YEN LE IDS FALL 2021

Capstone Project:

Preface:

The dataset provided for contains missing data, and as such, requires cleaning. Throughout my analysis, I handled the missing data by finding the null values in the data needed for computation and deleting the corresponding rows containing those null values. Additionally, for analysis that involves multiple variables that capture the same information, I applied a dimension reduction by performing a Principle Computation Analysis. I also choose the cut off alpha level of 0.05 throughout this analysis.

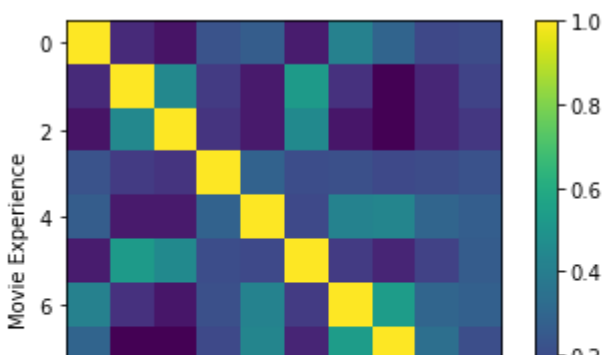
Content:

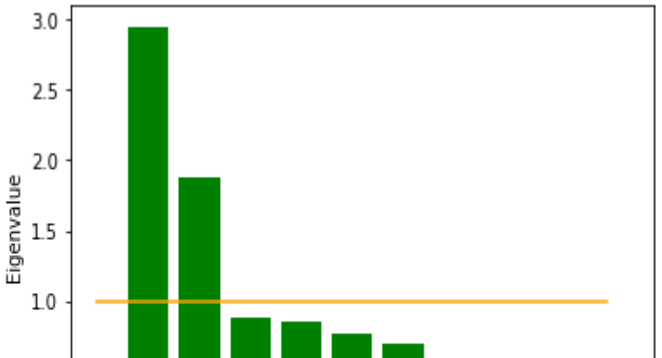
1) What is the relationship between sensation seeking and movie experience?

- The two features sensation seeking and movie experience have a correlation coefficient of r = 0.00125. The relationship between sensation seeking and movie experience is not that strong. So although there is a positive correlation between sensation seeking and movie experiences, I can say that there is no relationship between sensation seeking and movie experience.
- First, I will perform a PCA on the two features using a similar algorithm. I extract the data by indexing the columns, drop the NaN values and calculate a correlation matrix on the features, the movie experiences and the sensation seekings. (as shown in the graph below). In the process of doing a PCA, I normalize the data by transforming it into z-scored Data, initializing the PCA object and fitting it to my data. I also plot a Screeplot and use Kaiser criterion to evaluate the numbers of principal components that have the Eigenvalues of more than 1.
- In this example, we see that there are two main principle components, (graph in the graphs below), that can explain the Movie Experiences and there are 6 main principle components that can explain the Sensational Seeking feature.
- The r-value is calculated by using the built-in NumPy function np.corrcoef, calculating the correlation between two data frame representing the two features, giving us the answer of approximately 0.0125

For Movie Experiences:

- The correlation matrix shows variability so we can see there is some correlation between the questions. (Figure 1)





- Using Kaiser criterion, we derive at 2 main principle components:
- PC1: ‘How emotionally immersed you get when you watch movies’- mostly correlate to question 7,8
- PC2: ‘How likely you are to have troubles during watching movies’- mostly correlate to question 1,2

