Introduction to Data Wrangling

BE Chapman, PhD

2023-08-02

Housekeeping

- MIMIC access
  - Have you finished your training?
- Groups
  - Have you submitted your signed group contract?
  - Please let us know of any issues/difficulties connecting with peers
- Student representatives
  - Clemence Mottez (cmmottez@student.unimelb.edu.au)
  - Tanvesh Takawale (ttakawale@student.unimelb.edu.au)

Generating research-ready clinical data

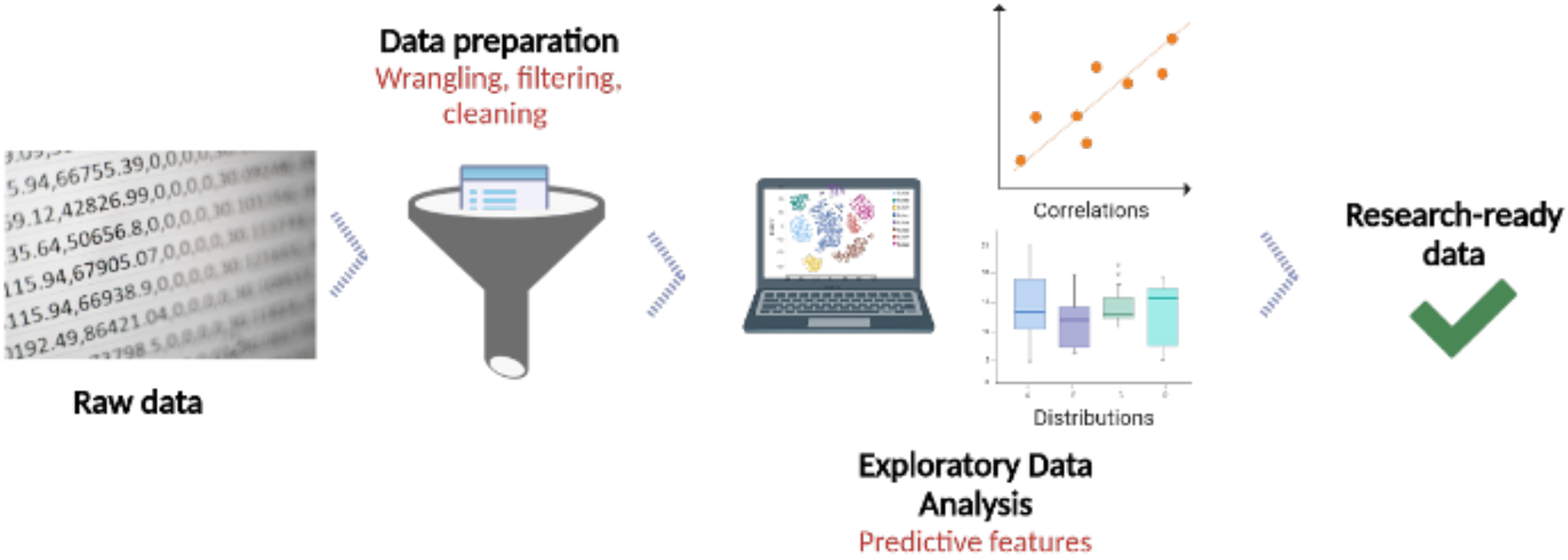


Figure 1: wrangling pipeline

Data Wrangling: a Simple Example

- What are all the "data" issues you might imagine with this tab delimited file?

```
Rank>---State>--Highest elevation>--Lowest elevation>---Average elevation$
1>--Colorado>---"14,440 feet">--"3,315 feet">---"6,800 feet"$
2>--Wyoming>"13,804 feet">--"3,099 feet">---"6,700 feet"$
3>--Utah>---"13,528 feet">--"2,000 feet">---"6,100 feet"$
4>--New Mexico>-"13,161 feet">--"2,842 feet">---"5,700 feet"$
5>--Nevada>-"13,140 feet">--479 feet>---"5,500 feet"$
6>--Idaho>--"12,662 feet">--710 feet>---"5,000 feet"$
7>--Arizona>"12,633 feet">--70 feet>"4,100 feet"$
8>--Montana>"12,799 feet">--"1,800 feet">---"3,400 feet"$
9>--Oregon>-"11,239 feet">--Sea level>--"3,300 feet"$
10>-Hawaii>-"13,796 feet">--Sea level>--"3,030 feet"$
11>-California>-"14,494 feet">---282 feet>--"2,900 feet"$
12>-Nebraska>---"5,424 feet">---840 feet>---"2,600 feet"$
13>-South Dakota>---"7,242 feet">---966 feet>---"2,200 feet"$
```

Figure 2: elevation

Data Wrangling: Another Example

- About how long is $3,175$ mm?

"What is the point?"

According to the ISO, the decimal sign is written eitehr as a comma or a point (a period), but in English the decimal sign is usually, although not always, written as a point. . . .The 22nd Conference Generale des Poids et Measures in 2003 repeated that "the symbol for the decimal marker shall be either the point on the line or the comma on the line." Further, it reaffirmed that when numbers are divided in groups of three in order to facilitate reading, "neither dots nor commas are ever inserted in the spaces between groups." (Grimvall 2011)

Data wrangling demo

https://github.com/chapmanbe/data_wrangling_demo

Generating research-ready clinical data

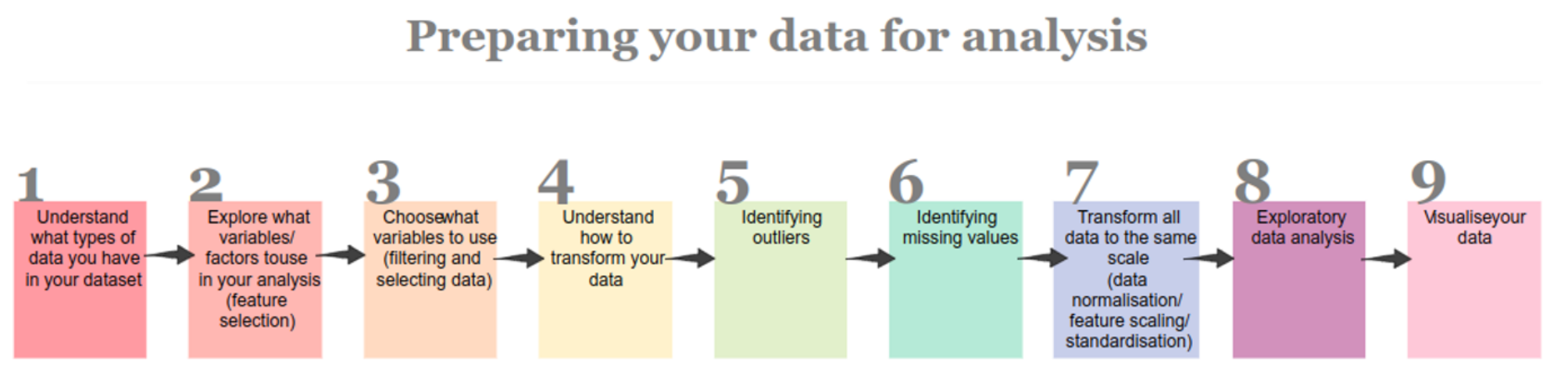


Figure 3: data prep

Data quality: common issues

When poll is active respond at **PollEv.com/brianchapman270**

Send **brianchapman270** to **22333**

Figure 4: q1

Data quality: Lack of standardisation

- "In talking with our clinical experts, we learned that normal blood glucose levels are between 3.9 and 5.5 and a value above 7 would indicate a diagnosis of diabetes."

