Homework 2

Part I

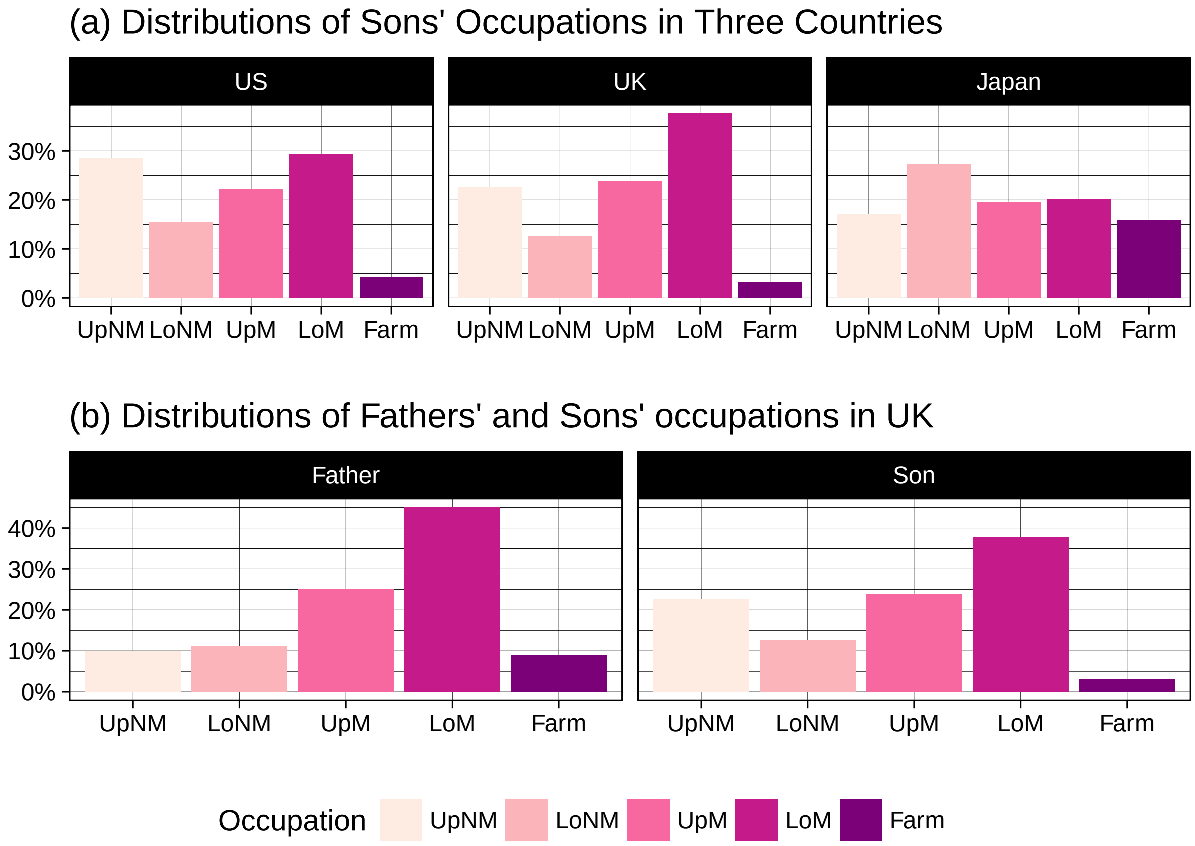
Ch4.5 Occupational Mobility

```
library("tidyverse")
library("grid")
library("gridExtra")
library("scales")
library("GGally")
```

```
data(Yamaguchi87, package="vcdExtra")
Yamaguchi87 <- tbl_df(Yamaguchi87)
```

```
grid_arrange_shared_legend <- function(..., ncol = length(list(...)), nrow = 1, position = c("bottom",
  plots <- list(...)
  position <- match.arg(position)
  g <- ggplotGrob(plots[[1]] + theme(legend.position = position))$grobs
  legend <- g[[which(sapply(g, function(x) x$name) == "guide-box")]]
  lheight <- sum(legend$height)
  lwidth <- sum(legend$width)
  gl <- lapply(plots, function(x) x + theme(legend.position="none"))
  gl <- c(gl, ncol = ncol, nrow = nrow)
  combined <- switch(position,
                     "bottom" = arrangeGrob(do.call(arrangeGrob, gl),
                                            legend,
                                            ncol = 1,
                                            heights = unit.c(unit(1, "npc") - lheight, lheight)),
                     "right" = arrangeGrob(do.call(arrangeGrob, gl),
                                           legend,
                                           ncol = 2,
                                           widths = unit.c(unit(1, "npc") - lwidth, lwidth)))
  grid.newpage()
  grid.draw(combined)
}
```

```
p45a <- Yamaguchi87 %>%
  select(Country, Son, Freq) %>%
  group_by(Country, Son) %>%
  summarise(Freq = sum(Freq)) %>%
  mutate(Percentage=Freq/sum(Freq)) %>%
  mutate(Occupation=Son) %>%
  ggplot(aes(x=Occupation, y=Percentage, fill=Occupation)) +
  geom_bar(stat="identity") +
  facet_grid(. ~ Country) +
  ggtitle("(a) Distributions of Sons' Occupations in Three Countries") +
  xlab("") + ylab("") +
  scale_y_continuous(labels=percent) +
  scale_fill_brewer(palette="RdPu") +
  theme_linedraw()

p45b <- Yamaguchi87 %>%
  filter(Country=="UK") %>%
  select(-Country) %>%
  gather(Generation, Occupation, -Freq) %>%
  group_by(Generation, Occupation) %>%
  summarise(Freq = sum(Freq)) %>%
  mutate(Percentage=Freq/sum(Freq)) %>%
  mutate(Occupation=factor(Occupation, levels = c("UpNM", "LoNM", "UpM", "LoM", "Farm"))) %>%
  ggplot(aes(x=Occupation, y=Percentage, fill=Occupation)) +
  geom_bar(stat="identity") +
  facet_grid(. ~ Generation) +
  ggtitle("(b) Distributions of Fathers' and Sons' occupations in UK") +
  xlab("") + ylab("") +
  scale_y_continuous(labels=percent) +
  scale_fill_brewer(palette="RdPu") +
  theme_linedraw()

grid_arrange_shared_legend(p45a, p45b, ncol = 1, nrow = 2)
```



a. US and UK have similar distribution of occupations, however Japan is different. We can see the bars of occupations in Japan are almost same height.