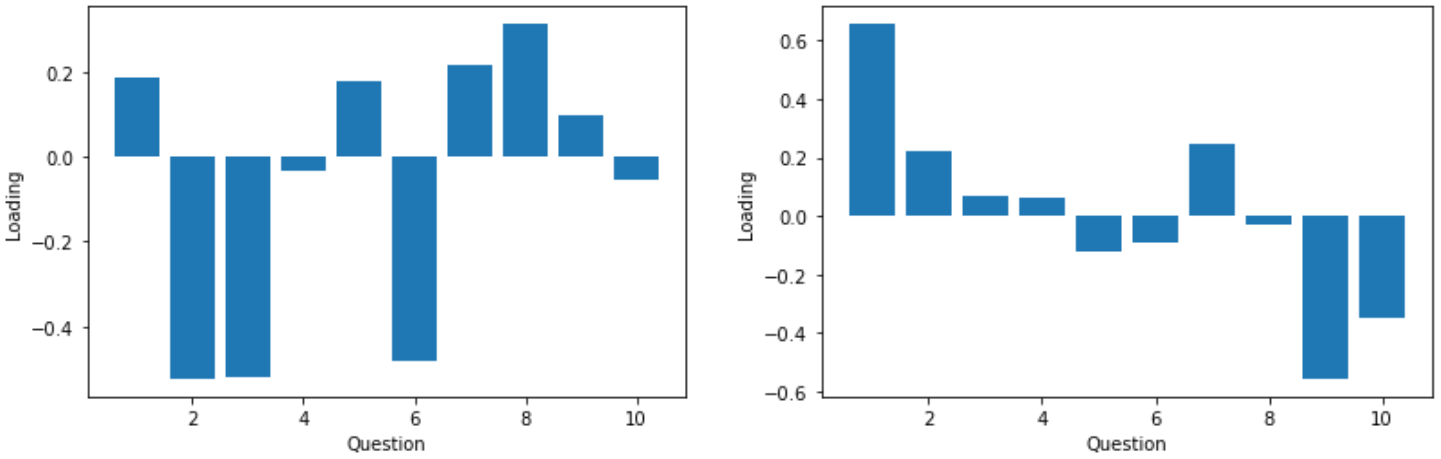


Figure 3: Principle Component (1st) and (2nd)

For Sensation Seeking:

- The correlation matrix shows variability so we can see there is some correlation between the questions. (Figure 2)

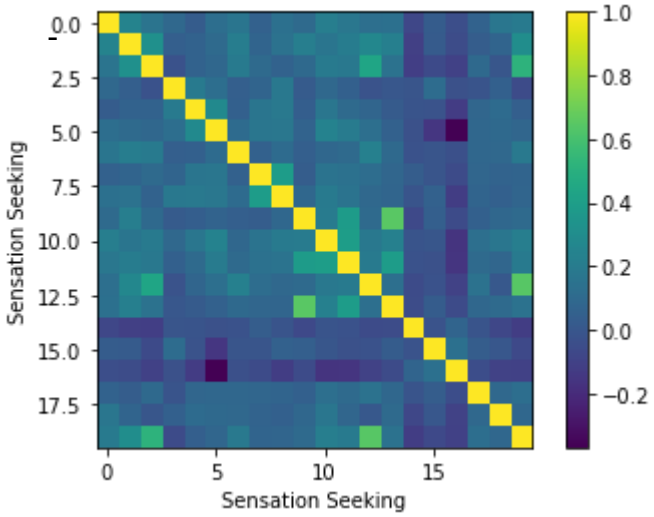


Figure 4: Sensation Seeking x Sensation Seeking correlation matrix

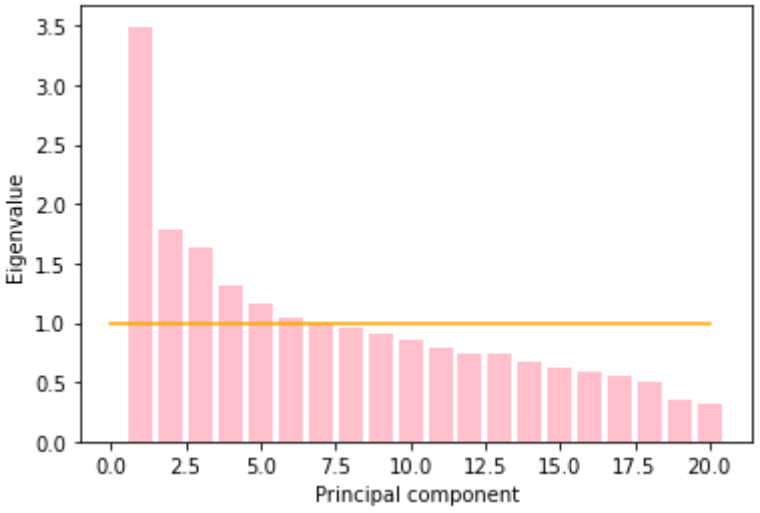


Figure 5: Screeplot for Sensation Seeking PCA

- Using Kaiser criterion, we derive at 6 main principle components:
- PC1: ‘Do you prefer crowded place’- mostly correlate to question 5,6,7
- PC2: ‘Do you enjoy extreme sports ’- mostly correlate to question 2,3,4
- PC3: ‘Do you get stress easily’- mostly correlate to question 1,15
- PC4: ‘Do you enjoy horror genre’- mostly correlate to question 10,14

