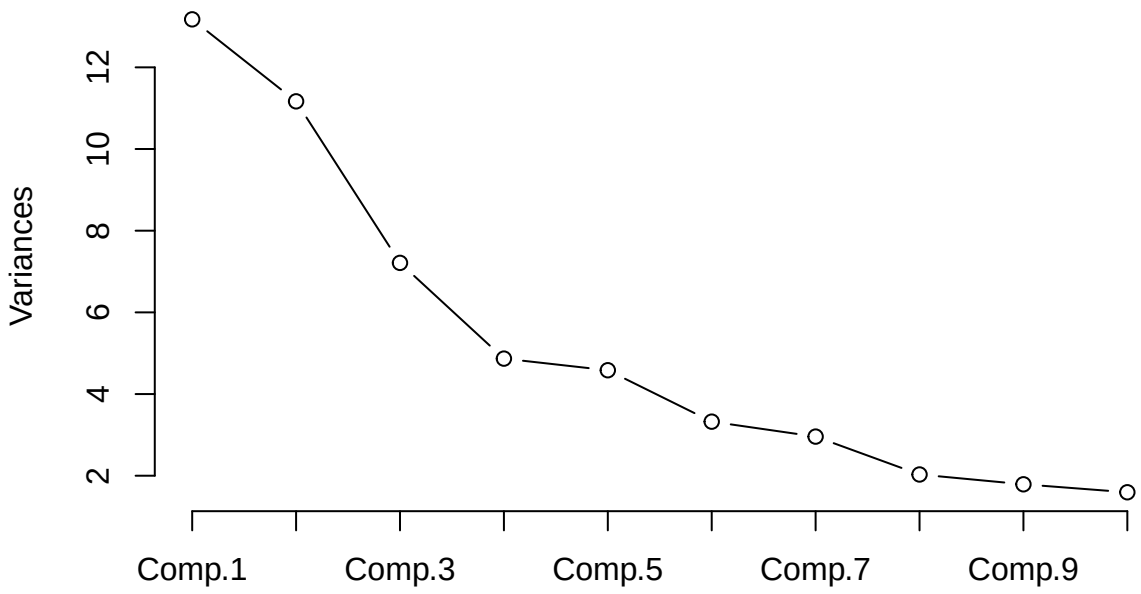

```
#do the PC analysis for the chemical data
chemical.pc <- princomp(chemical, cor=TRUE)

#print the summary, omit the loadings because
#the number of the principal component to fairly big.
summary(chemical.pc)
```

```
## Importance of components:
##                            Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     3.6297425 3.3417802 2.6854591 2.20626638 2.14076568
## Proportion of Variance 0.1937504 0.1642279 0.1060543 0.07158252 0.06739526
## Cumulative Proportion  0.1937504 0.3579783 0.4640326 0.53561511 0.60301037
##                            Comp.6     Comp.7     Comp.8     Comp.9
## Standard deviation     1.82321991 1.71987203 1.42481266 1.33796570
## Proportion of Variance 0.04888428 0.04349941 0.02985428 0.02632577
## Cumulative Proportion  0.65189464 0.69539405 0.72524833 0.75157410
##                           Comp.10    Comp.11    Comp.12    Comp.13
## Standard deviation     1.26287640 1.18496874 1.10581896 1.05153105
## Proportion of Variance 0.02345378 0.02064928 0.01798288 0.01626055
## Cumulative Proportion  0.77502788 0.79567715 0.81366003 0.82992058
##                           Comp.14    Comp.15    Comp.16    Comp.17
## Standard deviation     1.02062242 0.97635588 0.94393524 0.85499821
## Proportion of Variance 0.01531868 0.01401869 0.01310314 0.01075032
## Cumulative Proportion  0.84523926 0.85925795 0.87236109 0.88311141
##                           Comp.18     Comp.19     Comp.20     Comp.21
## Standard deviation     0.819395142 0.802659143 0.744128394 0.700500563
## Proportion of Variance 0.009873653 0.009474437 0.008143045 0.007216192
## Cumulative Proportion  0.892985068 0.902459504 0.910602550 0.917818741
##                           Comp.22    Comp.23     Comp.24     Comp.25
## Standard deviation     0.679014988 0.67607165 0.630045844 0.590711171
## Proportion of Variance 0.006780314 0.00672166 0.005837614 0.005131466
```

```
## Cumulative Proportion  0.924599055 0.93132072 0.937158329 0.942289795
##                            Comp.26     Comp.27     Comp.28     Comp.29
## Standard deviation     0.582333950 0.568323773 0.553265523 0.521397414
## Proportion of Variance 0.004986953 0.004749881 0.004501511 0.003997872
## Cumulative Proportion  0.947276749 0.952026630 0.956528141 0.960526012
##                            Comp.30     Comp.31     Comp.32     Comp.33
## Standard deviation     0.504949950 0.483038356 0.461154360 0.439207822
## Proportion of Variance 0.003749624 0.003431265 0.003127402 0.002836816
## Cumulative Proportion  0.964275636 0.967706902 0.970834304 0.973671120
##                            Comp.34    Comp.35     Comp.36     Comp.37
