Data preprocessing

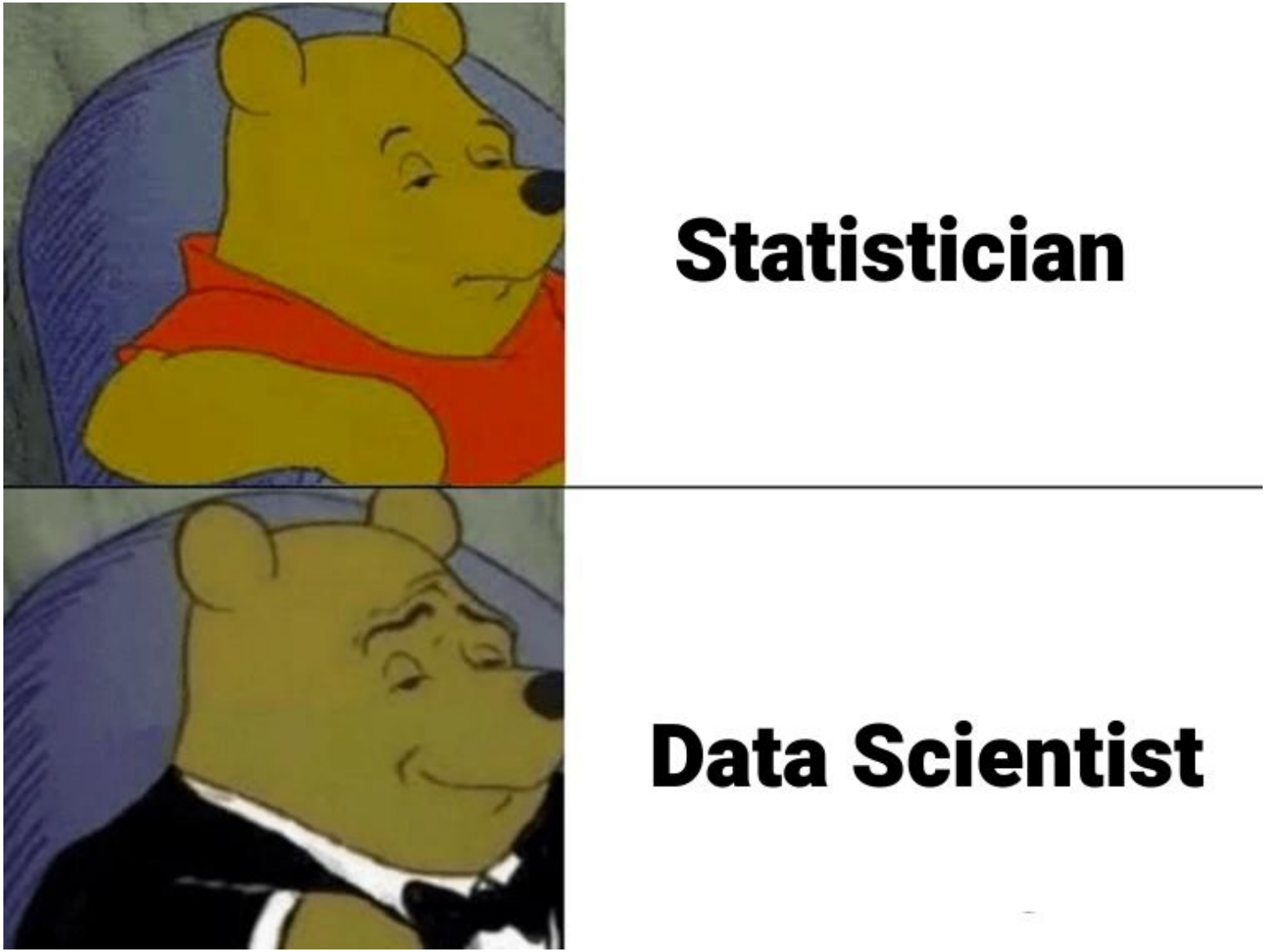
Elena Tuzhilina

Stanford University, Department of Statistics

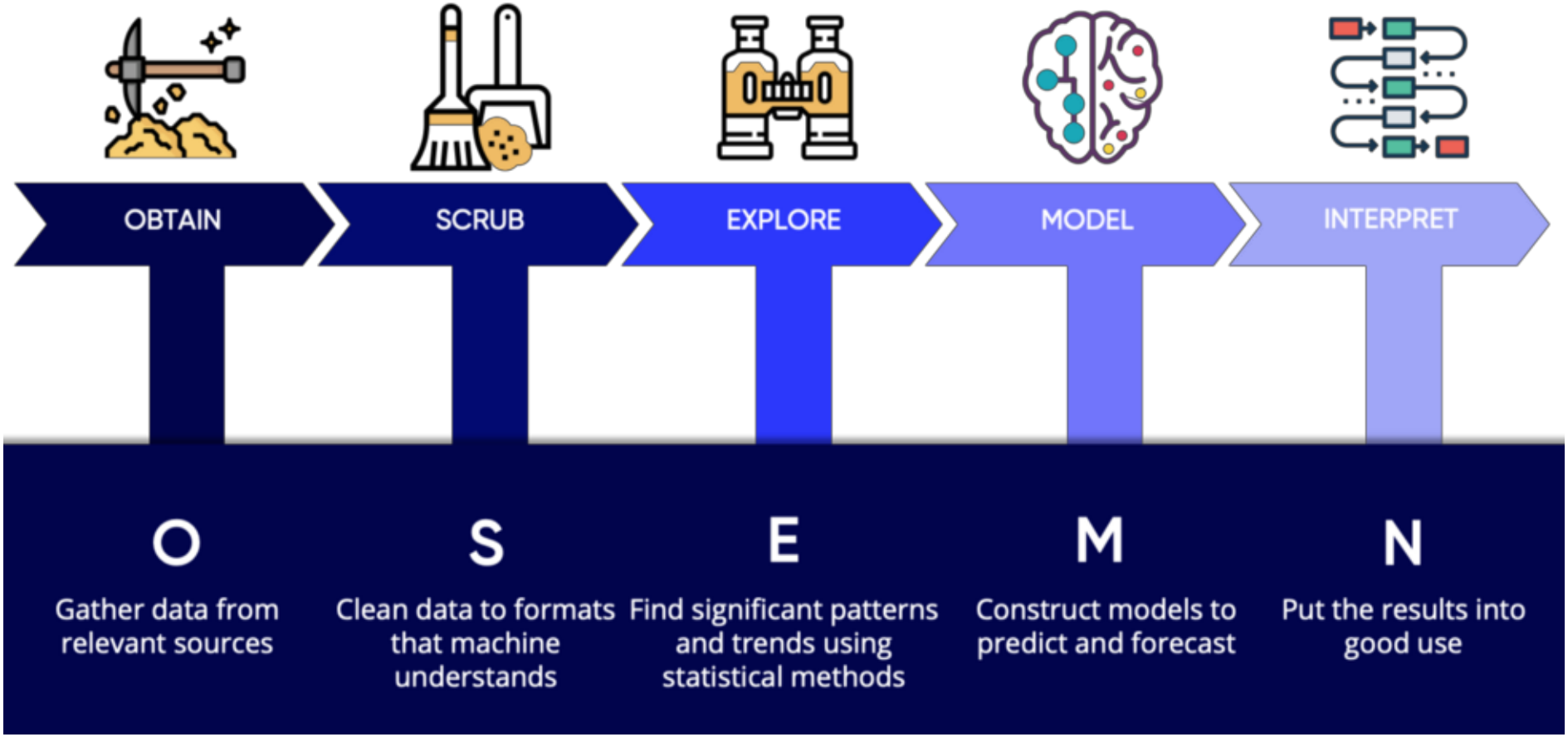
elenatuz@stanford.edu

January 18, 2022

