Real Estate Problem

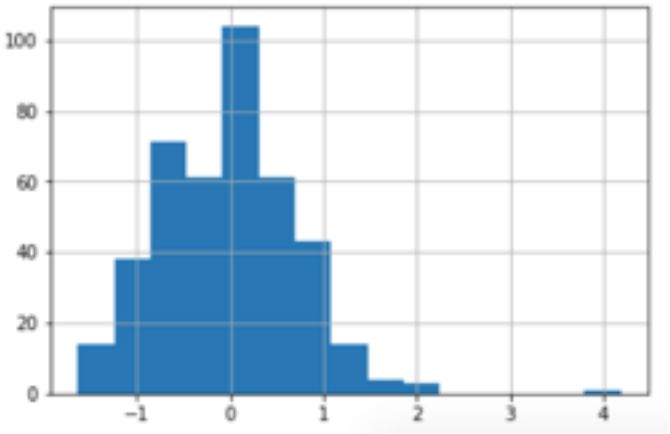
Standardisation :

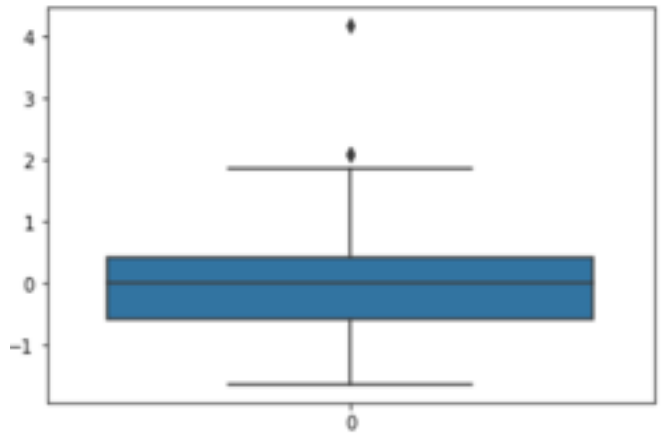
- Standardisation is implemented because the different features in our dataset have different ranges.
- The robust scaler is used for standardisation for this model, because most of the features in the dataset consist of outliers.
- Working of the robust scaler:
  - It scales the data according to the quantile range(IQR)
  - The centring and scaling statistics of this Scaler are based on percentiles and are therefore not influenced by a few numbers of huge marginal outliers.
- Outcome: Standardisation effected the model accuracy only a little.

Bin Count :

The bin count is chosen to be 15, because it gives the most clear and understandable visualisations.

Univariate Analysis for the "house price of unit area" :



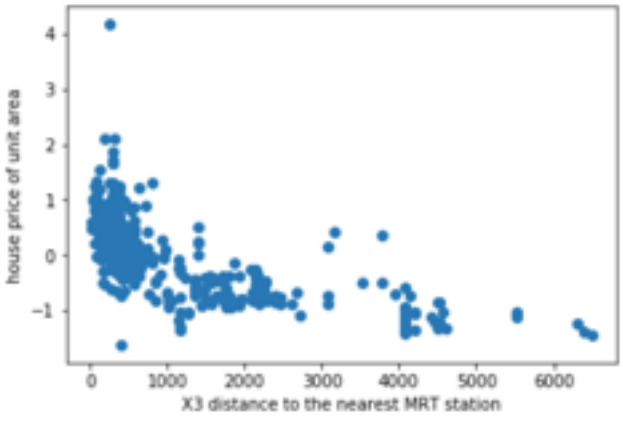


From the above histogram and box-plot, we can conclude that the dataset of "house price of unit area" is:

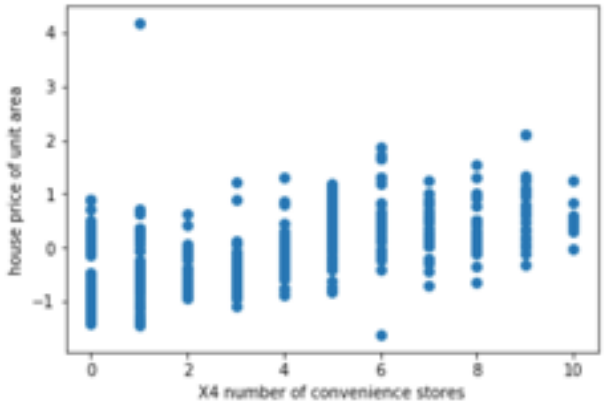
1. **Slightly right skewed**
2. **Has a few outliers**

Bivariate Analysis of different features with respect to the house price of unit area :

