Hamza Alshamy
DS 112-001
5/11/2023
Dr. Wallisch

Capstone Report

Pre-processing and data extraction

Generally, both Numpy and Pandas were used in this project. The primary use of Pandas was for cleaning (removing NaNs), extracting (smaller data frames), and organizing data (joining/concatenating) while the primary use of Numpy was using arrays in functions for more efficient processing. In the pre-processing stage, the data was imported using Pandas, extracted relevant rows and columns for easier inspection, and lastly, inspected the shape/dimensions of each extracted dataset to see if the extraction makes sense. As for Principal Component Analysis (PCA), a form of dimensionality reduction, it was used whenever necessary, such as in questions 8, 9, and 10. Further, there was additional pre-processing for each question that involved the concatenation of smaller dataframes of interest, dropping NaN rows to maintain as much data as possible, then conducting analyses. Lastly, a training/testing split was conducted in all regression models.

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

To answer the above question, hypothesis testing is advised to see whether a difference between the two types of art (classical and modern) is due to chance. The chosen test statistic is a Mann-Whitney U-test (Wilcoxon Rank Sum Test) because we are dealing with ranks and we do not know if the data is normally distributed. If the data were normally distributed a Welch T-test may have been appropriate as well. In Figure 1, the median rating for classical art ratings is one point higher than that of modern art. Let the Null Hypothesis be that there is no difference between the two ratings, the Alternative is that the median art ratings for classical art are higher than those of modern art, and a significance level $\alpha = 0.05$.

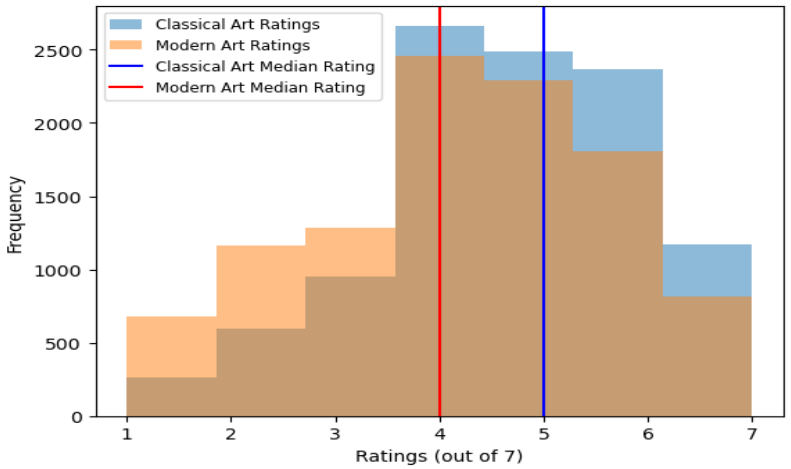

Figure 1: Classical Art Ratings and Modern Art Ratings

Hamza Alshamy
DS 112-001
5/11/2023
Dr. Wallisch

The P-value was 0.0 which is smaller than the chosen significance level, $\alpha = 0.05$. Therefore, it is not plausible that this outcome is due to chance alone, so we can reject the Null Hypothesis. We can say that classical art is more well-liked than modern art and that the observed difference is unlikely to be due to chance alone.

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

A hypothesis test was appropriate to answer the question. For the same reasons as question 1, a Mann-Whitney U-test was chosen. Let the Null Hypothesis be that there is no difference between art preference ratings for modern art and non-human art, Alternative is that there is a difference between the two, and a significance level $\alpha = 0.05$.

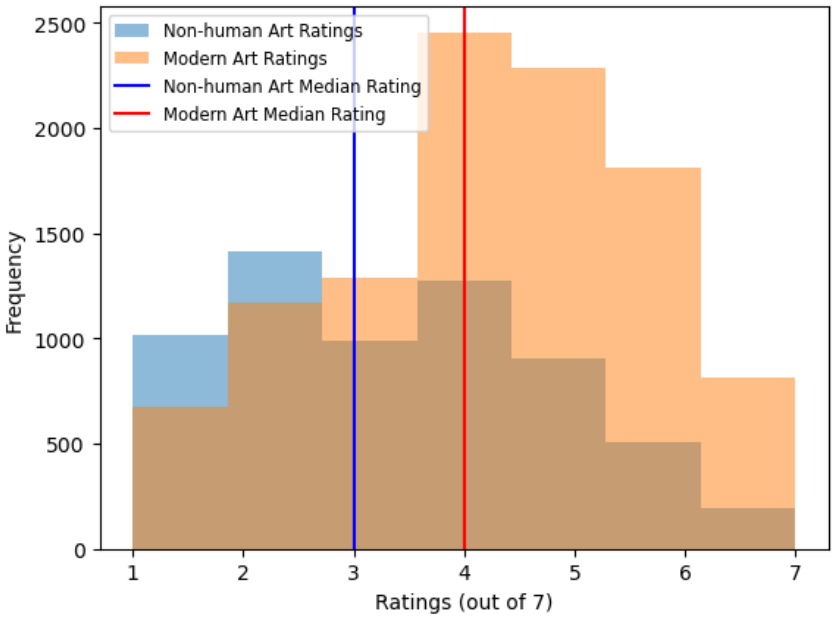

Figure 2: Modern Art Ratings and Non-human Art Ratings

As can be in Figure 2, the median rating of modern art ratings is one point higher than those of non-human art. After conducting the hypothesis testing, the P-value was 0.0 which is below the chosen significance level, $\alpha = 0.05$. Therefore, it is not plausible that this outcome is due to chance alone, so we can reject the Null Hypothesis. We can say that there is a difference between the preference for modern art and non-human art and that modern art is more well-liked, and that the observed difference is unlikely to be due to chance alone.

