# CHAPTER THREE METHODOLOGY

The aim of this thesis is to improve the accuracy of price predictions in the real estate market by using machine-learning methods combined with subjective evaluations. The market is complex and susceptible to the influence of economic forces and socio-demographic shifts. Traditional methods often encounter difficulties in understanding this complexity. The approach combines advanced machine learning algorithms with subjective evaluations from domain experts and real-world observations to provide a complete foundation for accurate housing price estimates. Accurate predictions provide benefits to buyers, sellers, investors, lawmakers, and urban planners by enabling educated decision-making and addressing concerns of cost and accessibility. The collaborative approach exceeds traditional boundaries, offering inventive answers to challenges in the real estate sector.

# 3.1 Data Collection

The evaluation of Singapore's housing market relies on comprehensive statistics, including the Singapore housing dataset, which provides essential criteria such as property attributes, geographical details, and historical pricing information. Public infrastructure, including hospitals, schools, and transportation, was also collected to determine the influence of the local environment on housing prices. Singapore has challenges in obtaining housing due to limited land availability and high pricing rivalry, leading many people to opt for buying pre-owned Housing Development Board (HDB) units. Undertaking research on the pricing and determinants impacting HDB resale prices is essential for individuals and families seeking new homes, as well as for analysts and economists studying the stability and performance of this sector.   
  
The dataset consists of four CSV files with transaction data spanning from 1990 to 2020. The dataset comprises many parameters pertaining to the sale of apartments, including the year of sale, location, kind of apartment, street name, block number, area of the apartment, lease information, and selling price. The remaining lease refers to the precise period, expressed in years, months, and days, that is left before the property lease expires. This computation is carried out throughout the resale flat application procedure and is rounded up to the nearest month for the explicit purposes of using CPF funds and applying for a HDB loan.   
  
The resale prices should be regarded as only indicative, since they are dependent on several variables. The data was acquired from Data.gov.sg and comprises more than 800,000 sales transactions during the period from 1990 to 2020. Supplemental information may be obtained to augment this dataset, and examining the data can improve understanding of Singapore's real estate industry. A set of preprocessing methods were done to guarantee the integrity and relevancy of the data. The process included data cleansing, standardization, and feature manipulation to address any discrepancies and ready the datasets for further analysis.

# 3.2 Preprocessing

Data preprocessing is an essential stage in any data analysis or modelling effort, since it establishes the basis for precise and dependable outcomes. This section will delineate the precise preprocessing measures that will be used on the gathered data to guarantee its quality and appropriateness for modelling.

## 3.2.1. Handling Missing Values

One of the first responsibilities in data preparation is to manage missing values, which may significantly influence the performance of machine learning models if left untreated. This research will use suitable imputation strategies to handle missing values in the dataset, such as substituting missing values with the median or mean of the corresponding feature. Special emphasis will be given to subjective variables, such as distance to amenities, where missing numbers may imply a lack of closeness to the associated facilities.

## 3.2.2 Outlier Detection

Outlier detection refers to the identification of data points that differ considerably from the majority of the sample. These outliers have the potential to distort statistical analyses and predictions made by models. Hence, outlier detection approaches will be used to accurately identify and manage outliers. One way to identify and eliminate outliers from the dataset is by visually examining boxplots or using statistical techniques like the Z-score or interquartile range (IQR).

## 3.2.3 Feature scaling

Feature scaling is a necessary step to ensure that all characteristics have equal importance in the modelling process, particularly when utilizing algorithms that rely on distances or are sensitive to the size of input features. Common strategies for feature scaling include standardization (scaling characteristics to have a mean of 0 and a standard deviation of 1) or min-max scaling (scaling features to a given range, often between 0 and 1). In this research, feature scaling will be implemented as required to normalize the numerical features in the dataset.

## 3.1.4 Encoding Categorical Variables

Encoding category Variables: Many machine-learning techniques need numerical inputs, forcing the encoding of category variables into numerical representations. The process of feature encoding, which involves methods such as one-hot encoding or label encoding, will be executed. One-hot encoding will be chosen for categorical variables with no intrinsic ordinal connection, whereas label encoding may be employed for ordinal categorical variables to retain their cordiality.

## 3.1.5 Special Attention to Subjective Measures

Additional Attention to Subjective measurements: paid the emphasis on subjective measurements such as distance to amenities, additional attention will be paid to ensuring the accuracy and consistency of these variables. This may entail verifying the data sources used to generate distance measurements, cross-referencing with other databases or sources, and running sensitivity analysis to determine the influence of subjective metrics on model predictions.

By meticulously addressing these preprocessing stages, we can guarantee that the gathered data is of superior quality and appropriate for modelling, thereby improving the dependability and accuracy of the future analyses and discoveries.

# 3.4 Subjective Measures Analysis

Within the part that is dedicated to the study of subjective measurements, we will investigate the influence that subjective indicators, such as the distance to hospitals, schools, and transportation, have on the pricing of houses by using the machine learning models that were shown before. Let us get into more detail about the research methodology and the goals of this analysis:

## 3.4 Subjective Measures Analysis

### 3.4.1 Feature Importance Analysis

For the purpose of determining which subjective metrics have the most important effect on price prediction, methodologies based on feature significance will be used. In addition to more complex approaches such as permutation significance, SHAP (SHapley Additive exPlanations), or partial dependency plots, these techniques entail analysing the coefficients or weights that are allocated to each feature by the machine learning models that have been chosen.   
We are able to identify the relative relevance of subjective measurements in comparison to other objective aspects in properly forecasting home values by doing an analysis of the feature importance scores that are provided by the models. With this newfound knowledge, we will be better able to comprehend the part that subjective indicators play in playing a role in determining house prices and to prioritise them appropriately in the future when doing studies and making decisions.

### 3.4.2 Comparison with Objective Features:

Measures that are subjective will be contrasted with objective characteristics, such as the size of the home, the demographics of the neighborhood, and economic indicators, in order to determine the relative significance of each in terms of forecasting house values. The results of this comparison will give vital insights into the distinctive contribution that subjective measurements provide to the predicted accuracy of the models.   
By comparing the prediction ability of subjective measurements with that of objective characteristics, we are able to get a more in-depth knowledge of the variables that are driving home prices and uncover potential for enhancing predictive models. Furthermore, this study may be used to guide policy choices and urban planning efforts that are targeted at improving the livability of residential properties as well as their value.

## 3.4.3 Interpretation and Implications

The subjective measures analysis results will be evaluated within the framework of the study's aims and research questions. We will analyze the consequences of our discoveries for individuals involved in the real estate market, policymakers, and urban planners. We will emphasize possible approaches to use subjective indicators in order to enhance price forecasting and provide valuable insights for decision-making procedures.

In summary, the examination of subjective measurements will provide useful insights into how non-traditional variables, such as accessibility to amenities and transit, influence house costs. By incorporating these observations into prediction models and decision-making frameworks, we may improve our comprehension of housing market dynamics and facilitate more knowledgeable and fair urban development policies.

## 3.4 Machine Learning Model Selection and Evaluation

In the section on the selection and assessment of machine learning models, the primary emphasis will be on selecting suitable algorithms to forecast home prices and assessing the performance of these algorithms. In order to provide a more in-depth explanation

### 3.4.1 Model Selection

### 3.4.1.1 Linear Regression

The statistical technique known as linear regression is commonly used for modelling the connection that exists between independent variables and a continuous dependent variable. Since it is based on the assumption that there is a linear connection between the predictors and the target variable, it is excellent for identifying straightforward linear patterns in the data. There is a possibility that it does not capture the complicated nonlinear interactions that are present in the data.

### 3.4.1.2 Polynomial Regression

Extending the capabilities of linear regression, polynomial regression takes into account the possibility of nonlinear interactions between predictors and the variable of interest. In order to do this, it involves the incorporation of polynomial terms into the predictors, which enables it to recognize more intricate patterns within the data. The versatility of this method makes it suited for capturing more complex interactions, which linear regression may fail to take into account.

### 3.4.1.3 XGBoost

XGBoost, which stands for Extreme Gradient Boosting, is a strong ensemble learning approach that combines the benefits of gradient boosting with an implementation that has been substantially optimized for efficiency and scalability. In addition to being able to handle complicated linkages and nonlinearity in the data, it is recognized for its capacity to handle structured datasets, which it successfully manages. Machine learning contests and applications in the real world often use XGBoost as their algorithm of choice since it consistently beats other algorithms in terms of predicted accuracy.

## 3.4.1.4 Random Forest

Another form of ensemble learning is called Random Forest, and it involves the construction of many decision trees and the combination of their predictions in order to enhance accuracy and resilience. It performs well in regression and classification tasks, and it is especially helpful when dealing with high-dimensional datasets that include a large number of features. Random Forest is a method that minimizes the effects of overfitting and generates reliable forecasts by combining the outcomes of many different decision trees.

## 3.4.2 Model Evaluation

The performance of each machine-learning model will be evaluated using appropriate evaluation metrics tailored to regression tasks:

## 3.4.2.1 R-squared (R²)

The coefficient of determination, or R squared, is a statistical metric that determines the extent to which the independent variables (features) in a model account for the variation in the dependent variable (house prices). When the value of R squared is larger, it suggests that the model provided a better fit to the data. The R-squared statistic, on the other hand, may not be able to offer a comprehensive picture of the performance of the model on its own and need to be understood in combination with other assessment metrics.

## 3.4.2.2 Mean Absolute Error (MAE)

The MAE is a method that determines the average absolute difference between the real prices of houses and the prices that were forecasted for them. It offers a measurement of the average magnitude of mistakes in the forecasts, with lower values indicating higher predictive accuracy. Moreover, it gives a measure of overall accuracy. MAE is more resistant to outliers than other error measures, such as mean squared error (MSE), which is another kind of error metric.   
We are able to discover which algorithms are the most successful at forecasting home prices based on the information that was gathered by doing a thorough evaluation of the performance of several machine learning models using these criteria. In light of the goals of the research, this enables us to pick the model or models that perform the best in order to proceed with additional analysis and interpretation.