## Standard deviation     0.396881224 0.38474479 0.382684385 0.370011091
## Proportion of Variance 0.002316393 0.00217689 0.002153637 0.002013356
## Cumulative Proportion  0.975987513 0.97816440 0.980318041 0.982331397
##                            Comp.38     Comp.39     Comp.40     Comp.41
## Standard deviation     0.347063451 0.339245446 0.314307102 0.306402603
## Proportion of Variance 0.001771368 0.001692463 0.001452779 0.001380626
## Cumulative Proportion  0.984102765 0.985795228 0.987248007 0.988628633
##                            Comp.42     Comp.43   Comp.44      Comp.45
## Standard deviation     0.292934930 0.286825616 0.2721749 0.2600270989
## Proportion of Variance 0.001261925 0.001209837 0.0010894 0.0009943249
## Cumulative Proportion  0.989890557 0.991100394 0.9921898 0.9931841189
##                            Comp.46      Comp.47      Comp.48      Comp.49
## Standard deviation     0.247118053 0.2289864925 0.2246512307 0.2018539327
## Proportion of Variance 0.000898049 0.0007711002 0.0007421791 0.0005991913
## Cumulative Proportion  0.994082168 0.9948532681 0.9955954472 0.9961946385
##                             Comp.50      Comp.51      Comp.52      Comp.53
## Standard deviation     0.1857786158 0.1767112172 0.1652781979 0.1476744680
## Proportion of Variance 0.0005075543 0.0004592184 0.0004017189 0.0003207022
## Cumulative Proportion  0.9967021928 0.9971614113 0.9975631302 0.9978838323
##                            Comp.54      Comp.55    Comp.56      Comp.57
## Standard deviation     0.1397441495 0.1323113766 0.12827563 0.1226679198
## Proportion of Variance 0.0002871828 0.0002574456 0.00024198 0.0002212856
## Cumulative Proportion  0.9981710151 0.9984284607 0.99867044 0.9988917262
##                            Comp.58     Comp.59      Comp.60      Comp.61
## Standard deviation     0.1164666187 0.114429669 0.1074384999 0.0937152432
## Proportion of Variance 0.0001994775 0.000192561 0.0001697505 0.0001291551
## Cumulative Proportion  0.9990912038 0.999283765 0.9994535152 0.9995826703
##                            Comp.62      Comp.63      Comp.64      Comp.65
## Standard deviation     8.198091e-02 7.927978e-02 6.993775e-02 6.800118e-02
## Proportion of Variance 9.883631e-05 9.243064e-05 7.193072e-05 6.800235e-05
## Cumulative Proportion  9.996815e-01 9.997739e-01 9.998459e-01 9.999139e-01
##                            Comp.66      Comp.67      Comp.68
## Standard deviation     5.332495e-02 4.251859e-02 3.471937e-02
## Proportion of Variance 4.181692e-05 2.658574e-05 1.772698e-05
## Cumulative Proportion  9.999557e-01 9.999823e-01 1.000000e+00
```

```
#print the scree plot
screeplot(chemical.pc, type = "lines")
```



