**Team 053 Final Report**

# **OBJECTIVE/PROBLEM**

**Project Title**  
**Changing Demographics and its Impact on Local Public Housing Resale Prices in Singapore**

**Background Information**

The Housing & Development Board (HDB) serves as Singapore's public housing authority and is entrusted with the task of developing housing estates in a manner that ensures affordability, quality, and a pleasant living environment. Currently, HDB flats houses more than 80% of Singapore’s resident population [1].

Prime Minister Lee acknowledges that there are supply-demand imbalances in HDB markets [2] which has resulted in higher resale prices for the past 3 years. The escalating housing prices have sparked concerns regarding housing affordability which is especially pressing for young couples who are eager to establish their families and desire to own their own homes. Unfortunately, the rapid increase in housing prices has outpaced salary growth, resulting in diminished affordability in public housing.

**Problem Statement**

Our project hopes to facilitate better policy planning and outcomes through the study of how the changing demographic structure in Singapore affects the prices of resale public housing flats.

**Research Questions**

Primary research question: “*How do local population demographics affect the resale public housing prices*?”

Supporting research questions**:**

1. Does the rate of increase in a particular age group result in a greater than proportionate rate of increase in public housing prices?
2. Does an ageing population result in overall higher public housing prices?

**Business Justification**

One of the main objectives of the government is to ensure that public housing remains affordable for all, regardless of background, age, or educational qualifications.

One metric to measure house affordability is the Housing Price to Income Ratio (PIR) [3].

|  |  |
| --- | --- |
| Period | PIR |
| 2001 – 2020 | 4.1 |
| 2021 | 4.5 |
| 2022 Q1 | 4.8 |

Table 1: Changes in housing price to income ratio between 2001 and 2022

The table shows a steep increase of PIR from 2020 to 2022 Q1. Generally, a PIR ranging from 4.1 to 5.0 indicates that housing is seriously unaffordable and going beyond this range signifies that housing is severely unaffordable [4].

A predictive model can be constructed to determine the demographic age group that influences housing prices most significantly, so that policymakers can implement policies tailored to the needs of this age group. Hence, this project aims to assist policy makers in devising more effective and efficient policies to tackle the rising public housing prices, focusing on the segments of the population that matter the most.

**Anticipated Conclusions/Hypothesis**

Our hypothesis suggests that changes in age group population increase the demand in public housing. As a result, this increased demand relative to housing supply will drive up housing prices.

**Expected impact and benefits from our analysis**

This project primarily benefits the government and policymakers as it enables them to utilize predictive analysis for formulating effective housing policies. By comprehending the relationship between demographics, inflation, and HDB housing prices, policymakers can implement appropriate measures to ensure housing affordability and maintain stability in the housing market. For instance, if the senior age group has a significant influence on housing prices and the aging population is increasing, policymakers may consider constructing more short-term lease retirement homes to cater to the needs of active, independent seniors.

Conversely, for investors, property developers, homebuyers, and sellers, this analysis provides valuable insights that facilitate the evaluation of long-term viability and profitability, risk management, and the adjustment of strategies to make well-informed decisions regarding property acquisitions and portfolio management. It also enhances public understanding of the resale market for HDB flats in Singapore.

# **DATASETS AND DATA WRANGLING**

**Dataset 1:** Our primary dataset is the HDB resale data obtained from Data.gov.sg, which serves as a government centralized platform hosting diverse government datasets across domains such as the economy, education, and society. We consolidated the data that was initially split as multiple files into one.

A picture containing text, screenshot, font, number

Description automatically generated

Figure 1: Snippet of the HDB resale dataset

**Dataset 2:** The second dataset that we used is the Population and Population Structure of Singapore, which includes information such as the Singapore Citizen population, Permanent Resident population, population growth, and population breakdown from 1950 to 2022 [6]. The data was transformed from a columnar to a row format for ease of our analysis and extra rows that were irrelevant were removed.

A picture containing text, font, number, line

Description automatically generated

Figure 2: Snippet of the Population and Population Structure of Singapore dataset

**Data Cleaning:**

The two datasets mentioned above were merged based on year. Each row of the final dataset now consists of HDB information along with the population numbers by age range for the corresponding year. Necessary data cleaning such as removing missing and duplicate values was done and the dataset was limited to recent years from 2015 to September 2022. Column headers were cleaned to eliminate any spaces and to ensure it does not start with a number. Lastly, categorical variables were one hot encoded.

