Executive Summary (Done by Nigel Wee)

1. Objective

To develop a predictive algorithm to determine the factors affecting resale prices of residential properties in Singapore. By understanding these factors, we aim to provide insights into strategies for curbing price inflation and improving housing affordability.

2. Dataset

Our dataset consists of 5 csv files, containing a total of 826,581 samples. The data spans from 1990 to 2020.

Key features include:

- Resale date (year and month)
- Flat type (e.g., 3 room, 4 room, executive)
- Location (town, block, street name)
- Storey range (floor level classification)
- Floor area (sqm) (size of the flat)
- Flat model (e.g., standard, DBSS, maisonette)
- Lease commencement date

Target variable: Resale price

Given the nature of the dataset, our features consist of:

- Continuous variables (e.g., floor area, remaining lease)
- Discrete variables (e.g., storey range, resale month)
- Categorical variables (e.g., flat type, location, flat model)

3. Methodology

Data preprocessing:

- Handled missing values and inconsistencies (e.g., standardizing categorical values like "MULTI GENERATION" vs. "MULTI-GENERATION").
- Created new features:
  - Extracted resale year and resale month from the resale date.
  - Computed remaining lease years (lease_commencement_date + 99 – resale_year)
  - Applied one-hot encoding to categorical variables (town, flat type, flat model).
- Removed redundant features: Dropped block and street name (as they are relative to town).

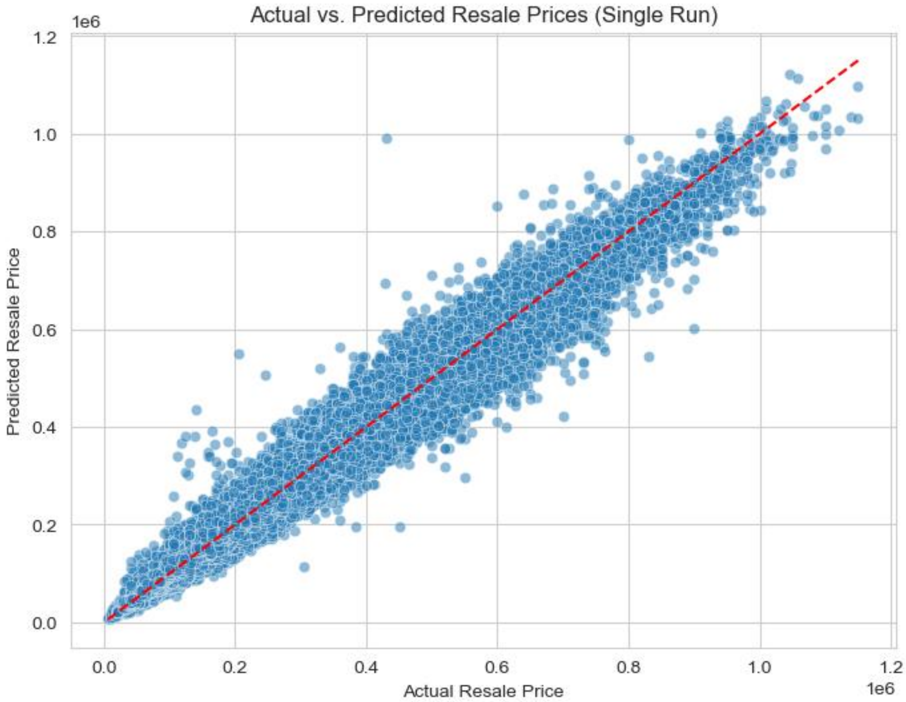
Model selection and evaluation:

- Trained and evaluated four machine learning models:
  - Random Forest
  - XGBoost
  - LightGBM
  - CatBoost
- Tree-based models were chosen because they can handle a mix of continuous, discrete and categorical features effectively.
- Used 5-fold cross-validation to ensure robustness.
- Tuned Random Forest hyperparameters using GridSearchCV (3-fold validation).
- Evaluated models using RMSE, MAPE, and R² Score.

