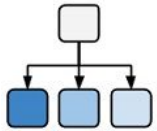
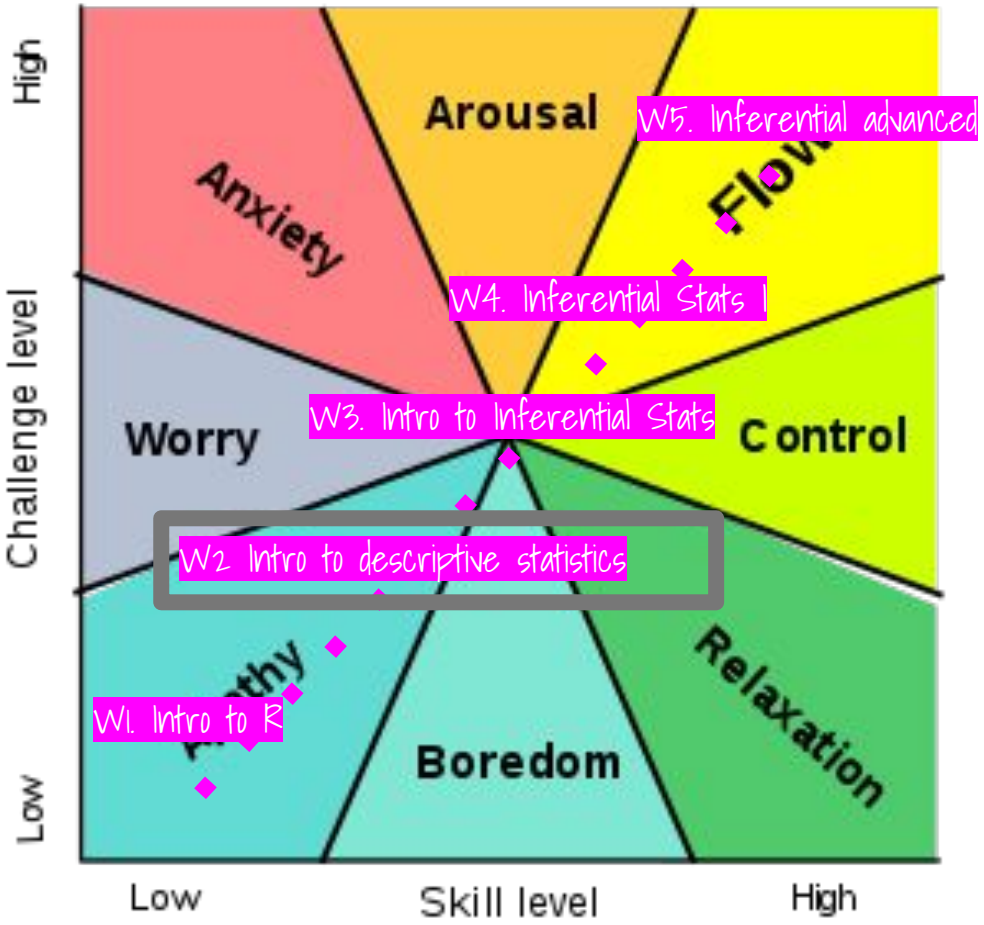
R Workshop II

