به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس استنباط آماری

پروژه اول-فاز دوم

دانیال ملکی

۸۱۰۱۹۵۲۰۰

خرداد ۱۳۹۶.

فهرست

Contents

دیتاست انتخابی

دیتاستی که برای این پروژه انتخاب شده است دیتاست فیلم های IMDB می باشد که به دلیل موضوع جالب و هیجان انگیز و آشنایی نسبی که با دیتاها و متغییر ها داشتم این موضوع و دیتاست را انتخاب کردم .

روابط جذابی که مابین میزان هزینه فیلم و امتیاز آن ها . هزینه و کارگردان فیلم . میزان محبوبیت بر اساس کشور سازنده و مابقی اطلاعاتی هیجان انگیزی که می توان از این دیتاست بدست آورد .

برای فراخوانی دیتاست مورد نظر دستورات زیر را باید در Rstudio وارد کنیم

```
library(readr)
movie_metadata <- read.csv("/movie_metadata.csv")
View(movie_metadata)
```

سوال ۱

در این سوال متغییر Categorical ای که انتخاب شده است متغییر color می باشد که دارای دو مقدار color و balck and white می باشد .

فرضی که برای این سوال در نظر میگیریم بدین صورت می باشد که مقدار p برابر با ۰,۵ می باشد بدین معنی که تعداد فیلم هایی که دارای مقدار color می باشند با تعداد فیلم هایی که دارای مقدار black and white می باشند برابر می باشد فرض ثانویه را نیز بدین صورت در نظر می گیریم که p مقداری مخالف با عدد ۰,۵ دارد .

$$H_0 : p = 0.5$$
$$H_A : p \neq 0.5$$

حال به بررسی این ادعا می پردازیم

مقداری که برای $\hat{p}$. n داریم از از طریق دستوران زیر که در Rstudio زده شده است به دست می آید

```
> color_movie=movie_metadata$color=="Color"
> list_of_color_movie=movie_metadata[color_movie,]
> nrow(list_of_color_movie)
[1] 4815
```

```
> number_of_movie=movie_metadata$color!=""
> movie_list2=movie_metadata[number_of_movie,]
> nrow(movie_list2)
[1] 5024
```

پس با توجه به داشتن مقادیر $\hat{p}$. n که در دستورات بالا بدست آمد مقدار $\hat{p}$ برابر است با

$$\hat{p} = \frac{4815}{5024} = 0.95 \, . n = 5024$$

حال به بررسی شرایط مورد نیاز برای بررسی این فرضیات می پردازیم

۱- استقلال : با توجه به این که تعداد فیلم های موجود در این دیتاست کمتر از ۱۰٪ فیلم های موجود در imdb می باشد این شرط برقرار می باشد .

۲- ۵۰۲۴×۰,۵ = ۲۵۱۲ که این مقدار از ۱۰ بزرگتر می باشد . و می توان چنین برداشت کرد که توزیع نیز نزدیک به نرمال می باشد .

$$\hat{p} = N(mean = 0.5 \, . SE = \sqrt{\frac{0.5 * 0.5}{5024}} = 0.007)$$

$$Z = \frac{0.95 - 0.5}{0.007} = 64.28$$

$$p_{value} = p(z > 64.28) * 2 = 0$$

پس نتیجه گیری که می توان از p_value داشت بدین صورت است که فرض صفر رد می شود و مقدار p عددی مخالف ۰,۵ می باشد .

سوال ۲

برای این سوال دو متغییر categorical ای که در نظر گرفته ایم color و language می باشد که به دلیل اینکه در language مقداری که داریم بیش از دو نوع می باشد مقدار را برای اینکه بتوانیم باینری در نظر بگیریم به این صورت تعریف می کنیم .

در صورتی که مقدار متغییر language مقداری برابر English داشت ۱ و در غیر این صورت ۰ در نظر می گیریم .

حال به بیان شرایط و محاسبه CI می پردازیم

۱- استقلال: با توجه به این که تعداد فیلم های موجود در این دیتاست کمتر از ۱۰٪ فیلم های موجود در imdb می باشد این شرط برقرار می باشد . همچنین باید توجه داشت که استقلال برون گروهی نیز برقرار می باشد و دو متغییر color و language از یکدیگر مستقل می باشند .

۲- می توان این فرض را برقرار دانست که توزیع تفاضل این دو متغییر تا حد زیادی به نرمال نزدیک می باشد .

حال به بررسی مقدار این دو نسبت و نحوه بدست آوردن آن ها می پردازیم .

$$CI: p_1 - p_2 \pm Z^* * SE$$

```
> english_movie=movie_metadata$language=="English"
> movie_list=movie_metadata[english_movie,]
> nrow(movie_list)
[1] 4704
```

```
> number_of_movie=movie_metadata$language!=""
> movie_list2=movie_metadata[number_of_movie,]
> nrow(movie_list2)
[1] 5031
```

```
> qnorm(0.05)
[1] -1.644854
```

برای متغیر color نیز مقادیری که در سوال قبل محاسبه شد استفاده خواهند شد .

$$\hat{p}_1 = \frac{4815}{5024} = 0.95 \,.\, n = 5024$$

$$\hat{p}_2 = \frac{4704}{5031} = 0.93 \,.\, n = 5031$$

$$(0.95 - 0.93) \pm 1.64 * \sqrt{\frac{0.95 * 0.05}{5024} + \frac{0.93 * 0.07}{5031}} = 0.02 \pm 0.007 = (0.012 \ . \ 0.028)$$

سوال ۳

برای این منظور متغییر country سطر ۱ تا ۱۵ را در نظر می گیریم تا خواسته های مسئله برقرار شود حال برای این متغییر مقدار USA را برابر ۱ و سایر مقادیر را برابر با ۰ در نظر می گیریم .

```
> nlevels(movie_metadata$country)
[1] 66
> levels(movie_metadata$country)
 [1] ""                    "Afghanistan"         "Argentina"           "Aruba"
"Australia"           "Bahamas"
 [7] "Belgium"             "Brazil"              "Bulgaria"            "Cambodia"
"Cameroon"            "Canada"
[13] "Chile"               "China"               "Colombia"            "Czech Repub
lic"       "Denmark"             "Dominican Republic"
[19] "Egypt"               "Finland"             "France"              "Georgia"
"Germany"             "Greece"
[25] "Hong Kong"           "Hungary"             "Iceland"             "India"
"Indonesia"           "Iran"
[31] "Ireland"             "Israel"              "Italy"               "Japan"
"Kenya"               "Kyrgyzstan"
[37] "Libya"               "Mexico"              "Netherlands"         "New Line"
"New Zealand"         "Nigeria"
[43] "Norway"              "Official site"       "Pakistan"            "Panama"
"Peru"                "Philippines"
```

```
[49] "Poland"              "Romania"             "Russia"              "Slovakia"
"Slovenia"            "South Africa"
[55] "South Korea"         "Soviet Union"        "Spain"               "Sweden"
"Switzerland"         "Taiwan"
[61] "Thailand"            "Turkey"              "UK"                  "United Arab
Emirates" "USA"                 "West Germany"
```

