Statistics final project presentation

電影票房分析

404410020 呂振麒
404410030 鄭光宇

結果

都滿足 → ７０％

預算大於七千一萬(鎂)

3,5,6,7,11,12月上映

youtube瀏覽大於10.9萬

有分級

FB讚大於4萬

資料來源

爬蟲：影片觀看次數

API：按讚數

爬蟲：各式電影資料

資料整理、分析

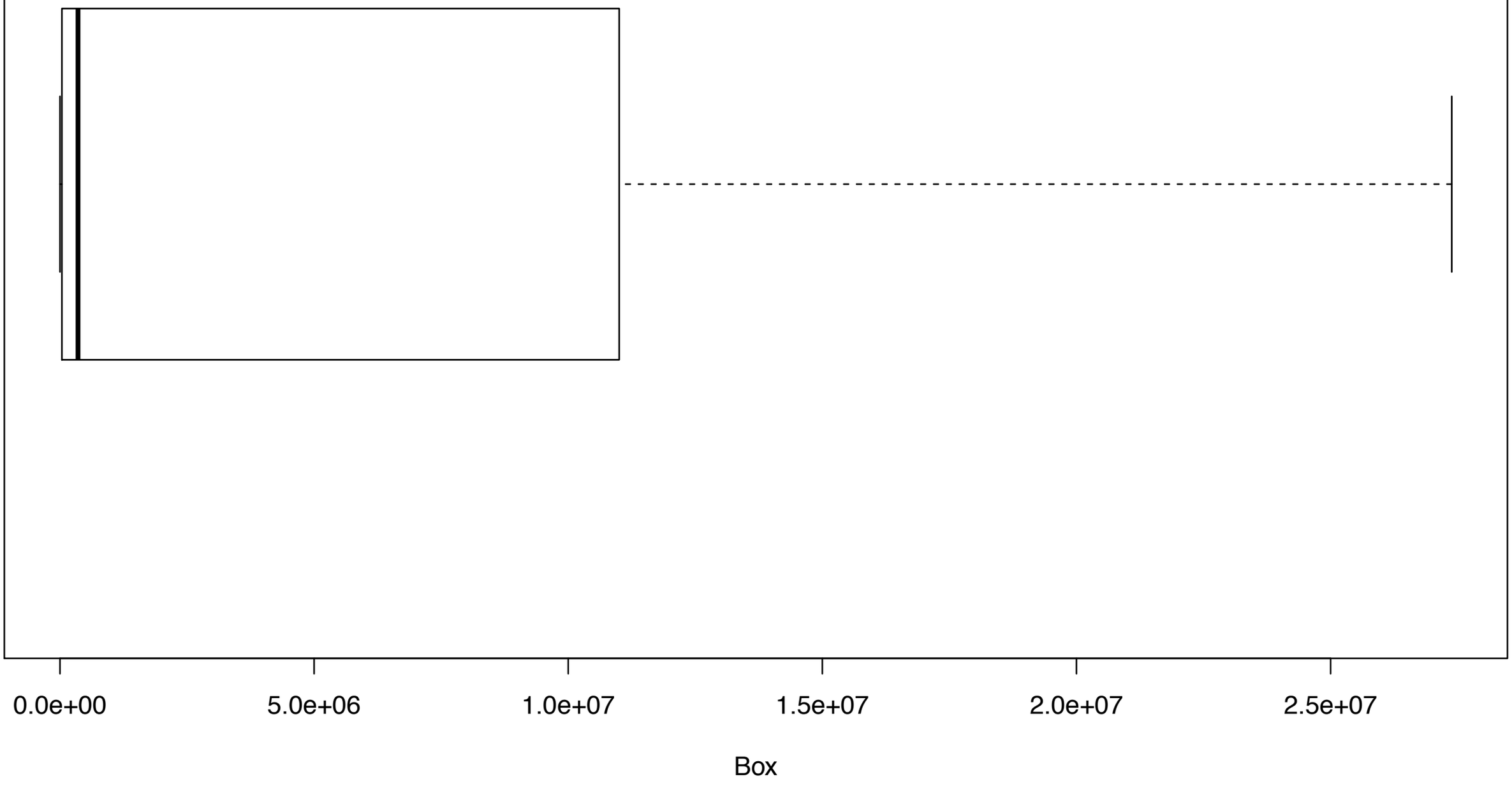
預算 類型 導演

電影長度 上映檔期

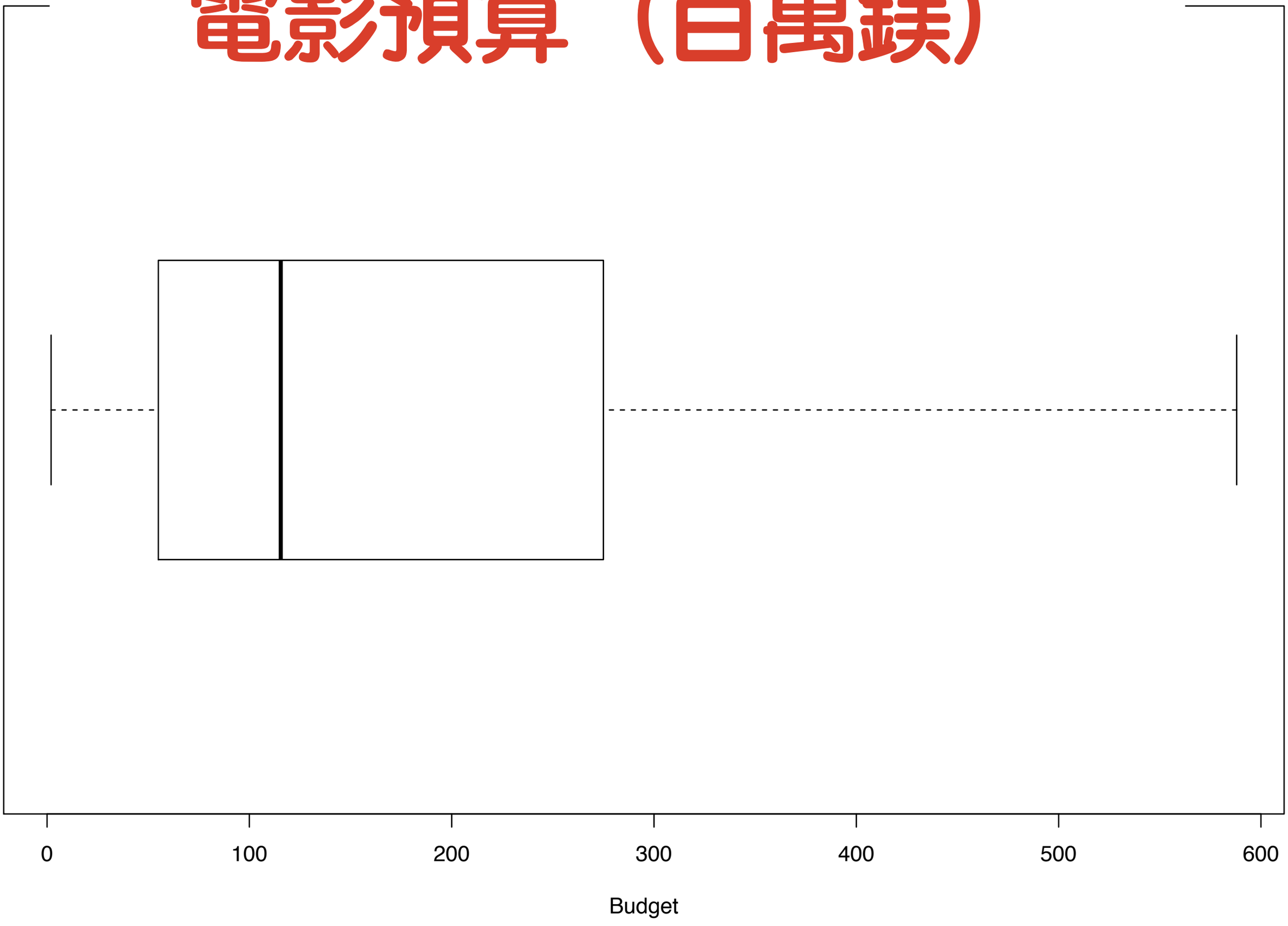
ＦＢ Youtube

BoxPlot BarPlot Decision Tree

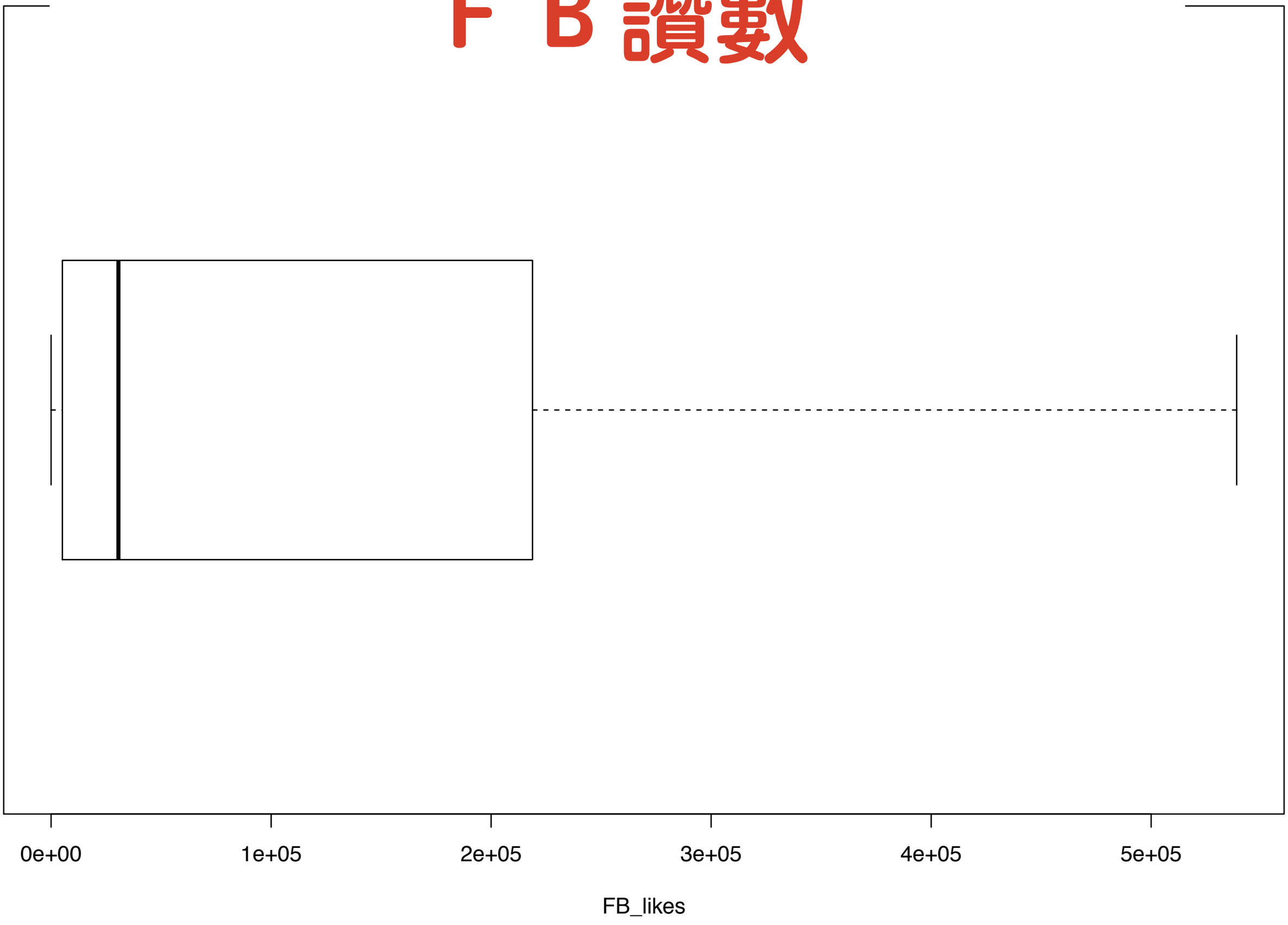
票房分佈（百萬鎂）



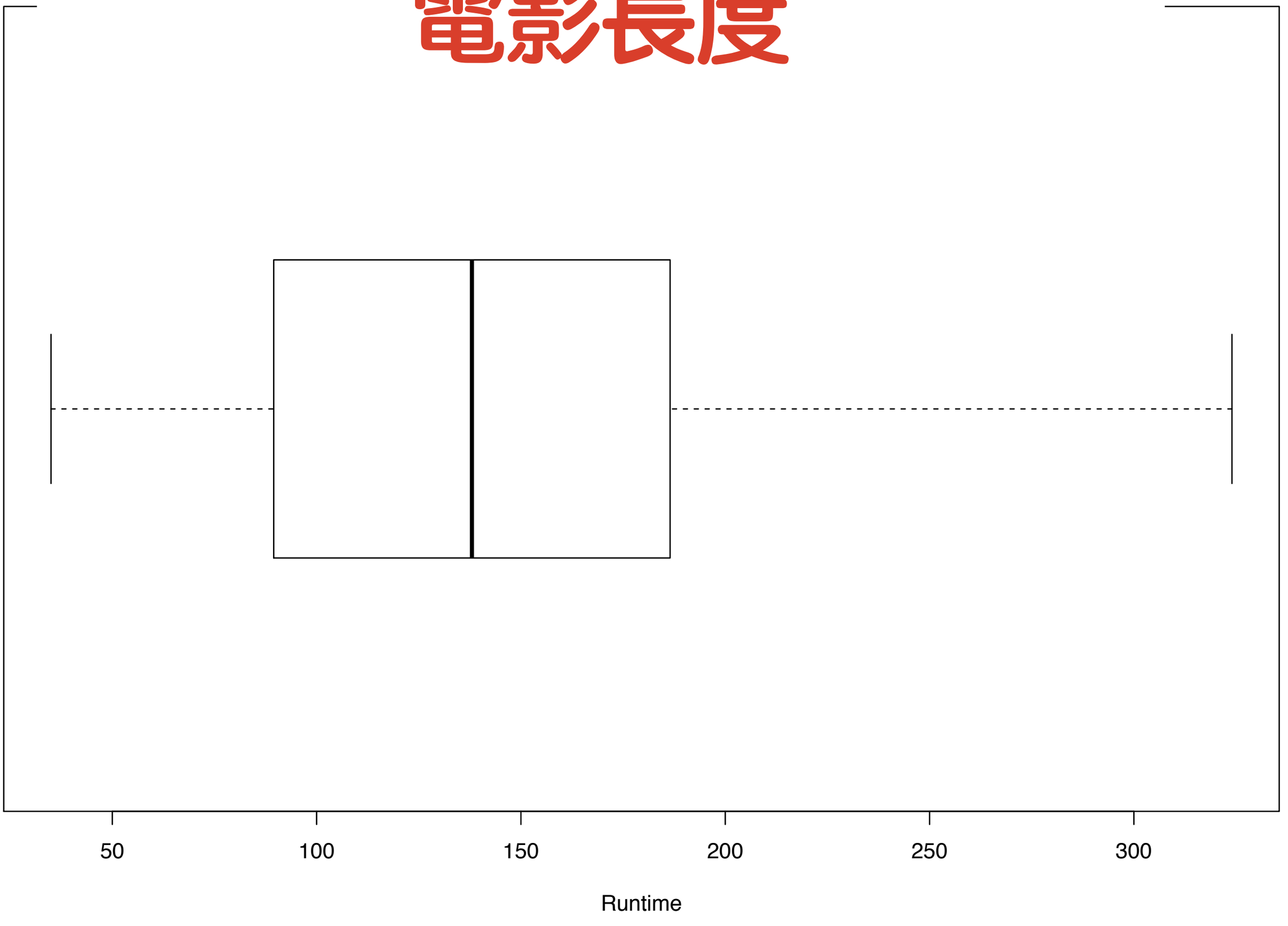
電影預算（百萬鎂）



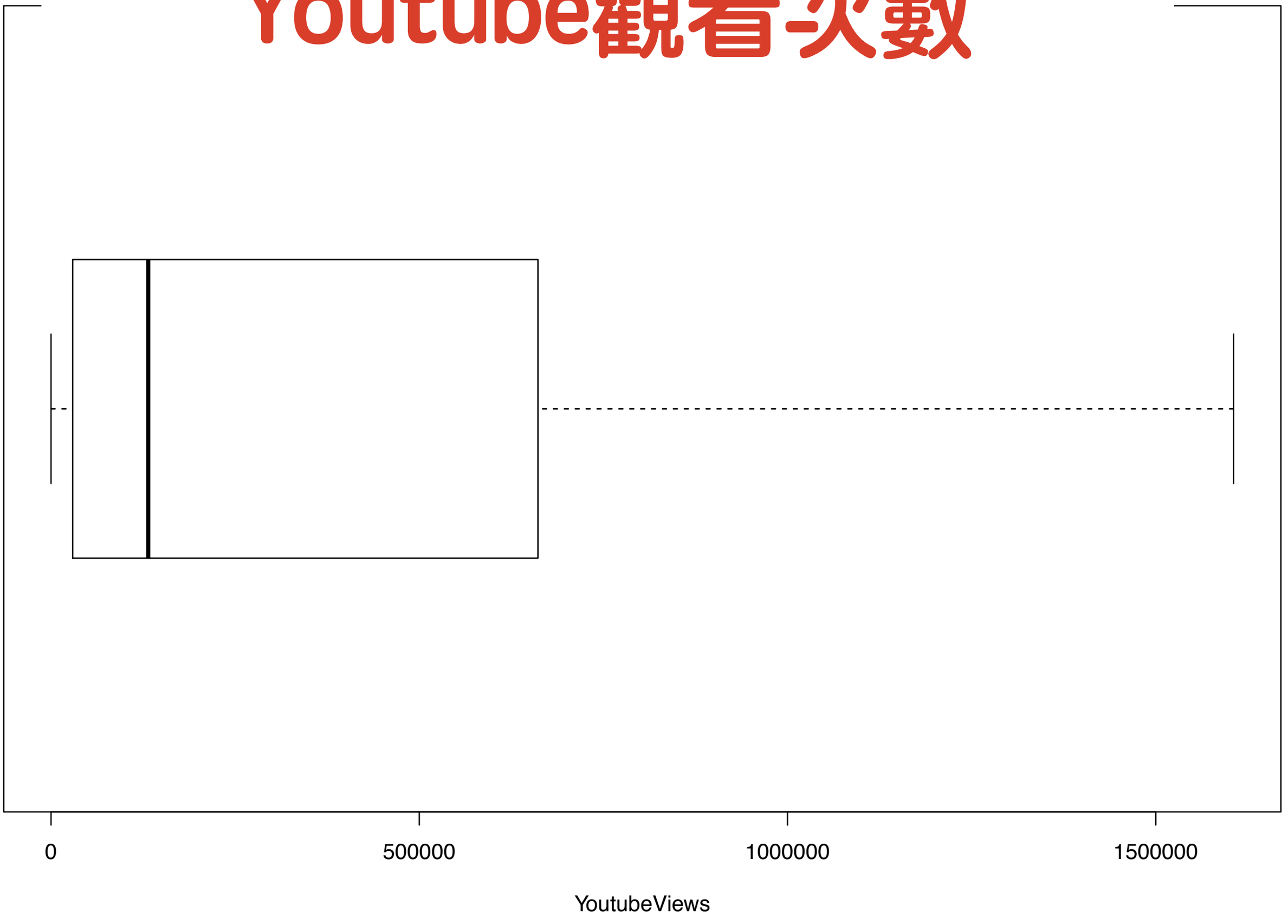
ＦＢ讚數



電影長度

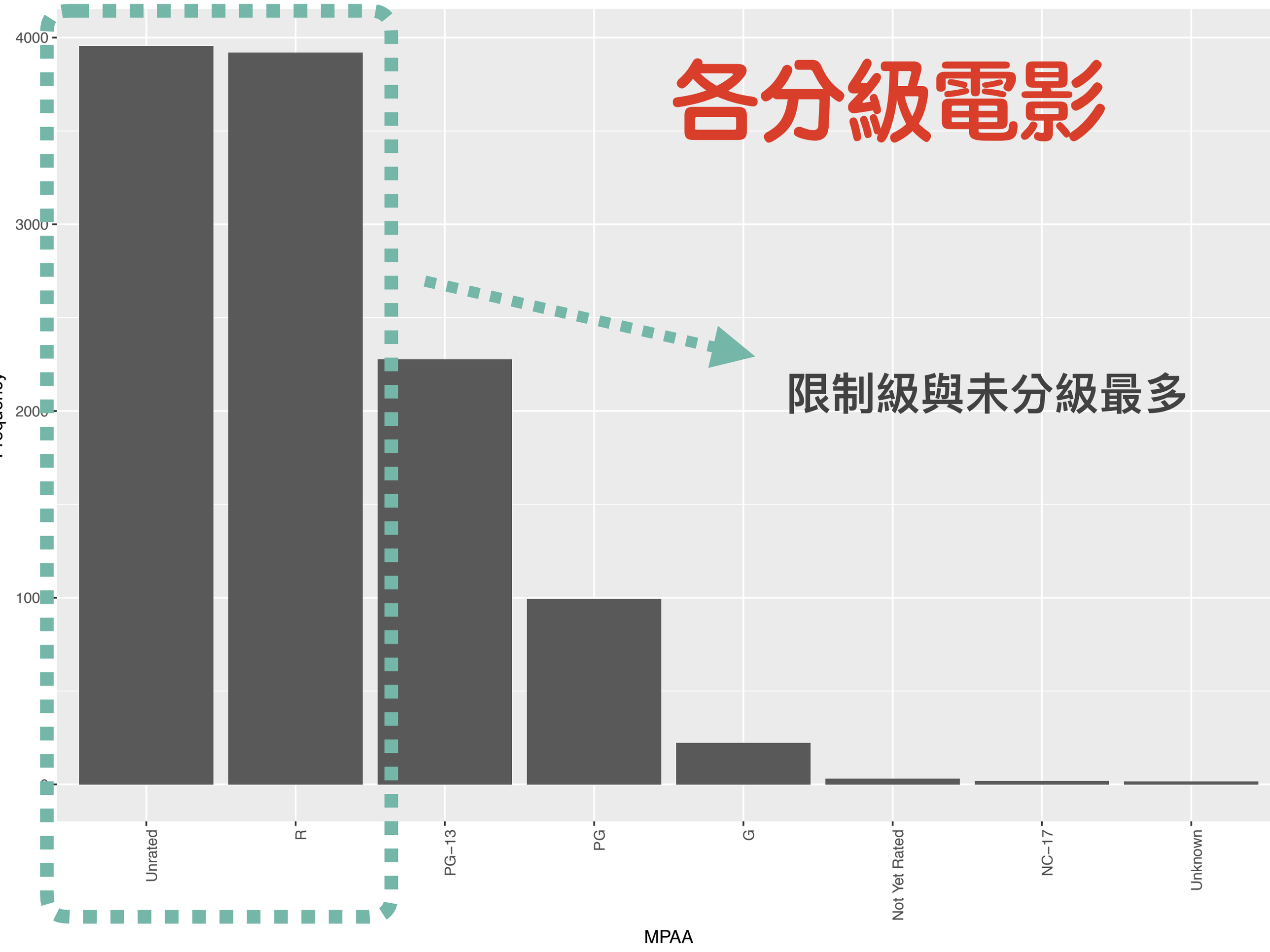


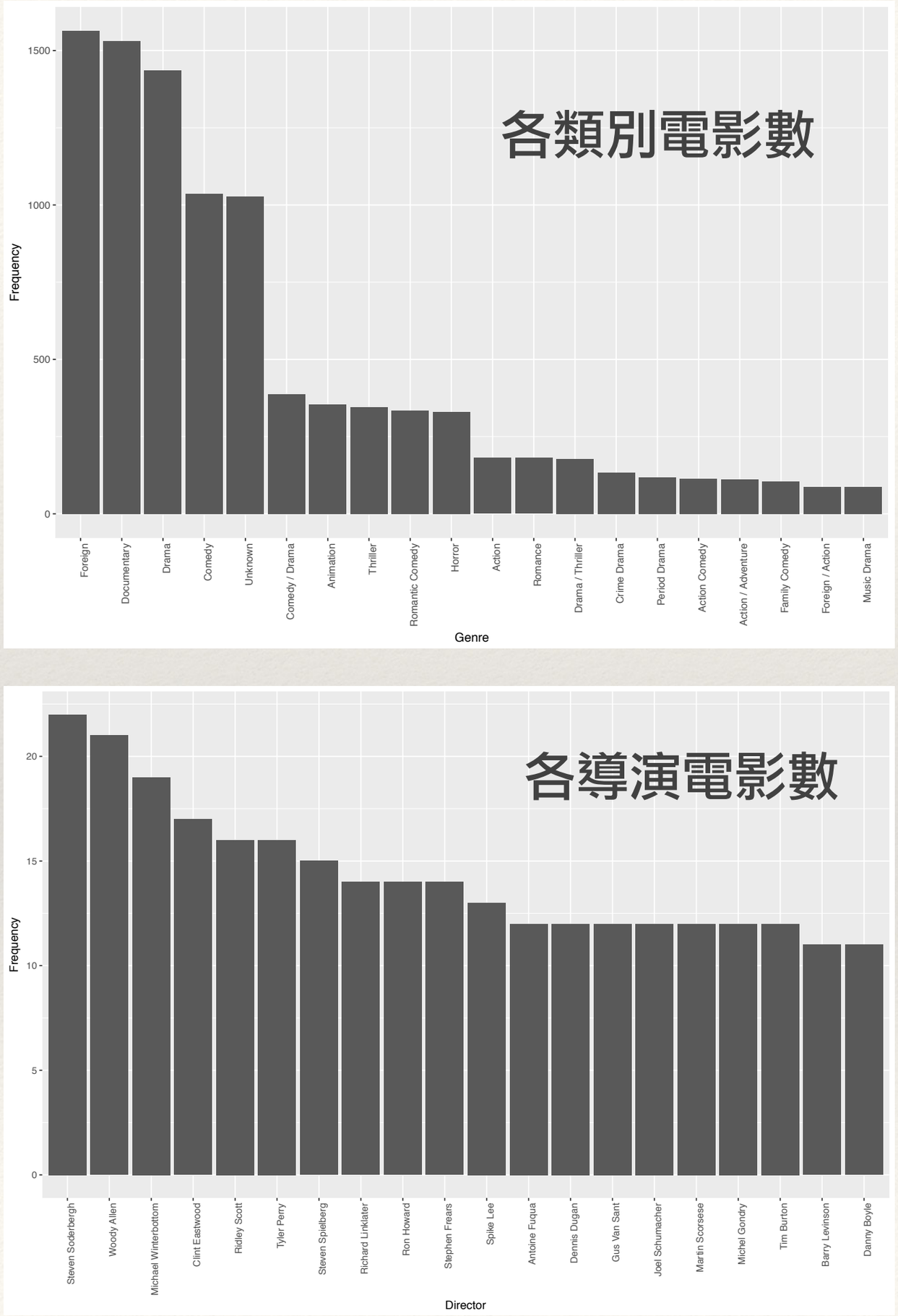
Youtube觀看次數

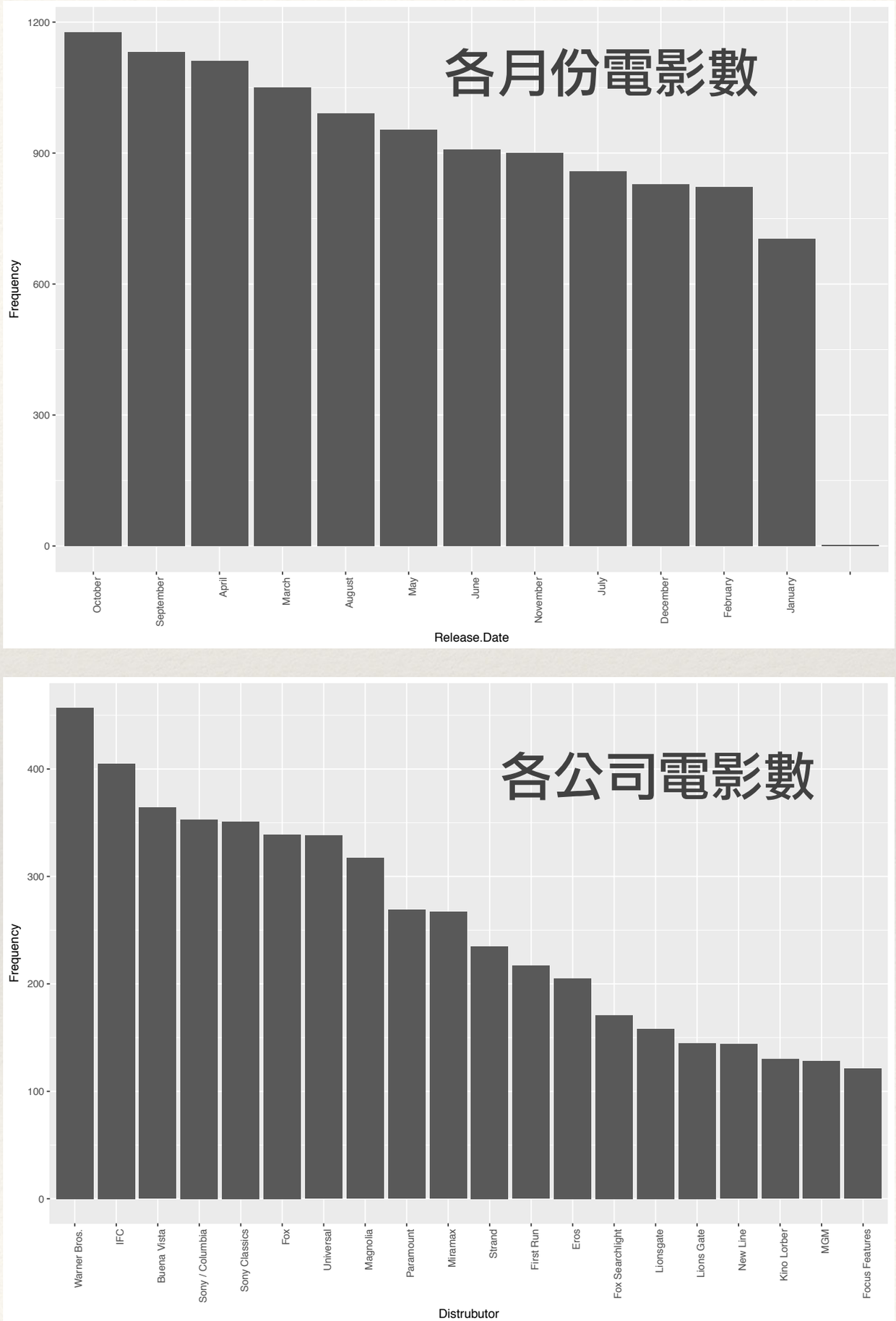


各分級電影

限制級與未分級最多




各類別電影數
Frequency
1500
1000
500
0
Foreign
Documentary
Drama
Comedy
Unknown
Comedy / Drama
Animation
Thriller
Romantic Comedy
Horror
Action
Romance
Drama / Thriller
Crime Drama
Period Drama
Action Comedy
Action / Adventure
Family Comedy
Foreign / Action
Music Drama
Genre


各月份電影數
Frequency
1200
900
600
300
0
October
September
April
March
August
May
June
November
July
December
February
January
Release.Date

各導演電影數
Frequency
20
15
10
5
0
Steven Soderbergh
Woody Allen
Michael Winterbottom
Clint Eastwood
Ridley Scott
Tyler Perry
Steven Spielberg
Richard Linklater
Ron Howard
Stephen Frears
Spike Lee
Antoine Fuqua
Dennis Dugan
Gus Van Sant
Joel Schumacher
Martin Scorsese
Michel Gondry
Tim Burton
Barry Levinson
Danny Boyle
Director

各公司電影數
Frequency
400
300
200
100
0
Warner Bros.
IFC
Buena Vista
Sony / Columbia
Sony Classics
Fox
Universal
Magnolia
Paramount
Miramax
Strand
First Run
Eros
Fox Searchlight
Lionsgate
Lions Gate
New Line
Kino Lorber
MGM
Focus Features
Distrubutor

統計量

```
> summary(data)
       Title              Box                       Distrubutor         Release.Date              Genre              Runtime
 Bluebeard:     2   Min.   :        30   Warner Bros.      : 457   October   :1176   Foreign          :1563   Min.   : 35.0
 Chaos    :     2   1st Qu.:     34637   IFC               : 405   September:1131   Documentary      :1530   1st Qu.: 90.0
 Evolution:     2   Median :    330050   Buena Vista       : 364   April     :1111   Drama            :1437   Median :100.0
 Heaven   :     2   Mean   : 17226091   Sony / Columbia: 353   March     :1050   Comedy           :1036   Mean   :103.9
 Logan    :     2   3rd Qu.: 10446436   Sony Classics     : 351   August    : 991   Unknown          :1027   3rd Qu.:112.0
 Room     :     2   Max.   :936662225   Fox               : 339   May       : 953   Comedy / Drama: 388   Max.   :729.0
