Project 2: HDB Price Estimator

Table of Contents

Problem Statement

In Singapore, a majority 77.9% of the population live in public housing flats. These are priced according to HDB themselves *by establishing the market value of the flat by looking at the prices of comparable resale flats nearby, which is influenced by* ***prevailing market conditions****, as well as the individual attributes of the flats.*

However, this way of pricing can leave first time home buyers paying amounts which are far above what the flat is actually worth due to 'prevailing market conditions'. Examples of such market conditions would be pent-up demand due to COVID, or an increase in prices of construction materials.

We aim to offer first time home buyers a way of getting a good gauge of property prices based on the attributes of the flat and property location, and not exacerbated costs due to external factors so that they can make informed choices on whether now is a good time to purchase the property or if they should wait till external factors are not affecting the price as much.

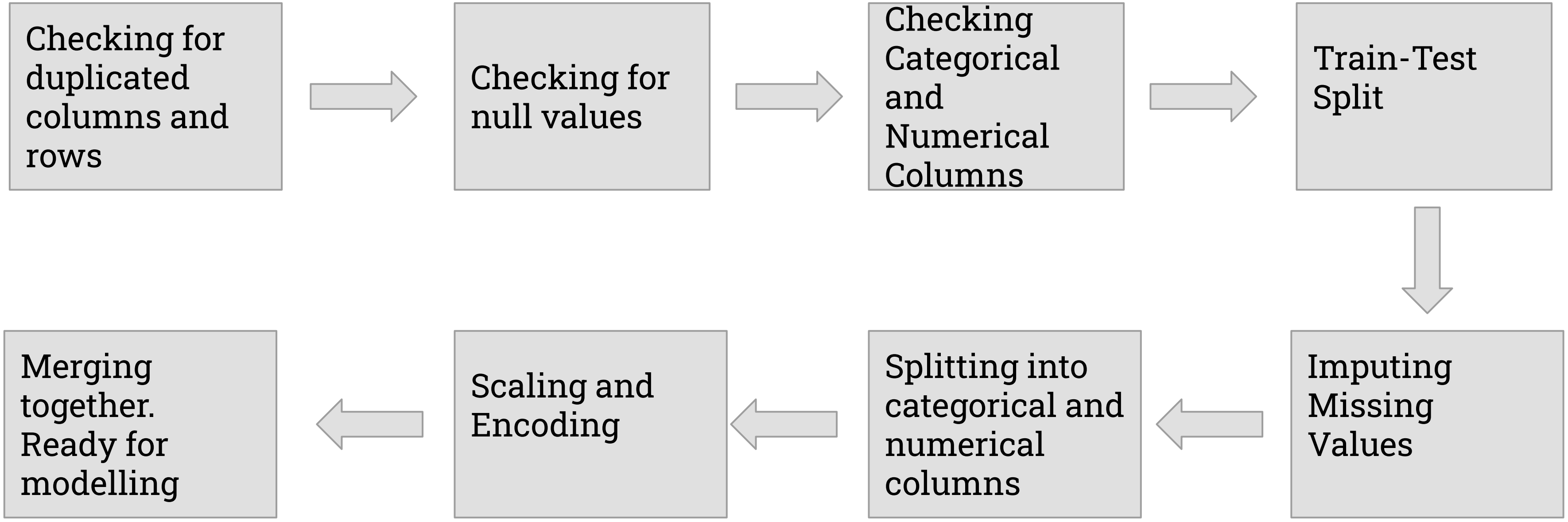
Data Importing & Cleaning

The dataset we used has over 70 features related to Singapore Housing.

These 70 features can be broken down into two main groups:

1. Property Location
    a. Some examples are the number of malls / hawker centres that are nearby the property.
2. Flat Attributes
    a. Some examples are size of the house, highest storey of the building, etc.

