Sarvesh Patki

# Movie Rating Analysis

# Part 1

*Hypothesis Testing of movie rating data*

### Dataset Description and Preprocessing -

We had the ratings data of 400 movies from 1097 participants. It is obvious that not everyone has watched every movie. Hence, there were a lot of missing values(nans) in the data which needed to be taken care of. The problem with row-wise data removal over the entire dataset is that we lose almost the entire data and are left with a handful of entries. Thus, element-wise removal was used wherever necessary and appropriate tests were performed. This gives us samples/groups of different sizes but it won't be an issue because all the tests that we have used can handle groups of different sizes. Since we are mostly dealing with ratings data, reducing the data to medians is more appropriate than reducing the data to means and we choose the tests accordingly.

*For all the tests conducted, level of significance ($\alpha$) is set to **0.005**

### 1) Are movies that are more popular (operationalized as having more ratings) rated higher than movies that are less popular?

First of all we count the number of ratings received for each movie and do a median-split to determine high vs low popularity movies. We define our **null hypothesis** that **movies that are more popular are not rated higher than movies that are less popular**. Our alternative hypothesis is that movies that are popular are rated higher than movies that are less popular. We then run the **one-tailed Mann-Whitney U test** because here we are comparing ordinal data (movie ratings) from 2 groups. We get a p-value $< 0.005$. Therefore, we **reject the null hypothesis**. Thus, **movies that are popular are rated higher than movies that are less popular**.

### 2) Are movies that are newer rated differently than movies that are older?

From the movie title, we extract the year for each movie and do a median split to determine the older and newer movies. We define our **null hypothesis** that **movies that are newer are not rated differently to the movies that are older**. We run the **two-sided Mann-Whitney U test** again and get a p-value $> 0.005$. Hence, we **fail to reject the null hypothesis**. Therefore, **movies that are newer are not rated differently than movies that are older**.

### 3) Is enjoyment of 'Shrek (2001)' gendered, i.e. do male and female viewers rate it differently?

We define the **null hypothesis** that **male and female viewers do not rate 'Shrek' differently**. We then run the **two-sided Mann-Whitney U test** on the ratings given by male and female viewers. The p-value we get is equal to 0.0505 which is $> 0.005$ . Hence, **we fail to reject the null hypothesis**. Therefore, **male and female viewers do not rate 'Shrek' differently (enjoyment of 'Shrek' is not gendered).**

## 4) What proportion of movies are rated differently by male and female viewers?

We define the **null hypothesis** that **male and female viewers do not rate a movie differently**. We then run the **two-sided Mann-Whitney U test** on the ratings given by male and female viewers for all the movies in the dataset. If the p-value for a particular movie is < 0.005, then we reject the null hypothesis, otherwise, we fail to reject it. After running the test on all movies we see that for 50 movies, the p-value is < 0.005, thus we reject the null hypothesis only for 50 movies. Therefore **12.5% (50/400) movies are rated differently by male and female viewers**.

## 5) Do people who are only children enjoy 'The Lion King (1994)' more than people with siblings?

We define the **null hypothesis** that **only children do not enjoy 'The Lion King' more than people with siblings**. We then run the **one-tailed Mann-Whitney U test** on the ratings given by viewers who are only children and viewers who have siblings. The p-value we get is equal to 0.977, which is > 0.005. Hence, **we fail to reject the null hypothesis**. Therefore, **people who are only children do not enjoy 'The Lion King' more than people with siblings**.

## 6) What proportion of movies exhibit an "only child effect", i.e. are rated differently by viewers with siblings vs. those without?

We define the **null hypothesis** that **only children and people with siblings do not rate a particular movie differently**. We then run the **two-sided Mann-Whitney U test** on the ratings given by viewers who are only children and viewers who have siblings. If the p-value for a particular movie is < 0.005, then we reject the null hypothesis, otherwise, we fail to reject it. After running the test on all movies we see that for 7 movies, the p-value is < 0.005, thus we reject the null hypothesis only for 7 movies. Therefore **1.75% (7/400) movies have an "only child effect"**.