Carlos Utrilla Guerrero
Institute of Data Science - Researcher



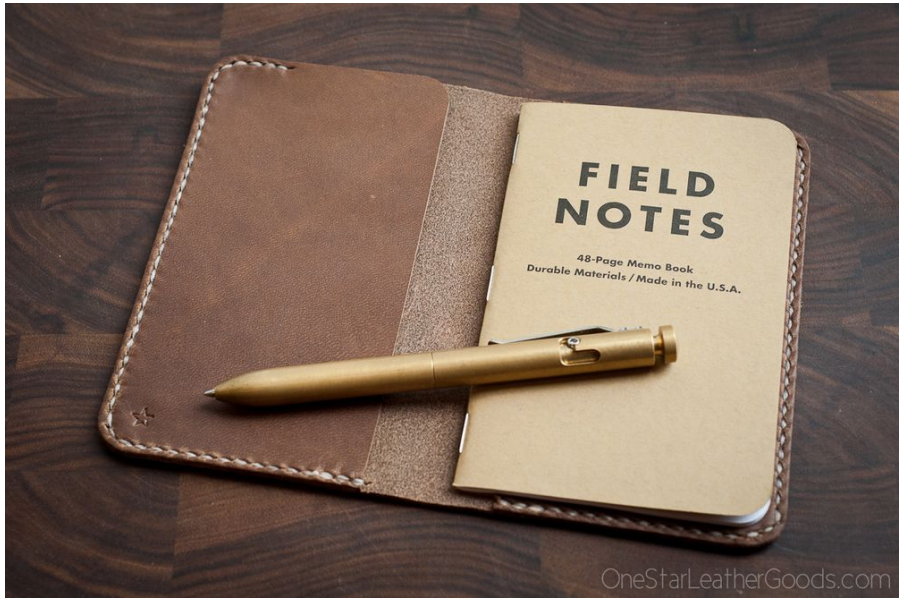
Course: VSK1004 Applied Researcher

Recap...

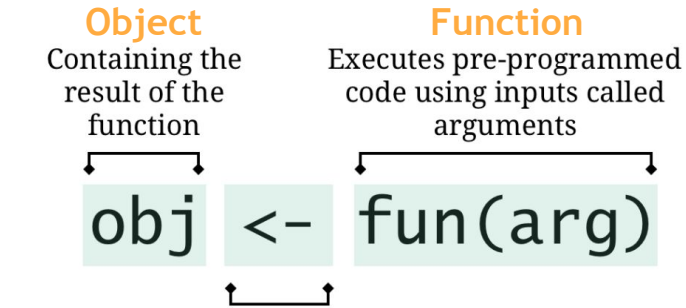

High
Challenge level
Low
Low
Skill level
High
Arousal
Anxiety
Flow
Worry
Control
Boredom
Relaxation
W5. Inferential advanced
W4. Inferential Stats I
W3. Intro to Inferential Stats
W2 Intro to descriptive statistics
W1. Intro to R

What we are covering today

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)



Lesson 1: Apply function to objects




```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
```

Lesson 2: Functions reside in packages

R: New Phone

R Packages:
Apps you can download

Lesson 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 package::name	package:::name
Datasets	**Help files** (Vignettes)
data(name)	?name ??name

Don't forget to find help with ?

```
?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

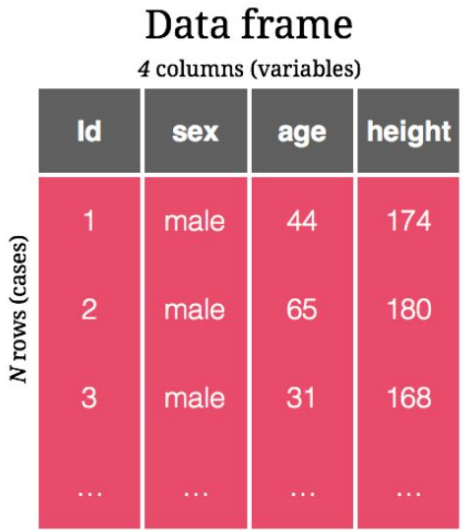
```
var(x, y = NULL, na.rm = FALSE, use)

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.

