Yong Li yl7123

Pascal Wallisch

Intro to Data Science

2 May 2023

Capstone Project

I. Data preprocessing, transformation, and visualization:

A. Data Cleaning: Entire datasets from user and art files were stored as 'df_user' and 'df_art'. Since the available data points are not so adequate(overall 300 users), NAN data points or missing values were only removed when it was necessary to do so in this project.

B. Dimension Reduction: In the datasets, there're some variables, such as Dark Personality, Action, and self Self - Image, that contain multiple subcolumns. These variables were z-scored and processed by PCA to reach the purpose of reducing dimension and the problem of overfitting.

C. For some questions, both normal data points and average values of them were considered and used to train the model. The purpose of doing two ways is to find the better and more accurate method.

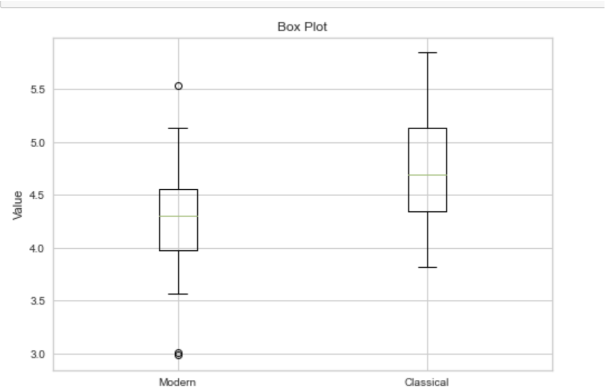
D. Matplotlib and Seaborn Library were used in this project to better visualize the data and help the readers to grasp the idea.

E. Random states for all models in this project were set as 12388176, which is my N number.

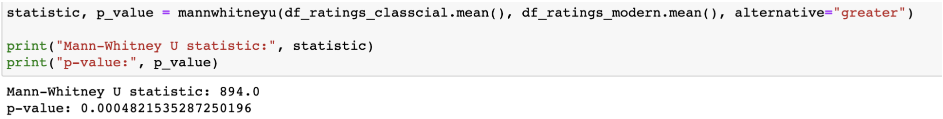
1.

To check whether classical art is more well-liked than modern art, I first extracted them separately from df_users and df_art forming them as two data frames.

I made a box plot to visually compare the difference between these two groups of art.



From the graph, I got the intuition that classical art should be preferred over modern art. I then took a further look at the significance test. Since preference rating is an ordinal variable, I decided to use the one-tail Mann-Whitney U test here to compare the median of these two groups of art pieces. The null hypothesis is classical and modern art get equal ratings.



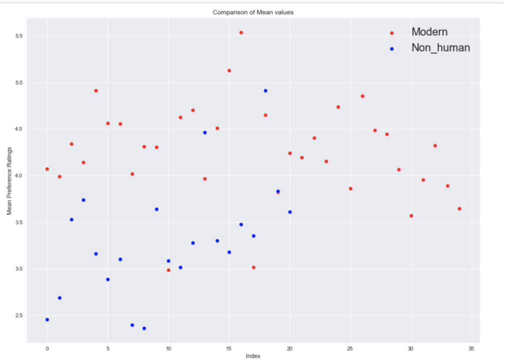
```
statistic, p_value = mannwhitneyu(df_ratings_classcial.mean(), df_ratings_modern.mean(), alternative="greater")

print("Mann-Whitney U statistic:", statistic)
print("p-value:", p_value)
```

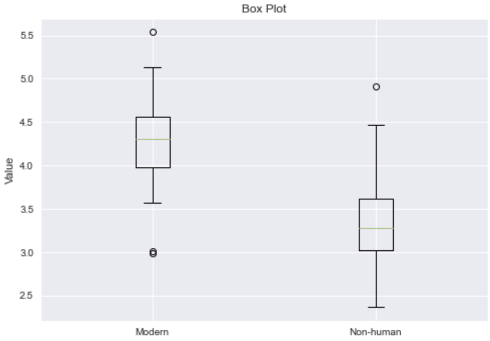
```
Mann-Whitney U statistic: 894.0
p-value: 0.0004821535287250196
```

According to the test result above, the p-value is only 0.0004821, which is significantly less than the alpha level of 0.05, so it's safe for me to reject the null hypothesis and claim that classical art is much more well-liked than modern art.

