IE1111R Industrial & Systems Engineering Principles & Practice I

By Group 5

Student Name	Student ID
Manish Monappa Kulal	A0301710B
Lim Xi	A0308478R
Lee Kang Hong	A0307058E
Neal Ng Seng Yew	A0307454E
Soh Wei Jie	A0307845X

Table of Contents

1. Introduction

1.1 Problem Statement

The Housing Development Board market is a cornerstone of Singapore's real estate market with its growth constantly on the rise. The high demand for HDB resale flats is fueled in part due to strong broad-based demand as well as supply tightness in the market. Just like any market, it is also influenced by a range of economic, social, and political factors. The volatile nature of HDB resale market thus makes it crucial for us to have a good prediction system model which accounts for all these vulnerabilities to better allow us to understand market trends and make more informed decisions. With homebuyers viewing HDB flats as both a necessity and a long-term investment, the market's behavior has become a key indicator of broader economic health in Singapore.

This report will first guide you through the methodology employed for data collection and cleaning, along with a detailed discussion of our rationale behind selecting and excluding various factors. Next you will learn about how we implemented linear regression to predict resale prices. Linear regression is our method of choice as it has proven an effective method for modelling relationships between a single dependent variable with a plethora of independent variables. Following that, we will outline the process behind our user-friendly interface and the VBA programming behind it. Lastly, we will wrap up with an evaluation showcasing the pros and cons of linear regression and other workarounds to raise accuracy.

1.2 Problem Description

HDB prices are influenced by various factors such as location, remaining lease period, and market demand. These fluctuations can lead to irregular pricing and potential overpricing of flats, complicating the decision-making process for buyers. Although resale prices occasionally dip, understanding the underlying reasons for these changes is crucial for prospective homeowners. From the buyer's perspective, the desire to secure an affordable HDB flat in a prime location is paramount. However, the unpredictability of market trends creates uncertainty about future pricing. Buyers need a reliable method to forecast HDB prices, enabling them to plan their purchases effectively and avoid financial strain.

1.3 Objective

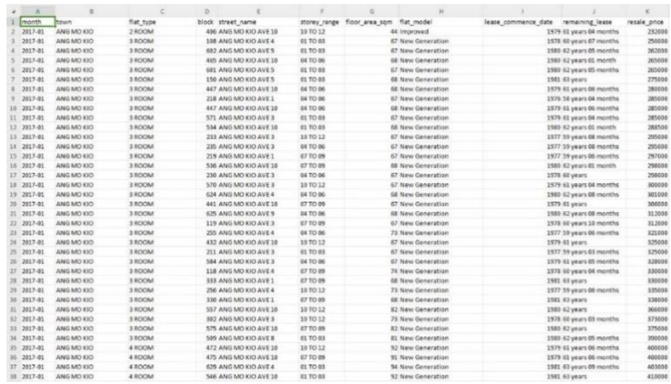
Our objective is to come up with a model that could predict future HDB resale prices based on various factors such as location, lease year commencement etc.

1.4 Target Audience

The primary target audience of our study are individuals who intend to purchase or sell a HDB resale unit. Our decision tool will allow them to gauge the estimated pricing of the HDB unit that they intend to purchase or sell so that it will help them make better informed decisions when purchasing or selling.