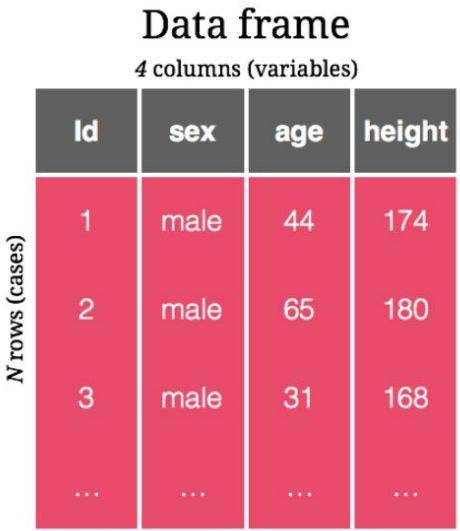
Lesson 3: Data reside in dataframes

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



Lesson 3: Data reside in dataframes

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

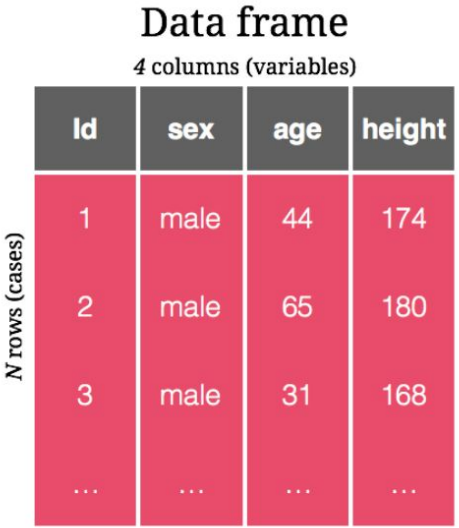



```
# inspect baselers via print
View(baselers)
```

	id	sex	age	height	weight	income
1	1	male	44	174.3	113.4	6300
2	2	male	65	180.3	75.2	10900
3	3	female	31	168.3	55.5	5100
4	4	male	27	209.0	93.8	4200
5	5	male	24	176.7	*NA*	4000
6	6	male	63	186.6	67.4	11400
7	7	male	71	151.6	83.3	12000
8	8	female	41	155.7	67.8	7600
9	9	male	43	176.1	69.3	8500
10	10	female	31	166.1	66.3	6100
11	11	female	42	157.8	51.9	8000

Lesson 3: Data reside in dataframes

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




```
# inspect baselers via print
View(baselers)
```

	id	sex	age	height	weight	income
1	1	male	44	174.3	113.4	6300
2	2	male	65	180.3	75.2	10900
3	3	female	31	168.3	55.5	5100
4	4	male	27	209.0	93.8	4200
5	5	male	24	176.7	*NA*	4000
6	6	male	63	186.6	67.4	11400
7	7	male	71	151.6	83.3	12000
8	8	female	41	155.7	67.8	7600
9	9	male	43	176.1	69.3	8500
10	10	female	31	166.1	66.3	6100
11	11	female	42	157.8	51.9	8000

```
# inspect baselers via print
baselers
```

```
## # A tibble: 10,000 x 20
##       id sex      age height weight
##    <int> <chr>  <int>  <dbl>  <dbl>
##  1     1 male      44   174.  113.
##  2     2 male      65   180.   75.2
##  3     3 female    31   168.   55.5
##  4     4 male      27   209    93.8
##  5     5 male      24   177.   NA
##  6     6 male      63   187.   67.4
##  7     7 male      71   152.   83.3
##  8     8 female    41   156.   67.8
##  9     9 male      43   176.   69.3
## 10    10 female    31   166.   66.3
## # ... with 9,990 more rows, and 15 more
## #   variables
```

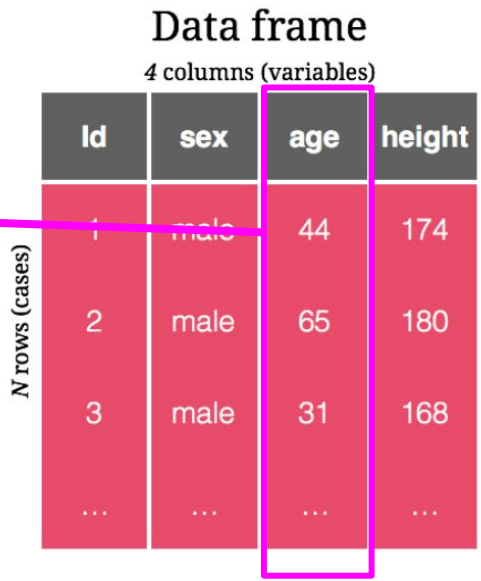
see The Elements of Data Analytic Style by Jeff Leek

Lesson 3: Data reside in dataframes

Select a column via $

```
# select age variable
baselers$age
```

```
##  [1] 44 65 31 27 24 63 71 41 43 31 42 31
## [13] 38 49 39 54 78 62 88 74
```



Lesson 4: Vector and data types

Select/Change/(Add) via []

```
# extract vector containing age
age <- baselers$age
age
```

