CM535 Data science development

Week 3 : Exploratory data analysis

Benjamin Lacroix
b.m.e.lacroix@rgu.ac.uk

February 11, 2020

Introduction

Exploratory data analysis

- Exploratory Data Analysis is a preliminary examination of the data
- No Specific Question is Answered, But you may have questions in mind
- Uses descriptive analytics
  - Descriptive statistics
  - Visualisation

Introduction

Aims

- To check any problems with the data
  - May suggest data preparation processes
- To find interesting patterns
  - EDA philosophy is open-minded
  - Find the unexpected as well as the expected
- To check if data is suitable for intended main analysis

Typical Exploration

- Univariate Analysis of Every Variable
  - Summary Statistics and Boxplot
  - Other plots of distribution
  - E.g. Histogram, Bar Chart, Dotplot
- Bivariate Analysis
  - Correlation Matrix
  - Scatterplots
  - Comparison of Subgroups of Interest

Example - Exploring mtcars

- Small data set:
  - 11 features (columns)
  - 32 instances (rows)
- Features stored as numerical variables:
  - 6 continuous features (mpg, disp, hp, drat, wt, qsec)
  - 3 integer features (cyl, gear, carb)
  - 2 binary categorical features (vsm, am)
- Features can be categorised as:
  - 8 design features
  - 3 perfromance measures (mpg, hp, qsec)

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

Univariate analysis

Describing the distribution of a single variable:

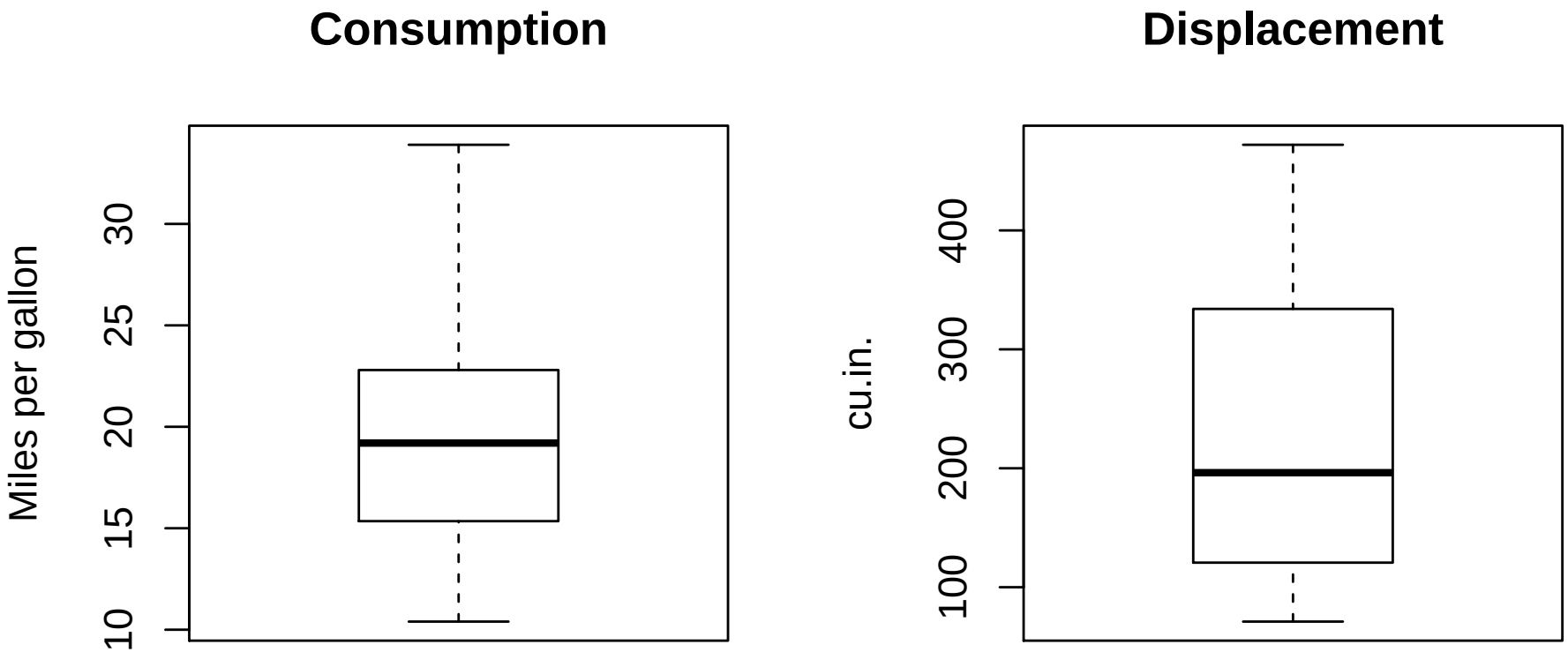
- Measures of location (mean and median)
- Measures of dispersion:
  - standard deviation - sd(x)
  - variance - var(x)
  - range - range(x) or min(x) and max(x)
  - quantiles - quantile(x, probs = c(0.25, 0.75))
- 5-number summary (min, Q1, Median, Q3, max)

```
summary(mtcars)
```

```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0  
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5  
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0  
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7  
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0  
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0  
##       drat             wt             qsec             vs        
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000  
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000  
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000  
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375  
```

