THE FIRST STEP
EXPLORATORY DATA ANALYSIS

Alejandro Ordonez

Assistant Professor - Department of Bioscience

Section for Ecoinformatics & Biodiversity

Center for Biodiversity Dynamics in a Changing World (BIOCHANGE)

EXPLORATORY DATA ANALYSIS
WHAT WE WILL TALK ABOUT TODAY

A very quick recap on **R** basics

Exploratory Data Analysis - Why?

Summary statistics – Get an overview of your data

Visualisation – A figure is worth 1000 words

Transformations – Making data symmetrical an comparable

Modelling – Establish simple relations between variables

Statistical Test – What are they doing?

Any questions?

Ready to start?

AN INTRODUCTION TO R

What is R?

- R is an...

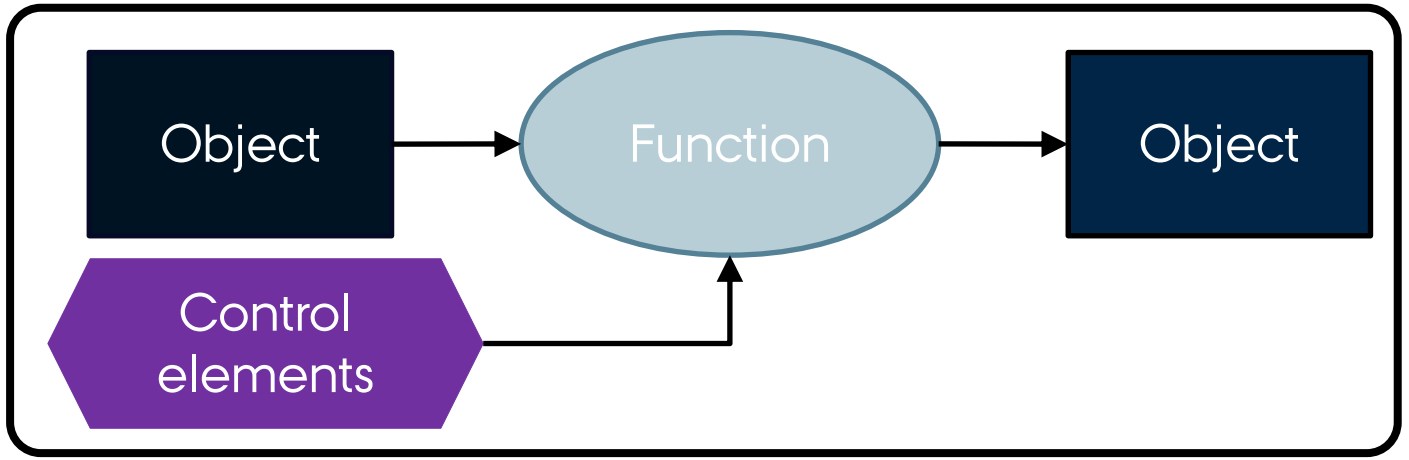
object-oriented statistical programming language.

Object-oriented programming is a programming model organised around **objects** rather than "actions" and data rather than logic.	A **programming language** is a vocabulary and set of rules for instructing a **computer** to perform specific tasks.

THE R WAY OF DOING THINGS
STRUCTURE – FUNCTIONS/METHODS

R focuses on three types of elements

- Functions/Methods: ***what do you want to do?***
- Control elements: ***when/how you want to do something?***
- Objects: ***on what do you want to use a function?***



THE R WAY OF DOING THINGS
STRUCTURE – FUNCTIONS/METHODS

Function/Method Argument Object

```
sqrt(x = 16)
```

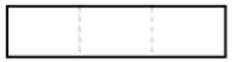
How this looks in R

```
sqrt(x = 16)
[1] 4
out = sqrt(x = 16)
out
[1] 4
```

THE R WAY OF DOING THINGS
STRUCTURE – OBJECTS

Vector

```
a = c(4,2,5,10)
```



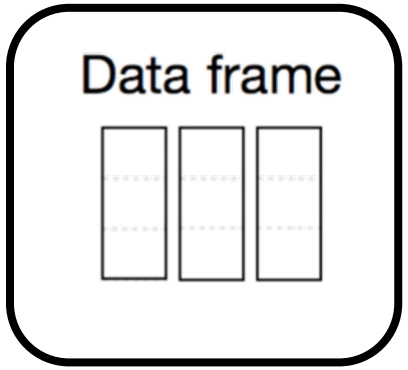
List

```
List <- list(id = c (1:5),
             name = c(”R","D","M","R", "G"),
             salary = c(62, 51, 61, 72, 84))
```

Matrix

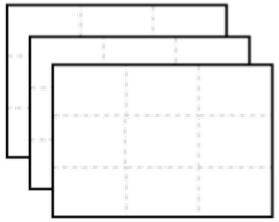
```
A = matrix(data = 0,
           nrow = 6,
           ncol = 5)
```

Data frame



The most useful as it stores information as tabular data and collates different information types into one object.

Array



```
Data <- data.frame(id = c (1:5),
                   name = c(”R","D","M","R", "G"),
                   salary = c(62, 51, 61, 72, 84))
```

THE R WAY OF DOING THINGS
READING AN SAVING INFORMATION

The generic Function [Generally you need to specify ALL arguments]

- `read.table()` ➔ generic function to read tables
- `write.table ()` ➔ generic function to save a table

The one you will use most [Often you need one specify file name]

- `read.csv()` ➔ reading coma separated files

```
myData = read.csv("some data.csv")
```

- `write.csv()` ➔ saving coma separated files

```
myData = read.csv(x = myData,
                  file = "updated data.csv")
```

THE R WAY OF DOING THINGS
PACKAGES

What are Packages?

- Packages are collections of functions that have a specific function.
- You need to install these and call these before you can use them → **only once**
- Many packages depend on other packages – dependencies

```
install.packages()
require(package.name)
```

Installing packages via the console

So far so good?

Any questions?

Ready to continue?

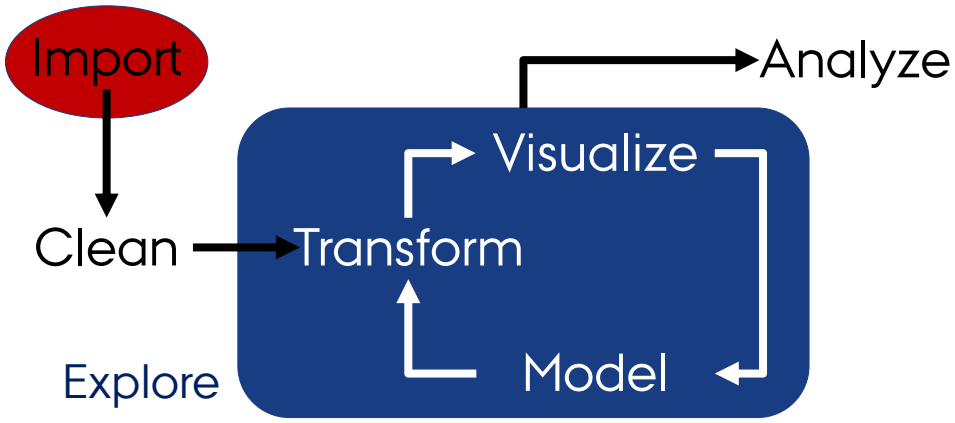
EXPLORATORY DATA ANALYSIS: WHY?

Exploratory Data Analysis (**EDA**) is *the art of looking at your data* to rapidly generate many promising leads that you can later explore in more depth.

EDA is the base for statistical analysis of multidimensional data by:

- Provide an overview of the data → is it meaningful?
- Establish if there are errors, outliers, or patterns.
- Help to determine the need to transform or recode some variables → are the analytical assumptions meet?
- Orient further analyses.

BEFORE EXPLORATORY DATA ANALYSIS:
GET YOUR DATA IN.



Before doing anything in R you need to get your data in the program.

External data is usually a txt or csv file that needs to be "read" using one of two functions.

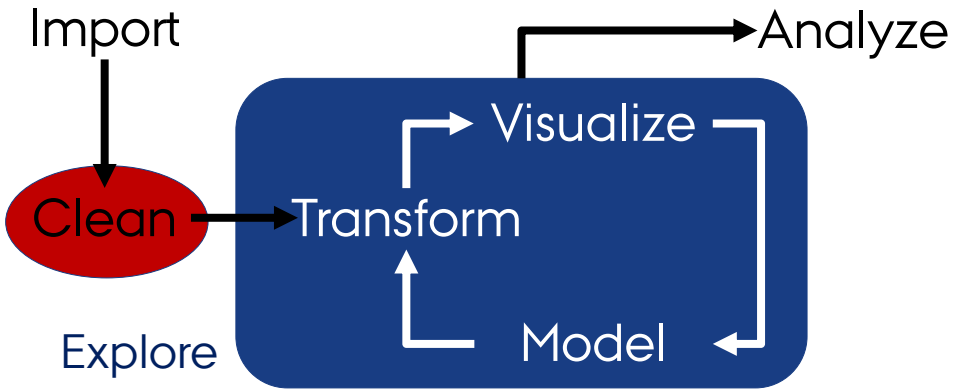
```
read.table?
read.csv?
```

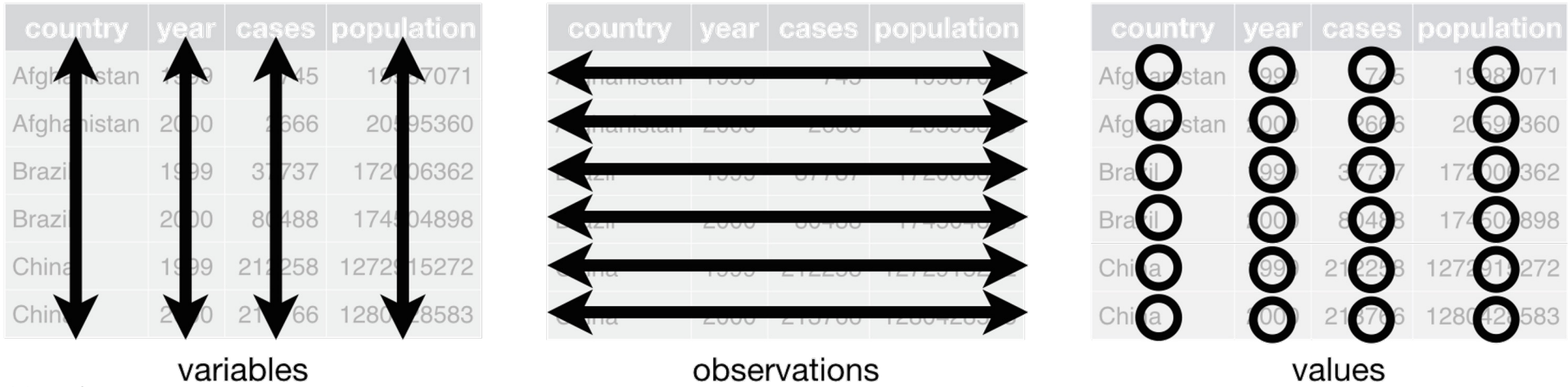
Reading is not enough, It needs to be stored it as an object!!

BEFORE EXPLORATORY DATA ANALYSIS: CLEAN YOUR DATA

There are three interrelated rules which make a dataset suited to be used in R:

- Each variable must have its own column.
- Each observation must have its own row.
- Each value must have its own cell.





AARHUS UNIVERSITY
DEPARTMENT OF BIOSCIENCE

STATISTICAL AND GEOSPATIAL MODELLING
DATA EXPLORATION

ALEJANDRO ORDONEZ GLORIA
ASSISTANT PROFESSOR

Figures: Grolemund and Wickham (2017) R for Data Science

BEFORE EXPLORATORY DATA ANALYSIS:
CLEAN YOUR DATA.

Step 1. Remove

- Leading/Trailing spaces.
- Comma, dots, hyphens and apostrophes.
- Spaces and Multiple spaces.

