MASTER IN DATA SCIENCE

GROUP ASSIGNMENT

COURSE CODE : **WQD7001**

COURSE TITLE : **PRINCIPLES OF DATA SCIENCE**

LECTURER : **ASSOCIATE PROF. DR. MAIZATUL AKMAR BINTI ISMAIL**

PROJECT TITLE : **MONTHLY RENT PRICE PREDICTION IN KUALA LUMPUR**

GROUP MEMBERS :

1. **Chow Kai Ern (S2193226)**
2. **Nur Hidayah binti Ahmad Shafii (22120931)**
3. **Hanisah Zulaikha binti Zulkifli (23072436)**
4. **Loh Bi Jia (23078886)**

Organisation Chart

Key Roles and Responsibilities

Roles	**Responsibilities**
Leader	Guiding the overall direction of the project, setting objectives, and defining the scope of work. Foster collaboration among team members. Act as the primary point of contact for stakeholders. Ensure project progress and adherence to deadlines.
Maker	Collect, clean, and analyze data. Apply statistical methods and machine learning algorithms. Create visualizations to communicate findings. Collaborate with team members on analysis methodologies. Ensure accuracy and reproducibility of analysis results.
Oracle	Provide insights into industry trends and market dynamics. Conduct market research and monitor industry developments. Share expertise on real estate analytics methodologies. Contribute to strategic decision-making and risk mitigation.
Detective	Uncover hidden patterns and anomalies within the dataset. Conduct exploratory data analysis and investigative techniques. Identify outliers and explore relationships between variables. Collaborate with team members to validate findings. Interpret results accurately to derive actionable insights.

Table of Contents

1. Project Background

Kuala Lumpur, the capital city of Malaysia, is at the heart of economic activities. The city's vibrant economy attracts a diverse population of both local and international residents who are in search of temporary accommodation. This has made the rental market a key component of the real estate sector.

In recent years, Kuala Lumpur rental market has experienced a great fluctuation due to influences by several factors including government policies (such as MM2H, changes in OPR, expansion of MRT/LRT network), global events like Covid pandemic, and evolving tenant preferences. The soaring demand for rental properties in urban centers like Kuala Lumpur has intensified the need for accurate rent price prediction models. With the rapid urbanization and influx of people into cities, understanding the dynamics of rental prices becomes paramount for both landlords and tenants. Hence, accurate rent prediction is essential for all stakeholders, be it landlords, tenants, investors and real estate professionals. By offering essential insights, rent prediction helps stakeholders make informed decisions, allowing them to navigate the complexities of the rental market with greater confidence and strategic foresight.

1.1 Problem Statement, Objectives and Project Domain

In Malaysia's dynamic real estate market, determining the monthly cost of renting or owning a residential unit involves various factors such as the number of bedrooms, location, amenities, and other pertinent details. However, there is a lack of a comprehensive tool or framework that accurately calculates the monthly unit cost based on these parameters. To aid tenants and property owners in estimating monthly rental costs accurately, there is a pressing need for a data-driven solution. The three main objectives for this project include,

1. To explore and analyze the collected data to identify key factors influencing rental prices in the Kuala Lumpur area, such as geographical location, property size, amenities, and market demand.
2. To implement machine learning algorithms, such as regression or ensemble methods, to train a model capable of predicting rental prices accurately based on the identified features.
3. To provide recommendations and insights derived from the predictive model to help property owners set competitive rental prices and assist potential tenants in making informed choices when searching for rental properties in the Kuala Lumpur area

This project focuses on developing a predictive model for monthly rent prices in Kuala Lumpur, Malaysia. Analysis will involve exploring patterns, trends, and correlations within the data to identify key factors influencing rental prices and market demand. With Kuala Lumpur being a dynamic urban center with high demand for rental properties, understanding the rental market's intricacies becomes imperative for stakeholders such as property developers, investors, and real estate agents. The project aims to provide insights into factors influencing rental prices in urban environments and to deliver a practical tool for landlords, tenants, and real estate professionals to make informed decisions regarding rental properties in Kuala Lumpur.