**Exploratory Data Analysis:**

Histograms, scatterplots, probability plots, boxplots and summary statistical analysis were conducted. Below are some noteworthy findings summarized in this section.

The left box plot below displays the adjusted price per square-meter for several types of flats, notwithstanding the year and location. It is observed that 3, 4, and 5 room flats have a higher proportion of outliers. Additionally, the box plots for 2 to 5-room and executive flats exhibit a longer right tail, indicating a positive skewness. It can be initially speculated that these patterns may be attributed to the increasing housing prices or other factors.

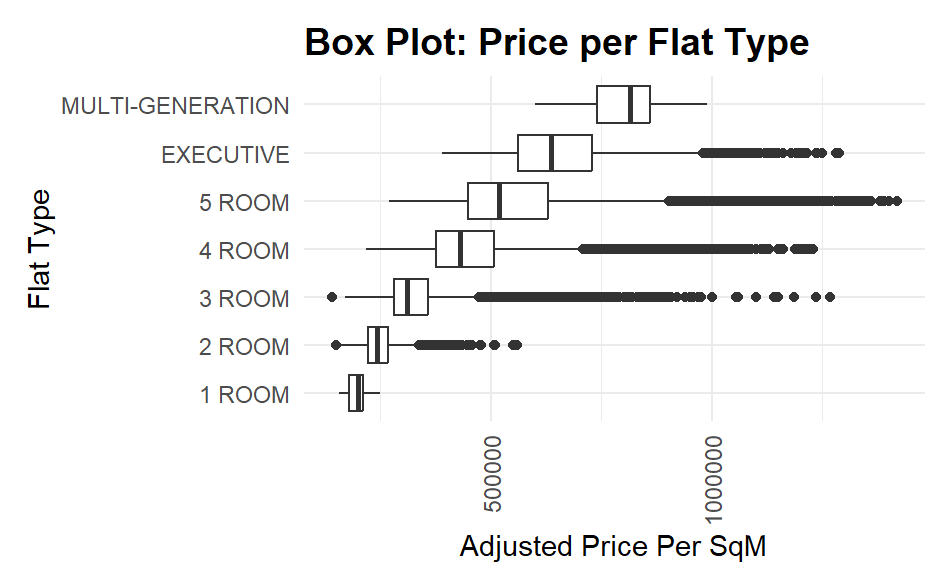
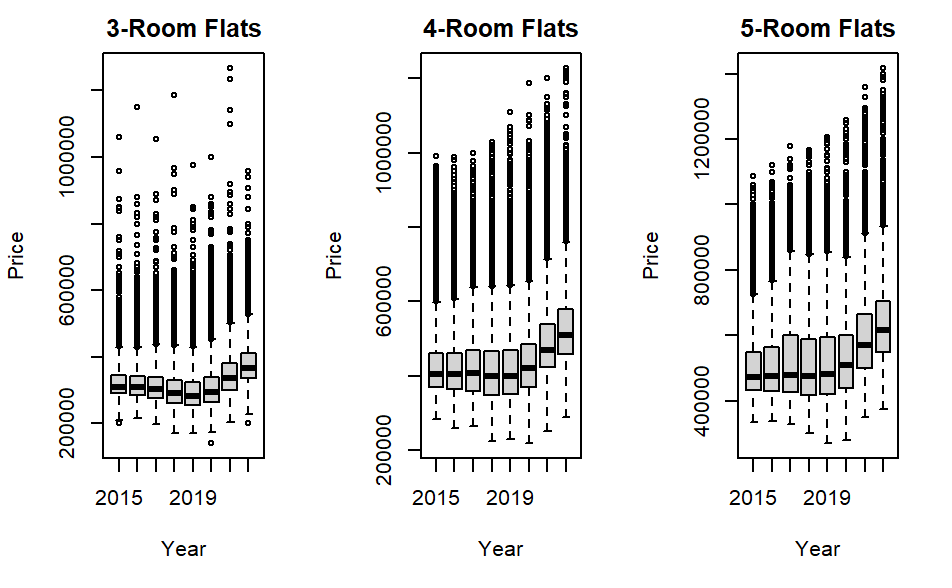
 

Figure 3: Box Plot of Price against Flat Type Figure 4: Box Plots of Price against the Year split by Flat Type

The data is further segregated by different years as shown on the right diagram above. Despite the segregation, outliers persist in the box plots, and there is an upward trend in the median values from 2020 to 2022.