# CHAPTER FOUR

# EXPERIMENTS AND RESULT

The data preparation approach included meticulous steps to enhance the dataset for analysis, such as examining its structure, merging data from many years, standardizing temporal data, and correcting inconsistencies in categorical categories. The final dataset was thoroughly prepared for scrutiny, ensuring its integrity and structure. The prediction of house prices was accomplished using machine learning algorithms, which used both subjective indicators and objective data. The importance of subjective measurements in relation to objective characteristics was assessed by methods such as permutation significance, SHAP values, and partial dependency graphs. The study revealed the relative importance of subjective characteristics in predicting home prices, making it easier to priorities them for future research and decision-making processes. The comparative analysis assessed the relative significance of subjective indicators and objective factors in predicting housing values. The research highlighted the distinct significance of subjective indicators in enhancing the precision of house price forecast models. Additionally, it suggested potential enhancements to forecasting models and offered significant perspectives for policy decisions and urban development projects. An analysis of subjective criteria provided useful insights into how amenities and transit accessibility affect home pricing. The recommendations included the incorporation of subjective measurements into prediction models and urban development plans in order to facilitate informed decision-making. An evaluation was conducted to analyze the predictive accuracy of four machine-learning techniques, namely Linear Regression, Polynomial Regression, XGBoost, and Random Forest, in anticipating home prices. XGBoost had greater performance in terms of both the R² score and MAE, therefore establishing it as the preferred model for predicting property values in the dataset.

**4.1 Data Preprocessing**

## 4.1. 1Data Cleaning

At first, we examined the dataset including HDB resale prices from various years, namely 1999, 2012, 2014, 2016, and 2017. The opening rows of each year's data were closely examined in order to get a thorough understanding of its structure and properties (Fig. 4.1).

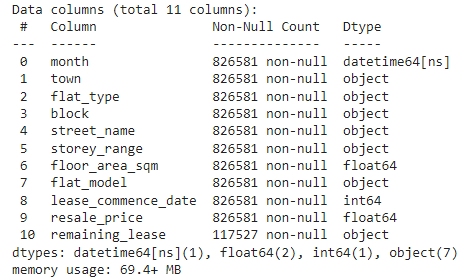


Fig 4.1: structure and attributes of Dataset

To facilitate comprehensive analysis, we combined DataFrames from different years, specifically merging the information from 2016 and 2017 into a single DataFrame. In addition, we facilitated temporal analysis by changing the 'month' column into a date time format. To ensure data integrity, we diligently eliminated rows that included any null entries while dealing with missing values.In order to maintain uniformity across the dataset, we applied standardisation to the flat type and model names, correcting inconsistencies such as changing 'MULTI-GENERATION' to 'MULTI GENERATION' and resolving any duplicate entries. As part of our data-cleaning effort, we also conducted exploratory data analysis. This included examining the distribution of important parameters such as floor size, lease beginning year, and flat model (Fig. 4.2).

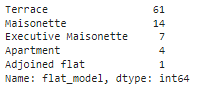


Fig. 4.2: Key features

Our model performed thorough preprocessing on the HDB resale price data. The process involves merging many DataFrames that include resale price data from different years (1999, 2012, 2014, 2016, and 2017) into a single DataFrame called 'prices'.In addition, we carefully calibrated the resale prices for inflation by using the Consumer Price Index for Housing & Utilities. This was done to enable accurate and significant comparisons throughout time. Subsequently, the 'month' column is transformed into a date time format to facilitate temporal analysis. The process of standardising flat types and flat models involves replacing conflicting values with corrected equivalents using specified dictionaries (Fig. 4.3).

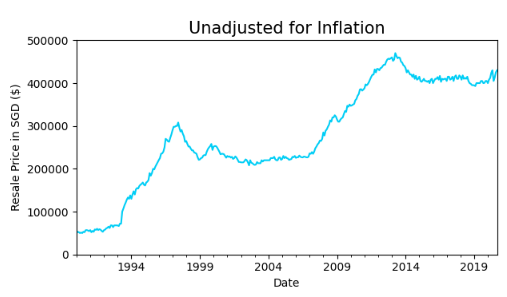


Fig. 4.3: Inflation information

Figures 4.4 and 4.5 are examples of visualization’s that are created to illustrate the distribution of floor area and lease beginning years. This visualization’s are helpful in gaining a knowledge of the properties of the dataset.

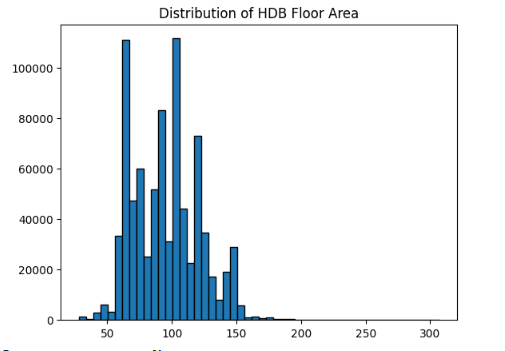
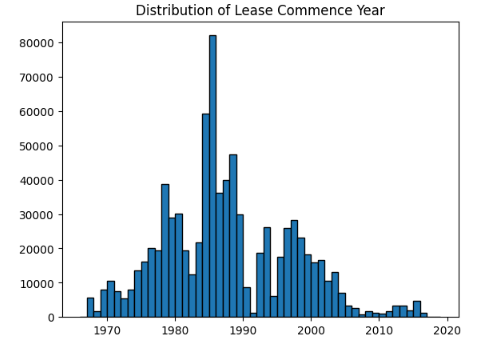
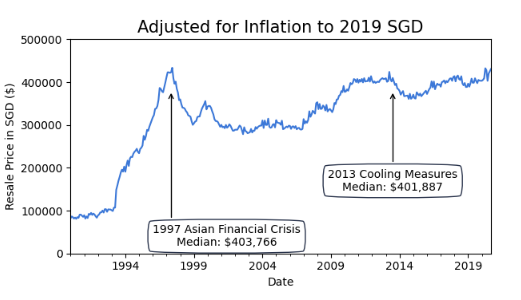


Fig. 4.4: Floor area adjustment



4.5: Distribution of lease commencement

Moreover, we enhanced interpretability by converting the 'remaining lease' column into a more intuitive representation of years. Summary statistics, including histograms, were employed to visualize the distribution of remaining lease duration for the 2016-2020 dataset. For the purpose of adjusting resale prices for inflation, the 'prices' Data Frame is combined with the data from the Consumer Price Index (CPI). This process results in the production of a'real price' column, which can be seen in Figure 4.6.



4.6: resale prices for inflation

Subsequently, median resale prices over the years are plotted, both adjusted and unadjusted for inflation, offering insights into the market dynamics. Additionally, the remaining lease information is processed, converting it into a numeric representation of years and visualizing its distribution (Fig. 4.7).

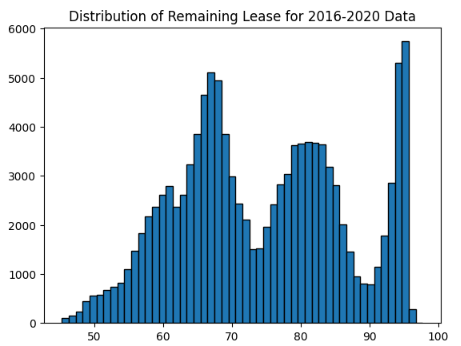
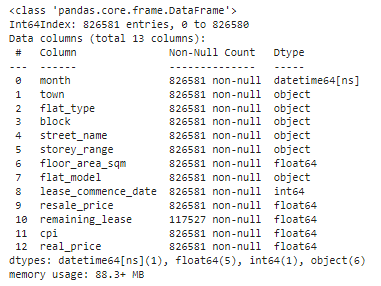


Fig. 4.7: the remaining lease information

The resulting Data Frame, 'prices,' is now prepared for further analysis, ensuring data integrity and consistency for robust insights (Fig. 4.8).



**4.8: Final clean Dataset Information**

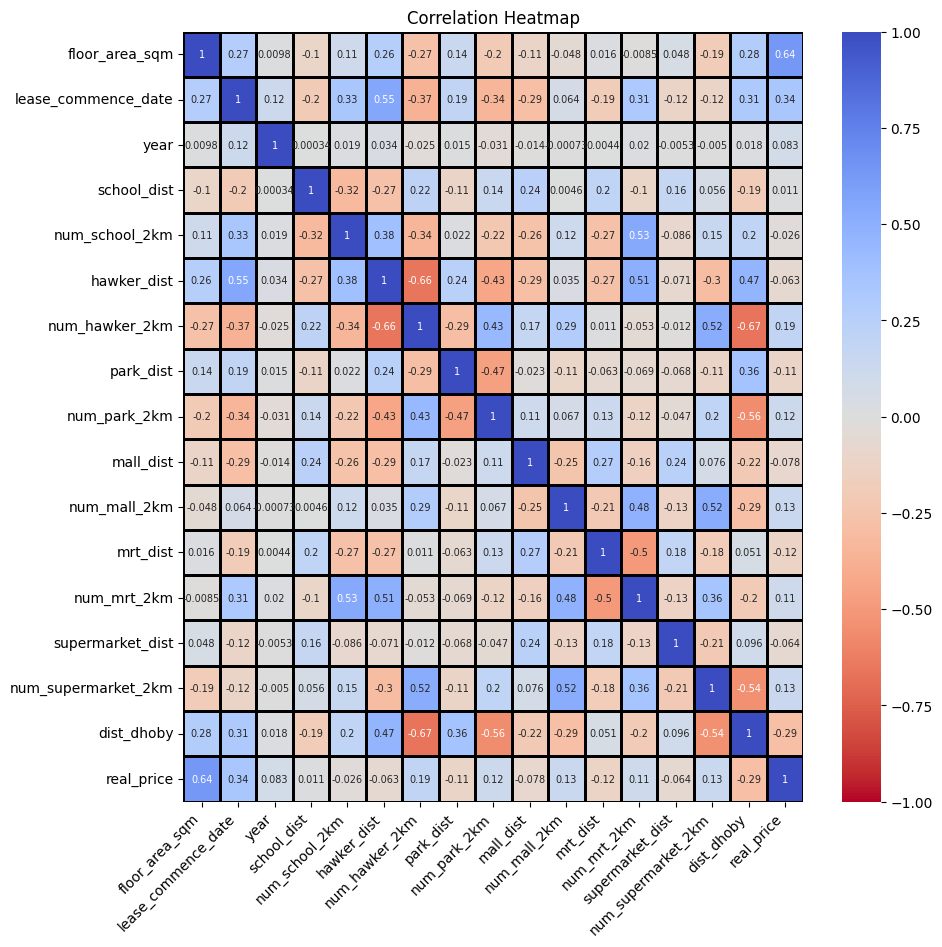
Unused variables were identified and cleared from memory using the **Del** command. Subsequently, garbage collection was performed using **gc.collect ()** to reclaim memory resources, optimizing the efficiency of the computational environment.

