Sean Wiryadi

DSUA112 Capstone Project

Preface:

The dataset provided for this project contains missing data, and as such, requires cleaning. Throughout my analysis, I handled the missing data by finding the null values in the data needed for computation and deleting the corresponding rows containing those null values. Additionally, for analysis that involves multiple variables that capture the same information, I applied a dimension reduction by performing a Principal Component Analysis. The purpose of this dimension reduction was to reduce the number of variables in the dataset while retaining as much of the original information as possible. This allowed me to more effectively analyze the data and make meaningful conclusions.

Question 1:

Distributions of the frequencies of the different ratings of art pieces were recorded for Classical, Modern and Non-Human Art.

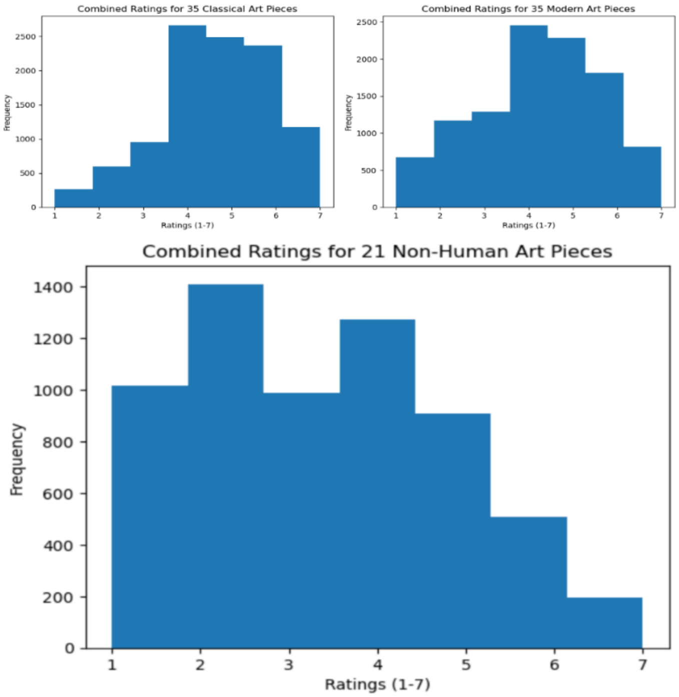


Figure 1: Distribution of frequencies of ratings

Based on the plots in Figure 1, the data are not normally distributed. Additionally, the data is comparing the art preference liking of individuals, comparing the means of the sample won’t make much sense. As a result, the medians of the sample will be used along with non-parametric test (Mann-Whitney U test) to answer the first 4 questions with a significance level of 0.05.

In the investigation to determine whether classical art is favored over modern art, the null hypothesis (H_0) is that the likability of classical and modern art is equal, whereas the alternative hypothesis (Hα) is that classical art is more preferred than modern art. The extremely small p-value (1.5881633286154516 × 10^(-97)), which is virtually zero, leads us to reject the null hypothesis. Therefore, we deduce that classical art is indeed more popular than modern art.

Question 2:

For the second question, the null hypothesis (H_0) suggested that the median preference for modern art is equivalent to that for non-human art, while the alternative hypothesis (Hα) says there is a difference between these two medians. The p-value of 8.742809791074804 x 10^(-264), being practically zero, clearly rejects the null hypothesis. Consequently, we infer a significant difference in the preference ratings between modern and non-human art.

Question 3:

In the third question, the 'user_gender' column in the dataset contained missing values and an extra 'Binary' category besides 'Men' and 'Women'. To ensure the accuracy of results, we excluded rows with missing values and the 'Binary' category. The median ratings for both men and women turned out to be 4. The null hypothesis (H_0) is that women rate art similarly to men, while the alternative hypothesis (Hα) is that women rate art higher than men. With a p-value of 0.1356455438, which is greater than 0.05, we do not have sufficient evidence to reject the null hypothesis. Thus, it appears that women and men give comparable preference ratings.

Sean Wiryadi

Question 4:

Regarding the fourth question, we grouped individuals with a '0' in the 'art_education' column as having no art experience, and those with non-zero values as having some art experience. The null hypothesis (H_0) stated that there is no difference in art preference between those with no art education and those with some art education, whereas the alternative hypothesis (Hα) proposed a difference. The p-value (1.0118570941459344 x 10^(-8)) being significantly less than 0.05, rejects the null hypothesis, suggesting a notable difference in art preference between the two groups.

Question 5:

A regression model was developed to predict preference ratings based on energy ratings. Preference ratings encompassed columns [1:91], and energy ratings covered columns [92:182]. The datasets X (Energy ratings) and Y (Preference ratings) were divided into two subsets: a Training set and a Test set. This division is critical to avoid overfitting and to subsequently evaluate the model's performance. A random 80% of the data was allocated to the Training set, with the remaining 20% assigned to the Test set. The resulting model displayed an R-Square score of -0.53386292, indicating a poor fit, performing worse than a model that always predict the means of the preference ratings. Additionally, the model reported a root mean squared error of 1.83693, which is considerably high given the 1-7 rating scale. This suggests that energy ratings are ineffective in predicting preference ratings, as evidenced by the low R-Square and high RMSE.

