STAT 230A Project Report
Predicting Resale Flat Prices in Singapore

Jangwon Lee, Yujian Zhou

May 13, 2025

1 Introduction

Singapore's housing market is globally distinctive, primarily due to the widespread presence of public housing managed by the Housing and Development Board (HDB). Over 80% of the resident population lives in HDB flats, making the resale market a critical area for real estate modeling and policy planning. Although the land is publicly owned, HDB flats are traded on an open resale market, leading to a hybrid system that blends public control with private market behavior. This hybrid structure presents an ideal setting for applying predictive modeling to understand price variation.

In this project, we aim to build models that accurately predict HDB resale prices based on a combination of temporal (e.g., remaining lease year), structural (e.g., flat characteristics), and spatial (e.g., transit access) features. We employ a range of modeling techniques including OLS, regularized regression (Lasso), and generalized linear models. While these models can provide insight into influential variables, our primary aim is predictive performance rather than causal inference.

2 Dataset

2.1 Original Dataset

Our primary dataset is sourced from Kaggle: Resale Flat Prices with Distance from Expressway. It contains information on the resale transactions of Housing & Development Board (HDB) flats in Singapore from 2015 to 2024, comprising over 220,000 observations. The key response variable is `resale_price`, accompanied by 10 predictors:

- `resale_price`: Transaction price (in Singapore dollars).
- `year`: Transaction year (2015–2024).
- `town`: Name of the town where the flat is located.
- `flat_type`: Flat type (e.g., 1-room, 2-room, 5-room, executive).
- `block`: HDB block number.
- `street_name`: Street address.
- `storey_range`: Floor range where the flat is located (e.g., 04 to 06).
- `floor_area_sqm`: Floor area in square meters.
- `remaining_lease_years`: Years remaining on the 99-year lease.
- `distance_from_expressway`: Distance to nearest expressway — $\leq$50m, 51–100m, 101–150m, 151–300m, 301–500m, >500m.

2.2 Integration of MRT/LRT Station Data

We enhanced our original dataset by integrating spatial information about MRT and LRT stations, sourced from the Kaggle dataset Singapore MRT/LRT Stations with Coordinates. This dataset includes the names and geographic coordinates (latitude and longitude) of all MRT and LRT stations in Singapore, allowing detailed spatial analyses of public transit accessibility.

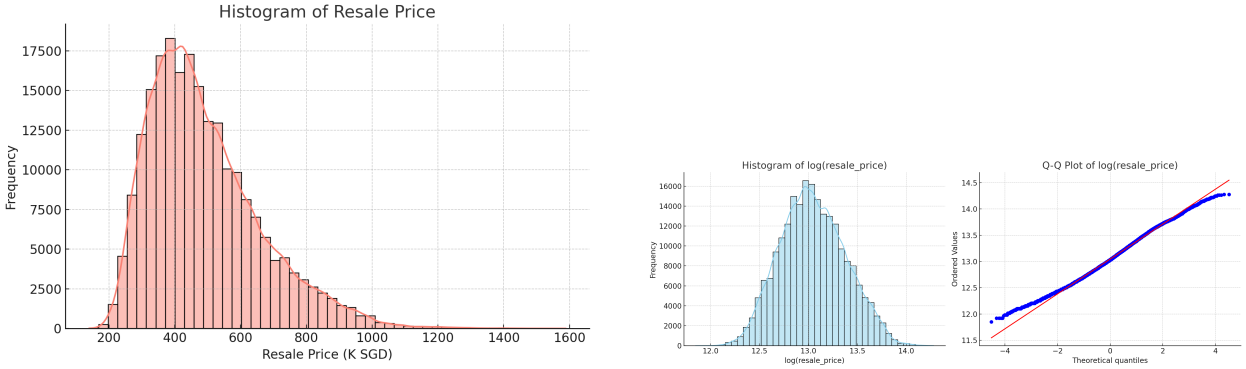
Using this spatial data, we derived two key accessibility metrics for each HDB flat:

1. `Distance to Nearest Station (within 1 km):` Capturing walkable proximity to public transit.
2. `Number of Stations within a 1 km radius:` Reflecting the density and accessibility of nearby transit options.

