Luke Psyhogios

20 December 2022

Intro to Data Science

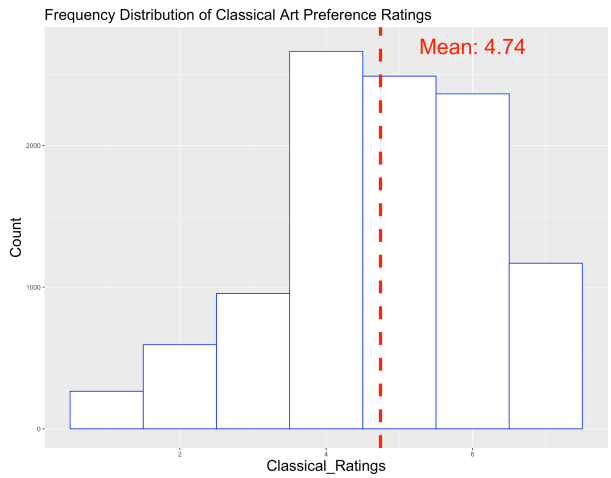
# Capstone Project

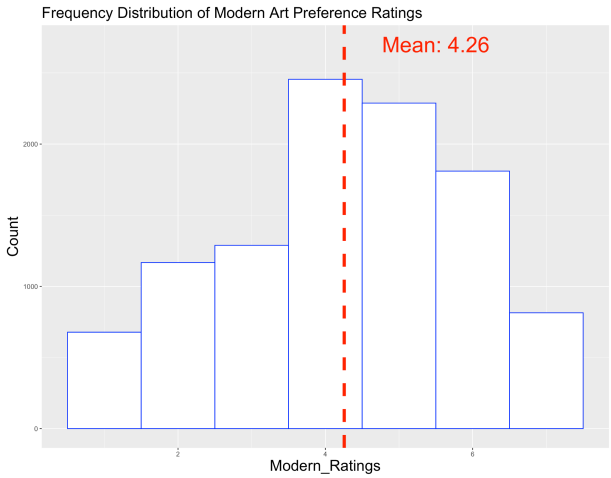
To complete this project I used Rstudio. To import the CSV into R I used the read_csv function from the tidyr package. I used tidyr's functions extensively to reshape and reformat my data. I got rid of NaNs whenever they were prevalent in a question by using the na.omit() function. I used the principal() function from the 'psych' package to perform dimension reduction by means of PCA.

All plots (excluding boxplots for which I used the built in R function) were made using ggplot2.

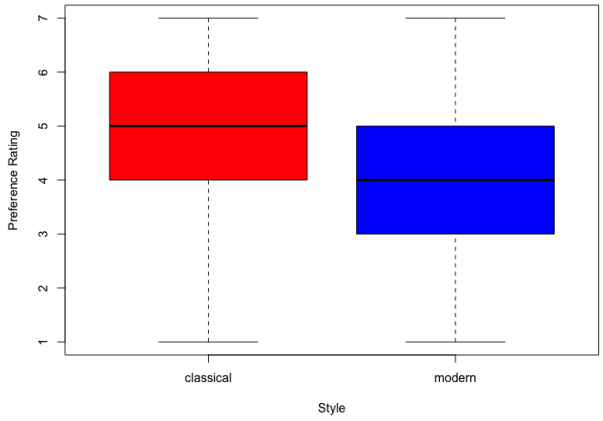
1) Is classical art more well liked than modern art?

I first organized all the classical art and modern art preference ratings into their own variables using the pivot_longer() function. I then had to decide which statistical test I would use to compare the central tendency statistics of my two variables, 'Classical_Ratings' and 'Modern_Ratings.' To make this decision, I checked to see if their distributions were normal.



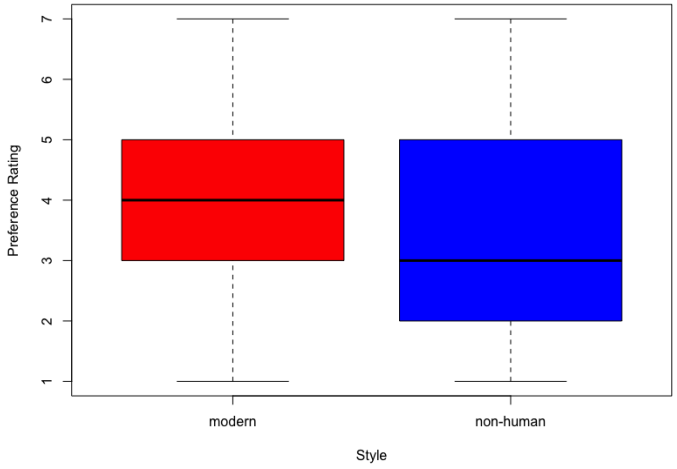
As shown by these figures, both distributions of preference ratings are left skewed, which led me to use a Wilcoxon rank sum test. My null hypothesis is that there is no difference between the medians of 'Classical_Ratings' and 'Modern_Ratings', and my alpha level is 0.05. The Wilcoxon rank-sum test resulted in a test statistic W = 64,145,482: W (in this case) is the same as the test statistic U from the Mann-Whitney U test (see here). The test gave a p-value of 3.18e-97. Based on the p-value which was far below that of my alpha-value, I reject the null hypothesis and conclude that **classical art is more well liked than modern art.** Below is a box plot of the two samples, with the median at Q2.




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

My first step was to get all the non-human preference ratings as observations into a variable titled 'Modern_Ratings.' I called the pivot_longer() function to filter all of my desired preference ratings into the variable. As shown by the modern art histogram in question one, the modern art preference ratings are not normally distributed, so I will use another Wilcoxon rank sum test.

