# Business Statistics Using R - Group Assignment (Proposal)

Abdul Hakim Bin Kamalur Rahman (24015257)
Annabel Ching Ke Xin (24002685)
Divani A/P Arumugam (19058908)
Harresh A/L Ragunathan (19076090)
Lai Woei Harng (20003158)

2024-12-16

## i. Narrative

### Research Question

How do structural, condition, and environmental factors influence housing prices, and can the inclusion of interaction terms enhance the accuracy of prediction models for the housing market?

### Introduction

Housing prices are central to both individual financial stability and broader economic growth. Homes serve as a foundational asset, influencing financial well-being and market trends (Yao & Feng, 2023). Rapid price fluctuations, as seen in Auckland's rise in the house price-to-income ratio from 6.4 in 2010 to 10.0 in 2016, highlight the profound impact of housing markets on personal finances and national economies (Greenaway-McGrevy & Phillips, 2021).

Traditionally, housing price determination relies on economic indicators such as employment levels and interest rates (Rico-Juan & Taltavull de La Paz, 2021). However, these fail to account for the complexity of housing markets, which are also shaped by sociodemographic dynamics and subtle factors like linguistic choices in real estate advertisements (Markowitz, 2023). While tools like the House Price Index (HPI) track pricing trends, they are limited by external shocks such as the COVID-19 pandemic and data inconsistencies (Aliefendioğlu et al., 2022; Sipan et al., 2018).

Machine learning offers a solution by capturing complex patterns in housing data, outperforming traditional models in predictive accuracy (Fang, 2023). This research explores how structural, condition, and environmental factors drive housing prices and evaluates whether interaction terms improve machine learning models for prediction.

### Research Questions

1. How do structural factors such as house area, number of bedrooms, bathrooms, stories and the availability of guestroom, basement and parking space influence the market price of a property?
2. How do house condition factors such as hot water heating, air conditioning and furnishing status impact housing prices?
3. What is the effect of environmental factors, including proximity to main roads and location in preferred areas on the market value of a property?
4. How do interaction terms improve the accuracy of housing price prediction models?

## Research Objectives

1. To examine the influence of structural factors including house area, number of bedrooms, bathrooms, stories and the availability of a guestroom, basement and parking space on the market price of a property.
2. To assess the impact of house condition factors such as hot water heating, air conditioning and furnishing status on housing prices.
3. To analyze the effect of environmental factors such as proximity to main roads and location in preferred areas on the market value of a property.
4. To evaluate the effectiveness of interaction terms in enhancing the accuracy of housing price prediction models.

## Motivation of Research

Accurate housing price predictions are critical in a volatile real estate market. Developers face rising construction costs and financial risks, while buyers and sellers struggle to navigate market dynamics (Khalid et al., 2018). Mismanagement often leads to abandoned projects, financial losses, and diminished market credibility (Mac-Barango, 2017). Buyers overpay due to insufficient market knowledge, and sellers risk pricing their properties uncompetitively (Wanjiku et al., 2021; Muñoz & Cueto, 2017).

This research addresses these challenges by developing advanced prediction models, enabling developers, agents, and buyers to make data-driven decisions and improve financial sustainability.

## Business-related dataset

The housing price dataset includes 545 observations and 13 columns, each representing a distinct property and its associated characteristics. This dataset focuses on structural, condition, and environmental factors influencing housing prices. These factors provide insights into the dynamics of real estate valuation and aid in building robust prediction models for housing markets.

URL: https://www.kaggle.com/datasets/yasserh/housing-prices-dataset/data

## Description of variables

The dataset captures a mix of numeric, categorical, and binary variables that describe each property's physical attributes, location, and additional features.

Table 1: Table: Description of Variables

| Variable Name | Description | Data Type |
|---|---|---|
| area | The size of the house in square feet | Numeric |
| bedrooms | The number of bedrooms | Numeric |
| bathrooms | The number of bathrooms | Numeric |
| stories | The number of stories | Numeric |
| main road | The location of the house, whether it is on the main road (yes, no) | Binary |
| guestroom | Whether there is a guest room (yes, no) | Binary |
| basement | Whether there is a basement (yes, no) | Binary |
| hot water heating | Whether there is hot water heating (yes, no) | Binary |
| air conditioning | Whether there is air conditioning (yes, no) | Binary |
| parking | The number of parking spaces | Numeric |
| preferarea | Whether it is in a preferred area (yes, no) | Binary |
| furnishing status | The furnishing status (furnished, semi-furnished, unfurnished) | Categorical |

| Variable Name | Description | Data Type |
|---|---|---|
| price | The price of the house | Numeric |

### Original Data Collection

The original curator collected the data from real estate transaction records and property listings. These sources typically include:

1. **Transaction Data**: Housing price data was likely sourced from public and private records, reflecting the finalized sale prices of properties within specific timeframes.
2. **Real Estate Listings**: Details like area, number of bedrooms, bathrooms, and features such as basement or air conditioning were extracted from property advertisements maintained by developers and real estate agents.
3. **Geographical Information**: Environmental factors such as proximity to main roads and preferred areas were derived from geospatial data or manually annotated based on neighborhood desirability.
4. **Developer or Agent Contributions**: Input from housing developers and agents provided additional context, such as furnishing status and amenities.

The combination of these data sources ensures that the dataset comprehensively represents the housing market while capturing the key factors affecting property valuation.

## ii. The Data

```
str(df)
```

```
## 'data.frame':    545 obs. of  13 variables:
##  $ price           : int  13300000 12250000 12250000 12215000 11410000 10850000 10150000 10150000 98
##  $ area            : int  7420 8960 9960 7500 7420 7500 8580 16200 8100 5750 ...
##  $ bedrooms        : int  4 4 3 4 4 3 4 5 4 3 ...
##  $ bathrooms       : int  2 4 2 2 1 3 3 3 1 2 ...
##  $ stories         : int  3 4 2 2 2 1 4 2 2 4 ...
##  $ mainroad        : chr  "yes" "yes" "yes" "yes" ...
##  $ guestroom       : chr  "no" "no" "no" "no" ...
##  $ basement        : chr  "no" "no" "yes" "yes" ...
##  $ hotwaterheating : chr  "no" "no" "no" "no" ...
##  $ airconditioning : chr  "yes" "yes" "no" "yes" ...
##  $ parking         : int  2 3 2 3 2 2 2 0 2 1 ...
##  $ prefarea        : chr  "yes" "no" "yes" "yes" ...
##  $ furnishingstatus: chr  "furnished" "furnished" "semi-furnished" "furnished" ...
```