Pre-processing

Pre-processing (dimensionality reduction, transformations, and data cleaning) was handled in the most appropriate way that was seen fit. To begin, in order to handle missing data or "NaNs", the rows with nans were deleted. Although the common argument against this decision resides in the fact that deleting full rows would remove a lot of data, only 22 rows were removed, which is only 7% of the data. In addition, removing rows with missing data was the best option because removing the nans would ruin the matrix, and replacing the nans with some other value, for example, such as the mean of each column, would create negative implications for clustering. In addition, in terms of dimensionality reduction, we did not reduce dimensions for the first question because the columns, or variables, for the main dataset were rows, or observations, in the art dataset. Since we are studying the observations, we cannot reduce all of the art columns in the main dataset just yet. In terms of transformation, we used the z-score of the data before we dimensionally reduced the data to answer other questions.

Question one

The data was not dimensionally reduced in this problem because the observations in the art dataset were columns in the main dataset. If we were to reduce the main dataset, the columns, which are also observations, would be lost. Since the data was not normally distributed and used ordinal data in the form of ratings, a t-test or any other parametric test would not be useful and if done, would not produce meaningful data. The data cannot be reduced to the sample mean because the mean is a normalized sum and presumes the unit of the data we are summing are equal when they are not. The psychological distance between each rating is not the same (the distance between a rating of 2 and 3 is not the same distance as a rating between 3 and 4). Since the median does not depend on distance, but rank, the Mann-Whitney U test is the optimal significance test. In addition, the question asked if people preferred classical or modern art better, but there was no column of modern art ratings and classic art ratings. Therefore, one had to be created. The ratings for each classical art piece were extracted and placed in a one dimensional array. The same process occurred for modern art. The Mann Whitney U test was then performed. The u test statistic was 38814531.5 and the p-value was 1.3928200744791768e-87. The p-value was significant at the alpha level of 0.05. Since modern art was the first group entered in the mann-whitney u test code, and the u test statistic was positive, the classical art was better liked than the modern art. In addition, the median score for classical art was 5 while the median score for modern art was 4.

Question two

The question asked if there was a difference between the preferences of modern art and non-human art. Since we are working with the same data as the previous question, where the data is holding the same assumptions, another mann-whitney u test was performed to determine if there was a difference between the two populations in terms of preference. The same method in question 1 was applied, where the preference was extracted using data from the art dataset and placed into one array for each art type. Then the test was performed. The u value was u = 19013098.0, and the p value was 1.3045360828419867e-243. Although this seems like a lot, the dataset had almost 10000 (actually 9625) points for each group, which makes sense because of the law of large numbers.

Question three

The question asked if women give higher preference ratings than men. Since we are still using the same dataset, which holds the same assumptions as the prior two questions (ordinal data with no normal distribution), we will use the Mann Whitney u test to determine if there is a difference between these two independent populations: men and women. Since the sample size is still very large (15925 for women and 8554 for men), the u score was 68961959.5 and the p value was 0.10113, which was not significant. Therefore, women do not give higher preference ratings than men because there is no difference in terms of preference between the two populations.