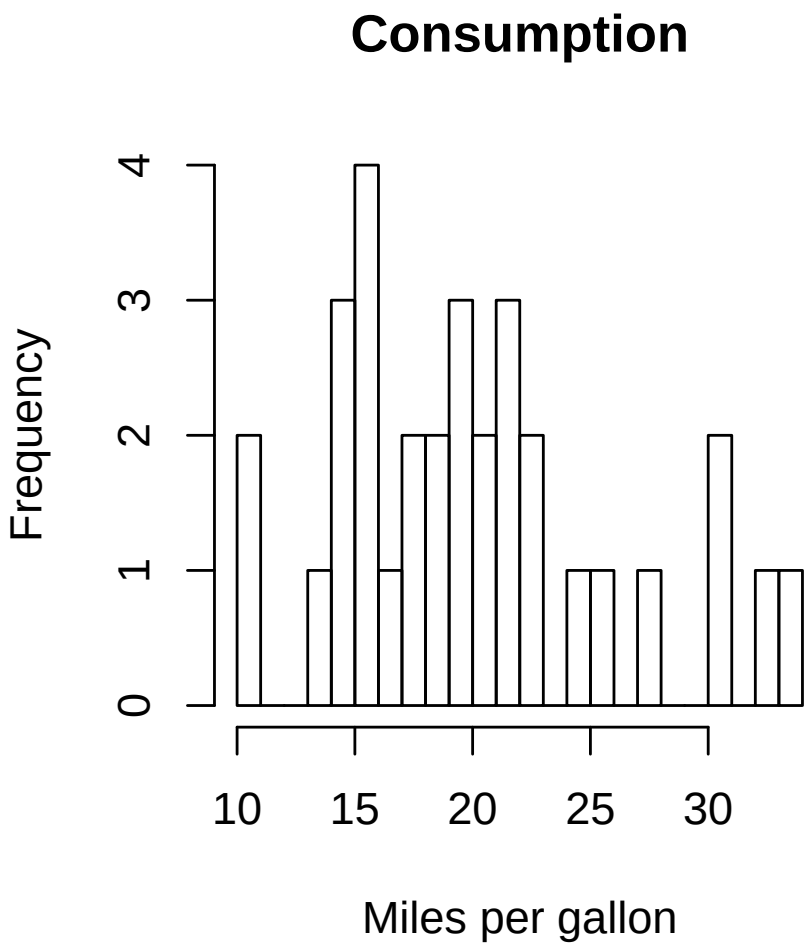
Univariate visualisations: Boxplots

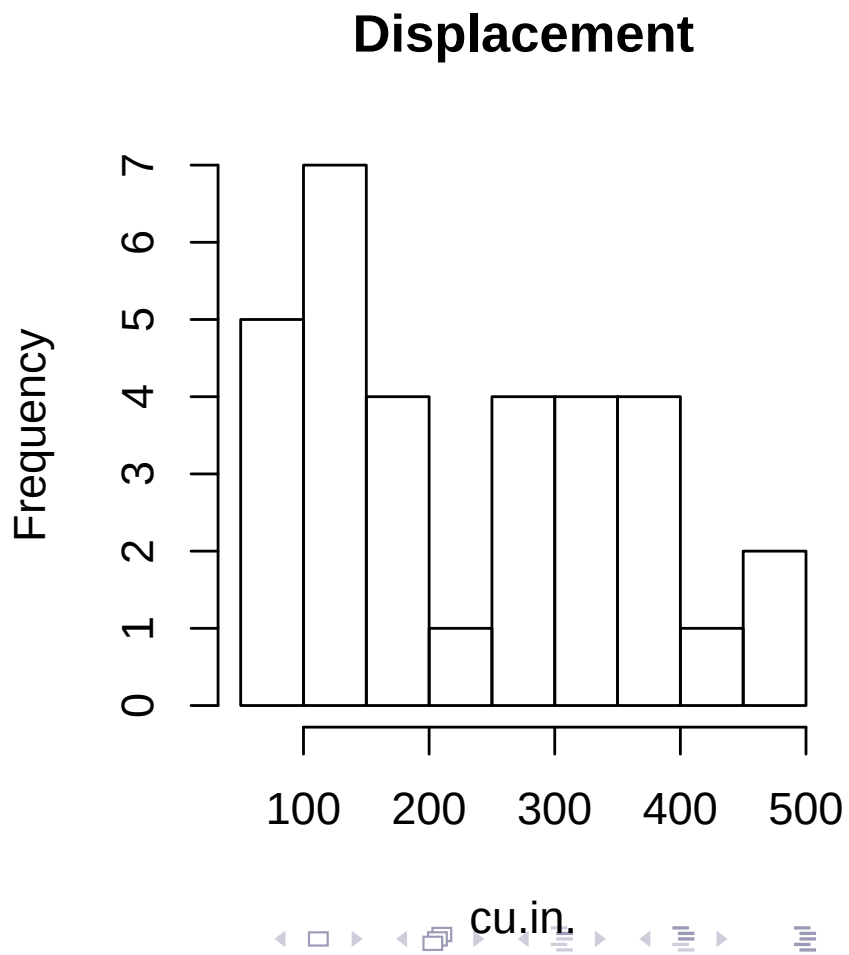
```
par(mfrow=(c(1,2)))
boxplot(mtcars$mpg, ylab = "Miles per gallon", main = "Consumption")
boxplot(mtcars$disp, ylab = "cu.in.", main = "Displacement")
```



