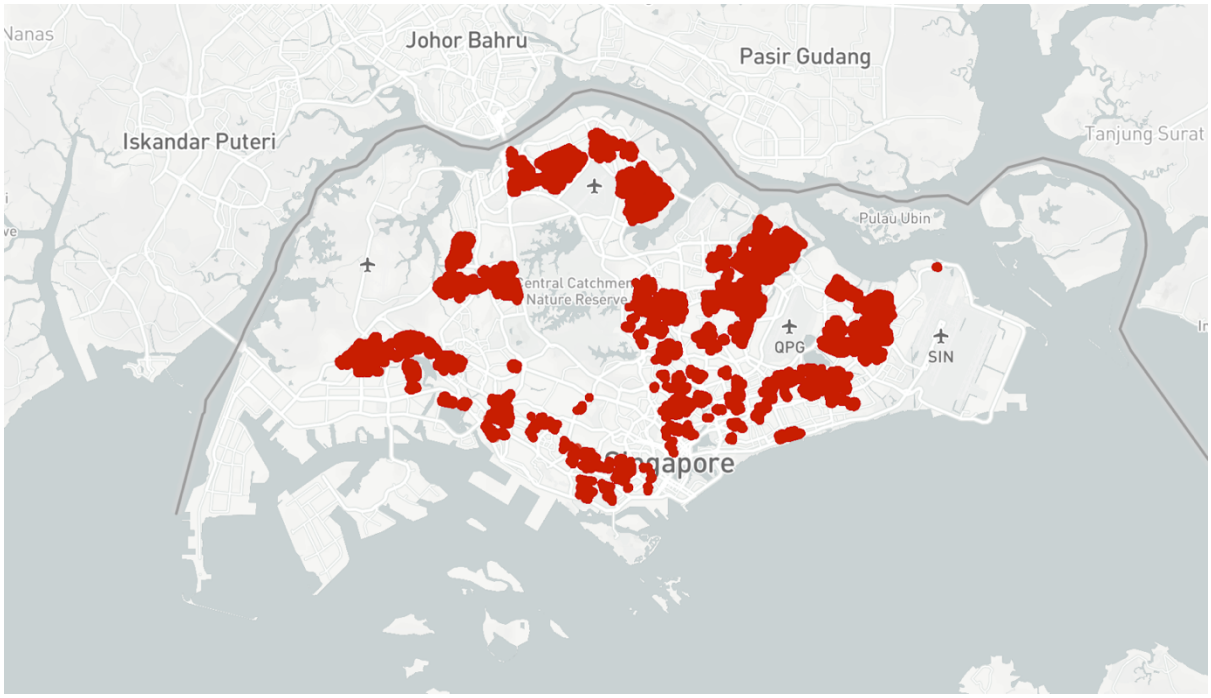
HDB Resales and Rentals
Price Prediction & Analysis



Final Report

Submitted on: 17 April 2024
Submitted for: IT5006 Fundamentals of data analytics

Team members:

s/n	Name	Student number
1	Hew Sock Fang	A0276573R
2	Lee Shu Ling Charlene	A0286974E
3	Huang Zichen	A0279979R
4	Seah Ee Wei	A0276567L

Section 1: Dataset and pre-processing

1. **Data cleaning**. The datasets provided were already mostly cleaned. We took this further by ensuring data types were appropriate and informative (e.g. converting month / rent approval date to datetime, including a column for year sold), filling in missing values (remaining lease for flats) and standardising some nomenclature (e.g. "Multi Generation" and "Multi-Generation" flats in flat models).

2. **Developing features**. For the purpose of exploratory data analysis, we developed two features from the existing resale dataset. First, price per square metre (resale price divided by floor area). This allowed us to assess price trends without the impact of the floor area, since larger flats tend to command higher prices. Second, we averaged the storey height from the storey range and binned it into four bins (1–5, 6–10, 11–20, 20+) as the existing ranges were overlapping (5 –10, 6–10) and inconsistent in range. As this was only for exploratory analysis, we chose more interpretable ranges for visualisation.