همانطور که از نتایج بالا مشخص می باشد تعداد ۶۶ فاکتور داریم که می توان چنین برداشت که احتمال اینکه مقدار متغییر country برای مسئله ما در صورتی که به صورت تصادفی country انتخاب شده باشد و برابر USA باشد با $\frac{1}{66}$ برابری می کند پس فرض اولیه و ثانویه به شکل زیر خواهد بود .

$$H_0 : p = \frac{1}{66}$$

$$H_A : p > \frac{1}{66}$$

حال مشاهده ای که داریم به صورت زیر می باشد

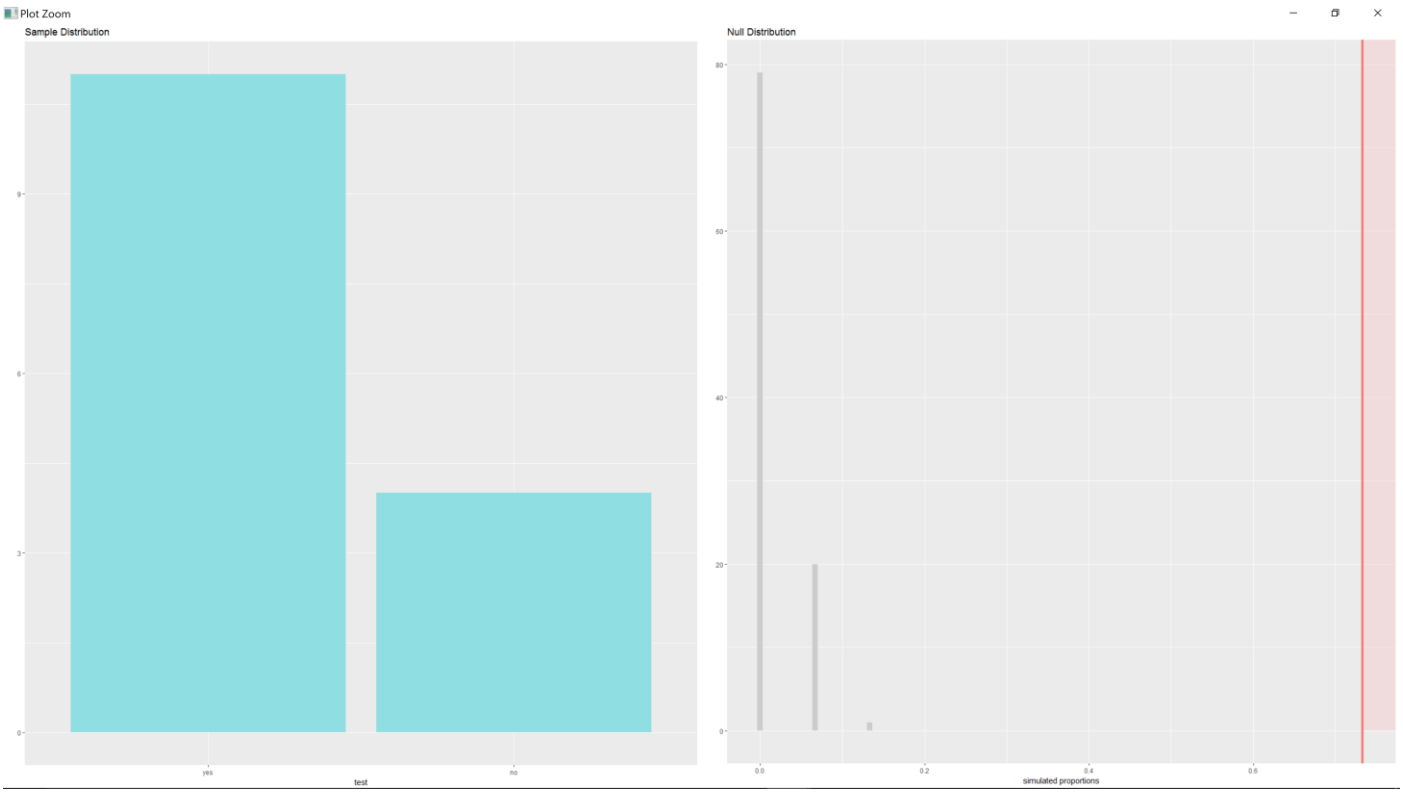
```
> movie_of_USA=movie_metadata[1:15,]$country=="USA"
> movie_of_USA
 [1]  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE FALSE
       TRUE  TRUE FALSE  TRUE  TRUE
```

که همانطور که در دستورات بالا نیز مشاهده می شود از ۱۵ مقدار بالا ۱۱ مقدار آن عبارتی برابر TRUE دارند و این یعنی مشاهده ما دارای نسبتی به اندازه $\frac{11}{15}$ می باشد .

حال به دلیل اینکه اندازه نمونه مقداری کوچک است در این سوال با استفاده از شبیه سازی سعی داریم تا بتوانیم نتیجه مورد نظر را بدست آوریم .

برای این منظور می توان اینطور در نظر داشت که ما می خواهیم یک تاس 66 وجهی را ۱۵ مرتبه پرتاب می کنیم و مقدار نسبت دیده شدن عدد ۱ را در آن مشاهده می کنیم این کار را به کرات انجام داده تا در نهایت مشاهده کنیم امکان اتفاق افتادن $p > \frac{11}{15}$ به چه میزان می باشد .

```
> test <- factor(c(rep("yes",11),rep("no",4)),levels = c("yes","no"))
> inference(test , data = movie_metadata , statistic = "proportion" , type = "ht" ,
        method = "simulation" , success = "yes" , null = 1/66 ,
        alternative = "greater", nsim = 100)
Single categorical variable, success: yes
n = 15, p-hat = 0.7333
H0: p = 0.0151515151515152
HA: p > 0.0151515151515152
p_value = < 0.0001
```



که همانطور که از نتایج مشخص می باشد فرض صفر یعنی تصادفی بودن country رد می شود و چنین فرضی درست نمی تواند باشد

سوال ۴

برای این سوال متغییر language را در نظر می گیریم که همان طور که از دستور زیر قابل مشاهده می باشد دارای ۴۷ (یکی از مقادیر حاوی null می باشد) حالت می باشد .

```
> nlevels(movie_metadata$language)
[1] 48
```

که مقادیر آن را در لیست زیر مشاهده می کنید .