2. Data Cleaning and Analysis

2.1 Dataset



	month	town	flat_type	block	street_name	storey_range	floor_area_sqm	flat_model	lease_commence_date	remaining_lease	resale_price
2	2017-01	ANG MO KIO	2 ROOM	406	ANG MO KIO AVE 10	10 TO 12	44	Improved	1979	61 years 04 months	232000
3	2017-01	ANG MO KIO	3 ROOM	108	ANG MO KIO AVE 4	01 TO 03	67	New Generation	1978	60 years 07 months	250000
4	2017-01	ANG MO KIO	3 ROOM	602	ANG MO KIO AVE 5	01 TO 03	67	New Generation	1980	62 years 05 months	262000
5	2017-01	ANG MO KIO	3 ROOM	465	ANG MO KIO AVE 10	04 TO 06	68	New Generation	1980	62 years 01 month	265000
6	2017-01	ANG MO KIO	3 ROOM	601	ANG MO KIO AVE 5	01 TO 03	67	New Generation	1980	62 years 05 months	265000
7	2017-01	ANG MO KIO	3 ROOM	150	ANG MO KIO AVE 5	01 TO 03	68	New Generation	1981	63 years	275000
8	2017-01	ANG MO KIO	3 ROOM	447	ANG MO KIO AVE 10	04 TO 06	68	New Generation	1979	61 years 06 months	280000
9	2017-01	ANG MO KIO	3 ROOM	218	ANG MO KIO AVE 1	04 TO 06	67	New Generation	1976	58 years 04 months	285000
10	2017-01	ANG MO KIO	3 ROOM	447	ANG MO KIO AVE 10	04 TO 06	68	New Generation	1979	61 years 06 months	285000
11	2017-01	ANG MO KIO	3 ROOM	571	ANG MO KIO AVE 3	01 TO 03	67	New Generation	1979	61 years 04 months	285000
12	2017-01	ANG MO KIO	3 ROOM	534	ANG MO KIO AVE 10	01 TO 03	68	New Generation	1980	62 years 01 month	288500
13	2017-01	ANG MO KIO	3 ROOM	233	ANG MO KIO AVE 3	10 TO 12	67	New Generation	1977	59 years 08 months	295000
14	2017-01	ANG MO KIO	3 ROOM	235	ANG MO KIO AVE 3	04 TO 06	67	New Generation	1977	59 years 08 months	295000
15	2017-01	ANG MO KIO	3 ROOM	219	ANG MO KIO AVE 1	07 TO 09	67	New Generation	1977	59 years 06 months	297000
16	2017-01	ANG MO KIO	3 ROOM	536	ANG MO KIO AVE 10	07 TO 09	68	New Generation	1980	62 years 01 month	298000
17	2017-01	ANG MO KIO	3 ROOM	230	ANG MO KIO AVE 3	04 TO 06	67	New Generation	1978	60 years	298000
18	2017-01	ANG MO KIO	3 ROOM	570	ANG MO KIO AVE 3	10 TO 12	67	New Generation	1979	61 years 04 months	300000
19	2017-01	ANG MO KIO	3 ROOM	624	ANG MO KIO AVE 4	04 TO 06	68	New Generation	1980	62 years 08 months	301000
20	2017-01	ANG MO KIO	3 ROOM	441	ANG MO KIO AVE 10	07 TO 09	67	New Generation	1979	61 years	306000
21	2017-01	ANG MO KIO	3 ROOM	625	ANG MO KIO AVE 9	04 TO 06	68	New Generation	1980	62 years 08 months	312000
22	2017-01	ANG MO KIO	3 ROOM	119	ANG MO KIO AVE 3	07 TO 09	67	New Generation	1978	60 years 10 months	312000
23	2017-01	ANG MO KIO	3 ROOM	255	ANG MO KIO AVE 4	04 TO 06	73	New Generation	1977	59 years 06 months	321000
24	2017-01	ANG MO KIO	3 ROOM	432	ANG MO KIO AVE 10	10 TO 12	67	New Generation	1979	61 years	325000
25	2017-01	ANG MO KIO	3 ROOM	211	ANG MO KIO AVE 3	01 TO 03	67	New Generation	1977	59 years 03 months	325000
26	2017-01	ANG MO KIO	3 ROOM	584	ANG MO KIO AVE 3	04 TO 06	67	New Generation	1979	61 years 05 months	328000
27	2017-01	ANG MO KIO	3 ROOM	118	ANG MO KIO AVE 4	07 TO 09	74	New Generation	1978	60 years 04 months	330000
28	2017-01	ANG MO KIO	3 ROOM	333	ANG MO KIO AVE 1	07 TO 09	68	New Generation	1981	63 years	330000
29	2017-01	ANG MO KIO	3 ROOM	256	ANG MO KIO AVE 4	10 TO 12	73	New Generation	1977	59 years 06 months	335000
30	2017-01	ANG MO KIO	3 ROOM	330	ANG MO KIO AVE 1	07 TO 09	68	New Generation	1981	63 years	338000
31	2017-01	ANG MO KIO	3 ROOM	557	ANG MO KIO AVE 10	10 TO 12	82	New Generation	1980	62 years	366000
32	2017-01	ANG MO KIO	3 ROOM	302	ANG MO KIO AVE 3	10 TO 12	73	New Generation	1978	60 years 03 months	373000
33	2017-01	ANG MO KIO	3 ROOM	575	ANG MO KIO AVE 10	07 TO 09	82	New Generation	1980	62 years	375000
34	2017-01	ANG MO KIO	3 ROOM	509	ANG MO KIO AVE 8	01 TO 03	81	New Generation	1980	62 years 05 months	390000
35	2017-01	ANG MO KIO	4 ROOM	472	ANG MO KIO AVE 10	10 TO 12	92	New Generation	1979	61 years 06 months	400000
36	2017-01	ANG MO KIO	4 ROOM	475	ANG MO KIO AVE 10	07 TO 09	91	New Generation	1979	61 years 06 months	400000
37	2017-01	ANG MO KIO	4 ROOM	629	ANG MO KIO AVE 4	01 TO 03	94	New Generation	1981	63 years 09 months	403000
38	2017-01	ANG MO KIO	4 ROOM	546	ANG MO KIO AVE 10	01 TO 03	92	New Generation	1981	63 years	410000

Fig. 1: HDB Resale Dataset

Our dataset contains data on HDB resale flat prices from 2017, updated till September 2024. The dataset consists of 11 data fields, consisting of: Month, Town, Flat Type, Block Number, Street Name, Story Range, Floor area, Flat model, Lease commencement date, remaining lease, and resale price.

The flat type of column indicates the classification of units by room size, the street name indicates the street the HDB is in, the flat model indicates the classification of units by the generation of which the flat was made.

2.2 Deciding on Essential Columns

We decided to remove the street name column from the analysis, as the analysis will be based on the general location of the HDB utilizing the town column, instead of its street. The remaining lease is also not included in the analysis as we noticed that the data field is dynamic and changes upon any updates or recalculations, making it harder to track long-term trends. Instead, we chose to utilize the Lease Commencement Date, as it provides a stable and fixed reference point that does not change over time regardless of any updates or recalculations, allowing for an accurate linear relationship between the Lease and the resale price. Block numbers are removed from analysis, as it has shown little to no correlation to resale prices as seen from Fig. 2.