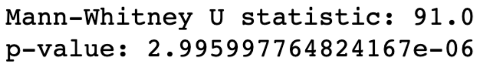
2.

To check whether modern art and non-human-generated art are the same, I followed the same logic in part 1. I made two plots — a box plot and a scatterplot (that plot each art piece's mean preference ratings based on 300 users) — to visualize the distribution of these two groups of data points here. Intuitionally, both of these two plots manage to show a great difference in the two groups.





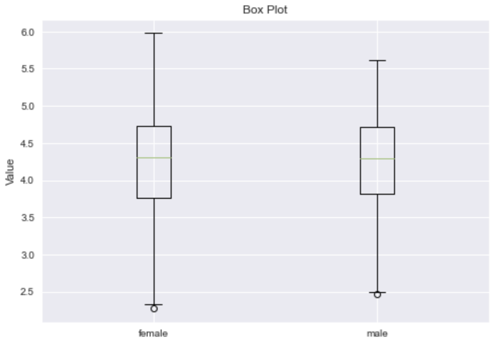
Again, the U test was used but this time is a two-tailed test. Since the p-value is so small — 2.995997764824167e-06, we can reject the null hypothesis and claim that the difference between the modern and nonhuman art pieces is too large to be reasonably consistent with chance.



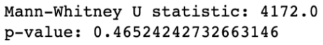
```
Mann-Whitney U statistic: 91.0
p-value: 2.995997764824167e-06
```

3.

To compare woman's and man's preference ratings, I first found the index of women users and the index of men users and then used these indexes to form two data frames with preference ratings in them. The box plot I made pointed out that these two groups' distributions are somewhat similar except for their ranges.



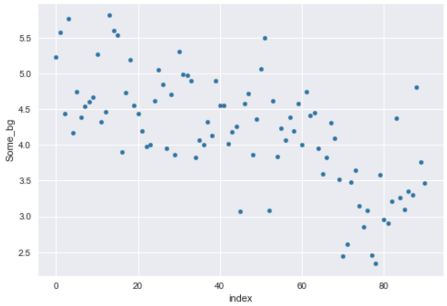
I then tried the one-tailed U test to test these two groups' preference ratings. The null hypothesis is that male preference ratings for art pieces are less than female preference ratings for art pieces. Based on the p-value from the test, which is much greater than the alpha level of 0.05, we fail to reject the null hypothesis. Thus, the slight difference between these two groups might due to chance alone.

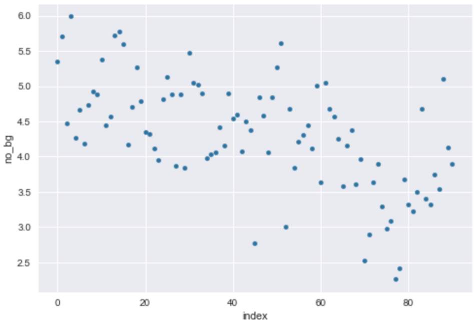


```
Mann-Whitney U statistic: 4172.0
p-value: 0.46524242732663146
```

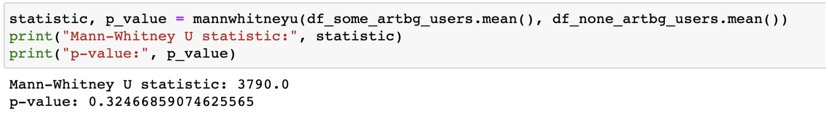
4.

I first extracted the data for no background and some background users and transformed them into two data frames. To check their overall distribution difference, I made two scatterplots to compare





Still, they seem quite similar intuitionally. I then tried a two-tailed U test. The null hypothesis is that there is a difference between these two groups' user ratings. Since the p-value is 0.032469…, which is much larger than 0.05, we fail to reject the null hypothesis, and thus claim that there's no difference in these two groups' ratings.



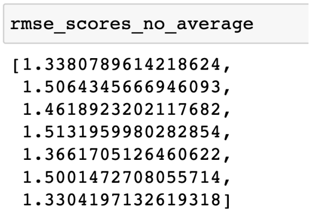
```
statistic, p_value = mannwhitneyu(df_some_artbg_users.mean(), df_none_artbg_users.mean())
print("Mann-Whitney U statistic:", statistic)
print("p-value:", p_value)
```

```
Mann-Whitney U statistic: 3790.0
p-value: 0.32466859074625565
```

5.

