Tidyverse etc.

Divij Sinha 01-23

Recap

- What is ML?
- How to use git/github

Prerequisites for ML

- Good question
- Good method for answering the question
- Relevant data for method
- Good data

Tidyverse + Data Wrangling

- What is `tidyverse`?
    - Tidyverse is a collection of R packages designed for data science that share a common design philosophy and grammar
- Why `tidyverse`?
    - Allows us to have a consistent language that runs across many different types of functions and libraries.
    - Easier to start using new libraries + easier to integrate into previous work

Data ETL

- Once you identify data, you generally want to perform (at least) some of the following steps to start using it.
- Extract the data from the source(s)
- Transform the data into a useful form *
- Load the data TO the storage location/format that we will be using
- `tidyverse` is helpful across all of these

`tidyverse` **extract**

- Reading in different filetypes *and* locations
- Most commonly, reading in from local - files you have already on your laptop/computer
- Where else could they be?
  - Databases (could be local?)
  - Web files
  - S3
  - APIs

tidyverse extract

- Reading in different filetypes *and* locations
- Most commonly, reading in from local - files you have already on your laptop/computer
- File formats
  - csv (& tsv etc.)
  - Excel
  - Fwf rds etc.
  - Images? Text? - coming back in second half!

`tidyverse` load

- Possible file formats to write to
  - CSV
  - RDS
  - Parquet
  - Database files
  - Excel?
  - Why pick one over the other?