Once upon a time I worked as a data scientist...



Data science process

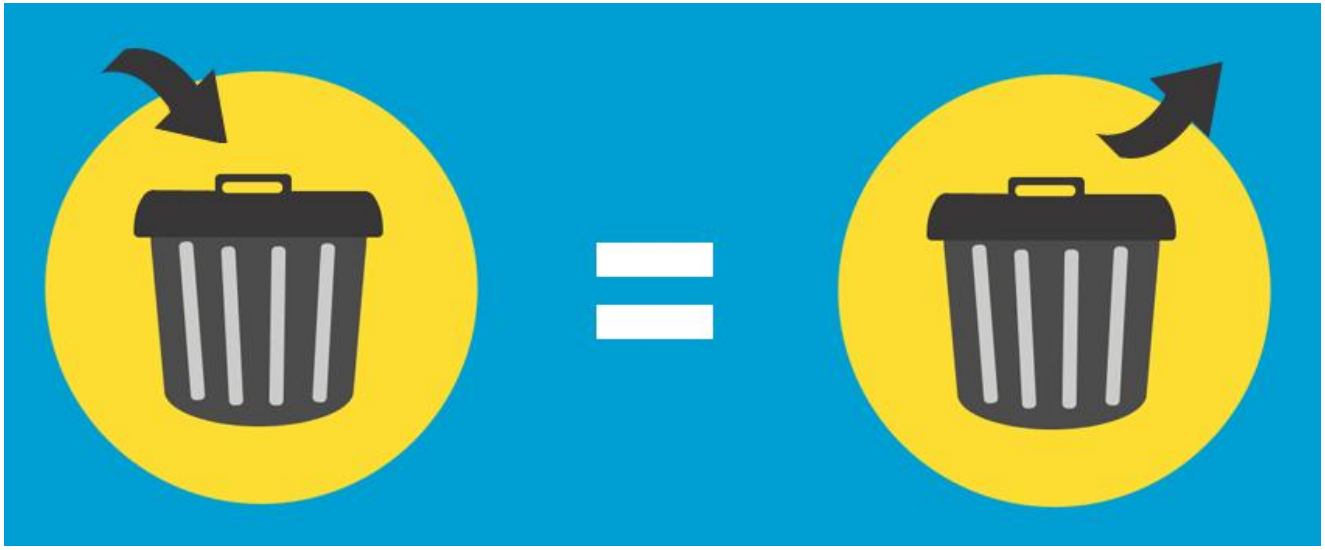


Definition

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Why is it important? Nonsense input data produces nonsense output.

