

Faculty of Information Science and Technology (FIST)

TDA3121 Data Analytics Fundamentals

Trimester 3 2022/2023

**Case Study (40%)**

**Case Study Title: Factor Analysis of Housing Melbourne Price Using Regression Approach**

Due date: **21st September 2023**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| No | Student Name | Student ID | Academic Program | Tutorial Session |
| 1 | LEE EN | 1191100578 | AI | 1B |
| 2 | FOO HAW LIANG | 1191101497 | ST | 1B |
| 3 | SIAH KAH CHUAN | 1191100577 | ST | 1B |
| 4 | GRAYSON GOH JIN YI | 1191101340 | ST | 1B |

**MARKING RUBRICS (40%)**

|  |  |  |  |
| --- | --- | --- | --- |
| **TASKS** |  | **MARKS** | |
| **Report (70%)** | Introduction |  | 5 |
| Methodology |  | 20 |
| Data Set |  | 10 |
| Data Analysis and Result |  | 20 |
| Conclusion |  | 5 |
| Format |  | 5 |
| References |  | 5 |

|  |  |  |  |
| --- | --- | --- | --- |
| **Presentation (30%)** | Clarity |  | 10 |
|  | Presentation Slides |  | 10 |
|  | Teamwork |  | 10 |
|  |  | Total | **/100** |

Contents

[Chapter 1: Introduction 3](#_heading=h.gjdgxs)

[1.1](#_heading=h.30j0zll) Background 3

[1.2](#_heading=h.1fob9te) Big Question & Sub Question 4

[Chapter 2: Methodology 6](#_heading=h.3znysh7)

[2.1 Model used (linear regression, and Random Forest) 6](#_heading=h.2et92p0)

[2.2 Experiment setup, evaluation metrics (r2) 7](#_heading=h.tyjcwt)

[2.3 SHAP 7](#_heading=h.3dy6vkm)

[Chapter 3: Dataset 9](#_heading=h.2s8eyo1)

[3.1 Data Dictionary 9](#_heading=h.17dp8vu)

[3.2 Data Preprocessing 10](#_heading=h.3rdcrjn)

[3.3 Descriptive Statistics 12](#_heading=h.26in1rg)

[Chapter 4: Data Analysis and Result 14](#_heading=h.lnxbz9)

[4.1 EDA 14](#_heading=h.35nkun2)

[4.2 Dropping Unnecessary Variables 18](#_heading=h.1ksv4uv)

[4.3 Feature Engineering 19](#_heading=h.44sinio)

[4.4 Explain model coefficient (linear regression), random forest dependence plot 21](#_heading=h.2jxsxqh)

[4.5 Ranking of both models 26](#_heading=h.z337ya)

[Chapter 5: Conclusion 27](#_heading=h.1y810tw)

[5.1 Finding 27](#_heading=h.4i7ojhp)

[5.2 Future Planning 27](#_heading=h.2xcytpi)

[Reference 28](#_heading=h.1ci93xb)

[DECLARATION 29](#_heading=h.3whwml4)

[**Group Member’s Declaration** 30](#_heading=h.2bn6wsx)

[**Group Member’s Declaration** 31](#_heading=h.qsh70q)

[**Group Member’s Declaration** 32](#_heading=h.3as4poj)

[**Group Member’s Declaration** 33](#_heading=h.1pxezwc)

**Chapter 1: Introduction**

* 1. **Background**

Real estate is the business of buying, selling, leasing, or managing land and buildings. It is one of the oldest and most important sectors of the economy, as it provides shelter, income, and wealth for individuals and societies. The background of real estate can be traced back to ancient times, when people started to claim ownership and rights over land and resources. Since then, real estate has evolved through different stages of development, such as feudalism, mercantilism, capitalism, and globalization.

The real estate industry is influenced by various factors, such as supply and demand, location, demographics, income, preferences, regulations, technology, and environment. These factors affect the price, value, and performance of different types of properties, such as residential, commercial, industrial, agricultural, and recreational. The real estate market is also affected by the macroeconomic conditions of a country or region, such as GDP growth, inflation, interest rates, exchange rates, and fiscal policies.

The real estate industry is a major source of employment, investment, and revenue for many countries. It also contributes to the social and environmental well-being of communities. The real estate industry faces many challenges and opportunities in the 21st century, such as urbanization, climate change, digitalization, innovation, and sustainability. To succeed in this dynamic and competitive field, real estate professionals need to have a solid background knowledge of the industry’s history, theory, practice, and trends.

However, real estate is not without its risks and uncertainties. Buyers and sellers often face various concerns when engaging in real estate transactions. Some of the concerns are:

* Market volatility: The real estate market can fluctuate due to various factors, such as the global pandemic, geopolitics, supply and demand, interest rates, and consumer preferences. This can affect the price, value, and performance of properties, as well as the availability and affordability of financing. Buyers and sellers need to be aware of the market trends and conditions and adjust their strategies accordingly.
* Financial issues: The real estate transactions involve significant amounts of money and require careful planning and budgeting. Buyers and sellers need to assess their financial situation and goals and determine the best financing options for their needs. They also need to consider the costs and benefits of different types of properties, such as maintenance, taxes, insurance, and appreciation. They also need to be prepared for unexpected expenses or changes in income or interest rates.
* Emotional issues: The real estate transactions can be stressful and emotional for both buyers and sellers. Buyers and sellers need to balance their rational and emotional factors, such as their needs, wants, expectations, and attachments. They also need to deal with the uncertainty and pressure of the market, the negotiation process, the inspection results, and the closing details. They also need to cope with the transition and adjustment of moving to a new place or leaving an old one.

To overcome these concerns, buyers and sellers can benefit from using data analysis in real estate. Data analysis is a powerful tool that can help real estate professionals make informed and profitable decisions, improve performance, and gain competitive advantage. Data analysis can provide valuable insights into the real estate market, such as the trends, patterns, and factors that influence the price, value, and performance of properties. Data analysis can also help real estate professionals optimize their operations, boost their revenues, improve their customer service, and reduce their risks. Some of the benefits of data analysis in real estate are:

* Risk mitigation: Data analysis can help us assess the commercial viability and potential returns of different properties and locations. It can also help us identify and avoid potential pitfalls, such as legal issues, environmental impacts, or market fluctuations.
* Simple and fast evaluations: Data analysis can help us evaluate properties quickly and accurately, using various metrics and models. It can also help us compare different properties and scenarios and generate reports and recommendations.
* Improved marketing strategies: Data analysis can help us design and implement effective marketing campaigns, using various channels and platforms. It can also help us measure and optimize the performance and impact of our marketing efforts.
* Market trend forecasting: Data analysis can help us anticipate and respond to the changing market conditions and customer demands. It can also help us identify new opportunities and niches and innovate new products and services.

Regression model is a good tool to explain the relationships between a dependent variable (Y) and a set of independent variables. Zhang (2021) analyse the factors that cause the increasing of median housing prices in Boston using a multiple linear regression approach. Recent trends show that using Machine Learning approach can be another alternative to analyse and predict the hosing prices. Their result indicates that both traditional hedonic and Random Forest models have their strength and weakness. Park and Bae (2015) compare several Machine Learning model to assess their accuracy in predicting the housing prices in Fairfax, Virginia. Rico-Juan and De La Paz (2021) conduct a regression analysis in Alicante, Spain to compare traditional hedonic model (linear) and Random Forest model in terms of accuracy in predicting the housing prices and how transparency the models are in convey useful information toward the potential stakeholders. Besides that, using SHAP analysis, they show Machine Learning model can provide useful insight toward buyer and seller through visualization of several variables.

The successful research carried tell indeed proof that regression analysis can provide insight for buyer and sellers on debating the housing prices. This motivates us to carry a factor analysis of housing price using both traditional linear and Machine Learning approaches.

* 1. **Big Question & Sub Question**

As mentioned in section 1.1, there are so many benefits from performing factor analysis on housing area. In this project, we selected Melbourne as study area to carry regression models to understand how real estate related variables affect the housing prices on there. Hence, we propose our big question as: **Factor Analysis of housing price in Melbourne using regression models**.

The reasons why we are choosing this as big question are:

* To identify the underlying factors that drive housing prices in Melbourne. The dataset contains a variety of data points about each property, such as its location, size, and amenities. By using factor analysis, we can identify the underlying factors that are most important in determining housing prices. This information can be used by buyers, sellers, and investors to make more informed decisions.
* To develop a model to predict housing prices in Melbourne. Once we have identified the underlying factors that drive housing prices, we can use this information to develop a model to predict housing prices in Melbourne. This model can be used by buyers, sellers, and investors to get a better idea of what a property is worth.
* To identify areas where the government can intervene to make housing more affordable. By understanding the underlying factors that drive housing prices, the government can identify areas where it can intervene to make housing more affordable. For example, by knowing the influential factors which cause unreasonable increasing housing price, government can later make beneficial decision on reducing burden of youth generation on becoming the first-time home buyer.