  PC5: ‘How organized is your life’- mostly correlate to question 16,17

- PC6: ‘Do you often take reckless risk’- mostly correlate to question 1,19

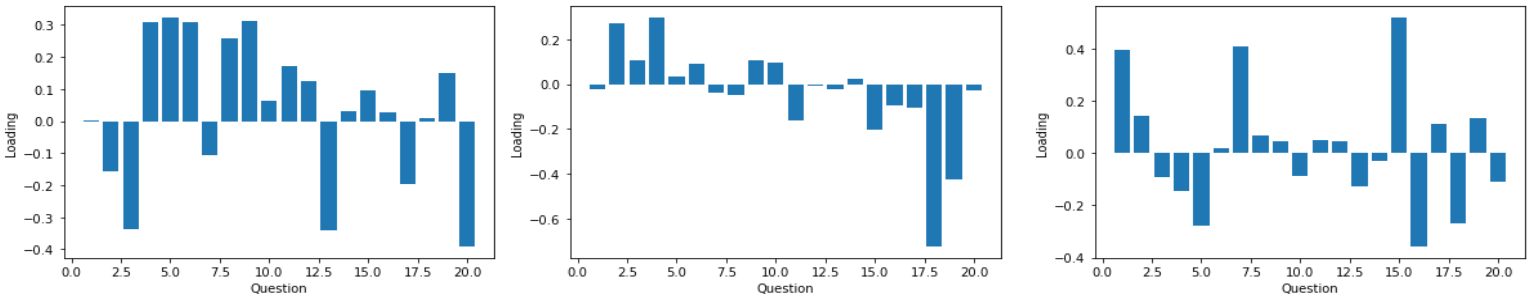


Figure 6: Principle Component (1st) and (2nd) and(3rd)

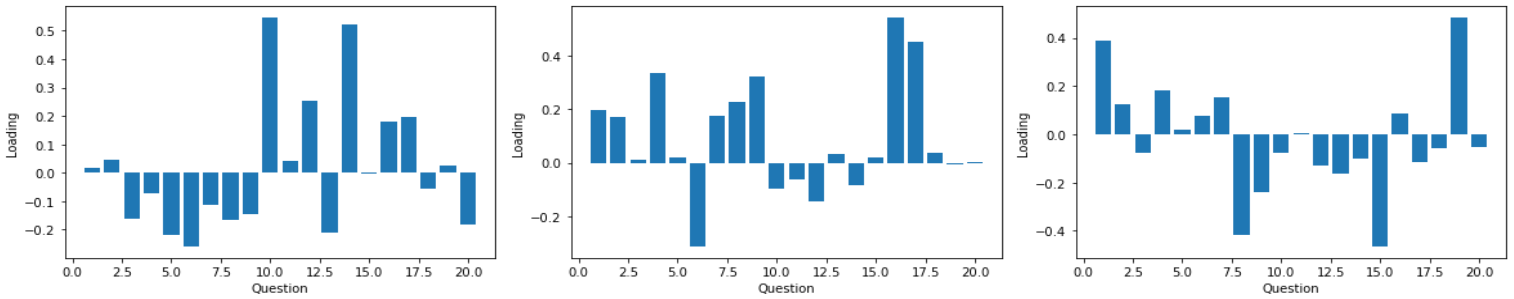


Figure 7: Principle Component (4rd) and (5rd) and(6rd)

<u>2) Is there evidence of personality types based on the data of these research participants? If so, characterize these types both quantitatively and narratively.</u>

Results:

- Using clustering and deriving at 2 clusters, there is evidence that there are two personality types. I used PCA on personality features and concluded that there are 8 main principle components that can explain personality features. Along with that, the optimal number of clusters are 2. Quantitatively, we can characterize this by splitting the PCs into 2 groups. Narratively, while it is hard to represent the figure in 2D, the two character types are characterized as social,active type and more reserved, organized type.

