STAT547_LAB_3

Zebosi Brian

9/9/2021

STAT547: Lab_3 Plotting Multivariate Data

Lab3 Purpose

- To learn how to produce plots of multivariate data, both static and interactive.
- Practice describing the patterns and relationships observed in the multivariate plots.

Music Clips Data

The music clips data is posted in music-plusnew-sub.csv. The data file has five quantitative variables contining audio information from 62 songs. The first two columns (Artist, Type) describe the artist and type of music. The raw data come from a time series for the sound produced by each music clip (track). For each time series the variance of amplitude, average amplitude, maximum amplitude, and two additional variables calculated from the spectral decomposition of the time series are calculated. The Type variable classifies the tracks as either Rock, Classical or New Wave, and there are 5 tracks that are not identified. Read the data into a data frame, indicating that the row names are in column 1 of the data file and that column is not a variable. The stringAsFactors=FALSE option prevents the first column from being converted to a factor.

Obtain information on the dimensions of the data frame. Also list the column names. List the first six columns of data.

```
setwd("D:/lecture_notes/STAT_575_Multivariate_Data_Analysis_BZ/Labs/Lab_3")
music <- read.csv("music-plusnew-sub.csv", row.names = 1, stringsAsFactors = FALSE)
colnames(music)
```

```
## [1] "Artist" "Type"   "LVar"   "LAve"   "LMax"   "LFEner" "LFreq"
```

```
dim(music)
```

```
## [1] 62  7
```

```
str(music)
```

```
## 'data.frame':    62 obs. of  7 variables:
##  $ Artist: chr  "Abba" "Abba" "Abba" "Abba" ...
##  $ Type  : chr  "Rock" "Rock" "Rock" "Rock" ...
##  $ LVar  : num  17600756 9543021 9049482 7557437 6282286 ...
##  $ LAve  : num  -90 -75.8 -98.1 -90.5 -89 ...
##  $ LMax  : int  29921 27626 26372 28898 27940 25531 14699 8928 22962 15517 ...
##  $ LFEner: num  106 103 102 102 100 ...
##  $ LFreq : num  59.6 58.5 124.6 48.8 74 ...
```

Compute summary statistics

```
summary(music)
```

```
##     Artist              Type                LVar                LAve
```

```
##  Length:62          Length:62          Min.   :   293608   Min.   :-98.063
##  Class :character   Class :character   1st Qu.:  2844213   1st Qu.: -6.253
##  Mode  :character   Mode  :character   Median :  8210359   Median : -5.662
##                                        Mean   : 19951792   Mean   : -7.807
##                                        3rd Qu.: 24547475   3rd Qu.:  1.962
##                                        Max.   :129472199   Max.   :216.232
##       LMax           LFEner           LFreq
##  Min.   : 2985   Min.   : 83.88   Min.   : 41.41
##  1st Qu.:16200   1st Qu.:101.69   1st Qu.: 99.18
##  Median :24431   Median :104.35   Median :175.29
##  Mean   :22486   Mean   :104.03   Mean   :231.39
##  3rd Qu.:29919   3rd Qu.:108.15   3rd Qu.:315.12
##  Max.   :32766   Max.   :114.00   Max.   :877.77
```

Compute a table of counts for each type of music.

```
table(music$Type)
```

```
##
## Classical  New wave      Rock
##        24         3        30
```

Compute a table of counts for each artist

```
table(music$Artist)
```

```
##
##      Abba   Beatles Beethoven      Eels      Enya    Mozart   Vivaldi
##        10        10         8        10         3         6        10
```

