Anni Zheng
DS-UA-112

Capstone Project

Introduction

This project used Python and NumPy, scikit-learn, Matplotlib libraries to respond several questions regarding the possible relationships or models of data from *movieReplicationSet.csv*. Element-wise and row-wise reduction were used accordingly. PCA was used for dimension reduction. Some large data were splatted into several small data sets if necessary. And during the use of logistic regression, median of a data set was a cut-off point to accommodate binary logic.

Question 1.

To find the relationship between sensation seeking and movie experience, I extracted of the two data and applied row-wise reduction of Nans. Since there were more than 10 variables for each category, I applied PCA for dimension reduction to only keep the variables that contribute significantly to the relationship.

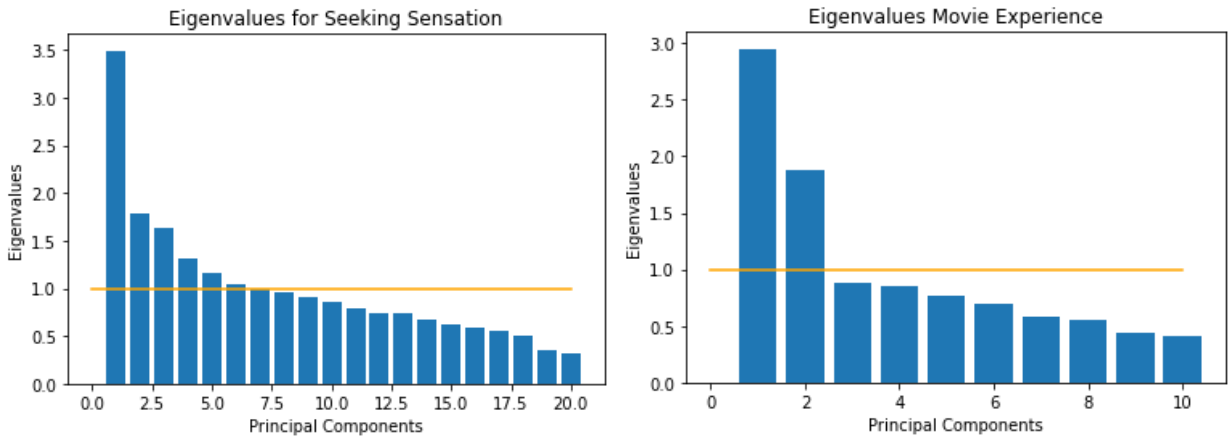

Figure 1

To select the critical principal components for examination, I used Kaiser criterion and decided to examine the first six principal components for those of seeking sensation and the first two principal components for those of movie experience (Figure 1). However, it was still hard to distinguish between the significance of the six principal components of seeking sensation, so I looped through the variances of the loadings of those six components, and decided to choose principal component 3, which with the greatest variance of 0.0498. My logic followed that of PCA's – with greater variance, one may find more useful data that did not align with each other.

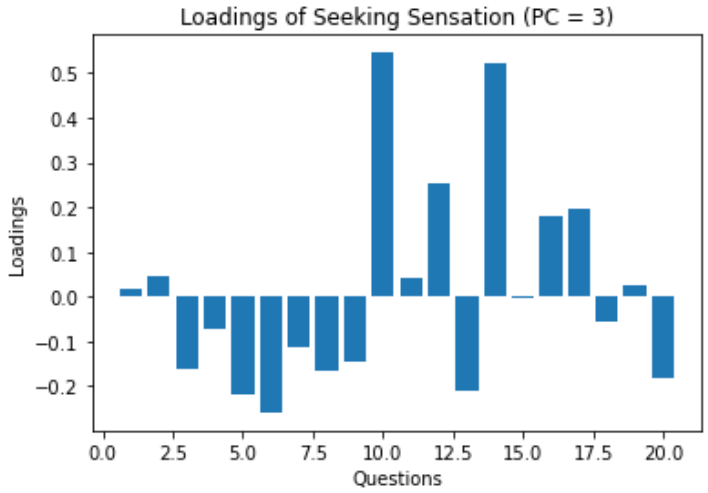

Figure 2

I examined the loadings of principal component 3 visually and found question 6, 10, and 14 being the two questions contributed to the score of seeking sensation significantly in two opposite directions (Figure 2). Note that question 6 was asking if one preferred to plan; question 10 was asking if one preferred festival; and question 14 was asking if one had parachuted. On the other hand for movie experience, I choose principal component 2 to examine the loadings since there are too many negative loadings in principal component 1 (Figure 3).

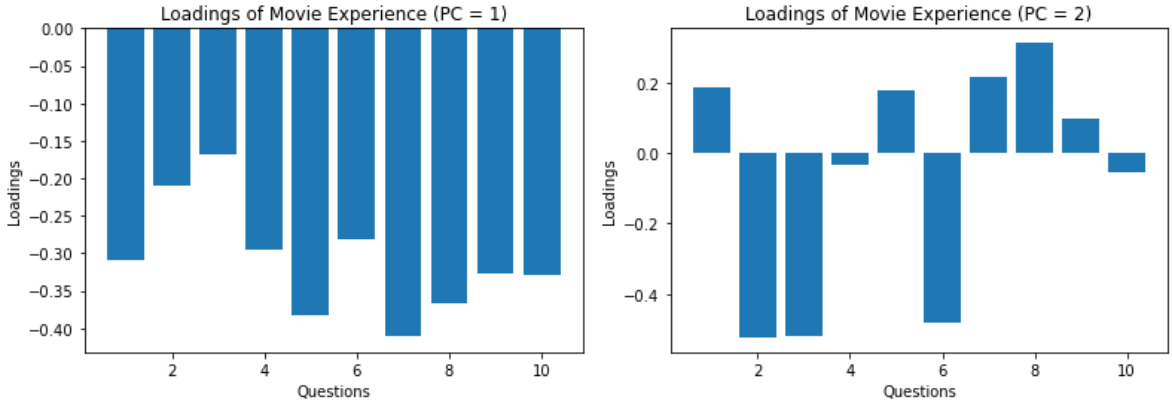

Figure 3

Fore movie experience, I decided to use question 2, which asked if one had forgotten the movie plot. Then, I used those components to create two spaces for multiple linear regression, and examined a R^2 of 0.733, or a R of 0.856 (Figure 4). The high R^2 indicated that it was possible to use sensation seeking to predict movie experience, there was a strong relationship between the two categories.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared (uncentered):                   0.733
Model:                            OLS   Adj. R-squared (uncentered):              0.732
Method:                 Least Squares   F-statistic:                              936.6
Date:                Sun, 21 Aug 2022   Prob (F-statistic):                   3.44e-293
Time:                        23:11:11   Log-Likelihood:                         -1894.6
No. Observations:                1029   AIC:                                      3795.
Df Residuals:                    1026   BIC:                                      3810.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             0.4635      0.027     17.414      0.000       0.411       0.516
x2             0.2328      0.038      6.054      0.000       0.157       0.308
x3             0.0947      0.041      2.330      0.020       0.015       0.175
==============================================================================
Omnibus:                       11.727   Durbin-Watson:                   1.831
Prob(Omnibus):                  0.003   Jarque-Bera (JB):               10.092
Skew:                           0.177   Prob(JB):                      0.00644
Kurtosis:                       2.668   Cond. No.                         6.03
==============================================================================

