Ghina Hasan Abdelmajid Al Shdaifat

Art Analysis: Capstone Project

Dimension Reduction:

- Dimension reduction started with creating a new dataset with the necessary variables. After that, the PCA was then fitted with the transformed data. I used PCA in more than one question to calculate eigenvalues but to also how well a regression model predicts outcomes using a specific number of components.

Data Cleaning:

- The entire dataset was set to a variable at the beginning called data. For each question, a new dataset was created from data, the variable name followed by the question number. Only the variables necessary would be taken from data and would be cleaned at this step to minimize the amount of data lost.

---

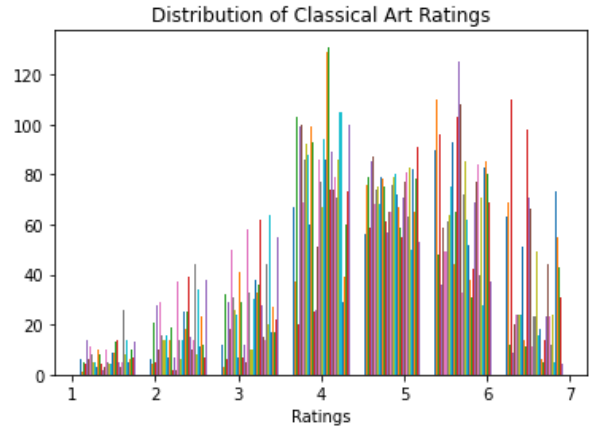
1) Is classical art more well-liked than modern art?

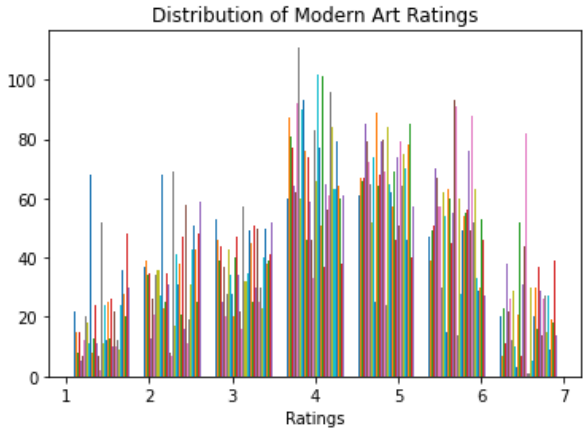
Using the theArt.csv dataset, column 6 specifies the "Type Code" of the art, where 1 = classical art, 2 = modern art, and 3 = non-human art. The first 34 columns in the theData.csv dataset represent the user preference likings of classical art paintings, while the following 34 columns represent the user preference likings of modern art paintings. I used slice indexing to store the ratings in separate variables, which I then calculated the total mean and median of each variable well and performed a Mann-Whitney U-statistic.

Since both variables represent categorical data (art preference **ratings**), I decided to calculate the median alongsides the means to ensure that the findings are representative of the data. The mean is a measure of central tendency that is used to summarize continuous data, not categorical data, which is why it may not be appropriate to use. However, I did find that by calculating the mean, median and Mann Whitney U-statsitic, I was able to understand and analyze the distribution as well as the difference of the variables.

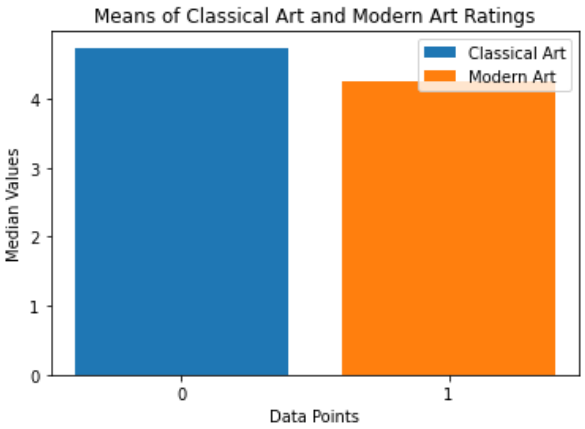
The mean of the classical art ratings carry a value of 4.742, which is larger than t**he mean of the modern art ratings – 4.257.** By performing a Mann-Whitney U-statsitic, I was able to calculate the p-value of both variables, which accounts for the difference in median values, where I got a **U-statistic value of 52313.757** and a **p-value of 0.109**.

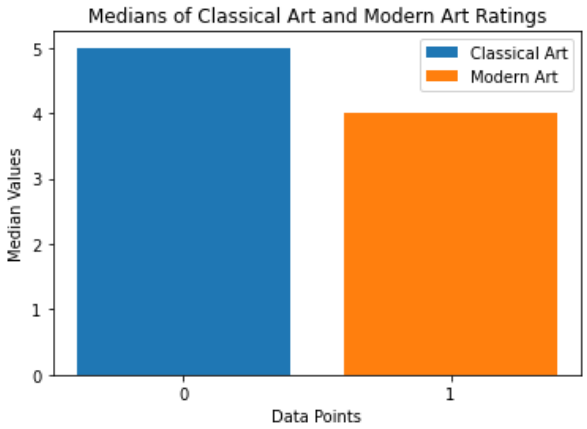
The numbers show that there is some difference between the two samples, but the difference is not statistically significant at the chosen level of significance (0.05), identifying that the difference is not large enough. This is reflected in the bar graphs below, where the data distribution of both variables is not visually evident.





