Analyzing resale prices of HDBs using Machine Learning Algorithms

ABSTRACT

The housing prices in Singapore has been on a rising trend in the recent years. However, these prices can vary significantly due to factors such as location, property type, and more. In this research paper, we use various machine learning algorithms namely k-Nearest Neighbour (KNN) regression, Decision tree regressor, Gradient boosting regressor, Linear regression and Support Vector Regression (SVR) to predict the median resale housing prices in Singapore.

I. INTRODUCTION

It has been reported that the resale prices of HDB flats has continued to increase for the 22nd consecutive month in April 2022. There were increases across all flat types and locations, with prices rising 1.1 percent in April 2022 month on month, and 11.9 percent on the year [1]. The dataset for median housing resale prices is readily available on data.gov.sg. In this paper, we will discuss about using various machine learning algorithms to determine if location, flat model, floor area and remaining lease may have a correlation to the median resale housing prices.

II. DATASET

The dataset "Resale flat prices based on registration date from Jan-2017 onwards" was retrieved from data.gov.sg [2]. It consists of 123,934 records of housing prices from 2017 to 2022, June. The below table shows a preview of the dataset provided. It includes attributes such as Month, Town, Flat type, Block, Street name, Storey range, Floor area sqm (sqm), Flat model, Lease commence date, Remaining lease, Resale price ($).

Due to the complexity and volume of the dataset, we will examine in particular, only the resale prices of the flats transacted between 2021, January, to 2022, June. There are in total, 40,345 records for this time period.

	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
87589	2021-01	ANG MO KIO	2 ROOM	170	ANG MO KIO AVE 4	07 TO 09	45.0	Improved	1986	64 years 01 month	225000.0
87590	2021-01	ANG MO KIO	2 ROOM	170	ANG MO KIO AVE 4	01 TO 03	45.0	Improved	1986	64 years 01 month	211000.0
87591	2021-01	ANG MO KIO	3 ROOM	216	ANG MO KIO AVE 1	04 TO 06	73.0	New Generation	1976	54 years 04 months	275888.0
87592	2021-01	ANG MO KIO	3 ROOM	223	ANG MO KIO AVE 1	07 TO 09	67.0	New Generation	1978	56 years 01 month	316800.0
87593	2021-01	ANG MO KIO	3 ROOM	223	ANG MO KIO AVE 1	10 TO 12	67.0	New Generation	1978	56 years	305000.0

Figure 1: Preview of the dataset provided by data.gov.sg, from month 2021, January onwards

III. ATTRIBUTES IN DATASET

Attributes	Description
month	Month in which the property was transacted.
town	Location of the property transacted.
flat_type	Can be in either 2-room, 5-room, Executive type of flats and etc
block	Block number of the resale flat
street_name	Street Address of which the resale flat is located in
storey_range	The range of storeys between which the resale flat is located (e.g. floors 1 to 3)
floor_area_sqm	Floor area of the resale flat, in square metres
flat_model	Various types of flat models are available, including "improved", "new generation", "Model A" and more
lease_commence_date	Date in which the flat was first built and leased out
remaining_lease	Remaining tenor before the flat is unavailable for lease.
resale_price	Price at which the flat was transacted, in Singapore Dollars

The below graphs have been plotted to observe the distribution of resale prices, from 2021, January to 2022, June.

