**ITD252 Assignment**

**Name: Timothy Taw**

**Admin No: 21a192w**

# Introduction

Location location location! Everyone believes location is the main factor that affects property prices. Or is it? In this assignment I will be investigating if there are other factors affecting resale HDB prices. The final machine learning model can also be used it to predict future prices and hopefully identify which bargain deals.

# Brief Introduction of Data Set

The collated data set contains HDB resale prices from 2011 till 2021 with 17 explanatory variables as shown below. Resale price is the target variable which we would like to predict.



*Figure 1*

On first inspection, some of the data needs to be transformed and cleaned before it can be further explored.

1. Storey Range was split into the lower and upper story by the delimiter “TO” and the average of them was found. The average was input in a new column called Estimated Storey.

Text

Description automatically generated



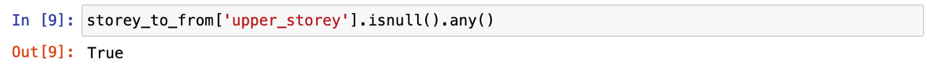
Graphical user interface

Description automatically generated with medium confidence

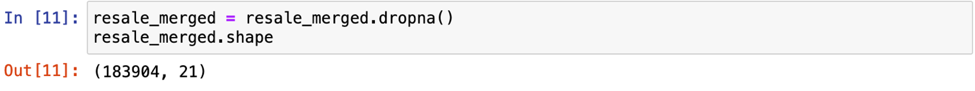
1. In addition, under the “Storey Range” column, one of its values 103.797. This is likely to be an error as the tallest HDB in Singapore is only about 50+ floors. Hence the data row was removed.

Table

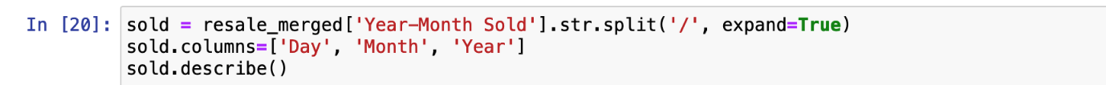
Description automatically generated=

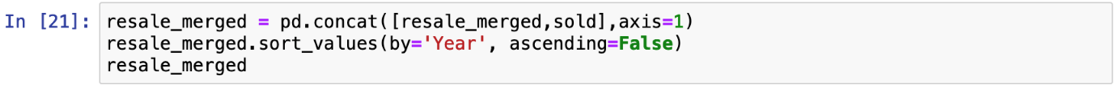


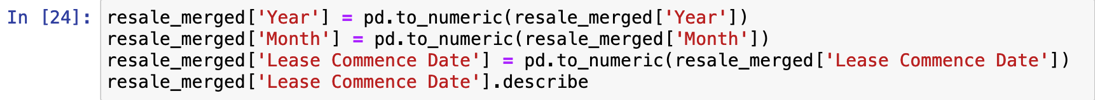
After splitting Storey Range by delimiter “TO”, the upper storey of the column with value 103.795 was empty. Hence it can be removed by dropping the null value.



1. Year-Month Sold column was also split into the year and month columns by splitting the cell with delimiter “/”. With the Year Sold and Lease Commence date, I could calculate the Lease Remaining when the HDB was sold by using the formula 99-(Year Sold-Lease-Commence Date+1).







Graphical user interface

Description automatically generated with medium confidence

1. I then checked if there were any null or zero elements in all the columns:Table

   Description automatically generated with medium confidence



I did the above for every column header, and there were no zero elements.

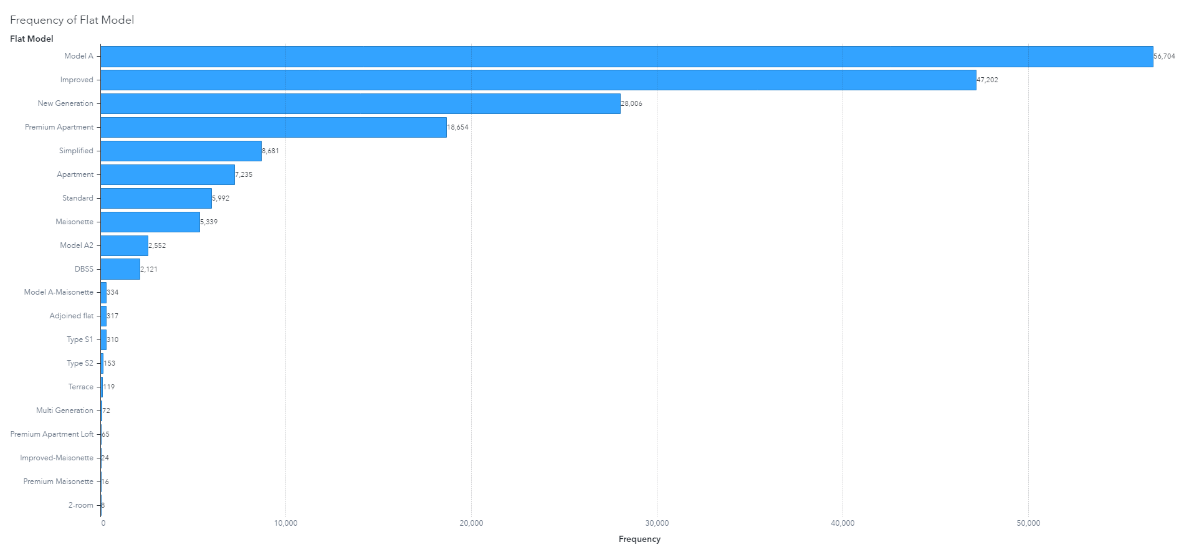
1. After all the above were done, the data frame from from jupyter notebook was exported and saved into a excel file for processing in SAS.



