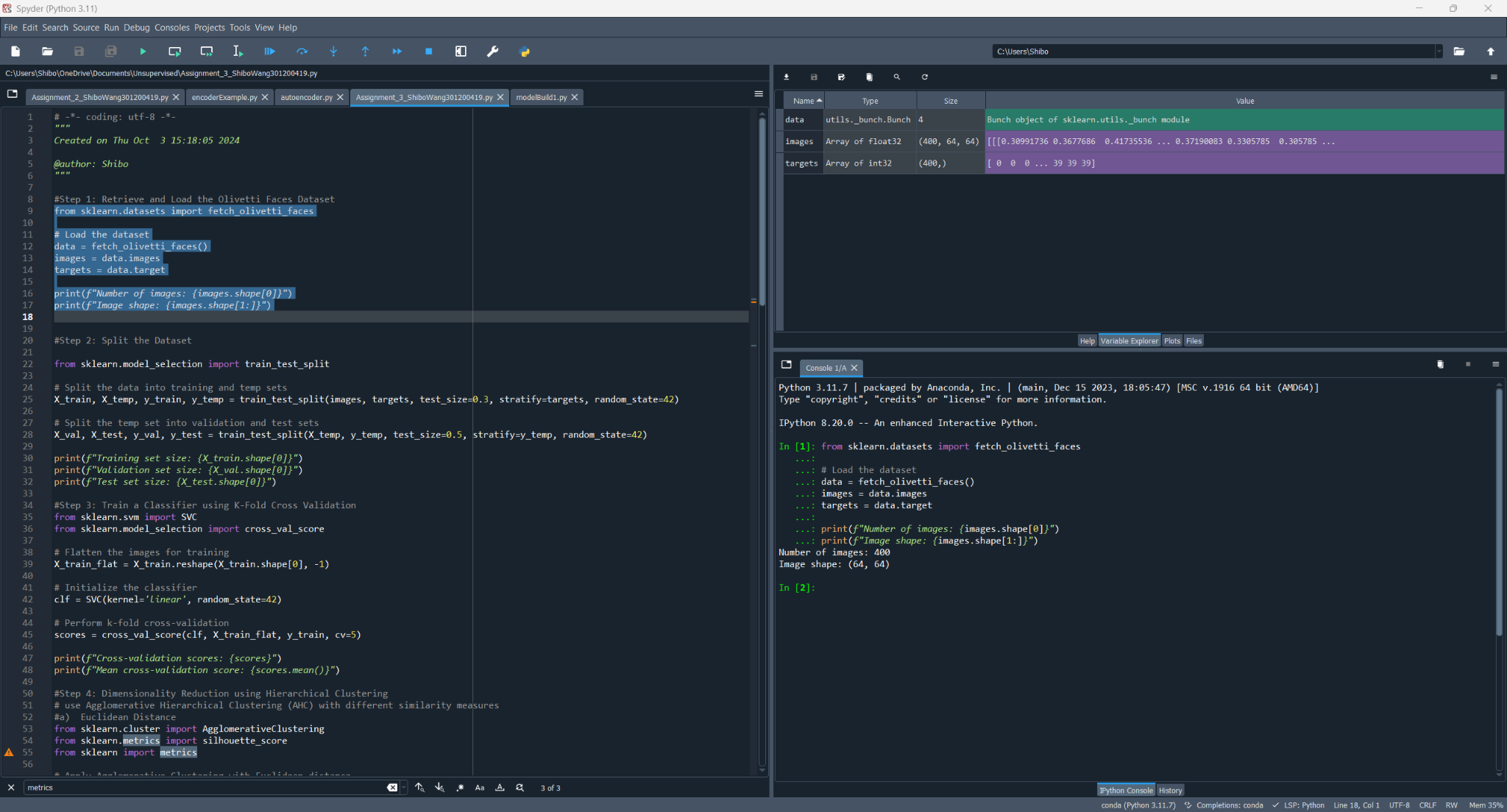