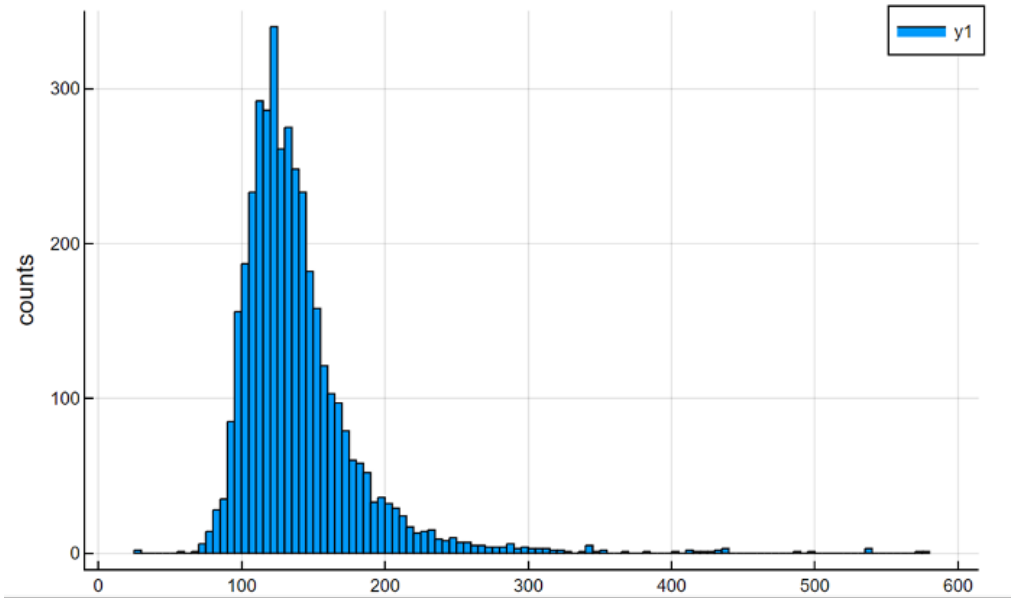


Figure 5: elevation

Data quality: Lack of standardisation

- Between 3.9 and 5.5 what?

