# IMDb Movies

| | genre | avg_vote | income |
|---|---|---|---|
| 1 | Crime, Drama, Thriller | 6.0 | 39.281 |
| 2 | Drama, Romance | 5.2 | 1.914 |
| 3 | Biography, Drama, Thriller | 8.0 | 233.556 |
| 4 | Action, Adventure | 5.4 | 7.455 |
| 5 | Comedy | 5.6 | 169.837 |
| 6 | Comedy, Romance | 4.9 | 100.375 |
| 7 | Action, Adventure, Drama | 7.6 | 710.645 |
| 8 | Action, Adventure, Drama | 5.8 | 161.502 |
| 9 | Action, Adventure, Sci-Fi | 5.6 | 1104.054 |
| 10 | Comedy, Crime | 6.3 | 107.645 |
| 11 | Comedy, Drama | 7.2 | 54.837 |
| 12 | Comedy, Drama | 6.2 | 4.511 |
| 13 | Action, Drama, Thriller | 6.2 | 53.260 |
| 14 | Comedy, Drama, History | 6.1 | 156.707 |
| 15 | Action, Biography, Drama | 7.3 | 547.426 |
| 16 | Adventure, Comedy, Drama | 5.9 | 212.902 |
| 17 | Adventure, Family, Sci-Fi | 5.8 | 45.681 |
| 18 | Biography, Drama | 7.5 | 1.862 |
| 19 | Comedy, Romance | 6.0 | 196.710 |
| 20 | Drama, Thriller | 5.2 | 17.534 |
| 21 | Action, Drama, Sci-Fi | 6.3 | 103.039 |
| 22 | Drama, Sport | 6.8 | 29.824 |

avg_vote (คะแนนโหวตเฉลี่ย) มีหน่วยเป็น คะแนน

income (รายได้ภาพยนตร์รวมทั่วโลก) มีหน่วยเป็น ล้านดอลลาร์สหรัฐ

```
> df <- read.csv("imdbm.csv")
> view(df)
```

# คำนวณหาค่าสถิติพื้นฐาน

คอลัมน์ avg_vote (คะแนนโหวตเฉลี่ย)

```
> getmode <- function(v) {
+   uniqv <- unique(v)
+   uniqv[which.max(tabulate(match(v, uniqv)))]
+ }
>
> mean(avg_vote)
[1] 6.4693
> median(avg_vote)
[1] 6.5
> getmode(avg_vote)
[1] 6.6
> sd(avg_vote)
[1] 0.9118108
>
> summary(avg_vote)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.400   6.000   6.500   6.469   7.100   8.600
```

Mean = 6.4693 คะแนน

Median = 6.5 คะแนน

Mode = 6.6 คะแนน

S.D. = 0.9118108

Min = 1.4 คะแนน

Max = 8.6 คะแนน

คอลัมน์ income(รายได้รวมทั่วโลก)

```
> getmode <- function(v) {
+   uniqv <- unique(v)
+   uniqv[which.max(tabulate(match(v, uniqv)))]
+ }
>
> mean(income)
[1] 156.462
> median(income)
[1] 50.423
> getmode(income)
[1] 39.281
> sd(income)
[1] 276.0558
>
> summary(income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.01   13.72   50.42  156.46  162.20 2797.80
```

