Executive Summary

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.

1. **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)

1. **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. 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.

1. **Results and findings**

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

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Mean RMSE** | **Std RMSE** | **Mean MAPE** | **Std MAPE** | **Mean R² Score** | **Std R² Score** |
| 22443 | 99.4 | 5.89% | 0.000153 | 0.977 | 0.000246 |

**A graph of a graph showing a line of blue dots

Description automatically generated with medium confidence**A graph with blue and white text

Description automatically generated with medium confidence

Feature Importance Analysis showed that the top 3 features were:

* 1. Resale year, highlighting the impact of 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.

1. **Key insights and recommendations**

* Resale year – resale prices have generally increased over time,