# Exploratory Data Analysis

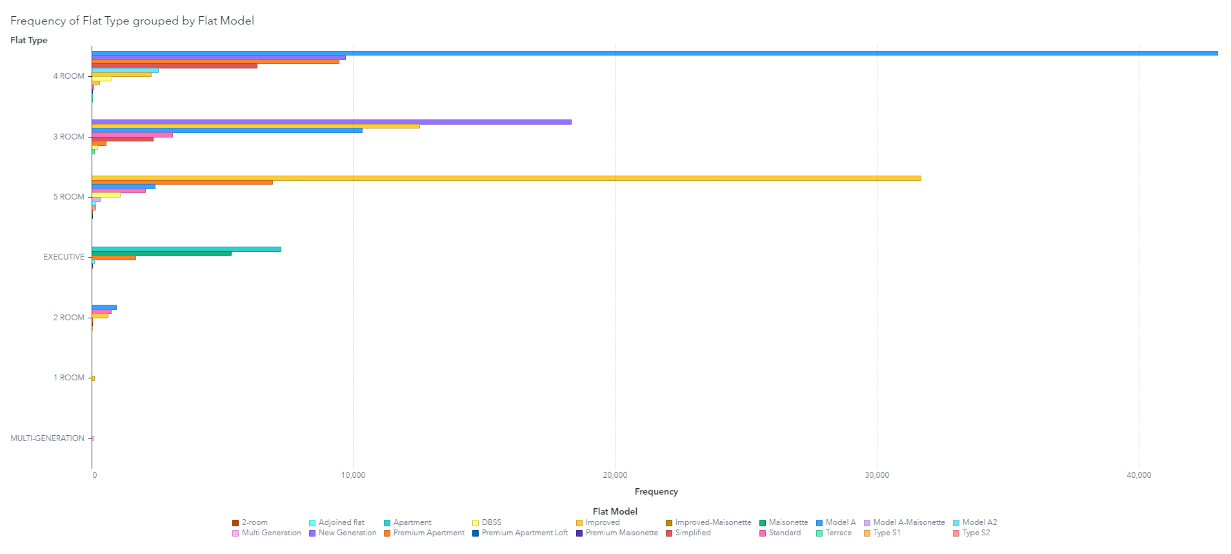
To better understand the data, I analysed the variables and plotted some simple graphs. The below are my findings:

1. Flat Model is a categorical data type with 20 different types. However several the categories are infrequently sold as shown below. Some like the DBSS and Terrace type have been phrased out too. I will be further exploring if removing these less commonly sold HDB types would make my model more robust.