Therefore, to create a simpler visual graphic of the distribution of both variables, I decided to create a bar graph that compares the median values of classical art ratings and modern art ratings where they carry a value of **5** and **4**, respectively.

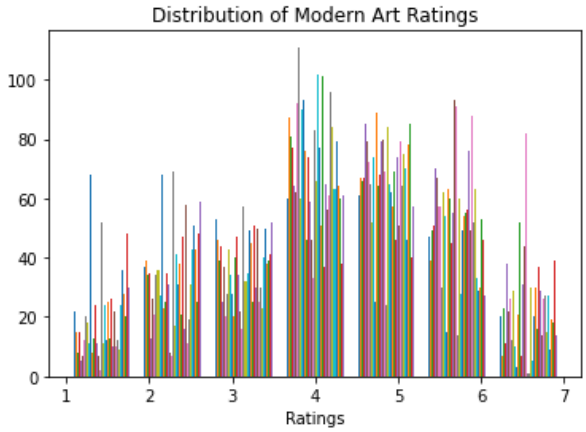


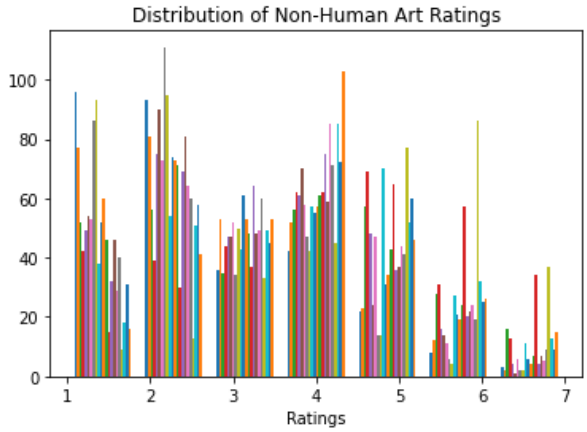


Since classical art has a higher median value and the distribution of the classical art ratings are more skewed to the right, indicating higher preference ratings, we can conclude that **classical art is more liked than modern art.**

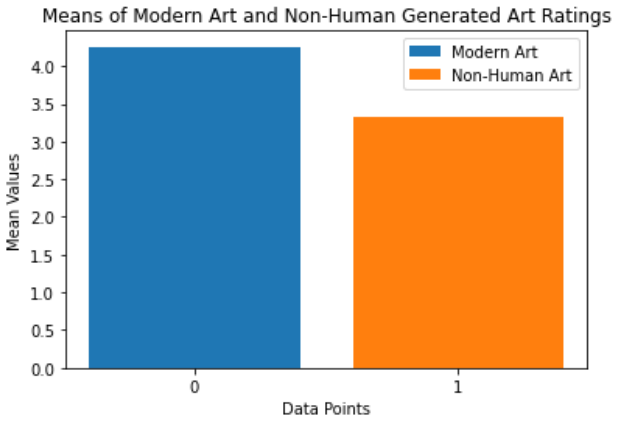
2) Is there a difference in the preference ratings for modern art vs. non-human (animals and computers) generated art?

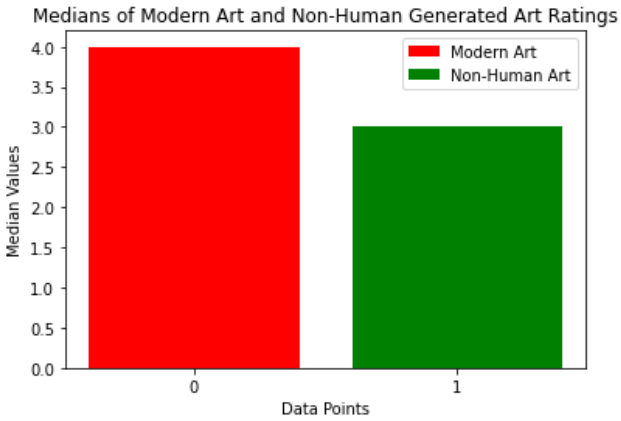
I additionally performed a Mann-Whitney U-statistic to determine the difference in preference ratings for modern art and non-human generated art, where I got **45308968.5 for the U-statistic** and **2.316e-259 for the p-value.** In this case, the p-value of 2.316e-259 is extremely small, indicating that there is a very low probability that the difference between the samples occurred by chance, and that there is a significant difference between the medians of the two samples. To visualize the difference, I graphed the distribution of both variables and represented their median values, as shown below.





Unlike the minor difference of distribution between classical art and modern art ratings in the previous question, the distribution of modern art ratings is skewed more towards the right, indicating high preference ratings, while the distribution of the non-human art ratings is more left skewed, indicating lower preference ratings, representing the difference of preference likings between both variables. This difference is further represented by the median and mean values of both variables, where **modern art has a median value of 4 and a mean value of 4.256571428571428**, while **non-human art has a median value of 3 and a mean value of 3.3334848484848485**, indicating that **modern-art is preferred to more than non-human art.**





