**PROJECT REPORT (Prepare by Koh Pee Sim)**  
*HDB Resale Flats Analysis and Price Prediction with Machine Learning*

**1. Introduction**

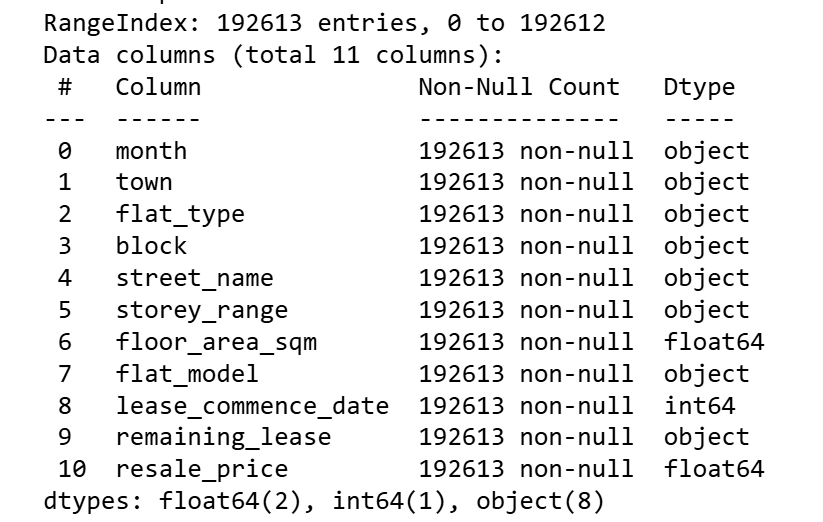
The Housing and Development Board (HDB) resale flat market in Singapore provides valuable data for understanding housing trends, price determinants, and demand patterns. This project aims to analyze HDB resale data using visualization techniques and machine learning to build a prediction model. Additionally, a user-friendly Tkinter-based tool was developed for data visualization and predictions.

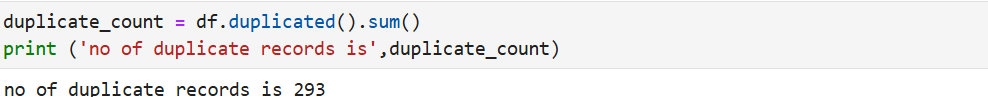
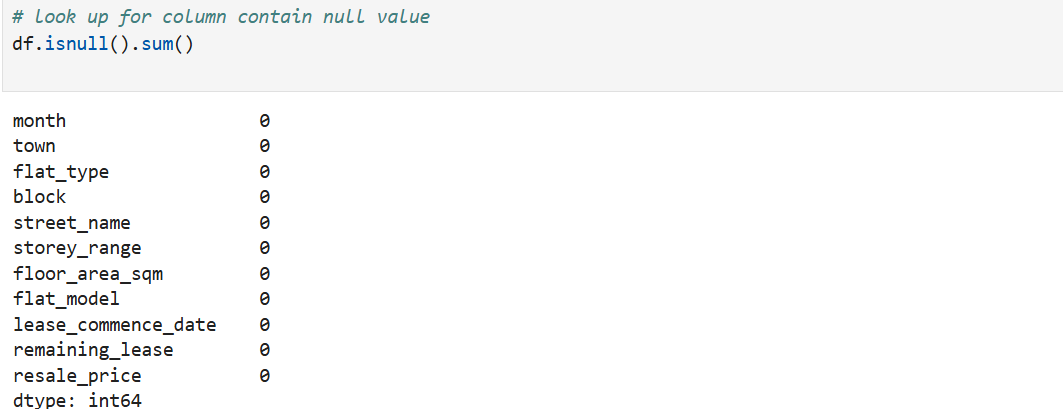
**2. Objectives**

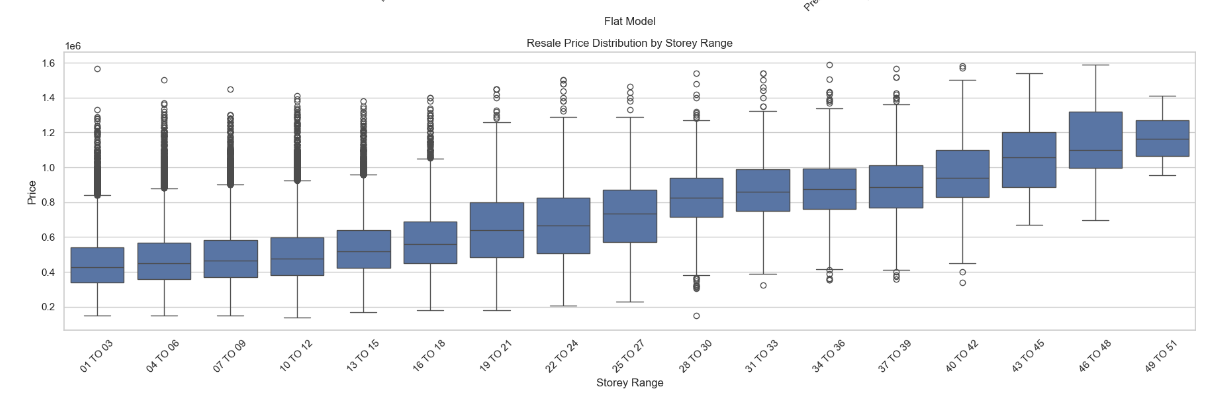
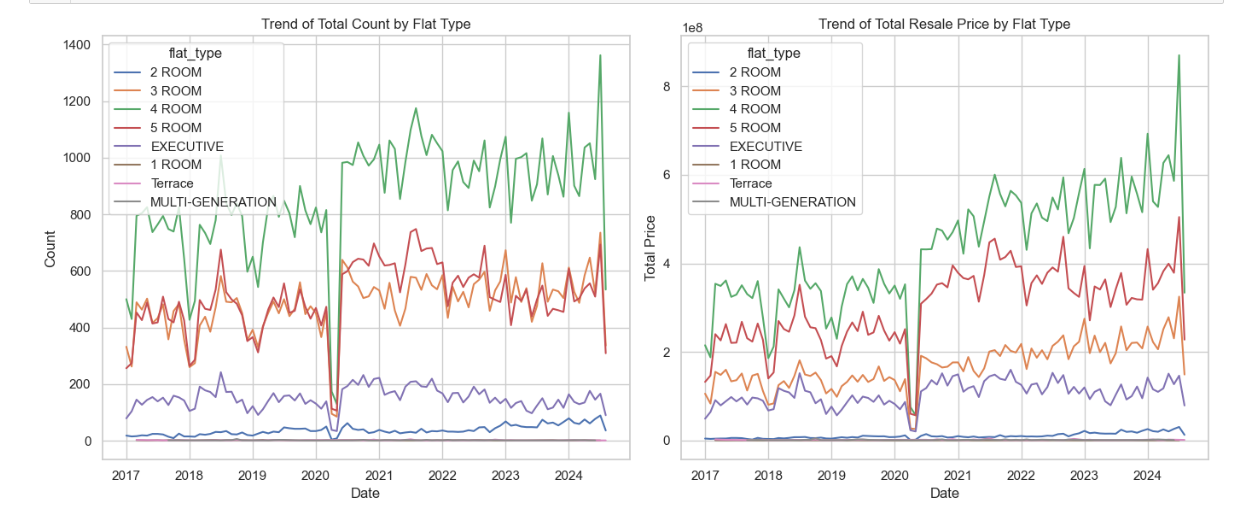
1. Perform data check and cleaning before data visualization.
2. Explore and visualize HDB resale data using various plot types.
3. Perform data preprocessing for machine learning readiness.
4. Train and compare three machine learning models: Linear Regression, Decision Tree Regressor, and Gradient Boosting Regressor.
5. Evaluate models using cross-validation and hyperparameter tuning.
6. Select the best model for prediction.
7. Implement a Tkinter-based application for data visualization and resale price prediction.
8. Create an interactive map visualization for one month of data with a toggle for flat types.