 (Other)  :11422                        (Other)           :9165   (Other)   :5022   (Other)          :4453   NA's   :376
             MPAA            Budget                            Director        FB_likes             YoutubeViews
 Unrated       :3955   Min.   :       2   Steven Soderbergh     :   22   Min.   :        0   Min.   :1.000e+00
 R             :3921   1st Qu.:      20   Woody Allen           :   21   1st Qu.:     2406   1st Qu.:2.793e+04
 PG-13         :2276   Median :      37   Michael Winterbottom:   19   Median :    17884   Median :1.300e+05
 PG            : 994   Mean   :   15216   Clint Eastwood        :   17   Mean   :   713202   Mean   :1.787e+06
 G             : 222   3rd Qu.:      80   Ridley Scott          :   16   3rd Qu.:   153851   3rd Qu.:6.715e+05
 Not Yet Rated:   32   Max.   :1000000   (Other)               :3645   Max.   :188623068   Max.   :2.078e+09
 (Other)       :   34   NA's   :8715      NA's                  :7694   NA's   :1652
```

ANOVA

```
> summary_movie_aov
                                      Df    Sum Sq   Mean Sq F value   Pr(>F)
FB_likes                               1 6.833e+17 6.833e+17 154.263  < 2e-16 ***
YoutubeViews                           1 5.662e+16 5.662e+16  12.783 0.000356 ***
Runtime                                1 1.078e+18 1.078e+18 243.358  < 2e-16 ***
Budget                                 1 9.038e+16 9.038e+16  20.405 6.54e-06 ***
FB_likes:YoutubeViews                  1 6.172e+17 6.172e+17 139.358  < 2e-16 ***
FB_likes:Runtime                       1 2.883e+16 2.883e+16   6.510 0.010783 *
YoutubeViews:Runtime                   1 1.023e+18 1.023e+18 231.047  < 2e-16 ***
FB_likes:Budget                        1 2.346e+14 2.346e+14   0.053 0.817995
YoutubeViews:Budget                    1 1.498e+15 1.498e+15   0.338 0.560869
Runtime:Budget                         1 1.624e+15 1.624e+15   0.367 0.544893
FB_likes:YoutubeViews:Runtime          1 1.819e+16 1.819e+16   4.106 0.042827 *
FB_likes:YoutubeViews:Budget           1 5.109e+15 5.109e+15   1.153 0.282917
FB_likes:Runtime:Budget                1 4.667e+14 4.667e+14   0.105 0.745504
YoutubeViews:Runtime:Budget            1 1.592e+16 1.592e+16   3.594 0.058082 .
FB_likes:YoutubeViews:Runtime:Budget   1 8.890e+14 8.890e+14   0.201 0.654190
Residuals                           2655 1.176e+19 4.429e+15
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
8763 observations deleted due to missingness
```

- 檢驗結果：預算會影響票房。
- 但不能確定 Youtube 瀏覽數、FB 讚數、電影長度個別是否會影響票房，因為在上頁，三者的交互作用是顯著的，不能對他們個別如何影響票房的假設。
- 只能猜測，Youtube 瀏覽數、FB 讚數、電影長度的組合可能會影響電影票房。

使用 Decision Tree

Decision Tree

- 資料先分成訓練集和驗證集
- 訓練：驗證 = 3 : 1
- 用驗證集計算模型準確度
- 準確度 ：
  (正確預測 true + 正確預測 false) / 總數量

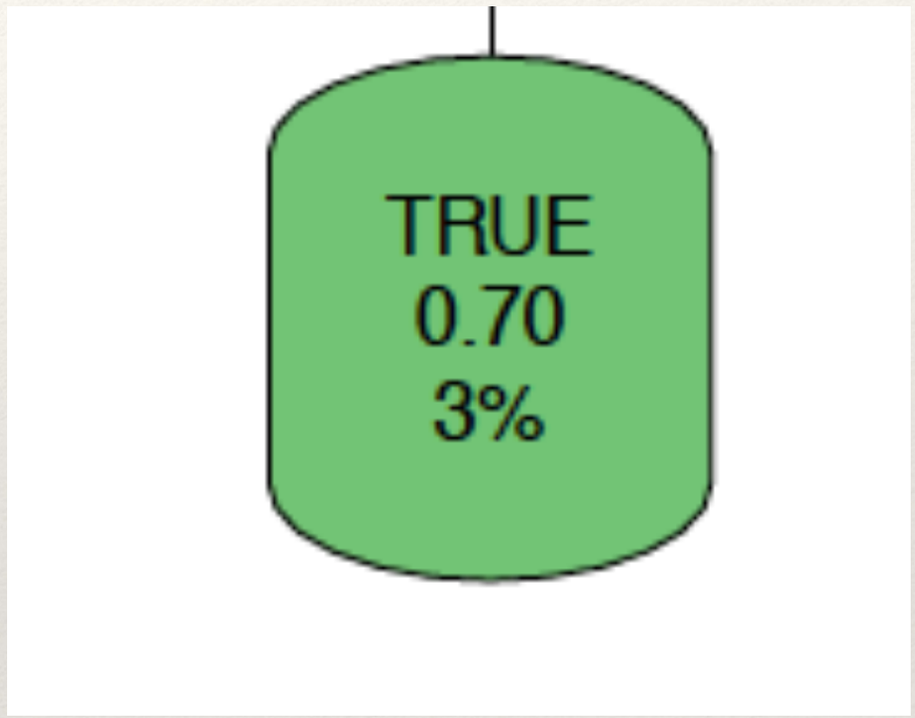
```
> valid.table
       Real
pred   FALSE TRUE
  FALSE  2690   69
  TRUE     45   54
```

96.01%

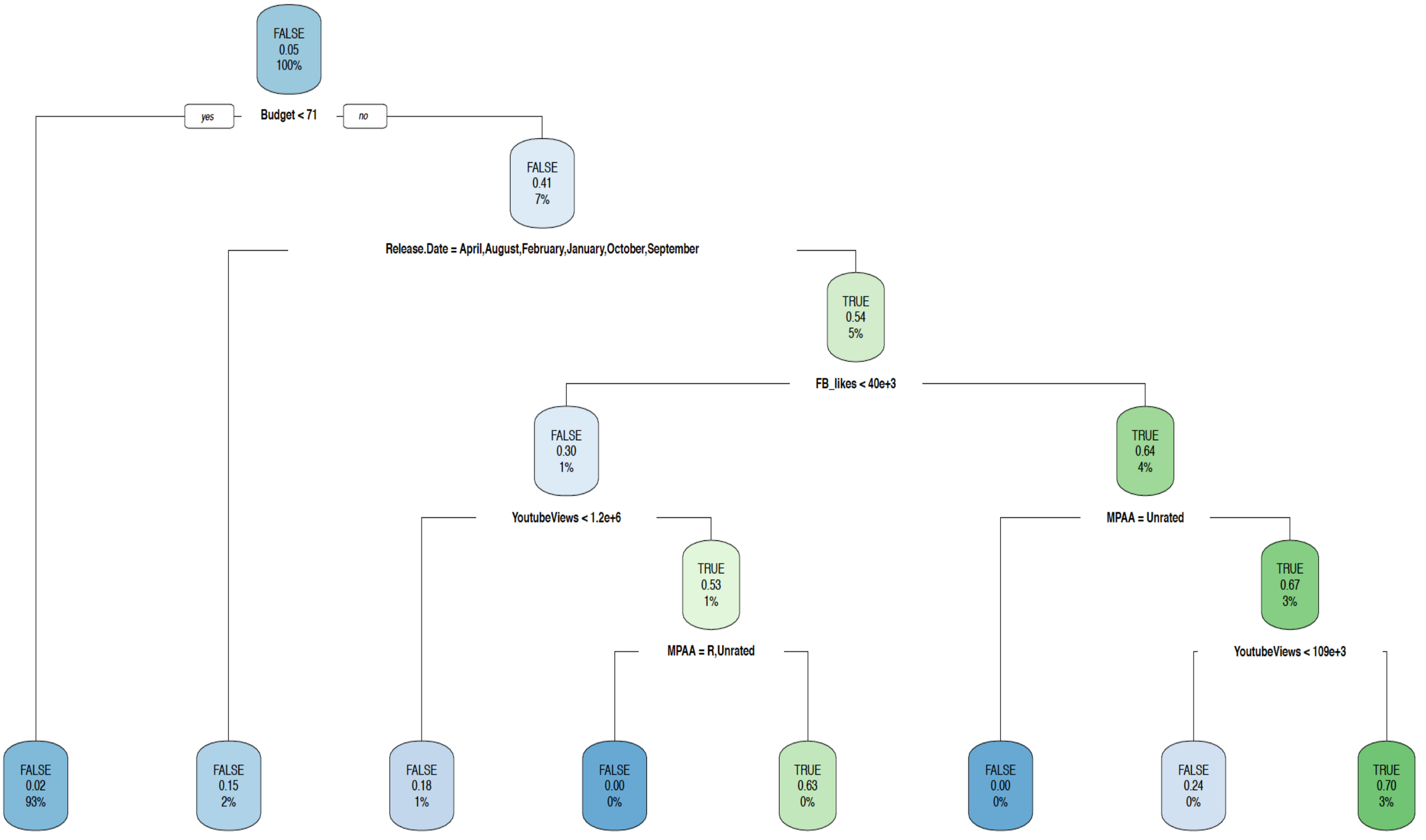
如果預測 True
喜憂參半

Decision Tree

- 閱讀方式
- 每個節點由上到下：
- 票房破億？ (True/False)
- 上述 True/False 正確的機率
- 在原始資料中，這樣的資料比例



Decision Tree

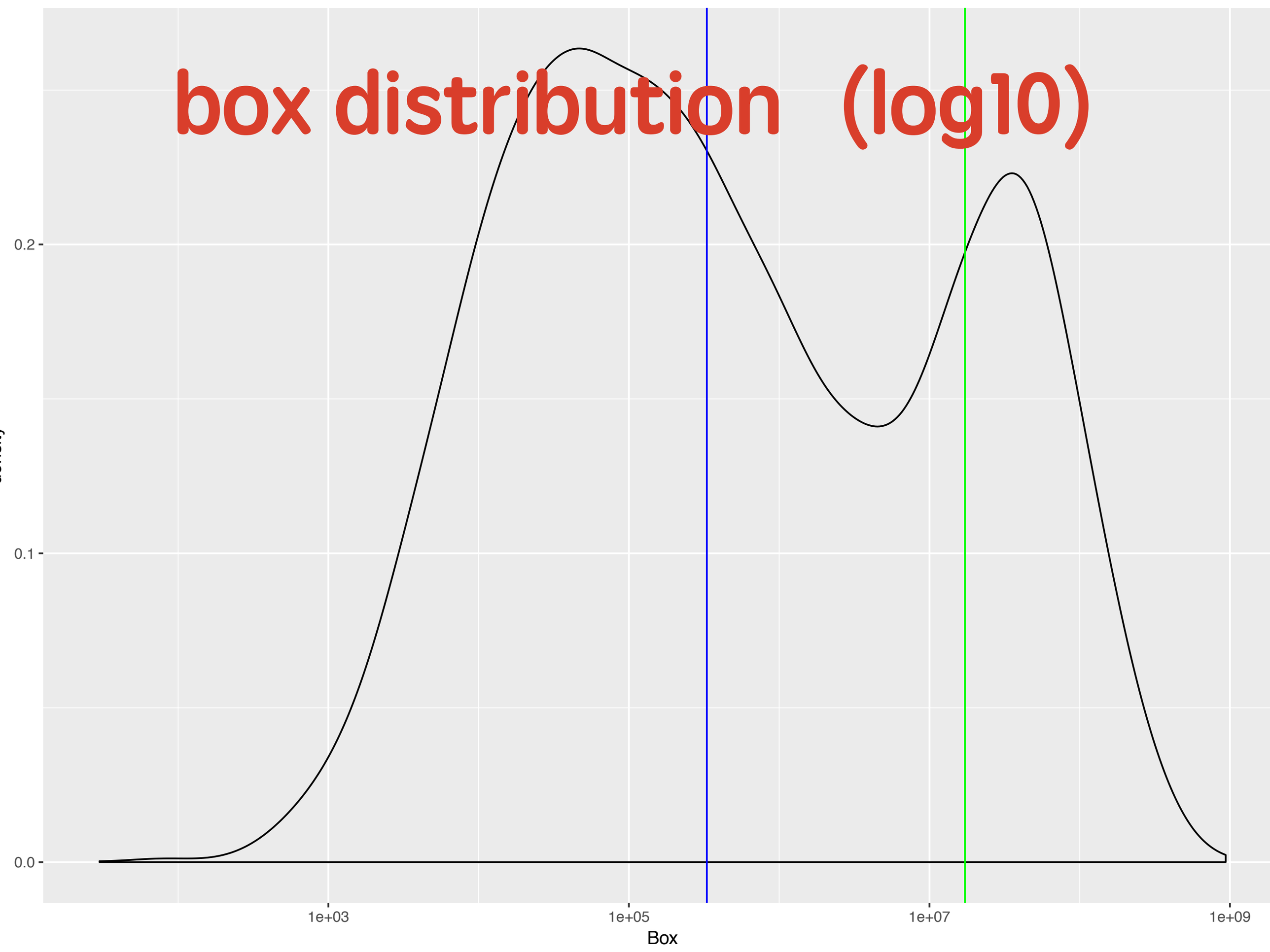


分工

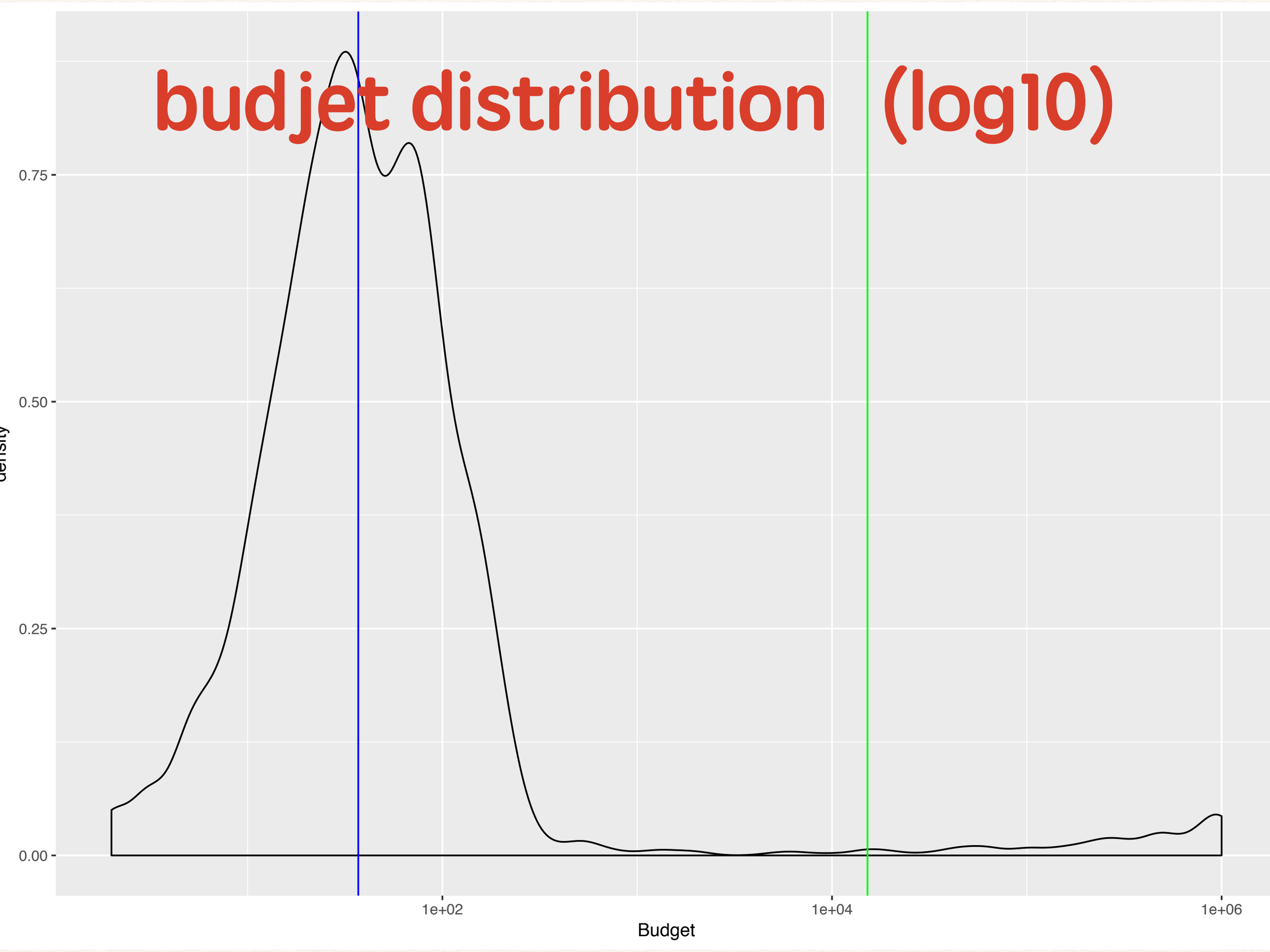
	FB api	爬蟲	R 程式	簡報
鄭光宇	>90%	65%	105%	<10%
呂振麒	<10%	35%	-5%	>90%

報告結束

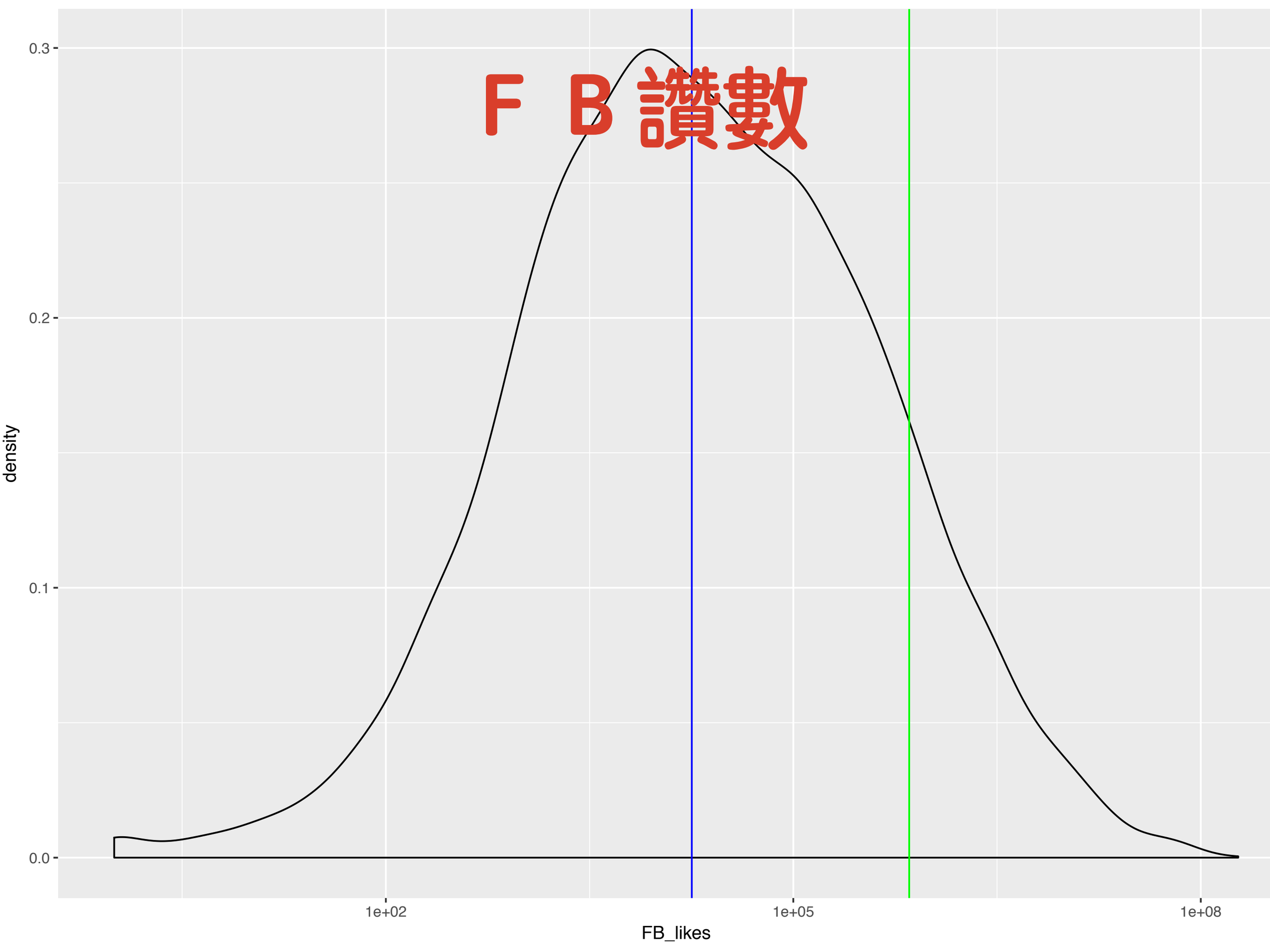
box distribution (log10)