Below are some sub questions derived from big question:

1. How does building areas affects housing prices?
2. How does the number of bathrooms affect housing prices?
3. How does the building age affect housing prices?
4. How distance affects housing prices?
5. Is there any nonlinear relationship between housing prices and independent variables selected?
6. What are the top 5 most influencing factor contributing to housing prices (linear and random forest)

To further enhanced the assumptions of non-linear relationships in 5th sub questions, we compared our SHAP analysis result with a similar study, carried by Teoh et al. (2022) which uses regression approach and SHAP analysis in the same dataset.

To answer the big question and sub questions listed above, we state our objectives as below:

1. **To identify the relationship between housing prices and a set of independent variables using Multiple Linear Regression (MLR) and Random Forest Regression (RFR)**. Every variable influences the housing price in their own way, it can have a positive, negative relationship, or even a mixture of them. Using MLR’s statistics result, we can identify the relationships by looking at the signs of the coefficients; while using Random Forest, although it exhibits ‘black box’ properties, by using SHAP model that built on top of it, we can easily visualize the relationships. This objective is meant to solve the first 5 sub questions. The 5th sub question can be answered by identifying the variable which does not pass statistical test in MLR and then visualize the relationship using the SHAP analysis on built RFR.
2. **To rank the independent variables that contribute to housing prices in Melbourne for both MLR and RFR**. For MLR, this can be known by computing the standardized beta coefficient. For RFR, by computing the feature importance of each variable, then we can rank them. This objective is meant to solve 6th sub question.

The goal of this project is to use regression models to find out the influential hour housing prices prediction, which can support buyers and sellers in decision making on acquiring an optimal price house.

Notice that we do not compare these two models, both models have their own advantages and disadvantages. The main task is to answer the big question. Following chapters are organized in this manner: Chapter 2 introduces the methodology used in this project, Chapter 3 introduces the dataset used in this project and some descriptive statistics, Chapter 4 discusses the result and discussion from our analysis and Chapter 5 states the conclusion of findings and future planning.

**Chapter 2: Methodology**

**2.1 Modelling**

Two models, MLR and RFR are selected in this phase. Both models can be used to investigate the relationship between the housing prices and independent variables. In this project, besides using them to answer the big question, their performance is compared too.

1. MLR

MLR is an application of OLS, which is typically used in the econometric field. Following are the linear models built:

where denoted the dependent variable (per capita expenditure), is the natural logarithm form, denoted the independent variables, denoted the coefficients, denoted the intercept and denoted the error term. The method used to estimate the is an ordinary least squares (OLS) method.

1. RFR

It is one of the common supervised Machine Learning algorithms that relies on ensemble learning to perform regression tasks. Multiple decision tree models predict the outcomes independently and then average them. For each decision tree model in a Random Forest model, a subset of the sample is selected independently to train it. Generally, RFR eliminates the overfitting problem due to it averaging properties.

RFR can rank the feature importance by finding out their impurity decrease. The feature with the highest impurity decrease is the most significant feature. Using the formula from Louppe et al., (2013), the equation of mean decrease impurity measure as follow:

where represents a variable, is the proportion of of samples reaching and is the variable that used when making a split, and is the total weighted impurity decrease for all nodes when considering variable.

**2.2 Experiment setup, evaluation metrics (r2)**

The dataset is initially divided into a training set and a testing set in a 4:1 ratio. This means that 80% of the data is used to train the model, with the remaining 20% used to test it. The training set will only be used to train the model and the testing set will be used to evaluate the model. The performance of trained model will be evaluated with coefficient of determination or more commonly, r-squared,. It measures the total explainable variance of the dependent variable from independent variables. The formula of is as follow:

where is the sum of the residuals squared, while is the sum of distance the samples are away from all the mean all squared.

**2.3 SHAP**

SHapley Additive exPlanations (SHAP) is an approach to explain the model’s predictions by interpreting the features’ contribution based on Shapley value. In 2017, Lundberg and Lee published the method which has helped a lot of researchers on discovering the black box properties behind the machine learning models (Lundberg & Lee, 2017).

It calculates Shapley values from each independent variable by considering all possible permutations of rows and columns. Shapley values are used to allocate a value to each player (feature) based on their contribution to a goal (model prediction). The idea is to measure how much each feature contributes on average across all possible feature combinations. Later, each cell then has its own Shapley value, and the figure can then be plotted to visualize the relationship between the dependent variable (natural logarithm of housing price) and the independent variables.

**Chapter 3: Dataset**

**3.1 Data Dictionary**

Table 3.1 introduces the variables available in the original dataset.

**Table 3.1: Data Dictionary**

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Suburb | Name of the suburb where the property is located |
| Address | Address of the property |
| Rooms | Number of rooms in the property |
| Type | Type of property (e.g., house, unit, townhouse) |
| Price | Price of the property |
| Method | Method used to sell the property   * S - property sold; * SP - property sold prior; * PI - property passed in; * PN - sold prior not disclosed; * SN - sold not disclosed; * NB - no bid; * VB - vendor bid; * W - withdrawn prior to auction; * SA - sold after auction; * SS - sold after auction price not disclosed; * N/A - price or highest bid not available; |
| SellerG | Name of the real estate agency or seller |
| Date | Date when the property was sold |
| Distance | Distance from the property to Melbourne central business district (CBD) in km |
| Postcode | Postal code of the area where the property is located |
| Bedroom2 | Number of bedrooms in the property (alternative) |
| Bathroom | Number of bathrooms in the property |
| Car | Number of car spaces or parking spaces |
| Landsize | Size of the land on which the property is built (sqm) |
| BuildingArea | Size of the building or house (sqm) |
| YearBuilt | Year in which the property was built |
| CouncilArea | Governing council for the area |
| Lattitude | Latitude coordinates of the property's location |
| Longitude | Longitude coordinates of the property's location |
| Regionname | Name of the region |
| Propertycount | Count of properties in the same region |

**3.2 Data Preprocessing**

**Check Duplicated**

Firstly, we check for the any duplicated value in the dataset. Figure 3.1 shows the screenshot of how we check the duplicated value using Excel.

