

# MASTER IN DATA SCIENCE

## GROUP ASSIGNMENT 2

**COURSE CODE** : WQD7001  
**COURSE TITLE** : PRINCIPLES OF DATA SCIENCE  
**LECTURER** : DR MAIZATUL  
**PROJECT TITLE** : MONTHLY RENT PRICE PREDICTION IN KUALA LUMPUR

**GROUP MEMBERS :**

1. Chow Kai Ern (S2193226) – *Data Modelling + Insights & Conclusion*
2. Nur Hidayah binti Ahmad Shafii (22120931) – *Data Modelling + Data Interpretation*
3. Hanisah Zulaikha binti Zulkifli (23072436) – *Data Modelling + Data Product*
4. Loh Bi Jia (23078886) – *Data Modelling + Reproducible Research + Project Bg&Objectives*

## Table of Contents

|  |           |
|--|-----------|
| <b>1.0 Project Background.....</b>             | <b>1</b>  |
| <b>2.0 Data Modelling.....</b>                 | <b>1</b>  |
| <b>3.0 Data Interpretation .....</b>           | <b>3</b>  |
| <b>4.0 Plan for Reproducible Research.....</b> | <b>6</b>  |
| <b>5.0 Deployment of Data Product.....</b>     | <b>7</b>  |
| <b>6.0 Insights and Conclusion.....</b>        | <b>8</b>  |
| <b>7.0 References.....</b>                     | <b>11</b> |

## 1.0 Project Background

Kuala Lumpur, the capital of Malaysia, is a center of economic activity and attracts many local and international residents. The high cost of housing sales leads people to rent (Raml et al. 2019). This makes the rental market an important part of the real estate industry. With increase of population, understanding rental price dynamics is critical for stakeholders such as landlords, tenants, investors and real estate professionals. The rental market of Kuala Lumpur is dynamic and complex. Recent volatility in the Kuala Lumpur rental market has been driven by factors such as government policies (such as the Malaysia My Second Home programme), changes in the Overnight Policy Rate (OPR) and the expansion of the MRT/LRT network. Global issues such as the COVID-19 pandemic have influenced the preferences of tenant towards properties, hence the rental market reshaped. These situations highlight that rental price prediction models are important. Due to variable and inconsistent market data, appraisers face significant challenges when estimating property prices and rental values (Abdul S. et al., 2021). Therefore, rental predictive models are important to maintaining a vibrant real estate market in cities.

The objectives of our project are:

i) To explore and analyze the collected data to identify key factors influencing rental prices in the Kuala Lumpur; ii) To compare the performance of machine learning algorithms on predicting rental prices based on the identified predicting factors; iii) To provide valuable insights to help investors or landlords and tenants to make accurate and informed decision in the rental market in Kuala Lumpur.

## 2.0 Data Modelling

In the data modeling process, we investigate the parameters that influence rental monthly rental prices by using Python. The most important features are determined by developing a predictive model using a dataset that has various attributes. The irrelevant attributes are removed from the cleaned dataset to improve the modelling. Next, some categorical variables must be converted into a numeric form. The 'Near KTM/LRT' column which indicates accessibility to public transport is converted from 'Yes'/'No' to 1/0. Also, the 'furnished' column which indicates the furnishing status is converted as 'Not Furnished' to 0, 'Partially Furnished' to 1 and 'Fully Furnished' to 2. The target variable ('monthly\_rent\_rm') was separated from the input variables by applying the "extractInputOutput" function to perform a reliable predictive model. Next, sanity check is performed to ensure the results obtained is closely with the expected proportion for testing data. A correlation heat map is applied to observed the correlation and linear relationships between the variables.

The next step is feature selection and data splitting, the Input Features (x) includes the selected features, which are the completion\_year, rooms, parking, bathroom, size\_sqft, furnished and Near KTM/LRT. The monthly rental price ("month\_rent\_rm") of residential properties is the target variable (y). The data is split into training (80%) and training sets (20%) for data modelling. After splitting the datasets, the Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R-squared ( $R^2$ ) were evaluated to understand and compare the performance of predicting machine learning model.

