Update-07/13

1.Calculating Significance

This figure shows the features that have significant relationship (p<0.05) with aging. In following process, I calculate the Biological Age with these features. Some meaningless features like when the interview was made were kept, but they weren't typed into certain categories or employed to calculate the Biological Age.

The text in this figure may be too small, you can zoom this PDF in the browser. The text labels in the upper subfigure are the codes in CLHLS, while the lower subfigure's text labels are their description.

TRUEAGE: Pearson Corr vs -logP

TRUEAGE: Pearson Corr vs -logP

Then I calculated the significance between sex and other features to determine if I need to calculate the Biological Age depending on different sex. This is the figure.
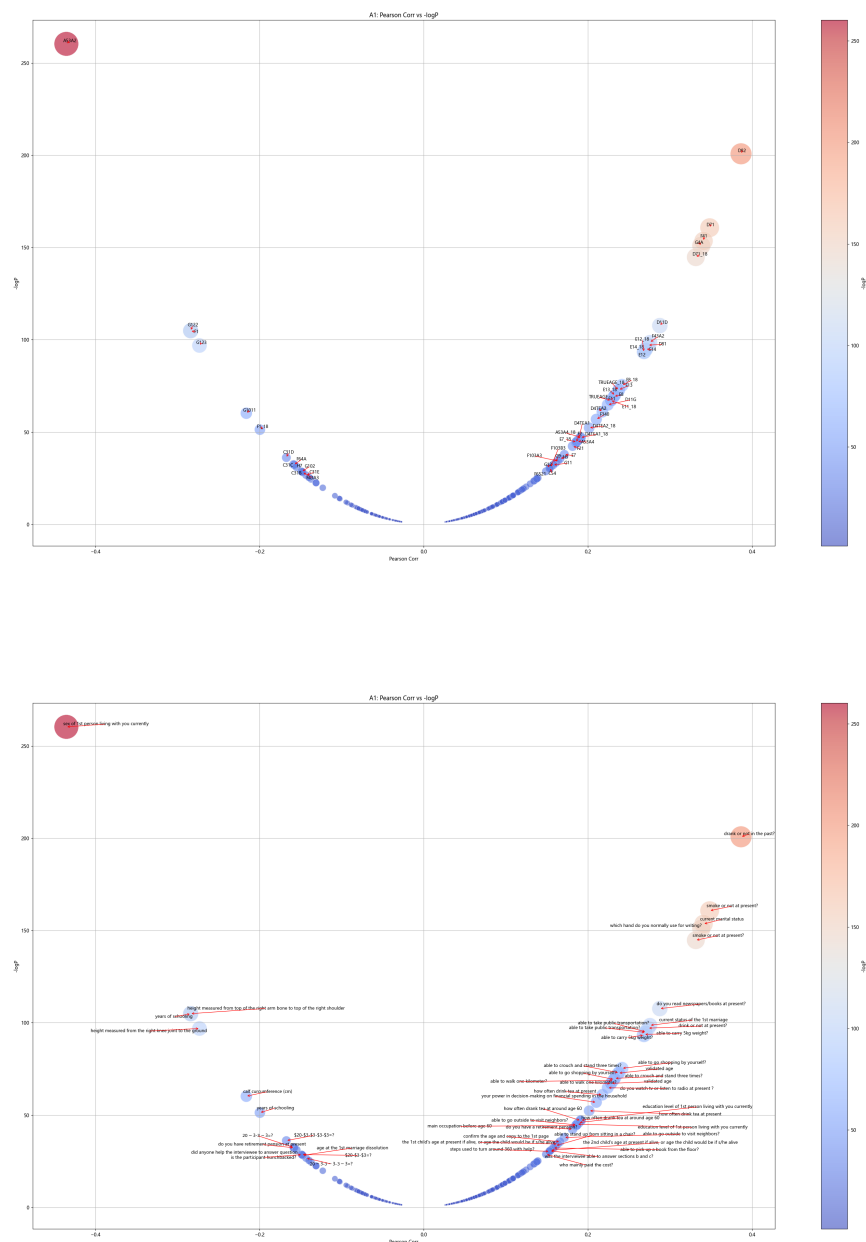
Some features show extreme correlation with sex, but sometimes they are meaningless. Like 'the sex of 1st person living with you currently'. Gladly, after comparing, these features don't have any correlation with age. Ignoring these features, the others really differs by sex.

Totally 136 features were kept because of their significance with age and 126 of them were typed into 11 categories.