Mean =  156.462 ล้านดอลลาร์สหรัฐ

Median = 50.423 ล้านดอลลาร์สหรัฐ

Mode = 39.281 ล้านดอลลาร์สหรัฐ

S.D. = 276.0558

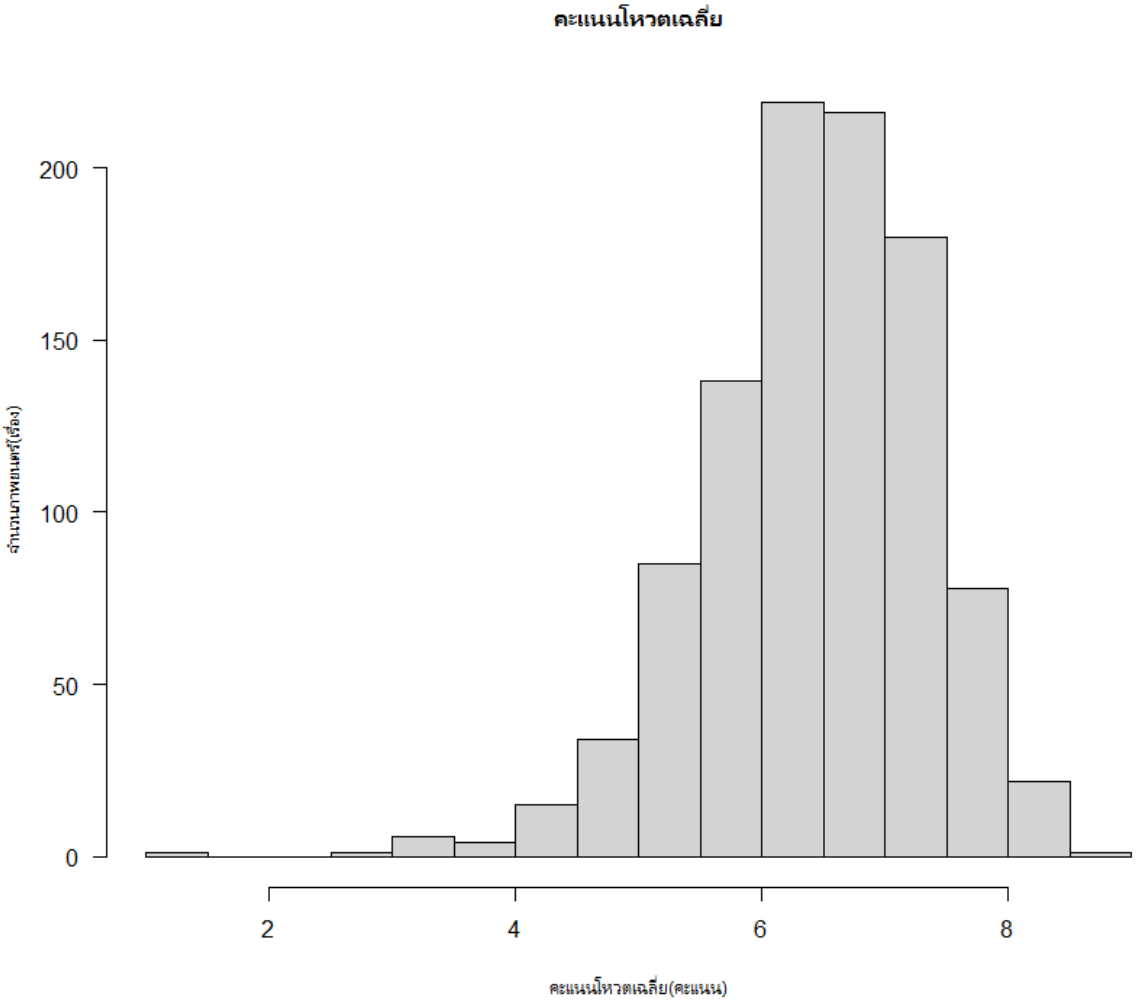
Min = 1.01 ล้านดอลลาร์สหรัฐ

Max = 2797.80 ล้านดอลลาร์สหรัฐ

# วาดกราฟ

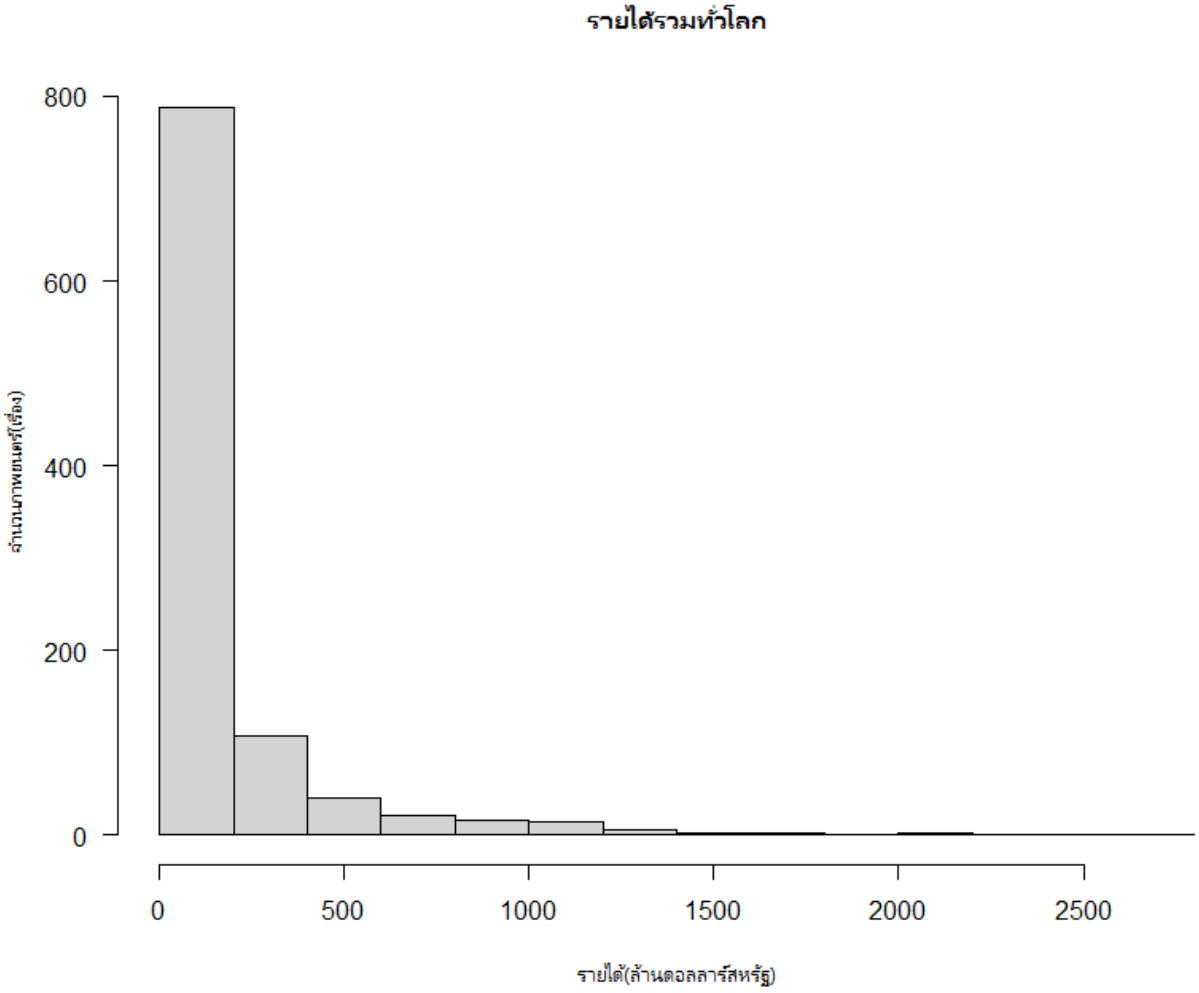
- Histogram

คอลัมน์ avg_vote (คะแนนโหวตเฉลี่ย)




```
hist(
