PREDICTING HOUSING PRICES AROUND SINGAPORE

A Project Report

s*ubmitted by Team 1*

Nguyen Ngoc Yen Nga	**A0279983B**
Nivash Sudalaimani	**A0280048H**
Shamik Banerjee	**A0276524Y**
Sricharan Sriram	**A0279517N**

Table of Contents:

1. Dataset and Preprocessing

1.1. Description of Datasets

For analysis and prediction of HDB resale and rental prices in Singapore, three main datasets from the open repositories of Singapore Government's public data were employed: one dataset from "*Renting Out of Flats 2023*" for rental price and two datasets from "*Resale Flat Prices*" for resale price.

Renting Out of Flats:

- This dataset contains data on rental prices of flats from January 2021 up to March 2024.
- It consists of 6 columns and 118593 rows, 6 columns include:

Column	*Data type*	*Description*
`rent_approval_date`	Object	Date when rent approved (YYYY-MM)
`town`	Object	Town where flat located
`block`	Object	Block number
`street_name`	Object	Street name
`flat_type`	Object	Type of flat
`monthly_rent`	Integer	Monthly rental price

Resale Flat Prices:

- The two main datasets contain data on resale prices of flats from 2012-2018 to 2019-2024.
- It consists of 9 columns and total 265878 rows, 9 columns include:

Column	*Data type*	*Description*
`month`	Object	Data of transaction (YYYY-MM)
`town`	Object	Town where flat located
`flat_type`	Object	Type of flat
`block`	Object	Block number
`street_name`	Object	Street name
`storey_range`	Object	Range of floors
`floor_area_sqm`	Float	Area of flat in square meters
`flat_model`	Object	Model of flat
`lease_commence_date`	Object	Year of the commenced lease
`resale_price`	Float	Resold price

For resale price's initial analysis, other historical datasets including periods from 1990-1999 and 2000-2011 were also concatenated to provide a comprehensive overview and visualization of the market trends and dynamics. However, due to the vast amount of data and the changing market conditions, for the resale price prediction, we opted to focus on more recent periods (2012-2018 and 2019-2024) to reflect the current market more accurately as well as to allocate the computational resources more efficiently.

Additional datasets were also employed to enhance the models' predictive power:

- MRT stations (existing and planned): Location data to calculate proximity to the current and future public MRT.
- Primary schools and shopping malls: Location data to calculate proximity to the amenities that could impact on flat prices due to its convenience factors.
- Zip code mapper: Mapping of street names and block numbers to geographic coordinates to facilitate grouping according to regions.

1.2. Preprocessing Steps

1.2.1. Cleaning Steps

Datasets for both resale and rental price analysis were examined to ensure no missing data, no duplicates, and consistent labels of categorical data. IQR (Interquartile Range) method was conducted to ensure that the percentage of outliers for each feature is within reasonable range.

1.2.2. Feature Engineering

New features were implemented to enhance the models and to reflect the evolving market conditions:

- Remaining lease: Calculated by subtracting current year from lease commence date plus the original lease term (99 years) to provide insight regarding property's value depreciation.
- Proximity to amenities: Used geographic data to compute approximate distance to the nearest MRT stations, primary schools, and malls.
- Regional mapping: Mapped town to its corresponding region to provide broader view of market dynamics between regions.
- CPI adjusted price: Used Consumer Price Index to adjust price to reflect inflation and provide a more accurate temporal comparision.

1.2.3. Transformation

- Dates (YYYY-MM) were converted into datetime format and to extract the year of sale feature.
- Categorical features (such as `flat_type`, `storey_range`, `flat_model`, `town`, `region`, ...) are converted from textual to an appropriate numeric format to ensure compatibility for machine learning models.

- Continuous numerical features (such as `floor_area_sqm`, `mrt_dist`, `upcoming_mrt_dist`, `school_dist`, `mall_dist`, ...) are scaled using standard scaler.
- Cyclical feature (`month`) was transformed into sine and cosine values to reflect its cyclical nature.

