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.

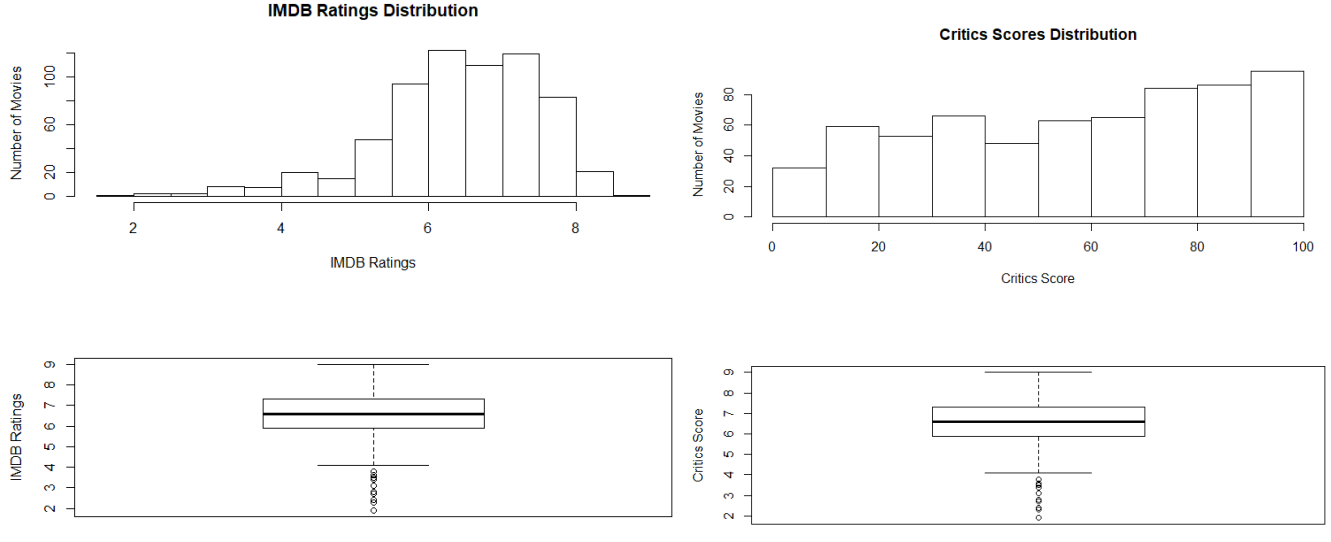
I get my data from http://www2.stat.duke.edu/~mc301/data/movies.html.

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.

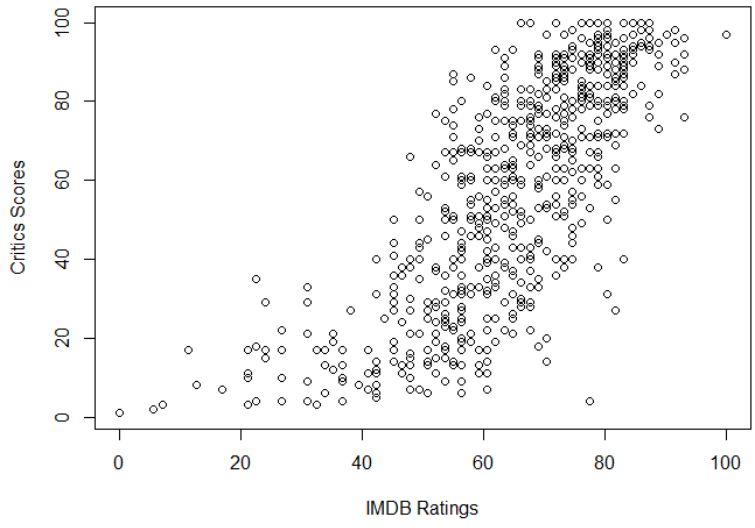
Is there a relationship between IMDB ratings (numerical variable) and critics score (numerical variable) on Rotten Tomatoes?

```
> par(mfrow = c(2,1))
> hist(imdb_rating, xlab='IMDB Ratings', ylab='Number of Movies', main = c('IMDB Ratings Distribution'))
> boxplot(imdb_rating, ylab = 'IMDB Ratings')
> summary(imdb_rating)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.900   5.900   6.600   6.493   7.300   9.000

> hist(critics_score, xlab='Critics Score', ylab='Number of Movies', main = c('Critics Scores Distribution'))
> boxplot(imdb_rating, ylab = 'Critics Score')
> summary(critics_score)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00   33.00   61.00   57.69   83.00  100.00
```



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.

```
> rescaled_imdb = rescale(imdb_rating, to = c(0, 100), from = range(imdb_rating, na.rm = TRUE, finite = TRUE))

> plot(rescaled_imdb,critics_score, xlab = 'IMDB Ratings', ylab = 'Critics Scores')
```

