KaitlinRMD

M. de Ferrante, K. Maciejewski, P. Batten

April 26, 2018

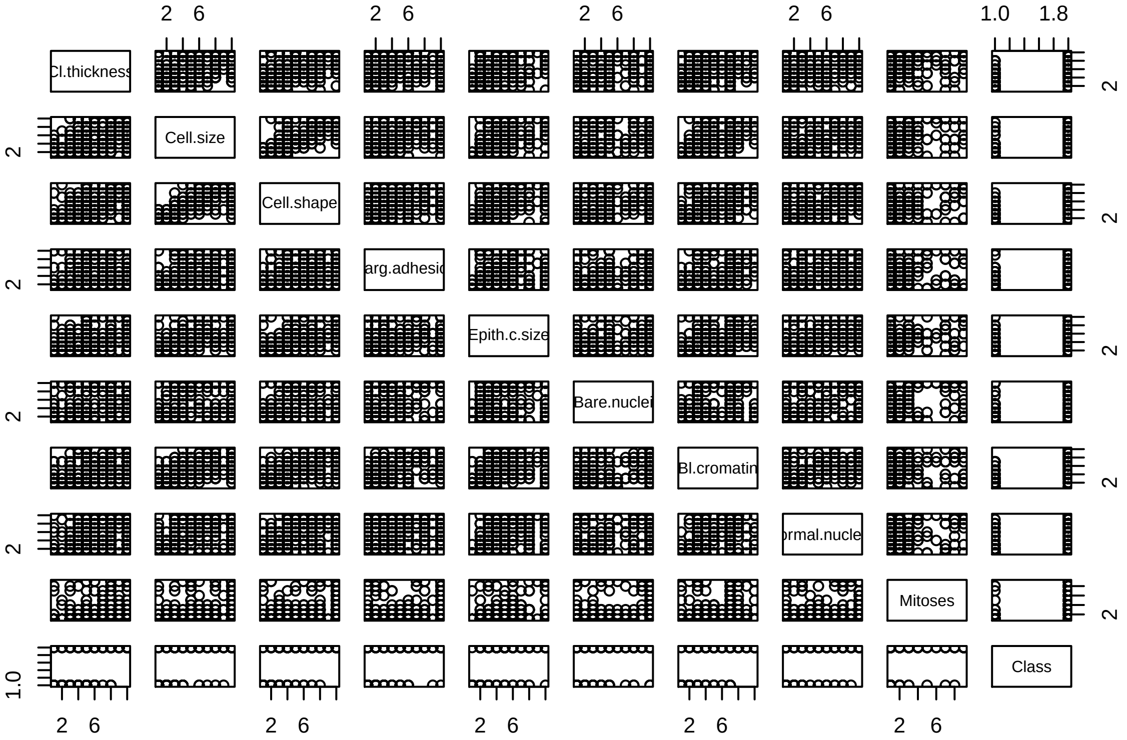
```
### Summary stuff for DS2 final project

library(mlbench)
data(BreastCancer)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
BreastCancer <- BreastCancer[,-1] # remove ID column
summary(BreastCancer) # note that everything is factor
```

```
##   Cl.thickness   Cell.size     Cell.shape  Marg.adhesion  Epith.c.size
##  1      :145   1      :384   1      :353   1      :407   2      :386
##  5      :130   10     : 67   2      : 59   2      : 58   3      : 72
##  3      :108   3      : 52   10     : 58   3      : 58   4      : 48
##  4      : 80   2      : 45   3      : 56   10     : 55   1      : 47
##  10     : 69   4      : 40   4      : 44   4      : 33   6      : 41
##  2      : 50   5      : 30   5      : 34   8      : 25   5      : 39
##  (Other):117   (Other): 81   (Other): 95   (Other): 63   (Other): 66
##   Bare.nuclei   Bl.cromatin  Normal.nucleoli    Mitoses          Class
##  1      :402   2      :166   1      :443     1      :579   benign   :458
##  10     :132   3      :165   10     : 61     2      : 35   malignant:241
##  2      : 30   1      :152   3      : 44     3      : 33
##  5      : 30   7      : 73   2      : 36     10     : 14
##  3      : 28   4      : 40   8      : 24     4      : 12
##  (Other): 61   5      : 34   6      : 22     7      :  9
##  NA's   : 16   (Other): 69   (Other): 69     (Other): 17
```