Decision Tree, Random Forest and Linear regression modelling are implemented to predict the rental price in Kuala Lumpur. The packages used for the modeling are shown below:

```
import pandas as pd
import seaborn as sns
import numpy as np
import matplotlib.pyplot as plt
import plotly.express as px
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
```

All the modeling was trained using the selected features. The coefficients from the linear regression model are analyzed to determine the degree of influence each feature has on rent prices. The coefficients were sorted by the absolute value of their coefficients to identify the most influential variables. Lastly, the coefficients' graphs were plotted for visualization. After the (model\_multi) is initialized and trained on the training data (X\_train, y\_train), predictions are made on the training set, and the first five predicted values are displayed. The Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and  $R^2$  generated to assess the performance of the predictive model. To understand each predictor's coefficients and statistical significance, a detailed summary of the regression model is printed using stats models. To visualise the performance of the machine learning model, predictions generated on the test sets is plotted and the actual values are plotted. All the steps are repeated for Random Forest and Decision Tree Prediction. Below is an example of Python code for the Linear Regression Model.

```
[ ] # Initialize the linear regression model
model_multi = LinearRegression()

# Fit the model to the training data
model_multi.fit(X_train, y_train)

# Predict monthly rent using the model for the training set
predicted_multi_train = model_multi.predict(X_train)

# Display the first 5 predicted values
print("First 5 predicted values:")
print(predicted_multi_train[:5])

# Calculate Mean Squared Error (MSE)
mr_mse = mean_squared_error(y_train, predicted_multi_train)

# Calculate Root Mean Squared Error (RMSE)
mr_rmse = np.sqrt(mr_mse)

# Print the MSE and RMSE
print(f"MSE (Multi-variable model): {mr_mse}")
print(f"RMSE (Multi-variable model): {mr_rmse}")

# Extract and print R-squared value
mr_r_squared = model_multi.score(X_train, y_train)
print(f"R-squared: {mr_r_squared}")

# Print the model summary
import statsmodels.api as sm

# Add a constant to the independent variables
X_train_with_const = sm.add_constant(X_train)

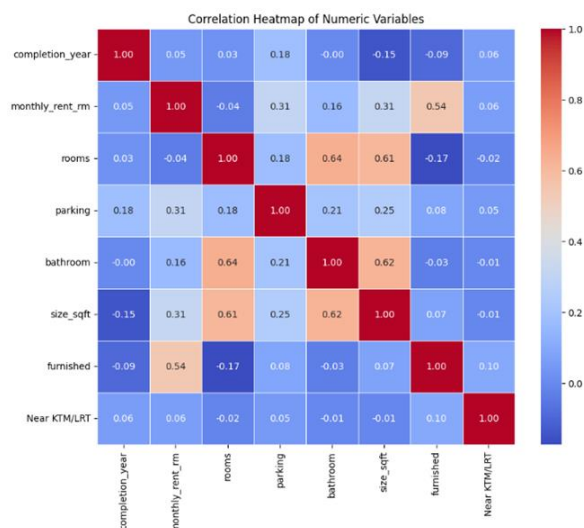
# Fit the OLS model
model_sm = sm.OLS(y_train, X_train_with_const).fit()

# Print the summary
print(model_sm.summary())

# Make predictions using the model for the test set
predicted_test = model_multi.predict(X_test)

# Plot the actual and predicted values
plt.figure(figsize=(10, 6))
sns.scatterplot(x=predicted_test, y=y_test, color='blue', alpha=0.6)
plt.plot([min(predicted_test), max(predicted_test)], [min(predicted_test), max(predicted_test)], color='red', linestyle='--')
plt.title("Actual vs Predicted Values")
plt.xlabel("Predicted Values")
plt.ylabel("Actual Values")
plt.show()
```

