Capstone Project

Introduction to Data Science

Joaquin Navarrete

May 10, 2023

Introduction

1. Is classical art more well-liked than modern art?
2. Is there a difference in the preference ratings for modern art vs. non-human (animals and computers) generated art?
3. Do women give higher art preference ratings than men?
4. Is there a difference in the preference ratings of users with some art background (some art education) vs. none?
5. Build a regression model to predict art preference ratings from energy ratings only. Use cross-validation methods to avoid overfitting and characterize how well your model predicts art preference ratings.
6. Build a regression model to predict art preference ratings from energy ratings and demographic information. Use cross-validation methods to avoid overfitting and comment on how well your model predicts relative to the "energy ratings only" model.
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?
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?
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.).
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. "nonleft" (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.

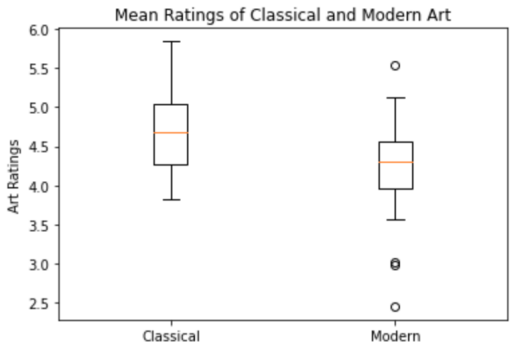
Data Cleaning: Before any analysis was performed, I cleaned the "art_data" dataset. I made the column titles in an ordered format and assigned appropriate names to each column to easily transverse and identify them. Throughout the analysis, more data cleaning was performed on the dataset to conduct the specific research (merging data frames, creating subset data frames to train models, creating new analytical tables with averages, etc.)

1)

Process of analysis:

- Central tendency differences between “classical art and modern art”
- One-sided Mann-Whitney U test to compare if the difference between a rating of classical and modern art is statistically significant

Findings:




```
Mean Clasical art:  4.707142857142856
Median Clasical art:  5.0
Mean Modern art:  4.210380952380954
Median Modern art:  4.0
```

Answer:

The ratings for classical art are significantly higher than those for modern art.

Explanation:

We understand that classical artwork's central tendency measures are higher than modern art's. Then we conducted a one-sided Mann-Whitney U test to know if this difference was statistical. I chose this test because it is more robust when we don’t know the specific distribution of the data. Based on the result, we rejected the null hypothesis. For this reason, there is statistically significant evidence based on these samples that classical art is more well-liked than modern art.

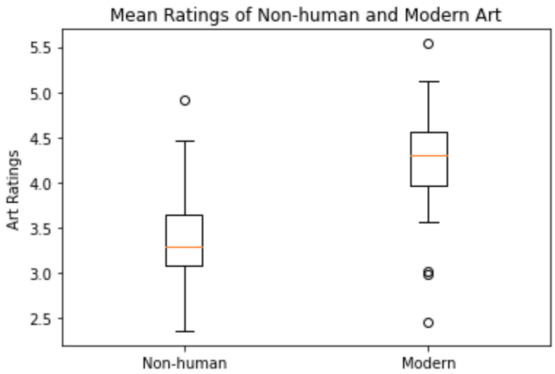
2)

Process of analysis:

- Central tendency differences between “classical art and modern art”

- Perform a 'two-sided' Mann-Whitney U test to show if there is a significant difference between ratings.

Findings:




```
Non-human Rating Mean:  3.375238095238095
Modern Art Rating Mean:  4.210380952380954
Non-human Rating Median:  3.0
Modern Art Rating Median:  4.0
```

Answer:

The ratings for 'non_human' and ‘modern artwork” are significantly different.