Bodyfat data

```
bodyfat <- read.csv("bodyfat.csv")
dim(bodyfat)
```

```
## [1] 252  14
```

```
head(bodyfat)
```

```
##   Percent.Body.Fat Ageyrs Weightlbs Heightinches NeckCircm ChestCircm
## 1             12.3     23    154.25        67.75      36.2       93.1
## 2              6.1     22    173.25        72.25      38.5       93.6
## 3             25.3     22    154.00        66.25      34.0       95.8
## 4             10.4     26    184.75        72.25      37.4      101.8
## 5             28.7     24    184.25        71.25      34.4       97.3
## 6             20.9     24    210.25        74.75      39.0      104.5
##   AbdomenCircm HipCircm ThighCircm KneeCircm AnkleCircm BicepsCircm
## 1         85.2     94.5       59.0      37.3       21.9        32.0
## 2         83.0     98.7       58.7      37.3       23.4        30.5
## 3         87.9     99.2       59.6      38.9       24.0        28.8
## 4         86.4    101.2       60.1      37.3       22.8        32.4
## 5        100.0    101.9       63.2      42.2       24.0        32.2
## 6         94.4    107.8       66.0      42.0       25.6        35.7
##   ForearmCircm WristCircm
## 1         27.4       17.1
## 2         28.9       18.2
## 3         25.2       16.6
```

```
## 4         29.4       18.2
## 5         27.7       17.7
## 6         30.6       18.8
```

```
str(bodyfat)
```

```
## 'data.frame':    252 obs. of  14 variables:
##  $ Percent.Body.Fat: num  12.3 6.1 25.3 10.4 28.7 20.9 19.2 12.4 4.1 11.7 ...
##  $ Ageyrs          : int  23 22 22 26 24 24 26 25 25 23 ...
##  $ Weightlbs       : num  154 173 154 185 184 ...
##  $ Heightinches    : num  67.8 72.2 66.2 72.2 71.2 ...
##  $ NeckCircm       : num  36.2 38.5 34 37.4 34.4 39 36.4 37.8 38.1 42.1 ...
##  $ ChestCircm      : num  93.1 93.6 95.8 101.8 97.3 ...
##  $ AbdomenCircm    : num  85.2 83 87.9 86.4 100 94.4 90.7 88.5 82.5 88.6 ...
##  $ HipCircm        : num  94.5 98.7 99.2 101.2 101.9 ...
##  $ ThighCircm      : num  59 58.7 59.6 60.1 63.2 66 58.4 60 62.9 63.1 ...
##  $ KneeCircm       : num  37.3 37.3 38.9 37.3 42.2 42 38.3 39.4 38.3 41.7 ...
##  $ AnkleCircm      : num  21.9 23.4 24 22.8 24 25.6 22.9 23.2 23.8 25 ...
##  $ BicepsCircm     : num  32 30.5 28.8 32.4 32.2 35.7 31.9 30.5 35.9 35.6 ...
##  $ ForearmCircm    : num  27.4 28.9 25.2 29.4 27.7 30.6 27.8 29 31.1 30 ...
##  $ WristCircm      : num  17.1 18.2 16.6 18.2 17.7 18.8 17.7 18.8 18.2 19.2 ...
```

```
summary(bodyfat)
```

```
##  Percent.Body.Fat      Ageyrs          Weightlbs       Heightinches
##  Min.   : 0.00    Min.   :22.00   Min.   :118.5   Min.   :29.50
##  1st Qu.:12.47    1st Qu.:35.75   1st Qu.:159.0   1st Qu.:68.25
##  Median :19.20    Median :43.00   Median :176.5   Median :70.00
##  Mean   :19.15    Mean   :44.88   Mean   :178.9   Mean   :70.15
##  3rd Qu.:25.30    3rd Qu.:54.00   3rd Qu.:197.0   3rd Qu.:72.25
##  Max.   :47.50    Max.   :81.00   Max.   :363.1   Max.   :77.75
##    NeckCircm       ChestCircm      AbdomenCircm       HipCircm
##  Min.   :31.10   Min.   : 79.30   Min.   : 69.40   Min.   : 85.0
##  1st Qu.:36.40   1st Qu.: 94.35   1st Qu.: 84.58   1st Qu.: 95.5
##  Median :38.00   Median : 99.65   Median : 90.95   Median : 99.3
##  Mean   :37.99   Mean   :100.82   Mean   : 92.56   Mean   : 99.9
##  3rd Qu.:39.42   3rd Qu.:105.38   3rd Qu.: 99.33   3rd Qu.:103.5
##  Max.   :51.20   Max.   :136.20   Max.   :148.10   Max.   :147.7
##    ThighCircm      KneeCircm       AnkleCircm    BicepsCircm     ForearmCircm
##  Min.   :47.20   Min.   :33.00   Min.   :19.1   Min.   :24.80   Min.   :21.00
##  1st Qu.:56.00   1st Qu.:36.98   1st Qu.:22.0   1st Qu.:30.20   1st Qu.:27.30
##  Median :59.00   Median :38.50   Median :22.8   Median :32.05   Median :28.70
##  Mean   :59.41   Mean   :38.59   Mean   :23.1   Mean   :32.27   Mean   :28.66
##  3rd Qu.:62.35   3rd Qu.:39.92   3rd Qu.:24.0   3rd Qu.:34.33   3rd Qu.:30.00
##  Max.   :87.30   Max.   :49.10   Max.   :33.9   Max.   :45.00   Max.   :34.90
##    WristCircm
##  Min.   :15.80
##  1st Qu.:17.60
##  Median :18.30
##  Mean   :18.23
##  3rd Qu.:18.80
##  Max.   :21.40
```

PISA Data

```
pisamath <- read.csv("pisamathmeans.csv")
dim(pisamath)
```

```
## [1] 10294     8
```

```
head(pisamath)
```

```
##   Gender      acc      acq      acs      acu      ape      apf      api
## 1 Female 432.0844 456.6989 444.5474 405.7564 441.1201 439.0949 415.8825
## 2 Female 422.0362 450.5453 457.3999 457.8673 448.2084 432.0065 411.5984
## 3 Female 527.0369 563.4912 554.9229 489.3364 516.9107 538.8768 519.5591
## 4 Female 436.2128 487.6227 449.4548 489.9595 492.6079 461.2946 435.4339
## 5   Male 631.1030 572.9942 563.8028 580.7836 627.6757 643.2544 580.1604
## 6 Female 424.3730 443.8464 390.5671 424.5287 456.9326 411.7542 441.9769
```

```
str(pisamath)
```

```
## 'data.frame':    10294 obs. of  8 variables:
##  $ Gender: chr  "Female" "Female" "Female" "Female" ...
##  $ acc   : num  432 422 527 436 631 ...
##  $ acq   : num  457 451 563 488 573 ...
##  $ acs   : num  445 457 555 449 564 ...
##  $ acu   : num  406 458 489 490 581 ...
##  $ ape   : num  441 448 517 493 628 ...
##  $ apf   : num  439 432 539 461 643 ...
##  $ api   : num  416 412 520 435 580 ...
```

```
summary(pisamath)
```

```
##     Gender               acc             acq             acs       
##  Length:10294       Min.   :162.8   Min.   :115.1   Min.   :115.8  
##  Class :character   1st Qu.:426.1   1st Qu.:412.2   1st Qu.:399.9  
##  Mode  :character   Median :494.2   Median :483.6   Median :467.5  
##                     Mean   :497.3   Mean   :483.0   Mean   :471.1  
##                     3rd Qu.:566.1   3rd Qu.:552.5   3rd Qu.:537.8  
##                     Max.   :827.2   Max.   :792.6   Max.   :809.7  
##                     NA's   :4978    NA's   :4978    NA's   :4978   
##       acu             ape             apf              api        
##  Min.   :165.2   Min.   :153.7   Min.   : 93.87   Min.   : 97.76  
##  1st Qu.:431.9   1st Qu.:420.8   1st Qu.:411.89   1st Qu.:428.03  
##  Median :495.6   Median :486.4   Median :480.07   Median :497.05  
##  Mean   :497.6   Mean   :487.5   Mean   :484.73   Mean   :498.71  
##  3rd Qu.:559.9   3rd Qu.:551.8   3rd Qu.:553.87   3rd Qu.:566.00  
##  Max.   :785.0   Max.   :778.7   Max.   :841.34   Max.   :837.44  
##  NA's   :4978    NA's   :4978    NA's   :4978     NA's   :4978
```

