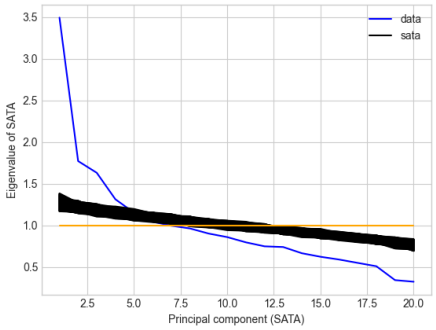
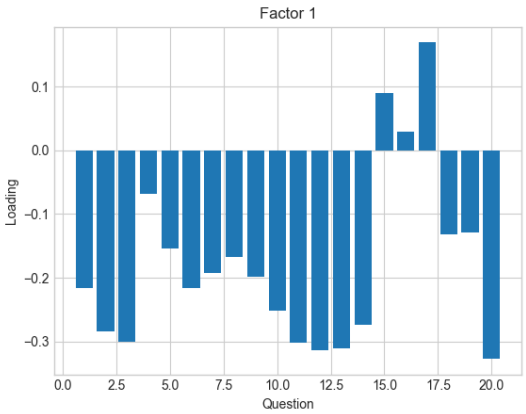
PRELIMINARY INFORMATION

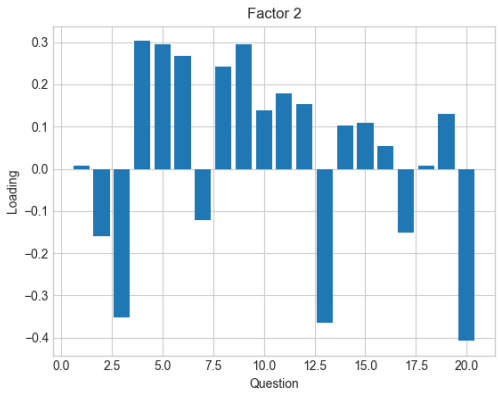
I have used PCA as my choice of dimension reduction. Missing values have been ignored. I have given all values to 2dp unless stated otherwise. Alpha levels have been set at 0.05 unless stated otherwise.

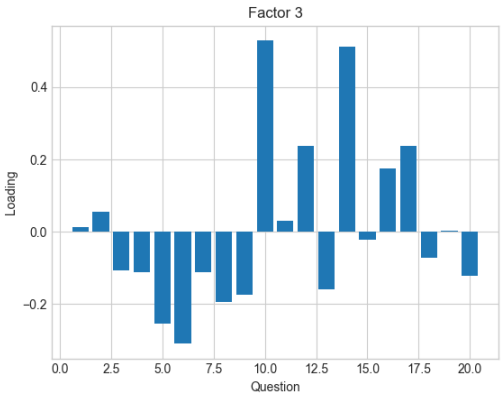
QUESTION 1

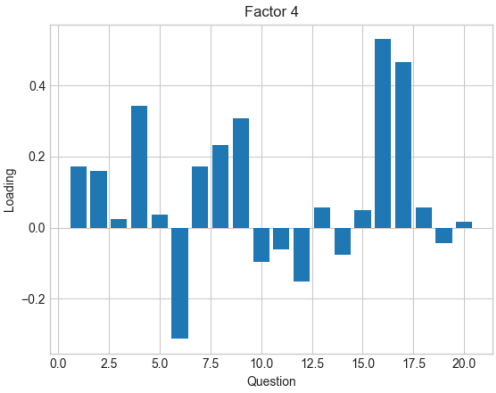
I have removed all rows with missing data, and z scored the data, and then performed PCA. Here is the scree plot using the horn method for sensation seeking. We can see that only the first four factors have passed this test. Below, we are able to see the loadings for each factor.