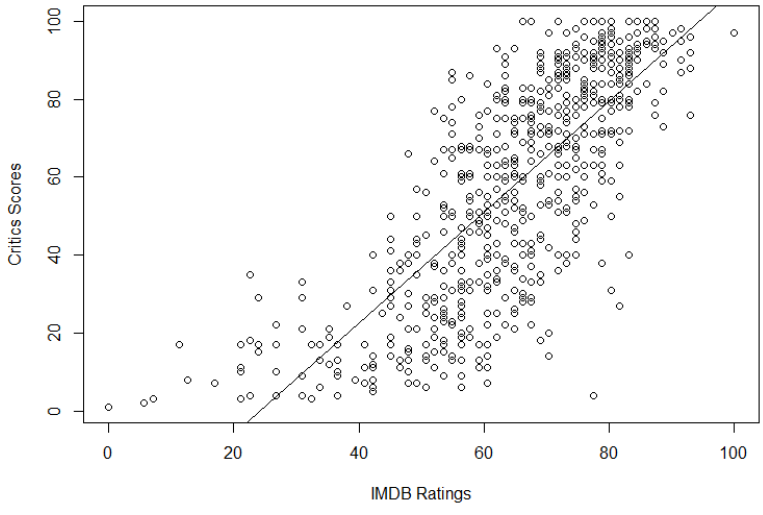


When we examine the scatter plot, we can obtain some ideas about the relationship between IMDB ratings and critics score on Rotten Tomatoes. As the circles on the scatter plot produce a lower left to upper right pattern, we can say that there is positive correlation between IMDB ratings and critics score 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: $\beta = 0$ (There is not linear relationship between movie rating on IMDB and movie critic score on Rotten Tomatoes)
- Ha: $\beta \neq 0$ (There is a linear relationship between movie rating on IMDB and movie critic score on Rotten Tomatoes)

```
> plot(rescaled_imdb,critics_score, xlab = 'IMDB Ratings', ylab = 'Critics Scores')
> out = lm(critics_score~rescaled_imdb)
> abline(out)
```




```
> cor(rescaled_imdb, critics_score)
[1] 0.7650355
```

```
> summary(out)

Call:
lm(formula = critics_score ~ rescaled_imdb)

Residuals:
    Min      1Q  Median      3Q     Max
-71.855 -12.803   3.116  13.694  43.196

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -34.31902    3.12366  -10.99   <2e-16 ***
rescaled_imdb   1.42225    0.04699   30.26   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.31 on 649 degrees of freedom
Multiple R-squared:  0.5853,    Adjusted R-squared:  0.5846
F-statistic: 915.9 on 1 and 649 DF,  p-value: < 2.2e-16
```

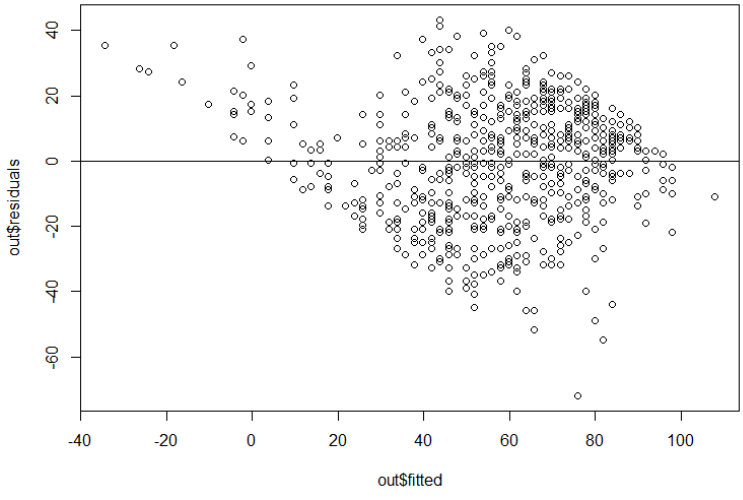
```
> anova(out)
Analysis of Variance Table

Response: critics_score
               Df Sum Sq Mean Sq F value    Pr(>F)
rescaled_imdb   1 306905  306905  915.91 < 2.2e-16 ***
Residuals     649 217469     335
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value = 2.2e-16 $\leq$ 0.05 $\rightarrow$ 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:

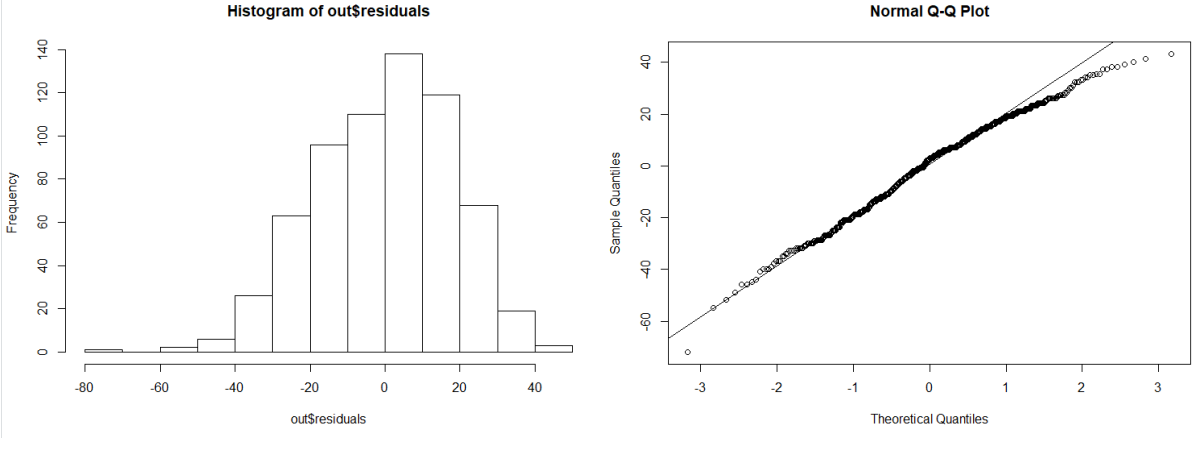
1) According to the scatter plot, there is a positive relationship between IMDB ratings and critics score on Rotten Tomatoes. They are linearly related.
2) Residuals don't have constant variance (constant spread for all values) and linear model is not appropriate. Plot doesn't have random pattern.

