Math3772/5772 Practical

By

Kevin Timothy Muller

A project report for

Multivariate and Cluster Analysis

Dataset : Nutrients

Date : 30th November, 2023

OVERALL THOUGHTS

The list of conclusions I have made about the dataset post-analysis are as follows :

1. The average woman is most deficient in Calcium

2. The average woman is most abundant in Protein

3. The nutrient with the highest variance is Vitamin A

4. A woman who is very deficient in a certain nutrient, generally implies that she is very deficient in other nutrients as well.

5. The nutrient protein has a somewhat high positive correlation with the other nutrients; meaning; low protein consumption implies low consumption of other nutrients and a high protein consumption implies high consumption of other nutrients as well.

6. Vitamins A and C must come from different food sources as there is hardly any correlation between the two.

7. The current average consumption of nutrients are not the same as the RDA values (Null Hypothesis)

8. Extra steps need to be taken to ensure that the consumption of Calcium, Iron and Protein move towards the RDA values as the RDA of those nutrients do not lie in their Simultaneous Confidence Intervals.

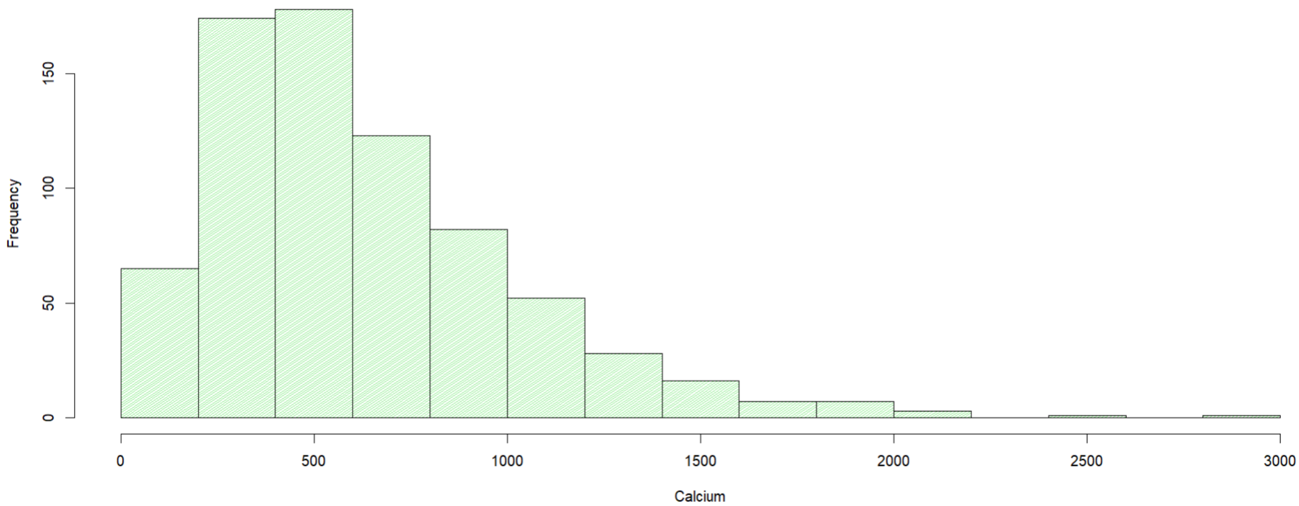
9. The current consumption of nutrients of each woman on average is not balanced (Profile Hypothesis)

10. Women who have a “low” intake of Vitamin C have a “low” intake of other nutrients as a whole and that women with a “high” intake of Vitamin C have a “high” intake of other nutrients as a whole.

QUESTION 1 :

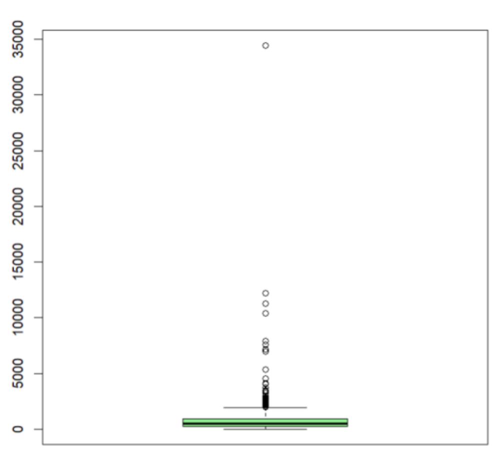
Upon initial inspection of the summary of the nutrients dataset, there is no trace of null or missing values. This allows us to confidently enter into exploratory analysis without any cause for concern.

Through this dataset, we are exploring the nutrient intake of 737 women of ages 25 to 50. More specifically, their intake of 5 nutrients namely : "Calcium", "Iron", "Protein", "VitaminA" and "VitaminC". There does exist certain women who seem to have no intake of a certain nutrient at all, except for **Calcium** with a minimum value of 7.44 mg ; which makes sense, considering milk is a staple drink of everyday life and a primary source of Calcium. However, it is also observed that Calcium is the nutrient which is consumed the least, in comparison to the RDA standards.