Notes:
[1] R² is computed without centering (uncentered) since the model does not contain a constant.
[2] Standard Errors assume that the covariance matrix of the errors is correctly specified.
0.8561541917201597
```

Figure 4

Question 2.

To find the personality type of research participant, I decided to use DBSCAN clustering method. I first extracted the personality information and remove the Nans. Notice that there were 44 variables (questions) to indicate personality, which was too much, so I applied PCA to reduce demotions. I repeated the steps introduced in Question 1 and used Kaiser criterion to choose the principal components to examine and found out that there are 8 of them I could choose (Figure 5). Again, it was too much.

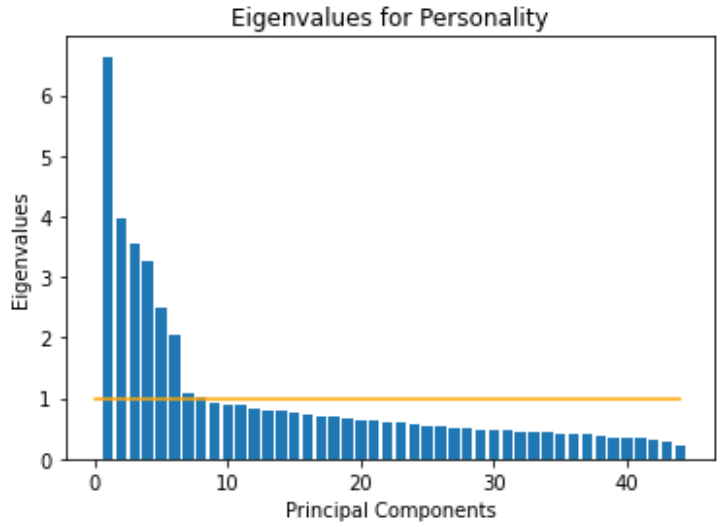

Figure 5

At this time, I decided to choose principal component 1 as not only was it having the highest eigenvalues (greatest scale), but the distribution of loadings also seemed great (Figure 6), so there was no need to look for others.

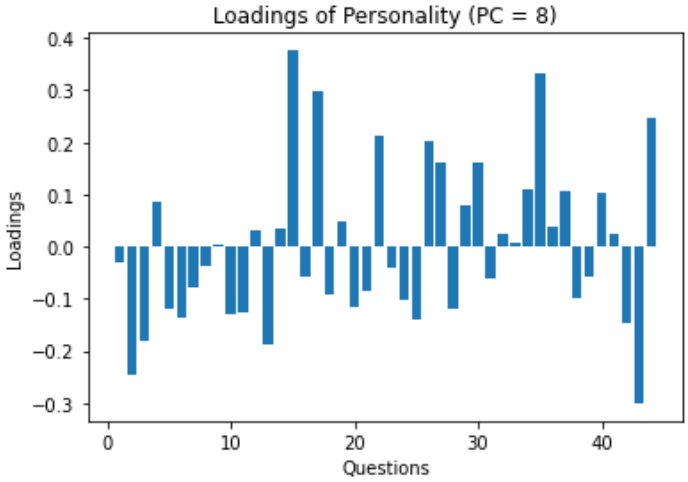

Figure 6

Based on the loadings, I plotted the graph for the spread of them, and applied DBSCAN to find clusters (Figure 7). Ultimately there were two clusters detected.

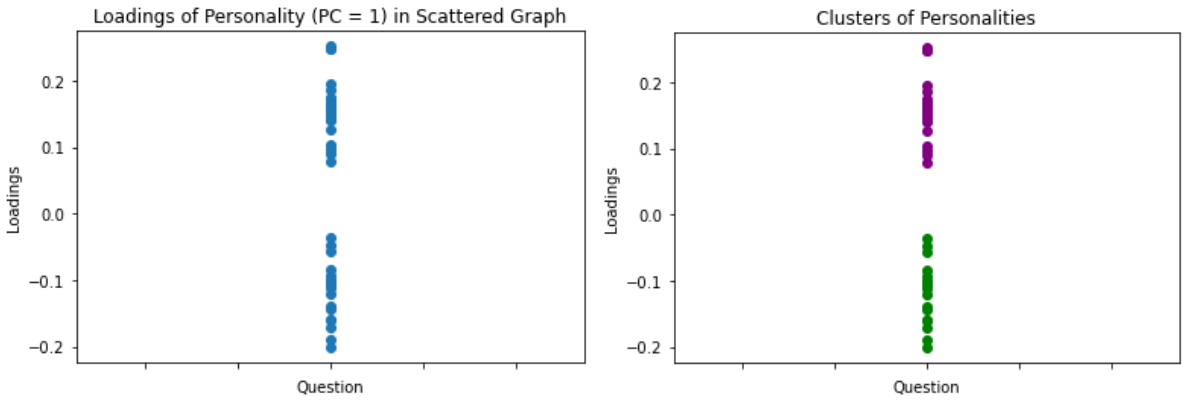

Figure 7

I then picked the questions based on their loading values. The questions with great positive loadings were asking if a person was extroverted or energetic, while questions with great negative loadings were asking if a person was reserved or depressed. So people were clustered by asking if they were introverted or extroverted, pessimistic or optimistic, etc.

Question 3.