### 3.0 Data Interpretation

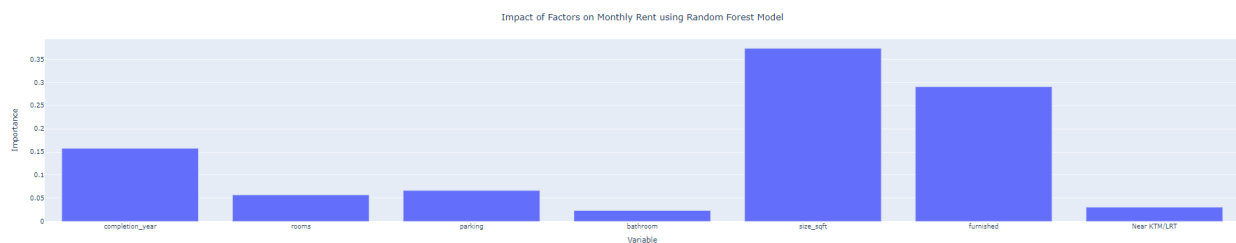


With a correlation coefficient of 0.54, the furnishing status has the strongest strong positive association on monthly rent, according to the correlation heatmap graphic above. It also means that tenants are prepared to pay a higher rent for the comfort and convenience of a furnished unit. Also, there is also a positive correlation between rental pricing and parking availability and property size. This highlights that sufficient parking and a larger property can enhance the value of a property. Moreover, the quantity of rooms is negatively correlated with monthly rent. Not to mention, the property's completion year and ease of access to public transport don't appear to have a significant impact on rental rates.

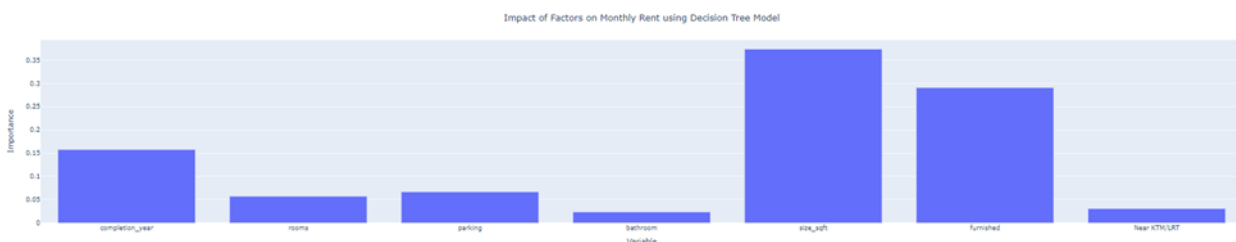
In order to maximise rental income and improve property management strategies, it is essential that investors and property owners understand the maximum coefficient and maximum importance.



The graph of linear regression coefficients indicates that the variables of furnishing status, parking availability, and the number of bathrooms have a positive effect on higher monthly rents. On the other hand, the number of rooms significantly negatively impacts rent, with a value of -178.94 which suggests the presence of unique market dynamics. The completion year, property size, and the accessibility to public transport have a smaller positive influence on monthly rental.



The analysis of importance from the Random Forest model graph indicates that property size and furnishing state are the most important factors in influencing monthly rent, with respective values of 0.37 and 0.29. The number of rooms, accessibility to public transport, and the number of bathrooms is less influential but still contribute to the overall prediction model.



From the Decision Tree model maximum importance graph above, we can see that the graph is similar to Random Forest model maximum importance graph. The graph indicates that the property's size and its furnishing state are the main factors in determining the monthly rent with values of 0.37 and 0.29 respectively. These values are also similar to Random Forest model maximum importance. Table 1 shows the performance of each machine learning Model.

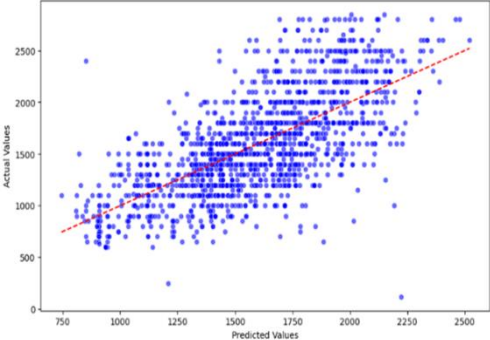
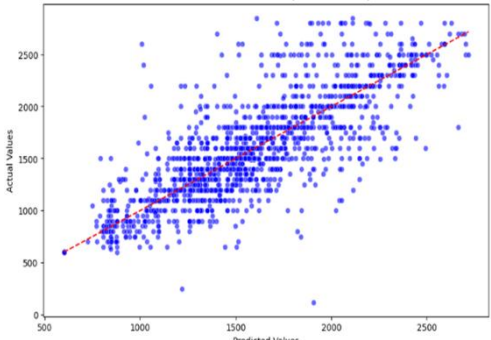
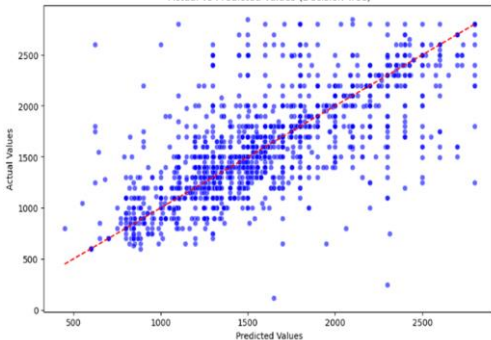
| Model                                | Linear Regression   | Random Forest   | Decision Tree   |
|--------------------------------------|---|---|---|
| The First Five Properties Prediction | RM 1118.78, RM 1473.17, RM 1891.66, RM 1803.98, and RM 2020.18                    | RM 1299.51, RM 1230.58, RM 1667.88, RM 1820.25, and RM 1714.33                      | RM 1302.50, RM 1200.00, RM 1675.00, RM 1700.00, and RM 1707.14                      |
| Mean Squared Error (MSE)             | 129320.06   | 30379.83  | 23615.55  |
| Root Mean Squared Error (RMSE) value | 359.61  | 174.30  | 153.67  |
| R-squared value                      | 0.452   | 0.8713  | 0.8999  |
| Graph                                |  |  |  |

