**EMRE OZBALTA**

**00228015**

**MA417 Fall 2019**

**Final Project**

My research questions are:

* Is there a relationship between IMDB ratings (numerical variable) and critics score (numerical variable) on Rotten Tomatoes?
* Is there an association between IMDB ratings (numerical variable) and genres of movies (categorical variable)?

As a result of my first research question, I expect to see positive correlation between two variables that I am going to use because if voters consider a movie as good or bad on IMDB, most of the same voters will vote it on Rotten Tomatoes too.

As a result of my second research question, I expect to see higher IMDB rating on certain genres such as Action & Adventure, Drama and Comedy because people mostly like watching movies with those genres. They are also most popular movie genres by total box office revenue.

I am studying the data of a population that is obtained from IMDB and Rotten Tomatoes.

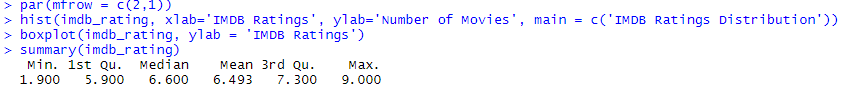
I will use critics\_score: “Critics score on Rotten Tomatoes” (numerical variable) and imdb\_rating: “Ratings on IMDB” (numerical variable) to help answering my first research question. I will use genre: “Genres of movies” (categorical variable) and imdb\_rating: “Ratings on IMDB” (numerical variable) to help answering my second research question.

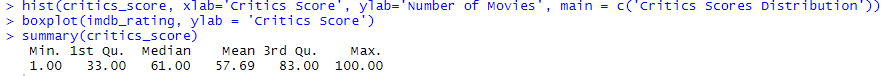
I get my data from <http://www2.stat.duke.edu/~mc301/data/movies.html>.

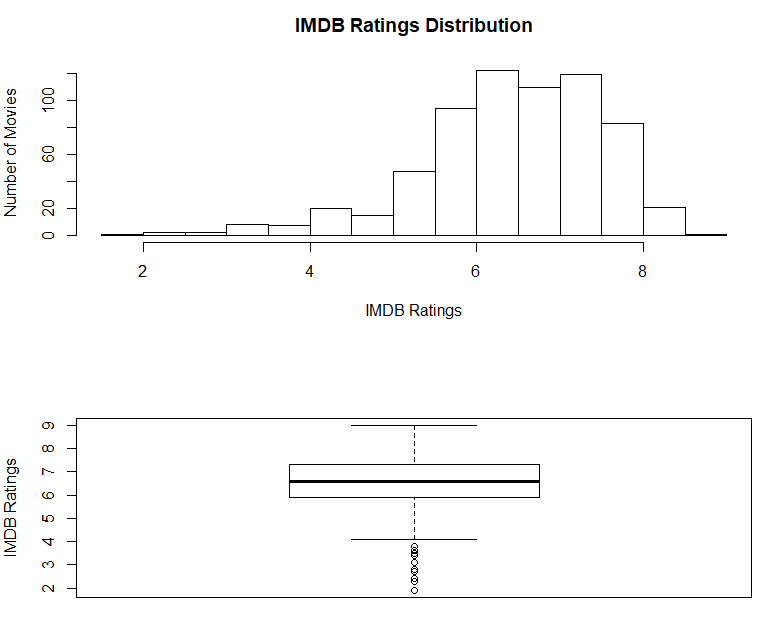
My data represents randomly sampled movies released between 1972 and 2014 in the United States.

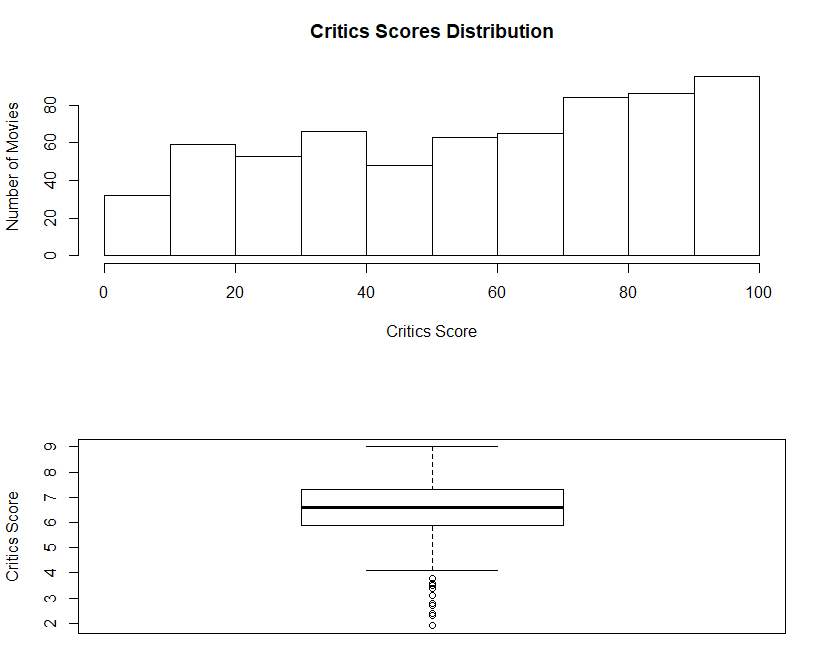
I will use a significance level of 5% for my study.

