Cyndi Liu
05/11/2023

Capstone Project

For this project, I used both NumPy and Pandas to deal with the data. To read the .csv files, I used NumPy to read the user data ("theData.csv") directly as it only contains numerical values in the file. Then I used Pandas to read the data describing the art pieces ("theArt.csv") and converted the dataframe to a NumPy array as it contains non-numerical values in the file. As most of the questions in this project involve users' preference ratings on art pieces, I extracted all preference ratings into a Numpy array and then checked if there are any NaNs in the array. I found that there is no NaNs in each column, so I used the original "preference_ratings" data for the project. To do dimension reduction for the dataset, I used Principal Component Analysis (PCA) when needed. I cleaned the data and removes NaNs for each question individually in order to minimize the impact of data loss. I cleaned the data row-wise, which means I remove the whole row if there is a NaN.

1) <u>Is classical art more well liked than modern art?</u>
   As preference ratings for art pieces are individual because they are rated based on each different art piece, and each art piece is unique. One might rate 1 and 7 for two pieces; the other might rate 4 and 4 for the same pieces. Both of them have the same mean of 4. Therefore, taking the mean here is not a good idea. Further, I conduct EDA to check the distribution of the data, they are not normally distributed. Thus, I chose to conduct the one-tailed Mann-Whitney U test that compares the medians from the two groups – preference ratings for classical art and for modern art, to see if the ratings for classical art are higher than for modern art. Computing the Mann-Whitney U test gave me a p-value = 1.5881633286154516e-97, which is smaller than the chosen alpha, 0.05. Therefore, the null hypothesis that classical art is not more well-liked than modern art is rejected. Classical art is more well-liked than modern art.

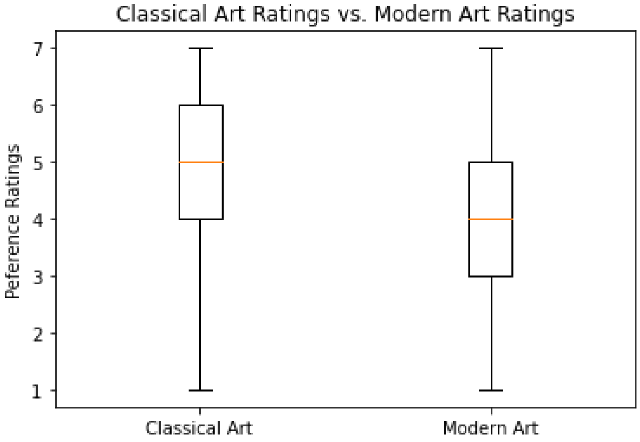


Figure 1: Box plot illustrating the median difference between ratings for classical art and modern art. It shows that ratings of classical art tends to spread higher than modern art.

2) Is there a difference in the preference ratings for modern art vs. non-human (animals and computers) generated art?
For the same reasons as in question 1, the data cannot be reduced to mean, and the distributions are not normal, I conduct the Mann-Whitney U test again to compare the medians of ratings for classical art and non-human art to see if a difference between their preference rating is significant. Running the two-tailed Mann-Whitney U test gave me a p-value=0.0, which is less than 0.05. Therefore, the null hypothesis that there is no difference in the preference ratings for modern art and for non-human art is rejected. There is a difference between the preference ratings for modern art and for non-human art.

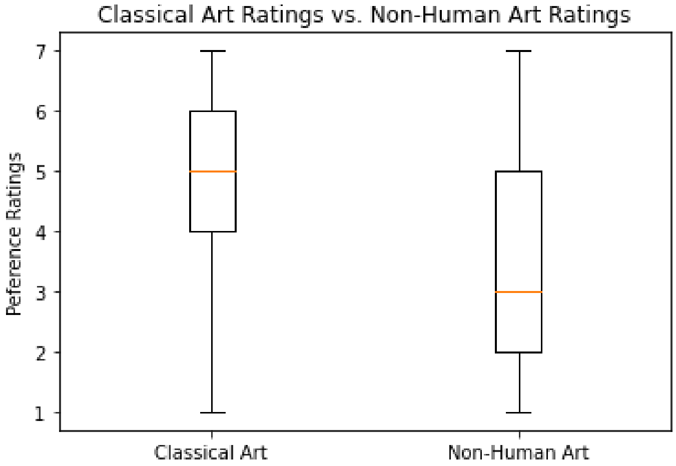


Figure 2: Box plot illustrating the difference between ratings for classical art and non-human art. It can be seen that the overall rating of non-human art is lower than classical art.

