**Introduction:**

As an individual reaching the age where housing will be an upcoming concern of mine, I decided to look into the factors that influence the resale price of a flat and train different models to be able to predict the resale price of a flat based on the factors and historical prices of previous sales. Unlike items such as laptops and clothes, the housing prices are heavily influenced by the market’s demand and other factors which makes the price range of housing flats very huge. Therefore, by creating the model to estimate the price of the house based on the factors, it will help me and hopefully others in their future purchases of their homes.

**Dataset:**

The original dataset was retrieved from data.gov.sg and contains 11 columns and 147,534 entries, with transactions recorded from Jan 2017 to Feb 2023. The column names and their description can be seen below.

|  |  |  |
| --- | --- | --- |
| **Name** | **Type** | **Description** |
| Month | Datetime “YYYY-MM” | Date of Recorded Sale |
| Town | Text | Town Of Flat |
| Flat\_type | Text | Flat Type |
| Block | Text | Block Number |
| Street\_name | Text | Street Name |
| Storey\_range | Text | Level Range Of Flat |
| Floor\_area\_sqm | Numeric | Flat Size in Square Meters |
| Flat\_model | Text | Flat Model |
| Lease\_commense\_date | Datetime “YYYY” | Lease Commence Date |
| Remaining\_lease | Text | Remaining Lease |
| Resale\_price | Numeric | Resale Price |

***Data Dictionary***

**Data Preparation:**

Since the flats are only categorized by their towns, I have created another column called “region” for further analysis to see if certain regions in Singapore are more expensive than others. As such the region column has 5 unique values: “Central”, “North”, “South”, “West” and “North-West” with the central region having the most (9) regions under it.

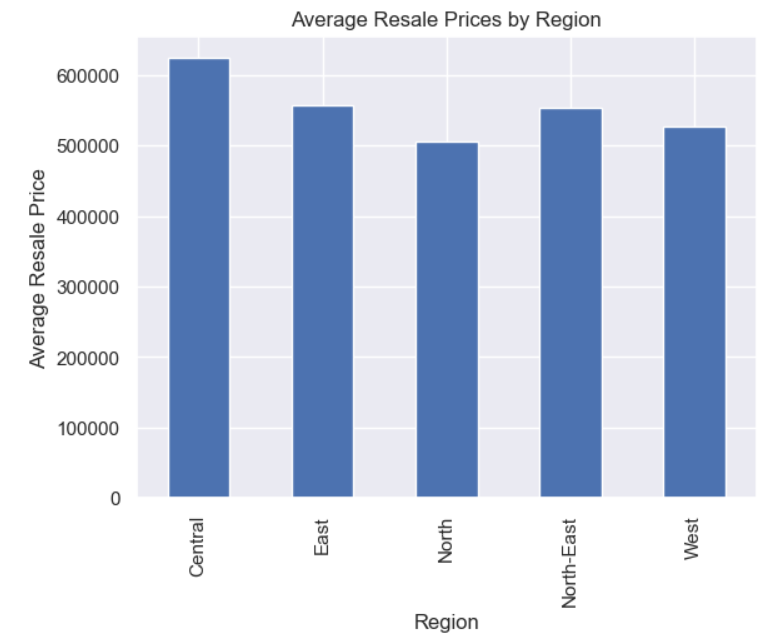
I have also separated the storey range column into two different columns: “Min Storey” and “Max Storey” and changed the type into numerical for ease of analysis.

In the original dataset, “remaining\_lease” column is in the format of “XX years XX months”. In order to analyse the data, I have converted the format of the column under a new column name, “remaining\_lease\_months”, to show the remaining lease in months only and changed the type to numerical.

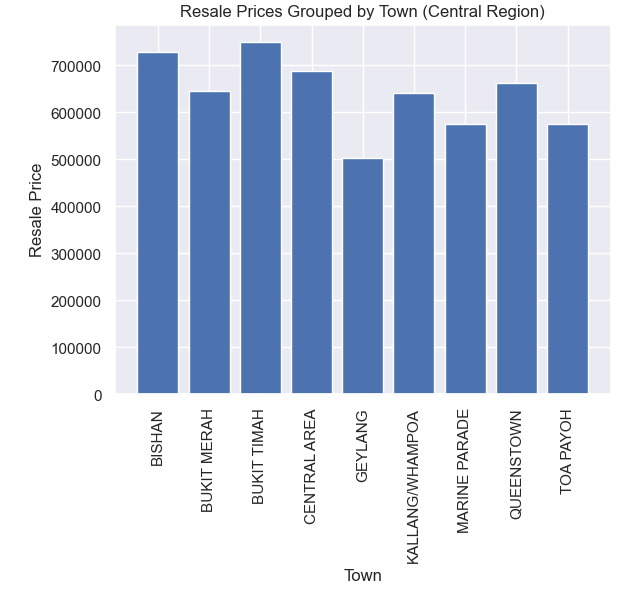
Lastly, since flat resale prices are highly dependent on recent historical data, only the past year of data has been selected for analysis. Therefore, I dropped all entries that were recorded before Feb 2022. Irrelevant columns such as “remaining\_lease” and “storey\_range” were dropped as well.

