# Intro to Statistics with R

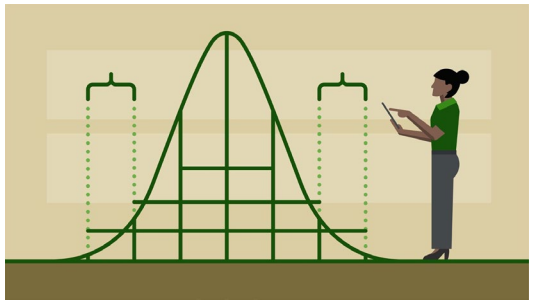
Workshop 2

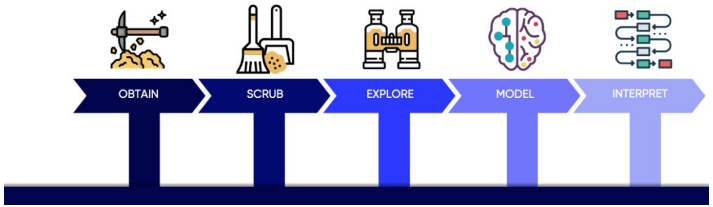
Course: VSK1004 Applied Researcher

# Our goal in the next 40 min

In this session, we will cover some of the **basic R lessons and principles of descriptive statistics**.

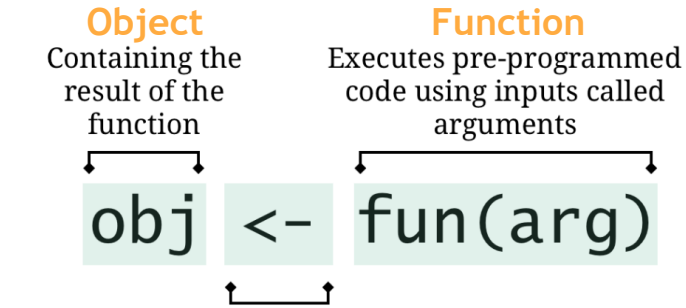
1. Four basic R lessons
2. Data scientific method
3. Data Cleaning
4. Data Exploration (Measures of Central Tendency and Variability)
5. Data Visualisation (Barplot, Boxplot, Histogram and Scatter Plot)





# 1. Four basic R lessons

# 1. Everything is an object




```r
# an object called x
x <- c(1,2,3,4)
```

```r
# an object that contains the mean() of x
mean_of_x <- mean(x)
```

```r
# print the object
print(mean_of_x)
```

```
[1] 2.5
```

# 2. Functions reside in packages

Install new package with install.packages()

```
# install package: Only do this once!
install.packages("dplyr")
```

Load existing packages with library()

```
# load package: EVERY TIME you write code
library(dplyr)
```

| Functions | Hidden functions |
| --- | --- |
| name<br>package::name | package:::name |
| **Datasets** | **Help files** (Vignettes) |
| data(name) | ?name<br>??name |

```
?cor
```

cor (stats) R Documentation

Correlation, Variance and Covariance (Matrices)

Description

`var`, `cov` and `cor` compute the variance of `x` and the covariance or correlation of `x` and `y` if these are vectors. If `x` and `y` are matrices then the covariances (or correlations) between the columns of `x` and the columns of `y` are computed.

`cov2cor` scales a covariance matrix into the corresponding correlation matrix efficiently.

Usage

```
var(x, y = NULL, na.rm = FALSE, use)

cov(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))

cor(x, y = NULL, use = "everything",
    method = c("pearson", "kendall", "spearman"))

cov2cor(V)
```