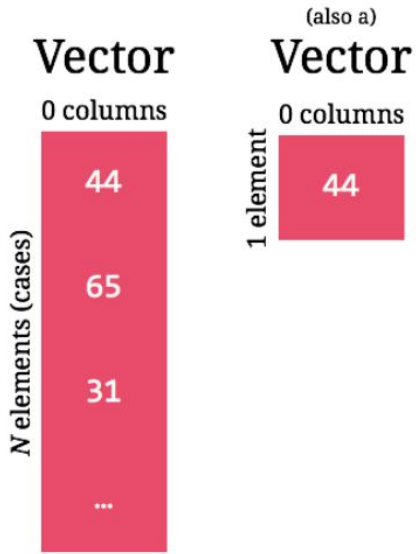
```
## [1]  88 130  62  54  48 126 142  82  86
```

```
# select value
age[2]
```

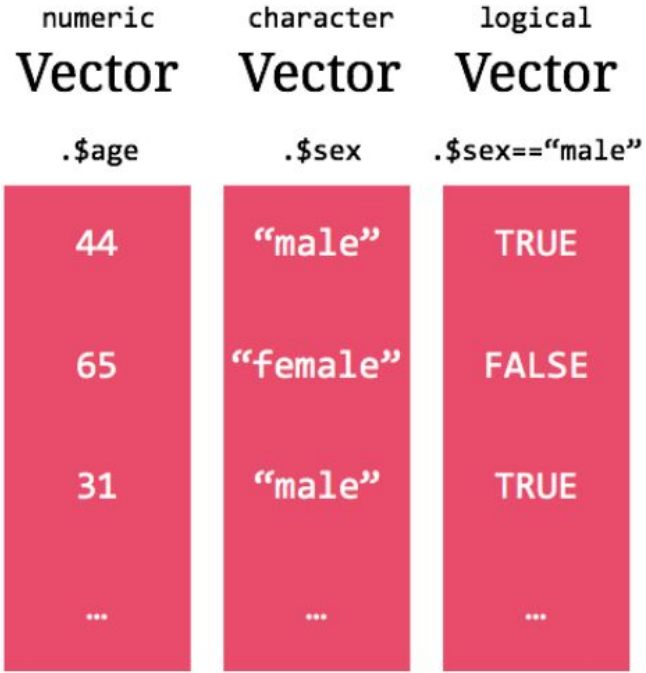
```
## [1] 130
```

```
# change value
age[2] <- 2
age
```

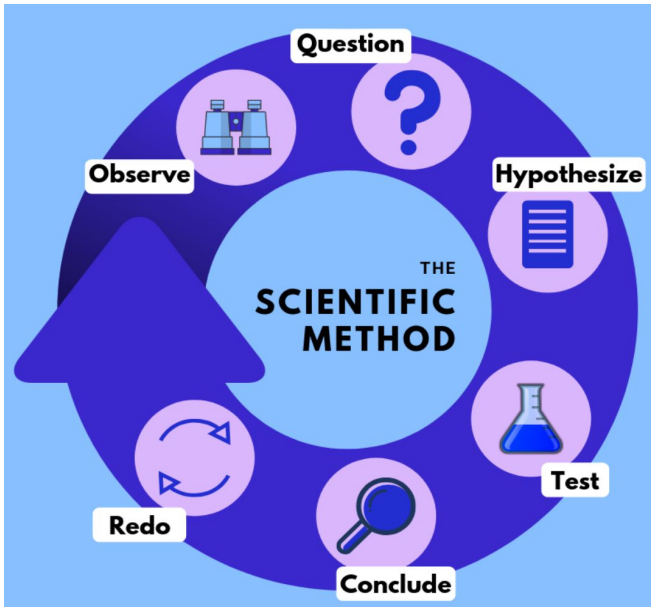
```
## [1]  88   2  62  54  48 126 142  82  86
```

