Machine Learning cycle

Dr. Jakub Kuzilek

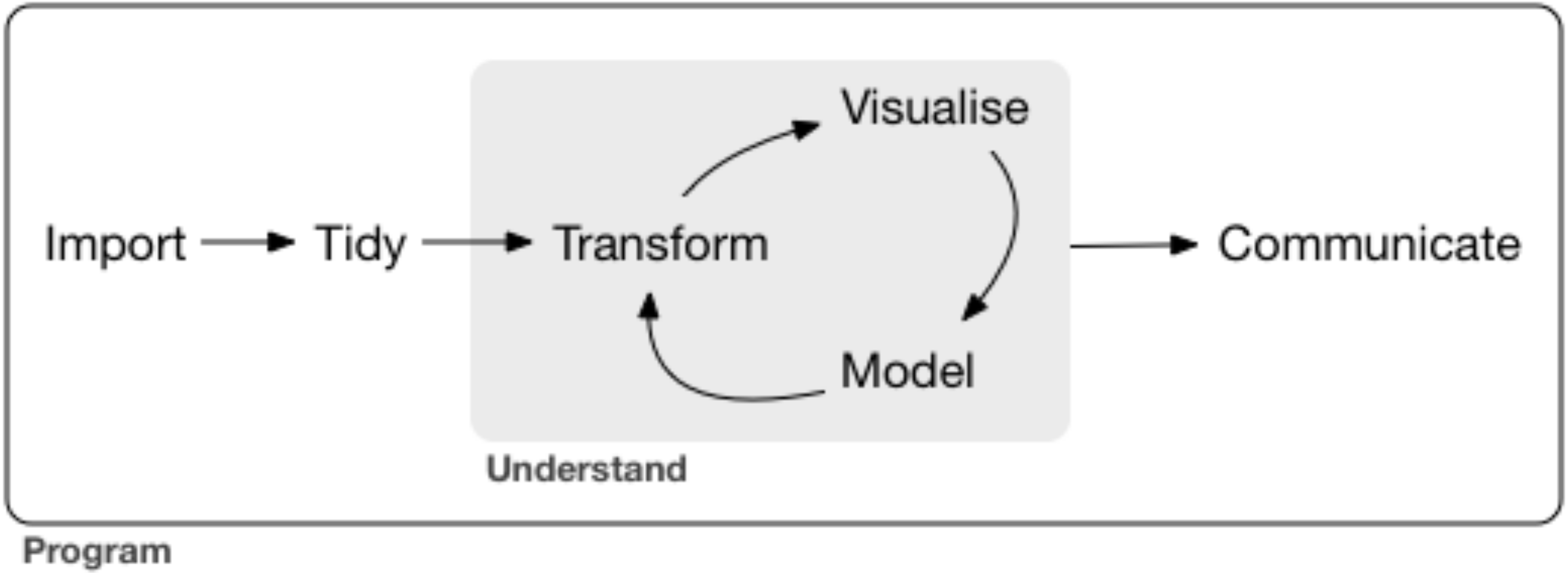
Introduction

Today you will learn:

- How normal ML experiment is conducted
- What are the steps in ML
- What you need to take care of and what you need to be cautious about.

Introduction

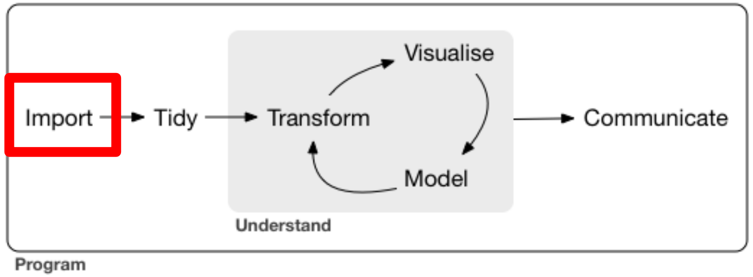
Machine Learning cycle (aka. Data Science process)



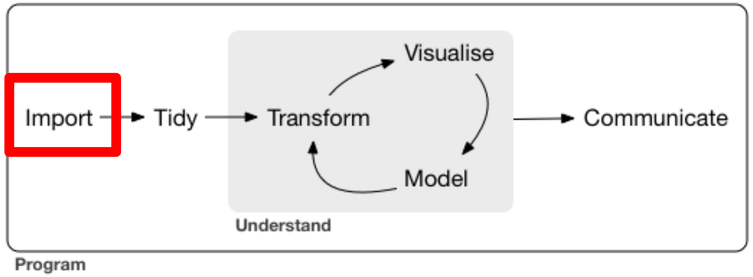
Import

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



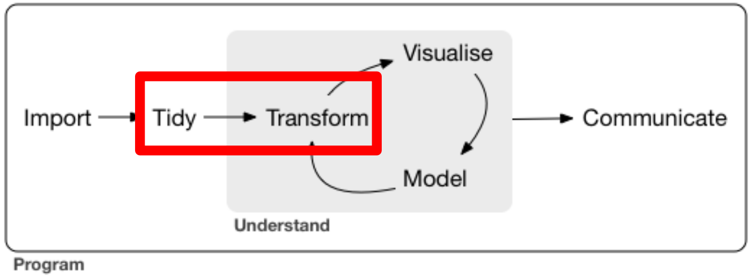
Import



Data sources types:

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

Wrangle

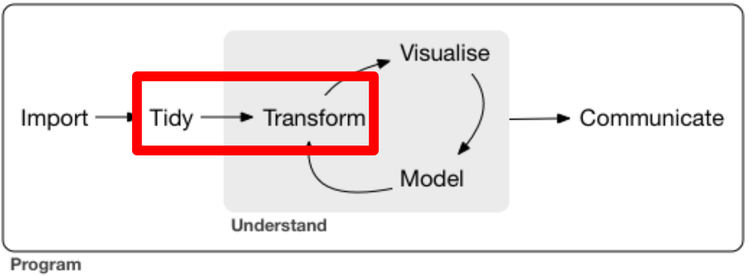


Tidy + Transform = Wrangle

↓

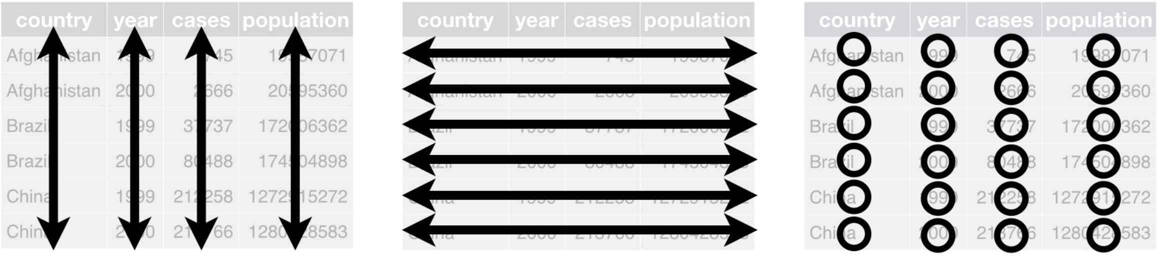
Prepare data for Visualisation and Modeling

Tidy



Tidy dataset:

- Each column represents one varia
- Each row represents one observa
- Each cell represent one value



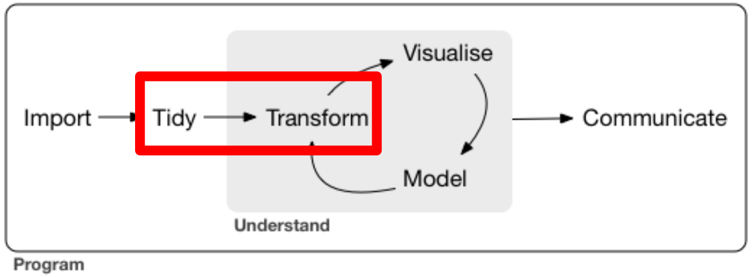
Problems:

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

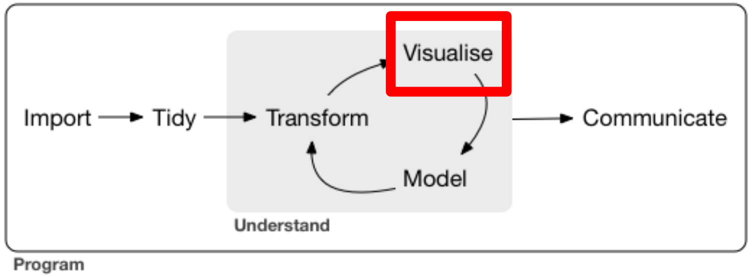
https://r4ds.had.co.nz/tidy-data.html

Transform

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

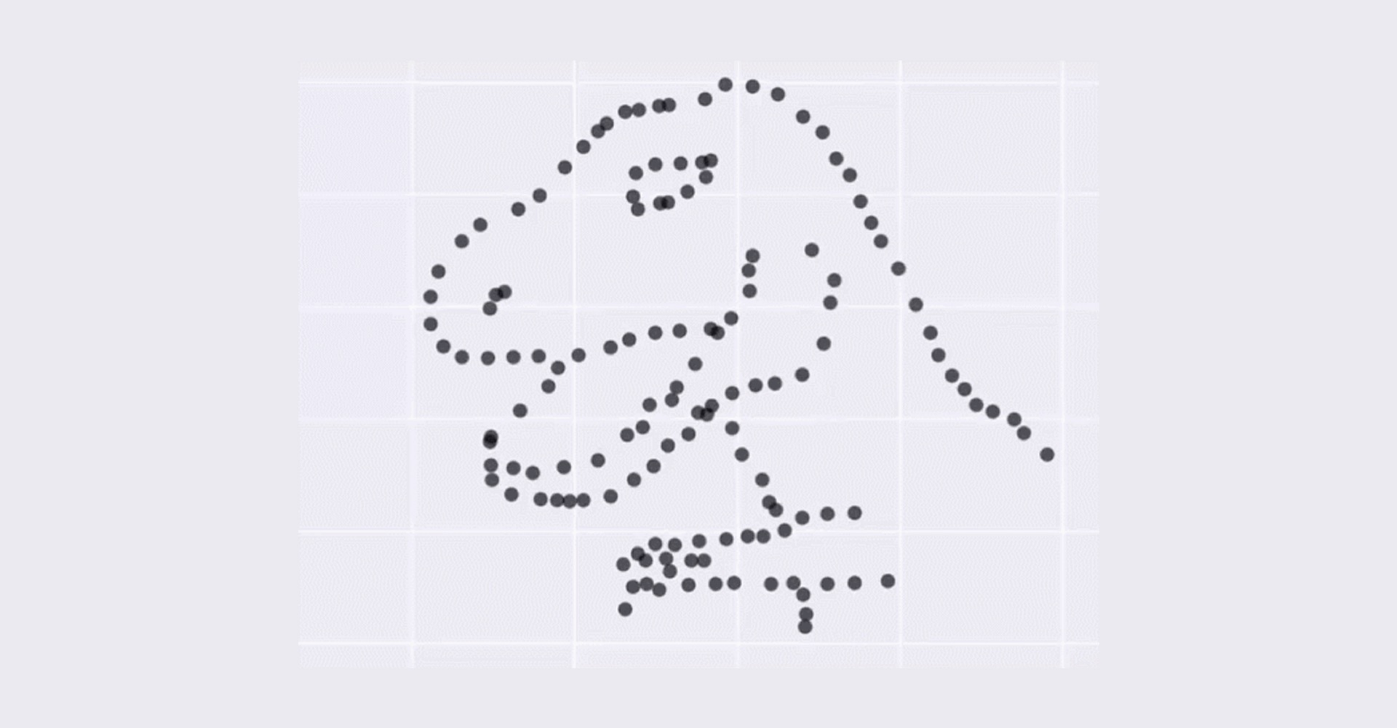


Visualize



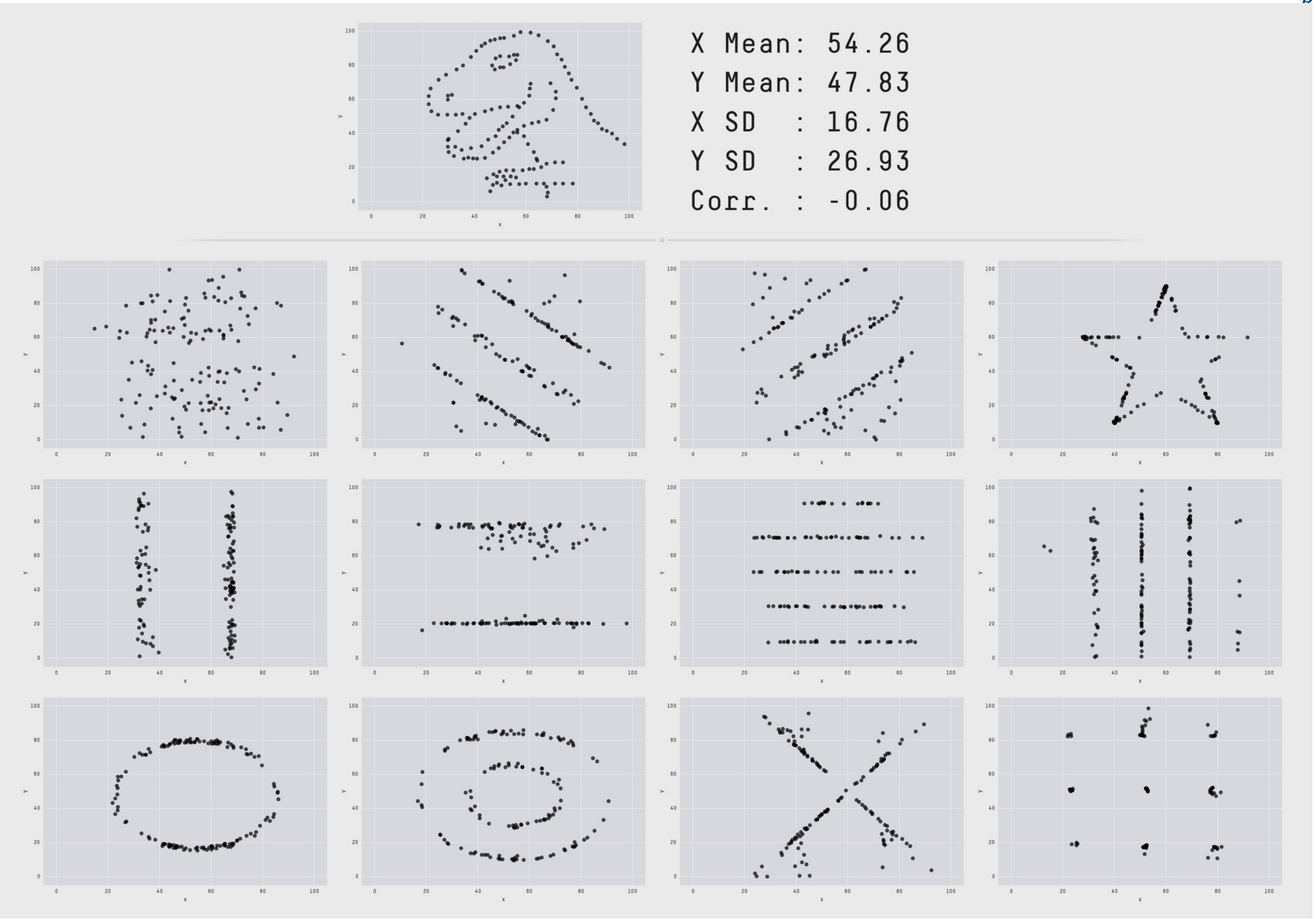
- Visualization is useful tool for providing the information to stakeholder but also during wrangling the data
- Helps to understand issues in the data
- Uncovers outliers
- Helps to identify relationships between variables

Datasaurus



