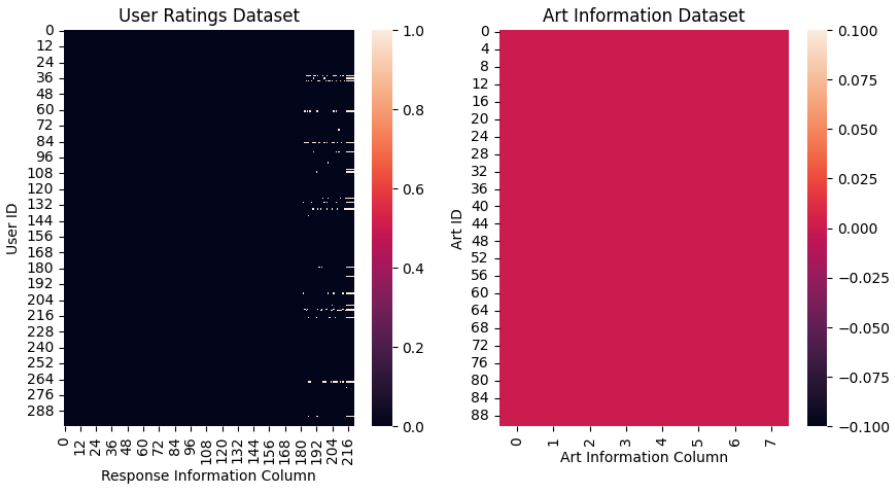
Before doing anything in this project, I created a heatmap of the base data to visualize missing entries, as shown below for the two csv files.



For the data cleaning, I mainly used row-wise removal of nan values as follows. I obtained two sets of columns of interest, merged them into one data frame, and dropped Nan-values, then separated them into two separate data frames. This would ensure that we do not lose extra data compared to just doing dropna on the entire dataset. For example, to identify if there is a difference in the preference ratings for modern art vs. non-human art, it would make sense to obtain those relevant columns and then do drop-na (though it does not seem like there are any nan-values ion columns 180 and below), that way nan-responses for non-columns of interest would not cause unnecessary loss of columns. I made a function for this (for combining art and theData in particular) as follows:

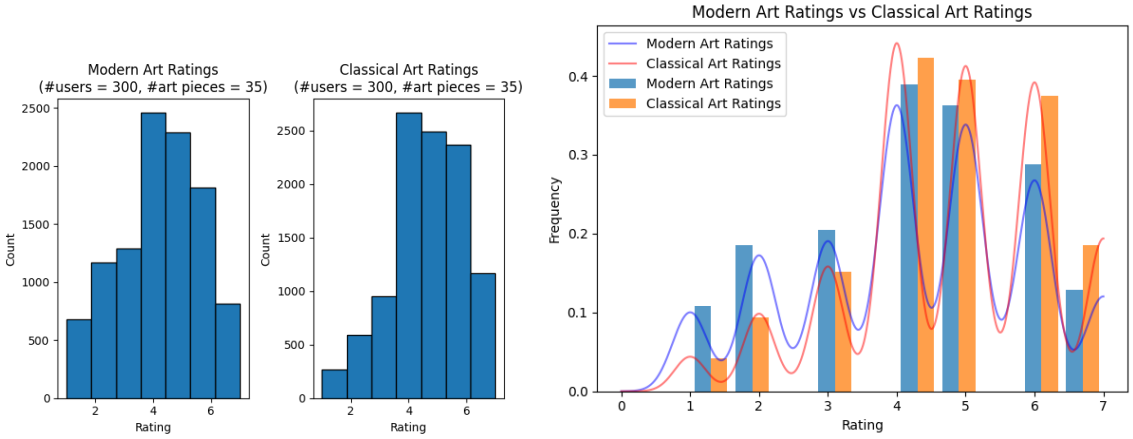
```python
# Helper function to join, drop na, then separate ratings from two indicies
def join_drop_na_separate_ratings(df, indicies1, indicies2):
    # Combine both indicies
    indicies = np.concatenate((indicies1, indicies2))

    # Get ratings for both indicies
    ratings = select_df(df, indicies)

    # Drop NA rows for both indicies
    ratings_no_na = ratings.dropna()

    # Select ratings with index1
    ratings1 = select_df(ratings_no_na, indicies1)

    # Select ratings with index2
    ratings2 = select_df(ratings_no_na, indicies2)

    return ratings1, ratings2
```

I removed outliers as needed in later portions, particularly in portion 9.

I only did PCA for questions requiring dimensional reduction for the user-ratings dataset with the given columns of interest using kaiser criterion.

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

I first had to select classical and modern art ratings to answer this question. To do this, I obtained the indices in the art_ratings dataset where the `source` column was 2 and 1 for modern and classical art, respectively. Then, I obtained ratings corresponding to those indices using the function specified in the pre-processing section above. We can visualize the ratings for each type of art piece with histograms as follows:



Firstly, I used a Kolmogorov-Smirnov test on the two sets of data with the following hypothesis:

H0: Modern and classical art ratings come from the same underlying distribution.
H1: Modern and classical art ratings do not come from the same underlying distribution.

The p-value and ks-statistic for the test were (***p-value*** = *1.1327956716601013e-72,* ***ks*** = *0.1256190476190476*). At a significance level of *alpha=0.05*, we can reject the null hypothesis in favor of the alternative. Evidence suggests that classical art is distributed differently than modern art for this set of responses.

Secondly, I used a Mann-Whitney U test on both sets of data. The Hypotheses were as follows:

H0: The population median of classical art is less than or equal to the population median of modern art,
H1: The population median of classical art is greater than the population median of modern art.

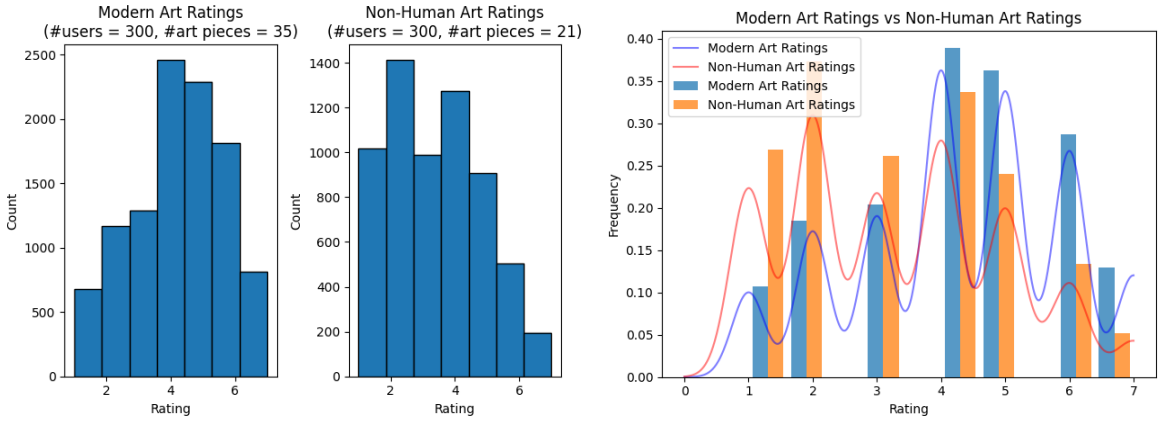
The p-value and u-statistic for the test were **(*p-value*=** *1.5881633286154516e-97* ***u*** **=** *64145482.0*). At a significance level of *alpha=0.05*, we can reject the null hypothesis in favor of the alternative. Evidence suggests that classical art is more liked than modern art for this set of responses.

We can conclude that from these data, classical art is more liked than modern art.

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

To answer this question, similar to the last one, I again had to select two sets of indices in the `source` column, those with values 2 and 3, corresponding to modern art and non-human art indices, respectively. Then I obtained the ratings for each of those indices using the same function specified in the pre-processing section.

The visualization for the data is as follows.



I again used the same two significance tests.

First, I tested to see if the data come from the same underlying distribution using the Kolmogorov-Smirnov test on the two sets of data with the following hypothesis:

H0: Modern and non-human art ratings come from the same underlying distribution.
H1: Modern and non-human art ratings do not come from the same underlying distribution.

The p-value and ks-statistic for the test were (***p-value =*** *6.521237590892692e-207,* ***ks =*** *0.24400000000000005*). At a significance level of *alpha=0.05*, we can reject the null hypothesis in favor of the alternative. Evidence suggests that classical art is distributed differently than modern art for this set of responses.

Then, I used a Mann-Whitney U test on both sets of data. The median is more robust to outliers, so having a non-equal sample size should not matter here. The Hypotheses were as follows:
H0: The population median of modern art is equal to the population median of non-human art,
H1: The population median of modern art is not equal to the population median of non-human art.

