HOST
Hochschule Stralsund

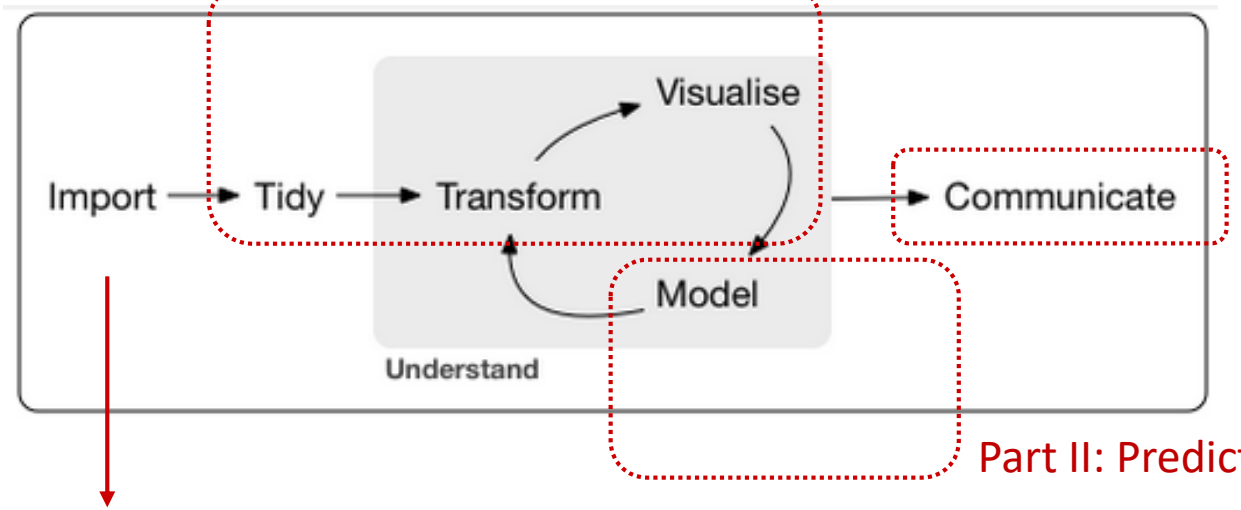
Data Science

Data Wrangling & Graphics

Gero Szepannek

Data Science Process

Part I: Data Engineering & Exploration

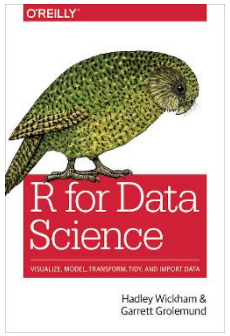


Final project exercise

Part II: Predictive Modelling

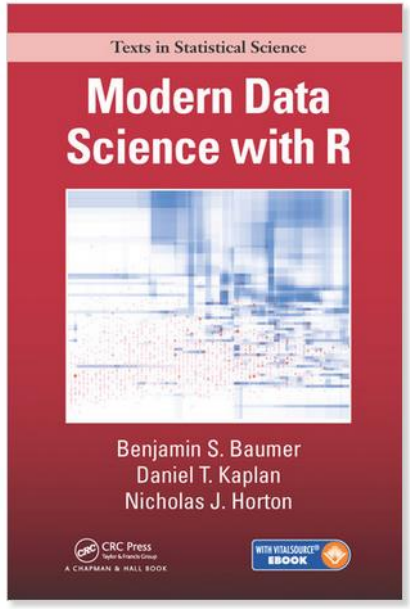
Today…

Iterative process!



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

Further Reading



https://mdsr-book.github.io/mdsr2e/

Traditional

vs.

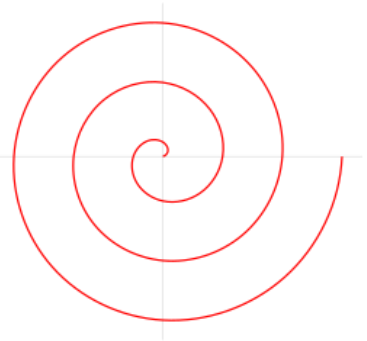
R packages for data science

The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

stackoverflow



Importing Data

read.table() → ASCII / .csv
load() → R objects

```
x <- read.table("C:/your/path/titanic.csv", sep = ";", header = TRUE)
load("C:/your/path/titanic.Robj")
```

Further useful Packages

- readr for reading .csv and fwf files.
- readxl / openxlsx for reading .xls and .xlsx files.
- haven for SAS, SPSS, and Stata files.
- httr for talking to web APIs.
- rvest for scraping websites.
- xml2 for importing XML files.

A first Glance at the Data…

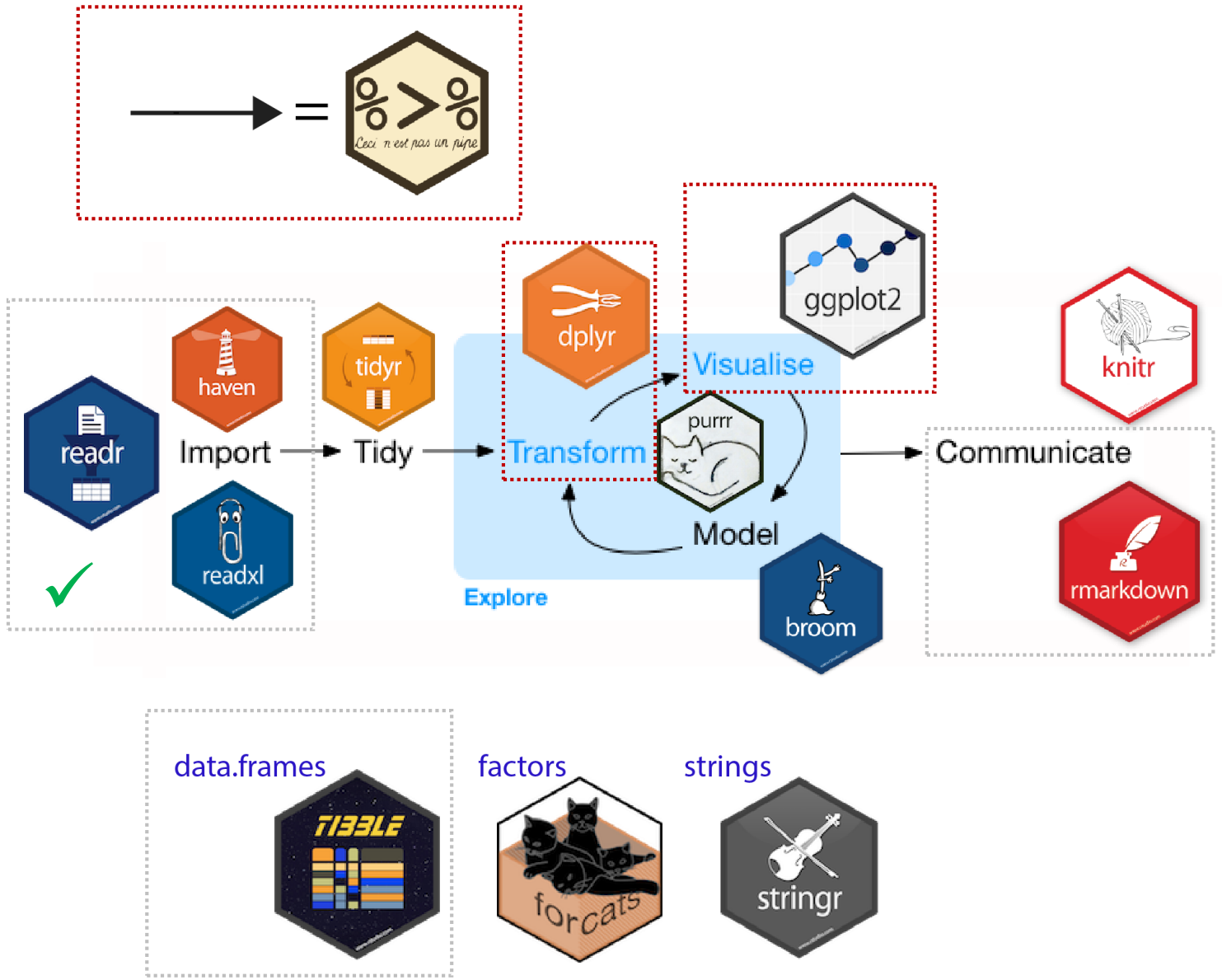
```
head(x)
str(x)
x$gender <- as.factor(x$gender)
x$survived <- as.factor(x$survived)
str(x)
```

```
> str(x)
'data.frame':   2207 obs. of  9 variables:
 $ gender  : Factor w/ 2 levels "female","male": 2 2 2 1 1 2 2 1 2 2 ...
 $ age     : num  42 13 16 39 16 25 30 28 27 20 ...
 $ class   : chr  "3rd" "3rd" "3rd" "3rd" ...
 $ embarked: chr  "Southampton" "Southampton" "Southampton" "Southampton" ...
 $ country : chr  "United States" "United States" "United States" "England" ...
 $ fare    : num  7.11 20.05 20.05 20.05 7.13 ...
 $ sibsp   : int  0 0 1 1 0 0 1 1 0 0 ...
 $ parch   : int  0 2 1 1 0 0 0 0 0 0 ...
 $ survived: Factor w/ 2 levels "no","yes": 1 1 1 2 2 2 1 2 2 2 ...
```

