Visit our website: <https://share.streamlit.io/alwaysnacy/housing_prediction_4231>

### **Data Preprocessing**

The original dataset contains the prices and features of resale HDB flats from 2017 to 2021. Since the inclusion of more data may lower estimation variance and improve the predictive performance of the model, we also experimented with the complete resale HDB price datasets from 1990 to 2022. Both datasets consist of 10 resale HDB flat features, namely the month, town, flat type, block, street name, storey range, floor area, flat model and lease commence date. The newer datasets contain an additional column of ‘remaining lease’. These datasets have both numerical and categorical features. The numerical features are floor area, lease commence date and resale price while the rest are categorical features. Categorical features will be numerically encoded for our regression model. Different encoding methods such as one-hot encoding and target encoding for the categorical features are tested to predict the resale HDB price.

#### 1.1 Data Exploratory Analysis

After plotting scatter diagrams to examine the relationship between the features and the resale price, we have decided to drop ‘block’, ‘street\_name’. Prior to that, the full address of the HDB is derived by concatenating these two features, which are used to geographically encode our data with latitude and longitude. The coordinates of the HDB allow us to obtain more useful features that are likely to influence buyers’ purchasing decisions such as the distance to the nearest public transport and the nearest shopping mall which will be discussed in the next section.

1.1.1 Month

Resale price hardly differs in relation to the months. Interestingly, the frequency of housing is very scarce in April, May 2020 and July 2021. This led to a lower resale price distribution due to the reduced sample size.

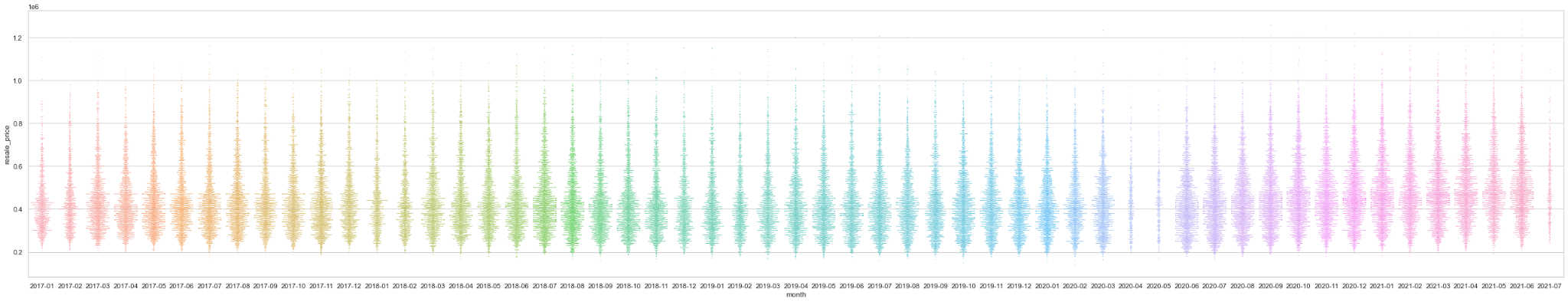


Figure 1. Month analysis

1.1.2 Town

In general, Punggol, Sengkang and Bishan have the top 3 highest resale HDB prices since 1990. Some towns with large resale HDB price ranges include Central area, Bishan and Bukit Timah.