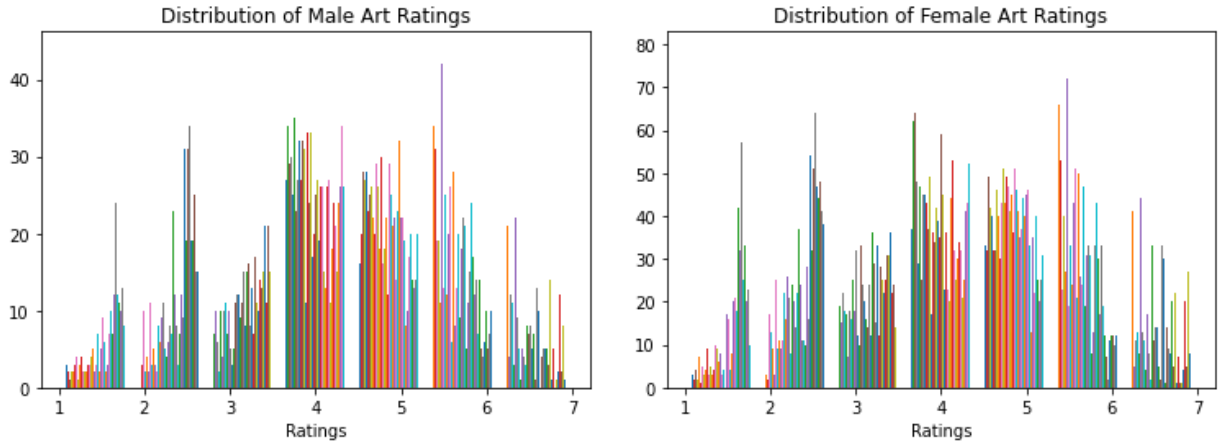
However, simultaneously, it is important to note that the modern-art ratings have a higher sample size (34) than non-human art ratings (20), which may affect statistical measurements. In this case, a smaller sample size can result in a median that is more influenced by extreme values, or outliers, in the sample, which can make the median less representative of the overall distribution of the sample. This may also result in a larger p-value, which means that it is more likely that the difference between the samples occurred by chance, and less likely that there is a significant difference between the samples, which may be a reason why the U-statistic is relatively large, confirming that there is less of a difference between the samples.

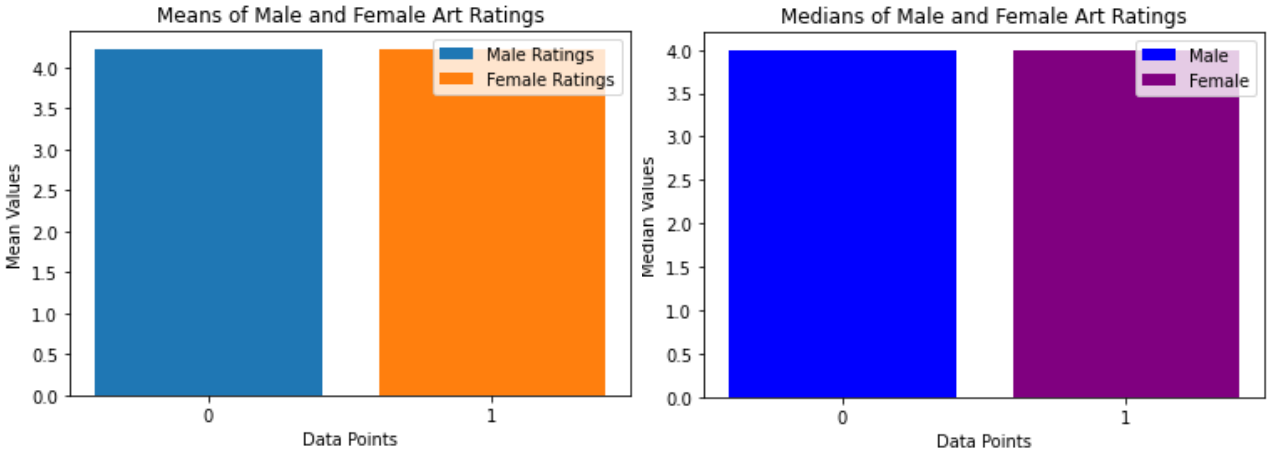
3) Do women give higher art preference ratings than men?

To determine whether women give higher art preference ratings than men, I first needed to separate the values rated by men and women from the dataset, where 1 = male, 2 = female, and 3 = non-binary. While looking at the dataset, I noticed that NaN values were present, which is why I decided to clean the data before beginning to statistically measure and analyze the variables. I did so by using the data.dropna() function, which then allowed me to separate the men and female ratings using the data.loc[] function, as shown in my code below.

```
#%% Q3: Do women give higher art preference ratings than men?
#Collect data
gender = user_data[:,216]
art_ratings = user_data[:,0:91]
#put variables in a dataframe to drop nan values
df_gender = pd.DataFrame(gender)
df_art_ratings = pd.DataFrame(art_ratings)
#join both dataframes together
df_art_ratings[91] = df_gender[0]
#drop nan values
gender_ratings_data = df_art_ratings.dropna()
#locate 1 for male, and 2 for female:
male_ratings = gender_ratings_data.loc[gender_ratings_data[91] == 1]
female_ratings = gender_ratings_data.loc[gender_ratings_data[91] == 2]
#delete values from dataframe:
del male_ratings[91]
del female_ratings[91]
#return to array:
female_ratings = female_ratings.to_numpy()
male_ratings = male_ratings.to_numpy()
```

After controlling for missing values, I performed a Mann-Whitney U-statsitic, and calculated the p-value – where I got a **U-statistic value of 8563.423**, indicating that there is not a significant difference between the medians of the two samples, and a **p-value of 0.375,** indicating that it is likely that the difference between the samples occurred by chance. Therefore, I calculated the **median values** where I got an equal number of **4** for both variables, respectively. Equal median values suggests that the distribution of the two samples is similar, but that does not mean that they are identical or that their means are equal. Hence, to gain a better understanding of the data, I decided to graph the distributions of both variables and calculate the mean values (**male mean: 4.214459224985541; female mean: 4.225735158696053**) to visualize the difference and give me more clarity on the data.

