ML types

* supervised
  + regression: (linear, polynomial) housing price prediction, market forecasting
  + classification: (logistic regression, decision trees, knn) fraud detection, image classification, diagnostics
* unsupervised:
  + clustering: (k-means) targetted marketing
  + association
  + dimensionality reduction
* semi-supervised (hybrid): text classification, lane finding on GPS data
* reinforcement: optimized marketing, driverless cars, robot navigation, game AI

Required codes

* tree -> DecisionTreeRegressor
* ensemble -> RandomForestRegressor
* preprocessing -> OrdinalEncoder
* preprocessing -> OneHotEncoder
* compose -> ColumnTransformer
* pipeline -> Pipeline
* impute -> SimpleImputer
* model\_selection -> cross\_val\_score
* metrics -> mean\_absolute\_error
* xgboost -> XGBRegressor

Missing values

* drop columns, rows (drop, dropna, isna)
* imputation (fillna)
* extension to imputation

Categorical variables

* drop
* ordinal encoding: assign a value like rating
* one-hot encoding

Pipelines (cleaner code, fewer bugs, easier to transform a model to something deployable at scale (prototype)

* better model validation: cross validation

Cross validation

XGBoost: gradient boosting

Data leakage

* target leakage
* train-test contamination