Univariate visualisations: Histograms

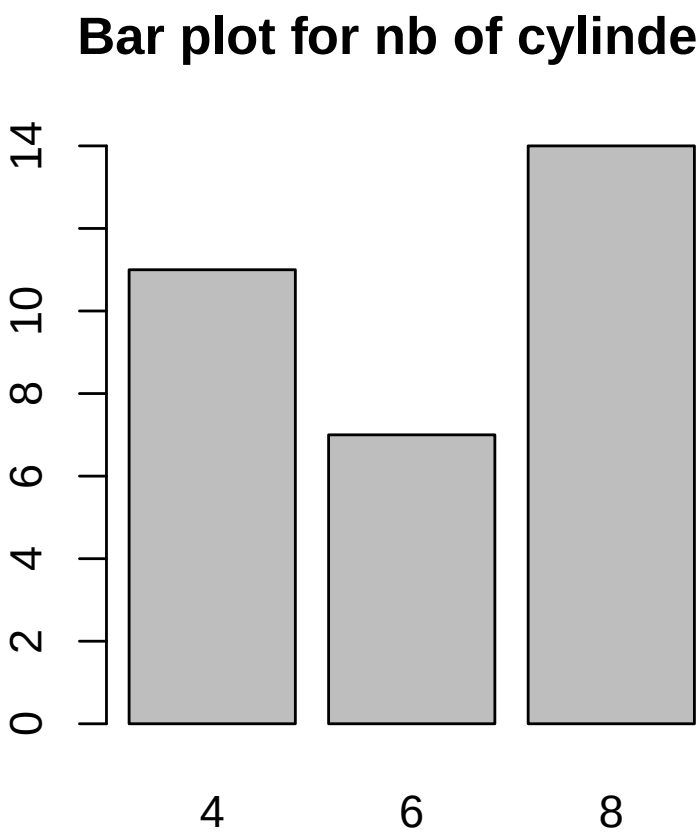
```
par(mfrow=(c(1,2)))
hist(mtcars$mpg, xlab = "Miles per gallon", main = "Consumption", breaks = 20)
hist(mtcars$disp, xlab = "cu.in.", main = "Displacement")
```

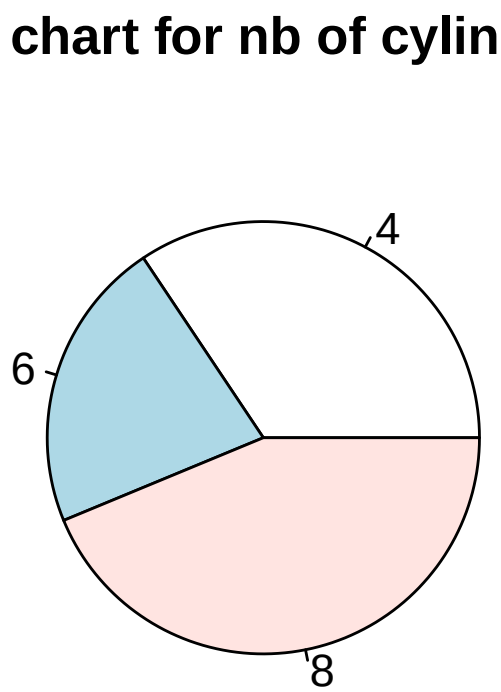




Univariate visualisations: Tabulating variables

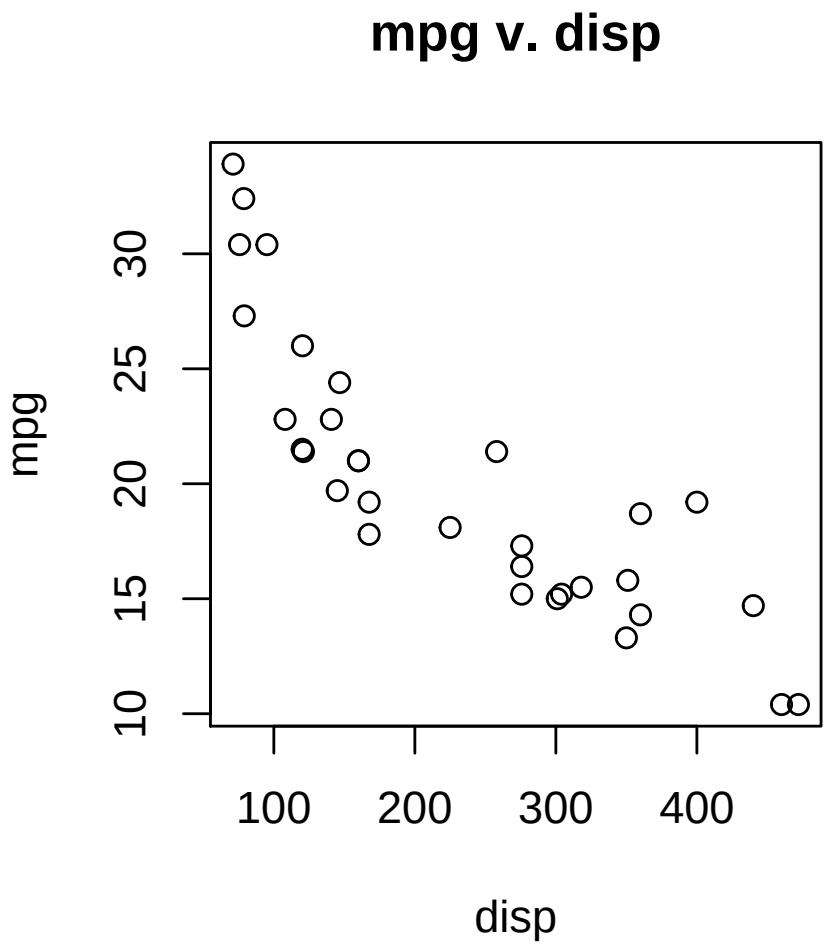
```r
cyl.table <- table(mtcars$cyl)

par(mfrow=(c(1,2)))
barplot(cyl.table, main = "Bar plot for nb of cylinders")
pie(cyl.table, main = "Pie chart for nb of cylinders")
```

