

Understanding Singapore's HDB Market through Data Analytics

Part 1: Identify Quantitative Variables

The dataset includes the following columns:

month, town, flat_type, block, street_name, storey_range, floor_area_sqm, flat_model, lease_commence_date, remaining_lease, and resale_price.

These can be grouped as follows:

- Quantitative (ready to use): floor_area_sqm, lease_commence_date, resale_price
- Quantitative (need transformation): storey_range, remaining_lease
- Categorical: month, town, flat_type, block, street_name, flat_model

Variables like storey_range and remaining_lease contain numeric information but must be transformed (e.g., extracting lower floor level or converting remaining years to numeric) before analysis.

Part 2: Distribution of Floor Area

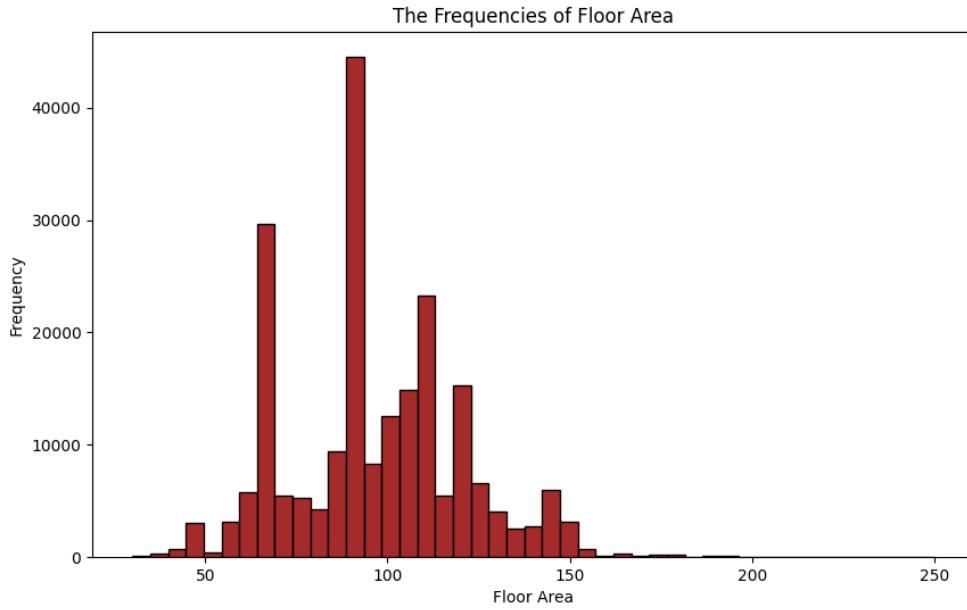


Figure 1:The Frequency of Floor Area

A histogram with intervals of 5 from 30 to 250 shows that most flats have areas between 50 and 150 sqm, with clear peaks around 65–70 sqm and 90–95 sqm. The distribution is slightly right-skewed, indicating

that large flats above 150 sqm are relatively rare.

These peaks likely correspond to standard floorplan designs commonly used in HDB estates.

