CS6140 Project 1

Yihan Xu
NUID: 001566238

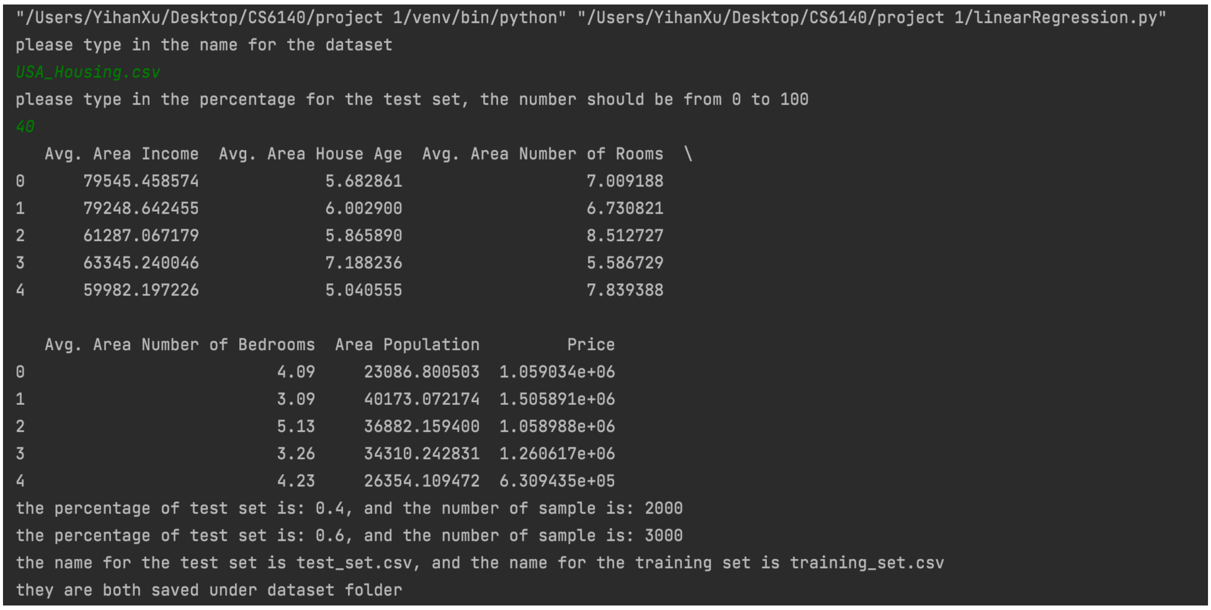
Overview:

In this project, I grabbed a dataset that contains the USA housing data, with 5 independent variables: avg. area income, area population, avg. area number of rooms, avg. area number of bedrooms, avg. house age. and 1 dependent variable: pricing, with 5000 rows. The file is saved in dataset folder named USA_Housing.csv

The models that are built include simple pair linear regression, multiple linear regression, ridge regression, 3rd order polynomial linear regression. This project also implements PCA Algorithm and tests the result on the projected data of PCA.

Dataset spliting:

I designed a function that takes in the name of the file and the percentage for the testing set, and randomly divides the dataset into a testing set and a training set based on the percentage provided, and saves them in the dataset folder. The parameter comes from the command line, here's a screenshot of the output:



```
"/Users/YihanXu/Desktop/CS6140/project 1/venv/bin/python" "/Users/YihanXu/Desktop/CS6140/project 1/linearRegression.py"
please type in the name for the dataset
USA_Housing.csv
please type in the percentage for the test set, the number should be from 0 to 100
40
   Avg. Area Income  Avg. Area House Age  Avg. Area Number of Rooms  \
0      79545.458574             5.682861                   7.009188
1      79248.642455             6.002900                   6.730821
2      61287.067179             5.865890                   8.512727
3      63345.240046             7.188236                   5.586729
4      59982.197226             5.040555                   7.839388

   Avg. Area Number of Bedrooms  Area Population         Price
0                          4.09     23086.800503  1.059034e+06
1                          3.09     40173.072174  1.505891e+06
2                          5.13     36882.159400  1.058988e+06
3                          3.26     34310.242831  1.260617e+06
4                          4.23     26354.109472  6.309435e+05
the percentage of test set is: 0.4, and the number of sample is: 2000
the percentage of test set is: 0.6, and the number of sample is: 3000
the name for the test set is test_set.csv, and the name for the training set is training_set.csv
they are both saved under dataset folder
```

Data plotting:

I also analyzed the dataset by plotting them in pairs with one dependent variable and one independent variable, here are the plots:

avg. area income vs. price

price

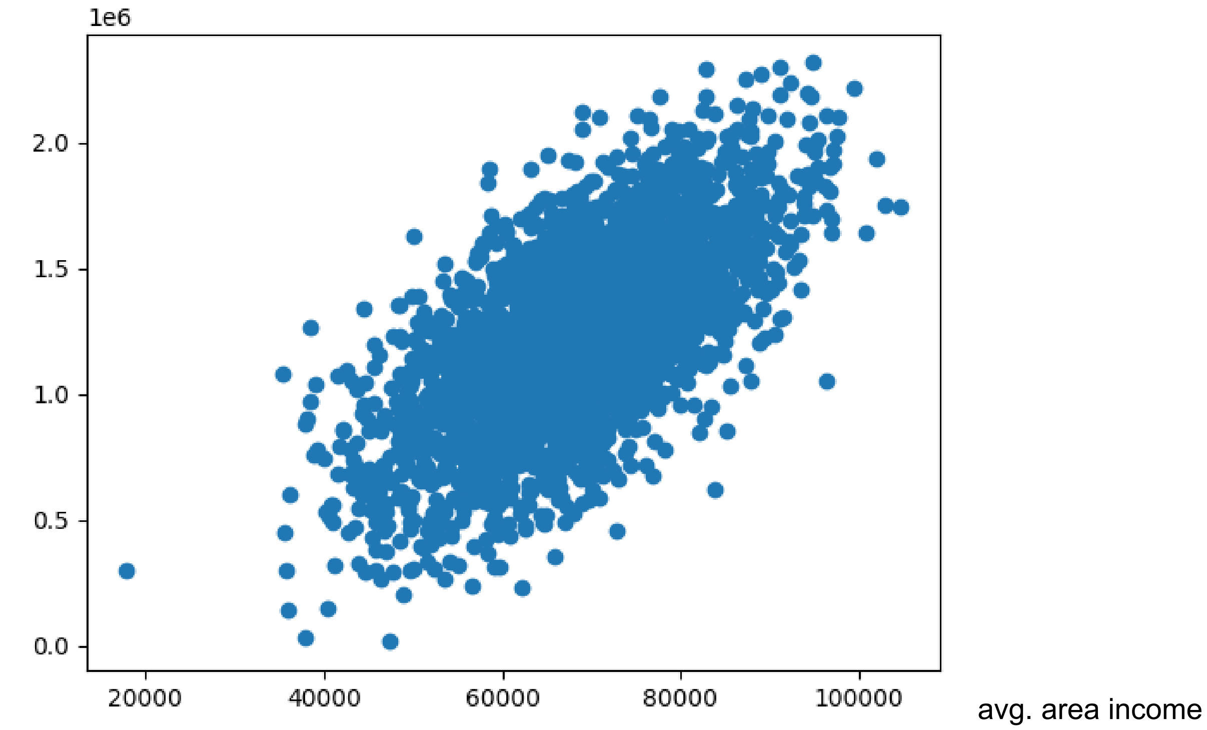


Figure 1

According to this graph, we can see the relationship between avg. area income and price is relatively convergent, with most of the data compiled together, and the overall trend shows that price increases when the avg. area income increases