Hamza Alshamy
DS 112-001
5/11/2023
Dr. Wallisch

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

Similar to questions 1 and 2, a hypothesis test will help determine if there is a difference in art preference ratings between men and women and whether it is due to chance. We conduct a Mann-Whitney U-test again because we are dealing with ranks and w do not know if the data is distributed normally. In this case, let the Null Hypothesis be that the median ratings of women are the same as men, Alternative is that the median ratings of women are higher than men, and a significance level $\alpha = 0.05$.

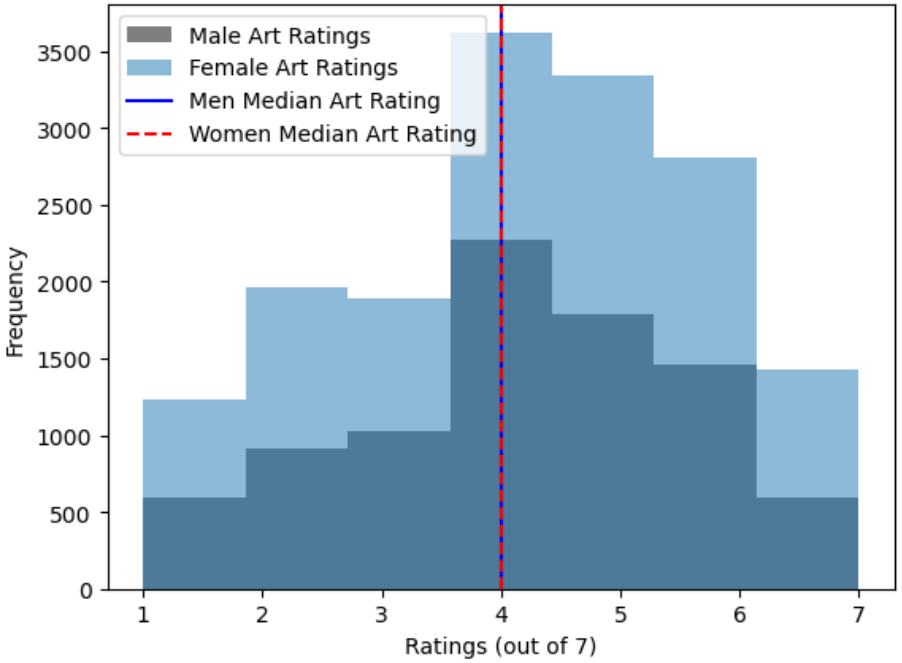

Figure 3: Male and Female Art Ratings

Observe from Figure 3 that both men and women have the same median art preference. After conducting the hypothesis test, the P-value obtained was 0.271 which is larger than the significance level, $\alpha = 0.05$. Therefore, it is plausible that this outcome is due to chance alone, so we fail to reject the Null Hypothesis. We can say that there is insufficient evidence to suggest that the median ratings of women are higher than men.

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

Similar to questions 1,2, and 3, it is appropriate here to use hypothesis testing. Since we do not know if the data is normally distributed and we are dealing with ranks, yet again, we use a Mann-Whitney U-test. Let the Null hypothesis be that the median art preference for both

those with some a no art background is the same, the Alternative being that there is a difference in the median art preference of those with some and those with no art background, and the significance level is $\alpha = 0.05$. Observe in Figure 4 that based on the data we have, the median art rating for both groups is 4.0 which gives the impression that there is no difference.

