Analyse et manipulation des données

DigitalLab@LaPlataforme_

Tools for data pre-processing

- Descriptive and inferential statistics tools
  - Univariate and multivariate analysis
- Data transformations: indexing, grouping and aggregation
- Feature Selection
- Combination of data sets

Now we add

- Encoding of categorical variables
- Dimensionality reduction with PCA
- Dimensionality reduction with LDA

Encodings

Machine learning algorithms require exclusively numerical data

We need to transform our categorical variables to some numerical format

One-hot encoding

Id	neighbourhood
1	Saint Vincent
2	Hill of the Roses
3	Maipú
4	Saint Vincent
5	Ituzaingó

Id	neighbourhood =Saint Vincent	neighbourhood =Hill of the Roses	neighbourhood =Maipú	neighbourhood =Ituzaingó
1				
2				
3				
4				
5				

One-hot encoding

Id	neighbourhood
1	Saint Vincent
2	Hill of the Roses
3	Maipú
4	Saint Vincent
5	Ituzaingó

Id	neighbourhood =Saint Vincent	neighbourhood =Hill of the Roses	neighbourhood =Maipú	neighbourhood =Ituzaingó
1	1	0	0	0
2				
3				
4				
5				

One-hot encoding

Id	neighbourhood
1	Saint Vincent
2	Hill of the Roses
3	Maipú
4	Saint Vincent
5	Ituzaingó

Id	neighbourhood =Saint Vincent	neighbourhood =Hill of the Roses	neighbourhood =Maipú	neighbourhood =Ituzaingó
1	1	0	0	0
2	0	1	0	0
3	0	0	1	0
4	1	0	0	0
5	0	0	0	1

The curse of dimensionality

By encoding the data in this way, we generate high-dimensional sparse vectors

- Takes up a lot of memory space
- The resulting vectors are orthogonal.
    - All vectors are the same distance from each other (if they have norm 1)
    - We cannot compute operations like the dot product.

Dimensionality reduction

Target

Reduce the number of columns or variables in our dataset

+

Preserve as much information as possible

What techniques do we know so far?

Mathematical formalization

Let's express the data set as a matrix X with n rows and m columns. Each row is a vector x_i that inhabits a mathematical space with m dimensions. Each dimension intuitively corresponds to a column.