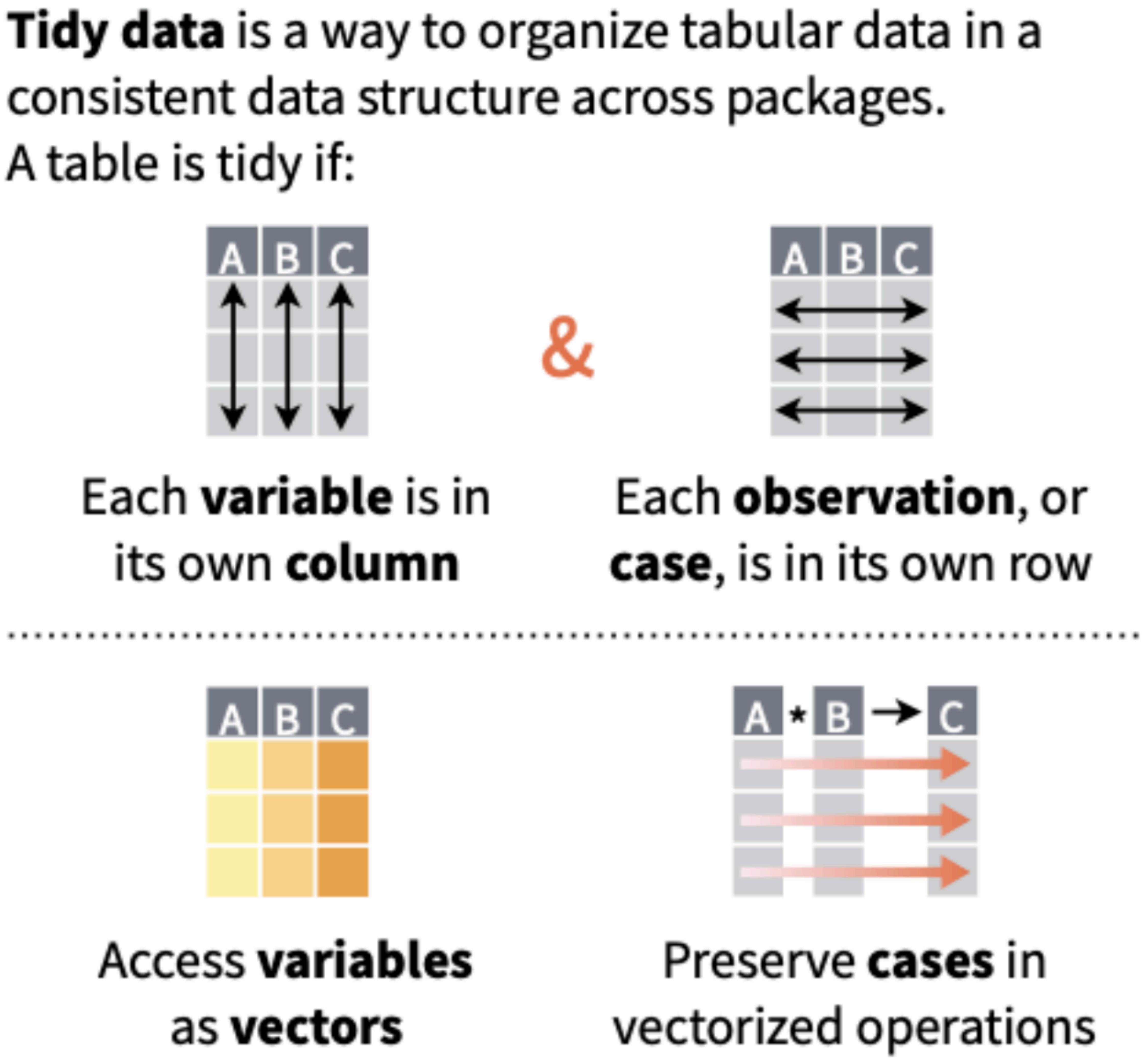
tidyverse transform

- Most common/time-consuming amongst the ETL
- Adding new columns - Create calculated columns using existing data
- Clean your data - remove duplicates, handle missing values etc.
- Tidy Data - What?

Tidy Data

- Why?

- https://vita.had.co.nz/papers/tidy-data.pdf



Tidy data is a way to organize tabular data in a consistent data structure across packages.
A table is tidy if:

Each **variable** is in its own **column**

&

Each **observation**, or **case**, is in its own row

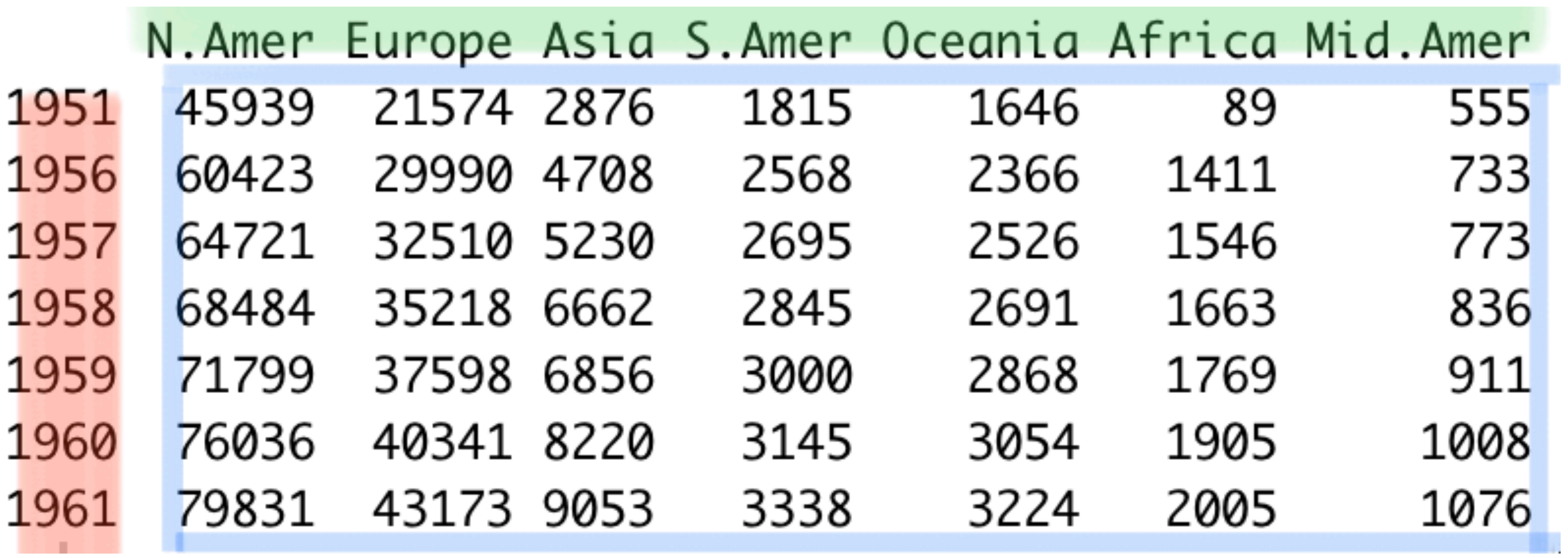
Access **variables** as **vectors**

Preserve **cases** in vectorized operations

Wide-form data

Number of Phones (in 1000) in each Region by Year

- Wide data



```
     N.Amer Europe Asia S.Amer Oceania Africa Mid.Amer
1951  45939  21574 2876   1815    1646     89      555
