Name: Suvir Wadhwa
Date: 20th December 2022
N-Number: 16395336

Introduction to Data Science: Capstone Project

As I worked through the questions for the Capstone Project, there were several instances where I has to clean the data, sort it, and structure it in a better way. Hence, to adhere to this, whenever needed, I would drop all the NaNs . There were also segments that had several dimensions of data. I performed PCA (Principal Component Analysis) to adhere to this.

The data for the user rating's did not have a header. Therefore, I had to set up a header for it. Similarly, the data for the art details had very long column names. To deal with this, I renamed the columns to make working with them easier. I used the means of certain data when needed to dealing with shaping issues for regressions and comparisons.

Question 1: Is Classical art more well liked than modern art?

To find out whether classical art is more well liked than modern art, I compared the means of the preference ratings for both of the categories. To do this, I used the art data to determine what ratings rows are for classical art and modern art, respectively. Following that, I took to mean of every column for each of the categories. To make comparisons easier, I took the mean of the means. I did the same as the aforementioned with the medians. My findings were:

Mean for Classical Music: 4.74
Mean for Modern Music: 4.26

Median for Classical Music: 5
Median for Modern Music: 4

As seen above the mean and median preference rating for classical music were higher, indicating classical could be more liked. The mean comparison can be represented in the bar graph below.

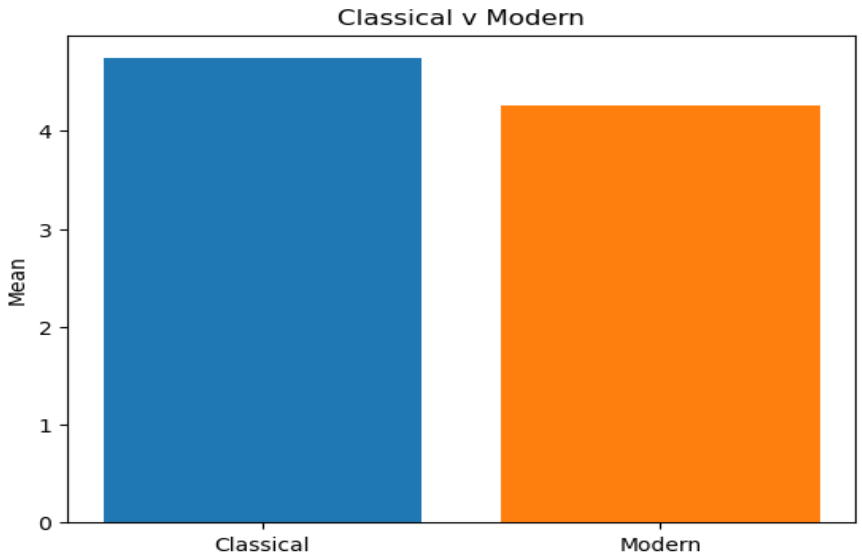

Figure 1: Difference between Classical and Modern Art Preferences

Question 2: Is there a difference in the preference ratings for modern art vs. non-human (animals and computers) generated art?

To find out whether there is a difference in the preference ratings for modern and nonhuman, I compared the means of the preference ratings for both of the categories. To do this, I used the art data to determine what ratings rows are for modern art and nonhuman art, respectively. Following that, I took to mean of every column for each of the categories. To make comparisons easier, I took the mean of the means. I did the same as the aforementioned with the medians. My findings were:

Mean for Modern Art: 4.26
Mean for Non-Human Art: 3.31

Median for Modern Art: 4
Median for Non-Human Art: 3

As seen above the mean and median preference rating for modern art were higher. The mean comparison can be represented in the bar graph below.