$$X \in \mathbb{R}^{n \times m}; x_i \in \mathbb{R}^m$$

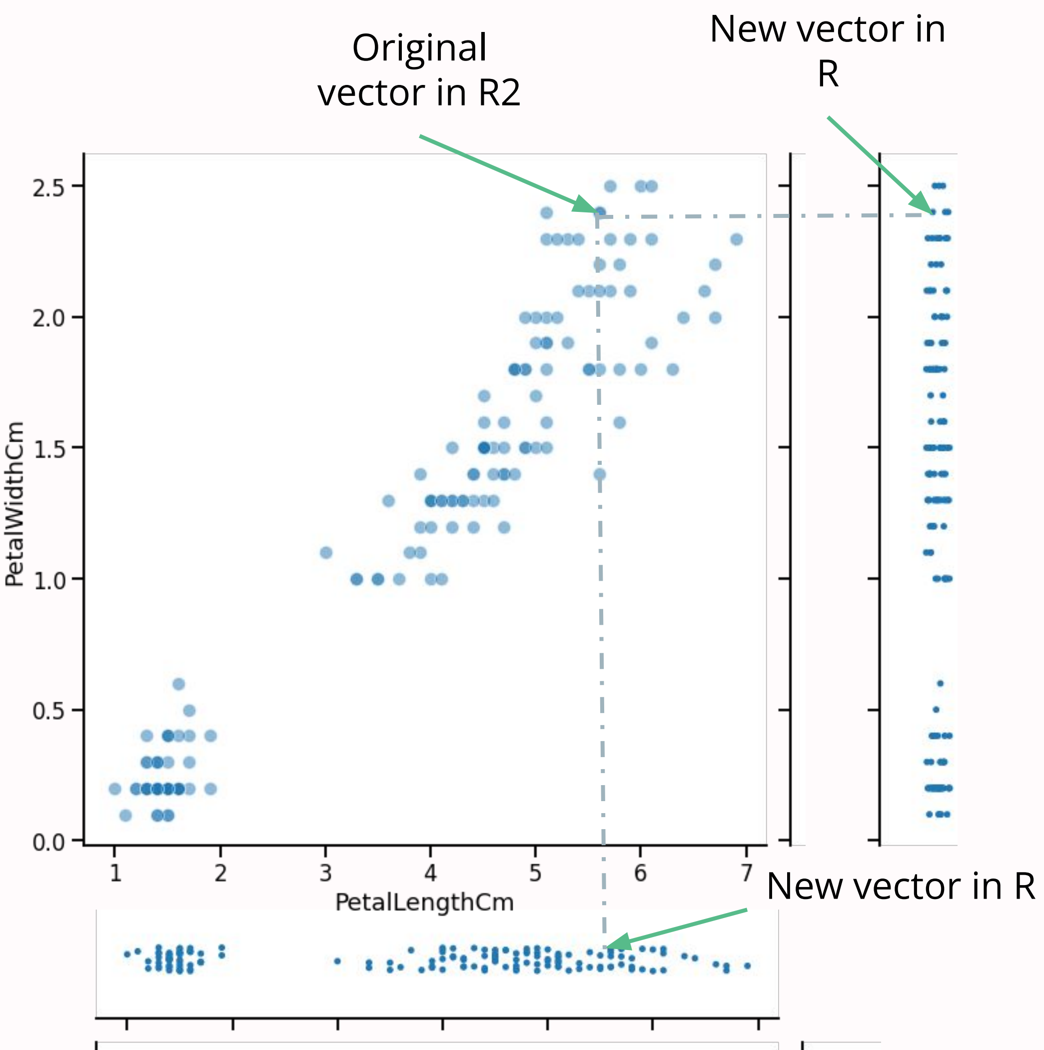
We want to obtain a new matrix Z that has the same number of rows, but a number of columns d much smaller than m.

$$Z \in \mathbb{R}^{n \times d}; d \ll m$$

Column deletion

Each row is a vector x in R2, that is, it has two dimensions.

If we remove any of them, we project the points to the direction of the x or y axis



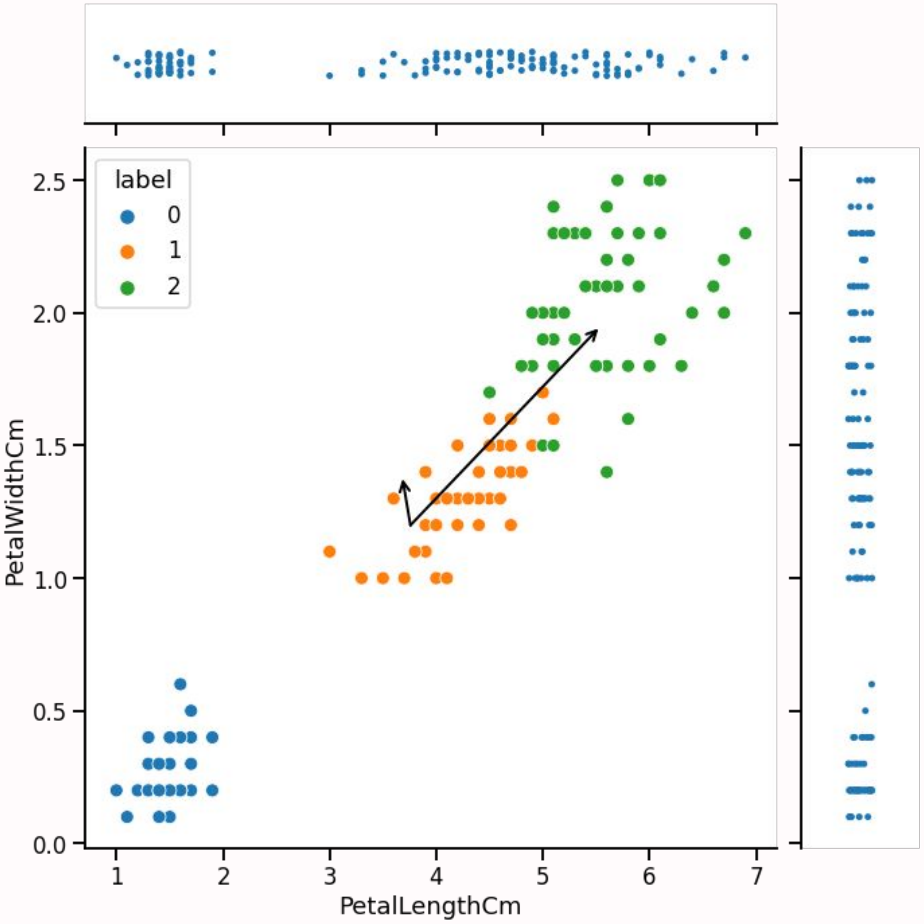
Principal Component Analysis (PCA)

