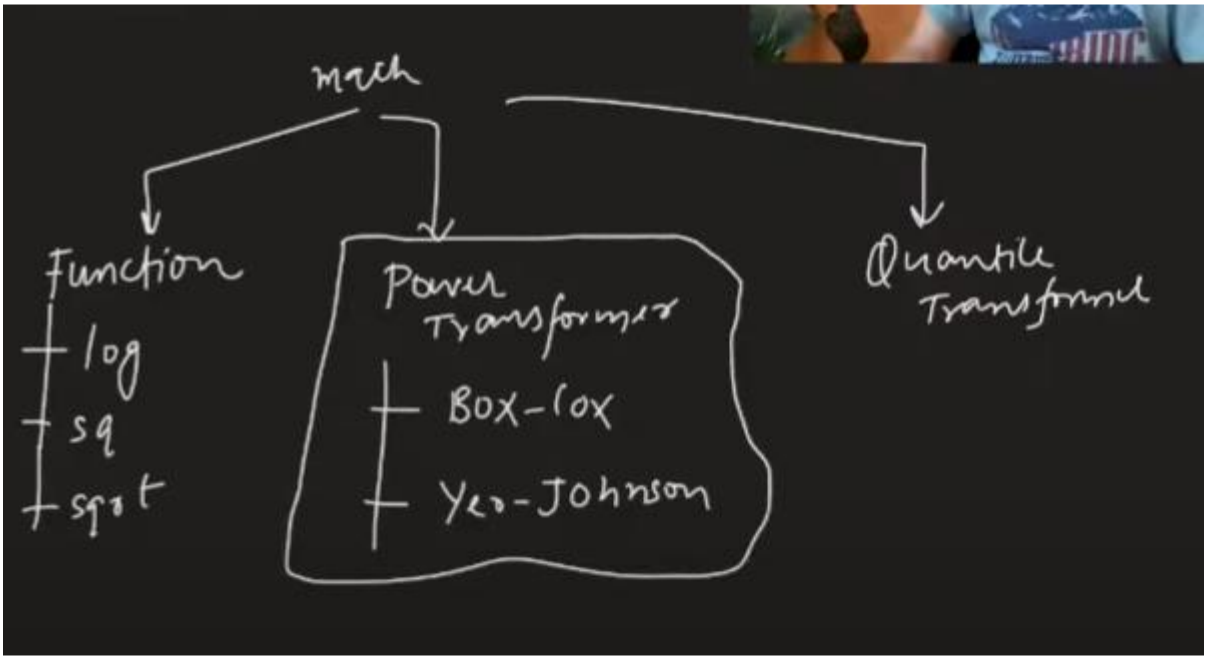
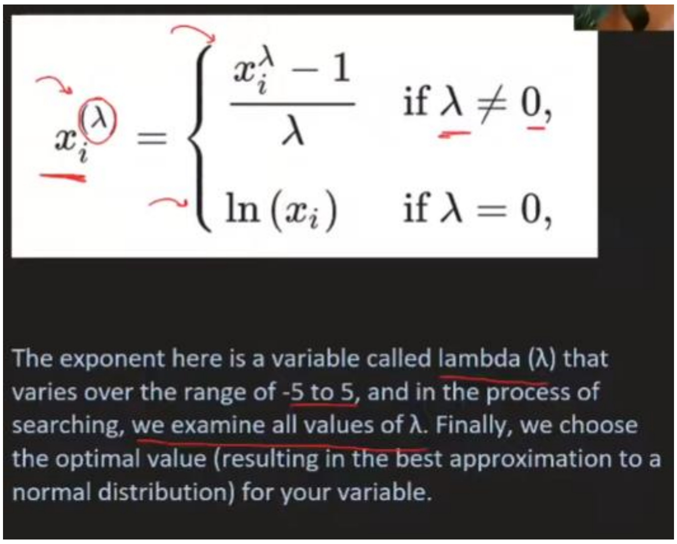
Machine Learning

Video 31:

Power Transformer | Box - Cox Transform | Yeo - Johnson Transform:



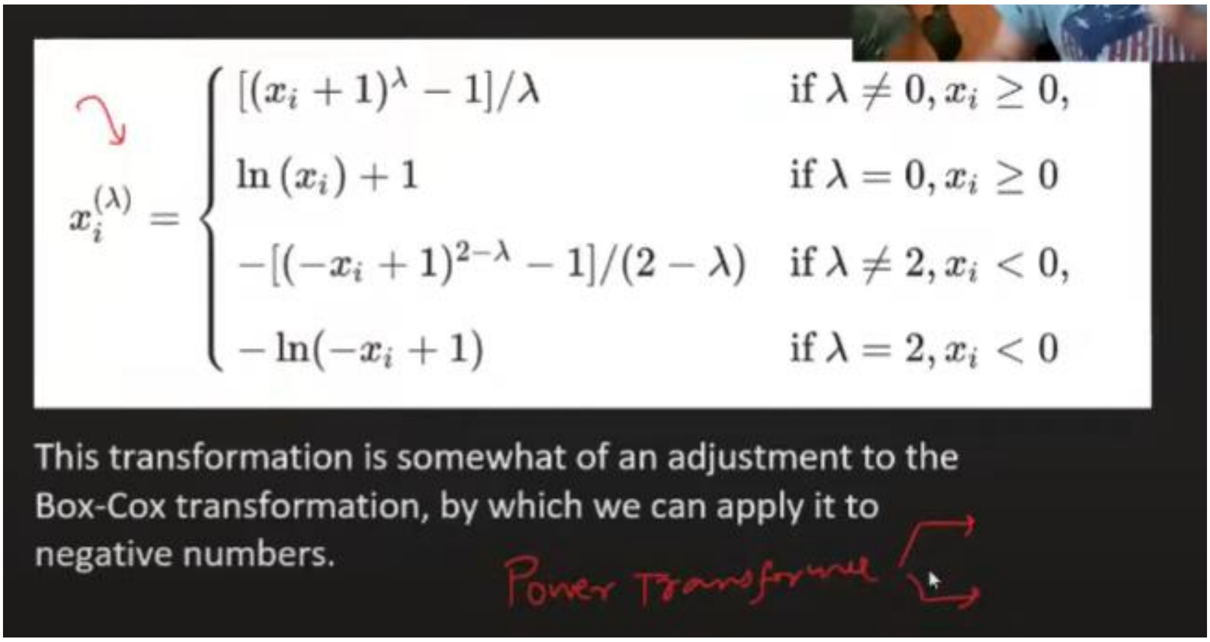
What is Box cox transform? It is applicable for n > 0



$$x_i^{(\lambda)} = \begin{cases} \dfrac{x_i^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln(x_i) & \text{if } \lambda = 0, \end{cases}$$

The exponent here is a variable called lambda (λ) that varies over the range of -5 to 5, and in the process of searching, we examine all values of λ. Finally, we choose the optimal value (resulting in the best approximation to a normal distribution) for your variable.

What is Yeo-Johnson transform?



$$x_i^{(\lambda)} = \begin{cases} [(x_i+1)^\lambda - 1]/\lambda & \text{if } \lambda \neq 0, x_i \geq 0, \\ \ln(x_i)+1 & \text{if } \lambda = 0, x_i \geq 0 \\ -[(-x_i+1)^{2-\lambda}-1]/(2-\lambda) & \text{if } \lambda \neq 2, x_i < 0, \\ -\ln(-x_i+1) & \text{if } \lambda = 2, x_i < 0 \end{cases}$$

This transformation is somewhat of an adjustment to the Box-Cox transformation, by which we can apply it to negative numbers.

Example:

Code link:

https://colab.research.google.com/drive/1S6nWYwPwM5nXlFf0h0pYaaDFhW35kRe2?usp=sharing

Data link:

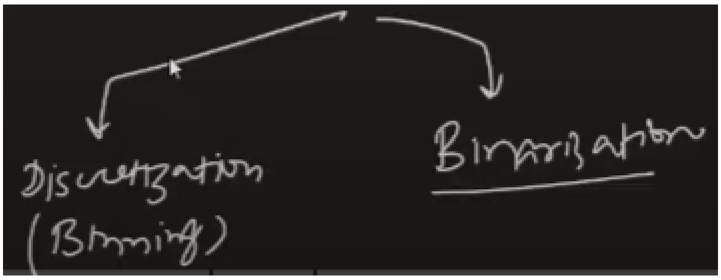
https://github.com/campusx-official/100-days-of-machine-learning/tree/main/day31-power-transformer