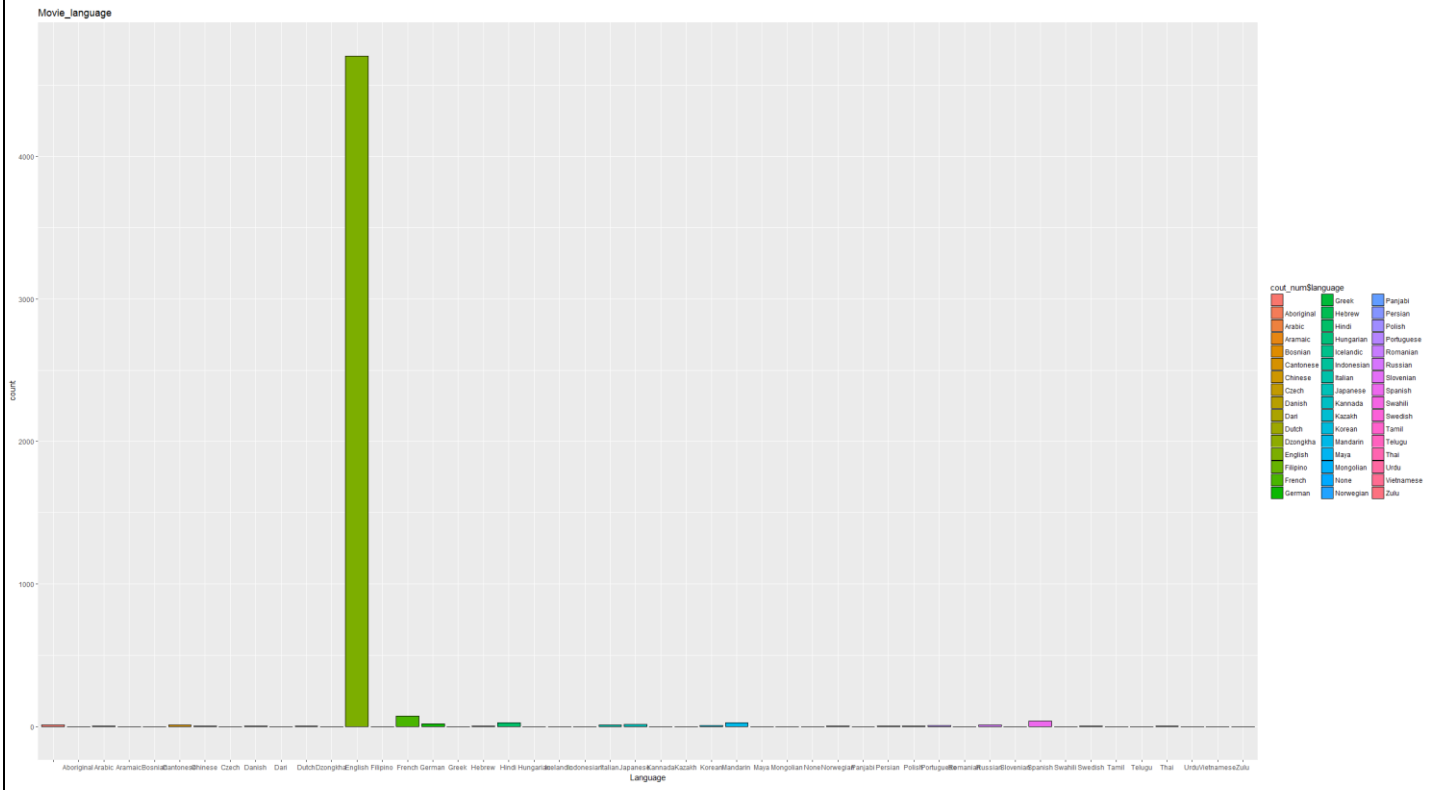
```
> library(plyr)
> count(movie_metadata,'language')
     language freq
1                12
2  Aboriginal    2
3      Arabic    5
4     Aramaic    1
5     Bosnian    1
6   Cantonese   11
7     Chinese    3
```

```
8       Czech    1
9      Danish    5
10       Dari    2
11      Dutch    4
12   Dzongkha    1
13    English 4704
14   Filipino    1
15     French   73
16     German   19
17      Greek    1
18     Hebrew    5
19      Hindi   28
20  Hungarian    1
21  Icelandic    2
22 Indonesian    2
23    Italian   11
24   Japanese   18
25    Kannada    1
26     Kazakh    1
27     Korean    8
28   Mandarin   26
29       Maya    1
30  Mongolian    1
31       None    2
32  Norwegian    4
33    Panjabi    1
34    Persian    4
35     Polish    4
36 Portuguese    8
37   Romanian    2
38    Russian   11
39  Slovenian    1
40    Spanish   40
41    Swahili    1
42    Swedish    5
43      Tamil    1
44     Telugu    1
45       Thai    3
46       Urdu    1
47 Vietnamese    1
48       Zulu    2
```

و همچنین Bar chart اطلاعات بالا نیز به صورت زیر خواهد بود .

```
> cout_num = count(movie_metadata,'language')

> ggplot(data = cout_num,aes(x=cout_num$language , y=cout_num$freq , fill=cout_num$lan
guage))+
        geom_bar(colour="black", stat="identity")+
        ggtitle("Movie_language")+xlab("Language")+ylab("count")
```



حال در قسمت دوم این سوال یک نمونه ۱۰۰ تای به صورت تصادفی از جامعه انتخاب میکنیم که دستورات زیر این کار را انجام خواهند داد

```
> r1=sample(1:5043,1)
> dd2=data.frame(movie_metadata[r1,])
> for (i in 1:100)
+ {
+   r1=sample(1:5043,1)
+   d1=data.frame(movie_metadata[r1,])
+   dd2=rbind(dd2,d1)
+ }
> count(dd2,'language')
    language freq
1    English   94
2     French    2
3      Hindi    2
4    Italian    1
5   Mandarin    1
6 Portuguese    1
```

