Naufil Bin Imran, N16790632

Pascal Wallisch

Intro to Data Science

12 May 2023

Capstone Project

Data Cleaning and handling nan values: Overall, the data did not have many nan values. However, there were some nan values. To use all of the data available whenever possible, I handled nan values by making subset data frames for each question from my main data frame that had all the data. To remove the nan values in the required questions, I used row-wise removal because only a tiny proportion of rows were missing, which could have had little impact on our analysis and the problem we were solving. Since we had 300 user ratings, in most cases, the data loss after row-wise nan removal was 10-20 rows, which is a deficient proportion of user ratings, and that is how nan values were handled.

Dimension reduction: To reduce the dimensionality of data wherever required, I used principal component analysis and calculated and graphed eigenvalues. Overall I used the Kaizen criterion for eigenvalues>1 to reduce the dimensionality of my dataset.

Data Transformation: For PCA, it was essential to standardize our data, so Z scores were used to standardize data wherever needed.

Before we start the answer questions, it is essential to understand what Python libraries were used in this project. This is a snapshot of what libraries were used in this project.

.

In [689]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import pingouin as pg
import scipy.stats as stats
from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
import seaborn as sns
from sklearn import metrics
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_samples
from sklearn.cluster import DBSCAN
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
from scipy.stats import zscore
from sklearn.decomposition import PCA
from sklearn.metrics import r2_score
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import roc_auc_score

#here we import all the libraries that we would be using for the whole project.
```

Question 1

For Q1, the null hypothesis was that there is no difference in the liking of classical art and modern art. However, the alternative hypothesis was that classical art is more liked than modern art. To do this question, I used a statistical test, it was necessary first to understand which parametric or nonparametric test must be used. To check for this, I reduced my data to sample means and plotted a normal distribution graph; however, the distribution was not expected; thus, I decided to use nonparametric tests. I had two data frames that individually stored the ratings from 300 users for modern and classical art. I used the Wilcoxon rank signed test as our sample sizes were small, and we were comparing if modern art or classical art is more liked, from the ratings from the same users, since we have two related groups of data, it was reasonable to use the Wilcoxon test. Through the test, we found that the p-value was less than 0.05, which means that there it is unlikely that this is due to chance alone, and we can reject the null hypothesis and conclude that classical art is more liked than modern art. I also drew a bar plot comparing the two means of all ratings of 300 users and modern and classical art and found that classical art has a higher mean.

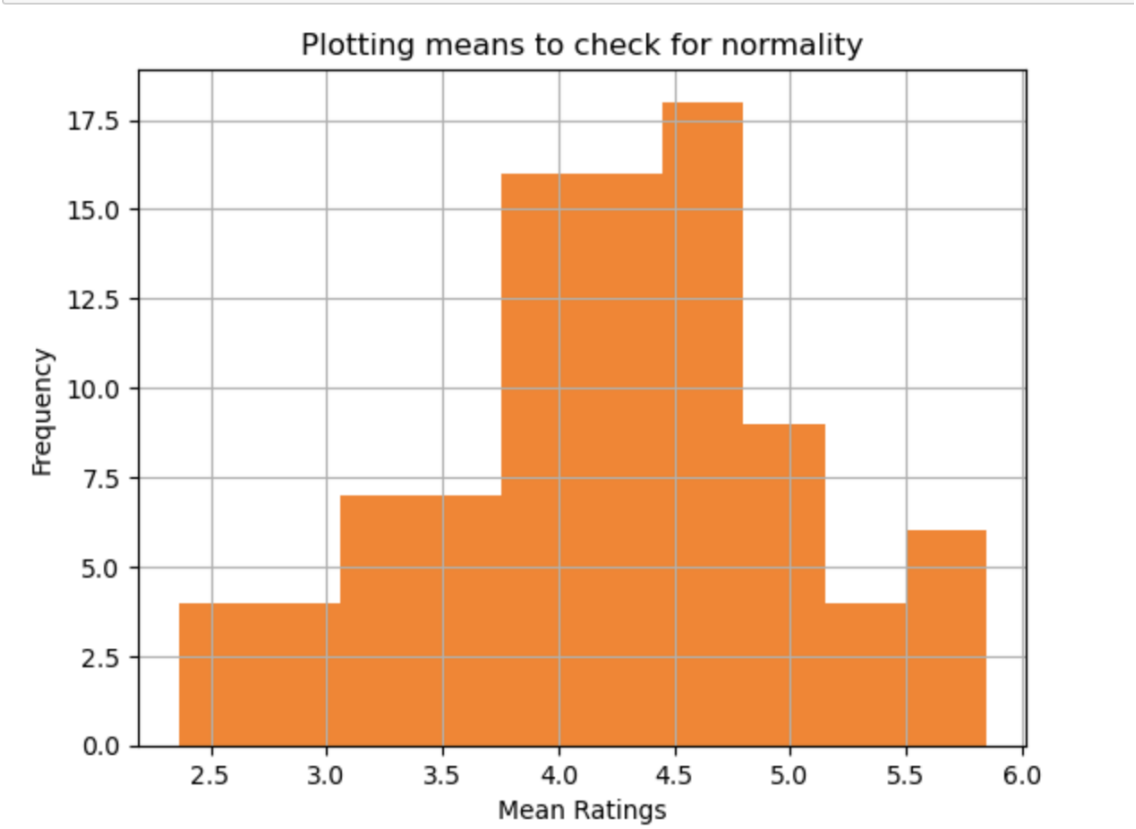

Figure 1 : This histogram shows that we could not find our data to be expected, that is, for all questions 1-4, nonparametric tests were used.

```
p-value: 3.208667577864861e-117
```

Figure 2: The P-value for the question which is less than 0.05, so we reject the null hypothesis.

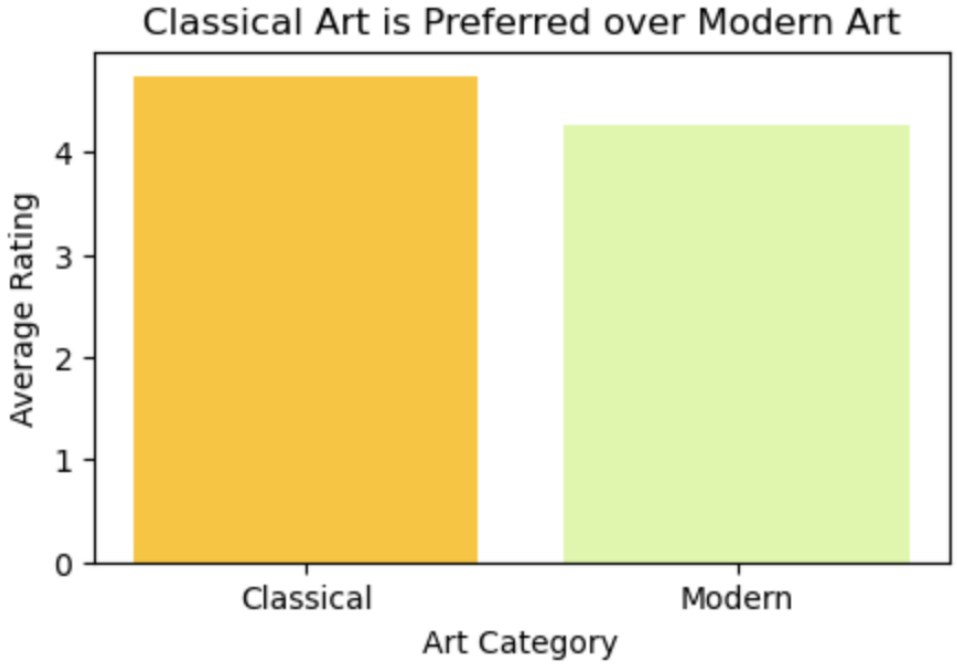

Figure 3: The bar chart showing the mean ratings for classical art is more than modern for the ratings provided by 300 users.

Question 2

For question 2, our null hypothesis was that there is no difference in preference for modern art vs. nonhuman art. The alternative hypothesis is that there is a difference between the preference. To test this hypothesis, we first followed that the data is not generally distributed, so we used nonparametric tests. To do nonparametric tests, we created two data frames one that had data of 300 users of all current data, and the second one had 300 user ratings for the nonhuman art pieces. The second data frame was created in two steps which can be seen in the code. To find the p-value, we used Mann Whitney U Test. We used the Man Whitney U test in this case because the data is not categorical, and only two groups are being compared in the design. It is reasonable to compare their central tendency as we have two independent groups. We get a p-value lower than 0.05, so we reject the null hypothesis that there is no difference in preference for modern vs. nonhuman art. If we compare the means for all 300 users for both groups, we find that the modern has a higher mean. This is shown in the bar chart below.

