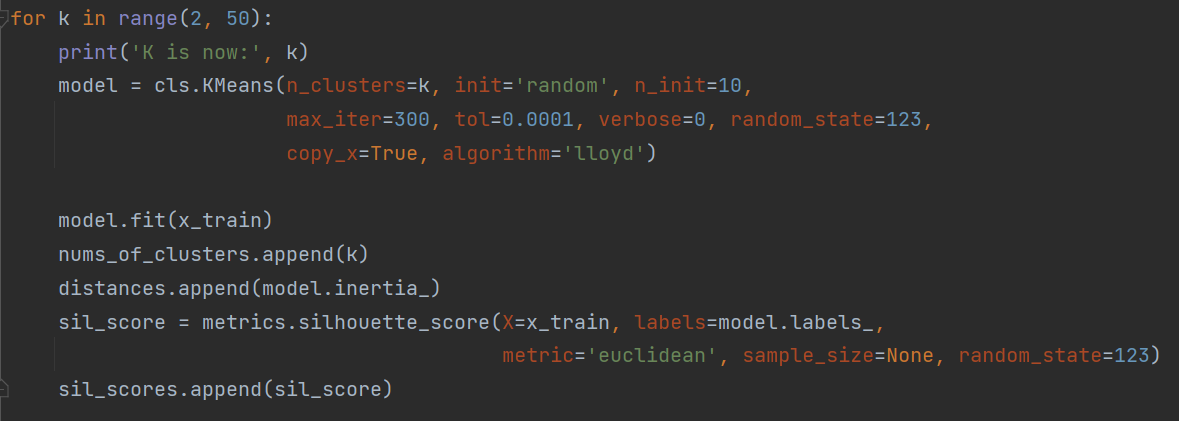
**USING K-MEANS TO CLUSTER AND PCA TO REDUCE THE NUMBER OF FEATURES IN THE DATA FOR USER’S FEELING OF YOUTUBER’S STYLE PREDICTION**

**Answer to required questions (detailed report below):**

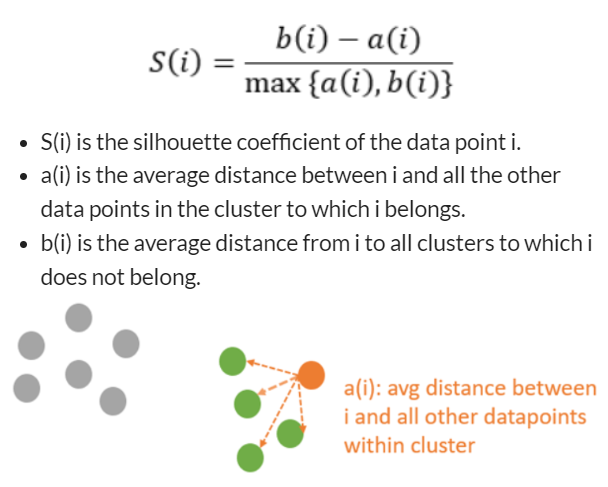
* K-Means:
  + Number of clusters to train on: from 2 to 49.
  + Method to choose k clusters: elbow and silhouette score.
  + Final result: the number of clusters chosen is k = 17 due to obvious elbow and high silhouette score.
* PCA:
  + After applying PCA, we got 13 principal components (PCs) rather than the original 15 features.
  + All 13 PCs contribute quite evenly in the variance of our data, so well that if we want 95% variance retained, we still need to use all 13 PCs.
  + Casual restaurant and glass container scenes and long food scenes with pot (presumably hotpots) play an important role in our data.
  + Casual and street restaurant short scenes, no matter the containers and scenes without box or glass containers don’t really contribute a lot in our data.
  + Based on a 2-feature PCA data plot, our data mostly don’t have fancy restaurant with glass or hotpot scenes. And for data that have either of them, it’s unlikely that those scenes get a 5/5 score in terms of viewer feeling of youtuber’s style.

