不同品牌类型麦片营养成分的探索

2018110760 李晨茜

一、背景与意义

燕麦片（英语：Oatmeal），又称麦片或麦皮，是由燕麦做成的食品。由于麦片食品的制作过程简单，而且省时，有些种类的麦片只要经过开水冲泡就可以食用，所以受到了许多人欢迎。燕麦在西方饮食中主要用来做麦片粥，或和果汁香料等混合一起焙制成一种干燥食品，泡在牛奶中做早餐。

很多不同的厂家生产了许多不同口味和规格的麦片，本文希望对不同品牌类型的麦片的营养成分进行探索，以探究什么样的麦片具有最高的营养成分，对人体最有益，从而为消费者提供购买参考和建议。

二、数据预处理

1.数据集基本信息

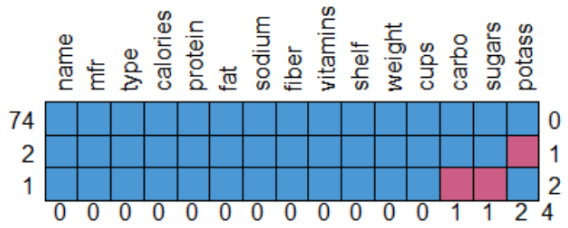
本文采用的数据集中共含有 77 个观测（代表着 77 种不同品牌类型的麦片），和 15 个变量（代表着对每种麦片的 15 个指标）。变量包含 1 个指示麦片种类的变量（name）和 4 个分类变量、10 个连续型变量（有 9 个与营养成分有关）。关于变量的具体信息整理如下表所示：

变量名	变量中文名	变量类型	变量解释	备注
name	商品名称			标识数据观测
mfr	厂商	分类变量	麦片制造商	
type	温度类型	分类变量	冷/热	
calories	卡路里	连续型变量	卡路里（每份）	
protein	蛋白质	连续型变量	蛋白质含量（克）	
fat	脂肪	连续型变量	脂肪含量（克）	
sodium	钠	连续型变量	钠含量（毫克）	
fiber	纤维	连续型变量	纤维含量（克）	
carbo	碳水化合物	连续型变量	碳水化合物含量（克）	
sugars	糖	连续型变量	糖含量（克）	
potass	钾	连续型变量	钾含量（毫克）	
vitamins	维他命	连续型变量	维他命和矿物质	0,25,100 表示占 FDA 推荐的含量的百分比
shelf	货架	分类变量	所在货架的层数	数值越低越接近地面

weight	重量	连续型变量	每份重量（盎司）	
cups	杯数	分类变量	每份含的杯数	

2.缺失值和异常值处理

经检查数据集中的缺失值，得到结果如下图所示：

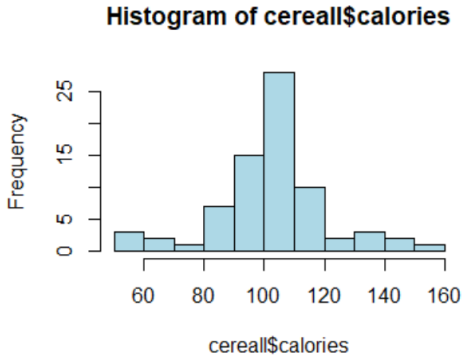


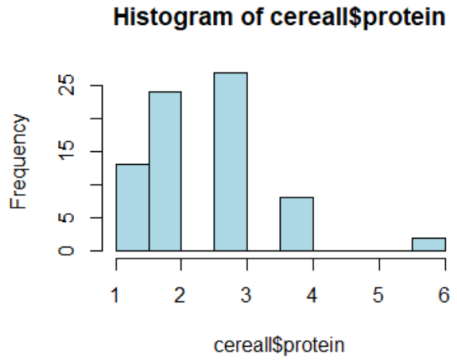
可以发现有 1 个观测同时缺少 carbo 和 sugars 两个变量的值，有两个观测缺少 potass 的值，共有 3 个观测中存在缺失值。由于缺失观测只占总观测数的 3.8%，因此采用个案删除法，将含有缺失值的观测从数据集中删除。