For this question, I first looped through the 400 movies and calculated their popularity based on the number of scored reported by participants. I also row-wisely reduce Nans. Then, I decided to use the median of each movie's score to represent the overall score of this movie. I used median instead of mean to remove the effect of score outliers (Figure 8). One may see from Figure 8, with purple line being the median and green line being the mean, that the mean was greatly influenced by the low ratings, which might not be a good reflection of a movie's score.

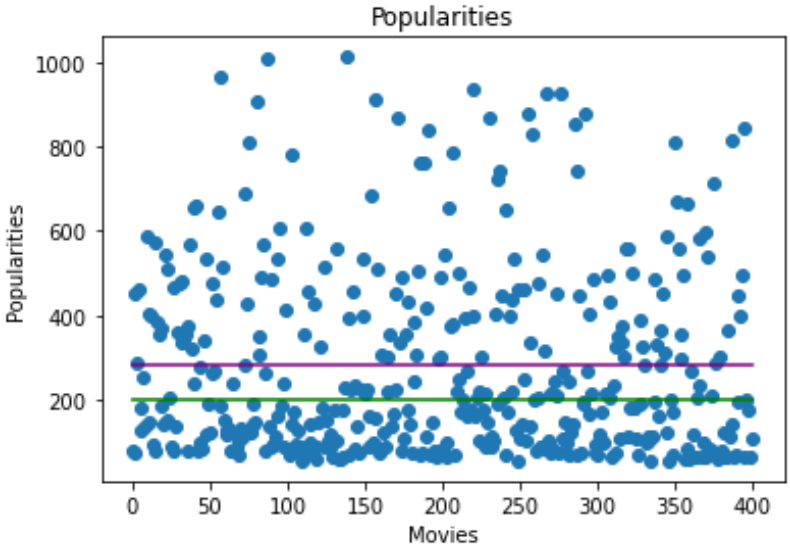

Figure 8

Then I used the median of each movie and its popularity to plot the graph, and a positive relationship was detected (Figure 9). So movies that were popular were rated highly

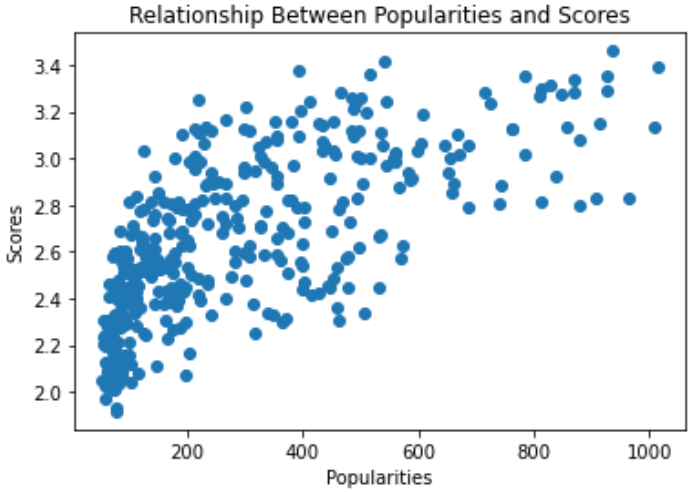

Figure 9

To go further, I used logistic regression, and I used the median popularity to cut off between popular and not-popular movies to confirm that movies that were popular were rated highly (Figure 10).

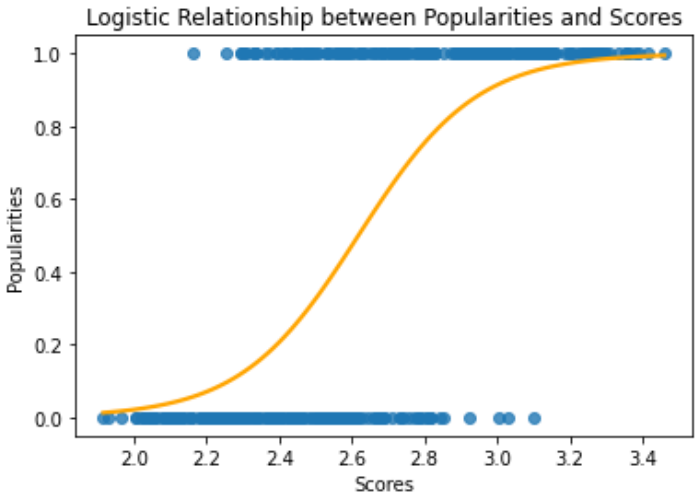

Figure 10

Question 4

To examine if the gender difference contributed to the ratings of Shrek (2001), I used the ANOVA test to examine the relationship between females' ratings, males' rating, and self-described people's ratings. I first extracted the information and did a row-wise reduction. Assuming the null hypothesis of *there was no relationship*, I then applied the ANOVA test, and a p-value of 0.376 was returned. The p-value was larger than 0.05, so I could not reject the null hypothesis (Figure 11). To confirm that there was indeed no relationship, I applied individual t-test to females' ratings and males' ratings, females' ratings and self-described people's ratings, and males' ratings and self-described people's ratings, and they each returned a p-value of 0.271, 0.349 and 0.562, which was again too large (Figure 11). So, I concluded that there was no relationship.