Table 1: Performance of Machine Learning Model

As can be seen from the above table, the Decision Tree model has the lowest MSE and RMSE values besides having a high R squared value when compared to other models. This points out that the most accurate prediction is Decision Tree model which followed by Random Forest and Linear Regression models. In terms of R-squared value, the Decision Tree model performs better than both other two models with a value close to 0.9. A smaller difference between the observed data and fitted values can be seen with a higher R-squared. As can be seen from the actual against predicted values graph of the Decision Tree model, 90% of data points lie on a straight line. In conclusion, the Decision Tree model is the best predictive model and the Linear Regression model is the worst predictive model among the three models in predicting monthly rental prices.

#### **4.0 Plan for Reproducible Research**

Reproducible research is research which can be reproduced by others based on the raw data, code and documentation given. Reproducible research is important in data science because this can ensure the results are valid and reliable.

The first step is to develop a well-documented protocol, which details the proposed study design and methods by including data collection methodology, data preparation, exploratory data analysis, machine learning, evaluation and interpretation of data. A transparent and replicate research process can be created.

Next, suitable materials and tools should be chosen. In this research, to ensure the data is reusable, open data formats are used, for example CSV. Python and R are used since they are widespread and strong support for data science tasks.

Furthermore, reproducible research must be systematically organized. All project files are centralized in Google Drive and OneDrive for this research. Raw data and processed data are stored separately in different folders, in addition, raw data must be set to read-only to preserve its integrity.

Additionally, keeping track of research activities ensures research transparency. A pre-registered study design and analysis plan, for example, a version control system is used to track changes to all files, especially analysis code, maintaining detailed records of research progress. Comprehensive documentation, including a readme file and data dictionary, will provide important context and description of the data and methods used.

Research outputs must be shared in reproducible research. The data, results and materials in the research should be available to the public. Licenses are applied on the code and data to ensure that the research can be legally accessed and utilized by the scientific community.

Lastly, reproducible data should be transparent with detail. All research procedures must be published clearly and comprehensively. All the details need to be published to allow others to reproduce the study. Excessive supplementary information should be avoided by combining essential information into primary publications or related repositories. A reputable platform that supports open access and transparency should be chosen.

In conclusion, the reproducibility of this research could be achieved by following this plan. This approach would increase the validation of findings by researchers and enhance the potential for future research in housing rental prediction.