- Algebraic method (does not depend on domain knowledge).
- Compute a set of addresses called principal components:
  - They are orthogonal (independent)
  - They are ordered according to the variance of the original data they capture.
- The matrix X is projected in the directions of its principal components
- The first k dimensions of the new projected matrix are selected.

Principal Components

The principal components of a matrix are the orthogonal directions of greatest variation of the data.

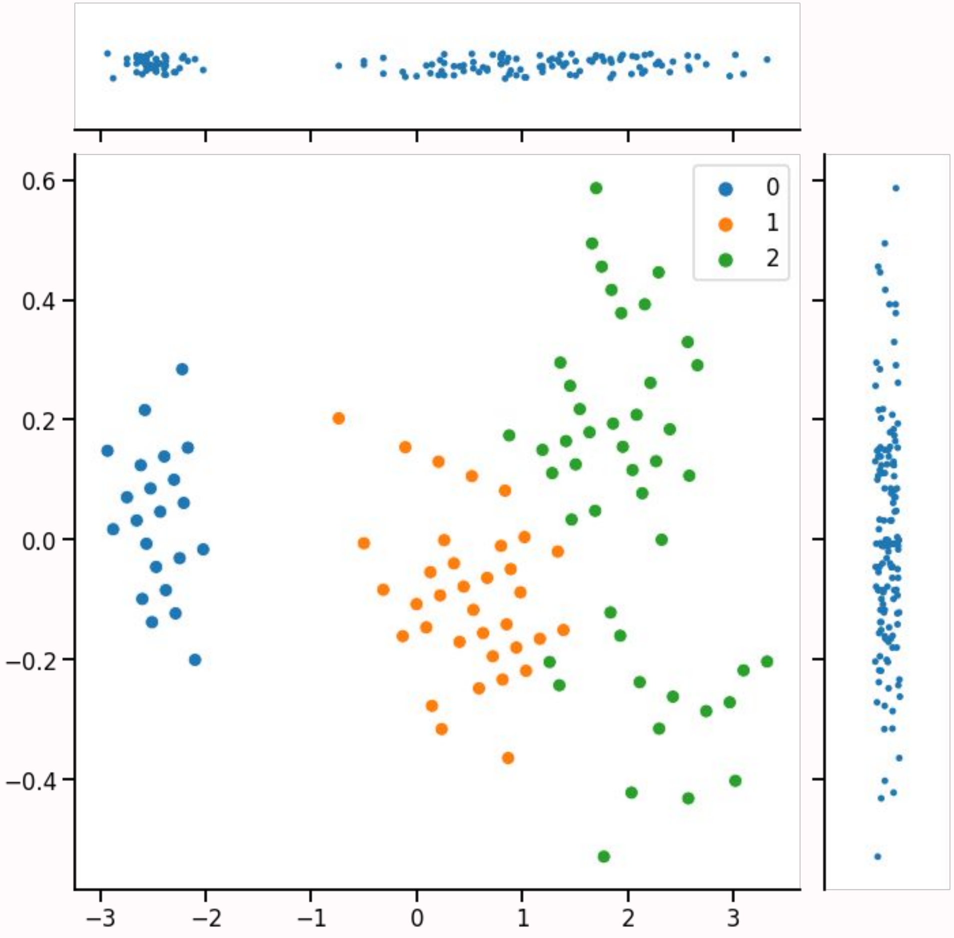
Why don't they "look" orthogonal?



New projections

We project each of the rows in the directions of the principal components.

Note that both representations of the data have exactly the same information.