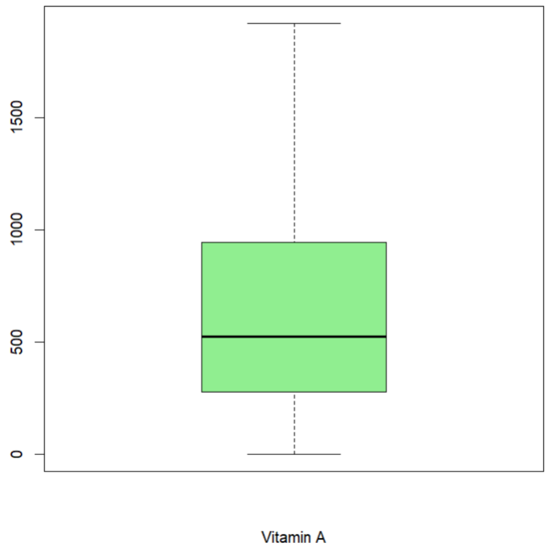


The Calcium intake of most women hovers around 250 to 750 mg, when it should be 1000mg.

Vitamin A is also a nutrient of interest as while it has the highest inter quartile range (numerically) out of all the nutrients, it also possesses the biggest outliers out of all the nutrients, making its variance more inflated than it already needs to be! This can be seen in the boxplot below :

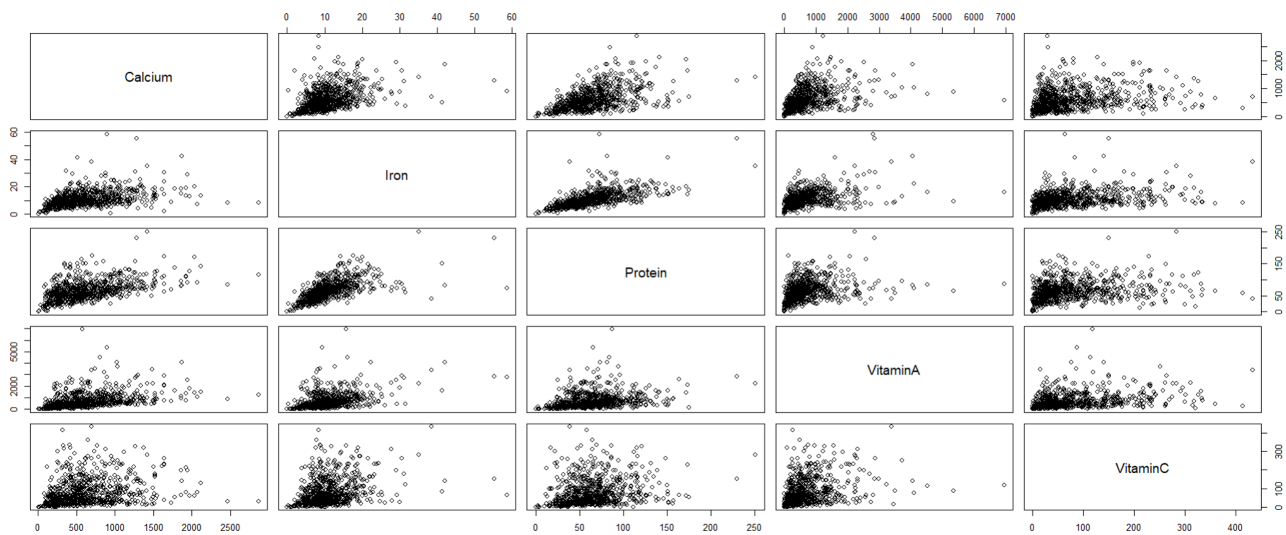


Without the existence of the many outliers in Vitamin A, we can begin to get a grasp of the majority of the data.



As for the distribution of the data, all nutrients are skewed right similar to that of Calcium and even worse in some cases like Vitamin A and Vitamin C due to the presence of high value outliers.

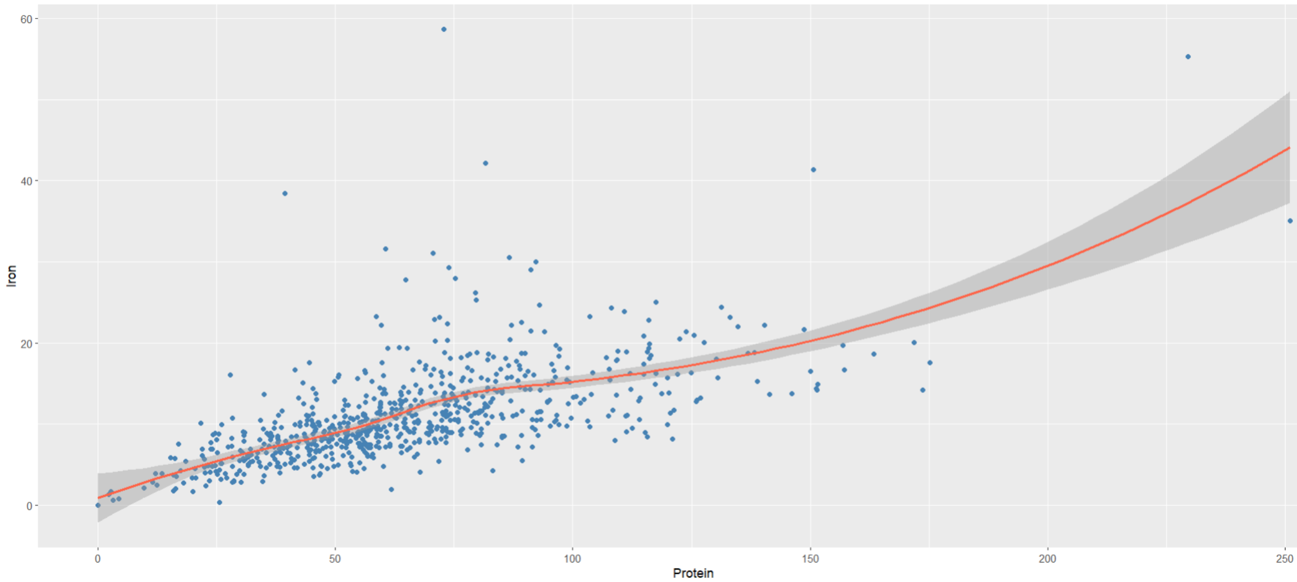
Shown below is the pairwise plot of the data.



