HTX xData Technical Test

Gorden Lim York Tee

Model Prediction Task 1

Model Used: LightGBM

Rationale:

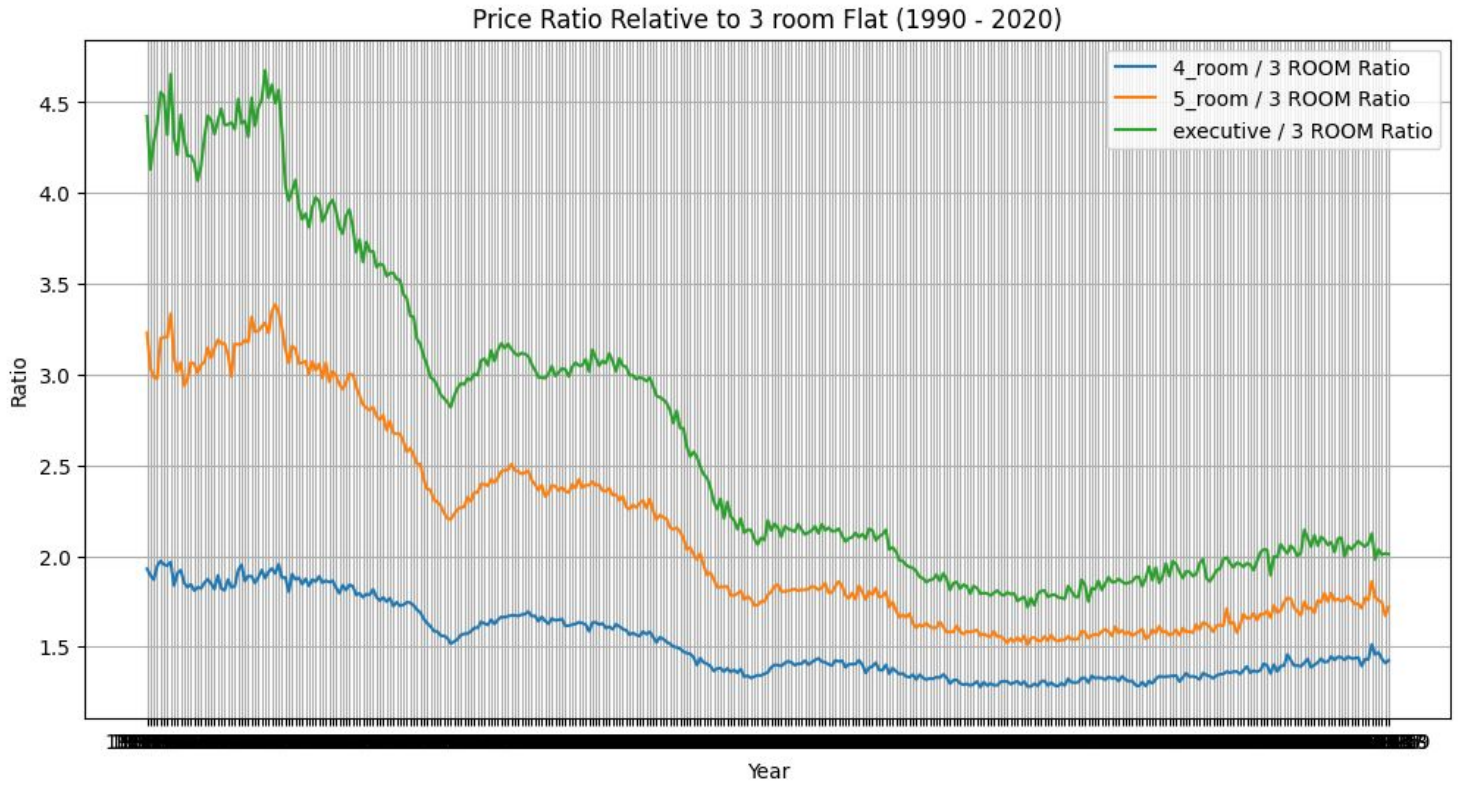
1) Explainability
2) Able to fit on CPUs
3) Relative Ease of deployment

Other Alternatives Considered: XGBoost, RandomForestRegressor (Baseline)

Rationale: LightGBM is able to ‘explain’ its features through evaluating the importance gain of each feature. It is also better suited to model complex non-linear relationships of the features. However, Random Forest Regression is used as a baseline model as it’s feature importance is easier to convey for non-technical users.