1956  60423  29990 4708   2568    2366   1411      733
1957  64721  32510 5230   2695    2526   1546      773
1958  68484  35218 6662   2845    2691   1663      836
1959  71799  37598 6856   3000    2868   1769      911
1960  76036  40341 8220   3145    3054   1905     1008
1961  79831  43173 9053   3338    3224   2005     1076
```

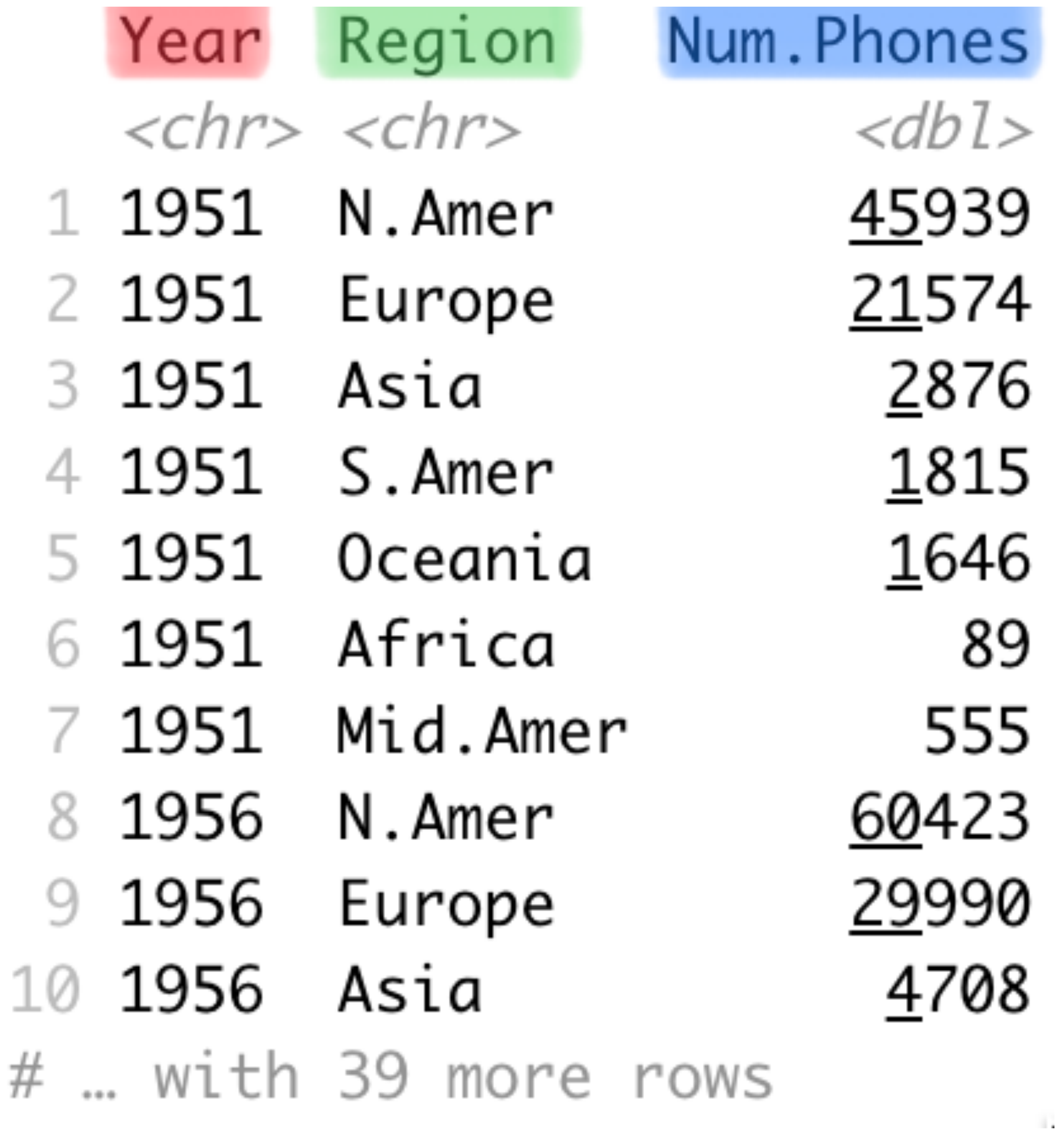
- Each row is a Year
- Each column is a Region
- Each cell is a Number of Phones