Explanation:

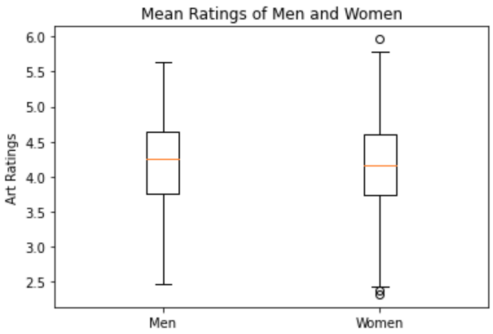
We observe the difference between central tendencies of 'non_human' vs 'modern' artwork ratings. Then we perform a 'two-sided' Mann-Whitney U test to conclude that the difference in ratings between non-human and modern art is statistically significant based on this dataset.

3)

Process of Analysis:

- Find central tendency differences between ratings by 'women' compared to ratings by "men"
- Perform the Welsh-t test to check if the difference is statistically significant.

Findings:




```
Men Rating Mean:  4.17912087912088
Women Rating Mean:  4.147852147852149
Men Rating Median:  4.0
Women Rating Median:  4.0
```

Answer:

Women do not rate higher than men at a statistically significant level.

Explanation:

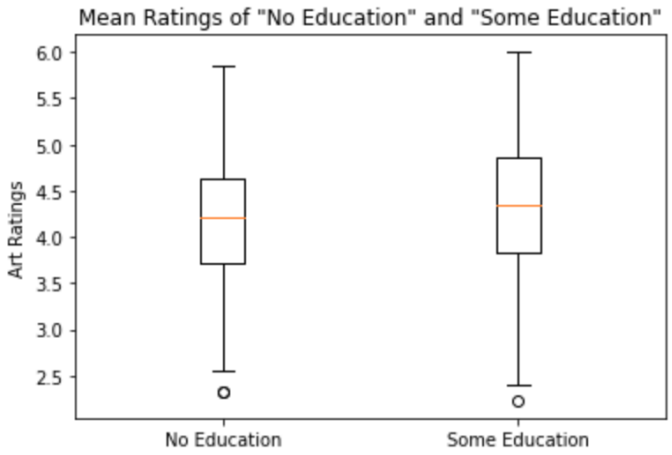
We understood the difference in central tendencies of "Men" ratings vs "Women" ratings and learned their means differed, but medians were equal. We then tested to see if this result was statistically significant using a one-tailed Welsh-t-test to understand the difference we found in their mean and determined the ratings by women are not significantly higher than men according to this dataset.

4)

Process of Analysis:

- Compare central tendency measures of ratings between people who have had some "art education" vs those who don't
- Test results with statistical significance tests

Findings:




```
Some Education Mean:  4.142615626132109
No Education Mean:  4.294529646177998
Some Education Median:  4.0
No Education Median:  4.0
```

Answer:

The ratings by 'some_education_ratings' are significantly different than 'no_education_ratings.'

Explanation:

After comparing central tendency measures between the ratings of those with some artistic education and those with none, we found a difference between the mean ratings. We then conducted a two-sided Welsh-t-test to test whether this difference was statistically significant.

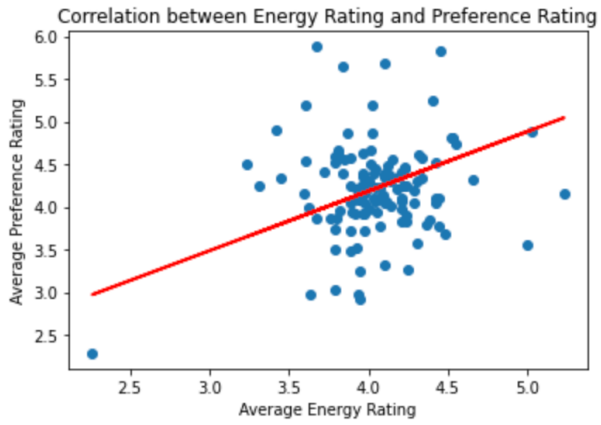
5)

Process of Analysis:

- Perform EDA: Find correlation between energy ratings and art preference ratings, outliers, and distribution of ratings
- Build a regression model to predict art preference ratings from energy ratings only
- Analyze findings

Findings:

```
Pearson correlation coefficient: -0.0156
Spearman coefficient: 0.1681
```




```
R^2 -0.0468221954358079
Root Mean Squared Error (RMSE): 0.5624
```

Answer:

Energy ratings by themselves are not reliable predictors of preference ratings.