Video 32:

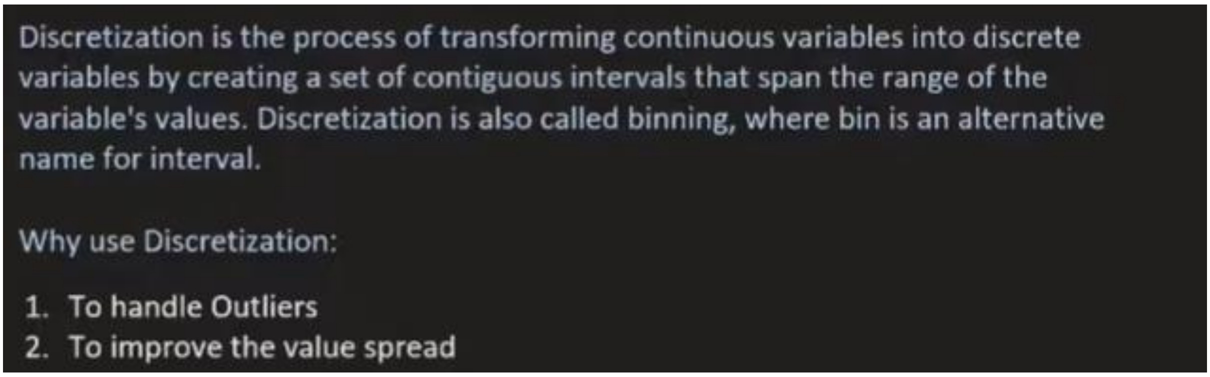
Binning and Binarization | Discretization | Quantile Binning | KMeans Binning:

How to Encode numerical features?

Method to do so:



What is Binning?

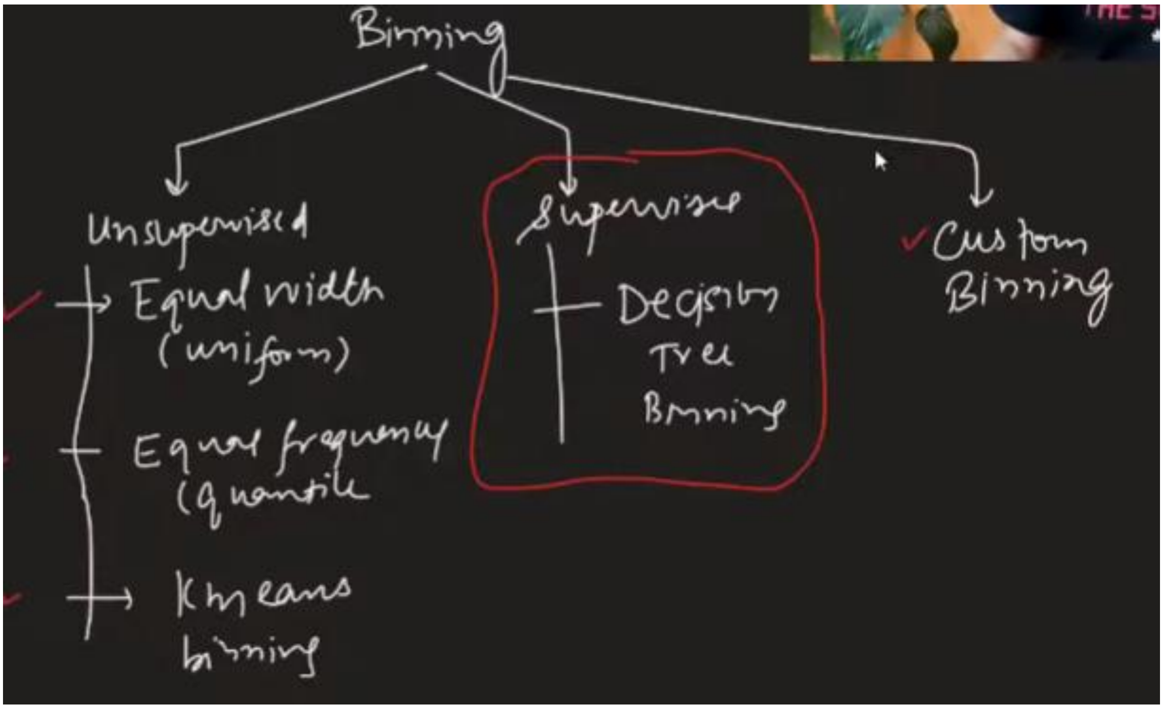


Discretization is the process of transforming continuous variables into discrete variables by creating a set of contiguous intervals that span the range of the variable's values. Discretization is also called binning, where bin is an alternative name for interval.