Question 6:

Demographic information comprised Age, Gender, and Political Orientation. We excluded any missing values in the dataset. These predictor variables were then consolidated with the energy rating values into a single dataframe. The same 80-20 split was used for the training and testing data, mirroring the approach in the fifth question. However, the inclusion of demographic information yielded an even worse model, with an R-Square score of -0.803011 and a Root Mean Squared Error of 2.00792. This suggests an average error of approximately 2 ratings on a 1-7 scale, equating to a 30% error rate each time the model is utilized to forecast art preference ratings. Therefore, the model's performance deteriorated with the inclusion of demographic data.

Question 7:

In order to reduce the data into a 2D Array, I decided to take the means of the Preference rating columns and Energy rating columns. In order to identify the clusters, KMeans would be used. Subsequently the sum of silhouette scores is used to identify the optimal number of clusters via the algorithm.

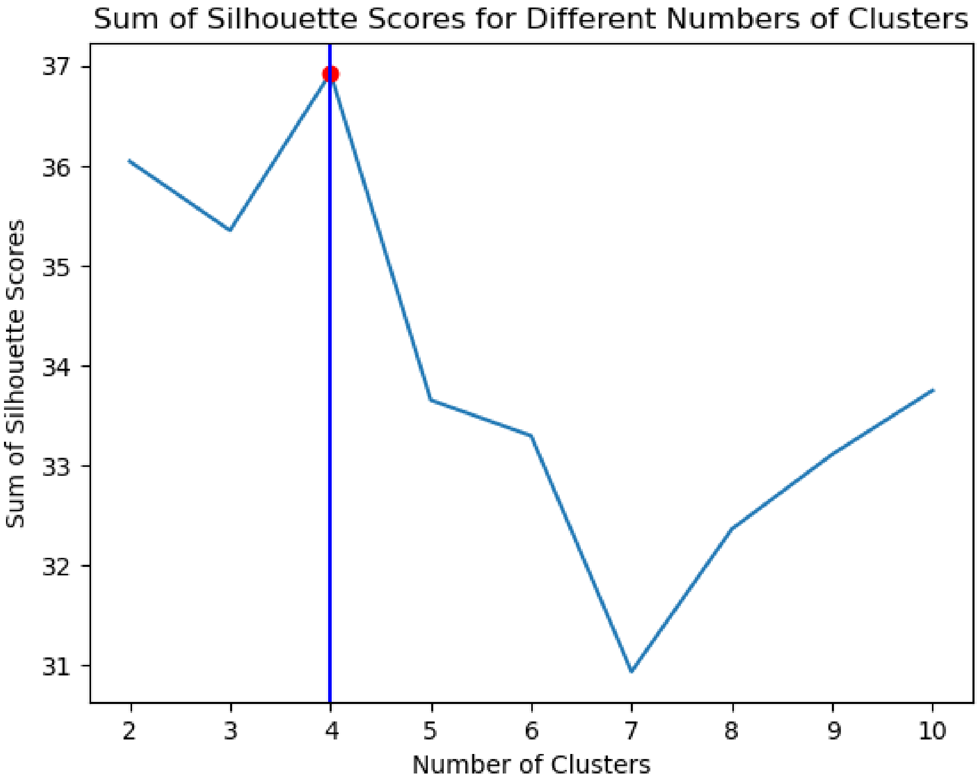


Figure 2: Sum of silhouette scores for different number of clusters

The optimal number of clusters is when the sum of the silhouette scores is the highest among the Number of clusters tested which is 4 with a sum of (36.9495140278378).

