Jonathan Zhou

Professor Wallisch

Introduction to data science

May 11th 2023

Capstone Project Report

Preprocessing:

I. The two datasets are imported into pandas dataframes, with the column names renamed for clarification, and each group of columns is extracted for future reference (such as preference rating columns, dark personality columns, etc). Nulls are not eliminated during preprocessing, but done before each question specifically over the dataset used for that question, to prevent unnecessary elimination of valuable data.

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

I. Methodology: two approaches are taken to answer this question, one where we do not reduce our data to user means, and one where we do. Both approaches will utilize a paired t-test as the same users are rating both classical art and modern art, and a Mann-Whitney rank-sum test to account for the possibility that the mean is not a good representation of our distribution due to unequal variance. Because we are interested in whether or not classical art is "more" well liked than modern art, we specifically use the one-sided Wilcoxon rank-sum test, to test our alternative hypothesis that classical art is more well liked than modern art (the two-sided p-value will also be considered).

II. Results: first, we plot the two sample groups and examine their distribution [Figure 1]; while they both seem to peak around a rating of 4, the distribution of classical art is noticeably more normal and centered.

A. Conducting our statistical tests on our method 1 data, our resulting p-values are 5.10×10^{-112} and 1.59×10^{-97}, for paired t-test and Wilcoxon rank-sum test, respectively. Both of the results are significant, as they are less than our alpha level of 0.05 (note that the two-sided Wilcoxon rank-sum test is also significant here). We can therefore reject our null hypothesis in favor of our alternative hypothesis that classical art is more well liked than modern art.

B. Repeating the same statistical tests as before, but using our method 2 data, which consists of two columns (classical and modern) of 300 rows of the average rating of the art in that category by each user. The resulting p-values are 4.95×10^{-23} and 4.018×10^{-15}, for paired t-test and one-tailed Wilcoxon rank-sum test, respectively. Once again, our tests yield significant results, indicating that users are, on average, likely to rate classical art higher than modern art.

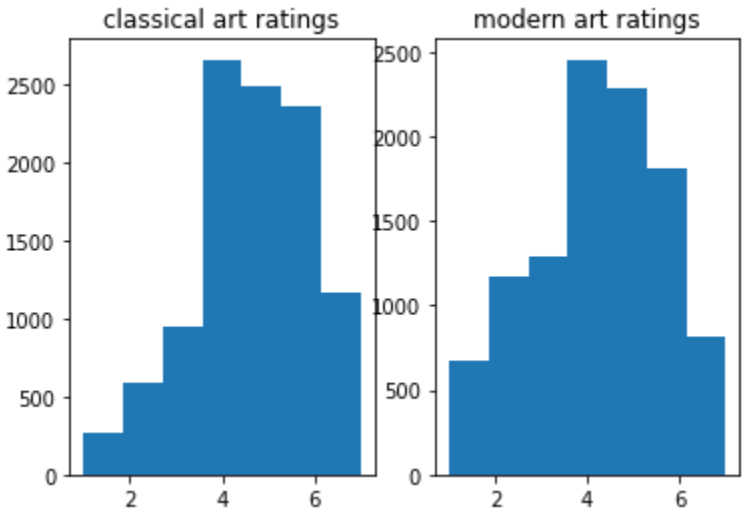


Figure 1

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

I. Methodology: similar to question 1, we repeat our two approaches to answer this question, one where we do not reduce our data to user means, and one where we do. Both approaches will utilize a paired t-test as the same users are rating both classical art and modern art, and a Mann-Whitney rank-sum test to account for the possibility that the mean is not a good representation of our distribution due to unequal variance. This time, we utilize a two-sided Mann-Whitney rank sum test as our new alternative hypothesis is that there is a difference in the preference ratings for modern art and nonhuman art.

II. Results: first, we plot the two sample groups and examine their distribution; the difference in their distribution is significant as the nonhuman art ratings are not evidently normal and highly skewed to the left [Figure 2].

   A. Conducting our statistical tests on our method 1 data, our resulting p-values are 5.38×10^{-270} and 8.74×10^{-264}, for paired t-test and Wilcoxon rank-sum test, respectively. Both of the results are significant, as they are less than our alpha level of 0.05 (note that the two-sided Wilcoxon rank-sum test is also significant here). We can therefore reject our null hypothesis in favor of our alternative hypothesis that there is a difference between preference ratings of modern art and nonhuman art.

   C. Repeating the same statistical tests as before, but using our method 2 data, which consists of two columns (modern and nonhuman) of 300 rows of the average rating of the art in that category by each user. The resulting p-values are 1.16×10^{-58} and 2.60×10^{-34}, for paired t-test and one-tailed Wilcoxon rank-sum test, respectively. Once again, our tests yield significant results, indicating that users are on average likely to rate modern art and nonhuman art differently.

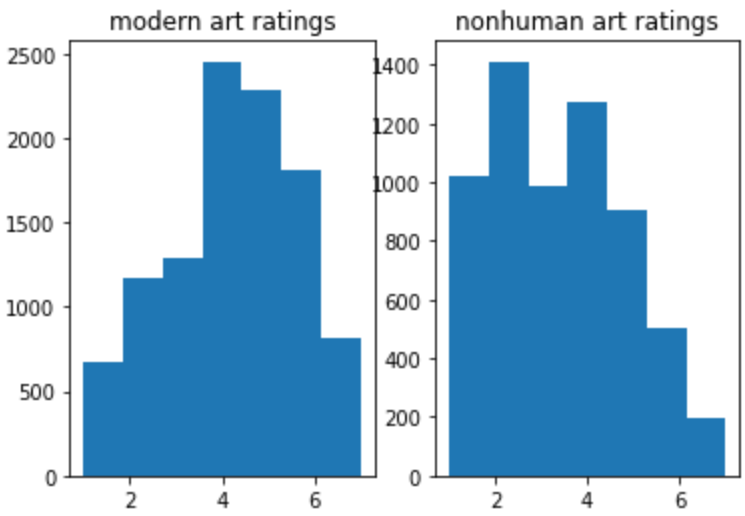


Figure 2

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

I. Methodology: to answer this question, we will first examine the two independent samples: ratings by women and ratings by men, to see if there is a difference in their distribution, and then compare their ratings across each source of art (classical, modern, nonhuman) to see if their preferences differ specifically in a given category. An independent t-test and one-sided rank sum test is conducted for each pair of samples.

II. First we examine the over distribution of preference ratings by women and by men, and recognize that while there are more ratings by women in our dataset, their ratings are less centered and have more spread, while the ratings by men are more centered and peak significantly in the middle [Figure 3].

   A. Women vs men overall: the independent t-test and one-sided rank sum test yielded p-values of 0.607 and 0.136, respectively. Comparing both to our alpha level of 0.05, we find that neither of the tests yielded significant results, and thus we fail to reject our null hypothesis that women's overall art preference ratings are the same as men's overall art preference ratings.

B. Women classical vs men classical: the independent t-test and rank sum test yielded p-values of 0.19 and 0.0325, respectively. Comparing both to our alpha level of 0.05, we find that while the one-sided rank sum test yielded significant results, our two-sided p-value would've been 0.067, which is not significant enough for us to justify the difference of the two groups, and hence we conclude that neither of the tests yielded significant results, thus we fail to reject our null hypothesis that women's classical art preference ratings are the same as men's classical art preference ratings.

C. Women modern vs men modern: the independent t-test and rank sum test yielded p-values of 0.016 and 0.0045, respectively. Comparing both to our alpha level of 0.05, we find that both of the tests yielded significant results, and thus we reject the null hypothesis and conclude that there is significant evidence that women's ratings of modern art is different from men's ratings of modern art.

D. Women modern vs men modern: the independent t-test and rank sum test yielded p-values of 0.00052 and ≅1.0, respectively. Comparing both to our alpha level of 0.05, we find that while the t-test indicates with strong evidence that our sample means are different, the rank sum test yielded an abnormally high p-value of almost 1.0. We fail to reject the null hypothesis that there is significant evidence that women's ratings of nonhuman art is higher than men's rating of nonhuman art. Upon reversing the direction of our hypothesis, we see that the rank-sum test yielded a p-value of 5.19×10^{-5}, which suggests that there is evidence for the opposite scenario, that women's rating of nonhuman art is lower than that of men.

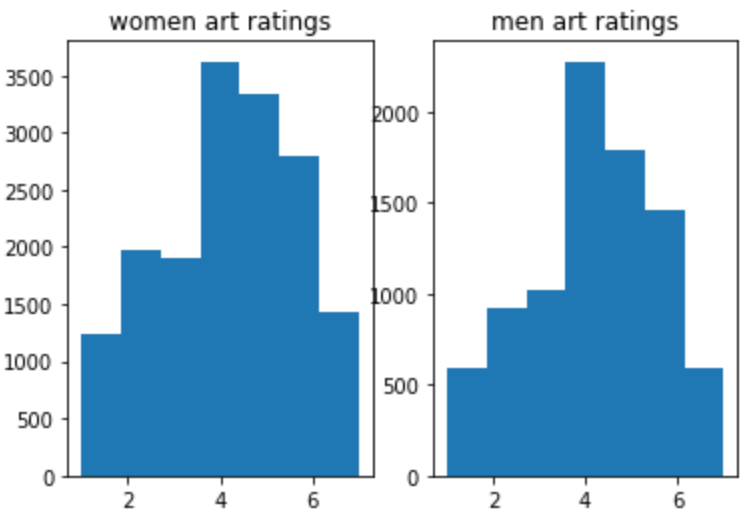


Figure 3

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

I. Methodology: to classify art education background, I chose to use column 219, "Art Education" as the characteristic to split up my samples. Following statistical tests will test the differences between the group with 0 art education v.s group with 1, 2, 3, respectively. Upon noticing the high variance in the average rating of each person, I have decided to not reduce each row to a sample statistic in order to preserve degrees of freedom in an attempt to prevent the sample mean from masking the underlying sample distribution. This concern is also addressed with additional Mann-Whitney tests. Our null hypothesis is that there is no difference in the ratings of users with no art education and some art background, while our alternative hypothesis is that there is a difference in the ratings of users with no art education and some art background.