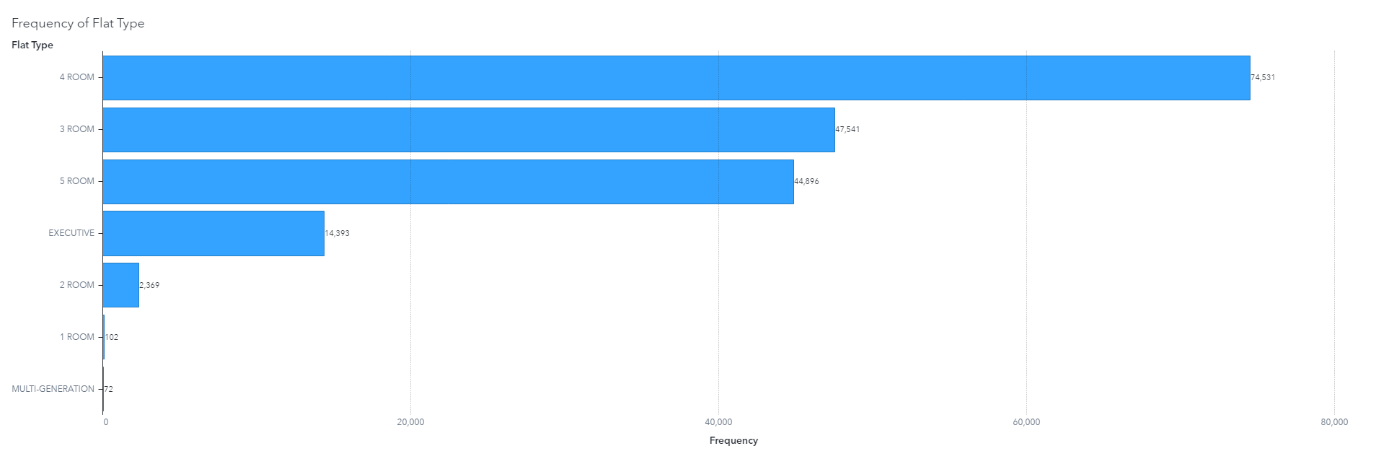


*Figure 2*

1. Flat Type is a categorical data type with 7 unique types. It is also closely related with Flat Model (Figure 3), ie most of the 5 room flats are Improved flat types and most of the 4 room flats are Model A. 1 and 2 room flats are also not in demand mostly due to their small size. Multi Generation flats are also not widely sold as they must be sold to another multi-generation family (Figure 4). Hence I would be disregarding this variable and instead use Flat Model as it is more specific and would avoid any multicollinearity problems.