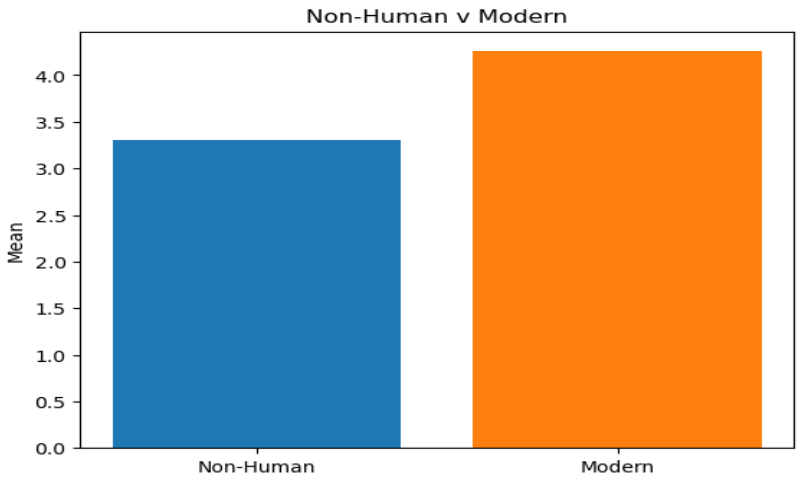

Figure 2: Difference between Non-Human and Modern Art Preferences

Question 3: Do women give higher art preference ratings than men?

To find out whether there is a difference in the preference ratings for Women and Men, I compared the means of the preference ratings for both of the categories. To do this, I used the gender column in theData.csv to determine what ratings rows are for Women and Men, respectively. Following that, I took to mean of every column for each of the categories. To make comparisons easier, I took the mean of the means. I did the same as the aforementioned with the medians. My findings were:

Mean for Women: 4.23
Mean for Men: 4.21

Median for Women: 4
Median for Men: 4

As seen above the mean preference rating for women was a higher by 0.02. This difference isn't large enough to conclude that Women had higher ratings than men.

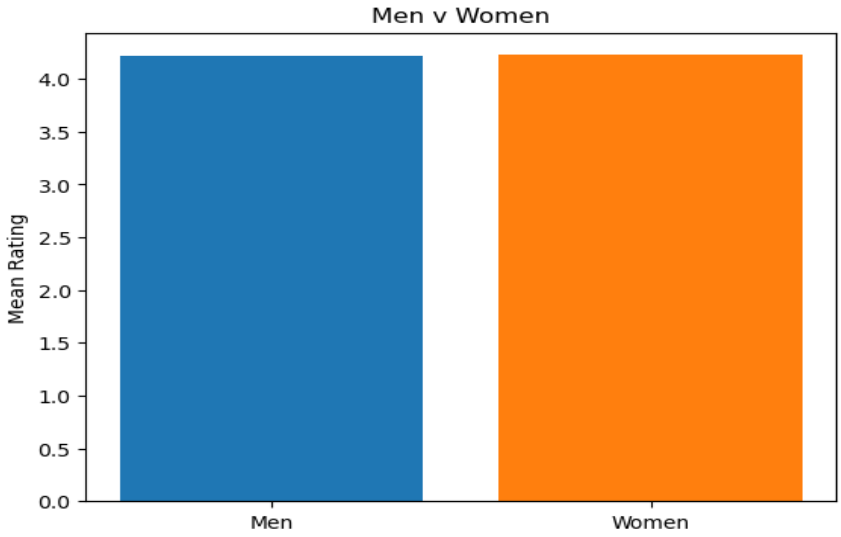

Figure 3: Difference between Preference Ratings from Men and Women

Question 4: Is there a difference in the preference ratings of users with some art background (some art education) vs. none?

To find out whether there is a difference in the preference ratings for users with some experience and no experience, I compared the means of the preference ratings for both of the categories. To do this, I used the experience column in theData.csv to determine what ratings rows are with no experience and used the other 2 for some experience, respectively. Following that, I took to mean of every column for each of the categories. To make comparisons easier, I took the mean of the means. I did the same as the aforementioned with the medians. My findings were:

Mean for Some Experience: 4.19
Mean for No Experience: 4.31

Median for Some Experience: 4
Median for No Experience: 4

As seen above the mean preference rating for people with no experience were higher, while the median was the same. This could perhaps be because they have less criteria when judging.

However, the difference wasn't large enough to come to a solid conclusion on the ratings. The mean comparison can be represented in the bar graph below.