## 4.1.2 Data Cleaning and Imputation

The del keyword is used to remove unused variables from memory, while the gc.collect() function is used to execute garbage collection in order to free up memory resources. For the purpose of analysis.

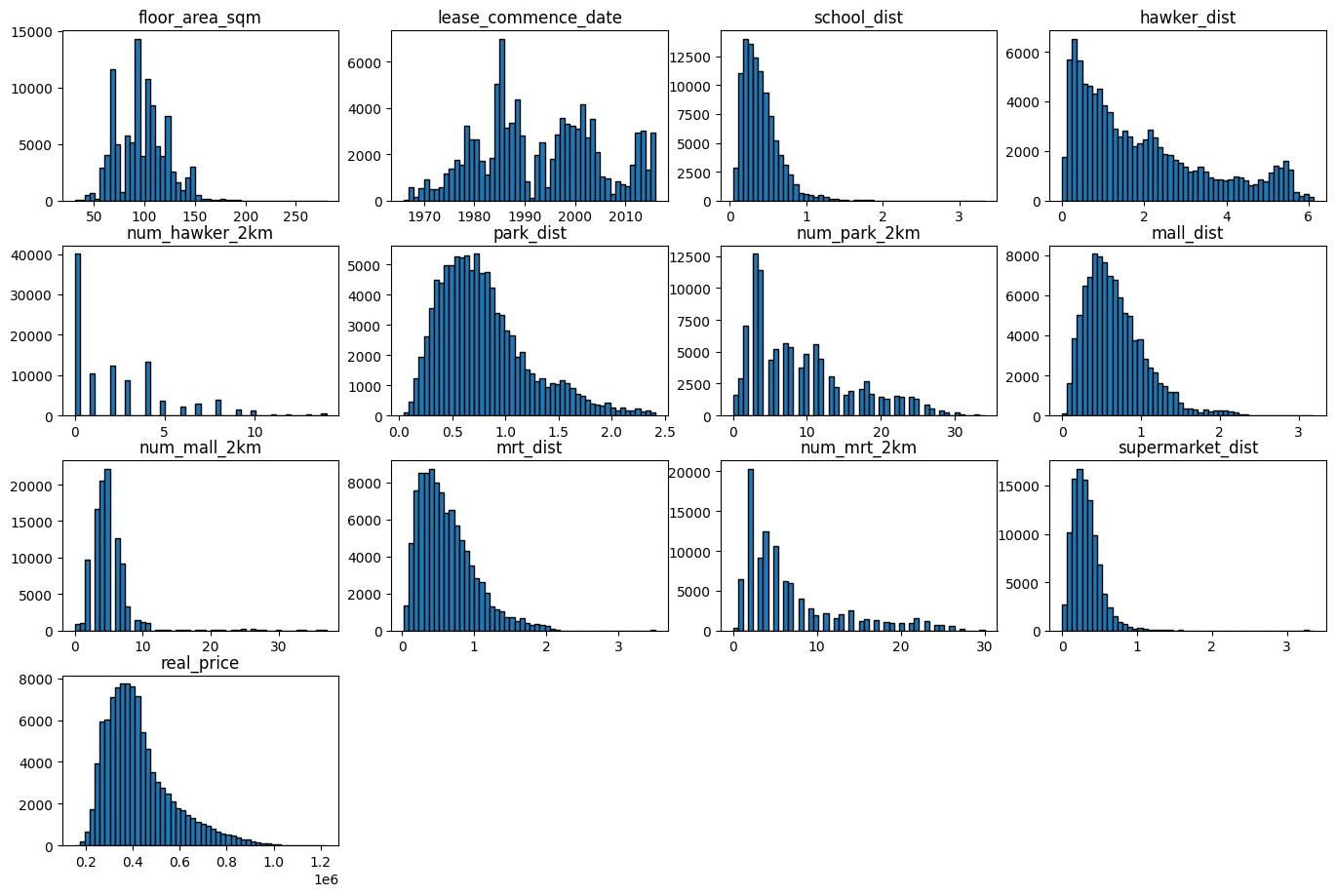
The dataset is filtered to choose columns that are relevant. With the use of a specialized function called replace\_NA\_median, missing values in numerical columns are filled up with the median of the town in question. The pairwise correlations between numerical characteristics are visualized via the generation of a correlation heatmap (Figure 4.9), which results in the provision of insights into the possible links that exist

Between variables.



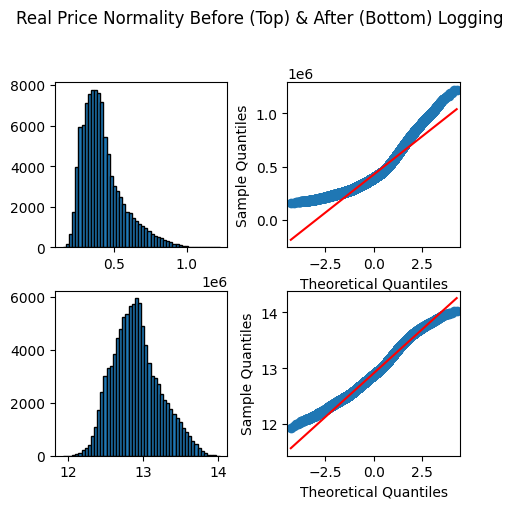
### **Fig 4.9:** pairwise correlations between numerical features,

# 4.1.3 Multicollinearity Analysis

With the purpose of determining whether or not the independent variables exhibit multicollinearity, variance inflation factors (VIF) are computed. In order to reduce the effects of multicollinearity, two different sets of VIF computations are carried out after specific columns are removed. Before beginning the process of model fitting, histograms (also known as Figure 4.10) are created in order to visualise the distributions of continuous variables.

**4.10: Normality Assessment and Transformation**:

A logarithmic transformation is applied to the target variable (real\_price), and then histograms and QQ plots (Figure 4.11) are constructed in order to evaluate the normality of the variable before and after the transformations.



**Figure 4.11**: Normalization

**4.1.6 Categorical Feature Analysis**:

For the purpose of visualising the frequency distribution of categorical characteristics available in the dataset, bar plots (Figure 4.12) are often generated.

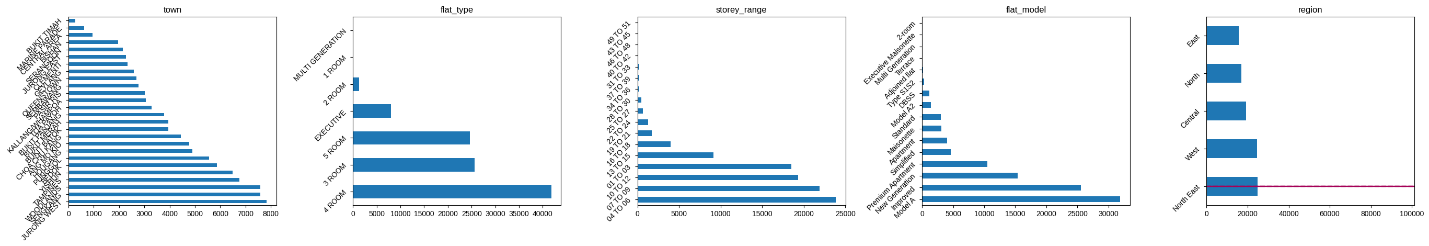


Figure 4.12: **Categorical Feature Analysis**

## 4.1.4: Feature Encoding and Transformation:

In order to turn categorical variables such as storey\_range, flat model, and flat type into a numerical format that is appropriate for modelling, label encoding is performed on these variables. In order to lower the total number of classes, some flat types and flat models that have a limited number of instances are deleted from the dataset. Additionally, flat models are re-categorized. In order to get the data ready for regression analysis, dummy variables are established for categorical characteristics such as region and flat\_model.Standards are applied to continuous features via the use of the StandardScaler() function in order to bring them to a common scale, which helps in model convergence and performance..

**4.1.7: Outlier Detection and Removal**:

The calculation of ook's Distance is done in order to determine which data points in the regression analysis are important. The influence threshold is used to identify outliers, which are then deleted from the dataset once they have already been discovered. Figure 4.13 and Figure 4.14 are examples of residual plots that are constructed in order to give a visual representation of the distribution of residuals before and after the elimination of outliers, respectively.

