EBA5002: Graduate Certificate in Business Analytics Practice Final Report

Property For Singaporean (PFS) - HDB Resale Price Prediction & Sentiment Analysis Of HDB's Cooling Measures

Team We R

Team Members

Alex Koh Jun Leng

Cai Shiying

Edwin Goh Duo Yao

Kerin Guo Chun'en

Zheng Xiao

Table of Contents

1 Background

Property For Singaporeans (PFS) is a fictional real estate agency that provides clients with access to information regarding various properties in Singapore. It offers a wide range of services including providing property listings, up-to-date property news, market insights and analysis with the ultimate goal of helping Singaporeans make informed decisions when buying and selling their homes. Being a pioneer in the Singapore real estate space, PFS has managed to gain a large percentage of the local market share since its inauguration in 2016. However, since then, PFS has been facing increased competition from rival property agencies who are more savvy in navigating the online social media landscape and who have leveraged social media as an effective marketing and branding tool to build and grow their market share. Wishing to stay ahead of the burgeoning competition, PFS wishes to incorporate more business analytics practices and methodologies into its business to provide better pricing recommendations and curated insights for its prospective and existing clients and to compete more effectively against its competitors.

2 Business Motivations

As Housing & Development Board (HDB) resale flats form the bulk majority of the housing market in Singapore, with 77.9% of the residential dwellings in Singapore being HDB flats, PFS wishes to start by focusing on the HDB resale market. The ability to accurately and efficiently perform property valuations, and forecast future pricing trends in response to changes in the property market and the macroeconomic environment is key to the business and PFS currently lacks the technical expertise to do so. In addition, PFS needs to stay abreast of current housing market sentiments so that it can respond swiftly to any changes in the housing property markets. Due to a tight budget, PFS is unable to conduct large-scale surveys and interviews to better understand public sentiments towards the current property markets and property cooling measures implemented by the government. Finally, PFS wishes to better understand its competitors' social media activity in order to compete more effectively in the social media landscape against its competitors. As such, our team of data analytics consultants has been hired to generate a forecast of HDB resale prices in the coming years, explore public sentiments towards the latest cooling measures introduced by the Singapore government on 29th September 2022 to moderate demand and ensure affordability of HDB resale flats, as well as to conduct competitor social media analysis.

3 Objectives

3.1 Business Objectives

PFS's main business model is to provide clients with access to information about the property market in Singapore, and there has been a lot of buzz in the community since a round of cooling measures was introduced by the Singapore government back in September 2022. As such, the business objectives of the project are: (1) to understand the general sentiments and concerns of Singaporeans (who may become PFS's potential clients) on the housing markets in light of the cooling measures implemented, as well as (2) to create a model to predict and forecast the trend of future prices of HDB resale flats in the coming few years and (3) to conduct competitor social media analysis to gather actionable insights that can be applied to improve PFS's social media presence.

This will greatly deepen PFS's understanding of its competitors' social media strategy, overall housing market sentiments and primary concerns with regards to the cooling measures, as well as the current state of HDB resale market, where the trend in resale flats' prices will be an indicator of demand within the market. Armed with this comprehensive and holistic set of insights, PFS can build its social media strategy, further tailor its service offerings and provide bespoke advice to assist prospective and existing clients in their decisions on whether they should purchase their new homes from the HDB resale market or other property markets.

3.1.1 Technical Objectives

The first technical objective is to create a predictive model that can be used as a forecast for prices of HDB resale flats in the future. We will first conduct a time series analysis to gain an understanding of the historical trend in prices of HDB resale flats sold in the past and identify potential time-related factors that may affect the prices of resale flats. We will then conduct feature engineering on the historical resale flat transaction dataset obtained from data.gov.sg to extract the most relevant features that could potentially affect the prices of HDB resale flats. Finally, we will pass these features as inputs into various machine learning models and evaluate which model is most effective in predicting future prices and factors influencing HDB resale flats.

The second technical objective is to conduct analysis of text data retrieved from popular social media platforms in Singapore with regards to the latest round of cooling measures implemented by the Singapore government. We will be scraping and cleaning at least five thousand social media posts using text pre-processing techniques, such as stopword removal and lemmatization, prior to analysis. Common sentiments and concerns will then be extracted from these posts using Natural Language Processing (NLP) techniques such as clustering and topic modelling.

The third technical objective is to conduct analysis of competitor social media posts. We will be scraping and cleaning at least five hundred posts using text pre-processing techniques. The most frequently posted topics will then be extracted using topic modelling.

4 Project Scope & Design

4.1 Project Data

4.1.1 Approved HDB Resale Transactions (1990 - 2023) Dataset

To predict future prices of HDB resale flats, it is essential to first understand the historical trends of HDB resale flat prices over the years and the factors influencing these prices. Hence, HDB's Resale Prices dataset published on data.gov.sg, was the most suitable candidate in fulfilling these two goals. The dataset contains a total of 892,418 approved HDB resale transactions dating from January 1990 to February 2023. It was noted that transactions from January 1990 to December 2014 contained only 10 columns, while an additional "*remaining_lease*" column was included to the transactions from January 2015 onwards. This is an important artefact that highlights the differences in how HDB has recorded resale transactions over the years and hence needs to be considered during the data preparation phase prior to modelling. The columns within the dataset are defined in the data dictionary below:

Column	Description	Data Type	Sample Value(s)
month	Year and Month when resale flat was sold	Date	2017-01
town	The residential region where the resale flat is located in Singapore	String	ANG MO KIO
flat_type	The housing type of resale flat	String	2 ROOM
block	The block number of resale flat	Integer	406
street_name	The name of street where resale flat is located	String	ANG MO KIO AVE 10

storey_range	The range of floors where resale flat unit is located	String	10 TO 12
floor_area_sqm	The area of the resale flat unit in square metres	Integer	44
flat_model	The housing model of the resale flat unit	String	Improved
lease_commence_date	The year when lease started for resale flat unit	Date	1979
remaining_lease	The number of years and months left for lease of resale flat unit	String	61 years 04 months
resale_price	The price at which resale flat unit was sold at	Integer	232000

Table 4.1.1: Data Dictionary of The Approved HDB Resale Transactions Dataset

4.1.2 Sentiment Analysis Dataset

To understand the general sentiments and concerns of Singaporeans with regards to property cooling measures, two online forums were selected as sources for data collection: HardWareZone and Reddit. These two platforms were chosen as they are discussion-based and comment-heavy platforms with active daily users, large membership bases, and have been identified to host the most relevant social conversations amongst Singaporeans. Data is collected over a period of 7 months between 29th September 2022 (announcement of 2022 cooling measure) and 25th April 2023 (before announcement of 2023 cooling measure).

Data Source	Reason for Choosing
HardWareZone	Contains a popular Singapore-based online forum for Singaporeans to discuss and share their thoughts on various topics, with over 700, 000 members
Reddit	Social news aggregation and discussion website with large number of user base in Singapore (e.g. 622, 000 members on r/singapore)

Table 4.1.2: Data Sources For Sentiment Analysis

4.1.3 HardWareZone

Python package BeautifulSoup was used to scrape data from manually identified forum threads on HardWareZone. Utilising HardWareZone's internal search engine and Google's search engine, a list of 83 relevant threads were manually identified with the search keyword "cooling measures". Only threads that contain at least 3 comments were included. Using the web scraper, a total of 9,849 comments were extracted from the shortlisted HardWareZone threads. Data fields that were retrieved includes thread title, comments, comment author, as well as comment date, which are defined in the data dictionary below:

Variable	Description	Data Type	Sample Value(s)
Thread Title	The title of thread	String	This is a title.
Comment Author	User name of commenter	String	John Doe
Comment	The content of the comment	String	This is a comment.
Comment Time	The time of comment	Datetime	00:00:00
Comment Date	The date of comment	Datetime	01/01/2023

Table 4.1.3: Data Dictionary of Comments Extracted From Shortlisted HardWareZone Threads

	Thread Title	Comment Author	Comment	Comment Time	Comment Date
1	HDB BTO oversupply not the answer to home affordability: Desmond Lee | HardwareZone Forums	fascist	Another stupid inconsistent policy. So manufacture an under supply and delay marriage and family planning. Then complain not enough babies - then import FTs.	9:29 AM	2023-02-08
2	HDB BTO oversupply not the answer to home affordability: Desmond Lee | HardwareZone Forums	wongkc	In short... they have no answer to the runaway property prices... letting it run auto mode is the best solution... and doing reactive actions with tiny bits of cooling measures along the way...	9:30 AM	2023-02-08
3	HDB BTO oversupply not the answer to home affordability: Desmond Lee | HardwareZone Forums	ng min teck	Should tie with ns those kind of thing seem actually both like is the need and they can't measure in price. After ns , minus land price and free 2 rm flat, if want 2 rm above need top up. Also like bto need wait 5 years.	9:35 AM	2023-02-08
4	HDB BTO oversupply not the answer to home affordability: Desmond Lee | HardwareZone Forums	abbakonghee	Sinkies want this and now kpkb? Well deserved	9:37 AM	2023-02-08
5	HDB BTO oversupply not the answer to home affordability: Desmond Lee | HardwareZone Forums	XiaoJinLing	All finding excuses push here n there... all same pattern... ok...	9:37 AM	2023-02-08

Figure 4-1: Example of Data Retrieved From HardWareZone

4.1.4 Reddit

In total, 933 comments were collected from Reddit using the PRAW module (Python Reddit API Wrapper). Search query term 'cooling measure' was used to obtain a list of threads within subreddits 'r/singapore', 'r/SingaporeRaw' and 'r/singaporefi. Irrelevant threads from the search result were manually identified and filtered off. Details of comments posted in shortlisted threads were then extracted and the attributes are defined in the data dictionary on the following page:

Variable	Description	Data Type	Sample Value(s)
comment_id	ID of the comment	String	abc123
comment_parent_id	ID of the parent comment	String	t1_abc123
comment_username	User name of commenter	String	John Doe
comment_body	The content of the comment	String	This is a comment.
comment_date	The date of comment	Datetime	01/01/2023
comment_upvotes	The number of upvotes for the comment	Integer	1
comment_post_id	The submission ID that the comment belongs to	String	t1_abc123

Table 4.1.4: Data Dictionary of Commented From Shortlisted Reddit Threads

	comment_id	comment_parent_id	comment_username	comment_body	comment_date	comment_upvotes	comment_post_id
0	jhssotz	t3_12zn0pa	AutoModerator	**The downvote is not a disagree button.** Please help to upvote articles that you want to see more discussion on, and downvote those that you feel has little value on the sub.\n\n\n*I am a bot, and this action was performed automatically. Please [contact the moderators of this subreddit](/message/compose/?to=/r/singapore) if you have any questions or concerns.*	26-04-2023	1	t3_12zn0pa
1	jhsuu3m	t3_12zn0pa	Elzedhaitch	This will really show if foreigners are the ones buying the houses. 60% is insane. You will almost never profit...\n\nYou will take decades to even break even now.	26-04-2023	474	t3_12zn0pa
2	jhswwk6	t3_12zn0pa	tougan-481	30 to 60% for foreigners, that's really significant. Wonder how much difference this will make for the whole housing situation	26-04-2023	165	t3_12zn0pa
3	jhswjkj	t3_12zn0pa	yang_	In other news, ICA sees a spike in applications for PR status..	26-04-2023	265	t3_12zn0pa
4	jht5c7v	t3_12zn0pa	buttermilkcrispy	"Easiest" way to get 0% ABSD as a foreigner is getting PR (or citizenship) in Switzerland, Norway, Iceland or Liechtenstein. You're exempted from SG ABSD then for the first property then.\n\nUS citizenship is also 0%, but not that easy as they don't allow US PR's. Unless you wanna give birth to a baby there and buy the property on the babies name lol. US automatically gives citizenship to anyone born there.\n\nAlternatively 5% ABSD for SG PR isn't too bad.\nHaving a Singaporean spouse is another option for 0%.	26-04-2023	60	t3_12zn0pa

Figure 4-2: Example of Data Retrieved From Reddit

4.1.5 Competitor Analysis Dataset

Data scraping of 696 posts was carried out from our client's two main competitor's Facebook pages over a period of 35 weeks. The two competitors chosen were PropertyGuruSg & 99dotco as they have a similar business model and are the two major players in the real estate industry with comparable market share to our client. Facebook was chosen as the social media platform to compete on due to its high number of daily active users and tightly-knitted communities, robust advertising options, and high number of businesses already doing commerce on their platform.

The data was scraped using the 'facebook-scraper' library, and includes many details such as datetime, post content, the image url, number of likes and reactions etc. However, not all columns were utilised for the purpose of our project.

4.2 Critical Success Factors

In order to fulfil PFS's overarching business objective of providing tailored services to clients and standing out from the competition, a critical success factor is the ability to translate the results of our proposed models into insights and recommendations for the business.

For instance, sentiment analysis of social media posts might reveal that one of the common concerns regarding the cooling measures is the prices of existing HDB properties amongst current homeowners who are looking to purchase a second property. However, the in-depth domain knowledge possessed by real estate sales experts are still required to better assuage their worries and advise them on next steps. Furthermore, to gain an edge over their competition with these insights, PFS will need to show potential clients that the business understands their concerns, especially through their marketing and sales efforts. Continuous consultations with individual clients are also necessary since they might have their own unique set of concerns while the insights generated from our analysis and models are merely an aggregation. With respect to the technical objectives of the project, critical success factors are identifying the most relevant features that affect the prices of resale flats and developing a predictive model that has the minimum Akaike information criterion (AIC) and Root Mean Square Error (RMSE). With regards to sentiment analysis, a critical success factor is the number of social media posts collected over a considerable period of time to avoid introducing bias into the dataset. Last but not least, compliance to best data management and analytics practices will ensure the quality of our work, together with the consistency and reproducibility of our results.

4.3 Key Deliverables

A Predictive Model of Future HDB Resale Prices: PFS will choose the predictive model most suitable for their business requirements and implement it on a platform where it is easily accessible by their real estate agents when they are meeting their clients.

A report that includes the following:

- Key findings from Exploratory Data Analysis (EDA)
- Comparison and evaluation of the analytical methods that will be discussed later in the report
- Summary of findings extracted from social media posts about the latest cooling measures
- Summary of findings from competitors' social media analysis
- Actionable insights and recommendations for PFS

Data Dictionaries: Provide a context for the datasets

5 HDB Resale Price Predictions

5.1 Data Preparation

Despite the fact that the quality of the dataset has been enforced by HDB and all approved resale transactions were recorded using a relatively consistent format, some cleaning needs to be conducted.

Upon further inspection of the "*flat_model*" column, it was observed that there was a change in the case for entries after January 2000. In other words, for transactions before January 2000, entries in that particular column were recorded in uppercase while those from January 2000 onwards were later recorded as capitalized entries. Hence, to prevent duplication of entries, the case of all entries in the "*flat_model*" column were all standardized to uppercase and stored in a new column labelled "*flat_model_standardized*". Similarly, there was also an inconsistency observed in the recording of entries in the "*flat_types*" column. One of the flat types was actually recorded in two different formats, where it was recorded as "MULTI-GENERATION" for transactions before January 2000 and as "MULTI GENERATION" for transactions from January 2000 onwards. As a form of format standardization, the hyphen for this particular flat type in entries of transactions before January 2000 were removed. Just as before, these standardized entries were stored in a new column labelled "*flat_type_standardized"*.

As mentioned in the previous section, the "*remaining_lease*" column was not present for transactions before January 2015. Since the lease for most (if not all) HDB flats are 99 years, this variable can easily be computed using the following formula:

$$99 - (Year\ of\ Transaction\ - lease\_commence\_date)$$

where *Year of Transaction* was extracted from the "*month*" column. The *remaining_lease* for transactions from January 2017 onwards were recorded in years and months, while for those transactions between January 2015 to December 2017 they were recorded only in years. For consistency, the "*remaining_lease*" were converted to months for all transactions and stored in a new column labelled "*remaining_lease_months*".

After the dataset has been cleaned, transformations and aggregations can then be introduced to enhance its predictive power. One such transformation is the conversion of the categorical "*storey_range*" column, which was recorded as a string, to a numeric column by taking the median storey number within the storey range, and labelling the new column as "*median_storey*". The concatenation of both the "*block*" and "*street_name*" columns generated the full "*address*" of the blocks where the HDB resale flats reside. The latitude and longitude of each of these HDB blocks was then obtained from its respective full "*address*" using the geocoding API services provided by Google Maps Platform and OpenStreetMap (OSM). To further enrich the dataset, the towns where these HDB blocks reside were agglomerated into regions as shown in the table below:

Region	Towns
CENTRAL	BISHAN, BUKIT MERAH, BUKIT TIMAH, CENTRAL AREA, GEYLANG, KALLANG/WHAMPOA, MARINE PARADE, QUEENSTOWN, TOA PAYOH
NORTH	SEMBAWANG, WOODLANDS, YISHUN
NORTHEAST	ANG MO KIO, HOUGANG, PUNGGOL, SENGKANG, SERANGOON
EAST	BEDOK, PASIR RIS, TAMPINES
WEST	BUKIT BATOK, BUKIT PANJANG, CHOA CHU KANG, CLEMENTI, JURONG EAST, JURONG WEST, LIM CHU KANG

Table 5.1.1: Region-Town Mapping (Source: https://www.hdb.gov.sg/about-us/history/hdb-towns-your-home)

Further transformations include the conversion of the floor area from square meters (*sqm*) to square feet (*sqft*) and dividing the "*resale_price*" by the floor area in terms of *sqm*/*sqft*.

In preparation for time series analysis, inflation in "*resale_price*" needs to be accounted for. This is achieved by obtaining Consumer Price Index (CPI) values since 1990 from the Monetary Authority of Singapore (MAS) website. Using these CPI values, resale prices adjusted for inflation can then be calculated using the following formula:

$$Resale\ Price\ Adjusted\ For\ Inflation\ = (CPIx * resale\_price)/\ CPIy$$

where x is taken to be the month of the latest transaction record, *i.e.* Feb 2023, and y is the month of the transaction that corresponds to the "*resale_price*".

5.2 Data Exploration

After the dataset has been cleaned and transformed, it is beneficial to explore and understand how the values in the numeric columns are distributed. The following table summarizes the distributions of values in all numeric columns, including derived quantities:

	floor_area_sqm	resale_price	remaining_lease_months	median_storey	resale_price_per_sqm	floor_area_sqft	resale_price_per_sqft	latitude	longitude
count	892418.000000	8.924180e+05	892418.000000	892418.000000	892418.000000	892418.000000	892418.000000	892418.000000	892418.000000
mean	95.709227	3.106401e+05	976.718090	7.646910	3206.550200	1030.200818	297.899629	1.361529	103.839445
std	25.904950	1.619536e+05	124.891342	4.768582	1426.189761	278.837280	132.497973	0.041681	0.073722
min	28.000000	5.000000e+03	516.000000	2.000000	161.290323	301.388108	14.984430	1.270369	103.685206
25%	73.000000	1.900000e+05	900.000000	5.000000	2269.230769	785.761853	210.819405	1.333670	103.773298
50%	93.000000	2.890000e+05	996.000000	8.000000	2890.109890	1001.039073	268.501227	1.354788	103.843656
75%	113.000000	4.050000e+05	1080.000000	11.000000	4042.553191	1216.316293	375.567205	1.382078	103.897814
max	307.000000	1.418000e+06	1212.000000	50.000000	14731.182796	3304.505327	1368.577948	1.457845	103.987631

Table 5.2.1: Summary Statistics of All Numeric Columns, Including Derived Quantities

The histogram plots for the above numeric columns, other than "*floor_area_sqft*", are found in the Appendix. Histogram plots for four additional categorical columns, namely "*flat_type_standardized*", "*flat_model_standardized*", "*region*" and "*town*", can also be found in the Appendix.