avg. population vs. price

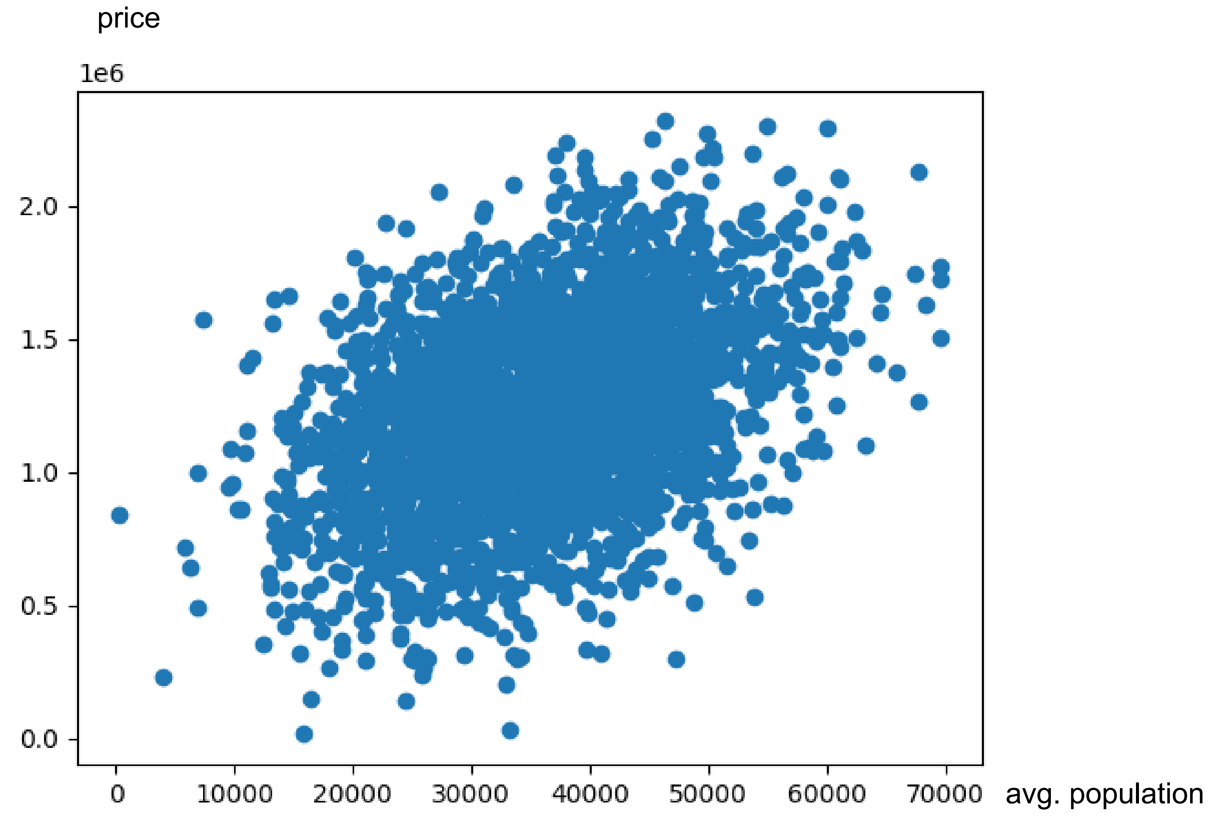
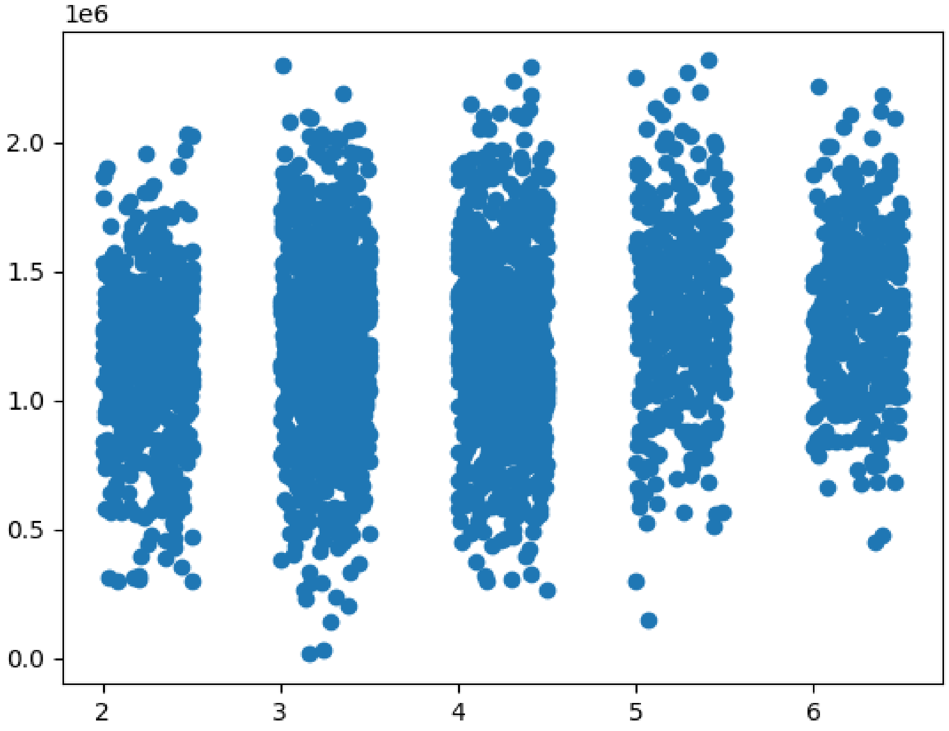


Figure 2

According to this graph, we can see the relationship between avg. population and price is less convergent than Figure 1,however, we can still see the linear relationship between these 2 variables, the overall trend shows that price increases when the avg. population increases

avg. number of bedroom vs. price

price



avg. num of bedroom

Figure 3

According to this graph, we can see a very weak relationship between avg. num of bedrooms and price, the relationship between them are not linear, however, the trend of price slightly increases when the avg. number of bedrooms increases.

avg. number of rooms vs. price

price

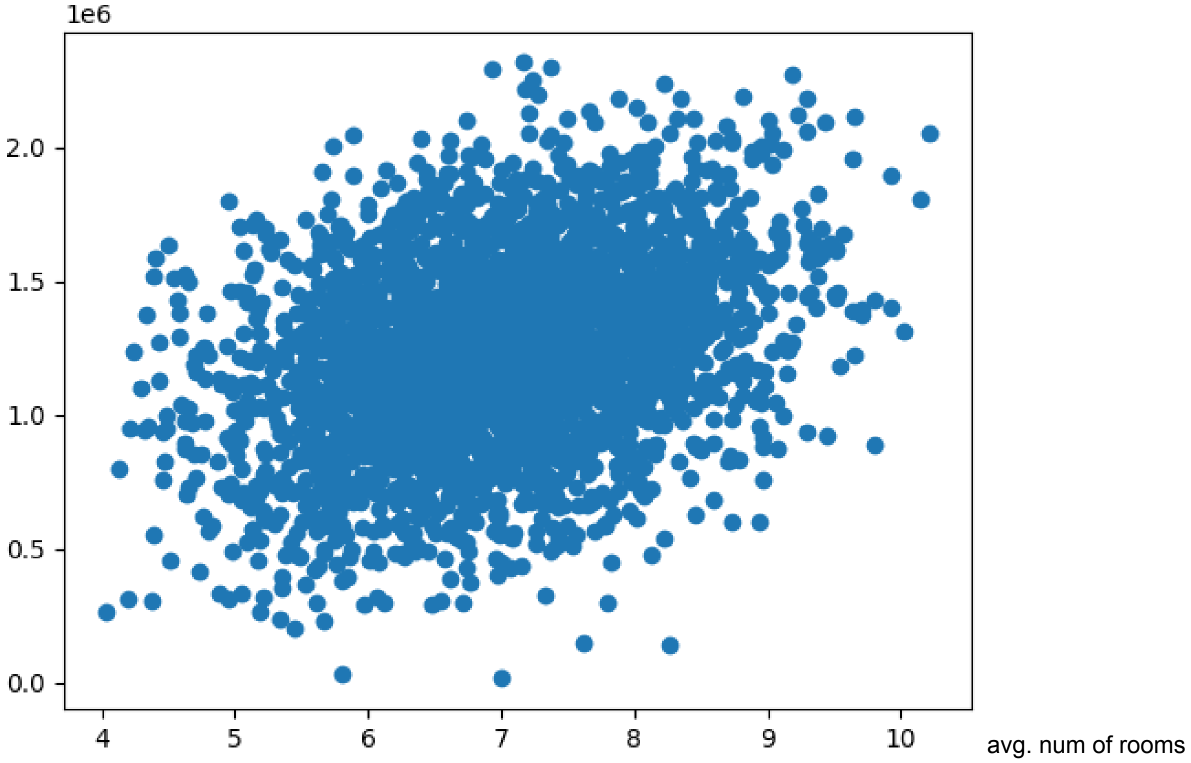


Figure 4

According to this graph, we can the relationship between avg. number of rooms and price is roughly as convergent as Figure 2, the trend of price slightly increases when the avg. number of rooms increases.

