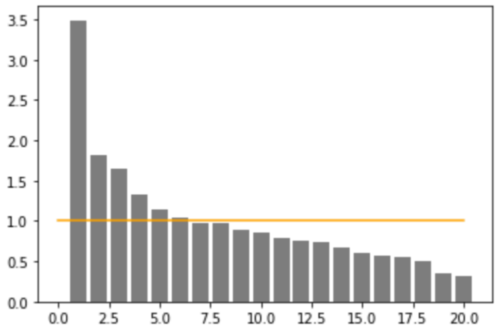
Peter Ze

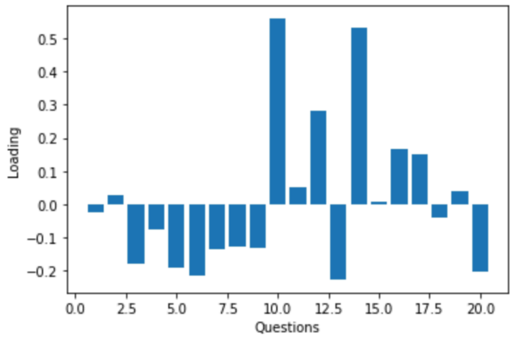
Capstone Project

In terms of data cleaning, when there are multiple movies or columns being compared, a row-wise removal is performed to keep the participants that voted for all columns. In cases where only one movie is examined, a simple removal of nulls is performed. With dimension reduction, a PCA is performed and the kaiser criterion is used to determine the number of factors, also a z-scoring is applied.

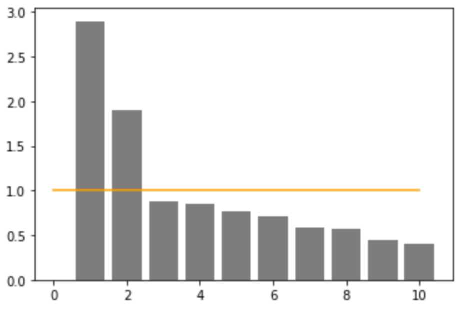
1)

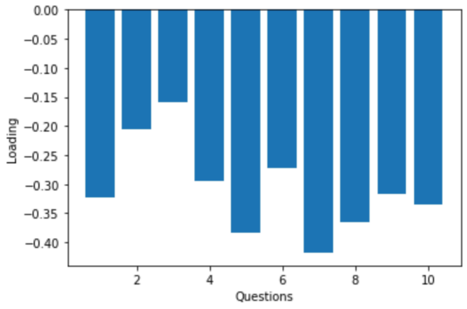
Images below describe the kaiser criterion and the factors that best describe "sensation taking." I used factor 3 or horror/jump-scare sensations to define "sensation seeking." I believe that these are the sensations that have relationship with movie experience



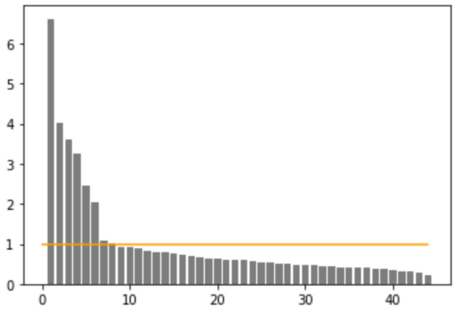


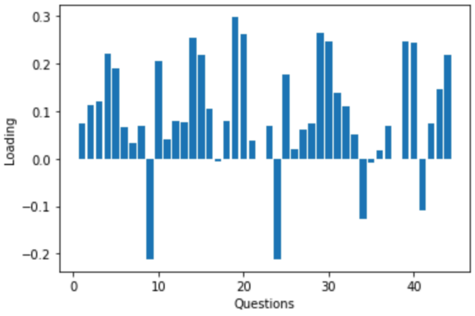
Images below describe the kaiser criterion and factor 1 for overall movie experience.



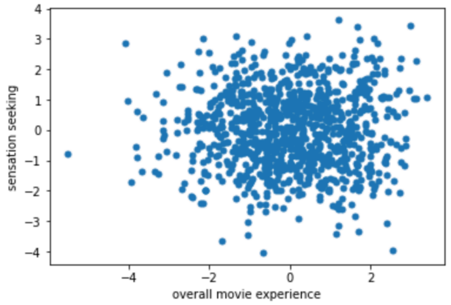


Personality could be a confound that can be reduced by partial correlation, thus showing factor 2, overthinking, characterizing questions like imagination or whether worries a lot.