The 1 km threshold was chosen based on urban planning literature, suggesting this distance as a generally acceptable walking range in urban contexts.

3 Model

Our exploratory data analysis shows that the distribution of the response variable `resale_price` exhibits a strong right skewness. As shown in Figure 1, most resale transactions cluster in the \$300k–\$500k range, but a significant right tail stretches beyond \$800k, indicating the presence of high-end resale units. We believe Gamma Regression with a log transformation would be a natural choice, as it is specifically designed to model positive continuous outcomes with skewness. More importantly, we believe this model would be robust to heteroskedasticity since Gamma regression assumes that the variance of response would be proportional to the square of the mean: $Var(Y_i|x_i) = \phi\mu_i^2$, where μ_i is expected value of the response and ϕ is a dispersion parameter. Besides, we also tried to make different judgment calls during the feature engineering process



(a) Resale price (SGD in thousands)

(b) Log-transformed resale price and Q-Q plot

Figure 1: Comparison of the original and log-transformed distributions of resale prices. The original distribution (left) is highly right-skewed, while the log-transformed version (right) is more symmetric and approximately normal.

and generated different design matrices. The major difference lies in how we treat the feature `street_name`: our dataset contains approximately 218,000 observations, and there are 566 unique values for `street_name`. Even the most frequently occurring street appears only 3,279 times, which is less than 2% of the total observations, while many other streets appear only a handful of times. Although `street_name` is a highly informative spatial feature, it is also strongly correlated with other spatial variables, such as `town`. To balance interpretability and predictive performance, we created two versions of the design matrix:

- **Version 1:** Exclude `street_name`. This version results in a more parsimonious model with fewer predictors, making it easier to interpret and more suitable for statistical inference.
- **Version 2:** Include `street_name` using encoding. While this version introduces multicollinearity and reduces interpretability, it typically achieves higher predictive accuracy due to finer spatial resolution.

This modeling choice represents a classic trade-off: accuracy versus interpretability. Our final decision depends on the application context—whether the goal is to explain price drivers or to maximize predictive precision.

Table 1: Comparison of Gamma Regression Models

Metric	**Without** `street_name`	**With** `street_name`
No. Predictors	33	446
Dispersion (Scale)	0.0142	0.0107
Pseudo R^2 (Cragg–Uhler)	0.9990	0.9999
Test RMSE (SGD)	58,332.58	**50,064.63**

To validate the core assumption of Gamma regression that the conditional variance of the response is proportional to the square of the mean we conducted an empirical diagnostic using the test data. We binned the predicted means $\hat{\mu}_i$ into 100 quantile-based intervals and computed the sample variance of the actual resale prices Y_i within each bin. As shown in Figure 2, the empirical variances increase with the predicted means, and the trend closely follows a fitted reference curve proportional to μ^2(red dashed line). This alignment indicate that the Gamma model with a log link is well-justified for this dataset.