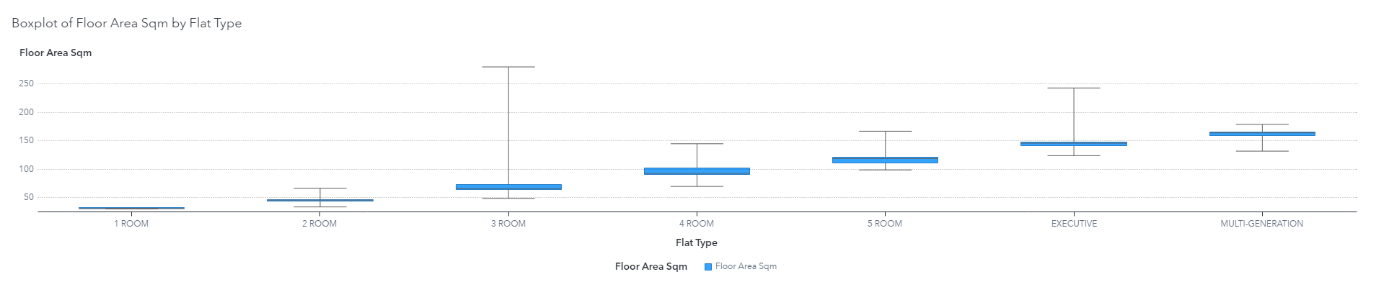


*Figure 3*

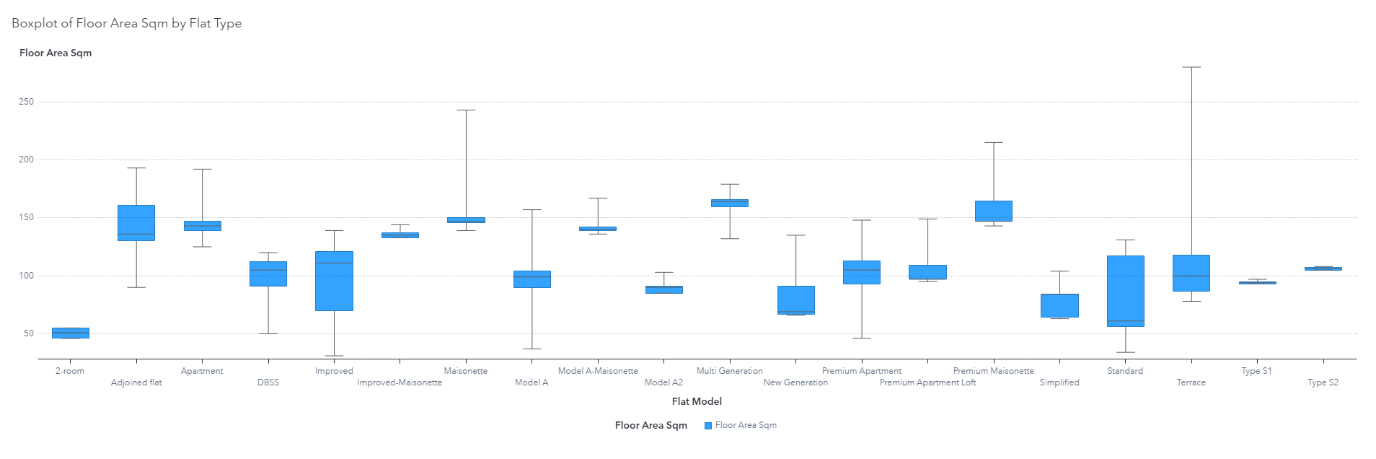


*Figure 4*

As shown in figure 5, Flat Types are closely related to Floor Area Square Meter (Sqm). However since Flat Model is not as closely related to Floor Area Sqm as Flat Types (figure 6), I will still be including Flat Model as a feature in the ML model.

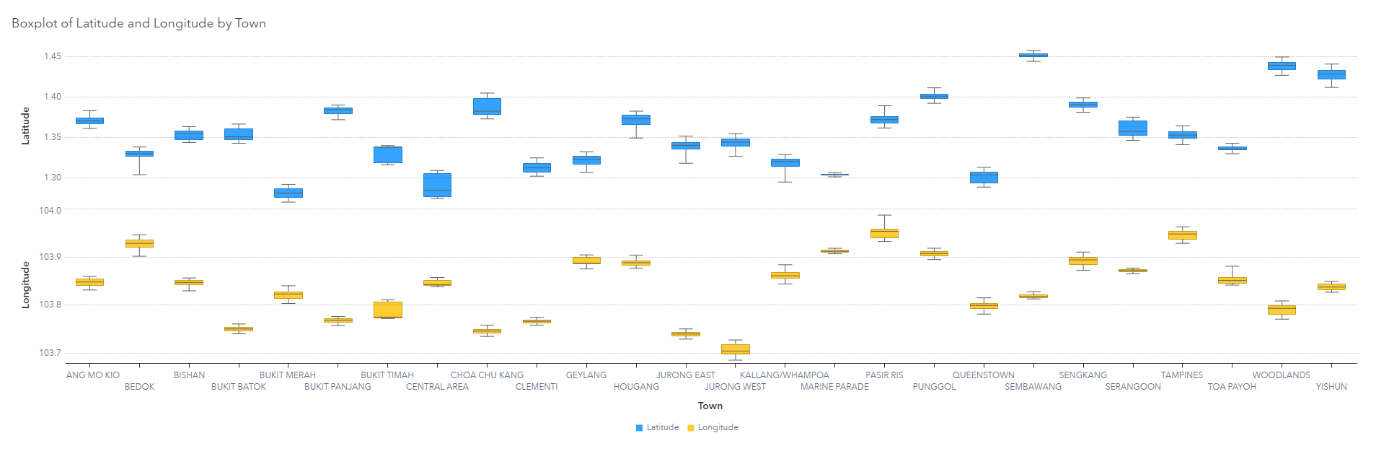


*Figure 5*

*Figure 6*

1. Town has 26 unique categories. It is also closely correlated with latitude and longitude based on knowledge and the below boxplot. One advantage of keeping the variable Town in a regression model would make its interpretation easier, ie a HDB in Town A is more expensive than a similar HDB in Town B. However I would prefer to use longitude and latitude over Town as it will drastically reduce the number of features and hence increase the robustness of the model.

Town Group is also correlated to Town and hence would be dropped as a feature.



*Figure 7*

1. Nearest Hawker and Nearest MRT have 56 and 128 unique values respectively. If I were to use them, it would introduce too many variables in the model. As per the curse of dimensionality, I would prefer to reduce the number of features as much as possible to get a more robust model. Nearest Hawker and Nearest MRT are also dependent on which Town the resale HDB is in, hence they will be dropped.
2. Nearest Primary School has 165 unique values. Like the above, I have decided not to use them as it will include in too many variables. Instead, I have created a list of top 10 most popular primary schools in 2021 from the website

(https://www.salary.sg/2021/best-primary-schools-2021-by-popularity/).

This way I can analyse whether proximity to a popular school would cause an increase in HDB resale price.