3) Do women give higher art preference ratings than men?
To separate women's ratings and men's ratings, I stored the women's ratings in an array if the value in the gender column from "user_data" is 2, and the men's ratings in another array if the value is 1. Then, I flatten them into 2 1-column arrays. For the same reasons as in question 1, the data cannot be reduced to mean, and the distributions are not normal, I conduct the one-tailed Mann-Whitney U test again to compare the medians of men's and women's preference ratings and see if women give higher ratings than men. Running the Mann-Whitney U test gave me a p-value=0.8643548652654933, which is >0.05, so the null hypothesis, which women do not give higher art preference ratings than men is not rejected. Thus, women do give higher art preference ratings than men.

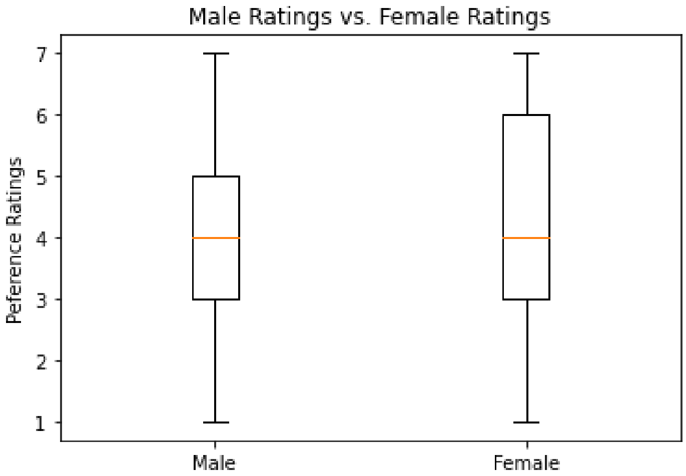


Figure 3: Box plot illustrating the different ratings from males and females. It can be seen that even though the two datasets have an equal median, the spread of females' ratings is higher than males'.

4) Is there a difference in the preference ratings of users with some art background (some art education) vs. none?
To determine whether the user has no art background or some art background, I classify the none-background group if the value is 0 in the "art education" column, and the some-background group if the value is greater than 0. I store these two groups in 2 1-D arrays. For the same reasons as in question 1, the preference rating data cannot be reduced to mean, and the distributions are not normal, I conduct the two-tailed Mann-Whitney U test to compare the median ratings from the none-art-background group and some-art-background group to see if the difference in the preference ratings from the two group is statistically significant. Running the Mann-Whitney U test gave me a p-value=3.0567413101500694e-09, which is way less than 0.05, so the null hypothesis that there is no difference in the preference ratings of users with some art background and no art background. There is a difference in preference ratings of users with some art background vs. none.

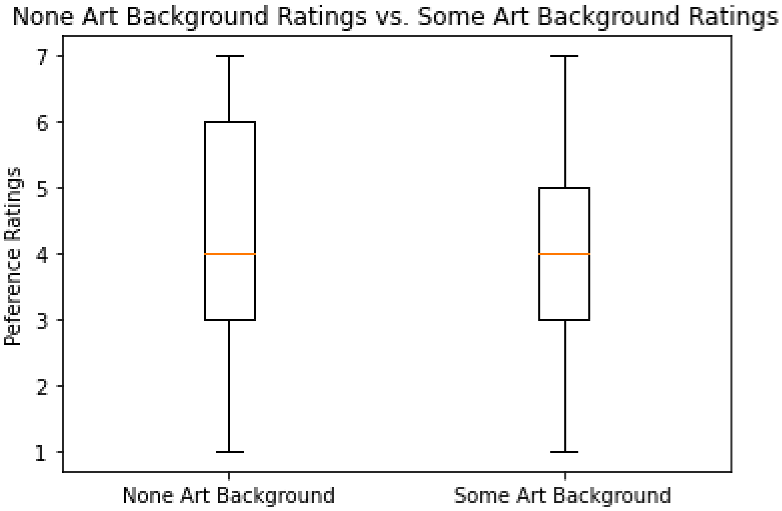


Figure 4: Box plot illustrating the difference in preference ratings between the non-art-background group and some-art-background group. It shows that non-art-background people usually rate higher than some-art-background people.

5) <u>Build a regression model to predict art preference ratings from energy ratings only. Make sure to use cross-validation methods to avoid overfitting and characterize how well your model predicts art preference ratings.</u>
   For this question, I input the users' energy rating means as predictors and the users' art preference rating as outcomes to build the regression model. As there is only one predictor, I chose to build a linear regression model. Once I got the linear regression model, I calculated the RMSE to assess the model. I got RMSE=0.600, meaning that the standard deviation of prediction error is 0.600. I also got R^2=0.068, meaning that the regression model only explains 6% of the variability of the art preference ratings.