GIGO: garbage in = garbage out

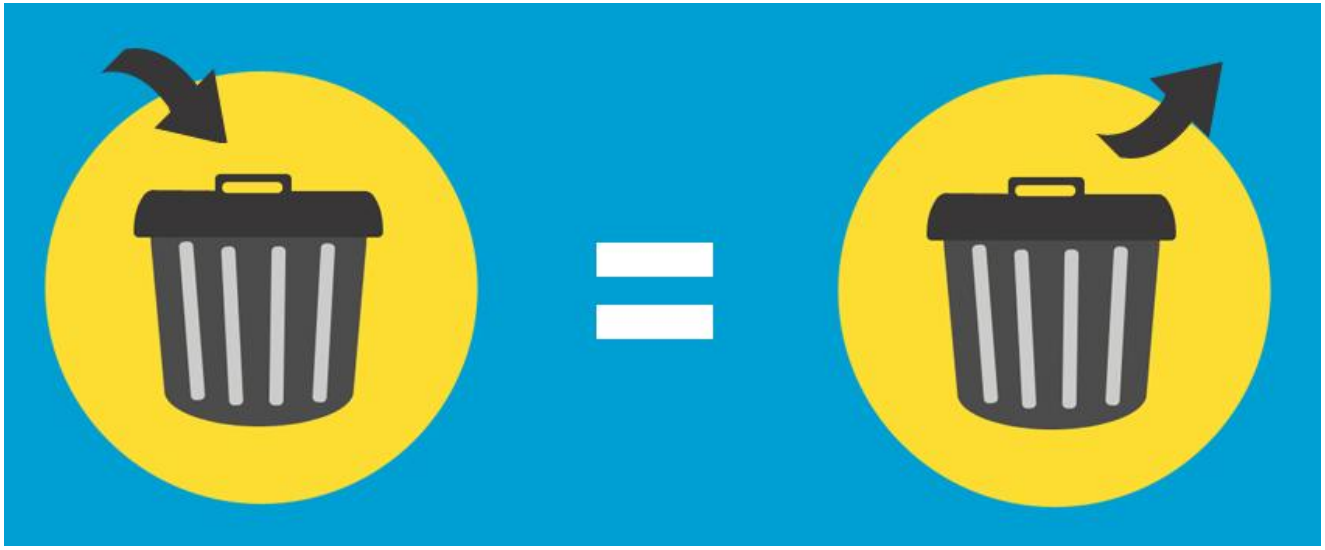


Definition

Data preprocessing is a data mining technique which is used to transform the raw data in a useful and efficient format.

Why is it important? Nonsense input data produces nonsense output.

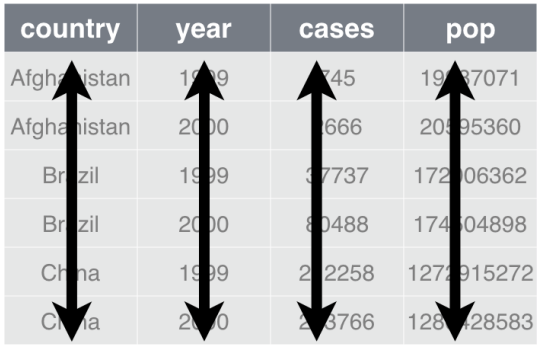
GIGO: garbage in = garbage out



What is the first step?

Data preprocessing stages

1. Sanity checks
2. Data reduction
   - remove uninformative features
   - remove correlated features
3. Data cleaning
   - remove duplicates
   - handle missing values
   - handle outliers
4. Data transformation
   - do binning
   - handle categorical variables
   - apply normalization/standardization
   - handle skewed distribution

