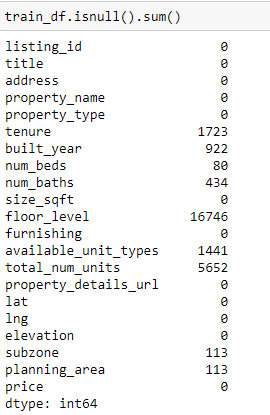
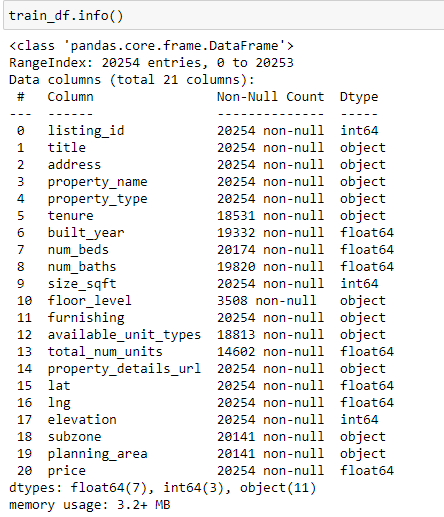
## Motivation

Our primary responsibilities for task 1 are those related to data preprocessing, such as handling missing data, categorical data transformation, and feature extraction. In the meantime, we concentrate on feature extraction (data mining) on the auxiliary dataset to gather additional data that could enhance the model's prediction performance. A few of our data preprocessing strategies will assume that we will use tree-base models for this task. In task 2, we intend to run recommendation models and evaluate the outcomes of various solutions using the previously extracted and transformed variables from task 1.



## Exploratory Data Analysis & Preprocessing

### Price

Figure X.1 shows that there are two unique price records that are outside of the price distribution. For building the model, we were concerned that the results of these 2 samples would be skewed, and we assumed that such irrational property prices were unlikely to be forecast for actual use. So, we made the decision to remove these two outstanding records. The price distribution of the reset records is presented in Figure X.2. Graphical user interface, application, Teams

Description automatically generated

### Title

The 'title' attribute always contains the following details: “{n bed} {property type} for sale in {location}”. Even though this information may overlap with those of other variables or be irrelevant for modeling, they can still be used to impute missing values or conduct sanity check for crucial attributes like the number of beds and property type. Figure x compares the features taken from the title to the current fields.

Graphical user interface, application

Description automatically generated

### Address

We simply discard this attribute from modeling in order to address the overfitting problem.

### Latitude & Longitude

We primarily concentrated on the incorrect records for latitude and longitude features. Given that the goal is to forecast Singapore house property prices, the latitude and longitude ranges should be close by (1.290270, 103.851959). However, a few records are located outside the Singapore region, and further research has revealed that those records (by title, address) belong to Singapore. Since there aren't many incorrect records (figure x.1), we decide to use Google to find the correct location (figure x.2) and manually amend the entries.

Graphical user interface, application

Description automatically generated