Jiewu Rao

Professor Pascal Wallisch

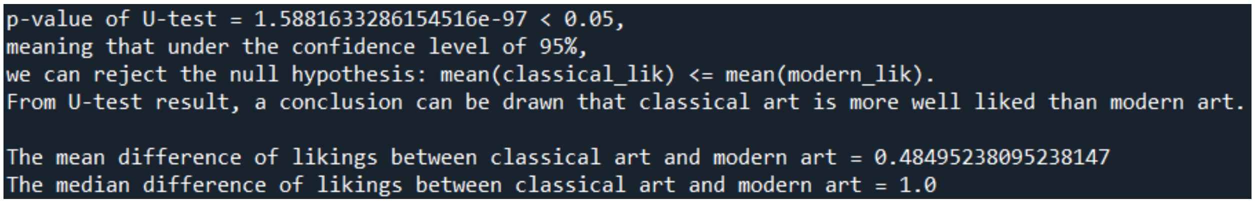
Intro to Data Science

Capstone Project Report

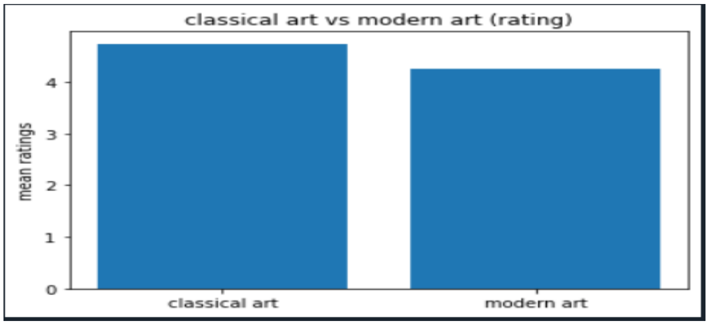
**Open Statement:**

The project consists of 91 individual art pieces and responses from 300 users. Also, I decided to replace the NaN value with its median of the column and remove some outliers in some questions. PCA is used for dimension reduction purposes. As for the first several questions, we cannot just compare their means or medians so I also do the U-test. For the entire project, I use the significance level of 0.05 for the hypothesis testing and test size of N, which is the number based on the random seed of my N number. The random_state may be different in some questions.

1. I first integrate their preference ratings into two lists accordingly. Both of them have 10500 preference ratings. Then I implement the Mann-Whitney U test. The null hypothesis is the mean rating for classical art is lower than or equal to the rating for modern art. This is a one-tailed test. The p-value of the U-test is lower than the threshold value 0.05, therefore we can reject the null hypothesis. From this result, we can conclude that classical art is more well liked than modern art. Next, I find their mean difference and median difference to assure my conclusion. The classical art has a higher value in both mean and median. Finally, the conclusion drawn from the U-test can be reconfirmed.

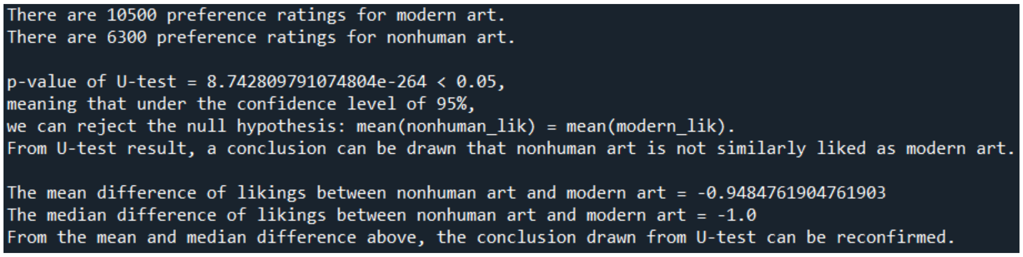


```
p-value of U-test = 1.5881633286154516e-97 < 0.05,
meaning that under the confidence level of 95%,
we can reject the null hypothesis: mean(classical_lik) <= mean(modern_lik).
From U-test result, a conclusion can be drawn that classical art is more well liked than modern art.

The mean difference of likings between classical art and modern art = 0.48495238095238147
The median difference of likings between classical art and modern art = 1.0
```