My null hypothesis is that there is no difference between the medians of 'Classical_Ratings' and 'Non_Human_Ratings' and I once again used a p-value of 0.05. The results of the Wilcoxon rank sum test were a p-value of 8.74e-264 and a W with a value of 43,486,536. Based on the p-value smaller than that of my alpha value, I reject the null hypothesis and conclude that **there is a difference in the preference ratings for modern art vs. non-human art. Modern art is more well liked than non-human art.** The box plot below supports these findings, as the medians are different.

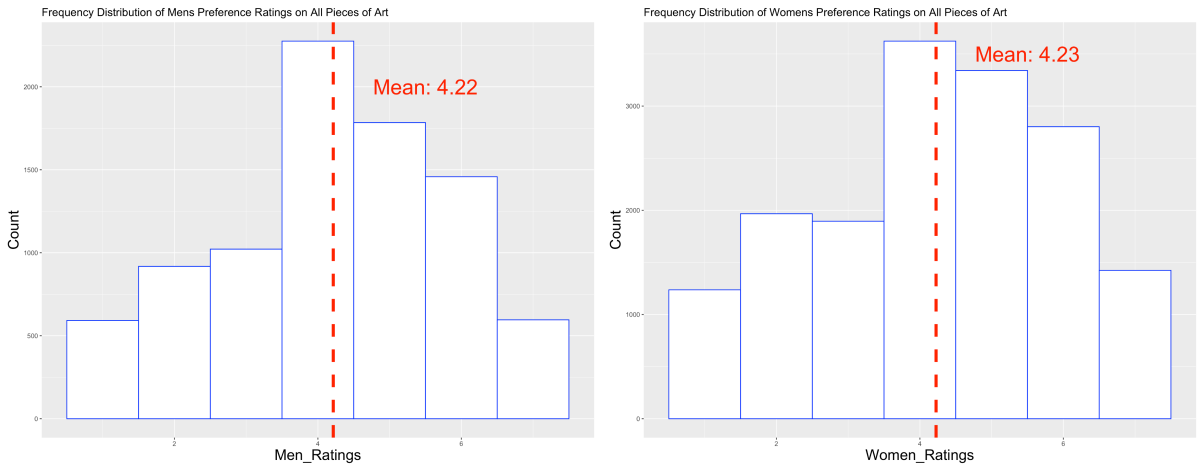



3) Do women give higher art preference ratings than men?

In order to answer this question I needed to isolate the art preference ratings of men and women into their own variables, titled 'men_preferences' and 'women_preferences'. I did so by executing the code block below:

```r
#Filter out non-binary participants and participants who did not disclose their gender
gender_df <- theData[!is.na(theData$Gender) & theData$Gender !=3,]

#Get the art preference ratings of men and women in their own data frames
Women_Ratings_df <- gender_df[gender_df$Gender == 2, 1:91]
Men_Ratings_df <- gender_df[gender_df$Gender == 1, 1:91]

#Get all observations of art preference ratings into their own variables.
men_df <- data.frame(matrix(nrow = 8645, ncol = 2))
colnames(men_df) <- c('Painting_Names',"Men_Ratings")
men_df <- pivot_longer(Men_Ratings_df,1:91,names_to = 'Painting_Names',values_to = 'Men_Ratings')

women_df <- data.frame(matrix(nrow = 16289, ncol = 2))
colnames(women_df) <- c('Painting_Names',"Women_Ratings")
women_df <- pivot_longer(Women_Ratings_df,1:91,names_to = 'Painting_Names',values_to = 'Women_Ratings')

#Get the variables of art preference ratings of men and women into the R global environment
women_preferences <- women_df$Women_Ratings
men_preferences <- men_df$Men_Ratings
```

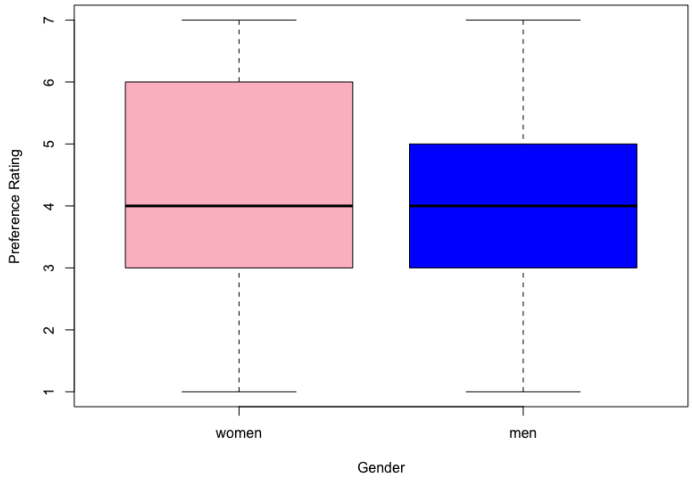
I then had to decide which statistical test to run to compare the preference ratings, so I plotted the distribution of my two variables again.




Once again, the preference ratings were not normally distributed, as both distributions are slightly left skewed, meaning another Wilcoxon rank sum test was in order.

My null hypothesis was that there is no difference between the medians of ‘women_preferences’ and ‘men_preferences’ and I used an alpha level of 0.05. The results of the Wilcoxon rank sum test were a p-value of 0.27 and a W equal to 70,994,990. Based on the given p-value which is larger than my alpha value, I fail to reject the null hypothesis and conclude that **women <u>do not</u>**

**<u>give higher art preference ratings</u> than men.** Below is a boxplot of women and men art preference ratings, and Q2 is equal for both variables, which does not necessarily prove my test results as accurate, but does support them.