```
the p value for question 2 is, 8.742809791074804e-264
```

Figure 4: The p-value for the man Whitney u test we do in question 2

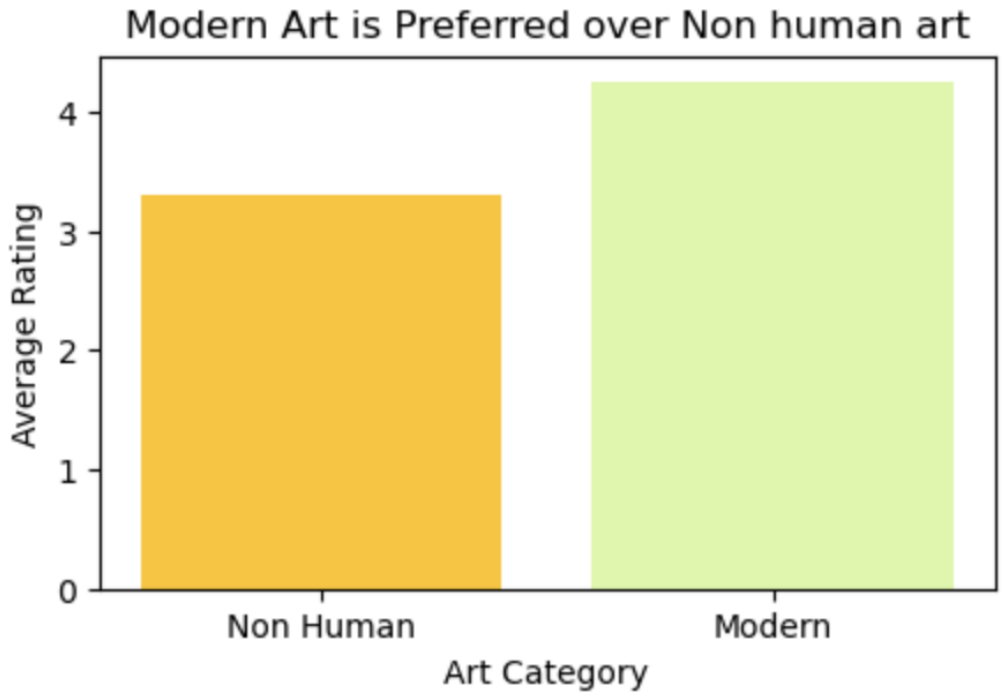

Figure 5: Showing the mean of average ratings of 300 users for modern art is more than non-human.

Question 3

To do Q3, we used statistical tests. Due to Figure 1, we concluded that data is not normally distributed, and we used nonparametric tests. To run a nonparametric test, we divided our data into two data frames, one with a rating of 95 males rating 91 art pieces and the other having data of 179 female ratings for 91 art pieces. We used Man Whitney u test, with the null hypothesis being that both men and females have the same art preferences and the alternative hypothesis being women have higher art preferences. We decided to use Man Whitney U Test as the data was not categorical, and it was not reasonable to reduce to sample means, we were comparing two groups, and the groups were independent of each other. We got a p-value greater than 0.05, which means we failed to reject the null hypothesis and do not have sufficient evidence that females give higher art preference than males.

```
Mann-Whitney U test statistic: 70994989.5
p-value: 0.2712910876266065
```

Figure 6: p-value is higher than 0.05 thus, we fail to reject the null hypothesis.

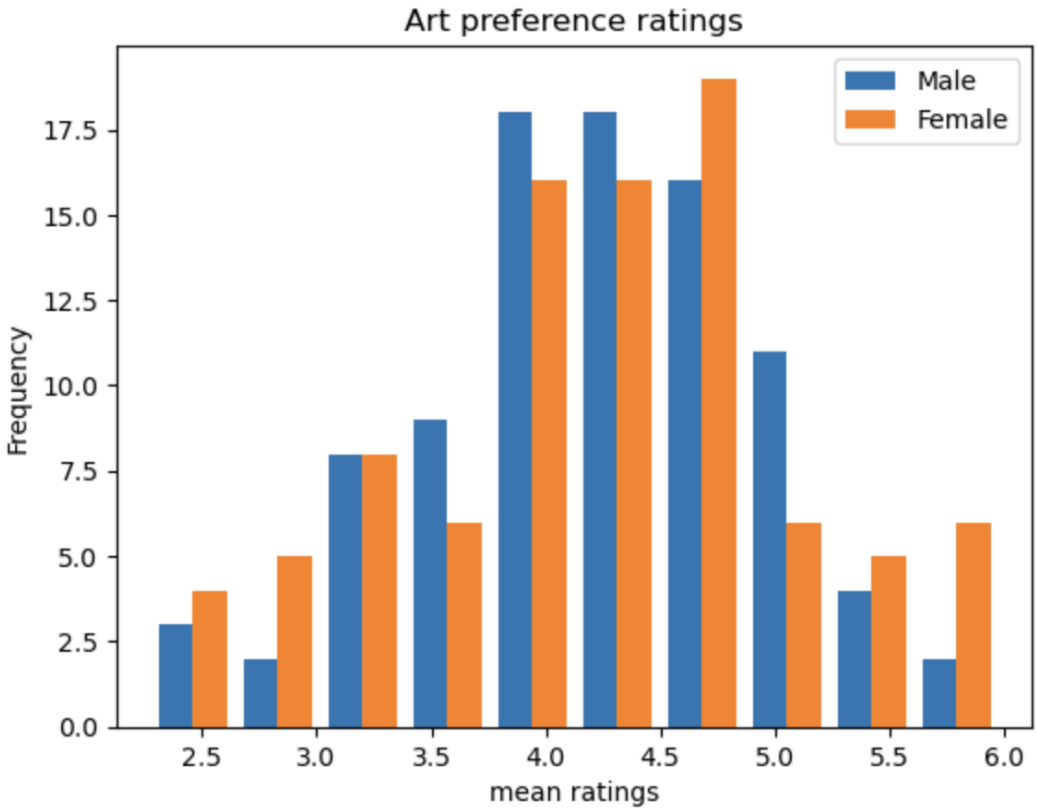

Figure 7: the difference between the mean ratings for men and female. This figure also proves our point that we can not conclude that females have a higher preference for art.

Question 4

The null hypothesis was that there is no difference in the preference ratings between people with art backgrounds and those without art knowledge. The alternative hypothesis was there was a difference in the preference ratings of people with some art knowledge. We decided to do this question using a nonparametric statistical test, as the data was not reduced to sample means. The data was divided into two data frames, one containing users with art knowledge and that without art knowledge after the row-wise removal as there were some nan values. After that Man Whitney U test was used as the data was not categorical and was independent as there were two groups, one with art knowledge and one without it. The central tendency was a good way to test our hypothesis. The results for man Whitney u test to show that the p-value was lower than 0.05, which means we can reject the null hypothesis that there is no difference in preference for art between people with some knowledge of art and people without knowledge of art.