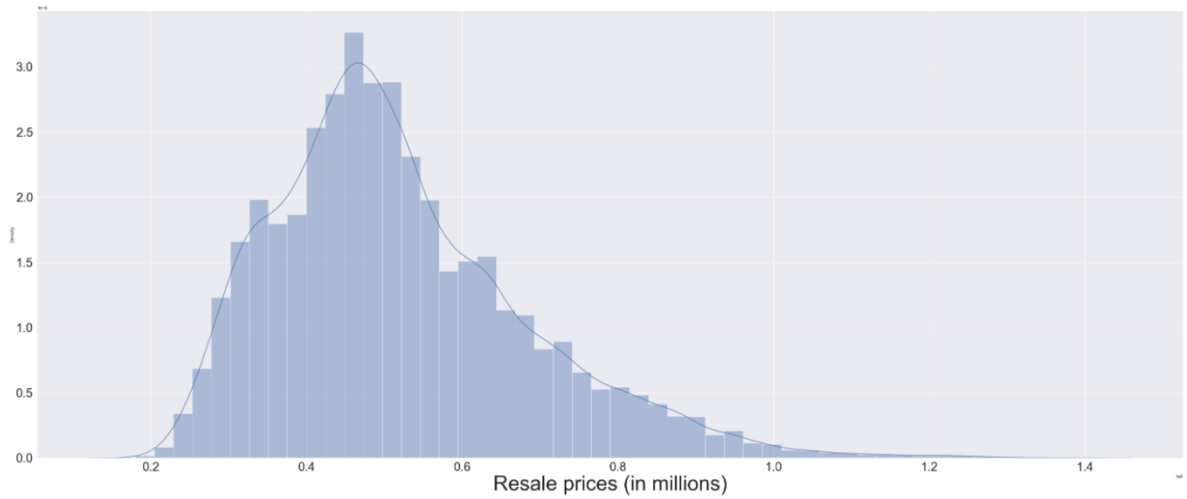

Figure 2.1 Distribution of resale prices of the flats (In millions, Singapore Dollars)

The distribution of resale prices is perceived to be somewhat negatively skewed, with most of the housing prices ranging between 200,000 Singapore dollars, to 1.0 million Singapore dollars.

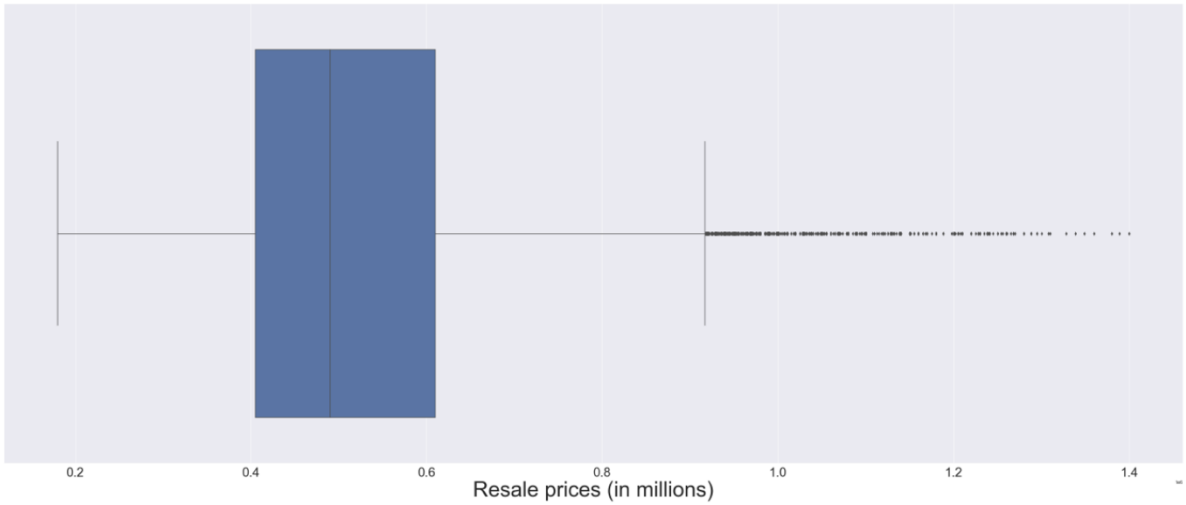

Figure 2.2 Boxplot of resale prices of the flats (In millions, Singapore Dollars)

Similarly, a boxplot was also plotted to examine the distribution of resale prices in Figure 2.2. As seen from the boxplot, there are numerous outliers in the resale prices. We will further examine other attributes below to determine the causes of these outliers, and prepare the data required for training the model.

Before building a model, the attributes were examined to see if there is a correlation to the resale prices. We started with the X-Y generic plots to get a preview of each features. The plots have been shown below.