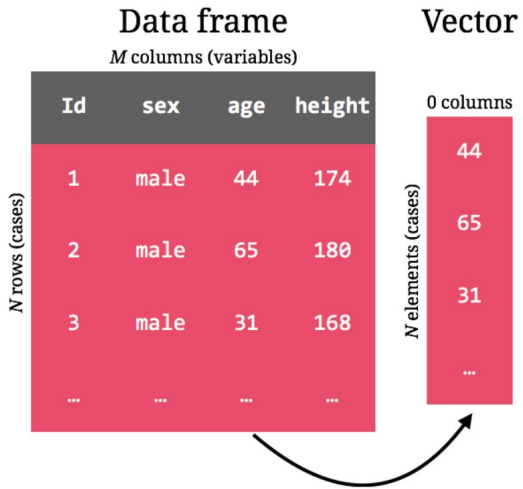
Arguments

| | |
| --- | --- |
| `x` | a numeric vector, matrix or data frame. |
| `y` | `NULL` (default) or a vector, matrix or data frame with compatible dimensions to `x`. The default is equivalent to `y = x` (but more efficient). |
| `na.rm` | logical. Should missing values be removed? |
| `use` | an optional character string giving a method for computing covariances in the presence of missing values. This must be (an abbreviation of) one of the strings `"everything"`, `"all.obs"`, `"complete.obs"`, `"na.or.complete"`, or `"pairwise.complete.obs"`. |
| `method` | a character string indicating which correlation coefficient (or covariance) is to be computed. One of `"pearson"` (default), `"kendall"`, or `"spearman"`: can be abbreviated. |
| `V` | symmetric numeric matrix, usually positive definite such as a covariance matrix. |

Don't forget to find help with ?

# 3. Data reside in data frames

Two-dimensional array where columns are variables and rows are the observations.



# 4. R data types

Get familiar with data structure/type. In R you can str() to check it out.

| *numeric* | *character* | *logical* |
|---|---|---|
| 42 | "a" | TRUE |
| 3.14 | "hi" | FALSE |

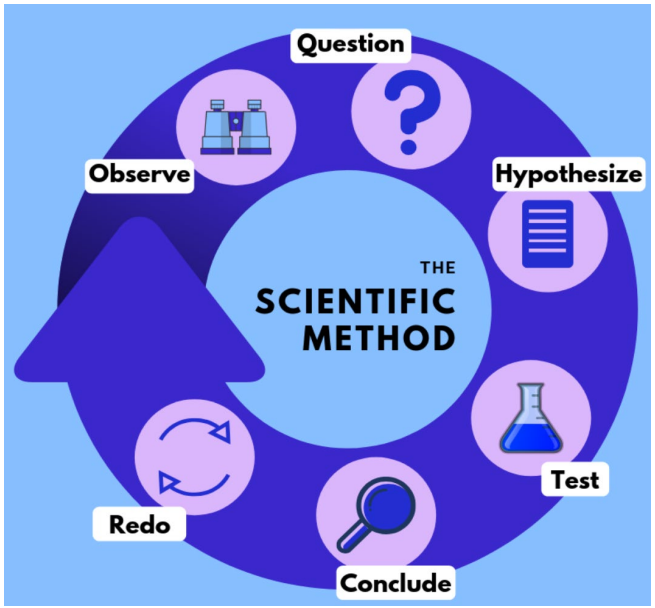
*factor*

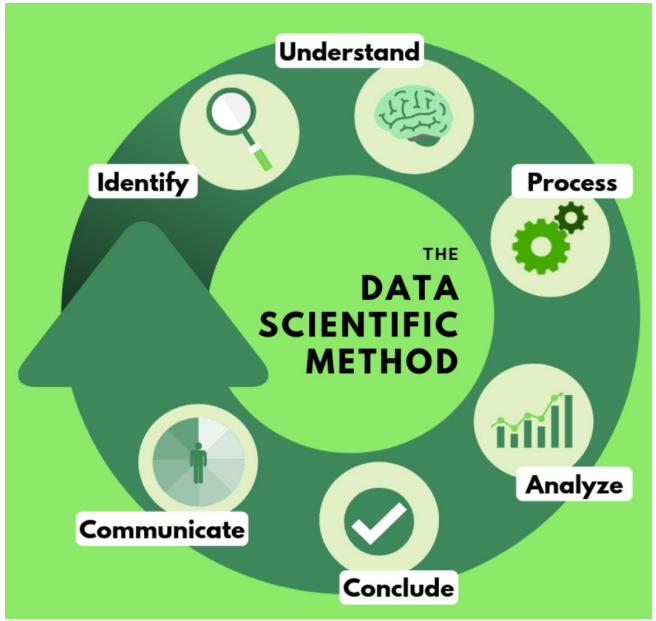
"a"