```
Mann-Whitney U test statistic: 68504156.0
p-value: 1.0118570941459344e-08
```

Figure 8: Test statistic and p-value for our man Whitney u test done in Python.

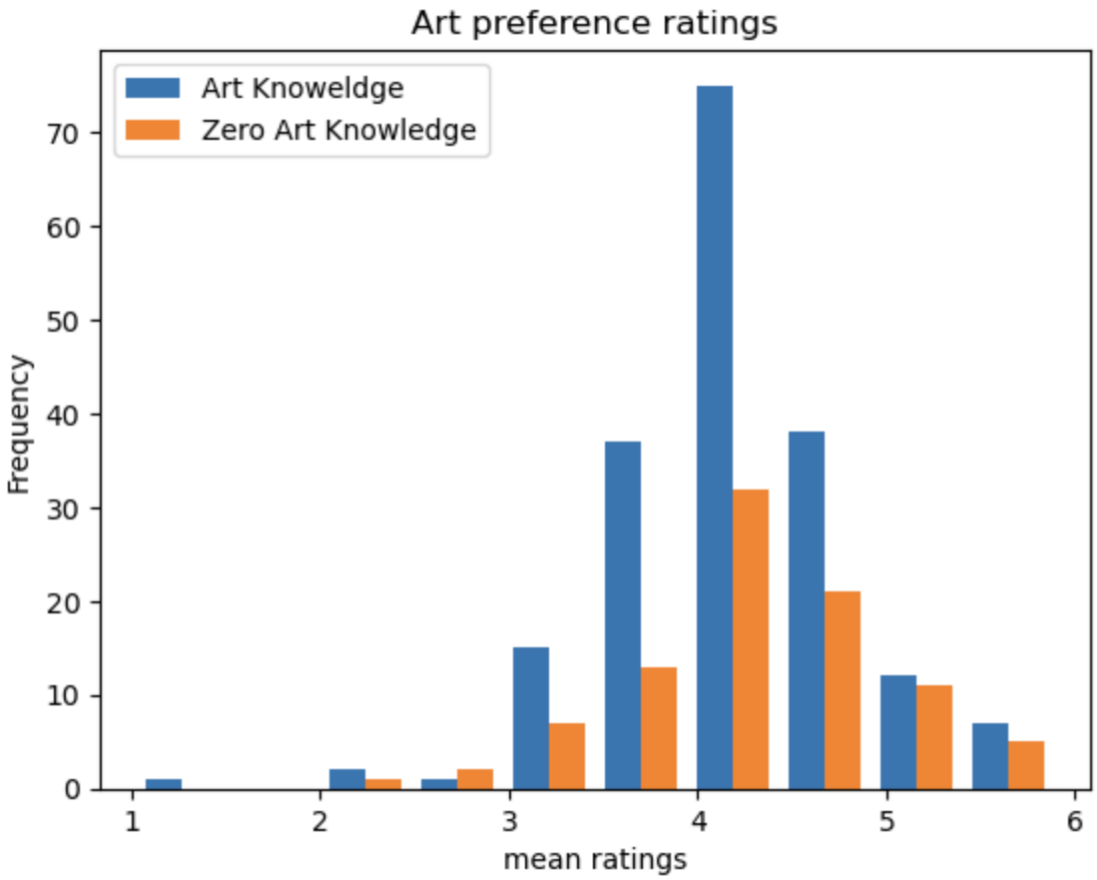

Figure 9: A simple bar chart showing the mean ratings for people with art knowledge vs people with zero art knowledge. The bar chart shows that there is a difference between the art preferences of people with art knowledge and zero art knowledge.

Question 5

In question 5, we had two data frames, one with 300 user ratings of 91 art pieces and the other with 300 energy ratings and 91 art preferences. To run the linear regression model, I imported all the required libraries. This data did not have any nan values, which were also checked. Once we had the two data frames, I took row-wise means to have data frames of one having 300*1 columns where each row represents the mean ratings given by a user for 91 art pieces, and the other data frame was reduced to sample means similarly. I took row-wise means because in this case, it was important to identify user average art preferences ratings in relation to the user average energy ratings. That is why row-wise means were taken. Once we had two data frames, a scatterplot was made for visualization, and then the linear regression model was fitted with a train/test of 0.7/0.3. This was done to avoid overfitting and ensure the model works fine for any

other data. The model was implemented, and the model had a rmse score of 0.585 which means the average difference between the actual value and the predicted value by the model is 0.585. The model also accounts for 20% of the variability in the response variable.

	Actual	Predicted
0	3.956044	4.145021
1	4.560440	4.330341
2	4.758242	4.169968

Figure 10: This shows the actual mean values of the art preference ratings and the predicted mean values of the art preference ratings by the model.

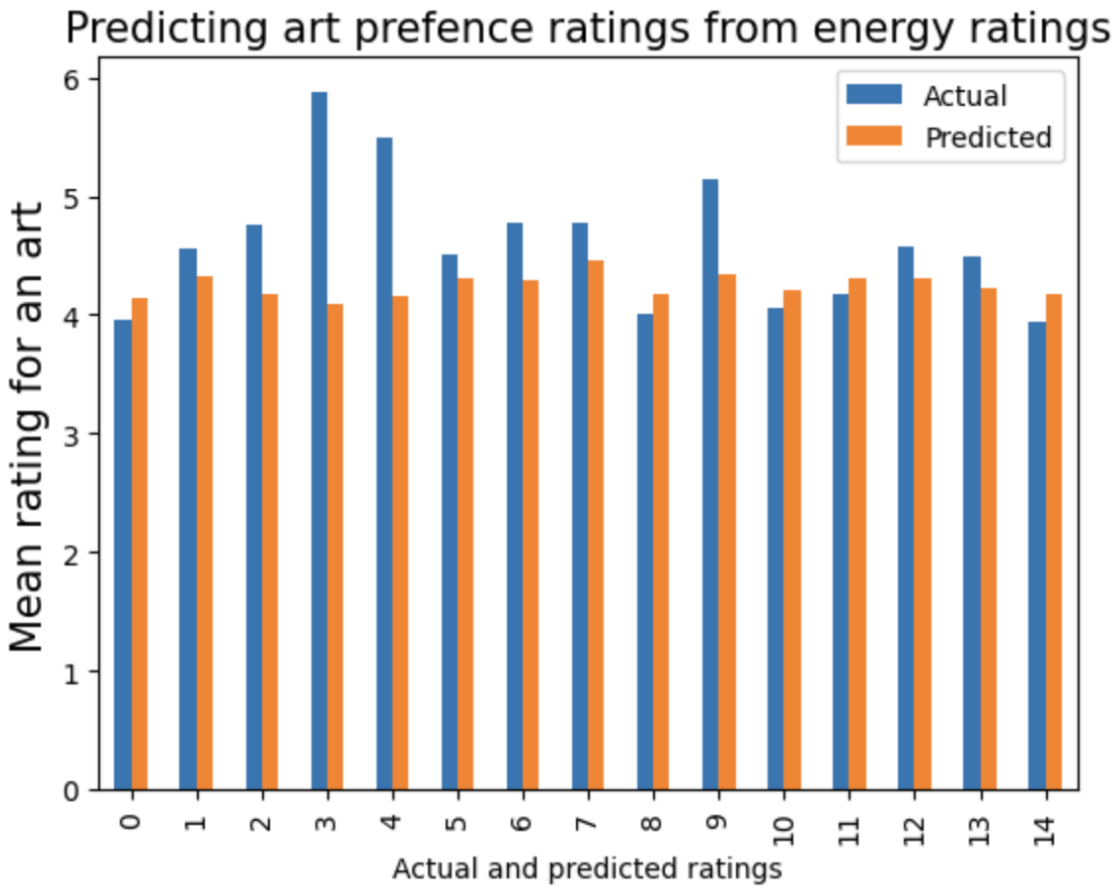

Figure 11: Figure 11 shows the difference between the actual and predicted mean ratings for art preferences by the model for our first 15 variables in the test set.

Question 6

To do this question we made a dataframe that contained row wise means of energy ratings of 91 paintings and the other two columns demographics included user age and gender. So dataframe 1 which was our x variable had three columns and 300 rows now. We catered for the nan by removing row wise nan and then storing indices so we can also remove the same indices from our dataset 2 which had the mean ratings of 300 users for 91 art preferences. So now we had two dataframes of 279*3 and 279*1. The row wise means was taken of energy ratings of 91 art pieces and user ratings and not of age and gender. This is because taking the mean of age and gender was categorical as the gender was given given a outcome of 1,2,3 and 1.5 for instance would mean nothing and for age they were range of differences was