Data Analysis Project 1

Group Number: 19

Group Members:

1. Rui Xu, rx2254
2. Sijun Wei, sw6155
3. Jingsheng Zhang, jz6464

Question1:

Based on our p-value=0 being less than the significance level (0.005) and the butterfly chart, we conclude that **popular movies have higher ratings than unpopular movies**.

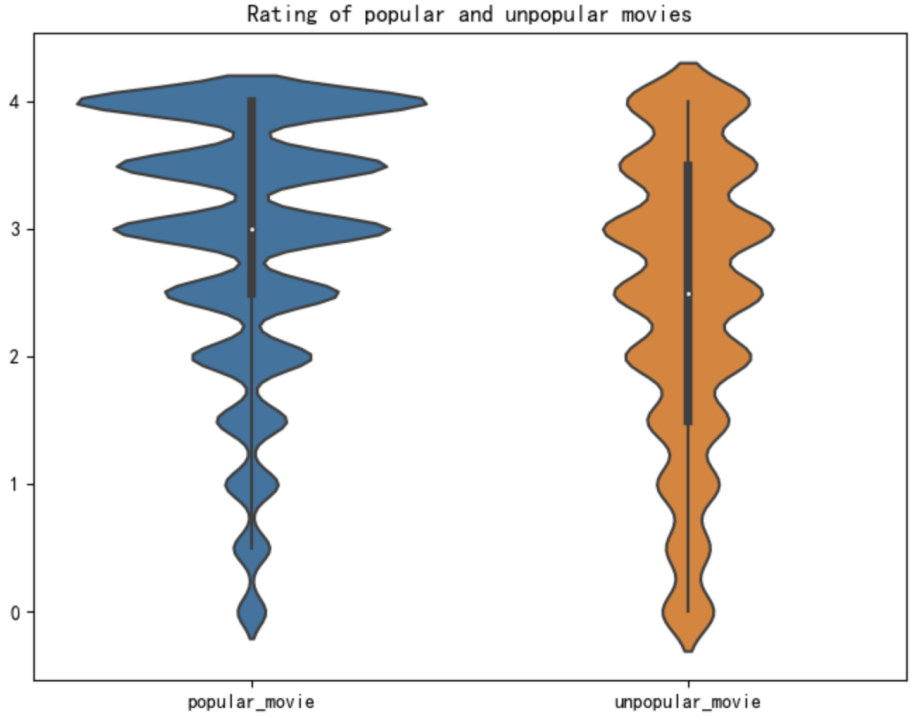
First, we counted the number of users who rated each movie to determine the popularity of the movie (the more users rate a movie, the more popular it is). We then sorted the movies based on the number of rating users from high to low and split them into two groups at the **median(197.5)**: a popular group and an unpopular group.

Next, we assumed that each user's rating for each movie is independent. So, we **flattened the movie matrix (n users * m movies) of each group into a vector (n*m)**. We chose to **drop all missing values** (i.e., when a user didn't rate a certain movie).

Afterward, we use Levene's test to derive that the variance between the two groups is significantly different. So we performed a Welch's t-test between the vector of the popular group and the vector of the unpopular group with
the null hypothesis H_0 = popular movie rating < unpopular movie rating,
the alternative hypothesis H_a = popular movie rating >= unpopular movie rating.

The resulting p-value=0 was less than the significance level alpha=0.005. Hence, we rejected the null hypothesis "popular movie rating < unpopular movie rating", concluding that the rating of popular movies >= the rating of unpopular movies.

Finally, we drew a butterfly chart based on the vectors of popular and unpopular movies, and it clearly shows that the rating of popular movies is higher than that of unpopular movies.



Question 2:

Based on our p-value=0.0024698 being less than the significance level (0.005), we conclude that **newer movie ratings are different from older movie ratings**.

First, we used the **re package in Python** to perform regular expression matching for years on all movie names in the column. After matching, we extracted the year for each movie and placed them in a list. **In the list, we sorted the years from the newest to the oldest, and we split the movie years at the median(year 1999) into a 'new group' and an 'old group'.**

Subsequently, we assumed that each user's rating for each movie is independent so we flattened the movie matrix (n users * m movies) of each group into a vector (n*m) and dropped all missing values.

Then, we use Levene's Test to derive that the variance between the two groups is not significantly different. So a t-test was performed between the vector of the new group and the vector of the old group:
H_0 = newer movies rating = old movies rating
H_a = newer movies rating != old movies rating

Resulting in a **two-sided p-value of 0.0024698 < 0.005**. Therefore, we rejected the null hypothesis, concluding that there is a difference in ratings between new and old movies.

Question 3:

Based on our p-value=0.27 being larger than the significance level (0.005), we **cannot conclude that there is a difference between male and female ratings in the movie ”Shrek (2001)”.**

We split the users based on gender, split into “male group” and “female group”.

Subsequently, we extract the column movie “Shrek (2001)” for two groups as two vectors.

Then, we use Levene's Test to derive that the variance between the two groups is not significantly different. So a t-test was performed between the vector of the new group and the vector of the old group:
H_0 = male rating for “Shrek (2001)” = female rating for “Shrek (2001)”
H_a = male rating for “Shrek (2001)” != female rating for “Shrek (2001)”

Resulting in a **two-sided p-value of 0.27 > 0.005**. Therefore, we conclude nothing about whether there is a difference between male and female ratings in the movie ”Shrek (2001)”.

Question 4:

We iterated through each movie and then applied the method from question 3 to analyze the ratings by men and women for each movie. We counted the number of movies where the ratings between men and women were significantly different. Ultimately, we divided this count(46) by the total number of movies(400) to get the proportion, which is 0.115.

Question 5:

Based on our p-value=0.02 being larger than the significance level (0.005), **we cannot conclude that the rating of people who are only children is higher than the rating of people with siblings rating in the movie “The Lion King (1994)”.**

We split the users based on if the people have siblings, split into “people who are only children” and “people who have siblings”.

Subsequently, we extract the column movie “The Lion King (1994)” for two groups as two vectors.

Then, we use Levene's Test to derive that the variance between the two groups are not significantly different. So a t-test was performed between the vector of the new group and the vector of the old group:

H_0 = people who is only child rating for “The Lion King (1994)” > people who have siblings rating for “The Lion King (1994)”

H_a = people who is only child rating for “The Lion King (1994)” <= people who have siblings rating for “The Lion King (1994)”

Resulting in a **two-sided p-value of 0.02 > 0.005**. Therefore, we cannot conclude that the rating of people who are only children is higher than the rating of people with siblings rating in the movie “The Lion King (1994)”.

Question 6:

We iterated through each movie and then applied the method from question 5 to analyze the ratings by people who are only children and people who have siblings for each movie. I counted the number of movies where the ratings between the two groups were significantly different. Ultimately, we divided this count(10) by the total number of movies(400) to get the proportion, which is 0.025.

Question 7:

Based on our p-value=0.0587 being larger than the significance level (0.005), **we cannot conclude that people who like to watch movies socially enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone.**

We divided the data into two groups based on their preference for watching movies: Group 1 (those who enjoy watching movies socially) and Group 2 (those who prefer watching alone). We then extracted the ratings for 'The Wolf of Wall Street (2013)' and conducted a statistical t-test to compare these ratings.

The division of respondents was necessary to distinctly evaluate the preferences of the two groups in question. A t-test was chosen because it's a standard method to determine if there's a statistically significant difference in means (ratings, in this case) between two independent groups.

The result of the t-test gave a p-value of approximately 0.0587. This number signifies the probability of observing the data, or something more extreme if there truly is no difference in movie ratings between the two groups.

H_0 = people who like to watch movies socially "The Wolf of Wall Street (2013)" < people who prefer to watch "The Wolf of Wall Street (2013)" alone

H_a = people who like to watch movies socially "The Wolf of Wall Street (2013)" >= people who prefer to watch "The Wolf of Wall Street (2013)" alone

Given that the p-value is approximately 0.0587, which is above the standard threshold of 0.05, we do not have enough evidence to conclusively say that one group enjoys 'The Wolf of Wall Street (2013)' more than the other. While there may be a difference in ratings, it's not statistically significant at the 0.005 level.

Question 8:

We iterated through each movie and then applied the method from question 5 to analyze the ratings by people who like to watch movies socially and people who prefer to watch movies alone for each movie. I counted the number of movies where the ratings between the two groups were significantly different. Ultimately, we divided this count(12) by the total number of movies(400) to get the proportion, which is 0.03.

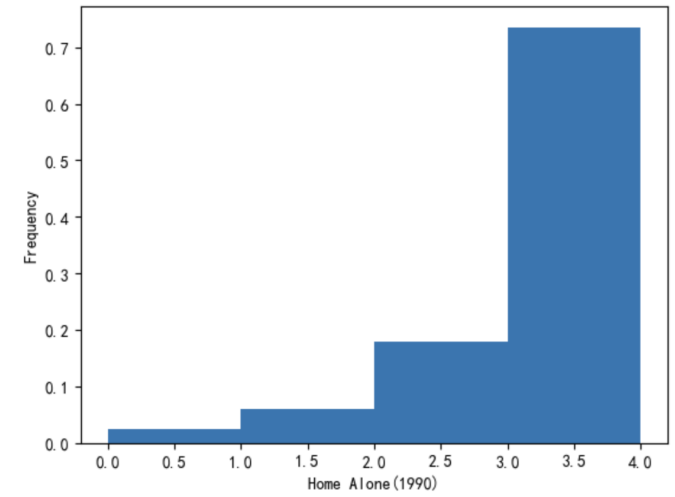
Question 9:

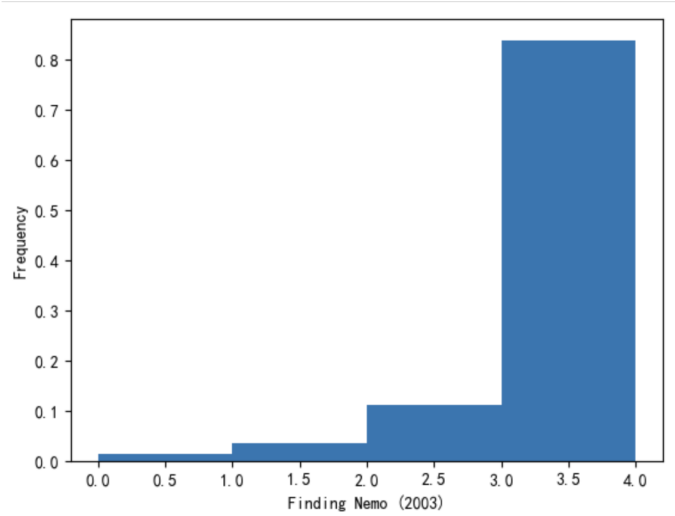
Based on our p-value = 6.379e-10 < 0.005, we can conclude that the distribution for ‘Home Alone (1990)’ and ‘Finding Nemo (2003)’ are different.

We assume the ratings of “Home Alone(1990)” and “Finding Nemo(2003)” are independent and both samples should be random and representative selections from their respective populations.

Because we have to compare two distributions of ratings of movies, we choose the KS test as our method. It does not compare means or medians, but simply compares the cumulative distribution function of their samples, which can better reflect the difference of two distributions.

Given the p_value=6.379e-10 from KS test, which is less than 0.005. We can conclude that the distribution for ‘Home Alone (1990)’ and ‘Finding Nemo (2003)’ are different.





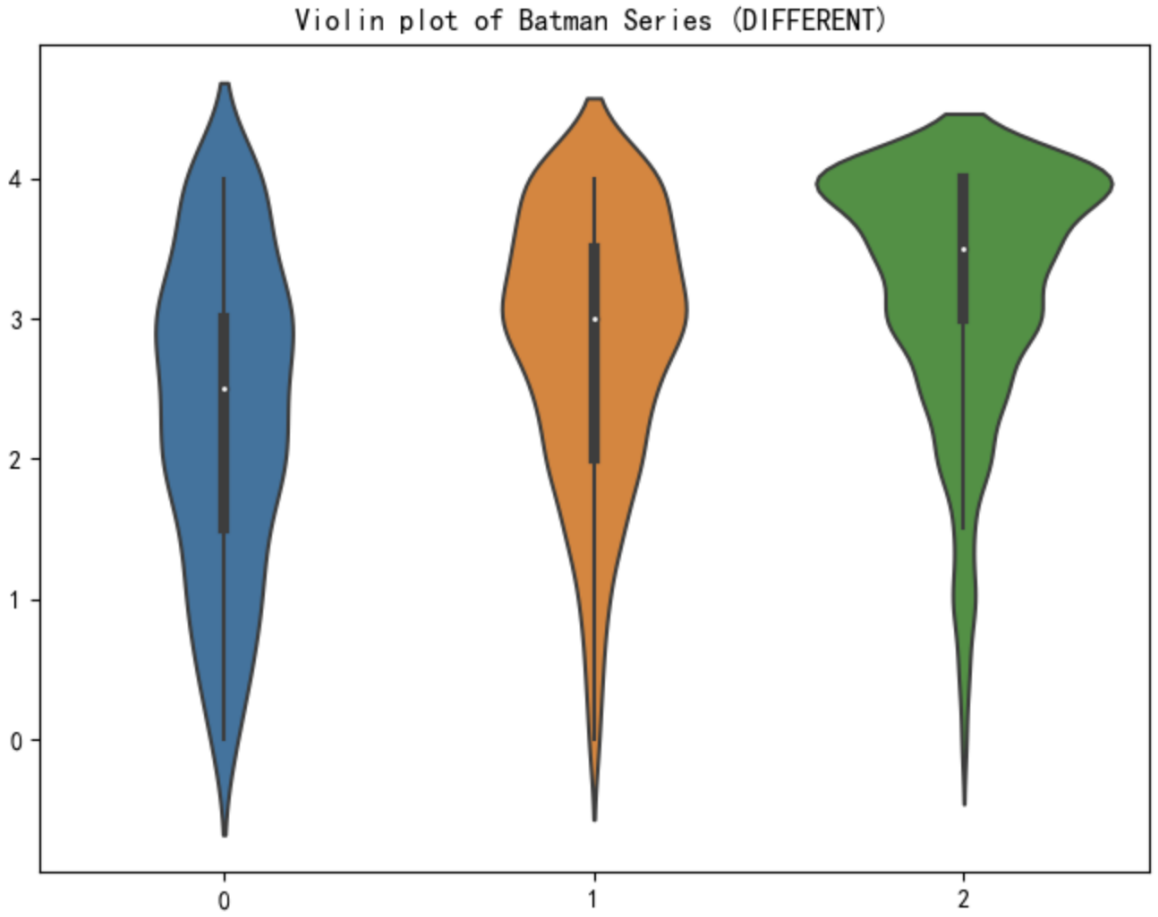
Question 10:

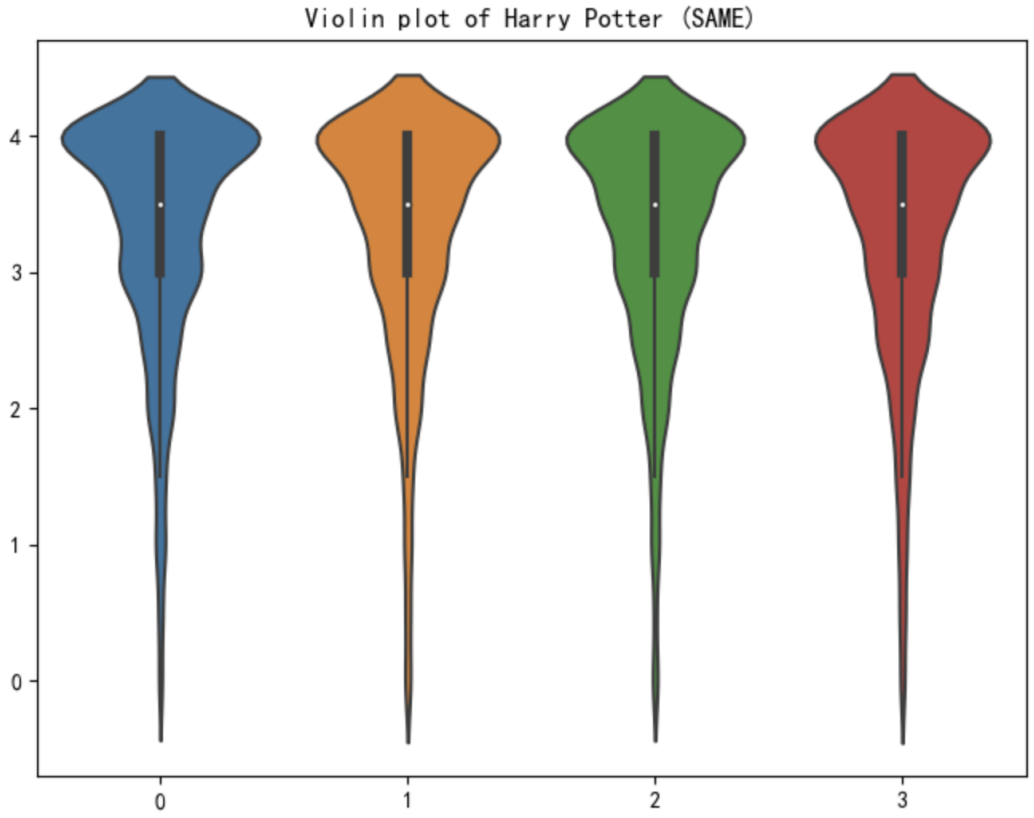
We assume that in every franchise the variances of ratings of the movies are similar and the ratings are independent.

Because there are more than 2 movies in one franchise, we choose Anova as our method to compare the differences, which actually extends the logic of t-test to more than 2 groups but it can simplify the calculation process by aggregating the comparisons between any 2 groups of the sample to just one test.

We use keywords of the movie's name to find the movie of the same franchise and use a dictionary to store them. Then in every franchise, we use Anova to check if the quality of the movies in the same series keeps consistent. After experiments, only the movies of "Harry Potter" keep almost the same quality in the whole series.

In conclusion, there are 7 movies that are of inconsistent quality: "Star Wars", "The Matrix", "Indiana Jones", "Jurassic Park", "Pirates of the Caribbean", "Toy Story", "Batman".




Violin plot of Harry Potter (SAME)
4
3
2
1
0
0
1
2
3