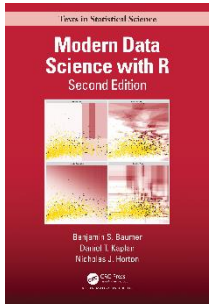
What is the difference between characters (`chr`) and factors (`Factor`)?

Exercise

1. Read the data custdata.csv into R!
2. How many observations have the data?
3. What is it about?


Ceci n'est pas un pipe
ggplot2
dplyr
knitr
haven
tidyr
Visualise
readr
purrr
Import
Tidy
Transform
Communicate
Model
readxl
Explore
broom
rmarkdown
data.frames
factors
strings
tibble
forcats
stringr

A Grammar for Data Wrangling

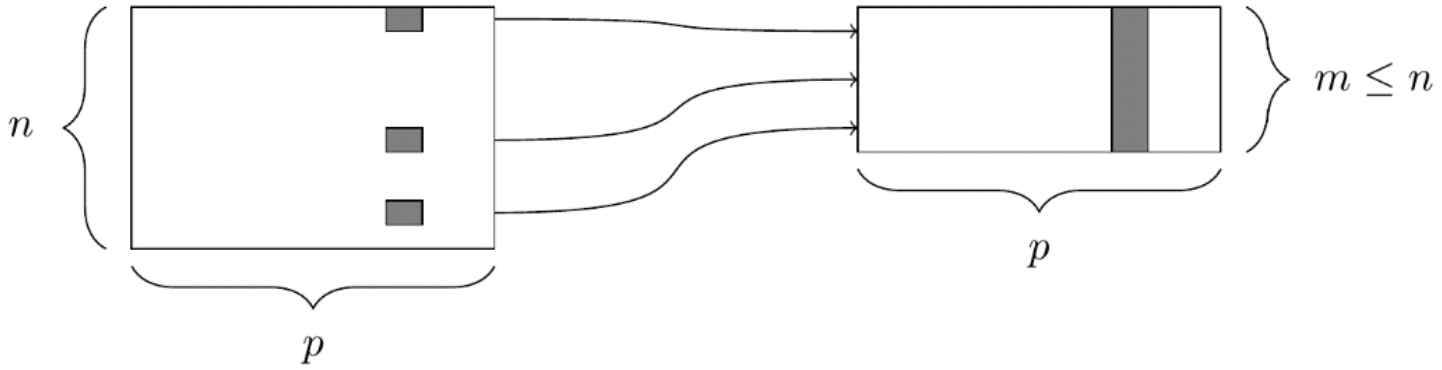


The verbs:

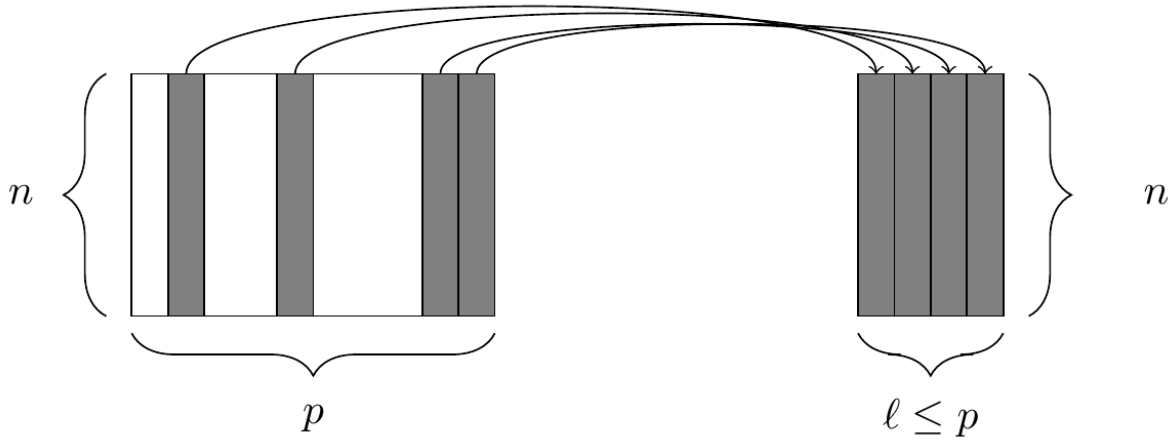
1. `select()` …a subset of columns (variables).
2. `filter()` …a subset of rows (observations).
3. `mutate()` Add or modify existing variables.
4. `arrange()` Sort the observations.
5. `summarize()` Aggregate the data accross observations according to some criteria.

Select and Filter

`filter()`



`select()`



Exercise

1. Import the data from the file presential.Robj.
2. **What is it about?**

3. Run:
```
head(presidential)
str(presidential)
```

4. **…What is different compared to the data import from custdata?**

5. Run:
```
# install.packages("dplyr")
library(dplyr)
select(presidential, name, party)
filter(presidential, party == "Republican")
```

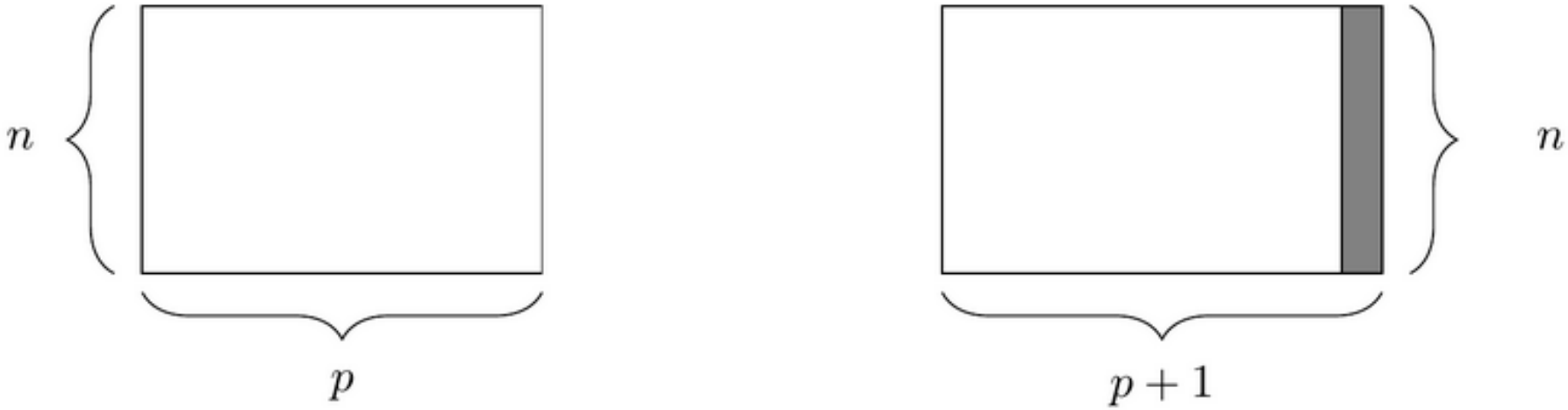
6. **What is the result? How many rows has presidential after running the code?**

