

# Real Estate Valuation

---

Kun Huang, Yizhou Jin, group 6

We aim to predict the house price `house_price` in New Taipei City, Taiwan using a data set containing the following variables,

- `trans_date`: the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
- `house_age`: the house age (unit: year)
- `distance_mrt`: the distance to the nearest MRT station (unit: meter)
- `stores`: the number of convenience stores in the living circle on foot (integer)
- `latitude`: the geographic coordinate, latitude. (unit: degree)
- `longitude`: the geographic coordinate, longitude. (unit: degree)

The response variable is `house_price`, the house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 meter squared).

The original contest can be see [here](#).

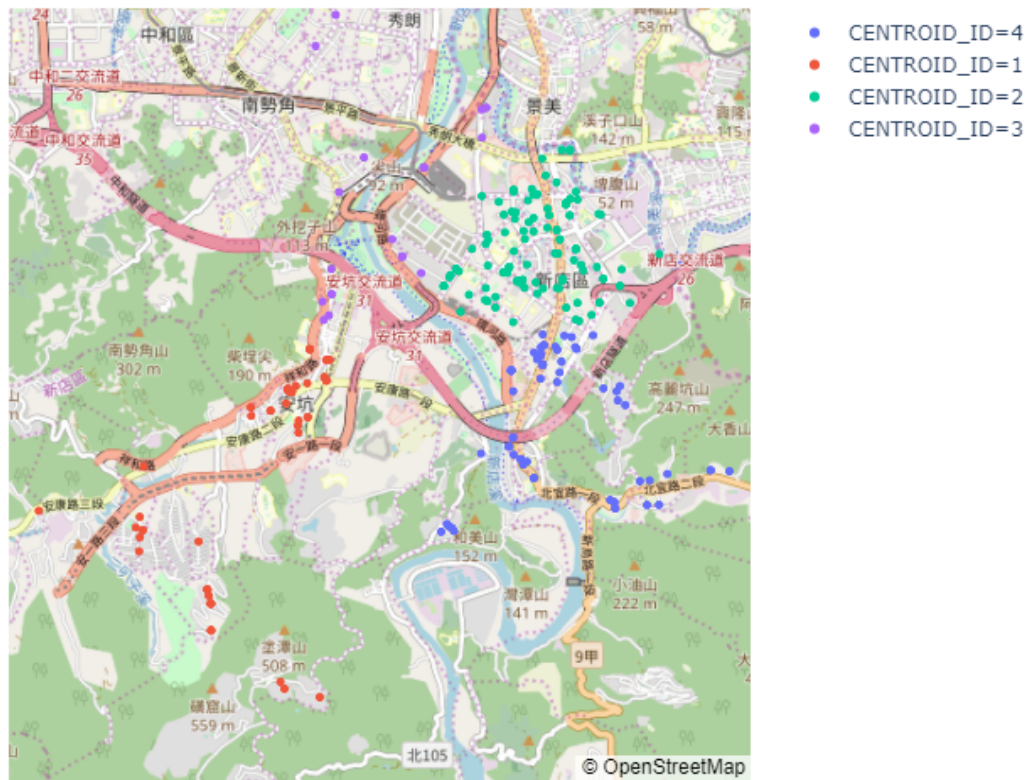
## Preprocessing and Exploratory data analysis

---

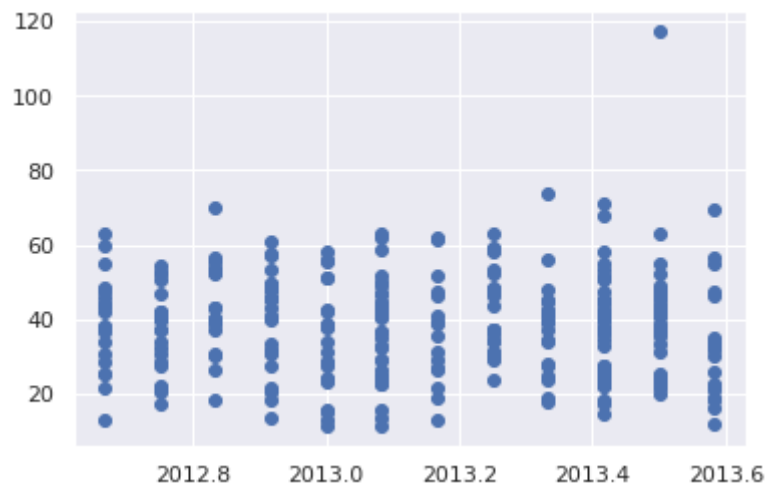
Since directly using the geographic coordinates as predictor variables will lead some problems, we transform the longitude and latitude into a categorical variable , `CENTROID_ID` to indicate the location of each observation.

After looking at the map, we roughly observed that there are 4 clusters in the map, hence we decided to cluster longitude and latitude based on data features into 4 groups by using K-means clustering. This step is imposed on the whole data set, i.e., the training set and test set.

## The Scatterplot of House Location



Besides, we decided to exclude predictor `trans_date` before building the model since based on the following figure, we can see the season of transaction dates had relatively small influence on the house price.











By using the K-means model, we successfully cluster the latitudes and longitudes into four groups as follows.

Davies–Bouldin index	1.0114
Mean squared distance	0.5062

## Numeric features

This table shows the centroid value for each feature. Use the select menu to view more numeric features.

Select features (2/2) ▾

Centroid Id	Count	latitude	longitude
1	79	 24.9547	 121.5068
2	188	 24.9773	 121.5404
3	34	 24.9836	 121.5243
4	113	 24.9609	 121.5429

Then we subset the result table by train group and test group and we used `CENTROID_ID`, `house_age`, `distance_mrt`, and `stores` as the predictors and `house_price` as the response variable to train the model and do the prediction.

## Model

We build a linear regression model using the BigQuery ML. Then we run the regression model and we use trained model to predict house price based on test data. The result is showed as follows.

↗

	predicted_house_price
0	39.871868
1	39.810054
2	40.970729
3	39.853762
4	40.077838
...	...
120	40.380154
121	41.192638
122	38.843652
123	38.795686
124	41.307743

125 rows × 1 columns

By submitting the data to Scriptedin, the score(MSE) is 13.1672 which is relatively good for a small dataset with size of 414.

Submission	Description	Score
<a href="#">results.csv</a>	result	13.1672