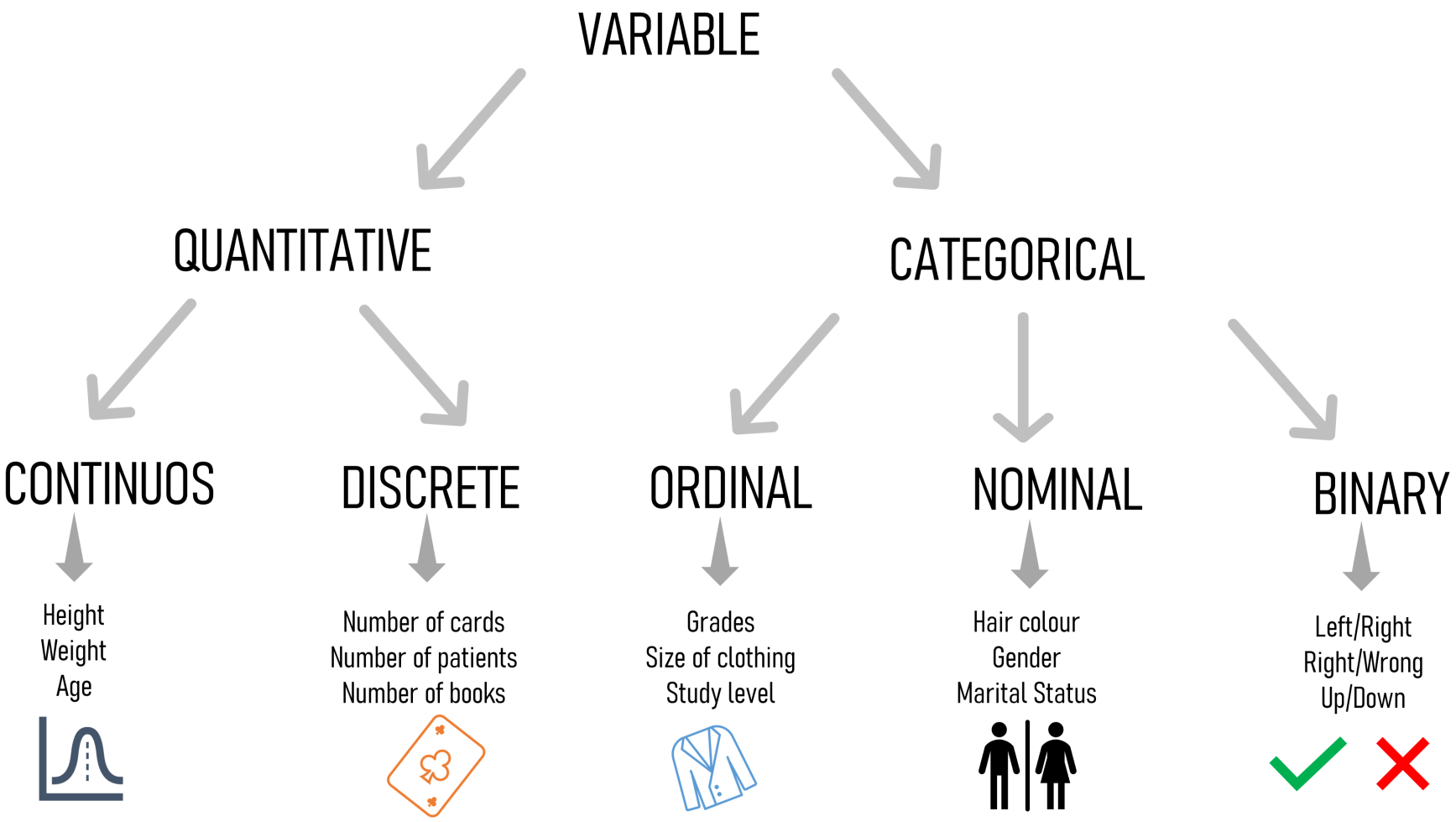


Variables

Observations

Preprocessing can be done on observation or variable level

Data types



Preprocessing strongly depends on variable type

1. Sanity checks

Example:

- out-of-range values, e.g. *Income: -100*
- impossible combinations, e.g. *Sex: Male; Pregnant: Yes*

Study the data and run some simple checks

2. Data reduction

Uninformative features

1. a lot of missing values, e.g. *contains* $> 70\%$ *of NAs*
2. categorical with too many values, e.g. *UserID: jkaj1daj*
3. constant/almost constant, e.g. *Year: 2021*

Correlated features

1. Check Pearson/Spearman correlation (numerical)
2. Check Cramer's V statistics (categorical)

You can remove some redundant features
or
you can do nothing and apply regularization

2. Data reduction

Uninformative features

1. a lot of missing values, e.g. *contains* $> 70\%$ *of NAs*
2. categorical with too many values, e.g. *UserID: jkaj1daj*
3. constant/almost constant, e.g. *Year: 2021*

Correlated features

1. Check Pearson/Spearman correlation (numerical)
2. Check Cramer's V statistics (categorical)

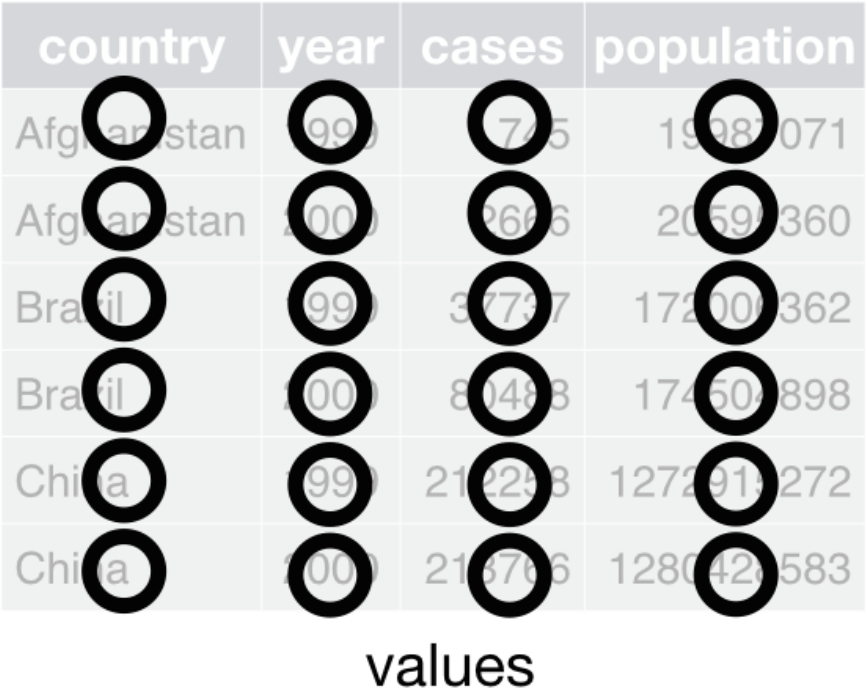
Do you remove features? If yes, what criteria do you apply?