4. Results and findings

Random Forest with tuned hyperparameters (5-fold validation):

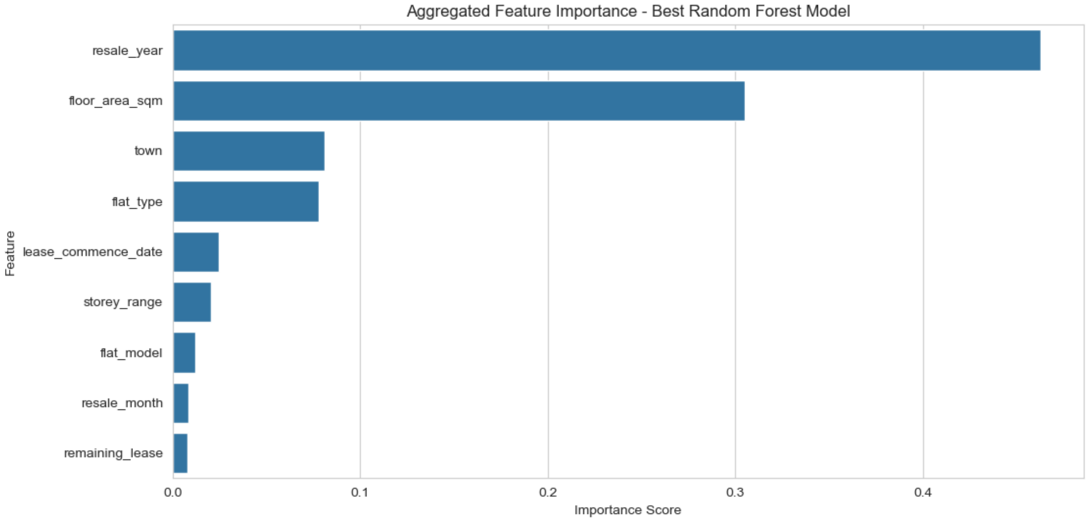
Metric	Mean	Standard Deviation
RMSE	22443	99.4
MAPE	5.89%	0.000153
R²	0.977	0.000246



Feature Importance Analysis showed that the top 3 features were:

1. Resale year, highlighting the impact of resale price inflation over time.
2. Floor area, with larger flats generally commanding higher resale values.

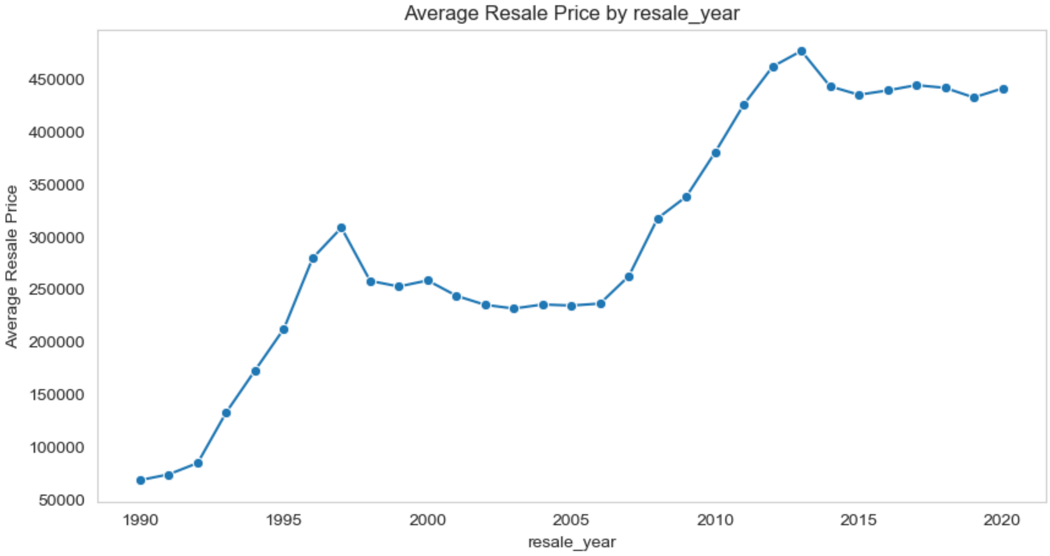
3. Town, indicating that location strongly influences property prices due to demand and accessibility.



