62010619 นายพัทธพล จันทร์ชู

HW1# - เสนอชุดข้อมูล

ชื่อชุดข้อมูล ชื่อคอลัมน์ข้อมูลที่สนใจ 3 คอลัมน์

- ชื่อชุดข้อมูล : IMDb Movies

- ชื่อคอลัมน์ : 1.ประเภทของภาพยนตร์

2.คะแนนโหวตเฉลี่ย

3.รายได้รวมทั่วโลก

Why is it interesting ?

- เนื่องจากผมเป็นคนชอบดูภาพยนตร์ จึงอยากรู้ว่าภาพยนตร์ประเภทไหน คนชอบดูเยอะที่สุด และมีรายได้สูงที่สุด

แหล่งที่มาของข้อมูล

- www.kaggle.com/stefanoleone992/imdb-extensive-dataset

คำอธิบายชื่อคอลัมน์ที่เลือก และวิธีการรวบรวมข้อมูล

- คำอธิบายชื่อคอลัมน์ : 1.ประเภทของภาพยนตร์ เช่น Action, Adventure, Comedy, Fantasy, Sci-Fi

2.คะแนนโหวตเฉลี่ย โดยคะแนนโหวตจะอยู่ในช่วง 1 – 10 คะแนน

3.รายรวมได้ทั่วโลก คือ นำรายได้ของภาพยนตร์ในเรื่องนั้นจากทั่วโลกมารวมกัน

- วิธีการรวบรวมข้อมูล : IMDb เป็นเว็บไซต์ภาพยนตร์ที่ได้รับความนิยมสูงสุด โดยจะรวมภาพยนตร์จากทั่วโลก โดยจะให้คนมาโหวตโดยการกดดาวให้ภาพยนตร์เรื่องนั้นตั้งแต่ 1 – 10 ดวง ตามความชอบ

62010619 นายพัทธพล จันทร์ชู

HW2# - Plots and Basic Statistics

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. = 0.9118 คะแนน

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 ล้านดอลลาร์สหรัฐ

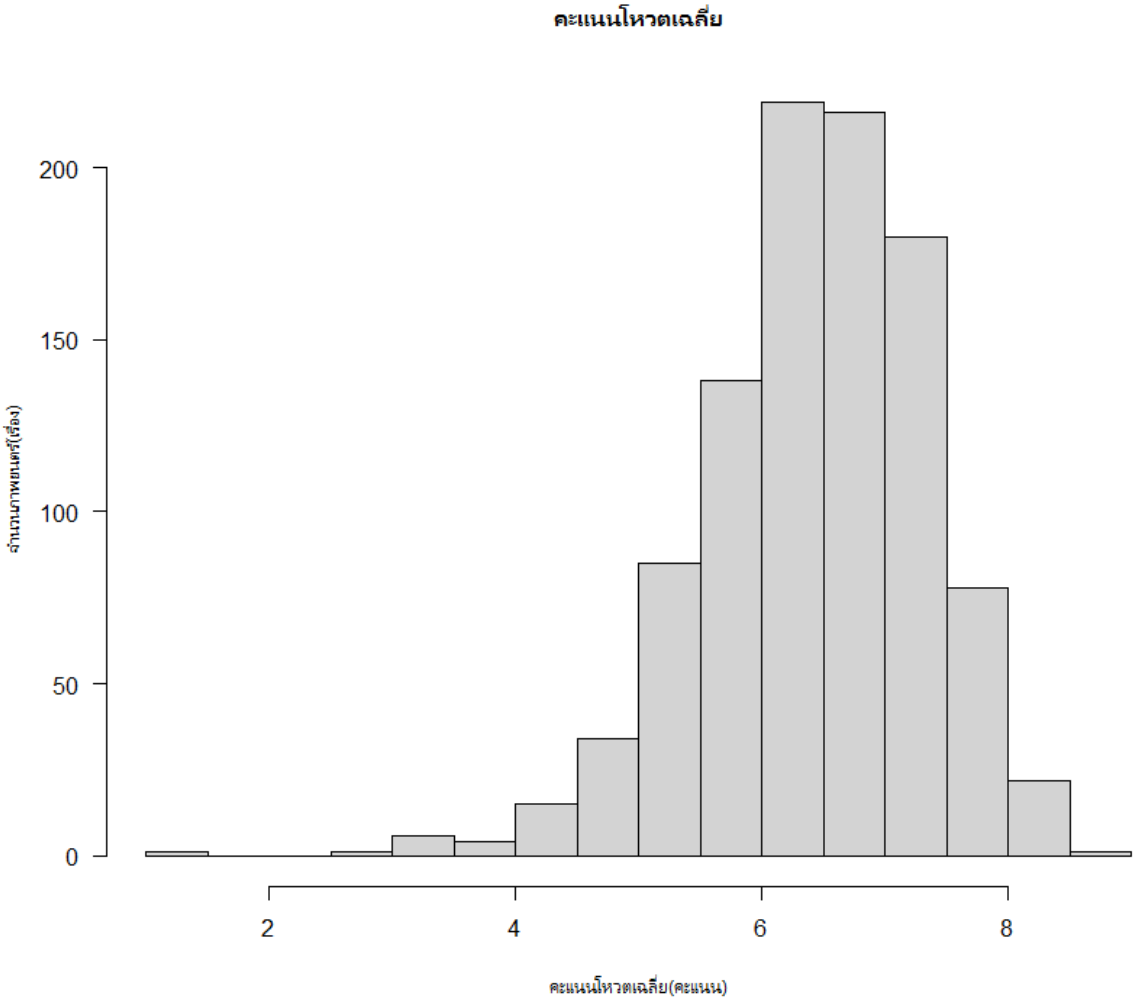
Min = 1.01 ล้านดอลลาร์สหรัฐ

Max = 2797.80 ล้านดอลลาร์สหรัฐ

วาดกราฟ

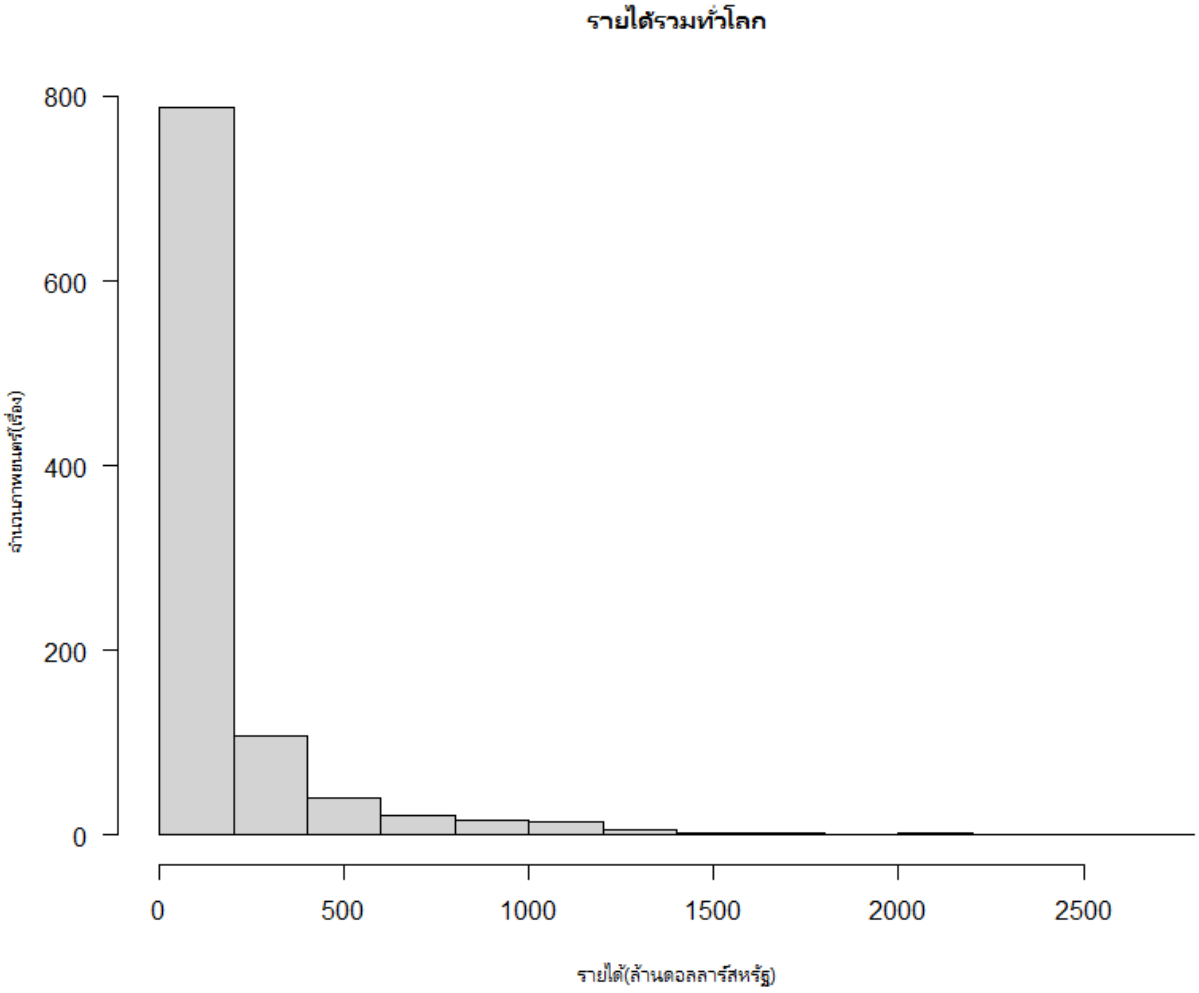
- Histogram

คอลัมน์ avg_vote (คะแนนโหวตเฉลี่ย)




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

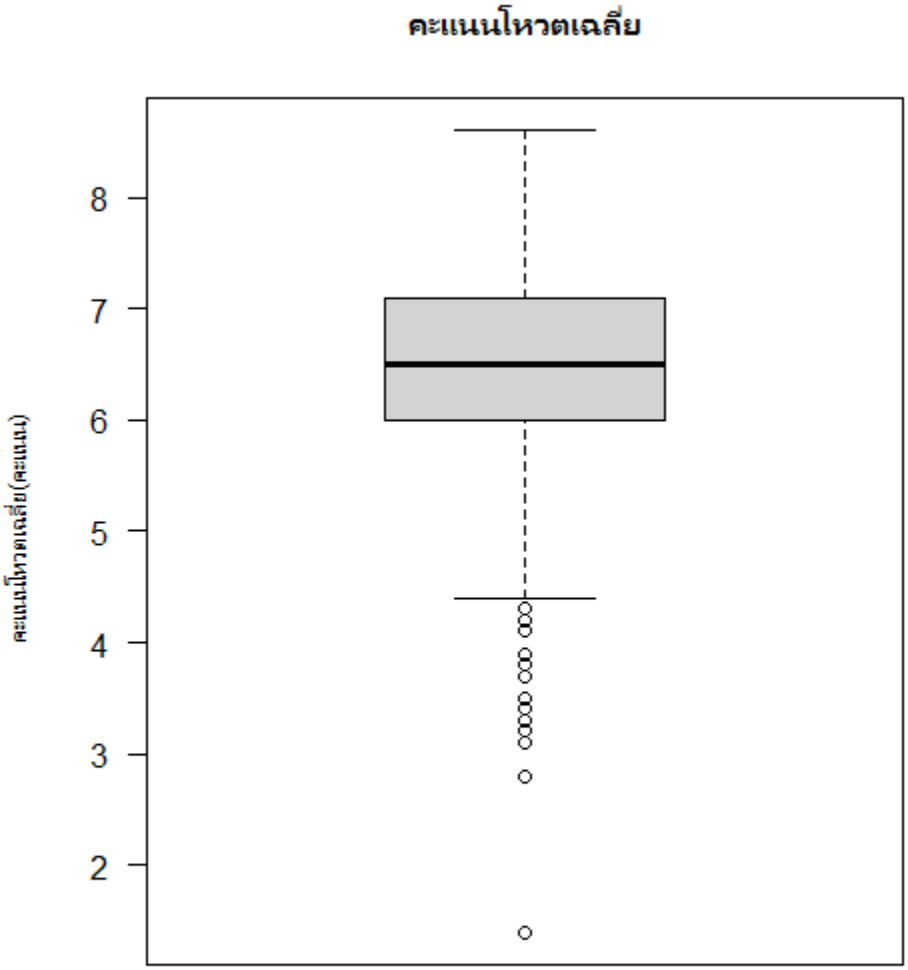
คอลัมน์ income(รายได้รวมทั่วโลก)