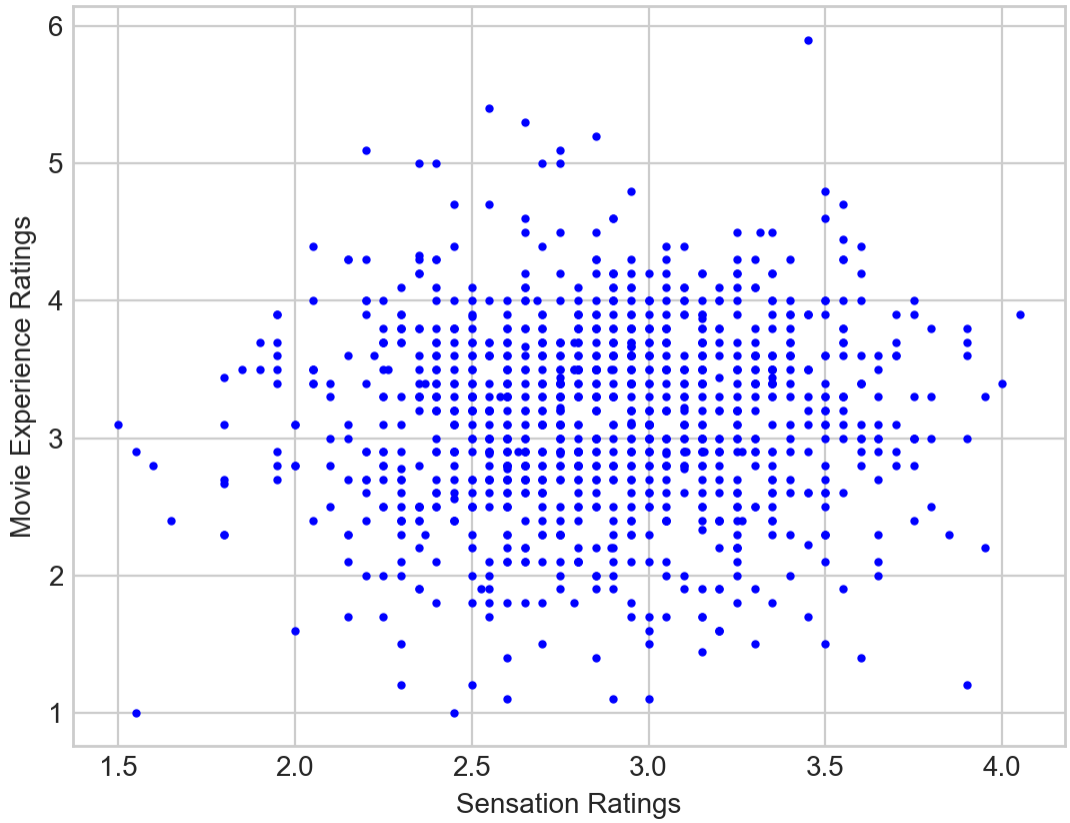








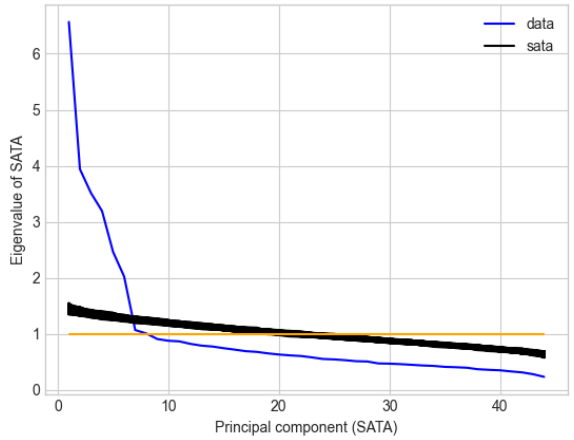
Alternatively, we can also just calculate the averages for each category and then correlate them.

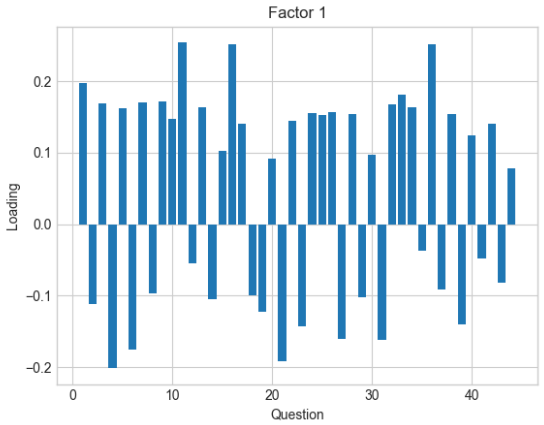


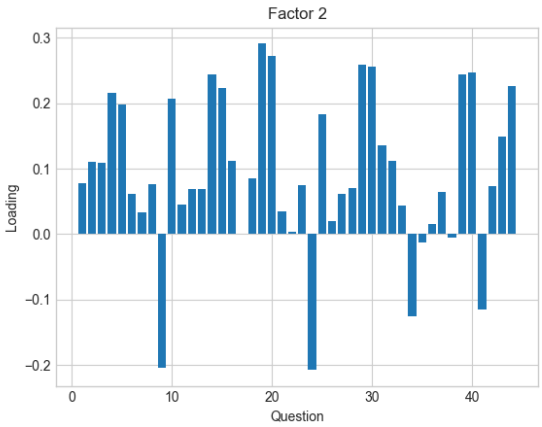
There is a correlation of 0.07. Therefore, we must conclude that based off this alternative form of dimension reduction, there seems to be no evidence of a relationship.