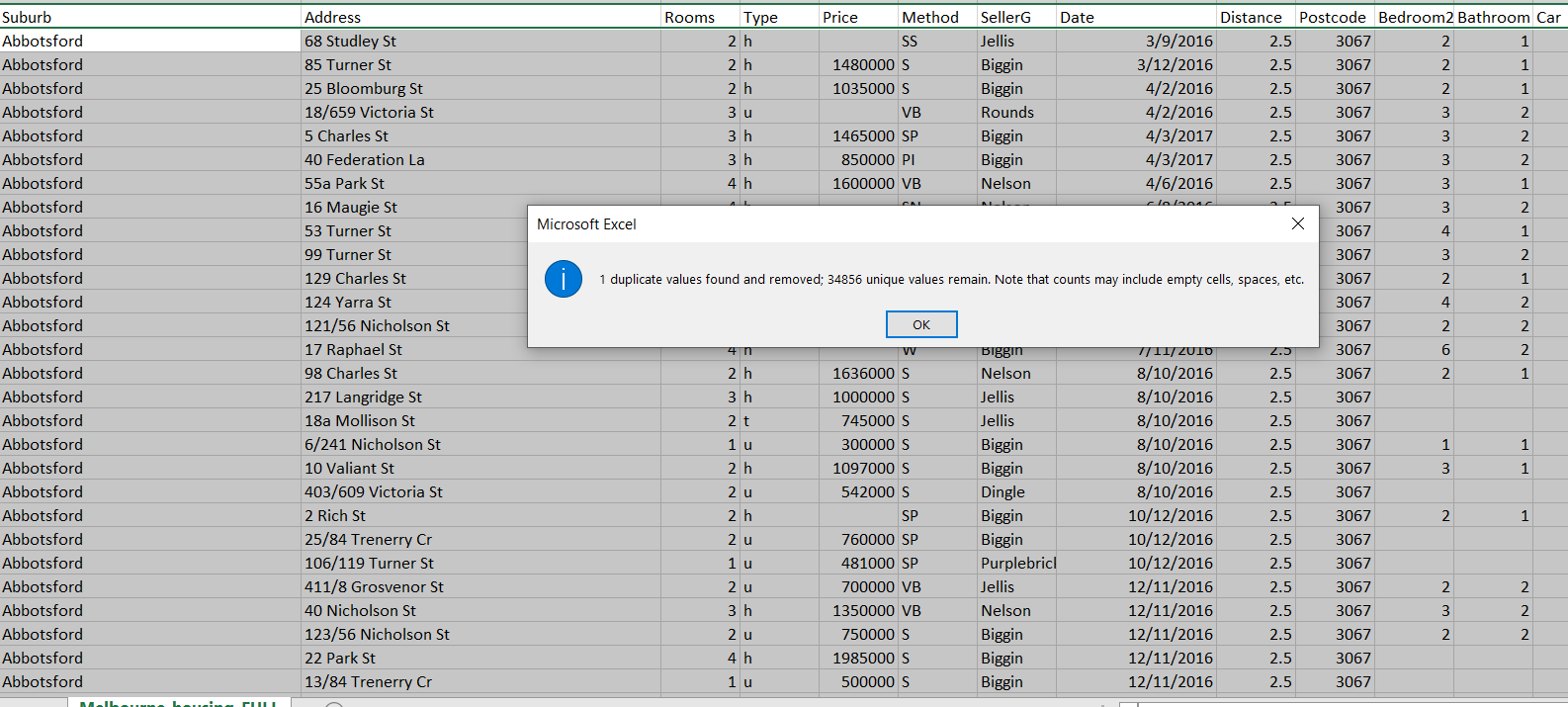
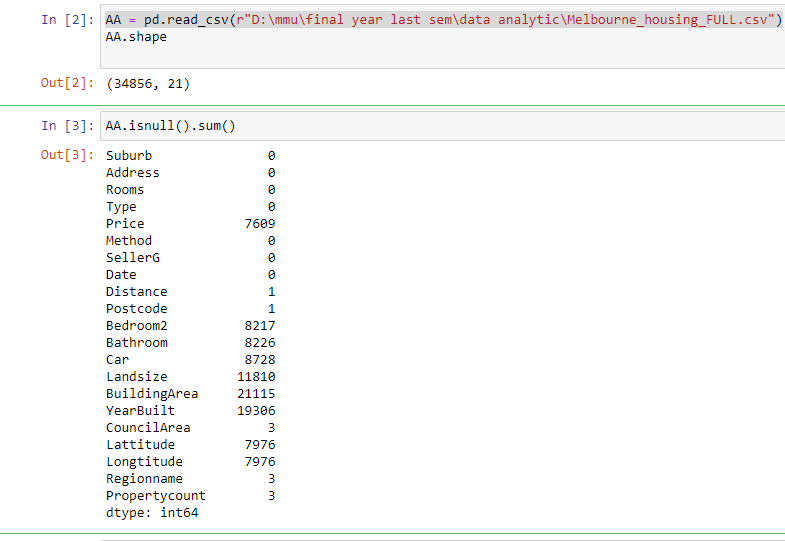


Figure 3.1: Checking the duplicated value using Excel.

**Data Filling**

After removing the duplicated data in the dataset, as there are some missing values in the datasets, we need to perform the data filling before we process the analysis. Firstly, after we import the raw datasets, we perform an initial inspection to identify missing values. The isnull().sum()function in Python was used to count the number of missing values in each column.



After checking each of the missing values in each column, we found the following 3 rows (figure 3.2) have a lot of missing values, so we decided to drop these rows.

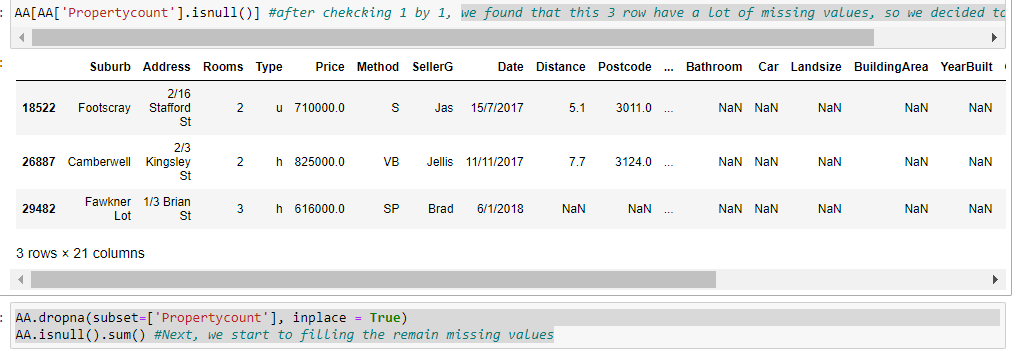


Figure 3.2: Missing value observed.

After, we need to start to perform the data filling for the rest of the columns. A correlation heatmap as shown in figure 3.3 was generated to understand the relationships between variables in the dataset. Based on the heatmap, below are observed:

* price strongly correlated with rooms, bedroom2, bathroom
* bedroom2 strongly correlated with rooms, bathroom, price
* bathroom strongly correlated with bedroom2, rooms, price
* car semi strongly correlated with rooms, bedroom2
* landsize semi strongly correlated with building area
* buildingarea semi strongly correlated with landsize
* yearbuilt semi strongly correlated with distance
* lattitude semi strongly correlated with postcode, longtitude
* longtitude semi strongly correlated with postcode, latitude

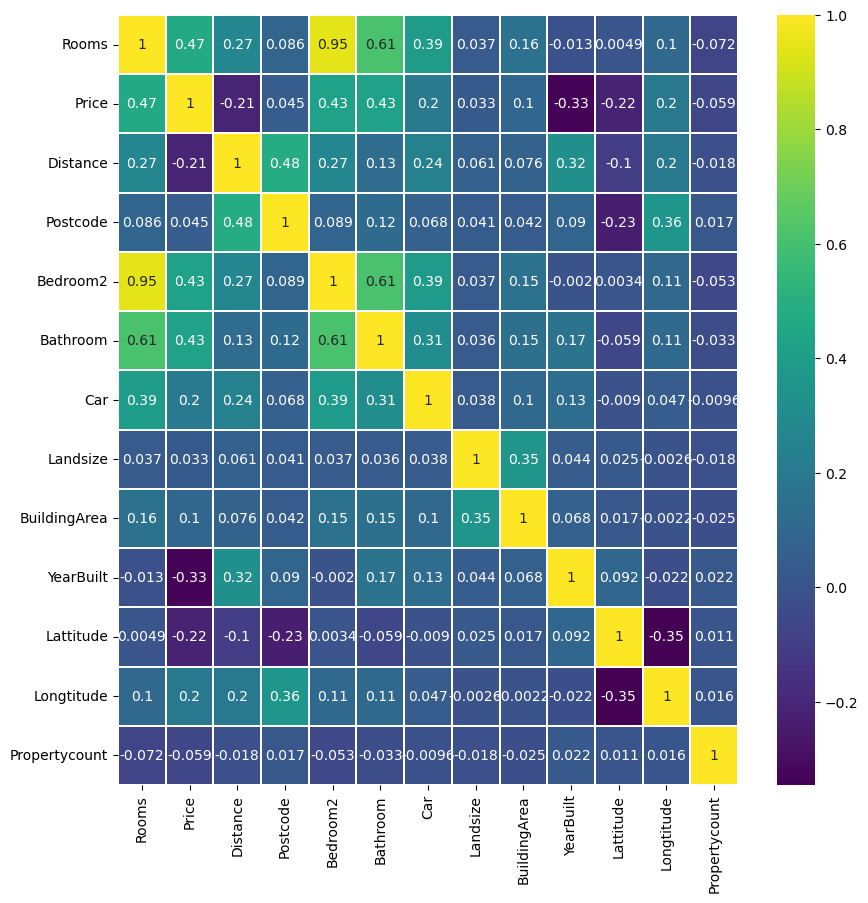


Figure 3.3: 1st Correlation matrix analysis for imputation purpose

After knowing the correlation between the variables, we input the missing values of each columns batch by batch based on the below table 3.1:

Table 3.1 Reference table for data filling method

|  |  |
| --- | --- |
| **Input Column** | **Imputation Method Used** |
| Price, Rooms, Bedroom2, Bathroom | IterativeImputer (from scikit-learn) |
| Car, Rooms, Bedroom2 | KNNImputer (from scikit-learn) |
| BuildingArea, Landsize | KNNImputer (from scikit-learn) |
| Lattitude, Longtitude, Postcode | KNNImputer (from scikit-learn) |

After completing the imputation steps, we performed a final check to ensure that there were or no remaining missing values in the dataset and export the processed datasets for further analysis. In this conclude, the missing data in the Melbourne Housing Dataset had been effectively addressed using iterative imputation and K-nearest neighbours imputation techniques. This ensures that the dataset is complete and more suitable for further analysis and modelling.

**3.3 Descriptive Statistics**

Table 3.4 introduces the descriptive statistics for the processed dataset for numerical variables while table 3.5 show the descriptive statistics for categorical variables.