The figure below depicts the top and bottom 5 areas’ change in price from 2015 to 2022. From here we observed that mature areas such as Bukit Merah and Bishan are selling at much higher prices as compared to non-mature areas such as Yishun and Woodlands. Furthermore, it is interesting to note that non-mature areas see a spike in prices between 2020 to 2022.

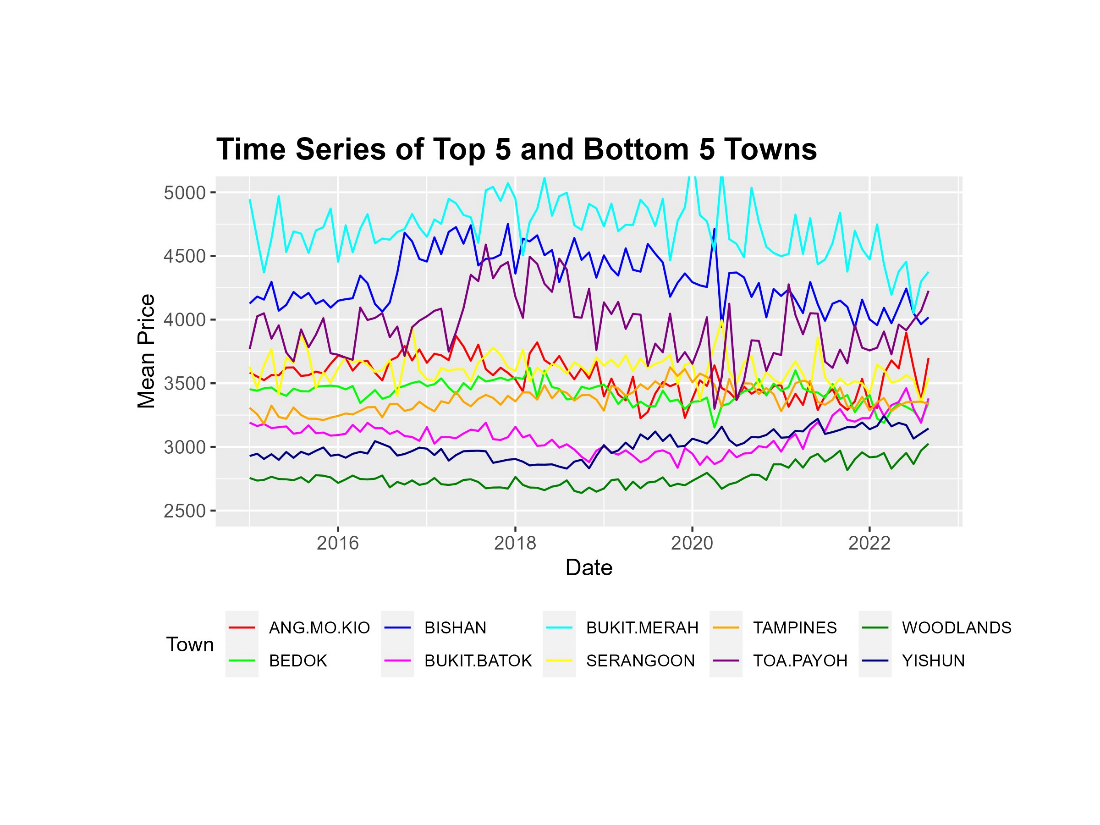


Figure 5: Line chart depicting the Price per Square Meters against Years

**Geospatial Visualization**

A map of singapore with different colored areas

Description automatically generated

Figure 6: Choropleth Map of Housing Price

Based on the geospatial visualization diagram above, it is evident that the central and southern regions of Singapore (Mature estates) exhibit higher mean housing prices per square meter and conversely the western and northern parts of Singapore (Non-mature estates) tend to have lower mean prices, further reinforcing our statement above. Hence, neighborhoods in these regions could be good candidates for subsequent modeling analysis.