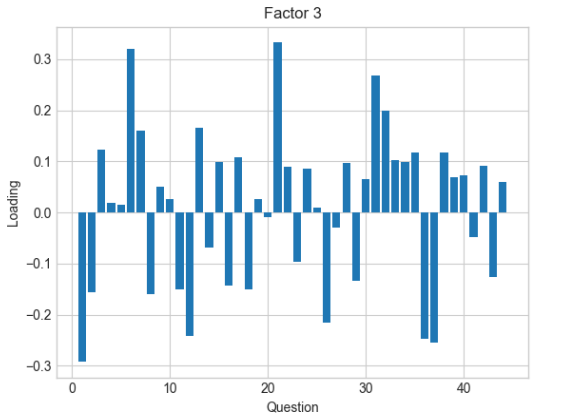
QUESTION 2

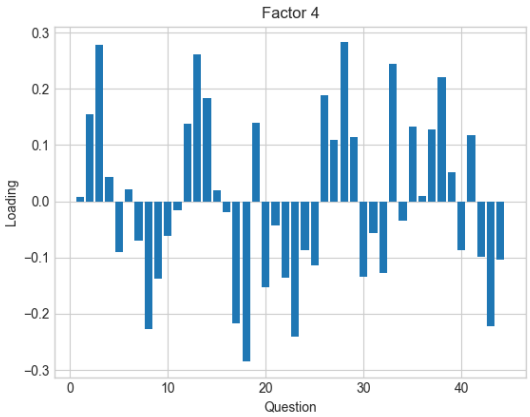
For personality, I have removed all rows with missing data, and z scored the data, and then performed PCA. Here is the scree plot using the horn method. We can see that only the first five factors have passed this test. Below, we are able to see the loadings for each factor.

