Influence of proximity to MRT stations and prestigious primary schools on HDB resale prices

Introduction and Problem Statement

HDB flats account for over 80% of homes in Singapore[1]. For aspiring HDB buyers and potentially future HDB sellers like us, there are many considerations that influence the value of a HDB flat. The objective of this study is to understand the impact of an HDB flat's accessibility (location with respect to an MRT station) and proximity to 'prestigious' primary schools on its resale price, and provide potential explanations for the observed results.

Dataset Description

The base HDB resale dataset[2] provided includes information on the address, storey, size, flat type, lease information, and resale price of every transacted HDB resale unit from 2000 to 2022.

The accessibility of a HDB resale unit, here defined as its proximity to an MRT station, requires location data of Singapore MRT stations. These data, in the form of latitude and longitudinal coordinates, were referenced from Lee's Singapore MRT locations dataset[3].

Latest information on primary school rankings in Singapore were referenced from Schoolbell[4], with 12 primary schools selected as 'prestigious'. A 'prestigious' school is defined as a school that has either a popularity of >= 100% (more applicants than places), or that offers the Gifted Education Programme (GEP). The latitude and longitudinal coordinates of these prestigious primary schools were recorded manually from Google Maps[5].

Each HDB unit in the base dataset was mapped to its corresponding postal code, latitude and longitude using Lee's HDB Postal Code Mapper dataset[6].

Methodology / Approach of Solution

All codes were run on Python 3.

Data Preprocessing was first carried out to ensure the quality of data is accurate, complete, and consistent. These include imputing missing dataset values such as those associated with remaining HDB lease, encoding categorical data such as flat type into numerical data, normalizing lease commence date to account for inflation over time, converting postal addresses into coordinates, amongst others.

To avoid multicollinearity issues where correlated features can impact the performance of machine learning models, a dummy feature (degree of freedom) was removed whenever categorical data was encoded using one-hot encoder.

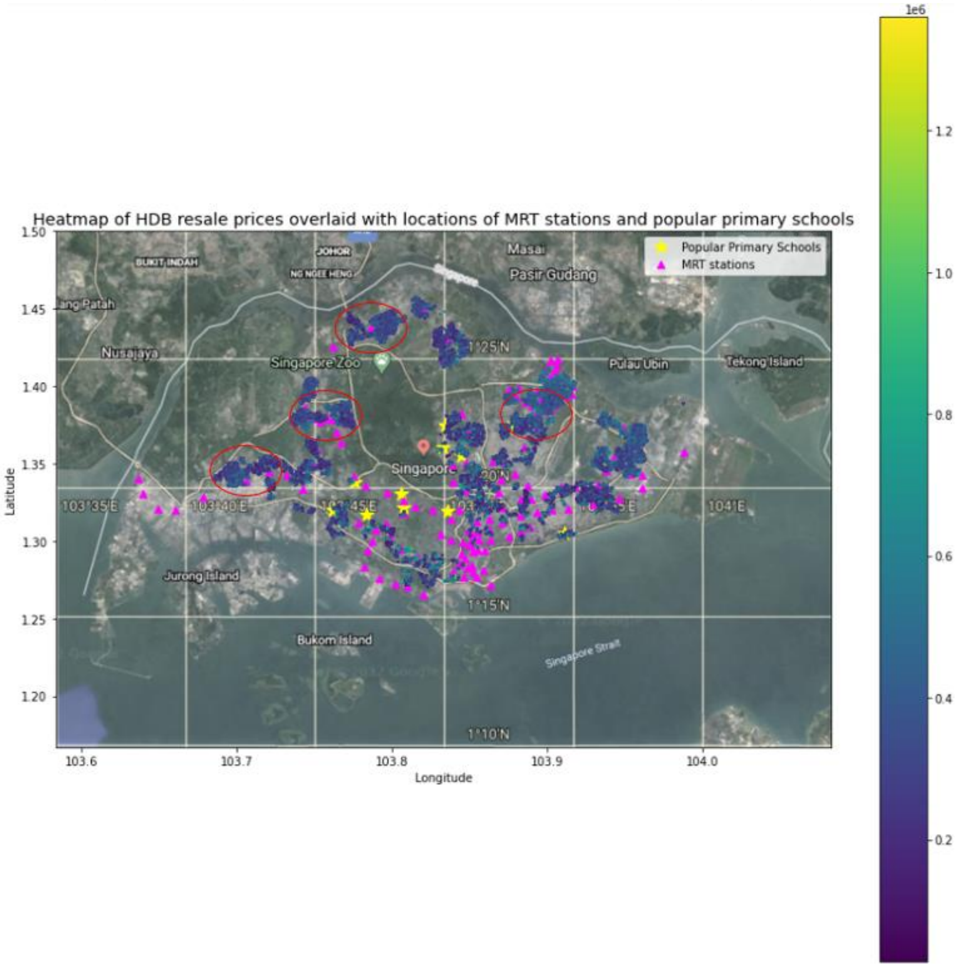
A custom transformer was built to perform the diverse types of data preprocessing, after which an HDB resale price prediction model was built, based on ordinary least squares (OLS) multiple variable linear regression. The OLS multiple variable linear regression model was selected due to its simple and satisfactory performance. More importantly, this is a white box model that can allow us to gain insights about how our features affect HDB resale prices.

