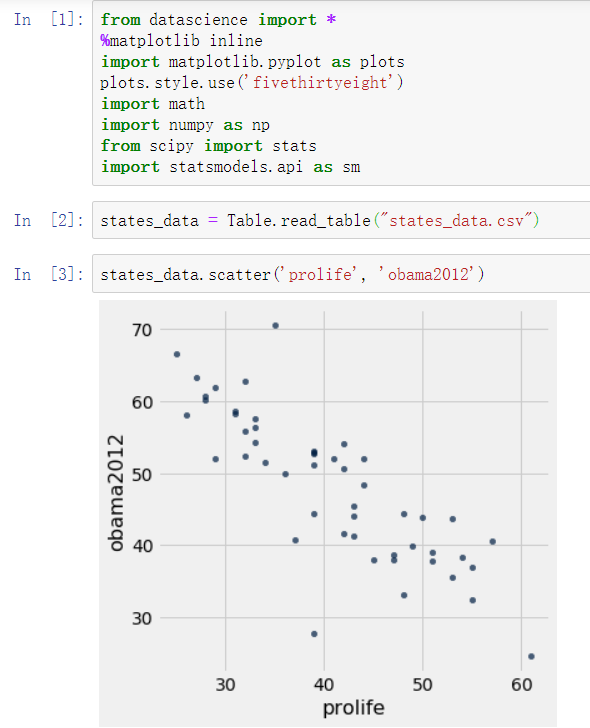
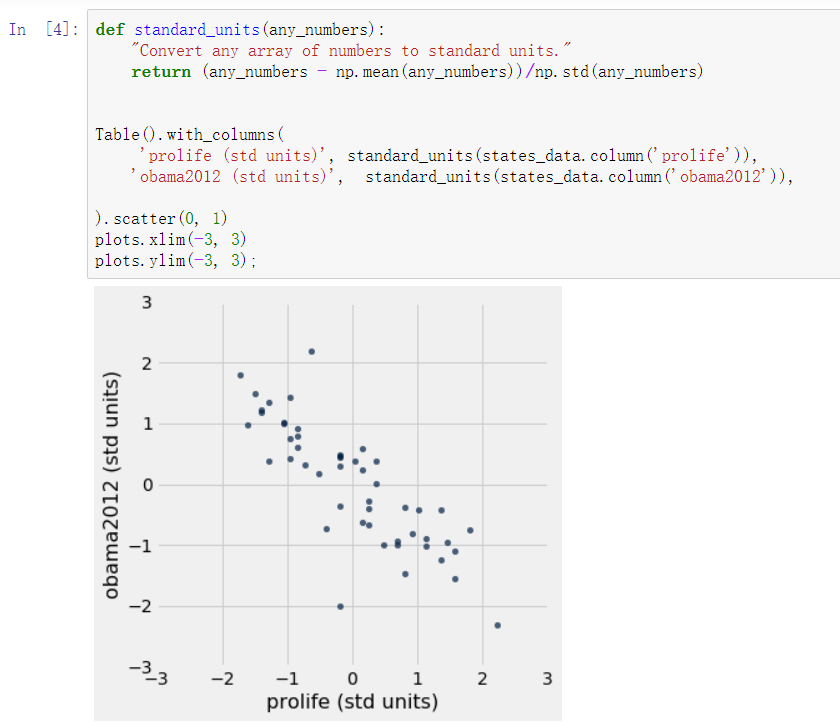
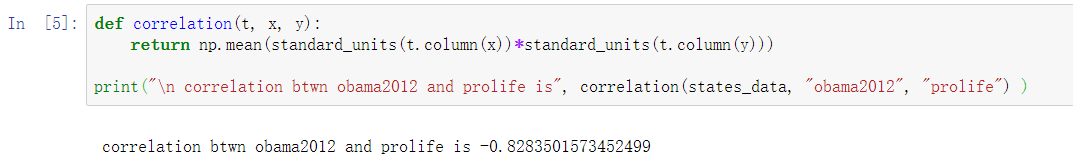
1. a)



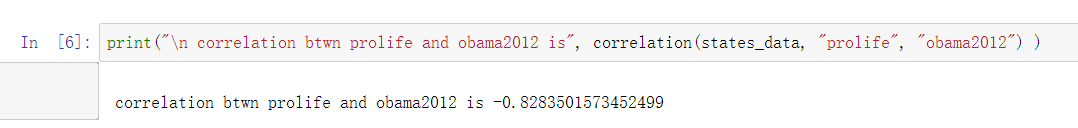
There’s a moderately strong negative relationship between the percentage of people in a state with pro-life views and the percentage of the presidential vote earned by Obama in 2012 in that state.

b) 

No. Because there’s no point at (0, 0) in the graph.

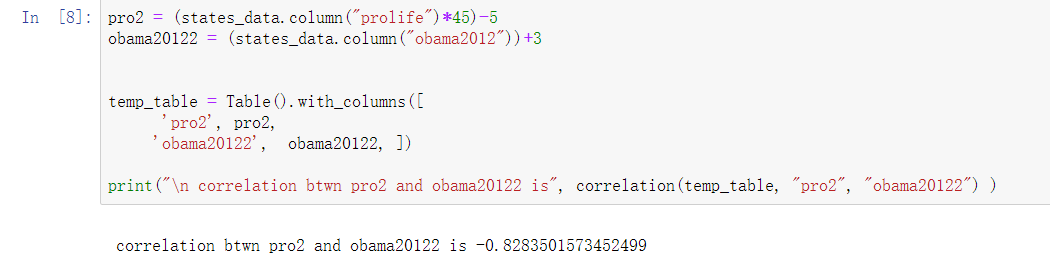
c) 

d)

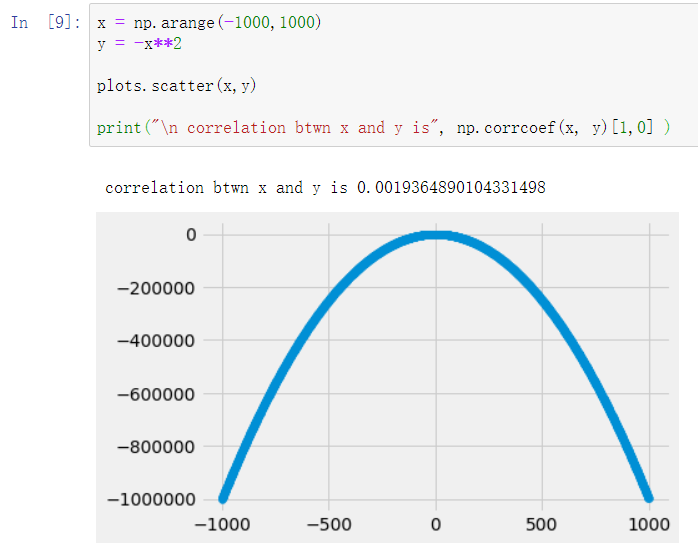


The two are the same, so the correlation between variables is symmetric.

e)