Data Importing & Cleaning

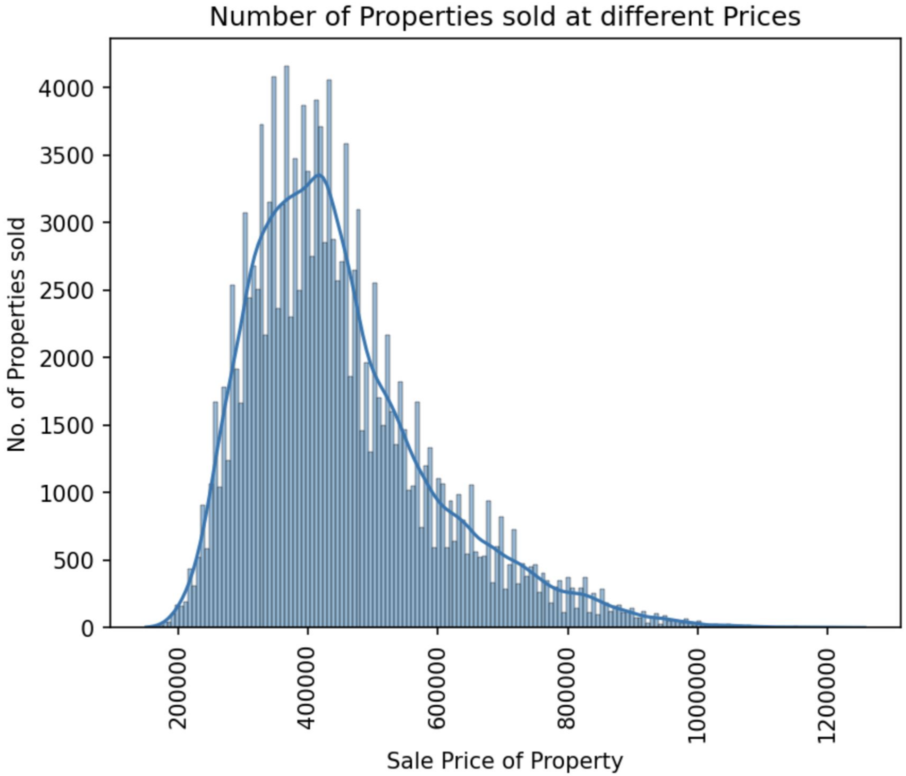


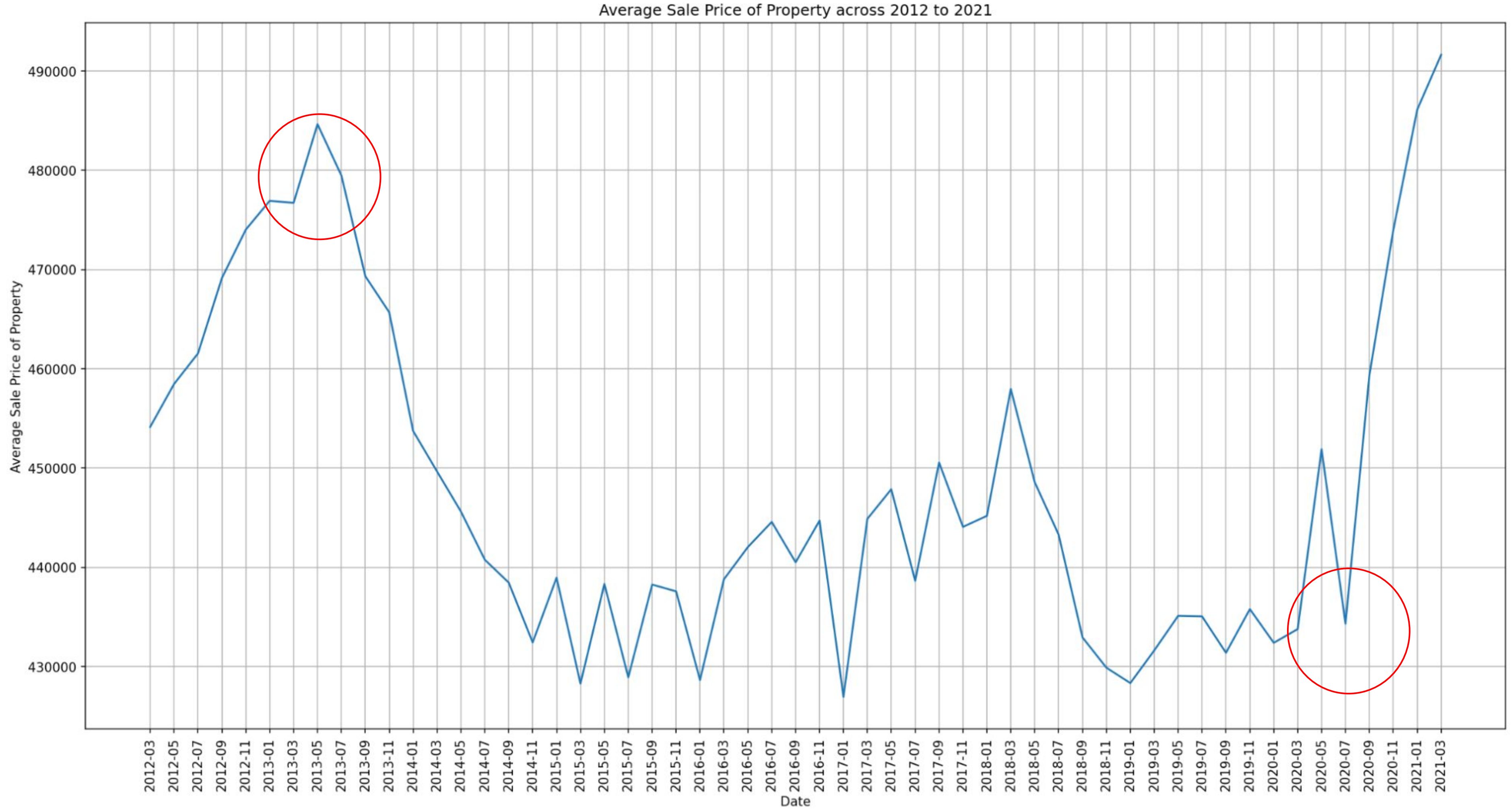
Exploratory Data Analysis

Histogram: Sale Price

1. Sale price is slightly right-skewed
2. Most houses are within price range: 200,000 to 600,000
3. Data suggests that there are expensive houses in data collected
4. Possible outliers in dataset

(HDB, possibly Bishan)