3. Data cleaning: duplicates

You can remove some redundant observations

3. Data cleaning: missing values

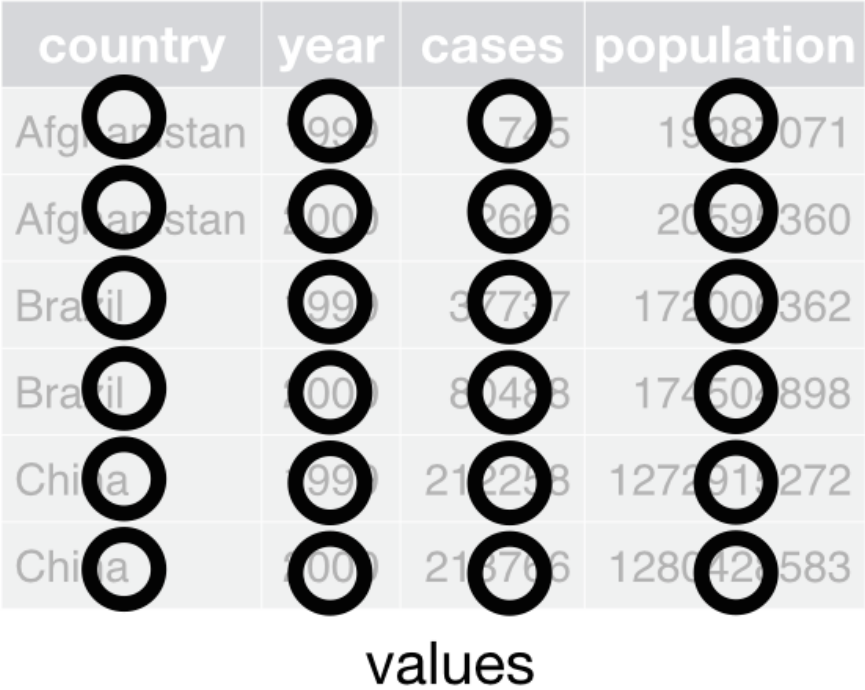
1. Remove observation
2. Impute numerical
   - use mean/median
   - use k nearest neighbours
   - use low-rank matrix approximation
3. Impute categorical
   - add new category "missing"
   - use most frequent values



You can impute missing values
however
some machine learning models can deal with them

3. Data cleaning: missing values

1. Remove observation
2. Impute numerical
   - use mean/median
   - use k nearest neighbours
   - use low-rank matrix approximation
3. Impute categorical
   - add new category "missing"
   - use most frequent values



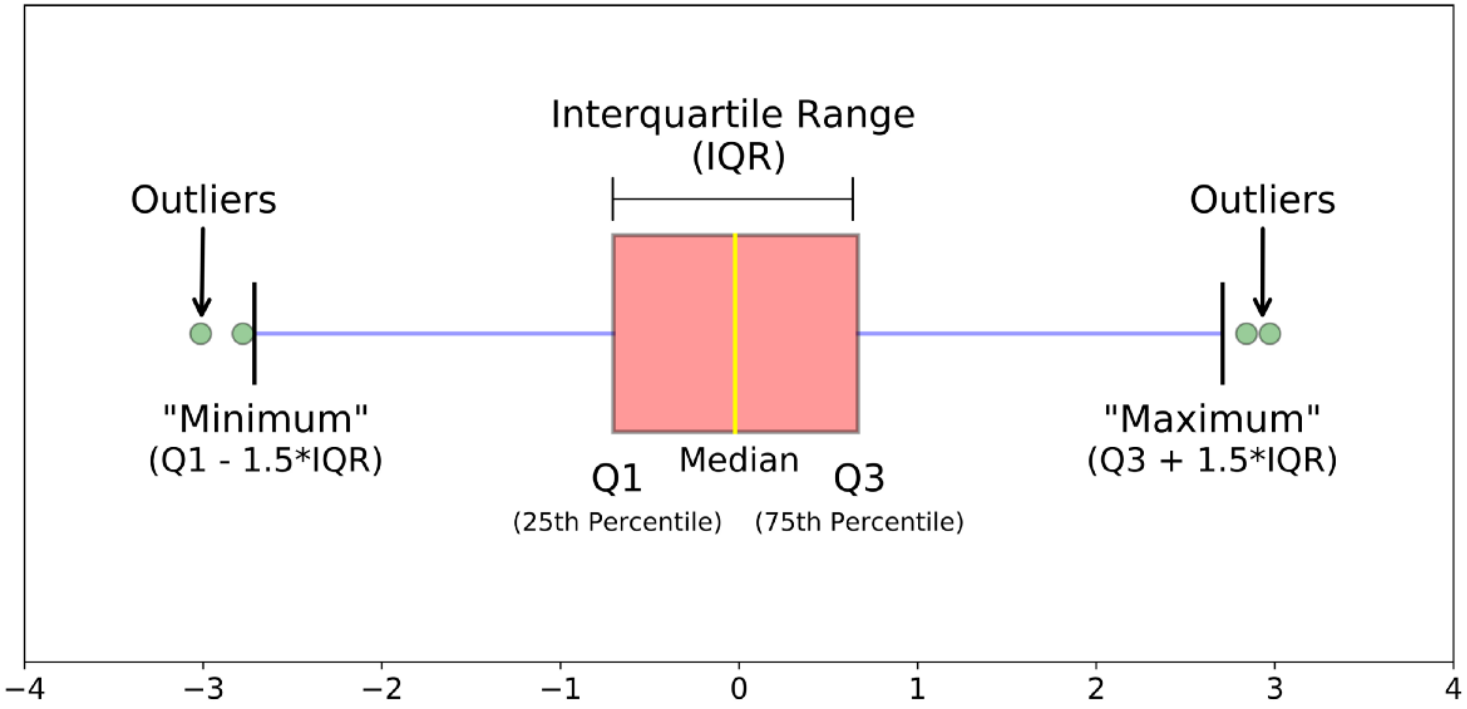
How do you handle missing values?