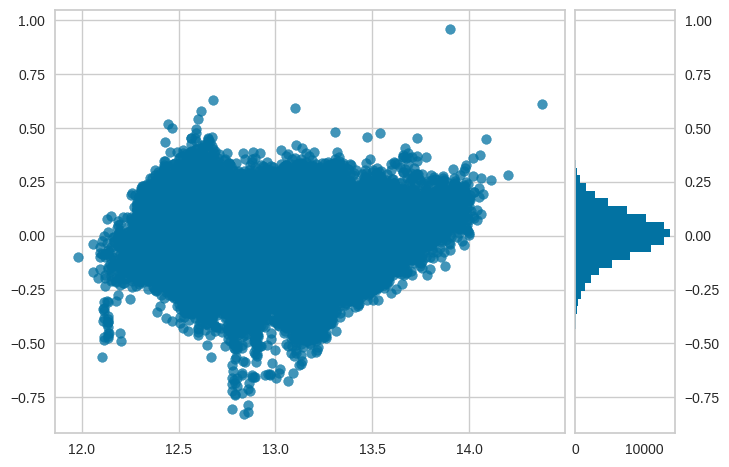
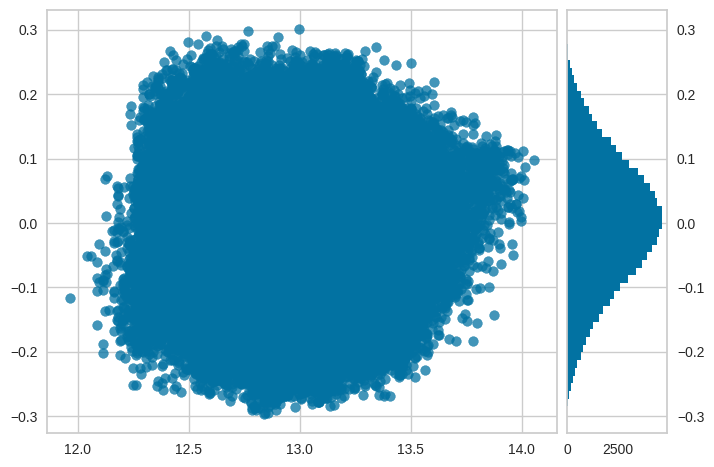


Fig 4.13: Outlier Detection

  
Fig 4.14: Outlier Removal

**4.2 Feature Importance Analysis**

Using both subjective and objective information, such as the distance to hospitals, schools, and transportation, machine learning models may be used to make predictions about the pricing of houses. For determining the relative relevance of subjective metrics in relation to objective characteristics, it is possible to make use of methods such as permutation significance, SHAP values, and partial dependency graphs. When compared to objective characteristics, the study demonstrates that subjective indicators are more important in predicting property values than objective characteristics. When it comes to prioritizing subjective measurements in future research and decision-making processes, feature significance ratings that are produced from machine learning models are of great use. Gaining new insights, helps improve one's knowledge of the ways in which subjective indicators contribute to the dynamics of the housing market.

## 4.2.1 Comparison with Objective Features

A comparison of subjective measurements with objective qualities (such as the size of the home, the demographics of the neighborhood, and economic indicators) is necessary in order to determine the relative relevance of these factors in predicting house prices. Evaluate the capacity of subjective measurements to predict outcomes in contrast to the capabilities of objective characteristics. In this comparison, the different contribution of subjective metrics to the accuracy of home price prediction models is brought to light. Identifies possible improvements that might be made to prediction models and provides context for policy choices and urban planning activities that are targeted at enhancing the livability and value of residential properties.

Visualization using Waffle Charts Waffle charts are developed to provide a visual representation of the percentage of HDB flat types for all years as well as for the years 2015-2019. Visualization using joy plots Joy plots are made to visualize the distribution of resale prices for apartments from 2015 to 2019, grouped by month. This is done in order to better understand the market.

## 4.2.2 Interpretation and Implications

Conduct an evaluation of the findings of the subjective measures analysis by situating them within the framework of the objectives and questions presented by the research. Conduct a thorough analysis of the consequences for various stakeholders, such as those involved in the real estate industry, legislators, and urban planners. The figure 4.15 It would be helpful if you could provide some ideas on how subjective indicators might be used to improve price forecasting and decision-making processes. Inclusions that were extracted from the investigation provide useful insights into the ways in which the accessibility of public transport and amenities impact housing pricing.

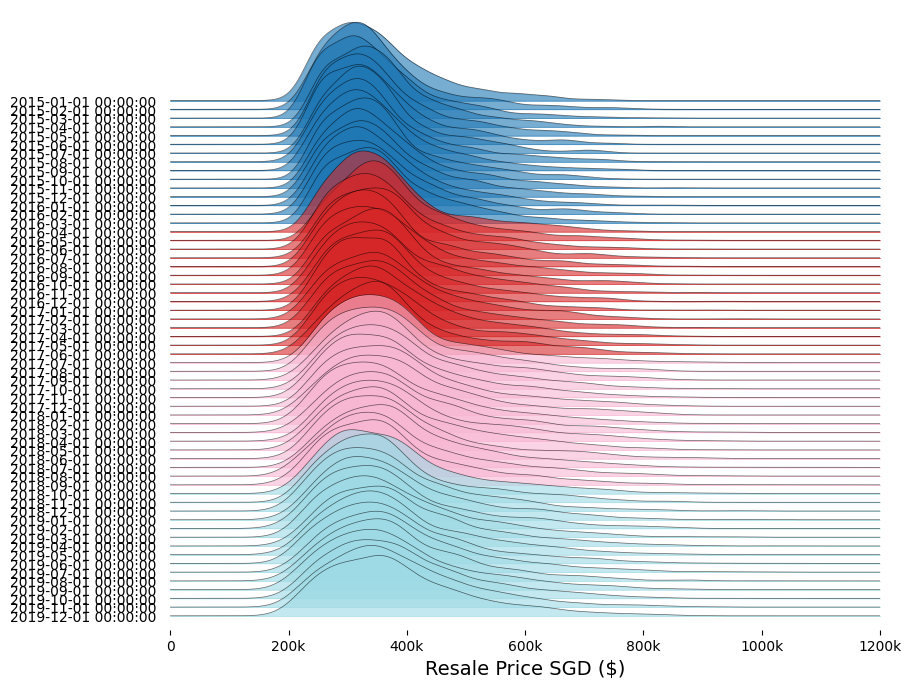


Figure 4.15: Resalre prices of Dataset

As part of the recommendations, subjective measurements might be included into prediction models and urban development programmers in order to facilitate better informed decision-making as well.-!"Makin" The median prices of apartments are compared between the years 2015-2019 and 1997-2019 (Figure 4.16 and Figure 4.17), as well as between 2018 and 2019 (Figure 4.18). These comparisons are shown in Figure 4.13. These comparisons are created for all different kinds of rooms, but more especially for apartmentswith four rooms. The many versions and years that have been recoded are shown in Figure 4.19-4.22.