From the model results, qualitative analysis was performed to explain the reason for the observed feature's impact on HDB prices.

Results and Findings

To visualize the data, a heatmap of HDB resale prices overlaid with locations of MRT stations and popular primary schools in Singapore was plotted.

Figure 1: Heatmap of HDB resale prices overlaid with locations of MRT stations and popular primary schools in Singapore



From the heatmap visualization, the impact of proximity to popular primary schools on HDB resale prices is not noticeably clear. On the other hand, as seen in some of the circled regions, the proximity to MRT stations does appear to have some impact on HDB resale prices, although the impact does not appear to be extremely strong as well.

To get a more quantitative visualization of the impact of the proximity to MRT stations and number of top primary schools within a 1 km radius on HDB resale prices, a plot of the regression coefficients for these features (together with other existing features), with their 95% confidence intervals, from most negative to most positive was generated as shown in Figure 2[7].

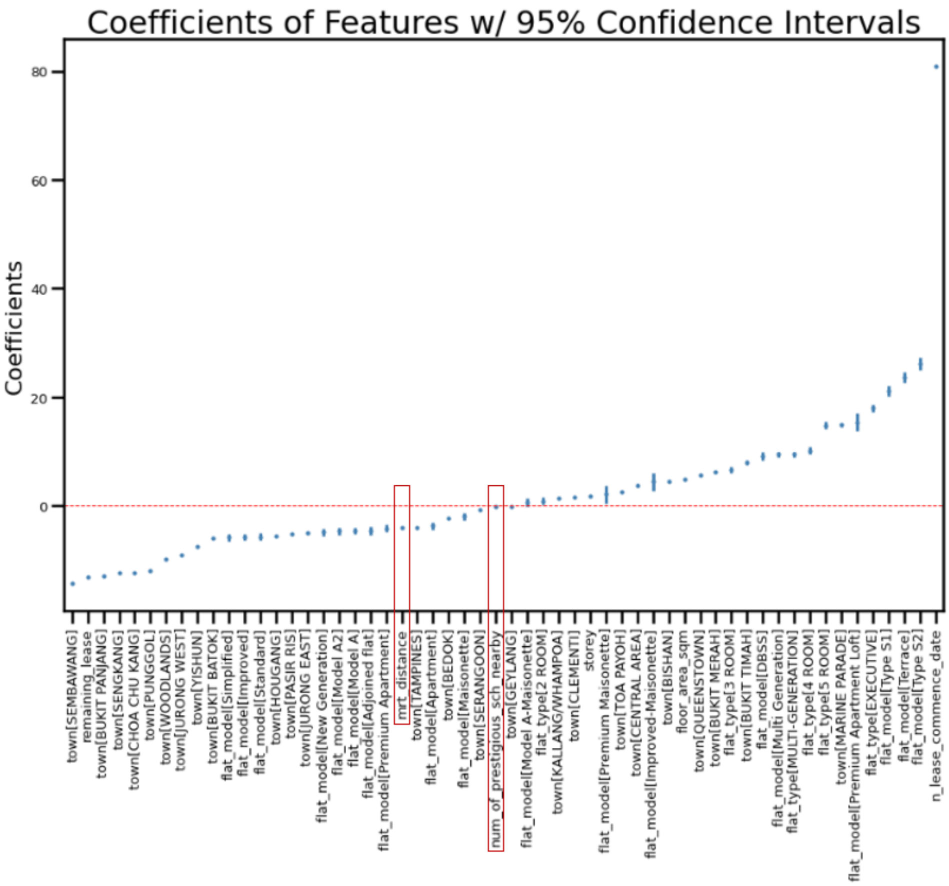
Impact of number of popular primary schools within 1 km radius on HDB resale value

Consistent with the observations from the heatmap, the regression coefficient of the number of popular primary schools within a 1 km radius is negligible, at only –0.2. However, the p-value for this coefficient being zero is negligible as well, as seen in Figure 3, which means that we can reject the null hypothesis and accept the alternative hypothesis that there is a non-zero coefficient for this feature.

