Data Handling

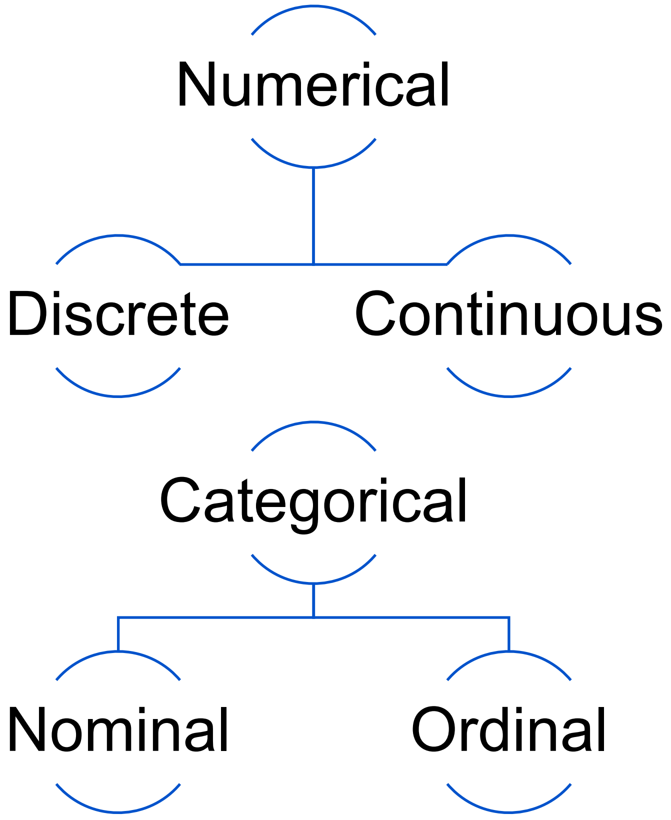
Data Handling

Common Tasks

- Identifying types of data
- Data cleaning
- Standardization
- Data Augmentation
- Visualization

Data Types

- **Numerical**: Numbers (duh)
  - Discrete or continuous
- **Ordinal**: Different states with a defined order
  - T-Shirt size: S < M < L
  - Low, medium, high
- **Nominal:** Multiple states without order
  - T-Shirt color
  - Gender


Numerical

Discrete

Continuous

Categorical

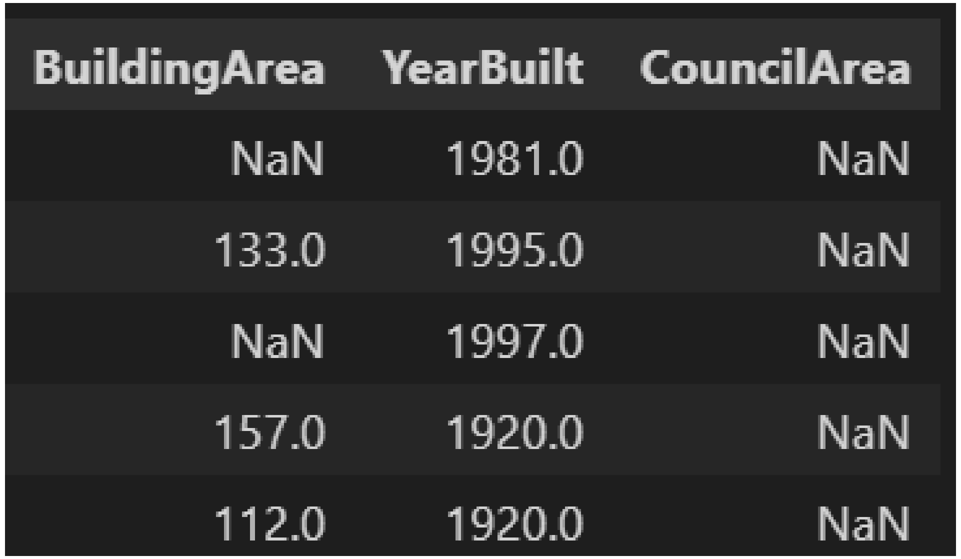
Nominal

Ordinal

Data cleaning

Missing values

- Strategies for handling missing values:
  - Ignore (☹)
  - Remove (losing statistical power)
  - Default values (e.g. 0)
  - Interpolate (e.g. mean, max)



