Data Sprint:

HDB Price Analysis & Predictor

Introduction & Objectives

- Problem statement:
  - Predict resale prices of HDB flats accurately
- Why it matters:
  - Helps WOW! agents offer data-driven advice
- Goal:
  - Build an accurate model to generate actionable insights

About the Data

- **Where the data comes from:**
  - ~150,000 rows and 78 columns of real HDB resale records, along with public info about nearby malls, hawker centres, schools, and MRT stations, etc.
- **What the data includes:**
  - Town and flat type (e.g. 3-room in Queenstown)
  - Floor size and storey level
  - Age of flat
  - Distance to MRT and other amenities
- **What we're predicting:**
  - The **resale price** of the flat in Singapore

Data Cleaning & Preparation

- Null handling

  Mall_Within_500m → 92,789 nulls

  Mall_Within_1km → 25,426 nulls

  Hawker_Within_500m → 97,390 nulls

  Hawker_Within_1km → 60,868 nulls

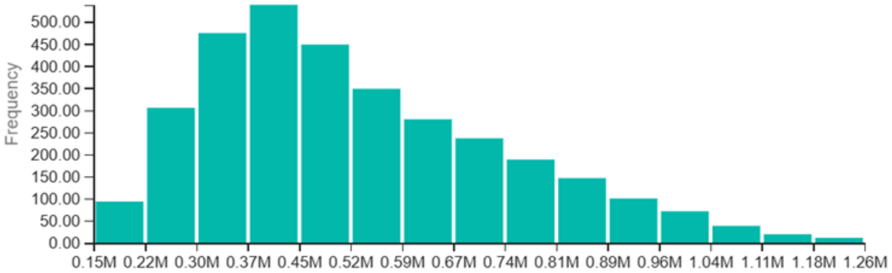
  Hawker_Within_2km → 29,202 nulls

  Imputed all missing values with 0

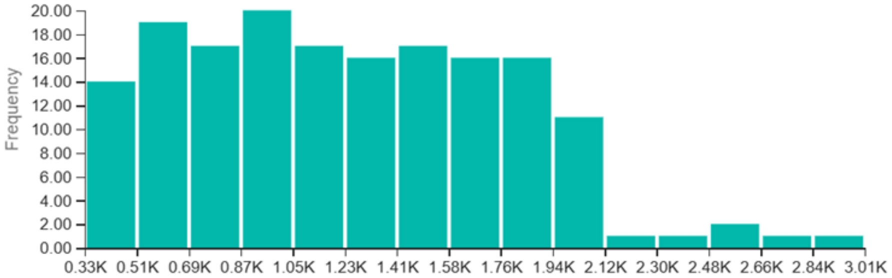
- Dropped low-relevance or duplicate columns

Exploratory Data Analysis

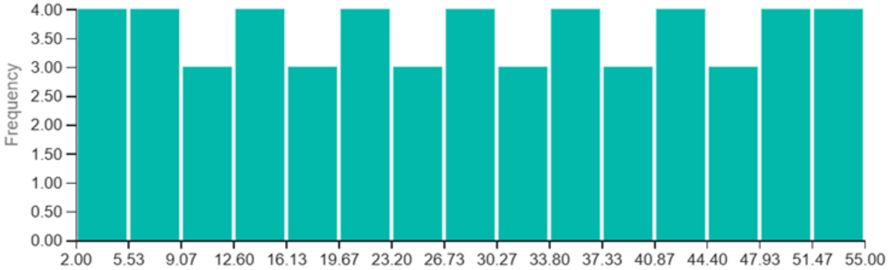
Distribution of Resale Prices



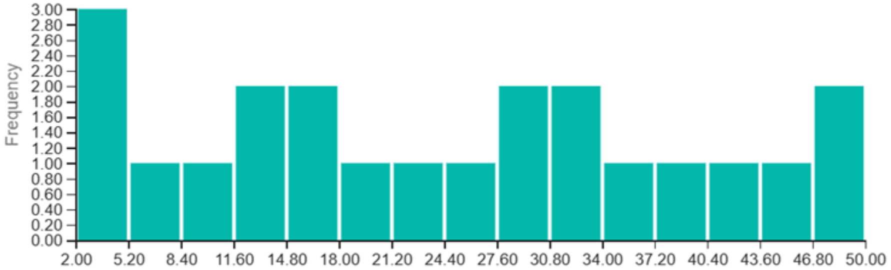
Distribution of Floor Area (per sq ft)



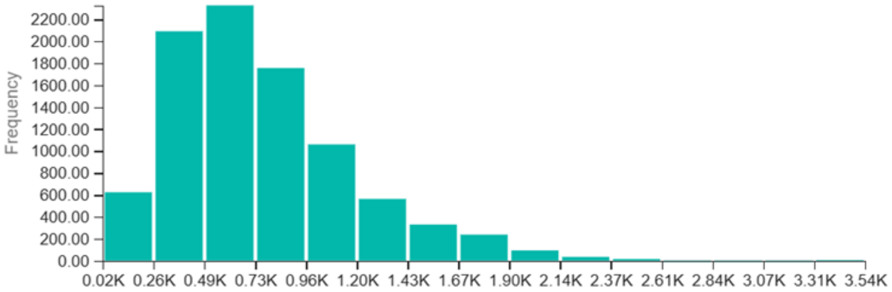
Distribution of Age of HDB Flat



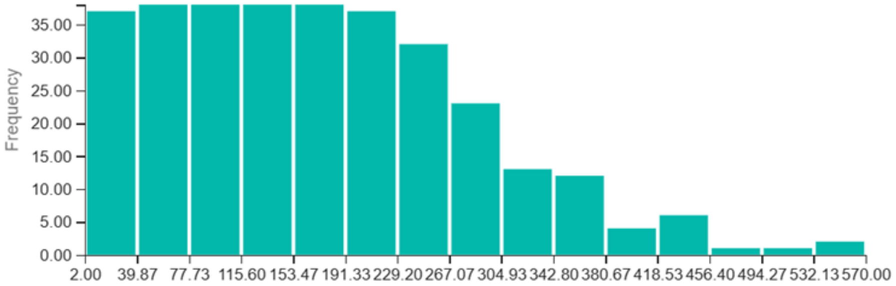
Distribution of HDB Flat Levels



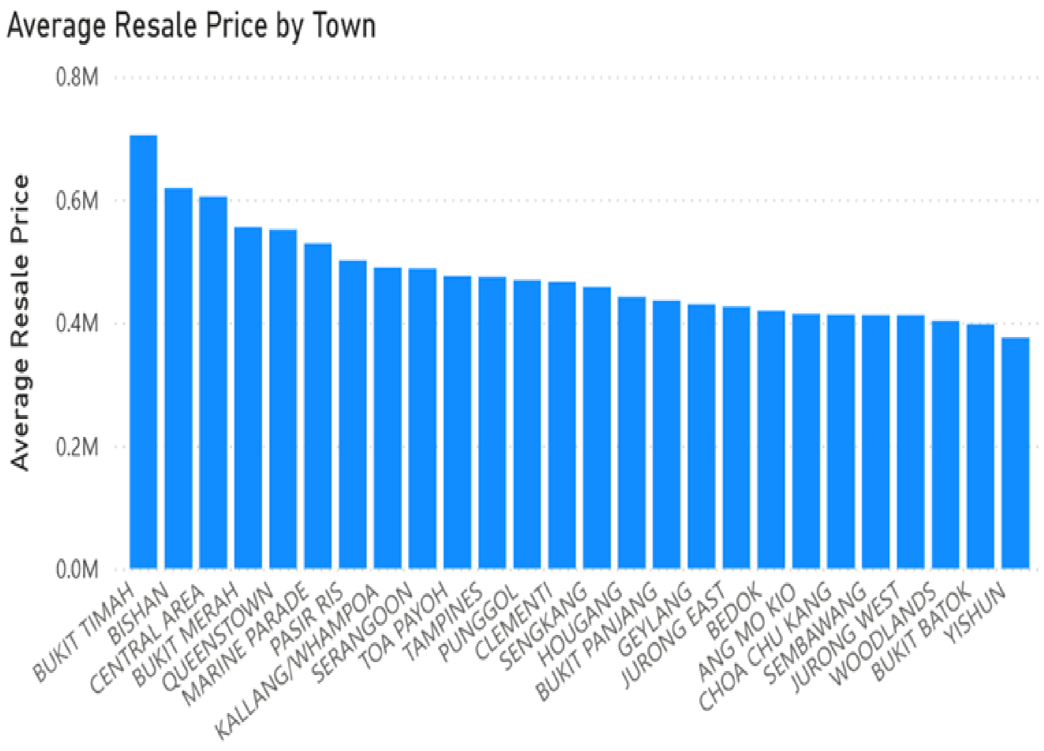
Distribution of Distance to Nearest MRT Station (m)