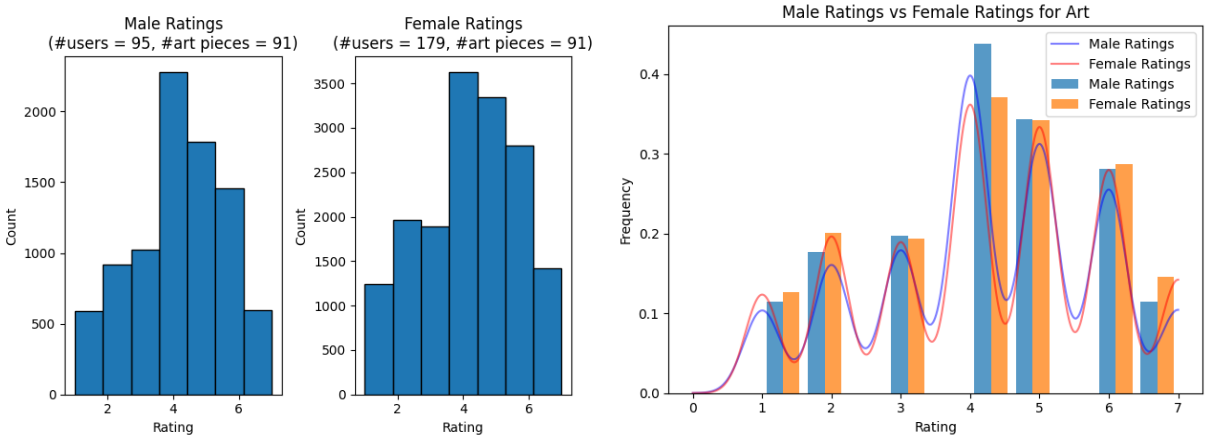
The p-value and u-statistic for the test were (***p-value =*** *8.742809791074804e-264,* ***u =*** *43486535.5*). At a significance level of *alpha=0.05*, we can reject the null hypothesis in favor of the alternative. Evidence suggests that the median rating for modern art is not the same as for non-human art for this set of responses.

There is a difference in the preference ratings for modern art vs. non-human (animals and computers) generated art.

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

To answer this question, I used the rating dataset only. I again had to select two sets of indices but, in column 216` in the user_ratings dataset with values 1 and 2, corresponding to male and female ratings, respectively. I did not need to do the join method mentioned in the intro section, as the male and female data rows differed. I selected the user ratings for males and females from these indices, then dropped na values.

The visualization for the data is as follows.



From these data, I did two significance tests.

First, I tested to see if the data came from the same underlying distribution using the Kolmogorov-Smirnov test on the two sets of data with the following hypothesis:
H0: male and female art ratings come from the same underlying distribution.
H1: Male and Female art ratings do not come from the same underlying distribution.

The p-value and ks-statistic for the test were (***p-value =*** *0.008197202819428897,* ***ks =*** *0.022029719765679773*). At a significance level of *alpha=0.05*, we reject the null hypothesis in favor of the alternative. Evidence suggests that male ratings are distributed differently than female ratings for the art pieces in this set of responses. This makes sense because, in at least one case, the male ratings are higher than the female ratings, and in at least one case, the converse is also true.

Then, I used a Mann-Whitney U test on both sets of data. As mentioned previously, the median is more robust to outliers, so having a non-equal sample size should not matter here. The Hypotheses were as follows:

H0: The population median of male ratings for art is equal to the population median of female ratings for art,
H1: The population median of male ratings for art is not equal to the population median of female ratings for art.

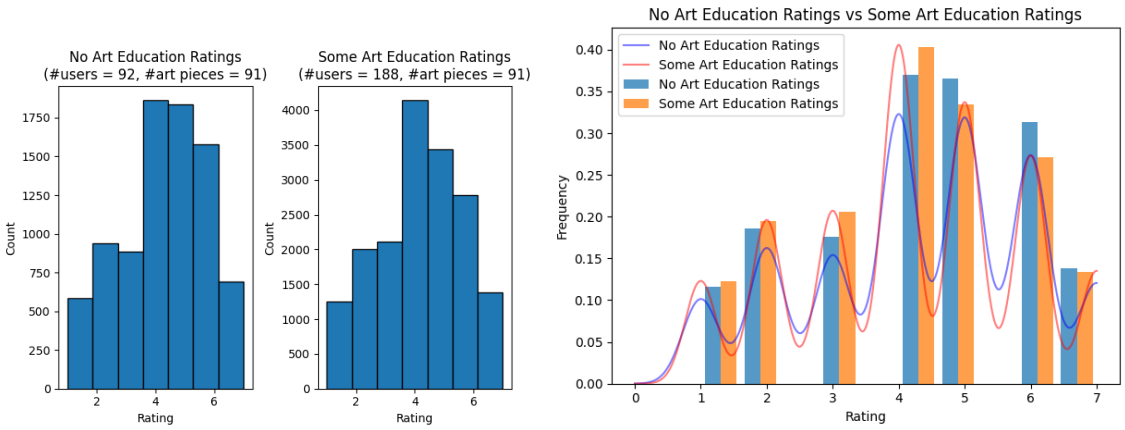
The p-value and u-statistic for the test were (***p-value*** **=** *0.13564554381330324,* ***u*** **=** *70994989.5*). At a significance level of *alpha=0.05*, we fail to reject the null hypothesis. These data suggest that it is not true that women give higher art preference ratings than men at *alpha=0.05*.

From these data, we can not conclude that women give higher preference ratings than men.

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

Similar to the last question, I used the rating dataset only. I again had to select two sets of indices but, in column `218` in the user_ratings dataset with values 0 and >0, corresponding to no-art-education and some-art-education, respectively. I did not need to do the join method mentioned in the intro section, as the rows for the education type were different. I selected the user ratings based on education level from these indices, then dropped na values.

The data are visualized as follows.



From these data, I did two significance tests.

First, I tested to see if the data came from the same underlying distribution using the Kolmogorov-Smirnov test on the two sets of data with the following hypothesis:

H0: no-art-education and some-art-education ratings come from the same underlying distribution.
H1: no-art-education and some-art-education ratings do not come from the same underlying distribution.

The p-value and ks-statistic for the test were (***p-value =*** *8.662710429908259e-11,* ***ks =*** *0.04603236726270954*). At a significance level of *alpha=0.05*, we reject the null hypothesis in favor of the alternative. Evidence suggests that no-art-education ratings are distributed differently than some-art-education ratings for the art pieces in this set of responses.

Then, I used a Mann-Whitney U test on both sets of data. As mentioned previously, the median is more robust to outliers, so having a non-equal sample size should not matter here. The Hypotheses were as follows:

H0: The population median of responses with no-art-education ratings for art is equal to the population median of responses with some-art-education ratings for art,
H1: The population median of responses with no-art-education ratings for art is not equal to the population median of responses with some-art-education ratings for art,

The p-value and u-statistic for the test were (***p-value =*** *1.0118570941459346e-08,* ***u =*** *74724020.0*). At a significance level of *alpha=0.05*, we can reject the null hypothesis in favor of the alternative. Evidence suggests that the median art rating for no-art-education is not the same as the median art rating for some-art-education for this set of responses. There is a difference in the preference ratings of users with some art background (some art education) vs. none.

From these data, we can conclude that the preference ratings of users with some art background (some art education) differ from those with no art background (no art education).

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.

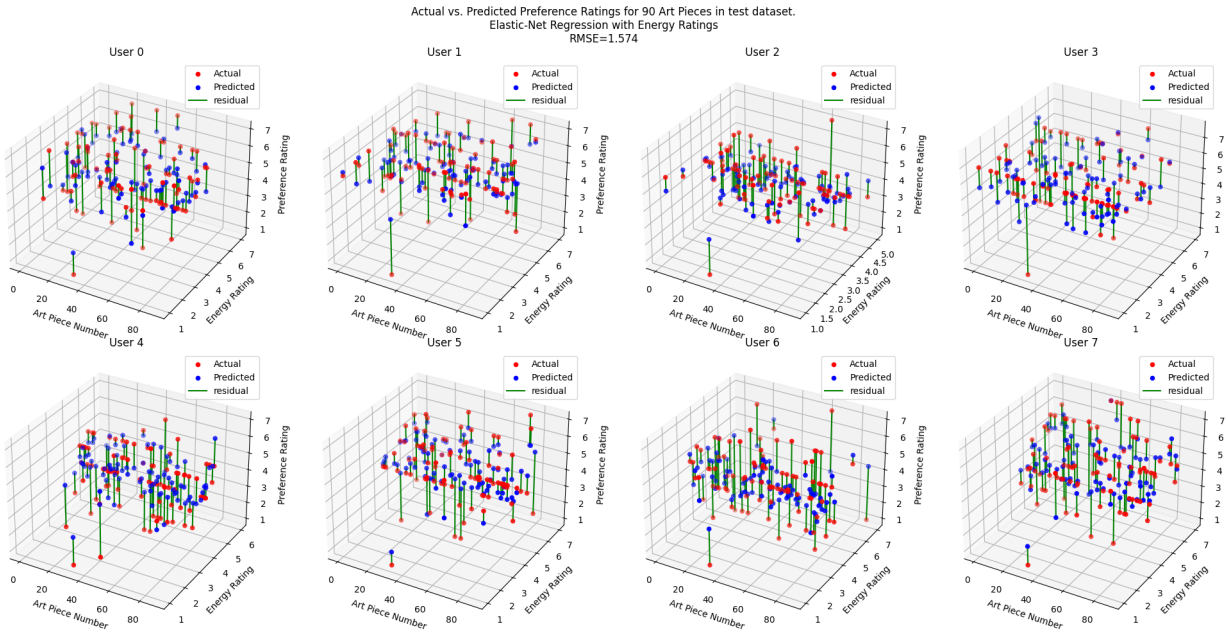
I came up with two approaches:

1. A regularized multivariate regression with an elastic-net model that combines the best of Ridge and Lasso regression. The input would be all of a user's energy ratings, and the output would be the predicted user's preference ratings for all 91 art pieces. I chose to do regularized regression due to collinearity concerns within the predictors.
2. Using another class of multi-regression with Gradient Boosting, which is more portable and has more features with sparse data. I trained this model on user responses in the training dataset to predict all of their art preference ratings from all their given energy ratings (i.e., a regression model for all art preference ratings, with the input as all of the energy ratings for all art pieces). This method is very compute-heavy but seems to work

well. We never learned gradient boosting for regression in class, so it was also a cool learning experience!

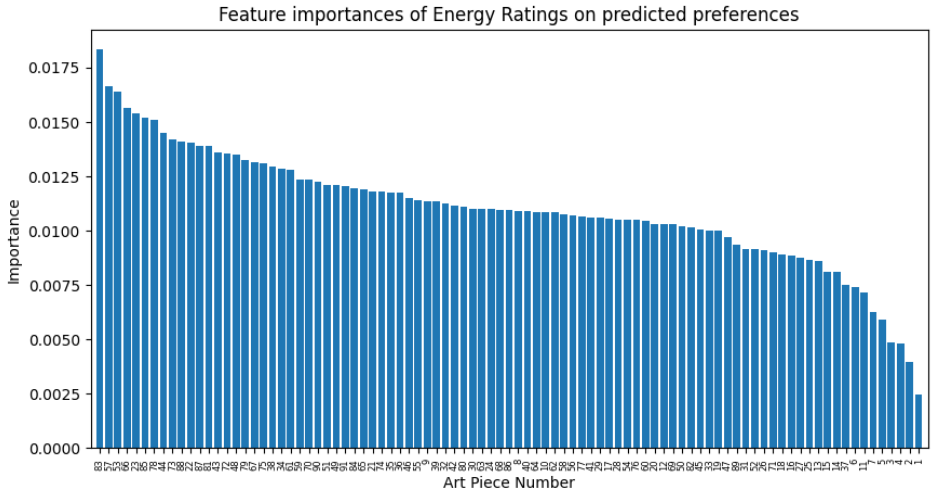
The first method has poorer performance compared to the second. For each of these methods, the data preprocessing is similar. I obtain both sets of ratings by selecting columns `0-181`, dropping na values, then I separate the resulting data frame into an energy rating data frame and a preference rating data frame. I obtain the values for each, then do a train_test_split using sklearn's model_selection module. The shape of my x_train, x_test, y_train, and y_test is ((210, 91), (90, 91), (210, 91), (90, 91)).

For the first method, I obtained the data mentioned above. From the training data, I fit a multivariate elastic-net regression model. The evaluation metrics are as follows:
RMSE: *1.573508151617326.* A visualization of the model's performance on the testing dataset can be seen as follows:

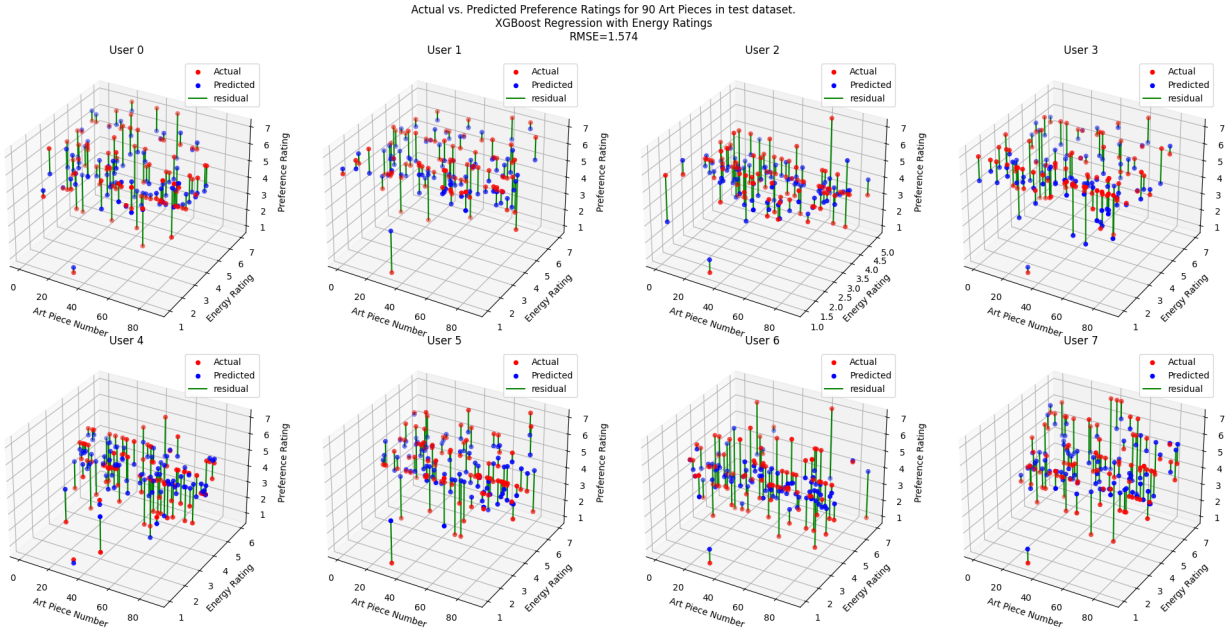


For the last method, I used Gradient Boosting regression for the reasons mentioned above. I used k-fold cross-validation on five splits. This was to reduce the variance of the model's estimated performance by running training and evaluating multiple subsets of the data. I did a regression to predict all art piece preferences. I generated metrics for the model:
Using sklearn.model_selection.cross_val_score, I generated a metric for the average k-fold cross-validated root mean squared error, resulting in
CV-RMSE: *1.5452142553315136*.

Feature Importance Visualization:



This was the most promising model. The performance on testing data is shown below:



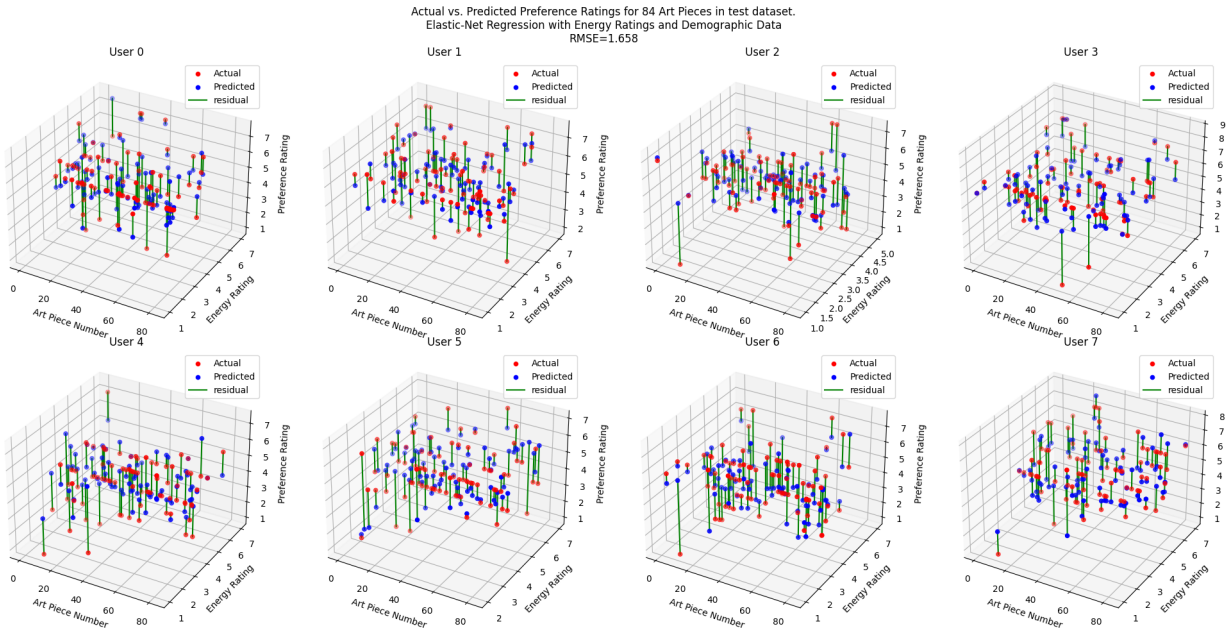
6. Build a regression model to predict art preference ratings from energy ratings and demographic information. Make sure to use cross-validation methods to avoid overfitting and comment on how well your model predicts relative to the “energy ratings only” model.