Process:

- I also did a PCA on the personality dataframe as the first question and derived at 8 main principle components, following the Kaiser criterion given by the graphs below. Next, I stack all the PCA-ed data together and perform clustering to find the classification of personality types. After using K-means and clustering, I found that 2 clusters are best for explaining the personality type. Since there are 8 principal components, I won’t be able to graph this in 2D. Therefore, in the graph below I will only use the first two principal components to visualize the clustering.

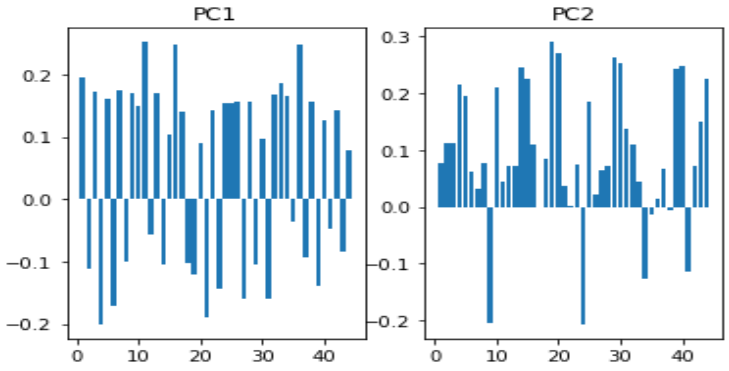


Figure 8: Principle Component (4rd) and (5rd) and(6rd)

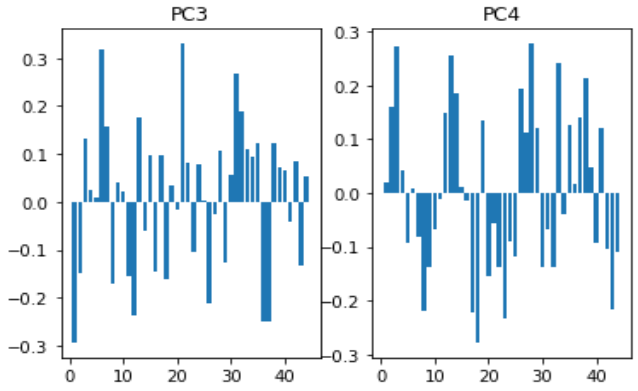


Figure 9: Principle Component (4rd) and (5rd) and(6rd)

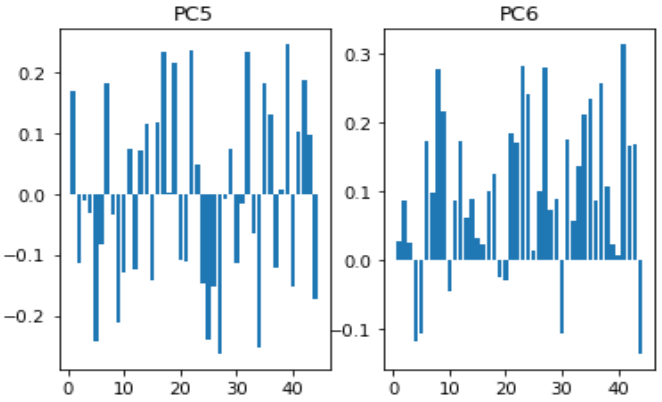


Figure 10: Principle Component (4rd) and (5rd) and(6rd)

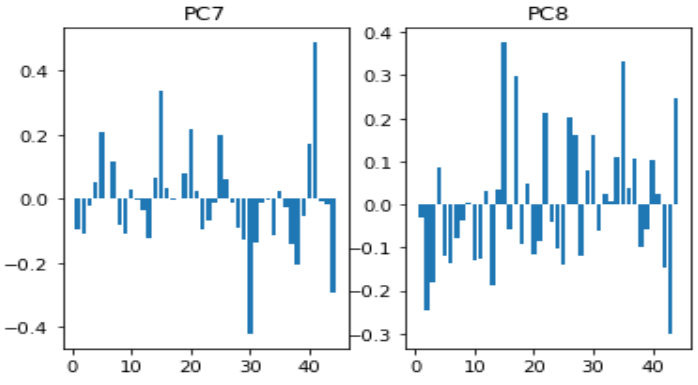


Figure 11: Principle Component (4rd) and (5rd) and(6rd)

- Using clustering and silhouette score, I find that the optimal number of clusters are 2, represented in the figure below. This is to show that there are 2 personality types.