Step 2. Standardise

- To lowercase/uppercase *Or* Capitalise.
- Put/remove a dots at the end and punctuation.
- Change "," to ";" in the text.
- Date formats.
- Precision [N° of decimals]

Step 3. Error checking

- Correcting or removing corrupt or inaccurate records.
- Eliminate duplicate entries.
- Data type consistency within a column.

Figures: Grolemund and Wickham (2017) R for Data Science

EXPLORATORY DATA ANALYSIS
WHAT TO DO BEFORE YOU START?

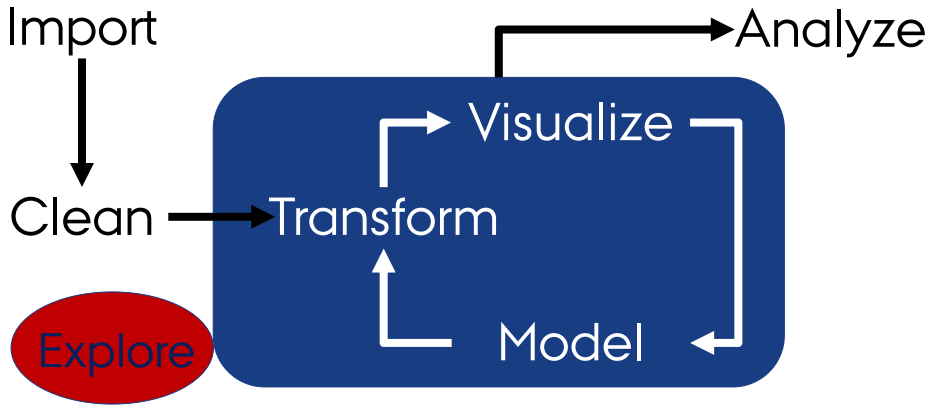
The two crucial questions you will **always** ask:

1. *What type of **variation** occurs within my variables?*

   Variation *is a description of how a random variable is dispersed, or spread out.*

2. *What type of **covariation** occurs between my variables?*

   Covariation *is a description of the relationship between two random variables*



EXPLORATORY DATA ANALYSIS: TOOLS

EDA uses **four** main tools to answer the two key questions:

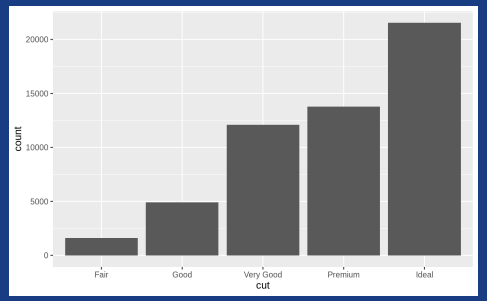
- **Descriptive statistics**: Quantitative descriptions or summaries of the features from a collection of information.
- **Visualisation**: The graphical representation of information and data to see and understand trends, outliers, and patterns.
- **Transformation**: process of changing the format, structure, or values of data.
- **Modelling**: simple low-dimensional summaries of a dataset, and trends assessments.

EXPLORATORY DATA ANALYSIS: TOOLS

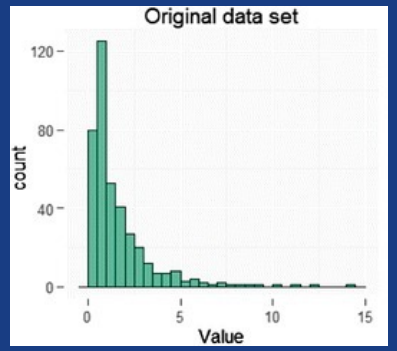
Descriptive statistics

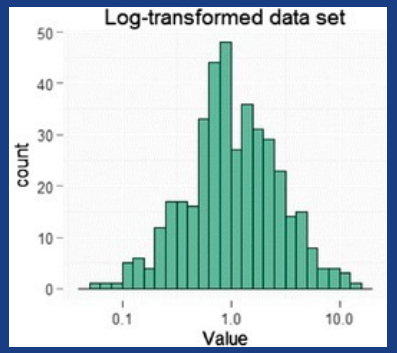
Country	Gender	Height	Basketball players
United Kingdom	Male	$1.95 \pm 7\%$	50% (1/2)
United Kingdom	Female	$1.8 \pm 0\%$	100% (1/1)
Canada	Male	NA	NA
Canada	Female	$1.8 \pm 14\%$	50% (1/2)
Germany	Male	$2.1 \pm 0\%$	100% (1/1)
Germany	Female	NA	NA

Visualisation

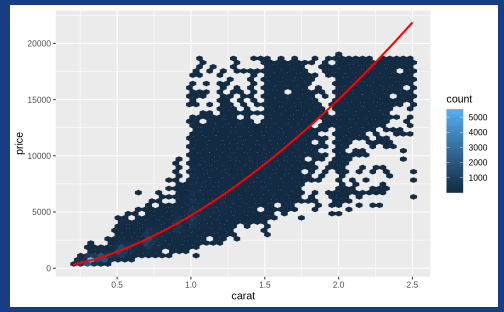


Transformation





Modelling



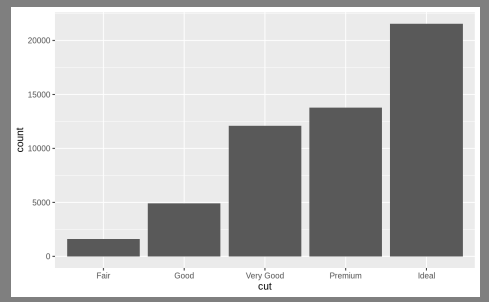
Figures: Grolemund and Wickham (2017) R for Data Science

EXPLORATORY DATA ANALYSIS: TOOLS

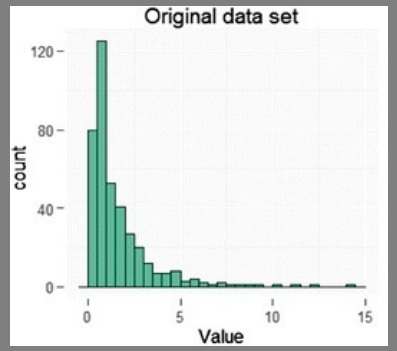
Descriptive statistics

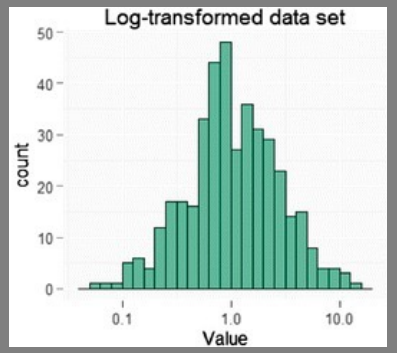
Country	Gender	Height	Basketball players
United Kingdom	Male	$1.95 \pm 7\%$	50% (1/2)
United Kingdom	Female	$1.8 \pm 0\%$	100% (1/1)
Canada	Male	NA	NA
Canada	Female	$1.8 \pm 14\%$	50% (1/2)
Germany	Male	$2.1 \pm 0\%$	100% (1/1)
Germany	Female	NA	NA

Visualisation

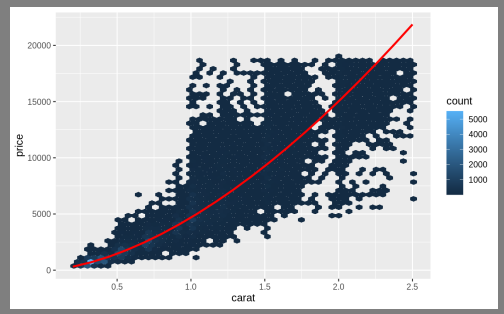


Transformation





Modelling



Figures: Grolemund and Wickham (2017) R for Data Science

EDA – DESCRIPTIVE STATISTICS

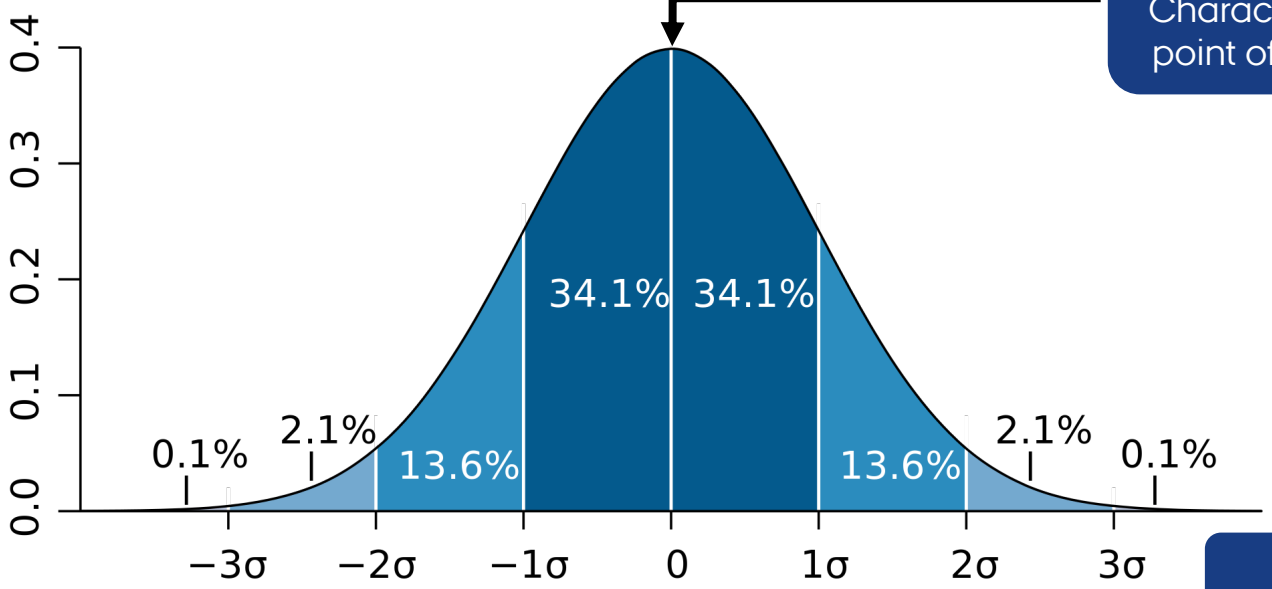
Summaries provide a good way to:

- Look at patterns and the quality of your data.
- Communicate the largest amount of information as simply as possible.

All your observations regarding a single variable can be described in terms of:

Central tendency (*mean, median, mode*).
Dispersion (*variance, standard deviation, coefficient of variation*).
Shape of the distribution (*skewness, kurtosis*).
Statistical dependence (correlation coefficients).

