Capstone Project

Introduction to Data Science

Spring 2023

Maya Nesen

Project Goals

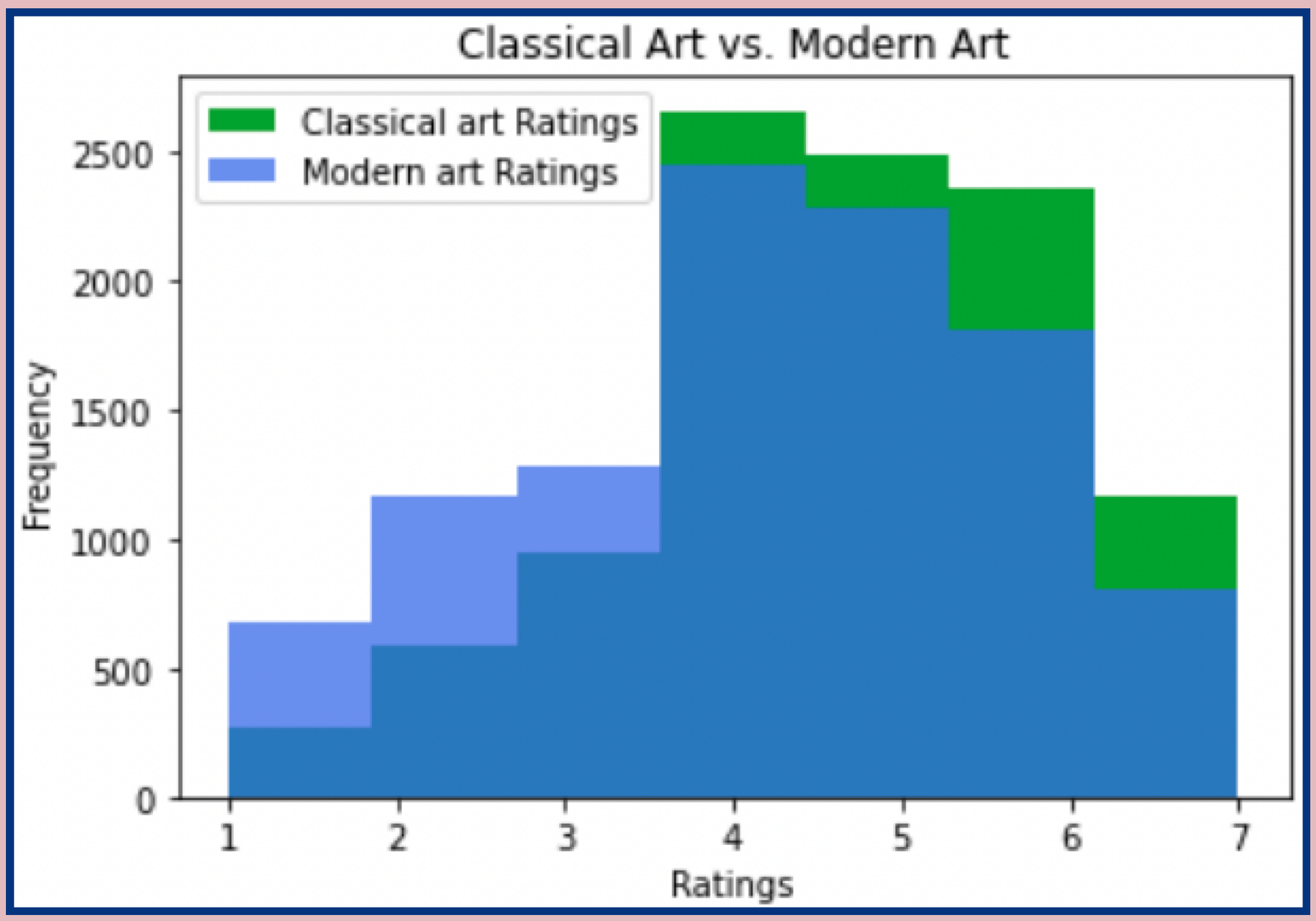
In Introduction to Data Science, I learned about various Data Science techniques, including statistical tests and machine learning algorithms. The goal of this project is to apply the skills and knowledge gained in class through hands-on work with data. Here, I analyze a given dataset of art preference ratings, user demographics, and other miscellaneous information. This project simulates working as a Data Scientist for an auction house that works with a major art gallery.

Pre-processing

To handle any NaN values, I either used a for loop to filter out at any rows that I needed, or I used the .dropna() function in pandas and converted any data back into numpy arrays so that I could perform statistical analyses with them. For art and energy ratings, I primarily used row means to dimensionally reduce the matrices, and Z-scored and used PCA to dimensionally reduce matrices based on questions, such as “dark personality” and “self-image”. For questions 1 through 4, I used an alpha-level of 0.01 and for PCA, I used the Kaiser criterion.

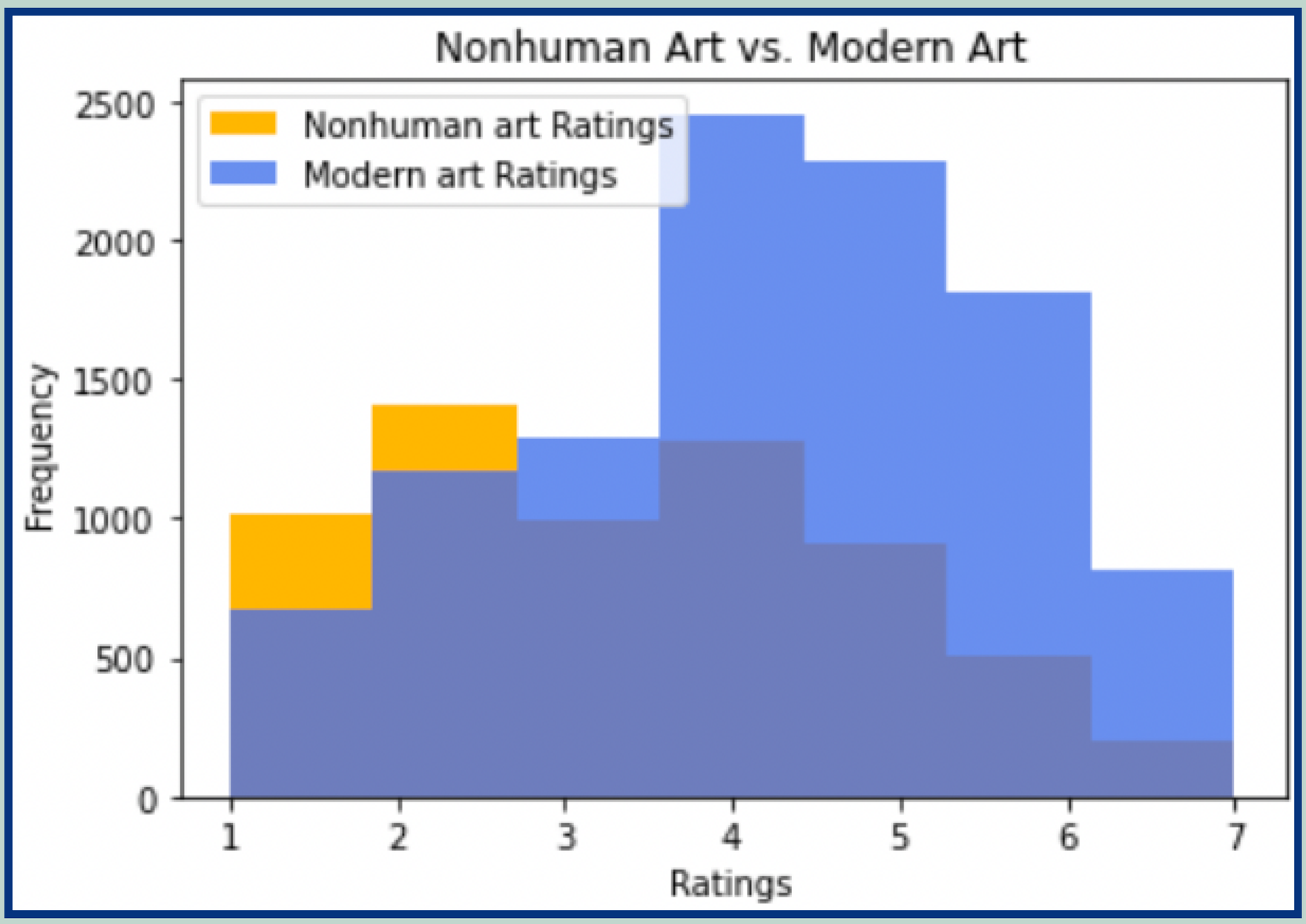
Question 1: Is classical art more well liked than modern art?

To compare the preference ratings of classical art versus modern art for our 300 participants, we first filter out the rows from our dataset for classical and modern art in separate arrays, followed by flattening the matrices. Then, we take the medians of the flattened matrices: the classical art median is a 5 out of 7, while the modern art median is a 4 out of 7. Therefore, the median of classical art is higher than the median for modern art. However, how likely is this due to chance? Let our null hypothesis be that the preference of classical art being higher than that of modern art be solely due to chance. Since our data here is ordinal, due to it being composed of ratings, and we are comparing medians, we use a Mann-Whitney U test to retrieve a p-value of 3.176e-97. Taking our alpha value to be 0.01, our p-value is largely below the threshold. Thus, our result is statistically significant and we can confidently drop the null hypothesis: classical art is indeed more well liked than modern art. We can additionally see this conclusion from the histogram: classical art has more high ratings than modern art.



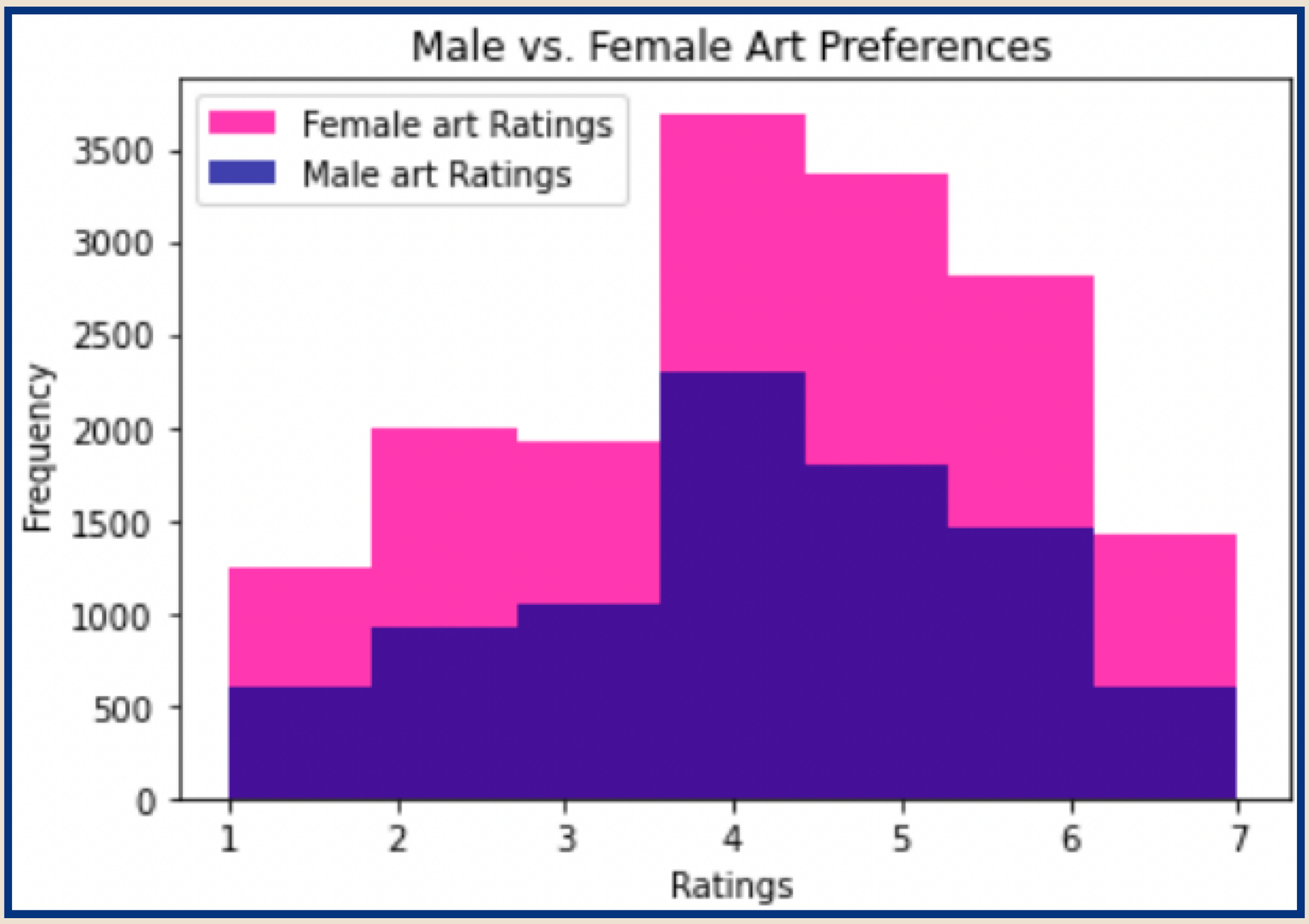
Question 2: Is there a difference in the preference ratings for modern art vs. non-human (animals and computers) generated art?

To compare the preference ratings for modern art versus non-human art, we perform the same analysis as the previous question, but using the rows from our dataset for non-human art and modern art. Comparing the medians, modern art still has a 4 out of 7, while non-human art is a 3 out of 7. Thus, the median for modern art is higher this time. However, this could be due to chance alone. Let our null hypothesis be that the preference of modern art being higher than that of non-human art is due to chance alone. Again, we use the Mann-Whitney U test, thus getting a p-value of 8.74e-264, which is largely below our alpha of 0.01. Thus, our result is statistically significant and we can confidently drop the null hypothesis: modern art is indeed more well liked than non-human art. We can see this conclusion as well from our histogram: non-human art has more low ratings than modern art and significantly less high ratings, clearly demonstrating an overall preference for modern art.