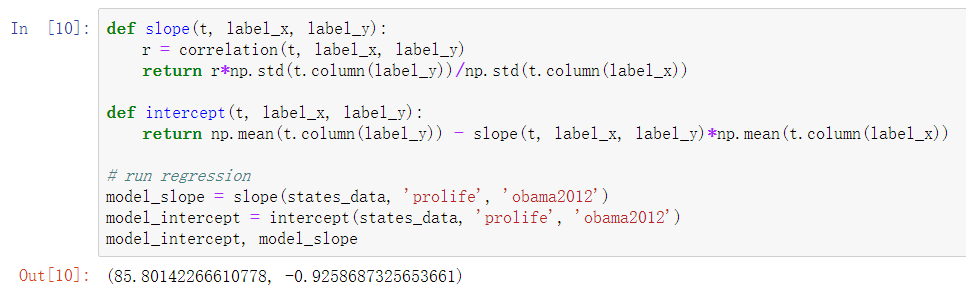


I found that correlation doesn’t change after rescaling. It’s because linear rescaling does not affect the correlation between the variables. When calculating the correlation coeﬃcient, the variables are standardized. So changes in scale does not matter.

f) 

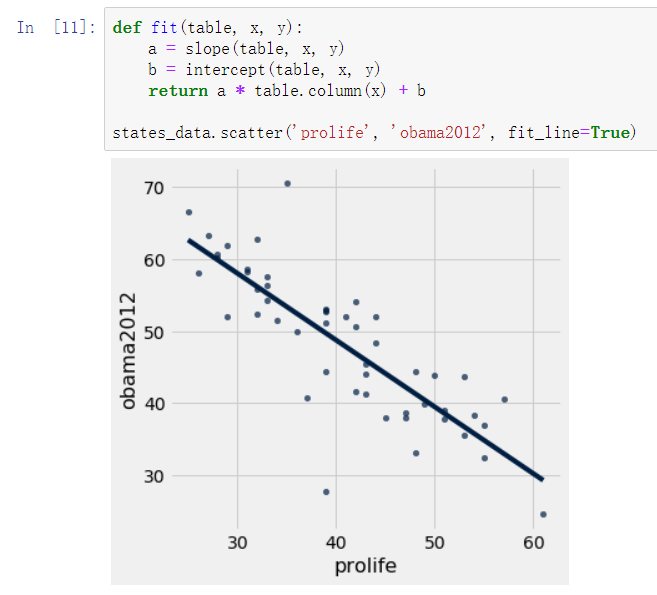
They are related, but it’s not a linear relationship. The correlation is basically zero. Because correlation only measures linear association. Variables that have strong non-linear association might have very low correlation.

g)

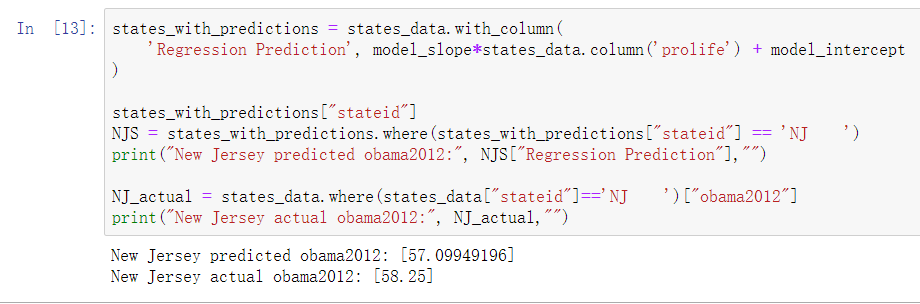


When the percentage of people in a state with pro-life views is zero, the percentage of the expected presidential vote earned by Obama in 2012 in that state would be basically 85.8014%. When the percentage of people in a state with pro-life views increases by one percent, the expected percentage of the presidential vote earned by Obama in 2012 in that state would decrease by basically 0.9259%.

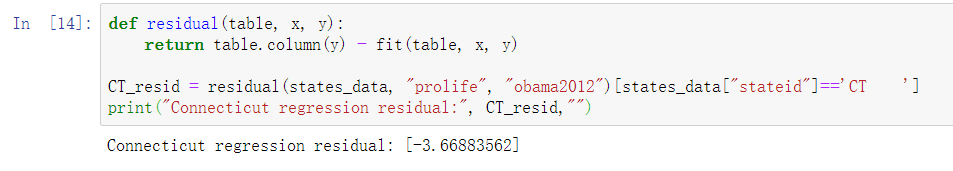
h)



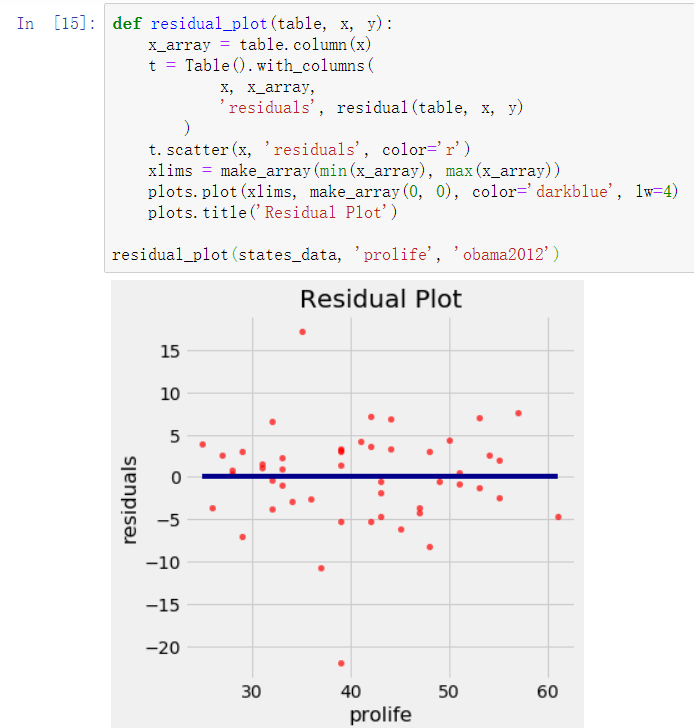
i)



j)

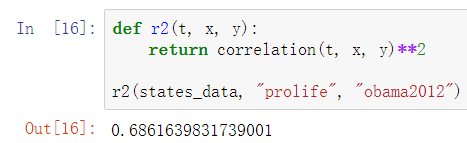


k)



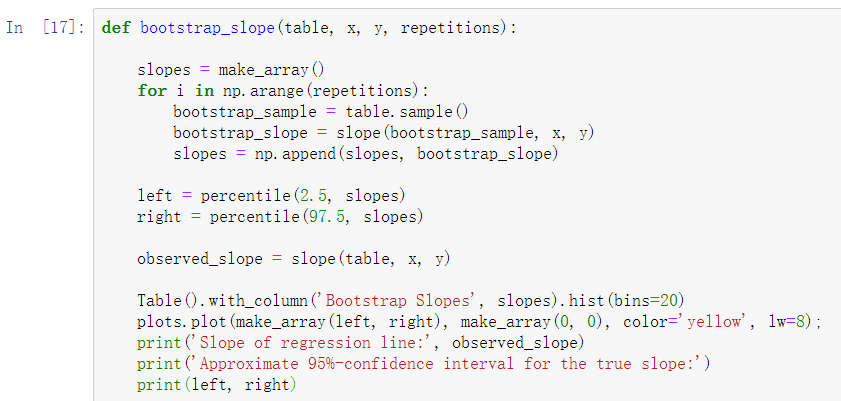
For range between 50% to 60% people in a state with pro-life views, the residual appears to be a cone-like shape. The residual is getting larger. The variability of “prolife” is unequal across the range.

