Olivia Lewandowski

Professor Wallisch

Intro to Data Science

11 May 2023

Capstone: Art Ratings

Preprocessing

Data preprocessing was performed insofar as each individual question required.

1. Dimensionality Reduction:
    a. I performed dimensionality reduction using PCA for questions 8 and 9 in order to reduce the self-image ratings and dark personality trait ratings to fewer, more principal components. I didn't choose which principal components, however, as the questions specified how many should be used for each regression.
2. Data Cleaning:
    a. There were NaN values within the user-specific questions (non art-rating questions), therefore I performed row-wise removal of the NaNs accordingly. For each analysis, NaNs would throw off not only the sample size but the analysis/comparison across variables, so it was necessary to remove the whole row. The dataset was generally easy to work with, though, as there were only around ~20 rows within the latter columns that contained NaNs.
3. Data Transformations:
    a. Standardization, or z-scoring of the data was necessary for PCA in questions 8 and 9, and also beneficial for question 6 in running multiple regression with many independent variables of different units.

1. Is classical art more well liked than modern art?

Firstly, the columns with classical art preference ratings (1-35) and the columns with modern art preference ratings (36-70) were divided into two distinct arrays - there was no need for NaN value removal, as none of the 70 columns included NaNs. Then, normality of the two distributions was tested using the Shapiro-Wilk test, which produced a p-value of ~0.0 for each distribution. As 0 is less than 0.05, it was concluded that both the classical art preference ratings and the modern art preference ratings were non-normal. The non-normality of the data, along with the ordinal nature of the data pointed to the necessity of a non-parametric test. Since both classical preference ratings and modern preference ratings came from the same exact subjects (300 users), a paired test was deemed appropriate due to the statistical dependence of the two variables at hand. As the data necessitated a non-parametric paired test, the Wilcoxon Signed-Rank Test was chosen over the paired t-test. The p-value elicited from the test was 1.6043337889324304e-117, infinitesimally smaller than the declared alpha-level of 0.05, therefore the null hypothesis, that there is no difference between the medians of the paired observations (median difference being zero), was rejected. As I performed a one-sided test (with classical ratings being group 1 and modern ratings being group 2), there was significant evidence to conclude that median classical art preference ratings were in fact higher than that of modern art preference ratings. In order to further determine if the classical art preference ratings were in fact higher, I observed both the ratings distributions and the medians of the two variables. From observing the two distributions, it is clear that classical ratings were skewed towards higher rating values, as classical ratings trump modern ratings for 4, 5, 6 and 7, while modern ratings are higher for 1, 2 and 3. Moreover, the median of the classical preference ratings was 5, while the median of the modern preference ratings was 4, further suggesting that classical art

preference ratings were higher. Therefore, the results from the Wilcoxon Signed-Rank Test, the two distributions, and the medians of the classical and modern art preference ratings do in fact suggest that classical art is more well liked than modern art (in respect to the scale of the data).

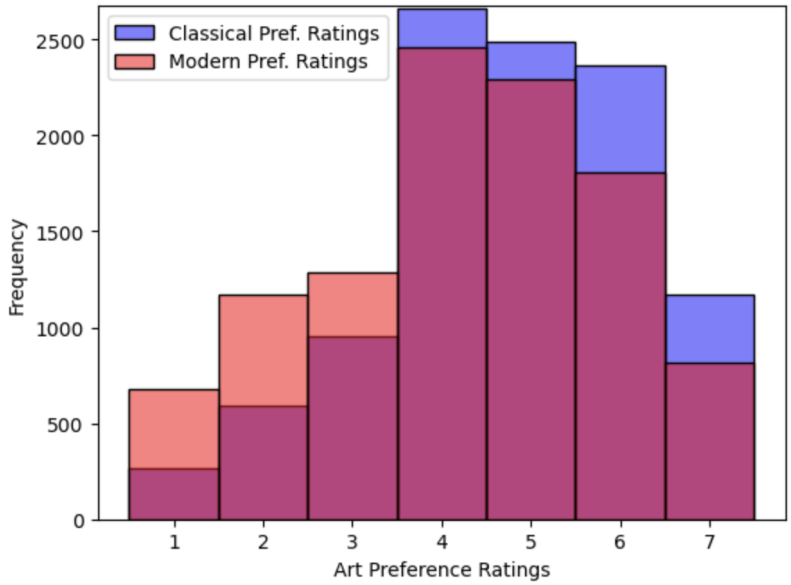


Fig 1: Histogram Plot of the Ratings Distributions (Classical vs. Modern)

2. Is there a difference in the preference ratings for modern art vs. non-human generated art? Firstly, the columns with modern art preference ratings (36-70) and the columns with non-human generated art preference ratings (71-91) were divided into two distinct arrays - again, there was no need for NaN value removal, as none of the 55 columns included NaNs. Then, the normality of the two distributions was tested using the Shapiro-Wilk test, which again concluded non-normality of the data due to the elicited p-values of ~0.0. Non-normality along with the ordinal nature of the data pointed to the necessity of a non-parametric test. Moreover, since the samples were of different sizes, a paired test was not appropriate nor feasible - this need for a non-parametric independent test led me to the Wilcoxon Rank Sum Test (AKA Mann-Whitney U-Test). The p-value produced from the test was 1.6275105785977077e-256. I performed a

two-sided test in order to fit the structure of the question, which only wanted to know about the difference between the two groups, not which group was higher. Since the elicited p-value was much lower than the significance level of 0.05/2 (as per a two-sided test), the null hypothesis was rejected, indicating that there was in fact evidence of a significant difference in the median preference ratings for modern art and non-human art. To further support this conclusion, comparable distributions of the two variables were produced, which clearly indicated that modern art preference ratings were higher in the larger categories (3, 4, 5, 6, and 7), while non-human art preference ratings were higher in the smaller categories (1 and 2). Therefore, the results from the Wilcoxon Rank-Sum Test and the two distributions of the modern and non-human art preference ratings do in fact suggest that there is a difference in the preference ratings for modern art vs. non-human art.

