University of Applied Sciences and Arts Northwestern Switzerland
School of Business

Data Science - Exercises

Holger Wache

Exercise A

The (build-in) Data Set “mtcars”

```
> mtcars
                   mpg cyl  disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6 160.0 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6 160.0 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4 108.0  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6 258.0 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8 360.0 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6 225.0 105 2.76 3.460 20.22  1  0    3    1
Duster 360        14.3   8 360.0 245 3.21 3.570 15.84  0  0    3    4
Merc 240D         24.4   4 146.7  62 3.69 3.190 20.00  1  0    4    2
Merc 230          22.8   4 140.8  95 3.92 3.150 22.90  1  0    4    2
Merc 280          19.2   6 167.6 123 3.92 3.440 18.30  1  0    4    4
Merc 280C         17.8   6 167.6 123 3.92 3.440 18.90  1  0    4    4
Merc 450SE        16.4   8 275.8 180 3.07 4.070 17.40  0  0    3    3
Merc 450SL        17.3   8 275.8 180 3.07 3.730 17.60  0  0    3    3
Merc 450SLC       15.2   8 275.8 180 3.07 3.780 18.00  0  0    3    3
Cadillac Fleetwood  10.4   8 472.0 205 2.93 5.250 17.98  0  0    3    4
Lincoln Continental 10.4   8 460.0 215 3.00 5.424 17.82  0  0    3    4
Chrysler Imperial   14.7   8 440.0 230 3.23 5.345 17.42  0  0    3    4
…
```

First Insights into Data Set “mtcars”

```
> str(mtcars)
'data.frame':    32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Compute the mean, median, and mode of column “wt”

```
> mean(mtcars$wt)
[1] 3.21725
```

```
> median(mtcars$wt)
[1] 3.325
```

```
> mode(mtcars$wt)
[1] "numeric"
```

Not correct; mode is another function

```
> y <- table(mtcars$wt)
> y

1.513 1.615 1.835 1.935  2.14   2.2  2.32 2.465  2.62  2.77  2.78 2.875
    1     1     1     1     1     1     1     1     1     1     1     1
3.15  3.17  3.19 3.215 3.435  3.44  3.46  3.52  3.57  3.73  3.78  3.84
   1     1     1     1     1     3     1     1     2     1     1     1
3.845  4.07  5.25 5.345 5.424
    1     1     1     1     1
> names(y)[which(y==max(y))]
[1] "3.44"
```

That would be the correct (statistical) mode

.. Or everything in one command for “mtcars”

```
> summary(mtcars)
      mpg             cyl             disp             hp             drat
 Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
 1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
 Median :19.20   Median :6.000   Median :196.3   Median :123.0   Median :3.695
 Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7   Mean   :3.597
 3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
 Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0   Max.   :4.930
       wt             qsec             vs              am
 Min.   :1.513   Min.   :14.50   Min.   :0.0000   Min.   :0.0000
 1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000   1st Qu.:0.0000
 Median :3.325   Median :17.71   Median :0.0000   Median :0.0000
 Mean   :3.217   Mean   :17.85   Mean   :0.4375   Mean   :0.4062
 3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :5.424   Max.   :22.90   Max.   :1.0000   Max.   :1.0000
      gear            carb
 Min.   :3.000   Min.   :1.000
 1st Qu.:3.000   1st Qu.:2.000
 Median :4.000   Median :2.000
 Mean   :3.688   Mean   :2.812
 3rd Qu.:4.000   3rd Qu.:4.000
 Max.   :5.000   Max.   :8.000
```

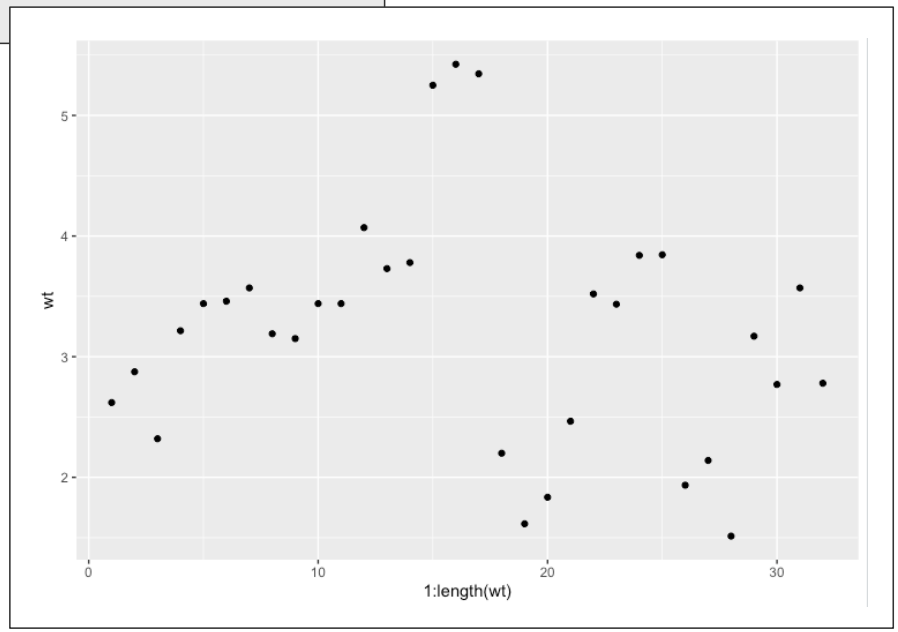
Draw (plot) the column “wt”

