**Linear Regression to predict house price per unit area.**

**Overview.**

The goal of this project is to predict real estate prices using linear regression. Prices vary for buying a house depending on a lot of factors that bring convenience to the buyers. Using all the available features and identifying that the distance of the metro station from the house was one of the dominating factors for varying prices for a house.

**Repository Navigation.**

**Code**:

1. EDA: Real\_Estate\_EDA.ipynb
2. Model: Linear\_Regression\_model\_Evaluation.ipynb

**Presentation:** Slide Deck

**Readme Navigation.**

**Data.**

Data Source: <https://www.kaggle.com/quantbruce/real-estate-price-prediction>

The dataset is obtained from Kaggle.com real estate price predictions with locations in Taiwan. The data has 8 columns and 415 rows. The data has no missing values or any other inconsistencies. The maximum price per unit area in this location is 117.5. The median price per unit area in the given dataset is 38.45.

A screenshot of a cell phone

Description automatically generated

**Modeling.**

Factors:

1. Age of house.
2. Distance from the metro station in units.
3. No. of convenience stores nearby.
4. Latitude.
5. Longitude.

**Results:**

Variance explained by the model is 56.10%.

The MSE is 71.78(varying) which displays the error of the model.

The relation of the distance from metro station against the prices is not linear but provides demonstration that the prices are high if the distance is low.

A screenshot of a cell phone

Description automatically generated

The relation between the number of convenience stores to the price of the house per unit area is shown below.

**A picture containing screenshot, drawing

Description automatically generated**

The relation between the age of house and price per unit area does not show any linearity.

A screenshot of a cell phone

Description automatically generated

**Future.**

The future goal is to make a model with more accuracy and data which can explain better variance of the data. There can be other factors like hospitals and distance from airports which can be brought into knowledge for a better understanding of the variation in prices due to their locations.

**Project Info:**

Contributors: Rushabh Shah.

Languages: Python.

Tools/IDE: Colab, Anaconda.

Libraries: pandas, matplotlib, sklearn, statsmodels.

Duration: September 2020.