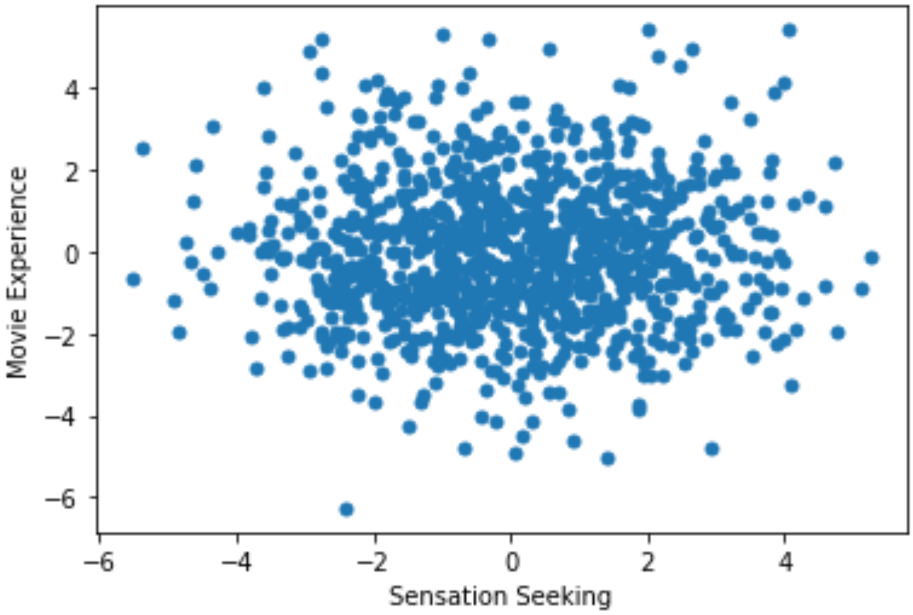
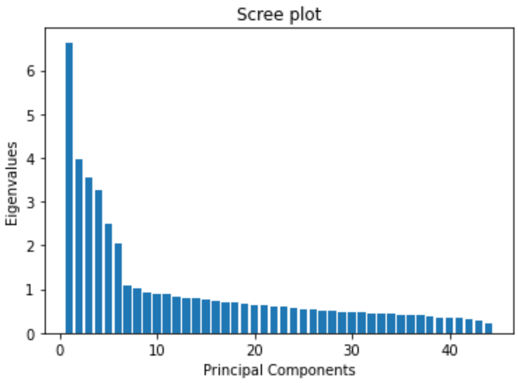
DS-UA 112 Capstone Project
Hanchuan Chen

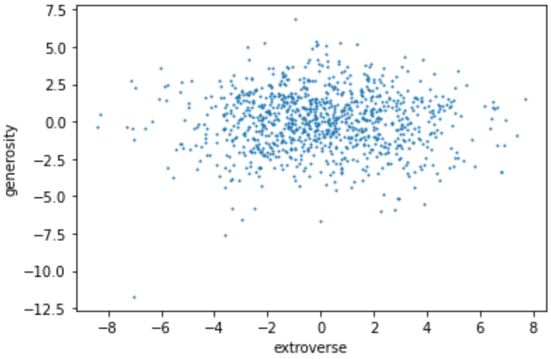
Brief Statement: In this project, I am handling a bunch of data about movie rantings. Since there might have many missing values, data cleaning is necessary. If we need to find the correlation between two data, such as question 1, I would like to combine two data together and clear the missing value row-wise, which means removing the row with NaN. For other data, I would like to clean missing value element-wise, which means I only need to remove the NaN element. Dimension reduction is another significant part. To reduce the dimension, I will z-score the data and run PCA, find the factors by using Kaiser rule, elbow rule, or 90% variance rule, depending on which rule can best interpret the original data. For data transformation, I first z-score the data and then get the rotated data by transforming the z-score data.

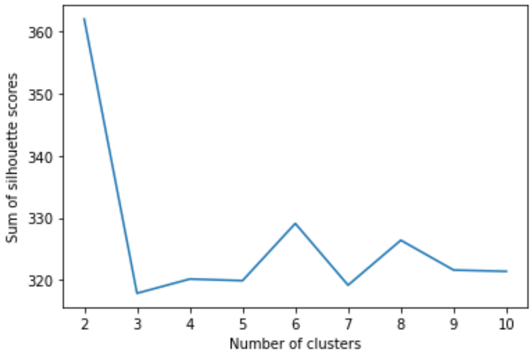
Question 1: In order to find the relationship between sensation seeking and movie experience, we will use the data that have already clean the missing value. Then I use PCA to reduce their dimensions. After doing PCA, I transform and rotate the data, and find the first principal component of sensation seeking and movie experience to do a correlation. The correlation coefficient r for these two data is 0.0124. Therefore, sensation seeking and movie experience might be almost uncorrelated.

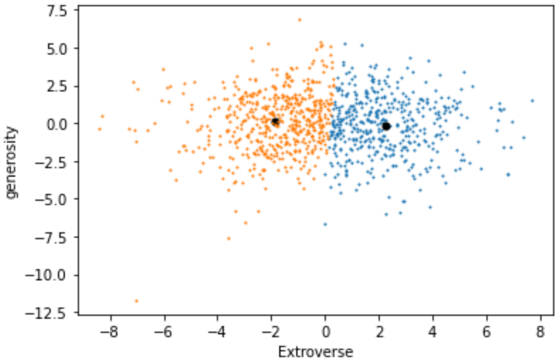


Question 2: To find the different type of personality, we will need to use the "personality" data. After cleaning the data element-wise, I do the dimension reduction and find that there are 6 principal component by using elbow rule. Then I use the first two principal component, rotated them, and plot the cluster, which x-axis is extrovert behavior and y-axis is generosity. Furthermore, I calculate that 2 centroid for this cluster might be optimal by using silhouette method. After finishing clustering, I find that there might have two different personalities: narratively, one is introvert but generous, and another is extrovert and generous. Quantitively, since the cluster only have two centroids, the left side have extrovert scale below 0 and vise-versa. On the generosity axis, since majority of data clustered around 0, so it is hard to determine if personality is generous or not.

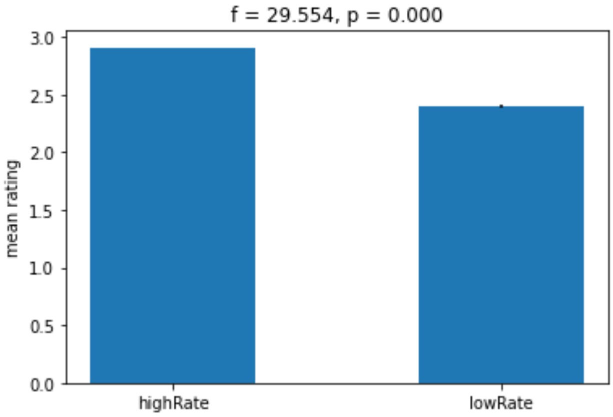




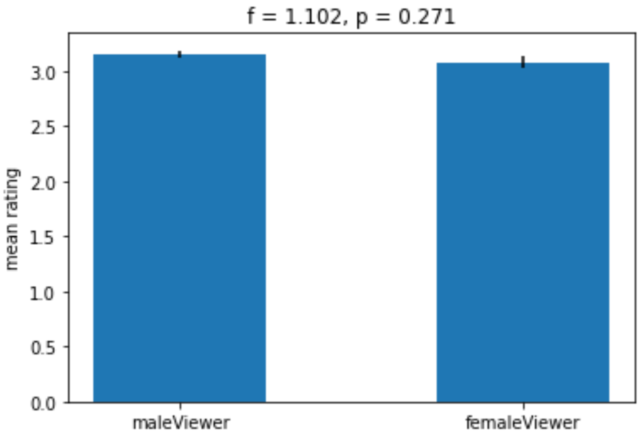




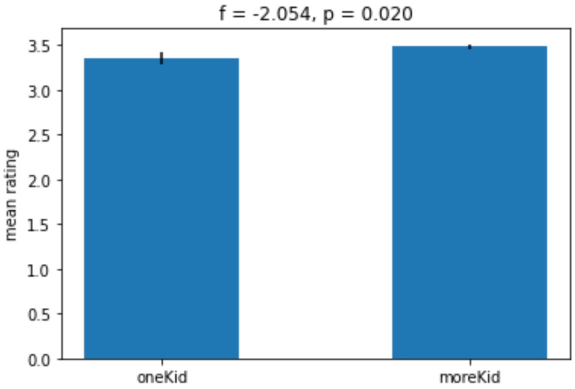
Question 3: For this question, first is to do a hypothesis test. My null hypothesis is that more popular movie has the same rate of less popular movie. Alternative hypothesis is more popular movie has higher rating than less popular movie. Then I do a median split to split the more popular movie and less popular movie. Moreover, I do the independent t test. From the test result, we get that t-test statistics is 29.53 and p-value is nearly approach to 0 for two tailed, so single tail is also 0 which is lower than alpha = 0.05, so we reject the null hypothesis. As a result, more popular movie rated higher than less popular movie.