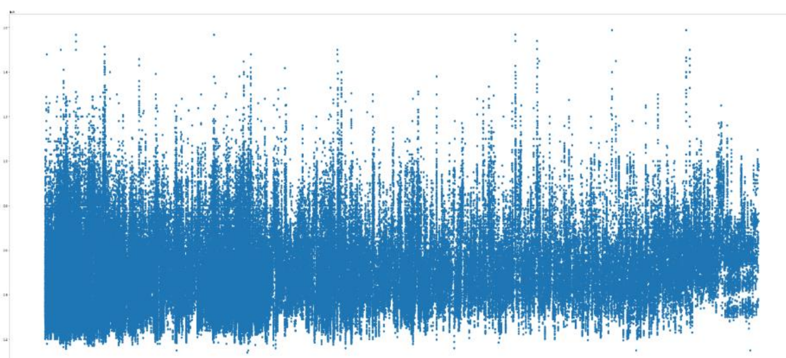


Fig. 2: Scatterplot graph that shows the relationship between Block number and resale price

2.3 Removing Duplicates and converting data into workable forms

After removing unnecessary data fields, we proceeded to remove duplicate rows through excel. Following that, we converted textual data fields like months, flat types, story range, and flat models into binary data fields. Doing so would smoothen the training process of our Linear Regression Model, as these models only work with numerical inputs, and categorizing these fields into binary values allow various categories to be represented as numerical vectors, allowing each feature to be interpreted in relation to its contributions to the model's predictions. We have also decided to split the original month column into 2 different columns, each containing the year and month of resale. This allows us to accurately visualize any seasonal trends and patterns in the resale prices of HDBs, while observing the general trend in resale prices throughout the years.

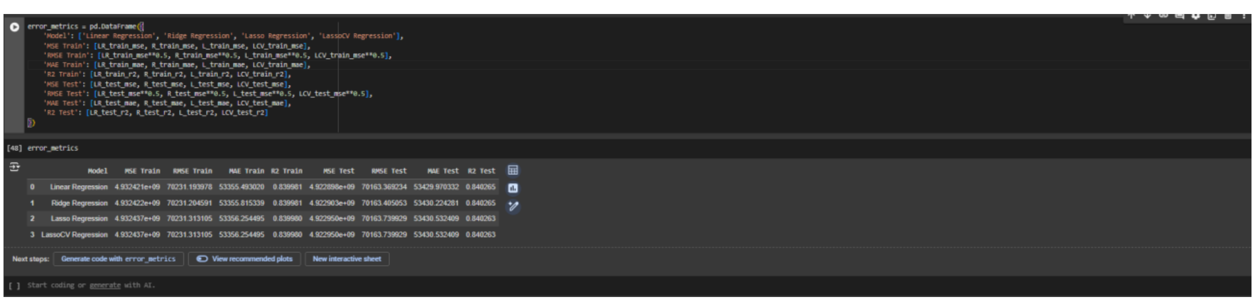
2.4 Utilizing Python and Sci-kit Learning to train Linear Regression Model

We decided to train the model using Python instead of Excel due to Python's ability to take in large datasets, and its ability to split data sets into training and testing datasets. After the completion of data cleaning and reformatting, we converted the refined excel sheet into a CSV file to be imported to Google Collab Notebook. Using Google Collab as an IDE, we imported Pandas library to read the CSV file and display data fields and rows as a dataframe. Utilizing a dataframe allows for the immediate monitoring of existing data following any modifications. We utilized pandas to check for any null values in any of the rows and dropped the rows accordingly.

From Sci-kit Learning libraries, we imported train_test_split function, the Linear Regression Model, and error metrics for the evaluation of model's performances. With the concise dataframe, we categorize the X axis as all independent variables excluding resale price, and Y as resale price. We proceeded to utilize train_test_split function to split X and Y into training and testing datasets, a ratio of 13:7 between training and testing datasets, shuffling the rows in random states of 20. It was essential to split training and testing sets in appropriate ratios such that there is sufficient training data to train the Linear Regression model, but not too much training data to result in an overfitting of the model. We proceeded to train the model and made predictions on resale prices using training X datasets and testing X datasets. Using the predictions, and respective training and testing Y datasets, we utilized error metric functions imported to measure mean squared error, root mean squared error, r2 value, and mean absolute error of the model, to evaluate the model's performance based on actual data. Using Linear Regression built-in methods, we retrieved the intercept and respective coefficients.

2.5 Results and Discussions

Accumulating evaluation values from tested variations of Linear Regression models, we stored all essential values in a pandas dataframe for visual comparisons. We concluded that the values of the original Linear Regression model performed the best, despite all values being relatively on par.



```
error_metrics = pd.DataFrame({
    'Model': ['Linear Regression', 'Ridge Regression', 'Lasso Regression', 'LassoCV Regression'],
    'MSE Train': [LR_train_mse, R_train_mse, L_train_mse, LCV_train_mse],
    'RMSE Train': [LR_train_mse**0.5, R_train_mse**0.5, L_train_mse**0.5, LCV_train_mse**0.5],
    'MAE Train': [LR_train_mae, R_train_mae, L_train_mae, LCV_train_mae],
    'R2 Train': [LR_train_r2, R_train_r2, L_train_r2, LCV_train_r2],
    'MSE Test': [LR_test_mse, R_test_mse, L_test_mse, LCV_test_mse],
    'RMSE Test': [LR_test_mse**0.5, R_test_mse**0.5, L_test_mse**0.5, LCV_test_mse**0.5],
    'MAE Test': [LR_test_mae, R_test_mae, L_test_mae, LCV_test_mae],
    'R2 Test': [LR_test_r2, R_test_r2, L_test_r2, LCV_test_r2]
})
```

