NAME : KEWAL JANI

NET-ID : KJ2062

Email : kj2062@nyu.edu

Project 3

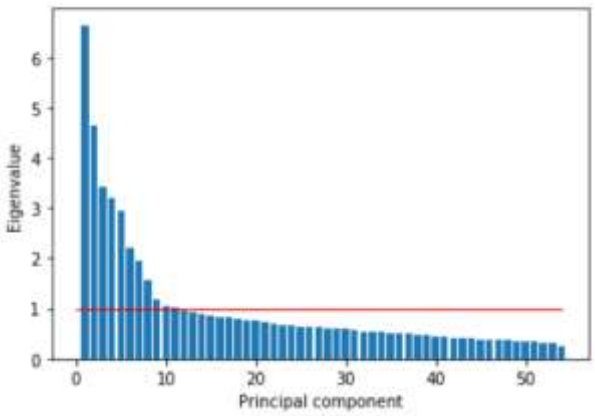
Q1 Apply dimension reduction methods – specifically a PCA – to the data in columns 421-474. As laid out above, these columns contain self-report answers to personality and how these individuals experience movies, respectively. It is us to you whether you do one PCA each for personality and movie experience, or one overall, but regardless of that, we would like you to:

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.**
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.**

ANS :

First I took the data and filled the NAN values with the Median values for the rating part. As mean would give float values which is not possible for a movie rating. After getting all the values I took the values from 421-474 and applied PCA on the column to check which columns are useful and contribute to the dataset. I Applied the dimension reduction with the help of PCA. To check which factors are useful. I used Kiser criteria that is the eigen value which is more than 1 is selected and other are neglected. The plotting is as shown below.

```
[6.64216455 4.64944788 3.41647703 3.19985232 2.94917137 2.20581848
 1.93939567 1.57019085 1.16565993 1.05101263 1.01787953 0.98823149
 0.91637216 0.88642271 0.84917938 0.84097796 0.81714563 0.78655086
 0.76067674 0.75196921 0.72304772 0.69467379 0.68004971 0.66461518
 0.65004819 0.64306721 0.62133318 0.61083002 0.60552362 0.59789216
 0.5643873  0.55297829 0.54656506 0.5287358  0.51332229 0.49904523
 0.49281453 0.47919076 0.45974593 0.45642376 0.42727389 0.42133516
 0.40724271 0.40042709 0.390619   0.38361227 0.37281803 0.36698192
 0.36026784 0.33683959 0.33131933 0.32265874 0.29998415 0.23900427]
```



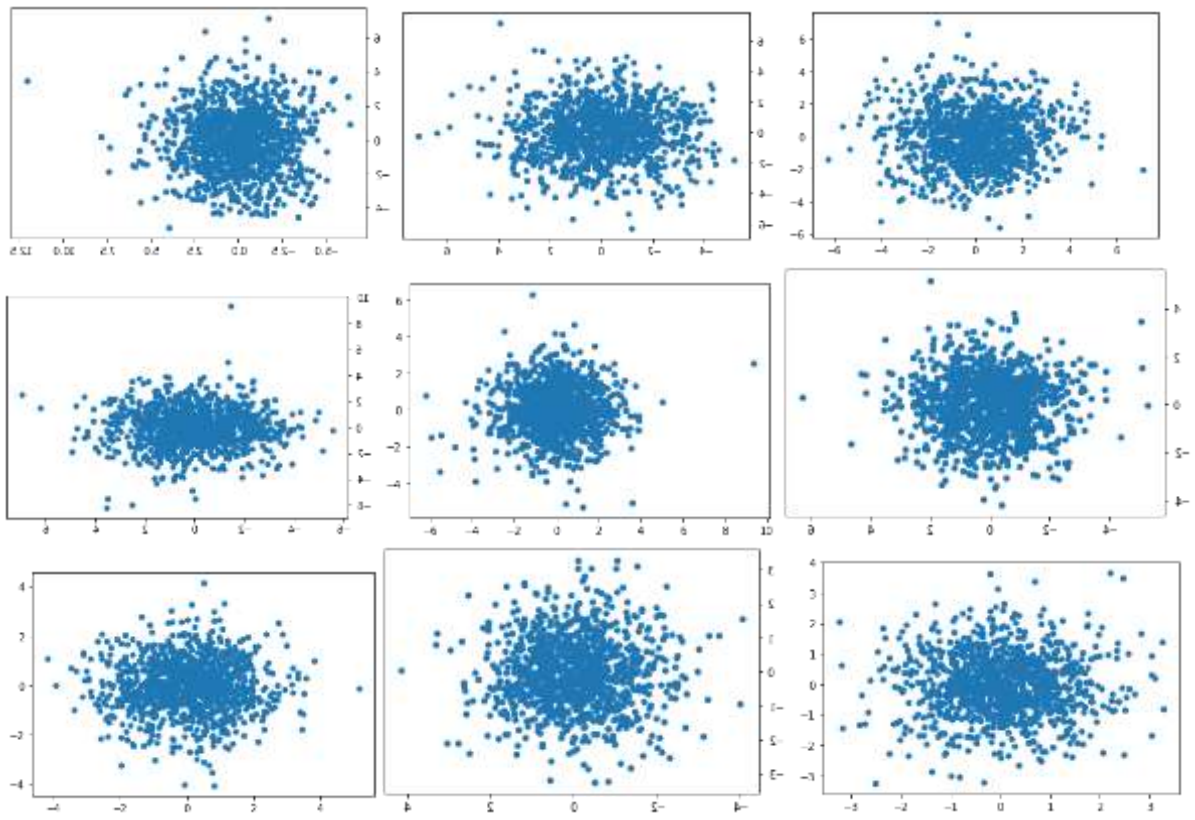
I found 11 eigen values to be more than 1 so for each eigen value I calculated which Question contribute the most. By plotting the contribution from loading matrix. I have included the plots in the Code which I have attached with this file.