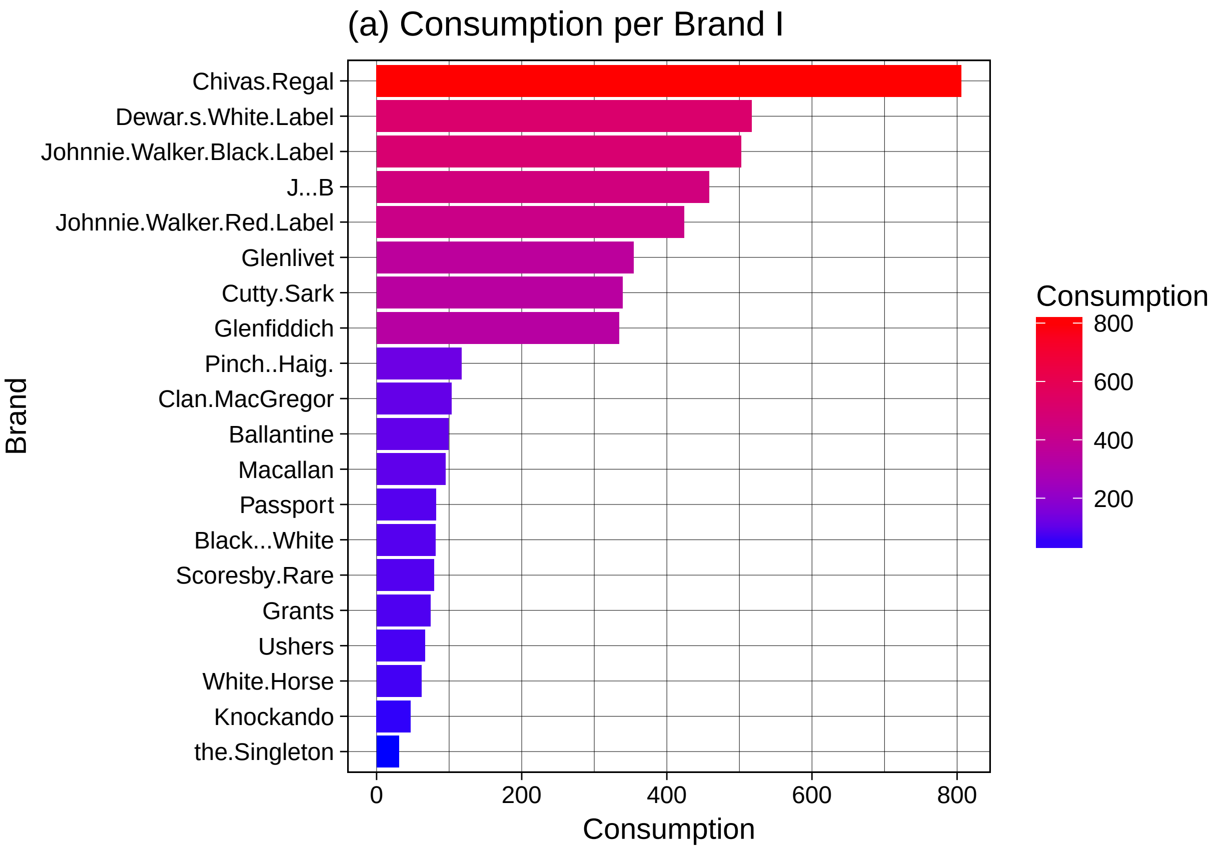
b. The distributions of Father and Son in UK look similar, but there are more Up Mon-Manual and less Low Manual and Farm in Son.

c. The results are what I would have expected. Data of these three countries were derived from surveys in early 1970s. At that time, UK and US are more developed and industrialized than Japan, so they have more people work in non-agriculture area as what we see in Figure (a). With the growth of economy and technology, machines replaced more and more human workers, so the percentage of people who work in farm and low manual area decresed and the percentage of people working in non-manual area increased, which match what we see in `Figure (b)`.

Ch4.6 Whisky

```
data(Scotch, package="bayesm")
Scotch <- tbl_df(Scotch)
```

```
Scotch %>%
  gather(Brand, Count) %>%
  filter(Brand!='Other.Brands') %>%
  group_by(Brand) %>%
  summarise(Consumption = sum(Count)) %>%
  ggplot(aes(x=reorder(Brand, Consumption), y=Consumption, fill=Consumption)) +
  geom_bar(stat="identity") +
  coord_flip() +
  xlab("Brand") +
  ggtitle("(a) Consumption per Brand I") +
  scale_fill_gradient(low="blue", high="red") +
  theme_linedraw()
```

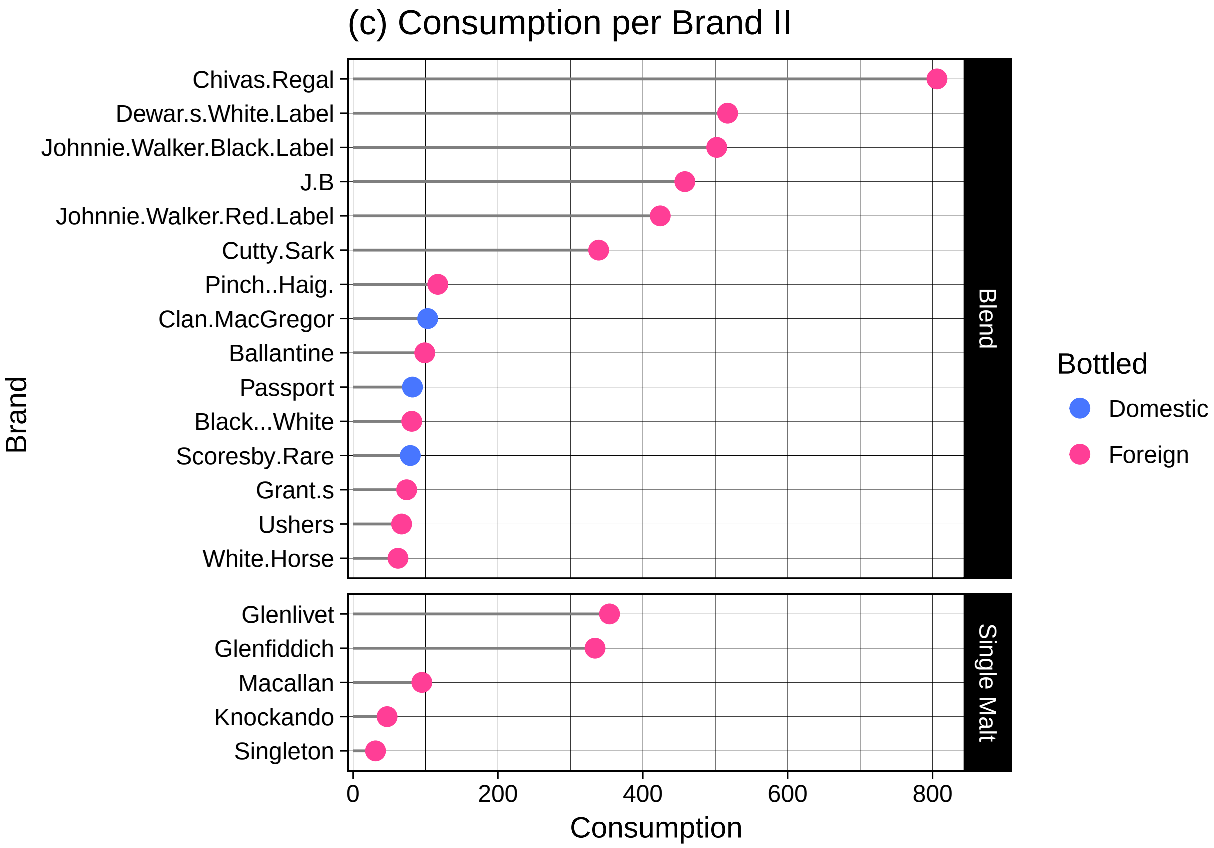


a. `Graph (a)` represents the amount of consumption in decrease order. `Chivas.Regal` is the best brand based on the information we have.