A. 'TOWN' VS 'RESALE_PRICE'

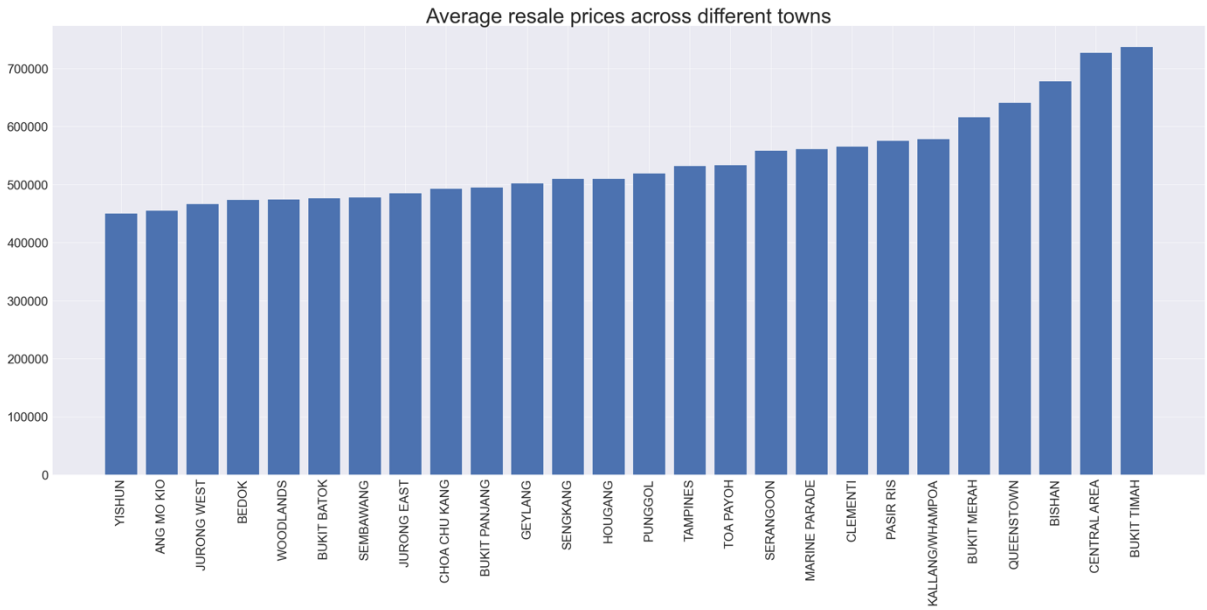

Figure 2.3 Barcharts of average resale prices, based on towns

In general, most of the average resale prices for various towns seem to range between 400,000 to 600,000 Singapore Dollars. However, houses in some of the areas such as Bishan, Central and Bukit Timah were seemingly more expensive, with average resale prices going up to 700,000 Singapore dollars.

B. ‘FLOOR_AREA_SQM’ VS ‘RESALE_PRICE’

Figure 2.4 Scatterplots of average resale prices, based on floor area (in square metres)

The floor area seems to have a positive correlation to the average resale prices. The higher the floor area, the higher the average resale prices.