Dramatic Price Shifts over the long-term

Price Ratio has changed drastically over the years



Limiting data from 2013 onwards for model training

Based on my data exploration, I have decided to limit the data used for training the model to HDB resale listings from 2013 onwards.

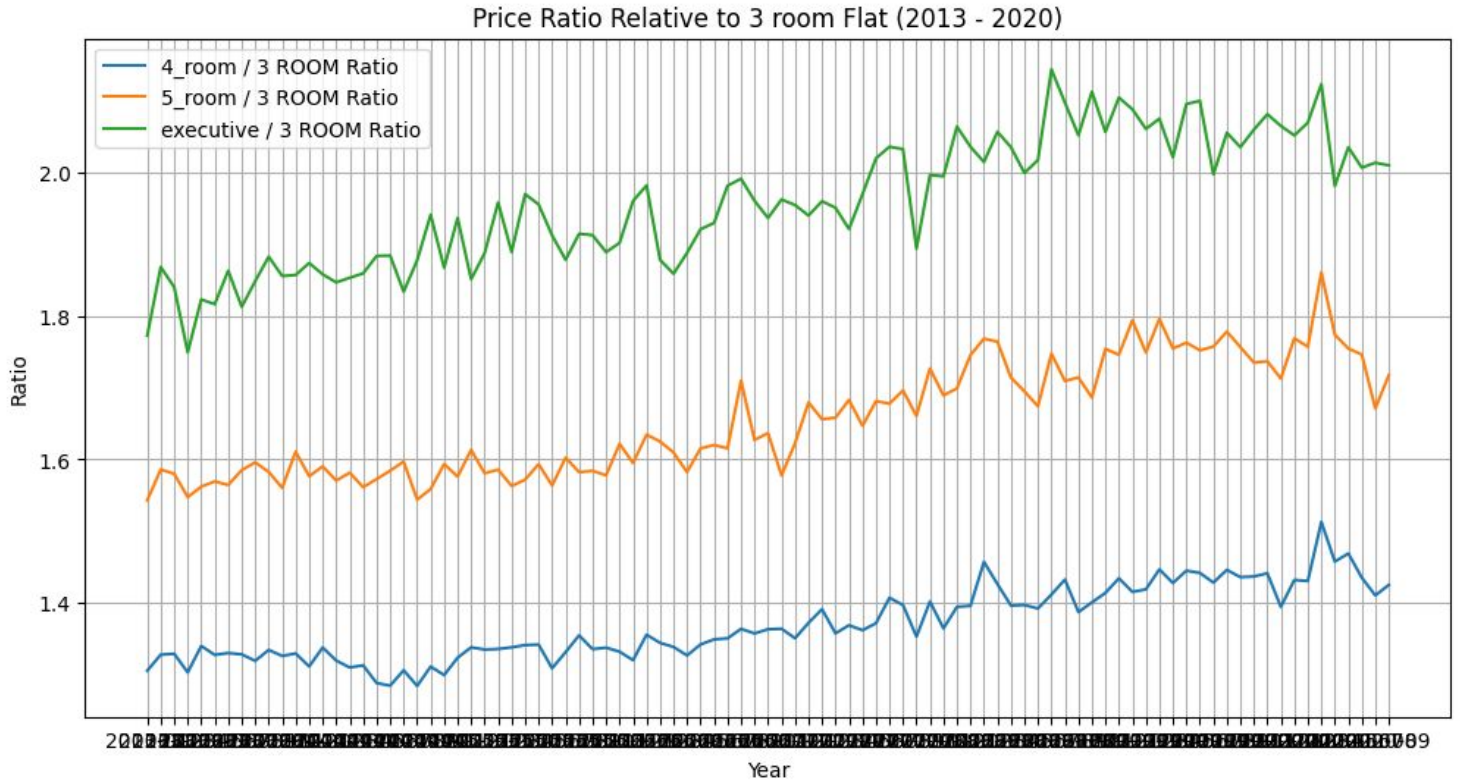
- The average resale price has changed drastically over the years, from slow and gradual growth from 1990 to 2007, to massive price spikes from 2008 to 2012, and finally a more stabilised average price from 2013 onwards. If we use data from 1990 to 2012, the model might learn a long-term upwards trend which is more muted in today's context.