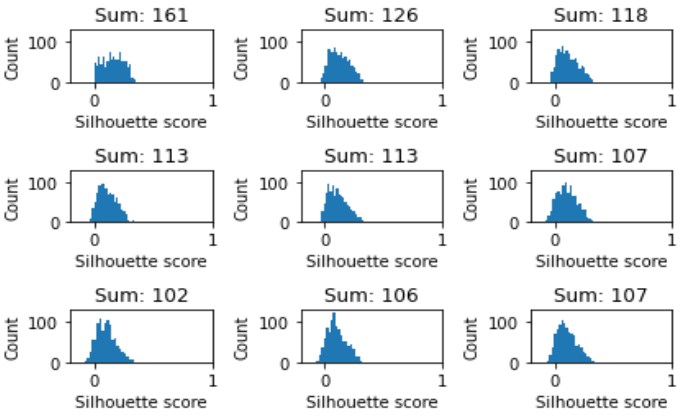


Figure 12: Sum of Silhouette Score

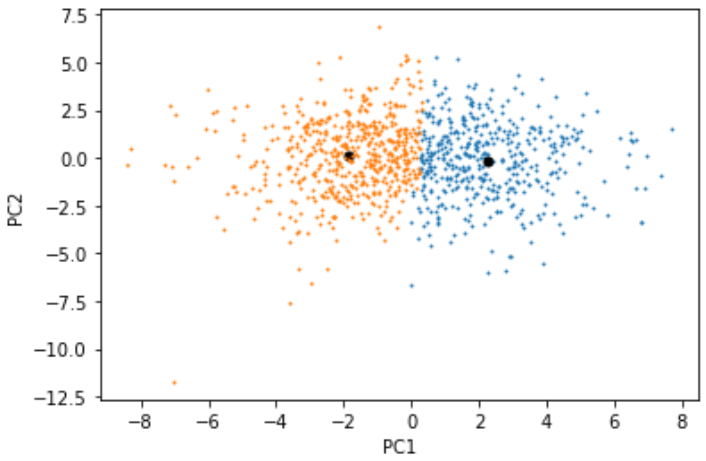


Figure 13: Clustering Plot in 2D

3) Are movies that are more popular rated higher than movies that are less popular?

- For this question, I extract the movie ratings and the gender columns. I first count the number of ratings in each column, calculating the median of ratings. Then specify popular movies as those who have more ratings than the grand median and unpopular movies as movies who have less ratings than the grand median. After storing them into two pieces of data, I calculate the median of median of the ratings values to see if the Popular Movies have higher ratings than Unpopular Movies. It shows that popular movies have a median ratings of 3.0, while unpopular movies have a median ratings of 2.5. This is to show that popular movies are rated higher than movies that are less popular.
- I think this is true because people tends to rate movies that they like. So popular movies, movies with more ratings, would usually receive high feedback. While in the case of unpopular movies, they received less ratings so each ratings accounts for a bigger proportion of variance in the median ratings. So if they have a low ratings, these low ratings will brought down the median by a lot, much more compared to popular movies (which have more ratings).

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

- In this question, I extract the data from the Shrek movies ratings and gender column and perform a Mann-Whitney test on the dataframe. This test gives us a U value of 96830.5 and p-value larger than 0.05, about 0.0503. Here, since the p-value is larger than the cut off level alpha of 0.05, we can conclude that we fail to reject the null hypothesis. This means that there is no difference between how males and females review the Shrek movies.

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

- In this question, I extract the data from The Lion King movies and the Only Child columns, performing the same Mann- Whitney test I did in question 3. This test gives me an U-values of 52929 and a p-values of 0.043. So here based on the p-values and the cut off level alpha of 0.05, I conclude that I reject the null hypothesis. So there is a difference in how only children and people with siblings rate the movie ' The Lion King'.
- Maybe people who have siblings can relate more to Simba, the main characters of the movie. First child tends to relate to Simba in the responsibilities he has to take and those who are second/third.. child understand Simba by looking up to their older siblings.