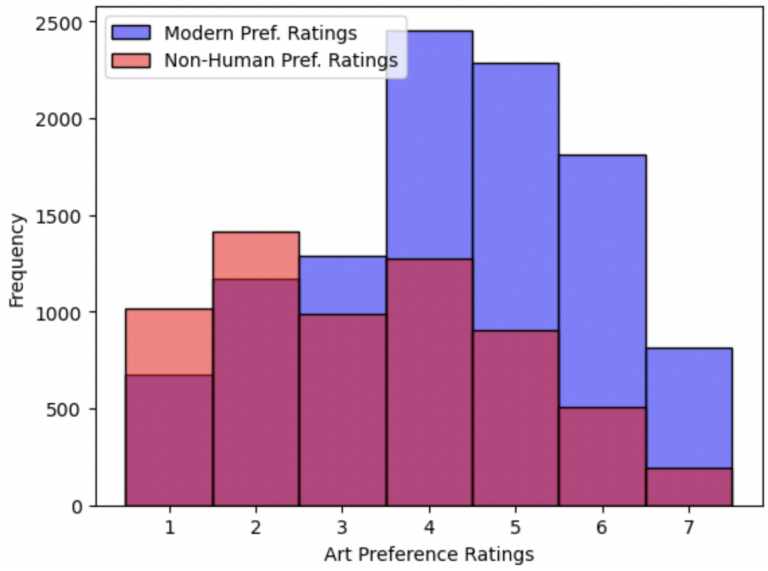


Fig 2: Histogram Plot of the Ratings Distributions (Modern vs. Non-Human)

3. Do women give higher art preference ratings than men?

Firstly, I dropped all columns in the data set other than preference ratings and gender, as they were irrelevant to the question. Then, I performed row-wise removal of the NaN values within the data set - as there were no NaN values within the preference ratings, the rows (users) were dropped that had NaN values in the gender column. Then, I separated the original data set into two new arrays based solely on gender - one for women's preference ratings and one for men's preference ratings. As the question pertains to the male and female gender only, the non-binary gender was excluded for this analysis. It is important to note that there were 95 males who rated the art pieces, and 179 females who rated the art preferences. Then, the normality of the distributions for male and female ratings was tested using the Shapiro-Wilk test, which concluded non-normality of the data due to the elicited p-values of ~0.0. Since the data was both non-normal and ordinal, a non-parametric test ws needed. Moreover, as the question was interested in the difference between two distinct groups, more specifically if women's ratings were higher, I used the one-sided Wilcoxon Rank-Sum Test, as it is a non-parametric independent test. This test produced the p-value of 0.13942219869082179, which was not lower than the significance level of 0.05. Consequently, the null hypothesis failed to be rejected, suggesting that there was no difference between the medians of the two groups, female and male preferences. Therefore, we can conclude that women do not give higher art preference ratings than men.

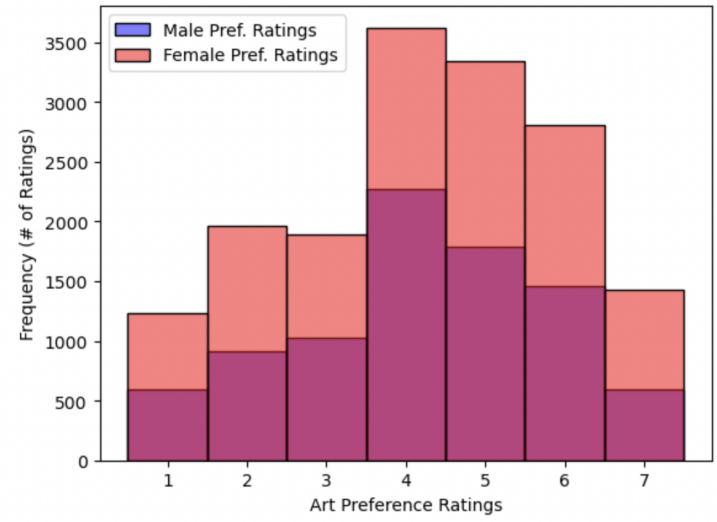


Fig 3: Histogram Plot of the Ratings Distributions (Male vs. Female)