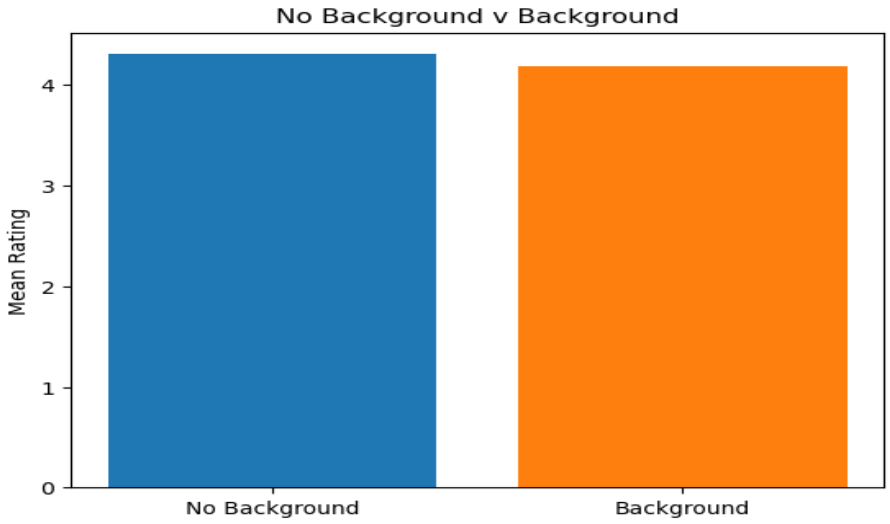

Figure 4: Difference in Preference Ratings from Users Without and With Art Background

Question 5: 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.

To predict art preference ratings from energy ratings, I used a linear regression. I found the mean rating for each of the 91 columns in the energy and preference ratings, respectively. I then carried out a train and test split for cross validation. As mentioned in the question, this helps with over fitting. My random state was the seed value generated from my N-Number and my test size was 30%. Following this I plotted a scatter plot with a line of best fit to visualize the results.

The measures to test how well the model predicts were:

Average MSE (Mean Squared Error): 0.515

Average MAE (Mean Absolute Error): 0.568

The plot below represents the relationship between the two ratings.

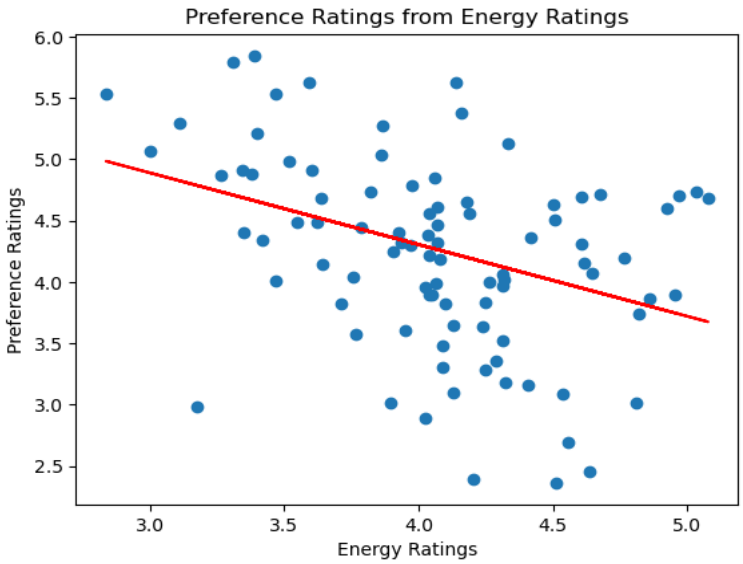

Figure 5: Scatter Plot Representing Preference Ratings from Energy Ratings

Question 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.

To predict art preference ratings using the energy and demographic ratings, I used a linear regression. Using KFolds, I was able to run 8 cycles of the linear regression, enabling me to get the r^2 score. The KFolds model had a random state of my seed and enabled shuffling. Finally, to get the r^2 score, I used a cross validation model, enabling me to prevent overfitting as required.

The **Average R^2 score** was: 0.74

I used the following box plot to represent the distribution of the samples.