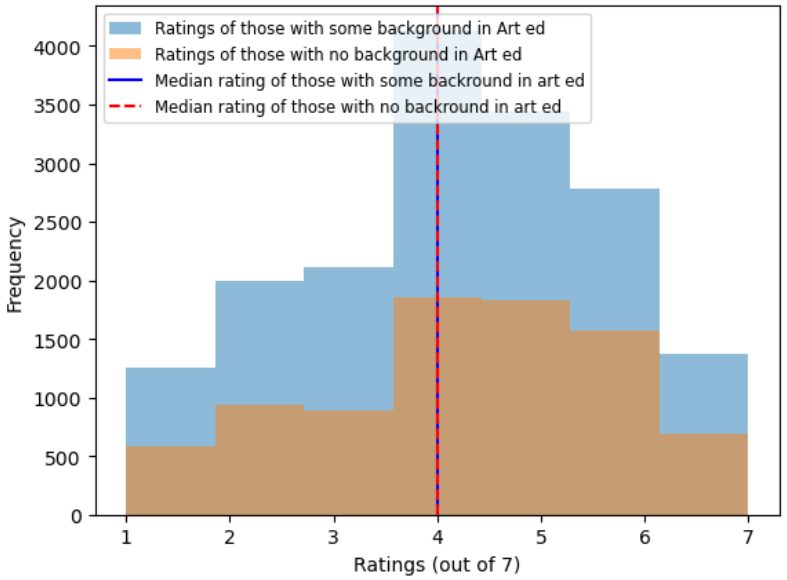

Figure 4: Art ratings of those with some and those with no art background

After conducting the hypothesis testing, we got a P-value of 0.0 which is smaller than the chosen significance level $\alpha = 0.05$. Therefore, we reject the null. We can say that there is significant evidence to suggest that the median ratings of users with no art background are not the same as the median ratings of users with some art background, even though the observed medians are both 4.0. This indicates that there might be an underlying difference between the two groups that is not captured by the observed median values.

Question 5: Regression model to predict art preference ratings from energy ratings

To build a reasonable prediction model, we used linear regression. As for cross-validation to avoid overfitting, we used a train-test-split test. Our predictor, X, is energy ratings, and our outcome, Y, is art preference ratings. The approach was to first join art ratings with energy ratings, drop NaN rows, and take the row mean (user means) of both user art ratings and user energy ratings. After this process, we are left with the predictor which is mean user energy ratings, and the outcome which is mean user energy ratings. At this point, we would want to split the data into an 80/20 split (80% training set and 20% test set) to assess the

generalizability of the model and overfitting. After this, we can fit the linear regression model with the training set (x_train, y_train). The plot of actual values and predicted values can be seen in Figure 5.

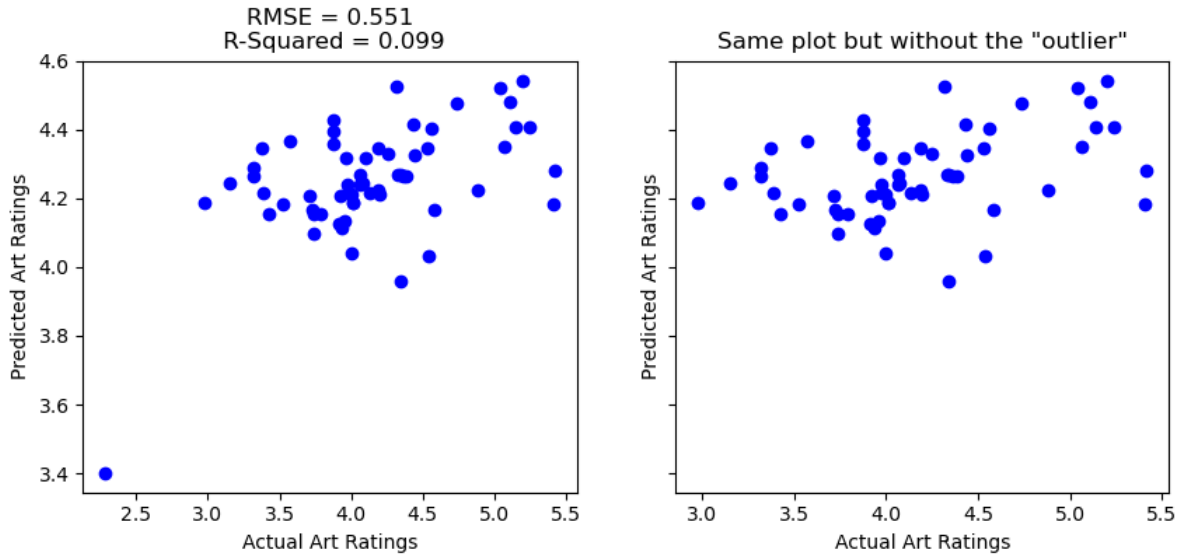

Figure 5: Actual Art Ratings Vs Predicted Art Ratings (**Right:** With "outlier"; **Left**: Without)

The chosen metrics to assess how the model performed is the root mean squared error (RMSE) and R-squared. The RMSE of our model is 0.55 which means that on average, the predicted values deviate from the actual values by approximately 0.55 units. R-squared, on the other hand, is approximately 0.1 which means that approximately 10% of the variance in the dependent variable (art preference) is explained by the independent variable (energy ratings) in the regression model. Notice that in Figure 5 (left), one of the data points was removed to better plot the data. That being said, the change in the RMSE was insignificant. Overall, the model does not have great explanatory power over the dependent variable.

Question 6: Regression model to predict art preference ratings from energy ratings and demographic information (age and gender)

For this question, the same algorithm for preprocessing as question 5 was followed. In this model, we have three predictors (X), average energy ratings, age, and gender, and one outcome which is average art preference ratings (Y). The graph of actual values on the x-axis and predicted values on the y-axis can be seen in Figure 6 below. As can be seen, the RMSE in this model is approximately 0.7 which means that on average, the predicted values deviate from the actual values by approximately 0.7 units. R-squared, on the other hand, is 0.14 which means that approximately 14% of the variance in the dependent variable (art

preference) is explained by the independent variables (energy ratings, age, gender) in the regression model.