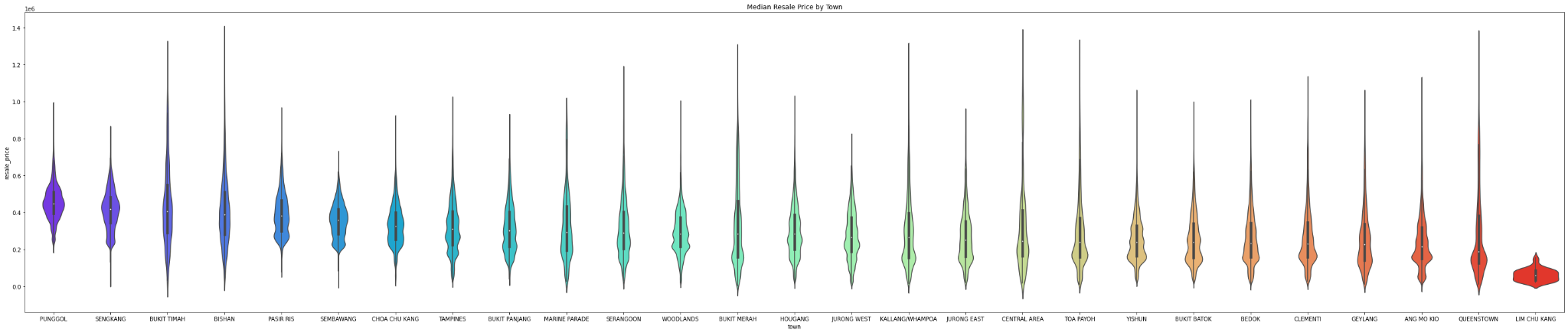


Figure 2. Town analysis

1.1.3 Flat type

As there is a small sample size of 1 room, 2 room and multi-generation flat types in the dataset, there is insufficient data to make a suitable conclusion from the relationship between flat type and resale price. Since the flat type has a natural order, it will be ordinally encoded.

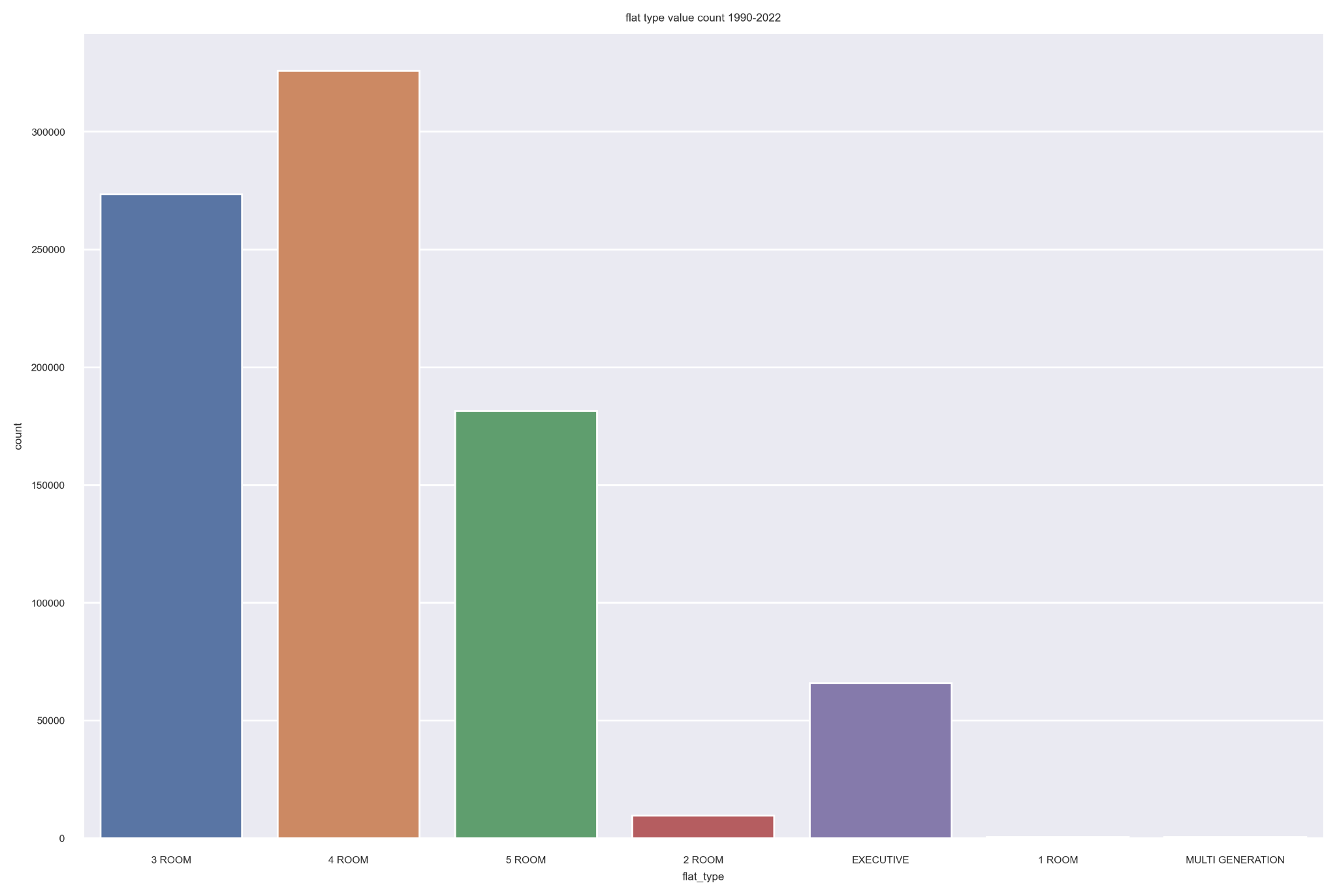


Figure 3. Flat type analysis

1.1.4 Storey range

We convert the storey range into a numerical variable by taking the mean of the lower and upper bound of each class. This also preserves the natural ordinal ordering of storey range of the HDB flat. By calculating the median value of the HDB range, the linear regression model yields a slightly better result as the R squared improves by ~0.001. However, the improvement is not guaranteed for another regression model such as the support vector regression.