Exercise 1

```
library(ggplot2)
```

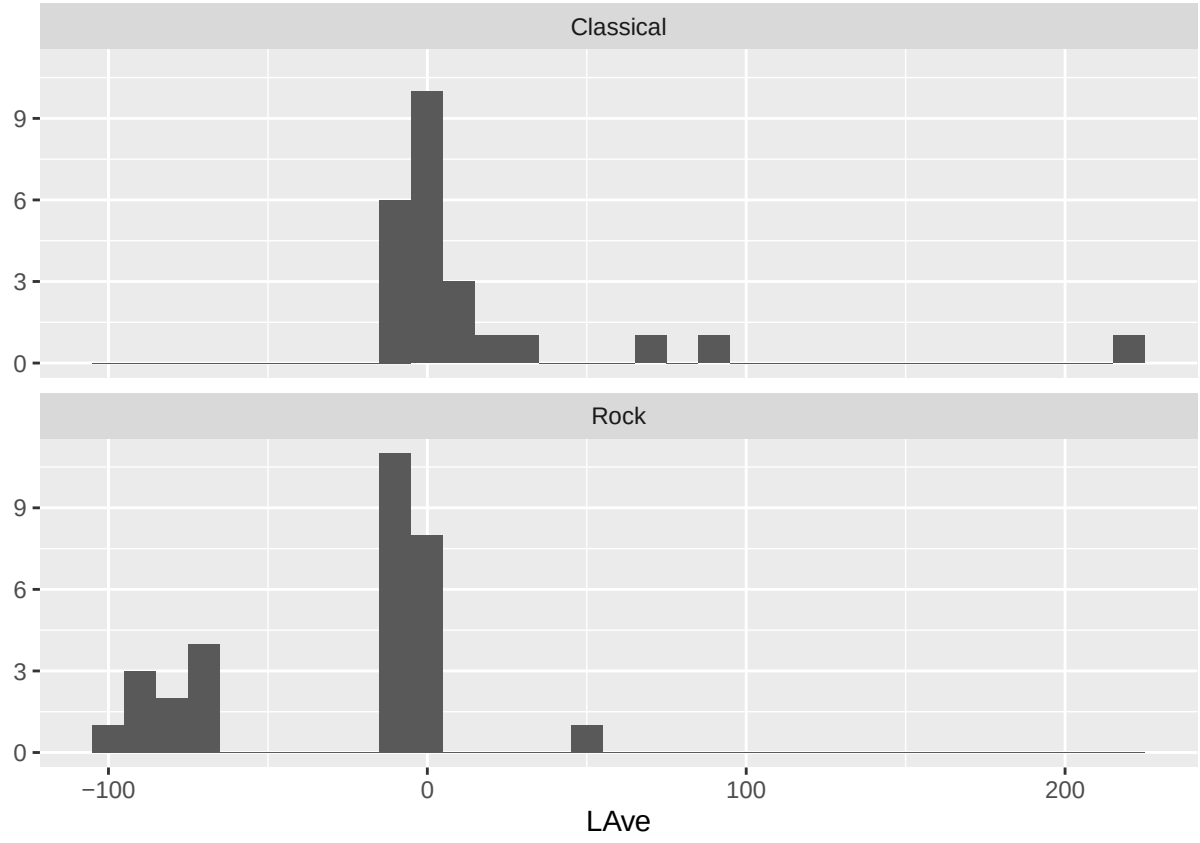
```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```
library(GGally)
```

