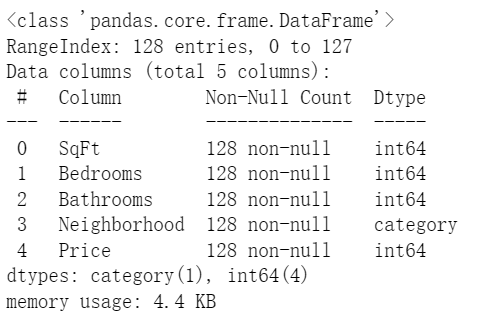
**2 Programming Report**

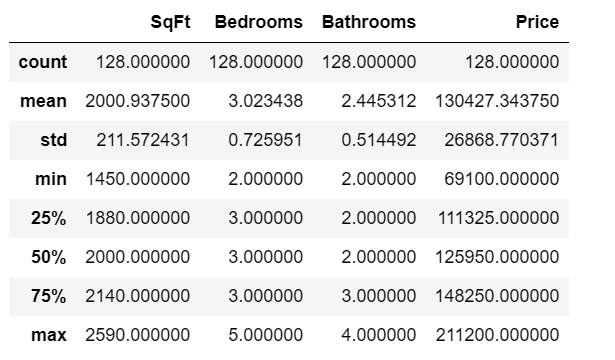
**Problem 1**

Description: We want to use appropriate attributes in “SqFt, Bedrooms, Bathrooms, Neighborhood” to predict

the attributes “Price”.

Step1: use panda to load the csv file.

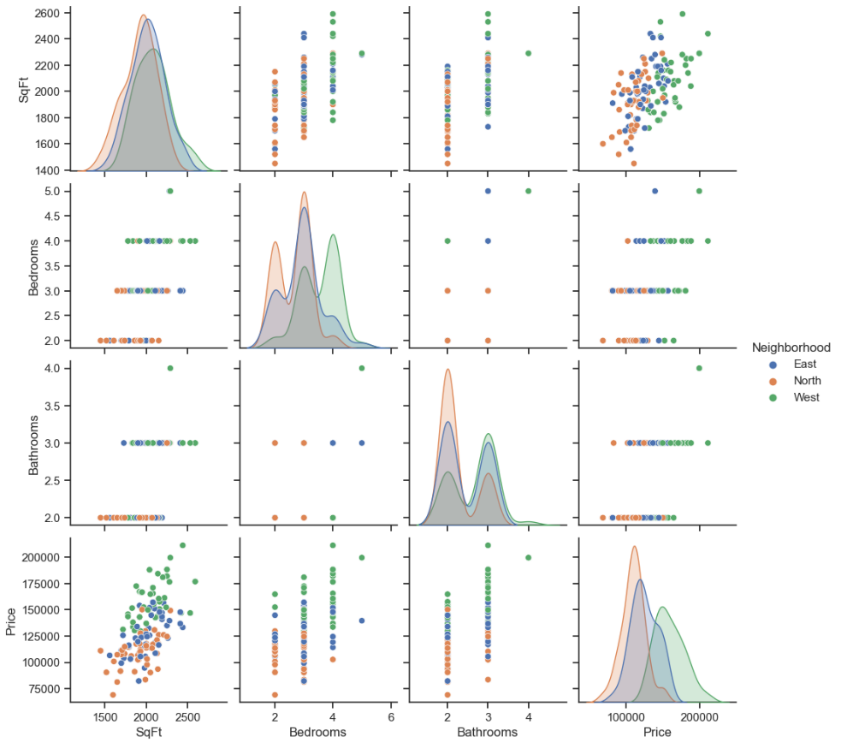




there are 128 data samples and 3 feature column, and I guess SqFt is the most important feature contributing to Pirce according to my experience. And also neighborhood is transformed to category to help machine learn it.

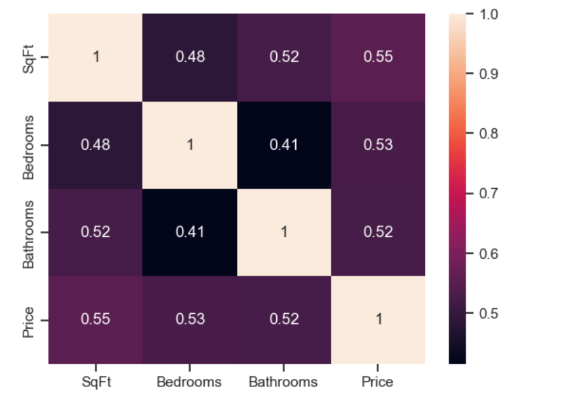
Step2: use seaborn library to visualize the dataset.

pairplot：



Just focusing on the last column, SqFt, Bedrooms, Bathrooms and Neighborhood are all correlated highly to Price. And as SqFt, Bedrooms, Bathrooms increases, Price increases. At last, the preference of neighborhood is West > East > North.