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

- In this question, I extract the data of movie ratings from the ‘ Wolf of Wall Street’ film and the Social Viewing Preference columns, perform a Mann-Whitney test on the dataframe and extract its U and p values. After performing the test, the U value is 56806 with the p-value of 0.113. Based on the cut off alpha level of 0.05, we fail to reject the null hypothesis . So we can conclude that there is no difference in the ratings of ‘The Wolf of Wall Street’ between the two groups of people: those who enjoy watching movies socially and those who enjoy watching movies alone.

7) 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?

- In this question, I performed a Kruskal Wallis D test between the movie ratings in each franchise to evaluate whether there are inconsistent quality between different parts of each franchise, as experienced by the viewers. I will record my results of D-values, p-values and Consistency in the table below :

Movies	D values	p values	Consistency
Star Wars	108,364	0.00…	Inconsistent
Harry Potter	4.354	0.113	Consistent
The Matrix	40.323	0.00..	Inconsistent
Indiana Jones	11.449	0.003	Inconsistent
Jurassic Park	49.427	0.00..	Inconsistent
Pirates of the Caribbean	6.66	0.036	Inconsistent
Toy Story	23.496	0.00..	Inconsistent
Batman	40.323	0.00	Inconsistent

- From this table, and based on the cut off alpha level of 0.05, we can see that there is quality consistency between these movies, except for Harry Potter. This inconsistency can be either explained by the budget of the movie, audience expectations or other factors. While Harry Potter does get affected by the same reason, one of the reasons why I think Harry Potter movies have a certain level of consistency in the movie quality is because this film is based on a best-selling novel. So, because people know what to

expect from the book and people love that book, we can easily say that people will enjoy the movie tool. Therefore there is consistency in quality in the ratings of Harry Potter.

8) Build a prediction model of your choice (regression or supervised learning) to predict movie ratings (for all 400 movies) from personality factors only. Make sure to use cross-validation methods to avoid overfitting and characterize the accuracy of your model.

- In this question, I will build a Random Forest with 100 trees to predict movie ratings (from all 400 movies) from personality factors only. Before training and testing the data, I did PCA on the personality factors, extracting 8 main principle components. Then use that data along with the movie ratings columns to build a random forest. Here, to cross-validate, I split my data into 20% test set and 80% train set. I also characterize the accuracy of my model by plotting the ROC curve.
- My model has a 52% accuracy. I also plot my ROC curve and my AUC value turns out to be 0.60.
- So my model predicts the outcome 52% of the time. This is not the highest percentage of accuracy I should get. In the future, I can try splitting the train set and test set differently or using different kinds of prediction model.
-

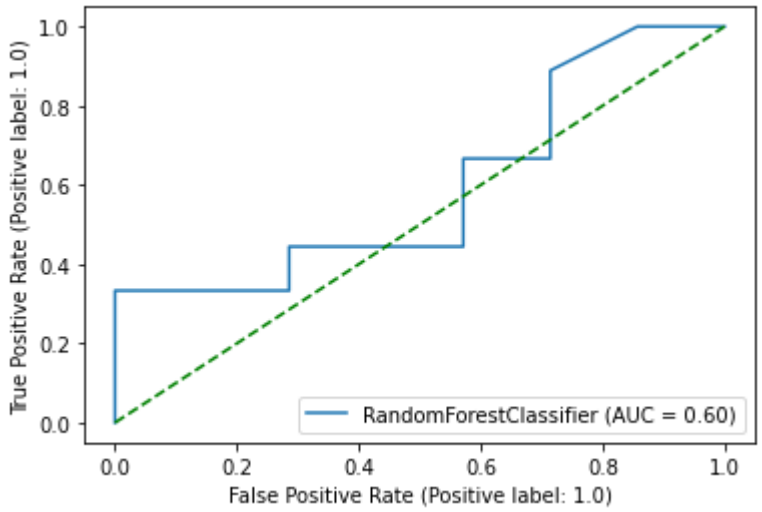


Figure 14: ROC curve and AUC Q8

9) Build a prediction model of your choice (regression or supervised learning) to predict movie ratings (for all 400 movies) from gender identity, sibship status and social viewing preferences (columns 475-477) only. Make sure to use cross-validation methods to avoid overfitting and characterize the accuracy of your model.