经检查，该数据集中不存在异常值。

三、数据可视化

为了初步观察不同麦片中不同营养成分含量的分布，画出各个营养成分的

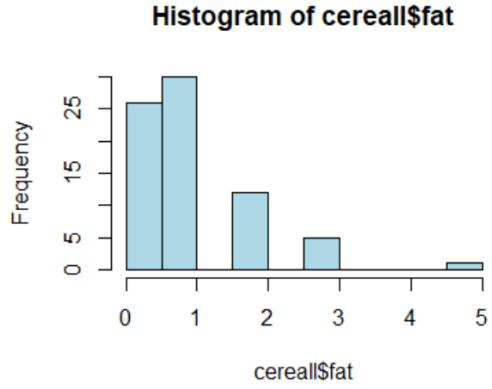


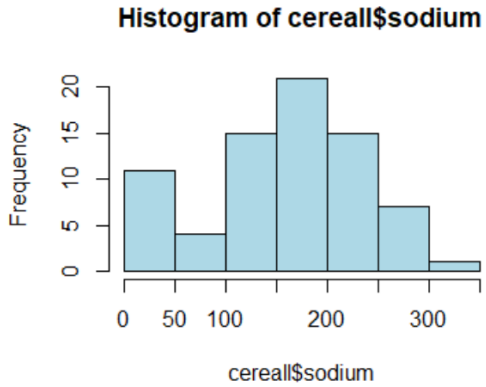


直方图如下：

可以发现大多数麦片的卡路里含量在 80-120 卡/份，表明大多数麦片食品是低卡的。猜测卡路里低至 60 卡/份的为专为减脂瘦身人群设计的麦片，而卡路里含量高至 160 卡/份的可能是因为麦片中为了提升口感而加入了坚果、果干等配料。

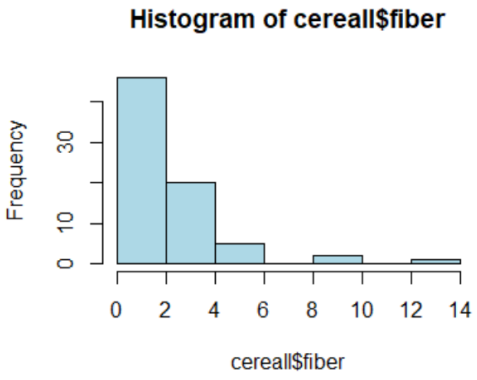
从蛋白质含量来说，大多数麦片的蛋白质含量在 2 克左右，只有极少数的麦片蛋白质含量高达 6 克。猜测可能加入了一些高蛋白的配料。

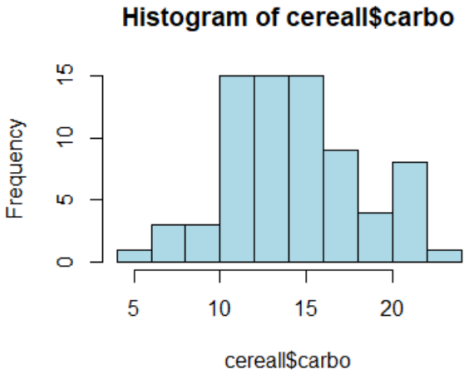




从脂肪含量来看，大多数麦片的脂肪含量较低。发现麦片确实是低脂低卡的健康食品。

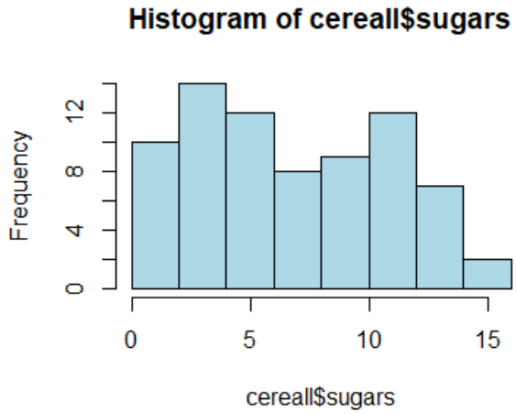
从钠含量来看，因为钠与钙在肾小管内的重吸收过程发生竞争，故钠摄入量高时，会相应减少钙的重吸收，而增加尿钙排泄。因尿钙丢失约为钙潴留的50%，故高钠膳食对钙丢失有很大影响。而大多数麦片的钠含量都在 250 毫克以下，中国居民膳食指南建议钠每天摄入不要超过 6g，食品标签中营养素参考值是钠 2000mg。说明麦片在钠含量上也是符合健康标准的。