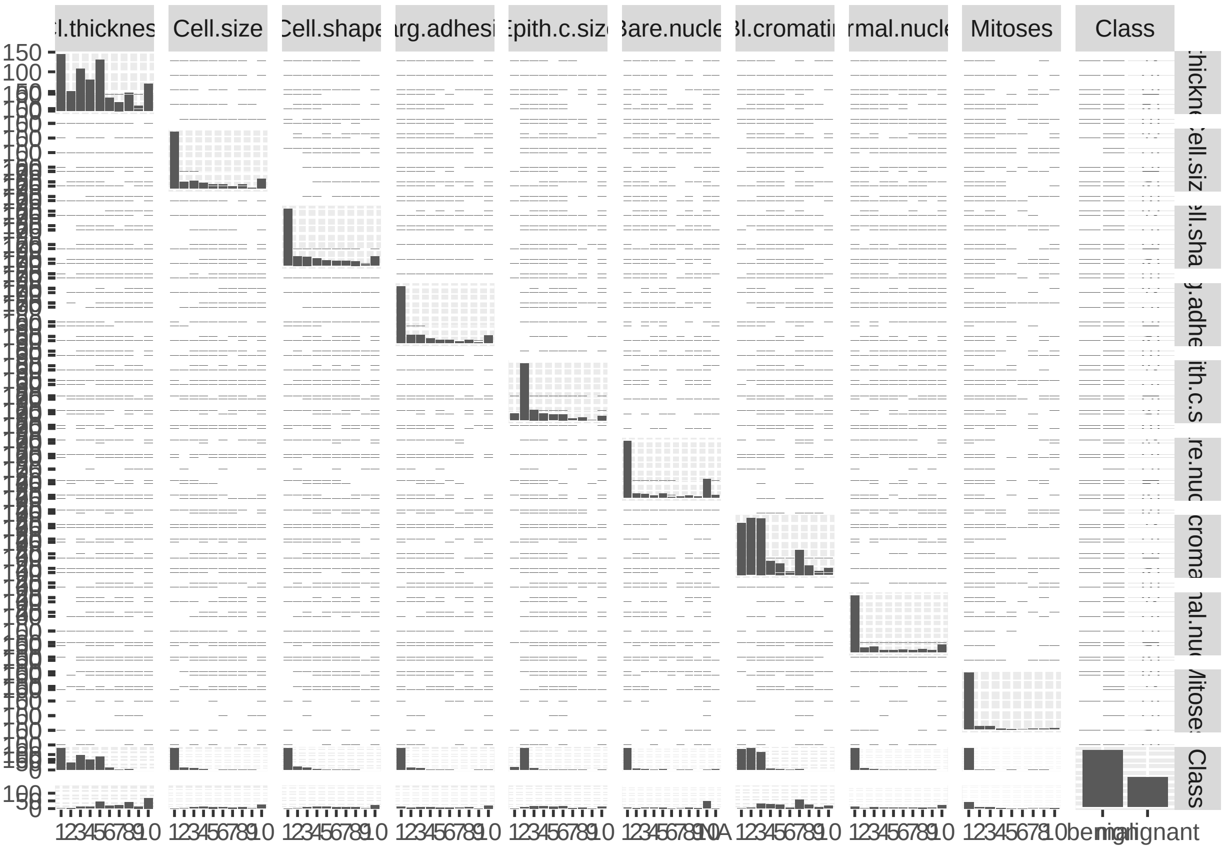
```
plot(BreastCancer)
```



```
library(GGally)
```