Question 4: For this question, hypothesis test is also necessary. My null hypothesis is that male and female viewer rate the same. The alternative hypothesis is that male viewer rate differently than female viewer. First, we need to clean the gender who is not male or female and remove the missing value. Then we do the independent t test. This result shows that t-test statistics is 1.101 and p-value for two-tailed is 0.27, p-value is still bigger than alpha = 0.05, so we fail to reject null hypothesis. Conclusion: male and female viewer rate the same.

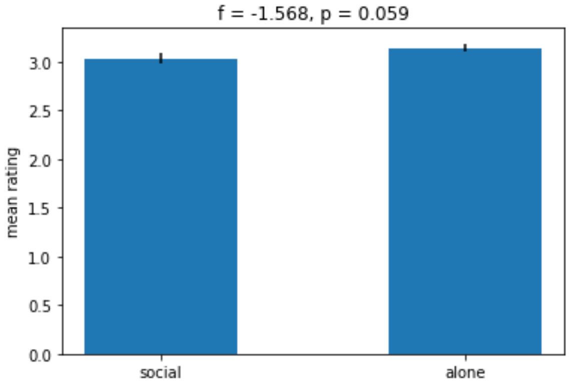


Question 5: Hypothesis test is also needed for this question. The null hypothesis is that people who are child only enjoy the same as people with siblings. The alternative hypothesis is that people who are child only enjoy more than people with siblings. First we clean the data with NaN and people who does not respond. Then let people who does not respond become NaN and remove missing value. Since it is a one-tailed t-test but it returns two-tailed, we should half it. This result shows that t-test statistic is -2.054 and p-value is 0.02 < alpha = 0.05. So we reject the null hypothesis. Conclusion: people who are child only enjoy more than people with siblings.

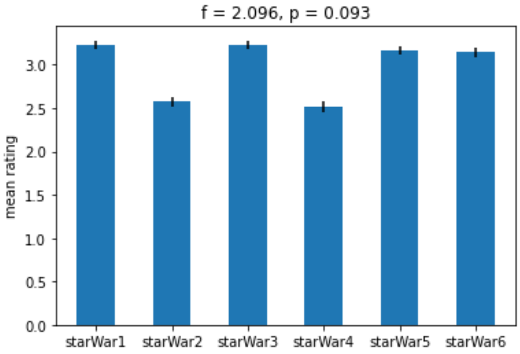


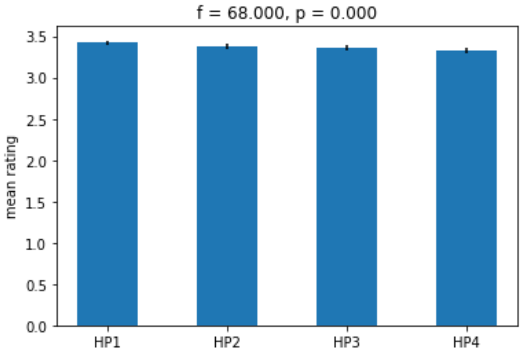
Question 6: Hypothesis test is also needed for this question. The null hypothesis is that people who like watch movie socially enjoy the same as people who watch alone

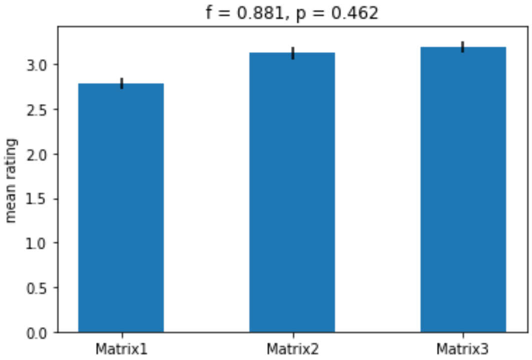
The alternative hypothesis: people who like watch movie socially enjoy more than people watch alone. As always, we clean the data first. Let people who does not respond become NaN and remove the missing value. Since it is also one-tailed t-test so we need to half the calculated p-value. From the result we can see the t-test statistic is -1.568 and p-value is 0.059. Since p-value > alpha = 0.05, we fail to reject null hypothesis. Conclusion: people like watch movie socially enjoy more than people watch alone.