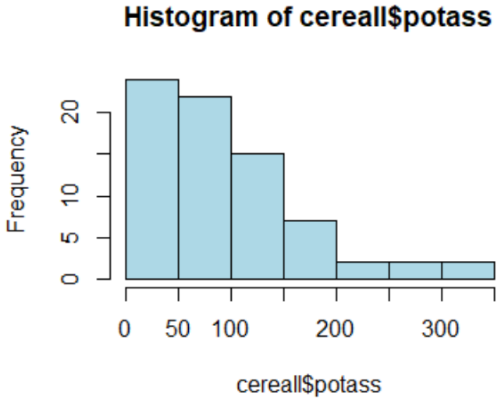




从纤维含量来看，多数麦片产品的纤维含量并不高。而经过查阅资料发现，燕麦片中的纤维含量通常在 5-6g，因此猜测可能由于麦片产品中加入了其他配料从而导致每单位产品的纤维含量低于纯燕麦片的纤维含量。

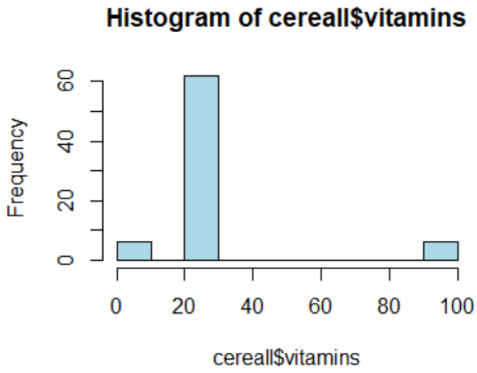
从碳水化合物来看，发现麦片中的碳水化合物含量大多处于 10-15 克之间，含量相对比较低，和上文发现的麦片低热量低脂肪的特征是相符的。





从含糖量来看，不同品牌类型的麦片的含糖量都在 15 克以下，且多数在 10 克以下。说明尽管因为配料不同、制作方法不同等原因导致不同品牌类型的麦片含糖量不同，但他们的含糖量都比较少，说明麦片不仅低脂低卡，并且低糖，是一种健康食品。

从含钾量来看，麦片中的钾含量多数在 150 毫克以下。中国居民膳食指南建议成年男女每日的钠的摄入量为 1875～5625 毫克。麦片是一种含钾量较低的食物，说明虽然麦片低脂低卡低糖，但在日常生活中只吃麦片也是不健康的。



从维生素和矿物质含量来看，绝大多数麦片的维生素和矿物质含量只占 FDA 推荐的量的百分之 25 左右。说明虽然麦片中富含一定量的维生素和矿物质，但相较于人体所需还是远远不够的。在日常饮食中还需要从其他食物中摄取必要的维生素和矿物质。

三、数据处理方法介绍

1. 主成分分析

主成分分析主要用于构造“综合指标”，以期最大程度地区分原始数据。主成分是标准化后的原始变量的线性组合，使这样的线性组合的标准差最大化的组合就是主成分。因为标准差刻画了一组数据与平均数的距离，标准差越大，数据的波动就越大，包含的信息就越多。