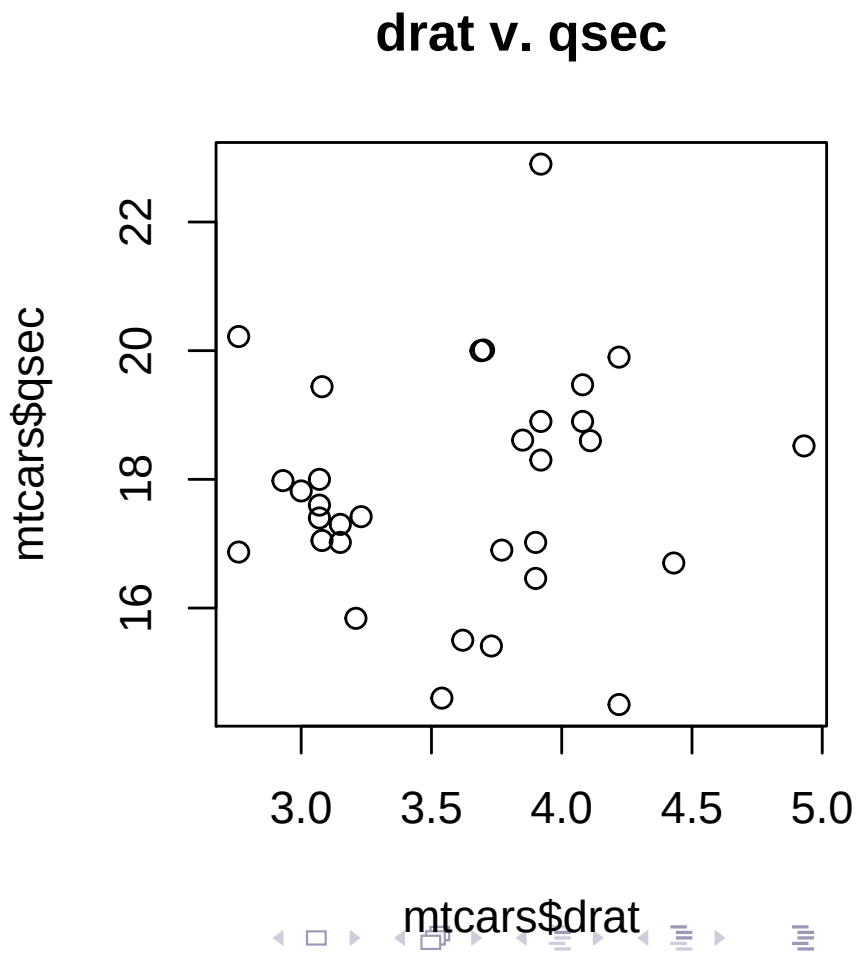




Bivariate analysis: Scatter plots for numeric variables

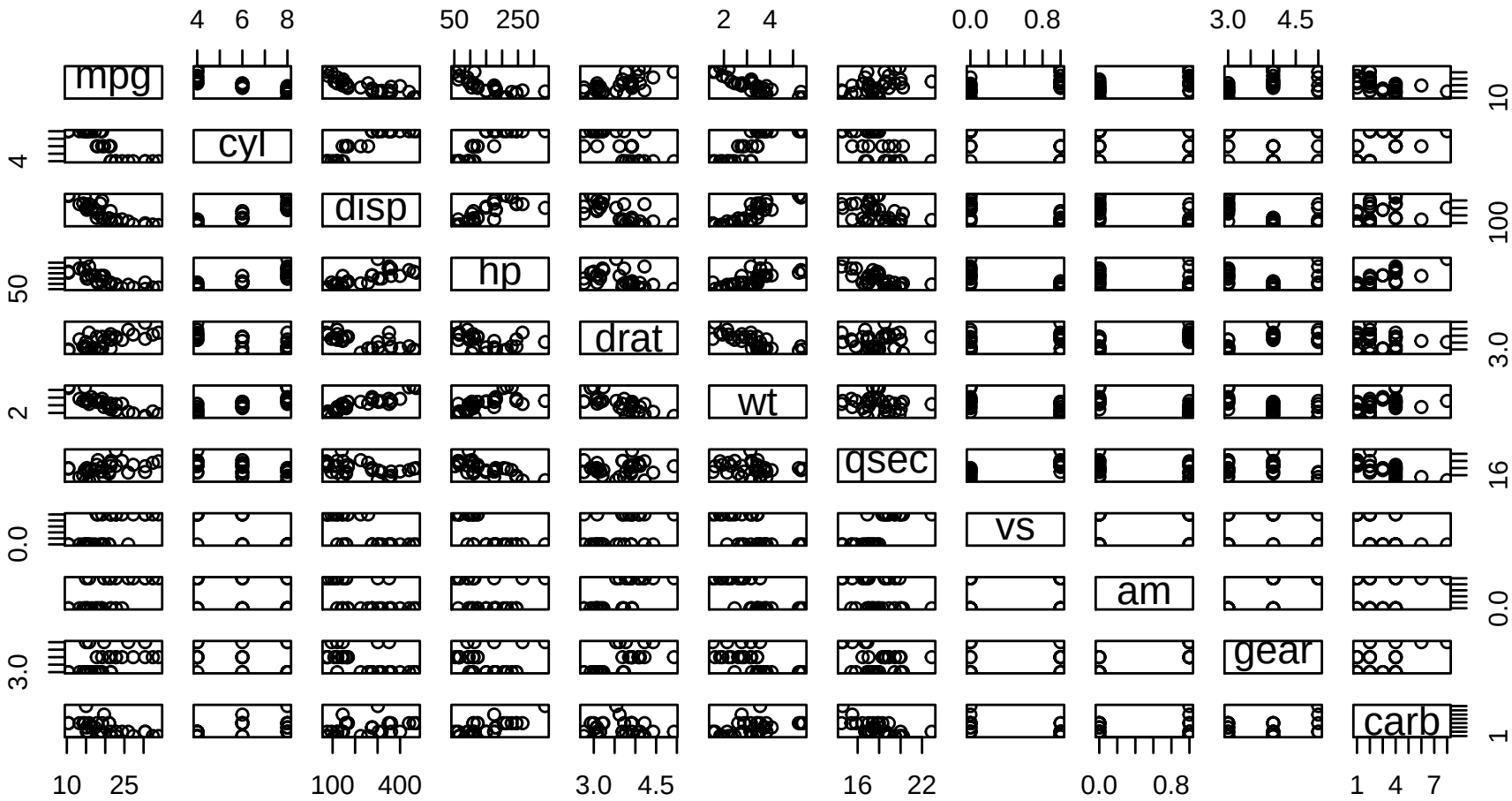
```
par(mfrow=(c(1,2)))
plot(mtcars$disp,mtcars$mpg, main = "mpg v. disp", ylab = "mpg", xlab = "disp")
plot(mtcars$drat,mtcars$qsec, main = "drat v. qsec")
```