```
> plot(out$fitted, out$residuals)
> abline(0,0)
```



3)

```
> hist(out$residuals)
> qqnorm(out$residuals)
> qqline(out$residuals)
> shapiro.test(out$residuals)

        Shapiro-Wilk normality test

data:  out$residuals
W = 0.98798, p-value = 3.49e-05
```

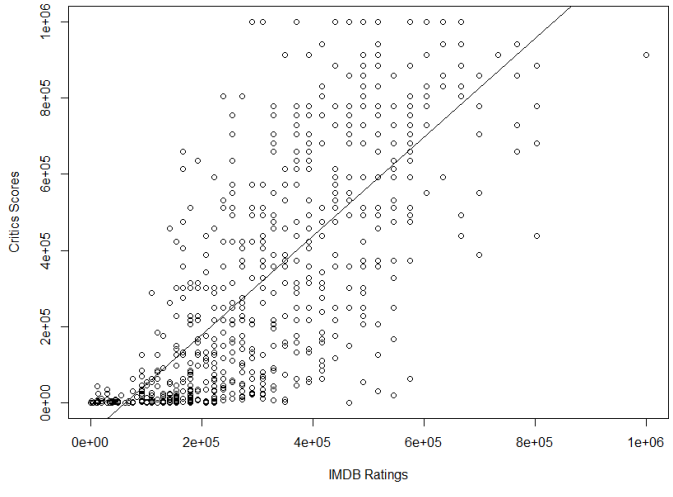


- Even though histogram and Q-Q plot look like that errors might be normally distributed, we can easily come up with the conclusion using shapiro test that errors are not normally distributed because p-value is less than 0.05.

- **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: $\beta = 0$ (There is not linear relationship between movie rating on IMDB and movie critic score on Rotten Tomatoes)
- Ha: $\beta \neq 0$ (There is a linear relationship between movie rating on IMDB and movie critic score on Rotten Tomatoes)

```
> cube_imdb = rescaled_imdb^3
> cube_critics = critics_score^3
> plot(cube_imdb,cube_critics, xlab = 'IMDB Ratings', ylab = 'Critics Scores')
> out = lm(cube_critics~cube_imdb)
> abline(out)
> cor(cube_imdb, cube_critics)
[1] 0.7404143
```




```
> summary(out)

Call:
lm(formula = cube_critics ~ cube_imdb)

Residuals:
    Min      1Q  Median      3Q     Max
-607179 -135080  -12390  114950  704169