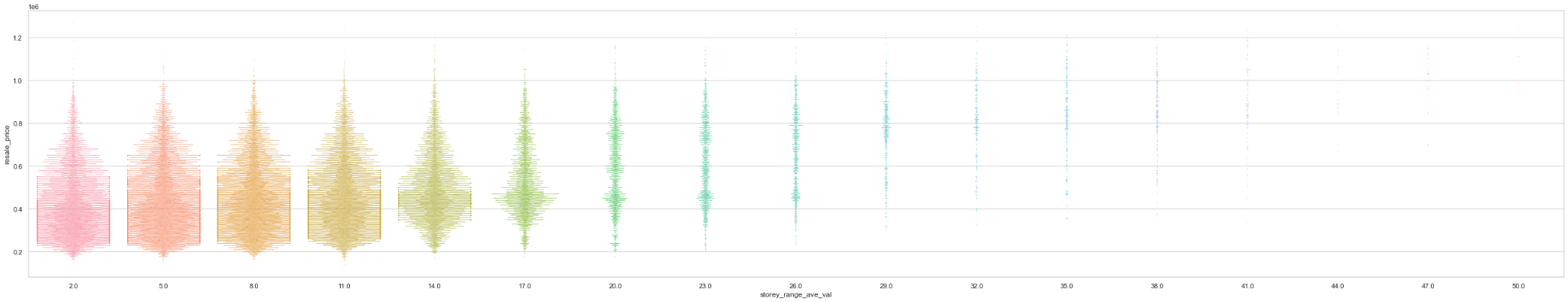


Figure 4. Storey range analysis

1.1.5 Floor area square meter

From the 2017-2020 data, a visible trend is that as floor area rises, the resale price is greater. However the same could not be said for the 1990-2022 data.

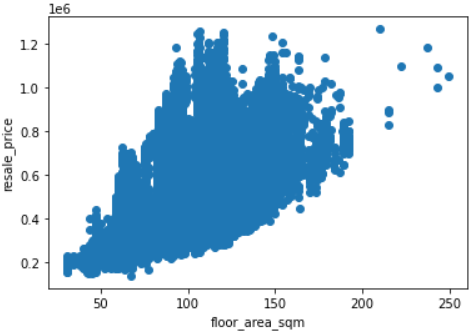
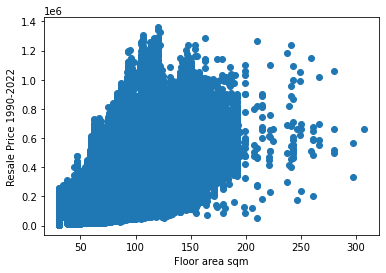


Figure 5. Area analysis

1.1.6 Flat model

We excluded the following flat model type: Terrace, Improved-Maisonette, Premium Maisonette, Multi Generation, Premium Apartment Loft, 2-room, due to their small sample size. As such, we discover a trend between flat model and resale price. For instance, New Generation, Simplified and Model A2 tend to fetch lower resale price in general.