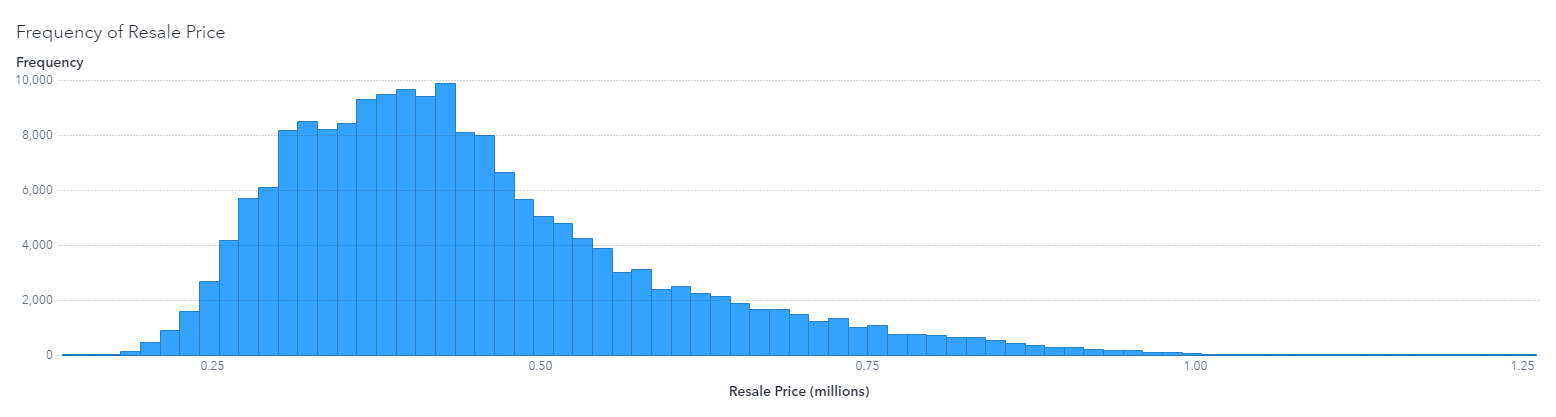
In excel, I created a list of unique Nearest Primary School values and used the countif function to check if its value is in my list of top 10 primary schools, if so then input 1 in a new column “Popular School”, otherwise input 0. I would then vlookup this table into the original data set.

Graphical user interface, application, table, Excel

Description automatically generatedTable

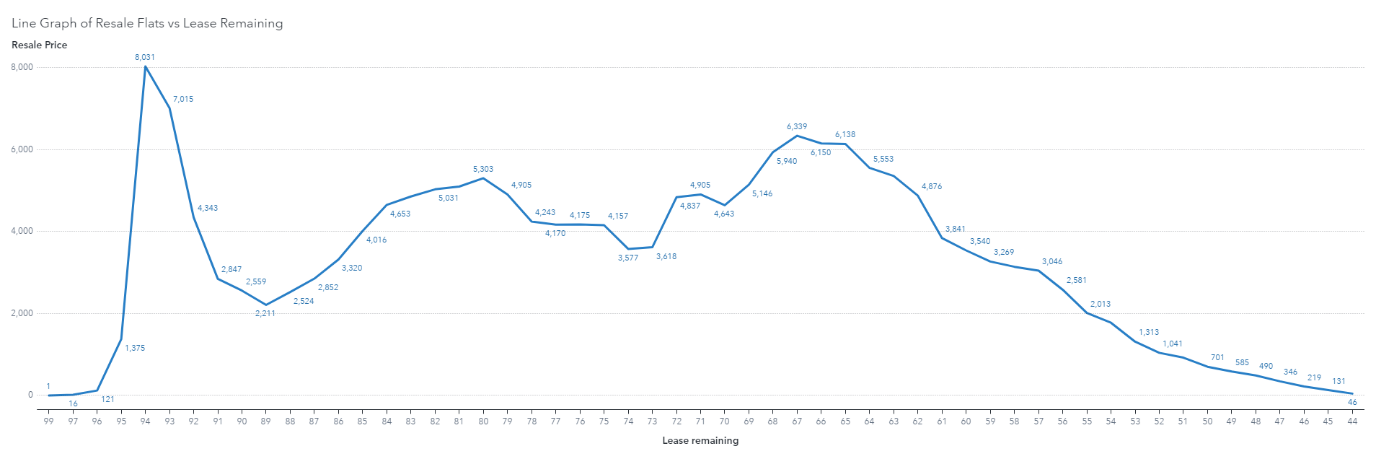
Description automatically generated

1. Resale price was slightly skewed to the left as shown below. I will explore whether it would upset the assumptions of a linear regression model and whether a log transformation would make the model more robust.



