Summary of Data Processing:

1. **Handling missing data**:
    a. communityAverage: fill null value with median
    b. constructionTime: fill null value with ‘0000’
    c. Important column/ data mismatch: drop null value row
    d. Since there are many unknown values, for those that should be in binary(1 for true and 0 for false), we changed from unknown to 2, also, for columns in need of a number, we changed them from unknown to the median value.
    e. For the column of district, there are unknown values, we assigned it into the district of 13, which represents an unknown district.

2. **Changing the unit of the columns**, since most data presented in the table are in dollars, changed those columns that are having unit as thousand dollars.

3. For machine learning, we created a new column for training of the model. Specifically, according to the price of the house, classify it into different ranges, such as 35000 would be classified into ‘30k_to_60k’.

4. **One-hot encoding**, to convert categorical variables, which are variables that have discrete, non-numeric values, into a binary representation that can be easily processed by machine learning algorithms. In this case,we did one-hot encoding for columns tradeYear','tradeMonth','floorLevel', 'buildingType', 'renovationCondition', 'buildingStructure' , 'fiveYearsProperty','elevator','subway','district'.

5. **Normalization** of numeric attributes can ensure all attributes are of equal importance during learning. In this case,we did normalization in columns ‘square', 'livingRoom','drawingRoom','kitchen','bathRoom', 'floor', 'ladderRatio','communityAverage'

6. **Feature selection:**
    We did the feature engineering, and the following is the feature we selected:
        price, square, livingRoom, drawingRoom, kitchen, bathRoom, floor, constructionTime, ladderRatio, communityAverage, priceRange, tradeYear, tradeMonth, floorLevel, buildingType, renovationCondition, buildingStructure, fiveYearsProperty, elevator, subway, district