3. **Adding features**. Both datasets were enriched with new features from publicly available datasets. We ensured that the feature values were renamed for consistency e.g. "Holland Close" vs "Holland Cl" to ensure accurate merging.

   a. Geographical data (postal code, longitude and latitude) from the OneMap API. This allowed us to assess geographical distributions of the data and calculate nearest distances to train stations / amenities. Only train stations existing at the time of sale were included in the calculation.

   b. Distances to the nearest train station, primary school, mall, hawker centre, and supermarket. The list of transport and amenities was obtained from publicly available data from government bodies on data.gov.sg. Thereafter, we called the OneMap API for location data and ran a k-nearest neighbours algorithm to find the nearest facilities for each address.

   c. Maximum floor level of each building (from HDB). Almost all buildings had maximum floor levels provided; for those which did not, we filled in those values with the median maximum floor level for buildings on the same street.

Section 2: Exploratory data analysis insights

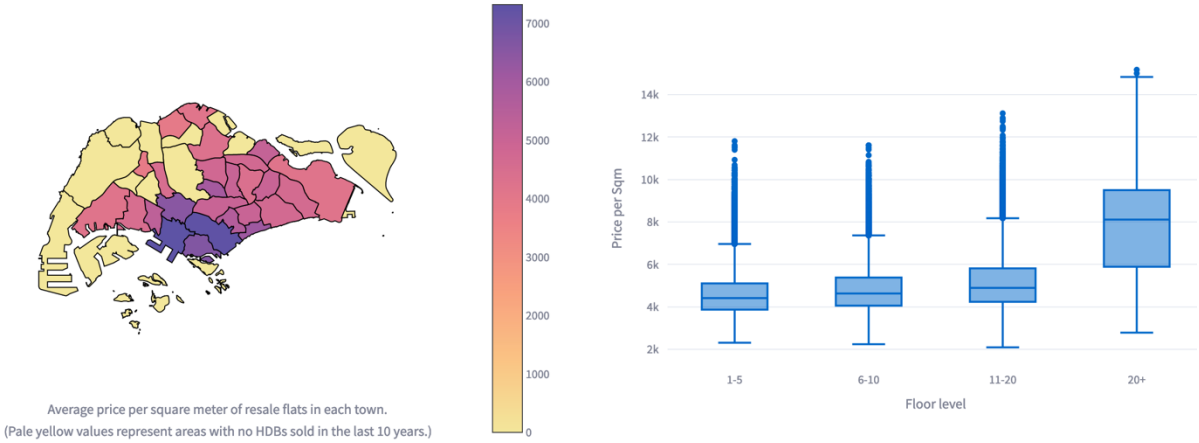
4. We explored the data using different data visualisation techniques. We focused on the past ten years (2014 to 2023) for the resale dataset to get more current insights and all years from the rental dataset.

5. **Resale prices are trending up, especially during and after COVID-19.** As can be seen from the charts below, while resale prices generally trended up in the last ten years, there was a marked increase in resale prices during and after COVID-19. In fact, the median price per square metre (psm) had actually decreased in the years leading up

to COVID-19 before markedly rising from mid-2020 onwards, shortly after the first circuit-breaker period in April 2020. There was an average 11% decrease in median psm in 2019 pre-COVID compared to an average increase of 22.2% in 2023 post-COVID (baseline 2014). Delays in BTO housing construction and demand for larger spaces due to work-from-home policies, may have driven this increase. [1] [2]

6. As seen below, the number of transactions has also increased, peaking in 2021. While the top three towns by price psm are located around the central areas, the top three towns by transactions are around the northeast (Sengkang), west (Jurong West), north (Woodlands) areas of Singapore. As more flats are built in newer towns and reach their minimum occupation period, we can expect more transactions in those areas. [3]