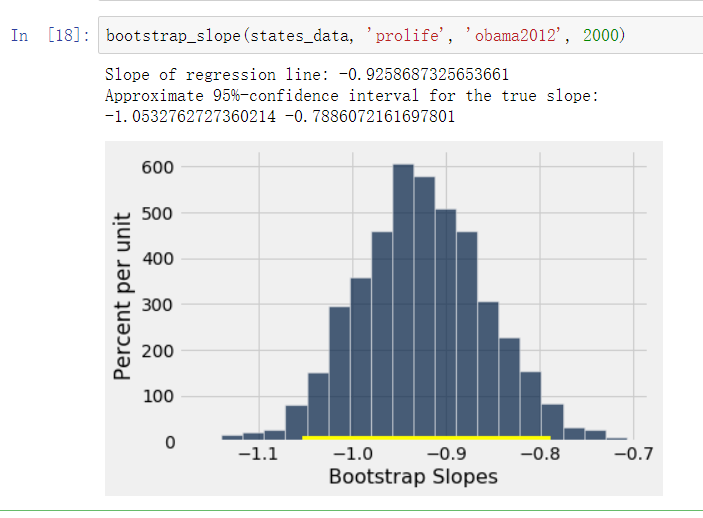
l)



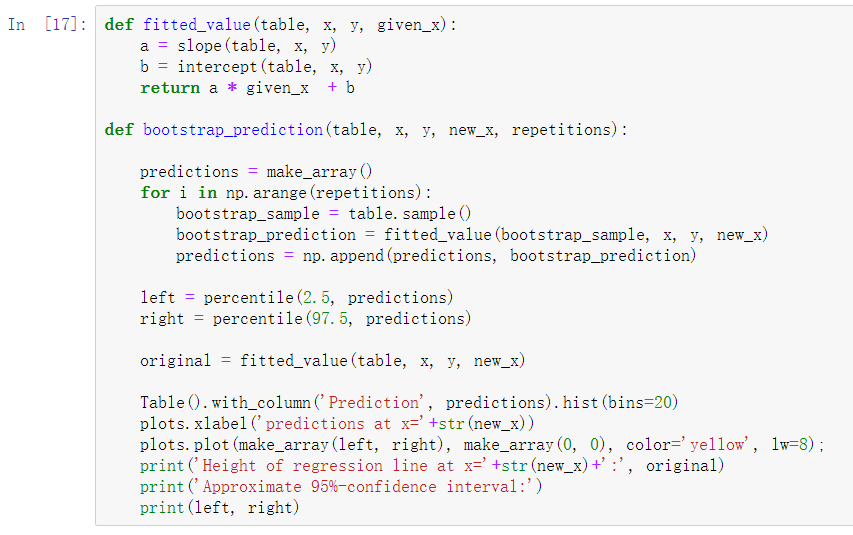
R^2 equals 0.6861639831739001 means the model explains basically 68.6164% of the variability of the response data around its mean.

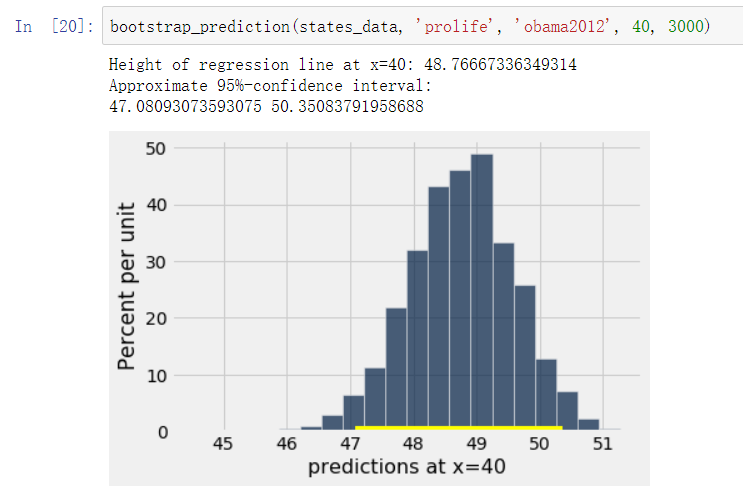
m)

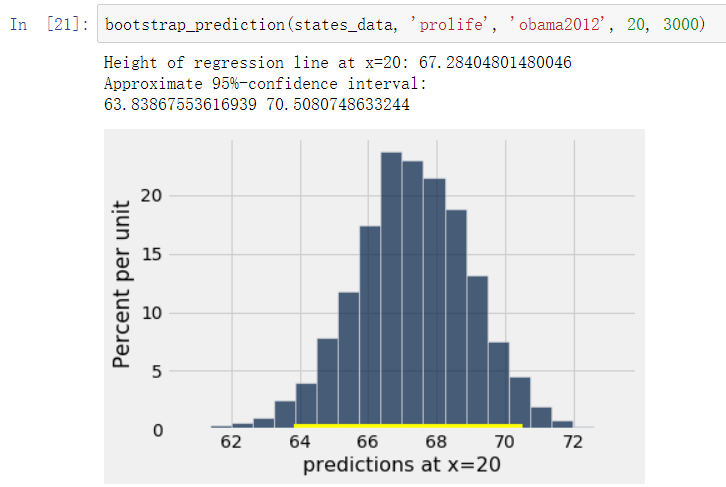




Yes, it is statistically significant. Because the 95% confidence interval for the true slope [-1.0532762727360214 -0.7886072161697801] includes the true slope of the regression line, -0.9258687325653661.

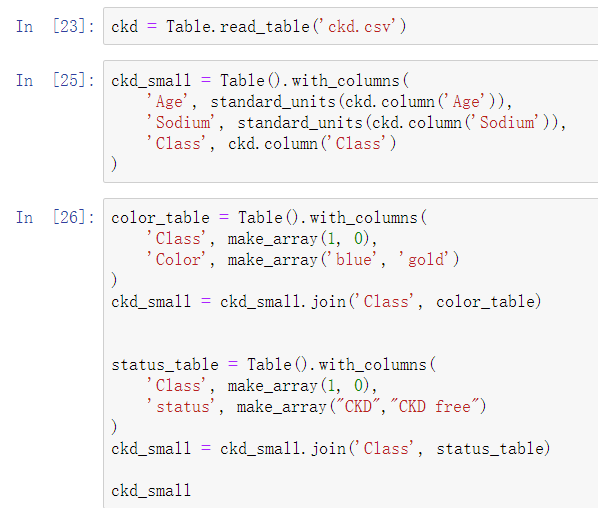
n) 