avg. house age vs. price

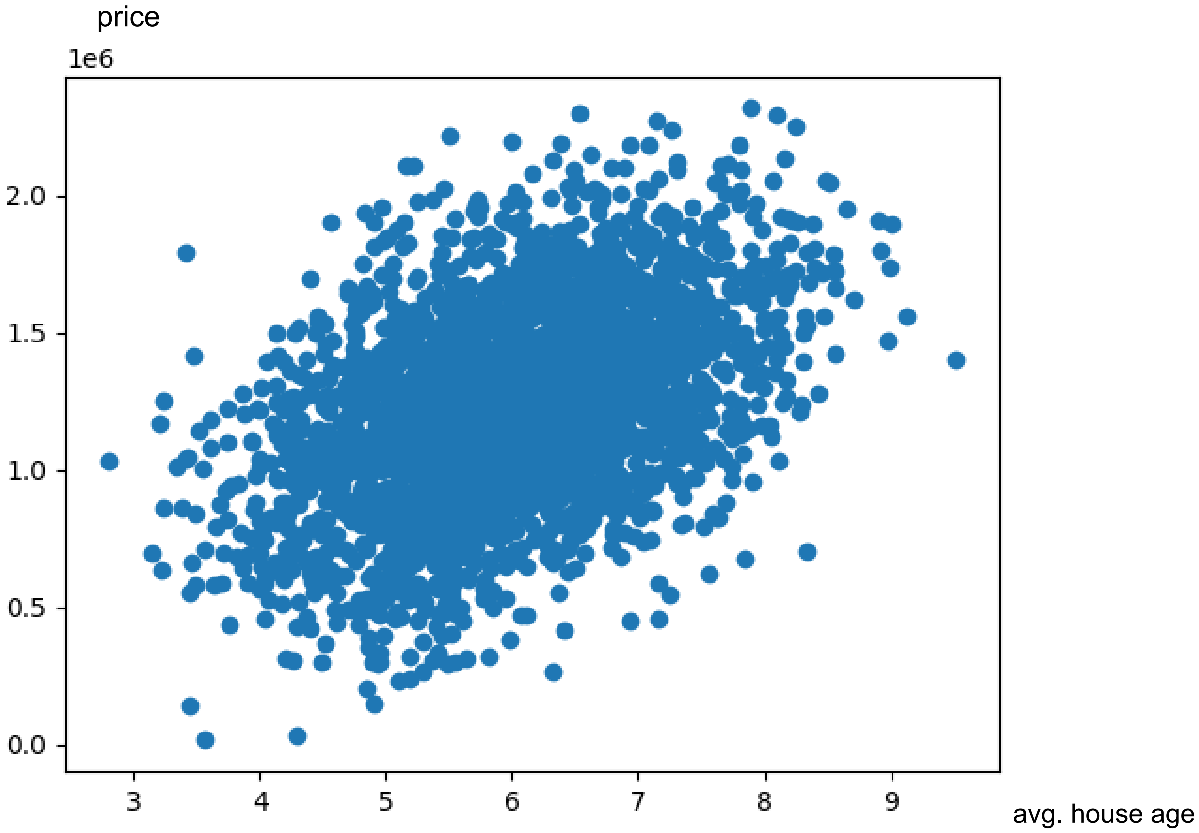


Figure 5

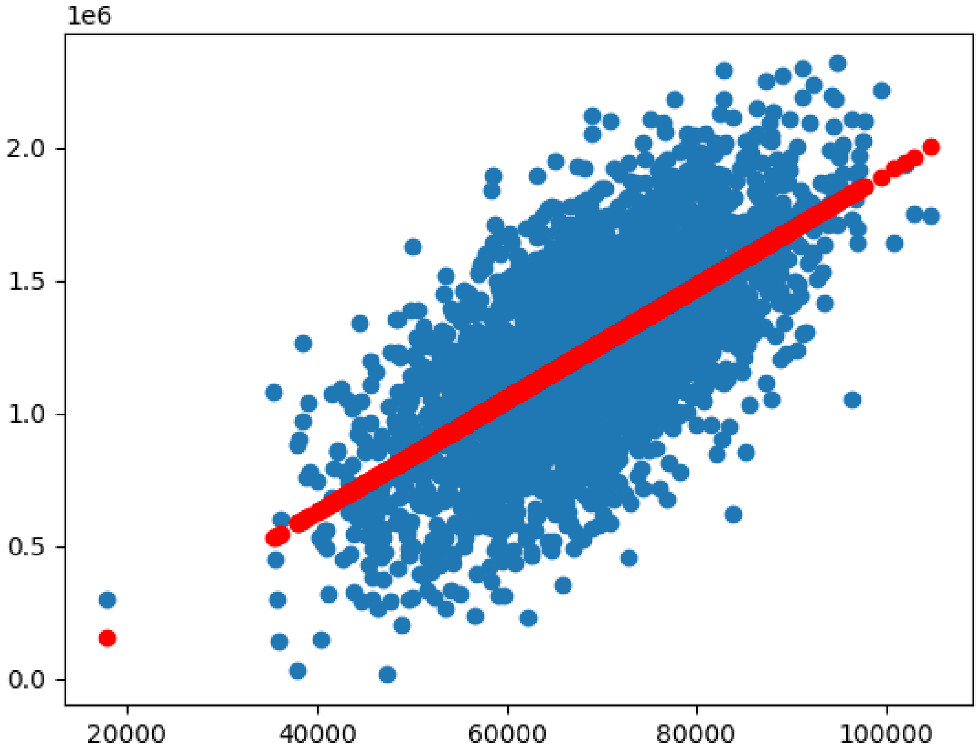
According to this graph, we can find the relationship between avg. number of rooms and price is roughly as convergent as Figure 2, the trend of price slightly increases when the avg. house age increases.

Simple linear regression:

For each independent variable, I executed a linear regression with the dependent variable using a line model (y = mx + b), please find the analysis of them as follows (red lines represent the best fit line):

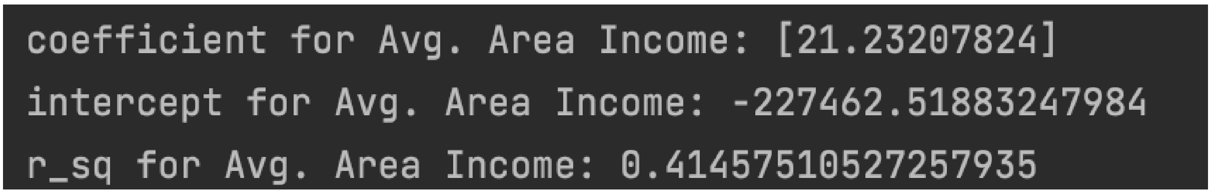
`avg. area income` vs. `price`

price



avg. area income

Figure 6



```
coefficient for Avg. Area Income: [21.23207824]
intercept for Avg. Area Income: -227462.51883247984
r_sq for Avg. Area Income: 0.41457510527257935
```

The coefficient is 21.19, and r^2 = 0.41, the slope is a reasonable number and r is close to 0.5, which means the avg. area income has some correlation with the house price and has some impact on the house price, but it is not a single strong feature.

`avg. population` vs. `price`

price

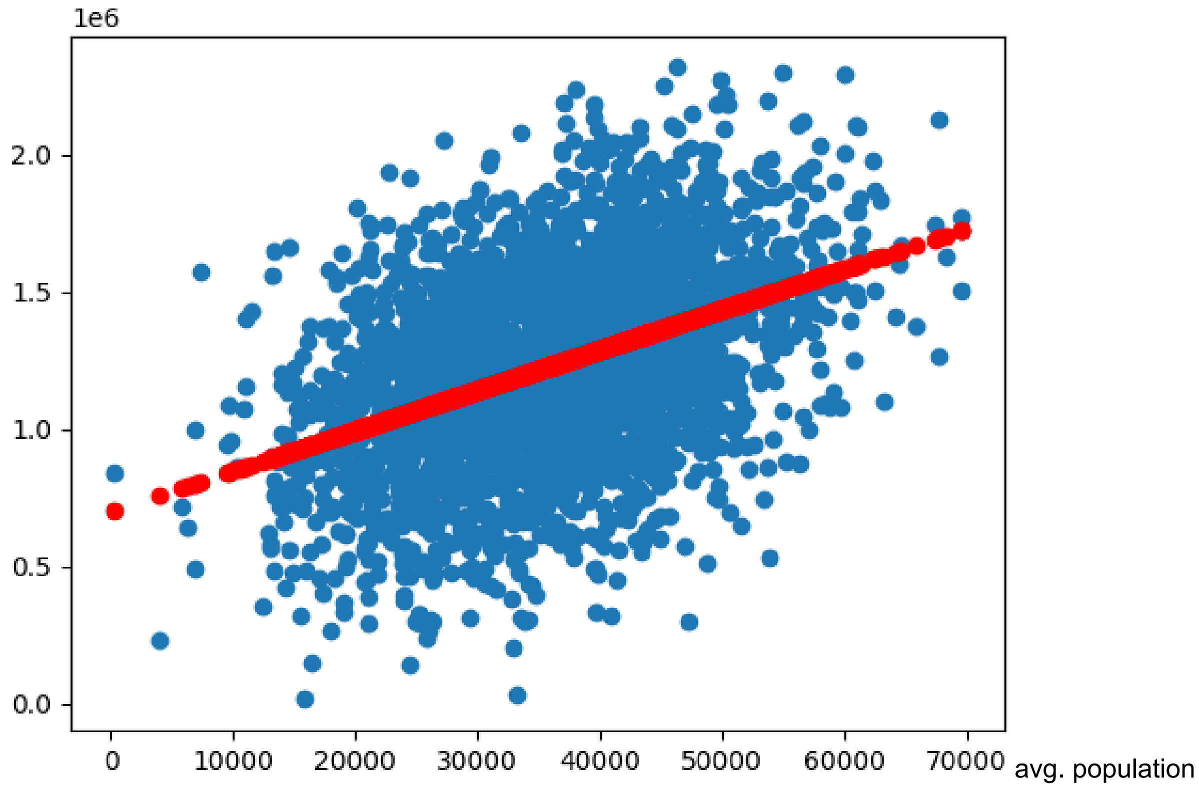
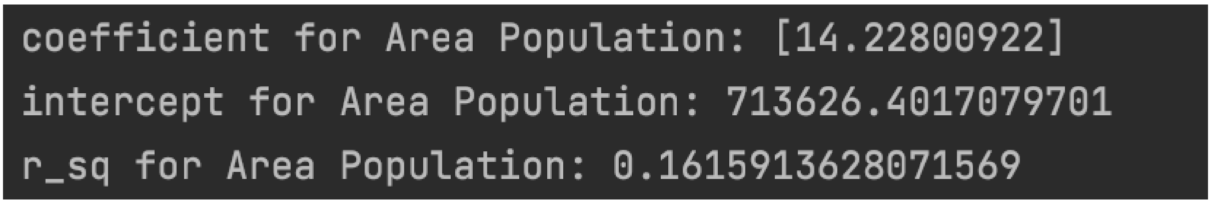