主成分分析的主要步骤：

①将原始变量做标准化，并计算其协方差矩阵

②求协方差矩阵的特征值与对应的特征向量

③把特征向量按其对应的特征值从大到小的顺序重新排列，取前 k 列组成矩阵（一般取累计方差占比大于 80%的 k 值）

2. Kmeans 聚类

聚类分析将研究对象根据一些特征指标的信息，把比较相似的研究对象按一定的方式归为同类。

Kmeans 聚类即先确定类别数K。确定之后,选取K个“种子”,然后看每个个体离哪个种子最近就归到哪一类。归类之后原来的种子就被每一个新类的“中心”代替。再重复上述的归类步骤,直到每个个体所属的类别不再变动为止。最终的种子（即最终聚类的中心点）可以用来刻画这一类的特征。

3. WSS 法确定聚类个数

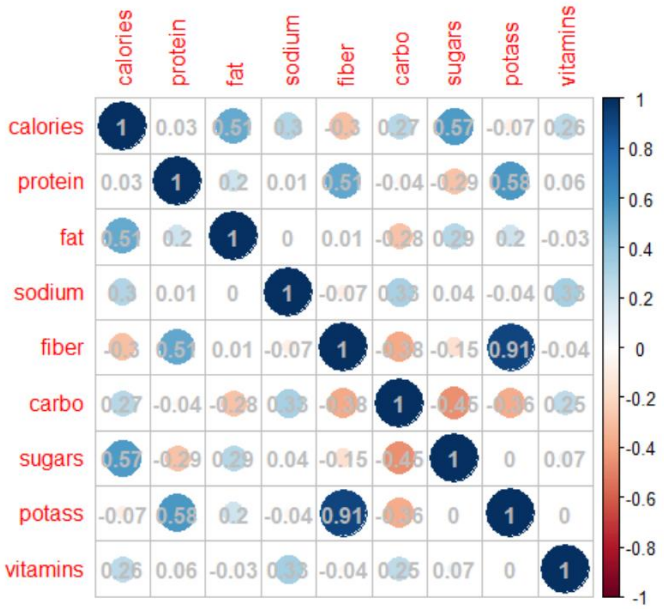
WSS 分数是集群中所有点的距离的平方的总和。为了使聚类的效果变好，希望每一类内的数据的特征尽可能相似，即每一类内数据间距离的平方和尽可能相对小。随着聚类数目增多，每一个类别中数量越来越少，距离越来越近，因此 WSS 值肯定是随着聚类数目增多而减少的，所以关注的是斜率的变化，当 WWS 减少得很缓慢时，就认为进一步增大聚类数效果也并不能增强，存在的这个“肘点”就是最佳聚类数目。

四、数据分析

1. 主成分分析

如上文所示，影响麦片营养成分的指标有很多，为了合并重复的信息，在降低现有变量的维度的基础上又不丢失重要信息，本文采用主成分分析的方法构造多个“综合指标”，希望能最大程度上的区分原始数据。

由于主成分本质是（标准化后的）原始变量的线性组合，因此对原始变量之间相关性的探索是有必要的。原始变量之间的相关系数图如下图所示：



由相关系数图可以发现相关系数绝对值大于 0.5 的有 5 个。其中 calories 和 fat 的相关系数为 0.51，calories 和 sugars 的相关系数为 0.57；protein 和 potass 的相关系数为 0.58，protein 和 fiber 的相关系数为 0.51；fiber 和

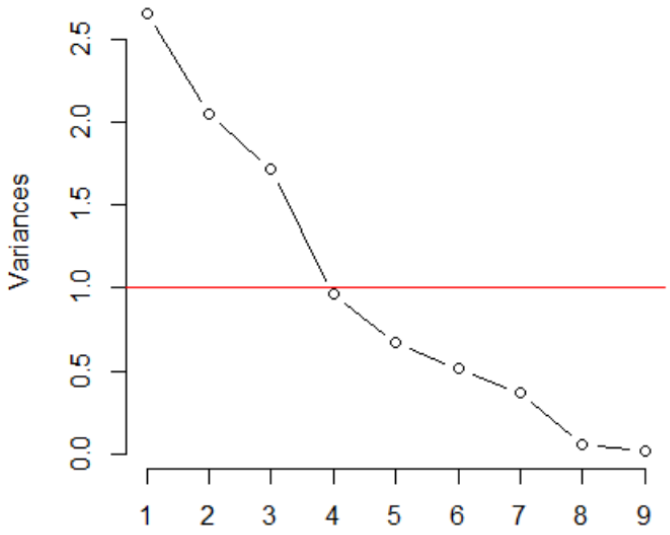
potass 的相关系数为 0.91。说明麦片中卡路里含量和含糖量、含脂肪量关系密切；麦片中含钾量、纤维含量和蛋白质含量关系密切。