## 7) Do people who like to watch movies socially enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone?

We define the **null hypothesis** that **people who watch movies socially do not enjoy 'The Wolf of Wall Street' more than those who prefer to watch them alone**. We then run the **one-tailed Mann-Whitney U test** on the ratings for the movie 'The Wolf of Wall Street' given by viewers who watch movies alone and viewers who like to watch movies socially. We get a p-value of 0.9436 which is > 0.005. Hence, **we fail to reject the null hypothesis**. Therefore, **people who like to watch movies socially do not enjoy 'The Wolf of Wall Street' more than those who prefer to watch movies alone**.

## 8) What proportion of movies exhibits such a "social watching" effect?

We define the **null hypothesis** that **people who watch movies socially do not enjoy a particular movie more than those who prefer to watch them alone**. We then run the **one-tailed Mann-Whitney U test** on the ratings given by viewers who watch movies alone and viewers who like to watch movies socially for all movies in the dataset. If the p-value for a particular movie is < 0.005, then we reject the null hypothesis, otherwise, we fail to reject it. After running the test on all movies we see that for 6 movies, the p-value is < 0.005, thus we

**reject the null hypothesis** only for 6 movies. Therefore **2.5% (10/400) of movies exhibit a "social watching" effect**.

### 9) Is the ratings distribution of 'Home Alone (1990)' different from that of 'Finding Nemo (2003)'?

In this case we are dealing with only 2 movies instead of the entire dataset of 400 movies.

Although we might lose some additional data as compared to element-wise removal, for this problem, **row-wise removal** of nans makes more sense because we want to take into consideration the individuals who have watched both movies. We define the **null hypothesis** that the **rating distribution of 'Home Alone' is not different from that of 'Finding Nemo'**. Since we want to compare the underlying distributions of the ratings of the 2 movies, we run the **two-sample Kolmogorov-Smirnov (KS) test**, as it is fair to consider that the ratings of the above 2 movies by a participant will be independent of each other. The p-value obtained is $< 0.005$ hence we **reject the null hypothesis**. Therefore the **ratings of distribution of 'Home Alone' and 'Finding Nemo' are different**.

### 10) There are ratings on movies from several franchises (['Star Wars', 'Harry Potter', 'The Matrix', 'Indiana Jones', 'Jurassic Park', 'Pirates of the Caribbean', 'Toy Story', 'Batman']) in this dataset. How many of these are of inconsistent quality, as experienced by viewers?

We define the **null hypothesis** that **the movies in a particular franchise are not of inconsistent quality**. For this problem, we have more than 2 samples hence we run the **Kruskal Wallis test** (which is a non-parametric version of ANOVA) on all the movies in each of the franchises. If we get a p-value $< 0.005$, we reject the null hypothesis and say that the movies in the franchise are of inconsistent quality. Except for the franchise "Harry Potter" (p-value = 0.1179) and "Pirates of the Caribbean" (p-value = 0.0357), we get the p-value for all other franchises $< 0.005$. Hence, we reject the null hypothesis for all the other franchises whereas we fail to reject the null hypothesis for "Harry Potter" and "Pirates of the Caribbean". Therefore, **"Harry Potter" and "Pirates of the Caribbean" are of consistent quality whereas all others franchises are of inconsistent quality, as experienced by viewers.**

### 11) *(Extra-Credit) Is the rating distribution for the movie 'Inception (2010)' different for people who have trouble following the story of a movie and those who do not?

We define the **null hypothesis** that **the ratings distribution for the movie Inception (2010), is not different for the people who have trouble following the story of a movie and those who don't**.

We have split the participants into groups for this problem. We have explored 2 ways of doing it:

1. We segregate the participants into 2 groups -
    1.1. Participants who have trouble following the story of the movie - Include the participants who answered either 1, 2 or 3 for the question 'I have trouble following the story of a movie'.
    1.2. Participants who do not have trouble following the story of the movie - Include

the participants who answered either 4 or 5 for the question 'I have trouble following the story of a movie'.

We run the **two-sample Kolmogorov-Smirnov (KS) test** to compare the rating distribution by these 2 groups of participants. We get a p-value > 0.005, hence we **fail to reject the null hypothesis**. Thus, the **rating distribution for the movie 'Inception (2010) is not different for people who have trouble following the story of a movie and those who do not**.