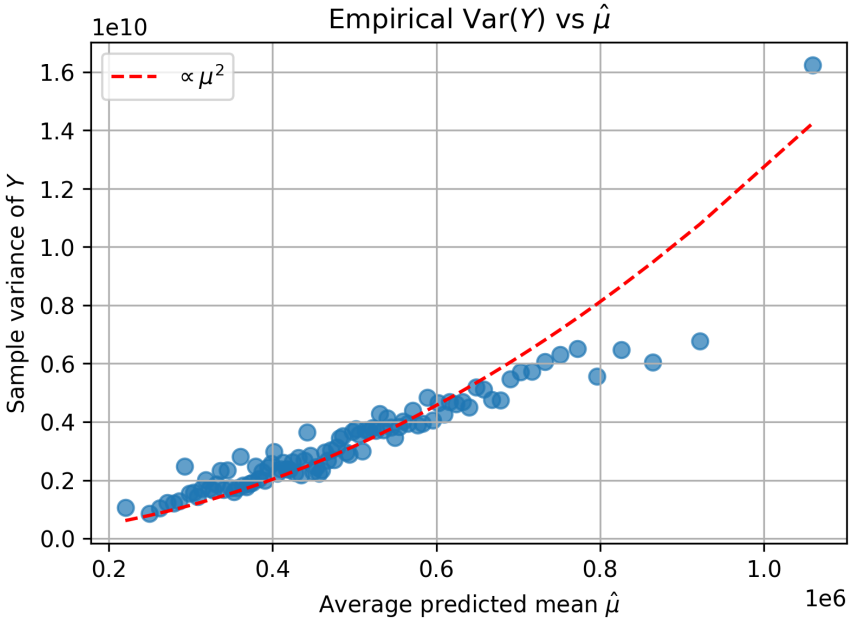


Figure 2: Empirical relationship between the average predicted mean and sample variance

4 Discussion

4.1 Results Interpretation

Although our main task is to build prediction models rather than perform causal inference, we can still interpret our Gamma regression results to understand how predictors are associated with changes in resale prices. Since we use a log link function, each coefficient reflects a multiplicative effect on the expected resale price, which we convert to a percentage. For example, remaining lease has a positive effect($\beta = 0.0106$), indicating that each additional remaining lease year is associated with a 1.06% increase in expected price. We also see meaningful effects from ordinal features: moving up one level in flat type (e.g., from 1-room to 2-room) increases expected price by 3.98%. Interestingly, greater distance from expressways is positively associated with price (1.26% per category), possibly reflecting preferences for quieter or less polluted environments. In terms of spatial variation, some towns stand out with large premiums over the baseline town. Overall, the model aligns well with market intuition: resale prices are higher for newer, larger

flats in central or prestigious towns with better transport connectivity. Although these associations are not causal, they provide valuable insights into the factors that drive housing market dynamics in Singapore.

4.2 Limitation

While our Gamma regression model demonstrates strong predictive performance and yields interpretable associations, several limitations should be noted regarding both the modeling process and the interpretation of results.

A central challenge in our analysis is the treatment of categorical variables, many of which have a high number of unique levels. Encoding such a feature using one-hot encoding leads to a large and sparse design matrix, which increases the risk of multicollinearity and can make the matrix singular. Additionally, some categorical features such as `distance_from_expressway` has a natural ordering($\leq 50, 51-100, 101-150$, etc), which we encoded using ordinal values for simplicity. However, this assumes equal spacing and consistent effects across levels, which may not hold in reality. For example, the impact of moving from ≤ 50 to 51–100 usually differ from moving from $51-100$ to $101-150$, even though both represent a one-unit increase. Such simplifications may lead to model misspecification or loss of nuanced spatial effects.

Previously we checked the empirical variances increase with the predicted means (Figure 2), however, for larger predicted values (especially above 800,000 SGD), the empirical variance tends to fall below the theoretical curve, except for extreme observations in the top percentile. To formally test this relationship, we considered a generalized variance model $Var(Y_i) = \phi\mu_i^{\alpha}$. This inspire us to regress log-squared residuals on log predicted means since we have $log((Y_i-\hat{\mu}_i)^2) \approx log(\phi)+\alpha log(\hat{\mu}_i)$. This model produced coefficient 1.3732 with a 95% confidence interval [1.306, 1.441]. This result indicates that the strict Gamma assumption($\alpha = 2$) may overstate the heteroskedasticity, particularly in the high-price range.

Finally, while all features in the model were statistically significant, the real economic significance is harder to evaluate due to a lack of domain knowledge. For example, our model indicates that each additional year of remaining lease is associated with a 1.06% increase in expected resale price. But we are unable to assess whether this magnitude will have economical influence on transaction market.