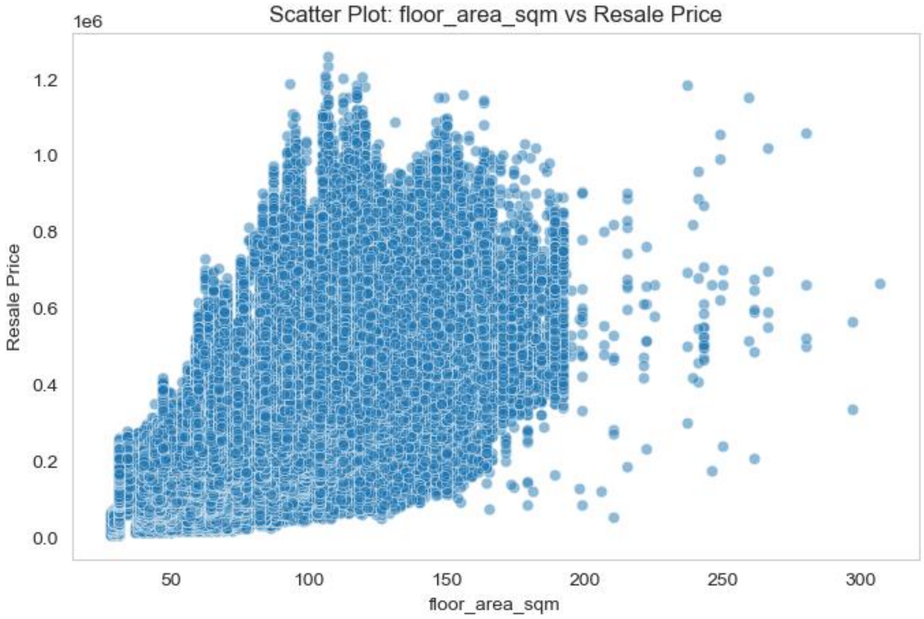
5. Key insights and recommendations

Insights

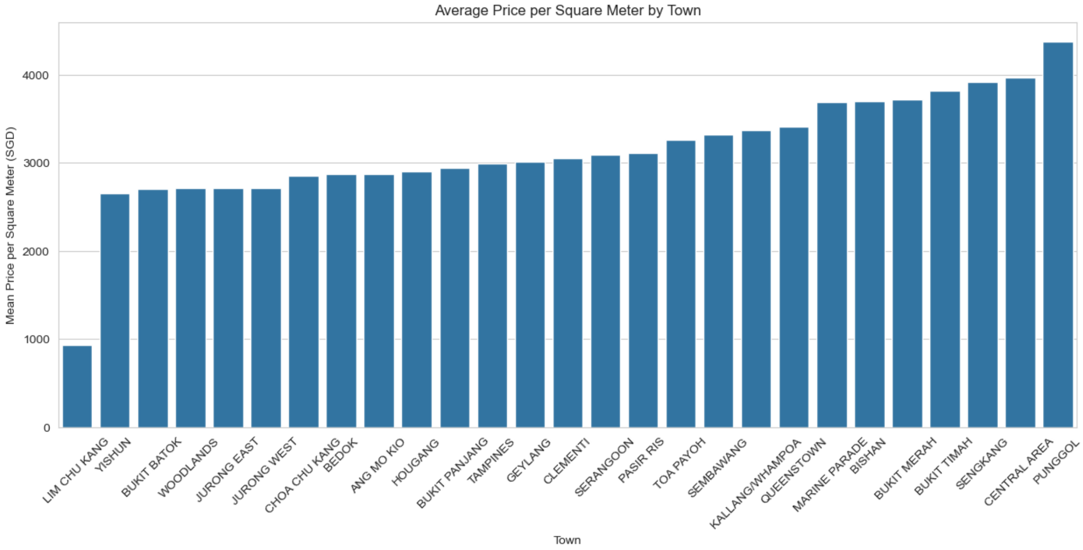
- Resale year – Resale prices have generally increased over time, likely due to economic inflation (assuming the dataset's resale prices are not inflation-adjusted). However, a more significant factor is the growing demand for housing outpacing supply.
  Notably, there was a sharp price surge between 2006 and 2007. This could be attributed to the discontinuation of the Walk-in Selection (WIS) scheme, which previously allowed buyers to purchase flats on a first-come, first-served basis [1]. Additionally, the oversupply of 17,500 new flats in 2001 may have temporarily suppressed prices, leading to a rebound in subsequent years.