- This question builds on the algorithm of question 8. However, we are predicting movie ratings based on three factors, gender identity, sibling status and social viewing preferences. Therefore, I ran a PCA on the three features and extracted only 1 principal component using the Kaiser criterion. Then, I use that data to build a Random Forest, using 20% of the data as a test set and the other 80% is the train size. This yields a 52% accuracy with an AUC values of 0.28

- So my model predicts the outcome 52% of the time. This is not the highest percentage of accuracy I should get. In the future, I can try splitting the train set and test set differently or using different kinds of prediction models.

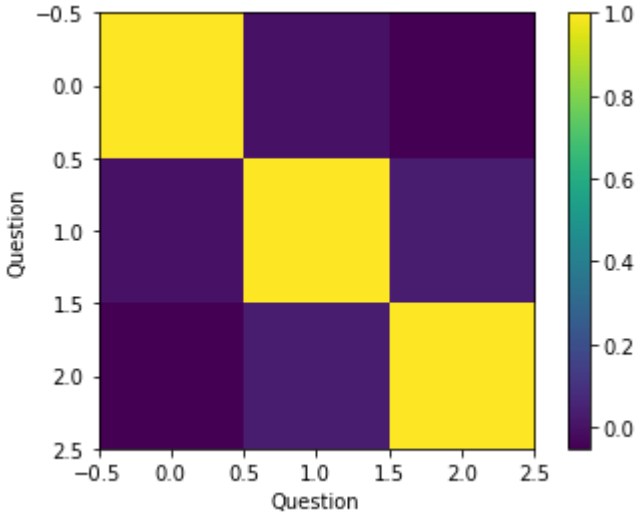


Figure 15: Correlation Matrix for Gender Identity, Sibling, Social Viewing Preference.

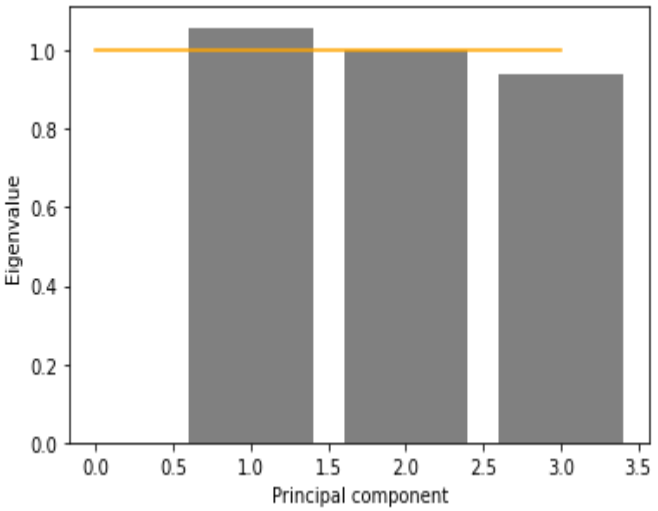


Figure 16: The Principal Component for Gender Identity, Sibling, Social Viewing Preference.

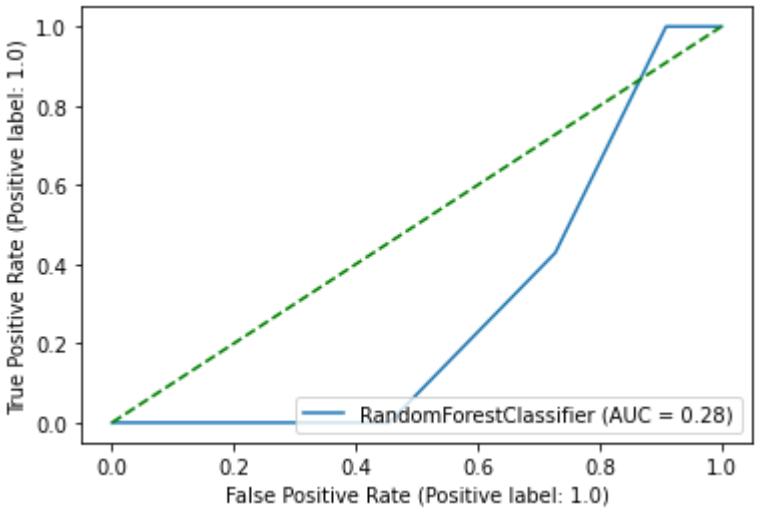


Figure 17: ROC curve and AUC Q9

<u>10) Build a prediction model of your choice (regression or supervised learning) to predict movie ratings (for all 400 movies) from all available factors that are not movie ratings (columns 401- 477). Make sure to use cross-validation methods to avoid overfitting and characterize the accuracy of your model.</u>

