Guide for Data Preparation and Cleaning

1- Remove duplicate rows if any

2- Remove columns that have the same single value

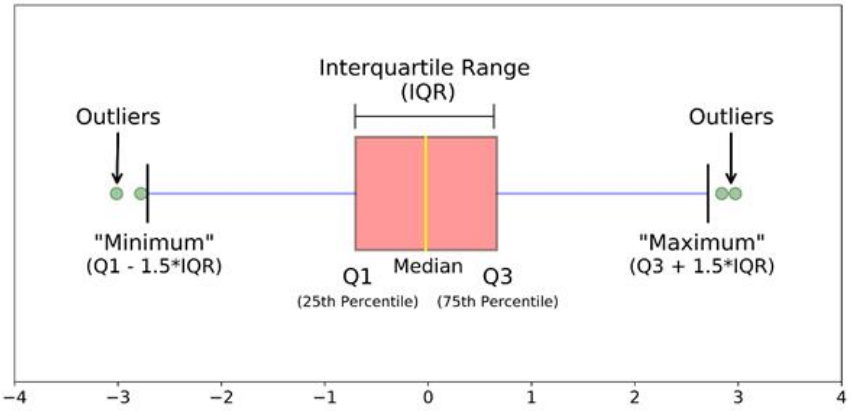
3- Check numerical columns that have very few unique values: drop or transform to categorical

4- Normalize highly skewed columns with log/power transform or Box-Cox

np.log(x) or np.log(1+x) if the column has 0 values

5- Deal with outliers:

Remove if enough data or replace values with the upper bound and the lower bound.



6- Deal with missing values:

- Drop all records if there is enough data
- Keep as NaN if missing values represent more than 60% of the observations
- Fill in with a random value
- Fill in with the mean or the median for numerical data, and the mode for categorical data
- Nearest neighbor: fill in with similar data points
- Heuristic-Based: make a reasonable guess based on knowledge of the underlying domain
- Interpolation: use a method like linear regression to predict the corresponding value

Data imputation: fill missing values

Advanced data imputation: fill missing values and add a new Boolean column that shows the location of the imputed entries.

Bed	Bath
1.0	1.0
2.0	1.0
3.0	2.0
NaN	2.0

Bed	Bath	Bed_was_missing
1.0	1.0	FALSE
2.0	1.0	FALSE
3.0	2.0	FALSE
2.0	2.0	TRUE