**Data Exploration:**

By comparing the average price of the flats in each region, we can determine if the region of the flats is a factor that affects the resale price of flats in Singapore. The graph below shows the average price of flats in each region.



As seen in the graph, flats in the Central region are on average more expensive than other regions while flats in the North region are generally cheaper. The towns of each region also affect the prices of the flats according to my analysis. In the five graphs below, the average resale price of a flat in each town has been compared for each of the five respective regions.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generatedChart, bar chart

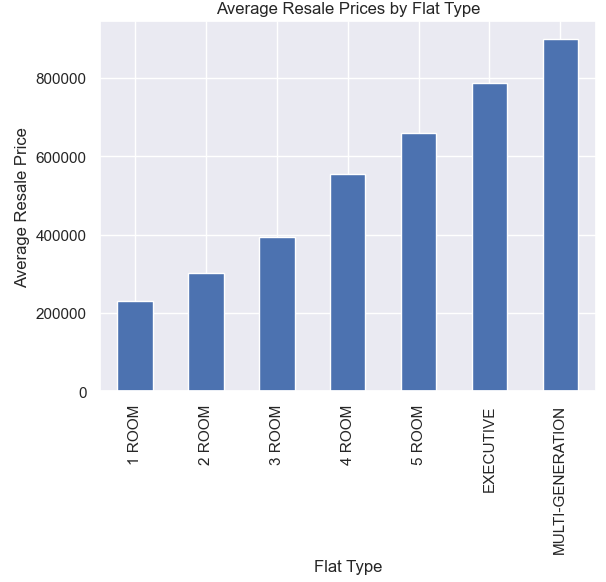
Description automatically generated

Chart, bar chart

Description automatically generated

As seen in all five graphs, there are differences in flat prices according to the towns that the flats are located in, therefore the towns of the flats are also a factor that should be taken into consideration.

Flat types are a category that describes the type of flat and more commonly the number of rooms the flat has. I have compared the average resale prices of all different types of flats in the graph below.



According to the graph, there is a clear distinction between the average flat prices for each flat type with one room flats costing the least and multigeneration flats costing the most. Therefore, flat types can be seen as a factor that affects the resale price.

In my exploratory analysis, it can also be determined that floor area and resale price have a strong positive correlation as shown in the graph below.

Chart, scatter chart

Description automatically generated

Resale prices and length of remaining lease has a weaker but still positive correlation as well as seen in the graph below.

Chart, scatter chart

Description automatically generated

It is also interesting to point out that the level that the flat is located at is also a determinant of the resale price. The average resale prices of flats located on each floor is shown below.

Chart, bar chart

Description automatically generated

It can be concluded that flats on higher floors are more likely to sell for a higher price as compared to lower-level flats. Lastly, flat models are also concluded to be determinants of resale price of a flat based on the graph below. Chart, bar chart

Description automatically generated

**Feature Selction:**

Before we train and test our models, we must choose our features to be used in our model. Based on the correlation matrix below, we can identify two pairs sof highly correlated columns. “mn\_storey” and “max\_storey”, and “lease\_commence\_date” and “remaining\_lease\_months”. Since both these pairs are perfectly correlated with a corerlation factor of 1, I have chose to drop “max\_storey” and “lease\_commence\_date” out as factors to prevent multicolinearity issues. Thus our final dataset that will be used to train and test the models will contain 7 independent varibales and 1 dependent variable.

**Training & Testing of Models:**

I have decided to go for a 80/20 split on the final dataset and have went to develop two models, a linear regression model and a decision tree model. The dataset was one-hot encoded to take into account the categorical variables in the final dataset and both models were trained and tested on the same test and train sets to compare the accuracy and results. The root mean square error and the R-Square score of both models are shown below.

|  |  |  |
| --- | --- | --- |
|  | **RMSE** | **R^2** |
| **Linear Regression** | 58370.40 | 0.88 |
| **Decision Tree** | 49995.44 | 0.92 |

The decision tree model slightly outperforms the linear regression model with a RMSE of $49995.44 and a R^2 Score of 0.92. Therefore, the decision tree model should be used for a more accurate estimation of resale flat prices.

**Final Conclusion:**

Despite the high accuracy on the model, there are a multitude of other factors that could affect a flat’s resale price that are not recorded in the original dataset. Certain ways that could be used to improve on the accuracy of the model would be to determine the distance from the flat to the nearest bus station using geolocation APIs as convenience of transport is a huge factor when determining resale prices of flats. Intangible factors that cannot be recorded down such as condition of the house are key to determining a resale price for a flat as well. Therefore, the estimation should be used in conjunction with a visit down to the house and not used solely on its own.