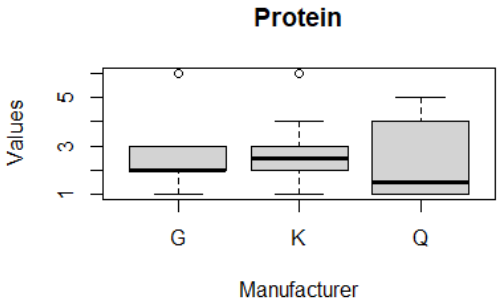
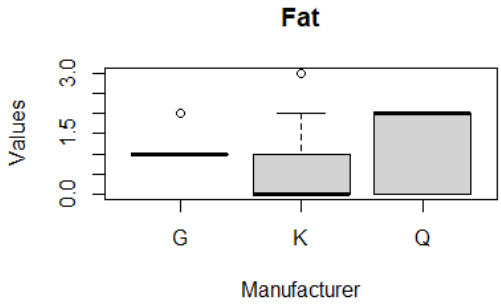
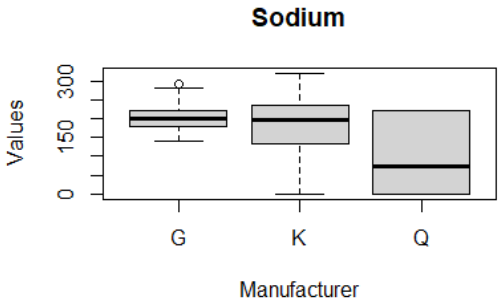
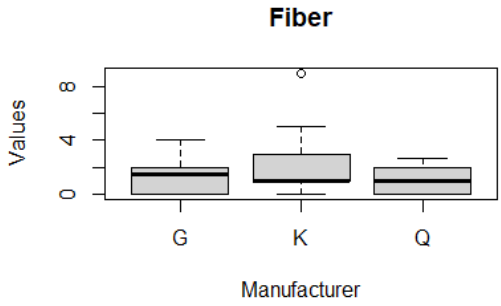
S1. Explore the distribution of each variables.