3. Data cleaning: outliers

An outlier is an observation that lies outside the overall pattern of a distribution

3. Data cleaning: outliers

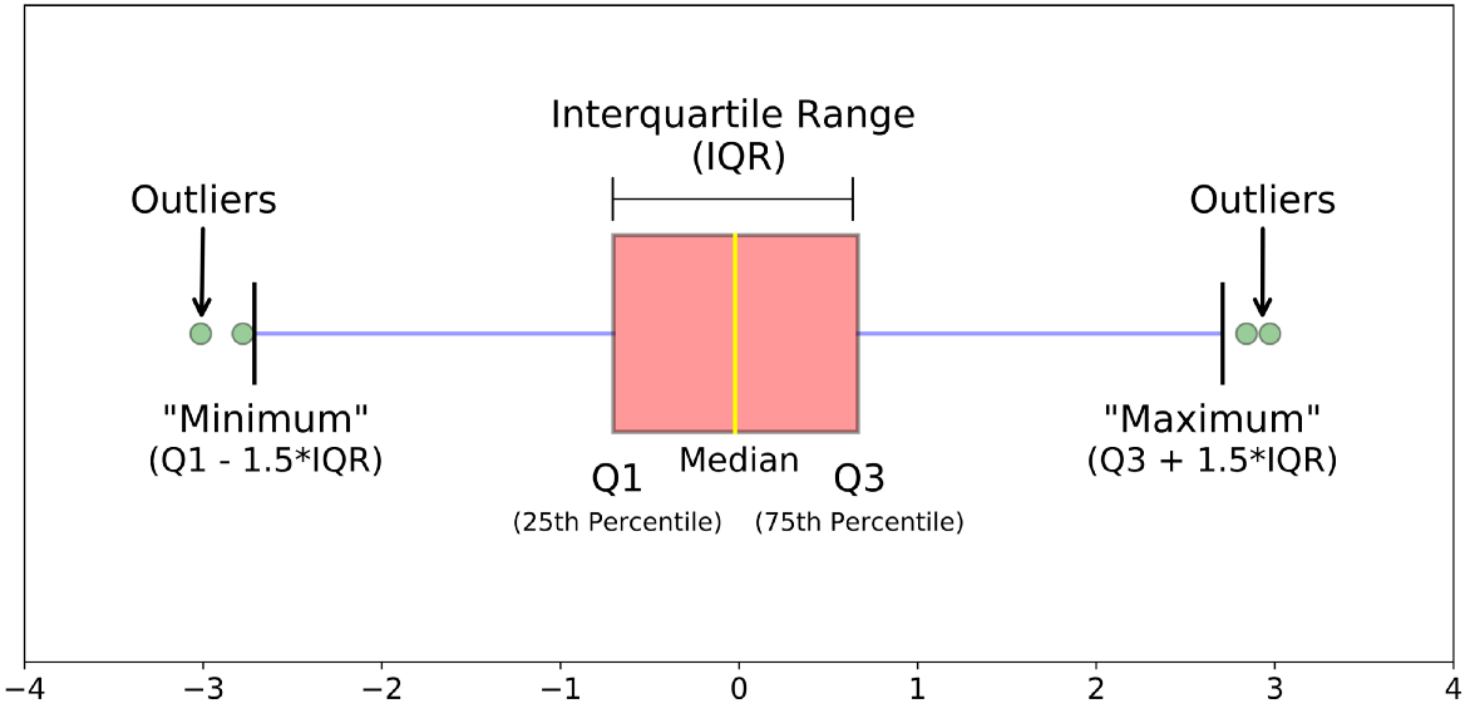
1. Remove observation
2. Change/trim the value
3. Apply transformation, e.g. log-transformation