II. Results: first, the graphical distribution of each sample group is examined [Figure 4], and it shall be noted that while we assumed the normality of our sample distributions, some samples are noticeably skewed. Each distribution noticeably peaks around the

4-5 rating range, which implies a possible commonly shared tendency to rate art pieces as average. Because we have a relatively large dataset, I am specifically opting to use an alpha value of 0.005 for this question. Upon conducting both an independent t-test and Mann-Whitney rank-sum test on 4 chosen sample combinations: group 0 vs 1, 0 vs 2, 0 vs 3, and finally 1 vs 2, our resulting p-values indicate the following:

A. 0 vs 1: while the difference in years of art education is one, this group yielded significant test statistics with p-value $\cong 0.0009$ for the independent t-test and p-value $\cong 0.00002$ for the Wilcoxon rank-sum test, both of which are below our alpha level of 0.005 and is therefore significant. We can therefore conclude that there is significant evidence that the distribution of ratings between users with no art education background and users with one year of art education background is significantly different.
B. 0 vs 2: this group yielded the most significant test statistics with p-value $\cong 1.1145023 \times 10^{-7}$ for the independent t-test and p-value $\cong 2.8659166124 \times 10^{-9}$ for the Wilcoxon rank-sum test, both of which are below our alpha level of 0.005 and is therefore significant. We can therefore conclude that there is significant evidence that the distribution of ratings between users with no art education background and users with two years of art education background is significantly different.
C. 0 vs 3: this group also yielded significant test statistics with p-value $\cong 0.00063$ for the independent t-test and p-value $\cong 0.0042$ for the Wilcoxon rank-sum test, both of which are below our alpha level of 0.005 and is therefore significant. We can therefore conclude that there is significant evidence that the distribution of ratings

between users with no art education background and users with three years of art education background is significantly different.

D. 1 vs 2: this group yielded test statistics with p-value ≅ 0.02 and p-value ≅ 0.04. Both p-values are higher than our alpha level of 0.005, and hence there is no significant evidence (note that it would've been significant at alpha = 0.05) that the distribution of ratings of users with one year of art education and users with two years of art education are different. This result is also hinted at by the relative similarity of the distributions depicted in the visualization.

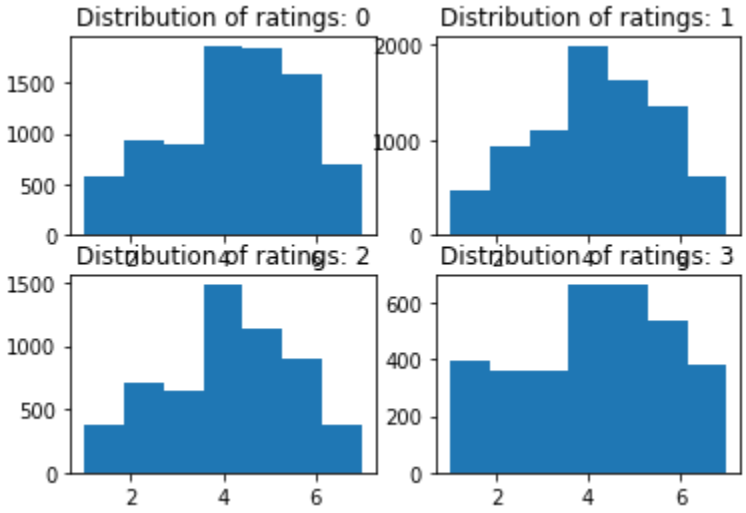


Figure 4

5) Build a regression model to predict art preference ratings from energy ratings only.

I. Methodology: the linear regression model will be approached in two different ways, one using individual energy rating to predict its corresponding preference rating, the other method using the average energy rating of each user to predict their average preference rating. Upon examining both the matrix of energy ratings and the matrix of preference ratings and confirming that there are no missing values, we can then flatten the matrix so that we arrive at two arrays that are individual ratings by one

person for one piece of art that can be used to build a model via method 1, using 300x91 energy ratings to predict its corresponding preference rating. For method 2, we condense both ratings into 300x1 arrays of average ratings for each person. No standardization of units is necessary as there is a single predictor and the units are the same.

II. Results:

A. Method 1: cross-validating the model with our test data set yielded a RMSE of ≅1.68, and our model itself has a R-squared value of ≅0.0002. Looking at a plot of our model, it seems that our model was unable to model the relationship between energy ratings and preference ratings, with an almost horizontal line that indicates little to no correlation [Figure 5.1], reflected by our ≅0 R-squared value. Because the data is originally rated on a 1-7 scale, our RMSE value indicates that we are expected to be around 1.68 units off with our prediction, which is very poor as that difference could be one of "love it" to "neutral". This demonstrates that there is no clear relationship between the preference rating of an art piece and its energy rating of an individual, which could be due to unpredictability in how people react to art and varying artistic preferences.