**3. Data Exploration and Preprocessing (For python codes please refer to KPSResaleflat.ipynb)**

**3.1 Data Exploration**

* Overview of dataset columns: .
* Summary statistics and null value and duplicate data analysis.

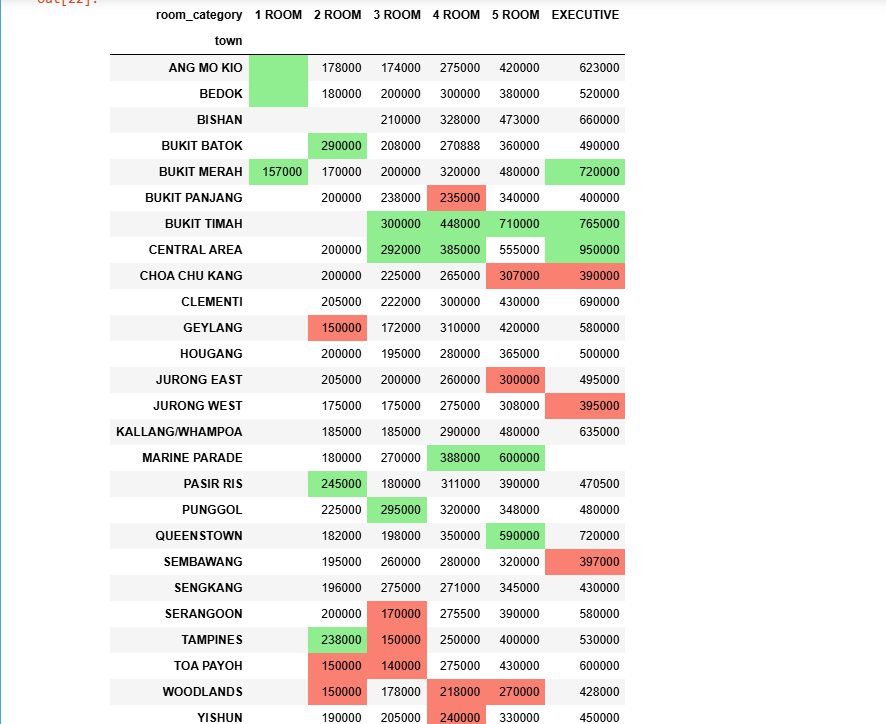
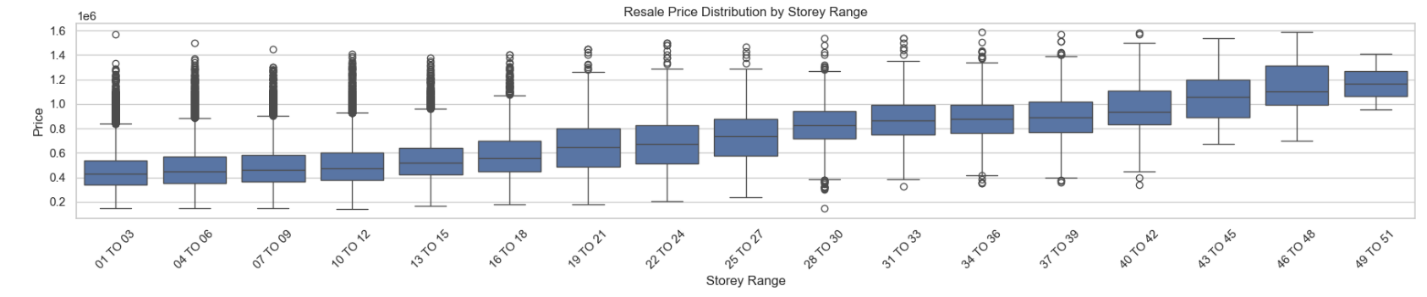
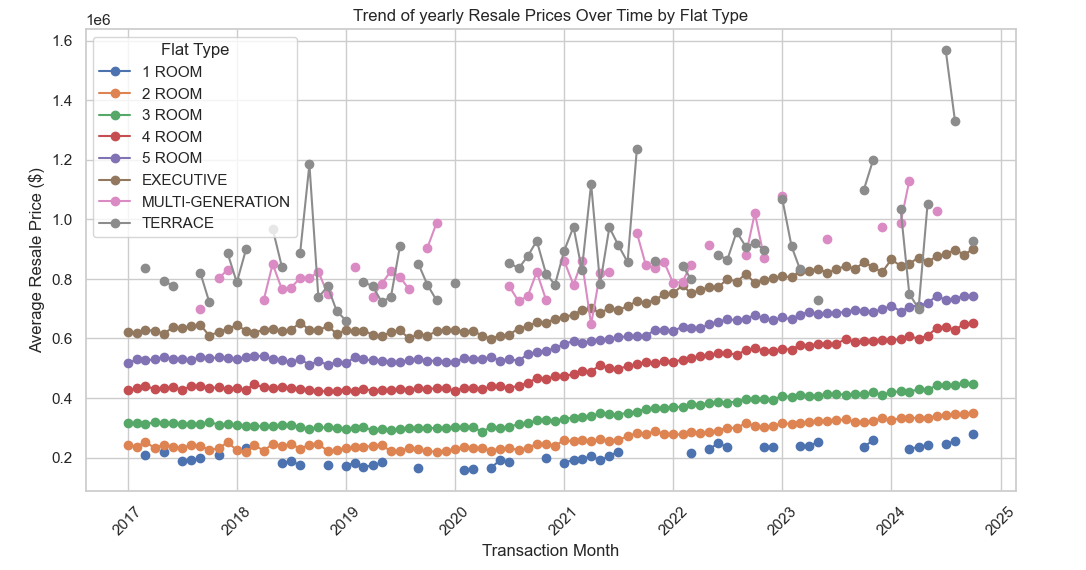
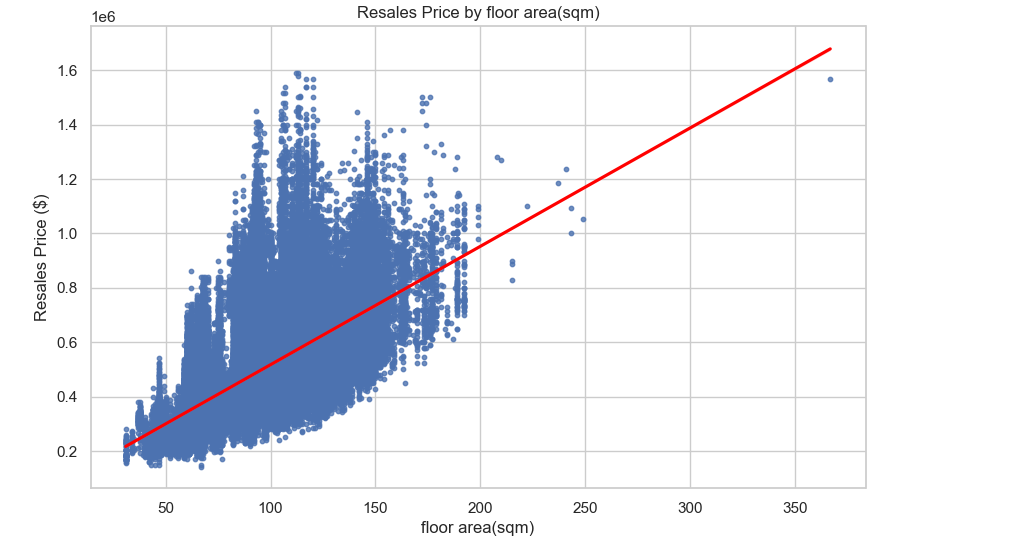
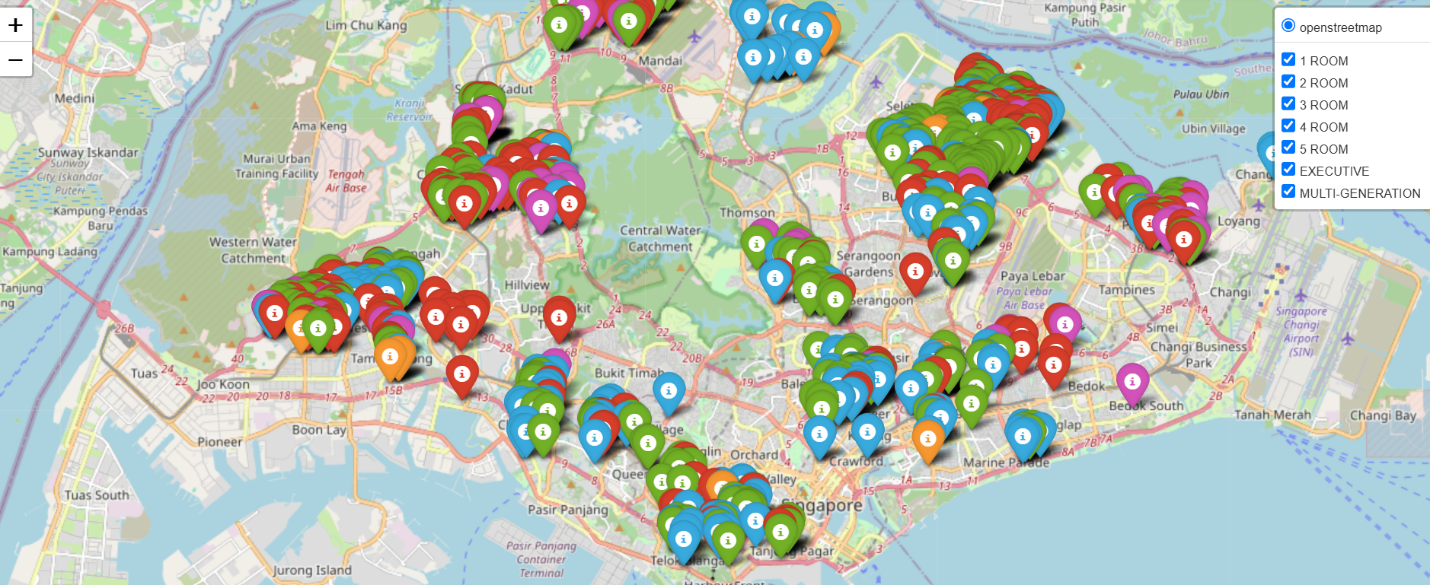
There are 192613 records and 11 columns in the dataset.There is no null values in every column but there are 293 duplicate rows.