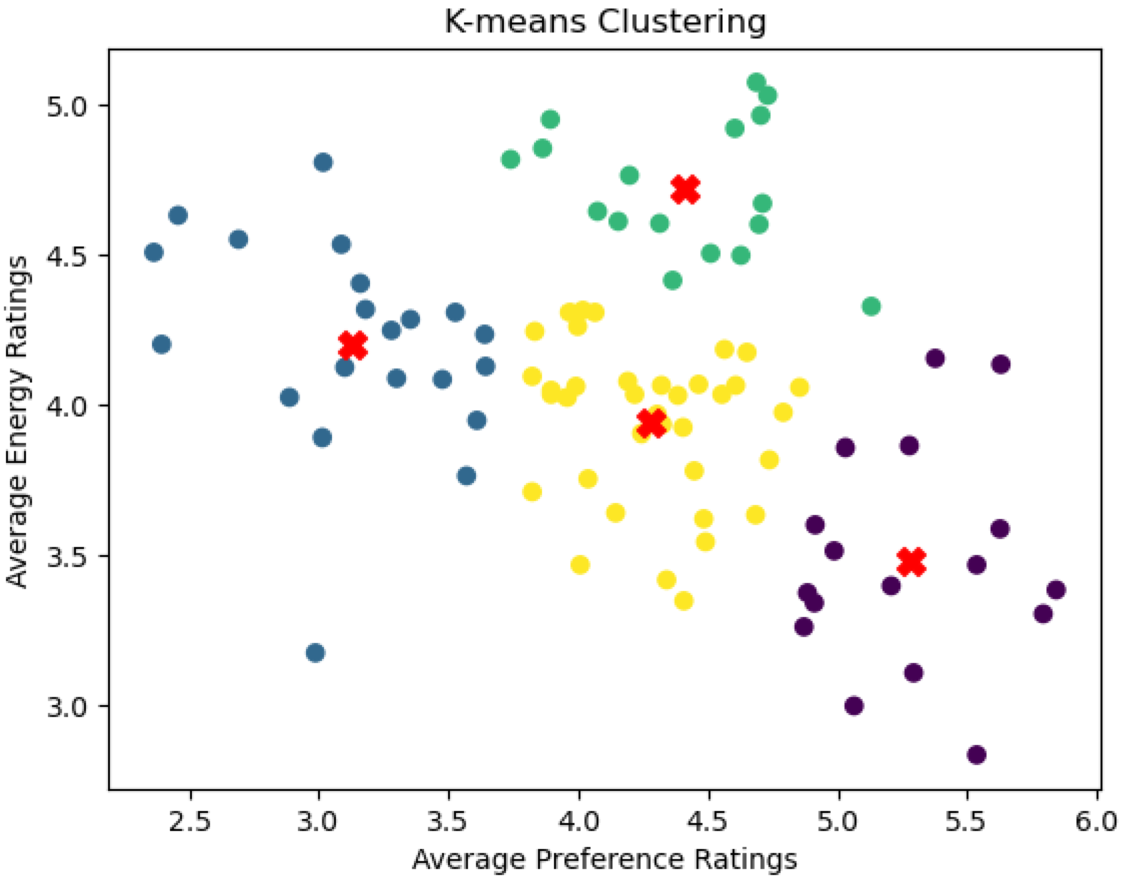
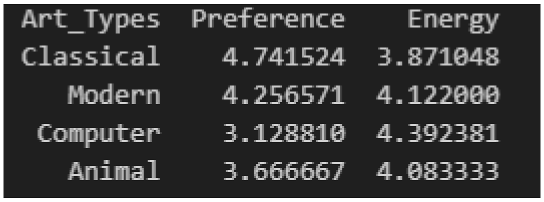


Figure 3: Clustering scatterplot



```
Art_Types  Preference    Energy
Classical    4.741524  3.871048
   Modern    4.256571  4.122000
 Computer    3.128810  4.392381
   Animal    3.666667  4.083333
```

Figure 4: Dataframe of mean preference and energy ratings

I computed the mean preference ratings for each category of art present in the dataset: Classical, Modern, and NonHuman (Computer and Animal). The Classical art category, represented by the purple cluster in the visual representation, has the highest mean preference rating. The Computer art category, signified by the blue cluster, aligns with the mean preference rating range of 2.5 to 3.5, mirroring the calculated mean preference rating of 3.12. Distinguishing between Modern and Animal art proved challenging due to the similarity in their data points. However, upon further analysis, I associated the green cluster with Modern art as its mean preference rating was noticeably higher compared to Animal art. This, in turn, led to the identification of the yellow cluster as Animal art. Although this identification of clusters is based on the mean ratings and patterns observed in the clusters, it's crucial to note the possibility of discrepancies. Specifically, the number of data points in each cluster may not correspond perfectly to those in the dataset.

Question 8:

Self-Image ratings has 10 predictors, a correlation matrix was carried out in order to test if there are any clusters between the predictors. Based on Figure 5, there is no evidence of a cluster but evidence of correlation hence allowing us to appropriately perform PCA analysis. The dataset contained missing values, so rows of missing values were dropped in order to accurately get a representation of the response of self-image ratings.

