Machine Learning Preprocessing Decisions

Prepared by: Farah Ahmed

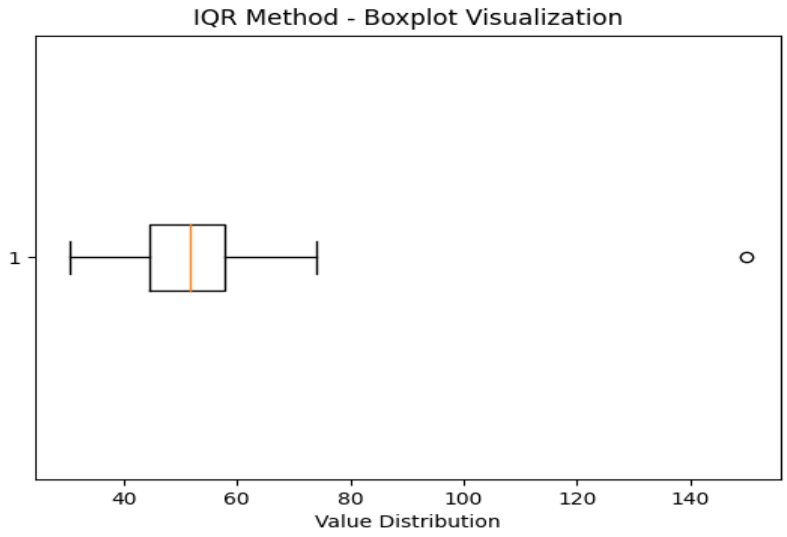
This document explains key preprocessing decisions including median vs mode imputation, IQR capping for outliers, and feature engineering strategies.

1. Why Median vs Mode?

Median is preferred for numerical data when distributions are skewed or contain outliers, because it is robust and not affected by extreme values. Mode is used for categorical variables since categories do not have numerical meaning and the most frequent value preserves dataset distribution.

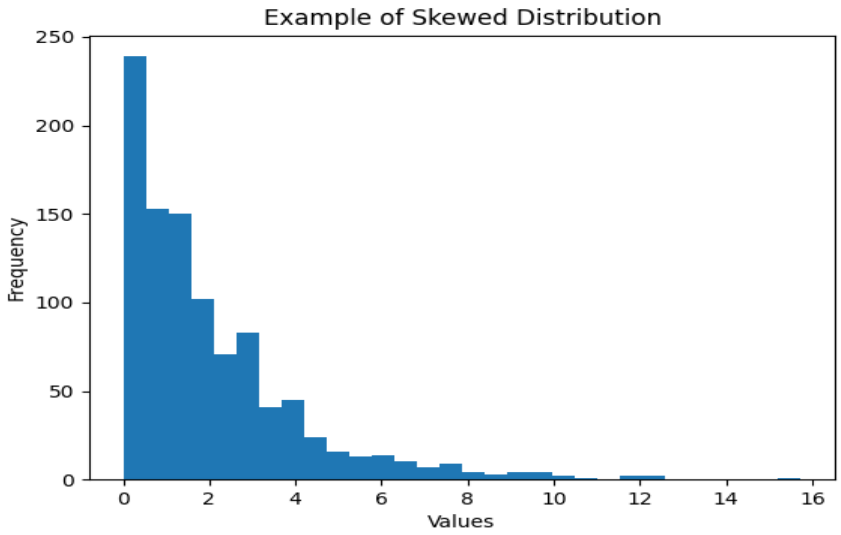
2. Why IQR Capping?

The IQR method identifies outliers using quartiles and does not assume normal distribution. Instead of removing extreme values, capping limits them to boundary thresholds to reduce distortion while preserving data.



3. Skewness and Why Median Works Better

In skewed distributions, extreme values stretch the mean toward the tail. The median remains centered and better represents the typical value. This histogram demonstrates a right-skewed distribution.



4. Feature Engineering Decisions

Engineered features included: Price per Square Meter (Price/Size) to normalize property value; Age Categories to reflect buyer perception; and Total Rooms to better represent usable space. These transformations improve predictive power and interpretability.