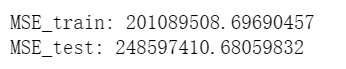
Heatmap:



From the last colum of the heatmap, it is clear that SqFt, Bedrooms, Bathrooms is of almost the same correlation.

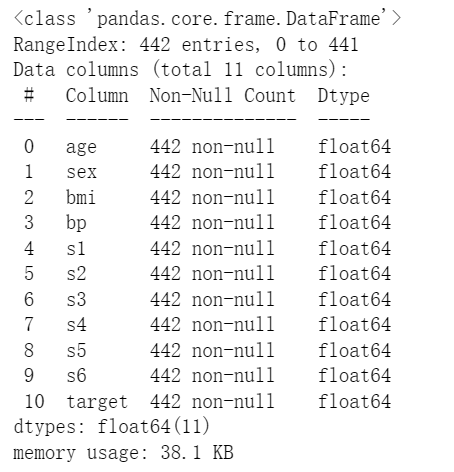
Step 3: Use sklearn library to process the category variable and use train\_test\_split in sklearn.model selection to randomly split the data.

Step4:



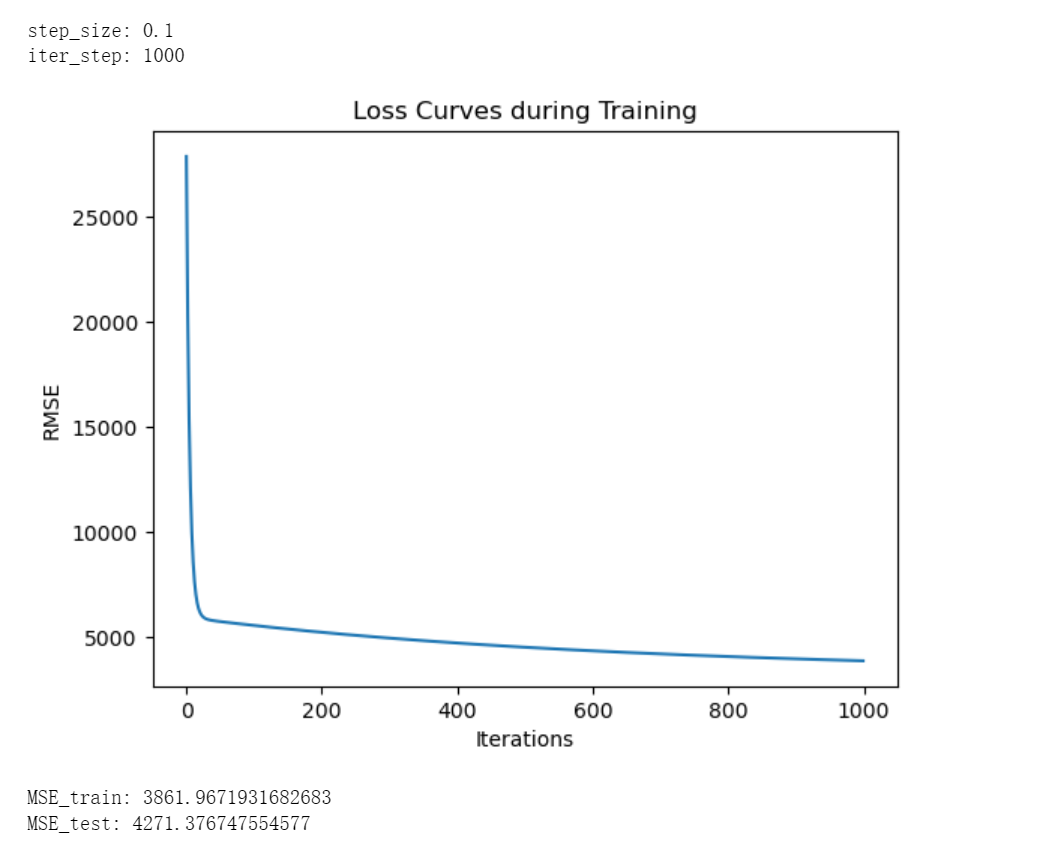
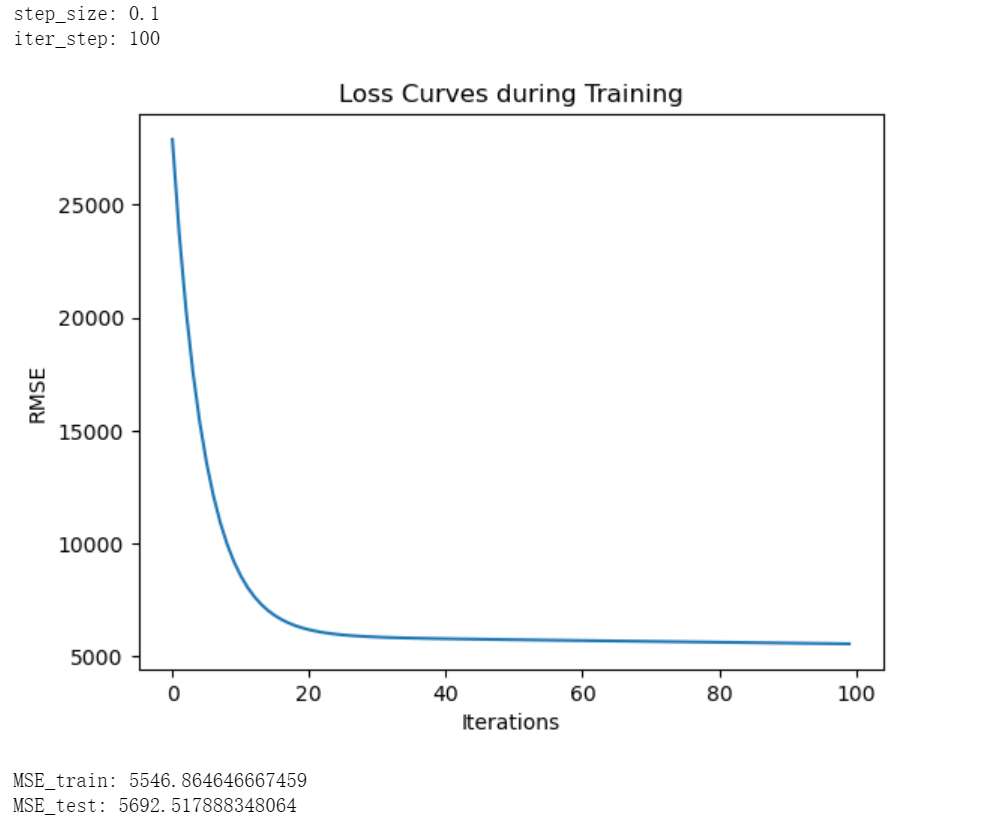
**Problem 2**

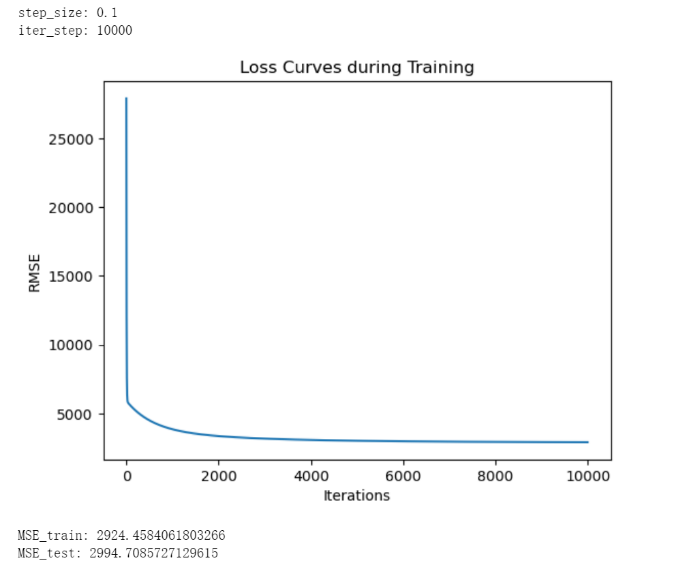
Description: Use gradient descent to train a linear regression model



