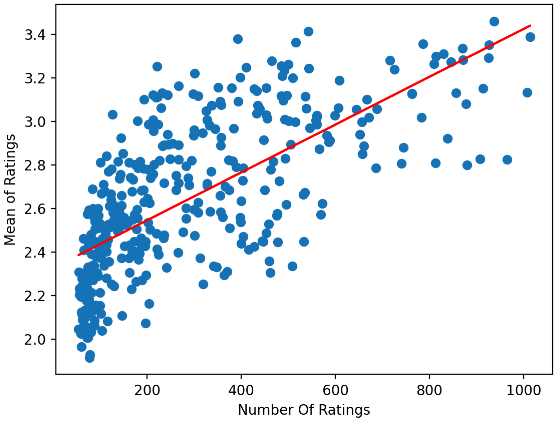
Data Analysis Project 1

1.

- We find that median number of ratings for our Movies is 197.5.
- So we split the movie into two groups; the popular movies (number of ratings greater than 197.5) and the less popular movies (number of ratings less than 197.5).
- We have a clean split with both groups having the same number of movies(200) but the number of ratings in Popular movies is 90214 and in the less Popular 22000.
- The Average rating for Popular movies is 2.92348 and the standard deviation is 1.01554. The Average rating for less Popular meetings is 2.45061 and the standard deviation is 1.12973.
- We can see that the two groups have different variance so we will employ Welch's t test for a one sided test. Our Null Hypothesis is that the difference of the mean between the group is zero. We are testing whether Popular movies have higher ratings than less popular movies.
- We find our test statistic, t = 56.74324, df = 31227
- Our p-value is incredibly small (< .00001)
- We can determine that our result here is significant using an alpha = .005
- We also find that the Correlation between the Number of Ratings a movie has and its average rating is 0.69916
- We also run a linear regression and get r^2 = 0.48883 (r = 0.69916 which matches our correlation ran with Pandas)

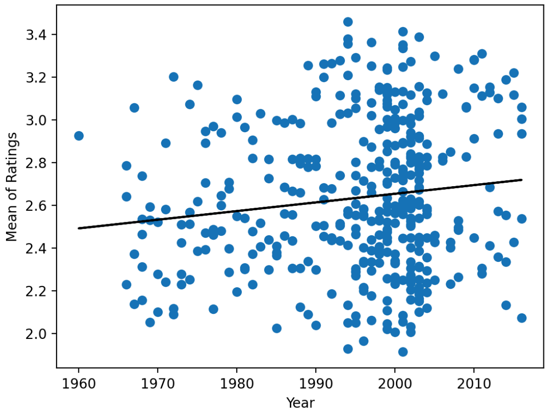


- We can use both our hypothesis test, correlation, and r^2 and to assume that popular movies have a higher rating
- The code for this starts at line 30

2.

- We find that the median year is 1999.

- We will split the data into two group; Older movies (before 1999) and Younger movies(after or in 1999)
- We have 57266 ratings for the older movies and 54948 ratings for the younger movies
- We calculate the Average of the Older movies to be 2.84011 and for Younger movies to be 2.82103, Our Older std is 1.05269 and the younger one is 1.05883. Our Variance seem very close to each other so we will assume that the Variances are the same and we will use a two sided Student's t-test. Our Null Hypothesis is that they have the same Mean.
- Our test statistic is 3.02708 which has a corresponding p-value of 0.00247 which is less than .0025, therefore since our p-value is less than our alpha we reject our null hypothesis
- While our Two means not might be exactly equal we can also deduce that they are very similar. I calculated the correlation between the Year the movie is released and average ratings of each movie, our Corr is 0.13115.
- We run a linear regression and r^2 = 0.01720 which shows very little linear dependency between our sets. Our linear regression also returns us a p-value for a Wald Test with t-distribution. The Null Hypothesis is that the Slope is 0. The p-value is 0.00864, with our alpha being .005, we fail to reject that hypothesis. Which means that the slight slope we do see is more likely (but not certainly) due to chance and there is no linear dependency.



- I conclude that there is a slight difference in rating for movies that are older vs younger our hypothesis test shows that this is statistically significant. But the difference (0.01908) is very slight.
- The code for this starts at line 73

3.

- We have 743 female ratings with an Average rating of 3.15545 and a Standard Deviation of 0.90655.
- We have 241 male ratings with an Average rating of 3.08299 and a Standard Deviation of 0.82498.
- We run a two sided t-test to see if the Average rating between males and females is different. Our Null Hypothesis is that they are the same. We calculate our test statistic to be

1.10167 with a p value 0.27088. This p-value is very much greater than our alpha = .0025. We fail to reject our hypothesis

- I conclude that we may need more male ratings for Shrek (2001) but with our current data we can not say that the genders rated the movie differently
- The code for this starts at line 123

4.

- We First calculate the mean and std by gender for each movie. We then run a two sided t test for each movie. Our Null Hypothesis is that the mean rating for each gender is the same.
- We calculated all the p values, found that only 8.5% of the movies were rated significantly different by gender. We failed to reject our hypothesis the other 91.5% of the time.
- The code starts at line 139

5.

- We calculate the mean rating and std for both only-children and those with siblings for Lion King. We are running a two sided t test with the Null hypothesis that the Mean rating for both groups is the same.
- We got a t statistic of 1.884 and a p value of 0.061 > .005 Therefore we fail to reject that these groups have equal means.
- The code starts at line 156

6.

- We First calculate the mean and std by only-child or not for each movie. We then run a two sided t test for each movie. Our Null Hypothesis is that the mean rating for both groups is the same.
- We calculated all the p values and found that only .5% of movies were rated significantly different. The other 99.5% of movies we failed to reject that there is not an “only child effect”.
- The code starts 171

7.

- We calculate the mean rating and std for both enjoy watching movies alone and those that don’t for Wolf of Wall Street. We are running a two sided t test with the Null hypothesis that the Mean rating for both groups is the same.
- We got a t statistic of -1.551 and a p value of 0.1214 > .005 Therefore we fail to reject that these groups have equal means.
- The code starts at line 189

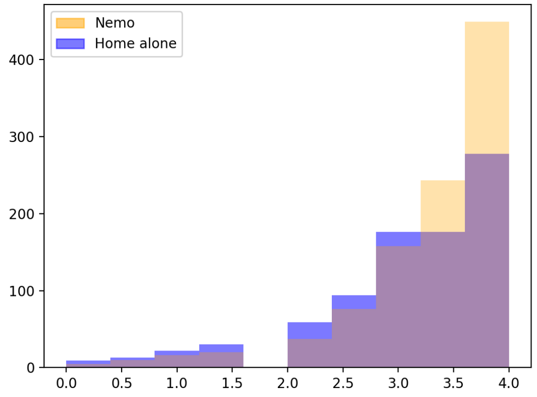
8.

- We First calculate the mean and std by whether they enjoy watching alone or not for each movie. We then run a two sided t test for each movie. Our Null Hypothesis is that the mean rating for both groups is the same.

- We calculated all the p values and found that only 0.5% of movies were rated significantly different. The other 99.5% of movies we failed to reject that there is not an “social watch effect”.
- The code starts at line 206

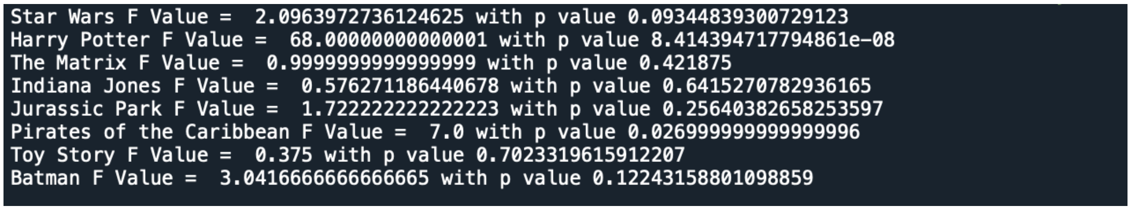
9.