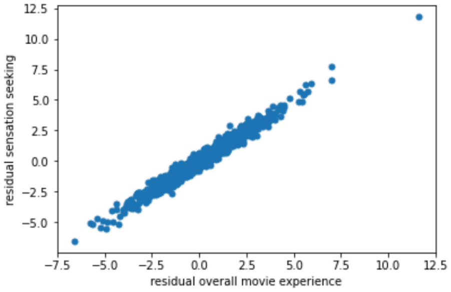




Before and after we partial-out personality, the correlation varied drastically, and it is concluded that there is a correlating relationship between sensation seeking and movie experience



```
correlation:  0.02711174272168862
```

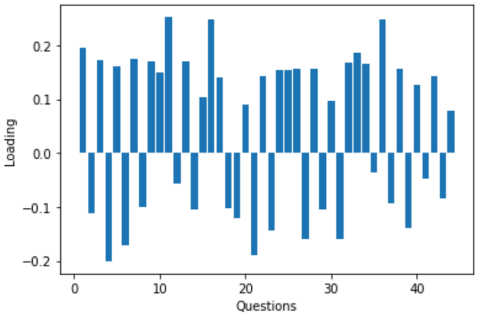


```
Partial correlation: 0.9846915454117291
```

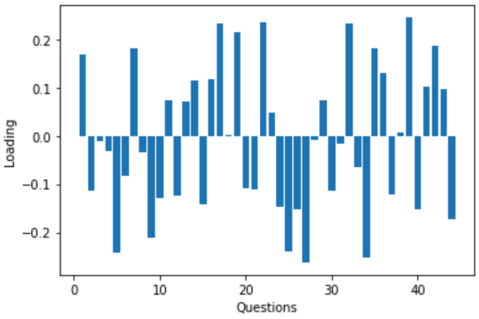
2)

Note that we have to redo the PCA on personality because some rows were removed in last question due to row-wise removal. I find two factors: one's sociability and one's openness to experience, they could best characterize personality. In short, sociability defines whether a person is extrovert/introvert, agreeableness/disagreeableness; opennesss to experience defines whether one is imaginative/conventional, abilities in work.

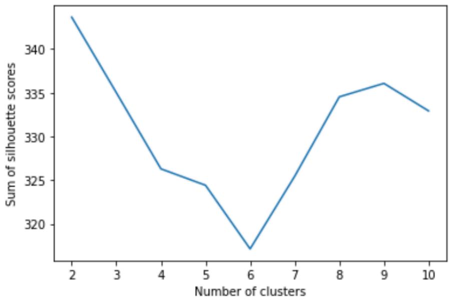
Sociability

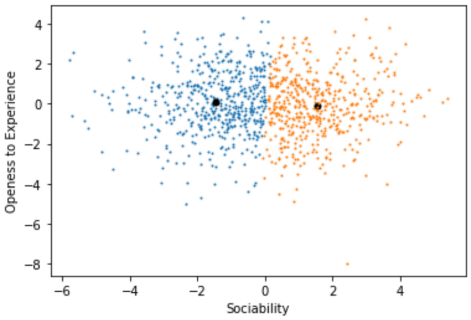


Openness to Experience



A silhouette method is used to determine a cluster of 2 and the cluster is shown below.

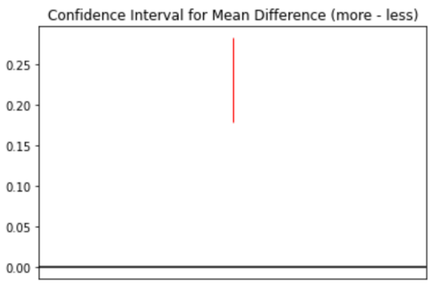


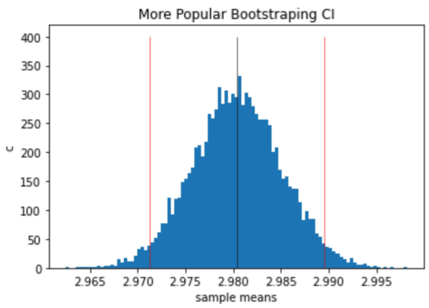


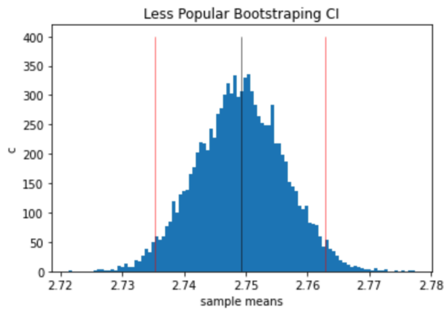
To conclude on the 2 clusters, they differ in terms of sociabilities, therefore there should be 2 groups in the participants in terms of personality type: extroverts vs introverts.