2. Literature Review

Rent is frequently described as the regular payment made to a landlord for the utilization of a property, reflecting the compensation for occupying the space. Nevertheless, economists perceive rent as the return obtained from the land itself. As numerous landlords regard property rental as an attractive investment prospect, the rental market has emerged as an appealing option for those seeking a more stable income stream compared to other investments (Wickramaarachchi, 2015).

According to Khan et al., (2011), large-scale property developers and investors often choose to rent out their properties to generate a consistent income stream. It is crucial to have significant sources of funding to purchase the land and construction costs. Ongoing expenses, including maintenance and managing agent fees, are typically covered by the rental income from tenants. Khan et al., (2011) further outlined several reasons individuals may opt to rent besides financial stability which may include, renting offers mobility and financial freedom compared to owning a property. Furthermore, budget flexibility enables individuals to relocate to more economical areas during hard times and upgrade as their income rises. This can free up money for other necessities like food, education, or emergencies. Moreover, renting is cost-effective for those adapting to change and not committed to long-term living in one place. It also avoids the long-term financial commitment and maintenance costs of owning a house. Renting lets individuals save for investments and send money home.

2.1 Factors Influencing Rental Value

The housing market is shaped by supply and demand, which change over time based on economic, social, cultural, geographical, and demographic factors. Meeting housing demand involves policies and economic conditions. Demand arises for consumption, investment, and wealth accumulation. Factors affecting house prices include income level, marital status, industrialization, interest rates, population growth, and migration. Housing prices impact socio-economic and national conditions (Kim and Park, 2005).

The housing prices in Kuala Lumpur have risen significantly over the years, being 40% higher than in other states (Mariadas et al., 2016; Yeee et al., 2021). The Malaysian property market, particularly in Kuala Lumpur, has been very attractive for investors due to the mix of lower purchase costs, higher rental revenue, and lower capital gain taxes (Sabit & Mohammad, 2008). Estimating the rental price is crucial since it influences investment decisions and offers information on the tax accruable (Embaye et al., 2021; Oshodi et al., 2019). In fact, predicting the rental price of a residential property is vital for real estate investors, buyers, financial institutions, and the government to make an important evaluation of its economic viability (Oshodi et al., 2019). There are four predictive variables in predicting the rental price: the characteristics of the property (numbers of bedrooms and bathrooms, surface area, etc.), geographical location (proximity to parks, schools, hospitals, etc.), socioeconomic status of the neighborhood, and mortgage rate (Embaye et al., 2021; Lenaers et al., 2024; Zulkifley et al., 2020). A study in Belgium using machine learning models predicted that the factors that may affect residential rental prices include the number of bedrooms and bathrooms, surface livability, and cadastral income (Lenaers et al., 2024). Similarly, another study in South Africa found that the determinants of rental values include the number of rooms, the availability of electricity, and toilet characteristics (Oshodi et al., 2019). As such current project predicted that property with more facilities, fully furnished conditions, closer to the city centre of the Kuala Lumpur will have a higher rental value compared to less facilities, unfurnished conditions and further away from the city centre.

Strong rental market in Kuala Lumpur, Malaysia whereby the rates rose by 5.5% in 2023 due to historical resilience, post-pandemic recovery, and demographic trends. Projections indicate sustained growth in urban centers as renting becomes more popular due to high homeownership costs. Investors can benefit from attractive gross rental yields (B., 2024). Wong and colleagues (2019) sugested that the increases in income and population is a significant factor influencing both the supply and demand dynamics within the housing market, which further contribute to the strong rental market in Kuala Lumpur in 2023. It is critical to forecast future rental value in Kuala Lumpur utilising a data-driven strategy. Accurate rental price forecast allows investors to make informed judgements about which residential properties to invest in, while tenants can alter their financial plans based on the expected rental market trend.

2.2 Predicting the Rental Price

Machine learning models have become important methods in real estate valuation and estimation (Breuer & Steininger, 2020). This is because the machine learning model makes it possible to make more accurate predictions considering they are more flexible and able to capture complex relationships (Lenaers et al., 2024; Valier, 2020). Advancement of algorithms for machine learning allows computers to learn from data and enhance their performance on specific tasks. This involves utilising data to identify trends, create predictions, and inform decision-making process (Embaye et al., 2021; Lenaers et al., 2024).