I came up with two methods to do a regression model

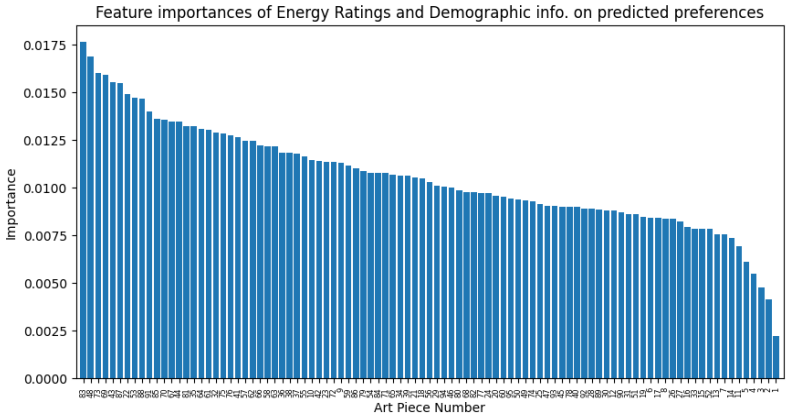
- A regularized multivariate regression with an elastic-net model that combines the best of Ridge and Lasso regression. The input would be all of a user's energy ratings combined with that user's age, gender, political orientation, and art education, and the output would be the predicted user's preference ratings for all 91 art pieces. I chose to do regularized regression due to collinearity concerns within the predictors.
- Use multi-regression with Gradient Boosting Regression, per user response in the training dataset, to predict their preference ratings, similar to the last question but with more parameters.

For each, in preprocessing, I select all columns from 0 to 218, then drop columns 182-214, and then drop any rows with na values. I then do a train_test_split using sklearn's model_selection module. The shape of my x_train, x_test, y_train, and y_test is((195, 95), (84, 95), (195, 91), (84, 91)).

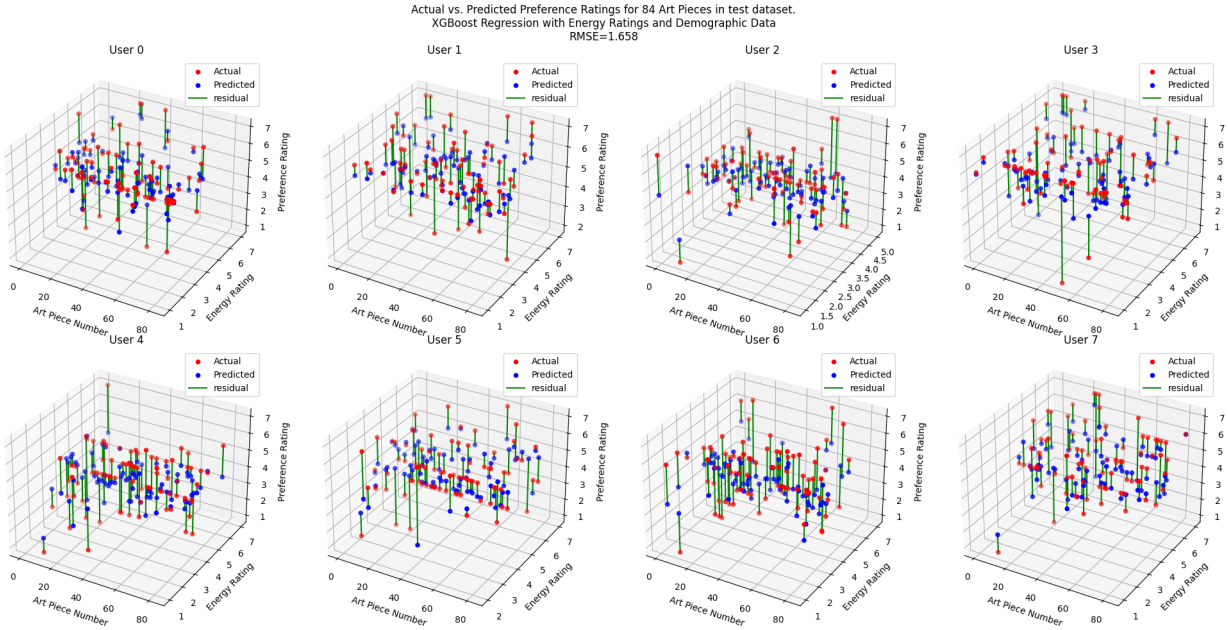
For the first method, I obtained the data, as mentioned above. I fit a multivariate elastic-net regression model on the training data with alpha=0.06. The evaluation metrics are as follows:
RMSE: *1.6576622894610127*.
The model somehow performed worse when accounting for demographic information. The results are visualized below:



I used Gradient Boosting regression again for the second method. The evaluation metrics were calculated the same way as in the previous question:
CV-RMSE: *1.5374840626535076*.
Feature Visualization:



This model also performed worse when accounting for demographic information. The results are visualized below:



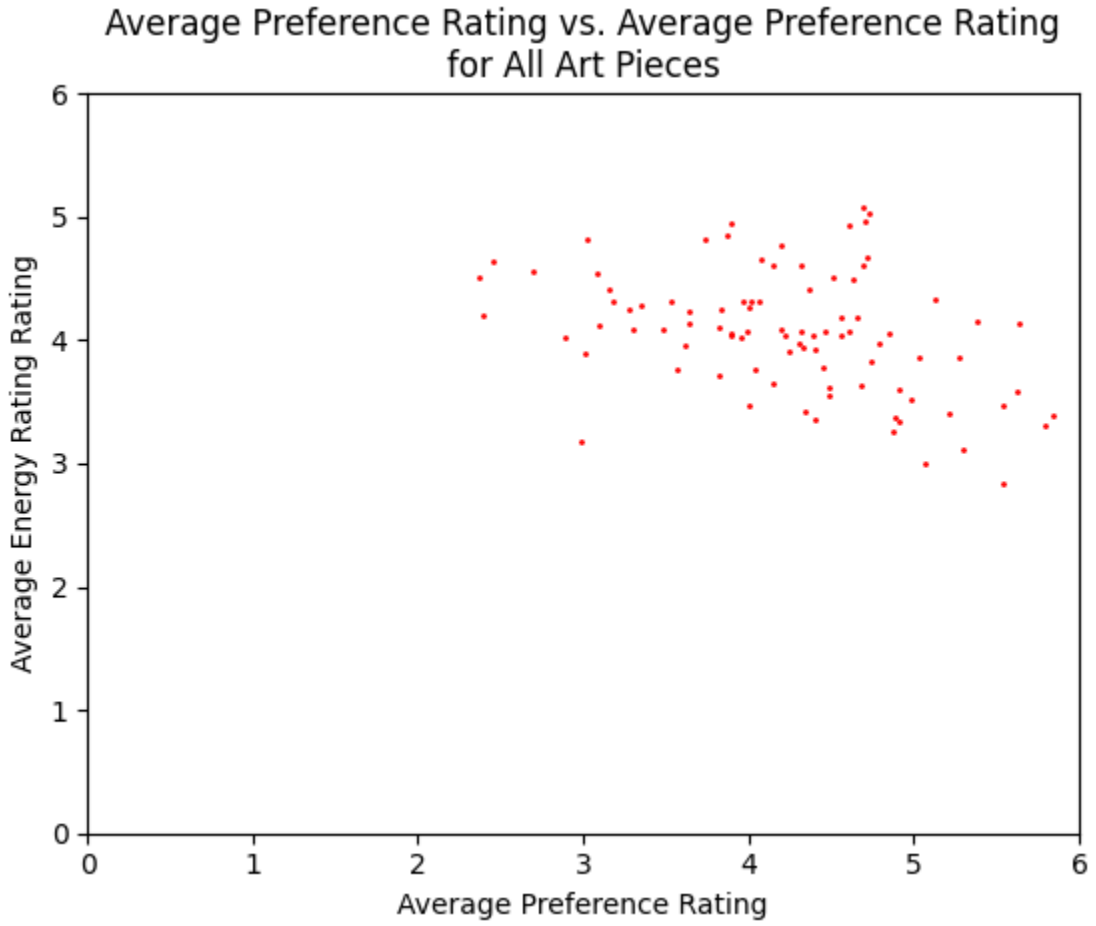
The RMSEs for each model are higher when accounting for demographic information, which makes sense. Intuitively, If an art piece is objectively bad, the difference in demographic information of a user would not be a good estimator of a preference rating.

The models here do not perform as well compared to the “energy ratings only” models.

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?

