Darien Nouri

Big Data Analysis Project

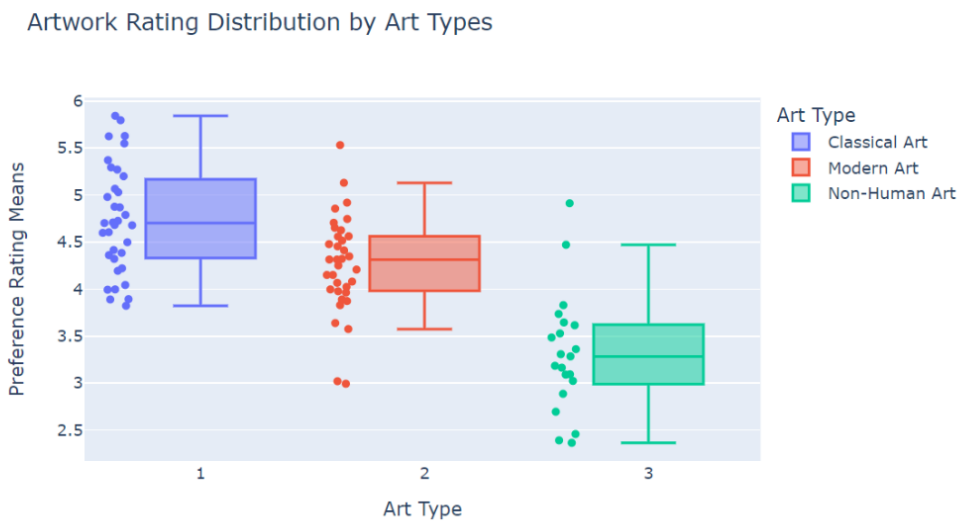
As part of the data preprocessing, I extracted the energy ratings and preference ratings from the dataset. I then used Ordinary Least Squares (OLS) Regression to examine the relationship between these two variables. The dataset contains 25 rows that contain at least one null value. Since not all columns will. To evaluate the performance of the initial model, I calculated several validation scores, including R2 (a measure of the proportion of total variation of outcomes explained by the model) and Mean Absolute Error (MAE; a measure of the average absolute difference between the observed and predicted results). The cross-validation scores for the initial model showed a consistent MAE across all K-fold splits, indicating that the model was well-trained.

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

To examine if classical art is more well-liked than modern art, I calculated the p-value with an alpha level of .05 using the Mann-Whitney U test. The methods of selection for a test take into consideration that the rating data in this set is ordinal. The calculations resulted in a u-score of 891.50 with a probability value of (P<.001). The null is rejected here in the observation of the preference rating distribution box plot in fig1. The box plot below depicts the results of this significance test reinforced as preference ratings for classical art are greater than that of modern art. Furthermore, preference ratios of classical art are more evenly normally distributed than that of modern art.

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

With the initial observation of the box plot distributions in fig1, preference ratings for non-human art are lower than that for classical and modern art. The delta between non-human art is quite marginal, thus being an early indication of a significant difference in the level and distribution of preference ratings of modern and non-human art. Statistical analysis with the Mann-Whitney U test reinforces this claim with a significance level of (P<.001).



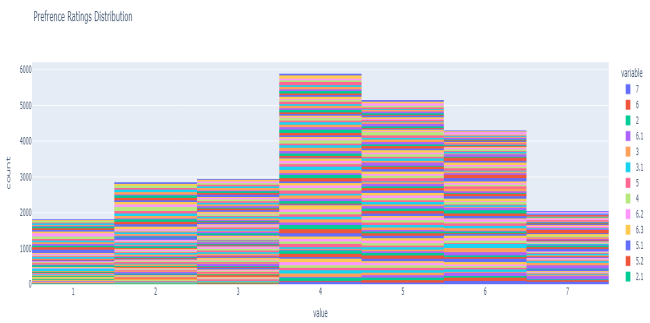
3. Do women give higher art preference ratings than men?