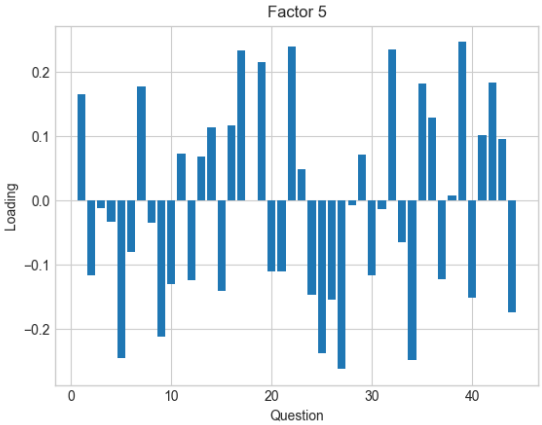




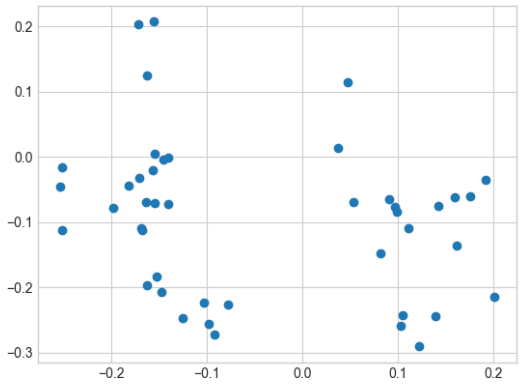








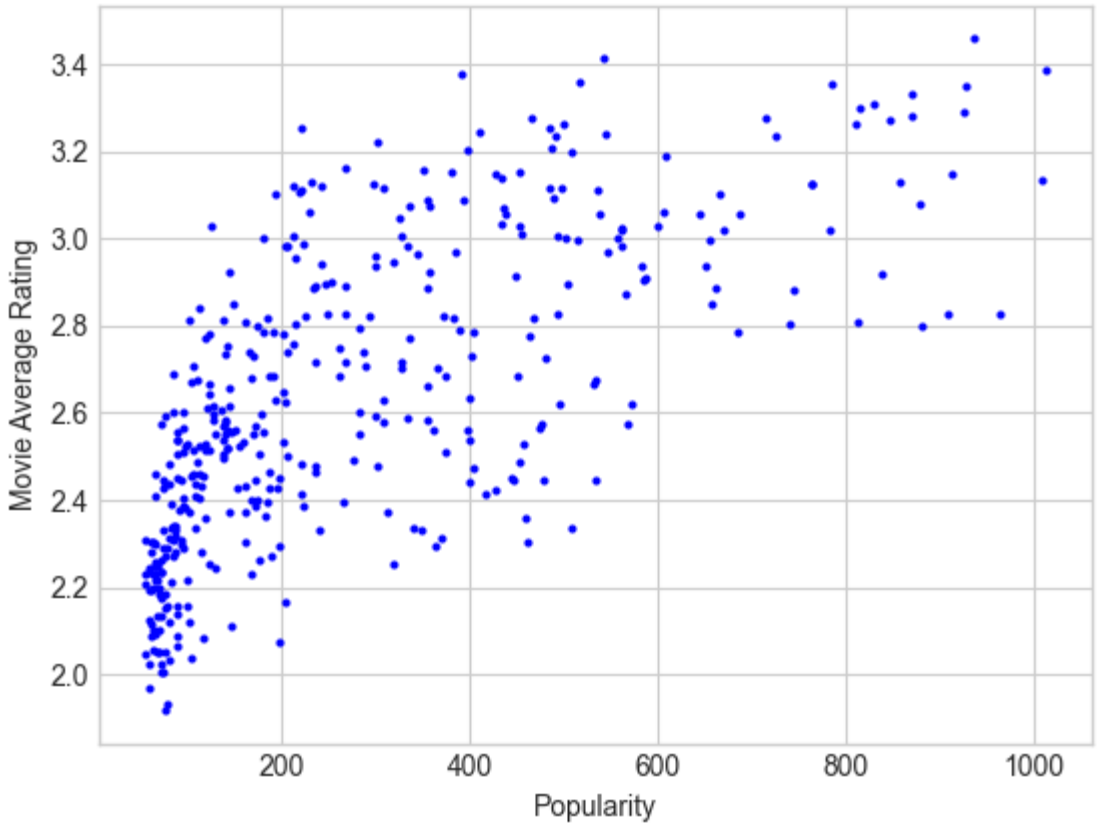
For the purposes of clustering, let us plot PC1 and PC2 on a 2d graph.



It seems quite apparent that there are two distinct clusters. Now that these personality types have been identified quantitatively, let us identify them narratively. If we examine the PCs and the questions, it seems give evidence of those who are enthusiastic and energetic, and those who are less so.

QUESTION 3

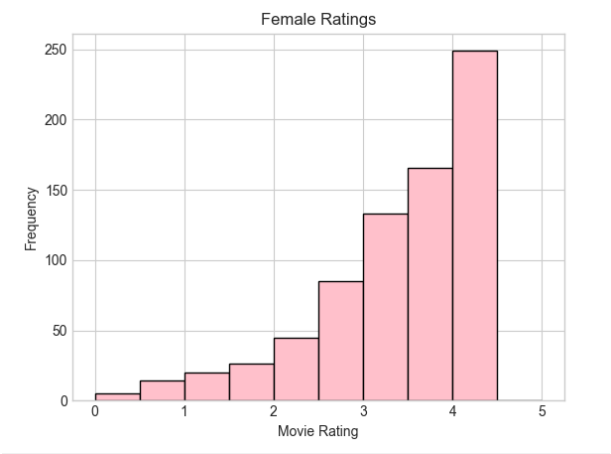
I have taken the average of each movie and counted the number of ratings for each movie. The plot for this data can be seen below.

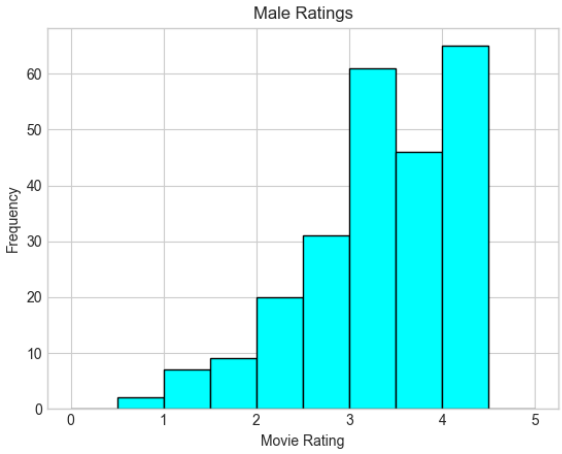


Using a spearman correlation, the r value is 0.76. The p value is 9.6*10^-77. Therefore, it is safe to conclude, with high confidence, that there is a strong relationship between movie popularity and rating.

QUESTION 4

For this question, I will only be comparing the male and female viewers. Below, I have shown the distributions of the male and female ratings.