Research has primarily utilized machine learning models like Random Forest, Ridge, Tree, Neural Network Model, etc., and hedonic price models, particularly in linear regression to predict rental prices (Embaye et al., 2021; Lenaers et al., 2024; Malpezzi et al., 2002; Oshodi et al., 2019). Existing studies comparing the machine learning model and the hedonic price model in predicting house rents and prices has consistently shown that the machine learning models are superior in accurately predicting housing prices and rents (Do & Grudnitski, 1992; Limsombunc et al., 2004; Valier, 2020). A study predicting the rental value of houses in household surveys in Tanzania, Uganda, and Malawi showed that most of the machine learning models (e.g., Lasso, Random Forest, Ridge, Tree, Bagging, and Boosting) outperformed the hedonic pricing model, specifically in Ordinary Least Squares, in predicting the house rental price (Embaye et al., 2021). Additionally, another study in the US applied the Ordinary Least Squares and Artificial Neural Network approach to predicting the residential home selling price in San Diego, California (Do & Grudnitski, 1992). Their results showed that the neural network approach provided the best prediction of housing values compared to the Ordinary Least Squares approach (Do & Grudnitski, 1992). As such, recent literature reveals that all machine learning models (neural network regression, random forest, gradient boosting regression, tree, etc.) are equally the best models for rent and price prediction (Breuer & Steininger, 2020; Embaye et al., 2021; Limsombunc et al., 2004; Valier, 2020). Based on previous research, it can be concluded that Machine Learning models will be the most accurate technique forecasting the housing and rental markets. This current project will use a machine learning approach to predict rental prices in Kuala Lumpur, Malaysia.

3. Methodology

3.1 Obtain – Data Collection

The data was obtained from mudah.my database where property listings are publicly available and compiled in https://www.kaggle.com/datasets/ariewijaya/rent-pricing-kuala-lumpur-malaysi/data. Regarding the reliability of the dataset, it is assumed the data provided on Mudah.my accurately reflects the data in the region and that the compilation process on Kaggle maintains the integrity of the original listings. The dataset collected comprises various attributes related to rental properties in the Kuala Lumpur and Selangor regions. These include categorical variables such as property name, location, facilities, and additional facilities, and numerical variables such as monthly rent, size, completion year, number of rooms, bathrooms, and parking spaces. This dataset provides a comprehensive overview of rental properties, facilitating analyses related to rental trends, property features, and market dynamics.

3.2 Scrub – Data Cleaning

Data cleaning is an essential step in the data analysis process. Several specific steps were taken during the data cleaning process:

i) **Duplicate Rows Removal**

Initially, duplicate rows were identified and removed to streamline the dataset. Duplicate rows were identified using the duplicated() method and removed from the dataset using the drop_duplicates() method. This involved establishing criteria for matching values across relevant columns, ensuring that only true duplicates were eliminated.

ii) Handling Missing Values

Missing values were handled through techniques such as imputation and deletion. Missing values were identified across columns using the isna() method. The sum() method was then used to calculate the total number of missing values for each column. For critical columns like 'prop_name' and 'monthly_rent', rows with missing values were dropped from the dataset using the dropna() method with the subset parameter specifying the columns of interest. Imputation using median values was employed for column 'completion_year' to retain data completeness.

iii) Addressing Data Inconsistencies

In the data cleaning process, formatting values played a crucial role in ensuring consistency and accuracy across the dataset. The formatting adjustments are:

a) Monthly Rent Formatting: The 'monthly_rent' column underwent cleaning to remove non-numeric characters such as "RM" and "per month". This was achieved using the str.replace() function to replace these characters with an empty string. Subsequently, the values were stripped of any whitespace characters and converted to float format using the astype(float) method.
b) Size Formatting: Similar to the 'monthly_rent' column, the 'size' column was formatted to remove the unit "sq.ft.". This was achieved using the str.replace() function to replace " sq.ft." with an empty string, followed by conversion to float format using the astype(float) method.
c) Location Name Standardization: Location names in the 'location' column were standardized by removing redundant prefixes such as "Kuala Lumpur -" and "Selangor -". This was accomplished using the str.replace() function to replace these prefixes with an empty string, ensuring uniformity and consistency in location representation.

iv) Outlier

The Interquartile Range (IQR) method identified outliers in numerical columns like 'monthly_rent_rm' and 'size_sqft'. Lower and upper bounds based on IQR were used to define outliers, which were then eliminated from the dataset to prevent undue influence on analysis outcomes. This approach aimed to enhance reliability and accuracy by stabilizing the dataset's distribution and mitigating the impact of outliers on model performance.