To analyze and understand the correlations between the numeric columns within the dataset, a plot of the correlation matrix is shown at the top of the following page:

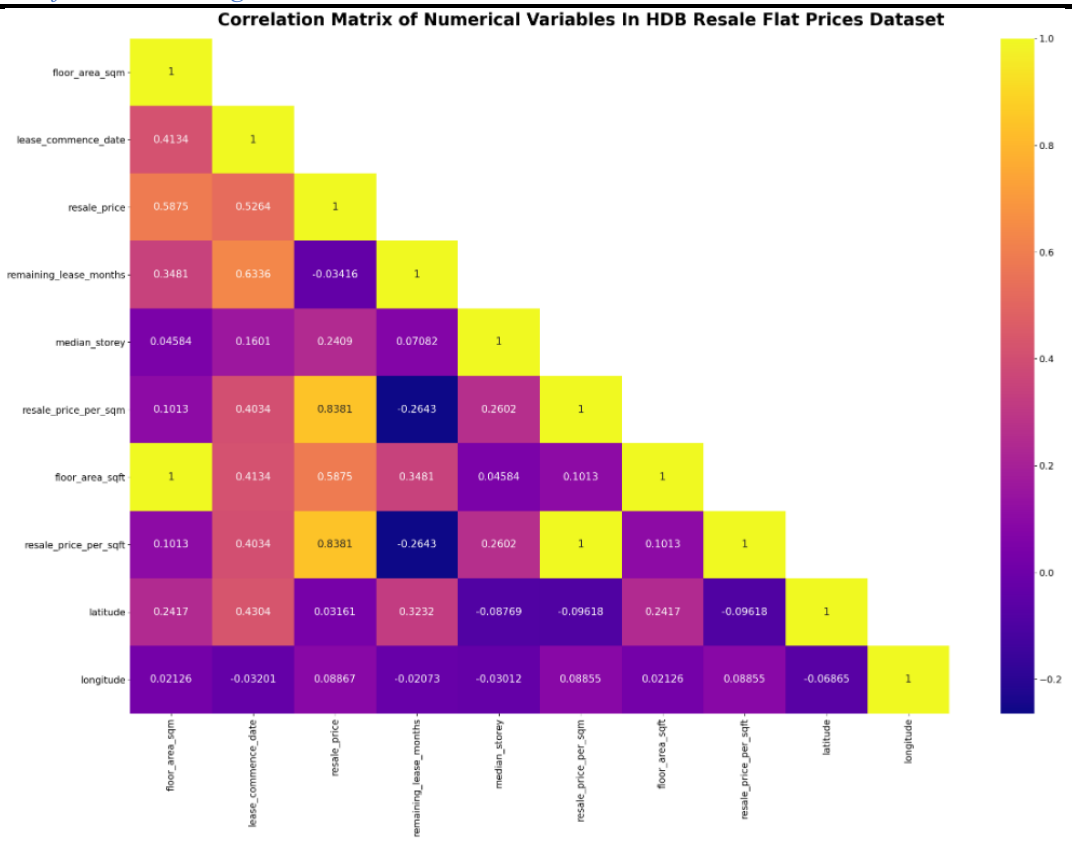


Figure 5-1: Correlation Matrix of All Numeric, Including Derived, Variables

5.3 Feature Selection & Model Preparation

After the data has been prepared, explored and understood, both the predictor (features) and response (target) variables need to be extracted from the dataset. Hence, the chosen predictor variables are shown in the screenshot below:

```
RangeIndex: 892418 entries, 0 to 892417
Data columns (total 10 columns):
 #   Column                      Non-Null Count   Dtype
---  ------                      --------------   -----
 0   Town                        892418 non-null  category
 1   Region                      892418 non-null  category
 2   Address                     892418 non-null  category
 3   Latitude                    892418 non-null  float64
 4   Longitude                   892418 non-null  float64
 5   Remaining Lease (Months)    892418 non-null  int64
 6   Median Storey No.           892418 non-null  int64
 7   Flat Model                  892418 non-null  category
 8   Flat Type                   892418 non-null  category
 9   Floor Area (Square Metres)  892418 non-null  float64
dtypes: category(5), float64(3), int64(2)
```

Figure 5-2: List of Predictors Extracted For Further Modelling

The chosen response variable is the "*resale_price*" since it is easier to evaluate the performance of the predictive models by comparing the actual resale price and its prediction.

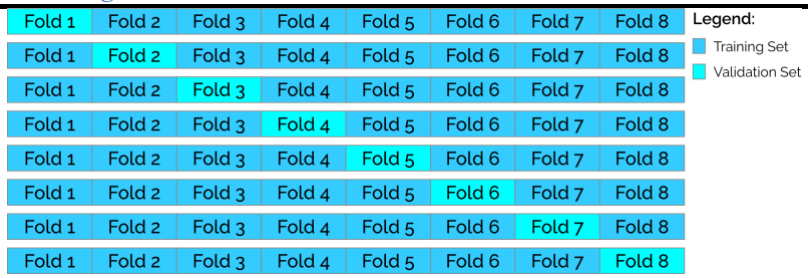
Before constructing the predictive models, the dataset containing the extracted predictor and response variables needs to be split into 2 different sets using a procedure known as the train-test split. In this procedure, the entries/rows in the dataset are shuffled and the resultant dataset is then split into 2 sets using a specified ratio. One of these sets is used to train the models and the other is known as the test set which acts as "unseen" data. A 80/20 ratio is used for the train-test split for all predictive models that will be discussed in the later sections, where 80% of the dataset will be assigned as the training set while the rest as the test set.

5.4 Hyperparameter Tuning

There are mainly two types of predictive models, namely parametric and non-parametric models. Parametric models capture the relationship between the input, *i.e.* predictors, and output, *i.e.* response, variables using a finite set of parameters. Unlike parametric models, non-parametric models do not make any specific assumptions on the form of the function mapping the predictor variables to the response variable.

Hyperparameters are another set of explicitly pre-defined parameters that control various aspects of the learning process. This set of parameters is particularly important for the construction of non-parametric models since their values determine how accurately the models are fitted to the actual data. To maximize the accuracy of model fit, these values cannot be inferred by simply fitting the model using trial and error, rather they are best determined using optimization techniques. Just as for all optimization techniques, there needs to be an objective function. By minimizing/maximizing the objective function, the optimal solution is obtained. In the case of hyperparameter optimization/tuning, the objective function is the selected evaluation metric, which is used to assess the performance of the fitted model. Since the prediction of future HDB resale prices is a regression problem, suitable evaluation metrics include Mean Absolute Error (MAE), Mean Squared Error (MSE), RMSE and Adjusted R^2. Due to its ease of interpretability and the continuity of the objective function, RSME was chosen to be the evaluation metric.

The optimal set of hyperparameter values can be determined using a variety of search techniques. For simplicity, grid search together with a k-fold cross-validation is used to tune the hyperparameters of the predictive models that will be discussed in the later sections. First, the algorithm searches through a grid containing the specified hyperparameter combinations. The k-fold cross-validation is then performed to prevent the model from overfitting the training data. The following illustration demonstrates how a 8-fold cross-validation is conducted:



Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8
Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7	Fold 8

Legend:
- Training Set
- Validation Set

Figure 5-3: Illustration Of How 8-Fold Cross Validation Is Conducted

As shown in the above illustration, the technique starts with splitting the training set into eight equal partitions/folds, where one of these partitions is used as the validation set while the rest is used to train the model. This is then repeated over 8 splits where the partition used as the validation set is not the same as that used in the previous split. The evaluation metric, which in this case is the RMSE, is then averaged over all 8 splits. This whole process is then repeated for all hyperparameter combinations given by the grid.

5.5 Time Series Analysis

To determine a suitable approach for forecasting of resale prices, data exploration is needed to understand the underlying pattern and trends of the time series data. To do so, a line graph of mean HDB resale prices was plotted against time to understand the general trend of HDB resale prices over the years. As shown in Figure , there is a general upwards trend for resale prices since 1990. However, these prices showed large fluctuations in the early years between 1990 and 2008. Given the current market and economic conditions has changed since this period, we decided to conduct time series forecasting with resale prices from the past 15 years onwards since more recent data would be more representative and relevant for future HDB resale price prediction.

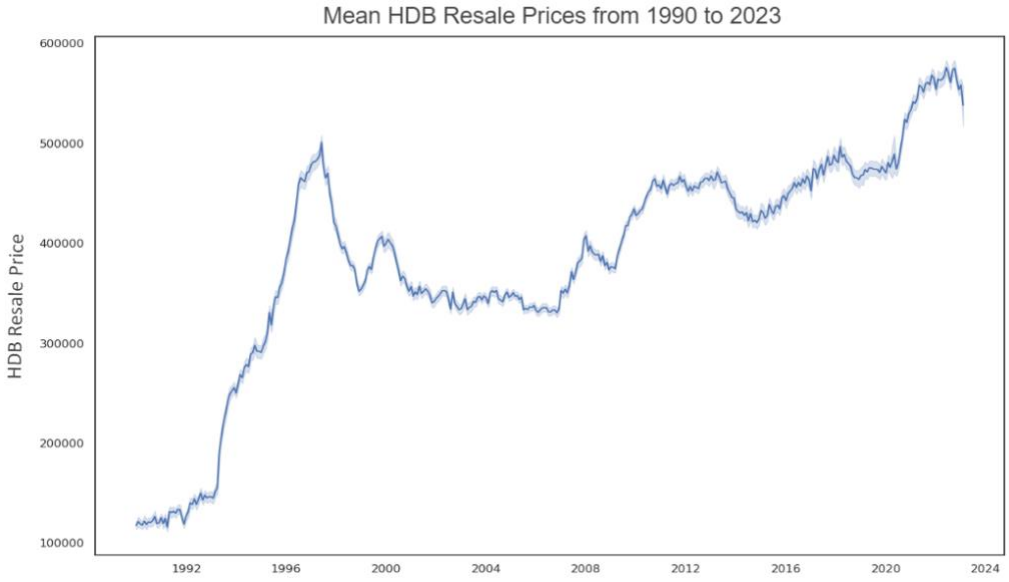


Figure 5-4: Graph Of Mean HDB Resale Prices Against Time

On further analysis of mean resale prices by region since 2008, we observed that the prices for all regions showed similar upwards trends with prices in the Central being the highest and the converse for prices in the North. This is shown in Figure 5-5. This may be due to people being more willing to spend more money for HDB in central areas to enjoy the convenience it brings. Moreover, the available spaces for building new HDBs in the central area are very limited, which also makes the resale market in the central area much hotter than in other regions.

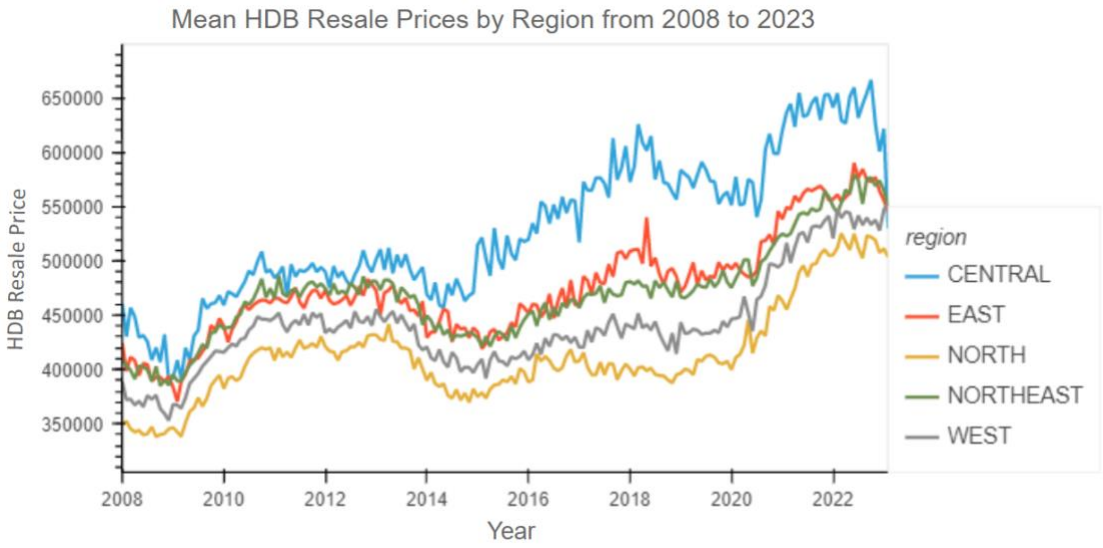


Figure 5-5: Graph Of Mean HDB Resale Prices By Region Against Time

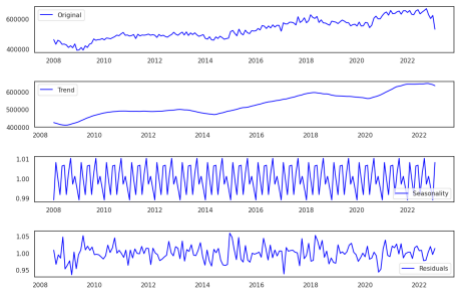
Based on the trends observed in Figure 5-5, the time series data is likely not stationary. To confirm this, Augmented Dickey-Fuller test was conducted using the *adfuller()* function from the Python library *statsmodels*. The null hypothesis for Augmented Dickey-Fuller test states that the time series contains a unit root and is non-stationary. As seen in Table 5.5.1, the p values for time series data for all regions are greater than 0.05. Hence, we failed to reject the null hypothesis. The time series data is confirmed to be not stationary.

Region	Central	East	North	Northeast	West
p-value	0.655825	0.749797	0.34	0.463878	0.471408

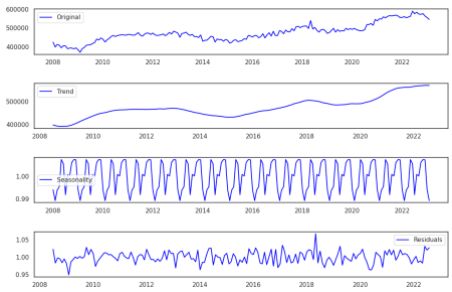
Table 5.5.1: p-values Of The Augmented Dickey-Fuller Test For Time Series Data By Region

Having confirmed that the time series data is not stationary, time series decomposition was conducted to further understand the underlying trend, seasonality and noise in the time series resale price data of each region. As shown by the following table, we further observed that the resale prices displayed seasonality and noise for all regions. Hence, the data needs to be seasonally differenced.

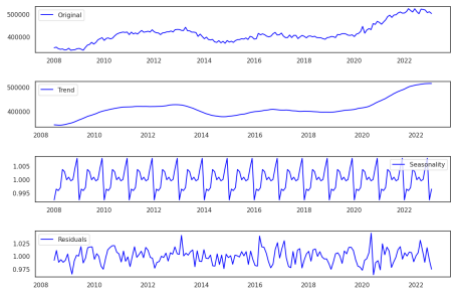
a) Central



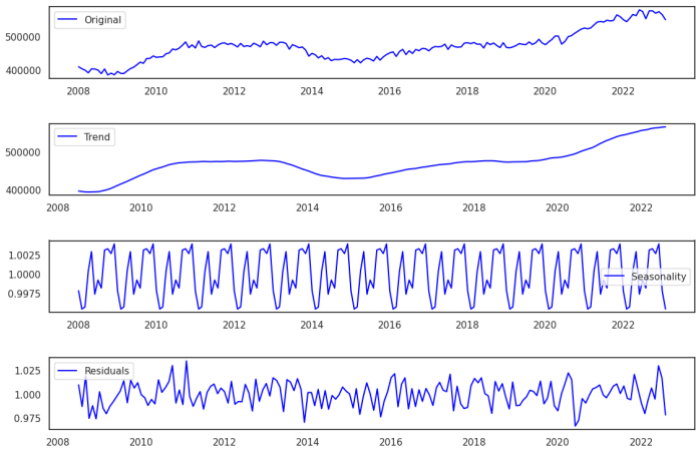
b) East



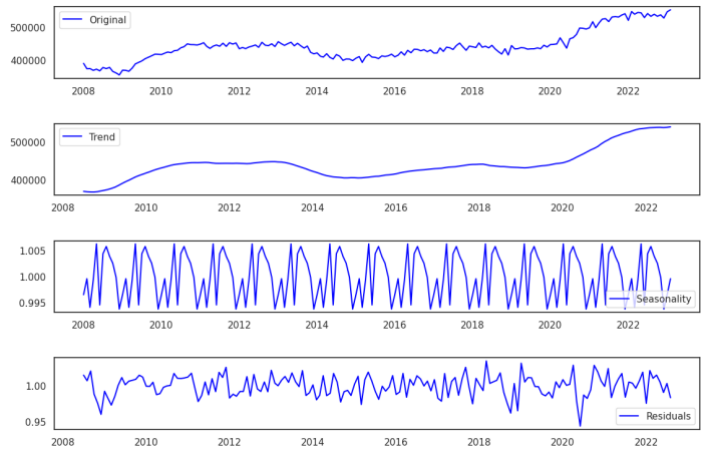
c) North



d) Northeast



e) West



The Augmented Dickey Fuller test was conducted on the residual data and the resulting p values are shown in Table 5.5.2 below. Since the p values are lesser than 0.05, we reject the null hypothesis. Hence, the residual data are stationary, random and do not have predictable patterns.

Region	Central	East	North	Northeast	West
p-value	2.467788E-23	3.132877E-22	3.469370E-12	2.406843E-23	0.000653

Table 5.5.2: p-values Of The Augmented Dickey-Fuller Test For Residual Data By Region

Based on the analysis above, time series resale price data for each region are fit into ARIMA models and are selected based on AIC values. This was achieved using the auto.arima() function in Python library pmdarima, where hyperparameter tuning of seasonal and non-seasonal Auto-Regressive (p), Integrated (d), Moving Average (q) was performed to find models with the lowest AIC values. These selected models were then fitted with test data and their performances were measured by RMSE. The final models and their respective AIC and RMSE values are shown in Table 5.5.3Figure 5-5: Graph Of Mean HDB Resale Prices By Region Against Time.

Region	Model	AIC	RMSE
Central	SARIMAX(1,1,0)(0,1,1)[50]	2,221.62	$32,048.60
East	SARIMAX(1,1,0)(0,1,1)[50]	2,128.05	$33,560.18
North	SARIMAX(1,2,1)(0,1,0)[60]	1,855.75	$43,165.28
Northeast	SARIMAX(0,2,1)(0,1,1)[55]	1,988.46	$17,690.06
West	SARIMAX(0,1,1)(0,1,1)[55]	1,959.84	$47,900.04

Table 5.5.3: Parameters, AIC And RMSE of The Selected Model For Each Region

5.6 Base Machine Learning Models

5.6.1 Multiple OLS Linear Regression

The first and most basic predictive model is the Multiple Linear Regression. Since it is a simple parametric model, the model has a clear functional form that maps the predictor (features) to the response (target) variables. In other words, the model is purely constructed and defined using the selected predictor variables shown in Figure 5-2. The pipeline of the model is shown in the following diagram:

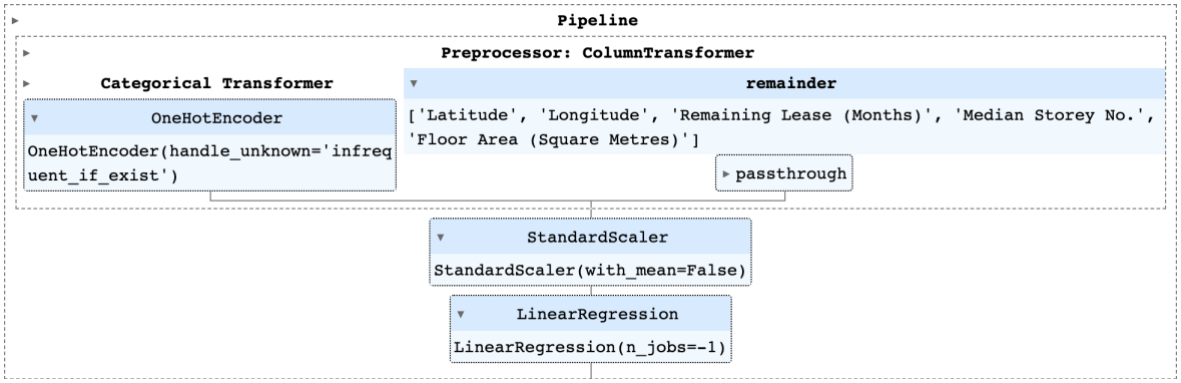


Figure 5-6: Final Pipeline of The Constructed Multiple Linear Regression Model

Since the model is unable to handle categorical variables directly, they need to be transformed using a one-hot encoder before they can be passed into the model. As seen in the pipeline above, the one-hot encoder has a parameter "*handle_unknown*" set to "*infrequent_if_exist*". This means that the transformer will map the one hot encoded column of an unknown category to the infrequent category if it exists and the infrequent category will be placed at the last position in the encoding.