very low and decided to not reduce to sample mean because it would have not adjusted with the rest of the dataframe in terms of dataframe if mean had been taken. Once we had two dataframes ready, we did a simple linear regression model, with the dataframe that has 279*3 structure was the x and mean art ratings was the y. An RMSE of 0.6338 was achieved which means the average difference between actual and predicted is 0.6338. Since the RMSE is higher than the previous RMSE in question 5, we can conclude that the performance of the model in q5 was better than this model performance in predicting the mean art ratings. The model explained 12.56% variability in response variable which is also lower than the previous question. To avoid overfitting the data was divided into training and testing with a ratio of 0.7/0.3. The random state was the N number according to the instructions.

	Actual	Predicted
0	3.802198	4.112069
1	4.340659	4.610403
2	4.263736	4.334717

Figure 12: This shows the actual mean values of the art preference ratings and the predicted mean values of the art preference ratings by the model.

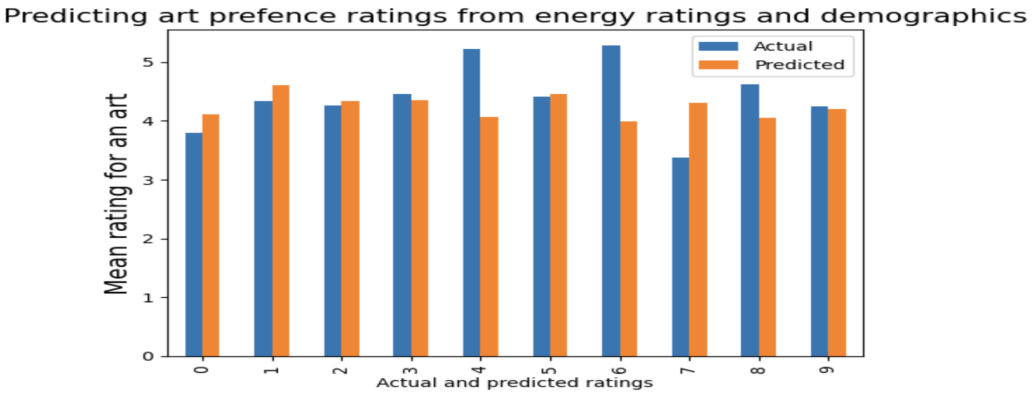

Figure 13: Figure 13 shows the difference between the actual and predicted mean ratings for art preferences by the model for our first 10 variables in the test set.

Question 7

To do question 7 it was important to first to have our data ready to be used. For this question we used column wise mean to have two datasets which has column wise mean of arts and energy ratings which result in we having two datasets of 91*1. To apply and determine the optimal value of clusters, we used z scores to standardize our energy and art means. To determine the optimal value of clusters, we used silhouette method instead of elbow method as we do not have to eye ball the number of clusters. The optimal number of clusters was determined as 4, which is was the highest sum of silhouette score. To identity the clusters, using Kmeans the data was plotted demonstrating which art and energy pair belonged in which cluster. To further understand the particular types of art, I used the np.where to see for instance which indices does the cluster represent and then compared it with my original art dataset. After doing the same process for all 4 clusters, cluster 1 represented modern and classical art, cluster 2 had ratings from modern and non human art. Cluster 3 represented modern and classical art and cluster 4 had ratings that were majority from modern and classical art as well.

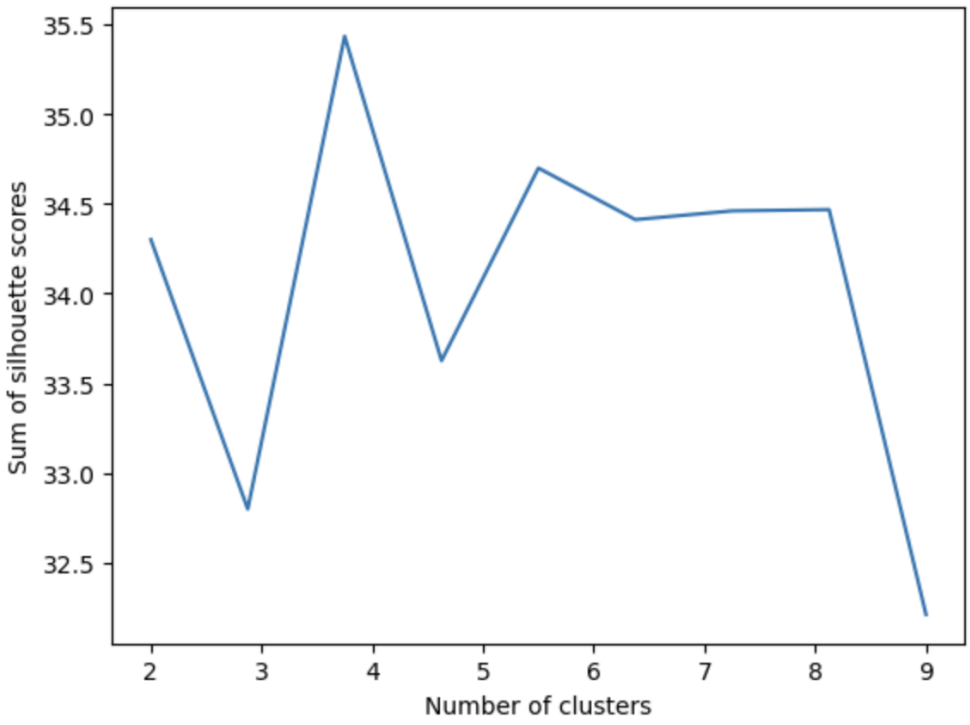

Figure 14: Figure 14 when sum of silhouette scores was plotted against number of clusters we found that the sum was max at 4 so that is why the sum 4 was chosen.

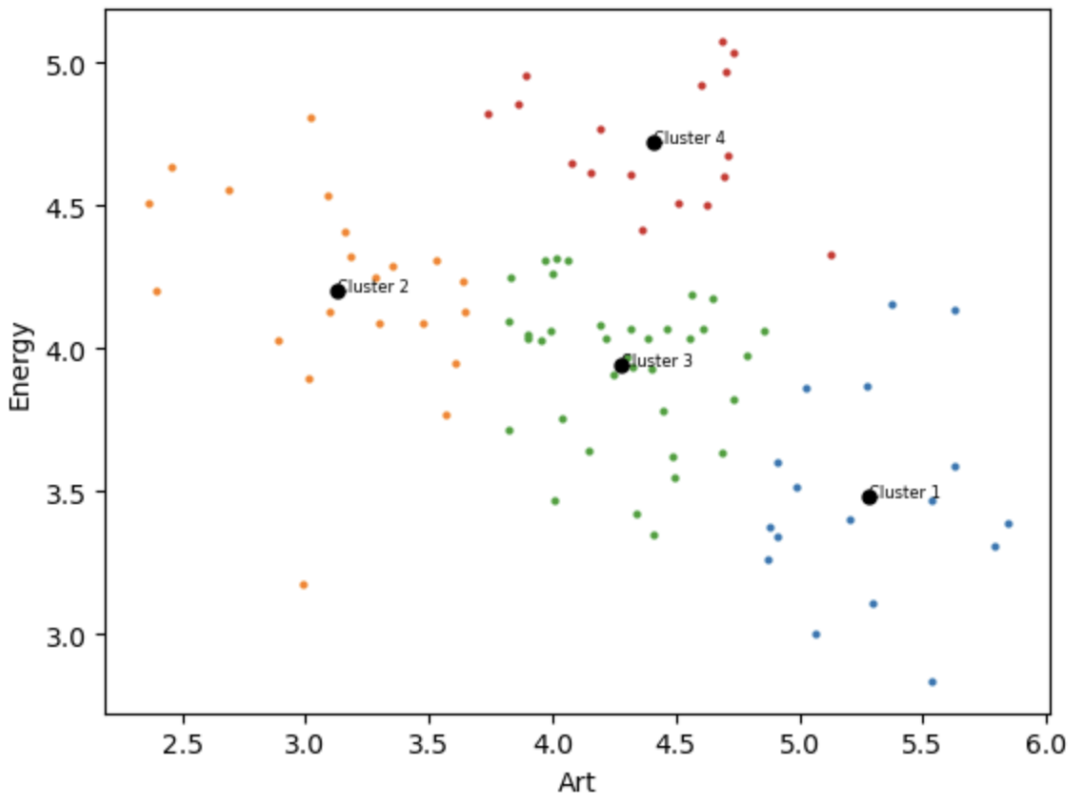

Figure 15: The figure shows that the mean ratings of art and energy taken columnwise for 91 different art pieces divided into 4 different clusters.