3.3 Exploratory Data Analysis

i) Describe the dataset

The feature consisted of:

Feature	Description
ads_id	Listing ID
prop_name	Name of the building/ property
completion_year	Completion/ established year of the property
monthly_rent	Monthly rent in Ringgit Malaysia (RM)
location	Property location in Kuala Lumpur region
property_type	Property type such as apartment, condominium, flat, duplex, studio, etc
rooms	Number of rooms in the unit

parking	Number of parking space for the unit
bathroom	Number of bathrooms in the unit
size_sqft	Total area of the unit in square feet
furnished	Furnishing status of the unit (fully, partial, non-furnished)
facilities	Barbeque area, Club house, Gymnasium, Jogging Track, Lift, Minimart, Multipurpose hall, Parking, Playground, Sauna, Security, Squash Court, Swimming Pool, Tennis Court, Air-Cond, Cooking Allowed, Internet, Near KTM/LRT, Washing Machine

ii) Univariate analysis

Monthly rental (N= 8293, M= 1593.89, SD= 485.16) is going to be predicted. Minimum month rental will be RM 100, maximum monthly rental will be RM 2850. The first quartile of the monthly rental will be RM 1250, second quartile will be RM 1500, and third quartile will be RM 1900.

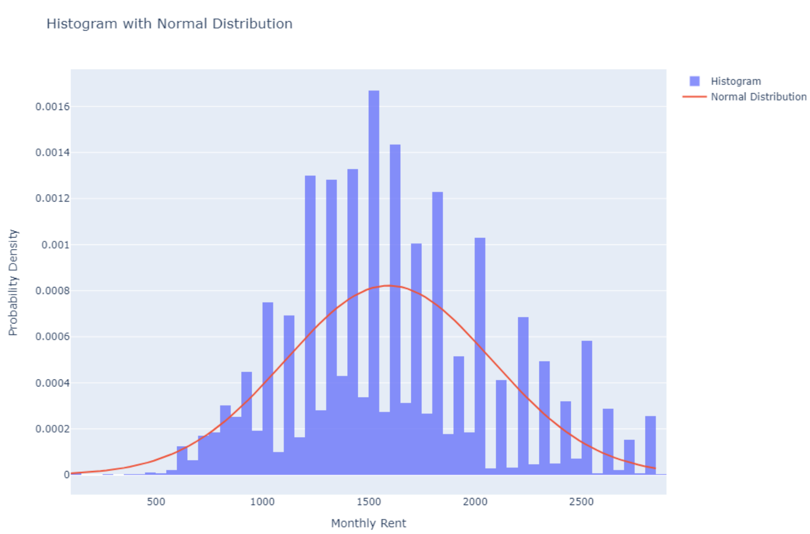

Figure 1: Histogram of Monthly rental distribution with normal distribution line

The histogram in Figure 1 shows the distribution of monthly rental in Kuala Lumpur. The distribution is normal distribution as compared to the normal distribution line plotted.

iii) Bivariate analysis

Bivariate analysis is separated into two parts: (i) Numerical variables analysis; and (ii) Categorical variables analysis.