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

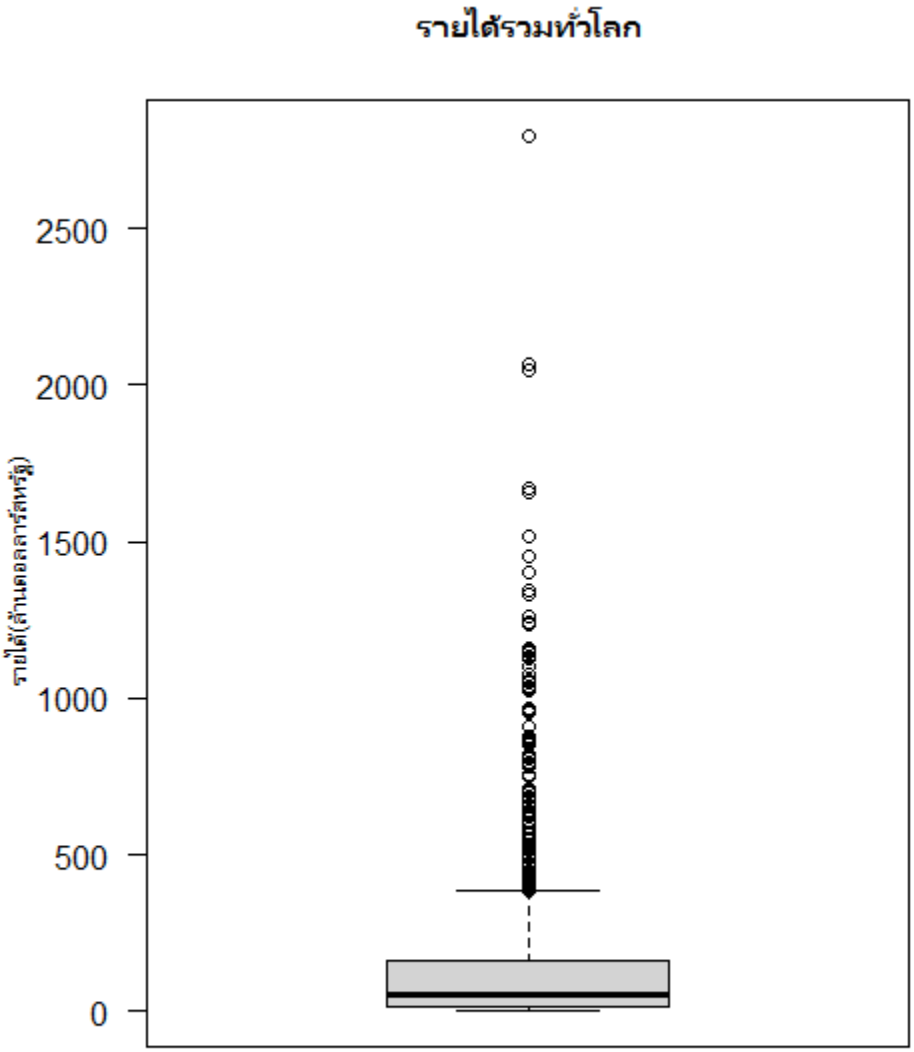
- Box Plot

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 | 00000001111111112222222222222222222233333333333333333334444444444444
  5 | 55555555555555556666666666666666666666666677777777777777777777777788+37
  6 | 00000000000000000000000000000000000011111111111111111111111111111+135
  6 | 55555555555555555555555555555555555555556666666666666666666666666666+148
  7 | 00000000000000000000000000001111111111111111111111111111111111111111+105
  7 | 55555555555555555555555556666666666666666666666666777777777777777888888+11
  8 | 00000000000111111111111223444444
  8 | 556
```

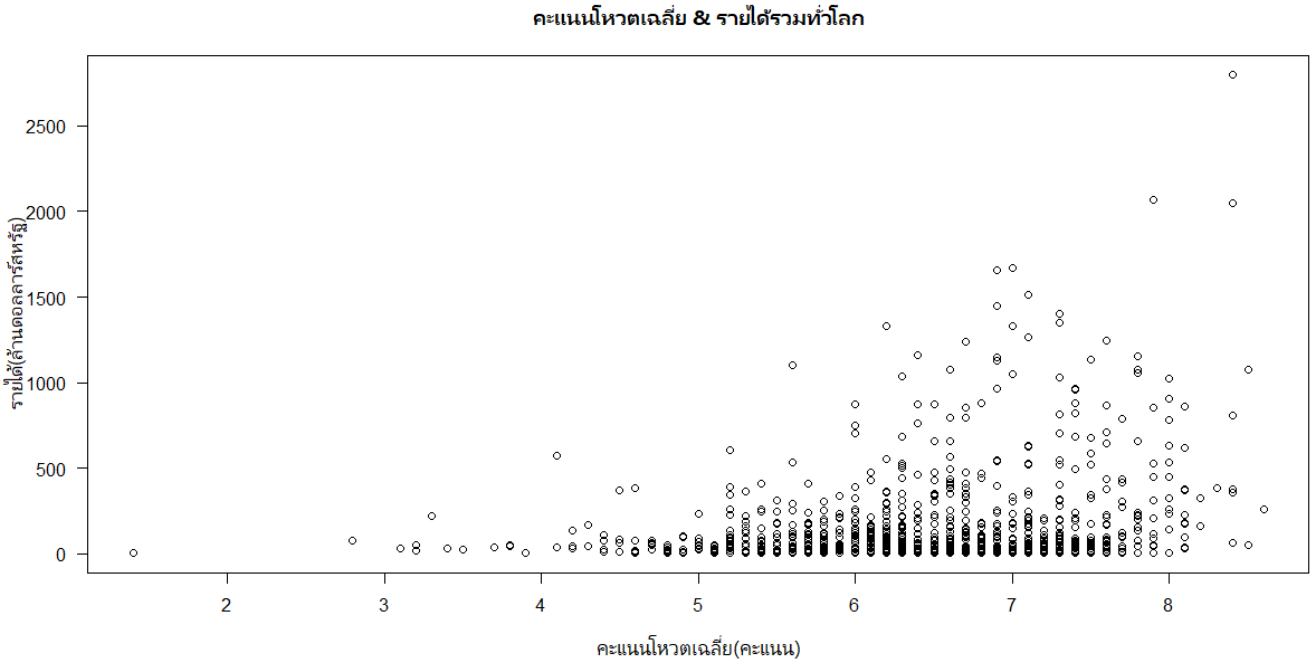
คะแนนโหวตเฉลี่ย(คะแนน)

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

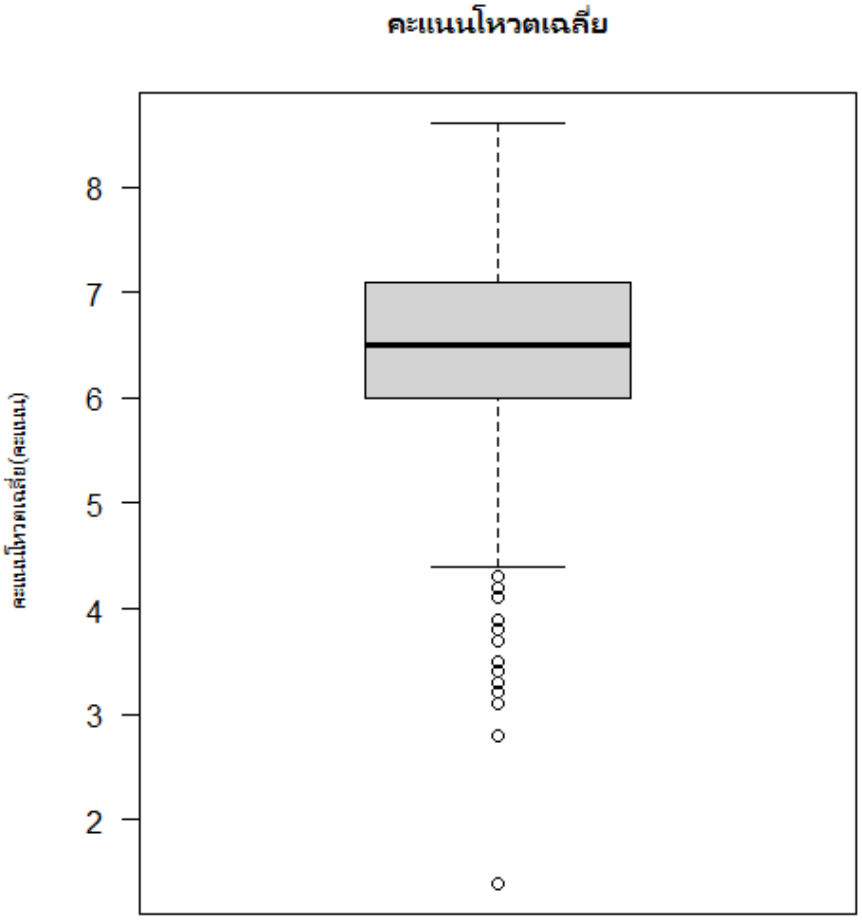


plot(avg_vote,income)

สาเหตุที่เลือก ตัวแปรต้นเป็น คะแนนโหวตเฉลี่ย และ ตัวแปรตามเป็น รายได้รวมทั่วโลก เพราะผมอยากรู้ว่าคะแนนโหวตจะส่งผลอย่างไรกับรายได้ของภาพยนตร์

-Outlier

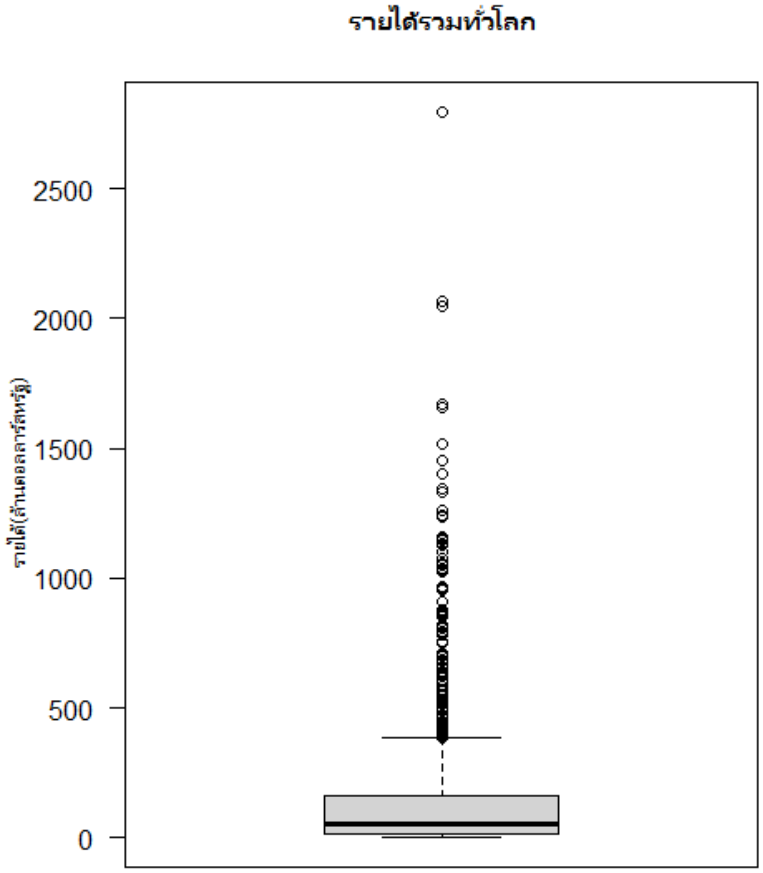
คอลัมน์ avg_vote (คะแนนโหวตเฉลี่ย)



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)

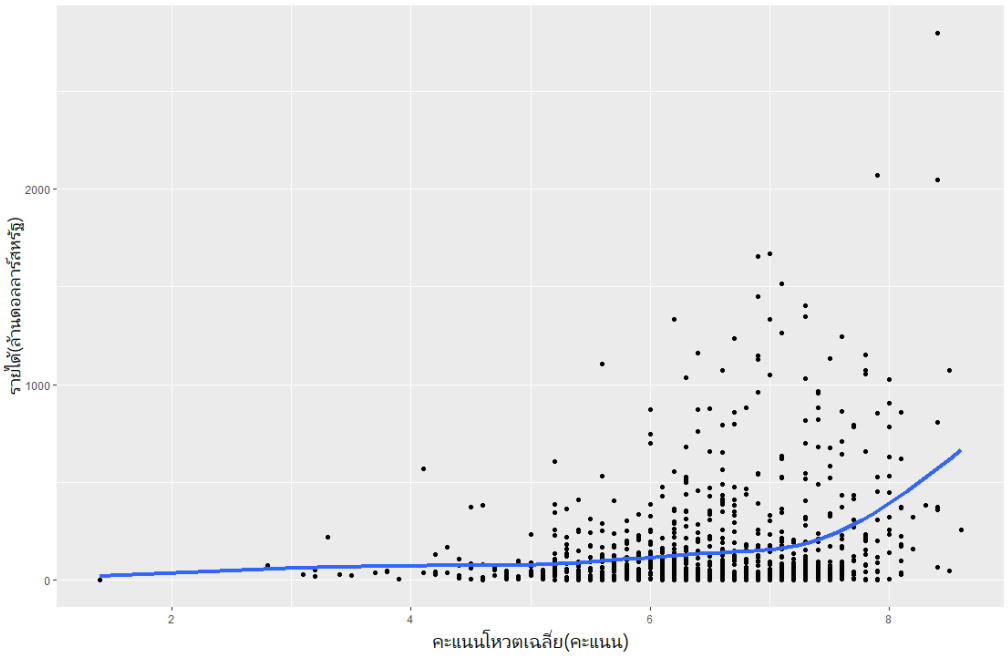


ค่าที่มีแนวโน้มจะเป็น outlier ได้แก่