```
In [951]: data_cluster_1 = art.iloc[np.where(cId == 0)]
```

Out[951]:

	Number	Artist	Title	Style	Year	Source (1 = classical, 2 = modern, 3 = nonhuman)	computerOrAnimal (0 = human, 1 = computer, 2 = animal)	Intent (0 = no, 1 = yes)
5	6	Rembrandt	Belshazzar's Feast	Baroque	1635	1	0	1
7	8	Ricci	Venus and Cupid	Late Baroque	1700	1	0	1
8	9	Chardin	Saying Grace	Realism	1740	1	0	1
9	10	Hogarth	Marriage A-la-Mode	Rococo	1744	1	0	1
11	12	Greuze	The Village Bride	Rococo	1761	1	0	1
17	18	Goya	The Sleep of Reason Produces Monsters	Romanticism	1798	1	0	1
35	36	Dubuffet	Propitious Moment	Art Brut	1901	2	0	1
43	44	Demuth	I Saw the Number Five in Gold	Futurism / Cubism	1928	2	0	1
46	47	Mondrian	Composition with Red, Blue, and Yellow	Neo-Plasticism	1930	2	0	1
47	48	Dali	The Persistence of Memory	Surrealism	1931	2	0	1
49	50	Miro	Painting	Surrealism / Dada / Experimental	1933	2	0	1
56	57	Matisse	Blue Nude II	Fauvism	1952	2	0	1
58	59	Johns	Three Flags	Abstract Expressionism / Neo-Dada & Pop Art	1958	2	0	1
60	61	Sikander	Walled States	Contemporary Art	1969	2	0	1
68	69	Taaffe	Large Cairene Window	Process - Based Abstraction	2010	2	0	1
73	74	Computer	Dress	Abstract	2017	3	1	0

Figure 16: This snapshot shows that the cluster 1 contains art ratings that are classical and modern. This step was done with all 4 clusters to evaluate which clusters correspond to particular type of art.

Question 8

To do question 8 first it was important to have the data for self image ratings made into a subset. Since this data had some nan values, there was a row wise removal of nan values with the indices being stored, so the same indices can be deleted from data art dataset as well. After this a correlation matrix was plotted to see the correlation between all the questions. The next steps involved zscoring which was done to standardize the values and PCA was done and eigenvalues for each principal was calculated and plotted against the principal component. For the purpose of this question, eigenvalues can be thought in terms of variance explained. The scree plot was visualized to see the variance explained by each question. By looking at the loading we were able to determine first question explains the most variance, which is on the whole i am satisfied with myself. Once we were able to