Data Handling

Data Handling

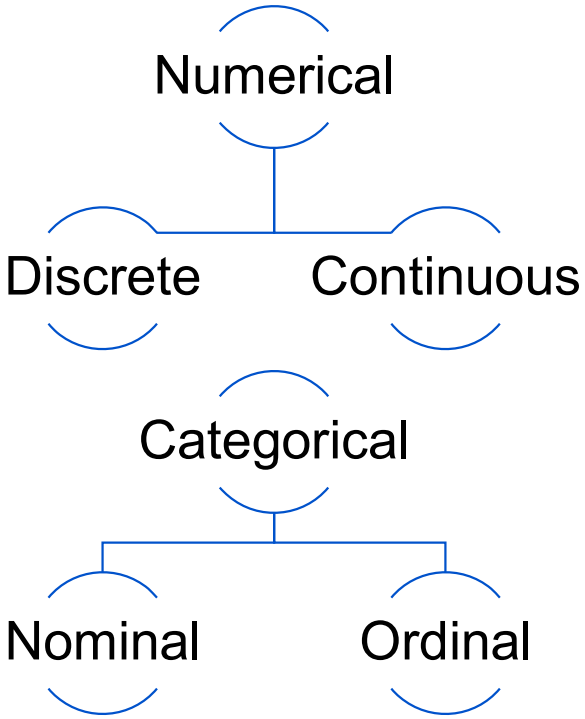
Common Tasks

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

NOSERYOUNG

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



Data cleaning

Missing values

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

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

NOSERYOUNG

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:

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

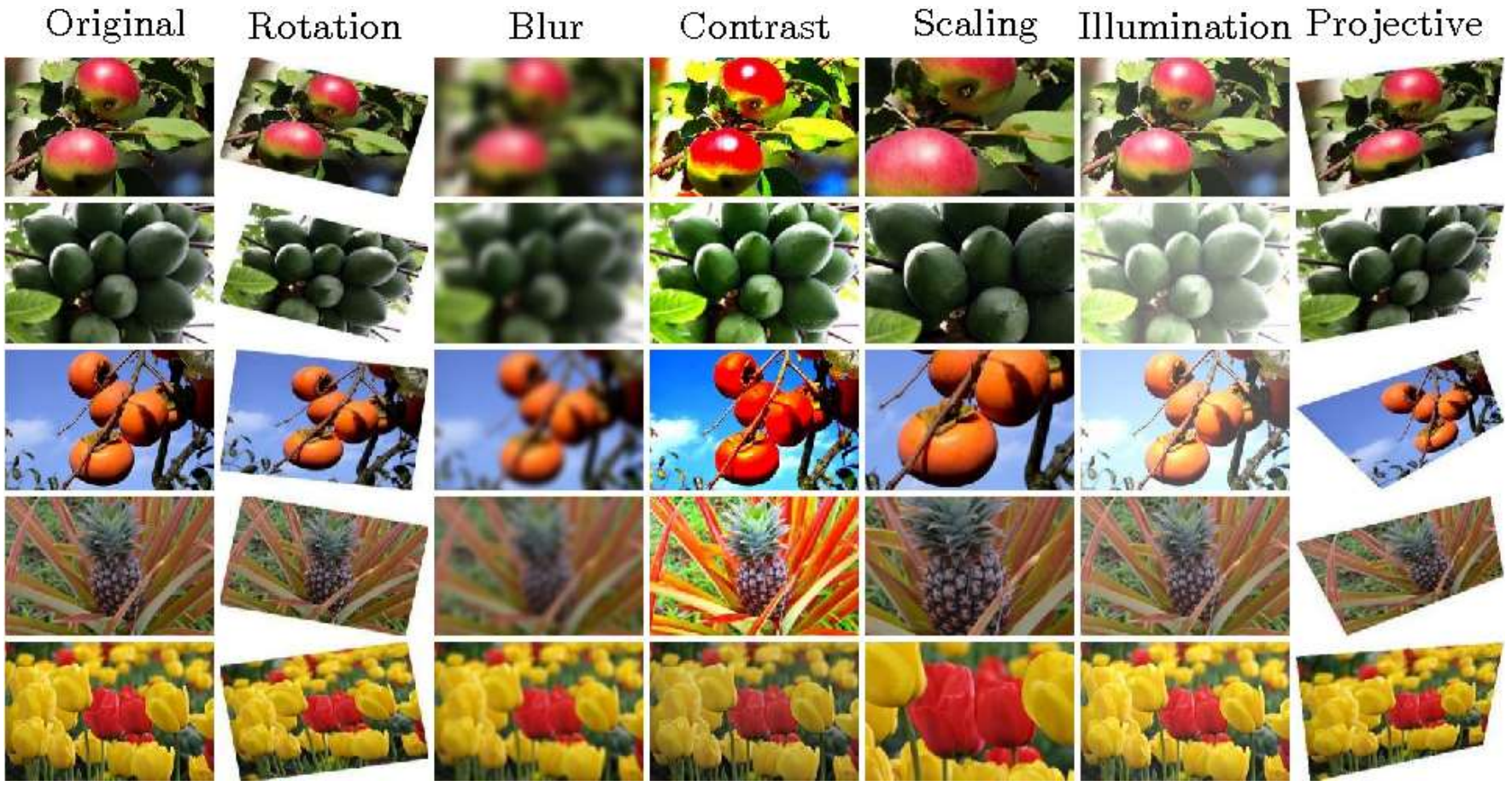
- Implemented by: sklearn.preprocessing.StandardScaler