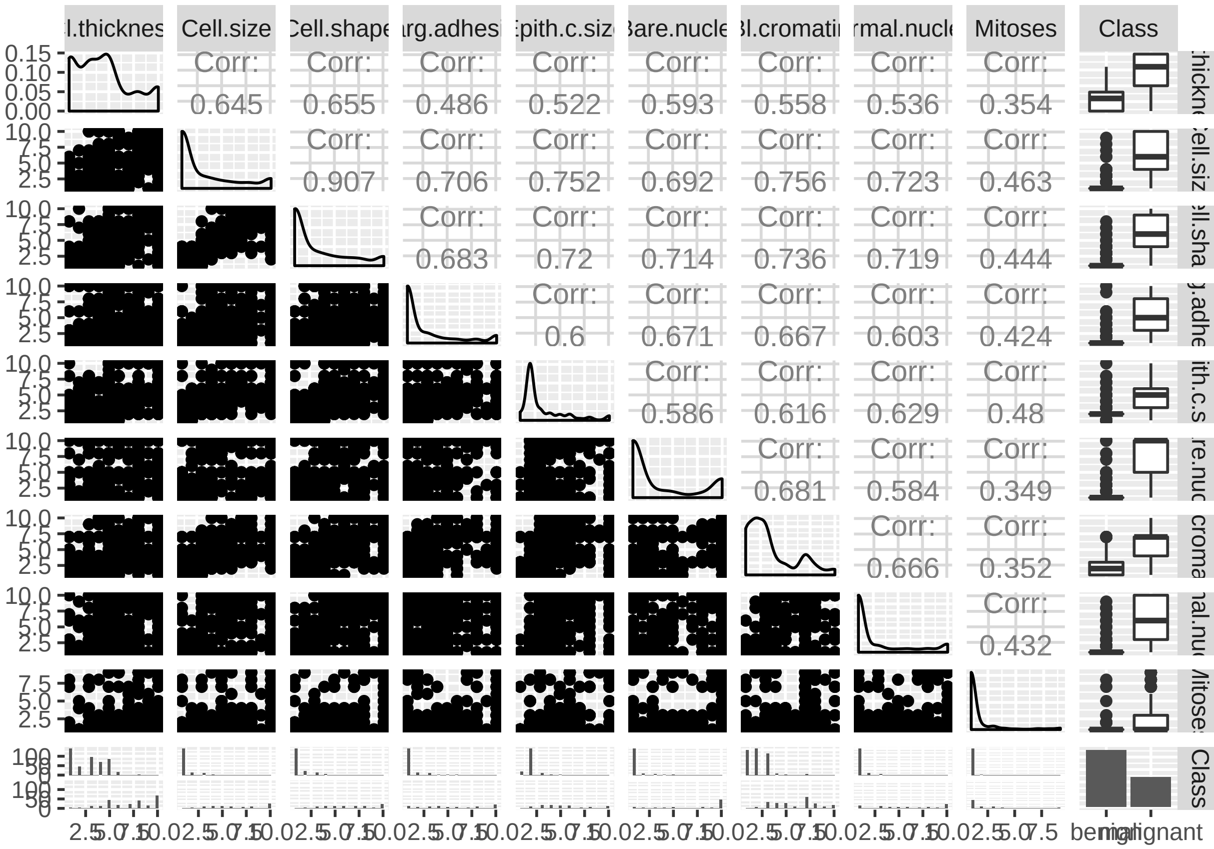
```
##
## Attaching package: 'GGally'

## The following object is masked from 'package:dplyr':
##
##     nasa
```

```
ggpairs(BreastCancer) # note all are factors
```

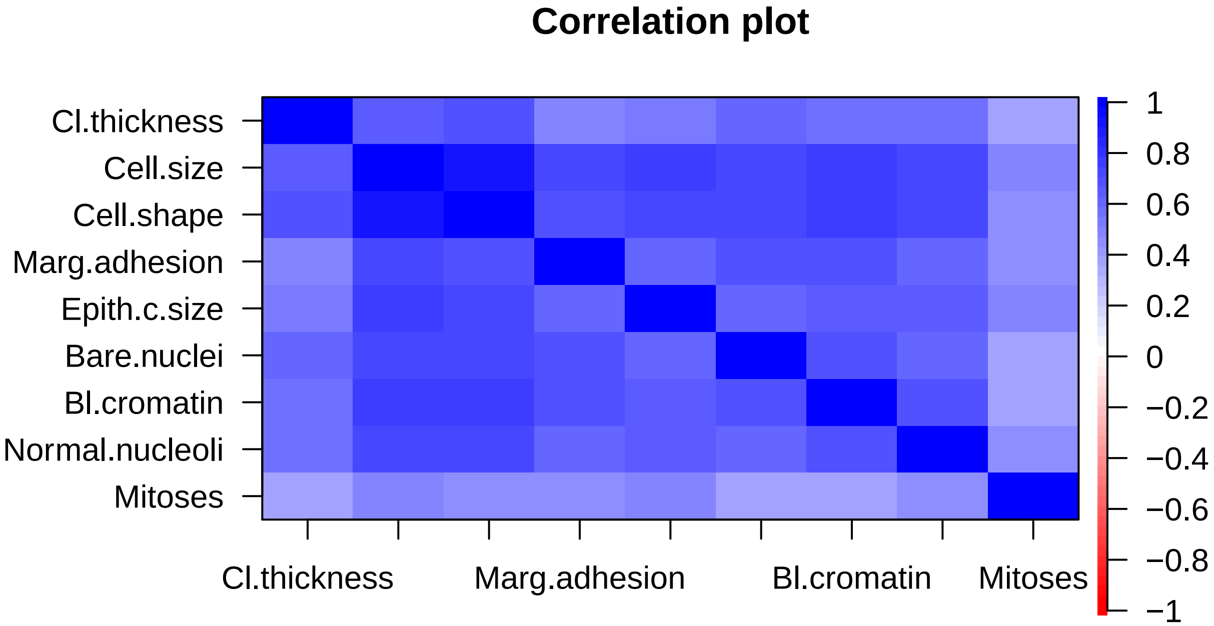


```r
BreastCancer = BreastCancer %>%
  mutate(Cl.thickness=as.numeric(Cl.thickness)) %>%
  mutate(Cell.size=as.numeric(Cell.size)) %>%
  mutate(Cell.shape=as.numeric(Cell.shape)) %>%
  mutate(Marg.adhesion=as.numeric(Marg.adhesion)) %>%
  mutate(Epith.c.size=as.numeric(Epith.c.size)) %>%
  mutate(Bare.nuclei=as.numeric(Bare.nuclei)) %>%
  mutate(Bl.cromatin=as.numeric(Bl.cromatin)) %>%
  mutate(Normal.nucleoli=as.numeric(Normal.nucleoli)) %>%
  mutate(Mitoses=as.numeric(Mitoses))

ggpairs(BreastCancer) # all but response are numeric
```