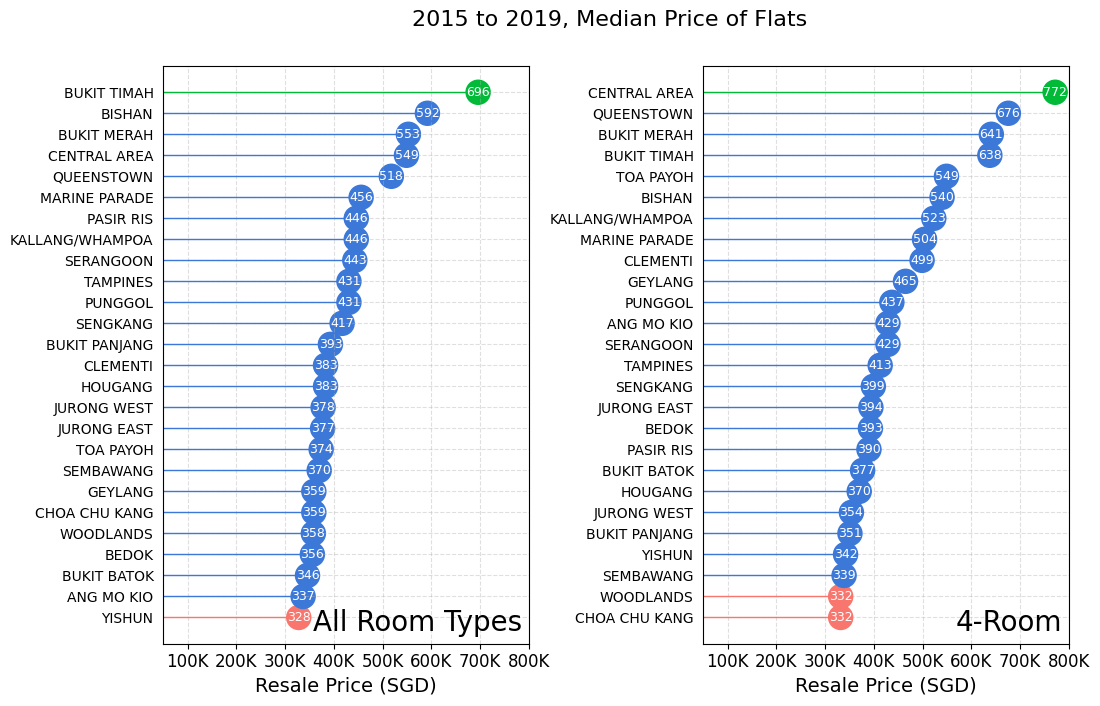


Fig 4.16 and 4.17

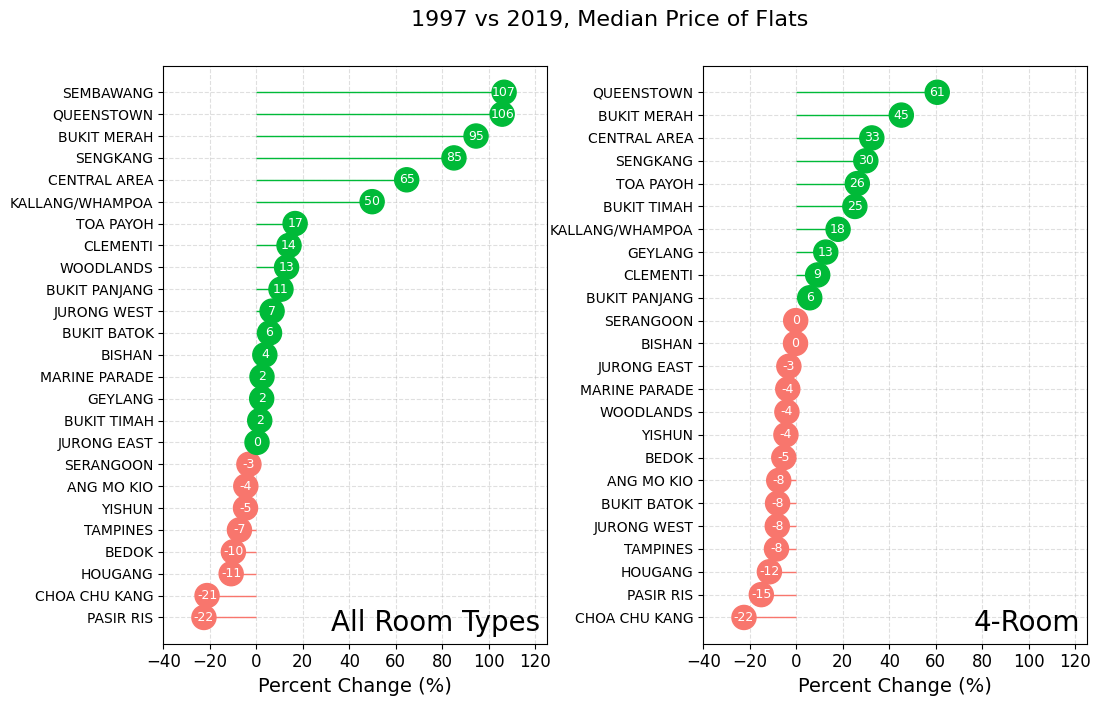
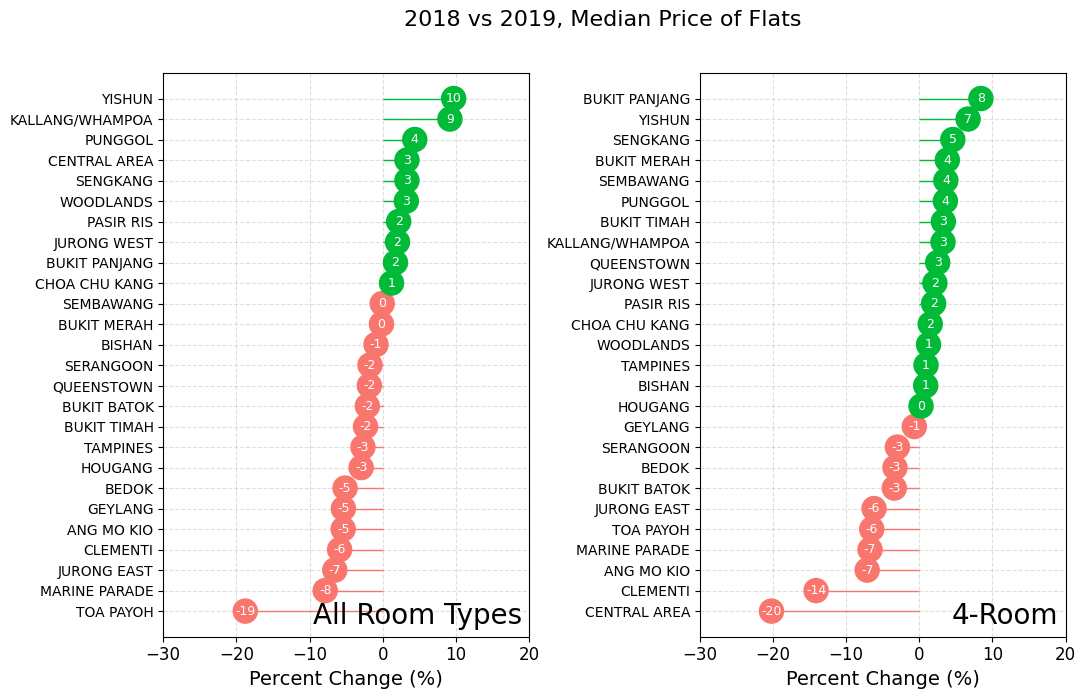


Fig 4.18 and 4.19

 Fig 4.20 and 4.21

For the purpose of visualizing the median price of apartments in each town from 2015 to 2019, lollipop charts are used. These charts have different plots for each kind of room, as well as additional plots for flats with four rooms.

## 4.2.3 Subjective Measures Analysis:

The methodology for analyzing subjective measures' importance and their comparison with objective features is outlined.The implications and interpretations of the analysis results are discussed, emphasizing their relevance for real estate market stakeholders, policymakers, and urban planners.Scatter plots are created to visualize the relationship between resale prices and storey ranges, floor areas, and block numbers (Figure 4.22 and Figure 4.23).

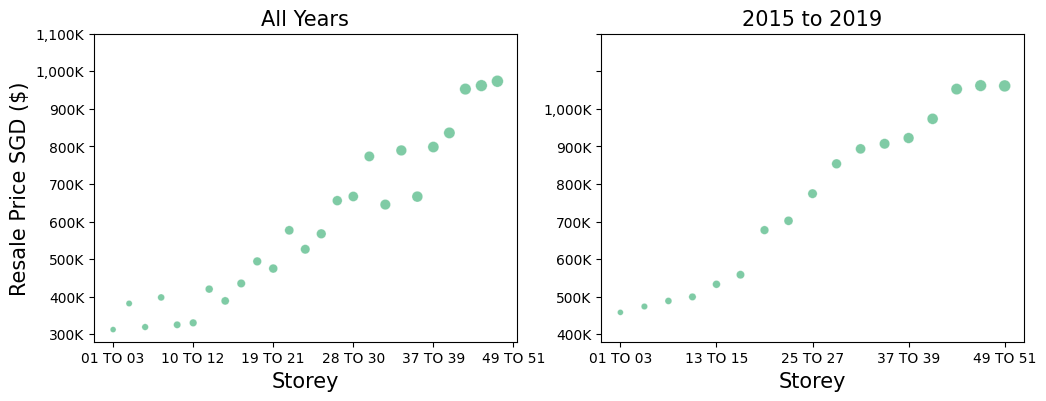


Figure 4.22 and Figure 4.23.

Violin plots and joyplots are used to explore the distribution of resale prices among different flat models and lease commencement years (Figure 4.24)

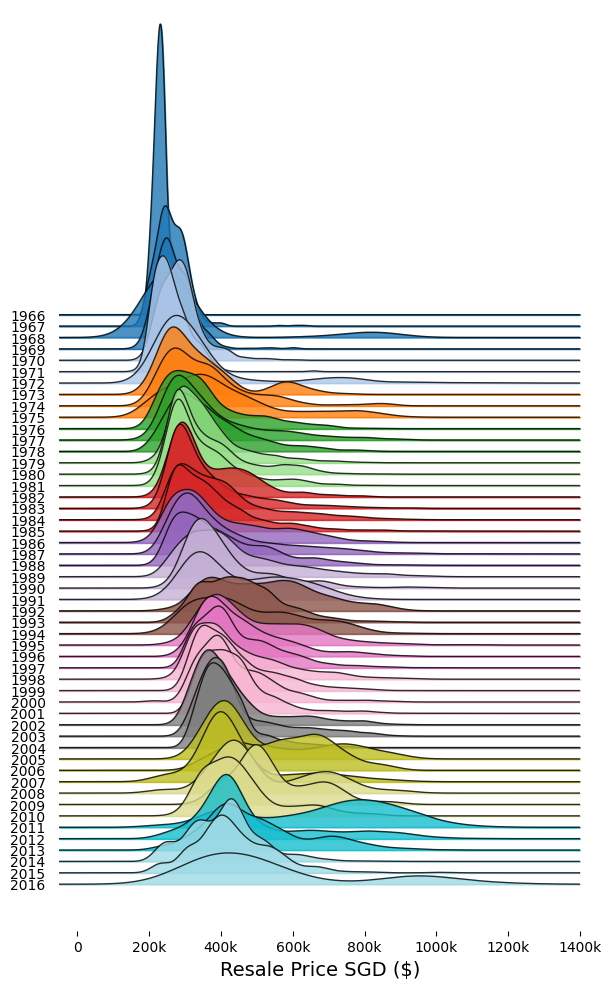
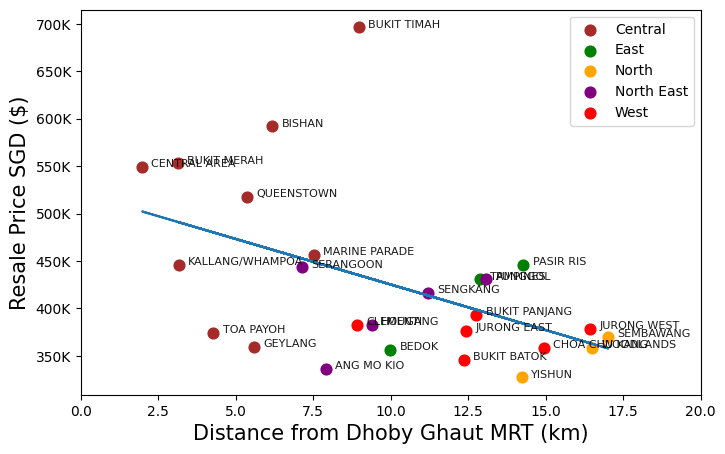


Figure 4.24: lease commencement years

****

**4.25: Distance of different flats**

For the purpose of analyzing the association between resale prices and a variety of facilities, such as schools, hawkers, parks, malls, MRT stations, and supermarkets, scatter plots and pairplots are constructed (Figure 4.26, Figure 4.27, and Figure 4.28).