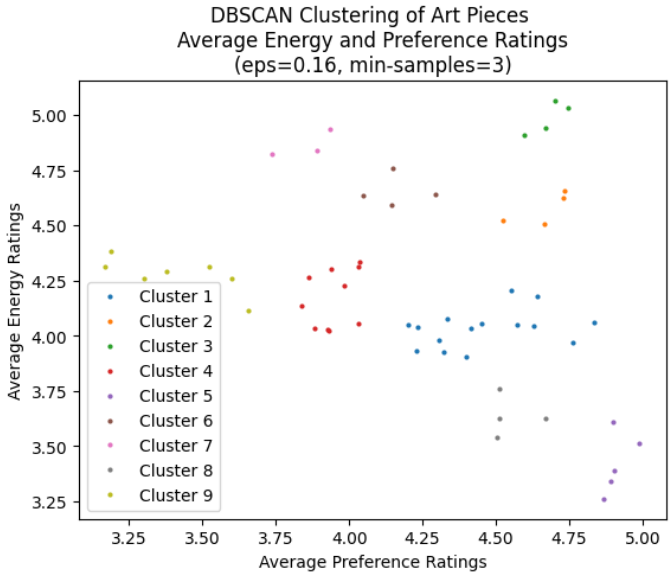
To answer this question, I had to do a similar kind of join, dropna, then a separate method as mentioned in the introduction, but with only the rating dataset. I obtained the columns corresponding to energy ratings and preference ratings, dropped the na values, then separated the data into two data frames corresponding to energy ratings and preference ratings.

From there, I calculated the average energy rating along each column and the average preference rating along each column. Plotting this, I saw an outlier with a low average energy rating and a low average preference rating, as outlined below.



I removed the data point in the lower left corner and then combined the average energy and average preference ratings as two columns, where rows represented a point. I then tried two clustering methods: DBSCAN and K-Means.

I ran DBSCAN with the following parameters: (epsilon=0.16, min-samples=3) resulting in 9 clusters as follows:

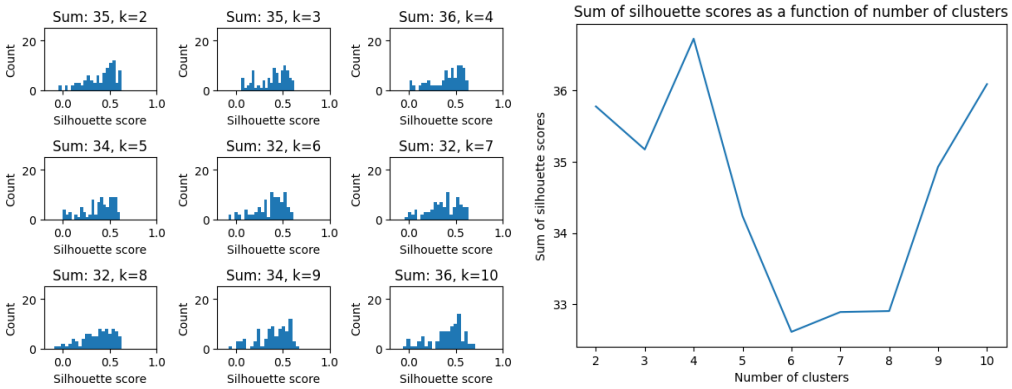


I visually characterized each cluster as follows

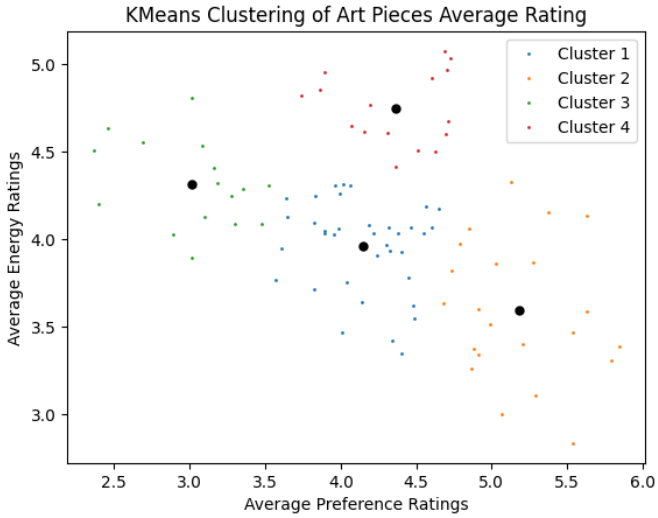
Cluster Number	Art pieces	Average Preference Rating Rank (lower rating -> lower rank)	Average Energy Preference Rating Rank (lower rating -> lower rank)	Visual Classification of Cluster Identity
1	5,7,20,21,22,29,41,42,45,53,55,57,61,67,83	4	7	Paintings with decent energy rating but with high preference. Some self-portraits and abstract art with some structure.
2	6,18,46,49	3	4	Higher energy with higher preference. Again some self-portraits and abstract art. Well-known paintings
3	8,9,10,47	2	1	Historical art with very high energy ratings along with one abstract art piece.
4	24,28,35,37,43,48,54,64,66,89	7	6	Mainly abstract art without much contrast in the piece. There are a few historical paintings with many actors for some reason.
5	27,33,34,40,88	1	9	Art pieces depicting some historical events but with very little energy, and also one abstract art piece with little energy.
6	36,44,56,58	6	3	General abstract art with a lot of energy.
7	60,68,73	8	2	Abstract art with a lot of energy and contrast but low preference
8	13,25,62,63	5	8	Art with low energy but high preference. It generally uses contrast pretty well and has some kind of centering in the piece.
9	69,72,74,79,82,85,87	9	5	Abstract art with various preference ratings (generally low) and average energy ratings around 4.25.

However, DBSCAN required hyperparameter-tuning the epsilon size, and boundary ambiguity was a problem. A lot of data was also removed from clustering. Following this, I tried clustering with K-Means to find the optimal number of centroids.

I used the silhouette method to identify the number of clusters algorithmically. Below is the silhouette plot of k-Means at varying numbers of clusters, as well as the sum of all the silhouette scores. The optimal number of clusters without overfitting according to this method is four.



The resulting clustering is shown below:



From these clusters, we can identify the identity of each as follows:

Cluster #	Art Pieces	Avg Preference Rating Rank	Avg Energy Preference Rating	Visual Classification of Cluster Identity
1	3,5,7,13,17,20,21,22,23,24,28,30,35,37,38,39,41,42,43,45,48,53,54,55,57,62,63,64,65,66,67,69,79,83,89,90	3	3	Medium preference and medium energy (mainly abstract art without too much energy)

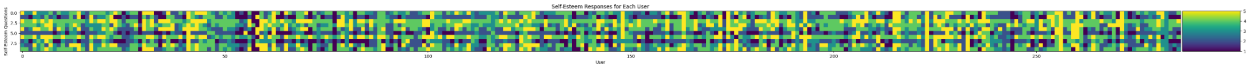
				Modern art
2	1,2,4,11,14,15,16,19,25,26,27,29,31,32,33,34,40,50,51,59,61,88	1	4	High preference with varying levels of energy (generally high). Usually famous historic art and some abstract art. Classical art
3	52,70,71,72,74,75,76,77,78,80,81,82,84,85,86,87	4	2	Low preference ratings with generally high energy ratings. Mainly abstract art with very high energy. Machine generated art
4	6,8,9,10,12,18,36,44,46,47,49,56,58,60,68,73	2	1	Medium-high preference rating with very high energy ratings. Famous historic images with high energy and some abstract art with high energy. Modern art,

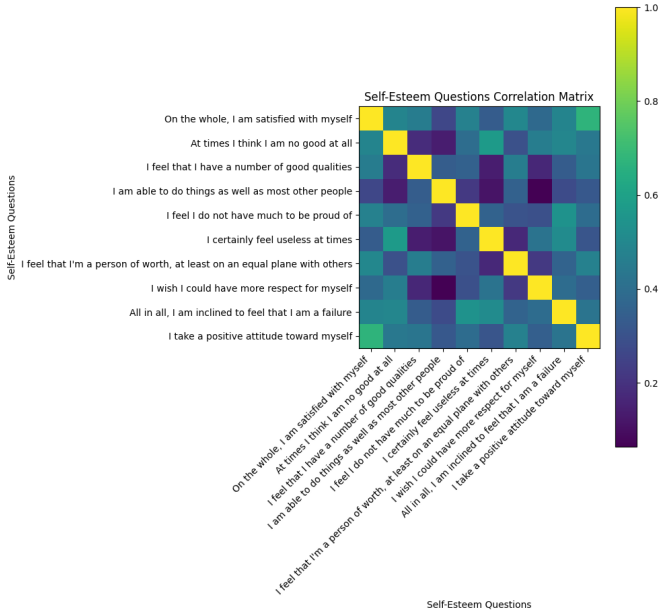
However, I can't say I agree with this because there aren't any distinctively separable or clustered regions of data. This is because we are clustering over the average, which may not represent an art piece's true rating. The metric is more prone to outliers in the data. Generally, it should not work because the `psychological` distance between two ratings may not be the same, even if there is some ordering relation between ratings.

Nevertheless, I can algorithmically identify 4 clusters in this space. Each cluster seems to correspond to types of art as outlined above.