```
> 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

คะแนนโหวตเฉลี่ยและรายได้รวมทั่วโลกของภาพยนตร์



จากข้อมูลจากกราฟความสัมพันธ์ระหว่างคะแนนโหวตเฉลี่ย รายได้และสาเหตุที่เลือก ตัวแปรต้นเป็น คะแนนโหวตเฉลี่ย และ ตัวแปรตามเป็น รายได้รวมทั่วโลก เพราะผมอยากรู้ว่าคะแนนโหวตจะส่งผลอย่างไรกับรายได้ของภาพยนตร์ สามารถวิเคราะห์ได้ว่าเมื่อคะแนนโหวตสูง รายได้ของภาพยนตร์ก็จะสูงด้วย ซึ่งผมคิดว่าอาจจะเกิดจากคนดูภาพยนตร์จะดูคะแนนโหวตของภาพยนตร์ก่อนไปดูภาพยนตร์เรื่องนั้น 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))
```

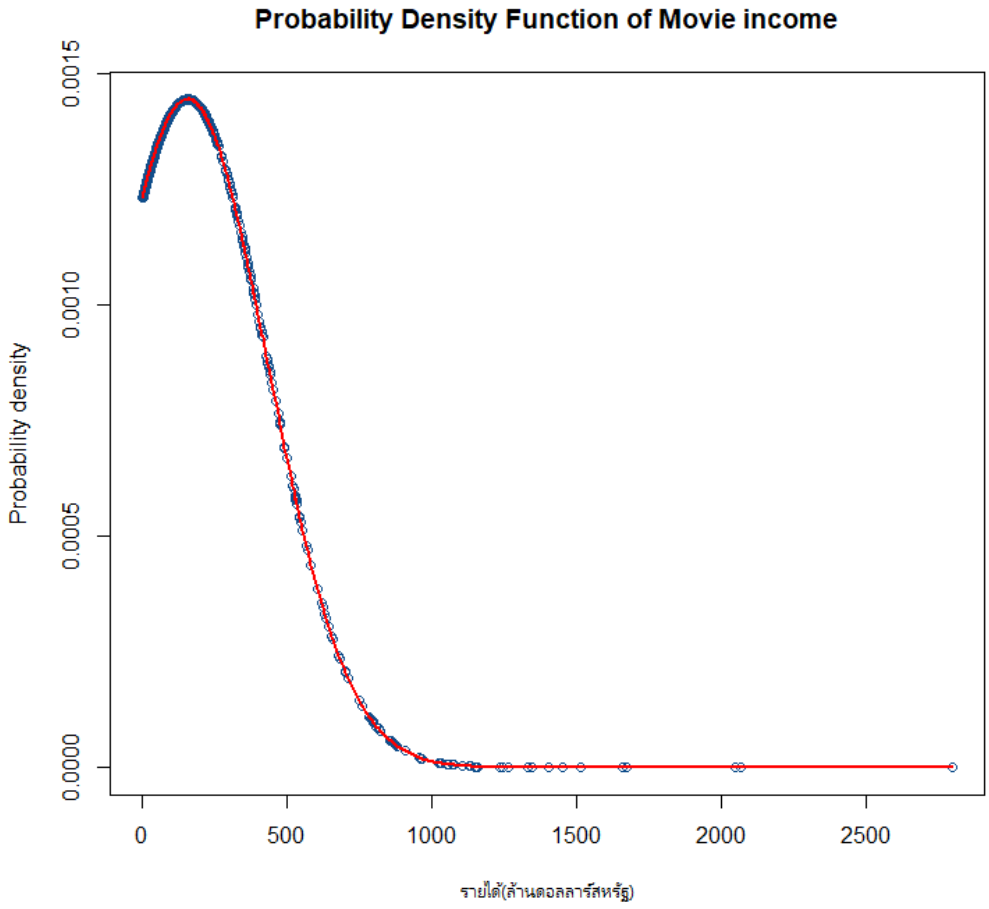
62010619 นายพัทธพล จันทร์ชู

HW#3 - Probability Density Function/Cumulative Prob Function

IMDb Movies

Probability Density Function

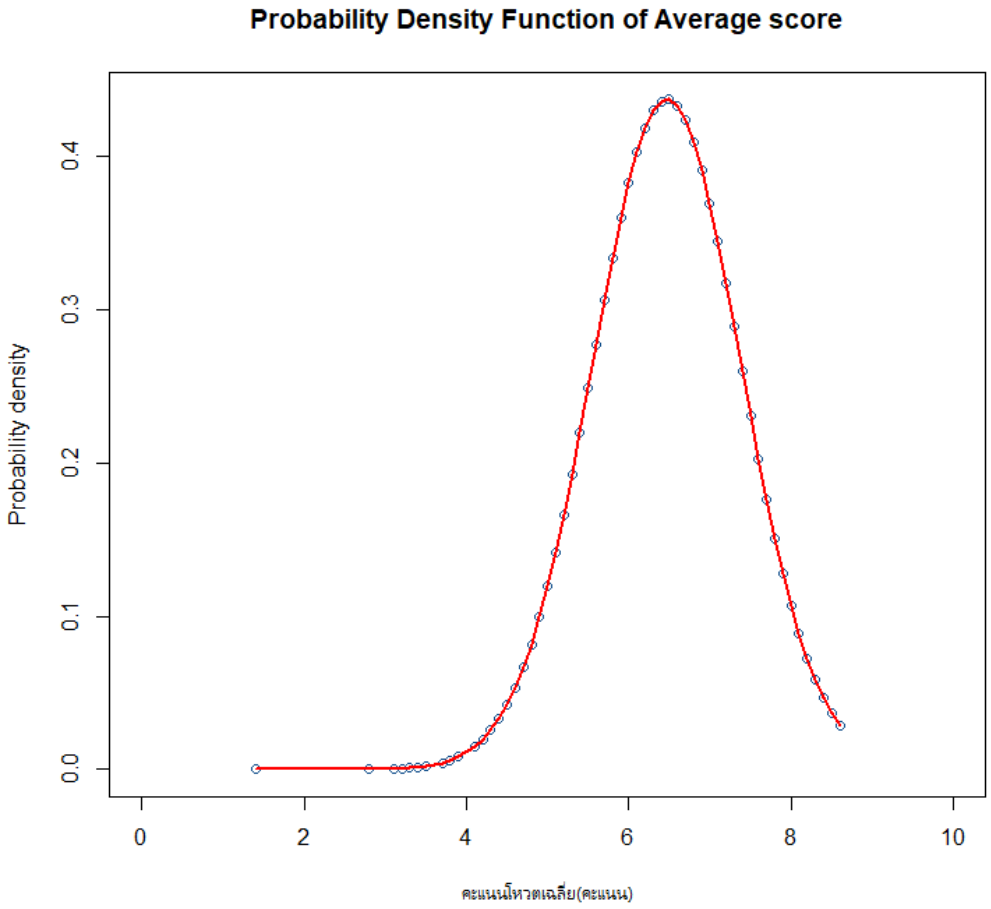
-รายได้ภาพยนตร์รวมทั่วโลก