- This question builds on the algorithm of question 8. Here, we are predicting movie ratings based on all factors. Therefore, I ran a PCA on all the features and extracted 18 principal components using the Kaiser criterion. Then, I use that data to build a Random Forest, using 20% of the data as a test set and the other 80% is the train size. This yields a 53% accuracy with an AUC value of 0.48.

- So my model predicts the outcome 53% of the time. This is not the highest percentage of accuracy I should get. In the future, I can try splitting the train set and test set differently or using different kind of prediction models.

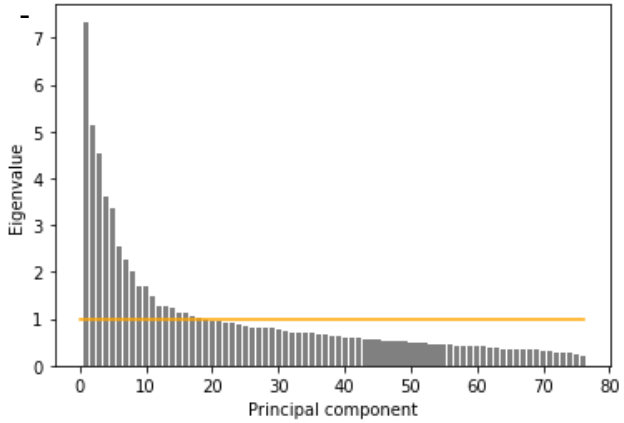


Figure18: Screeplot for all feature PCA

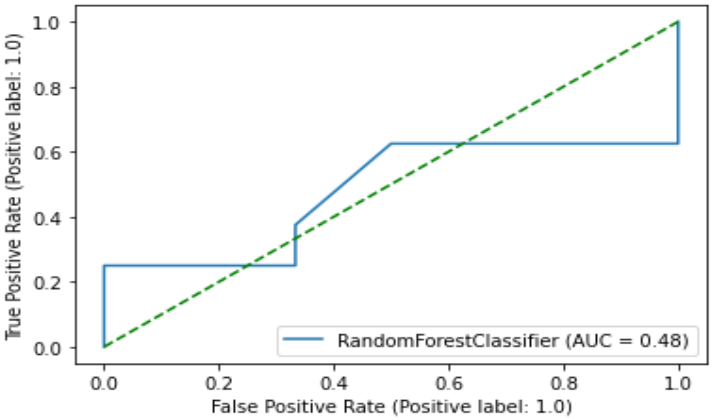


Figure 19: ROC curve and AUC Q10

Extra credit: Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions.

One interesting thing about this dataset is that the viewing ratings for The Wolf of Wallstreet is different among only children and people who have siblings. This is one particular interest of mine because my friend (who is an Only Child) views the film drastically differently from me (who has 3 siblings). While I regard Jordan Belfort as an ambitious and daring man who is willing to take risks, my friend views him as an irresponsible and unorthodox person. I think this stems from the fact that as an only child, she has to take on more responsibility. And as a result, she relies more on orthodox thinking and following her parents path. While I am more free to do whatever I want and I don't have that mental constraint she has. Putting this in our family's strict Asian family background, I think this would be a plausible reason. In the dataset, I use the Mann-Whitney test to find the U-value and P-value. As a result, the p-value is approximately 0.4. Given the alpha level of 0.05 or 0.005, we fail to reject the null hypothesis. So, there are no difference in the viewing and rating of the film between Only Child and People who have Siblings.

This is to show that my example can be one small outlier example compared to this dataset. Or, our graphical background is different. Maybe if I have the possibility to record data in my hometown, the result could be different.

Thank you very much !