[48] error_metrics

	Model	MSE Train	RMSE Train	MAE Train	R2 Train	MSE Test	RMSE Test	MAE Test	R2 Test
0	Linear Regression	4.932421e+09	70231.193978	53355.493020	0.839981	4.922898e+09	70163.369234	53429.970332	0.840265
1	Ridge Regression	4.932422e+09	70231.204591	53355.815339	0.839981	4.922903e+09	70163.405053	53430.224281	0.840265
2	Lasso Regression	4.932437e+09	70231.313105	53356.254495	0.839980	4.922950e+09	70163.739929	53430.532409	0.840263
3	LassoCV Regression	4.932437e+09	70231.313105	53356.254495	0.839980	4.922950e+09	70163.739929	53430.532409	0.840263

Next steps: Generate code with error_metrics | View recommended plots | New interactive sheet

[ ] Start coding or generate with AI.

Fig. 3: Linear Regression Model Error Metrics

3. Regression

3.1 Linear regression

We aim to use linear regression to model the relationship between the average resale price of an HDB flat and several predictor variables: the purchase date (X1), floor area in square meters (X2), lease commencement date (X3), location or town (X4…X30), flat type (X31…X38), flat model (X39…X61), and story range (x62…X78). The resale price was selected as our dependent variable with X1 to Xn being our independent variables. Many of our variables such as Purchase date, floor area, story range are either continuous in nature or can be encoded in a way that they maintain an ordinal relationship. Since linear regression is effective for predicting prices for variables with continuous relationships it remains an effective method to predict the resale price.

$$Y = -60452540.4 + 25831.72x1 + 1516.46x2 + 4171.61x3 + 4219.08x4 + 74867.86x5 + 58396.81x6 + 174425.11x7 - 4589.08x8 + \cdots + 112049.46xn$$

4. Excel-Based Decision Tool

4.1 UserForm

Our Userform can predict the prices based on the coefficients and the inputs provided. We require users to enter the most accurate data they have for the form to predict the price of the HDB flat. We have utilized combo boxes for users to choose the month in which they intend to buy/sell, lease commencement date of the HDB, the location users want to buy/sell their HDB, flat type of their HDB, flat model that they intend to buy/sell and the floor range of the HDB. We have also provided two textboxes for the users to key in the purchase/selling year and to enter the floor squared area. There are two action buttons, namely the calculate button which calculates the predicted price with +- Root Mean Squared Error (RMSE) deviations and displays it and a close application button which will exit the application. The reason as to why we used RMSE for our deviation is because RMSE is sensitive to large errors because it squares the differences between predicted and actual values before averaging. In predicting HDB resale prices, large prediction errors are more concerning. RMSE emphasizes these large errors and helps focus on reducing them. Hence, we used RMSE for our deviation.

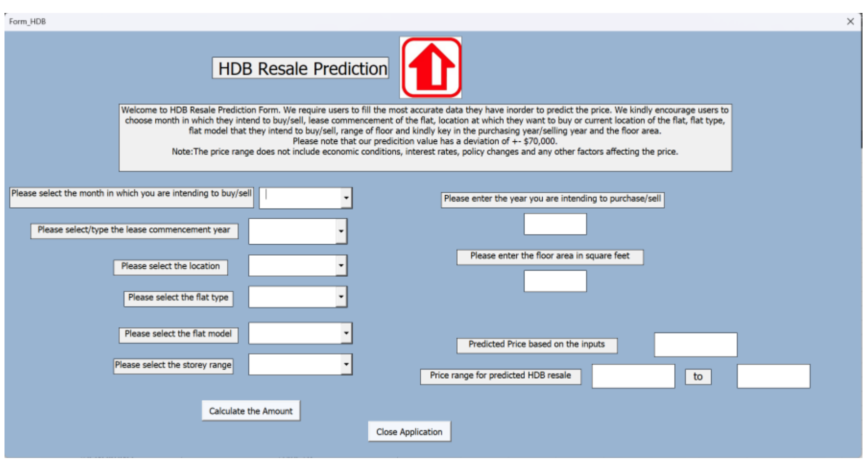


Fig. 4: HDB Resale Prediction UserForm

4.2 VBA

After the users have entered the data and clicked 'Calculate the Amount' our VBA will start calculating the predicted price of the HDB resale. We have implemented several functions and variables to validate the inputs, ensuring that they are either accepted as valid or flagged as invalid. Firstly, we have defined variables that will store the inputs from the combo box and the text box.

```
Dim selectedMonth As String
Dim selectedLeaseCommencementDate As Integer
Dim selectedLocation As String
Dim selectedFlatType As String
Dim selectedFlatModel As String
Dim selectedStoreyRange As String
Dim selectedYear As Integer
Dim selectedFloorArea As Double
Dim Final price As Double
```

Fig. 5: VBA Code Defining Variables

Once these values are initialized, we then define sub functions for each input to calculate their respective coefficients and return the value.