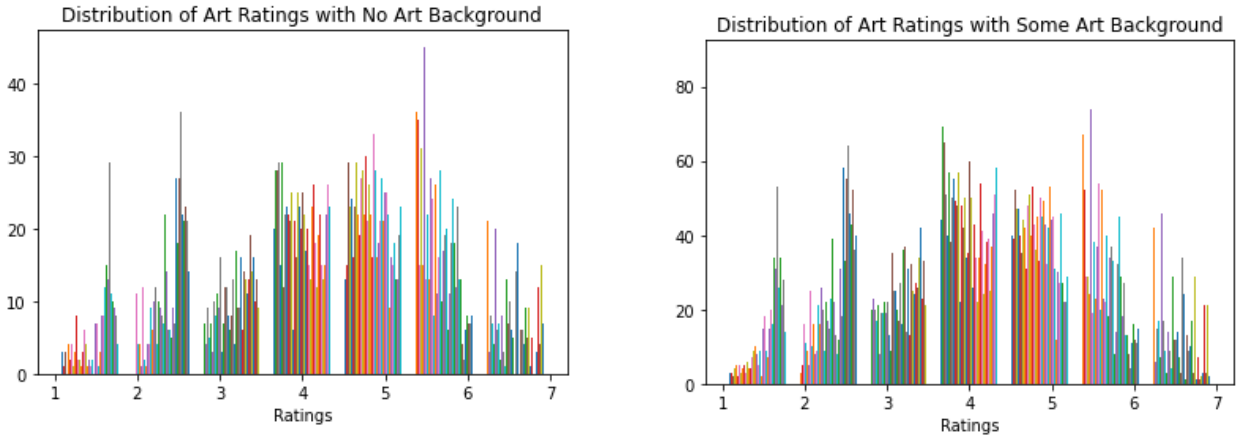




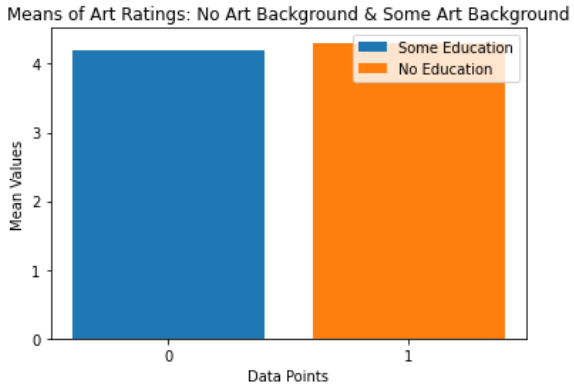
As seen above, there is not a significant difference between the distributions of both variables and a 0.01127593371 difference in the mean values. Since the median values are equal and the distribution of both samples are very similar, we can conclude that **women do not necessarily give higher preference ratings than men**.

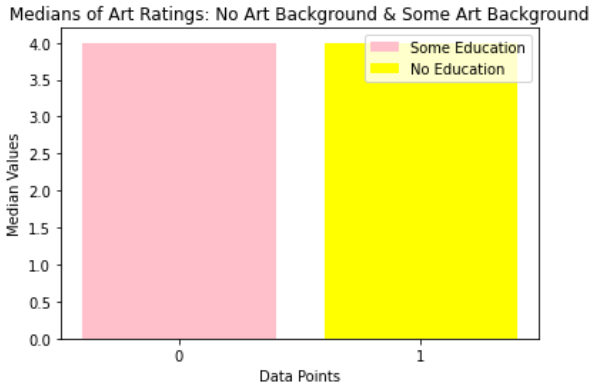
4) Is there a difference in the preference ratings of users with some art background (some art education) vs. none?

Similar to the previous question, I needed to account for NaN values for ratings with some art background and no art knowledge. Therefore, I cleaned the data using the same method of the previous question and dropped all NaN values. After doing so, I performed a Mann-Whitney U-statistic and calculated the p-value where I got a value of **9051.440** and **0.455**, respectively. This indicated that there is not a significant difference between the medians of the two samples and that it is likely that the difference between the samples occurred by chance due to the high p-value. Hence, I decided to visualize the distribution of the variables by graphing the data:



As shown in the graphs above, there is a slight difference between the ratings of users with no art background and with some art background – where individuals with no art background tend to give higher art ratings than users with some art background. This minor difference is shown by the overall shape of the graph, where the one on the left is more skewed to the right than the other graph. However, the minor difference of art preference ratings is not significant, which is why I decided to calculate and graph the median and mean values:

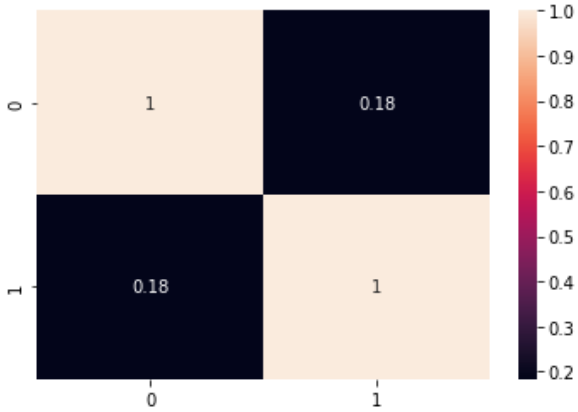


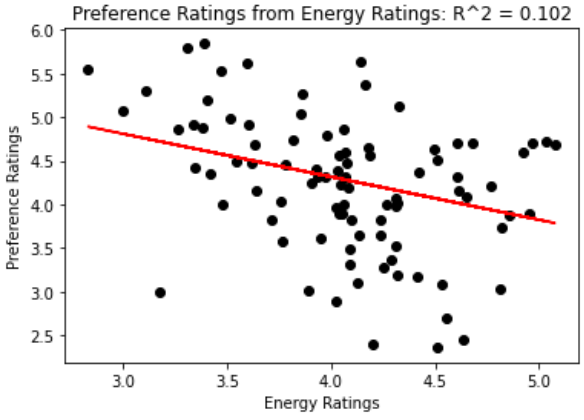


As shown in the graph, both samples have **equal median values of 4**, respectively, suggesting that the central tendency of the two samples is similar. Although there is a 0.11470097894 increase in the 'no education' art ratings mean values where – **mean of 'no education': 4.305781175346393 & mean of 'some education': 4.191080196399345 – the difference is not significant**.

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 first began with calculating the person's correlation matrix between both variables to determine if there is a relationship there or not. As seen from the matrix and heatmap below, **a correlation value of 0.1837** indicates a **moderate positive correlation** between the two variables. This means that as the values of energy ratings increase, the value of art preference ratings are also likely to increase, and vice versa. Hence, I decided to do a linear regression model to predict art preference ratings from energy ratings.





I used a train-test split to evaluate the model's performance on the test set and to get a better understanding of how well the model will perform on new, unseen data, which is fundamental to predict art preference ratings. Additionally, using a train-test split helps to prevent overfitting, which allows us to avoid poor generalizations and poor performances on new data. Hence, I calculated the following statistical measurements to better understand how well the linear regression model fits the data as well as analyze the relationship between energy ratings and art preference ratings:

Mean Squared Error: 0.9100046792760648
Mean Absolute Error: 0.752092028392483
R^2: 0.10180796650785162
Cross Validation Score: 0.08478785718380148

The statistical values above suggest that the linear regression model is not a good fit to the data, as the errors and the R^2 are relatively high and the Cross Validation Score is relatively low. This could be due to the lack of linear relationship between the dependent and independent variables, which can be seen in the linear regression scatter plot.

The scattered plots represent the predicted art preference rating values of the energy ratings variable for a given value of the predictor variable. If the plots are closer to each other, it means that the predicted values are relatively close to the actual art preference rating values, which indicates that the model is able to accurately capture the relationship between the variables. However, in this case, the plots are widely scattered, which could indicate that **the model does not accurately capture the relationship of the art preference ratings and energy ratings**, or that there may be other factors that are affecting the relationship, or that there is simply no relationship between the variables.

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.

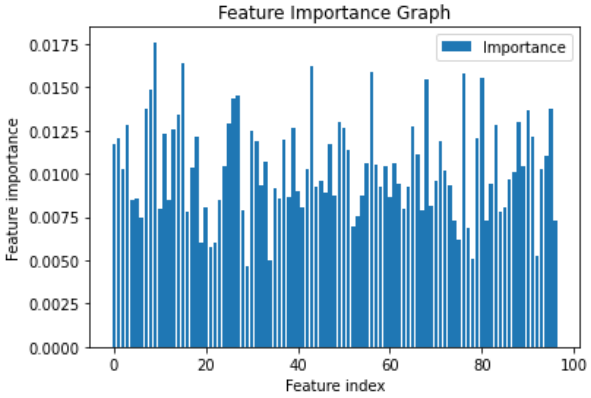
Since I will be predicting more than one feature (energy ratings and demographic information), I decided to use a random forest regression model, in which each feature is trained on a random subset of the data. They are able to handle high-dimensional data and large number of features while simultaneously being resistant to overfitting. Therefore, the use of a random forest regressor can help to improve the accuracy and robustness of the predictions.

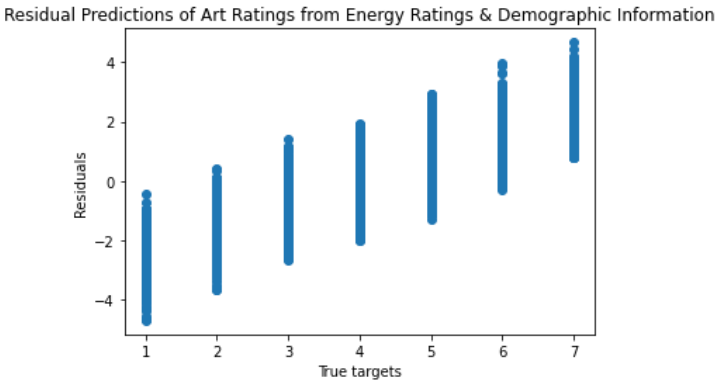
I first began by collecting the demographic data, which includes age and gender. I then concatenated the data and cleaned it from NaN values. Since the number of NaN values was relatively high, I decided to "fill" the NaN values with the median instead of "dropping" them to ensure that there is enough data to model and predict. After doing so, I calculated the following measurements:

Mean Absolute Error: 1.1259026687598117
Mean Squared Error: 1.9948640502354789
Root Mean Squared Error: 1.4123965626676804
R^2: 0.051939403305666904