- The Price Ratio Plot also shows that the relative value in flat types has changed drastically over the years. For example the "premium" of a 5-room flat over a 3-room flat was at almost 3 times the price, whereas it is closer to 1.8 times today.

Since we are trying to find ways to curb housing price inflation and not model long-term historical inflation, we will drop old data as they will likely introduce more noise and bias.

‘Stable’ average prices in recent years

Price Ratio has a slow gradual change in recent years

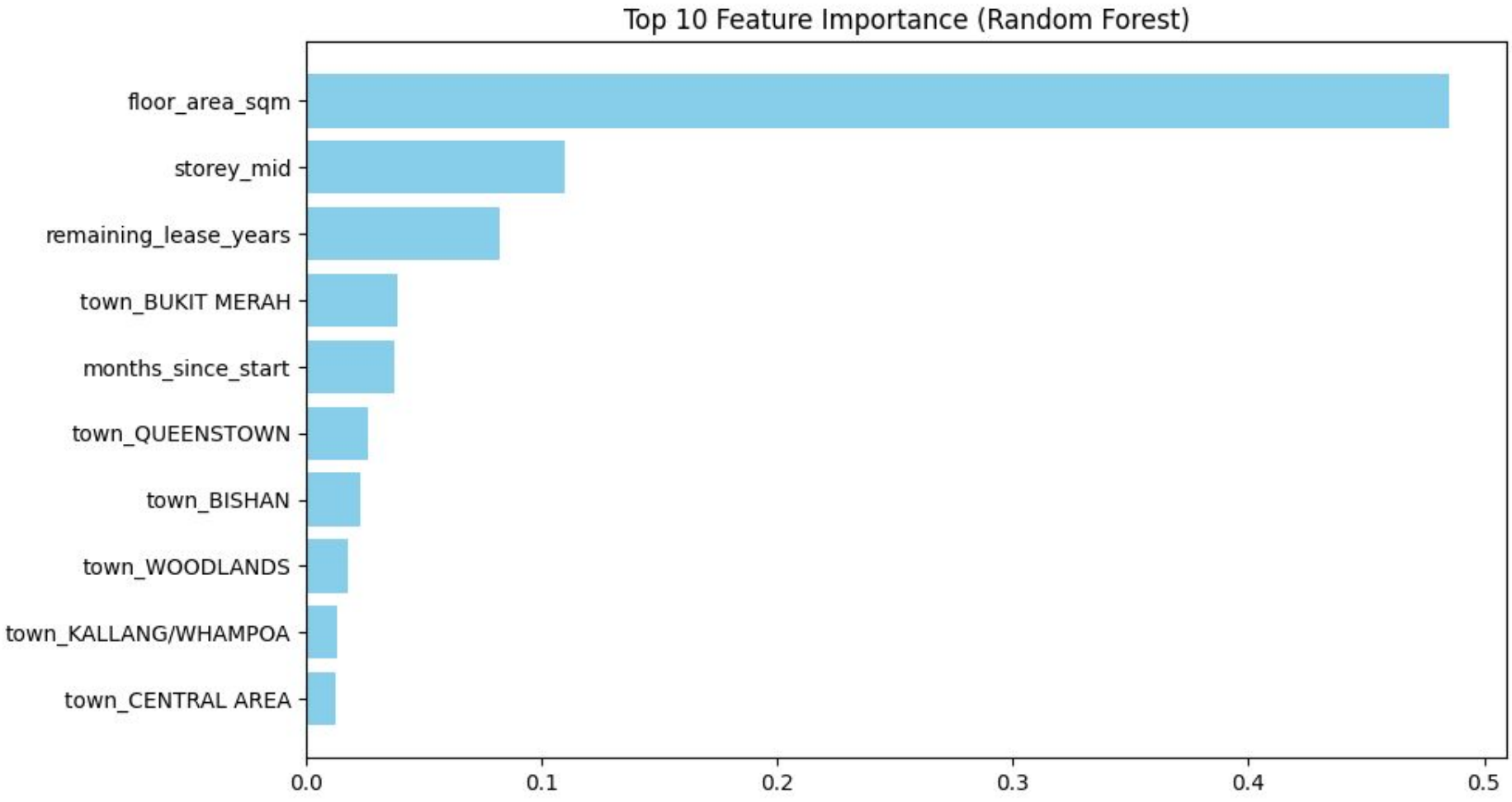


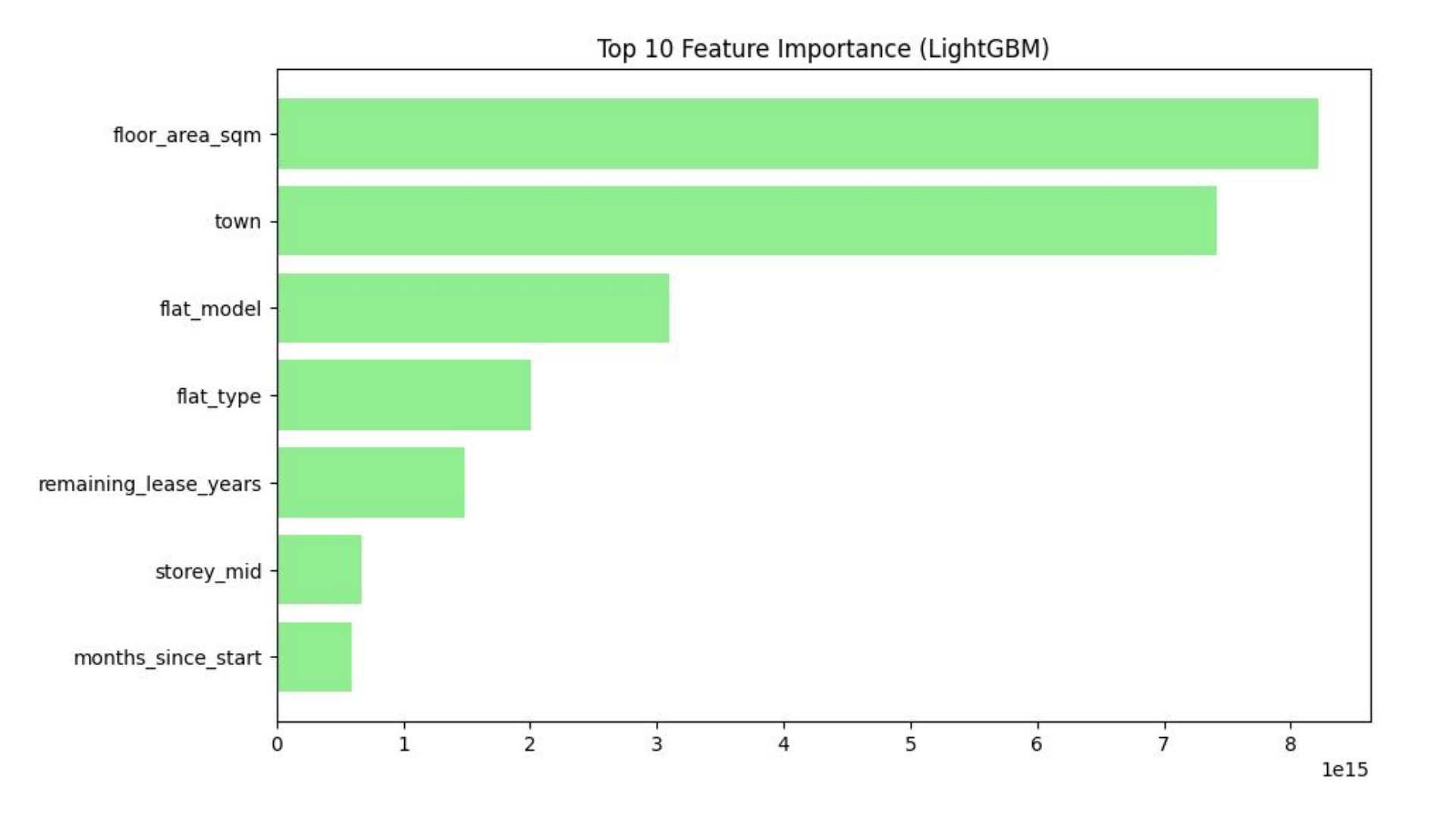
Data Preprocessing

- Feature Engineering
  - remaining_lease_years extracted using columns ‘month’ & ‘lease_commence_date’
  - storey_mid which is the average storey level extracted from text column ‘storey_range’
  - month_since_start is the number of months since the earliest month in the dataset
- Data Split
  - Since the model will likely be used to predict future resale prices, the train set and test set will be split using a temporal split to prevent leakage.