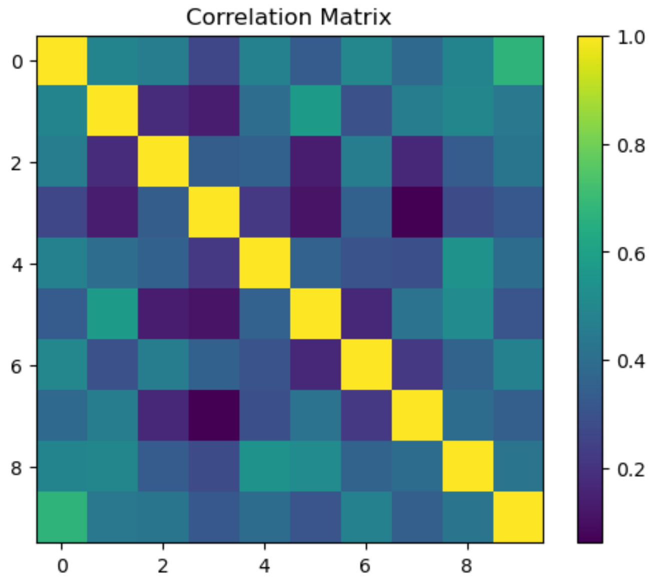


Figure 5: Correlation Matrix of 10 Questions about self-image

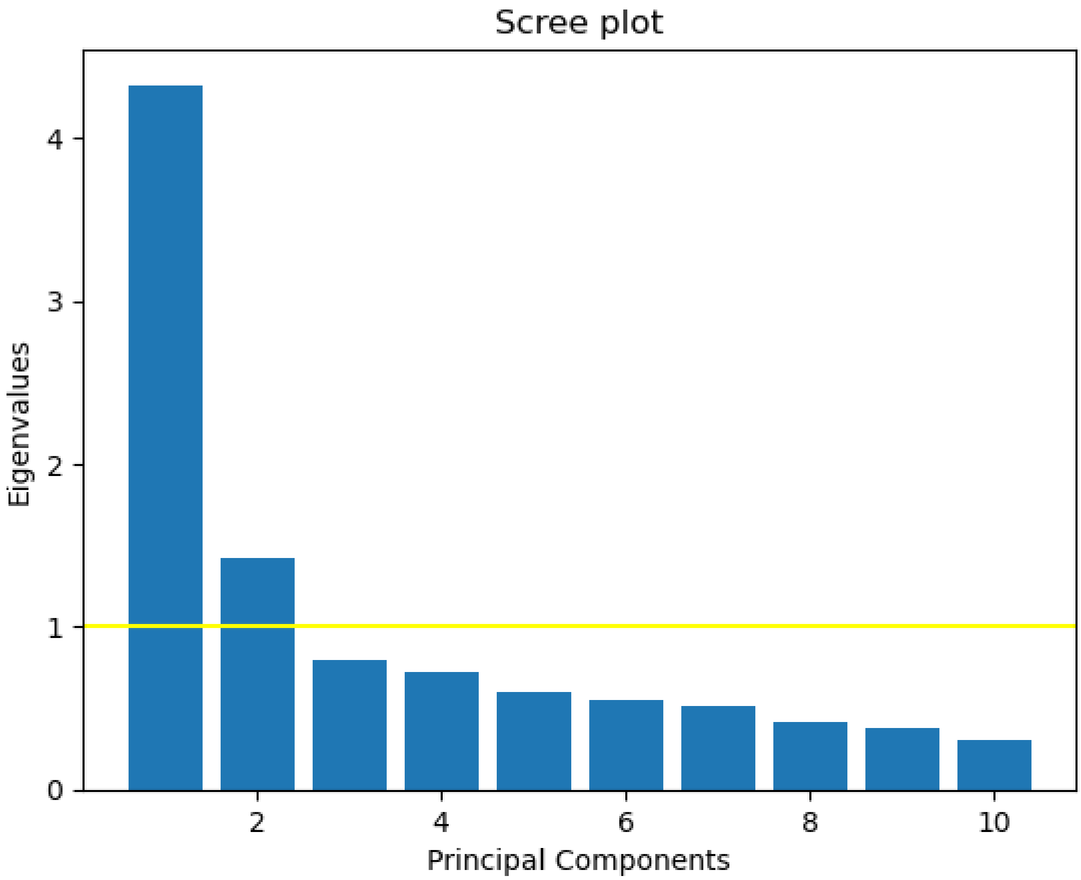


Figure 6: Scree Plot of 10 Questions about self-image

We found that by the Kaiser criterion that there are 2 columns which have eigenvalues over 1. However, the question states to only consider the first principal component alone and thus only the first will be used in order to build the regression model to predict art preference ratings. To facilitate the operation of the linear regression with Python's scikit-learn module, a reduction in the number of columns by taking the mean was necessary to achieve a [286x1] dimensionality. The model's effectiveness was evaluated using the Root Mean Squared Error (RMSE) metric, which yielded a value of 0.636667411. Given that the art preference ratings range from 1 to 7, the deviation of 0.636667411/7, representing approximately 10% of the rating range, suggests that the model may not be highly precise. In addition to this, an 'Index versus Residuals' scatterplot was created. As depicted in Figure 7, a substantial number of data points exhibit errors close to the zero-horizontal line, indicating that the model makes reasonably accurate predictions for these points. Nonetheless, a significant number of data points show large errors, suggesting that the model's predictions for these points are far off. Given the rating scale (1-7), the wide range of errors from -3 to 1.5 points towards potential shortcomings in the model.