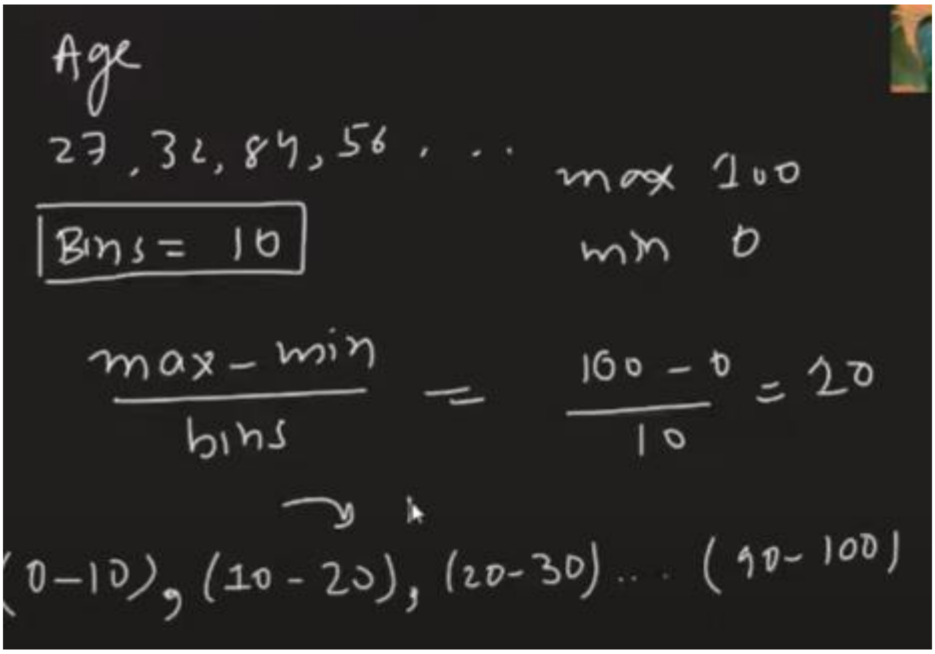
Why use Discretization:

1. To handle Outliers
2. To improve the value spread

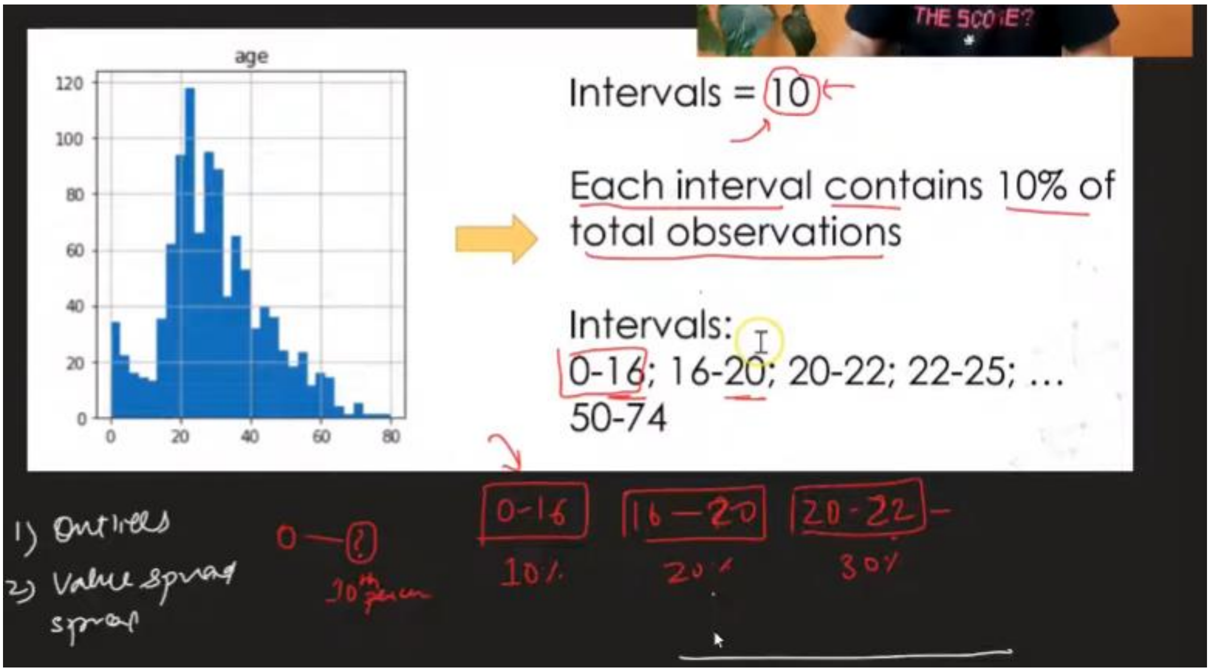
Types of binning?