BuildingArea	YearBuilt	CouncilArea
NaN	1981.0	NaN
133.0	1995.0	NaN
NaN	1997.0	NaN
157.0	1920.0	NaN
112.0	1920.0	NaN

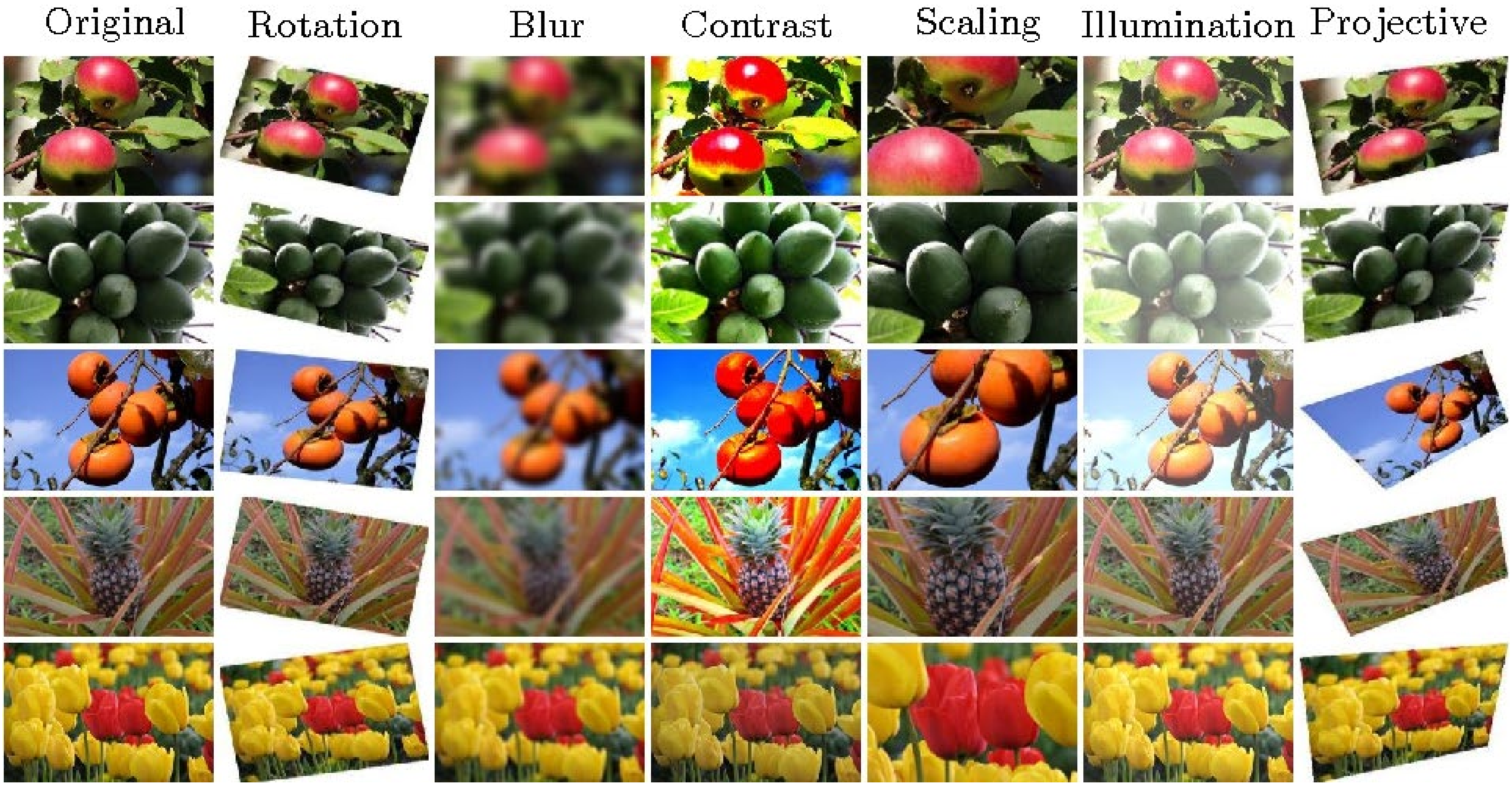
Standardization

- **Problem:**
  - Features with a large scale are interpreted as having more weight (e.g. grams vs. kg).
  - Features with a large variance are interpreted as more informative.
- **Idea**: Scale features to same mean and variance:

$$standard(x_i) = \frac{x_i - mean(x)}{stdev(x)}$$

- Implemented by: sklearn.preprocessing.StandardScaler

Data Augmentation



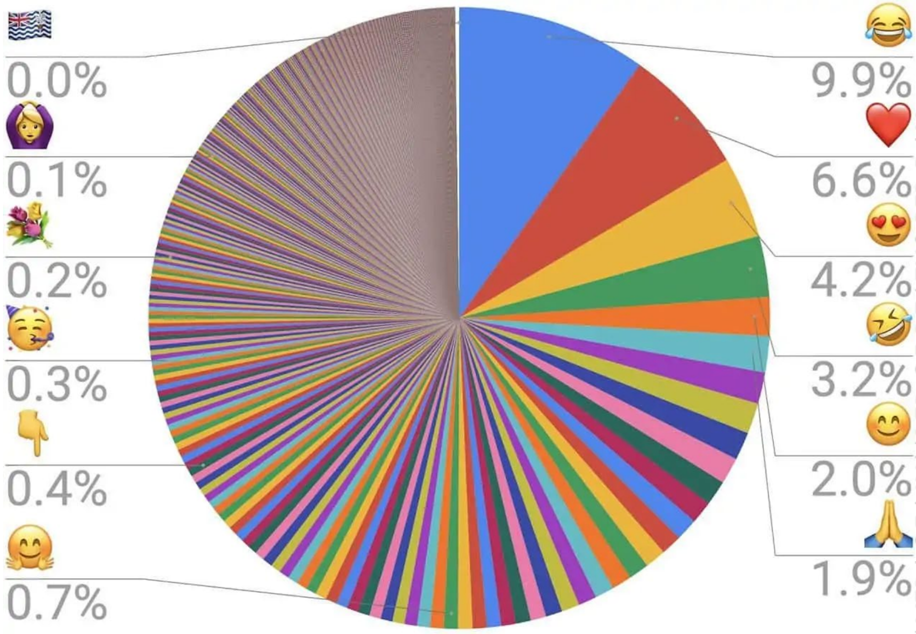
Data Augmentation

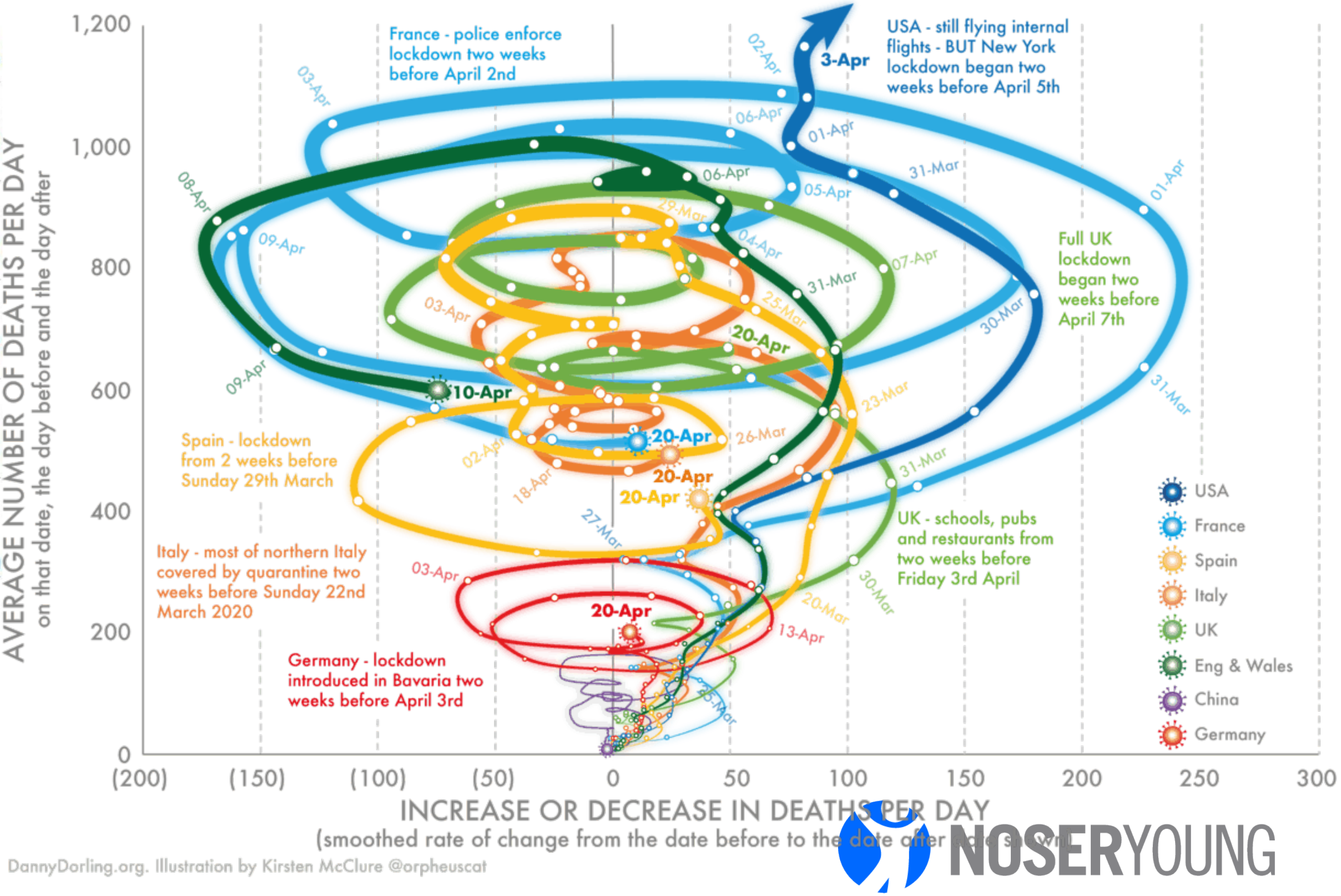
- **Idea**: **Modify data to augment the dataset**
- Used to make a training set more robust by introducing more variation in the Dataset
- **Improving model prediction accuracy** by increasing generalizability and overall increasing the size of the training dataset.