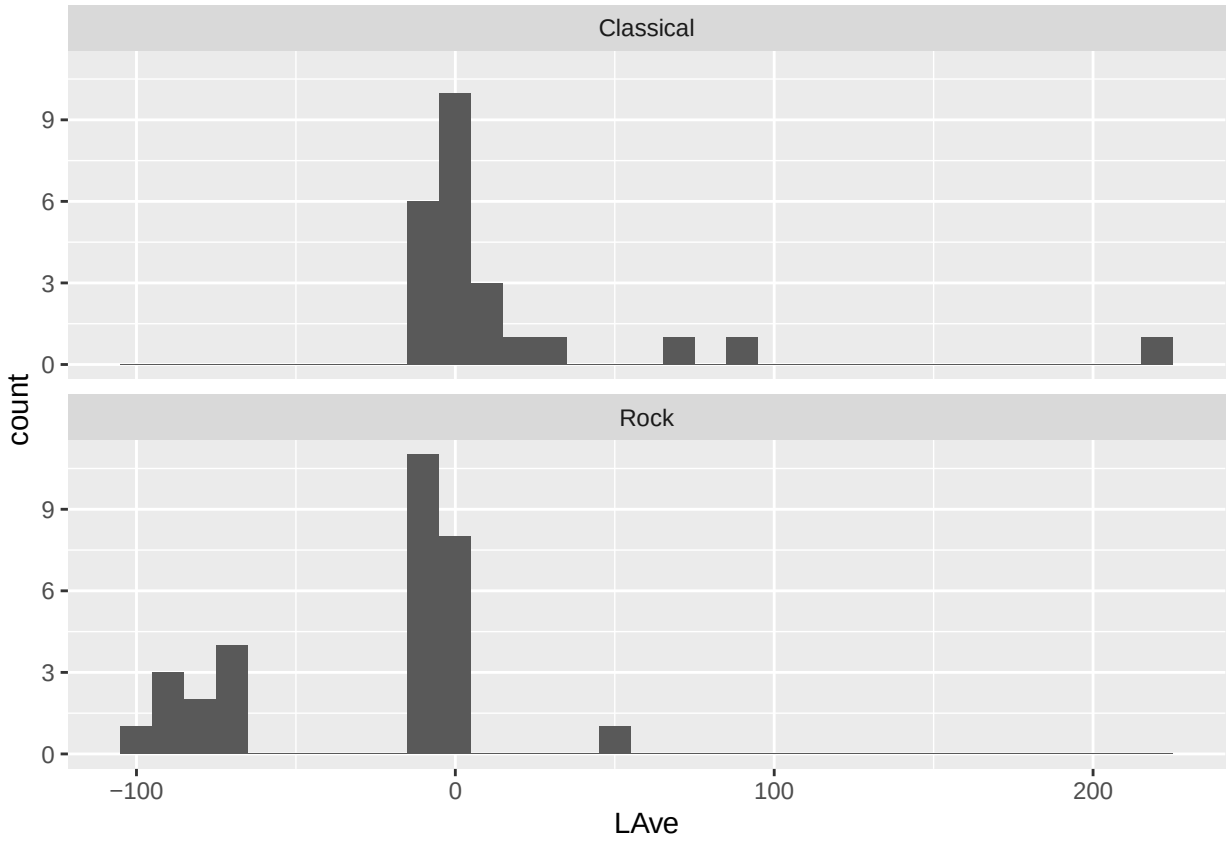
```
## Warning: package 'GGally' was built under R version 4.0.5
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

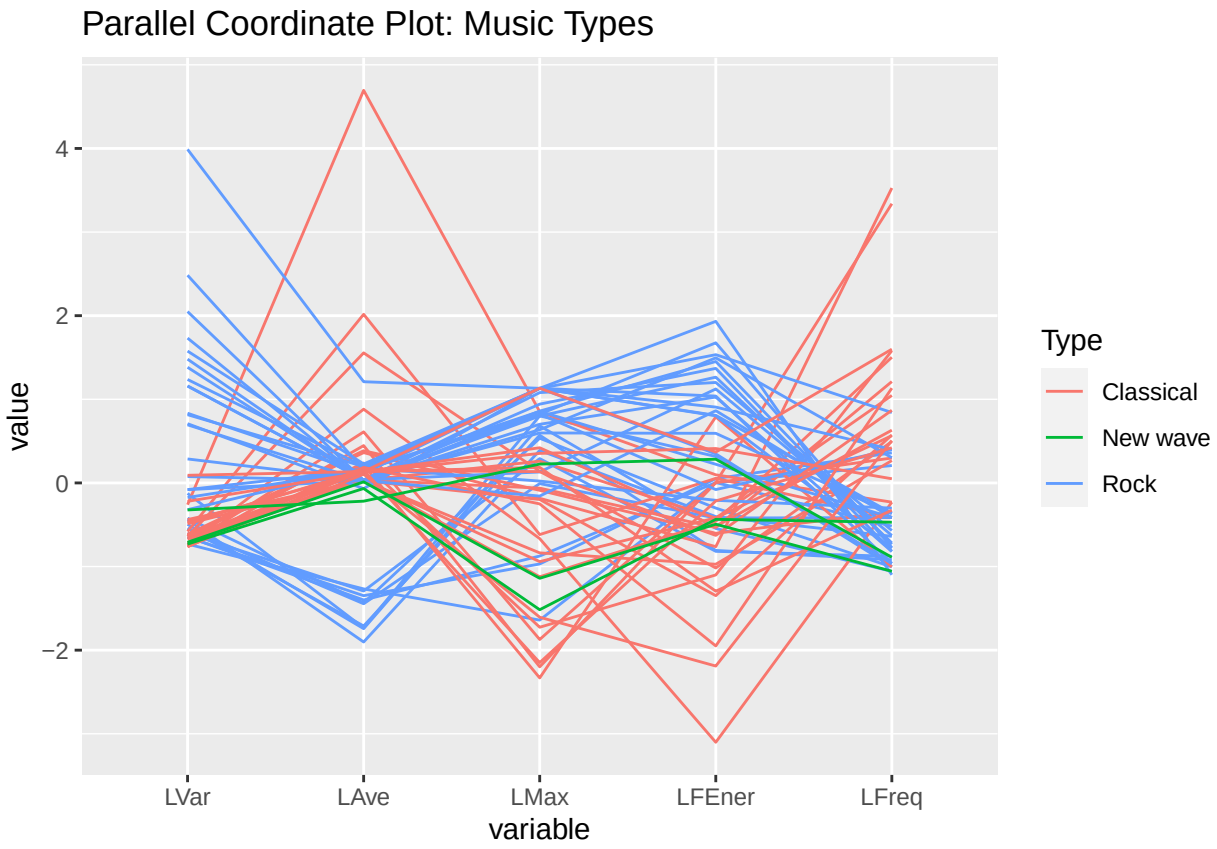
```r
music.sub <- subset(music, Type == "Rock" | Type == "Classical")