identify the first principal component, we used the values in in as the independent variable in our linear regression model and the mean average of art ratings row wise as our dependent variable. The model was divided into training and testing to avoid for crossfitting.I took row wise means as it give a sense of the average rating given by a user for all pieces, which we are trying to achieve by this model. After doing the linear regression model, an RMSE of 0.611 was obtained telling us that the average difference between the actual and predicted value by our model is 0.611.

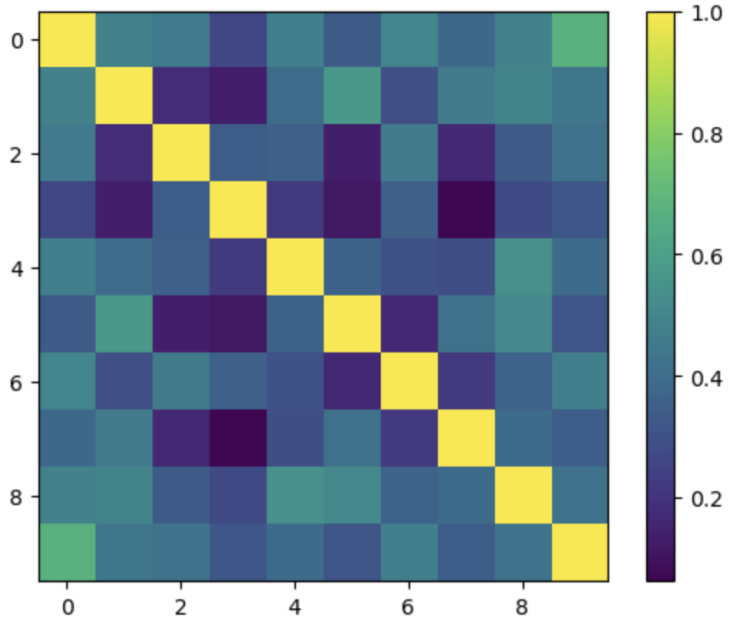

Figure 19: Correlation between all the 10 questions for the self image important to visualize before doing the PCA

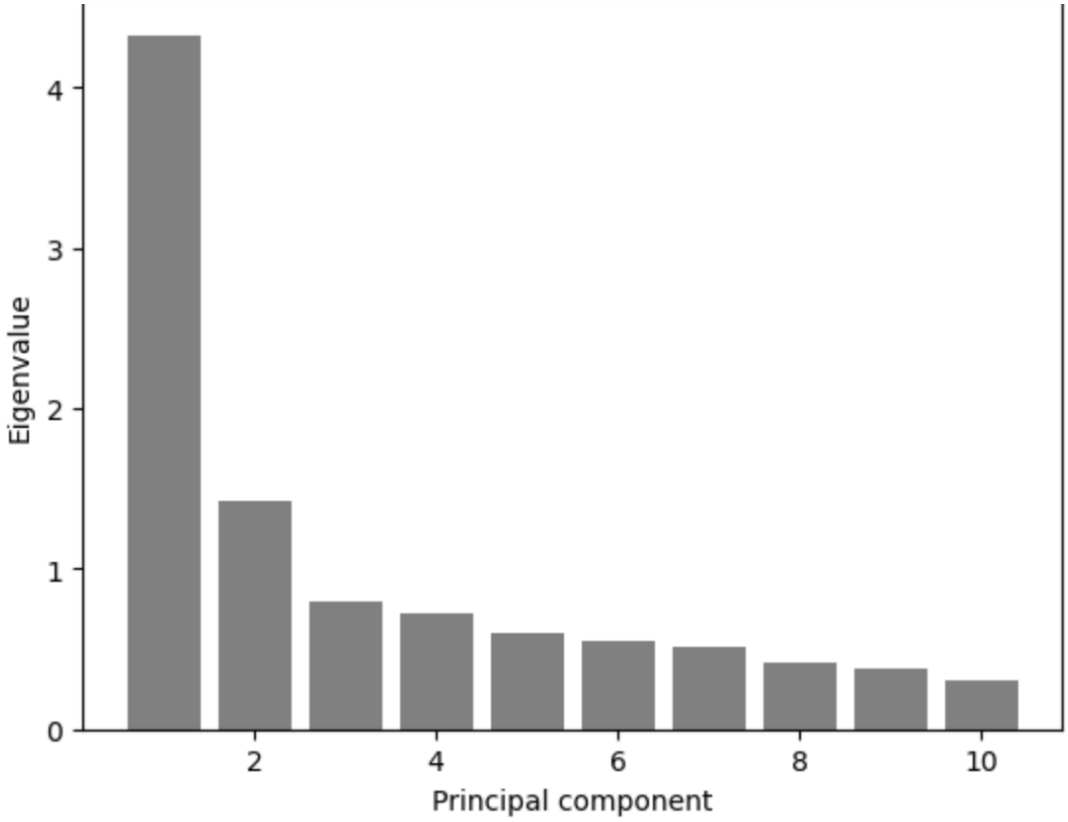

Figure 20: Eigenvalue plotted against the principal component. The sum of all eigen values in this case equal 10.

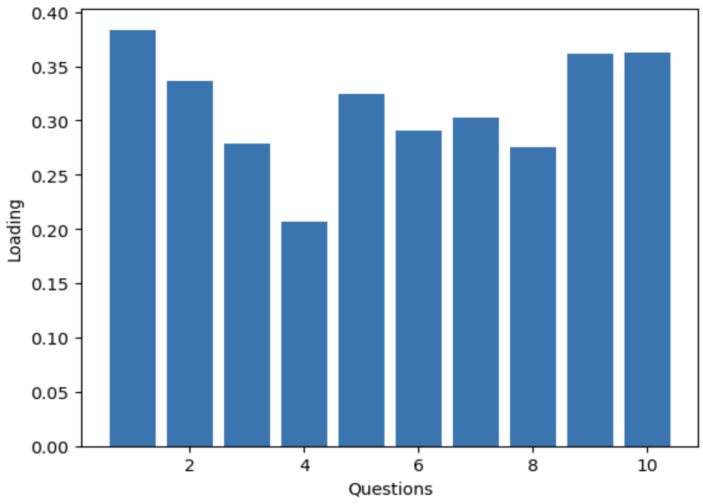

Figure 21: Loading shows that question No 1 on the whole i am satisfied with myself is the first principal component.

Question 9

To do question 9 first it was important to have the data for dark personality traits made into a subset. Since this data had some nan values, there was a row wise removal of nan values with the indices being stored, so the same indices can be deleted from data art dataset as well. After this a correlation matrix was plotted to see the correlation between all the questions. The next steps involved zscoring which was done to standardize the values and PCA was done and eigenvalues for each principal was calculated and plotted against the principal component. For the purpose of this question, eigenvalues can be thought in terms of variance explained. The scree plot was visualized to see the variance explained by each question. For this question kaiser criteria also showed that 3 principal components would have an eigenvalue of 1, which is why we would be considering 3 principal components for our linear regression model. Loadings data was plotted 3 times, against questions, to know which 3 questions would be our principal components. From the 3 loading against questions, we inferred that question 1, question 8 and question 9 of dark personality traits explained the most variance and we would be using data from