Explanation:

The linear regression model is not a reliable model to predict preference ratings from energy ratings. Before performing the regression I analyzed the correlation between these 2 variables. Because the linear correlation by the Pearson coefficient was so low, I also calculated the Spearmen in order to better understand how these two variables were related if not linear. The monotonic relationship coefficient was higher than the linear correlation so this was already a strong indicator that the regression model was not going to be that accurate. The result of the regression gave a very low R^2 and considerably high RMSE, confirming the linear relationship of energy and preference ratings is reliable.

6)

Process of Analysis:

- Assuming demographic information refers to these three columns: Column 216: User age Column 217: User gender (1 = male, 2 = female, 3 = non-binary) Column 218: Political orientation (1 = progressive, 2 = liberal, 3 = moderate, 4 = conservative, 5 = libertarian, 6 = independent)
- Conduct correlation tests to first understand possible relationships between demographics and energy ratings as well as preference ratings, this could be helpful to reduce confounders
- Conduct multivariable linear regression to predict ratings from energy ratings as well as demographics

Findings:

	avg_preference_ratings	avg_energy_ratings	age	gender
avg_preference_ratings	1.000000	0.320264	-0.108275	0.040645
avg_energy_ratings	0.320264	1.000000	0.156726	0.017537
age	-0.108275	0.156726	1.000000	-0.122822
gender	0.040645	0.017537	-0.122822	1.000000


```
R-squared: 0.20508343516610827
RMSE: 0.6247885167384617
```

Answer:

Predicting preference ratings from demographics and energy ratings is more reliable than using energy ratings alone, but it's still not a robust model.

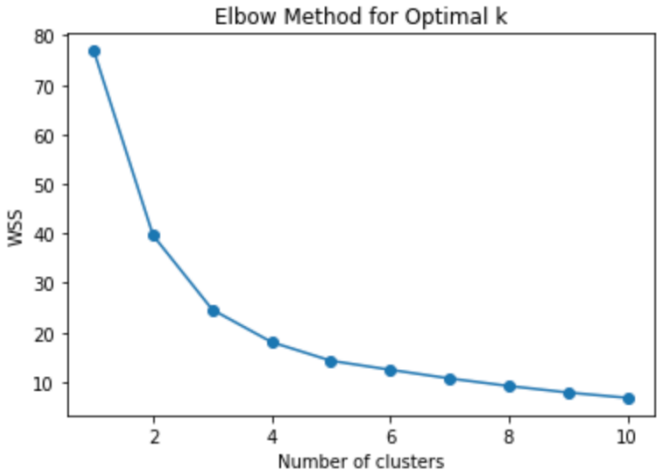
Explanation:
The R^2 of the model is much higher than the model predicting preference ratings from energy ratings alone, but it still only explains 20% of the variance of preference ratings. The RMSE also indicates that the error of the model is larger than .5 of ratings, which indicates it is not very accurate.

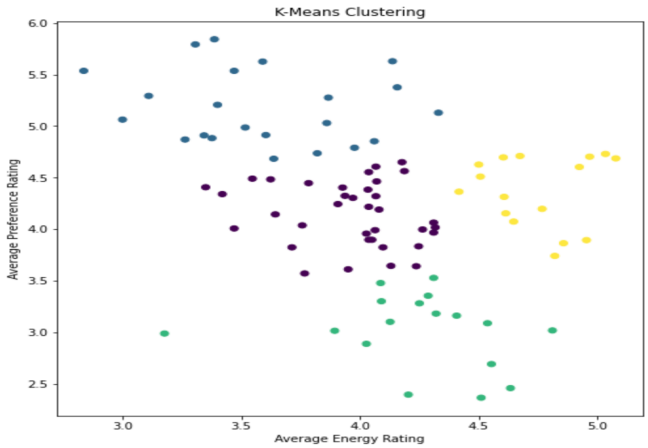
7)