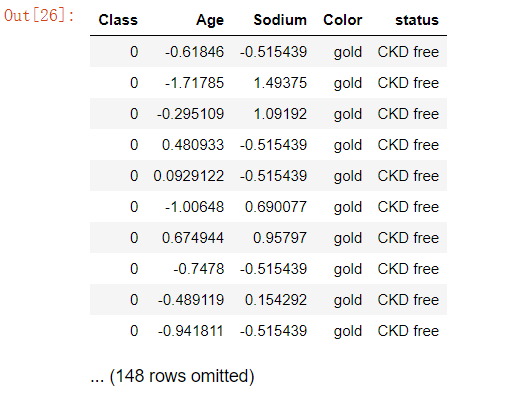


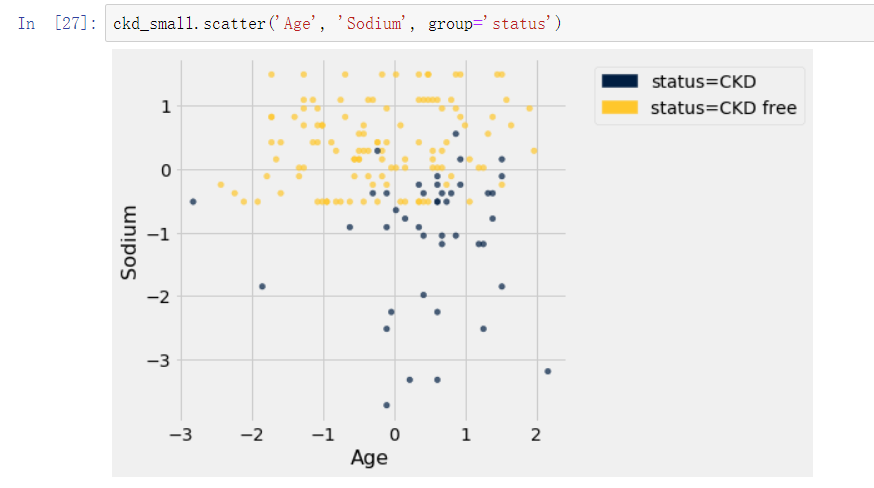


The one with x=20 is wider. Because with a larger sample size x=40, we can make a more specific prediction for confidence interval.

1. a) 

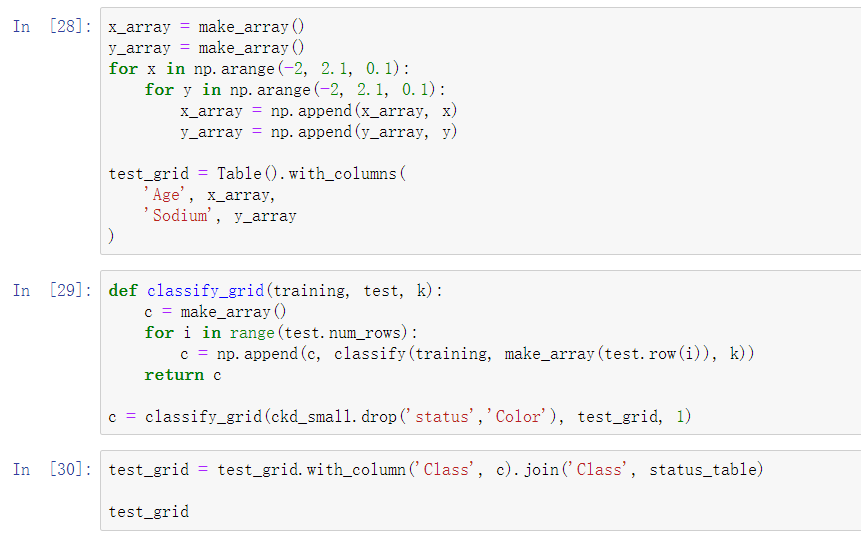


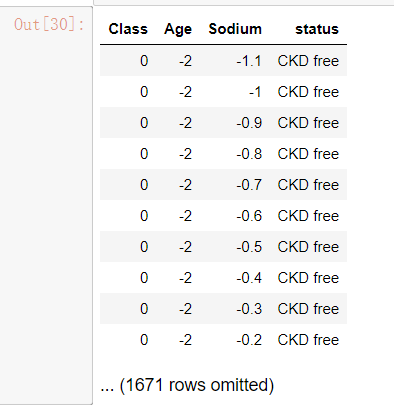


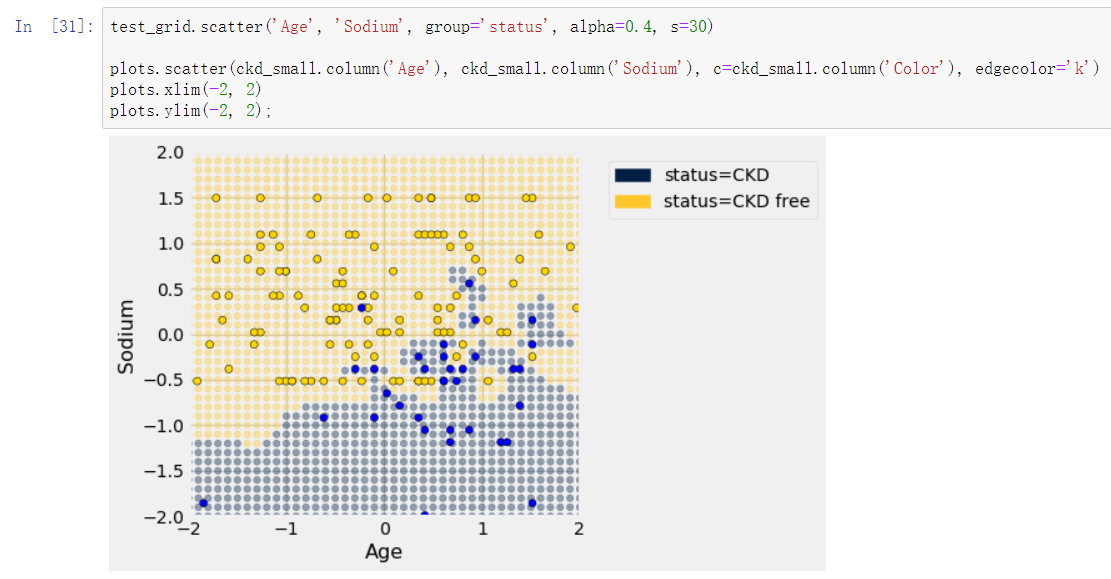


Generally speaking, sodium and age are both key factors of CKD. Low Sodium and high age is a bad combination.

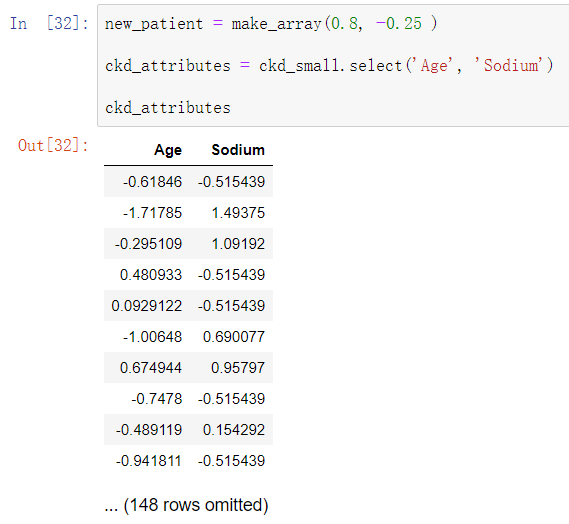
b)