The weak regression coefficient despite a low p-value is explainable. HDB resale buyers may or may not have children who are entering primary school, thus it is possible for the current data set to contain HDB resale transactions that were independent of this feature, if being near to a popular primary school did not matter to a buyer.

Overall, what we can infer from this analysis is that the number of popular primary schools within a 1 km radius of a HDB unit should have an impact on the HDB unit's resale value. However, more data will be required for us to get a representative coefficient, which extends beyond the scope of this project. These data may include the demographics of the resale buyer, specifically whether the buyer has children that are entering primary school soon.

Figure 2: Regression Coefficients of features with 95% Confidence Intervals



Impact of proximity to MRT stations on HDB resale value

From Figures 2 and 3, the proximity of an MRT station is inversely related to the resale price, with a regression coefficient of ~-4.1239. In other words, a HDB unit that is 1 km further away from an MRT station is worth about $41,239 less than another HDB unit that is 1 km nearer to an MRT station, all else being equal. The p-value for this coefficient being zero is negligible as well, as seen in Figure 3, which means that we can reject the null hypothesis and accept the alternative hypothesis that there is a non-zero coefficient for this feature.

Qualitatively, the results are in line with expectations, since a larger distance from an MRT station would reduce commuting convenience and lower the value of a HDB resale unit.

Figure 3: OLS Regression Results Summary

OLS Regression Results

Dep. Variable:	resale_price	R-squared (uncentered):	0.973
Model:	OLS	Adj. R-squared (uncentered):	0.973
Method:	Least Squares	F-statistic:	3.020e+05
Date:	Sat, 16 Apr 2022	Prob (F-statistic):	0.00
Time:	11:55:26	Log-Likelihood:	-1.4982e+06
No. Observations:	462083	AIC:	2.997e+06
Df Residuals:	462028	BIC:	2.997e+06
Df Model:	55		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
storey	1.6931	0.010	172.960	0.000	1.674	1.712
floor_area_sqm	4.8062	0.042	113.358	0.000	4.723	4.889
remaining_lease	-13.2129	0.016	-813.148	0.000	-13.245	-13.181
n_lease_commence_date	80.8623	0.104	780.520	0.000	80.659	81.065
flat_type[2 ROOM]	0.8823	0.329	2.686	0.007	0.238	1.526
flat_type[3 ROOM]	6.5963	0.321	20.530	0.000	5.967	7.226
flat_type[4 ROOM]	10.1244	0.332	30.457	0.000	9.473	10.776
flat_type[5 ROOM]	14.8027	0.351	42.193	0.000	14.115	15.490
flat_type[EXECUTIVE]	17.9560	0.379	47.420	0.000	17.214	18.698
flat_type[MULTI-GENERATION]	9.4373	0.228	41.335	0.000	8.990	9.885
town[BEDOK]	-2.3153	0.057	-40.864	0.000	-2.426	-2.204
town[BISHAN]	4.4911	0.081	55.622	0.000	4.333	4.649
town[BUKIT BATOK]	-5.9980	0.063	-95.122	0.000	-6.122	-5.874
town[BUKIT MERAH]	6.2190	0.068	91.599	0.000	6.086	6.352
town[BUKIT PANJANG]	-12.8376	0.072	-179.023	0.000	-12.978	-12.697
town[BUKIT TIMAH]	7.9456	0.190	41.866	0.000	7.574	8.318
town[CENTRAL AREA]	3.7441	0.131	28.569	0.000	3.487	4.001
town[CHOA CHU KANG]	-12.2591	0.066	-185.704	0.000	-12.389	-12.130
town[CLEMENTI]	1.5966	0.069	23.012	0.000	1.461	1.733
town[GEYLANG]	-0.1455	0.074	-1.964	0.049	-0.291	-0.000
town[HOUGANG]	-5.5009	0.060	-91.965	0.000	-5.618	-5.384
town[JURONG EAST]	-4.9205	0.072	-68.023	0.000	-5.062	-4.779
town[JURONG WEST]	-9.0403	0.059	-154.254	0.000	-9.155	-8.925
town[KALLANG/WHAMPOA]	1.3129	0.073	17.969	0.000	1.170	1.456
town[MARINE PARADE]	14.8735	0.122	121.675	0.000	14.634	15.113
town[PASIR RIS]	-5.1350	0.071	-72.806	0.000	-5.273	-4.997
town[PUNGGOL]	-11.9869	0.082	-146.262	0.000	-12.148	-11.826
town[QUEENSTOWN]	5.6220	0.075	75.278	0.000	5.476	5.768
town[SEMBAWANG]	-14.3767	0.084	-171.537	0.000	-14.541	-14.212
town[SENGKANG]	-12.4379	0.070	-176.578	0.000	-12.576	-12.300
town[SERANGOON]	-0.7502	0.075	-9.946	0.000	-0.898	-0.602
town[TAMPINES]	-4.0484	0.057	-70.446	0.000	-4.161	-3.936
town[TOA PAYOH]	2.5306	0.071	35.747	0.000	2.392	2.669
town[WOODLANDS]	-9.9175	0.059	-169.419	0.000	-10.032	-9.803
town[YISHUN]	-7.5923	0.058	-130.100	0.000	-7.707	-7.478
flat_model[Adjoined flat]	-4.5405	0.389	-11.686	0.000	-5.302	-3.779
flat_model[Apartment]	-3.7214	0.353	-10.529	0.000	-4.414	-3.029
flat_model[DBSS]	9.1408	0.364	25.107	0.000	8.427	9.854
flat_model[Improved]	-5.7792	0.336	-17.209	0.000	-6.437	-5.121
flat_model[Improved-Maisonette]	4.4285	0.843	5.255	0.000	2.777	6.080
flat_model[Maisonette]	-1.9496	0.355	-5.499	0.000	-2.645	-1.255
flat_model[Model A]	-4.5946	0.332	-13.839	0.000	-5.245	-3.944
flat_model[Model A-Maisonette]	0.6240	0.397	1.571	0.116	-0.155	1.403
flat_model[Model A2]	-4.6403	0.344	-13.490	0.000	-5.314	-3.966
flat_model[Multi Generation]	9.4373	0.228	41.335	0.000	8.990	9.885
flat_model[New Generation]	-4.8365	0.334	-14.473	0.000	-5.491	-4.182
flat_model[Premium Apartment]	-4.1633	0.336	-12.391	0.000	-4.822	-3.505
flat_model[Premium Apartment Loft]	15.4035	0.833	18.500	0.000	13.772	17.035
flat_model[Premium Maisonette]	2.0793	0.808	2.574	0.010	0.496	3.663
flat_model[Simplified]	-5.8619	0.341	-17.182	0.000	-6.531	-5.193
flat_model[Standard]	-5.6851	0.344	-16.538	0.000	-6.359	-5.011
flat_model[Terrace]	23.6362	0.466	50.721	0.000	22.723	24.550
flat_model[Type S1]	21.0669	0.509	41.377	0.000	20.069	22.065
flat_model[Type S2]	26.1752	0.629	41.608	0.000	24.942	27.408
mrt_distance	-4.1239	0.029	-142.632	0.000	4.181	-4.067
num_of_prestigious_sch_nearby	-0.2059	0.039	-5.280	0.000	0.282	-0.129