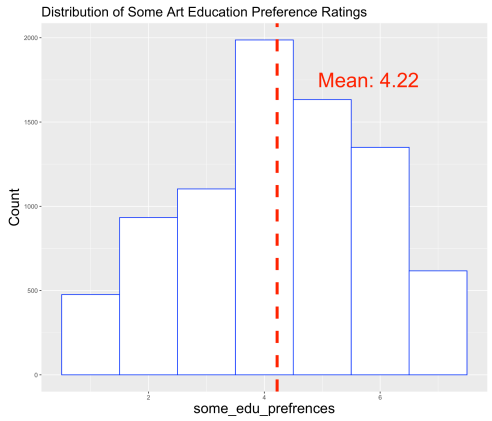


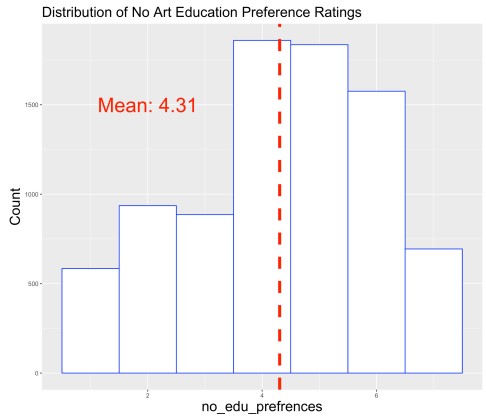

4) Is there a difference in the preference ratings of users with some art background (some art education) vs. none?

The first course of action when answering this question is to determine what qualifies as "some art background" based on the data given. I opted to include both those participants with one year and two years of art education, mainly so I could work with a larger sample size of preference ratings than only including one of the two groups. I didn't include those with three years of art education as I thought their education exceeded the threshold of 'some.' I then had to isolate two groups of art preference data, those with no previous art education and those with 1-2 years of it. I did so by executing the following block of code.

```r
#Filter out those who had zero or 3 years of art education, as well as those who did not respond and place in one df
some_education_df <- theData[!is.na(theData$`Art education`) & theData$`Art education` !=3 & theData$`Art education`
#Filter out those who had any art education and those who did not respond and place the obs. into another df
no_education_df <- theData[!is.na(theData$`Art education`) & theData$`Art education` !=3 &
                           theData$`Art education` != 1 & theData$`Art education` != 2,]

#Isolate preference ratings from both data frames
some_df <- some_education_df[some_education_df$`Art education` == 1 & 2, 1:91]
none_df <- no_education_df[no_education_df$`Art education` == 0, 1:91]

#Place all observations of art preference ratings into their own variables
some_pref_df <- data.frame(matrix(nrow = 8099, ncol = 2))
colnames(some_pref_df) <- c('Painting_Names', 'Some_Edu_Ratings')
some_pref_df <- pivot_longer(some_df,1:91,names_to = 'Painting_Names',values_to = 'Some_Edu_Ratings')

none_pref_df <- data.frame(matrix(nrow = 8372, ncol = 2))
colnames(none_pref_df) <- c('Painting_Names', 'No_Edu_Ratings')
none_pref_df <- pivot_longer(none_df,1:91,names_to = 'Painting_Names', values_to = 'No_Edu_Ratings')

#Get the variables of art preference ratings of those with some art education and those
# with no art education into the R global environment
no_edu_prefrences <- none_pref_df$No_Edu_Ratings
some_edu_prefrences <- some_pref_df$Some_Edu_Ratings
```

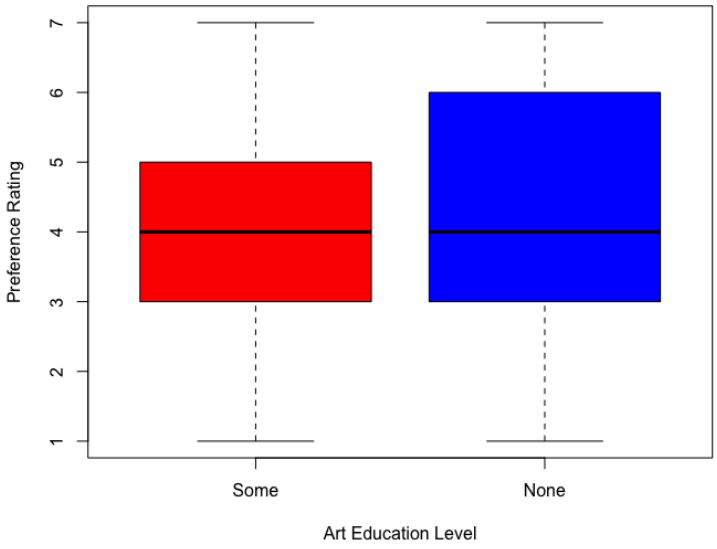
The next course of action was to check if the preference ratings were normally distributed, as this would determine whether I use a t-test or a Wilcoxon rank sum test once again.




While the preference ratings of those with some art education is approaching normal, those with no art education clearly have left skewed preference ratings so I will be opting for a Wilcoxon rank sum test. My null hypothesis is that there is no difference in the median art preference rating of those with some art education vs those with no art education and I set my alpha value at 0.05. The results of the Wilcoxon rank sum test are a W with the value of 32,632,726 and a p-value of 2.34e-05. Due to the p-value being smaller than the alpha value I reject the null hypothesis and

conclude that **there is a difference in the preference ratings of users with some art background vs those with no art background.** Those with no art education give higher art preference ratings than those with some art background. This makes sense as those with some art background will likely be harsher critics. Below is a box plot of the two groups. It's important to note that despite the two groups having the same median in this sample, the true difference between their medians is **not zero**, as shown by the low p-value.



