Foundations of Machine Learning 1st Assignment

Dataset: Real State Dataset

Mahan Madani – 99222092

1. Overview

This dataset is comprised of 414 records of houses, containing information in 8 different columns. All of the columns are numerical and include information on each individual house's price, coordinates, and local access.

The goal of this project is to process the data, run a number of statistical tests, and train a simple regression model to predict house prices.

2. Dataset Exploration & Analysis

The dataset stores 8 different attributes. Each record in the dataset belongs to one specific house.

Below you can find a list of all the attributes:

- **No:** The index of the record, starting from 1.
- **X1 transaction date:** The date of the house, stored as a float.
- **X2 house age:** The age of the house.
- **X3 distance to the nearest MRT station:** The distance to the nearest Mass Rapid Transit station.
- **X4 number of convenience stores:** Number of nearby stores.
- **X5 latitude:** Indicates the latitude of the coordinates.
- **X6 longitude:** Indicates the longitude of the coordinates.
- **Y house price of unit area:** Price is the target variable.

3. Data Preprocessing

3.1 Check for Duplicate Records:

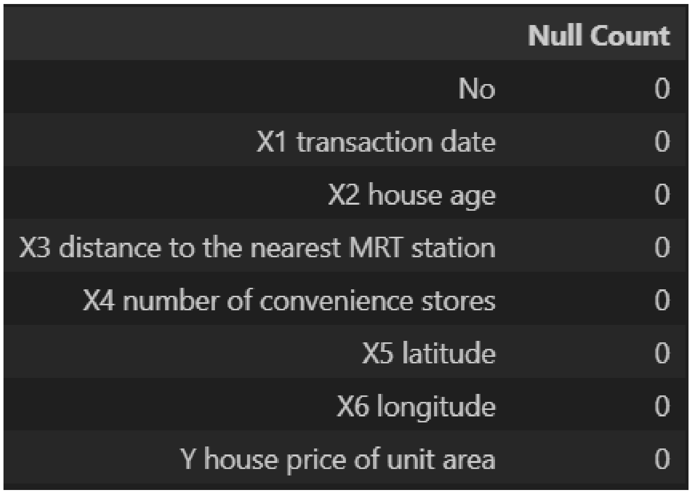
To ensure the dataset contains no duplicate records, a combination of all features (except ‘No’) can be used. If two or more records share the same value for these columns, that means they are potentially duplicates.

Even if the records aren’t completely identical, they may still contradict each other and need to be handled. **After testing the dataset, it seems that it does not contain any duplicate records**.

3.2 Handle Null Values:

As seen in this table, none of the dataset’s attributes contain null or missing values.

Therefore, no data imputation is required for this project.



	Null Count
No	0
X1 transaction date	0
X2 house age	0
X3 distance to the nearest MRT station	0
X4 number of convenience stores	0
X5 latitude	0
X6 longitude	0
Y house price of unit area	0

3.3 Detect Outlier Values:

Using the z-score method, outlier data can be identified. Outliers can be detected from the ‘Y house price of unit area’ column, and exactly 1 outlier record was detected based on this feature. Due to its insignificance and the low amount of data available, I decided not to drop that record.

3.4 Feature Generation:

Based on the available data, the following feature can be added to be used for hypothesis testing.

- **Age:** A categorical value describing the age of the house with four different categories (very old to very new) using the quartiles.

4. Visualization

Figure 4.1: Histogram of house age

This plot displays a histogram of house ages distributed in 20 bins.

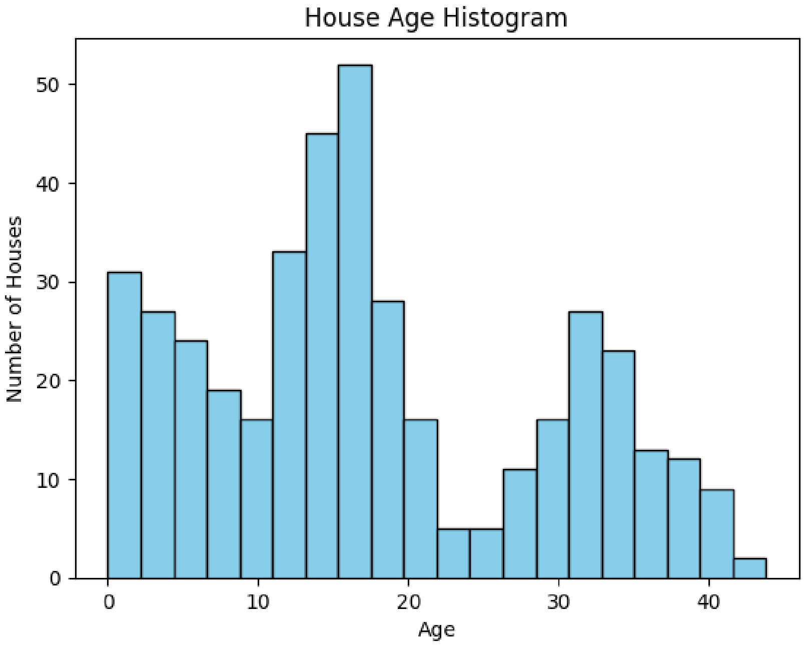


Figure 4.2: Density plot of house price

This plot displays the density of house prices. We can infer that the majority of house prices fall somewhere around 40 to 50.

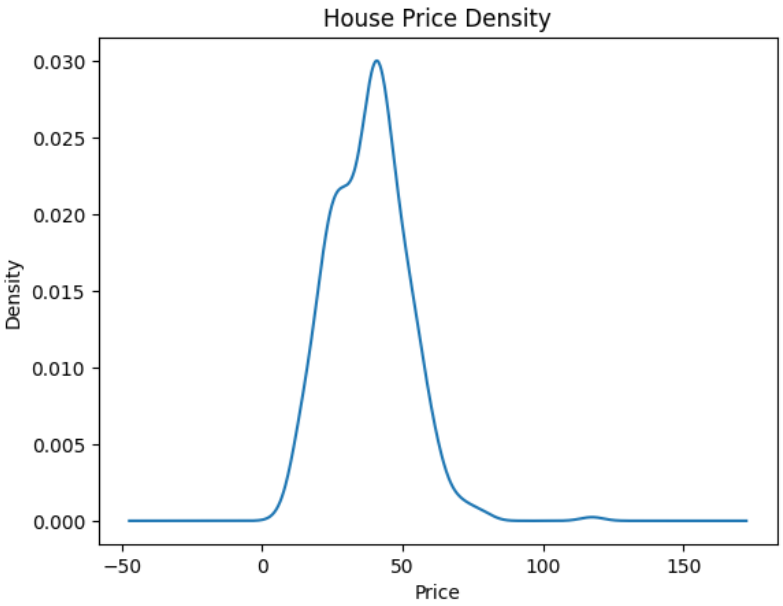
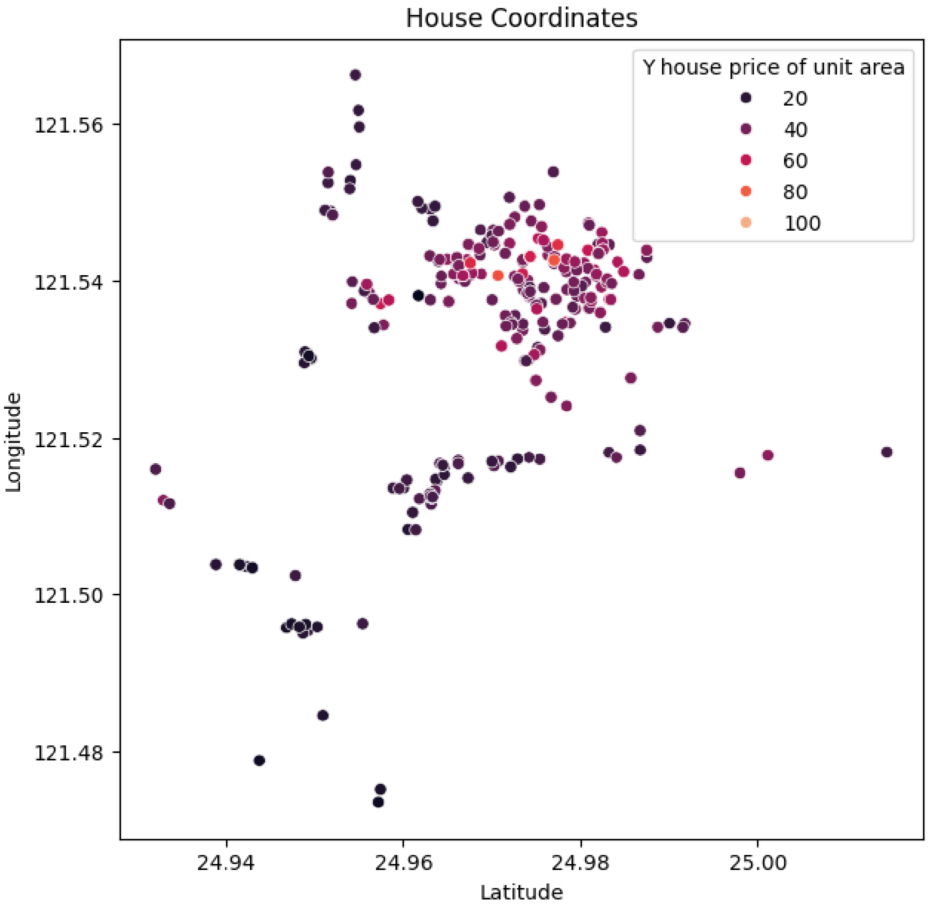


Figure 4.3: Scatter plot of house coordinates

This scatter plot uses the values from the latitude and longitude features of the dataset to display a simulation of the map where the houses are located. Each dot represents a house.

The hue of the scatter plot indicates the price range of a house. We can see that the majority of the more expensive houses are grouped together, creating a neighborhood.



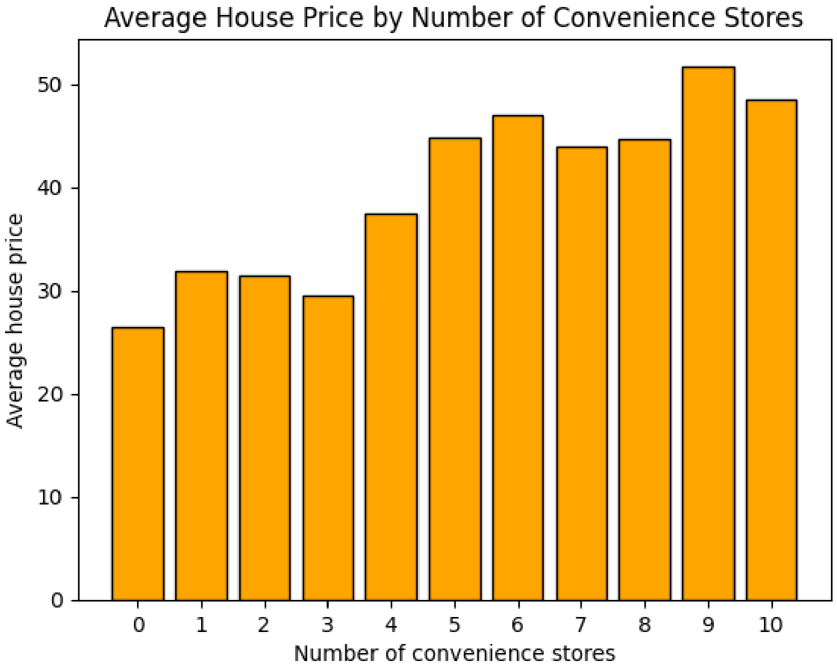


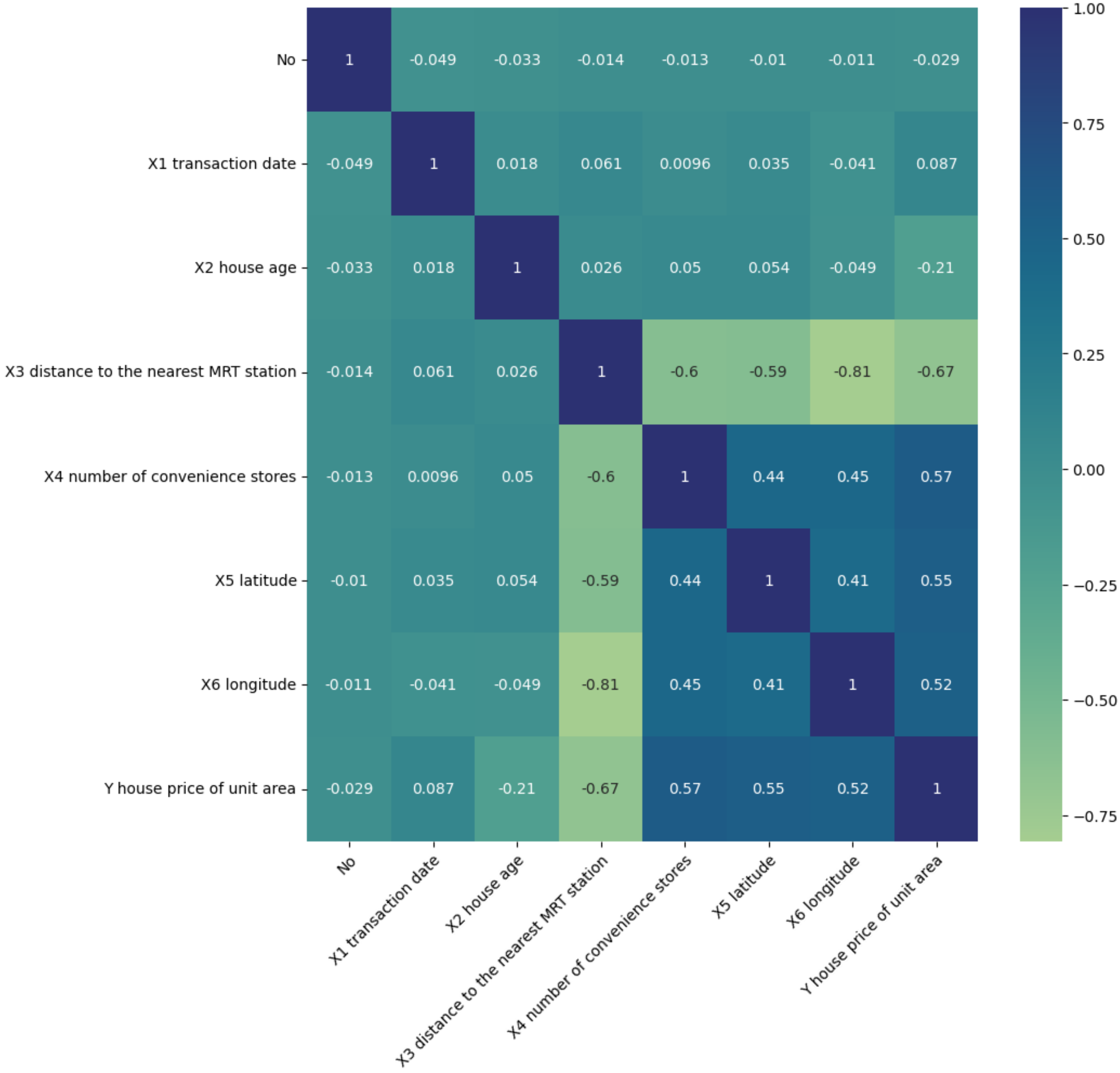
Figure 4.4: Bar plot of Average House Price by Number of Convenience Stores

To create this bar plot, the data was grouped based on the number of convenience stores near a house, and then the average house price was calculated for each group.

Figure 4.5: Heatmap of the Correlation Matrix

This heatmap visualizes the correlation between the various attributes of the dataset. Some strong correlations can be found here, such as the correlation between the ‘number of convenience stores’ and the ‘price of the house’.

The ‘distance to the nearest MRT station’ seems to have the biggest negative correlation with the house price.



5. Statistical Tests

For hypothesis testing, a number of statements will be presented and then by utilizing a statistical test, their credibility is evaluated.