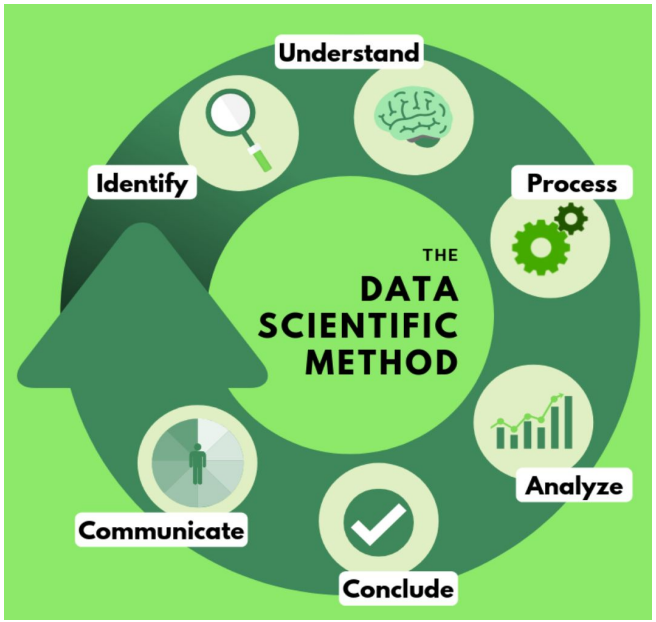


Lesson 4: Vector and data types



Standardize the process of conducting experiments with data-intensive methods





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

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.

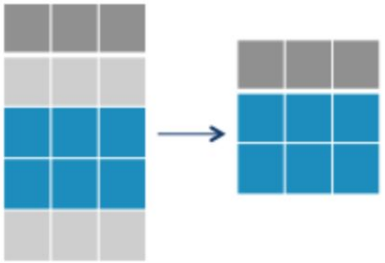
In R, missing values are often represented by NA or some other value that represents empty responses (i.e. -99).

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.



Filtering data: return rows with matching conditions

Eye_colour	Hight	Weight	Age
Blue	1.8	65	31
Brown	1.9	73	34
Blue	1.7	74	64
Blue	1.9	87	45

What is the average age for people that have blue eyes?

Filtering data: return rows with matching conditions

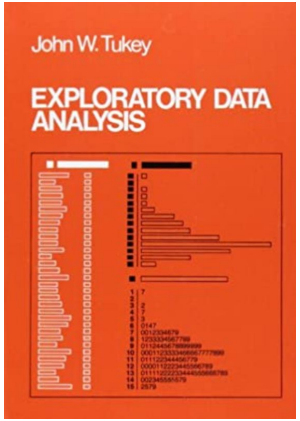
Eye_colour	Hight	Weight	Age
Blue	1.8	65	31
Brown	1.9	73	34
Blue	1.7	74	64
Blue	1.9	87	45

What is the average age for people that have blue eyes?

```
blue_eyes_data <- filter(mydat, Eye_color == "Blue")  #filter mydat for specific eye colour

mean(blue_eyes_data$Age)  #average/mean age of eye colour people
```

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



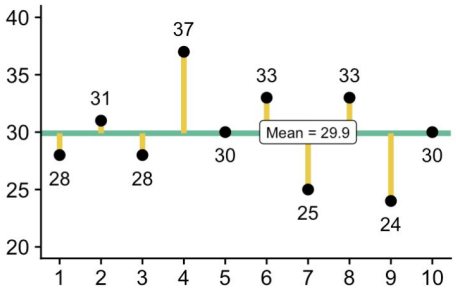
"*Helping you in the discovery process*"
Classic EDA book, Tukey (1977)