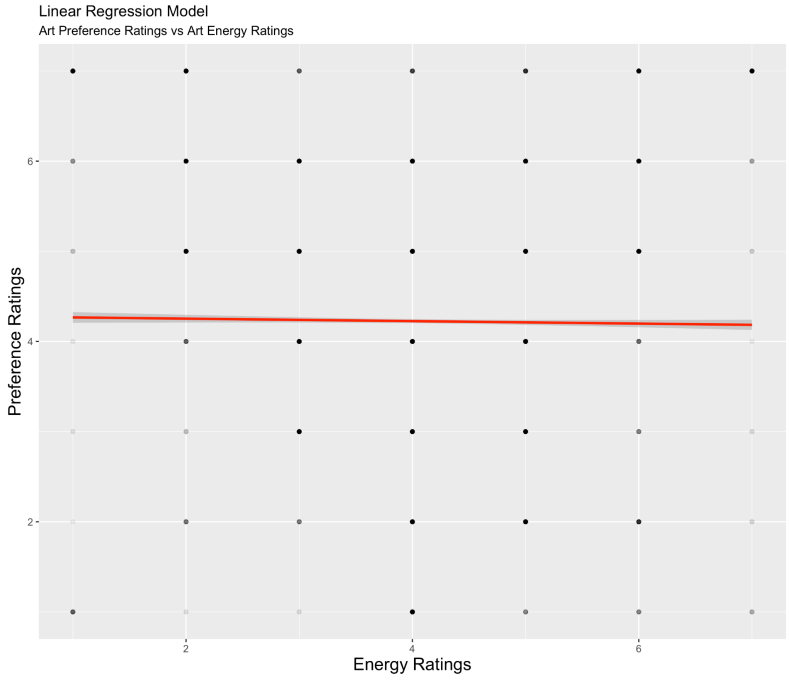

5) Build a regression model to predict art preference ratings from energy ratings only. Make sure to use cross-validation methods to avoid overfitting and characterize how well your model predicts art preference ratings.

I had to install the caret package in order to use cross-validation methods, as I did for any question henceforth where I had to split my data into a testing and training set. I first isolated the preference ratings and energy ratings into their own variables titled 'Preference_Ratings' and 'Energy_Ratings.' Upon generating my variables I checked their correlation which gave an **$R^2$ of -0.016**. While the negative correlation was to be expected since a 7 in preference rating

corresponds to loving the piece of art while a 7 in energy rating corresponds to being agitated, I was not expecting such a correlation so close to zero. This made me suspect that the model would not have much predictive power.

To cross-validate, I split my data set into a training and test set with the assistance of the 'caret' package. I called the createDataPartition() function and passed through my dependent variable, 'Preference_Ratings' and arbitrarily chose p = 0.8, meaning 80 percent of my original data would go into the training set.

I created my model using the lm() function. My linear regression model had an **$R^2$ of 0.0001043036, meaning that my model explained slightly more than .001% of the variance in preference rating. This model has virtually no predictive power.** Below is the graph of my linear model, showing the very weak predictive power of my independent variable on my dependent variable. The opacity of each dot represents the relative frequency of that result.



I then calculated the **RMSE** of the predicted values which gave a result of **1.67**. Since the ratings are on a scale from 1-7, this RMSE value is large, as the predicted values are 1.67 rating points away from the actual values on average. This is in line with my other findings that **the model does not make accurate predictions**.

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 decided to use the final six columns of the data set to make up my demographic information, and use each column as a predictor variable. I cleaned rows with NaNs in these variables using the na.omit() function. Because each person only has one observation per demographic information (i.e. a person can only be one age) but have 91 separate rating observations for both preference and energy, I reduced both their preference and energy ratings down to their means. This way when I put all my data for my linear regression model into a dataframe I would have an equal number of observations across my variables.

I then ran a correlation matrix to see if there were any independent variables I should not include in my model.

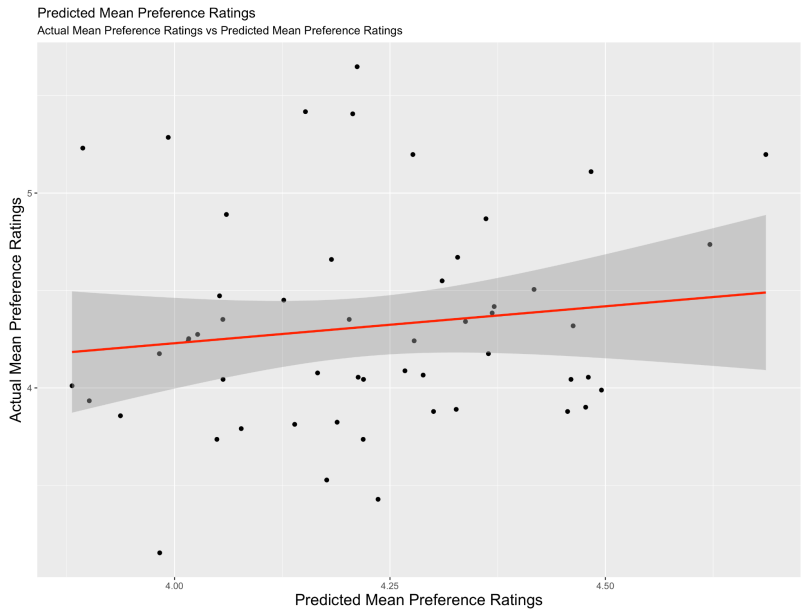
| | Age | Gender | Political_orientation | Art_education | Sophistication |
|---|---|---|---|---|---|
| Age | 1.00000000 | -0.12181943 | 0.0726273048 | -0.02234857 | 0.081834320 |
| Gender | -0.12181943 | 1.00000000 | -0.1296665464 | 0.06240393 | 0.085198538 |
| Political_orientation | 0.07262730 | -0.12966655 | 1.0000000000 | -0.06009449 | 0.091768362 |
| Art_education | -0.02234857 | 0.06240393 | -0.0600944942 | 1.00000000 | 0.151801134 |
| Sophistication | 0.08183432 | 0.08519854 | 0.0917683616 | 0.15180113 | 1.000000000 |
| An_artist_themselves | -0.05591775 | 0.10100818 | -0.1213486903 | 0.34832148 | -0.031967597 |
| energy_means | 0.15949594 | 0.01367762 | 0.0006675179 | -0.08936149 | -0.002060951 |
| preference_means | -0.10176529 | 0.04155229 | -0.0670338163 | -0.07514720 | -0.054187758 |