**Is there a relationship between IMDB ratings (numerical variable) and critics score (numerical variable) on Rotten Tomatoes?**

****

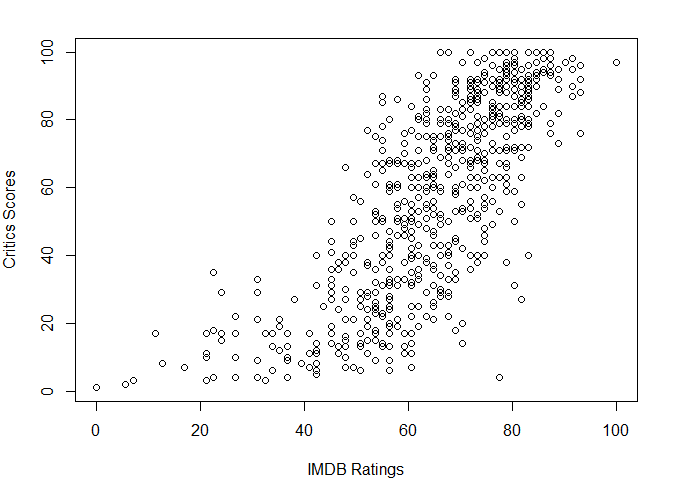
****

****



I rescaled imdb\_rating data to be able to compare with critics\_score data. IMDB ratings were in the range 0-10 but critics score on Rotten Tomatoes were in the range 0-100. I used the R code to rescale imdb\_rating so I could compare it with critics\_score.

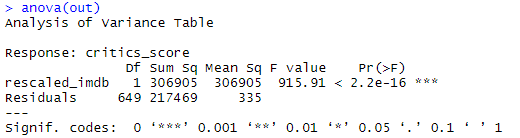
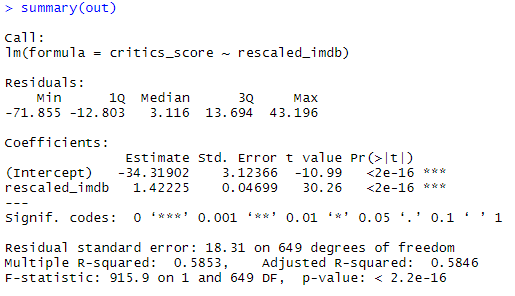
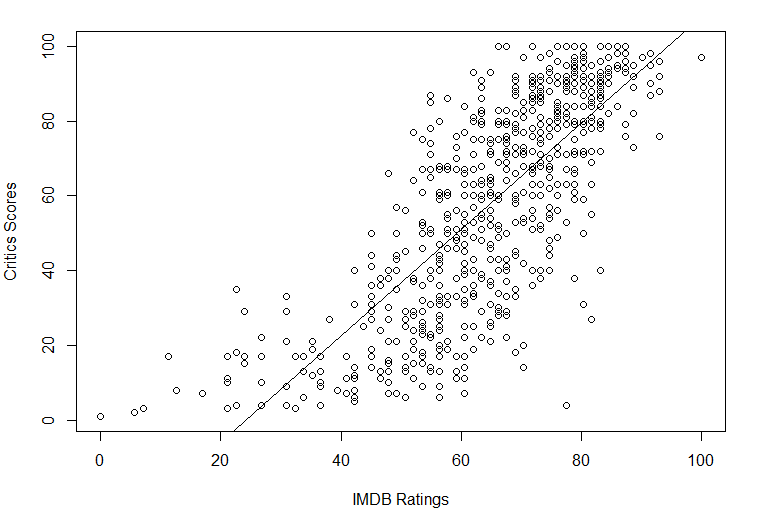




When we examine the scatter plot, we can obtain some ideas about the relationship between IMDB ratings and critics score on Rotten Tomatoes. As the circles on the scatter plot produce a lower left to upper right pattern, we can say that there is positive correlation between IMDB ratings and critics score on Rotten Tomatoes.

I used **Simple Linear Regression** test as my hypothesis test to see if, at a 5% significance level, there is a positive linear relationship between movie rating on IMDB and movie critic score on Rotten Tomatoes.