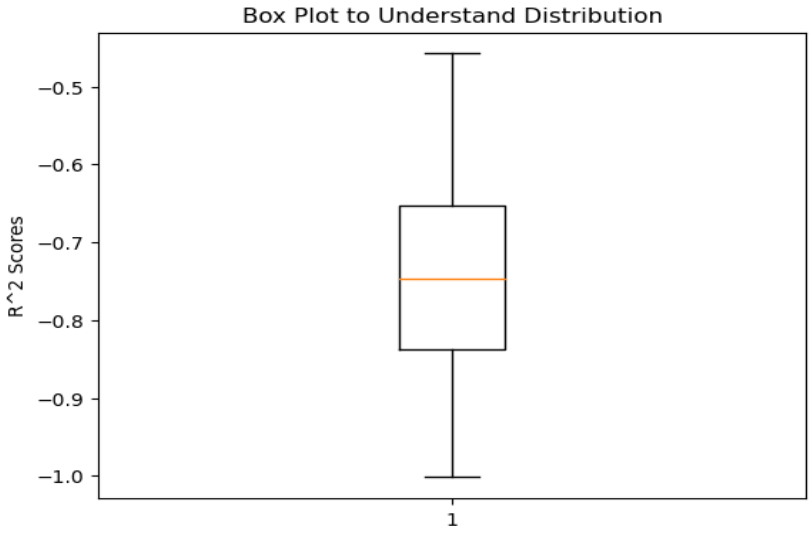

Figure 6: Box Plot

Question 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?

To do the above question, I did a Kmeans scatter plot. To start, I took the mean of each of the 91 ratings for preference and energy ratings respectively. It was important to determine the number of k values required for the kMeans. To do this I tested to silhouette score for different score and found the k value with the highest sum value. The k score I found was three using the graphs below.

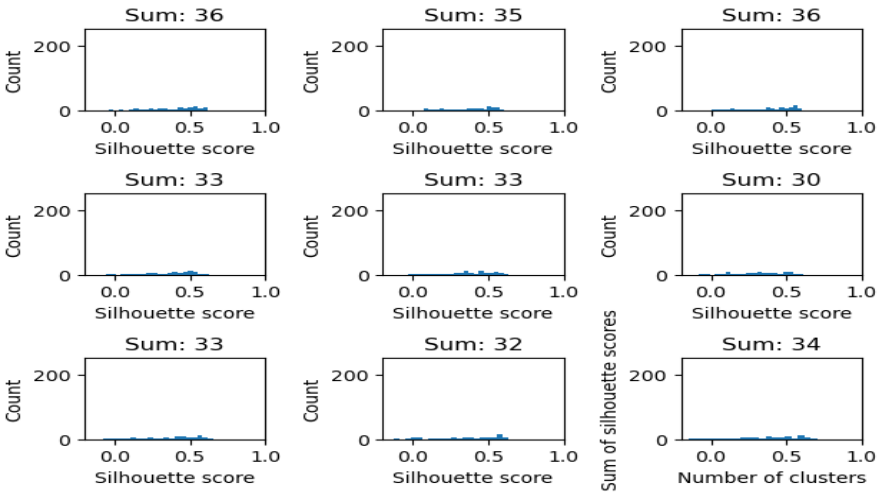

Figure 7: Silhouette Sum Plots

To determine the identity of the clusters, I plotted 2 different cluster graphs. One with the predicted ratings data and one with the known.

