A53271697

Bosi Cheng

1. from 1.0 to 5.0: (judged by if, elif, else)

[211, 343, 1099, 1624, 4137, 8797, 16575, 12883, 4331]

2.



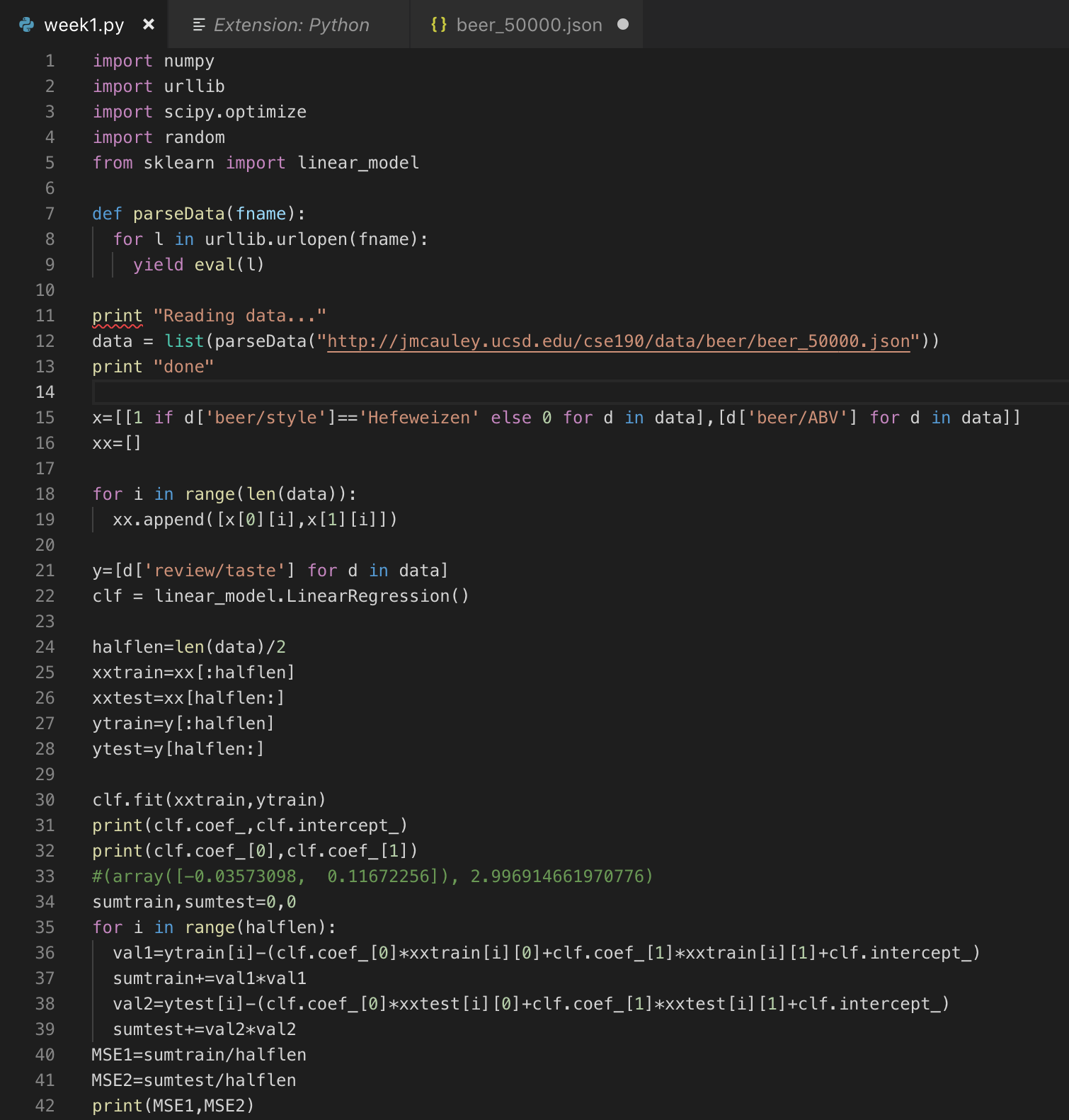
linear regression coefficient(θ1 and θ2) and its intercept (θ0):

(array([-0.05637406, 0.10877902]), 3.1179508424739546)

3.