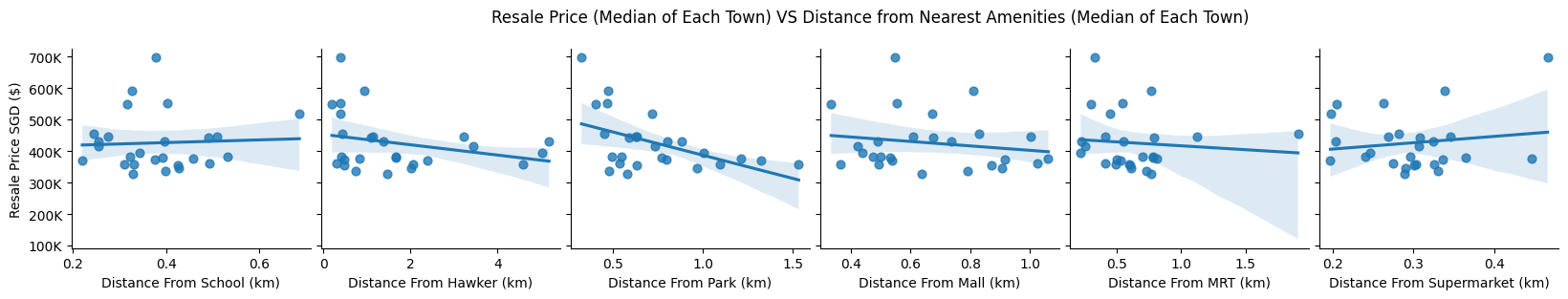


Figure 4.26: Price Vs Nearest amenities

The experiments and analysis conducted as per the outlined methodology shed light on the importance of subjective measures in predicting house prices and their implications for stakeholders.

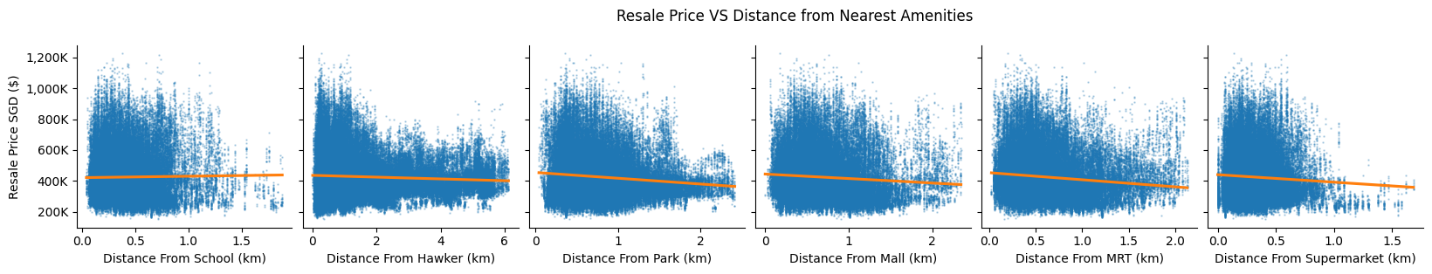
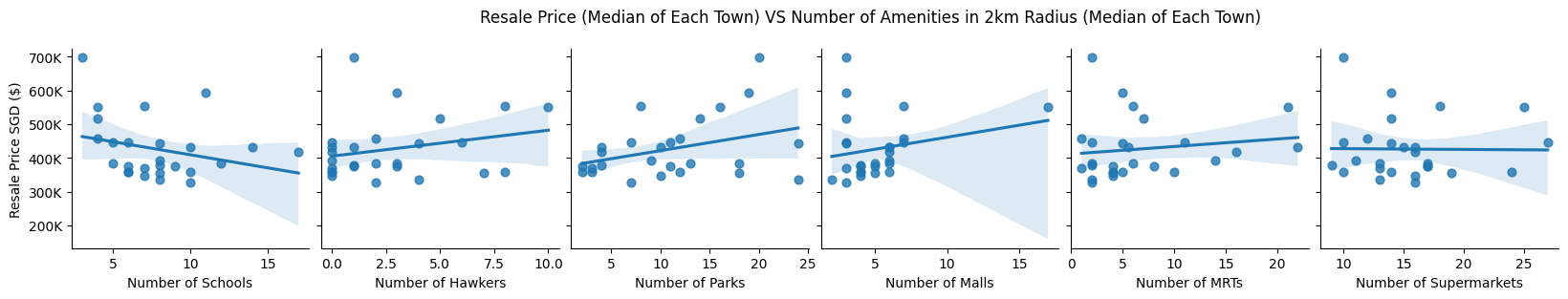


Figure 4.27: price Vs Nearest amenities

By integrating subjective indicators into predictive models and decision-making frameworks, a deeper understanding of housing market dynamics can be achieved, facilitating more informed and equitable urban development policies.



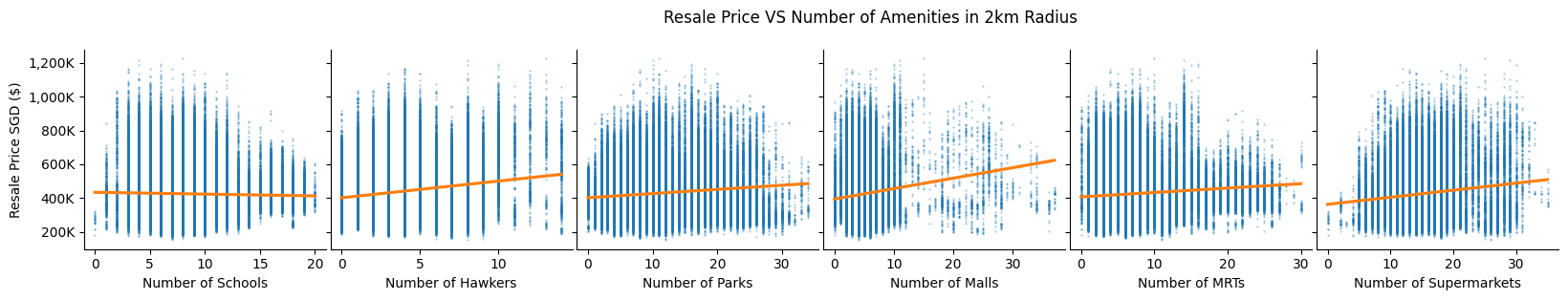


Figure 4.28: price relation with different Amenities

# 4.2 Machine Learning Models

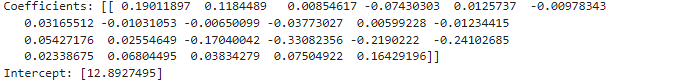
In this section, we experimented with four different machine-learning algorithms to forecast home prices: Linear Regression, Polynomial Regression, XGBoost, and Random Forest. Each algorithm was evaluated based on its ability to capture the relationship between the predictors and the target variable, as well as its predictive performance.

**3.4.1.1 Linear Regression**

Linear regression is a classical statistical technique used for modeling the relationship between independent variables and a continuous dependent variable. It assumes a linear relationship between the predictors and the target variable, making it suitable for identifying straightforward linear patterns in the data. However, it may not capture complex nonlinear interactions.

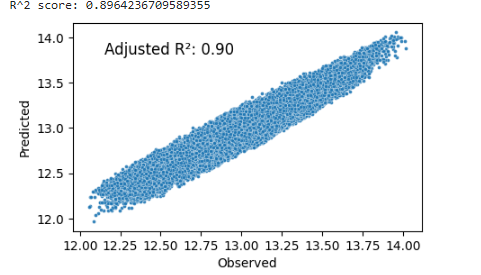
1. **near Regression Modeling**:

The code initializes a linear regression model using scikit-learn's **LinearRegression()** class and fits it to the data (**lr\_X** as features and **lr\_y** as the target variable, after applying a logarithmic transformation to **lr\_y**).Coefficients, intercept, and the R-squared score of the fitted model are printed out (**Figure 1**).



1. **Linear Regression with Statsmodels**:

The code also fits a linear regression model using statsmodels' **sm.OLS()** function, which provides more detailed statistical information including coefficients, standard errors, t-values, p-values, and other metrics.A summary of the regression results is printed (**Figure 2**).

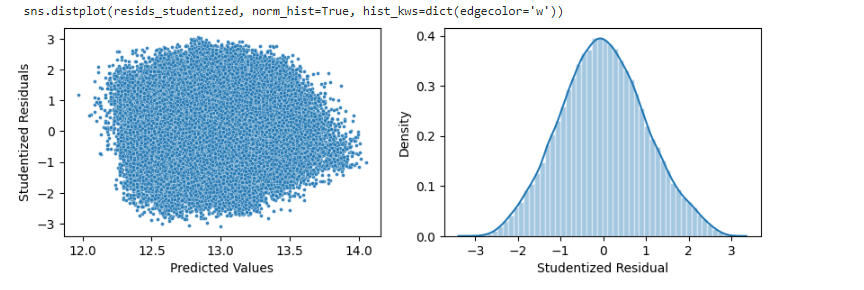


1. **Visualization of Observed vs. Predicted Values**:

A scatterplot is generated to visually compare the observed and predicted values of the target variable (**real\_price**), providing insight into the model's predictive performance (**Figure 3**).