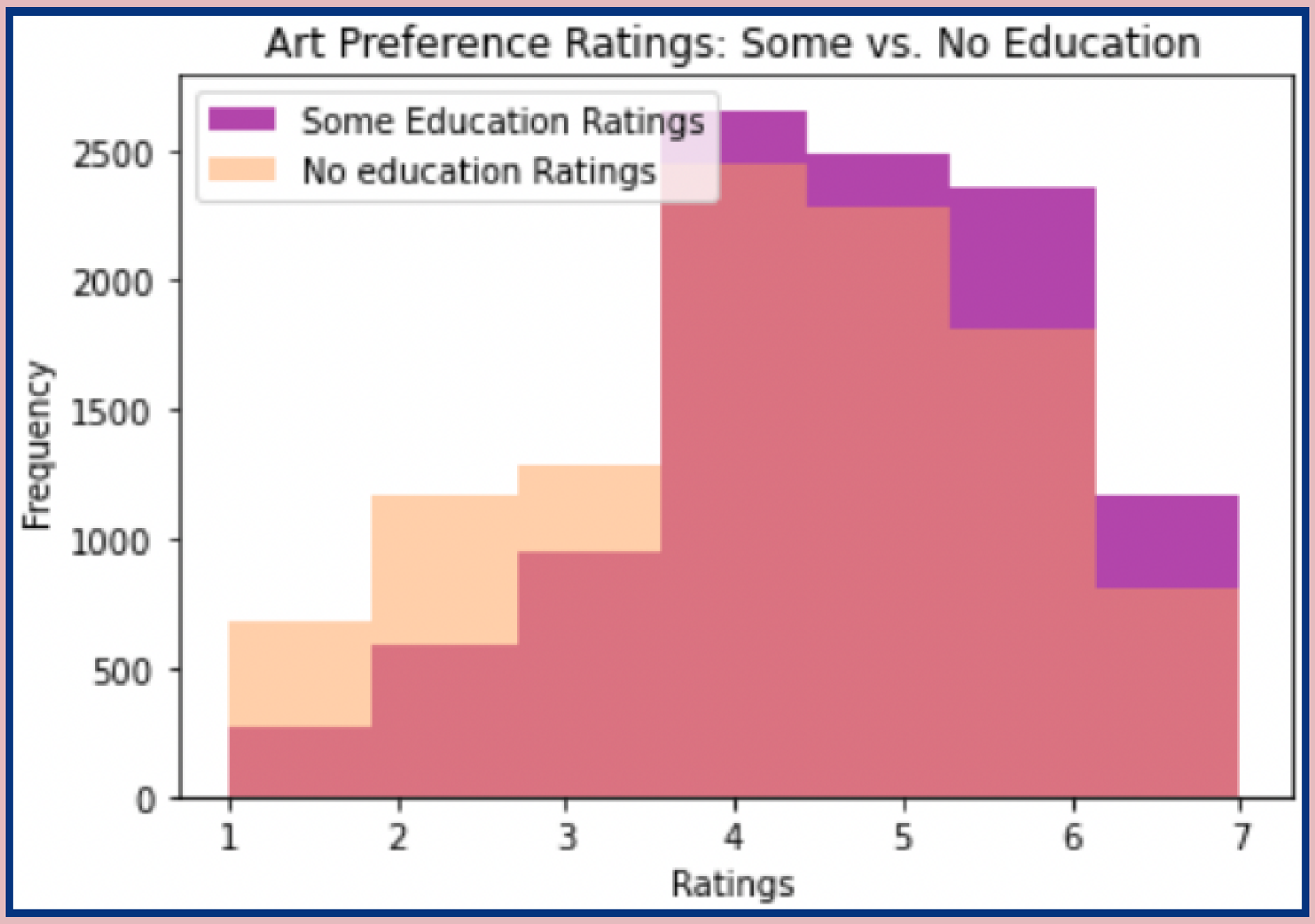
Question 3: Do women give higher art preference ratings than men?

To compare the preference ratings from women compared to men, we perform the same analysis as the previous question, but using the rows from our dataset based on the values in the column for the participants' gender (1 is man, 2 is woman, 3 is non-binary). Using a for loop, I was able to filter out the rows, avoiding any entries in the gender column which had NaN values. Once again, we flattened the matrices and compared the medians: both the male and female medians are a 4 out of 7. Let our null hypothesis be that our medians being the same is solely due to chance. Using the Mann-Whitney U test to determine statistical significance, we get a p-value of 0.373, which is above our alpha of 0.01. Therefore, our result is not statistically significant and we fail to reject the null hypothesis. This signifies that there is not enough evidence to conclude that the median male art preference is different from the median art preference of women. Our histogram shows that although the median for both female and male art ratings is 4, there are more female art ratings than male, so we cannot clearly conclude that the medians are significantly different.



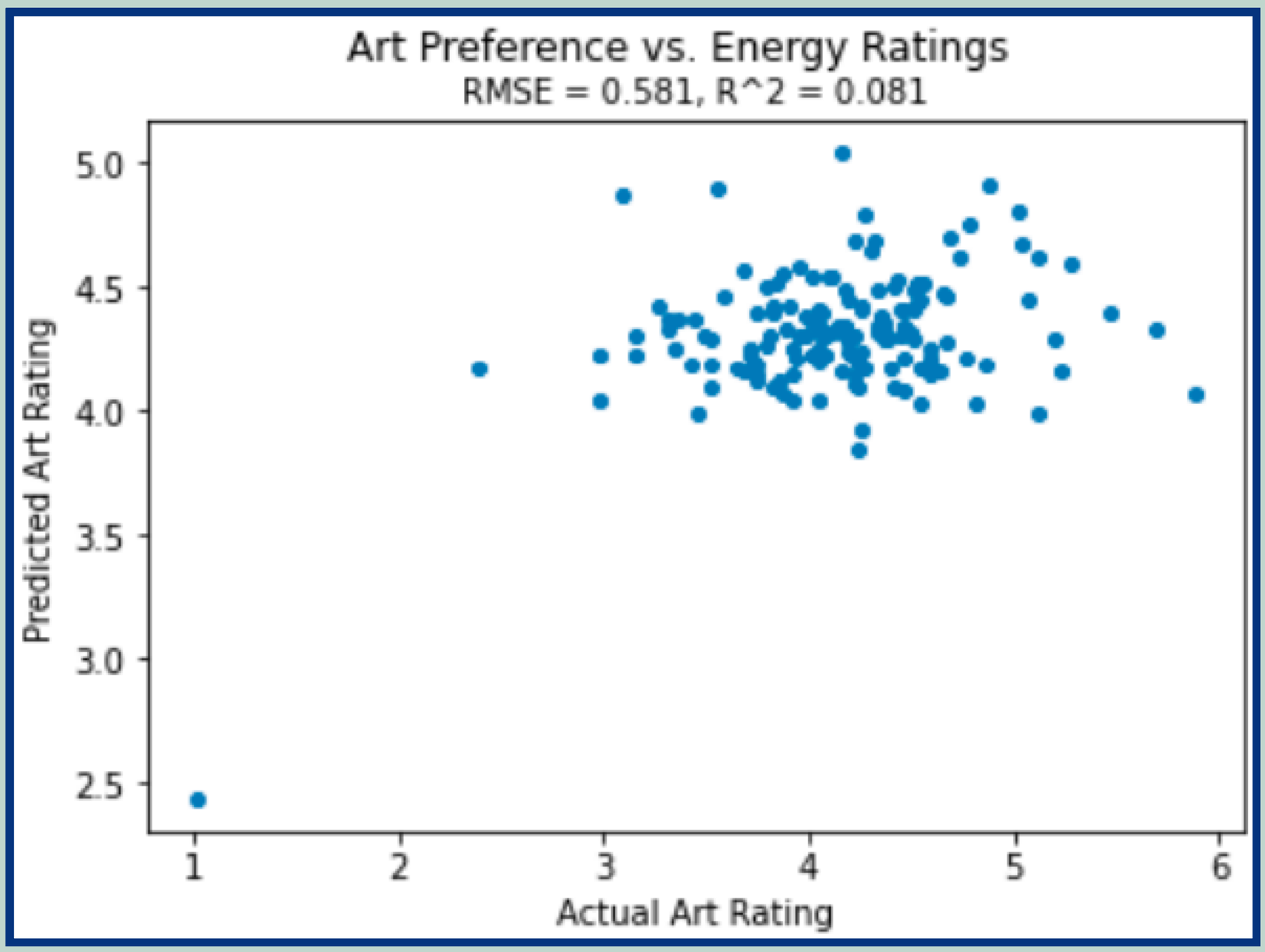
Question 4: Is there a difference in the preference ratings of users with some art background (some art education) vs. none?