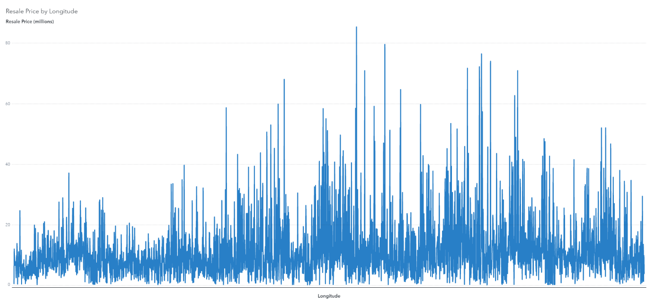
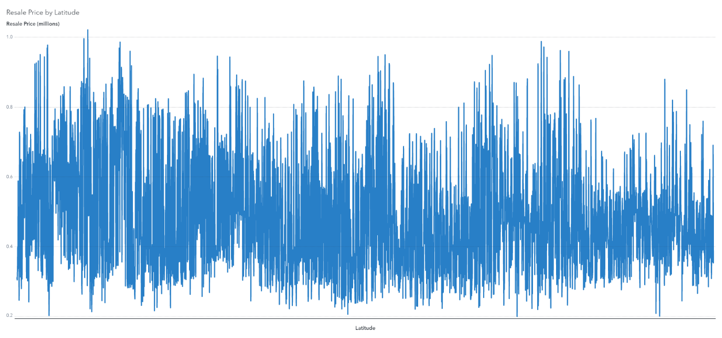
*Figure 8*

1. Based on the below figure, there were also some resale flats that were sold with more than 94 years of lease remaining. From HDB’s website (<https://www.hdb.gov.sg/residential/selling-a-flat/eligibility>), most flats require a Minimum Occupation Period (MOP) of 5 years. Flats sold before the MOP period could be due to various reasons such as bankruptcy or divorce and should be considered as a distress sale. This may impact the price of the unit sold. Another interesting point is that many home owners sell their flat immediately once their MOP is up, likely to cash in on a booming resale market.



*Figure 9*

1. I would imagine resale prices would be lower on the extreme ends of latitude and longitutde and higher in the centre, however based in the figures below, it shows quite an even distribution of housing prices at all latitudes; and high resale prices in the in central Singapore and also on the eastern side.



*Figure 10 Figure 11*

# Machine Model Building

I explore 3 machine learning methods that are used to predict the target variable, Resale HDB prices in Singapore, namely linear regression, decision tree and neural network.

The linear regression will be my baseline model as it clearly explains which variable affects Resale Prices and its magnitude. I will tune its features to find the model with the highest predictive strength, or adjusted R Square (R2), while not flouting the assumptions of the model. As the name suggests, it assumes linearity between the independent variables and dependent variable.

Decision Tree Regression model is included as it is commonly used and easily understandable. It breaks down a dataset into smaller and smaller subsets and in the process builds an inverted tree. The higher the variable in the tree, the more important it is.

Lastly, I have also included a single multi-layer perceptron neural network model as it is currently very popular. The input data goes through a layer of “filters” or hidden layers and the model tries to recognize underlying relationships in the data by minimizing the loss function in the “filters” matrices.

Decision Tree and Neural Network models also have an added advantage over linear regression as it can capture non-linear relationships in the variables.

# Regression Model Overview

After removing all the variables that are correlated to each other, I have the following independent variables left: Flat Model, Town, Floor Area Sqm, Latitude, Longitude, Hawker Dist, Mrt Dist, School Dist, Estimated Storey, Lease Remaining, Popular School.

I first compared whether I should include the Town categorical variable or Latitude and Longitude variables. I should not include both in as the variables are correlated with each other (as mentioned above).

In both models, all features are significant in explaining the dependant variable. The Longitude and Latitude regression model has a lower adjusted R2 of 0.810 and Average Square Error (ASE) of 3.949x10^9 compared to the Town regression adjusted R2 of 0.870 and ASE of 2.648x10^9. Despite the lower R2 and ASE, I would prefer the Longitude and Latitude regression model because is more precise in pin pointing the location of the resale HDB than just specifying the general area the HDB is in. The lower R2 and ASE is also likely due to the much lesser degrees of freedom the model has: 28 compared to 51. The Longitude and Latitude model also shows that resale HDBs near popular primary schools command a premium, which is reinforced by general knowledge. The Town model instead infers that resale HDBs near non popular primary schools command a premium. Hence in conclusion the Longitude and Latitude regression model will be used.