حال برای ایجاد نمونه به صورت bias به این صورت عمل می کنیم که ابتدا فیلم هایی که duration آن ها کمتر از ۷۰ می باشد را انتخاب و از میان آن ها ۱۰۰ مورد اول را انتخاب می کنیم . دستورات زیر این کار را برای ما انجام خواهند داد .

```
> bias = movie_metadata$duration<70
> number=number[1:100,]
```

حال فرکانس تکرار مقادیر آن را به مانند بالا محاسبه می کنیم .

```
> count(number,'language')
  language freq
1             3
2   Danish    1
3  English   83
4   French    1
5  Italian    1
6 Japanese    1
7   Polish    3
8     <NA>    7
```

حال به محاسبه Goodness of Fit برای هر دو نمونه می پردازیم .

نمونه اول (به صورت تصادفی و بدون bias)

	English	French	Hindi	Italian	Mandarin	Portuguese	Other
Expected	93.27%	1.4%	0.55%	0.22%	0.51%	0.15%	3.9%
	92	1	1	1	1	1	3
Observed	94	2	2	1	1	1	0

$$X^2 = \sum \frac{(O-E)^2}{E} = \frac{(94-92)^2}{92} + \frac{(2-1)^2}{1} + \frac{(2-1)^2}{1} + \frac{(1-1)^2}{1} + \frac{(1-1)^2}{1} + \frac{(1-1)^2}{1} + \frac{(0-3)^2}{3}$$

$$= \frac{4}{92} + \frac{1}{1} + \frac{1}{1} + 0 + 0 + 0 + \frac{9}{3} = 5.043$$

$$k = 7 - 1 = 6$$

```
> pchisq(5.043,6,lower.tail = FALSE)
[1] 0.53831
```

نتیجه ای که می توان از دستور بالا داشت به این صورت می باشد که فرض صفر که تفاوت نداشتن بین مقادیر مورد انتظار و مشاهده شده می باشد را رد نمی توان کرد .

نمونه دوم (به صورت bias دار و غیرتصادفی)

	Danish	English	French	Italian	Japanese	Polish	Other
Expected	0.1%	93.27%	1.4%	0.22%	0.3%	0.08%	4.63%
	1	91	1	1	1	1	4

Observed	1	83	1	1	1	3	7+3

$$X^2 = \sum \frac{(O-E)^2}{E} = \frac{(1-1)^2}{1} + \frac{(83-91)^2}{91} + \frac{(1-1)^2}{1} + \frac{(1-1)^2}{1} + \frac{(1-1)^2}{1} + \frac{(3-1)^2}{1} + \frac{(10-4)^2}{4}$$

$$= 0 + \frac{8}{91} + 0 + 0 + 0 + 4 + \frac{36}{4} = 13.8$$

$$k = 7 - 1 = 6$$

```
> pchisq(13.08,6,lower.tail = FALSE)
[1] 0.04178299
```

با توجه به مقداری که p_value دارد می توان این برداشت را داشت که فرض صفر که متفاوت نبودن بین مقادیر مشاهده شده و مقادیر مورد انتظار می باشد رد خواهد شد .

سوال ۵

در این سوال دو متغییر categorical ای که در نظر گرفته شده color و language می باشد که جدول contengency آن ها را در زیر می بینید .

```
> CrossTable(movie_metadata$language,movie_metadata$color,prop.t = TRUE , prop.r = TRUE , prop.c = TRUE)

   Cell Contents
|-------------------------|
|                       N |
| Chi-square contribution |
|           N / Row Total |
|           N / Col Total |
|         N / Table Total |
|-------------------------|

Total Observations in Table:  5043

                    | movie_metadata$color
movie_metadata$language |       Black and White |                 Color |             Row Total |
------------------------|-----------------------|-----------------------|-----------------------|
                        |                     6 |                     3 |                     3 |                    12 |
                        |               784.308 |                12.594 |                 6.243 |                      |
                        |                 0.500 |                 0.250 |                 0.250 |                 0.002 |
                        |                 0.316 |                 0.014 |                 0.001 |                      |
                        |                 0.001 |                 0.001 |                 0.001 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
             Aboriginal |                     0 |                     0 |                     2 |                     2 |
                        |                 0.008 |                 0.083 |                 0.004 |                      |
                        |                 0.000 |                 0.000 |                 1.000 |                 0.000 |
                        |                 0.000 |                 0.000 |                 0.000 |                      |
                        |                 0.000 |                 0.000 |                 0.000 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
                 Arabic |                     0 |                     0 |                     5 |                     5 |
                        |                 0.019 |                 0.207 |                 0.011 |                      |
                        |                 0.000 |                 0.000 |                 1.000 |                 0.001 |
                        |                 0.000 |                 0.000 |                 0.001 |                      |
                        |                 0.000 |                 0.000 |                 0.001 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
                Aramaic |                     0 |                     0 |                     1 |                     1 |
                        |                 0.004 |                 0.041 |                 0.002 |                      |
                        |                 0.000 |                 0.000 |                 1.000 |                 0.000 |
                        |                 0.000 |                 0.000 |                 0.000 |                      |
                        |                 0.000 |                 0.000 |                 0.000 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
                Bosnian |                     0 |                     0 |                     1 |                     1 |
                        |                 0.004 |                 0.041 |                 0.002 |                      |
                        |                 0.000 |                 0.000 |                 1.000 |                 0.000 |
                        |                 0.000 |                 0.000 |                 0.000 |                      |
                        |                 0.000 |                 0.000 |                 0.000 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
              Cantonese |                     0 |                     2 |                     9 |                    11 |
                        |                 0.041 |                 5.230 |                 0.215 |                      |
                        |                 0.000 |                 0.182 |                 0.818 |                 0.002 |
                        |                 0.000 |                 0.010 |                 0.002 |                      |
                        |                 0.000 |                 0.000 |                 0.002 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
                Chinese |                     0 |                     0 |                     3 |                     3 |
                        |                 0.011 |                 0.124 |                 0.006 |                      |
                        |                 0.000 |                 0.000 |                 1.000 |                 0.001 |
                        |                 0.000 |                 0.000 |                 0.001 |                      |
                        |                 0.000 |                 0.000 |                 0.001 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
                  Czech |                     0 |                     0 |                     1 |                     1 |
                        |                 0.004 |                 0.041 |                 0.002 |                      |
                        |                 0.000 |                 0.000 |                 1.000 |                 0.000 |
                        |                 0.000 |                 0.000 |                 0.000 |                      |
                        |                 0.000 |                 0.000 |                 0.000 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
                 Danish |                     0 |                     1 |                     4 |                     5 |
                        |                 0.019 |                 3.033 |                 0.125 |                      |
                        |                 0.000 |                 0.200 |                 0.800 |                 0.001 |
                        |                 0.000 |                 0.005 |                 0.001 |                      |
                        |                 0.000 |                 0.000 |                 0.001 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
                   Dari |                     0 |                     0 |                     2 |                     2 |
                        |                 0.008 |                 0.083 |                 0.004 |                      |
                        |                 0.000 |                 0.000 |                 1.000 |                 0.000 |
                        |                 0.000 |                 0.000 |                 0.000 |                      |
                        |                 0.000 |                 0.000 |                 0.000 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
                  Dutch |                     0 |                     0 |                     4 |                     4 |
                        |                 0.015 |                 0.166 |                 0.009 |                      |
                        |                 0.000 |                 0.000 |                 1.000 |                 0.001 |
                        |                 0.000 |                 0.000 |                 0.001 |                      |
                        |                 0.000 |                 0.000 |                 0.001 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
                Dzongkha |                     0 |                     0 |                     1 |                     1 |
                        |                 0.004 |                 0.041 |                 0.002 |                      |
                        |                 0.000 |                 0.000 |                 1.000 |                 0.000 |
                        |                 0.000 |                 0.000 |                 0.000 |                      |
                        |                 0.000 |                 0.000 |                 0.000 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
                 English |                    12 |                   184 |                  4508 |                  4704 |
                        |                 1.848 |                 0.615 |                 0.062 |                      |
                        |                 0.003 |                 0.039 |                 0.958 |                 0.933 |
                        |                 0.632 |                 0.880 |                 0.936 |                      |
                        |                 0.002 |                 0.036 |                 0.894 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
                Filipino |                     0 |                     1 |                     0 |                     1 |
                        |                 0.004 |                22.171 |                 0.955 |                      |
                        |                 0.000 |                 1.000 |                 0.000 |                 0.000 |
                        |                 0.000 |                 0.005 |                 0.000 |                      |
                        |                 0.000 |                 0.000 |                 0.000 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
                  French |                     0 |                     3 |                    70 |                    73 |
                        |                 0.275 |                 0.000 |                 0.001 |                      |
                        |                 0.000 |                 0.041 |                 0.959 |                 0.014 |
                        |                 0.000 |                 0.014 |                 0.015 |                      |
                        |                 0.000 |                 0.001 |                 0.014 |                      |
------------------------|-----------------------|-----------------------|-----------------------|-----------------------|
```