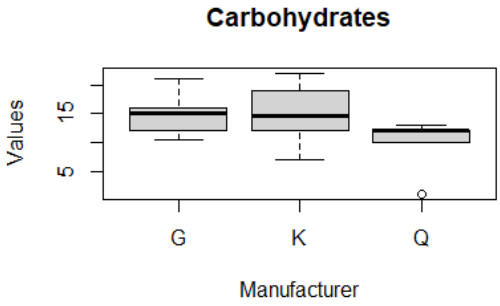
*Boxplots

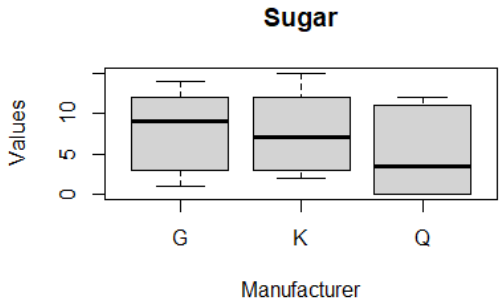



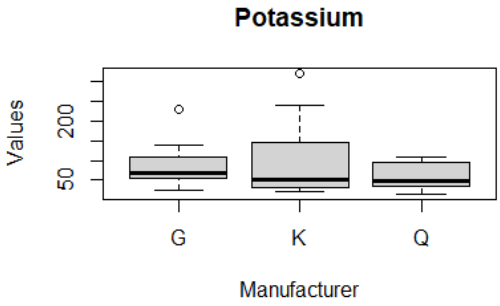



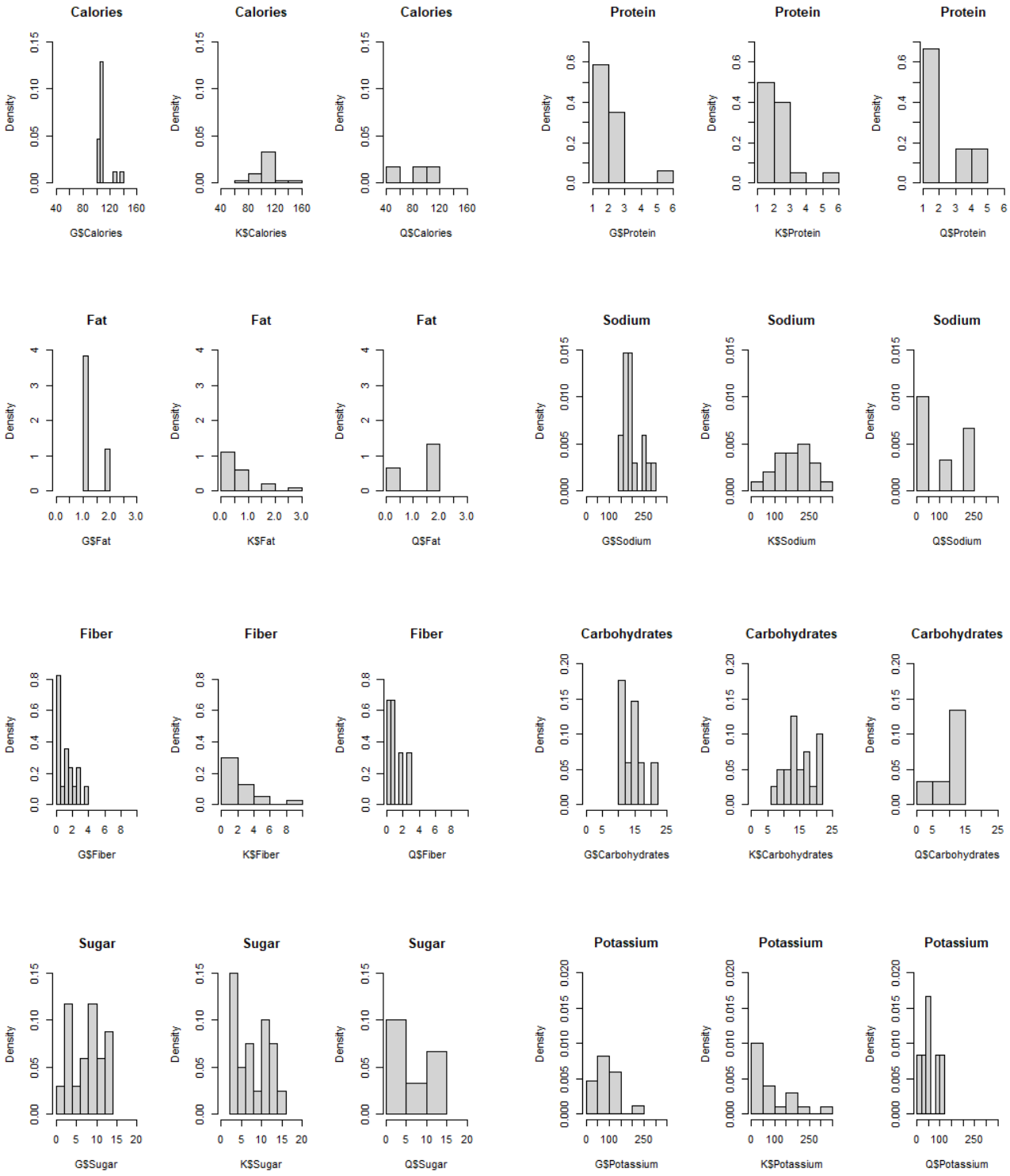




*Histograms




**Calories:**

Boxplots: 由箱形圖可看出 G 廠的中位數與 Q1 以及 Q3 幾乎相等，箱子被壓扁為一條線，推測可能原因為資料量不多且現有的數據除了離群值之外皆非常集中。K 廠顯示較多的離群值，Q 廠沒有離群值，但 Q 廠的箱子較其他兩廠商長很多，可能原因為 Q 廠樣本數據少，因此箱型受單獨個體影響明顯。總體