b. Since there's an obvious gap between `Glenfiddich` and `Pinch..Haig.` on the graph and the number of consumption for these two are `334` and `117`, I would pick a number in between, let's say 200 as the cutoff for big brand.

```
data(whiskey, package="flexmix")

w1 <- tbl_df(data.frame(whiskey$Incidence * whiskey$Freq)) %>%
  gather(Brand, Count) %>%
  group_by(Brand) %>%
  summarise(Consumption = sum(Count))

w2 <- tbl_df(whiskey_brands) %>%
  mutate(Brand=gsub("([^a-zA-Z])", ".", Brand))

whiskey_all <- left_join(w1, w2, by=c("Brand" = "Brand"))
```

```
whiskey_all %>%
  filter(Brand!='Other.brands') %>%
  ggplot(aes(x=Consumption, y=reorder(Brand, Consumption))) +
  geom_segment(aes(yend=Brand), xend=0, color='grey50') +
  geom_point(size=3, aes(color=Bottled)) +
  facet_grid(Type ~ ., scales='free_y', space='free_y') +
  ylab("Brand") +
  ggtitle("(c) Consumption per Brand II") +
  scale_color_manual(values = c("royalblue1", "violetred1")) +
  theme_linedraw()
```

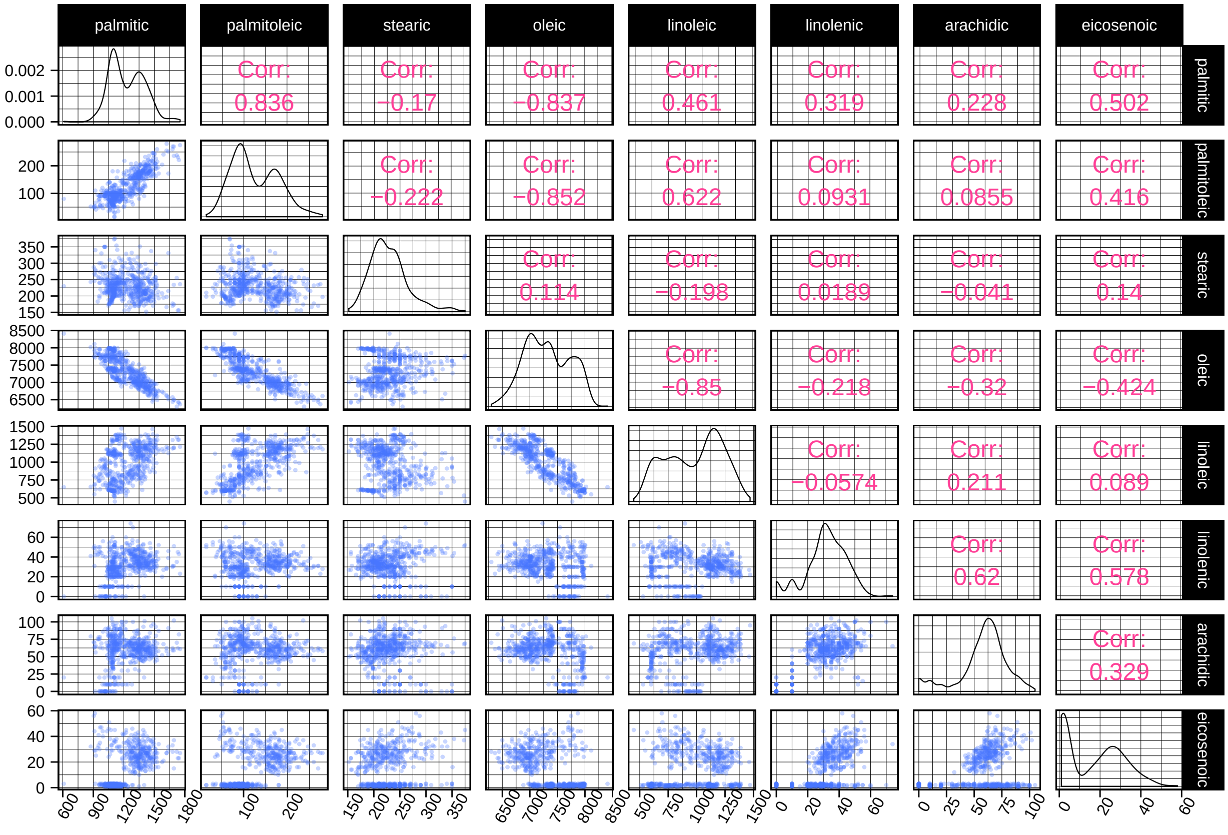


d. By coloring different `Bottled` category, we see that most Brands come outside of United States, so I think "whisky" is better.

Ch5.8 Olive oils from Italy

```
data(olives, package="extracat")
olives <- tbl_df(olives)
```

```
olives %>%
  ggpairs(columns=3:10,
          lower=list(continuous=wrap("points", alpha=0.3, size=0.1, color="royalblue1")),
          diag=list(continuous=wrap("densityDiag", alpha=0.7, size=0.2), axisLabels='none'),
          upper=list(continuous=wrap("cor", size=rel(3), color="violetred1"))
          ) +
  theme_linedraw() +
  theme(
    text=element_text(size = 7),
    axis.text=element_text(size = 6),
    axis.text.x=element_text(angle = 60))
```



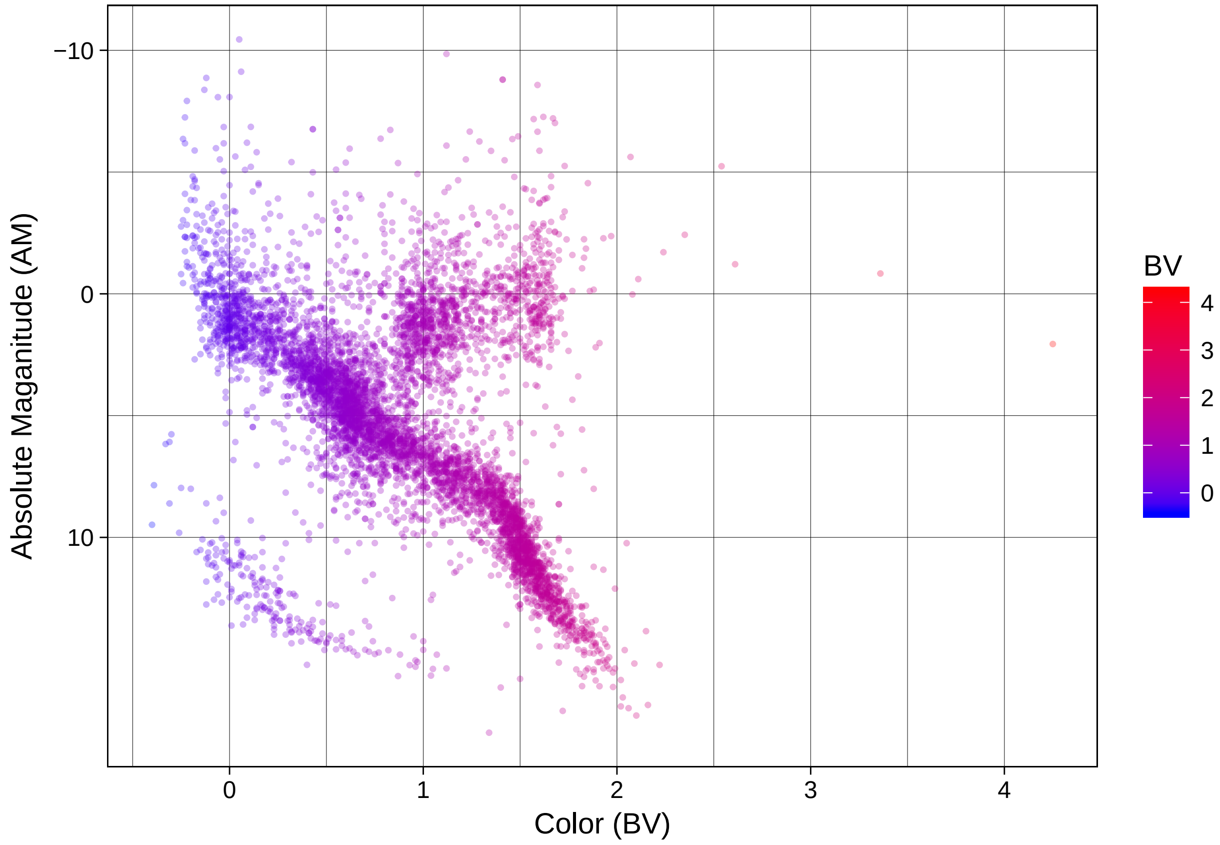
a. From the plot above we can tell that `Palmitoleic` and `Palmitic` are strongly positively associated. `Oleic` and `Palmitic`, `Oleic` and `Palmitoleic` are strongly negatively associated.