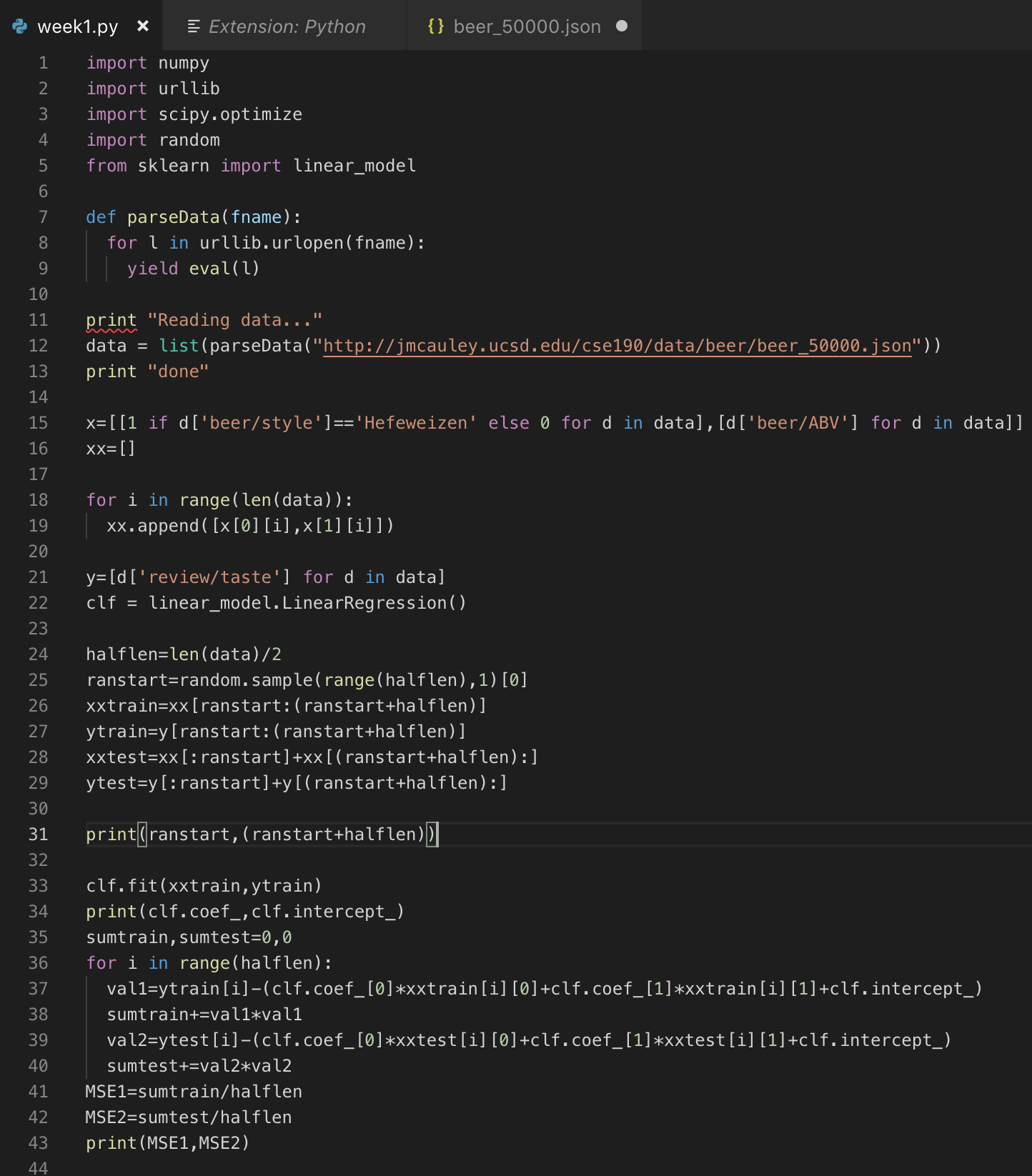
linear regression coefficient and its intercept:

(array([-0.03573098, 0.11672256]), 2.996914661970776)



MSE of training and MSE of testing:(0.4839680560133444, 0.4237065211985007)

4.



start and end of the training part(14487, 39487)

linear regression coefficient and its intercept:

(array([-1.23361769, 0.09547546]), 3.3326207702584405)

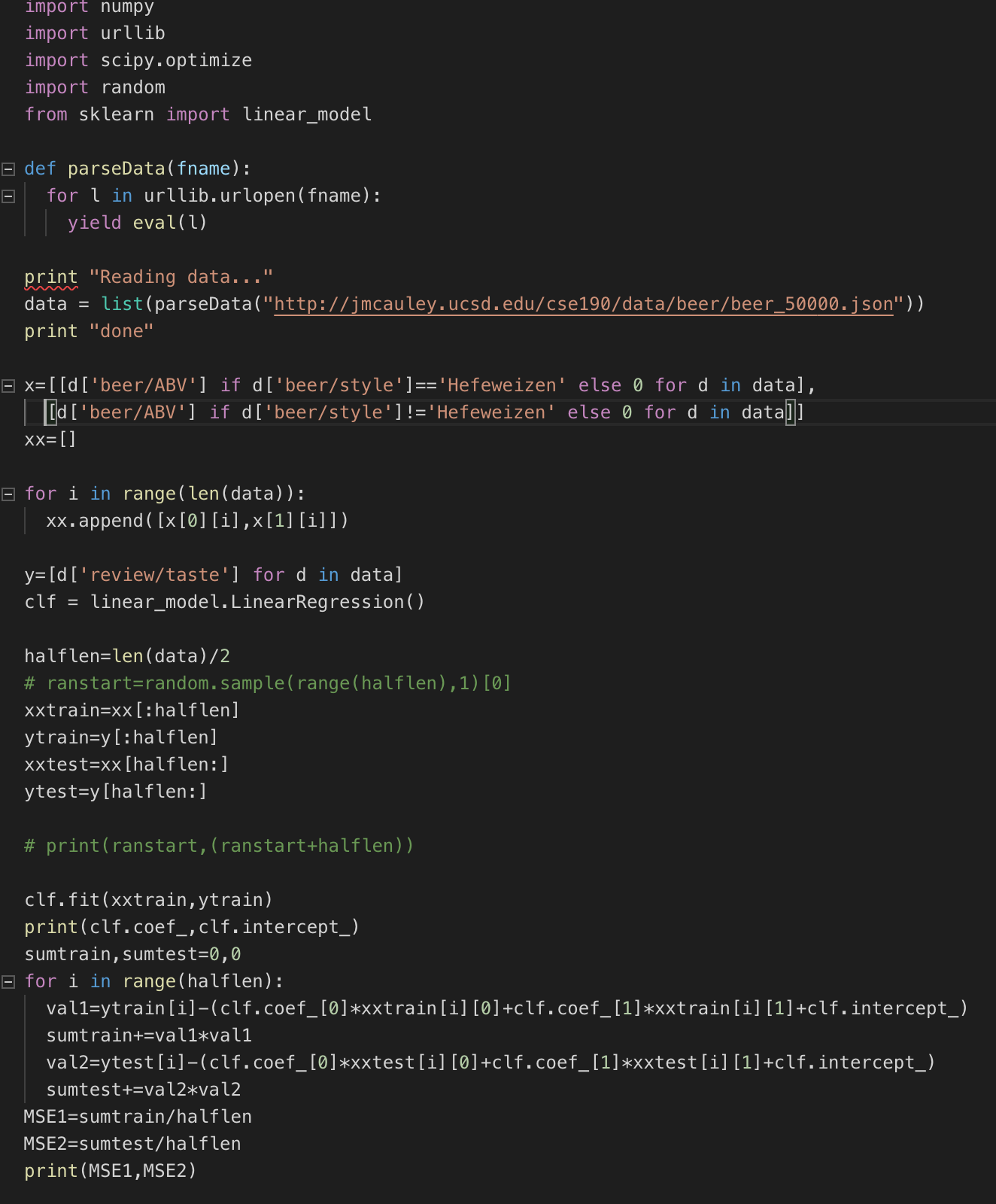
MSE of training and MSE of testing:

(0.3595337172103225, 0.5945455303770442)

One possible reason:

Around the time the training part's ending the beer producer proposed some strategies to help the following selling situation (the test part) more organized or correlated to some certain variables. However, the random division of the training and testing is unorganized or totally accident.

5.



linear regression coefficient and its intercept:

(array([0.11005028, 0.1167231 ]), 2.9969206914347235)