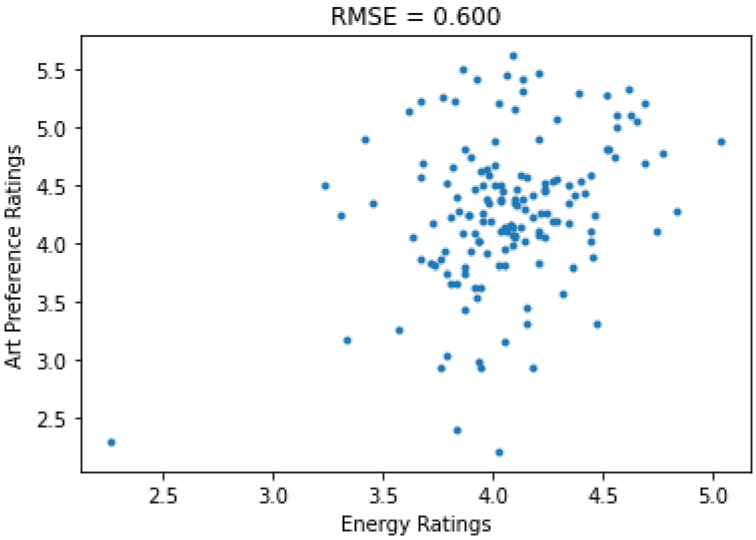


Figure 5: Linear regression model that predicts art preference ratings from energy ratings.

6) Build a regression model to predict art preference ratings from energy ratings and demographic information. Make sure to use cross-validation methods to avoid overfitting and comment on how well your model predicts relative to the "energy ratings only" model.
For this question, I input the demographic information and users' mean of energy ratings as predictors and art preferences ratings as outcomes. I built a multiple regression and a ridge regression model. For the multiple regression model, I got RMSE=0.663, meaning that the standard deviation of prediction error is 0.663. For the ridge regression model, I got the R^2=0.153, meaning that the model only explains 15.3% of the variability of the actual art preference rating. Compare to the "energy rating only" regression model, this ridge regression model performs better because its R^2 is greater than the R^2 of the "energy rating only" model, meaning that it accounts for more proportion of the variability of art preference rating.

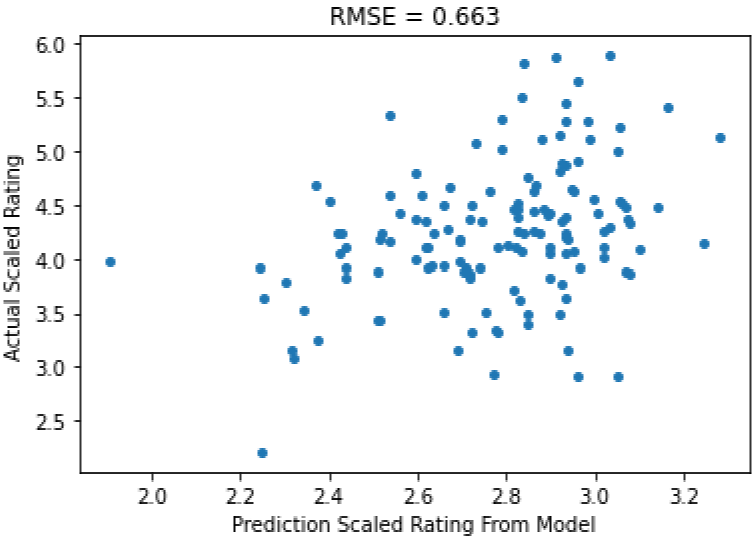


Figure 6: Multiple regression model predicting the art preference rating from demographic information and energy rating.

7) Considering the 2D space of average preference ratings vs. average energy rating (that contains the 91 art pieces as elements), how many clusters can you – algorithmically - identify in this space? Make sure to comment on the identity of the clusters – do they correspond to particular types of art?
First, I stored the average preference ratings and average energy ratings of 91 art pieces in a 2D array. In order to determine how many clusters, K, to ask for, I used the silhouette method and define the range of plausible values K as 2-9. Running the silhouette method,

I got histograms of s(i) values of all data points for all values of K. After getting the histograms of the silhouette coefficients, I plot a line graph that shows the sum of the silhouette scores as a function of the number of clusters.