# Histogram
# Short version
qplot(LAve, data = music.sub, geom = "histogram", binwidth = 10) +
  facet_wrap(~ Type, ncol = 1)
```



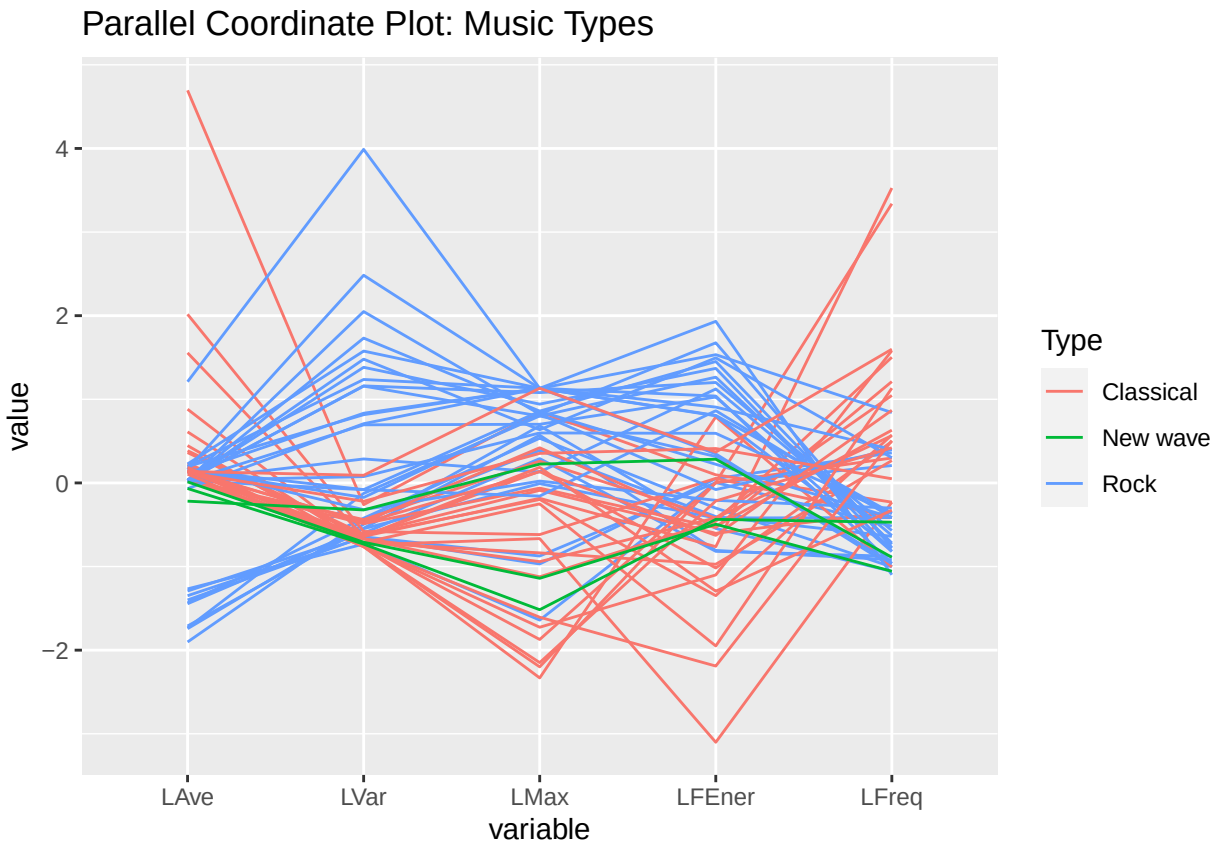

```r
# Full version
ggplot(music.sub, aes(x = LAve)) +
  geom_histogram(binwidth = 10) +
  facet_wrap(~ Type, ncol = 1)
```



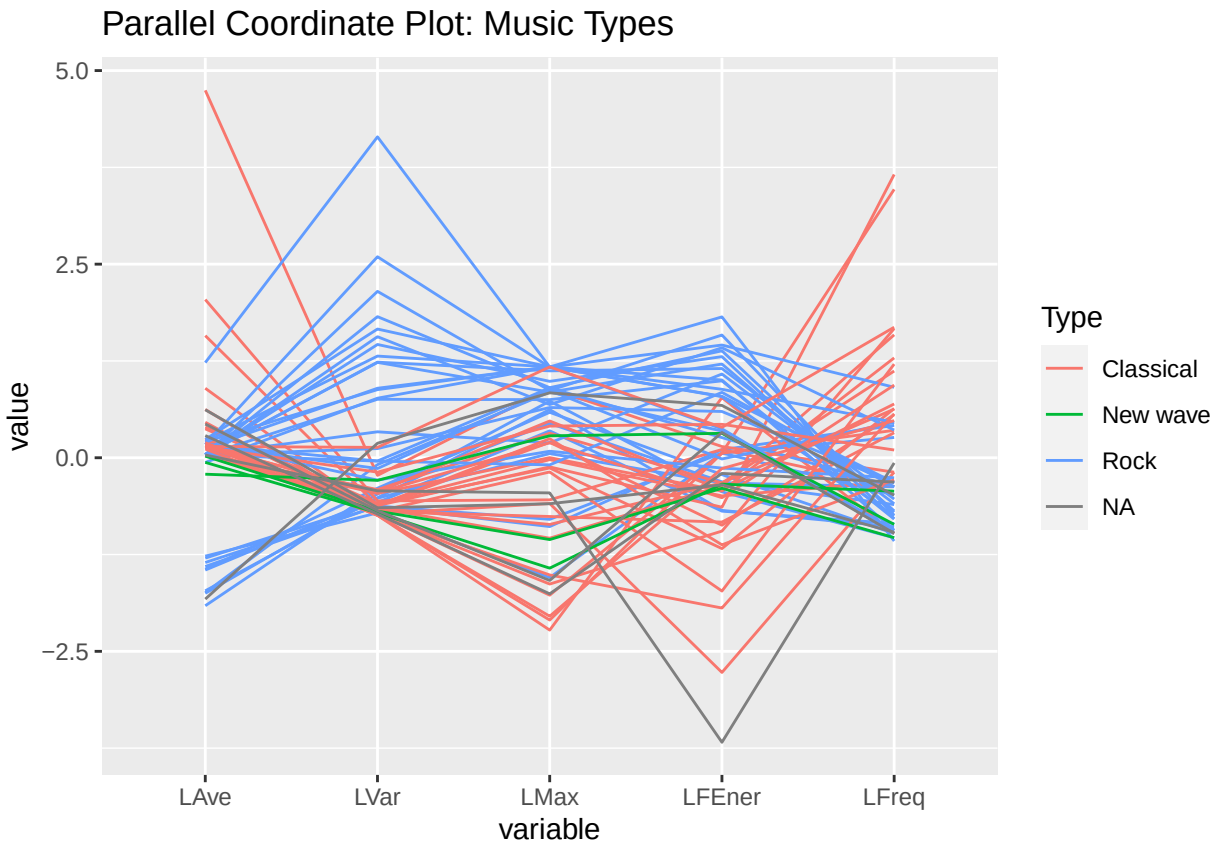