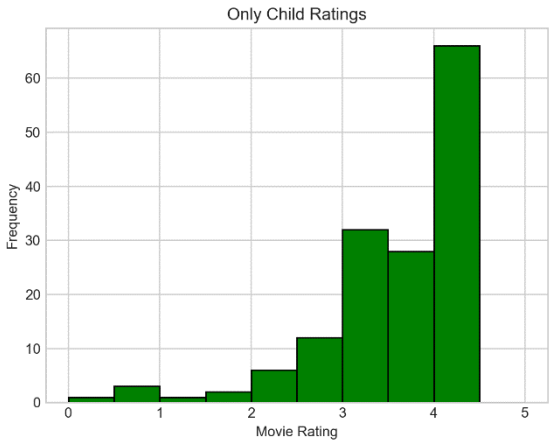


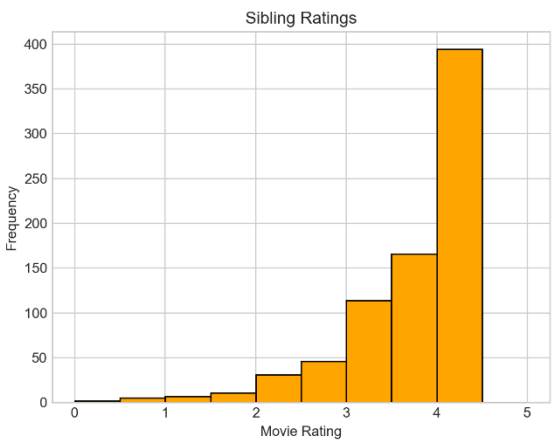


We can see that the data is not distributed normally. Thus, I have opted to use a non-parametric test. I will be using the Mann-Whitney U test. The null hypothesis is that the enjoyment of Shrek is not gendered i.e. male and female viewers do not rate it differently. The p-value is 0.05, so we fail to reject the null hypothesis. Therefore, we can conclude that the enjoyment of Shrek is not gendered.

QUESTION 5

For this question, I will only be comparing those who have answered whether they have siblings or are an only child. Those who have no response will be ignored. Below, I have shown the distributions of the only child and sibling ratings.



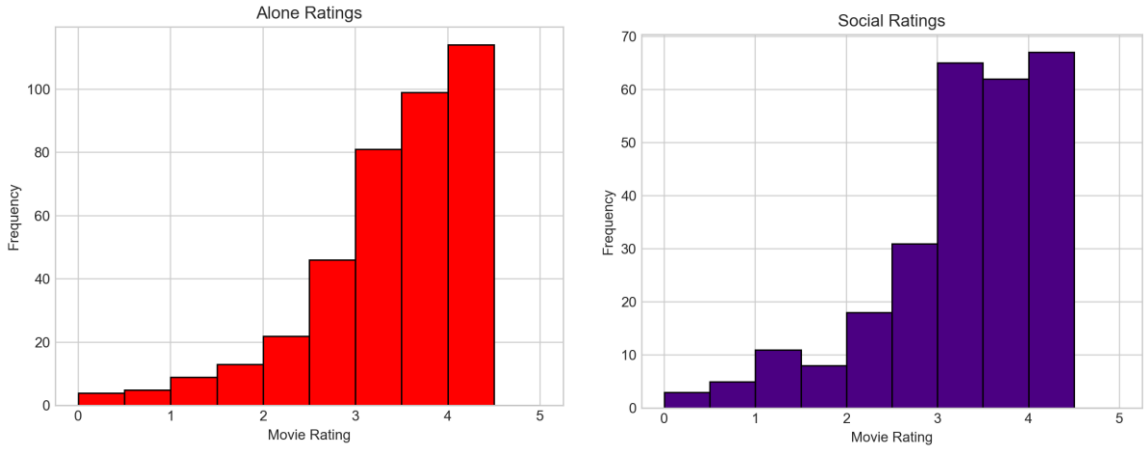


We can see that the data is not distributed normally. Thus, I have opted to use a non-parametric test. I will be using the Mann-Whitney U test. The null hypothesis is that the enjoyment of Lion

King is not influenced by whether someone is an only child or not. The p-value is 0.04 and is significant, so we reject the null hypothesis. Therefore, we can conclude that those who are and only child rate Lion King differently than those with siblings.

QUESTION 6

For this question, I will only be comparing those who have answered whether they prefer watching alone or socially. Those who have no response will be ignored. Below, I have shown the distributions of the alone and social ratings.



We can see that the data is not distributed normally. Thus, I have opted to use a non-parametric test. I will be using the Mann-Whitney U test. The null hypothesis is that the enjoyment of The Wolf of Wall Street is not influenced by whether someone prefers watching alone or socially. The p-value is 0.04 and is significant, so we reject the null hypothesis. Therefore, we can conclude that those who are an only child rate Lion King differently than those with siblings.

QUESTION 7

In order to answer this question, I have compared each franchise movie against each other using the Kruskal-Wallis test. I have subset the movie dataframe to only contain movies of the franchise. Then, I have removed all NaN values, meaning that only people who have seen all movies in the franchise are left in the subset.