To improve the performance of the model and speed up training, a standard scaler step was included into the pipeline. The standard scaler standardizes the data in numeric columns using the following equation:

$$z = \frac{x - \mu}{\sigma}$$

where z is the z-score for the column, μ is the mean of all entries within the column, and σ is the standard deviation of all entries within the column. Since the categorical variables have already been one-hot encoded, the resultant matrix from the "*ColumnTransformer*" pre-processing step is a sparse matrix, which consumes a large amount of memory. Since the standard scaler step is unable to handle sparse matrices, numeric columns cannot be centred around their respective means and hence the parameter "*with_mean*" was set to *False*.

Once the data has passed through all pre-processing steps, it can finally be fitted in the Ordinary Least Squares (OLS) Linear Regression model.

5.6.2 Decision Tree Regression

Unlike Multiple Linear Regression, decision tree regression is a non-parametric model. Instead of being constructed purely using the predictor variables, the model seeks to learn simple decision rules by inferring patterns that might exist in the data. Based on these decision rules, the model will then build a binary tree where every parent node has at most 2 child/leaf nodes. At each split, the model aims to minimize the MSE which is equal to the reduction in variance of the randomly selected predictor. Since decision tree regression is a non-parametric model, it contains a number of tuneable hyperparameters that can be used to optimize the model fit. For simplicity and the ease of model interpretability, the maximum depth (*max_depth*) of the constructed decision tree was the only hyperparameter being optimized. Using the algorithm discussed in Hyperparameter Tuning and the same pre-processing steps as the Multiple OLS Linear Regression model the pipeline of the decision tree regression is shown below:

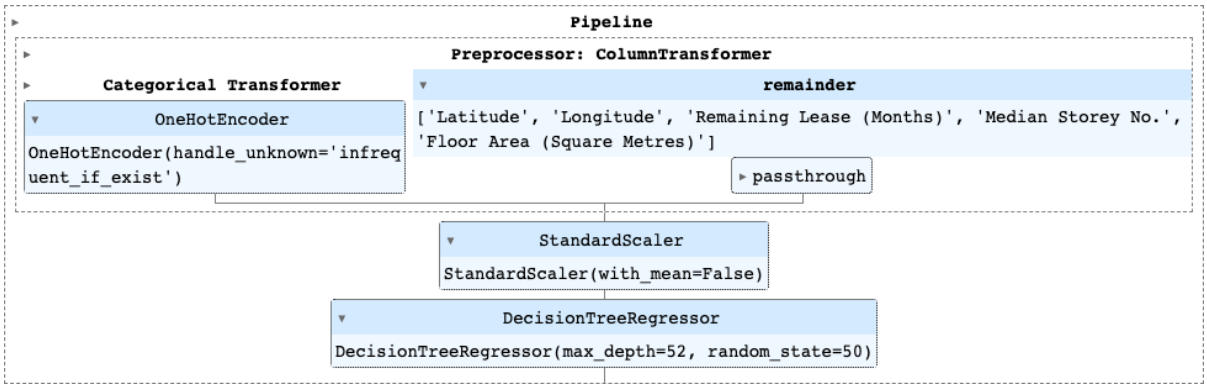


Figure 5-7: Final Pipeline of The Constructed Decision Tree Regression Model

As discussed in Hyperparameter Tuning, the RMSE is the chosen objective function for evaluating the goodness of the model fit. As shown by the above pipeline, the optimum *max_depth* of the constructed decision tree is 52. For this hyperparameter combination, the RMSE of the fitted model was minimized as shown by the plot and table on the following page:

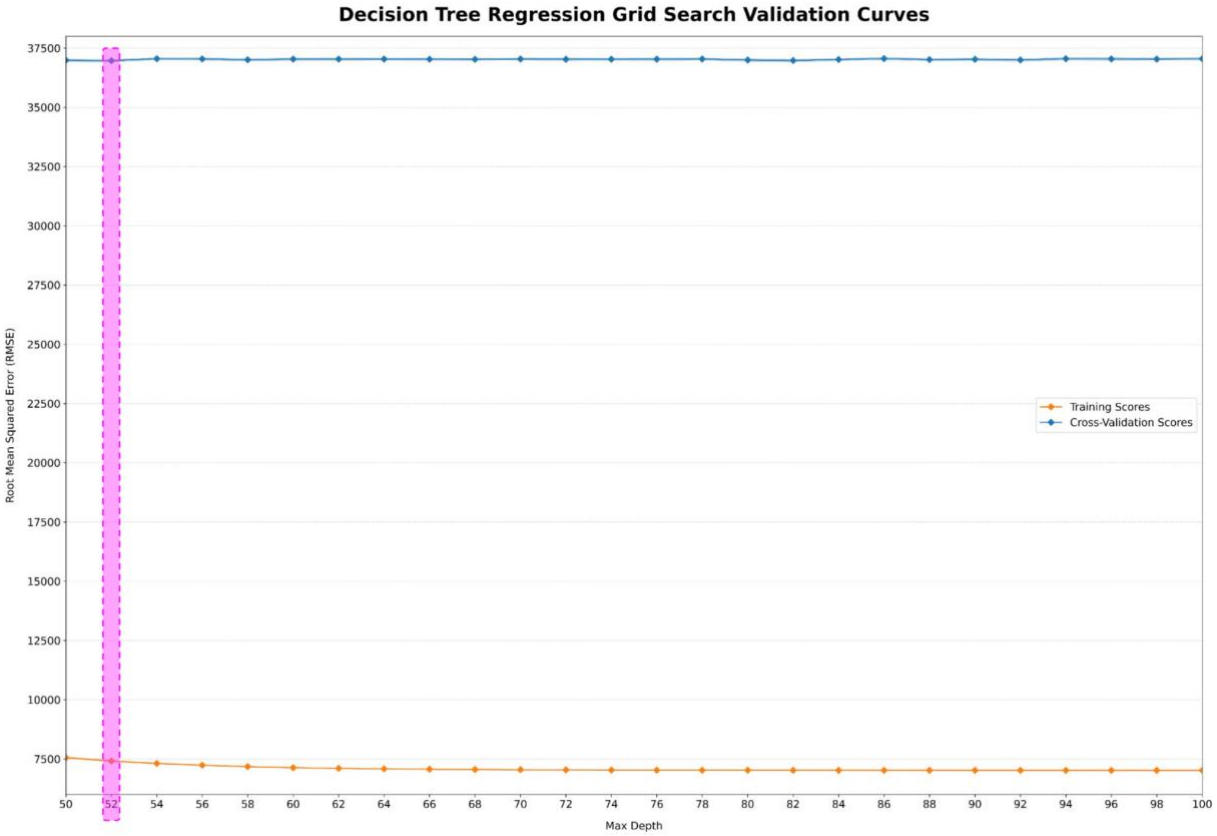


Figure 5-8: Training And Validation Curves From Decision Tree Regression Grid Search

50	**52**	**54**	**56**	**58**	**60**
$(36,982.13 ± 53.85)	$(36,967.72 ± 50.49)	$(37,049.30 ± 42.25)	$(37,038.91 ± 50.54)	$(37,006.59 ± 52.22)	$(37,032.34 ± 50.27)
62	**64**	**66**	**68**	**70**	**72**
$(37,034.16 ± 52.28)	$(37,034.14 ± 44.05)	$(37,027.91 ± 51.72)	$(37,023.34 ± 40.50)	$(37,037.86 ± 42.35)	$(37,025.46 ± 55.52)
74	**76**	**78**	**80**	**82**	**84**
$(37,028.84 ± 44.45)	$(37,029.72 ± 57.37)	$(37,037.51 ± 44.58)	$(36,994.41 ± 57.21)	$(36,973.60 ± 54.12)	$(37,016.42 ± 39.30)

86	**88**	**90**	**92**	**94**	**96**
$(37,054.43 ± 56.34)	$(37,014.17 ± 44.07)	$(37,025.41 ± 51.97)	$(36,998.36 ± 49.57)	$(37,049.14 ± 55.06)	$(37,039.38 ± 59.43)
98	**100**				
$(37,033.97 ± 44.19)	$(37,047.91 ± 51.84)				

Table 5.6.1: RMSE Values of The Fitted Models Based On "max_depth" Hyperparameter

5.7 Ensemble Techniques

For basic decision tree models, some of the major drawbacks include instability and susceptibility to overfitting. These models are highly sensitive to variations in the training data, where small changes to the data can result in disproportionately large changes to the decision rules used in the construction of the models. Furthermore, they have a tendency of becoming too complicated, especially if the constructed decision trees are too deep and the *max_depth* hyperparameter was not properly tuned, and may not be able to generalize to new "unseen" data well.

To minimize the impact of these drawbacks, ensemble techniques, namely random forest and gradient boosted decision trees, will be introduced in the following sections. The main objective of these techniques is to enhance model predictive power and generalization (to new "unseen" data) using an ensemble of base models/learners. The details on how different ensemble techniques can achieve the aforementioned objective will be shared in the following sections.

5.7.1 Random Forest Regression

The first ensemble technique is the random forest regression model, where an ensemble of decision trees is constructed in parallel and independent of each other. During the construction of each decision tree, a sample is randomly drawn with replacement from the training set, which acts as a bootstrap sample. After the ensemble has been fully constructed, the predictions from all decision trees are then aggregated together, using unweighted voting, to generate the final model prediction.

Similar to what was discussed in the previous section, the introduced randomness results in the construction of a diverse set of decision trees. Therefore, the prediction from each individual decision tree becomes somewhat independent of that from another decision tree. By taking the average of all these predictions, some errors in these predictions get somewhat cancelled out resulting in a variance reduction.

Since the number of decision trees will affect the diversity of the ensemble, it is another hyperparameter, in addition to the *max_depth* hyperparameter, that needs to be optimized. Using the hyperparameter tuning algorithm shared in Hyperparameter Tuning and the same pre-processing steps as the previous two models, the random forest regression model is constructed using the following pipeline:

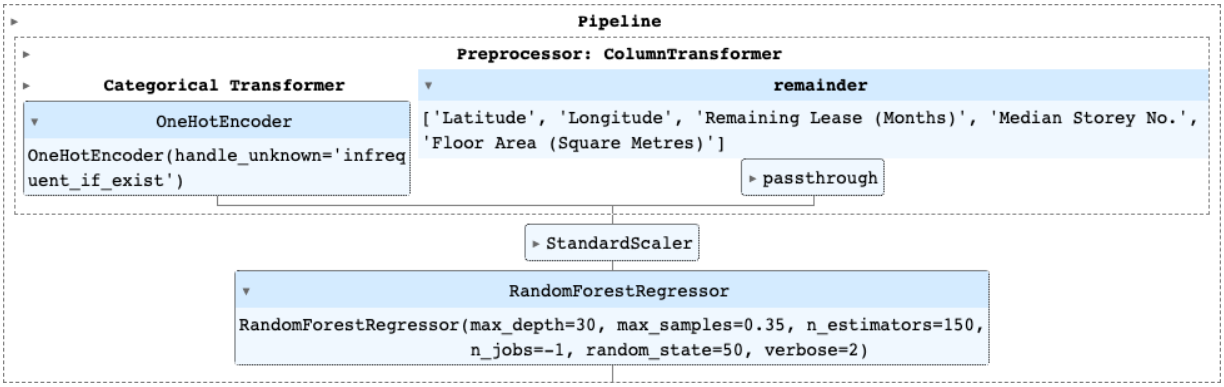


Figure 5-9: Final Pipeline of The Constructed Random Forest Regression Model

	50	75	100	125	150
5	$(104,296.97 ± 77.22)	$(104,349.36 ± 67.53)	$(104,349.51 ± 66.94)	$(104,320.37 ± 68.09)	$(104,349.51 ± 65.58)
10	$(69,649.60 ± 29.10)	$(69,628.10 ± 24.68)	$(69,586.44 ± 20.66)	$(69,567.00 ± 22.44)	$(69,583.17 ± 20.30)
15	$(50,521.51 ± 29.44)	$(50,470.84 ± 29.46)	$(50,428.91 ± 29.94)	$(50,402.94 ± 30.74)	$(50,424.02 ± 31.36)
20	$(38,807.30 ± 23.91)	$(38,709.59 ± 26.45)	$(38,662.73 ± 26.99)	$(38,614.34 ± 27.56)	$(38,627.49 ± 27.97)
25	$(33,954.85 ± 18.53)	$(33,827.68 ± 21.36)	$(33,767.45 ± 21.00)	$(33,715.73 ± 21.85)	$33,705.60 ± 21.83)
30	$(32,438.55 ± 17.06)	$(32,297.39 ± 19.16)	$(32,222.47 ± 18.24)	$(32,170.42 ± 20.69)	$(32,151.74 ± 20.54)

Table 5.7.1: Decision Tree Regression Hyperparameter Grid With The RMSE Values Of The Fitted Models

5.7.2 Light Gradient Boosting Machine (LightGBM)

The LightGBM is a distributed gradient boosting framework designed with a focus on efficiency and scalability. Unlike most decision tree algorithms which grow trees level/depth-wise, the LightGBM algorithm grows trees leaf-wise, where the algorithm evaluates the loss for each leaf node and then chooses the one with highest loss to grow. The algorithm is summarized by the following illustration:

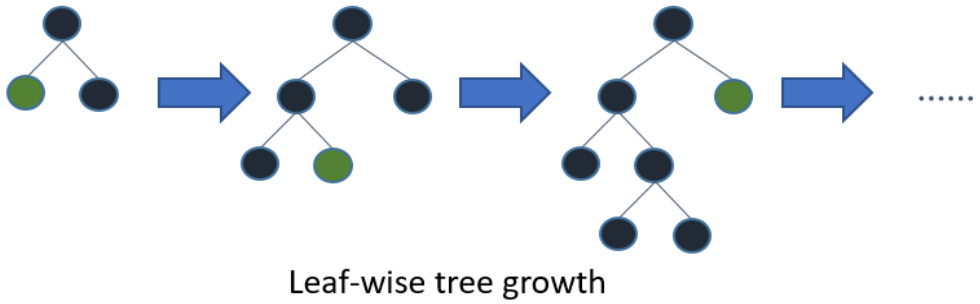


Figure 5-10: Decision Tree Growth Based On The LightGBM Algorithm

It is believed that the reduction in MSE achieved by leaf-wise growth algorithms tends to be lower than that achieved by level-wise algorithms. However, it is also believed that such algorithms tend to produce much more complex trees, hence increasing the likelihood of overfitting. This potential pitfall in the algorithm can be avoided by regulating a number of hyperparameters, such as the maximum depth of each tree (*max_depth*) and the maximum number of leaves in each tree (*num_leaves*). The maximum number of leaves (*num_leaves*) is actually related to the maximum depth (*max_depth*) via the following formula:

$$num\_leaves = 2^{max\_depth}$$

As seen by the above relation, the tuning of the *num_leaves* hyperparameter indirectly optimizes the *max_depth* hyperparameter. This in turn regulates tree growth and hence prevents the likelihood of overfitting. To simplify the construction of the pipeline, only two hyperparameters will be optimized, namely the number of boosted trees (*n_estimators*) and the maximum number of leaves in each tree (*num_leaves*).

Since LightGBM is a histogram-based algorithm, it is able to handle categorical variables directly, without the need of one-hot encoding. For numerical variables, the algorithm discretizes the continuous values in these columns into bins which accelerates model training and reduces memory consumption. As a result of the aforementioned simplifications, the pipeline of the LightGBM model becomes very straightforward, as shown in the following diagram:

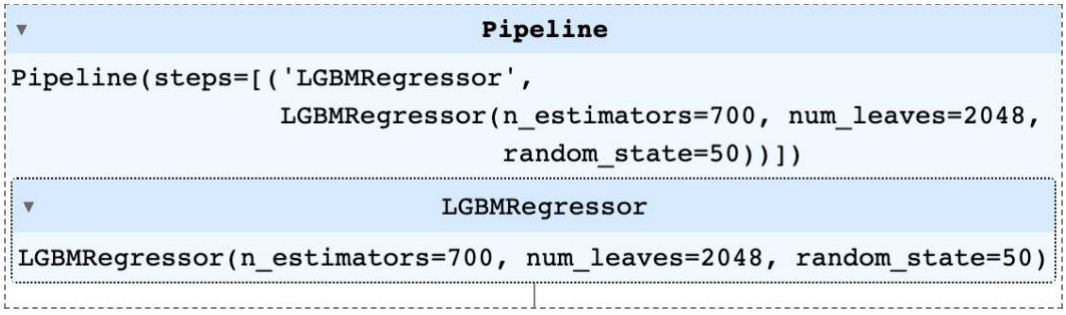


Figure 5-11: Final Pipeline of The Constructed LightGBM Model

	512	1024	2048	4096
100	\$(33,124.74 ± 26.45)	\$(30,388.22 ± 16.10)	\$(28,444.60 ± 13.65)	\$(27,199.64 ± 8.86)
200	\$(30,464.54 ± 21.03)	\$(28,193.68 ± 16.52)	\$(26,879.63 ± 14.53)	\$(26,247.73 ± 12.12)
300	\$(29,172.12 ± 22.50)	\$(27,301.13 ± 18.54)	\$(26,393.52 ± 9.50)	\$(26,128.75 ± 12.63)
400	\$(28,311.70 ± 25.18)	\$(26,862.00 ± 16.07)	\$(26,185.61 ± 7.55)	\$(26,131.91 ± 13.99)
500	\$(27,740.01 ± 25.73)	\$(26,566.33 ± 13.77)	\$(26,085.24 ± 10.41)	\$(26,186.15 ± 15.12)
600	\$(27,397.79 ± 24.44)	\$(26,382.00 ± 14.22)	\$(26,042.24 ± 11.51)	\$(26,256.62 ± 15.96)
700	\$(27,145.67 ± 20.25)	\$(26,240.31 ± 13.54)	\$(26,030.48 ± 11.64)	\$(26,329.38 ± 16.11)

Table 5.7.2: LightGBM Hyperparameter Grid With The RMSE Values Of The Fitted Models

5.8 Overall Model Evaluation

After all the models have been constructed, the performance of each individual model needs to be evaluated and compared with one another. Therefore, an evaluation metric needs to be selected for a fair comparison between models. Using the same argument as that discussed in Hyperparameter Tuning, the chosen evaluation metrics were the RMSE and Adjusted R^2. An additional point of comparison is the time taken to train the model, including hyperparameter tuning. The complete comparison of all four models is shown in the table below:

Model	RMSE	Adjusted R^2	Training Duration[1]
Multiple OLS Linear Regression	\$59,482.42	0.865311	3.76 secs
Basic Decision Tree Regression	\$35,768.35	0.951297	2 hrs 21 mins 42 secs
Random Forest Regression	\$31,101.60	0.963177	24 hrs 3 mins 4 secs

[1] Models are trained on a 14" Macbook Pro 2021 (Apple M1 Pro chip with 10-Cores CPU, 16-Cores GPU & 16-GB RAM)

LightGBM	$25,590.15	0.975071	6 hrs 48 mins 20 secs

Table 5.8.1: Model Comparison Using Different Evaluation Metrics

As inferred from the table above, the best performing model is the Light Gradient Boosting Machine (LightGBM). Since the algorithm does not require categorical variables to be one-hot encoded, the resultant model can easily be interpreted, which is the "icing on the cake". Hence, it is straightforward to identify what the key factors influencing the prices of HDB resale flats are, which will further be discussed in Market Forecast and Features Analysis.