* Identification of trends and outliers. 

**3.2 Data Cleaning and Wrangling**

* Remove the duplicate records and remove those transaction is not in open resale market.
* Converted categorical columns (month, flat\_type, storey\_range, remaining\_lease) into numerical representations. And for town and flat\_model using one-hot encoding.
* Created new features: binned flat\_type and storey\_range.

**3.3 Data Visualization**

* Created plots to understand patterns and insights (please refer to KPSResaleflat.ipynb for python code and insights):
  + **Table charts**: Average resale prices by flat type and town. 
  + **Boxplots**: Resale price distribution across storey\_range.
  + **Trend lines**: Average resale price trends over months. 
  + **Scatter plots**: Relationship between floor\_area and resale\_price.
  + **Point plots**: Visualizing one-month resale transactions on a Singapore map, with toggle functionality for flat types. (python code please refer to sgpmap\_toggle.ipynb)

**4. Machine Learning Workflow**

**4.1 Data Preprocessing**

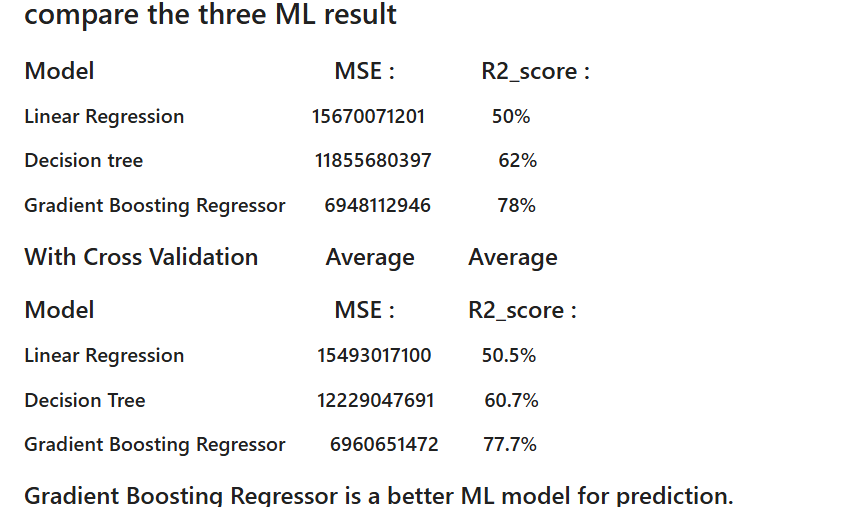
* **Feature selection 1**: month, flat\_type, storey\_range , floor\_area, remaining\_lease.
* **Feature selection2**: month, flat\_type, storey\_range , floor\_area, lease\_commencement\_date , remaining\_lease, town and flat\_model (with one\_hot encoding).
* Split the data into training and testing sets: 70% train, 30% test.

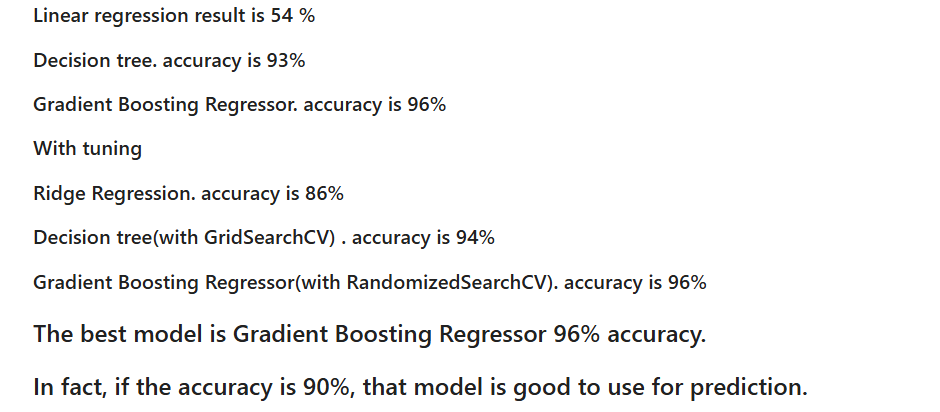
**4.2 Model Training**

The following models were trained with the two sets of feature selection: (use r2 score as the accuracy of the model.)

1. **Linear Regression**: A baseline model for predicting resale prices.
2. **Decision Tree Regressor**: Captures non-linear relationships.
3. **Gradient Boosting Regressor**: Ensemble learning for enhanced accuracy.

**4.3 Cross-Validation and Hyperparameter Tuning**

* **Cross-validation**: Performed k-fold (k=5) cross-validation to detect overfitting or underfitting. 
* **Hyperparameter tuning**: Utilized grid search or RandomizedSearchCV for optimal parameters.
  + Decision Tree: Tuned max\_depth, min\_samples\_split, min\_samples\_leaf, max\_features.
  + Gradient Boosting: Tuned learning\_rate, n\_estimators, max\_depth, min\_samples\_leaf, max\_features and loss.