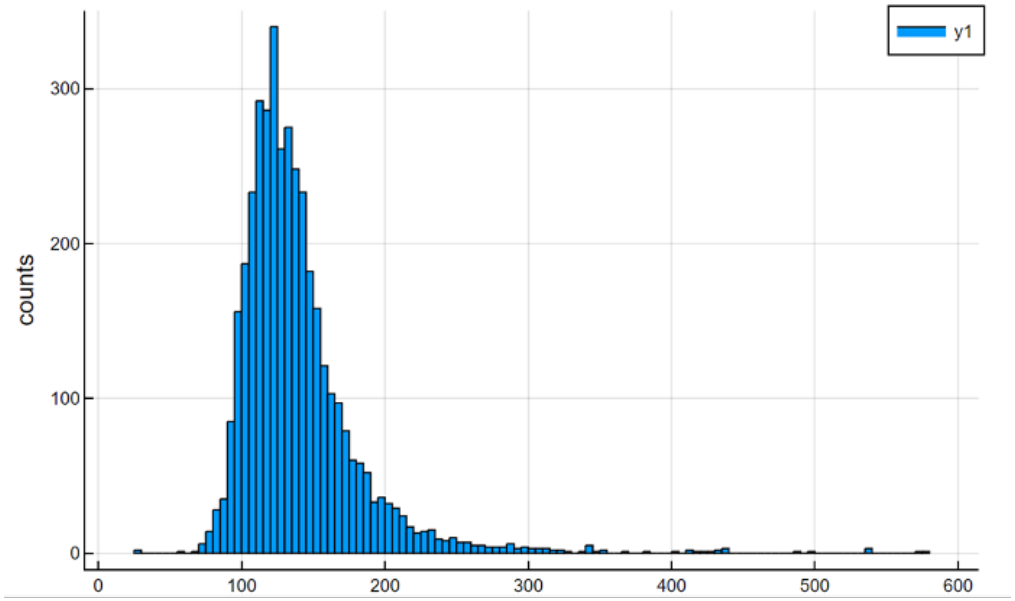


Figure 6: elevation

Data quality: Lack of standardisation

- Between 3.9 and 5.5 what?
- mmol/L

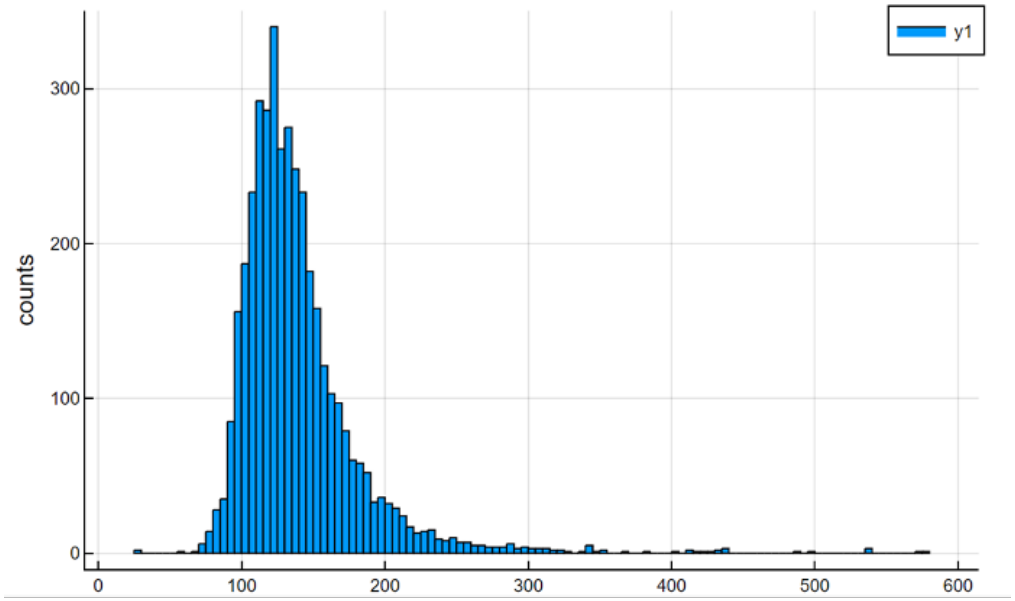


Figure 6: elevation

Data quality: Lack of standardisation

- Between 3.9 and 5.5 what?
- mmol/L
- MIMIC data units: mg/dL

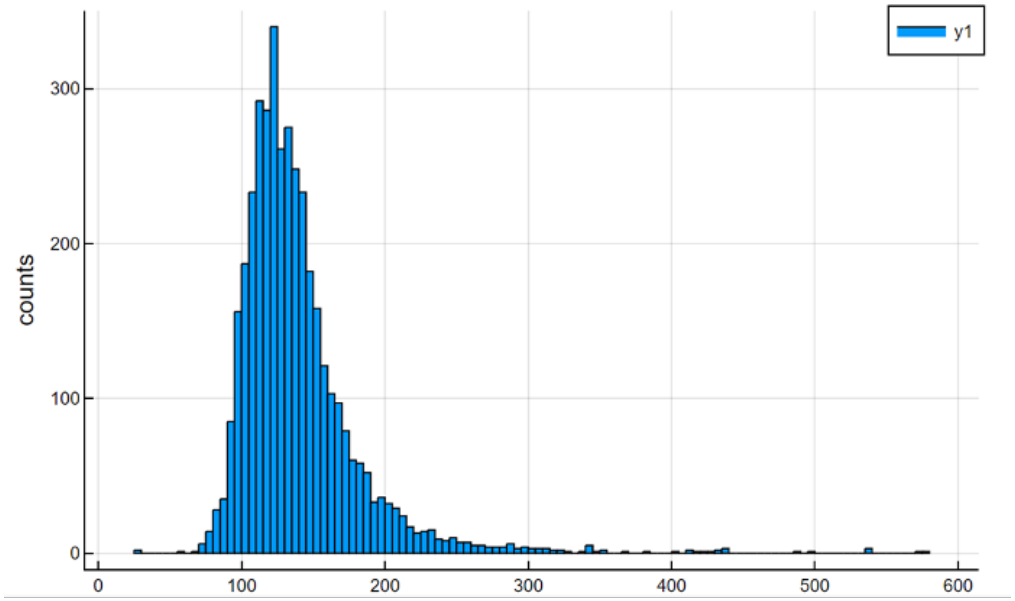


Figure 6: elevation

Data quality: Lack of standardisation

- Between 3.9 and 5.5 what?
- mmol/L
- MIMIC data units: mg/dL
- Convert!

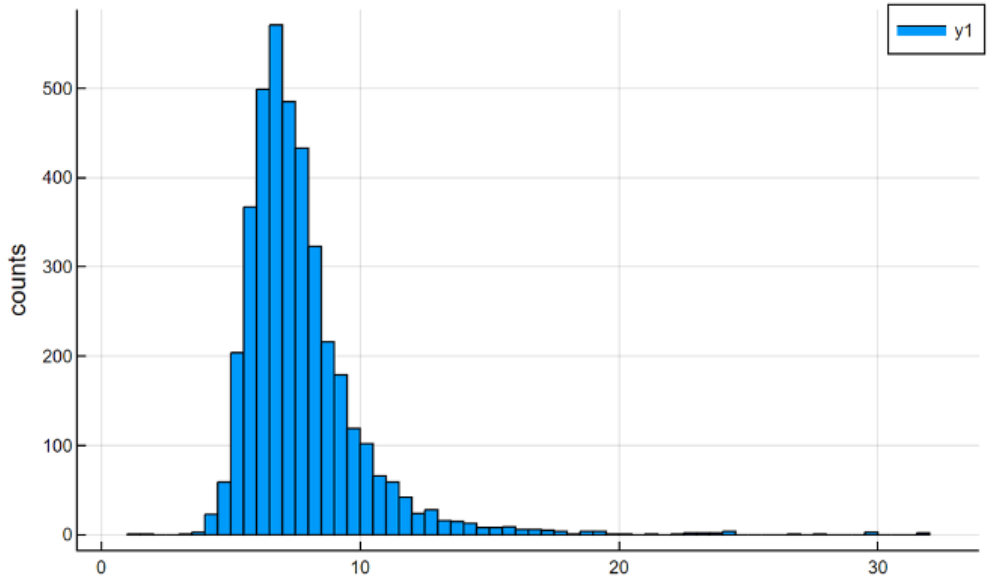


Figure 7: elevation

Does your machine learning algorithm care...

...whether it is mmol/L or mg/dL?

Data Wrangling: an Example[1]

2.11 Systolic Blood Pressure Histogram

```
select bucket, count(*) from (
  select width_bucket(value1num, 0, 300, 300) as bucket
    from mimic2v26.chartevents ce,
         mimic2v26.d_patients dp
   where itemid in (6, 51, 455, 6701)
     and ce.subject_id = dp.subject_id
     and months_between(ce.charttime, dp.dob)/12 > 15
  ) group by bucket order by bucket;
```

!

[1] *MIMIC II SQL Cookbook,* Daniel J. Scott and Ikaro Silva

Data quality: outliers and errors

- What is an outlier?
- An observation that diverges or is distant from an overall pattern of samples/observations
- Data points lying outside the overall distribution of a data set

