Arjun Rajpal
Professor Wallisch
Introduction to Data Science

Capstone Project

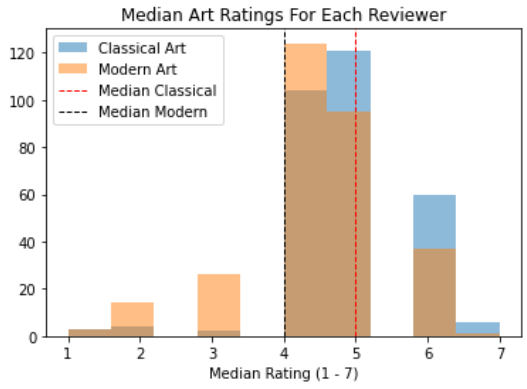
Important Notes

About Data Cleaning/Dimension Reduction/Data Transformations

- **I seeded the random numbers in my file with my N number at the top**
- Not to mention - cross validation with 80/20 split was done throughout the project to fit regression models
- **Dimension Reduction** was done with PCA for the questions responses about normalized Dark Personality Traits, Action Preferences, and Self-image/self-esteem data. I made sure to z-score my matrix before doing PCA as PCA requires normalized data.
- **Data Cleaning** was done row-wise to remove reviewers/participants who failed to answer every question/provide a rating for all art works (I removed rows with missing data). **I refer to this with statements like "row-wise" removal of nans.**
- **Data Transformations:**
    - **Preference Ratings/Energy Ratings** - for several of the prompts below, I used the median score that each reviewer gave as a representation of Preference Ratings. **I refer to this process as getting the row-wise median.** I chose to do this because it is not practical to fit 91 different models for each column of data (this would theoretically be required if I didn't do this).
    - **PCA** - I z-scored/normalized the data before doing my PCA analysis

1. **Is classical art more well liked than modern art?**

To answer this question, I gathered the columns of each respective art type (modern/classical). I combined these two arrays into one array and cleaned the data by removing rows with missing values. From here, I proceed to calculate the median rating that each reviewer gave for each respective art group. Note that the mean is not reasonable to use to summarize preference ratings, hence medians were chosen.

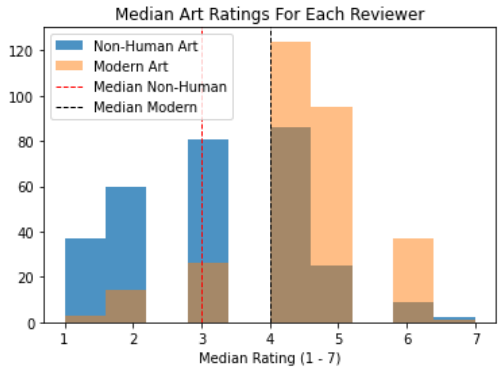


Considering that the classical art sample has a median of 5 and that the sample of modern art had a median of 4 (they are different), I decided to use a Mann-Whitney U test to compare the medians of the two samples and got a p-value of approximately 1.272e-07 (pretty close to 0), which is lower than the alpha level of 0.05. With this information, I would reject the null hypothesis that there is no difference between the medians of these two distributions. With this

information I would be inclined to argue that classical art is more well-liked than modern art but, with a visual inspection, I would add that the distribution of ratings would suggest that the difference isn't that big.

2. **Is there a difference in the preference ratings for modern art vs. non-human (animals and computers) generated art?**

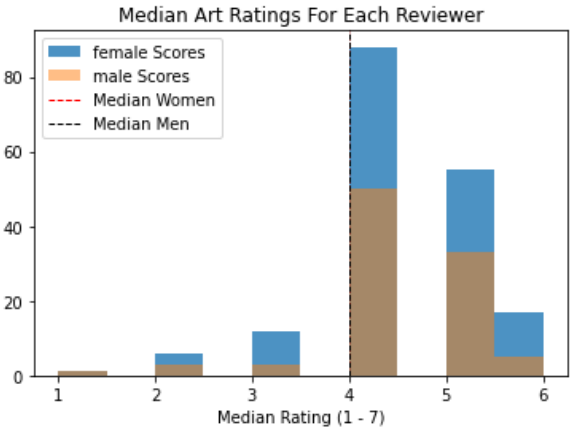
In order to answer this question, I gathered the columns of each respective art type (modern/non-human). I combined these two arrays into one array and cleaned the data row-wise. From here, I proceed to calculate the median rating that each reviewer gave for each respective art group. Note that the mean is not reasonable to use to summarize preference ratings, hence medians were chosen.