| | An_artist_themselves | energy_means | preference_means |
|---|---|---|---|
| Age | -0.05591775 | 0.1594959369 | -0.10176529 |
| Gender | 0.10100818 | 0.0136776205 | 0.04155229 |
| Political_orientation | -0.12134869 | 0.0006675179 | -0.06703382 |
| Art_education | 0.34832148 | -0.0893614933 | -0.07514720 |
| Sophistication | -0.03196760 | -0.0020609514 | -0.05418776 |
| An_artist_themselves | 1.00000000 | 0.0496162993 | 0.03858786 |
| energy_means | 0.04961630 | 1.0000000000 | 0.34149775 |
| preference_means | 0.03858786 | 0.3414977511 | 1.00000000 |

Since none of my independent variables had a very high correlation with each other, I kept in these six variables for my linear regression model. I once again split my dataset into a training and testing set to cross-validate, again using p = 0.8.

I called the lm() function to create my model, passing in my dependent variable 'preference_means' first followed by my seven independent variables. My linear regression model had an **$R^2$ of 0.17, meaning that my model was able to explain 17% of the variance in mean preference rating. This model is able to predict preference rating significantly better than my model in question five.** While an $R^2$ of 0.17 is not large by any means, it is much more

The $R^2$ of the predicted mean preference rating and actual mean preference in my testing set was 0.13, and the RMSE was 0.56. These results are about in line with what I would expect based on the results of the training set, so I feel happy with the results of my model. Below is a scatterplot of the actual mean preference ratings vs the predicted mean preference ratings in my test set, where we can see the slight positive correlation.




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?

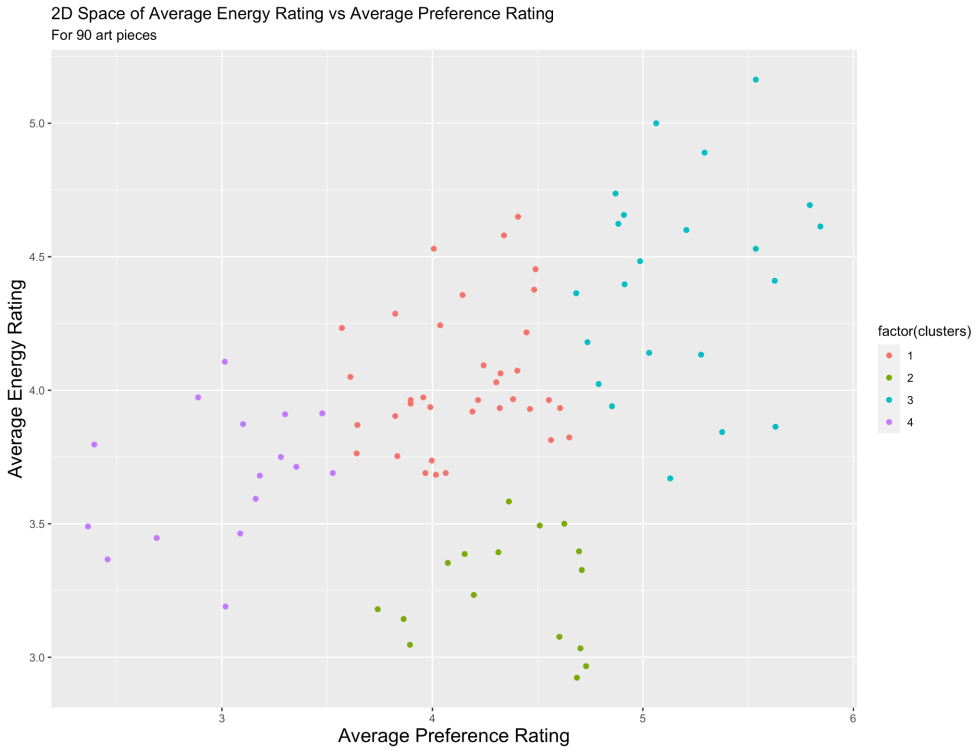
I installed the 'cluster' package to use clustering functions in the process of answering this question. I first reversed the scale of energy ratings. I did this because a 1 for preference ratings corresponds to low preference whereas a 1 for the energy ratings corresponds to feeling agitated. While this doesn't change the process of my clustering, it will likely give the 2D space a slight positive correlation, and make clusters easier to identify for me. I first had to calculate the average preference ratings and average energy ratings, which I did using the colMeans() function. I then created my new data frame with three variables, 'avg_pref' 'avg_energy', both holding the mean preference and energy rating for each piece of art, and 'clusters', which was left empty for now. At this time I plotted the variables to get an idea of how many clusters I might want, and look for outliers that would mess up the clustering algorithm. Artwork 46 titled

“Counter - Composition” by Doesburg was a clear outlier, with a mean preference of 2.99 and mean energy rating of 4.82, so I eliminated it from my data frame.