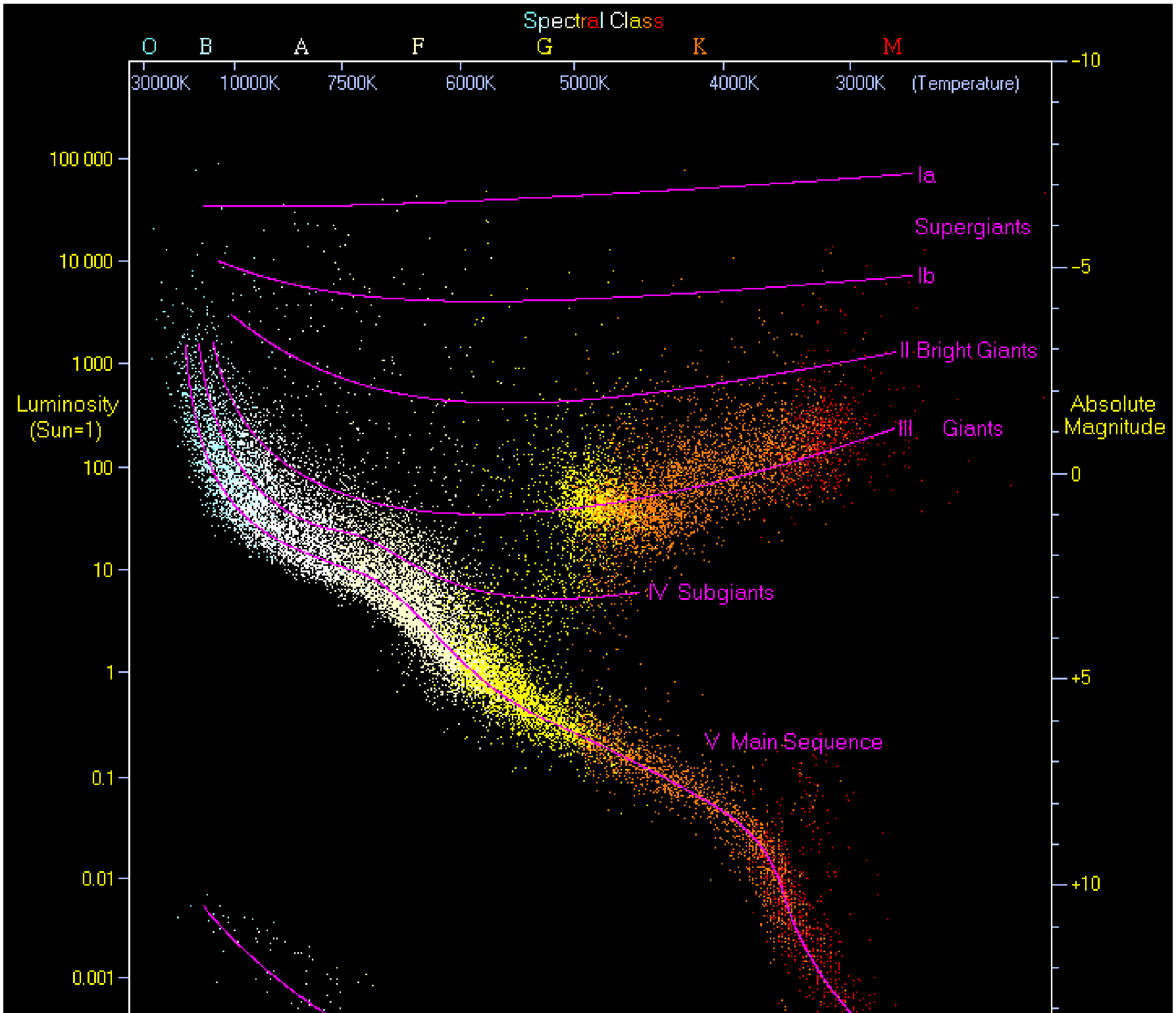
b. Yes. All the scatter plots of `Eicosenoic` associated with others have many outliers locate at the bottom, like a line.

Ch5.10 Hertzsprung-Russell

```
data("HRstars", package="GDAdata")
HRstars <- tbl_df(HRstars)
```

```r
HRstars %>%
  mutate(AM = V+5*(1+log10(Para))) %>%
  ggplot(aes(x=BV, y=AM, color=BV)) +
  geom_point(alpha=.3, size=.5) +
  scale_color_gradient(low="blue", high="red") +
  xlab("Color (BV)") + ylab("Absolute Maganitude (AM)") +
  scale_y_reverse() +
  theme_linedraw()
```


−10
0
10
Absolute Maganitude (AM)
0
1
2
3
4
Color (BV)
BV
4
3
2
1
0


Spectral Class
O
B
A
F
G
K
M
30000K
10000K
7500K
6000K
5000K
4000K
3000K
(Temperature)
100 000
10 000
1000
100
10
1
0.1
0.01
0.001
Luminosity
(Sun=1)
Ia
Supergiants
Ib
II Bright Giants
III Giants
IV Subgiants
V Main Sequence
−10
−5
0
+5
+10
Absolute
Magnitude

a. Comparing to the graph from Wikipedia, the patterns are similar. Graph drawn from `HRstars` data clearly shows `White Dwarfs`, `Main Sequence` and `Giants`. But it is probabaly due to the small amount of data we have, it is not as clear as the one on Wikipedia, and lacks of details.

