Exploring the Data Set

Test Data

* both data sets are targeted at predicting the price of houses sold in 2017.
* data is reasonably clean

HDB Training Data

* contains houses sold from early 1990s to 2017
* contains potential features that are categorical as well as numerical
* data appears to have some outliers
* data is reasonably clean

Private Housing Training Data

* contains houses sold from early 1990s to 2017
* contains potential features that are categorical as well as numerical
* contains some outliers (e.g. some houses with large floor area, but low price). These records are removed from training data set manually.
* contains dirty data (e.g. tenure descriptions are incomplete)
* needs to derive new meaningful values (e.g. remaining tenure years)

Figure HDB Training Data

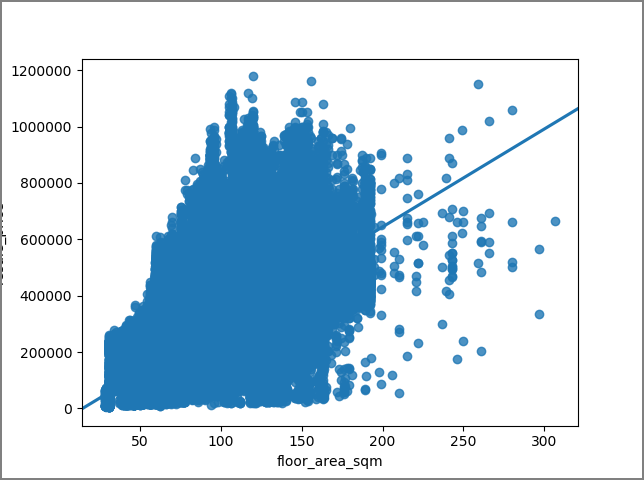
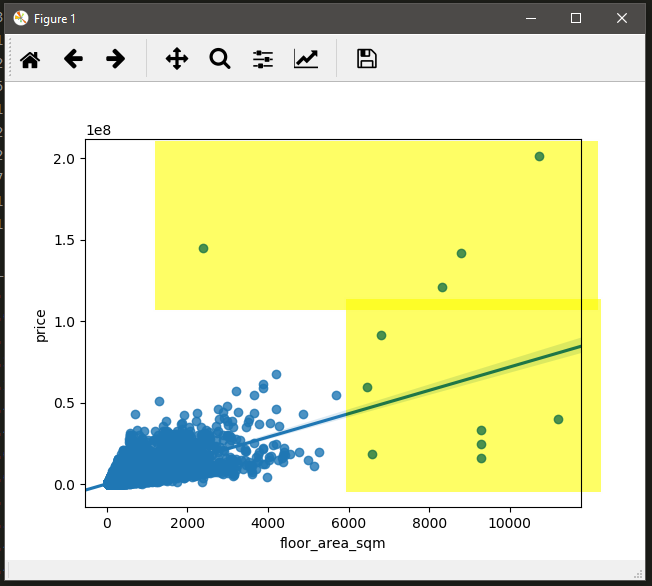


Figure Private Housing Training Data with outliers highlighted



Most correlated features (in descending order)

|  |  |
| --- | --- |
| HDB | Private Housing |
| 1. resale\_price 2. flat\_type 3. floor\_area\_sqm 4. month 5. lease\_commence\_date 6. floor 7. storey\_range 8. longitude 9. latitude 10. town 11. street\_name 12. block 13. postal\_code 14. flat\_model | 1. price 2. floor\_area\_sqm 3. tenure 4. type\_of\_sale 5. property\_type 6. floor\_num 7. address\_block 8. month 9. area 10. contract\_date 11. longitude 12. address\_street (derived from address) 13. project\_name 14. unit\_num 15. completion\_date 16. postal\_district 17. postal\_sector 18. postal\_code 19. latitude 20. region |

Data Preparation

* text values are label-encoded as integer values in both data sets
* block number and street names are derived and label-encoded from address in Private Housing dataset
* tenure years are derived and label-encoded from tenure description in Private Housing

Training Data Selection

* Since test data set is aiming at housing prices in 2017, we decided to train our models using only 2016 and 2017 training data. Our assumption is that buyers and sellers do not need to know old house price. But they need to know house prices that are based on recent house prices in current market so that they can buy or sell at competitive price range.

Feature Selection

* We rely mostly on the correlation characteristics of training data sets.
* For private housing, we also experimented with only following features.
  + Floor\_area\_sqm
  + Latitude
  + Longitude
  + Floor\_num

Training the models

* We chose GradientBoostingRegressor and XGBRegressor algorithms
* Both algorithms undergone Cross Validation Score with 10 KFold
* The resultant models are then ensembled together to predict the housing price
* We wrote two separated Python programs, one for HDB and one for Private Housing
* We then combine the prediction results together