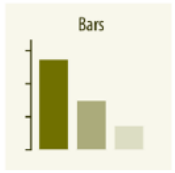
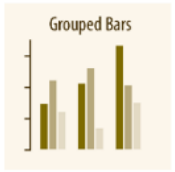
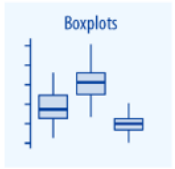
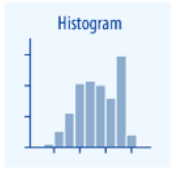
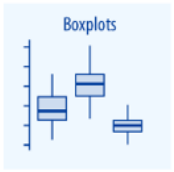
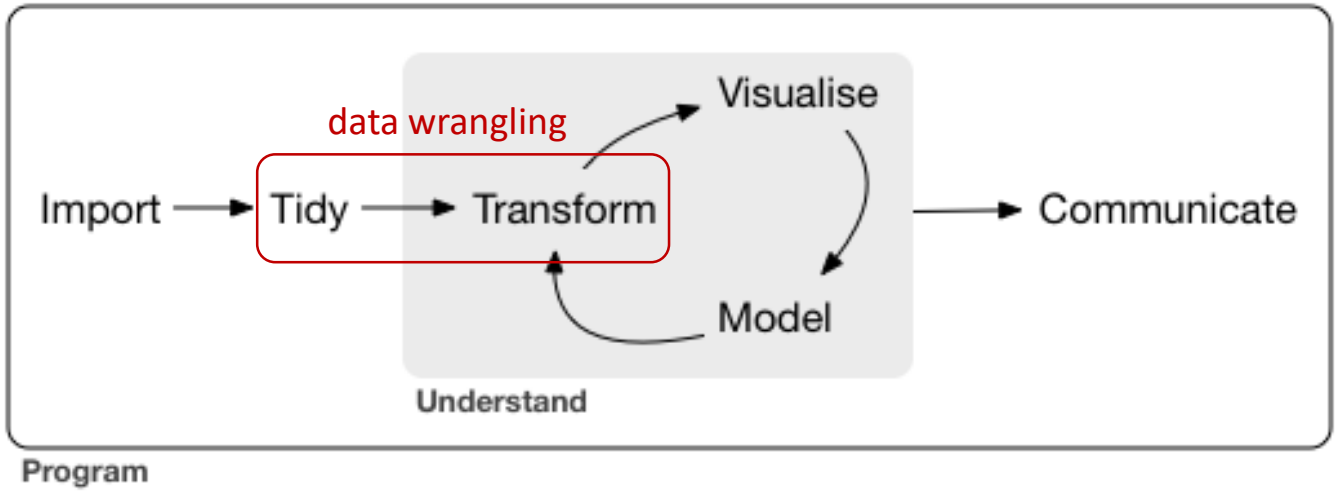
Data Analysis

Data Wrangling

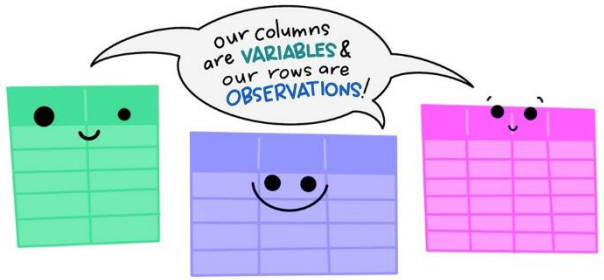
Prof. Dr. Gero Szepannek
Statistics, Business Mathematics & Machine Learning
Stralsund University of Applied Sciences

Variable 1	no 2nd variable	2nd variable categorical	2nd variable numeric
categorical	 Frequency tables	   Contingency tables χ^2 Test Cramer's V	 Boxplots Mean | group Variance | group
numeric	  Histogram Kernel density plot mean Variance / standard deviation Inter quartile range	 Boxplots Mean | group Variance | group	 Correlation

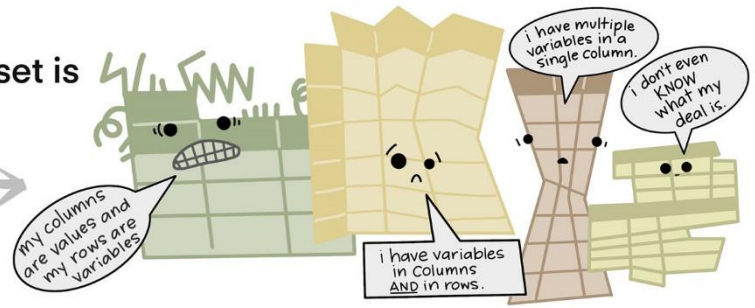


https://r4ds.had.co.nz/introduction.html

The standard structure of tidy data means that "tidy datasets are all alike..."