Levels: a

# 2. Data Scientific Method

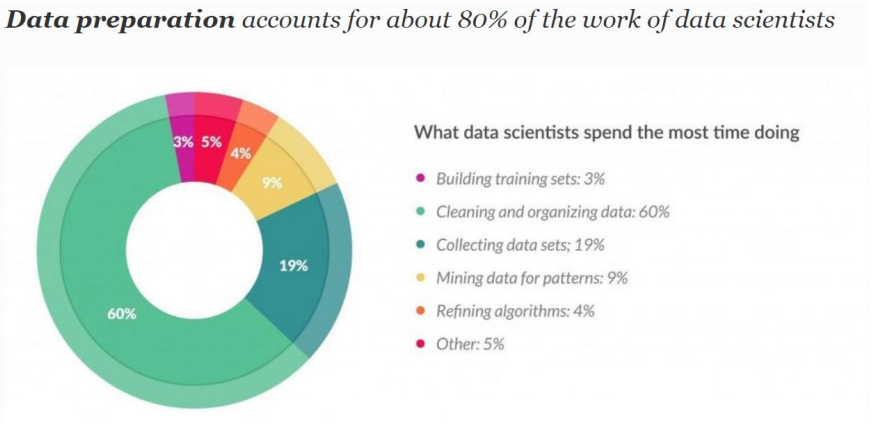
# Standardize the process of conducting experiments with data-intensive methods




https://towardsdatascience.com/a-data-scientific-method-80caa190dbd4

# Cleaning and organizing data




https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#22eaab266f63

# 3. Data Cleaning

Before we start exploring our data, we need to perform a set of data cleaning steps in order to enhance the quality of our dataset.

| **Steps** | **Actions** |
|---|---|
| Variable names | Removing inappropriate column names |
| Missing values | Checking how complete is your dataset |
| Categorical variables | Converting to dummy and factor variable |
| Data manipulation | Filtering subset of data |

Before we start exploring our data, we need to perform a set of **data cleaning** steps in order to enhance the quality of our dataset.

| Steps | Actions |
|---|---|
| Variable names | Removing inappropriate column names |
| **Missing values** | **Checking how complete is your dataset** |
| Categorical variables | Converting to dummy and factor variable |
| **Data manipulation** | **Filtering subset of data** |

# Missing values affect statistics and cause bias.

Missing values are those observations in your dataset that are empty.

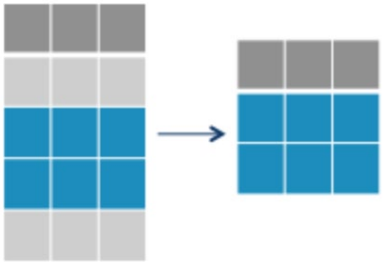
If the missing values are not handled properly, then we might end up drawing invalid conclusions about our data.

In R, missing values are often represented by NA or some other value that represents empty responses (i.e. -99).

Explore the best strategy to deal with missing data (i.e. imputation methods).

# Filtering data: return rows with matching conditions

Process of choosing a smaller part your data and using that subset for analysis.



Filtering generally is used to:

Look at records from particular period.
Exclude errors or “bad” observations from your analysis.

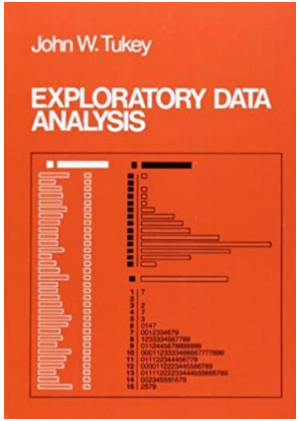
You need to specify the rule or logic to identify cases:

```
filter(starwars, species == "Human")
filter(starwars, mass > 1000)
```

# 5. Data Exploration

Once we ‘clean’ the data, we always look for ways to understand our dataset. Some of the common measurements in descriptive statistics are central tendency and variability:

| Type | Examples |
|---|---|
| Central Tendency | Mean, mode, median |
| Variability | Variance, standard deviation |