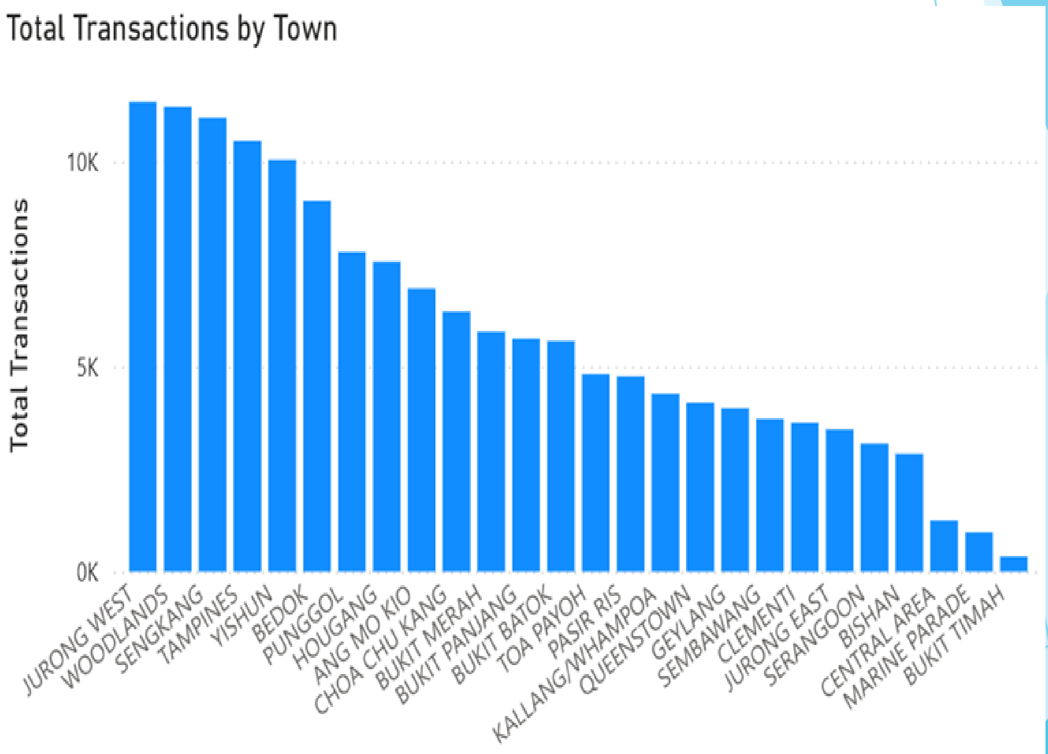


Distribution of Total Units in a HDB Flat

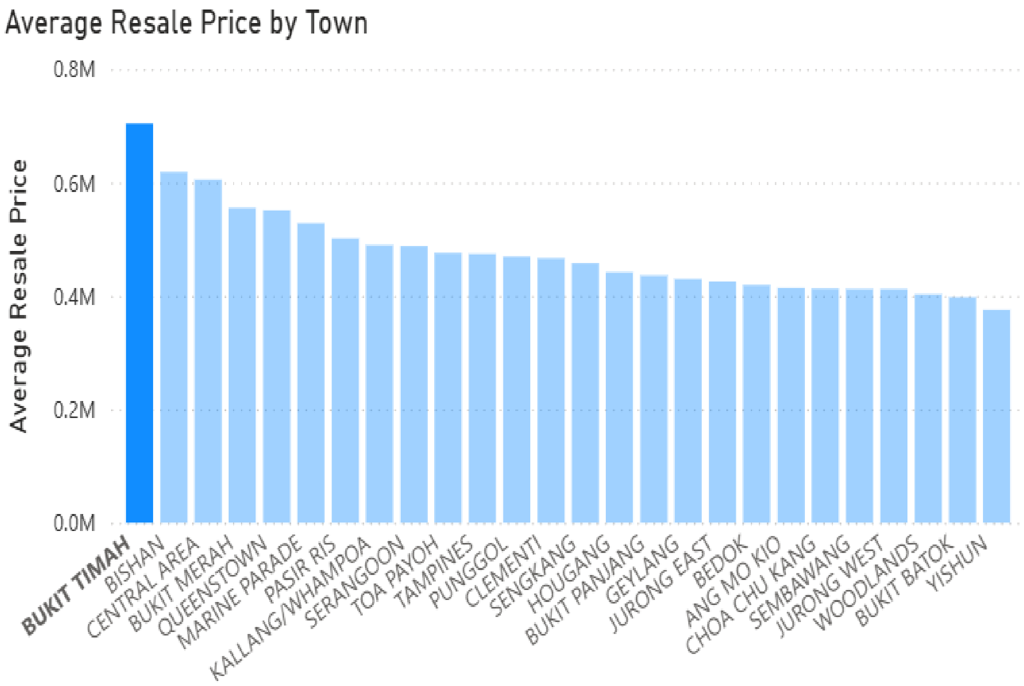


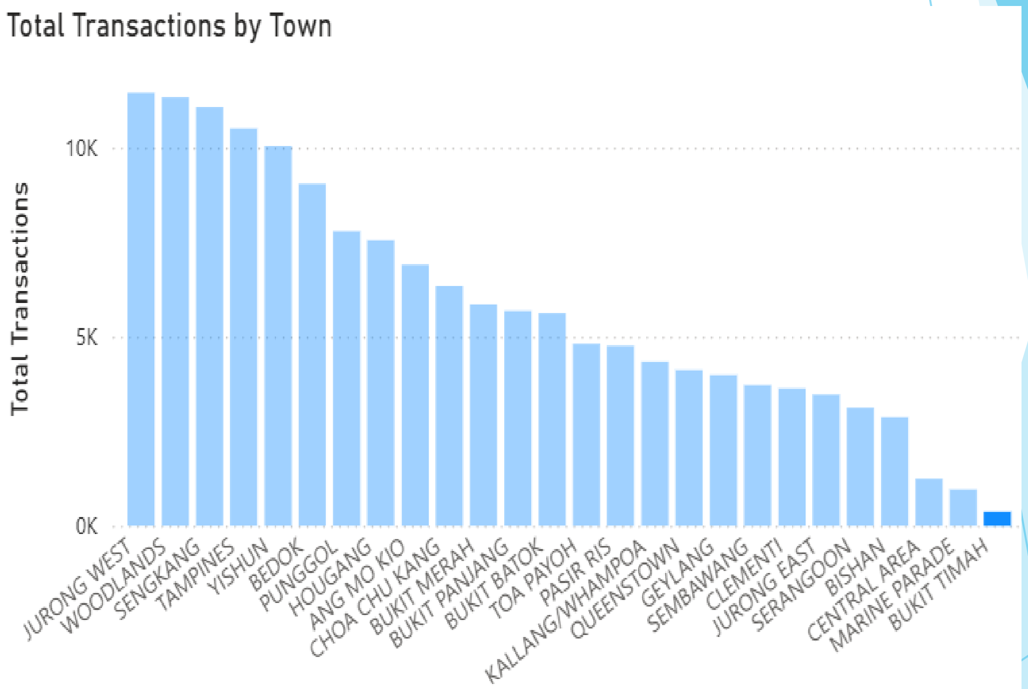
Analysis of Resale Prices



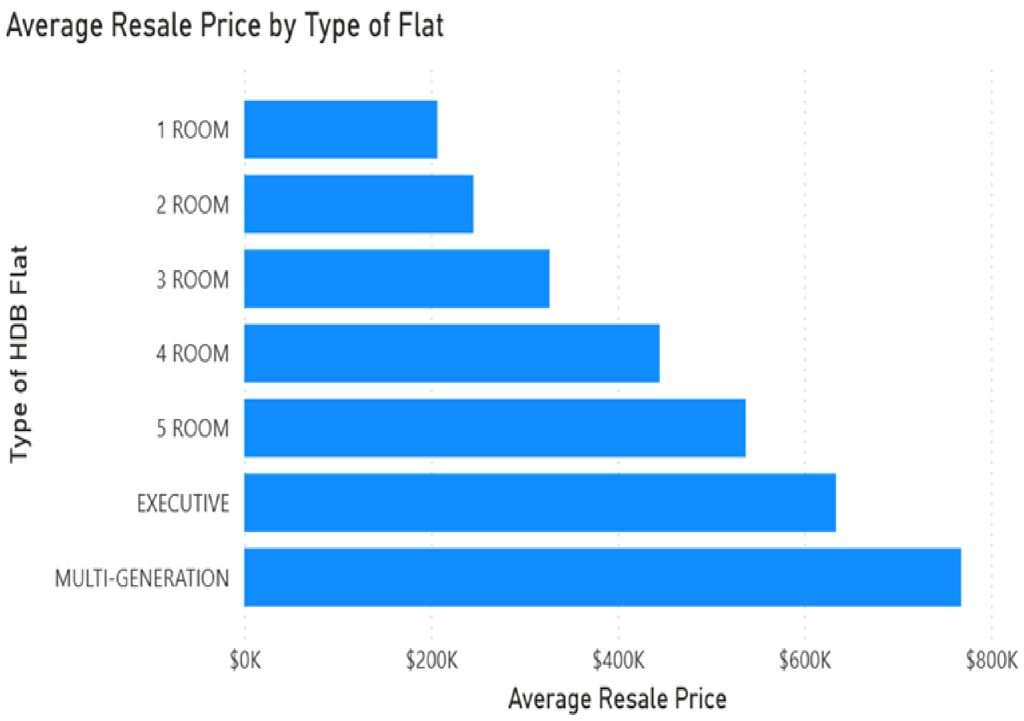


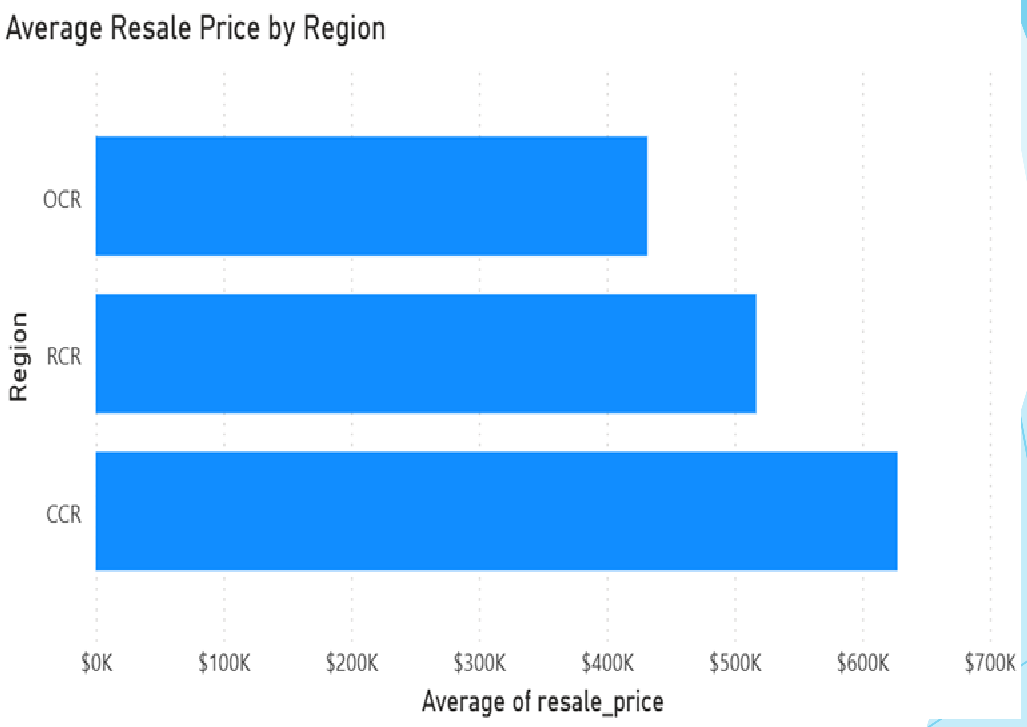