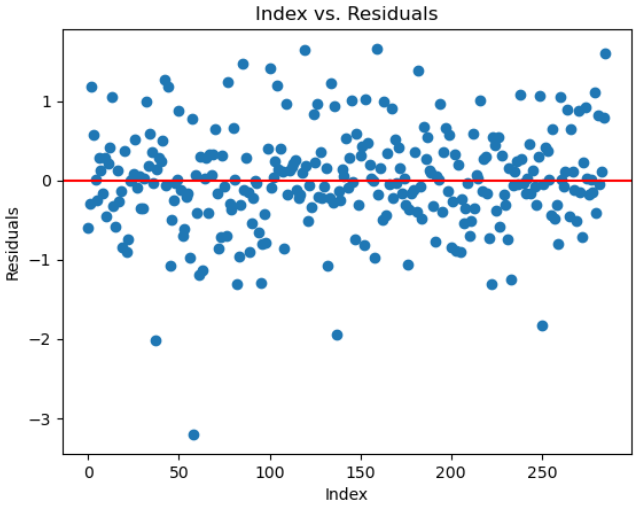


Figure 7: Scatterplot of Residuals

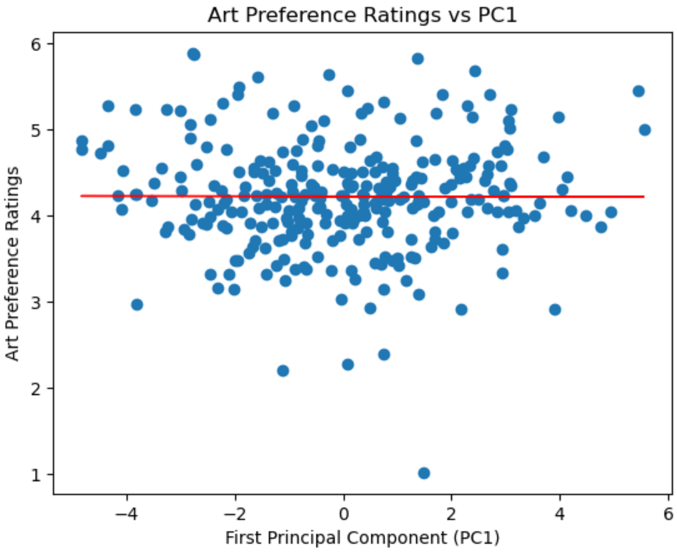


Figure 8: Scatterplot with best fit line

Sean Wiryadi

Question 9:

As per Figure 9, the correlation matrix does not exhibit any discernible clusters. To reduce the dimensionality of the various factors, a Principal Component Analysis (PCA) was performed. The Kaiser criterion identified three columns with eigenvalues exceeding one. As a result, these three columns were utilized to establish a regression model to predict art preference ratings. Similar to the question 8, the model's R-squared value is 0.02474864 and the Root Mean Squared Error (RMSE) is 1.46. These metrics indicate that only 2% of the variability is explained by the three factors derived from the PCA. Furthermore, the high RMSE of 1.46, relative to the rating scale, suggests a high average error rate for the model.

To test the significance of the predictors, a hypothesis test was conducted for the coefficients of the predictor variables. A loop was created to establish an Ordinary Least Squares (OLS) model for each art column against the three predictor variables, and to derive the p-values of each coefficient. The null hypothesis (H_0) states that components 1, 2, and 3 do not predict art preference rating, whereas the alternative hypothesis ($H\alpha$) posits that these components significantly predict the art preference rating. As all the p-values of the coefficients are less than 0.05, the null hypothesis is rejected in favor of the alternative hypothesis, asserting that all three components significantly predict all 91 art preference ratings.

To infer the potential identities of the factors extracted from the PCA, we examined the questions with the highest loadings. Based on Figure 11, the first factor is primarily associated with Questions 1 and 4, which suggest a personality trait of "**Manipulativeness**". The second factor is linked with Questions 5, 6, and 7, which imply a "**Narcissistic**" personality. Lastly, the third factor is tied to Questions 1, 4, and 12, which infer an "**Entitled**" personality.