2. We segregate the participants into 5 groups -
   - 2.1. Participants who marked 1 for the question 'I have trouble following the story of a movie'.
   - 2.2. Participants who marked 2 for the question 'I have trouble following the story of a movie'.
   - 2.3. Participants who marked 3 for the question 'I have trouble following the story of a movie'.
   - 2.4. Participants who marked 4 for the question 'I have trouble following the story of a movie'.
   - 2.5. Participants who marked 5 for the question 'I have trouble following the story of a movie'.

If a participant marks 5, he strongly agrees with the statement that he/she has trouble following the story of a movie.

We run the **Kruskal Wallis test** (which is a non-parametric version of ANOVA) on the above 5 samples and it is reasonable to assume that these samples are independent. We get a p-value > 0.005 and hence **we fail to reject the null hypothesis**. Thus, **rating distribution for the movie 'Inception (2010) is not different for people who have trouble following the story of a movie and those who do not**.

If we wanted to investigate if people who have trouble following the story of a movie and those who don't, rate the movie 'Inception (2010)' differently?, we could also run the **two-sided Mann Whitney U-test**. In this case, our null hypothesis is that **people who have trouble following the story of a movie and those who don't, do not rate the movie 'Inception (2010)' differently**. We get the p-value > 0.005, so **we fail to reject the null hypothesis**. Thus, **people who have trouble following the story of a movie and those who don't, do not rate the movie 'Inception (2010)' differently.**

Thus, we can see that even if we have used different tests and different methods of splitting the participants, we get the same results in each case. Christopher Nolan movies are usually considered to be intellectually stimulating, complex and often hard to keep up with. Even the people who don't have trouble following the story of a movie might get lost while watching Chris Nolan's movies. This might be the reason that there was no difference in ratings between participants who have trouble following the story of a movie and those who don't. So, it is reasonable to say that nobody understands Chris Nolan's movies, lol. Had it been a different movie that was somewhat easier to comprehend and keep up with (probably with fewer aspects of Physics), the rating distribution and ratings given by those groups would be different.

# Part 2

*Correlation and Regression of Movie Ratings Data*

### Dataset Description and Preprocessing -

There are a lot of missing values in the dataset. The missing values occur when a user has not rated a particular movie or has not answered a particular personality question. The missing values for the movie ratings are expected as it is not reasonable to assume that all users have watched each and every movie in the dataset. Also, a few participants have not answered some of the personality questions which further introduces some missing values. We handle all the missing values by imputing them with their means. We have used the sklearn 'SimpleImputer' class for this task. All the missing values (nans) in each column (movie rating or personality question) are replaced by the mean of the rest of the values in the column.

We divide our dataset into 2 parts - df_rate and df_pers:

Columns 1 to 400 contain the ratings of all users for the 400 movies. We call this df_rate. The dimension of df_rate is 1097 x 400.

Columns 401 to 474 contain responses to self-assessments on sensation-seeking behaviors, personality questions, and self-reported movie experience ratings. We call this df_pers and its dimension is 1097 x 74.

### 1) For all missing values in the data, use the average of the corresponding column so to fill in the missing data. In this problem, under the most correlated, we consider the largest correlation in the absolute value.

#### 1. For every user in the given data, find its most correlated user.

To find the correlation between users, we will consider only the movie rating part of the data i.e df_rate. We will compute the correlation matrix using the .corr() function in pandas and consider the absolute values of correlation. The size of the correlation matrix is 1097x1097 because we have 1097 users. The elements along the diagonal of this matrix represent the values of self-correlation for the users. This matrix is symmetric in nature. The value of self-correlation for each user is equal to 1. This will be the highest value of correlation for each user. But since we are interested in finding the maximum correlation of each user with other users, let us set this value to -1 and then find the value of maximum correlation for each user using the max() function and store it in a variable max_correlation_each_user. This gives us the index of the maximum value, which in turn gives us the user who is most correlated. The variable **most_correlated_user_idx** stores the values of the most correlated users for each user.

#### 2. What is the pair of the most correlated users in the data?

The most correlated pair of users are users #**831** and #**896**. We use the max() function to find maximum values of correlation for each user and store it in a variable max_correlation_each_user. Now, we use the idxmax() function to get the index of the user

which has the highest value of correlation. Once we get this index, we get one of the users who is a part of the pair. To find the other corresponding user in the pair, we will find the index associated with this user using the value of the 'most_correlated_user_idx' variable at the index that we have calculated above (for one of the users in the pair that we are interested in). Effectively what we are doing is finding the pair of users that has the highest correlation from the correlation matrix.