Analysis of Resale Prices



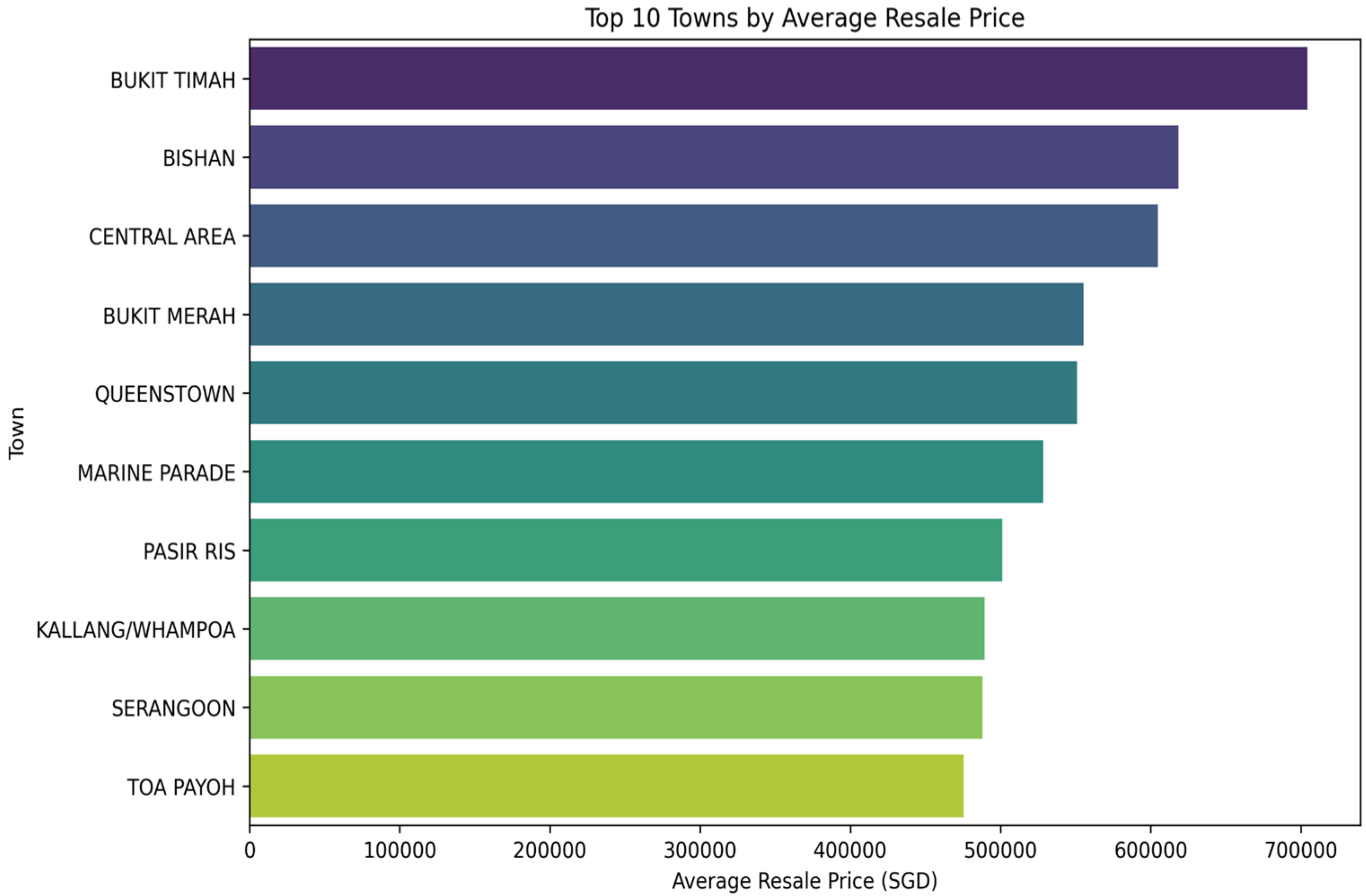


Analysis of Resale Prices

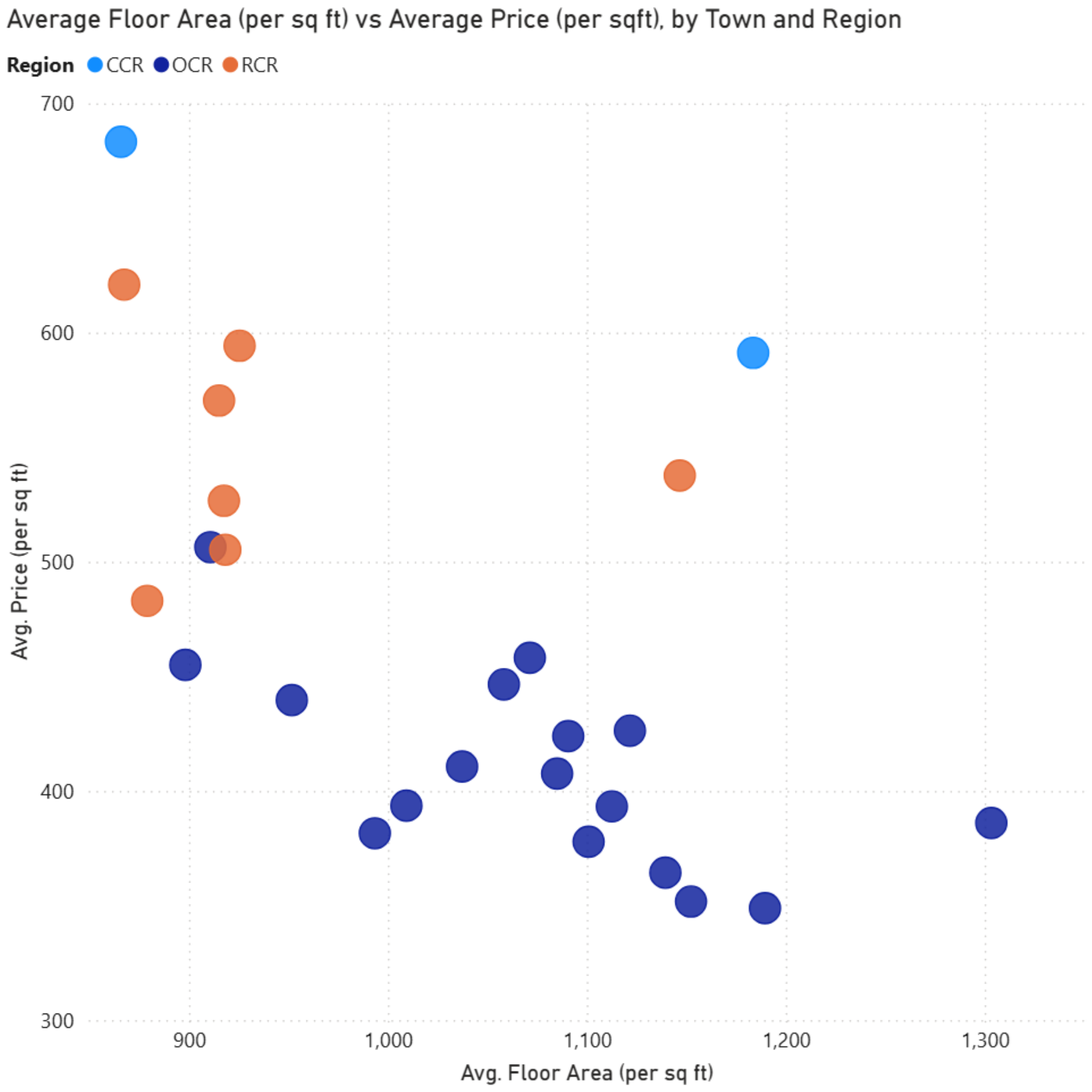




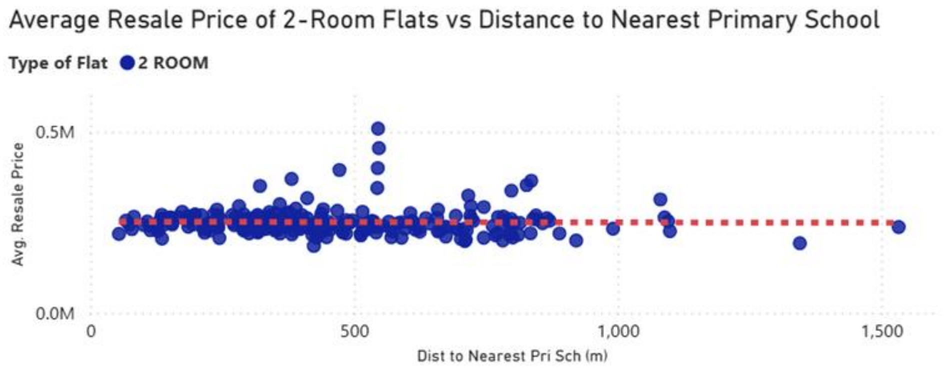
Top 10 Towns by Average Resale Price

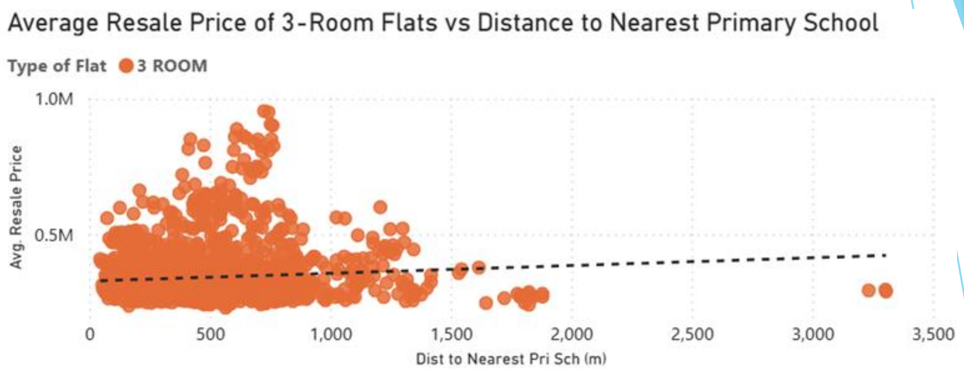


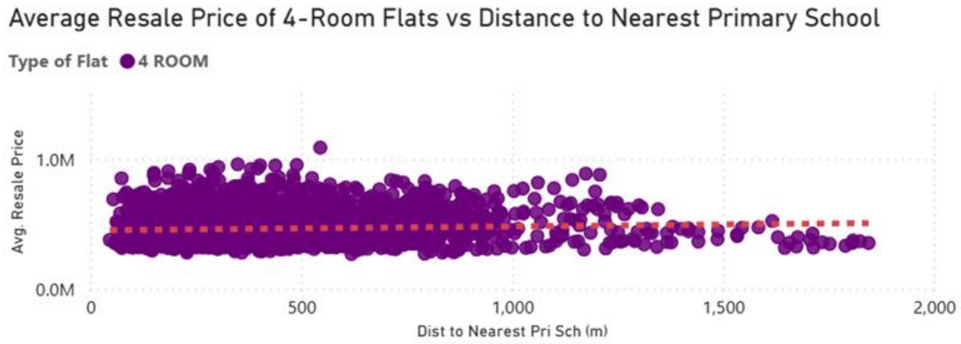