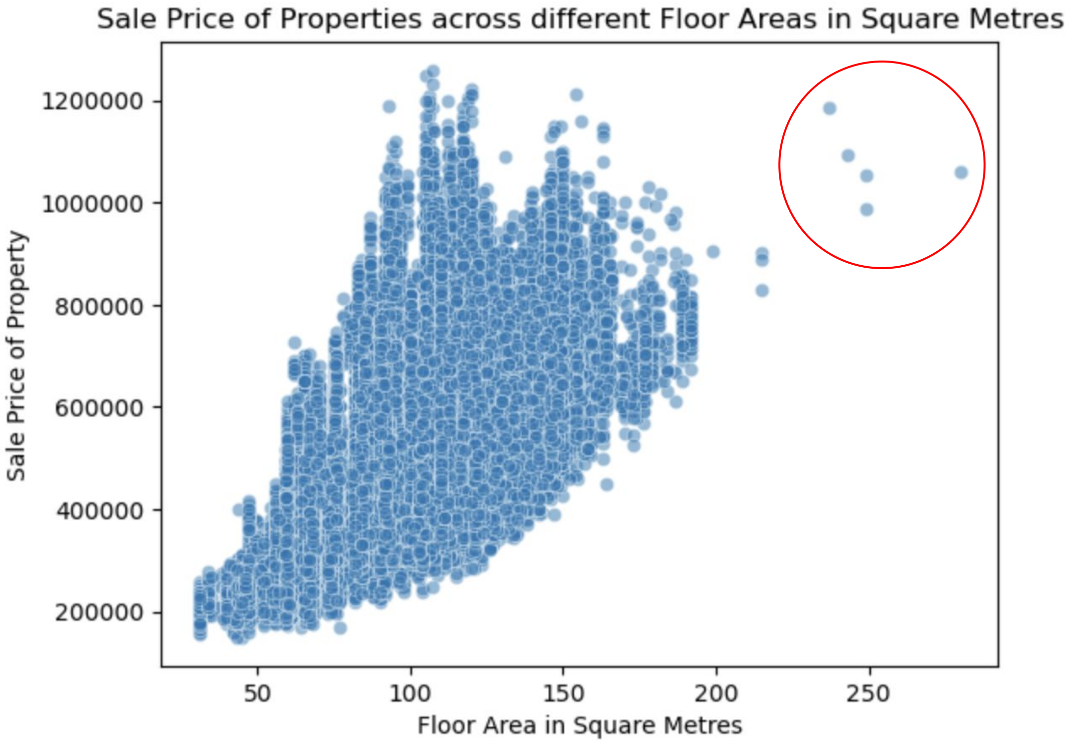


Average Sale Price of Property across 2012 to 2021
Average Sale Price of Property
Date
490000
480000
470000
460000
450000
440000
430000
2012-03
2012-05
2012-07
2012-09
2012-11
2013-01
2013-03
2013-05
2013-07
2013-09
2013-11
2014-01
2014-03
2014-05
2014-07
2014-09
2014-11
2015-01
2015-03
2015-05
2015-07
2015-09
2015-11
2016-01
2016-03
2016-05
2016-07
2016-09
2016-11
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
2018-05
2018-07
2018-09
2018-11
2019-01
2019-03
2019-05
2019-07
2019-09
2019-11
2020-01
2020-03
2020-05
2020-07
2020-09
2020-11
2021-01
2021-03

Scatter Plot: Resale Price vs Floor Area

1. Positive correlation observed between floor area and resale price
2. As number floor area increases, price of property increase
3. Possible Outliers

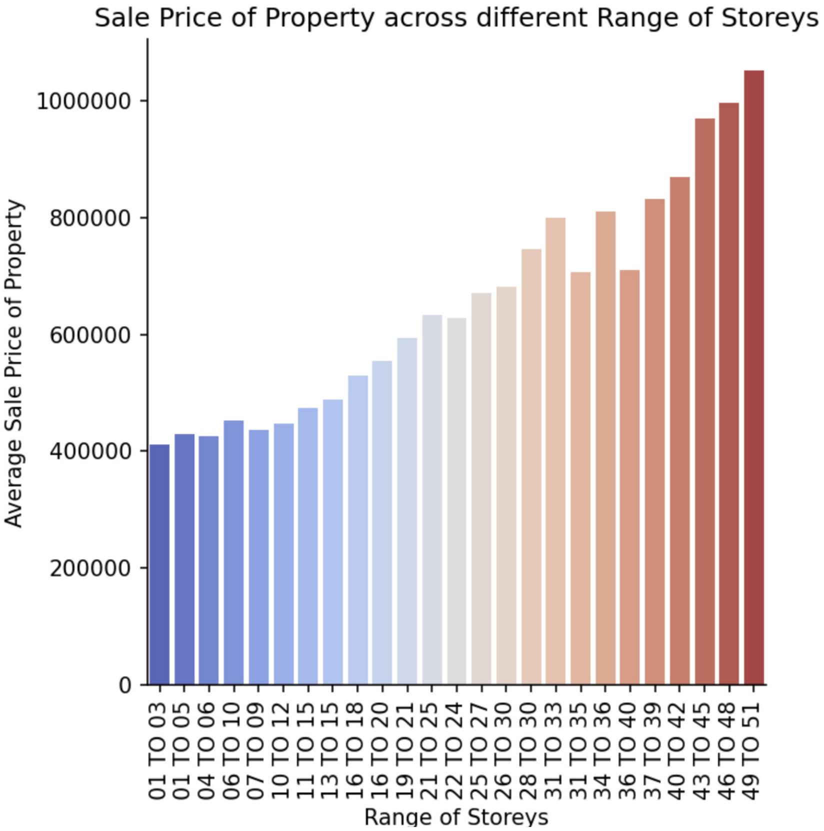
(HDB, my future home layout)