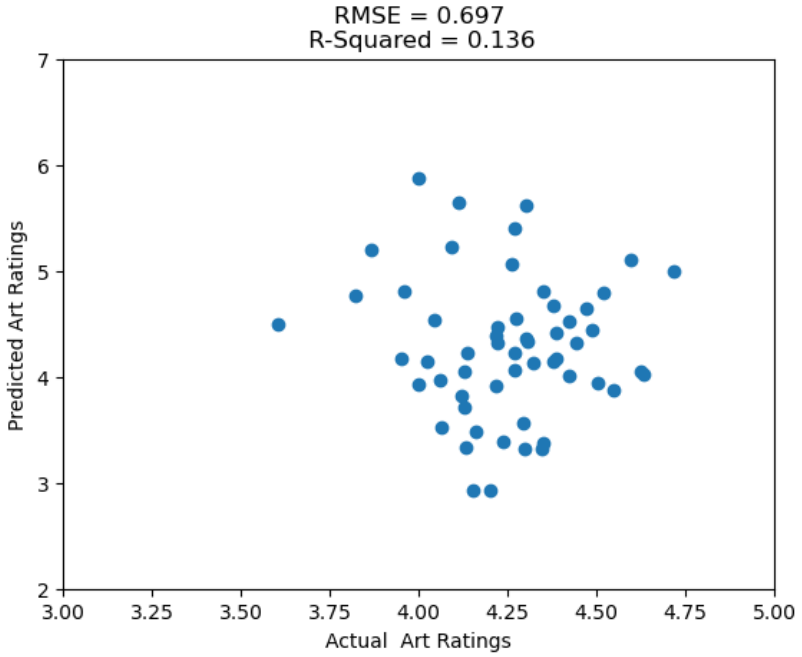

Figure 6: Actual Art Ratings Vs Predicted Art Ratings (Multiple Regression)

Overall, the improvement in R-squared in this model compared to the model in question 5 can be offset by the disimprovement in the RMSE.

Question 7: K-means clustering for average art preference ratings and average energy ratings

The goal is to identify clusters of average preference ratings and average energy ratings. One of the effective methods to do that is through K-means clustering, a form of unsupervised learning. The algorithm to do K-means clustering is as follows – first, join all predictors in one array/matrix. Second, find the optimal number of clusters. Lastly, do the actual clustering.

After concatenating the predictors (art and energy ratings) and dropping NaN rows, we can start finding the optimal number of clusters. To do so, we plotted the number of clusters on the x-axis and the corresponding sum of silhouette scores as seen in Figure 7 below. The conclusion is that the optimal number of clusters is K = 4 corresponding to the highest silhouette score.

Hamza Alshamy
DS 112-001
5/11/2023
Dr. Wallisch

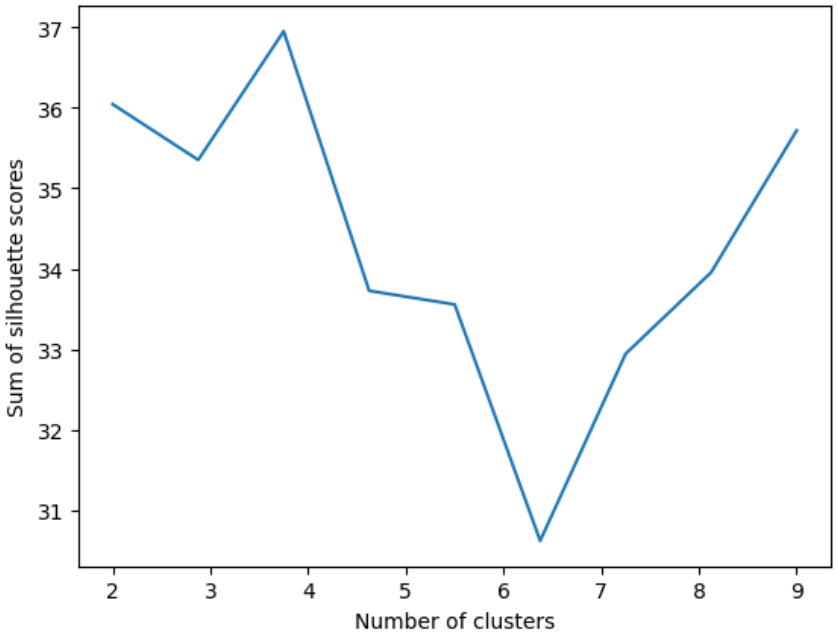

Figure 7: Number of clusters and silhouette scores

For the actual clustering, the algorithm minimizes the summed distances between a cluster center and its members. Once the minimum has been found (regardless of starting position), it stops. Doing the actual clustering result in Figure 8 below.