C. ‘FLAT_TYPE’ VS ‘RESALE_PRICE’

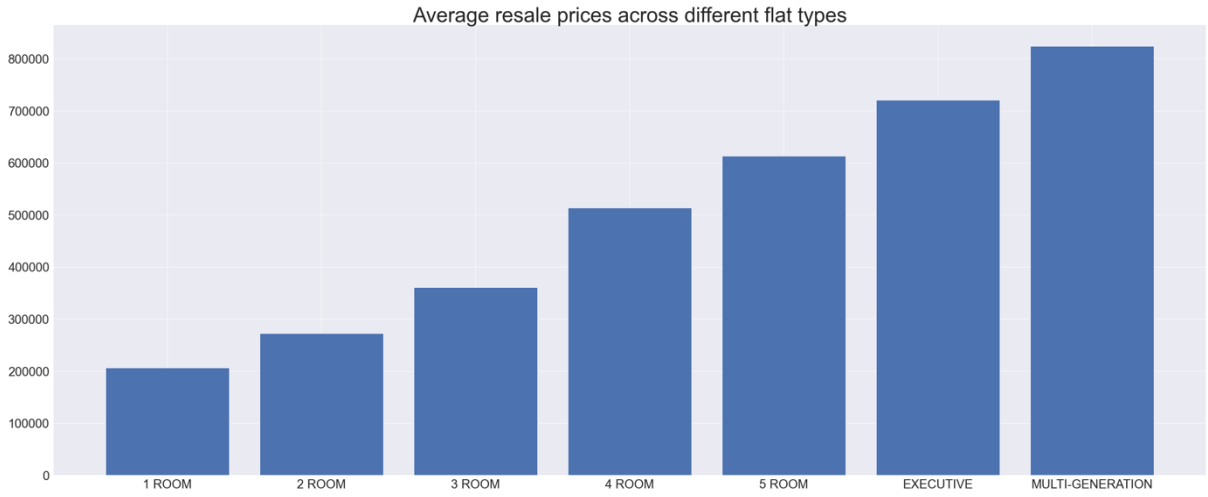

Figure 2.5 Barcharts of average resale prices, based on flat types

Similarly, for flat types, it seems that 1 room flat types have the lowest average resale prices, whilst better built flat types like multi-generation flat type has the highest average resale prices.

D. ‘FLAT_MODEL’ VS ‘RESALE_PRICE’

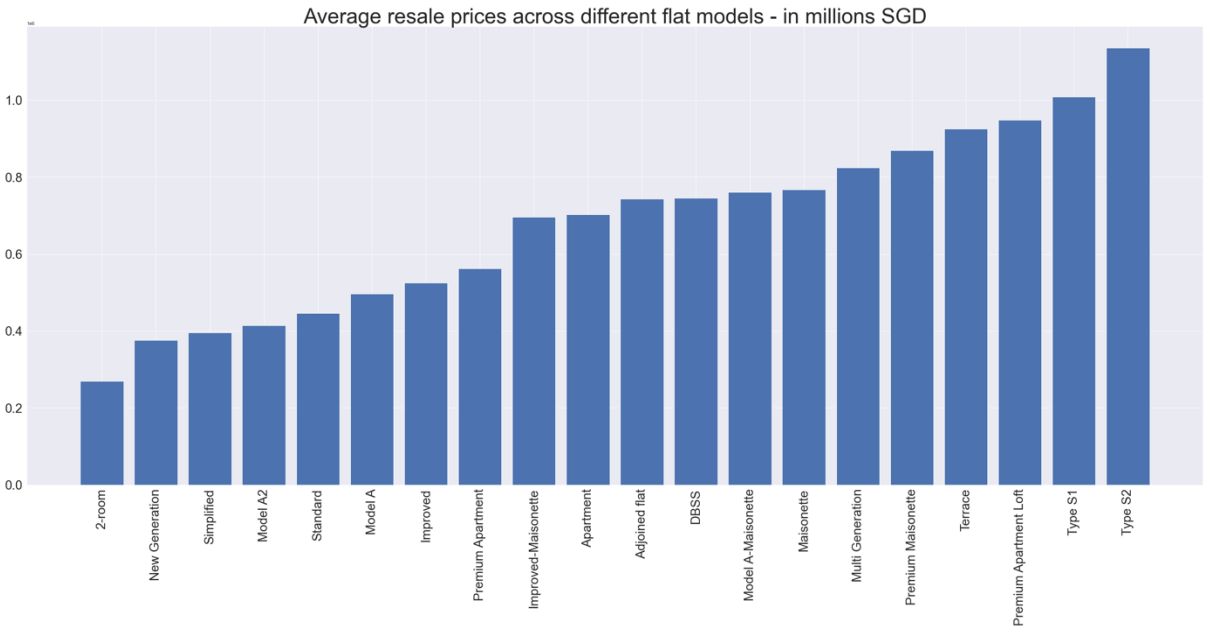

Figure 2.6 Barcharts of average resale prices, based on flat models

The flat models have also been sorted according to its average resale prices, with 2-4oom being the cheapest, and Type S2 being the most expensive, on average.