对上述与麦片营养成分有关的因素标准化后进行主成分分析，结果如下表所示：

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
calories	0.228	-0.573	0.16	-0.21	-0.02	0.426	-0.08	-0.27	0.54
protein	-0.392	-0.084	0.373	-0.37	-0.14	0.064	0.72	-0.01	-0.14
fat	-0.093	-0.514	-0.09	-0.49	-0.01	-0.55	-0.32	0.018	-0.27
sodium	0.17	-0.175	0.469	0.287	0.724	-0.3	0.148	0.017	0.015
fiber	-0.562	-0.003	0.146	0.221	0.091	0.138	-0.31	-0.67	-0.2
carbo	0.339	0.183	0.501	-0.28	-0.01	0.386	-0.33	0.115	-0.5
sugars	0.098	-0.526	-0.34	0.37	0.022	0.334	0.227	0.1	-0.54
potass	-0.547	-0.165	0.164	0.126	0.065	0.216	-0.3	0.675	0.182
vitamins	0.129	-0.179	0.444	0.468	-0.67	-0.3	-0.04	0.005	0.013

同时可以得到各个主成分的标准差、方差占比、累计方差占比和碎石图如下所示：

	Standard deviation	Proportion of Variance	Cumulative Proportion
PC1	1.6288	0.2948	0.2948
PC2	1.4308	0.2275	0.5222
PC3	1.3086	0.1903	0.7125
PC4	0.9806	0.1068	0.8193
PC5	0.81659	0.07409	0.89342
PC6	0.71611	0.05698	0.9504
PC7	0.60708	0.04095	0.99135
PC8	0.24271	0.00655	0.99789
PC9	0.13776	0.00211	1



可以发现前四个指标就可以累计包含原数据中 80%以上的信息，因此只采用前四个指标就可以解释说明麦片中的影响因素，这四个指标分别可以表示为：

$$PC1 = 0.228calories - 0.392protein - 0.093fat + 0.17sodium - 0.562fiber + 0.339carbo + 0.098sugars - 0.547potass + 0.129vitamins$$

$$PC2 = -0.573calories - 0.084protein - 0.514fat - 0.175sodium - 0.003fiber + 0.183carbo - 0.526sugars - 0.165potass - 0.179vitamins$$

$$PC3 = 0.16calories + 0.373protein - 0.09fat + 0.469sodium + 0.146fiber + 0.501carbo - 0.34sugars + 0.164potass + 0.444vitamins$$

$$PC4 = -0.21calories - 0.37protein - 0.49fat + 0.287sodium + 0.221fiber - 0.28carbo + 0.37sugars + 0.126potass + 0.468vitamin$$