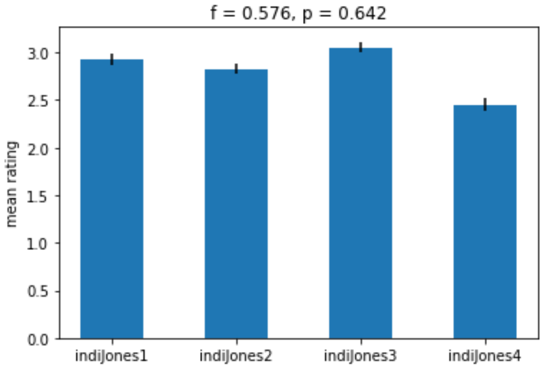


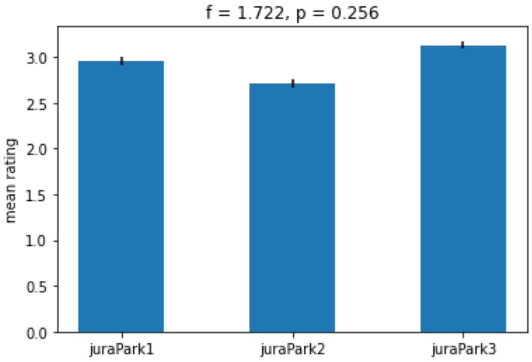
Question 7: Before figuring the question, we need to define inconsistent quality: the mean rating of each season in its franchise has huge difference that means we should use ANOVA to do this hypothesis test. Null hypothesis: each movie in its franchise has no inconsistent quality. Alternative hypothesis: the movie in its franchise has inconsistent quality. There are eight franchises, so we will plot eight graphs. Based on the plots and their separate p-values, Star War, Matrix, Indiana Jones, Jurassic Park, Toy Story, and Batman have inconsistency quality.