6 Sentiment Analysis of HDB Cooling Measures

6.1 Data Preparation

For ease of analysis, raw data from both HardWareZone and Reddit were combined to create a single dataset. Two columns were included in this final dataset, namely "comment" and "comment_date".

	comment	comment_date
0	I'm sorry to say he has no credibility. He said at a [Jan 2021 REDAS event](https://www.businesstimes.com.sg/property/singapore-government-ensure-property-market-line-economic-fundamentals) that "THE Singapore government is monitoring the developments in the property market "very closely", and will adjust policies if necessary, to maintain a stable and sustainable property market for Singaporeans". What followed was over 24% increase in resale flat prices in two years.\n\n> "That said, given our track record, investors may also remain bullish on our property market as economies reopen."\n\nThis I agree. His track record of managing property prices helped spark FOMO buying.	2023-01-13
1	Notice how he said "Price growth moderated" and not "Price moderated"? \n\nSlower rate of price growth is still price growth.	2023-01-13
2	Havnt even pass one month of 2023 and they start to project price growth become moderate. Wait till we see constant growth after 1H and they show #SurprisePikachuFace.	2023-01-13
3	Already at all time peak and we're kind of told that the "growth rate moderated" and be happy with it. Gov doesn't want the price to drop and indirectly encouraging the stupid assumption of "property always goes up". \n\nJust postponing the problem and waiting for disaster in future.	2023-01-13
4	# *"we will continue to monitor the current situation very closely to ensure a readily available supply of homes for future generations"*	2023-01-13

Figure 6-1: Example of The Combined Dataset

To prepare the text data for analysis, text pre-processing was performed to clean, format and tokenize the combined data. The following steps were taken to clean the data:

i. Removal of irrelevant comments

Irrelevant comments such as comments made by bot accounts, as well as comments that have been deleted or removed (labelled as '[deleted]' and '[removed]' respectively) were filtered off from the dataset. Additionally, it was assumed that comments that were too short in length (≤ 5 words) were likely to be irrelevant to the topic of interest and were also removed.

ii. Replacement of emojis

As emojis play a significant role in expressing sentiments, they were replaced with the expression they represent, in text.

Example: 😁 = happy

iii. Expansion of contractions and short forms

To standardise the text data, contractions and abbreviations were expanded to their original form.

Example: won't = will not, gov = government

iv. Removal of HTML tags

As HTML tags do not add any value towards the analysis of text, they were removed as part of text pre-processing.

v. Removal of quotes

Block quotes of other user's comments would result in duplicated comments. As such, they were removed from the dataset.

vi. Removal of user mentions and hashtags

Unnecessary strings such as user mentions and hashtags were removed from the text data.

Example: @user, #hashtag

vii. Removal of repeating characters

Characters that were repeated more than necessary were dealt with by reducing the repetition.

Example: hdbbbb

viii. Removal of special characters

Not only do special characters not add any value to text understanding, they could also possibly induce noise into algorithms. Hence, it is crucial to strip the text data off these characters.

Example: !?&*^$123

ix. Conversion of all strings to lowercase

To ensure that all strings follow a consistent format, all words were converted to lowercase.

x. Removal of stopwords

While stopwords are helpful in understanding the structure of a sentence, it has little to no significance in understanding the semantics of the sentence. Standard English language stopwords list from NLTK

(Natural Language Toolkit) in python was used to filter stopwords from the data. Singlish terms such as 'lah', 'leh', 'lor', etc were also added to the list of stopwords.

xi. Tokenization

Tokenization is an essential step when working with text data. It involves breaking down sentences into smaller units (words) that can be easily understood by a machine. Comments were tokenized using the tokenize module of NLTK.

xii. Lemmatisation

To ensure standardisation of text, lemmatization was performed to remove word affixes, trimming words to its root form.

6.2 Approach to Analysis

A three-pronged approach was taken to analyse the extracted social media posts, threads and comments. Firstly, initial analysis was performed for data understanding and gives a brief overview of the extracted data along with some initial findings. Secondly, thematic analysis was performed to uncover hidden and underlying themes in the extracted data by grouping the comments into common clusters or topics and identifying important keywords. Finally, qualitative analysis was performed in order to contextualise the keywords identified and obtain a better understanding of the commonly discussed themes and sub-themes in social media discussions.

6.3 Data Exploration and Analysis

Initial exploratory analysis was performed on the extracted comments and thread/post titles to get a sense of the common topics of discussion.

For an exploratory overview of the extracted comments, a word cloud summarising the top 200 most commonly occurring words is shown in Figure 6-2 below. This provides a glimpse into general public sentiments and potential concerns with regards to the cooling measures and housing property markets. For instance, words like "hdb", "price" and "buy" appear to stand out, highlighting housing price as a major concern. Words about the various property markets (e.g. hdb, resale, private, build-to-order (BTO), rental) are also prominent, indicating that discussions not only center on HDB resale market, but also on private property, BTO and rental markets. Finally, terms like "interest rate", "inflation", "expensive" and "affordable" express worry about the impact of inflation, as well as the servicing of housing loans amidst a high interest rate environment.

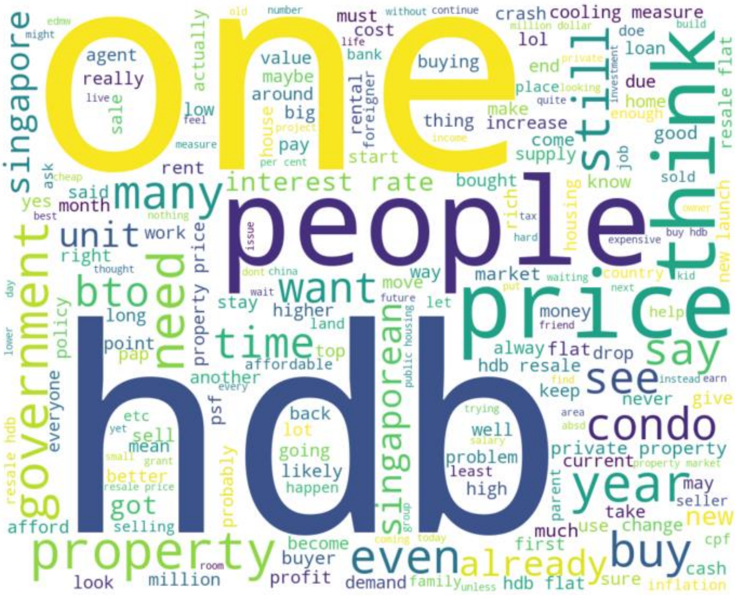


Figure 6-2: Top 200 Most Commonly Occurring Words In The Extracted Comments

In addition, to identify the hottest topics of discussion, HardWareZone threads and Reddit posts with more than 100 comments were chosen for further analysis. Based on the titles of top commented threads and posts (refer the Appendix, it is observed that popular social media discussions on cooling measures and housing markets tend to center on the following topics: property price speculation, cooling measures and housing policies discussion, impact of cooling measures on private property owners and rental markets, and million dollar flats.

6.4 Thematic Analysis

Given a fixed timeline and the large number of comments extracted for analysis, it may be too impractical and time-consuming to manually review every single comment to gather insights from the data. To address this challenge, three natural language processing (NLP) methods - namely K-means clustering, agglomerative hierarchical clustering and topic modelling - were used to analyse the extracted HardWareZone and Reddit comments automatically. These NLP methods help to simplify the vast amount of data that is collected into common clusters or topics and keywords which can be utilised for further analysis.

6.4.1 K-Means Clustering

The first technique used was K-means clustering, which is an unsupervised machine learning technique that groups un-labelled dataset into different clusters. To implement the K-means algorithm, the number of clusters (K) to be created needs to be determined and elbow method was used to have an indication of clusters. In the elbow method, the number of clusters was varied from 1 - 80. For each value of K, WCSS (Within-Cluster Sum

of Square) was calculated, as shown in Figure 6-3. However, as seen from the elbow curve, the elbow point is ambiguous and appears to be between 35 - 45.

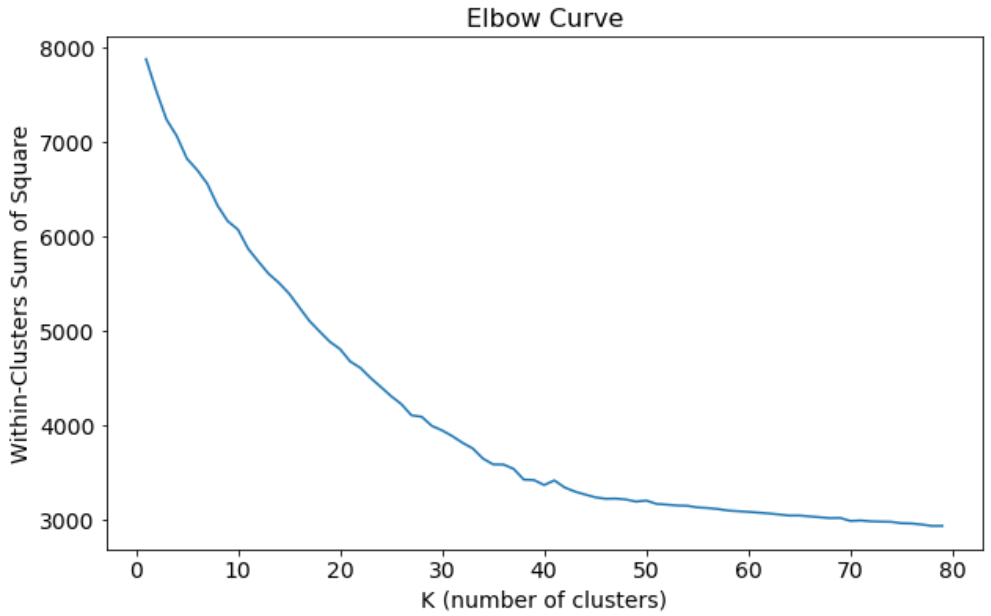


Figure 6-3: Elbow Method For Identification Of Optimal K Value In K-Means Clustering

To further validate the value of K, silhouette scores for 35 to 45 clusters were calculated (Figure 6-4) Out of this range, 45 clusters had the highest score at 0.29, indicating the optimal k for this dataset is 45.

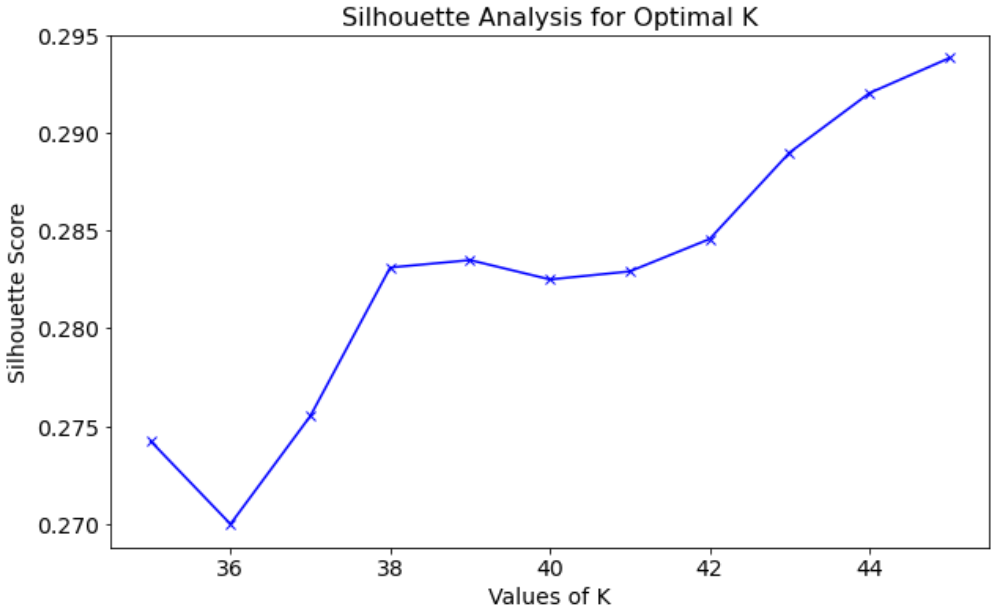


Figure 6-4: Silhouette Scores For 35-45 Clusters

Hence, K = 45 was used to fit into the K-means algorithm and silhouette analysis was conducted to evaluate the quality of these clusters. Firstly, the average silhouette score is relatively low at 0.29, denoting overlapping clusters. In addition, from Figure 6-5(a), it was observed that there are presence of clusters with below average silhouette score, presence of samples with negative silhouette scores, as well as large variations in the size of the clusters. This indicates that the clustering of the dataset is not well handled by K-means clustering, since the ideal conditions of silhouette analysis are not met. As such, agglomerative hierarchical clustering was attempted next.

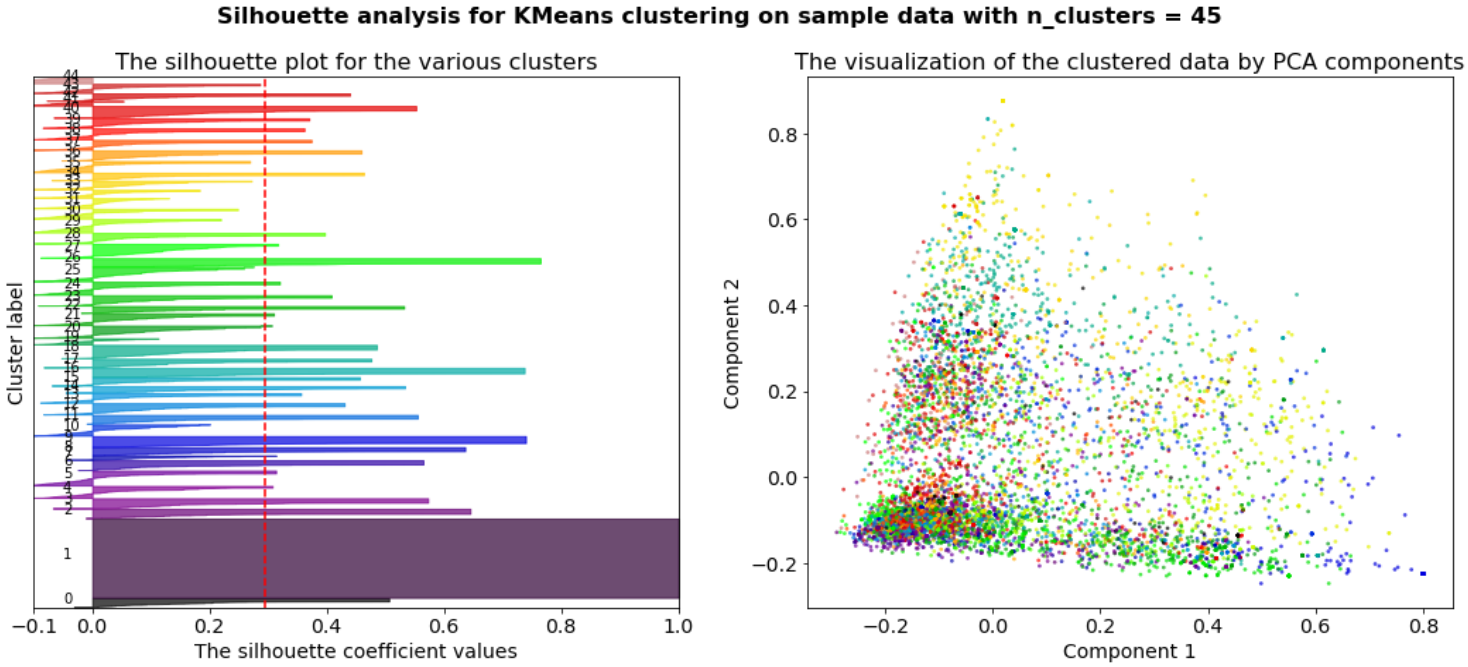


Figure 6-5: (a) Silhouette Plot of 45 Clusters; (b) Visualization Of The Clustered Data

6.4.2 Agglomerative Hierarchical Clustering

Hierarchical clustering, which is also a method of cluster analysis, groups data into a hierarchy of clusters. Unlike K-means clustering, the number of clusters do not need to be predefined in hierarchical clustering. Using Ward's method to build the clusters, the following dendrogram was generated (Figure 6-6). Based on the dendrogram, 3 clusters were selected to fit into the model.

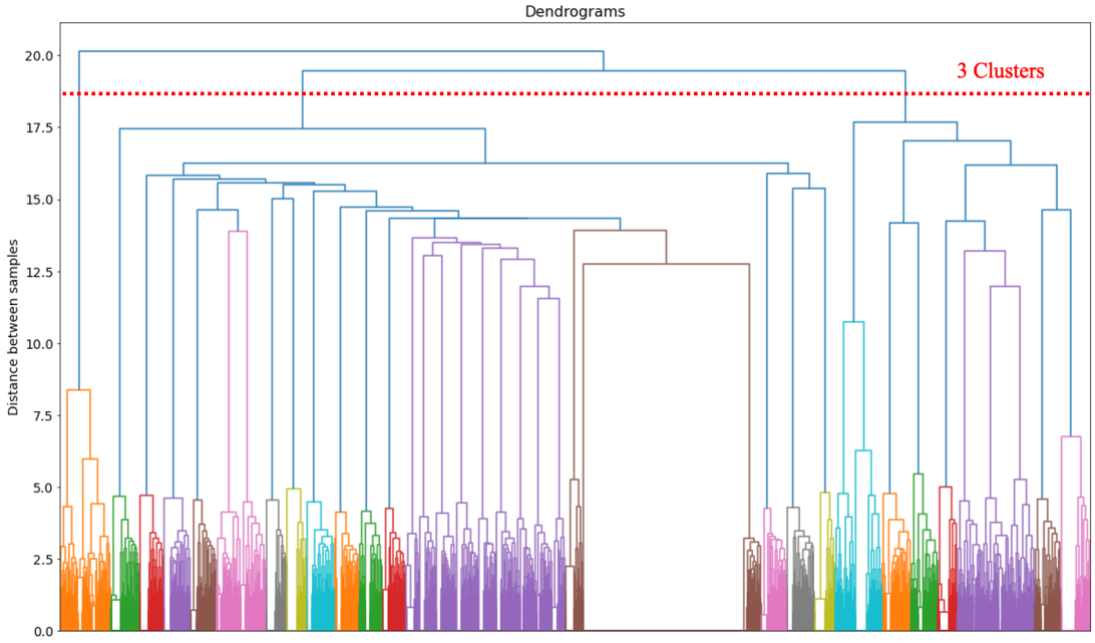


Figure 6-6: Dendrogram Of Agglomerative Hierchical Clustering

After building the model, silhouette analysis was conducted to evaluate the quality of the 3 clusters. The average silhouette score is even lower in hierarchical clustering at 0.077, compared to K-means clustering (0.29). With a value near 0, this means that the distance between clusters is not significant and could indicate overlapping clusters. Similar to K-means clustering, not only are there presence of clusters with silhouette scores below average, there are also presence of samples with negative silhouette scores and an unbalanced classification of data (Figure 6-7(a)). As the result of hierarchical cluster analysis is not ideal, it was not used in the analysis of the dataset.