Question four

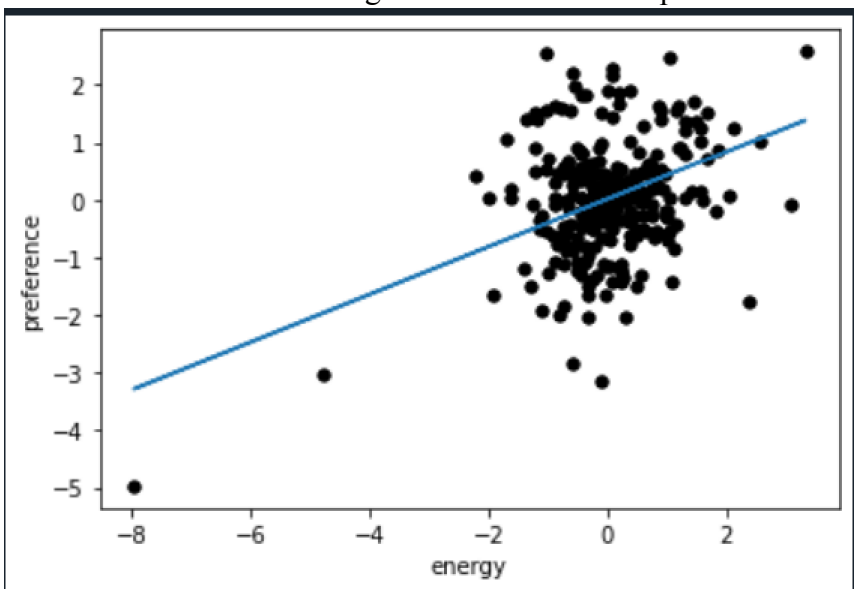
The question asked if there is a difference in terms of preference between users who have some art background versus users who have none. I conceptualized some art background as anything but none, which translates in the dataset as any input that was not 0 (which represents no art background). I extracted preferences from users with some art background by looping through the dataset and adding each user's preference into its designated array. Then, since the data still holds the same assumptions (data is ordinal and does not have normal distribution), I used the Mann Whitney U test to determine if there was a difference between the populations. The u score was 65646419.5, the p value is 3.7774486070036225e-08, which means it is significantly different at the 0.05 alpha level. The size of the art educated group is 16926, and the size of the art non-educated group is 8099, which are extremely large sample sizes. In addition, the art education group is twice as large as the art non educated group, giving their group a better representation according to the law of large numbers. Since the other group was not as represented, it could present issues in terms of interpreting the data but 8099 is still large enough to interpret the data meaningfully.

Question five

The question asked to predict preference ratings from energy ratings only. Because the question asked to use a linear regression model, which assumes the data is continuous (when rating data is not), I had to convert the data to the mean of each art piece rating. Again, extracting the mean from rating data is not informative, however, it must be done to draw any conclusion from the linear analysis model. Once the mean was extracted from each user for both preference and rating of the art piece, I cross validated and interpreted the statistics. For the model using average rating of each art piece, splitting the data into ¼ test set and 3⁄4 training set and doing this 3 times, I received cross validation scores of 0.05043686, 0.14668896, -0.0259419 and a mean score of 0.057. The figure below shows the standardized raw data.

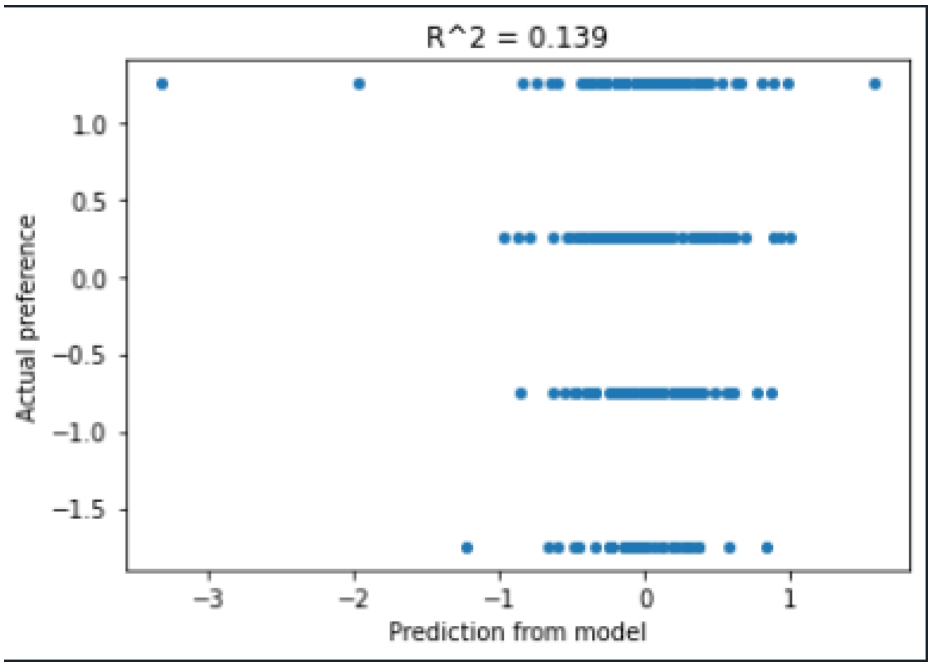
For this analysis, the mean squared error was 0.96 and its standard deviation was 0.15. Although the variance is not well explained in this model and the testing score is negative (-0.13403638013411556), this lack of accuracy can be explained by the fact that the data is ordinal and not continuous, which means that a classification algorithm would have worked

better. The blue line is the regression line with a slope of 0.4131678 and an intercept of 0.0063.



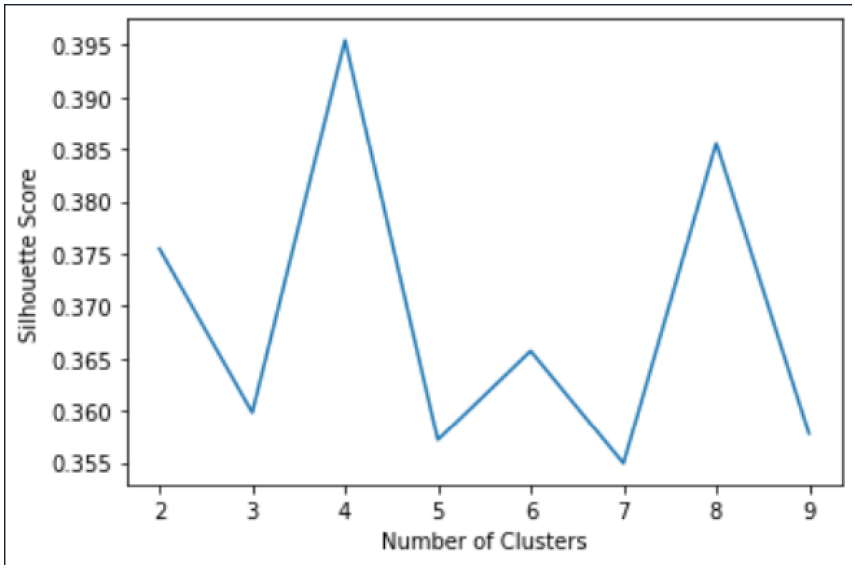
Question 6

The question asks to build a regression model to predict art preference ratings from demographic ratings and energy ratings. It also asks to use cross-validation methods and describe how well the model did compared to the previous one. First, I extracted the columns we wanted to use (energy and demographics). For energy columns, instead of using the energy rating for each column, which would be very noisy, I used the mean energy rating. The reason I chose to use the mean instead of the median is because although the data is ordinal, we are using linear regression, which works best with continuous data. In addition, the prediction, preferences, were reduced to the mean preference of each user for all paintings for the same reason. The rest of the predictors, however, are still categorical (ex: 1 for women and 0 for man). This makes linear regression a really bad model to predict preferences from using these types of predictors. This lack of accuracy is shown in the mean score after cross validation: 0.05. On average, only 5% of the variance is explained by this model. The figure above uses an r squared value 0f 13% because it is specific to that iteration. Compared to the prior model that only used energy ratings, this model still does not perform better with a similar mean mean squared error of 0.92 and its standard deviation 0.12. This lack of accuracy is accounted for by the type of data being used–ordinal rating data and categorical demographic data, which do not respond well to linear regression. For example, the data in the graph isn't even continuous because of the categorical demographic information used in the model.