```
# Parallel coordinate plot
music.sub2 <- subset(music, Type == "Rock" | Type == "Classical" | Type == "New wave")
ggparcoord(music.sub2, columns = 3:7, groupColumn = "Type",
           title = "Parallel Coordinate Plot: Music Types")
```




```
ggparcoord(music.sub2, columns = c(4, 3, 5, 6, 7), groupColumn = "Type",
           title = "Parallel Coordinate Plot: Music Types")
```




```
ggparcoord(music, columns = c(4, 3, 5, 6, 7), groupColumn = "Type", missing = "exclude",
           title = "Parallel Coordinate Plot: Music Types")
```



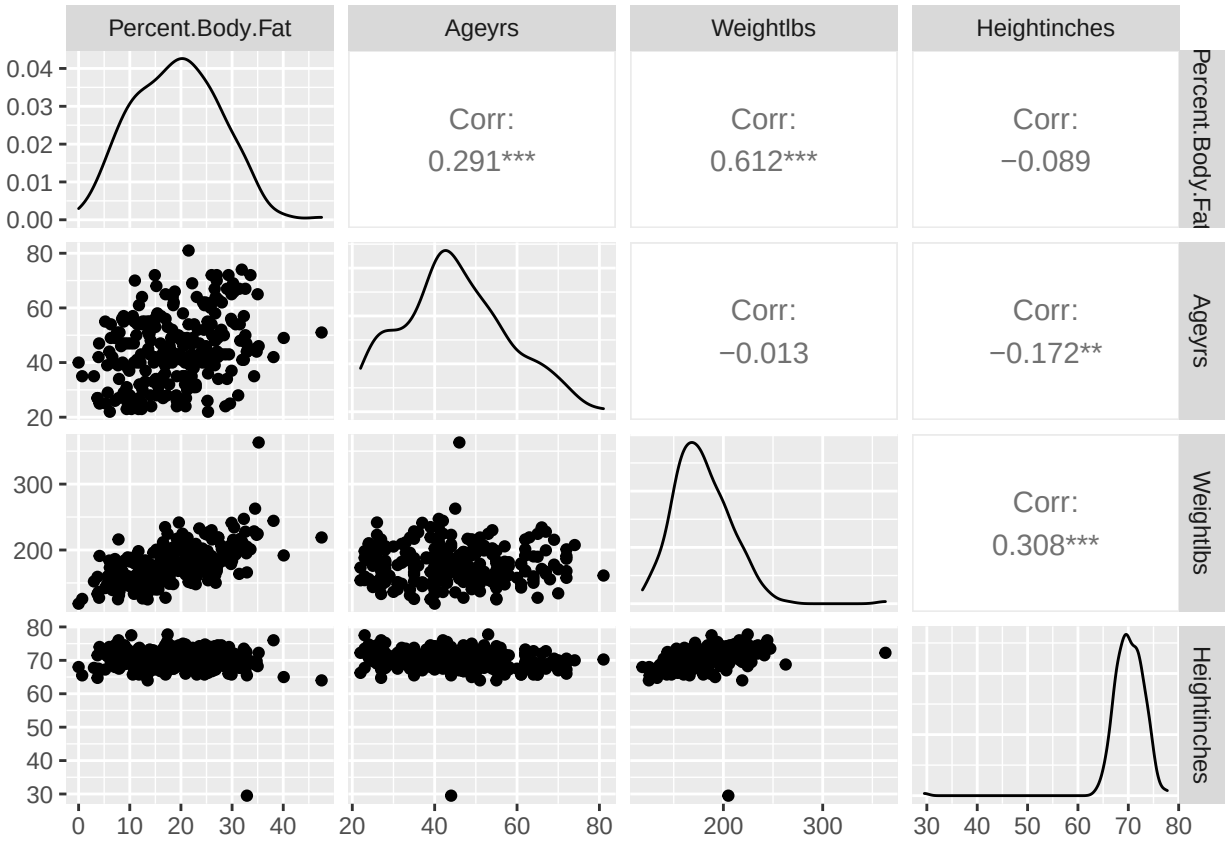
(a)

(b)

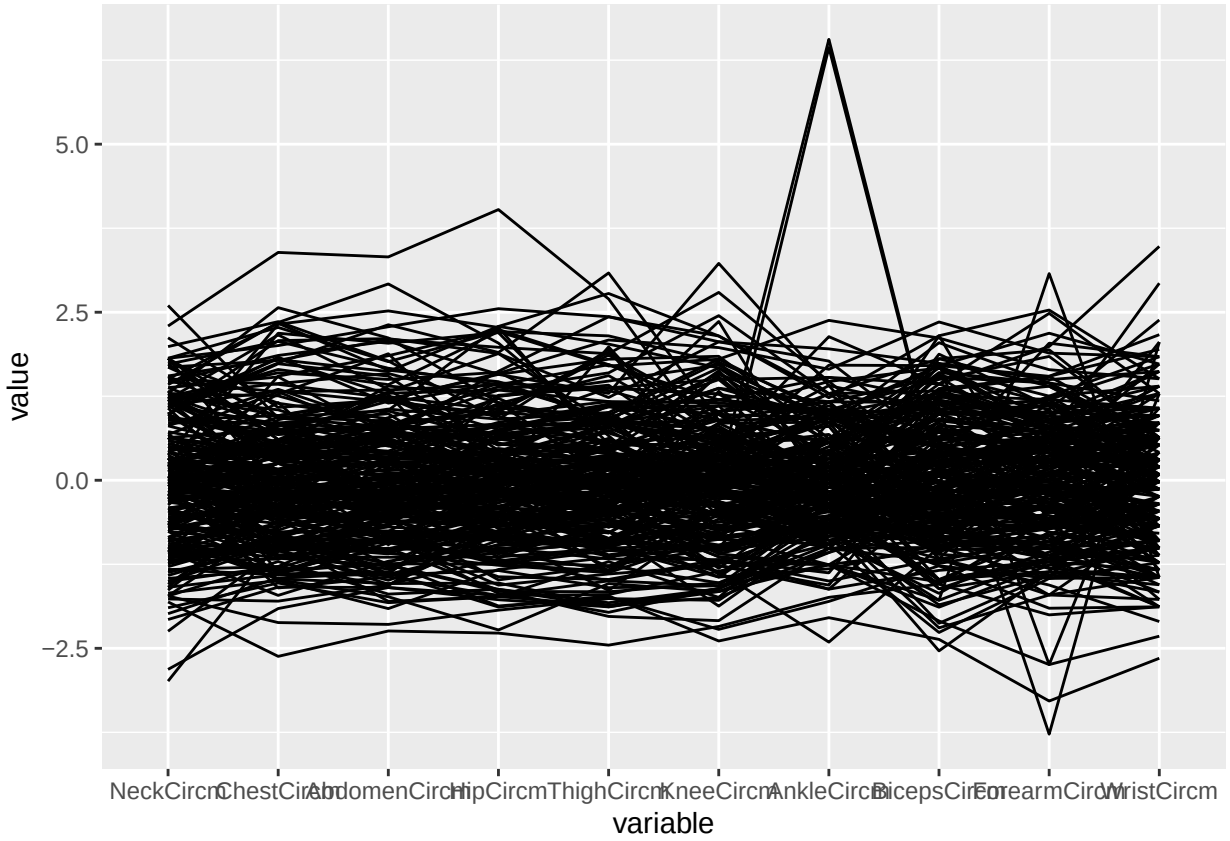
(c)

Exercise 2

```
# Scatterplot matrix
ggpairs(bodyfat[, 1:4])
```




```
bf.sub <- subset(bodyfat, Heightinches > 60 & Weightlbs < 300)

# Parallel coordinate plot
ggparcoord(bf.sub, 5:14)
```




```
summary(bf.sub$AnkleCircm)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.10   22.00   22.80   23.07   24.00   33.90
```

```
sort(bf.sub$AnkleCircm)
```

```
##   [1] 19.1 19.7 20.1 20.2 20.4 20.4 20.5 20.6 20.8 20.9 21.0 21.0 21.0 21.0 21.0
##  [16] 21.0 21.3 21.3 21.4 21.4 21.4 21.4 21.5 21.5 21.5 21.5 21.5 21.5 21.5 21.5
##  [31] 21.6 21.6 21.6 21.7 