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.065e+04  1.660e+04  -4.858 1.49e-06 ***
cube_imdb    1.298e+00  4.625e-02  28.063  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 207900 on 649 degrees of freedom
Multiple R-squared:  0.5482,    Adjusted R-squared:  0.5475
F-statistic: 787.5 on 1 and 649 DF,  p-value: < 2.2e-16
```

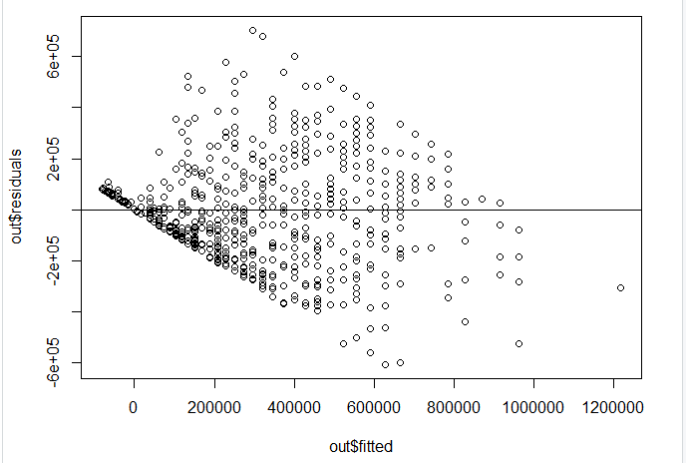
```
> anova(out)
Analysis of Variance Table

Response: cube_critics
            Df     Sum Sq    Mean Sq F value    Pr(>F)
cube_imdb    1 3.4038e+13 3.4038e+13  787.52 < 2.2e-16 ***
Residuals 649 2.8051e+13 4.3222e+10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

p-value = 2.2e-16 $\leq$ 0.05 $\rightarrow$ 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:

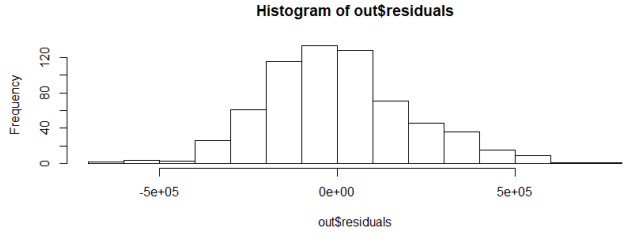
1) According to the scatter plot, there is a positive relationship between IMDB ratings and critics score on Rotten Tomatoes. They are linearly related.

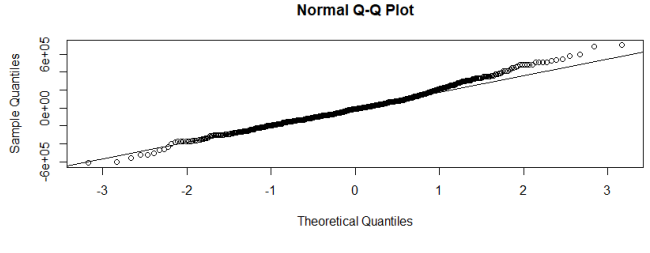
2) Residuals don't have constant variance (constant spread for all values) and linear model is not appropriate. Plot doesn't have random pattern.



3)

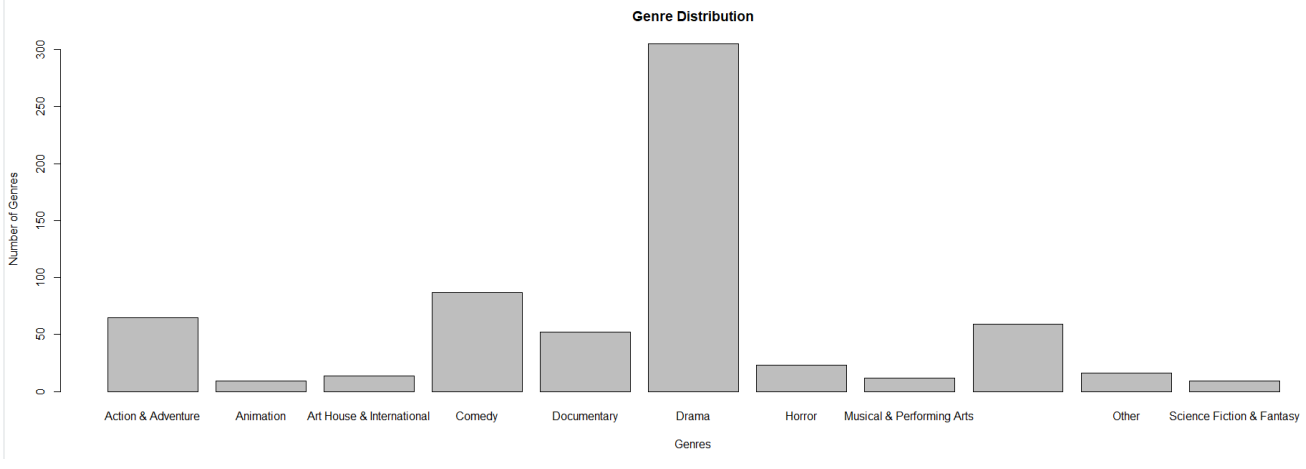
```
> hist(out$residuals)
> qqnorm(out$residuals)
> qqline(out$residuals)
> shapiro.test(out$residuals)

        Shapiro-Wilk normality test

data:  out$residuals
W = 0.98953, p-value = 0.0001351
```