Table 2: Interpretation of Gamma Regression Coefficients

Predictor	$\boldsymbol{\beta}$	$e^{\boldsymbol{\beta}}$	Change in Expected Price
Year	0.0437	1.0447	+4.47%
Floor area (sqm)	0.0083	1.0083	+0.83%
Remaining lease (years)	0.0106	1.0106	+1.06%
# Nearby MRTs	0.0084	1.0084	+0.84%
Distance to MRT	-0.0001	0.9999	-0.01%
Flat type (ordinal)	0.0390	1.0398	+3.98%
Distance to expressway (ordinal)	0.0125	1.0126	+1.26%
Storey range (ordinal)	0.0233	1.0236	+2.36%
Selected Town effects (relative to reference category ANG MO KIO)			
Marine Parade	0.4678	1.596	+59.6%
Central Area	0.3022	1.352	+35.2%
Bukit Timah	0.2734	1.314	+31.4%
Bishan	0.1598	1.173	+17.3%
Queenstown	0.1395	1.150	+15.0%
Jurong West	-0.2409	0.785	-21.5%
Woodlands	-0.3278	0.721	-27.9%
Sengkang	-0.3620	0.696	-30.4%
Choa Chu Kang	-0.3707	0.690	-31.0%
Punggol	-0.3488	0.706	-29.4%

5 Conclusion

In this project, we aimed to develop an interpretable model to predict the resale prices of Housing Development Board (HDB) flats in Singapore based on temporal and spatial features as well as flat attributes. We applied multiple linear based model and pick Gamma Regression as our final approach. This model is appropriate for the task, given that the response is right skewed and exhibit heteroskedasticity. We also validated the model's assumptions through residual diagnostics and variance checks, confirming that the variance increases with the mean, though may not be strictly quadratic.

The model is practical for predicting future resale prices. It can be used as a reliable forecasting tool for policymakers, urban planners, and real estate stakeholders. While more sophisticated machine learning models like gradient boosting or random forests could potentially achieve better performance than our simple GLM model, our model offers a strong balance between accuracy and interpretability. Its interpretability and predictability make it useful for evaluating how changes in flat characteristics can influence resale values.

A. Additional Work and Alternative Approaches

A1. Feature Engineering Part

Initially, we considered incorporating detailed commuting times to Singapore's Central Business District (CBD) via both private vehicles and public transit to comprehensively measure urban accessibility. However, practical constraints, including high API usage costs and substantial prediction uncertainty, led us to abandon these plans. Specifically, Google Maps API costs were prohibitive, and alternative datasets, such as public transport smart card data[1], were inaccessible.

Given Singapore's extensive MRT/LRT network and the policy-driven limitations on car ownership, we subsequently excluded private vehicle accessibility analyses from our final modeling. Furthermore, based on findings from Agarwal (2020)[1], which indicate that working adults in Singapore tend to prefer the MRT system over buses due to its superior speed and reliability, we chose to exclude buses from our analysis. Instead, we focused exclusively on MRT/LRT accessibility, manually calculating Haversine distances between each HDB flat and nearby stations using coordinates obtained from Singapore's OneMap API. Although this approach lacks the precision of actual walking distances, it was a practical compromise given the resource limitations.

These additional explorations highlight both the complexity and practical challenges involved in quantifying urban accessibility rigorously.

A2. Other Models

We used OLS as the baseline model. However, this model exhibited signs of high multicollinearity among predictors. We applied regularization including Ridge and Lasso. Given that our dataset involves a relatively small number of predictors compared to the number of observations($p << n$), the benefits of regularization were limited. In particular, cross-validation will select penalty coefficient close to 0.

To complement our continuous modeling approach, we also explored a classification-based framework by discretizing the resale prices and applying a logistic regression model with multiple ordered categories. This approach is meaningful in the context of real estate and urban planning, where classifying flats into broad price tiers (e.g., low, medium, high) can help stakeholders make quick, interpretable decisions without requiring precise price predictions. However, this approach introduced two key challenges: First, we lack domain knowledge to set cut points. Ideally, discretization boundaries would reflect meaningful thresholds in market segmentation (e.g., policy eligibility, affordability, or buyer behavior). Without related expertise, we used decile-based cuts and such a split may limit the real-world relevance of the classification. Second, When fitting the ordinal logistic regression model, we encountered computational issues related to the Hessian matrix. Depending on how high-cardinality categorical variables were encoded, the model's design matrix sometimes led to a non-invertible Hessian. This instability underscores the importance for careful feature engineering.

