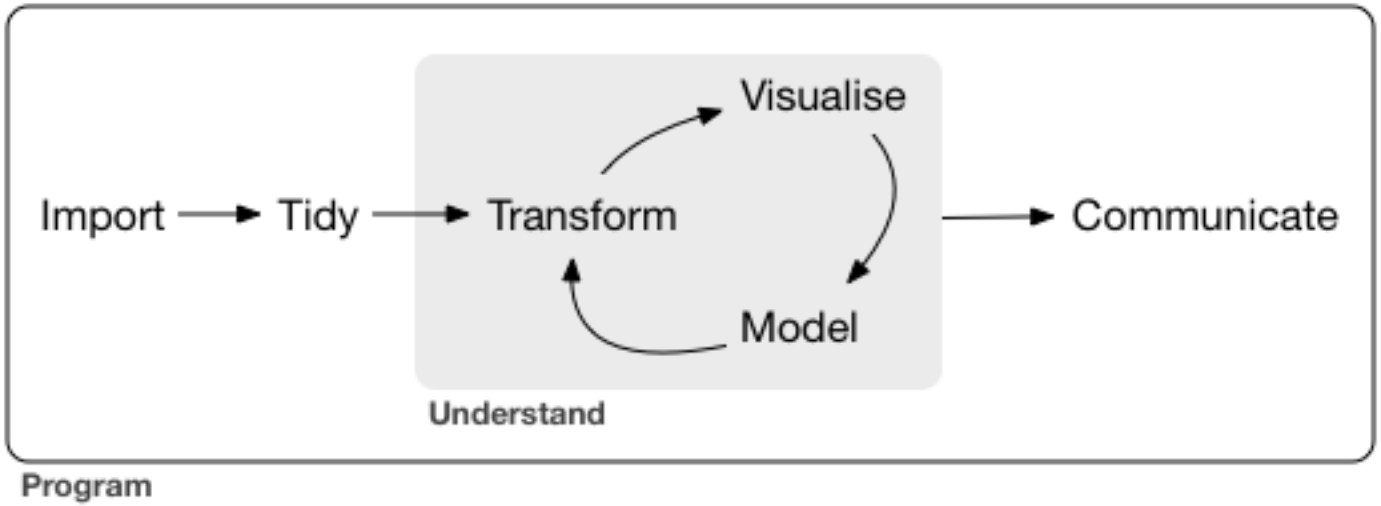

# Data Science process

# Data Science process



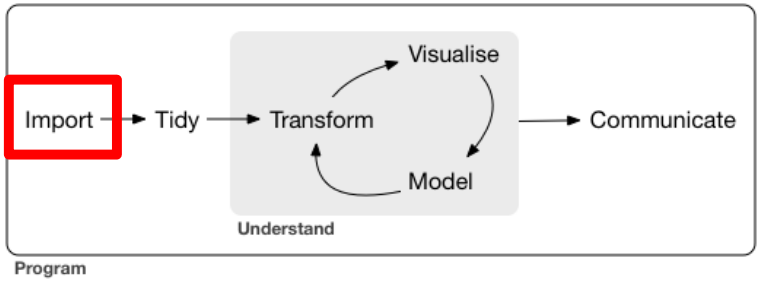

Tidy + Transform = Wrangle

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

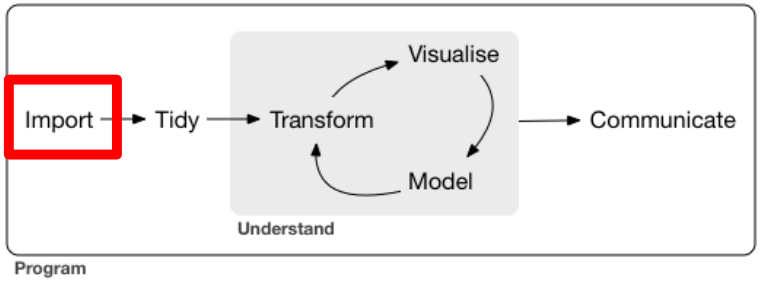
# Import

Data sources in education:

- Student record data
- Staff data
- Admissions & applications data
- Financial data
- Alumni data
- Course data
- Estates and facilities data
- Virtual Learning Environments
- Assessment data
- Forum data



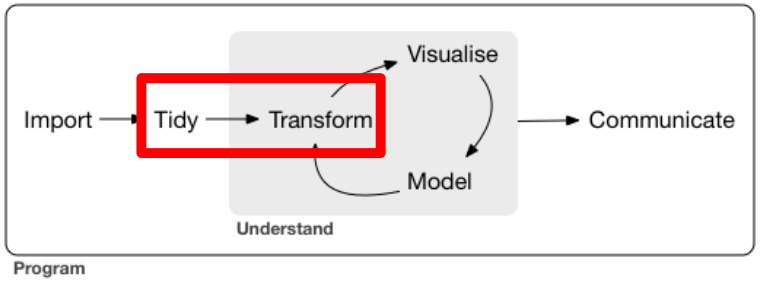
# Import




Data sources types:

- Files
  - CSV, XML, JSON
- Databases
  - SQL, NoSQL (key-value, graph-based, document-based,...)
- API (Social media,...)