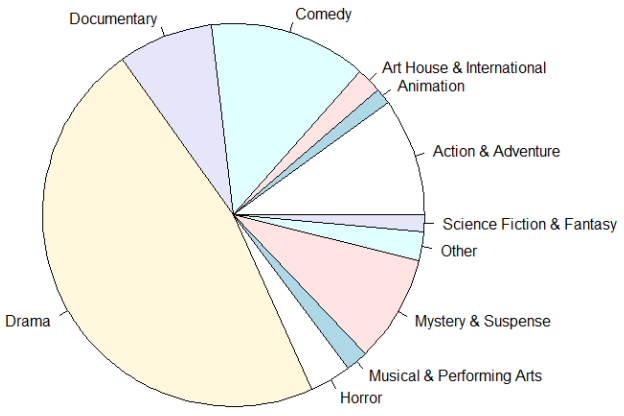
- Even though histogram and Q-Q plot look like that errors might be normally distributed, we can easily come up with the conclusion using shapiro test that errors are not normally distributed because p-value is less than 0.05.

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

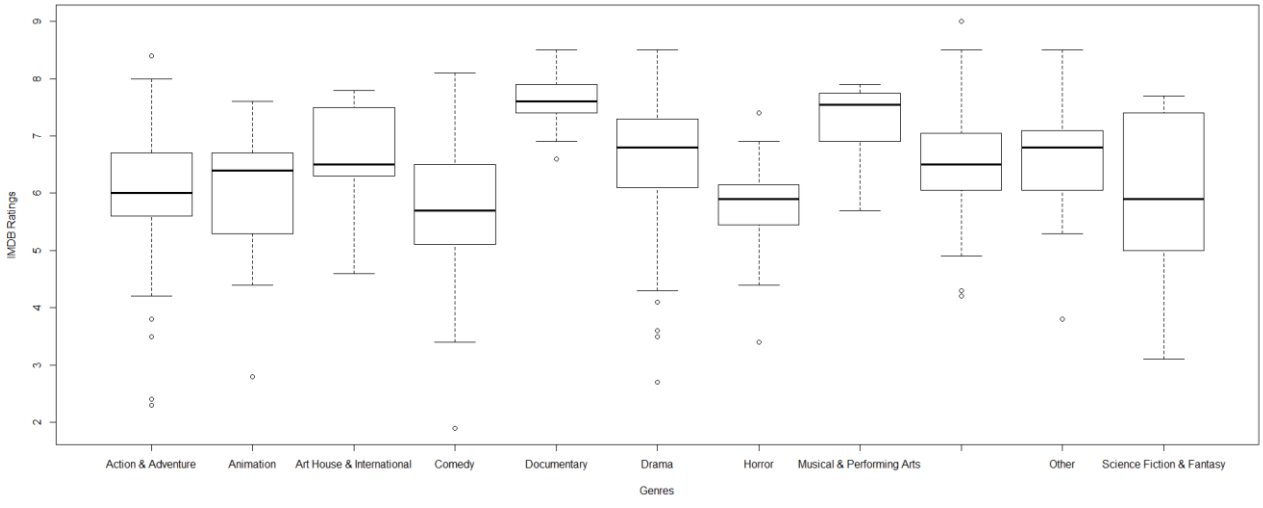



```
> relfreq_genre = genre_tb / sum(genre_tb)
> relfreq_genre
genre
       Action & Adventure                 Animation Art House & International                    Comedy
                0.09984639                0.01382488                0.02150538                0.13364055
               Documentary                     Drama                    Horror Musical & Performing Arts
                0.07987711                0.46850998                0.03533026                0.01843318
       Mystery & Suspense                     Other Science Fiction & Fantasy
                0.09062980                0.02457757                0.01382488
```

```
> pie(genre_tb, labels = c('Action & Adventure', 'Animation', 'Art House & International', 'Comedy', 'Documentary', 'Drama', 'Horror', 'Music
al & Performing Arts', 'Mystery & Suspense', 'Other', 'Science Fiction & Fantasy'))
```




```
> by(imdb_rating, genre, summary)
genre: Action & Adventure
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.300   5.600   6.000   5.971   6.700   8.400
------------------------------------------------------------
genre: Animation
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    2.8     5.3     6.4     5.9     6.7     7.6
------------------------------------------------------------
genre: Art House & International
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.600   6.300   6.500   6.614   7.475   7.800
------------------------------------------------------------
genre: Comedy
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.900   5.100   5.700   5.745   6.500   8.100
------------------------------------------------------------
genre: Documentary
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.600   7.400   7.600   7.648   7.900   8.500
------------------------------------------------------------
genre: Drama
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.700   6.100   6.800   6.673   7.300   8.500
------------------------------------------------------------
genre: Horror
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.400   5.450   5.900   5.761   6.150   7.400
------------------------------------------------------------
genre: Musical & Performing Arts
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  5.700   6.950   7.550   7.300   7.725   7.900
------------------------------------------------------------
genre: Mystery & Suspense
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.20    6.05    6.50    6.48    7.05    9.00
------------------------------------------------------------
genre: Other
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.800   6.125   6.800   6.631   7.050   8.500
------------------------------------------------------------
genre: Science Fiction & Fantasy
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  3.100   5.000   5.900   5.756   7.400   7.700
```