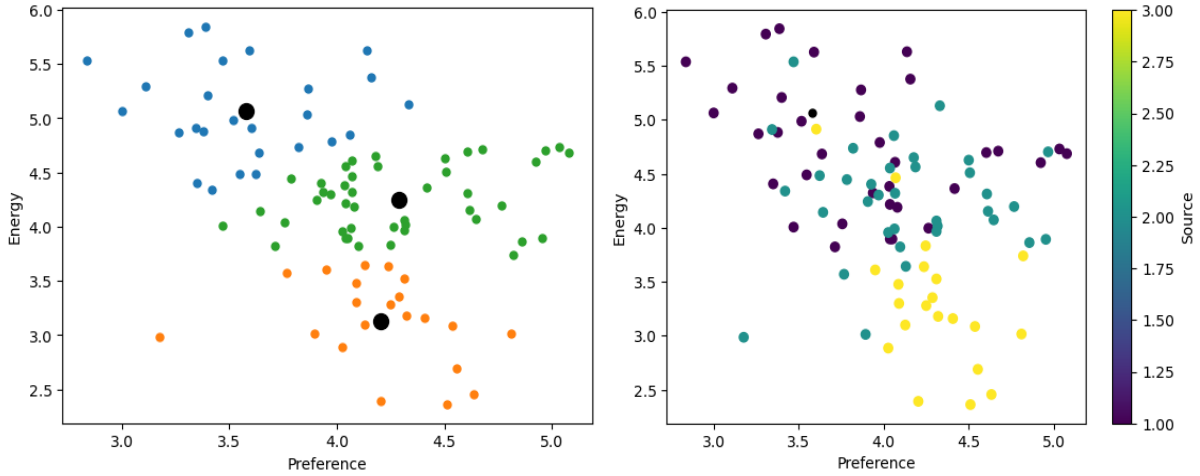

Figure 8: Comparison of Predicted Scatter and Known Scatter

Using the source data in the Arts data frame: 1 = Classical Art, 2 = Modern Art, and 3 = Non-Human Art

Comparing the two plots we have above, we can see some similarities between the predicted scatter plot and the known scatterplot. The blue group in the predicted plot corresponds to

Classical art in the Known Scatter. The green group in the predicted plot corresponds to Classical art in the Known Scatter. Finally, the orange group in the predicted plot corresponds to non-human art in the Known Scatter.

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 calculate the first PCA, I took the first component from the PCA matrix for the self-image ratings data. I then plotted a linear regression for between the first PCA and the means of all the art preference ratings. I used cross validation to ensure I could prevent overfitting.

The below plot represents the strength of each PCA and its variance on the predictions.

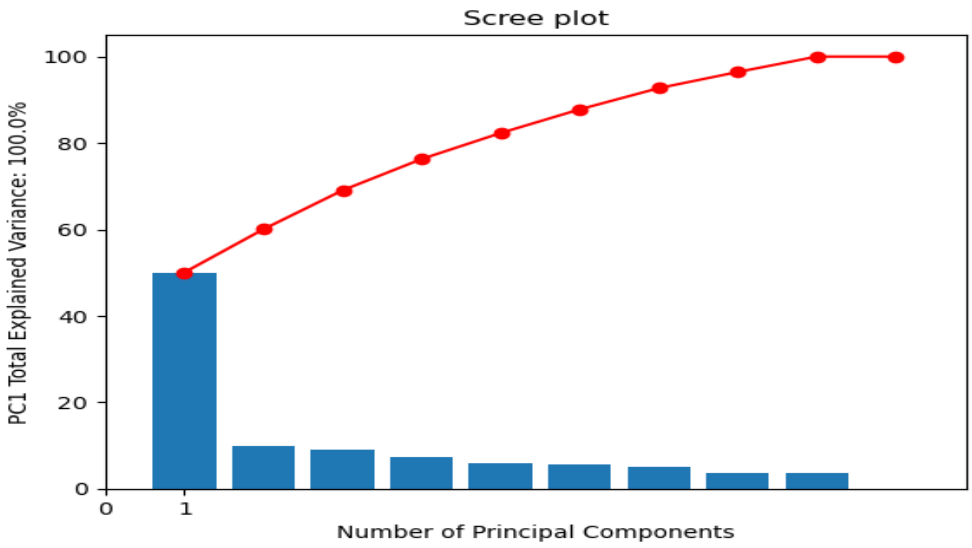

Figure 9: PCA Analysis

The measures to test how well the model predicts were:

Average MSE (Mean Squared Error): 0.476

Average MAE (Mean Absolute Error): 0.462

R^2: 0.054

The plot below represents the relationship between the two ratings.