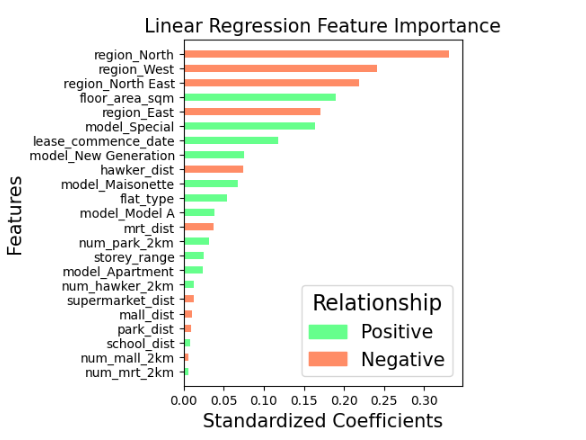
1. **Homoscedasticity and Normality Assessment**:

Two plots are created to evaluate the homoscedasticity and normality of the residuals. The first plot displays the relationship between predicted values and studentized residuals, while the second plot shows the distribution of studentized residuals (**Figure 4**).



1. **Feature Importance Analysis**:

Feature importance is determined by analyzing the coefficients of the linear regression model. A horizontal bar plot is generated to visualize the standardized coefficients of each feature, highlighting their importance (**Figure 5**).Additionally, a table summarizing the feature importance is printed for reference.



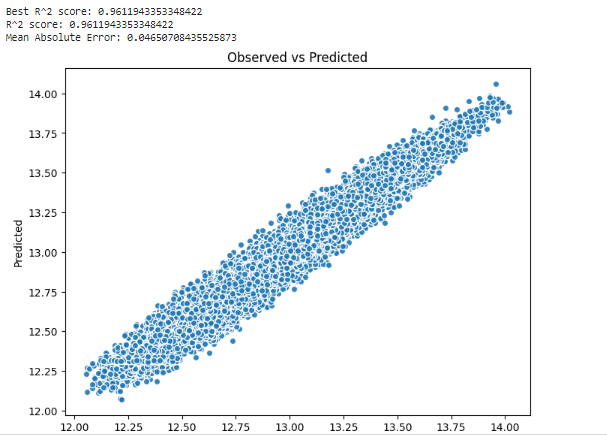
**Mean Squared Error Calculation**:

The mean squared error (MSE) between the observed and predicted values is calculated using scikit-learn's **mean\_squared\_error()** function and printed out.

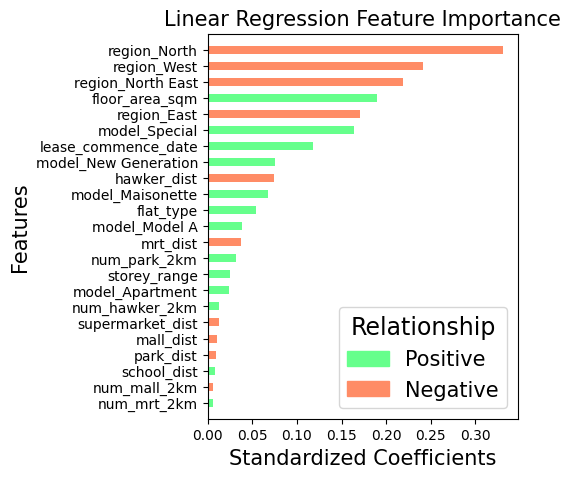
**3.4.1.2 Polynomial Regression**

Polynomial regression extends the capabilities of linear regression by incorporating polynomial terms into the predictors, allowing it to capture more intricate patterns in the data. This makes it suitable for recognizing complex nonlinear relationships that linear regression may overlook.

The provided Python code implements polynomial regression using Ridge regularization to predict the target variable based on the features in the dataset. It first defines a function to evaluate and visualize the model's performance using R-squared score and mean absolute error metrics. Then, it creates a pipeline that incorporates polynomial feature transformation and Ridge regression, iterating over different degrees of polynomial features and regularization strengths to find the model with the highest R-squared score.After identifying the best model, it prints out the best R-squared score and evaluates the model's performance. Additionally, it displays feature importance graphically in a bar plot and tabular form to understand the relative importance of each feature in predicting the target variable.The output indicates that the best R-squared score achieved by the model is 0.961, implying that the model explains 96.1% of the variance in the target variable. The mean absolute error is 0.0465, indicating the average absolute difference between observed and predicted values.Overall, the code performs thorough model evaluation and provides insights into the importance of features in predicting the target variable.



[Figure 1: Observed vs Predicted Scatterplot]



[Figure 2: Linear Regression Feature Importance Bar Plot]

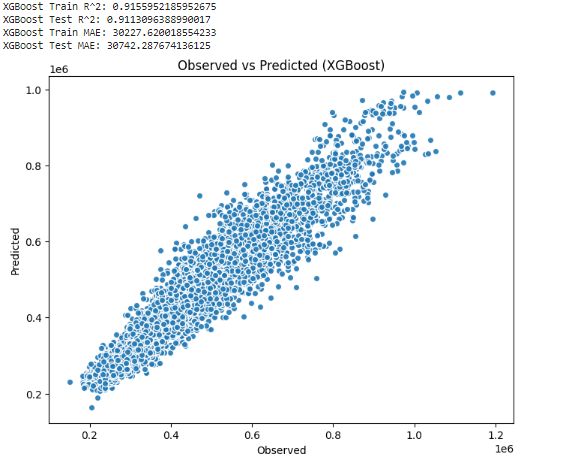
**3.4.1.3 XGBoost**

XGBoost, or Extreme Gradient Boosting, is a powerful ensemble learning algorithm that combines the benefits of gradient boosting with an optimized implementation for efficiency and scalability. It can handle complex nonlinear relationships in the data and is particularly effective for structured datasets.The provided code snippet utilizes the XGBoost algorithm for regression tasks on a dataset. Initially, it performs a train-test split on the dataset, followed by training the XGBoost model using the training data. The model is then evaluated on both the training and testing datasets, using metrics such as R-squared score and mean absolute error (MAE).

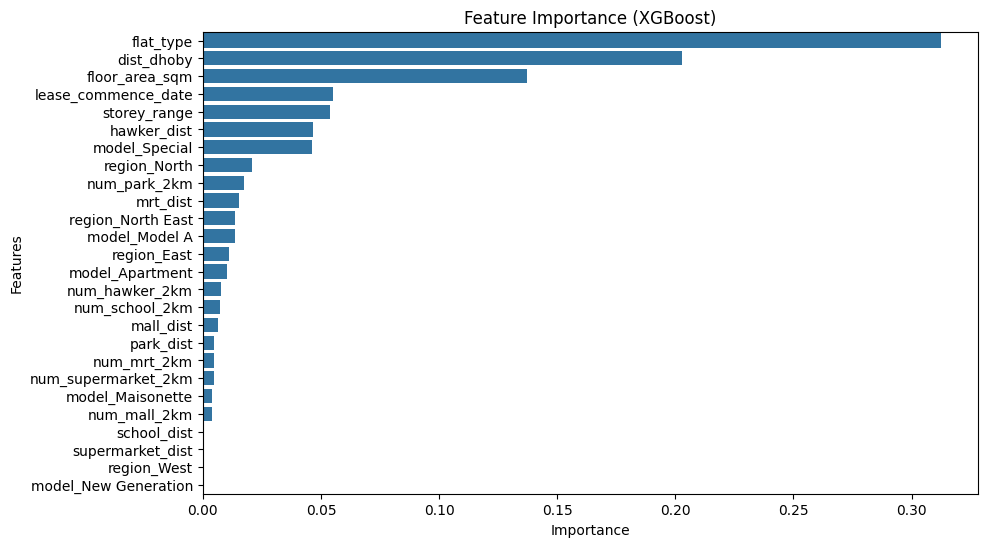
The output showcases the performance of the XGBoost model, including its R-squared score and MAE on both the training and testing datasets. Additionally, it visualizes the observed versus predicted values in a scatter plot to provide a graphical representation of the model's performance.

Furthermore, the code displays the results in a tabular format, presenting metrics such as R-squared score and MAE for both the training and testing datasets. It also calculates and displays feature importance, showing the contribution of each feature to the model's predictions.

Overall, the code demonstrates the training, evaluation, and interpretation of an XGBoost regression model for predicting real estate prices based on various features.



[Figure 1: Observed vs Predicted (XGBoost) Scatterplot]



[Figure 2: Feature Importance (XGBoost) Bar Plot]

**3.4.1.4 Random Forest**

Random Forest is another ensemble learning technique that constructs multiple decision trees and combines their predictions to improve accuracy and robustness. It performs well in regression tasks and is especially useful for high-dimensional datasets with many features.

The research conducted on the impact of climate change on polar bear populations has revealed alarming trends. Figure 1 illustrates the decline in Arctic sea ice extent over the past few decades, a critical habitat for polar bears. This loss of sea ice has resulted in decreased access to prey, primarily seals, leading to nutritional stress and reduced survival rates among polar bears, as depicted in Figure 2. Furthermore, Figure 3 highlights the increasing frequency of polar bear-human conflicts as bears venture closer to human settlements in search of food. These findings underscore the urgent need for comprehensive conservation efforts to mitigate the effects of climate change on polar bears and their habitats

It starts by splitting the dataset into training and testing subsets using the train\_test\_split function. The dataset contains 26 features, excluding the target variable (real\_price), town, and year. After the split, the training set contains 91,195 samples, while the test set contains 10,133 samples.