Standard Deviation of the Cross Validation Scores: 0.01917414767266982

The standard deviation of cross validation scores is a measure of the variability of the scores obtained on the different folds. A larger standard deviation indicates a greater variability in the scores, however the standard deviation score is relatively low with a value of 0.01917414767266982 . The values represented above suggest that the random forest regression model is performing relatively poorly, as the errors and the R squared are relatively high and the standard deviation of the cross-validation scores is relatively low. This may indicate that the model is not able to accurately capture the relationship between the predictor and response variables. Therefore, I decided to gain a better understanding of the data and the different features by plotting a features importance graph and displaying the residual predictions of art ratings from energy ratings and demographic information:

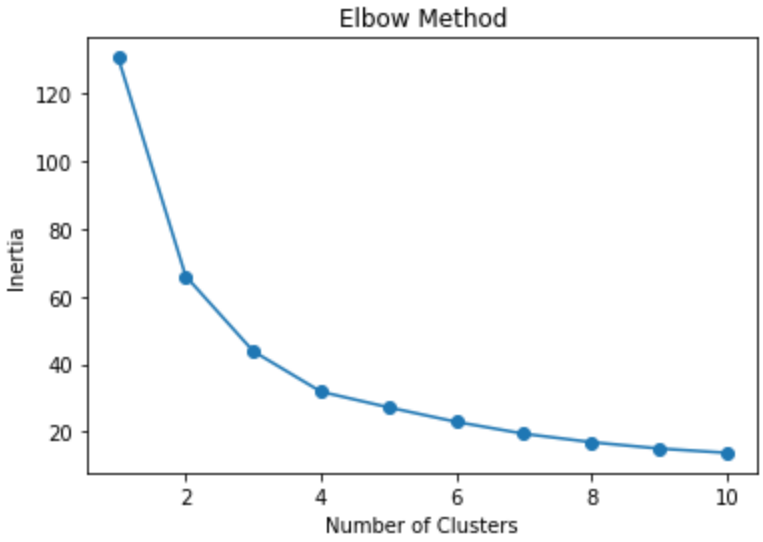




Unfortunately, both graphs do not provide us with much information on the energy ratings and demographics, confirming that the model is not able to accurately capture the relationship between art preference predictions from energy ratings and demographic information. Additionally, this further indicates that there is no relationship between the art preference predictions and energy ratings.

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?

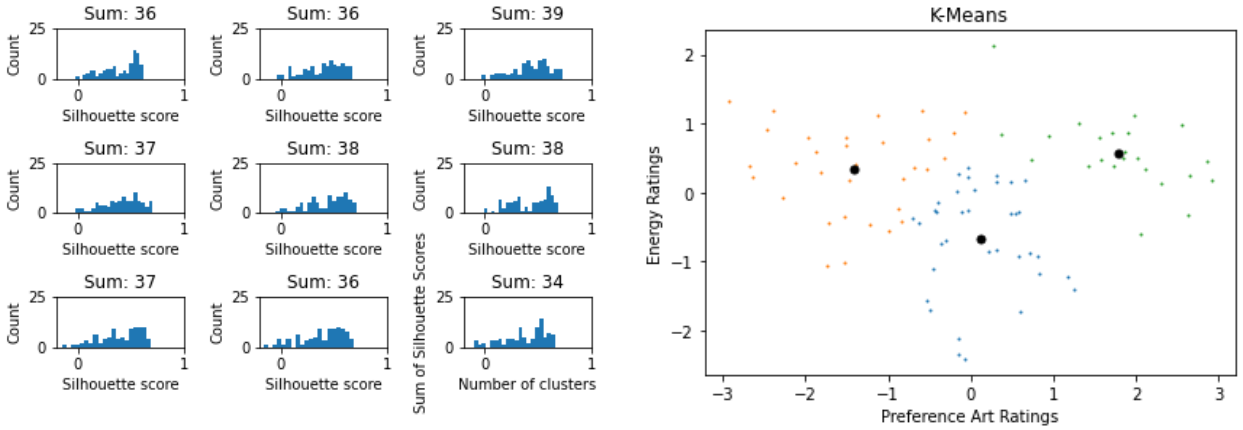
I performed a K-Means model to identify the number of clusters between average preference ratings and average energy ratings. I first began with collecting and separating my data, and inputting them in separate variables: art ratings, energy ratings and art types. To determine the optimal number of clusters, I created a knee-elbow plot, which is a graphical representation of the within-cluster sum of squares (WCSS). As seen in the graph below, the point at which the WCSS starts to flatten out and the rate of improvement begins to decrease is around 3 clusters.



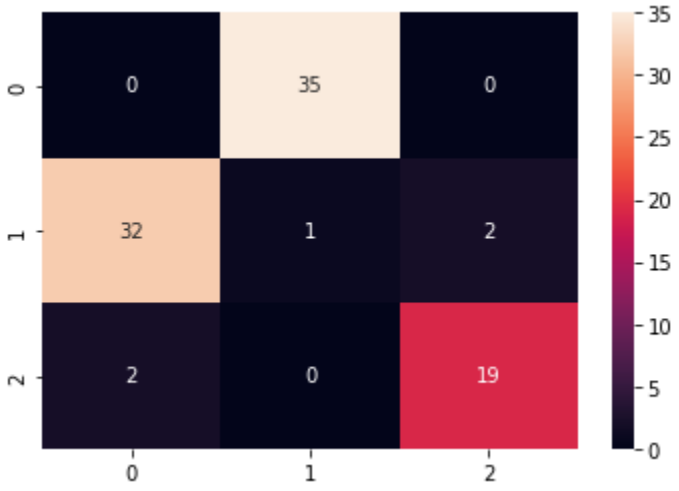
However, to ensure that 3 is the accurate number of clusters, I decided to perform a principal component analysis (PCA) as a statistical technique to reduce dimensionality of the dataset while preserving as much of the variance as possible. In order to do so, I calculated the z-scores to standardize the data by transforming each variable to have a mean of zero and a standard deviation of one. This will give the art and energy ratings an equal weight, ensure that the data is in the same units, and reduce the influence of outliers. Since both energy and art ratings range from 1-7, the variables already carry an equal weight and are in the same units, hence calculating the z-scores was unnecessary.