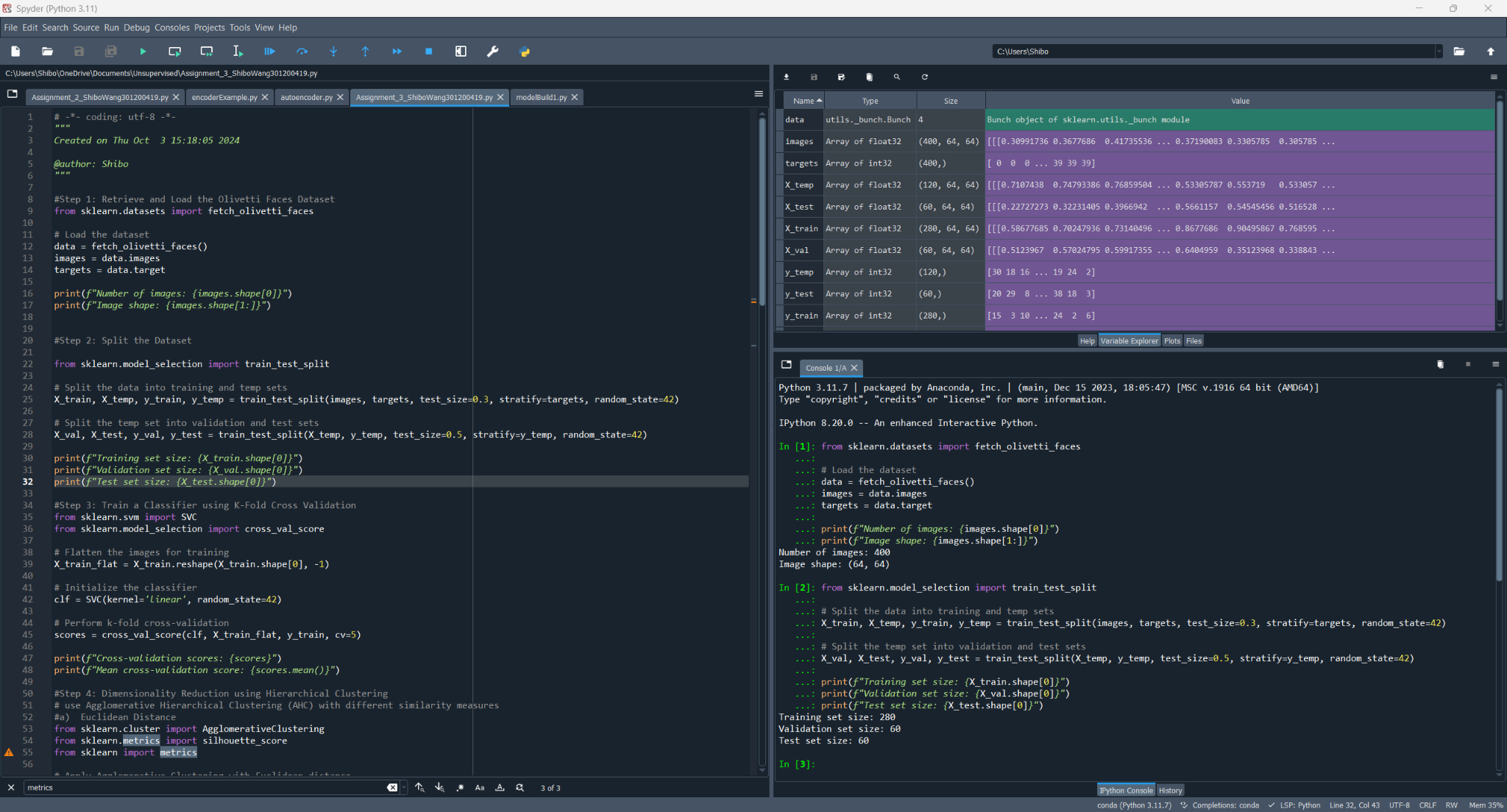
Assignment 3: Hierarchical Clustering