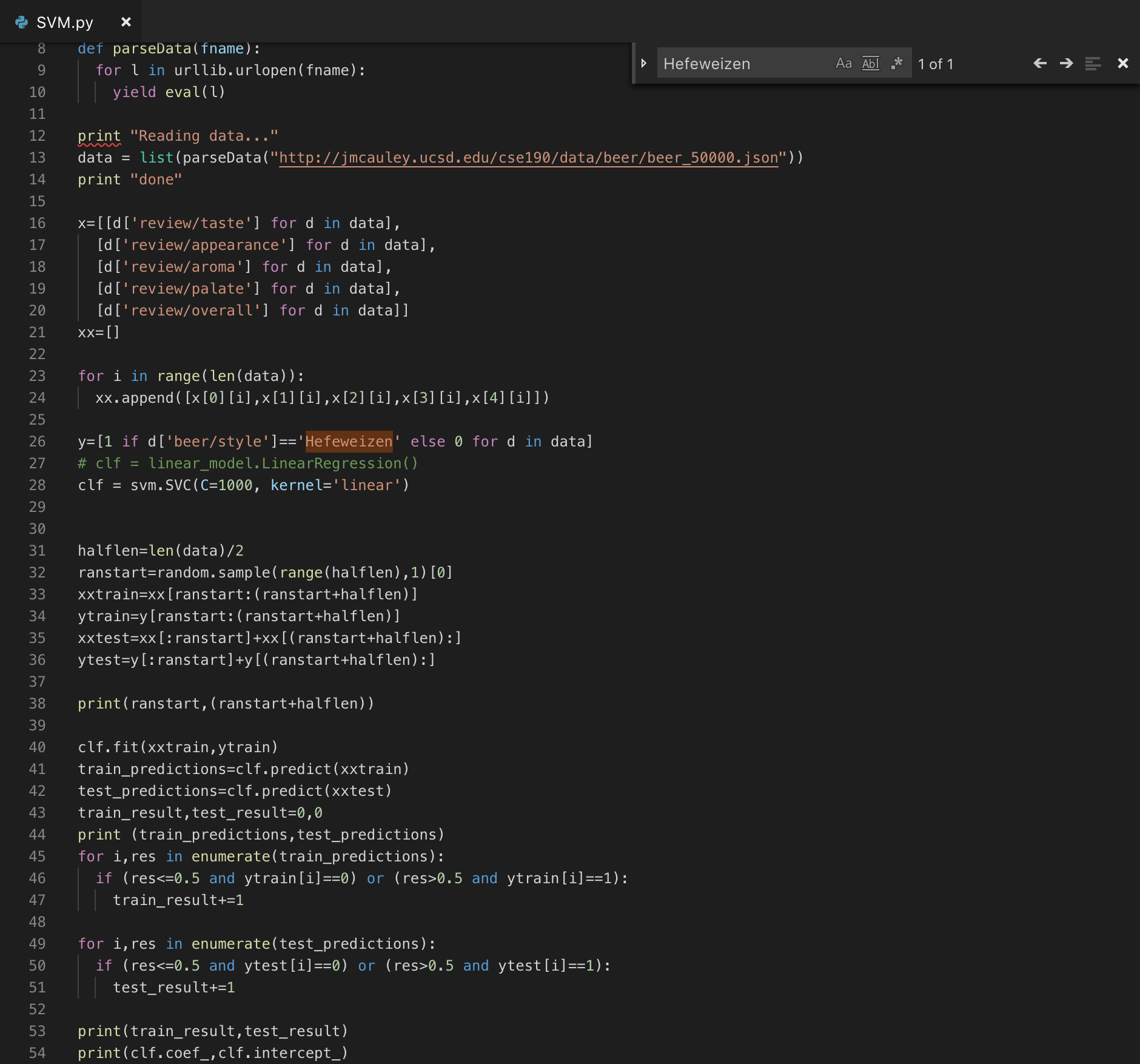
MSE of training and MSE of testing:

(0.4839671462417231, 0.42369661406846515)

6.

Although The model from Question 5 uses the same two features as the model from Questions 2-4 and has the same dimensionality, they are not identical. Moreover, the two variables in Question 5 is more like "yes or no", and the extent it impacts the 'review/taste' is more powerful.

7.

**

start and end of the training part (1054, 26054)

linear regression coefficient and its intercept:

(array([[ 6.31252769e-05, 5.16767614e-06, 3.27652087e-05,

4.49432991e-05, -1.99214555e-05]]), array([-1.0004203]))

the number of correct predictions and its percentage of training and testing data:

(24699, 24683) 98.80% 98.73%

8.

My opinion:

The result in the last case is not accurate simply because the percentage of 'Hefeweizen' is so low that the percentage of correct classification seems high. It is not an appropriate case in which the variable 'Hefeweizen' is so weak. Adapting 'linear' as the kernel of svm.SVC is usually in less-variable and fewer-data situation. Besides, the feature 'review/overall' is meaningless because it's somewhat based on other four features.

So when changing the kernel to 'rbf' and deleting the feature 'review/overall', the number of correct predictions and its percentage of training and testing data come to:

(24708, 24675) 98.83% 98.7%



But after running it I would rather 'linear' kernel because the 'rbf' takes much longer time to finish.