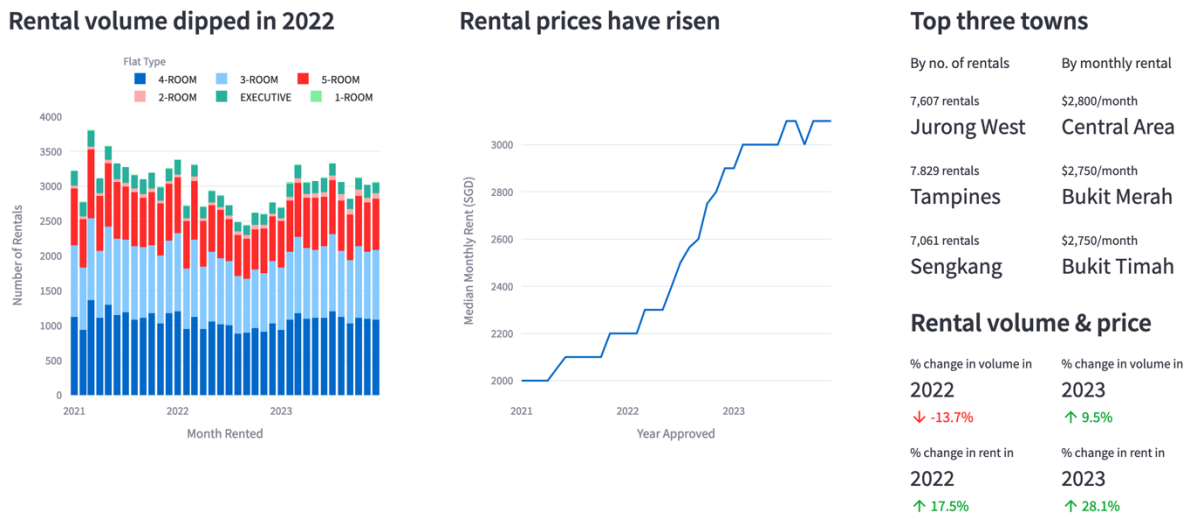
7. **Central areas and flats on a higher storey tend to command higher resale prices**. This was an unsurprising insight, given that the centrality of property location, with its accessibility to the financial, government and shopping districts, has always commanded higher prices. Flats on a higher storey may be more popular because they are also less likely to encounter pests and with unobstructed views.



Average price per square meter of resale flats in each town.
(Pale yellow values represent areas with no HDBs sold in the last 10 years.)

8. For the rental dataset, we noted that the rental volume took a brief dip in 2022 but rental prices steadily increased. The median monthly rent increased 17.5% in 2022 and

28.1% in 2023, compared to a 13.7% decrease in rental volume in 2022 and 9.5% increase in 2023. Similar to the resales data, the highest prices were observed in the central areas and the most rental transactions were concluded in the satellite towns.



9. **Weak correlations between distances to transport, schools and amenities** (defined as the minimum distance to a supermarket, hawker centre or mall). The correlations for the resale dataset were -0.22, 0.12 and -0.19 respectively. For the rental dataset, it was even more negligible at -0.068, -0.006 and -0.008 respectively.

10. This was fairly surprising. Good urban planning may explain this. We noticed that almost all the flats in Singapore were located within 2km to the nearest MRT station or primary school, and within 1km to the nearest amenity. This is also in line with LTA's 2040 Master Plan to create 20-minute towns and 45-minute cities. [3] Another reason may be that we did not rank the primary schools or malls near each flat due to lack of an objective metric. Popular primary schools are likely to be more popular with parents looking to take advantage of the distance requirements for primary school enrolment and may have correlated better with price.