4. Is there a difference in the preference ratings of users with some art background vs. none? Firstly, I dropped all columns in the data set other than preference ratings and art education as they were irrelevant to the question. Then, I performed row-wise removal of the NaN values within the data set - as there were no NaN values within the preference ratings, the rows (users) were dropped that had NaN values in the art education column. Then, I separated the data set into two new arrays based solely on art education - one for preference ratings of users with no art education (0 years), and one for preference ratings of users with some art education (1+ years). It is important to note that there were 188 users with some art education that rated the art pieces, and 92 users with no art education that rated the art preferences. Then, the normality of the distributions for some art ed. and no art ed. ratings was tested using the Shapiro-Wilk test, which concluded non-normality of the data due to the elicited p-values of ~0.0. Since the data was both non-normal and ordinal, a non-parametric test ws needed. Moreover, as the question was interested in the difference between two distinct groups, more specifically users with some art ed. vs. no art ed., I used the two-sided Wilcoxon Rank-Sum Test, as it is a non-parametric

independent test. This test produced the p-value of 1.708219981286751e-08, which was lower than the significance level of 0.05/2 (as per the two-sided test). Consequently, the null hypothesis was rejected, suggesting that there was a significant difference between the medians of the two groups, users with some art education and users with no art education. Therefore, we can conclude that there is a difference in the preference ratings of users with some art background versus users with no art background, in respect to the scale of this data. I would like to mention that I was slightly skeptical of this result, as the two group distributions in *Fig. 4* appear fairly similar, with the biggest differences being that the sample size for users with no art education was lower. The p-value elicited from the test was understandably larger than the p-values for the other tests I performed that suggested significance, which makes sense because the distributions between the two groups were much more similar. Regardless, based on the test I performed, there is a difference between the preference ratings of the two groups.

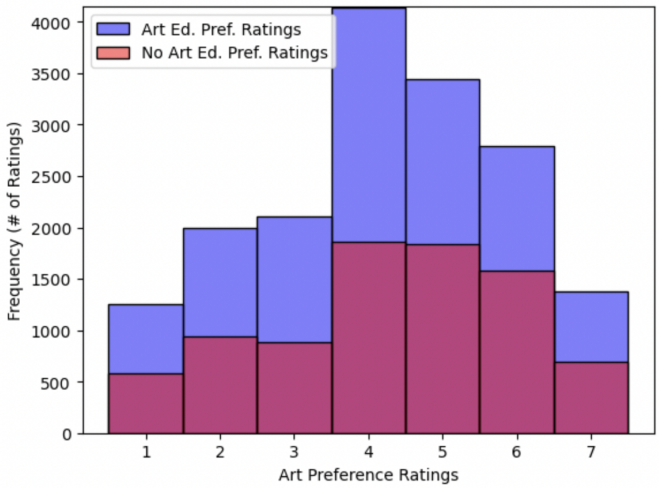


Fig 4: Histogram Plot of the Ratings Distributions (Art Ed. vs. No Art Ed.)

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

In order to approach this question, I first examined the nature of the data and quickly realized that aggregation was necessary in order to run regression. The data was ordinal in nature, therefore using means was immediately ruled out - the mean is a normalized sum, which presumes that the ratings being considered are equal, therefore it wouldn't be fit to accurately characterize the central tendency of the ratings data. Medians would be a more apt characterization for ordinal data, as it would represent the middle value of the data. Moreover, I interpreted the question as asking to predict the art preference ratings based on each user rather than each art piece, as that would only make sense (especially in the context of Question 7). Therefore, for both preference and energy ratings, I reduced each row to one median, a median which represented the middle value of the user's ratings for all 91 art pieces. It is important to note that aggregation wasn't absolutely necessary for this question, as it would be reasonable to run regression with each individual energy rating as the predictor, corresponding to each individual preference rating as the outcome; however, I decided to aggregate for the sake of comparing this model with Question 7's model, which necessitated aggregation. Furthermore, due to the data being ordinal, linear regression would not be optimal due to the fact that the dependent variable is not continuous. However, running random forest regression elicited a negative r-squared value, and running ordinal logistic regression elicited a model score of -0.18, meaning they were worse at predicting the outcomes than the mean itself. Because of the apparent uselessness of these models, I turned to the linear regression model, while also taking into consideration that it may not be very fit for the data.

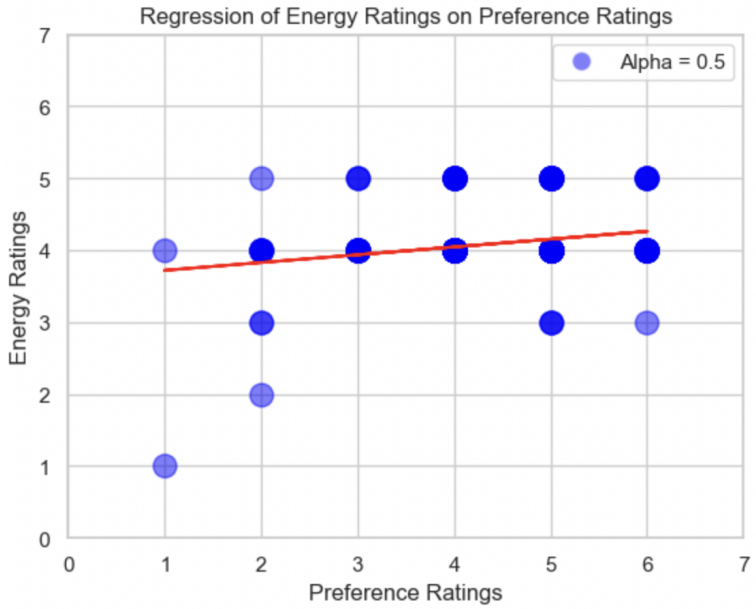


Fig 5: OLS Regression of Energy Ratings on Preference Ratings

In order to consider overplotting, I set the transparency of the points to 0.5, so overlapping points would be much darker in color. The regression model generated an r-squared value 0.05203 (0.97 without the constant), which means that 5% of the variance could be accounted for by the model. The RMSE was 0.41, which could be generally interpreted as good, as the predictions weren't extremely far from the actual values. In order to cross-validate this model, I performed k-fold cross-validation with splits/folds of 1 through 5, in order to assess the RMSE of the model when the data has a higher amount of splits. The RMSE for fold 1 was 0.4821, and it generally decreased as folds increased, with the RMSE for fold 5 being 0.3579. The mean RMSE for the 5 folds was 0.4092. From these statistics we can conclude that although the RMSE was generally good, the regression is not a good predictor of energy ratings, in that it accounts for an extremely small amount of the variance.