- Categorical Features
  - Flat Type will be transformed using OrdinalEncoder to preserve the ordinal nature of the size of the flat types.
  - Town and Flat Model will be one-hot encoded using OneHotEncoder

Model Results

Model	RMSE	MAE	R^2
Random Forest Regressor (Baseline)	$45,656.91	$30,838.15	0.9104
LightGBM	$41,375.18	$29,000.15	0.9266


Top 10 Feature Importance (Random Forest)
floor_area_sqm
storey_mid
remaining_lease_years
town_BUKIT MERAH
months_since_start
town_QUEENSTOWN
town_BISHAN
town_WOODLANDS
town_KALLANG/WHAMPOA
town_CENTRAL AREA
0.0
0.1
0.2
0.3
0.4
0.5


Top 10 Feature Importance (LightGBM)
floor_area_sqm
town
flat_model
flat_type
remaining_lease_years
storey_mid
months_since_start
0
1
2
3
4
5
6
7
8
1e15

Insights & Strategy to Curb Inflation

1. Address the floor area premium (Factor: floor_area_sqm)

Observation: Floor area is among single biggest determinant of price (44% contribution via RandomForestRegressor).

Strategy: Increase the supply of larger flat types (4-Room, 5-Room) in new BTO launches. This suggests a scarcity premium on space; alleviating this scarcity in BTO launches can reduce pressure on the resale market. A middle-ground strategy could be 'flexible' 4-room layouts that allow for easy conversion, meeting the demand for space without the footprint of a traditional 5-room unit.

2. Targeted cooling measures in areas with high importance (Factor: town)

Observation: Specific towns (Bukit Merah, Queenstown, Bishan) appear explicitly in the top 10 features, indicating location premiums that act independently of flat quality.

Strategy: Implement location-specific cooling measures for resale transactions in these areas to prevent them from setting a high benchmark for the rest of the market.

Insights & Strategy to Curb Inflation

3. Dynamic MOP for HDBs based on location (Factor: lease_commence_date)

Observation: Newer flats command significantly higher prices.

Strategy: Dynamically adjust the Minimum Occupation Period (MOP) for BTO builds based on location and the aforementioned feature importance to have targeted dampening of speculative demand for newer launches.

