House Price Prediction and Attribute Analysis

Group 4

Introduction to datasets

The dataset was originally used to predict house prices in California. It contains more than 20,000 data samples and each of which consists of ten variables. The following are the explanations of the variables in terms of meaning:

Longitude/ latitude: Together form coordinates that represent particular places on Earth.

housing_median_age: The median age of the house.

total_rooms: The number of rooms in the house.

total_bedrooms: The number of bedrooms in the house

population: Population in the area where the house is located.

households: The number of families suitable for living in the house.

median_income: Median income in the area where the home is located.

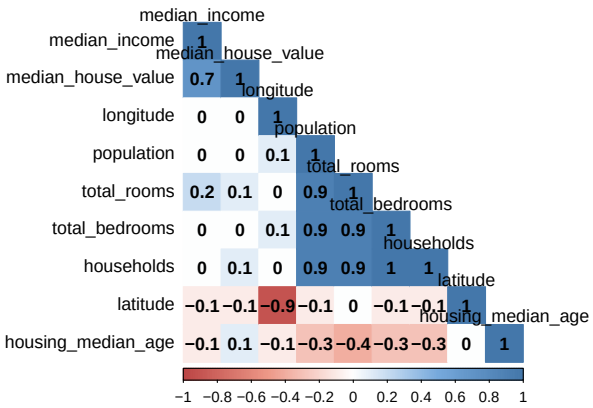
median_house_value: The median of the house value.

Recall that our goal is to 1) Fit a model for house price prediction. 2) Explore the relationship between various variables, estimate and simulate the distribution and characteristics of sample data, and calculate their errors

Data analysis

To preserve the analysis is feasible, the step is to normalize all the attributes to eliminate the influences bring by scale. And after that, for the model fitting, a correlation matrix should be plotted for observing the relationship between variables.

```
## corrplot 0.92 loaded
```



By the plot, we obtain that there are strong correlations between the four variables (total_rooms, population, total_bedrooms, and households) in pairs.

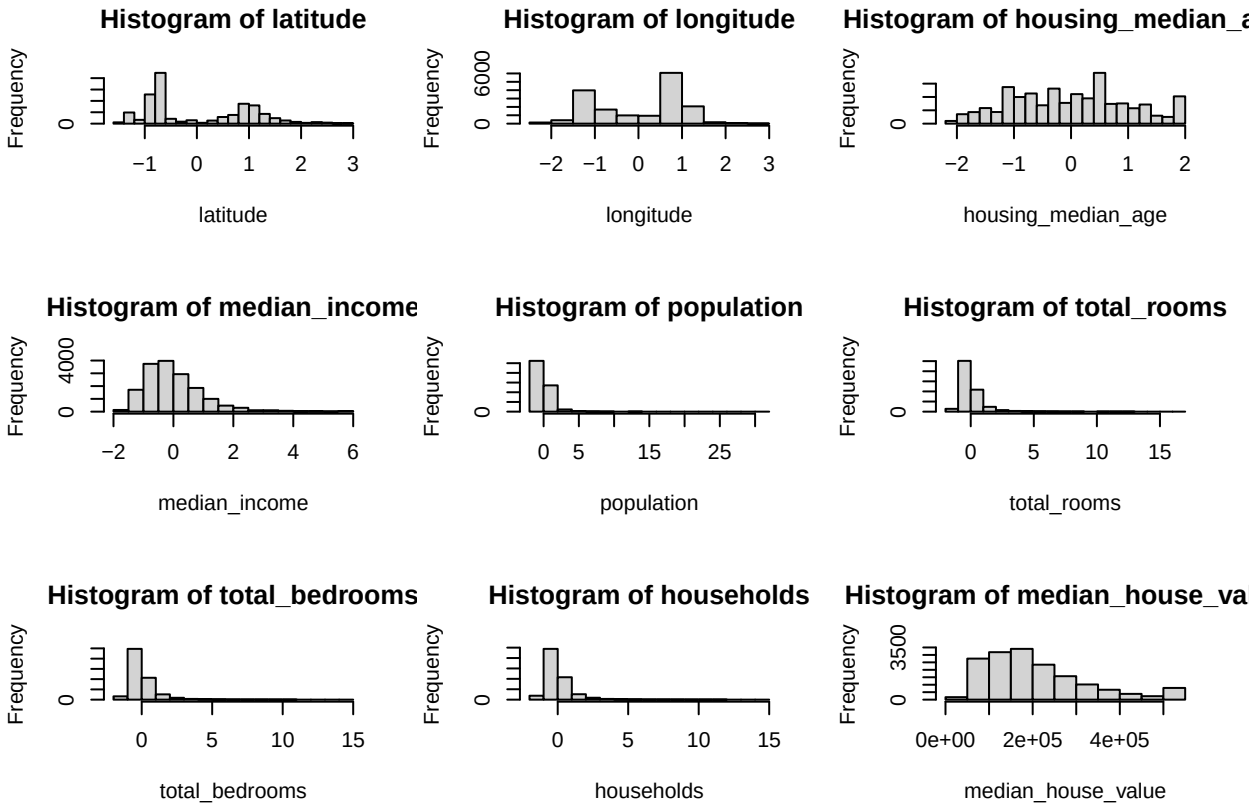
Meanwhile, for the second target, we need to know what distribution latitude and longitude follow. And to better understand them, we also construct statistics for longitude and latitude in terms of kurtosis and skewness.

Variable	Skewness	Kurtosis
Longitude	-0.288392	1.66327
Latitude	0.4614616	1.88424
House Median Price	9.76E-01	3.319279

Skewness and Kurtosis

Figure 1: avatar

To verify if the results are meaningful, we further plot the histogram of the above three variables.



The histograms prove that the test of *House_Median_Price* makes sense because it approximately follows a normal distribution, while for latitude and Longitude, quite the reverse is true. For a better exploration of the distribution they follow, we plot a scatter plot (x-longitude, y-latitude). And find that the houses are approximately located as two clusters close to (-1, 1) and (1, -1). The reason why it will perform like that is that the distribution they follow is Mixture Gaussian Distribution, which will be introduced later.

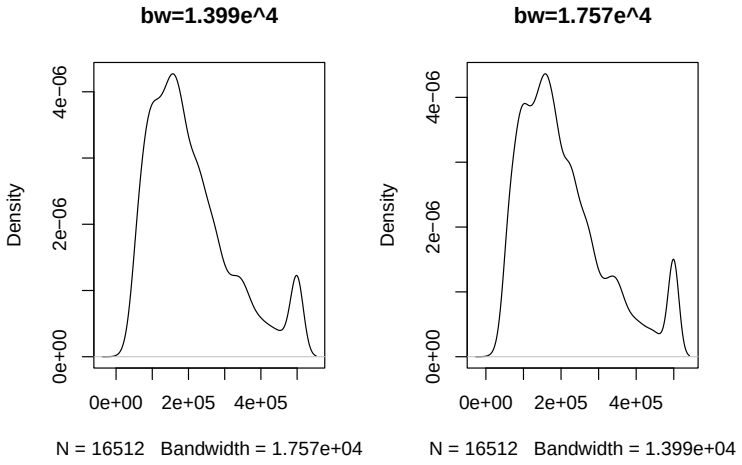
Introduction to kernel density estimate

Kernel Density Estimation (KDE) is used to estimate unknown density functions and is one of the nonparametric test methods. It uses a kernel function to fit the observed data and use it to simulate the population's probability density function.

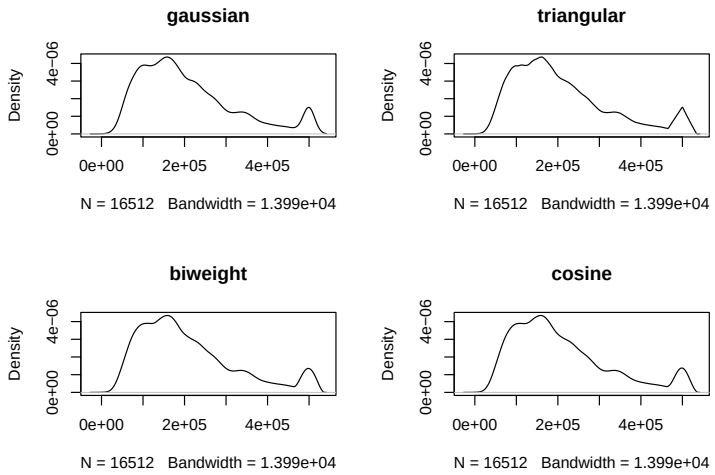
In the case that we know the probability distribution of a certain thing, if a certain number appears many times in the observation, we can think that the probability density of this number is very large, and the probability density of the number closer to this number will also be larger, and those numbers farther away from this number will have a lower probability density. The kernel function is to help us estimate the distribution probability for this property. Based on this idea, for each number in the observation, we can use the kernel function K to fit the probability density of the far small near large in our imagination. Take the average of multiple probability density distribution functions fitted to each observation. The fitted density function is:

$$\frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h})$$

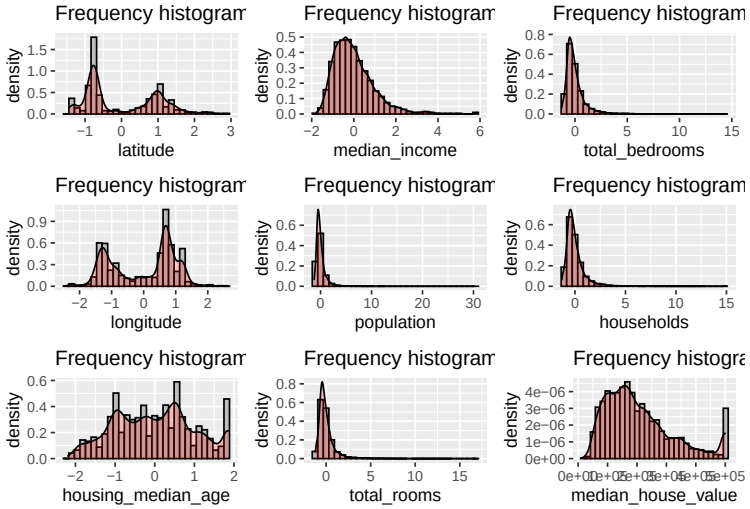
Where h is smoothing parameter. The density function drawn by different h has different degrees of smoothness. Below are plots of kernel density estimates for different bandwidths of house price data.



And also for the house price data, we select some different kernel function with the same bandwidth to do the KDE. As a result, the difference between them is not so obvious.



For each attribute, we draw the kernel density estimation diagram attch to the histogram. As follow:



Formula explanation Take the kernel function Gaussian distribution as an example. x makes a difference for each sample point (x_i), if x is closer to x_i, the smaller the difference, the larger the result of the kernel function. Conversely, if it is farther from the sample point, the result is closer to 0. Finally, average the output of the kernel function of all the sample points of this point, which means that the higher the kernel function result can affect the result of the store.

Core Concept Kernel density estimation uses the data and bandwidth of each data point as the parameters of the kernel function through the kernel function and obtains N kernel functions. Then linearly superimpose them to form the estimation function of the kernel density. After normalization, the result is the kernel density probability distribution.

MLE and Sample Method estimation of variables' mean and variance

To better glimpse the population means and variance of chosen variables, we perform maximum likelihood estimation and sample method estimation respectively.

Sample Method's explanation By using the data itself, it is possible to estimate changes in statistics calculated from that data. Calculates the error of a sample statistic in estimating a population statistic by simply using the sample data at hand without making any assumptions about the distribution of the population. There are some main steps:

1. Using repeated sampling techniques to draw a certain number of samples from the original sample with replacement.
2. Calculate the statistic T^* to be estimated based on the sample drawn.
3. Repeat the above N times to obtain N statistics T^*.
4. Obtain the original sample statistic T and N groups T^* of the bootstrap, and use how T^* changes around T to infer how T changes around the overall estimator.

MLE explanation Maximum likelihood estimation (MLE) is a method of estimating the parameters of an assumed probability distribution. For an arbitrary distribution, how well the sample data follow the distribution relies on the value of distribution parameters. What maximum likelihood estimation does is to find the parameters that have the greatest likelihood of a given observation.

Basic flow:

1. Writing the likelihood function: $L(\theta) = f(x_1, x_2, ..., x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$, which is the product of the probabilities that each observation fits the distribution.
2. Taking the logarithm of the likelihood function: $l(\theta) = ln(L(\theta)) = \sum_{i=1}^{n} log(f(x_i|\theta))$. In general, this step can convert the product in the expression into summation, which is more convenient for derivation later.
3. Taking derivatives correspond to parameters and make the result equal to 0 $\frac{\partial l}{\partial \theta} = 0$ to find the optimal value of θ when $l(\theta)$ is the largest.
4. Obtain the parameter estimators.

Deduction of normally distributed estimator:

1. $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
2. $L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = (2\pi\sigma)^{-\frac{n}{2}} e^{-\sum_{i=1}^{n}(\frac{(x-\mu)^2}{2\sigma^2})}$
3. $l(\mu, \sigma^2) = ln(L(\mu, \sigma^2)) = ln((2\pi\sigma)^{-\frac{n}{2}} e^{-\sum_{i=1}^{n}(\frac{(x-\mu)^2}{2\sigma^2})}) = -\frac{n}{2}(ln2\pi + ln\sigma^2) - \frac{1}{2\sigma^2}\sum_{1=2}^{n}(x_i - \mu)^2$
4. Finally, we can get:

$$\begin{cases} \frac{\partial l}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \\ \frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i - \mu)^2 = 0 \end{cases} \Rightarrow \begin{cases} \mu = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \\ \sigma = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \end{cases}$$

Deduction of exponential distribution estimator:

1. $f(x) = \lambda e^{-\lambda x}$
2. $L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$
3. $l(\lambda) = ln(L(\lambda)) = ln\lambda^n + lne^{-\lambda\sum_{i=1}^{n} x_i} = nln\lambda + \lambda\sum_{i=1}^{n} x_i$
4. $\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i \Rightarrow \lambda = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$

Expectation Maximization Algorithm (EM) explanation The EM algorithm is an iterative method to find maximum likelihood estimates of parameters in statistical models (pdf in our case), where the model depends on unobserved latent variables. Every iteration can be separated into E (expect) Step and M (maximum) Step. E step creates a function for log-likelihood expectation evaluated via current parameters' estimations. M step maximizing the expected log-likelihood and compute the parameters correspondingly. The intuition explanation in our experiment is to firstly initialize parameters corresponding to observations. And then estimate the missing data depending on the parameters calculated in the last step. Repeating this process until convergence. In our application scenario, we assume the pdf of Longitude and Latitude are Gaussian Mixed Model, based on the observation histogram, with initial parameter guess as follows.

$$p(x) = \sum_{k=1}^{K} \alpha_k N(x|\mu_k, \alpha_k)$$

Parameter	Explanation	Value
k	# of different normal distribution	2
α_i	mixture coefficient	0.5
σ_i	σ of ith normal distribution	random # form uniform distribution
μ_i	μ of ith normal distribution	random # form uniform distribution

Figure 2: avatar

Therefore, we have totally six parameters (each distribution three) to be estimated. And at beginning, the log-likelihood function should be calculated. Which is:

$$-\sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k, \sigma_k)}{\sum_j \pi_j M_j(x_n|\mu_j, \sigma_j)} \sigma_k^{-1}(x_n - \mu_k)$$

And let the expression equals to zero, we then are capable to obtain parameters.

$$\mu_k = \frac{1}{N_k} = \gamma_{nk} x_n, \sigma_k = \frac{\sum_{n=1}^{N} \gamma_{nk}(x_n - \mu_k)(x_n - \mu_k)^T}{N_k}, \alpha_k = \frac{N_k}{N}$$

$$\gamma_{nk} = \frac{\alpha_k N(x_n|\mu_n, \sigma_n)}{\sum_{j=1}^{K} \alpha_k N(x_n|\mu_j, \sigma_j)}, N_k = \sum_{n=1}^{N} \gamma_{nk}$$

Afterward, we can iteratively execute the above computation until meet the convergent conditions to obtain the result.

Monte Carlo Simulation

The Monte Carlo method is a computational algorithm, that relies on random sampling, and can be used for parameter estimation. The process is to first generate a Monte Carlo sample based on the distribution assumption. And then using the proper estimate method to calculate estimators. In our scenario, we assume the Median house value follows the Normal distribution while Longitude, Latitude follow the Gaussian Mixture distribution.

Besides, the estimate method we choose is still the sampling method and maximum likelihood method. Both these methods will be used for the estimation of mean and variance, in the meantime, the comparison between the two methods' results based on mean square errors under different scale samples will be conducted as well. Meanwhile, for the estimators, we construct the confidence intervals.

Experiment

1. Sample Method and Maximum Likelihood Estimation Bootstrap is used when the sampling method is implemented. We set the iteration times to be 2000, and the number of data selected each time to be n divided by 5 where n represents the number of total samples. For every variable, we execute the Bootstrap algorithm to obtain 2000 estimated parameters. And then take the mean of these estimated parameters as the result.

Attributes	Mean	Variance
Longitude	-0.06225	1.00128
Latitude	0.00522	1.00075
Median_house_value	207222.9	115602.7

sample method estimation

Figure 3: avatar

When performing Maximum likelihood estimation, the pre-calculated MLE estimators are involved. And the variable (Median_house_value) mean and variance can be obtained easily. Initially, we prefer to use exponential distribution as the assumed pdf. However, the results are unacceptable. Hence, we switch to a normal distribution, and this time, the estimators work well. For the two excluded variables, we've introduced the EM method for the estimation. The Assumption is that the Gaussian Mixture Model consists of two Gaussian distributions with initial parameters set as α of both distributions equal to 0.5, σ, μ equal to random number from a uniform distribution. In detail, in terms of algorithm, the maximum iteration is settled to be 200, and the convergent condition is that the subtraction of parameters of the last step and current step is smaller than the threshold (1e-5).

Attributes	Mean	Variance
Longitude (EM)	-0.06251	0.75555
Latitude (EM)	0.00528	1.08272
Median_house_value	207194.7	115619.1

maximum likelihood estimation

Figure 4: avatar

2. MC Stimulation For House Price, we set up two sample sizes (500 and 10000) and perform 1000 iterations. For each generated MC sample, both the sampling method (200 iterations) and the MLE method are used. And the result varies corresponding to a different scale. Since the distribution is assumed to be the normal distribution, the MC generation can be conducted with the assistance of the R-provide function (rnorm).

For Longitude and Latitude, we need to write our function because R does not provide the random function for Mixture Gaussian distribution. Instead of using the random guessing parameters, we used those calculated

House Value	Mean	Sd
Standard	207194.7	115622.6
Sample Method	204323.9	115406.5
MLE	208104	117439.7
MSE	27334667	24911221

Sample Size:500

House Value	Mean	Sd
Standard	207194.7	115622.6
Sample Method	207971.3	115800.1
MLE	207371.2	115131.1
MSE	1093212	1016847

Sample Size:10000

Figure 5: avatar

in the previous step as initialization. Besides, for the variables that follow Mixture Gaussian distribution, MLE cannot solve the estimation directly, hence EM method will be conducted. And the results are as follows:

Latitude	Mean	Sd
Standard	0.005279	1.000367
Sample Method	-0.03139	1.033477
MLE	0.005407	1.055481
MSE	0.003368	0.002926

Sample Size:500

Latitude	Mean	Sd
Standard	0.005279	1.000367
Sample Method	-0.003656	0.997969
MLE	0.006439	1.07578
MSE	0.000185	0.006789

Sample Size:10000

Longitude	Mean	Sd
Standard	-0.006251	1.001154
Sample Method	0.023002	1.019586
MLE	1.58E-05	0.750576
MSE	0.002195	0.073074

Sample Size:500

Longitude	Mean	Sd
Standard	-0.006251	1.001154
Sample Method	-0.002232	0.99928
MLE	-0.006009	0.755711
MSE	1.21E-04	5.94E-02

Sample Size:10000

Figure 6: avatar

Modeling

1. Best Model Selection To select the best model, we use the Jackknife (leave one out) method. When verifying the model fitting ability, one sample is left for cross-validation, and the rest of the data is used to fit the model. Then the average error of each validation is used to evaluate the model fitting ability.