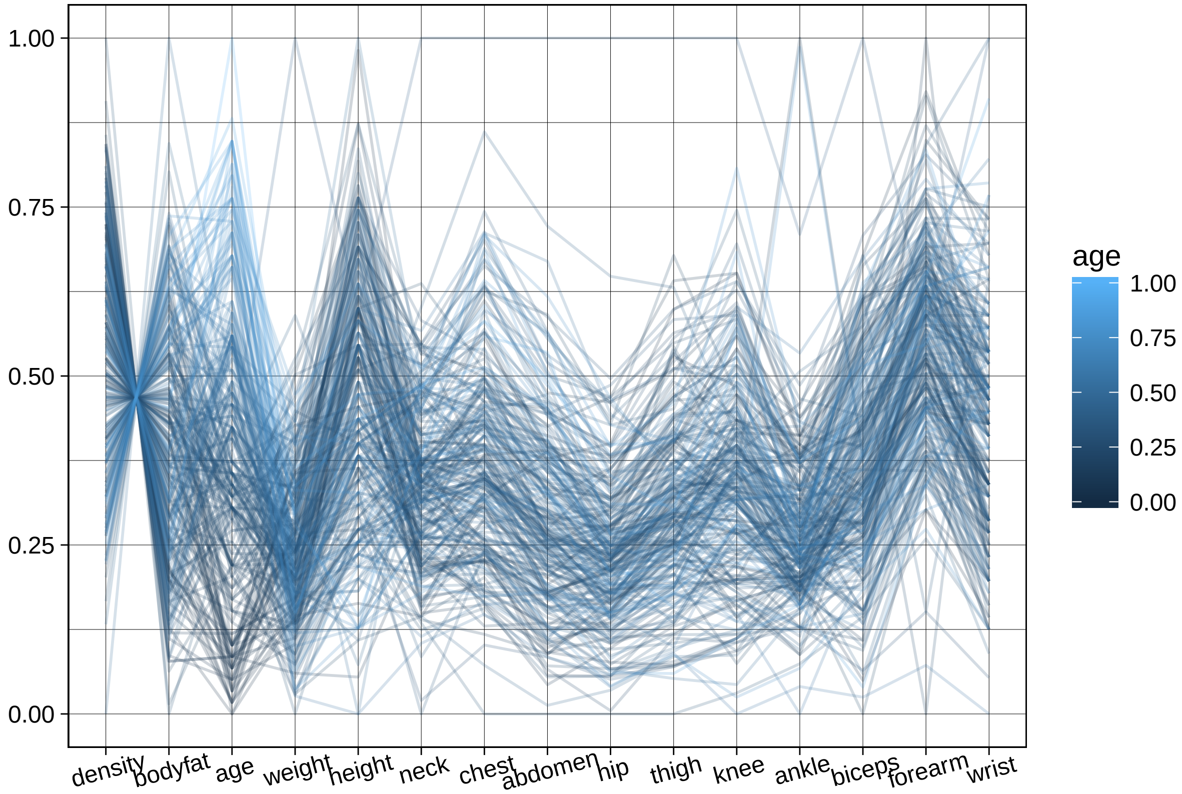
b. It seems like the graph from Wiki has more data for `Giants`, so the shape of `Giants` is a longer line.

c. I colored points by `BV`. And in the graph, there is a very obvious trend that more blue points have been drawn on the left and more red on the right.

Ch6.5 Bodyfat

```
data(bodyfat, package="MMST")
bodyfat <- tbl_df(bodyfat)
```

```
bodyfat %>%
  ggparcoord(alphaLines=0.2, scale="uniminmax", groupColumn="age") +
  xlab("") + ylab("") +
  # scale_color_gradient(low="blue", high="red") +
  theme_linedraw() +
  theme(axis.text.x=element_text(angle = 15))
```



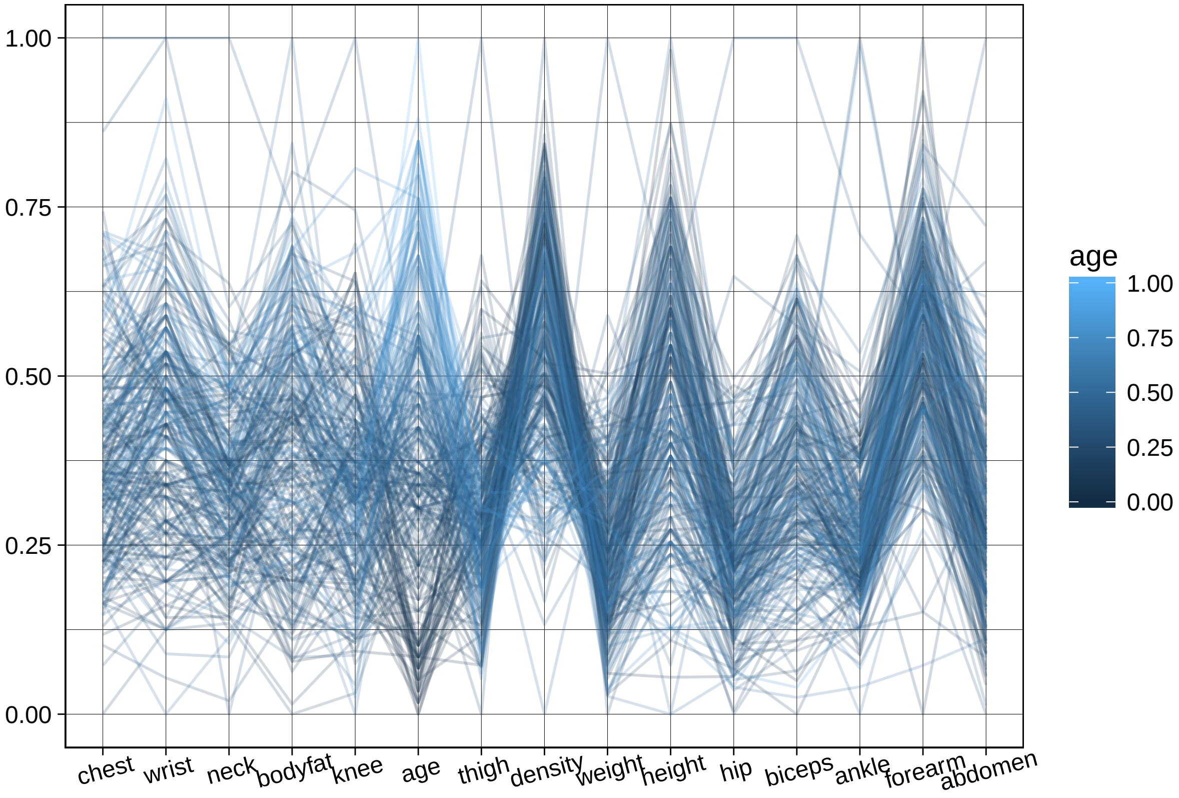
a. Yes, there are outliers. Individual outliers can be seen on most of the variables (`weight`, `neck`, `chest`, `abdomen`, `hip`, `thign`, `knee`, `ankle` and `biceps`) and usually are extreme high values.

b. `Height` has many small subgroups and positive correlation with `weight` and `neck`.

c. `density` and `bodyfat` are strongly negative correlated.