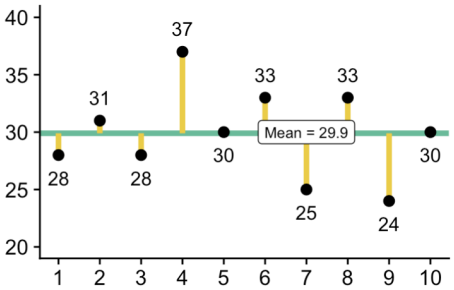

*“Helping you in the discovery process”*
**Classic EDA book, Tukey (1977)**

# Central Tendency

It describes your data with a single value that represents the centre of its distribution. The main measures of central tendency are:

## Mean

It is the sum of the observation divided by the sample size. It is affected by extreme values and missing values. In R you can use mean().

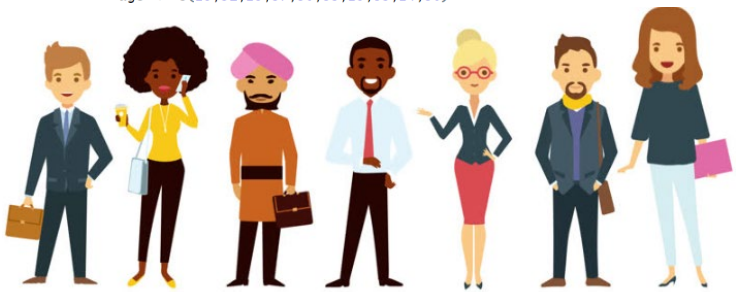



$$Mean = \frac{28 + 31 + 28 + \ldots}{10} = 29.9$$

## Median

It is the middle value of your data. It splits the data in half and called 50th percentile. In R, you can use median().

```
# Age of the  participants
age <- c(28,31,28,37,30,33,25,33,24,30)
```



tly in the

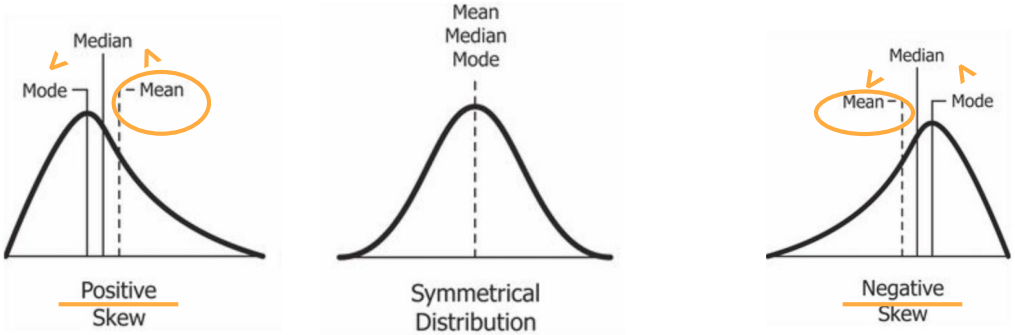
```
uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

How old are you? n = 10 participants

