# Lab 3 Plotting Multivariate Data

Report due: Friday, September 10, 2020, by 11:59 pm.

Electronically submit one solution per group, including the Rmd file and the output file (either pdf or html). Please record the names of the group members who attended the lab.

# 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.

The following four data sets are examined in this lab.

### 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.

music <- read.csv("music-plusnew-sub.csv", row.names=1, stringsAsFactors=FALSE)

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

dim(music)

## [1] 62 7

colnames(music)

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

head(music)

## Artist Type LVar LAve LMax LFEner LFreq  
## Dancing Queen Abba Rock 17600756 -90.00687 29921 105.9210 59.57379  
## Knowing Me Abba Rock 9543021 -75.76672 27626 102.8362 58.48031  
## Take a Chance Abba Rock 9049482 -98.06292 26372 102.3249 124.59397  
## Mamma Mia Abba Rock 7557437 -90.47106 28898 101.6165 48.76513  
## Lay All You Abba Rock 6282286 -88.95263 27940 100.3008 74.02039  
## Super Trouper Abba Rock 4665867 -69.02084 25531 100.2485 81.40140

Compute summary statistics

summary(music)

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

Comput 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

The body fat data bodyfat.csv contains various measurements on 14 variables for 252 typical American males.

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

## [1] 252 14

colnames(bodyfat)

## [1] "Percent.Body.Fat" "Ageyrs" "Weightlbs"   
## [4] "Heightinches" "NeckCircm" "ChestCircm"   
## [7] "AbdomenCircm" "HipCircm" "ThighCircm"   
## [10] "KneeCircm" "AnkleCircm" "BicepsCircm"   
## [13] "ForearmCircm" "WristCircm"

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

### PISA Data

Mathematics test scores for 15 year olds. Only the USA measurements are examined. These are the scores for the different types of math skills.

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

## [1] 10294 8

summary(pisamath)

## Gender acc acq acs   
## Female:5142 Min. :162.8 Min. :115.1 Min. :115.8   
## Male :5152 1st Qu.:426.1 1st Qu.:412.2 1st Qu.:399.9   
## 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

# Software

We will examine some aspects of the ggplot2 package and GGally package in R. The main web page for ggplot2 is <http://had.co.nz/ggplot2>, but this page <http://www.cookbook-r.com/Graphs/> has very good practical examples.

Only the manual pages exist for help for the GGally package. To install the GGally package, run install.packages("GGalley") in the Console window.

Attach the functions in the ggplot2 and the GGally packages.

library(ggplot2)  
library(GGally)

# Exercise 1 Music clips data:

This exercise uses the qplot function in the ggplot2 package to make a panel of histograms and a scatterplot. First select a subset of the data that contains only classical and rock music.

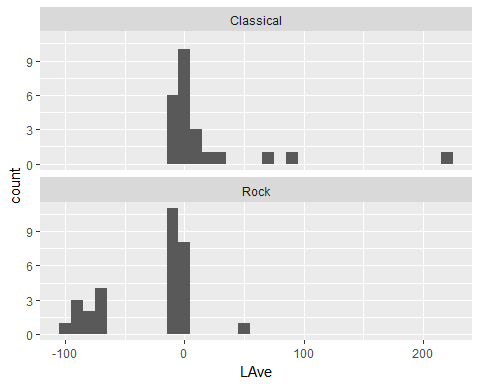
table(music$Type)

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

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

A. For classical and rock music make histograms for the avergae amplitude variable (LAve) facetted by Type. Set the binwidth to units of 10. How do the distributions of average amplitude values differ between classical and rock music?

qplot(LAve, data=music.sub, geom="histogram", binwidth=10.0) + facet\_wrap(~Type, ncol=1)



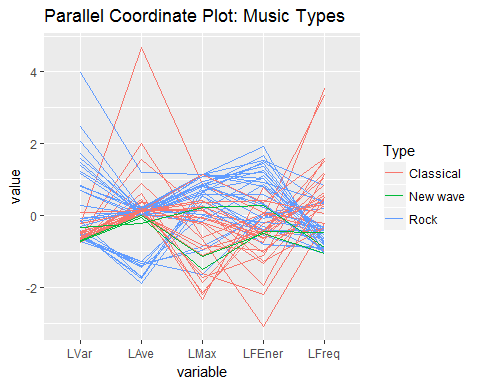
B. Make a scatterplot of LVar vs LAve, with points colored by the type of music. Describe differences between the patterns of the points on the plot corresponding to Rock and Classical music.

