In this project we need to predict real estate prices based on categorical and numerical features, Data preprocessing is the most important part of machine learning projects, it takes about 80% of the whole modeling process. I prepared the data for machine learning models as fellow:

1. **Handle missing data**;

* Drop the feature if it has more than 40% missing records.
* Compute the missing values with mean, median or mode.

1. **Handle the inconsistency and redundant in category’s features;**

* Clean each feature records from duplicate data.
* Unit the records format for each feature.

1. **Handle rare categories;**

* Rare categories are same as outliers in numerical features.
* Remove the rare category if its proportion in data are small otherwise;
* Combine the rare categories of one feature and substitute it with another category (such as others)

1. **Handling outlier in numerical features**

* Handle the outlier carefully based on its important
* Use inter quantile ratio or standard deviation to remove it.

1. **Encode category features**

* Since model understand only number, all the categories should be converted to numbers.
* Use one hot encoding, count frequency or mean encoding (nominal category)

1. There is no correlation between the features, so no **feature selection** step is performed
2. dataset is ready to be used.