**Table 3.4 Descriptive statistics for Categorical Columns**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Suburb** | **Address** | **Type** | **Method** | **Seller** | **CouncilArea** | **Regionname** |
| **Count** | 34852 | 34852 | 34852 | 34852 | 34852 | 34852 | 34852 |
| **Unique** | 350 | 34005 | 3 | 9 | 388 | 33 | 8 |
| **Top** | Reservoir | Charles ST | h | S | Jellis | Boroondara City Council | Southern Metropolitan |
| **Freq** | 844 | 6 | 23977 | 19743 | 3357 | 3675 | 11836 |

**Table 3.5: Descriptive statistics for Numerical Columns**

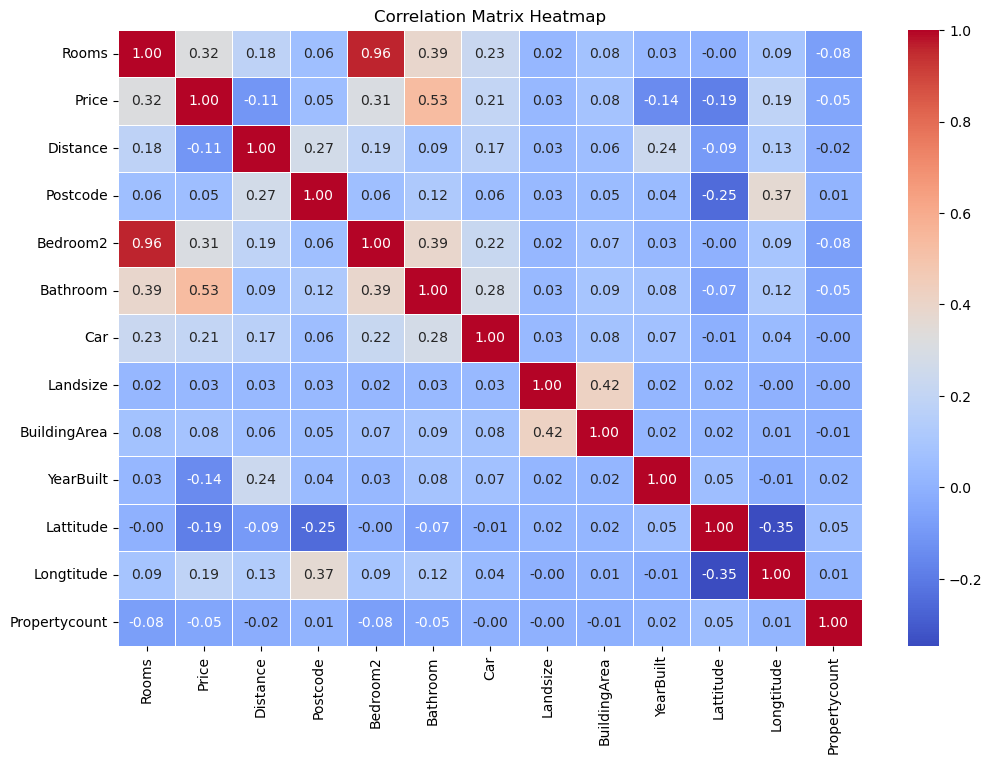
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Rooms** | **Price** | **Distance** | **Postcode** | **Bedroom2** | **Bathroom** | **Car** |
| **count** | 34582 | 34582 | 34582 | 34582 | 34582 | 34582 | 34582 |
| **Mean** | 3.03 | 1066132.13 | 11.19 | 3116.06 | 3.02 | 1.59 | 1.76 |
| **Std** | 0.97 | 593024.23 | 6.79 | 109.03 | 0.98 | 0.69 | 0.92 |
| **Min** | 1.00 | 85000 | 0.00 | 3000 | 0.00 | 0.00 | 0.00 |
| **25%** | 2.00 | 685000 | 6.40 | 3051 | 2.00 | 1.00 | 1.00 |
| **50%** | 3.00 | 906250 | 10.30 | 3103 | 3.00 | 1.00 | 2.00 |
| **75%** | 4.00 | 1340000 | 14.00 | 3156 | 4.00 | 2.00 | 2.00 |
| **max** | 16.00 | 11200000 | 48.10 | 3978 | 30.00 | 12.00 | 26.00 |

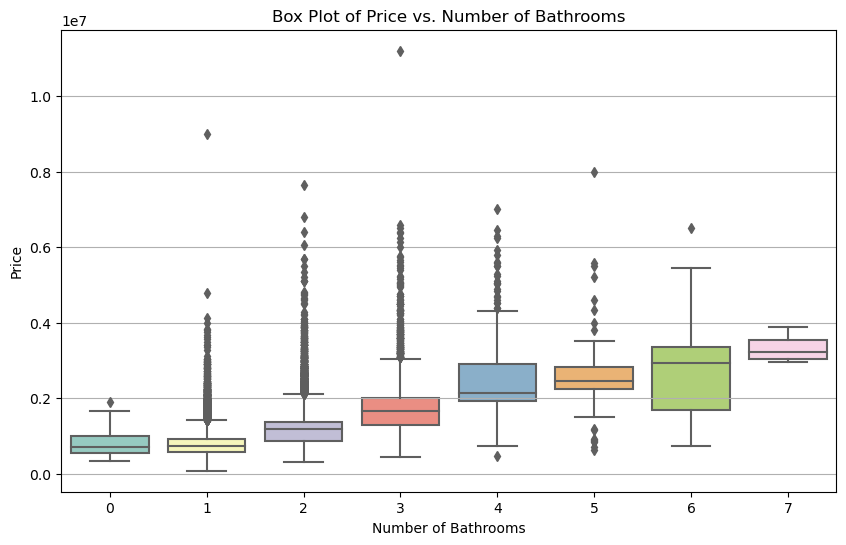
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Landsize** | **BuldingArea** | **YearBuilt** | **Lattitude** | **Longtitude** | **Propertycount** |
| **count** | 34582 | 34582 | 34582 | 34582 | 34582 | 34582 |
| **Mean** | 590.98 | 162.70 | 1964.22 | -37.8112 | 145.0031 | 7572.99 |
| **Std** | 2765.39 | 304.18 | 37.53 | 0.09 | 0.12 | 4428.19 |
| **Min** | 0.00 | 0.00 | 1196 | -38.1904 | 144.4238 | 83 |
| **25%** | 343.00 | 123.20 | 1940 | -37.8640 | 144.9363 | 4385 |
| **50%** | 593.60 | 160.30 | 1970 | -37.8082 | 145.0097 | 6763 |
| **75%** | 605.00 | 169.40 | 1999 | -37.7538 | 145.0725 | 10412 |
| **max** | 433014.00 | 44515.00 | 2019 | -37.3902 | 145.5264 | 21650 |

**Chapter 4: Data Analysis and Result**

**4.1 EDA**

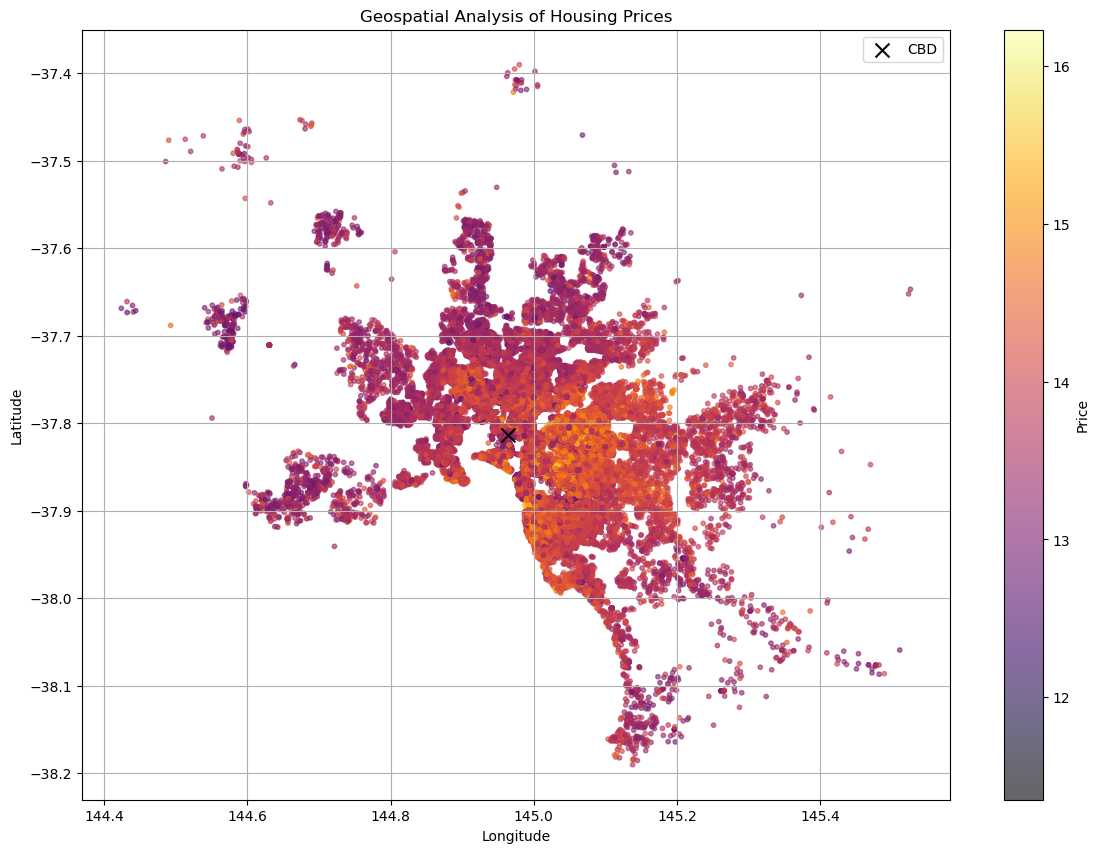
A correlation heatmap is a powerful visual tool that allows us to explore the relationships between different variables within a dataset. It provides insights into how these variables are interconnected, helping us understand patterns, dependencies, and potential influences. In our cases, we perform feature importance analysis to identify which features (columns) in dataset have the most significant impact on housing prices, as shown in figure 4.1. We have created a correlation matrix and visualize it as a heatmap to identify which features have the strongest linear relationships with price. In our observation, we see that *price* strongly correlated with *rooms, bedroom2, bathroom*. Also, we have found several variables like *Distance, YearBuilt* which having the negative correlation with the *Price* variable.