Considering the image on the right and the visual difference between these two distributions and a clear difference, I decided to use a Mann-Whitney U test to compare the medians of the two distributions. This test yielded a p-value of approximately 2.214e-32 which is well below the alpha level of 0.05, which indicates that the two distributions of median art preference ratings are significantly different. With this information and also a visual inspection of the graph above, I would argue that there is a difference between modern art vs. non-human (animals and computers) generated art.

3. **Do women give higher art preference ratings than men?**

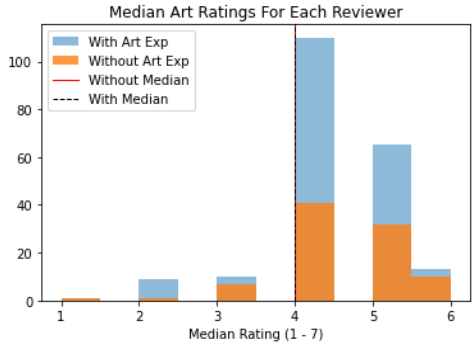
In order to answer this question, I first split the preference data into two samples by gender (male/female) and calculated the average (median) reviews of each respective reviewer. On the right is a plot of these samples. I did this split in this manner because the mean would be poorly suited to handle ordinal data like user ratings.



Then, I ran a Mann-Whitney U test to compare the two medians of the two samples and got a p-value of 0.987, which is well above the threshold of 0.05. This test then indicates that we failed to reject the null hypothesis that women do not give higher art preference ratings. With this (and a visual inspection of the plot on the right and the fact that the medians are also the same at 4), I would argue that men and women give similar ratings.

Arjun Rajpal
Professor Wallisch
Introduction to Data Science

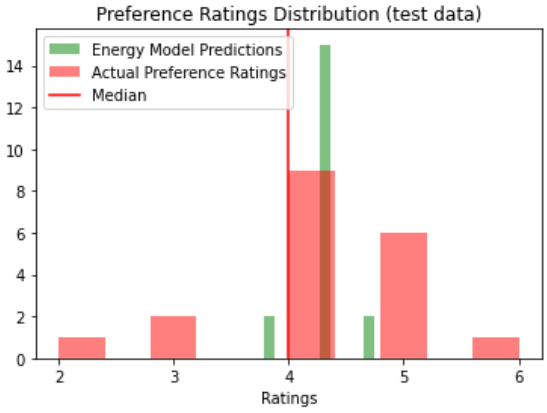
4. **Is there a difference in the preference ratings of users with some art background (some art education) vs. none?**



To answer this question, I first split the preference data into two samples by art experience (some art education/without art education) and calculated the average (median) reviews of each respective reviewer. See the plot on the right for the distribution of these median art preference ratings. I did this because the mean would be poorly suited to handle ordinal data like user ratings.

To see if these two samples come from the same population, I ran a Mann-Whitney U test to compare the medians and got a p-value of 0.189 which is well above the alpha value of 0.05. This would indicate that I fail to reject the null hypothesis that there is no difference in the preference ratings (the distributions of the medians) between users with some art background and no background. Also given the figure on the right and the fact that the medians are also the same (at a rating of 4), I would argue that there is no difference in the preference ratings of users.

5. **Build a regression model to predict art preference ratings from energy ratings only. Make sure to use cross-validation methods to avoid overfitting and characterize how well your model predicts art preference ratings.**



I built a regression model to predict art preference ratings based on energy ratings. I first cleaned and summarized the data by calculating the median ratings for both the energy and preference ratings (row-wise for each reviewer). I then assessed the relationship between these two variables using the Spearman and Pearson coefficients and found that the relationship was weak. I split the data into a training and testing set (80/20 split) and fit a linear regression model on the training set. When I evaluated the model using the testing set, I found that it had a Root Mean Squared Error (RMSE) of approximately 0.8697, which is not surprising given the weak relationship between the variables and the domain knowledge that scores can only take discrete numerical values between 1 and 7. I also plotted the predicted and actual values for the testing

set to visualize the fit of the model. (see model above). Overall, my model does a mediocre job of predicting values.

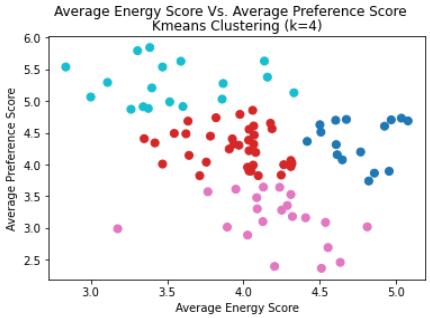
Note: I could have fit 91 models for each art piece but, since the question asked for a single model, I was forced to reduce each row of data to a summary statistic. Not to mention, when I fit a classifier model to this data I was able to get better results (perhaps this would be a better model but, since the question explicitly asked for regression I didn't use something like a random forest classifier)