6. Regression model to predict art preference ratings from energy ratings and demographic information

In order to carry out multiple regression I first had to select which variables to use as the predictors. I opted to use all six demographic variables - age, gender, political orientation, art education, general education, and artist - along with energy ratings in order to see the impact that each of them had on preference ratings. It was necessary to aggregate the preference ratings and energy ratings to medians, as justified in Question 5, in order to reduce them to one variable/column per user, so they could be compared with the demographic variables. I then performed a row-wise removal of the NaNs so that each row used in the regression had a value - 21 rows within the demographic columns included NaNs, so 279 rows were left. In order to make sure that there weren't scaling issues among the variables, I standardized them. The multiple linear regression produced an RMSE of 0.8475 and an r-squared of 0.07641, indicating that it was an even poorer model than the single linear regression model, which had a much lower RMSE. Since there were multiple coefficients within the data I opted to use LASSO Regression in order to completely eliminate certain variables, in hopes to improve the model.

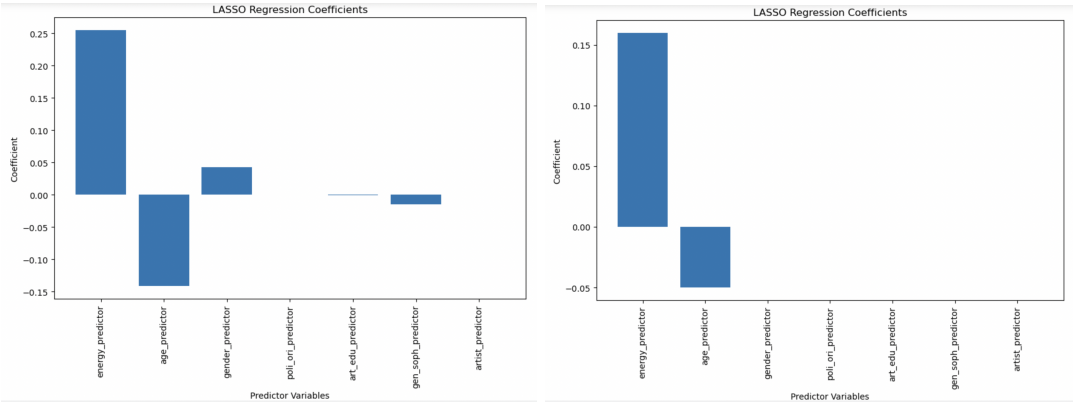


Fig. 6: LASSO Coefficients for an alpha of 0.01 and 0.1

As seen in *Fig. 6*, setting alpha to 0.01 eliminated the political orientation and artist predictors, and produced an RMSE of 0.7671. Setting the alpha to 0.1, however, eliminated all variables other than the energy predictor and the age predictor, and produced an RMSE of 0.7186. Although LASSO regression helped to improve the model by lowering the RMSE, especially with the increase in alpha, the RMSE of the single linear regression model was still lower, and therefore better. From this, we can conclude that energy ratings are a much better predictor of preference ratings than the demographic variables are, as the more we decreased the betas of the demographic variables, the better our model became.

7. K-means clustering for 2D space of average preference ratings vs. average energy rating

In order to prepare the data for k-means clustering, I created a new array of 182 columns with solely the art preference and energy rating columns. I then averaged the user preference ratings of the first 91 columns, producing a singular column of art preference rating means for each of the 91 art pieces. Then, I averaged the user energy ratings of the second 91 columns, producing a second distinct column of art energy rating means for each of the 91 art pieces. It is important to note that NaN value removal wasn't necessary, as there were no NaN values in the preference ratings. Then, in order to determine which k to use for k-means clustering, I used the silhouette method to produce silhouette scores for k values 1-30. From this method, the best k-value was 19, with a silhouette score of 0.438 - as silhouette scores range from -1 to 1, this score indicates that the data points are reasonably close to their respective clusters and far from others. However, in the context of the question, a k-value of 19 is meaningless. We are trying to classify, or identify the type of art that the clusters correspond to - the art is categorized by source, so it would only make sense to try to identify if the clusters correspond to classical, modern, and

non-human generated art. Yes, the art pieces also are categorized by style, but there are so many styles included among the pieces, with some pieces even having their own distinct style, it wouldn't make sense to try to cluster them based on style - we would need more data for that. Therefore, I decided on a k-value of 3 for the sake of logically interpreting the clusters, which still has a reasonable silhouette score of 0.38848546174510756.