3)
Through a median split for all 400 movies, all movies are classified either as popular (>197.5) or less popular (<=197.5). After cleaning out two data containing rating means of both groups of movies, a t-test, that assumes a normal distribution, is performed. The null hypothesis is that there is no significant difference in ratings between more and less popular movies. Additionally, whether greater or not could be seen from confidence intervals.

- P-value under t-test: 3.798399239327428e-18
- Confidence interval for the difference in means (more - less) under normal distribution
- Resampling confidence interval from bootstrapping when we do not assume normality.


Confidence Interval for Mean Difference (more - less)

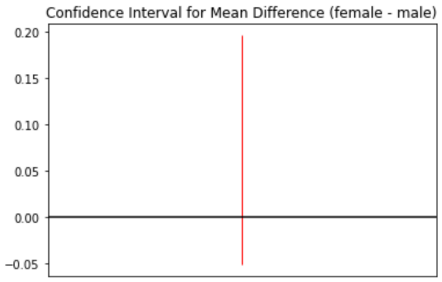

More Popular Bootstraping CI


Less Popular Bootstraping CI

Thus, we could reject the null hypothesis and show that there is a significant difference. Through the confidence intervals, we could see that there is a significant positive difference between mean of higher populated movie means - lower populated, thus they are rated higher.

4)
After indexing the 'Shrek' dataset to male rate sand female rates, a t-test is performed to investigate the null hypothesis: there is no significant difference in means of both groups.

- P-value under t-test: 0.27087511813734183
- Confidence interval for the difference in means (female - male) under normal


Confidence Interval for Mean Difference (female - male)

Since the question asks of 'Shrek' is gendered or not, other non-parametric tests or permutations might be a difference approach for the investigation

- P-value under mann-whitney: 0.050536625925559006
- P-value under permutation test (test-static: female mean - male mean): 0.13083

In conclusion, these tests all show the difference to be not significant, thus the null hypothesis is not rejected and 'Shrek' is not gendered.

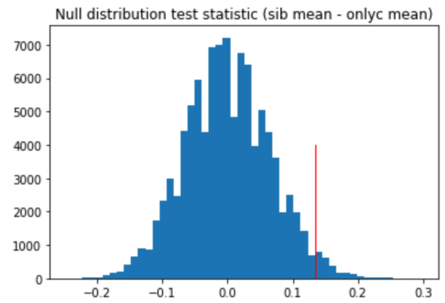
5)

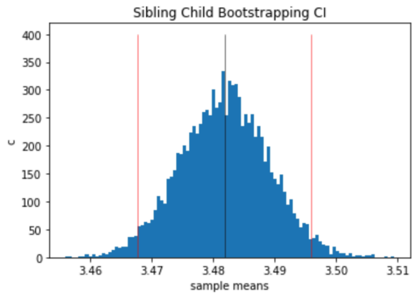
After indexing ‘Lion King’ into only child rates and sibling rates, a t-test is performed to test the hypothesis: there is no significant difference between only child ratings and sibling ratings.

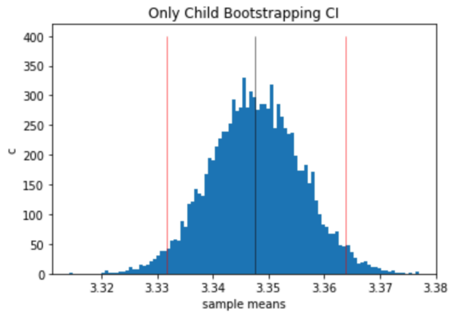
- P-value under t-test: 0.04026705526268264

While according to the t-test there is a significance, we do not know which one is larger, and the previous assumption was under normal distribution only. Thus, a permutation test and a bootstrap confidence interval is investigated

- P-value under permutation test (test-static: sibling mean - only mean): 0.02163
- Confidence interval for both groups using bootstrap


Null distribution test statistic (sib mean - onlyc mean)