I opted to use k-means for my clustering algorithm. Using a for loop, I then calculated the average silhouette value for one through ten clusters.

```
[1] "k = 2  average silhouette: 0.397532862890333"
[1] "k = 3  average silhouette: 0.390815952008798"
[1] "k = 4  average silhouette: 0.408110313054914"
[1] "k = 5  average silhouette: 0.378614334381111"
[1] "k = 6  average silhouette: 0.37774027750644"
[1] "k = 7  average silhouette: 0.374185524451668"
[1] "k = 8  average silhouette: 0.394621784618397"
[1] "k = 9  average silhouette: 0.385146474463533"
[1] "k = 10  average silhouette: 0.407996001986863"
```

I opted to use four clusters as it had the largest average silhouette value meaning it clustered the data into the most distinct clusters.

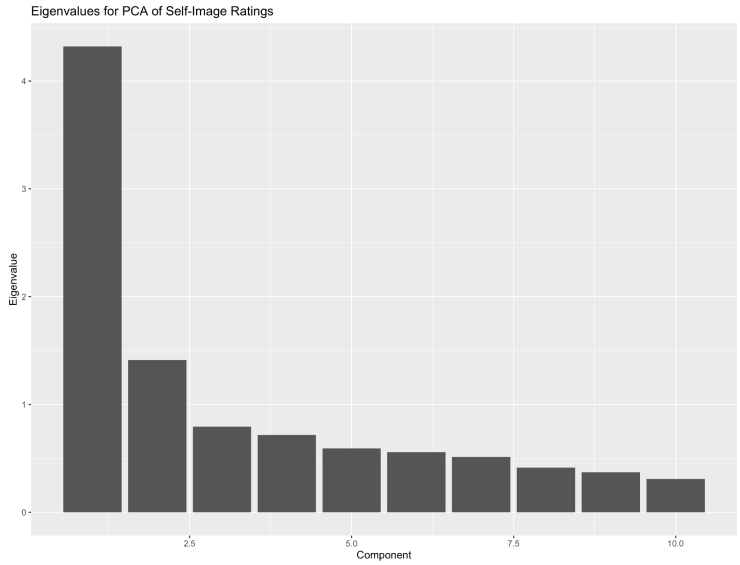


With the exception of one art piece, cluster four all AI or animal generated art. Subjectively speaking, this makes sense as these are the pieces of art with the ugliest content. Cluster three, which has the best ratings, is dominated by classical art. Cluster one is mostly modern art, but there are a substantial number of classical pieces in this cluster as well. Cluster two is a pretty even mix of modern and classical art, but these pieces tended to agitate people the most.

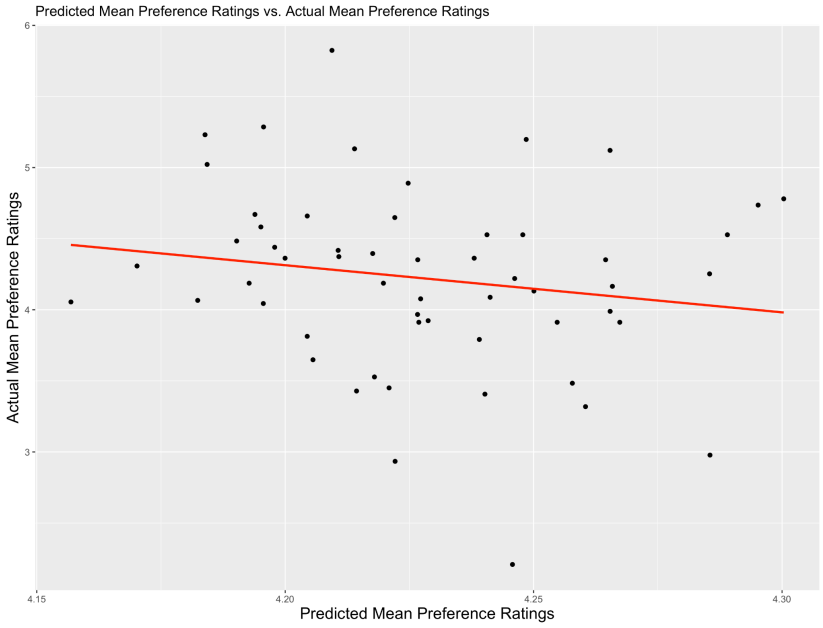
8) Considering only the first principal component of the self-image ratings as inputs to a regression model – how well can you predict art preference ratings from that factor alone?

For this question I had to load in the package ‘psych’ to perform principal component analysis. I first merged preference ratings and self-image ratings into their own data frame. I then cleaned the NaNs and calculated the mean of each row, corresponding to the mean preference rating for each user who filled out all the self-image questions.

I then used the principal() function from the psych package to perform my principal component analysis on self-image ratings. I stored the scores of the first component in its own variable titled 'first_pca_component'. Following this, I extracted and plotted the eigenvalues, even though we are only meant to use the first component.