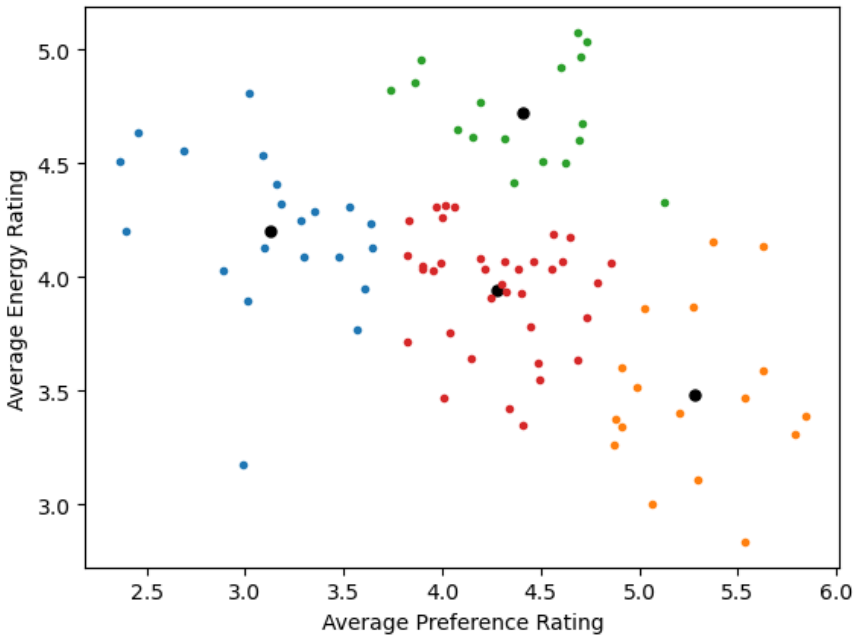

Figure 7: K-means clustering between art and energy ratings

We can interpret each of the clusters as follows. *Cluster 1 (Blue)* represents a low average preference rating and medium average energy rating. *Cluster 2 (Red)* represents a medium average preference rating and a medium average energy rating. *Cluster 3 (Greem)* represents a medium average preference rating and a high average energy rating. Lastly, *cluster 4 (Yellow)* represents a high average preference rating and a low average energy rating.

Question 8: Predicting art preference using the first principal component of self-image ratings

The process here is to first conduct the principal component analysis (PCA) to find the principal component (PC). Second, choose the PCs based on the Kaiser criterion and a Scree plot. However, in this question, we are only using the first PC. Third, interpret the PC. Lastly, fit a regression model with the PC and the outcome (art preference ratings). Looking at the Scree plot in Figure 8, we would choose 2 PCs, however, we are instructed to only use the first.

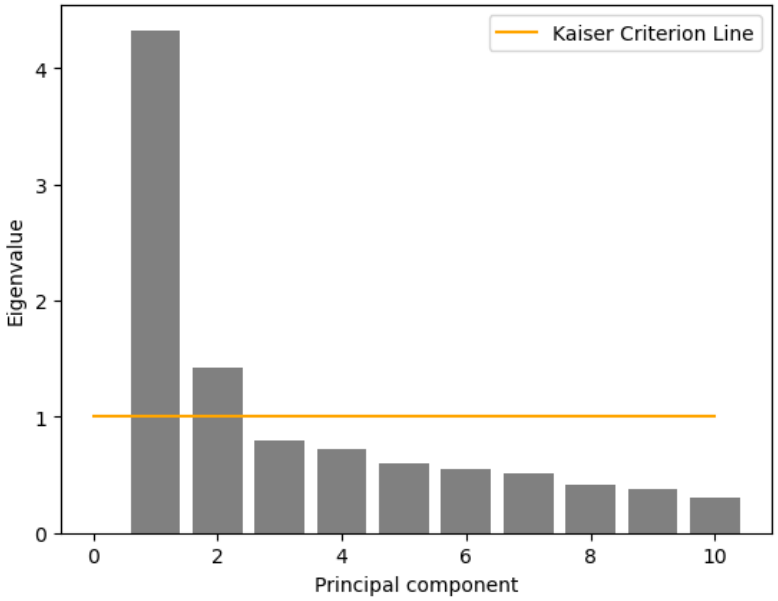

Figure 8: Scree Plot

To interpret the first PC, we must look at the loadings of the questions as in Figure 9 below. Since all the loadings are in a positive direction, and consulting with the original questions, we can interpret PC 1 as “Overall, I am happy with myself”. We would expect/hope that this question/PC captures much of the variance.