these 3 questions for our linear regression model. The three questions that were selected were reflective of an individual manipulating others as wel trying others to like them and admire them. Thus manipulativness can be called as an important identity of these factors. The independent variable for our linear regression was the responses to 284 users to these 3 questions, as 16 rows had been deleted to due having nans. Our dependent variable was the row wise mean ratings of 284 users who had rated 91 arts. A linear regression model with a random state of my N number and training and testing size if 0.7 and 0.3 was used to prevent overfitting. The RMSE of the model was calculated which was 0.628 meaning that the average difference between the actual and predicted value by our model is 0.628.

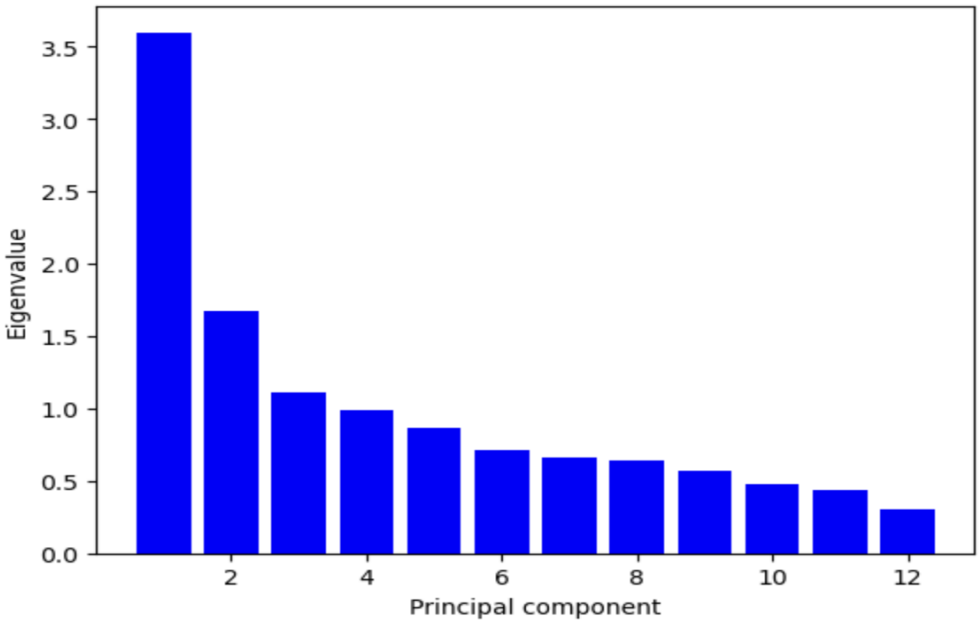

Figure 22: The eigen value against principal component shows that 3 principal comoponet have a value higher than 1 which was also confirmed later in the code by Kaizer criterion.

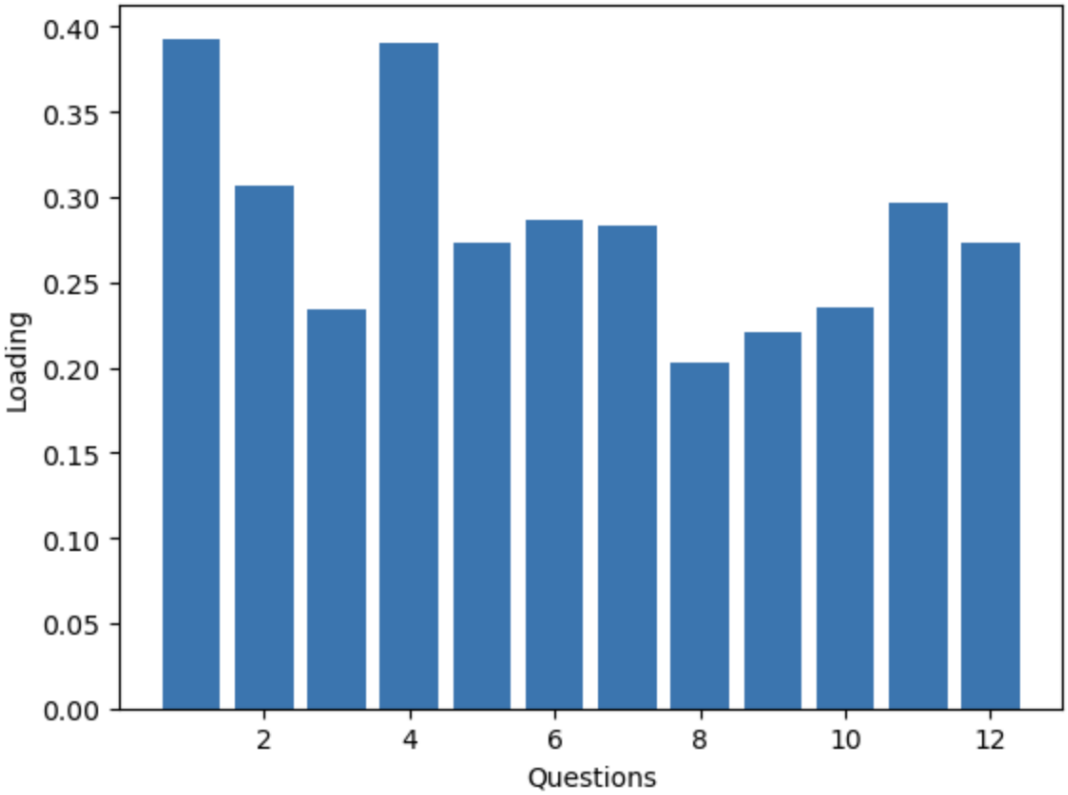

Figure 23: The loading against questions shows that question 1 shows the most variance and was determined as the first principal component.

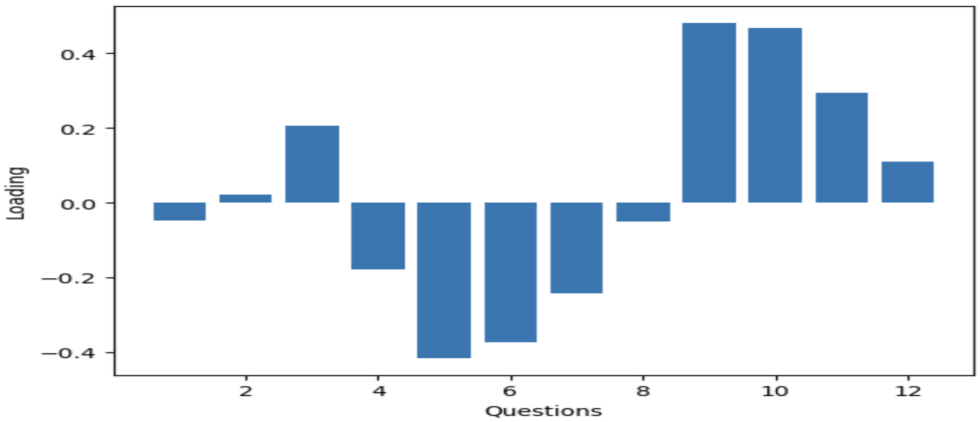

Figure 24: Question 9 had the highest loading value thus was determined as the second principal

Component.

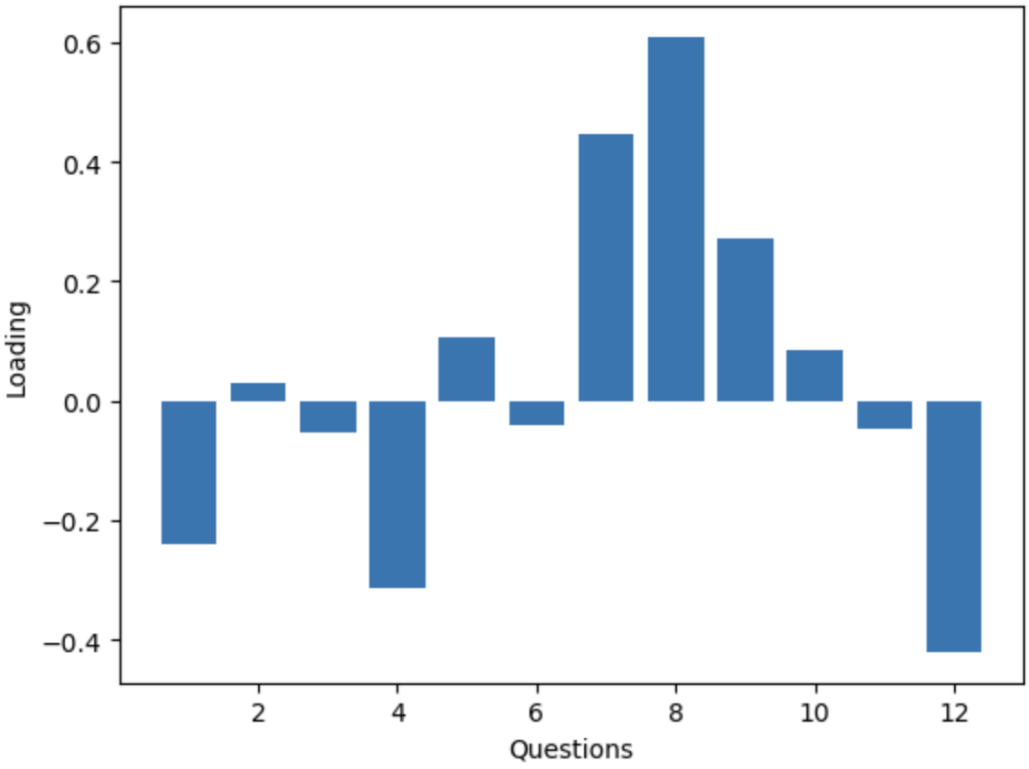

Figure 25: Question 8 had the highest loading value thus was selected as the third principal component.

Question 10

To start question 10 it was important to divide our data into binary data where a value of 1 represented non left which contained all the political orientation of users that rated higher than 2 and value of 0 represented all the ratings that were left and had ratings less than or equal to 2. For this classification we decided to use logistic regression as one of the main reasons was that we were only classifying into two categories it was important to not over fit the data, that is why logistic regression was used. Furthermore, logistic regression is fast in training and predicting as well its simple to implement combined with the fact that it is robust, which in our case age was was an outlier as it had very higher values in comparison to others, led to the choice