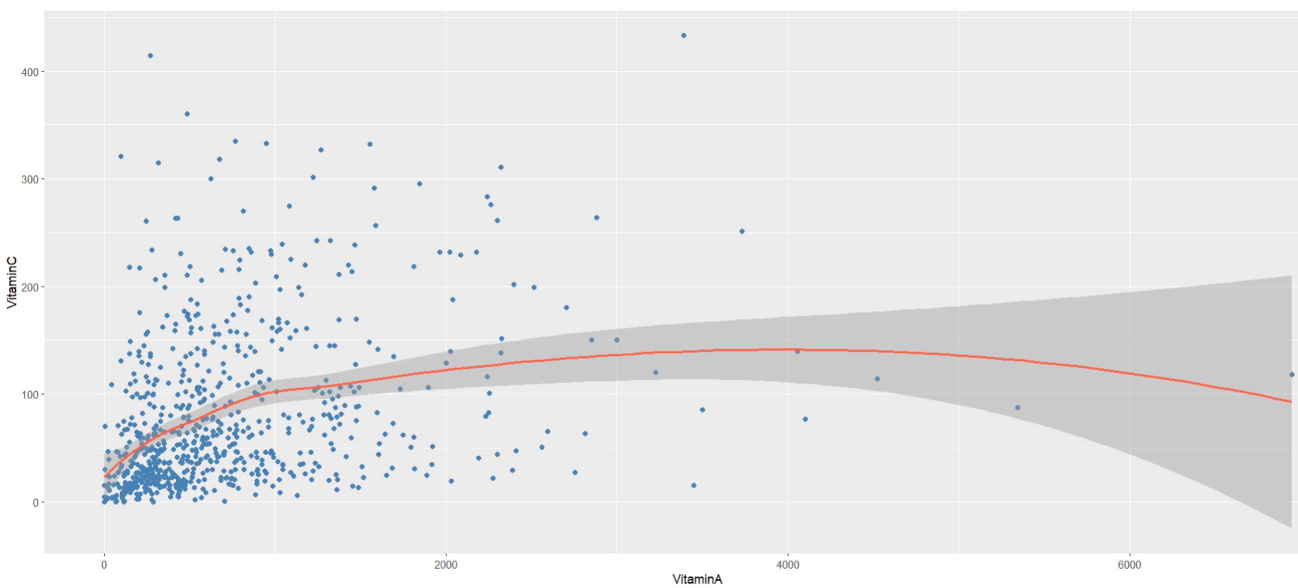
From the pairwise plot, we observe that all nutrients have a mostly positive correlation with one another, especially at the lower end of the X axis. We can hence infer that women who are **generally deficient in a certain nutrient, are very likely to be deficient in all other nutrients as well.** As the nutrient consumption increases, there is a general increase in variance of all nutrient consumption; much like a flower boquet. Out of all the nutrients, **Protein** seems to have the most positive correlation with other nutrients, implying that women with a high intake of protein, will also have a high intake of other nutrients. Another interesting observation is that the correlation between **Vitamin A and C** is mostly non-existent, especially in higher values of Vitamin C intake, which is surprising to say the least.

Also, do note that for the sake of clarity in visualization, rows with Vitamin A outlier values greater than 7000 have been removed in the pairwise plot.

Protein vs Iron (High positive Correlation) :



VitaminA vs VitaminC (Negligible Correlation) :



QUESTION 2 :

Upon applying the Hotelling's T square test on the data, we get a p value of 2.988651e-191 which is much lesser than 0.5 percent. Hence, we reject the null hypothesis and conclude that there is no similarity between the nutrient means of the dataset and the RDA values; which in turn implies that women are not getting the recommended level of nutrients on average.

Differences between the average nutrient intake of women and the RDA values suggest that on average, women are Calcium and Iron deficient, but consume an abundance of Protein, Vitamin A and Vitamin C.

```
> xmm = xbar - mu0
> xmm
    Calcium        Iron     Protein    VitaminA    VitaminC
-375.950746   -3.870100    5.803441   39.635346    3.928446
```

The simultaneous confidence intervals (at 95% confidence level) of each of the nutrients are as follows :
Calcium : (575.0912, 673.0073)
Iron : (10.39244, 11.86735)
Protein : (62.03547, 69.57141)
Vitamin A : (638.3278, 1040.9429)
Vitamin C : (69.85901, 87.99788)

These values suggest with 95% confidence that the mean of future random samplings will lie in this range. In the case of Calcium, Iron and Protein, their Simultaneous Confidence Intervals lie outside the RDA, which means that it is highly unlikely that the mean of the nutrients will ever come close to the RDA values. However, the opposite is the case for Vitamin A and Vitamin C, meaning that the future means of these nutrients might progress towards the RDA values.