German	0	5	14	19
	0.072	22.536	0.945	
	0.000	0.263	0.737	0.004
	0.000	0.024	0.003	
	0.000	0.001	0.003	
Greek	0	0	1	1
	0.004	0.041	0.002	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	
Hebrew	0	0	5	5
	0.019	0.207	0.011	
	0.000	0.000	1.000	0.001
	0.000	0.000	0.001	
	0.000	0.000	0.001	
Hindi	0	1	27	28
	0.105	0.022	0.003	
	0.000	0.036	0.964	0.006
	0.000	0.005	0.006	
	0.000	0.000	0.005	
Hungarian	0	0	1	1
	0.004	0.041	0.002	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	
Icelandic	0	0	2	2
	0.008	0.083	0.004	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	

Indonesian	0	0	2	2
	0.008	0.083	0.004	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	
Italian	0	1	10	11
	0.041	0.649	0.024	
	0.000	0.091	0.909	0.002
	0.000	0.005	0.002	
	0.000	0.000	0.002	
Japanese	0	1	17	18
	0.068	0.086	0.002	
	0.000	0.056	0.944	0.004
	0.000	0.005	0.004	
	0.000	0.000	0.003	
Kannada	0	0	1	1
	0.004	0.041	0.002	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	
Kazakh	0	0	1	1
	0.004	0.041	0.002	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	
Korean	0	0	8	8
	0.030	0.332	0.017	
	0.000	0.000	1.000	0.002
	0.000	0.000	0.002	
	0.000	0.000	0.002	

Mandarin	1	3	22	26
	8.306	3.430	0.321	
	0.038	0.115	0.846	0.005
	0.053	0.014	0.005	
	0.000	0.001	0.004	
Maya	0	0	1	1
	0.004	0.041	0.002	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	
Mongolian	0	0	1	1
	0.004	0.041	0.002	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	
None	0	0	2	2
	0.008	0.083	0.004	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	
Norwegian	0	0	4	4
	0.015	0.166	0.009	
	0.000	0.000	1.000	0.001
	0.000	0.000	0.001	
	0.000	0.000	0.001	
Panjabi	0	0	1	1
	0.004	0.041	0.002	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	

Persian	0	0	4	4
	0.015	0.166	0.009	
	0.000	0.000	1.000	0.001
	0.000	0.000	0.001	
	0.000	0.000	0.001	
Polish	0	1	3	4
	0.015	4.198	0.176	
	0.000	0.250	0.750	0.001
	0.000	0.005	0.001	
	0.000	0.000	0.001	
Portuguese	0	0	8	8
	0.030	0.332	0.017	
	0.000	0.000	1.000	0.002
	0.000	0.000	0.002	
	0.000	0.000	0.002	
Romanian	0	0	2	2
	0.008	0.083	0.004	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	
Russian	0	2	9	11
	0.041	5.230	0.215	
	0.000	0.182	0.818	0.002
	0.000	0.010	0.002	
	0.000	0.000	0.002	
Slovenian	0	0	1	1
	0.004	0.041	0.002	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	