So, these were the Questions that contributed the most

Is full of energy
The emotions on the screen "rub off" on me - for instance if something sad is happening I get sad or if something frightening is happening I get scared
Is reserved
Has an assertive personality
I have trouble following the story of a movie
Has few artistic interests
As a movie unfolds I start to have problems keeping track of events that happened earlier
When watching a movie I get completely immersed in the alternative reality of the film
Gender identity (1 = female; 2 = male; 3 = self-described)
Values artistic/aesthetic experiences
Is ingenious/a deep thinker

Q2 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). Hint: If you identified more than 2 meaningful factors, it is a good idea to create several 2D (X vs. Y) subplots for better interpretability.

Ans : As I identified 11 factors the total plots I will get for each question will be 11*11 that is 121 but I am showing random 9.

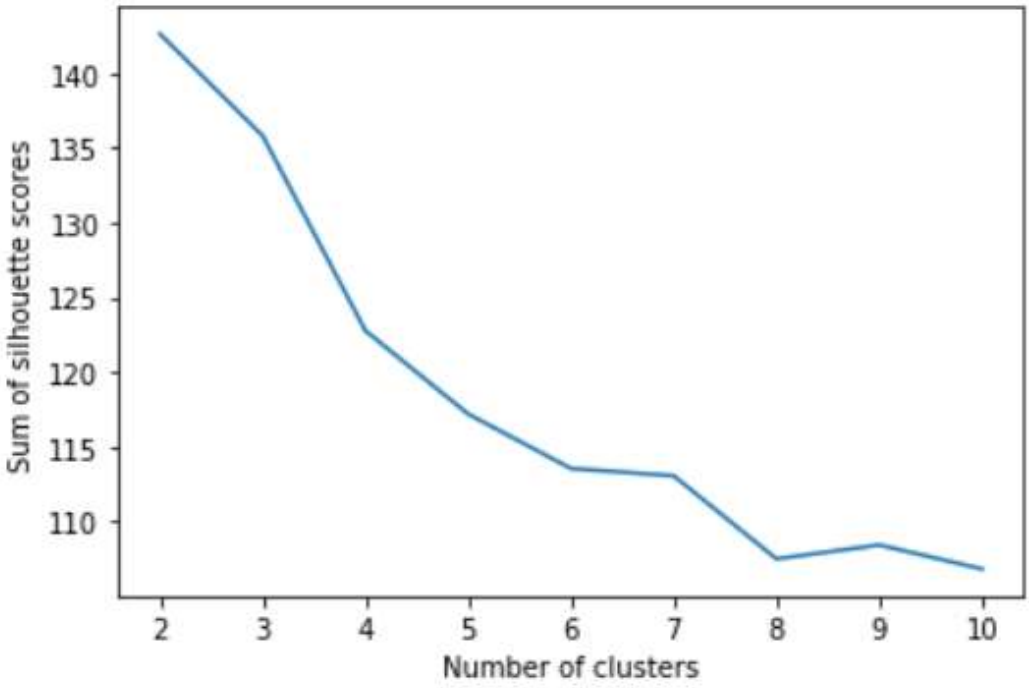


3**) Identify clusters in this new space. Use a method of your choice (e.g. kMeans, DBScan, hierarchical clustering) to do so. Determine the optimal number of clusters and identify which cluster a given user is part of.**

ANS :

Now after getting all the relevant Questions, I Collected them in on array and used Kmeans library to plot the graph but before that I have to select the optimum number of clusters needed for doing so

For that I ran the silhoute scores for cluster from 2-10 and got graph as shown below



This sates that the optimum number of cluster which I can select is 2 now I applied the cluster ing algorithm and got 2 clusters of the data divided in {0: 639, 1: 456}. We can give value of user to the Fitted model and will get output of the cluster the user is part of.