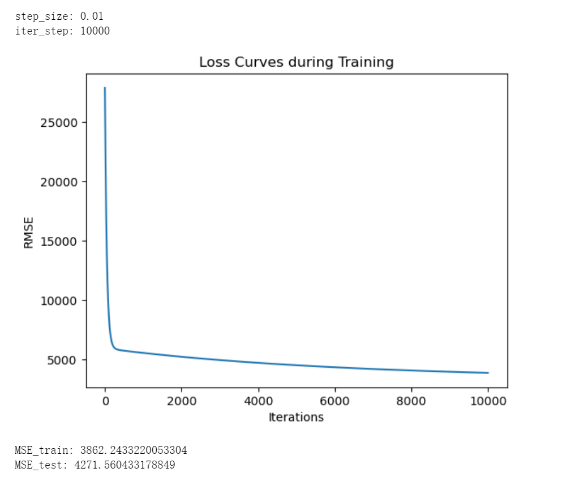
There are 442 samples and 10 features to predict target.

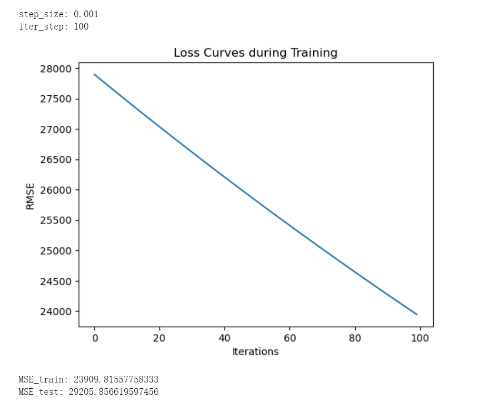
**Stopping criterion**: I choose when the iteration times is reached because it helps to find the relationship between loss and iteration.

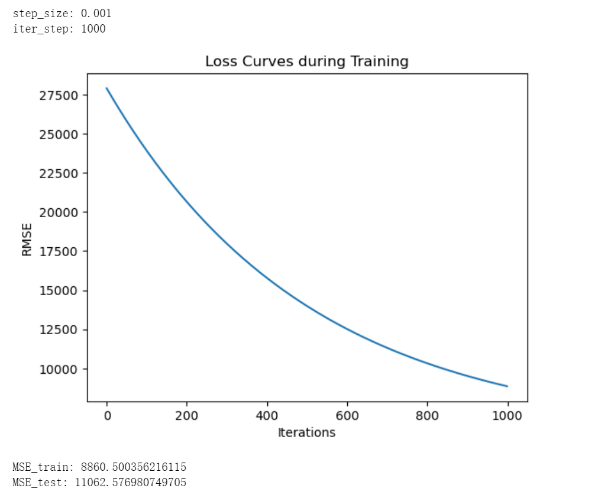
****

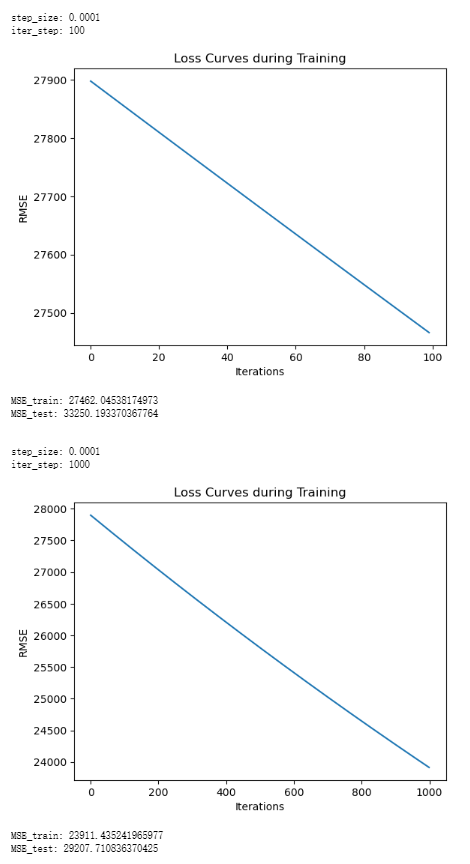
****

****

****

****

****

****

**Discovery:** when step size is relatively large, the loss decreases sharply. When step size is relatively small, it is hard to reach an converge but the accuracy is better. When the number of iteration is large, it actually waste some time since it does not contribute much to the decrease of loss. When the number of iteration is small, it may diverge.