The Pipe Operator


```r
filter(presidential, party == "Republican")
# ... is the same as:

# install.packages("magrittr")
library(magrittr)
presidential %>% filter(party == "Republican")
```

Mutate and Rename

`mutate()`



Exercise

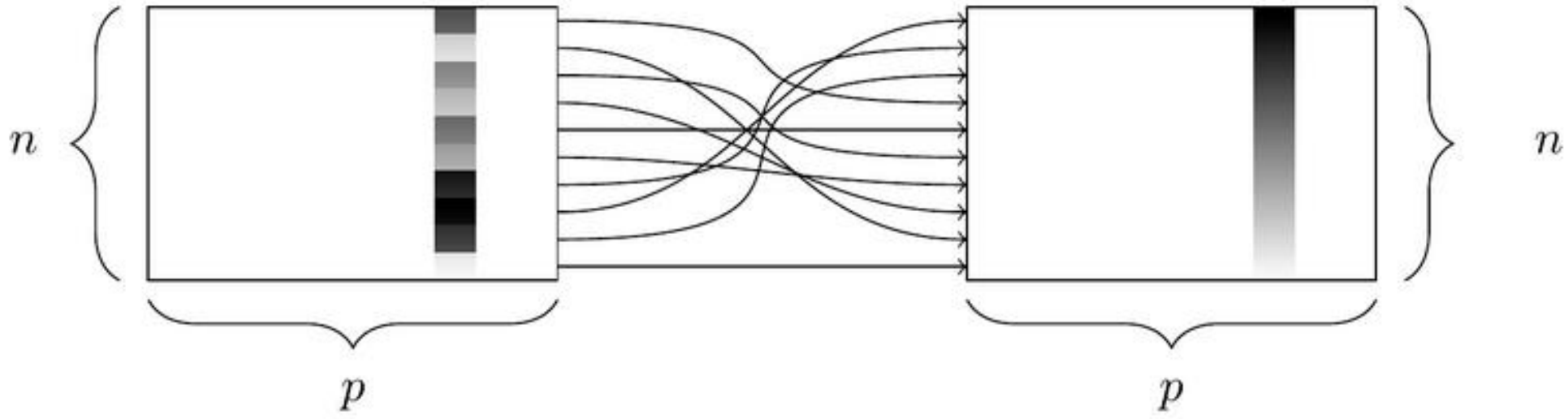
1. Run the following code:


```
# install.packages("lubridate")
library(lubridate) # for operating with time formats

howlong <- presidential %>% mutate(term.length = interval(start, end) / dyears(1))
```

2. **What is computed, here?**
3. Use the function `rename()` to rename the variable `term.length` into `duration`!

Arrange

`arrange()`



Exercise

1. Run the following code:

```
howlong
sort(howlong$duration)
order(howlong$duration)
```

2. **What does the function `order()` return?**
3. Use the function `order()` to re-arrange the data set howlong according to the duration!
4. …Note: You can add, `decreasing = TRUE` to invert the sorting order.

Exercise

1. Run the following code:

```
howlong
sort(howlong$duration)
order(howlong$duration)
```

2. **What does the function `order()` return?**
3. Use the function `order()` to re-arrange the data set howlong according to the duration!
4. …Note: You can add, `decreasing = TRUE` to invert the sorting order.

```
# ...using dplyr:
arrange(howlong, desc(duration), party, name)
```

Summarize and Group_by

Which party has been elected more often / longer?

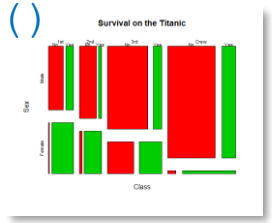
```
howlong %>% group_by(party)

howlong %>% group_by(party) %>% summarize(times_elected = n(),
                                          total_years = sum(duration),
                                          first_time = min(year(start)),
                                          average_duration = mean(duration)
                                          )
```

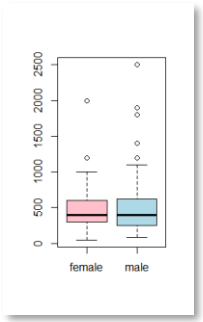
```
# ...in base R
aggregate(howlong$duration, by = list(howlong$party), FUN = sum)
```

A Grammar for Exploratory Data Analysis

factor	numeric	numeric	**Data Type:**
factor	factor	numeric	Variable 1 Variable 2
Contingency tables Barplots Mosaic plots	Means (& variances) by category Boxplots	Correlation ρ Scatterplot	**Method**

`table()`
`prop.table()`
`barplot()`
`mosaicplot()`
	`aggregate()`
	`boxplot(y~x)`
		`cor()`
		`plot()`	





The Power of Boxplots

1. Import the data `Default.Robj` and run the following code:

```
str(Default)
head(Default)

boxplot(income ~ default, data = Default)
boxplot(balance ~ default, data = Default)
```

2. Which of the two variables income or balance is more suitable to predict creditworthiness? …Try to explain, why!

Understanding Mosaicplots

1. Load the titanic data and run the following code:

```
head(titanic)

           table(titanic$gender, titanic$survived)
prop.table(table(titanic$gender, titanic$survived), 1)
mosaicplot(table(titanic$gender, titanic$survived), col = c("red", "green"))
```

2. **What does the function `prop.table()` return?**
3. **What can you see in the plot? Is there a dependency between survival and gender?**
4. Modify the code to analyze the dependency between survival and the class!

Numeric Data: Scatterplot & Correlation

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2 \cdot \sum_{j=1}^{n}(y_j - \bar{y})^2}}$$

numeric

numeric

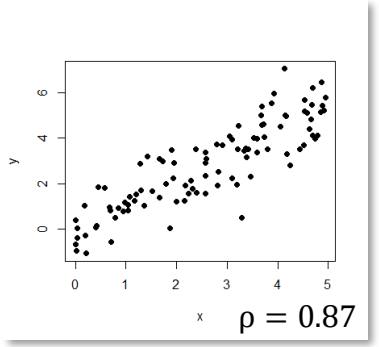
$-1 \leq \rho \leq 1$

Correlation ρ

Scatterplot

Correlation	Interpretation
$\rho > 0$	Positive dependency
$\rho = 0$	No (linear) dependency
$\rho < 0$	Negative dependency

```
cor()
plot()
```



Numeric Data: Scatterplot & Correlation

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\underbrace{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2 \cdot \sum_{j=1}^{n}(y_j - \bar{y})^2}}}$$