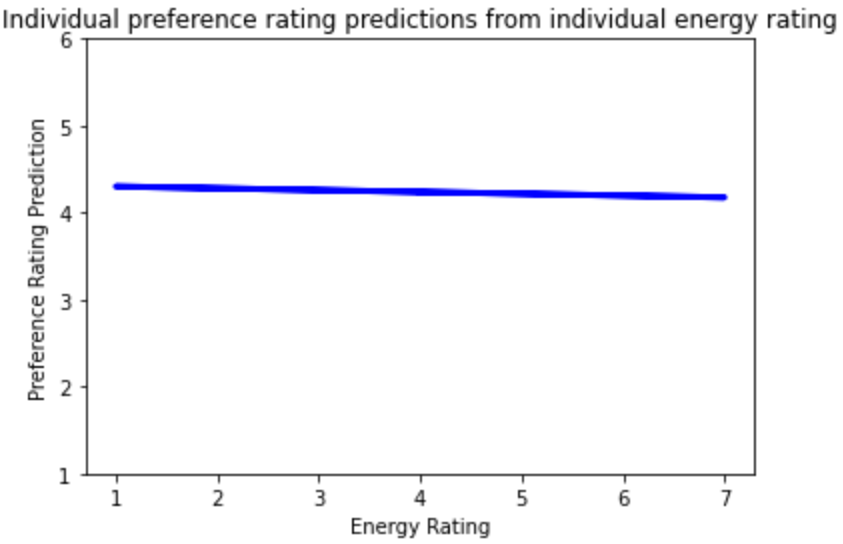


Figure 5.1

B. Method 2: cross-validating the model with our test data set yielded a RMSE of ≅0.51, and our model itself has a R-squared value of ≅0.10. This is a noticeable improvement compared to our method 1 results, which is largely attributed to the significant decrease in variation of our outcomes, which are now average preference ratings. The relationship becomes more significant and is noticeable with a steeper linear model [Figure 5.2].

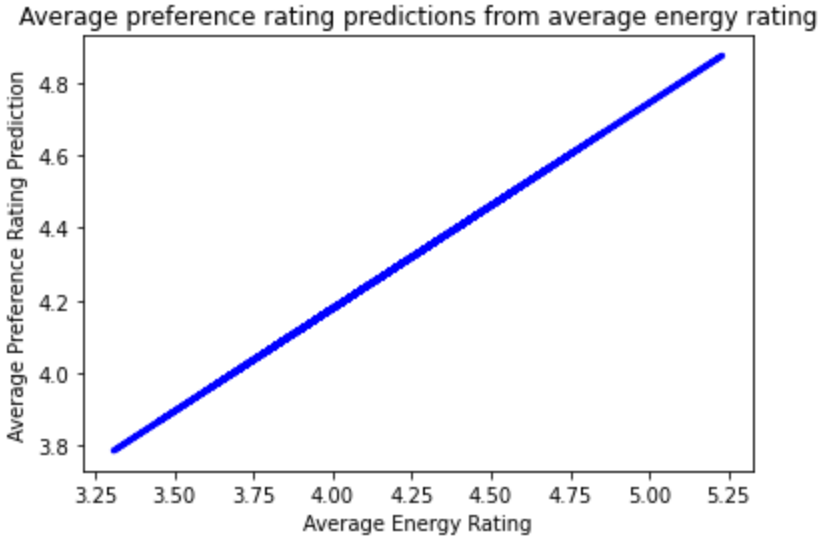


Figure 5.2

III. Regularization: In the concern of overfitting that could've led to the large RMSE between our test outcomes and predicted outcomes for method 1, we'll apply both the ridge and lasso regression. Unsurprisingly, neither of the regularization methods yielded significant improvements to our model, as the RMSE improves with increasing alpha levels, indicating that a horizontal line is potentially a better predictor. The improvements are also minimal, at 3.66×10^{-9} for alpha=10, indicating that overfitting is not the reason for the poor accuracy of our model, and that energy rating itself is not correlated enough with preference rating for it to be a useful predictor. Regularizing method 2 model with Ridge regression (alpha = 4)improved our RMSE by ≅ 0.0005, whereas using Lasso regression to penalize our R-squared to 0 yielded a 0.03 increase in RMSE, which indicates that our method 2 model is at least better than a horizontal line predictor.

6) Predict art preference ratings from energy ratings and demographic information

I. Methodology: building upon our previous model (method 2), two demographic information fields will be added as predictors: age and gender. As there are null values, all fields are combined and all rows with null values are removed to ensure that no rows have missing values (resulting in 279 rows). Because there exists multiple predators and there are different scales for the predictors, they are standardized (z-scored). The 279x3 matrix of predictors is then used to fit 279 outcomes.