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

Datasaurus



X Mean: 54.26
Y Mean: 47.83
X SD : 16.76
Y SD : 26.93
Corr. : -0.06

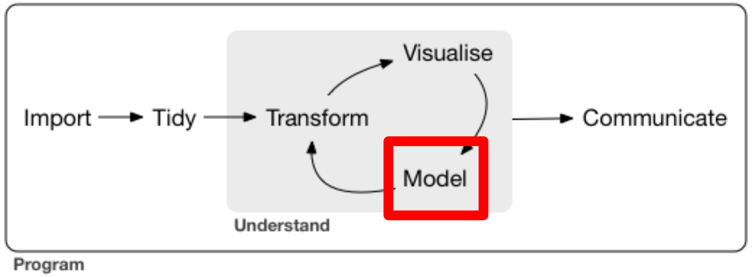
https://www.autodesk.com/research/publications/same-stats-different-graphs

Visualize

Do not trust your data blindly

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

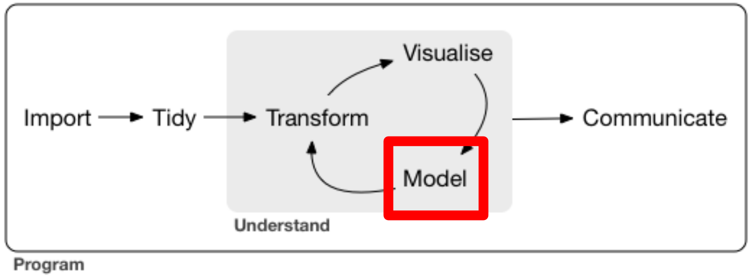
Model



What is Machine Learning?

What types of Machine Learning you know?

Model



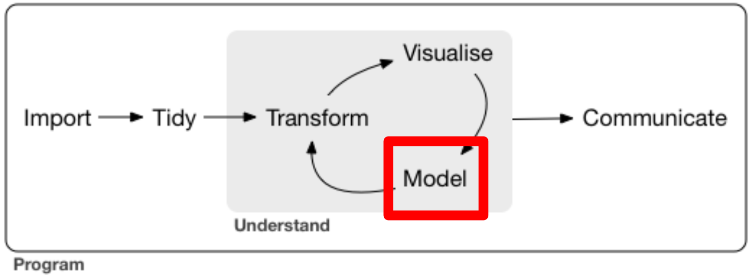
Machine Learning is:

"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*

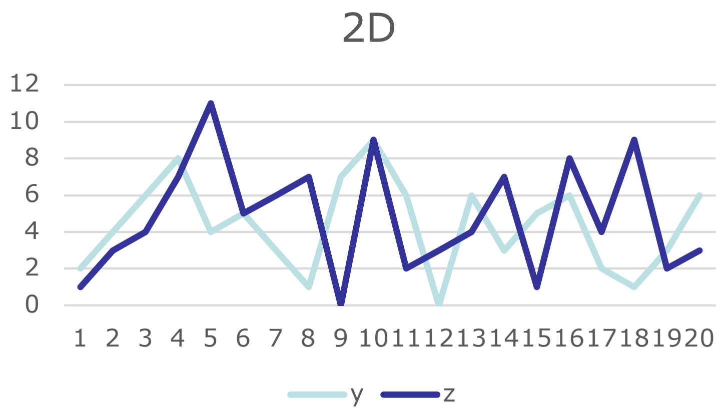
Model

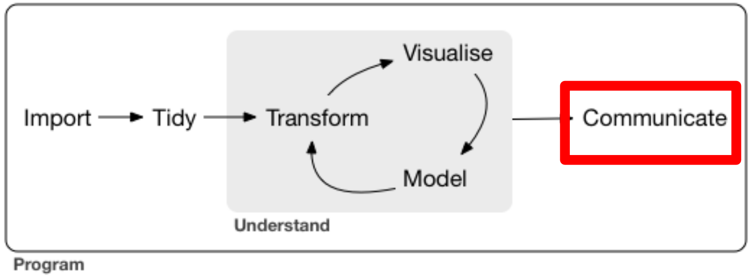


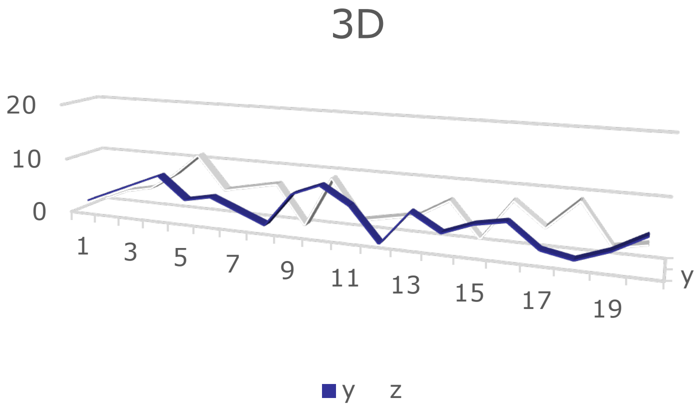
ML task categories based on “feedback” available to learning system:

- Supervised learning
  - We know the right answers
- Unsupervised learning
  - We do not know right answers
- Reinforcement Learning
  - Machine interacts with dynamic environment in which it needs to achieve certain goal without teacher telling it if it is close to the goal or not.