Process of Analysis:

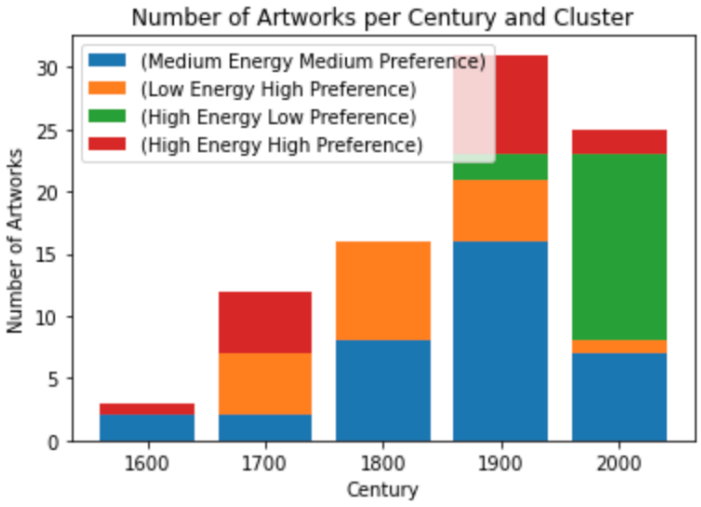
- I used the elbow method to identify the number of clusters to use in the K-means algorithm.
- After Identifying the optimal number of clusters for my algorithm I will perform a K-means classification to distinguish both clusters and make inferences on how they are different

Findings:





Extra Credit*



Answer:

There are four meaningful clusters in the 2D space of average preference ratings vs. average energy ratings. They represent artworks with ratings of, (Medium Energy Medium Preference), (Low Energy High Preference), (High Energy Low Preference), (High Energy High Preference).

Explanation:

Based on the elbow method, we understand there are four meaningful clusters for this analysis. When understanding these clusters we see they are different because of their energy and preference ratings. To understand more about the identity of these clusters I associated the individual artworks that belonged in each cluster and identifies their corresponding date. I then analyzed what was the modal century where each cluster's artworks were made (the most common century for each cluster). This led to the understanding of how the time when the artwork was constructed relates to how people make their preferences and energy ratings. I found these results, which could lead to further understanding of the relationship of the artworks date with how people feel about the artwork, there were significant differences between the modal time period of each cluster.

Cluster 0 (Medium Energy Medium Preference): Modal Decade of artworks in this cluster: 20

Cluster 1 (Low Energy High Preference): Modal Decade of artworks in this cluster: 18

Cluster 2 (High Energy Low Preference): Modal Decade of artworks in this cluster: 20

Cluster 3 (High Energy High Preference): Modal Decade of artworks in this cluster: 19

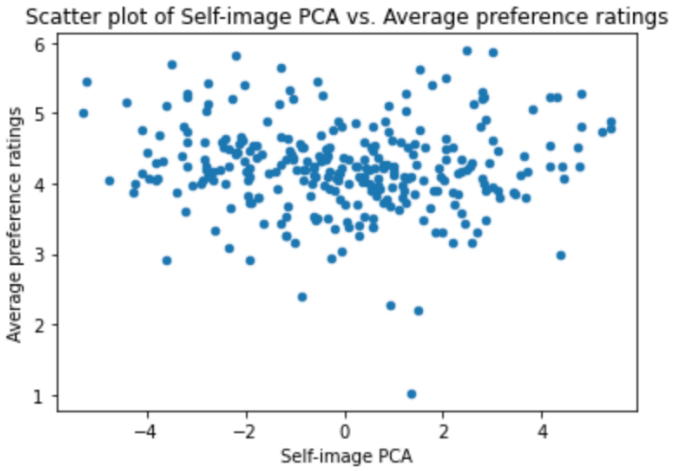
8)

Process of Analysis:

- Identify "self_image" ratings
- Perform PCA of this inputs
- Create Screeplot
- Understand what is the first principal component and isolate this data to become useful for analysis
- Perform regression from only that factor

- 6) Analyze regression results

Findings:




```
RMSE: 0.6361
R-squared: 0.0017
```

Answer:

The first PCA component of self-image ratings is not a reliable predictor of preference ratings.