Spanish	0	0	40	40
	0.151	1.658	0.086	
	0.000	0.000	1.000	0.008
	0.000	0.000	0.008	
	0.000	0.000	0.008	
Swahili	0	1	0	1
	0.004	22.171	0.955	
	0.000	1.000	0.000	0.000
	0.000	0.005	0.000	
	0.000	0.000	0.000	
Swedish	0	0	5	5
	0.019	0.207	0.011	
	0.000	0.000	1.000	0.001
	0.000	0.000	0.001	
	0.000	0.000	0.001	
Tamil	0	0	1	1
	0.004	0.041	0.002	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	
Telugu	0	0	1	1
	0.004	0.041	0.002	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	
Thai	0	0	3	3
	0.011	0.124	0.006	
	0.000	0.000	1.000	0.001
	0.000	0.000	0.001	
	0.000	0.000	0.001	

Urdu	0	0	1	1
	0.004	0.041	0.002	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	
Vietnamese	0	0	1	1
	0.004	0.041	0.002	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	
Zulu	0	0	2	2
	0.008	0.083	0.004	
	0.000	0.000	1.000	0.000
	0.000	0.000	0.000	
	0.000	0.000	0.000	
Column Total	19	209	4815	5043
	0.004	0.041	0.955	

در گام دوم قصد داریم تا با استفاده از ازمون Chi-square مستقل بودن و یا وابسته بودن این دو متغییر را با یکدیگر بررسی کنیم . برای این منظور داریم

در صورتی که در قطعه کد بالا مقدار chisq=TRUE قرار دهیم ازمون independency را نیز برای ما انجام خواهد داد به این ترتیب داریم

```
> my_table = CrossTable(movie_metadata$language,movie_metadata$color,
        prop.t = TRUE , prop.r = TRUE , prop.c = TRUE , chisq = TRUE)
```

که جواب نهایی برای این سوال به شکل زیر خواهد بود

```
Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  912.9922     d.f. =  94     p =  2.431142e-134
```

همانطور که مشخص می باشد مقدار p_value برای این سوال که مقدار بسیار پایین دارد می توان چنین برداشت کرد که دو متغییر color و language از یکدیگر مستقل نمی باشند و فرض صفر که بیانگر استقلال این دو متغییر بود رد می شود .

سوال ۶

در این سوال تاثیر متغیر های budget , color , imdb_score را بر روی متغییر gross بررسی خواهیم کرد

دستورات لازم برای این کار به صورت زیر می باشد .

```
> data = lm(movie_metadata$gross ~ movie_metadata$color +
        movie_metadata$imdb_score + movie_metadata$budget)
> summary(data)
```

که نتایج بدست امده به شکل زیر می باشد .

```
Call:
lm(formula = movie_metadata$gross ~ movie_metadata$color + movie_metadata$imdb_score +
    movie_metadata$budget)

Residuals:
       Min         1Q     Median         3Q        Max
-419464301  -41865209  -17362723   16729444  682278319

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                        -4.857e+07  4.834e+07  -1.005    0.315
movie_metadata$color Black and White -1.987e+07  4.829e+07  -0.411    0.681
movie_metadata$colorColor             6.612e+06  4.793e+07   0.138    0.890
movie_metadata$imdb_score             1.432e+07  1.036e+06  13.820  < 2e-16 ***
movie_metadata$budget                 2.975e-02  4.888e-03   6.085 1.27e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 67760000 on 3886 degrees of freedom
  (1152 observations deleted due to missingness)
Multiple R-squared:  0.05859,  Adjusted R-squared:  0.05762
F-statistic: 60.47 on 4 and 3886 DF,  p-value: < 2.2e-16
```

معادله خط رگرسیون برای این سوال نیز به شکل زیر خواهد بود

$$gross = (-4.857e+07) - (1.987e+07 * Black\ and\ White) + (6.612e+06 * color) + (1.432e+07 * imdbScore) + (2.975e-02 * budget)$$

عرض از مبدا: بیان می دارد که در صورت صفر بودن همه متغیر های Black and White , color , imdbScore , budget مقدار Gross عدد 4.857e+07- خواهد بود .

شیب Black and White :این مقدار بیان می دارد در صورت ثابت بودن مابقی متغیر ها در صورت یک بودن این متغییر میزان gross به میزان 1.987e+07 واحد کاهش خواهد داشت .

شیب Color : بیان می دارد به ازای یک بودن این مقدار و ثابت بودن باقی متغیر ها مقدار gross به میزان 6.612e+06 واحد افزایش خواهد داشت .

شیب imdbScore : بیان می دارد به ازای یک بودن این مقدار و ثابت بودن باقی متغیر ها مقدار gross به میزان 1.432e+07 واحد افزایش خواهد داشت .

شیب budget : بیان می دارد به ازای یک بودن این مقدار و ثابت بودن باقی متغیر ها مقدار Gross به میزان 2.975e-2 واحد افزایش خواهد داشت .

ازمون فرض برای متغییر budget

$$H_0: \beta_1 = 0$$

$$H_A: \beta_1 \neq 0$$

$$T = \frac{2.975e-02 \ - \ 0}{4.888e-03} = 6.08 \rightarrow p_{value} = 1.27e-9$$

پس با داشتن چنین p-value ای می توان چنین برداشت کرد که فرض صفر رد می شود و شیب برای مقدار متغییر budget عددی مخالف صفر می باشد و در تعیین مقدار gross تاثیر گذار می باشد .

سوال ۷

برای این سوال متغیر هایی که در نظر گرفتیم تاثیر imdb_score , country , genres , duration , Color بر روی متغیر gross می باشد . در بخش اول سعی بر این می باشد که با استفاده از الگوریتم backwards elimination و Adjusted R^2 مل صرفه جو را برای این متغیر ها پیدا کنیم . S

Step	Variable included	Adjusted R^2
Full	Color , duration , country , imdb_score , aspect_ratio	0.1296
Step 1	Duration , country , imdb_score , aspect_ratio	0.1265
	Color , country , imdb_score , aspect_ratio	0.09708
	Color , duration , imdb_score , aspect_ratio	0.07934
	Color , duration , country , aspect_ratio	0.1046
	Color , duration , country , imdb_score	0.1297
Step 2	duration , country , imdb_score	0.1261
	Color , country , imdb_score	0.0909
	Color , duration , country	0.106

در قدم بعدی با استفاده از الگوریتم forward selection و p_value مدل صرفه جو را برای این سوال بدست می آوریم .