Category	Feature Number
Diseases and Medicare	16
Fitness	14
Lifestyle (Entertain and Exercise)	29
Cognition	24
Diet	26

Category	Feature Number
Psychological Condition	5
Cardiovascular Condition	2
Drinking or not	2
Smoking or not	2
Teeth	5
Vision	1

2.Choosing Method to Calculate Biological Age

By reading relevant articles, I learnt that there are 3 main methods to calculate Biological Age. These methods function well on NHANES which covers interviewees of all age and more quantitative biochemical features. However, CLHLS only provides limited features of the elderly and mostly not quantitative, so some methods don't function on it.

Homeostatic Dysregulation(HD)

This method doesn't need the true age of the interviewees to fit the biological age function of this group. It provides someone's Biological Age by calculating Mahalanobis Distance of this one's biomarkers from those of a young and healthy model.

However, choosing some elderly in CLHLS as the standard model seems improper and subjective. I abandoned this method.

PhenoAge

PhenoAge is based on the death rate of interviewees to calculate their Biological Age. In CLHLS(2014~2018), if the elderly died during these years, that seems he had a higher Biological Age while those remain alive had the lower. I tried this method and got a barely useful outcome. But I didn't choose it due to these reasons:

1. Death is complex and random especially for the elderly. A simply fall down can lead to death, and CLHLS doesn't provide the death reason of the elderly to drop those dying of casual death, so there can be too many outliers.

2. This method functions well on NHANES and NHANES covers interviewees of all age. If a young person died naturally, we can say his lifestyle and hereditary factors makes his Biological Age very high. But an 100-year-old person passed away, it's not proper to say his lifestyle and other factors is not positive for his lifespan.

Klemera-Doubal Method(KDM)

This method was used in the given article. KDM gets Biological Age by the regressions of series of biomarkers. The main limitation of employing KDM on CLHLS is that CLHLS covers only the elderly and the features are not sufficient.

But compared with HD and PhenoAge, KDM is the only workable choice. The following processes are based on the Biological Age calculated by KDM.

KDM also has its limitation. For calculating Biological Age, my process must have some shortcomings. Hope we can discuss this point at the next meeting.

3.Calculating Relationships between Biological Age and True Age

In the given article, it makes use of an open-source project BioAge to employ KDM. It's set as an R-script lib, so its expandability for datasets other than NHANES is not good. I modified some codes to make it function on our cleaned dataset.

Some categories have too few features (like Vision and Teeth, only 1~3 features), their R-squares are lower than 0.1. Other categories (like Diseases and Medicare) show R-squares lower than 0.1, it implies calculating Biological Age by KDM with categories of these features is not proper though these factors can really effect how fast someone ages or their health condition in reality.

This figure shows the only 4 categories with which KDM provides the acceptable Biological Age.

What's to focus on was the Cognition Category, in spite of such R-square, we can find someone's Biological Age is too high and everyone's Biological Age is always higher than a certain line.

I think people with too high Biological Age can be totally dementia, they can not answer any question with logic. They are severely different from those can answer question properly who are considered to be biologically young. After checking the data of those of too high age, I found they can't answer any questions correctly. It's consistent with my conjecture.

As for the lower limit line or the asymptote, I think it's due to a similar conjecture. Being unable to tell the names of daily foods is a severe sign of dementia. Few elderly can reach that stage when alive. Since most people can answer all the questions correctly and absolutely most interviewees are above 60, there can naturally be the lower limit.