4.2.1 Year Value

In Fig. 6, we defined CheckYearValue function which returns a float (Double) value. We set the yearValue to the selectedYear which was what the user input and check if it is a number and whether it is greater than 0. If the conditions are satisfied, the year value will be calculated by using the coefficient for the year from our Linear regression multiplying it with the yearValue and return the float value.

```
Function CheckYearValue() As Double
    Dim yearValue As Variant
    Dim yearCoeff As Double

    yearValue = selectedYear

    If IsNumeric(yearValue) And yearValue > 0 Then
        yearCoeff = yearValue * 25831.72
        CheckYearValue = yearCoeff
    End If
End Function
```

Fig. 6: VBA CheckYearValue() Function

Similarly, we have defined multiple sub functions which process the data entered and return a value that is used to calculate the final amount of the HDB.

4.2.2 Calculate_Click()

Calculate_Click() is called when the "Calculate Amount" button is clicked. It checks for any blank inputs and prompts the user to fill them in if necessary. Furthermore, we have defined IsValidYear() function that determines if the input year they intend to buy/sell is before the lease commencement date or if the year they intend to buy/sell is more than the 99 years lease from the lease commencement date. We do not want the users to be able to buy/sell a HDB before it is even built and since HDB collects back the HDB flat after the 99 years lease, we also do not want the users to be able to buy/sell the HDB unit after its 99 years lease. Hence, we defined this function to prevent this from occurring.

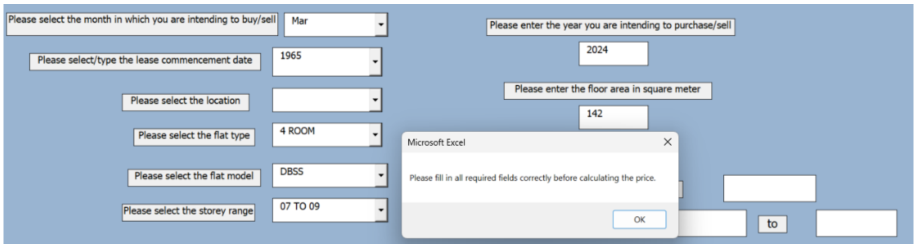


Fig. 7: UserForm Error 1

4.2.3 CalculateFinalPrice()

CalculateFinalPrice() is a function that calculates the final predicted price buy adding all the values returned from individual sub functions like CheckFloorValue(), CheckFlatModelValue() etc. After calculating the value it will display the value in the textboxes provided.

```
Function CalculateFinalPrice(flatmodelCoeff As Double, _
                             flatCoeff As Double, _
                             locationCoeff As Double, _
                             monthCoeff As Double, _
                             storeyCoeff As Double, _
                             floorCoeff As Double, _
                             intercept As Double, _
                             yearCoeff As Double, _
                             lcdCoeff As Double) As Double

    CalculateFinalPrice = flatmodelCoeff + flatCoeff + locationCoeff + monthCoeff + storeyCoeff + floorCoeff + intercept + yearCoeff + lcdCoeff
End Function
```

Fig. 8: VBA CalculateFinalPrice() Function

We call the CalculateFinalPrice() inside Calculate_Click() after checking if all the conditions for the inputs are satisfied. It then takes in all the arguments, calculates the price and returns the value. Since we are counting for +-70,000 deviation based on our RMSE, we defined two more textboxes each one displaying a lower and higher range of the predicted price based on our deviation. The function also checks if final price – 70,000 is 0 or smaller, if it is, it prints the first final price as the lower range.

```
Private Sub Calculate_Click()

    If Not IsValidInput() Then
        MsgBox "Please fill in all required fields correctly before calculating the price."
        Exit Sub
    End If

    If Not IsValidYear() Then
        MsgBox "Entered year must be greaten than lease commencement year."
        Exit Sub
    End If

    Dim flatmodelCoeff As Double
    Dim flatCoeff As Double
    Dim locationCoeff As Double
    Dim monthCoeff As Double
    Dim storeyCoeff As Double
    Dim floorCoeff As Double
    Dim intercept As Double: intercept = -60452540.44
    Dim finalPrice As Double
    Dim yearCoeff As Double
    Dim lcdCoeff As Double


    flatmodelCoeff = CheckFlatModelValue()
    flatCoeff = CheckFlatValue()
    locationCoeff = CheckLocationValue()
    monthCoeff = CheckMonthValue()
    storeyCoeff = CheckStoreyValue()
    floorCoeff = CheckFloorValue()
    lcdCoeff = CheckLeaseCommencementDateValue()
    yearCoeff = CheckYearValue()

    finalPrice = Abs(CalculateFinalPrice(flatmodelCoeff, flatCoeff, locationCoeff, monthCoeff, storeyCoeff, floorCoeff, intercept, yearCoeff, lcdCoeff))
    Display_Price.value = Format(finalPrice, "$0.00")
    PositiveDev.value = Format(finalPrice + 70000, "$0.00")
    If (finalPrice - 70000) > 0 Then
        MinusSeventy.value = Format(finalPrice - 70000, "$0.00")
    ElseIf (finalPrice - 70000) < 0 Then
        MinusSeventy.value = Format(finalPrice, "$0.00")
    End If

End Sub
```

Fig. 9 VBA Calculate_Click() Sub

5. Evaluation

5.1 Quantitative evaluation

Mean squared error (MSE) measures the average squared difference between the actual and predicted resale prices and root mean squared error (RMSE) is just the square root of the value to bring it back to the same units as the data. Using MSE allows larger errors to contribute disproportionately to the final value due to squaring which essentially magnifies larger errors and minimizes smaller errors making it a good representation of the model's accuracy (DataTechNotes, 2019).