II. Results: cross-validating our multi regression model yielded an RMSE of ≅ 0.54, with an R-squared value of 0.15 over the dataset. Comparing this result with our previous RMSE of ≅ 0.51, it's surprising that adding demographic information as additional

predictors did not make a noticeable improvement in the accuracy of our model, this may be due to overfitting, or simply a lack of correlation between demographic information and preference rating. Judging from our distribution of predicted vs actual average ratings, our model [Figure 6] was able to detect a correlation between our predictors and the outcome, but was unable to construct an effective regression model that can accurately predict each person's average preference ratings.

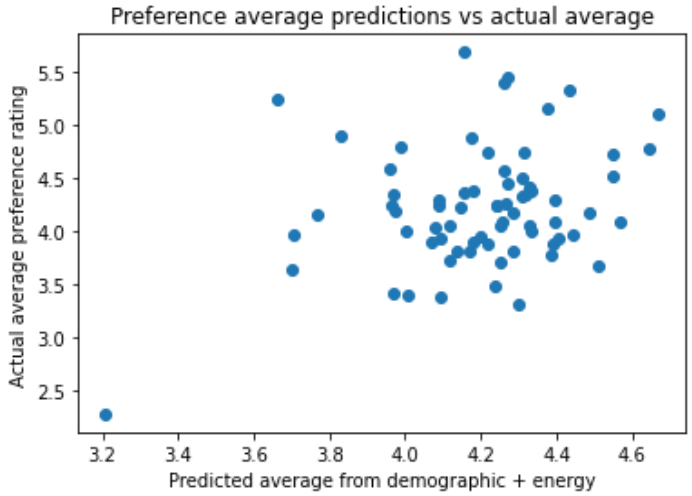


Figure 6

III. Regularization: Using the ridge regression to prevent overfitting of our predictors resulted in an improved RMSE of ≅ 0.538, which is minimized at alpha between 8 and 9. Applying penalty terms weighed by alpha = 8 resulted in a decreased R-squared value of 0.144, which suggests that our model was likely overfitting on our predictors. Utilizing the lasso regression, the same alpha yields an R-squared of 0.0, with a RMSE of ≅0.561. These results suggest that our predictors are poor predictors that, while they can help predict preference ratings more accurately, are ineffective. The alternative of simply guessing based on the average of average ratings is not significantly worse.

7) Cluster Classification

I. Methodology: we begin by calculating the average energy and preference rating of each art piece (averaging column-wise). We will then use the silhouette method to calculate the optimal number of clusters algorithmically, and then interpret the results to classify each cluster.

II. Results: the silhouette method suggests that the optimal number of clusters is 4 [Figure 7.1], so we apply K-means classification on our dataset and look at the 4 clusters visually [Figure 7.2] and interpret what each of the clusters might be classified as.

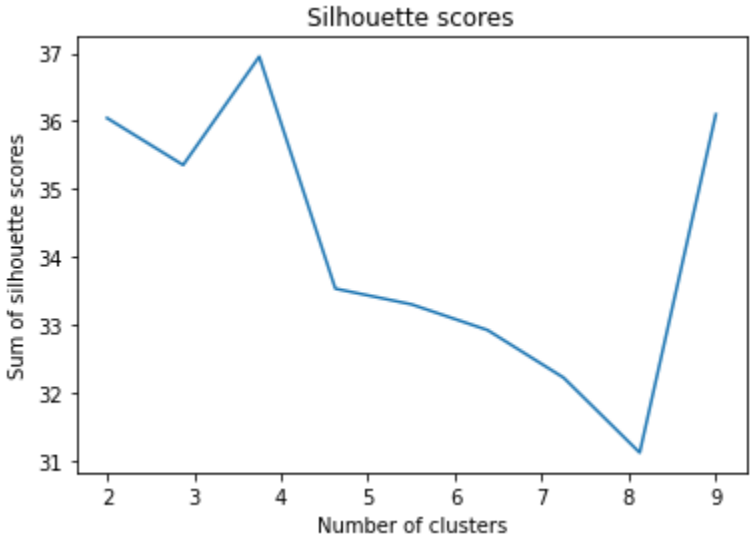


Figure 7.1

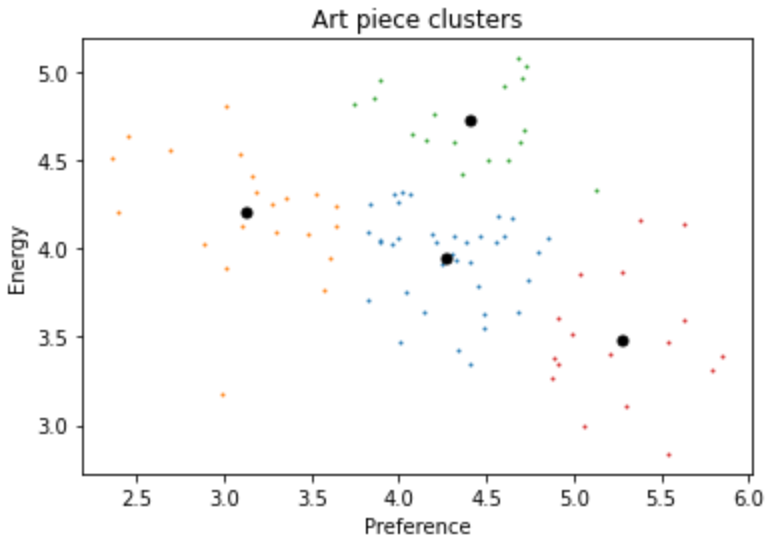


Figure 7.2

III. Interpretation: the four clusters will be addressed as follows: orange (leftmost) = cluster 1, blue (middle bottom) = cluster 2, green (middle top) = cluster 3, red (rightmost) = cluster 4. The classification of each of these clusters will be determined with the "Style" of the art works in their cluster and their "Type" of art.

