

ST3131 Assignment: HDB Resale Prices

Introduction

This report aims to propose the optimal linear model for the prediction of HDB resale prices using the dataset 2021 HDB Resale Prices as released by government. Our objective is to produce a strong-performing linear predictor while identifying and correcting common model adequacy problems. To achieve the above goals, this report will adopt the following strategies:

1. Select and justify a set of variables influential to HDB resale prices to be adopted into the model as regressors.
2. Explore response variable to understand distribution and association with regressor variables and apply transformations if required.
3. Fit linear model and diagnose for inadequacies. Then proposing treatment via transformation, Weighted Least Squares (WLS)/Generalised WLS and/or Ridge regression
4. Compare competing models and select final best-performing linear model

The dataset consists of 11 variables, including resale price the response variable. Based on preliminary inspection and domain knowledge, three variables—block, street name, and lease commence date — were excluded because they either do not directly influence resale prices or their effects are already captured by other variables (e.g. remaining lease and town). This leaves seven explanatory variables: five categorical variables (month, town, flat type, storey range, and flat model) and two numerical variables (floor area and remaining lease).

Exploratory Data Analysis

Response Variable: Resale Prices (In Thousands)

The response variable in this dataset is the HDB resale prices, a numerical variable. Noting that HDB resale prices are large, they are scaled down, represented in the thousands (resale_scaled). We first visually inspect the distribution of Resale Prices using histograms and QQ-plot.

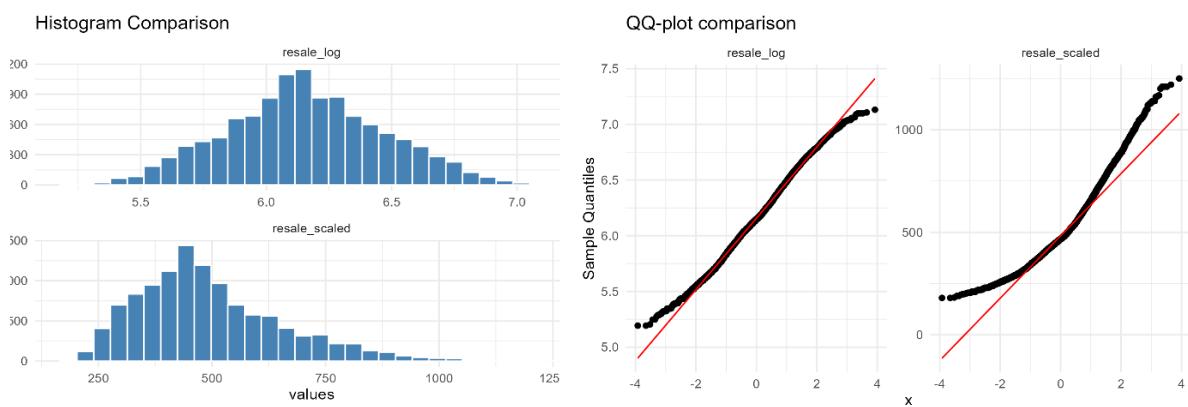


Figure 1 (left, histogram) and Figure 2 (right, QQ-plot)

The histogram indicates that the scaled resale price distribution is right-skewed, with a long tail of higher-priced flats. The QQ-plot further reveals a significant deviation of tails from the reference line, especially the upper tail, indicating the presence of extreme values and deviation from normality. This may violate key assumptions of linear regression, particularly the assumptions of linearity and constant variance of errors.

To address this, a logarithmic transformation of the resale price (`resale_log`) was applied. This reduces the impact of extreme values and produces a more symmetric distribution with more stable variance, as shown in the transformed histogram and QQ-plot. These adjustments help satisfy the assumptions of linear regression.

Relationship between Resale Prices and Numeric Regressors

We have two numeric variables, floor area (`sqm`) and remaining lease (years).