**Figure 4.1 Correlation Matrix Heatmap of numeric variables**



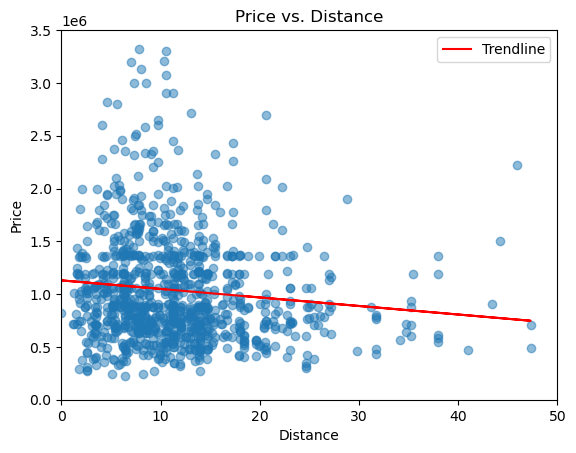
**Figure 4.2 Box Plot of Price vs Numbers of Bathrooms**

As the *Bathroom* variables have the strongest correlation with the *Price*, we have decided to use the box plot to have clearer visualization of the relationship between two variables. In the box plot analysis of property prices across different bathroom categories showed in figure 4.2, it's evident that the boxes skew toward the upper end of the whiskers, indicating that most properties within each bathroom category tend to have higher prices. This observation suggests a positive correlation between the number of bathrooms and property prices. As the number of bathrooms increases, median property prices generally rise, highlighting the influence of bathroom count on property valuations. However, an important aspect to note is the presence of numerous outliers in the data. Outliers are individual data points that fall beyond the whiskers of the boxes and have prices significantly higher or lower than the typical prices within their respective bathroom categories. These outliers, both above and below the whiskers, introduce substantial variability and complexity to the dataset.

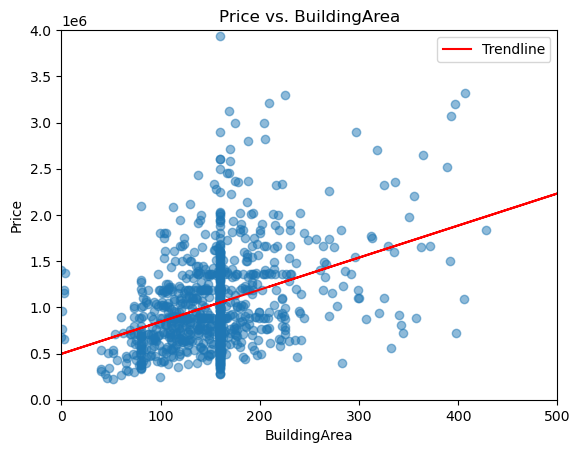


**Figure 4.3 Geospatial Analysis of Housing Prices**

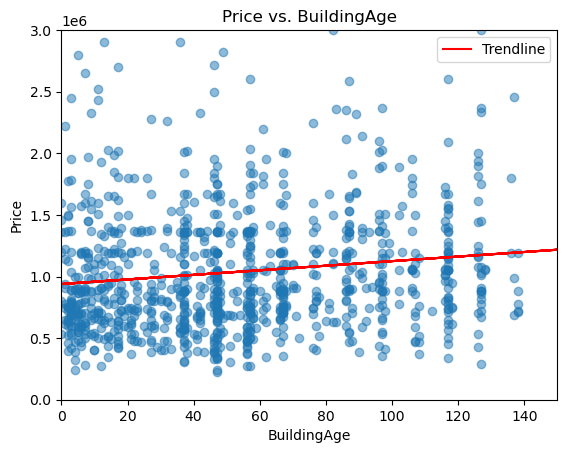
The geospatial scatter plot analyses the relationship between geographic location and housing prices using latitude and longitude coordinates. Each data point on the scatter plot represents a property listing, with the x-axis representing longitude, the y-axis representing latitude. The colour of each point on the scatter plot represents the price of the corresponding property. The colour is determined by the logarithm of the property's price (Log\_Price), with darker colors indicating lower prices and lighter colours indicating higher prices. The reason using Log\_Price is because prices in real estate datasets often exhibit a wide range, with some properties having significantly higher prices than others. We take the logarithm of prices helps to normalize the data. This normalization reduces the impact of extreme values and allows for a more balanced representation in the plot. It can make it easier to observe patterns and relationships among properties with a wider range of prices. Also, there is a black 'X' marker on the plot that represents the Central Business District (CBD) location. By observing the scatter plot, we can have general understanding about the relationship of distance of property to the CDB and the property price. The most notable pattern observed is that properties located closer to the Central Business District (CBD), represented by the black 'X' marker, tend to have substantially higher prices. Also, the price of property decreases as move farther away from the central urban area. This pattern aligns with a common trend seen in real estate markets worldwide. In many cities, properties situated in or near central urban hubs are more expensive due to their close proximity to business centres, essential amenities, and various services. The convenience and accessibility offered by these locations often command premium prices.



**Figure 4.4 Scatter plot of Price vs Distance**



**Figure 4.5 Scatter Plot of Price vs Building Area**



**Figure 4.6 Scatter Plot of Price vs BuildingAge**

The above three scatter plot (figures 4.4, 4.5, 4,6 show the relationship between *Log\_ Price* with three different factors which is *Distance* (distance from the property to Melbourne central business district (CBD) in km), *BuidingArea*, and *BuildingAge* in the dataset. In this analysis, each of the scatter plots are using a sample of 1000 data points.

The first scatter plot (figure 4.4) explores the relationship between property prices and their distance from CBD. The trendline within this plot exhibits a discernible negative slope, indicating a negative correlation as we have also seen that in the above correlation matrix heatmap. As the distance increases, prices tend to decrease. This negative correlation is common happen in real estate as Properties located in close proximity to essential amenities, job centres, and transportation hubs often command higher prices because there is more convenience and accessibility. Instead, as one moves further away from these central areas, property prices tend to decrease. This information is valuable to sellers and buyers for discerning how location will impact property values.

The second scatter plot (figure 4.5) represents the relationship between price and the size of the building area. The trendline within this plot show a significant positive slope, signifying a strong positive correlation. As the building area increases, property prices tend to rise. Larger properties commonly have more living space and potential for facility, which make them more attractive to prospective buyers. Thus, they often carry higher price tags. This positive relationship shows the significance of square footage in determining property values. Buyers more willing to pay a premium for increased space and comfort.

In the third scatter plot (figure 4.6), we analyse the relationship between property prices and the age of buildings. The trendline, while relatively shallow, presents a slight positive slope. This suggests that as the age of the property (number of years since construction) increases, property prices also exhibit a little increase. This positive correlation between price and building age can be due to various factors. Older properties may have unique historical or architectural features that add to their attractiveness and value. In addition, they might be situated in established neighbourhoods with character and history, attractive to specific buyers. However, it is important to note that the influence of building age on price is relatively minor when compared to factors like location and building area.