To ensure that 3 is the correct number of clusters, I calculated the silhouette scores and the sum of the silhouette scores to understand the data further. As seen in the graph below, the sum of the silhouette scores reaches its maximum value (39) at position number three of the iteration, confirming that the number of clusters present in the data is indeed 3.

```
zscoredData = stats.zscore(data)
# Initialize PCA object and fit to our data:
pca = PCA().fit(zscoredData)
# Eigenvalues: Single vector of eigenvalues in decreasing order of magnitude
eigVals = pca.explained_variance_
# Loadings (eigenvectors): Weights per factor in terms of the original data.
loadings = pca.components_*-1
# Rotated Data - simply the transformed data:
origDataNewCoordinates = pca.fit_transform(zscoredData)*-1
#Silhouette Score
x = np.column_stack((origDataNewCoordinates[:,0],origDataNewCoordinates[:,1]))
```



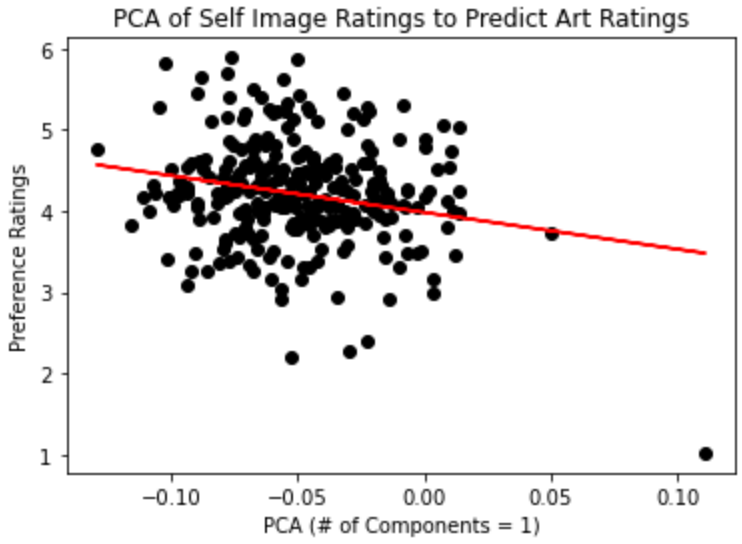
To identify the identity of each cluster, I decided to create a confusion matrix that displays the Type I and Type II errors – the number of correct and incorrect predictions made by the model for each art type (classical, modern and non-human generated art).



As seen by the confusion matrix above, the model did a relatively good job at predicting the art types, where the model was able to correctly identify classical art, incorrectly predicted two modern art as non-human art, and one modern art as classical art, as well as incorrectly predicted two non-human generated art as modern art. As for the identity of the clusters, we were able to distinguish that classical art is more well liked and preferred more than modern art, and simultaneously, modern art is more preferred to non-human generated art from the previous questions. Hence, the right cluster with the higher preference ratings can be identified as the classical art cluster, the middle represents the modern art category, while the left cluster is the non-human generated art ratings.

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

I performed the first Principal Component Reduction on the self-image ratings and inputted the findings through a linear regression model:



To further understand the model and analyze how well the model predict art preference ratings, I calculated the following statistical values:

Mean Absolute Error: 0.4351452339676892
Mean Squared Error: 0.3655152067830631
R^2: -0.06028021211338319

The MSE value of 0.365 indicates that the average squared error of the art preference predictions is relatively small, hence the differences between the predicted values and the true values were small and that the model is able to capture the general trend in the data reasonably well. This is also seen by the spread of the plots on the linear regression model, where the plots are close to each, capturing the relationship between the variables. The MAW value of 0.4351 additionally indicates that the average absolute error of the art preference predictions is relatively. In other words, the model is able to make art preference predictions that are close to the true values, on average.

However, the R^2 value of -0.060 indicates that the regression model does not fit the data very well. Although a negative R^2 value is not very meaningful, in this case, this could be due to the concept of outliers. It is important to recognize that the scattered data plots identify extreme outliers. Extreme values can pull the regression line away from the majority of the data points, resulting in a poorer fit falsely indicating that the model is not performing as well as it could be. The MSE and MAE values are similarly sensitive to the presence of the outliers which could be contributing to higher MAE and MSE values.

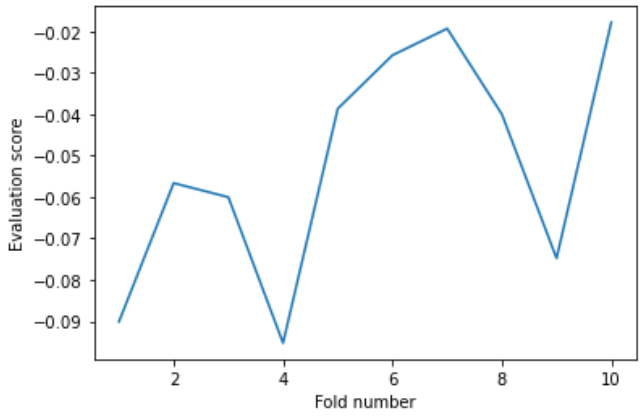
It is worth noting that the MSE, MAE and R^2 values are all based on a single component PCA. This means that the reduction of dimensionality to one single component could potentially limit the model's ability to capture the complexity of the data.