Part 3: Frequency of Flat Types

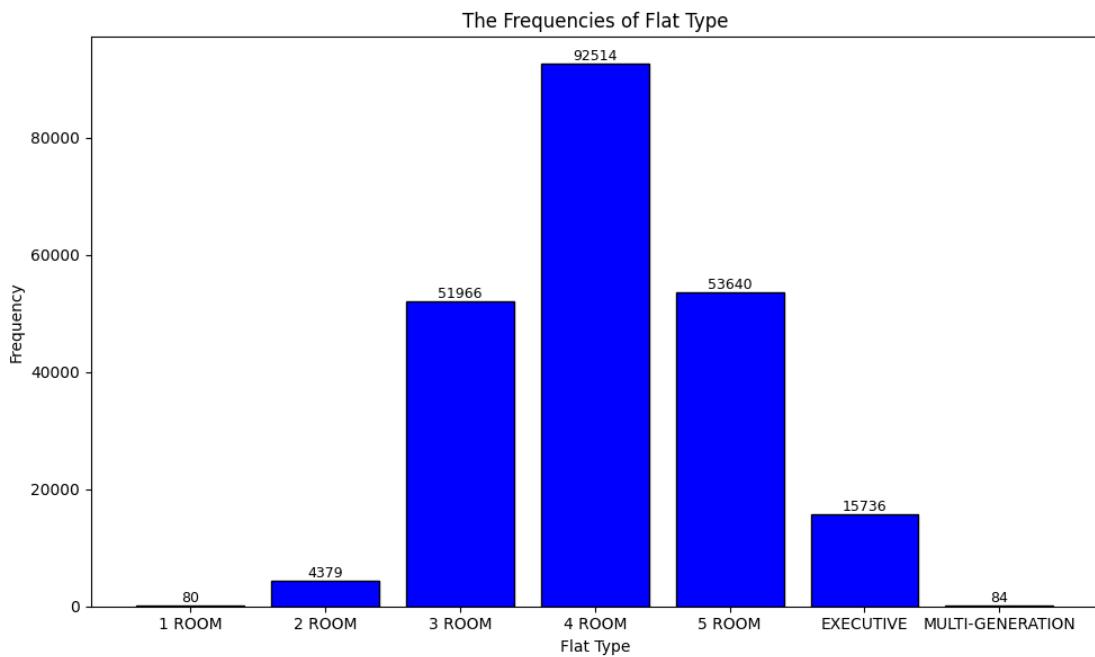


Figure 2: The Frequency of Flat Type

The bar chart shows that “4 ROOM” flats are the most common (over 40%), followed by “3 ROOM” and “5 ROOM” types. Only a small number of “2 ROOM” and “MULTI-GENERATION” flats exist in the sample, which may limit statistical reliability for those categories.

This pattern reflects the general market preference for medium-sized family units in Singapore.

Part 4: Comparison of Floor Area by Flat Type

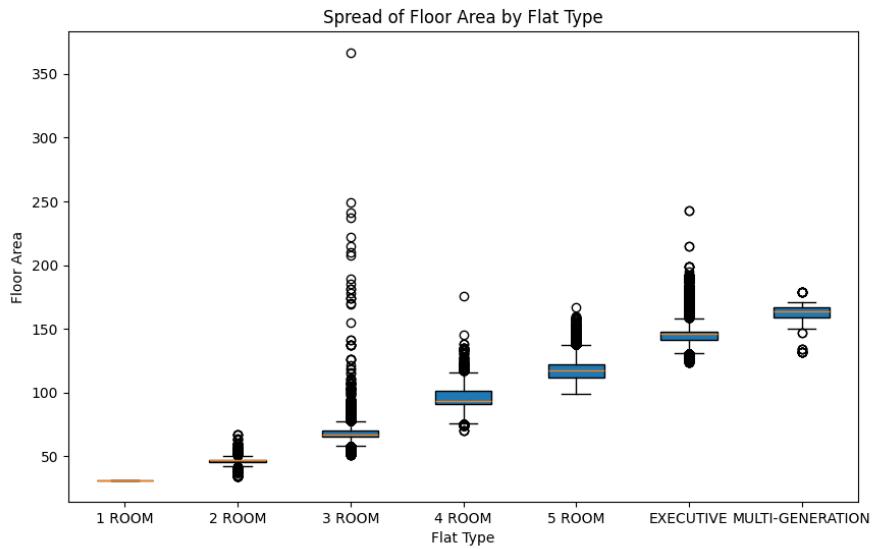


Figure 3: Spread of Floor Area by Flat Type

The boxplot clearly shows increasing median and range with larger flat types:

2-room < 3-room < 4-room < 5-room < Executive.

The interquartile range (IQR) also widens with size, meaning larger flats vary more in area.

The “1 ROOM” and “MULTI-GENERATION” type have too few samples for meaningful comparison.

Part 5: Comparison of Floor Area by Flat Type

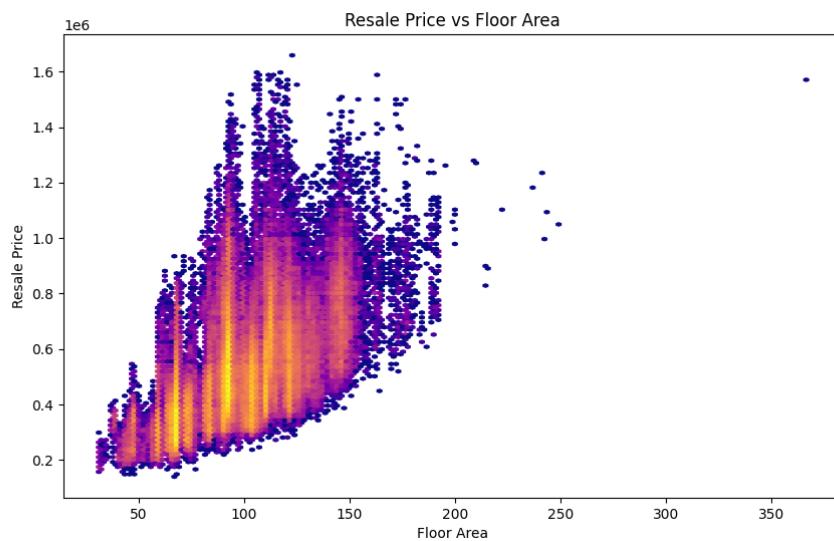


Figure 4: Resale Price vs Floor Area

A hexbin plot of `floor_area_sqm` against `resale_price` shows a clear upward trend — larger flats tend to be more expensive.