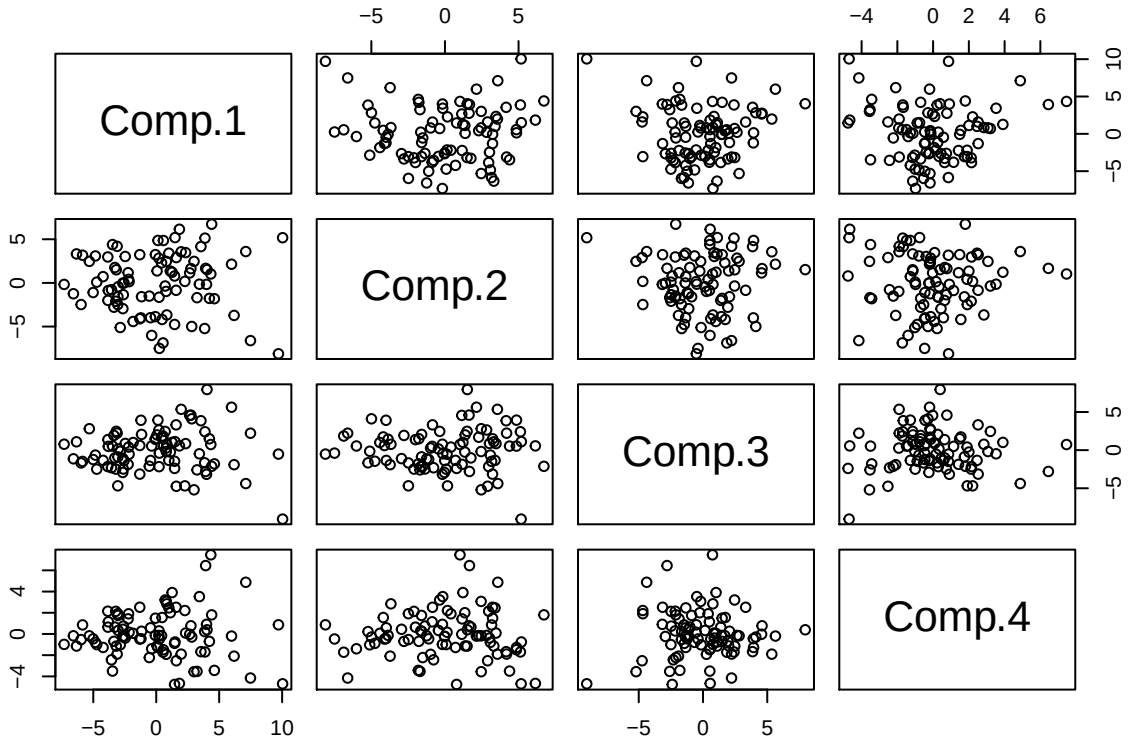

```
#although from the summary of the pca analysis, we
#can see that there are many principal components
#that have a fairly large variance, we just analyze
#that most 4 important PC that can see from the scree plot.
chemical.pc$scores[,1:4]
```

```
##            Comp.1      Comp.2     Comp.3      Comp.4
##  [1,]  7.48675471 -6.61001401  2.2211597 -4.15531088
##  [2,] -4.72489936  0.08487628 -2.4959658 -1.00242322
##  [3,]  2.65767949  1.14332549  4.5601552 -0.02659537
##  [4,] -0.06624913  3.23841162  1.4897161  1.48781487
##  [5,] -2.15612379  0.14191231 -0.5363271  0.16359183
##  [6,] -3.04368786 -1.59211217 -0.8831364  1.75098388
##  [7,]  3.87282367  5.12746856  2.4123758 -1.69595326
##  [8,]  3.23950044 -1.69383475 -2.5890152 -3.52947588
##  [9,] -7.31674564 -0.17465085  0.7762194 -0.98180903
## [10,]  0.06625574  1.37370984  1.9993011 -1.12849306
## [11,] -5.01153809 -1.10463917 -1.2292787 -0.46127469
## [12,]  3.55184234  4.15811320  3.8588188 -1.67003350
## [13,]  0.56764094  4.85464233  0.5318961 -1.48526581
## [14,] -1.21252628 -4.00483386  3.9185178 -0.44496811
## [15,] -0.36211967 -6.01479417  0.5453262 -1.39669059
## [16,] -3.83627869 -2.00464236  1.3804057  0.64223750
## [17,] -1.30487182 -4.10915307  0.6064853 -0.48007370
## [18,] -4.83397879  3.08649936 -0.6514748 -0.79027959
## [19,]  0.78819617 -0.97635099  0.2688603 -1.91877752
## [20,]  0.22388414 -1.64430797 -2.1679880  0.28929864
## [21,] -2.19870499  0.40562572  0.4021639  1.90223832
## [22,]  1.12194246  1.35002005 -0.2517700  2.00335691
## [23,] -5.85196123  3.16862840 -1.4796822  0.85512329
## [24,]  1.57935877 -0.84600423  3.3191337  2.51765894
## [25,]  1.00216499  3.40187904 -1.3389676  2.45938517
```

```
## [26,] -1.30593631  3.23111548 -3.1451787  2.52200337
## [27,] -1.16561826 -1.58320512  2.0605399  0.25896799
## [28,] -3.10452793  4.18742598  2.2283263 -1.89510489
## [29,]  1.98027121  3.56378830  5.3544061 -1.90052630
## [30,] -2.55704192 -0.01054469 -1.3803961 -0.22364567
## [31,]  4.61811561 -1.80892735 -1.8007156 -3.43509573
## [32,] -0.55029609 -1.50949557 -2.1092211 -2.22492394
## [33,] -3.47188089  4.38845140  0.5292045 -3.50106692
## [34,]  5.98212132  2.14061303  5.6351179 -0.19806902
## [35,] -3.28553810  1.76269847 -1.0221959  0.64769802
## [36,] -2.46431920  3.02574905  