**4.2 Dropping Unnecessary Variables**

According to figure 4.1, variable *bedroom2* is highly correlated with *rooms*. To prevent multicollinearity problems, we selected to drop bedroom2. Besides that, we also dropped *Latitude* and *longitude* because they have no true linear relationships with *Log\_Price*. The Melbourne geographical map of housing price in figure 4.3 clearly shows that housing price reaches the peak in the middle of the latitude and longitude. Although as we mentioned in Chapter 1, RFR can handle nonlinearity issues, but analysis on latitude and longitude is hard to relate in the perspective of buyer, hence we decided to drop them.

**4.3 Feature Engineering**

We preferred and selected the natural logarithm of price (*Log\_Price*) instead of the original *price* variable because it improves the model performance after evaluation using r2 as metric. Generally, economists favour using the natural logarithm form of monetary value as they tend to improve the model performance (Lütkepohl & Xu, 2010).

Besides that, we created *MonthSold* and year variables *YearSold* to understand how they contribute to housing prices in Melbourne too. We created the *BuildingAge* variable also to become another independent variable because building age generally has a positive relationship with housing prices in every country.

Meanwhile we also created dummy variables for two categorical variables, *Regionname* and *Type*. Using this method, every unique value inside these two variables becomes a new Boolean variable, this solves not only the non-numbering issue in regression modelling, but also ease our analysis after modelling.

Table 4.1 shows the variable we used in section 4.4:

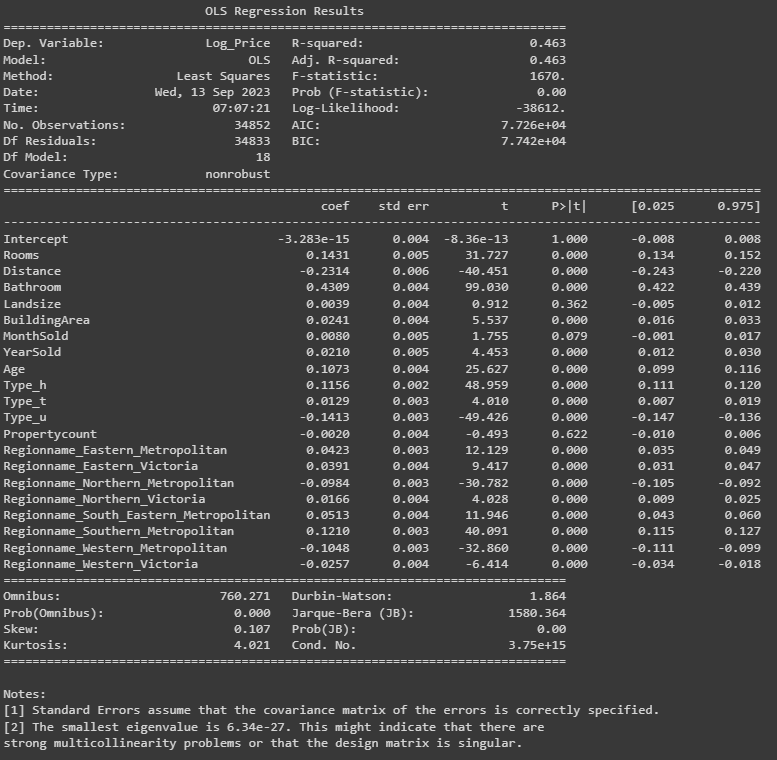
|  |
| --- |
| **Variables** |
| Rooms |
| Distance |
| Bathroom |
| Car |
| Landsize |
| BuildingArea |
| YearBuilt |
| Propertycount |
| MonthSold |
| YearSold |
| BuildingAge |
| LogPrice |
| type\_h |
| type\_t |
| type\_u |
| region\_Eastern Metropolitan |
| region\_Eastern Victoria |
| region\_Northern Metropolitan |
| region\_Northern Victoria |
| region\_South-Eastern Metropolitan |
| region\_Southern Metropolitan |
| region\_Western Metropolitan |
| region\_Western Victoria |

Table 4.1: Dataset used in modelling phase

**4.4 Relationships understanding to answer the sub question(MLR and RFR)**

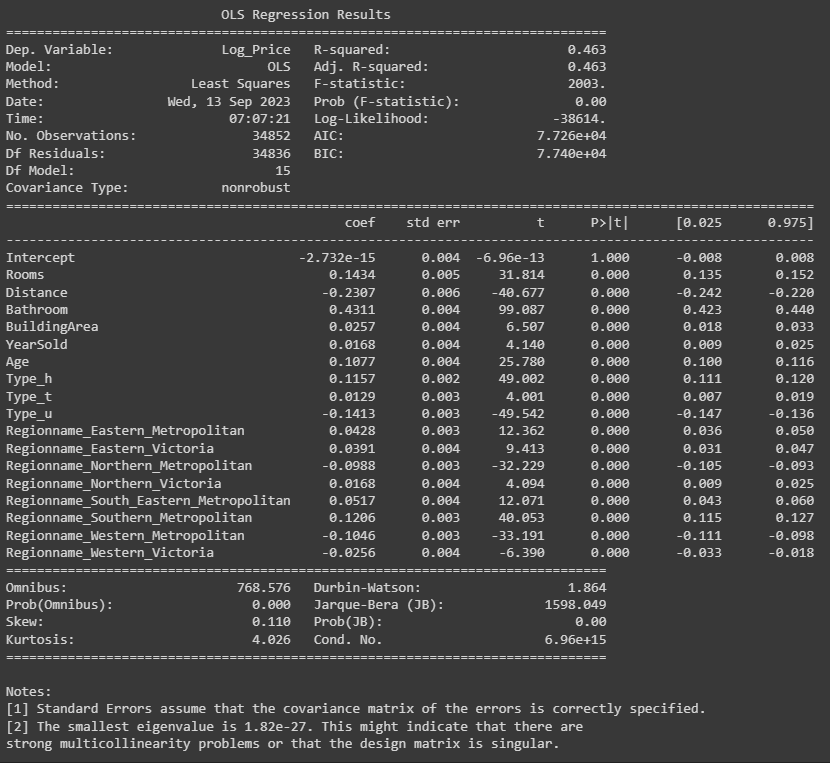
This section discusses the regression results from two models. This section answers the 1-5th sub questions.

**Multiple Linear Regression**



**Figure 4.7: QLS Regression Result**

Figure 4.7 shows the statistics result from the first MLP experiment. From figure 4.7 above, we notice that p-values for *landsize, propertycount and monthSold* variables are not significant, therefore, we decide to drop it and re-run the analysis again. From here, we can then question: is there exists non-linear relationships between these three independent variables and *Log\_Price*? This will be answered later.



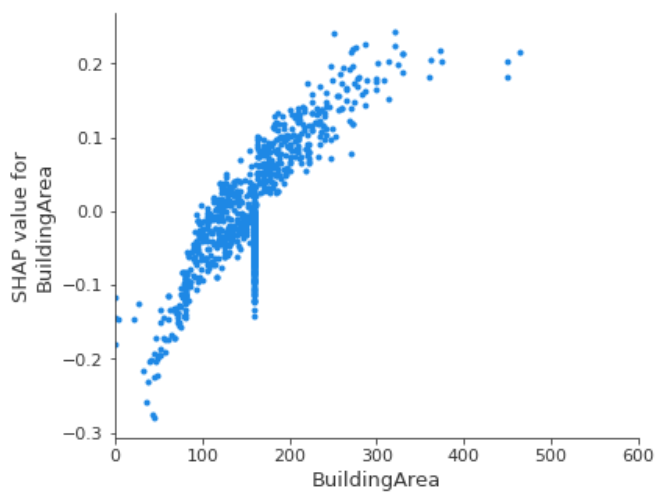
**Figure 4.8: QLS Regression Result**

From figure 4.8 above, it provides lots of valuable information for us to make informed decisions. R square of 0.463 indicates that 46.30% of variability in dependent variable can be explained by the independent variable. This shows a moderate relationship between independent variable and dependent variable. Since all the p-values of independent variables are less than 0.05, we can safely assume that all variables are significant.

We can interpret the approximate change in *Log\_Price* from the coefficient of each independent variable. Noticed that the coefficient is the standardized beta coefficients, it means they are measured in the standard deviation unit, this is because we would like to use it to rank the variable importance in section 4.5. However, we still can still interpret the relationship. Firstly, as the *BuildingArea* increases, the *Log\_Prices* also increase. Every increase in one standard deviation in *BuildingArea* increase the *Log\_Price* by 0.0257. Second, as *Bathroom* increases, the *Log\_Prices* also increase. This implies a positive relationship. Every increase in one standard deviation in *Bathroom* increase the *Log\_Price* by 0.4311. Third, as *BuildingAge* (Refer variable *Age* from figure 4.8) increases, the *Log\_Prices* also increase. This implies a positive relationship. Every increase in one standard deviation in *BuildingAge* increase the *Log\_Price* by 0.1077. Lastly, as *Distance* increases, the *Log\_Prices* decrease. This implies a negative relationship. Every increase in one standard deviation in BuildingAge decreases the Log\_Price by 0.2307.