```
pdfIncome = dnorm(x=income,mean=mean(income),sd = sd(income))

plot(income,pdfIncome,col = "dodgerblue4" ,main = "Probability Density Function of Movie income",ylab="Probability density"
     ,xlab = "รายได้(ล้านดอลลาร์สหรัฐ)")
lines(smooth.spline(income,pdfIncome), col='red',lwd=2)
```

แกน x เป็นรายได้(ล้านดอลลาร์สหรัฐ) แกน y เป็น ค่าความหนาแน่นที่สอดคล้องกับ mean และ sd ของรายได้(ล้านดอลลาร์สหรัฐ)

-คะแนนโหวตเฉลี่ย

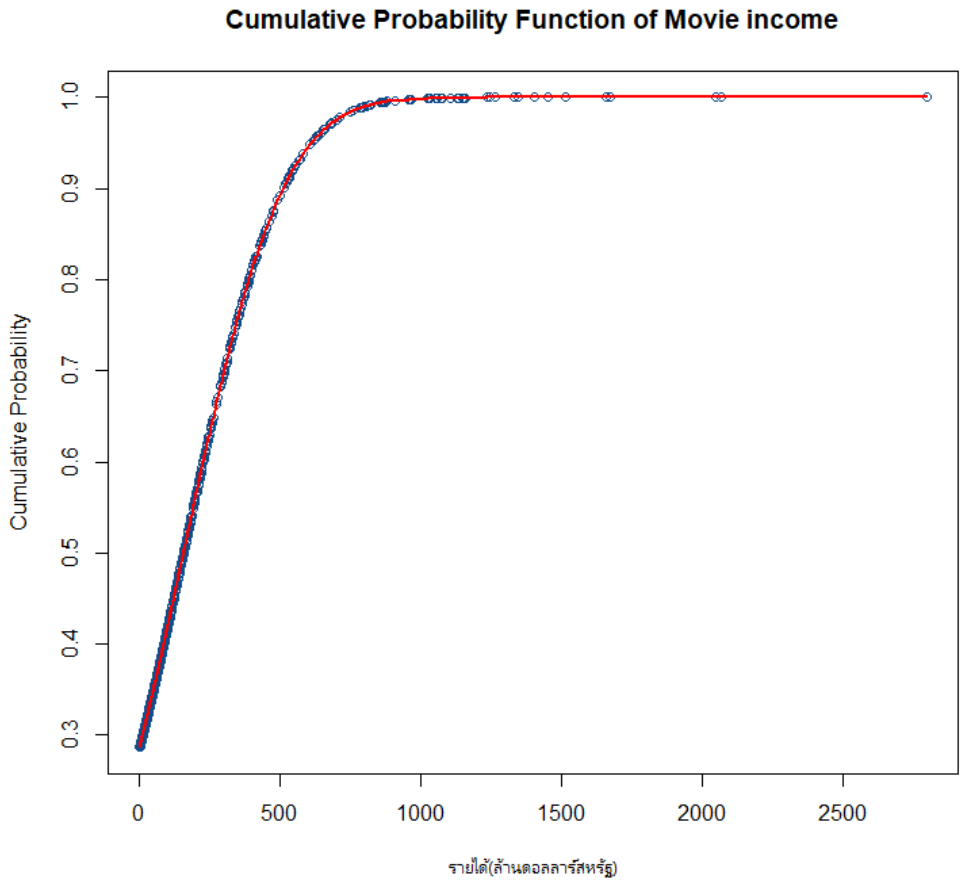



```
pdfAvgVote = dnorm(x=avg_vote,mean = mean(avg_vote),sd=sd(avg_vote))

plot(avg_vote,pdfAvgVote,xlim = c(0,10),col = "dodgerblue4" ,main = "Probability Density Function of Average score",ylab="Probability density"
     ,xlab = "คะแนนโหวตเฉลี่ย(คะแนน)")
lines(smooth.spline(avg_vote,pdfAvgVote), col='red',lwd=2)
```

แกน x เป็นคะแนนโหวตเฉลี่ย(คะแนน) แกน y เป็น ค่าความหนาแน่นที่สอดคล้องกับ mean และ sd ของคะแนนโหวตเฉลี่ย(คะแนน)

Cumulative Probability Function

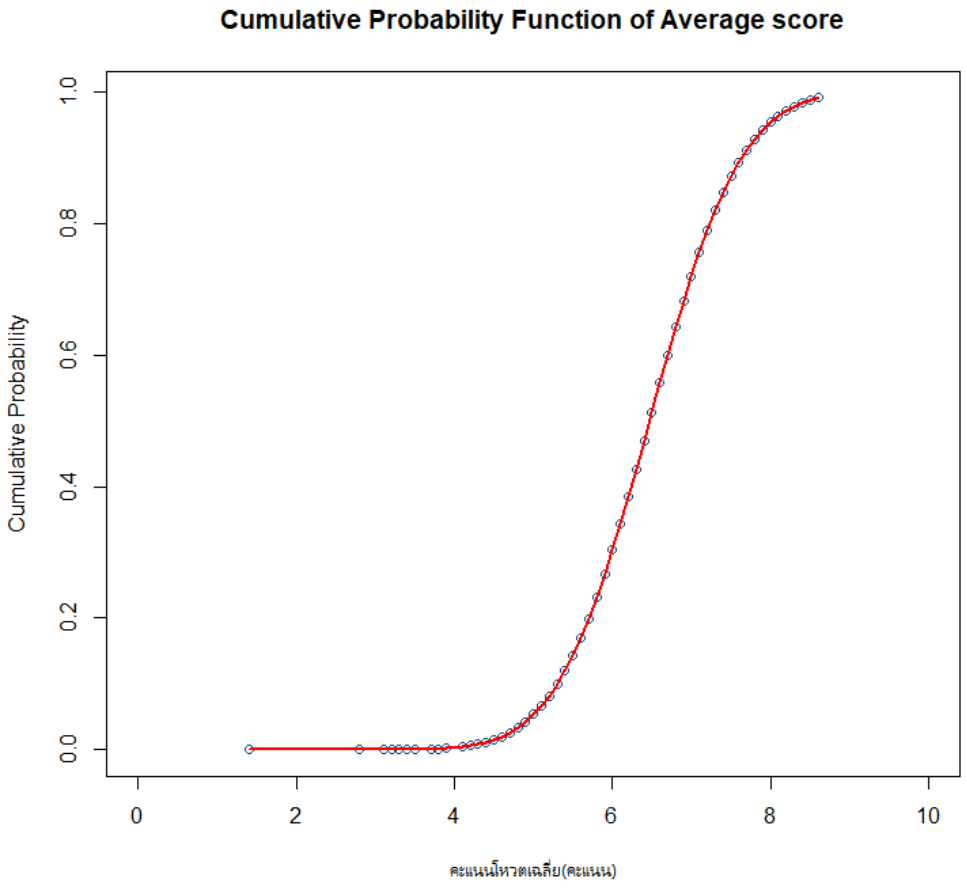
-รายได้ภาพยนตร์รวมทั่วโลก




```
cpfIncome = pnorm(q=income,mean = mean(income),sd=sd(income))

plot(income,cpfIncome,col = "dodgerblue4" ,main = "Cumulative Probability Function of Movie income",ylab="Cumulative Probability"
     ,xlab = "รายได้(ล้านดอลลาร์สหรัฐ)")
lines(smooth.spline(income,cpfIncome), col='red',lwd=2)
```

แกน x เป็นรายได้(ล้านดอลลาร์สหรัฐ) แกน y เป็น ค่าสะสมที่สอดคล้องกับ mean และ sd ของรายได้(ล้านดอลลาร์สหรัฐ)

-คะแนนโหวตเฉลี่ย




```
cpfAvgVote = pnorm(q=avg_vote,mean = mean(avg_vote),sd=sd(avg_vote))

plot(avg_vote,cpfAvgVote,xlim = c(0,10),col = "dodgerblue4" ,main = "Cumulative Probability Function of Average score",ylab="Cumulative Probability"
     ,xlab = "คะแนนโหวตเฉลี่ย(คะแนน)")
lines(smooth.spline(avg_vote,cpfAvgVote), col='red',lwd=2)
```

แกน x เป็นคะแนนโหวตเฉลี่ย(คะแนน) แกน y เป็น ค่าสะสมที่สอดคล้องกับ mean และ sd ของคะแนนโหวตเฉลี่ย(คะแนน)

62010619 นายพัทธพล จันทร์ชู

บทวิเคราะห์ข้อมูลจากกราฟ

Probability Density Function

- จากกราฟรายได้ของภาพยนตร์ จะเห็นได้ว่าในช่วงรายได้ประมาณ 100-400 ล้านดอลลาร์สหรัฐจะมีค่าความหนาแน่นมากที่สุด และในช่วงรายได้ประมาณ 400 ล้านดอลลาร์สหรัฐเป็นต้นไปก็จะค่อยๆลดลงเรื่อย ๆ วิเคราะห์ได้ว่า รายได้ของภาพยนตร์ส่วนใหญ่จะอยู่ในช่วงประมาณ 100-400 ล้านดอลลาร์สหรัฐ

-จากกราฟคะแนนโหวตเฉลี่ย จะเห็นได้ว่าในช่วงคะแนนโหวตเฉลี่ย 0-4 คะแนนค่าความหนาแน่นเพิ่มขึ้นน้อยมาก และเพิ่มขึ้นอย่างรวดเร็วในช่วงคะแนนโหวตเฉลี่ย 4-6 คะแนน และในช่วงคะแนนโหวตเฉลี่ย 6-7 คะแนนจะมีค่าความหนาแน่นจะมากที่สุด และในช่วงคะแนนโหวตเฉลี่ย 7 คะแนนเป็นต้นไปค่าความหนาแน่นก็จะลดลงเรื่อย ๆ วิเคราะห์ได้ว่า คะแนนโหวตเฉลี่ยของภาพยนตร์ส่วนใหญ่จะอยู่ในช่วง 6-7 คะแนน

Cumulative Probability Function

- จากกราฟรายได้ของภาพยนตร์ จะเห็นได้ว่าในช่วงรายได้ 0-500 ล้านดอลลาร์สหรัฐค่าสะสมจะเพิ่มขึ้นอย่างรวดเร็ว ในช่วง 500-1000 ล้านดอลลาร์สหรัฐค่าสะสมก็จะเพิ่มขึ้นช้าลง และในช่วง 1000 ล้านดอลลาร์สหรัฐเป็นต้นไปค่าสะสมจะเพิ่มขึ้นน้อยมาก วิเคราะห์ได้ว่า รายได้ของภาพยนตร์ส่วนใหญ่จะอยู่ในช่วง 0-500 ล้านดอลลาร์สหรัฐ

-จากกราฟคะแนนโหวตเฉลี่ยจะเห็นได้ว่าในช่วงคะแนนโหวตเฉลี่ย 0-5 คะแนนค่าสะสมที่เพิ่มขึ้นน้อยมาก แต่ในช่วงคะแนนโหวตเฉลี่ย 5-8 คะแนนค่าสะสมเพิ่มขึ้นอย่างรวดเร็ว และในช่วงคะแนนโหวตเฉลี่ย8 คะแนนขึ้นไปก็จะค่อยเพิ่มขึ้นช้าลง วิเคราะห์ได้คะแนนโหวตเฉลี่ยส่วนใหญ่จะอยู่ในช่วง 5-8 คะแนน

62010619 นายพัทธพล จันทร์ชู

Source Code

```
setwd("~/CE2D-2/git/Propstat")

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

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


pdfIncome = dnorm(x=income,mean=mean(income),sd = sd(income))

plot(income,pdfIncome,col = "dodgerblue4" ,main = "Probability Density Function of Movie income",ylab="Probability density"
     ,xlab = "รายได้(ล้านดอลลาร์สหรัฐ)")
lines(smooth.spline(income,pdfIncome), col='red',lwd=2)



pdfAvgVote = dnorm(x=avg_vote,mean = mean(avg_vote),sd=sd(avg_vote))

plot(avg_vote,pdfAvgVote,xlim = c(0,10),col = "dodgerblue4" ,main = "Probability Density Function of Average score",ylab="Probability density"
     ,xlab = "คะแนนโหวดเฉลี่ย(คะแนน)")
lines(smooth.spline(avg_vote,pdfAvgVote), col='red',lwd=2)



cpfIncome = pnorm(q=income,mean = mean(income),sd=sd(income))

plot(income,cpfIncome,col = "dodgerblue4" ,main = "Cumulative Probability Function of Movie income",ylab="Cumulative Probability"
     ,xlab = "รายได้(ล้านดอลลาร์สหรัฐ)")
lines(smooth.spline(income,cpfIncome), col='red',lwd=2)


cpfAvgVote = pnorm(q=avg_vote,mean = mean(avg_vote),sd=sd(avg_vote))

plot(avg_vote,cpfAvgVote,xlim = c(0,10),col = "dodgerblue4" ,main = "Cumulative Probability Function of Average score",ylab="Cumulative Probability"
     ,xlab = "คะแนนโหวดเฉลี่ย(คะแนน)")
lines(smooth.spline(avg_vote,cpfAvgVote), col='red',lwd=2)
```

HW4# - Confidence Interval (CI) of Mean

IMDb Movies

หา Confidence Interval (CI) ของคอลัมน์ คะแนนโหวตเฉลี่ย IMDb Movies

เนื่องจากภาพยนตร์ในโลกนี้มีเยอะมากไม่สามารถรวบรวมภาพยนตร์ทั้งหมดมาได้ ผมจึงสมมติว่าภาพยนตร์ทั้งหมดมี 1000 เรื่อง จะได้ค่า population mean (mu) ของคะแนนโหวตเฉลี่ย = 6.4693 คะแนน