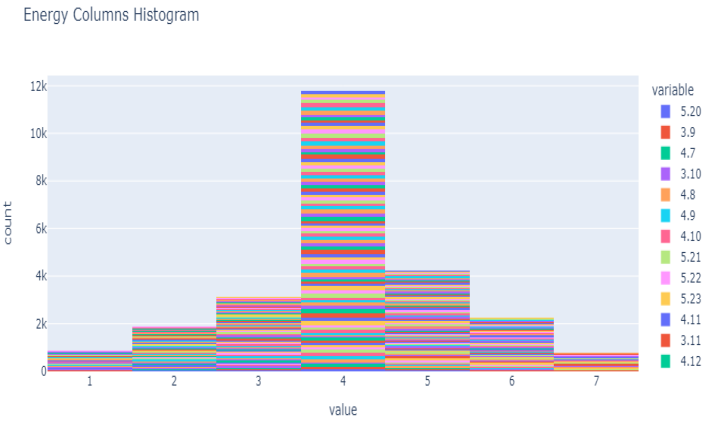
Before any calculations are undergone, it is important to observe the summary statistics with respect to men and women. Within this dataset, 98 observations of ratings were by women, and 178 observations of ratings were submitted by men. Regarding the mean, the difference between the ratings of men and women is minimal, with a 4.21 average for women and a 4.23 average for men. Testing with the Mann-Whitney U test yields (P=.44), which further reveals a likely lack of difference between the preference ratings between men's and women's significance levels.

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

The average preference rating for users with some art background is 4.19. In contrast, the average preference rating of users without an art background is 4.31, leaving a difference of 0.12 between the average preference ratings of those with some art background and those with none.

	0	u-score	p-value	Group 1 Average	Group 2 Average
0	Classic v. Modern	891.50	0.0005350258162413517	4.74	4.26
1	Modern v. Non-Human	644.50	1.4358021617813987e-06	4.26	3.31
2	Men v. Women	4192.00	0.4429416449781828	4.23	4.21
3	With Art v. Without Art Edu	3761.00	0.8575406189479926	4.19	4.31





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

The methodology behind this process was designed to assess many variations of data and model configurations. Model Evaluation functions were designed to minimize repetitive tasks, including test-train-split configuration, full model training, methods of cross-validation, and model analysis visualizations. A standard Linear Regression model was selected over Regression models such as Logistic, Ridge, and Lasso as the input data, energy, and rating scores, are of Ordinal type, of the same scale of rank 1-7. With the relatively small size of the input data, features <100, and observation<300, K-Fold cross-validation with a (n_split=5) is selected to avoid overfitting the model. With the K-Fold approach, a subset of the input data is first selected, then the selected data is further segmented using a test-train split. The model is then trained, and various validation scores are calculated and appended to a list. This process is repeated 4 more times across the remaining K-Fold data subsets. Once all models have been trained, the stored lists to subsets of the data are iterated. Through the pipeline, it iterates by splitting the data into training, and testing segments with training Linear Regression models The pipeline begins with variations of independent and dependent variable structure data.

full model, the creation of test-train-split, and the complexity of code. Firstly, as part of data pre-processing, energy and preference ratings were extracted from the dataset.
To investigate the relationship between energy scores and rating scores, I performed an Ordinary Least Squares (OLS) Regression. The dataset includes energy scores and preference ratings of all participants across the 91 preference ratings. With no nan values among the initial model's dataset, no additional data pre-processing was needed. Cross-validation scores of the initial