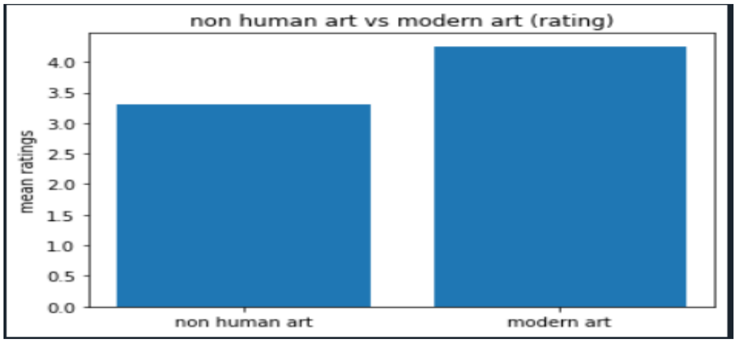


2. The process is kind of similar to Q1.
There are 10500 preference ratings for modern art and 6300 preference ratings for nonhuman art. I only created a list for non-human art in this question as I have done it for modern art in Q1. Then I implement the Mann-Whitney U test. The null hypothesis is that the mean rating for modern art is equal to the rating for non-human art. This is a two-tailed test. The p-value of the U-test is lower than the threshold value 0.05, therefore

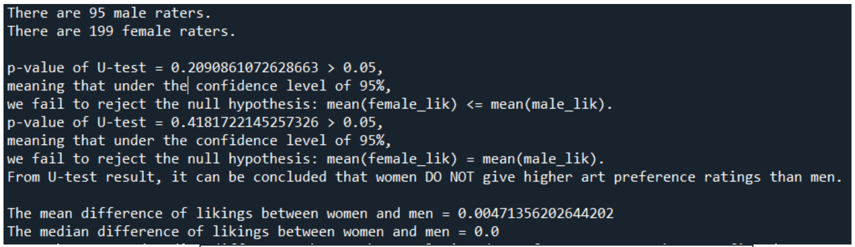
we can reject the null hypothesis. From this result, we can conclude that there is a difference in the preference ratings for modern art vs. non-human art. Next, I find their mean difference and median difference to assure my conclusion. There is a difference in their mean and median. Finally, the conclusion drawn from the U-test can be reconfirmed.



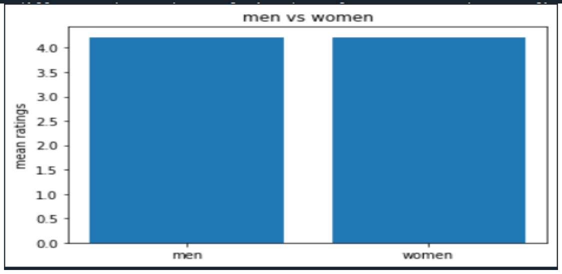
```
There are 10500 preference ratings for modern art.
There are 6300 preference ratings for nonhuman art.

p-value of U-test = 8.742809791074804e-264 < 0.05,
meaning that under the confidence level of 95%,
we can reject the null hypothesis: mean(nonhuman_lik) = mean(modern_lik).
From U-test result, a conclusion can be drawn that nonhuman art is not similarly liked as modern art.

The mean difference of likings between nonhuman art and modern art = -0.9484761904761903
The median difference of likings between nonhuman art and modern art = -1.0
From the mean and median difference above, the conclusion drawn from U-test can be reconfirmed.
```



3. There are 95 male raters and 199 female raters. Then I implement the Mann-Whitney U test. The null hypothesis is the mean rating of women is lower than or equal to the men's rating. This is a one-tailed test. The p-value of the U-test is greater than the threshold value 0.05, therefore we fail to reject the null hypothesis. After this, I conduct another U-test which has the null hypothesis as there is no difference between women's rating and men's rating. This is a two-tailed test. Again, the p-value is greater than 0.05. We fail to reject this hypothesis either. It can be concluded that women DO NOT give higher art preference ratings than men. Next, I find their mean difference and median difference to assure my conclusion. The mean difference between men and women is tiny and there is no difference in the median. Therefore, the conclusion drawn