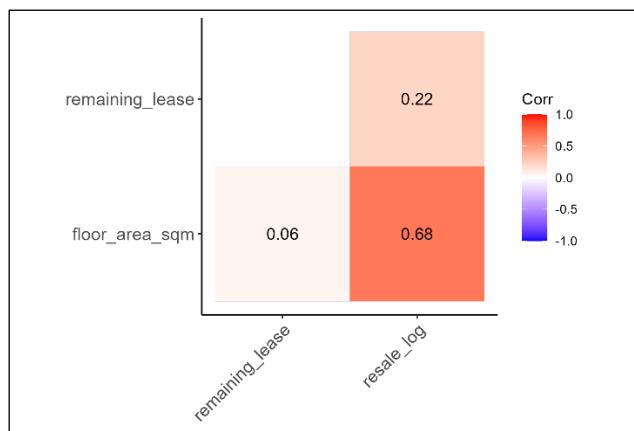


Figure 3: Correlation Plot for Numeric Variables

The correlation analysis shows that floor area has a strong positive linear association with resale prices ($r = 0.68$), confirming intuition that larger flats command higher resale values. Remaining lease shows a weaker linear correlation with resale price ($r = 0.22$), suggesting that remaining lease is not as influential. Since the correlation between the two predictors is low ($r = 0.06$), multicollinearity is likely negligible.

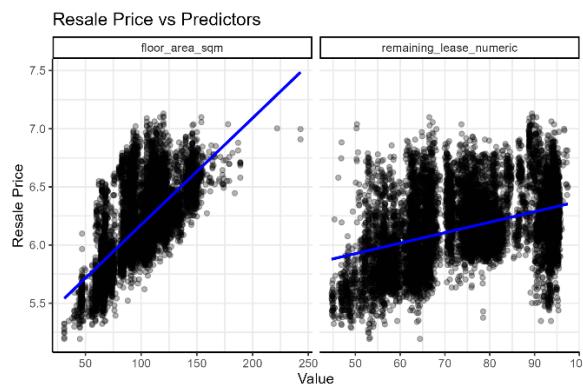


Figure 4: Scatter plot of Resale Price (log transformed) against Numeric Predictors

From the scatterplots, we can visually assess that there is low concern of non-constant variance or deviation from linearity. There are 3 large outliers for Floor Area (sqm) but it is a

negligible number compared to total sample size of 11527, thus data can be retained without distorting model significantly.

Relationship between Resale Prices and Categorical Regressors

ANOVA was conducted to test the significance of categorical predictors on the resale prices, using eta-squared (η^2) as the measure of influence. Eta-squared represents the proportion of total variation in resale prices that can be explained by differences between the categories of a predictor. Thus, a higher η^2 indicates a more influential categorical variable.

categorical variable	flat_type	flat_model	storey_range	town	town_region	month
F-value	2114	307.3	107.5	58.9	240.8	6.856
P-value	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16	2.14E-06
SSR	607.2	390.1	150.6	131.5	46.5	3.4
SST	1158.6	1158.7	1158.6	1158.7	1158.7	1158.6
ETA SQ (SSR/SST)	0.5240808	0.3366704	0.12998446	0.113489255	0.04013118	0.002934576

Figure 5: ANOVA Table for Categorical Regressors

The results show that flat type ($\eta^2 \approx 0.52$) and flat model ($\eta^2 \approx 0.33$) have high eta-squared values, indicating their high influence. Storey range ($\eta^2 \approx 0.13$) also shows a meaningful positive effect. For town, regrouping towns into three regions based on their distance from Central Area (CCR, RCR, OCR) results in a moderate effect size of $\eta^2 \approx 0.04$. This is lower than the original $\eta^2 \approx 0.11$ when using individual towns, but the reduction in dummy variables (from over 20 to 3) can be considered a reasonable trade-off to improve model simplicity and interpretability. Month has a negligible effect ($\eta^2 \approx 0.002$), of less than 1% of explaining the variance despite having a large 12 categories. With minimal explanatory power and 12 categories, it adds unnecessary complexity and is therefore excluded from further modelling.