The linear correlation coefficient is 0.5742, suggesting a moderate positive linear relationship.

The fitted simple linear regression equation is:

$$\text{resale_price} = 4429.16 \times \text{floor_area_sqm} + 93508.55$$

Although the overall trend is clear, the data density reveals multiple vertical clusters, implying that other factors such as town, flat type, or remaining lease also influence resale prices.

Part 6: Linear Regression Model

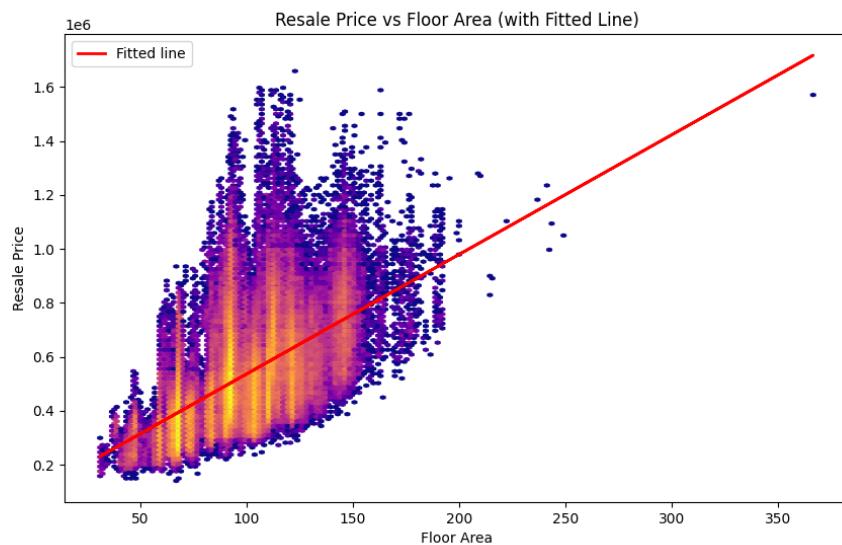


Figure 5: Fitted Line

To quantify the relationship, a simple linear regression model was fitted using `floor_area_sqm` as the explanatory variable and `resale_price` as the response variable.

- MSE: 23013642553.60631
- MAE: 113646.11777501665
- R²: 0.32973184088928553

The low R² indicates that the model explains only about 33% of the total variation in resale prices. While it confirms a positive relationship between size and price, it also shows that floor area alone is insufficient to predict resale prices accurately.

Residual plots (not shown) would likely reveal heteroscedasticity and unaccounted variance due to omitted factors such as location and lease balance.

In summary, the simple linear regression model offers a basic understanding but limited predictive accuracy.

Part 7: Extension – Using a Machine Learning Model (LightGBM)

To further explore model performance, a LightGBM regression model was trained. The dataset was divided into 80% training and 20% testing sets to properly evaluate prediction ability.

LightGBM can model non-linear and interaction effects without explicit transformations.

After tuning and early stopping at 748 iterations, the test-set performance was:

- MSE: 6752249856.297144
- MAE: 66570.30674334243
- R²: 0.800950740975843

Compared with the simple linear model ($R^2 = 0.33$), the LightGBM model captures much more of the variation in resale prices and achieves far lower prediction error.

This improvement indicates that the true relationship between area and price is non-linear, and that complex interactions exist in the dataset.

Overall, the LightGBM experiment demonstrates the advantage of modern machine-learning models for predictive accuracy, while the linear regression approach remains valuable for statistical interpretation and understanding variable influence.

Conclusion

Through descriptive statistics, visualization, and modeling, we found that:

- Flat type strongly influences floor area distribution.
- Floor area has a moderate linear correlation with resale price.
- A simple linear model gives limited explanatory power, while a more flexible LightGBM model achieves strong predictive performance ($R^2 \approx 0.80$).

These findings suggest that incorporating multiple factors and non-linear modeling techniques can substantially improve resale-price prediction in Singapore's HDB market.