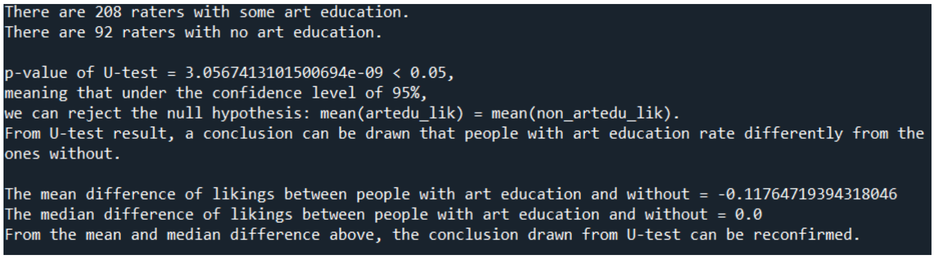


```
There are 95 male raters.
There are 199 female raters.

p-value of U-test = 0.2090861072628663 > 0.05,
meaning that under the confidence level of 95%,
we fail to reject the null hypothesis: mean(female_lik) <= mean(male_lik).
p-value of U-test = 0.4181722145257326 > 0.05,
meaning that under the confidence level of 95%,
we fail to reject the null hypothesis: mean(female_lik) = mean(male_lik).
From U-test result, it can be concluded that women DO NOT give higher art preference ratings than men.

The mean difference of likings between women and men = 0.00471356202644202
The median difference of likings between women and men = 0.0
```

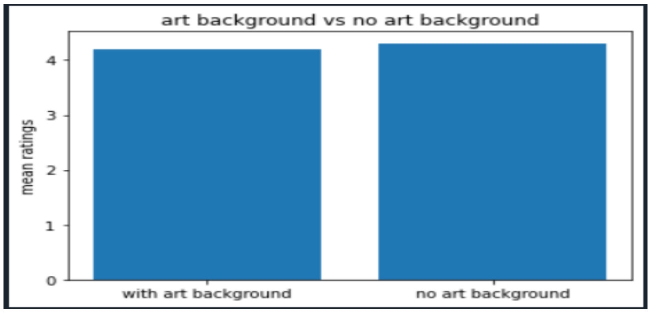


from the U-test can be reconfirmed.  We can see the difference is tiny from the above visualized comparison.

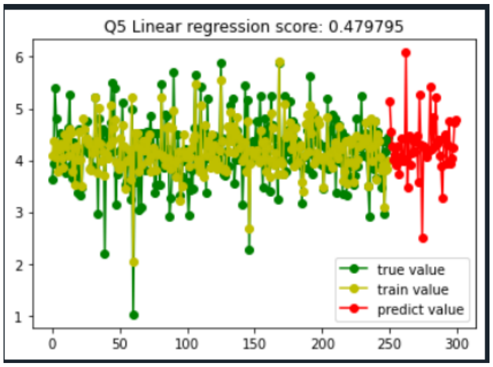
4. There are 208 raters with some art education and 92 raters with no art education. Then assign to two lists to store their preferences. Then I implement the Mann-Whitney U test. The null hypothesis is that the mean rating of raters with some art education has no difference compared to the mean rating of raters with no art education. This is a two-tailed test, which is similar to Q2. The p-value of the U-test is lower than the threshold value 0.05, therefore we can reject the null hypothesis. From this result, we can conclude that there is a difference in the preference ratings between raters with some art education and raters without art education. Next, I find their mean difference and median difference to assure my conclusion. There is a difference in their mean and median. Finally, the conclusion drawn from the U-test can be reconfirmed.

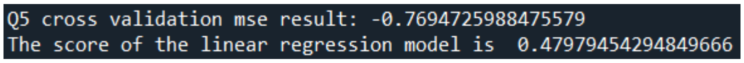


```
There are 208 raters with some art education.
There are 92 raters with no art education.

p-value of U-test = 3.0567413101500694e-09 < 0.05,
meaning that under the confidence level of 95%,
we can reject the null hypothesis: mean(artedu_lik) = mean(non_artedu_lik).
From U-test result, a conclusion can be drawn that people with art education rate differently from the
ones without.

The mean difference of likings between people with art education and without = -0.11764719394318046
The median difference of likings between people with art education and without = 0.0
From the mean and median difference above, the conclusion drawn from U-test can be reconfirmed.
```



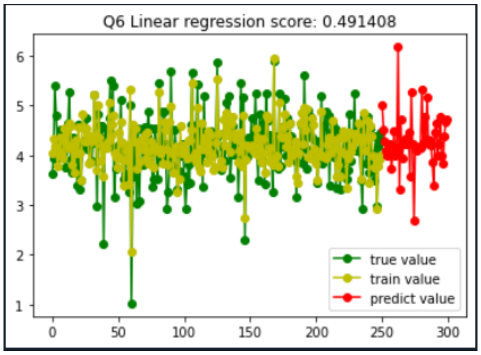
5. First split out the X, Y and assigned names accordingly for them. The ratio between the training and testing set is 5:1. The score is 0.479795, and also the r square, which is how much of the outcome the predictor accounts for. The cross_val_score method returns the negative version of the MSE. The cross validation mse result is -0.769, showing the model is not overfitting the data.