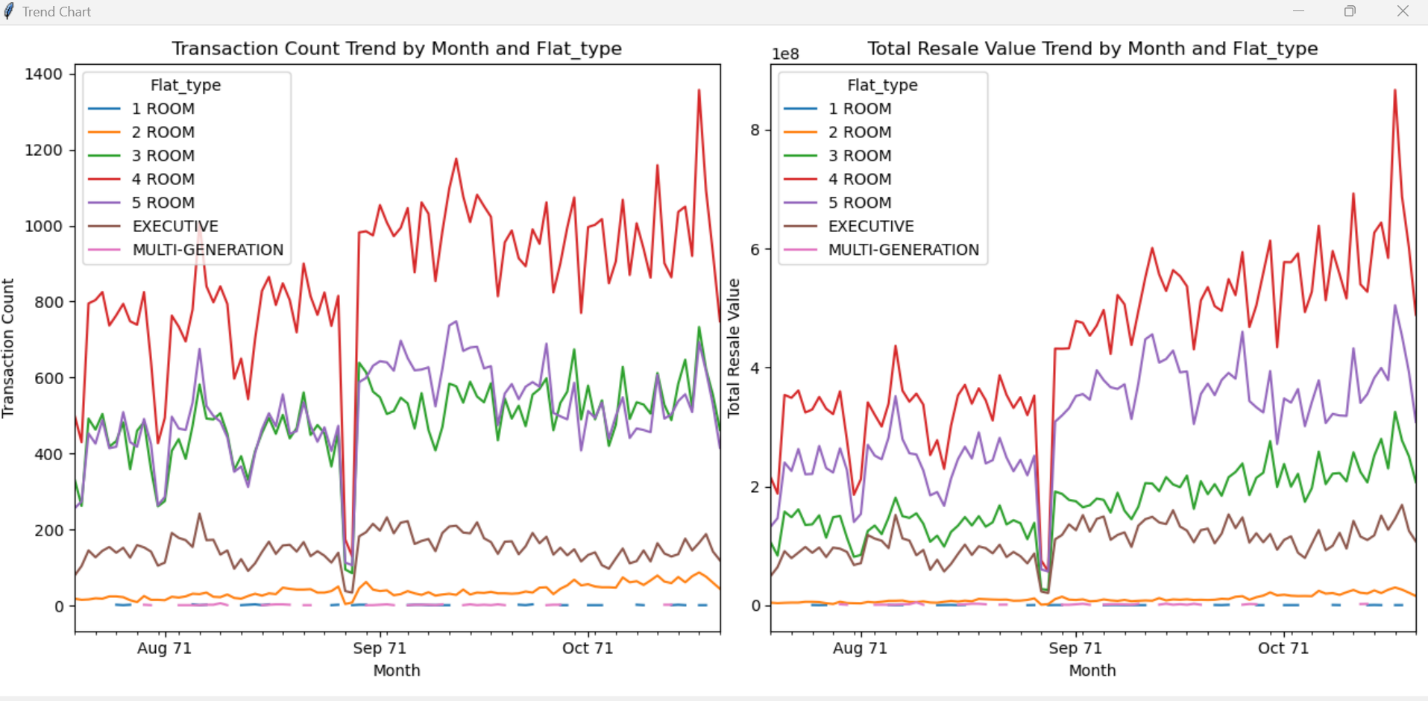
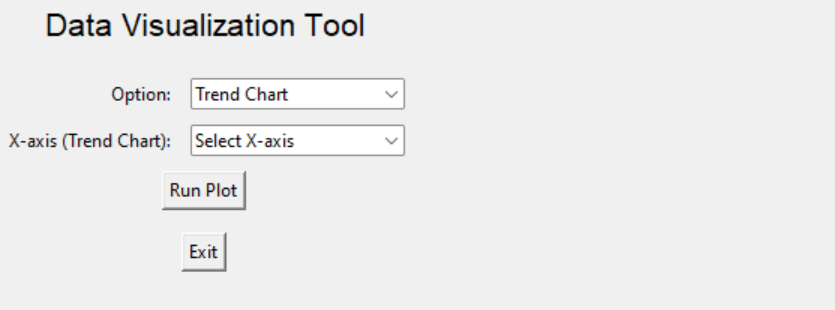
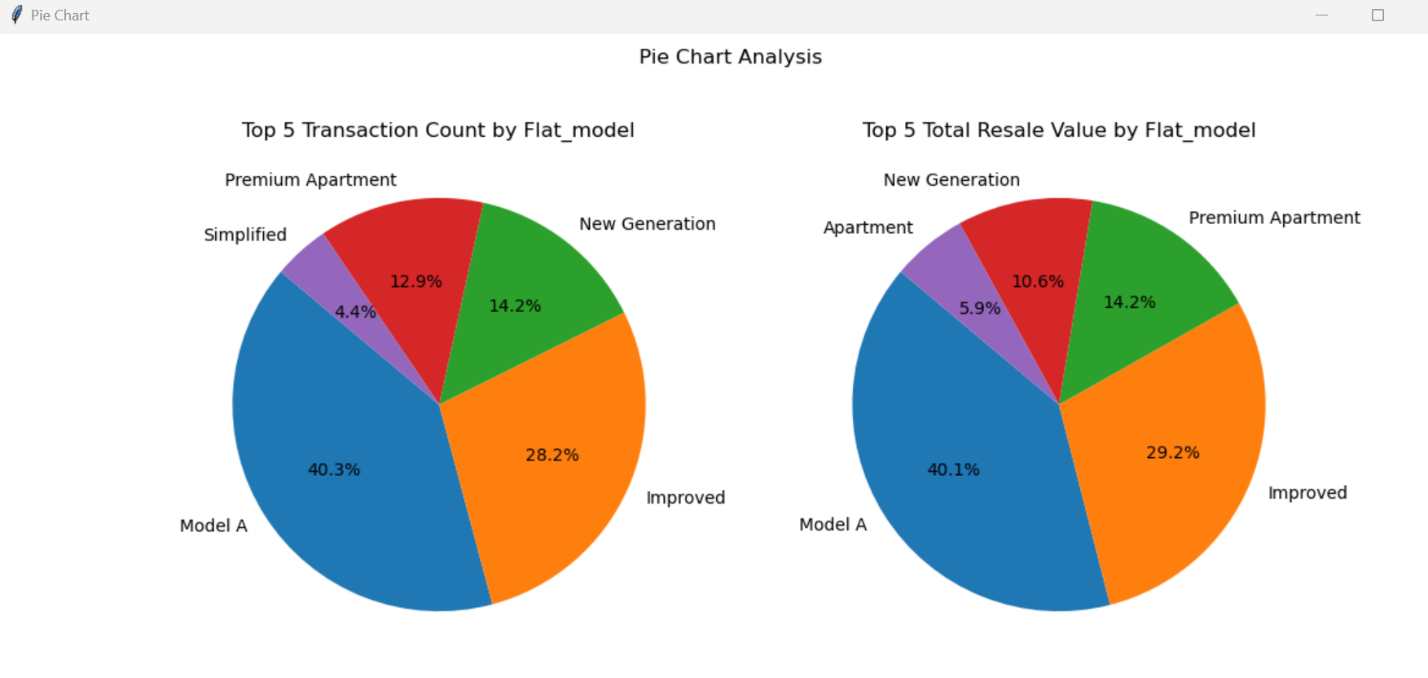
