Capstone

IDS

Julian Agolini

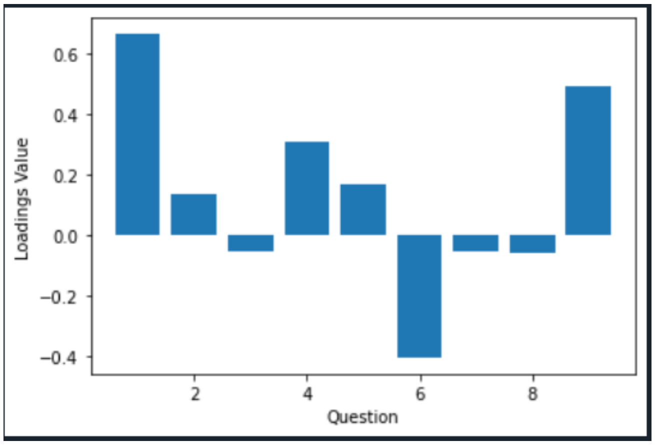
18 May 2022

Movie Analysis Report

In this report I will detail solutions to the nine questions that were brought to me. Before I get into specifics, first I will explain some generalities that apply to the entire report. In terms of data transformation, I took the given data set and split it up into 5 subsets of: movies (Columns 1-400), sensation seeking behaviors (401-420), personality questions(421-464), movie experience ratings(465-474), and personal questions(475-477). When ever two of these data sets needed to interact, I would concatenate the data frames so they could be operated on as one unit. I used row-wise removal in order to clean the data across the report. There is ample data to use with small enough variability that keeping equal sample sizes was ultimately more important than keeping a few extra rows of data. Only two respondents responded to every question, so crucially I only removed missing data in the context of when it concerned a question, in other words any given person is more likely to have answered any two given questions than all questions. Finally, a note on dimensionality reduction. In order to respond accurately to some questions I needed to distill the essence of many variables into a few new variables. This was done by implementing a principle component analysis for the sensation seeking behaviors, personality questions, and movie experience ratings data sets. A Kaiser criterion was applied to all such analyses.

1. What is the relationship between sensation seeking and movie experience?

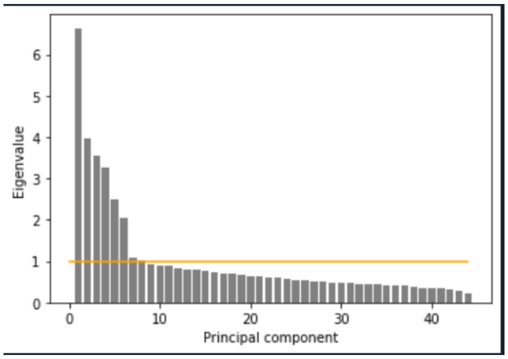
First of all, after completing a pca on movie experience, I found that there are two main contributors to the different respondents. I categorized these as the empty- headed viewer and the emotionally unstable movie viewer. From there, I completed a pca of these components combined with the pca of the sensation seeking category in order to find which sensation seeking categories most aligned with the movie experience categories. To no surprise, the emotionally unstable movie viewer was more likely to score highly in destructive sensation seeking behavior such as being alone as a child and gambling while the empty-headed movie viewer was more likely to enjoy healthier activities such as rock climbing and motorcycle riding. As you can see below question 1, (Are you an emotionally unstable movie viewer?) Aligns highly with question 6 and 9, questions asking about gambling and being alone as a child. On the flip side, not aligning with the more healthy activities.