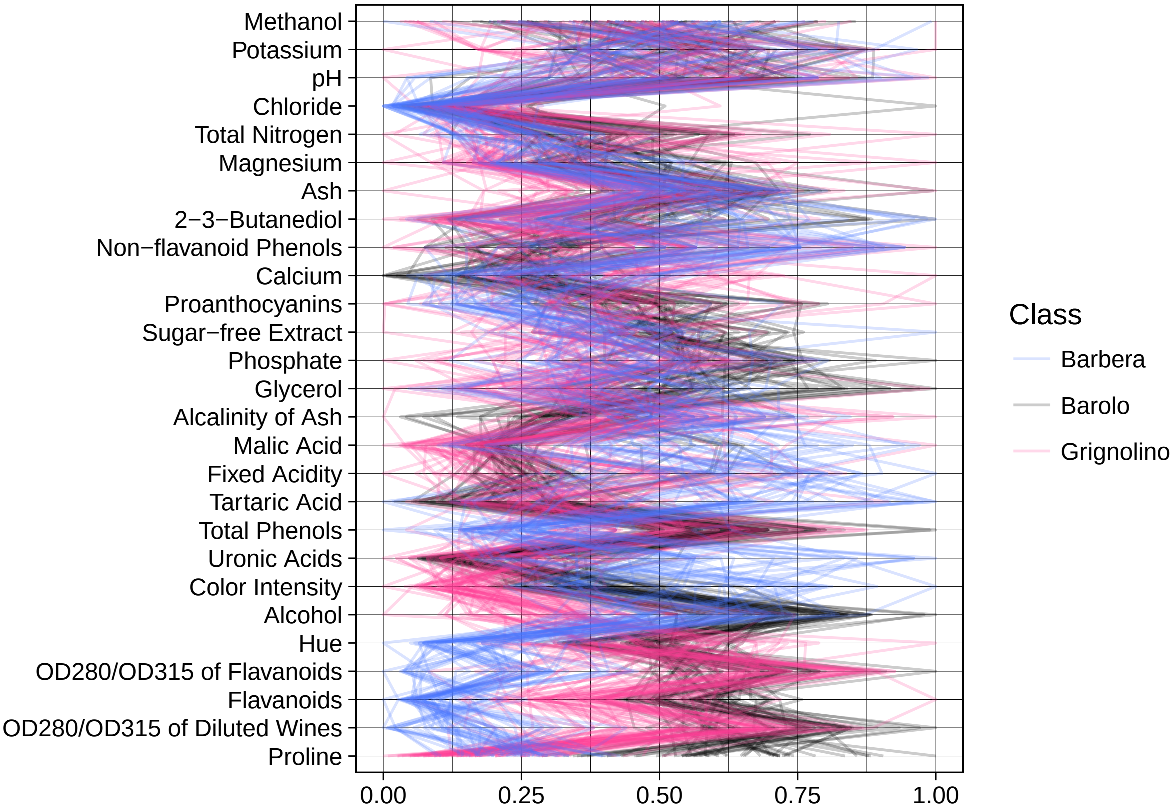
d. Yes. Because if we put the first two variables far away from each other, it will be impossible to see the negative correlation. The new graph made after reordering does a better job in showing correlations.

```
bodyfat %>%
  ggparcoord(alphaLines=0.2, scale="uniminmax",
             order=c(7,15,6,2,11,3,10,1,4,5,9,13,12,14,8),
             groupColumn="age") +
  xlab("") + ylab("") +
  theme_linedraw() +
  theme(axis.text.x=element_text(angle = 15))
```



Ch6.7 Wine

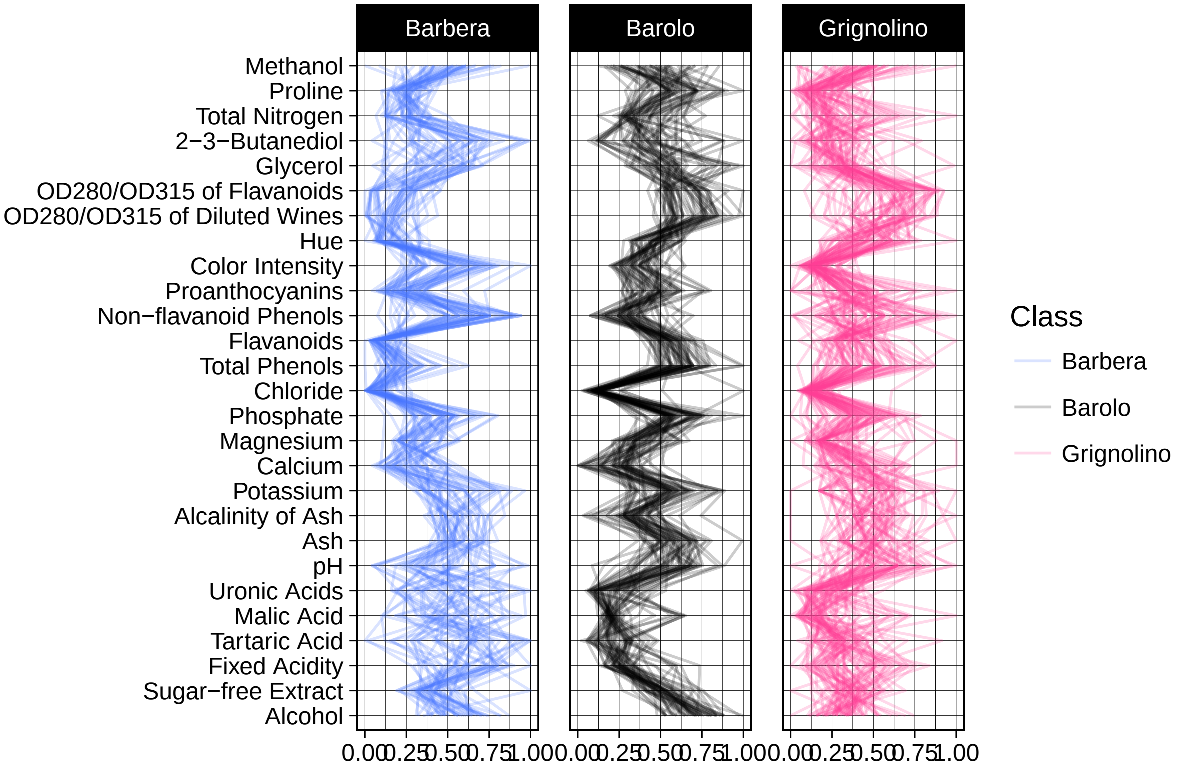
```
data(wine, package="MMST")
wine_mmst <- tbl_df(wine)

data(wine, package="pgmm")
wine_pgmm <- tbl_df(wine)

wine_pgmm <- wine_pgmm %>%
  mutate_all(funs(as.numeric)) %>%
  mutate_at(vars(Type), funs(as.factor))

wine_classname <- wine_mmst %>%
  select(Class=class, classdigit) %>%
  distinct()

wine_all <- left_join(wine_pgmm, wine_classname, by=c("Type" = "classdigit"))
wine_all %>%
  ggparcoord(columns=2:28, groupColumn="Class", alphaLines=0.2, scale="uniminmax", order="anyClass") +
  xlab("") + ylab("") +
  coord_flip() +
  scale_colour_manual(values = c("royalblue1", "black", "violetred1")) +
  theme_linedraw()
```



a. In the pcp graph, we can see that there are evidences which several variables can be used to seperate classes. `Flavanoids` seems to be able to separate all three clases. `Proline`, `OD280/OD315 of Diluted Wines`, `OD280/OD315 of Flavanoids`, `Hue`, `Alcohol` maybe helpful to separate one class from the other two.

b. Yes, there are outliers. Many variables have extreme value of the rightside in the pcp graph, such as `Flavanoids`, `hue`, `Chloride` and `Calcium`.

```
wine_all %>%
  ggparcoord(columns=2:28, alphaLines=0.2, scale="uniminmax", groupColumn="Class") +
  xlab("") + ylab("") +
  ggtitle("") +
  facet_grid(~Class) +
  coord_flip() +
  scale_colour_manual(values = c("royalblue1", "black", "violetred1")) +
  theme_linedraw() +
  theme(panel.spacing.x=unit(0.8, "lines"))
```