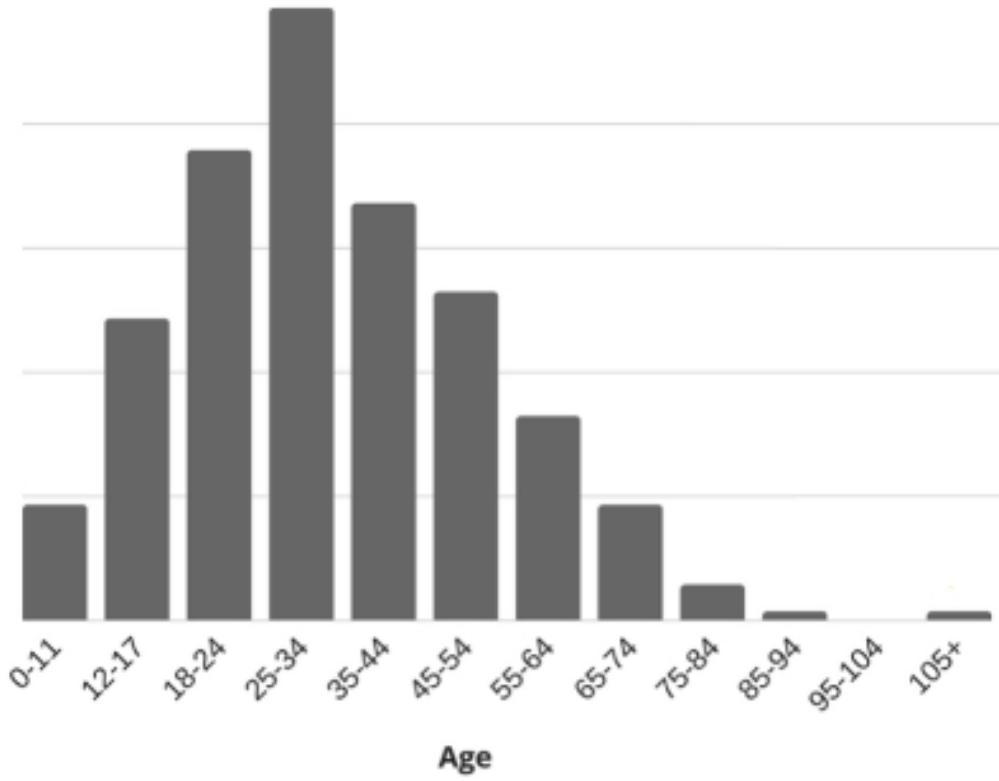
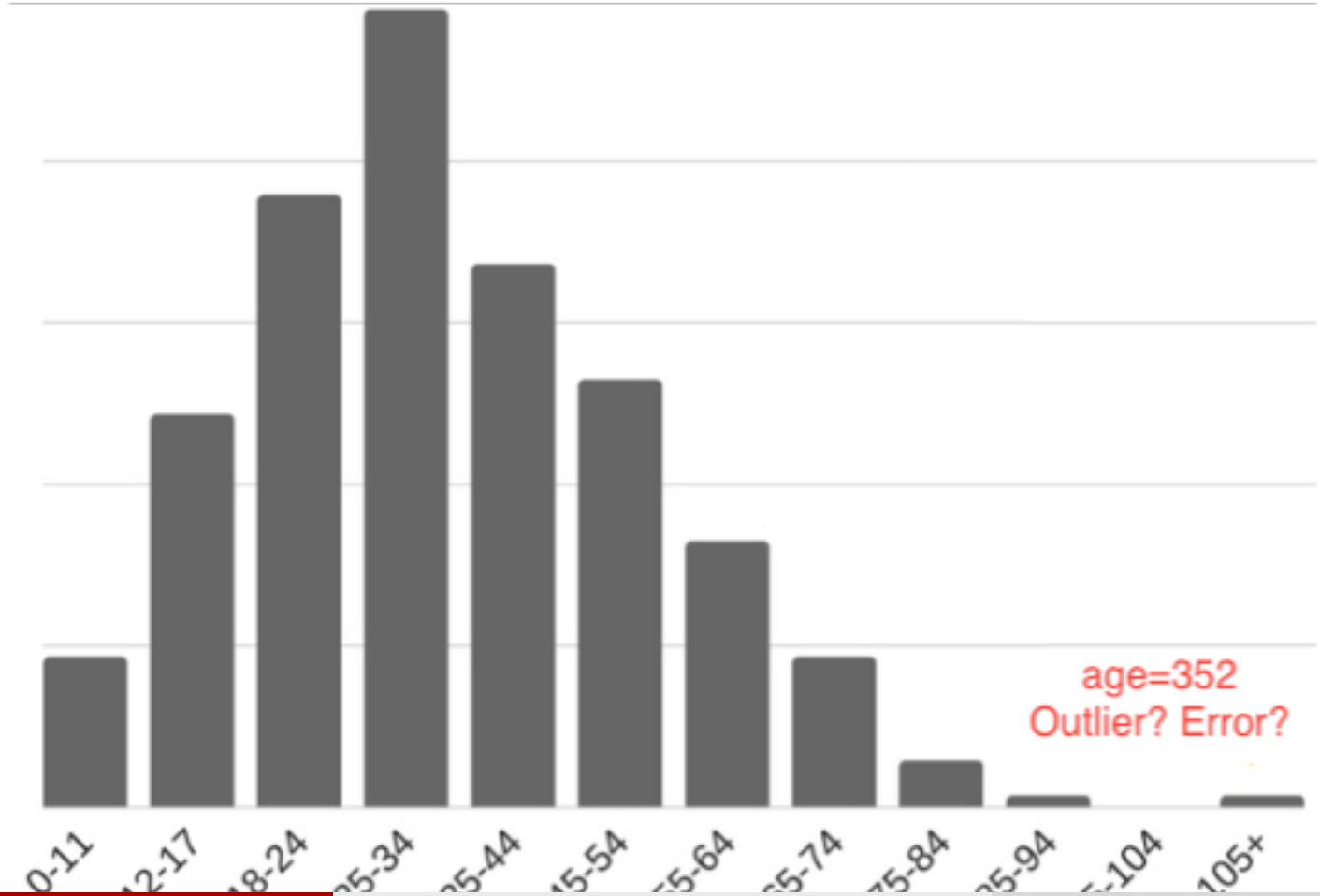
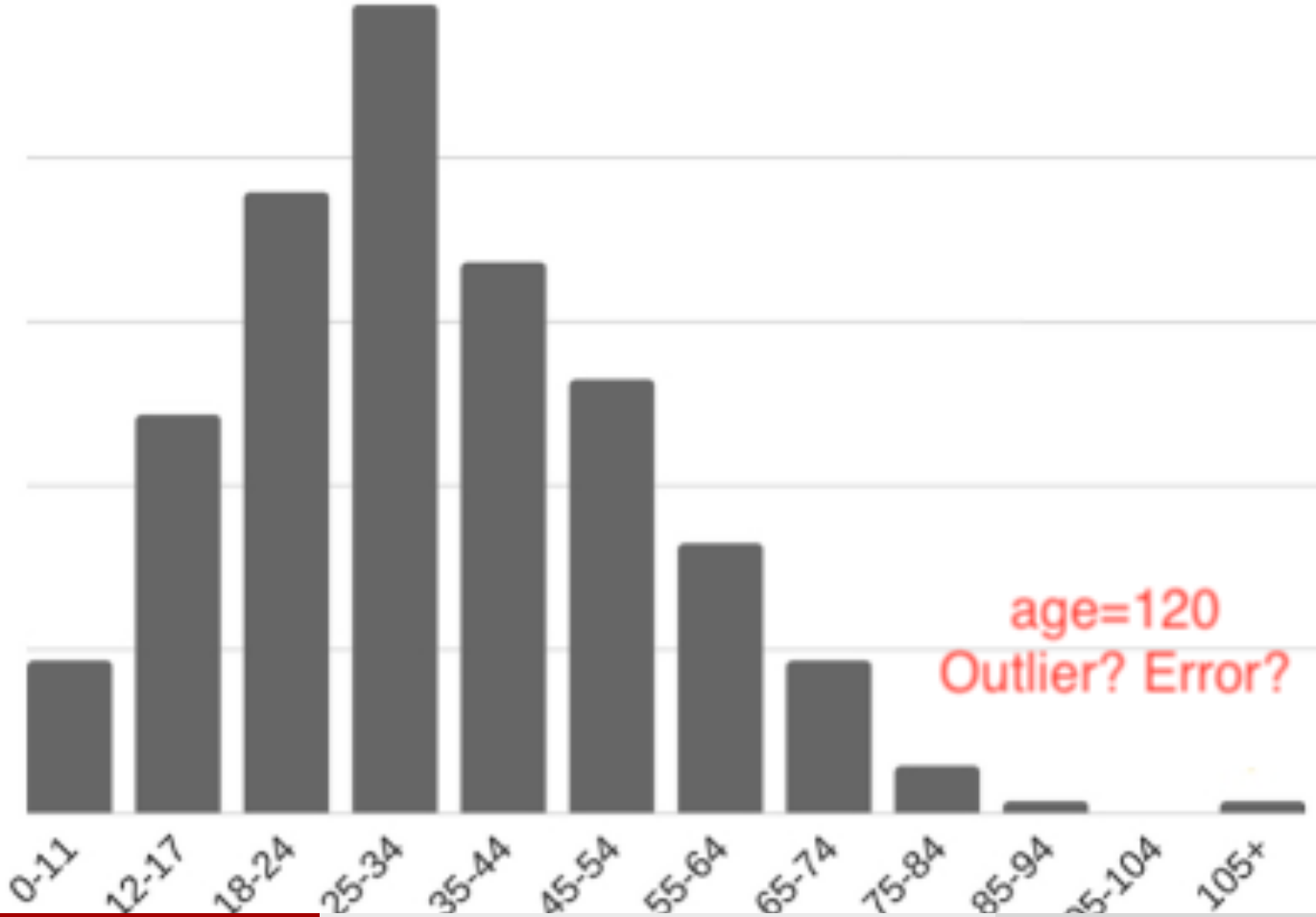