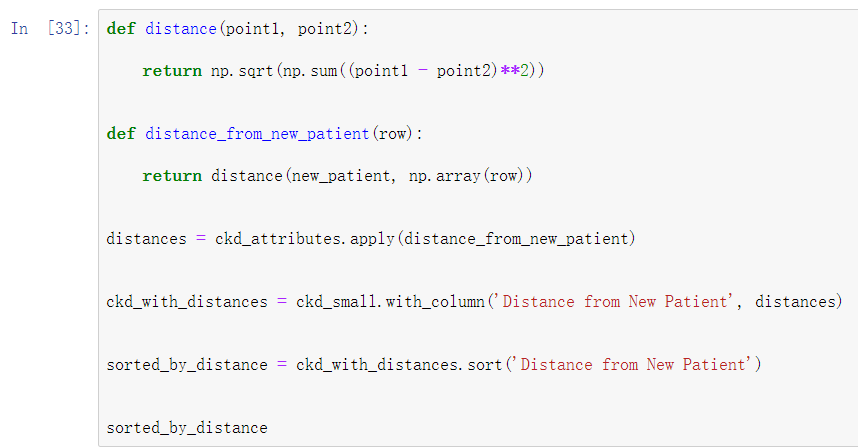


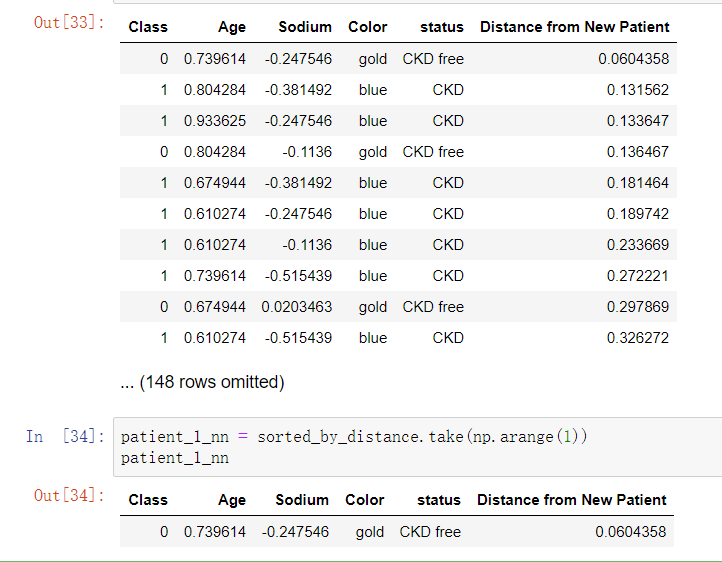




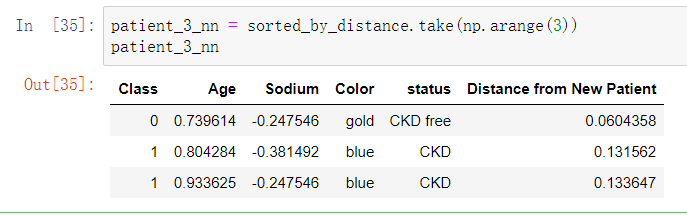
c)



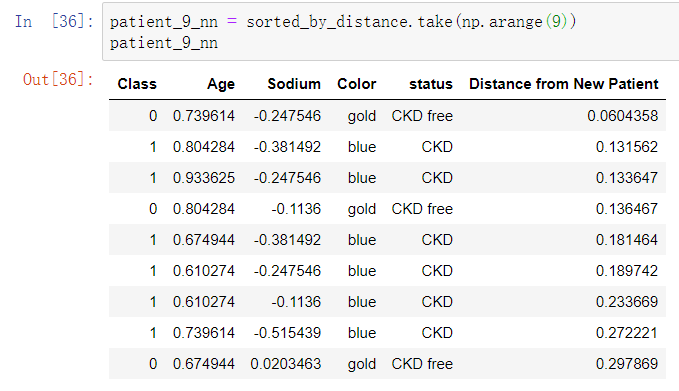




d)

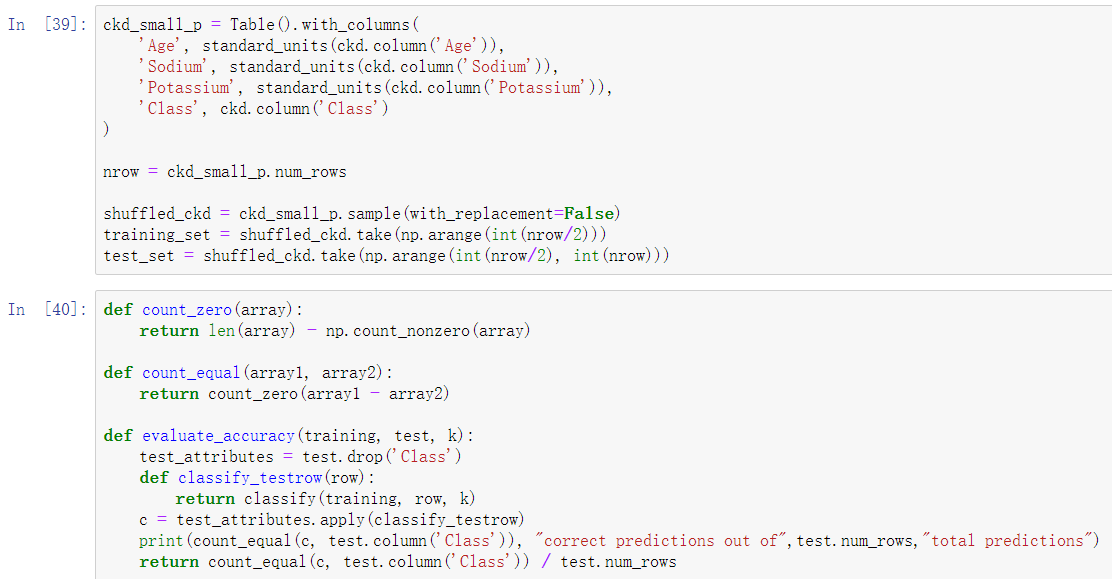


e)

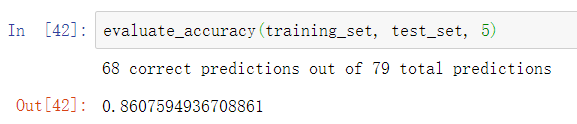


f)







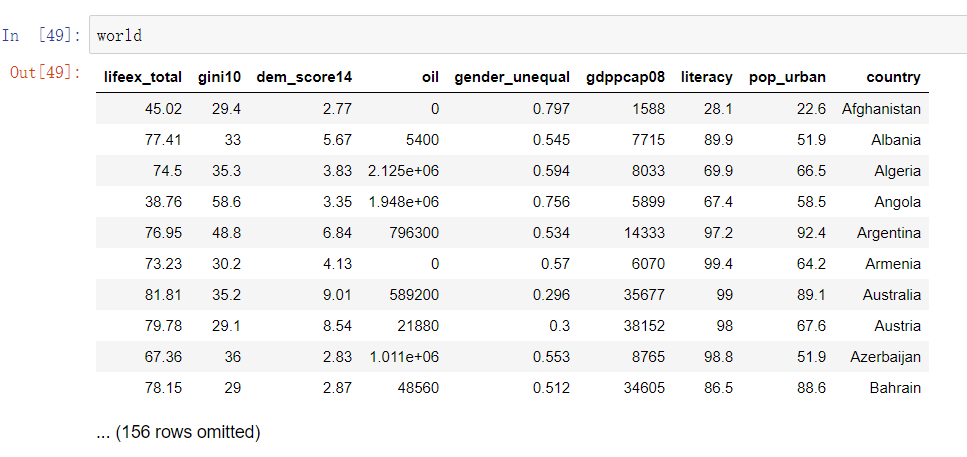


Knn classifier made 68 correct predictions out of 79 total predictions.

g) The accuracy is 0.8607594936708861.

