Leo Ding
Professor Wallisch
Introduction to Data Science

Capstone Project

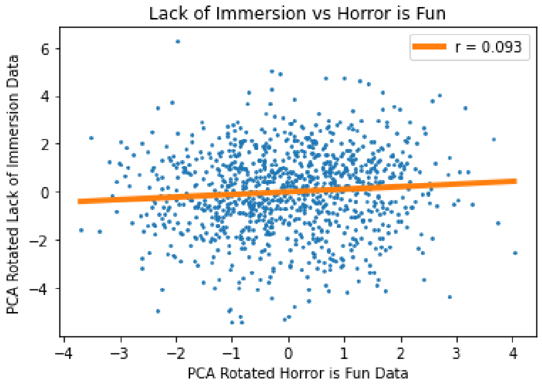
Introduction:

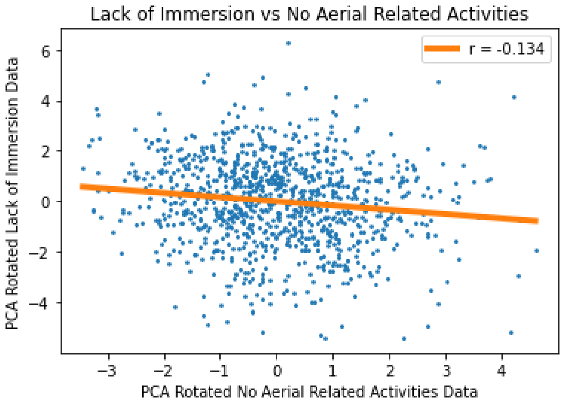
Dimension reduction, through Principal Component Analysis, was used over the sensation seeking, personality, and movie experience sections of data individually. For analysis of data, row-wise removal (in regards to participants) was used in order to remove missing data points.

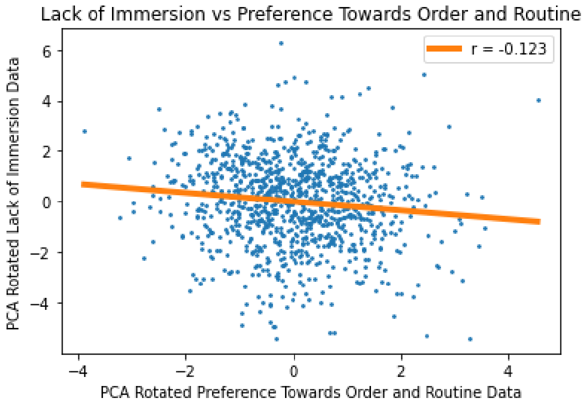
1: What is the relationship between sensation seeking and movie experience?

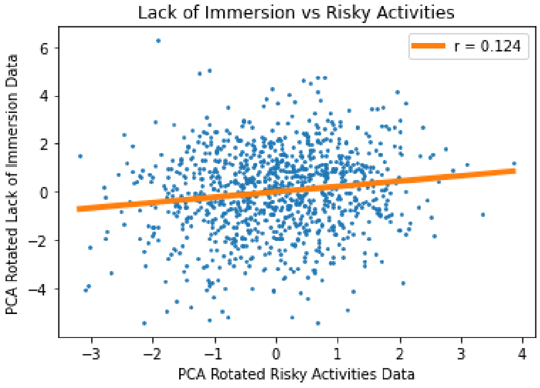
First I took a PCA of the sensation seeking and movie experience columns respectively to reduce the number of dimensions that I would be working with. Using the Kaiser criterion for both PCAs, I found that sensation seeking has 6 principal components while movie experience only has 2 principal components. After finding all of the principal components, I found the spearman correlation coefficient between all possible pairs of sensation seeking and movie experience principal components (12 in total).

What I found was that the second principal component of movie experience, which was determined to be “Good Storytelling from Movie” had no significant correlation with any of the principal components from sensation seeking. The first principal component “Lack of Immersion”, though, had a statistically significant correlation with 4 out of 6 sensation seeking principal components, although none of the correlations were particularly strong.





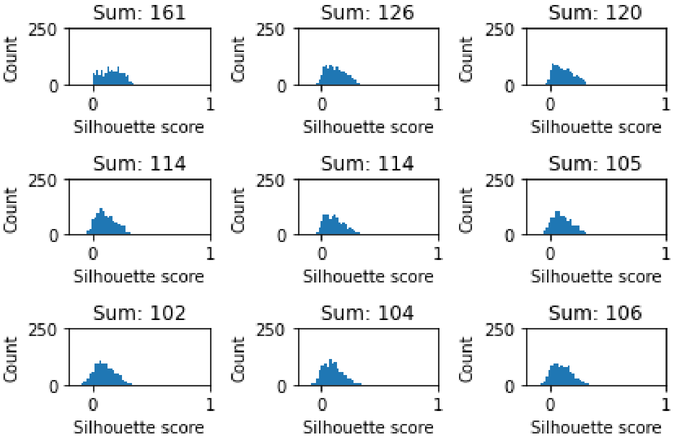




2: Is there evidence of personality types based on the data of these research participants? If so, characterize these types both quantitatively and narratively.