Figure 8: example distribution

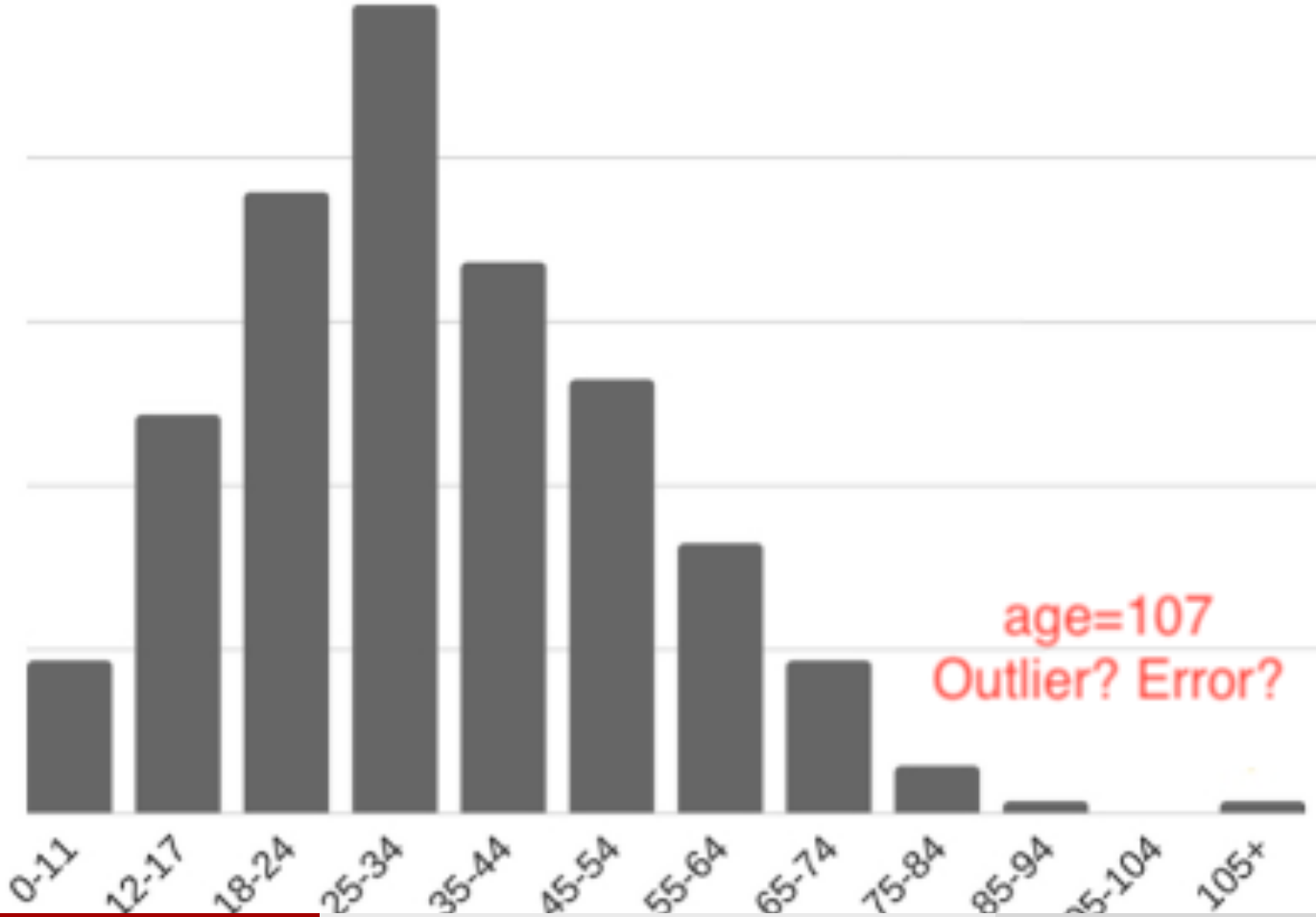
Data quality: outliers and errors



Data quality: outliers and errors



Data quality: outliers and errors



Outliers and implications for statistics & ML

- Girls under 16 basketball teams 2019

Outliers and implications for statistics & ML

- Girls under 16 basketball teams 2019
- Is the photo a mistake?

Outliers and implications for statistics & ML

- Girls under 16 basketball teams 2019
- Is the photo a mistake?
  - Red/Blue: USA

Outliers and implications for statistics & ML

- Girls under 16 basketball teams 2019
- Is the photo a mistake?
  - Red/Blue: USA
  - Blue/White: El Salvador

Outliers and implications for statistics & ML

- Girls under 16 basketball teams 2019
- Is the photo a mistake?
  - Red/Blue: USA
  - Blue/White: El Salvador
  - USA won 114-19

Outliers and implications for statistics & ML

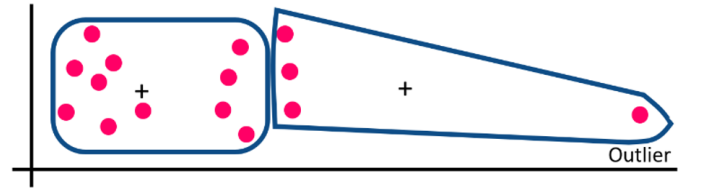


Figure 12: Outlier example 1a

Figure 13: Outlier example 1b

Where do outliers come from?

- Data entry errors (human errors)
- Measurement errors (instrumental errors)
- Experimental errors (data extraction or experiment planning/executing errors)
- Intentional (dummy outliers made to test detection methods)
- Data processing errors (due to data manipulation or unintended mutations in data)
- Sampling errors (extracting or mixing data from wrong or various sources)
- Natural (not an error at all, but rather **novelties** in the data)

How to detect outliers? (non-exhaustive)

- Z-Score or Extreme Value Analysis (parametric)
- Probabilistic and Statistical Modeling (parametric)
- Linear Regression Models (prinicpal component analysis (PCA), least means squares (LMS))
- Proximity Based Models (non-parametric)
- Information Theory Models
- High Dimensional Outlier Detection Methods
- Visualization?
- **Domain knowledge?**

Why do I need to worry about outliers?

