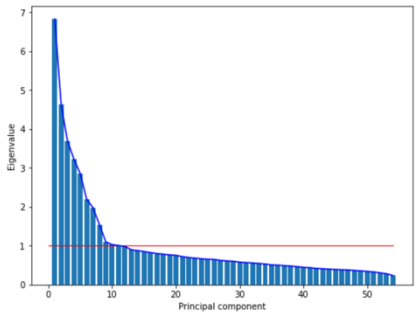
# Data analysis Project 3

Yoon Tae Park(yp2201@nyu.edu)

**1) 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.**
**Answer)**

- Meaningful number of factors: 8
- The elbow criterion: Pick only factors left of the biggest/sharpest drop(Scree plot included below)
- From the second biggest drop point, I picked left of the factors, which were 8 in total




**Explanation)**

- Firstly, I've normalized the dataset and conducted PCA. I found eigenvalues in decreasing order of magnitude. While plotting eigenvalues, I tried Kaiser criterion, but there were several pca features of having eigenvalues just above 1. I needed to optimize more, so used elbow criterion.
- By checking the biggest drop point, pca feature 1 to feature 2 had the biggest drop point. However, only using pca feature 1 was not enough, since it cannot explain much (only 12.6% of covariance explained). So I chose the second biggest drop point, and picked 8 features which were left of the drop point. 8 features have about 50% of covariance explained.

**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**
**Answer)**

By inspecting the eigenvector matrix, I found a relationship among questions and names the factors as below

- Feature 0: Energetic
- Feature 1: Emotional
- Feature 2: Quiet
- Feature 3: Persevere
- Feature 4: Uninventive
- Feature 5: Unartistic
- Feature 6: Forgets easily
- Feature 7: Empathy

**Explanation)**

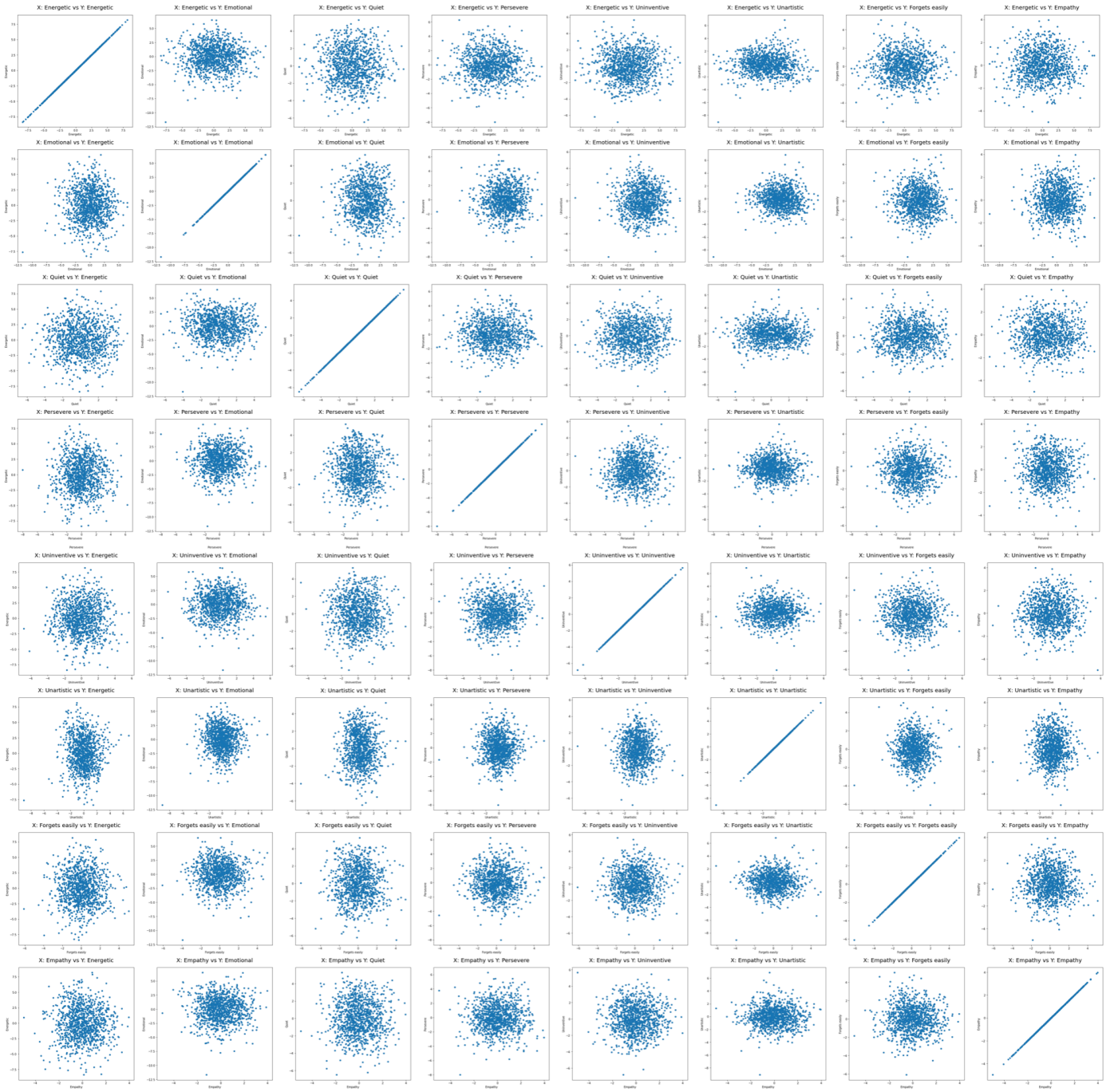
- By iterating group of eigenvectors, I checked top 3 and bottom 3 questions that have correlation with given pca components. By grouping and characterizing given questions, I was able to name 8 features
- For example, feature 0 had top 3 question of ['Is full of energy', 'Generates a lot of Enthusiasm', 'is outgoing/sociable'], and bottom 3 question of ['Is depressed/Blue', 'Tends to be quiet', 'Can be cold and aloof']. Feature 0 can be named as 'Energetic' since top 3 questions are about energetic and bottom 3 questions are about 'Lethargic'.