EDA – DESCRIPTIVE STATISTICS



Central tendency

Characterises the central point of the observations

Mean
Median
Mode

Dispersion

Characterises the variability of the observations

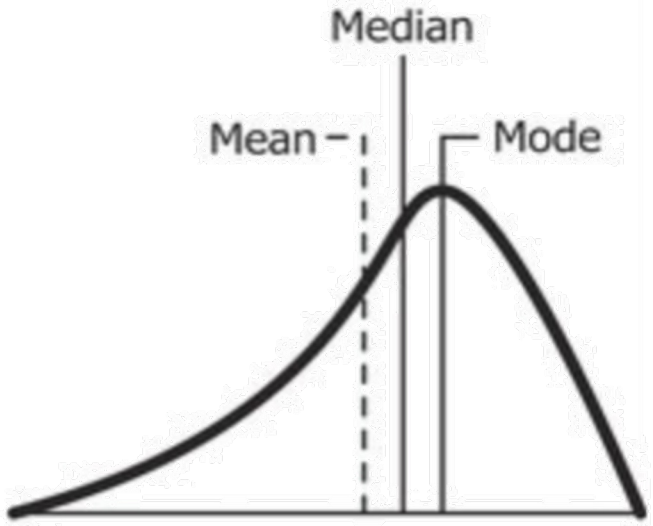
σ^2 = Variance

σ = Standard Deviation

$\frac{\sigma}{\sqrt{n}}$ =Standard Error

$$\sigma^2 = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$

EDA – DESCRIPTIVE STATISTICS



Negative Skew

$\mu <$ Md or Mo

Shape

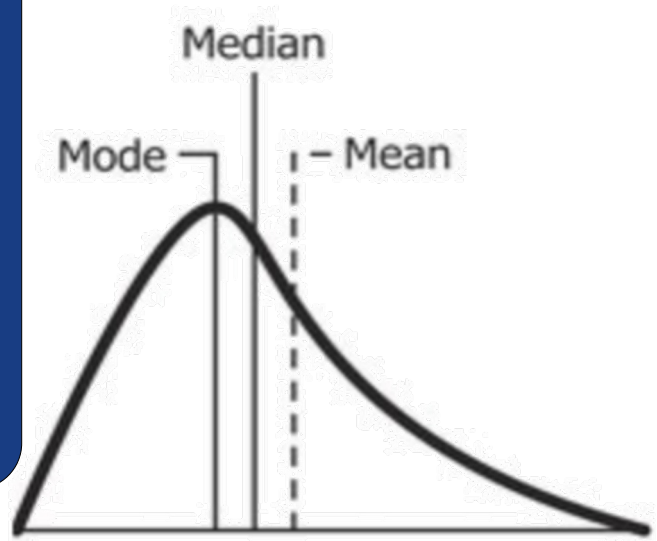
Characterize the deviation from a normal (symmetrical) distribution

Skewness

$$Sk_1 = \frac{\mu - Mo}{\sigma^2}$$

$$Sk_2 = \frac{3(\mu - Md)}{\sigma^2}$$

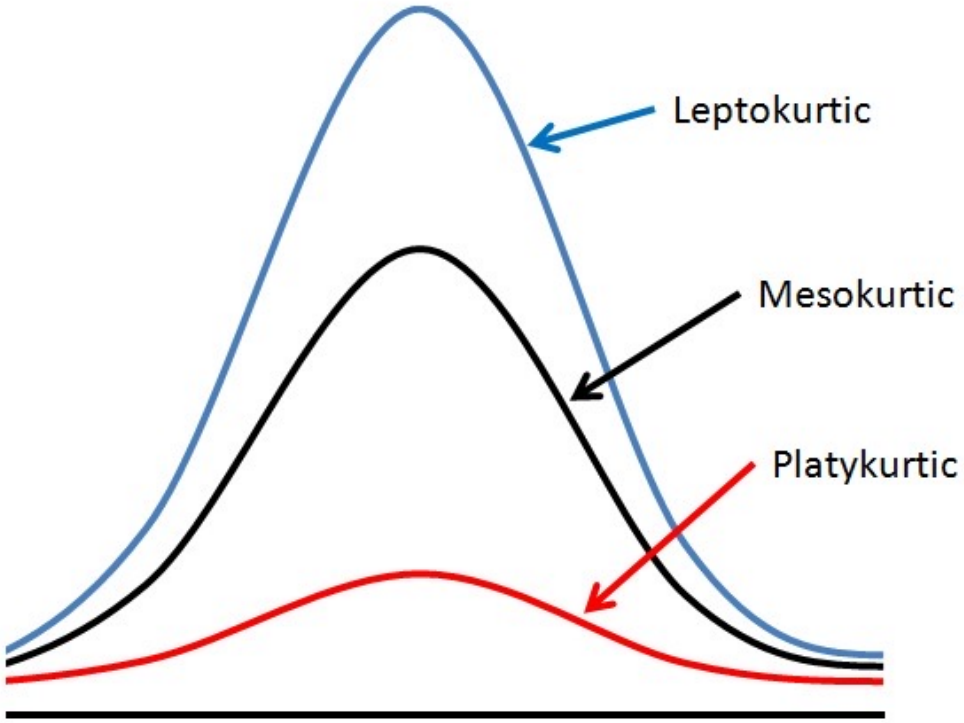
Skewness direction given by the sign



Positive Skew

$\mu >$ Md or Mo

EDA – DESCRIPTIVE STATISTICS



Shape

Help to identify outlier problems.

$$g_2 = \frac{m_4}{m_4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4\right]^2} - 3$$

Kurtosis

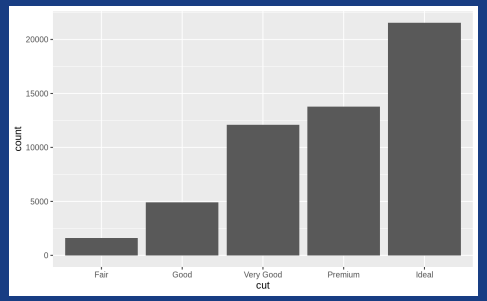
Larger kurtosis = serious outlier problem

EXPLORATORY DATA ANALYSIS: TOOLS

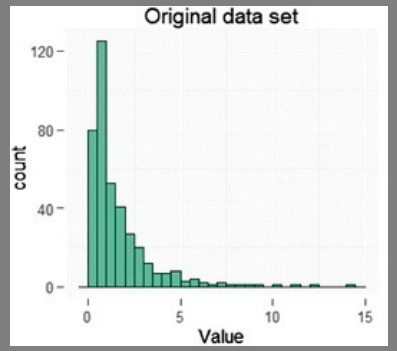
Descriptive statistics

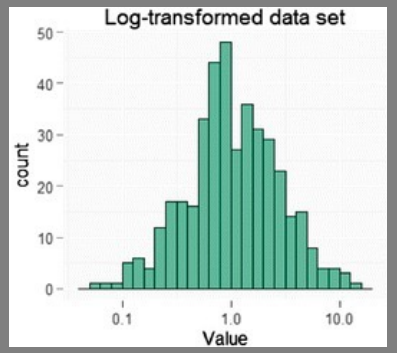
Country	Gender	Height	Basketball players
United Kingdom	Male	$1.95 \pm 7\%$	50% (1/2)
United Kingdom	Female	$1.8 \pm 0\%$	100% (1/1)
Canada	Male	NA	NA
Canada	Female	$1.8 \pm 14\%$	50% (1/2)
Germany	Male	$2.1 \pm 0\%$	100% (1/1)
Germany	Female	NA	NA

Visualisation

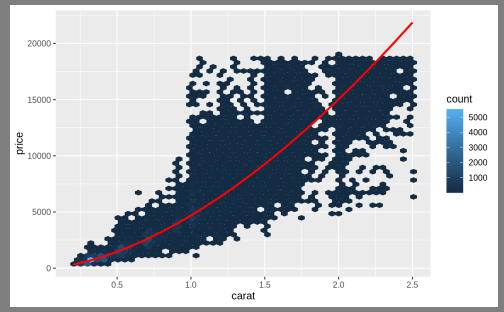


Transformation





Modelling



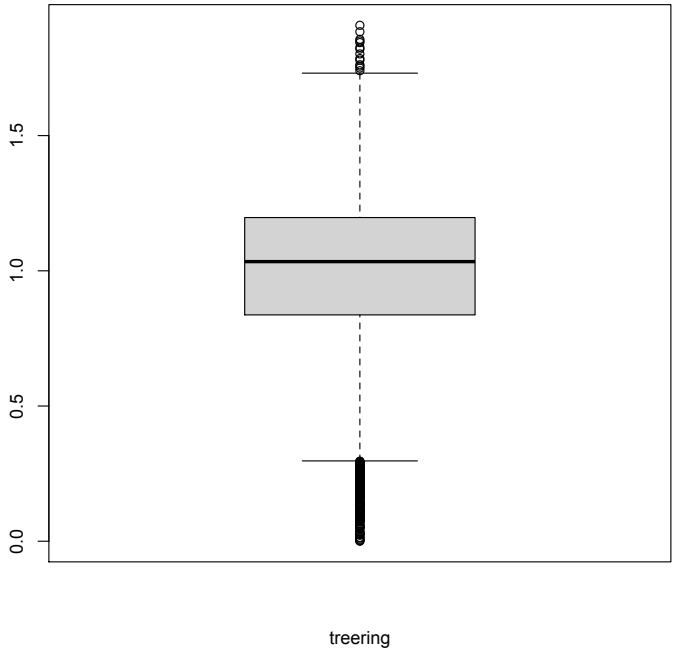
Figures: Grolemund and Wickham (2017) R for Data Science

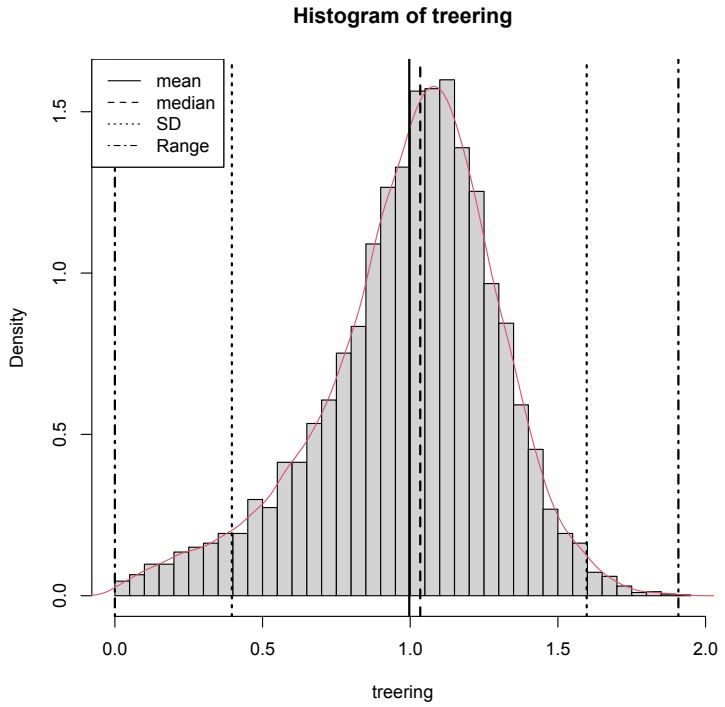
EDA – VISUALIZATION

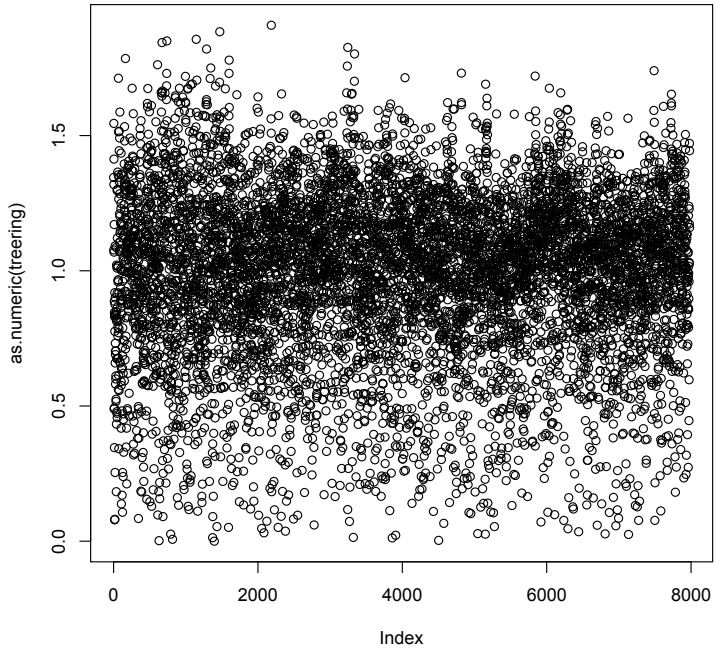
The Three main functions of graphical displays in data analysis:

- **Explore:** Checking data for unusual values, making sure the data meet the assumptions of the chosen analysis and occasionally deciding what analysis (or model) to use.
- **Analyse:** Checking assumptions and ensuring that the chosen model is a realistic fit to the data.
- **Present:** Summarizing numerical information and graphical displays for specific statistical purposes.