```
> getmode(age)
[1] 28
```

# Symmetry in data distribution: Skewness

It is the degree of distortion from a symmetrical or normal distribution:



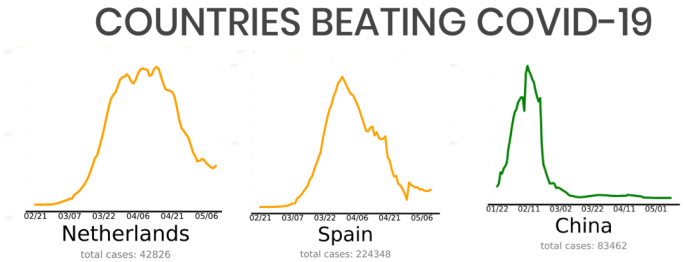
# Variability

It represents the amount of dispersion of your dataset. How spread out are the values?

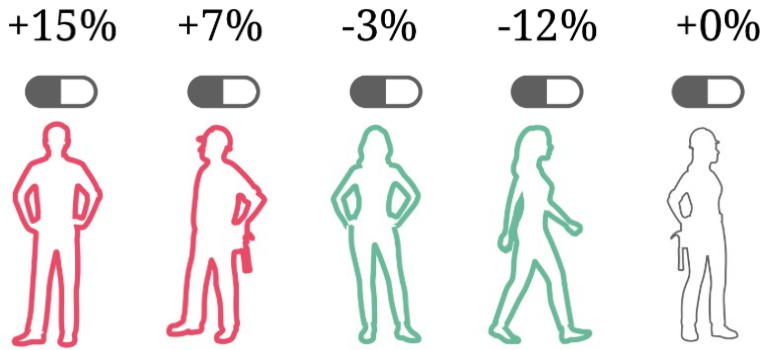
All interesting data processes have variability:

- Daily new cases of *COVID-19* change over time.
- Individual patients respond to drugs differently.

If there was no variability, statistics would be not longer needed.




https://www.endcoronavirus.org/countries#action



# Variability

The most common measures of statistical variability (or dispersion) are:

## Variance

- It helps determine the size of the data spread.
- Average of the squared differences from the mean.

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$S^2$ = sample variance
$x_i$ = the value of the one observation
$\bar{x}$ = the mean value of all observations
$n$ = the number of observations

You can use the var() function in R.

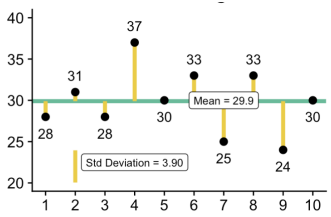
## Standard Deviation

- It measures the absolute variability of the dispersion.