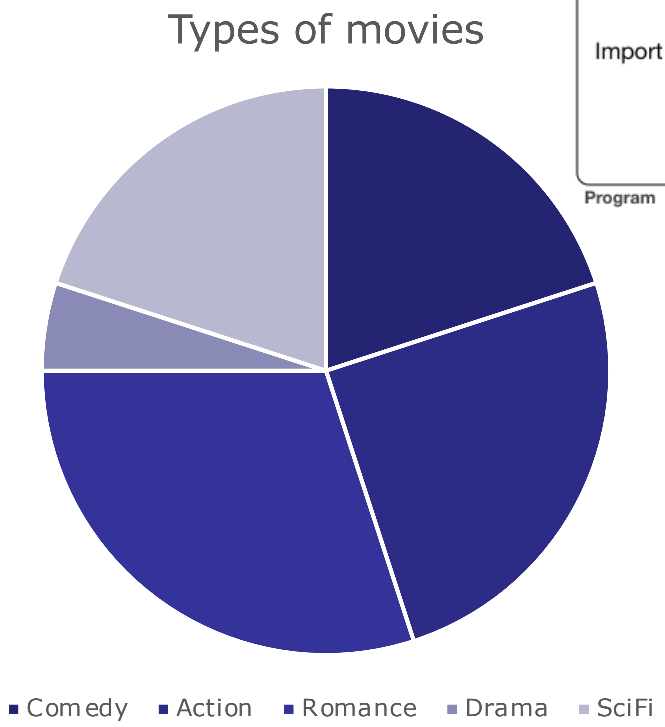
Communicate: No 3D graphs

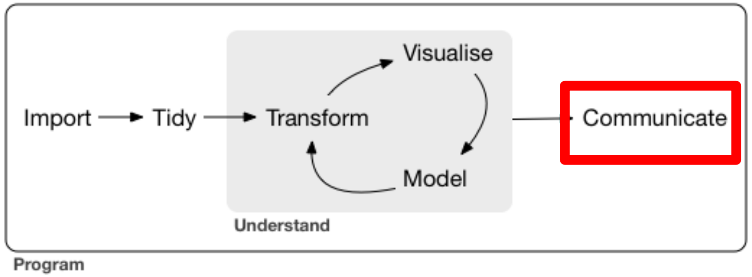






Communicate: No pie charts




Import → Tidy → Transform

Visualise

Model

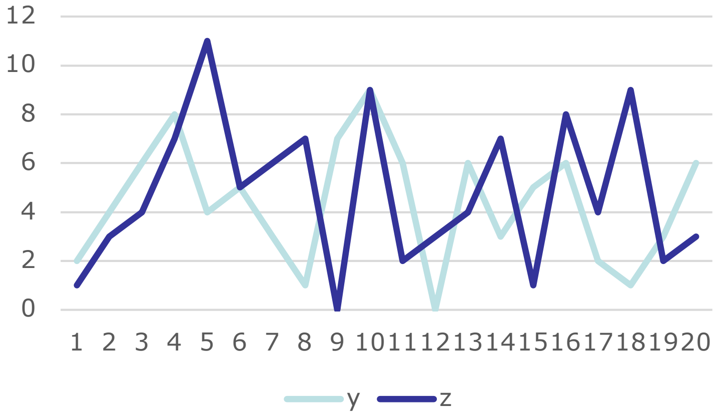
Understand

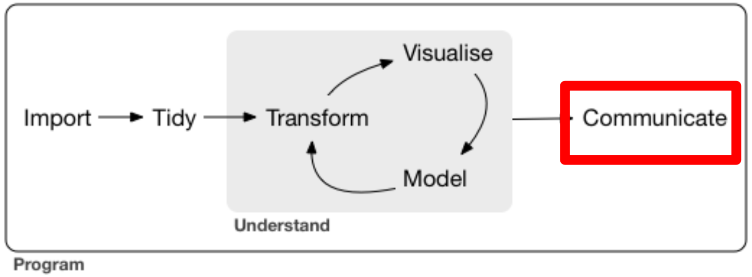
Communicate

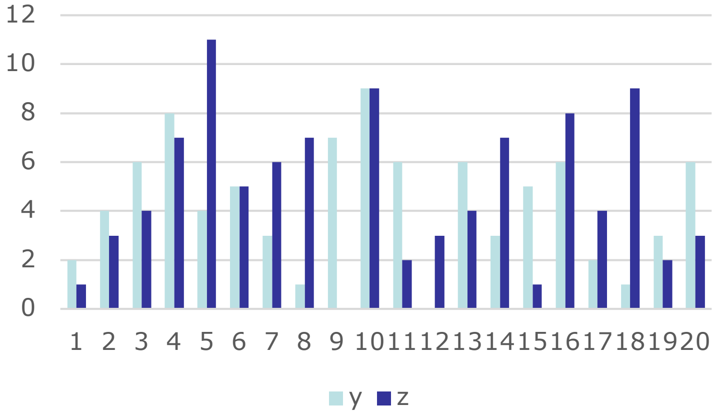
Program

Pre-attentive characteristics does not help with showing exact quantitative differences

