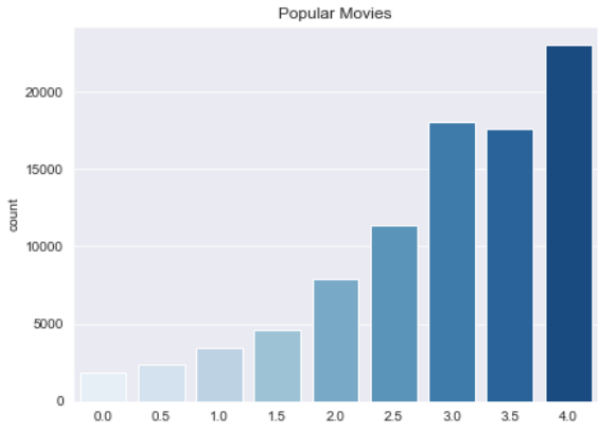
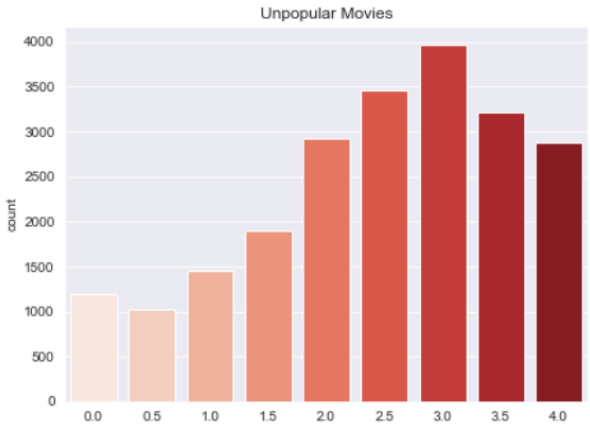
Data analysis project 1

1) Are movies that are more popular (operationalized as having more ratings) rated higher than movies that are less popular?

The first step to solve this question is to divide the data into popular and unpopular movies. To do that, I found the median number of ratings per movie, which was 197.5 ratings, and assigned all movies with less than 197.5 ratings into the unpopular movies dataset and the movies with more than 197.5 ratings into the popular movies dataset.





Now that we have the data ready, we must decide which significance test to perform.

A z-test is not appropriate because we do not have the population parameters. A t-test is not appropriate either because the data consists of ratings and therefore the difference between ratings is psychologically different (0 to 0.5 is not the same distance as 4.5 to 5). KS-test is not appropriate as we are not interested in the samples' distributions. Lastly, the Mann Whitney U-test is appropriate as comparing the medians would show which sample is rated higher.

Let's state our null and alternative hypothesis and perform the significance test.

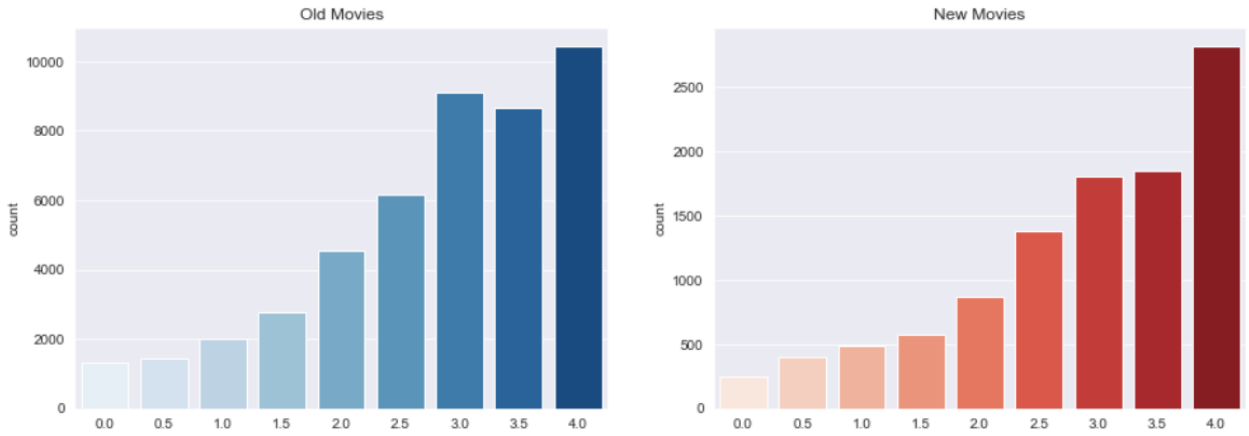
H_0: Popular movies are not rated higher than unpopular movies.
H_1: Popular movies are rated higher than unpopular movies.