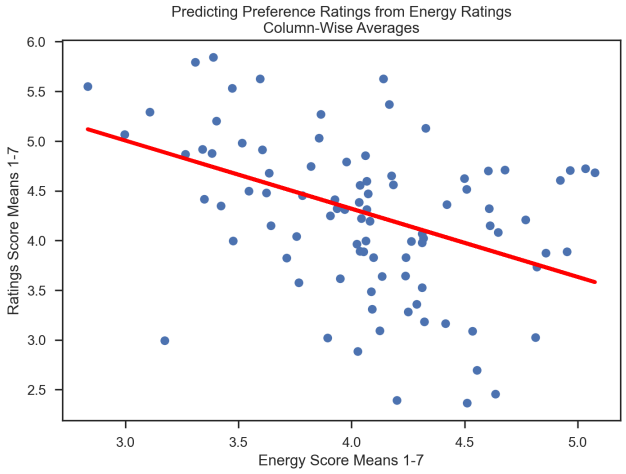
model demonstrate a steady MAE across all K-fold splits, signaling a well-trained model. Model Error was measured with metrics including the sum of residuals squared R^2 and Mean Absolute Error (MAE). Both metrics are variations in calculating the difference between expected and predicted values. In this linear regression model, R^2 is a measurement of how well the model predicts the actual observed outcomes, based on the proportion of total variation of outcomes explained by the model. The Mean Absolute Error (MAE) calculates the average absolute difference between the actual observed results and the predicted results. The Mean Absolute Error remains within the scale of the dataset, making it especially useful in ordinal datasets when scaling remains unchanged before and during modeling.

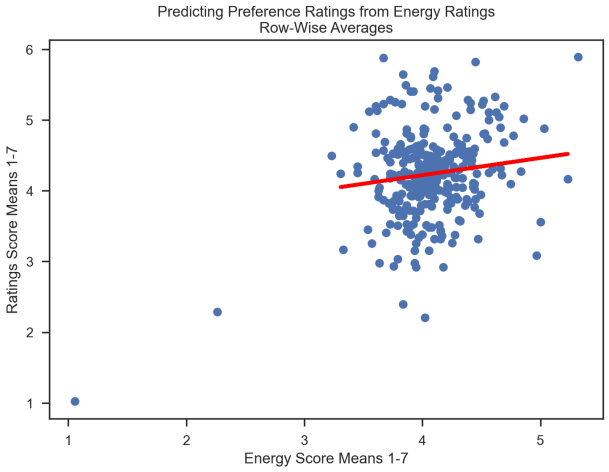
In this model, there is an average error of 1.473, meaning the model predicted artwork preference ratings on average 1.473 away from their true values, which could have been anywhere between 1-7. On the other hand, the R^2 metric for this model was negative -0.703. A negative R^2 value may either allude to inaccuracies in creating the model or when the mean of actual values is a better fit than the model itself. Therefore I created another model to cross reference for predicting the average preference ratings by training the model with the average energy scores given by each participant and the average ratings by each participant per participant from the average energy scores of the same individuals. The second model yielded an average MAE = 0.458 and a positive $R^2 = 0.1$. Although yielding considerably better results than the first model, the decrease in error variance can be explained by the distribution of the new dataset, excluding much of the variance compared
to the original model.

	test_r2	test_neg_mean_absolute_error	test_neg_mean_squared_error	test_explained_variance
0	-0.73	-1.48	-3.80	-0.70
1	-0.58	-1.56	-3.90	-0.55
2	-0.74	-1.40	-3.35	-0.70
3	-0.56	-1.40	-3.10	-0.54
4	-0.77	-1.49	-3.71	-0.73

```
Mean  R2                   -0.68
Mean -MAE                  -1.47
Mean -MSE                  -3.57
Mean  Explained Variance   -0.64
```