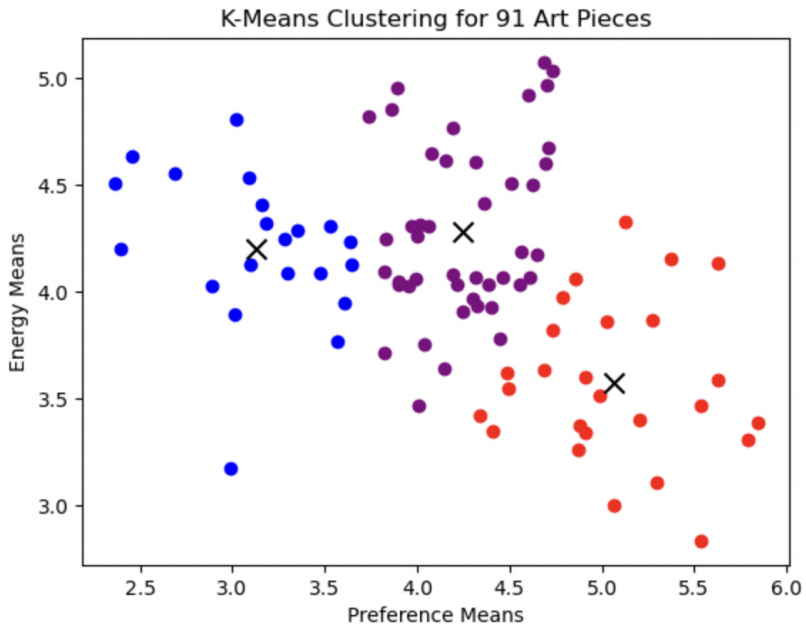


Fig. 7: K-means Clustering of Preference vs. Energy Means for 91 Art Pieces

After observing the pattern of the preference means and energy means of each art piece, I noticed that they do somewhat correspond to the resulting k-means clusters. Of course, the classification isn't spot on by any means, but it seems as though classical art pieces could be identified in cluster 3 (red cluster), modern art pieces in cluster 2 (purple cluster), and non-human generated art pieces in cluster 1 (blue cluster).

8. Using 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?

My first step towards PCA was creating a new data set from the original with only the 10 columns of self-image statement ratings. Then, I dropped all of the rows with NaN values so we could compare each statement equally. In order to visualize the user ratings for each statement, I produced a visual image of the 2D data. At a glance, it is clear that statements 2, 6, and 8 have lower user ratings, while statements 3 and 7 have higher user ratings.

Fig. 8: User Ratings for 10 Self-Image Statements

In order to get a better idea of how the statements correlated, I created a correlation matrix. in which there doesn't seem to be any exceptional correlation other than statements 1 and 10 being highly correlated and statements 8 and 4 being weakly correlated.

Fig. 9: Correlation Matrix for 10 Self-Image Statements

An essential step in the process of PCA is z-scoring the data, so each column has a mean of zero. After z-scoring the 10 columns, I initialized the PCA object and fit the data accordingly. I extracted the loadings in the form of a matrix, then rotated the data in order to represent the users as columns and variables as rows. I then created a scree plot to display the Eigenvalues of each of the 10 principal components. In order to manually examine the variance explained for each principal component, I divided the respective eigenvalue by the sum of eigenvalues, multiplied by 100.

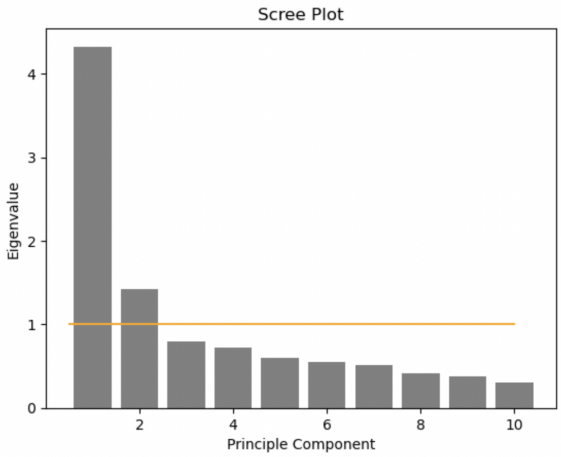


Fig. 10: Scree Plot for Self-Image Principal Components

It is clear that the first principal component accounts for a large amount of variance explained in the data - 43.137% to be exact. The second principal accounts for only 14.183%, illustrating the considerable drop off after the first component. In order to further analyze the first principal component, I examined its loadings, or the weights per statement (variable) in terms of the original data. I reached into the loadings matrix and produced a plot to identify which statements had the most impact on the first principal component.