To detect outlies you can use IQR

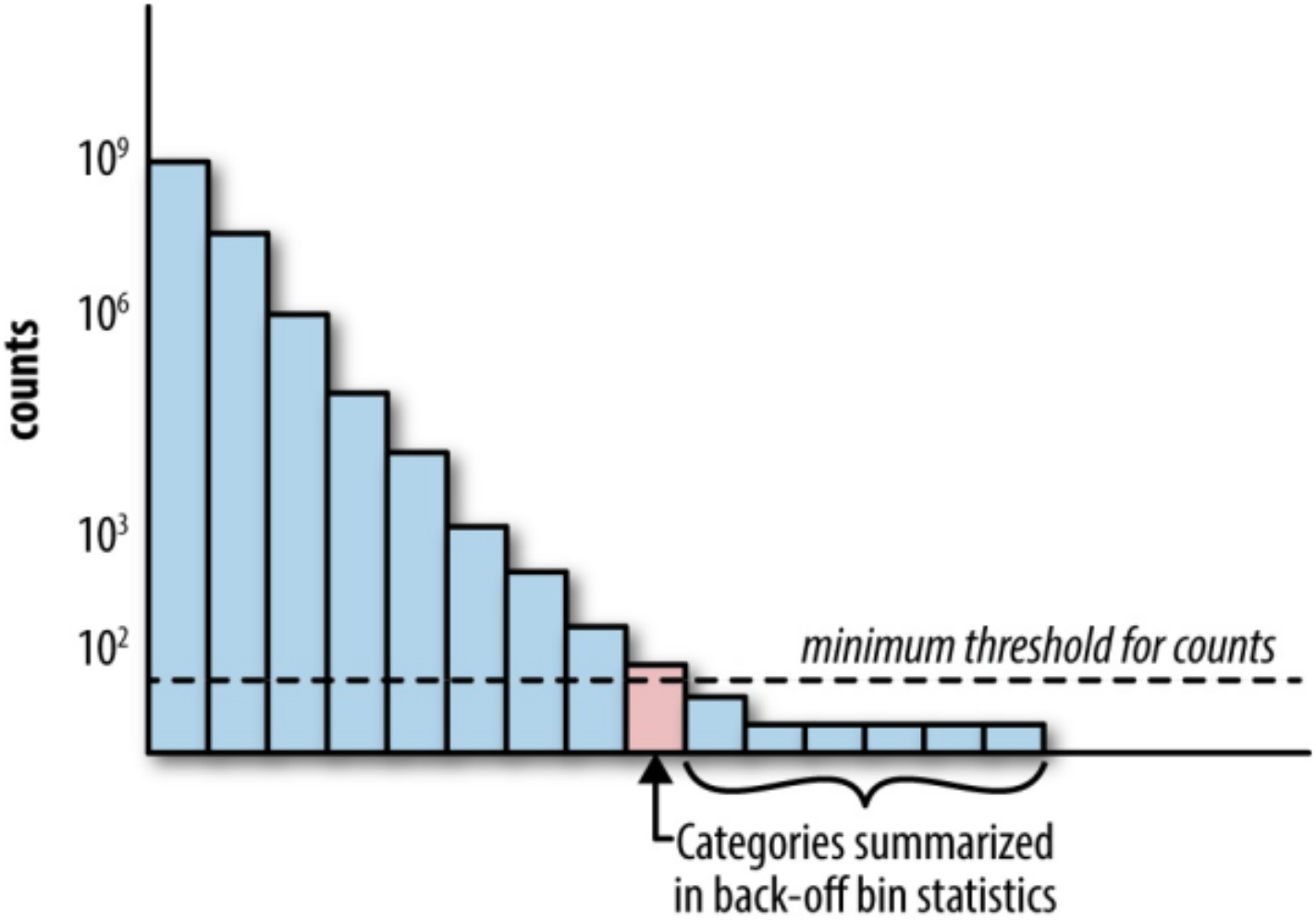
3. Data cleaning: outliers

1. Remove observation
2. Change/trim the value
3. Apply transformation, e.g. log-transformation



How do you deal with outliers?

4. Data transformation: binning



You can apply binning to reduce number of different values of a categorical feature

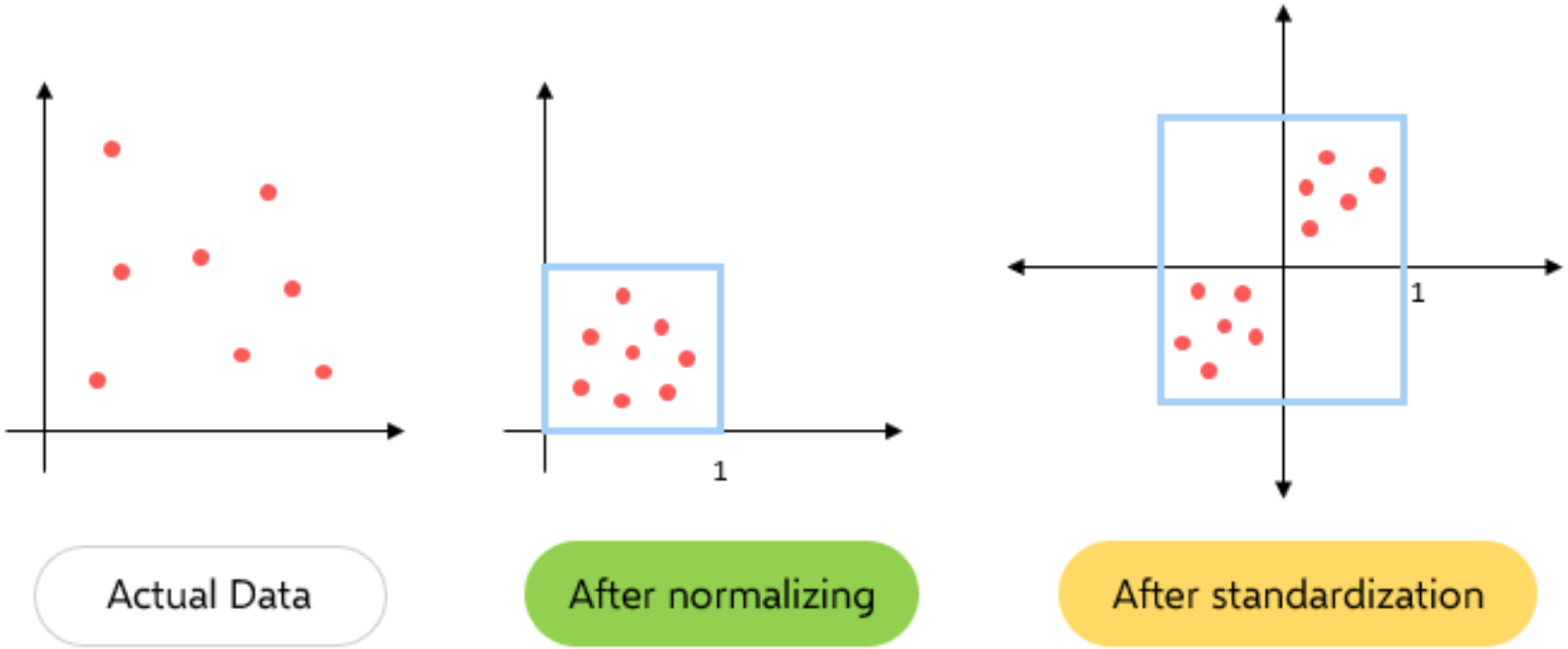
4. Data transformation: categorical variable