(i) Numerical variables analysis

Discrete numerical variables include: 'rooms', 'parking', 'bathroom'

Continuous numerical variables include: 'completion_year', 'monthly_rent_rm', 'size_sqft'

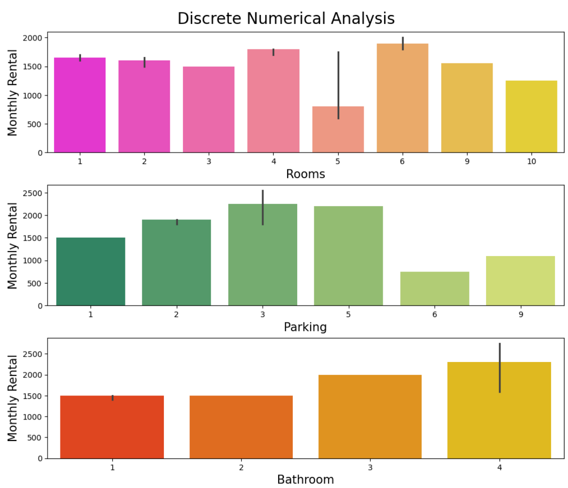

Figure 2: Barchart of Discrete numerical variables

From Figure 2 above, the average house monthly rental is in the range 1500-2000 for different numbers of rooms in the house except with 5 rooms. The monthly rental increases as the number of parking increases until the number of parking reaches 5 and above parking. The monthly rental increases as the number of bathrooms increases. From these three features, the numbers of rooms and parking should not be selected for further analysis as neither feature are not directly proportional to the monthly rental.

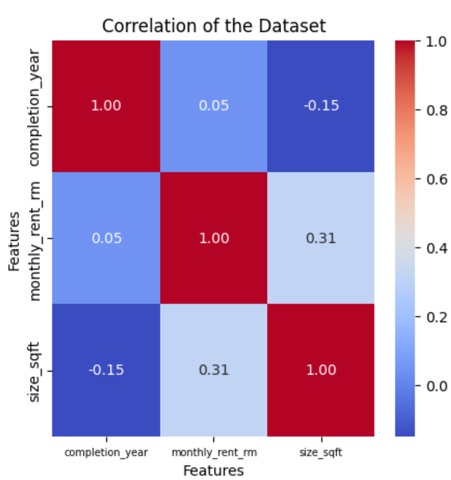

Figure 3: Heatmap of Continuous numerical variables

The heatmap in Figure 3 indicates both completion year and size of house have weak positive correlation (<1) with monthly rent which means both features are not directly correlated to monthly rental.

(ii) Categorical variables analysis

Categorical variables include:

'location', 'property_type', 'furnished', 'Barbeque area', 'Club house', 'Gymnasium', 'Jogging Track', 'Lift', 'Minimart', 'Multipurpose hall', 'Parking', 'Playground', 'Sauna', 'Security', 'Squash Court', 'Swimming Pool', 'Tennis Court', 'Air-Cond', 'Cooking Allowed', 'Internet', 'Near KTM/LRT', 'Washing Machine'