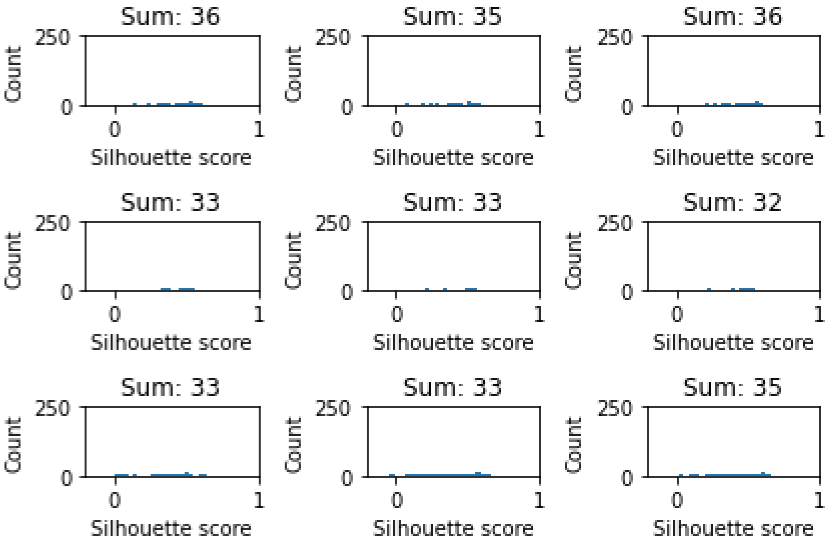


Figure 7: The silhouette coefficient histograms for all 9 values of K.

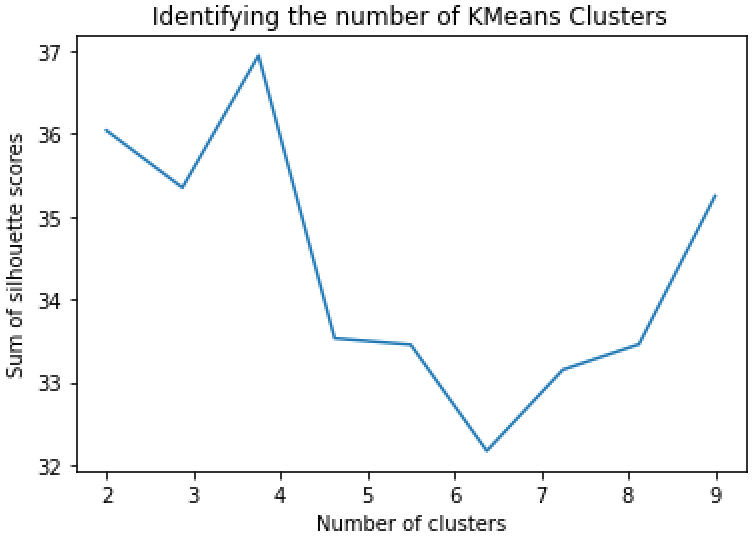


Figure 8: Line graph of the sum of silhouette scores for each value of clusters.

From figure 8, it can be seen that the peak of the line graph occurs when the number of clusters is 4, which means the sum of the silhouette score has the highest value when there are 4 clusters. Therefore, 4 clusters allow for optimal clustering in the 2D space of average preference ratings vs. average energy ratings.

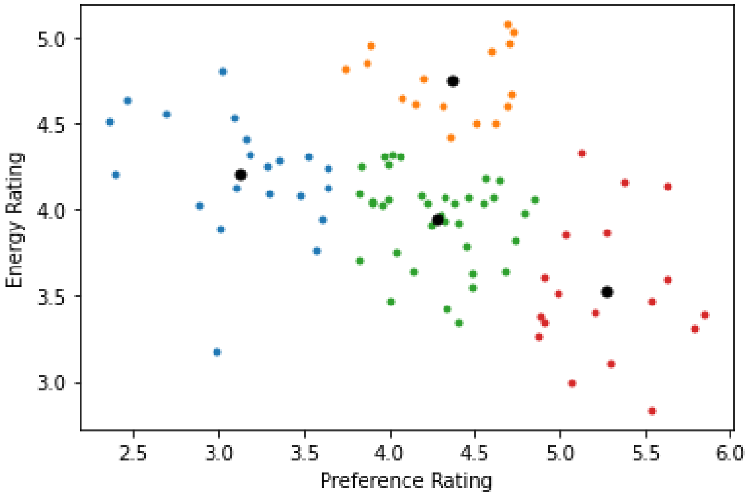
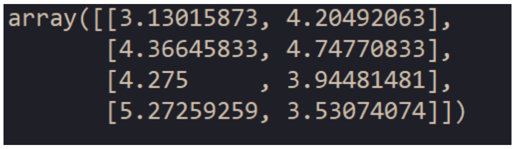
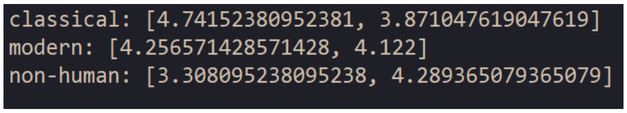


Figure 9: Color-coded average preference ratings vs. average energy ratings data after clustering.