- Square root of the variance.

$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}$$

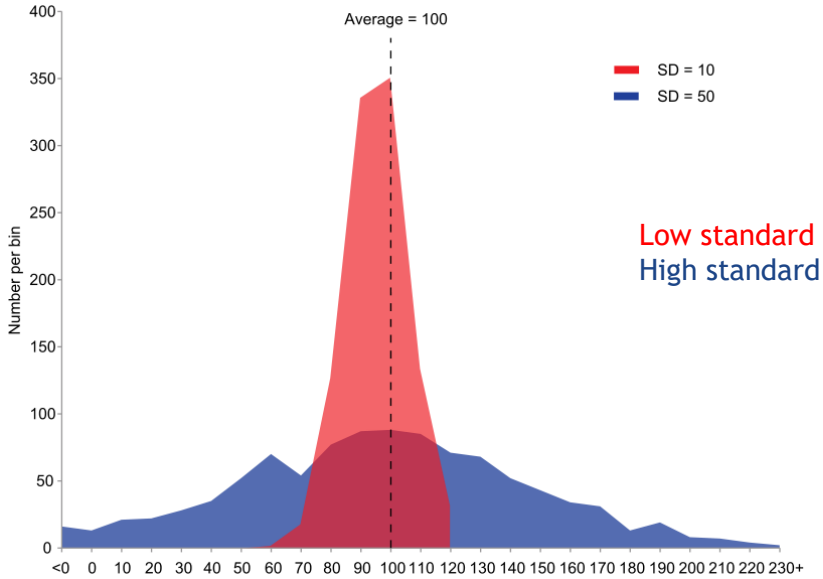
$s$ = sample standard deviation
$N$ = the number of observations
$x_i$ = the observed values of a sample item
$\overline{x}$ = the mean value of the observations

You can use the sd() function in R.




$$Stand.\ Dev. = \sqrt{\frac{(28-29.9)^2 + (31-29.9)^2 + \ldots}{10-1}} = 3.90$$

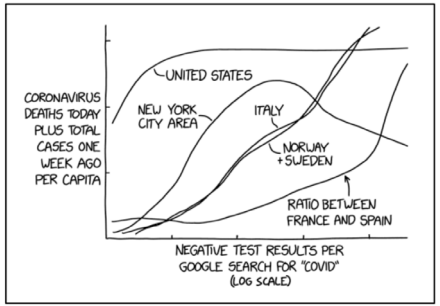
# Example: two samples with same mean but different variances




Low standard deviation (SD) values -> most of data are very close to the mean
High standard deviation (SD) values -> your observations are spread out

# 5. Data Visualisation

Once we explore the data with descriptive statistics, we can use **graphs** to show and **capture** some (un)expected aspects of our dataset, **synthesize information** and **communicate efficiently.**




I'M A HUGE FAN OF WEIRD GRAPHS, BUT EVEN I ADMIT SOME OF THESE CORONAVIRUS CHARTS ARE LESS THAN HELPFUL.

https://xkcd.com/

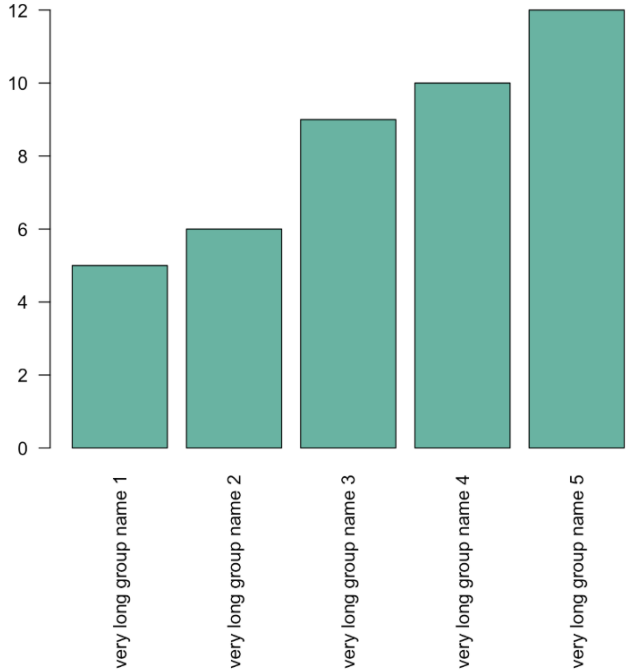
# Bar plots

Comparison of categorical data.

2-dimensional:
- category axis:: *group*