A. Cluster 1: Characterized by the lowest average rating of all the clusters and a relatively high average energy rating, this group is surprisingly predominantly computer generated, "Abstract" art. Upon closer examination, most of these images are extremely abstract, convey little to no meaning (as they are not generated by humans), and elicit a wide range of emotional responses, with some bright images with high color contrast that are agitating. The images of this category can be described as **abstract nonhuman images**.

B. Cluster 2: this cluster is another very diverse cluster, with styles ranging from neoclassicism and romanticism to contemporary and pop art. This cluster contained many classics like Michelangelo's Adam, and some rather abstract pieces too. A possible common trait is that the pieces depict familiar concepts

relating to culture, identity, urban life, etc. This cluster can then reasonably be classified as **normal everyday art**.

C. Cluster 3: characterized by the highest average energy ratings and the second highest average preference ratings, this cluster contains diverse art styles: baroque, abstract, rococo, cubism, and many contemporary art styles, to name a few. Upon examining the art works themselves, we notice that the art pieces are relatively emotional, with the people depicted having dramatic expressions, and the abstracted pieces having very complicated symbolisms that challenge the viewer to be intellectually engaged. As such, this cluster can be described as **expressive, emotional art**.

D. Cluster 4: this cluster has the highest average rating, and contains a mixture of styles including renaissance, neoclassicism and realism. The pieces here mostly depict familiar ideas and images such as the mona lisa; as such, the high preference rating and medium energy rating of this group becomes reasonable and the pieces can be described as **popular classics**.

8) Considering only the first principal component of the self-image ratings as inputs to a regression model

I. Methodology: self-esteem questions and art preference ratings first need to be cleaned for null values together, then PCA is conducted on the self-esteem columns to reduce the columns to a 1-dimensional predictor, which is then used to predict average art ratings of each person through simple linear regression. Note that we are predicting the average preference rating of an individual as the self-esteem questions are related to individuals.

II. Our model yielded a negative R-squared of ≅ 6.867×10^{-5}, which is indication of the poor predictivity of our model. Compared to our model in question 5 method 2 and question 6, cross-validation on our new model yielded an even higher RMSE of ≅ 0.583. These results suggest that the reduced one-dimensional information of an individual's self-esteem is not a suitable predictor of their average preference rating.

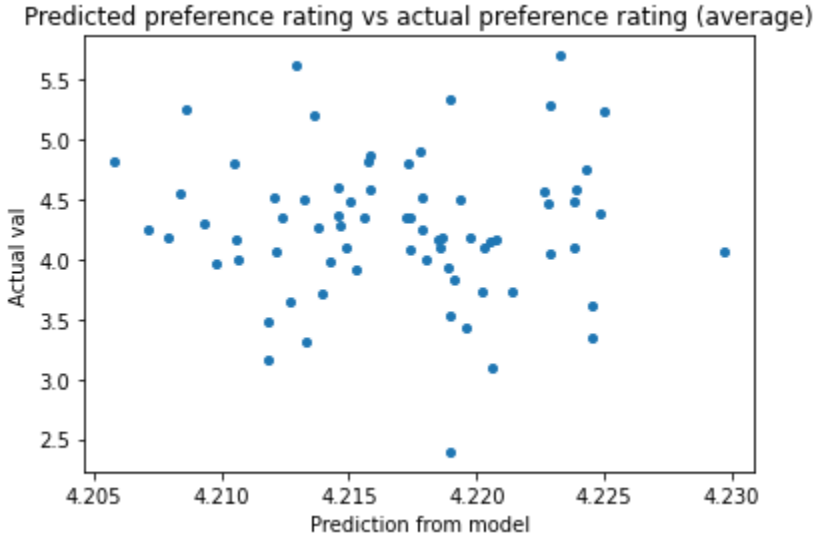


Figure 8.1

III. Regularization: applying both Ridge and Lasso regression, we notice that the improvement to the model is barely noticeable. As alpha increases to infinity, the RMSE continues to decrease as the R-squared approaches 0, with the Lasso regression having an RMSE of 0.5828 when R-squared becomes 0, which indicates that our model is failing to make accurate predictions, and a blind prediction model would perform more favorably [Figure 8.2].