We compare the preference ratings from users with some art education background versus none by filtering out the rows based on the values in the column for participants' education (1, 2, or 3 years of education or 0 for no education) using a for loop, which also avoids any NaN values. We flatten the matrices and compare the medians: both medians are a 4 out of 7. Let our null hypothesis be that our medians being the same is solely due to chance. Using the Mann-Whitney U test to determine statistical significance, we get a p-value of 0.005, which is less than our alpha of 0.01. Therefore, our result is statistically significant and we can reject the null hypothesis. Thus, this means that the medians of ratings from some education versus none are likely the same, which we can also see by our histogram: the frequencies of the ratings are distributed very similarly.



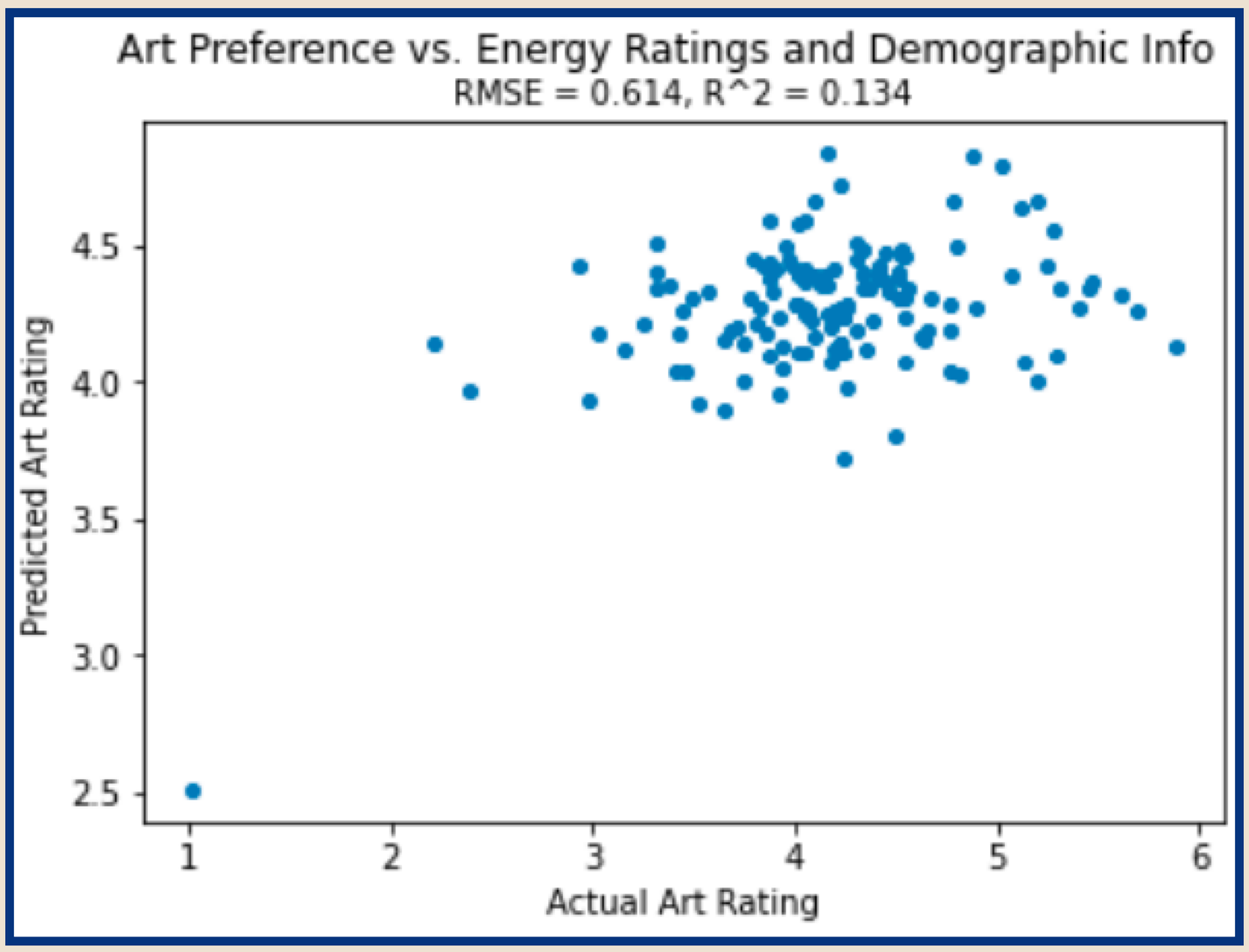
Question 5: Build a regression model to predict art preference ratings from energy ratings only. Make sure to use cross-validation methods to avoid overfitting and characterize how well your model predicts art preference ratings.

To predict art preference ratings from energy ratings, we use linear regression where art preference ratings are used as the outcome or dependent variable, and energy ratings are the independent variable or predictor. We first drop the NaN values from the energy ratings matrix, then dimensionally reduce the energy and art matrices by taking the row means since we want to compare the ratings per user. We then use cross-validation to avoid overfitting through a 50-50 train-test split, and run our regression model. We use RMSE and R^2 to assess how well our regression model predicts our outcomes. Our RMSE is 0.581, meaning that, on average, our predicted art rating values deviate from the actual art rating values by 0.581 units. Our R^2 of 0.081 indicates that approximately 8% of the variance in the art ratings is explained by the energy ratings in this regression model. This means that our model has limited predictive power and a fairly weak relationship between energy ratings and demographic information versus art ratings. The majority of the variance is not captured by the model. Thus, our model is likely not explaining the underlying relationships between our variables and needs to be changed or improved to fit the data better.