The assumptions made in order to accept the validity of the T square test are that of Independence between the nutrients and that the overall trend of each of the nutrient data is similar to a normal distribution. Although there is high correlation between some of the nutrients, I do not think this impacts the test as this is just a feature of certain nutrients and not an error in the data procurement process. However, none of the nutrients possess a normal distribution due to the presence of outliers as discussed in the 1st Question. For this reason, we cannot consider this a valid test on the current dataset.

By removing certain outliers from the dataset and normalizing the values, we can convert this data into a normal distribution in order to conduct a valid test. But whether the results of such a test would be beneficial for the United States Department of Agriculture is not a guarantee.

QUESTION 3 :

In order to test the profile hypothesis, we first divide each of the nutrient columns by their respective RDA values. Assume a 5*4 matrix A such that matrix multiplying our current dataset with the transpose of this matrix, it computes the differences between each of the column means :

	[,1]	[,2]	[,3]	[,4]	[,5]
[1,]	1	-1	0	0	0
[2,]	0	1	-1	0	0
[3,]	0	0	1	-1	0
[4,]	0	0	0	1	-1

That is, “Calcium – Iron”, “Iron – Protein”, “Protein – VitaminA” and “VitaminA – VitaminC”. This is considered to be our xbar and mu0 for this case is a zero vector.

Upon applying the Hotelling’s T square test on the data, we get a p value of 8.89301e-138 which is much lesser than 0.5 percent. Hence, we reject the profile hypothesis and conclude that there is no similarity between the ratios of nutrient means of the dataset and RDA values. This hypothesis is definitely one of nutritional interest as it aims at proving whether the women are consuming a balanced diet. Since we have rejected the hypothesis, we imply that the women are consuming nutrients at different proportions.

Although we learned from Question 2 that the women’s nutrient intake is dissimilar to that of the proposed RDA values, it does not necessarily mean that they are unhealthy as each woman will only require the amount of nutrients as per her weight. In this case, the profile hypothesis gives us a much clearer understanding of whether the women are healthy or not as women of all weights should consume a balanced diet.

The simultaneous confidence intervals (at 95% confidence level) of each of the nutrients are as follows :
Calcium - Iron : (-0.16784275, -0.06804535)
Iron -Protein : (-0.4010850, -0.3083764)
Protein – Vitamin A : (-0.1842869, 0.2786466)
Vitamin A – Vitamin C : (-0.2418458, 0.2361756)

Variables “Protein – Vitamin A” and “Vitamin A – Vitamin C” have their standard confidence intervals lie in a range which includes zero meaning that there is a chance that the future procurement of data could result in their ratios with their respective RDAs being the same. This is however, not the case for “Calcium – Iron” and “Iron -Protein”, meaning that this hypothesis will almost never (95% confidence) be true even in future samplings.

QUESTION 4 :

Rather than providing a single number as the RDA of a nutrient, I think it would be more helpful to provide the RDA as a range of values based on the woman's weight. This way, the non average woman in terms of weight, does not need to strive to consume the recommend amount of nutrients of a woman of average weight.

Another suggestion would be that, in order to promote the consumption of a balanced diet, it would be worthwhile to advertise foods containing nutrients which the average woman is lacking; namely Calcium and Iron. It would also be helpful to promote staple foods which are a balanced diet on it's own such as milk. Provision of a list of foods rich in just particular nutrient(s) would be helpful for women who want to make up for their lost nutrient consumption.

In these ways, we can reduce the number of outliers and the women can lead healthy lives.

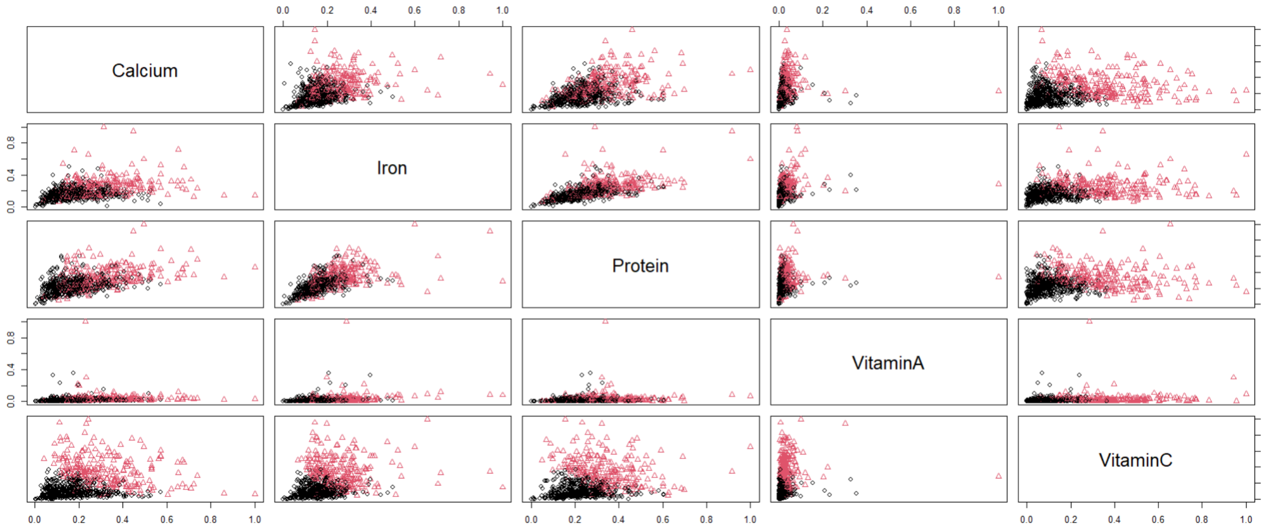
QUESTION 5 :

Since K means clustering makes use of Euclidian distances to find its clusters, I have normalized the dataset by dividing each nutrient by it's greatest value so that all nutrients approximately have a standard deviation of 1 while still maintaining the relative distances between the values. Now, the high value outliers which would usually skew the Euclidian distance heavily in it's favor is no longer a problem. Hence, with the use of this modified dataset, we can come up with more meaningful clusters than what we would get without the normalization of data.