E. ‘STOREY_RANGE’ VS ‘RESALE_PRICE’

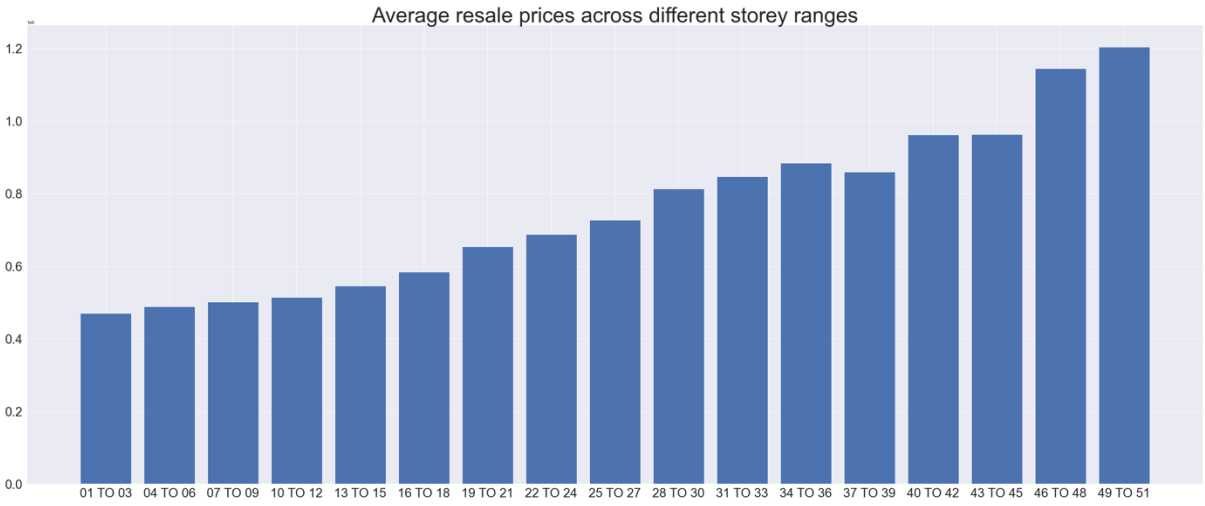

Figure 2.7 Barcharts of average resale prices, based on storey range

Similar to floor area, the higher the storey range of the flat, the more expensive it becomes. Flats situated between 1st to 3rd floors are expected to be the least expensive, whilst flats on higher storeys from 49 to 51, are the most expensive on average.

IV. METHODOLOGY

A. APPLYING TRANSFORMATION TO DATASET

After understanding various features of the data and its correlation to the housing prices, we will now proceed to clean the training dataset. Since the dataset does not come with null values, we will process the outliers in the resale prices first. This is done by adding a column in the dataset and applying numpy.log to the resale prices. This makes the distribution of resale prices seemingly more evenly distributed, after transformation.

Figure 3.1 Distribution of resale prices of the flats after transformation of the data