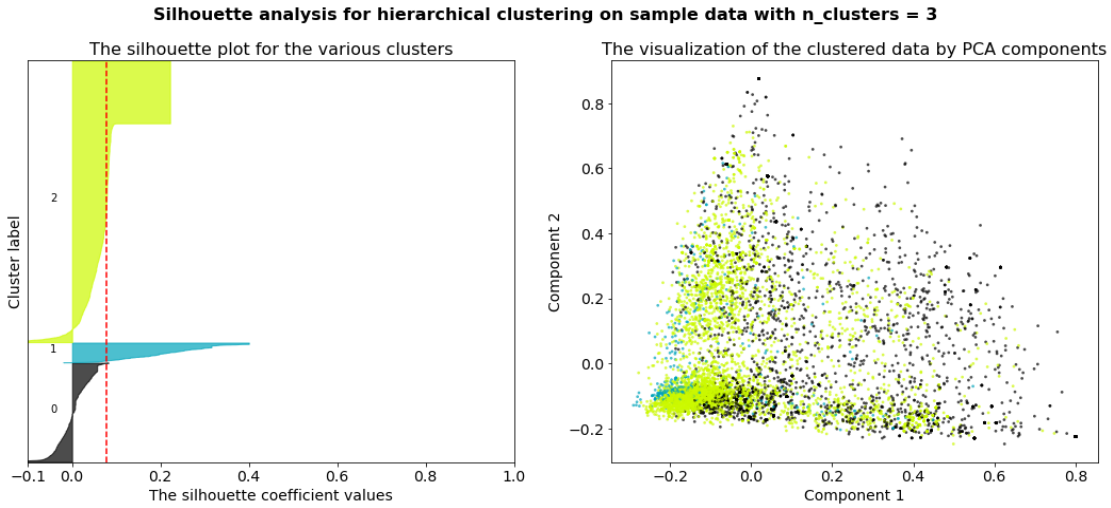


Figure 6-7: (a) Silhouette Plot Of 3 Clusters; (b) Visualization Of The Clustered Data

6.4.3 Topic Modelling

The third technique used was topic modelling - an unsupervised machine learning technique that detects patterns in words and phrases within documents and groups them by topic. Latent Dirichlet allocation (LDA), a popular topic modelling method used in analysing social media data was used. To determine the optimal number of topics (k), a series of LDA models were created (with k ranging from 2 to 20) and evaluated using coherence scores. From Figure 6-8, it is observed that CV is the highest at lower topic numbers (2 to 4 topics), while UMass decreased as the number of topics increased. After balancing for coherence of topics and interpretability of topics, a 4-topic model was chosen as the optimal model.

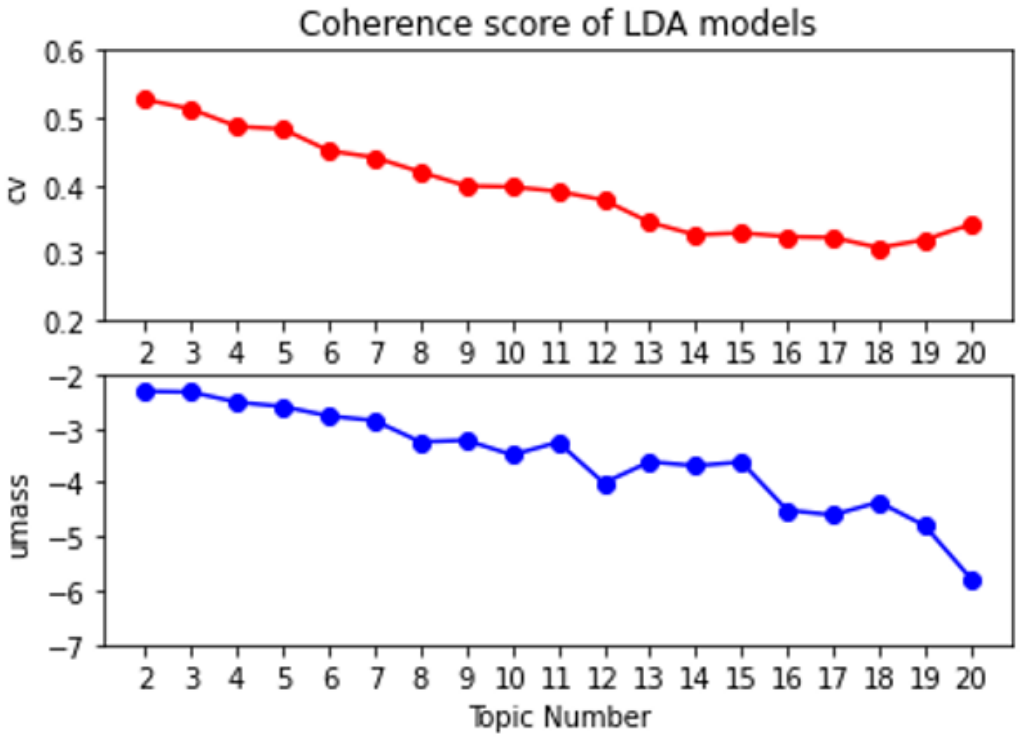


Figure 6-8: Coherence Score Of The LDA Models

Figure 6-9, on the following page, summarises the four topics identified and ten representative keywords for each topic. The first topic identified is the wide ranging impact of cooling measures on various property markets. The second topic is property price speculation. The third topic is housing policy debates while the last topic identified is that of housing-related financial concerns.

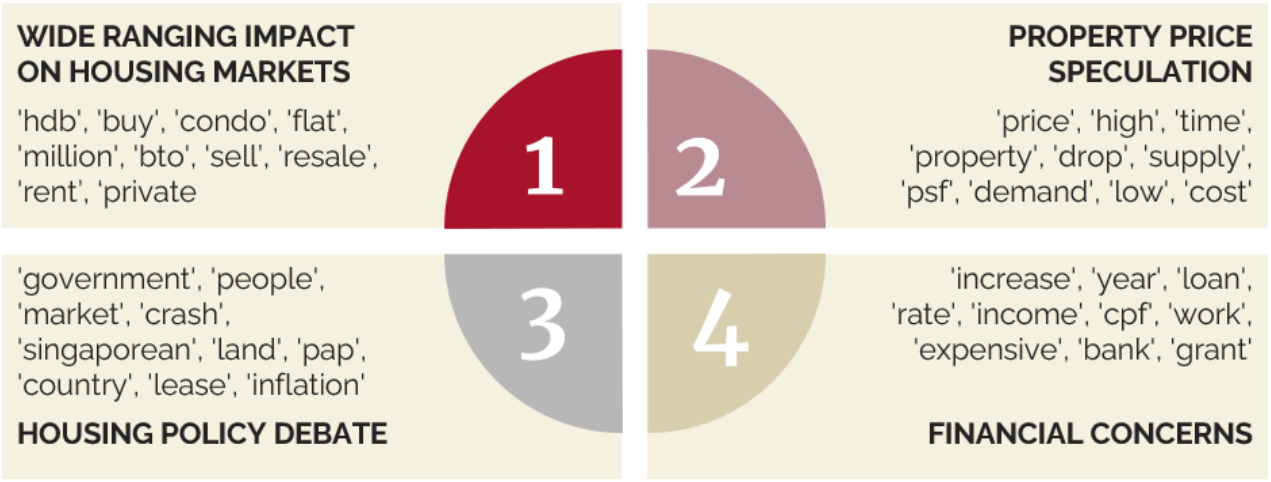


Figure 6-9: 10 Most Important Keywords By Topic

6.5 Qualitative Analysis

Following the identification of four main topics and their representative keywords, further analysis was performed in an attempt to understand keywords in the context of the comment made. For each keyword, 10 comments (with comment length of >50 words) were sampled randomly and reviewed to give contextual meaning to the keywords. This helps to deepen the understanding of issues discussed and allow for common themes and sub-themes to be identified. Incorporating the findings from this review process and the earlier analysis performed, six final themes were developed as shown in Table 6.5.1 below.

S/N	Theme	Sub-themes
1	Impact of Cooling Measures	• Moderate prices and demand in HDB resale flats • Private property owners looking to switch to HDB resale flats • Higher rental demand and prices
2	Commentary on Cooling Measures	• Effectiveness of cooling measures • Treating HDB flat as a home or investment • Curbing housing demand (e.g. foreign and investment-based) and increasing housing supply (e.g. BTO)
3	Advice Seeking and Decision Making	• When to buy or sell property (now or wait) • Buy or rent • Financial prudence

4	Worries and Concerns related to Housing Markets	• Financial challenges and concerns (e.g. financing loans, high interest rate, inflation, overleveraging, economic downturn) • Limited housing supply • High demand from foreigners, investors and new citizens driving property prices up
5	Housing Price Speculation	• Predicting future housing price trends (e.g. HDB resale, private property, new launches)
6	Property News	• Million dollar HDB flats • New launches (e.g. BTO, Condominium)

Table 6.5.1: Themes And Sub-Themes Identified From HardWareZone And Reddit Comments

As seen from Themes 1 and 2, majority of the online social media discussions on HardWareZone and Reddit naturally revolved around the cooling measures put in place by the Singapore government in September 2022. In an attempt to cool the HDB resale market, measures introduced were targeted towards increasing the difficulty of borrowing and dampening HDB resale flat demand from private property owners by introducing a 15-month waiting period. With these measures implemented, homeowners and prospective buyers will need time to figure out their next steps. Depending on whether the housing loan is taken from private financial institutions or HDB, prospective buyers need to take into account the new maximum loan quantum limits, notably the Total Debt Servicing Ratio (TDSR), Mortgage Servicing Ratio (MSR) or Loan-to-value (LTV) limit. In this case, financial calculators and loan guides will be particularly useful for those impacted. In addition, the waiting period enforced also increased the demand for rental flats as private property owners still need a place to stay before they can switch from private property to HDB resale flats. This was confirmed with ample discussions on rising rental costs.

Theme 3 focuses on advice seeking and decision making, with some commenters describing their housing and financial situation and looking for advice for their next steps from the community. Common questions include when to buy or sell a property, whether to do so immediately or to continue to wait for better market conditions, and whether to buy private property, resale property or to rent instead. Prospective buyers and sellers, especially those with limited knowledge or existing financial concerns, may be blindsided by the cooling measures and need to seek advice from property agencies like PFS to guide them in their decision-making process.

Theme 4 covers various worries and concerns expressed towards the housing market. In particular, financial challenges and concerns are brought up repeatedly in social media discussions. With the current high inflation

and high market interest rate environment, borrowing costs for home purchases are expected to maintain at a high level. Reviewing of own finances, ensuring prudent borrowing and the avoidance of overleveraging is key for existing and prospective homeowners as they attempt to navigate an ever-changing housing market. This information is useful for PFS to better understand overall market sentiments and concerns and offer targeted information and tips in their client interactions (e.g. face-to-face, social media) that address these worries and concerns.

Theme 5 involves discussions on housing price speculation for various housing markets (e.g. HDB resale, private property, new launches). With housing property being a major asset for the majority of Singaporeans, there is always a close eye on how property prices will change over time. This is especially so for prospective or existing homeowners who want to make a housing transaction in the immediate future. Being able to make an accurate prediction of future housing prices is important for prospective or existing homeowners in deciding when to make the buy/sell transaction in order to maximise their profit gains or minimise costs. This can be satisfied by our predictive model for HDB resale flat prices which can be implemented as an interactive tool on PFS's website, social media and in face-to-face client interactions.

Finally, Theme 6 centres around discussions on property news. In particular, million dollar HDB resale flat transactions attract a lot of attention and are seen by some as a proxy for current housing market conditions and a way to assess the effectiveness of cooling measures. With housing supply in Singapore being tightly controlled, it is also not surprising to see high interest levels in new BTO and private property launches every year. Similarly, PFS can provide analysis and content on these attention-grabbing topics on its website and social media.

7 Analysis of competitor's Facebook activities

7.1 Data Preparation

Our final dataset for competitor's Facebook analysis consists of 3 columns of scraped data, namely "time", "reactions" and "post_text", as well as one column "competitor" that was added to distinguish between the competitors' posts. Since the data was scraped separately from two different Facebook Groups, there was a need to merge them into a neat table for analysis. The data dictionary of the merged table can be seen below:

Column	Description	Data Type	Sample Value(s)
time	Datetime of when post was uploaded	Datetime	2023-02-20 18:00:03

competitor	Tell us which competitor posted it	String	PropertyGuruSG
reactions	Dictionary of number of likes & reactions	Dictionary	{'like':4'}
post_text	Textual contents of post	String	From new private home sales surging…

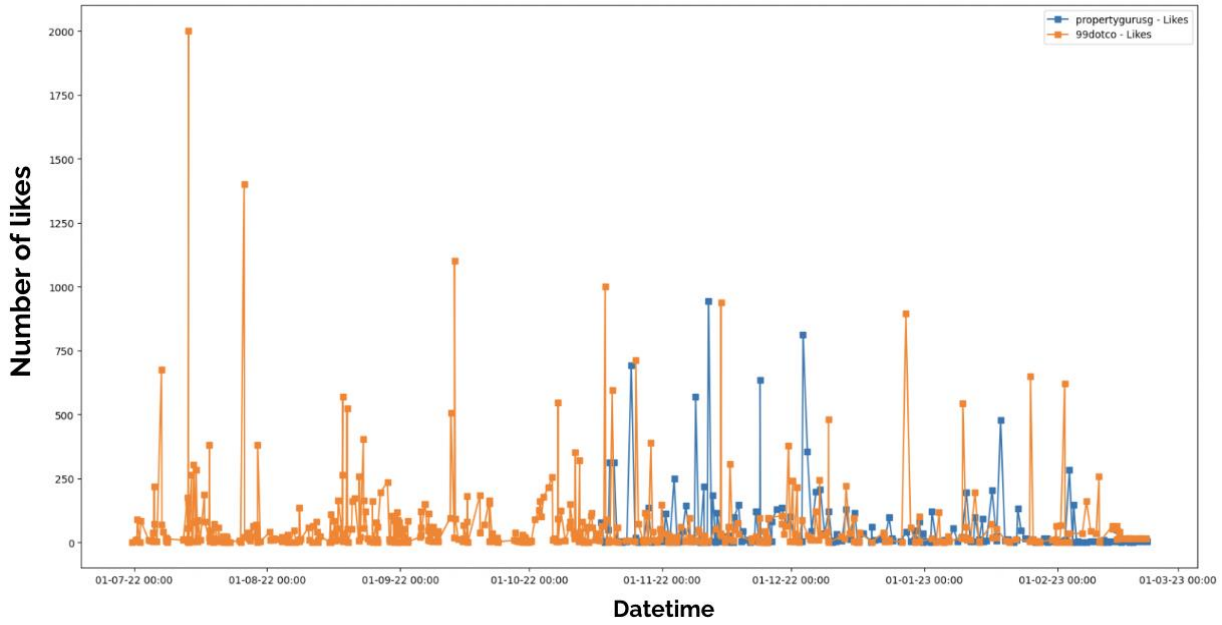
Table 7.1.1: Data Dictionary of Competitor's Facebook Posts

	time	competitor	reactions	post_text
1	2023-02-20 18:00:03	propertygurusg	{'like': 4}	From new private home sales surging % month-on-month (MoM) in January to the CP
2	2023-02-19 18:00:01	propertygurusg	{'like': 8, 'love': 1}	For those considering buying an HDB BTO or a resale flat, this ultimate checklis
3	2023-02-19 12:00:01	propertygurusg	{'like': 3}	Planning to buy a new flat with an HDB loan? We answer some of your burning ques
4	2023-02-18 18:00:01	propertygurusg	{'like': 3}	Here are some tips for those who want to be 'strategic' about applying for a BTO
5	2023-02-18 12:00:34	propertygurusg	{'like': 2}	When Biek's husband needed to relocate to Singapore for a new job, they moved in
0	2023-02-21 15:00:20	99dotco	{'like': 16}	Do you know any friends or family members planning to apply for an SBF this year
1	2023-02-20 20:00:51	99dotco	{'like': 16}	We also recommend using the Property Value Tool to check your potential gain!\n#
2	2023-02-20 08:00:19	99dotco	{'like': 16}	Within the first seven days of February , condos were sold with recorded capita
3	2023-02-19 16:00:29	99dotco	{'like': 16}	In early February , condos were sold above their projects' previous all-time-hi
4	2023-02-18 16:00:23	99dotco	{'like': 16}	Based on HDB resale transaction registrations in February (so far), HDB resale

Prior to NLP modelling, the text data has to be cleaned. Any advertisements or links were removed, along with HTML tags or special characters. Words that were not impactful to the meaning of the sentence were identified as stopwords and removed. The text was then tokenized, before being lemmatised to their base form. Words that appeared less than 3 times were also removed.

7.2 Data Exploration & Analysis

We used number of likes as a proxy for performance, and the number of likes for each post for both competitors was plotted against datetime as shown below. This illustrates how 99dotco has a lot more likes/posts compared to PropertyGuruSg, at least over the time period of interest.



Our analysis section will be split into 4 parts, with 7.3.1 analysing the frequency of competitor's posts. This will be a point of parity where PFS will need to match this posting frequency in order for their posts to not be drowned out by those of the competition. 7.3.2 will analyse the popular hashtags that were used in competitor's top liked posts, with 7.3.3 analysing the topics posted by competitors. Last but not least, the analysis in 7.3.4 will focus on the popular locations mentioned in competitors' posts, and a visual heatmap will be generated to illustrate this easily to stakeholders. These analyses will provide PFS with points of differentiation to help them stand out from the competition.

7.2.1 Frequency of Posts

On average, PropertyGuruSg posted an average of 10.3 posts/week, and 99dotco posted an average of 14.3 posts/week. This means that to even be the average of the competitors, PFS will need to post a minimum of 12.3 times/week. This works out to be a daily average of approximately two posts a day by PFS.

