Reference: https://www.kaggle.com/alexisbcook/pipelines

- Definition
  - A simple way to keep your data preprocessing and modeling code organized
  - Bundles preprocessing and modeling steps so you can use the whole bundle as if it were a single step.

- Benefit
  - Cleaner code: don't have to manually keep track of training and validation data at each step
  - Fewer bugs: less likely to forget or confuse a processing step
  - Easier to productionize: more research needed ??
  - More Options for model validation: next tutorial
  - Easier to deal with data contains both categorical data and columns with missing values
- How to construct the pipeline ?
  - Step 1:
    - Define the preprocessing steps using ColumnTransform class; we can bundle different preprocessing step
    - Example: preprocessing for data contains both categorical and missing value
      - imputes missing values in ***numerical*** data, and
      - imputes missing values and applies a one-hot encoding to ***categorical*** data.
  - Step 2: define the model
  - Step 3: create and evaluate the pipeline