به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

درس استنباط آماری

پروژه اول

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 .

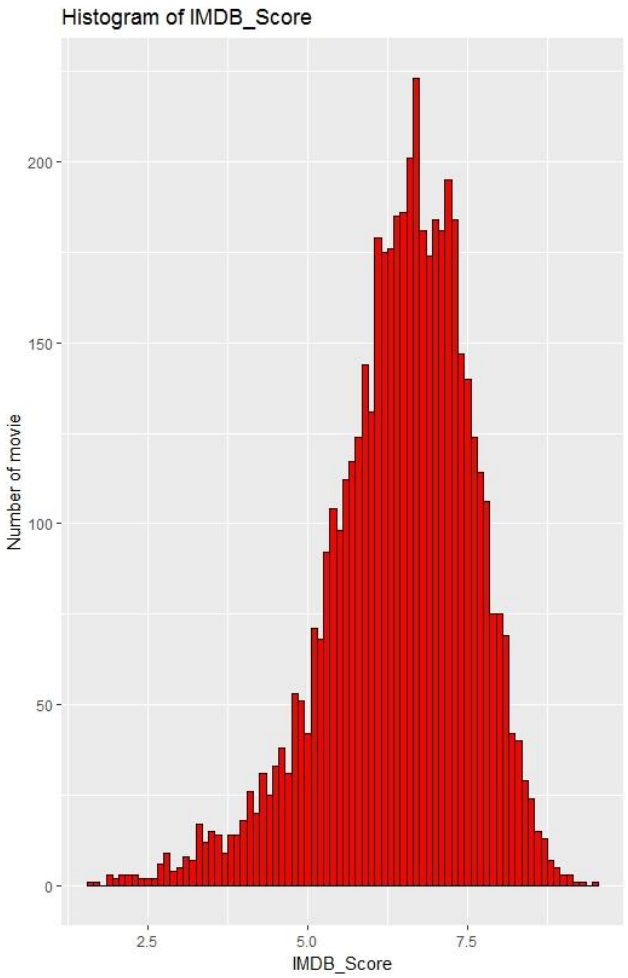
برای فراخوانی دیتاست مورد نظر دستورات زیر را باید در Rstudio وارد کنیم

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

سوال اول

برای این سوال متغییر IMDB_Score را انتخاب کردم که موارد خواسته شده به شکل زیر می باشد .

```
> ggplot(data=movie_metadata, aes(movie_metadata$imdb_score)) +
      geom_histogram(binwidth = 0.1 , fill="red" , col="black")+
      labs(title="Histogram of IMDB_Score" , x="IMDB_Score" , y="Number of movie")
```



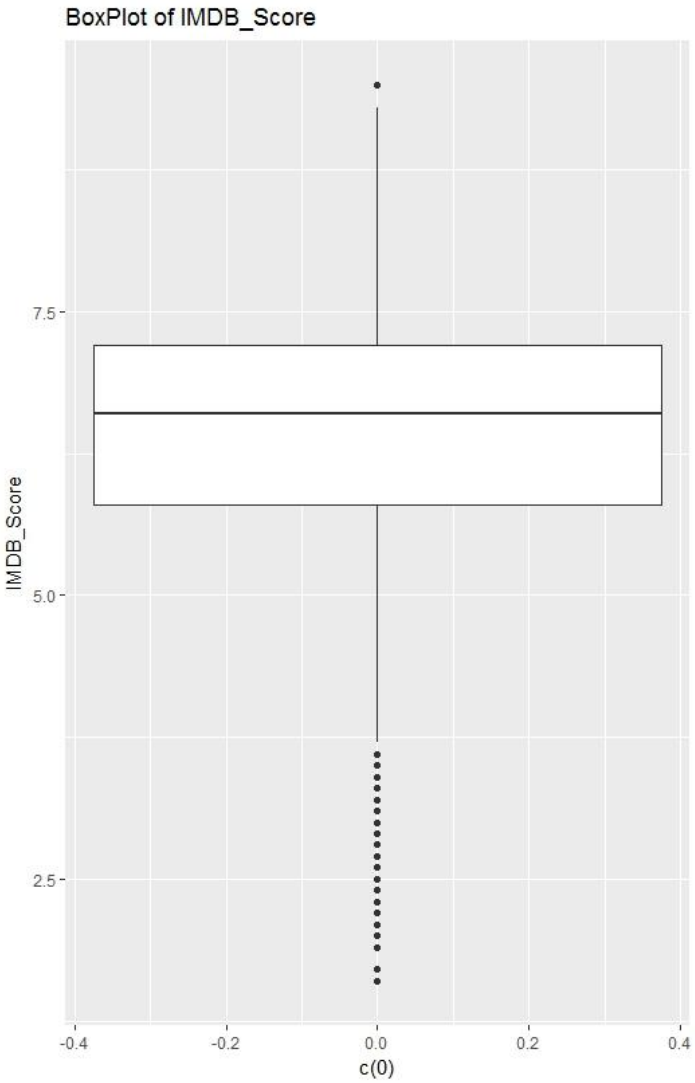
که به نظر می رسد با توجه به نمودار بالا bimodal بودن نمودار امری واضح باشد .

برای بدست آوردن میزان چلوگی می توان از R کمک گرفت و با استفاده از دستورات زیر میزان چولگی نمودار را بدست آورد .

```
> install.packages('e1071', dependencies=TRUE)
> library(e1071)
> skewness(movie_metadata$imdb_score)
[1] -0.7410303
```

که تحلیل ما از اندازه و علامت آن نشان دهنده چولگی به چپ (left skew) دارد که مقدار آن نیز ۰,۷- می باشد .

```
> ggplot(data=movie_metadata, aes(y=movie_metadata$imdb_score , x= c(0))) +
      geom_boxplot()+
      labs(title="BoxPlot of IMDB_Score" ,y="IMDB_Score")
```

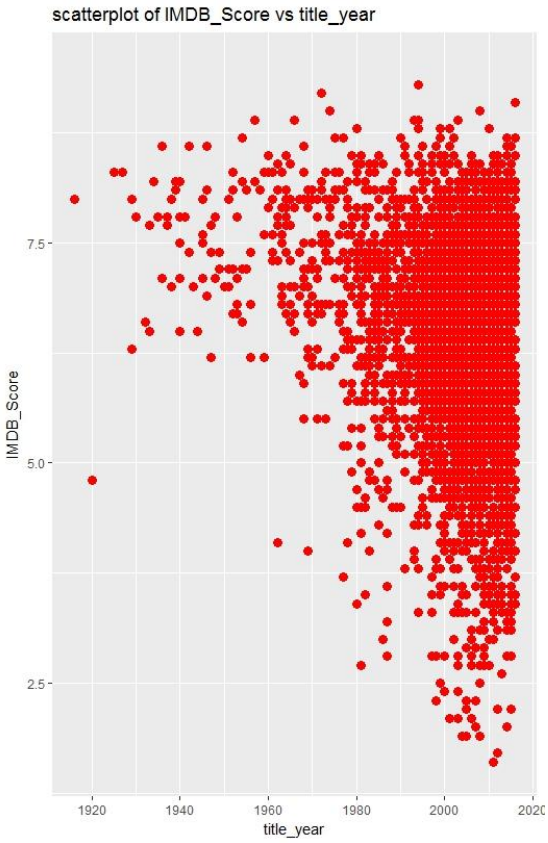


همانطور که در شکل نیز مشخص است تعداد outlier های موجود در این متغییر ۲۱ عدد می باشد که با دایره های توپر مشکی مشخص شده اند .

سوال ۲

برای این سوال دو متغییر عددی ای که برای مقایسه و رسم Scatterplot در نظر گرفته شده است IMDB_Score و title_year می باشد . برای این منظور دستور زیر قادر به انجام چنین کاری می باشد .

```
> ggplot(data=movie_metadata, aes(y=movie_metadata$imdb_score , x=movie_metadata$title_year)) +
 geom_point(col="red" , size=3)+
labs(title="scatterplot of IMDB_Score vs title_year" , y="IMDB_Score" , x="title_year")
```



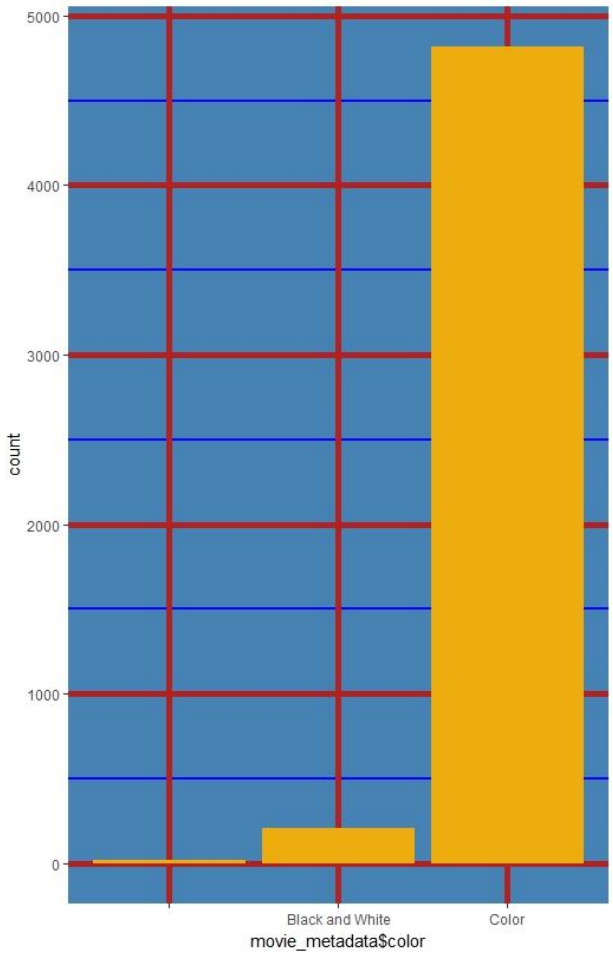
اولین برداشتی که از این نمودار می توان داشت بدین صورت است که در سال های اولیه تعداد فیلم های تولید بسیار محدود بوده ولی در سال های اخیر این تعداد به مراتب بسیار بیشتر شده است . همچنین فیلم هایی که در سال های ابتدایی تولید شده است از کیفیت مطلوبی برخوردار است ولی در سال های اخیر تنوع کیفیت فیلم ها بسیار بالا تر رفته است و همه نوع فیلمی قابل مشاهده است . یکی از دلایل این مسئله می تواند بدین صورت باشد که به دلیل هزینه های بالای فیلمبرداری فیلم ها پس از بررسی های زیاد و مشخص شدن با کیفیت بودن و مخاطب پسند بودن آن ها شروع به تولید آن ها می شده است.

سوال ۳

```
> table(movie_metadata$color)/length(movie_metadata$color)*100

                Black and White           Color
      0.3767599       4.1443585      95.4788816
```

در شکل بالا frequency table را برای متغییر country رسم کردیم که به شکل بالا می باشد .

```
> plot1 = ggplot(movie_metadata , aes(x=movie_metadata$color))+
geom_bar(fill="darkgoldenrod2")+
theme(panel.background = element_rect(fill ='steelblue'),panel.grid.major = element_line(colour
= "firebrick",size=2),panel.grid.minor = element_line(colour = "blue",size=1))
> print(plot1)
```



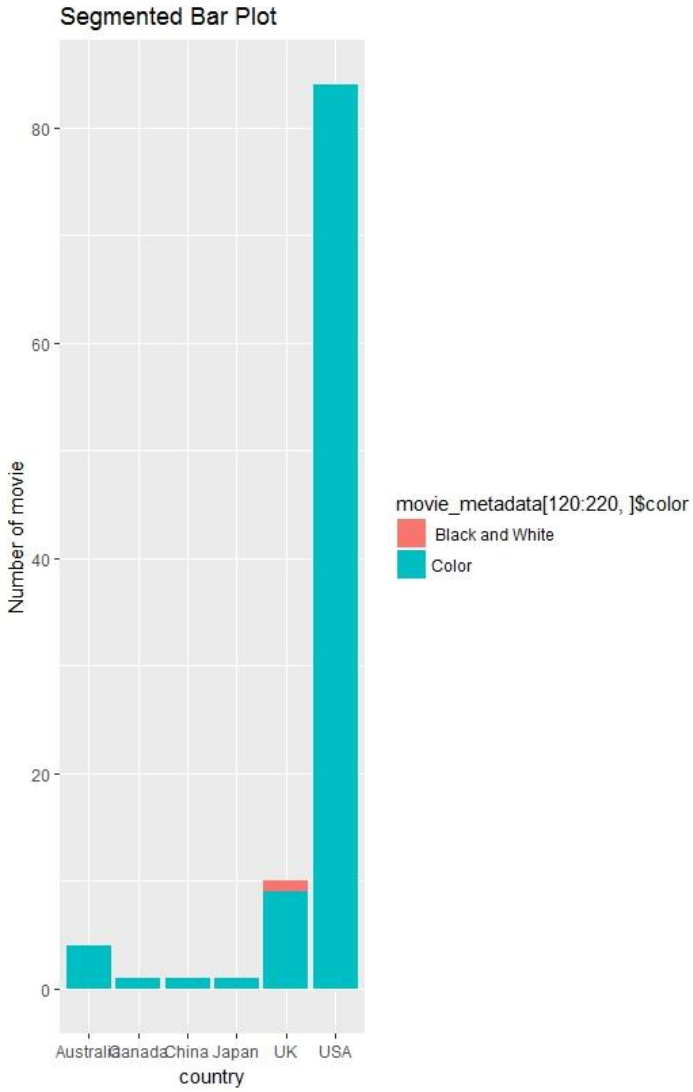
سوال ۴

```
> CrossTable(movie_metadata[120:220,]$color, movie_metadata[120:220,]$country, 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:  101

                            | movie_metadata[120:220, ]$country
movie_metadata[120:220, ]$color | Australia |    Canada |     China |     Japan |        UK |       USA | Row Total |
-------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                Black and White |         0 |         0 |         0 |         0 |         1 |         0 |         1 |
                                |     0.040 |     0.010 |     0.010 |     0.010 |     8.199 |     0.832 |           |
                                |     0.000 |     0.000 |     0.000 |     0.000 |     1.000 |     0.000 |     0.010 |
                                |     0.000 |     0.000 |     0.000 |     0.000 |     0.100 |     0.000 |           |
                                |     0.000 |     0.000 |     0.000 |     0.000 |     0.010 |     0.000 |           |
-------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                        Color |         4 |         1 |         1 |         1 |         9 |        84 |       100 |
                                |     0.000 |     0.000 |     0.000 |     0.000 |     0.082 |     0.008 |           |
                                |     0.040 |     0.010 |     0.010 |     0.010 |     0.090 |     0.840 |     0.990 |
                                |     1.000 |     1.000 |     1.000 |     1.000 |     0.900 |     1.000 |           |
                                |     0.040 |     0.010 |     0.010 |     0.010 |     0.089 |     0.832 |           |
-------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
                  Column Total |         4 |         1 |         1 |         1 |        10 |        84 |       101 |
                                |     0.040 |     0.010 |     0.010 |     0.010 |     0.099 |     0.832 |           |
-------------------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
```

در این سوال دو متغییر color– categorical و country را در جدول contingency اوردیم که نتایج آن در شکل بالا قابل مشاهده می باشد . البته جدول Crosstable حاوی اطلاعات بیشتری نیز می باشد که در legend جدول اورده شده است .

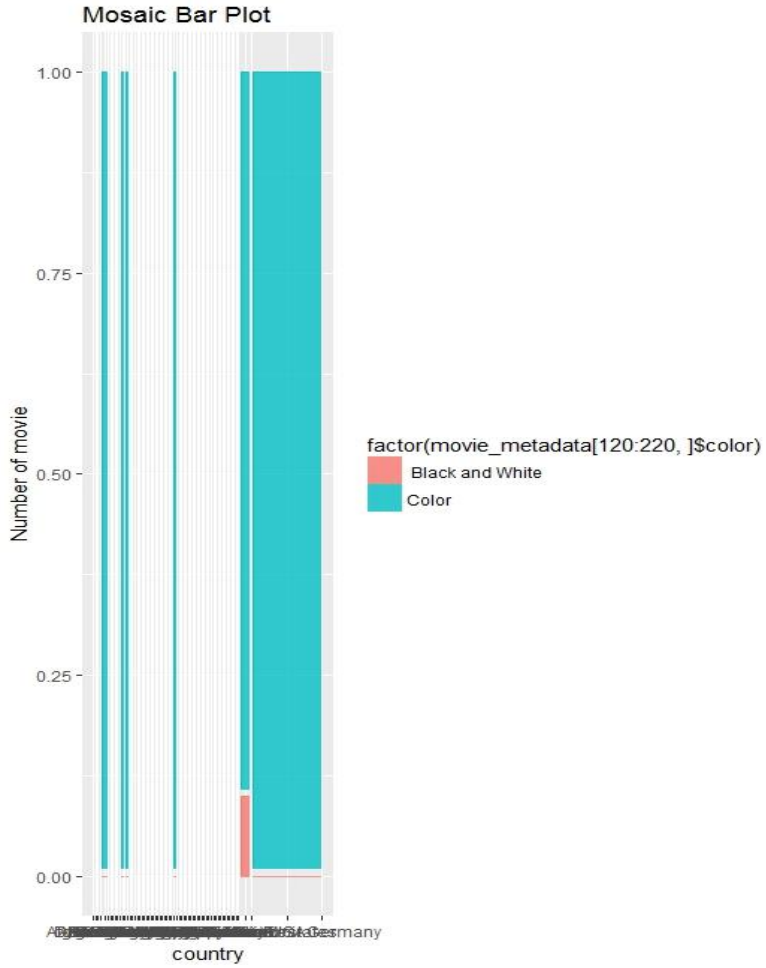
در قسمت بعدی برای این دو متغییر نمودار segmented Bar را رسم خواهیم کرد .

```
> plot1= ggplot(movie_metadata[120:220,] , aes(x=movie_metadata[120:220,]$country))+
       geom_bar(aes(fill=movie_metadata[120:220,]$color))+
       labs(title="Segmented Bar Plot", x="country" ,y="Number of movie")
> print(plot1)
```



در قسمت بعدی دستورات و شکل مربوط به mosaic plot را خواهیم دید .

```
> ggplot(data = movie_metadata[120:220,]) +
   geom_mosaic(aes(weight = 2, x = product(movie_metadata[120:220,]$country),
   fill=factor(movie_metadata[120:220,]$color)))+
   labs(title="Mosaic Bar Plot", x="country" ,y="Number of movie")
```



سوال ۵

```
> mean(na.omit(movie_metadata$duration))
[1] 107.2011
> qnorm(0.005)
[1] -2.575829
> var(na.omit(movie_metadata$duration))
[1] 634.911
> sqrt(var(na.omit(movie_metadata$duration)))
[1] 25.19744
> length(na.omit(movie_metadata$duration))
[1] 5028
```

$$CI = \bar{X} \pm z^* SE = 107.2 \pm 2.57\left(\frac{25.19}{\sqrt{5028}}\right) = 107.2 \pm 0.93$$

سوال ۶

$$H_0: \mu = 105$$

$$H_A: \mu \neq 105$$

حال موارد مورد نظر را برای بدست اوردن P_value محاسبه خواهیم کرد .

طبق مقادیری که در سوال بالا بدست آورده شد داریم

$$s = 25.19 \, . n = 5028 \rightarrow SE = 0.35$$

$$T = \frac{107.2 - 105}{0.35} = 6.28$$

$$P(t > 6.28) = 0$$

```
> pt(6.28,df=5027,lower.tail = FALSE)*2
[1] 3.67071e-10
```

پس با توجه به پایین بودن مقدار P_value می توان نتیجه گرفت که فرض اول رد خواهد شد .

```
> power.t.test(n=5028,delta = 107.2 , sd=25.19 , sig.level = 0.05, type = "one.sample"
,alternative = "two.sided")

     One-sample t test power calculation

              n = 5028
          delta = 107.2
             sd = 25.19
      sig.level = 0.05
          power = 1
    alternative = two.sided
```

سوال ۷

برای این سوال دو متغییر duration و num_critic_for_review را انتخاب کرده ام . با دستورات زیر اندازه این دو متغییر را بررسی می کنیم .

```
> length(na.omit(movie_metadata$imdb_score))
[1] 5043
> length(na.omit(movie_metadata$duration))
[1] 5028
```

همانطور که از دستورات زیر قابل دیدن می باشد اندازه این دو متغییر با یکدیگر متفاوت می باشد .

حال فرضیات را برای این دو متغییر می نویسیم .

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_A: \mu_1 - \mu_2 \neq 0$$

حال به بررسی شرایط ، رد یا عدم رد شدن فرض صفر خواهیم پرداخت .

شرایط مورد نیاز برای این مسئله

استقلال:

درون گروهی

به دلیل اینکه که تعداد نمونه های ما کمتر از ۱۰٪ جامعه فیلمی IMDB می باشد بنابراین استقلال درون گروهی وجود دارد

در مورد تصادفی بودن این نمونه باید توجه داشت که فرض بر این قرار داده شده است که نمونه ای که در این دیتاست در اختیار داریم چنین شرطی را داشته باشد تا بتوانیم در ادامه با این دیتاست کار کنیم .

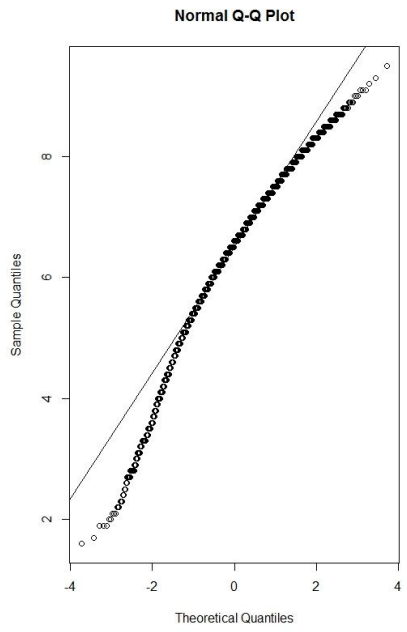
بین گروهی

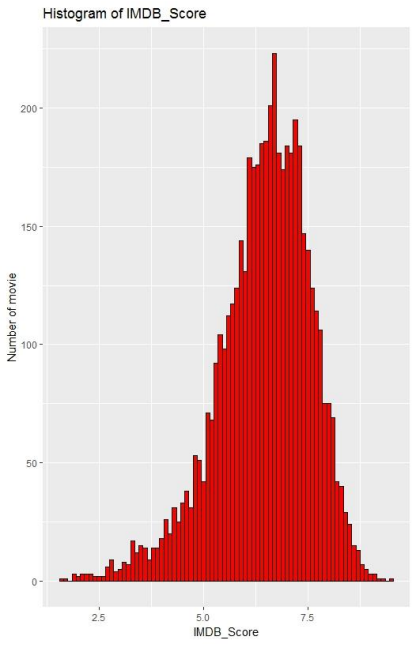
در این مورد چنین استقلالی نیز برقرار می باشد و دو متغییری که انتخاب شده است دارای چنین خصلتی می باشند .

اندازه نمونه و چولگی:

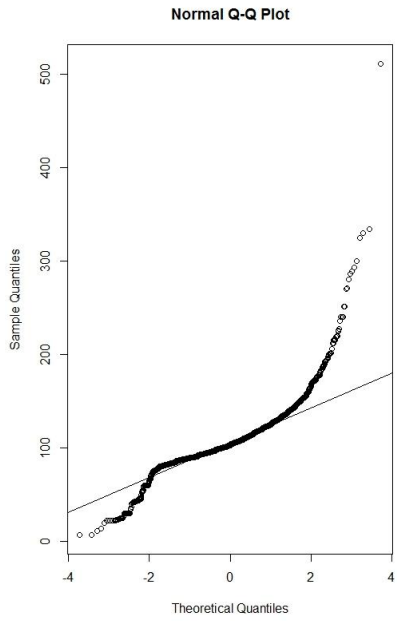
این شرط نیز برقرار می باشد زیرا اندازه این نمونه به اندازه کافی بزرگ است که بتواند به نمودار نرمال نزدیک باشد و چولگی پایینی داشته باشد . به بررسی این شرط با استفاده از qqnorm() و qqline() خواهیم پرداخت

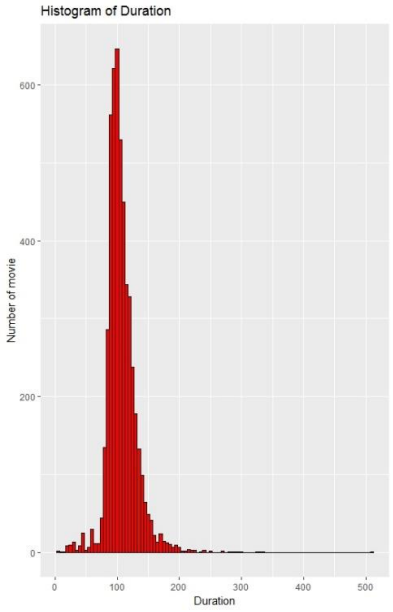
```
> qqnorm(movie_metadata$imdb_score)
> qqline(movie_metadata$imdb_score)
```






```
> qqnorm(movie_metadata$duration)
> qqline(movie_metadata$duration)
```





حال پس از این مراحل به بررسی فرضیات می پردازیم

مقدار میانگین را در ابتدا بدست می آوریم

```
> mean(na.omit(movie_metadata$imdb_score))
[1] 6.442138
```

```
> mean(na.omit(movie_metadata$duration))
[1] 107.2011
```

برای انجام این کار باید انحراف از معیار را بدست بیاوریم که دستوارت زیر این کار را برای ما انجام خواهند داد .

```
> sqrt(var(na.omit(movie_metadata$imdb_score)))
[1] 1.125116
```

```
> sqrt(var(na.omit(movie_metadata$duration)))
[1] 25.19744
```

حال با داشتن اندازه نمونه و انحراف از معیار می توانیم SE را نیز محاسبه کنیم

فرمول محاسبه SE به شکل زیر می باشد .

$$SE_{(\bar{x}_1-\bar{x}_2)} = \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$$

```
> SE=sqrt(((var(na.omit(movie_metadata$imdb_score)))/5043)+
        ((var(na.omit(movie_metadata$imdb_score)))/5028))
> SE
[1] 0.02242288
```

همچنین برای میزان درجه آزادی نیز طبق فرمولی که از قبل داشتیم

$$df = min(n_1 - 1, n_2 - 1)$$

خواهیم داشت $df = 5027$

پس از بدست آورین مقادیر خواسته شده به محاسبه T خواهیم پرداخت

$$T = \frac{(107.2 - 6.44) - 0}{0.022} = 4580$$

```
> pt(4580 , df =5027 , lower.tail = FALSE)*2
[1] 0
```

که با توجه به کمتر بودن این مقدار از ۰,۰۵ می توان اینطور برداشت داشت که فرض صفر رد خواهد شد و میانگین این دو متغییر با یکدیگر برابر نخواهند بود

برای محاسبه power نیز دستورات زیر را وارد و نتایج زیر بدست امده است

```
> power.t.test(n=5028 , delta = 100.76 ,sd = 0.022 ,sig.level = 0.05 , alternative = "
two.sided" )

     Two-sample t test power calculation

              n = 5028
          delta = 100.76
             sd = 0.022
      sig.level = 0.05
          power = 1
    alternative = two.sided

NOTE: n is number in *each* group
```

سوال ۸

در این قسمت از سوال باید میانگین چند گروه (بیش از دو گروه) مختلف را با یکدیگر مقایسه کنیم که برای این کار باید از تکنیک ANOVA استفاده کنیم .برای این بخش متغییر های عددی imdb_score و Duration و director_facebook_likes را با یکدیگر مقایسه میکنیم . فرض اولیه و ثانویه برای این بخش از سوال به شکل زیر می باشد .

$$H_0: \mu_1 = \mu_2 = \mu_2$$

H_A: حداقل یکی از میانگین گروه ها با یکدیگر متفاوت باشد

```
> data_anova = data.frame(cbind(movie_metadata$duration,movie_metadata$director_facebo
ok_likes,movie_metadata$imdb_score))
> data_stack = stack(data_anova)
> aov_test = aov(aov(values~ind,data=data_stack))
> summary(aov_test)
               Df    Sum Sq   Mean Sq F value Pr(>F)
ind             2 1.340e+09 669915718   257.2 <2e-16 ***
Residuals   15007 3.909e+10   2604556
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
119 observations deleted due to missingness
```

با مشاهده دستوارت بالا و نتایج بدست آمده از آن ها می بینیم که مقدار pr بسیار کوچک می باشد و بر همین اساس فرض صفر یعنی برابر بودن میانگین این سه گروه با یکدیگر رد خواهد شد .