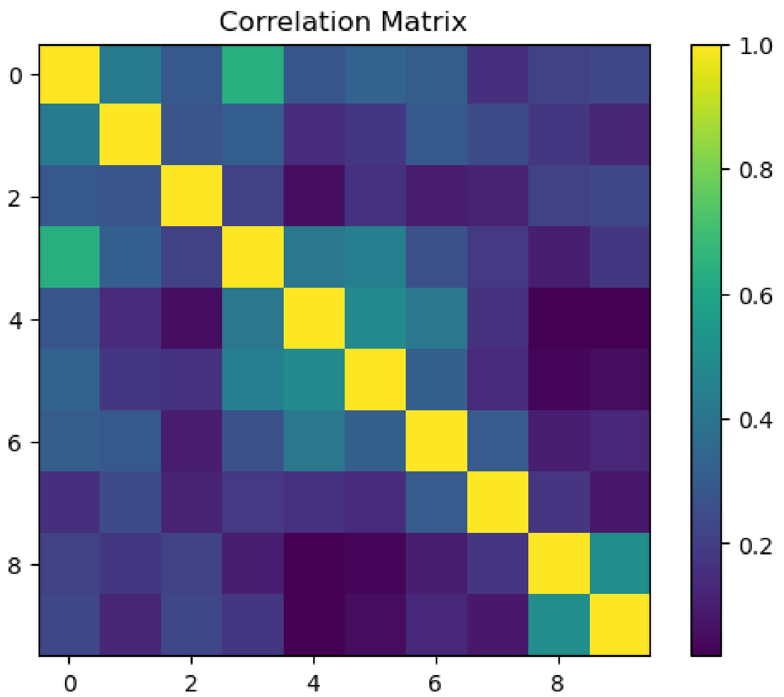


Figure 9: Correlation matrix of Dark personality traits

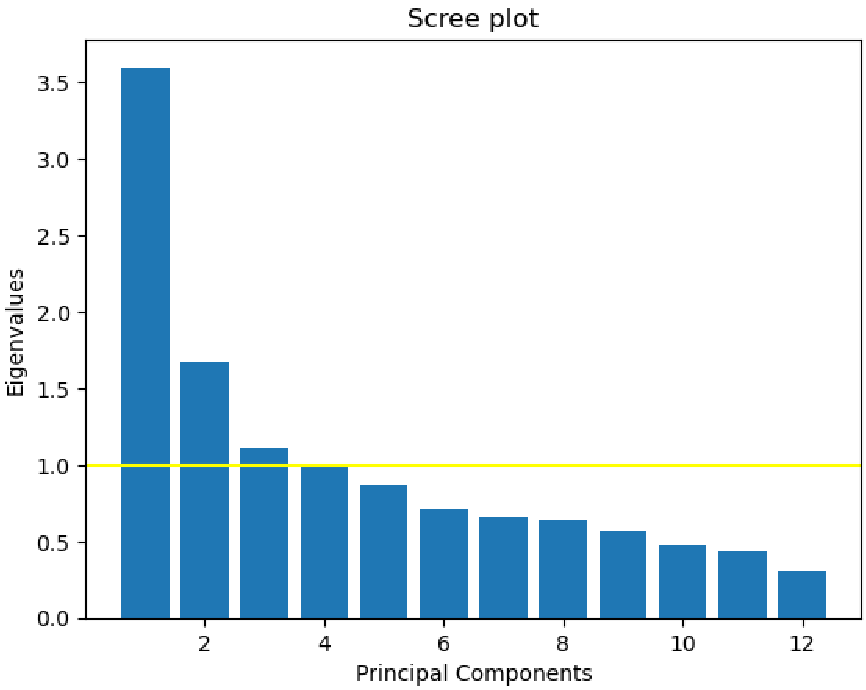


Figure 10: Scree Plot of Dark personality traits

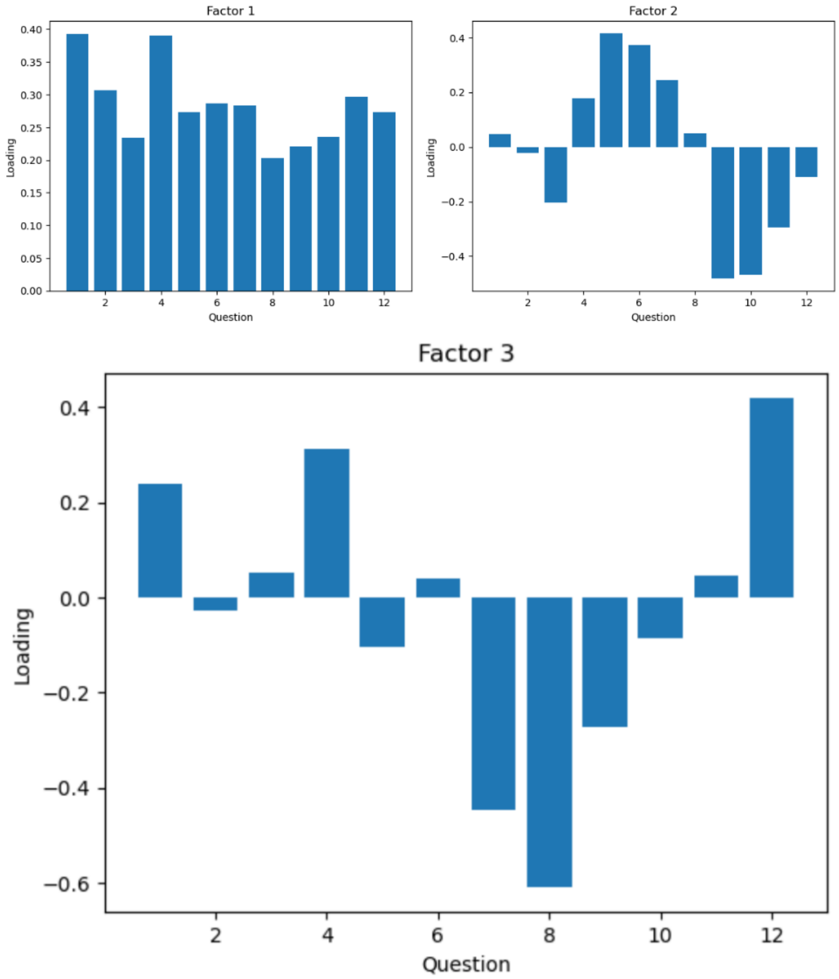


Figure 11: Question against loadings for Factor 1,2,3

Question 10:

The logistic regression model was implemented with X representing all columns except "Political_Orientation", and y representing the "Political_Orientation" column. As depicted in Figure 12, the model resulted in 35 True Positives, 13 False Negatives, 15 False Positives, and 20 True Negatives. An analysis of the classification report revealed the model's performance. The model's accuracy stands at 66%, denoting it performs moderately better than sheer guessing. The F1-scores for classes 0 and 1 are 0.71 and 0.59, respectively. These scores suggest a satisfactory trade-off between precision and recall for both classes, albeit with class 0 exhibiting slightly superior performance. Moreover, the ROC value of 0.72 implies that the model possesses a decent capability to distinguish between positive and negative instances.

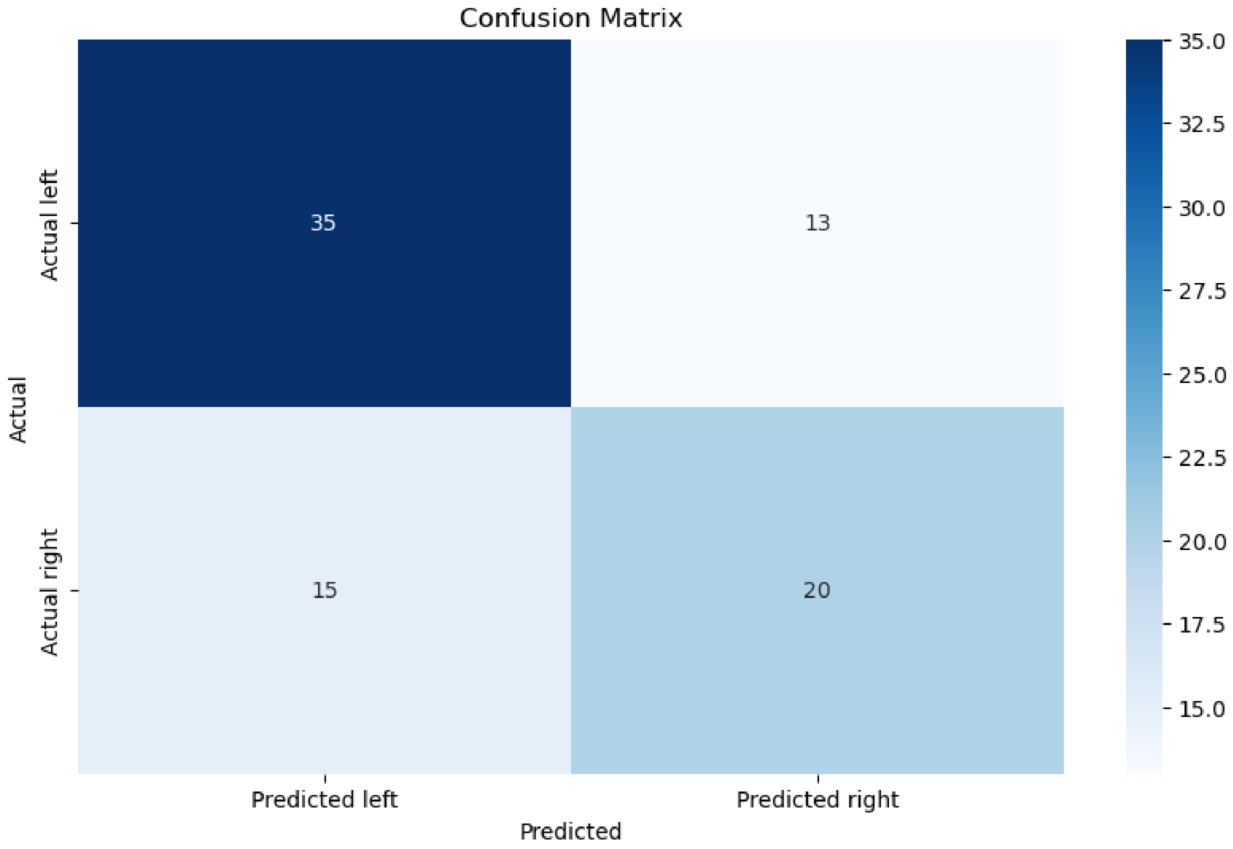
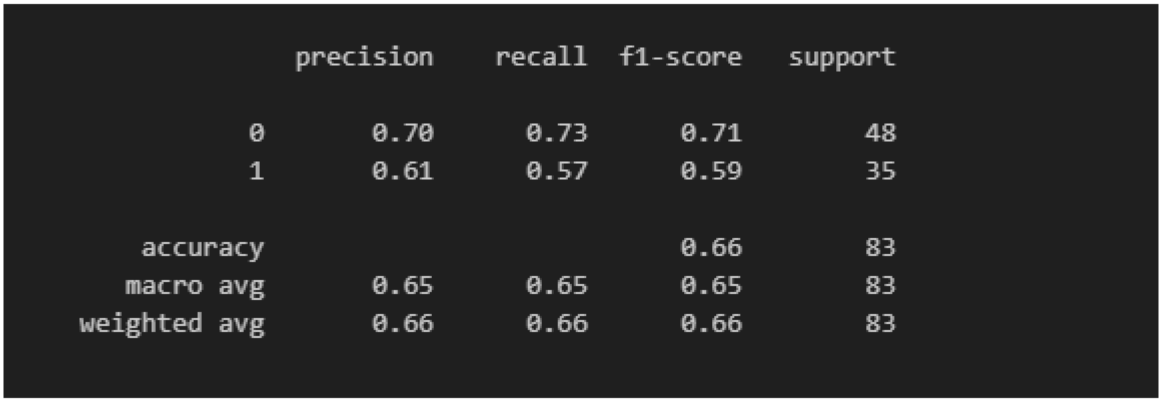