而言，資料集中度為 G 廠最高，K 場次之，Q 廠最低。另外，由三個廠商的箱子高低可以看出 G 廠和 K 廠卡路里相差不遠，而 Q 廠熱量則比 G,K 兩廠低一些。

Histograms: 由直方圖可見 G 廠平均分佈最集中在 110 大卡，而 K 廠商也是集中於 100~120 大卡之間，並且大致呈現對稱型態，最後 Q 廠資料平均分佈於三項數值間，且最高值在 G,K 集中的 120 大卡。

**Protein:**

Boxplots: 可見 G 廠普遍呈現左偏型態，外加一筆離群資料，B 廠呈現對稱分佈，也另外外加一筆離群值，而 Q 廠仍是三者中箱子最大的，表示其數值較 G,K 兩廠分散。另外，觀察中位數可看出 G 廠商除離群值外，其餘樣本皆集中在 2 公克；K 廠平均集中在 2~3 公克之間；Q 廠則是 1~2 公克居多。

Histograms: 直方圖可明顯看出三家廠商皆發生以 1~3 公克為大部分蛋白質公克數，偶有少數超出 3 公克的情況，而 Q 廠則發現較多此情況，因此未被歸類於離群值。

**Fat:**

Boxplots: 明顯看出 G 廠商幾乎將脂肪含量精準控制在 1 公克；K 廠有半數集中在少於 1 公克的範圍內；Q 廠為三廠商中箱子最長的，代表其樣本分布較廣，中位數也顯示其脂肪含量多位居三者之首。

Histograms: 相較箱型圖，直方圖可明顯看出資料集中位置的細微變化，像是 G 廠除了佔多數的 1 公克外，其餘被列為離群值的皆為 2 公克；K 廠則較鬆散的分佈於 0 至 1 之間；Q 廠則生產較多含 2 公克脂肪的產品，剩餘皆為 0 公克。

**Sodium:**

Boxplots: 可以看出資料集中度為 G 廠最高，K 廠次之，Q 廠最分散。從四分位距及中位數推測鈉含量大致而言也是 G 廠最高，K 廠次之，Q 廠最低。

Histograms: 同時從直方圖也可以再次驗證箱型圖看出的三家廠商資料集中度，除此之外，可更細部觀察出 G 廠商鈉含量多集中在 150~200 公克之間，K 廠幾乎呈現常態分佈，Q 廠以 50 公克以下及 200 以上為主。

**Fiber:**

Boxplots: 由箱型圖觀察可得相較於前四個變數，纖維素的含量三家廠商有差異不大的結果，其中 K 廠的纖維素含量在部分產品具有突出的表現，總的來說，K 廠在纖維素的含量分佈上有稍大的分散。

Histograms: 由直方圖也可以明顯看出 G,Q 廠相較 K 廠的數據要集中一些，G 廠商多數產品的纖維素含量界於 0~1 之間，Q 廠則是多數皆低於 3 公克。

**Carbohydrates:**

Boxplots: 透過中位數觀察出 G 廠跟 K 廠有相似的分佈，但兩者之中，G 廠碳水化合物的含量控制得較為精確。Q 廠商則較 G,K 兩廠有更低的碳水化合物含量，其中有部分品牌製造出碳水化合物含量極低的產品。

Histograms: 三家廠商由直方圖觀察分別集中在右邊、左邊、中間，G 廠大部分碳水化合物含量集中在 10 至 17 公克之間；K 廠則穩定分佈於中間地帶；Q 廠的碳水化合物含量業多集中在 10 至 15 公克之間，其餘少數分佈於 10 公克以下，為三廠商中含量普遍最低者。