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
```

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.

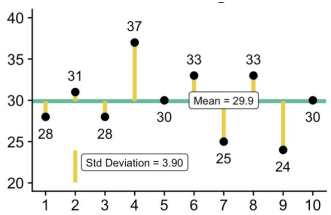
Standard Deviation

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

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

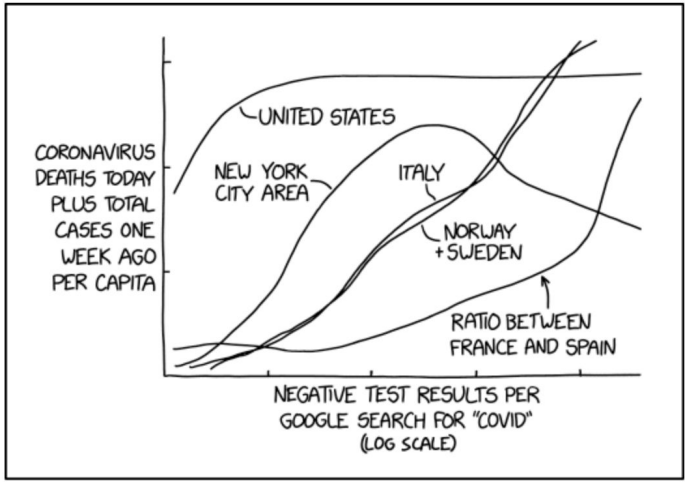
s = sample standard deviation
N = the number of observations
x_i = the observed values of a sample item
$\bar{x}$ = the mean value of the observations

You can use the sd() function in R.



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

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/

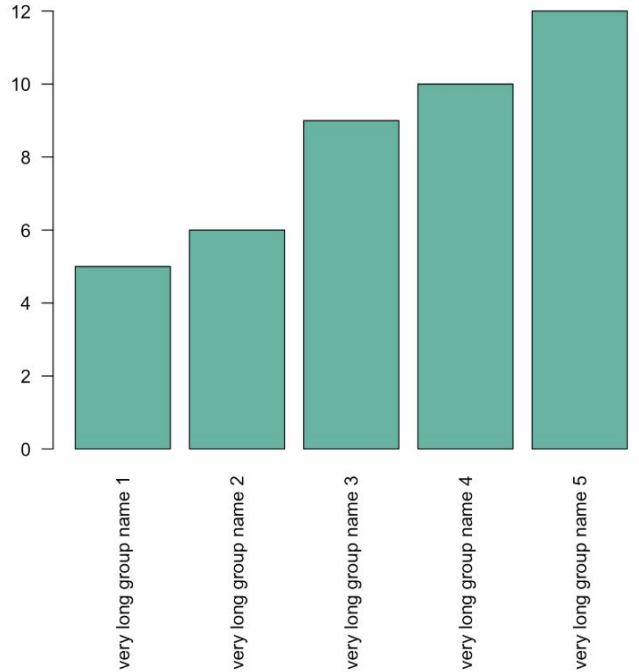
Bar plots

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.