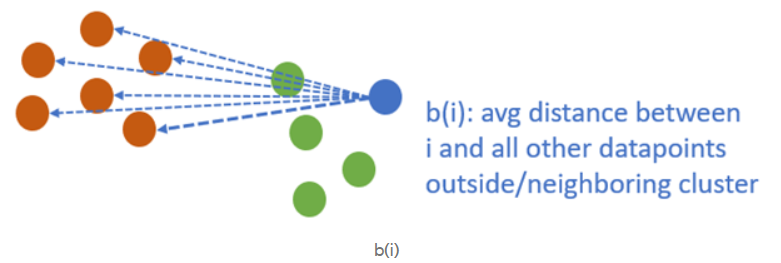
**Detailed report**

**K-Means:**



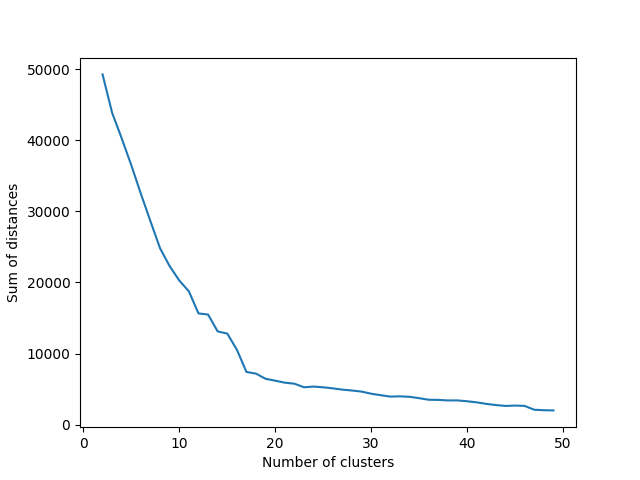
* We are going to different K-Means models based on the number of clusters for testing (from 2 to 49 clusters).
* Then, based on the sumation of squared distances of samples to their closest cluster center of each model, we’ll determine a good range of clusters to use.
* Another metric we are going to use is silhouette score, which is as following:



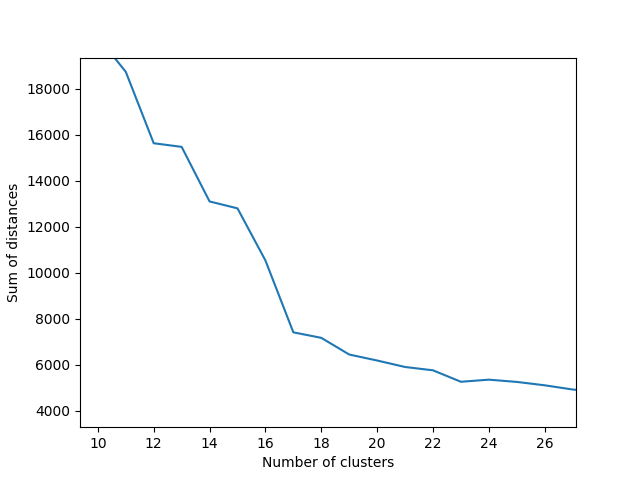


* The higher the silhouette score is, the better the number of clusters to choose from.
* Finally, we will plot out the result for easy comparision.

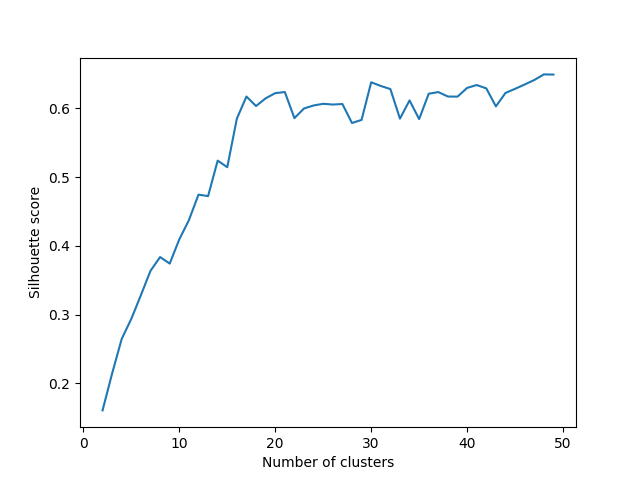
**The result of 48 K-Means models:**



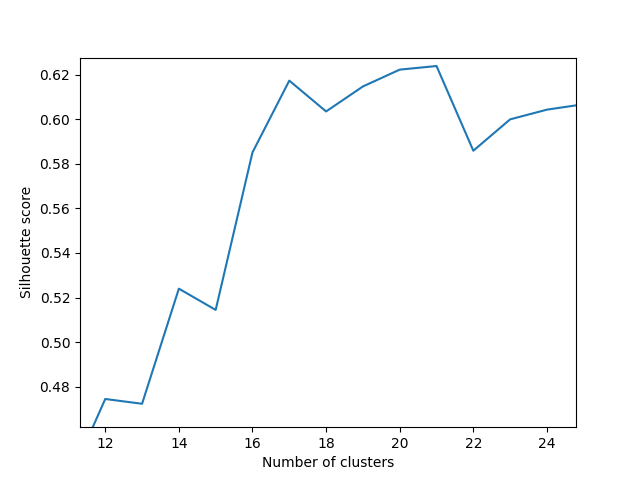
* In this image, we can see how the sum of distances rapidly decreases when our number of clusters ranges from 2 to 15-ish.
* Let’s take a closer look into this image and see where the elbow is.



* So, the elbow is this graph is obviously at 17 clusters.
* Now, let’s have a look over the silhouette score of each model we have trained.



* Although the graph can be quite the mess at the right half, we still have a valuable information: when the number of clusters is less than 15, the score is very low. Thus, the number of clusters should be larger than 15.



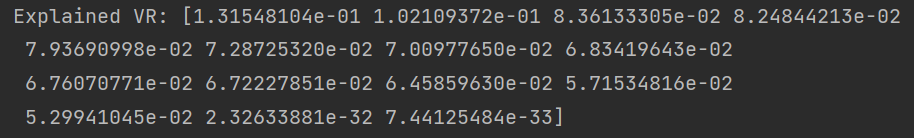
* In this zoomed image, we can see how 17 and 21-cluster models have the better silhouette scores than other models. And considering how 17-cluster model is the best model in terms of elbow evaluation and our original data has only 5 labels, we should choose 17 clusters for our K-Means model.

**PCA:**

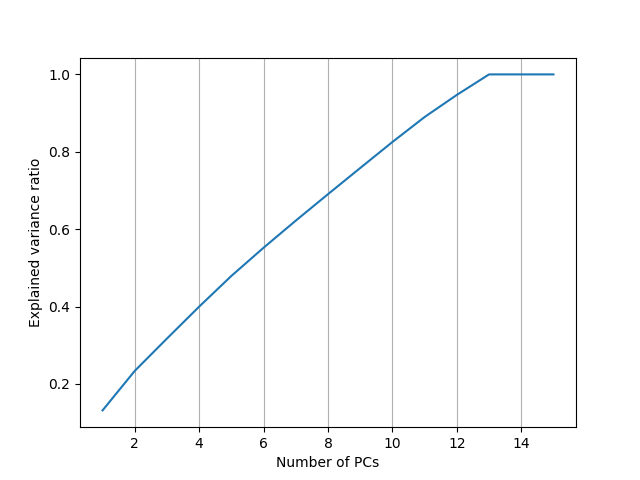
* Let’s run a quick PCA on our data with 15 features:



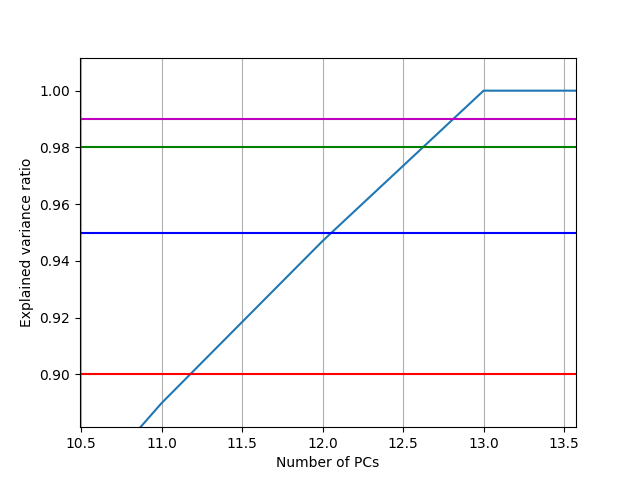
* The result:



* It’s easy to see that we should only need the first 13 principal components rather than the original 15 features.
* The other 13 principal components contribute quite evenly in our data set.
* Let’s plot the numbers for a better visuallization.