```
nSample = 50
sampleAvg = sample(avg_vote,nSample)
sampleMean = mean(sampleAvg)
sampleSD = sd(sampleAvg)
```

ทำการสุ่มภาพยนตร์ตัวอย่างมา 50 เรื่อง จะได้ sample mean = 6.552 คะแนน และ sd = 0.8179716 คะแนน

-หา Confidence Interval (CI) ของแต่ละ Confidence Level

```
getCI <- function(cl,n,x){
  m <- mean(x)  # mean
  s <- sd(x)  # standard deviation

  # 1.standard error (SE)
  se <- s / sqrt(n)
  # 2.z-score
  z <- qnorm(cl)
  # 3.margin error
  me <- se * z
  # 4.confidence interval
  ci <- c(m - me, m + me)
  return(ci)
}
```

Confidence Level = 90%

-90% confidence interval = [6.403752, 6.700248] คะแนน

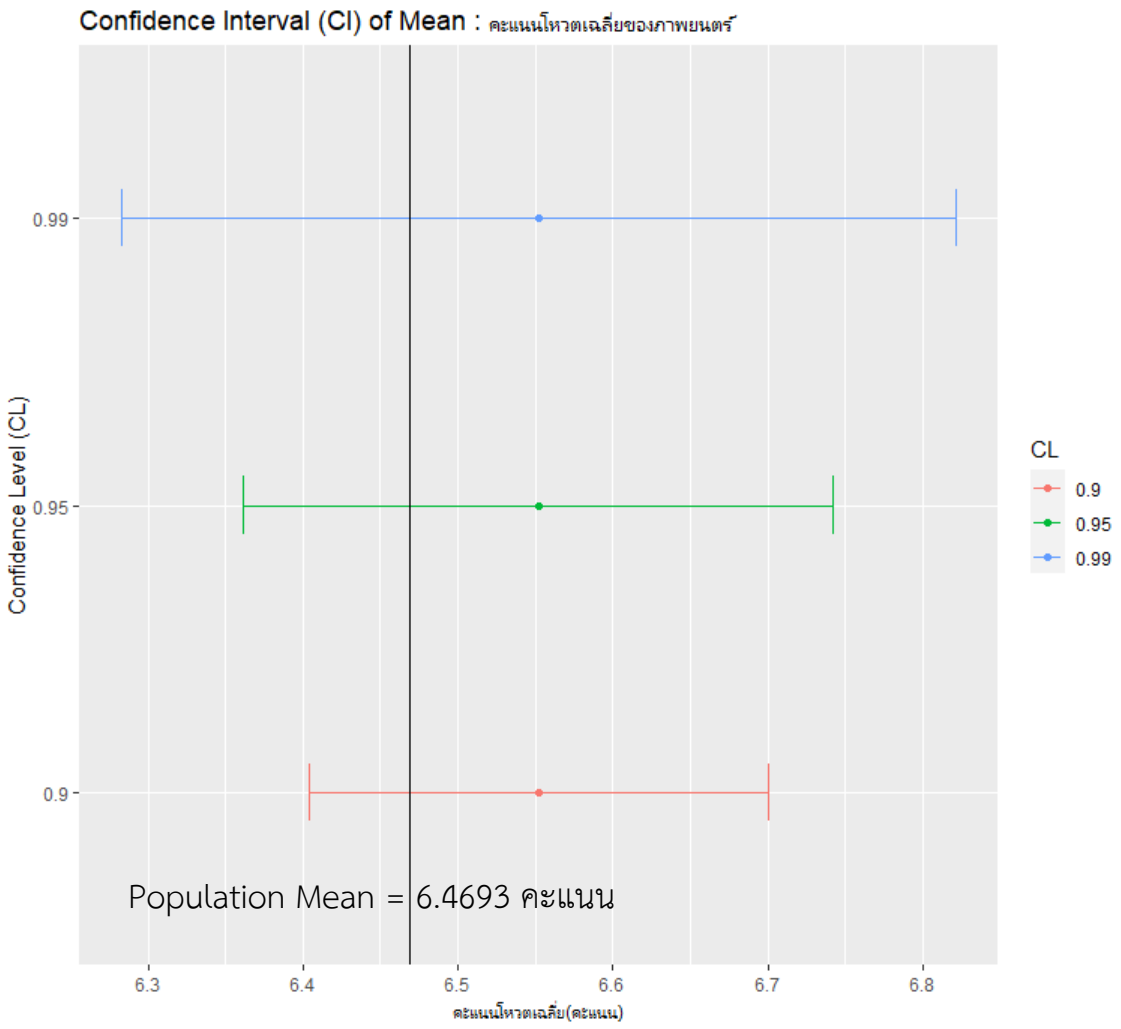
Confidence Level = 95%

-95% confidence interval = [6.361726, 6.742274] คะแนน

Confidence Level = 99%

-99% confidence interval = [6.282891, 6.821109] คะแนน

กราฟ Confidence Interval (CI) of Mean



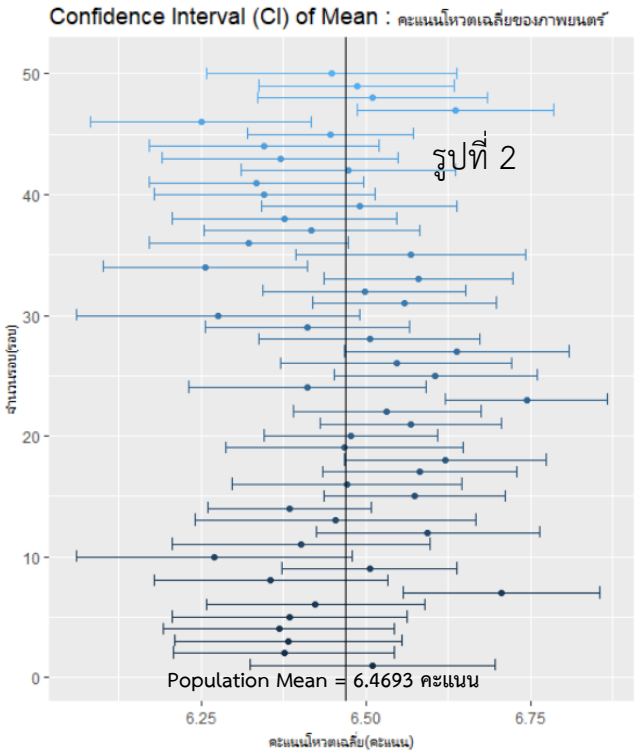
รูปที่ 1

```
cl = c(0.9,0.95,0.99)
d = data.frame(
  CL = c("0.9","0.95","0.99"),
  Mean = c(sampleMean,sampleMean,sampleMean),
  lower = c(getCI(cl[1],50,sampleAvg)[1],getCI(cl[2],50,sampleAvg)[1],getCI(cl[3],50,sampleAvg)[1]),
  upper = c(getCI(cl[1],50,sampleAvg)[2],getCI(cl[2],50,sampleAvg)[2],getCI(cl[3],50,sampleAvg)[2])
)

qplot(x    = Mean ,
      y    = CL,
      color = CL,
      data = d,main = "Confidence Interval (CI) of Mean : คะแนนโหวตเฉลี่ยของภาพยนตร์",xlab = "คะแนนโหวตเฉลี่ย(คะแนน)",
      ylab = "Confidence Level (CL)") +

  geom_errorbar(aes(
    xmin  = lower,
    xmax  = upper,
    width = 0.2))+ geom_vline(xintercept = mean(avg_vote))
```

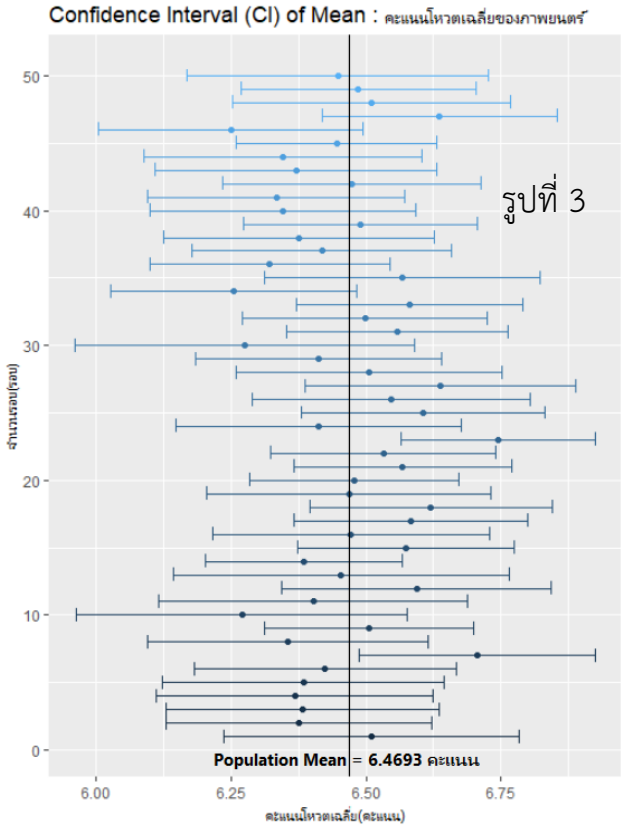
หา Confidence Interval

จากข้อมูลประชากร ทำการสุ่มข้อมูลมาจำนวน 50 รอบๆ ละ 50 ตัวอย่าง จะได้

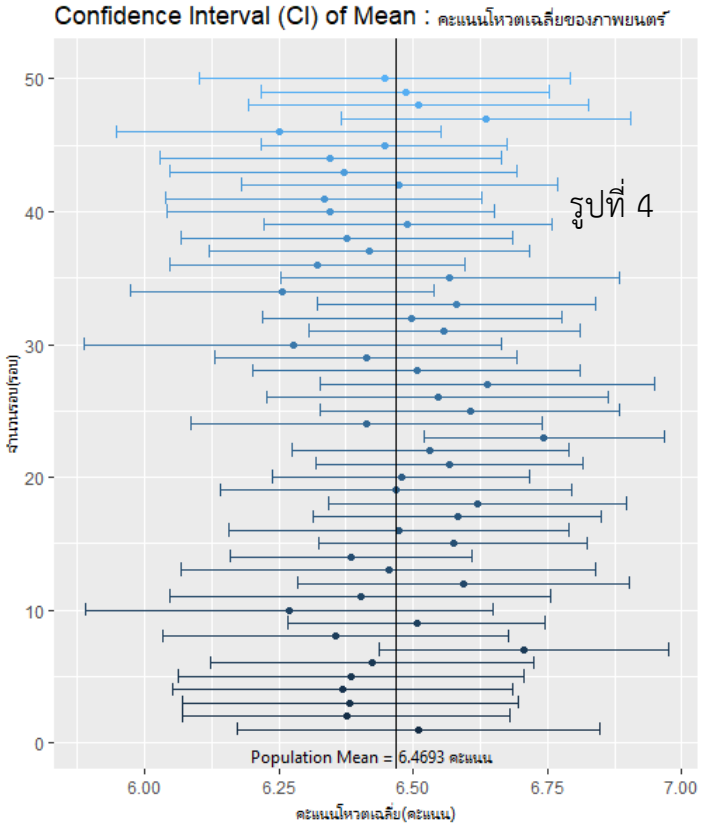
Confidence Interval จาก Confidence Level = 90% ของการสุ่มตัวอย่างรอบที่ 1 – 50



Confidence Interval จาก Confidence Level = 95% ของการสุ่มตัวอย่างรอบที่ 1 – 50



Confidence Interval จาก Confidence Level = 99% ของการสุ่มตัวอย่างรอบที่ 1 – 50



บทวิเคราะห์ข้อมูลจากกราฟ

วิเคราะห์ข้อมูลจากกราฟรูปที่ 1

จากกราฟ Confidence Interval (CI) of Mean ของคะแนนโหวตเฉลี่ยของภาพยนตร์ ซึ่งทำการสุ่มภาพยนตร์มาจำนวน 50 เรื่อง มี sample mean = 6.552 คะแนน และ sd = 0.8179716 คะแนน

จาก confidence interval ที่สร้างขึ้นมา ค่า population mean = 6.4693 คะแนน จะอยู่ในช่วง confidence interval ที่สร้างขึ้นมาทั้ง 3 ค่า

ถ้าค่า Confidence Level เยอะกว่าจะทำให้ confidence interval กว้างกว่า Confidence Level ที่มีค่าน้อยกว่า

วิเคราะห์ข้อมูลจากกราฟรูปที่ 2-4

ทุกครั้งที่เราสุ่มตัวอย่างใหม่ ค่าสถิติทั้งหมดไม่ว่าจะเป็นค่า mean, sd รวมถึง confidence interval ก็จะเปลี่ยนไปเรื่อย ๆ แต่ถ้าเราสุ่มซ้ำหลายๆครั้ง เช่น ทำซ้ำ 50 ครั้งและทำทุกอย่างเหมือนเดิม

จาก Confidence Level = 90% มี 45 ครั้ง ใน 50 ครั้งที่ ค่า population mean อยู่ในช่วง confidence interval ที่สร้างขึ้นมา หรือคิดเป็น 90 % และมี 10 % ที่ค่า population mean ไม่ได้อยู่ในช่วง confidence interval

จาก Confidence Level = 95% มี 48 ครั้ง ใน 50 ครั้งที่ ค่า population mean อยู่ในช่วง confidence interval ที่สร้างขึ้นมา หรือคิดเป็น 96 % และมี 4 % ที่ค่า population mean ไม่ได้อยู่ในช่วง confidence interval

จาก Confidence Level = 99% มี 49 ครั้ง ใน 50 ครั้งที่ ค่า population mean อยู่ในช่วง confidence interval ที่สร้างขึ้นมา หรือคิดเป็น 98 % และมี 2 % ที่ค่า population mean ไม่ได้อยู่ในช่วง Confidence Level

สามารถวิเคราะห์ได้ว่า ค่า Confidence Level = x % หมายถึง มีโอกาส x % โดยประมาณที่ confidence interval ที่สร้างขึ้นมาจะครอบคลุมค่า population mean

Source Code