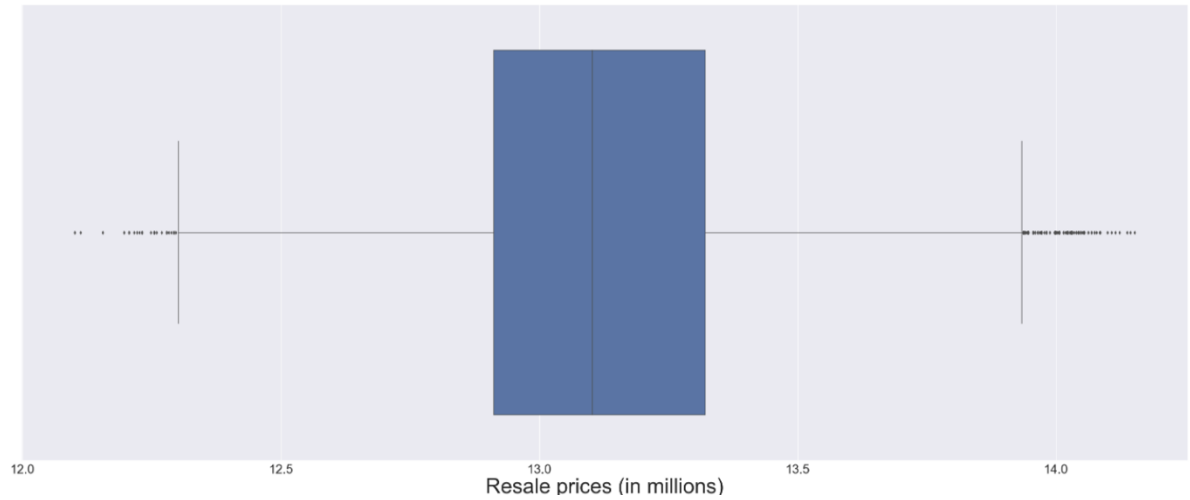

Figure 3.2 Boxplot of resale prices of the flats after transformation of data

B. DROPPING COLUMNS SUCH AS ‘MONTH’, ‘BLOCK’, 'STREET_NAME', 'LEASE_COMMENCE_DATE'

Next, we will proceed to drop these 4 columns, namely ‘month’, ‘block’, 'street_name', 'lease_commence_date'. The ‘month’ column has been dropped since most of the transactions were performed within a short time frame (i.e. 2021 to 2022, June), and ‘block’ and ‘street_name’ columns were dropped as it has little relevance to the resale prices. Further, the

'lease_commence_date' was also dropped since similar information can be obtained from 'remaining_lease'.

C. CONVERTING OBJECT VALUES FOR 'TOWN', 'FLAT_MODEL', 'FLAT_TYPE', 'STOREY_RANGE' TO INTEGER VALUES

Lastly, we will proceed to assign values for the columns: 'town', 'flat_model', 'flat_type', 'storey_range'. Values are assigned based on their average resale prices. For instance, the values for 'flat_type' are assigned values from 0 to 6, since there are 7 different 'flat_types' in the dataset. The values are assigned according to its mean sale prices, with 1 room flats given a value of 0, and Multi-Generation flat types given a value of 6.

```
#Transforming objects in the data to values for town, flat_type, storey_range, flat_model

train_test_data.replace({'town':{'YISHUN':0, 'ANG MO KIO':1, 'JURONG WEST':2, 'BEDOK':3, 'WOODLANDS':4,
       'BUKIT BATOK':5, 'SEMBAWANG':6, 'JURONG EAST':6, 'CHOA CHU KANG':7,
       'BUKIT PANJANG':8, 'GEYLANG':9, 'SENGKANG':10, 'HOUGANG':11, 'PUNGGOL':12,
       'TAMPINES':13, 'TOA PAYOH':14, 'SERANGOON':15, 'MARINE PARADE':16, 'CLEMENTI':17,
       'PASIR RIS':18, 'KALLANG/WHAMPOA':19, 'BUKIT MERAH':20, 'QUEENSTOWN':21, 'BISHAN':22,
       'CENTRAL AREA':23, 'BUKIT TIMAH':24},
                                  'flat_model':{'2-room':0, 'New Generation':1, 'Simplified':2, 'Model A2':3, 'S
       'Model A':5, 'Improved':6, 'Premium Apartment':7, 'Improved-Maisonette':8,
       'Apartment':9, 'Adjoined flat':10, 'DBSS':11, 'Model A-Maisonette':12,
       'Maisonette':13, 'Multi Generation':14, 'Premium Maisonette':15, 'Terrace':16,
       'Premium Apartment Loft':17, 'Type S1':18, 'Type S2':19},
                                  'flat_type':{'1 ROOM':0, '2 ROOM':1, '3 ROOM':2, '4 ROOM':3, '5 ROOM':4, 'EXECU
       'MULTI-GENERATION':6},
                                  'storey_range':{'01 TO 03':0, '04 TO 06':1, '07 TO 09':2, '10 TO 12':3, '13 TO
       '19 TO 21':6, '22 TO 24':7, '25 TO 27':8, '28 TO 30':9, '31 TO 33':10, '34 TO 36':11,
       '37 TO 39':12, '40 TO 42':13, '43 TO 45':14, '46 TO 48':15, '49 TO 51':16}},
                                  inplace=True)
```

Figure 3.3 Values assigned to data that are identified as objects in the dataframe