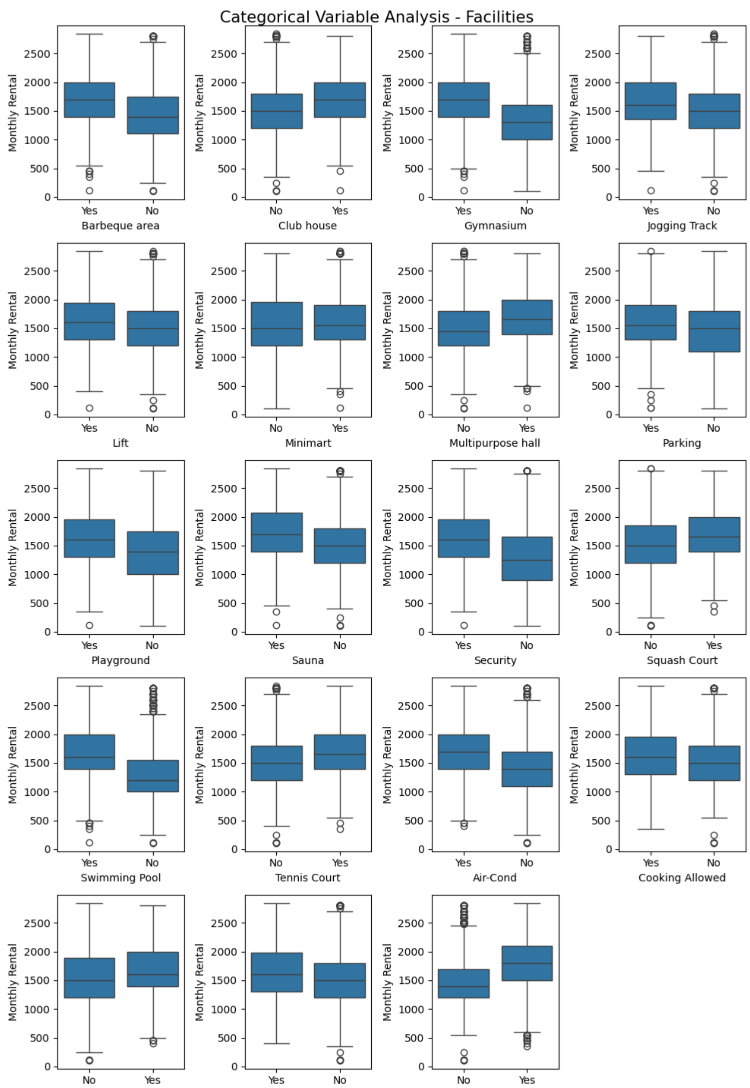

Figure 4: Boxplot of Categorical variables (Facilities)

The boxplot in Figure 4 shows the monthly rental for houses with facilities are higher than without facilities.

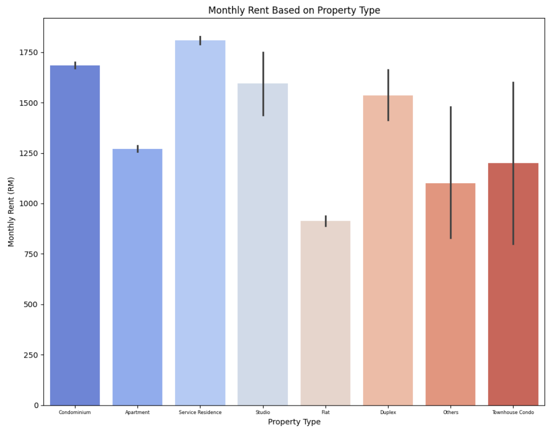

Figure 5: Barchart of Monthly Rent by Property type

Figure 5 indicates the average monthly rental by different property types. The top 3 properties with high monthly rental is condominiums, service residence and studio.

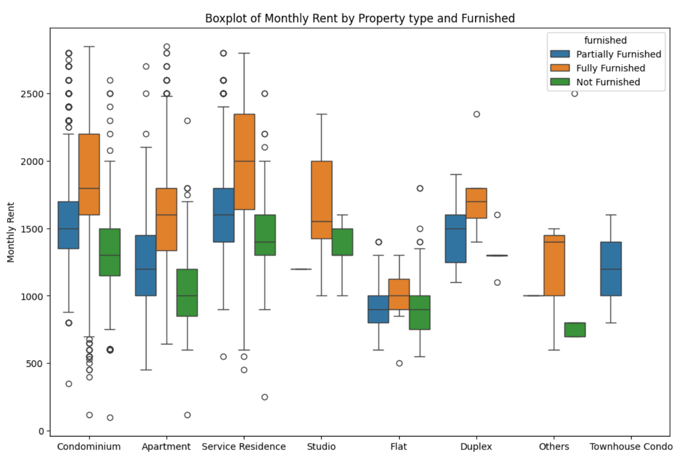

Figure 6: Monthly Rent by Categorical variables (Property type and Furnished)

The boxplot in Figure 6 indicates the relationship between property type and furnished type of house with monthly rental. The monthly rental of fully furnished houses is greater than partially furnished while not furnished houses with lowest monthly rental even in different types of property types.