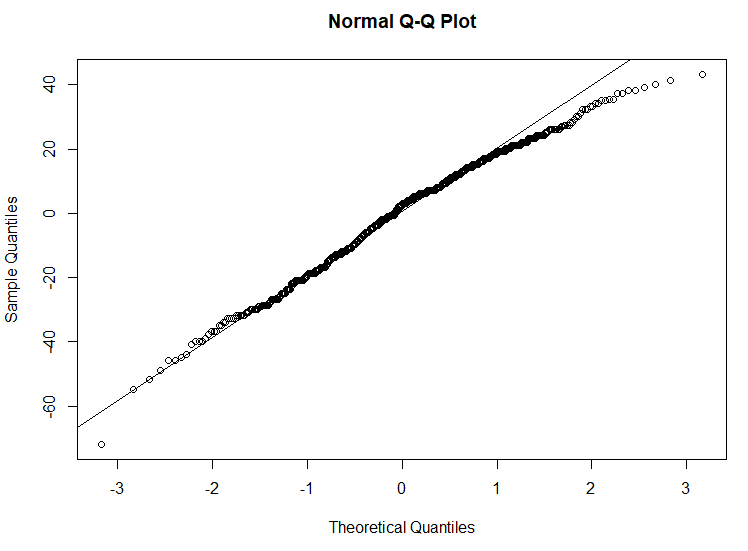
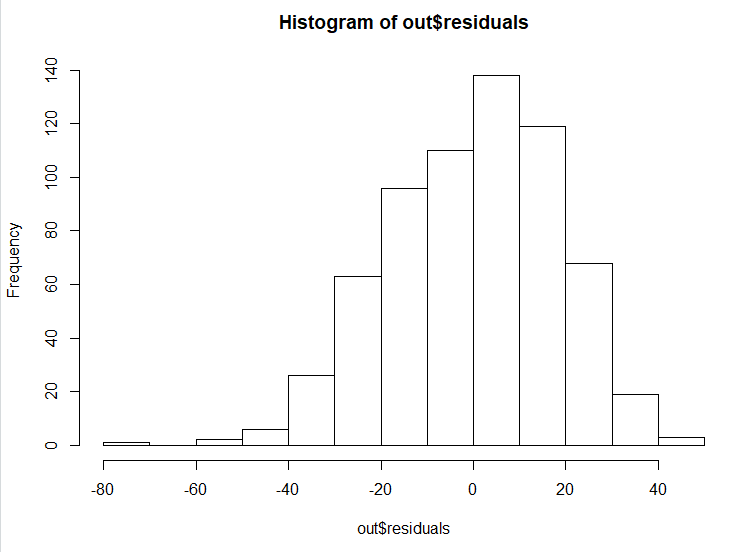
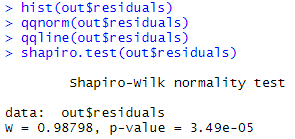
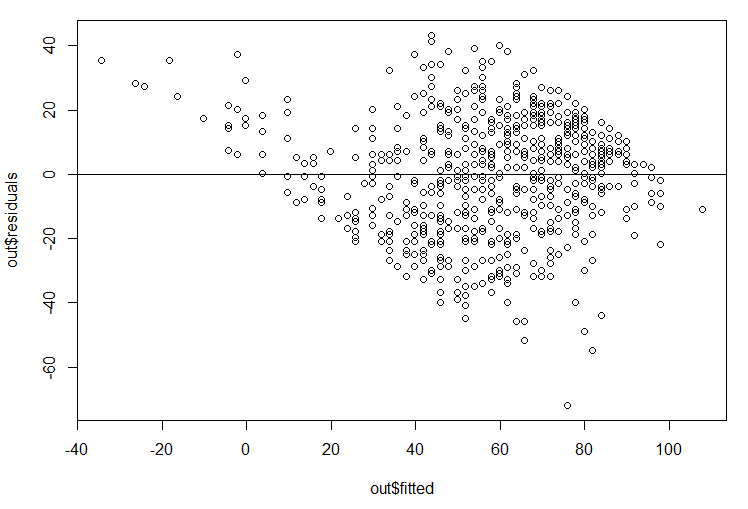
* Ho: β = 0 (There is not linear relationship between movie rating on IMDB and movie critic score on Rotten Tomatoes)
* Ha: β ≠ 0 (There is a linear relationship between movie rating on IMDB and movie critic score on Rotten Tomatoes)



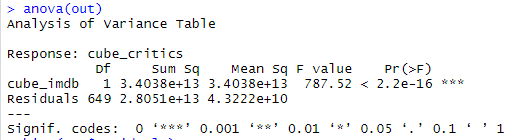
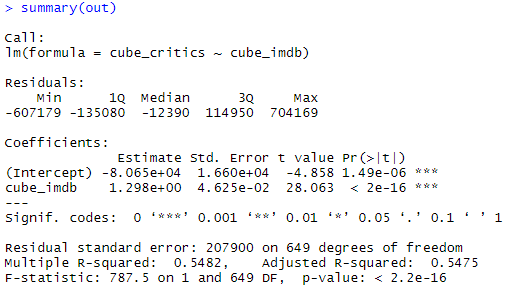
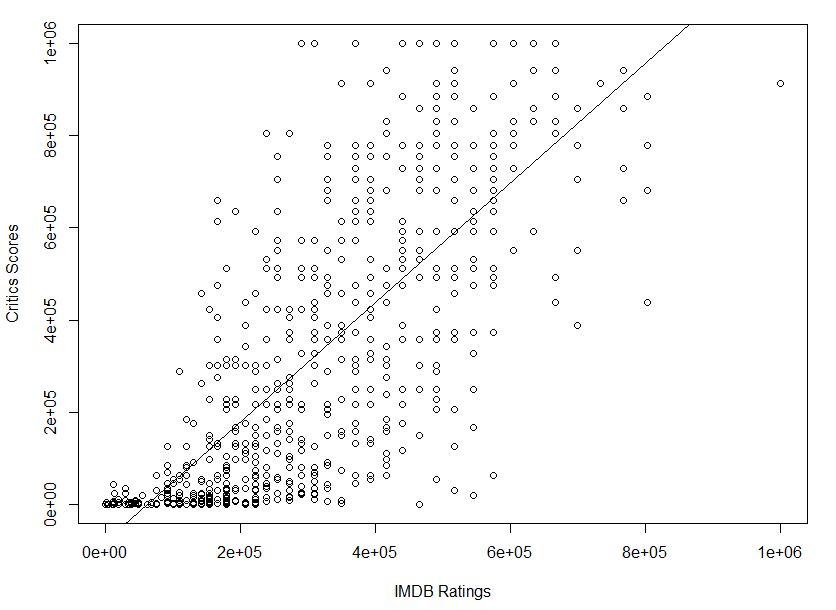
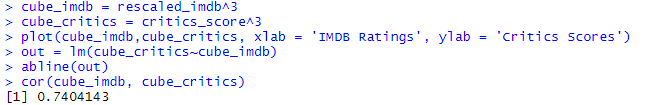
p-value = 2.2e-16 ≤ 0.05 → At 5% significance level, there is enough evidence to support that there is a positive linear relationship between movie rating on IMDB and movie critic score on Rotten Tomatoes.