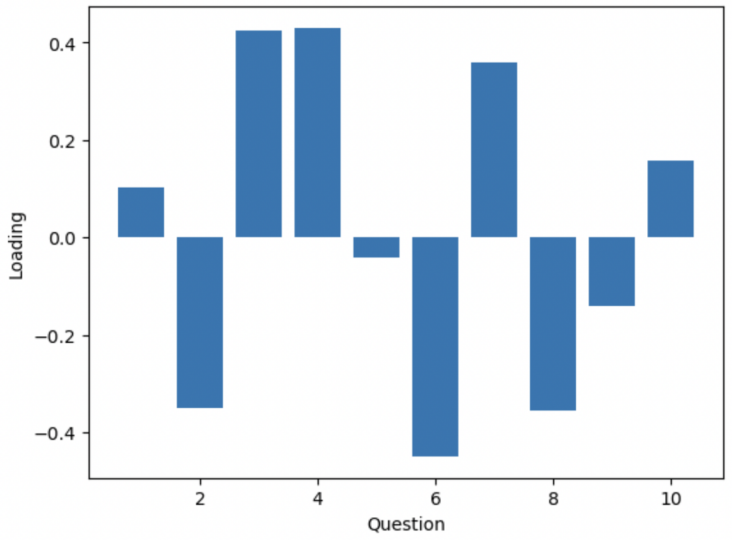


Fig. 11: Loadings for the First Principal Component

It is clear from *Fig. 11* that statements 3, 4, and 7 were responsible for the first principal component, as they had the highest loading. Statements 3, 4, and 7 were "I feel that I have a number of good qualities," "I am able to do things as well as most other people," and "I feel that I'm a person of worth, at least on an equal plane with others," respectively. An apt identity of these variables, which could describe the principal component, might be competency, or dignity. The ability of the first principal component to predict preference ratings was then tested with linear regression - although the data is ordinal, PCA assumes linearity as well so it would be only fit to continue with linear regression. In order to prepare the data for regression, I returned to the original data set and removed the entire rows for self-image columns with NaN values, so the art preference data would correspond accordingly. As justified earlier, for the regression, medians (of all 91 pieces, per user) were used in order to display the central tendency of the ordinal art preference ratings.

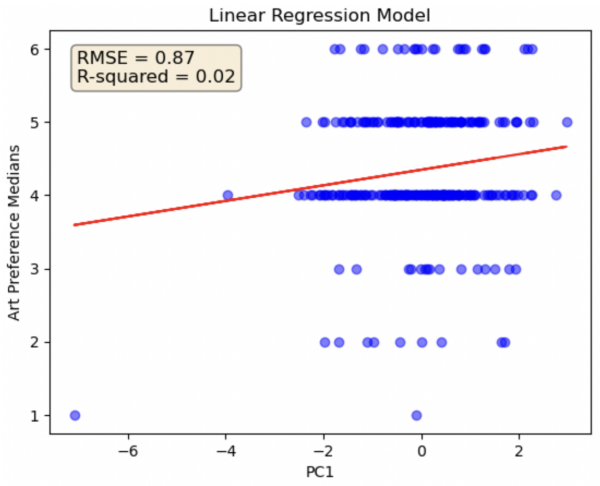


Fig. 12: Simple Linear Regression for PC1 on Art Preference Medians

From the regression, an RMSE of 0.87 was produced, which is the standard deviation of the residuals - since 0.87 is a relatively high RMSE, the predictions are relatively far from the actual values. Furthermore, the r-squared, or coefficient of determination, was only 0.02, which indicated that only 2% of the variance could be accounted for by the model. Based on these two statistics, the first principal component isn't necessarily a good predictor of art preference ratings (medians), because the variance accounted for is extremely small while the predictions themselves were poor as well.

9. Using 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.).

My first step towards PCA was creating a new data set from the original with only the 12 columns of dark personality trait statement ratings. Then, I dropped all of the rows with NaN values in order to remove missing data and compare each statement equally. In order to visualize

the user ratings for each statement, I produced a visual image of the 2D data. At a glance, it is clear that statements 4, 5, and 6 have lower user ratings, while statements 9 and 10 have higher user ratings.

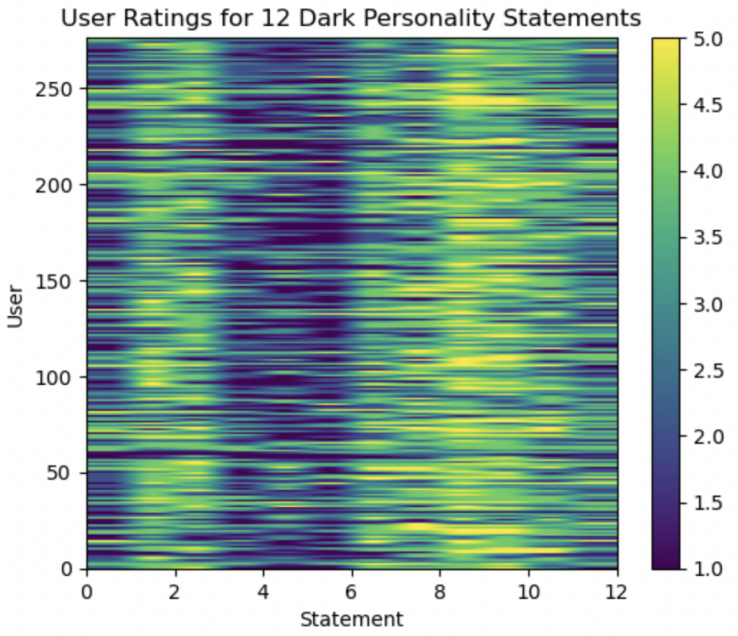


Fig. 13: User Ratings for 12 Dark Personality Statements

I then created a correlation matrix in order to observe the relationship between each statement. Statements 1 and 4 were highly correlated, while statements 5 and 9 were weakly correlated.