- 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 : $\mu_{ad} = \mu_{an} = \mu_{ar} = \mu_{c} = \mu_{do} = \mu_{dr} = \mu_{h} = \mu_{mu} = \mu_{my} = \mu_{o} = \mu_{s}$
- Ha : At least 1 mean is different

μ_{ad} = Average IMDB rating for Action & Adventure movies

μ_{an} = Average IMDB rating for Animation movies

μ_{ar} = Average IMDB rating for Art House & International movies

μ_{c} = Average IMDB rating for Comedy movies

μ_{do} = Average IMDB rating for Documentary movies

μ_{dr} = Average IMDB rating for Drama movies

μ_h = Average IMDB rating for Horror movies

μ_{mu} = Average IMDB rating for Musical & Performing Arts movies

μ_{my} = Average IMDB rating for Mystery & Suspense movies

μ_o = Average IMDB rating for Other movies

μ_s = Average IMDB rating for Science Fiction & Fantasy movies

```
> result = aov (imdb_rating ~ genre)
> summary(result)
              Df Sum Sq Mean Sq F value Pr(>F)
genre         10  174.5  17.446   18.91 <2e-16 ***
Residuals    640  590.4   0.922
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 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.

```
> pairwise.t.test(imdb_rating, genre, p.adjust = "bonferroni")

        Pairwise comparisons using t tests with pooled SD

data:  imdb_rating and genre

                       Action & Adventure Animation Art House & International Comedy  Documentary Drama   Horror  Musical & Performing Arts Mystery & Suspense Other
Animation              1.00000            -         -                        -       -           -       -       -                         -                   -
Art House & International 1.00000         1.00000   -                        -       -           -       -       -                         -                   -
Comedy                 1.00000            1.00000   0.09601                  -       -           -       -       -                         -                   -
Documentary            < 2e-16            3.3e-05   0.02073                  < 2e-16 -           -       -       -                         -                   -
Drama                  6.6e-06            0.96565   1.00000                  4.5e-13 1.7e-09     -       -       -                         -                   -
Horror                 1.00000            1.00000   0.49326                  1.00000 9.9e-13     0.00072 -       -                         -                   -
Musical & Performing Arts 0.00068         0.05503   1.00000                  1.1e-05 1.00000     1.00000 0.00044 -                         -                   -
Mystery & Suspense     0.18313            1.00000   1.00000                  0.00038 1.7e-08     1.00000 0.13347 0.39477                   -                   -
Other                  0.76986            1.00000   1.00000                  0.04039 0.01272     1.00000 0.30430 1.00000                   1.00000             -
Science Fiction & Fantasy 1.00000         1.00000   1.00000                  1.00000 3.8e-06     0.26760 1.00000 0.01580                   1.00000             1.00000

P value adjustment method: bonferroni
```

- $\mu_{ad} \neq \mu_{do}$ (p-value ≤ 0.05), $\mu_{ad} \neq \mu_h$ (p-value ≤ 0.05), $\mu_{ad} \neq \mu_{mu}$ (p-value ≤ 0.05), $\mu_{an} \neq \mu_{do}$ (p-value ≤ 0.05), $\mu_{ar} \neq \mu_{do}$ (p-value ≤ 0.05), $\mu_{ar} \neq \mu_{do}$ (p-value ≤ 0.05), $\mu_c \neq \mu_{do}$ (p-value ≤ 0.05), $\mu_c \neq \mu_{dr}$ (p-value ≤ 0.05), $\mu_c \neq \mu_{mu}$ (p-value ≤ 0.05), $\mu_c \neq \mu_{my}$ (p-value ≤ 0.05), $\mu_c \neq \mu_o$ (p-value ≤ 0.05), $\mu_{do} \neq \mu_{dr}$ (p-value ≤ 0.05), $\mu_{do} \neq \mu_h$ (p-value ≤ 0.05), $\mu_{do} \neq \mu_{my}$ (p-value ≤ 0.05), $\mu_{do} \neq \mu_o$ (p-value ≤ 0.05), $\mu_{do} \neq \mu_s$ (p-value ≤ 0.05), $\mu_{dr} \neq \mu_h$ (p-value ≤ 0.05), $\mu_h \neq \mu_{mu}$ (p-value ≤ 0.05), $\mu_{mu} \neq \mu_s$ (p-value ≤ 0.05),