**Assumptions:**

1. According to the scatter plot, there is a positive relationship between IMDB ratings and critics score on Rotten Tomatoes. They are linearly related.
2. Residuals don’t have constant variance (constant spread for all values) and linear model is not appropriate. Plot doesn’t have random pattern.



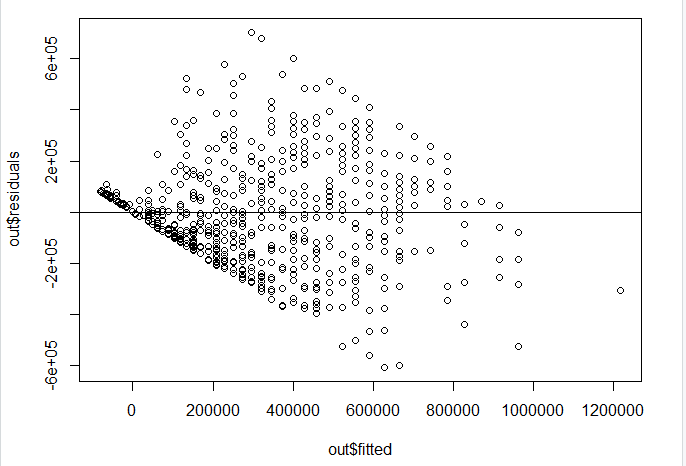
* Even though histogram and Q-Q plot look like that errors might be normally distributed, we can easily come up with the conclusion using shapiro test that errors are not normally distributed because p-value is less than 0.05.
* **Since some of the assumptions are not met, I tried to transform the data by taking the cube of both variables. I repeat the analysis like below but assumptions are not still met:**
* Ho: β = 0 (There is not linear relationship between movie rating on IMDB and movie critic score on Rotten Tomatoes)
* Ha: β ≠ 0 (There is a linear relationship between movie rating on IMDB and movie critic score on Rotten Tomatoes)

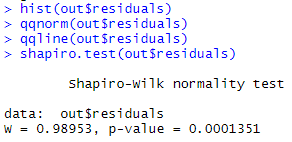


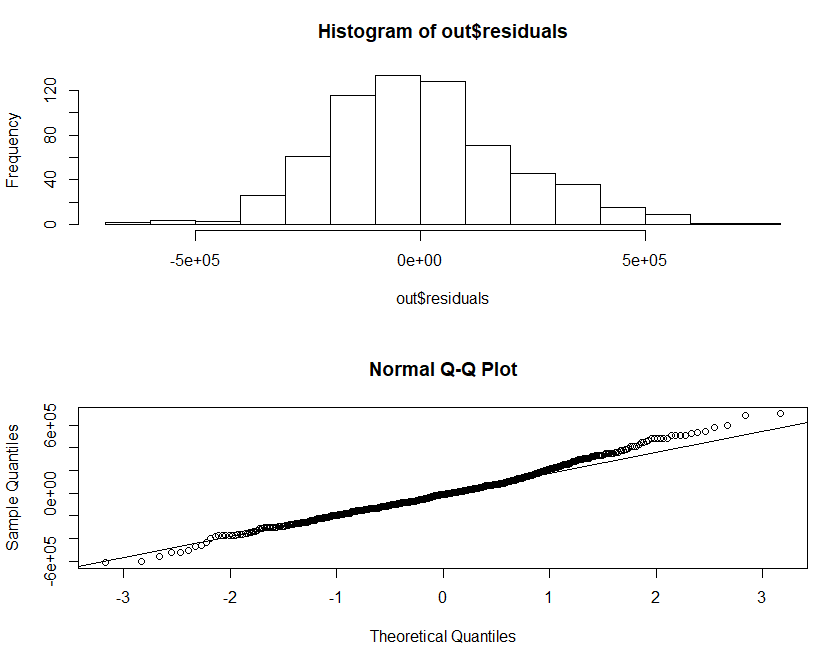
p-value = 2.2e-16 ≤ 0.05 → At 5% significance level, there is enough evidence to support that there is a positive linear relationship between movie rating on IMDB and movie critic score on Rotten Tomatoes.

**Assumptions:**

1. According to the scatter plot, there is a positive relationship between IMDB ratings and critics score on Rotten Tomatoes. They are linearly related.
2. Residuals don’t have constant variance (constant spread for all values) and linear model is not appropriate. Plot doesn’t have random pattern.