Exploring Price by Region

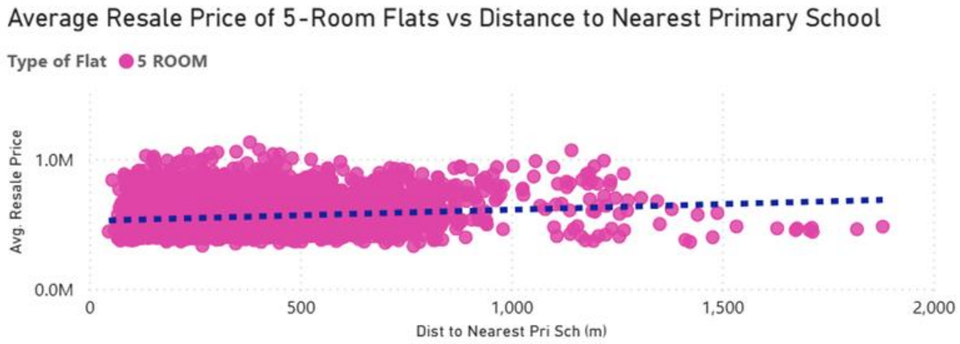


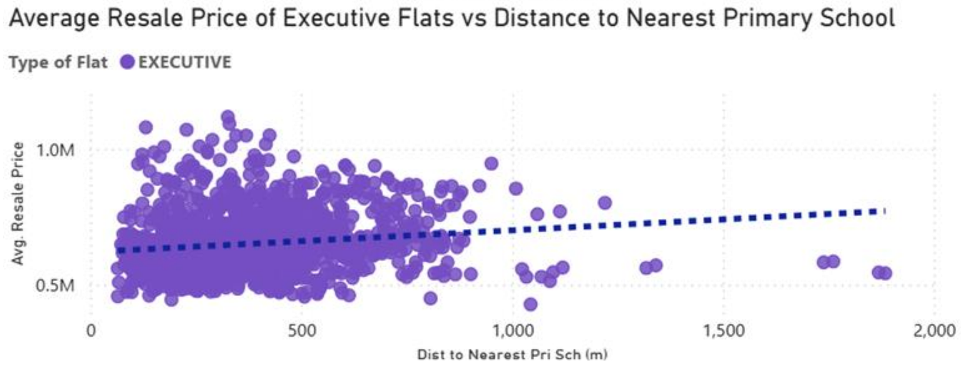
Resale Prices vs Dist to Nearest Pri Sch

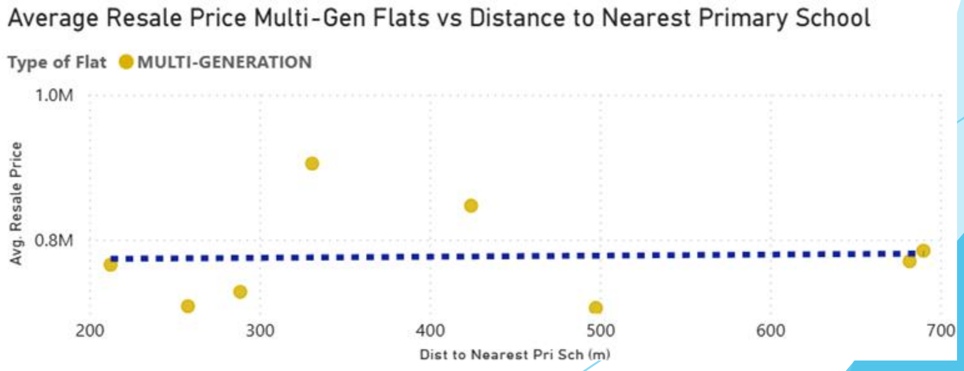






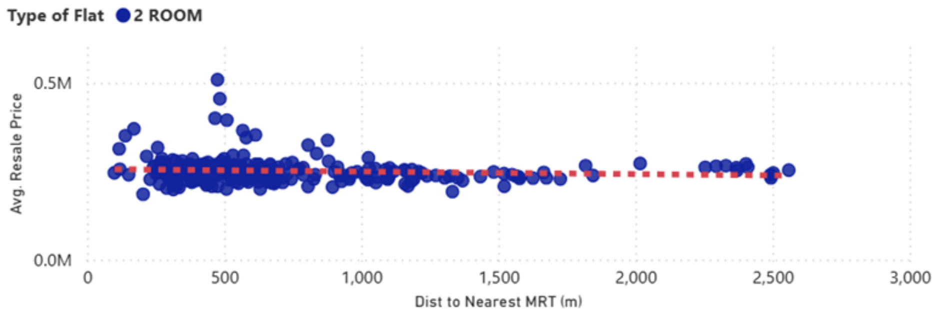




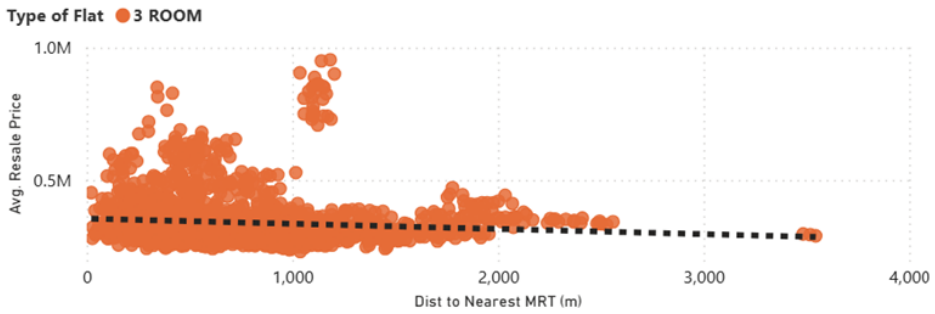