2. Exploratory Analysis Insights

2.1. Overview of Combined Resale and Rental Price Analysis (2021-2023)

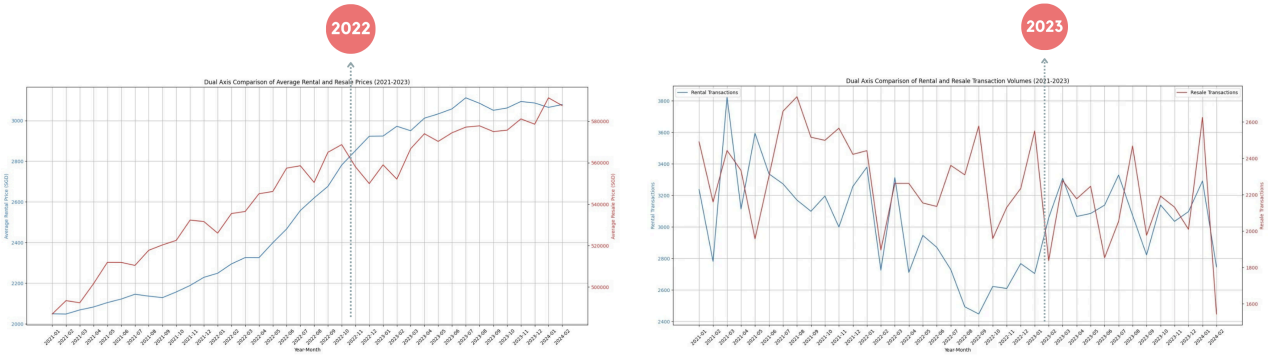

Figure 1. Dual-axis Crossover Graphs for Rental vs Resale Prices and Transactions (2021-2023)

The dual-axis graph displays a general upward trajectory in both rental and resale prices over the observed period. For transaction volumes, the dual-axis graph shows more dynamic movements with variances in which could be related to seasonal or broader economic factors.

Around October 2022, a significant event is reflected in the inflection point in prices. The transaction volume trend until 2023 shows a pattern where individuals seem to be purchasing homes and exiting the rental market, which might indicate a market that favored buying over renting. The sharp rise in the number of rental transactions post-2022 might witness the recovery phase where there is a surge in non-residents population[1] due to easing of pandemic-related restrictions[2]. The resale trends might reflect the market saturation post sellouts of 1000+ unit mega-developments (e.g. Treasure at Tampines, Normanton Park,...) in previous year and 2022 saw no new launches of equally significant size)[3].

2.2. Resale Price Analysis

2.2.1. Visualizations

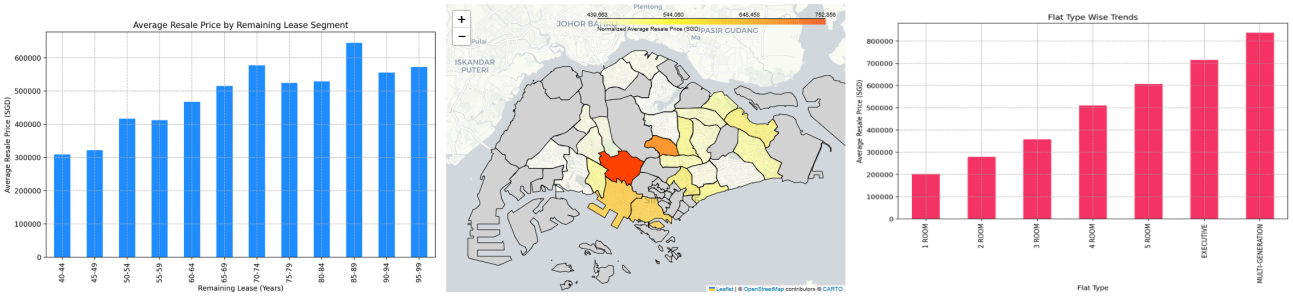

Figure [...]. Multifaceted View of Resale Market

The initial exploratory data analysis presents a multifaceted view of the resale market, illustrating clear trends and distribution of prices among various flat types and towns:

- The first bar graph shows a direct positive correlation between the remaining lease and the resale prices, which shows the general trend that more the remaining lease for a flat, the better the resale price is.
- The second visualization is that of a geospatial distribution, which gives us a better understanding of how towns and regions play a vital role in determining resale prices. As seen in the plot, the towns located in the Central region (Bukit Timah, Bishan, CWC, Queenstown) have higher resale prices compared to the others. (This is plotted for the time period 2019-2024)
- The third plot visualizes the trends between flat types and the average resale prices, which indicates that there is a direct correlation - more the rooms, more the resale prices (which can be intuitive, but the plot validates this hypothesis)