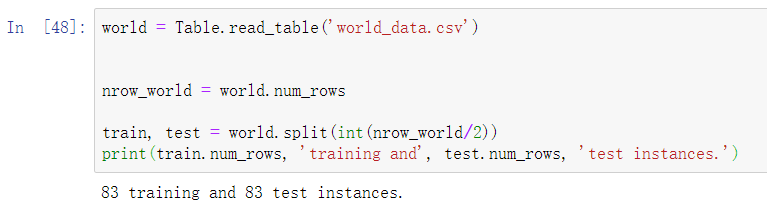
h) 

1. a)



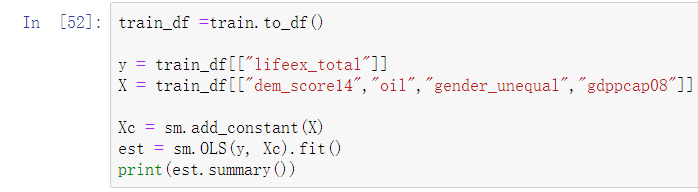
Albania is the second country in the data

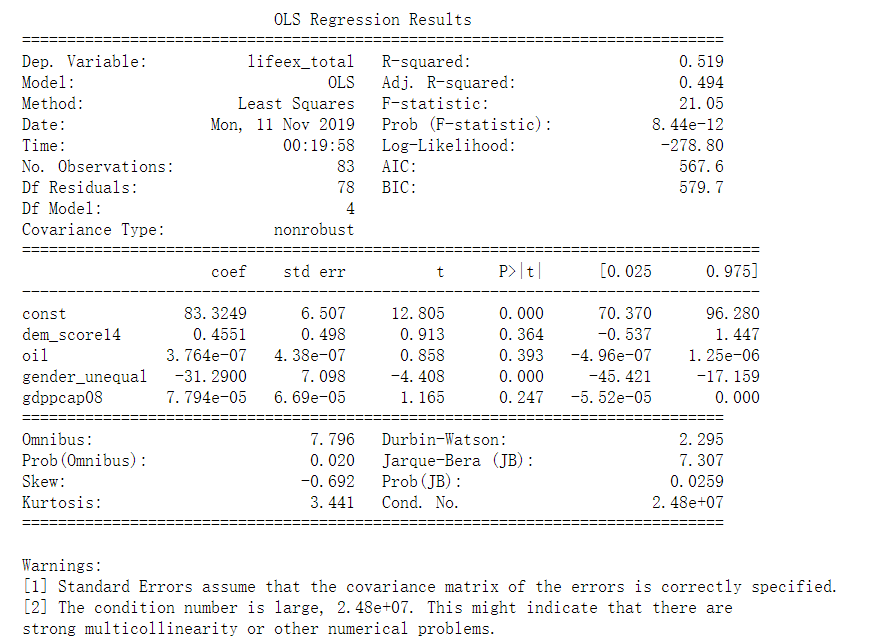
b)



There are 83 observations in each.

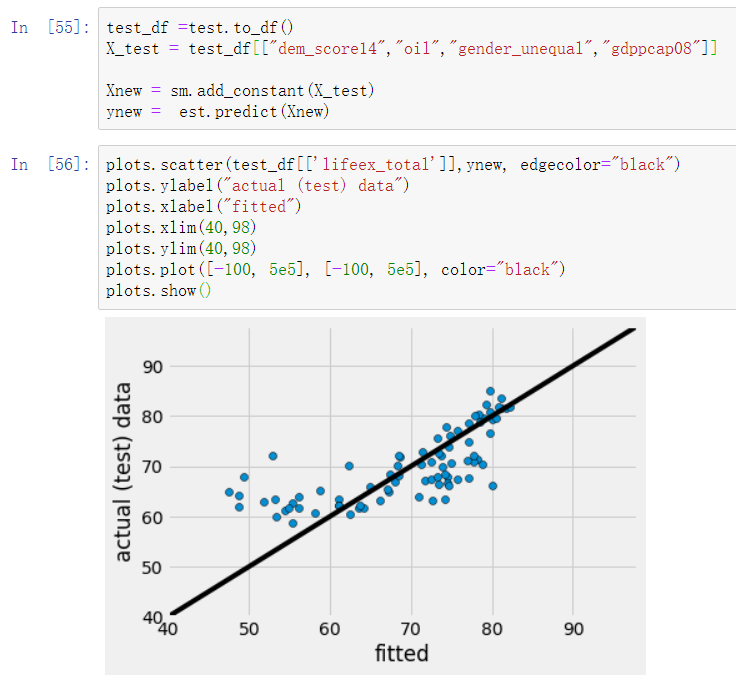
c)



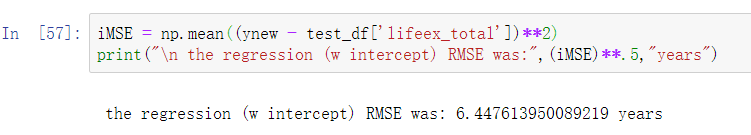


d) For each unit increase in GDP per capita, predict life expectancy in years would increase by 7.794e-05.

e)



The model does well between 60-80 years LE, but not so well between 40 to 60.

f)

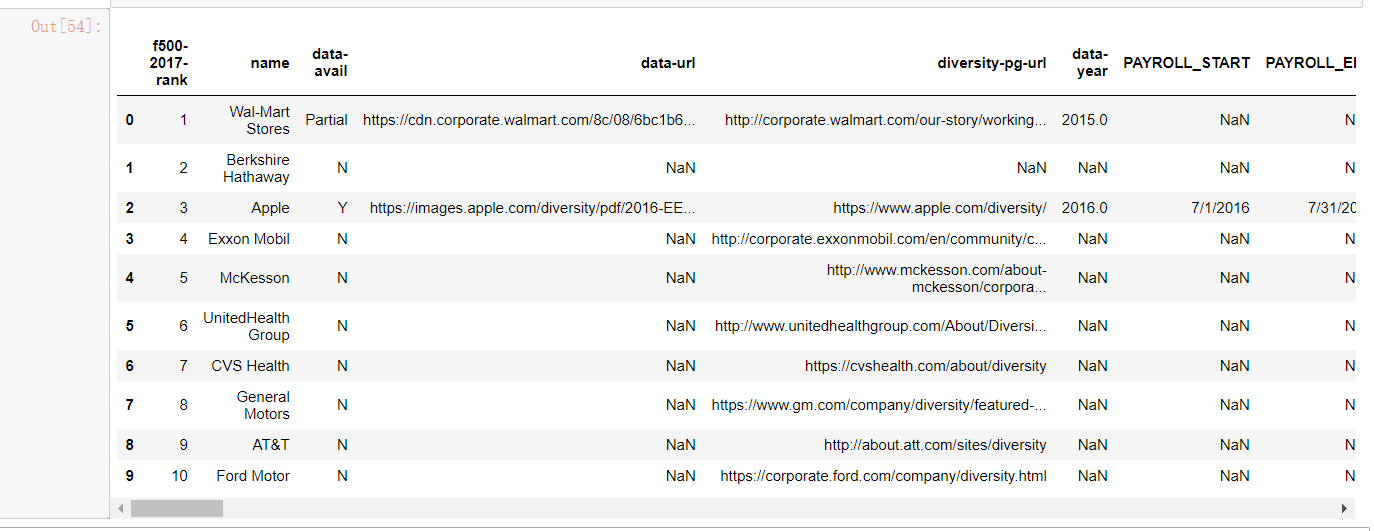
g)