```
> TukeyHSD(result)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = imdb_rating ~ genre)

$genre
```

	diff	lwr	upr	p adj
Animation-Action & Adventure	-0.07076923	-1.17426554	1.03272708	1.0000000
Art House & International-Action & Adventure	0.64351648	-0.27065108	1.55768404	0.4531723
Comedy-Action & Adventure	-0.22594164	-0.73461436	0.28273107	0.9392805
Documentary-Action & Adventure	1.67730769	1.10005291	2.25456247	0.0000000
Drama-Action & Adventure	0.70267339	0.27880872	1.12653806	0.0000065
Horror-Action & Adventure	-0.20989967	-0.96265475	0.54285542	0.9982143
Musical & Performing Arts-Action & Adventure	1.32923077	0.35439598	2.30406556	0.0006338
Mystery & Suspense-Action & Adventure	0.50889179	-0.04901432	1.06679789	0.1110706
Other-Action & Adventure	0.66048077	-0.20540140	1.52636294	0.3278897
Science Fiction & Fantasy-Action & Adventure	-0.21521368	-1.31870998	0.88828263	0.9999255
Art House & International-Animation	0.71428571	-0.61131156	2.03988299	0.8138538
Comedy-Animation	-0.15517241	-1.24156725	0.93122242	0.9999960
Documentary-Animation	1.74807692	0.62793011	2.86822374	0.0000322
Drama-Animation	0.77344262	-0.27592245	1.82280770	0.3805318
Horror-Animation	-0.13913043	-1.35902634	1.08076547	0.9999995
Musical & Performing Arts-Animation	1.40000000	0.03185941	2.76814059	0.0397251
Mystery & Suspense-Animation	0.57966102	-0.53063850	1.68996054	0.8422233
Other-Animation	0.73125000	-0.56152134	2.02402134	0.7635642
Science Fiction & Fantasy-Animation	-0.14444444	-1.60704825	1.31815936	0.9999999
Comedy-Art House & International	-0.86945813	-1.76290761	0.02399135	0.0645769
Documentary-Art House & International	1.03379121	0.09959256	1.96798986	0.0164306
Drama-Art House & International	0.05915691	-0.78887923	0.90719305	1.0000000
Horror-Art House & International	-0.85341615	-1.90515007	0.19831778	0.2398838
Musical & Performing Arts-Art House & International	0.68571429	-0.53486217	1.90629075	0.7711320
Mystery & Suspense-Art House & International	-0.13462470	-1.05699299	0.78774359	0.9999951
Other-Art House & International	0.01696429	-1.11848979	1.15241837	1.0000000
Science Fiction & Fantasy-Art House & International	-0.85873016	-2.18432743	0.46686712	0.5830125
Documentary-Comedy	1.90324934	1.35939943	2.44709924	0.0000000
Drama-Comedy	0.92861504	0.55150643	1.30572364	0.0000000
Horror-Comedy	0.01604198	-0.71141239	0.74349635	1.0000000
Musical & Performing Arts-Comedy	1.55517241	0.59973927	2.51060556	0.0000107
Mystery & Suspense-Comedy	0.73483343	0.21156594	1.25810092	0.0003535
Other-Comedy	0.88642241	0.04244281	1.73040202	0.0301430
Science Fiction & Fantasy-Comedy	0.01072797	-1.07566687	1.09712280	1.0000000
Drama-Documentary	-0.97463430	-1.44012979	-0.50913881	0.0000000
Horror-Documentary	-1.88720736	-2.66416617	-1.11024854	0.0000000
Musical & Performing Arts-Documentary	-0.34807692	-1.34172054	0.64556670	0.9886993
Mystery & Suspense-Documentary	-1.16841591	-1.75857183	-0.57825999	0.0000000
Other-Documentary	-1.01682692	-1.90383127	-0.12982257	0.0104479
Science Fiction & Fantasy-Documentary	-1.89252137	-3.01266818	-0.77237455	0.0000037
Horror-Drama	-0.91257306	-1.58347035	-0.24167576	0.0006642
Musical & Performing Arts-Drama	0.62655738	-0.28655041	1.53966516	0.4937278
Mystery & Suspense-Drama	-0.19378161	-0.63505498	0.24749177	0.9435995
Other-Drama	-0.04219262	-0.83794063	0.75355539	1.0000000
Science Fiction & Fantasy-Drama	-0.91788707	-1.96725214	0.13147801	0.1506853
Musical & Performing Arts-Horror	1.53913043	0.43425724	2.64400363	0.0004162
Mystery & Suspense-Horror	0.71879145	-0.04390190	1.48148480	0.0854332
Other-Horror	0.87038043	-0.13966603	1.88042690	0.1666448
Science Fiction & Fantasy-Horror	-0.00531401	-1.22520991	1.21458190	1.0000000
Mystery & Suspense-Musical & Performing Arts	-0.82033898	-1.80286827	0.16219030	0.2033887
Other-Musical & Performing Arts	-0.66875000	-1.85359450	0.51609450	0.7659844
Science Fiction & Fantasy-Musical & Performing Arts	-1.54444444	-2.91258503	-0.17630386	0.0127829
Other-Mystery & Suspense	0.15158898	-0.72294682	1.02612478	0.9999752
Science Fiction & Fantasy-Mystery & Suspense	-0.72410546	-1.83440498	0.38619406	0.5728422
Science Fiction & Fantasy-Other	-0.87569444	-2.16846578	0.41707689	0.5140992

- $\mu ad \neq \mu do$ (p-value ≤ 0.05), $\mu ad \neq \mu h$ (p-value ≤ 0.05), $\mu ad \neq \mu mu$ (p-value ≤ 0.05), $\mu an \neq \mu do$ (p-value ≤ 0.05), $\mu ar \neq \mu do$ (p-value ≤ 0.05), $\mu ar \neq \mu do$ (p-value ≤ 0.05), $\mu c \neq \mu do$ (p-value ≤ 0.05), $\mu c \neq \mu dr$ (p-value ≤ 0.05), $\mu c \neq \mu mu$ (p-value ≤ 0.05), $\mu c \neq \mu my$ (p-value ≤ 0.05), $\mu c \neq \mu o$ (p-value ≤ 0.05), $\mu do \neq \mu dr$ (p-value ≤ 0.05), $\mu do \neq \mu h$ (p-value ≤ 0.05), $\mu do \neq \mu my$ (p-value ≤ 0.05), $\mu do \neq \mu o$ (p-value ≤ 0.05), $\mu do \neq \mu s$ (p-value ≤ 0.05), $\mu dr \neq \mu h$ (p-value ≤ 0.05), $\mu h \neq \mu mu$ (p-value ≤ 0.05), $\mu mu \neq \mu s$ (p-value ≤ 0.05),