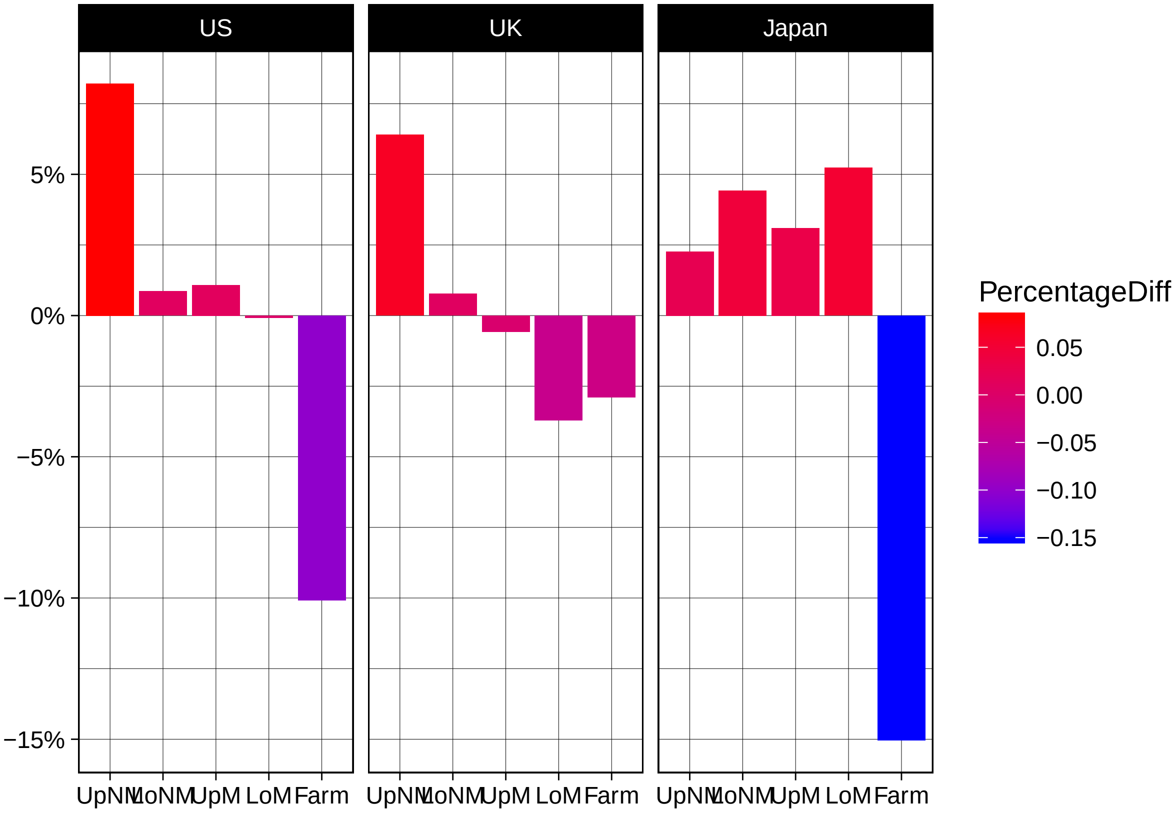


c. We deduces several variables in each `Class` which have subgroups from the pcp graph: `Color Intensity` in class `Barbera`, `Malic Acid` in class `Barolo` and `Total Phenols` in class `Grignolino`

Part II

Yamaguchi87 Dataset

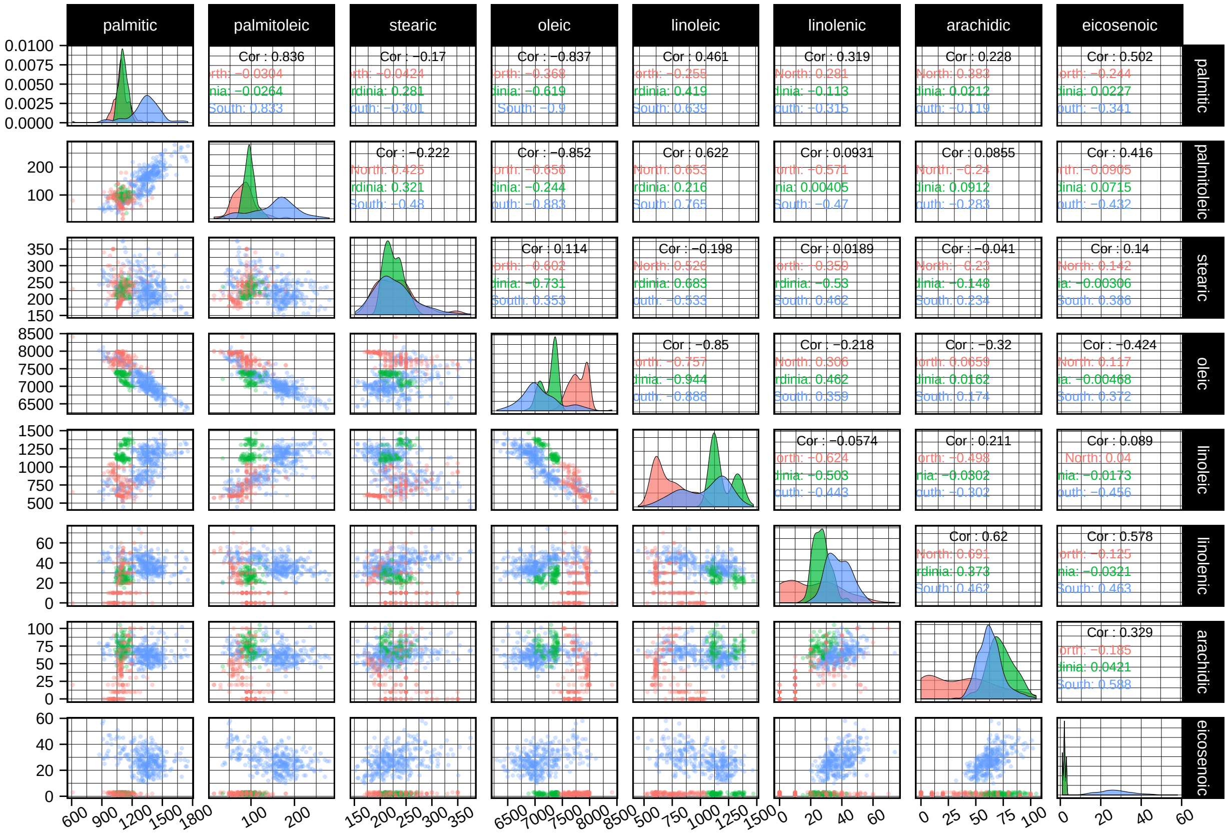
```
Yamaguchi87 %>%
  gather(Generation, Occupation, -Freq, -Country) %>%
  group_by(Country, Generation, Occupation) %>%
  summarise(Freq = sum(Freq)) %>%
  spread(Generation, Freq) %>%
  mutate(Diff = Son - Father) %>%
  mutate(PercentageDiff=Diff/sum(Father + Son)) %>%
  mutate(Occupation=factor(Occupation, levels = c("UpNM", "LoNM", "UpM", "LoM", "Farm"))) %>%
  ggplot(aes(x=Occupation, y=PercentageDiff, fill=PercentageDiff)) +
  geom_bar(stat="identity") +
  facet_grid(. ~ Country) +
  xlab("") + ylab("") +
  scale_y_continuous(labels=percent) +
  scale_fill_gradient(low="blue", high="red") +
  theme_linedraw()
```



The y axis of this bar chart is the difference of `Freq` between `Father` and `Son` in percentage, positive number means more people for variable `Father`, negative number means the opposite. The x axis represents the occupations. There is a general trend for all these three regions which is the percetage of people who work in more tranditional areas reduces and the percetage of people who work in more "advanced", or say non-manual areas increases for `Son` compare to `Father`. The change of occupation structure reflects the developments in economy and technology at 60s and early 70s.