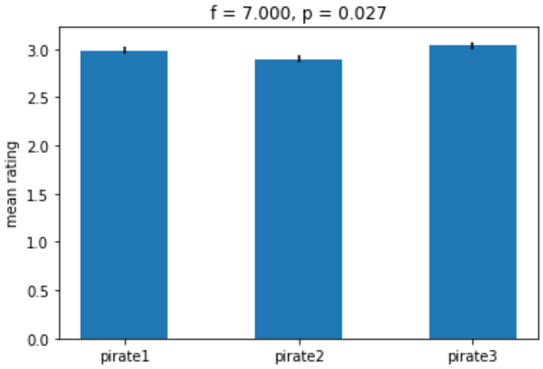


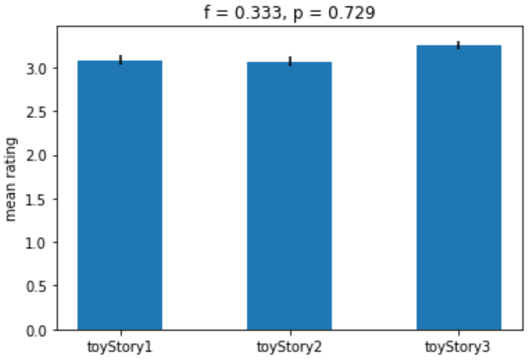


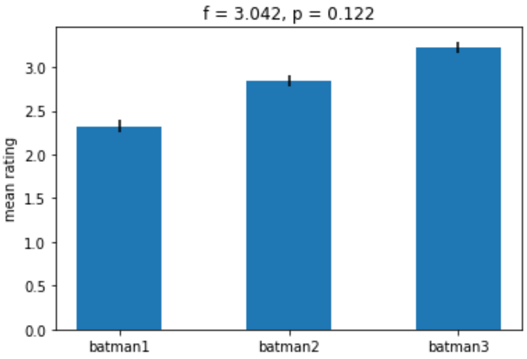




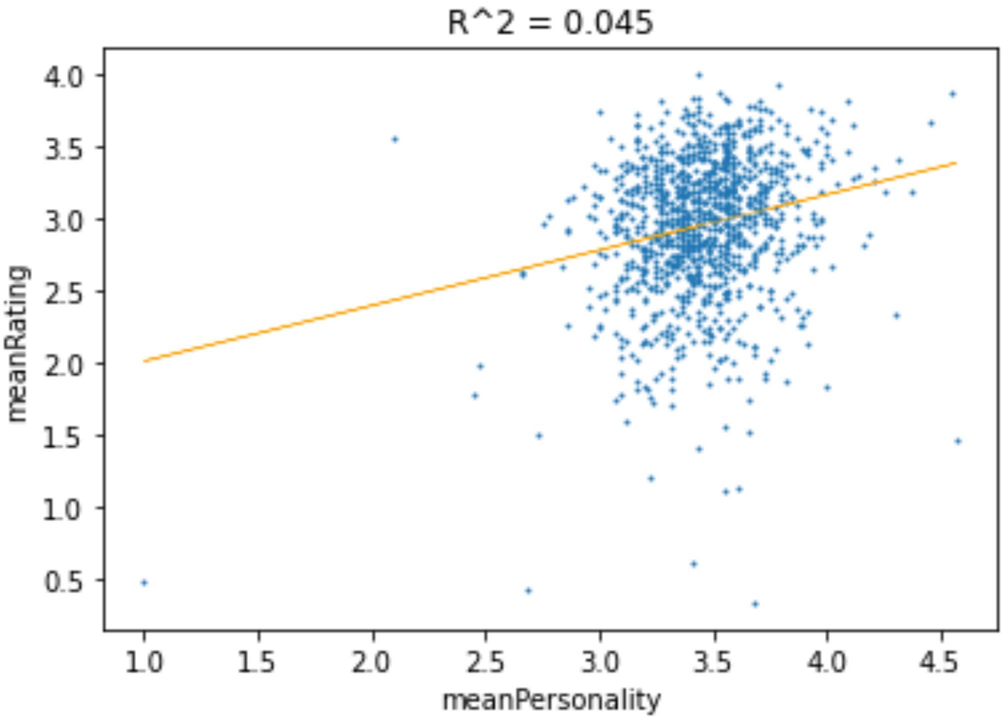




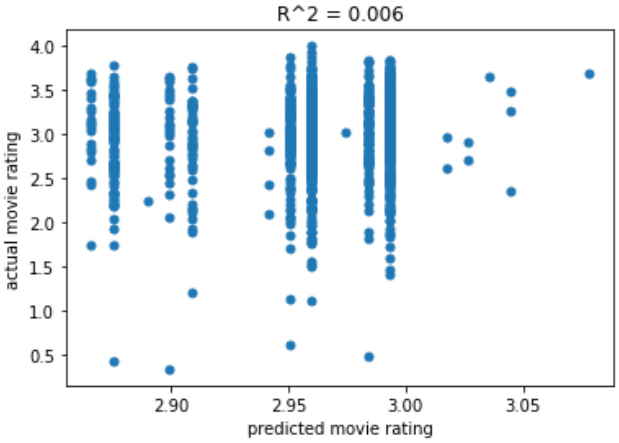




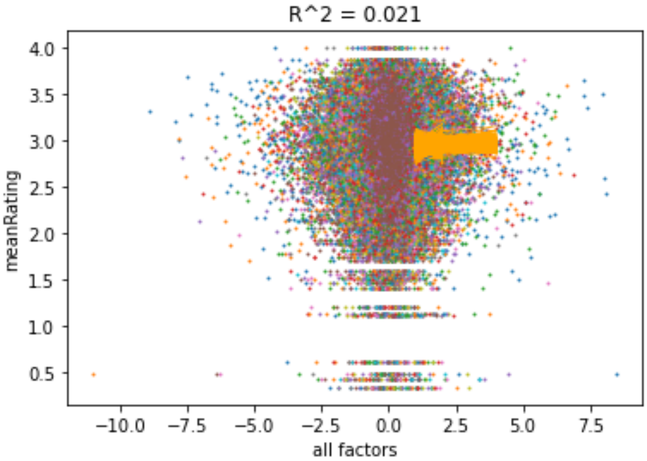
Question 8: In order to predict movies from personalities, I find the mean rating of 400 movies for 1097 people, and mean personality for 1097 people. So one mean rating corresponding to one mean personality. Based on these data I do the simple linear regression for prediction. After cleaning the missing value and build the model, the r^2 is only 0.045, which means the mean personality can only explain 4.5% of mean movie rating, so it might not be a good prediction model.



Question 9: Since this question ask us to find predict the movie rating based on gender, sibship status, and social viewing preference. I think the optimal method is to do the multiple regression. I take the gender, sibship status, and social viewing preference as predictor, and mean rating of each movie as actual outcome. Finally, I plot the scatter plot that x-axis is predicted mean movie rating and y-axis is actual mean movie rating. Since the r^2 is only 0.006, so the predictor only explains less than 1% of outcome, and this prediction model might not be a well-fit model.



Question 10: for this question, I choose to use linear regression to build the prediction model. First I calculate the mean movie rating for every users so there are 1097 ratings. Then I combine this rating with all other factors and do the data cleaning row-wise to remove the missing value. After that, I do the PCA for all factors and use its first principal component as predictor to and take the linear regression. The plot is wired and r^2 is only 0.021, which means by using this model, all factors can only explain 2.1% of mean movie rating. So this might also be a bad prediction model.



Extra Credit: In the movie rating dataset (first 400 columns in the data), people tend to give the moderate score such as 3 or 3.5. On the contrast, they prefer not to rate the movie with lowest score which is 0.