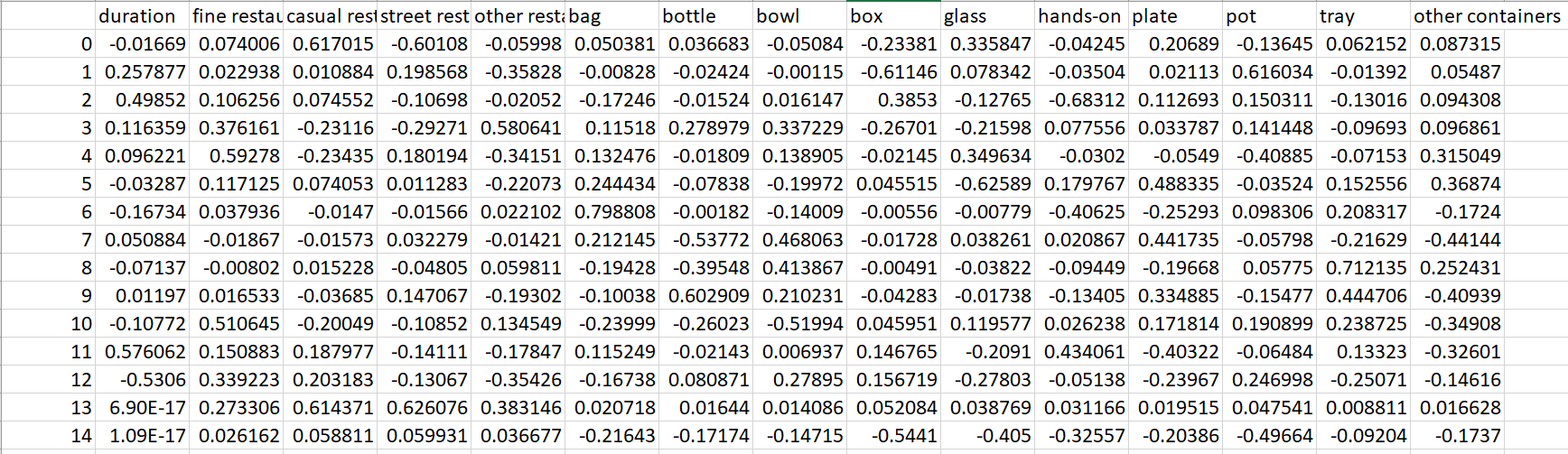


* Zoom it in a little bit and add some lines:



* If our intention was to keep over 99% variance of the original data, we should keep 13 principal components. The same goes for 98% - 13 PCs, 95% - 13 PCs, 90% - 12 PCs.

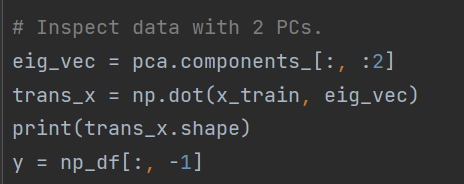
**Old and new feature relations:**



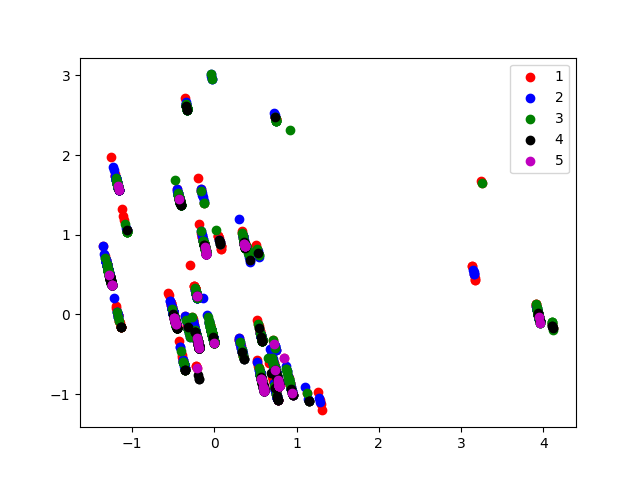
* Starting with the best feature according to variance: casual restaurant and glass container play an important role in this feature, which should capture food scenes with drinks in a decent restaurant.
* The second-best feature focuses more on the pot feature in a food scene, and some bit of duration in that scene.
* The third-best feature actually represent how most street food in Vietnam is not. This feature focuses a lot on long food scene that doesn’t use hands-on food.
* Let’s have a look at the last 2 features: it seems that casual and street restaurant short scenes, no matter the containers (feature 14) and scenes without box or glass (feature 15) containers don’t really contribute a lot in our data.

**Plot out PCA-ed data in 2 dimensions:**

* Let’s try to calculate our data in the 2 most important features (about 23% variance retained):



* Then plot this new data based on their labels.



* We can see how the samples with different labels are overlapping each other rather than the expected separable data.
* Since this is just a plot of our data with only 2 PCs (23% variance retained), we can’t not obtain much information from this plot. Except for the fact that our data mostly don’t have fancy restaurant with glass or hotpot scenes. And for data that have either of them, it’s unlikely that those scenes get a 5/5 score in terms of viewer feeling of youtuber’s style.