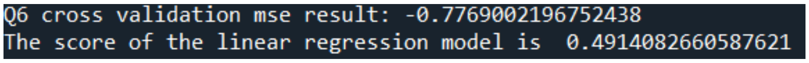


```
Q5 cross validation mse result: -0.7694725988475579
The score of the linear regression model is  0.47979454294849666
```

6. The logic and process are similar to Q5, except that the X has more columns in it. The new linear regression model yields a score of 0491408, which has a slight improvement.

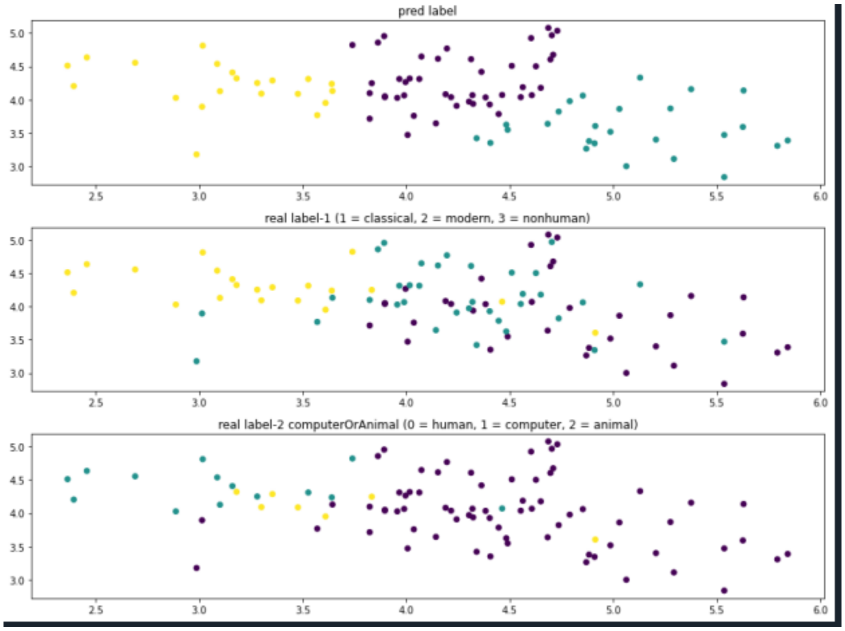


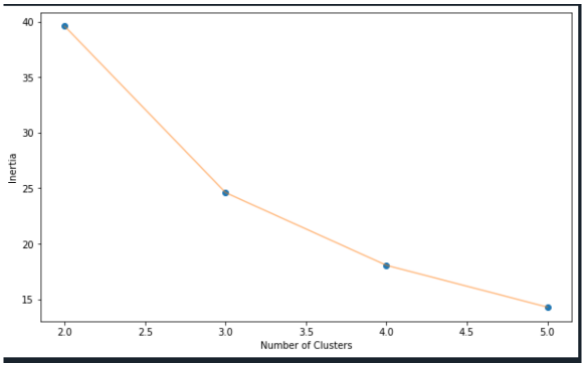
The cross_val_score method returns the negative version of the MSE. The cross validation mse result is -0.777, showing the model is not overfitting the data.



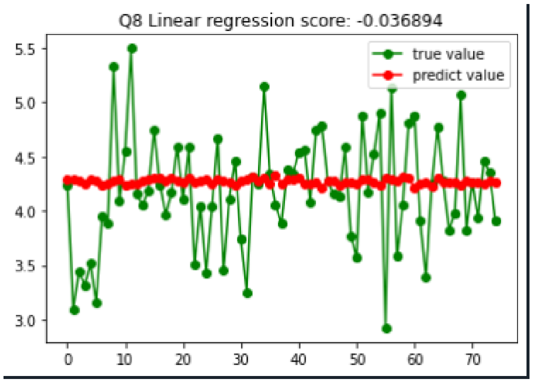
```
Q6 cross validation mse result: -0.7769002196752438
The score of the linear regression model is  0.4914082660587621
```