After determining the number of clusters, I got the color-coded data after clustering in 4 clusters. In order to determine which particular art type each cluster is corresponding to, I took the means of preference ratings and energy ratings for the 3 art types (classical art, modern art, and non-human art). Then compare them to the "cCoords" array, which is the array of the centroids for each cluster, to see which pairs are similar. By comparing and locating the means of preference rating and energy rating from the 3 art types in the graph, I determined that the orange region is classical art, the green region is modern art, and the red region is non-human art.



```
array([[3.13015873, 4.20492063],
       [4.36645833, 4.74770833],
       [4.275     , 3.94481481],
       [5.27259259, 3.53074074]])
```



```
classical: [4.74152380952381, 3.871047619047619]
modern: [4.256571428571428, 4.122]
non-human: [3.308095238095238, 4.289365079365079]
```

Figure 10: The centroid of each cluster (left) and the average of preference ratings and energy ratings for the 3 art types (right).

8) Considering only the first principal component of the self-image ratings as inputs to a regression model – how well can you predict art preference ratings from that factor alone?

To answer this question, I computed PCA to approach dimension reduction in order to get the first principal component of the self-image ratings.

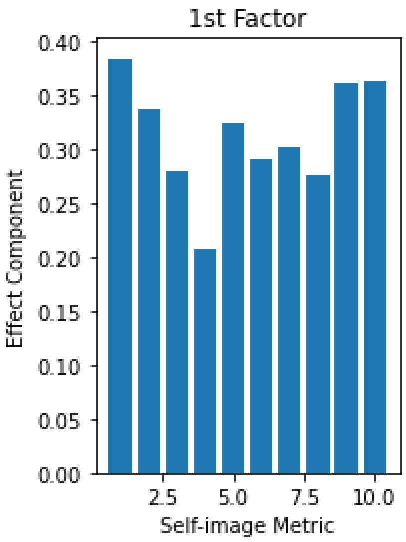


Figure 11: The loading of the first principal component

After getting the loading matrix of the first principal component, I used the original data in new coordinates as input to build a linear regression model to predict art preference ratings. Once I got the regression model, I calculated the RMSE to assess the model. The RMSE, the standard deviation of the prediction error, is **0.619**.

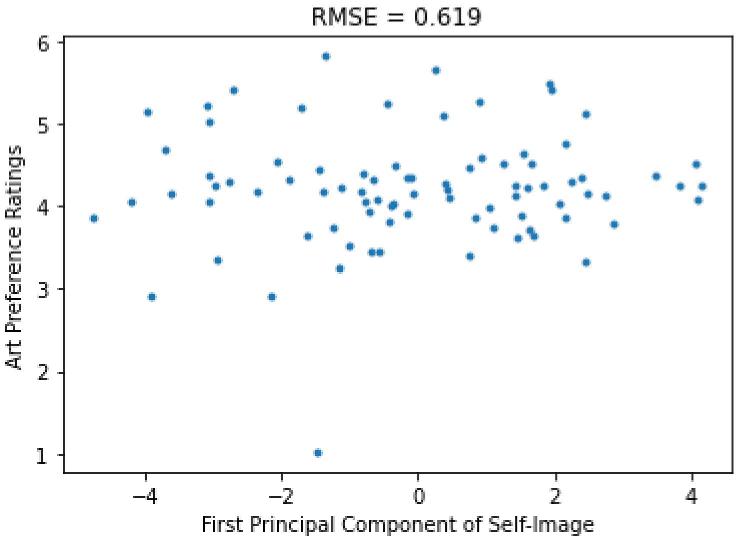


Figure 12: Linear regression model that inputs the first principal component of self-image to predict the art preference ratings.

9) Consider the first 3 principal components of the "dark personality" traits – use these as inputs to a regression model to predict art preference ratings. Which of these components significantly predict art preference ratings? Comment on the likely identity of these factors (e.g. narcissism, manipulativeness, callousness, etc.).