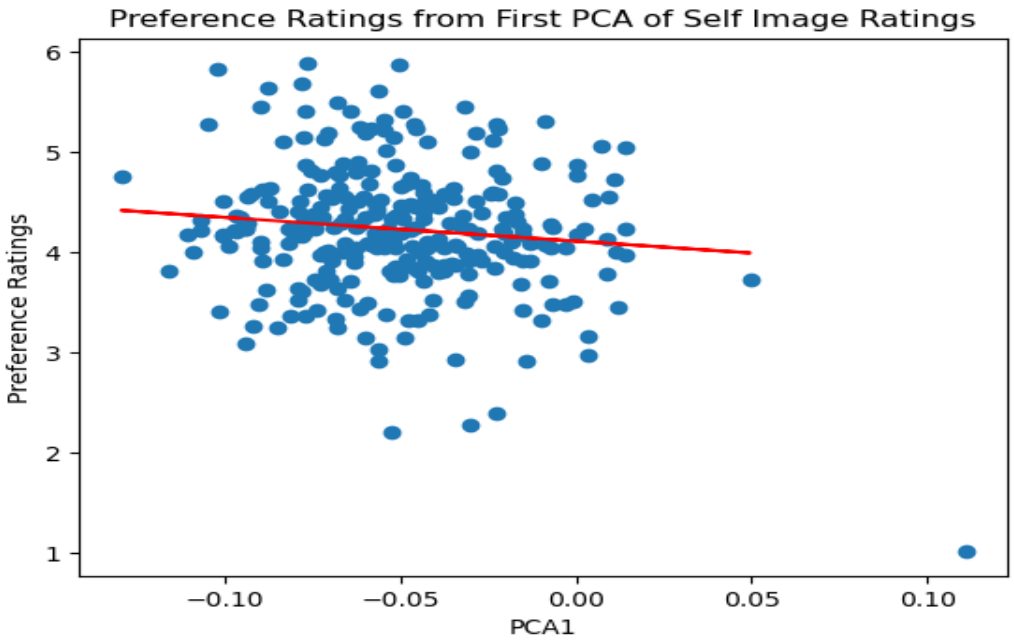

Figure 10: Preference Ratings from PCA2 of Self Image

Looking at the scatter plot above that shows the relation between the first component of self-ratings and the preference ratings, we can see that there isn't much of a relation. The first component isn't the best predictor of preference ratings. This can also be seen from the R^2 score of 0.054.

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.)

To get the first 3 principal components of the dark personality traits, I set the PCA components on my PCA module to 3. I then used kFold to take 10 different splits of the data set for my model. The kFold had a random state of my seed value and had shuffling enables. I then plotted a linear regression to predict the art preference ratings from my PCA data.

The below plot represents the strength of each PCA and its variance on the predictions.

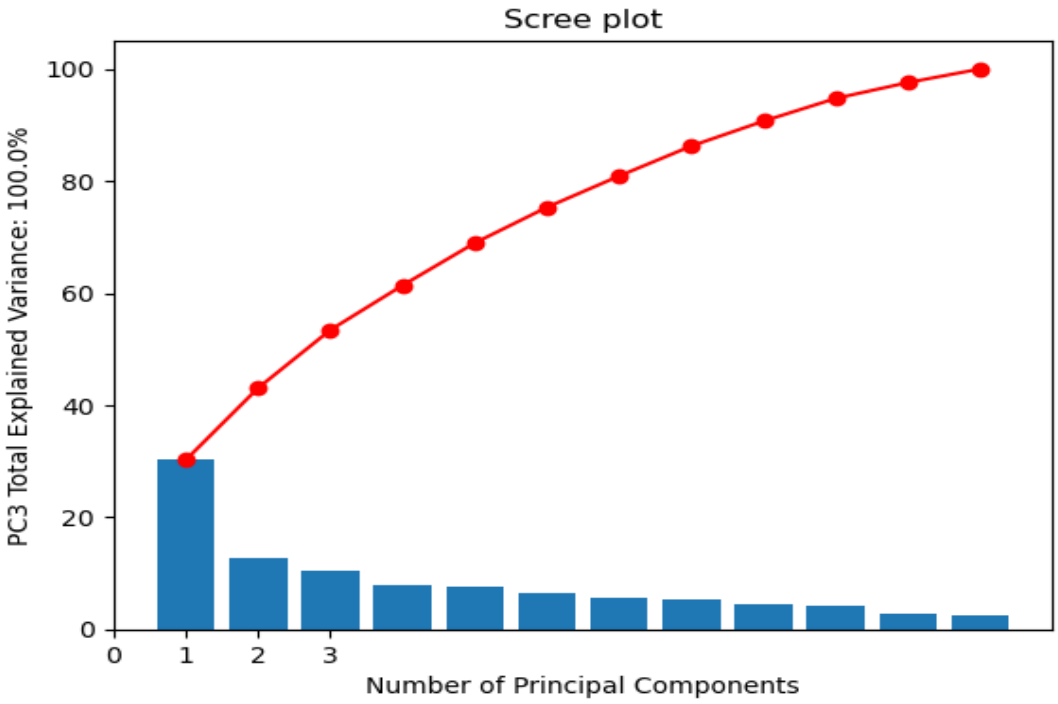

Figure 11: PCA Analysis

Below is a plot of all of the R^2 scores for the 10 KFold splits.