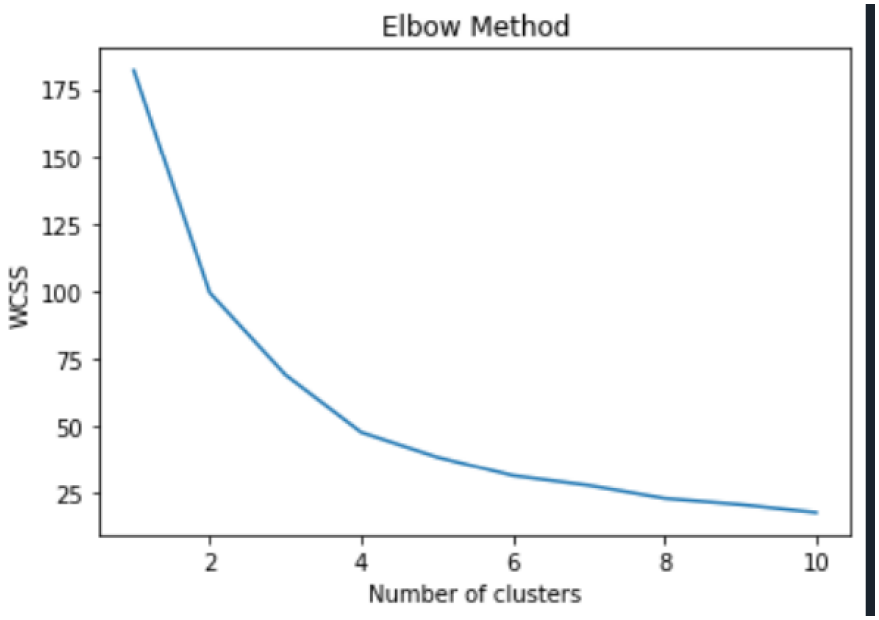


Question 7

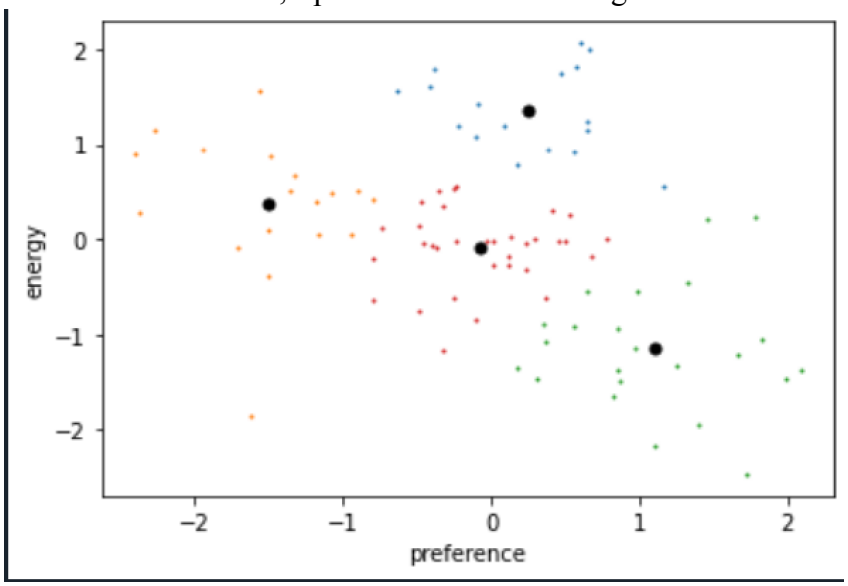
This question asks to consider the 2D space of average preference ratings vs. average energy rating, algorithm identify a number of clusters, and interpet what each of those clusters could be. First, I found the mean preference of each painting, and the mean energy of each painting and placed each in their own array using a for loop. Then, I scored the means because the preferences and energies use different scales. Next, I used the silhouette method to find out what number of clusters I should use for this data. The silhouette score was as follows:



A high silhouette score means that the data is well matched to its own cluster and poorly matched to neighboring clusters. 4 clusters gives the highest silhouette score. Next, I used the elbow method to verify that 4 clusters is sufficient.



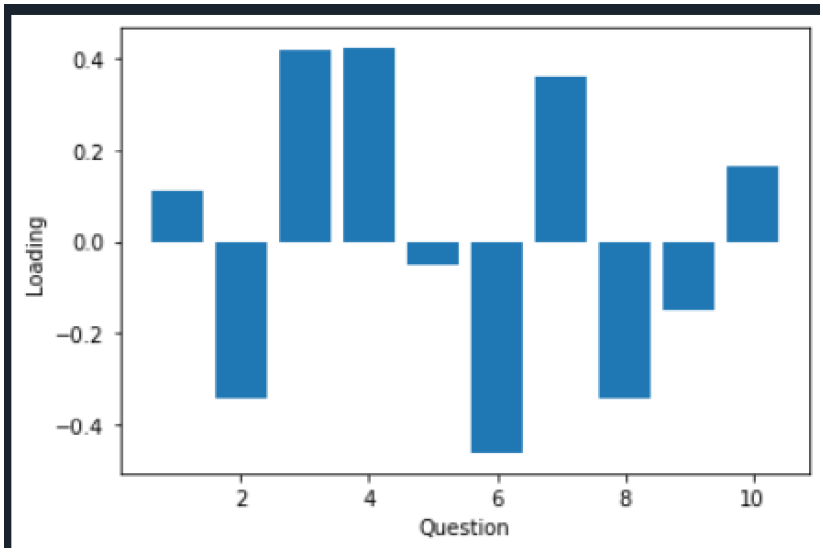
The elbow method also showed that 4 clusters are sufficient because the line starts to level off at about 4 clusters. Next, I perform the k-means algorithm on the data using 4 clusters.