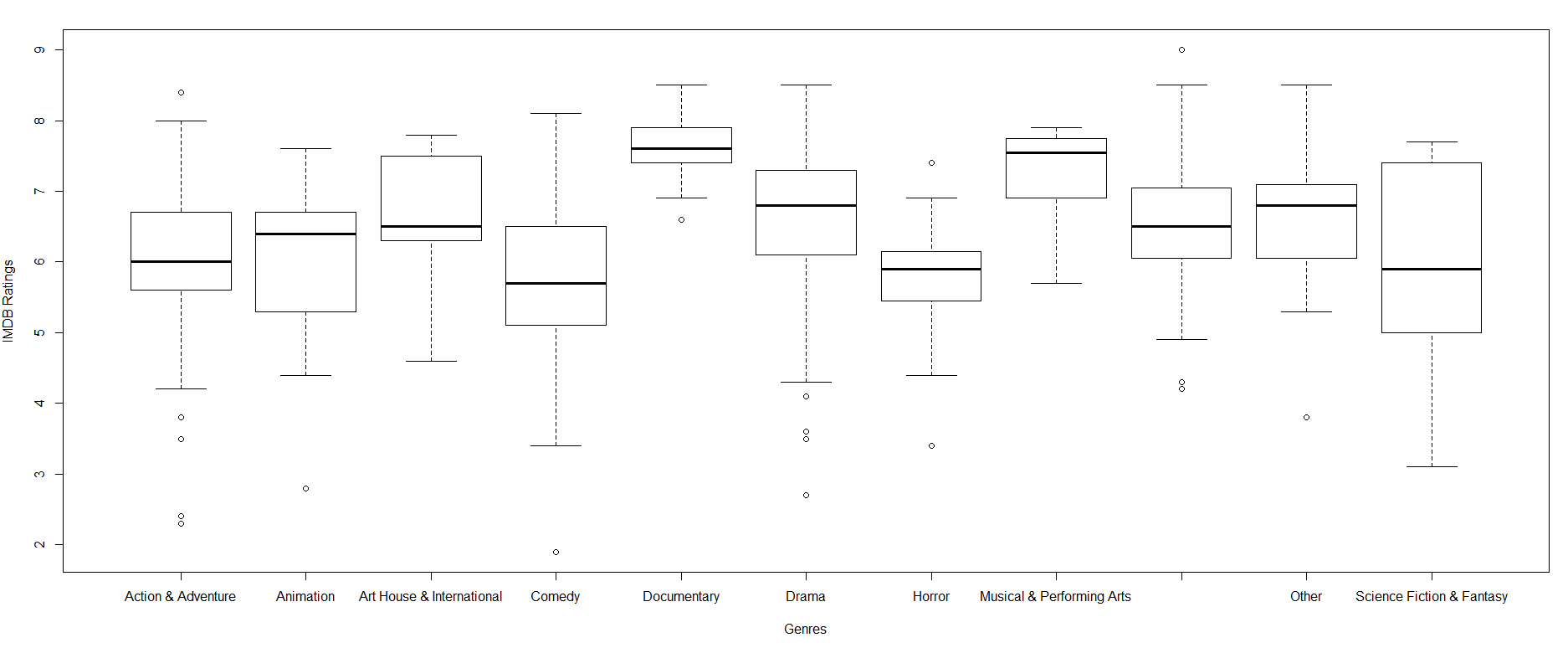
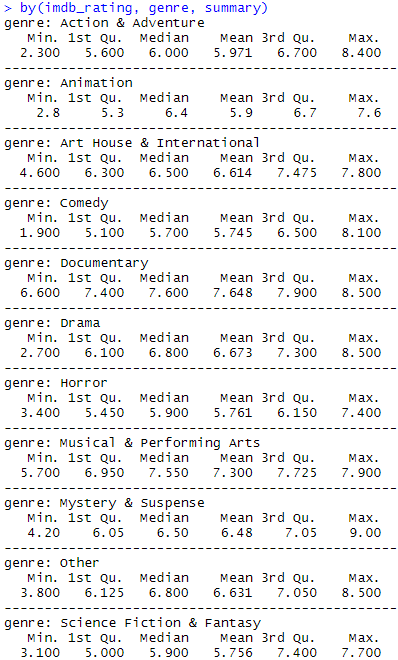
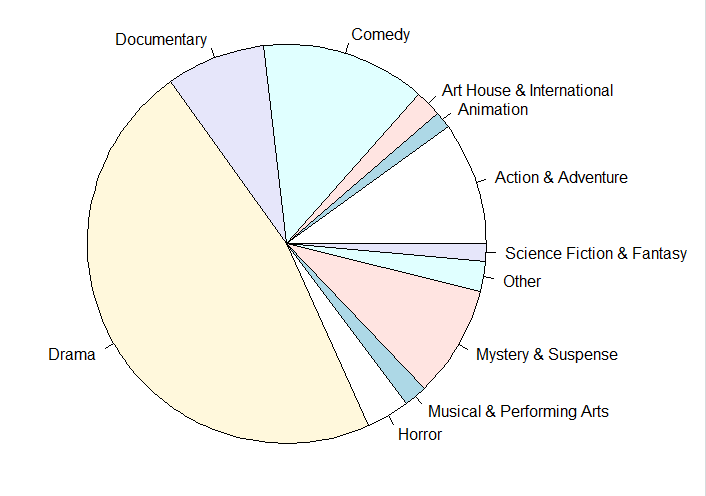
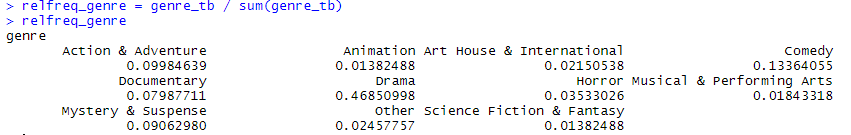
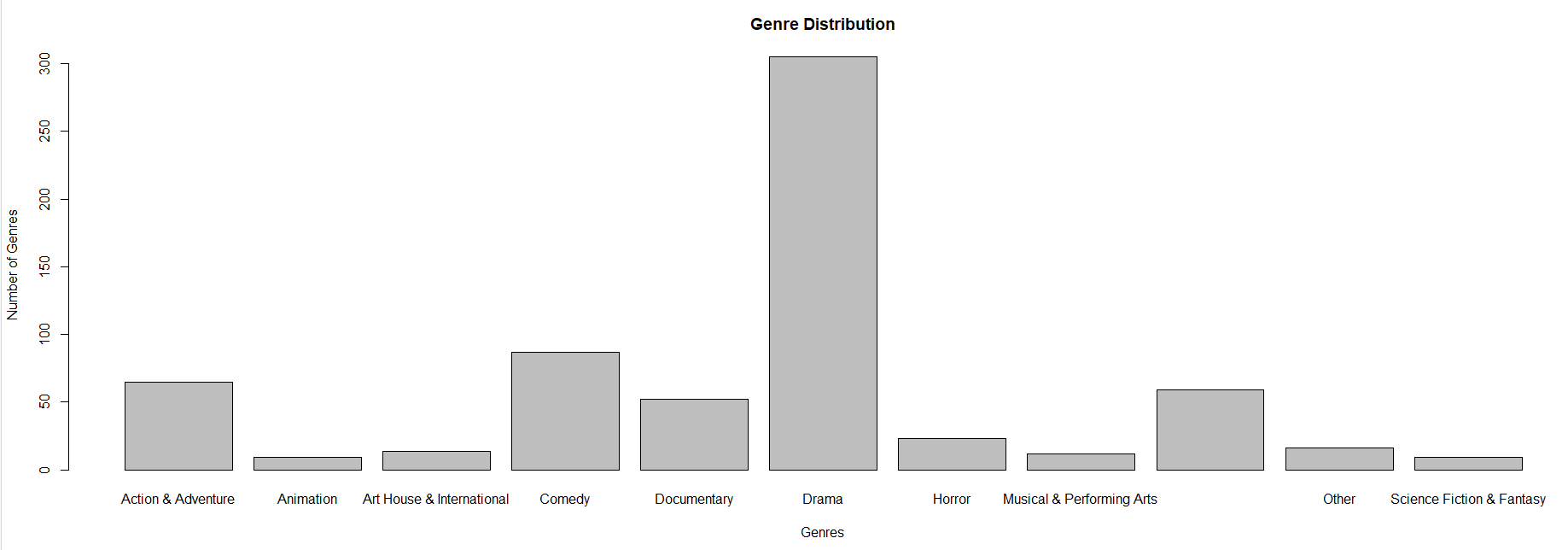