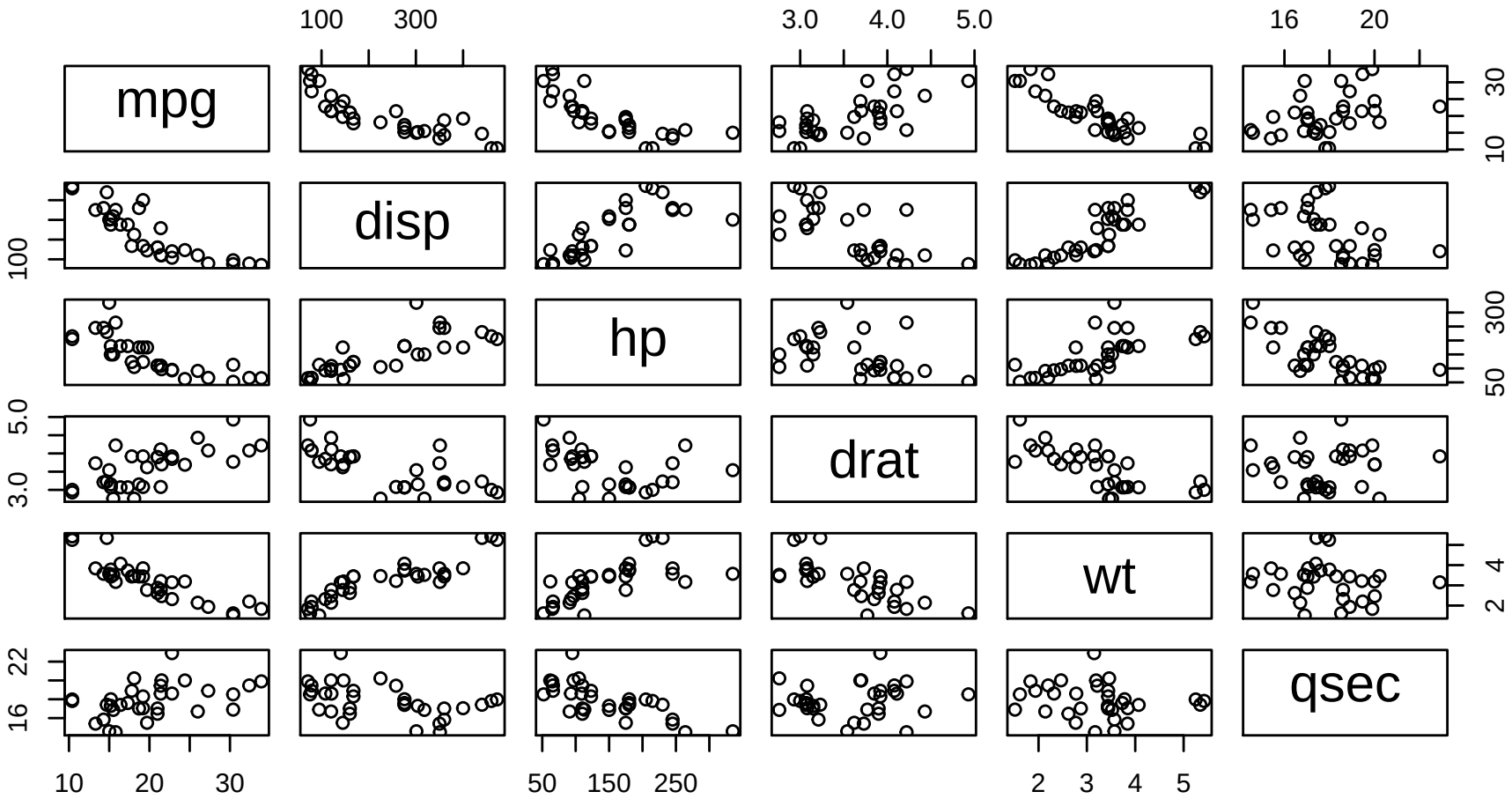
Bivariate analysis

```
pairs(mtcars)
```