VISUALIZATION: BASIC TYPES





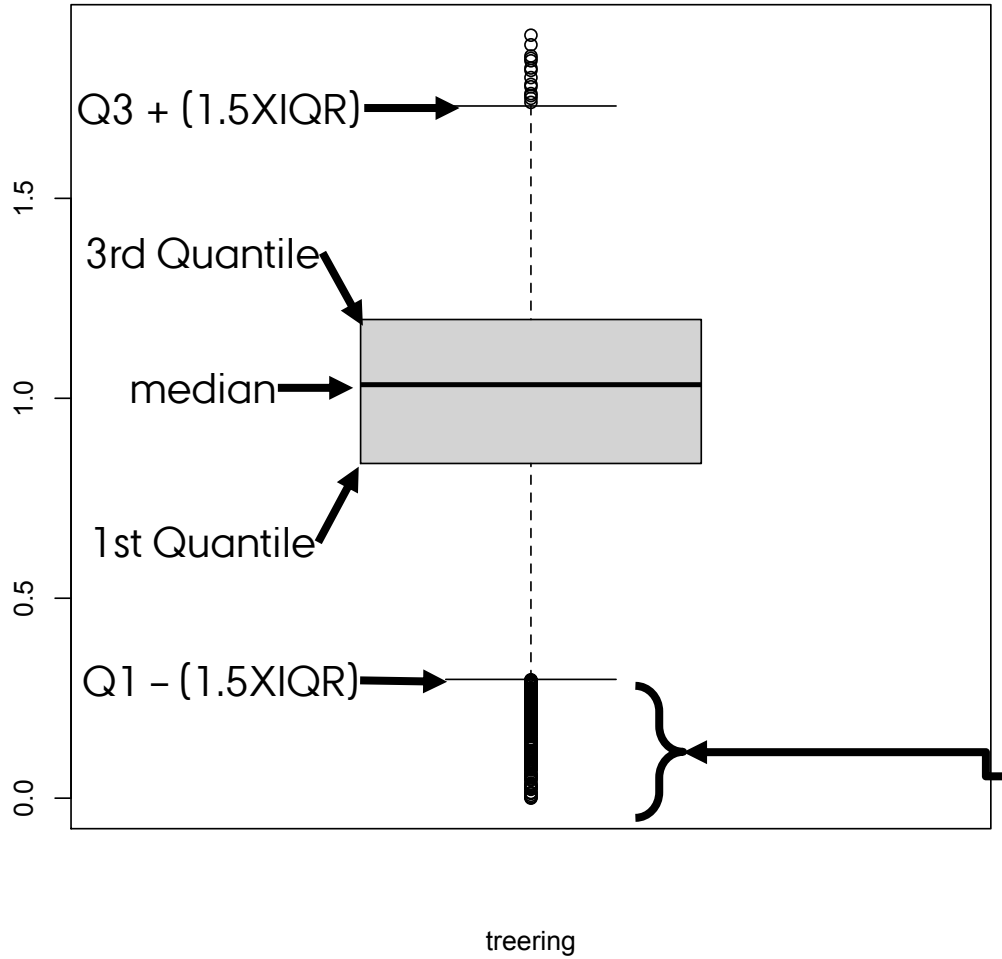


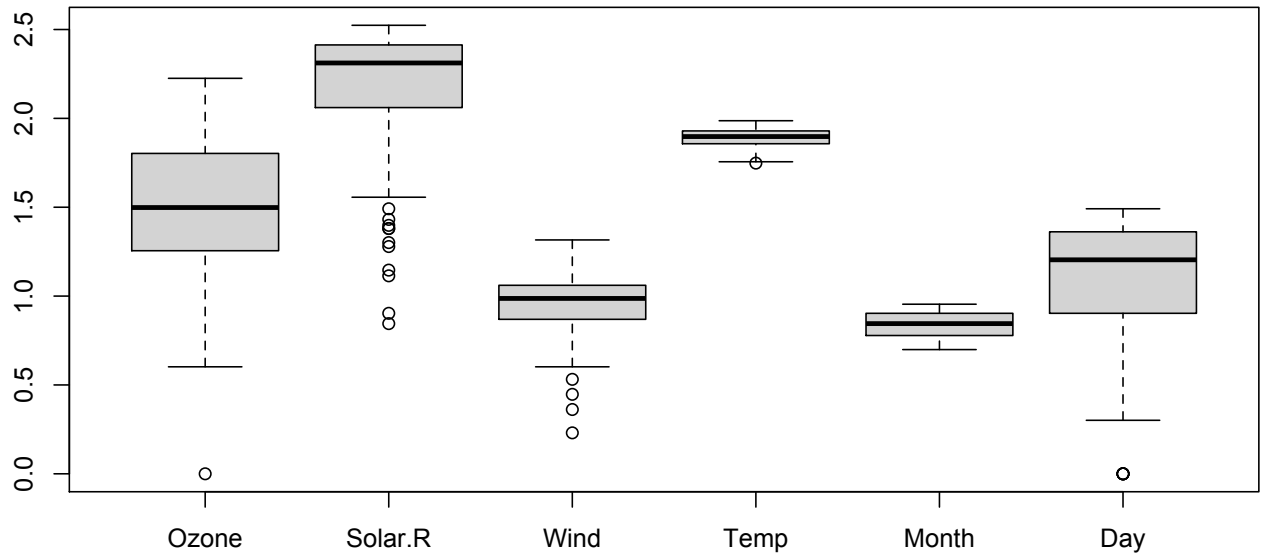
Box-plots
graphical representation of different descriptive statistics

Histograms
graphical representation of a frequency (or density) of the observations.

Dotplots
Each observation is represented by a single dot or symbol.

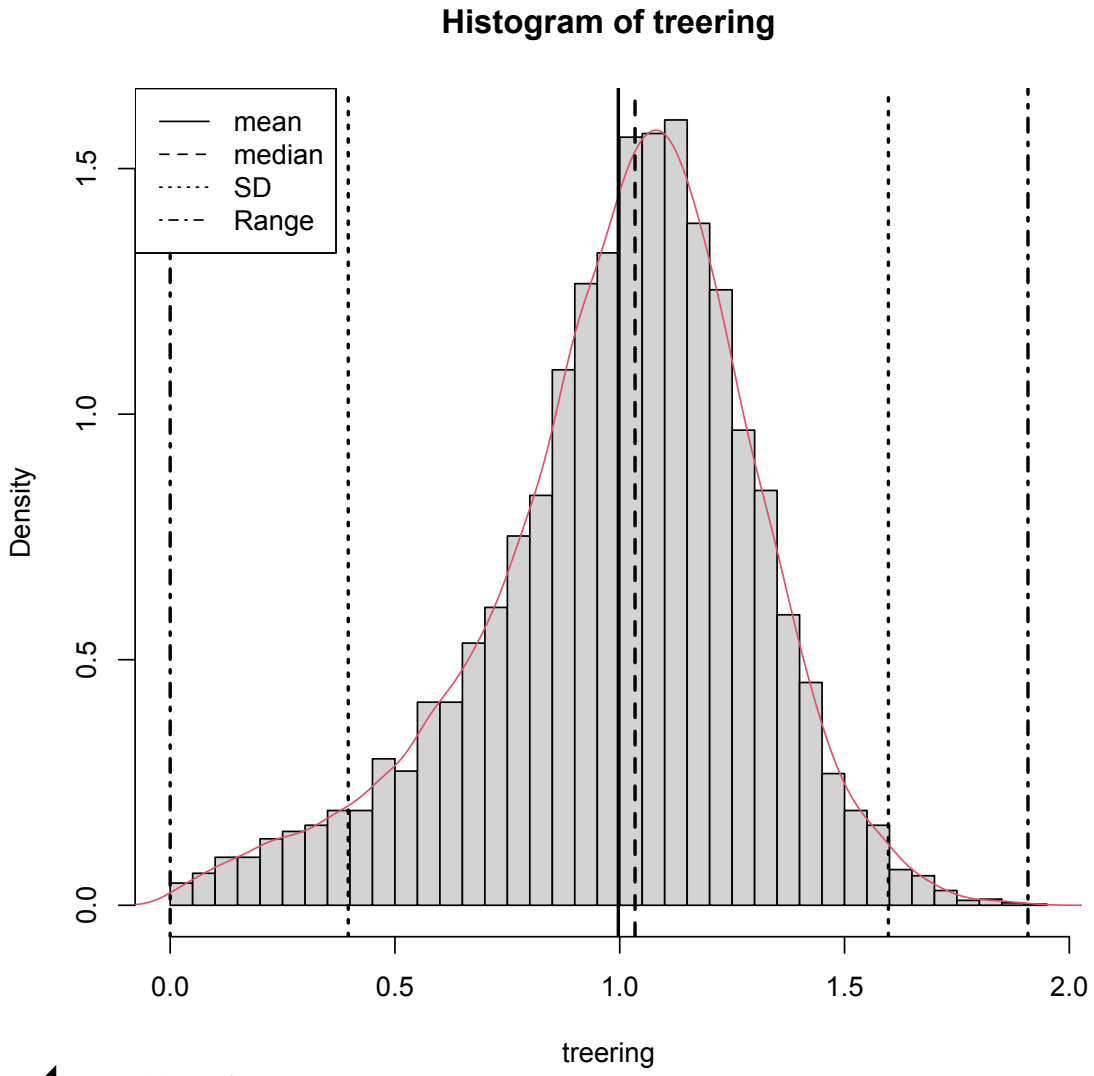
VISUALIZATION: BOX-PLOTS





How would identify and outlier using a Boxplot

VISUALIZATION: HISTOGRAM



Which is the shape of the distribution?

This is the most important point we want to know about our observations, and therefore about the population from which our data came

Why?: Most of "common" the statistical tests assume that the variables being analysed have normal distributions

R FUNDAMENTALS: NORMALITY TEST

Normality of a series of observations can be **tested** by using the Shapiro–Wilk test

`shapiro.test()`

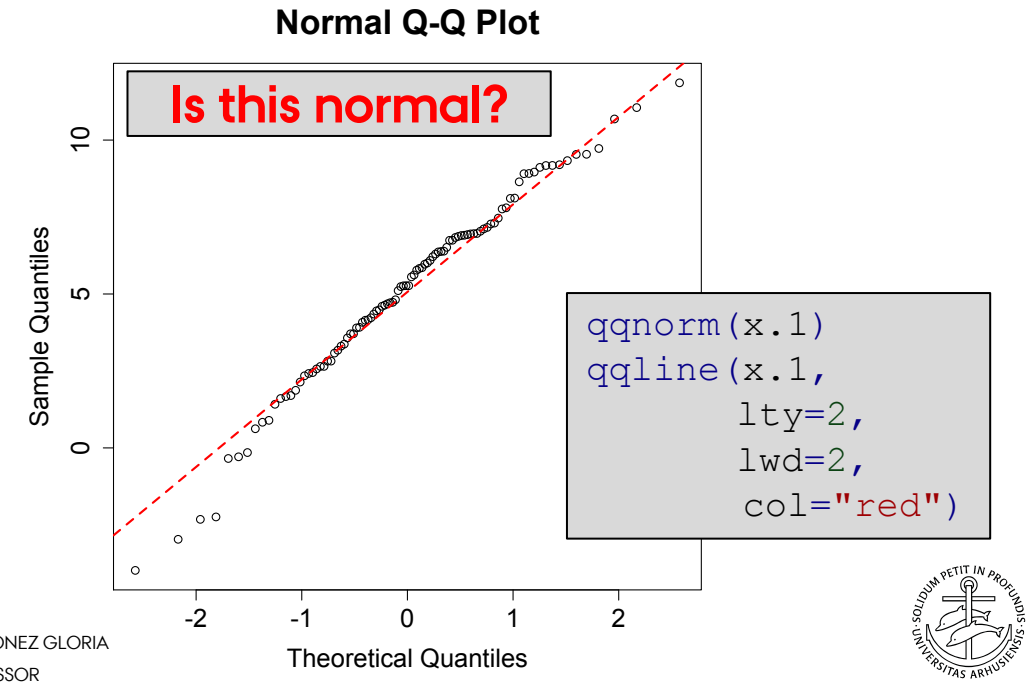
If the test is Non significant you have normality!

```
> # Normality Tests
> ### test using a Normal distribution
> shapiro.test(rnorm(100, mean = 5, sd = 3))
Shapiro-Wilk normality test
data:  rnorm(100, mean = 5, sd = 3)
W = 0.98122, p-value = 0.1654
> ### test using a Poisson distribution
> shapiro.test(rpois(1000, lambda=5))
Shapiro-Wilk normality test
data:  rpois(1000, lambda = 5)
W = 0.97227, p-value = 6.871e-13
```