Significance Level (**α**) = 0.05

If the calculated p-value is smaller than **α**, the null hypothesis is rejected and the alternative hypothesis is accepted.

5.1 Hypothesis Test 1

Null hypothesis: There is no significant difference between the average price per unit area of houses above the median age and those below the median age.

‘house price of unit area’ is quantitative and roughly follows a normal distribution. Dividing all of the data into 2 groups (below median age, above median age) allows us to use a two-sample t-test to find the mean of the two samples.

Test type: **Two-sample T-Test**, sample size = **200**

T-Statistic	P-Value	Result
-4.98324	9.22205e-07 < α	Reject Null Hypothesis

Alternative Hypothesis: There is significant difference between the average price per unit area of houses above the median age and those below the median age.

5.2 Hypothesis Test 2

Null hypothesis: There is no significant difference between the average price per unit area and the number of convenience stores.

Comparing a quantitative feature with a categorical feature with more than 2 samples can be done using the ANOVA (analysis of variance) test.

Note that the data must be grouped by the categorical feature (which is the number of convenience stores). The values of 'Y house price of unit area' are then compared between these different groups.

Test type: **One Way ANOVA**

Statistic	P-Value	Result
24.92075	1.17810e-36 < α	Reject Null Hypothesis

Alternative Hypothesis: There is significant difference between the average price per unit area and the number of convenience stores.

5.3 Hypothesis Test 3

For this test, I chose the ‘Age’ and ‘Number of Convenience Stores’ features as categorical features. ‘Age’ was created earlier to categorize house price into 4 different categories.

Null hypothesis: There is no significant association between the age of the house and the number of convenience stores.

Comparing two categorical features can be done using the Chi-2 test.

Test type: **Chi-2**

Statistic	P-Value	Result
135.17212	2.65536e-15 < α	Reject Null Hypothesis

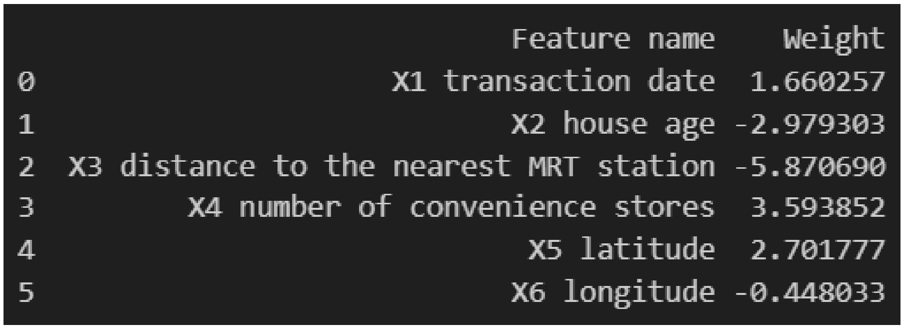
Alternative Hypothesis: There is a significant association between the age of the house and the number of convenience stores.

6. Training a Regression Model

I trained a simple linear regression model on 80 percent of the data and used the remaining 20 percent to test the model. I used MSE (Mean Squared Error) for evaluation and the MSE value was: 53.3047

Note that I used a standard scaler to scale all of the train and test data. The target variable (price) remains untouched.

Extracting the model's weights shows some interesting data, as you can see below:



```
                            Feature name    Weight
0                    X1 transaction date  1.660257
1                           X2 house age -2.979303
2  X3 distance to the nearest MRT station -5.870690
3        X4 number of convenience stores  3.593852
4                            X5 latitude  2.701777
5                           X6 longitude -0.448033
```

The feature with the highest weight appears to be the number of convenience stores, which already had a high correlation with the price. Additionally, the distance to the nearest MRT station was a feature with a high negative correlation with price, and it also has the smallest weight in the model. Longitude surprisingly has a negative weight; despite being highly correlated with price.

This may potentially be due to the **Multicollinearity** of the features. Longitude and latitude are highly correlated with each other. This can lead to unstable coefficient estimates, causing the sign of the coefficient for the highly correlated feature to flip in the presence of other correlated features.