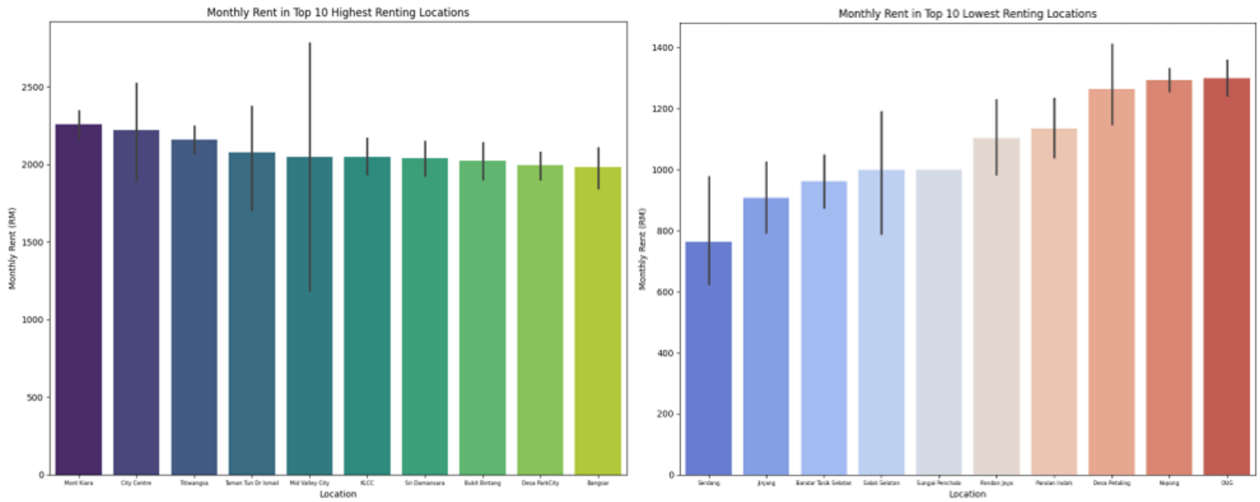

Figure 7: Top 10 and bottom 10 monthly rentals by location.

Figure 7 indicates the top 10 locations with high monthly rental and bottom 10 locations with low monthly rental out of 56 locations. Overview, houses in Mont Kiara with the highest monthly rental (RM 2256.71) while in Serdang with the lowest monthly rental (RM 763.75).

3.4 Exploratory Data Analysis Conclusion

In conclusion, our analysis of monthly rental prices in Kuala Lumpur reveals key insights. Firstly, rentals exhibit a normal distribution pattern. Secondly, while room numbers and parking don't directly correlate with rent, factors like facilities, property type, and furnishing significantly influence prices. Additionally, completion year and house size show weak correlations with rent. Lastly, Mont Kiara tops the list for highest rents, while Serdang has the lowest. These findings underscore the complexity of factors influencing rental prices in Kuala Lumpur and provide valuable insights for further predictive modeling and decision-making in the real estate market.

4. Ethical Consideration

Responsible data scientists must address key ethical considerations when using online datasets. Firstly, they should ensure data privacy and consent by obtaining permission from data owners and adhering to data protection regulations. The personally identifiable information (PII) in the data should be anonymised or aggregated so that the privacy of individuals represented in the data is protected.

Secondly, the data should not be biased. The potential biases in the online dataset may be due to sample selection, data collection methods, or historical inequalities. Biases in the data will influence analyses and perpetuate inequalities. Hence, we are responsible for identifying and mitigating biases in datasets to ensure the fairness and accuracy of our analyses.

Lastly, transparency and accountability are key when using third-party datasets. It's necessary to disclose this information in the analysis report to ensure transparency about data provenance. This requires maintaining clear records of data sources and methods, providing accurate and interpretable analyses, and ensuring that our findings are used ethically and responsibly to inform decision-making processes. Prioritizing ethical considerations over purely commercial interests is essential when considering the potential impact of our analyses on individuals or communities.

5. Impact

This project has the potential to have a significant impact on society. Firstly, accurate predictions of rent prices can provide valuable insights for landlords, tenants, and property managers, enabling them to better understand market trends. By leveraging data-driven models, they can make more informed decisions about setting rental rates, ultimately contributing to a fairer and more transparent rental market. For tenants and property owners, rent price predictions offer a valuable tool for financial planning. Tenants can anticipate rent increases and budget accordingly, while property owners can make smarter decisions about their investments. This helps create more stable and predictable relationships between landlords and tenants, reducing the risk of financial strain or displacement. This project's insights into excessively high rental areas can inform policymakers and housing advocacy groups, facilitating efforts to promote fair housing. By leveraging this information, they can advocate for changes to enhance housing affordability for low-income families. Ultimately, these initiatives offer greater opportunities for everyone to secure affordable housing options. Predictions of rent prices enable investors to identify profitable areas for property rental, facilitating informed decision-making and maximizing investment returns. When investors make smart choices, it helps keep the property market stable and strong for everyone. Furthermore, data-driven insights give tenants more power to stand up for themselves in the rental market. When tenants have information about rent affordability and unfair practices, they can make better choices and speak up if something isn't right. This makes the rental market fairer for everyone, ensuring that all voices are heard and respected. In conclusion, this project has far-reaching implications for housing markets and society.