written reports

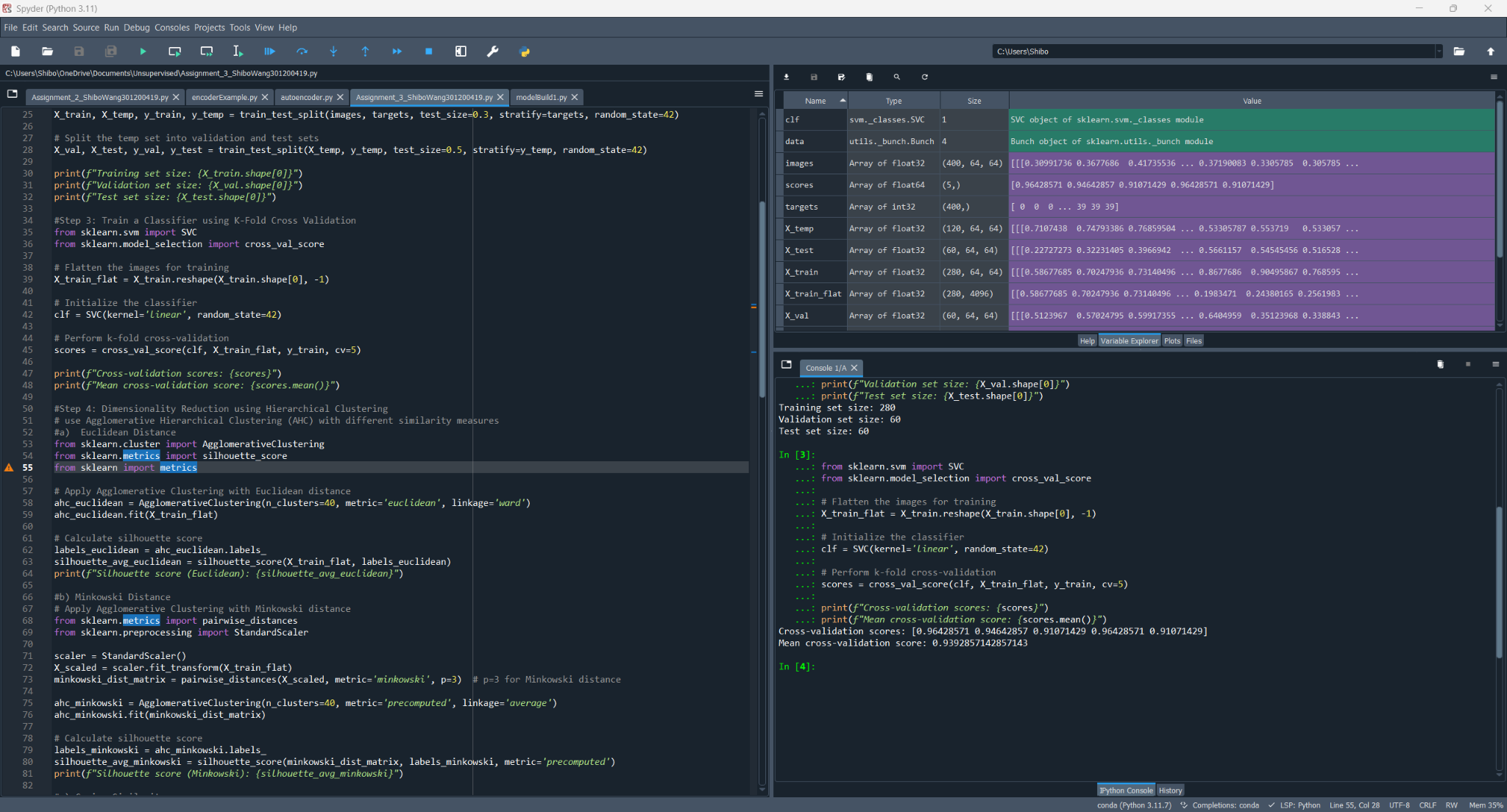
1. Retrieve and load the Olivetti faces dataset [0 points]



1. Split the training set, a validation set, and a test set using stratified sampling to ensure that there are the same number of images per person in each set. [0 points]



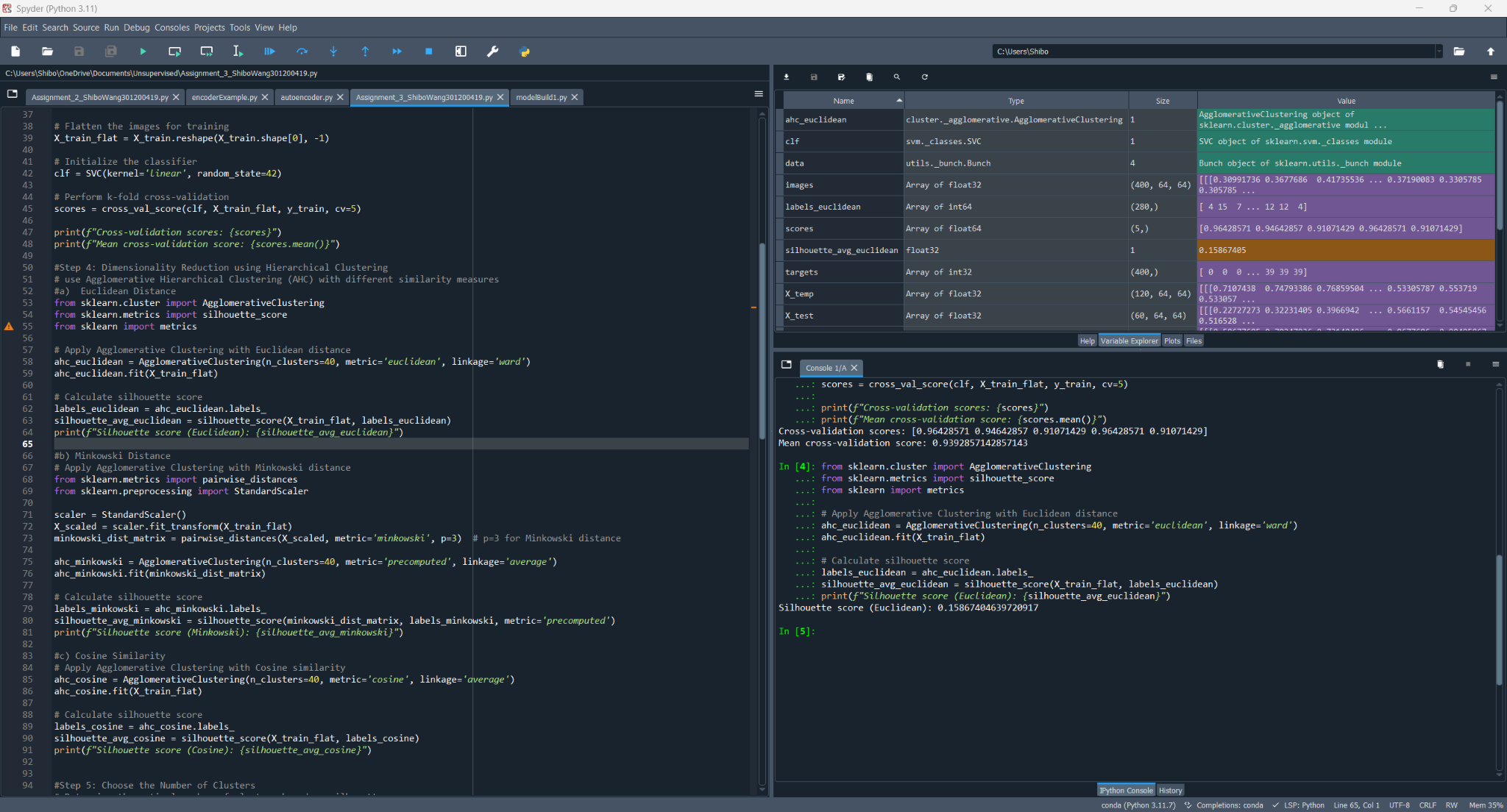
1. Using k-fold cross validation, train a classifier to predict which person is represented in each picture, and evaluate it on the validation set. [0 points]



1. Using either Agglomerative Hierarchical Clustering (AHC) or Divisive Hierarchical Clustering (DHC) and using the centroid-based clustering rule, reduce the dimensionality of the set by using the following similarity measures:

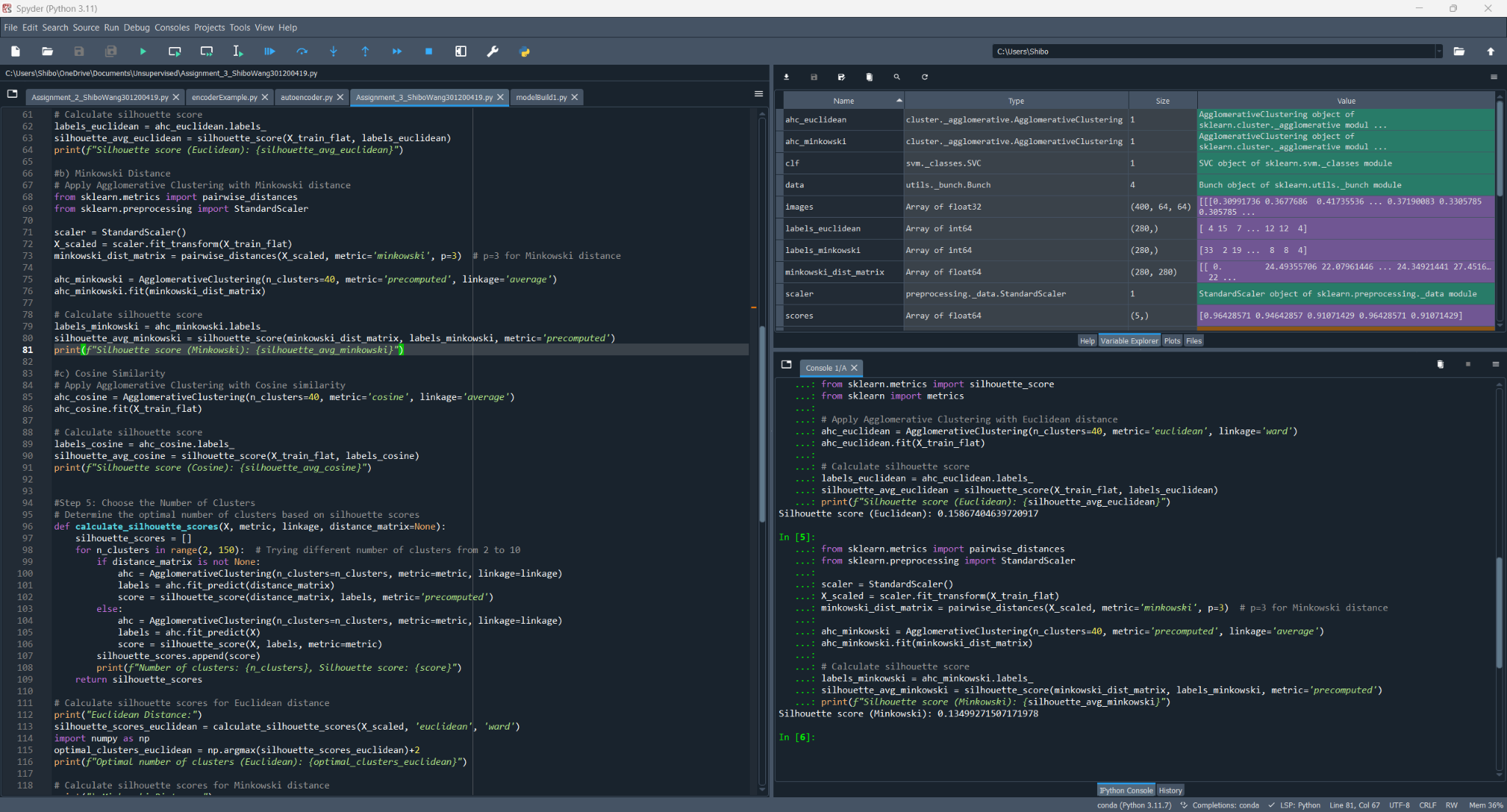
a) Euclidean Distance [20 points]

The dataset is using 40 different people’s faces, so I set the default n\_clusters=40.