…only for scaling $-1 \leq \rho \leq 1$

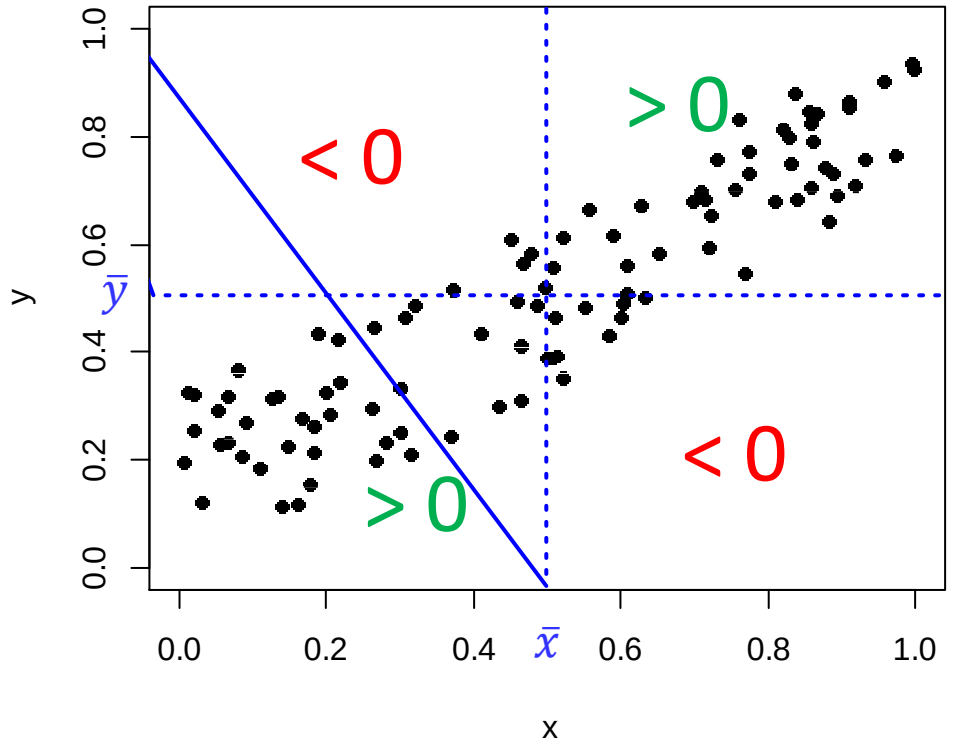
Correlation	Interpretation
$\rho > 0$	Positive dependency
$\rho = 0$	No (linear) dependency
$\rho < 0$	Negative dependency

Numeric Data: Scatterplot & Correlation

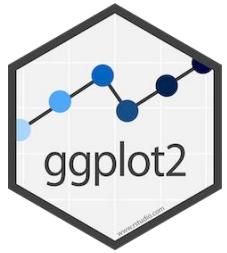
$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\underbrace{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2 \cdot \sum_{j=1}^{n}(y_j - \bar{y})^2}}}$$

…only for scaling $-1 \leq \rho \leq 1$

Correlation	Interpretation
$\rho > 0$	Positive dependency
$\rho = 0$	No (linear) dependency
$\rho < 0$	Negative dependency



ggplot2 – A Grammar for Graphics



Basic elements:

- Base function `ggplot()`.
- Data: data frame providing the data.
- Aesthetics: `aes()` -- mapping from data to the plot.
- Geoms: `geom_*()` -- type of the plot.
- …stats: `stat` can be used to summarize the data.

```
library(ggplot2)
ggplot(Default, aes(x = default, y = balance)) + geom_boxplot()
ggplot(Default, aes(x = income, y = balance)) + geom_point(aes(color = default))
```

https://ggplot2.tidyverse.org/

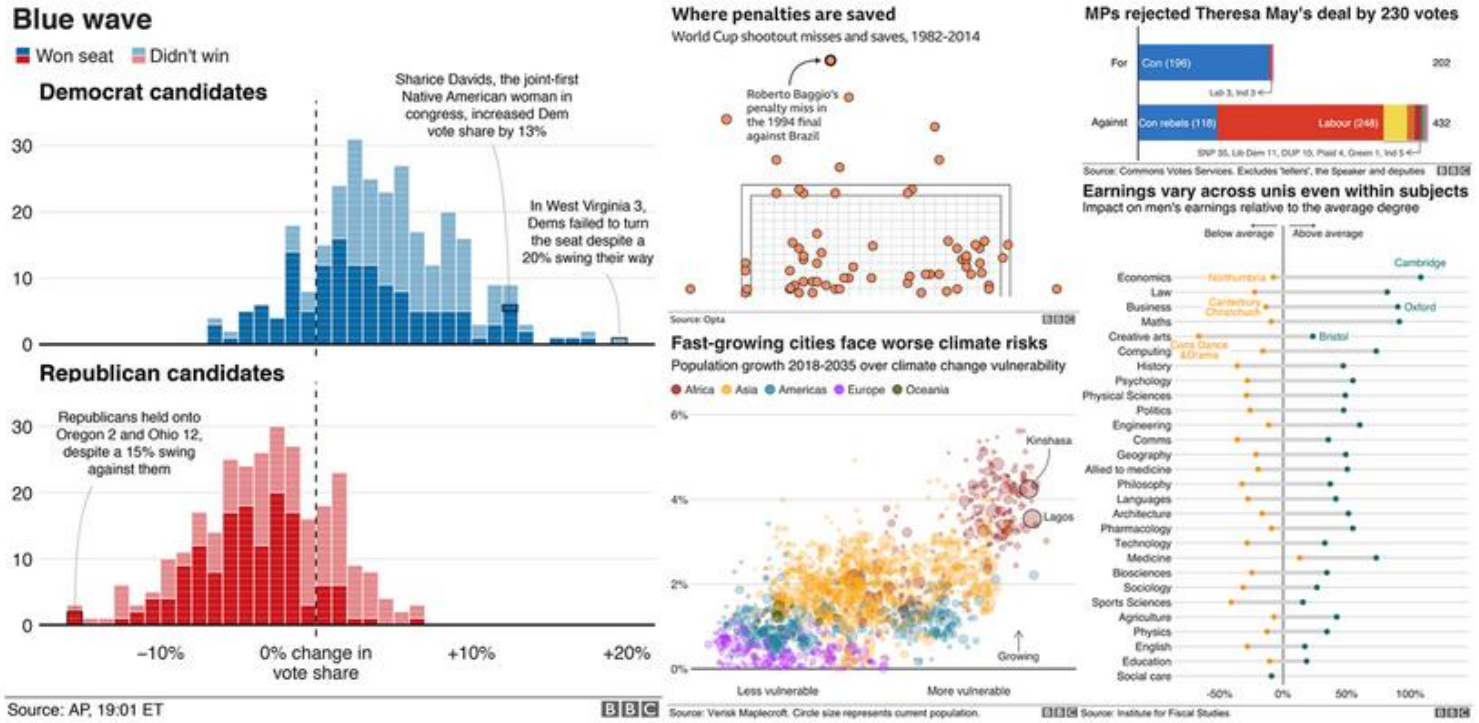
BBC Visual and Data Journalism cookbook for R graphics

Last updated: 2019-01-24

How to create BBC style graphics

At the BBC data team, we have developed an R package and an R cookbook to make the process of creating publication-ready graphics in our in-house style using R's ggplot2 library a more reproducible process, as well as making it easier for people new to R to create graphics.

The cookbook below should hopefully help anyone who wants to make graphics like these:



We'll get to how you can put together the various elements of these graphics, but **let's get the admin out of the way first...**

https://bbc.github.io/rcookbook/

Case Study: NYC Flights

Exercise…

1. Load the file NYCflights.Robj!
2. **Which data sets does it contain?**
3. **What kind of information do the data sets contain?**

4. Create a barplot of the variable flights$day!
5. **What do you observe? Explain!**
6. **Which airport (variable: origin) has the most flights?**
7. **What does the following code compute?**

```
flights %>% group_by(origin) %>% summarize(MAT = mean(air_time, na.rm=T))
```

8. **…What happens if you remove `na.rm = T` ?**

Merging Tables

```r
head(airports) # look up table

# base R
nd <- merge(flights, airports, by.x="dest", by.y = "faa", all.x = TRUE)

# dplyr
named_dests <- left_join(flights, airports, by = c("dest" = "faa"))
named_dests <- rename(named_dests, dest_airport = name)
# ...note the difference in computation time!
```

Exercise:

1. Filter the flights from airport „JFK“ to „Miami Intl“!
2. Use the `group_by()` and `summarize()` to compute the number of flights in 1. for each month!