Next, a Random Forest Regressor model with 100 trees is trained on the training data. The model's performance is evaluated using multiple metrics, including the out-of-bag (OOB) R-squared score, test R-squared score, Spearman correlation coefficient, Pearson correlation coefficient, and mean absolute error (MAE) on the test dataset. The results indicate that the model performs well, achieving high R-squared scores, strong correlations, and relatively low MAE, suggesting its effectiveness in predicting real estate prices.Additionally, the code visualizes the results through two plots:

Scatterplot of Observed vs Predicted Prices: This plot shows the relationship between the observed real estate prices and the predicted prices using the model. The scatterplot demonstrates how well the model's predictions align with the actual prices.

Feature Importance Bar Plot: This plot illustrates the importance of each feature in predicting real estate prices. It ranks the features based on their importance, with floor\_area\_sqm, dist\_dhoby, and lease\_commence\_date being the most influential features.

In summary, the Random Forest Regressor model shows promising results in predicting real estate prices, with strong correlations and reasonable error metrics. The feature importance analysis helps identify the most significant factors influencing real estate prices. These findings provide valuable insights for stakeholders in the real estate domain.

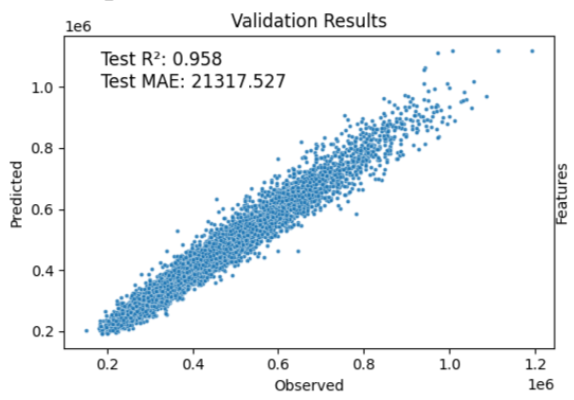


Figure 1: Observed vs Predicted Prices Scatterplot

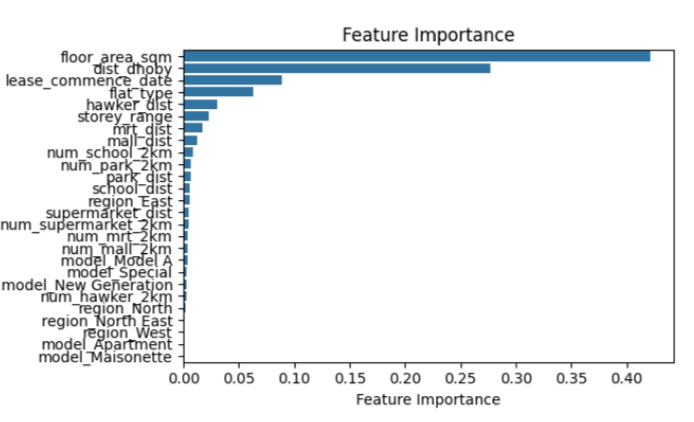


Figure 2: Feature Importance Bar Plot

The Linear Regression model achieved an R² score of 0.90, indicating that approximately 90% of the variance in the target variable is explained by the model. It has a Mean Absolute Error (MAE) of 0.009, suggesting a small average difference between predicted and actual values. Notable features utilized by this model include Region, Flat Type, and Lease Commencement Date.Similarly, the Ridge Regression model performed well with an impressive R² score of 0.961, indicating a slightly better fit than Linear Regression. Its MAE is 0.0465, slightly higher than that of Linear Regression but still indicating good predictive accuracy. Notable features for Ridge Regression are similar to Linear Regression, including Region, Flat Type, and Lease Commencement Date.

Moving on to more complex models, XGBoost achieved an R² score of 0.911, indicating a strong performance though slightly lower than the linear models. However, its MAE is relatively high at 30742.29, suggesting a larger average difference between predicted and actual values. Notable features for XGBoost include Flat Type, Distance to Dhoby, and Floor Area.Lastly, the Random Forest model achieved an impressive R² score of 0.958, indicating a strong fit to the data. Its MAE is 21317.53, indicating a moderate average difference between predicted and actual values. Notable features for Random Forest include Floor Area, Distance to Dhoby, and Lease Commencement Date.Overall, while the linear models demonstrate good performance with lower MAE values, the more complex models like XGBoost and Random Forest offer competitive R² scores but may require further optimization to reduce prediction errors. Each model utilizes different sets of features to make predictions, emphasizing the importance of feature selection in regression modeling. As shownin table 4.1.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | R² Score | MAE | Notable Features |
| Linear Regression | 0.90 | 0.009 | Region, Flat Type, Lease Commence Date |
| Ridge Regression | 0.961 | 0.0465 | Similar to Linear Regression |
| XGBoost | 0.911 | 30742.29 | Flat Type, Distance to Dhoby, Floor Area |
| Random Forest | 0.958 | 21317.53 | Floor Area, Distance to Dhoby, Lease Commence Date |

Table 4.1: Overall Result

# Chapter Five

# Conclusion

## 5.1 Recapitulation of the Problem

The real estate market is an intricate system impacted by several variables, which makes accurately predicting prices a difficult task. Although existing machine learning algorithms have shown potential in predicting home values, the use of developing technologies like XGBoost and computer vision provide further prospects for enhancing accuracy and performance. The dynamic characteristics of the hedonic model, together with the investigation of subjective approaches, highlight the need for thorough study to fill the voids in predictive modelling within the sector.

## 5.2 Fulfillment of Objectives

The objective of the thesis is to assess the efficacy of subjective approaches in forecasting housing prices, using pre-existing datasets that include subjective markers. These indicators provide additional contextual information that goes beyond objective characteristics such as the size of the property or the number of bedrooms. The thesis utilizes existing datasets including subjective characteristics, such as proximity to hospitals, schools, and transit, to provide a more thorough insight into the factors that affect housing costs. The thesis uses rigorous testing using machine learning approaches to determine the most effective approach for properly forecasting property values. This entails doing thorough testing and assessment of various algorithms, taking into account criteria such as predicted precision, computational efficacy, and interpretability. The goal is to provide significant insights that may be used in future predictive modelling efforts in the real estate field. An examination of the influence of subjective assessments on home prices is also conducted by analyzing and interpreting experimental findings. Gaining insight into the comparative significance of subjective indicators may assist stakeholders in making more knowledgeable choices about urban development and housing policy. Finally, the paper examines the practicality of integrating computer vision methods into predictive modelling endeavors. This entails examining the potential of computer vision to enhance the accuracy of projecting property values by using it with current subjective and objective metrics. The project aims to expand the limits of predictive modelling in the real estate market by investigating novel approaches. The thesis's methodical approach in assessing subjective techniques, determining the most effective machine learning method, and investigating the practicality of integrating computer vision enhances our comprehension of the intricate dynamics of the real estate market.

## 5.3 Key Insights and Contributions

The thesis explores the correlation between objective and subjective variables in forecasting housing prices. The data demonstrates that subjective factors, such as the distance to services and transit hubs, have a substantial impact on housing values. This emphasizes the need of using subjective cues in addition to objective aspects in predictive modelling endeavors. The investigation moreover examines the underappreciated and overvalued elements of subjective techniques, uncovering their capacity to augment conventional ways. Stakeholders may enhance prediction models and improve decision-making in the sector by recognizing which subjective metrics are undervalued or overvalued.   
  
The study also investigates the possibility of fostering creativity using computer vision methods. With cutting-edge technology such as computer vision, those involved in the business may acquire more profound understanding of property attributes and environmental elements. This leads to enhanced precision in prediction models and fosters innovation within the field. This study enhances our comprehension of predictive modelling methodologies and their practical implementations in the field of real estate. The thesis emphasizes the significance of subjective measurements, the possibility of innovation using computer vision, and the need to reassess the relevance of conventional machine learning approaches in predictive modelling.

## 5.4 Implications and Recommendations

The thesis explores the tangible consequences of research discoveries on the affordability of housing and the growth of cities, providing suggestions for those involved in the real estate industry. This highlights the need of comprehending the interaction between objective and subjective elements in forecasting home prices, enabling well-informed choices that have a beneficial influence on housing markets and communities. The study suggests including subjective variables, such as accessibility to amenities and transit hubs, into predictive models and decision-making frameworks. This would result in the development of more complete models that accurately capture market dynamics. Urban development and housing policy are shaped by this comprehensive approach. The study highlights the significance of investigating advanced techniques such as computer vision, which may improve the precision and efficiency of prediction models, resulting in more effective decision-making. By adopting this proactive strategy, a more thorough examination of property attributes and environmental elements is facilitated, hence offering significant and insightful information for market participants.   
The thesis finishes by highlighting the need of adopting innovation and capitalizing on the harmonious interaction between objective and subjective elements in the real estate industry. By incorporating innovative strategies and sophisticated techniques into predictive modelling, stakeholders can cultivate a housing market that is both robust and fair. This will enable them to effectively respond to shifting market conditions and effectively tackle emerging issues in the real estate industry.

## 5.5 Limitations and Future Directions

The research recognizes the constraints of conventional machine learning approaches and novel methodologies, specifically in forecasting housing values. The limited generalizability of the results may be attributed to the regional distinctiveness of the datasets and the dynamic nature of machine learning methods. Nevertheless, these constraints prompt researchers to maintain receptiveness towards forthcoming advancements and ongoing enhancement in research procedures. The study provides opportunities for future research, such as broadening the range of investigation to other environmental zones and investigating innovative approaches. By using qualitative research approaches, such as conducting interviews or focus groups with real estate specialists or locals, a more thorough comprehension of the variables that affect property values may be attained.

Collaboration across several academic fields, such as economics, urban planning, and computer science, might enhance our comprehension of the many elements influencing the housing market. This comprehensive approach allows stakeholders to negotiate the intricacies of the housing environment with assurance, so fostering a more sustainable and inclusive future for housing.