7.2.2 Popular Hashtags

Since number of likes was our proxy for performance, we ranked the posts in descending order of likes. The hashtags were then extracted, grouped according to hashtag, and the mean number of likes for each hashtag was calculated. The average like for each hashtag across the posts was used as this allowed us to account for differences in performance across different posts for the same hashtag. The ten top performing hashtags are included in the following diagram, and they happen to fall into three distinct categories: "being environmentally friendly", "financing homes", and "sharing stories of their homes".

hashtags	
SustainableLiving	478.000000
GreenLiving	478.000000
EcoFriendly	478.000000
Mortgage	311.000000
HomeLoans	311.000000
HomeFinancing	311.000000
GuidesGuru	296.800000
pghomestories	290.276596
HomesOfSG	290.276596
capitalgains	284.146341

Being environmentally friendly

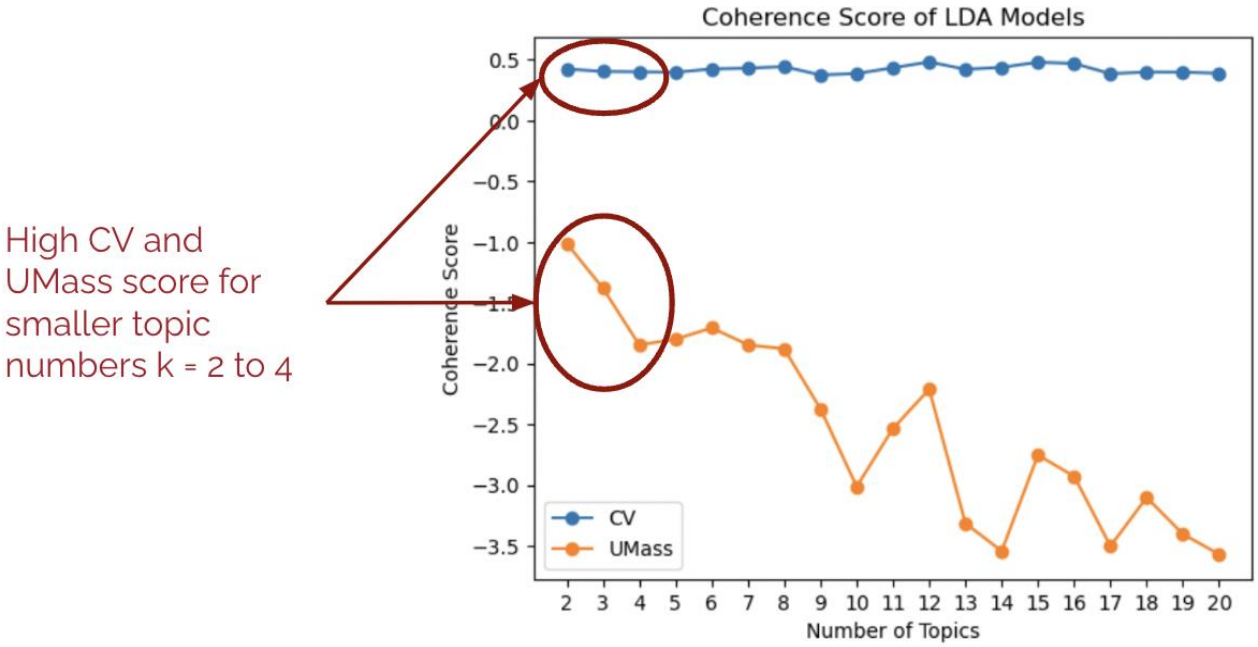
Financing homes

Sharing stories of their homes

Figure 7-1: Top 10 best Performing Hashtags

7.2.3 Topic Modelling

For topic modelling, we can see that CV remained largely the same even when the number of topics increased, but UMass dropped off significantly as the number of topics increased. After balancing for coherence of topics and interpretability of topics, we decided to go with 4 topics in our topic modelling.



The four topics that were interpreted from the posts can be seen below, with property-related topics such as why Woodlands property is a good buy, launch of new flat in the west and development units in Ang Mo Kio taking up three out of the four topics. The last topic identified was that of financing expensive units.

Figure 7-2: Top 4 Most Discussed Topics Identified From Competitor's Facebook Posts

7.2.4 Popular MRT locations

In order to better understand the location of the real estate properties the competition was focusing on, the unique MRT stations were extracted from each post before being converted into a Lat-Lon data and plotted as a heat map. In the diagram below, each python list corresponds to a Facebook post, with each python string corresponding to a unique MRT station that was mentioned in the post.

```
['Kallang', 'Tampines', 'Serangoon', 'Tengah', 'Queenstown', 'Bedok'],
['Kallang', 'Tampines', 'Serangoon', 'Tengah', 'Queenstown', 'Bedok'],
['Tampines'],
['Tampines'],
['Kallang', 'Queenstown', 'Yishun', 'Tengah'],
['Kallang', 'Queenstown', 'Yishun', 'Tengah'],
['Kallang', 'Queenstown', 'Tengah'],
['Kallang', 'Queenstown', 'Tengah'],
['Tengah'],
```

Figure 7-3: MRT Stations Mentioned In Competitors' Facebook Posts

From the heatmap, it is clear that the most commonly mentioned properties belonged to the west and south of Singapore.

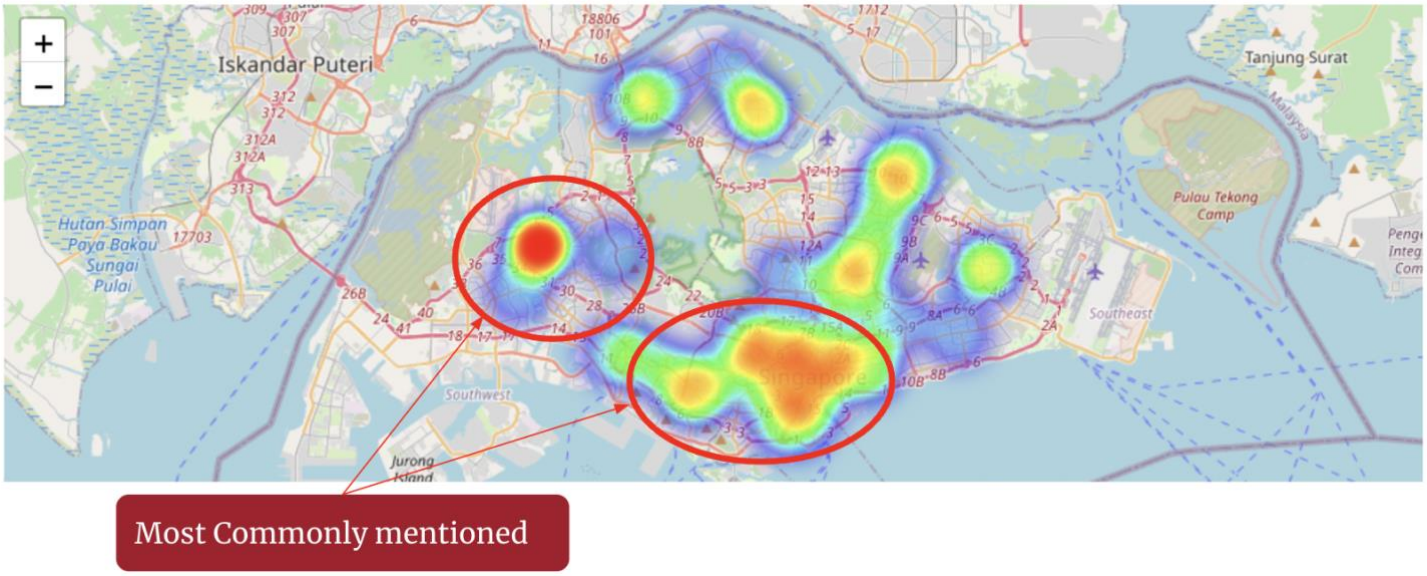


Figure 7-4: Heatmap Of Unique MRT Stations Mentioned In Competitor's Facebook Posts

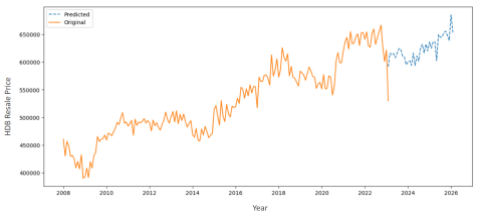
8 Recommendations & Potential Future Work

8.1 Market Forecast and Features Analysis

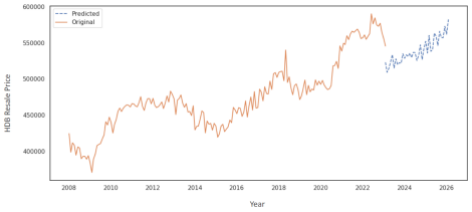
A self -service model derived from the predictive models discussed earlier will provide PFS the ability to generate HDB resale price predictions by region autonomously. Coupled with the identified key factors affecting resale prices, this self-service model will also empower PFS to seek out deeper market insights and generate a suitable set of recommendations to its clients.

Using the price prediction models developed for each individual region, a forecast of future HDB resale prices by region has been generated for the next three years. As seen in the table at the top of the following page, HDB resale prices in the Central and Northeast regions were predicted to increase while resale prices in the other regions were forecasted to have rather lack-lustre performance. Based on this forecast, if potential growth in resale prices is of interest, clients would need to set their sights on a HDB resale flat in the Central or Northeast region. However, individual needs and priorities may vary from client to client. For instance, a young couple working in the North may have a preference for resale flats located around Woodlands or Yishun which is close to their working place. Since the travel time from their potential new home to the Central region may not be of a high priority, they will most likely purchase a resale unit in the neighbourhood areas within the North region of Singapore.

a) Central

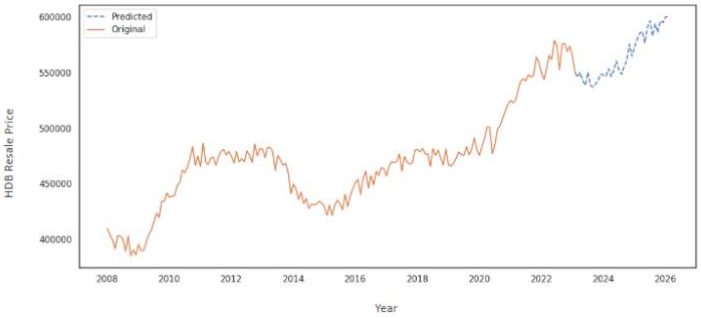


b) East

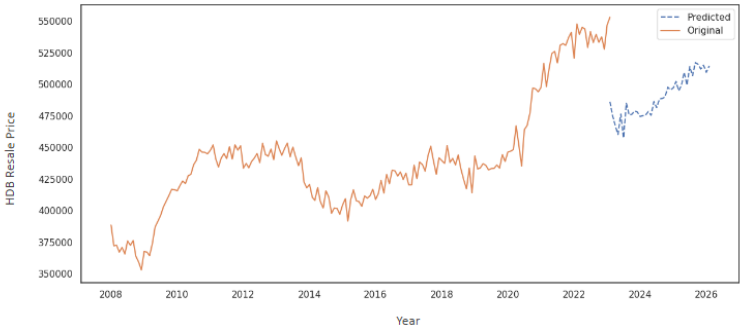


c) North

d) Northeast



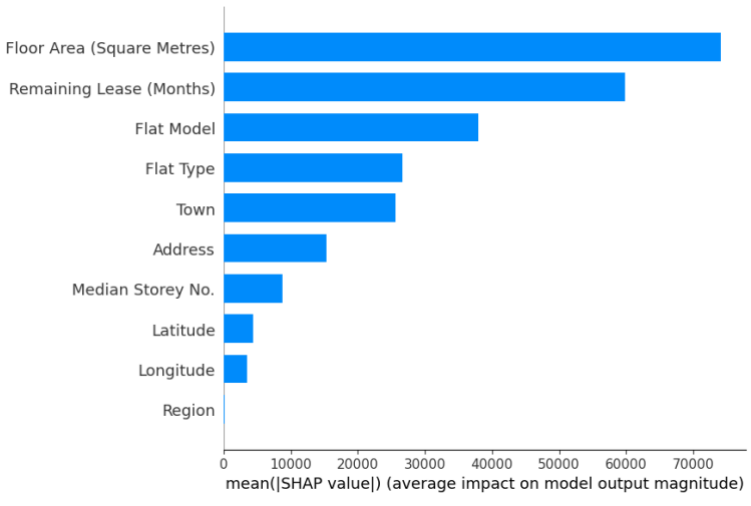
e) West



For this group of clients, since they already have an idea of where they want their new homes to be located at, they may be more interested in the key features of resale units and their influence on resale prices. As mentioned previously in Overall Model Evaluation, the LightGBM algorithm was chosen due to its ease of interpretability and since it was identified as the best performing model. A technique to interpret the output of the model is to plot the relative contributions of individual features using SHAP (SHapley Additive exPlanations) library, as shown by the two summary plots below:

1) Summary Plot Type: Dot

2) Summary Plot Type: Bar



From the bar plot, it can be clearly inferred that the two main features affecting the resale price are the floor area and the remaining lease of the resale unit. It is important to note that the floor area of the resale flat is related to the flat type and model, which happened to be the next two major contributors influencing resale price. As mentioned earlier, different clients will have different sets of needs and priorities. To further elaborate on the previous example, the young couple have also expressed an intention of settling down over the long term and starting a family. Given this particular set of requests, PFS understands that the couple would appreciate affordable moderately sized recommendations from less mature HDB estates in the North region of Singapore. With this in mind, a good set of recommendations would need to include 4-room, or larger, resale flats with at least 960 months in remaining lease located in Sembawang, Yishun and Woodlands.

8.2 Overall Social Media Strategy

Firstly, PFS needs to decide on a consistent brand messaging, tone, and aesthetics across all social media platforms for it to compete more effectively in the social media landscape.

Secondly, with regards to the type of content to post, PFS needs to find a good balance between providing promotional content (e.g. listings) and information that users will find helpful in making their purchase decisions (e.g. educational content such as market analysis, market updates, home buyer or seller guides). To stand out from its competition, PFS can consider ways to differentiate itself from competitors. For example, with regards to promotional content, PFS can consider giving more attention to properties in the East and Central, which tends to be the more expensive and luxurious homes compared to competitors who focus on the West and South regions. In addition, PFS can also consider focusing more on providing educational content and storytelling versus its competitors who tend to focus on new launches and sales. This would help to build PFS's reputation as a trustworthy and authoritative figure in the real estate industry whose clients can rely on for comprehensive and up-to-date information on property markets as well as tailored property advice.

Thirdly, on the frequency of posting, PFS needs to at least match or exceed the high activity level of competitors on each social media platform (e.g. post at least twice per day on Facebook). This is especially important in the beginning stages of building PFS's social media since newer accounts take more effort and time to gain traction and following.

Finally, PFS should always utilize hashtags when posting and aim to use the most popular hashtags for related content. This helps PFS improve their visibility on the platform, better organise their posted contents, while also making it easier for social media users to find similar content that they may be interested in.

8.3 Social Media Content Ideas

Based on the insights from the predictive model, feature analysis, public sentiment analysis and competitor analysis performed, PFS can utilise the following social media content ideas to create valuable and curated content for its platform:

1. Hotly discussed topics
    - Cooling measures and housing policies. Social media attention and engagement is typically highest when new housing policies are first announced. PFS should post a rapid analysis within one day to capture this interest, and make sure that the analysis is targeted towards the impacted demographics - e.g. for cooling measures announced in September 2022, impacted parties include prospective property buyers, private property sellers intending to switch to HDB resale, landlords and renters
2. Information that addresses worries and concerns expressed towards the housing market
    - For financial concerns, provide financial tips and advice (e.g. importance of financial prudence), housing loan guides and useful financial calculators for different demographics - first-time home buyers or sellers, nationality and type of property
    - For housing price concerns, provide quarterly market outlook updates and analysis for each market segment - HDB resale, BTO, condominium and rental markets. PFS can also provide insights to the most important features affecting HDB resale price
    - For housing supply concerns, provide information and analysis for the new launches of the year - BTO and condominium
3. Housing market updates (e.g. million dollar flat transactions, top performing districts, housing demand changes over time) and forecasts (e.g. projected HDB resale prices for different regions) to highlight interesting or noteworthy trends
4. Interviews with subject matter experts on current market outlook, insights and future projections (e.g. interest rate, overall macroeconomic environment)
5. Popular topics posted by competitors (e.g. tips on sustainable living, client success stories)

Finally, a social media content calendar (Figure 8-1) that outlines the various types of content and posting frequency is created for PFS.

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Week 1	Home buyer tip	Analysis on latest cooling measure	- Listing - Video walkthrough of listing	Client testimonial	- Market trend - Interview with subject-matter experts	New launches	- Neighbourhood guide - Eco-friendly home tips
Week 2	Analysis on new launches	- Listing - Factors affecting resale flats prices (insights from predictive model)	Rental market update	Rental market outlook	- Renovation tips - Million dollar flat transactions	- Listing - Guide on housing loans	- Neighbourhood guide - Customer success stories
Week 3	Client testimonial	Sharing statistics on housing demands over the years	HDBs market update	- HDBs market outlook - Findings of time series analysis	- Listing - Home buyer tip	- Customer success stories - Guide on financial calculators	- Neighbourhood guide - Top performing districts
Week 4	New launches	Analysis on new launches	Private property market update	Private property market outlook	- Client testimonial - Interview with subject-matter experts	- Listing - Video walkthrough of listing	- Neighbourhood guide - Customer success stories

Figure 8-1: Example of Social Media Content Calendar

8.4 Potential Future Work

With the help of feature engineering and selection, the predictive models discussed in HDB Resale Price Predictions, have attained excellent performance. However, additional features could still be included to further improve model performance, such as distance to nearest MRT/LRT station or other key amenities in the vicinity. In addition, other types of Machine Learning models can be considered and experimented with. One such model that could be well suited for predicting HDB resale prices is K-Nearest Neighbours (KNN) algorithm. The algorithm uses the idea of "feature similarity" to make predictions, which in this case predicting the price of a HDB resale flat based on features of similar resale units.

Another potential area for future work is to combine predictions from both time series and feature-based analyses together using a weighted average. This combined analysis is believed to be a more holistic approach to understanding the key factors driving the HDB resale market, resulting in more accurate and realistic resale price predictions.

While LDA is useful in uncovering underlying patterns and topics within text data, one limitation of the technique is that it does not account for semantic understanding. To overcome this, one potential future work for the NLP solutions is to experiment with deep learning models such as BERTopic (topic modelling with BERT - Bidirectional Encoder Representations from Transformers). As BERTopic takes into consideration the relationships between words, it could potentially improve topic quality of which the content is semantically correlated.

9 Conclusion

With the help of a predictive model that is capable of forecasting future HDB resale prices, PFS is better

equipped to generate a customized and data-driven set of pricing recommendations for its clients. In addition, with insights from both public sentiment and competitor analysis, PFS is now in a better position to succeed in its social media marketing against growing competition and offering better service to existing and potential clients.

References

1) Monetary Authority of Singapore. (2022, September 29). Measures to Promote Sustainable Conditions in the Property Market by Ensuring Prudent Borrowing and Moderating Demand. https://www.mas.gov.sg/news/media-releases/2022/measures-to-promote-sustainable-conditions-in-the-property-market-by-ensuring-prudent-borrowing-and-moderating-demand

2) Singapore Department of Statistics. (2023, February 9). Households - Latest Data. Retrieved May 8, 2023 from https://www.singstat.gov.sg/find-data/search-by-theme/households/households/latest-data

[Predictive model]

3) Resale Flat Prices. (2023, January 16). Retrieved January 16, 2023 from https://data.gov.sg/dataset/resale-flat-prices

4) *Welcome to LightGBM's documentation!* (n.d.). LightGBM. https://lightgbm.readthedocs.io/en/v3.3.2/

5) *Cross-validation: evaluating estimator performance.* (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/cross_validation.html#cross-validation

6) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *JMLR*, 12, pp. 2825-2830

7) Saxena, S. (2020, April 20). A Beginner's Guide to Random Forest Hyperparameter Tuning. https://www.analyticsvidhya.com/blog/2020/03/beginners-guide-random-forest-hyperparameter-tuning/

8) *API Reference.* (n.d.) Scikit-learn. https://scikit-learn.org/stable/modules/classes.html#

[HardwareZone info]

9) SPH Media Solutions. (n.d.). HardwareZone.com. Retrieved May 8, 2023 from https://www.imsph.sg/hardwarezone-com/

[Reddit info]

10) r/singapore (n.d.). Reddit.com. Retrieved May 11, 2023 from https://www.reddit.com/r/singapore/

[Example of text analytics papers using social media data]

11) Chong, M. & Choy, M. (2018). The social amplification of haze-related risks on the Internet. *Health Communication*. 33, (1), 14-21. Research Collection Lee Kong Chian School Of Business. https://ink.library.smu.edu.sg/lkcsb_research/4343

12) Chipidza W., Krewson C., Gatto N., Akbaripourdibazar E., & Gwanzura T. (2022) Ideological variation in preferred content and source credibility on Reddit during the COVID-19 pandemic. *Big Data Soc*. 2022 Mar 9; 9(1):20539517221076486. doi: 10.1177/20539517221076486.