Figure 7



```
coefficient for Area Population: [14.22800922]
intercept for Area Population: 713626.4017079701
r_sq for Area Population: 0.1615913628071569
```

The coefficient is 14.23, and r^2 = 0.16, the slope is a reasonable number and r is relatively small, which means the avg. Area population has small correlation with the house price and has a slight impact on the house price.

`avg. number of bedroom` vs. `price`

price

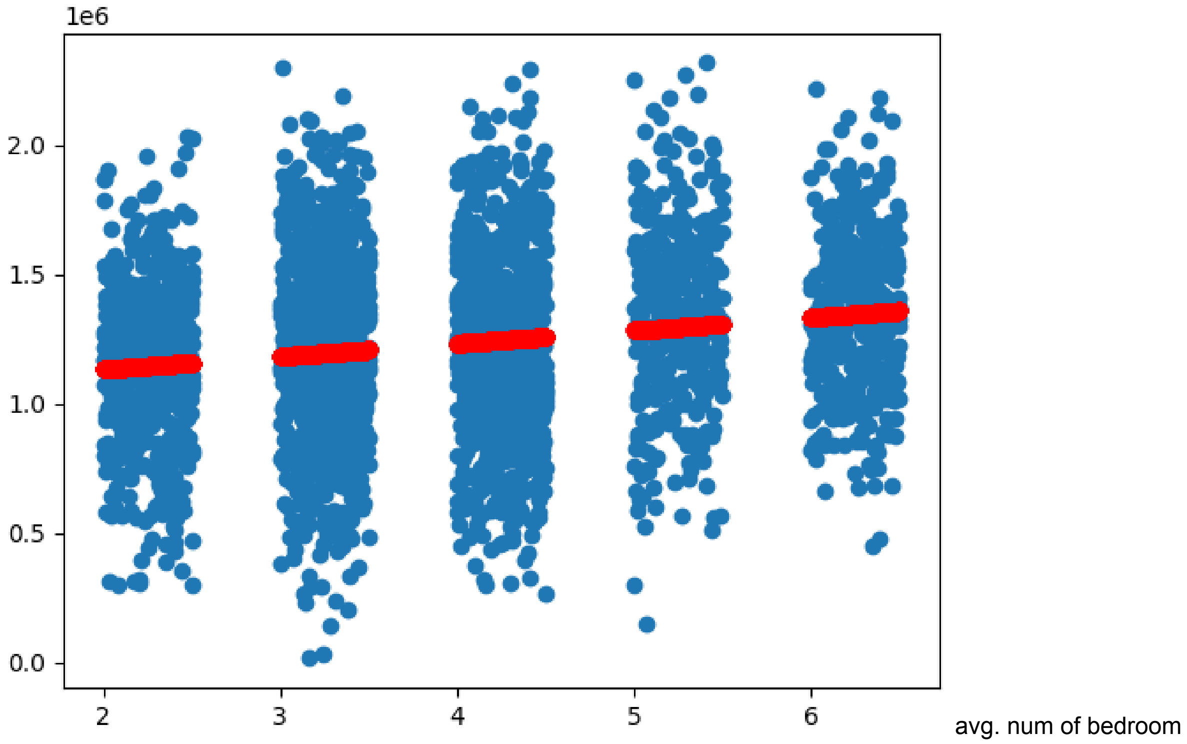
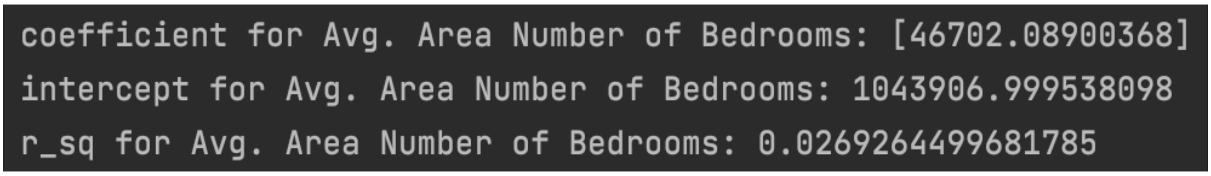


Figure 8



```
coefficient for Avg. Area Number of Bedrooms: [46702.08900368]
intercept for Avg. Area Number of Bedrooms: 1043906.999538098
r_sq for Avg. Area Number of Bedrooms: 0.0269264499681785
```

The coefficient is 46702, and r^2 = 0.026, the slope is large and r is close to 0, which means the avg. number of bedrooms nearly has no correlation with the house price and has nearly no impact on the house price.

avg. number of rooms vs. price

price

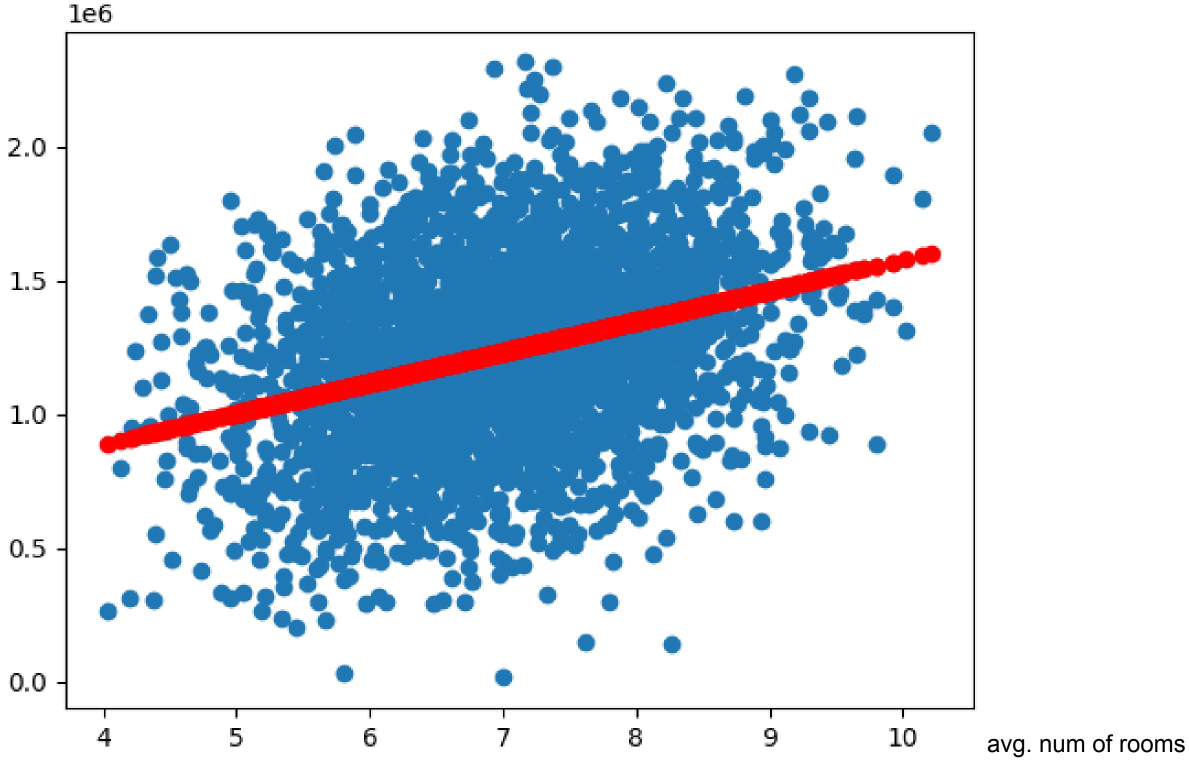
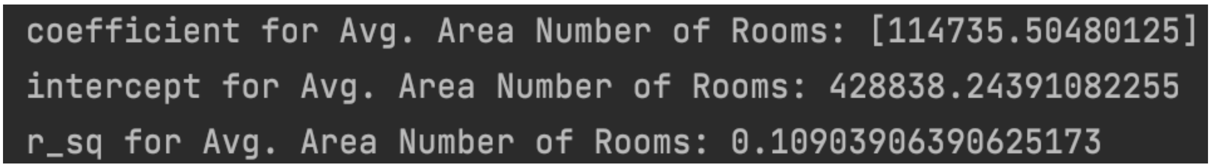