B. Exploratory Data Analysis (EDA)

Due to space constraints, we include only a subset of our analysis here; for additional exploratory results, please refer to `Part1_EDA.ipynb`.

B1. Overview of Key Variable Distributions

Our exploratory analysis begins with examining the distribution of the response variable `resale_price`, which exhibits a strong right skew. As shown in Figure 1, most resale transactions cluster in the \$300k–\$500k range, but a significant right tail stretches beyond \$800k, indicating the presence of high-end resale units. This skewness motivates the later use of both logarithmic transformation and Gamma regression to better capture model fit and account for heteroskedasticity.

We also explored correlations and distributions among the main covariates. Among structural features, `floor_area_sqm` shows a clear positive association with resale price, while `flat_type` demonstrates categorical effects that align with expected price gradients (e.g., 5-room flats tend to cost more than 3-room flats). Temporal variables like `remaining_lease_years` also have significant explanatory power, as lease decay reduces eligibility for financing and government grants, affecting market valuation.

On the locational side, we observed nuanced patterns. Units closer to MRT stations generally show higher resale prices, consistent with expectations about improved accessibility. Conversely, units within 100m of an expressway often have slightly lower prices, potentially reflecting disamenity effects such as noise and air pollution.

These initial findings support the inclusion of both structural and spatial variables in our modeling framework, and motivate further statistical exploration of how each factor contributes to overall housing value.

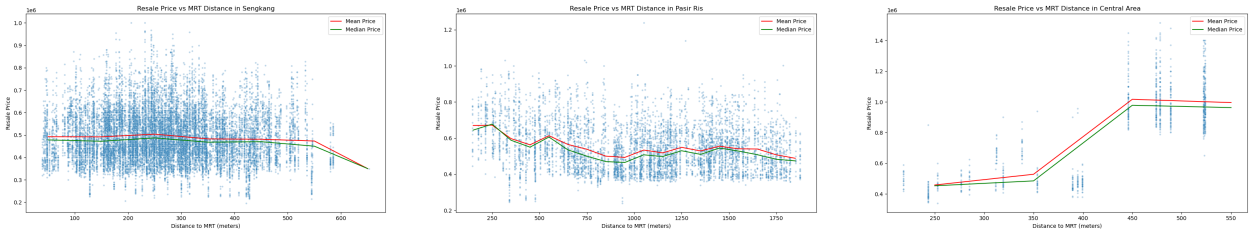
B2. Transit Proximity Heterogeneity Across Towns

Our EDA revealed that the relationship between MRT/LRT proximity and resale prices varies markedly by town. We highlight three illustrative examples in Figure 3.

Sengkang. Located in the northeastern periphery of Singapore, Sengkang was developed beginning in the early 2000s to absorb rapid population growth and relieve housing shortages. From its inception, Sengkang's HDB estates were built in a ring around the new MRT stations and the circular LRT loop that feeds into them. As a result, almost all flats lie within 500 m of a rail stop, and resale prices remain remarkably uniform across distance (Figure 3 a). This strong spatial planning has effectively neutralised proximity as a price differentiator in Sengkang.

Pasir Ris. At Singapore's northeastern edge, Pasir Ris is served by only one MRT station on the East –West line. Here, flats within 300 m of the station command a clear price premium, with both mean and median resale prices declining by 10–15% beyond that threshold (Figure 3 b). The single-station topology concentrates the accessibility effect into a tight catchment area.

Pasir Ris. At Singapore's northeastern edge, Pasir Ris is served by only one MRT station on the East–West line. In this setting, flats located closer to the station tend to command higher resale prices, while prices



(a) Sengkang (b) Pasir Ris (c) Central Area

Figure 3: Resale price vs. distance to nearest MRT/LRT, by town. Red and green lines show mean and median prices in distance-bins.