### 3. What is the value of this highest correlation?

The highest value of correlation is **0.9987**.

The variable max_correlation_each_users contains the maximum values of correlation for each user. Finding the maximum value of this array will give us the maximum value of correlation that is present. Alternatively, we can also directly find the highest value of correlation directly from the correlation matrix.

### 4. For users 0, 1, 2, \dots, 9, print their most correlated users.

The variable most_correlated_user_idx which we calculated in 1.1 has the list of the most correlated users for each user. To find the most correlated users for users 0,1, …, 9 we just have to check the first 10 entries of this array. The results are as follows:

| User | Most Correlated User |
|---|---|
| 0 | 118 |
| 1 | 831 |
| 2 | 896 |
| 3 | 19 |
| 4 | 784 |
| 5 | 990 |
| 6 | 1071 |
| 7 | 1074 |
| 8 | 821 |
| 9 | 1004 |

2) We want to find a model between the ratings and the personal part of the data. Our main task is to model:

df_pers = function(df_rate)

First of all we split the original data into training and testing data in the ratio 0.8 : 0.2. For this, we use the test_train_split function.

1. Model df_pers = function(df_rate) by using linear regression. What are the errors on: (i) the training part; (ii) the testing part?

After splitting the data into training and testing sets we ran a linear regression model and we computed two types of errors:

| Training Error | Testing Error |
|---|---|
| 0.6127612374234509 | 3.2509647714850916 |

2. Model df_pers = function(df_rate) by using the ridge regression with hyperparameter values alpha from [0.0, 1e-8, 1e-5, 0.1, 1, 10]. For each of the previous values for alpha, what are the errors on: (i) the training part; (ii) the testing part? What is the best choice for alpha?

In this question, we will now run the ridge regression and input different values of the hyperparameter alpha. For each value of alpha, we have computed the training and testing errors.

| Alpha | Training Error | Testing Error |
|---|---|---|
| 0.0 | 0.612761237423450 | 3.250964771485118 |
| 1e-08 | 0.612761237423451 | 3.250964762320853 |
| 1e-05 | 0.612761237424664 | 3.250955607294804 |
| 0.1 | 0.612867547463894 | 3.166127659797673 |
| 1 | 0.617659618995400 | 2.718440736635615 |
| **10** | **0.668106371145381** | **1.893212275643253** |

The training error increases slightly as the model generalizes. The testing error decreases as we increase alpha. The best choice of alpha is the value for which the testing error is the lowest. Hence, in this case, the best choice of alpha is 10, as the testing error is the lowest for it.

3. Model df_pers = function(df_rate) by using the lasso regression with hyperparameter values alpha from [1e-3, 1e-2, 1e-1, 1]. For each of the previous values for alpha, what are the errors on: (i) the training part; (ii) the testing part? What is the best choice for alpha?

In this question, we will now run the lasso regression and input different values of the hyperparameter alpha. For each value of alpha, we have computed the training and testing errors.

| Alpha | Training Error | Testing Error |
|---|---|---|
| 0.001 | 0.635907800772781 | 2.2804960457572663 |
| 0.01 | 0.893235190614389 | 1.3742334226815167 |
| **0.1** | **1.210230986955849** | **1.2534246581541626** |
| 1 | 1.226153957114822 | 1.2646356285936624 |

As in the previous case for the ridge, we see that the training error increases as the model generalizes. The testing error decreases as we increase alpha. The best choice of alpha is the value for which the testing error is the lowest. Hence, in this case, the best choice of alpha is 0.1, as the testing error is the lowest for it.