21.7 21.7 21.7 21.8 21.8 21.8 21.8 21.8 21.8 21.8 21.8
##  [46] 21.8 21.8 21.9 21.9 21.9 21.9 21.9 21.9 21.9 21.9 22.0 22.0 22.0 22.0 22.0
##  [61] 22.0 22.0 22.0 22.0 22.0 22.0 22.1 22.1 22.1 22.1 22.1 22.1 22.1 22.2 22.2
##  [76] 22.2 22.2 22.3 22.3 22.3 22.3 22.3 22.3 22.3 22.4 22.4 22.4 22.4 22.4 22.4
##  [91] 22.4 22.4 22.5 22.5 22.5 22.5 22.5 22.5 22.5 22.5 22.5 22.5 22.5 22.6 22.6
## [106] 22.6 22.6 22.6 22.6 22.6 22.6 22.6 22.6 22.6 22.7 22.7 22.7 22.7 22.7 22.7
## [121] 22.7 22.7 22.7 22.8 22.8 22.8 22.8 22.8 22.9 22.9 22.9 22.9 22.9 22.9 22.9
## [136] 23.0 23.0 23.0 23.0 23.0 23.0 23.1 23.1 23.1 23.1 23.1 23.2 23.2 23.2 23.2
## [151] 23.2 23.2 23.2 23.2 23.2 23.3 23.3 23.3 23.3 23.4 23.4 23.4 23.4 23.4 23.4
## [166] 23.4 23.4 23.5 23.5 23.5 23.5 23.6 23.6 23.6 23.6 23.7 23.7 23.7 23.8 23.8
## [181] 23.8 23.8 23.8 23.9 23.9 24.0 24.0 24.0 24.0 24.0 24.0 24.0 24.0 24.1 24.1
## [196] 24.1 24.1 24.1 24.1 24.2 24.4 24.4 24.5 24.5 24.5 24.5 24.6 24.6 24.6 24.6
## [211] 24.6 24.6 24.6 24.6 24.7 24.7 24.7 24.7 24.7 24.7 24.7 24.8 24.8 24.8 24.8
## [226] 24.8 24.9 25.0 25.0 25.0 25.0 25.1 25.2 25.2 25.2 25.2 25.4 25.4 25.5 25.5
## [241] 25.6 25.6 25.8 25.9 26.0 26.3 26.6 27.0 33.7 33.9
```