Upon completing the clustering process with k = 2, the K means algorithm has divided the dataset into two clusters of 501 items in the 1st cluster and 236 items in the 2nd cluster with the 1st cluster generally having lower means of nutrient value and the 2nd cluster having higher means of nutrient value.

	Calcium	Iron	Protein	Vitamin A	Vitamin C
[1,]	0.1698001	0.157179	0.2253851	0.01800273	0.09980453
[2,]	0.3194133	0.258769	0.3402055	0.03792978	0.35692902

The next page shows the pairwise plot of each of the nutrients against each other, with black diamonds denoting items from the first cluster and red triangles denoting the second.

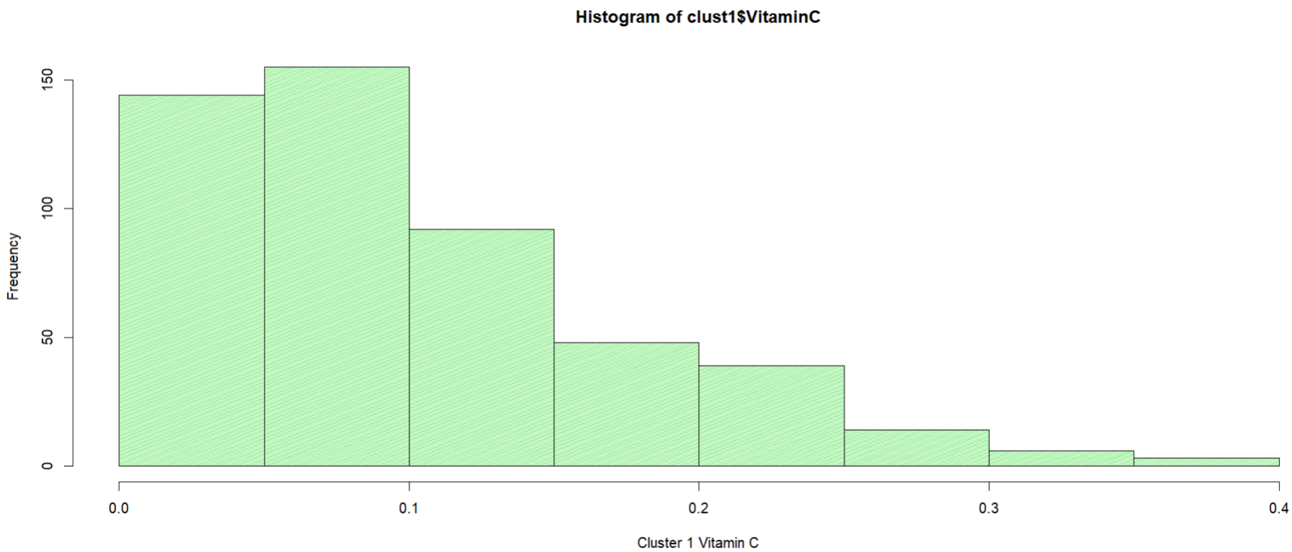


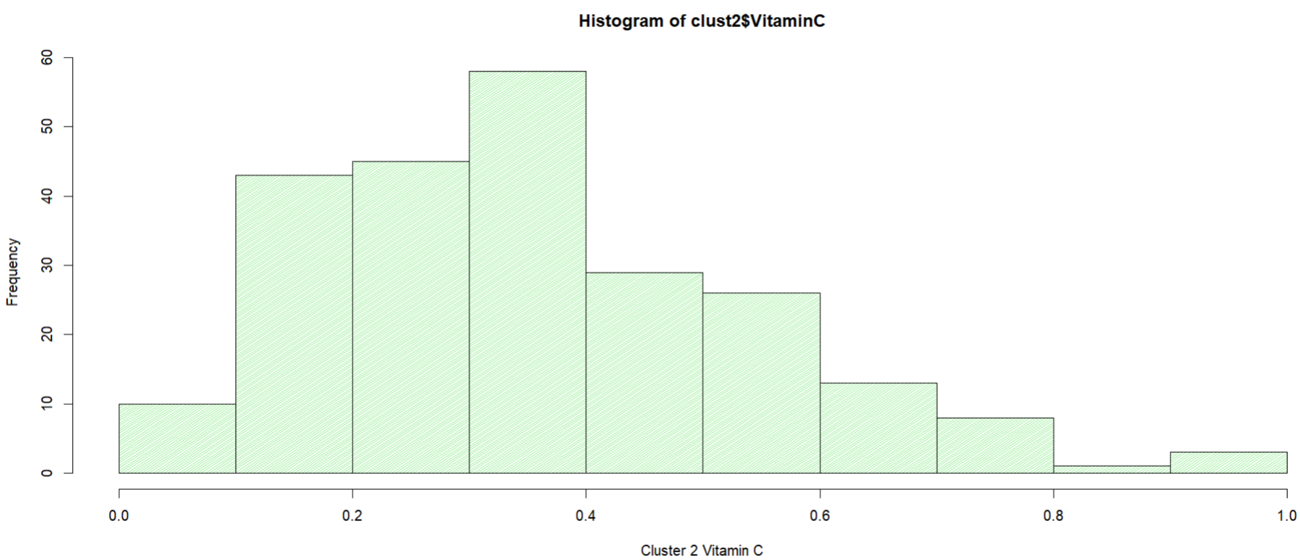
The clustering shows a good split of data around the middle points of each of the graphs. This can be further improved upon removing certain high value outliers such as that of Vitamin A which would still skew the centroid of its cluster somewhat towards itself, but clusters formed this way might lose integrity in real-life applications such as the one we are dealing with now.