Based on domain knowledge, flat type, storey range and region (distance from central) can be ordered logically and thus treated as ordinal categories that can be represented numerically instead of having to use many indicator variables. To justify this, we check for consistent trends across the categories.

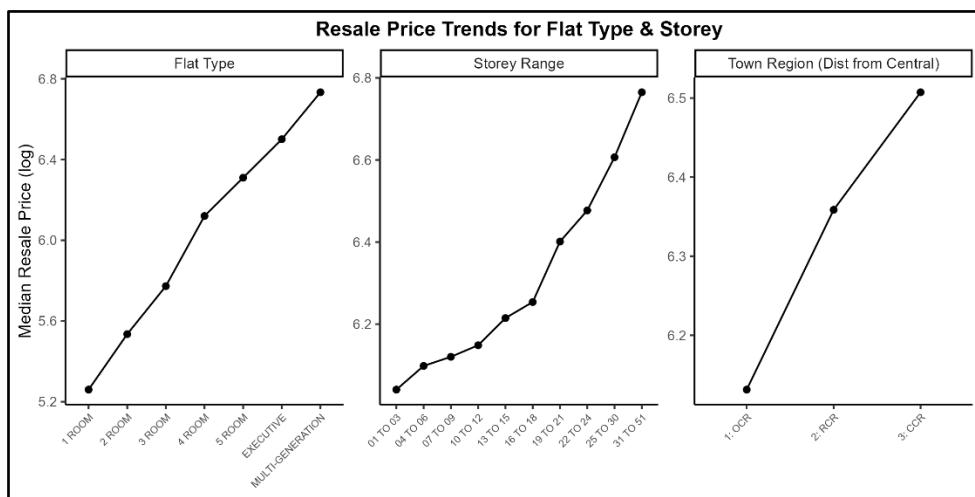


Figure 6: Ordinal Trend across Flat Type, Storey Range and Town Region (Distance from Central)

Flat type shows a clear trend: resale prices increase with flat size, from 1–5 room flats, followed by Executive and Multi-Generation. Storey range required regrouping because the original data is highly right-skewed, with sparse samples at the highest floors (e.g., 49–51 floor had only one flat). To reduce distortion from extreme values in small sample sizes, bins were merged to ensure each range had at least 1% of the total sample (≥ 115 flats out of 11,527). After regrouping, a consistent increasing trend is observed, higher storey ranges correspond to higher resale prices. Town Region (Distance from Central) also shows a consistent increase in resale prices, with flats closer to the city centre (OCR < RCR < CCR) commanding higher prices. Figure 6 visually confirms these ordinal trends, supporting the use of numeric encoding for both flat type and storey range in subsequent modeling. Only flat model remains as discrete categorical as there is no natural way to order its categories.

Building Models

Initial Model M_0

Interpreting coefficient estimation significance

We construct M_0 with 6 initial regressors: Flat Model (Reference Category: “2-room”), Town Region, Flat Type, Storey Range, Remaining Lease, and Floor Area.