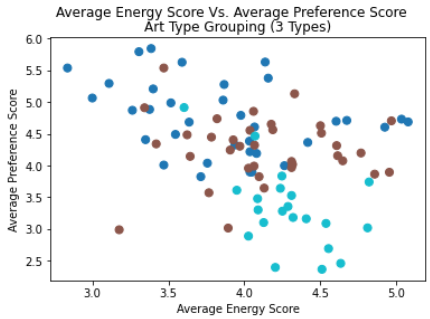
6. **Build a regression model to predict art preference ratings from energy ratings and demographic information. Make sure to use cross-validation methods to avoid overfitting and comment on how well your model predicts relative to the "energy ratings only" model.**

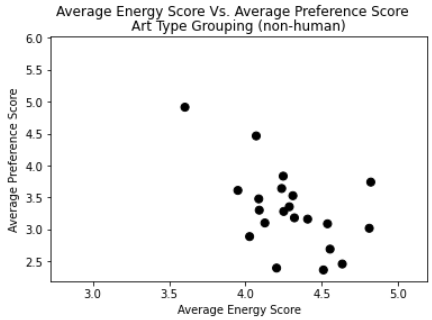
I cleaned and processed the preference ratings and energy ratings by simplifying them to row-wise medians and combined them into a single array along with the demographic data. During this step I dropped rows with missing data. I then checked the correlations of all the predictors to the output feature and found that none of them had a strong linear correlation (visually with a correlation matrix plot). I then split the data into an 80/20 train-test split and trained a linear regression model on the training data. When I used this model to predict the preference scores for the test data, I obtained a root mean squared error (RMSE) of approximately 0.907, which is higher than the energy model in question 5. This suggests that the energy model may be more effective at predicting preference scores than the energy + demographic information model that was fitted for this question. The distribution of predicted values and correct values for the test data is shown in the plot above.

Arjun Rajpal
Professor Wallisch
Introduction to Data Science

7. **Considering the 2D space of average preference ratings vs. average energy rating (that contains the 91 art pieces as elements), how many clusters can you – algorithmically - identify in this space? Make sure to comment on the identity of the clusters – do they correspond to particular types of art?**



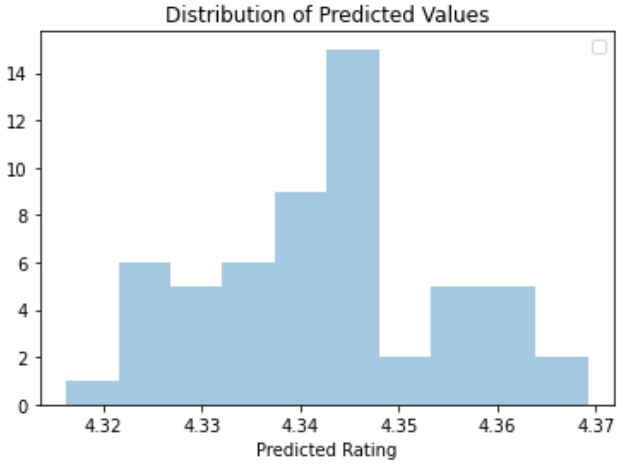




After cleaning the data as described in the first section, I first calculated a column of means for both the preference ratings and energy ratings. With this data, I fit several Kmeans models with different k-values and I evaluated the Silhouette Scores of each of these models to find the best k-value. With this analysis, I determined that the best k value was 4 with a silhouette score of 0.40578338 (see clustering with image on the right). From a visual perspective, I can see that the two clusters are rather different - the only exception being that the turquoise and pink points seem to be similarly grouped. From a visual inspection, these turquoise data points represent non-human art, as seen through a comparison of the turquoise points and the bottom graph which just shows only the nonhuman reviews. From this, I would conclude that the clusters correspond to only non-human art despite the k value of 4 and there being three art types.

Arjun Rajpal
Professor Wallisch
Introduction to Data Science

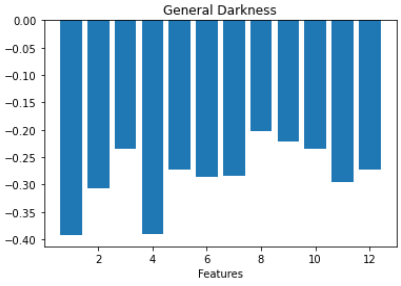
8. **Considering only the first principal component of the self-image ratings as inputs to a regression model – how well can you predict art preference ratings from that factor alone?**