In order to get the first 3 principal components of the "dark personality" trains, I conducted PCA to do dimension reduction of the "dark personality" traits. After getting the loading matrix of the first principal component, I used the original data in new coordinates as input to build a multiple regression model to predict art preferences ratings from the first 3 principal components. Once I got the multiple regression model, I calculated the RMSE=0.622, meaning that the standard deviation of the prediction error is 0.622. I also compute LASSO regression to check which components significantly predict art preference ratings. However, by running the LASSO regression model, it gives me that all 3 coefficients equal 0, meaning that none of these coefficients deem to significantly predict the art preference ratings.

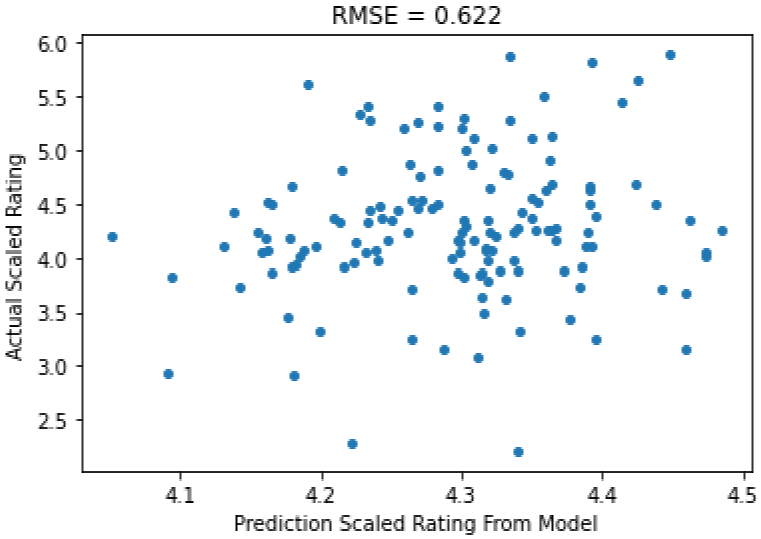


Figure 13: The multiple regression model predicting art preference ratings from the first 3 principal components of "dark personality" traits.

By looking at the heights of columns from the graphs of loading matrices for the first 3 principal components, I can conclude that the first component is the overall "darkness" of personality; the second component takes account of questions 9, 10, and 11, which points to narcissism; the third component takes care of questions 7 and 8, which points to callousness.

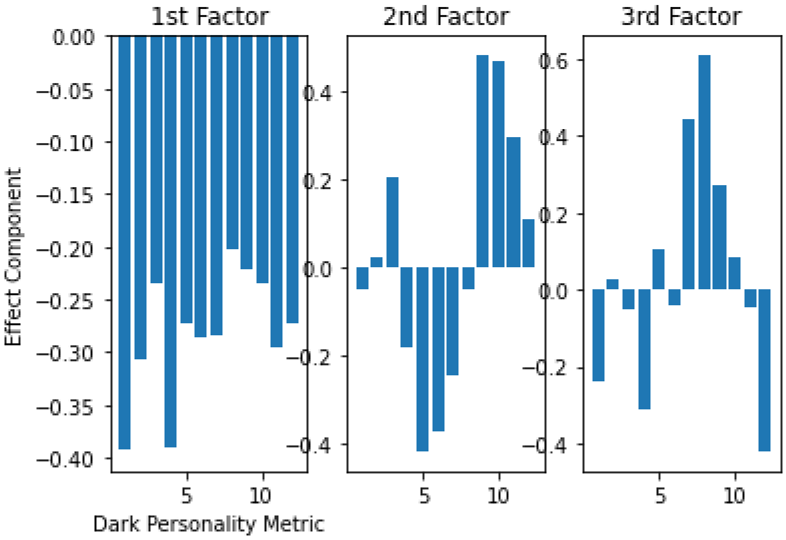


Figure 14: Loading matrices for the first 3 principal components

10) Can you determine the political orientation of the users (to simplify things and avoid gross class imbalance issues, you can consider just 2 classes: “left” (progressive & liberal) vs. “non-left” (everyone else)) from all the other information available, using any classification model of your choice? Make sure to comment on the classification quality of this model.

For this question, to determine/predict the political orientation of the users, I decided to build models that predict the political orientation from self-image, dark personality traits, and action preferences respectively. I store users' political orientation in a NumPy array in which 0 represents non-left and 1 represents left orientation. Therefore, I will build 3 random forest models to determine the political orientation from these predictors. I first compute PCA for the 3 predictors.

a) Predicting political orientation from self-image:
From the scree plot, it can be seen that the eigenvalue for the first 2 principal components is higher than 1. Therefore, based on the Kaiser criterion, I will use the first 2 components of self-image as input to predict the political orientation.