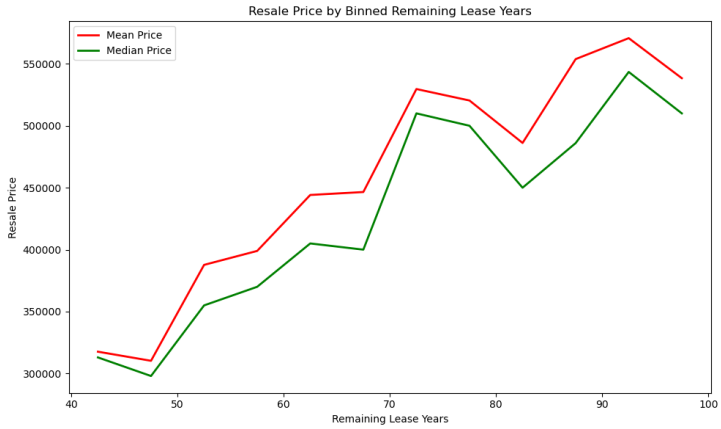


Figure 4: Mean and median resale price by binned remaining lease years.

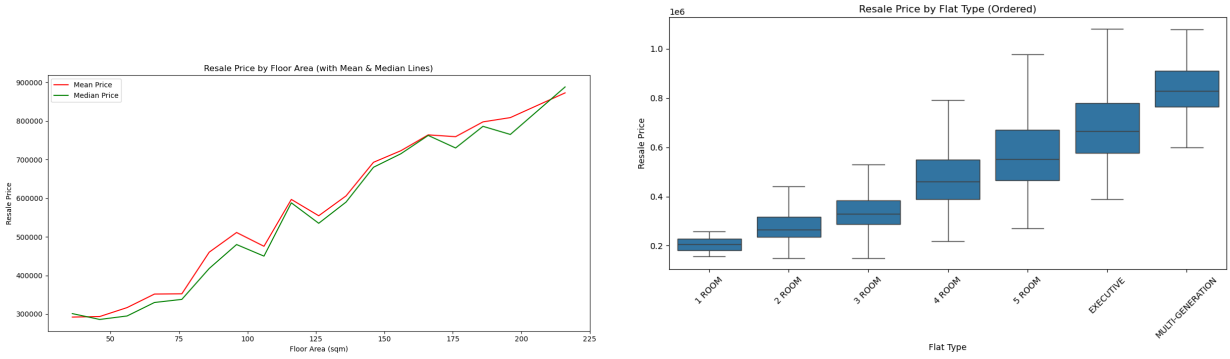
gradually decline as distance increases (Figure 3 b). The single-station topology concentrates the accessibility effect within a limited radius, reinforcing the value of proximity in transit-sparse areas.

Central Area. By contrast, the Central Area is already saturated with MRT, LRT, and bus links, and alternative transport modes (e.g., taxis) abound. Interestingly, the plot for this region shows a slight upward trend in both mean and median resale prices with increasing distance from the nearest MRT station (Figure 3 c). Although the limited sample size makes it difficult to draw definitive conclusions, one possible explanation is that flats very close to central MRT stations may suffer from noise, congestion, or other disamenities, slightly lowering their residential appeal. This suggests that in highly urbanised areas, extreme proximity may not always be desirable for housing.

B3. Effect of Lease Decay on Resale Price

Figure 4 plots the mean and median resale prices against remaining lease years, binned in five-year intervals. A stark monotonic pattern emerges: flats with shorter residual leases exhibit substantially lower prices. The median line closely tracks the mean, underscoring that this decline is not driven by a few outliers but reflects a genuine market premium on long leases.

This relationship is economically intuitive. HDB leasehold properties lose economic value as the lease term shortens: financing becomes more restrictive, CPF grants taper off, and potential buyers factor in the diminishing tenure and higher exit risk. By capturing this lease decay non linearly—via binned intervals—we allow the regression models to flexibly account for accelerating price drops as leases enter their latter decades.



(a) Mean and median resale price by floor area (10-sqm bins)

(b) Boxplot of resale price by flat type

Figure 5: Relationship between dwelling size and resale price using (a) continuous floor area and (b) flat type categories.

B4. Impact of Dwelling Size on Resale Price

A strong, intuitive relationship between home size and resale price is clearly visible in Figure 5a and Figure 5b. Larger dwellings consistently command higher prices in both continuous and categorical specifications of size.