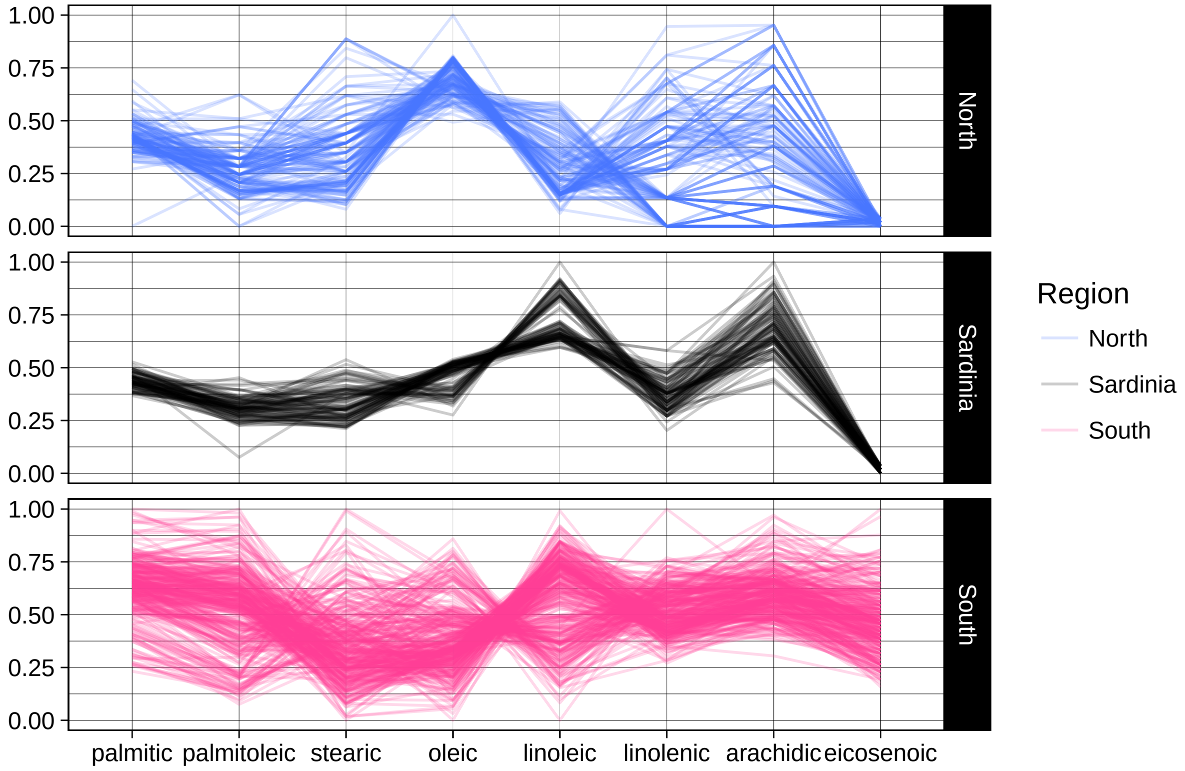
Olives Dataset

```
olives %>%
  ggpairs(columns=3:10,
          lower=list(continuous=wrap("points", alpha=0.3, size=0.1)),
          diag=list(continuous=wrap("densityDiag", alpha=0.7, size=0.1), axisLabels='none'),
          upper=list(continuous=wrap("cor", size=1.8)),
          ggplot2::aes(colour=Region)
          ) +
  theme_linedraw() +
  theme(
    text=element_text(size = 7),
    axis.text=element_text(size = 6),
    axis.text.x=element_text(angle = 30))
```



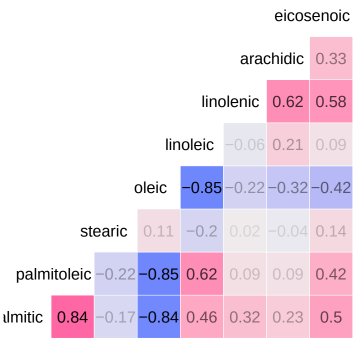
After coloring `olives` data by `region`, the evidences of difference pattern among regions show up. The most obvious one is in the last row, which represents `eicosenoic` on x-axis, all red and green points (region `North` and `Sardinia`) locate at the bottom of y-axis like a line, however the blue points (`South`) have total different pattern. Actually not only for this one feature, most of the features have distinct pattern in each `region`. Parallel coordinate plot is one good way to show such differences.

```
olives %>%
  ggparcoord(columns=3:10, alphaLines=0.2, scale="uniminmax", groupColumn="Region") +
  xlab("") + ylab("") +
  ggtitle("") +
  facet_grid(Region ~.) +
  # coord_flip() +
  scale_colour_manual(values = c("royalblue1", "black", "violetred1")) +
  theme_linedraw() +
  theme(panel.spacing.x=unit(0.8, "lines"))
```

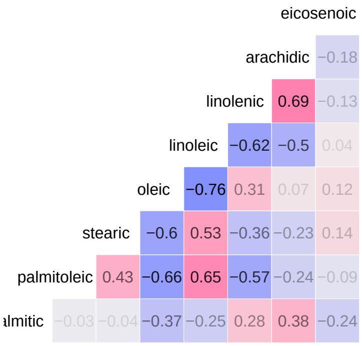


```
p1 <- olives %>%
  select(-Area, -Region, -Test.Training) %>%
  ggcorr(label = TRUE, label_size = 2, label_round = 2, label_alpha = TRUE, low = "royalblue1", mid = "
  ggtitle("All Regions")

p2 <- olives %>%
  filter(Region=="North") %>%
  select(-Area, -Region, -Test.Training) %>%
  ggcorr(label = TRUE, label_size = 2, label_round = 2, label_alpha = TRUE, low = "royalblue1", mid = "
  ggtitle("North")

p3 <- olives %>%
  filter(Region=="Sardinia") %>%
  select(-Area, -Region, -Test.Training) %>%
  ggcorr(label = TRUE, label_size = 2, label_round = 2, label_alpha = TRUE, low = "royalblue1", mid = "
  ggtitle("Sardinia")

p4 <- olives %>%
  filter(Region=="South") %>%
  select(-Area, -Region, -Test.Training) %>%
  ggcorr(label = TRUE, label_size = 2, label_round = 2, label_alpha = TRUE, low = "royalblue1", mid = "
  ggtitle("South")

grid.arrange(p1,p2,p3,p4)
```

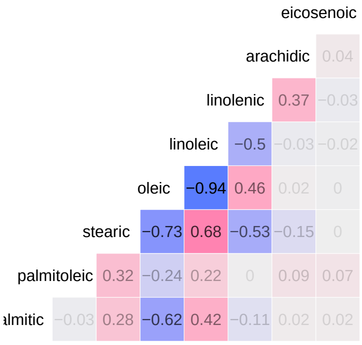
All Regions



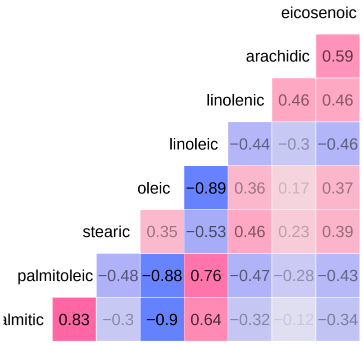
North



Sardinia



South



Furthermore, I noticed that not all the combination of variables has linear relations and I presented the evidence by showing the actual value of correlation and highlighting with color and alpha.