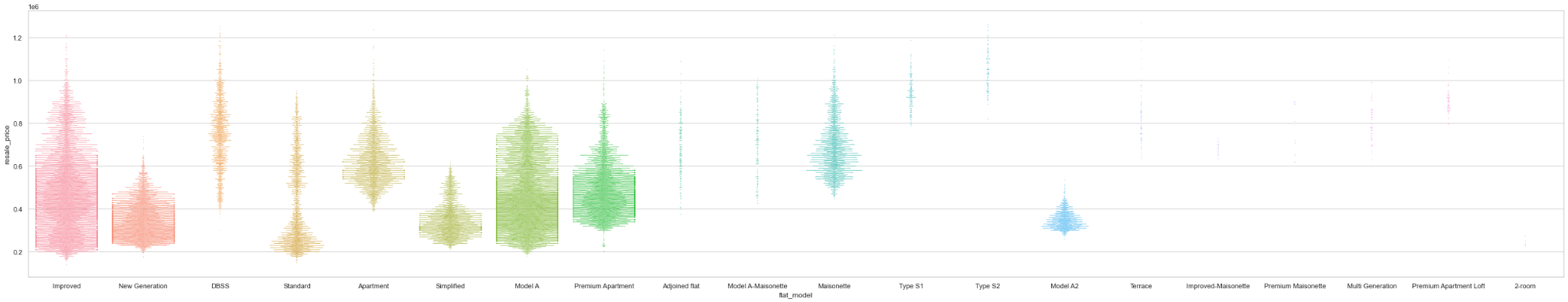


Figure 6. Flat model analysis

1.1.7 Lease Commence Date

Lease commence date data was initially presented in years. Generally HDB with later lease commence dates are valued at a greater resale price.

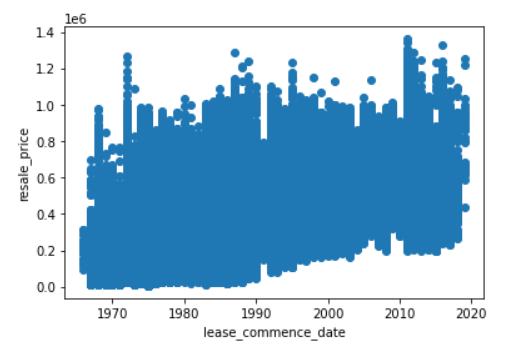


Figure 7. Lease analysis

1.1.8 Resale price

We see that the resale HDB price has risen significantly from 1990 to 2019. The property cooling measures introduced in 2013 and 2018 have led to a slight dip in the resale HDB prices.

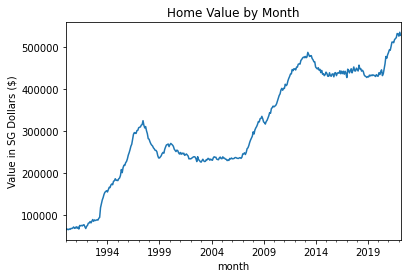
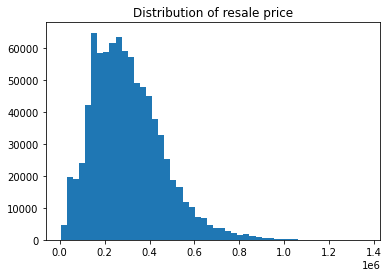


Figure 8. Resale price analysis

#### 1.2 Feature Synthesis

We also consider other factors that influence the HDB resale price, such as accessibility to public transport, shopping and lifestyle amenities and maturity of the town. Therefore, we have obtained relevant data to increase the accuracy of the model.

1.2.1 Proximity to MRT

The minimum distance to MRT of the resale HDB is obtained by geodesic measure with the coordinates of the HDB address and the coordinates of the list of MRT stations. The list of MRT stations is scraped from [MRT Map SG](https://mrtmapsingapore.com/mrt-stations-singapore/). The coordinates of the MRT and HDB address are acquired with the Onemaps API.

1.2.2 Proximity to Commercial Buildings

The proximity to the nearest shopping mall is obtained with a similar method as to how the proximity to MRT is calculated.