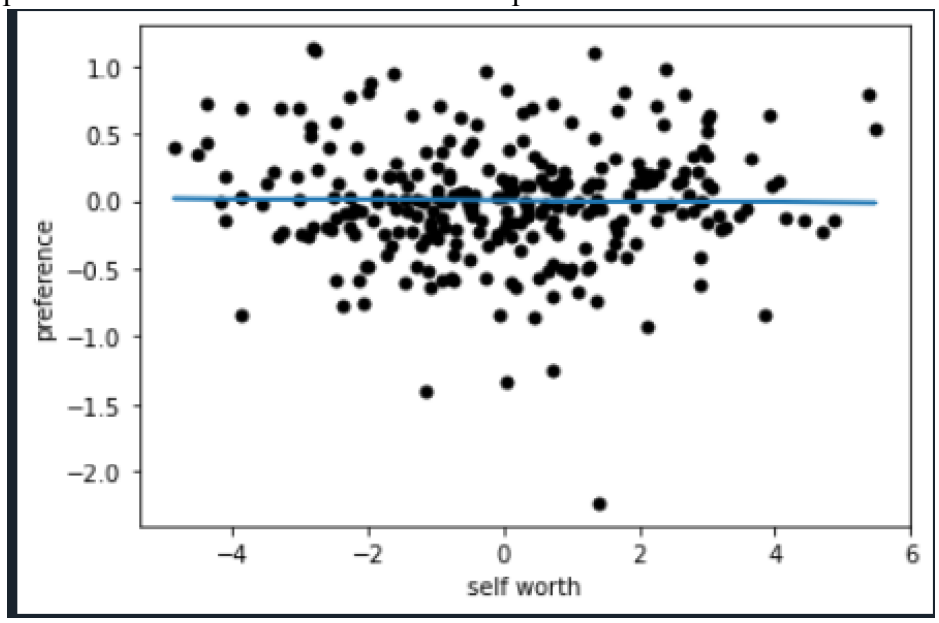
Now, I interpret what each cluster could be, and if they correspond to any particular types of art. I would say red is non-human art, which is about average. For the rest, calming art of nature in classical art would be low energy high preference like green. Orange would be modern art. Blue must be classic art that is more intense.

Question 8

The question asks to consider only the first principal component of the self-image ratings as inputs to a regression model, and asks about the validity of the model. First, I did PCA. The pca had loadings from questions 2,6,8 with reverse polarity and 3,4,7 with polarity. Therefore, I concluded the first principal component is general self-worth (how I feel about myself).

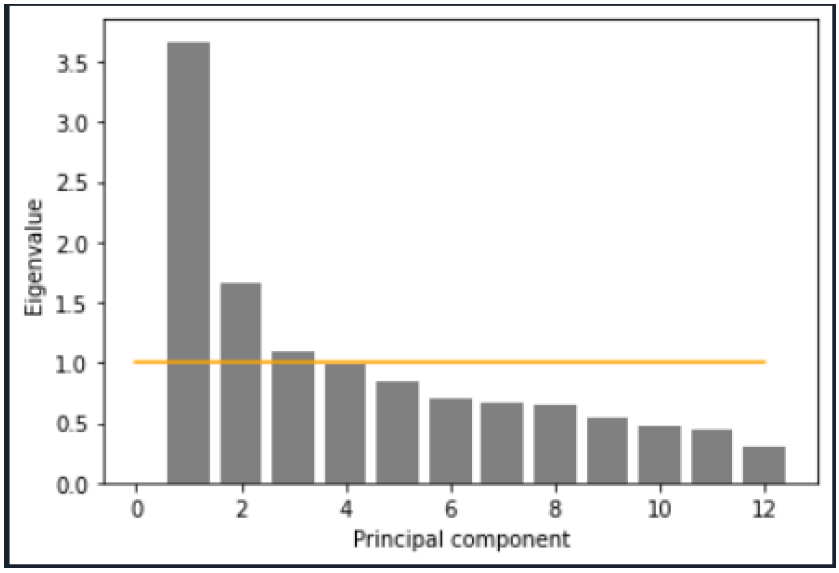


Next, I performed a linear regression model to predict preference ratings (mean preference rating). The intercept was almost 0, the coefficient was -0.0032, and the mean cross validation score was -0.05115430908258822, meaning the model performed worse than just guessing the mean. The model performed badly because there is no relationship between self worth and art preference. The data was scored before pca.