Figure 5a plots average and median resale prices against floor area (in square meters), binned into 10-sqm intervals. The relationship is close to linear, with both lines rising steadily across the spectrum. This suggests that each additional unit of space contributes positively to housing value.

Complementing this, Figure 5b shows a boxplot of resale prices by flat type, which is a discrete proxy for floor size. As expected, prices increase with room count: from 1-room to 5-room units, and further to Executive and Multi-Generation flats. The boxplots reveal substantial price dispersion within each type, but the median price aligns closely with type ordering, reinforcing the explanatory power of flat size.

B5. Resale price by proximity to expressway

Figure 6 plots average resale prices against six distance bands from the nearest expressway. A clear upward trend is observed: units located within 50 meters of an expressway show the lowest resale values, while prices increase steadily with greater separation. This pattern is consistent with prevailing assumptions about environmental disamenities: flats adjacent to expressways tend to suffer from higher noise levels, air pollution, and reduced aesthetic appeal, all of which can deter potential buyers and depress market value.

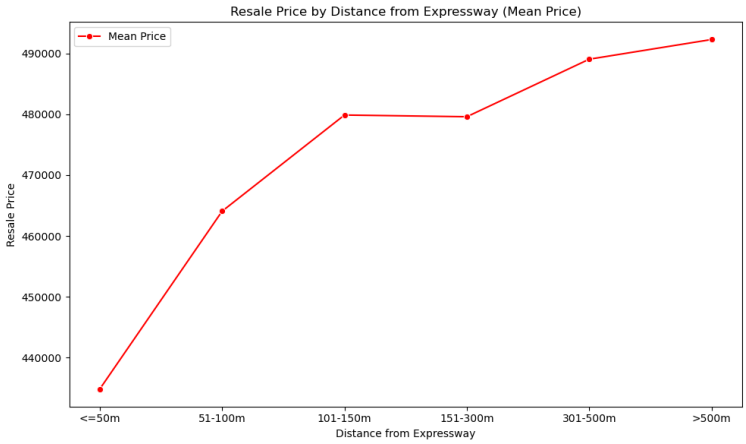


Figure 6: Mean resale price by distance from the nearest expressway.

C. Statement of Contribution

Throughout the project, both team members maintained active communication and jointly contributed to all major decisions, including the research design, model selection, and feature engineering strategy. The report itself was collaboratively written, with each section undergoing mutual review and editing.

- **Jangwon Lee:** Conducted literature review, curated and preprocessed datasets, designed and implemented feature engineering—particularly for spatial accessibility features—and contributed substantially to exploratory data analysis (EDA) and wrote corresponding sections of the final report.
- **Yujian Zhou:** Led the development and implementation of modeling and regression methods, including OLS, regularization (Ridge and Lasso), and alternative modeling strategies. I also contributed to the interpretation of model results and wrote corresponding sections of the final report.

References

[1] Sumit Agarwal et al. "Preferences of public transit commuters: Evidence from smart card data in Singapore". In: *Journal of Urban Economics* 120 (2020), p. 103288. DOI: 10.1016/j.jue.2020.103288. URL: https://doi.org/10.1016/j.jue.2020.103288.

[2] D. Bax, T. Zewotir, and D. North. "A gamma generalised linear model as an alternative to log linear real estate price functions". In: *Journal of Economic and Financial Sciences* 12.1 (2019), a476. DOI: 10.4102/jef.v12i1.476.

[3] Peng Ding. *Linear Model and Extensions*. https://arxiv.org/abs/2401.00649v1. Accessed May 2025. 2024.

[4] Institute of Real Estate and Urban Studies. *Aging and Decaying Leases of Residential Properties*. https://ireus.nus.edu.sg/2019/04/06/aging-and-decaying-leases-of-residential-properties/. National University of Singapore. 2019.

[5] Yong Tu and Grace K. M. Wong. "Public Policies and Public Resale Housing Prices in Singapore". In: *International Real Estate Review* 5.1 (2002), pp. 115–132.