```r
income <- df$income
avg_vote <- df$avg_vote

mean(avg_vote)
sd(avg_vote)

rounds=50
nSample = 50
arraySampleAvg = c()
arraySampleMean = c()
arraySampleSD = c()

for(i in 1:rounds){
  sampleVote = sample(avg_vote,nSample)
  arraySampleAvg[i] = c(data.frame(sampleVote))
  arraySampleMean[i] = c(mean(sampleVote))
  arraySampleSD[i] = c(sd(sampleVote))
}

getCI <- function(cl,n,x){
  m <- mean(x)  # mean
  s <- sd(x)  # standard deviation
  # 1.standard error (SE)
  se <- s / sqrt(n)
  # 2.z-score
  z <- qnorm(cl)
  # 3.margin error
  me <- se * z
  # 4.confidence interval
  ci <- c(m - me, m + me)
  return(ci)
}
```

```r

lowerOf90 = c()
upperOf90 = c()
meanOf90 = c()

for (i in 1:rounds) {
  lowerOf90[i] = getCI(0.90,nSample,arraySampleAvg[[i]])[1]
  upperOf90[i] = getCI(0.90,nSample,arraySampleAvg[[i]])[2]
  meanOf90[i] = mean(arraySampleMean[i])
}

d90 = data.frame(roundsArr,meanOf90,lowerOf90,upperOf90)
roundsArr = c(1:rounds)

qplot(x    = meanOf90 ,
      y    = roundsArr,
      color = roundsArr,
      data = d90,main = "Confidence Interval (CI) of Mean : คะแนนโหวดเฉลี่ยของภาพยนตร์",
      xlab = "คะแนนโหวดเฉลี่ย(คะแนน)",
      ylab = "จำนวนรอบ(รอบ)") +

  geom_errorbar(aes(
    xmin  = lowerOf90,
    xmax  = upperOf90,
    width = 1))+ geom_vline(xintercept = mean(avg_vote))

```

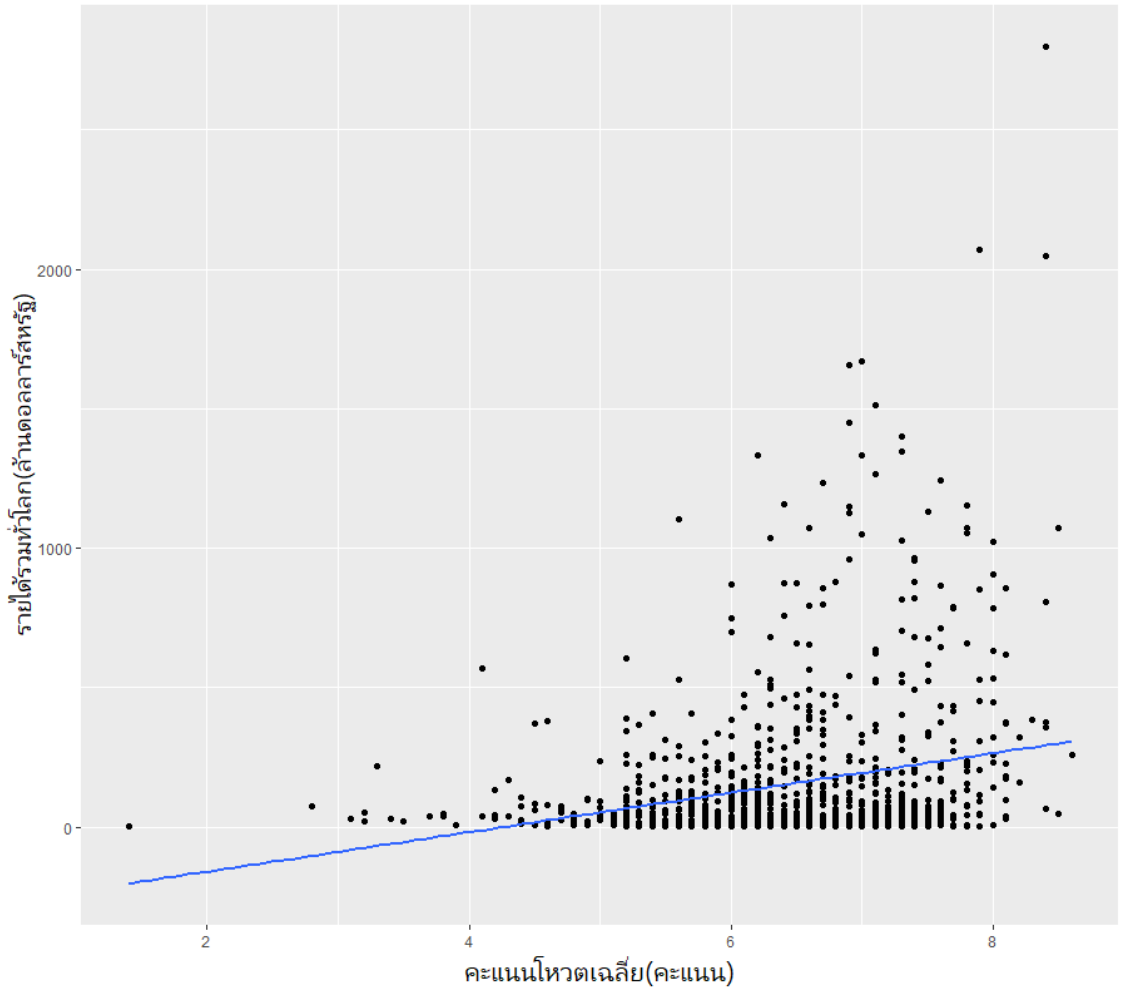
```r
cl = c(0.9,0.95,0.99)
d = data.frame(
  CL = c("0.9","0.95","0.99"),
  Mean = c(sampleMean,sampleMean,sampleMean),
  lower = c(getCI(cl[1],50,sampleAvg)[1],getCI(cl[2],50,sampleAvg)[1],getCI(cl[3],50,sampleAvg)[1]),
  upper = c(getCI(cl[1],50,sampleAvg)[2],getCI(cl[2],50,sampleAvg)[2],getCI(cl[3],50,sampleAvg)[2])
)

qplot(x    = Mean ,
      y    = CL,
      color = CL,
      data = d,main = "Confidence Interval (CI) of Mean : คะแนนโหวดเฉลี่ยของภาพยนตร์",
      xlab = "คะแนนโหวดเฉลี่ย(คะแนน)",
      ylab = "Confidence Level (CL)") +

  geom_errorbar(aes(
    xmin  = lower,
    xmax  = upper,
    width = 0.2))+ geom_vline(xintercept = mean(avg_vote))
```

HW5# - Linear Regression

IMDb Movies

เปรียบเทียบกราฟถดถอยเชิงเส้นกับกราฟ XY(Scatter) Plot ของข้อมูล คะแนนโหวตเฉลี่ย (independent) และ รายได้รวมทั่วโลก(dependent)

Linear Regression-(คะแนนโหวตเฉลี่ย และ รายได้รวมทั่วโลก) ของภาพยนตร์




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

```r
model <- lm(income ~ avg_vote,data = df)
model
```

```r
tidy(model)
```

term	estimate	std.error	statistic	p.value
<chr>	<dbl>	<dbl>	<dbl>	<dbl>
(Intercept)	-301.48906	60.875621	-4.952542	8.597981e-07
avg_vote	70.78831	9.317919	7.597009	6.960195e-14

จะได้สมการ y = -301.49 + 70.79 * x

y คือ income(ล้านดอลลาร์สหรัฐ) , x คือ avg_vote(คะแนน)

Correlation Coefficient