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

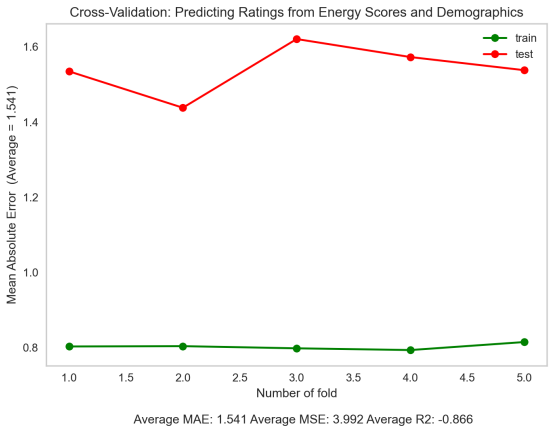




First, the column, including demographic data, is merged with the energy and rating columns. In dealing with nan values, demographic columns contain 20 rows of nan values, which are dropped from the DataFrame. Although a smaller dataset, I chose to drop nan rows rather than any imputation as there is already an increase in data included with the use of K-fold cross-validation as well as inaccuracies that may stem from the inclusion of an unproportionate amount of imputed data on a smaller K-fold test segment. I then fit a Linear Regression Model and evaluate the returned metrics. The table below displays a set of test metrics for each cross-validation model as well as the averages of the metrics across the set of models. After adding the columns containing demographics data, the model resulted in performance scores worse than just energy scores as the predictor. This model scored an MAE of -1.53, whereas the energy ratings only model's MAE was -1.47. In relation to only using energy scores as predictors, this model yielded an *R*2 value of -0.82. Which is 10% lower. Both of the previous models had a negative mean explained variance as well. Therefore, the data poorly fits the model. Better predictions would be made by simply predicting the mean of the preference ratings.

	test_r2	test_neg_mean_absolute_error	test_neg_mean_squared_error	test_explained_variance
0	-0.79	-1.49	-3.69	-0.72
1	-0.83	-1.50	-3.70	-0.80
2	-0.58	-1.40	-3.27	-0.55
3	-1.07	-1.64	-4.60	-1.05
4	-0.83	-1.63	-4.34	-0.75

```
Mean  R2                    -0.82
Mean -MAE                   -1.53
Mean -MSE                   -3.92
Mean  Explained Variance    -0.77
```

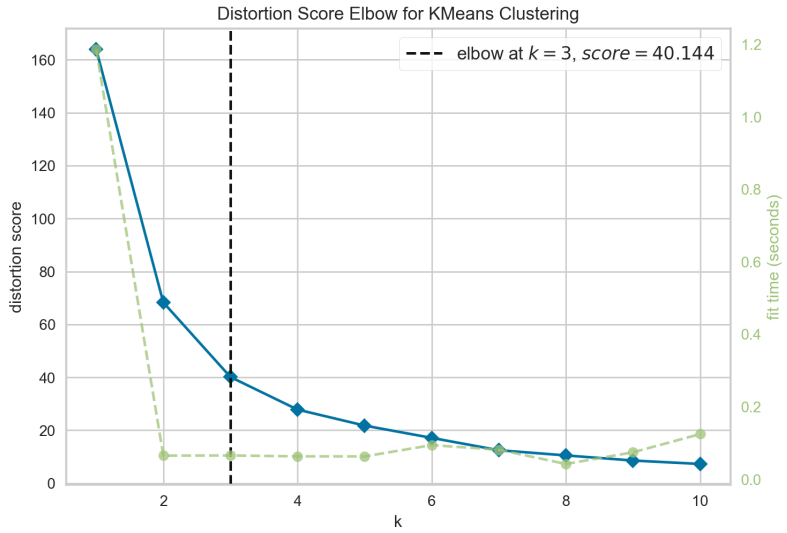


7. Considering the 2D space of average preference ratings vs. average energy rating (that

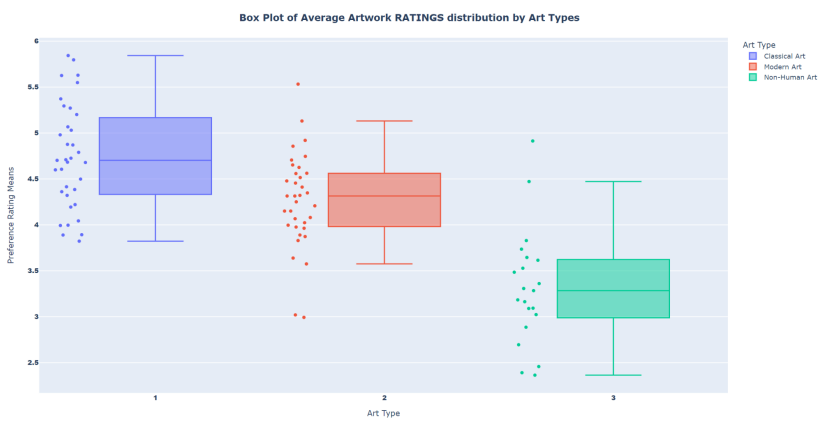
contains the 91 art pieces as elements), how many clusters can you – algorithmically - identify
in this space? Make sure to comment on the identity of the clusters – do they correspond to particular types of art?