13) Kang, B. N. Y. (2019). A quantitative exploration of repeated advice-seeking on Reddit's r/relationships. Master's thesis, Nanyang Technological University, Singapore. https://hdl.handle.net/10356/137027

14) Khoo, S. Z. T., Ho, L. H., Lee, E. H., Goh, D. K. B., Zhang, Z., Ng, S. H., Qi, H., & Shim, K. J. (2020) Social media analytics: A case study of Singapore General Election 2020. 2020 IEEE International Conference on Big Data: Virtual conference, December 10-13: Proceedings. 5730-5732. Research Collection School Of Computing and Information Systems. https://ink.library.smu.edu.sg/sis_research/5649

15) Markides, B. R., Laws, R., Hesketh, K., Maddison, R., Denney-Wilson, E., & Campbell, K. J. (2022). A thematic cluster analysis of parents' online discussions about fussy eating. *Maternal & Child Nutrition*, *18*, e13316. https://doi.org/10.1111/mcn.13316

[Text pre-processing]

16) Yadav, D. (2020, April 6). NLP: Building Text Cleanup and PreProcessing Pipeline. https://towardsdatascience.com/nlp-building-text-cleanup-and-preprocessing-pipeline-eba4095245a0

[Clustering]

17) Saji, B. (2023, April 26). Elbow Method for Finding the Optimal Number of Clusters in K-Means. https://www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/#What_Is_the_Elbow_Method_in_K-Means_Clustering?

18) Tomar, A. (2022, November 18). Stop Using Elbow Method in K-means Clustering, Instead, Use this! https://towardsdatascience.com/elbow-method-is-not-sufficient-to-find-best-k-in-k-means-clustering-fc820da0631d

19) Hu, Y. X., Li, K., Meng, A. (2018, December 7). Agglomertive Hierarchical Clustering using Ward Linkage.https://jbhender.github.io/Stats506/F18/GP/Group10.html#:~:text=There%20are%20four%20methods%20for,analyzes%20the%20variance%20of%20clusters.

[LDA]

20) Asmussen, C.B., Møller, C. Smart literature review: a practical topic modelling approach to exploratory literature review. *J Big Data* **6**, 93 (2019). https://doi.org/10.1186/s40537-019-0255-7

21) Sutherland, I., Sim, Y., Lee, S. K., Byun, J., & Kiatkawsin, K. (2020). Topic Modeling of Online Accommodation Reviews via Latent Dirichlet Allocation. *Sustainability*, *12*(5), 1821. https://doi.org/10.3390/su12051821

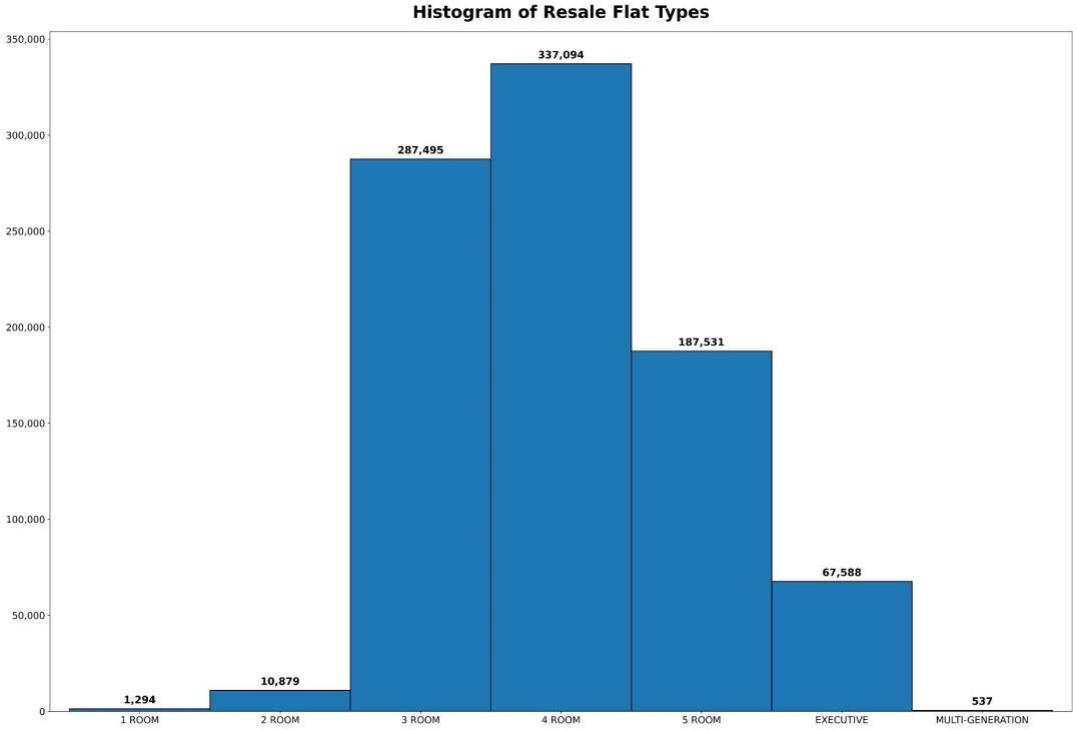
[BERTopic]

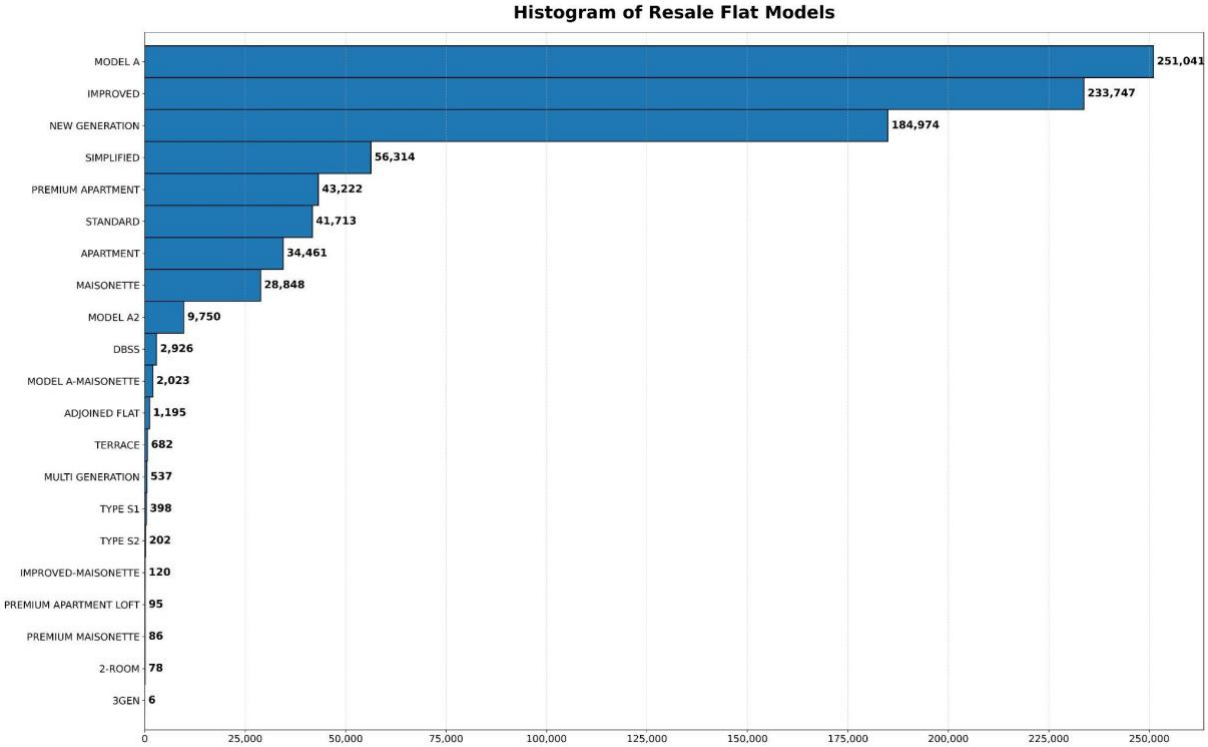
22) Andronikou, K. (2022, October 21). An In-Depth Introduction to Topic Modelling Using LDA and BERTopic.https://www.theanalyticslab.nl/an-in-depth-introduction-to-topic-modeling-using-lda-and-bertopic/

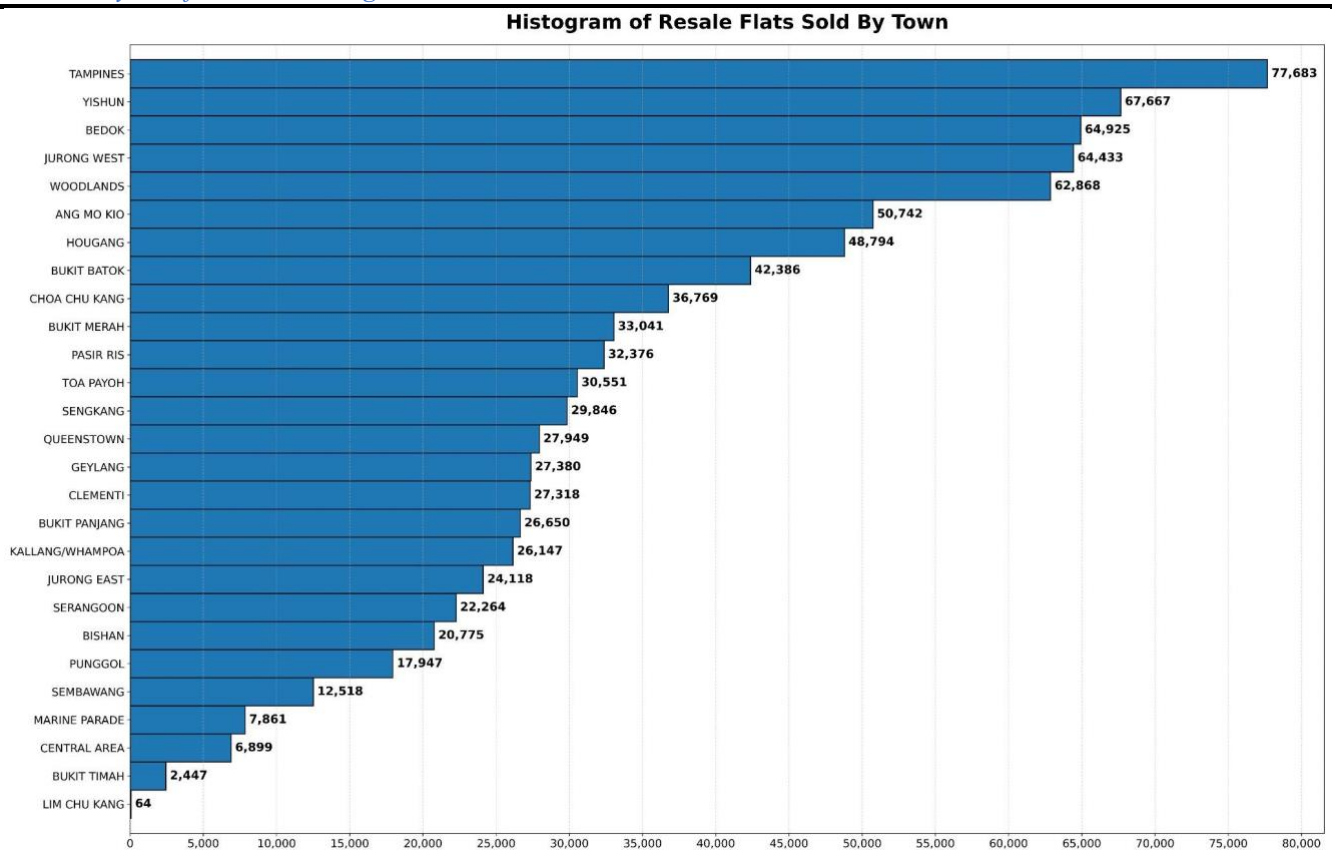
23) Teng, A. (2023, February 14). Topic Modeling with BERT. https://medium.com/@angelamarieteng/topic-modeling-with-bert-2e3218723373

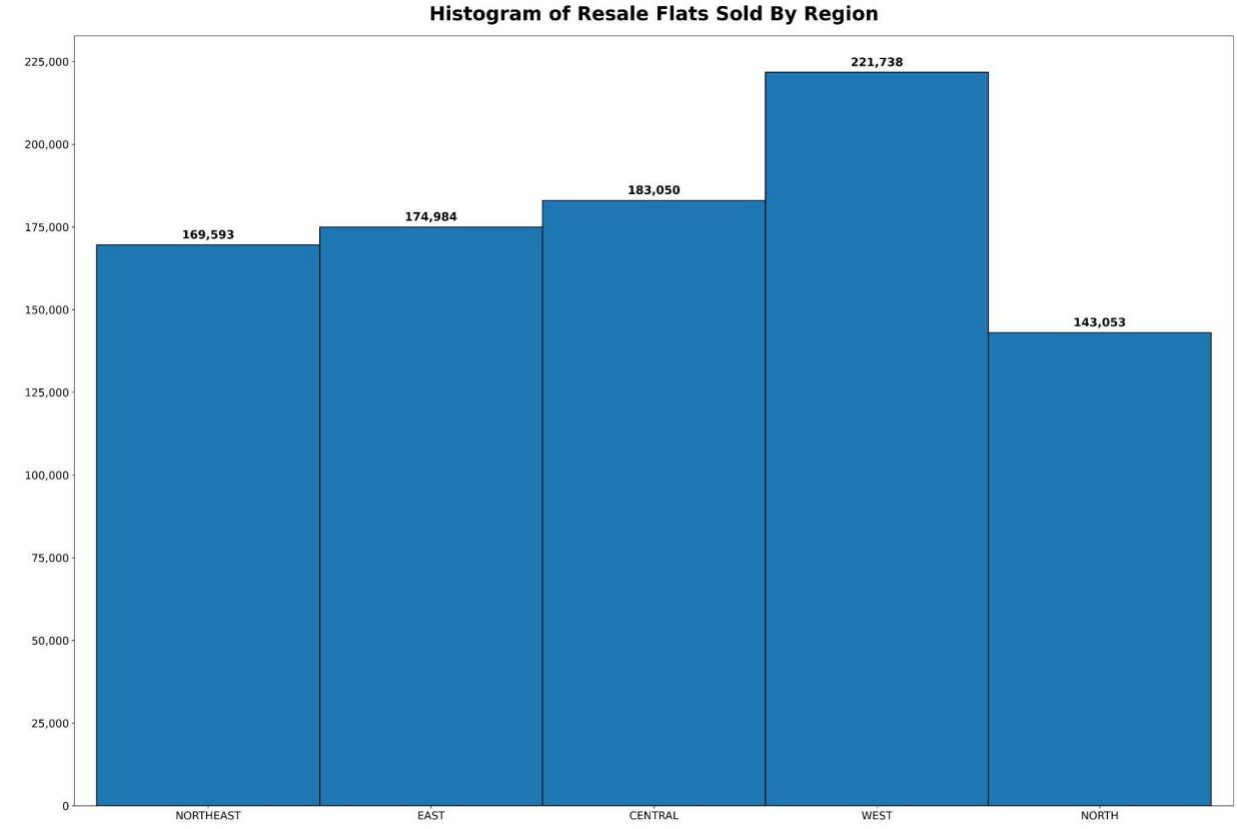
Appendix

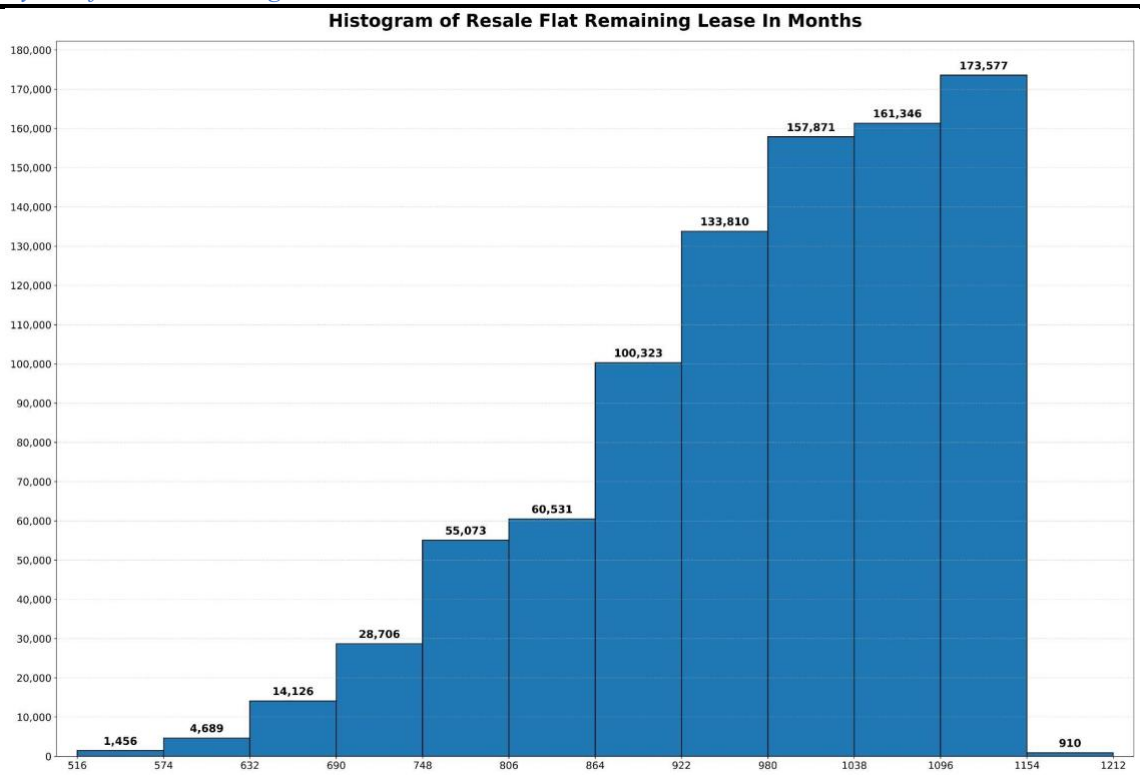
Exploratory Data Analysis (EDA) Using Histogram Plots