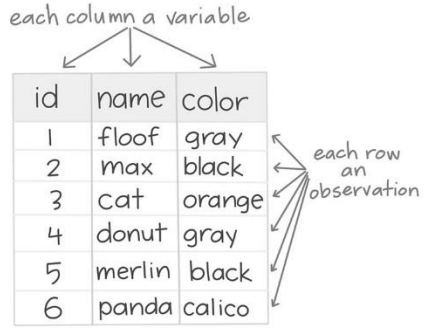
"...but every messy dataset is messy in its own way."
–HADLEY WICKHAM



"TIDY DATA is a standard way of mapping the meaning of a dataset to its structure."
–HADLEY WICKHAM

In tidy data:

- each variable forms a column
- each observation forms a row
- each cell is a single measurement



id	name	color
1	floof	gray
2	max	black
3	cat	orange
4	donut	gray
5	merlin	black
6	panda	calico

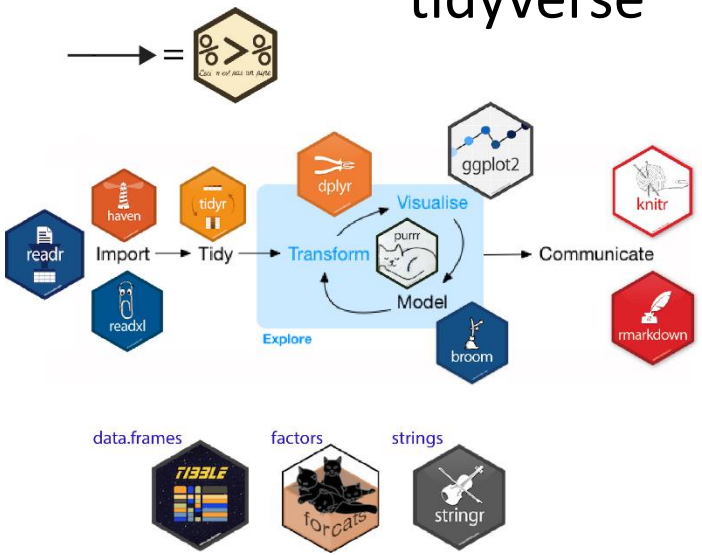
Wickham, H. (2014). Tidy Data. Journal of Statistical Software 59 (10). DOI: 10.18637/jss.v059.i10

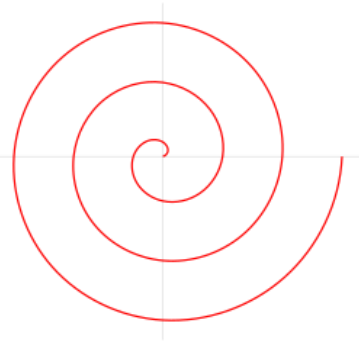
https://allisonhorst.com/other-r-fun

Traditional

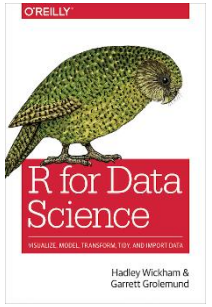
vs.

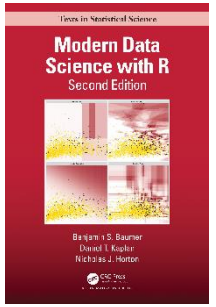
“tidyverse“





A Grammar for Data Wrangling





The verbs:

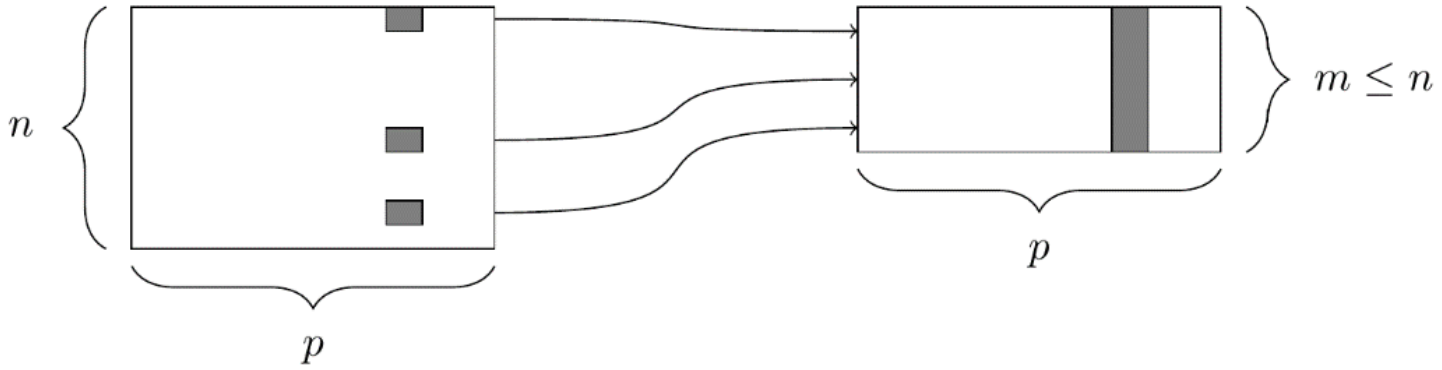
1. `select()` …a subset of columns (variables).
2. `filter()` …a subset of rows (observations).
3. `mutate()` Add or modify existing variables.
4. `arrange()` Sort the observations.
5. `summarize()` Aggregate the data accross observations according to some criteria.

General usage:

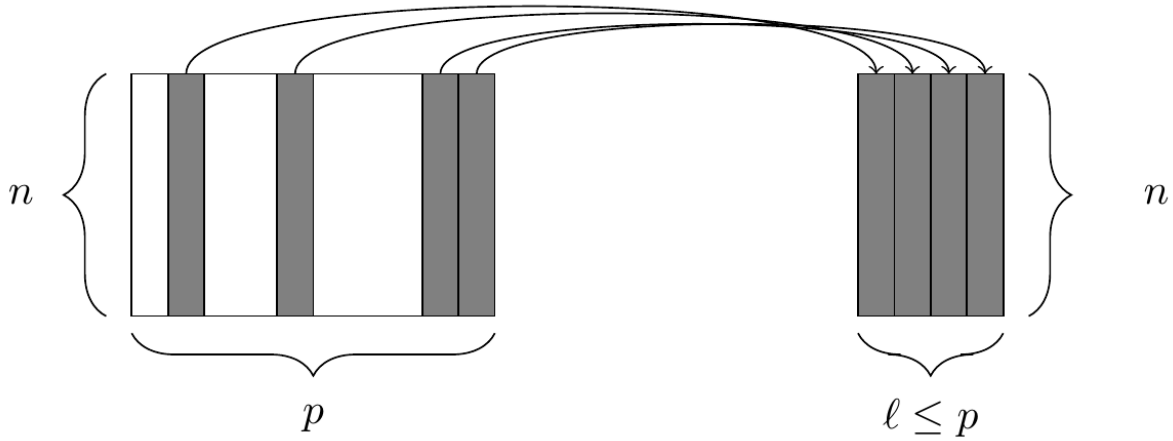
1. The first argument is a data frame.
2. The subsequent arguments describe what to do with the data frame, using the variable names (without quotes!).
3. The result is a new data frame.