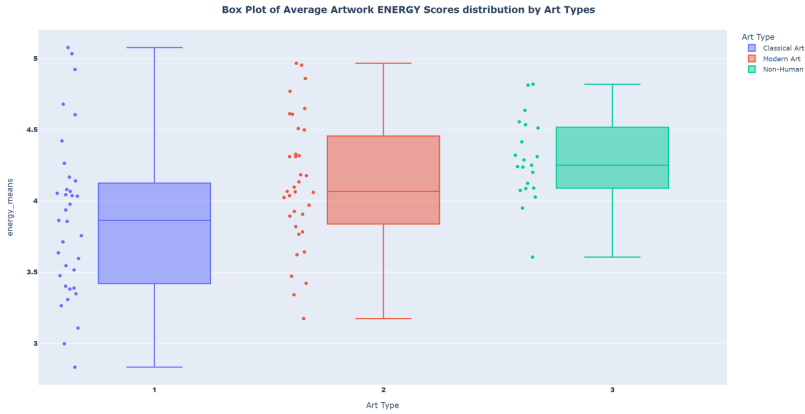
Firstly, average preference and energy ratings were calculated as the mean rating given to each artwork. In the data pre-preprocessing for this question, the 'energy_means',' rating_means', and 'actual_art_type' columns are merged into a container DataFrame. Art Types are classified as follows: {1: Classical Art, 2: Modern Art, 3: Non-Human Art}. The lack of identifiers or insight into the 2D space of average preference and energy ratings found a K-Means algorithm to be an appropriate model given the unsupervised nature of this dataset. The elbow method was then used to identify an ideal number of K-clusters = 3.

The model is trained with K = 3 clusters.