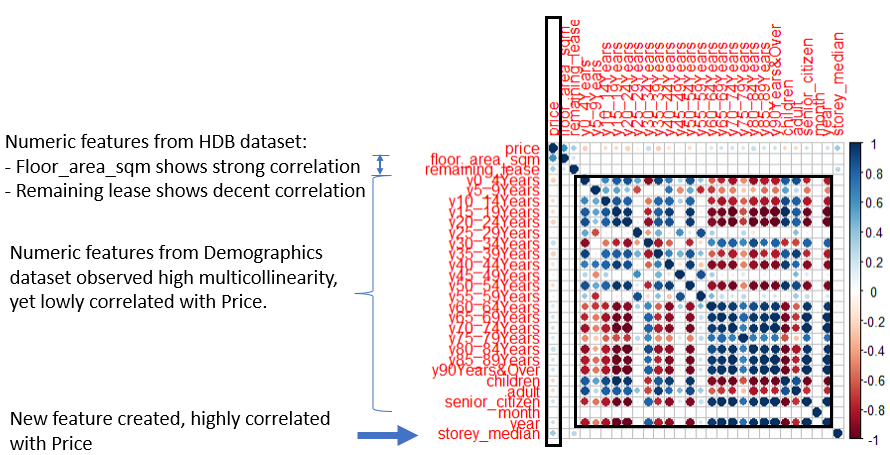
**Correlation between Response and Predictor Variables**

A graph with numbers and lines

Description automatically generated

Figure 7: Boxplot of Price by Storey Range

The box-plot analysis of the "storey-range" variable reveals a strong correlation with price. Hence, a new feature, “storey\_median” was created using the median value of each storey range.

Figure 8: Correlation Matrix of numeric features against Price

From the correlation matrix, “floor\_area\_sqm” shows very strong correlation with the dependent variable, followed by the "remaining lease" and “storey\_median” variable. On the other hand, numeric features from the Demographics dataset exhibit low correlations with the dependent variable, despite the presence of high multicollinearity.

**Feature Selection and Preprocessing**

Features were selected based on previous exploratory data analysis and correlation with the dependent variable. The features that exhibited the highest positive or negative correlations were chosen. Here are the nineteen selected features:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Town Features | | Flat Model | Flat related Numeric | Demographic |
| Bukit Merah | Toa Payoh | DBSS | Storey median | 30\_34 Years |
| Queenstown | Clementi | Type 3 room | Remaining lease | 80\_84 Years |
| Central Area | Kallang/Whampoa | Model Type S1 | Floor area sqm |  |
| Woodlands | Chua Chu Kang | Model Type S2 | Month |  |
| Bishan |  |  |  |  |

Table 2: Selected features for modeling

To mitigate multicollinearity, only two demographic features were selected. Principal Component Analysis (PCA) was attempted but the modeling results were not satisfactory. Based on the principal component regression, the RMSEP does not really improve after 25 components hence using 25 components, R2 value of 0.64 This could be attributed to the fact that most of the demographic features did not exhibit significant correlations with the dependent variable. The dataset consisting of the selected features was then randomly split into training, validation, and test sets in a ratio of 70:15:15, respectively.

**Modeling**

A baseline model was created using the mean value of the training set's housing price. This baseline serves as a reference for comparing the performance of other models. A negative R-squared indicates that the mean value performs poorly when predicting the housing prices on the validation dataset as seen in table 3 below.

Next, three models were developed. The first model is a multiple linear regression that incorporates all the features identified in the earlier feature selection. The second model is also a multiple linear regression, but the dependent variable was log-transformed to normalize the dataset. This transformation is expected to help address skewness and improve the linearity assumption of the regression model. Lastly, a decision tree model was built as an alternative to the linear regression models, to provide a different approach to modeling and allows for a comparison with the linear regression models.

By developing these three models, a more complete evaluation of the data and different modeling techniques can be achieved.

**Modeling Evaluation**

Root Mean Square Error (RMSE) and Coefficient of Determination (R2) were used to evaluate the models. RMSE quantifies the average magnitude of prediction errors while R2 measures the proportion of variance explained by the model.