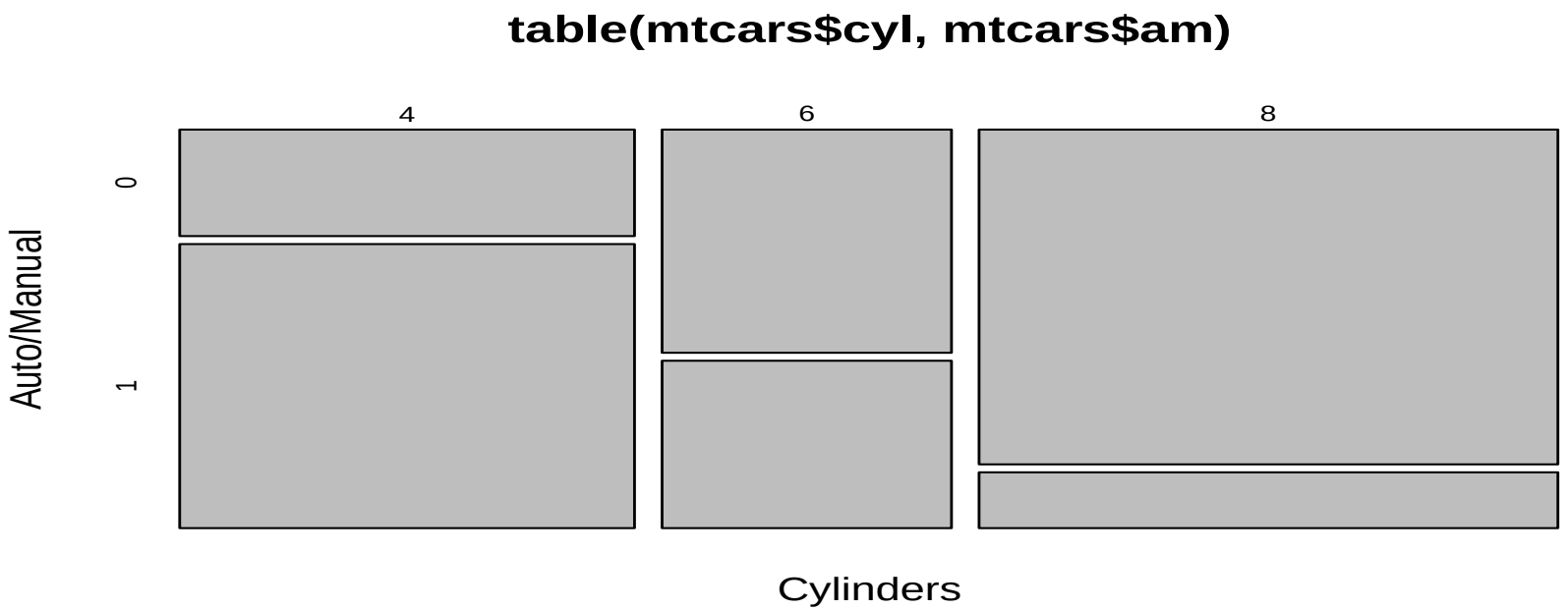
Bivariate analysis

```
pairs(mtcars[,c(1,3,4,5,6,7)])
```



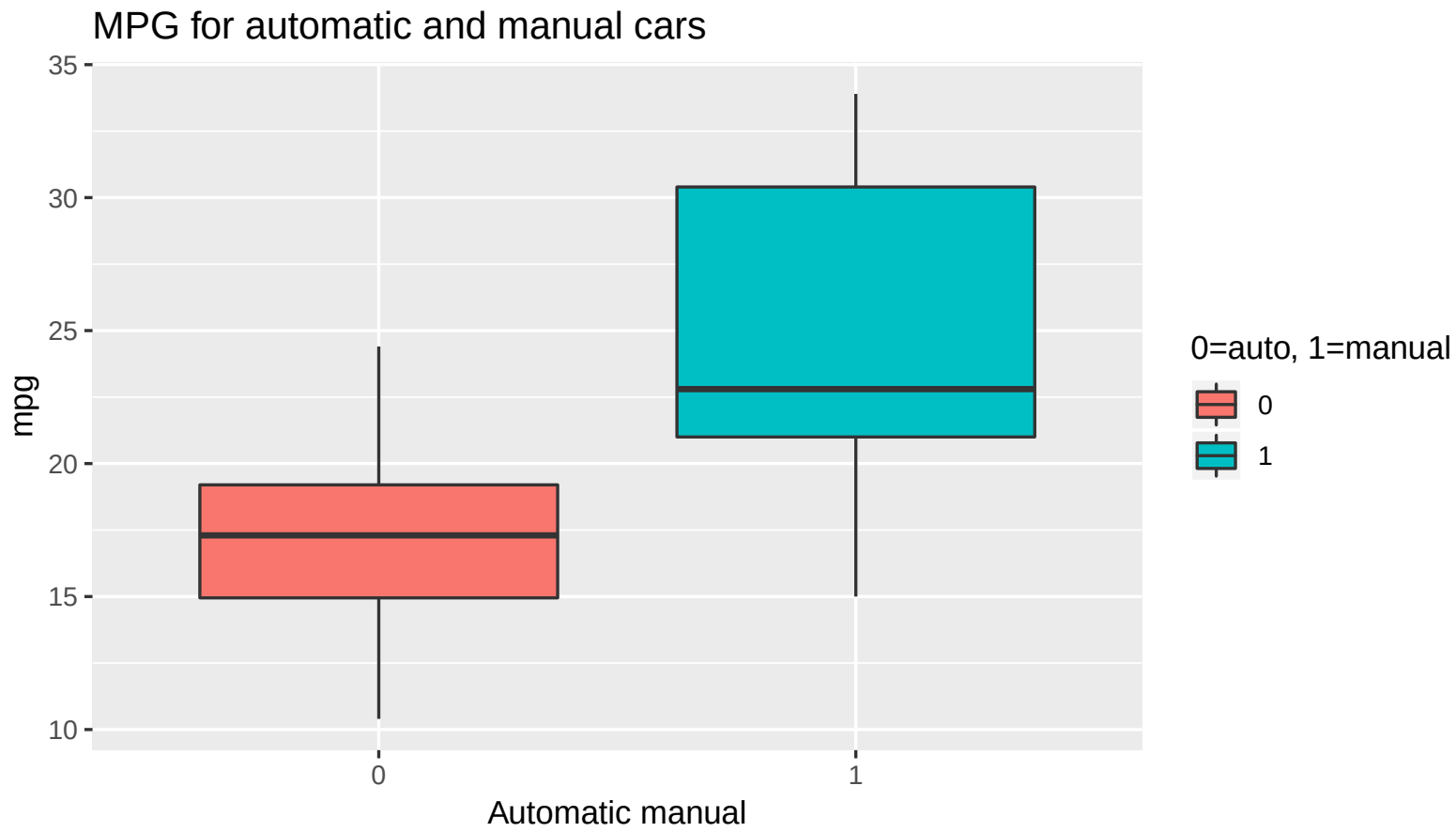
Bivariate analysis: cross tabulation for discrete variables

```
table(mtcars$cyl,mtcars$am)

##
##      0  1
##   4  3  8
##   6  4  3
##   8 12  2

mosaicplot(table(mtcars$cyl,mtcars$am), ylab = "Auto/Manual", xlab = "Cylinders")
```