What is Equal width / uniform binning?



What is Equal frequency / quantile binning?



Intervals = 10

Each interval contains 10% of total observations

Intervals:
0-16; 16-20; 20-22; 22-25; …
50-74

1) Outliers
2) Value spread
spread

0 — ?
10th percen

0-16 | 16-20 | 20-22 -
10% | 20% | 30%

What is K means binning?

It makes clusters.

Example:

Code link:

https://colab.research.google.com/drive/1S6nWYwPwM5nXlFf0h0pYaaDFhW35kRe2?usp=sharing

Data link: https://github.com/campusx-official/100-days-of-machine-learning/tree/main/day32-binning-and-binarization

What is binarization? Special case of discretization.

We convert a continuous value into binary.

Video 33:

Handling Mixed Variables | Feature Engineering:

What is mixed data?

In machine learning, mixed data refers to datasets containing both numerical (e.g., age, salary) and categorical (e.g., gender, color) variables. Handling mixed data requires preprocessing techniques like normalization for numerical data and encoding for categorical data, ensuring that both types of features can be effectively used in models.

Example:

Code link:

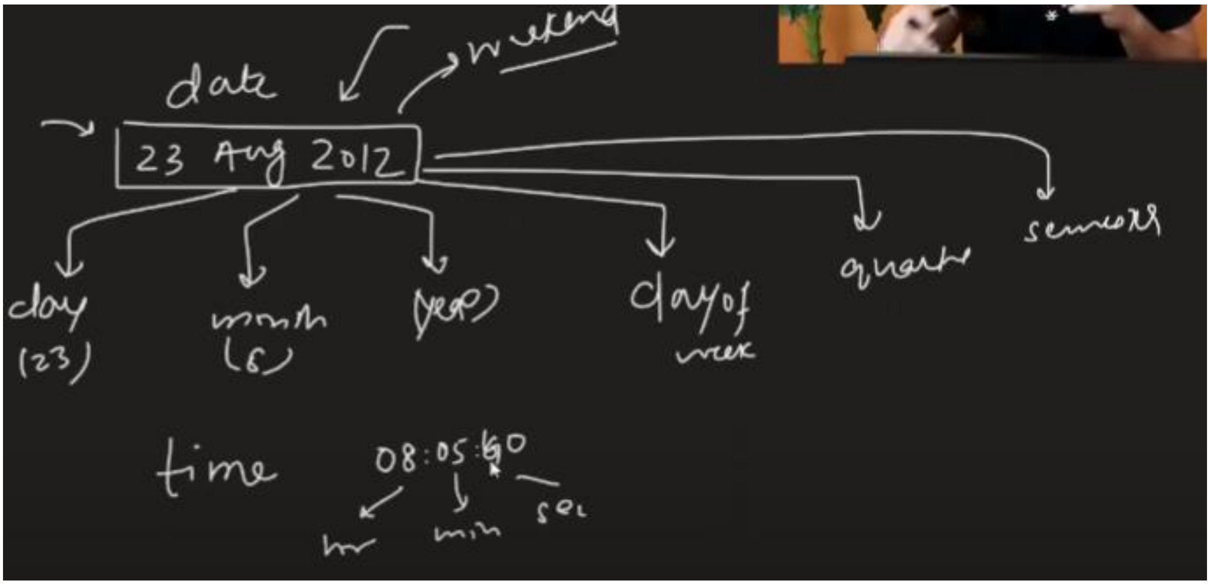
https://colab.research.google.com/drive/1DtLc0S6D1lXGxA_zo0FEHeLbpsiMHxZ8?usp=sharing