  avg_vote,
  main = "คะแนนโหวตเฉลี่ย",
  xlab = "คะแนนโหวตเฉลี่ย(คะแนน)",
  ylab = "จำนวนภาพยนตร์(เรื่อง)",
  las = 1
  )
```

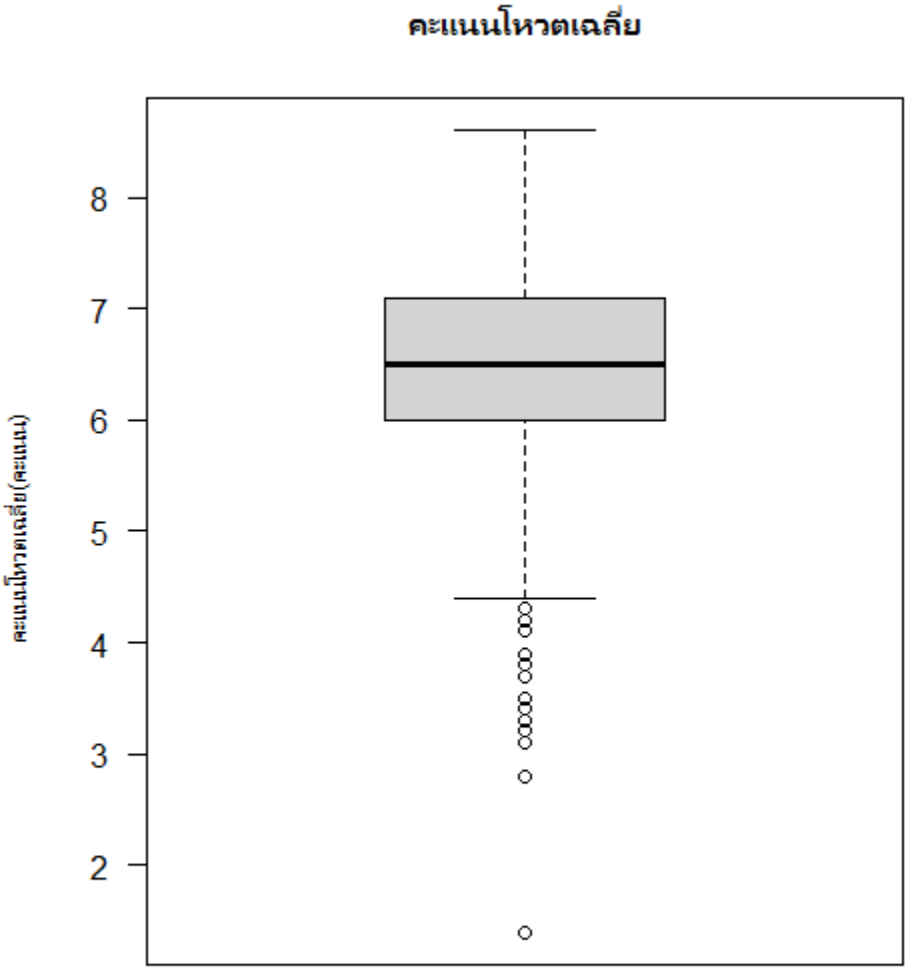
คอลัมน์ income(รายได้รวมทั่วโลก)




```
hist(
  income,
  main = "รายได้รวมทั่วโลก",
  xlab = "รายได้(ล้านดอลลาร์สหรัฐ)",
  ylab = "จำนวนภาพยนตร์(เรื่อง)",
  las = 1
  )
```