```
> ggplot(mtcars) + geom_point(mapping = aes(x = 1:length(wt), y = wt))
```

Advanced plotting of data mtcars

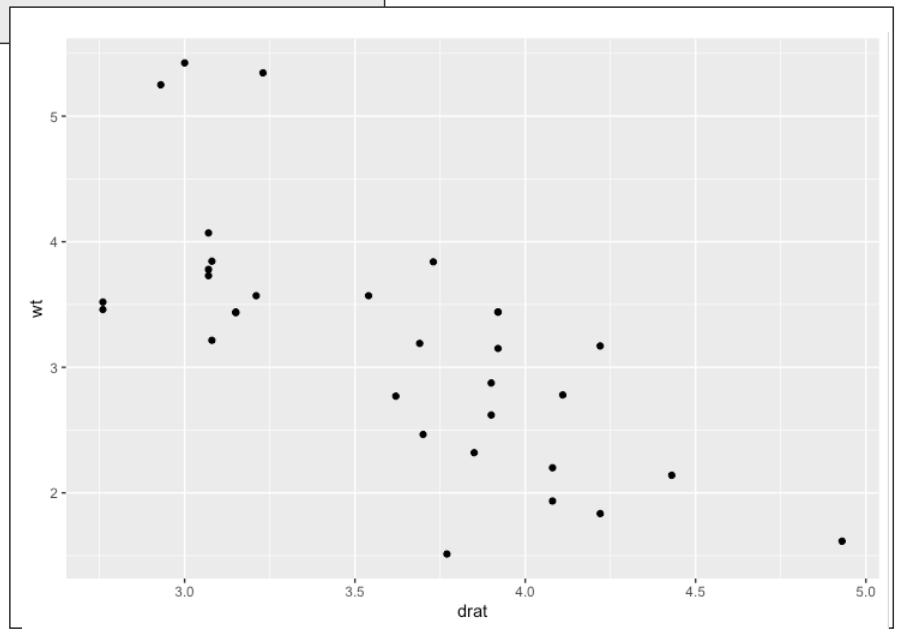
Adding a layer

Draw points a the x and y coordinates



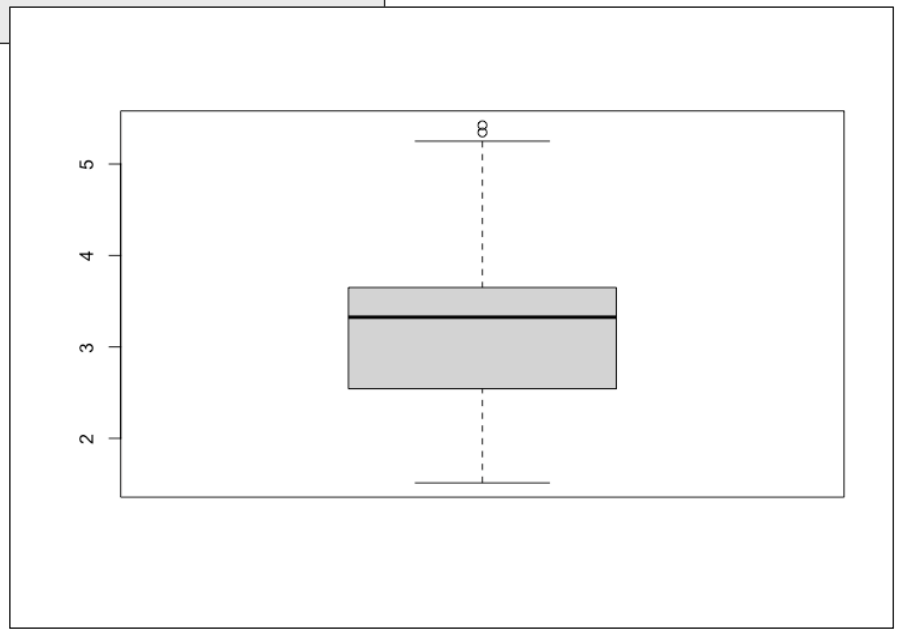
Draw (plot) the column “wt” against column “drat”