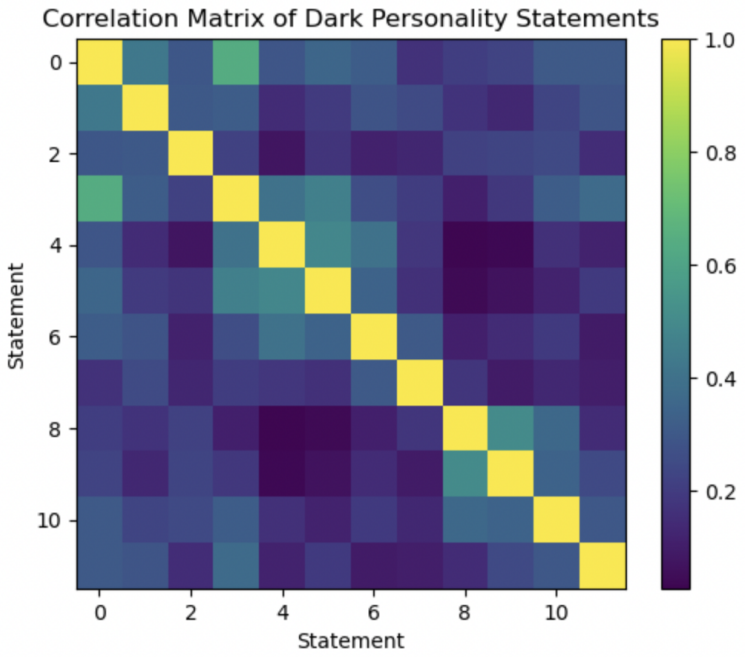


Fig. 14: Correlation Matrix for 12 Dark Personality Statements

An essential step in the process of PCA is z-scoring the data, so each column has a mean of zero. After z-scoring the 12 columns, I initialized the PCA object and fit the data accordingly. I extracted the loadings in the form of a matrix, then rotated the data in order to represent the users as columns and variables as rows. I then created a scree plot to display the Eigenvalues of each of the 12 principal components. In order to manually examine the variance explained for each principal component, I divided the respective eigenvalue by the sum of eigenvalues, multiplied by 100.

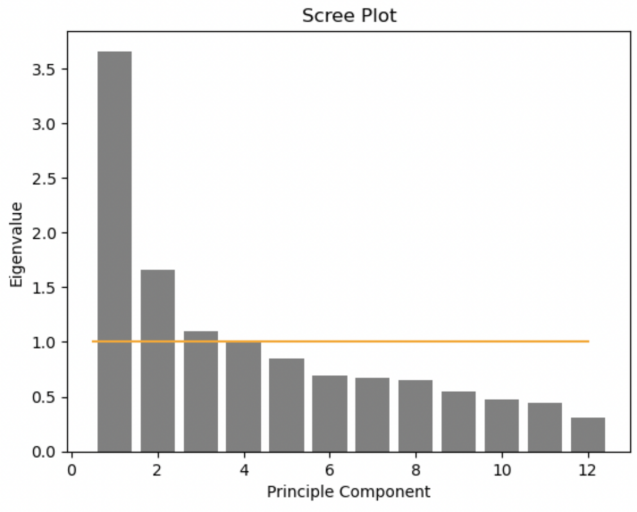
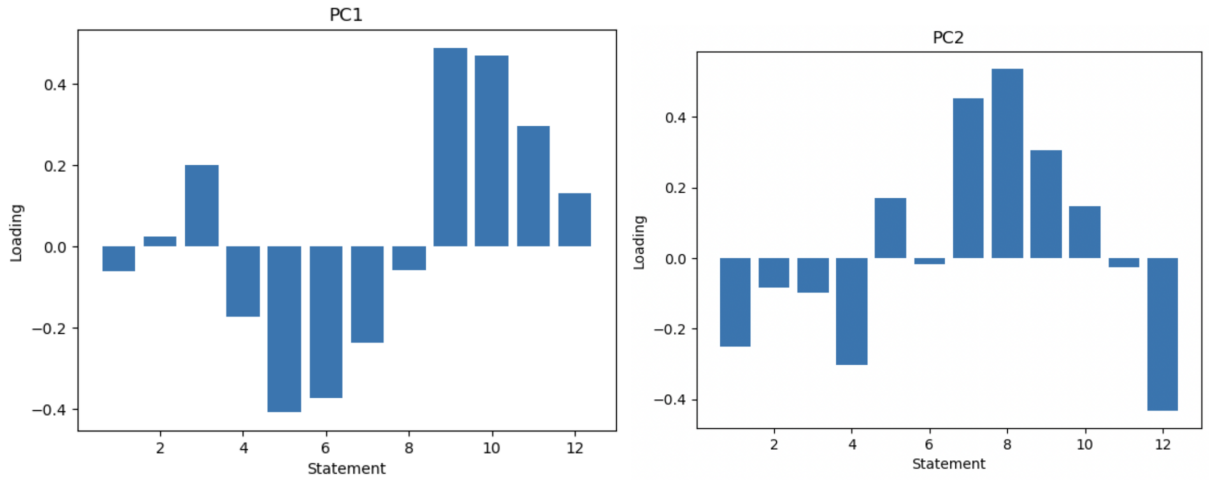


Fig. 15: Scree Plot for Dark Personality Principal Components

It is clear that the first principal component accounts for a large amount of variance explained in the data - 30.365% to be exact. The second principal accounts for only 13.738%, illustrating the considerable drop off after the first component, and the third component accounts for 9.127%. In order to further analyze the first principal components, I examined their loadings, or the weights per statement (variable) in terms of the original data. I reached into the loadings matrix and produced three plots to identify which statements had the most impact on the first, second, and third principal components.