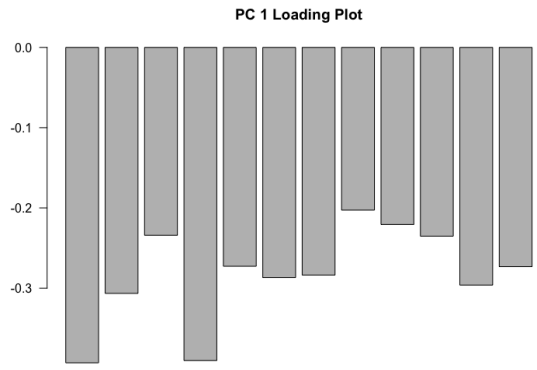

Next I created a new data frame with only the first component and mean preference ratings in order to create my testing and training set, using the createDataPartition() function once again, then passing it into the two sets. I made my model on my training set, with the first principal component as my independent variable and mean preference ratings as my dependent variable, and calculated the predicted mean art preference ratings using the testing set. My model had an **$R^2$ of 0.0022, meaning that the first principal component of self-image ratings was able to explain .22% of the variance in mean preference rating. The RMSE of the predicted mean preference rating values compared to the actual values was 0.64, using the testing set data.** Overall, you can not predict art preference ratings from the first principal component of self-image ratings very well. Below are the predicted and actual mean preference ratings from my test set on a scatter plot.

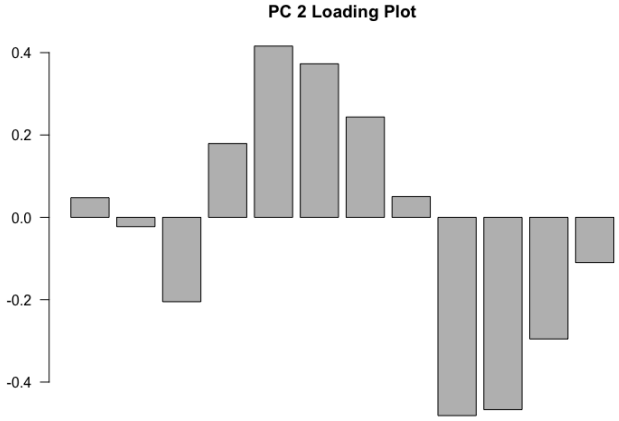


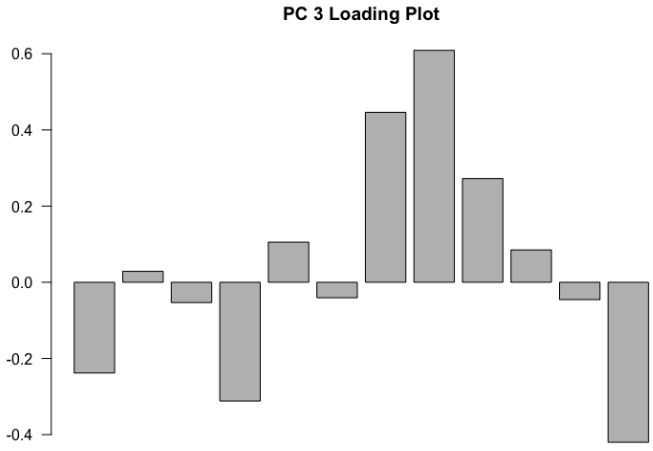

9) Consider the first 3 principal components of the "dark personality" traits – use these as inputs to a regression model to predict art preference ratings. Which of these components significantly predict art preference ratings? Comment on the likely identity of these factors (e.g. narcissism, manipulativeness,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

Using the same process as I used in question eight, while replacing self-image traits for dark personality traits, I got a clean data frame with only the preference ratings and dark image ratings. I once again took the mean of each participant's overall preference ratings.

I then performed my principal component analysis on the dark personality traits.

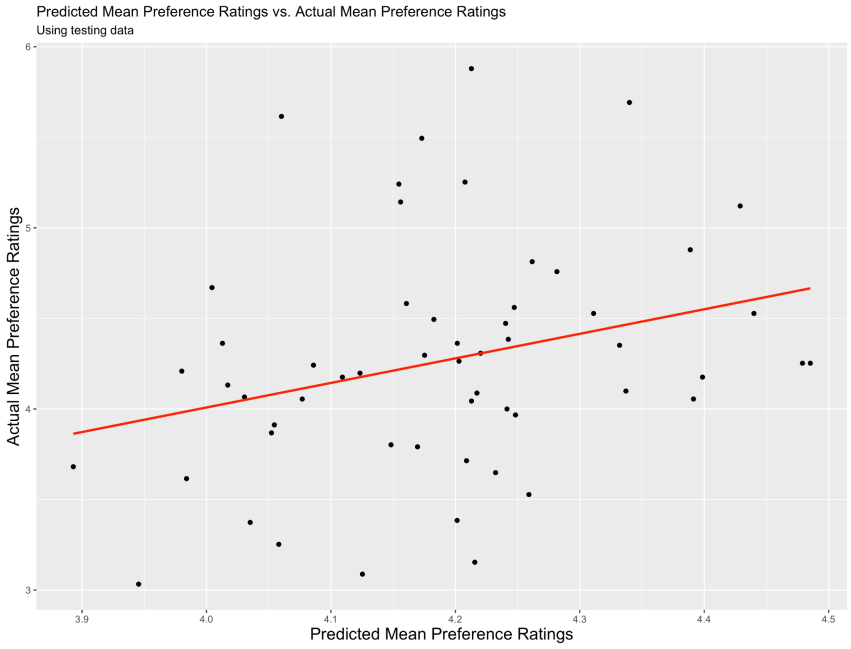

PC 1 Loading Plot
0.0
-0.1
-0.2
-0.3


PC 2 Loading Plot
0.4
0.2
0.0
-0.2
-0.4


PC 3 Loading Plot
0.6
0.4
0.2
0.0
-0.2
-0.4

I then looked at the loading plots for each of the first three principal components. Principal component two has the largest loading values at variables 4-7, peaking at "I tend to lack remorse" and "I tend to be unconcerned with the morality of my actions." The identity of this factor is a lack of empathy. Principal component two has peaking loading values between variables 7-9. These correspond to "I can be callous or insensitive", "I tend to be cynical", and "I tend to want others to admire me." While there is not an easy underlying adjective to chalk these three up I think general cynicism captures them pretty well, especially variables seven and eight.