Hamza Alshamy
DS 112-001
5/11/2023
Dr. Wallisch

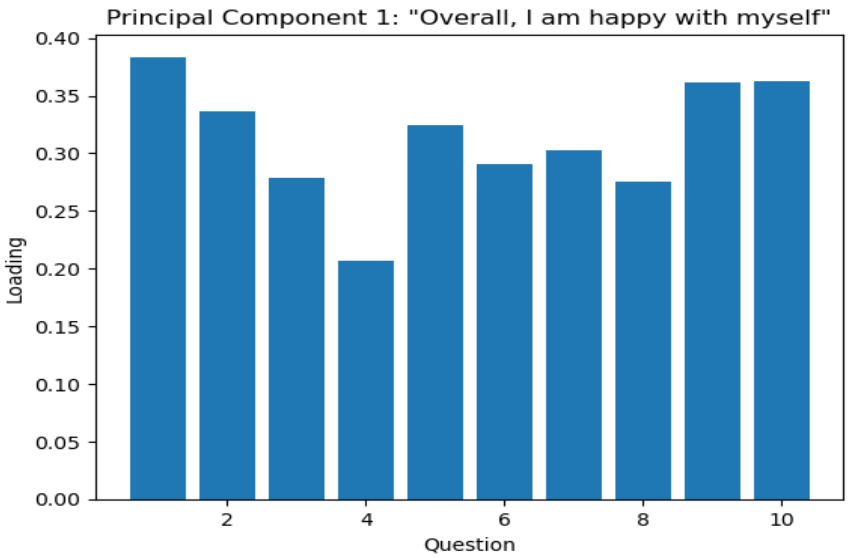

Figure 9: Loadings and interpretation of PC 1 of self-image

Finally, we can fit the regression model with the training sets composed of PC 1 as the predictor (X) and the art preference ratings as the outcome (Y). The actual art ratings compared to the predicted art ratings can be seen in Figure 10 below.

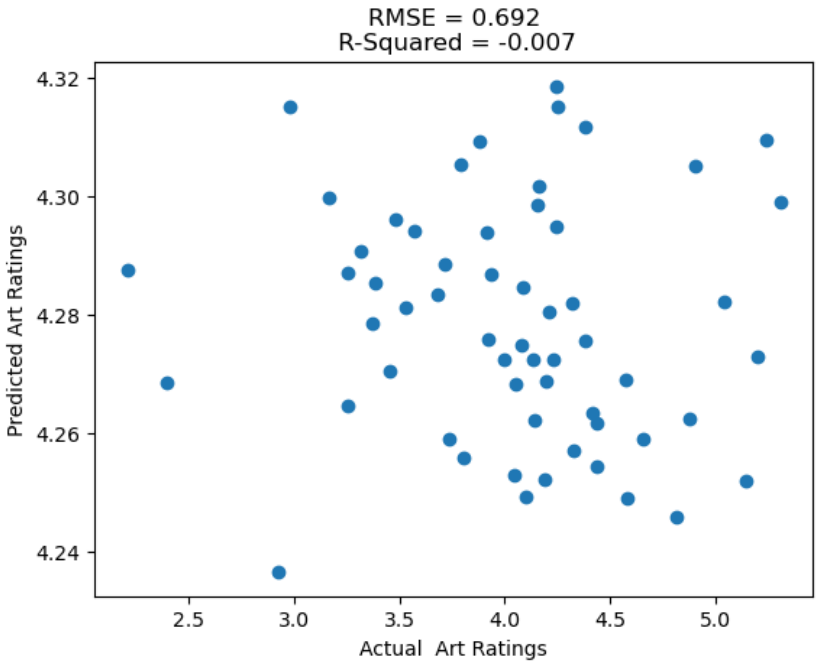

Figure 10: Actual Art Ratings Vs Predicted Art Ratings

As can be seen, the RMSE in this model is approximately 0.7 which means that on average, the predicted values deviate from the actual values by approximately 0.7 units. As for

R-squared, it is -0.007, approximately 0. However, the fact that it is negative indicates very poor predictive power.

Question 9: Predicting art preference using the first three principal components of dark-personality trait

We will follow the same process as in question 8. First, the Scree plot is represented in Figure 11 representing the amount of variance explained by each factor or principal component.

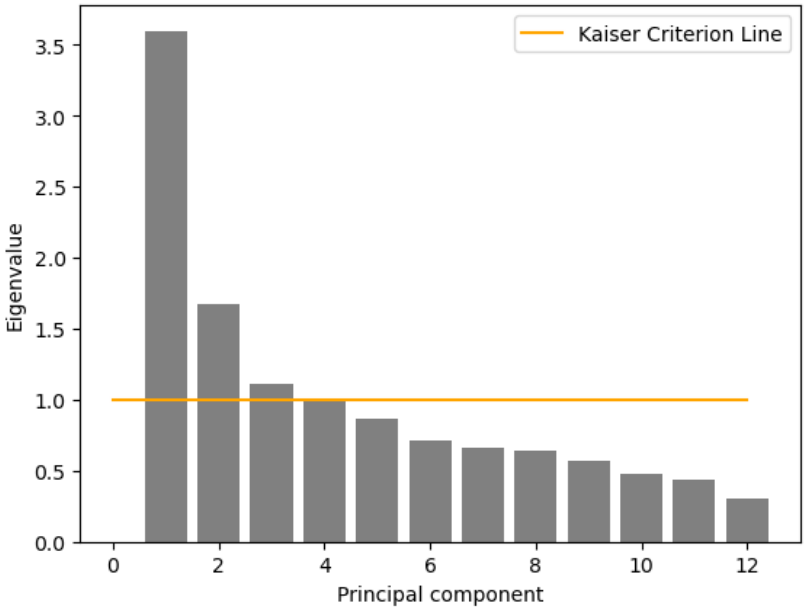

Figure 11: Scree plot

Based on the Kaiser criterion, we would choose 4 PCs (4 would be reasonable as well). However, this question asks us to conduct the analysis with 3 principal components. To interpret each of the three PCs, the loadings of each question are represented in Figure 12.