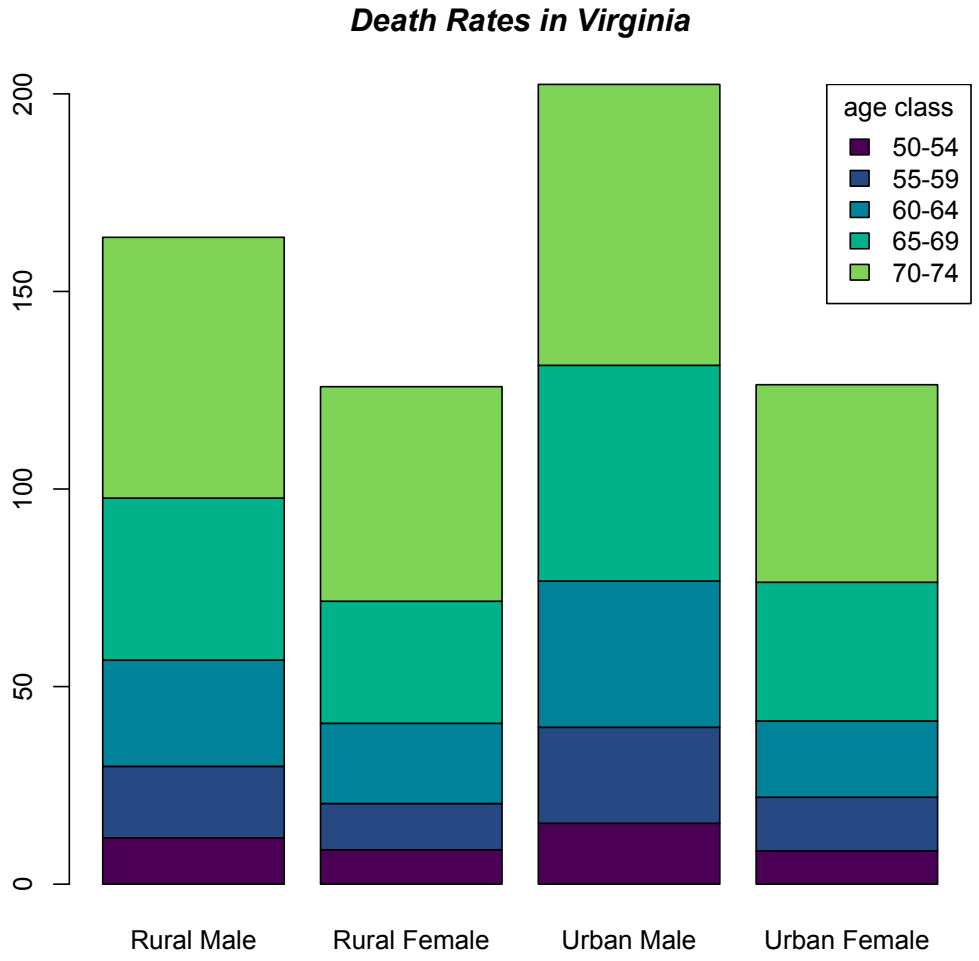
Normality of a series of observations can be **evaluated** by using quantile–quantile plots

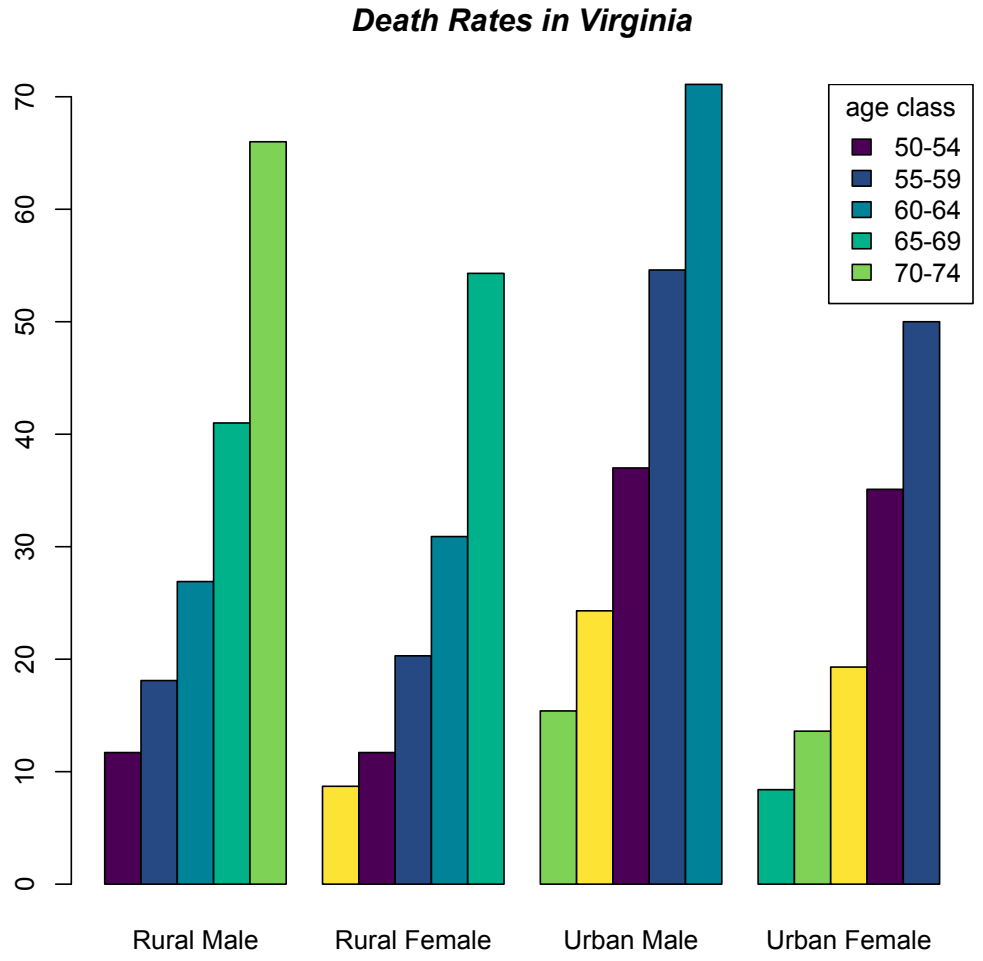
`qqnorm()` and `qqline()`

Departures from normality show up as S-shapes or banana shapes



VISUALIZATION: BAR PLOTS

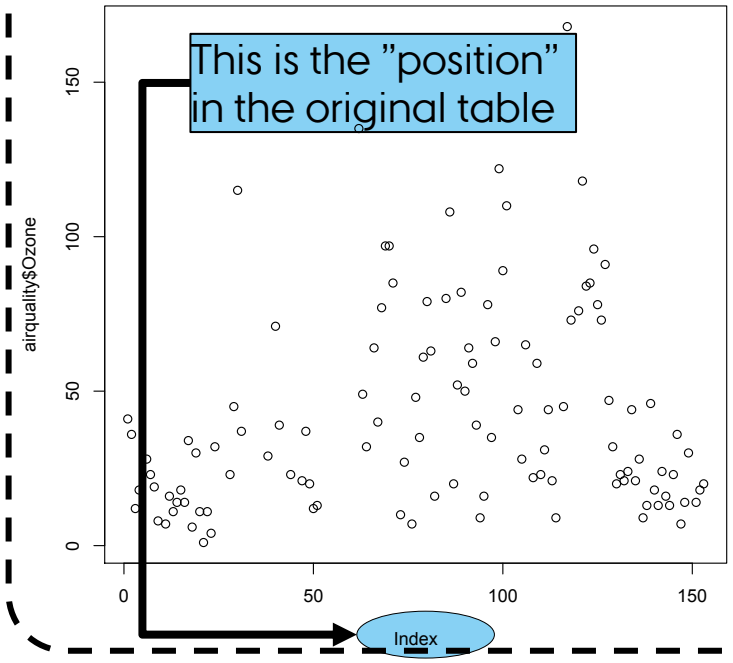




VISUALIZATION: DOTPLOT

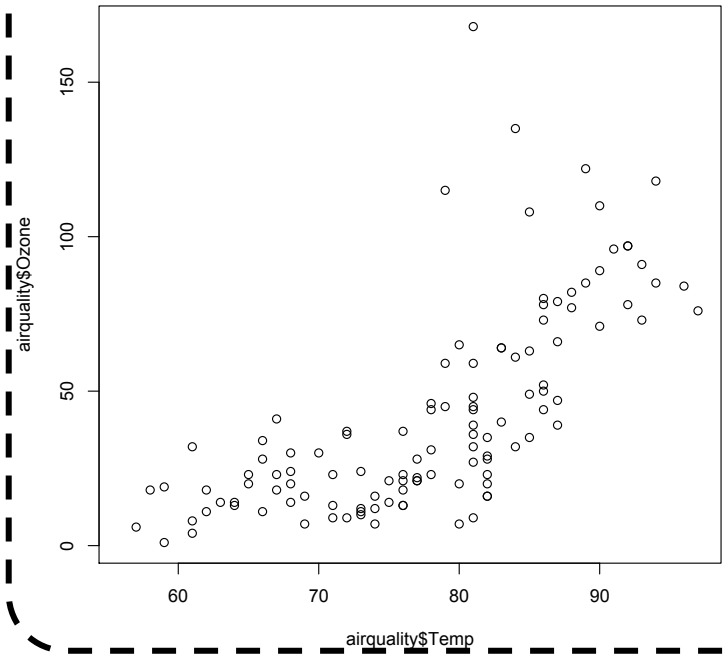
One variable

Dotplot



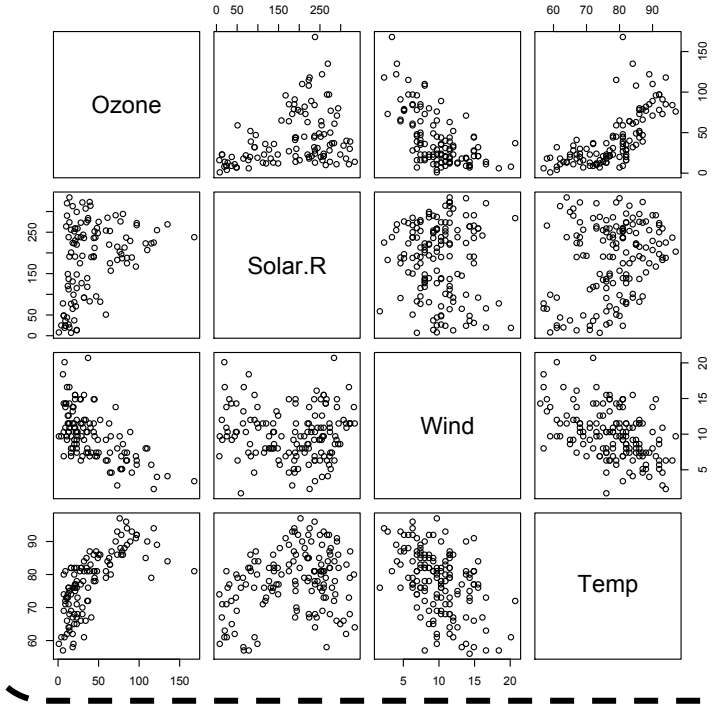
Two variables

Scatterplot



Multiple variables

Scatterplot

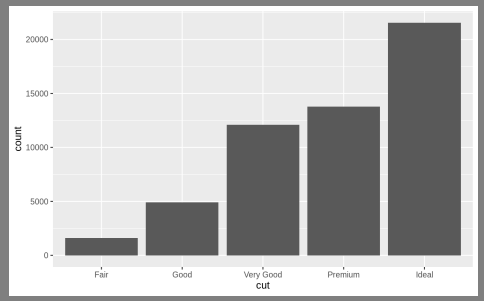


EXPLORATORY DATA ANALYSIS: TOOLS

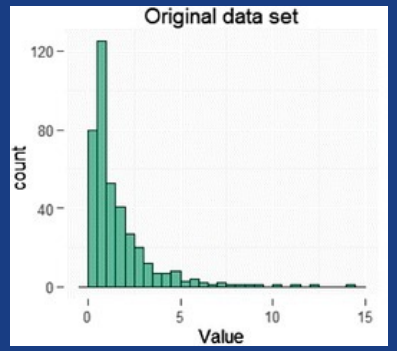
Descriptive statistics

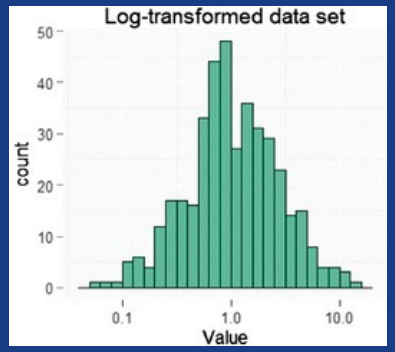
Country	Gender	Height	Basketball players
United Kingdom	Male	$1.95 \pm 7\%$	50% (1/2)
United Kingdom	Female	$1.8 \pm 0\%$	100% (1/1)
Canada	Male	NA	NA
Canada	Female	$1.8 \pm 14\%$	50% (1/2)
Germany	Male	$2.1 \pm 0\%$	100% (1/1)
Germany	Female	NA	NA