Figure 9



```
coefficient for Avg. Area Number of Rooms: [114735.50480125]
intercept for Avg. Area Number of Rooms: 428838.24391082255
r_sq for Avg. Area Number of Rooms: 0.10903906390625173
```

The coefficient is 114735.50, and r^2 = 0.11, the slope is large and r is small, which means the avg. number of rooms has slight correlation with the house price and has a slight impact on the house price.

avg. house age vs. price

price

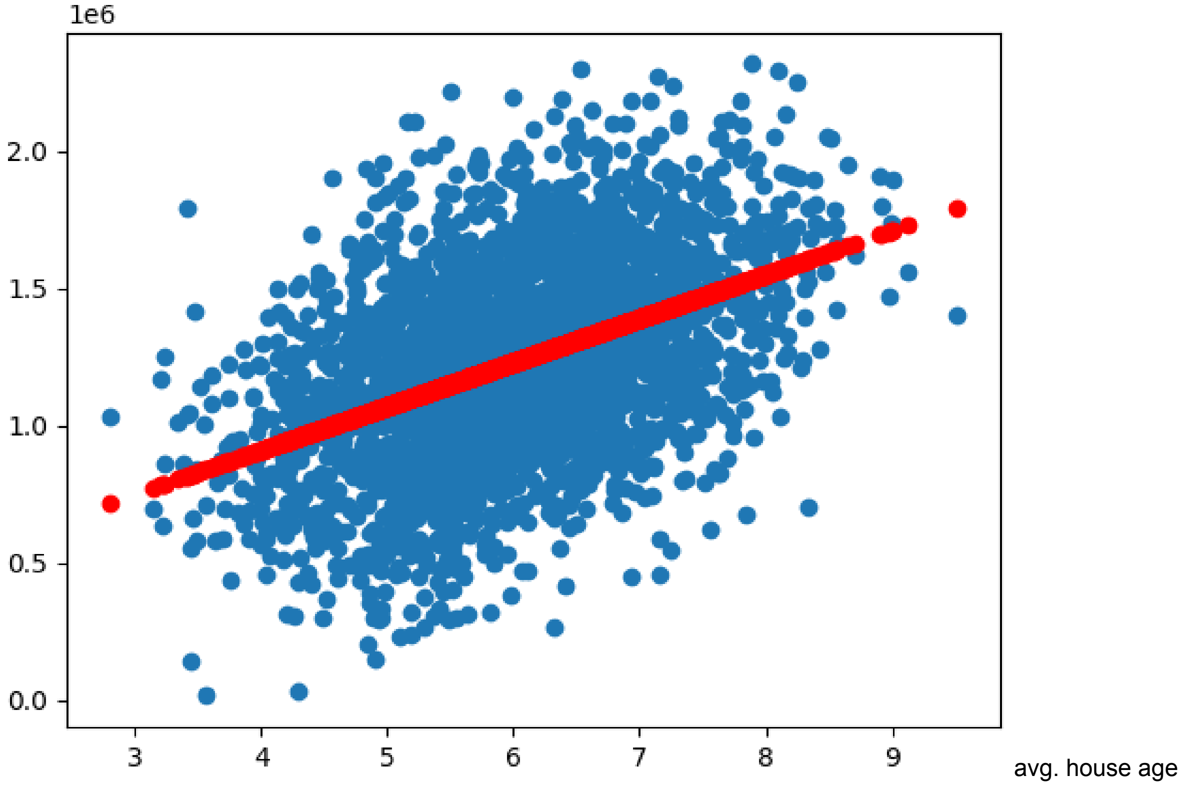
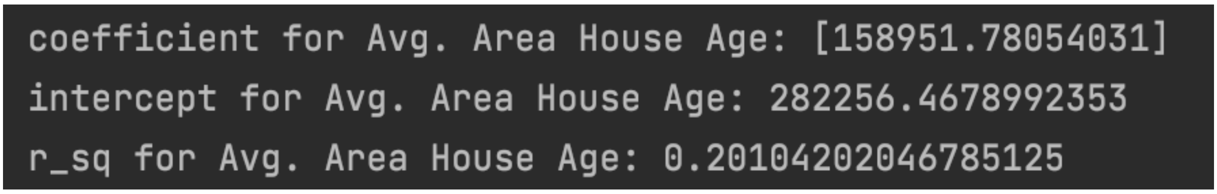


Figure 10

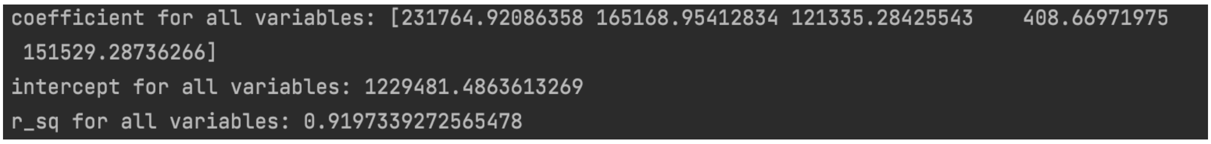


```
coefficient for Avg. Area House Age: [158951.78054031]
intercept for Avg. Area House Age: 282256.4678992353
r_sq for Avg. Area House Age: 0.20104202046785125
```

The coefficient is 158951.78, and r^2 = 0.20, the slope is large and r is relatively small, which means the avg. Area population has relatively small correlation with the house price and has a relatively small impact on the house price.

Multiple linear regression:

I also executed multiple linear regression on the data set with all of the independent numerical variable, please find the coefficient and r^2 as follows:



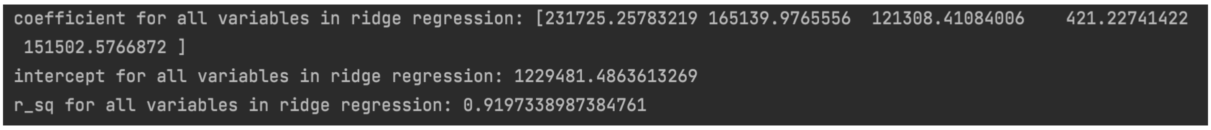
```
coefficient for all variables: [231764.92086358 165168.95412834 121335.28425543    408.66971975
 151529.28736266]
intercept for all variables: 1229481.4863613269
r_sq for all variables: 0.9197339272565478
```

the coefficient for all variables:
avg. area income: 231764.92
avg. area house age: 165168.95
avg. area number of rooms: 121335.28
avg. area number of bedrooms: 408.67
area population: 151529.29

r_sq = 0.92
The r_sqicient is 0.92, which means the multiple linear regression fits the data pretty well. Among all the independent variables, the avg. area income is the most correlated variable,

and the avg. area number of bedrooms is very slightly correlated, and the other variables are somewhat correlated but not as strong as avg. area income. Avg. area income has the biggest impact on the price.

I also executed ridge linear regression on the data set with all of the independent numerical variable:



```
coefficient for all variables in ridge regression: [231725.25783219 165139.9765556  121308.41084006    421.22741422
 151502.5766872 ]
intercept for all variables in ridge regression: 1229481.4863613269
r_sq for all variables in ridge regression: 0.9197338987384761
```

the coefficient for all variables:
avg. area income: 231725.26
avg. area house age: 165139.97
avg. area number of rooms: 121308.41
avg. area number of bedrooms: 421.23
area population: 151502.58

r_sq = 0.92
The r_sqicient is 0.92, it is the same as linear regression, which means the ridge linear regression fits the data pretty well. And the difference between coefficients of the same variable is really small, which means ridge and OLS linear regression produce similar results on my dataset, which means most of the data in my dataset are useful and there is a low level of overfitting.