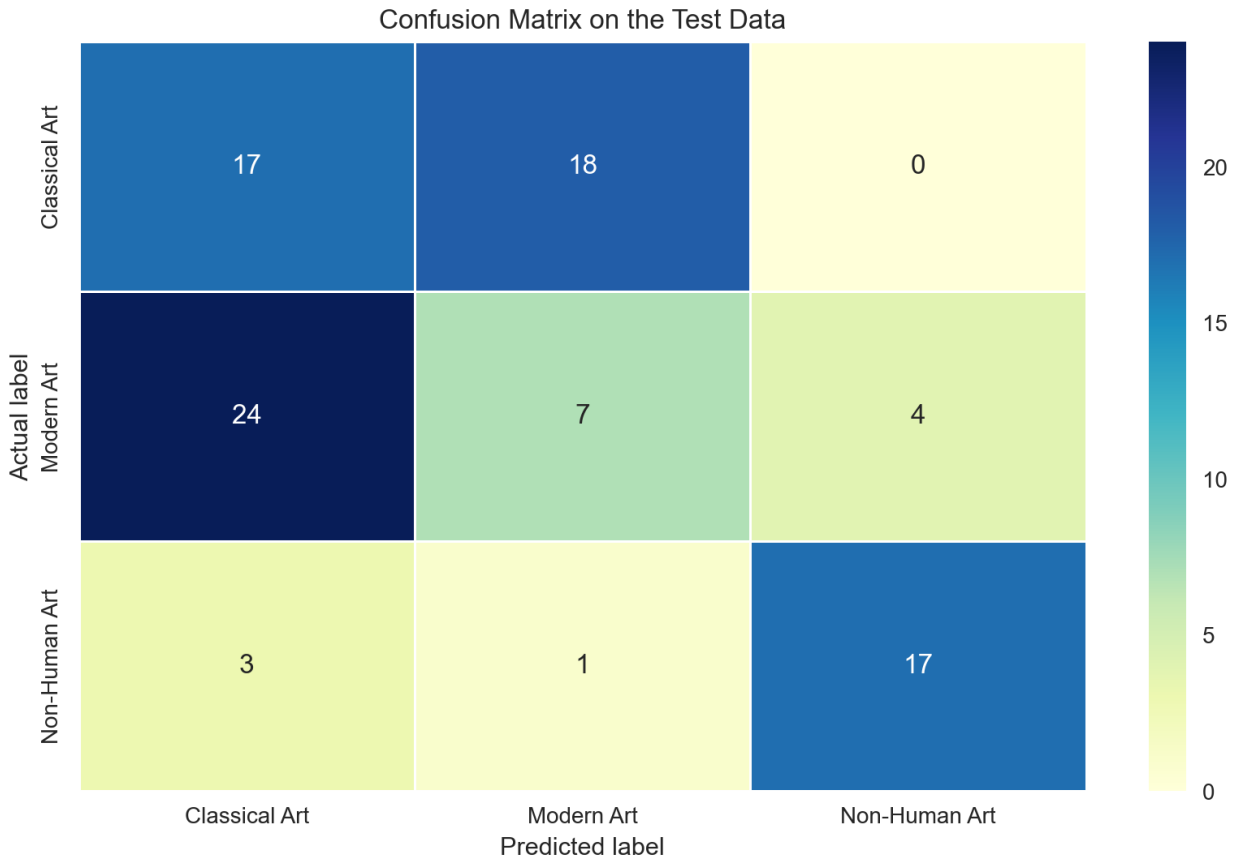


The left-side scatter plot displays the average energy and preference ratings with the KMeans model predictions denoted by the color of the points. In contrast, the right side color scheme denotes the actual classification of the artwork type.





To get a better understanding of the data before selecting modeling methods and parameters, I created the following box plot distribution to serve as a visual representation of the current dataset. At first glance, there are numerous distinctions in the distributions. The artwork classifications in the rating plot display a downward trend in mean from classical to non-human. On the contrary, the average energy scores per art type hold an upward trend. On a deeper level, classic and modern art seemingly rated more similarly than to non-human art ratings. They share more overlap and seem to both have a smoother distribution.



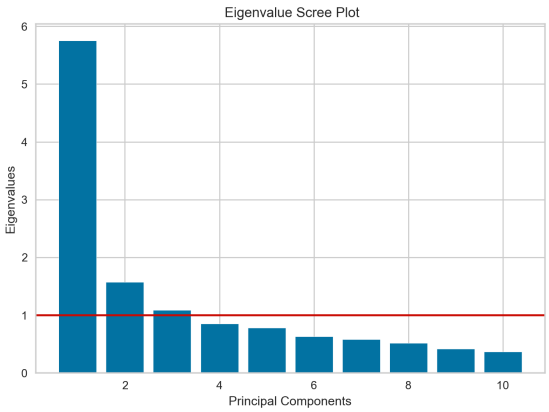
For the 'classical art' class, the model made 44 predictions. Of these, 17 were correct (true positive), and 24 were incorrect (false positive). For the 'modern art' class, the model made 12 predictions. Of these, seven were correct (true positive), and 4 were incorrect (false positive). Finally, for the 'non-human art' class, the model made 21 predictions. Of these, 17 were correct (true positive), and three were incorrect (false positive).

Overall, the model seems to perform better at predicting the 'classical art' and 'non-human art' classes, as these classes have a higher number of true positive predictions. The 'modern art' class has the lowest number of true positive predictions, indicating that the model may need help to accurately predict this class. To make sense of what happened, we need to look at the data. As previously shown in the box plot distributions, there is a significant overlap between the preference ratings and energy scores of classical and modern art. Yet, strangely this asymmetry does not continue into artworks classified as non-human. There are plenty of plausible reasons for the isolation of non-human art in relation to the overlap between Classical and Modern art. Over the course of the ever-evolving world of art, the forefront of new and current styles retain, for the most part, aspects of its generations before. Slowly evolving, yet still present generations on. It would not be unwise to attribute, in some form, the overlapping evolution of art as a confounding aspect of the significant overlap of Classical and Modern art in this dataset, as witnessed in the confusion matrix above. The striking aspect, however, is the isolation of art classified as non-human. For instance, there was no false positive case of classifying classical as non-human, and only one instance of modern art being misclassified as non-human. Unlike the evolution of art, which seemingly grows off the branches of its predecessors, it seems we may be witnessing the birth of an entirely new tree.