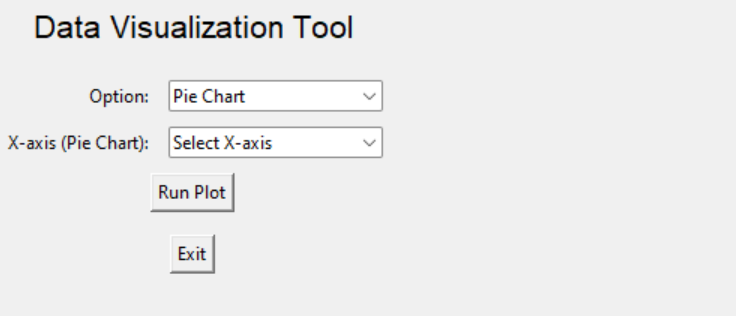
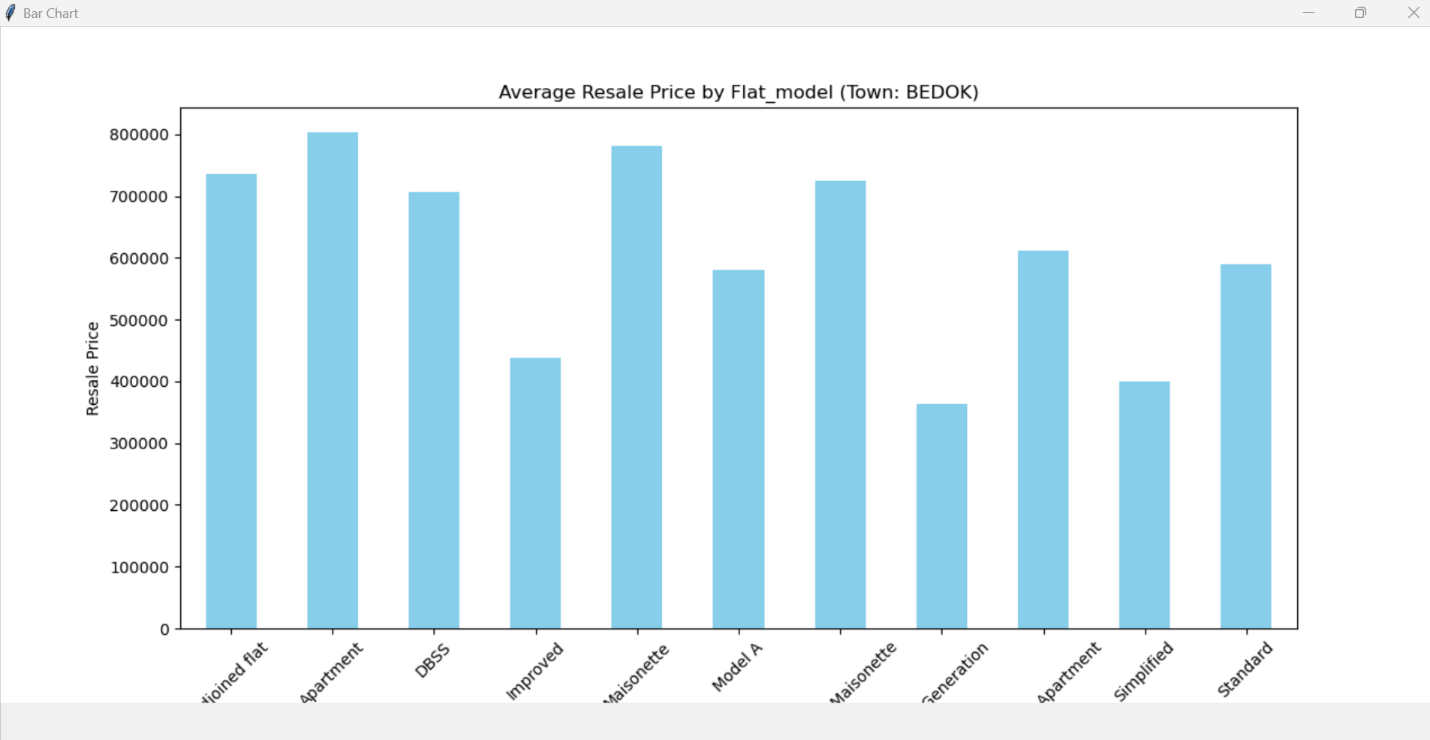
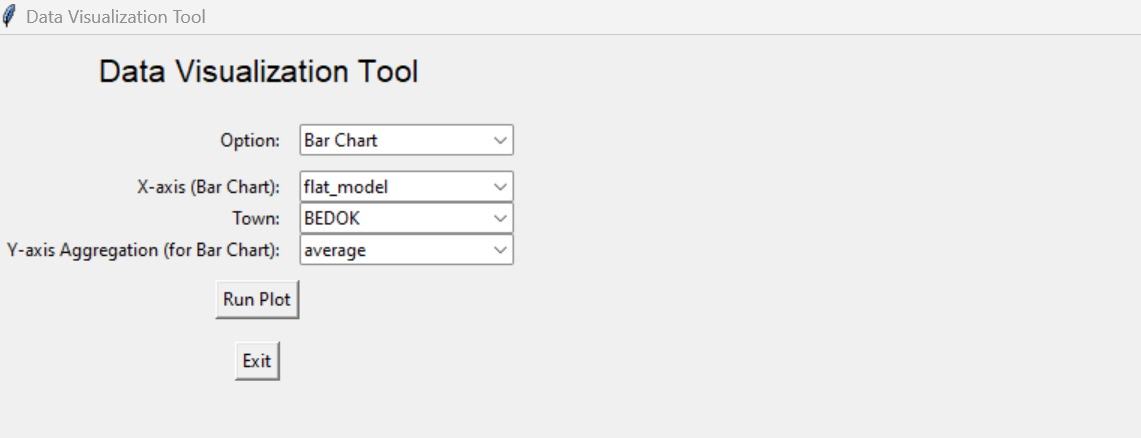