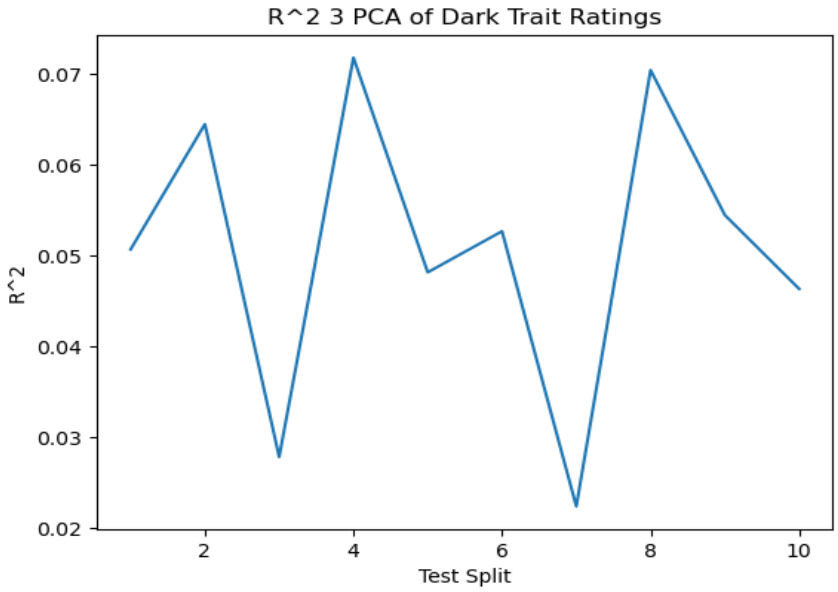

Figure 12: Dark Train R^2 Analysis

All of the R^2 scores are very low, indicating that the first 3 PCA aren't the best fit/predictors.

To determine the most significant PCA predictors I matched the max values of the PCA value predictors to the feature names.

Below are the identities of the 3 most significant predictors:

	0	1
0	PC0	I tend to manipulate others to get my way
1	PC1	I tend to lack remorse
2	PC2	I tend to be cynical
3	PC3	NaN
4	PC4	NaN
5	PC5	NaN
6	PC6	NaN
7	PC7	NaN
8	PC8	NaN
9	PC9	I tend to lack remorse
10	PC10	NaN
11	PC11	NaN

Figure 13: Significant Predictors

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.

To determine the political orientation of the users we need to use a classification model. I used a Random Forrest Classifier. To set up the political data, I set all of the left rows in the Political column to 0 and made the non-left set 1. For my predictor data, I used every other column in the ratings data set. I then used cross validation to ensure there is no overfitting of the data. Training the Random Forest with 200 tree estimator.

The measure of Accuracy was calculated using the score of the model.

Accuracy: 0.55

Below is a confusion matrix representing the Type 1 and 2 errors along with the accurate results.