2.2.2. Statistical Analysis

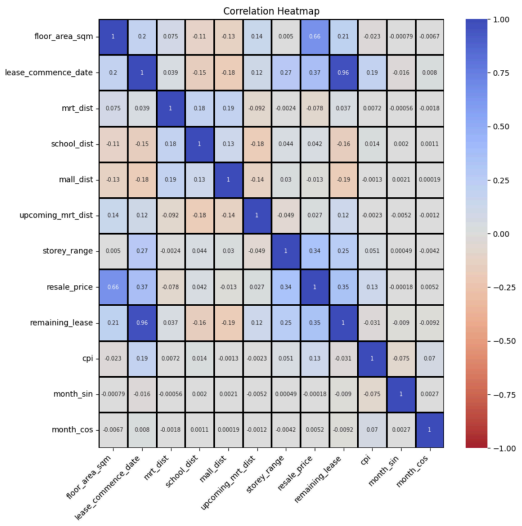

Figure 2. Correlation Map for Resale Features

The correlation analysis reveals several significant relationships between features that could influence the strategy to implement predictive models:

- A notably strong correlation (0.963) between lease commence date and remaining lease indicates that these features could be streamlined to enhance efficiency.
- The correlation between year and CPI (0.797) could reflect the inflation or other economic changes over time that could affect the property values.
- The substantial correlation (0.656) between floor area and resale price underscores the intuitive market dynamics in which larger flats tend to have higher resale prices.
- Proximity features such as MRT distance and Mall distance displays moderate correlations with each other (0.189), and their individual correlations with resale prices even though less prominent but still notable, suggesting a certain modest influence on the flat valuations.

2.2.3. Initial Findings

The analysis highlights the importance of floor area in predicting the resale prices. The insights from the analysis indicates a necessity for further feature selection (drop redundant features) or for regularization techniques (Lasso/Ridge) to handle multicollinearity, especially between highly correlated features, to stabilize the models and enhance further generalizability.

2.3. Rental Price Analysis

2.3.1. Visualizations

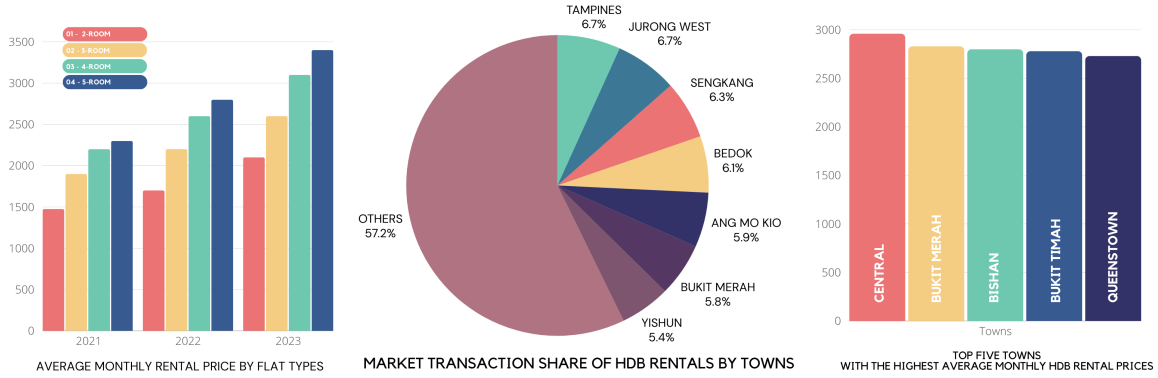

Figure [...]. Multifaceted View of Rental Market

The initial exploratory data analysis presents a multifaceted view of the rental market, illustrating clear trends and distribution of prices among various flat types and towns:

- The first bar graph shows a direct positive correlation between the flat size and the rental prices where more rooms in a flat indicate higher rents.
- The pie chart emphasizes the notable market shares of key major towns in the number of rental transactions, as nearly half of all transactions occur in Tampines, Jurong West, Sengkang, Bedok, Ang Mo Kio, Bukit Merah, and Yishun. This highlights the importance of location in rental popularity.
- The final bar chart identifies the top five most expensive regions for rentals, showing the central region with the highest concentration of highest rental costs, indicating the impact of regional preference in influencing the rental landscape.