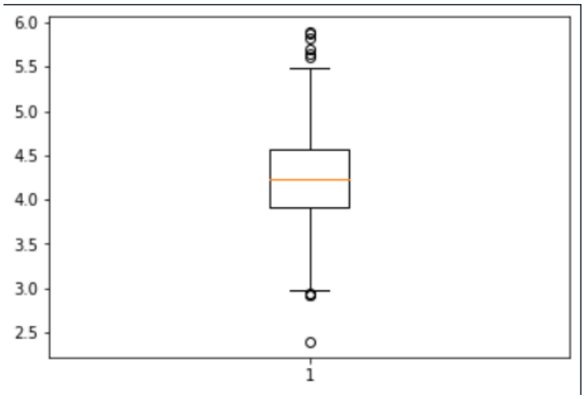
7. For this question I use the elbow method, which is the most common way to do it with eyeballing. The number of clusters is 3. Each clustering predicted result is therefore considered to correspond to a particular type of art.

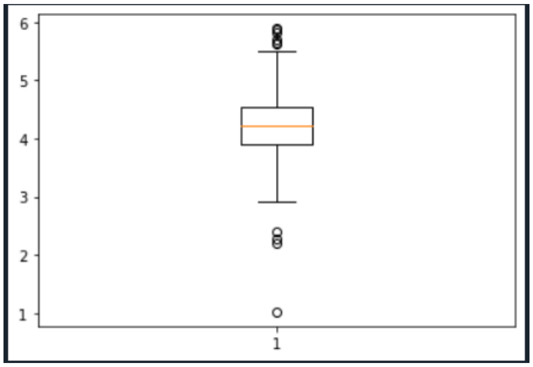




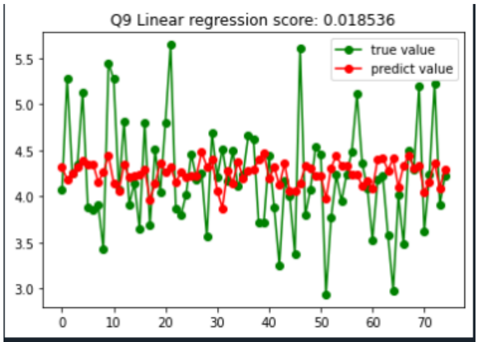
8. First I eliminate outliers based on the standard deviation to improve the score. Then I do the z score, transform on the data as the professor shows in the code session. The score of the linear regression model is -0.036894, which is below 0. This shows the model has a very weak predictive ability. When n_component = 1, the model score is low. These two shows the difference in data before and after the removal of outliers.

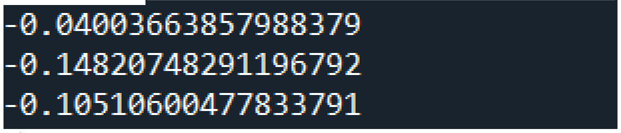






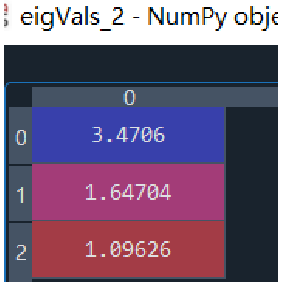
9. This time n_component = 3, the score is better at 0.018536 but is still not considered a good score. The following screenshot shows the correlation score of the three components. Neither one has a positive score, which shows their weakness to predict. The same for the data in Q8 as its correlation coefficient is low. As shown in screenshots, the percentage of the variance explained by the first component is the most significant. Therefore, the first of the dark traits, manipulation, plays a bigger role in predicting the rating.





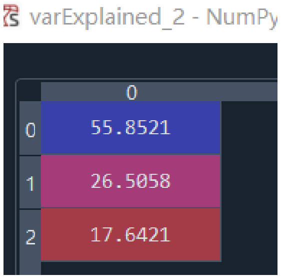
```
-0.040036638579883799
-0.14820748291196792
-0.10510600477833791
```

For question 8-9, I also tried the decision tree regression model. But the model has a score -0.69324, which is lower than the linear regression model. So I stick to linear regression.