Select three mucis tpes. The othe songs have missing values for the music type.

music.sub2 <- subset(music, Type == "Rock" | Type == "Classical" | Type=="New wave")

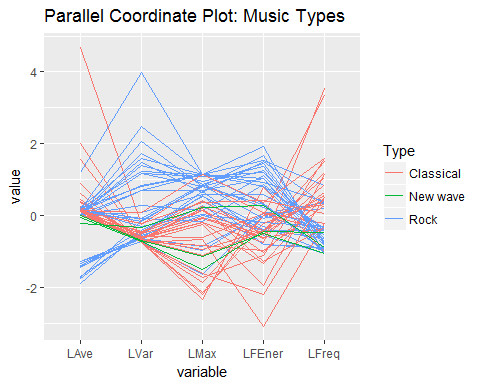
Make a parallel coordinate plot

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



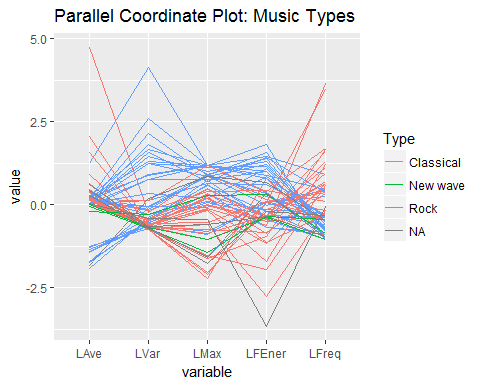
Reorder how the varibales appear on the plot.

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



The following plot includes a category called "NA" for for songs for which the song type is missing.

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

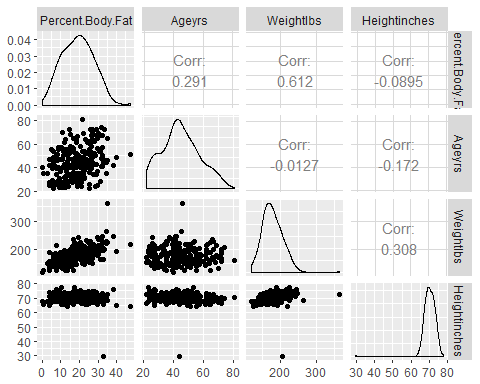


1. Describe how the types of music differ. Is there evidence that rock music should be subdivided into two subtypes? Explain

# Exercise 2 Body fat data: Using GGally.

This exercise uses the ggpairs function in GGally to create a scatterplot matrix for the first four columns of data. This display uses all cases in the data set.

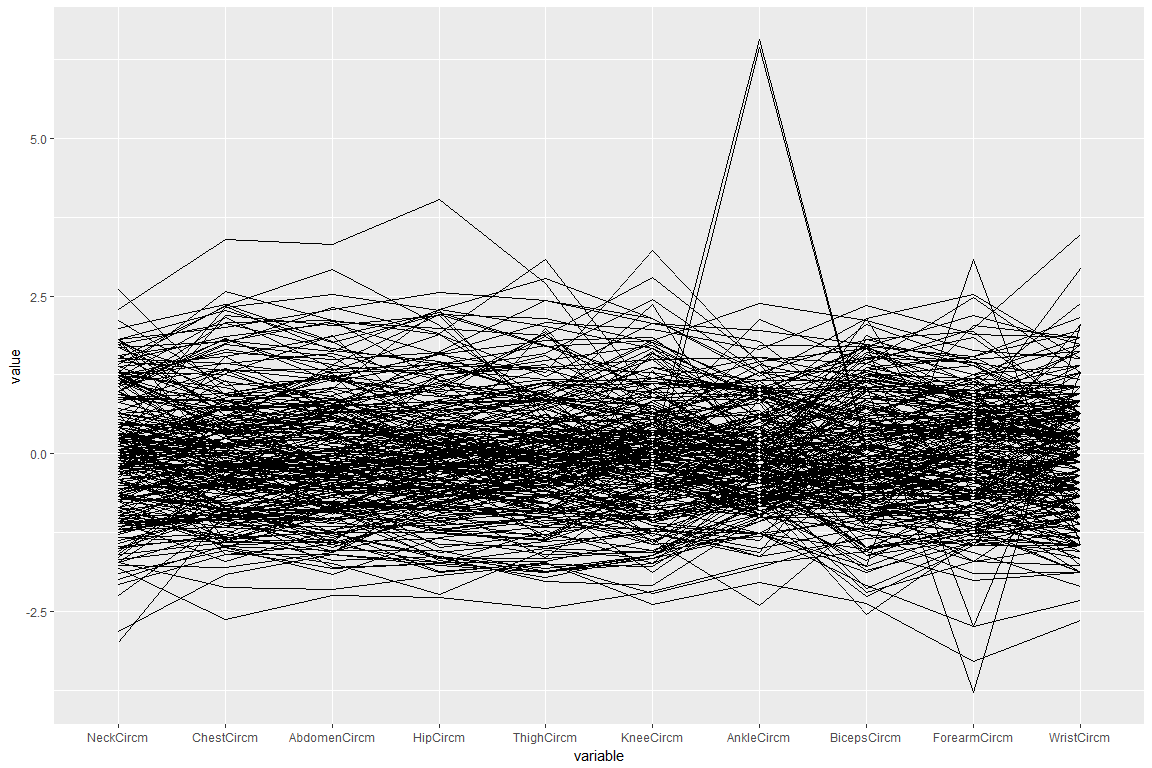
ggpairs(bodyfat[,1:4])