It is also worth noting that since we are dealing with 5 dimensional data, there is some overlap between the clusters. However, out of all the nutrients, Vitamin C seems to be least affected by such a problem; that is, it's clusters are mostly well separated. This means that Vitamin C has a good density of data across it's minimum and maximum along with decent correlation with the other nutrients despite having a high variance. The same could also be said for Calcium, however, it's plot against Iron is quite messy. While Vitamin A has the highest variance due to its outliers, most of it's data lies below its 35th percentile, hence it does not have much say in the clustering.

We can hence infer that women who have a low intake of Vitamin C have a low intake of other nutrients as a whole and that women with a high intake of Vitamin C have a high intake of other nutrients as a whole.

Histogram plots showcasing the clean split of Vitamin C over the clusters are showcased below :
(next page)


Histogram of clust1$VitaminC
Frequency
0
50
100
150
0.0
0.1
0.2
0.3
0.4
Cluster 1 Vitamin C


Histogram of clust2$VitaminC
Frequency
0
10
20
30
40
50
60
0.0
0.2
0.4
0.6
0.8
1.0
Cluster 2 Vitamin C

APPENDIX :

```
nutrients=read.csv("http://www1.maths.leeds.ac.uk/~john/3772/nutrients.csv")
attach(nutrients)


req = nutrients[,c("Calcium", "Iron", "Protein", "VitaminA", "VitaminC")]
rec=data.frame(Calcium        = 1000,Iron =  15,Protein        = 60 ,VitaminA =      800 ,VitaminC =
        75)
```

#Question 1 :

```
summary(req)

plot (req$Calcium,req$Protein,xlab = "Calcium",ylab ="Protein" )
#install.packages("ggplot2")
library(ggplot2)

hist(req$Calcium,xlab = "Calcium", col = "lightgreen", border = 129,density = 100)
hist(req$Iron)
hist(req$Protein)
hist(req$VitaminA)
hist(req$VitaminC)
boxplot(req$Calcium)
boxplot(req$Iron)
boxplot(req$Protein)
boxplot(req$VitaminA,col = "lightgreen", border = 129,density = 100,outline = TRUE,xlab =
"Vitamin A")
boxplot(req$VitaminA,col = "lightgreen", border = 129,density = 100,outline = FALSE,xlab =
"Vitamin A")
boxplot(req$VitaminC)
var(req)
cov(req)
cor(req)
forvis = req[req["VitaminA"]<7000,]
forvis
summary(forvis)
pairs(forvis)

ggplot(req, aes(x = Protein, y = Iron)) +
  geom_point(color= "steelblue") +
  geom_smooth(color = "tomato")

ggplot(forvis, aes(x = VitaminA, y = VitaminC)) +
  geom_point(color= "steelblue") +
```

```
  geom_smooth(color = "tomato")

heatmap(cor(req))
cor(req)
```

#Outputs :

```
> var(req)
             Calcium       Iron   Protein     VitaminA   VitaminC
Calcium  157829.4439  940.08944 6075.8163   102411.127  6701.6160
Iron        940.0894   35.81054  114.0580     2383.153   137.6720
Protein    6075.8163  114.05803  934.8769     7330.052   477.1998
VitaminA 102411.1266 2383.15341 7330.0515 2668452.371 22063.2486
VitaminC   6701.6160  137.67199  477.1998    22063.249  5416.2641
> cov(req)
             Calcium       Iron   Protein     VitaminA   VitaminC
Calcium  157829.4439  940.08944 6075.8163   102411.127  6701.6160
Iron        940.0894   35.81054  114.0580     2383.153   137.6720
Protein    6075.8163  114.05803  934.8769     7330.052   477.1998
VitaminA 102411.1266 2383.15341 7330.0515 2668452.371 22063.2486
VitaminC   6701.6160  137.67199  477.1998    22063.249  5416.2641
> cor(req)
           Calcium      Iron   Protein  VitaminA  VitaminC
Calcium  1.0000000 0.3954301 0.5001882 0.1578060 0.2292111
Iron     0.3954301 1.0000000 0.6233662 0.2437905 0.3126009
Protein  0.5001882 0.6233662 1.0000000 0.1467574 0.2120670
VitaminA 0.1578060 0.2437905 0.1467574 1.0000000 0.1835227
VitaminC 0.2292111 0.3126009 0.2120670 0.1835227 1.0000000
```