Step	Variable included	P_value
Step 1	Color	0.01924
	Duration	2.2e-16>
	Country	0.124<
	Imdb_score	2.2e-16>
	Aspect_ratio	9.9e-6
Step 2	Imdb_score , color	0.670<
	Imdb_score , duration	2.2e-16>
	Imdb_score , country	0.1194<
	Imdb_score , aspect_ratio	2.21e-05
Step 3	Imdb_score , duration , color	0.6474
	Imdb_score , duration , country	0.1343
	Imdb_score , duration , aspect_ratio	0.0224
Step 4	Imdb_score , duration , aspect_ratio , color	0.6980<
	Imdb_score , duration , aspect_ratio , country	0.16489<

پس در نهایت Imdb_score , duration , aspect_ratio بهترین متغیر ها از مجوعه متغیر های اولیه می باشد

سوال ۸

در این سوال نیز تاثیر متغیر های duration , language , imdb_score را بر روی متغیر Categorical color بررسی می کنیم . با استفاده از دستور glm نتایج زیر بدست می آید .

```
> summary(glm(movie_metadata$color ~   movie_metadata$duration +
        movie_metadata$imdb_score + movie_metadata$language ,
        data=movie_metadata ,family=binomial))

Call:
glm(formula = movie_metadata$color ~ movie_metadata$duration +
    movie_metadata$imdb_score + movie_metadata$language, family = binomial,
    data = movie_metadata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.5219   0.0492   0.0637   0.0787   1.1903

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -4.135e+00  1.728e+00  -2.392   0.0168 *
movie_metadata$duration          2.233e-02  1.158e-02   1.929   0.0537 .
movie_metadata$imdb_score        3.433e-01  1.932e-01   1.777   0.0755 .
movie_metadata$languageAboriginal  1.985e+01  1.251e+04   0.002   0.9987
movie_metadata$languageArabic    1.995e+01  7.851e+03   0.003   0.9980
movie_metadata$languageAramaic   1.958e+01  1.773e+04   0.001   0.9991
movie_metadata$languageBosnian   2.039e+01  1.773e+04   0.001   0.9991
movie_metadata$languageCantonese  2.021e+01  5.287e+03   0.004   0.9969
movie_metadata$languageChinese   2.018e+01  1.018e+04   0.002   0.9984
movie_metadata$languageCzech     1.964e+01  1.773e+04   0.001   0.9991
movie_metadata$languageDanish    2.025e+01  7.624e+03   0.003   0.9979
movie_metadata$languageDari      1.988e+01  1.234e+04   0.002   0.9987
movie_metadata$languageDutch     1.966e+01  8.730e+03   0.002   0.9982
movie_metadata$languageDzongkha  1.971e+01  1.773e+04   0.001   0.9991
movie_metadata$languageEnglish   5.760e+00  7.440e-01   7.742 9.81e-15 ***
movie_metadata$languageFilipino  1.945e+01  1.773e+04   0.001   0.9991
movie_metadata$languageFrench    2.005e+01  2.048e+03   0.010   0.9922
movie_metadata$languageGerman    1.970e+01  3.888e+03   0.005   0.9960
movie_metadata$languageGreek     2.010e+01  1.773e+04   0.001   0.9991
movie_metadata$languageHebrew    2.001e+01  7.896e+03   0.003   0.9980
movie_metadata$languageHindi     1.937e+01  3.329e+03   0.006   0.9954
movie_metadata$languageHungarian  1.927e+01  1.773e+04   0.001   0.9991
movie_metadata$languageIcelandic  1.943e+01  1.027e+04   0.002   0.9985
movie_metadata$languageIndonesian  1.978e+01  1.254e+04   0.002   0.9987
movie_metadata$languageItalian   2.000e+01  5.192e+03   0.004   0.9969
movie_metadata$languageJapanese  1.990e+01  3.957e+03   0.005   0.9960
movie_metadata$languageKazakh    2.014e+01  1.773e+04   0.001   0.9991
movie_metadata$languageKorean    1.946e+01  6.136e+03   0.003   0.9975
movie_metadata$languageMandarin  2.702e+00  1.241e+00   2.178   0.0294 *
movie_metadata$languageMaya      1.892e+01  1.773e+04   0.001   0.9991
movie_metadata$languageMongolian  1.938e+01  1.773e+04   0.001   0.9991
movie_metadata$languageNone      1.973e+01  1.251e+04   0.002   0.9987
movie_metadata$languageNorwegian  2.013e+01  8.838e+03   0.002   0.9982
movie_metadata$languagePanjabi   1.929e+01  1.773e+04   0.001   0.9991
movie_metadata$languagePersian   1.989e+01  8.776e+03   0.002   0.9982
```

```
movie_metadata$languagePolish      2.063e+01  8.773e+03   0.002   0.9981
movie_metadata$languagePortuguese  1.968e+01  6.192e+03   0.003   0.9975
movie_metadata$languageRomanian    1.985e+01  1.246e+04   0.002   0.9987
movie_metadata$languageRussian     2.034e+01  5.162e+03   0.004   0.9969
movie_metadata$languageSlovenian   2.065e+01  1.773e+04   0.001   0.9991
movie_metadata$languageSpanish     2.005e+01  2.756e+03   0.007   0.9942
movie_metadata$languageSwahili     2.082e+01  1.773e+04   0.001   0.9991
movie_metadata$languageSwedish     1.939e+01  7.620e+03   0.003   0.9980
movie_metadata$languageTamil       1.949e+01  1.773e+04   0.001   0.9991
movie_metadata$languageTelugu      1.827e+01  1.773e+04   0.001   0.9992
movie_metadata$languageThai        1.941e+01  9.488e+03   0.002   0.9984
movie_metadata$languageVietnamese  1.915e+01  1.773e+04   0.001   0.9991
movie_metadata$languageZulu        1.993e+01  1.251e+04   0.002   0.9987
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 238.70  on 5027  degrees of freedom
Residual deviance: 183.14  on 4980  degrees of freedom
  (15 observations deleted due to missingness)
AIC: 279.14

Number of Fisher Scoring iterations: 19
```

نتایج بالا نشان می دهد که به دلیل اینکه در متغیر دسته ای زبان English دارای *** می باشد دلیل بر اهمیت این متغیر و غیر قابل حذف بودن آن دارد . همچنین سایر متغیر ها (duration , imdb_score) نیر به دلیل اهمیتی که دارند و مقدار p_value پایینی که دارند قابلین حذف شدن ندارند .

با توجه به مقادیر بالا می توان نتایج زیر را در نظر داشت .

در صورت صفر بودن همه مقادیر متغیر ها مقدار gross برابر 4.135- می باشد .