As a result, this is the final processed dataset, before it is split into train and test datasets, using 80% dataset as training dataset, and 20% dataset as test dataset.

	town	flat_type	storey_range	floor_area_sqm	flat_model	remaining_lease	resale_price	resale_price_log
87589	1	1	2	45.0	6	64	225000.0	12.323856
87590	1	1	0	45.0	6	64	211000.0	12.259613
87591	1	2	1	73.0	1	54	275888.0	12.527750
87592	1	2	2	67.0	1	56	316800.0	12.666026
87593	1	2	3	67.0	1	56	305000.0	12.628067

Figure 3.4 Clean dataset ready to be split into train and test sets.

We also examine the correlation between the different attributes below, as illustrated in the heatmap below.

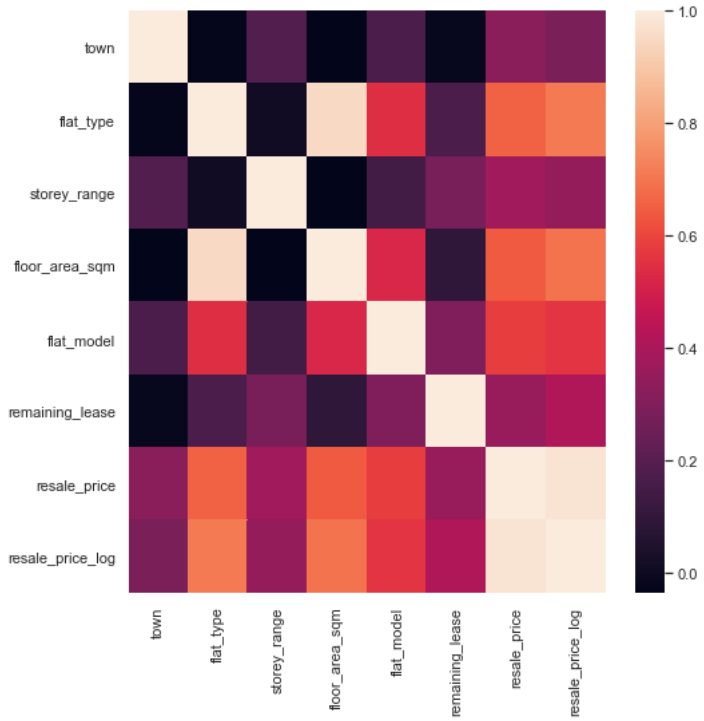

Figure 3.5 Heatmap showing the correlation between various features

V. EXPERIMENT

Next, we will perform five different machine learning algorithms on the datasets – namely the Decision tree regressor, KNeighborsRegressor, Ensemble Gradient Boosting Regressor, Support Vector Regression (SVR) and Linear Regression. Below are the results derived from applying the five different types of algorithms, sorted in descending order based on its training set scores.

VI. RESULTS

Models	r^2 Score - Training	r^2 Score - Test
Decision Tree	0.977888	0.921495
KNeighborsRegressor	0.956013	0.918892
Gradient Boosting Regressor	0.944872	0.937175
SVR	0.833920	0.828555
Linear Regression	0.762400	0.755621

Figure 4.1 Results (r^2 scores) derived from the training and test datasets for each of the algorithms used

VII. Conclusion

Decision Tree Regressor, KNeighborsRegressor and Ensemble Gradient Boosting Regressor produced fairly similar and commendable results for both Training and test datasets. Taking into consideration of both the training and test r^2 scores performance, Decision Tree Regressor seems to be the best algorithm to be applied in this problem; it achieved a decent r^2 score of 97.78% and 92.15% respectively for the train and test datasets.

VIII. REFERENCES

1. Vivian, T., "HDB resale prices continue upward trend in April as sales volumes flatten: SRX, 99.co". Retrieved, June 4, 2022 from https://www.businesstimes.com.sg/real-estate/hdb-resale-prices-continue-upward-trend-in-april-as-sales-volumes-flatten-srx-99co

2. Data.gov.sg: resale-flat-prices-based-on-registration-date-from-jan-2017-onwards.csv [Online]. Available: https://data.gov.sg/dataset/resale-flat-prices