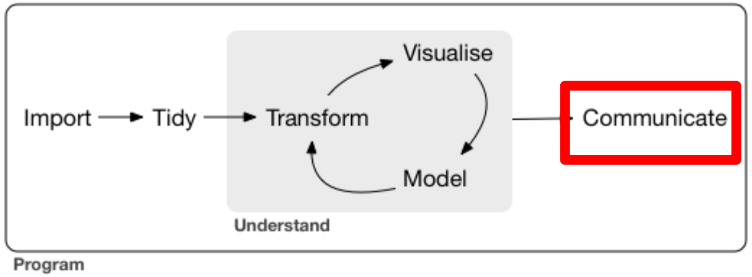
Communicate: Use common sense







Communicate: Don’t misled the users



Import

Tidy

Transform

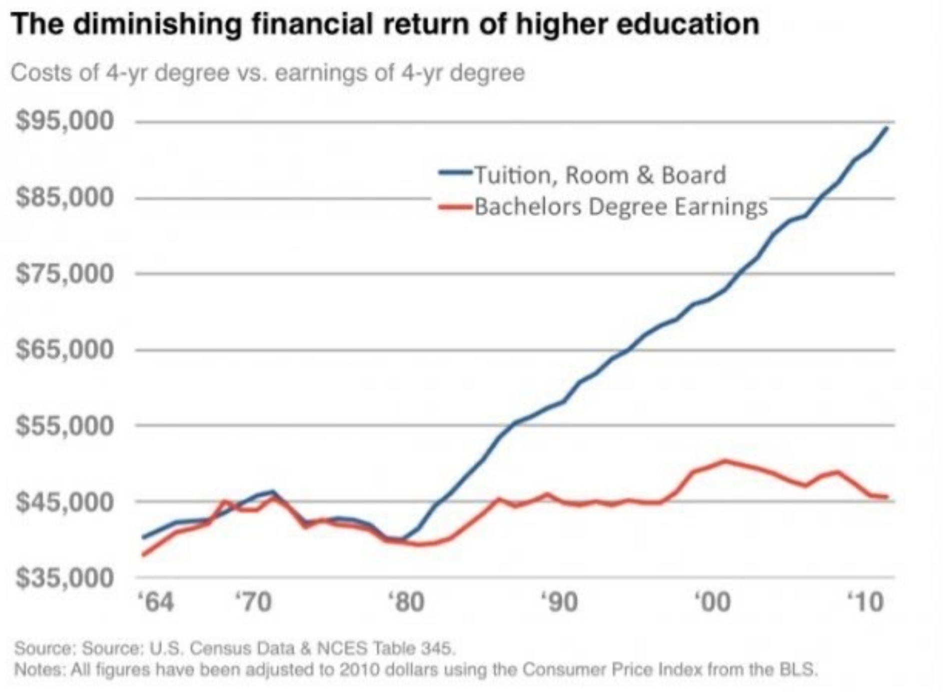
Visualise

Model

Understand

Communicate

Program

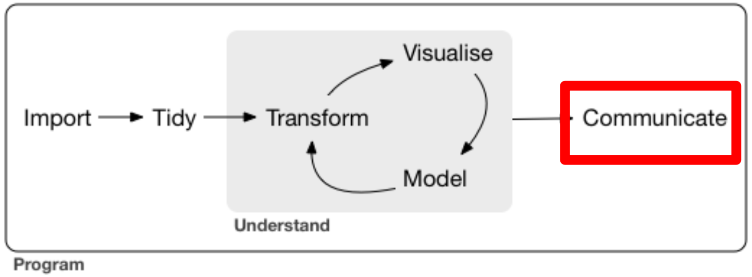


The diminishing financial return of higher education

Costs of 4-yr degree vs. earnings of 4-yr degree

Source: Source: U.S. Census Data & NCES Table 345.
Notes: All figures have been adjusted to 2010 dollars using the Consumer Price Index from the BLS.