The significance test outputs a p-value of $1.23*10^{-15}$ therefore we reject the null hypothesis and conclude that popular movies are rated higher than unpopular movies.

2) Are movies that are newer rated differently than movies that are older?

First, in order to create the two samples of old and new movies, I extracted the release year of every movie and determined the median year of the dataset, which happened to be 1999. Then, the median was used to divide the initial dataset into old and new movies.

It is worth mentioning that the movies released in 1999 (the cutoff) where assigned to the new movies dataset in order to have a smaller difference between sample sizes.



Now that we have the data ready, we must decide which significance test to perform.

Again, a z-test and t-test would not be appropriate because of the reasons mentioned in the previous question. But in this case, where we are interested in the difference of ratings between the two samples, we have several options. A chi-squared test would be appropriate but we would need the exact same sample size in both samples, so we won't use it. A KS-test would be appropriate as we are interested in comparing the underlying distributions of our samples and also a Mann-Whitney could be used (even though it will not be as reliable) if a two tailed model is used.

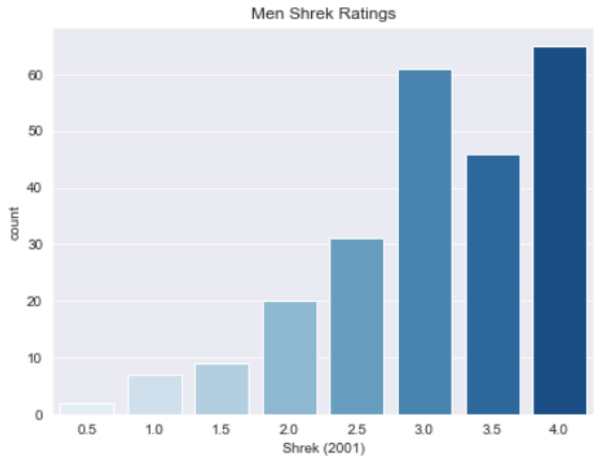
Let's state our null and alternative hypothesis and perform the significance test.

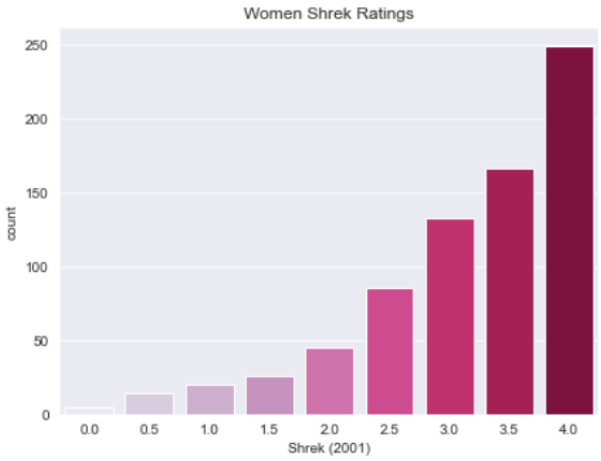
H_0: There is no difference between the ranks of new and old movies.
H_1: There is a difference between the ranks of new and old movies.

The Mann Whitney U-test outputs a p-value of $1.28*10^{-6}$ and the KS-test outputs a p-value of 0.002 therefore we reject the null hypothesis and conclude that there is difference between the ranks of old and new movies.