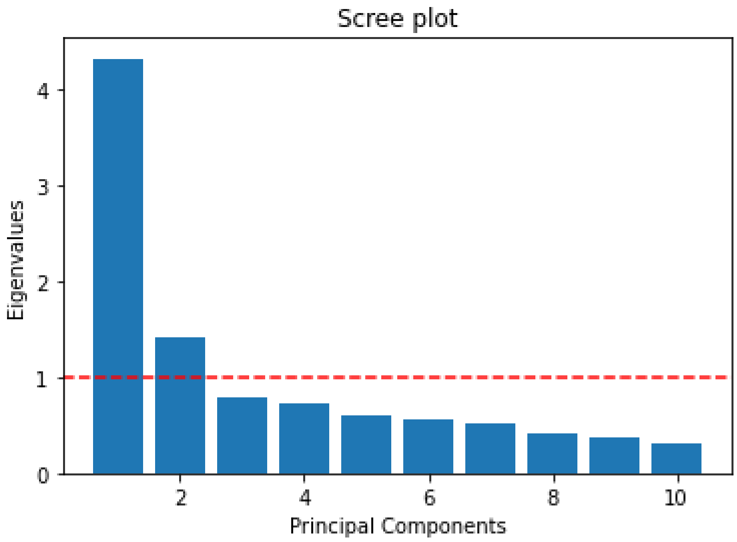


Figure 15: Screeplot of principal components of self-image.

After conducting the random forest model using RandomForestClassifier from sklearn.ensemble, I got the accuracy of the random forest model as 0.5178571428571429, which is 52% accuracy. Then I got the confusion matrix of the model to visualize the confusion matrix.

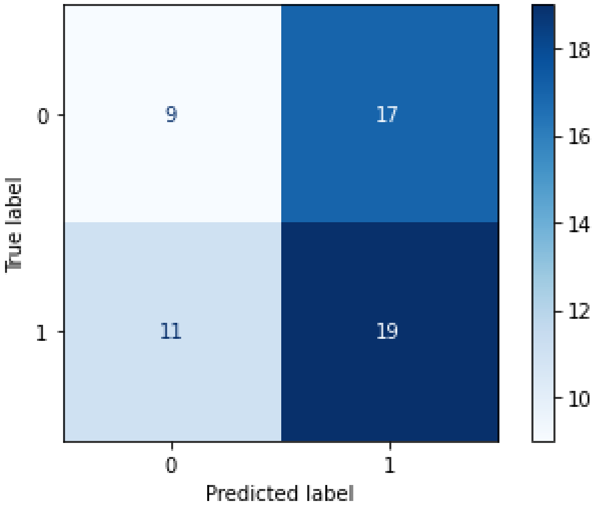


Figure 16: Confusion matrix of random forest model for political orientation vs. self-image.

b) Predicting political orientation from action preference:
From the scree plot, it can be seen that the eigenvalue for the first 3 principal components is higher than 1. Therefore, based on the Kaiser criterion, I will use the first 3 components of action preference as input to predict the political orientation.

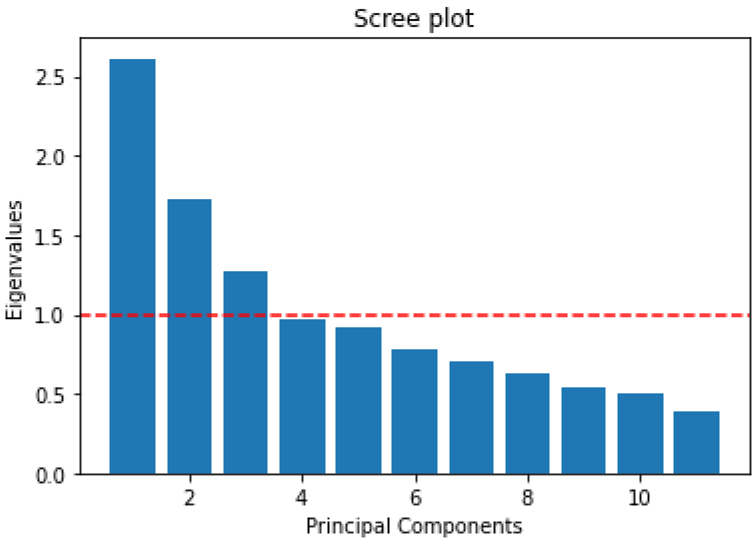


Figure 17: Screeplot of principal components of action preference.