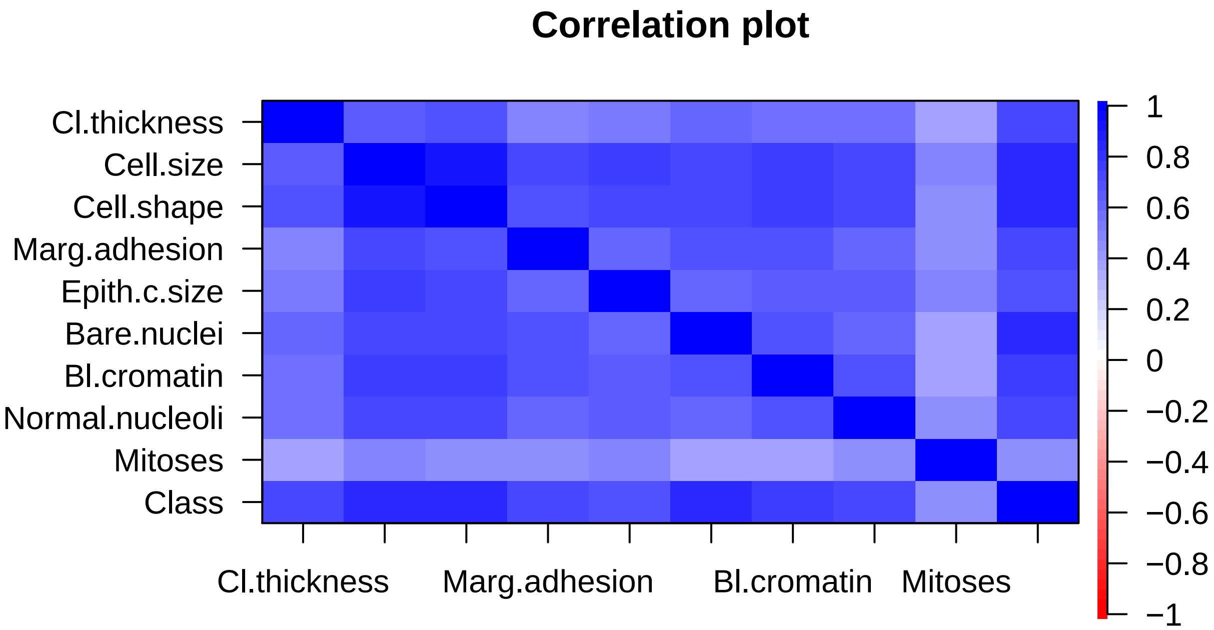
```
## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 16 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 16 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 16 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 16 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 16 rows containing missing values

## Warning: Removed 16 rows containing missing values (geom_point).
## Warning: Removed 16 rows containing missing values (geom_point).

## Warning: Removed 16 rows containing missing values (geom_point).

## Warning: Removed 16 rows containing missing values (geom_point).

## Warning: Removed 16 rows containing missing values (geom_point).

## Warning: Removed 16 rows containing non-finite values (stat_density).

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 16 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 16 rows containing missing values

## Warning in (function (data, mapping, alignPercent = 0.6, method =
## "pearson", : Removed 16 rows containing missing values

## Warning: Removed 16 rows containing non-finite values (stat_boxplot).

## Warning: Removed 16 rows containing missing values (geom_point).

## Warning: Removed 16 rows containing missing values (geom_point).

## Warning: Removed 16 rows containing missing values (geom_point).

## `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`.

## Warning: Removed 16 rows containing non-finite values (stat_bin).

## `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`.
```



```
library(psych)
cor.plot(BreastCancer[,-10]) # correlation not including response
```



```
BreastCancer <- BreastCancer %>% mutate(Class = as.numeric(Class)) # numeric response
cor.plot(BreastCancer[,])
```


Correlation plot
Cl.thickness
Cell.size
Cell.shape
Marg.adhesion
Epith.c.size
Bare.nuclei
Bl.cromatin
Normal.nucleoli
Mitoses
Class
Cl.thickness
Marg.adhesion
Bl.cromatin
Mitoses
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1