Data link:

https://github.com/campusx-official/100-days-of-machine-learning/tree/main/day33-handling-mixed-variables

Video 34:

Handling Date and Time Variables:



Example:

Code link:

https://colab.research.google.com/drive/1DtLc0S6D1lXGxA_zo0FEHeLbpsiMHxZ8?usp=sharing

Data link:

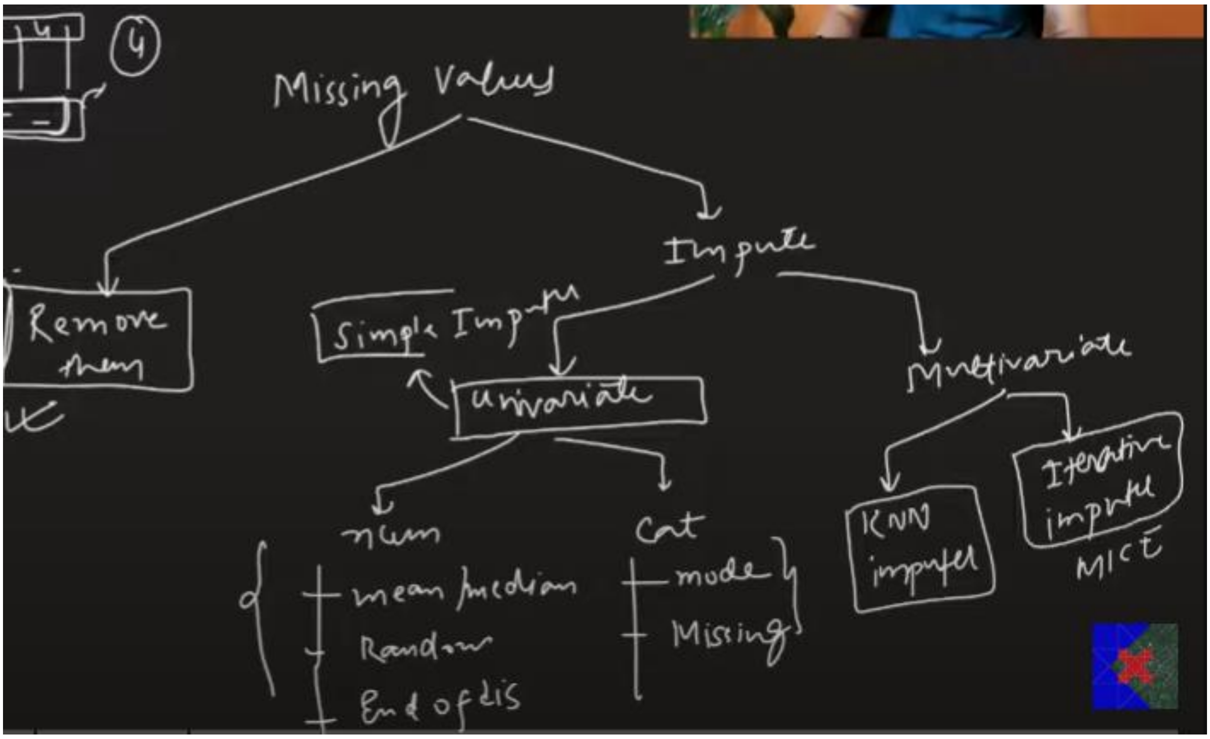
https://github.com/campusx-official/100-days-of-machine-learning/tree/main/day34-handling-date-and-time

Video 35:

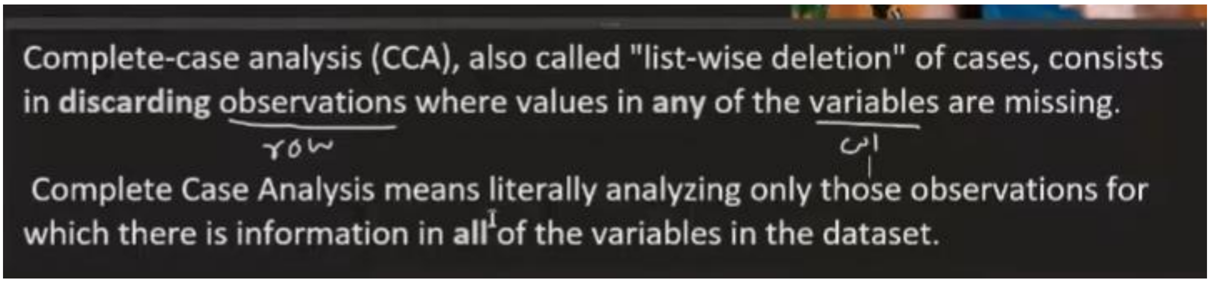
Handling Missing Data: Part 1

What to do when we have missing data?

1. Remove them – not much preferred
2. Impute – to fill
   a. Univariate
   b. Multivariate



What is Complete-case analysis?



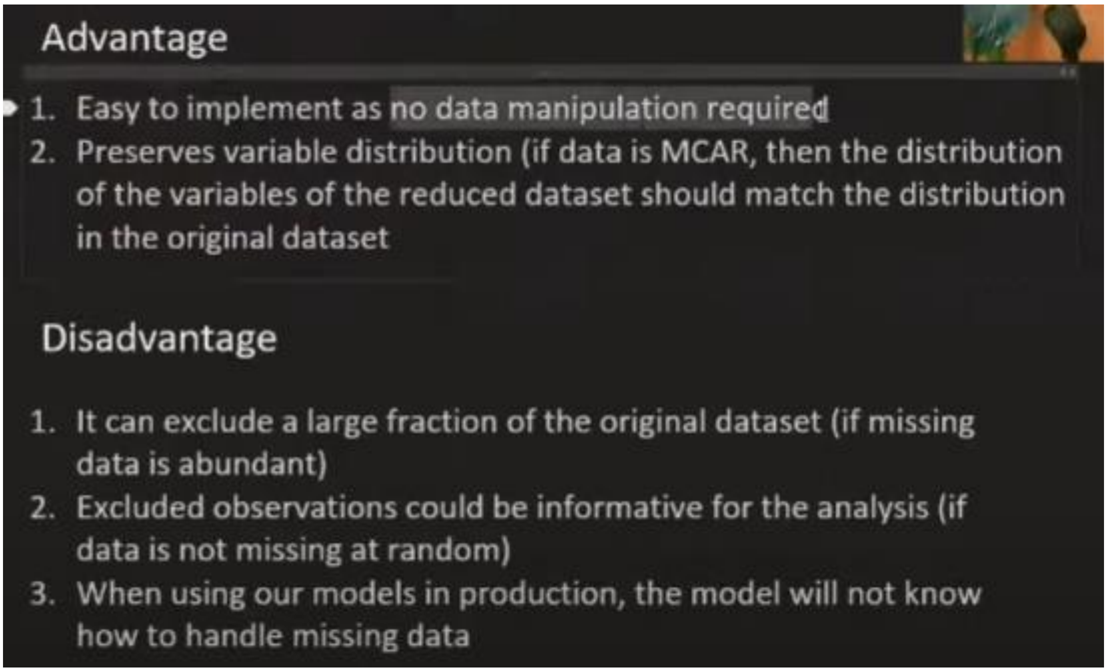
Complete-case analysis (CCA), also called "list-wise deletion" of cases, consists in discarding observations where values in any of the variables are missing.

Complete Case Analysis means literally analyzing only those observations for which there is information in all of the variables in the dataset.

Assumptions for CCA:

1. Data will be missing completely at random (MCAR).

Advantages and disadvantages of this method:



Advantage

1. Easy to implement as no data manipulation required
2. Preserves variable distribution (if data is MCAR, then the distribution of the variables of the reduced dataset should match the distribution in the original dataset

Disadvantage

1. It can exclude a large fraction of the original dataset (if missing data is abundant)
2. Excluded observations could be informative for the analysis (if data is not missing at random)
3. When using our models in production, the model will not know how to handle missing data

When to use CCA?

1. MCAR
2. Less than 5% of data is missing

Example:

Code link:

https://colab.research.google.com/drive/11JZVpyUxzHKk_Mdec1e-XLdPQo2Xws5K?usp=sharing

Data link:

https://github.com/campusx-official/100-days-of-machine-learning/tree/main/day35-complete-case-analysis