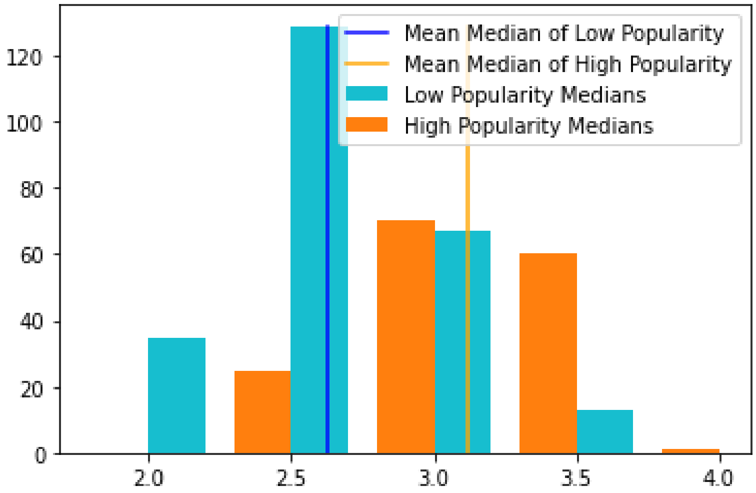
For this I ran a PCA on the personality data to reduce the amount of dimensions I would be working with. By the Kaiser criterion, the personality data has 8 principal components. With the rotated data for the 8 principal components, I used the silhouette method, iterating from 2 to 10 clusters, to determine the optimal amount of k clusters that should be applied onto the data. The highest sum from the simulations came out to be for 2 clusters. Thus, I ran a KMeans clustering algorithm with two 2 clusters to find their centroid positions. The centroid positions were mainly just the opposite sides of the first principal component spectrum. One being around 1.86 and the other being around -2.27. The first principal component is based on extroversion.

So overall, it seems to suggest that there are two personalities: extroverts and introverts.



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

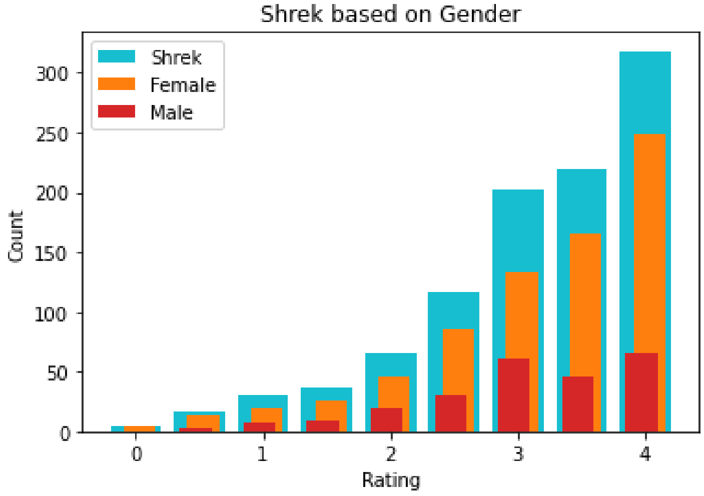
Due to the fact that movie ratings aren't suited for being reduced down to their means, I first calculated every movie's median rating. I then proceeded to count the number of ratings per movie to determine how "popular" they were. I then split the medians based on whether they were more popular or less popular than average. After that, I ran a Kolmogorov-Smirnov test to see whether the underlying distributions were the same for popular and unpopular movies. The p-value for the test was 1.27e-23, which is far below 0.05 thus suggesting that popular and unpopular movies are rated differently. Looking at the histogram of the medians, we can see that the more popular movies are rated, on average, higher than less popular movies.



4: Is enjoyment of “Shrek (2001)” gendered?

I started by clearing away all of the missing data values across both the movie and the gendered column, making sure to also remove the rows in which gender had a response of 3 or self-described. After that, I split the data to create two datasets, one that contained all of the ratings submitted by females and the other containing all of the ratings submitted by males.