Categorical Plot: Resale Price vs Range of Storeys

1. Positive correlation
2. Properties located on a higher storey are sold for higher price
3. Approx. 60% price increase

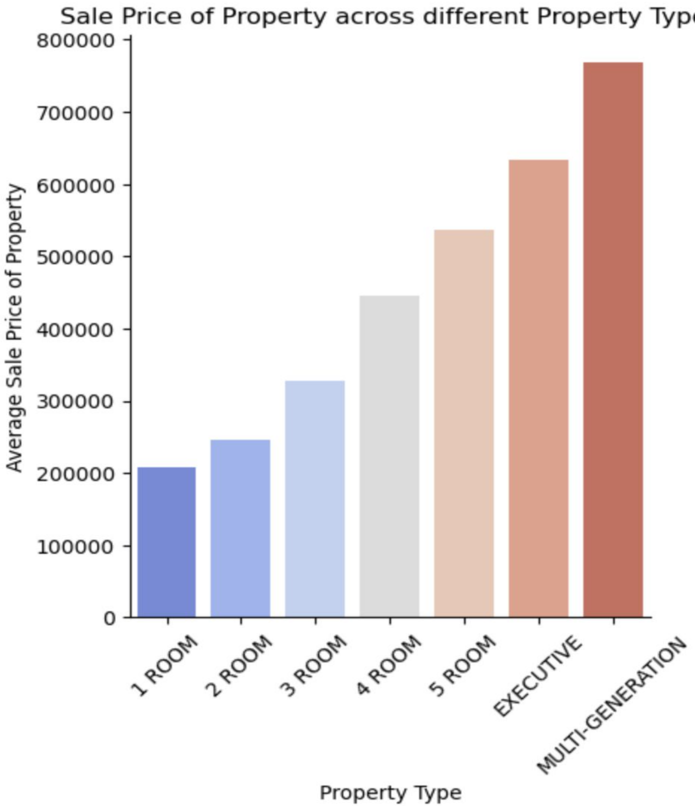
(Heatherwick Studio's Singapore skyscraper)



Categorical Plot: Resale Price vs Property type

1. Positive correlation
2. Approx. 75% increase in property price for larger properties.

(Serangoon, home of champions)



Baseline Model

The baseline model we used was multiple linear regression, and it produced the following results.

Model	R-Squared Score - Training Set	R-Squared Score - Test Set	Cross-Validated R-Squared Score folds = 5	RMSE
Multiple Linear Regression	0.90	0.90	0.90	44827

The initial results were promising with an R-Squared score of 0.90 on both training and test set suggesting that there is no overfitting.

The RMSE score of 44827 is also a good score when prices of properties can go up into the millions.

Baseline Model

1. From the scatterplot, we can see that our predictions are close to the actual resale price.

Regression Models with Regularisation

We next ran the dataset through regression models with regularisation as the dataset had many features. The results of these models are summarised in the table below.

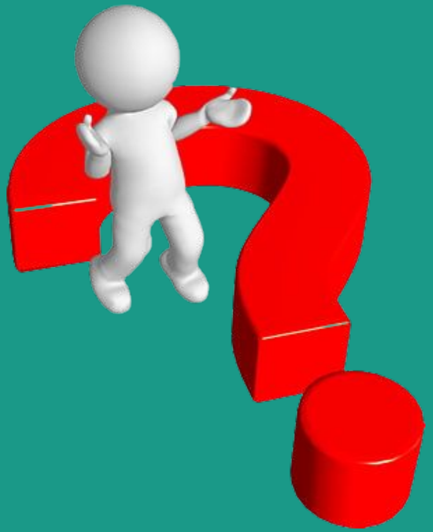
Model	R-Squared Score - Training Set	R-Squared Score - Test Set	Cross-Validated R-Squared Score folds = 5	RMSE
Multiple Linear Regression	0.90	0.90	0.90	44827
Lasso alpha = 93.6	0.89	0.89	0.89	47068
Ridge alpha = 1	0.90	0.90	0.90	44818
ElasticNet l1_ratio = 1 alpha = 93.6	0.89	0.89	0.89	47068