3rd order polynomial regression

I also executed the 3rd order polynomial regression on one of the numerical independent variables (avg. area income) in my data set, here's the graph of best fit:

`avg. area income` vs. `price`

price

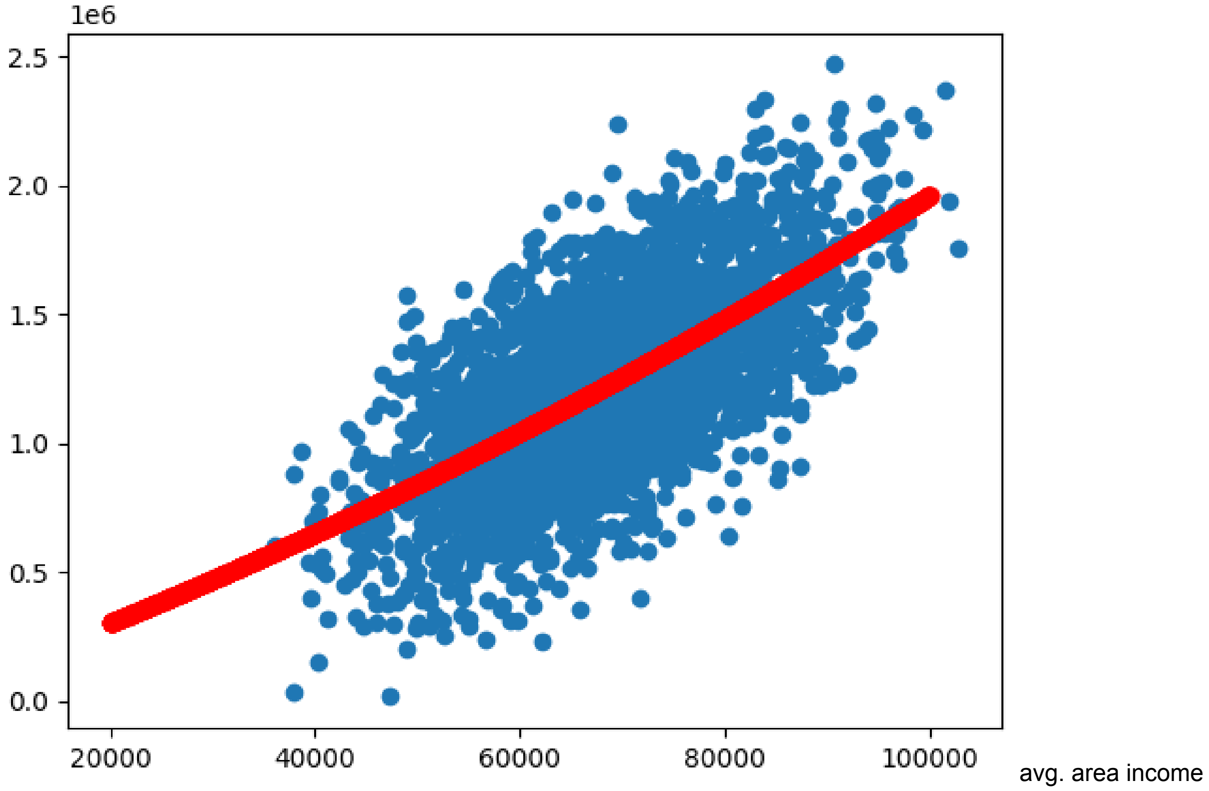
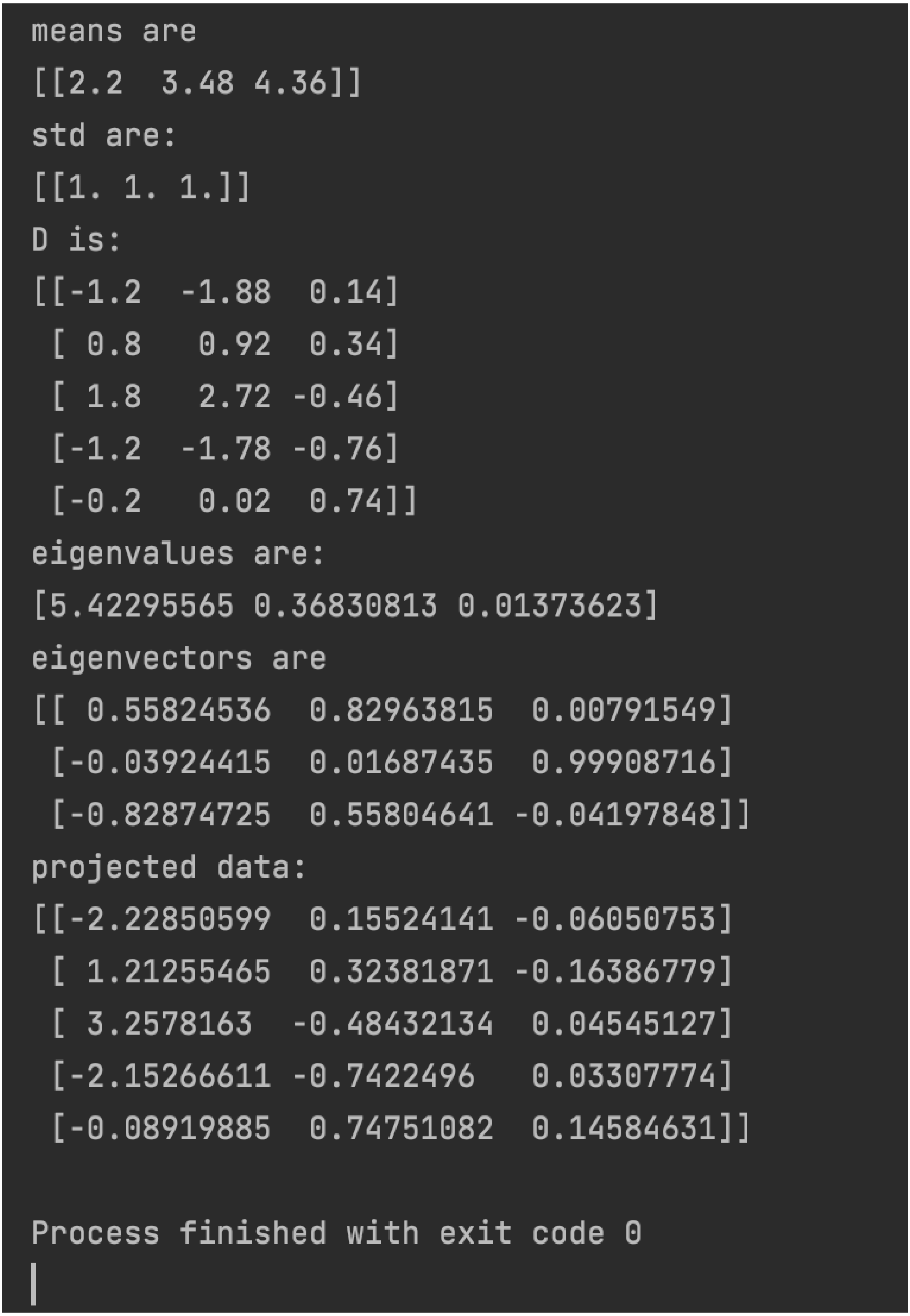


Figure 11

PCA

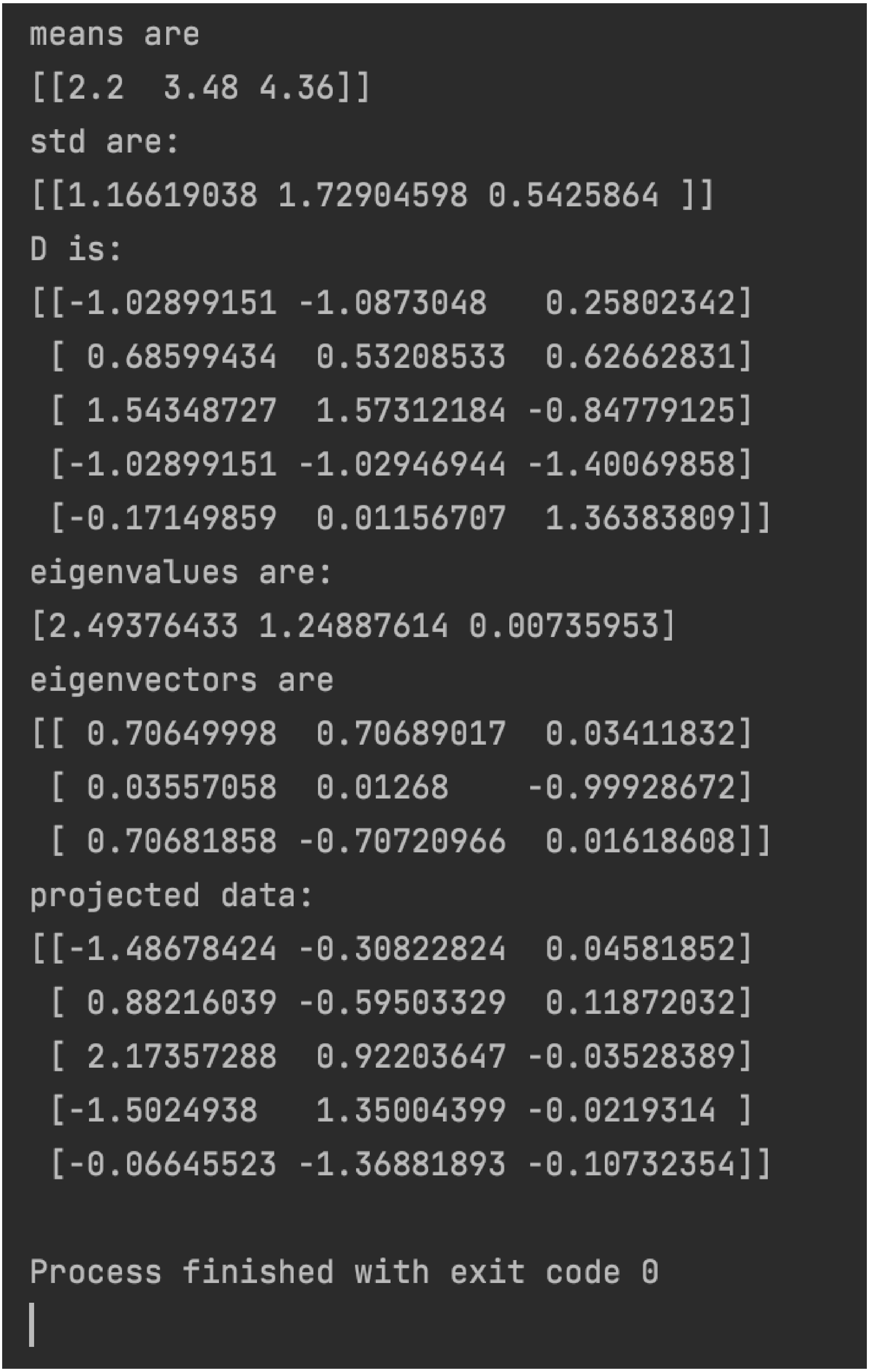
Please find the result for PCA function on the non_whitened data:



```
means are
[[2.2  3.48 4.36]]
std are:
[[1. 1. 1.]]
D is:
[[-1.2  -1.88  0.14]
 [ 0.8   0.92  0.34]
 [ 1.8   2.72 -0.46]
 [-1.2  -1.78 -0.76]
 [-0.2   0.02  0.74]]
eigenvalues are:
[5.42295565 0.36830813 0.01373623]
eigenvectors are
[[ 0.55824536  0.82963815  0.00791549]
 [-0.03924415  0.01687435  0.99908716]
 [-0.82874725  0.55804641 -0.04197848]]
projected data:
[[-2.22850599  0.15524141 -0.06050753]
 [ 1.21255465  0.32381871 -0.16386779]
 [ 3.2578163  -0.48432134  0.04545127]
 [-2.15266611 -0.7422496   0.03307774]
 [-0.08919885  0.74751082  0.14584631]]

Process finished with exit code 0
```