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

To answer this question, the initial dataset was divided into ranks by male viewers and ranks by female users, specifically the ranks for the movie Shrek.





Now that we have the data ready, we must decide which significance test to perform.

This problem is similar to question 2 as we must find if the difference between two samples is statistically significant, so the same rules to select a significance test apply.

A KS-test would be appropriate as we are interested in comparing the underlying distributions of our samples and also a Mann-Whitney could be used (with less reliability) if a two tailed model is used.

Let's state our null and alternative hypothesis and perform the significance test.

H_0: There is no difference between the way men and women rate the movie Shrek.
H_1: There is a difference between the way men and women rate the movie Shrek.

The Mann Whitney U-test outputs a p-value of 0.051 and the KS-test outputs a p-value of 0.056 therefore we fail to reject the null hypothesis and conclude that there is no difference between the way men and women rate the movie Shrek.

4) What proportion of movies are rated differently by male and female viewers?

To answer this question, we must loop over all movies, divide it into two samples of male and female viewers and perform a KS-test (and Mann Whitney U-test) to check whether they are statistically different or not. During this loop, we must keep track of the number of movies whose ratings between men and women are indeed statistically different and output its proportion.

```
titles = movies.columns.tolist()
significant, significant2 = 0, 0
for title in titles:
    movie_genders = df.iloc[:,[df.columns.get_loc(title),474]].dropna()
    movie_men = movie_genders[movie_genders['Gender identity (1 = female; 2 = male; 3 = self-described)'] == 2.0][title]
    movie_women = movie_genders[movie_genders['Gender identity (1 = female; 2 = male; 3 = self-described)'] == 1.0][title]
    test4 = kstest(movie_men, movie_women, alternative='two-sided')
    if test4.pvalue < 0.005:
        significant = significant + 1
    test4 = mannwhitneyu(movie_men, movie_women, alternative='two-sided')
    if test4.pvalue < 0.005:
        significant2 = significant2 + 1

print('The proportion of movies rated differently by male and female viewers is {}% (based on KS-test)'
      .format(100 * significant/len(titles)))
print('The proportion of movies rated differently by male and female viewers is {}% (based on Mannwhitney U-test)'
      .format(100 * significant2/len(titles)))
```

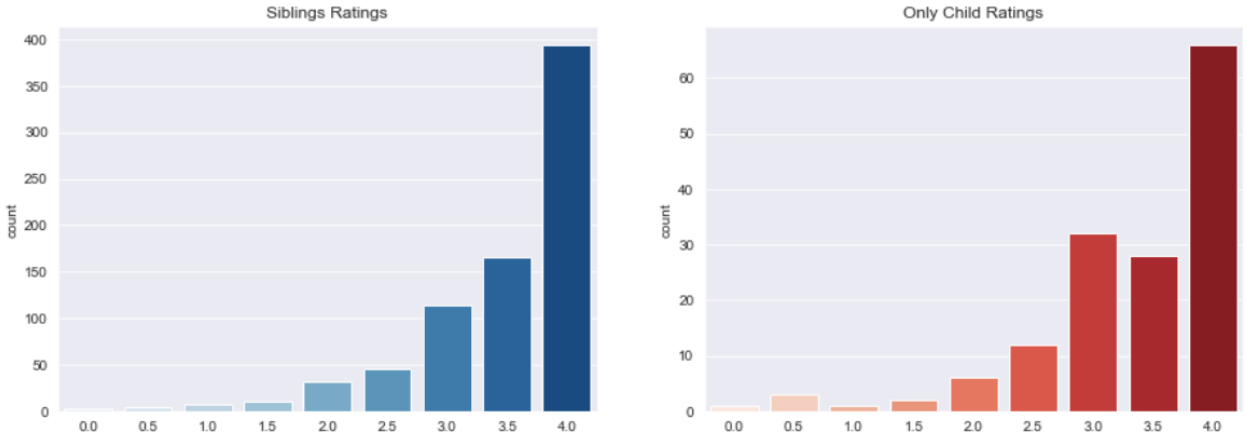
```
The proportion of movies rated differently by male and female viewers is 6.25% (based on KS-test)
The proportion of movies rated differently by male and female viewers is 12.5% (based on Mannwhitney U-test)
```

As we can see in the output of the code, the proportion of movies rated differently by male and female users is 6.5% according to the KS-test and 12.5% according to the Mann Whitney test.

