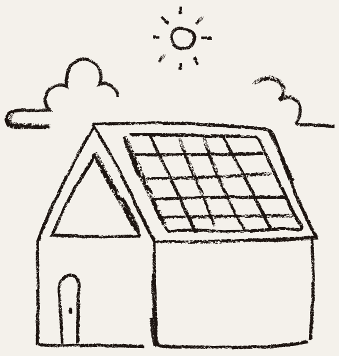
Advancing HOUSE PRICE PREDICTION

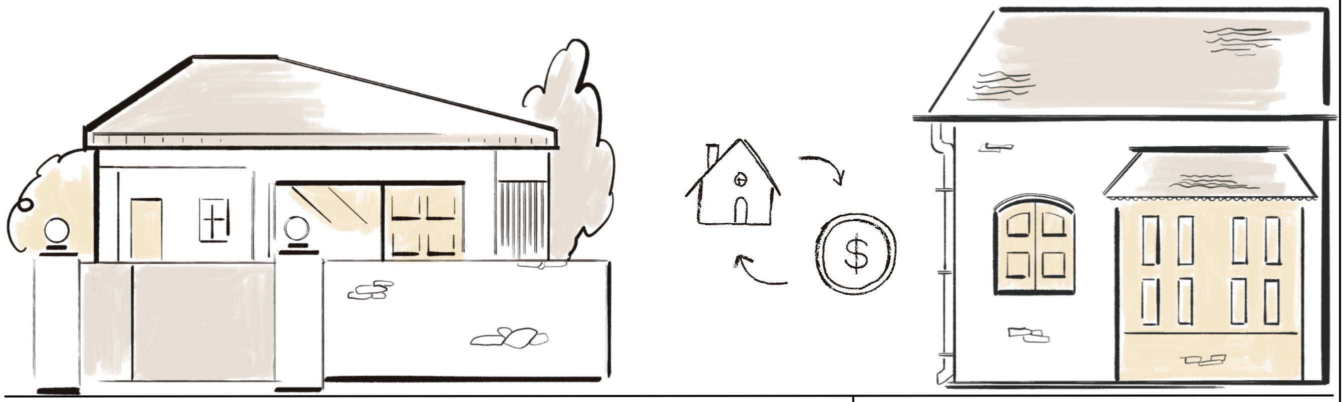
THROUGH **MACHINE LEARNING**



Group 7

2602190816 – Kimberly Kayla Dewi
2602199582 – Fiona Maharani Nugraha
2602227775 – Keitaro Alexander Herman

Department of Data Science, School of Computer Science, BINUS University @Kemanggisan Campus, Jakarta, Indonesia.



Problem Statement

The real estate market is a complex and dynamic system influenced by various factors such as location, size, amenities, economic conditions, and market trends.

Accurate prediction of housing prices is **crucial** for buyers, sellers, and real estate professionals to make **informed decisions**.

Traditional methods often fall short in capturing the intricate patterns and dependencies present in the housing market, making it an **ideal candidate for machine learning-based** solutions.

Objective

Develop a **machine learning model** capable of accurately **predicting housing prices** based on a set of relevant features.

The goal is to **find the best predictive model** that outperforms traditional methods, providing a reliable tool for stakeholders in the real estate industry.

Data Description

- **Price**: The price of the house.
- **Area**: The total area of the house in square feet.
- **Bedrooms**: The number of bedrooms in the house.
- **Bathrooms**: The number of bathrooms in the house.
- **Stories**: The number of stories in the house.
- **Mainroad**: Whether the house is connected to the main road (Yes/No).
- **Guestroom**: Whether the house has a guest room (Yes/No).
- **Basement**: Whether the house has a basement (Yes/No)
- **Hot water heating**: Whether the house has a hot water heating system (Yes/No).
- **Airconditioning**: Whether the house has an air conditioning system (Yes/No).
- **Parking**: The number of parking spaces available within the house.
- **Prefarea**: Whether the house is located in a preferred area (Yes/No).
- **Furnishing status**: The furnishing status of the house (Fully Furnished, Semi-Furnished, Unfurnished).

DATA UNDERSTANDING

The dataset comprises **545 rows** and **13 columns**, encompassing both **integer** and **categorical data types**. The objective is to **predict house prices** through **regression analysis.**

Notably, there are **no missing values** in the dataset, ensuring completeness.

Five features, namely **'area,' 'bathrooms,' 'stories,' 'airconditioning,'** and **'parking'** are identified as having a substantial impact on housing prices.