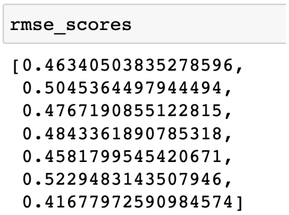
The first idea raised in my mind is using all 300 rows and 91 columns of the energy ratings data frame as input to train the model —— so each user's energy ratings for each art piece as independent variables and each user's preference ratings for each art piece as dependent variables. However, I think it might also be a good choice if we take the average energy ratings and preference ratings(from all users) of each art piece to train the model, which in my view can be less

volatile and accurate. I decided to use the mean but not the median because the median for art pieces could be highly similar, which might lead to a loss of information. So to compare these two methods, I trained the model in these two conditions respectively and recorded their RMSE. Also, I used k-folder cross-validation, with k = 7 here, to prevent the problem of over-fitting.



```
rmse_scores_no_average
```

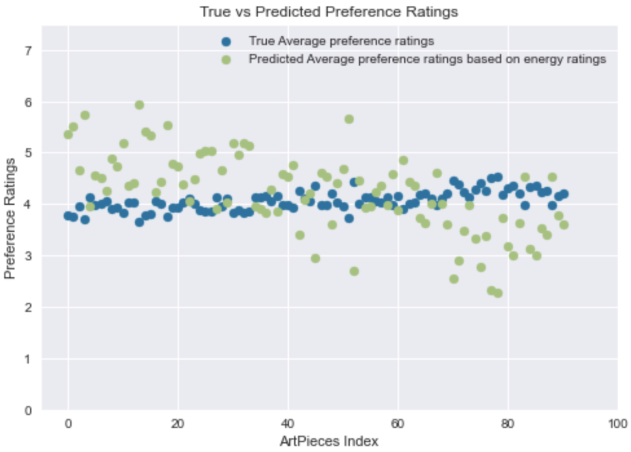
```
[1.3380789614218624,
 1.5064345666946093,
 1.4618923202117682,
 1.5131959980282854,
 1.3661705126460622,
 1.5001472708055714,
 1.3304197132619318]
```



```
rmse_scores
```

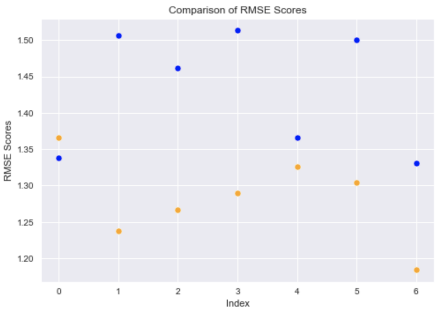
```
[0.46340503835278596,
 0.5045364497944494,
 0.4767190855122815,
 0.4843361890785318,
 0.4581799545420671,
 0.5229483143507946,
 0.41677972590984574]
```

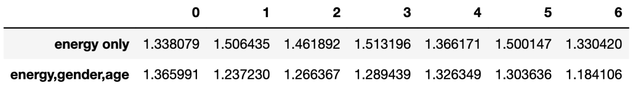
The resulting RMSE lists from the two methods are shown above. It's obvious that the method that is based on average is much more accurate than the one not based on. I also made a scatterplot to compare the true average preference ratings for each art piece and predicted average ratings for each art piece, it seems like our prediction has a close value but a much bigger variance compared to the true average.



6.

I first extracted two more demographic variables "Age" and "Gender", and put them together with energy ratings into the input data frame. The output data frame is still 300 rows and 91 columns of preference rating data. This time, using average energy ratings (mean values for 91 art pieces from all users) as one of the predictors is not appropriate in my perspective because age and gender are variables that differ from user to user. So, I used the whole input data frame along with the output data to train the linear regression model. To prevent the issue of overfitting, I used k-folders (7 folders here) to train test split my datasets. I then compared the RMSE from this model and the model that only uses energy ratings to predict.