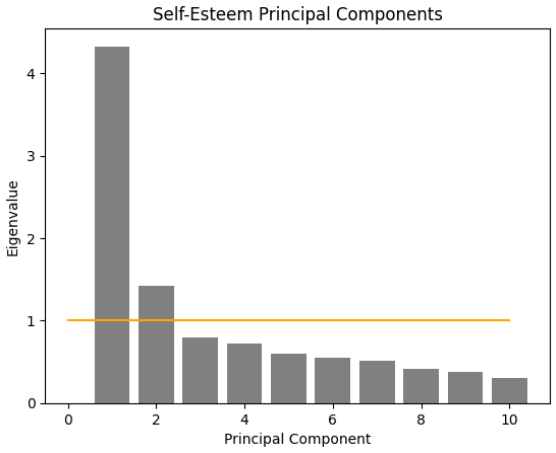
8. Considering only the first principal component of the self-image ratings as inputs to a regression model, how well can you predict art preference ratings from that factor alone?

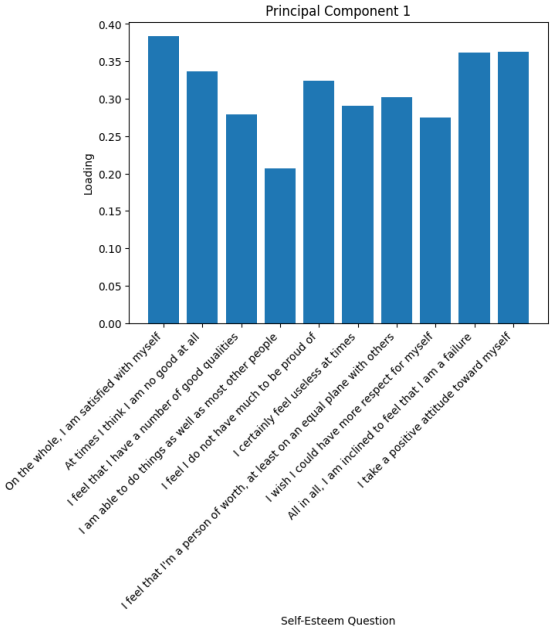
To answer this question, I did a similar method as the previous questions, selecting columns 0-214, dropping the columns 91-204, then dropping na rows on the remaining columns. I chose not to impute data as there's no good baseline for self-esteem. Even if there are rows with only 1 column of data missing and not the entire self-esteem portion, I did not know of a good imputation value, so I removed all such rows. From there, I created a preference rating and esteem response data frame. The image and correlation matrix for esteem responses are shown below:



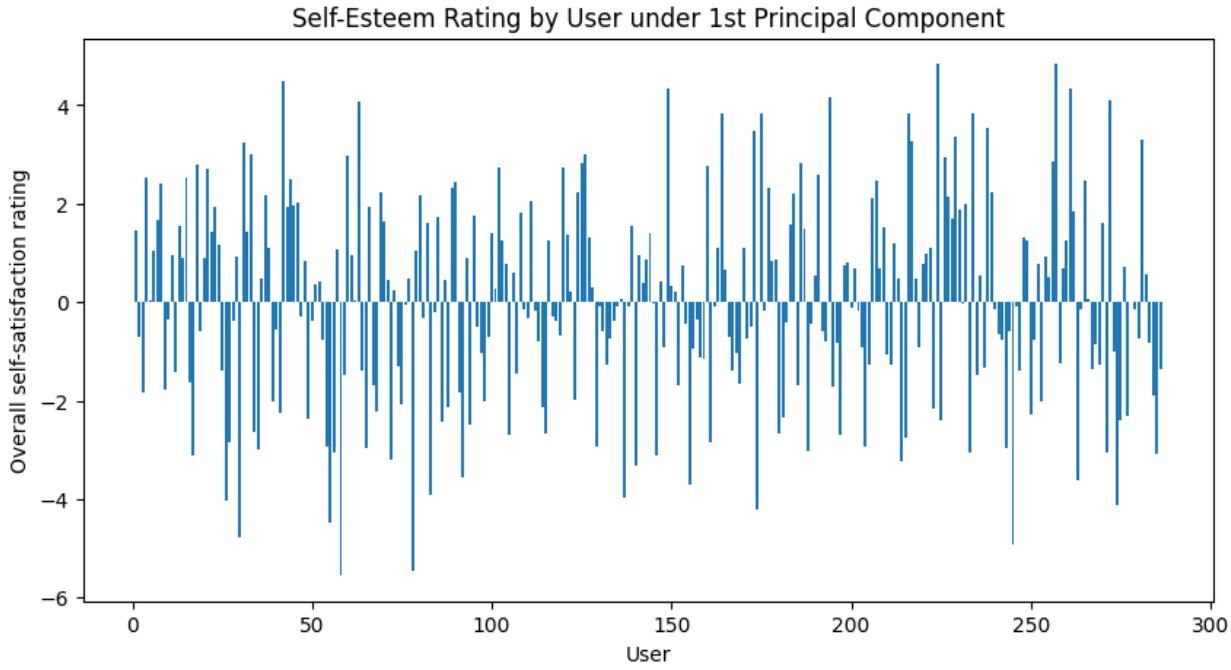


Some data are correlated with each other. To generate the principal components, I normalized the esteem responses by z-score and did a PCA on the normalized data. The 1st principal component explains 43.14% of the variance in self-esteem questions. The scree plot and loadings are listed below.





The first principal component essentially refers to a user's overall self-satisfaction. The actionable space of the transformed data is shown below.



I decided to use a linear regression that uses the 1st principal component of a user's self-esteem to predict the predicted user's preference ratings for all 91 art pieces.

For data preprocessing, I obtain the 1st principal component of the esteem responses, then do a train_test_split using sklearn's model_selection module. After splitting the data into training and testing, the shape of my x_train, x_test, y_train, and y_test is ((200, 1), (86, 1), (200, 91), (86, 91)).

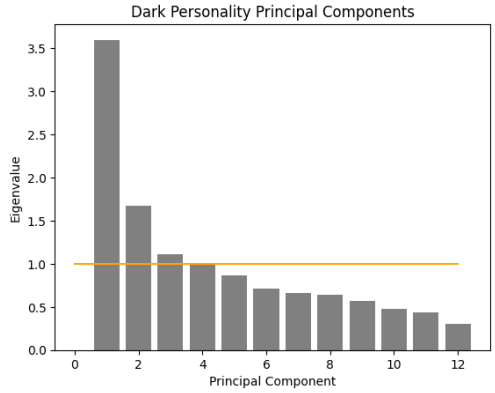
For the linear model, The evaluation metrics are as follows:
RMSE: *1.4657848702799772,* ***R^2:*** *-0.013851959977821855.* . Since R^2 is negative, the mean of the preference ratings provides a better estimate of the outcomes than the linear model.

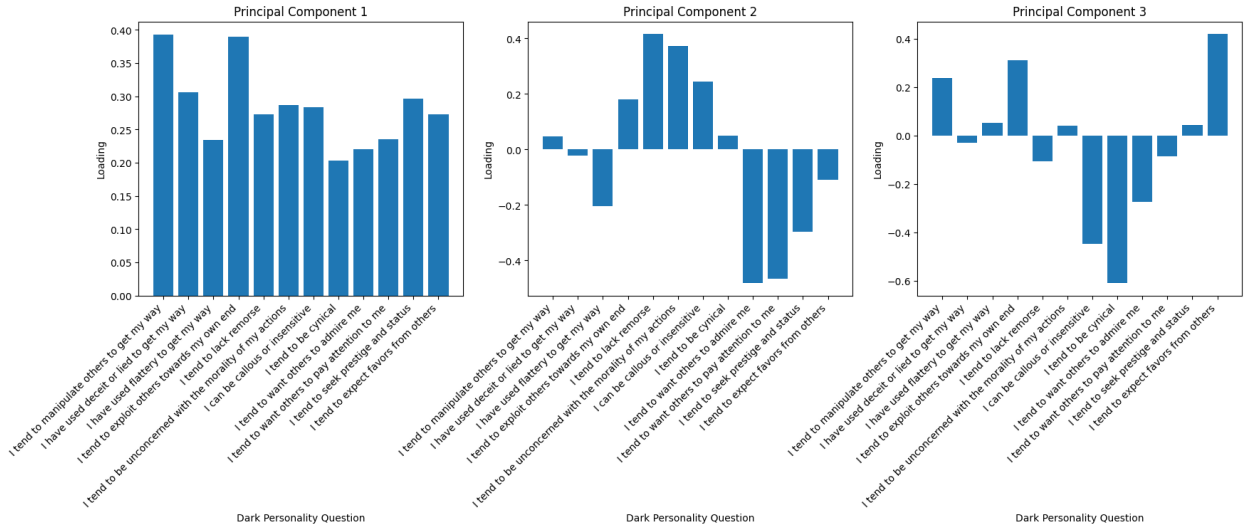
Therefore, I cannot predict art preference ratings well from this factor alone.

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.).

