the shortcut × moz://a

DATA SCIENCE

TRAINING PROGRAM

Lab 3: Data Manipulation

Larissa Leite, Kodit.io

kodit.io selko SILO.AI

Dataset 1

- Real estate transactions
  https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data
- Register to Kaggle and download
- Features:
  - SalePrice: sale price in dollars (target variable to predict)
  - LotArea: lot size in square feet
  - Utilities: type of utilities available
  - BldgType: type of dwelling
  - YearBuilt: original construction date
  - ...

Lab 3

- Which variables have missing values?
  - Should they be removed? Or replaced?
- Are there duplicate entries in the dataset?
- Are there wrong values or mistakes in any of the variables?
- Numerical data
  - Which variables can be scaled or standardized?
  - Which variables can be discretized?
  - Which variables can be logarithmic or exponentially transformed?
- Categorical data
  - Apply encoding
- Feature selection
  - Any non-informative variables that can be dropped?
  - Any highly correlated variables that can be dropped?
  - Any variables that you would **create** based on (derive from) the existing ones?