Exercise: Analyzing Delays

1. Use `mutate()` and the function `wday()` (from the package lubridate) to create a new variable weekday from the variable time_hour in the flights data set!
2. Now use `group_by()` and `summarize()` to compute the average departure delay (variable: dep_delay) per weekday!

3. What are typical delays and flight durations? Create boxplots…
4. Describe your results!

5. Create a subset without missing data!

```
flights_complete <- flights[complete.cases(flights),]
```

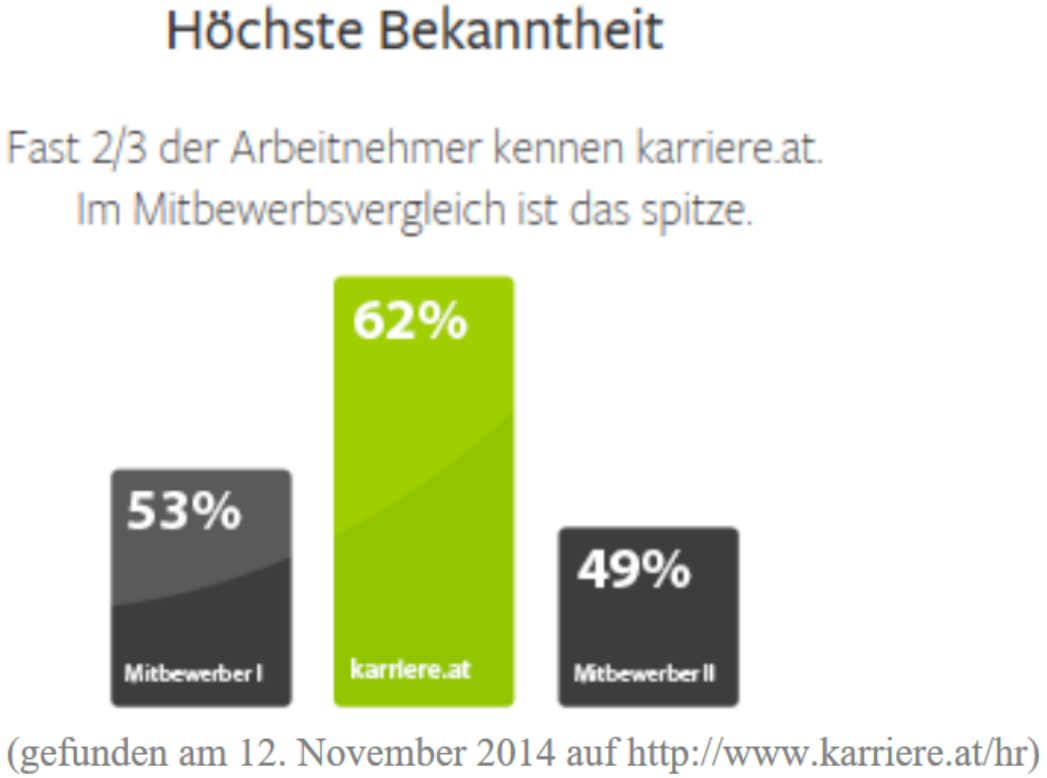
6. … **How many flights are removed from the data?**
7. … **Which variables have the highest dependency with the departure delay?**

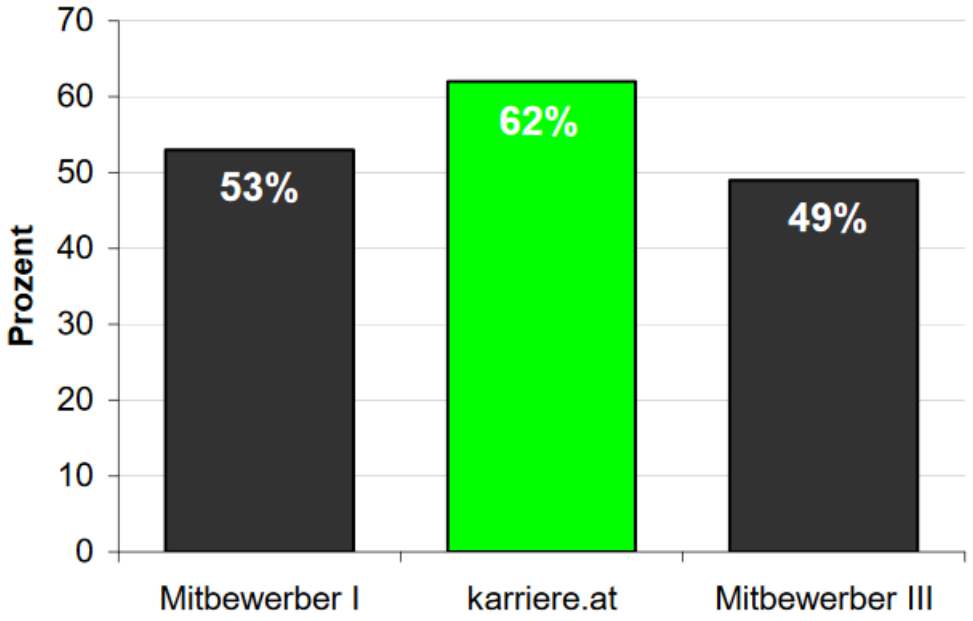
Cheating with Statistics…

What‘s wrong, here?

Source: A. Quatember, http://www.jku.at/ifas/content/e101235/e101334/e259008/bertriebenerVorsprunggegenberMitbewerberNov2014.pdf

What‘s wrong, here?

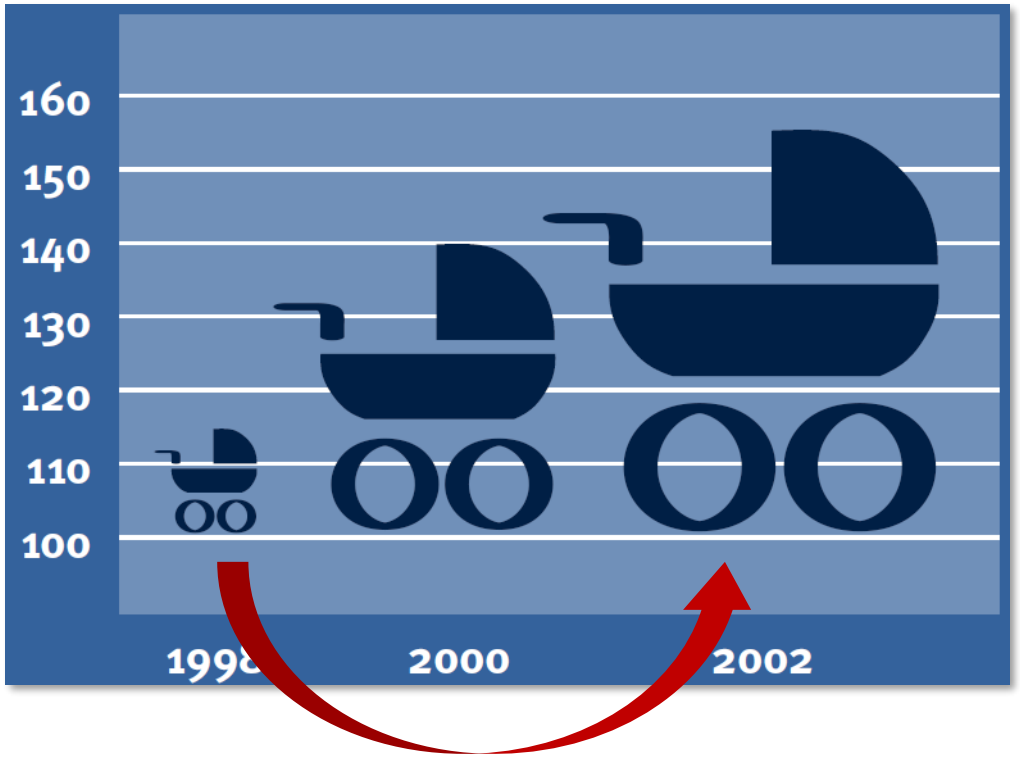




Source: A. Quatember, http://www.jku.at/ifas/content/e101235/e101334/e259008/bertriebenerVorsprunggegenberMitbewerberNov2014.pdf