6. References

B., K. (2024, January 15). *Rents to rise after 5.5 pct increase in 2023, pointing to a stronger*

Breuer, W., & Steininger, B. I. (2020). Recent trends in real estate research: a comparison of recent working papers and publications using machine learning algorithms. In *Journal of Business Economics* (Vol. 90, Issue 7, pp. 963–974). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/s11573-020-01005-w

Do, A. Q., & Grudnitski, G. (1992). A Neural Network Approach to Residential Property Appraisal. *Real Estate Appraiser*, *58*(3), 38.

Embaye, W. T., Zereyesus, Y. A., & Chen, B. (2021). Predicting the rental value of houses in household surveys in Tanzania, Uganda and Malawi: Evaluations of hedonic pricing and machine learning approaches. *PLoS ONE*, *16*(2 February). https://doi.org/10.1371/journal.pone.0244953

housing market: New Straits Times. NST Online. https://www.nst.com.my/property/2024/01/1001527/rents-rise-after-55-pct-increase-2023-pointing-stronger-housing-market

Khan, Firoz & Scholtz, Louise. (2011). Rental Housing.

Kim, K., & Park, J. (2005). Segmentation of the housing market and its determinants: Seoul and its neigh-bouring new towns in Korea. Australian Geographer, 36(2), 221–232

Lee, C.L. and Ting, K.H. (2011), “Linkage between malaysian houisng prices, property companies and stocks”, Pacific Rim Property Research Journal, Vol. 17 No. 2, pp. 287-312, doi: 10.1080/ 14445921.2011.11104329.

Tan, Y.K. (1999), “An hedonic model for house prices in Malaysia”, Proceedings of the PRRES Conference, Kuala Lumpur, available at: www.prres.net, pp. 1-15.

Wong, W.L., Chin, L. and Seow, S.K. (2019), “Housing prices in peninsular Malaysia: supported by income, foreign inflow or speculation?”, International Journal of Housing Markets and Analysis, Vol. 12 No. 3, pp. 424-441, doi: 10.1108/IJHMA-01-2018-0001.

Lenaers, I., Boudt, K., & De Moor, L. (2024). Predictability of Belgian residential real estate rents using tree-based ML models and IML techniques. *International Journal of Housing Markets and Analysis*, *17*(1), 96–113. https://doi.org/10.1108/IJHMA-11-2022-0172

Limsombunc, V., Gan, C., & Lee, M. (2004). House Price Prediction: Hedonic Price Model vs. Artificial Neural Network. *American Journal of Applied Sciences*, *1*(3), 193–201. https://doi.org/10.3844/ajassp.2004.193.201

Malpezzi, S., Gibb, K., & O'sullivan, A. (2002). *Hedonic Pricing Models: A Selective and Applied Review*. http://www.bus.wisc.edu/realestate

Mariadas, P. A., Selvanathan, M., & Hong, T. K. (2016). A Study on Housing Price in Klang Valley, Malaysia. *International Business Research*, *9*(12), 103. https://doi.org/10.5539/ibr.v9n12p103

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

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

Valier, A. (2020). Who performs better? AVMs vs hedonic models. In *Journal of Property Investment and Finance*(Vol. 38, Issue 3, pp. 213–225). Emerald Group Holdings Ltd. https://doi.org/10.1108/JPIF-12-2019-0157

Wickramaarachchi, N. (2015). Determinants of rental value for residential properties; A land owner's perspective for boarding homes.

Yeee, G. F., Sufahani, S. F., Wahab, M. H. A., & Idrus, S. Z. S. (2021). Factors influence the housing price in Kuala Lumpur by using AHP. *Journal of Physics: Conference Series*, *1793*(1). https://doi.org/10.1088/1742-6596/1793/1/012027

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