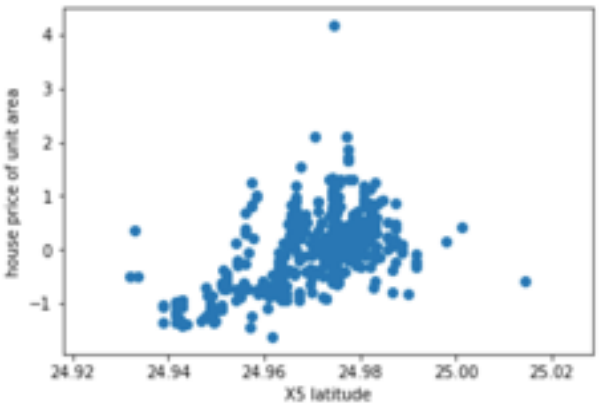
1. Distance to nearest MRT store :



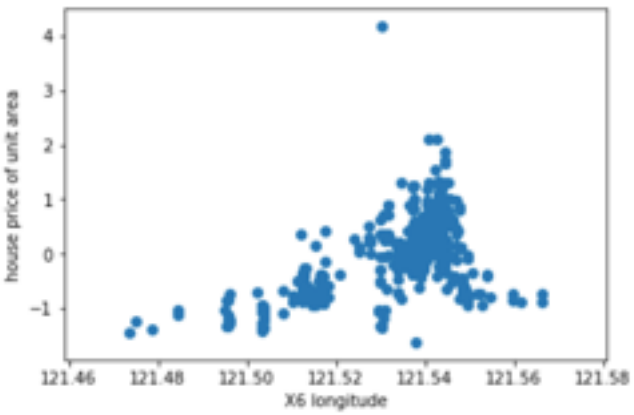
2. Number of convenience stores:



3. Latitude:



4. Longitude:



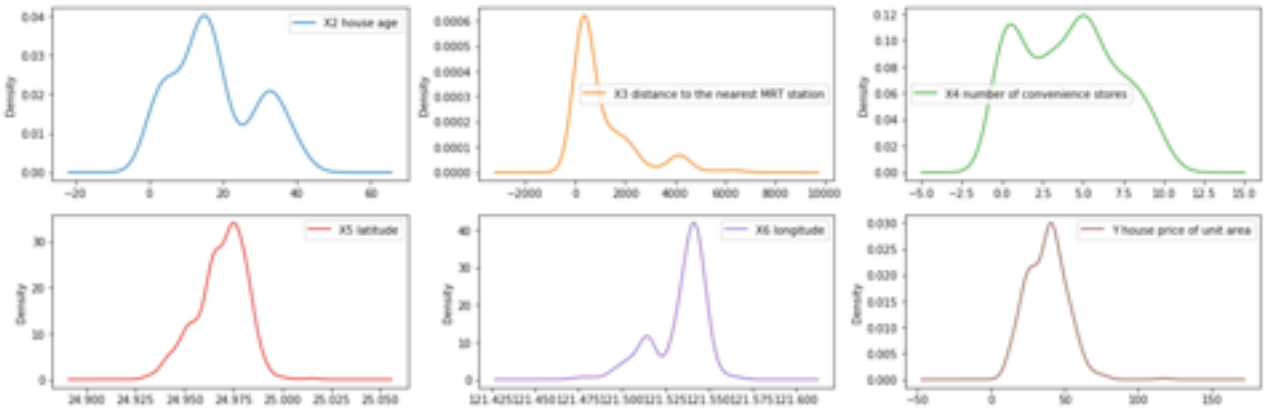
The correlation for the above features relating to their respective scatterplots is given below:

FEATURES	PEARSON CORRELATION
Distance to MRT station	`-0.67361(Negative corr.)`
Number of convenience stores	`0.57100(Positive corr.)`
Latitude	`0.5463(Positive corr.)`
Longitude	`0.52328(Positive corr.)`

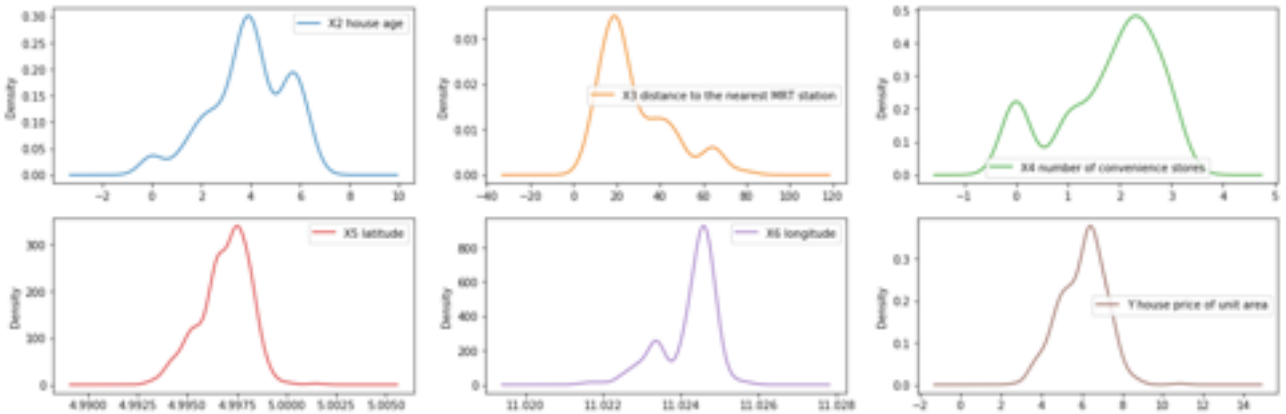
Handling the outliers:

After doing research related to the factors effecting the real estate cost price, it is seen that the outliers in the various features can make sense.
Thus, we cannot completely remove them from the dataset. Even if we do remove the outliers, we get no change in the accuracy of the model.

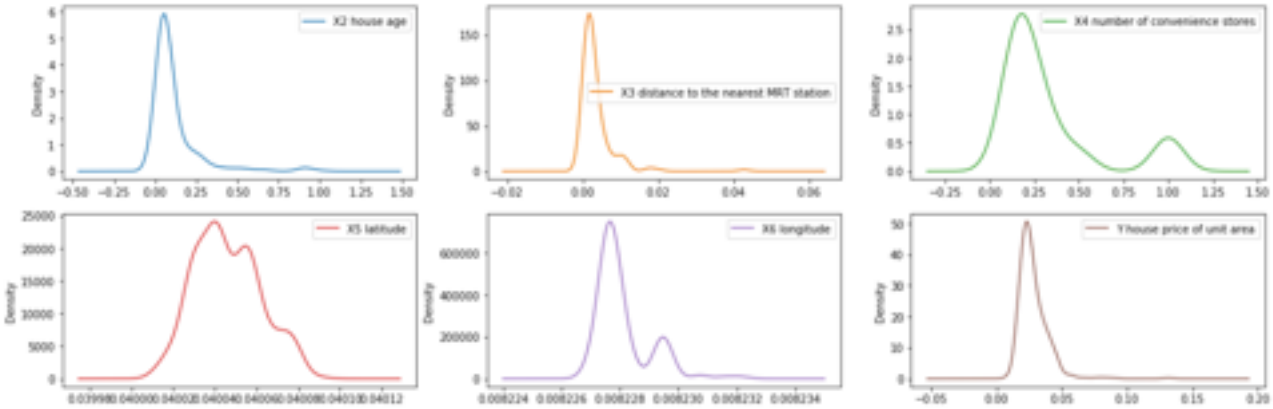
The density plots for all the features BEFORE TRANSFORMATION are given below :



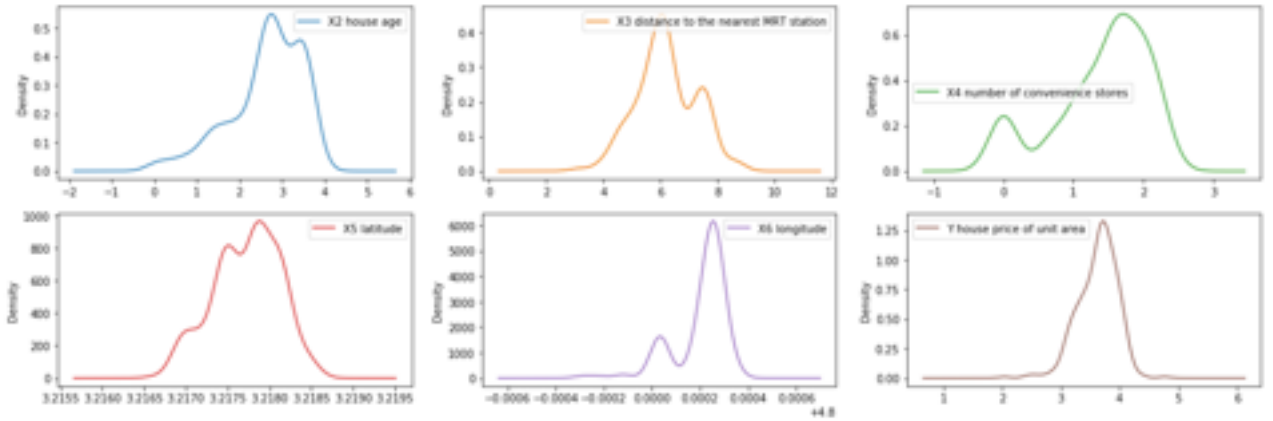
1. Square root transformation :



2. Reciprocal Transformation :



3. Log transformation:



From the above density plots, log transformation is most suited. It also gives an increase in the prediction accuracy of the model.