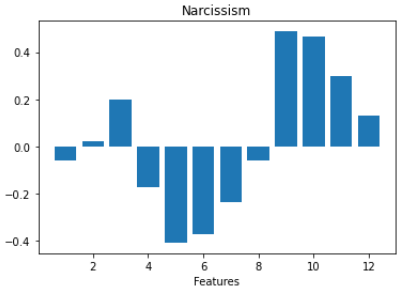


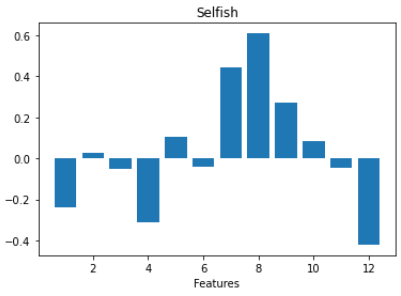
After cleaning and gathering data for self-image ratings, preference ratings were also collected and reduced to their row-wise medians. All values were z-scored and a PCA model was fit to the data. Using the Kaiser Criterion method, it was determined that the first two components of the model accounted for over 90% of the variance. The first principal component was found to represent an overall positive self-image, as all values in the loading were above .2. A linear regression model was then fit to the data using the first principal component. This model had a root mean squared error (RMSE) of approximately 0.990. It is worth noting that this model may not be an appropriate way to model ordinal data on a scale of 1 through 7 because people may have their own understanding of the rating system and the difference between ratings may not be consistent. Additionally, a high RMSE close to 1 is generally not desirable, even at this scale. So this model doesn't predict art preference ratings particularly well, not to mention that it will predict continuous values which are hard to interpret when I consider that ratings are given on a scale of 1 to 7. With other methods like classification, we might get better results than with using regression (perhaps a random forest classification model would work better - this will apply to other regression based questions as well).

Arjun Rajpal
Professor Wallisch
Introduction to Data Science

9. **Consider the first 3 principal components of the "dark personality" traits – use these as inputs to a regression model to predict art preference ratings. Which of these components significantly predict art preference ratings? Comment on the likely identity of these factors (e.g. narcissism, manipulativeness, callousness, etc.).**



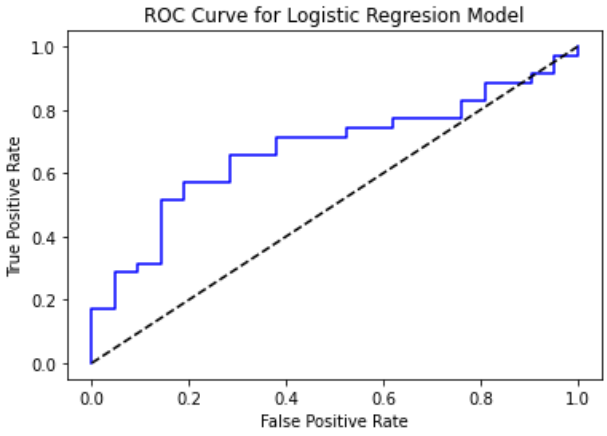




After cleaning and gathering data for dark personality traits and preference scores, the data was z-scored and a PCA model was fit. Using the Kaiser Criterion method, it was determined that the first three components of the model accounted for over 90% of the variance (which is consistent with what the question was asking for). With the loadings of each component, I was able to interpret the meaning of each component. The first component was found to represent general darkness (all features), the second component represented narcissism (features 5, 6 9, 10), and the third component represented selfishness (features 5, 6, 9, 10). I interpreted these results in this way but it is in a way a little subjective as to what these components represent. With this information I did an 80/20 split of the data, I fit a ridge regression model using these three components using the training split, with the alpha value optimized. This model had a root mean squared error (RMSE) of approximately 0.765 when I used the test data to predict values, which is better than the models previously produced in this report. The coefficients for the three components were -0.00402499, 0.06288727, and 0.08698904, respectively, indicating that narcissism and selfishness significantly predict art preferences as these two coefficients are further away from zero compared to the first one.

Arjun Rajpal
Professor Wallisch
Introduction to Data Science

10. **Can you determine the political orientation of the users (to simplify things and avoid gross class imbalance issues, you can consider just 2 classes: "left" (progressive & liberal) vs. "non- left" (everyone else)) from all the other information available, using any classification model of your choice? Make sure to comment on the classification quality of this model.**



To do this, I cleaned the data (removal of nans row-wise) and I converted the political feature data to 1s and 0s (1 representing left and 0 representing non-left). Then, I used a PCA model to try and reduce the number of features of the dataset (using all the features - z-scored). Using the Kaiser Criterion method, I determined that the first 64 components accounted for over 90% of the variance. In regards to this and the fact that the political data is binary, I decided to fit a Logistic Regression model to predict political orientation. I fit this model using an 80/20 split of the data for cross validation purposes. When testing the model's accuracy with the test data, it produced an AUC score of approximately 0.638, which means it is better than randomly guessing (see ROC Curve on the right). But it is not by much, based on this information, it would be difficult to predict the political affiliation of the users with all the other information in this data set.