MSE=$\frac{1}{n}\sum_{i=1}^{n}(Y_a - Y_p)^2$ where Y_a represents actual values and Y_p represents predicted values

RMSE=$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_a - Y_p)^2}$

Mean absolute error measures the average absolute difference between actual and predicted values. A lower MAE suggests that the model's predictions are, on average, closer to the actual values. Unlike MSE MAE takes the absolute value thus it is not as sensitive to outliers which are exceptionally big. This makes it suitable for datasets where you want a metric that reflects overall predictive performance without being skewed by a few extreme predictions. Thus, the MAE directly shows on average what is the absolute price difference between the predicted and actual value giving a clear representation of the reliability of the model (DataTechNotes, 2019).

MAE=$\frac{1}{n}\sum_{i=1}^{n}|Y_a - Y_p|$

The R^2 coefficient is a statistical measure that indicates how well the independent variables in a regression model explain the variability of the dependent variable. The R^2 range goes from 0 to 1 where a R^2 value of 1 means every change in the independent variable 100 percent predicts the change in resale price while a R^2 value of 0 hints there is absolutely no correlation between the resale price and the independent variables (DataTechNotes, 2019).

$R^2=1-\frac{\sum_{i=1}^{n}(Y_a-Y_p)^2}{\sum_{i=1}^{n}(Y_a-Y_{avg})^2}$ where Y_{avg} represents the average of the actual resale price $Y_{avg}=\frac{1}{n}\sum_{i=1}^{n}Y_a$

However, it is important to note that high R^2 value may not necessarily translate to more accurate results due to occurrence of overfitting. Overfitting occurs when a model learns not just the underlying patterns in the training data but also the random variations in data that do not represent any trend. Overfitting typically occurs when there are too many predictor variables including irrelevant ones leading to the model mapping resale price to these unnecessary independent variables. This limits the model's accuracy when it comes to new data as the model has tailored itself too closely to the training data rather than learning the generalizable patterns.

5.2 Qualitative evaluation

Incorporating qualitative methods into the analysis of HDB resale prices offers valuable insights that extend beyond what quantitative data can capture, especially considering the complexities of real-world markets. Qualitative data delves into non-numeric aspects, capturing motivations, preferences, and market sentiments that can be challenging to quantify but are crucial for a comprehensive and more accurate understanding.

One effective approach is to conduct in-depth interviews with recent and potential buyers and sellers to gain firsthand insights into their decision-making processes. Understanding buyers/sellers motivations and price sensitivity through direct engagement can expose underlying themes that are less easily captured in numerical data, such as lifestyle preferences, long-term investment perspectives, or location-based sentiments.

Additionally, consulting with experienced real estate agents could further enrich our understanding. Real estate agents, being seasoned experts can offer insights into market dynamics, such as how various government policies or economic changes have historically influenced resale prices and consumer behaviour. With their insights, we can better identify recurring patterns that are often overlooked in purely quantitative approaches, allowing us to account for complexities and subtle factors impacting the HDB resale market. Thus, these qualitative methods enhance our understanding of the housing market ultimately providing a more balanced and precise view of the factors driving HDB resale prices.

6. Assumptions and Limitations

6.1 Assumptions

- Price recommendations are only based on past data from 2017 onwards due to lack of information
- For 'central area' we grouped places such as Queen Street, Jalan Kukoh, Rowell, Waterloo, Newmarket under it
- Each independent variable influencing the resale prices of HDB flats are highly unrelated and are not multi-collinear
- The resale price of one HDB flat is independent of another HDB flat, and there is no spillover impact from nearby property sale.
- Changes in each independent variable affect the resale prices of HDB flats consistently and proportionally

6.2 Limitations

- External factors, including economic conditions, interest rates and policy changes are excluded as they are difficult to quantify and include in the training of Linear Regression model. However, these factors play a crucial role in influencing HDB resale prices. The exclusion results in a less accurate model
- Dataset used only contained price recommendations from 2017 onwards, hence resulting in a margin of error derived from the lack of information
- There is an oversimplification of the housing market as HDB resale prices are often influenced by a myriad of factors that are difficult to quantify. Adopting a linear regression model, with limited independent factors affecting the resale price reduces the accuracy of its subsequent predictions
- Final model is entirely trained based on past data, and will not be able to adapt to major economic events in the future, and account for the large deviations

7. Conclusion

In conclusion utilizing linear regression to predict resale prices provides an insightful and efficient approach to the relationship between the various factors and the resale prices. Linear regression's straightforward implementation and interpretability makes it a good choice for analysis and insight into this matter. However as linear regression assumes a linear relationship among variables and does not handle interactions or non-linear dynamics well, this makes it hard to implement in the Singapore housing market. Due to the complex nature of Singapore's housing market characterized by government interventions, volatile economy, people's different mindsets, these emotional and psychological aspects make it hard to predict accurately just with the use of a model. Thus, to strengthen our model we have to account for nonlinear aspects perhaps through the use of decision trees in addition our key variables scope can be expanded to include variables like distance form key institutions like schools, hospitals etc.