Out of the 3 models evaluated, Linear-Linear Multiple Regression model performs the best, with consistent R2 value for training, validation, and test dataset. This implies minimum overfitting. Root Mean Square Error (RMSE) value is also the lowest among all the models. None of the features exhibited a Variance Inflation Factor (VIF) greater than 5, indicating the absence of multicollinearity among the explanatory variables. Additionally, all variables demonstrated p-values less than 0.001, signifying statistical significance at a 0.1% significance level.

See below for the model evaluation summary table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model**    **Metric for**  **each dataset** | **Base Model** | **Model 1: Multiple Linear Regression**  **(Linear-Linear)** | **Model 2: Multiple Linear Regression**  **(Log-Linear)** | **Model 3: Decision Tree** |
| Prediction is based on mean value of prices from the training dataset | All independent variables identified from earlier feature selection section. | Log-linear of Model 1 | minsplit = 2000, maxdepth = 7,  cp of 0.01.  Based on hyperparameter tuning |
| Training R2 | - | 0.7701 | 0.7991 | 0.5657 |
| Training Adj R2 | - | 0.7701 | 0.7990 | - |
| Training RMSE | - | 76,815 | 495,022 | 105,582 |
| Validation R2 | -1.538373e-05 | 0.7732 | 0.7615 | 0.5637 |
| Validation RMSE | 160,453 | 76,412 | 494,503 | 105,979 |
| Test R2 |  | 0.7686 |  |  |
| Test RMSE | 76,941 |  |  |

Table 3: Evaluation of all three models using R2 and RMSE

The performance of the Log-Linear Multiple Linear Regression model was unexpected, as it exhibited a higher RMSE despite having somewhat similar R2 values compared to the linear-linear model. This indicates that the log-linear model may not effectively capture the true relationship between the predictors and the target variable. One possible reason could be interpretation differences. RMSE reflects the prediction errors on the original scale of the target variable. If the log-linear model fails to accurately capture the relationships, it can lead to larger errors when back-transforming the predictions to the original scale.

A collage of graphs and diagrams

Description automatically generated A collage of graphs and charts

Description automatically generated

Figure 9: Linear-Linear Multiple Regression Figure 10: Log-Linear Multiple Regression

Above is the diagnostics plots comparison between linear-linear and log-linear model. It is visible that log-linear has more normalized Q-Q plot and evenly distributed residual vs fitted values, despite much higher RMSE value.

Decision Tree model on the other hand has limited ability to capture complex relationships between variables. It partitions the data based on binary splits at each node, which may not effectively capture intricate patterns in housing price.

**Modeling Evaluation – Final Model (Model 1: Linear-Linear Multiple Linear Regression)**

Further examination predicted y against true y is generally linear with a 1:1 relationship, except for a few outliers as seen in Figure 11 below.

A graph with blue dots

Description automatically generated

Figure 11: Prediction distribution plot for Multiple Linear Regression Model

A graph with a bar chart

Description automatically generated

Figure 12: Top predictors based on Regression Coefficients

The top predictors, including features such as type of flat model (S2 and S1), the location of Queenstown, DBSS flat model, Bishan, central area, Kallang and Whampoa town, Clementi town, and Toa Payoh town, are strongly associated with higher housing prices. Conversely, homes located in Woodlands and Choa Chu Kang are found to have negative influence on housing prices.

This aligns with the findings from the exploratory data analysis, suggesting that homes with these features tend to fetch higher prices. Notably, proximity to central areas has a significant impact on housing prices.

|  |  |
| --- | --- |
| Features | Coefficients |
| Flat model Type S2 | 295162 |
| Flat model Type S1 | 224029 |
| Queenstown | 187264 |
| Flat Model DBSS | 180079 |
| Bukit Merah | 174072 |
| Bishan | 162372 |
| Population of 30-34 years old | 2.613 |
| Population of 80-84 years old | 0.839 |

Table 4: Coefficients for each feature

Coefficients in the above table help to quantify the impact of each variable. Example, a S2 flat type will increase housing price by $295,162. S1 flat type will increase housing price by $224,029. Homes located in Queenstown shall increase housing price by $187,264. Likewise, one population count increase in age group 30-34 years old, will just increase home price by $2.61.

**Conclusion**

Demographic variables, such as the population of individuals aged 30-34 and 80-84, show statistically significant correlations with housing prices, but their actual impact on housing prices is minimal, based on their regression coefficients. One unit increase in the age group 30-34 years old will increase home prices by $2.61, while one unit increase in the age group 80-84 years old will increase home prices by $0.84. An aging population does not result in higher public housing prices.

This could be due to most Singapore citizens purchasing their home upon marriage rather than at old age. Seniors may face restricted housing options mandated by the government, making it less financially sensible for them to purchase homes at an older age [7]. Newlywed couples will tend to buy more affordable Build-To-Order (BTO) homes [8] that are heavily subsidized by the government, rather than from the resale market.

As such, our hypothesis that suggests that changes in age group population increases the demand in public housing is invalid.

**Recommendations**

Our recommendation to the government and policymakers is to focus on building more flats in towns such as Queenstown, Bukit Merah, Toa Payoh, Clementi, Kallang, and Bishan. They are more desirable based on the incremental price derived from our predictive analysis, attributed to their centralized and convenient location. By developing more housing options in these towns, we can address the demand for homes in popular areas, promote affordability, and contribute to a more stable housing market.

Another recommendation for the government is to place priority on constructing residential homes featuring flat models of types S2, S1, and DBSS in Woodlands and Choa Chu Kang. Currently, these towns are experiencing high negative coefficients, indicating lower demand compared to other estates. To rectify this situation, it is essential to introduce flat types with notably positive coefficients, which can potentially stimulate increased demand for housing in these less desirable towns. This approach could create a more equitable housing landscape and alleviate any over-demand in towns with high positive coefficients.

As government and policymakers utilize our model, we recommend periodically updating the data used in the model to incorporate more recent information. This will help ensure the accuracy and relevance of the model's predictions and capture any changes in the impact of each variable over time.

On the other hand, for investors and property developers, it is recommended to consider homes in Choa Chu Kang and Woodlands, if they are targeting low-middle-income buyers since the price is much lower compared to other areas.

# **Appendix**

# **References**

[1] “HDB | Public Housing – A Singapore Icon - Housing & Development Board,” [*https://www.hdb.gov.sg/about-us/our-role/public-housing-a-singapore-icon*](https://www.hdb.gov.sg/about-us/our-role/public-housing-a-singapore-icon) (accessed June 1, 2023).

[2] “Commentary: Supply-demand imbalances in HDB market will need time to clear,” [*https://www.channelnewsasia.com/commentary/hdb-public-housing-bto-affordability-accessibility-supply-demand-3259186*](https://www.channelnewsasia.com/commentary/hdb-public-housing-bto-affordability-accessibility-supply-demand-3259186)(accessed June 15, 2023)

[3] “The Unassuming Economist,” *The Unassuming Economist A Look at Housing Affordability in Asia Comments* [*https://unassumingeconomist.com/2019/12/a-look-at-housing-affordability-in-asia/*](https://unassumingeconomist.com/2019/12/a-look-at-housing-affordability-in-asia/) (accessed June 16, 2023)

[4] “Is Public Housing Still Affordable For The Average Singaporean Couple?” [*https://blog.seedly.sg/is-public-housing-still-affordable-for-the-average-singaporean-couple/*](https://blog.seedly.sg/is-public-housing-still-affordable-for-the-average-singaporean-couple/) (accessed June 17, 2023)

[5] “Resale Flat Prices” *Data.gov.sg.* [*https://data.gov.sg/dataset/resale-flat-prices*](https://data.gov.sg/dataset/resale-flat-prices) (accessed June 12, 2023)

[6] “Population and Population Structure,” *Base.* [*https://www.singstat.gov.sg/find-data/search-by-theme/population/population-and-population-structure/latest-data*](https://www.singstat.gov.sg/find-data/search-by-theme/population/population-and-population-structure/latest-data) (accessed June 18, 2023)

[7] “A Seniors’s guide to Buying an HDB Flat,” *Mynicehome by HDB.* <https://www.mynicehome.gov.sg/hdb-how-to/buy-your-flat/a-seniors-guide-to-buying-an-hdb-flat/> (accessed July 19, 2023)

# [8] “Demand for BTOs doubles from 2018 to 2020 as more get married, choose to have their own flat: HDB,“ *Today*. <https://www.todayonline.com/singapore/demand-btos-doubles-2018-2020-more-get-married-choose-have-their-own-flat-hdb> (accessed July 19, 2023)