Question 9

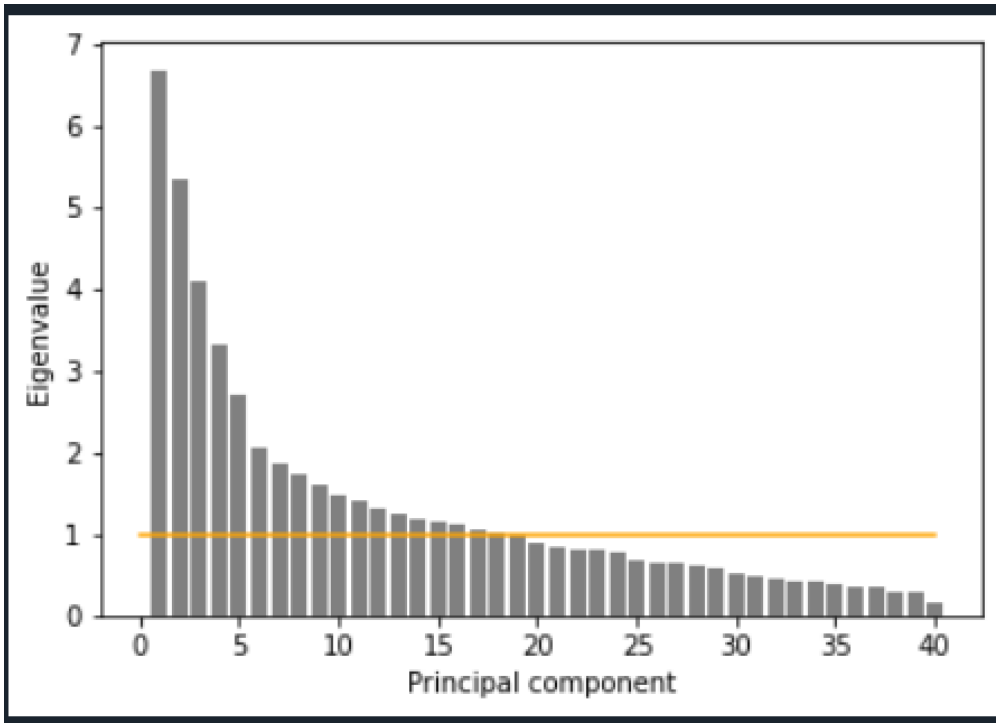
This question asks to) consider the first 3 principal components of the “dark personality” traits – use these as inputs to a regression model to predict art preference ratings, which of these components significantly predict art preference ratings, and to comment on the likely identity of these factors (e.g. narcissism, manipulativeness, callousness, etc.). First, I did pca with z scored data.



The 3 pca factors in order were narcissism, callousness, and manipulation. The linear regression model did poorly, with a testing score of 0.086. Callousness, the third principal component, explains most of the three, with a r-squared of 0.036.

Question 10

The question asks to determine the political orientation of the users using a classification model and commenting on the quality of the model. First I did pca, z scored everything except for classification questions, and used 5 components as per the elbow method because 16 components would not be worth it per kaiser criterion.



Next, I identified the identities of all 5 components used in order: manipulation of person, enjoying time with people make you sympathize with them, like video games and mind activities

opposed to physical, preference of indoor activities like art instead of outdoor like hiking, and how badly you want other people to like you. Using five components as predictors, I used a logistic regression model. I used all of the data, with the preference and energy for art as means instead of 91 columns. I computed the auc score as 0.68. The positive (1) was not liberal and 0 was liberal. Below is the rock curve. The auc score was not too bad, therefore the model is not terrible. We can predict from other data the political orientation of the user. This makes sense because linear regression does not work well with ordinal or categorical data the same way classification models do. The classification data violates assumptions of linear regression, yet we use it all the time.