- They might reflect errors in our data that we need to correct.
- If they are novelty, they might influence downstream analysis that we carry out.
- Many different summary statistics (e.g., mean, standard deviation) are actually sensitive to outliers, meaning they will be heavily influenced by them.

What are missing values?

Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest. The problem of missing data is relatively common in almost all research and can have a significant effect on the conclusions that can be drawn from the data. Accordingly, some studies have focused on handling the missing data, problems caused by missing data, and the methods to avoid or minimize such in medical research. (Kang 2013)

Where do missing values come from?

Real-world, routinely collected clinical data often has lots of missing values and handling them adequately is a critical step for data cleaning. There can be multitude reasons why they occur:

- Human/data entry errors
- Optional fields/responses in surveys (e.g., patient weight might not be always recorded in all consultations)
- Incorrectly acquired data (e.g., from errors in sensor readings)
- Software bugs in data processing pipelines
- And many others!

Why do we need to handle them adequately?

- Absence of data reduces statistical power
  - Probability that the test will reject the null hypothesis when it is false
- Can cause bias in the estimation of parameters
- Can reduce the representativeness of the samples
- It may complicate the analysis of the study

Handling missing values

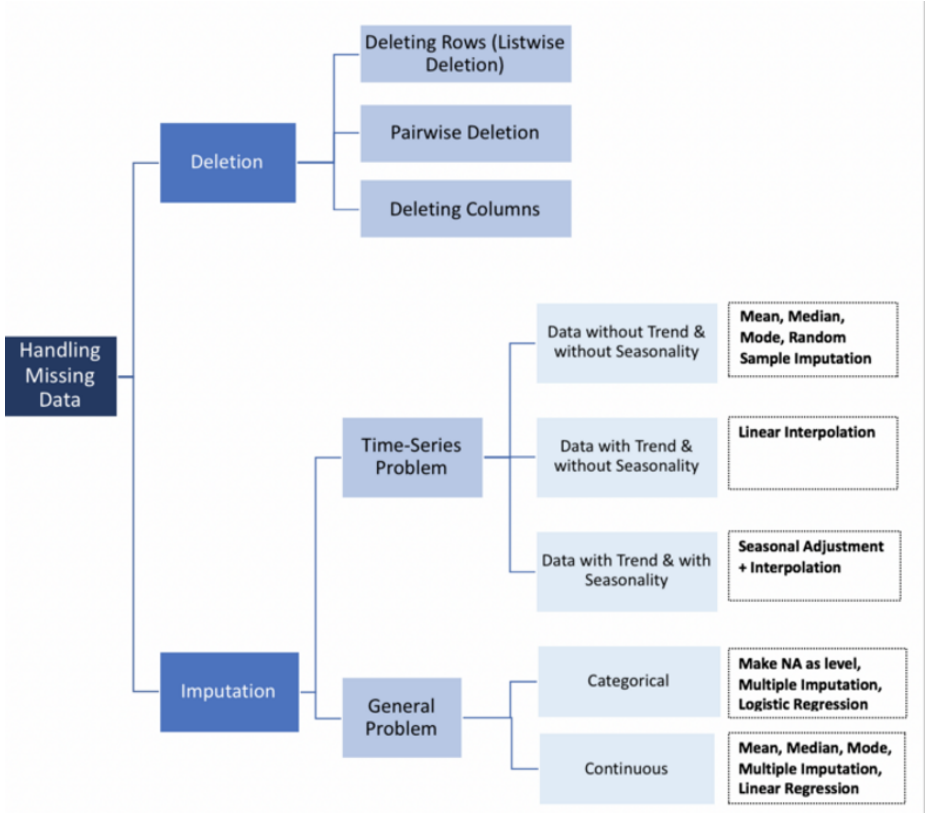


Figure 14: Missing values decision tree

Missing data vs. “Dark data”

Figure 15: David J. Hand

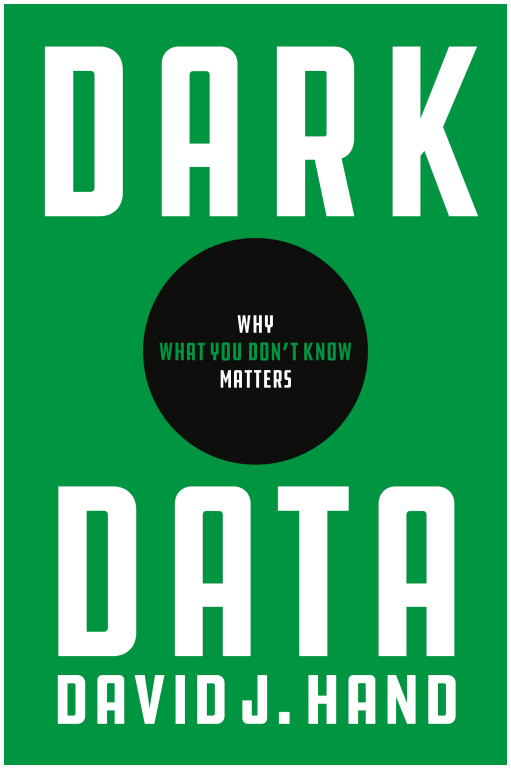


Figure 16: Data Data

Dark Data: A Taxonomy of Dark Data

- DD-Type 1: Data We Know are Missing
- DD-Type 2: Data We Don't Know are Missing
- DD-Type 3: Choosing Just Some Cases
- DD-Type 4: Self-Selection
- DD-Type 5: Missing What Matters
- DD-Type 6: Data Which Might Have Been
- DD-Type 7: Changes with Time
- DD-Type 8: Definitions of Data
- DD-Type 9: Summaries of Data
- DD-Type 10: Measurement Error and Uncertainty
- DD-Type 11: Feedback and Gaming
- DD-Type 12: Information Asymmetry
- DD-Type 13: Intentionally Darkened Data
- DD-Type 14: Fabricated and Synthetic Data
- DD-Type 15: Extrapolating beyond Your Data

(Hand 2020)

Dark Data

The map is not the territory!