Next Steps

- Instead of an arbitrary temporal split, we should adopt a TimeSeriesSplit to create multiple folds. Allowing us to test our model on 2014, 2015 and so on. This would better prove our model has the ability to work consistently over time.
- Perform hyperparameter tuning using our validation set via GridSearch for better model performance
- Experiment dropping unimportant features to reduce noise
- Create subsets of validation sets based on flat type or town to ensure the model performs well on all towns or flat types

Model Prediction Task 2

When building an in-house model for users, the following factors and considerations come to mind. (Assumption here that 'in-house' here refers that the model will only be used internally)

1. Data Governance

As the model will most likely enriched using non-public data, it becomes more important to not expose internal datasets. Using the Task example, data such as specific buyer/seller demographics etc.

2. Latency & Scalability

Both latency and scalability become not a priority, thus we can devote fewer resources towards serving the predictions.

3. Explainability vs Accuracy

Public-facing models have a heavy emphasis on the accuracy of the model, whereas in-house models might have need for explainability to inform of current or shifting feature importances. This tasks is one such example where we might value having a more explainable model to inform us of possible inflation curbing policies.

Model Prediction Task 2

4. Model Drift & Retraining

An in-house model needs a data strategy for retraining to ensure that the data is kept up-to-date with the latest shifts in the housing market. Such a system will also need failsafes in the event that the retrained model performs worst on unseen data.

5. Batch vs Inference

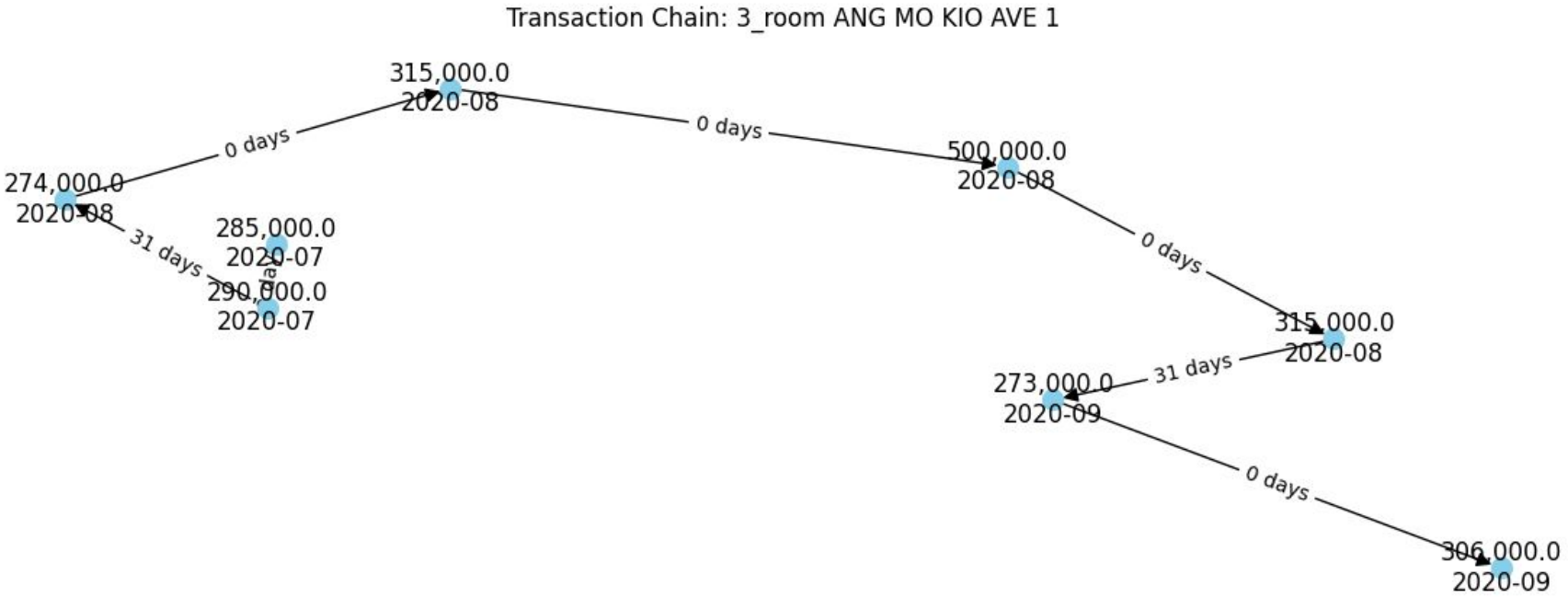
Will need to understand how the model will be used internally as well which would affect how it is deployed. If users need real-time pricing, then the model will have to be hosted on an API service to serve users (latency & scalability also have to be reassessed), if its a monthly report, then scheduling tools such as Airflow can be used to trigger the model to generate predictions.