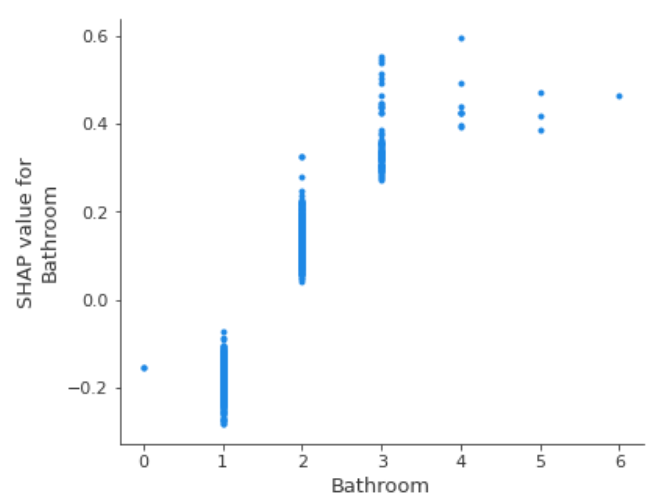
**Random Forest Regression**

This section visualizes the relationships between 4 selected variables from sub questions and *Log\_Price*.



**Figure 4.9: Random Forest SHAP dependence plot for the feature BuildingArea**

Figure 4.9 illustrates the relationship found between *BuildingArea* and *Log\_Price*. Based on the Shapley value computed by SHAP, from here we observed a generally positive relationship between *BuildingArea* and *LogPrice*. We can generally say that there exists a positive relationship.



**Figure 4.10: Random Forest SHAP dependence plot for the feature Bathroom**

From figure 4.9, we observed a generally positive relationship between *Bathroom* and *Log\_Price*. However, when the number of bathrooms is equal or larger than 3, the increasing effect then become constant. Hence, we can conclude that as the number of bathrooms increase until 3, the housing price in Melbourne then increase. After number of bathroom larger than 3, it shows neither increasing or decreasing effect, or at least very small increasing/decreasing effect. This relationship is quite similar with the SHAP dependence plot found by (Teoh et al., 2022) too.

A graph with blue dots

Description automatically generated

**Figure 4.11: Random Forest SHAP dependence plot for the feature BuildingAge**

Based on the Shapley value computed by SHAP, from here we observed a generally positive relationship between *BuildingAge* and *Log\_Price*. At start, the *BuildingAge* seems to have no relationship with housing price. Until age 40, we can observe increasing trend until *BuildingAge* of 80. We can generally say that there exists a positive relationship, but the increasing effect become obvious starting from *BuildingAge* of 40 till 80. This result shows part of identical with (Teoh et al., 2022), which states that buyers are more willing to purchase the properties which passes *70BuildingAge*.

A graph with blue dots

Description automatically generated

**Figure 4.12: Random Forest SHAP dependence plot for the feature Distance**

Based on the Shapley value computed by SHAP, from here we observed a generally negative relationship between *Distance* and *Log\_Price*. Strong negative relationship is found between *Distance* of 0 till 20. Starting from 20, Distance seems to have no relationship over housing price. We can generally say that there exists a negative relationship. Again, this part shows similar result with (Teoh et al., 2022).

The 1st – 4th sub questions have been answred using two regression models. We continue to answer 5th sub question.

A graph with blue dots

Description automatically generated

**Figure 4.13: Random Forest dependence plot for the feature Landsize**

Based on the Shapley value computed by SHAP, from here we observed a generally positive relationship between *Landsize* and *Log\_Price*. But the strange vertical dots observed results the distribution looks to be more nonlinear. Recall back the p-value significance test in MLR, and the relationship observed here, we can then conclude this *Landsize* variable has a nonlinear relationship with *Log\_Price*.

A graph with numbers and blue dots

Description automatically generated

**Figure 4.14: Random Forest dependence plot for the feature Propertycount**

Based on the Shapley value computed by SHAP, we observed a relationship closely to quadratic. There is an increasing trend from 0 to 10000, so we can say that as *Property Count* increases (until 10000), the *Log\_Price* increases too. This is no true starting from 10000, as we can see there is a decreasing trend that greatly affect the *Log\_Price*. Hence, this tells us that *Propertycount* has a nonlinear relationship but a quadratic trend toward LogPrice and the reason why *Propertycount* did not pass the significance test.

As a conclusion, we can answer the 5th sub questions, *Landsize* and *Propertycount* have nonlinear relationships with *Log\_Prices*.

A graph showing different colored lines

Description automatically generated with medium confidence

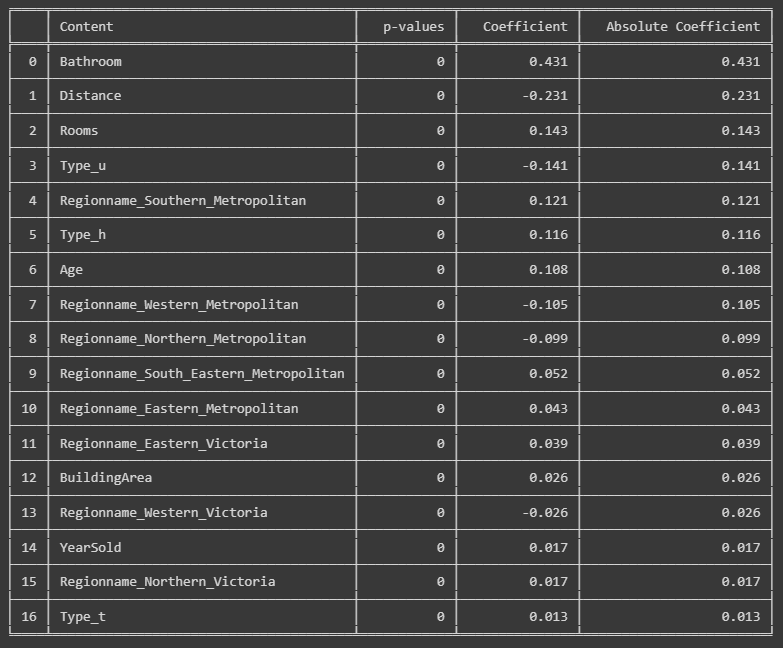
**Figure 4.15: Summary plot for the SHAP values of built random forest model**

Figure 4.15 shows the SHAP summary plot. It provides an overall picture of Shapley Value inside Random Forest. The relationship can be determined by observing the direction of colour change. If the SHAP value of a variable increases as the corresponding feature value increase, then this independent variable is said to have a positive impact on dependent variable or vice versa.

As an example, in *Bathroom*, blue colour in left side change to red colour as moving to right side, which means a positive relationship. Besides that, another example is, the positive relationship found in *Rooms*. *Distance* has a negative relationship with *Log\_Price*, as the colour changes is red to blue from left to right side. Others seems to have unclear relationships. The detail of some variables has been explained in the dependence plots.

**4.5 Ranking of both models**

This section answered the 6th sub question: What are the top 5 most influential factors affecting the housing prices in Melbourne? Both models have their own rankings.



**Figure 4.16: The ranking of variables of MLR.**

Figure 4.16 shows the ranking of variables according to the standardized beta coefficient of independent variable (absolute coefficient). From this table, we get to know that *Bathroom* variable is having the most significant effect onto the changes of dependent variable which is *Log\_Price* variable. Apart from that, *Distance* variable also having great influence onto the log price variables. It shows us that as one house is getting close to the city, the log price is higher. The 3rd important variable is Rooms, followed by *Type\_u* and lastly, the *Regionname\_Southern\_Metropolitan*.

A screenshot of a computer

Description automatically generated

**Figure 4.17:** **The ranking of variables of RFR.**

Figure 4.17 shows the ranking of variables according to the feature importance. For each independent variable, the RFR use its algorithm to keeps track of the overall Gini impurity decreases on average when that variable is used to split nodes (refer to the trees in RFR) across all trees in the forest. Firstly, identical to MLR’s result, the most important variable is *Bathroom*. The second important is *BuildingArea*. *Distance* scored the 3rd places. These two variables’ set are interchange in RFR model as compared to MLR model, but this does not change the fact that they are powerful variable in changing the housing price in Melbourne. The 4th important variable is *Landsize*, which is the variable dropped in MLR experiment. Lastly, the 5th important variable in RFR is the *Rooms*, which is also identical to MLR result.

**Chapter 5: Conclusion**

**5.1 Finding**

The project aims to analyse the factors that affect the house price in Melbourne, using two different models: LRL and RFR. The project uses a dataset that contains information about the location, type, size, price, and features of various properties sold in Melbourne from 2016 to 2018. The project performs factor analysis, coefficient analysis, and SHAP value analysis to identify the key features that influence the house price prediction. The project also compares the accuracy and performance of the two models.