Graphical user interface, application

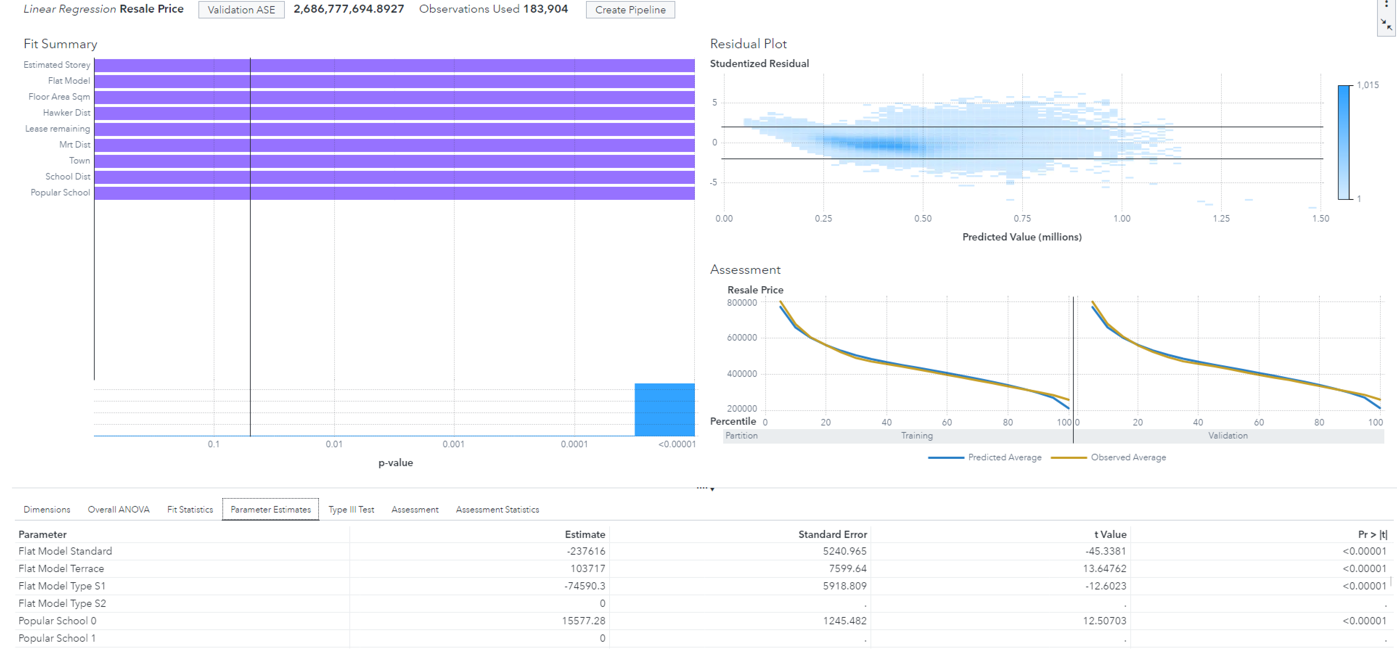
Description automatically generated

*Figure 12. Longitude and Latitude Regression Model*

Chart

Description automatically generated

*Figure 12.1. Longitude and Latitude Regression Model with histogram of residual plot*



*Figure 13. Town Regression Model*

From the above histogram of the residual plot (Figure 12.1), the residuals are normally distributed with only a slight skewed to the left, hence a log transformation of the resale prices is unnecessary. It will only complicate the model and make its interpretation harder.

On analysing the residual plot, most of the data points have equal variance and are not fanned out, indicating the residuals are independent and uncorrelated. Upon further investigation, most of the outliers are at the extreme ends of the resale prices. This is mostly likely because of the low demand of old and small HDBs and the million dollar HDBs in prime locations.

Filter 1:

To further improve the model, I will remove Model A Maisonette, Adjoined Flat, Type S1, Type S2, Terrance, Multi Generation, Premium Apartment Loft, Improved Maisonette, Premium Maisonette and 2 room as some of these Flat Models are not built anymore and have very low demand (Figure 2). Observations only decreased about 1500 out of about 184,000. On the other hand, I will not be removing the millon dollar HDBs from the model as they are an ever-increasing important part of the market. According to Straits Times, the number of millon dollar HDBs transacted tripled in 2021 (https://www.straitstimes.com/singapore/housing/record-261-million-dollar-hdb-flats-in-2021-resale-prices-rise-in-december-as-volume-dips).

Graphical user interface

Description automatically generated

*Figure 14. Regression model removing various Flat Models*

After removing the various Flat Models, the Adjusted R2 had a slight drop to 0.800 while the ASE slightly improved to 3.878x10^9. It did not make much of a difference in the final model. Hence I will not be filtering out the variables for my final linear regression model.

Filter 2:

After filtering out resale HDBs that are sold within their MOP of 5 years, there wasn’t a significant change in the model as shown below. Adjusted R2 and ASE decreased slight to 0.809 and 3.911x10^9 respectively. We can conclude that the flats sold within the MOP period were not sold at a considerable difference, likely due to strong demand of resale flats.

I will not include this filter in the final linear regression model.