Data-Visualization

Data-Visualization

How NOT to do it





France - police enforce lockdown two weeks before April 2nd

USA - still flying internal flights - BUT New York lockdown began two weeks before April 5th

Full UK lockdown began two weeks before April 7th

Spain - lockdown from 2 weeks before Sunday 29th March

Italy - most of northern Italy covered by quarantine two weeks before Sunday 22nd March 2020

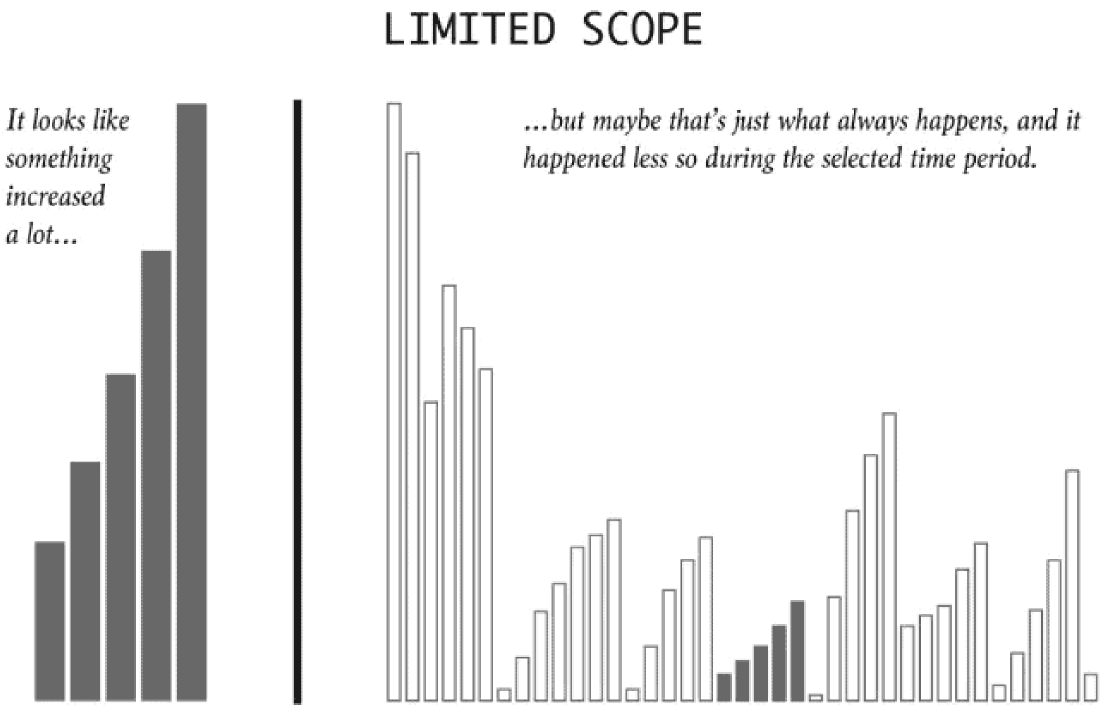
Germany - lockdown introduced in Bavaria two weeks before April 3rd