Select and Filter

`filter()`

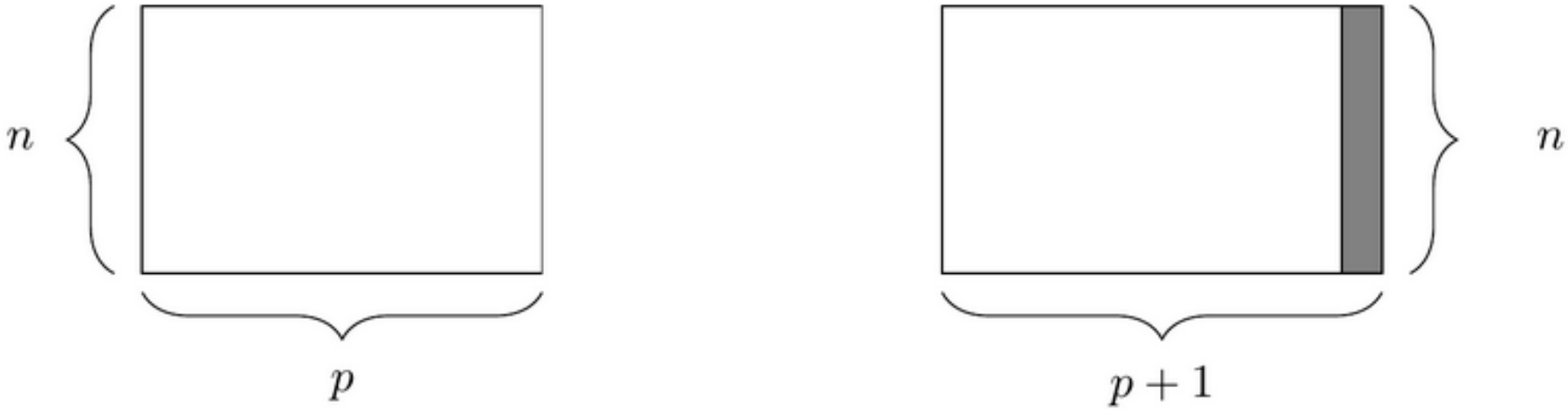


`select()`



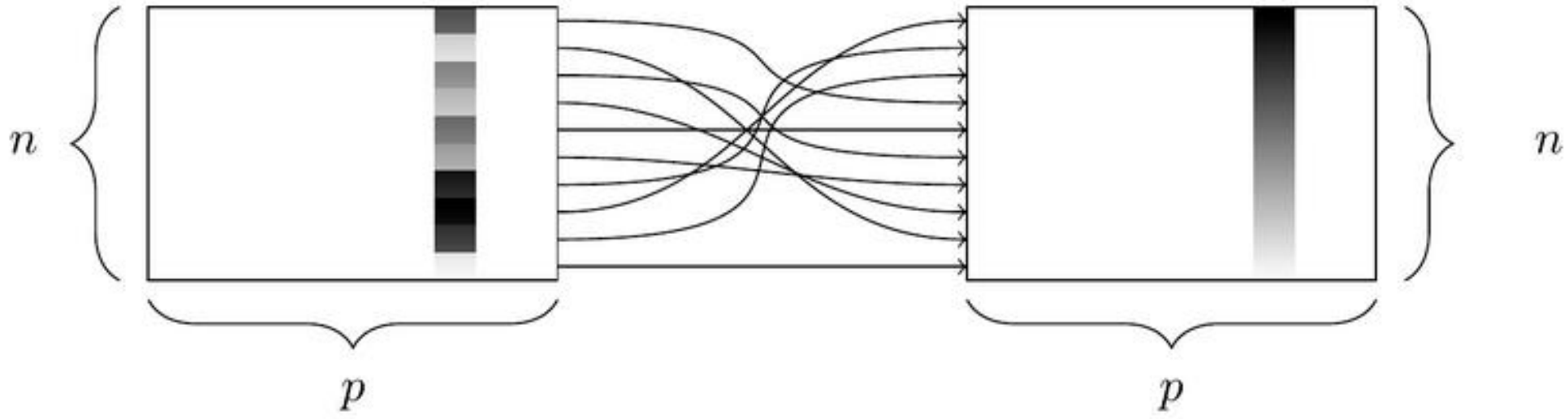
Mutate and Rename

`mutate()`



Arrange

`arrange()`




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

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

```
by_day <- group_by(flights, year, month, day)
summarise(by_day, delay = mean(dep_delay, na.rm = TRUE))
```

useful summary functions:

- `n()`
- `sum()`
- `mean()` — location
- `median()` — location
- `sd()` — variation
- `IQR()` — variation
- `mad()` — variation
- `min()`
- `max()`
- `quantile(x, 0.25)`
- `first()`
- `last()`
- `nth(x,2)`
- `n_distinct()`
- `sum(!is.na(x))`

Depending on the context there are 3 ways to deals with outliers:

1. Keep them
2. Remove them
3. Replace them

"…It's good practice to repeat your analysis with and without the outliers.
If they have a substantial effect on your results, you shouldn't drop them without justification. Try to figure out what caused them and disclose that you removed them in your write-up."

Always try to answer:

- Which values are the most common? Why?
- Which values are rare? Why? Does that match your expectations?
- Can you see any unusual patterns? What might explain them?

