Grace Lim

Professor Pascal Wallisch

DS-UA 112

Capstone Project

Dimension reduction: Unless otherwise stated, I did a row-wise removal of the required subset of the data, then z-scored the data to normalize the data, then ran a PCA to reduce the dimension of the data. In general, based on the Kaiser Criterion, I found 6 PC's for Sensation Seeking (SS), 2 for Movie Experience (ME), and 8 for Personality (P).

Data Cleaning: I did element-wise removal of NaNs when the data subset is only a column (a movie), and row wise removal of NaNs when the data subset encompasses multiple columns.

Data Transformation: I extracted the necessary data (in columns) for each question

Additionally, I use the conventional choice of alpha at 0.05.

Below is a table to explain the PC's (to be used for Questions 1 and 2).

Name	Questions that mattered (in descending order)	PC interpreted
ME1	-7,-5,-8	overall negative experience
ME2	-2,-3,-6	easiness to follow and remember movie
SS1	-20, -13, -12, -11, -3	prefers safe and secure activities
SS2	-20, -13, -3	avoids height related risky activities
SS3	10, 14	enjoys scary movies and experiences
SS4	16, 17	prefers ordered and predictable life (because of stressful reality)
SS5	19, -15, 1, -8	individualistic, risky habits (probably due to upbringing)
SS6	-18	havent ridden motorcycle
P1	11, 36, 16	extroverted
P2	19, 20, 29, 30, 40, 14, 39	full of emotions
P3	21, 6	reserved

P4	-18, 28, 3	good work ethic
P5	-27, -34, 39, -5, -25, 22, 17, 32	hyperactive and warm
P6	41, 23, 27, 8	passive/few interests in life/ demotivated in life
P7	41, -30	not interested in art
P8	15, 35, -43	focused/passionate in their work

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

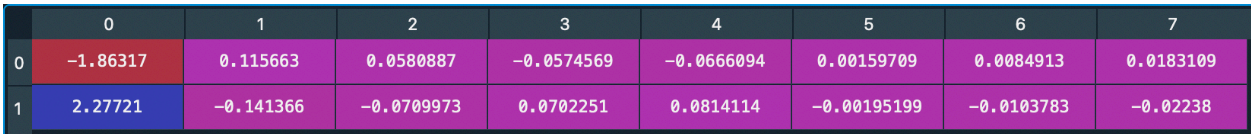
As I did not assume a linear relationship between any of the factors, I did a correlation to characterize the relationship and found Spearman's r for each of the SS PC's with each of the ME PC's, finding 12 r's. Below is a table to show the results:

PC's correlated	Spearman's r	p-value
ME1 and SS1	0.027	0.38729
ME1 and SS2	-0.134	0.38729
ME1 and SS3	0.093	0.00273
ME1 and SS4	-0.123	0.00008
ME1 and SS5	0.124	0.00007
ME1 and SS6	-0.021	0.50510
ME2 and SS1	-0.039	0.21358
ME2 and SS2	0.015	0.62116
ME2 and SS3	0.038	0.22265
ME2 and SS4	0.007	0.81797
ME2 and SS5	-0.051	0.10229
ME2 and SS6	-0.041	0.18653

Using a p-value cut off of 0.05, I found that SS and ME are related in 3 ways (as highlighted), although the relationship is not strong (as shown by small r): Users who had an overall negative experience (ME1) enjoy scary movies and experiences (SS3), do not prefer an ordered and predictable life (SS4), but do show individualistic and risky habits (SS5).

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

As I had found 8 PC's for P, I first ran a silhouette on the first 8 columns of the data in terms of the PC score to compute K, and found that K is 2 with a sum of 161. I then ran a K-mean cluster and found these coordinates:

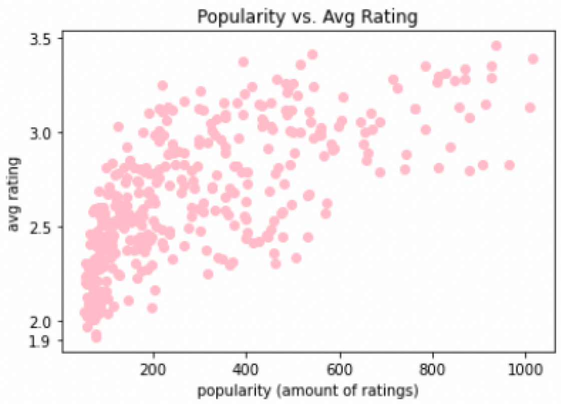


	0	1	2	3	4	5	6	7
0	-1.86317	0.115663	0.0580887	-0.0574569	-0.0666094	0.00159709	0.0084913	0.0183109
1	2.27721	-0.141366	-0.0709973	0.0702251	0.0814114	-0.00195199	-0.0103783	-0.02238

Based on how I interpreted the P PC's, the first personality type is not extroverted, is full of emotions, is reserved, does not have a good work ethic, is not hyperactive and not warm, has few interests in life, is not interest in art, but focused in their work. The second personality type is the complete opposite. Although I could not plot a graph in 8D, I concluded that the 2 clusters can be best described as Introverted and Extroverted.