We can conclude that these franchises are of inconsistent quality because the p value is less than 0.05, therefore we can reject the null hypothesis that these franchises are of consistent quality – Batman, Star Wars, The Matrix, Indiana Jones, Jurassic Park, Pirates of the Caribbean, Toy Story

```
Comparing Batman movies: Index(['Batman & Robin (1997)', 'Batman (1989)',
       'Batman: The Dark Knight (2008)'],
      dtype='object')
KruskalResult(statistic=84.65778425637279, pvalue=4.1380499020034183e-19)
```

```
Comparing The Matrix movies: Index(['The Matrix Revolutions (2003)', 'The Matrix Reloaded (2003)',
       'The Matrix (1999)'],
      dtype='object')
KruskalResult(statistic=40.32303905969196, pvalue=1.7537323830838066e-09)
```

```
Comparing Star Wars movies: Index(['Star Wars: Episode IV - A New Hope (1977)',
       'Star Wars: Episode II - Attack of the Clones (2002)',
       'Star Wars: Episode V - The Empire Strikes Back (1980)',
       'Star Wars: Episode 1 - The Phantom Menace (1999)',
       'Star Wars: Episode VII - The Force Awakens (2015)',
       'Star Wars: Episode VI - The Return of the Jedi (1983)'],
      dtype='object')
KruskalResult(statistic=166.7790826672854, pvalue=5.136543575941175e-35)
```

```
Comparing Indiana Jones movies: Index(['Indiana Jones and the Last Crusade (1989)',
       'Indiana Jones and the Temple of Doom (1984)',
       'Indiana Jones and the Raiders of the Lost Ark (1981)',
       'Indiana Jones and the Kingdom of the Crystal Skull (2008)'],
      dtype='object')
KruskalResult(statistic=54.19395477406098, pvalue=1.020118354785894e-11)
```

```
Comparing Jurassic Park movies: Index(['The Lost World: Jurassic Park (1997)', 'Jurassic Park III (2001)',
       'Jurassic Park (1993)'],
      dtype='object')
KruskalResult(statistic=49.42733030275783, pvalue=1.8492328391686058e-11)
```

```
Comparing Pirates of the Caribbean movies: Index(['Pirates of the Caribbean: Dead Man's Chest (2006)',
       'Pirates of the Caribbean: At World's End (2007)',
       'Pirates of the Caribbean: The Curse of the Black Pearl (2003)'],
      dtype='object')
KruskalResult(statistic=6.660021086485515, pvalue=0.035792727694248905)
```

```
Comparing Toy Story movies: Index(['Toy Story 2 (1999)', 'Toy Story 3 (2010)', 'Toy Story (1995)'], dtype='object')
KruskalResult(statistic=23.496729938969775, pvalue=7.902234665149812e-06)
```

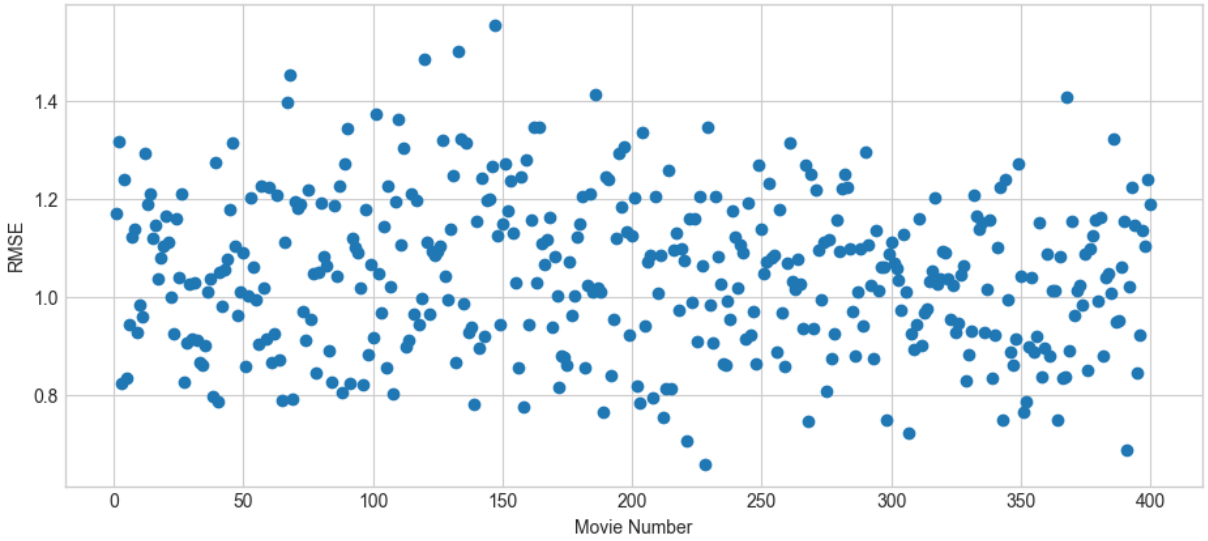
We can conclude that this franchise is of consistent quality because the p value is greater than 0.05, therefore we fail to reject the null hypothesis that this franchises is of consistent quality – Harry Potter.

```
Comparing Harry Potter movies: Index(['Harry Potter and the Sorcerer's Stone (2001)',
       'Harry Potter and the Deathly Hallows: Part 2 (2011)',
       'Harry Potter and the Goblet of Fire (2005)',
       'Harry Potter and the Chamber of Secrets (2002)'],
      dtype='object')
KruskalResult(statistic=0.8156012598972984, pvalue=0.6651114689124273)
```