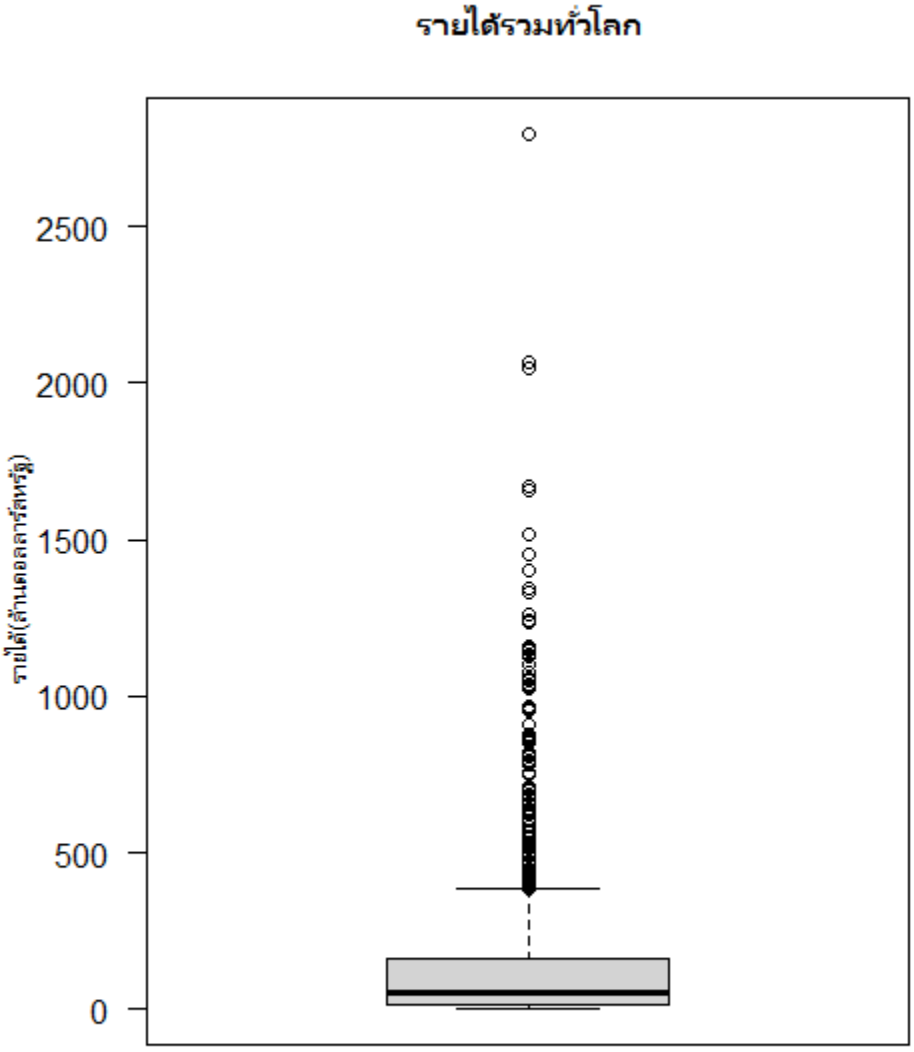
- Box Plot

คอลัมน์ avg_vote (คะแนนโหวตเฉลี่ย)




```
boxplot(avg_vote, main = "คะแนนโหวตเฉลี่ย",
        ylab =  "คะแนนโหวตเฉลี่ย(คะแนน)",
        las=1
        )
```