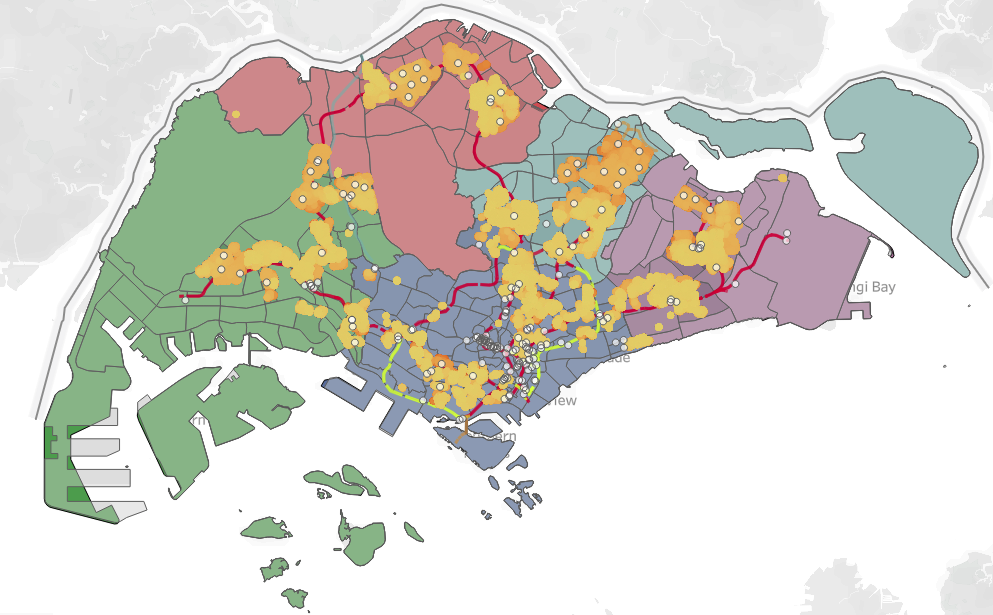


Figure 9. Spatial features analysis (details in appendix)

1.2.3 Availability of Markets or Hawker Centers

The availability of markets or hawker centers of the resale HDB is retrieved from the [HDB](https://data.gov.sg/dataset/hdb-property-information) [property information dataset](https://data.gov.sg/dataset/hdb-property-information). By mapping the addresses of the resale HDB price dataset, the presence of market or hawker centers is binarily encoded.

1.2.4 Town Maturity

The list of mature towns is obtained from various [house property websites](https://www.propertyguru.com.sg/property-guides/non-mature-vs-mature-bto-55760). Mature towns usually have a more extensive network of public transport and more developed facilities. Town maturity is included as a predictor as it can imply the convenience and accessibility of the HDB flats, which affects buyers’ purchasing decisions. However, there is no clear explanation in the classification of town maturity. For instance, towns like Punggol and Sengkang, which have well-established public transport and amenities can be listed as non-mature. The town maturity is also binarily encoded.

1.2.5 Region

Since there are a total of 27 towns, one hot encoding would lead to an increase of 27 dimensions and sparsity of the data. The performance of the model deteriorates with higher dimensions due to the curse of dimensionality. Hence, towns can be grouped into their respective regions (central, east, north, northeast or west regions) to reduce the number of dimensions of the model.

#### 1.3 Feature Engineering

##### 1.3.1 Multicollinearity

Multicollinearity occurs when two or more independent variables are highly correlated with one another in a regression model. It can pose an issue in interpreting the impact of one feature on its HDB resale price. However, multicollinearity may not affect the model’s predictive accuracy.

To check the collinearity between categorical variables, they must first be converted into numerical variables. This can be done with target encoding, where the categorical value is substituted with the mean of the resale price. Next, multicollinearity between predictors can be examined with variance inflation factor (VIF) or correlation matrix. In this case, we see a high correlation between floor area and flat type. Thus, one of the variables with high collinearity may be dropped to improve the model’s interpretability.

##### 1.3.2 Selected Features

There are a total of 12 selected features used for our model. The features are as follows: ‘month’, ‘flat\_type’, ‘floor\_area\_sqm’, ‘flat\_model’, ‘lease\_commence\_date’,’ availability of market/hawker’, ‘town maturity’, ‘region’, ‘distance to the nearest MRT station’, ‘distance to the nearest shopping mall’ and ‘storey\_range’.

##### 1.3.3 Feature Scaling

From the distribution plot of the resale HDB price, we observe that the resale HDB price is rightly skewed. Therefore, we perform log transformation on the resale HDB price before passing it into the model as the linear regression model assumes that the residuals are normally distributed. Though feature scaling on the independent variables has minimal impact on predictive accuracy, we decided to standardize the coefficients to better interpret the effect each predictor has on resale HDB price.

### **Model**

We have implemented different regression models to predict the resale HDB price. The regression models are run with the same features as mentioned above. The deep learning approach is also used to compare against the performance of regression models.

2.1 Linear Regression

Linear regression serves as the baseline of our models. The performance metric of the model is ordinary least squares. Considering that the trend of resale HDB price might be cyclical, we split the testing and training data into different periods and compare the performance of the model. Testing with 6, 12, 36, and 60 periods respectively, the best R square for prediction is 0.9372 where the 60th month is used for testing and the previous 59 months are reserved for training. Therefore, we set the test period to be 60 in our subsequent experimentation. As the R square of the model is lower than the prediction R square, we did not regularize the model.