- Natural, human readable, makes sense to us

Long-form data

Number of Phones (in 1000) in each Region and Year

- Long data



```
   Year  Region       Num.Phones
   <chr> <chr>             <dbl>
 1 1951  N.Amer            45939
 2 1951  Europe            21574
 3 1951  Asia               2876
 4 1951  S.Amer             1815
 5 1951  Oceania            1646
 6 1951  Africa               89
 7 1951  Mid.Amer            555
 8 1956  N.Amer            60423
 9 1956  Europe            29990
10 1956  Asia               4708
# … with 39 more rows
```

- Each row represents a Number of Phones in ” Year-Region”
- Each cell is a value of Year, Region, Number of Phones depending on the column

- Longer, harder to quickly glance
- Each row has a specific meaning
- MUCH easier to code with (for eg., can filter on Year, Region and Number of Phones and get the subset of rows we care about easily!)

Exploratory Data Analysis

Have to know what you are working with!

EDA

6.2 The gist of EDA
A few things to consider. In many ways, EDA is a “pre-flight checklist”—a chance to kick the tires on
the data and a proposed project. Even before touching the data, consider formulating a set of questions that
you would like to have answered. These answers should determine if you should proceed with a data project
or if the EDA itself is the project.

EDA - things to think about

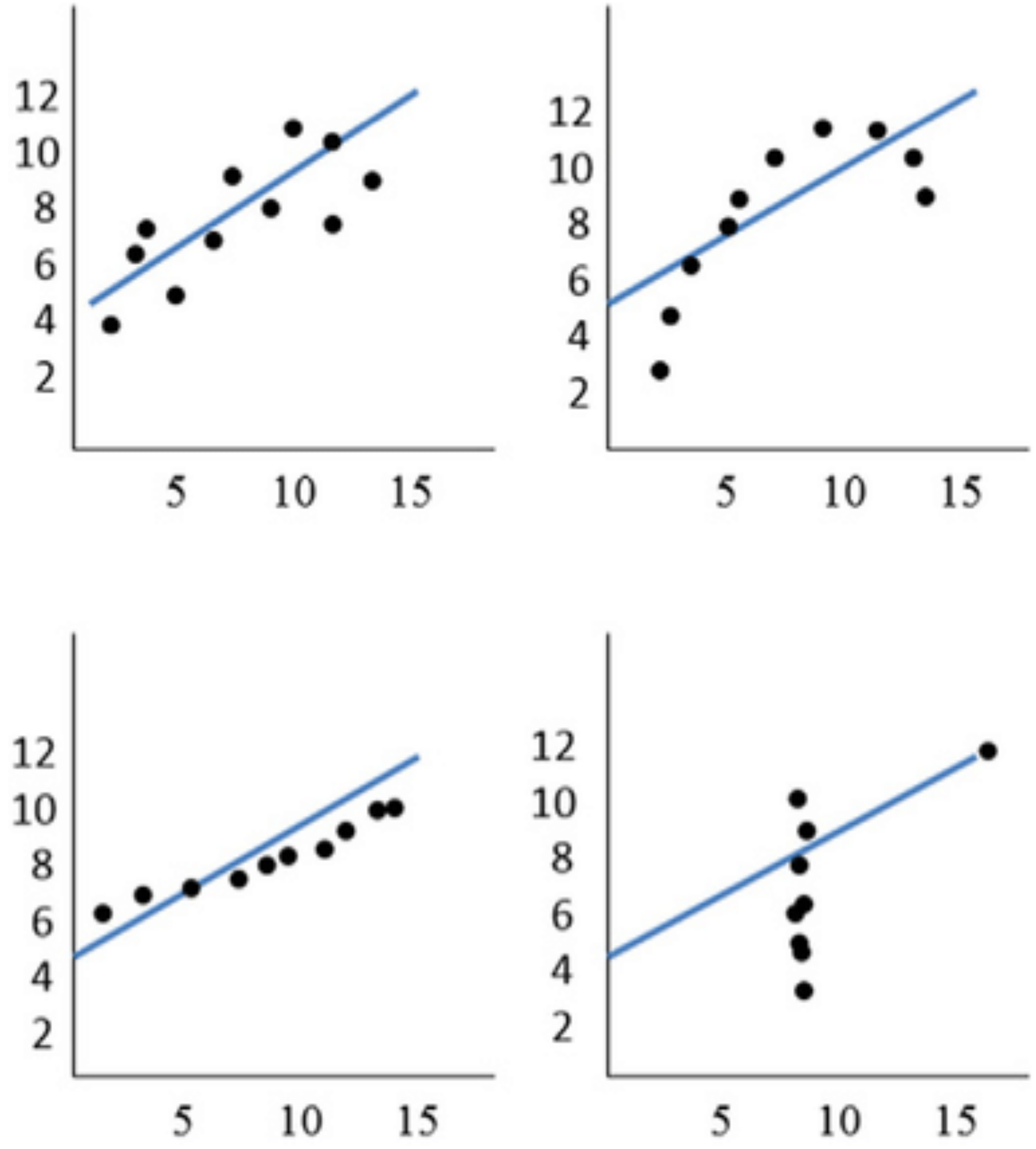
- Is data in the right format?
- Is some of the data missing?
- Do the distributions “look right”?
- Ask common sense questions, expect common sense answers

Visualisations

- Very important first step
- Allows us to visually confirm our instincts
- Why not rely on just summary statistics?
- NOT for publications yet - quick and dirty are good, the more the merrier!

Anscombe's Quartet of Different XY Plots of Four Data Sets Having Identical Averages, Variances, and Correlations

Anscombe's Quartet



Property	Value
Mean of X (average)	9 in all 4 XY plots
Sample variance of X	11 in all four XY plots
Mean of Y	7.50 in all 4 XY plots
Sample variance of Y	4.122 or 4.127 in all 4 XY plots
Correlation (r)	0.816 in all 4 XY plots
Linear regression	y = 3.00 + (0.500 x) in all 4 XY plots

Data sets for the 4 XY plots

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	5.76
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	8.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	7.26	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Source: Adapted from Anscombe (1973, pp. 19-20)

ggplot2

Plotting stuff

- Grammar of Graphics
  - Built on layers - canvas, geoms, facets etc.
- Think simple

sf

Mapping stuff

- Data Types - POINTS, LINESTRINGS, POLYGONS
- CRS
- Where to find and what to look for?
- Spatial Queries
- Spatial Aggregations
- Geometry Manipulation

Can AI solve a problem?

What does it mean for AI to "solve" a problem?

What problems is it good at solving?

- Narrow vs. General AI
- Disease
- Drug Discovery - alphafold
- Image recognition
- Personalized Recommendations
- Algorithmic Trading

NOT THE SAME AXIS AS HUMANS

Deep and narrow vs shallow and broad

Examples

- Counting - tokenization
- Image description - units of novelty
- Facial Recognition - Sensitivity vs specificity
- Image generation - example
- Chatbots - human like?