QUESTION 8

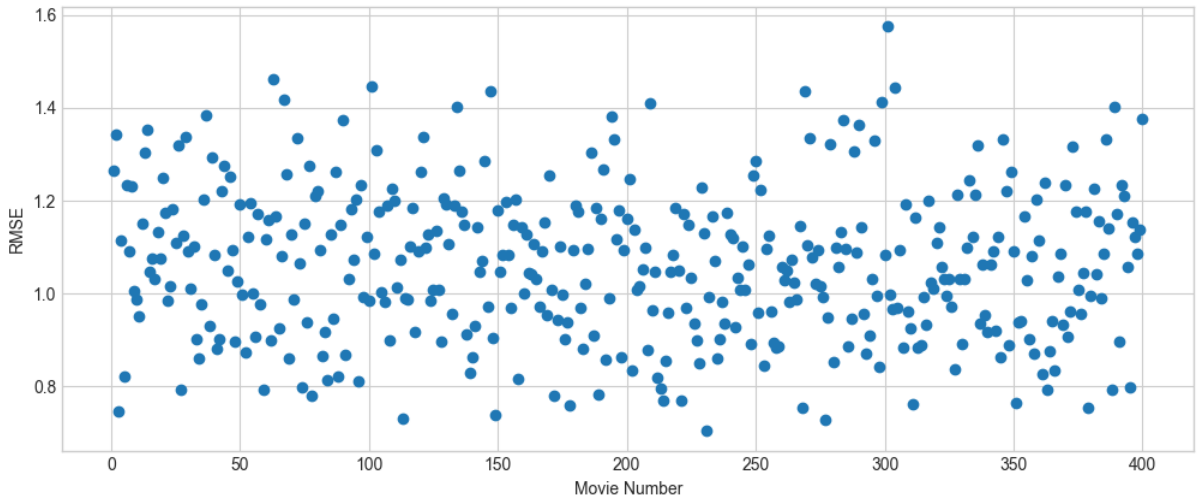
I have chosen to use multiple regression to answer this question. There are 400 target variables and 44 personality variables. I have performed PCA on the personality variables and decided to use 4 PCs for the multiple regression. I have split the movies into 80% training set and 20% test set. I have used the K-folds cross validator. I have run a multiple regression using the aforementioned parameters on each of the 400 movies. As we can see below, the RMSEs are

spread out and differ by movie. Each movie's training set is unique, because NaN values were removed in accordance with each movie's responses.



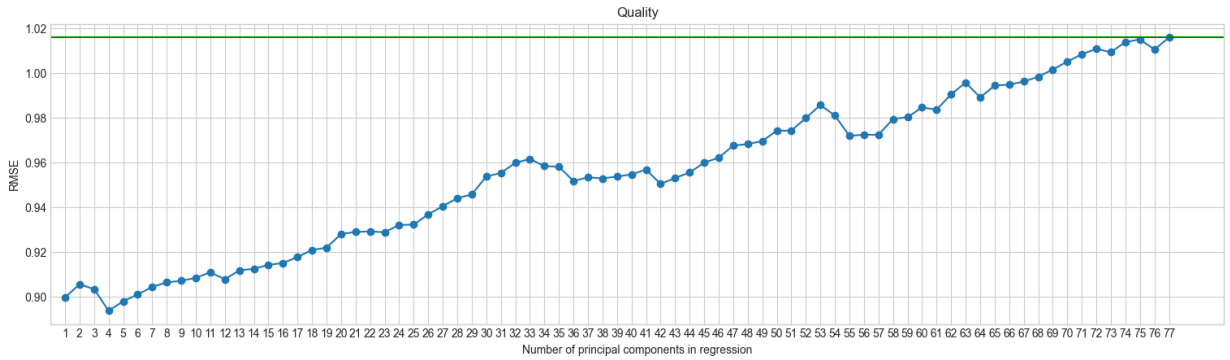
QUESTION 9

I have chosen to use multiple regression to answer this question. There are 400 target variables, and only three predictor values, which are gender, only child status, and viewing preference. I have split the movies into 80% training set and 20% test set. I have used the K-folds cross validator. I have run a multiple regression using the three mentioned predictors on each of the 400 movies. We can see the RMSEs of each movie below. Each movie's training set is unique, because NaN values were removed in accordance with each movie's responses.