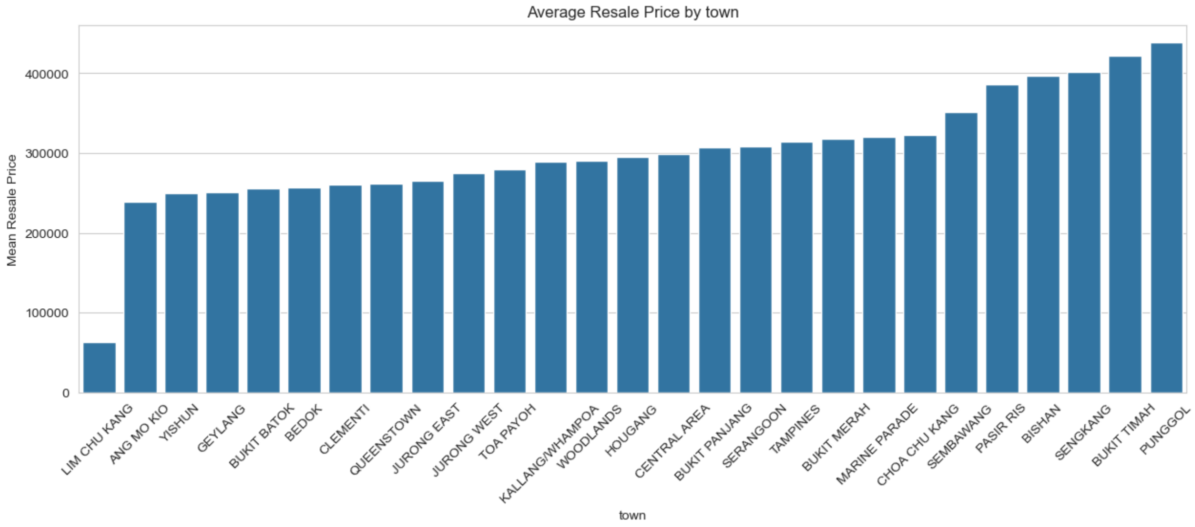
- Floor area - Floor area is a key determinant of resale prices, with larger flats generally commanding higher prices. This trend is expected as larger living spaces offer greater functionality and comfort, making them more desirable to buyers.



- Town - Flats in mature and centrally located towns such as Queenstown, Bukit Merah, and the Central Area generally have a higher price per square meter, driven by better amenities, transport connectivity, and proximity to the city center. Conversely, non-mature towns like Bukit Batok, Woodlands, and Lim Chu Kang tend to have lower resale prices, reflecting longer commuting distances and fewer amenities.



Interestingly, while Queenstown and the Central Area command some of the highest prices per square meter, their average resale prices remain relatively lower. This suggests that flats in these towns are generally smaller in size, leading to higher price density despite a lower absolute resale price.