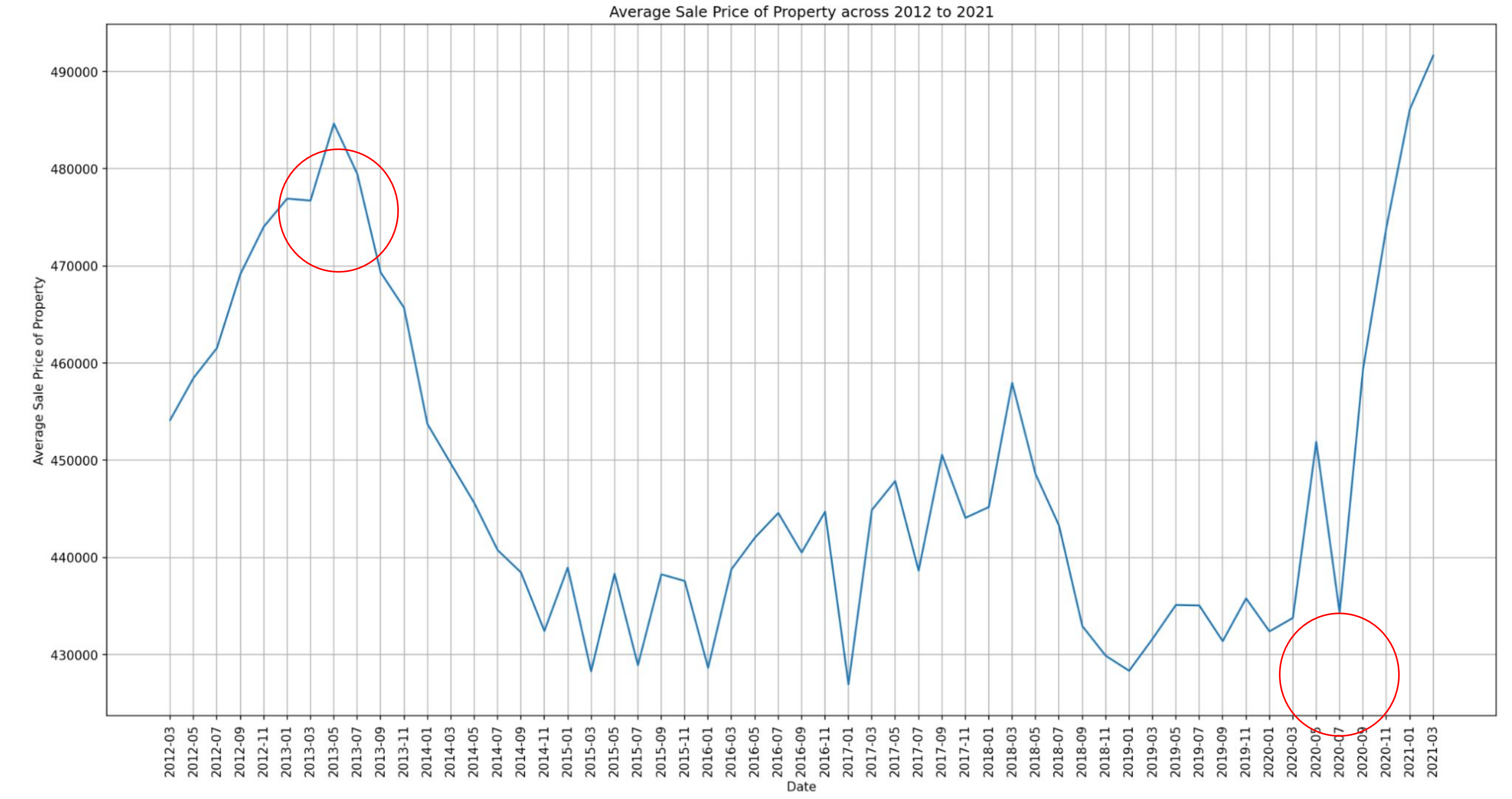
The results from the regularisation regression models are similar to the baseline model which is surprising as we felt that with a model with this many features, that performance would improve with regularisation. This result could mean that all the features are useful.