Coefficients:						
β	(Intercept)	3.9246422	0.0927033	42.336	< 2e-16	***
1.	flat_model\Adjoined flat	0.1715531	0.0966094	1.776	0.07580	.
	flat_model\Apartment	0.0401387	0.0923943	0.434	0.66399	.
	flat_model\BSS	0.3715495	0.0922078	4.029	5.63e-05	***
	flat_model\Improved	0.0514082	0.0920298	0.559	0.57644	.
	flat_model\Improved-Maisonette	0.3822388	0.1590991	2.403	0.01630	*
	flat_model\Maisonette	0.1159734	0.0924758	1.254	0.20983	.
	flat_model\Model A	0.0935222	0.0918768	1.018	0.30874	.
	flat_model\Model A-Maisonette	0.2362754	0.0951898	2.482	0.01307	*
	flat_model\Model A2	0.0247309	0.0927729	0.267	0.78980	.
	flat_model\Multi Generation	0.1048291	0.1033972	1.014	0.31068	.
	flat_model\New Generation	0.1950836	0.0920219	2.120	0.03403	*
	flat_model\Premium Apartment	0.0954395	0.0919688	1.038	0.29941	.
	flat_model\Premium Apartment Loft	0.2718314	0.0955423	2.845	0.00445	**
	flat_model\Premium Maisonette	-0.0119980	0.1592504	-0.075	0.93995	.
	flat_model\Simplified	0.1482263	0.0922392	1.607	0.10809	.
	flat_model\Standard	0.0520301	0.0923889	0.563	0.57333	.
	flat_model\Terrace	0.6153791	0.1047127	5.877	4.30e-09	***
	flat_model\Type S1	0.1197895	0.0953403	1.256	0.20898	.
	flat_model\Type S2	0.0558476	0.0975273	0.573	0.56690	.
2.	town_region_ord	0.2865468	0.0032272	88.791	< 2e-16	***
3.	flat_type_ord	0.0652099	0.0058828	11.085	< 2e-16	***
4.	storey_ord	0.0288595	0.0006793	42.482	< 2e-16	***
5.	remaining_lease_numeric	0.0083416	0.0001368	60.964	< 2e-16	***
6.	floor_area_sqm	0.0079746	0.0002126	37.510	< 2e-16	***

Analysis of Variance Table

Response: resale_log	Df	Sum Sq	Mean Sq	F value	Pr(>F)
flat_model	19	390.05	20.53	1219.3	< 2.2e-16 ***
town_region_ord	1	44.04	44.04	2615.5	< 2.2e-16 ***
flat_type_ord	1	417.64	417.64	24805.2	< 2.2e-16 ***
storey_ord	1	47.97	47.97	2849.0	< 2.2e-16 ***
remaining_lease_numeric	1	41.62	41.62	2472.3	< 2.2e-16 ***
floor_area_sqm	1	23.69	23.69	1407.0	< 2.2e-16 ***
Residuals	11502	193.66	0.02		

Figure 7 (top): Coefficient Table & Figure 8 (bottom): ANOVA Table

The results indicate the 6 chosen predictors included in the model are statistically significant in explaining variation in resale (log) prices. From Figure 7, the estimate, t-value, and p-value show both the direction and strength of each regressor's effect. For instance, Flat Type has a estimate of 0.065 indicating that higher flat types are associated with higher resale prices.

Similarly, storey range, town region, remaining lease and floor area have significant positive coefficients, confirming that higher floors, proximity to central regions, longer remaining lease, and larger floor area increase resale prices. Notably, the p-values for all the above 5 regressors < 1% indicating high significance.

Although several individual flat model indicator variables have p-values > 1%, indicating less significance, the ANOVA table confirms that flat model as a categorical predictor is highly significant ($F\text{-value} = 1219.3$, $p < 2.2\text{e-}16$), justifying its inclusion in the model. Overall, these results confirm that all six predictors contribute meaningfully to explaining resale price variation, and the signs of the estimates are consistent with domain knowledge.

Model Adequacy Check

The adequacy of the initial model (M_0) was evaluated through a series of checks.