Explanation:

After performing the PCA algorithm on self-image ratings, I created a new data frame with average preference ratings for each artwork as well as the corresponding PCA rating. After constructing the data frame I found the correlation between this PCA component and the ratings, which was considerably low: -0.04140. After I performed the regression Which was not considered accurate, with a very low R^2, showing that the model was only explaining .0017 of the variance of preference ratings.

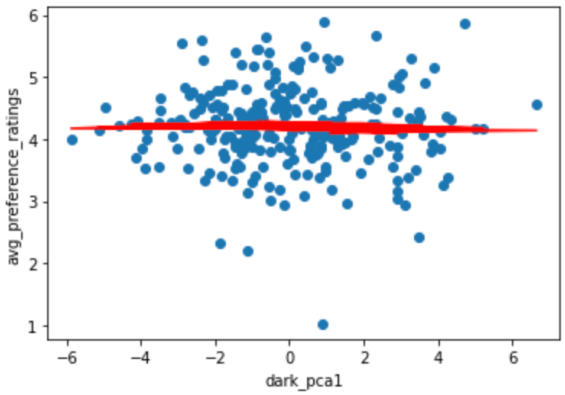
9)

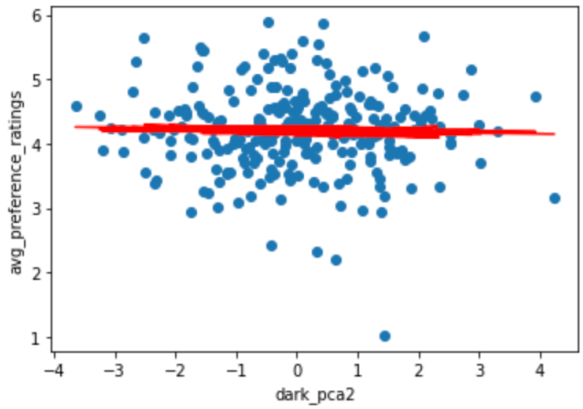
Process of Analysis:

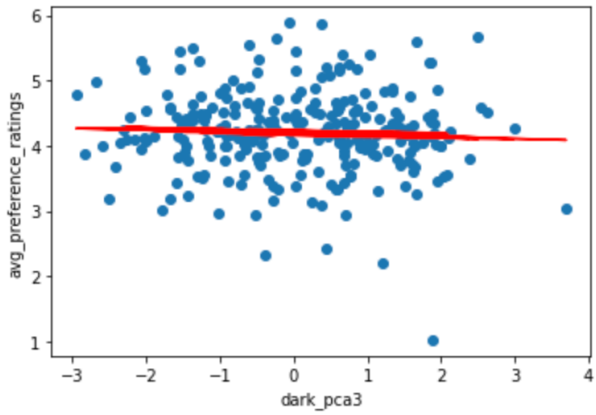
- Identify the first 3 principal components of the “dark personality” traits
- Analyze each loadings matrix of the principal components
- Prepare regression data frame and perform EDA
- Build multilinear regression model