DATA PREPARATION

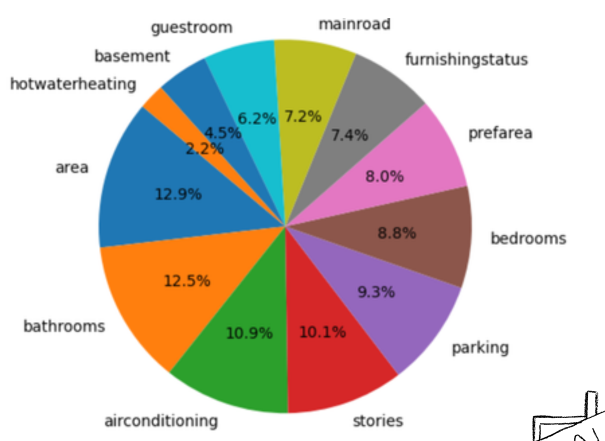
Categorical data transformation into binary format is implemented. This process involves converting categorical variables into binary indicators, facilitating **their integration into regression models**.

Additionally, **outliers in the 'price' and 'area' columns are removed**. Outliers can significantly impact the performance of regression models, and their removal ensures a more robust analysis.

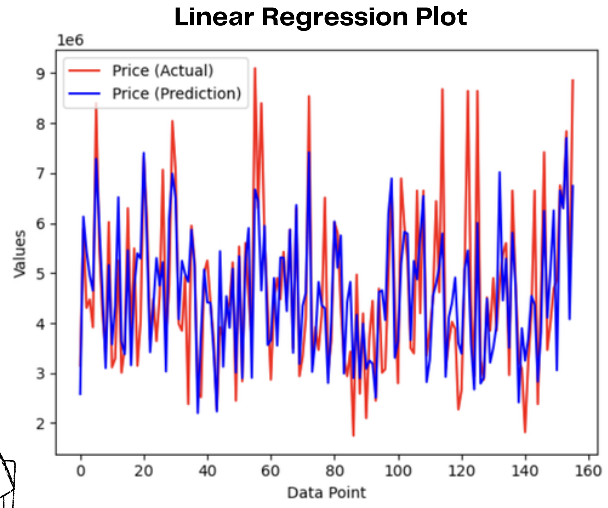
MODELS

- KNN Regression
- Linear Regression
- Decision Tree Regression
- K-Means Regression

Correlation between price and other features



Linear Regression Plot



price = a.area + b.bathrooms + c.stories + d.airconditioning + e.parking

Evaluation

Linear Regression Metrics	RMSE: 958852.46047 MAE: 736616.72228 R^2 : 0.60716
KNN Regression Metrics	RMSE: 1289323.55612 MAE: 993955.12820 R^2 : 0.28972
Decision Tree Regression Metrics	RMSE: 1145557.49021 MAE: 869808.05570 R^2 : 0.43929

K-Means Regression Metrics

```
For n_clusters = 2 The average silhouette_score is : 0.6151371089761382
For n_clusters = 3 The average silhouette_score is : 0.5607074950099619
For n_clusters = 4 The average silhouette_score is : 0.5483556539076002
For n_clusters = 5 The average silhouette_score is : 0.5461958285988512
For n_clusters = 6 The average silhouette_score is : 0.5231147032633476
For n_clusters = 7 The average silhouette_score is : 0.536888560402004
For n_clusters = 8 The average silhouette_score is : 0.5421936784743625
For n_clusters = 9 The average silhouette_score is : 0.5392657767916794
```

The K-Means results, where clusters 2-5 have very similar values but a **significant difference is observed when entering cluster 6**, it suggests that the algorithm has identified a natural grouping of data points up to a certain point (clusters 2-5)
.

After that, the differences become more pronounced, indicating a **distinct separation** in the data. The convergence criteria mentioned above will stop the algorithm when these conditions are met, ensuring a stable and meaningful partitioning of the data into clusters.

Conclusion!

Based on the results of evaluating model performance metrics, it can be concluded that **Linear Regression is the best choice for predicting house prices**. With an R^2 of around 0.60716, **Linear Regression is able to explain around 60.7% of the variation in house prices** using the five selected features. The lower RMSE and MAE values further indicate that the Linear Regression model provides predictions that are **more accurate and closer to the true value when compared to other models**. The advantages of Linear Regression, including **its simplicity and ease of interpretation**, are essential in home price prediction.

Additionally, it is worth noting that **the results between Linear Regression and K-Means Regression are almost similar. K-Means regression performs better when the number of clusters (n_clusters) is set to 2** because it is **better suited to modeling patterns in a data set that contains two distinct groups** that can be identified well by the model. We use this number of clusters because **there is a decrease in performance when n_clusters is set to more than 2** which highlights the limitations of the K-Means Regression model in handling structures or variations that are more complex and cannot be represented clearly.

Therefore, **Linear Regression remains a stable and consistent choice**.