Create a second scatter plot matrix that only uses cases that are taller than 60 inches and weight less than 300 pounds. This removes points from the plot for one person who is under 40 inches tall and another person who weighs more than 300 pounds.

The next chunk uses ggparcoord to create a parallel coordinates plot for the data in columns 5 through 14, using the subset of cases created in the previous chunk.

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  
## [15] 21.0 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  
## [29] 21.5 21.5 21.6 21.6 21.6 21.7 21.7 21.7 21.7 21.8 21.8 21.8 21.8 21.8  
## [43] 21.8 21.8 21.8 21.8 21.8 21.9 21.9 21.9 21.9 21.9 21.9 21.9 21.9 22.0  
## [57] 22.0 22.0 22.0 22.0 22.0 22.0 22.0 22.0 22.0 22.0 22.1 22.1 22.1 22.1  
## [71] 22.1 22.1 22.1 22.2 22.2 22.2 22.2 22.3 22.3 22.3 22.3 22.3 22.3 22.3  
## [85] 22.4 22.4 22.4 22.4 22.4 22.4 22.4 22.4 22.5 22.5 22.5 22.5 22.5 22.5  
## [99] 22.5 22.5 22.5 22.5 22.5 22.6 22.6 22.6 22.6 22.6 22.6 22.6 22.6 22.6  
## [113] 22.6 22.6 22.7 22.7 22.7 22.7 22.7 22.7 22.7 22.7 22.7 22.8 22.8 22.8  
## [127] 22.8 22.8 22.9 22.9 22.9 22.9 22.9 22.9 22.9 23.0 23.0 23.0 23.0 23.0  
## [141] 23.0 23.1 23.1 23.1 23.1 23.1 23.2 23.2 23.2 23.2 23.2 23.2 23.2 23.2  
## [155] 23.2 23.3 23.3 23.3 23.3 23.4 23.4 23.4 23.4 23.4 23.4 23.4 23.4 23.5  
## [169] 23.5 23.5 23.5 23.6 23.6 23.6 23.6 23.7 23.7 23.7 23.8 23.8 23.8 23.8  
## [183] 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 24.1  
## [197] 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  
## [225] 24.8 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  
## [239] 25.5 25.5 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

1. Make a scatterplot matrix of the first four variables, % Body Fat, Age, Weight and Height. There's a problem with the data. What is it?
2. Fix the problem with the data and remake the scatterplot matrix. Describe the association between the variables.
3. Make a parallel coordinate plot of the last 10 variables, the circumference variables. There are some outliers in these measurements. Explain the ones that you see.
4. Remove the outliers and remake the parallel coordinate plot. Describe the structure of the data.