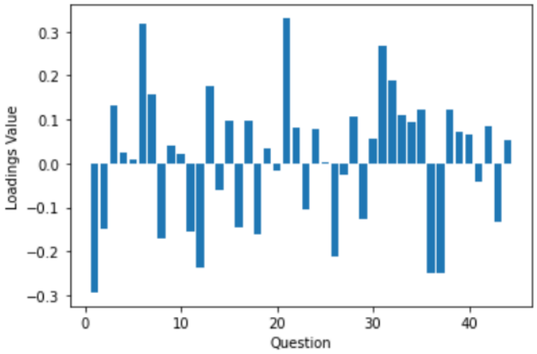


2. Is there any evidence of personality types based on the data of these research participants?

After doing a principle component analysis of the personality questions, it becomes apparent that

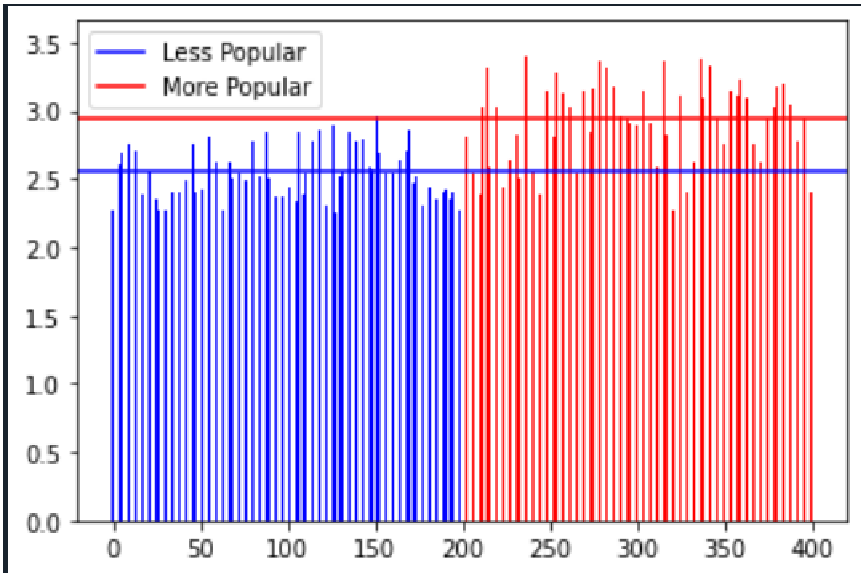
out of the 44 total questions given to the participants, only 8 meaningfully contribute to the differences among them. This gives evidence to suggest that there are 8 personality types that can be substantively differentiated amongst the participants. After finding that there are 8 main contributors to uniqueness in personality, I looked at the loadings value for each question for every 8 main components in order to further demystify the categories. From doing so, I arrived at the conclusion that the 8 personality types among participants are: The worrier, The quiet one, The Persevered, The cold one , The coy one, The grounded, The thinker, The naive. Below are figures which show the principal components along with an example of the loadings value per question for principal component one. As you can see, question 21 most contributed to this principal component so it is chosen to represent that personality type





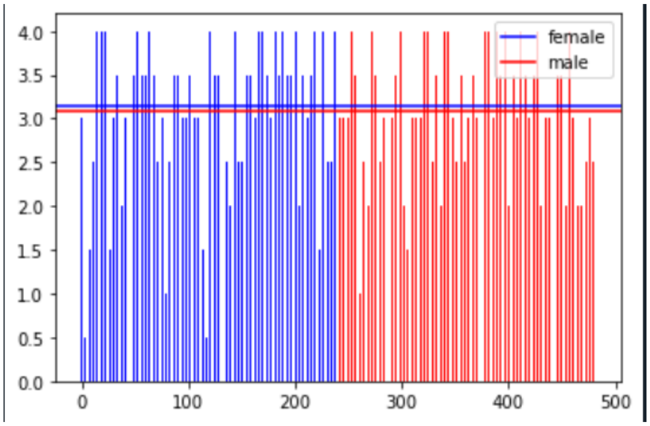
3. Are movies that are more popular rated higher than movies that are less popular?

First, to operationalize movies that are more popular, I selected the group of movies which had been rated more than average. To do this, I looked at the number of non missing ratings of every movie, found that the median number of reviews was 187.5, so for every movie with more than or equal to 188 or more ratings I put together in a group and for every movie with less than or equal to 187 ratings I put together in a group and compared those two groups. Within these groups, I found that the mean for the more popular movies was 2.9 whereas the mean rating for less popular movies was 2.4. As you can see below, this is a significant result. Both groups have their ratings with their means shown.