****



* Even though histogram and Q-Q plot look like that errors might be normally distributed, we can easily come up with the conclusion using shapiro test that errors are not normally distributed because p-value is less than 0.05.

**Is there an association between IMDB ratings (numerical variable) and genres of movies (categorical variable)?**



* According to boxplot and summary of IMDB ratings and genres, we can forecast that at least some of means of IMDB ratings are different from the others depending on their genres.

I used **ANOVA** test as my hypothesis test to see if, at a 5% significance level, the average IMDB ratings differed based on movie genre.

* Ho : µad = µan = µar = µc = µdo = µdr = µh = µmu = µmy = µo = µs
* Ha : At least 1 mean is different

µad = Average IMDB rating for Action & Adventure movies

µan = Average IMDB rating for Animation movies

µar = Average IMDB rating for Art House & International movies

µc = Average IMDB rating for Comedy movies

µdo = Average IMDB rating for Documentary movies

µdr = Average IMDB rating for Drama movies

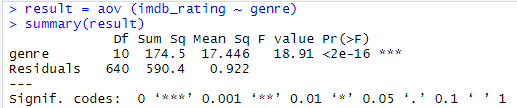
µh = Average IMDB rating for Horror movies

µmu = Average IMDB rating for Musical & Performing Arts movies

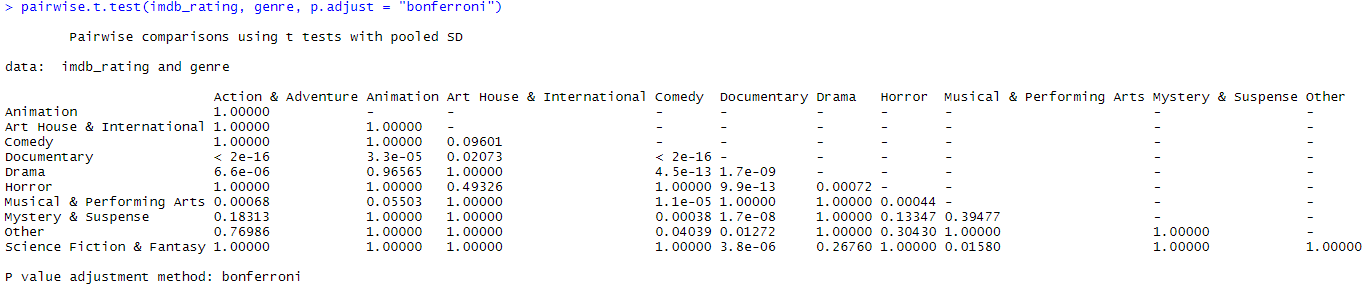
µmy = Average IMDB rating for Mystery & Suspense movies