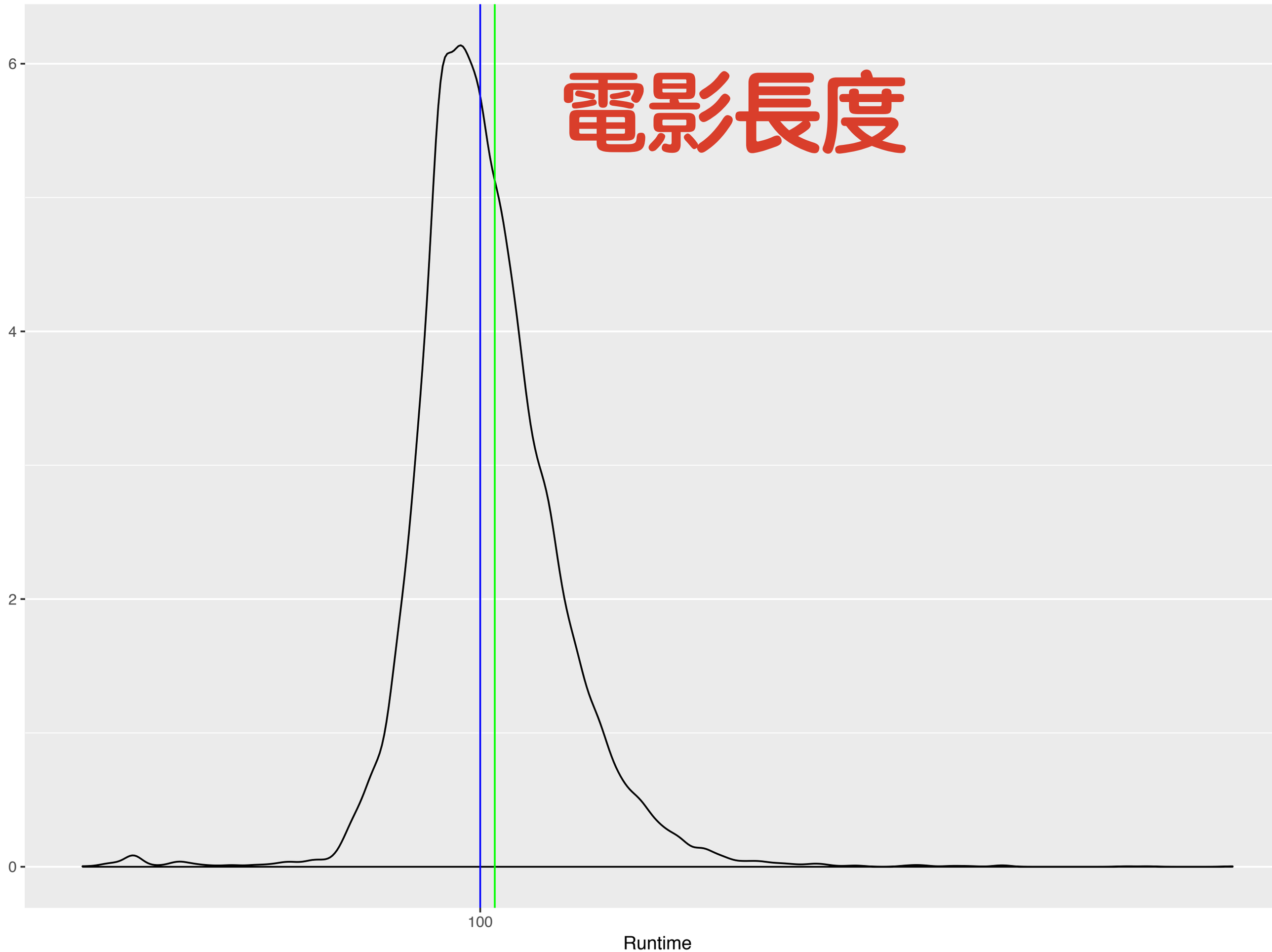


budjet distribution (log10)

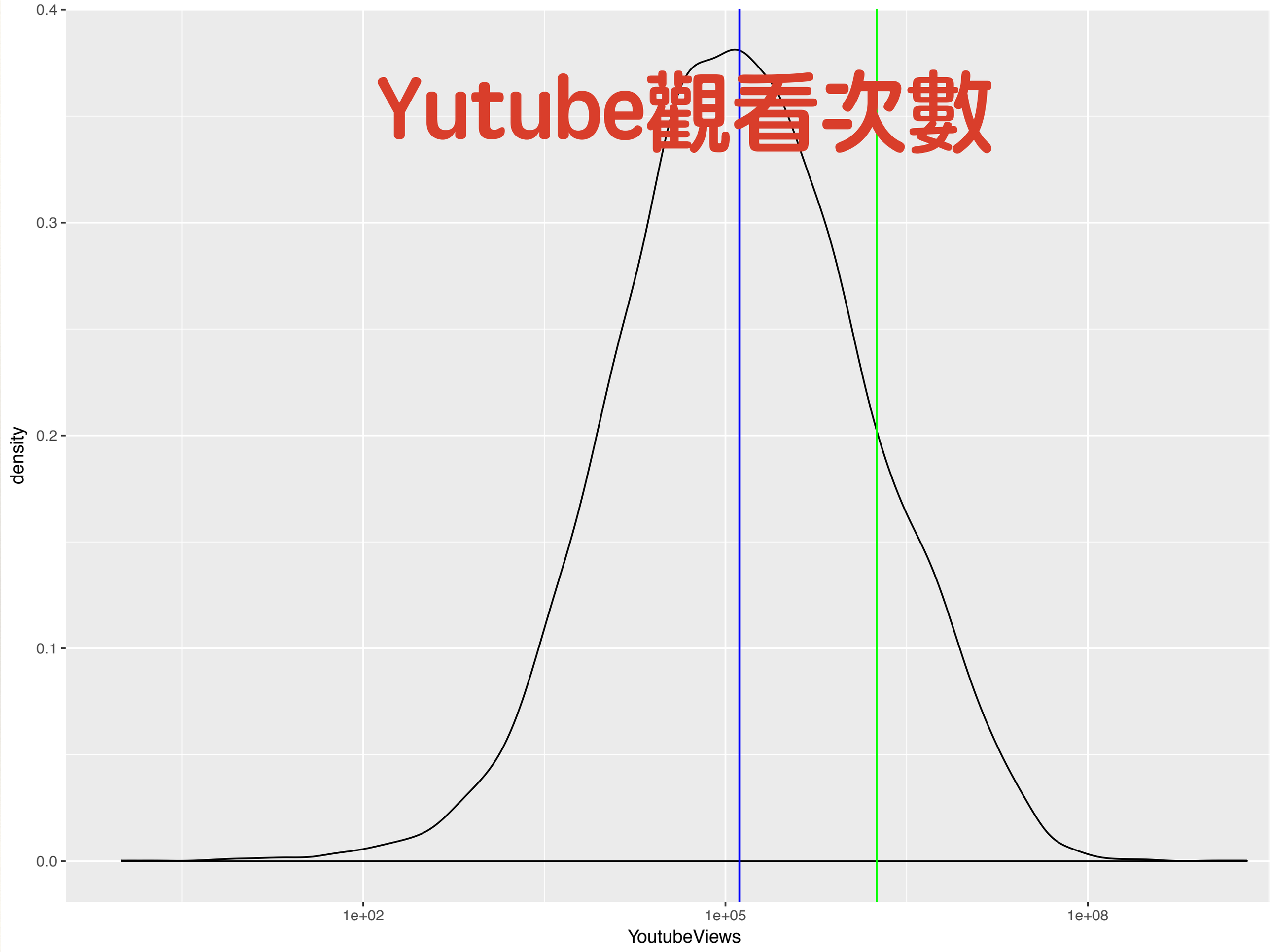


ＦＢ讚數




電影長度
6
4
2
0
100
Runtime

Yutube觀看次數



Branch: master ▾ **Statistics_Final** / Movie_Parser / **DownloadPkg_0.R** Find file | Copy path


peter0749 foo 93df361 7 days ago

1 contributor

3 lines (2 sloc) | 105 Bytes Raw | Blame | History

```
1  rm(list=ls(all=TRUE))
2  install.packages(c('XML', 'bitops', 'RCurl', 'NLP', 'Rfacebook', 'httr', 'chron'))
```