Link Analysis Task 1

Assumption: Housing paths are path-dependent. A buyer will look at the last transacted price in the street as a benchmark. This "transaction history" link is the strongest causal relation. As such I chose to use a directed graph to model this relationship.

Node	Represents a Single Listing Attributes: Resale Price, Sale Month, Flat Type, Town, Street Name
Edges	Connects Same Street & Flat Type from one listing to another in sequence. The weight of the edge is the number of days between the previous listing to the next listing.

Sample Transaction Chain

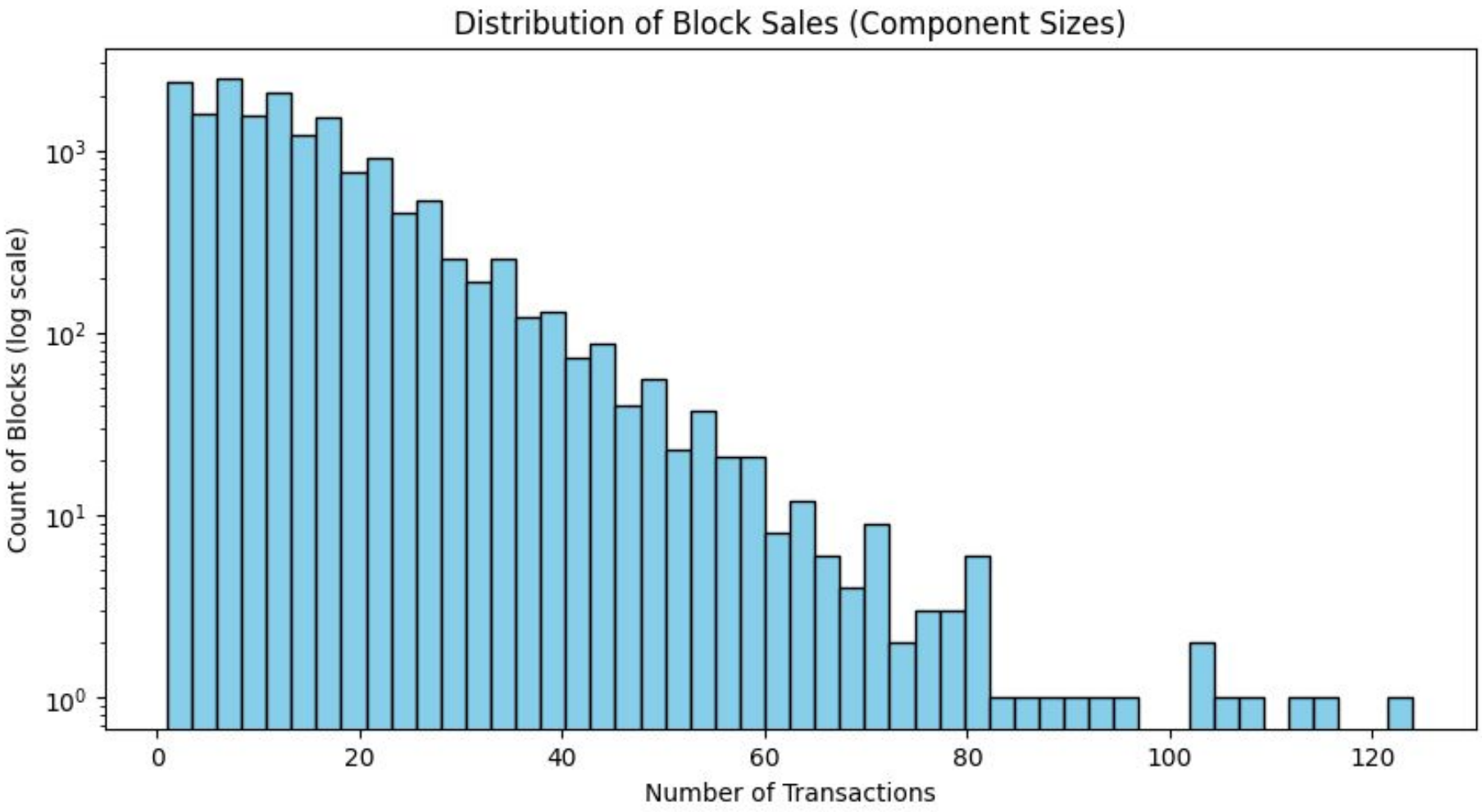


Link Analysis Task 2

Algo 1: Weakly Connected Components (WCC)

The algorithm finds the number of subgraphs by traversing (using breadth-first search) from an arbitrary node to every neighbouring node along the edges (both ways), once it is unable to flag any new nodes, it jumps to an undiscovered node and repeats the process until all the nodes are discovered.

The output shows how many streets have extremely high liquidity (very long chains) and which streets have low liquidity (short chains). The plot of the distribution shows is currently more left skewed, which indicates a long tail of highly liquid streets. These large components should show steady incremental price growth, versus the smaller components where a single high sale could cause an inflation shock.


Distribution of Block Sales (Component Sizes)
Count of Blocks (log scale)
10^3
10^2
10^1
10^0
0
20
40
60
80
100
120
Number of Transactions

Top 5 Sample of Most Liquid Streets and Flat Types:

Street: ANG MO KIO AVE 10, Flat Type: 3_room, Transactions: 1169