As plots shown in the first subfigure, fitting them with a quadratic curve can be a better choice. People can get dementia or some cognitive diseases like AD easily when they get older, and this tendency actually accelerates when getting older. This general knowledge is also consistent with the quadratic curve conclusion.
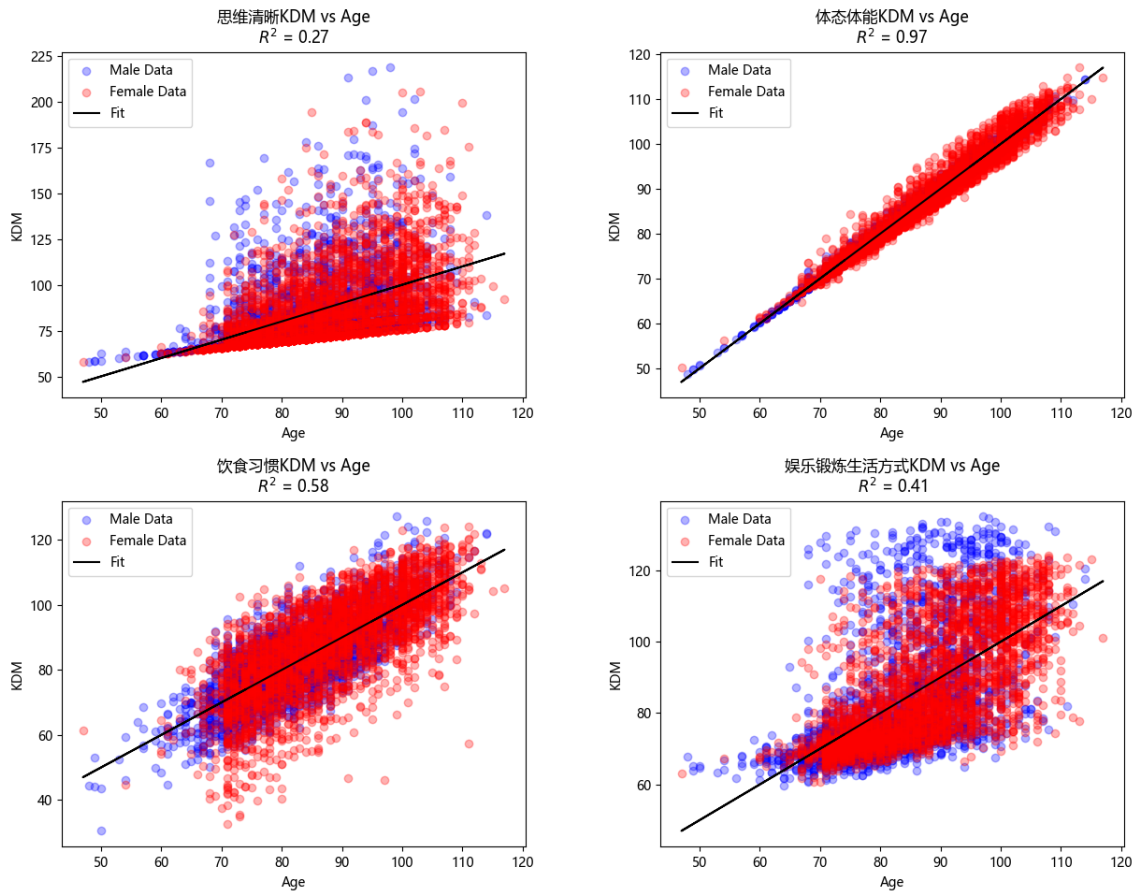
4.Distribution of Aging Rate by Different Categories

Aging Rate is calculated by this:

$$AgingRate = \frac{BiologicalAge - TrueAge}{TrueAge}$$

Before showing the correlation heatmap of these four categories' aging rate, here is a figure showing the distribution of Aging Rate of different categories. It's another kind of visualization that shows the effect of calculating Biological Age by different categories. Because of larger skewness, cognition category can lead to higher Biological Age, so as lifestyle. Fitness and diet can make people biologically younger.
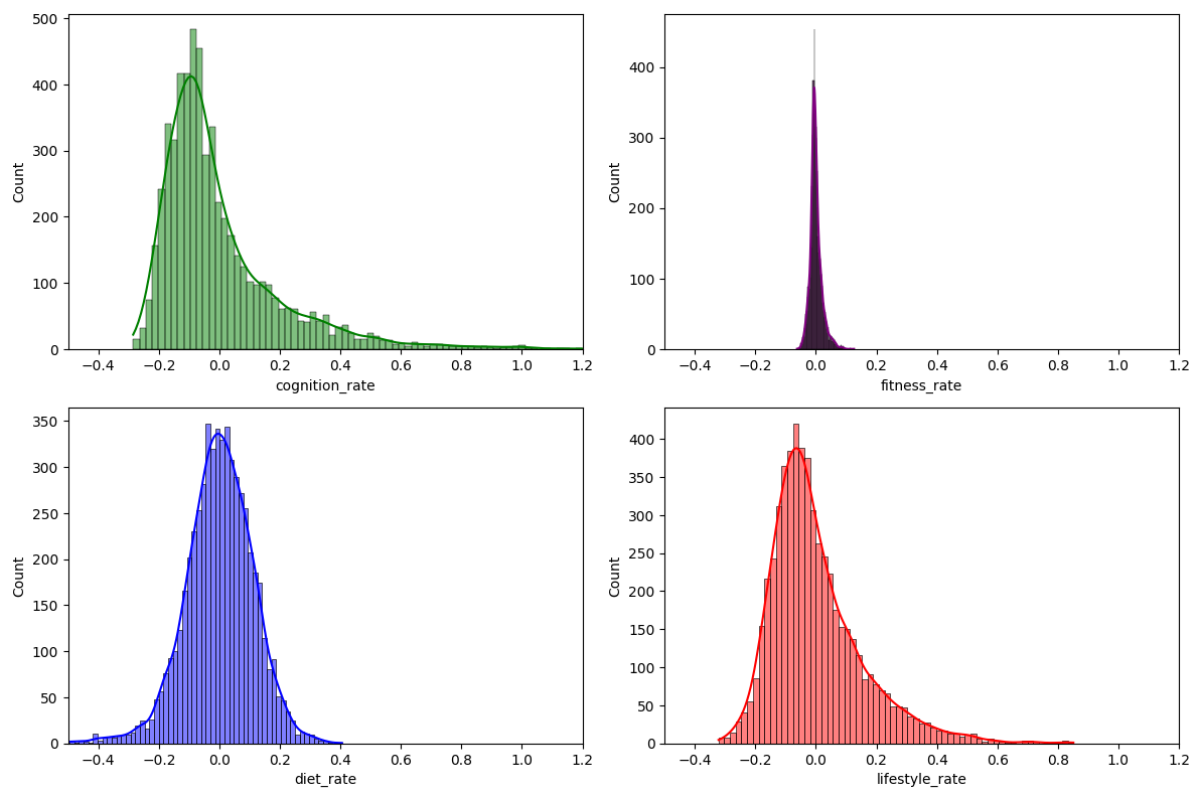
As for diet, the features are just the frequencies how interviewees have some food. And larger numbers means they have fewer food of this kind, so this chart actually shows the effect of slowing down physiological aging. The types of food are not specific, and so called junk food are not included, so my conclusions may be not convincible. At least it seems the protein intake (meat, fish, egg, milk) and the combination of different sources of protein really matter when slowing down the aging process. Fibers, pickles or oil types are not significant here.

```
与抗衰老关系最密切的列及其相关系数:
60岁_蘑菇或藻类           0.150444
60岁_维生素 (a/c/e)      0.147575
60岁_药用植物            0.144071
目前_蘑菇或藻类           0.122604
60岁_坚果制品           0.120587
60岁_豆类食品           0.116030
60岁_奶制品            0.113248
60岁_鱼              0.100244
60岁_鸡蛋             0.099272
目前_药用植物            0.094711
60岁_肉              0.091387
60岁_水来源           -0.094200
Name: rate, dtype: float64
潜在的协同作用列:
60岁_蘑菇或藻类: ['60岁_奶制品', '60岁_坚果制品', '目前_蘑菇或藻类', '60岁_蘑菇或藻类', '60岁_药用植物']
60岁_维生素 (a/c/e): ['60岁_奶制品', '60岁_维生素 (a/c/e)', '60岁_药用植物']
60岁_药用植物: ['60岁_蘑菇或藻类', '60岁_维生素 (a/c/e)', '目前_药用植物', '60岁_药用植物']
目前_蘑菇或藻类: ['目前_奶制品', '目前_坚果制品', '目前_蘑菇或藻类', '60岁_蘑菇或藻类', '目前_药用植物']
60岁_坚果制品: ['60岁_坚果制品', '60岁_蘑菇或藻类']
60岁_豆类食品: ['60岁_肉', '60岁_鱼', '60岁_鸡蛋', '60岁_豆类食品']
60岁_奶制品: ['60岁_鸡蛋', '60岁_奶制品', '60岁_蘑菇或藻类', '60岁_维生素 (a/c/e)']
60岁_鱼: ['60岁_肉', '目前_鱼', '60岁_鱼', '60岁_鸡蛋', '60岁_豆类食品']
60岁_鸡蛋: ['60岁_肉', '60岁_鱼', '60岁_鸡蛋', '60岁_豆类食品', '60岁_奶制品']
目前_药用植物: ['目前_蘑菇或藻类', '目前_药用植物', '60岁_药用植物']
60岁_肉: ['60岁_肉', '60岁_鱼', '60岁_鸡蛋', '60岁_豆类食品']
60岁_水来源: ['60岁_水来源']
```

The antagonistic effects also exist. In CLHLS, drinking boiled water is marked as 1 while unboiled as 2, so we find that drinking unboiled water makes people physiologically older. However, intaking nutritious food can counteract the effect of drinking unboiled water. This can be explained by better nutrition level equips people with better immunity to negative conditions in water, including harmful mocrobes, especially for the elderly.

```
潜在的拮抗作用列：
60岁_蘑菇或藻类：['60岁_水来源']
60岁_奶制品：['60岁_水来源']
60岁_鱼：['60岁_水来源']
60岁_肉：['60岁_水来源']
60岁_水来源：['60岁_肉'，'60岁_鱼'，'60岁_奶制品'，'60岁_蘑菇或藻类']
```

5.Calculating Correlations between Different Categories

This heatmap shows several conclusions. The Aging Rate calculated by Category Diet has almost no correlation with others. The other 3 categories have high correlations between each other. Fitness and Lifestyle has the highest correlation because they are sometimes complementary to each other. The aged one with weak body can't work out naturally, while those have the habit to work out **or just not too old to remain the ability to do something** should have a better body condition. As for Cognition, clear cognition is the prerequisite for maintaining the lifestyle that includes exercise and entertainment.

Aging Rate Correlation Heatmap