```r
r <- SSxy/(sqrt(SSxx)*sqrt(SSyy));r
```

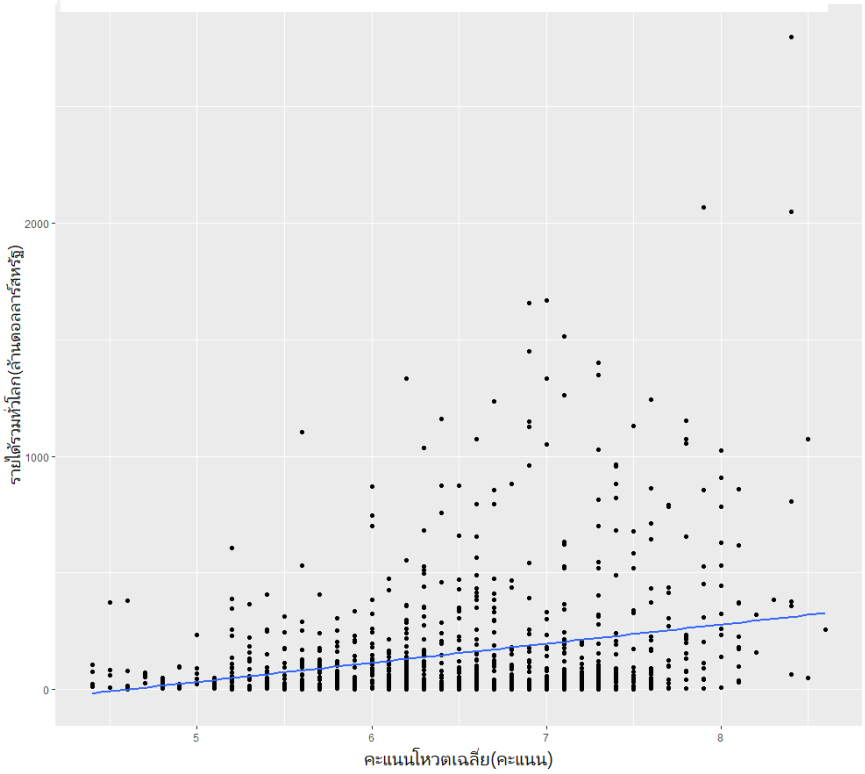
```
[1] 0.2338134
```

ได้ r = 0.233 เป็น Weak or No Correlation (r เป็นบวกและมีค่าเข้าใกล้ 0)

เนื่องจากค่า r มีค่าเข้าใกล้ศูนย์ซึ่งไม่ค่อยมีความสัมพันธ์ในแนวเส้นตรง ผมจึงเอา outlier ของ คะแนนโหวตเฉลี่ยออกได้ดังนี้

```
outliers <- function(x) {
  Q1 <- quantile(x, probs=.25)
  Q3 <- quantile(x, probs=.75)
  iqr = Q3-Q1
  upper_limit = Q3 + (iqr*1.5)
  lower_limit = Q1 - (iqr*1.5)
  x > upper_limit | x < lower_limit
}
remove_outliers <- function(df, cols = names(df)) {
  for (col in cols) {
    df <- df[!outliers(df[[col]]),]
  }
```

```
new_df = remove_outliers(df, c('avg_vote'))
```




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -377.46      68.92  -5.477 5.50e-08 ***
avg_vote       82.04      10.48   7.829 1.28e-14 ***
---
```

จะได้สมการ y = -377.46 + 82.04 * x

y คือ income(ล้านดอลลาร์สหรัฐ) , x คือ avg_vote(คะแนน)

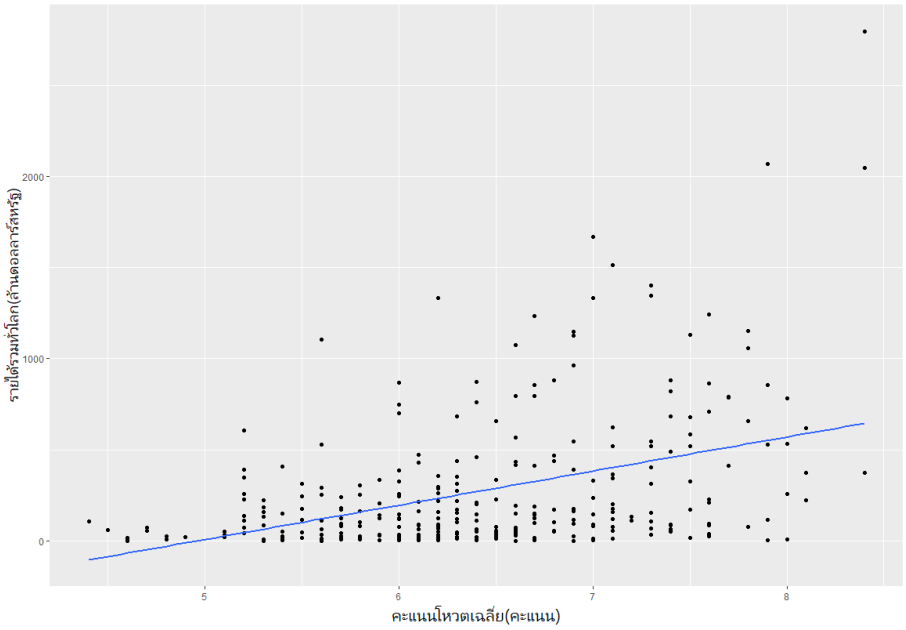
```
> SSxy <- sum(x*y) - n*xbar*ybar; SSxy
[1] 54376.78
> SSxx <- sum (x^2) - n*xbar^2;SSxx
[1] 662.7681
> SSyy <- sum(y^2) - n*ybar^2;SSyy
[1] 75724239
> r <- SSxy/(sqrt(SSxx)*sqrt(SSyy));r
[1] 0.2427254
```

ได้ r = 0.243 ซึ่งดีกว่าเดิมนิดนึงแต่ก็ยังไม่เป็นที่น่าพอใจ ผมจึงคิดว่าควรจะหาวิธีใหม่

แบ่งตามประเภทของภาพยนตร์ได้ดังนี้

```
df_genre = new_df %>%
  filter(grepl('Action',genre))
```

1.Action




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -929.4      161.0  -5.774 1.94e-08 ***
avg_vote       187.5       24.8   7.561 5.02e-13 ***
```

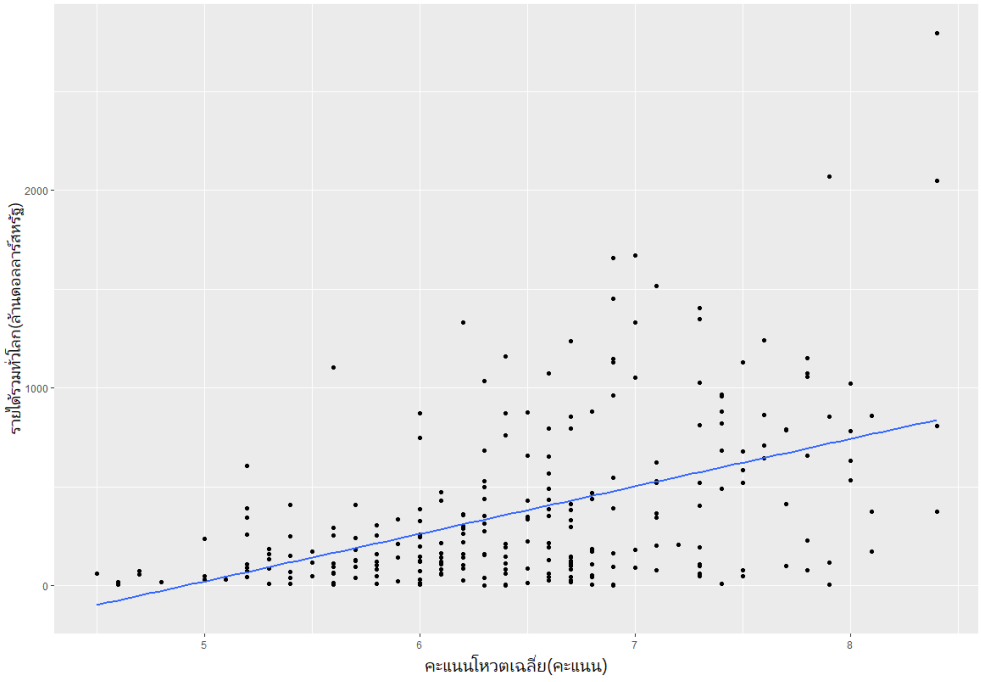
จะได้สมการ y = -929.4 + 187.5 * x

y คือ income(ล้านดอลลาร์สหรัฐ) , x คือ avg_vote(คะแนน)

```
> x=avg_vote
> y=income
> xbar <- mean(x)
> ybar <- mean (y)
> n <- length(y)
>
> SSxy <- sum(x*y) - n*xbar*ybar; SSxy
[1] 37730.99
> SSxx <- sum (x^2) - n*xbar^2;SSxx
[1] 201.2543
> SSyy <- sum(y^2) - n*ybar^2;SSyy
[1] 43822064
> r <- SSxy/(sqrt(SSxx)*sqrt(SSyy));r
[1] 0.4017716
```

ได้ r = 0.402

2. Adventure




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1181.35     186.30  -6.341 1.03e-09 ***
avg_vote      240.48      28.46   8.449 2.26e-15 ***
```

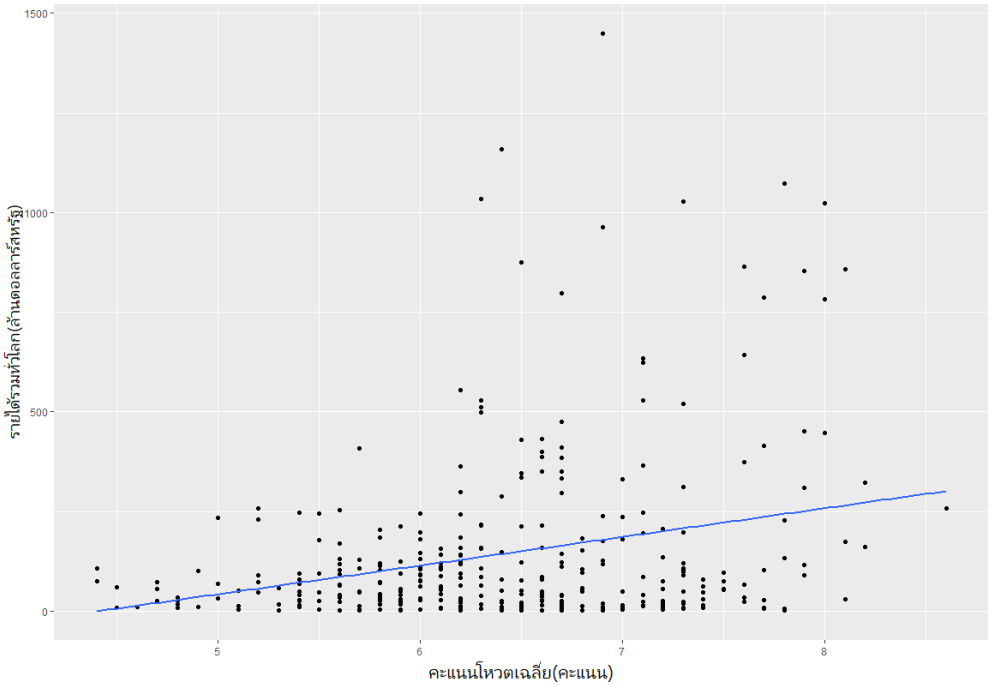
จะได้สมการ y = -1181.35 + 240.48 * x

y คือ income(ล้านดอลลาร์สหรัฐ) , x คือ avg_vote(คะแนน)

```
> SSxy <- sum(x*y) - n*xbar*ybar; SSxy
[1] 42795.5
> SSxx <- sum (x^2) - n*xbar^2;SSxx
[1] 177.956
> SSyy <- sum(y^2) - n*ybar^2;SSyy
[1] 47051103
> r <- SSxy/(sqrt(SSxx)*sqrt(SSyy));r
[1] 0.4676888
```