Findings:

	avg_preference_ratings	dark_pca1	dark_pca2	dark_pca3
avg_preference_ratings	1.000000	-0.032880	-0.038361	-0.050866
dark_pca1	-0.032880	1.000000	0.027226	0.000077
dark_pca2	-0.038361	0.027226	1.000000	0.040767
dark_pca3	-0.050866	0.000077	0.040767	1.000000


avg_preference_ratings
dark_pca1


avg_preference_ratings
dark_pca2


avg_preference_ratings
dark_pca3

```
R-squared for dark_pca1: 0.004923759431549546
RMSE for dark_pca1: 0.6324553743470674
R-squared for dark_pca2: 0.00419778693945716
RMSE for dark_pca2: 0.6326860408387585
R-squared for dark_pca3: 0.003129521774344779
RMSE for dark_pca3: 0.6330253126747974
R-squared for all predictors combined: 0.004923759431549546
RMSE for all predictors combined: 0.6324553743470674
```

Answer:

After performing the PCA algorithm on the “dark personality” traits, I analyzed each loadings matrix to identify the individual questions that had the most correlation with each PCA. Although the PCA analysis was successful to reduce the dimensions of the model, the results of the multilinear regression did not show that these PCA components were strong predictors of preference ratings.

PC 1:
Questions with the most influence on PC 1: 1: 'I tend to manipulate others to get my way', 4: 'I tend to exploit others towards my own end'
PC 2:
Questions with the most influence on PC 2: 3: 'I have used flattery to get my way', 5: 'I tend to lack remorse', 10: ' I tend to want others to pay attention to me'
PC 3:
Questions with the most influence on PC 3: 8: 'I tend to be cynical'

Explanation:

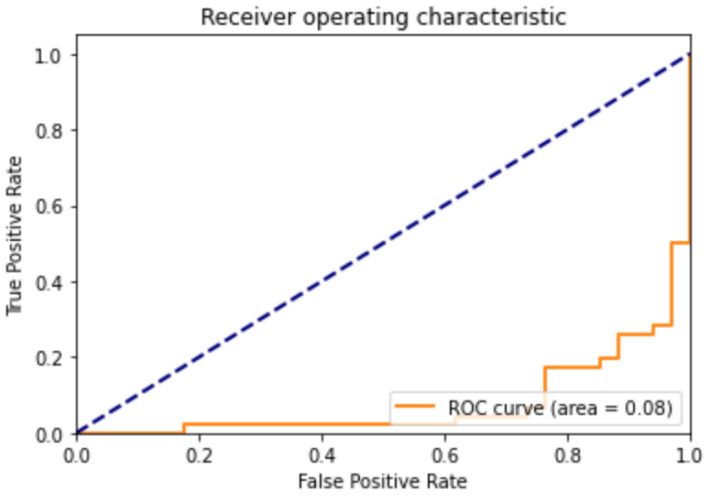
After identifying the first three principal components of the “dark personality” traits, I tested the correlation between each individual component and preference ratings of artworks. There was no component that showed a strong correlation with preference ratings individually which was a first indicator that the linear regression model was not going to be significant. The results of the regression confirmed this hypothesis as the variance explained by the model as a whole was extremely low and there was no significant linear relationship between any of the components with preference ratings.

10)

Process of Analysis:

- Choose appropriate classification model
- Perform PCA analysis on action traits
- Build data frame with all necessary data for model
- Train model and analyze results

Findings:




```
Sensitivity: 0.7647058823529411
Specificity: 0.8913043478260869

Accuracy: 0.8375
```

Answer:

While the accuracy of the model is considerably high, there are other metrics that show that the model has significant deficiencies.

Explanation:

I chose to use logistic regression because we were looking for a binary classification system. Logistic regression is an accurate system for a binary classification system because it gives the probability of every data point falling into each of the categories, in this case, left or non-left political association. We can understand that while the model has an accuracy of over 80% and individually it is able to predict true positives with 76% accuracy and false negatives with a high 89% accuracy, the area under the curve analysis of the ROC is very low. This means that the model is not dependable when the thresholds of the model change. The ROC curve measures the results of the model with different thresholds and the larger the AUC the more robust the model is. The conclusion to this analysis is that although the model can make considerably accurate predictions, it is likely not dependable enough for different datasets or different scales.