2.3.2. Statistical Analysis

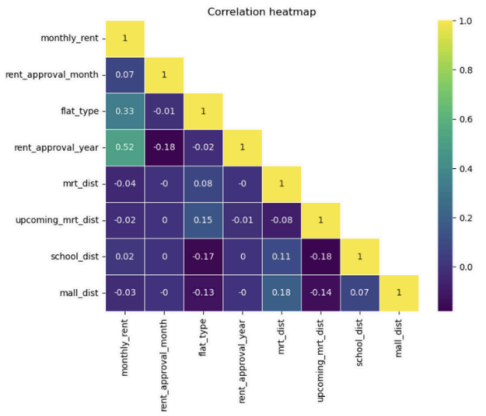

Figure 2. Correlation Map for Resale Features

The correlation analysis shows a relative low linear relationship between the features in the rental dataset:

- There is a moderate positive correlation (0.40) between rental approval year and monthly rental price. This could indicate an annual increase in rental prices, however, this alone is not sufficient enough for accurate predictions.
- The weak negative correlation between monthly rental price and proximity to MRT stations (`mrt_dist`: -0.05) suggests that being farther from MRT might slightly reduce price.
- In contrast, a slight positive correlation with `school_dist` (0.02) suggests that being closer to educational facilities might indicate higher rents.

Nevertheless, overall, the lack of stronger correlations between key factors indicates that the current rental dataset might not be able to fully capture the impactful features influencing rental prices, which in turn limits the predictive strength of the models.

2.3.3. Initial Findings

The initial exploratory analysis of the rental dataset underscores the important role of flat size and location in affecting the rental prices as larger flats and central regions align with higher rents. However, the statistical analysis reveals weak correlations between features, which might suggest an insufficient capture of influential features or certain data quality issues. The current rental dataset might not be robust enough for effective implementation of predictive models. This entails a necessity for richer data that could further capture the complexities of the rental market.

3. Approach for the Prediction Task

3.1. Resale Price Prediction

3.1.1 Feature Selection

Based on the correlation analysis and initial findings, key steps in feature selection are implemented to ensure that selected features are relevant to target and that highly correlated features are avoided.

Features selected:

- Geographical features: Proximity measures (`mrt_dist`, `upcoming_mrt_dist`, `school_dist`, `mall_dist`) are kept as these reflect key factors that potential buyers consider due to the convenience factors.
- Physical feature: `floor_area_sqm` is included to reflect the size of the flat.
- Temporal features: `month` and `remaining_lease` are transformed to reflect its cyclical effect and the decreasing value of lease over time, respectively.
- Categorical features: `flat_type` is included to capture the flat design variations; `town`, `region` are included to reflect the location-based price variations.

Features excluded:

- `lease_commence_date` and `year`: removed to avoid high correlation with `remaining _lease`
- `street_name` and `block`: removed to avoid overfitting to narrow geographical data

Validation technique: The dataset was partitioned into training and testing sets using `train_test_split`.

3.1.2. Model Development

The models tested ranged from simple linear models to more complex ensemble methods. Linear Regression is initially used, then being compared with Ridge Regression to test the multicollinearity through regularization. Other advanced trees-based models like RandomForestRegressor and AdaBoostRegressor are also used to capture the complex nonlinear relationships between features. GridSearchCV is used whenever possible considering computation expense, specifically for tuning parameters like number of neighbors in KNeighborsRegressor.

Model	*Tuning parameters*	*Model's strengths*	*Model's weaknesses*	*Implementation notes*
Linear Regression	OLS-based	Simple to implement and interpret	Sensitive to outliers	Applied log transformation to the target variable
Ridge Regression	`alpha`=0.1	Reduce multicollinearity by penalization	Not suitable for large number of categorical features	Used GridSearchCV for tuning `alpha`
K Neighbors Regressor	`n_neighbors`=8	Non-parametric and flexible	Computationally expensive	Scaled features were essential
Random Forest Regressor	`min_samples_leaf`=20, `max_depth`=50	Handle non-linear data well	Can be overfitting if not tuned properly	
AdaBoost Regressor	no max depth	Improve accuracy by combining weak learners	Sensitive to noisy data and outliers	Used with Decision Tree Regressor as the base estimator
Gradient Boosting Regressor	`n_estimators`=100	Robust to outliers and scalable	Can be slow to train on large datasets	
Voting Regressor	Weight of each model	Reduce variance by averaging multiple models	Less interpretable	Combined RandomForest and AdaBoost
Stacking Regressor	Models used as estimators and final_estimator	Leverage strengths of individual models	Complex model structure	Used a mix of linear and tree-based models

3.1.3. Performance Metrics

Models are evaluated based on R2 Score and RMSE to compare accuracy and error.