Bivariate analysis: categorical and numerical variables

```
library(ggplot2)
ggplot(mtcars, aes(factor(am),mpg, fill = factor(am))) + geom_boxplot() +
  ggtitle("MPG for automatic and manual cars") + xlab("Automatic manual") +
  labs( fill = "0=auto, 1=manual")
```



Bivariate analysis: Measures of association - Covariance

Measures how much two variables vary together

$$\sigma(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n}$$

Interpretation:

- Positive: there is a positive association between variables X and Y.
  - x tends to be higher than average when y is higher than average.
- Negative: there is a negative association between variables X and Y.
  - x tends to be lower than average when y is higher than average.
- 0 : no association between X and Y

Bivariate analysis: Measures of association - Covariance

```
cov(mtcars)
```

```
##              mpg         cyl        disp          hp         drat
## mpg    36.324103  -9.1723790  -633.09721 -320.732056   2.19506351
## cyl    -9.172379   3.1895161   199.66028  101.931452  -0.66836694
## disp -633.097208 199.6602823 15360.79983 6721.158669 -47.06401915
## hp   -320.732056 101.9314516  6721.15867 4700.866935 -16.45110887
## drat    2.195064  -0.6683669   -47.06402  -16.451109   0.28588135
## wt     -5.116685   1.3673710   107.68420   44.192661  -0.37272073
## qsec    4.509149  -1.8868548   -96.05168  -86.770081   0.08714073
## vs      2.017137  -0.7298387   -44.37762  -24.987903   0.11864919
## am      1.803931  -0.4657258   -36.56401   -8.320565   0.19015121
## gear    2.135685  -0.6491935   -50.80262   -6.358871   0.27598790
## carb   -5.363105   1.5201613    79.06875   83.036290  -0.07840726
##               wt         qsec          vs          am        gear
## mpg   -5.1166847   4.50914919  2.01713710  1.80393145   2.1356855
## cyl    1.3673710  -1.88685484 -0.72983871 -0.46572581  -0.6491935
## disp 107.6842040 -96.05168145 -44.37762097 -36.56401210 -50.8026210
## hp    44.1926613 -86.77008065 -24.98790323  -8.32056452  -6.3588710
## drat  -0.3727207   0.08714073  0.11864919  0.19015121   0.2759879
## wt     0.9573790  -0.30548161 -0.27366129 -0.33810484  -0.4210806
## qsec  -0.3054816   3.19316613  0.67056452 -0.20495968  -0.2804032
## vs    -0.2736613   0.67056452  0.25403226  0.04233871   0.0766129
## am    -0.3381048  -0.20495968  0.04233871  0.24899194   0.2923387
```

Bivariate analysis: Measures of association - Correlation

Not obvious what size of covariance means. When is it close to zero?
Need to normalise using the standard deviation of the variables. The correlation coefficient is:

$$R(x, y) = \frac{\sigma(x, y)}{\sigma(x)\sigma(y)}$$

Returns a value beteween -1 and 1

- 1: perfect linear relationship with positive gradient
- 0.7 to 1: Strong positive linear relationship
- 0.3 to 0.7: Weak positive linear relationship
- -0.3 to 0.3: no (linear) relationship
- -0.7 to -0.3: Weak negative linear relationship
- -1 to -0.7: Strong negative linear relationship
- -1: perfect linear relationship with negative gradient
- 0: no (linear) relationship

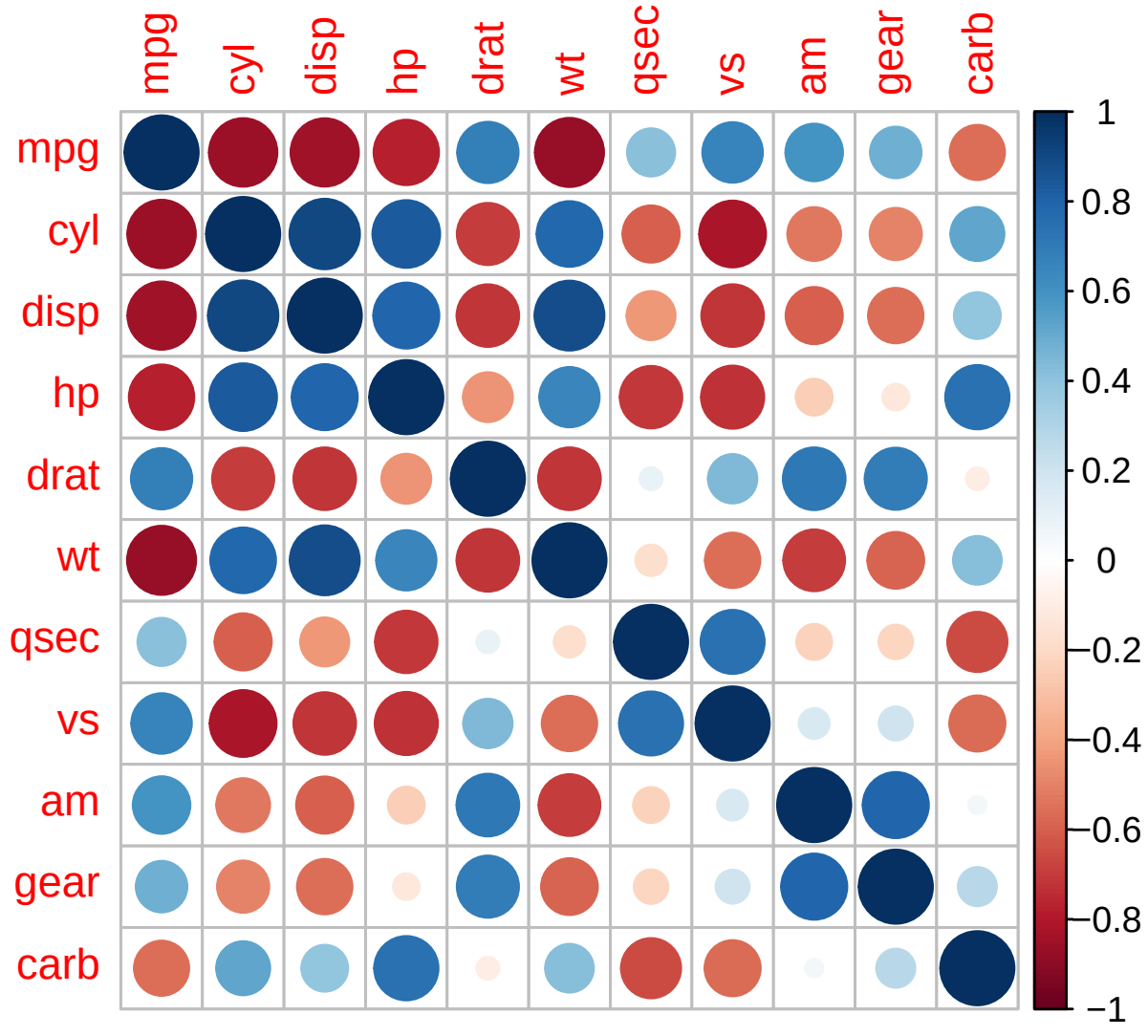
Bivariate analysis: Measures of association - Correlation

```
cor(mtcars)
```

```
##             mpg        cyl       disp         hp        drat          wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
```