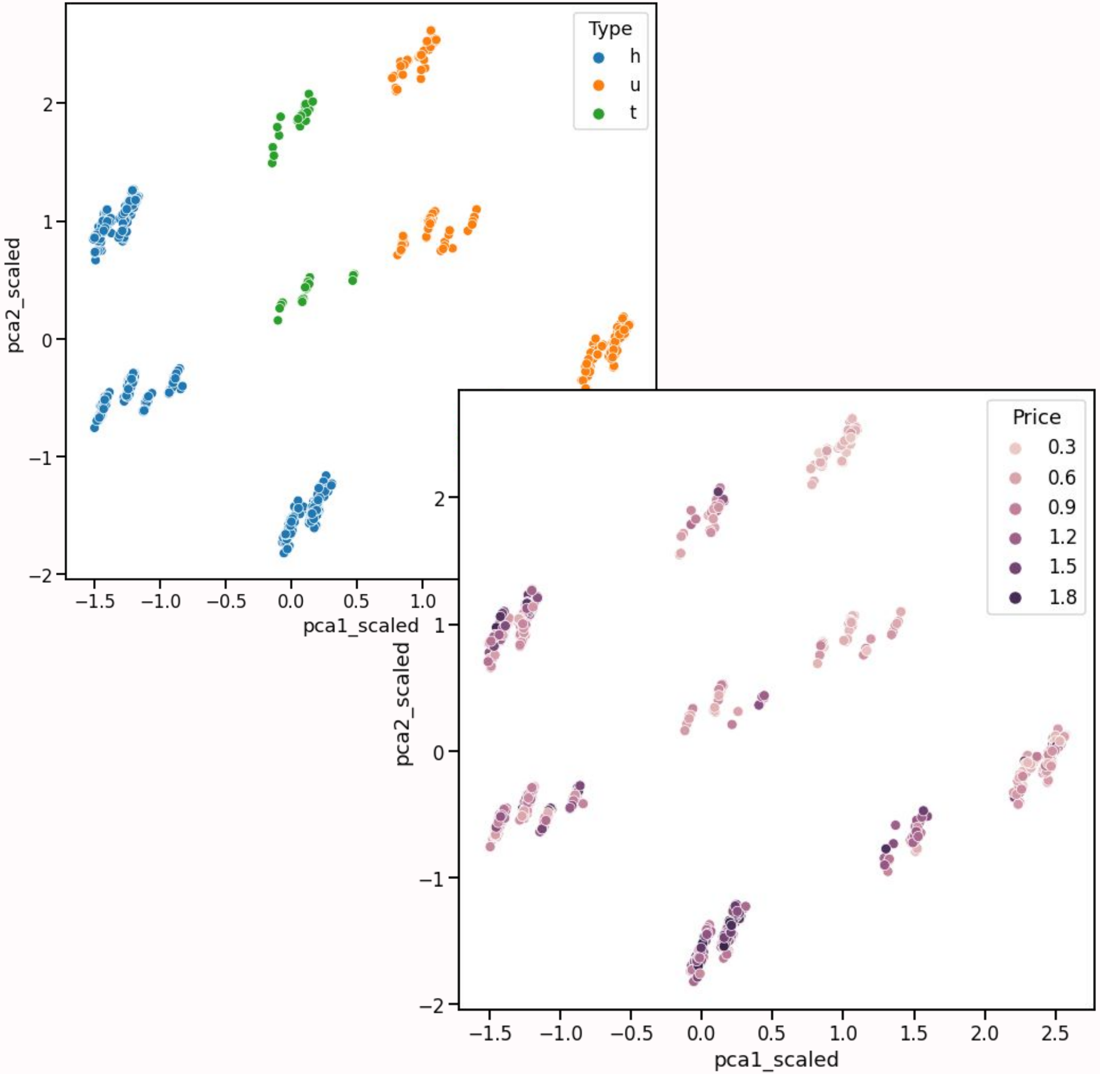
Demo notebook
03_PCA_toy_example.ipynb

Demo notebook 04_encodings_and_PCA_in_Melbourne.ipynb

Result

In the melbourne data set, the main components separate property types very well, and price to a lesser extent.

If the type is closely related to the components of the PCA, does it help us to add this new information?