Model	*R2 Score*	*RMSE*	*Performance evaluation*	*Potential improvements*
Linear Regression	0.873	0.108 (of log)	Good fit for the data with high R2	Feature selection to reduce noise and improve simplicity
Ridge Regression	0.873	0.108 (of log)	Similar to Linear Regression, robust to collinearity	Hyperparameter tuning to optimize regularization strength
K Neighbors Regressor	0.747	0.174 (of log)	Lower R2, model struggles with complex patterns	Increase neighbors or adjust weights
Random Forest Regressor	0.97	30858.924	Excellent R2, high variance capture	Adjust tree depth and number of trees to reduce overfitting
DecisionTreeRegressor	0.945	34351.3	High R2 but potentially overfitted	Pruning and setting max depth
AdaBoost Regressor	0.968	26408.929	High R2, improves over baseline decision trees	Adjust learning rate and base estimator complexity
Gradient Boosting Regressor	0.892	48202.569	Moderate R2, needs careful tuning	Increase number of boosting stages, tune learning rate
Voting Regressor	0.969	25511.236	High R2, ensemble strength	Optimize weights and explore other combining methods
Stacking Regressor	0.962	28645.890	High R2, complex model structure	Fine-tune base and final estimators, cross-validate more

3.2. Rental Price Prediction

3.2.1. Feature Selection

Based on the correlation analysis and initial findings, key steps in feature selection are implemented to ensure that selected features are relevant to target and that highly correlated features are avoided.

Features selected:

- Geographical features: Proximity measures (`mrt_dist`, `upcoming_mrt_dist`, `school_dist`, `mall_dist`) are kept as these reflect key factors that potential buyers consider due to the convenience factors.
- Temporal features: `month` is transformed to reflect its cyclical effect, and 'year' for the annual effect.
- Categorical features: `flat_type` is included to capture the flat design variations; `town`, `region` are included to reflect the location-based price variations.

Features excluded:

- `street_name` and `block`: removed to avoid overfitting to narrow geographical data

Validation technique: The dataset was partitioned into training and testing sets using `train_test_split`.

3.2.2. Model Development

Akin to the resale price predictions, the models tested ranged from simple linear models to more complex ensemble methods. Linear Regression is initially used, then being compared with Ridge Regression to test the multicollinearity through regularization. Other advanced trees-based models like RandomForestRegressor and AdaBoostRegressor are also used to capture the complex nonlinear relationships between features.

GridSearchCV is used whenever possible considering computation expense, specifically for tuning parameters like number of neighbors in KNeighborsRegressor.

Model	*Tuning parameters*	*Model's strengths*	*Model's weaknesses*	*Implementation notes*
Linear Regression	OLS-based	Simple to implement and interpret	Sensitive to outliers	Applied log transformation to the target variable
Ridge Regression	`alpha`=0.1	Reduce multicollinearity by penalization	Not suitable for large number of categorical features	Used GridSearchCV for tuning `alpha`
K Neighbors Regressor	`n_neighbors`=10	Non-parametric and flexible	Computationally expensive	Scaled features were essential, used GridSearchCV for best params
Random Forest Regressor	`n_estimators`=100	Handle non-linear data well	Can be overfitting if not tuned properly	
AdaBoost Regressor	`n_estimators`=100 `learning_rate`=0.1	Improve accuracy by combining weak learners	Sensitive to noisy data and outliers	Used with Decision Tree Regressor as the base estimator, used GridSearchCV for best params
Gradient Boosting Regressor	`n_estimators`=300 `learning_rate`=0.1 `max_depth`=5	Robust to outliers and scalable	Can be slow to train on large datasets	Used GridSearchCV for best params

3.2.3. Performance Metrics

Models are evaluated based on R2 Score and RMSE to compare accuracy and error.