UK - schools, pubs and restaurants from two weeks before Friday 3rd April

DannyDorling.org. Illustration by Kirsten McClure @orpheuscat

Data-Visualization

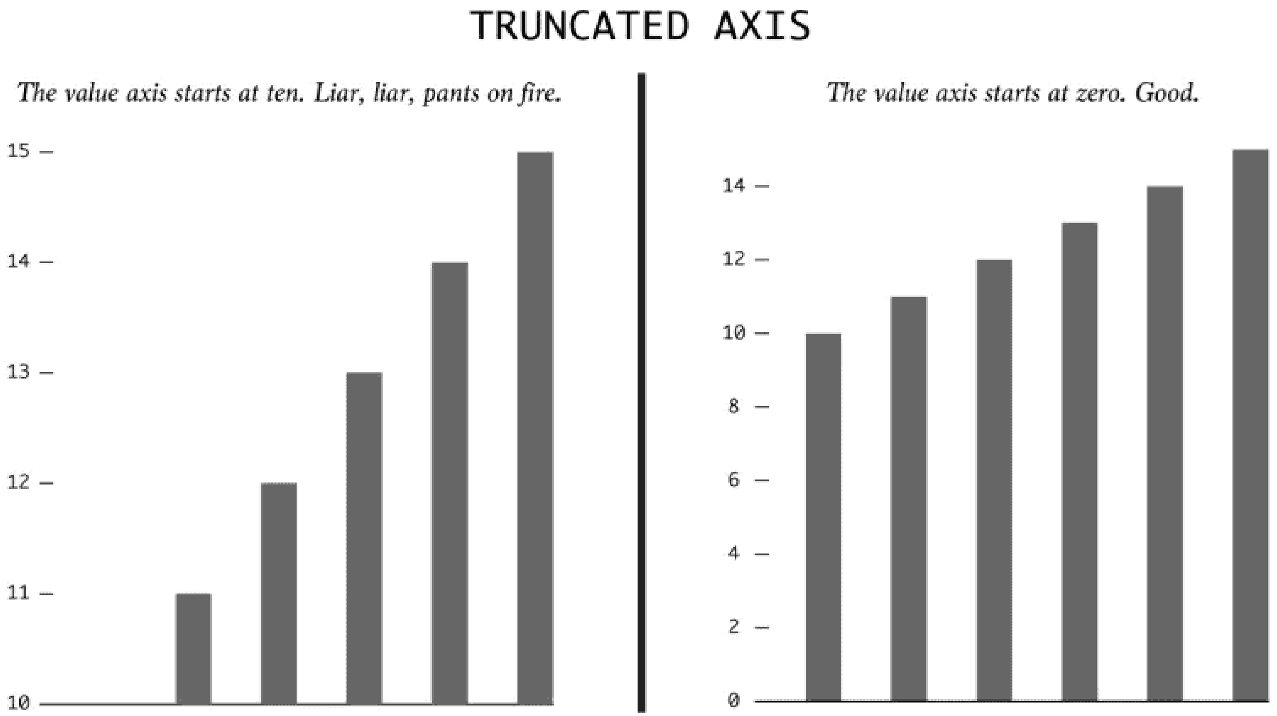
How NOT to do it



LIMITED SCOPE

It looks like something increased a lot…

…but maybe that's just what always happens, and it happened less so during the selected time period.