คอลัมน์ income(รายได้รวมทั่วโลก)




```
boxplot(income, main = "รายได้รวมทั่วโลก",
        ylab = "รายได้(ล้านดอลลาร์สหรัฐ)",
        las=1
)
```

-Stem and Leave

คอลัมน์ avg_vote (คะแนนโหวตเฉลี่ย)

```
> stem(avg_vote)

  The decimal point is at the |

  1 | 4
  1 |
  2 |
  2 | 8
  3 | 12234
  3 | 57889
  4 | 11222334444
  4 | 55556666667777788888888888999999
  5 | 00000001111111112222222222222222222223333333333333333334444444444444
  5 | 55555555555555556666666666666666666666666677777777777777777777777788+37
  6 | 00000000000000000000000000000000000011111111111111111111111111111111+135
  6 | 55555555555555555555555555555555555555556666666666666666666666666666+148
  7 | 00000000000000000000000000001111111111111111111111111111111111111111+105
  7 | 55555555555555555555555566666666666666666666666677777777777777888888+11
  8 | 00000000000111111111111223444444
  8 | 556
```

คะแนนโหวตเฉลี่ย(คะแนน)

คอลัมน์ income(รายได้รวมทั่วโลก)

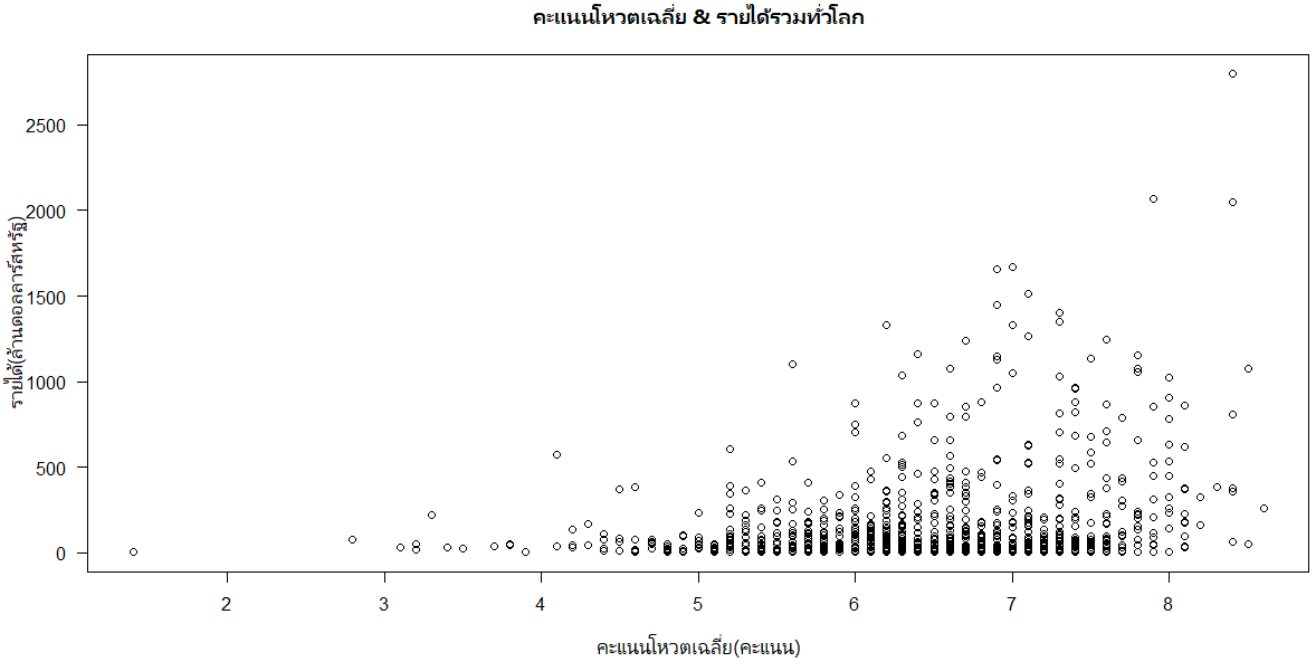
```
> stem(income)

  The decimal point is 2 digit(s) to the right of the |

   0 | 00000000000000000000000000000000000000000000000000000000000000000000+701
   2 | 00000000001111111111222222333333334444444455555566666666778999000000+32
   4 | 00111123334444556777799012223333333445577 8
   6 | 122334566888001568999
   8 | 01125666778881667
  10 | 23356777033556
  12 | 446335
  14 | 052
  16 | 67
  18 |
  20 | 57
  22 |
  24 |
  26 |
  28 | 0
```