Next, I binded three variables all containing the scores of the first principal components with the mean art preference ratings. I then split my data to cross-validate and ran my linear regression model on my training data. The linear regression model which used the first three principal components of the dark energy ratings to predict mean preference ratings had an **$R^2$ of 0.044, meaning that these three principal components explain 4.4% of the variance in mean preference rating. When looking at the predictions made using the testing data, the RMSE of the predicted mean preference rating values compared to the actual values was 0.62.**



To look at which of the components significantly predicted mean preference ratings, I called the lm.beta function, which I got from calling the lm.beta package at the beginning of my code. The function returned beta values of -0.14, 0.15, 0.03, for principal components 1-3 respectively. While I wouldn't exactly qualify any of these as "significantly" predicting art preference ratings, components one and two had some predictive power.

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.

To answer this question I first split everyone into the left and nonleft by mutating a variable called 'left' onto a copy of the original data, using a ifelse statement. The variable would contain a 1 if the person was on the left and a 0 otherwise. I then removed all the preference and energy ratings, transformed political orientation into NULL (would spoil the model if left in the data set), and removed NaNs using na.omit() once again. I then scaled the data using I then created my training and testing set, this time using the sample() function because the dataset was too large for createDataPartition. I split my data into a training and testing set, and decided to run a linear regression model. This was largely just to get the coefficients to see which trait variables had the largest coefficients, or in other words predicted whether someone was on the left or not. Unsurprising to see that yogie's and artists are mostly on the left.

| | linear_model.coefficients |
|---|---|
| (Intercept) | 0.593088008 |
| I.like.to.do.yoga | 0.321617977 |
| An_artist.themselves | 0.307618881 |
| I.feel.that.I.m.a.person.of.worth..at.least.on.an.equal.... | 0.301331923 |
| I.like.to.take.walks.on.the.beach | 0.205037963 |
| I.like.to.play.video.games | 0.154609599 |
| I.feel.that.I.have.a.number.of.good.qualities | 0.128353663 |
| I.tend.to.exploit.others.towards.my.own.end | 0.121046440 |
| I.wish.I.could.have.more.respect.for.myself | 0.117374514 |
| Gender | 0.075596064 |
| I.am.able.to.do.things.as.well.as.most.other.people | 0.075040574 |
| I.have.used.flattery.to.get.my.way | 0.070702780 |
| I.like.to.meditate | 0.057211531 |
| I.like.to.take.walks.in.the.forest | 0.054159371 |
| I.tend.to.want.others.to.pay.attention.to.me | 0.048491737 |
| I.tend.to.lack.remorse | 0.034396979 |
| I.tend.to.want.others.to.admire.me | 0.029284485 |
| At.times.I.think.I.am.no.good.at.all | 0.002299194 |
| I.tend.to.manipulate.others.to.get.my.way | -0.005016477 |
| I.have.used.deceit.or.lied.to.get.my.way | -0.007094982 |
| I.like.to.do.paintball | -0.007624238 |
| I.tend.to.be.cynical | -0.011946133 |
| Age | -0.066416714 |
| I.like.to.play.board.games | -0.070220651 |
| I.like.amusement.parks | -0.073895835 |
| I.like.to.ski | -0.077985002 |
| I.tend.to.expect.favors.from.others | -0.084429327 |
| I.can.be.callous.or.insensitive | -0.095227523 |
| On.the.whole..I.am.satisfied.with.myself | -0.097915977 |
| Art.education | -0.107066562 |
| I.feel.I.do.not.have.much.to.be.proud.of | -0.111037228 |
| I.like.to.hike | -0.138001982 |
| I.certainly.feel.useless.at.times | -0.166820414 |
| All.in.all..I.am.inclined.to.feel.that.I.am.a.failure | -0.174099979 |
| Sophistication | -0.175124728 |
| I.take.a.positive.attitude.toward.myself | -0.186702911 |
| I.tend.to.be.unconcerned.with.the.morality.of.my.acti... | -0.191919239 |
| I.like.to.play.role.playing..e.g..D.D..games | -0.292873522 |
| I.tend.to.seek.prestige.and.status | -0.377083106 |

I then made my classification model. I used the glm() function and had to specify that I wanted to make a logistic regression model. I passed through the ‘left’ variable as my dependent variable and all the variables in my training set as the independent variables. I used the predict() function to get the predicted values. I calculated the RMSE of the logistic regression model which gave a value of 1.55. Unfortunately, the package necessary for calculating the AUC (“pROC”), was unable to be downloaded on my computer. I troubleshooted for a while but ultimately had to

decide to take the points off for being unable to report the AUC for my model. I did, however, plot my logistic regression model, which gives me a rough idea that the AUC was likely a relatively large value, so I feel subjectively happy with the quality of my model.