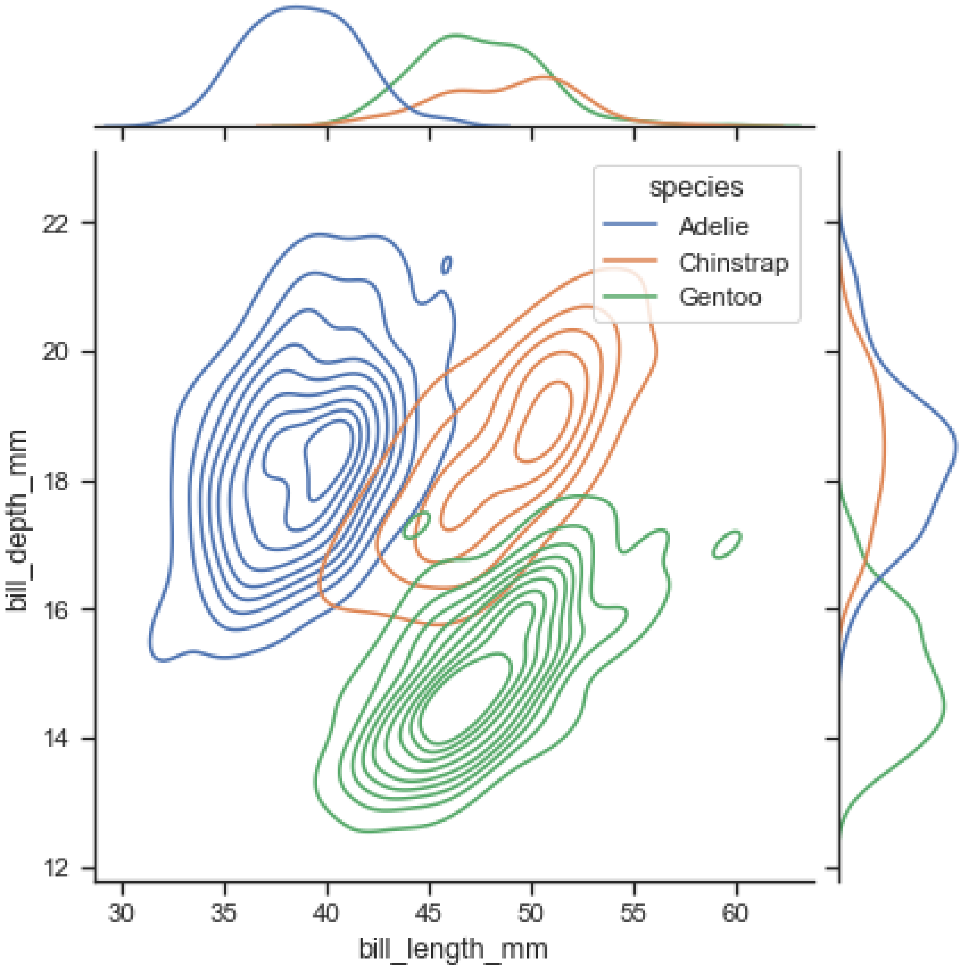
TRUNCATED AXIS

The value axis starts at ten. Liar, liar, pants on fire.

The value axis starts at zero. Good.

Data-Visualization

- **Visualizing your data should be the first and last thing you do!**
- Communicating results is a difficult but important part of Data Science
- The more complex the data the more important good and accurate data visualization becomes

