In part 3 clustering method , I would introduce the method I use.
Hierarchichal clustering is the method I used in this project.
DTW distance is a distance measure I used to measure distance between time series
——

In hierarchical clustering, we need to define the distance between two object and the distance between to cluster.Interestingly ,we do not need coordinate in this method.

In this project , I choose DTW as distance measure ,and complete linkage method as cluster distance then apply the Agglomerative clustering.

complete linkage is one of the way to calculate cluster distance, and It’s defined by this formula.
In this formula we can see that, given two

Agglomerative clustering follow this simple step to work.
——
In this slide I’ll show you the step of hierarchical clustering.

Suppose we want to cluster these six data into two group.

In the beginning ,all points belong to its own group.Then in every step,we merge the closest two cluster.

In this case we choose two closest cluster and merge
In this case we choose two closest cluster and merge

In this case ,each cluster has two data, so we would apply the complete linkage method to calculate the distance between cluster.(For each element in two cluster ,we find the farthest distance)
——
And the raw-based-method I is a simple way to solve time series clustering task.It consists of two step. The first step is define the distance measure.The second step is apply any clustering method you know.
It is noteworthy that In this method distance measures are usually more important than clustering method. And It’s general for almost every domain.
——
In this slide I would talk about feartures of DTW distance and how to compute it.
Given two time series with different length, The DTW measure is defined by the following formula.
In this formula ,we can see that the DTW in i,j is defined by the distance between p,j plus the minimum of former DTW ,So one should recursively solve the formula to find DTW measure.

DTW can compare the stretched or compressed time series with different time length, which means that as DTW closer to zero, the series you compare might have the same shape no matter how one of the series is streched or compressed.
——
So after we define the distance between two time series, we can calculate the distance matrix.
——
Once we have the distance matrix we can apply the hierarchical clustering to time series.
——

In Part 4 result ,I compare the 4-group result in each method.And I present the data result in this way. In the left plot ,each color represent one group , and the darker line in each plot represent the mean of each group.

In the right plot,I present the mean v.s. standard deviation plot , and the color of the dot represent the same group information in the left plot.This dotted line is the efficient frontier of 77 mutual fund process. and this dotted line is the efficient frontier of 4 mean series process.

In the left plot we can see that each group seems have it's trend and compare to the mean vs std plot, we can see that
the pink group consist of mutual funds which had low risk and low return,
the blue group consist of mutual funds which had high risk and low return,
the red group consist of mutual funds which had high risk and high return,
the yellow group consist of mutual funds which had high risk and minus return,

But the efficient frontier for the mean portfolio is not good enough
———
So I did the following things, I form the N-Nearest portfolio with DTW-HAC method.

For example 1-Nearest portfolio is consist of sereis which is nearest to the mean series in each group.

Then I found that It can form the following plot,these dotted plot are represented efficient frontier of 1-Nearest to 18 Nearest portfolio , and it can be used to choosing portfolio in tremendous assets for specific number.