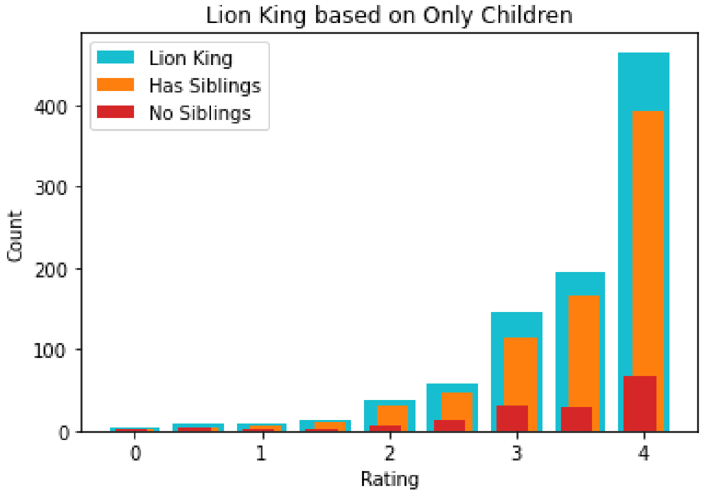
I then ran a Kolmogorov-Smirnov test to ensure that their underlying distributions were the same. Even though they both came from the “Shrek” data, there could’ve been a difference between the two distributions that make up “Shrek”. After the test returned a p-value of 0.056, which is statistically insignificant meaning that the distribution were the same, I ran a Mann-Whitney U Test to see if there was a difference between the two sets. This test outputted a p-value of 0.051 which is also not statistically significant suggesting that the enjoyment of “Shrek” is not gendered.



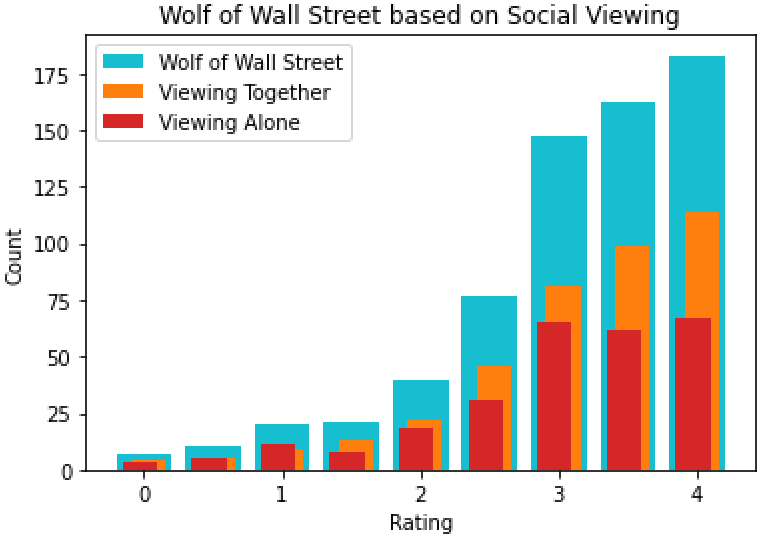
5: Do people who are only children enjoy “The Lion King (1994) more than people with siblings?

I cleaned the data first making sure to also remove rows in which there was no response or a ‘-1’ for the siblings question. I then split “The Lion King” data into two sets, one containing ratings from only children, and the other containing ratings from children with siblings.

I proceeded to run a Kolmogorov-Smirnov test to check that only children and children with siblings had the same distribution when it comes to rating “The Lion King”. The outputted p-value of 0.154 suggests that they do have the same underlying distribution. Because of this, I proceed with a Mann-Whitney U Test to test for a significant difference between the sets. The p-value that was outputted from the Mann-Whitney U Test was 0.043, suggesting that there is actually a difference, based on a 0.05 alpha level, between the ratings given to “The Lion King” by only children and children with siblings.



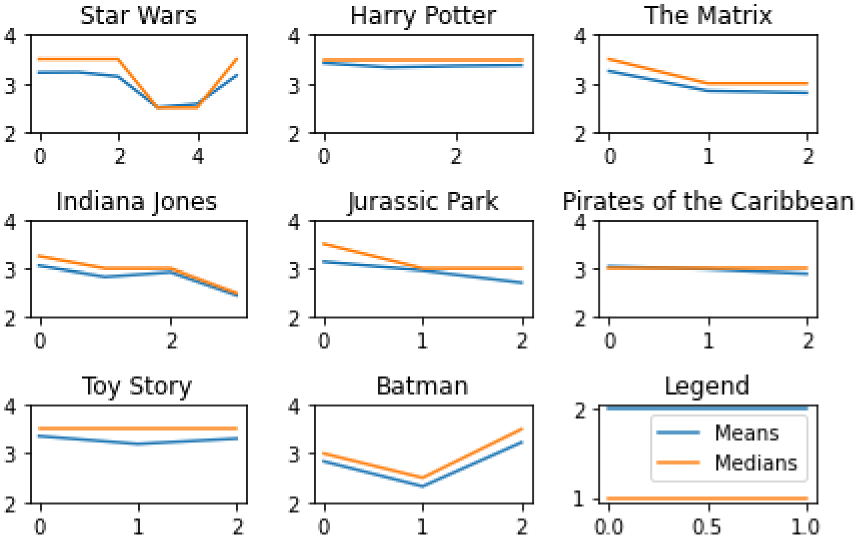
6: Do people who like to watch movies socially enjoy "The Wolf of Wall Street (2013) more than those who prefer to watch them alone?