Visualisation

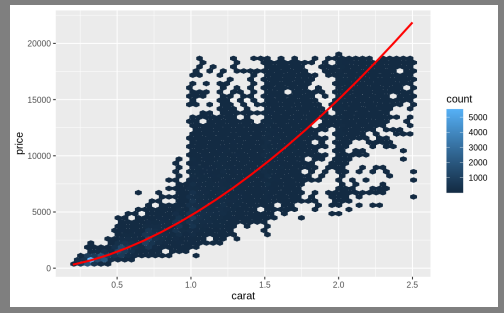


Transformation





Modelling



Figures: Grolemund and Wickham (2017) R for Data Science

EDA – TRANSFORMATIONS

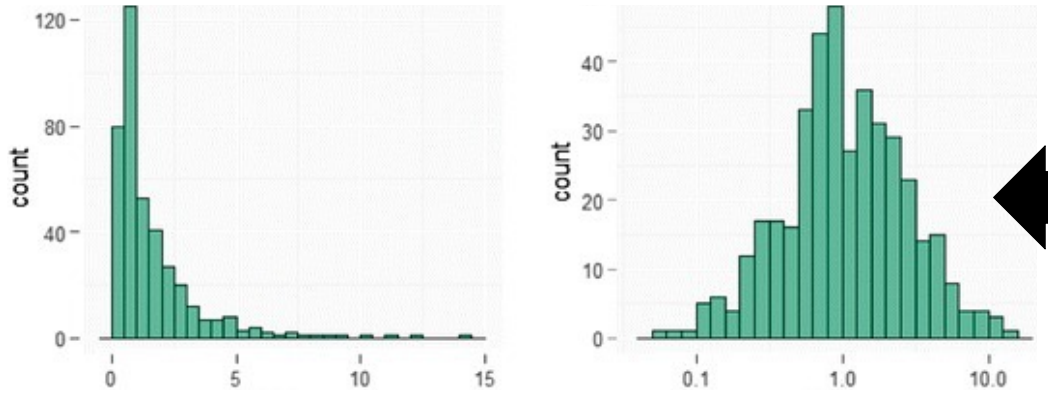
Data transformation: Converting data from one format or structure into another format or structure.

Aims of data transformations are:

1. Make the data and the model error terms closer to a normal distribution.
2. Reduce any relationship between the mean and the variance.
3. Reduce the influence of outliers, especially when they are at one end of a distribution.
4. Make effects that are multiplicative on the raw scale additive on a transformed scale.

EDA – TRANSFORMATIONS

Data transformation: Converting data from one format or structure into another format or structure.

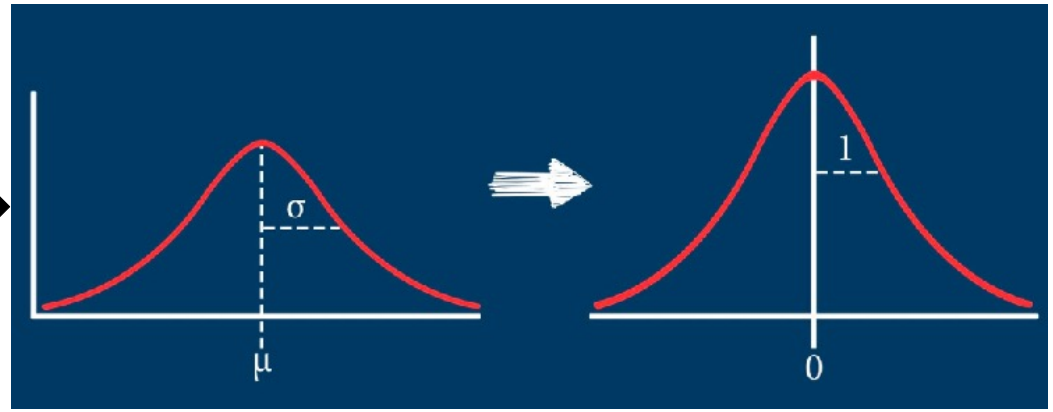


Normalizing

Make the distribution of the data symmetrical

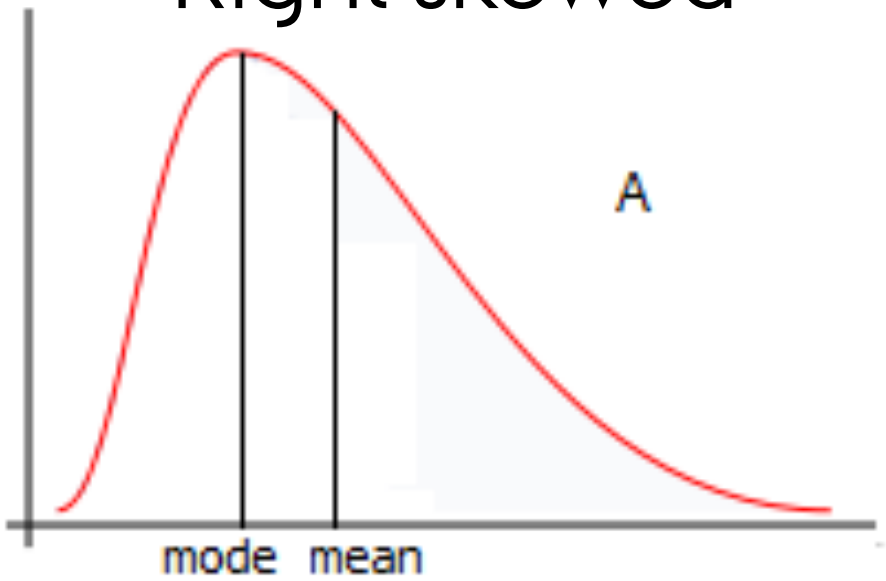
Standardising

Put all variable in units of variation [mean of 0 and a standard deviation of 1]

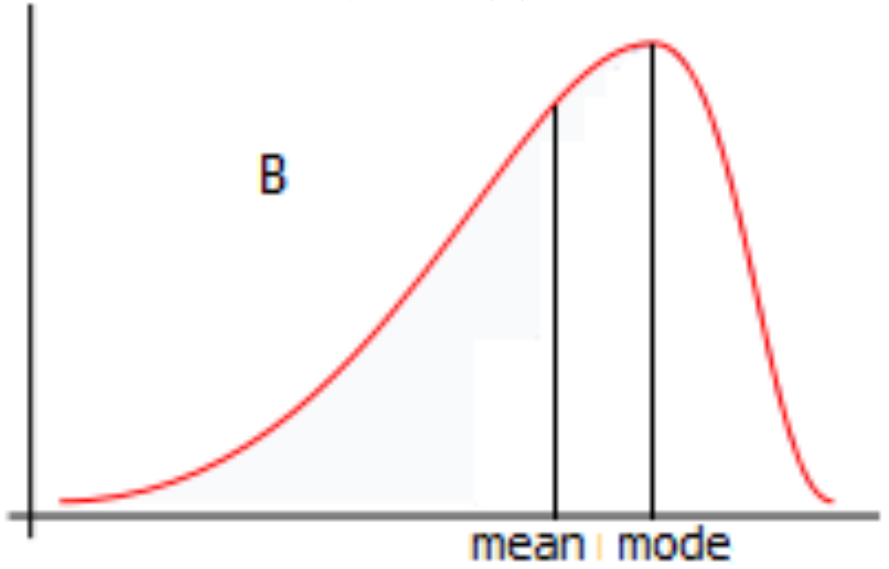


EDA – NORMALIZING

Right skewed



Left skewed



Normalizing these distributions is useful in statistical modelling because:

1. Makes data and errors **Normal**.
2. Reduces relations between Mean and variance.
3. Reduces the influence of outliers.
4. Makes multiplicative effects on a raw scale liner on the transformed scale.

EDA – NORMALIZING

Power [Y^p]: Useful for long tails data

- The Root transformation [$Y^{1/p}$]: Useful for counts data **Which is not normal!!**

Log [$log\ (Y+c)$]: make positively skewed (hump on the left side) distributions more symmetrical

Logit [$Log\left(\frac{p}{1-p}\right)$]: Use to make linear portion or probability data **Which is not normal!!**

- **Arcsine** [$sin^{-1}(Y^{0.5})$]

The Box–Cox family of transformations

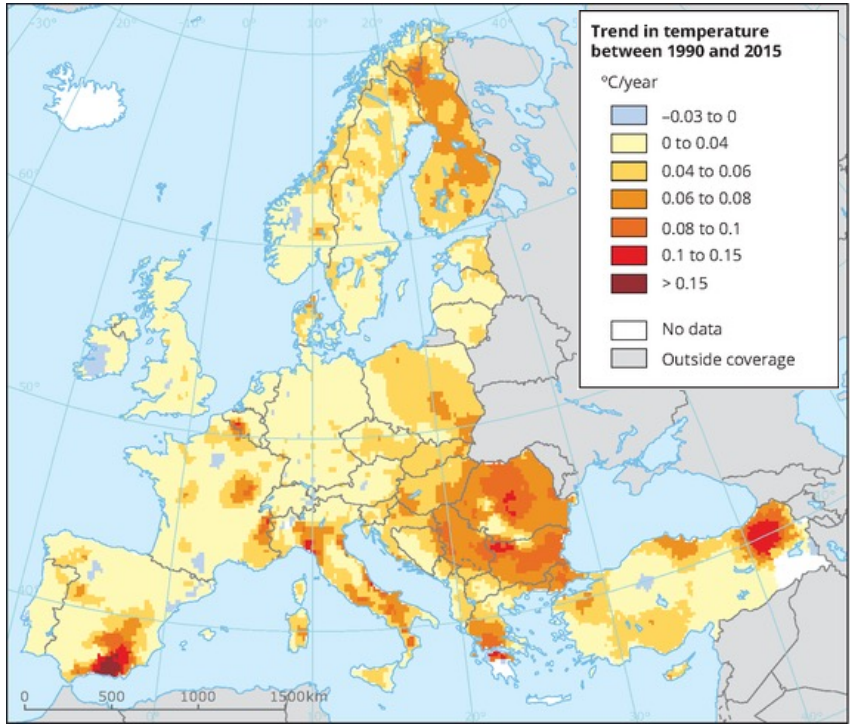
Can also be used to find the best transformation