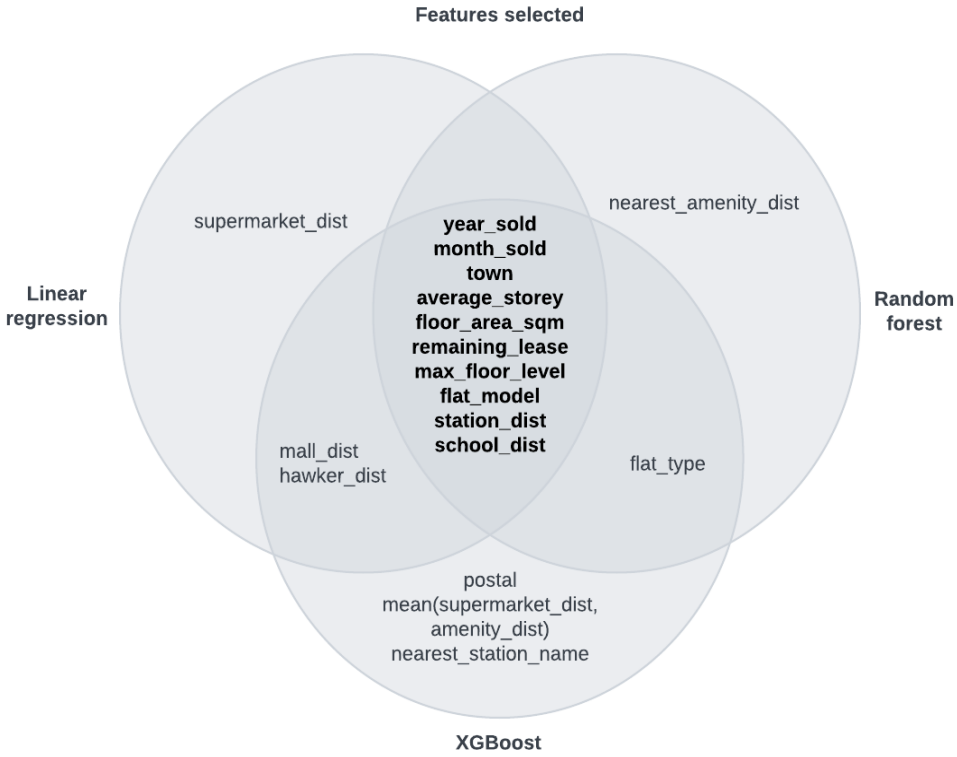
Section 3: Approach to price prediction

11. For resales, we ran three models: ordinary least squares linear regression, random forest, and XGBoost. For rentals, we used linear regression and random forest regression. To gauge performance, we used R^2 (explained variance), root mean squared error (RMSE) and mean absolute percentage error (MAPE). Ultimately, we chose random forest regression for its good performance and interpretability.

Resale dataset

12. **Choice of features**. Based on our exploratory data analysis, we selected features that were likely predictive, namely those relating to time of sale, storey height, town, remaining lease, square area footage, and the distances to nearest amenities. We then

tuned the models, adjusting and encoding features based on the model's performances. For all models, we found that the features in bold were consistently predictive:



13. **Overall performance**. With a random train-test split, all models performed fairly well with R^2 scores of at least 0.85 and RMSE of around \$20k–\$65k. However, random forest and XGBoost performed significantly better than linear regression.

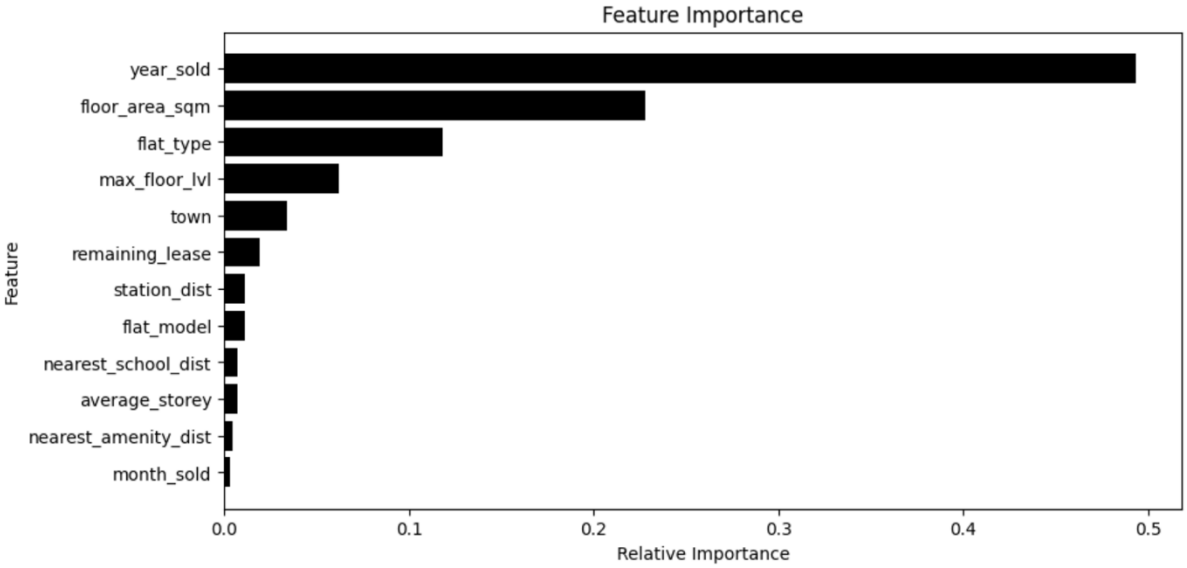
Model	Train		Test	
Metric	R^2	RMSE	R^2	RMSE
Linear regression	0.85	\$65,031	0.85	\$65,122
Random forest regression	0.98	\$22,380	0.98	\$24,916
XGBoost	0.98	\$21,396	0.98	\$21,905