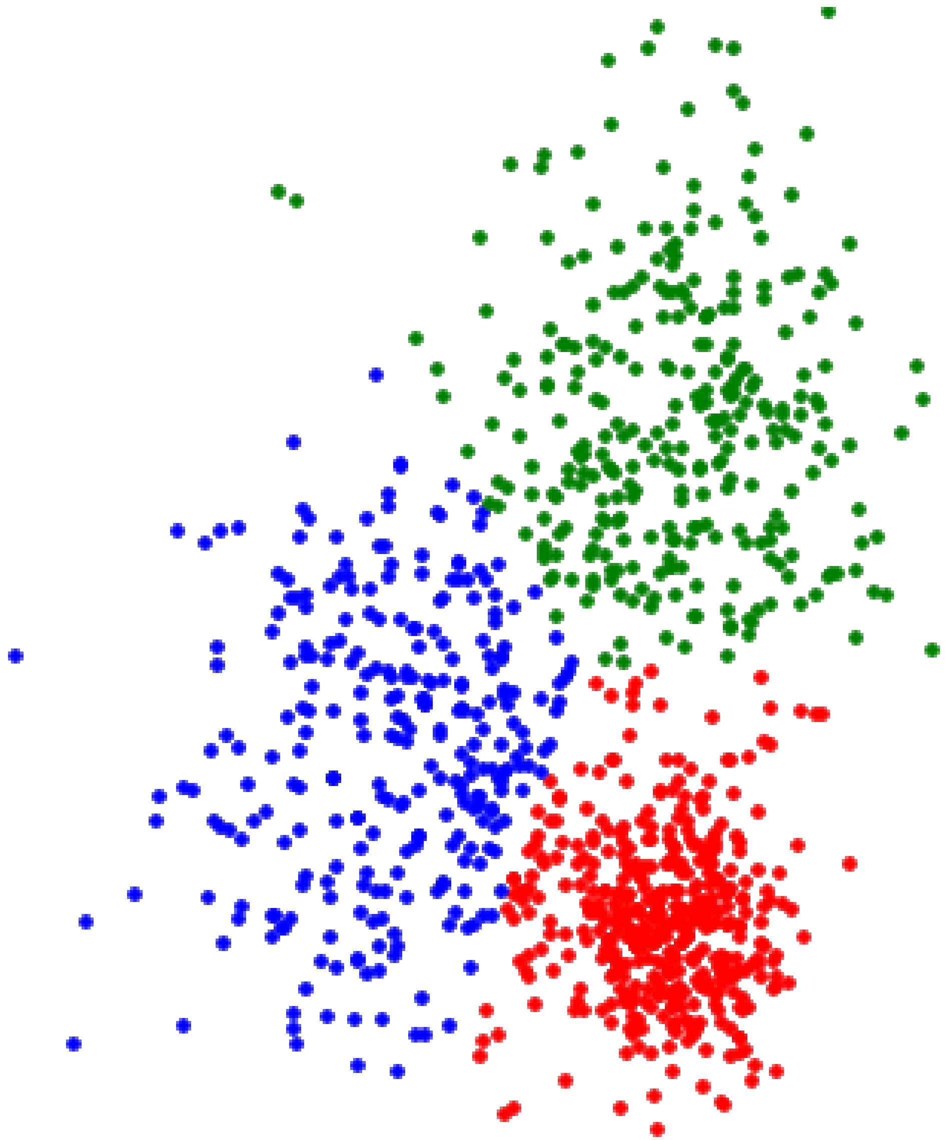


Hands-On

Part 2

Explore the Dataset “melb_data.csv” of the Melbourn Housing Market

1. Clean the dataset
2. Standardize the data
3. Think about how you would augment this dataset
4. Visualize and present an aspect of the dataset you find interesting



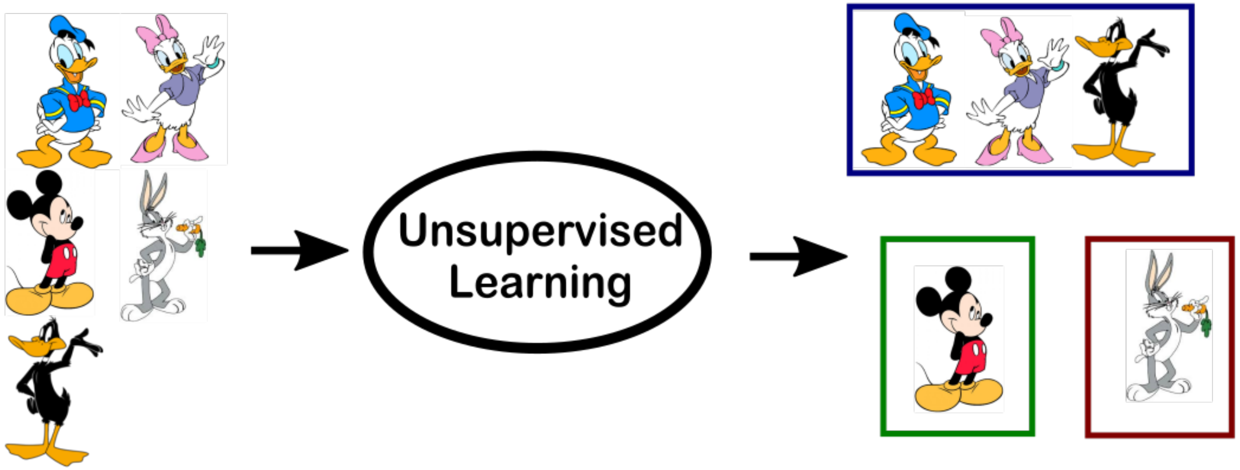
”

Unsupervised Machine Learning

NOSERYOUNG

Unsupervised ML

- **Idea: Find patterns & trends** in the data, without any prior knowledge
- These patterns may give us new insights into our data
- **Main Types:**
  - **Clustering**
  - **Dimensionality reduction**



NOSERYOUNG