Q4 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.

ANS

For Question 4 I am using the principal components that I obtained from the PCA and trained the independent variables. It was hard to get multilabel classification from the data as there are fewer scikit libraries but I found random forest that predict the output as multiclass.

I divided the data in 80:20 ratio and then trained the 80% data. I tested the remaining testing data and got the accuracy of 83%.

Q5 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.

ANS

The input to this neural network is vector of length 77. It produces the output vector of length 400. Each element of a vector represents the movie rating prediction. I have used relu as the non-linearity to delete all the negative values. As movie prediction ratings cannot be negative. I have introduced dropout of probability 0.3 to zero 30 percent neurons. Due to this other neighboring neuron won't just copy their neighbours. They will learn by themselves instead of just copying. I have used MSE loss in order to find how much far away my prediction is from the correct output. By back propagation and updating weights I am reducing the loss and making my prediction come closer to the correct output. I got an average accuracy of 64%.

Q6 EXTRA QUESTION which movies contribute the most to the dataset?

ANS: I applied the PCA to the data set on 1-400 columns and repeated the above steps which I performed in PCA and found the most correlated movie rating. So instead of asking the user dataset of all the movies we can ask user these 90 movies to obtain the relevant data or replicate the dataset

Escape from LA (1996)
Wing Commander (1999)
Snatch (2000)
Planet of the Apes (2001)
Uptown Girls (2003)
Shutter Island (2010)
Poltergeist (1982)
The Game (1997)
Harry Potter and the Goblet of Fire (2005)
Harry Potter and the Sorcerer's Stone (2001)
Cocktail (1988)
Star Wars: Episode IV - A New Hope (1977)
Pearl Harbor (2001)
Just Married (2003)
21 Grams (2003)
Big Fish (2003)
The Game (1997)
On Golden Pond (1981)
Friday the 13th Part III (1982)
Shutter Island (2010)
House of Sand and Fog (2003)
Can't Hardly Wait (1998)
Flowers in the Attic (1987)
L.A. Confidential (1997)
Rambo: First Blood Part II
Saving Private Ryan (1998)
The Lord of the Rings: The Fellowship of the Ring (2001)
Billy Madison (1995)
Beetle Juice (1988)
Independence Day (1996)
28 Days Later (2002)
The Lord of the Rings: The Two Towers (2002)
Rambo: First Blood Part II
Sorority Boys (2002)
The Mist (2007)
Ed Wood (1994)
Interstellar (2014)
Anger Management (2002)

The Girl Next Door (2004)
13 Going on 30 (2004)
Baby Geniuses (1999)
10 Things I Hate About You (1999)
Batman & Robin (1997)
The Holiday (2006)
Armageddon (1998)
The Jungle Book (1967)
Schindler's List (1993)
Gladiator (2000)
Love Story (1970)
Bend it Like Beckham (2002)
La La Land (2016)
Grease (1978)
Star Wars: Episode II - Attack of
Die Another Day (2002)
The Iron Giant (1999)
The Mummy (1999)
Avatar (2009)
Ghost (1990)
Ed Wood (1994)
Wild Wild West (1999)
Saving Private Ryan (1998)
The Last Samurai (2003)
The Matrix Revolutions (2003)
The Mask (1994)
Bend it Like Beckham (2002)
Billy Madison (1995)
The Karate Kid Part II (1986)
Bend it Like Beckham (2002)
Mission: Impossible II (2000)
The Rock (1996)
The Blair Witch Project (1999)
The Last Samurai (2003)
Baby Geniuses (1999)
The Blair Witch Project (1999)
A Beautiful Mind (2001)
Spirited Away (2001)
Cast Away (2000)
The Sixth Sense (1999)
Top Gun (1986)
Austin Powers: The Spy Who Shagge
Ghostbusters (2016)
On Golden Pond (1981)
Let the Right One In (2008)
The Fast and the Furious (2001)
The Blue Lagoon (1980)
Zoolander (2001)
One Flew Over the Cuckoo's Nest (
Eternal Sunshine of the Spotless
The Cabin in the Woods (2012)
Scream (1996)