The project provides valuable insights into the real estate market in Melbourne and can help home buyers and sellers of house asking price. The project also demonstrates the advantages and limitations of different models and techniques for data analysis. The MLR come with simplicity, stable, common and easier interpreted. On the other hand, RFR come with greatly enhanced in performance, nonlinearity capturable ability. We can interpret the approximate change of a single independent variable toward the housing price in MLR model, which is unachievable in RFR model. RFR able to incorporate almost every independent variable, including the independent variable found to be nonlinear.

Our finding shows that, the building area, number of bathrooms, the building age have generally positive relationship toward the housing price in Melbourne. Both MLR and RFR shows the identical result. Second, the land size of property and the property count are two variables found to have nonlinear relationships with housing prices using SHAP dependence plot analysis. Finally, number of bathrooms, number of rooms, distance toward CBD are three power variables in changing the housing prices in Melbourne from both models’ ranking. *Type\_U* (duplex or unit house) and Southern Metropolitan areaare another two influential factors ranked by MLR while land size and building area are the two important variables ranked by RFR.

**5.2 Future Planning**

The project aims to improve and expand its scope and impact in the following ways. First, it plans to extend the dataset to include more recent and diverse data on the properties sold in Melbourne, such as the features, amenities, ratings, reviews, and images of the properties. This will allow for a more comprehensive and accurate analysis of the house price trends and factors in Melbourne. Second, it intends to explore other models and techniques for data analysis, such as neural networks, clustering, classification, and natural language processing, and compare their performance and results with the current models. This will enable a deeper and richer understanding of the data and its implications for the house price prediction. Third, it hopes to create interactive and dynamic visualizations and dashboards for the data analysis results, such as maps, charts, graphs, and tables, that can allow users to filter, sort, and customize the data according to their preferences and needs. This will enhance the user experience and engagement with the data and its insights. Fourth, it aims to develop a web application or a mobile app that can provide users with real-time and personalized recommendations and predictions for the house price in Melbourne, based on their input and feedback. This will offer a practical and useful tool for users who are interested in buying or selling properties in Melbourne. Fifth, it plans to conduct a user study and evaluation to assess the usability, usefulness, and satisfaction of the project’s outputs and outcomes, and to identify the strengths, weaknesses, opportunities, and challenges of the project. This will provide valuable feedback and suggestions for further improvement and refinement of the project.

**Reference**

Lütkepohl, H., & Xu, F. (2010). The role of the log transformation in forecasting economic variables. *Empirical Economics, 42*(3), 619–638. <https://doi.org/10.1007/s00181-010-0440-1>

Lundberg, S., & Lee, S. (2017). *A unified approach to interpreting model predictions. Neural Information Processing Systems, 30*, 4768–4777. <https://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>

Park, B., & Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. *Expert Systems With Applications,* 42(6), 2928–2934. <https://doi.org/10.1016/j.eswa.2014.11.040>

Rico-Juan, J. R., & De La Paz, P. T. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. *Expert Systems With Applications, 171,* 114590. <https://doi.org/10.1016/j.eswa.2021.114590>

Teoh, E. Z., Yau, W., Ong, T. S., & Connie, T. (2022). Explainable housing price prediction with determinant analysis. *International Journal of Housing Markets and Analysis, 16(5)*, 1021–1045. https://doi.org/10.1108/ijhma-02-2022-0025

Zhang, Q. (2021). Housing price prediction based on multiple linear regression. *Scientific Programming*, 2021, 1–9. <https://doi.org/10.1155/2021/767893>

**DECLARATION**

We agree that all members deserve equal marks for this project. We confirm that we have contributed equally to produce an original report in our best effort.

Write down your ID (clearly) and sign (by all members):

A black line on a white background

Description automatically generated

Student ID: 1191100578 \_\_\_\_Signature:



Student ID: 1191101497 Signature:

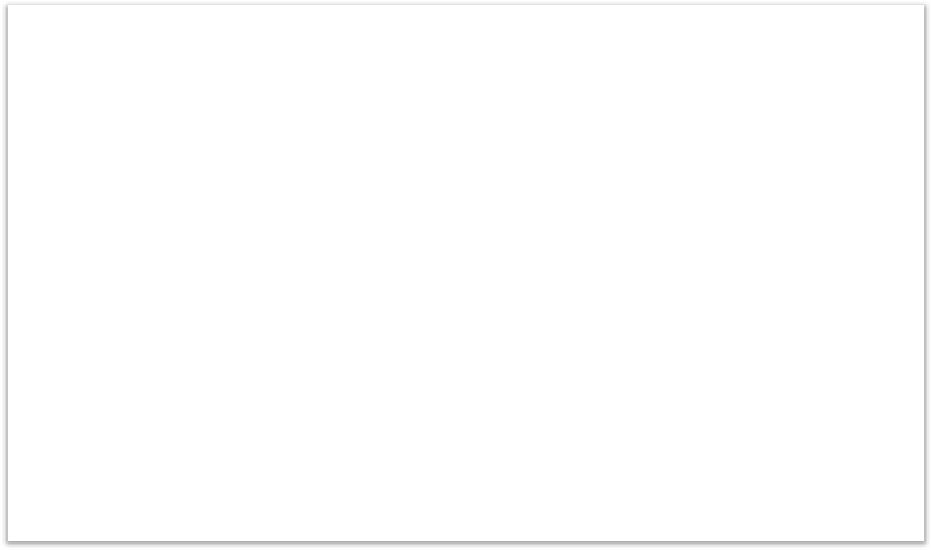


Student ID: 1191100577 Signature: 

Student ID: 1191101340 Signature:

A black line on a white background

Description automatically generated



**Declaration by Group Leader**

I hereby declare that all group members’ names are correctly included in the above section. I hold a copy of this assignment which I can produce if the original is lost or damaged. I certify that not part of this assignment has been copied from any other student’s work or from any other source except where due acknowledgement is made in the assignment/project/etc.

Group Leader’s Signature: Group Leader’s Name: Lee En Group Leader’s ID: 1191100578 Date: 21/8/2023

**Group Member’s Declaration**

(Each group member, including the group leader, must individually fill up and submit this form. This form has to be attached together with the assignment/ project submission.)

Group member’s name: Lee En

Student ID: 1191100578

For the purpose of completing this assignment, I have performed the following tasks:

Coding (Random Forest)

Chapter 4.1 and 4.2

I hereby declare that I have assessed the final submission and I take full responsibility should there be any inaccuracies, incompleteness, omissions, delays or non- submission.

A black line on a white background

Description automatically generated

Group member’s signature:

Group member’s name: Lee En

Group member’s ID: 1191100578

Date: 21/8/2023

**Group Member’s Declaration**

(Each group member, including the group leader, must individually fill up and submit this form. This form has to be attached together with the assignment/ project submission.)

Group member’s name: Foo Haw Liang

Student ID: 1191101497

For the purpose of completing this assignment, I have performed the following tasks:

Introduction

Conclusion

Report Compiling

Explain model coefficient (linear regression), random forest dependence plot

Ranking of both models

I hereby declare that I have assessed the final submission and I take full responsibility should there be any inaccuracies, incompleteness, omissions, delays or non- submission.



Group member’s signature:

Group member’s name: Foo Haw Liang

Group member’s ID: 1191101497

Date: 17/9/2023

**Group Member’s Declaration**

(Each group member, including the group leader, must individually fill up and submit this form. This form has to be attached together with the assignment/ project submission.)

Group member’s name: Siah Kah Chuan

Student ID: 1191100577

For the purpose of completing this assignment, I have performed the following tasks:

Multiple Linear Regression Model

Feature engineering

Interpret result.

Variable ranking

I hereby declare that I have assessed the final submission and I take full responsibility should there be any inaccuracies, incompleteness, omissions, delays or non- submission.

A signature on a white background

Description automatically generated

Group member’s signature:

Group member’s name: Siah Kah Chuan

Group member’s ID: 1191100577

Date: 21/8/2023

**Group Member’s Declaration**

(Each group member, including the group leader, must individually fill up and submit this form. This form has to be attached together with the assignment/ project submission.)

Group member’s name: Grayson Goh Jin Yi

Student ID: 1191101340

For the purpose of completing this assignment, I have performed the following tasks:

Data Preprocessing,

EDA coding

Report:

Chapter 3.1,

Chapter 3.2,

Chapter 3.3,

Chapter 4.1

I hereby declare that I have assessed the final submission and I take full responsibility should there be any inaccuracies, incompleteness, omissions, delays or non- submission.



Group member’s signature:

Group member’s name: Grayson Goh Jin Yi

Group member’s ID: 1191101340

Date: 21/8/2023