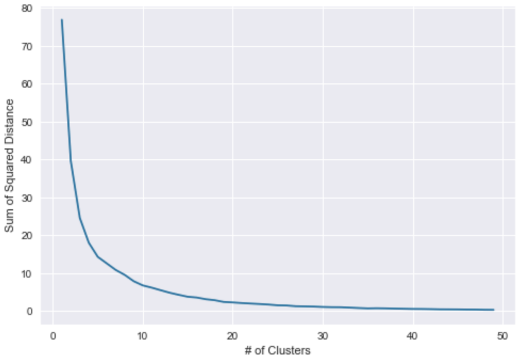


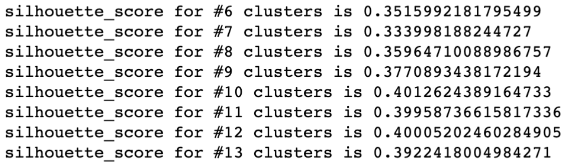
	0	1	2	3	4	5	6
energy only	1.338079	1.506435	1.461892	1.513196	1.366171	1.500147	1.330420
energy,gender,age	1.365991	1.237230	1.266367	1.289439	1.326349	1.303636	1.184106

We can see from the above RMSE values comparison tables and scatterplots for 7 folders, the overall RMSE of the model with age and gender is somehow smaller (except folder 1). But considering the scale of the target variable – preference ratings – is from 1 to 7, the RMSE of the new model with age, gender, and energy ratings as a predictor is still not a great value.

7.

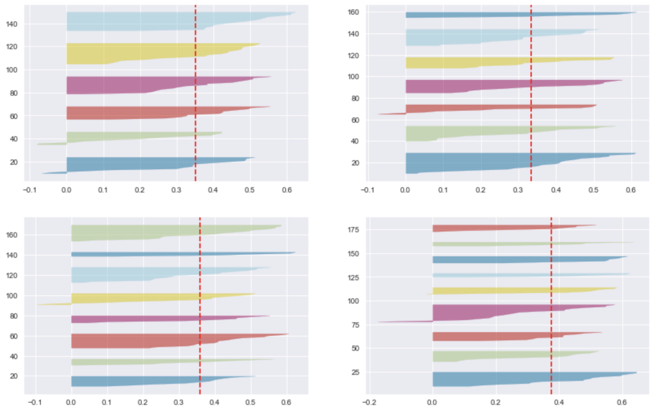
I decide to use K-means clustering here to investigate the potential clusters that exist between the 2D space of average preference ratings and average energy ratings for these 91 art pieces. Before using the Silhouette Score to find the appropriate K value, I first used the Elbow method to determine the appropriate interval of k that can be used. By testing K from range 1 to 50 (since the available art pieces are 91, 50 is already a quite big number of clusters), I made a plot below. We can see from it that the sum of square distance declines sharply from 1 cluster to 7 or 8 clusters, and then enters a slow stage declination from 7 or 8 clusters to 10 clusters and even more clusters.

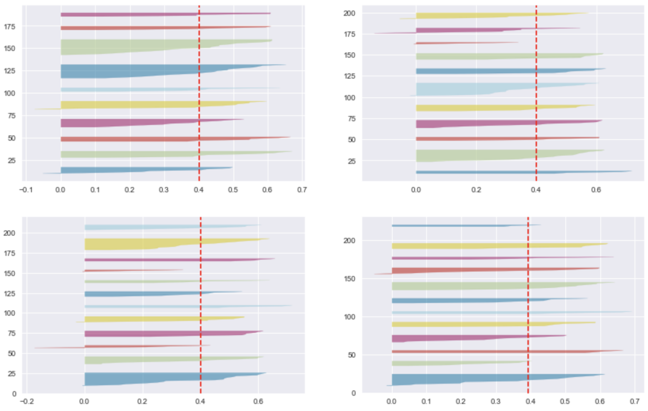




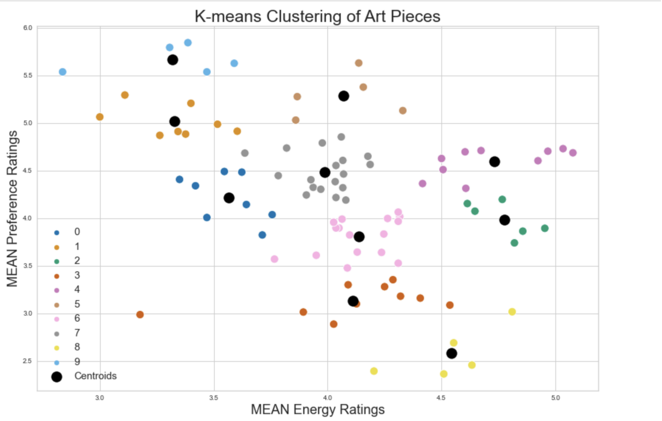
```
silhouette_score for #6 clusters is 0.3515992181795499
silhouette_score for #7 clusters is 0.333998188244727
silhouette_score for #8 clusters is 0.35964710088986757
silhouette_score for #9 clusters is 0.3770893438172194
silhouette_score for #10 clusters is 0.4012624389164733
silhouette_score for #11 clusters is 0.39958736615817336
silhouette_score for #12 clusters is 0.40005202460284905
silhouette_score for #13 clusters is 0.3922418004984271
```