Branch: master ▾ **Statistics_Final** / Movie_Parser / **Functions.R** Find file | Copy path


peter0749 foo 227a6c1 7 days ago

1 contributor

2 lines (1 sloc) | 119 Bytes Raw | Blame | History

```
1  pureDigit <- function(X){ as.numeric(gsub('[^[:digit:]]','', as.character(X))) } #convert string to valid numeric type
```

peter0749 foo f2a61ab 7 days ago

1 contributor

67 lines (62 sloc) | 2.79 KB Raw Blame History

```
1  rm(list=ls(all=TRUE))
2  require(XML)
3  require(bitops)
4  require(RCurl)
5  require(NLP)
6  require(httr)
7  require(chron)
8  source('./Functions.R')
9
10 alldata = read.csv('testcsv.csv')
11 orgURL = 'http://www.boxofficemojo.com'
12 fulldata = data.frame()
13
```

```r
myHttpheader<- c(
  "User-Agent"="Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.103 Safari/537.36"
  "Accept"="text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8",
  "Connection"="keep-alive",
  "Accept-Charset"="big5,GB2312,utf-8;q=0.7,*;q=0.7",
  "Accept-Encoding"="gzip, deflate, sdch",
  "Accept-Language"="zh-TW,zh;q=0.8,en-US;q=0.6,en;q=0.4",
  "Upgrade-Insecure-Requests"="1",
  "Cache-Control"="max-age=0",
  "Cookie"="__utmt=1; __utma=137419939.1443367072.1468586224.1468720076.1468727749.10; __utmb=137419939.3.10.1468727749; __utmc
  "Host"="www.boxofficemojo.com",
  "Referer"="http://www.boxofficemojo.com/yearly/"
)

i=1
for( i in i:length(alldata$X))
{
  yahooURL <- paste(orgURL, alldata$Path[i], sep='')
  #yahooURL <- iconv(yahooURL, "big5", "utf8")
  #Encoding(yahooURL) = "UTF-8"
  print(yahooURL)
  URLExist = url.exists(yahooURL)
  print(URLExist)
  if(URLExist)
  {
    html = getURL(yahooURL, ssl.verifypeer = FALSE, encoding='UTF-8', httpheader = myHttpheader)
    xml = htmlParse(html, encoding='UTF-8')
    text = xpathSApply(xml, '//tr[@bgcolor="#ffffff"]/td[@valign="top"]/b', sessionEncoding='UTF-8', xmlValue)
```

```
    if (length(text)<6) next
    hasDirector = xpathSApply(xml, '//td[2]//div[@class="mp_box_content"]//tr[1]//td[1]//font[@size="2"]//text()', sessionEncod
    director = NA
    if (length(hasDirector) > 0 && substring(hasDirector,1, 8)=='Director') {
      director = xpathSApply(xml, '//td[2]//div[@class="mp_box_content"]//tr[1]//td[2]//font[@size="2"]//text()', sessionEncodi
    }
    testframe = data.frame(t(text), director)
    names(testframe) = c("Distrubutor","Release Date","Genre","Runtime","MPAA","Budget", "Director")
    testframe = cbind(alldata[i,-1],testframe)
    fulldata = rbind(fulldata, testframe)
  }
}

#Post-processing
fulldata$Runtime = gsub(" hrs. ",":",fulldata$Runtime)
fulldata$Runtime = gsub(" min.",":00",fulldata$Runtime)
fulldata$Runtime = chron::times(fulldata$Runtime)
fulldata$Runtime = chron::hours(fulldata$Runtime)*60 + minutes(fulldata$Runtime)
fulldata$'Release Date' = gsub(",","",fulldata$'Release Date')
fulldata$'Release Date' = gsub(" |[0-9]","",fulldata$'Release Date')
#fulldata$Budget = substring(fulldata$Budget,2)
#fulldata$Budget = gsub(" million","",fulldata$Budget)
fulldata$Budget = pureDigit(fulldata$Budget)
fulldata$Budget = as.numeric(fulldata$Budget)

write.csv(fulldata,"Fulllist.csv")
```

Branch: master ▾ **Statistics_Final** / Movie_Parser / **MovieList_1.R**

Find file | Copy path

xm35p4fu6 uhh — e309765 7 days ago

2 contributors

67 lines (58 sloc) | 2.37 KB

Raw | Blame | History

```
require(XML)
require(bitops)
require(RCurl)
require(httr)
rm(list=ls(all=TRUE))
source('./Functions.R')

yahoourl = "http://www.boxofficemojo.com/yearly/chart/?page="
appendurl_1 = "&view=releasedate&view2=domestic&yr="
appendurl_2 = "&p=.htm"
testurl = ""
testvector = c()
testframe = data.frame()

```

```
myHttpheader<- c(
  "User-Agent"="Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/51.0.2704.103 Safari/537.36"
  "Accept"="text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8",
  "Connection"="keep-alive",
  "Accept-Charset"="big5,GB2312,utf-8;q=0.7,*;q=0.7",
  "Accept-Encoding"="gzip, deflate, sdch",
  "Accept-Language"="zh-TW,zh;q=0.8,en-US;q=0.6,en;q=0.4",
  "Upgrade-Insecure-Requests"="1",
  "Cache-Control"="max-age=0",
  "Cookie"="__utmt=1; __utma=137419939.1443367072.1468586224.1468720076.1468727749.10; __utmb=137419939.13.10.1468727749; __utm
  "Host"="www.boxofficemojo.com"
)

for(j in 1996:2017)
{
  for(i in 1:10)
  {
    testurl = paste(yahoourl,i,appendurl_1,j,appendurl_2,sep='')
    testexist = url.exists(testurl)
    print(testurl)
    print(testexist)
    if(testexist)
    {
      html = getURL(testurl, ssl.verifypeer = FALSE, encoding='UTF-8', httpheader = myHttpheader)
      xml = htmlParse(html, encoding='UTF-8')
      path = xpathSApply(xml,'//table//td[2]/b/font[@size="2"]/a/@href',sessionEncoding='UTF-8')
      if(length(path)<1) break
      title = xpathSApply(xml,'//table//td[2]/b/font[@size="2"]/a[@href]',sessionEncoding='UTF-8',xmlValue)
```

```
    if(length(title)<1) break
    box = xpathSApply(xml,'//table//td[4]/font[@size="2"]/b',sessionEncoding='UTF-8',xmlValue)
    if(length(box)<1) break

    testlen = length(path)

    if(length(title)!=testlen || length(box)!=testlen) next
    tempframe = data.frame(title,path,box)
    testframe = rbind(testframe,tempframe)
  }
  else print(paste("assert: URL:",testurl,"not exist!"))
}

names(testframe) = c("Title","Path","Box")

#Post-processing and sorting
#testframe$Box = substring(testframe$Box,2)
#testframe$Box = gsub(",","",testframe$Box)
testframe$Box = pureDigit(testframe$Box)
testframe$Box = as.numeric(testframe$Box)
testframe = testframe[order(testframe$Box,decreasing=TRUE),]

write.csv(testframe,"testcsv.csv")
```

xm35p4fu6 沒bug就是有bug 9784668 7 days ago

2 contributors


```
rm(list=ls(all=TRUE))
require(XML)
require(bitops)
require(RCurl)
require(NLP)
require(httr)
require(chron)
require(Rfacebook)
source('./Functions.R')

tok = 'your token'
alldata = read.csv("Fulllist.csv")

youtubeSRC = 'https://www.youtube.com/results?q='
yAppendURL = '%20trailer&sp=CAA%253D'
youtubeURL = ''

fulldata = data.frame()

```

```
"User-Agent"="Chrome/51.0.2704.103",
"Upgrade-Insecure-Requests"="1"
)

for( i in 1:length(alldata$X))
{
testframe = data.frame( 'FB_likes' = Rfacebook::searchPages(alldata$Title[i], tok, n=1)$likes , 'YoutubeViews'=NA)
youtubeURL <- paste(youtubeSRC, alldata$Title[i], yAppendURL, sep='')
youtubeURL <- gsub(" ","%20",youtubeURL)
#youtubeURL <- iconv(youtubeURL, "big5", "utf8")
#Encoding(youtubeURL) = "UTF-8"
  print(youtubeURL)
URLExist = url.exists(youtubeURL)
  print(URLExist)
if(URLExist)
  {
html = getURL(youtubeURL, ssl.verifypeer = FALSE, encoding='UTF-8', httpheader = myHttpheader)
xml = htmlParse(html, encoding='UTF-8')
text = xpathSApply(xml,'//li/div/div/div[2]/div[2]/ul/li[2]/text()', sessionEncoding='utf8', xmlValue)
if(length(text)<1) next
#text = substring(text,6)
#text <- gsub(",","",text)
text <- pureDigit(text)
text <- as.numeric(text)
testframe$'YoutubeViews' = t(text)[1]
testframe = cbind(alldata[i,-1],testframe)
fulldata = rbind(fulldata, testframe)
  }
```