14. **Model 1: Ordinary least squares linear regression.** We first tried linear regression. Predictors such as month sold, town, flat type and flat model were encoded as categorical variables to avoid introducing biases. This is especially so for month sold, as the relationship between sales/rentals and month appeared seasonal rather than linear. We used residual plots for each predictor to assess potential linear relationships to tune the appropriate variables. However, the plots indicated heteroscedasticity, violating the homoscedasticity assumption of linear regression. To address this, we applied exponential and squared transformations to the data as indicated by the residual patterns. This led to slight improvements but we ultimately needed more sophisticated models to capture the non-linear relationships within the data.

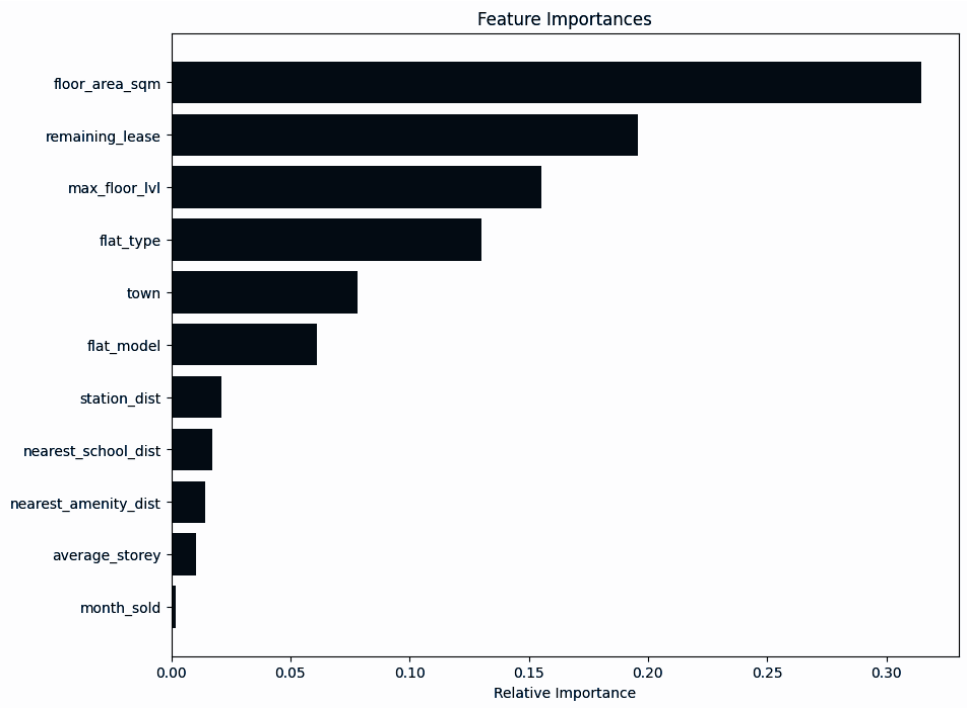
15. **Model 2: XGBoost.** We optimised this model by focusing on location features. First, we carried out a k means clustering (n=5) based on latitude and longitude to add a

location cluster to each data point. Second, we target encoded the town and the nearest station. These measures introduced spatial context and significantly boosted our model's accuracy. The inclusion of XGBoost's regularization and pruning not only mitigated overfitting but also optimized the model's efficiency, aligning with our project's goals to capture and interpret the complex dynamics influencing prices effectively.

16. **Final chosen model: Random Forest regression**. For random forest regression, we one-hot encoded the categorical features, ordinal encoded 'year_sold' and carried out standardisation and normalisation of the numerical features. Although the use of 'postal' gave a better prediction than 'town', the improvement was insignificant. Given that 'town' would be a more convenient input for app users, we adopted 'town' instead. We were limited by our computing resources and had to use a systematic approach to tune the hyperparameters. That said, the model provided good overall performance and did not require extensive manipulation/transformation of the data values, making it more interpretable than XGBoost's. The overall test RMSE was around $25k, translating to a mean absolute percentage error of around 6.4%. We thus chose it as the basis for our app prediction. From the feature importance values below, we see that the year sold, floor area, flat type, and maximum floor level were most predictive.