รายได้(ล้านดอลลาร์สหรัฐ)

-XY (Scatter) Plot

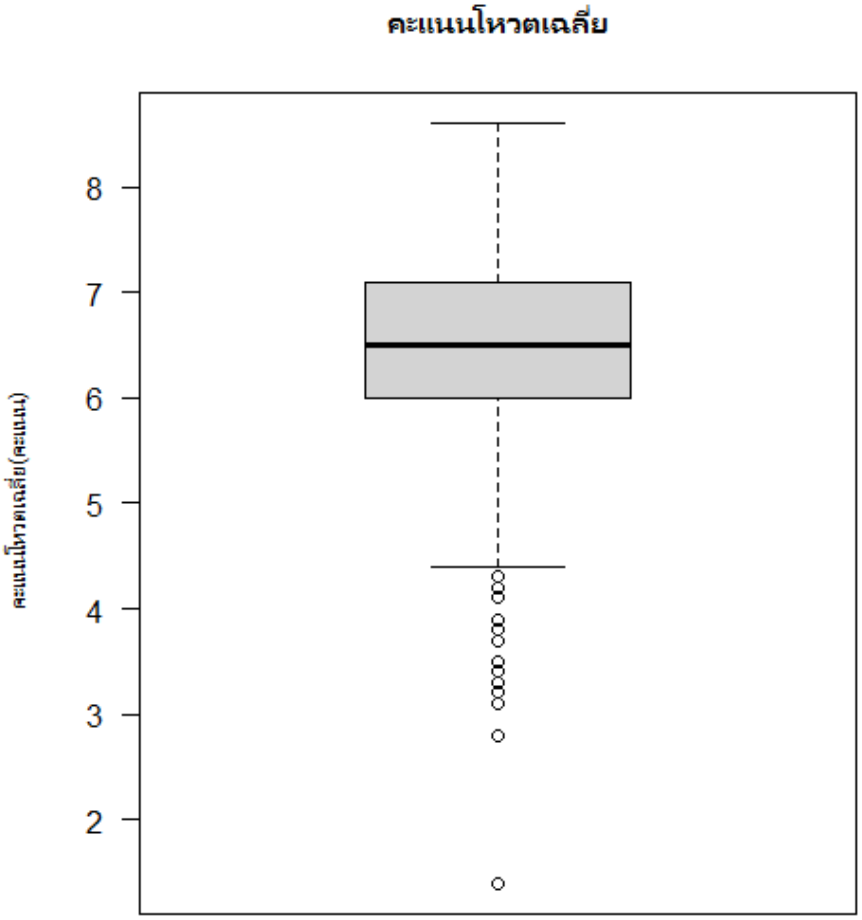



plot(avg_vote,income)

สาเหตุที่เลือก ตัวแปรต้น (x) เป็น คะแนนโหวตเฉลี่ย และ ตัวแปรตาม (y) เป็น รายได้รวมทั่วโลก เพราะผมอยากรู้ว่ารายได้ภาพยนตร์จะขึ้นอยู่กับคะแนนโหวตหรือไม่ เช่น ถ้าคะแนนเยอะรายได้ก็จะเยอะ คะแนนน้อยรายได้ก็จะน้อย