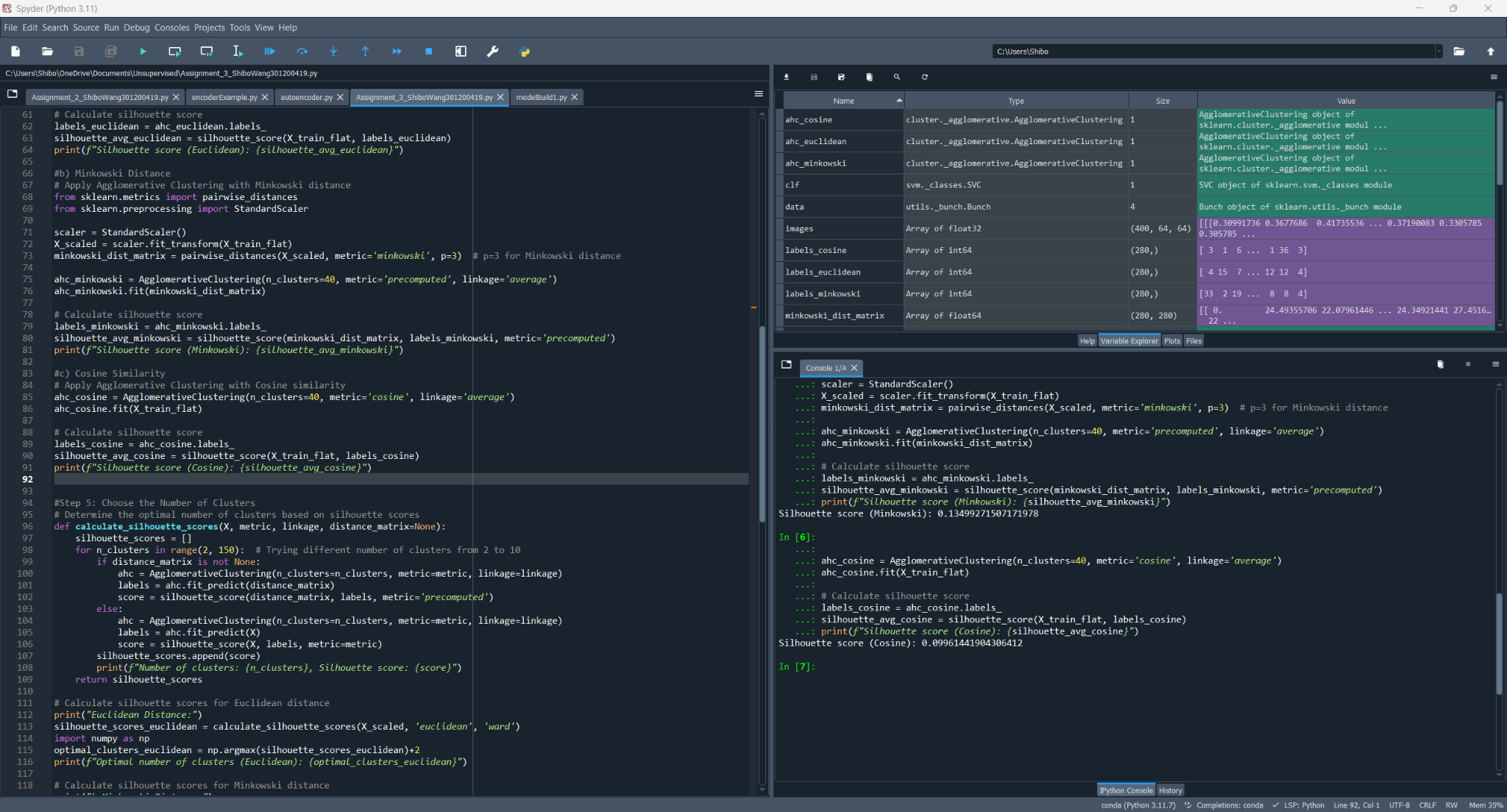


b) Minkowski Distance [20 points]

Can not set metric= ‘minkowski’, so I applied Agglomerative Clustering with Minkowski distance