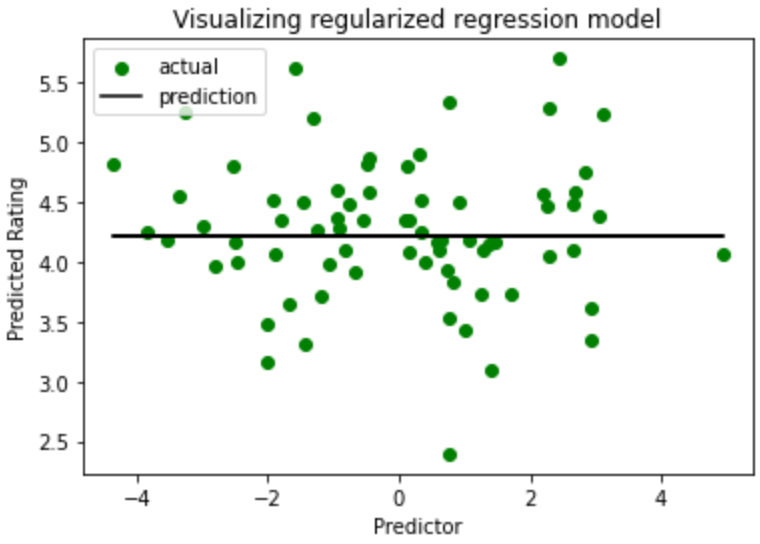


Figure 8.2

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

I. Methodology: PCA is conducted on "dark personality" columns to reduce it to 3 dimensions, and average preference ratings is calculated for each individual for the predictors to be paired with a single numerical outcome.

II. Results: The 3 PCA components have an explained variance ratio of 0.298, 0.139 and 0.092 each. The multi regression model yielded a cross-validation RMSE of ≅ 0.654 with an R-squared of ≅ 0.0426. While the model doesn't seem to be an accurate predictor of average preference rating, utilizing statsmodels regression summary shows that one of our components significantly predict art ratings: principle component 2 [Figure 9.1], as it had a P-value of 0.001 which is less than our alpha level of 0.05. Looking at the loading graphs, principle component 1 [Figure 9.2] correlates most with the questions: "I tend to manipulate others to get my way" and "I tend to exploit others towards my own end", so its identity is likely **narcissism**. Principal component 2 [Figure 9.3] corresponds to questions: "I tend to want others to

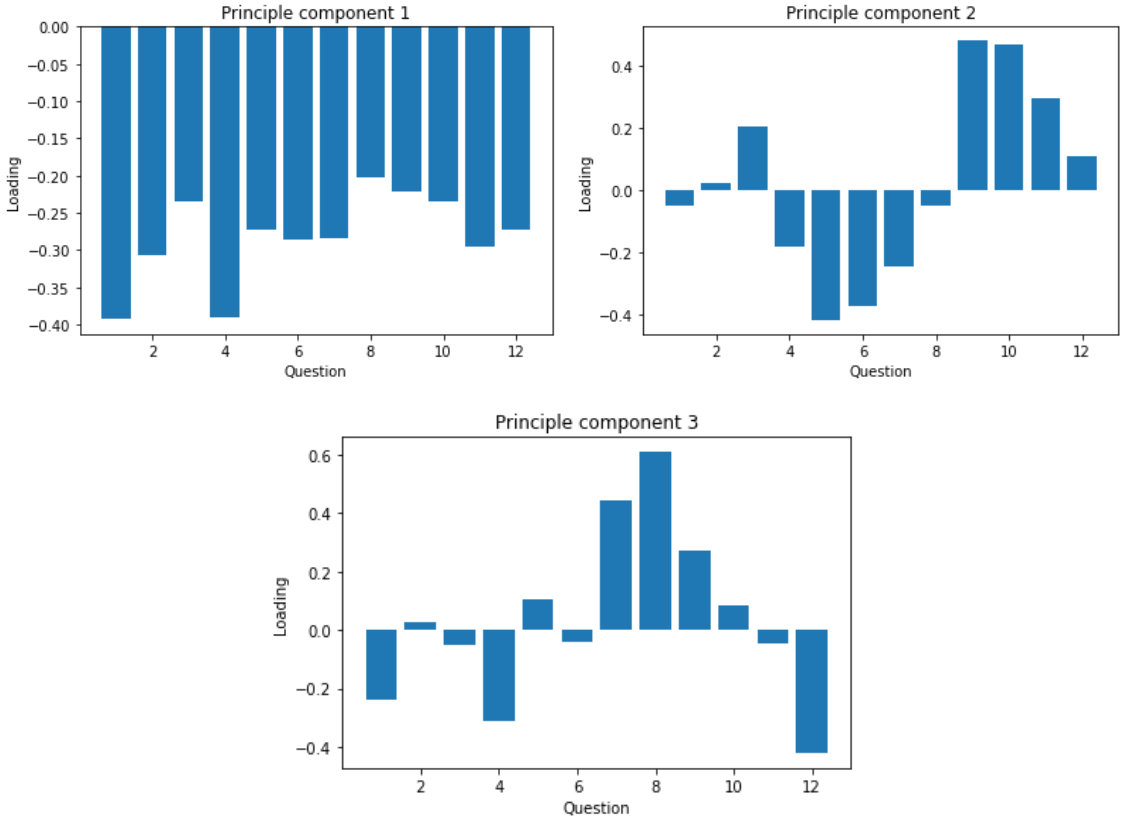
admire me" and "I tend to want others to pay attention to me" which indicates it can be identified as **attention seeking**. Finally, principle component 3 [Figure 9.4] corresponds to the questions "I can be callous or insensitive" and "I tend to be cynical", which indicates that its identity is likely **callousness.**