Sibling Child Bootstrapping CI

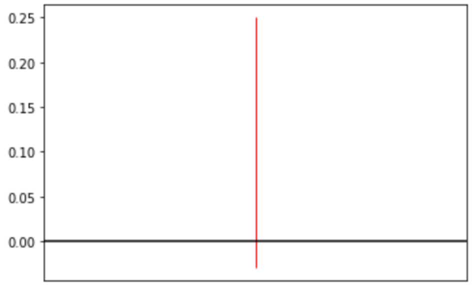

Only Child Bootstrapping CI

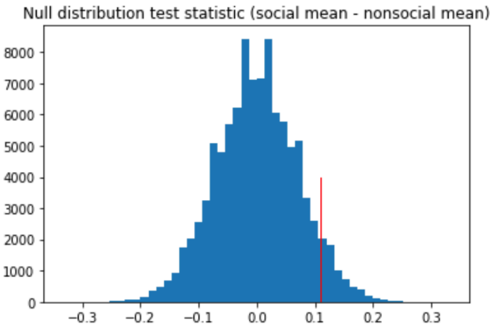
Thus, we can conclude that people who are siblings are more likely to enjoy ‘Lion’ and reject the null hypothesis. But, as the question asks if the only child like it more, the final answer to the question is no.

6）

After indexing ‘Wolf of Wall Street’ into social and non social rates, a t-test is performed to test the hypothesis: there is no significant difference between social and non social ratings.

- P-value under t-test: 0.11738913665664574
- Confidence interval for difference between social means and nonsocial means under normal
- P-value under permutation test (test-statistic: social mean - nonsocial mean): 0.05553



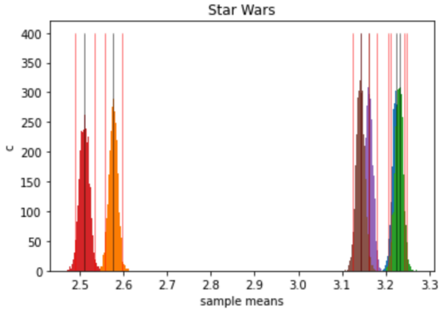

Null distribution test statistic (social mean - nonsocial mean)

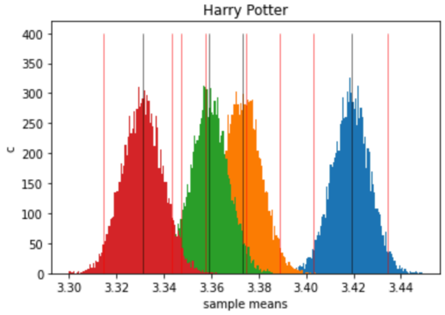
After assessing p-values with t-test under normal distribution and permutation test with null distribution, we could not reject the null hypothesis, as there is no significance. Therefore, the answer to the question is no.

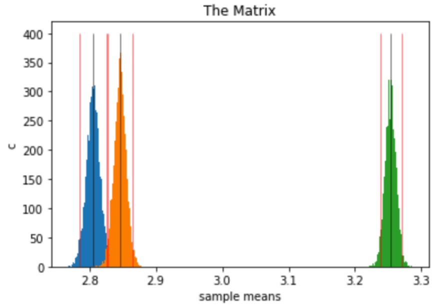
7)

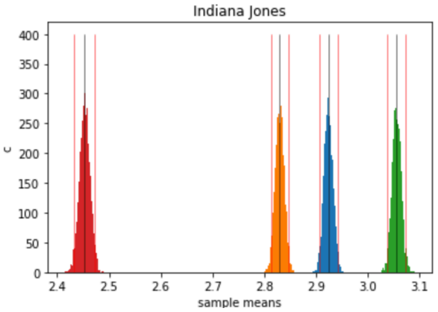
For each franchise, the movies included are cleansed and tested with annova on the null hypothesis of there is no significance between each movie in the franchise.