- We first want to see what the histogram of the distribution look like to get an idea. They seem similar but Nemo has more ratings (1014 to be exact) compared to Home Alone (857).
- We will run a two sided discrete Kolmogorov-Smirnov test (KS). The KS test compares the empirical cdf of both samples and determines its KS statistics and will return a p-value associated with that.
- Our KS test statistics is 0.14311 with p value 3.26265e-10 < .0025. We strongly reject the hypothesis that these two movies have the same rating distribution.



10.

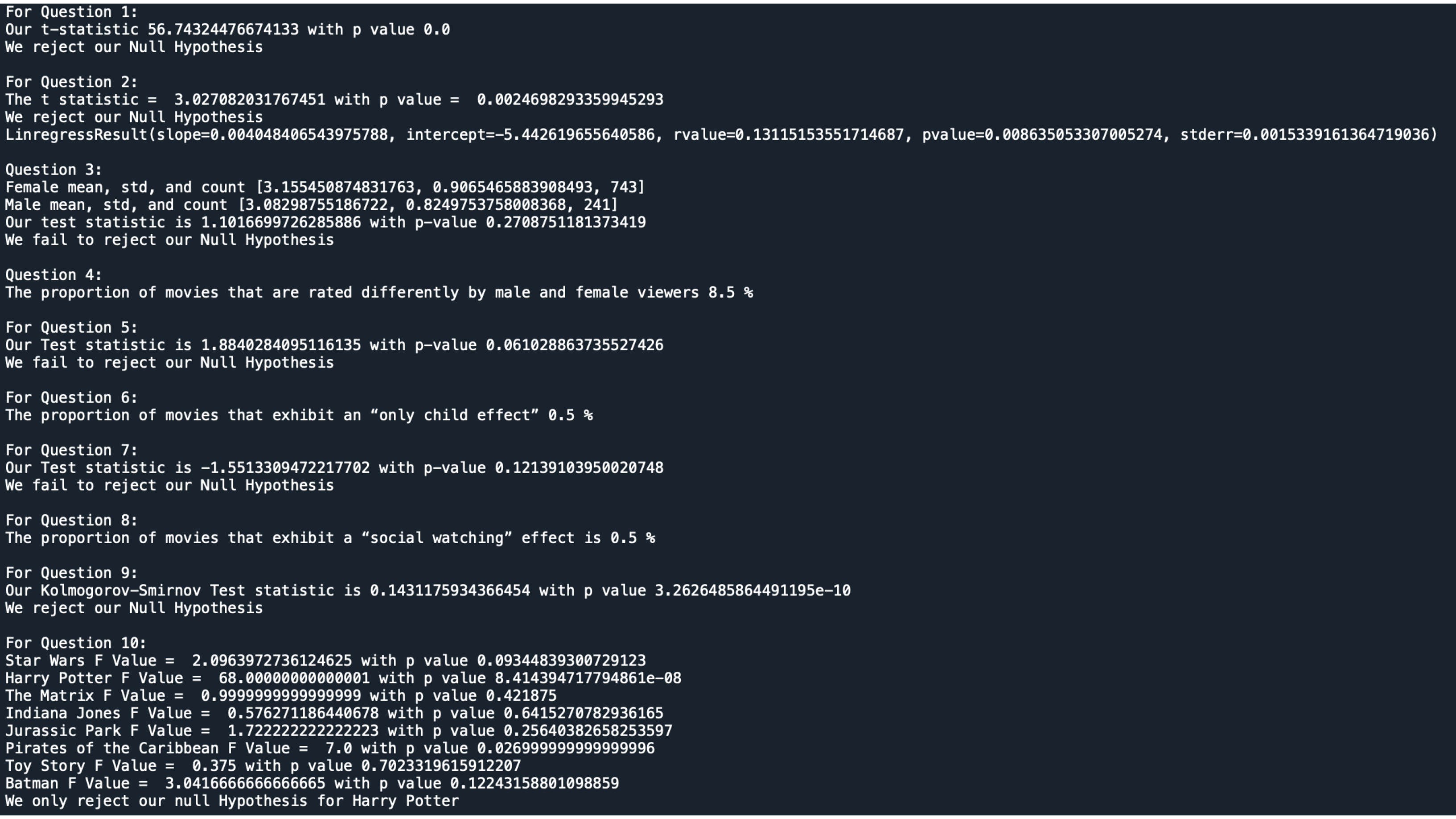
- We want to find if movies in the same Franchise are of consistent quality in the eyes of our consumers. We can run a anova test with a null hypothesis that each movie has an equal mean rating.
- We run our anova testing and we get:



```
Star Wars F Value =  2.0963972736124625 with p value 0.09344839300729123
Harry Potter F Value =  68.00000000000001 with p value 8.414394717794861e-08
The Matrix F Value =  0.9999999999999999 with p value 0.421875
Indiana Jones F Value =  0.576271186440678 with p value 0.6415270782936165
Jurassic Park F Value =  1.722222222222223 with p value 0.25640382658253597
Pirates of the Caribbean F Value =  7.0 with p value 0.026999999999999996
Toy Story F Value =  0.375 with p value 0.7023319615912207
Batman F Value =  3.0416666666666665 with p value 0.12243158801098859
```