# Data Analysis Project 3

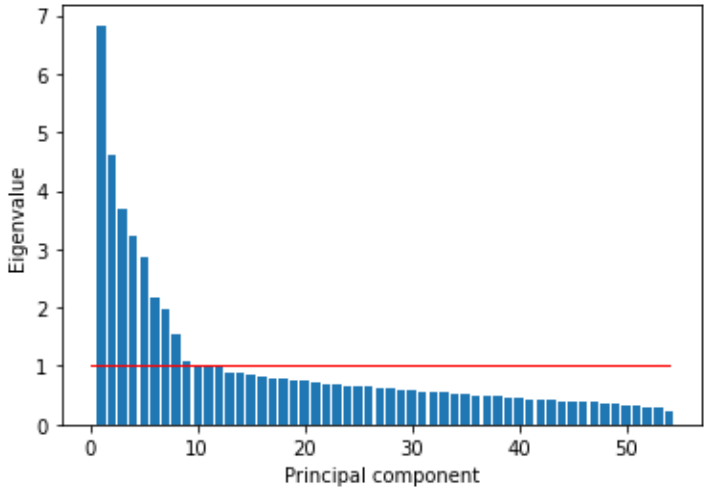
*Applying Machine Learning Methods to Movie Ratings Data*

1) Apply dimension reduction method i.e. 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.

We will apply PCA to the entire dataset. First of all, we need to normalize the data in before running PCA on it. We do that by finding the z-score of all the entries in the data and then run PCA over it. Next, we find all the eigenvalues and the loadings i.e the weights per factor in terms of the original data. After this, we compute the rotated/transformed data matrix and calculate the co-variance explained by each factor from the eigenvalues that we computed earlier.

a) To determine the number of factors (principal components) that can be interpreted meaningfully

From the eigenvalues that we computed, we can see that there are 11 eigenvalues that are greater than 1. Hence, by the **Kaiser criterion**, we will choose 11 principal components.




Scree Plot - with Kaiser Line in red

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.

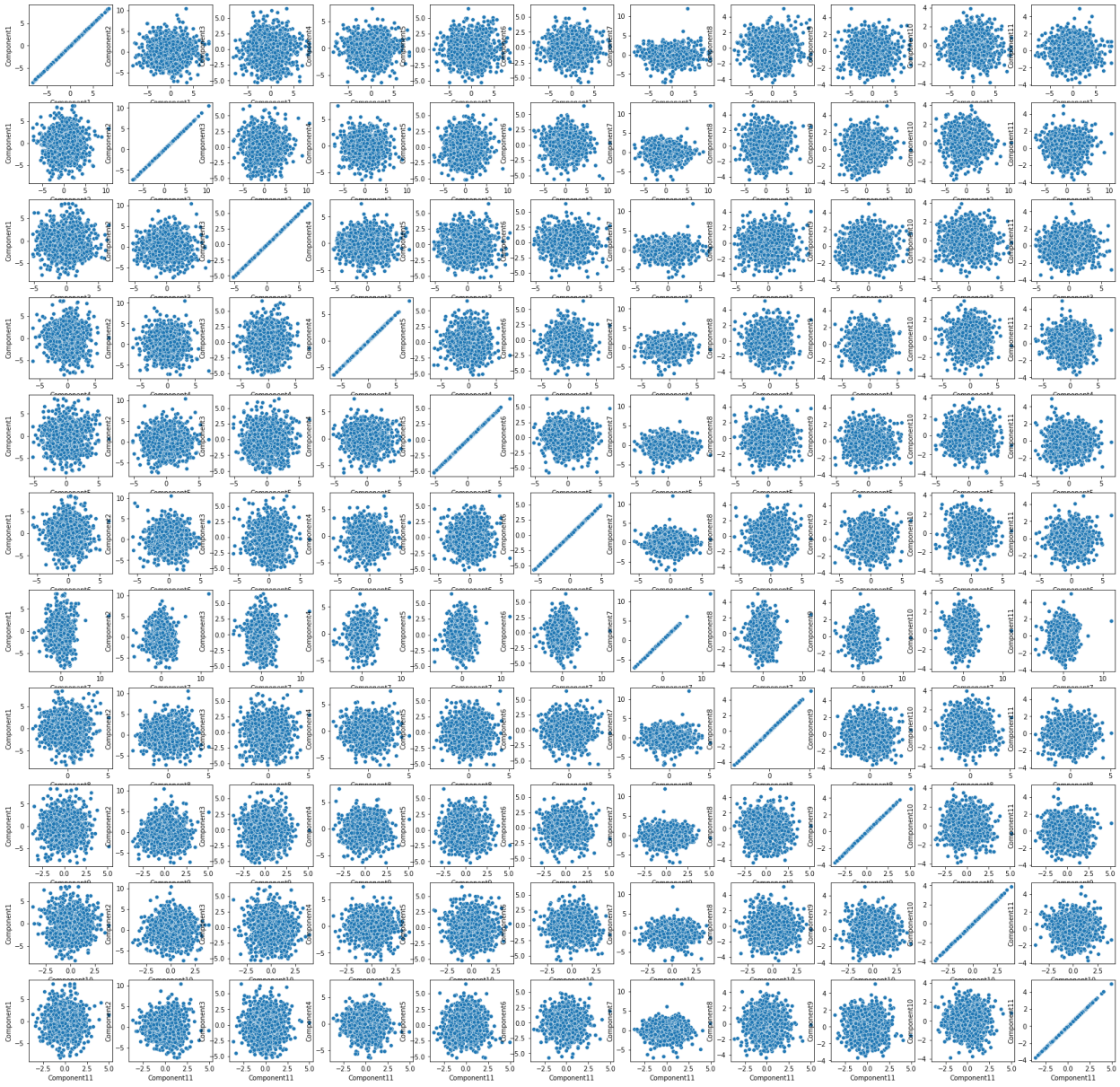
The loading matrix will help is to identify which principal component corresponds/points towards which of the 54 personality questions. We have selected the first 11 principal components according to the Kaiser criterion. The maximum value of the covariance for each factor in the loadings matrix corresponds to the direction in which the principal component points. Hence, for the first 11 principal components, we find this maximum value to get the associated factor in our original data.