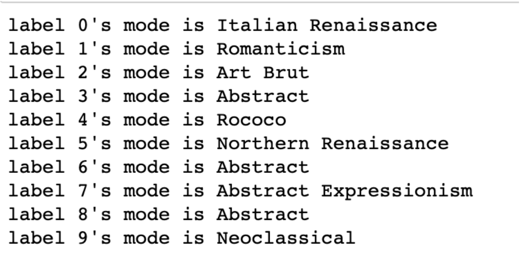
I thus went through over 6 clusters to 13 clusters, and found out that the highest Silhouette Score achieved is when k = 10. Below is the visualized Silhouette Score graph (the left one is from 6 clusters to 9 clusters, and the right one is from 10 clusters to 13 clusters) and stats.





After determining the value of k, data were put in the K-means model and trained.



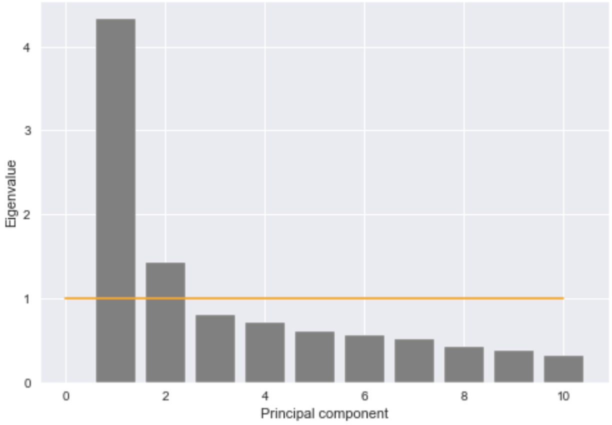


```
label 0's mode is Italian Renaissance
label 1's mode is Romanticism
label 2's mode is Art Brut
label 3's mode is Abstract
label 4's mode is Rococo
label 5's mode is Northern Renaissance
label 6's mode is Abstract
label 7's mode is Abstract Expressionism
label 8's mode is Abstract
label 9's mode is Neoclassical
```

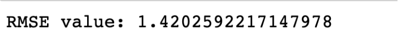
The scatterplot above shows how art pieces were clustered based on the 10 centroids. I then took a further step to check the mode of each cluster's art style. According to the list below, it seems like the output labels of k mean clustering are still not so accurate(we can see lots of overlapping in styles between label 3, label 6, and label 8 —— their mode of art style is the same: Abstract). But still, the 10 clusters can still somehow reflect their particular art styles.

8.

The self-image variable contains multiple columns, so using PCA is necessary. After standardizing the data frame of the self-image, I did PCA on it and found out that indeed 2 components here satisfy the Kaiser criterion line.

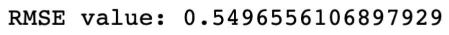


Using only the first component here, I transformed the self-image dataset to the new coordinate space and put it together with the target variable preference ratings to do the linear regression. The resulting RMSE value is 1.4202592217147978. Given the scale of the target variable ratings (1-7), I regard the RMSE value as quite large. So, using the first component of self-image ratings to make a prediction on preference ratings is not a quite good choice indeed.



```
RMSE value: 1.4202592217147978
```

Besides this perspective, I also tried to use the first component to predict the average mean preference ratings for each user.