Correlation vs. causation

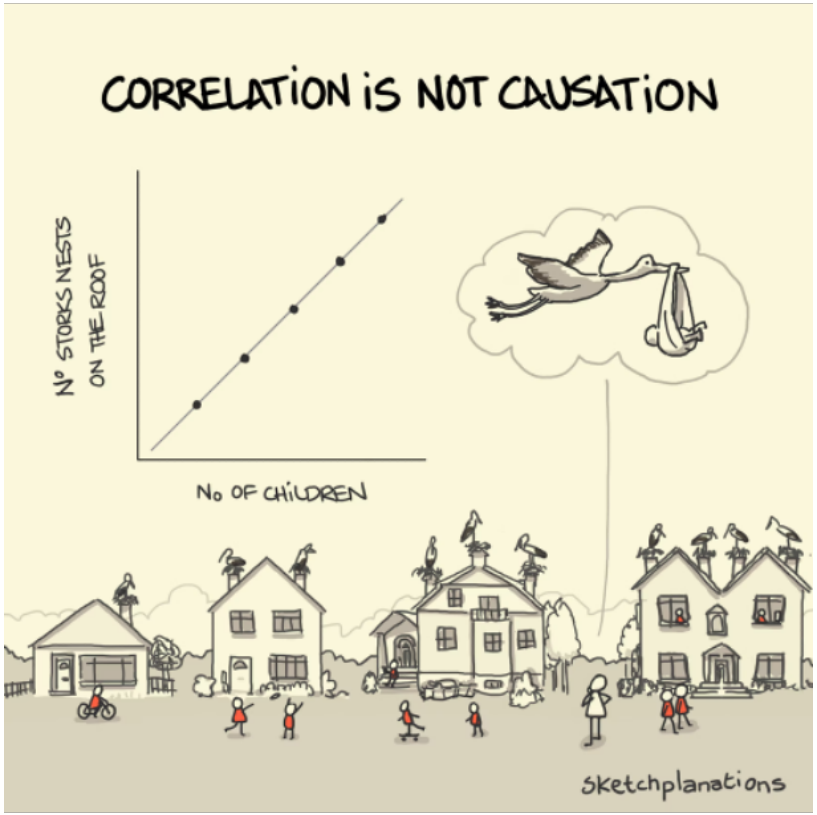


Figure 17: correlation vs. causation

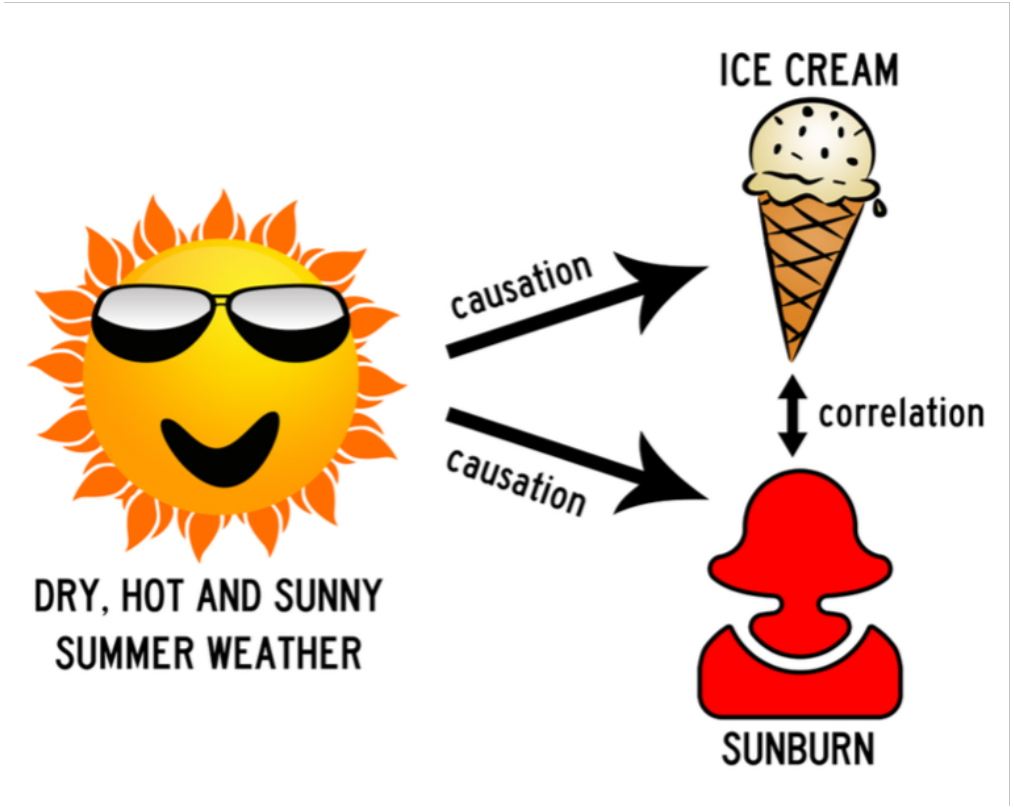


Figure 18: correlation vs. causation

exploratory Data Analysis (EDA)

- Task of analysing data using simple tools from statistics to simple plotting tools.
  - Discover patterns
  - Identify outliers and errors
  - Check assumptions
  - Identify promising variables for predictive modelling

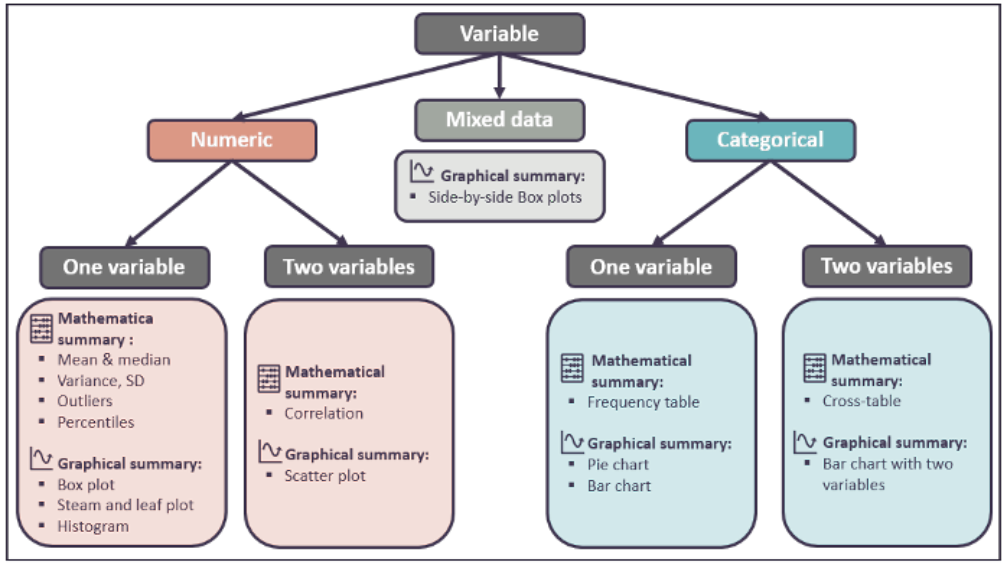


Figure 19: EDA

Exploratory Data Analysis (EDA)

- Before going from analytic to predictive
  - Treating ML as a black box can be dangerous
- Rational selection and investigation of potential features
  - Leads to more interpretable models
  - Less complex (minimal set of features) and more generalisable predictive models (Occam's Razor principle)