#Question 2 :

```
#install.packages("DescTools")
library("DescTools")
mu0 = c(1000, 15, 60,800,75)
xbar = colMeans(req)
rbind(xbar, mu0)
xbar/mu0
xbar
S = var(req)
S
R = cor(req)
R
xmm = xbar - mu0
xmm
tsq = 737 * t(xmm) %*% solve(S) %*% xmm
tsq
fstat = tsq * (737-5)/(5*736)
pf(fstat, df1 = 5, df2 = 732, lower.tail=F)
#Which is less than 0.5 percent
#So we reject the null hypothesis
```

#Meaning the women arent taking the right amount of nutrients

```
#SCI
t5 = qf(0.05, df1 = 5, df2 = 732, lower=F)*736*5/732
con = sqrt((1/737)*S[1,1]*t5)
c(xbar[1]-con,xbar[1]+con)
rec[1]
con = sqrt((1/737)*S[2,2]*t5)
c(xbar[2]-con,xbar[2]+con)
rec[2]
con = sqrt((1/737)*S[3,3]*t5)
c(xbar[3]-con,xbar[3]+con)
rec[3]
con = sqrt((1/737)*S[4,4]*t5)
c(xbar[4]-con,xbar[4]+con)
rec[4]
con = sqrt((1/737)*S[5,5]*t5)
c(xbar[5]-con,xbar[5]+con)
rec[5]
```

#Outputs :

```
> S = var(req)
> S
            Calcium       Iron   Protein    VitaminA   VitaminC
Calcium  157829.4439  940.08944 6075.8163  102411.127  6701.6160
Iron        940.0894   35.81054  114.0580    2383.153   137.6720
Protein    6075.8163  114.05803  934.8769    7330.052   477.1998
VitaminA 102411.1266 2383.15341 7330.0515 2668452.371 22063.2486
VitaminC   6701.6160  137.67199  477.1998   22063.249  5416.2641
> R = cor(req)
> R
           Calcium      Iron   Protein  VitaminA  VitaminC
Calcium  1.0000000 0.3954301 0.5001882 0.1578060 0.2292111
Iron     0.3954301 1.0000000 0.6233662 0.2437905 0.3126009
Protein  0.5001882 0.6233662 1.0000000 0.1467574 0.2120670
VitaminA 0.1578060 0.2437905 0.1467574 1.0000000 0.1835227
VitaminC 0.2292111 0.3126009 0.2120670 0.1835227 1.0000000
> xmm = xbar - mu0
> xmm
    Calcium        Iron     Protein    VitaminA    VitaminC
-375.950746   -3.870100    5.803441   39.635346    3.928446
> tsq = 737 * t(xmm) %*% solve(S) %*% xmm
> tsq
         [,1]
[1,] 1758.541
> fstat = tsq * (737-5)/(5*736)
> pf(fstat, df1 = 5, df2 = 732, lower.tail=F)
              [,1]
[1,] 2.988651e-191
```

#Question 3 :

```
means = colMeans(req)
matplot(means,pch = "*",
        xlab = "Nutrient", ylab = "Mean Value")
req3 = mapply("/",req,rec)
req3
colMeans(req3)
A = matrix(c(1,-1,0,0,0,0,1,-1,0,0,0,0,1,-1,0,0,0,0,1,-1),nrow=4,byrow=T)
A
req3 = req3%*%t(A)

as.matrix(means)
t(as.matrix(rec))
ans = as.matrix(means)/t(as.matrix(rec))
ans
mu0 = c(0,0,0,0)
xbar = as.matrix(colMeans(req3),nrow=4)
xbar
S = var(req3)
R = cor(req3)
R
xmm = xbar - mu0
xmm
tsq = 737 * t(xmm) %*% solve(S) %*% xmm
tsq
fstat = tsq * (737-4)/(4*736)
pf(fstat, df1 = 4, df2 = 733, lower.tail=F)
#Which is lesser than 0.5 percent
#So we reject the null hypothesis
#Meaning that everyone consumes a food with different proportions of nutrients

#SCI
t5 = qf(0.05, df1 = 4, df2 = 733, lower=F)*736*4/733
con = sqrt((1/737)*S[1,1]*t5)
c(xbar[1]-con,xbar[1]+con)
mu0[1]
con = sqrt((1/737)*S[2,2]*t5)
c(xbar[2]-con,xbar[2]+con)
mu0[2]
con = sqrt((1/737)*S[3,3]*t5)
c(xbar[3]-con,xbar[3]+con)
mu0[3]
con = sqrt((1/737)*S[4,4]*t5)
c(xbar[4]-con,xbar[4]+con)
mu0[4]
```