After conducting the random forest model using RandomForestClassifier from sklearn.ensemble, I got the accuracy of the random forest model as 0.6964285714285714, which is 70% accuracy. Then I got the confusion matrix of the model to visualize the confusion matrix.

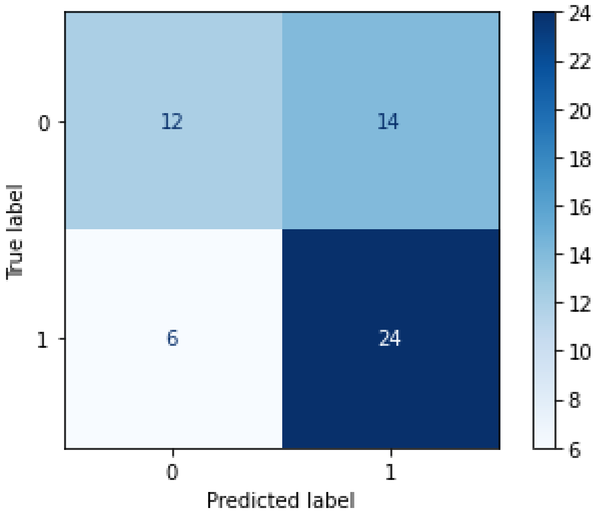


Figure 18: Confusion matrix of random forest model for political orientation vs. action preference.

c) Predicting political orientation from dark personality traits:
From the scree plot, it can be seen that the eigenvalue for the first 3 principal components is higher than 1. Therefore, based on the Kaiser criterion, I will use the first 3 components of dark personality traits as input to predict the political orientation.

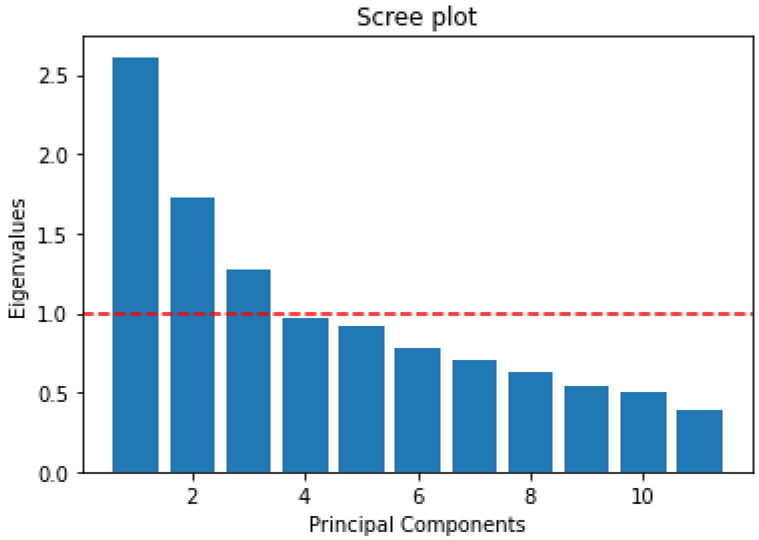


Figure 19: Screeplot of principal components of dark personality traits.

After conducting the random forest model using RandomForestClassifier from sklearn.ensemble, I got the accuracy of the random forest model as 0.44642857142857145, which is 45% accuracy. Then I got the confusion matrix of the model to visualize the confusion matrix.

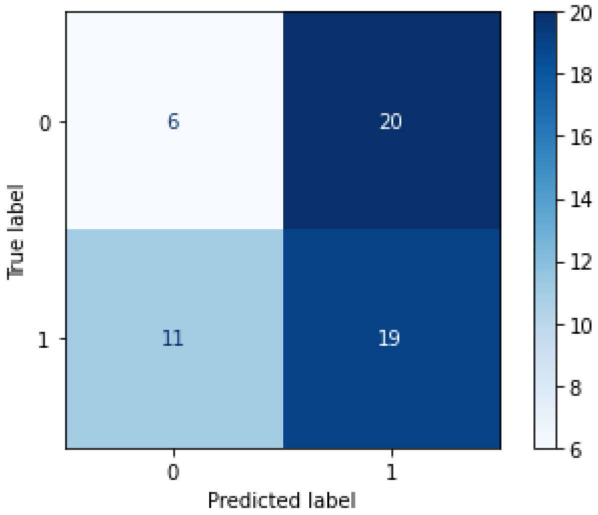


Figure 20: Confusion matrix of random forest model for political orientation vs. dark personality traits.