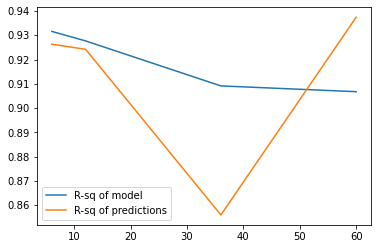


Figure 9. R square comparison

#### 2.2 Polynomial Regression

Polynomial regression is used to examine if the variables and resale HDB price have a nonlinear relationship. The R square of the model has improved while the R square for prediction has decreased drastically. This may be caused by overfitting. Hence, we test ridge regression with polynomial features to address overfitting.

#### 2.3 Ridge Regression

Ridge regression introduces a regularization parameter to penalize overfitting in a complicated model. The parameter alpha in ridge regression which determines the amount of regularization is tuned to find the optimal one for our model. Ridge regression with 5 fold validation gives an average R square of . An exhaustive search of the best parameter with GridSearchCV from the range of 0.5 to 50 shows that the best parameter is .

R-sq of model: 0.9288333581576479

RMSE: 142578826.84

R-sq of predictions: 0.9260819260913342

#### 2.4 Support Vector Regression

For support vector regression, we set epsilon to be 0 and test the regularization parameter with 1, 10 and 50.

C = 10

R-sq of model: 0.911352613085745

RMSE: 0.03

R-sq of predictions: 0.9109760158614247

#### 2.5 Multi-Layer Perceptron Model (Deep Learning)

MLP Model is used to compare with the linear models. It is trained using the same data with similar features, so it will have 34 nodes in the input layer. Before training, we scaled the data using the StandardScaler to make the model converge faster and more stable. The first version of the model comprises 4 layers with the largest layer having 200 nodes. We trained it for 500 epochs without any dropout layer.

R-sq of model: 0.979699

RMSE: 23467.35616 (without log transform)

R-sq of predictions: 0.9783296

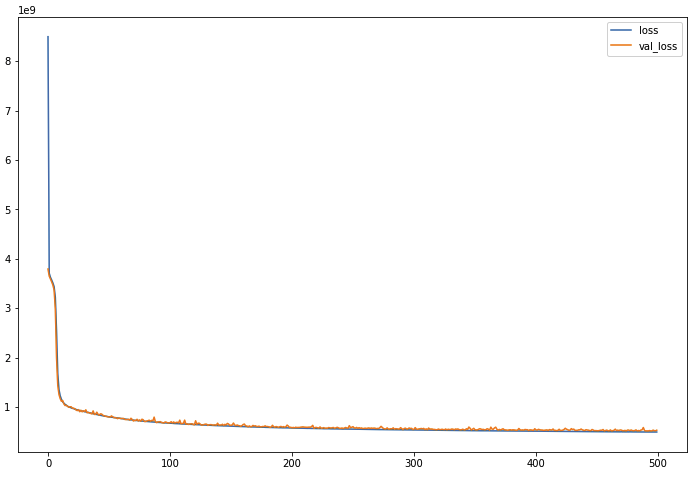


Figure 10. Deep learning loss graph

The loss decreases rapidly after 10-15 epochs, then the model converges slowly. It can be observed that the loss is still going down when we stop the training, which means it can be reduced further if we increase the number of epochs. We also tried the model with more layers and more nodes, details are in the appendix.

### **Conclusion**

| Regression Model | Linear | Polynomial | Ridge | Support Vector | MLP Model |
| --- | --- | --- | --- | --- | --- |
| R-sq of model |  | 0.94056 | 0.92883 | 0.879464 | 0.979699 |
| RSME |  | 0.22 (Log transform) | 142578826.84 | 215613853.84 | 23467.356 |
| R-sq of predictions | 0.9372 | 0.83 | 0.92608 | 0.888218 | 0.9783296 |

Table 1. Performance Comparison

#### 3.1 Evaluation of Model

Among the linear models that we have tried, taking into account the different R2 values, the best model appears to be the polynomial regression. However, high R2 value does not imply best fit.

Deep learning models are proven to be more efficient in learning the data, hence, giving more accurate predictions.

However, given that we have a large dataset with just 34 features (after encoding), the performance of the linear model (linear regression) is mostly as good as the deep learning model.