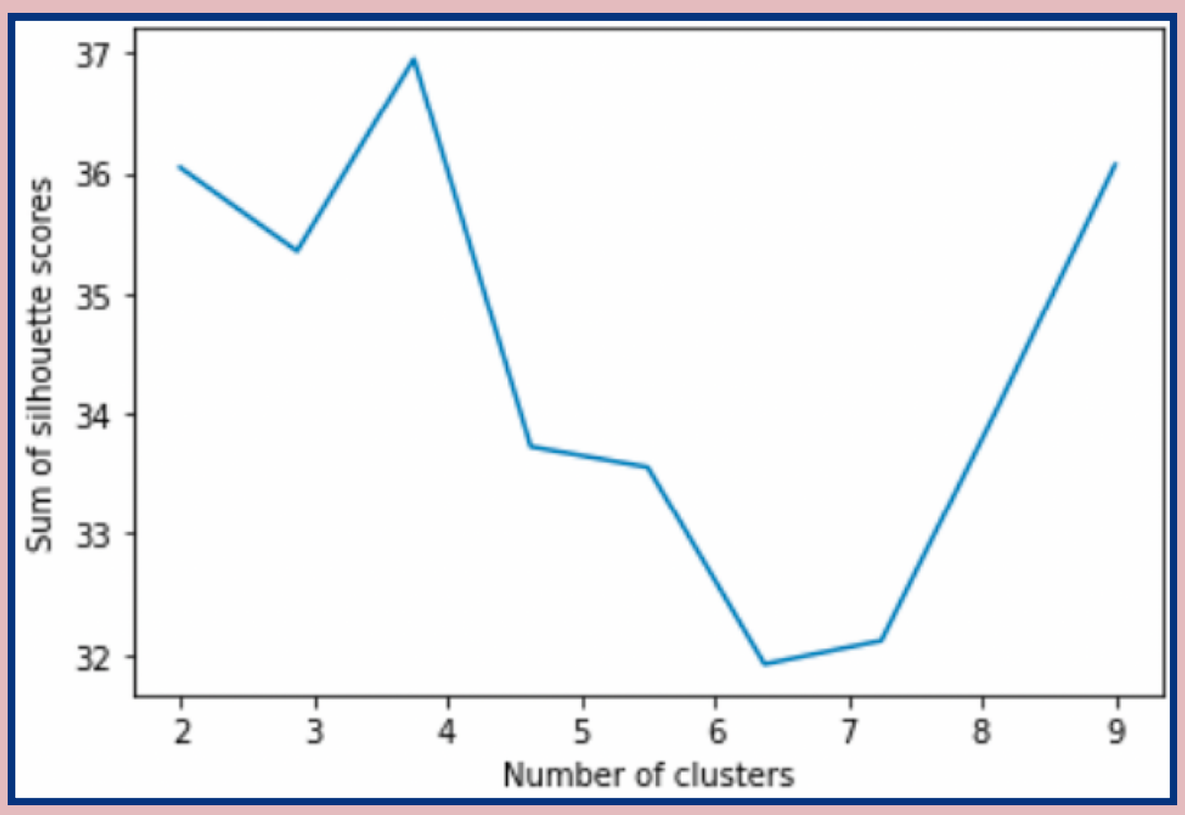
Question 6: Build a regression model to predict art preference ratings from energy ratings and demographic information. Make sure to use cross-validation methods to avoid overfitting and comment on how well your model predicts relative to the "energy ratings only" model.

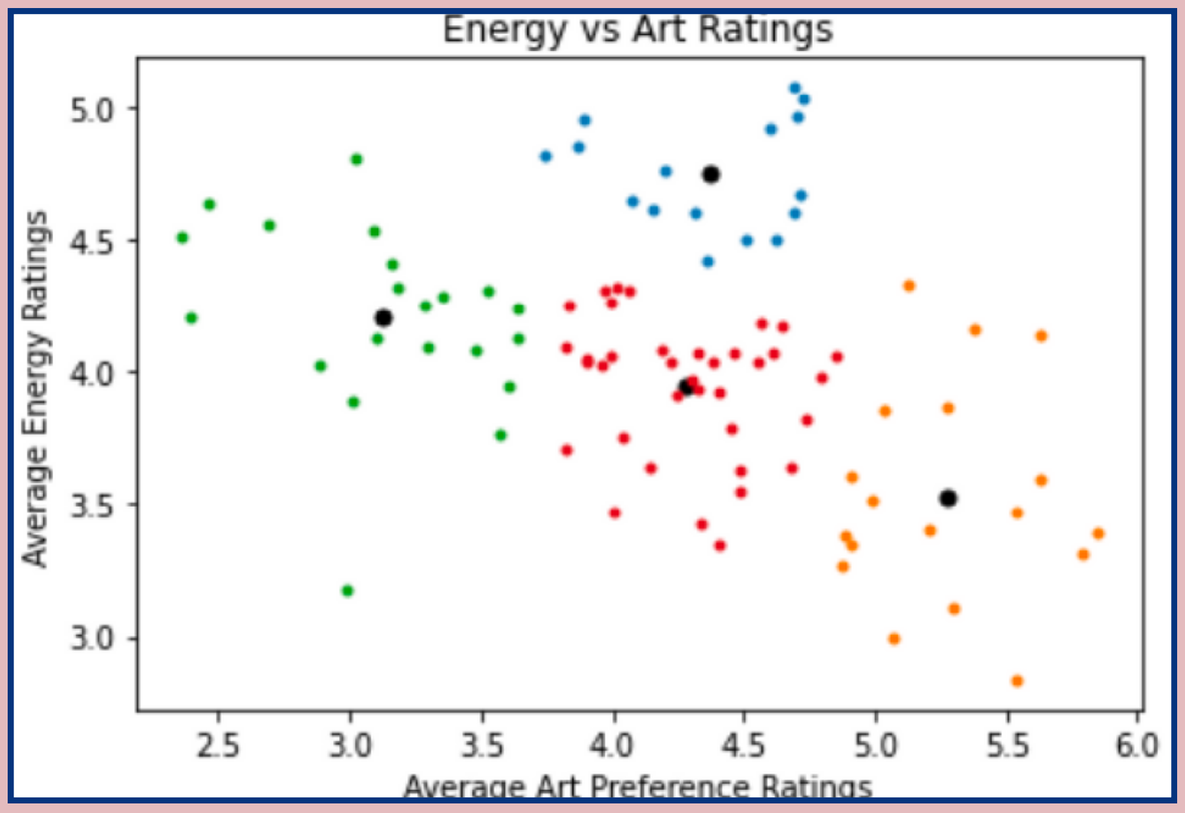
To predict art preference ratings from energy ratings and demographic information, we use multiple regression where art preference ratings are used as the outcome or dependent variable, and energy ratings and demographic information (age and gender) are the independent variables or predictors. We first drop the NaN values from the rows of energy ratings and the age and gender columns by concatenating the columns and matrices. We then dimensionally reduce the energy and art ratings by taking the row means. We concatenate our new energy, age, and gender into one to use as our input for a 50-50 train test-split, and then run our regression model. We get an RMSE of 0.614, meaning that, on average, our predicted art rating values deviate from the actual art rating values by 0.614 units. We get an R^2 value of 0.134, which indicates that approximately 13% of the variance in the dependent variable is explained by the independent variables in the regression model. Thus, our model again has limited predictive power and a fairly weak relationship between energy ratings and demographic information versus art ratings. The majority of the variance is not captured by the model. Our model likely is not explaining the underlying relationships between our variables and needs to be changed or improved to fit the data better.