It is the same as the expected result.

Result for PCA function on the whitened data:



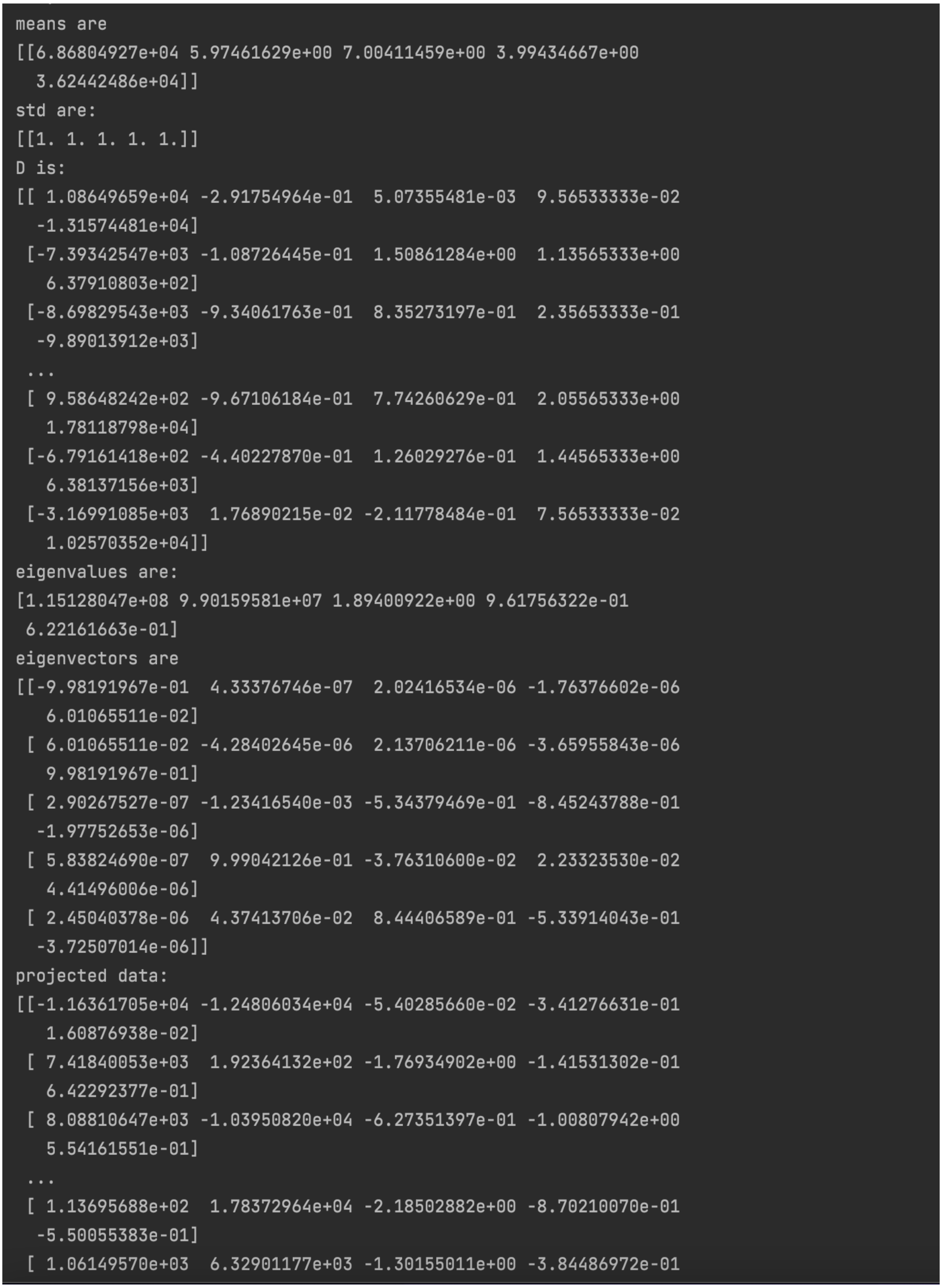
```
means are
[[2.2  3.48 4.36]]
std are:
[[1.16619038 1.72904598 0.5425864 ]]
D is:
[[-1.02899151 -1.0873048   0.25802342]
 [ 0.68599434  0.53208533  0.62662831]
 [ 1.54348727  1.57312184 -0.84779125]
 [-1.02899151 -1.02946944 -1.40069858]
 [-0.17149859  0.01156707  1.36383809]]
eigenvalues are:
[2.49376433 1.24887614 0.00735953]
eigenvectors are
[[ 0.70649998  0.70689017  0.03411832]
 [ 0.03557058  0.01268    -0.99928672]
 [ 0.70681858 -0.70720966  0.01618608]]
projected data:
[[-1.48678424 -0.30822824  0.04581852]
 [ 0.88216039 -0.59503329  0.11872032]
 [ 2.17357288  0.92203647 -0.03528389]
 [-1.5024938   1.35004399 -0.0219314 ]
 [-0.06645523 -1.36881893 -0.10732354]]

Process finished with exit code 0
```

It is the same as the expected result.

Difference in eigenvalues:
the eigenvalues for non_whitened data are much bigger than eigenvalues for the whitened data, this is because after whitening, the data gets scaled down and has zero mean and unit variance, the divergence of data gets smaller and the dataset gets normalized, which lead to a smaller eigenvalues.