I first cleaned the data making sure to remove rows where there was a '-1' or no response for the social viewing preference question. I then spilt the ratings data for "The Wolf of Wall Street" based on social viewers and single viewers.

I ran a Kolmogorov-Smirnov test first to determine the same underlying distribution for each of the split data sets. The p-value output of the test was 0.498, which suggests that the distributions are the same. Because of this, I proceeded to run a Mann-Whitney U Test to test the difference between the sets, and this test had a p-value output of 0.113, suggesting that the two groups do not rate "The Wolf of Wall Street" differently.

7:There are ratings on movies from several franchises. How many of them are of inconsistent quality, as experienced by viewers?



For each of the movie franchises, their individual movies were compiled together into one set. Then all missing data was removed row-wise.

For each franchise, I then ran a Kruskal-Wallis test to test for differences between each movie within the franchise. The p-value outcomes of the Kruskal-Wallis test showed that the Harry Potter franchise was the only with a p-value (0.118) greater than 0.05 suggesting that that is the only franchise that did not have significant differences between movies. I did, though, then plot the means and medians of each franchise in chronological order to see the change in ratings over time. I agree with the Kruskal-Wallis test's result for the Harry Potter movies, but I would argue that the Pirates of the Caribbean movies are also fairly consistent. To be fair though, their p-value was the closest to crossing over the barrier to have statistically insignificant difference with 0.0358.

8: Build a prediction model to predict movie ratings from personality factors only.

For each of the 400 movies, I ran a multiple regression analysis against the personality factors. Each iteration consisted of row-wise removal between the movie and all of the personality factors, then using the personality PCA that was created in question 2 to rotate the personality data, then I split the data into 80-20 train test splits, to avoid overfitting, using a random state of 42, and then running a multiple regression with personality as the predictor and the movie rating as the prediction.

The best regression model was for the movie X Men 2 (2003) as it had an RMSE value of 0.718. On the other hand the worst regression model was for The Doom Generation (1995) with an RMSE value of 17132976194153.709. Overall though, running movie regressions against personality was fairly successful with 75 different movies having an RMSE value below 1, although the spread was quite high as can be seen by the RMSE value of the worst model. Thus, personality itself may not be the strongest predictor of movie ratings.

9: Build a prediction model to predict movie ratings from gender identity, sibship status and social viewing preferences.

I ran a multiple regression analysis for each of the 400 movies against gender identity, sibship status, and social viewing preferences. Each iteration started with row-wise removal between the movie and the other factors, then I split the data into 80-20 train test splits, to avoid overfitting, using a random state of 42, and then ran multiple multiple regressions with the non-movies as predictors and the movie rating as the predictions.

The best regression model was Harry Potter and the Sorcerer's Stone (2001) with an RMSE value of 0.705, while the worst regression model was for Ran (1985) with an RMSE value of 1.560. Further, a total of 147 models had an RMSE below 1. Thus, gender identity, sibship status, and social viewing preferences seem to be a decently good predictor of movie ratings.

10: Build a prediction model to predict movie ratings from all available factors that are not movie ratings.

I ran 400 multiple regression analyses, one for each of the 400 movies, against all of the non-movie factors in the data set. Each iteration started with row-wise removal between movies and all other factors, then I used the three PCAs created in questions 1 and 2 to rotate the sensation seeking, movie experience, and personality data, then I split the data into 80-20 train test splits, to avoid overfitting, using a random stat of 42, and lastly I ran the actual multiple regression with all non-movie factors as the predictor and the movie rating as the prediction.

The best regression model was for The Lion King (1994) with an RMSE of 0.676 while the worst regression model was for Equilibrium (2002) with an RMSE of 101.876. In total, there were 60 models with an RMSE under 1. Thus, using all of the non-movie factors to predict the movie ratings is okay, but it’s not great by any means.