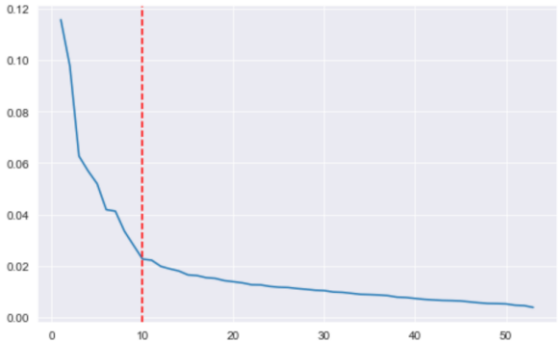
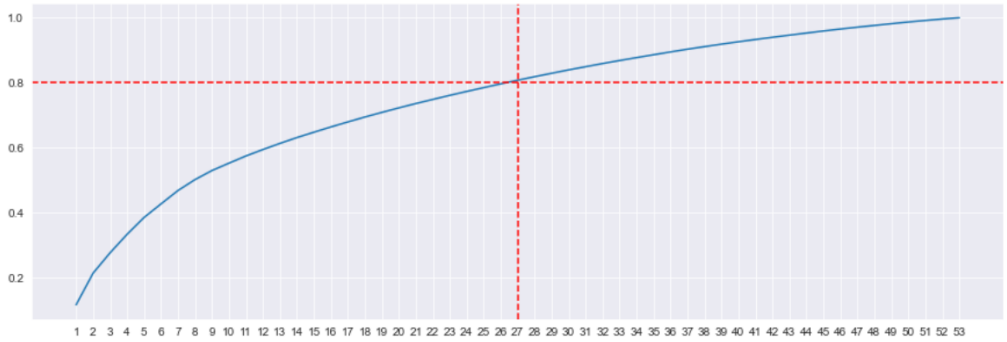
Data analysis Project 3

Apply dimension reduction methods – specifically a PCA – to the data in columns 421-474.

- a) Determine the number of factors (principal components) that you will interpret meaningfully (by a criterion of your choice – but make sure to name that criterion). Include a Scree plot in your answer.

Before performing PCA on the dataset, I filled in the missing values using the mean value per column and I mean centered in order to have a covariance matrix in this form A^TA. I used the pca function in sklearn, which has a attribute called acc_variance (which is basically the eigenvalues themselves) that show the variance explained by each eigenvector.





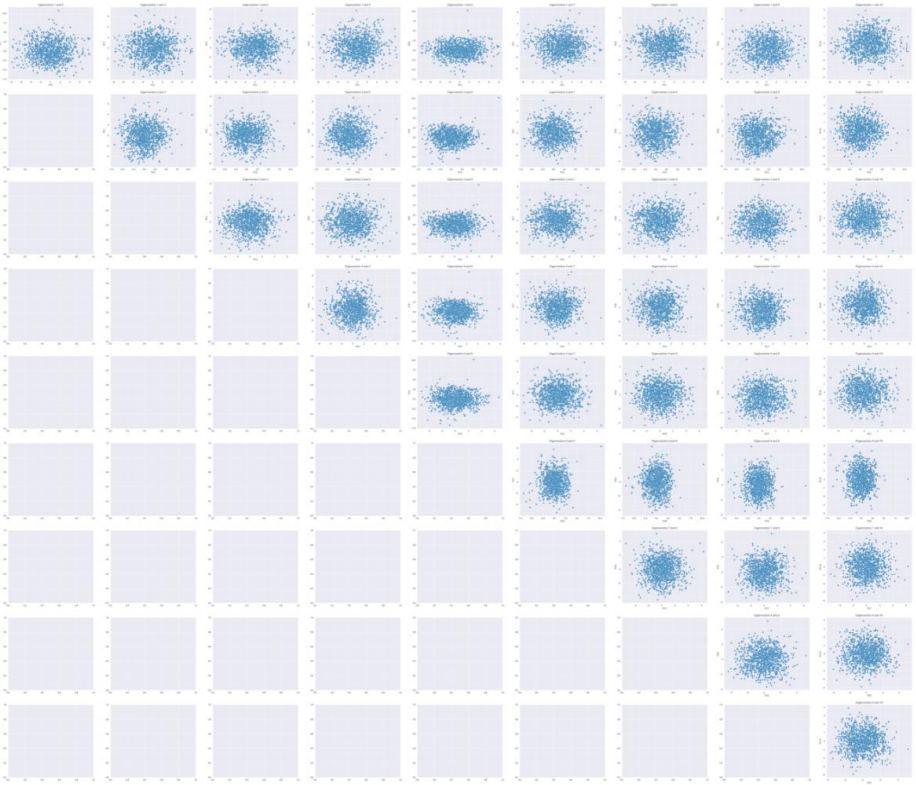
I checked how many eigenvectors accounted for 80% of the data and I obtained 27. As I had to plot them on the next question, I decided to choose the result from the Scree plot, which was 10 eigenvectors.

- b) Semantically interpret what those factors represent (hint: Inspect the loadings matrix). Explicitly name the factors you found and decided to interpret meaningfully in 1a). Be creative.

After analyzing the coefficients per column of every major eigenvector (by major I mean the top 10 eigenvectors that account for the most variance), I reached this semantic interpretation:

PC1: Positive/Negative mentality
PC2: Emotionally fragile/emotionally stable
PC3: Assertive/Meekness
PC4: Forgiving/ruthless
PC5: Structured/Unstructured
PC6: Emotional/Cold
PC7: Insecure/Empathetic
PC8: Artsy/Non-artsy
PC9: Social-watcher/Unsocial-watcher
PC10: Passionate/Dispassionate

2) Plot the data from columns 421-474 in the new coordinate system, where each dot represents a person, and the axes represent the factors you found in 1).



This is a plot of all the principal components against each other. I expected some clusters in this graphs but none appeared.

3) Identify clusters in this new space. Use a method of your choice (e.g. kMeans, DBScan, hierarchical clustering) to do so. Determine the optimal number of clusters and identify which cluster a given user is part of.

To determine the number of clusters I used the silhouette method. Some clusters were found even though the silhouette score was not high enough to be confident. Regardless, and for the sake of he exercise, I followed the scores provided by the silhouette method.

4) Use these principal components and/or clusters you identified to build a classification model of your choice (e.g. logistic regression, kNN, SVM, random forest), where you predict the movie ratings of all movies from the personality factors identified before. Make sure to use cross-validation methods to avoid overfitting and assess the accuracy of your model by stating its AUC.

I put together the training and testing sets with the features and targets required but I was not able to generate a AUC score (I believe it is due to the format of predicting a whole matrix and I would have had an AUC per predicted column, which I don't think is what is being asked in this question).