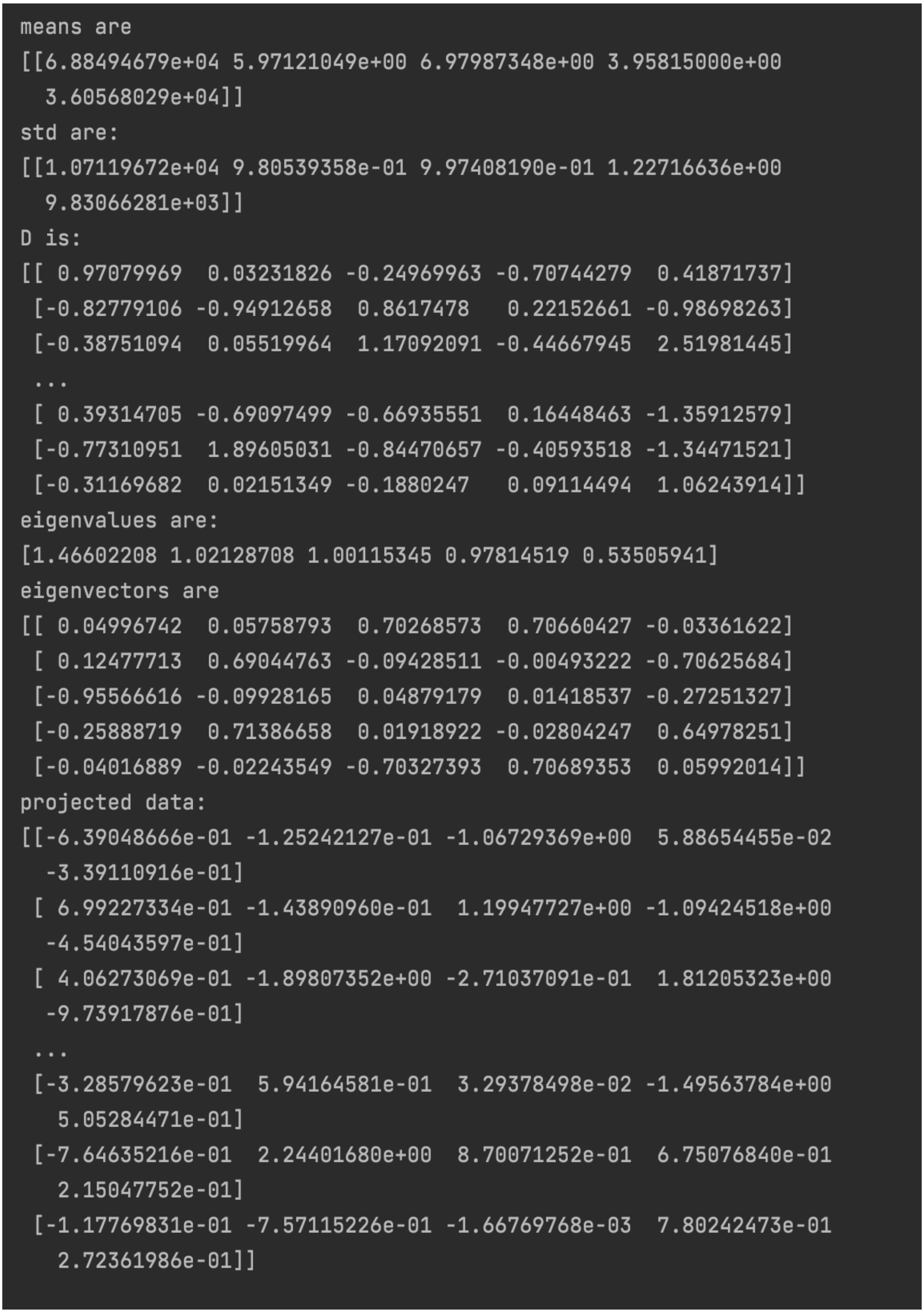
Apply PCA on USA_Housing dataset

result for non_whitened data:



```
means are
[[6.86804927e+04 5.97461629e+00 7.00411459e+00 3.99434667e+00
  3.62442486e+04]]
std are:
[[1. 1. 1. 1. 1.]]
D is:
[[ 1.08649659e+04 -2.91754964e-01  5.07355481e-03  9.56533333e-02
  -1.31574481e+04]
 [-7.39342547e+03 -1.08726445e-01  1.50861284e+00  1.13565333e+00
   6.37910803e+02]
 [-8.69829543e+03 -9.34061763e-01  8.35273197e-01  2.35653333e-01
  -9.89013912e+03]
 ...
 [ 9.58648242e+02 -9.67106184e-01  7.74260629e-01  2.05565333e+00
   1.78118798e+04]
 [-6.79161418e+02 -4.40227870e-01  1.26029276e-01  1.44565333e+00
   6.38137156e+03]
 [-3.16991085e+03  1.76890215e-02 -2.11778484e-01  7.56533333e-02
   1.02570352e+04]]
eigenvalues are:
[1.15128047e+08 9.90159581e+07 1.89400922e+00 9.61756322e-01
 6.22161663e-01]
eigenvectors are
[[-9.98191967e-01  4.33376746e-07  2.02416534e-06 -1.76376602e-06
   6.01065511e-02]
 [ 6.01065511e-02 -4.28402645e-06  2.13706211e-06 -3.65955843e-06
   9.98191967e-01]
 [ 2.90267527e-07 -1.23416540e-03 -5.34379469e-01 -8.45243788e-01
  -1.97752653e-06]
 [ 5.83824690e-07  9.99042126e-01 -3.76310600e-02  2.23323530e-02
   4.41496006e-06]
 [ 2.45040378e-06  4.37413706e-02  8.44406589e-01 -5.33914043e-01
  -3.72507014e-06]]
projected data:
[[-1.16361705e+04 -1.24806034e+04 -5.40285660e-02 -3.41276631e-01
   1.60876938e-02]
 [ 7.41840053e+03  1.92364132e+02 -1.76934902e+00 -1.41531302e-01
   6.42292377e-01]
 [ 8.08810647e+03 -1.03950820e+04 -6.27351397e-01 -1.00807942e+00
   5.54161551e-01]
 ...
 [ 1.13695688e+02  1.78372964e+04 -2.18502882e+00 -8.70210070e-01
  -5.50055383e-01]
 [ 1.06149570e+03  6.32901177e+03 -1.30155011e+00 -3.84486972e-01
```

result for whitened data:



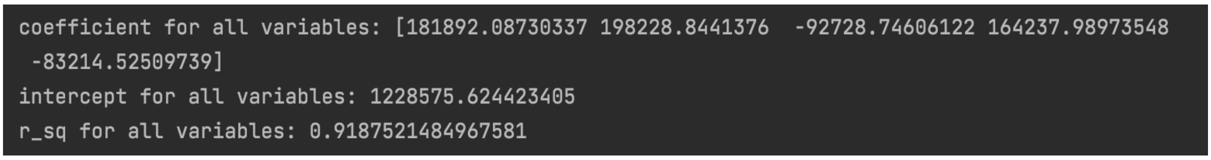
```
means are
[[6.88494679e+04 5.97121049e+00 6.97987348e+00 3.95815000e+00
  3.60568029e+04]]
std are:
[[1.07119672e+04 9.80539358e-01 9.97408190e-01 1.22716636e+00
  9.83066281e+03]]
D is:
[[ 0.97079969  0.03231826 -0.24969963 -0.70744279  0.41871737]
 [-0.82779106 -0.94912658  0.8617478   0.22152661 -0.98698263]
 [-0.38751094  0.05519964  1.17092091 -0.44667945  2.51981445]
 ...
 [ 0.39314705 -0.69097499 -0.66935551  0.16448463 -1.35912579]
 [-0.77310951  1.89605031 -0.84470657 -0.40593518 -1.34471521]
 [-0.31169682  0.02151349 -0.1880247   0.09114494  1.06243914]]
eigenvalues are:
[1.46602208 1.02128708 1.00115345 0.97814519 0.53505941]
eigenvectors are
[[ 0.04996742  0.05758793  0.70268573  0.70660427 -0.03361622]
 [ 0.12477713  0.69044763 -0.09428511 -0.00493222 -0.70625684]
 [-0.95566616 -0.09928165  0.04879179  0.01418537 -0.27251327]
 [-0.25888719  0.71386658  0.01918922 -0.02804247  0.64978251]
 [-0.04016889 -0.02243549 -0.70327393  0.70689353  0.05992014]]
projected data:
[[-6.39048666e-01 -1.25242127e-01 -1.06729369e+00  5.88654455e-02
  -3.39110916e-01]
 [ 6.99227334e-01 -1.43890960e-01  1.19947727e+00 -1.09424518e+00
  -4.54043597e-01]
 [ 4.06273069e-01 -1.89807352e+00 -2.71037091e-01  1.81205323e+00
  -9.73917876e-01]
 ...
 [-3.28579623e-01  5.94164581e-01  3.29378498e-02 -1.49563784e+00
   5.05284471e-01]
 [-7.64635216e-01  2.24401680e+00  8.70071252e-01  6.75076840e-01
   2.15047752e-01]
 [-1.17769831e-01 -7.57115226e-01 -1.66769768e-03  7.80242473e-01
   2.72361986e-01]]
```

The eigenvalues for non_whitened value and whitened_value are much different, the non_whitened eigenvalues are much larger than whitened eigenvalues, this is because by normalizing the values, it scales down the data to unit variant and make the data distributed in a more standard way. So it makes sense to do whitening.

According to the eigenvectors, it has 5 significant dimensions, and most of them are equal to or larger than 1, one of them close to 0.5, which means most of the features are correlated to the data. According to the eigenvectors, the 5 columns have relatively the same sum of eigenvalues and have similar impact on the pricing.

Apply whitened data on multiple linear regression

I applied multiple linear regression on the whitened data, please find the result as follows:



```
coefficient for all variables: [181892.08730337 198228.8441376  -92728.74606122 164237.98973548
 -83214.52509739]
intercept for all variables: 1228575.624423405
r_sq for all variables: 0.9187521484967581
```

After PCA, the coefficient changes because the data are projected, and most of them are important to the pricing, and the r_sq is 0.92, which means this model can perfectly reflect the data. And it is almost the same as the result for multiple linear regression without PCA.