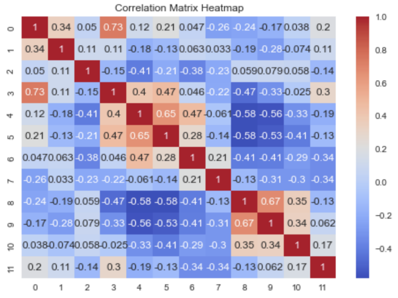


```
RMSE value: 0.5496556106897929
```

The value of RMSE in this condition decreased a lot. Given the scale of the target variable ratings (1-7), I regard the RMSE value as quite small and acceptable. So, using the first component of self-image ratings to make a prediction on each user's median preference ratings for all art pieces is a quite good choice.

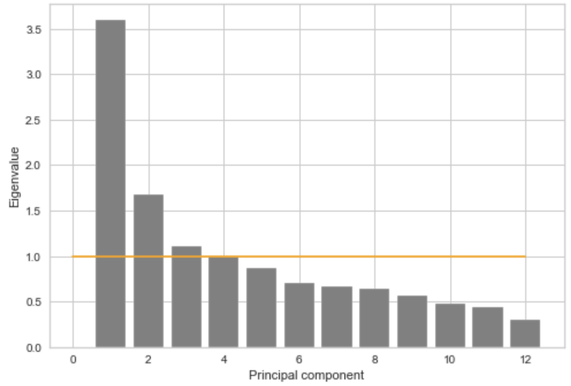
9.

Since dark personality is also characterized by multiple columns(variables), It's still important for us to do PCA here. I first did a correlation heatmap here to better visualize the relationship. Three components here satisfy the Kaiser Criterion.

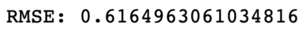


Correlation Matrix Heatmap

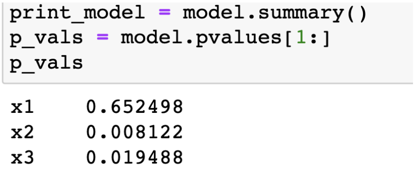
	0	1	2	3	4	5	6	7	8	9	10	11
0	1	0.34	0.05	0.73	0.12	0.21	0.047	-0.26	-0.24	-0.17	0.038	0.2
1	0.34	1	0.11	0.11	-0.18	-0.13	0.063	0.033	-0.19	-0.28	-0.074	0.11
2	0.05	0.11	1	-0.15	-0.41	-0.21	-0.38	-0.23	0.059	0.079	0.058	-0.14
3	0.73	0.11	-0.15	1	0.4	0.47	0.046	-0.22	-0.47	-0.33	-0.025	0.3
4	0.12	-0.18	-0.41	0.4	1	0.65	0.47	-0.061	-0.58	-0.56	-0.33	-0.19
5	0.21	-0.13	-0.21	0.47	0.65	1	0.28	-0.14	-0.58	-0.53	-0.41	-0.13
6	0.047	0.063	-0.38	0.046	0.47	0.28	1	0.21	-0.41	-0.41	-0.29	-0.34
7	-0.26	0.033	-0.23	-0.22	-0.061	-0.14	0.21	1	-0.13	-0.31	-0.3	-0.34
8	-0.24	-0.19	0.059	-0.47	-0.58	-0.58	-0.41	-0.13	1	0.67	0.35	-0.13
9	-0.17	-0.28	0.079	-0.33	-0.56	-0.53	-0.41	-0.31	0.67	1	0.34	0.062
10	0.038	-0.074	0.058	-0.025	-0.33	-0.41	-0.29	-0.3	0.35	0.34	1	0.17
11	0.2	0.11	-0.14	0.3	-0.19	-0.13	-0.34	-0.34	-0.13	0.062	0.17	1



After transforming the dark personality dataset into the new coordinate, I used it to predict each user's average preference ratings for all art pieces (it provides less RMSE than for each art piece's preference ratings).



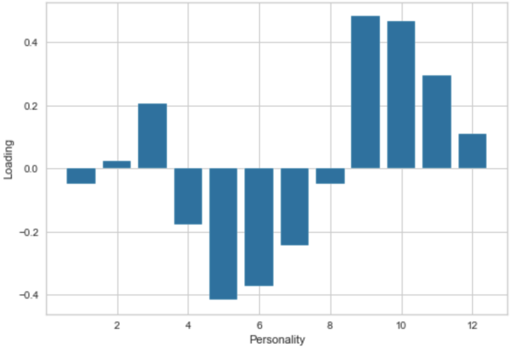
```
RMSE: 0.6164963061034816
```

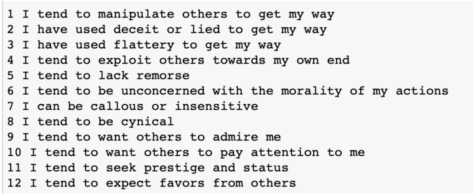


```
print_model = model.summary()
p_vals = model.pvalues[1:]
p_vals
```