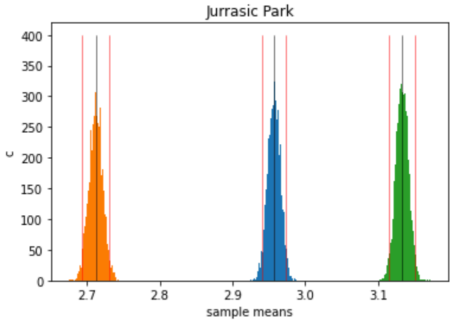
Star Wars p-value under annova: 2.399595163532992e-38, significant
Harry Potter p-value under annova: 0.2275340290918136, not significant
The Matrix p-value under annova: 1.2957225925356723e-08, significant
Indiana Jones p-value under annova: 5.20425425762115e-12, significant
Jurassic Park p-value under annova: 3.542127514286409e-10, significant
Pirates of the Caribbean p-value under annova: 0.03207932803269902, significant
Toy Story p-value under annova: 0.0005193828629536134, significant
Batman p-value under annova: 1.6410731510652519e-18, significant

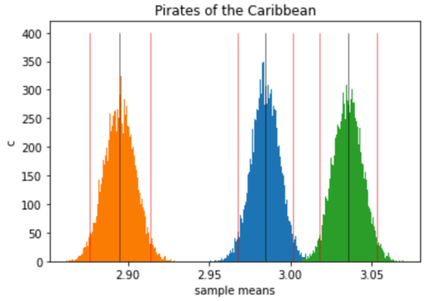


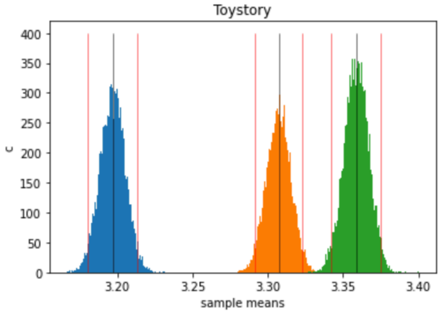


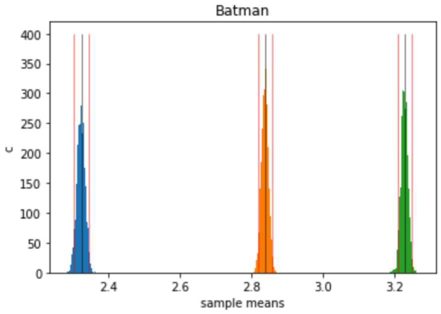








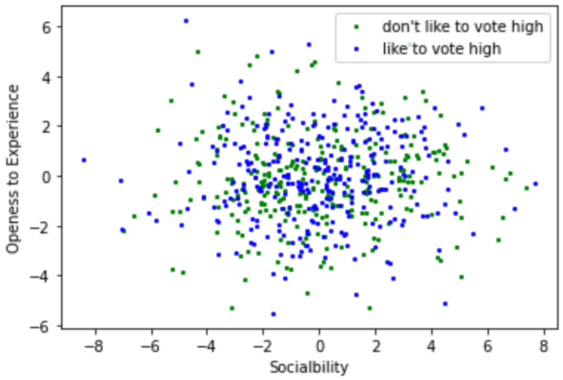


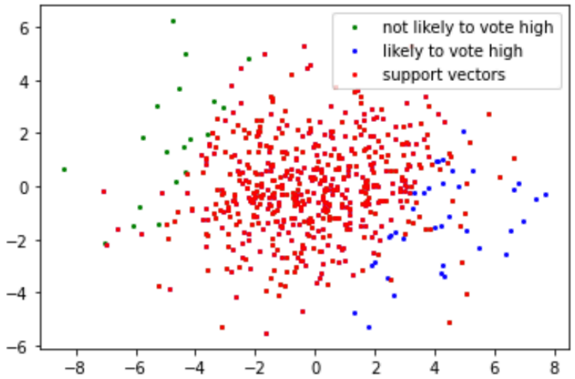


In conclusion, we can see that only the Harry Potter franchise has a consistent rating, with it not having a significant difference in mean ratings between each movie. Shown in the bootstrapping CIs as well, compared to others, Harry Potter is the only with fully overlapping distributions.

8)
The prediction model used a supervised learning approach through svm classification. After a PCA on personality, two factors (the same as our clustering) were picked to characterize personality as a whole and that is our x in the model. The y is a binary variable in which I characterize as a person's likeness to vote high. It is derived from the mean movie-rating per person and a medium split is performed to determine if high or not. Note that cross validation has been used for the model with splitting train/test for a more precise accuracy.