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

I characterized a more popular movie as one that has a higher amount of ratings. Although we learned that it is not appropriate to reduce a dataset to its mean, I decided to plot a scatter plot of average rating against popularity as it is the easiest way to visualize and interpret the data. I found a Spearman's rank correlation of 0.761 with a p-value of 9.602e-77, meaning that there is a strong positive correlation between popularity and average ratings, and it could not be due to chance alone.

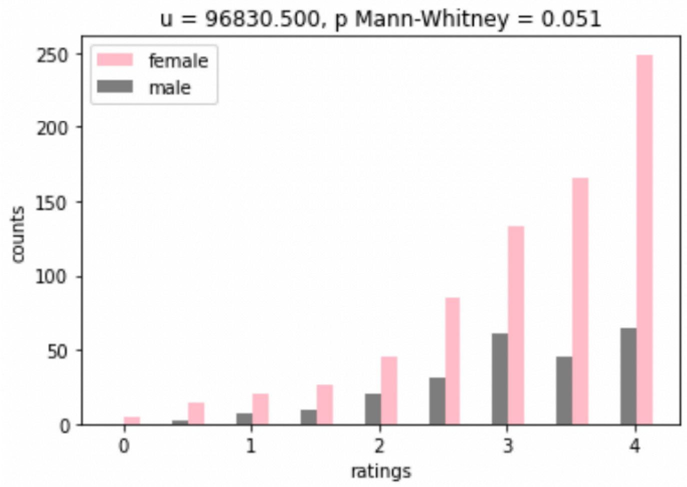


I also checked the maxima and minima for amount of ratings (popularity) and average ratings:

	Movie Name	Amount of Ratings	Average Rating
Maximum Rating	The Lion King (1994)	937	3.460
Minimum Rating	3000 Miles to Graceland (2001)	77	1.916
Most Popular	Finding Nemo (2003)	1014	3.388
Least Popular	Best Laid Plans (1999)	54	2.046

As shown, with higher amount of ratings, the average increases, and vice versa.

4) Is enjoyment of ‘Shrek (2001)’ gendered, i.e. do male and female viewers rate it differently?

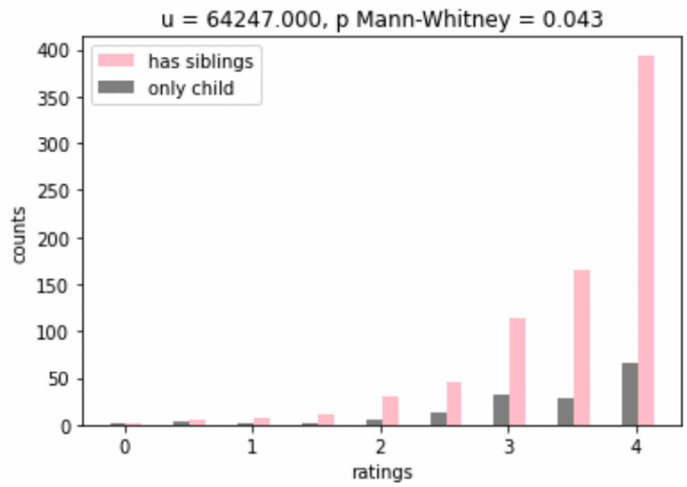


I did questions 4, 5, 6 using the same function.

I first separated the ratings into 2: for male and female. To the left is a grouped histogram to visualize the data:

As there is an unequal amount of male and female users, and that the units of the ratings are unequal (as they are ratings), I compared the median of the two groups using the Mann-Whitney U test. With a p-value of 0.051, I fail to reject the null hypothesis that the difference in ratings is due to chance alone. As the results are not statistically significant, I am reasonably sure that the difference in ratings can be explained by chance alone.

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



I did questions 4, 5, 6 using the same function.

I first separated the ratings into 2: users who are the only child and users who had siblings. To the left is a grouped histogram to visualize the data:

As there is an unequal amount of only child and has siblings users, and that the units of the ratings are unequal (as they are ratings), I compared the median of the two groups using the Mann-Whitney U test. With a p-value of 0.043, I reject the null hypothesis and observe that the difference in ratings could not have been due to chance alone.