Figure 12: Confusion Matrix



```
              precision    recall  f1-score   support

           0       0.70      0.73      0.71        48
           1       0.61      0.57      0.59        35

    accuracy                           0.66        83
   macro avg       0.65      0.65      0.65        83
weighted avg       0.66      0.66      0.66        83
```

Figure 13: Classification Report

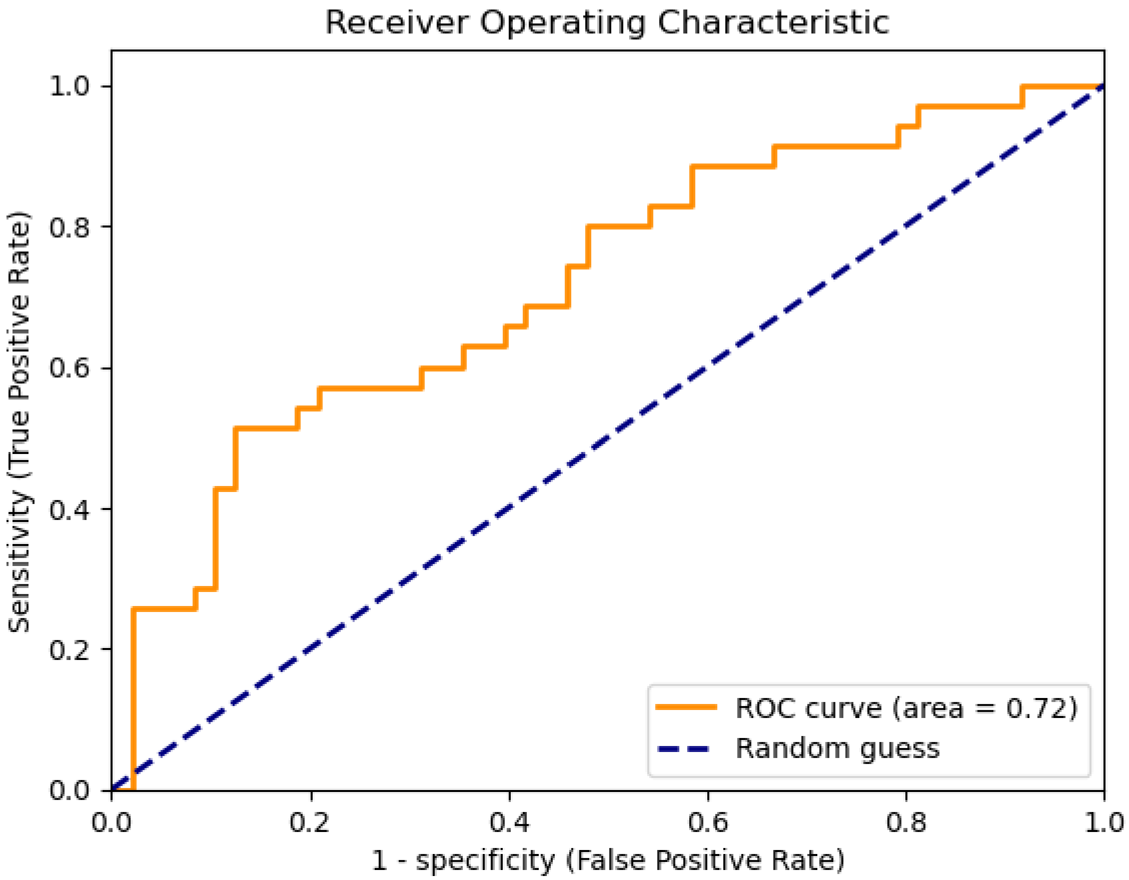


Figure 14: ROC Curve