17. Given that the values of 'year_sold' are unique and have direct impact on flat prices, we explored the redistribution of feature importance and model's performance without this dominant predictor. The model retained its performance – the decrease in R^2 was marginal, from 0.98 to 0.95. As seen below, the top 5 features after 'year_sold' remained consistent. Notably, the remaining lease and maximum floor level of the block gained relative importance. Thus, other than the year sold, the remaining chosen features could still sufficiently explain the variance in the data.



Rental dataset

18. The approach to the rental dataset was similar to the resale dataset. We trained two models, linear regression and random forest regression, and systematically calibrated the features selected and hyperparameters based on model performance. The best performing model was ultimately random forest regression with an overall R^2 value of 0.5 and a mean absolute percentage error of around 16.1%. The most predictive factors were year sold, flat type, and maximum floor level.

19. Ultimately, the models were unable to generate results that were as good as those from the resale dataset. We theorise a few reasons for this:

a. The dataset was relatively small, with only 108,898 entries over a 3-year period from 2021 to 2023. This contrasts with the vast amount of data we had for resales data. The dataset is likely too small to effectively train the model, hence leading to insufficient learning and generalization capabilities and made it difficult to provide good predictions.

b. The rental statistics are self-reported by flat owners and not verified by HDB [5]. The data may not be factually representative of the rental rates e.g. there were entries that clearly deviates from the norm, which seems to reflect prices for a single room instead of the entire unit. However, HDB officer was not responsive to our query and we were unable to confirm via any other data sources. Notwithstanding, we noted that the ranking of feature importances was consistent with that of the resale dataset.

Section 4: Reflections and conclusions

20. In this project, we analysed the rental and resale data and developed a web application presenting our insights from exploratory data analysis with a price prediction component. To enrich the existing HDB rental and resale datasets, we added geographical data, data on nearby transport and amenities, and building characteristics to each flat. For the price prediction, we developed linear regression, random forest regression and XGBoost models and eventually selected the random forest regression models for both datasets due to its relatively good performance and interpretability. Over the course of the project, we learnt the importance of ensuring good quality data and the benefits of training various models to observe the differences in performance.

21. It is worth noting that the intrinsic property features and location could already provide substantial predictive power. The addition of features not only did not significantly improve the model's predictive power, it also burdened the computational efficiency. This is commonly known as the 'curse of dimensionality', The balance between model complexity, predictive power, and computational efficiency is a classic trade-off in machine learning. As long as there are sufficient and reliable data points, simpler models with fewer features can be easier to interpret, quicker to train and nearly as effective as models with a large number of features – underlining the importance of feature selection and the understanding of the domain when building predictive models.

References

[1] I. Gafoor, "Commentary: There is greater demand for bigger homes in a pandemic but will it last?," CNA, 22 Jun 2021. [Online]. Available: https://www.channelnewsasia.com/commentary/bigger-flats-record-price-hdb-private-pandemic-whampoa-terrace-1935776. [Accessed 9 Apr 2024].

[2] S.-A. Tan, " HDB resale prices hit new record high after rising 8.9% this year: Flash data," The Straits Times, 1 Oct 2021. [Online]. Available: https://www.straitstimes.com/singapore/housing/hdb-resale-prices-rise-27-in-q3-flash-data. [Accessed 9 Apr 2024].

[3] C. H. Min, " No longer 'ulu': Why resale prices in Sembawang have risen faster than any other neighbourhood in Singapore," Channel News Asia, 22 Feb 2024. [Online]. Available: https://www.channelnewsasia.com/singapore/sembawang-hdb-resale-flats-prices-towns-4135751. [Accessed 7 Apr 2024].

[4] "Land Transport Master Plan 2040," Land Transport Authority, [Online]. Available: https://www.lta.gov.sg/content/ltagov/en/who_we_are/our_work/land_transport_master_plan_2040.html. [Accessed 9 Apr 2024].

[5] Data.gov.sg, Renting Out of Flats 2024 (CSV) [Online]. Available: https://beta.data.gov.sg/datasets/d_c9f57187485a850908655db0e8cfe651/view. [Accessed 15 Apr 2024].