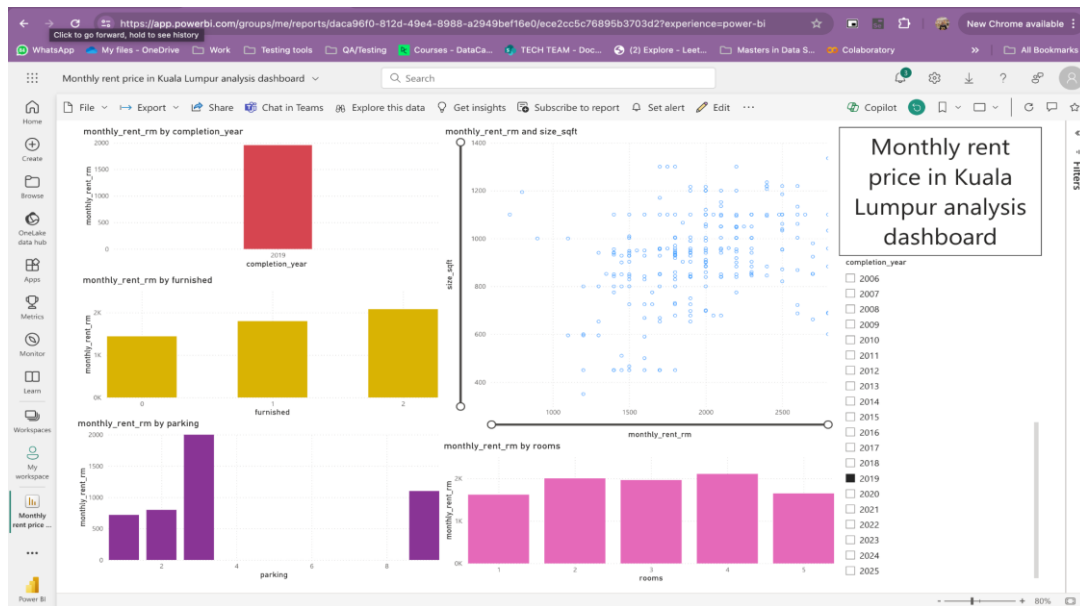
## 5.0 Deployment of Data Product

We have decided to use Power BI as a tool for our data product. Power BI is great at analyzing and predicting monthly rental prices in Kuala Lumpur. The software has strong ability for data integration, advanced analytics, interactive visualizations, and interfaces that are user friendly which collectively offer a full solution for data analysis and insight development.

Besides that, Power BI has the ability to integrate coherently with a diverse array of data sources. It has the ability to integrate with cloud services, SQL databases, Excel spreadsheets, and other platforms. Power Query Editor in Power BI enables for easy data cleaning and transformation. This includes managing null values, transforming data types, and generating computed columns to ensure that our dataset, which includes attributes such as completion year, monthly rent, number of rooms, parking availability, bathrooms, property size, furnishing status, and proximity to KTM/LRT stations, is prepared for precise and comprehensive analysis.

Power BI is highly skilled in generating sophisticated and dynamic visualizations, which are essential for detailed data understanding. It has the ability to generate histograms, line charts, maps, scatterplots and various other visualizations that will assist with conveying our data into meaningful information. Furthermore, Power BI gives us the enablement to customize our dashboards to achieve our goals and requirements. It has the ability to create reports that displays the most important outcome for our users. Displayed below, users have the option to observe and examine the connections and data between monthly rental costs and several attributes including the year of construction, furnishing status, availability of parking, size in square feet, and the number of rooms. This assist their ability to analyze the impact of these features on monthly rental pricing. To add to that, we have the ability to add slicers and filters to provide reports that are both dynamic and interactive. Users can utilize this interactive feature to thoroughly examine the data and acquire profound understanding of the variables that influence rental costs in Kuala Lumpur. As an illustration, we have implemented a completion year slicer that enables users to choose analyze the data based on particular years, facilitating their understanding of trends over time.

Power BI assists with sharing of insights to users in a more user friendly way. We have deployed it to the Power BI service which then creates a link that gives access to the public. Users have the ability to access and engage with these reports on the internet or via mobile devices.



## 6.0 Insights and Conclusion

This study has been conducted to provide valuable insights to allow investors, landlords and tenants accurately predict the future rental price in Kuala Lumpur and to make a better financial or investment decisions (Oshodi et al., 2019). Moreover, investors has been shifting their investment plan to investing in properties for renting and hosting airbnb purposes, as it is a more predictable and lower risk investment compared with other investments (Sabit & Mohammad, 2008; Wickramaarachchi, 2016). As a results, a data-driven approach is need to help landlords, investors, and tenants to accurately predict the monthly rental price in Kuala Lumpur. To fill the research gap, this current study aims to compare the predictive performance of linear regression, random forest, and decision trees model on the residential property rental price in Kuala Lumpur, Malaysia. This study focuses on evaluating several important factors that could affect the rental prices, including the property completion years, number of rooms, parking, and bathroom, the size of the property, the furnishing conditions of the property, as well as whether there is public transportation, mainly KTM or LRT, nearby the property.