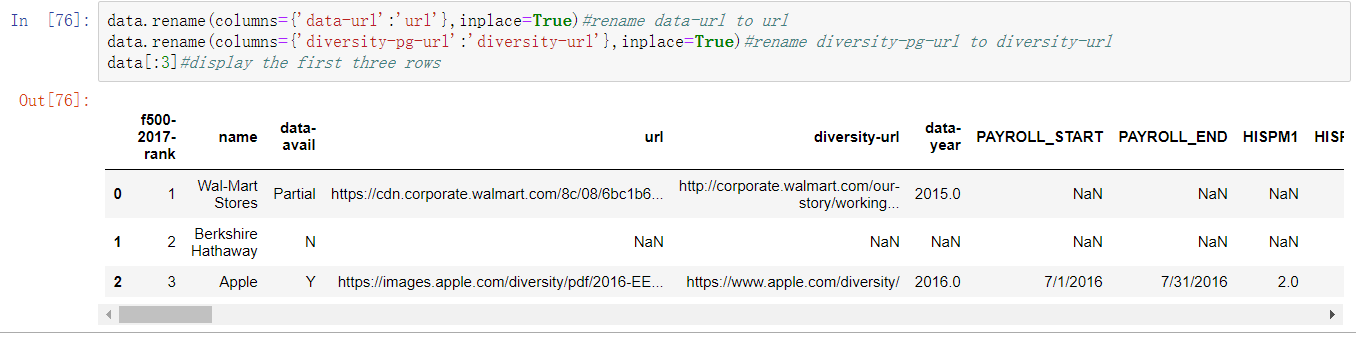
It fits worse than the regression.

* 1. In general, we expect the probability of a positive test result given the disease to be higher than the probability of having the disease given a posiive test result.
  2. Pr(D|test+)= Pr(test+|D)\* Pr(D)/ Pr(test+)=0.95\*0.001/0.02=0.475.
  3. Pr(not D|test+)=1- Pr(D|test+)=0.525

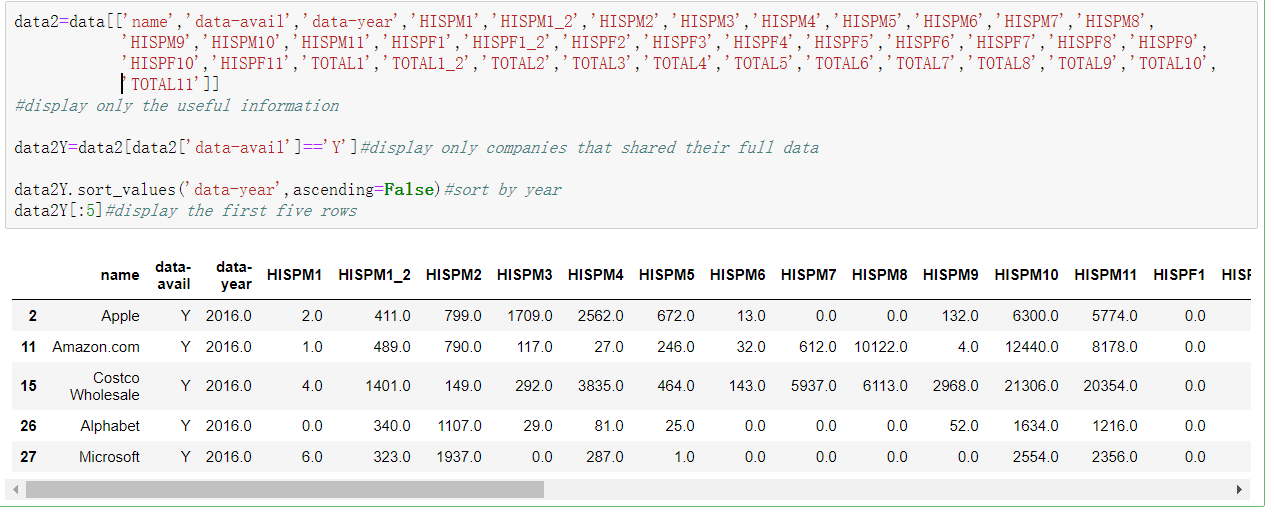
1. a) 



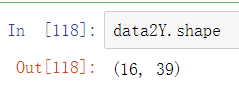
b)



c)

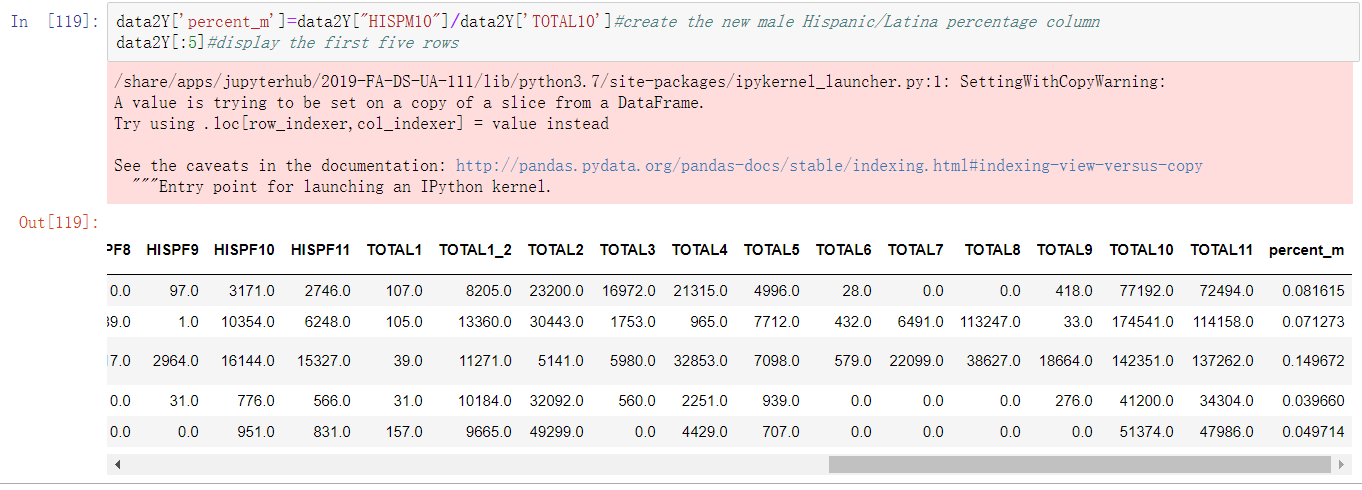


d)