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?

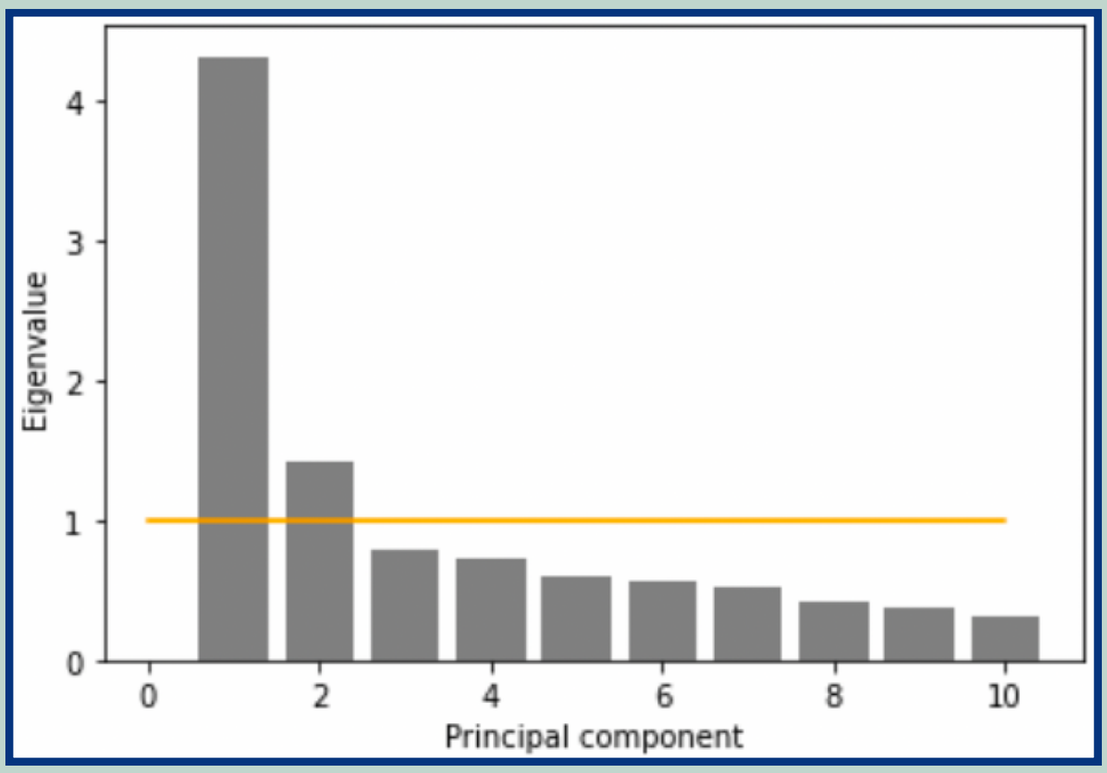
To compare average preference art ratings to average energy ratings, we begin by dropping any NaN values in each matrix and reducing the data by taking the column means so that we can compare the ratings for the 91 art pieces. To use k-means clustering, we first use the silhouette method to find the ideal number of clusters. In the plot of the sum of the silhouette scores, the peak is at k = 4, so we use 4 clusters. Using the clustering algorithm, we see in our plot our 4 clusters. Note that energy ratings are on a scale of 1 (“it calms me down a lot”) to 7 (“it agitates me a lot”). The orange and blue clusters are particularly informative. The orange cluster, corresponding to mostly classical art pieces, shows that higher average art ratings have lower average energy ratings, signifying that people who enjoy these art pieces feel more calm. The blue cluster, corresponding to mostly modern art pieces, shows that art pieces that are “ok” have higher energy ratings, signifying that people who look at these pieces get more agitated when they see them. The green cluster corresponds to mostly non-human art pieces, which have high energy ratings, meaning they agitate the users more, and thus have low art preference ratings