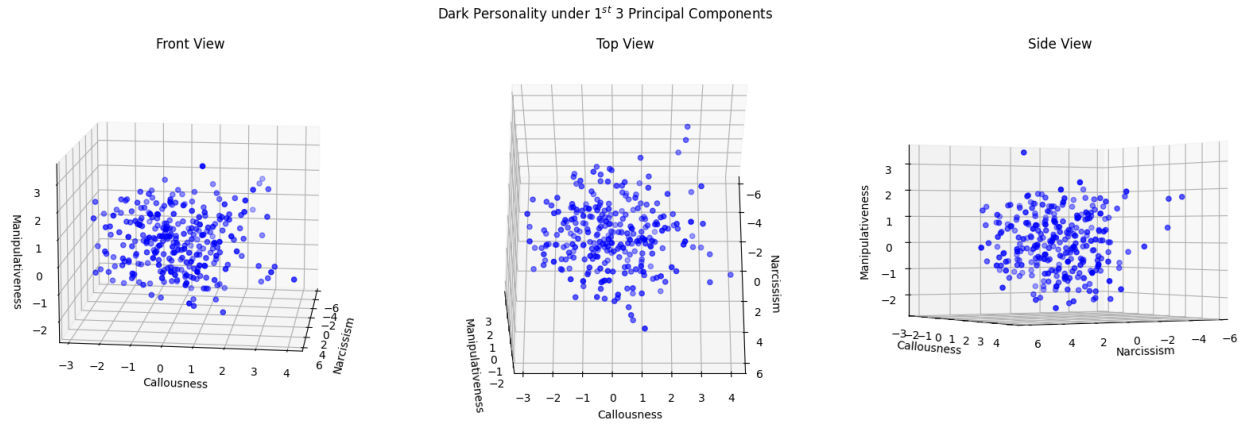
I Selected columns `0-193`, dropped na rows, then selected preference ratings and dark responses from the respective columns. I also chose not to impute data and dropped rows with Na values instead. From there, I created a preference rating and dark-personality-question-response dataframe. The responses image and correlation matrix are shown below:

There are some data correlated with each other. To generate the principal components. I normalized the dark-response questions by z-score and did a PCA on the normalized data. The 1st, 2nd, and 3rd principal components explain 29.87%, 13.90%, and 9.23% of the variance, respectively. The scree plot and loadings are listed below.



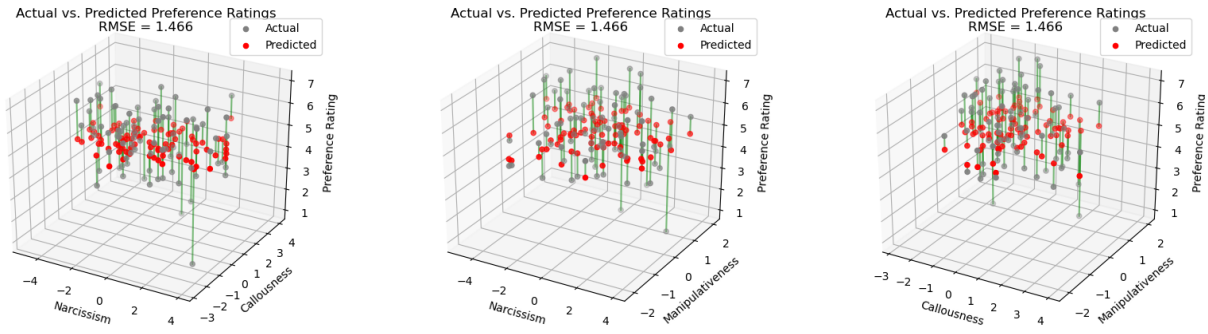


The first principal component refers to a user's degree of narcissism, having an inflated sense of self-importance. The second principal component refers to a user's Callousness/ Overall insensitivity, not understanding or caring about the feelings of others. The third principal component refers to a user's degree of manipulativeness/Opportunism, being deceitful and manipulative. From these 3 components, the actionable space is visualized below.



From these, in my pre-processing, I obtain the 1st three principal components and perform a column stack with them, while also collecting the preference ratings for all art pieces per user response. I do a train_test_split using sklearn's model_selection module. The shape of my x_train, x_test, y_train, and y_test is ((198, 3), (86, 3), (198, 91), (86, 91)).

For the regression, I chose to use a multivariate elastic-net regularized regression model on the training data with an alpha=0.01. The evaluation metric is as follows:
RMSE: *1.4663798706418214*.
This has a lower RMSE than any previous regression done in this paper. The results are visualized below for all user's responses to the 1st art piece (chosen arbitrarily).



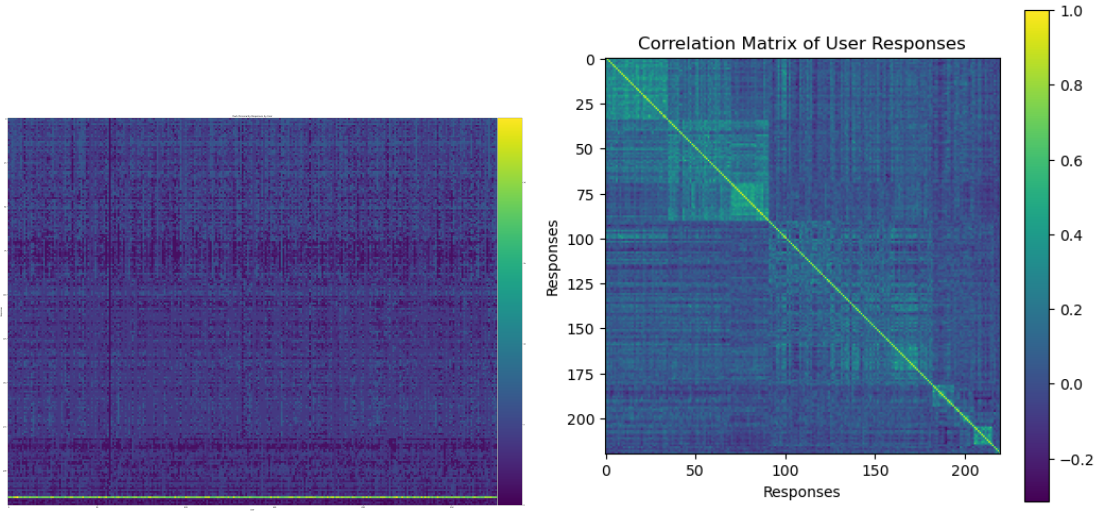
To determine which principal component significantly predicts art preference ratings, I performed a separate elastic-net regression for each. I did a test_train split for each, with the predictor as either narcissism, callousness, or manipulativeness. The shape of my x_train, x_test, y_train, and y_test was ((198, 1), (86, 1), (198, 91), (86, 91)). The RMSEs for each are listed below.

Principal Component	RMSE	R^2	Explained Variance Score
1 - Narcissism	1.4708634210613674	-0.013473982921138059	-0.00019398143260736627
2 - callousness	1.4693293502957276	-0.013473982921138059	-0.00019398143260736627
3 - manipulativeness	1.4683653790919957	-0.008069225911974252	0.004314946382436067

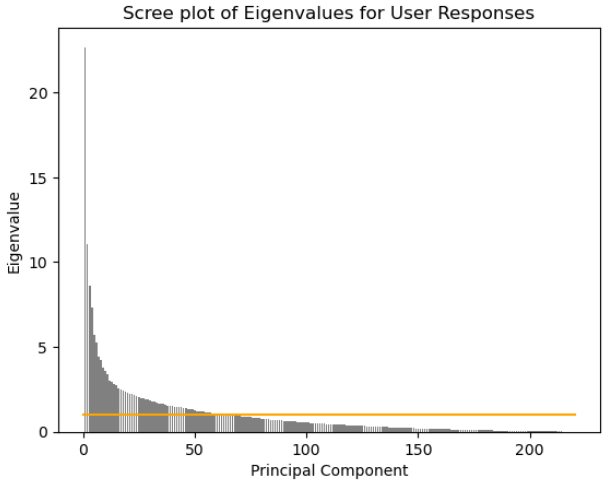
To determine which principal component significantly predicts art preference ratings, I looked at the summary statistics of each model. The manipulativeness model is the only one with a positive explained variance score and the least negative R^2 score. I did not prove “statistical” significance, but these factors suggest that manipulativeness, the third principal component, significantly predicts art preference ratings compared to the 1st two.

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.