OLS Regression Results

Dep. Variable:	y	R-squared:	0.067
Model:	OLS	Adj. R-squared:	0.053
Method:	Least Squares	F-statistic:	4.978
Date:	Mon, 08 May 2023	Prob (F-statistic):	0.00234
Time:	16:48:28	Log-Likelihood:	-198.93
No. Observations:	213	AIC:	405.9
Df Residuals:	209	BIC:	419.3
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4.2359	0.043	99.294	0.000	4.152	4.320
x1	0.0175	0.023	0.777	0.438	-0.027	0.062
x2	-0.1145	0.035	-3.268	0.001	-0.184	-0.045
x3	-0.0749	0.040	-1.890	0.060	-0.153	0.003

Figure 9.1



Figures 9.2-9.4

III. Regularization: applying Ridge and Lasso regression to evaluate overfitting, we see a similar result to model 8, where the RMSE decreases as alpha increases, eventually minimized when R = 0 by Lasso regression. This indicates the poor predictive power of our model.

10) Can you determine the political orientation of the users vs. “non left” (everyone else)) from all the other information available, using any classification model of your choice?

I. Methodology: to create our model, 2 core values are considered: simplicity and accuracy. Upon creating multiple correlation matrices to test how closely each column is correlated with political orientation, 4 columns were chosen as predictors: artist status, sophistication, and two dark personality trait questions were selected: “I tend to be unconcerned with the morality of my actions” and “ I tend to seek prestige and status.” These columns all had a relatively high correlation coefficient with political orientation (≅ 0.1). These columns are standardized and used to predict categorized political orientation through multiple logistic regression. Note that despite these factors exhibiting some of the highest correlations with political orientation, their values range from 0.09 to 0.15, which is still incredibly low, but through combining these factors, we hope to gain better predictive power.

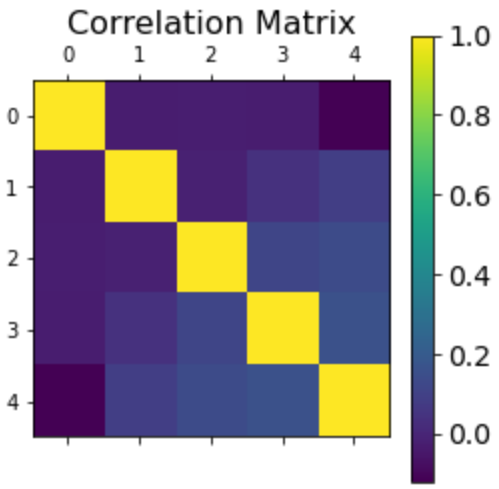


Figure 10.1

II. Results: the logistical regression yielded an accuracy of 0.7, with 2 of our predictors resulting in a significant p-value [Figure 10.2]: artist status with a p-value of 0.002 and personality question 11 (prestige seeking) with a p-value of 0.004. Cross validating our results yielded a minimum accuracy of ≅ 0.46, an average accuracy of ≅ 0.65, and a maximum accuracy of ≅ 0.79. While the overall accuracy of our model is not significantly better than making a blind prediction, which would yield a 0.5 accuracy, our model still evidently improves our odds of making a correct prediction.

Logit Regression Results

Dep. Variable:	y	No. Observations:	279
Model:	Logit	Df Residuals:	275
Method:	MLE	Df Model:	3
Date:	Mon, 08 May 2023	Pseudo R-squ.:	0.05228
Time:	17:23:36	Log-Likelihood:	-180.41
converged:	True	LL-Null:	-190.36
Covariance Type:	nonrobust	LLR p-value:	0.0001776

	coef	std err	z	P>|z|	[0.025	0.975]
x1	0.3923	0.128	3.065	0.002	0.141	0.643
x2	-0.1938	0.126	-1.535	0.125	-0.441	0.054
x3	-0.2294	0.127	-1.811	0.070	-0.478	0.019
x4	-0.3690	0.129	-2.865	0.004	-0.621	-0.117

Figure 10.2

Extra Credit:

I. While exploring the correlation between the statistics in the dataset, I noticed that the correlation coefficient between years of art education and artist status is about 0.35, which is reasonable as people who are receiving an art education are likely going to become an artist. But what surprised me was that the correlation between “sophistication” and artist status is -0.028. I had expected a positive correlation between sophistication and being an artist, assuming that people who consume art more often are likely to be artists, and vice versa, but our data does not seem to support that assumption.