Improved Models

1. Removing outlier years

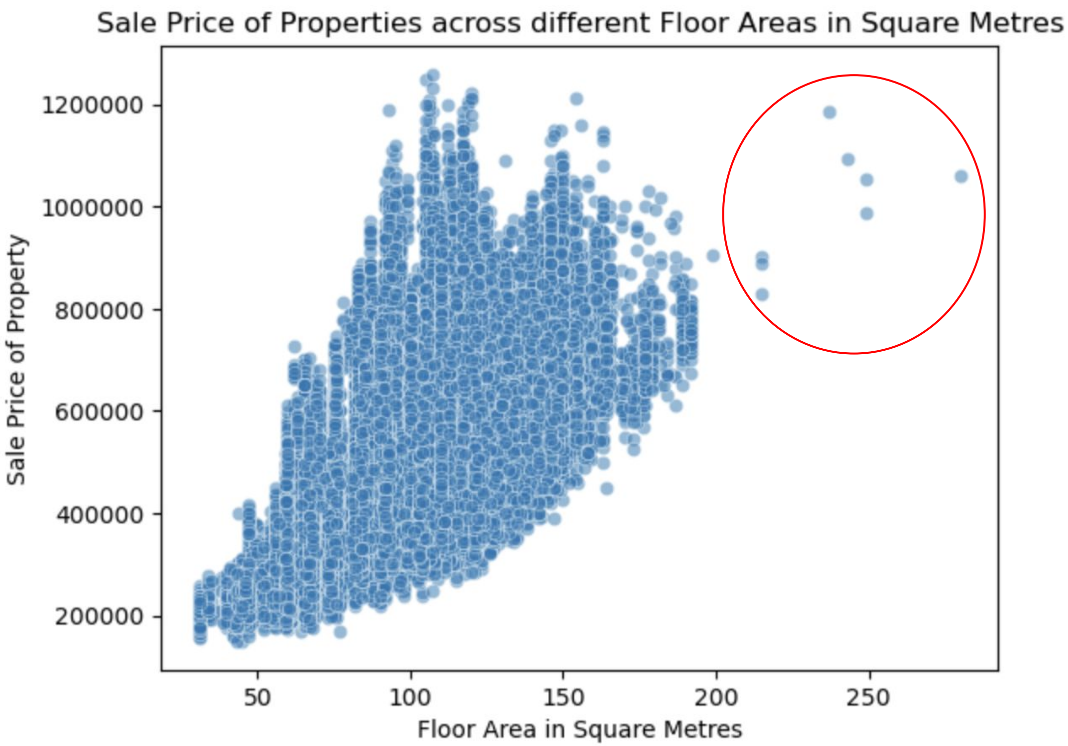
2. Removing outlier data points




Average Sale Price of Property across 2012 to 2021
Average Sale Price of Property
490000
480000
470000
460000
450000
440000
430000
Date
2012-03
2012-05
2012-07
2012-09
2012-11
2013-01
2013-03
2013-05
2013-07
2013-09
2013-11
2014-01
2014-03
2014-05
2014-07
2014-09
2014-11
2015-01
2015-03
2015-05
2015-07
2015-09
2015-11
2016-01
2016-03
2016-05
2016-07
2016-09
2016-11
2017-01
2017-03
2017-05
2017-07
2017-09
2017-11
2018-01
2018-03
2018-05
2018-07
2018-09
2018-11
2019-01
2019-03
2019-05
2019-07
2019-09
2019-11
2020-01
2020-03
2020-07
2020-11
2021-01
2021-03

Scatter Plot: Resale Price vs Floor Area

1. Positive correlation observed between floor area and resale price
2. As number floor area increases, price of property increase



Improved Models Results

Model	R-Squared Score - Training Set	R-Squared Score - Test Set	Cross-Validated R-Squared Score folds = 5	RMSE
Multiple Linear Regression	0.90	0.90	0.90	44828
Lasso alpha = 93.6	0.89	0.89	0.89	47068
Ridge alpha = 1	0.90	0.90	0.90	44816
ElasticNet l1_ratio = 1 alpha = 93.6	0.89	0.89	0.89	47068