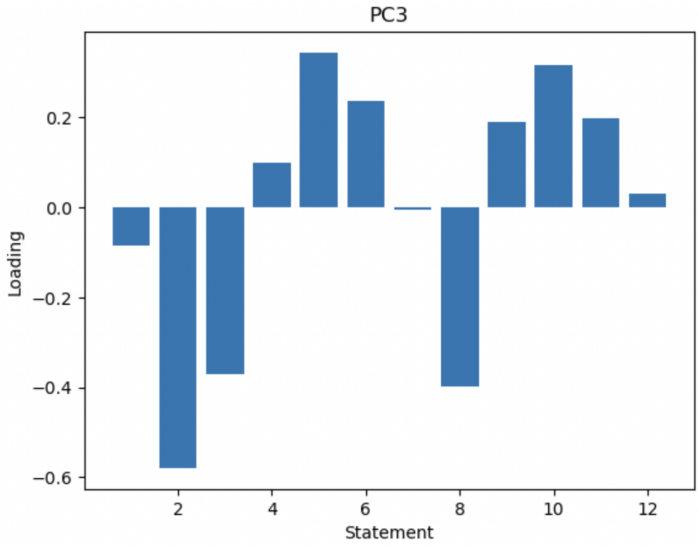


Fig. 16: Loadings for the First Three Principal Components

It is clear from *Fig. 16* that statements 9, 10, and 11 were responsible for the first principal component, as they had the highest loading. Statements 9, 10, and 11 were "I tend to want others to admire me," "I tend to want others to pay attention to me," and "I tend to seek prestige and status," respectively. An apt identity of these variables, which could describe the principal component, might be histrionic, as histrionic individuals are characterized by their need for attention and approval. Statements 7, 8, and 9 were highest for the second PC, which were, "I can be callous or insensitive," "I tend to be cynical," and "I tend to want others to admire me," respectively. The dark personality trait that could characterize these variables might be nihilistic, although nihilistic individuals likely don't care about how others perceive them. So narcissism might be a more apt description of these three statements. Lastly, statements 5, 6, and 10 were highest for the third principal component, which were "I tend to lack remorse," "I tend to be unconcerned with the morality of my actions," and "I tend to want others to pay attention to me." Sociopathic seems to be the most fit description of these statements, as sociopaths lack regard for what is right and wrong, but still crave attention nonetheless.

The ability of the first three principal components to predict preference ratings was then tested with linear regression - although the data is ordinal, PCA assumes linearity as well so it would be only fit to continue with linear regression. In order to prepare the data for regression, I returned to the original data set and removed the entire rows for dark personality columns with NaN values, so the art preference data would correspond accordingly. As justified earlier, for the regression, medians (of all 91 pieces, per user) were used in order to display the central tendency of the ordinal art preference ratings. From the regression, an RMSE of 0.88 was produced, which is the standard deviation of the residuals - since 0.88 is a relatively high RMSE, the predictions are relatively far from the actual values. Furthermore, the r-squared, or coefficient of determination, was only 0.01, which indicated that only 1% of the variance could be accounted for by the model. Based on these two statistics, the first three principal components aren't necessarily a good predictor of art preference ratings (medians), because the variance accounted for is extremely small while the predictions themselves were very poor as well.

10. Logistic regression to classify political orientation of the users

In order to classify the political orientation of the users, using all of the other data provided, I opted to use logistic regression, as it is optimal for classification with binary outcomes. In preprocessing the data, I removed all of the rows that had any NaN values, since every single column would be taken into account in this model. In order to make the political orientation column binary, I turned all values of 1 and 2 into 0, in order to represent the left, and values of 3-6 into 1, to represent non left political orientation. I then made the predictor variables every single column other than political orientation, and set the outcome variable to the political orientation column.

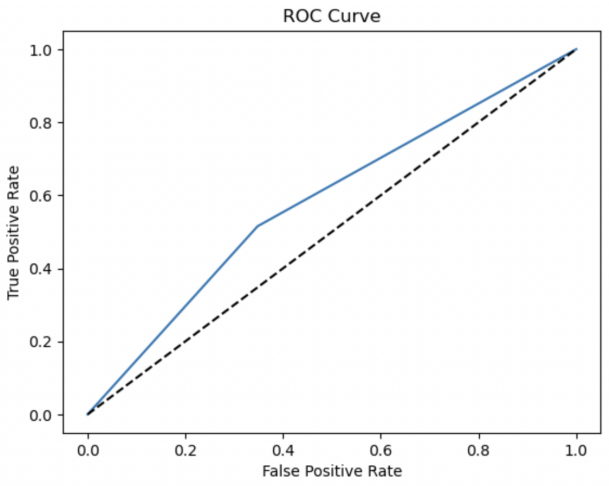


Fig. 17: ROC Curve for Logistic Regression

In order to assess how good the logistic classification was, I analyzed the ROC curve and produced the stat for the AUC, or area under the ROC curve. The AUC score was 0.5836 - which means that there is a relatively high probability of false positives, or incorrectly predicting (non left) political orientation. Although not great by any means, the predictions of this model are better than random guessing (or a random classifier), which would produce an AUC of 0.5. So, to answer the question, you cannot straight-up determine the political orientation (left or non left) of the users based on action preferences, art preference ratings, art energy ratings, self-image ratings, dark personality trait ratings, and demographic variables combined.