Model	*R2 Score*	*RMSE*	*Performance evaluation*	*Potential improvements*
Linear Regression	0.475	519.98	Good fit for the data.	Feature selection to reduce noise and improve simplicity
Ridge Regression	0.474	523.92	Similar to Linear Regression, robust to collinearity	Hyperparameter tuning to optimize regularization strength
K Neighbors Regressor	0.495	517.55	Higher R2, model seems to work well for the data	Increase neighbors or adjust weights
Random Forest Regressor	0.477	518.71	Good R2, high variance capture	Adjust tree depth and number of trees to reduce overfitting
AdaBoost Regressor	0.432	543.27	Low R2, potentially due to more columns and lesser data	Adjust learning rate and base estimator complexity
Gradient Boosting Regressor	0.555	484.9	Best R2, with tuning	Increase number of boosting stages, tune learning rate

4. Reflection and Conclusion

4.1. Inferences

4.1.1. Inferences for Resale Price Prediction

- Region seems to play a very important role in resale prices, with high feature importance shown in linear regression
- Towns, floor area, flat models, flat types and amenities come next, with significant coefficients
- CPI, remaining lease and month are the next most important features
- Random Forest Regressor brought out the best predictions with the highest R2 value of 0.97
- Linear Regression by itself was great in accounting for the resale price behavior, with a good R2 score of 0.873
- Boosted trees and CV showed a good performance but not a good increase in R2, owing to the already well performing base models

4.1.2 .Inferences for Rental Price Prediction

- Region seems to play a very important role in rental prices, with high feature importance shown in linear regression
- Features such as nearest amenities had little influence on improving predictive capability of our models
- CPI which was expected to play a role in impacting rental prices surprisingly did not affect much
- XGBoost Regressor relatively performed the best out of all the models with a R2 score of 0.555
- Mostly every other model had around the same R2 scores, typically in the range of 0.4 - 0.5

- AdaBoost Regressor proved to be the worst model for rental price prediction.

4.2. Key Learnings

- **Importance of Data Preprocessing**: The extensive data cleaning and feature engineering steps undertaken ensured the reliability of the predictive models. Handling of outliers, transformation of categorical features into numeric formats, and adjustments for inflation (CPI adjusted price) were crucial for accurate predictions.
- **Integration of Diverse Data Sources**: The project adeptly combined various data sources from external sources. Identified and merged relevant features from disparate datasets to create a unified dataset that better represented the variables influencing housing prices. This integration underscored the value of diverse datasets in strengthening real estate predictive modeling.
- **Impact of Location and Amenities**: The significant impact of geographical factors on housing prices. Proximity to amenities such as MRTs, schools, and malls proved to be influential in both resale and rental markets, showcasing the importance of location in housing price determination.
- **Model Complexity vs. Performance**: The exploration of various modeling techniques from simple linear regression to more complex ensemble methods demonstrated a balance between model complexity and predictive performance. This revealed the importance of selecting the right model based on the nature of data and the prediction task.
- **Predictive Power of Different Features:** Different features displayed varying degrees of influence on housing prices. For example, the remaining lease period and the floor area were key predictors in resale price models, while the size of the flat and its location were more influential in rental price predictions.
- **Learning from Model Comparisons**: Comparing different models provided insights into how different algorithms handle biases and variance. For instance, ensemble methods like Random Forest and XGBoost generally offered better performance by effectively capturing complex patterns in the data.
- **Challenges in Predictive Modeling**: Dealing with multicollinearity among features ensured that models did not overfit. Regularization techniques and careful feature selection were necessary to address these issues.

4.3. Future Directions

- More amenities can be explored, such as hawker centers, hospitals, local attractions, etc
- Making predictions with the behavioral pattern of the specific individual/family searching
- Using APIs such as OpenStreetMaps for obtaining accurate driving distances instead of Haversine
- A more intelligent implementation of nearest amenities
    - Is the nearest MRT for a house the same in all years?
    - Chances that the nearest MRT might not have existed in that year
- Upcoming projects in areas in proximity to the housings can be analyzed

5. References

1. Dataset sources:
    a. HDB Resale and Rental Data - data.gov.sg
    b. Existing and planned MRT stations, primary schools and malls - data.gov.sg
    c. CPI - SingStat
    d. Coordinates of various amenities and HDBs - OpenStreetMaps and Google Maps

5.1. Citations

1. Population in Brief 2021: Key Trends, Population in Brief 2022: Key Trends, Population in Brief 2023: Key Trends
2. 14 Current Mega Developments And How They've Performed (1,000 Units And Above) (thefinance.sg)
3. Singapore COVID - Coronavirus Statistics - Worldometer (worldometers.info)