DATA PREPROCESSING: HANDLING MISSING VALUES

Something you have to deal, will need to apply human intelligence to solve

pandas - isnull().sum() - for each column and row
- can dropna() on the dataframe (rows)
- can drop columns dropna(axis=1)
  - can apply thresholds, limits on drops, etc

Rows: Too much data lost: overfitting, wider confidence intervals, ect.
May bias sample

Columns: May lose information on features (underfitting)

Before dropping or imputing missing values, ask:
- What are the mechanisms causing the missing values?
- Are these missing values missing at random?
- Are there rows or columns missing you are not aware of?

Missing at random? Impute
estimate using mean, median, most frequent (categorials), other val from business context

sklearn Imputer

Advanced Methods - Imputing Missing Values
MICE - multiple imputation by chained equations
sklearn.impute.MICEImputer (v0.20)

Python fancyimpute package
KNN impute
Soft Impute
MICE
etc