Variance:

In general, how different are the values from the mean

NOSERYOUNG

Data Augmentation



NOSERYOUNG

Data Augmentation

- **Idea**: **Modify data to augment the dataset**
- **Improving model prediction accuracy** by increasing generalizability and increasing the size of the training dataset.
  - *E.g. the model should still work with black and white images*

Generalizability:
Ability of a model to be applied to a wide variety of real world problems

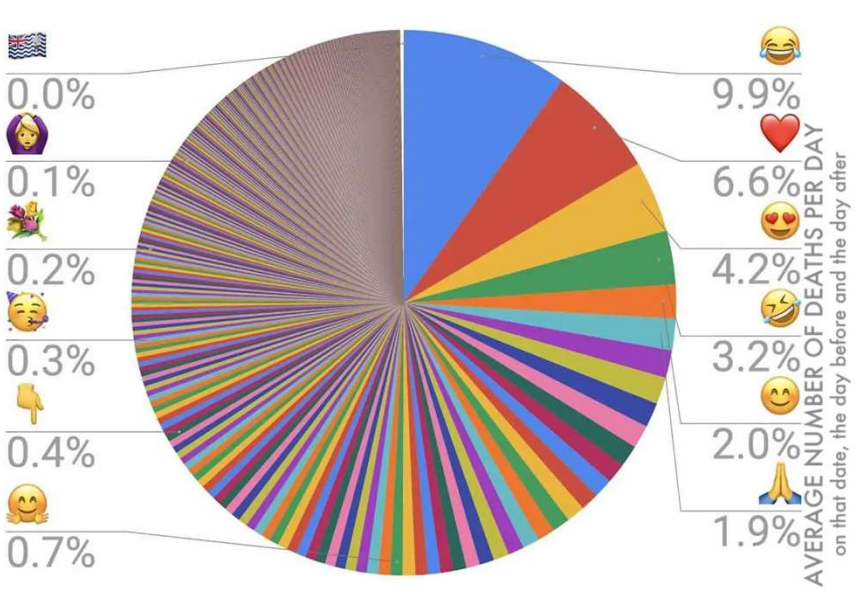
NOSERYOUNG

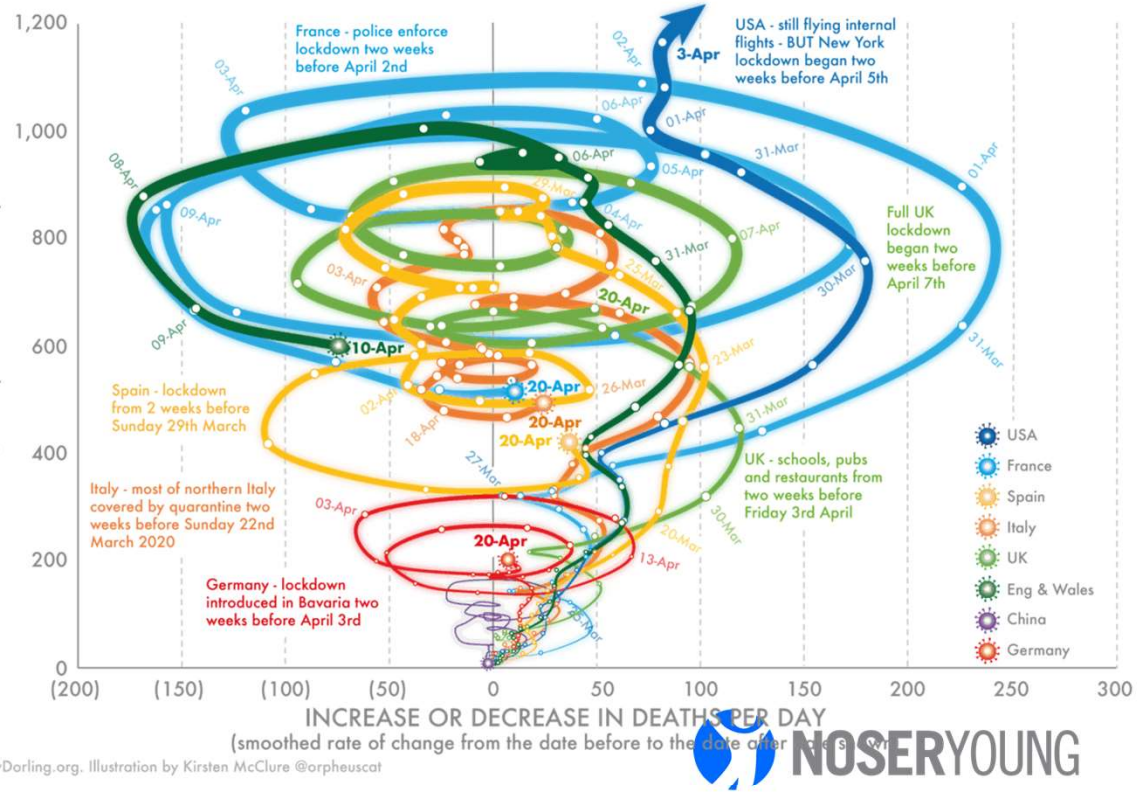
Data-Visualization

NOSERYOUNG

Data-Visualization