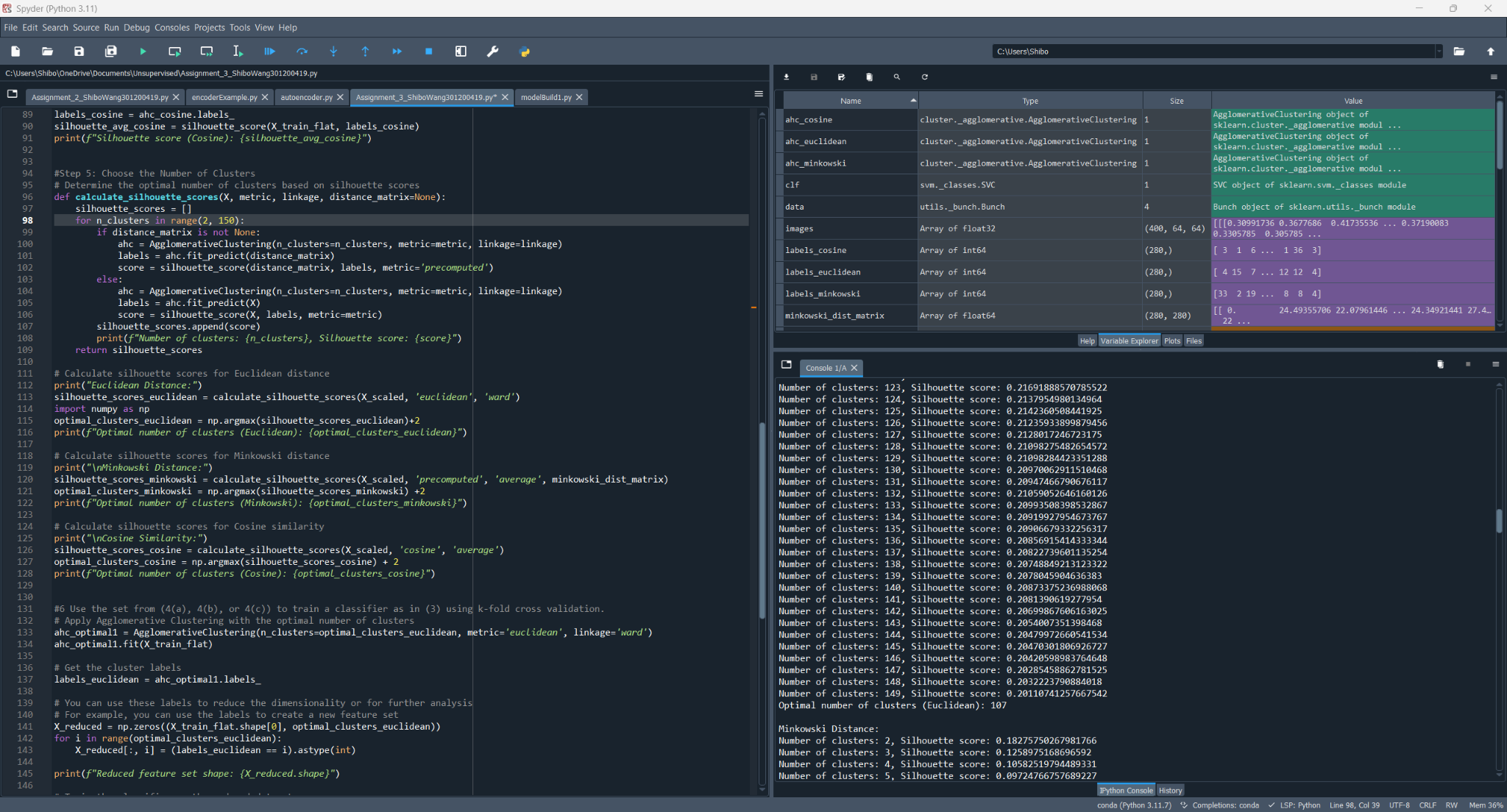


c) Cosine Similarity [20 points]



1. Discuss any discrepancies observed between 4(a), 4(b), or 4(c).

Use the silhouette score approach to choose the number of clusters for 4(a), 4(b), and 4(c). [10 points]