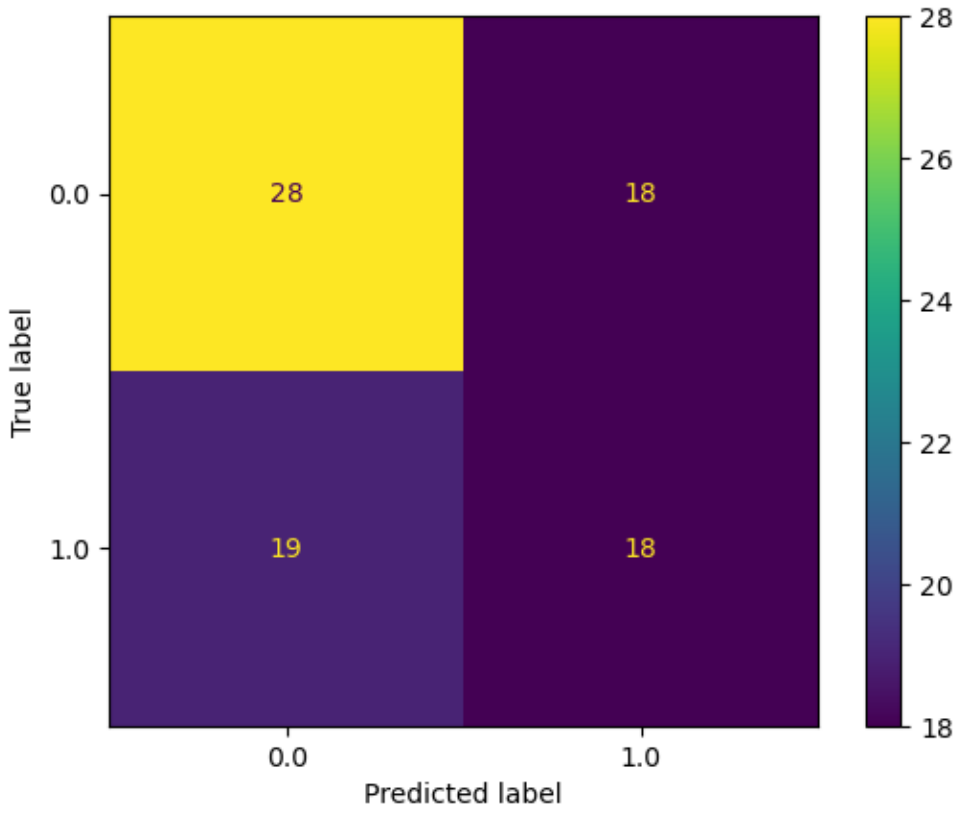

Figure 14: Confusion Matrix for Random Forest

Extra Credit: Compare the preference rating differences between women, men and non-binary.

To find out the difference in preference ratings for Women, Men, and Non-Binary, I compared the means of the preference ratings for all 3 of the of the categories. To do this, I used the gender column in theData.csv to determine what ratings rows are for Women, Men, and Non-Binary, respectively. Following that, I took to mean of every column for each of the categories. To make comparisons easier, I took the mean of the means. I did the same as the aforementioned with the medians. My findings were:

Mean for Women: 4.23

Mean for Men: 4.21

Mean for Non-Binary: 4.55

Median for Women: 4

Median for Men: 4

Median for Non-Binary: 4.5

As seen through the mean and median, there is a difference between the ratings for Non-binary in comparison to men and women. Non-Binary has a higher mean and median and hence could be said to have higher overall preference ratings.

The below graph represents the difference in means between the three genders.

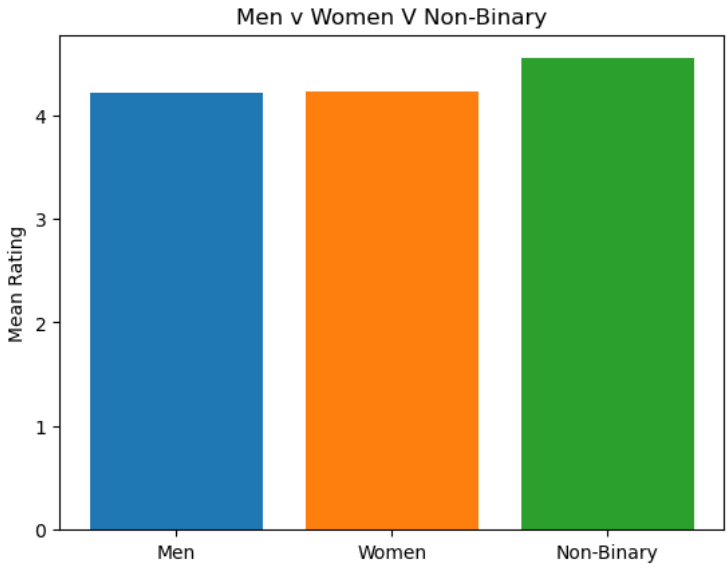

Figure 15: Difference between Preference Ratings from Men, Women, and Non-Binary