-Outlier

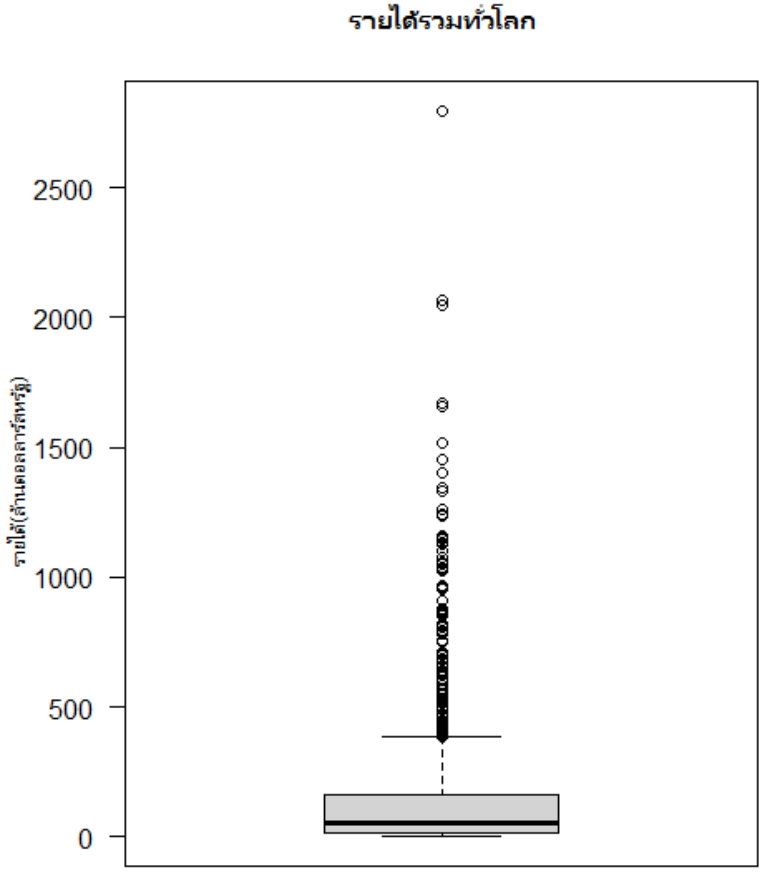
คอลัมน์ avg_vote (คะแนนโหวตเฉลี่ย)




ค่าที่มีแนวโน้มจะเป็น outlier ได้แก่

```
> boxplot(avg_vote)
> boxplot.stats(avg_vote)$out
 [1] 3.1 4.1 1.4 4.3 3.4 4.2 4.1 3.9 3.5 4.2 3.3 4.3 3.8 3.8 4.2 3.2 3.7 3.2
[19] 2.8
```

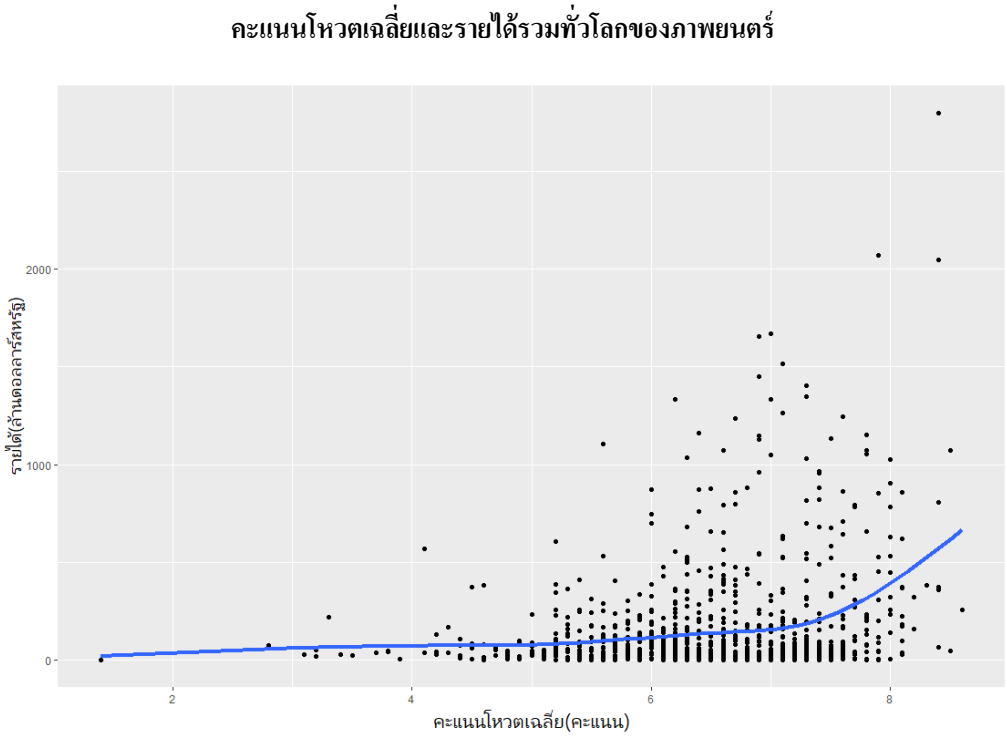
คอลัมน์ income(รายได้รวมทั่วโลก)