Basic flow: 1. Use a stepwise way to remove multicollinearity between variables, but we find the data does not have attributes that need to be eliminated.

2. Model assumptions. Model1 is the exponential model; Model2 to Model6 use the best fit line in the scatter plot of each attribute and house price as a reference and try to use grid search to filter out some bad model. In addition, there is a skill can be used that plot the model diagnostics graph before Jackknife test. Finally, we assume six models.

3. Use the LM function to fit the data and leave one sample for verification each time to calculate the average prediction error. Model1 and Model2 are linear models and exponential models. Model3 and model4 use the best fit line in the scatter plot of each attribute and house price as a reference. For the data, the relationship between households, total bedrooms and total bedrooms and y is not a straight line but a curve like a quadratic function, we set these items to the second power term.

4. As a result of the last step, the minimum error of mean squared error is still a bit high. Thus, we can try to perform feature engineering on the data by principal component analysis (PCA) to reduce the dimension of the data and select the first 6 principal components that can represent the data. We try to use 6 new pc attributes to fit the model.

Then we do the model assumption again. Model1 is the exponential model; Model2 to Model6 use the best fit line in the scatter plot of each PC attribute and house price as a reference and try to use grid search to filter out some bad model. In addition, there is a skill can be used that plot the model diagnostics graph before Jackknife test. Here we use jackknife method to select model again and we find the model6 has the smallest prediction error. The model is:

$$log(y) = PC1 + log(PC2) + log(PC3) + log(PC4) + PC5^2 + log(PC6) + b$$

Model	Error	Select
Model1	0.43156	×
Model2	0.90651	×
Model3	0.87827	×
Model4	0.31671	×
Model5	0.70014	×
Model6	0.22991	✓

Figure 7: avatar

5. The final step is model diagnosing, we can plot the model and observe the fitting effect. Most importantly, we can find the outliers and extreme points which can influence the model fitting. Then just remove them to improve the model.

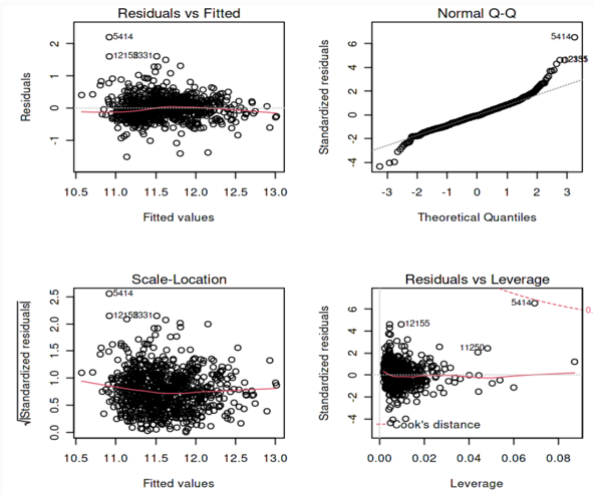


Figure 8: avatar

2. Estimate Bias and Standard Error In this model, we have 8 estimators. For each estimator, we use the bootstrap method to estimate the coefficient. Base on the ordinary parameter and estimated value, we can get the bias and standard error.

Coefficient	Bias	Standard Error
Intercept	1.36E-03	0.0322893
PC1	-1.19E-04	0.0071357
log(PC2)	-3.98E-05	0.0170491
log(PC3)	-9.45E-04	0.0136392
log(PC4)	4.38E-04	0.0109445
$PC5^2$	-9.20E-04	0.0312148
log(PC6)	5.46E-04	0.0112096

Figure 9: avatar

Conclusion

In this project, we clearly describe and explain the concept of some estimation methods. And we perform data visualization, parameters estimation, and distribution estimation for the sample data to analyze it from various angles. Meanwhile, we also fit a model for house price prediction as well as figure out the house distribution in terms of latitude and longitude.