All in all, linear regression is an effective method, as our data set grows gradually to improve its accuracy and by adopting methods to account for nonlinear effects our model can be enhanced greatly.

8. Bibliography

Housing and Development Board. (2021). Resale flat prices based on registration date from Jan-2017 onwards (2024) [Dataset]. data.gov.sg. Retrieved from https://data.gov.sg/datasets/d_8b84c4ee58e3cfc0ece0d773c8ca6abc/view

DataTechNotes. (2019, February 14). *Regression Model Accuracy (MAE, MSE, RMSE, R-squared) Check in R*. Retrieved from DataTechNotes: https://www.datatechnotes.com/2019/02/regression-model-accuracy-mae-mse-rmse.html#google_vignette

(Python Code) Google Colab:

https://colab.research.google.com/drive/1fyOEC5Dd2WdV4WyZouDfr9zaGoxL5jbC?usp=sharing

9. Appendix

9.1 VBA

Fig. 10 shows the initialization of the variables and once the user inputs a data, the data will be stored according to their variables as shown in Fig. 11. These variables will then be retrieved by the functions to calculate the coefficients.

```vba
Private Sub Month_Change()
    If Month.Text = "" Then
        selectedMonth = ""
    Else
        selectedMonth = CStr(Month.Text)
    End If
End Sub
Private Sub LeaseCommencement_Change()
    If LeaseCommencement.Text = "" Then
        selectedLeaseCommencementDate = 0
    Else
        selectedLeaseCommencementDate = CInt(LeaseCommencement.Text)
    End If
End Sub
Private Sub Location_Change()
    If Location.Text = "" Then
        selectedLocation = ""
    Else
        selectedLocation = CStr(Location.Text)
    End If
End Sub
```

Fig. 10: VBA Code 1

```vba
Dim selectedMonth As String
Dim selectedLeaseCommencementDate As Integer
Dim selectedLocation As String
Dim selectedFlatType As String
Dim selectedFlatModel As String
Dim selectedStoreyRange As String
Dim selectedYear As Integer
Dim selectedFloorArea As Double
Dim Final_price As Double
```

Fig. 11: VBA Code Defining Variables 2

9.1.1 Flat Type Value

For the binary data, we have also defined the sub functions like CheckFlatValue() which returns a float(Double). Here we use Select Case to check if the chosen flat type matches with the data we have.

```vba
Function CheckFlatValue() As Double
    Dim flatValue As Variant
    Dim flatCoeff As Double
    Dim TWO_ROOM As Double: TWO_ROOM = -31499.41
    Dim THREE_ROOM As Double: THREE_ROOM = -4092.7
    Dim FOUR_ROOM As Double: FOUR_ROOM = 7881.97
    Dim FIVE_ROOM As Double: FIVE_ROOM = 18171.38
    Dim EXECUTIVE As Double: EXECUTIVE = 10929.31
    Dim ONE_ROOM As Double: ONE_ROOM = -65723.34
    Dim MULTI_GENERATION As Double: MULTI_GENERATION = 64332.79

    flatValue = selectedFlatType

        Select Case flatValue
            Case "1 ROOM"
                flatCoeff = ONE_ROOM
            Case "2 ROOM"
                flatCoeff = TWO_ROOM
            Case "3 ROOM"
                flatCoeff = THREE_ROOM
            Case "4 ROOM"
                flatCoeff = FOUR_ROOM
            Case "5 ROOM"
                flatCoeff = FIVE_ROOM
            Case "EXECUTIVE"
                flatCoeff = EXECUTIVE
            Case "MULTI-GENERATION"
                flatCoeff = MULTI_GENERATION
            Case Else
                flatCoeff = 0
        End Select
    CheckFlatValue = flatCoeff
End Function
```

Fig. 12: VBA CheckFlatValue() Function

Here, the conditions must be satisfied before returning the coefficients which we have defined at the start of the function. These coefficients are obtained from the linear regression. The same logic has been implemented for other functions to return the coefficient values.

9.1.2 IsValidInput()

IsValidInput() is a function we defined which returns a Boolean value. This function helps to solve the issue of incorrect input and empty inputs.

```
Private Function IsValidInput() As Boolean
    IsValidInput = True

    If selectedMonth = "" Then IsValidInput = False
    If selectedLeaseCommencementDate = 0 Then IsValidInput = False
    If selectedLocation = "" Then IsValidInput = False
    If selectedFlatType = "" Then IsValidInput = False
    If selectedFlatModel = "" Then IsValidInput = False
    If selectedStoreyRange = "" Then IsValidInput = False
    If selectedYear <= 0 Then IsValidInput = False
    If selectedFloorArea <= 0 Then IsValidInput = False
End Function
```

Fig. 13: VBA IsValidInput() Function

IsValidYear()

We have also accounted for entering year smaller than lease commencement date. For example, we do not want to buy or sell a HDB before it is even built. So, we defined a function called IsValidYear() that returns a Boolean value.