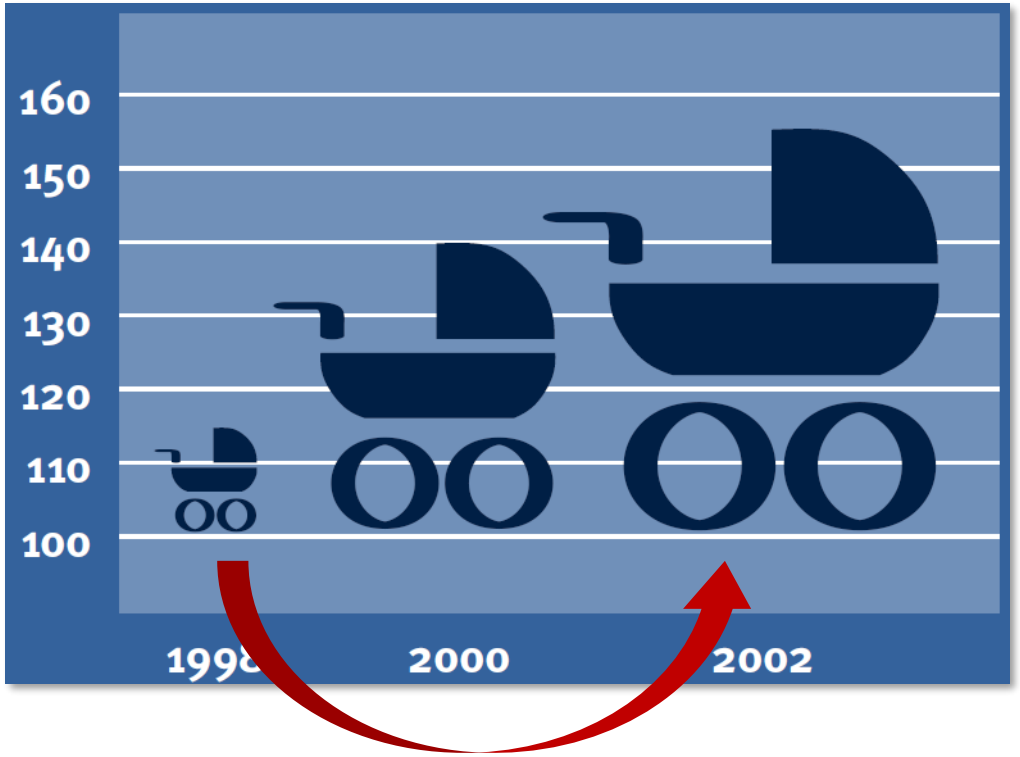
Cheating with Statistics II…

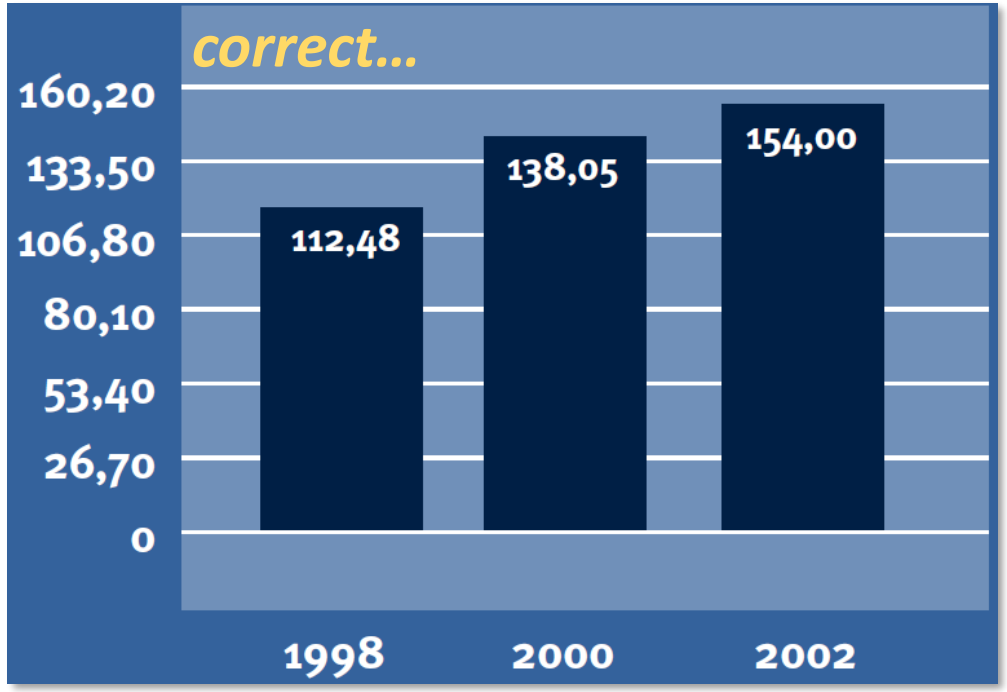
Estimate from the graph: By which factor did Kindergeld increase in Germany from 1998 to 2002?



Cheating with Statistics II…

Estimate from the graph: By which factor did Kindergeld increase in Germany from 1998 to 2002?

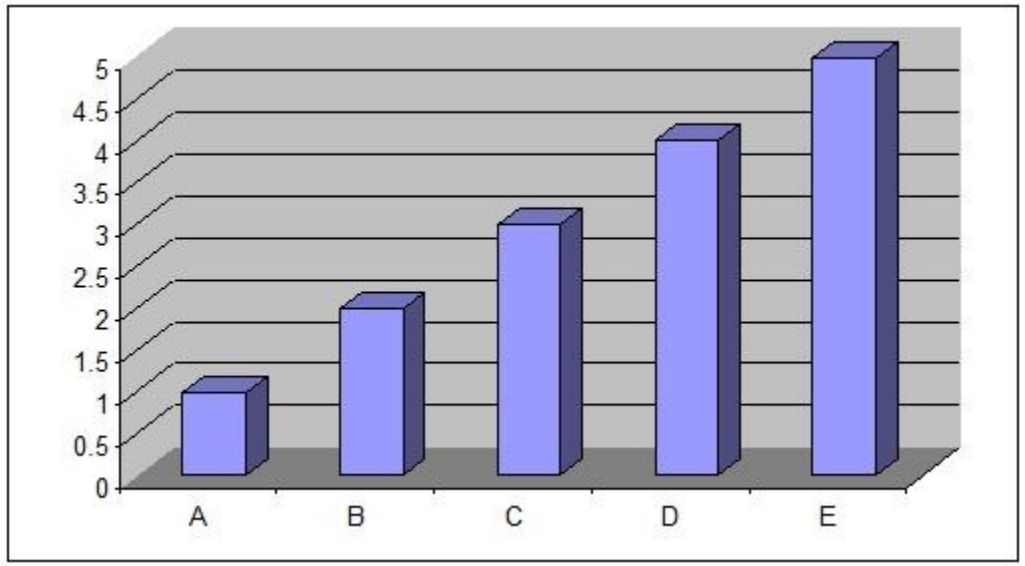




Source: http://www.wdr.de/tv/applications/fernsehen/wissen/quarks/pdf/Q_Zahlen.pdf