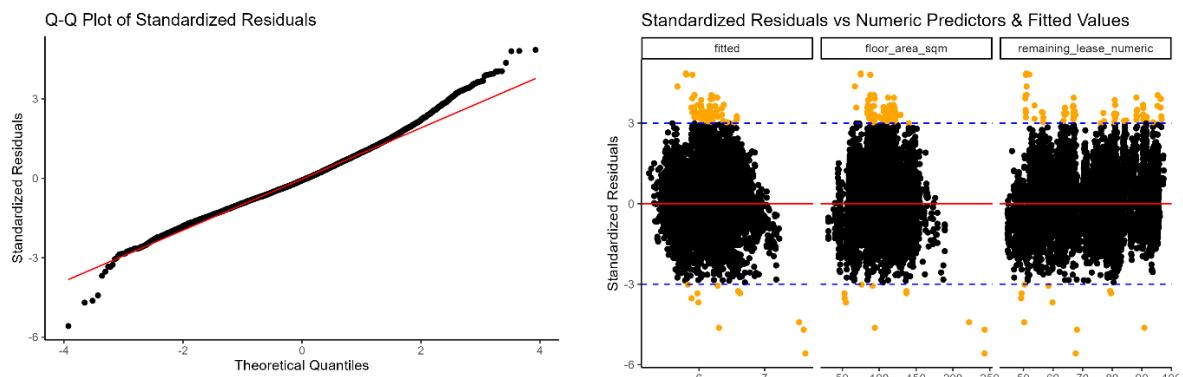


Figure 9 (left): QQ plot of SR and Figure 10 (right): SR against Numeric Predictors and Fitted Values

First, the normality of residuals was assessed using a QQ plot. Most standardised residuals (SRs) closely followed the reference line, indicating approximate normality. Mild deviations were observed in the tails, reflecting the presence of both high and low outliers, but can be accepted in a large dataset.

Next, linearity and constant variance were examined. The plot of SRs against fitted values showed no curvature or funnel shape, suggesting that the linearity and constant variance assumption likely hold. SRs plotted against the numeric predictors (floor area and remaining lease) also showed no usual pattern, further supporting the adequacy of the model.

Outliers were identified (orange points in Figure 10) with SRs exceeding an absolute value of 3, with slightly more positive outliers than negative ones. To assess their influence, a cleaned model ($M_0\text{_clean}$) was fitted after removing these 89 outlying observations. Comparison between M_0 and $M_0\text{_clean}$ (Figure 7 and 11) showed that coefficient estimates changed by less than 0.02, indicating the model is robust to these extreme values. R^2 increased slightly from 0.8329 to 0.8463, and the residual sum of squares decreased from 193.66 to 174.97. These minimal improvements do not justify excluding the outliers, particularly in the absence of a clear reason for the outliers.

	Coefficients:					
	(Intercept)	Estimate	Std. Error	t value	Pr(> t)	
1.	flat_modelAdjoined flat	0.1715531	0.0966094	1.776	0.07580 .	
	flat_modelApartment	0.0401387	0.0923943	0.434	0.66399	
	flat_modelDBSS	0.3715495	0.0922078	4.029	5.6e-05 ***	
	flat_modelImproved	0.0514082	0.0920298	0.559	0.57644	
	flat_modelImproved-Maisonette	0.3822388	0.1590991	2.403	0.01630 *	
	flat_modelMaisonette	0.1159734	0.0924758	1.254	0.20983	
	flat_modelModel A	0.0935222	0.0918768	1.018	0.30874	
	flat_modelModel A-Maisonette	0.2362754	0.0951898	2.482	0.01307 *	
	flat_modelModel A2	0.0247309	0.0927729	0.267	0.78980	
	flat_modelMulti Generation	0.1048291	0.1033972	1.014	0.31068	
	flat_modelNew Generation	0.1950836	0.0920219	2.120	0.03403 *	
	flat_modelPremium Apartment	0.0954395	0.0919688	1.038	0.29941	
	flat_modelPremium Apartment Loft	0.2718314	0.0955423	2.845	0.00445 **	
	flat_modelPremium Maisonette	-0.0119980	0.1592504	-0.075	0.93995	
	flat_modelSimplified	0.1482263	0.0922392	1.607	0.10809	
	flat_modelStandard	0.0520301	0.0923889	0.563	0.57333	
	flat_modelTerrace	0.6153791	0.1047127	5.877	4.30e-09 ***	
4.	flat_modelType S1	0.1197895	0.0953403	1.256	0.20898	
5.	remaining_lease_numeric	0.0083416	0.0001368	60.964	< 2e-16 ***	
6.	floor_area_sqm	0.0079746	0.0002126	37.510	< 2e-16 ***	