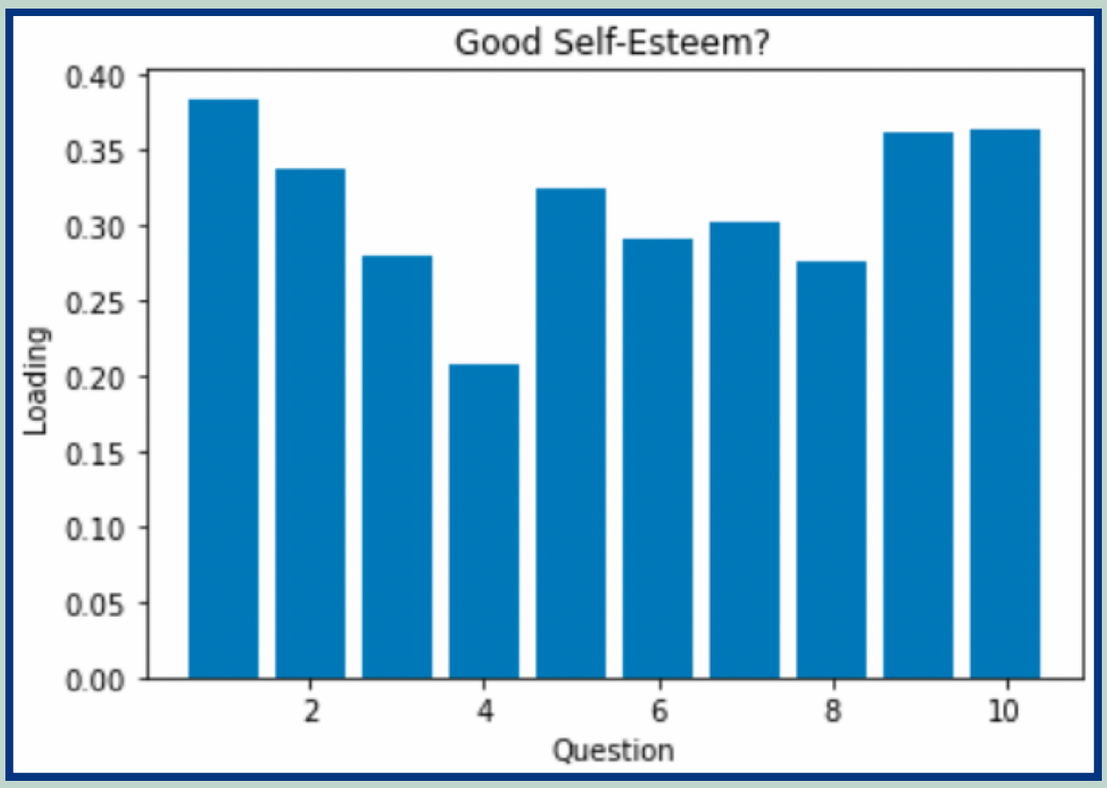


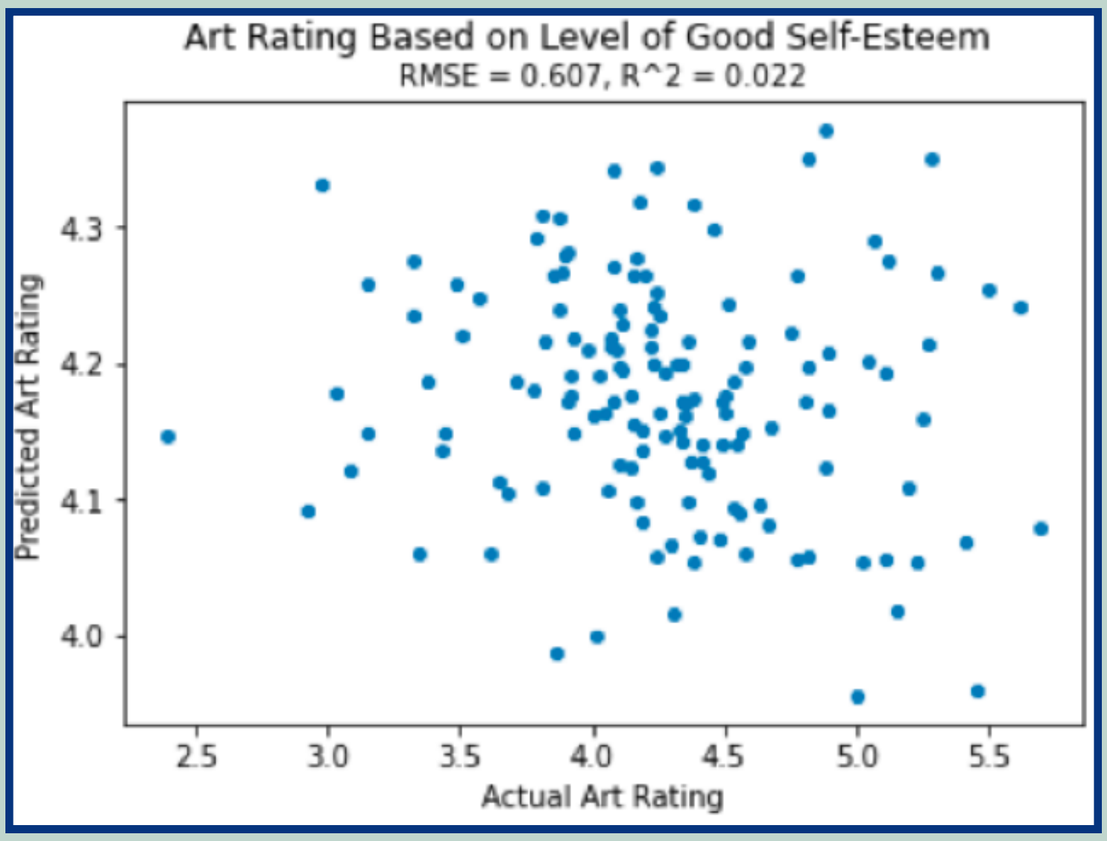


Question 8: Considering only the first principal component of the self-image ratings as inputs to a regression model – how well can you predict art preference ratings from that factor alone?

To predict art preference ratings based on the first principal component, we need to perform principal component analysis (PCA). The self-image ratings are based on 10 questions, so we want to reduce these questions to one that can encapsulate them well. After dropping any NaN values, we Z-score the data and perform PCA. From the scree plot, we see our eigenvalues sorted from high to low: the first principal component is dominant above the others and the Kaiser criterion of 1, representing about 43% of the variance explained. This signifies that the first out of the 10 questions ("On the whole, I am satisfied with myself") is a fairly good representative of the other nine, which we can put under the category of "Do I have an overall positive self-esteem/self-image?" This is further shown by the bar plot based off of the loadings matrix: all the questions are positively similar, based on the first principal component. After doing a 50-50 train-test split with the first column of the rotated data matrix, we perform linear regression. We get an RMSE of about 60%, which signifies that, on average, the predicted values deviate from the actual values by approximately 0.6 units. We get an R^2 value of 0.022, which indicates that approximately 2% of the variance in the dependent variable is explained by the independent variables in the regression model. Therefore, once again, our model seems to have little explanatory power and a moderate level of error. The model is likely not capturing the underlying patterns of our data here, and should be improved as well in terms of model fit and prediction accuracy.