I then found the descriptive stats:

	Has Siblings	Only Child
Amount of users	776	151

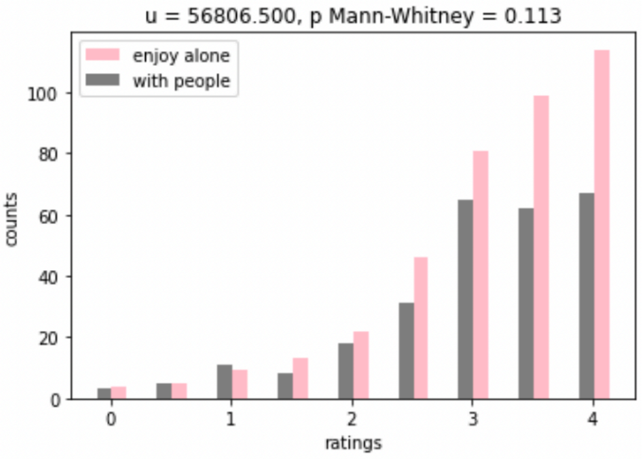
Average Rating	3.482	3.348
Median Rating	4.0	3.5
STD	0.718	0.814
SEM	0.026	0.066

As the average and median rating for the only child group is lower, and that the amount of users is also significantly lower (implying that there is less interest/popularity), I conclude that people who are only children do not enjoy The Lion King (1994) more than people 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 did questions 4, 5, 6 using the same function.

I first separated the ratings into 2: for users who enjoy them alone and those who watch socially. To the left is a grouped histogram to visualize the data:



As there is an unequal amount of lone watchers and social watchers, and that the units of the ratings are unequal (as they are ratings), I compared the median of the two groups using the Mann-Whitney U test. With a p-value of 0.113, I fail to reject the null hypothesis that the difference in ratings is due to chance alone. As the results are not statistically significant, I am reasonably sure that the difference in ratings can be explained by chance alone.

7) There are ratings on movies from several franchises (['Star Wars', 'Harry Potter', 'The Matrix', 'Indiana Jones', 'Jurassic Park', 'Pirates of the Caribbean', 'Toy Story', 'Batman']) in this dataset. How many of these are of inconsistent quality, as experienced by viewers?

As the ANOVA allows us to compare more than two groups without risking an exploding amount of false positives, and that the one-way ANOVA is considered a robust test against the normality assumption, I chose to do an ANOVA. I characterized inconsistent quality of movies in a franchise as one that has a statistically significant difference in average ratings between the movies. The table below is a summary of my findings:

Franchise Name	Amount of Movies in Data Set	Amount of Users who watched all the movies	p-value
Star Wars	6	333	0.0
Harry Potter	4	710	0.22753
The Matrix	3	260	0.0
Indiana Jones	4	244	0.0
Jurassic Park	3	398	0.0
Pirates of the Caribbean	3	561	0.03208
Toy Story	3	757	0.00052
Batman	3	219	0.0

As shown in the table, all of the franchises had p-values that were below the cutoff of 0.05, meaning that the result of differing average ratings could not have been due to chance alone (i.e. they come from different groups, in this case movies). Therefore, all of the franchises except for Harry Potter had inconsistent quality.

8) Build a prediction model of your choice (regression or supervised learning) to predict movie ratings (for all 400 movies) from personality factors only. Make sure to use cross-validation methods to avoid overfitting and characterize the accuracy of your model.

I did questions 8, 9, 10 using the same function.

For each movie in the dataset, I did a PCA on the Personality factors as different movies would attract a different audience. I then created a multiple regression model based on how many PC's are found for that movie, and cross-validated it by using an 80-20 train-test-split. The R^2 of the model before cross validation ranges from 0.01632 to 0.38462, while the RMSE of the model after cross validation ranges from 0.58186 to 1.45374.

9) Build a prediction model of your choice (regression or supervised learning) to predict movie ratings (for all 400 movies) from gender identity, sibship status and social viewing preferences (columns 475-477) only. Make sure to use cross-validation methods to avoid overfitting and characterize the accuracy of your model.

I did questions 8, 9, 10 using the same function.

For each movie in the dataset, I did a PCA on the last 3 factors (gender identity, sibship status, and social viewing) as different movies would attract a different audience. I then created a multiple regression model based on how many PC's are found for that movie, and cross-validated it by using an 80-20 train-test-split. The R^2 of the model before cross validation ranges from 0 to 0.09827, while the RMSE of the model after cross validation ranges from 0.47002 to 1.5599.

10) Build a prediction model of your choice (regression or supervised learning) to predict movie ratings (for all 400 movies) from all available factors that are not movie ratings (columns 401- 477). Make sure to use cross-validation methods to avoid overfitting and characterize the accuracy of your model.

I did questions 8, 9, 10 using the same function.

For each movie in the dataset, I did a PCA on all the factors as different movies would attract a different audience. I then created a multiple regression model based on how many PC's are found for that movie. I limited the model to having a maximum of 13 determining PC's. I then cross-validated it by using an 80-20 train-test-split. The R^2 of the model before cross validation ranges from 0.02818 to 0.483, while the RMSE of the model after cross validation ranges from 0.63484 to 1.64323.