Recommendations

- Increase Housing Supply in High-Demand Areas
    - The government could increase BTO launches in mature estates with high demand, such as Queenstown and Bukit Merah, to moderate resale prices.
    - Releasing more land for public housing in high-demand areas could reduce the premium attached to these locations.
    - Encourage en-bloc redevelopment of older flats to create higher-density housing while maintaining affordability.
- Developing Non-Mature Estates to Improve Affordability and Reduce Price Inflation (Assumptions are to be made for the reasons, surveys will need to be done in future to get the general public's opinion for why there is no demand for such estates.)
    - Expanding MRT lines, express bus services, and integrated transport hubs can significantly reduce commuting times, making non-mature estates more accessible and attractive.
    - Developing business parks, commercial districts, and co-working spaces in non-mature estates can create job opportunities closer to residential areas, reducing the need for long commutes.
    - Increasing the availability of shopping malls, healthcare facilities, and recreational spaces can enhance the overall quality of life, making non-mature estates more desirable for families and young professionals.
    - This would provide a better long-term effect of curbing housing prices inflation.

6. Factors and Considerations in Building an In-House Predictive Model for Users

- Data collection and preprocessing
    - Data sources need to be reliable, this could involve data from company databases, external databases, measurements (e.g., sensor readings).

  - Feature engineering to create meaningful features that are more representative of the target. This would require domain knowledge that could include people from different disciplines (e.g., healthcare/military projects may require someone with such domain knowledge).
  - Handling of missing data, how are these missing data going to be handled?
  - Data normalization for models that require it for better performance.
  - Encoding for categorical data (e.g., one-hot encoding).
  - Is regularisation needed for more generalisation and model robustness?
- Model selection and justification
  - Type of models would depend on our objectives, dataset, complexity and resources (time and cost) available.
  - If speed is a priority (e.g., real-time fraud detection), lightweight models like logistic regression or decision trees may be preferable.
  - Simple models (Linear Regression, Decision Trees) are easier to interpret but may not capture complex relationships.
  - Complex models (XGBoost, Deep Learning) provide better predictive accuracy but are harder to interpret.
- Hyperparameter tuning and model evaluation
  - Performance metrics: Model evaluation should be based on the problem type:
    - Regression Models: RMSE, MAPE, MAE, R^2 Score
    - Classification Models: Accuracy, Precision, Recall, F1 Score
  - Cross-validation: Using k-fold cross-validation ensures robustness and reduces the risk of overfitting.
  - GridSearchCV and RandomizedSearchCV help find optimal hyperparameters like learning rate, tree depth, and number of estimators.
  - Bayesian optimization can be used for more efficient tuning in high-dimensional spaces.
- Deployment and user accessibility
  - Local deployment: Suitable for sensitive data (e.g., healthcare, banking).
  - Cloud deployment (AWS, GCP, Azure): Ensures scalability and ease of access for multiple users.
  - Batch predictions: Suitable for periodic analysis (e.g., weekly sales forecasts).
  - Real-time API: Needed for instant predictions (e.g., fraud detection in banking transactions).
  - Provide interactive dashboards (Tableau, Streamlit, Flask apps) to visualize predictions. Will our users be non-technical
  - Will model retraining be done? How often should the retraining be done?
  - Will there be data drift over time? Will our old dataset be different from the newly evolved trends? How do we detect such drifts?

All in all, building an in-house predictive model requires careful consideration of key factors such as business objectives, data availability, and computational resources. Additionally, domain expertise plays a crucial role in ensuring the model is tailored to the organisation's specific needs, enabling more accurate insights and better alignment with strategic goals.

References

[1] F. MUHAMMAD and W. S. YING, Public housing choices in singapore - ISOCARP | case study platform, https://www.isocarp.net/Data/case_studies/315.pdf (accessed Feb. 14, 2025).