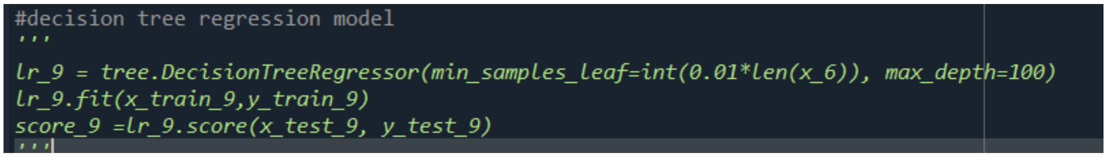
eigVals_2 - NumPy obje

| | 0 |
|---|---|
| 0 | 3.4706 |
| 1 | 1.64704 |
| 2 | 1.09626 |



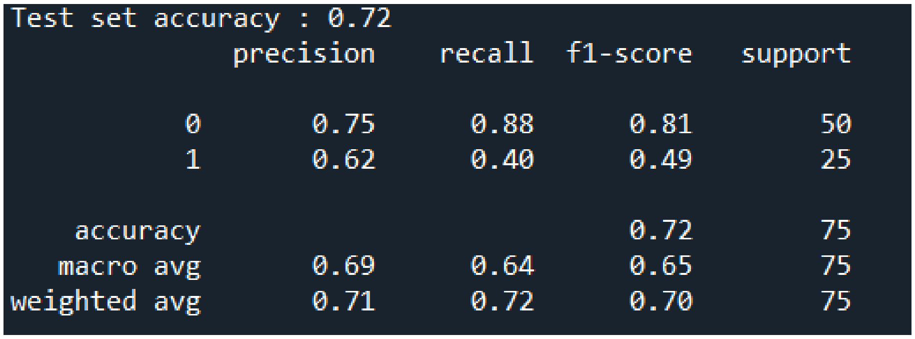
varExplained_2 - NumPy

| | 0 |
|---|---|
| 0 | 55.8521 |
| 1 | 26.5058 |
| 2 | 17.6421 |



```
#decision tree regression model
'''
lr_9 = tree.DecisionTreeRegressor(min_samples_leaf=int(0.01*len(x_6)), max_depth=100)
lr_9.fit(x_train_9,y_train_9)
score_9 =lr_9.score(x_test_9, y_test_9)
'''
```

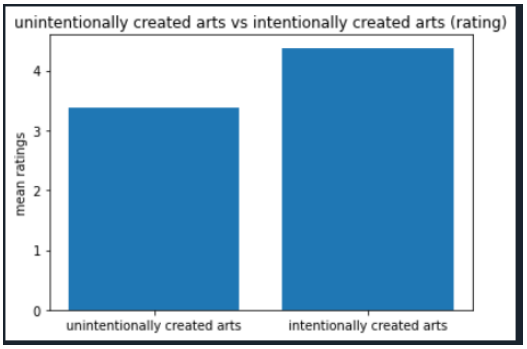
10. I use the random forest algorithm for this question and set the test_size = N as well. The model has an accuracy of 0.72, which is a pretty good score. Therefore, with all other information, it is possible to determine the political orientation of users considering the model accuracy.



```
Test set accuracy : 0.72
              precision    recall  f1-score   support

           0       0.75      0.88      0.81        50
           1       0.62      0.40      0.49        25

    accuracy                           0.72        75
   macro avg       0.69      0.64      0.65        75
weighted avg       0.71      0.72      0.70        75
```

11. **Extra Credit: I discuss the topic: The intentionally created artworks are rated higher than unintentionally created artworks.**
I test the above null hypothesis to see if it is true, since people often say the best art comes from random inspirations. There are more intentionally created artworks than unintentionally created artworks in the data. Then I implement the Mann-Whitney U test.

The null hypothesis is the mean rating for intentionally created art is higher than or equal to the rating for unintentionally created art. This is a one-tailed test. The p-value of the U-test is higher than the threshold value 0.05, therefore we fail to reject the null hypothesis. Next, I find their mean difference and median difference to assure my conclusion. The internationally created art has a higher value in both mean and median. Finally, the conclusion drawn from the U-test can be reconfirmed. It is pretty surprising, at least to me, to see that intentionally created artworks have a better rating than the unintentional ones.


unintentionally created arts vs intentionally created arts (rating)
mean ratings
unintentionally created arts
intentionally created arts