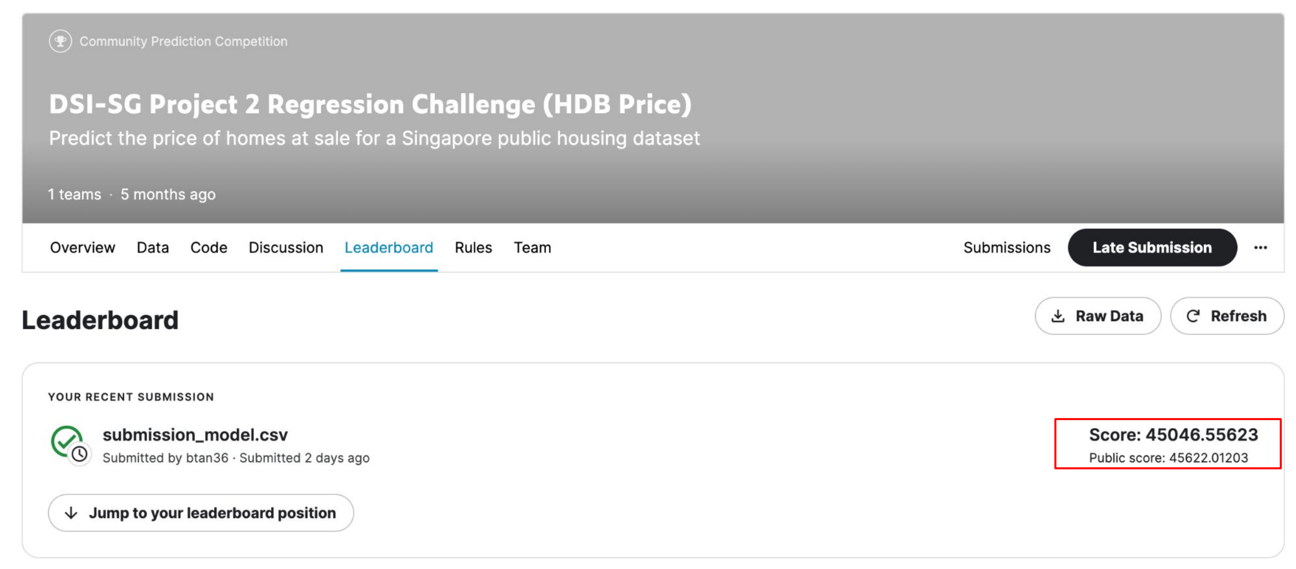
Model after removing outliers and 0 weighted features by Lasso	R-Squared Score - Training Set	R-Squared Score - Test Set	Cross-Validated R-Squared Score folds = 5	RMSE
Multiple Linear Regression	0.90	0.90	0.90	45379
Lasso alpha = 93.6	0.89	0.89	0.89	46429
Ridge alpha = 1	0.90	0.90	0.90	45383
ElasticNet l1_ratio = 1 alpha = 93.6	0.89	0.89	0.89	46429

Final Model Selection

- Surprisingly, after removing the outlier years and data points, the new models are performing almost exactly the same and maybe slightly even worse based on the RMSE.
- This could mean that the outliers were not actually having as big of an impact on the model that we thought;
- Or that the outliers actually contain valuable information in predicting the target variable.

Taking this into consideration, we will stick to the initial data with outlier years/data points for the final model used for Kaggle submission.

Kaggle Submission



Future Works & Conclusion

Conclusion:

1. We accomplished the goal of the problem statement of offering first time home buyers a way of getting a good gauge of property prices based on the attributes of the flat and property location, and not exacerbated costs due to external factors so that they can make informed choices on whether now is a good time to purchase the property or if they should wait till external factors are not affecting the price as much.

Recommendations for Future Works :

- To include financial data
    - To include macroeconomic parameters such as CPI
- To include demographics of populations of respective areas
    - Include average age of residents in a particular area
- Include decibel level as a factor.
    - Noise pollution could possibly affect property price
- Proximity to parks / green spaces.
    - Properties closer to parks may help fetch higher price
- To include Closeness to healthcare
    - With an aging population, this might be an increasingly importing factor