The most important factors predicting residential rental prices in Kuala Lumpur are the size of the properties, followed by the furnishing condition, the number of bathrooms, parking spaces and . The results of important predictors of rental prices are in line with recent studies (Ho et al., 2021; Uzut & Buyrukoglu, 2020; Zulkifley et al., 2020). The structural characteristics of a property are important rental or housing price predictors, especially the size of the property (Zulkifley et al., 2020). The number of rooms, bathrooms, and parking, as well as whether the property is near KTM or LRT, have a significantly lower impact on predicting the monthly rental. However, the signifiant positive correlation of the number of parking spaces, bathrooms

and the proximity to KTM or LRT with the monthly rental price in Kuala Lumpur, shows that properties with parking space available, with more number of bathrooms, and near to KTM or LRT station still has a higher investment valued compared with those properties without this facilities.

The number of rooms shows a negative correlation with the rental price, which suggests that the rental value of the property significantly drops when there are more than 2 rooms. Unlike the findings in countries such as Japan, Indonesia, Turkey, and China etc., this research indicates that house age has no influence over the rental prices of housing properties in Kuala Lumpur, Malaysia. One possible explanation is that these countries suffers from several severe natural disasters including earthquake, typhoon and volcanic eruption (Uzut & Buyrukoglu, 2020). Modern houses in these countries will have a higher resistance over natural disasters comparing to the old houses (Uzut & Buyrukoglu, 2020). On the other hand, Malaysia is a geologically stable country which is not at risk of severe natural disasters. Because of its safe from natural disaster, this can explain why housing age is not an important factors in predicting rental price in Kuala Lumpur. In summary, this study suggests that investors should consider investing in larger size, fully furnished properties but with fewer number of rooms for better investment returns. In addition, investors, tenants, and landlords can also take into account the number of parking spaces and bathrooms in a property as well as the proximity to public transport while negotiating the rental price.

From the modeling results discussed earlier, this study suggested that the decision trees (DT) regressor model explains about 90% of the variance in the targeted variables and performed slightly better than the random forest (RF), which explains about 87.1% of the variance in targeted values. The linear regression (LR) model is significantly the least accurate model to predict rental price, as it only explains about 45.2% of the variance in the targeted variables. The results obtained from this study is aligned with several existing studies that applied machine learning model to predict the rental and housing price in India, China, Hong Kong and Malaysia (Nadhirah et al., 2022; Zulkifley et al., 2020). The LR model is a popular predictive model used to predict the relationship between different variables (Abdul S. et al., 2021; Begum et al., 2022). Recent studies found that the LR model performed significantly worse in predicting housing and rental prices as compared with other machine learning models (Abdul S. et al., 2021; Ho et al., 2021; J. McCluskey et al., 2014). McCluskey and colleagues (2014) found that the LR model ( $R^2 = 0.70$ ) is significantly worse at predicting the rental price as compared with the boosted regression tree ( $R^2 = 0.91$ ).

DT and RF are the most popular ML algorithms used in real estate price prediction (J. McCluskey et al., 2014; Nadhirah et al., 2022; Sai et al., 2023; Sharma et al., 2024). DT is a

well known algorithm for classification, prediction and regression of data as it is excellent at capturing complex interactions between variables (Li, 2024; Muhamad Harussani et al., 2021). This result is supported by several recent studies focusing on housing price predictions and observing the performance of different machine learning models (Abdul S. et al., 2021). For instance, a research by Sharma and colleagues in Bangalore predicted the housing price by comparing the performance of different regression models, including DT, RT, XGBoost and Support Vector Regression concluded that DT model is significantly outperformed other models in predicting housing price with an overall accuracy of 99%. Furthermore, Sai and colleagues (2023) also suggested that the DT model provides an accuracy of 71.63% in predicting the housing price, whereas the RF model only provides an accuracy of 63.17%. This current study consists of categorical data with different numbers of levels, and the RF model may be biased in favor of attributes with more values, which resulted in significantly less accuracy in predicting rental or housing prices (Prajwala, 2015). In addition, the DT model is nonparametric, so it is easy to accommodate a variety of numeric or categorical data layers (Prajwala, 2015). The limitations of the RF model could explain why the DT model outperformed the RT model in predicting the rental price.