The factors that we found can be seen in the table below:

| Factor |
|---|
| Movies change my position on social-economic or political issues |
| Has a forgiving nature |
| I have trouble remembering the story of a movie a couple of days after seeing it |
| Does things efficiently |
| As a movie unfolds I start to have problems keeping track of events that happened earlier |
| When watching movies things get so intense that I have to stop watching |
| Worries a lot |
| Gets nervous easily |
| When watching a movie I feel like the things on the screen are happening to me |
| Generates a lot of Enthusiasm |
| When watching a movie I get completely immersed in the alternative reality of the film |

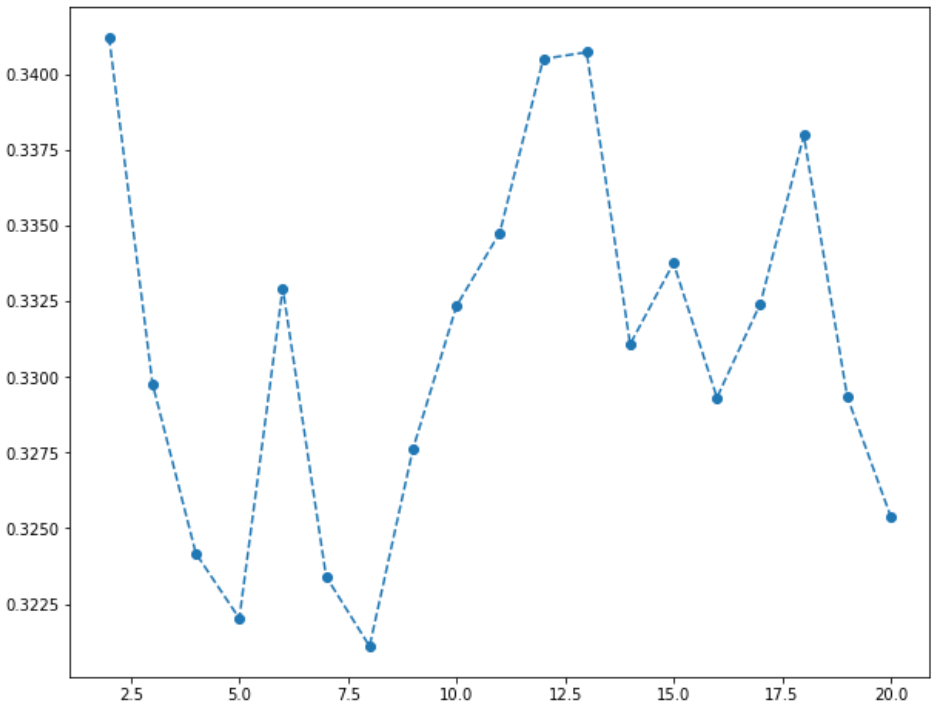
2) Now we plot the data from columns 421-474 in the new coordinate system, where each dot represents a person, and the axes represent the factors we found in 1).