Omnibus:	23095.801	Durbin-Watson:	1.998
Prob(Omnibus):	0.000	Jarque-Bera (JB):	32807.440
Skew:	0.468	Prob(JB):	0.00
Kurtosis:	3.910	Cond. No.	1.20e+16

Conclusion

An ordinary least squares multiple linear regression model was built to predict HDB resale prices, using features such as flat types, model, and storey from HDB, as well as additional geological data from external sources to include the impact of proximity to MRT stations and popular primary schools. A satisfactory model performance was observed (R^2 of 0.97). There is an observable inverse relationship between the resale value of an HDB unit, and the distance between the unit and an MRT station. Although a negligible regression coefficient was obtained for the feature of the number of popular primary schools within 1 km of a HDB resale unit, we cannot reject the null hypothesis that the coefficient is 0 and will require additional surveys on the demographics of HDB buyers for further analysis.

References

1. Public Housing – A Singapore Icon, accessed 16 April 2022, <https://www.hdb.gov.sg/about-us/our-role/public-housing-a-singapore-icon#:~:text=The%20flats%20spell%20home%20for,optimal%20living%20environment%20for%20residents>

2. Resale Flat Prices (2022), accessed 16 April 2022, <https://data.gov.sg/dataset/resale-flat-prices>

3. Lee Y.X. (2019), Singapore Train Station Coordinates, accessed 16 April 2022, <https://www.kaggle.com/datasets/yxlee245/singapore-train-station-coordinates/metadata>

4. Primary School Ranking [2022], accessed 16 April 2022, <https://schoolbell.sg/primary-school-ranking>

5. Google Maps, accessed 2 April 2022, <https://www.google.com/maps>

6. Lee M.Y. (2020), Singapore HDB Postal Code Mapper, accessed 16 April 2022, <https://www.kaggle.com/datasets/mylee2009/singapore-postal-code-mapper>

7. Baldini, J.F. (2020), Create Your Own Coefficient Plot Function in Python, accessed 16 April 2022, <https://medium.com/analytics-vidhya/create-your-own-coefficient-plot-function-in-python-aadb9fe27a77>