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

The first step to solve this question is to extract from the data the ratings for the movie The Lion King divided into viewers that are only childs and viewers that have siblings.

I noticed that the data is highly unbalanced, with 177 viewers that are only children and 894 viewers with siblings.



Now that we have the data ready, we must decide which significance test to perform.

This question is similar to question 1 as we must find if a sample has higher ratings than another sample, therefore the criteria to select a significance test is similar.

We rule out KS-test because we are not interested in comparing the general distribution of the data, also Chi-squared because the samples do not have equal size. The appropriate test would be Mann Whitney because comparing its medians would be a fair method to determine if one has higher ratings than the other.

Let's state our null and alternative hypothesis and perform the significance test.

H_0: Only child viewers do not enjoy The Lion King more than viewers with siblings.
H_1: Only child viewers enjoy The Lion King more than viewers with siblings.

The significance test outputs a p-value of 0.978 therefore we fail to reject the null hypothesis and conclude that only child viewers do not enjoy The Lion King more than viewers with siblings.

6) What proportion of movies exhibit an “only child effect”, i.e. are rated different by viewers with siblings vs. those without?

To answer this question, we must loop over all movies, divide it into two samples of only child viewers and viewers with siblings and perform a KS-test (and Mann Whitney U-test) to check whether they are statistically different or not. During this loop, we must keep track of the number of movies whose ratings between only children and viewers with siblings are indeed statistically different and output its proportion.

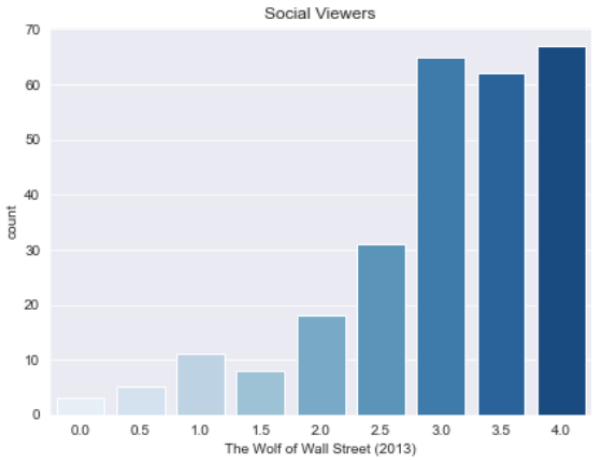
```
titles = movies.columns.tolist()
significant, significant2 = 0, 0
for title in titles:
    movie_family = df.iloc[:,[df.columns.get_loc(title),475]].dropna()
    movie_onlychild = movie_family[movie_family['Are you an only child? (1: Yes; 0: No; -1: Did not respond)'] == 1][title]
    movie_siblings = movie_family[movie_family['Are you an only child? (1: Yes; 0: No; -1: Did not respond)'] == 0][title]
    test6 = kstest(movie_onlychild, movie_siblings, alternative='two-sided')
    if test6.pvalue < 0.005:
        significant = significant + 1
    test6 = mannwhitneyu(movie_onlychild, movie_siblings, alternative='two-sided')
    if test6.pvalue < 0.005:
        significant2 = significant2 + 1

print('The proportion of movies that exhibit an only child effect is {}% (based on KS-test)'
      .format(100 * significant/len(titles)))
print('The proportion of movies that exhibit an only child effect is {}% (based on Mann Whitney U-test)'
      .format(100 * significant2/len(titles)))
```