When we project we change the properties of the data, we want to project in a way that helps understand/classify

Other possible projections

Free text analysis

Suburb	closest_airbnb_neighborhood_overview
Melton South	Close to the CBD, 30-60 minutes from top Victorian beaches and suitable for day trips out to the beautiful Victoria countryside...
Oakleigh	Close to Chadstone Shopping centre, Oakleigh Centro, Walking distance approx 500m to Oakleigh and Huntingdale train station .Bus stops are easily available a couple of streets away...
Balwyn	Filled with gorgeous parks, award winning restaurants and shops and leading Deli's across Melbourne. It's close to the city- 15 minute tram ride into the city or 12 minutes into Richmond...

Text encoding in bags of words

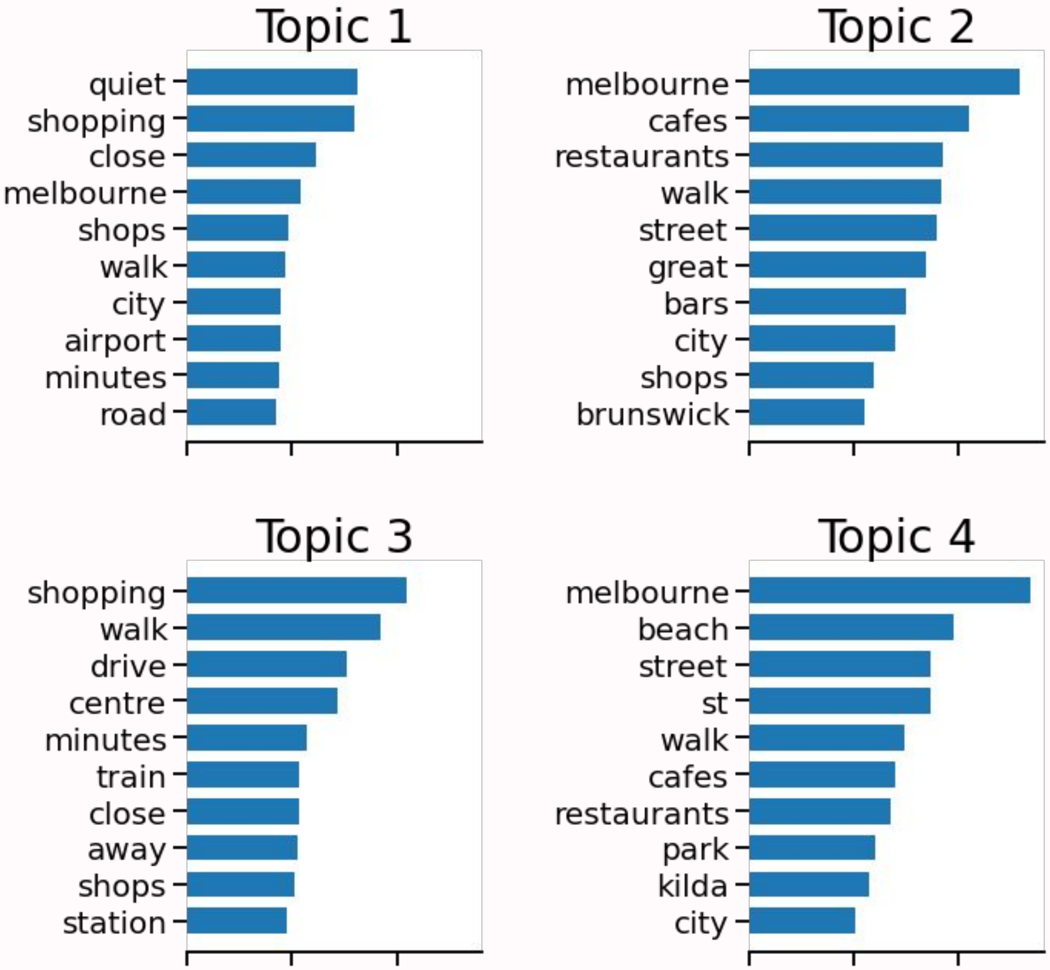
Id	comment
1	No traffic no
2	Near the airport
3	airport traffic
4	Near the beach

Id	no	traffic	near	the	airport	beach
1	2	1	0	0	0	0
2	0	0	1	1	1	0
3	0	1	0	0	1	0
4	0	0	1	1	0	1

Topic modeling with LDA

LDA or Latent Dirichlet Allocation is a model that assumes that each text talks about an unknown subject or topic.