0.8775827 -0.38469811
## [37,]  1.45809774  0.79175469 -2.3968620 -4.75764712
## [38,] -3.17324384  1.50305736  2.4724063 -1.14632034
## [39,] -5.31437098  2.46355894  2.8106917 -0.17456223
## [40,]  0.42139044  2.29352526  1.6966125 -0.15708526
## [41,]  4.22175458 -1.77272893  1.4908759 -0.69571945
## [42,] -2.23324070  1.16579223 -1.8658880  1.42216265
## [43,]  0.84026327 -3.67047410  0.7692823  2.84156207
## [44,] -3.05329189 -2.48279659 -4.6809224  1.91101405
## [45,]  3.42756721 -0.14933655 -0.4814896  3.52444589
## [46,] -3.21591278 -2.10052925 -2.4867579  2.13172749
## [47,] -2.61404965 -1.35848644 -2.4740002  0.74325989
## [48,]  1.26224869  1.22083524  1.0161121  3.90563163
## [49,]  4.33609840  1.02175575  0.7385160  7.46546756
## [50,] -2.65890924 -0.02944569 -2.9289533 -0.23459302
## [51,]  0.71063104 -0.35719095  2.4566727  3.20614215
## [52,] -6.33360925  3.31869563  1.0933665 -1.14122801
## [53,] -4.22053038  0.71630484 -1.3414706 -1.27292698
## [54,] -3.80685900 -0.87089324  0.4966586  1.14562064
## [55,] -3.01867223 -0.44148870  0.1574801  0.33080457
## [56,]  2.82121683 -4.99289756  4.1088264 -0.28852990
## [57,]  1.49452439  5.19013981  1.1158649 -0.75593935
## [58,] -3.36071166 -2.78506482  1.9309750 -0.69676760
## [59,]  0.24926064 -7.50275335 -0.3884372 -0.47300164
## [60,] -6.57764438 -1.24442418 -1.1359633 -0.17238628
## [61,]  4.03264908  1.51957041  7.9550588  0.39467394
## [62,]  2.27443931  0.17218163 -4.6862173  2.21082748
## [63,]  0.46415490 -4.18960571  1.7336645  1.55402348
## [64,]  0.13616466  4.86136925  3.9363526 -1.27364772
## [65,] -2.85830844 -5.10964988 -0.1476371 -1.02221971
## [66,]  2.75644350  1.63839007  4.5951591  0.77184165
## [67,]  0.20967049  2.83147691  2.6463832 -0.13677230
## [68,] -0.07191477 -3.86375541  1.4269929 -0.13448291
## [69,]  6.18443622 -3.72414674 -1.9076423 -2.09043893
## [70,] -1.83459271 -4.42084725  0.9976835  0.06090390
## [71,]  3.99584970 -0.16984430 -3.1665361  0.90925671
## [72,]  2.33396979  3.46811867  0.8558419  0.08575916
## [73,] -3.55507309 -0.77278630 -2.3097211 -2.43775723
## [74,]  1.45806139 -4.76099427 -1.4540739 -0.88471603
## [75,] -0.45367800 -3.98635985 -1.3503161  0.57751886
## [76,]  3.00400323  2.46406249 -5.2056540 -3.55752896
## [77,] -5.96837349 -2.47999036 -1.6784331 -0.52918135
## [78,]  9.70753864 -8.12011278 -0.5173177  0.86450200
## [79,]  4.40441454  6.72337557 -2.1228870  1.80305026
```