- value axis:: *value (e.g. number of students)*

Use bar plot when you have many categories.

Order categories to transmit a clear message.

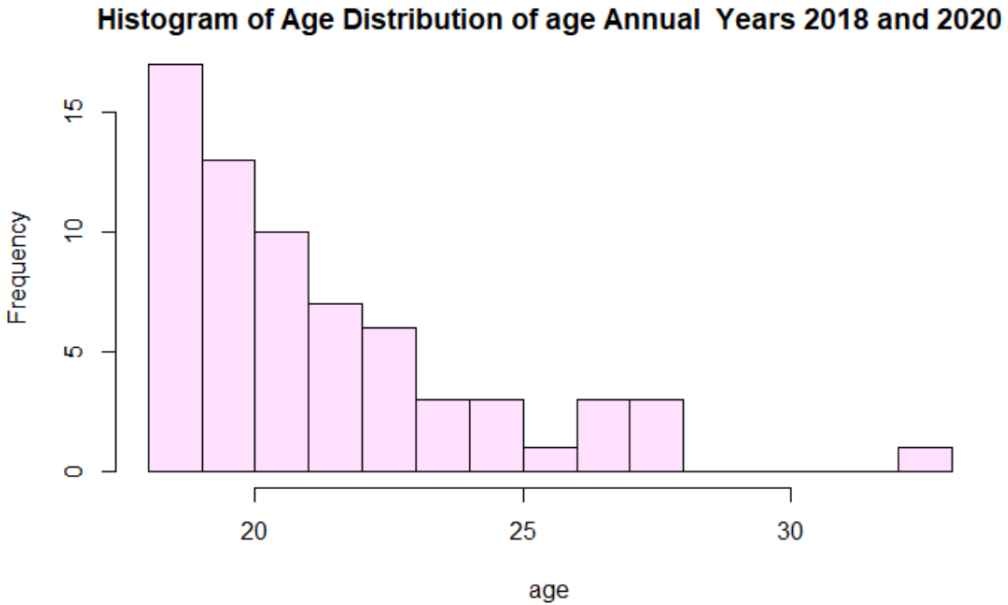


# Histograms

Similar to bar plot but it represents a numerical (i.e. age) variable.

x-axis:: scale of measurements (age)
y-axis:: number of times value occurred

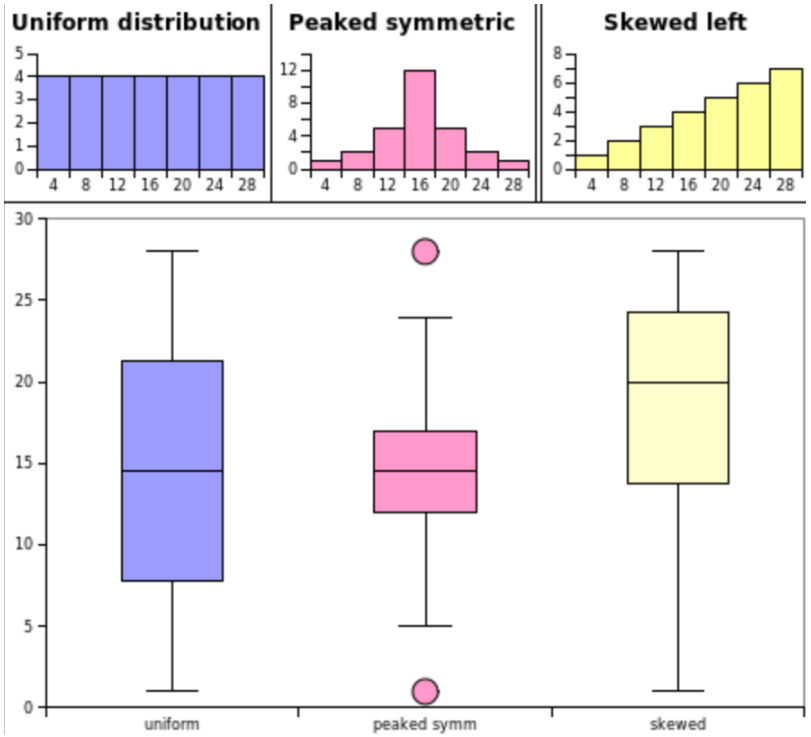
Visual representation of data distribution (e.g. mean, median, outliers)



# Box plots

Descriptive values of your dataset (minimum value, first quantile, the median, the third quartile and the maximum value)

Display boxplot and histogram together provides greater insights of your data distribution.



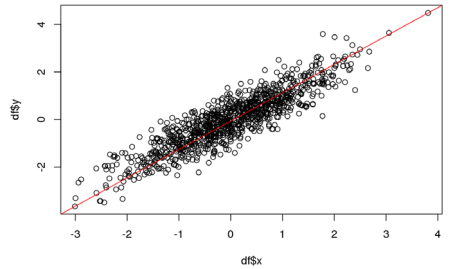
# Bivariate Scatter Plot

Axes = variables.

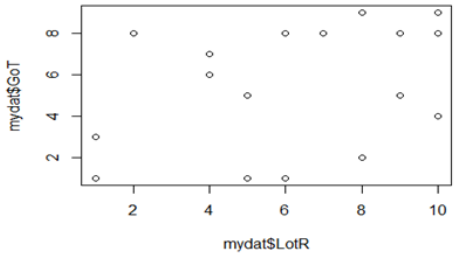
Points in two-dimensional space.

Useful for small-medium size dataset.

Look for structure patterns: circular or linear relationship.

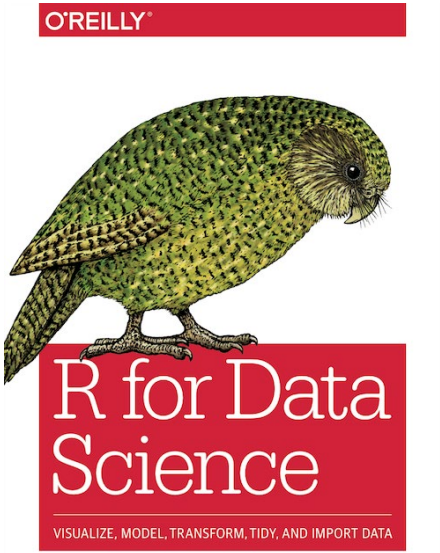



**Scatter plot - Linear association**




**Scatter plot - No association**

# Recommended book




https://r4ds.had.co.nz/

# Let's practise!

# Questions?