Data Analytics Program: Data Science Track

Capstone Project: Prediction of Real Estate Prices

Student Name: Hazem Kamal Al Hendawi

Date: 13th Jan 2025

* Introduction

## 

## Real estate pricing is a critical component of the housing market and urban planning. Accurate predictions of property prices can empower investors, developers, and individuals to make informed decisions about purchasing or investing in real estate. This project aims to develop a predictive model for estimating the price per unit area of real estate properties based on key factors, such as property age, distance to public transportation, and nearby amenities. By analyzing these variables, we aim to identify the primary drivers of real estate prices and build a reliable tool for price prediction.

* Problem

## The challenge in real estate pricing lies in accounting for the multitude of factors influencing property values. These include physical characteristics (like property age and location) and external factors (such as proximity to convenience stores and public transportation). Variability in these factors introduces significant uncertainty in price estimation, making it difficult for buyers and sellers to gauge fair market values. This project addresses the need for a robust model capable of predicting property prices based on a range of features, providing clarity and confidence to all stakeholders in the real estate market.

* Data set

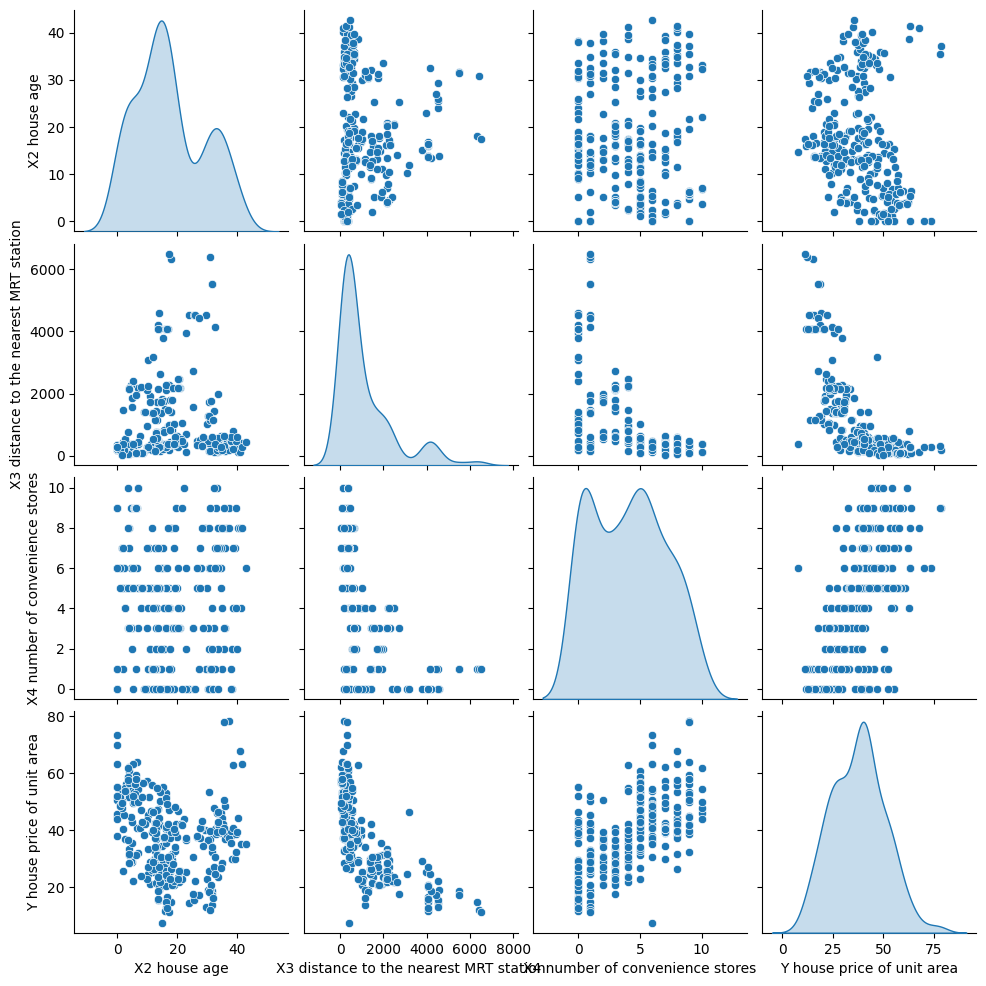
Source : Real estate

Link : https://drive.google.com/file/d/1gGYsdOr408BqyYdRsXCUOq6SXxd7di08/view?usp=sharing

## 

## Attribute Information

* Transaction Date: The date of the property transaction.
* House Age: The age of the house in years.
* Distance to Nearest MRT Station: The distance to the closest MRT (Mass Rapid Transit) station in meters.
* Number of Convenience Stores: The count of convenience stores within the vicinity of the property.
* Latitude: The latitude coordinate of the property.
* Longitude: The longitude coordinate of the property.
* House Price of Unit Area: The dependent variable representing the price per unit area of the property.
* Data visualization



Analysis of the relationship between variables using matrix plot

The relationship between variables was analyzed using matrix plot to understand the distribution and correlation between variables.

Key findings:

Distribution of variables:

Housing unit prices (Y) are concentrated at values ​​less than 50.

Distance to MRT station (X3) is often short with limited number of large values.

Bilateral relationships:

There is a clear inverse relationship between housing unit price (Y) and distance to MRT station (X3).

The relationship between house age (X2) and unit price (Y) is non-linear, with variation in prices.

There is no strong relationship between the number of nearby shops (X4) and unit price.

Importance of analysis:

This analysis helps to identify influential variables, such as distance to MRT station, to be used in building predictive models.

A graph of blue bars

Description automatically generated

The bar graph shows the relationship between the number of stores and the distance to the nearest MRT station. It can be seen that areas with more stores tend to be closer to MRT stations, indicating that densely populated areas have better facilities.

A graph of blue dots

Description automatically generated

The graph shows the relationship between the distance to the nearest MRT station and the price of a residential unit. It can be seen that properties closer to MRT stations tend to have higher prices, indicating that proximity to public transportation increases the value of the property.

A graph of a number of convenience stores

Description automatically generated

The graph shows the relationship between the number of nearby stores and the price of a residential unit. It can be seen that the more nearby stores the higher the price of a residential unit, indicating that proximity to amenities increases the value of the property.

A graph showing a line

Description automatically generated

The graph shows the trend of housing prices over time. Fluctuations in prices can be observed with a general trend of increase over the specified period, indicating growth in the real estate market.

* Model

# K-nearest neighbors (KNN)

* + Random forest
* Results

**K-Nearest Neighbors (KNN):**

* + Mean Squared Error: 54.35606746987952
  + R-squared: 0.6759886132079893

**Random Forest:**

* + Mean Squared Error: 71.20551591407651
  + R-squared: 0.5755506416768443
* Explanation

Two models were evaluated for predicting real estate prices: K-Nearest Neighbors (KNN) and Random Forest. The KNN model showed better performance with a Mean Squared Error (MSE) of 54.356 and a coefficient of determination (R-squared) of 0.676, indicating that it explains about 67.6% of the variance in the data. On the other hand, the Random Forest model showed an MSE of 71.206 and an R-squared of 0.576, indicating poorer performance compared to KNN. Based on these results, it can be concluded that the KNN model is more effective in predicting real estate prices in this case.

* Github:

**https://github.com/DivHazem/GSGProject**