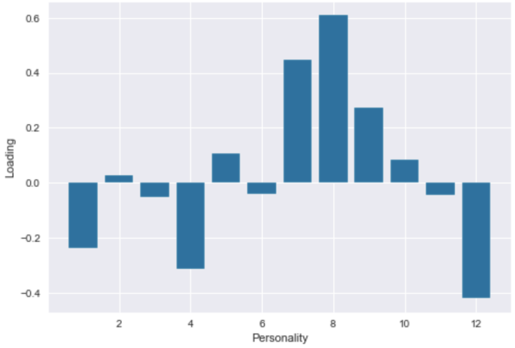
```
x1    0.652498
x2    0.008122
x3    0.019488
```

The p-value for the second and the third components here is less than the alpha level of 0.05, meaning that they are significant. I then plotted the potential identities of these two components.





```
1 I tend to manipulate others to get my way
2 I have used deceit or lied to get my way
3 I have used flattery to get my way
4 I tend to exploit others towards my own end
5 I tend to lack remorse
6 I tend to be unconcerned with the morality of my actions
7 I can be callous or insensitive
8 I tend to be cynical
9 I tend to want others to admire me
10 I tend to want others to pay attention to me
11 I tend to seek prestige and status
12 I tend to expect favors from others
```

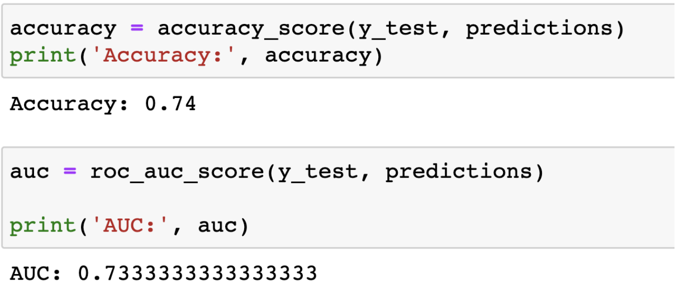


For the first component, factors like “9. I tend to want others to admire me”, “10. I tend to want others to pay attention to me”, and “11. I tend to seek prestige and status” should be potential identities.

For the second components, for the second component, factors like “7. I can be callous or insensitive”, and “8. I tend to be cynical” should be potential identities.

10.

I think Independent variables that were used to predict political tendencies here can be classified into three types: categorical variables, variables that needed to be processed by PCA, and variables that needed to be standardized. All the categorical variables in the datasets are already converted numerically into numbers like “1, 2, and 3”, so I did not use one hot encoder to process them. Variables like dark personality, action, and self-image were processed by PCA. Since we do not standardize categorical variables and variables derived from PCA, I only standardized the rest variables like average preference ratings for 91 art pieces, age, etc. I did not choose to perform PCA for all the independent variables here again as they were already quite independent (If we perform PCA again, The AUC score is 0.4899999 which is much lower than the AUC score without PCA again below). The finalized independent variables were therefore created. The target variable – political orientation – was transformed into a binary variable: 0 means left and 1 means left. Independent variables and target variables were put into a logistic regression model. The accuracy of it is 0.74 and the AUC score of it stands at 0.7333. It’s thus a model that can perform much better than a random guess.



```
accuracy = accuracy_score(y_test, predictions)
print('Accuracy:', accuracy)
```

```
Accuracy: 0.74
```

```
auc = roc_auc_score(y_test, predictions)

print('AUC:', auc)
```

```
AUC: 0.7333333333333333
```

- A meaningful point about this dataset

The ordinal variables: preference and energy ratings are on a scale of 1 to 7.

Calculating the mean in this condition(user rating condition) leads to a normalized sum that presumes that the units of items being summed are equal in order to be meaningful. This might not be so accurate in reflecting users' level of affection or hatred don different styles of art pieces, but using the median here as the target variable might also lead to lots of duplicates among various users who indeed contain quite divergent characteristics.

So, I still decided to use mean to convert the preference and energy ratings when I need to get summary stats for each user’s preference and energy ratings in this project.