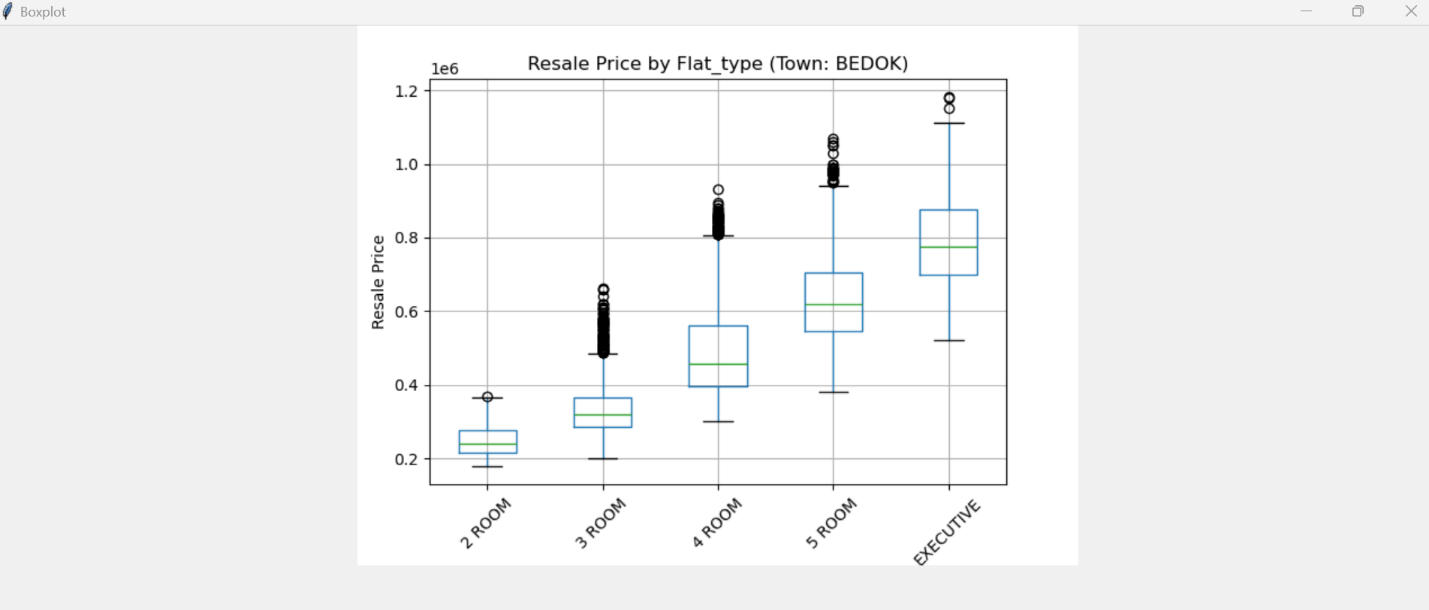
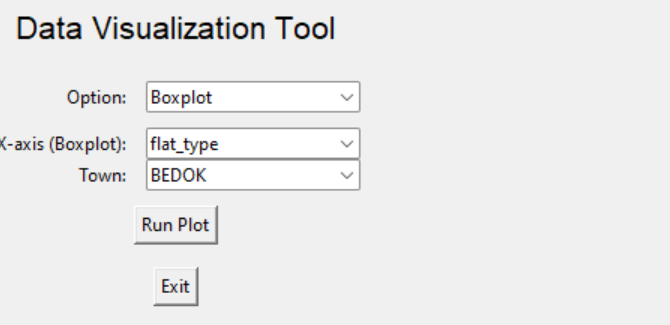
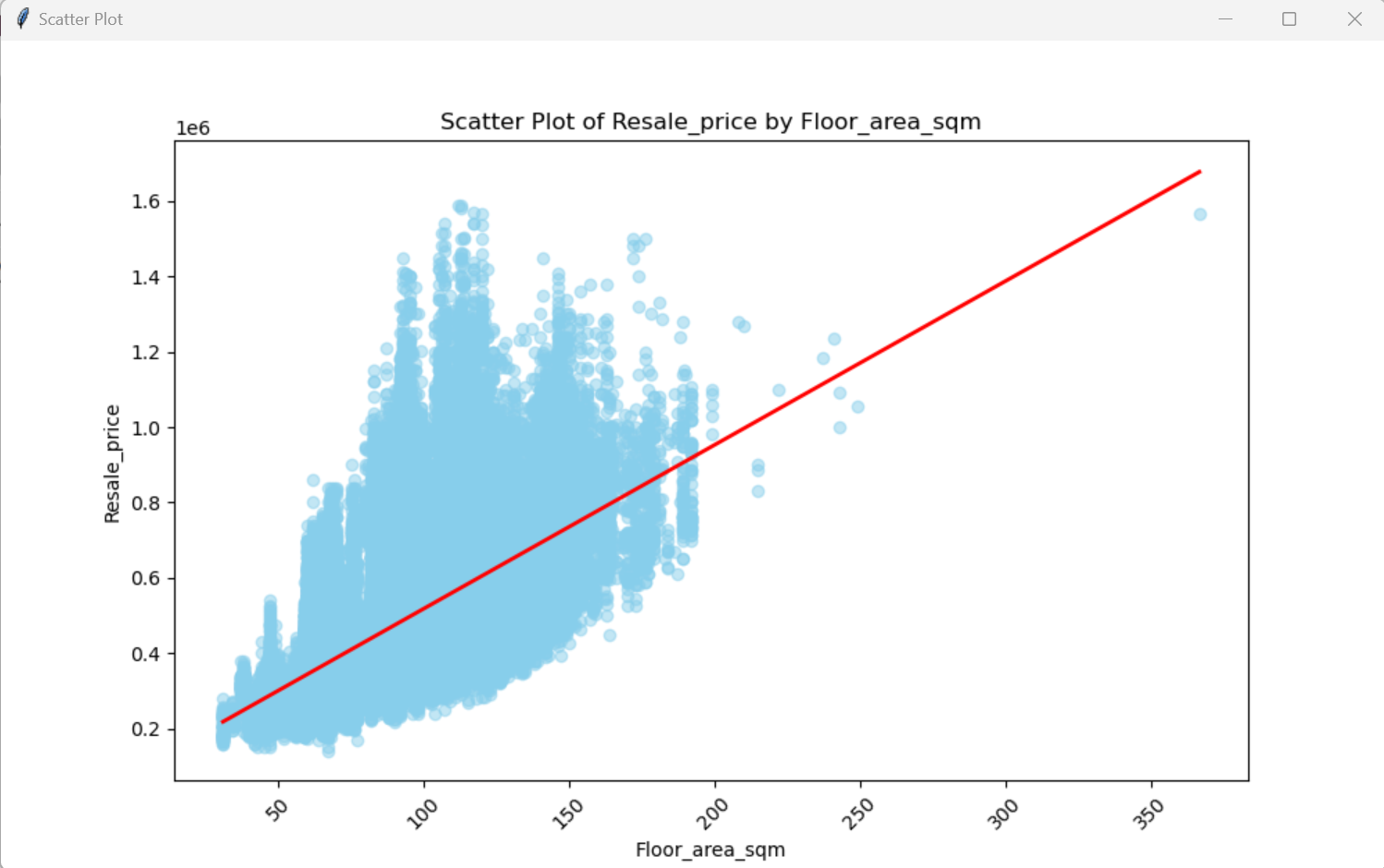