# Data analysis Project 3

Yoon Tae Park(yp2201@nyu.edu)

**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)**

- Plotting the data as multiple subplots of 8 features. I will use a new space of x-axis: Feature 0 vs y-axis: feature 1
- The new space will be relationship between ‘Energetic’ and ‘Emotional’

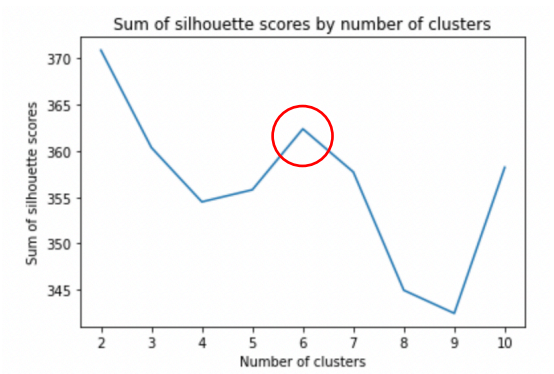


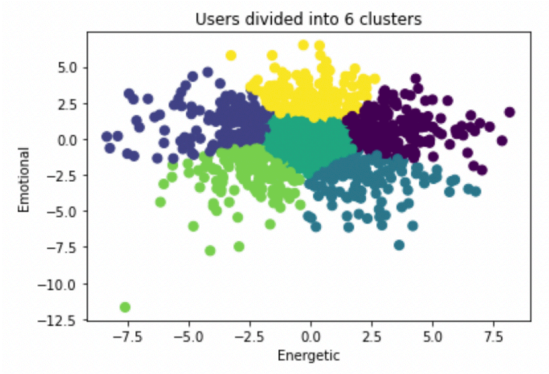
# Data analysis Project 3

Yoon Tae Park(yp2201@nyu.edu)

**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**

- By using KMeans, and checking the peak silhouette scores, I chose optimal clusters as 6




- Users are divided into 6 clusters: Cluster 0: Energetic, Cluster 1: Lethargic + Emotional, Cluster 2: Energetic + Emotionless, Cluster 3: Neutral, Cluster 4: Lethargic + Emotionless, Cluster 5: Emotional


**4) Use these principal components and/or clusters you identified to build a classification model of your choice, 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.**

- By using Logistic regression model, I got average auc score: 0.5936.
- I've chose logistic regression, since it can be easily interpreted. I've used 8 pca components as features, and 400 movie ratings as target values. Since movie ratings are not binomial, I've transferred movie ratings into 0 or 1.
- Assumption: If movie ratings are more or equal to 2.0, then it is a good rated movie(=1), else it is a bad rated movie(=0).
- Firstly, I've created logistic regression model without any cross-validation or hyper parameter tuning. By iterating over 400 movies, I got average auc score: 0.6022, Average accuracy score: 0.9520
- Then, I used cross-validation method by using grid search cv, with having 5 folds, and optimizing its hyper parameter. By iterating over 400 movies, I got average auc score: 0.5936, Average accuracy score: 0.9520

**5) Create a neural network model of your choice to predict movie ratings, using information from all 477 columns. Make sure to comment on the accuracy of this model.**

- By using neural network model, I got average accuracy of this model of 0.9521
- I've used data in columns 421-474 as features, and 400 movie ratings as target values. Since movie ratings are not binomial, I've transferred movie ratings into 0 or 1.
- Assumption: If movie ratings are more or equal to 2.0, then it is a good rated movie(=1), else it is a bad rated movie(=0)
- Firstly, I've created CNN class: (x * 128) * (128 * x). Then I've initializing CNN model: (x * 128) * (128 * 16) and Train by 16 indexes: (16 * 128) * (128 * 16). Finally, I predict on test data and calculate accuracy of this model. By iterating over 400 movies, I got average accuracy score: 0.9521

**Extra Credit: Use machine learning methods to tell us something interesting and true about the movies in this dataset that is not already covered by the questions above [for an additional 10% of the grade score].**

- In terms of accuracy score, my logistic regression model to predict movie ratings was 0.9520. Also, the neural network model had accuracy score of 0.9521. It is quite interesting to see that traditional machine learning model(logistic regression) and cutting-edge model(neural network model) doesn't show a lot of difference in this case.
- This implies that traditional classification method can be also used to predict the target value even in this era of 'deep learning'. Sometimes, traditional method can be more powerful in terms of interpretation.