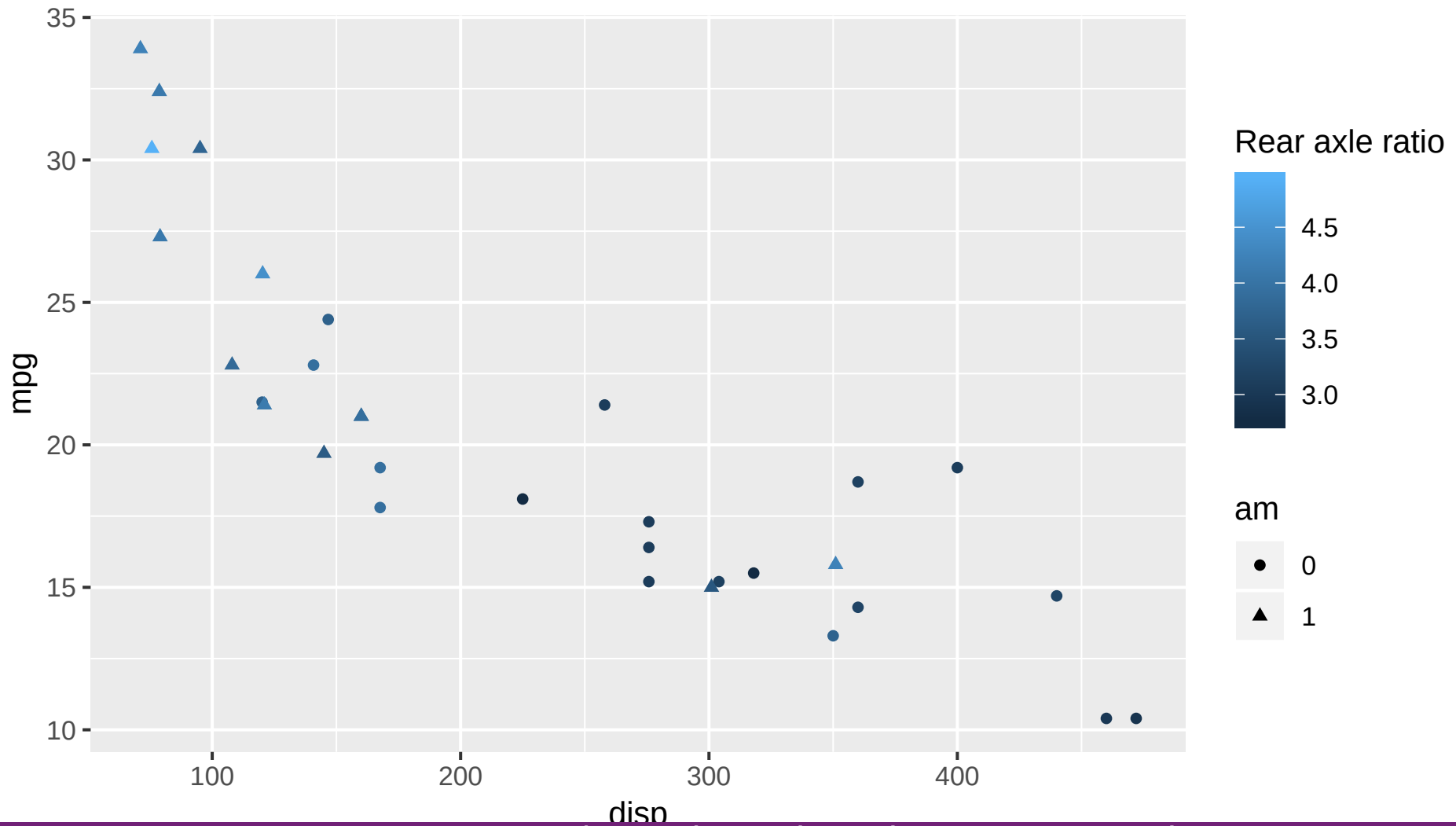
Bivariate analysis: correlation

```
library(corrplot)
corrplot(cor(mtcars))
```



Multivariate Visualisation

```
ggplot(mtcars, aes(x = disp, y = mpg, color = drat, shape = factor(am))) +
  geom_point() +
  labs(color = "Rear axle ratio", shape = "am")
```



Up next

Today's lab in N533

- Data preparation and data cleaning in R
- Loading data
- Detecting NA
- Data transformation

Next week:

- Exploratory data analysis