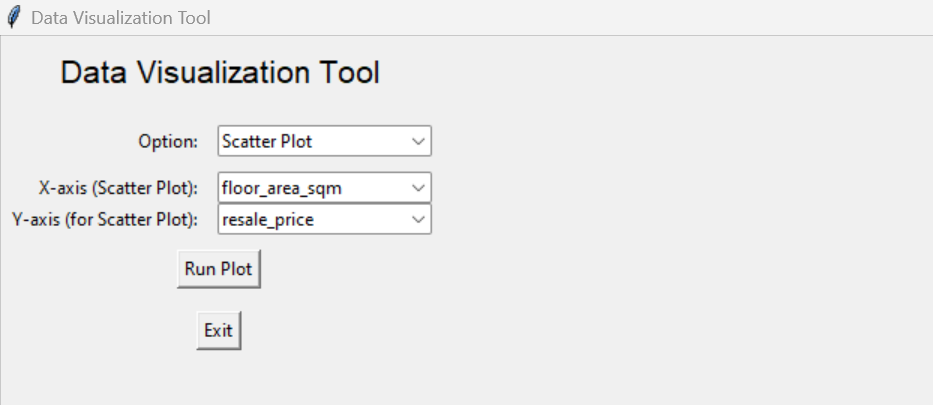
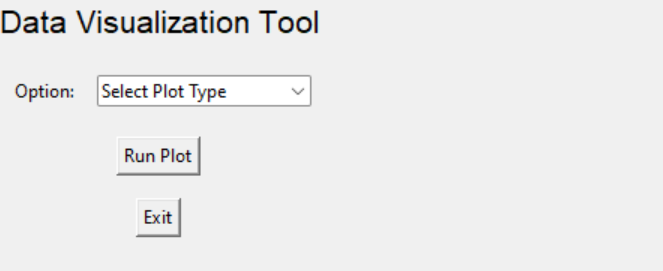
**4.4 Model Evaluation and Comparison**