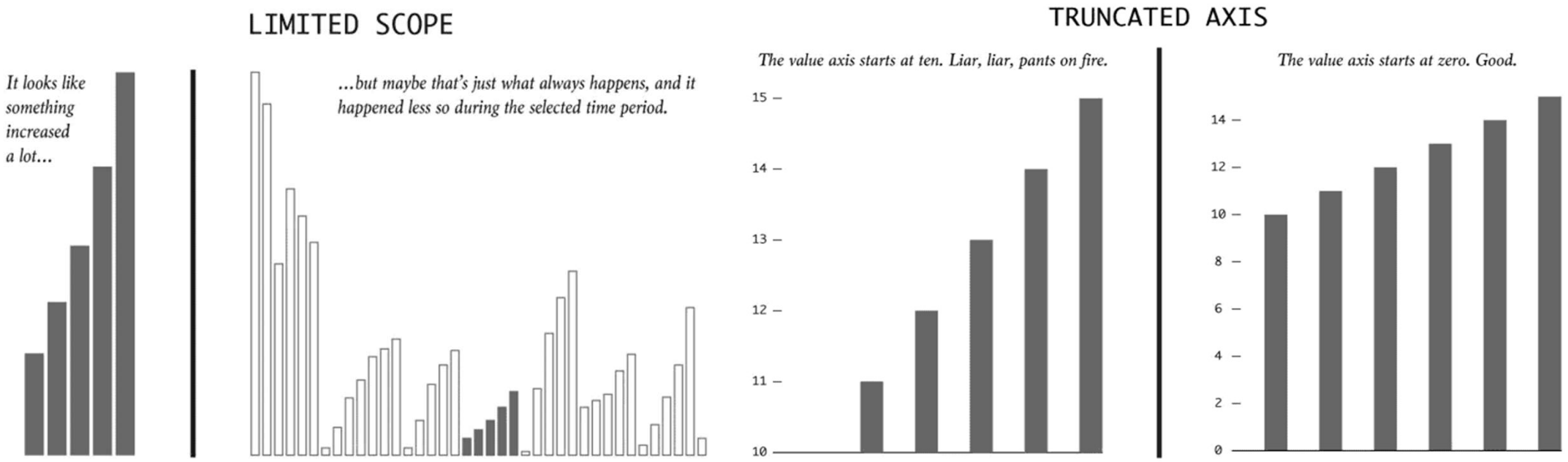
How NOT to do it





Data-Visualization

How NOT to do it

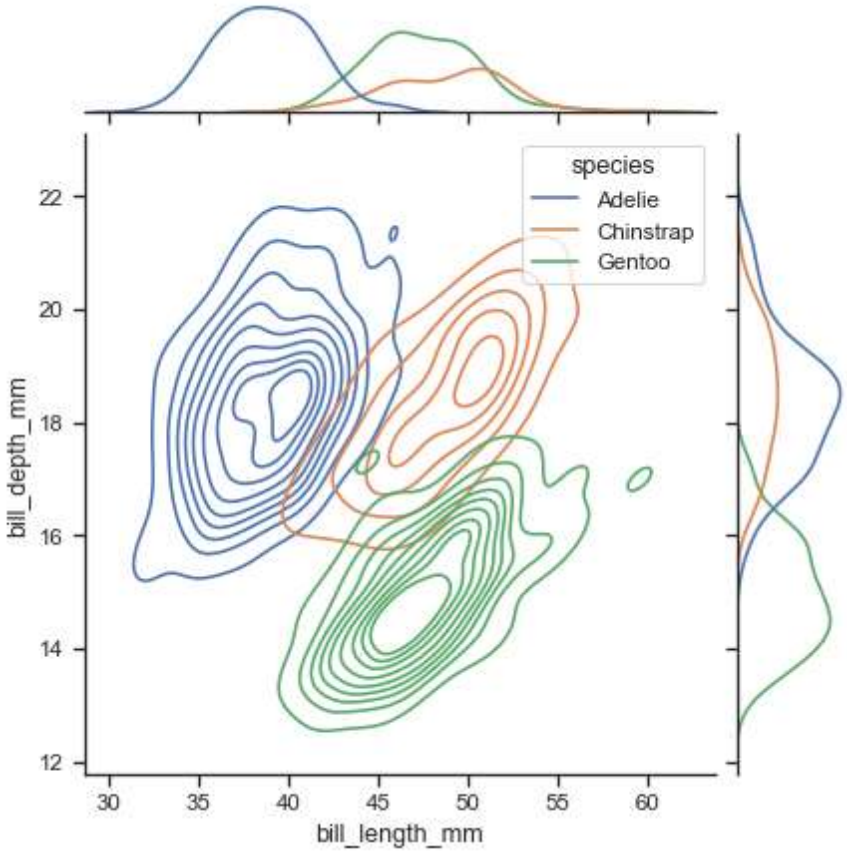


NOSERYOUNG

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

https://seaborn.pydata.org/examples/index.html



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

NOSERYOUNG