We have plotted 2D plot for each pair of the 11 meaningful factors. Each factor is perfectly correlated to itself as can be seen from the plots along the diagonal.



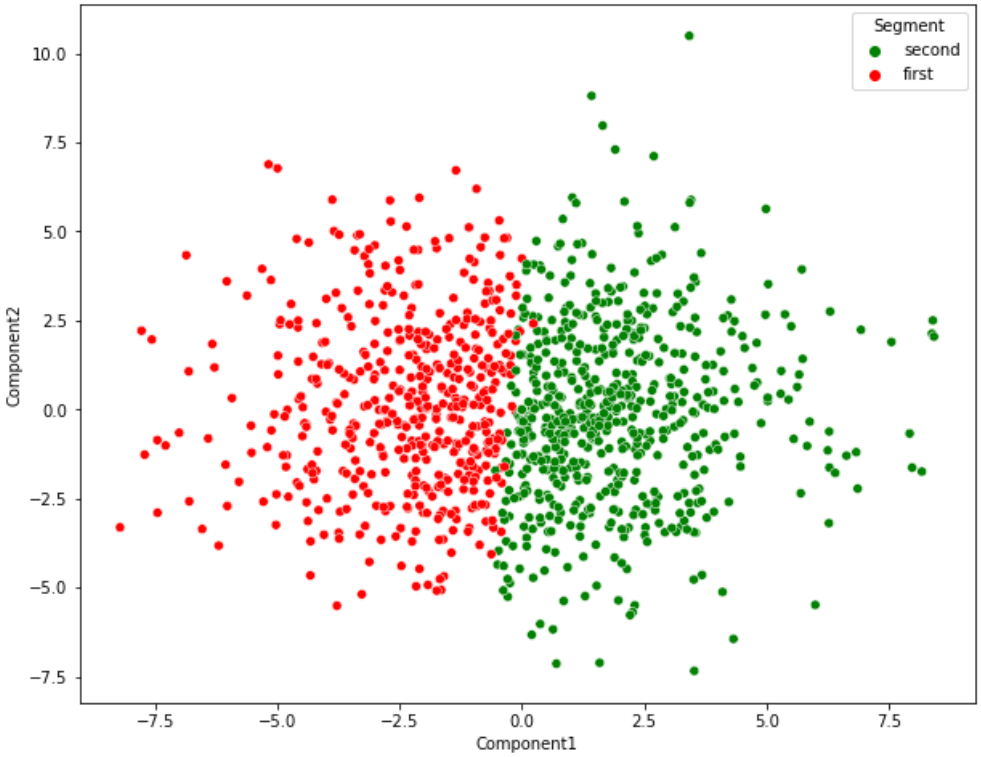
### 3) Identifying clusters in this new space and finding the optimal number of clusters.

We have used the **k-means** algorithm for clustering. We will use the **silhouette method** to determine the optimal number of clusters for our data. We have plotted the silhouette coefficients for different cluster sizes (2 to 20) and we can see that the plot peaks at k=2. Hence, the **optimal number of clusters is 2.**




Silhouette coefficients for the different number of clusters(2-20)

Now we run the k-means algorithm with k=2 and we get the clusters as follows.



### 4) Neural network model to predict movie ratings, using information from all 477 columns.

We have created a neural network model using the Keras framework. The input to our model is the answers to the personality questions from columns 421-474. We convert the ratings from continuous to categorical. We use the floor() function to round the ratings to the lower integer and finally all ratings belong to one of the following categories: 0,1,2,3,4. Now, we use a multi-class classification model. There are 54 personality questions hence input dimension is 54. The first hidden layer has 12 neurons, and the second hidden layer has 5 neurons. Both these layers have relu activation functions. And the final output layer has 5 neurons(one for each of the rating categories - 0,1,2,3,4) and we use a softmax activation for the output layer.

We use an adam optimizer for this model and a categorical_crossentropy and the metric used is accuracy for evaluating the model. We run the model for **10 epochs** for each of the 400 movies. The mean accuracy for all the 400 movies is 78.21%.