Graphical user interface

Description automatically generated

*Figure 15. Regression model removing flats sold within MOP.*

# Decision Tree Model

Graphical user interface, application

Description automatically generated

*Figure 16*

The decision tree model was even less accurate compared to the linear regression model, with a much higher ASE of 5.636x10^9. It also predicted that popular schools and school distance were not significant, which goes against what we know.

# Neural Network Model

Chart

Description automatically generated with medium confidence

*Figure 17*

The Neural Network Model was even worse, with a much higher ASE of 217x10^9. Hence it would not be used.

# Model Comparison and Evaluation

Graphical user interface

Description automatically generated

*Figure 18*

The final chose model will be the linear regression with Longitude and Latitude with no filtering as there was no significant improvement after filtering. The linear regression model also had the lowest ASE among the models.

# Conclusion

Location is still very important in influencing the price of a resale HDB. It is also fueling the growth of million dollar resale flats. However, if one can’t get a centrally located HDB like the Duxton, there are many other good options like those in Queenstown, Bishan, Kallang/Whampoa (Appendix 3). Other factors like a higher storey, size of the HDB, and distance to a hawker centre or MRT also play a part. On the contrary, distance to a school doesn’t matter. It is more important to be closer to a top school. In addition, when considering a resale HDB with 65 years of lease left, one needs to be cautious as its resale price drops drastically. From my understanding, there will likely be a diminising market for such old properties. Most people would rather pay more for a newer property than get a cheaper and older property. Moreover the ability to use CPF to pay for these older properties is increasingly restricted.

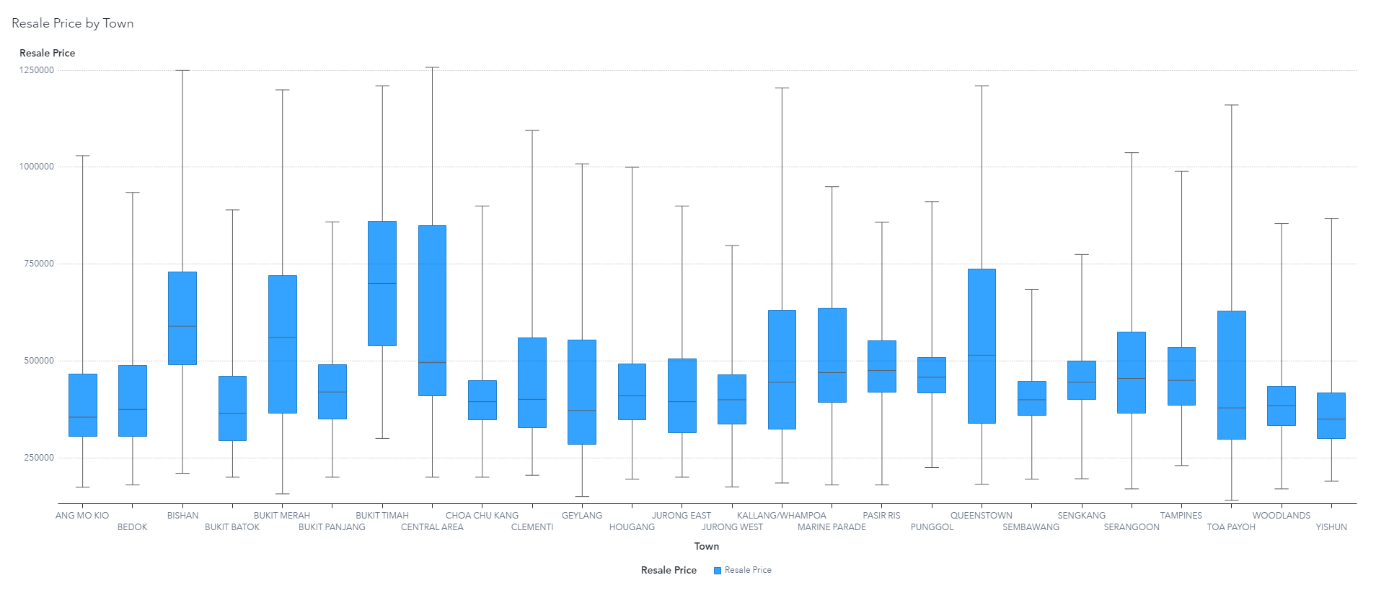
# Further Study and Recommendations

Million dollar resale flats have always been a hot topic for a couple of years that never seems to be losing any steam. More analysis can be done on this subset of flats to see what variables are driving their prices up and how they differ from the generic HDB flat.

# Appendix



*Appendix 1*



*Appendix 2*

Chart, box and whisker chart

Description automatically generated

*Appendix 3*