AI_534
MACHINE LEARNING

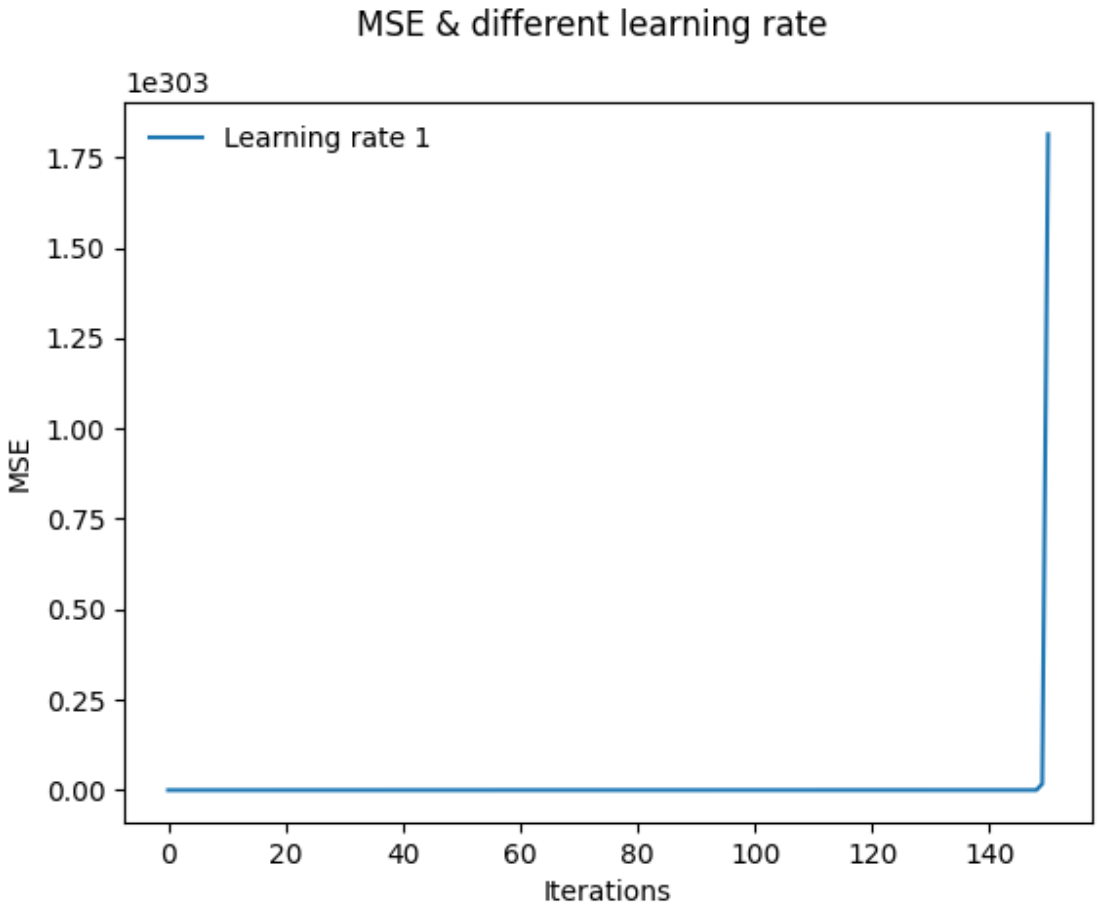
IA1 Report

CHI-CHIEH WENG
TSU-CHING LIN

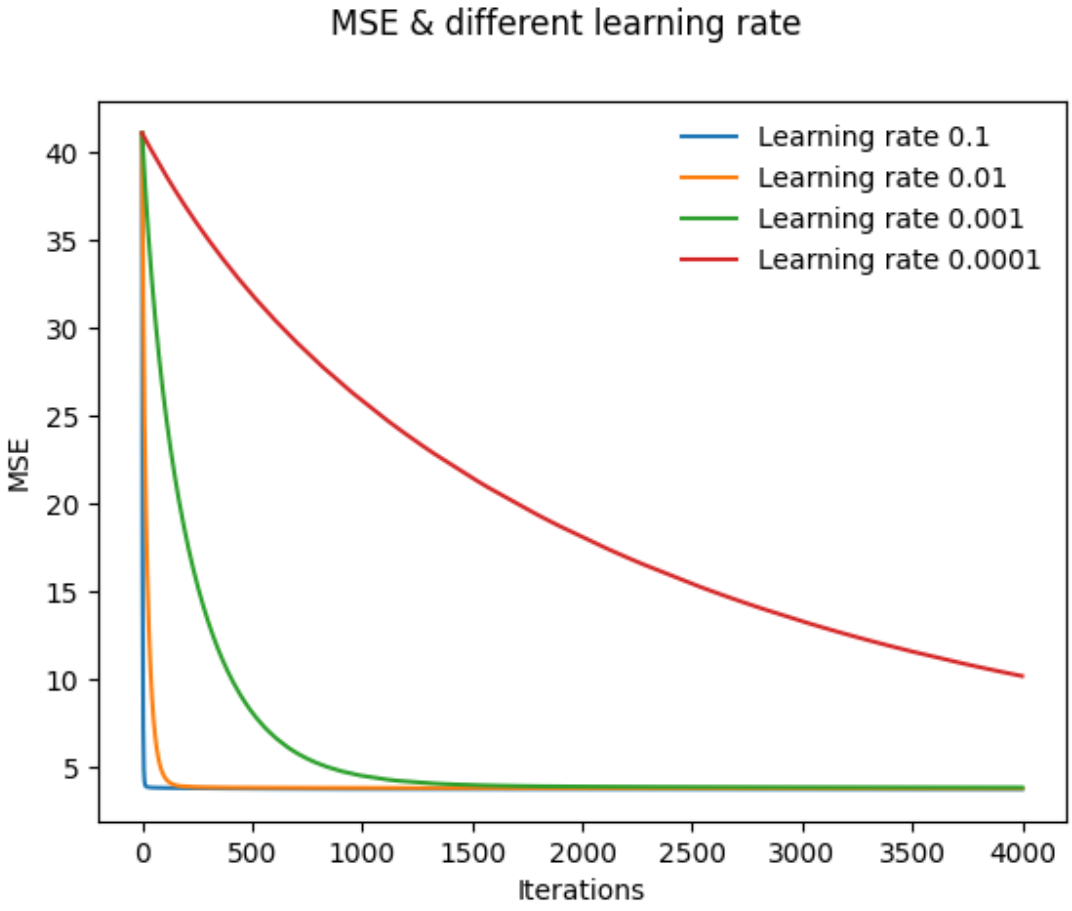
Part 1

(a)

Learning rate 10^0



Learning rate $10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$



We believe the better learning rate for the data set will be 0.1.
When we use 1 as our data set learning rate which cause gradient descent to diverge.

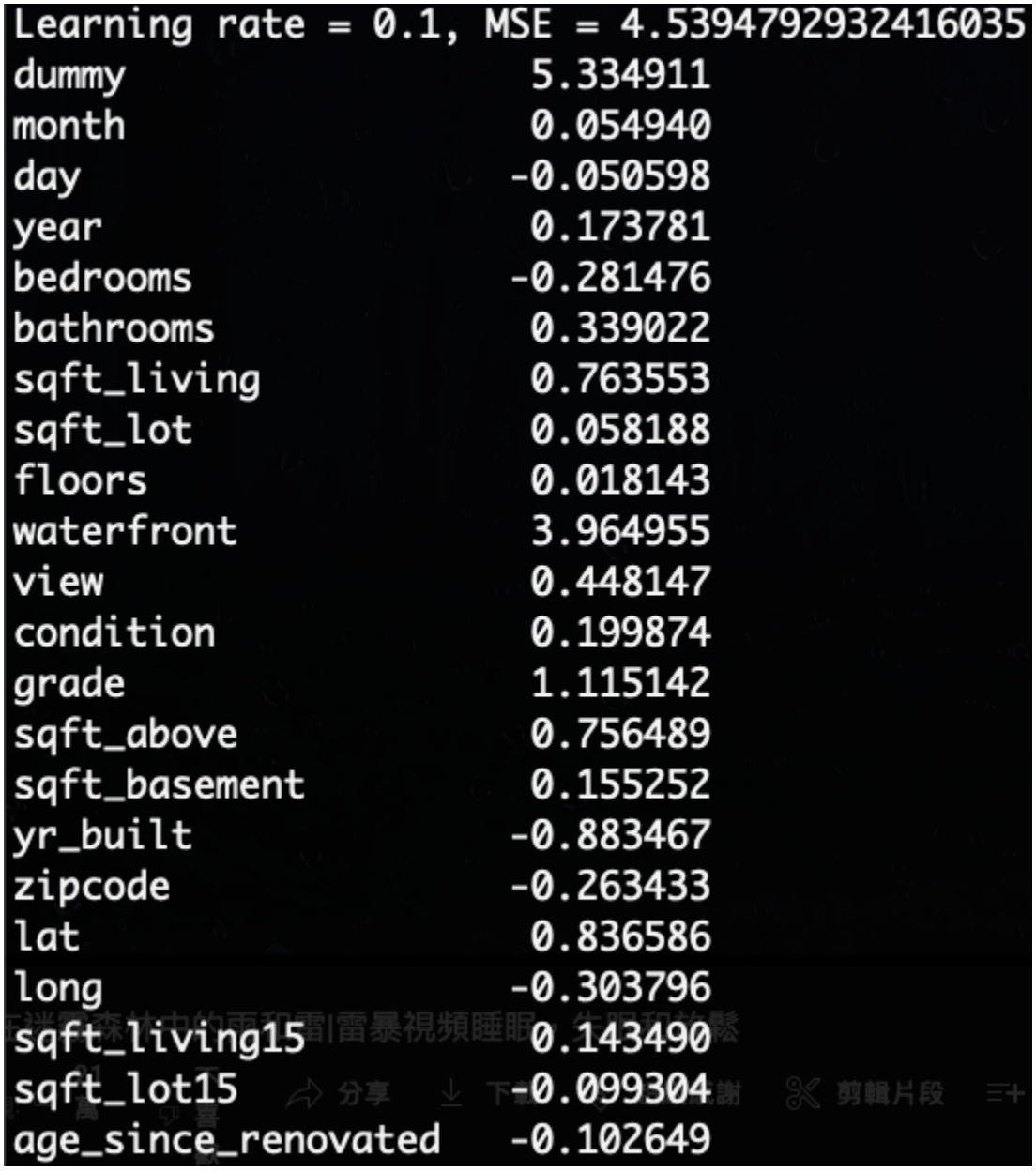
(b)

```
Learning rate=1, compute_val_data_MSE=inf
Learning rate=0.1, compute_val_data_MSE=4.5394792932416035
Learning rate=0.01, compute_val_data_MSE=4.691057732929848
Learning rate=0.001, compute_val_data_MSE=4.8108446020504365
Learning rate=0.0001, compute_val_data_MSE=11.772283565262052
```

The best learning rate for validation MSE in our experience is 0.1.

For different convergent learning rates, we would first choose the low-loss one, and if they're all similar, then we choose the bigger one. That's because the bigger learning means a simpler function.

(c)

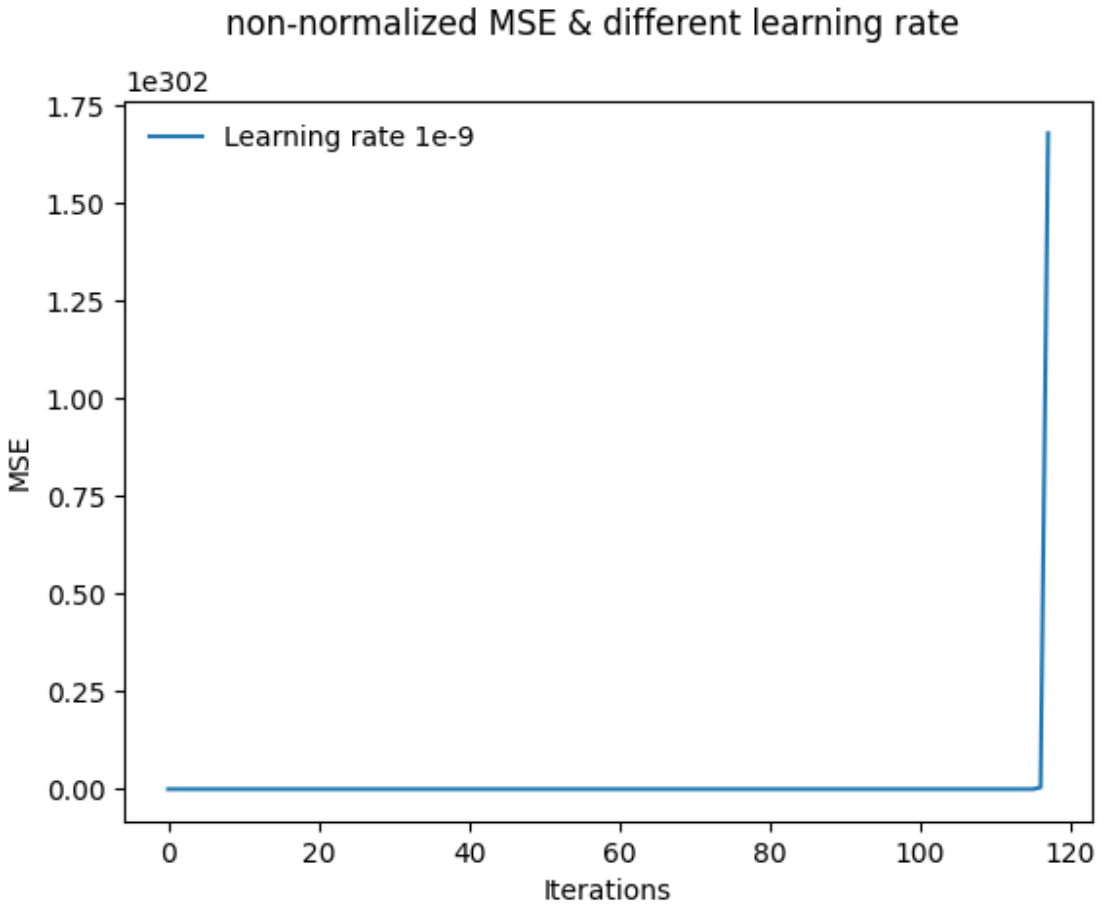


The most important attribute is grade(excluding dummy and waterfront).

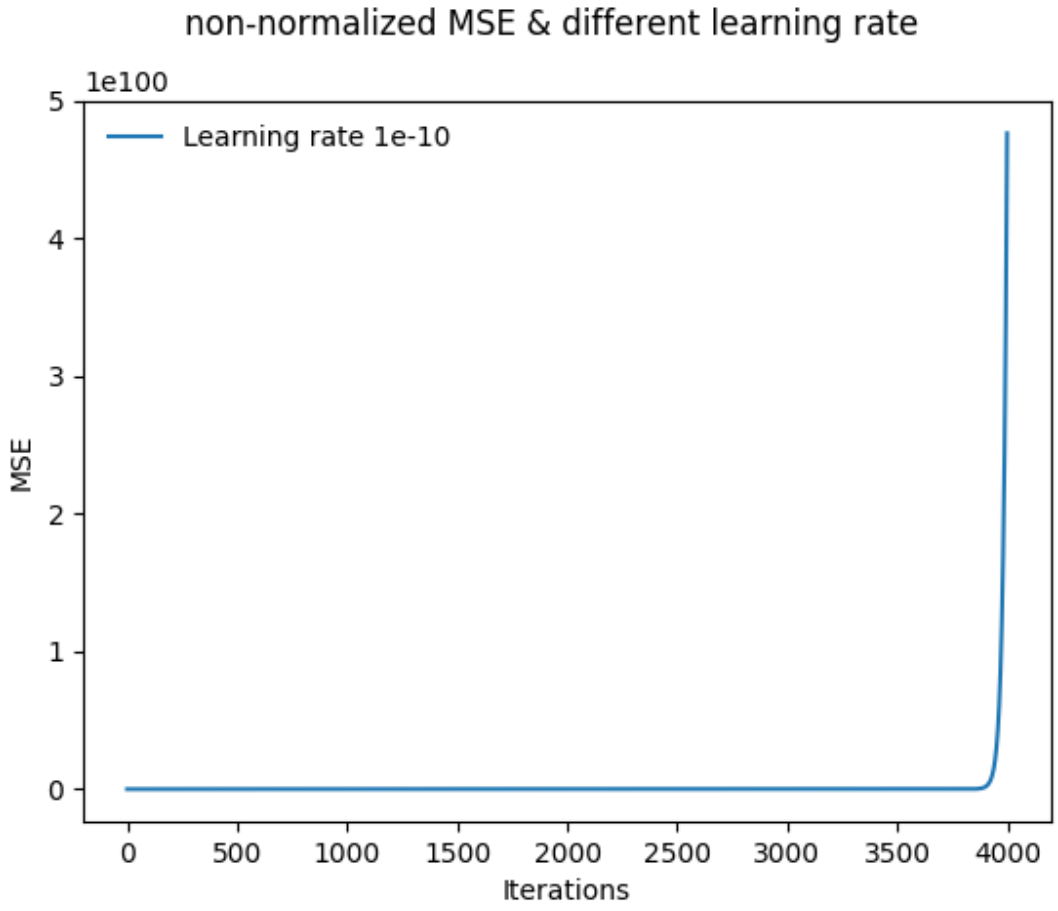
Part 2a

(a)

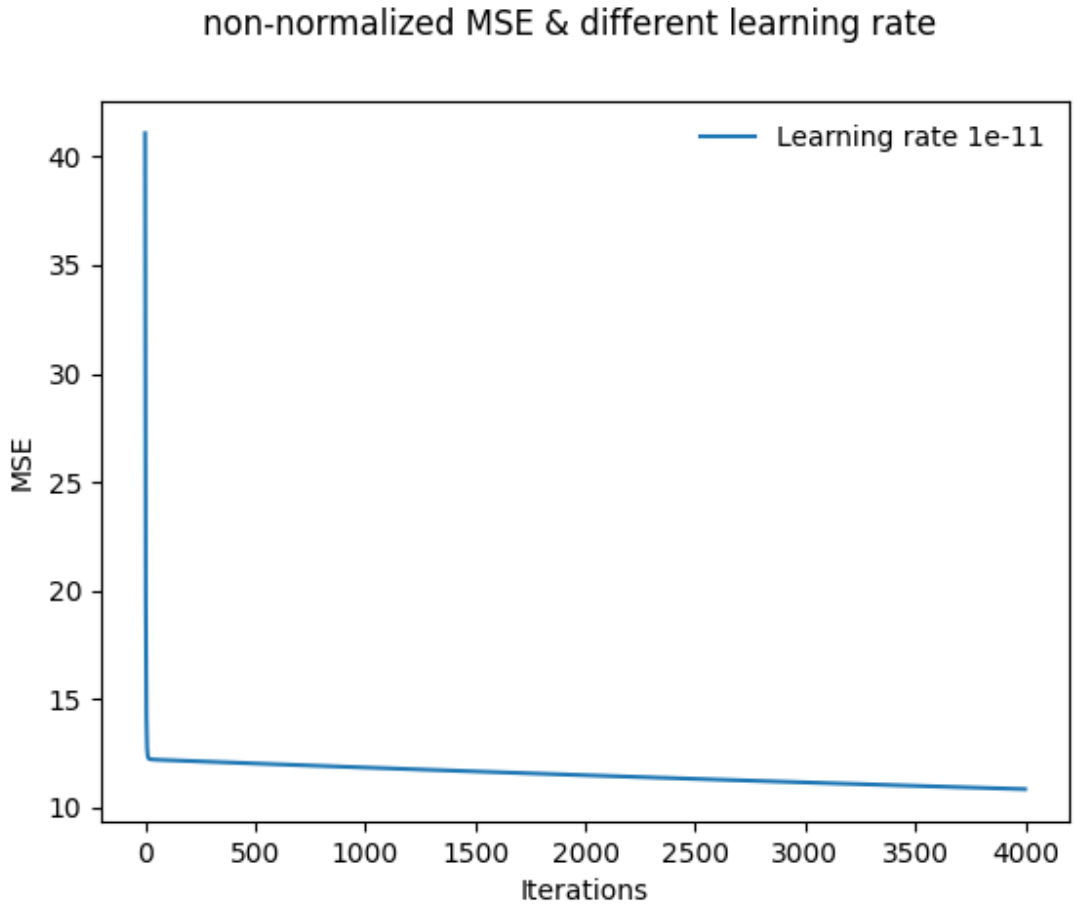
Learning rate 10^{-9}



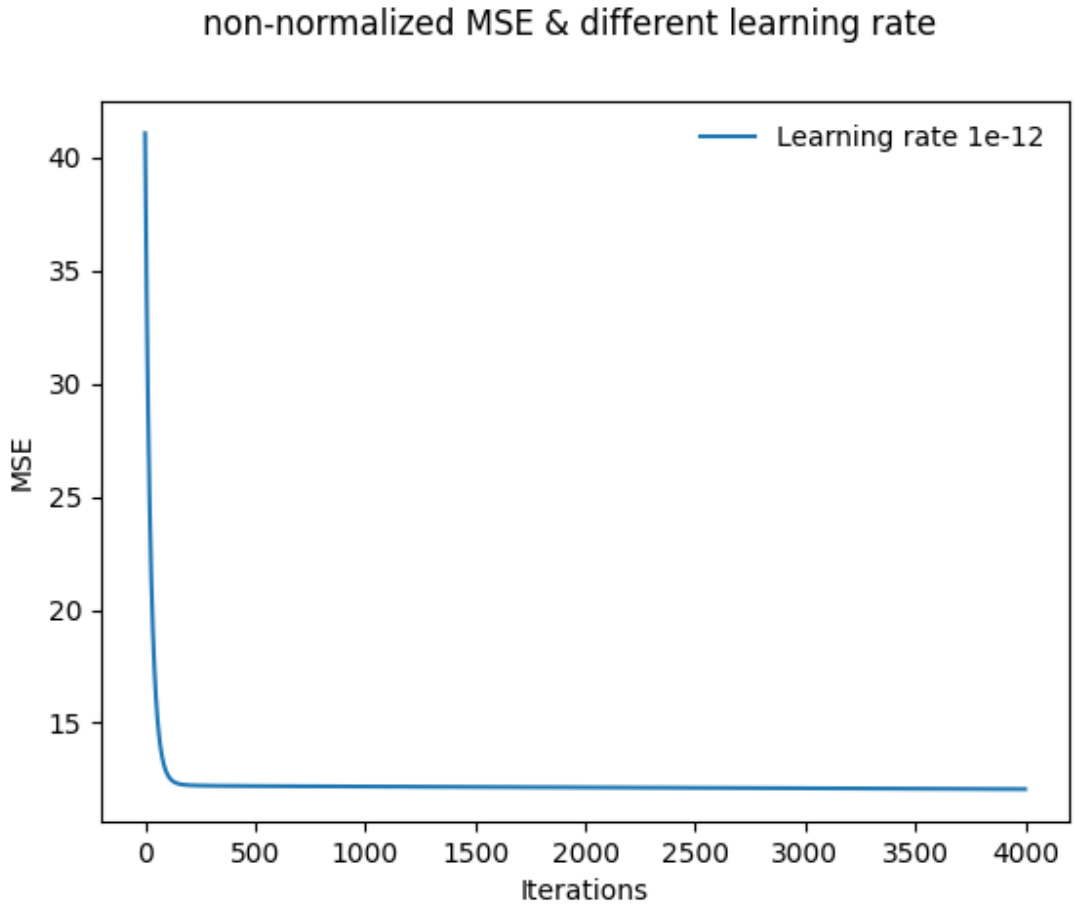
Learning rate 10^{-10}



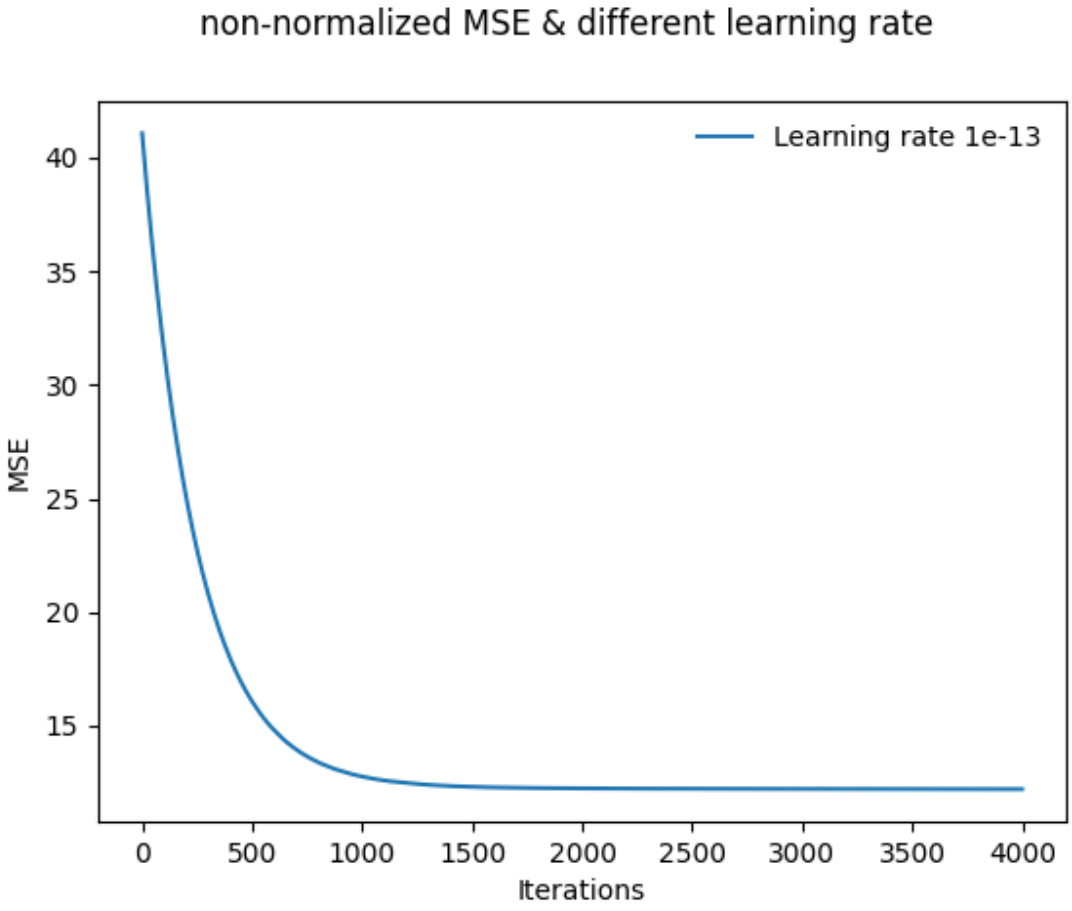
Learning rate 10^{-11}



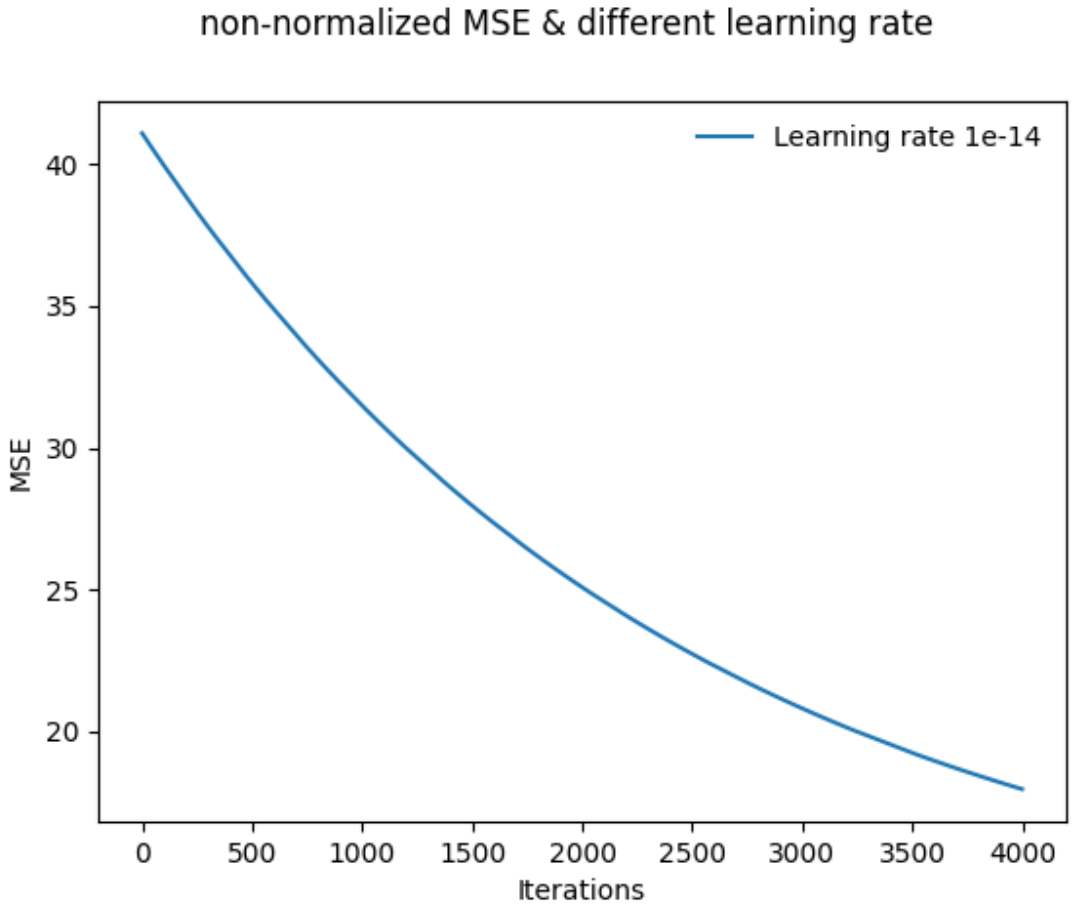
Learning rate 10^{-12}



Learning rate 10^{-13}



Learning rate 10^{-14}



(b)

The best learning rate for validation MSE in our experience is 10^{-10}.

```
Learning rate=1e-09, compute_val_data_MSE=6.2237123462057 31e+304
Learning rate=1e-10, compute_val_data_MSE=4.7456856568459 18e+100
Learning rate=1e-11, compute_val_data_MSE=12.871130796848968
Learning rate=1e-12, compute_val_data_MSE=14.178794532426927
Learning rate=1e-13, compute_val_data_MSE=14.31195181643134
Learning rate=1e-14, compute_val_data_MSE=20.660566947697564
```

(c)

month	1.409409e+41
day	3.380909e+41
year	4.300949e+43