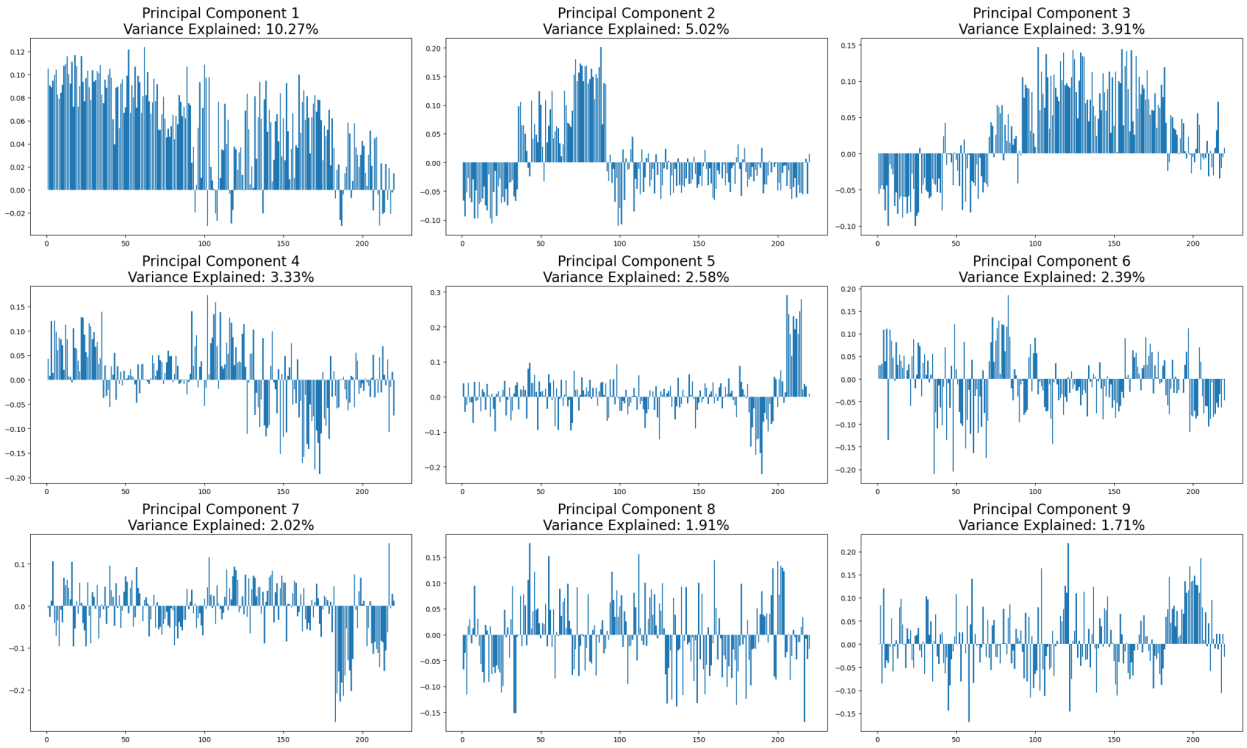
To answer this question, I had to select all of the user response columns, dropna values, then separate them into the columns we would do PCA on (not column 217) and the ground truth (column 217) that was re-codedto 0 for “left” and 1 for “non-left.” From here, I had to do a Principal Component Analysis on the predictors. An image of the responses and correlation matrix are listed below.



We can clearly see that some responses are correlated. From here, I plotted the scree plot and selected the number of principal components (64) by the kaiser criterion.

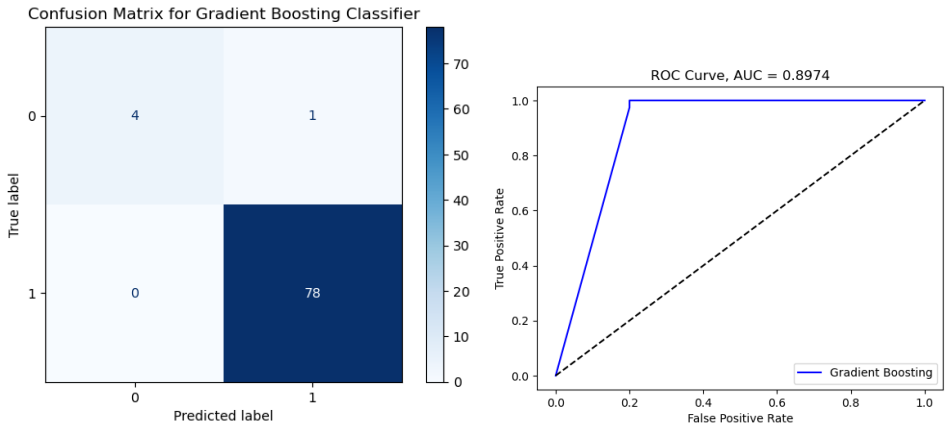


The first 64 principal components cumulatively account for 77.08% of the variance in the user responses. A plot of the first 9 loadings is listed below.



After selecting the 64 principal components, I do a train_test_split using sklearn's model_selection module. The shape of my x_train, x_test, y_train, and y_test is (193, 220) (83, 220) (193, 1) (83, 1). From here, I trained a Gradient Boosting classification model. Gradient boosting trains a sequence of weak learners to form a strong learner. I chose this method because random forests, while less prone to overfitting, do not perform well on imbalanced datasets, and logistic regression was very simplistic. I trained this using StratifiedKFold to help with the issues of class imbalance, by keeping the same proportion of TN and TP data in each fold.

The correlation matrix and AUC curve for the training data are shown below.



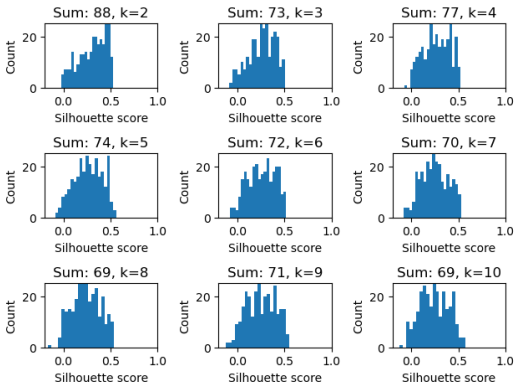
I also generated the following metrics:
Precision: *0.9873417721518988*
Recall: *1.0*
F1: *0.9936305732484078*

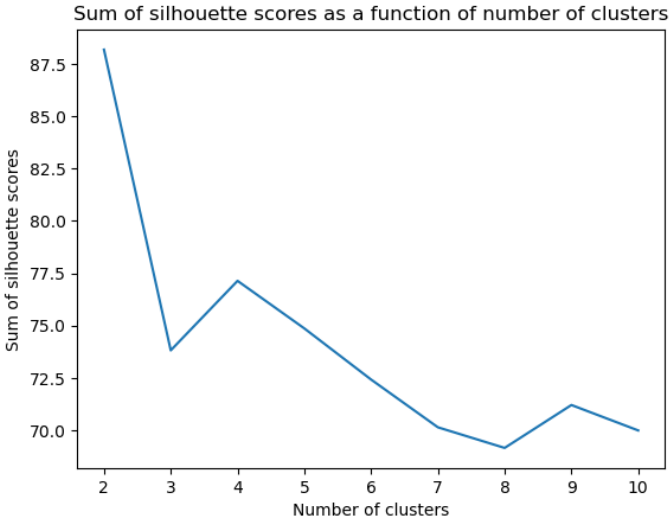
I can classify pretty well on testing data, but because of class imbalance issues, such as having only 19 “left” labels and 257 “non-left” labels, I am hesitant to say I can determine the total political orientation of the users from all other information available. The classification quality of the model is pretty good, considering the AUC.

11. Extra credit: Tell us something interesting about this dataset that is not trivial and not already part of an answer (implied or explicitly) to these enumerated questions.

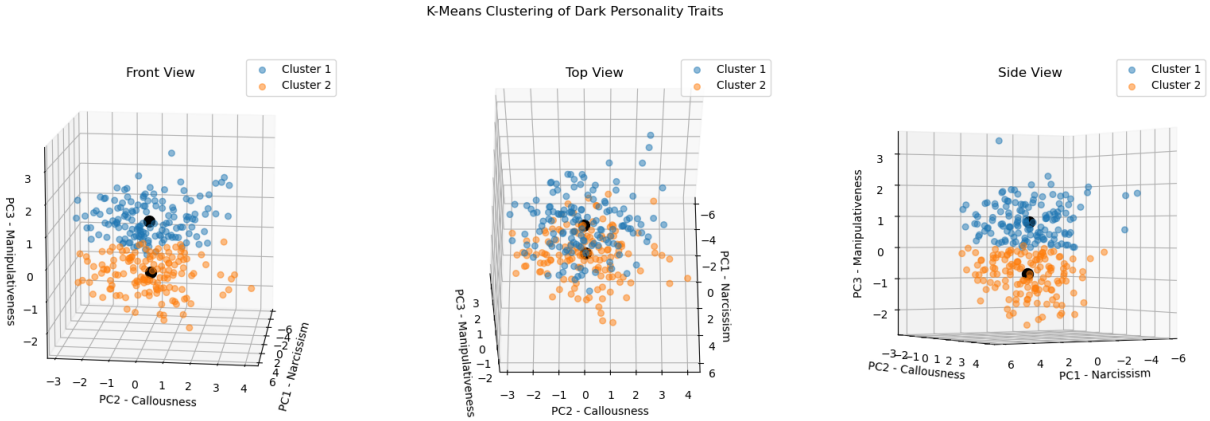
How many clusters can be identified in the first three principal components of the “dark personality traits”?

The data of the principal components were loaded, as stated in question 9. I used the silhouette method to identify the number of clusters algorithmically. Below is the silhouette plot of k-Means at varying numbers of clusters, and the sum of all silhouette scores. According to the model, the optimal number of clusters without overfitting is two.





The resultingf clustering is shown below:



It was surprising to see that the cluster boundaries are directly along PC3, not on PC1 or PC2. I expected the decision boundary to be non-coplanar with two principal components, which was surprising.