06

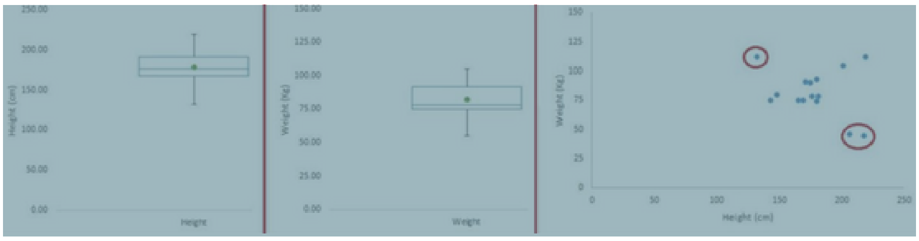
Outlier, Noise, and Missing Value

Outliner

Outlier is an observation that appears far away and diverges from an overall pattern in a sample.

Outliner Types

Univariate outlier : a univariate outlier is a data point that consists of an extreme value on one variable.
Multivariate outlier : a multivariate outlier is a combination of unusual scores on at least two variables/in an n-dimensional space



What is the Impact of Outliers on a Data Set?

It increases the error variance and reduces the power of statistical tests. If the outliers are non-randomly distributed, they can decrease normality. They can bias or influence estimates that may be of substantive interest. They can also impact the basic assumption of regression, ANOVA and other statistical model assumptions.

What Causes Outliers?

- Artificial(error)/non-Natural
- Natural

Most common causes of outliers on a data set:

- Data entry errors (human errors)
- Measurement errors (instrument errors)
- Experimental errors (data extraction or experiment planning/excuting errors)
- Intentional (dummy outliers made to test detection methods)
- Data processing errors (data manipulation errors)
- Sampling errors (exctracting or mixing data from wrong or various sources)
- Natural (not an error, novelties in data)

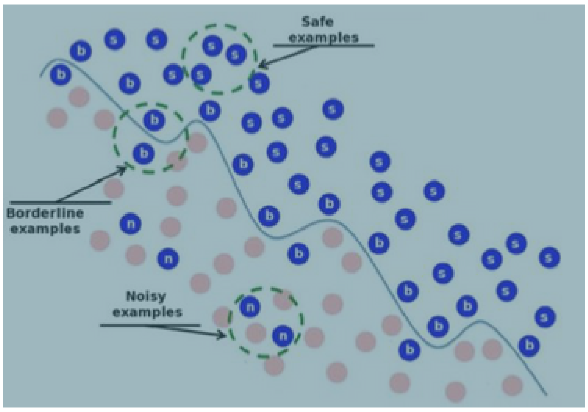
How to Remove the Outlier?

The common techniques used to deal with outliers are:

- Deleting observations
- Transforming and binning values
- Imputing
- Treat outliers separately

Data Noise

Noisy data is data with a large amount of additional meangingless information in it called noise. This includes corrupted data. It also includes any data that a user system cannot understand and interpret correctly..



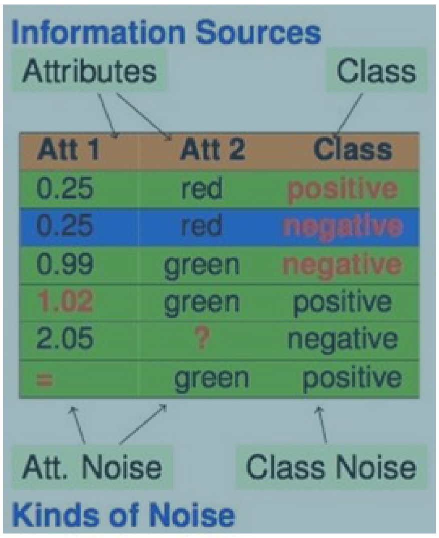
Noise Types

Class noise :

- Contradictory examples
- Mislabeled examples

Attribute noise :

- Erroneous values
- Missing values,
- Don’t care values


Information Sources
Attributes
Class
Att. Noise
Class Noise
Kinds of Noise

Att 1	Att 2	Class
0.25	red	positive
0.25	red	negative
0.99	green	negative
1.02	green	positive
2.05	?	negative
=	green	positive

Missing Value

In statistics, missing data, or missing values, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and can have a significant effect on the conclusions that can be draw from the data.

Why do Data Have Missing Value?

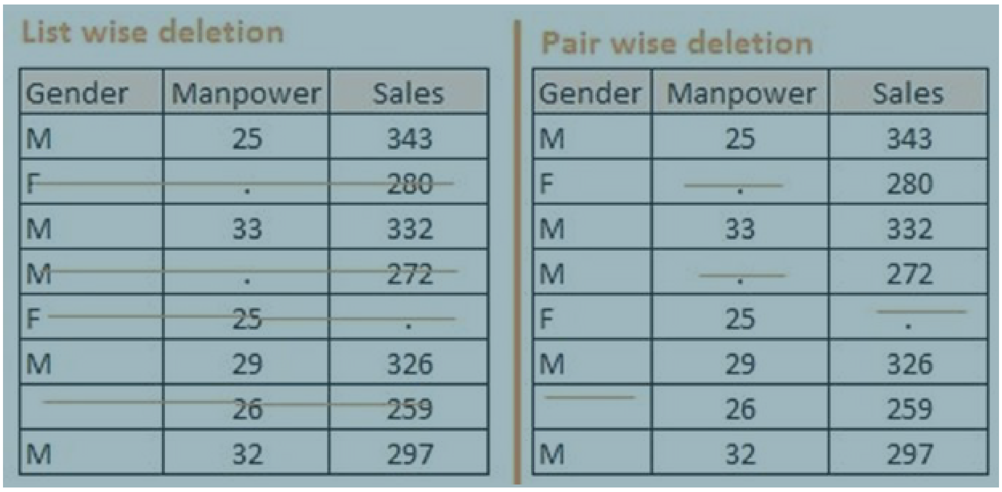
Data extraction

Data collection

missing completely at random, missing at random, missing that depends on unobserved predictors, missing that depends on the missing value itself

Which are the Methods to Treat Missing Values?

1. Deletion



List wise deletion

Gender	Manpower	Sales
M	25	343
~~F~~	~~.~~	~~280~~
M	33	332
~~M~~	~~.~~	~~272~~
~~F~~	~~25~~	~~.~~
M	29	326
	~~26~~	~~259~~
M	32	297

Pair wise deletion

Gender	Manpower	Sales
M	25	343
F	~~.~~	280
M	33	332
M	~~.~~	272
F	25	~~.~~
M	29	326
	26	259
M	32	297

2. Mean/Mode/Media Imputation
Generalized imputation and similar case imputation

3. Prediction Model

4. KNN Imputation

Picture Source

- unsplash.com
- pexels.com
- pixabay.com
- https://analyticsvidhya.com/
- https://medium.com/
- https://sci2s.ugr.es/noisydata