ได้ r= 0.468

3. Comedy




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -317.00      91.19  -3.476 0.000576 ***
avg_vote       71.89      14.08   5.105 5.57e-07 ***
---
```

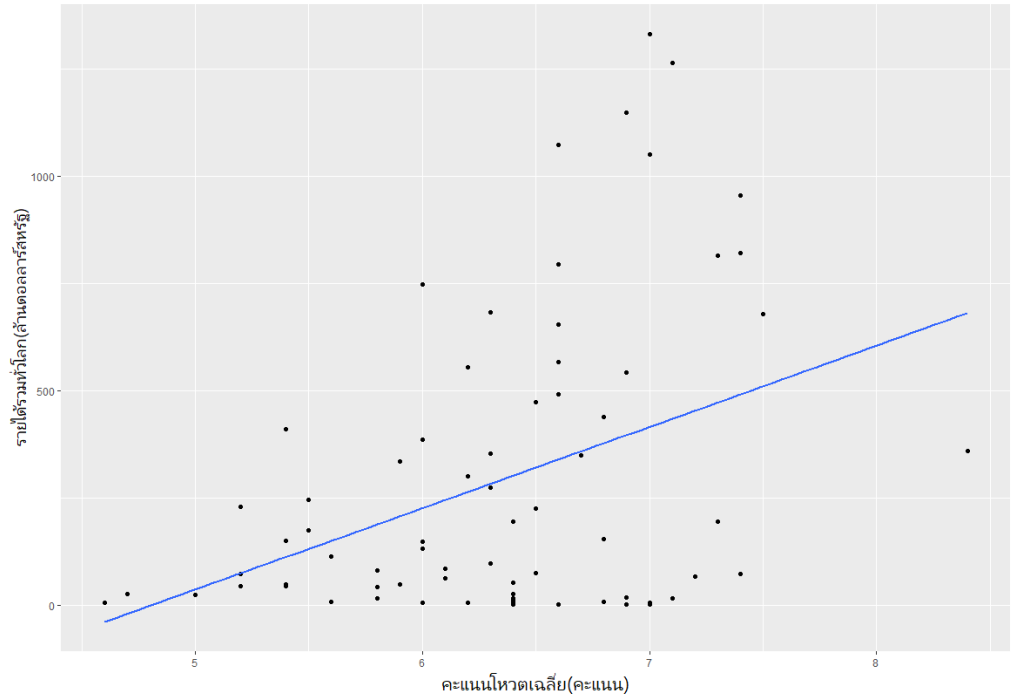
จะได้สมการ y = -1181.35 + 240.48 * x

y คือ income(ล้านดอลลาร์สหรัฐ) , x คือ avg_vote(คะแนน)

```
> SSxy <- sum(x*y) - n*xbar*ybar; SSxy
[1] 16045.15
> SSxx <- sum (x^2) - n*xbar^2;SSxx
[1] 223.1947
> SSyy <- sum(y^2) - n*ybar^2;SSyy
[1] 15937029
> r <- SSxy/(sqrt(SSxx)*sqrt(SSyy));r
[1] 0.2690284
```

ได้ r = 0.269

4.Fantasy




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -912.3      340.1  -2.682 0.009120 **
avg_vote       189.6       53.3   3.558 0.000676 ***
---
```

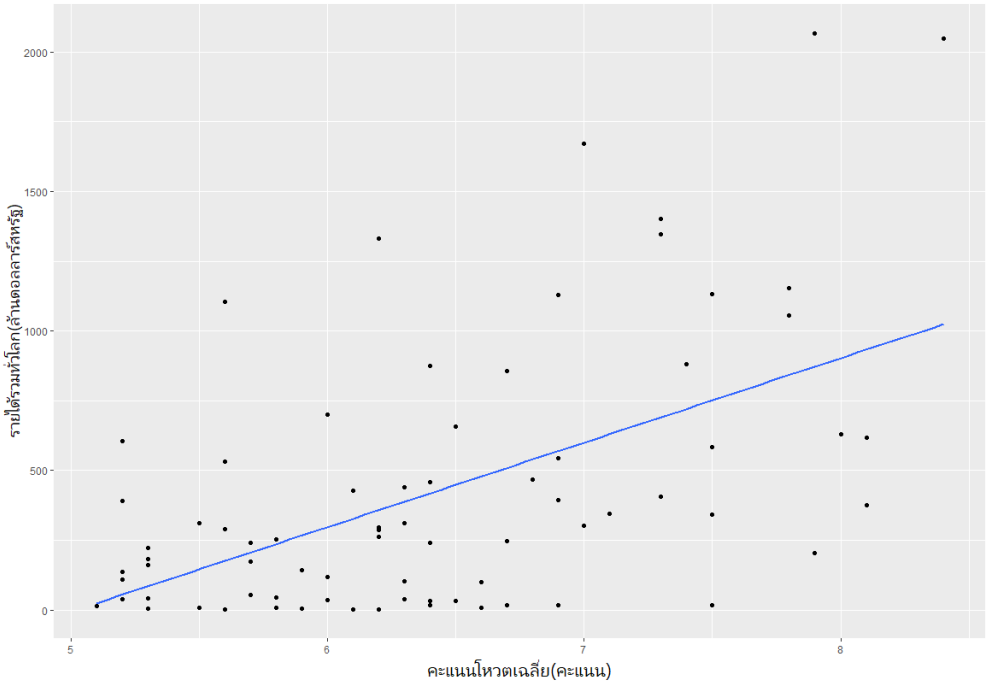
จะได้สมการ y = -912.3 + 189.6 * x

y คือ income(ล้านดอลลาร์สหรัฐ) , x คือ avg_vote(คะแนน)

```
> SSxy <- sum(x*y) - n*xbar*ybar; SSxy
[1] 6992.718
> SSxx <- sum (x^2) - n*xbar^2;SSxx
[1] 36.875
> SSyy <- sum(y^2) - n*ybar^2;SSyy
[1] 8657970
> r <- SSxy/(sqrt(SSxx)*sqrt(SSyy));r
[1] 0.391356
```

ได้ r= 0.391

5.Sci-Fi




```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1521.20     355.59  -4.278 5.50e-05 ***
avg_vote      302.99      55.09   5.500 5.04e-07 ***
---
```

จะได้สมการ y = -1521.20 + 302.99 * x

y คือ income(ล้านดอลลาร์สหรัฐ) , x คือ avg_vote(คะแนน)

```
> SSxy <- sum(x*y) - n*xbar*ybar; SSxy
[1] 16752.06
> SSxx <- sum (x^2) - n*xbar^2;SSxx
[1] 55.28987
> SSyy <- sum(y^2) - n*ybar^2;SSyy
[1] 17661087
> r <- SSxy/(sqrt(SSxx)*sqrt(SSyy));r
[1] 0.5360887
```

ได้ r = 0.536

จากที่ได้ลองแบ่งตามประเภทของภาพยนตร์ ค่า r ที่ได้ดีขึ้นกว่าเดิมพอสมควร

62010619 นายพัทธพล จันทร์ชู

บทวิเคราะห์ข้อมูลจากกราฟ

จากข้อมูลจากกราฟ Linear Regression - (คะแนนโหวตเฉลี่ย และ รายได้รวมทั่วโลก) ของภาพยนตร์วิเคราะห์ได้ว่า เนื่องจากกราฟมีความชันเป็นบวกเมื่อ คะแนนโหวตมีค่าเพิ่มขึ้น รายได้รวมก็จะมีค่าเพิ่มขึ้นด้วย

จากข้อมูลดิบ ได้ r = 0.234 เป็น Weak หรือ No Correlation ซึ่งไม่ค่อยมีความสัมพันธ์ในเชิงเส้นตรง

จากการตัด Outlier ของคะแนนโหวตเฉลี่ยออก ได้ r=0.243 ซึ่งถือว่าดีกว่าเดิม

จากการแบ่งตามประเภทภาพยนตร์ ซึ่งทำให้ได้จำนวนกราฟหลายกราฟ เนื่องจากภาพยนตร์มีหลายประเภท ซึ่งได้ค่า r ดังนี้

1.Action ได้ r = 0.402

2. Adventure ได้ r = 0.468

3.Comedy ได้ r = 0.269

4.Fantasy ได้ r = 0.391

5.Sci-Fi ได้ r = 0.536

ซึ่งถือว่าค่า r ดีกว่าเดิมมาก ทำให้ linear correlation ดีขึ้นกว่าเดิม

ผมคิดว่าเวลาจะวิเคราะห์ข้อมูลต่างๆเราควรมองหลายๆปัจจัย เพื่อช่วยให้เราวิเคราะห์ข้อมูลได้ง่ายและดีขึ้น

Source Code

```
library(tidyverse)
library(broom)
library(psych)
library(modelr)
library(ggfortify)
setwd("~/CE2D-2/git/Propstat")

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

income <- df$income
avg_vote <- df$avg_vote
```

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

```
model <- lm(income ~ avg_vote,data = df)
model
```

```
tidy(model)
```

```
x=avg_vote
y=income
```

```
xbar <- mean(x)
ybar <- mean (y)
n <- length(y)
```

```
SSxy <- sum(x*y) - n*xbar*ybar; SSxy
```

```
SSxx <- sum (x^2) - n*xbar^2;SSxx
```

```
SSyy <- sum(y^2) - n*ybar^2;SSyy
```

```
r <- SSxy/(sqrt(SSxx)*sqrt(SSyy));r
```

```
outliers <- function(x) {
  Q1 <- quantile(x, probs=.25)
  Q3 <- quantile(x, probs=.75)
  iqr = Q3-Q1
  upper_limit = Q3 + (iqr*1.5)
  lower_limit = Q1 - (iqr*1.5)
  x > upper_limit | x < lower_limit
}
remove_outliers <- function(df, cols = names(df)) {
  for (col in cols) {
    df <- df[!outliers(df[[col]]),]
  }
  df
}
```

```
new_df = remove_outliers(df, c('avg_vote'))
```

```
df_genre = new_df %>%
  filter(grepl('Action',genre))
```

62010619 นายพัทธพล จันทร์ชู

สรุปผลการศึกษาและเสนอแนะแนวทางการศึกษาเพิ่มเติม

จากการศึกษาข้อมูล IMDb Movies จากภาพยนตร์จำนวน 1000 เรื่องสามารถสรุปได้ว่า

คะแนนโหวต และ ประเภทของภาพยนตร์ มีผลต่อรายได้ของภาพยนตร์ คือภาพยนตร์ที่มีคะแนนโหวตเยอะก็จะมีรายได้เยอะ และ ภาพยนตร์ที่มีรายได้เยอะส่วนใหญ่จะเป็นภาพยนตร์แนว Action , Adventure และ Sci-Fi

เสนอแนะแนวทางการศึกษาเพิ่มเติม

1)ควรจะศึกษาปัจจัยด้านอื่น ๆ ของภาพยนตร์เพิ่มเติม เพื่อให้สามารถวิเคราะห์ข้อมูลได้ดียิ่งขึ้น เช่น ภาพยนตร์เรื่องนี้มาจากประเทศอะไรเป็นต้น

2)ควรศึกษาข้อมูลเกี่ยวกับเศรษฐกิจในช่วงที่ภาพยนตร์แต่ละเรื่องแสดง เพื่อใช้ในการวิเคราะห์ข้อมูลเกี่ยวกับรายได้ของภาพยนตร์ได้ดียิ่งขึ้น

3)ถ้าทำงานเกี่ยวกับสถิติในอนาคตผมคิดว่าเราควรมองปัจจัยอื่นๆของเรื่องที่ทำ และเรื่องอื่นๆที่อาจจะเกี่ยวของกับเรื่องที่ทำ เพื่อให้สามารถวิเคราะห์ข้อมูลได้ดียิ่งขึ้น