```
> ggplot(mtcars) + geom_point(mapping = aes(x = drat, y = wt))
```



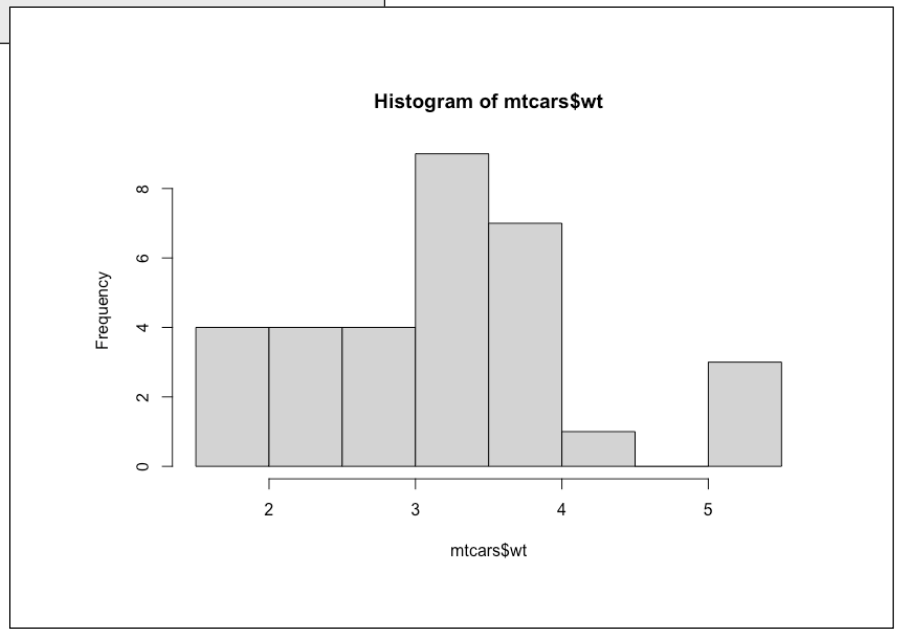
Boxplot of column “wt”

```
> boxplot(mtcars$wt)
```



Histogram of column “wt” and help for hist()

```
> hist(mtcars$wt)
```




```
> ?hist
```

hist {graphics} R Documentation

Histograms

Description

The generic function `hist` computes a histogram of the given data values. If `plot = TRUE`, the resulting object of class `"histogram"` is plotted by `plot.histogram`, before it is returned.

Usage

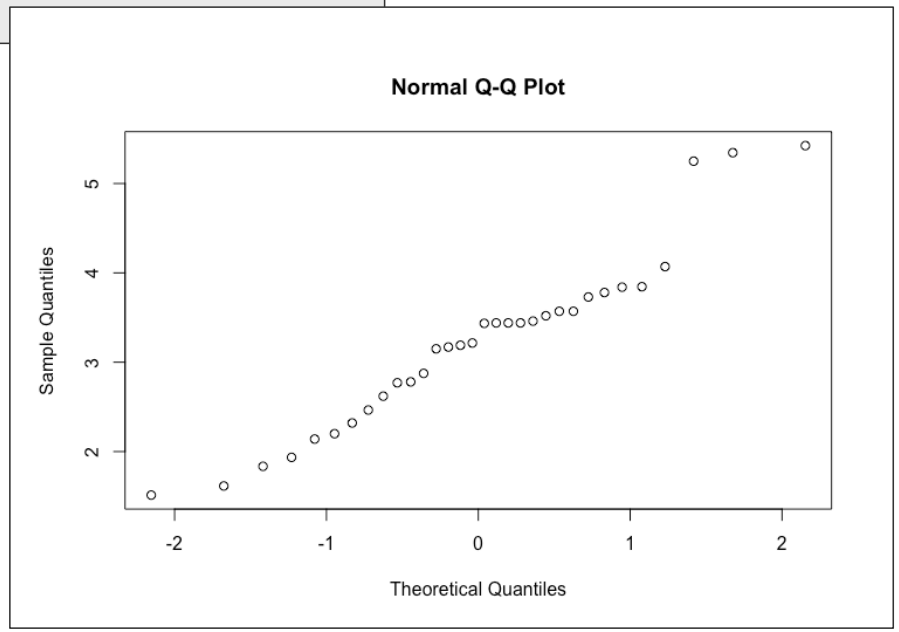
```
hist(x, ...)