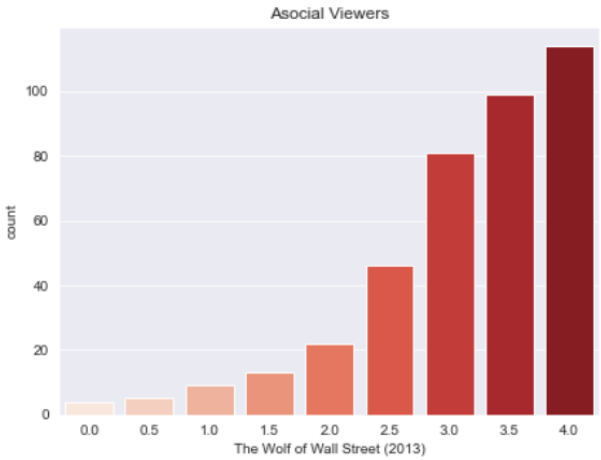
```
The proportion of movies that exhibit an only child effect is 0.75% (based on KS-test)
The proportion of movies that exhibit an only child effect is 1.75% (based on Mann Whitney U-test)
```

As we can see in the output of the code, the proportion of movies rated differently by only children and viewers with siblings is 0.75% according to the KS-test and 1.75% according to the Mann Whitney test.

7) Do people who like to watch movies socially enjoy 'The Wolf of Wall Street (2013)' more than those who prefer to watch them alone?

The first step to solve this question is to extract from the initial dataset the ratings for the movie The Wolf of Wall Steet divided into viewers that enjoy watching movies by themselves and viewers that prefer watching movies with other people.





Now that we have the data ready, we must decide which significance test to perform.

This question is similar to question 1 as we must find if a sample has higher ratings than another sample, therefore the criteria to select a significance test is similar.

We rule out KS-test because we are not interested in comparing the general distribution of the data, also Chi-squared because the samples do not have equal size. The appropriate test would be Mann Whitney because comparing its medians would be a fair method to determine if one has higher ratings than the other.

Let's state our null and alternative hypothesis and perform the significance test.

H_0: Viewers who like to watch movies socially do not enjoy the Wolf of Wall Street movie more than viewers who prefer to eatch movies alone.
H_1: Viewers who like to watch movies socially enjoy the Wolf of Wall Street movie more than viewers who prefer to eatch movies alone.

The significance test outputs a p-value of 0.944 therefore we fail to reject the null hypothesis and conclude that viewers who like watching movies socially do not enjoy the Wolf of Wall Street movie more than viewers who prefer to watch movies alone.

8) What proportion of movies exhibit such a “social watching” effect?

To answer this question we must loop over all movies, divide it into two samples of social and asocial viewers and perform a KS-test (and Mannwhitney U-test) to check whether they are statistically different or not. During this loop, we must keep track of the number of movies whose ratings between social and asocial viewers are indeeed statistically different and output its proportion.