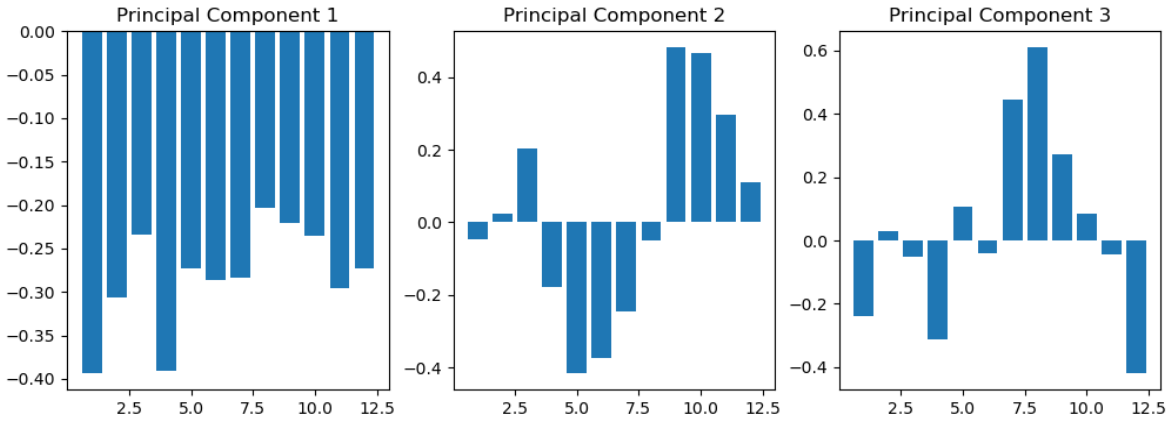

Figure 12: Scree plots for PC 1, 2, and 3

After consulting with the original questions, PC 1 is interpreted as “I tend to be benevolent”, PC 2 can be interpreted as “I tend to seek validation from people”, and lastly, PC 3 can be interpreted as “I tend to expect the worst from people”.

Finally, we can fit the regression model with the training sets composed of PC 1, 2, and 3 as the predictor (X) and the art preference ratings as the outcome (Y). Plotting the actual and predicted art preferences is seen in Figure 13 below.

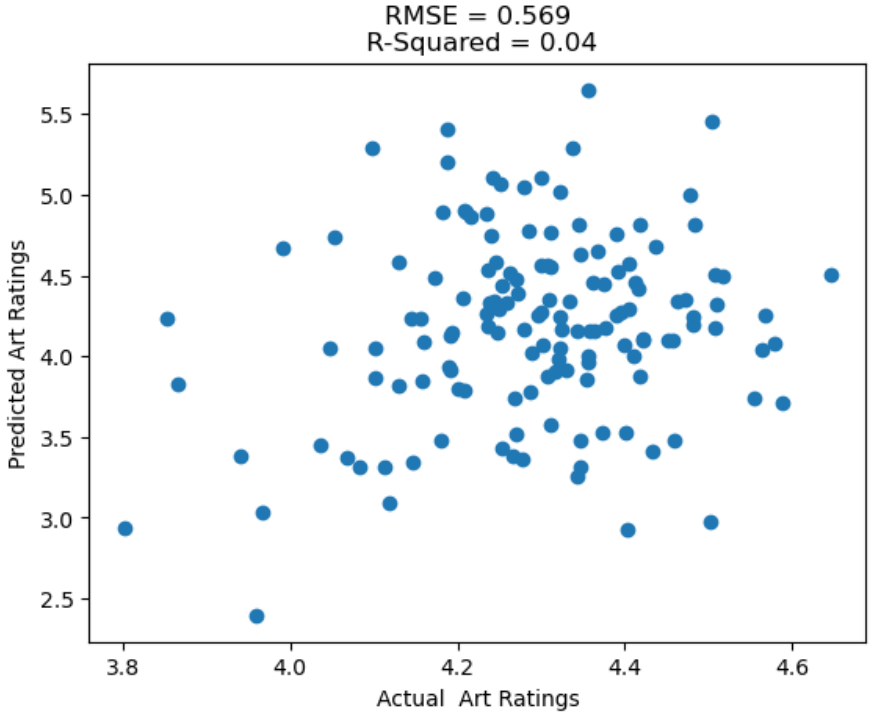

Figure 13: Actual Art Ratings Vs Predicted Art Ratings

As can be seen, the RMSE in this model is approximately 0.569 which means that on average, the predicted values deviate from the actual values by approximately 0.569 units. As for R-squared, it is 0.04 which means that approximately 4% of the variance in the dependent variable (art preference) is explained by the independent variables (PC1, PC2, and PC3) in the regression model. Overall, the predictive power of this model based on the metrics is relatively poor.