Resale Prices vs Dist to Nearest Pri Sch

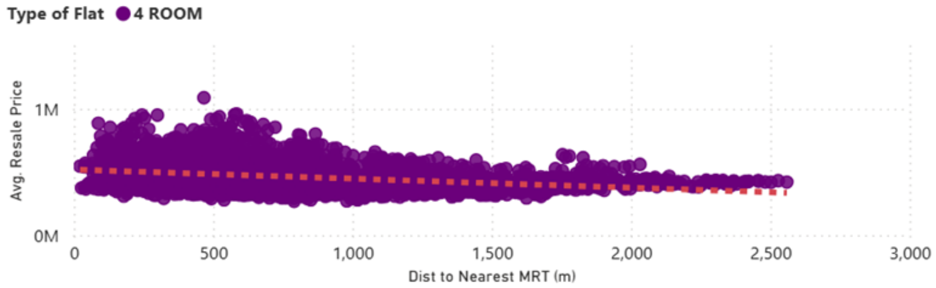
Average Resale Price of 2-Room Flats vs Distance to Nearest MRT Station



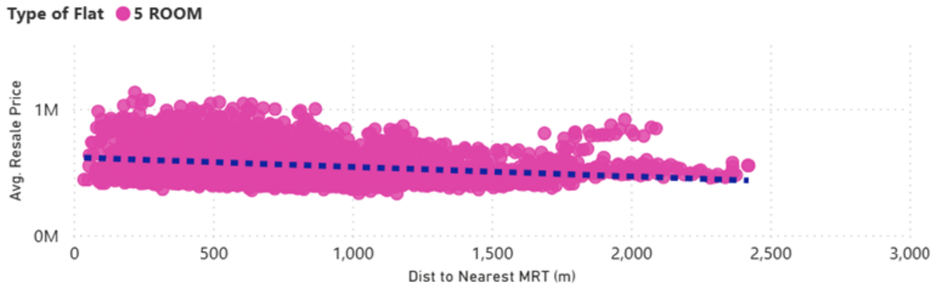
Average Resale Price of 3-Room Flats vs Distance to Nearest MRT Station