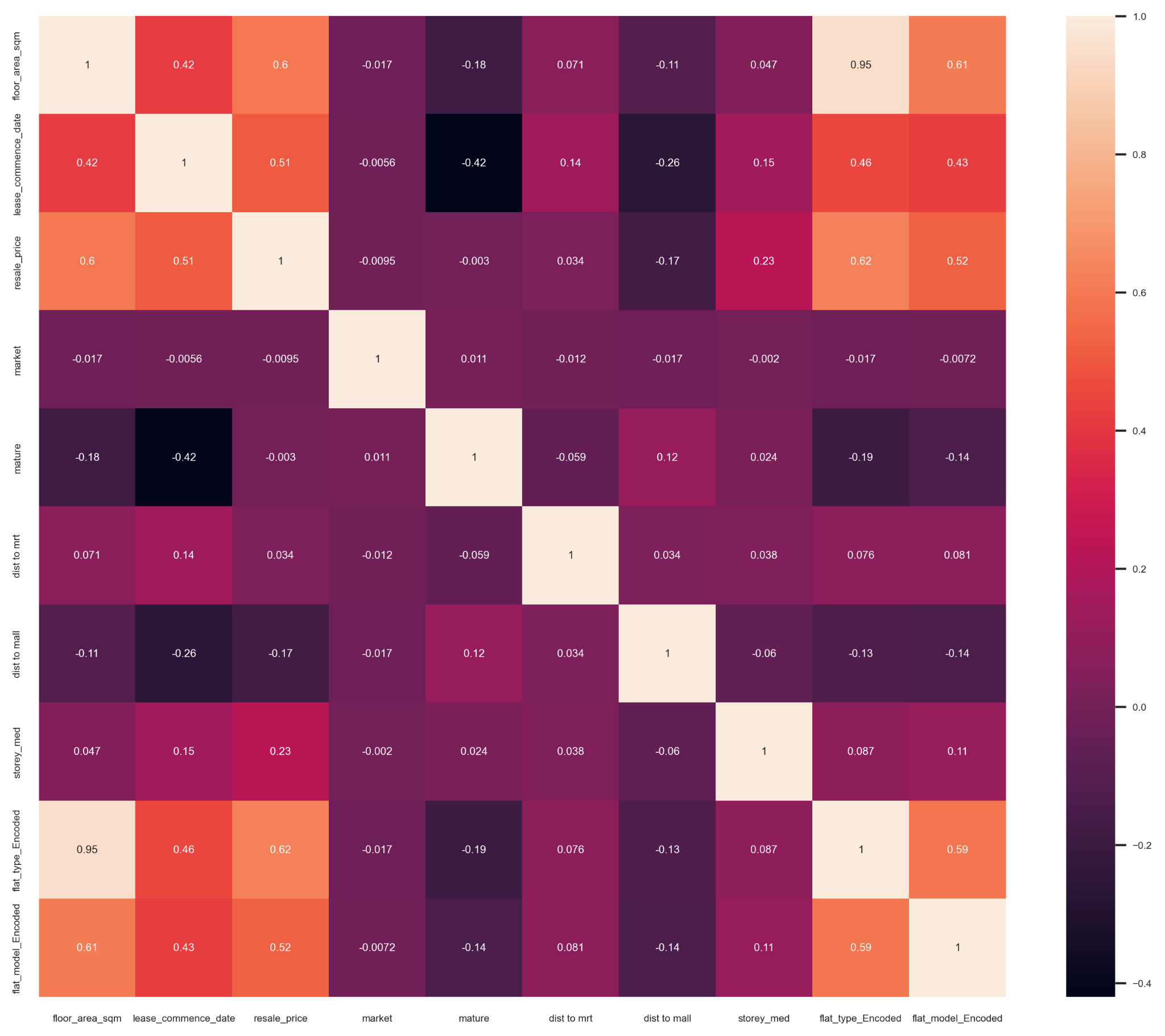
**Link:**

Code:

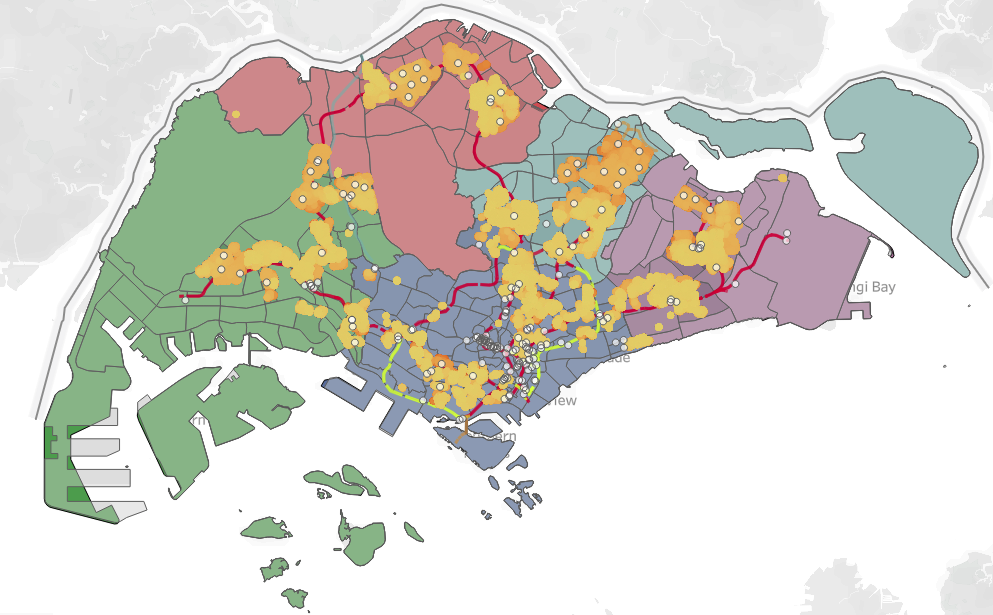
<https://github.com/EE4802-IE4213-CA1/housing-prediction>

**Appendix:**

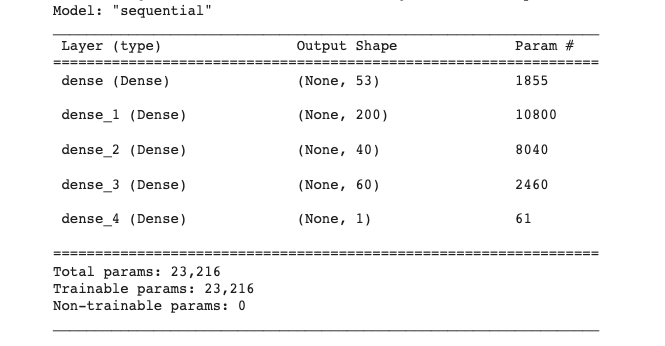
**Collinearity Matrix**



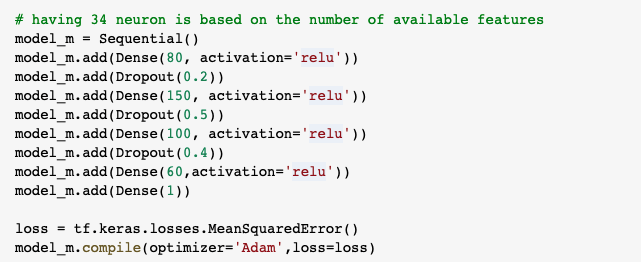
**Spatial Data Analysis using Tableau:**



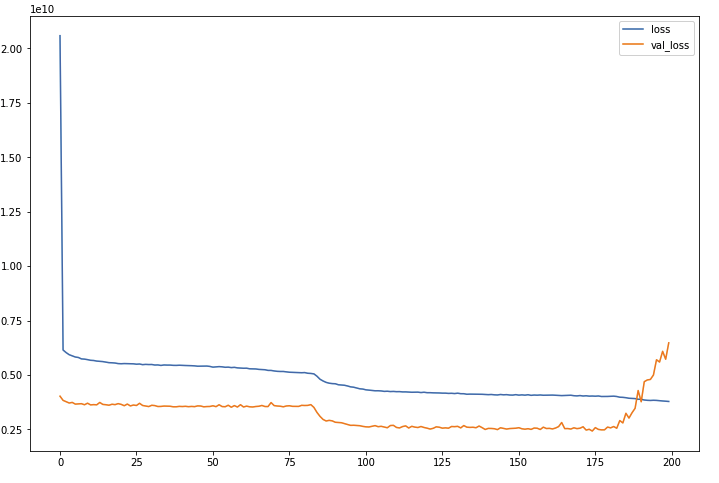
The yellow dots are all the HDBs and the shade of them reflects their prices, the darker the dot, the higher it is valued. The white dots are malls and lines seen under the dots are the rail lines of Singapore. It can be noticed that most of the HDB are concentrated near the MRT lines. We can all notice the 5 regions from east to west. The shade of the color of the region reflects the total population of the sub areas under that region. Tampines is the most crowded town where dark purple can be observed.

**MLP Model Summary (Report model)**

**Deeper MLP Model (more nodes and more layers)**



We added Dropout layers to prevent overfitting. We trained with batch size 512 and 200 epochs.



After epoch 175, validation loss started to increase.