## Wrangle



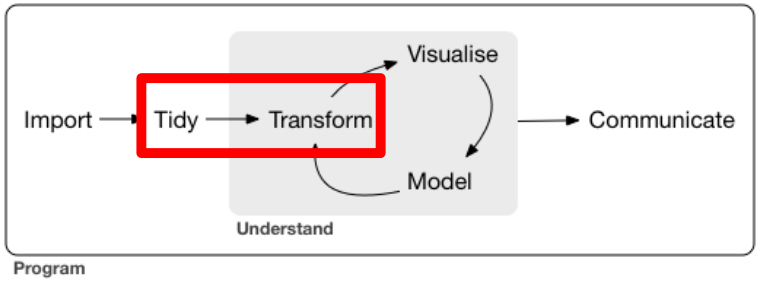

# Tidy + Transform = Wrangle

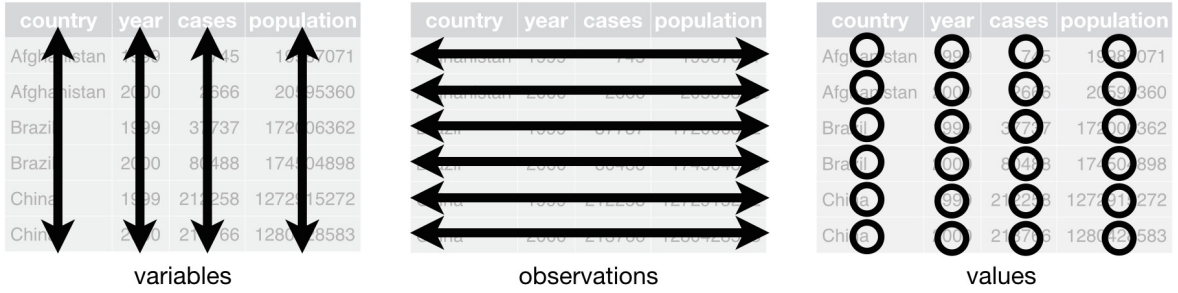
Prepare data for Visualisation and Modeling

# Tidy

Tidy dataset:

- Each column represents one variable
- Each row represents one observation
- Each cell represent one value



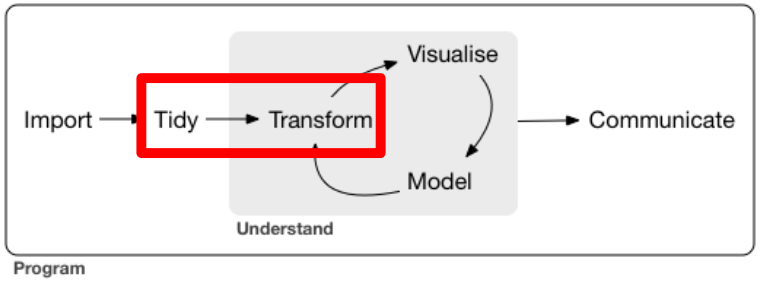

Problems:

- Multiple data sources with different unique identifier
- Values in one column represents multiple variables
- One observation spreads in multiple rows

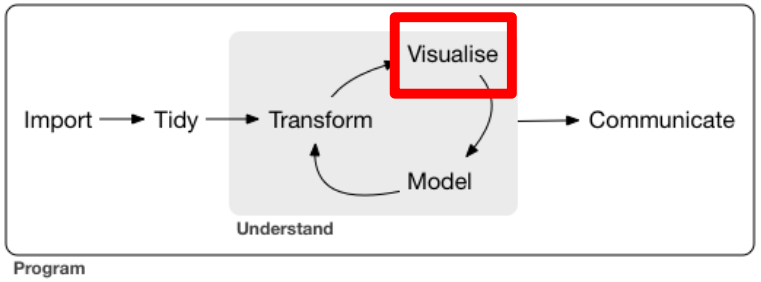
https://r4ds.had.co.nz/tidy-data.html

# Transform

- Handle missing data
- Inconsistent data types
- Outliers
- Encoding
- Filtering the data
- Aggregation of the data
- Transforming values
- Handling texts and dates