در هر واحد افزایش duration در صورت تغییر نکردن باقی متغیر ها میزان gross به میزان 2.233e-02 افزایش خواهد داشت .

در هر واحد افزایش imdb_score در صورت تغییر نکردن باقی متغیر ها میزان gross به میزان 3.433e-01 افزایش خواهد داشت .

برای متغییر language نیز می توان چنین نگاهی داشت به این ترتیب برای مثال زبان Aboriginal در صورت ثابت بودن باقی مقادیر و یک بودن مقدار Aboriginal مقدار gross به میزان 1.985e+02 افزایش خواهد داشت .

باقی زبان ها نیز چنین ویژگی ای دارند و چنین برداشتی از آن ها می توان داشت .

سوال ۹

در این دیتاست که شامل اطلاعات پایه از یک نمونه فیلم های IMDB بود نتایج جالبی که می توان انتظار داشت و البته برای خود من نیز هیجان انگیز بود بررسی اطلاعات بدست آمده از دستور زیر می باشد .

```
summary(lm(movie_metadata$imdb_score ~ movie_metadata$country +
        movie_metadata$genres + movie_metadata$language +
        movie_metadata$director_name))
```

یعنی بررسی امتیاز IMDB بر اساس کشور سازنده ژانر فیلم زبان و تهیه کننده فیلم . که نتایج جالب توجهی داشت که در زیر به برخی از آن ها اشاره خواهم کرد

فیلم هایی که توسط ایران و افغانستان تولید شده باشند یکی از مهم ترین کشور هایی می باشد که باعث می شود نتوان country را از این لیست حذف کرد و باعث تغییر امتیاز IMDB مورد نظر می شود یعنی Estimated بالایی دارد .

```
movie_metadata$countryIndonesia   1.0604704  1.1065087  0.958 0.337987
movie_metadata$countryIran        3.8698904  1.4312584  2.704 0.006916 **
movie_metadata$countryIreland     1.7272468  1.0226318  1.689 0.091380 .
movie_metadata$countryIsrael      1.7284161  1.3933526  1.240 0.214955
```

مورد بعدی که جلب توجه می کند ژانر های تاثیر گذار بر روی IMDB_score می باشد که در شکل زیر مهم ترین آن ها آورده شده است .

```
movie_metadata$genresAction|Adventure|Comedy|Family|Sci-Fi      -1.5074007  0.5488822  -2.746 0.006084 **
movie_metadata$genresAction|Adventure|Comedy|Fantasy            -0.2624201  0.6751267  -0.389 0.697544
movie_metadata$genresAction|Adventure|Comedy|Fantasy|Mystery    -0.3797231  1.0034792  -0.378 0.705171
movie_metadata$genresAction|Adventure|Comedy|Fantasy|Romance    -1.5961511  0.8291758  -1.925 0.054381 .
movie_metadata$genresAction|Adventure|Comedy|Fantasy|Sci-Fi     -5.3961511  0.8291758  -6.508 9.74e-11 ***
movie_metadata$genresAction|Adventure|Comedy|Fantasy|Thriller   -1.8037118  0.9299792  -1.940 0.052587 .
movie_metadata$genresAction|Adventure|Comedy|Music|Thriller     -3.0961511  0.8291758  -3.734 0.000194 ***
movie_metadata$genresAction|Adventure|Comedy|Musical            -2.0961511  0.8291758  -2.528 0.011553 *
movie_metadata$genresAction|Adventure|Comedy|Romance            -0.9178920  0.7454658  -1.231 0.218364
movie_metadata$genresAction|Adventure|Comedy|Romance|Sci-Fi     -2.6961511  0.8291758  -3.252 0.001168 **
```

همان طور که در شکل نیز مشخص است ژانر های Action , Adventure , comedy , fantasy , Sci-Fi و ژانر Action , adventure , comedy , music , Thriller از پر اهمیت ترین ژانر هایی می باشد که باعث می شود امتیاز فیلم در IMDB پایین بیاید و حتی در شکل زیر نمونه ای دیگر از این ژانر ها را می توان مشاهده کرد

```
movie_metadata$genresAction|Adventure|History|Western           -2.8961511  0.8291758  -3.493 0.000489 ***
movie_metadata$genresAction|Adventure|Horror|Sci-Fi             -0.0897411  0.5769624  -0.156 0.876412
movie_metadata$genresAction|Adventure|Horror|Sci-Fi|Thriller    -0.6736461  0.7020467  -0.960 0.337407
movie_metadata$genresAction|Adventure|Horror|Thriller           -1.2269754  1.1113037  -1.104 0.269696
movie_metadata$genresAction|Adventure|Mystery|Romance|Thriller  -2.9961511  0.8291758  -3.613 0.000310 ***
movie_metadata$genresAction|Adventure|Mystery|Sci-Fi            -1.0634570  0.7313194  -1.454 0.146068
```

نکته جالب توجه دیگر اهمیت زبان فارسی می باشد که در شکل زیر این مورد تا حدودی مشخص می باشد

```
movie_metadata$languageNorwegian  -0.664034   0.881915  -0.753 0.451550
movie_metadata$languagePanjabi    -1.462983   0.883827  -1.655 0.097987 .
movie_metadata$languagePersian    -2.162983   0.883827  -2.447 0.014459 *
movie_metadata$languagePolish      0.408295   0.574353   0.711 0.477223
```

که همانطور که مشخص می باشد باعث خواهد شد تا امتیاز فیلم پایین بیاید !!! .

در مورد تهیه کنندگان نیز اکثر تهیه کنندگانی که بر روی امتیاز فیلم تاثیر دارند تاثیر منفی می گذارند . تنها تعداد محدودی از تهیه کنندگان می باشند که باعث می شوند امتیاز فیلم بالا برود از این تهیه کنندگان می توان به Amal Al-Agroobi اشاره کرد که باز هم نکته جالب توجهی می باشد .

تحلیل : در کل می توان پس از بررسی های انجام شده به این نتیجه رسید که اهمیت بعضی پارامتر ها بر روی دیگر پارامتر ها می تواند بیشتر یا کمتر باشد همچنین گاها متغییر ها می توانند با هم correlation داشته باشند که این موضوع نیز در این پروژه بررسی و تحلیل شد . گاها مقادیر مشاهده شده با انچه مورد انتظار می باشد متفاوت می باشد و این نشان از اهیمت بررسی اماری و تحلیل آن ها دارد .