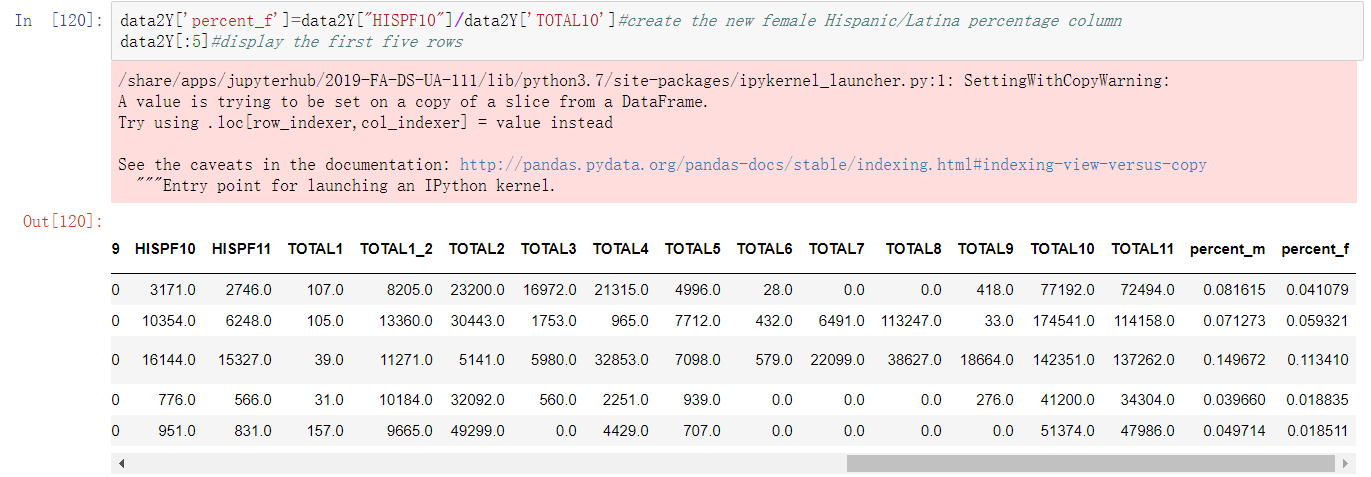


There are 16 rows and 39 columns.

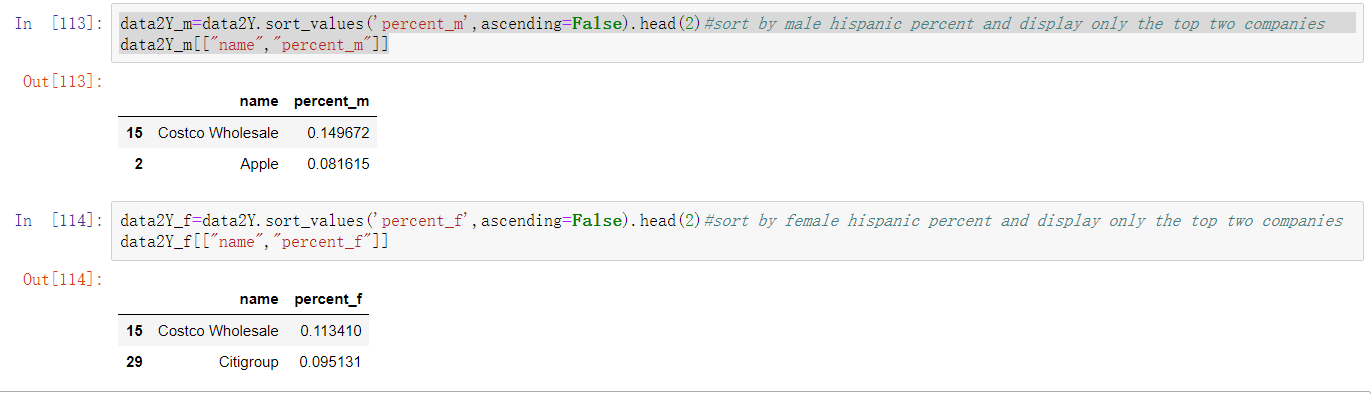
e)



f)



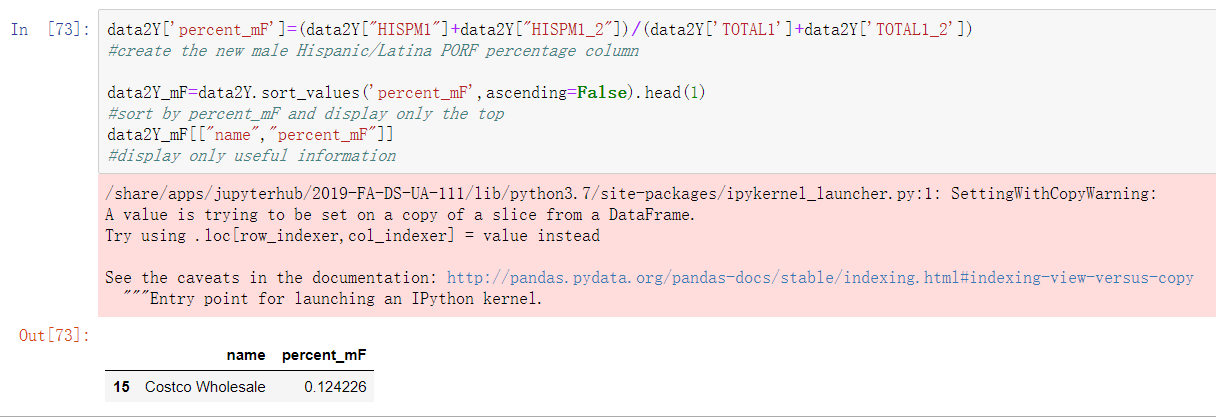
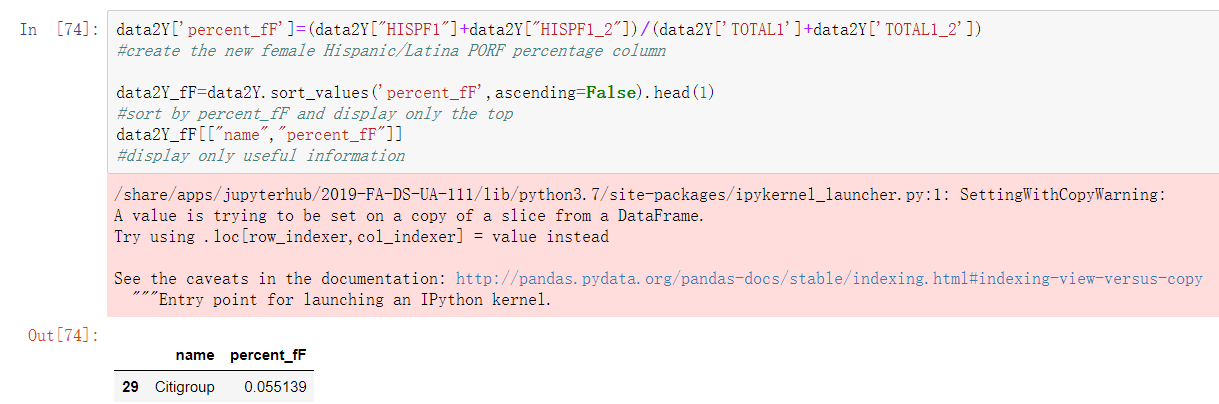
g)



Costco Wholesale and Apple has the highest percentage of male Hispanic/Latino employees, with basically 14.9672% for Costco Wholesale and 8.1615% for Apple. Costco Wholesale and Citigroup has the highest percentage of female Hispanic/Latino employees, with basically 11.3410% for Costco Wholesale and 9.5131% for Citigroup.

h) The two top employment levels should be 1-Senior OFF AND MGRS and 1\_2 Mid OFF AND MGRS. Percentage of employees in these two levels best capture “leadership” because it displays Hispanic role in professional level.

i)

Costco Wholesale has the highest representation Hispanic/Latino male in leadership with a percentage of basically 12.4226%. Citigroup has the highest representation Hispanic/Latino female in leadership with a percentage of basically 5.5139%.

j) Selection error and statistical error might influence. Selection error shows the sample we choose might not be representative for the true population. Statistical error shows that there might be recorded incorrectly that might influence the result.

k) If I have unlimited resources, I could take those companies that I currently don’t have enough data into account. I could study the population in stead of only companies that provide all data. The limitation of my method is it takes much more time to study the whole population instead of taking a sample.