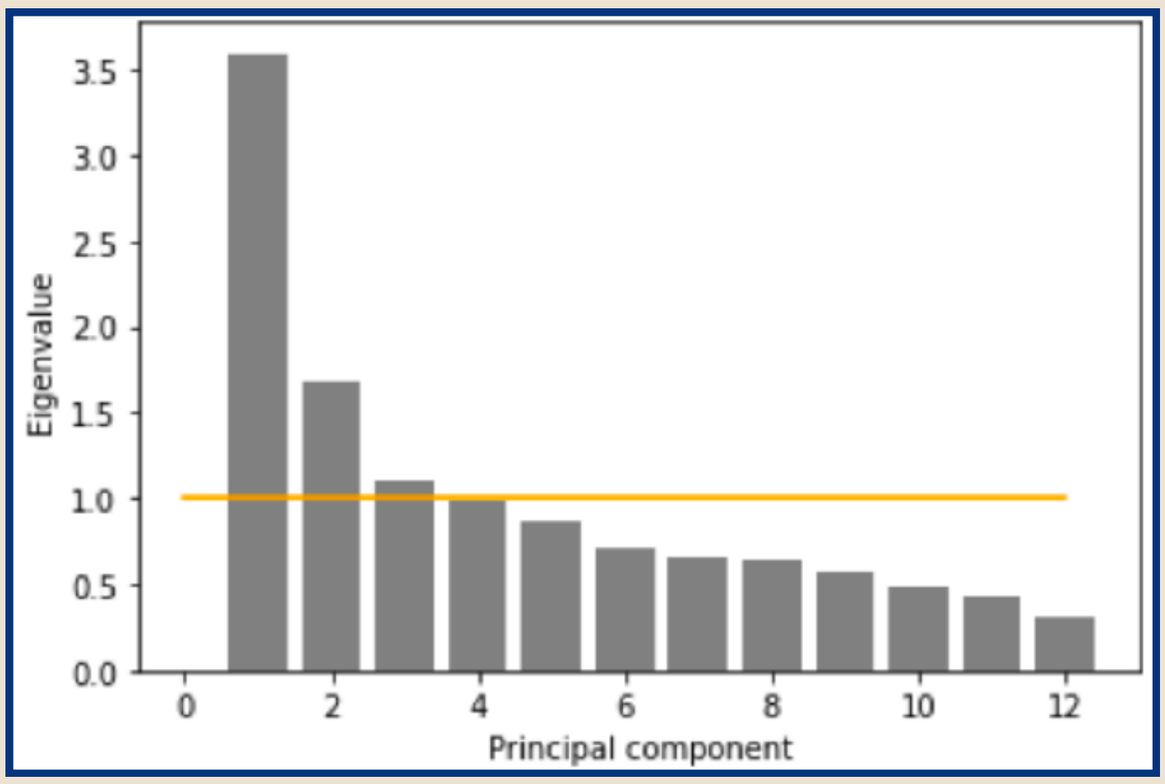


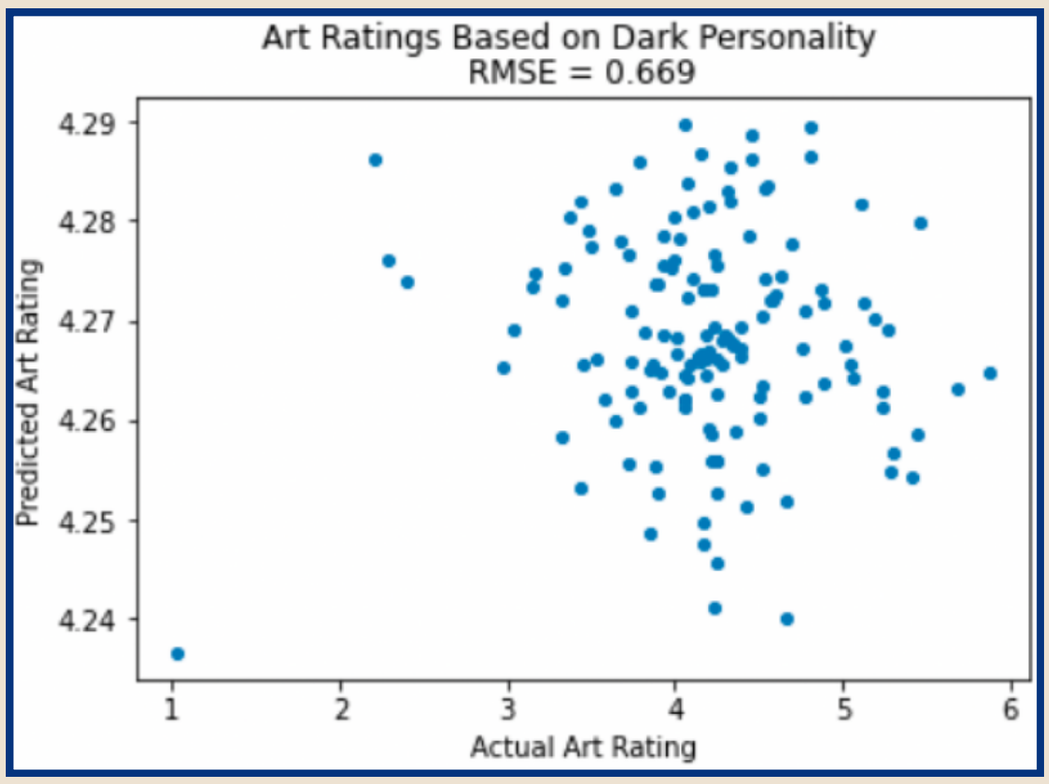


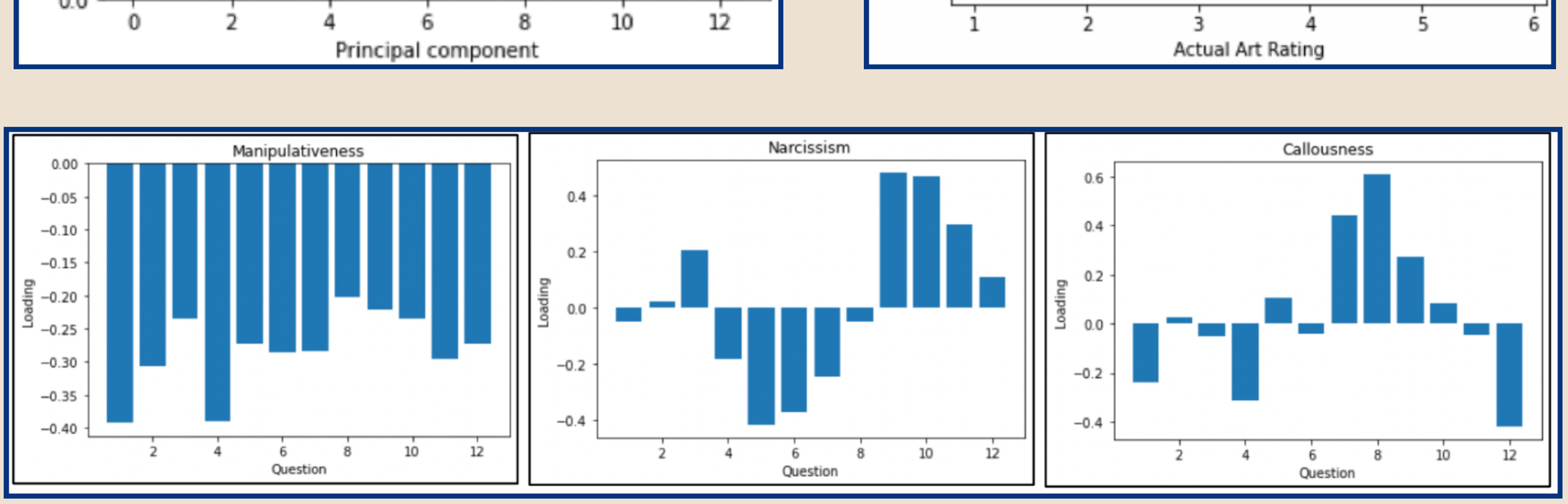


Question 9: Consider the first 3 principal components of the "dark personality" traits – use these as inputs to a regression model to predict art preference ratings. Which of these components significantly predict art preference ratings? Comment on the likely identity of these factors (e.g. narcissism,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

Similarly to the previous question, we perform PCA to determine the first three principal components of the "dark personality" traits, which is composed of ratings on a scale of 1 ("strongly disagree") to 5 ("strongly agree") based on 12 statements. We want to reduce the 12 statements to three that are representative of the others. Based on our scree plot, our first three principal components are significant since they are above the Kaiser criterion of 1, with each component representing about 29%, 14%, and 9% of the variance respectively.Based on our bar plots of the individual principal components and the 12 statements, we can say that the first principal component is "manipulativeness", the second is "narcissism", and the third is "callousness". We then use a multiple regression model, using our 3 principal components as our independent variables, and the art ratings as our dependent variable. We get an RMSE of about 67%, meaning that, on average, the predicted values deviate from the actual values by about 0.669 units. Once again, our model has limited predictive power and is not explaining underlying relationships very well.







Question 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 use logistic regression to predict a binary outcome ("left" or "non-left") based on the rest of our dataset. We first handle NaN values using the pandas .dropna() function. We then turn our "political orientation" column into binary so that 0 represents "left" and 1 represents "non-left" using a for loop: if the values in the column are 1 or 2, then we mark it as 0 for "left, and anything else is 1 for "non-left". Then, we perform three PCAs: one for "dark personality", one for "self-image", and one for "action preferences". Based on the scree plots, I decided to use the first three principal components for "dark personality" and "action preferences", and the first two principal components for "self-image". We additionally take the means of our art and energy ratings. After re-organizing our data, which is now dimensionally reduced and without the politics column, we use a 50-50 train-test split to cross-validate our data to use for our multiple logistic regression model. For our regression plot, as an example, we use the first column of our test data, which is our average art preference ratings. The accuracy score of 62.3% indicates that our model correctly predicts the political leaning of our users 62.3% of the time. This may seem reasonable, but we need to additionally consider the AUC score. The AUC score is 0.5079, which is very close to 0.5, meaning the score is for a random model, and cannot distinguish well between whether our users are "left" or "non-left".