```
> by(imdb_rating, genre, shapiro.test)
genre: Action & Adventure

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.95401, p-value = 0.01686

------------------------------------------------
genre: Animation

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.87807, p-value = 0.1498

------------------------------------------------
genre: Art House & International

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.92032, p-value = 0.2222

------------------------------------------------
genre: Comedy

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.98318, p-value = 0.3196

------------------------------------------------
genre: Documentary

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.97809, p-value = 0.4482

------------------------------------------------
genre: Drama

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9485, p-value = 7.505e-09

------------------------------------------------
genre: Horror

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.95011, p-value = 0.2942

------------------------------------------------
genre: Musical & Performing Arts

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.83006, p-value = 0.02101

------------------------------------------------
genre: Mystery & Suspense

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.95345, p-value = 0.02442

------------------------------------------------
genre: Other

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.95163, p-value = 0.5159

------------------------------------------------
genre: Science Fiction & Fantasy

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.91054, p-value = 0.3197
```

- 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.

```
> leveneTest(imdb_rating, factor(genre))
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group  10  5.3334 1.385e-07 ***
      640
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- 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:**

```
> cube_imdb = imdb_rating^(3)
> by(cube_imdb, genre, shapiro.test)
genre: Action & Adventure

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9678, p-value = 0.08831

------------------------------------------------
genre: Animation

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.95611, p-value = 0.7569

------------------------------------------------
genre: Art House & International

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.93975, p-value = 0.4151

------------------------------------------------
genre: Comedy

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.95651, p-value = 0.005241

------------------------------------------------
genre: Documentary

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.97486, p-value = 0.3356

------------------------------------------------
```

```
------------------------------------------
genre: Drama

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.99291, p-value = 0.1576

------------------------------------------
genre: Horror

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.97091, p-value = 0.7109

------------------------------------------
genre: Musical & Performing Arts

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.8672, p-value = 0.06023

------------------------------------------
genre: Mystery & Suspense

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.89641, p-value = 0.0001109

------------------------------------------
genre: Other

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.96547, p-value = 0.7609

------------------------------------------
genre: Science Fiction & Fantasy

        Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.90229, p-value = 0.2655
```

- 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.

```
> leveneTest(cube_imdb, factor(genre))
Levene's Test for Homogeneity of Variance (center = median)
       Df F value   Pr(>F)
group  10  2.3627 0.009539 **
      640
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- 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.