**Sugar:**

Boxplots: 三家廠商皆具有較長的箱子，代表三者對於糖類的含量控制可能較不嚴苛，因此糖類含量變動大，形成較分散的箱型圖。

Histograms: 如同箱型圖所觀察到的，三家廠商的糖類含量在直方圖中也呈現較為平均分散的情況，其中 Q 廠商有較 G,K 廠稍低的糖類含量。

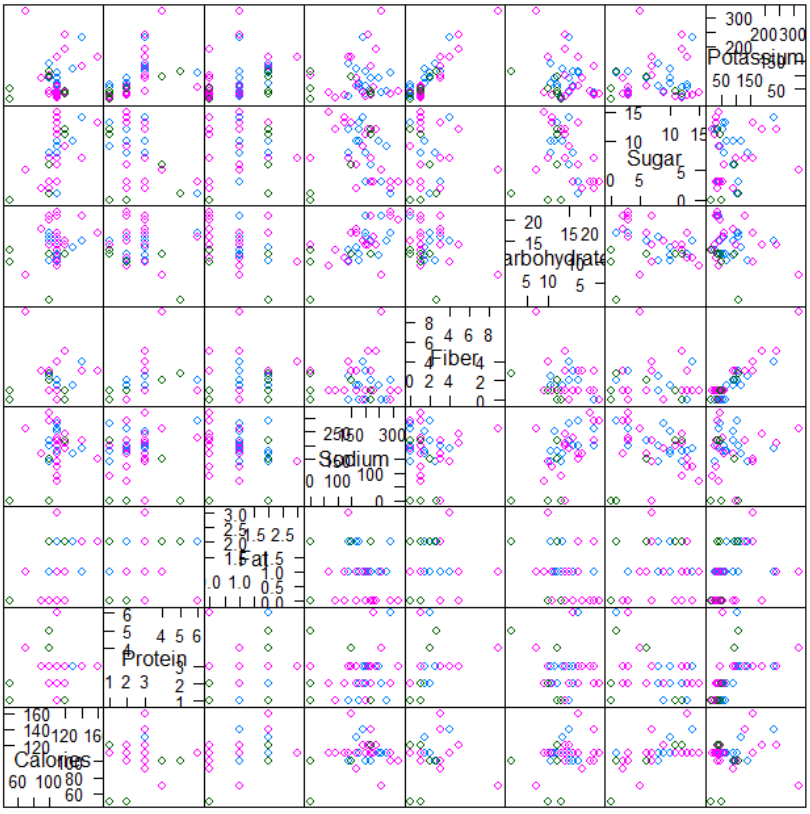
**Potassium:**

Boxplots: G 廠及 Q 廠有較相似的箱型圖，G 廠的鉀含量又比 Q 廠再稍微高出一些，另外也有部分離群值產生。K 廠商為三者之中資料分佈較為分散者，代表 K 廠製造的不同品牌的麥片其鉀含量較不一致，他的離群值也比 G 廠商的高出許多。

Histograms: 直方圖中 Q 廠較 G 廠看起來有更高的集中度，可明顯指出 Q 廠鉀含量多數集中在 50 毫克上下。

S2. Explore the association among the eight nutritional components.

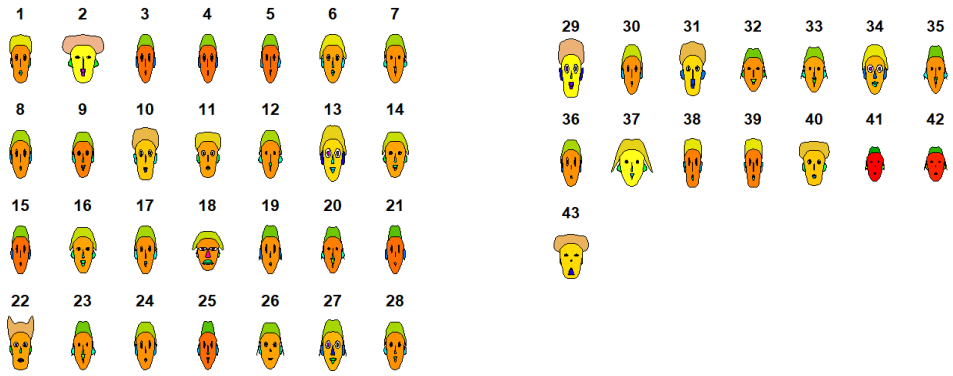
* Scatterplot Matrix and Chernoff Faces




Scatter Plot Matrix

Scatterplot matrix:

從矩陣散步圖可以看出 Fiber 跟 Potassium 之間有線性正相關，這表示當麥片中的鉀含量上升時，纖維素成分也會上升。其餘變數之間無顯著相關。




Chernoff faces:

變數對應特徵:

1. 臉及頭髮高度: Calories
2. 臉及頭髮寬度: Protein
3. 臉部結構、髮型: Fat
4. 嘴巴及鼻子高度: Sodium
5. 嘴巴及鼻子寬度: Fiber
6. 微笑程度、耳朵寬度: Carbohydrates
7. 眼睛及耳朵高度: Sugar
8. 眼睛寬度: Potassium

說明: 直觀 43 張臉，其中較為突出有 2 號的臉部及頭髮寬度,18 號的眼睛寬度、嘴巴及鼻子寬度、臉及頭髮的寬度,22 號的臉部結構及髮型,34 號的眼睛寬度，表示對應編號的品牌具有較突出的上述對應的變數數值。另外可以看到屬於 Q 廠商的 41,42 兩種品牌有特別小的臉部結構及鼻子、嘴巴，表示這兩種品牌有較低的卡路里、脂肪、鈉。綜觀而言，三家廠商並沒有特別明顯屬於自己廠商的共同特徵。

S3. Conclusions

整體而言，G 廠及 K 廠有較多的相似度，並且 G 廠商對於所有成分的控制力較強，大多較 K 廠商的成分含量更為集中。Q 廠由於資料較少，因此每筆數據影響力度較強，並不完全適合用來與其他兩家廠商共同比較。從實際營養與否角度看三家廠商的數據，K 廠與 G 廠比較有較低的脂肪、較高的纖維素及蛋白質，但同時 G 廠也得到較低的碳水化合物含量，選擇哪家就要依據消費者較關注哪種變數為主。