Question 10: Determine political orientation based on all available information

To determine political orientation, we simplified the values of the political orientation column to binary outcomes, “0” referring to non-liberal and “1” referring to liberal. Since we have

binary outcomes, logistic regression would be appropriate. First, we reduced the dataset from 221 columns to only 16 columns. The process involved the following – first, reducing art ratings to means (1), energy ratings to means (1), reducing "dark-personality" (3), self-image (2), and action (3) to principles components, joining remaining columns (6), and lastly, converting the political orientation column to binary. This process leaves a total of 1+1+3+2+3+6 = 16 columns. Thus, our predictor (X) is a total of 15 variables and our outcome is the binary political orientation column. (*Note: For all the details of PCA for dark-personality, self-image, and action columns, refer to the code file.)

At this point, we can fit the regression model with the training sets composed of the 15 columns as the predictor (X) and political orientation (binary) as the outcome (Y). The logistic regression plot can be seen in Figure 14 below.

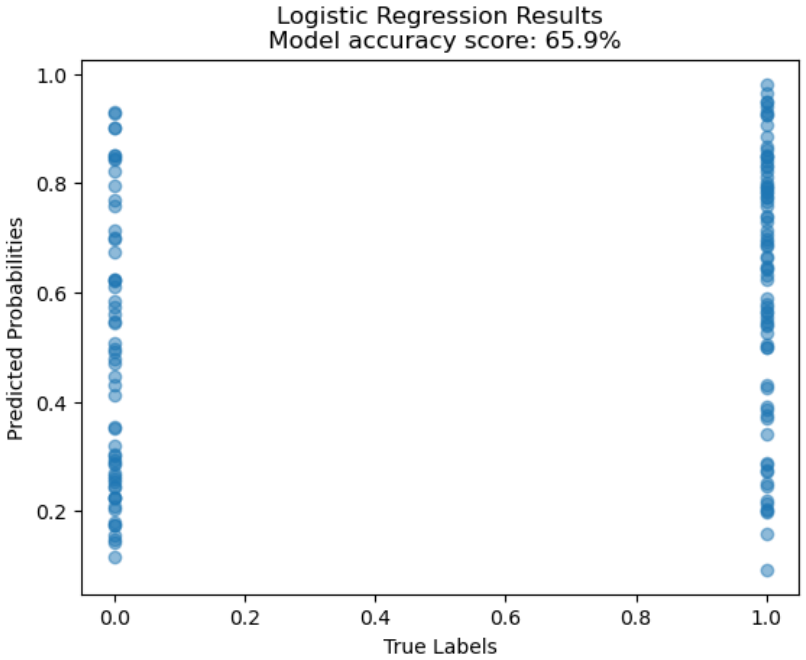

Figure 14: True labels Vs predicted probabilities

The model's accuracy is about 66%. If the model guessed everyone was liberal (1), its accuracy would be 57.0% ((number of liberal)/ (number of liberal + number of non-liberal)). Therefore, the model is better than just guessing.

Some further analyses on this question included RandomForest Classification which resulted in a 66% accuracy rate, and Adaptive Boosting (AdaBoost) with an accuracy of 57.25%. Observe that both logistic regression and random forest classification models gave around the same accuracy.

Hamza Alshamy
DS 112-001
5/11/2023
Dr. Wallisch

Glossary (*ChatGPT's Defenition****)***

1. **AdaBoosting (Adaptive Boosting)**: A boosting algorithm used as an ensemble method in machine learning to create a strong classifier from a number of weak classifiers.
2. **Alternative Hypothesis**: This is the hypothesis that is contrary to the null hypothesis. It is the hypothesis that the researcher is trying to prove.
3. **Cross-validation**: A statistical method used to estimate the skill of machine learning models. It is commonly used in applied machine learning to compare and select a model for a given predictive modeling problem.
4. **Hypothesis Test:** A statistical method that is used to make decisions or draw conclusions about populations based on sample data.
5. **K-means Clustering**: An unsupervised learning algorithm that groups similar data points together to discover underlying patterns.
6. **Null Hypothesis**: In statistical hypothesis testing, this is the hypothesis that there is no significant difference between specified populations, any observed difference being due to sampling or experimental error.
7. **Principal Component Analysis (PCA)**: A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
8. **Random Forest Classification**: A machine learning method that involves training multiple decision trees on various sub-samples of the dataset and using their average to improve the predictive accuracy and control overfitting.
9. **Root Mean Squared Error (RMSE)**: A measure of the differences between values predicted by a model and the values actually observed. It squares the differences before averaging them to penalize larger errors.
10. **R-squared:** A statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model.
11. **Scree Plot**: A plot used in exploratory data analysis to display the proportion of total variance in a dataset accounted for by each component in PCA.
12. **Significance Level**: The probability of rejecting the null hypothesis in a statistical test when it is true. It is denoted by the Greek letter alpha (α) and is also called the alpha level.
13. **Silhouette Score**: A measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation).
14. **Test Statistic**: A quantity derived from sample data to decide whether to accept or reject the null hypothesis in a hypothesis test.
15. **Unsupervised Learning**: A type of machine learning algorithm used to draw inferences from datasets consisting of input data without labeled responses.