	Coefficients:					
	(Intercept)	Estimate	Std. Error	t value	Pr(> t)	
	flat_modelAdjoined flat	3.9173730	0.0884700	44.279	< 2e-16 ***	
	flat_modelApartment	0.0455673	0.0881659	0.517	0.60528	
	flat_modelDBSS	0.3756077	0.0879841	4.269	1.98e-05 ***	
	flat_modelImproved	0.0515960	0.0878160	0.588	0.55685	
	flat_modelMaisonette	0.1244455	0.0882452	1.410	0.15850	
	flat_modelModel A	0.0912608	0.0876645	1.041	0.29789	
	flat_modelModel A-Maisonette	0.2323048	0.0908274	2.558	0.01055 *	
	flat_modelModel A2	0.0300729	0.0885245	0.340	0.73408	
	flat_modelMulti Generation	0.1183300	0.0986887	1.199	0.23054	
	flat_modelNew Generation	0.1988587	0.0878076	2.265	0.02355 *	
	flat_modelPremium Apartment	0.1001365	0.0877547	1.141	0.25385	
	flat_modelPremium Apartment Loft	0.2664580	0.0911616	2.923	0.00347 **	
	flat_modelSimplified	0.1494009	0.0880229	1.697	0.08967 .	
	flat_modelStandard	0.0329098	0.0881821	0.373	0.70900	
	flat_modelTerrace	0.6862611	0.1015029	6.761	1.44e-11 ***	
	flat_modelType S1	0.1105031	0.0909692	1.215	0.22449	
	flat_modelType S2	0.0526913	0.0930617	0.566	0.57127	
	town_region_ord	0.2918120	0.0030973	94.215	< 2e-16 ***	
	flat_type_ord	0.0551502	0.0058907	9.362	< 2e-16 ***	
	storey_ord	0.0292185	0.0006519	44.820	< 2e-16 ***	
	remaining_lease_numeric	0.0083947	0.0001324	63.422	< 2e-16 ***	
	floor_area_sqm	0.0083287	0.0002135	39.011	< 2e-16 ***	

Figure 7 (left): original M0 coefficient table and Figure 11: M0_clean coefficient table

Leverage values were also checked using the diagonal of the hat matrix. A total of 327 observations exceeded the threshold of $2p/n$, indicating potential leverage. However, none of these points had a Cook's Distance greater than 1, suggesting that no observation exerted concerning influence on the fitted model. Consequently, all data points were retained in the final model.

Overall, M_0 satisfies the assumptions of linear regression adequately, with residuals approximately normal, variance constant and no evidence of influential observations distorting the model. The next step was to assess multicollinearity among predictors using the Variance Inflation Factor (VIF). For categorical predictors with multiple levels, namely Flat Model, the Generalized VIF (GVIF) was used.

	GVIF	Df	GVIF^(1/(2*Df))
flat_model1	7.518895	19	1.054525
town_region_ord	1.393092	1	1.180293
flat_type_ord	19.477575	1	4.413341
storey_ord	1.200412	1	1.095633
remaining_lease_numeric	2.665842	1	1.632741
floor_area_sqm	17.326046	1	4.162457

Figure 12: GVIF for 6 regressors

Figure 12 shows the $\text{GVIF}^{(1/(2\cdot Df))}$ values for the six regressors. All values are below 5, indicating low risk of multicollinearity and no corrective treatment is required.

Final Model M_N

Based on variable selection and adequacy checks, the final model M_N retains the same predictors as M_0 . Thorough investigation confirmed that the assumptions of multiple linear regression are satisfied, and thus M_N is effectively equivalent to the initial model.