#Outputs :

```
> colMeans(req3)
  Calcium      Iron   Protein  VitaminA  VitaminC
0.6240493 0.7419933 1.0967240 1.0495442 1.0523793
> A = matrix(c(1,-1,0,0,0,0,1,-1,0,0,0,0,1,-1,0,0,0,0,1,-1),nrow=4,byrow=T)
> A
     [,1] [,2] [,3] [,4] [,5]
[1,]    1   -1    0    0    0
[2,]    0    1   -1    0    0
[3,]    0    0    1   -1    0
[4,]    0    0    0    1   -1
> req3 = req3%*%t(A)
> as.matrix(means)
              [,1]
Calcium  624.04925
Iron      11.12990
Protein   65.80344
VitaminA 839.63535
VitaminC  78.92845
> t(as.matrix(rec))
         [,1]
Calcium  1000
Iron       15
Protein    60
VitaminA  800
VitaminC   75
> ans = as.matrix(means)/t(as.matrix(rec))
> ans
              [,1]
Calcium  0.6240493
Iron     0.7419933
Protein  1.0967240
VitaminA 1.0495442
VitaminC 1.0523793
```

```
> mu0 = c(0,0,0,0)
> xbar = as.matrix(colMeans(req3),nrow=4)
> xbar
             [,1]
[1,] -0.117944052
[2,] -0.354730710
[3,]  0.047179834
[4,] -0.002835103
> S = var(req3)
> R = cor(req3)
> R
            [,1]        [,2]        [,3]        [,4]
[1,]  1.00000000 -0.39891026  0.05074899 -0.04091942
[2,] -0.39891026  1.00000000 -0.21656195  0.03465978
[3,]  0.05074899 -0.21656195  1.00000000 -0.88185997
[4,] -0.04091942  0.03465978 -0.88185997  1.00000000
> xmm = xbar - mu0
> xmm
             [,1]
[1,] -0.117944052
[2,] -0.354730710
[3,]  0.047179834
[4,] -0.002835103
> tsq = 737 * t(xmm) %*% solve(S) %*% xmm
> tsq
         [,1]
[1,] 1030.795
> fstat = tsq * (737-4)/(4*736)
> pf(fstat, df1 = 4, df2 = 733, lower.tail=F)
             [,1]
[1,] 8.89301e-138
```

#Question 5 :

```
# This file defines the local function clustering
# for use in Math 5772, Exercise Sheet 2
# updated 29 October 2015 to use multiple starts in kmeans

library(mclust)
cluster.descriptions=function(x,x.cl) {
  #  x (n by p) = data
  #  x.cl n-vector of cluster labels
  k=max(x.cl) # assumes cluster labels range from 1:k
  p=ncol(x)
  for(i in 1:k){
    cat("Cluster", i, "consists of\n")
    print(names(x.cl[x.cl==i]))
  }
  means=matrix(0,k,p)
  for(i in 1:k) means[i,]=apply(x[x.cl==i,,drop=FALSE],2,mean)
  cat("cluster means: rows=clusters; columns=variables\n")
```

```
  print(means)
  means
}

clustering=function(x,method,k=0) {
  # general clustering function
  x=as.matrix(x)
  if(k==0 &method !="mixture") return(cat("clustering failed; needs a value for k
\n"))
  if(method=="single"| method=="complete" | method=="average") {
    hc=hclust(dist(x),method=method); x.cl=cutree(hc,k)
  }
  if(method=="kmeans") x.cl=kmeans(x,k,nstart=100)$cluster
  if(method=="mixture") {
    x.mix=Mclust(x)
    x.cl=x.mix$classification
    print.Mclust(x.mix)
  }
  pairs(x,col=x.cl,pch=x.cl)
  means=cluster.descriptions(x,x.cl)
  k=max(x.cl); count=rep(0,k)
  for(i in 1:k) count[i]=sum(x.cl==i)
  list(labels=x.cl, means=means,count=count)
}

clust = data.frame("Calcium" = req$Calcium/max(req$Calcium),"Iron" =
req$Iron/max(req$Iron),"Protein" = req$Protein/max(req$Protein),"VitaminA" =
req$VitaminA/max(req$VitaminA), "VitaminC" = req$VitaminC/max(req$VitaminC))
done = clustering(clust,"kmeans",k=2)
clust1 = clust[done$labels==1,]
clust2 = clust[done$labels==2,]
clust1
clust2

cor(clust1)
cor(clust2)

hist(clust1$VitaminC,xlab = "Cluster 1 Vitamin C", col = "lightgreen", border = 129,density = 100)
hist(clust2$VitaminC,xlab = "Cluster 2 Vitamin C", col = "lightgreen", border = 129,density = 100)
```

#Outputs :

```
> clust = data.frame("Calcium" = req$Calcium/max(req$Cal
max(req$Protein),"VitaminA" = req$VitaminA/max(req$Vitam
> done = clustering(clust,"kmeans",k=2)
Cluster 1 consists of
NULL
Cluster 2 consists of
NULL
cluster means: rows=clusters; columns=variables
          [,1]     [,2]      [,3]       [,4]       [,5]
[1,] 0.1698001 0.157179 0.2253851 0.01800273 0.09980453
[2,] 0.3194133 0.258769 0.3402055 0.03792978 0.35692902
> cor(clust1)
              Calcium      Iron    Protein  VitaminA     VitaminC
Calcium   1.000000000 0.2247388 0.33034187 0.1446986 -0.004915414
Iron      0.224738811 1.0000000 0.60595664 0.2242191  0.121181360
Protein   0.330341871 0.6059566 1.00000000 0.1173967  0.043727486
VitaminA  0.144698604 0.2242191 0.11739666 1.0000000  0.117928192
VitaminC -0.004915414 0.1211814 0.04372749 0.1179282  1.000000000
> cor(clust2)
             Calcium        Iron    Protein   VitaminA    VitaminC
Calcium   1.00000000  0.20131868  0.3865359 0.03049623 -0.32783758
Iron      0.20131868  1.00000000  0.4695636 0.15532057 -0.09497055
Protein   0.38653589  0.46956360  1.0000000 0.04307580 -0.28327407
VitaminA  0.03049623  0.15532057  0.0430758 1.00000000  0.04613452
VitaminC -0.32783758 -0.09497055 -0.2832741 0.04613452  1.00000000
```