IsValidYear() is a function that determines if the input year they intend to buy/sell is smaller than the lease commencement date. We do not want the users to buy or sell a HDB before it is even built hence, we defined this function in preventing this from occurring.

```
Private Function IsValidYear() As Boolean
    IsValidYear = True
    If selectedYear < selectedLeaseCommencementDate Or selectedYear > (selectedLeaseCommencementDate + 100) Then IsValidYear = False
End Function
```

Fig. 14: VBA IsValidYear() Function

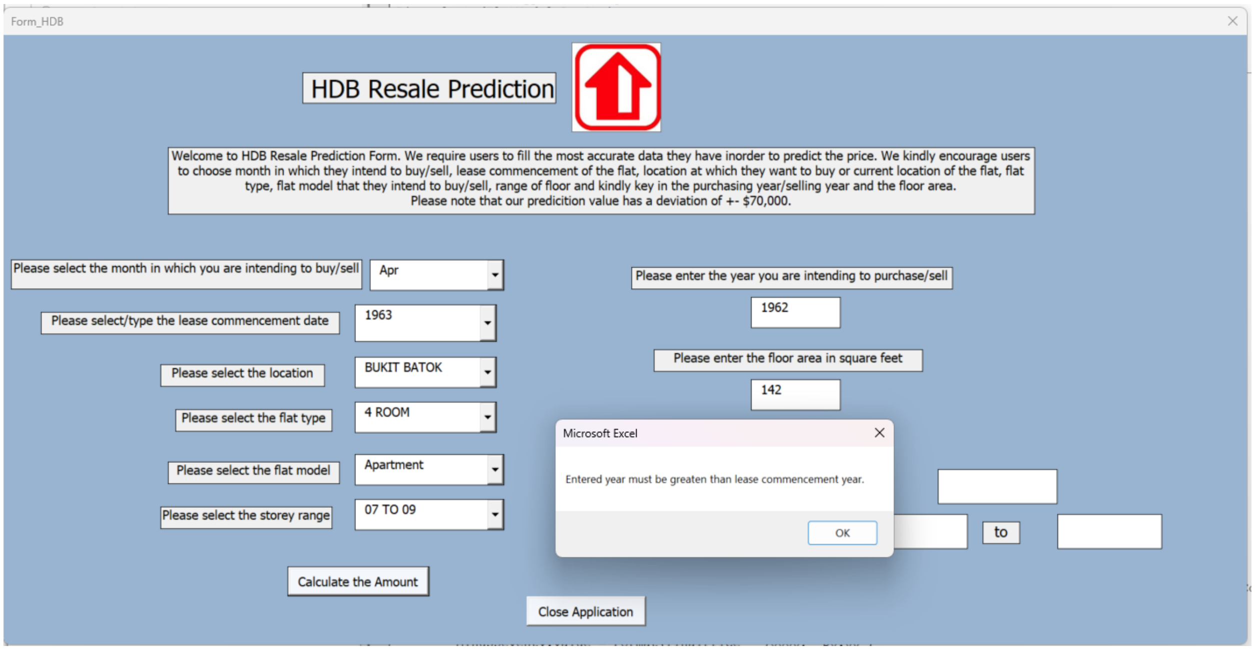


Fig. 15: UserForm Error 2

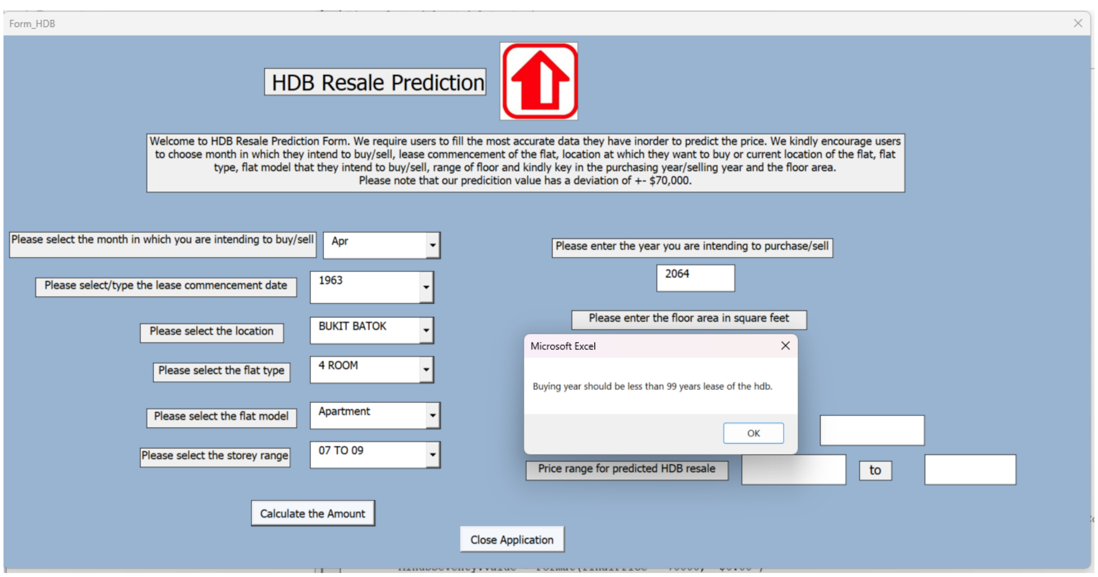


Fig. 16: UserForm Error 3