R Coding:

```
# clear all variables
rm(list = ls(all = TRUE))
graphics.off()

# load data
Cereal <- read.csv("C:/Users/user/Documents/Cereal.csv")

summary(Cereal[1:17,3:10])

# S1.1 Plot box plot
boxplot(Cereal$Calories~Cereal$Manufacturer, xlab = "Manufacturer", ylab =
"Values", main = "Calories")
boxplot(Cereal$Protein~Cereal$Manufacturer, xlab = "Manufacturer", ylab =
"Values", main = "Protein")
boxplot(Cereal$Fat~Cereal$Manufacturer, xlab = "Manufacturer", ylab = "Values",
main = "Fat")
boxplot(Cereal$Sodium~Cereal$Manufacturer, xlab = "Manufacturer", ylab =
"Values", main = "Sodium")
boxplot(Cereal$Fiber~Cereal$Manufacturer, xlab = "Manufacturer", ylab = "Values",
main = "Fiber")
boxplot(Cereal$Carbohydrates~Cereal$Manufacturer, xlab = "Manufacturer", ylab =
"Values", main = "Carbohydrates")
boxplot(Cereal$Sugar~Cereal$Manufacturer, xlab = "Manufacturer", ylab = "Values",
main = "Sugar")
boxplot(Cereal$Potassium~Cereal$Manufacturer, xlab = "Manufacturer", ylab =
"Values", main = "Potassium")

# S1.2 Histograms
install.packages("lattice")
library(lattice)

G <- subset(Cereal, Cereal$Manufacturer == "G", select = Manufacturer:Potassium)
K <- subset(Cereal, Cereal$Manufacturer == "K", select = Manufacturer:Potassium)
Q <- subset(Cereal, Cereal$Manufacturer == "Q", select = Manufacturer:Potassium)
par(mfrow = c(1, 3))
```

```
# Calories
hist(G$Calories,freq = FALSE,xlim = c(40,160),ylim = c(0,0.15),main = "Calories")
hist(K$Calories,freq = FALSE,xlim = c(40,160),ylim = c(0,0.15),main = "Calories")
hist(Q$Calories,freq = FALSE,xlim = c(40,160),ylim = c(0,0.15),main = "Calories")
# Protein
hist(G$Protein,freq = FALSE,xlim = c(1,6),ylim = c(0,0.7),main = "Protein")
hist(K$Protein,freq = FALSE,xlim = c(1,6),ylim = c(0,0.7),main = "Protein")
hist(Q$Protein,freq = FALSE,xlim = c(1,6),ylim = c(0,0.7),main = "Protein")
# Fat
hist(G$Fat,freq = FALSE,xlim = c(0,3),ylim = c(0,4),main = "Fat")
hist(K$Fat,freq = FALSE,xlim = c(0,3),ylim = c(0,4),main = "Fat")
hist(Q$Fat,freq = FALSE,xlim = c(0,3),ylim = c(0,4),main = "Fat")
# Sodium
hist(G$Sodium,freq = FALSE,xlim = c(0,350),ylim = c(0,0.015),main = "Sodium")
hist(K$Sodium,freq = FALSE,xlim = c(0,350),ylim = c(0,0.015),main = "Sodium")
hist(Q$Sodium,freq = FALSE,xlim = c(0,350),ylim = c(0,0.015),main = "Sodium")
# Fiber
hist(G$Fiber,freq = FALSE,xlim = c(0,10),ylim = c(0,0.9),main = "Fiber")
hist(K$Fiber,freq = FALSE,xlim = c(0,10),ylim = c(0,0.9),main = "Fiber")
hist(Q$Fiber,freq = FALSE,xlim = c(0,10),ylim = c(0,0.9),main = "Fiber")
# Carbohydrates
hist(G$Carbohydrates,freq = FALSE,xlim = c(0,25),ylim = c(0,0.2),main =
"Carbohydrates")
hist(K$Carbohydrates,freq = FALSE,xlim = c(0,25),ylim = c(0,0.2),main =
"Carbohydrates")
hist(Q$Carbohydrates,freq = FALSE,xlim = c(0,25),ylim = c(0,0.2),main =
"Carbohydrates")
# Sugar
hist(G$Sugar,freq = FALSE,xlim = c(0,20),ylim = c(0,0.15),main = "Sugar")
hist(K$Sugar,freq = FALSE,xlim = c(0,20),ylim = c(0,0.15),main = "Sugar")
hist(Q$Sugar,freq = FALSE,xlim = c(0,20),ylim = c(0,0.15),main = "Sugar")
# Potassium
hist(G$Potassium,freq = FALSE,xlim = c(0,350),ylim = c(0,0.02),main = "Potassium")
hist(K$Potassium,freq = FALSE,xlim = c(0,350),ylim = c(0,0.02),main = "Potassium")
hist(Q$Potassium,freq = FALSE,xlim = c(0,350),ylim = c(0,0.02),main = "Potassium")

# S2.1
# Scatterplot Matrix
```

```
splom(Cereal[, 3:10], groups=Cereal$Manufacturer, data=Cereal)

# install and load packages
libraries = c("aplpack")
lapply(libraries, function(x) if (!(x %in% installed.packages())) {
   install.packages(x)
})
lapply(libraries, library, quietly = TRUE, character.only = TRUE)
x <- Cereal[,3:10]
ncolors = 20

# face plot
faces(x, nrow = 4, face.type = 1, scale = TRUE, col.nose = rainbow(ncolors), col.eyes =
rainbow(ncolors,

start = 0.6, end = 0.85), col.hair = terrain.colors(ncolors), col.face =
heat.colors(ncolors),
        col.lips = rainbow(ncolors, start = 0, end = 1), col.ears = rainbow(ncolors,

start = 0, end = 0.8), plot.faces = TRUE)
```