# Exercise 3 PISA Maths scores

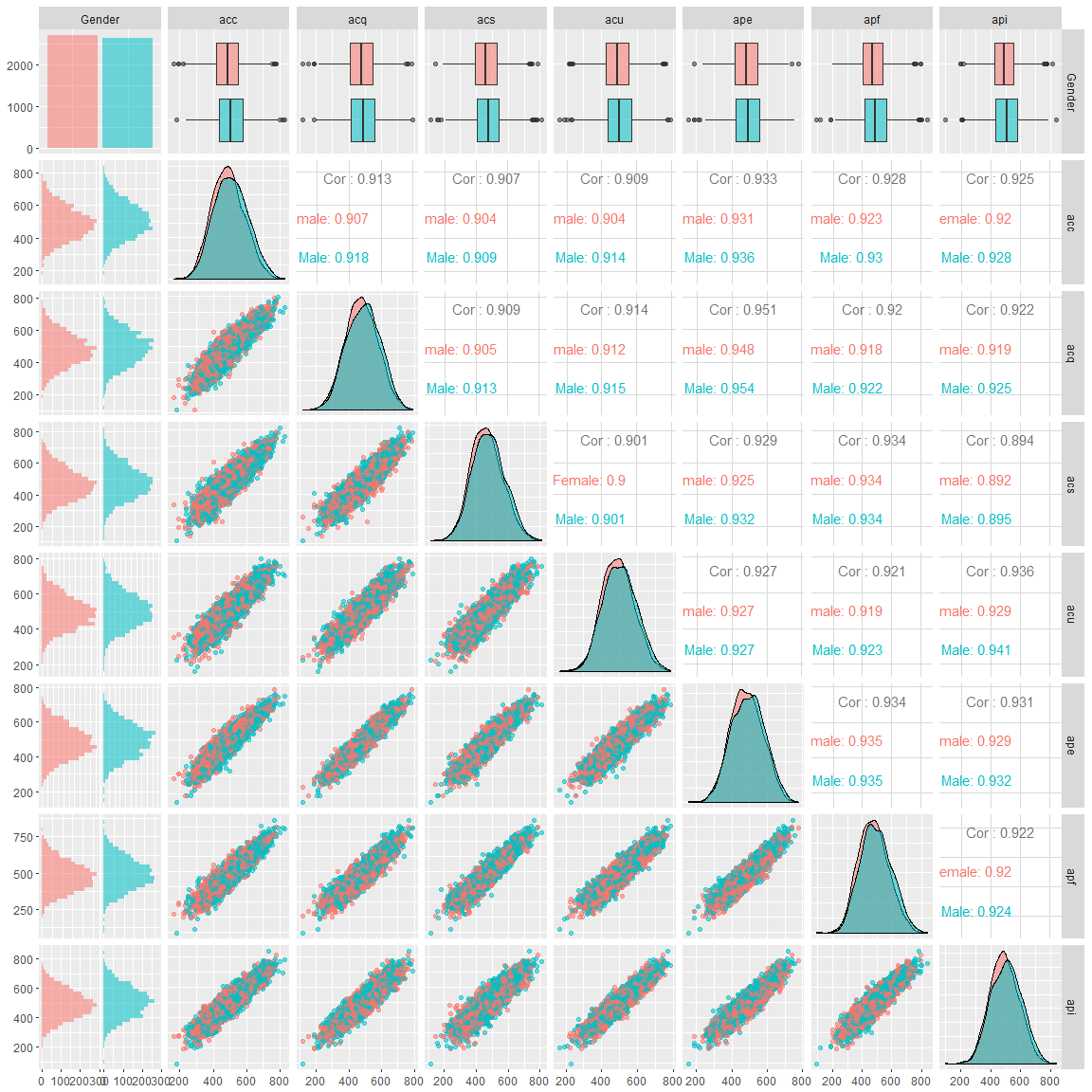
This exercise uses the ggpairs function in GGally to create a scatterplot matrix that includes information on correlations.

pisamath2 <- subset(pisamath, !is.na(acc))  
summary(pisamath2)

## Gender acc acq acs   
## Female:2689 Min. :162.8 Min. :115.1 Min. :115.8   
## Male :2627 1st Qu.:426.1 1st Qu.:412.2 1st Qu.:399.9   
## 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   
## 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

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`.



1. How many missing values for variable acc? Remove the missing values in the data.
2. Make a scatterplot matrix of the variables, with points coloured by Gender. Describe the association between the variables.

# Final Instructions

Put the names of all members of your group who participated in the Lab on Lab Report before you knit it.

Report due: Friday, September 10, 2020, by 11:59 pm.