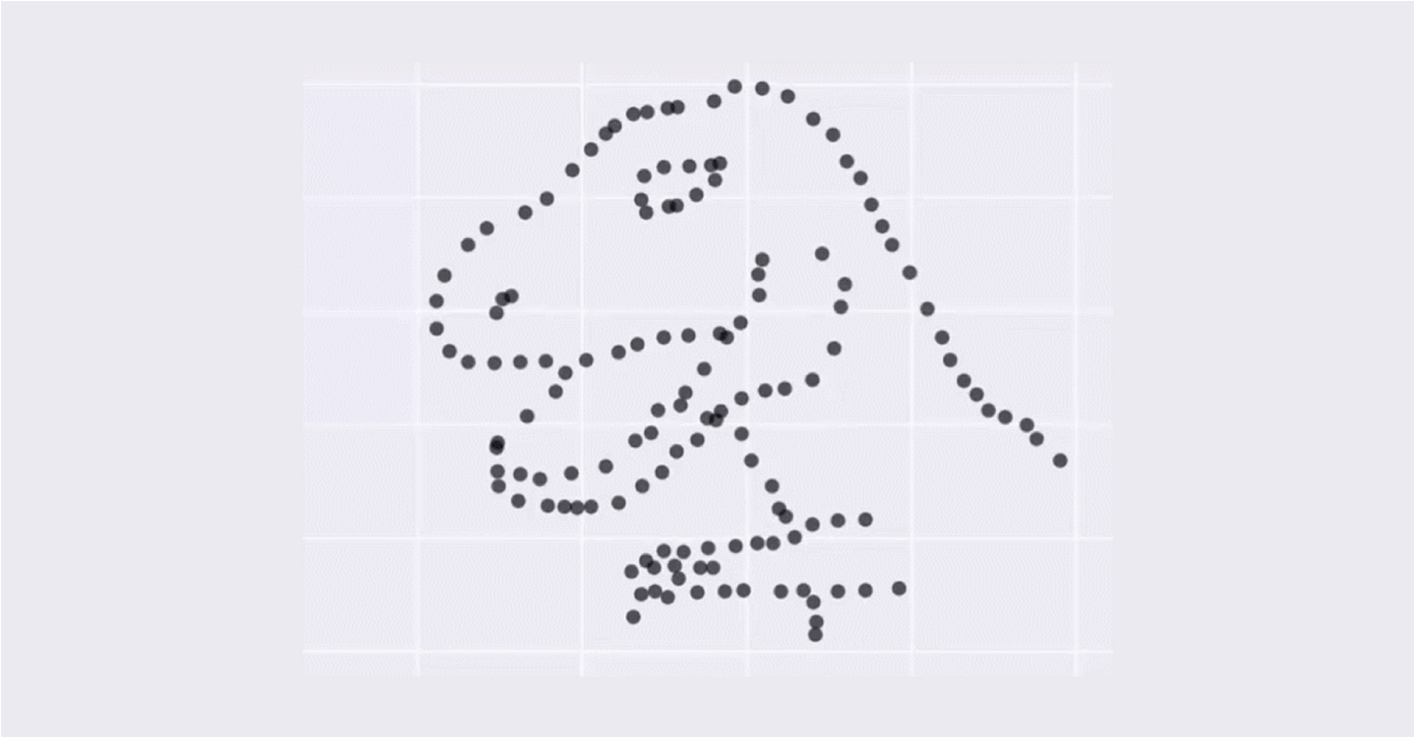


## Visualize



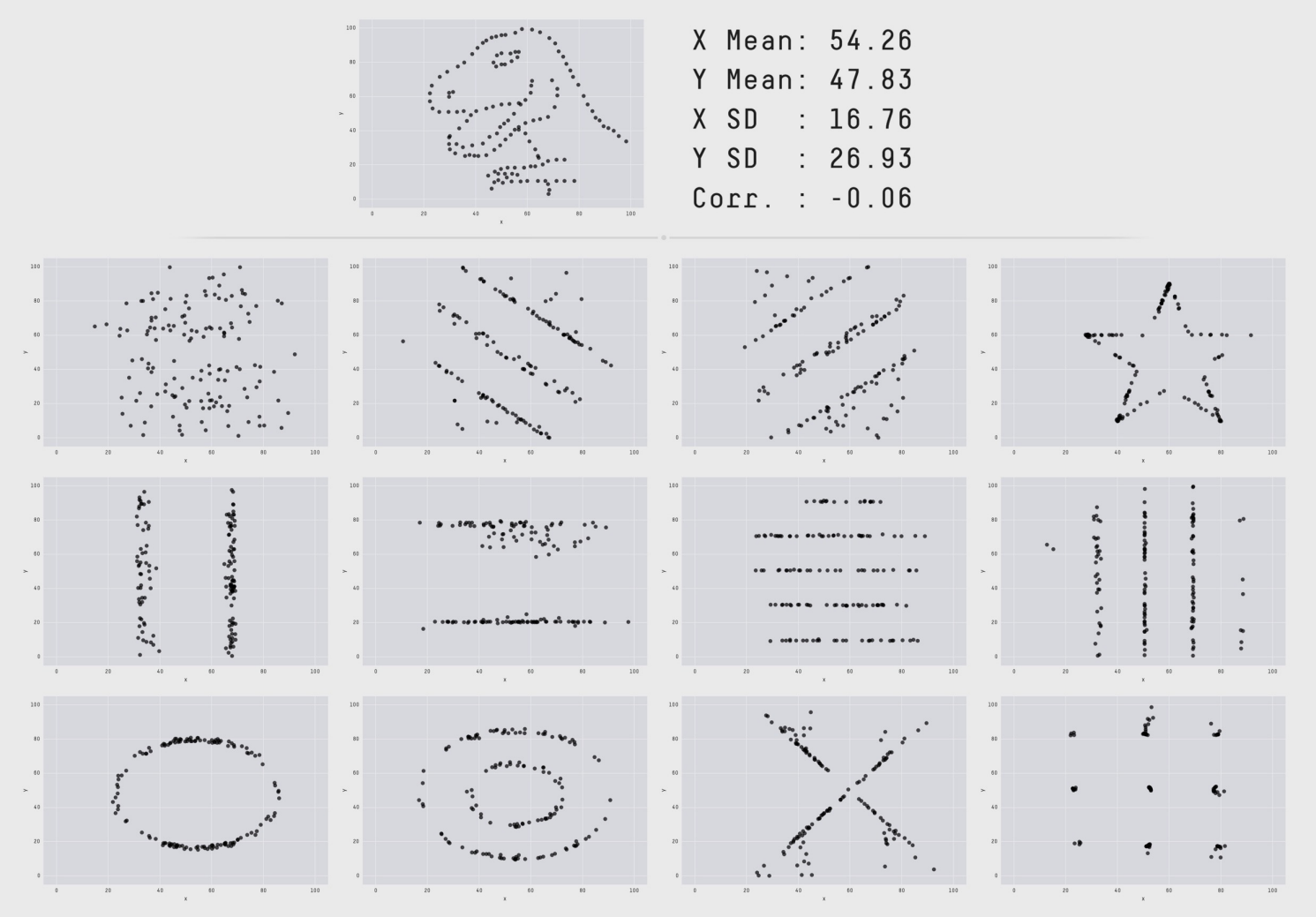

- Visualization is useful tool for providing the information to stakeholder but also during wrangling the data
- Helps to understand issues in the data
- Uncovers outliers
- Helps to identify relationships between variables

# Datasaurus



https://www.autodesk.com/research/publications/same-stats-different-graphs

# Datasaurus

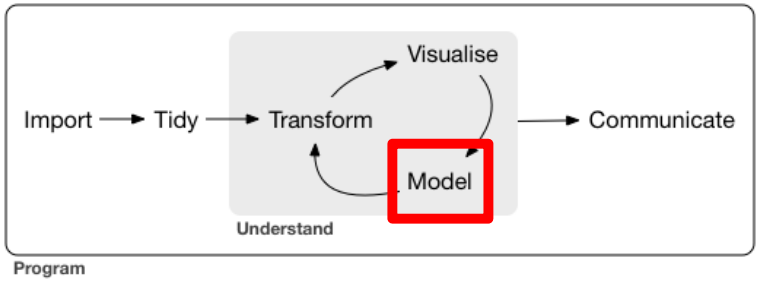



https://www.autodesk.com/research/publications/same-stats-different-graphs

# *Do not trust your data blindly*

- **Always check data visually.**
- **Statistics can be misleading.**

# Model




- Tidy data can be used for creating of model
- For that Machine Learning methods can be used

Machine Learning is:

*"Field of study that gives computers the ability to learn without being explicitly programmed"*

~ Arthur Samuel, 1959

- Machine Learning is subfield of Computer Science
- Objective: *Generalize from experience*

## Communicate

Communicate -> Dashboard

- Deliver information to stakeholder in interactive way
- Automation of the analysis
- Includes possibility for user to adjust some parameters

- Challenges:
  - Scalability
  - Data quality
  - User interface
  - Evaluation

- Issues:
  - Too much colour
  - Too much details
  - Useless decorations
  - Poor visualisations