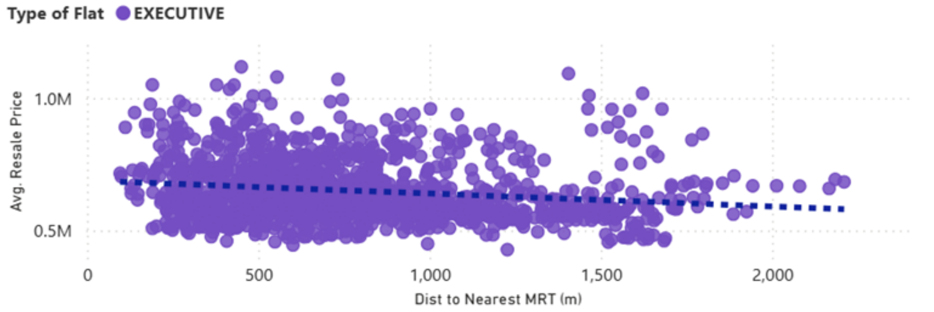
Average Resale Price of 4-Room Flats vs Distance to Nearest MRT Station



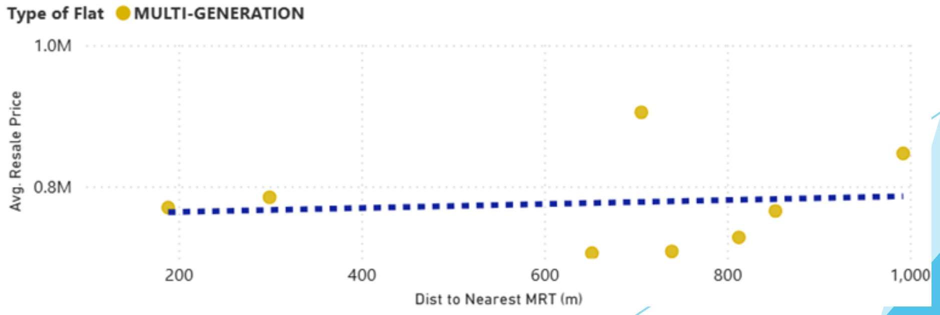
Average Resale Price of 5-Room Flats vs Distance to Nearest MRT Station