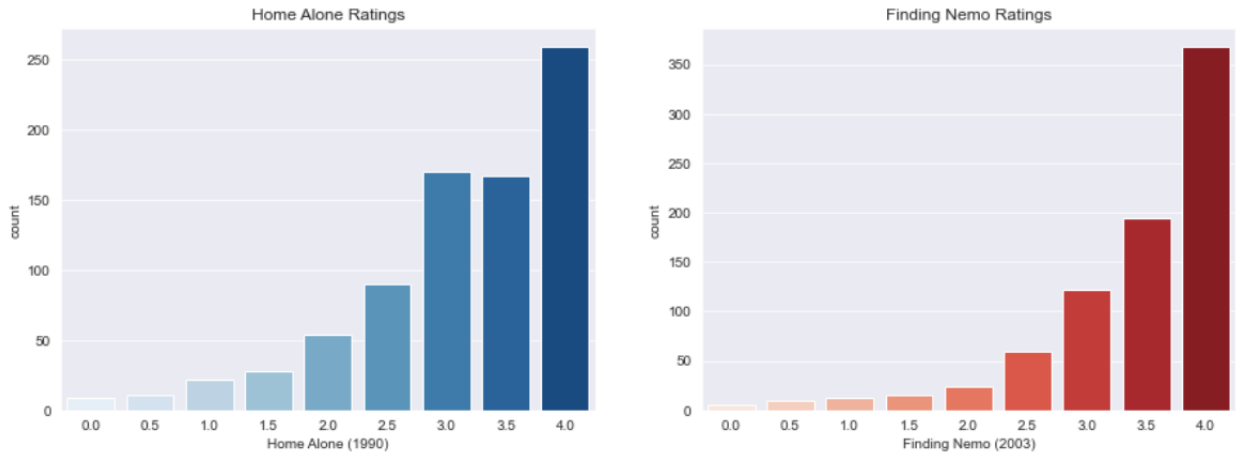
```
titles = movies.columns.tolist()
significant, significant2 = 0, 0
for title in titles:
    movies_ = df.iloc[:,[df.columns.get_loc(title),476]].dropna()
    movie_social = movies_[movies_['Movies are best enjoyed alone (1: Yes; 0: No; -1: Did not respond)'] == 0.0][title]
    movie_asocial = movies_[movies_['Movies are best enjoyed alone (1: Yes; 0: No; -1: Did not respond)'] == 1.0][title]
    test6 = kstest(movie_social, movie_asocial, alternative='two-sided')
    if test6.pvalue < 0.005:
        significant = significant + 1
    test6 = mannwhitneyu(movie_social, movie_asocial, alternative='two-sided')
    if test6.pvalue < 0.005:
        significant2 = significant2 + 1

print('The proportion of movies that exhibit a social watching effect is {}% (based on KS-test)'
      .format(100 * significant/len(titles)))
print('The proportion of movies that exhibit a social watching effect is {}% (based on Mann Whitney U-test)'
      .format(100 * significant2/len(titles)))
```

```
The proportion of movies that exhibit a social watching effect is 0.5% (based on KS-test)
The proportion of movies that exhibit a social watching effect is 2.5% (based on Mann Whitney U-test)
```

As we can see in the output of the code, the proportion of movies rated differently by social viewers compared to asocial viewers is 0.5% according to the KS-test and 2.5% according to the Mann Whitney test.

9) Is the ratings distribution of 'Home Alone (1990)' different than that of 'Finding Nemo (2003)'?

To answer this question, I extracted the rating for both the movie of Home Alone and the movie of Finding Nemo into two separate arrays of samples.



Now that we have the data ready, we must decide which significance test to perform.

This problem is similar to question 2 as we must find if the difference between two samples is statistically significant, so the same rules to select a significance test apply.

A KS-test would be appropriate as we are interested in comparing the underlying distributions of our samples and also a Mann-Whitney could be used (with less reliability) if a two tailed model is used.

Let's state our null and alternative hypothesis and perform the significance test.

H_0: The movie Home Alone has the same distribution compared to the movie Finding Nemo.
H_1: The movie Home Alone has a different distribution compared to the movie Finding Nemo.

The Mann Whitney U-test outputs a p-value of $2.44*10^{-12}$ and the KS-test outputs a p-value of $2.2*10^{-10}$ therefore we fail to reject the null hypothesis and conclude that there is no difference between the distributions of Finding Nemo and Home Alone.

10) 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?

To solve this question, we must first organize the data. In order to do this, I created a dictionary with each franchise as the keys and the movies of those franchises as the keys' values.