Communicate: Do not lie



Import → Tidy → Transform

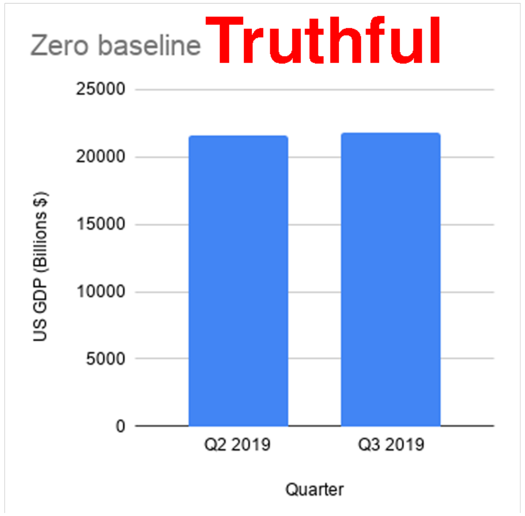
Visualise

Model

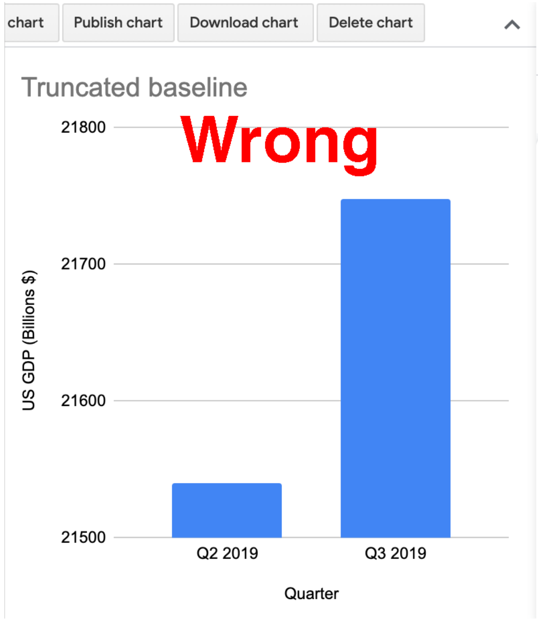
Understand

Communicate

Program



Truthful

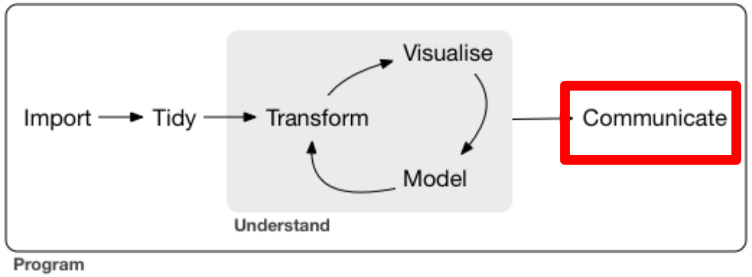


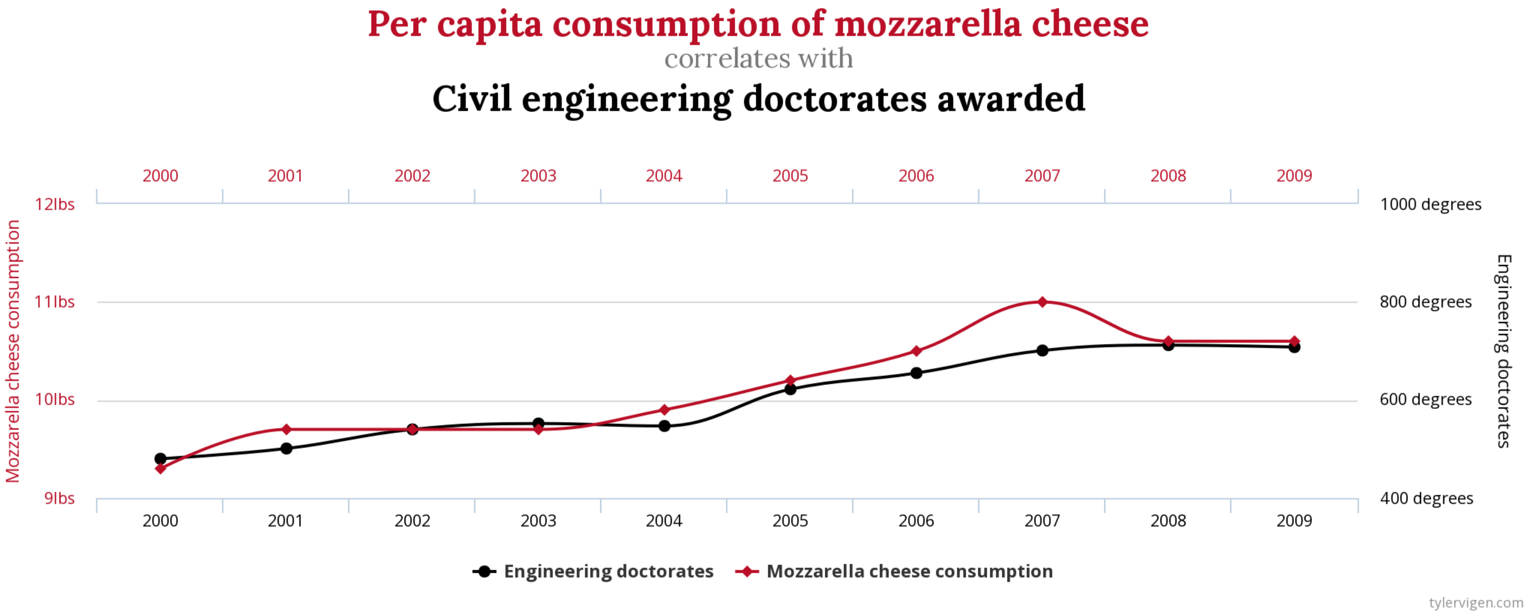
Wrong

https://handsondataviz.org/how-to-lie-with-charts.html#fig:gdp-baseline-merged

Communicate: Correlation is not causation

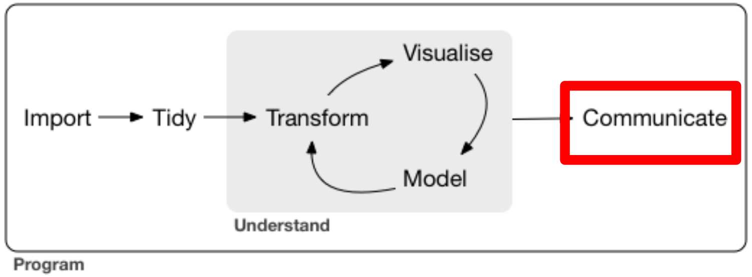
Correlation is not causation





http://www.tylervigen.com/spurious-correlations

Communicate



Visualise + 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

Questions?