Find the vectors that correspond to the topics that would best explain the data



Projection with LDA

Then, LDA is used to estimate the conditional probability that a text is talking about each of the topics.

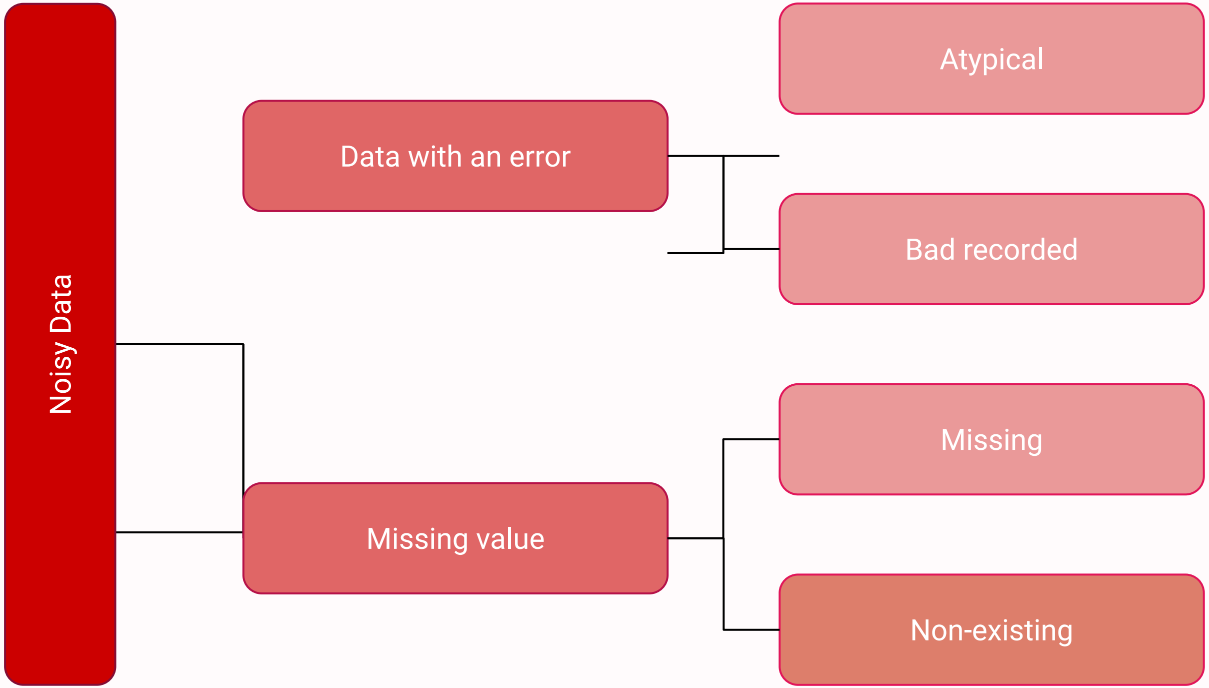
We can now represent each text with a combination of different themes

closest_airbnb_neighborhood_overview	topic0	topic1	topic2	topic3
Our house is located in a very small, quiet and safe court in the bayside suburb of Moorabbin, with no through traffic, so you are undisturbed by traffic noise. The local shopping centre and cafes is 10 minute's walk from the house The large Southland (Westfield) Shopping Centre is 2.6Km away and easily accessible by a bus which is a few minutes walk from our home. Chadstone is a bus ride away. Brighton Beach is 6Km from the house and easily accessed by public transport, where you can enjoy a walk or swim, or a meal of fish and chips on the foreshore.	0,001	0,001	0,934	0,062