Average Resale Price of Executive Flats vs Distance to Nearest MRT Station

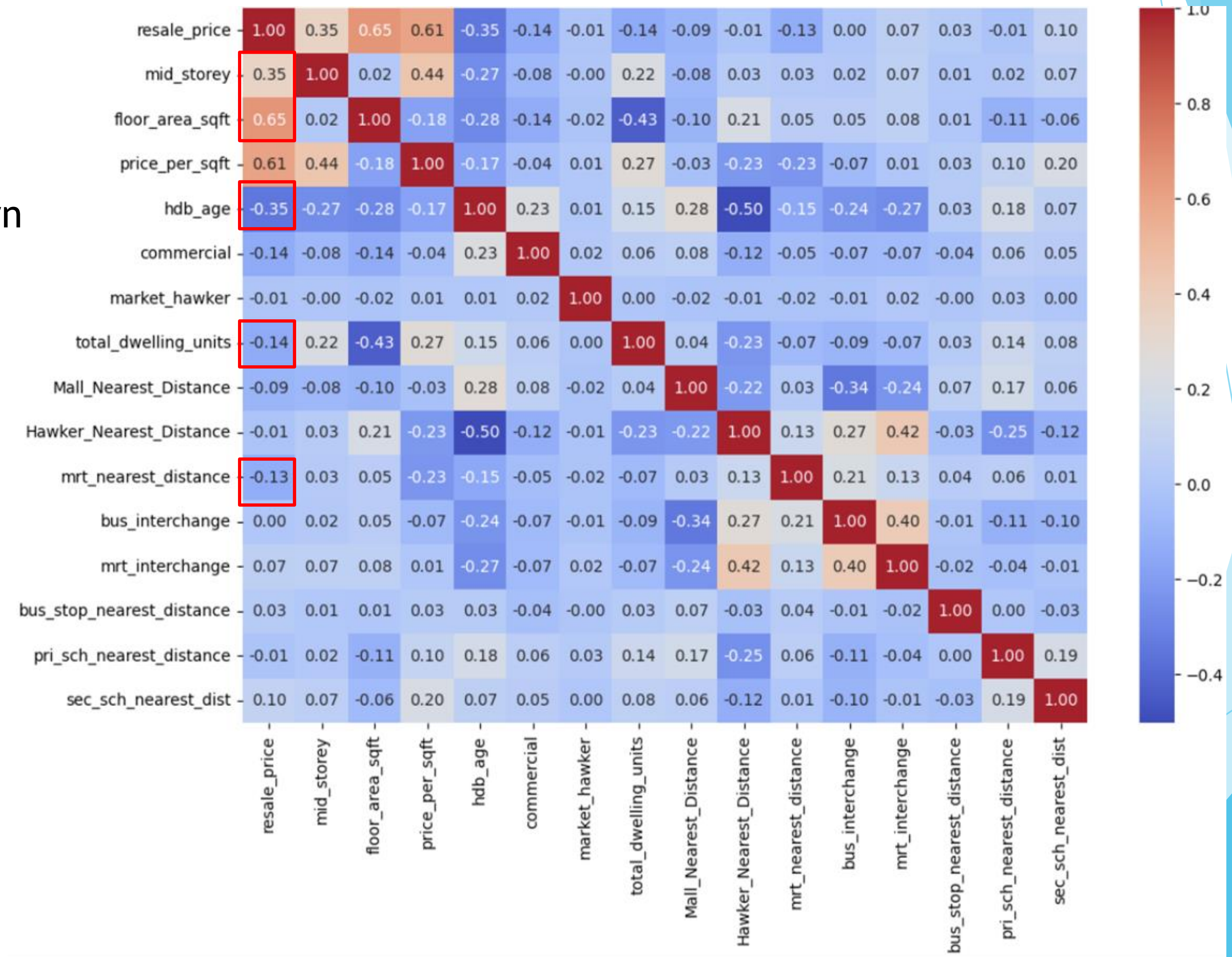


Average Resale Price of Multi-Gen Flats vs Distance to Nearest MRT Station

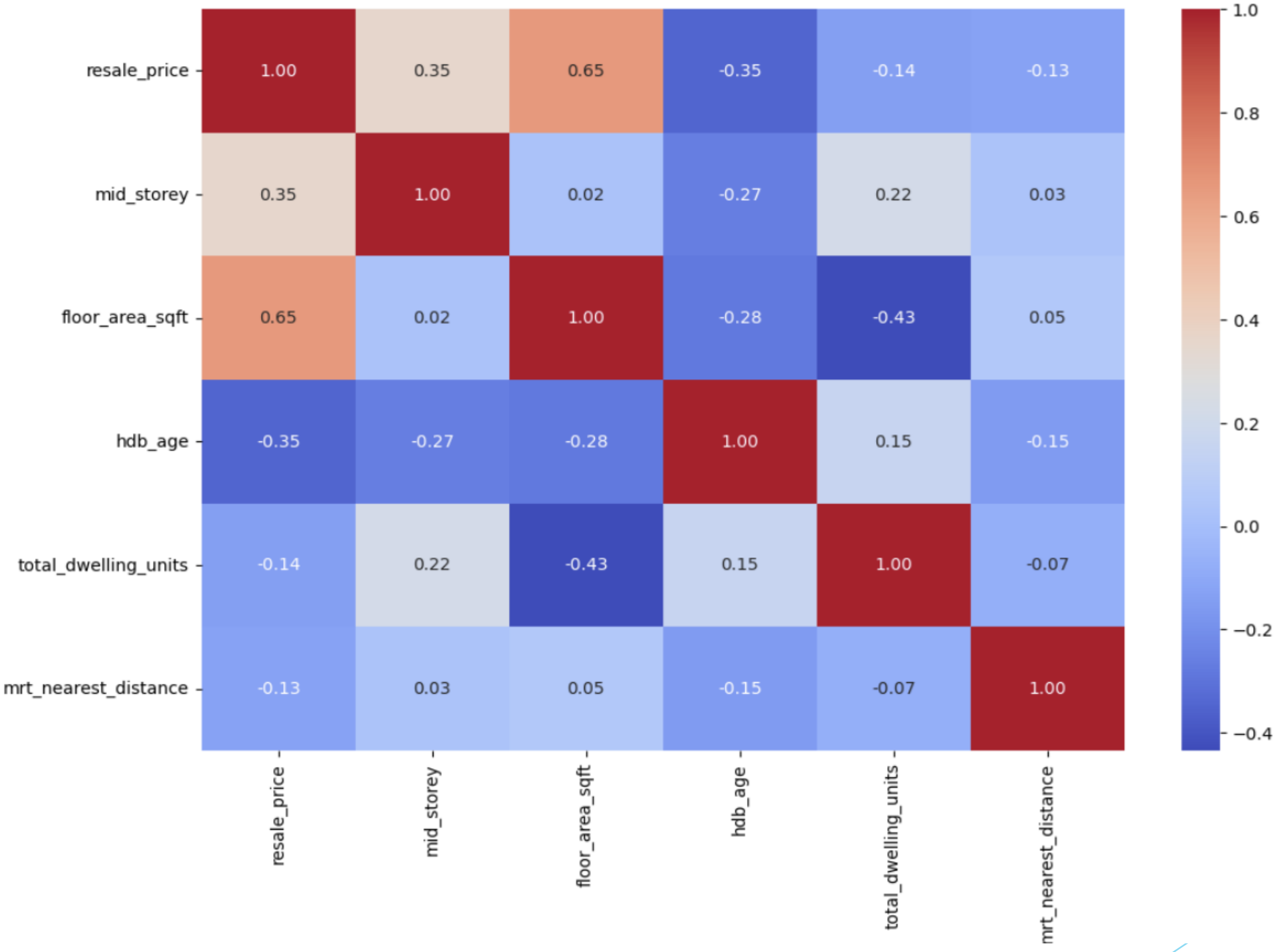


Correlation Between Features & Target

- We narrowed down to 5 key features

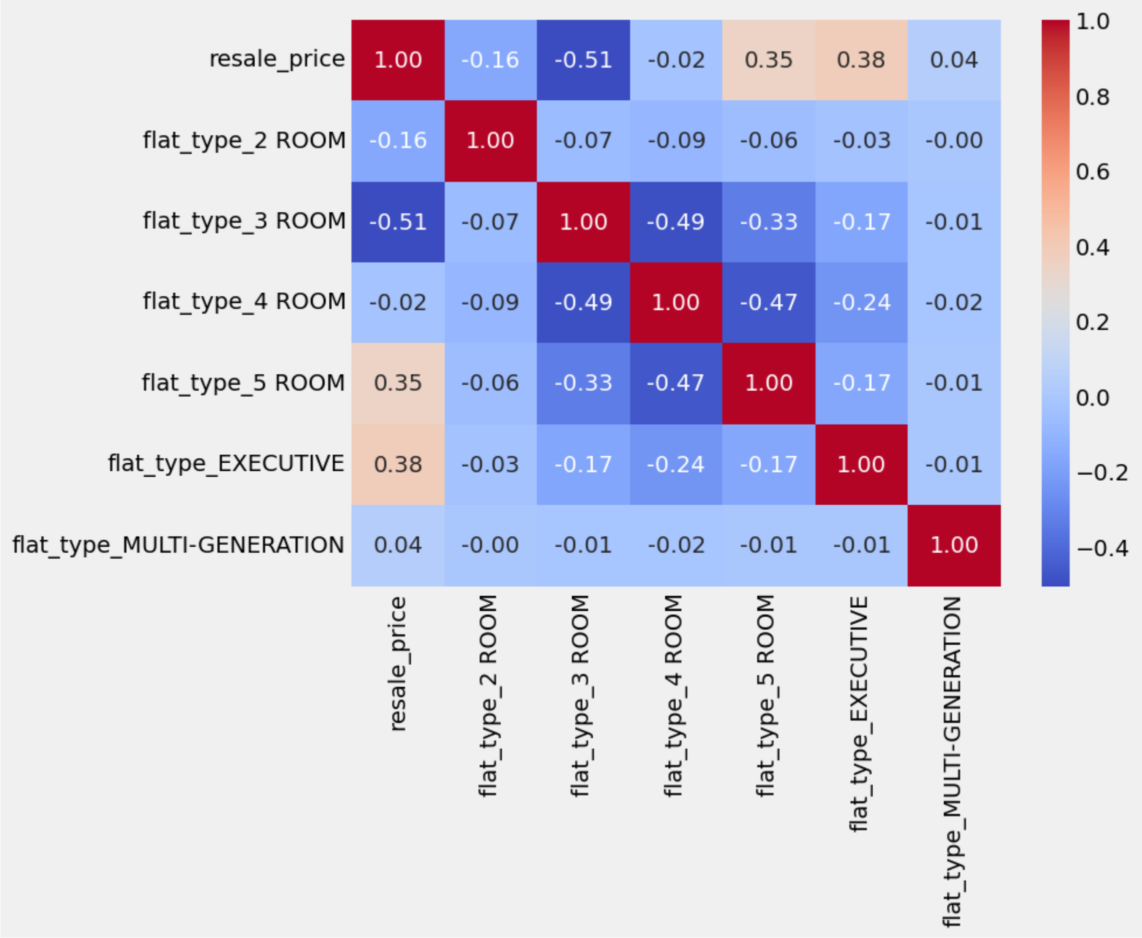


Correlation Between Features & Target



Correlation Between Features & Target

- One-hot encoded that flat type feature to check if flat type is of value to the model.
- It showed range from -0.51 to 0.38 which we deem to be significant and used for modelling.
- While we understand this is correlated with flat size, we included in the model as we designed this with a price predictor for buyers at the end.



Final Selected Numerical Features

- **Floor Area sqft vs Resale Price:**
  - **Strong positive correlation** (r ≈ 0.65): Larger flats tend to be more expensive.
- **Mid Storey vs Resale Price:**
  - **Moderate positive correlation** (r ≈ 0.35): Higher floors often fetch better prices due to views and privacy.
- **HDB Age vs Resale Price:**
  - **Moderate negative correlation** (r ≈ -0.35): Older flats tend to have lower resale prices; newer ones are more desirable.
- **Total Dwelling Units vs Resale Price:**
  - **Slight negative correlation** (r ≈ -0.14): May indicate that larger estates are less exclusive or older.
- **MRT Nearest Distance vs Resale Price:**
  - **Mild negative correlation** (r ≈ -0.13): Flats closer to MRTs generally sell for slightly higher prices.

Modelling Approach

- First decided on Linear Regression to study the potential linear relationship between the features and resale prices.