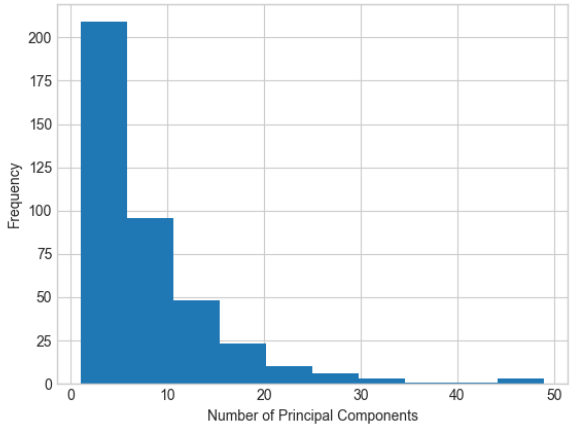


QUESTION 10

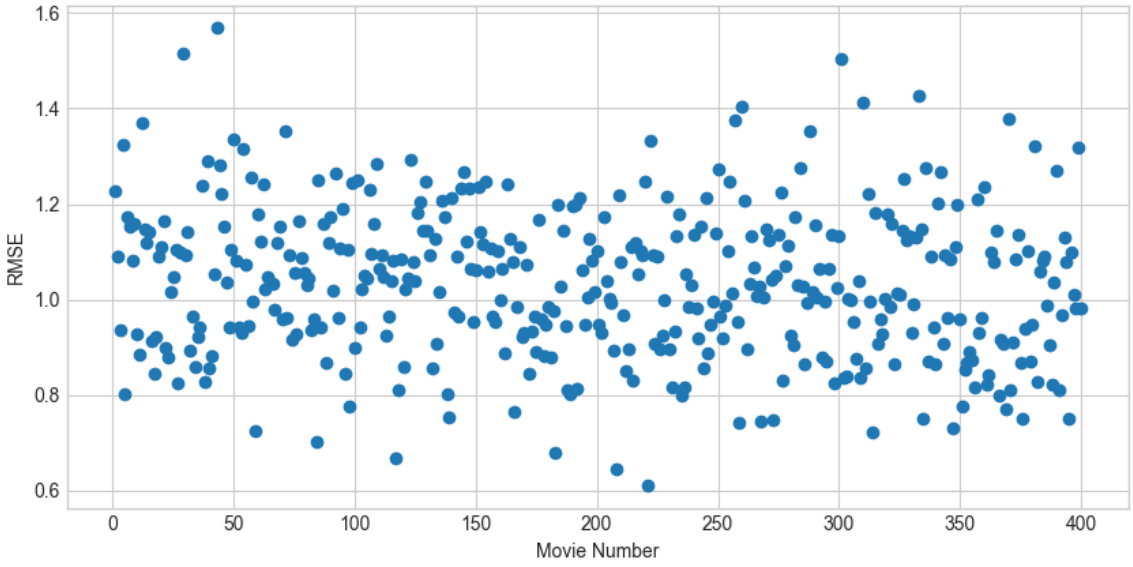
I have tackled this question in a similar manner as question 8. I have done PCA on all the available variables other than movie ratings. Below we can see an example of how RMSE varies with the number of PCs.



This is for one movie. I have repeated this for each time to get 400 of the lowest PCs.



We can see that 4 PCs are the ideal number of PCs to use for the regression model. I have split the movies into 80% training set and 20% test set. I have used the K-folds cross validator. I have run a multiple regression using the three mentioned predictors on each of the 400 movies. We can see the RMSEs of each movie below. Each movie's training set is unique, because NaN values were removed in accordance with each movie's responses.



EXTRA CREDIT QUESTION

I have chosen to ask: “Are old movies rated differently than new movies?”

To answer this question, I have opted to classify any movie older than 2000 as old, and any movie 2000 or newer as new. There are 244 movies older than 2000 and 156 movies that are 2000 or newer. Each movie’s ratings have been averaged and added to an array. One array contains the averages for the old movies, and the other contains averages for the new movies. After performing a Mann-Whitney U test, we can see from the p-value below, that we fail to reject the null hypothesis, so there is no evidence to suggest that old movies are rated differently to new movies.

	U-val	alternative	p-val	RBC	CLES
MWU	19546.0	two-sided	0.64889	-0.027007	0.513504