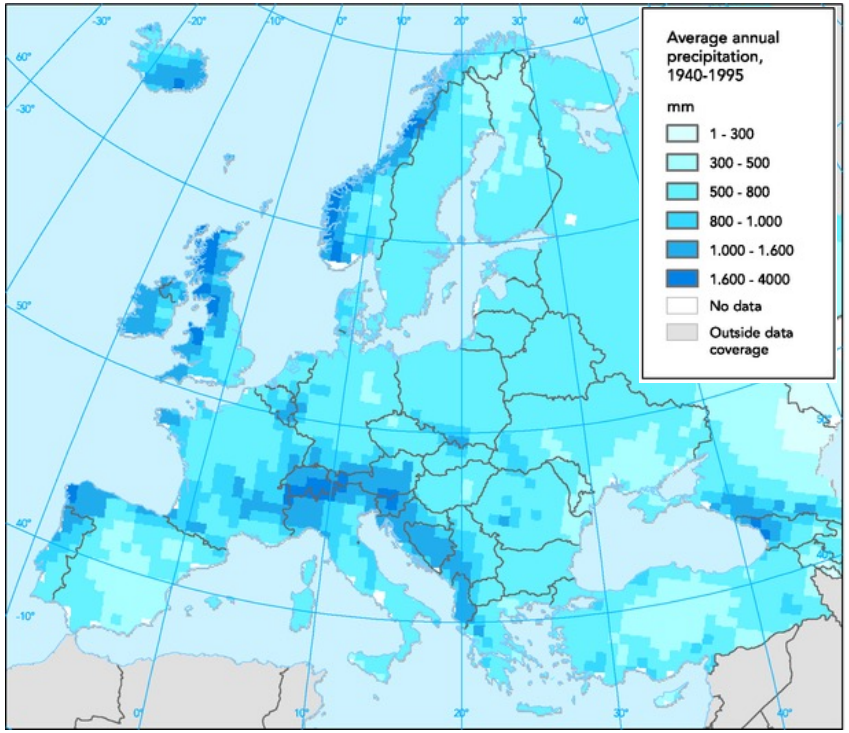
$\frac{Y^{\lambda}-1}{\lambda}$ when $\lambda \neq 0$

$Log(Y)$ when $\lambda = 0$

The goal is achieving normality

EDA – STANDARDISATIONS





Stabilization allows contrasting the effects of variables measured in different units in modelling frameworks

EDA – STANDARDISATIONS

Different standardisations are possible:

Centring: Making the mean Zero

$$y_i = y_i - \bar{y}$$

Ranging: Marking the observations values to range between zero (minimum) to one (maximum)

$$y_i = \frac{y_i}{y_{max}} \text{ or } y_i = \frac{y_i - y_{min}}{y_{max} - y_{min}}$$

Scaling: Changing a variable so it has a mean of zero and a standard deviation (and variance) of one

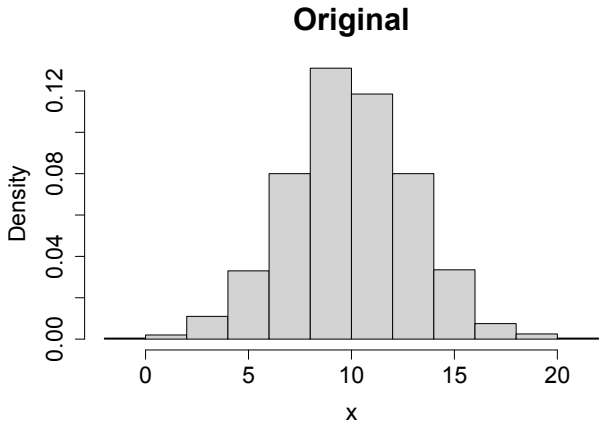
$$y_i = \frac{y_i - \bar{y}}{s}$$

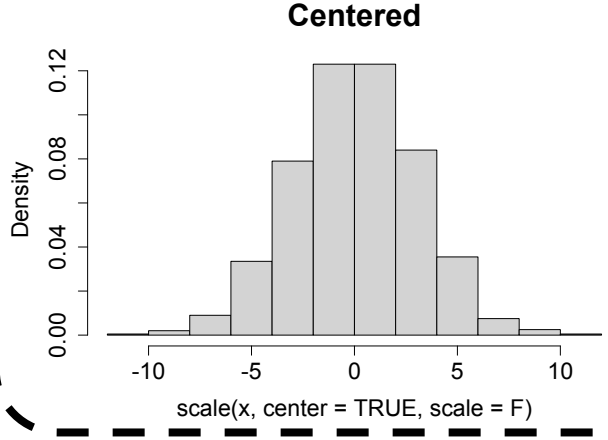
Centring and Ranging
useful as standardisations of abundance data before multivariate analyses

Scaling
useful to convert all the variables to a similar scale

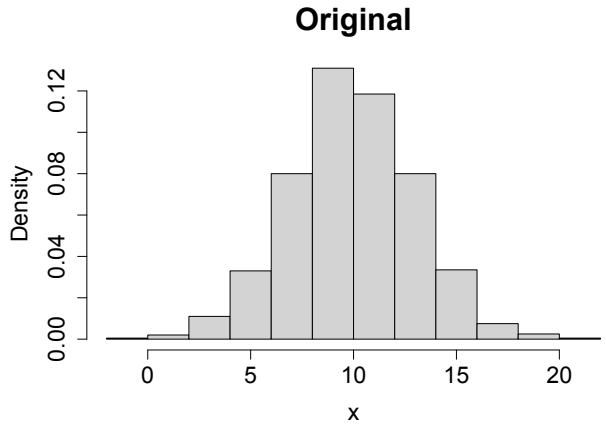
EDA – STANDARDISATIONS

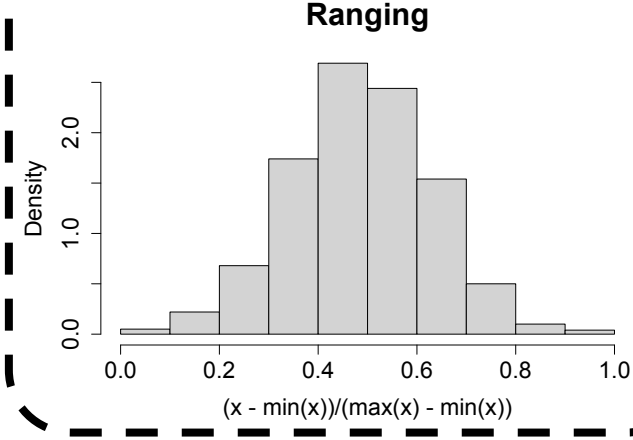
Centring



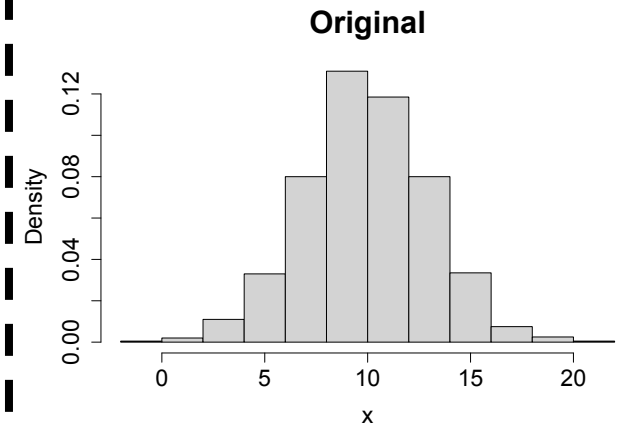


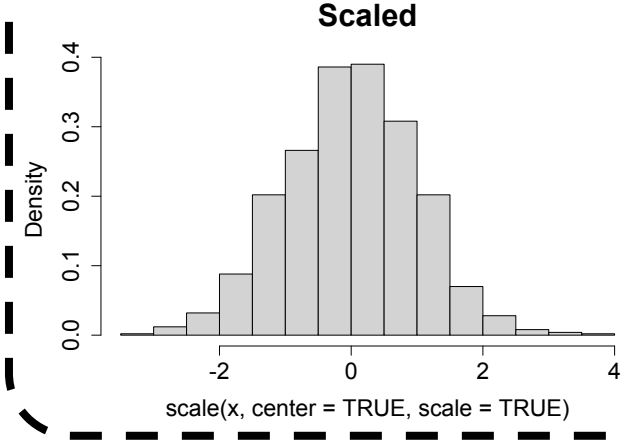
Ranging





Scaling



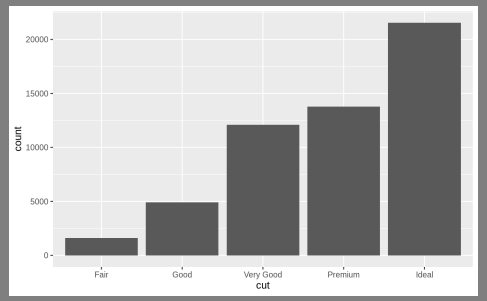


EXPLORATORY DATA ANALYSIS: TOOLS

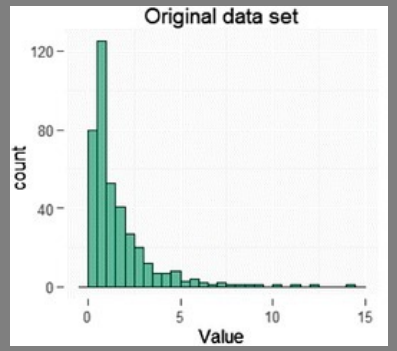
Descriptive statistics

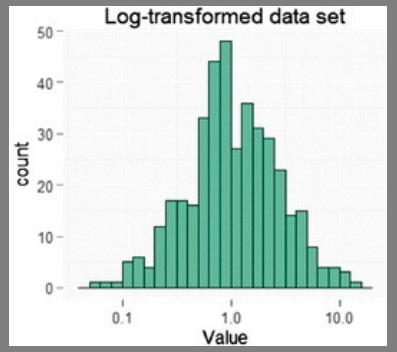
Country	Gender	Height	Basketball players
United Kingdom	Male	$1.95 \pm 7\%$	50% (1/2)
United Kingdom	Female	$1.8 \pm 0\%$	100% (1/1)
Canada	Male	NA	NA
Canada	Female	$1.8 \pm 14\%$	50% (1/2)
Germany	Male	$2.1 \pm 0\%$	100% (1/1)
Germany	Female	NA	NA

Visualisation

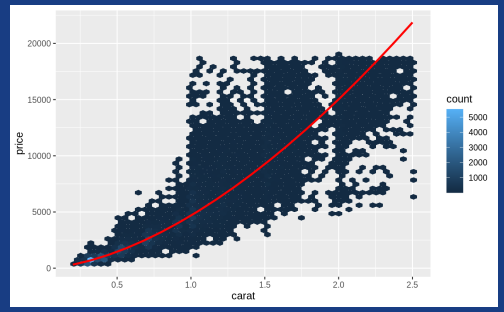


Transformation





Modelling



Figures: Grolemund and Wickham (2017) R for Data Science

EDA – MODELLING

Modelling: Converting data from one format or structure into another format or structure.

