Predicting Houses price using linear regression

Introduction:

The goal of this project is that we are going to build a linear regression model that predicts the prices of the houses based on the features. The data was scraped from http://house.speakingsame.com/ website. We will do a comprehensive analysis with all data cleaning, exploration, visualization, feature selection, model building, and evaluation.

Question:

Can this model really predict the houses prices based on the features?

Dataset Description:

The dataset has 33.7k rows and 19 columns.

Column names are:

- **address** : Street Address
- **suburb** : Name of the suburb
- **price** : Price on last sold date
- **bedrooms** : Number of bedrooms
- **bathrooms** : Number of bathrooms
- **garage** : Number of garages/car spaces
- **land_area** : Land area in square meters
- **floor_area** : Floor area in square meters
- **buil_year** : Year the house was built
- **CBD_dist** : Distance to Central Business District
- **nearest_stn** : Nearest train station
- **nearest_stn_dst** : Distance to the nearest train station
- **date_sold** : Date the property was last sold
- **POSTCODE** : Postal code of the suburb
- **LATITUDE** : Latitude of address
- **LONGITUDE** : Longitude of address
- **NEAREST_SCH :** Nearest ATAR-applicable school
- **NEAREST_SCH_DIST** : Distance to nearest ATAR-applicable school

- **NEAREST_SCH_RANK** : Rank of ATAR-applicable school

Tools:

• Pandas and NumPy packages to manipulate data.

• Matplotlib library for visualizing data.

• Linear Regression from the sklearn.linear_model class of the sklearn module

• mean_squared_error from the sklearn.metrics module.

• mean_absolute_error from the sklearn.metrics module to measure the accuracy of the model.

• Jupyter notebook to execute the code.