There are many assumptions and flaws in the predictions stated above. They reflect my options as an extension of this data analysis project rather than any basis in testing or confirmation. Flaws must be considered. Firstly, my take assumes the validity of the data and its collection methods. Also, is it strange why the non-human artwork class is a mixture of animal, and AI-generated art, as animal and AI-generated art may be in classes of their own? Granted, both types are non-human and foster their bases, but the impact of animal-generated art on the future of human art is likely negligible. In contrast, the noninformative and distant nature of AI-generated art, ever likely primitive, is already inserting influence in the contemporary art world.

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 start, the self-esteem columns contain 15 rows with NaN values which first had to be mitigated. With the loss of 15 rows not being substantial, equating to 5% of the data set, I concatenated the Self-Esteem and Rating columns and then dropped all NaN rows. However, to mitigate the loss in Rank, I increased the K parameter in KFolds cross-validation by one. With the dataset prepped, I began the PCA by selecting the self-esteem columns and plotting their eigenvalues on the scree plot below.



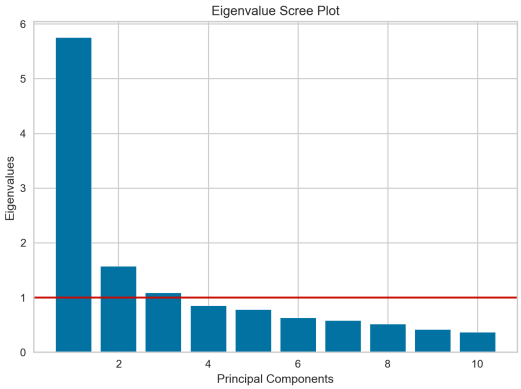
In this case, the first principal component substantially gains over the second and explains more than 50% of the variance. In selecting a suitable reduction of dimensions, although the second and third principal component passes the threshold of the Kaiser Criterion, the substantial loss from the first to second components marks a clear elbow, making the first suitable as well. I then fit a linear regression model using the first principal component of the self-esteem columns to predict the preference rating columns.

9. Consider the first three 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.)

For this question, I created a new data frame with 12 columns related to "dark personality" traits. Then I ran a PCA of the first three principal components to determine the level of variance explained by each. The Scree Plot below shows that the first principal component accounts for roughly 40% of explained variance, and the first three principal components account for 61.43%

of the variance. From this PCA, I cross-referenced the weights of each principal component with the list of Dark Personality questions to find the features of the most importance. As shown in the table below, questions regarding manipulation, lack of remorse, and being cynical are the three features that account for 84% of the variance of the 12 questions.

First, I ran a PCA to determine the variance distribution across the 12 dimensions to minimize multicollinearity. By plotting the eigenvalues, I utilized the Kaiser Criterion and elbow criterion denoted by the horizontal line at y=1. Based on this, I reduced the dimensions to the first three principal components. I then fit a Linear Regression model with the reduced dark personality data. The scoring metrics reveal that the model accounts for 10% of the variance in preference ratings.



	test_r2	test_neg_mean_absolute_error	test_neg_mean_squared_error	test_explained_variance
0	0.02	-1.17	-2.13	0.05
1	0.08	-1.14	-1.99	0.10
2	0.11	-1.14	-2.05	0.13
3	0.06	-1.13	-1.95	0.09
4	0.10	-1.20	-2.25	0.12

```
Mean  R2                  0.07
Mean -MAE                -1.16
Mean -MSE                -2.07
Mean  Explained Variance  0.10
```

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

I started by re-processing the data by selecting all columns except for those containing the political data. With the political column needing to be split into the respective progressive, liberal, then the rest, I applied an ordinal encoder and mapped the data to their respective class. I

then split the data by group type (‘ratings,’ ‘action,’ etc.) and then dropped all null rows (25 total). Following this, I applied PCA to the below groups denoted as ‘#PCA,’ selecting a principal component level in line with the Kaiser and Elbow Criterions.