Demo notebook 05_encodings_for_text_and_LDA.ipynb

Missing Values

Missing values or noisy data



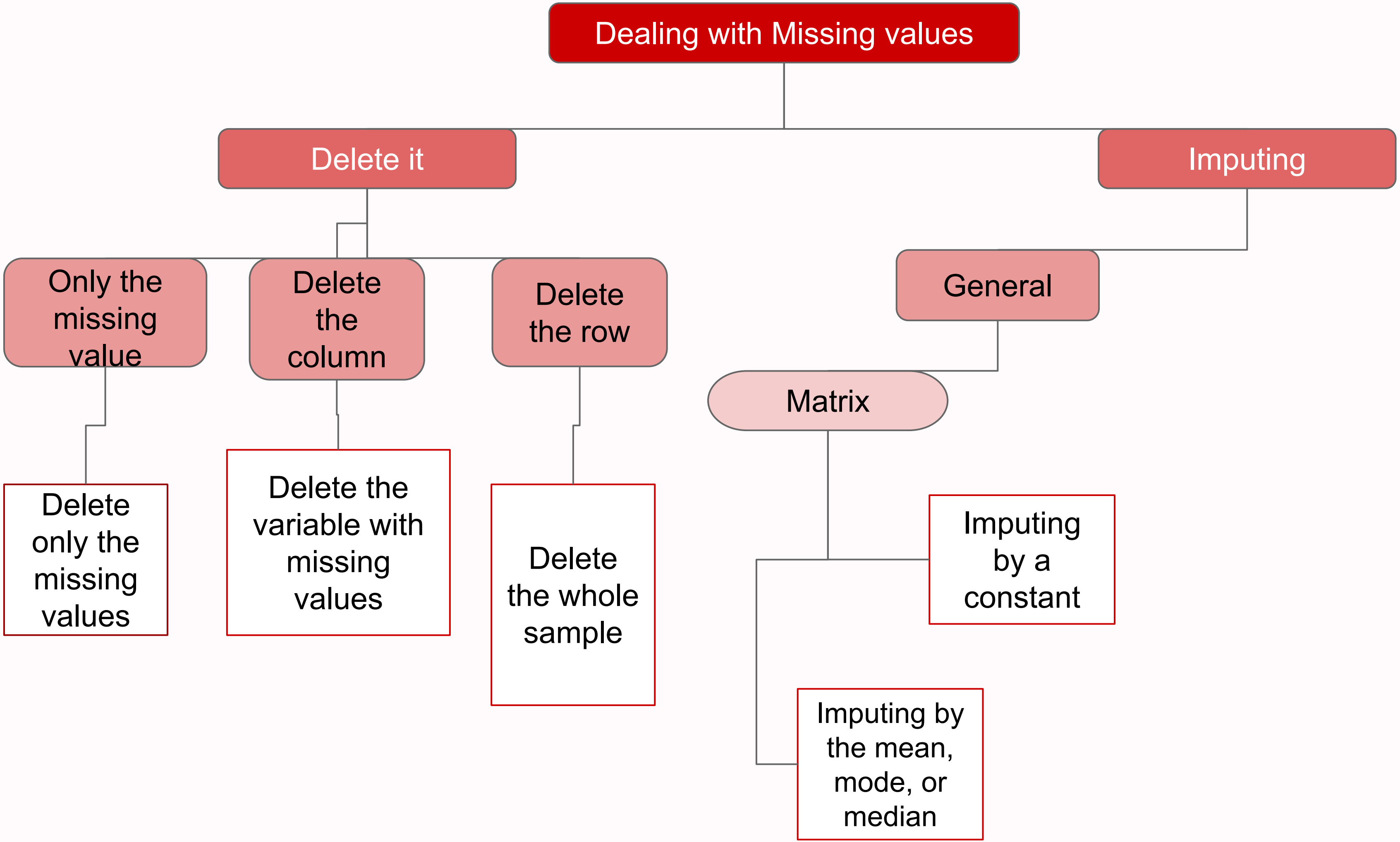
Missing values

In statistics

- to predict is to give value to data that has not yet been sampled,

- to impute is to estimate a value that may have been sampled but is not known.

If one manages to make a prediction model with the data that does not have noise … we impute the missing values by means of predicting that data.


Dealing with Missing values
Delete it
Imputing
Only the missing value
Delete the column
Delete the row
General
Matrix
Delete only the missing values
Delete the variable with missing values
Delete the whole sample
Imputing by a constant
Imputing by the mean, mode, or median

Some useful links

- Scikit-learn tutorial on different types of decompositions
- Video about PCA.