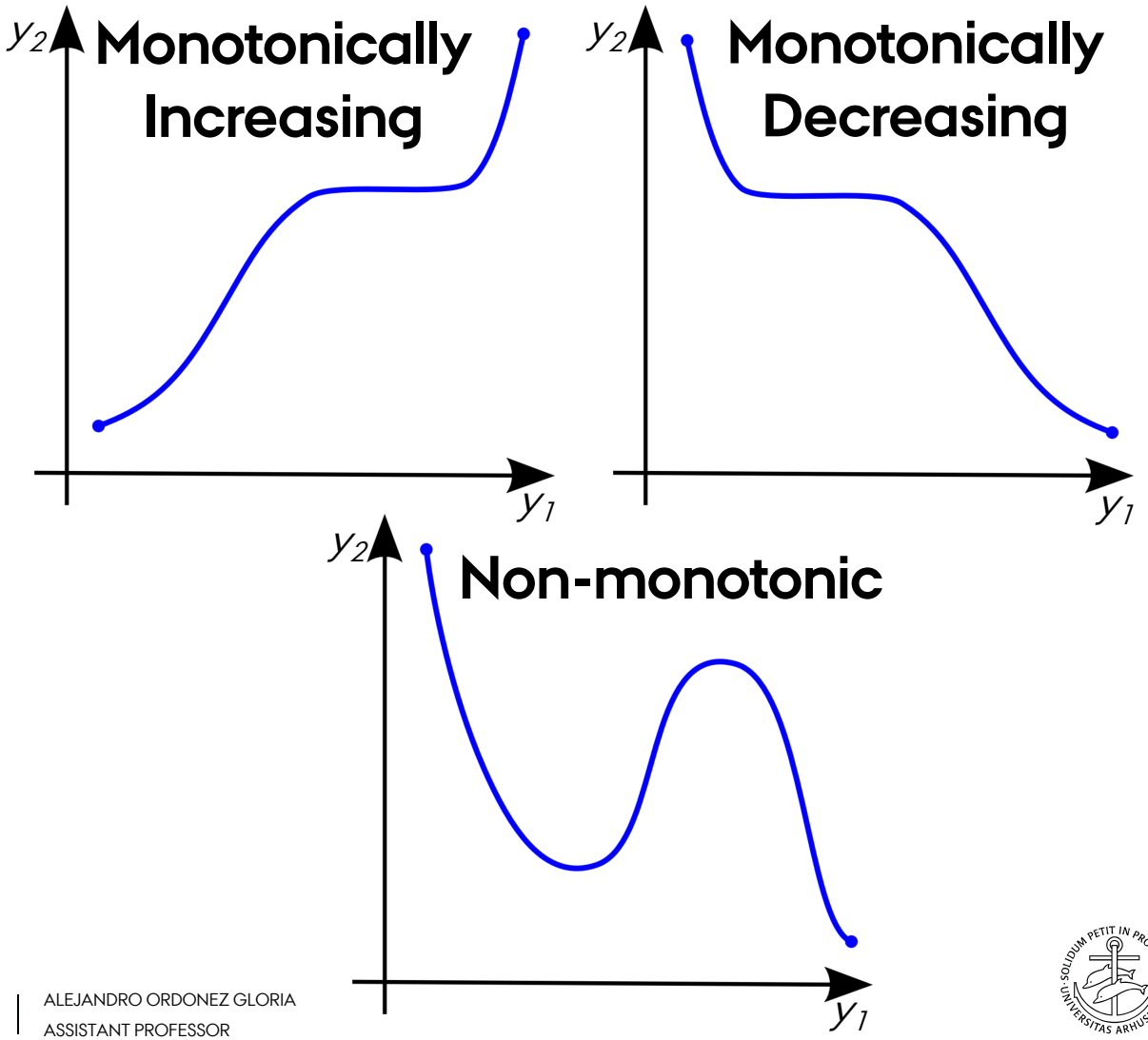
Aims of Modelling before doing any analyses are:

1. Assess regressions assumptions (**more on this next week**).
2. Determine (in)dependence between variables → correlations
3. Establish additive effects if variables are dependent → linearity

MODELLING – CORRELATIONS

Correlation coefficients is a statistical measure which determines the **'strength'** of the **co-relationship** or **association** of two continuous variables (Y_1 and Y_2).

- Correlation coefficients assume **monotonic** relations → always changes in the same direction
- Correlation coefficients bounded between -1 and 1



MODELLING – CORRELATION

Parametric indices

Pearson correlation (r)

$$r_{Y_1Y_2} = \frac{\sum_{i=1}^{n}[(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)]}{\sqrt{\sum_{i=1}^{n}(y_{i1} - \bar{y}_1)^2 \sum_{i=1}^{n}(y_{i2} - \bar{y}_2)^2}}$$

Other assumptions: normality of variables and linearity

Correlation **DOES NOT** mean directionality or causality.

Non-Parametric indices

Spearman's rank correlation (r_S)

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}.$$

n is the number of pairs of measurements

d_i is the difference of the i^{th} pair of rankings

Kendall's rank correlation(τ_A)

$$\tau_A = \frac{n_c - n_d}{n_0}$$

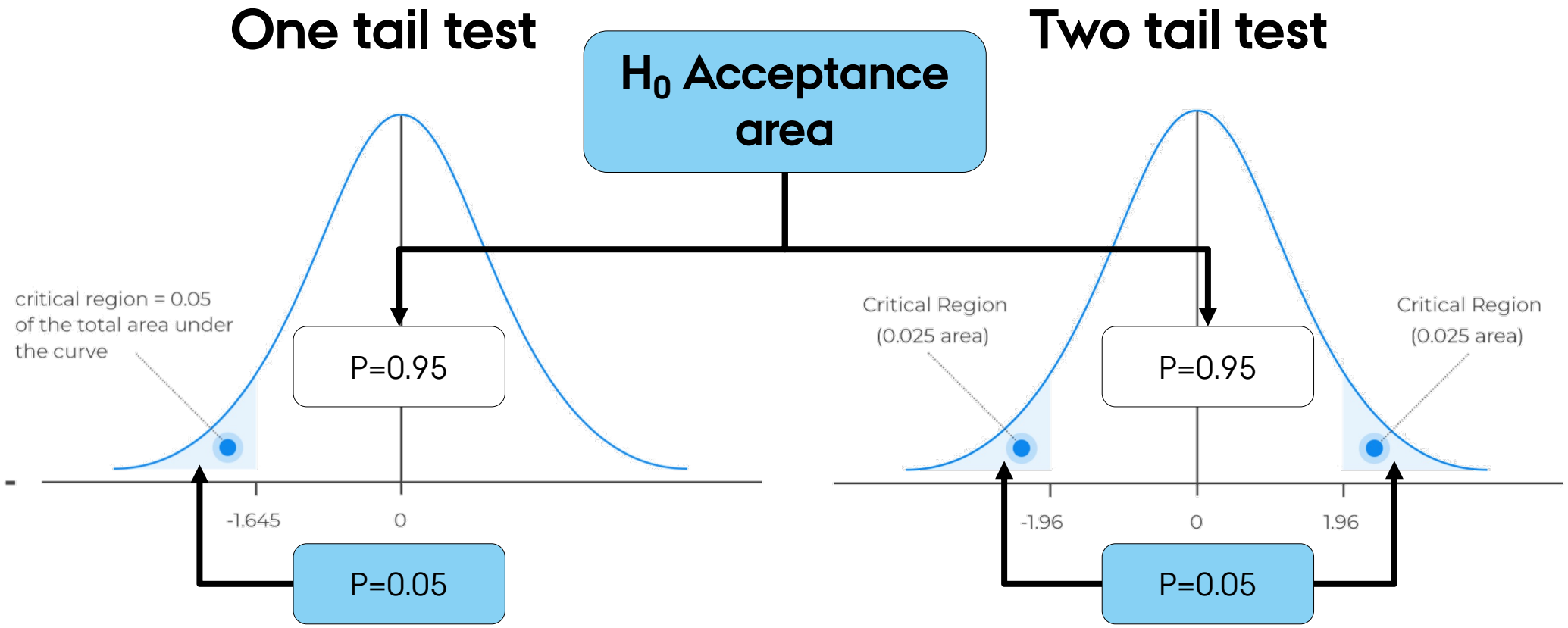
$n_0 = n(n-1)/2$

$n_c =$ Number of concordant pairs

$n_d =$ Number of discordant pairs

MODELLING – SIGNIFICANCE

Statistically significant means that if the **Null Hypothesis** (H_0) is false.

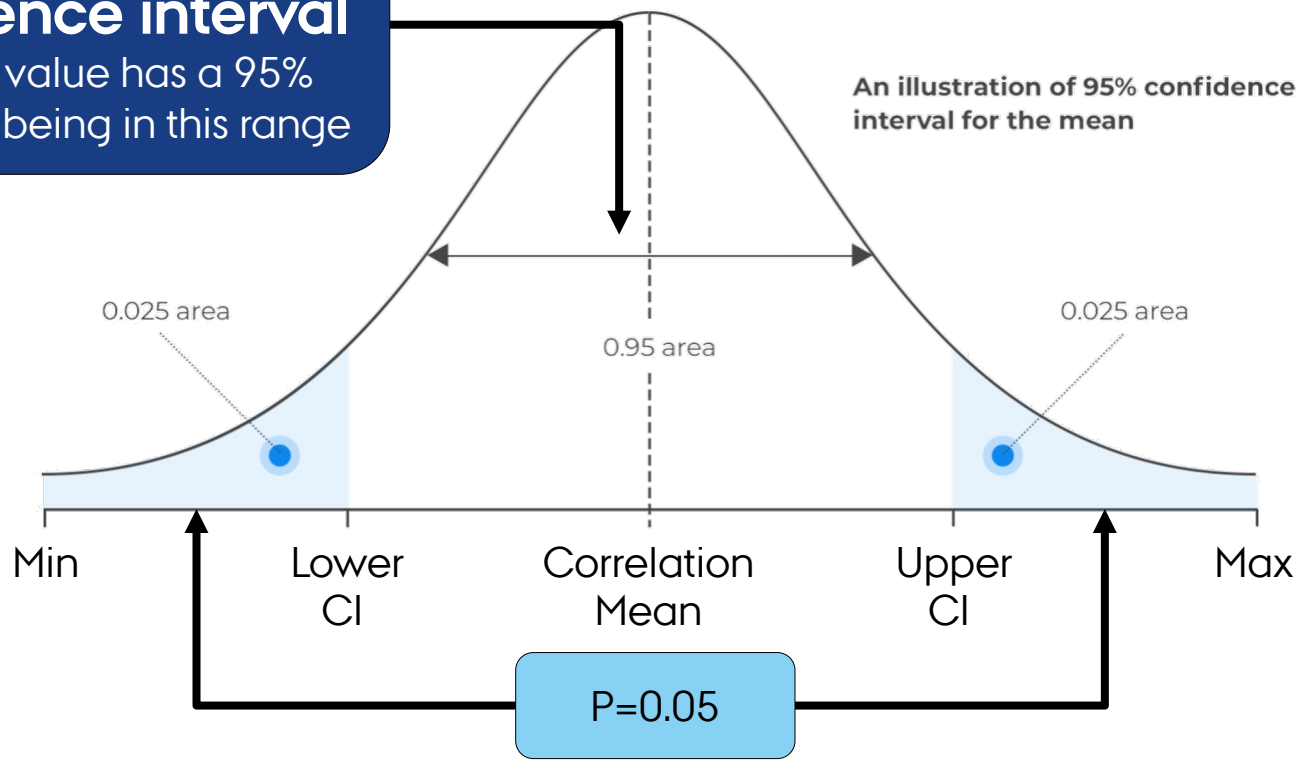


MODELLING – SIGNIFICANCE & CORRELATION

Estimate 95% confidence interval
The **true** value has a 95% change of being in this range

H_0 = Correlation estimate 95% CI *includes 0*

Two tail test



Confidence Interval Formula

$$CI = \mu \pm t_{\frac{\alpha}{2}} \times \frac{S}{\sqrt{n}}$$

$\frac{S}{\sqrt{n}}$ =Standard deviation

$t_{\frac{\alpha}{2}}$ = *T*-score

μ =Mean or estimate

So far so good?

Any questions?

Ready to finish?

SUMMARY

Exploratory Data Analysis (**EDA**) is the art of looking at your data

EDA is also the base for statistical analysis of multidimensional data

Before you stat you need to import and clean your data

The two crucial questions you will **always** ask:

What type of variation occurs within my variables?

What type of covariation occurs between my variables?

SUMMARY

EDA uses **four** main tools:

- **Descriptive statistics**: Quantitative descriptions or summaries of the features from a collection of information.
- **Visualisation**: The graphical representation of information and data to see and understand trends, outliers, and patterns.
- **Transformation**: process of changing the format, structure, or values of data
- **Modelling**: simple low-dimensional summaries of a dataset, and trends assessments.

AARHUS
UNIVERSITY