```
## [80,]  7.11430342  3.58988685 -4.3852709  4.87752393
## [81,] 10.03856348  5.17550158 -9.0474692 -4.69903023
## [82,]  1.84058630  6.14289659  0.5589052 -4.66386235
## [83,] -2.64784837 -2.95058987 -1.1084465  0.44274109
## [84,]  3.92627272  1.66526335 -2.7989515  6.44997099
## [85,]  0.80014395  2.45719669 -0.1824441  3.06984886
## [86,] -3.84750048  2.96930912 -2.1562369  2.14881369
## [87,]  3.84791477 -5.22668922 -1.6621436  0.22505726
## [88,]  0.53863825 -6.87686976  1.8532607 -1.72130855
## [89,]  1.59335579  2.89928953 -4.7399002 -2.52630230
```

```
pairs(chemical.pc$scores[,1:4])
```



sub-problem 4

We first interpret the flavor data and then interpret the chemical data.

For the flavor data, there are 8 component(except the variety variable) in total. After doing the PC analysis and we can see that the first 5 PC almost have all the portion of the total variance, and from the scree plot we find the 'elbow' at the third principal component, so the first 2 or 3 PC's are the most important among the 8.

From the scatter plots between 2 of the first 3 PC scores, because the PC analysis use the correlation matrix and we cna use the scatter plots to determine the outliers. For each scatter plot we can set a value and compute the statistical distance of 2 points and then eleminate the points that the statistical distance is too far away from the original point.

Then we interpret the chemical data.

From the summary of the PC analysis, we can see that unlike the flavor data, the chemical data cannot be discribed or represented by a very few components. This is due to the number of the total number of the

PC's –68 to some extent. From the summary we can find that in order to get the 98% of the variance of the data, we should use the first 36 PC's.

Although there are many PC's, we can also see the importance of the different PC's from the scree plot of the scree plot. Although we should use a fairly large nubmer of the PC's to get a high variance of the original data, each of the PC do not have so much variance. The variace of the first 4 PC's decrease very fast and then began to decrease much smoothly.

As for the scatter plot of the PC scores, we can also use the plots to do the outlier detection.