Color
Red
Red
Yellow
Green
Yellow

Red	Yellow	Green
1	0	0
1	0	0
0	1	0
0	0	1
0	1	0

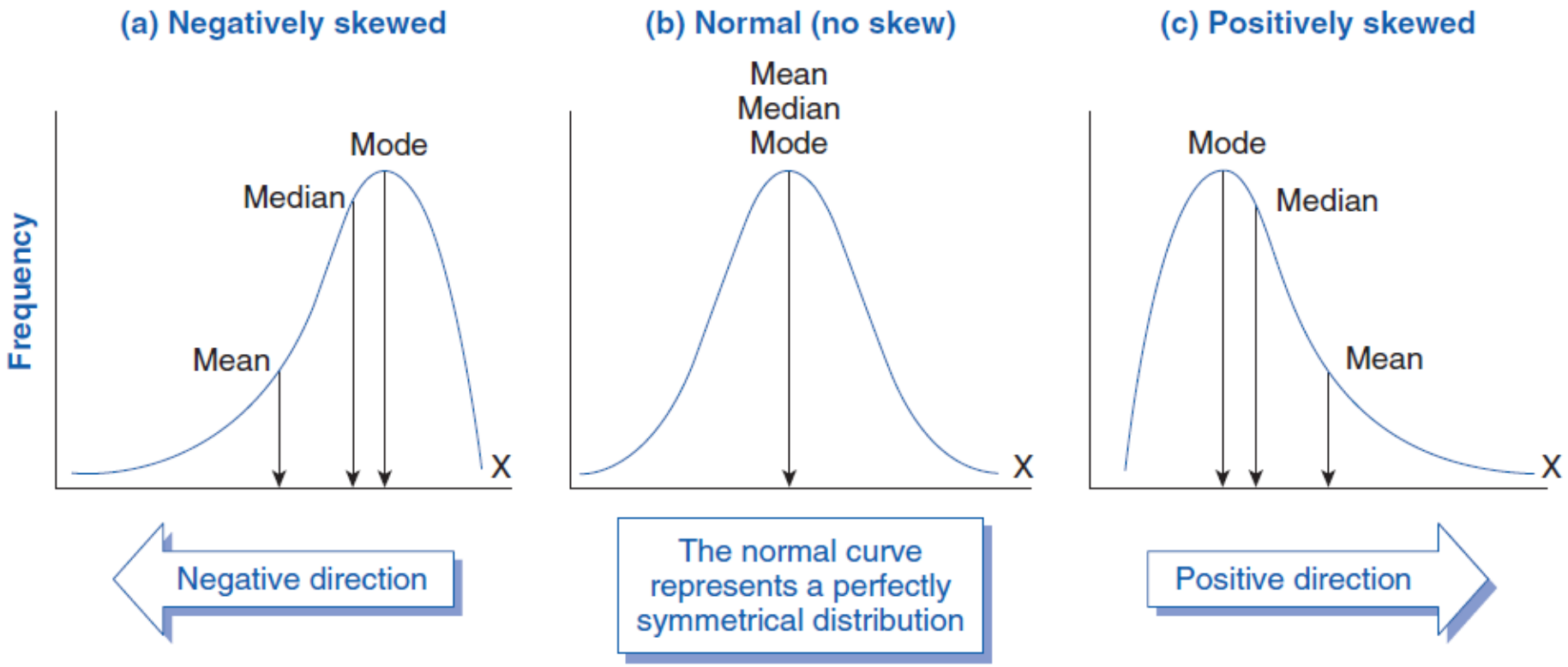
You need to convert all categorical variables to numeric format

4. Data transformation: normalization/standardizaiton



Some ML techniques requires data scaling and centering

4. Data transformation: skewed distribution



Some ML techniques does not work well for very skewed distributions
You can apply log-transformation to these features

Questions

1. **Do you do any data preprocessing? How complex is it?**
2. **From your point of view, how important is data preprocessing step?**
3. **What tools do you use for data preprocessing?**

Images used

1. Data Scientist
2. OSEMN
3. GIGO
4. Variables, Observations, Values
5. Not sure I can trust this data
6. Duplicates
7. Outliers
8. Boxplot
9. Binning
10. Feature encoding
11. Data standardization
12. Skewed distribution