Once the dictionary was created, I crated a loop where each franchise was analyzed for consistency. During the loop, all the movies for a given franchise were compared to each other using a KS-test to determine if their underlying distributions were different. Here is the code with its output:

```
for franchise in franch_dict:
    significant = 0
    for title1 in franch_dict[franchise]:
        for title2 in franch_dict[franchise]:
            if title1 == title2:
                continue
            else:
                test10 = kstest(franch_dict[franchise][title1], franch_dict[franchise][title2], alternative='two-sided')
                if test10.pvalue < 0.005:
                    significant = significant + 1

    if significant == 0: print('{} is of consistent quality!'.format(franchise))
    else: print('{} is of inconsistent quality because {} pairs movies are significantly different from each other.'
                .format(franchise, int(significant/2)))
```

```
Star Wars is of inconsistent quality because 8 pairs movies are significantly different from each other.
Harry Potter is of consistent quality!
The Matrix is of inconsistent quality because 2 pairs movies are significantly different from each other.
Indiana Jones is of inconsistent quality because 4 pairs movies are significantly different from each other.
Jurassic Park is of inconsistent quality because 3 pairs movies are significantly different from each other.
Pirates of the Caribbean is of consistent quality!
Toy Story is of inconsistent quality because 1 pairs movies are significantly different from each other.
Batman is of inconsistent quality because 3 pairs movies are significantly different from each other.
```

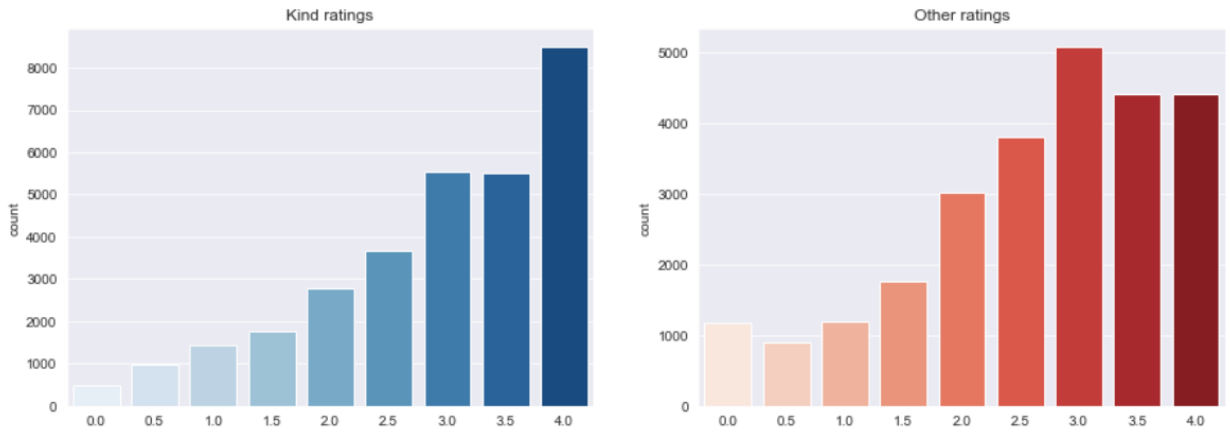
If a franchise had a pair of movies that was statistically different based on the KS-test, the whole franchise was categorized as of inconsistent quality. Therefore, all franchises were considered of inconsistent quality but two: Harry Potter and Pirates of the Caribbean.

Extra Credit: Tell us something interesting and true (supported by a significance test of some kind) about the movies in this dataset that is not already covered by the questions above [for 5% of the grade score].

After looking into the several personal questions in the dataset, I found this one: “Is considerate and kind to almost everyone” and came up with the following Null and Alternate hypothesis:

H_0: Viewers considered kind to almost everyone do not give higher ratings than other viewers
H_1: Viewers considered kind to almost everyone give higher ratings than other viewers

The first step is to extract from the initial dataset all the ratings from both “kind” viewers and the other viewers and store them in two separate arrays.



Now, the most appropriate significance test for this problem is Mann Whitney U-test (less reliable).

The Mann Whitney U-test outputs a p-value of $1.816*10^{-205}$ therefore we reject the null hypothesis and conclude that viewers considered kind to almost everyone give higher ratings than other viewers.