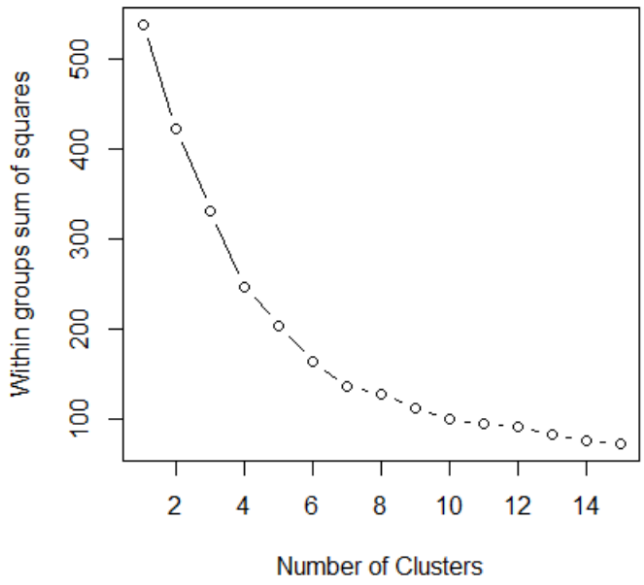
指标 1 可以表示麦片中纤维素、脂肪、蛋白质和其他营养素的相对差距，指标 1 越大说明该麦片中纤维素、脂肪、蛋白质的含量相对较少，而其他营养素相对较多；指标 2 可以表示麦片中碳水化合物和其他营养指标的差距，指标 2 越大说明该麦片中的碳水化合物相对越多、其他营养素相对越少；指标 3 可以表示麦片中糖与脂肪和其他营养素的含量的差异，指标 3 越大说明该麦片中的糖与脂肪相对较少、其他营养素相对较多；指标 4 可以表示麦片中卡路里、蛋白质、脂肪、碳水化合物和其他营养素含量的相对差距，指标 4 越大说明卡路里、蛋白质、脂肪、碳水化合物含量相对较少，其他营养素含量相对较多。

综上所述，上述四个指标从不同方面描述了麦片的营养成分含量，而综合上述四个指标则能很大程度上解释该麦片的营养成分水平。

2. 聚类分析

为了进一步区分不同麦片之间营养成分的差异和共性，基于上文主成分分析的结果进行 Kmeans 聚类分析。

为了确定聚类的类别数量，采用 WSS 方法进行确定，画出 WSS 随类别数 k 的变化趋势图如下所示：



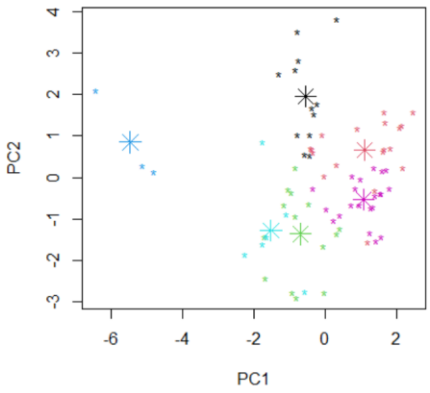
发现 k=6 之后的 WSS 下降相对不是很明显了，因此选择 k=6 作为聚类的类别数。聚类结果如下所示：

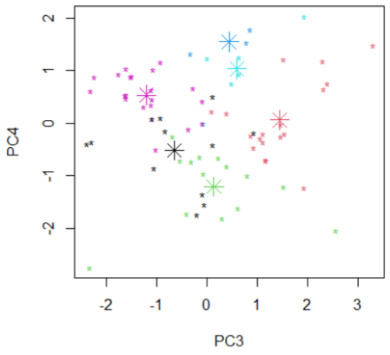
类别	1	2	3	4	5	6
观测个数	12	18	14	3	6	21

并且这 6 类的中心点的四个指标的数据分别如下所示：

	PC1	PC2	PC3	PC4
1	-0.54849	1.951578	-0.64924	-0.52202
2	1.09644	0.665867	1.454349	0.060644
3	-0.70389	-1.36789	0.141366	-1.20968
4	-5.46602	0.850031	0.4404	1.546232
5	-1.52791	-1.27647	0.599135	1.034532
6	1.060283	-0.53073	-1.20393	0.536302