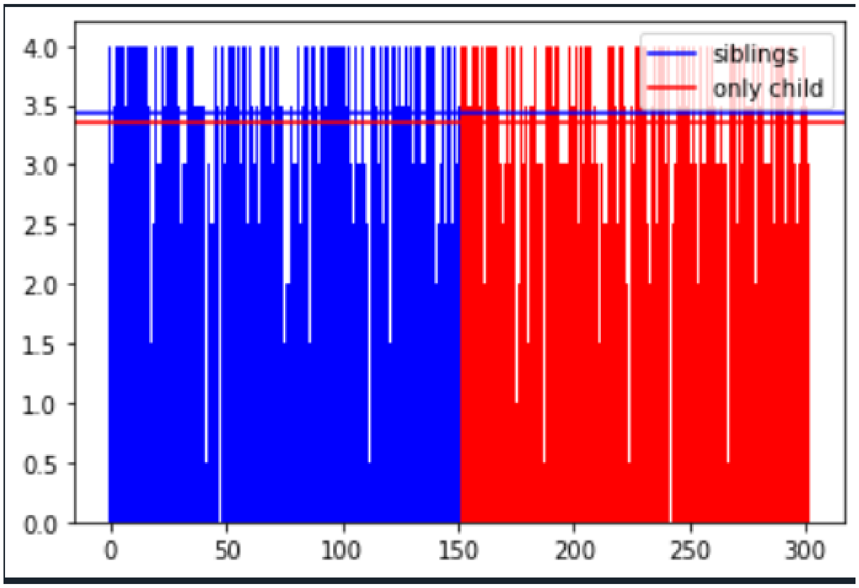
4. Is the enjoyment of 'Shrek (2001)' gendered?

In order to answer this question, I first selected all of the respondents that had watched both Shrek and had responded that they were either male or female in the survey. With these results, I compared the male viewers and the female viewers with an independent t-samples test to see if there was a statistically significant change in ratings by gender. Ultimately, I found that there was no statistically significant difference at the p=.05 level with a test statistic of .65 and a p value of .5. Below you can see the results, participants separated by gender and their ratings, the mean rating for each is included. As you can see, not much difference.



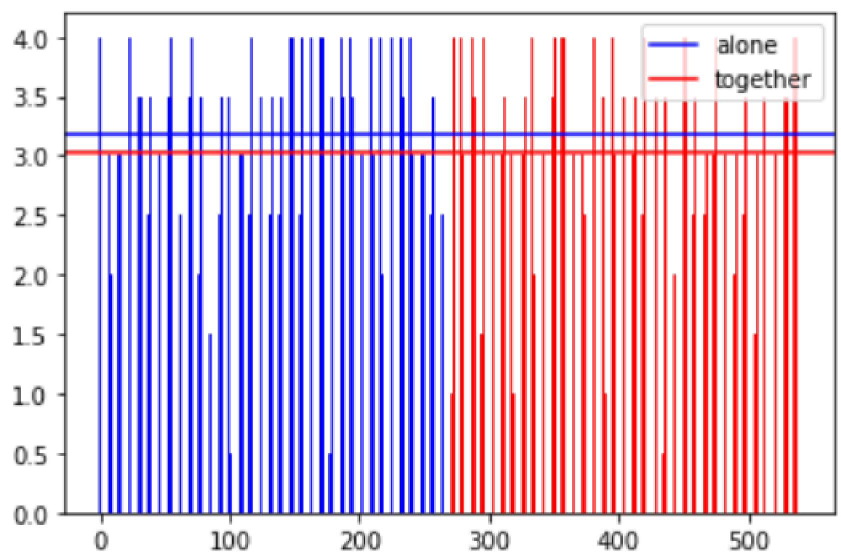
5. Do people who are only children enjoy "The Lion King (1994)" more than people with siblings?