Figure 20: https://xkcd.com/1838/

Exploratory Data Analysis via Visualization

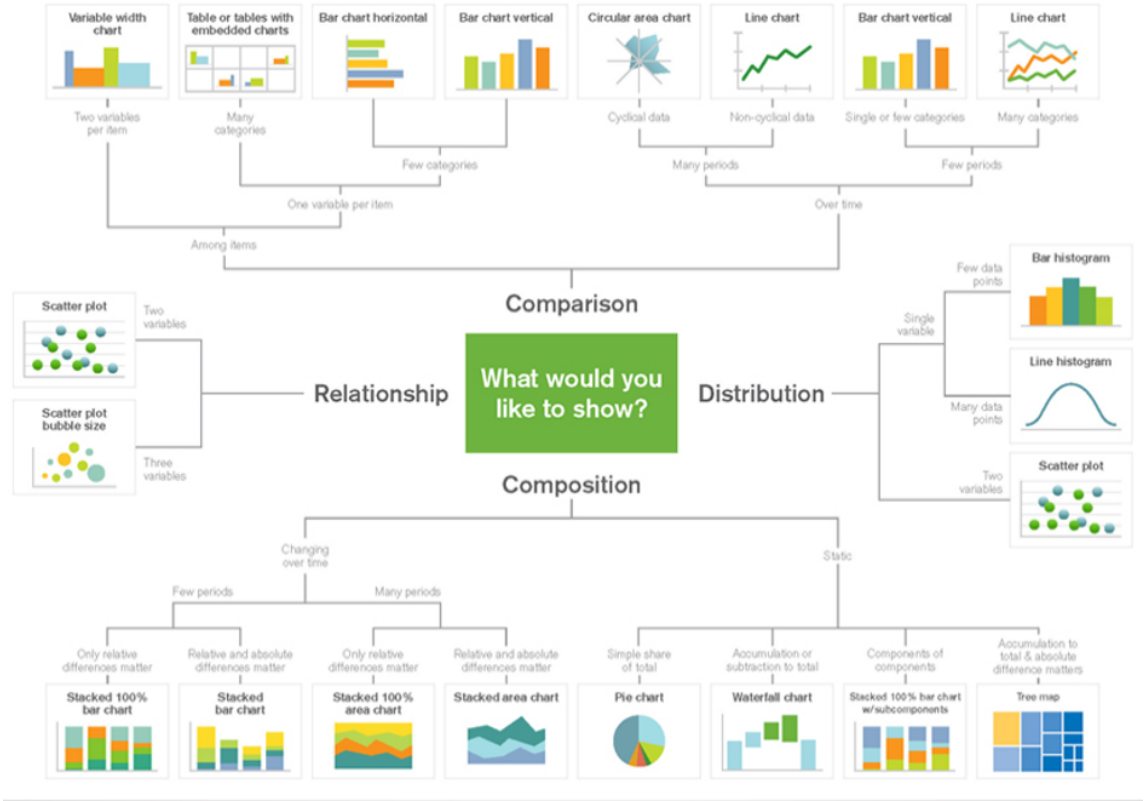


Figure 21: EDA visualization

Reproducibility

- Data wrangling has lots of steps, lots of humanness. How do we make it reproducible?
- This is a hard problem!

Reproducibility: Specifying Anaconda environments

`environment.yml` files

```yaml
name: dwd
channels:
  - pyviz
  - conda-forge
dependencies:
  - python=3.11
  - pandas=2.0.3
  - hvplot=0.8.4
  - jupyterlab=4.0.3
```

Reproducibility: Docker (Podman, Singularity)

```
FROM jupyter/base-notebook:5cb1a915c4bf
MAINTAINER chapmanbe <brian.chapman@utah.edu>
USER root

RUN apt-get update && apt-get upgrade -y && apt-get install -y \
      locales-all \
  && rm -rf /var/lib/apt/lists/*

RUN conda update conda -y && conda install -c conda-forge -c pyviz -y \
    pandas=2.0.3 \
    hvplot=0.8.4

WORKDIR /home/jovyan

COPY Resources/ Resources/
ADD Pandas-DataWrangling.ipynb .

# RUN nbstripout --install
CMD ["start-notebook.sh"]
```

Reproducibility

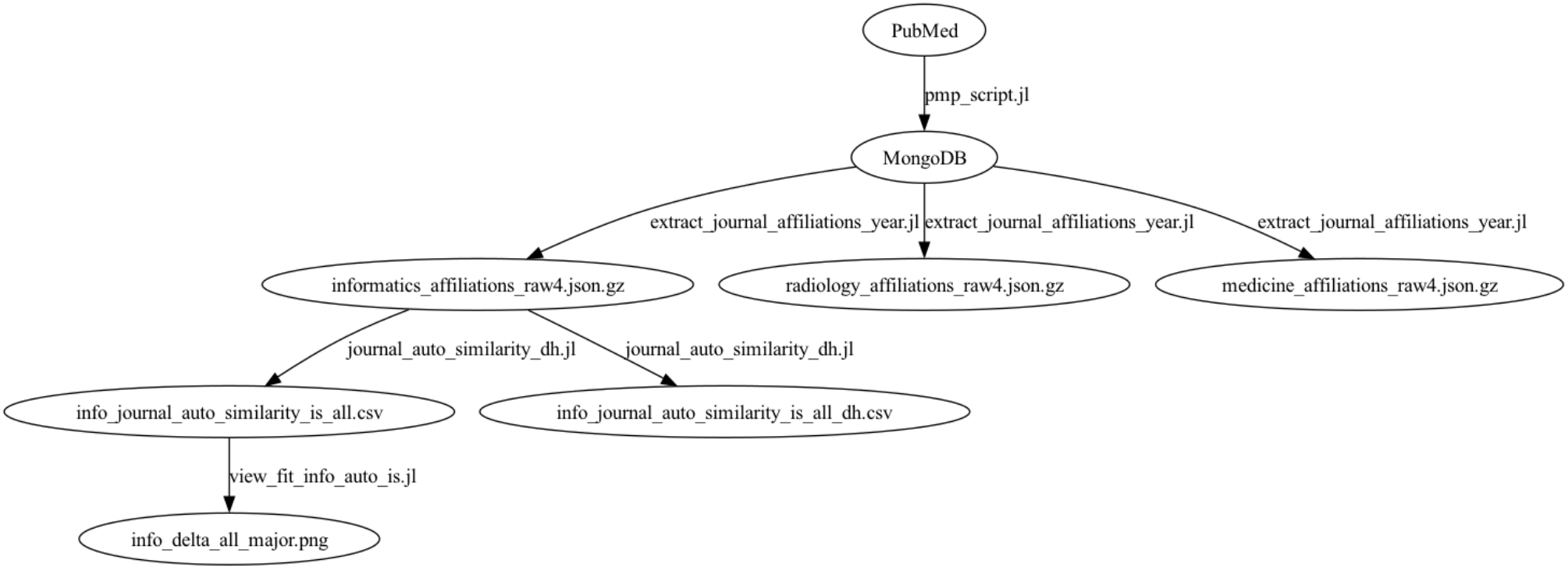


Figure 22: my data flow

Reproducibility

- Reproducible Data Science with Python
- Reproducible Data Analysis in Jupyter

References I

Grimvall, G. 2011. *Quantify!: A Crash Course in Smart Thinking*. Johns Hopkins University Press. https://books.google.com.au/books?id=5PqTZbgl-7QC.

Hand, D. J. 2020. *Dark Data: Why What You Don't Know Matters*. Princeton University Press. https://books.google.com.au/books?id=FL2mDwAAQBAJ.

Kang, H. 2013. "The prevention and handling of the missing data." *Korean J Anesthesiol* 64 (5): 402–6.