- We can observe that the only Franchise where we fail to reject that each movie has the same mean is Harry Potter.

This the exact code output:



```
For Question 1:
Our t-statistic 56.74324476674133 with p value 0.0
We reject our Null Hypothesis

For Question 2:
The t statistic =  3.027082031767451 with p value =  0.0024698293359945293
We reject our Null Hypothesis
LinregressResult(slope=0.004048406543975788, intercept=-5.442619655640586, rvalue=0.13115153551714687, pvalue=0.008635053307005274, stderr=0.0015339161364719036)

Question 3:
Female mean, std, and count [3.155450874831763, 0.9065465883908493, 743]
Male mean, std, and count [3.08298755186722, 0.8249753758008368, 241]
Our test statistic is 1.1016699726285886 with p-value 0.2708751181373419
We fail to reject our Null Hypothesis

Question 4:
The proportion of movies that are rated differently by male and female viewers 8.5 %

For Question 5:
Our Test statistic is 1.8840284095116135 with p-value 0.061028863735527426
We fail to reject our Null Hypothesis

For Question 6:
The proportion of movies that exhibit an “only child effect” 0.5 %

For Question 7:
Our Test statistic is -1.5513309472217702 with p-value 0.12139103950020748
We fail to reject our Null Hypothesis

For Question 8:
The proportion of movies that exhibit a “social watching” effect is 0.5 %

For Question 9:
Our Kolmogorov-Smirnov Test statistic is 0.1431175934366454 with p value 3.2626485864491195e-10
We reject our Null Hypothesis

For Question 10:
Star Wars F Value =  2.0963972736124625 with p value 0.09344839300729123
Harry Potter F Value =  68.00000000000001 with p value 8.414394717794861e-08
The Matrix F Value =  0.9999999999999999 with p value 0.421875
Indiana Jones F Value =  0.576271186440678 with p value 0.6415270782936165
Jurassic Park F Value =  1.722222222222223 with p value 0.25640382658253597
Pirates of the Caribbean F Value =  7.0 with p value 0.026999999999999996
Toy Story F Value =  0.375 with p value 0.7023319615912207
Batman F Value =  3.0416666666666665 with p value 0.12243158801098859
We only reject our null Hypothesis for Harry Potter
```