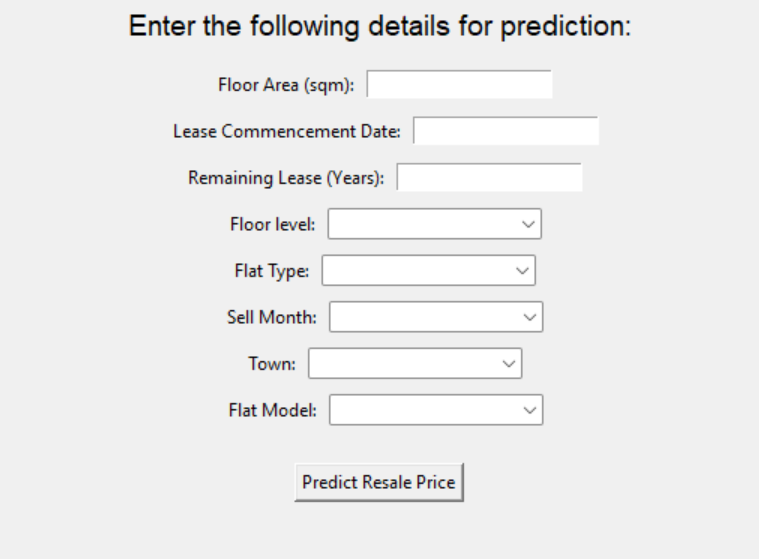
* Metrics: Mean Squared Error (MSE), R² score(accuracy) was used to compare between models.
* Gradient Boosting Regressor outperformed others with the lowest MSE and highest R² score(accuracy). So Selected **Gradient Boosting Regressor** for deployment.

**5. Deployment: Tkinter-Based Application**

A Tkinter application was developed with the following features:

1. **Visualization Menu**: Options for scatter plot, boxplot, bar chart, pie chart and trend plot.(Python code please refer to enquiry\_menu.ipynb. Additional visuals can be viewed by running the program with different selections.)



1. **Prediction Tool**: Users can input features floor\_area, lease commencement date, remaining\_lease, floor level, flat\_type, intended Sell month , town and flat\_model to predict the resales price using the Gradient Boosting model.( Python code please refer to prediction.ipynb)

**6. Conclusions and Recommendations**

* Gradient Boosting Regressor provides the best accurate resale price predictions.
* Visualizations reveal critical trends in the HDB resale market.
* The interactive map is a valuable tool for exploring location-based trends.

**7. Future Enhancements**

* Include external economic factors (e.g., interest rates), price per sqm and the pattern of annual or quarterly percentage changes in average resale prices in the analysis.
* Expand the Tkinter prediction application with more validation and better data input using just Block and street name.
* Integrate a web-based interface for wider accessibility.