- SVM roc_auc_score: 0.5703533294991374



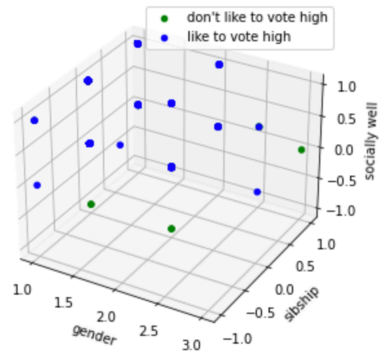


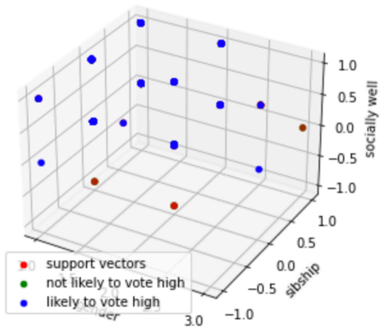
In conclusion, it makes logical sense that people who have high openers to experience and low sociability like to vote lower, since they have better imaginations as well as less likely to express themselves, and vice versa for those that like to vote high.

9)
The prediction model used a supervised learning approach through random forest or svm. With the 3 predictors that are independent of each other, there is no need for a dimension reduction. I used both classification methods because of difficulties in displaying a clear diagram under random forest. Note that cross validation has been used by splitting test/train data with a 20/80 split.

- Random Forest roc_auc_score: 0.5367262847772312
- SVM roc_auc_score: 0.522770227076906
- Results below represent the SVM model

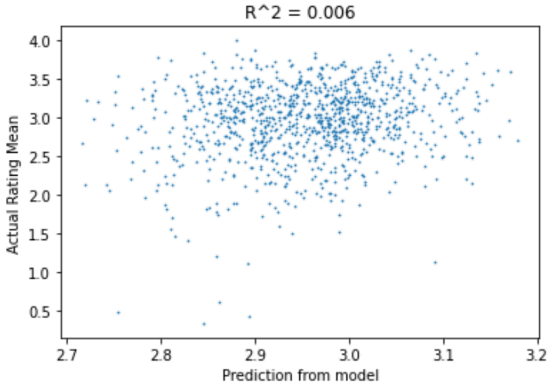
Please note that the diagrams are not as representative due to color overlays upon one point.





10)
The predictor that I choose are the movie-experience columns. After a PCA, like that shown in question 1, the columns were dimensionally reduced to 2 factors, overall film experience and cognitive experience. By using these two factors and the mean of movie ratings for each individual, a multiple linear regression model was fitted to predict a person's average rating mean. Note that from a cross validation, a 50-50 split, the RMSE is calculated.

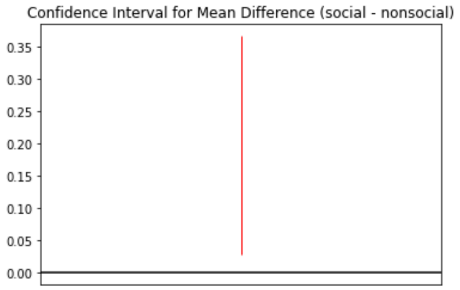
- RMSE: 0.4885966278622591

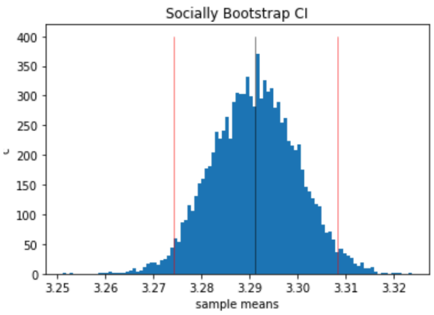


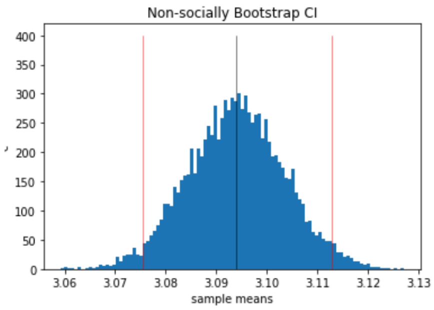
In conclusion, the model is not great, predicting the means of a person's rating linearly is not an ideal approach.

EC)
My extra investigation is on: do people who like to watch movie socially enjoy 'Pulp Fiction' more than those that do not. The null hypothesis to investigate is that there is no significance between two groups.

- P-value under t-test: 0.020507467783240796
- Confidence Interval for the difference of means (social - non social) under normal
- P-value under permutation test (test-statistic: social mean- nonsocial mean):0.00936







Thus, we can reject the null hypothesis, and by comparing the CI under both normal and null distribution, we can conclude that the answer is yes.