In order to answer this question I used a very similar approach to that of question 4. I selected all of the respondents that had watched both this lion king movie and responded that they were either an only child or had siblings. Next I compared the reviews of only children and siblings with an indpentet t-samles test to see if there was a statistically significant change in ratings by childhood family status. Ultimately, I found that there was no statistically significant difference at the p=.05 level with a test statistic of -1 and a p value of .3.Below you can see the results, participants separated by siblingship status and their ratings, the mean rating for each is included. As you can see, not much difference, but more than that of the difference of genders and shrek above.



6. Do People who like to watch movies socially enjoy “The Wolf of Wall Street (2013)” more than those who prefer to watch them alone?

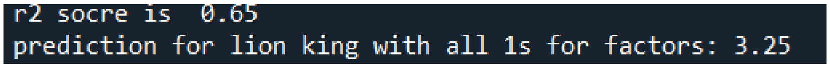
In order to answer this question I used a very similar approach to that of question 4 and 5. I selected all of the respondents that had watched both the wolf of wall street and those that had responded to the watching movies alone or socially question in a binary way and compared those to groups. I used an independent t-test to see if there was a statistically significant difference between the groups for ratings of this film. Ultimately, I found that there were a significant difference at the p=.05 level with a test statistic of 2 and a p value of .04. Below you can see the results, participants separated by with whom they like to watch movies and their ratings, the mean rating for each is included. As you can see, a more significant difference.



Finally, I built three prediction models using multiple linear regression to predict movie ratings based on 1: personality factors only, 2: personal questions only, 3: all available non movie factors. In each of the models I used cross validation methods and characterized the accuracy of my model with an r^2 score. The details for each model can be found below.

Model 1: Personality questions only.

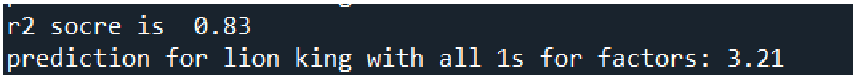
To build this model, I first took the 8 questions that best represented each of the 8 personality types as outlined above. For every movie, the model finds the respondents that have watched the movie and answered these 8 key personality questions. From there, I implemented multiple linear regression to make a prediction of the movie rating based on available personality data. I cross validated the model to avoid overfitting , using a train test split of 20 80 resulting of an r^2 score of .65 for lion king to characterize the accuracy of my model. The results for the example movie Lion King can be seen below although the model can produce a result for any given movie.



```
r2 socre is  0.65
prediction for lion king with all 1s for factors: 3.25
```

Model 2: Personal questions only.

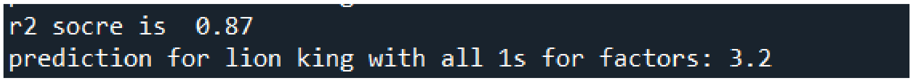
To build this model, I used a very similar approach as outlined above. I first found all the respondents that have watched any given movie and have responded to all three personal questions in a meaningful way. From there, I implemented multiple linear regression to make a prediction of the movie rating based on available personal question data. I cross validated the model to avoid overfitting , using a train test split of 20 80 resulting in a r^2 score of .84 for lion king to characterize the accuracy of my model. The results for an example movie of Lion King can be seen below although the model can produce a result for any given movie.



```
r2 socre is  0.83
prediction for lion king with all 1s for factors: 3.21
```

Model 3: All available factors.

To build this model, if I simply took all the pca components of all the non movie factors, I would be left with 24 independent variables for my multiple linear regression. Of course, this is far too many, and we know that some of this data is highly correlated as outlined in question 1. To accomplish this task, I took a pca of all of these factors and arrived at 10 principal components to use in my model. From there, I took all the respondents that have watched any given movie and have responded to all 10 key survey questions . From there, I implemented multiple linear regression to make a prediction of the movie rating based on the available data. I cross validated the model to avoid overfitting, using a train test split of 20 80 resulting in a r^2 score of .87 for lion king. The results for an example movie lion king can be seen below.



```
r2 socre is  0.87
prediction for lion king with all 1s for factors: 3.2
```