Example (III) of a **Bad** Bar Plot…

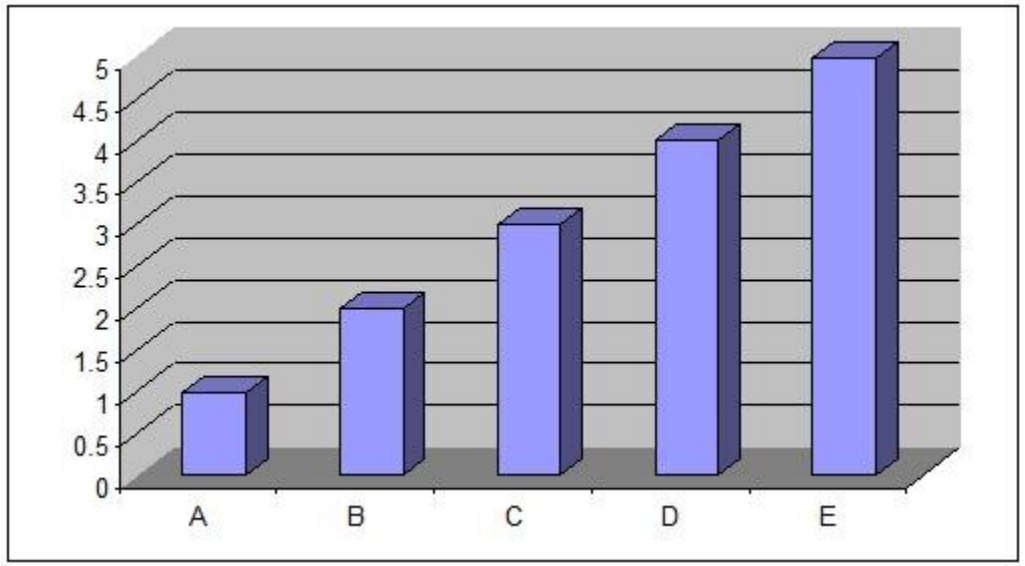
What value takes C?

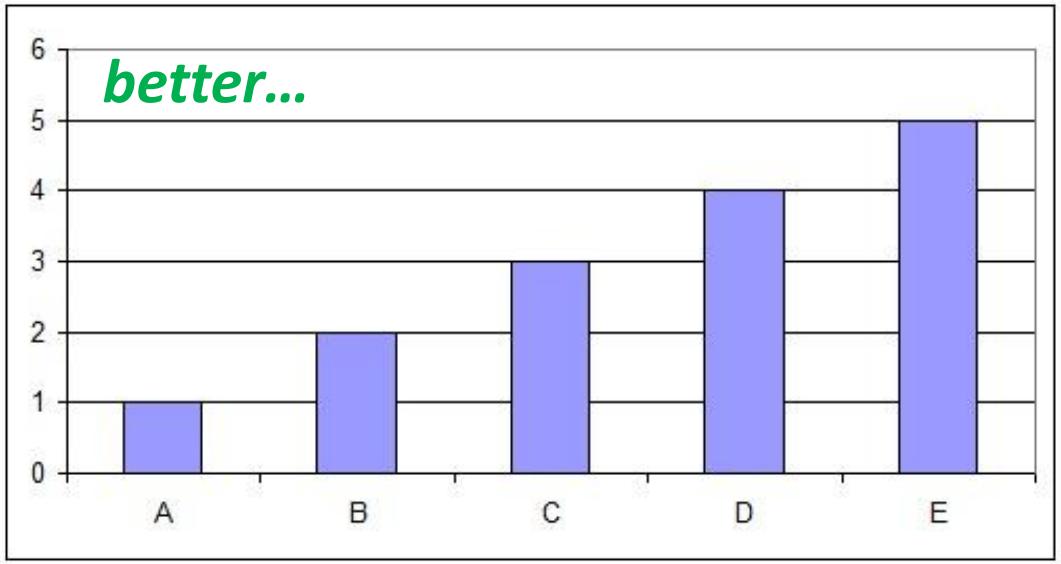


Source: http://consultantjournal.com/blog/use-3d-charts-at-your-own-risk

Example (III) of a **Bad** Bar Plot…

What value takes C?



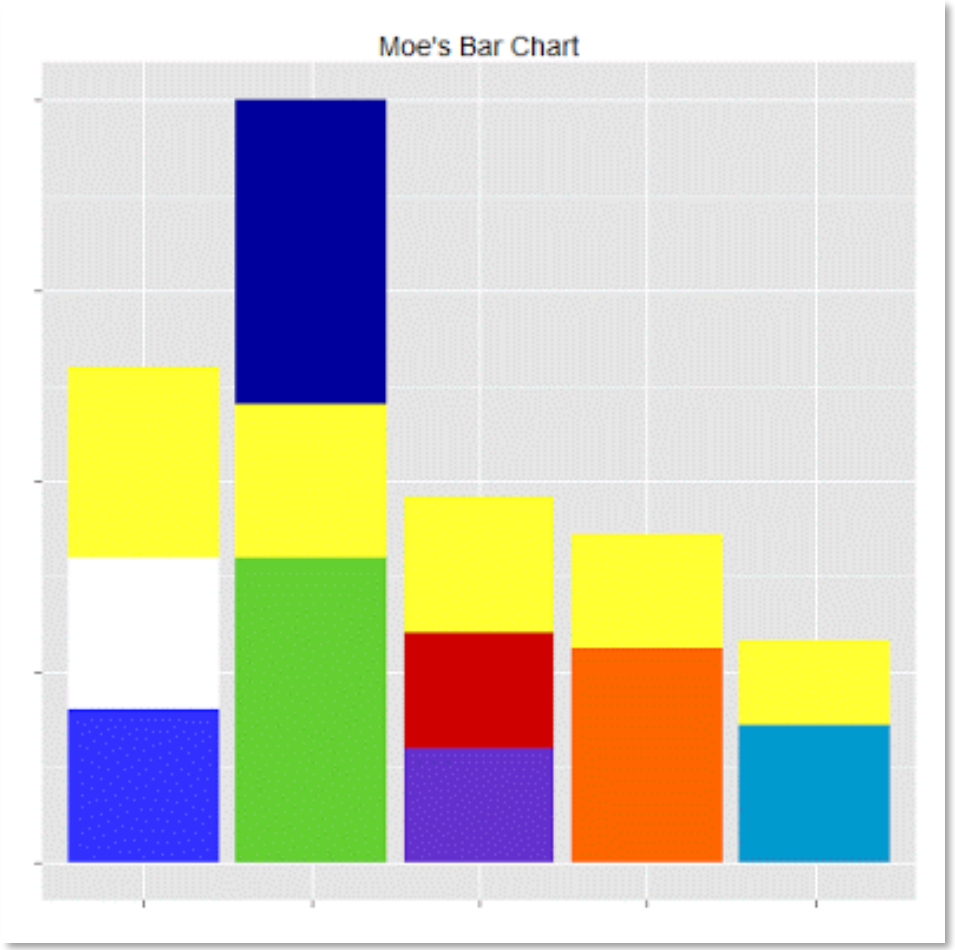


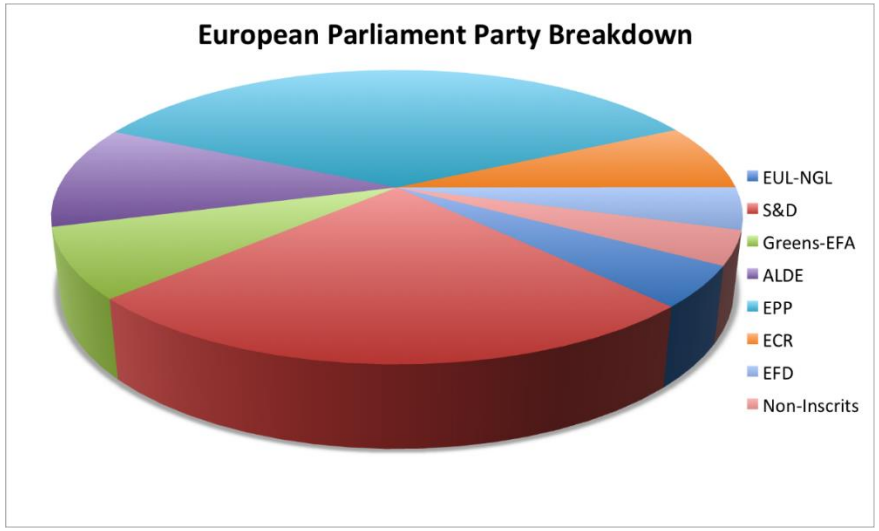
Why is this graph in 3D?