ค่าที่มีแนวโน้มจะเป็น outlier ได้แก่

```
> boxplot(income)
> boxplot.stats(income)$out
  [1] 710.645  1104.054 547.426  657.868  956.020  498.781  414.352  458.864
  [9] 1670.401 519.312  440.604  542.351  532.951  658.344  858.071  473.991
 [17] 386.042  1159.443 569.651  880.675  682.717  1402.809 2068.224 474.800
 [25] 1515.048 630.162  439.049  677.718  746.847  782.612  389.682  521.171
 [33] 1028.571 875.458  1023.789 873.638  966.552  814.039  543.934  408.579
 [41] 634.155  1153.332 643.347  1056.058 446.486  415.485  553.810  821.847
 [49] 681.872  701.796  435.086  794.879  880.167  962.102  409.232  807.084
 [57] 1332.540 1264.064 619.021  605.425  490.720  1034.799 853.979  566.653
 [65] 527.966  863.756  1236.005 410.903  526.949  870.325  856.085  1148.486
 [73] 436.189  582.894  906.885  1347.281 428.028  511.596  1242.805 392.925
 [81] 654.856  2048.360 467.990  530.259  1331.958 791.120  622.674  528.584
 [89] 785.794  529.324  399.907  451.183  404.853  1073.395 521.800  1074.144
 [97] 386.600  1128.276 2797.801 1450.027 491.730  430.051  433.005  1656.964
[105] 1050.694 1131.928 759.057  1074.251 473.093  699.857  796.576
```

# บทวิเคราะห์ข้อมูลจากกราฟ




จากข้อมูลจากกราฟความสัมพันธ์ระหว่างคะแนนโหวตเฉลี่ยและรายได้ วิเคราะห์ได้ว่า เมื่อคะแนนโหวตสูง รายได้ของภาพยนตร์ก็จะสูงด้วย ซึ่งผมคิดว่าอาจจะเกิดจากคนดูภาพยนตร์จะดูคะแนนโหวตของภาพยนตร์ก่อนไปดู ถ้าคะแนนเยอะก็จะไปดู ทำให้รายได้ของภาพยนตร์เรื่องนั้นๆสูง ถ้าคะแนนน้อยก็จะไม่ดู ทำให้รายได้ของภาพยนตร์เรื่องนั้นๆต่ำ

## Source Code

```r
setwd("~/CE2D-2/git/Propstat")
library(formattable)
library(ggplot2)

df <- read.csv("imdbm.csv")
View(df)

income <- df$income
avg_vote <- df$avgVote

getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

mean(avgVote)
median(avgVote)
getmode(avgVote)
sd(avgVote)
summary(avgVote)

mean(income)
median(income)
getmode(income)
sd(income)
summary(income)

hist(
  income,
  main = "รายได้รวมทั่วโลก",
  xlab = "รายได้(ล้านดอลลาร์สหรัฐ)",
  ylab = "จำนวนภาพยนตร์(เรื่อง)",
  las = 1
  )
```

```r

hist(
  avg_vote,
  main = "คะแนนโหวตเฉลี่ย",
  xlab = "คะแนนโหวตเฉลี่ย(คะแนน)",
  ylab = "จำนวนภาพยนตร์(เรื่อง)",
  las = 1
  )

boxplot(avg_vote, main = "คะแนนโหวตเฉลี่ย",
        ylab =  "คะแนนโหวตเฉลี่ย(คะแนน)",
        las=1
        )

boxplot(income, main = "รายได้รวมทั่วโลก",
        ylab = "รายได้(ล้านดอลลาร์สหรัฐ)",
        las=1
)

stem(avg_vote)
stem(income)

plot(avg_vote,income,xlab="คะแนนโหวตเฉลี่ย(คะแนน)",
     ylab = "รายได้(ล้านดอลลาร์สหรัฐ)",
     las = 1,
     main = "คะแนนโหวตเฉลี่ย & รายได้รวมทั่วโลก",
     cex.lab=1.5,  cex.main=1.5
     )

boxplot(avg_vote)
boxplot.stats(avg_vote,coef=5)$out
boxplot(income)
boxplot.stats(income,coef=5)$out

ggplot(df,aes(x=avg_vote,y=income))+geom_point()+
       geom_smooth(method="gam",se=F, size = 1.5, alpha = 1)+
       xlab("คะแนนโหวตเฉลี่ย(คะแนน)") + ylab("รายได้(ล้านดอลลาร์สหรัฐ)")+
         theme(axis.title = element_text(size = 20))
```