- Features used:
  - mid_storey
  - floor_area_sqft
  - hdb_age
  - total_dwelling_units
  - mrt_nearest_distance
  - flat_type: one-hot encoded
  - town: one-hot encoded

Model	RMSE	R^2 Score	Remarks
Linear Regression	57,111	0.840	Fast baseline, but limited to linear trends

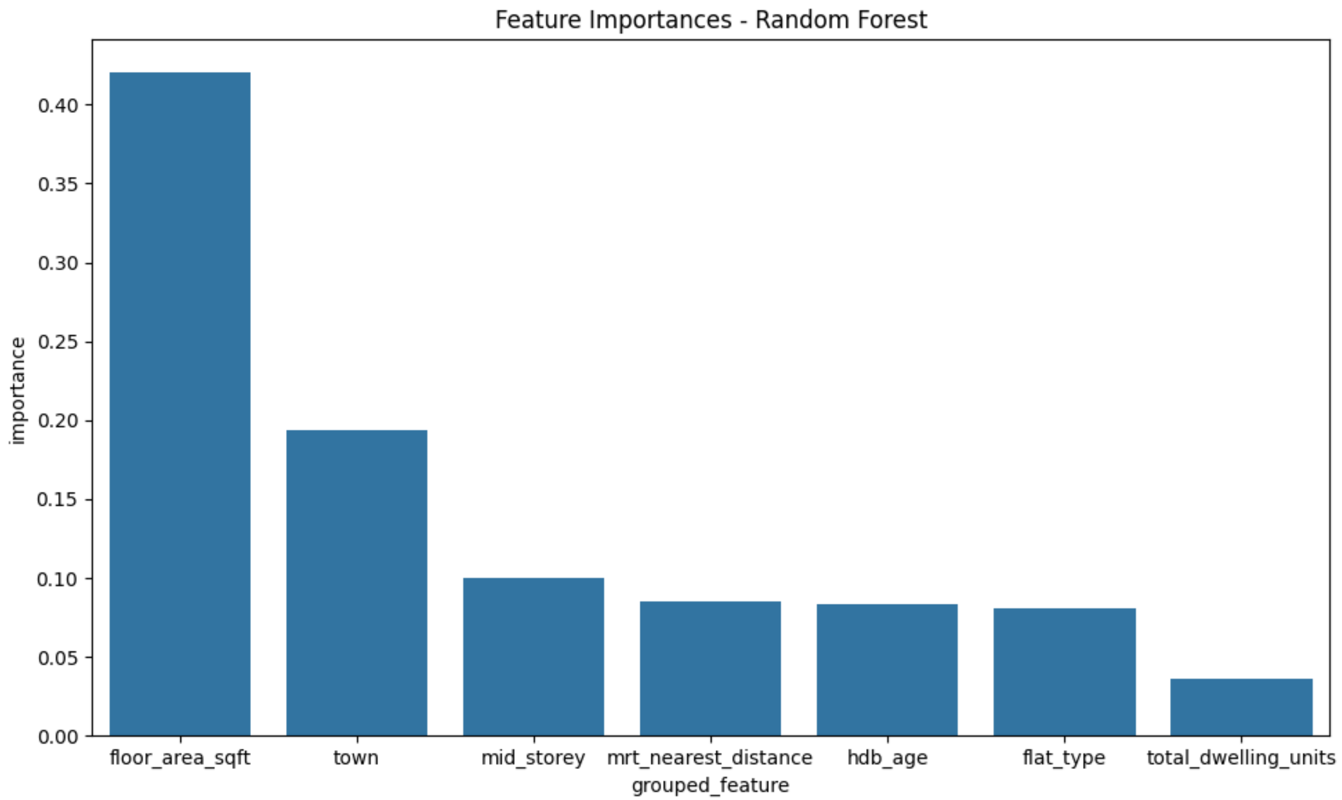
Modelling Approach & Choice

- Further explored 2 other models to test if a better performing model can be found.

Model	RMSE	R^2 Score	Remarks
Ridge Regression	57,111	0.840	Regularized linear model, no gain in performance
Random Forest	**39,375**	**0.924**	Best performer; captured non-linear relationships well

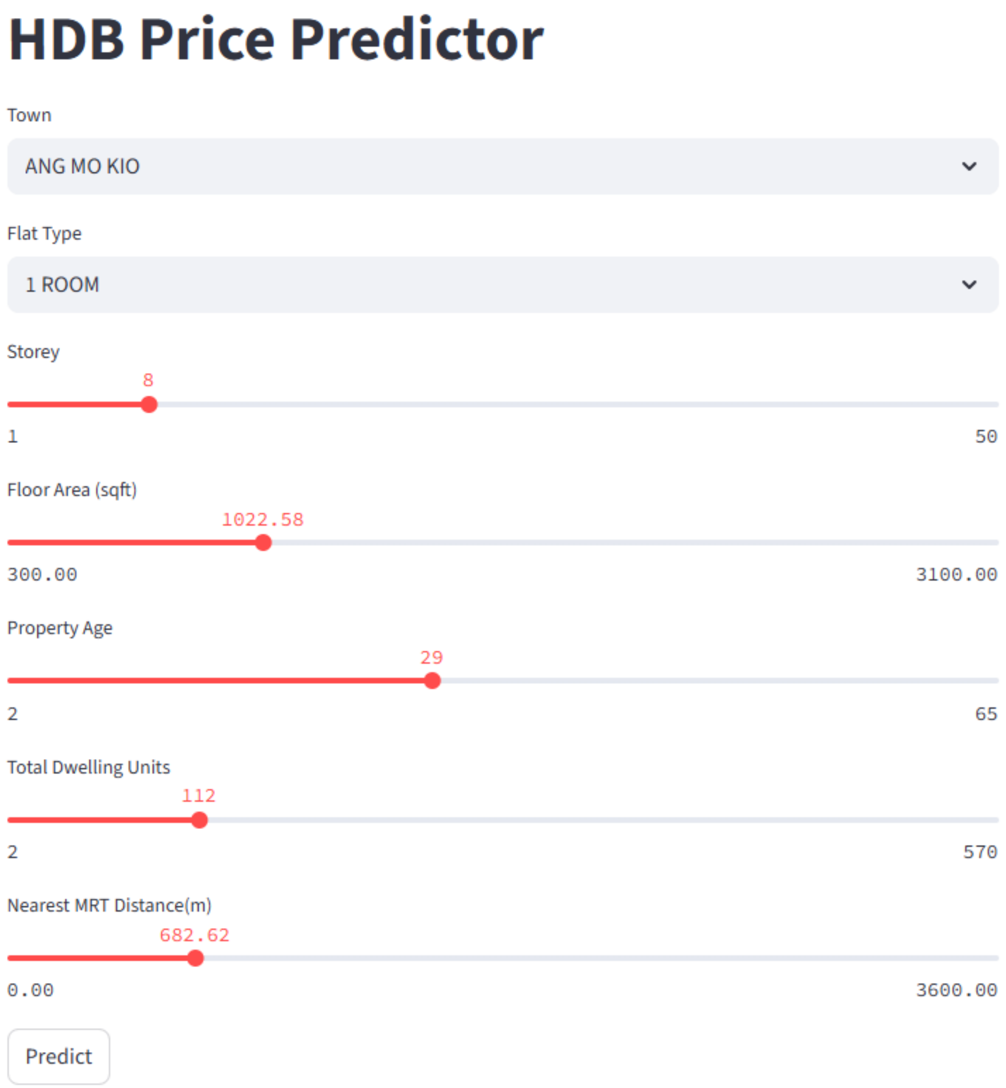
What Affects Price the Most?

- Most important features:
  - floor_area_sqft
  - town
  - mid_storey



HDB Price Predictor

- Created app using the trained Random Forest model



HDB Price Predictor

https://hdb-price-predictor-team-2.streamlit.app/

Recommendations for WOW!

- **Deploy the price predictor app**
  - User-friendly interface with dropdowns and sliders
  - Real-time predictions using trained Random Forest model
  - Easy for agents to use in the field or during client
  - meetings

- **Train agents to interpret top price drivers**
  - Example: features like floor area, town and HDB age
  - Empowers agents to justify price recommendations with data

- **Future enhancements**
  - Integrate real-time resale transaction feeds
  - Enable feedback loops to improve model accuracy

Limitations

- **The past doesn't Represent the Future:**
  - Training data reflects past market; it does not account for policy changes or future demand shifts

- **Lack of Buyer Profile Data:**
  - Without demographic information, the model cannot account for differences in preferences or affordability across buyer segments.

- **Excludes Policy Impact:**
  - Government interventions like cooling measures or BTO launches are not factored in, even though they significantly affect market behaviour.

Future Improvements

Future Improvements:

- Integrate Market Sentiment Analysis:
  - Analyze news, forums, and social media to capture public perception and emotional factors affecting pricing.

- Segment Buyers by Profile:
  - Incorporate buyer demographics such as family size, age group, or income tier to reflect different price sensitivities.

- Factor in Government Policy Changes:
  - Include variables or flags for cooling measures, BTO launches, and housing grants, which directly influence resale demand and pricing.