	GEOID	NAME	variable	estimate	moe
1	01	Alabama	income	24476	136
2	01	Alabama	rent	747	3
3	02	Alaska	income	32940	508
4	02	Alaska	rent	1200	13
5	04	Arizona	income	27517	148
6	04	Arizona	rent	972	4
7	05	Arkansas	income	23789	165
8	05	Arkansas	rent	709	5
9	06	California	income	29454	109
10	06	California	rent	1358	3
11	08	Colorado	income	32401	109
12	08	Colorado	rent	1125	5
13	09	Connecticut	income	35326	195
14	09	Connecticut	rent	1123	5
15	10	Delaware	income	31560	247
16	10	Delaware	rent	1076	10
17	11	District of Columbia	income	43198	681
18	11	District of Columbia	rent	1424	17

```
us_rent_income_wide <- pivot_wider(us_rent_income,
                                   names_from = variable,
                                   values_from = c(estimate, moe))
```

	GEOID	NAME	estimate_income	estimate_rent	moe_income	moe_rent
1	01	Alabama	24476	747	136	3
2	02	Alaska	32940	1200	508	13
3	04	Arizona	27517	972	148	4
4	05	Arkansas	23789	709	165	5
5	06	California	29454	1358	109	3
6	08	Colorado	32401	1125	109	5
7	09	Connecticut	35326	1123	195	5
8	10	Delaware	31560	1076	247	10
9	11	District of Columbia	43198	1424	681	17

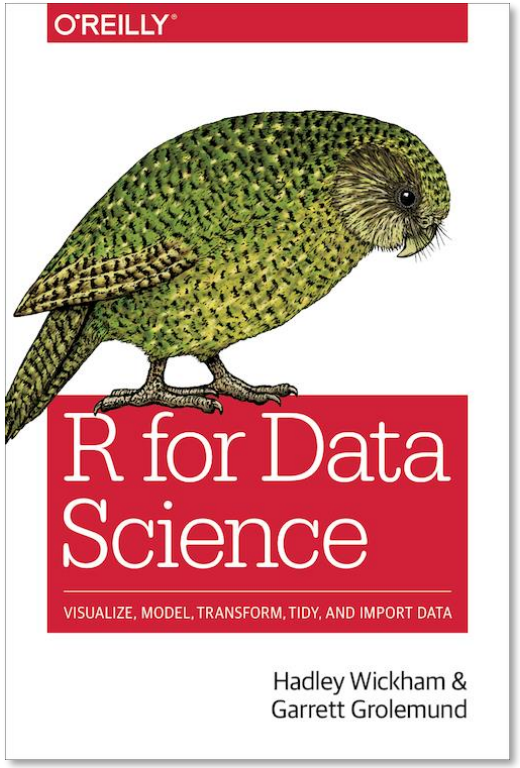
- `pivot_wider()` Eine (oder mehrere) qualitative Variable wird verwendet um weitere Spalten zu erzeugen und mehrere Beobachtungen zu einer zusammenzufassen
- `pivot_longer()` Mehrere Variablen werden zu einer Variable zusammengefasst. Dadurch entstehen mehr Beobachtungen und eine qualitative Variable, die angibt, welche der Ursprungsvariablen in der zusammengefassten Variable eingetragen sind

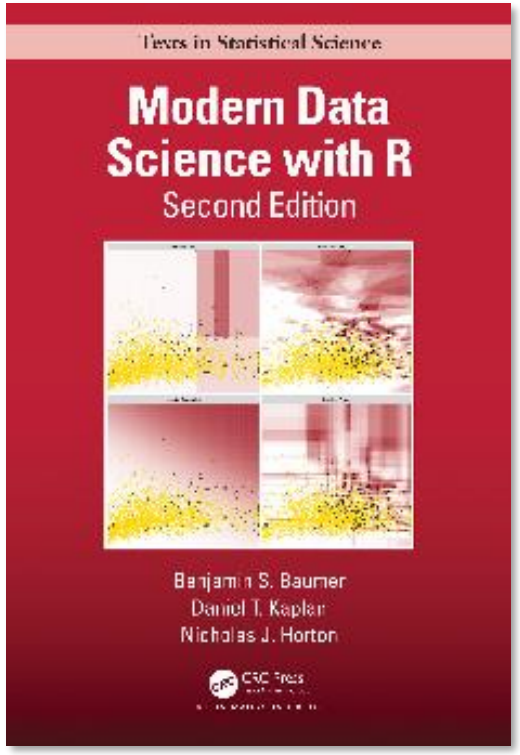
taken from Lars Koppers @ stRalsund R Usertreffen 6/2022

```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!
```





https://r4ds.had.co.nz/index.html
chapter 3: Data transformation with dplyr
chapter 9: Tidy Data with tidyr
chapter 10: Relational Data with dplyr
(online: chapter 5, 7.3, 12 & 13)