聚类结果的图像如下图所示：





结合四个指标的实际意义可以发现：

第一类麦片中碳水化合物、糖、卡路里相对占比较多，其他占比较少；

第二类麦片中碳水化合物相对占比较多，其他占比较少；

第三类麦片中卡路里、纤维、蛋白质相对占比较多，其他占比较少；

第四类麦片中碳水化合物、脂肪、蛋白质、纤维相对占比较多，其他占比较少；

第五类麦片中蛋白质、纤维相对占比较多，其他占比较少；

第六类麦片中糖、脂肪相对占比较多，其他占比较少；

五、结论

麦片是一种相较其他食品低脂低卡低糖的健康食物。麦片中包含供能的糖、碳水化合物、脂肪等营养成分；维生素和微量元素等营养成分；促进胃肠蠕动的纤维类营养成分。

在本数据集中，不同品牌类型的麦片可以大致根据营养成分含量分为六类：

第一类麦片中碳水化合物、糖、卡路里相对占比较多，其他占比较少；

第二类麦片中碳水化合物相对占比较多，其他占比较少；

第三类麦片中卡路里、纤维、蛋白质相对占比较多，其他占比较少；

第四类麦片中碳水化合物、脂肪、蛋白质、纤维相对占比较多，其他占比较少；

第五类麦片中蛋白质、纤维相对占比较多，其他占比较少；

第六类麦片中糖、脂肪相对占比较多，其他占比较少；

六、附录

```
library(ggplot2)
library(mice)
library(corrplot)
library(gridExtra)
library(factoextra)

cerealorigin<-read.csv("D://学习//大三上//探索性数据分析
//hw3//cereals.csv",header=T)
###数据预处理
#缺失值
head(cerealorigin)
md.pattern(cerealorigin,rotate.names = TRUE)
cereal<-cerealorigin[complete.cases(cerealorigin),]
cereall<-cereal[,c(-1,-2,-3,-13,-14,-15)]

#异常值
boxplot(cereall$calories)
boxplot(cereall$protein)
boxplot(cereall$fat)
boxplot(cereall$sodium)
boxplot(cereall$fiber)
boxplot(cereall$carbo)
boxplot(cereall$sugars)
boxplot(cereall$potass)
hist(cereall$vitamins)

###可视化
score.corr <- round(cor(cereall), 3)
corrplot(score.corr, addCoef.col = 'grey')
ggplot(cereal,aes(x=factor(1),fill=mfr))+geom_bar()+coord_polar(the
ta="y")+scale_fill_manual(values=alpha(c("#99CCFF","#9999FF","#6666
FF","#0033CC","#0099CC","#660099","#000066"), 0.65))

p1<-hist(cereall$calories,col = "lightblue")
ggplot(cereal)+geom_histogram(aes(x=calories,fill=mfr),binwidth =
17)
p2<-hist(cereall$protein,col = "lightblue")
```

```r
p3<-hist(cereall$fat,col = "lightblue")
p4<-hist(cereall$sodium,col = "lightblue")
p5<-hist(cereall$fiber,col = "lightblue")
p6<-hist(cereall$carbo,col = "lightblue")
p7<-hist(cereall$sugars,col = "lightblue")
p8<-hist(cereall$potass,col = "lightblue")
p9<-hist(cereall$vitamins,col = "lightblue")


###主成分
score.pca <- prcomp(cereall, scale=T, retx=T)
summary(score.pca)
score.pca$rotation
score.pca
#主成分得分
cerealpca<- predict(score.pca)
cerealpca
#碎石图
plot(score.pca,type="lines",main="")
abline(h=1,col='red')
box()

###聚类
set.seed(1234)
cerealpcak<-cerealpca[,1:4]
wssplot <- function(data, nc=15, seed=1234){
  wss <- (nrow(data)-1)*sum(apply(data, 2, var))
  for(i in 2:nc){
    set.seed(seed)
    wss[i] <- sum(kmeans(data, centers = i)$withinss)
  }
  plot(1:nc, wss, type = "b",xlab = "Number of Clusters",
       ylab = "Within groups sum of squares")
}
wssplot(cerealpcak)

#kmeans 聚类
set.seed(1234)
cerealkmeans <- kmeans(cerealpcak,centers = 6)
table(cerealkmeans$cluster)
cerealkmeans$centers

plot(cerealpca[,1:2],col=cerealkmeans$cluster,pch="*")
points(cerealkmeans$centers[,1:2],pch=8,col=1:6,cex=2)
```

```
plot(cerealpca[,3:4],col=cerealkmeans$cluster,pch="*")
points(cerealkmeans$centers[,3:4],pch=8,col=1:6,cex=2)



fviz_cluster(cerealkmeans, data = cerealpcak)
```