µo = Average IMDB rating for Other movies

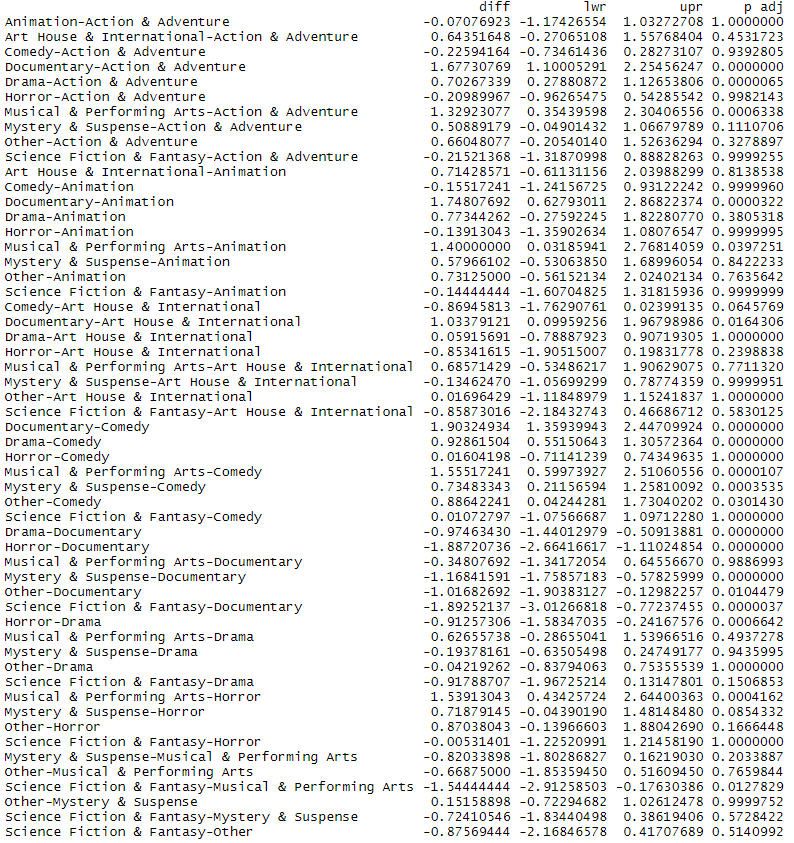
µs = Average IMDB rating for Science Fiction & Fantasy movies



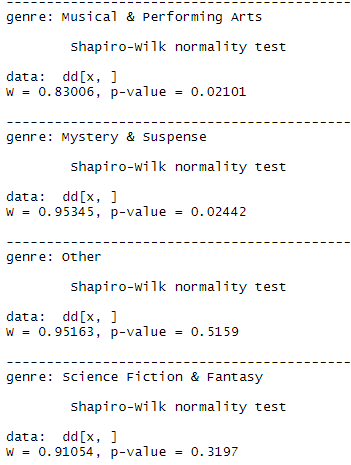
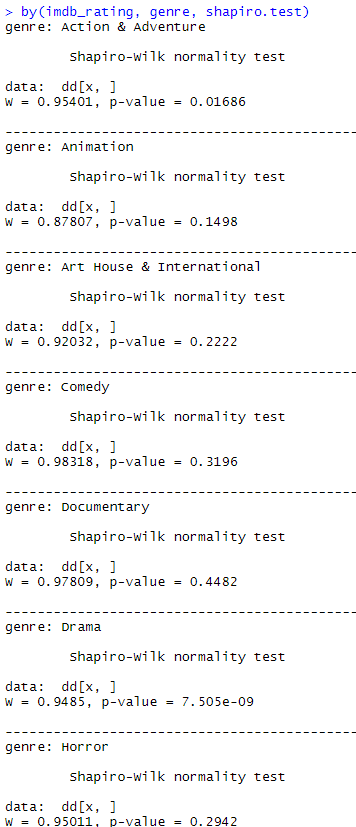
* p-value ≤ 0.05 → At a 5% significance level, we have enough evidence to support the average IMDB ratings differed based on movie genre. There is an association between IMDB ratings and genres of movies.



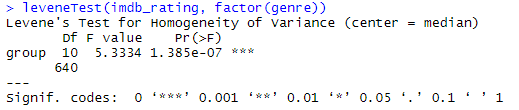
* µad ≠ µdo (p-value ≤ 0.05), µad ≠ µh (p-value ≤ 0.05), µad ≠ µmu (p-value ≤ 0.05), µan ≠ µdo (p-value ≤ 0.05), µar ≠ µdo (p-value ≤ 0.05), µar ≠ µdo (p-value ≤ 0.05), µc ≠ µdo (p-value ≤ 0.05), µc ≠ µdr (p-value ≤ 0.05), µc ≠ µmu (p-value ≤ 0.05), µc ≠ µmy (p-value ≤ 0.05), µc ≠ µo (p-value ≤ 0.05), µdo ≠ µdr (p-value ≤ 0.05), µdo ≠ µh (p-value ≤ 0.05), µdo ≠ µmy (p-value ≤ 0.05), µdo ≠ µo (p-value ≤ 0.05), µdo ≠ µs (p-value ≤ 0.05), µdr ≠ µh (p-value ≤ 0.05), µh ≠ µmu (p-value ≤ 0.05), µmu ≠ µs (p-value ≤ 0.05),