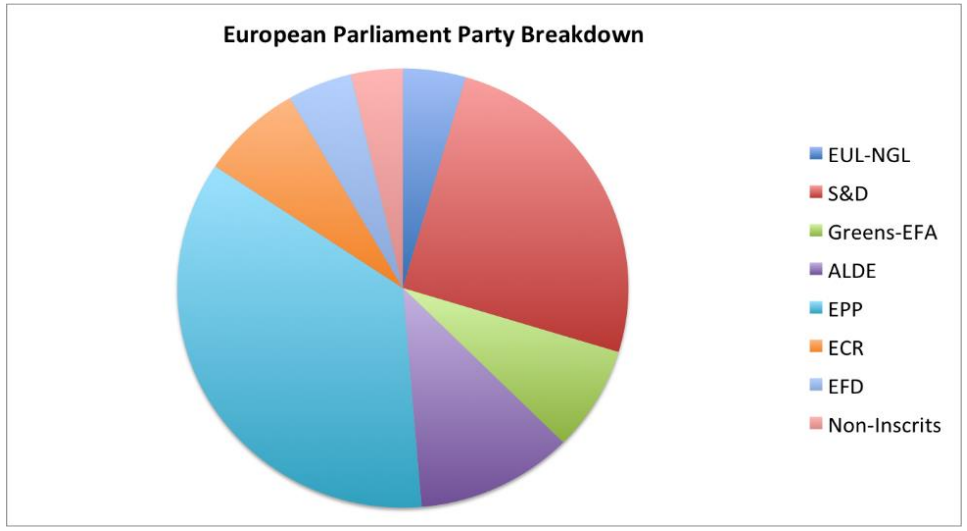
Is there any additional information in the left chart?

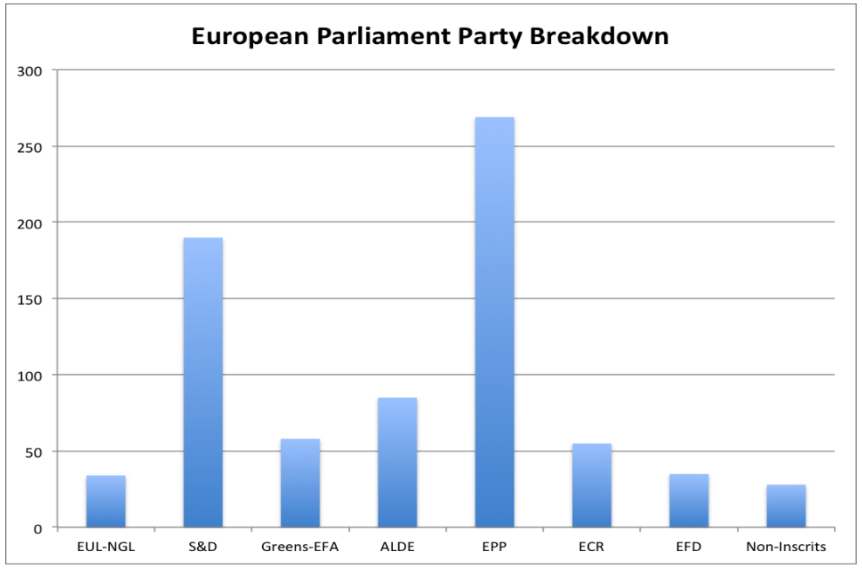
Source: http://consultantjournal.com/blog/use-3d-charts-at-your-own-risk

Example of a well-made Barplot





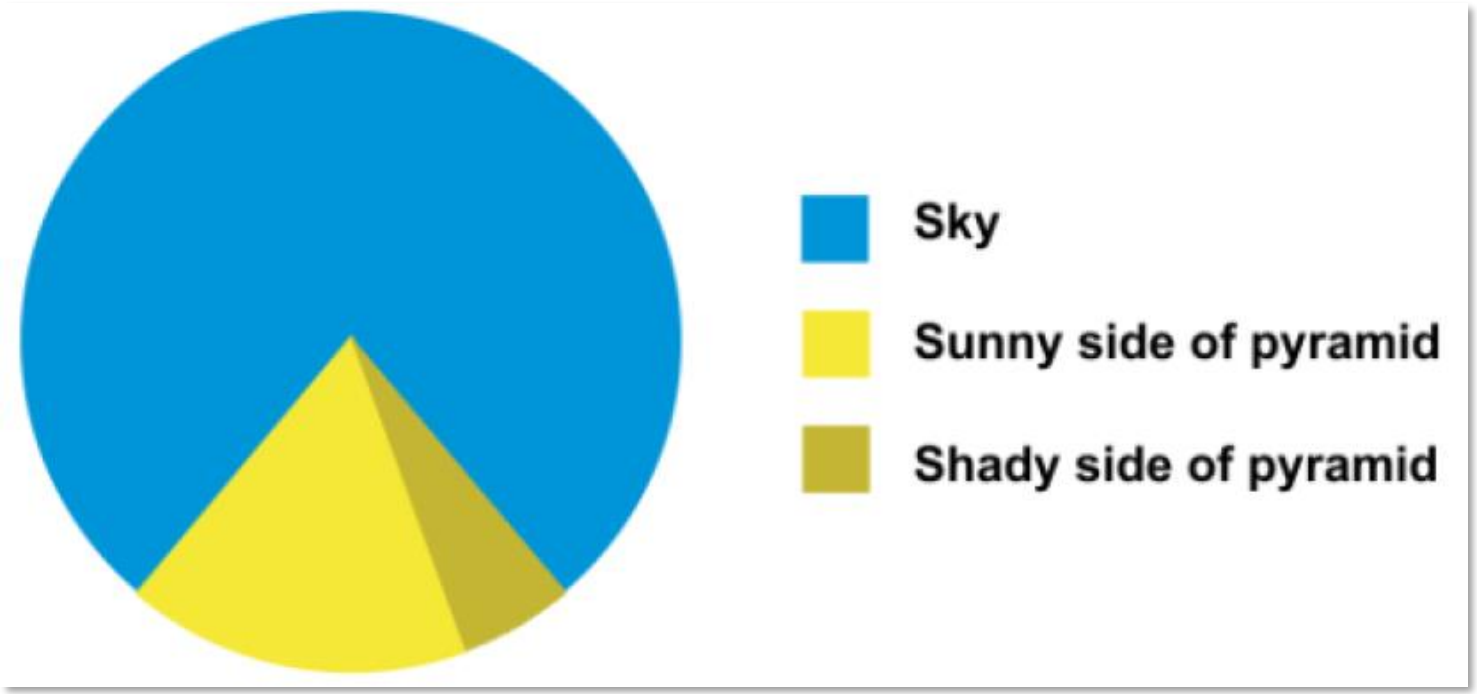




Source: http://www.businessinsider.com/pie-charts-are-the-worst-2013-6?IR=T

Takeaways

- Steps of the data science process
- Base R and the tidyverse
- Importing data
- A grammar for data wrangling
- A grammar for exploratory data analyses
- Using ggplot2 to create craphics


Sky
Sunny side of pyramid
Shady side of pyramid