Exploratory data analysis

-EDA: investigating, organizing and analyzing datasets -> summarize main characteristics

-6 EDA practices:

- Discovering: familiarize with each column meaning
- Structuring: Transform and organize raw data to be more easily visualized, explained and modelled. It's important to avoid bias (grouping data that does not represent the whole population)
- Cleaning: Remove errors that may distort data or make it less useful
- Joining: augmenting or adjusting data by adding values from other datasets
- Validating: verify that data is consistent and high quality
- Presenting: make your cleaned dataset / data visualizations available to others for analysis or further modelling

-> The process is iterative and non-sequential. A real example:

Visual example

Imagine you are assigned a dataset that has only 200 rows and five columns of data about trees in a coniferous forest in Norway. You know that to complete your full analysis you'll need more than 1,000 rows and at least two more columns. Even without much more detail than that, your entire EDA process might look something like this: