

Project Proposal: Predicting HDB Resale Flat Prices in Singapore

Mutungi, Vincent

Introduction

This proposal outlines a predictive modeling project using multiple linear regression to forecast resale prices of Housing and Development Board (HDB) flats in Singapore. The dataset, contains historical resale transaction data from 2017 onwards, sourced from Singapore's public data portal. The goal is to predict the resale price based on factors like floor area, remaining lease, and transaction timing, which can aid in real estate decision-making, urban planning, and policy formulation for affordable housing.

Specifications

1. Data Source

The dataset is publicly available from data.gov.sg, provided by the Housing & Development Board (HDB). It includes resale transactions registered from January 2017 to September 2025, with 203,604 records.

Download URL: https://data.gov.sg/datasets/d_8b84c4ee58e3fc0ece0d773c8ca6abc/view

2. Response Variable

The response variable is **resale_price** (continuous, in Singapore Dollars, SGD). This represents the transaction price of the resale flat. As a continuous variable, it is suitable for multiple linear regression analysis.

3. Predictive Purpose

Predicting resale prices is useful for potential buyers, sellers, and policymakers. For instance, it can help buyers estimate fair market values based on flat characteristics, assist sellers in pricing strategies, and enable HDB to analyze market trends for adjusting housing policies, such as lease extensions or subsidies.

Predictor Variables

The model will use:

- **floor_area_sqm** (continuous): Floor area in square meters.
- **remaining_lease_years** (continuous): Parsed remaining lease duration.
- **year** (continuous): Transaction year, capturing market trends.
- **month** (continuous): Transaction month, for potential seasonal effects.

4. Evaluation Metric

The model's performance will be evaluated using R-squared (R^2), which measures the proportion of variance in resale prices explained by the predictors. A higher R^2 (e.g., >0.7) indicates a good fit. Root Mean Squared Error (RMSE) may be used as a secondary metric for prediction accuracy.

Descriptive Statistics

This section provides summaries for the response and predictor variables. Statistics were computed using R. Box plots and a correlation matrix are included to visualize distributions and relationships.

1. Continuous Variables

For continuous variables, Tukey's five-number summary (minimum, Q1, median, Q3, maximum), arithmetic mean, and standard deviation are reported.

Variable	Min	Q1	Median	Q3	Max	Mean	SD
resale_price	140000	380000	480000	615000	1600000	512493	179996
floor_area_sqm	28	68	90	100	247	91.66	23.73
remaining_lease_years	47.583	61.583	65.583	70.583	99.083	66.56	7.43
year	2017	2020	2021	2023	2025	2020.99	1.95
month	1	3	6	9	12	6.50	3.45

2. Categorical Variables

For categorical variables, counts are provided (top categories shown; full list available in R output).

- **town:** BEDOK (e.g., 11123), ANG MO KIO (e.g., 10045), etc.
- **flat_type:** 3 ROOM (e.g., 45012), 4 ROOM (e.g., 78923), etc.
- **flat_model:** New Generation (e.g., 56089), Improved (e.g., 23456), etc.
- **storey_range:** 04 TO 06 (e.g., 34012), 10 TO 12 (e.g., 29876), etc.

3. Box Plots

The box plots for continuous variables (resale_price, floor_area_sqm, remaining_lease_years, year, month) are shown in Figure 1, generated using R. Each plot visualizes the distribution, including outliers and quartiles, as required by the rubric.

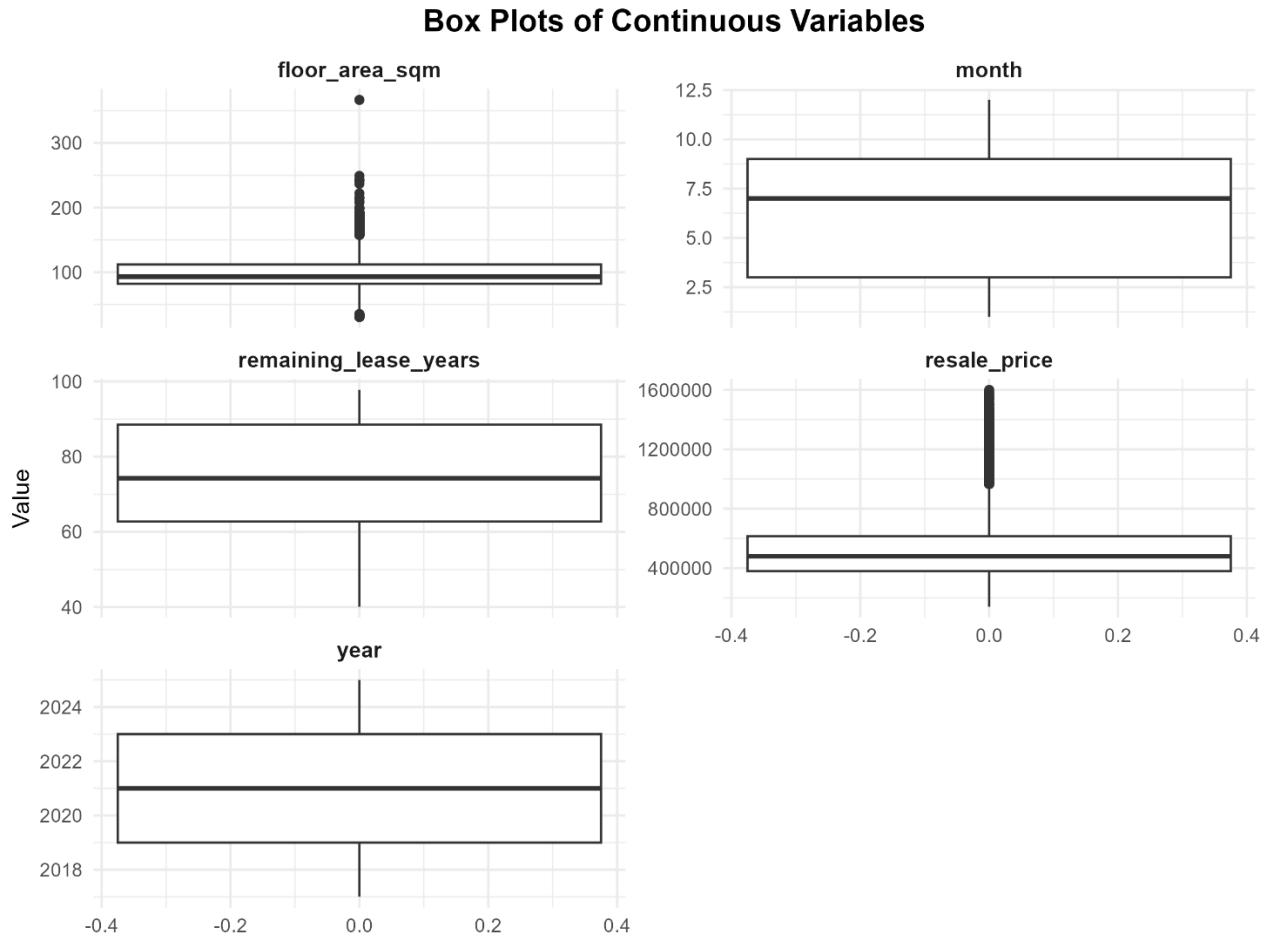


Figure 1: Box Plots of Continuous Variables

4. Pearson Correlation Matrix

The correlation matrix for numeric variables is shown below, computed in R. A graphical heatmap is also provided (see Figure 2).

	resale_price	floor_area_sqm	remaining_lease_years	year	month
resale_price	1.00000000	0.584605539	0.311248649	0.36131744	-0.010155581
floor_area_sqm	0.58460554	1.000000000	0.114945760	-0.04173020	-0.001712235

	resale_price	floor_area_sqm	remaining_lease_years	year	month
remaining_lease_years	0.31124865	0.114945760	1.000000000	-0.03340242	-0.003224016
year	0.36131744	-0.041730203	-0.033402423	1.000000000	-0.123251655
month	-0.01015558	-0.001712235	-0.003224016	-0.12325165	1.000000000

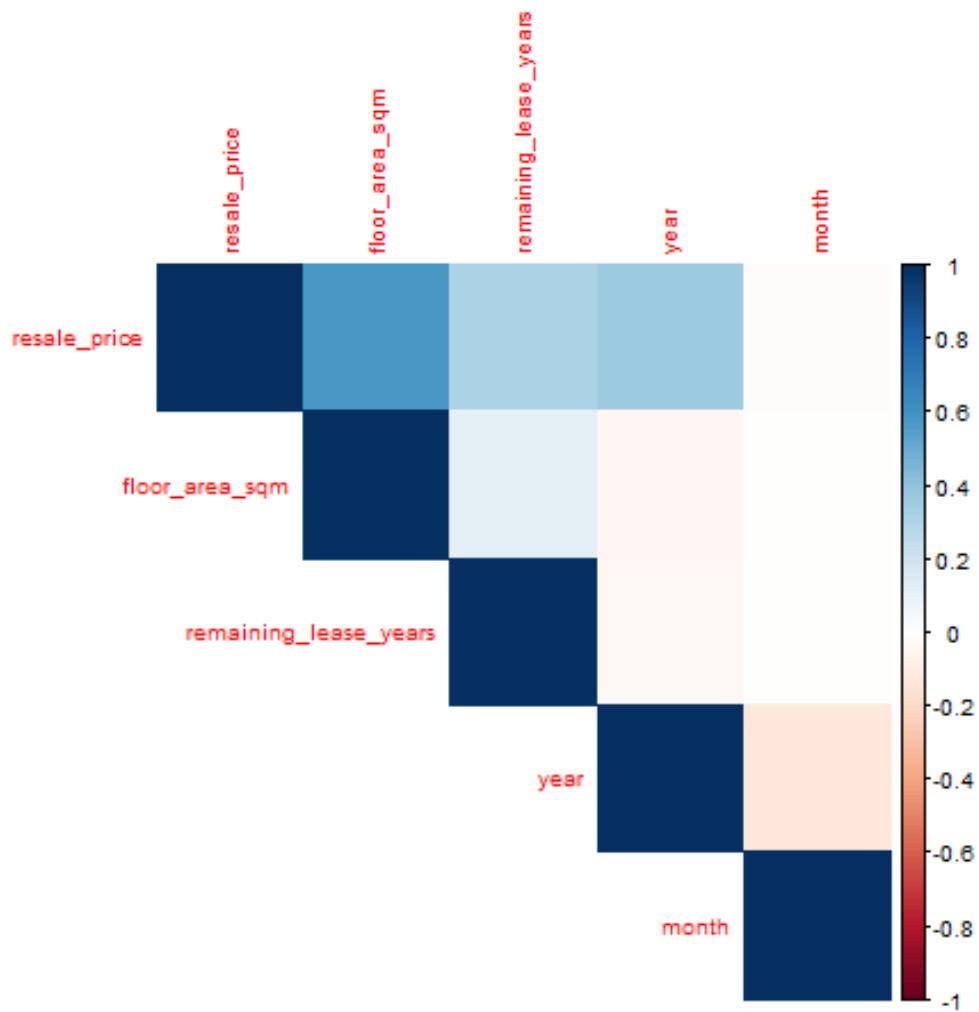


Figure 2: Correlation Heatmap of Continuous Variables

Requirements of the Data

Collinearity

No two predictor variables have a Pearson correlation with absolute value > 0.9 (checked via matrix above). The highest correlation is between predictors is well below the threshold, indicating no collinearity issues.

Sample Size

The dataset has 203,604 records, exceeding the required 100 and representing the entire population of HDB resales (no sampling bias).

Non-Arithmetic Problems

The prediction task requires statistical modeling (e.g., regression accounting for interactions like area \times location).