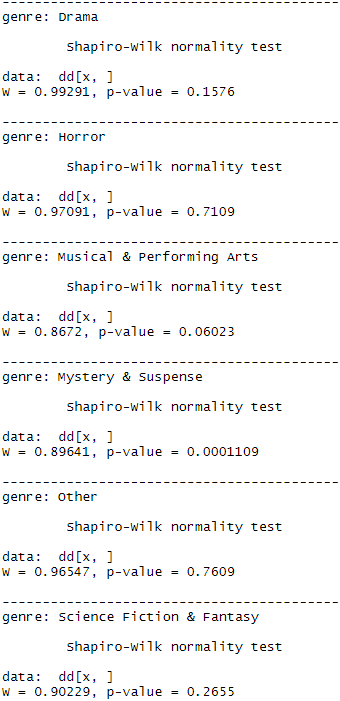
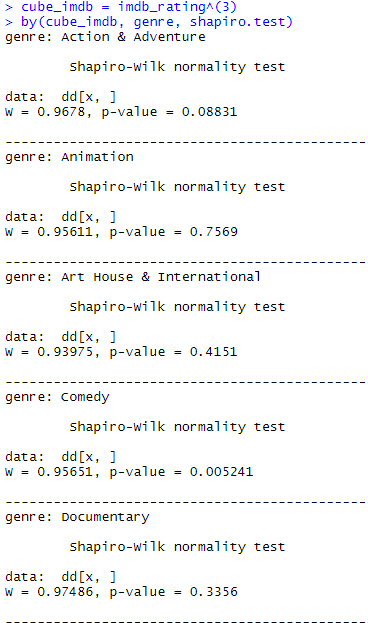
* µad ≠ µdo (p-value ≤ 0.05), µad ≠ µh (p-value ≤ 0.05), µad ≠ µmu (p-value ≤ 0.05), µan ≠ µdo (p-value ≤ 0.05), µar ≠ µdo (p-value ≤ 0.05), µar ≠ µdo (p-value ≤ 0.05), µc ≠ µdo (p-value ≤ 0.05), µc ≠ µdr (p-value ≤ 0.05), µc ≠ µmu (p-value ≤ 0.05), µc ≠ µmy (p-value ≤ 0.05), µc ≠ µo (p-value ≤ 0.05), µdo ≠ µdr (p-value ≤ 0.05), µdo ≠ µh (p-value ≤ 0.05), µdo ≠ µmy (p-value ≤ 0.05), µdo ≠ µo (p-value ≤ 0.05), µdo ≠ µs (p-value ≤ 0.05), µdr ≠ µh (p-value ≤ 0.05), µh ≠ µmu (p-value ≤ 0.05), µmu ≠ µs (p-value ≤ 0.05),



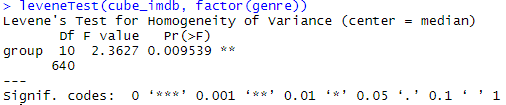
* p-value (Animation) > 0.05, p-value (Art House & International) > 0.05, p-value (Comedy) > 0.05, p-value (Documentary) > 0.05, p-value (Horror) > 0.05, p-value (Other) > 0.05, p-value (Science Fiction & Fantasy) > 0.05 → IMDB ratings for 7 of 11 genres are normally distributed.
* p-value (Action & Adventure) ≤ 0.05, p-value (Drama) ≤ 0.05, p-value (Musical & Performing Arts) ≤ 0.05, p-value (Mystery & Suspense) ≤ 0.05 → IMDB ratings for 4 of 11 genres are not normally distributed.



* p-value = 1.385e-07 ≤ 0.05 → IMDB ratings for all genres don’t have same variance.
* **Since some of the assumptions are not met, I tried to transform the data by taking the cube of IMDB ratings. I repeat the analysis like below but assumptions are not still met:**



* p-value (Animation) > 0.05, p-value (Art House & International) > 0.05, p-value (Action & Adventure) > 0.05, p-value (Documentary) > 0.05, p-value (Horror) > 0.05, p-value (Other) > 0.05, p-value (Science Fiction & Fantasy) > 0.05, p-value (Drama) > 0.05, p-value (Musical & Performing Arts) > 0.05 → IMDB ratings^(3) for 9 of 11 genres are normally distributed.
* p-value (Comedy) ≤ 0.05, p-value (Mystery & Suspense) ≤ 0.05 → IMDB ratings^(3) for 2 of 11 genres are still not normally distributed.



* p-value = 0.0095 ≤ 0.05 → IMDB ratings^(3) for all genres still don’t have same variance.

Conclusions on my research questions are:

* There is a relationship between IMDB ratings (numerical variable) and critics score (numerical variable) on Rotten Tomatoes according to **Simple Linear Regression** test that I used.
* There is an association between IMDB ratings (numerical variable) and genres of movies (categorical variable) according to **ANOVA** test that I used.
* The results from sample can be generalized to the population it represents because sample is randomly taken from its bigger population (all of the movies released between 1972 and 2014 in the US).
* Even though IMDB ratings and critics score are related or IMDB ratings and genres of movies are associated, it doesn’t mean that one of the variables causes the other variable to happen. That’s why we can’t make causal statements.