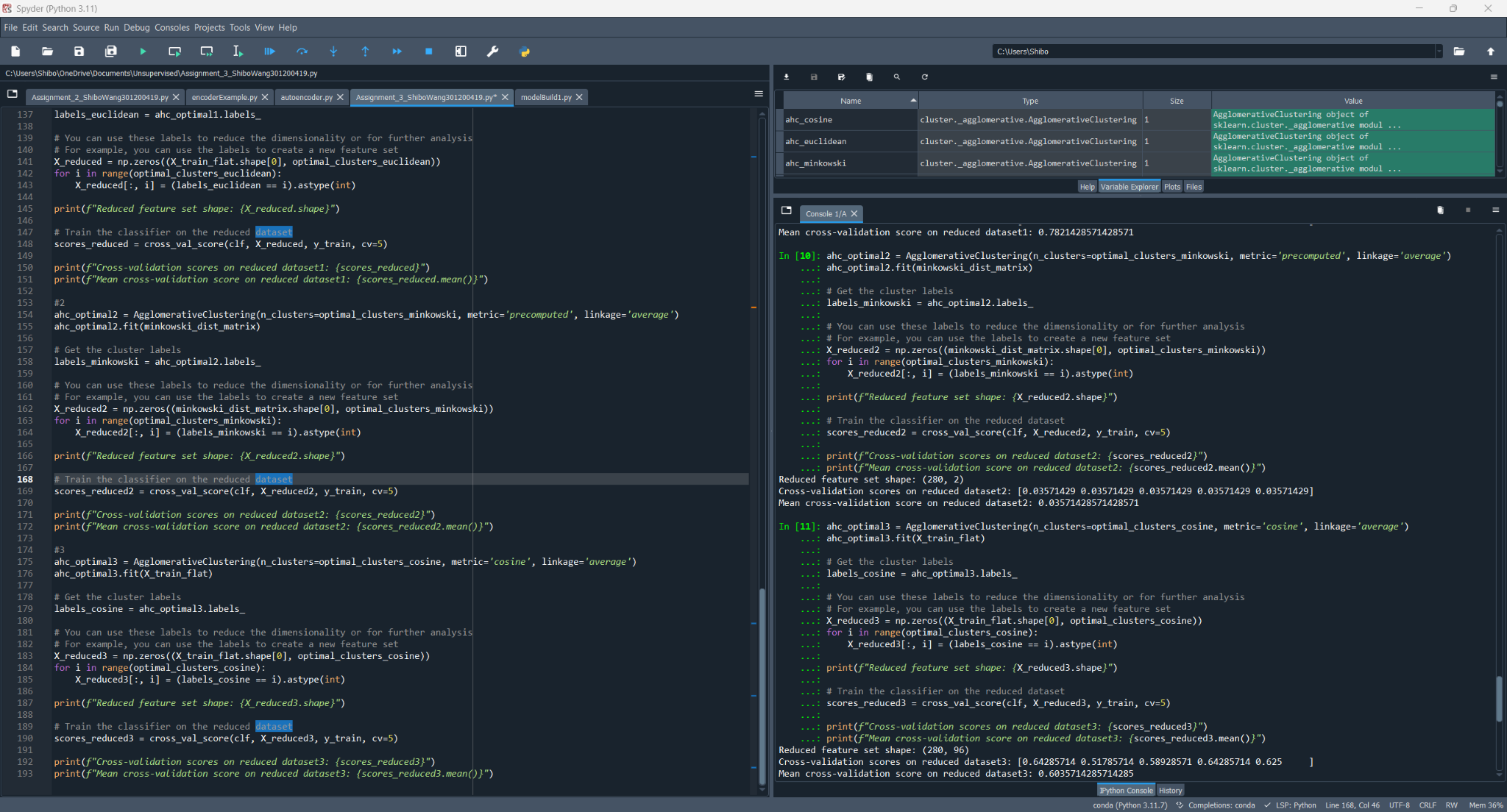
I wrote a function to calculate the best n\_clusters in range 2-150.

For Euclidean, Optimal number of clusters (Euclidean): 107

For Minkowski, Optimal number of clusters (Minkowski): 2

For Cosine, Optimal number of clusters (Cosine): 96

1. Use the set from (4(a), 4(b), or 4(c)) to train a classifier as in (3) using k-fold cross validation. [30 points]



Then I use best optimal\_clusters for each type of metrics, got the cross-validation scores:

Euclidean: Mean cross-validation score on reduced dataset1: 0.7821428571428571

Minkowski: Mean cross-validation score on reduced dataset2: 0.03571428571428571

Cosine: Mean cross-validation score on reduced dataset3: 0.6035714285714285

Euclidean works better.