Street: YISHUN RING RD, Flat Type: 4_room, Transactions: 1091

Street: ANG MO KIO AVE 3, Flat Type: 3_room, Transactions: 975

Street: YISHUN RING RD, Flat Type: 3_room, Transactions: 893

Street: FERNVALE RD, Flat Type: 4_room, Transactions: 876

Top 5 Sample of Least Liquid Streets and Flat Types:

Street: TAMPINES ST 86, Flat Type: 5_room, Transactions: 1

Street: MOH GUAN TER, Flat Type: 4_room, Transactions: 1

Street: HENDERSON RD, Flat Type: 3_room, Transactions: 1

Street: HO CHING RD, Flat Type: 5_room, Transactions: 1

Street: TAMPINES ST 86, Flat Type: 4_room, Transactions: 1

Link Analysis Task 2

Algo 2: Node Assortativity

Assortativity measures the correlation between a node's attributes and its neighbors' attributes. Specifically, we calculate the Price Correlation between a listing and its predecessor.

It does this by going through each edge and forming a pair of its attributes of the connecting nodes (price_A, price_B) for example from Node A to node B. It then calculates the correlation from these values.

The result is a high price assortativity correlation of 0.9270, this confirms that there is strong price momentum, it statistically proves that the price of a unit is heavily influenced to the previous sale in that same block. If the correlation is high, record-breaking sales are dangerous because they immediately pull up the baseline for the entire chain. If the correlation were low, it would imply that other unit attributes (floor, renovation) matter more than the building's recent history.

Conversely, a policy intervention that lowers price will also cascade to future purchases.