bedrooms	7.211475e+40
bathrooms	4.544260e+40
sqft_living	4.479596e+43
sqft_lot	4.274546e+44
floors	3.203345e+40
waterfront	1.511642e+38
view	5.044121e+39
condition	7.292627e+40
grade	1.639010e+41
sqft_above	3.860749e+43
sqft_basement	6.188467e+42
yr_built	4.208938e+43
zipcode	2.094145e+45
lat	1.015472e+42
long	-2.609407e+42
sqft_living15	4.271916e+43
sqft_lot15	3.421090e+44
age_since_renovated	8.714477e+41

1. The dataset is not standardized, so large numbers of attributes can be complicated.
2. It would be easier to normalize the data. Because the attribute complexity is low.
3. The results of each weight without normalized data are higher than the results of each weight of 1c, so we believe that this phenomenon may be explained by an imbalance in the magnitude of the different features. Also, we believe that during the training period, the weight of the large feature must be very small to counteract the

dominance of the large value and try to balance the contribution of the small value, so that the large value contributes more to Δw than the small value.

Part 2b

We tried learning rate of i equal one to four, and when i equal to one gives us smaller MSE on validation set (10^{-i}). Hence we choice of learning rate 1e-1 as our learning rate, and the MSE is 4.372.

```
itr=4000, lr=0.1, loss=4.372
dummy                  5.425434
month                  0.036456
day                   -0.048783
year                   0.187219
bedrooms              -0.220782
bathrooms              0.399568
sqft_living            0.716019
sqft_lot               0.077090
floors                 0.023838
waterfront             6.795080
view                   0.442931
condition              0.207406
grade                  1.397781
sqft_above             0.632915
sqft_basement          0.291381
yr_built              -0.788699
zipcode               -0.341916
lat                    0.865500
long                  -0.347157
sqft_lot15            -0.089167
age_since_renovated   -0.124216
dtype: float64
```

Compare to 1(c), it looks like it doesn't have a lot of change of sqft_living when we drop spft_living15. Thus, considering the situation, when two features x1 and x2 are redundant, we expect the weights (w1 and w2) learned by both features would be very similar, so they very small change when we drop the redundant one. Although the redundant features are very similar to each other, they are still slightly different, so we believe that the more repetitions of the training, the better the performance should be.