```
bf.sub[bf.sub$AnkleCircm > 33, ]
```

```
##    Percent.Body.Fat Ageyrs Weightlbs Heightinches NeckCircm ChestCircm
```

```
## 31             11.9     32       182        73.75      38.7      100.5
## 86             26.6     67       167        67.50      36.5       98.9
##    AbdomenCircm HipCircm ThighCircm KneeCircm AnkleCircm BicepsCircm
## 31         88.7     99.8       57.5      38.7       33.9        32.5
## 86         89.7     96.2       54.7      37.8       33.7        32.4
##    ForearmCircm WristCircm
## 31         27.7       18.4
## 86         27.7       18.2
```

(a)

(b)

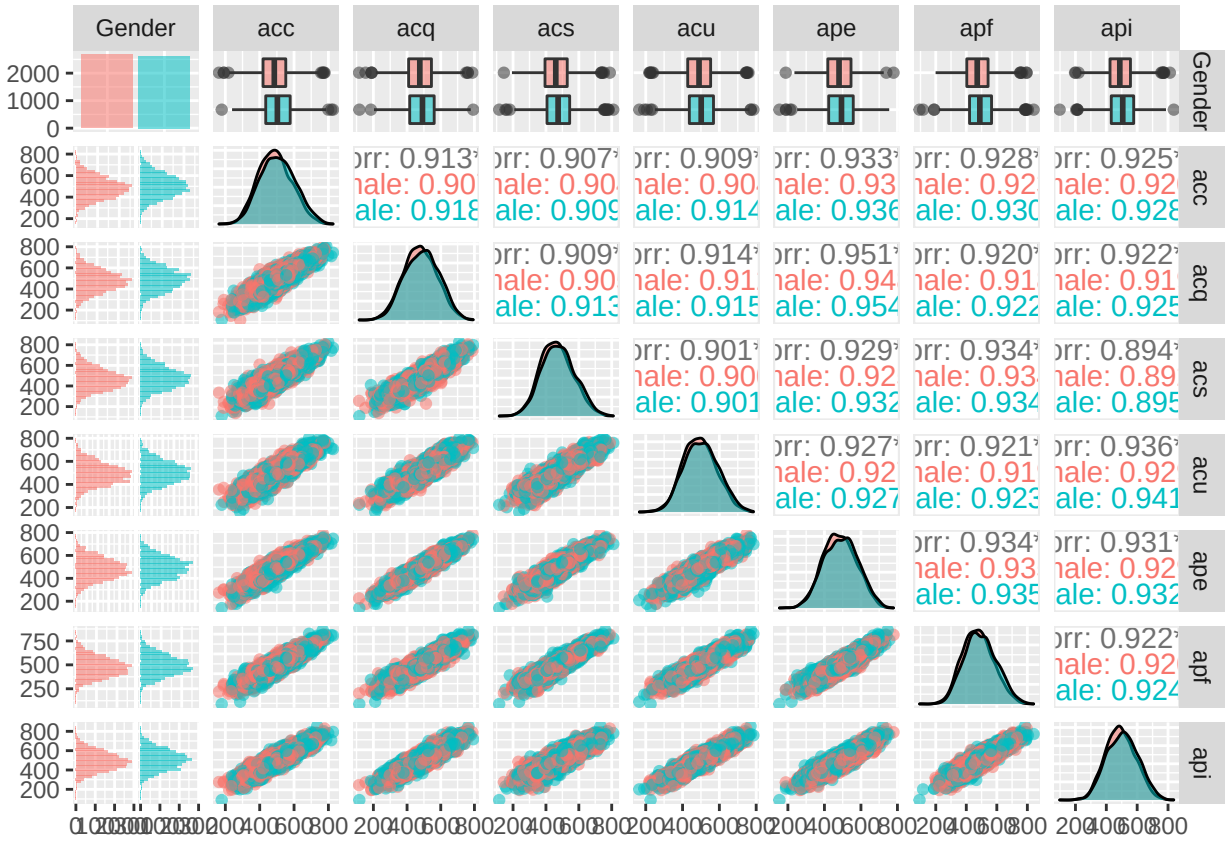
(c)

(d)

Exercise 3

```
# Scatterplot matrix
pisamath2 <- subset(pisamath, !is.na(acc))
ggpairs(pisamath2, mapping = aes(color = Gender, alpha = 0.5))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Gender
acc
acq
acs
acu
ape
apf
api
2000
1000
0
800
600
400
200
750
500
250
0
200
400
600
800

(a)

(b)