of logistic regression. Since we already have our dependent variable dataset, it was important to combine our independent dataset to run a logistic regression. To handle the nan values row wise removal oh whole dataset was done, as it did not result in high loss of data and loss of data was very less. Row wise mean was taken of art ratings and energy ratings which led to our combined dataframe being 276*2 by now. Since we had already done PCA on dark personality traits, the three questions that were selected in Q9, their data was merged to the combined dataframe. To reduce

the dimensionality of action, a PCA was done it was determined that 3 principal components had a value higher than eigenvalue of 1 and using the loadings, the 3 questions that explained most variance were selected and their data was merged into our combined dataframe. For self image, the PCA was already done in Q8, and using the Kaizen technique, we saw that there were 2 factors and the data from those 2 questions was merged into the combined dataframe along with all the remaining columns except of political orinetation. To avoid overftting the data was divided into training and test split and a logistic regression. The Auc value achieved was 0.629 Which means that given we have 100 ratings, the model would be able to accurately predict political orientation of 63 users and thus model can be able to distinuguish with some level of accuracy that is better than random prediction as a value of AUC 0.5 is considered random guessing.

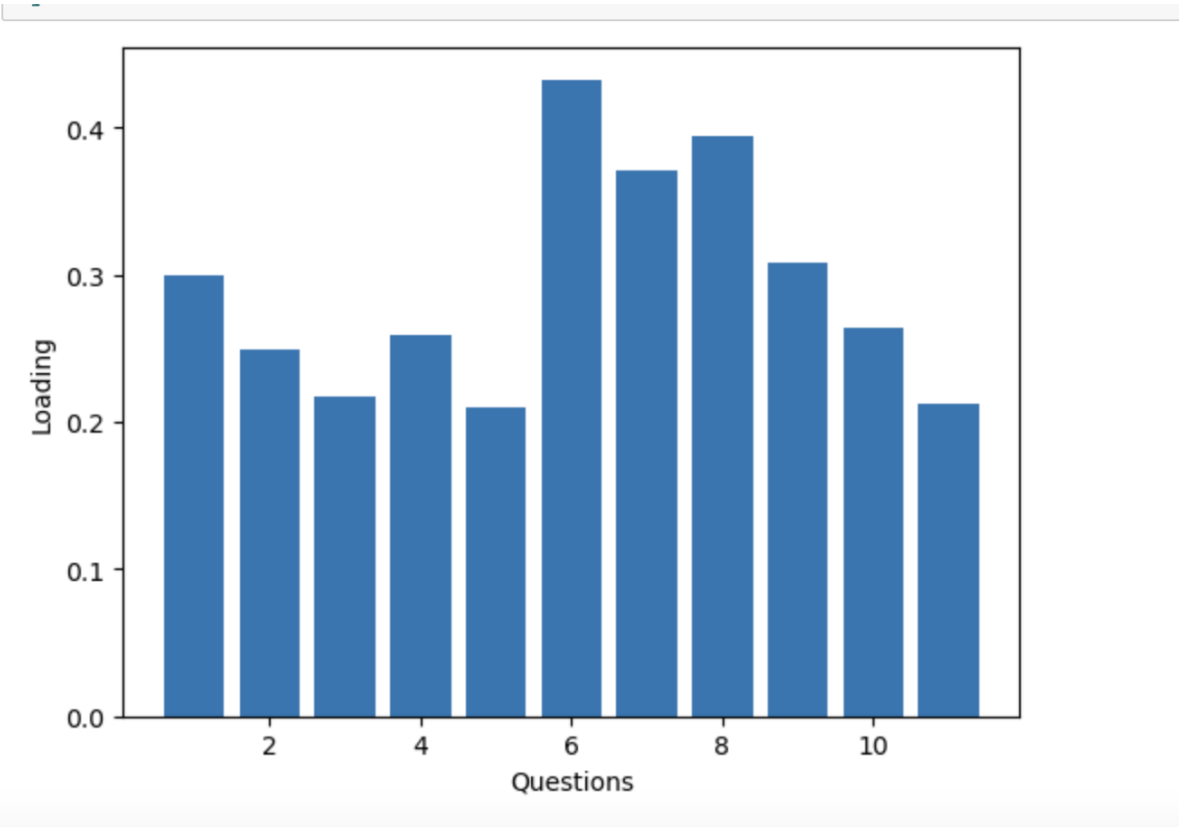

Figure 26: PCA for first component of action did in question 10 as well.

Out[970]:

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
0	3.626374	3.912088	4.0	2.0	4.0	4.0	4.0	2.0	5.0	2.0	19.0	2.0	0.0	2.0	2.0
1	3.934066	3.901099	4.0	3.0	4.0	5.0	3.0	4.0	3.0	2.0	20.0	1.0	1.0	3.0	1.0
2	5.406593	3.901099	3.0	3.0	4.0	5.0	4.0	3.0	3.0	2.0	18.0	2.0	1.0	1.0	2.0
3	4.802198	4.010989	2.0	1.0	4.0	1.0	1.0	1.0	5.0	2.0	21.0	2.0	1.0	3.0	0.0
4	4.230769	3.747253	2.0	3.0	3.0	3.0	3.0	3.0	3.0	2.0	22.0	1.0	0.0	3.0	0.0

Figure 27: Combined dataframe that was used as our independent variable.

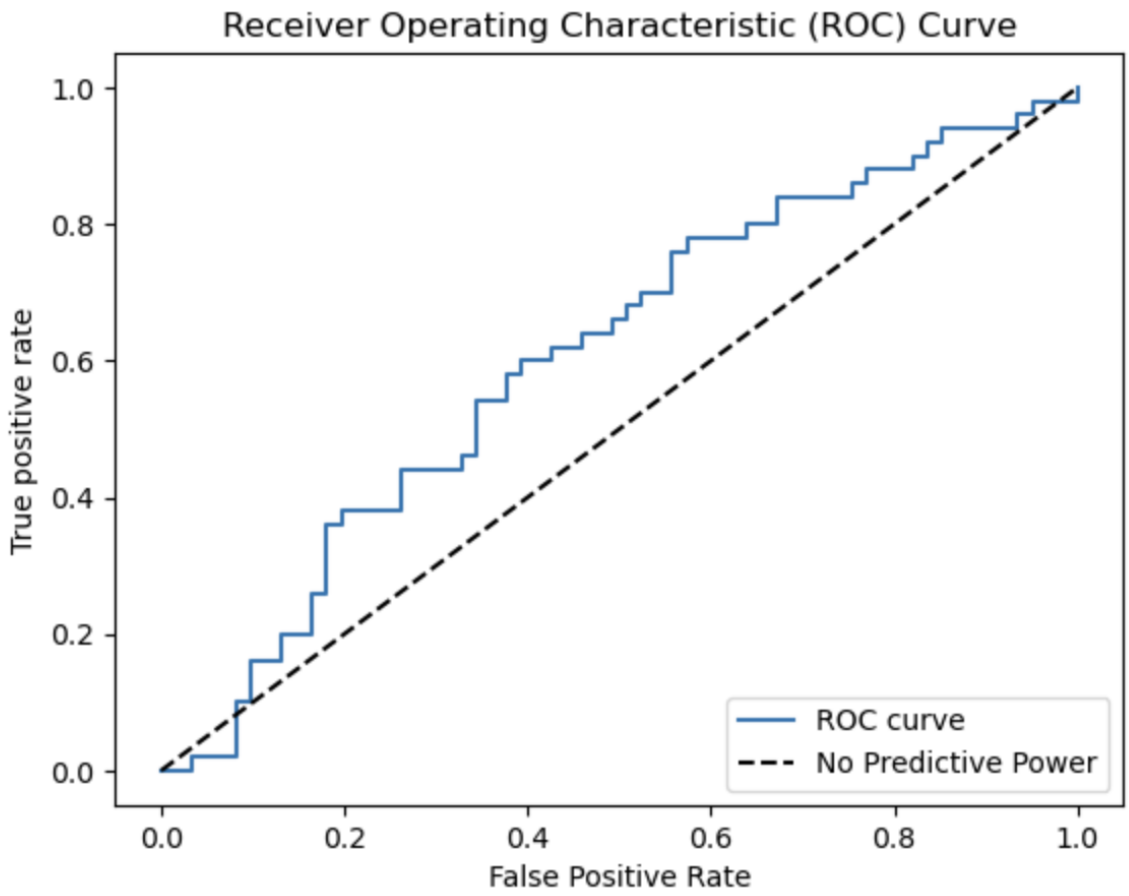

Figure 28: A graphical representation to draw the ROC curve to evaluate the performance of our binary model.