Q10

Data Pre-processing

```
data = pd.read_csv(data_file)
political = pd.DataFrame(data['3.24'])
encoder= ce.OrdinalEncoder(cols=['3.24'],return_df=True,
                           mapping=[{'col':'3.24','mapping':{1:0, 2:0, 3:1, 4:1, 5:1, 6:1}}])
political['encoded'] = encoder.fit_transform(political)

ratings = data.iloc[:,:91]
energy = data.iloc[:,91:182]
dark = data.iloc[:,182:194] #PCA
action = data.iloc[:,194:205] #PCA
esteem = data.iloc[:,205:215] #PCA
demographics = data[['19','2.45','0','2.46','2.47']]

all = pd.concat([ratings,energy,dark,action,esteem,demographics,political.encoded],axis=1).dropna(axis=0)

ratings = all.iloc[:,:91]
energy = all.iloc[:,91:182]
dark = all.iloc[:,182:194] #PCA
action = all.iloc[:,194:205] #PCA
esteem = all.iloc[:,205:215] #PCA
demographics = all[['19','2.45','0','2.46','2.47']]
political_y = all.encoded
```

[172] ✓ 0.7s

Once the principal components had been selected in their respective groups, I then began trying out different models and parameter configurations. To speed up the testing process, I created three interchangeable modes, each with its parameter configuration. I then continued to tweak and tune various predictors and model parameters to find the best model. I found that different models reacted differently to changes, and all tested models performed better than the Linear Regression model across all attempted variations in this last section.

Logistic Regression: The Logistic Regression model performed better when the energy and preference rating columns were removed. The Logistic Regression’s test accuracy came out as one of the best models and had a 0.654 test accuracy.

Logistic Regression Model

```
model = LogisticRegression()
cv = RepeatedKFold(n_splits=4, n_repeats=3, random_state=14369331)
scoring = ['accuracy']
n_scores = cross_validate(model, X, y, scoring=scoring, cv=10)
for key, score in n_scores.items():
    print(f'{key}: {mean(score):.3f}'.format(key,mean(score)))

[145] ✓ 0.2s
```

```
fit_time: 0.021
score_time: 0.001
test_accuracy: 0.654
```

Gaussian Naive Bayes Model:

The Naive Bayes model seemingly increased in accuracy after removing ratings and energy features. The model scored a test accuracy = 0.647 and a test AUC = 0.674.

Gaussian Naive Bayes Model

```
gnb = GaussianNB()
cv = RepeatedKFold(n_splits=4, n_repeats=3, random_state=14369331)
scoring = {'acc':'accuracy','r2':'r2', 'mae':'neg_mean_absolute_error','auc':"roc_auc"}
n_scores = cross_validate(gnb, X, y, scoring=scoring, cv=10)
for key, score in n_scores.items():
    print(f'{key}: {mean(score):.3f}'.format(key,mean(score)))

[149] ✓ 0.1s
```

```
fit_time: 0.002
score_time: 0.005
test_acc: 0.647
test_r2: -0.446
test_mae: -0.353
test_auc: 0.674
```

Support Vector Machines:

Next, I used Support Vector Machines to make predictions and score the accuracy. The support vector also didn't like the energy and rating columns, even when the 2-dimensional array of mean ratings and energy was used over the 181 dimensions of the full ratings and energy scores. The Support Vector model performed exceptionally well in relation to the other models tested. In hindsight, it is important to note the ordinality of the data features and the binary classification of the dependent variable during the model selection and development process.

Support Vector Model

```
clf = SVC(kernel='linear')
model = RandomForestClassifier()
cv = RepeatedKFold(n_splits=4, n_repeats=3, random_state=14369331)
scoring = {'acc':'accuracy','r2':'r2', 'mae':'neg_mean_absolute_error','auc':"roc_auc"}
n_scores = cross_validate(model, X, y, scoring=scoring, cv=10)
for key, score in n_scores.items():
    print(f'{key}: {mean(score):.3f}'.format(key,mean(score)))
'''
Model Accuracy Decreased After Removing Ratings and Energy Features
'''

[157] ✓ 1.8s
```

```
fit_time: 0.149
score_time: 0.030
test_acc: 0.673
test_r2: -0.339
test_mae: -0.327
test_auc: 0.718
```