```
F_onewayResult(statistic=0.9781679760055932, pvalue=0.37635698516296434)
Ttest_indResult(statistic=1.1016699726285888, pvalue=0.27087511813734183)
Ttest_indResult(statistic=0.9366215198198362, pvalue=0.3492488598175757)
Ttest_indResult(statistic=0.5805522632613999, pvalue=0.5620393557212684)
```

Figure 11

Question 5

I used the same strategy for this question as that of question 4 – I first extracted the data, did the row-wise reduction of Nans, and separated the data between rating data from those with siblings and rating data from those without siblings. Assuming the null hypothesis of *no* relationship, I then applied the individual t-test, and a p-value of 0.0403 was returned. The p-value was smaller than 0.05 so the null hypothesis was rejected. There was indeed a relationship existed between the score given by viewers with siblings and those given by viewers without siblings. I plotted a box graph to examine the actual difference and found out that those with siblings gave higher ratings to The Lion King (1994) (Figure 12).

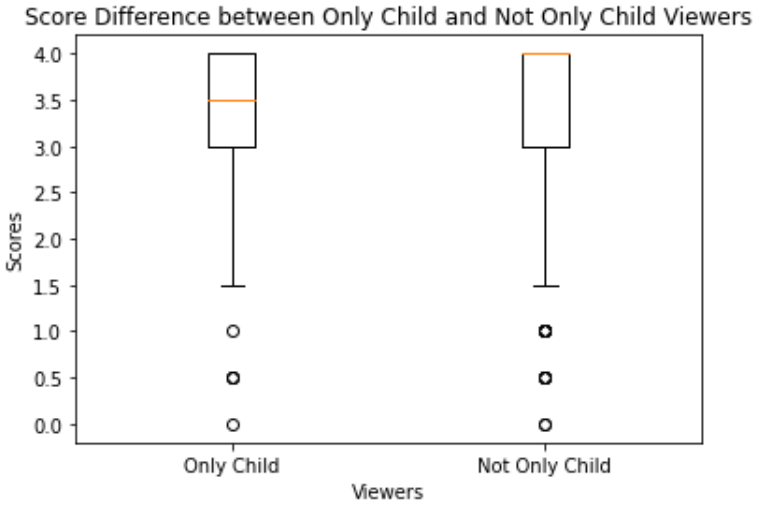

Figure 12

Question 6

I again used the same method as that of last question – I extracted the data, applied row-wise reduction of Nans, separated between the ratings from those who preferred to be alone and those who preferred to be not alone (and ignored those who did not answer). I assumed the null hypothesis of there was no relationship between social preference and the ratings given by a person, and the p-value given by the individual t-test was 0.117, which was larger than 0.05, so I could not reject the null hypothesis (Figure 13). There was no relationship.

```
Ttest_indResult(statistic=1.567873874504994, pvalue=0.11738913665664574)
```

Figure 13