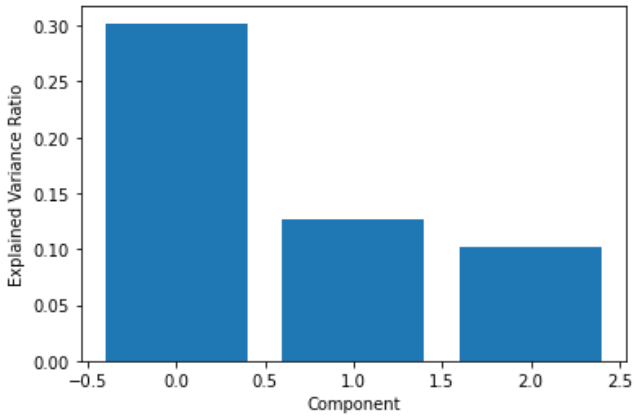
9) Consider the first 3 principal components of the "dark personality" traits – use these as inputs to a regression model to predict art preference ratings. Which of these components significantly

predict art preference ratings? Comment on the likely identity of these factors (e.g., narcissism, manipulativeness, callousness, etc.).

Similarly to the previous question, I performed a PCA of dark personality traits as inputs to predict art preference ratings, but this time, with three principal components instead of one and performed the model through the K-Fold cross validation method. The reason why I decided to use the KFold method is because it helps to reduce the risk of overfitting as the linear regression model is trained and evaluated on different subsets of the data that have larger proportionality and hence, reduces the variance of the evaluation metric. Considering that we are identifying three different PCA components, I found the KFold an appropriate method to use.

As I was analyzing the data, I realized that there are a lot of NaN values for "dark personality" traits. Considering that the data set only has 12 columns, I decided to "fill" the NaN values with the median instead of "dropping" them to keep as much data as possible.





I decided to calculate the variance ratio to determine the model's performance across the different fold of the data in the KFold cross-validation process. The **total sum of my variance ratio** came out to be **0.9999999999999997**, which indicates that the variance of the model's performance is very close to the mean of the performance across the folds. This could suggest that the model is performing relatively consistently across different folds of the data, and that it is generally highly variable.

To identify the three most common features for each principal component, I calculated the loadings of each feature on the principal components by multiplying the components by the inverse of the standard deviation of each feature as shown in my code below:

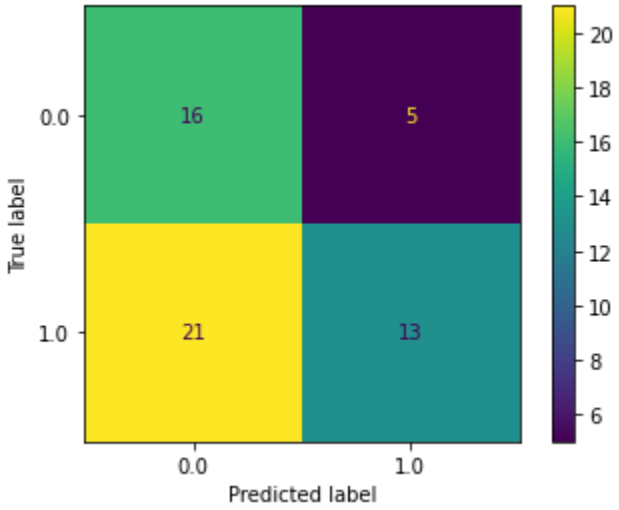
```
#%%
Q9pca.fit(dark)
# Loadings
loadings9 = Q9pca.components_ @ np.diag(1 / dark.std(axis=0))
# Sort the loadings in decreasing order to identify the features with the highest loadings
sorted_loadings9 = np.abs(loadings9).argsort(axis=1)[:, ::-1]
# Print the indices of the three most common features for each principal component
for i in range(3):
    print(f"Principal component {i+1}:")
    print(sorted_loadings9[i, :1])
```

The Three Most Common Features for Each Principal Component:
Principal component 1: [0] - “I tend to manipulate others to get my way”
Principal component 2: [4] - “I tend to lack remorse”
Principal component 3: [7] - “I tend to be cynical”

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 first began with splitting the political orientation ratings into two separate dataframes: 0 – “left” (progressive and liberal), and 1 – “right” (everyone else). Considering that we are looking at the different political orientations of the users, I decided to use a Random Forest Classifier because the method can handle a large number of features and is not sensitive to scaling of the data. Hence, the method is resistant to overfitting, as the combination of many decision trees helps to smooth out the effect of any single tree that may be overfitting the training data.

I decided to use accuracy as a measure of how the model is able to predict the correct political orientation. In this case, I got an **accuracy score of 0.53**, which means that the model is able to correctly predict the output for approximately 53% of the test data. To visualize this, I performed a confusion matrix to show when the model was able to currently and incorrectly predict the outcome:



Although an accuracy score of 0.53 may be considered moderate, it is not a great value. One reason for this is the model may not be complex enough to capture the complexity of the data. In other words, since I am looking at political orientation only, the model may be too simple, which may not be able to accurately capture the relationships between the features and the target variable; hence a low accuracy score.