## Default S3 method:
hist(x, breaks = "Sturges",
     freq = NULL, probability = !freq,
     include.lowest = TRUE, right = TRUE,
     density = NULL, angle = 45, col = "lightgray", border = NULL,
     main = paste("Histogram of" , xname),
     xlim = range(breaks), ylim = NULL,
     xlab = xname, ylab,
     axes = TRUE, plot = TRUE, labels = FALSE,
     nclass = NULL, warn.unused = TRUE, ...)
```

Arguments

x	a vector of values for which the histogram is desired.

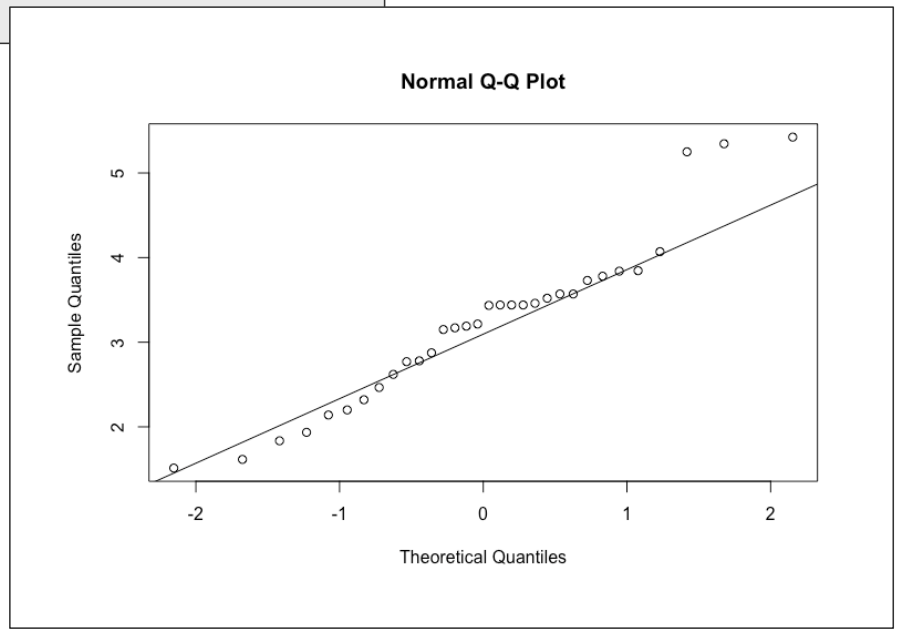
Q-Q Plot of column “wt”

```
> qqnorm(mtcars$wt)
```



Q-Q Plot with assumed line of column “wt”

```
> qqline(mtcars$wt)
```



Correlation of columns of data set “mtcars”

```
> cor(mtcars)
…
> round(cor(mtcars),2)
       mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
mpg   1.00 -0.85 -0.85 -0.78  0.68 -0.87  0.42  0.66  0.60  0.48 -0.55
cyl  -0.85  1.00  0.90  0.83 -0.70  0.78 -0.59 -0.81 -0.52 -0.49  0.53
disp -0.85  0.90  1.00  0.79 -0.71  0.89 -0.43 -0.71 -0.59 -0.56  0.39
hp   -0.78  0.83  0.79  1.00 -0.45  0.66 -0.71 -0.72 -0.24 -0.13  0.75
drat  0.68 -0.70 -0.71 -0.45  1.00 -0.71  0.09  0.44  0.71  0.70 -0.09
wt   -0.87  0.78  0.89  0.66 -0.71  1.00 -0.17 -0.55 -0.69 -0.58  0.43
qsec  0.42 -0.59 -0.43 -0.71  0.09 -0.17  1.00  0.74 -0.23 -0.21 -0.66
vs    0.66 -0.81 -0.71 -0.72  0.44 -0.55  0.74  1.00  0.17  0.21 -0.57
am    0.60 -0.52 -0.59 -0.24  0.71 -0.69 -0.23  0.17  1.00  0.79  0.06
gear  0.48 -0.49 -0.56 -0.13  0.70 -0.58 -0.21  0.21  0.79  1.00  0.27
carb -0.55  0.53  0.39  0.75 -0.09  0.43 -0.66 -0.57  0.06  0.27  1.00
```

Plot of the correlation of columns of data set “mtcars”

```
> install.packages("corrplot")
> library(corrplot)
> corrplot(cor(mtcars), type = "upper", order = "hclust",
tl.col = "black", tl.srt = 45)
```