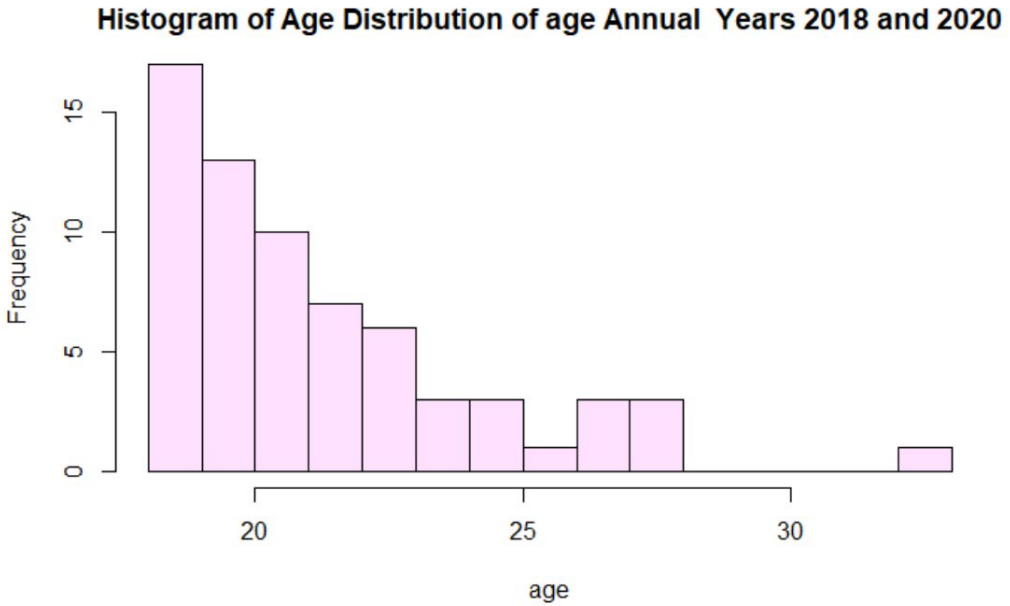


Histograms

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

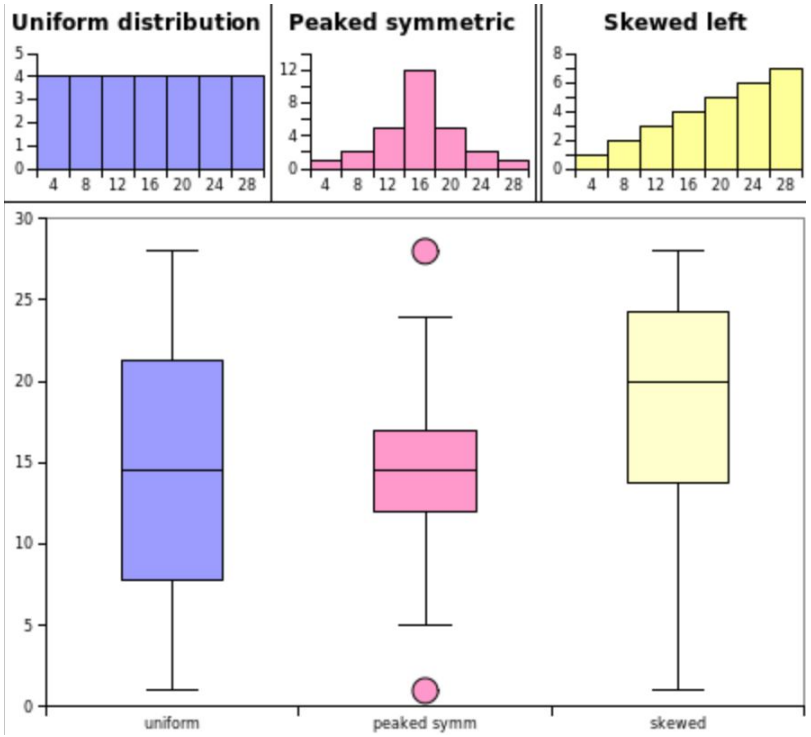
Visual representation of data distribution (e.g. mean, median, outliers)



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.**



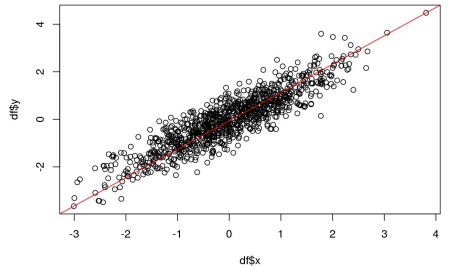
Bivariate Scatter Plot

Axes = variables.

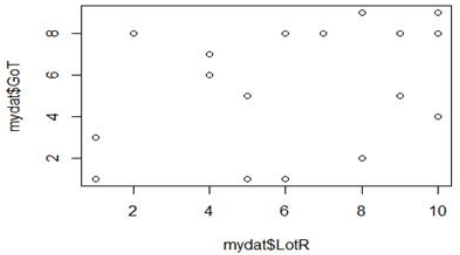
Points in two-dimensional space.

Useful for small-medium size dataset.

Look for structure patterns: **circular** or **linear** relationship.

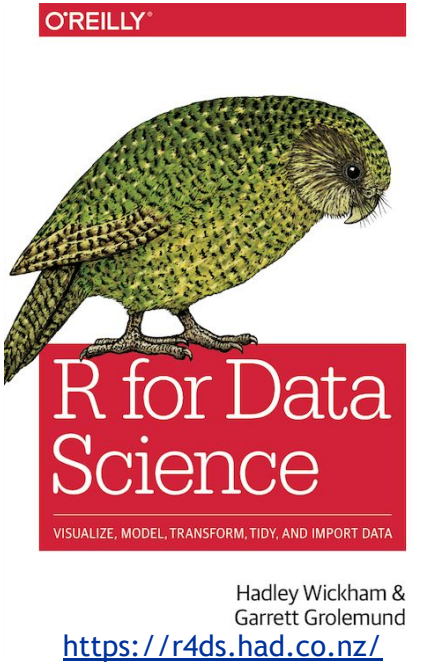


Scatter plot - Linear association



Scatter plot - No association

Recommended book



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