Future research can examine this area further to determine if comparable results can be obtained using the same online data sets, enabling a better understanding of the limitations of the DT and RT models in predicting rental prices. Future research can evaluate other machine learning models on the same dataset, i.e., SVM. This study analyzed properties that are near KTM and LRT but did not take into account the rental price of the properties that are near MRT. MRT is the most advance and convenient public transport in Klang Valley, future research should include properties near to MRT station to get an up-to-date insights. In conclusion, this current study has provided market insights into the significant predictors and the best machine learning algorithm of predicting residential rental prices in Kuala Lumpur.

## 7.0 References

- Abdul S., M. H., Mohd, T., Masrom, S., Johari, N., & Mohamad Saraf, M. H. (2021). *Machine Learning Algorithms on Price and Rent Predictions in Real Estate: A Systematic Literature Review*.
- Begum, A., Kheya, N. J., & Rahman, Md. Z. (2022). Housing Price Prediction with Machine Learning. *International Journal of Innovative Technology and Exploring Engineering*, 11(3), 42–46. <https://doi.org/10.35940/ijitee.C9741.0111322>
- Fang, Y., Li, T., & Zhao, H. (2022). Random Forest Model for the House Price Forecasting. *2022 IEEE 14th International Conference on Computer Research and Development, ICCRD 2022*, 140–143. <https://doi.org/10.1109/ICCRD54409.2022.9730190>
- Ho, W. K. O., Tang, B. S., & Wong, S. W. (2021). Predicting property prices with machine learning algorithms. *Journal of Property Research*, 38(1), 48–70. <https://doi.org/10.1080/09599916.2020.1832558>
- J. McCluskey, W., Zulkarnain Daud, D., & Kamarudin, N. (2014). Boosted regression trees. *Journal of Financial Management of Property and Construction*, 19(2), 152–167. <https://doi.org/10.1108/jfmpc-06-2013-0022>
- Nadhirah, N., Kamarulzaman, A. B., Syafiqah, N., Shamsudin, B., Zahidah, S., Rashidi, B., Fairos, W., Yaacob, W., & Sobri, M. (2022). Predicting House Rental Value Among Students in Higher Institution Using Data Mining Techniques. In *Applied Mathematics and Computational Intelligence* (Vol. 11, Issue 2).
- Oshodi, O. S., Thwala, W. D., Odubiyi, T. B., Abidoye, R. B., & Aigbavboa, C. O. (2019). Using neural network model to estimate the rental price of residential properties. *Journal of Financial Management of Property and Construction*, 24(2), 217–230. <https://doi.org/10.1108/JFMPC-06-2019-0047>
- Prajwala, T. R. (2015). A Comparative Study on Decision Tree and Random Forest Using R Tool. *IJARCCCE*, 196–199. <https://doi.org/10.17148/ijarcce.2015.4142>
- Sabit, M. T., & Mohammad, H. (2008). *Sustaining the Means of Sustainability: The Need for Accepting Wakaf (Waqf) Assets in Malaysian Property Market*. <https://www.researchgate.net/publication/316830006>
- Sai, P., Reddy, M., & Praveen Chandar, J. (2023). “Decision Tree Regressor Compared with Random Forest Regressor for House Price Prediction in Mumbai.” In *Journal of Survey in Fisheries Sciences* (Vol. 10, Issue 1S).
- Sharma, M., Sharma, D., Burle, R., Patil, P., Joge, I., & Puri, C. (2024). Predicting House Price Model : A Comprehensive Analysis with Random Forest and Decision Tree Method. *2024 3rd International Conference for Innovation in Technology (INOCON)*, 1–6. <https://doi.org/10.1109/INOCON60754.2024.10511732>

- Uzut, Ö. G., & Buyrukoglu, S. (2020). PREDICTION OF REAL ESTATE PRICES WITH DATA MINING ALGORITHMS. *Euroasia Journal of Mathematics, Engineering, Natural & Medical Sciences International Indexed & Refereed*, 8(9), 77–84.  
<https://orcid.org/0000-0001-7844-3168>
- Wickramaarachchi, N. (2016). Determinants of rental value for residential properties: A land owner's perspective for boarding homes. *Built-Environment Sri Lanka*, 12(1), 10.  
<https://doi.org/10.4038/besl.v12i1.7612>
- Zulkifley, N. H., Rahman, S. A., Ubaidullah, N. H., & Ibrahim, I. (2020). House price prediction using a machine learning model: A survey of literature. *International Journal of Modern Education and Computer Science*, 12(6), 46–54.  
<https://doi.org/10.5815/ijmecs.2020.06.04>