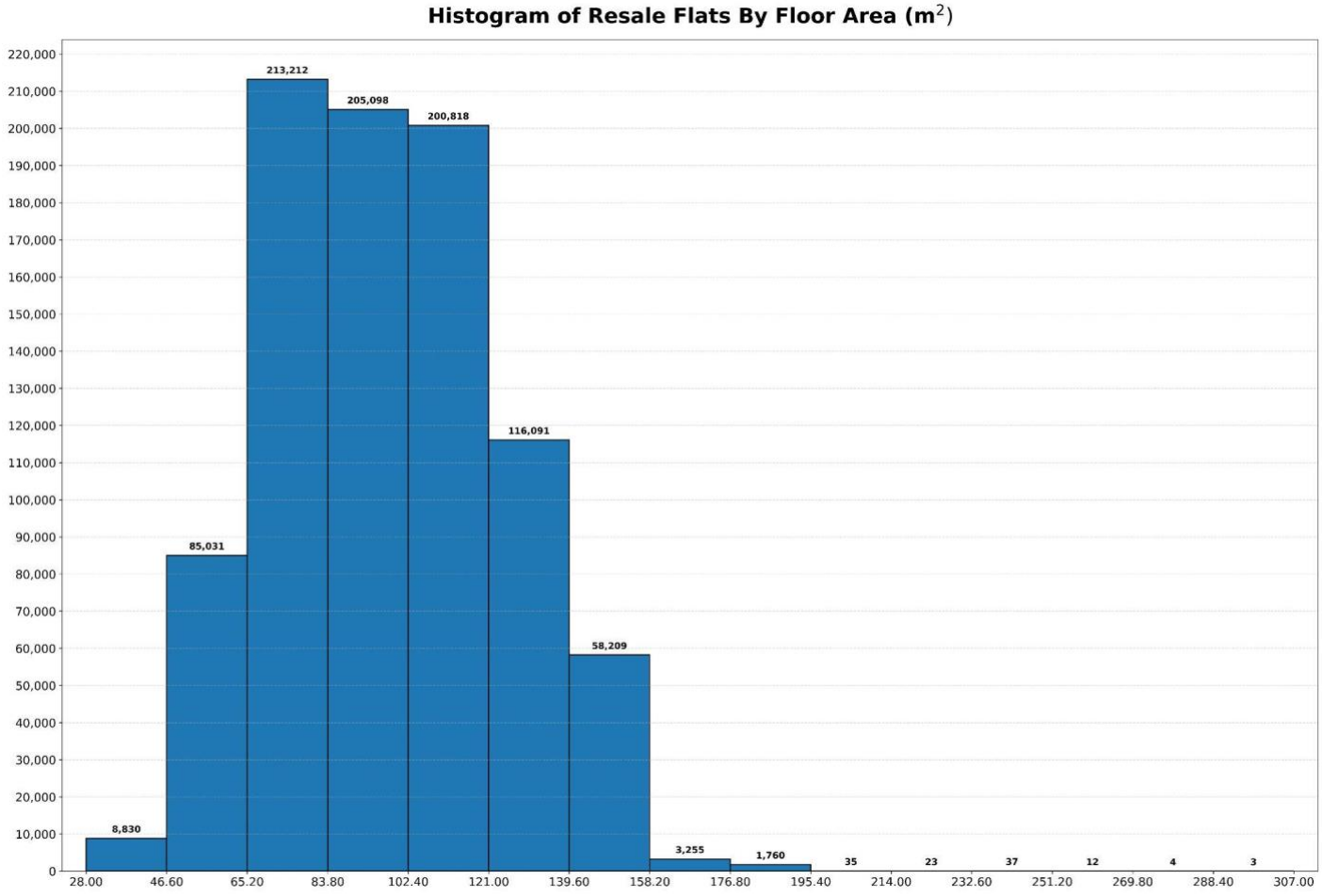


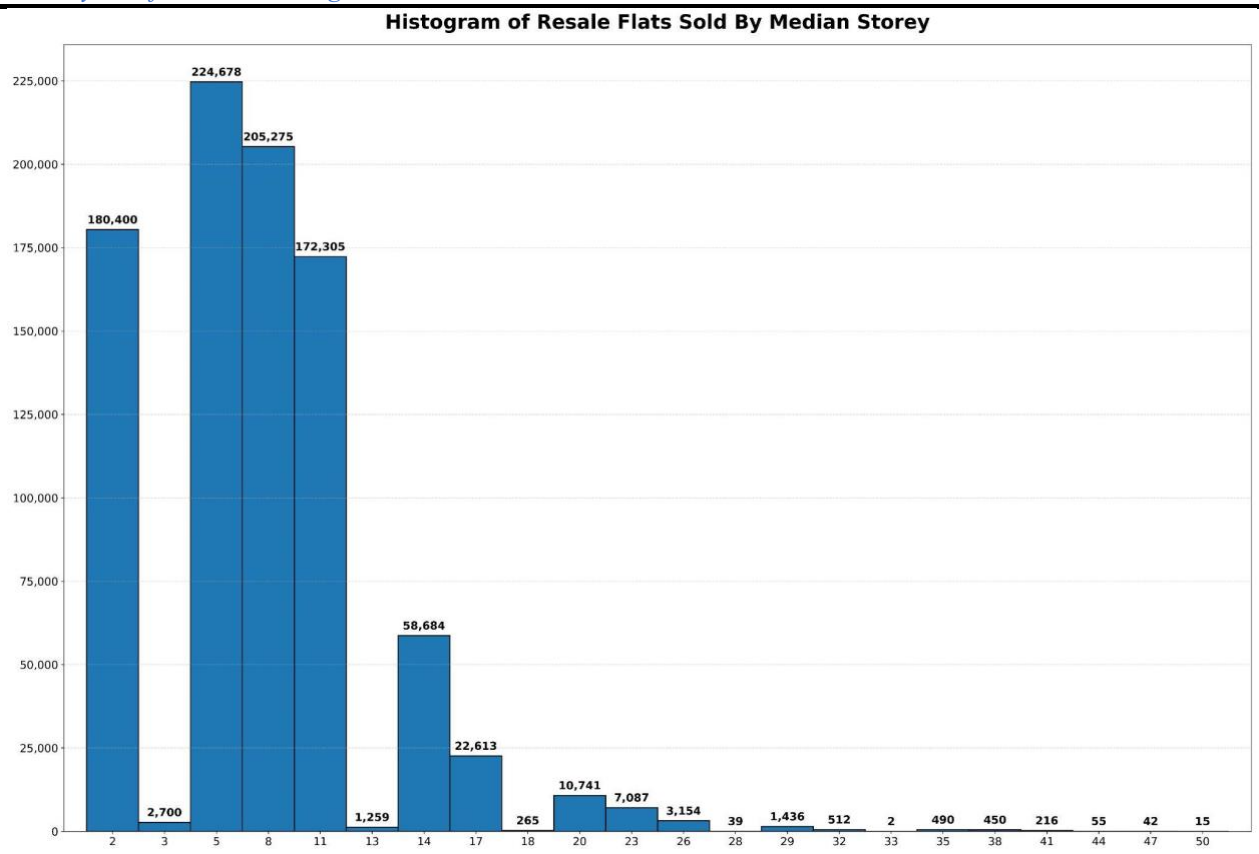


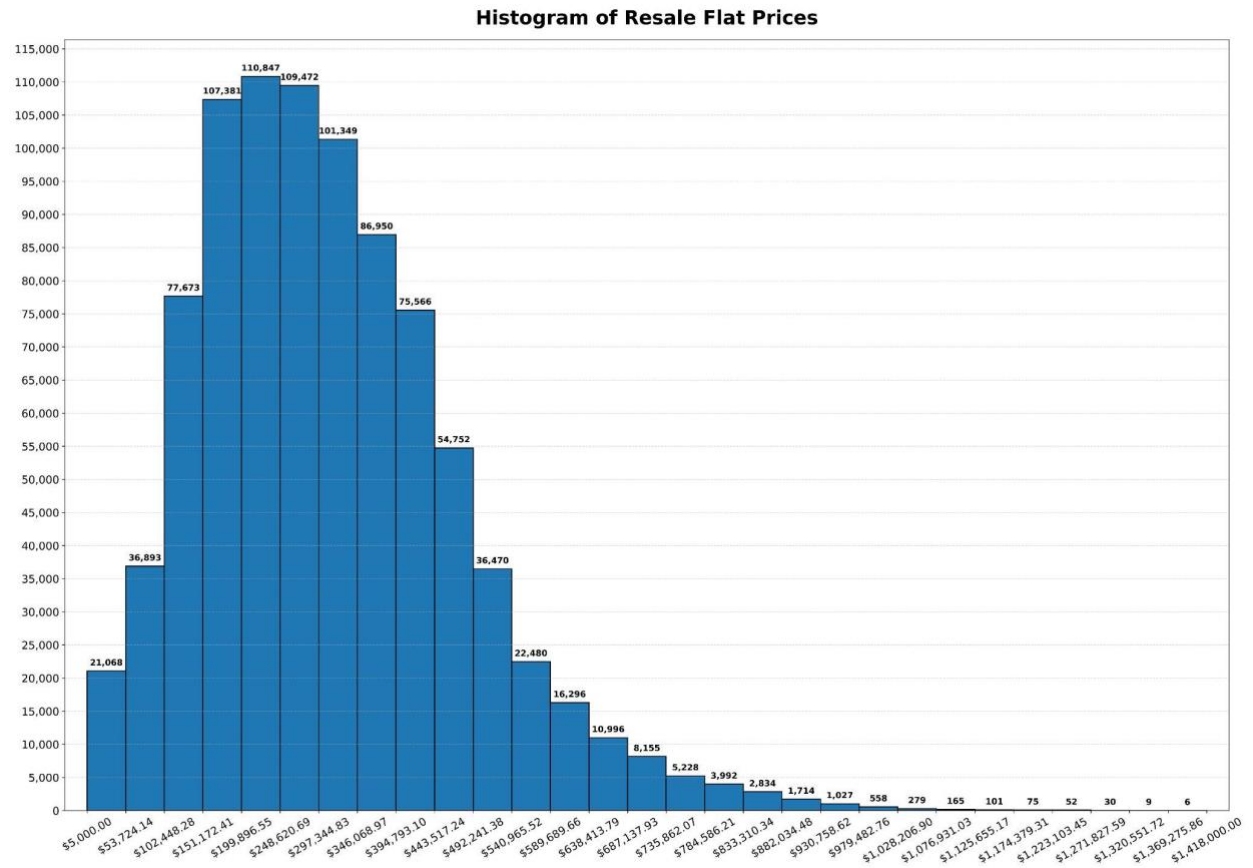

Histogram of Resale Flats Sold By Town
TAMPINES 77,683
YISHUN 67,667
BEDOK 64,925
JURONG WEST 64,433
WOODLANDS 62,868
ANG MO KIO 50,742
HOUGANG 48,794
BUKIT BATOK 42,386
CHOA CHU KANG 36,769
BUKIT MERAH 33,041
PASIR RIS 32,376
TOA PAYOH 30,551
SENGKANG 29,846
QUEENSTOWN 27,949
GEYLANG 27,380
CLEMENTI 27,318
BUKIT PANJANG 26,650
KALLANG/WHAMPOA 26,147
JURONG EAST 24,118
SERANGOON 22,264
BISHAN 20,775
PUNGGOL 17,947
SEMBAWANG 12,518
MARINE PARADE 7,861
CENTRAL AREA 6,899
BUKIT TIMAH 2,447
LIM CHU KANG 64
0 5,000 10,000 15,000 20,000 25,000 30,000 35,000 40,000 45,000 50,000 55,000 60,000 65,000 70,000 75,000 80,000

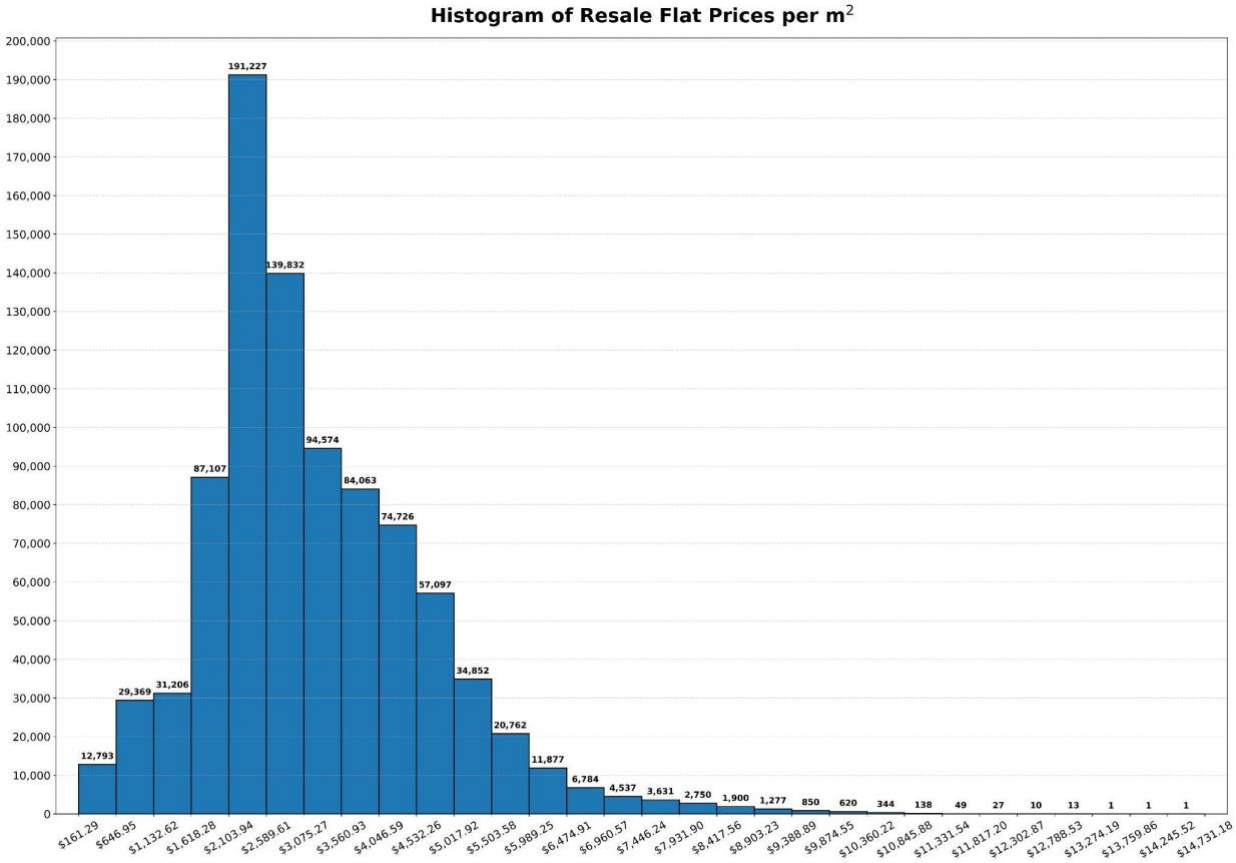

Histogram of Resale Flats Sold By Region
225,000 200,000 175,000 150,000 125,000 100,000 75,000 50,000 25,000 0
169,593
174,984
183,050
221,738
143,053
NORTHEAST EAST CENTRAL WEST NORTH


Histogram of Resale Flat Remaining Lease In Months
1,456
4,689
14,126
28,706
55,073
60,531
100,323
133,810
157,871
161,346
173,577
910
516
574
632
690
748
806
864
922
980
1038
1096
1154
1212


Histogram of Resale Flats By Floor Area (m^2)
8,830
85,031
213,212
205,098
200,818
116,091
58,209
3,255
1,760
35
23
37
12
4
3
28.00
46.60
65.20
83.80
102.40
121.00
139.60
158.20
176.80
195.40
214.00
232.60
251.20
269.80
288.40
307.00


Histogram of Resale Flats Sold By Median Storey
225,000
200,000
175,000
150,000
125,000
100,000
75,000
50,000
25,000
0
180,400
2,700
224,678
205,275
172,305
1,259
58,684
22,613
265
10,741
7,087
3,154
39
1,436
512
2
490
450
216
55
42
15
2
3
5
8
11
13
14
17
18
20
23
26
28
29
32
33
35
38
41
44
47
50


Histogram of Resale Flat Prices
115,000
110,000
105,000
100,000
95,000
90,000
85,000
80,000
75,000
70,000
65,000
60,000
55,000
50,000
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
21,068
36,893
77,673
107,381
110,847
109,472
101,349
86,950
75,566
54,752
36,470
22,480
16,296
10,996
8,155
5,228
3,992
2,834
1,714
1,027
558
279
165
101
75
52
30
9
6
$5,000.00
$53,724.14
$102,448.28
$151,172.41
$199,896.55
$248,620.69
$297,344.83
$346,068.97
$394,793.10
$443,517.24
$492,241.38
$540,965.52
$589,689.66
$638,413.79
$687,137.93
$735,862.07
$784,586.21
$833,310.34
$882,034.48
$930,758.62
$979,482.76
$1,028,206.90
$1,076,931.03
$1,125,655.17
$1,174,379.31
$1,223,103.45
$1,271,827.59
$1,320,551.72
$1,369,275.86
$1,418,000.00



Titles of HardwareZone Threads & Reddit Posts With More Than A Hundred Comments

S/N	Title of Thread or Post	Number of Comments	Date of creation	Source
1	Which Direction Will Property Prices Go ?	2353	26/9/2020	hwz
2	Government imposes new property cooling measures: Maximum loan quantum limits tightened, loan-to-value limit lowered - CNA	164	29/9/2022	reddit
3	[BREAKING] Housing loan limits - including for HDB loans - tightened to ensure prudent borrowing, moderate demand in property market	721	30/9/2022	hwz
4	[serious] instead of building more hdb to support the increase in population. yet more cooling measures when underlying	170	30/9/2022	hwz

	demand still unresolved.			
5	Condo owners shelving plans to downgrade, others who sold homes are stuck, say agents after cooling measures	275	30/9/2022	hwz
6	From today onwards, hdb resale prices will be falling	204	30/9/2022	hwz
7	15-month wait for private home owners to buy resale HDB 'a form of deterrence for buyers with deep pockets'	120	30/9/2022	hwz
8	Hdb flippers are screwed. Private property prices to remain high.	189	30/9/2022	hwz
9	New Cooling Measures 30 Sept 2022	247	30/9/2022	hwz
10	[GLGT]"I'm in hot water right now," said a potential buyer who just sold his condominium unit 3 weeks ago	267	1/10/2022	hwz
11	Those take floating house loan how ah? SORA chiong to 4.4% liao.	255	3/10/2022	hwz
12	Don't let people own HDB flat and private property at the same time	260	4/10/2022	hwz
13	'I was so shocked': New property cooling measures scupper some private owners' plans to sell homes, downgrade to HDB	121	5/10/2022	reddit
14	The astronomical increase in 3 room resale prices in the 4 cheapest non-mature estates from '19 to '22 vs wage growth for youths and seniors from '19 to '21	177	8/10/2022	reddit
15	HDB flat owners who go on to purchase a private property should have to sell their HDB flat since they no longer require a subsidised home	185	11/10/2022	hwz

16	[GLGT] Condo, HDB rents up in September; strong demand ensures landlord's market	133	12/10/2022	hwz
17	Will property market see a correction soon with recession and rising interest rates?	119	15/10/2022	hwz
18	[GPGT] Reddit predict SG property market to drop by 30%	145	5/11/2022	hwz
19	So many layoffs will we finally see a property correction of 10-15%?	136	12/11/2022	hwz
20	how prepared are you for 5% mortgage rate by 31st Dec 2022?	166	20/11/2022	hwz
21	CNY I tio stun when i visit my relative 4 room hdb flat	153	24/1/2023	hwz
22	[GLGT]Sinkie will not have to worry about having an affordable home to call their own: PM Lee	155	8/2/2023	hwz
23	Ban HDB from renting, hdb is for staying not for generating revenue.....	158	8/2/2023	hwz
24	HDB BTO oversupply not the answer to home affordability: Desmond Lee	190	8/2/2023	hwz
25	HDB is affordable. Not sarcasm.	310	8/2/2023	hwz
26	[For discussion] HDB housing bubble possible?	103	17/2/2023	hwz
27	BREAKING: Sengkang 5 room HDB SOLD for almost ONE MILLION SGD	155	18/2/2023	hwz
28	What is stopping the government from imposing a price cap on HDB.	145	19/2/2023	hwz

29	5-room Pinnacle@Duxton unit sold for $1.38m, buyer paid about $120k cash over valuation	118	25/2/2023	reddit
30	Lawrence Wong admits the government is keeping HDB prices high by ownself setting the land costs at 'market value' and refuses to lower it.	127	25/2/2023	hwz
31	Brace Yourself: Is the Singapore Property Market About to Crash?	231	1/3/2023	hwz
32	Fearing the worst for SG Property	109	13/3/2023	hwz
33	Homeless soon so hard to buy resale	124	1/4/2023	hwz