**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

**HW2# - Plots and Basic Statistics**

**IMDb Movies**



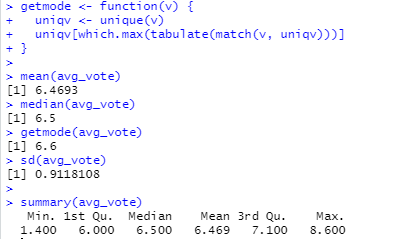
avg\_vote (คะแนนโหวตเฉลี่ย) มีหน่วยเป็น คะแนน

income (รายได้ภาพยนตร์รวมทั่วโลก) มีหน่วยเป็น ล้านดอลลาร์สหรัฐ



**คำนวณหาค่าสถิติพื้นฐาน**

คอลัมน์ avg\_vote (คะแนนโหวตเฉลี่ย)



Mean = 6.4693 คะแนน

Median = 6.5 คะแนน

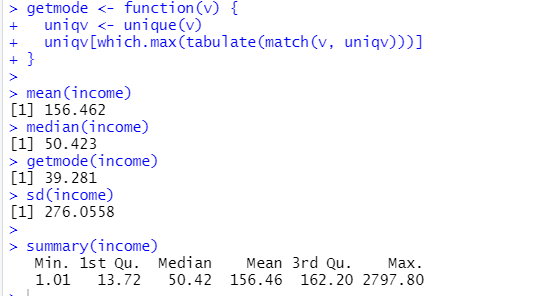
Mode = 6.6 คะแนน

S.D. = 0.9118 คะแนน

Min = 1.4 คะแนน

Max = 8.6 คะแนน

คอลัมน์ income(รายได้รวมทั่วโลก)



Mean = 156.462 ล้านดอลลาร์สหรัฐ

Median = 50.423 ล้านดอลลาร์สหรัฐ

Mode = 39.281 ล้านดอลลาร์สหรัฐ

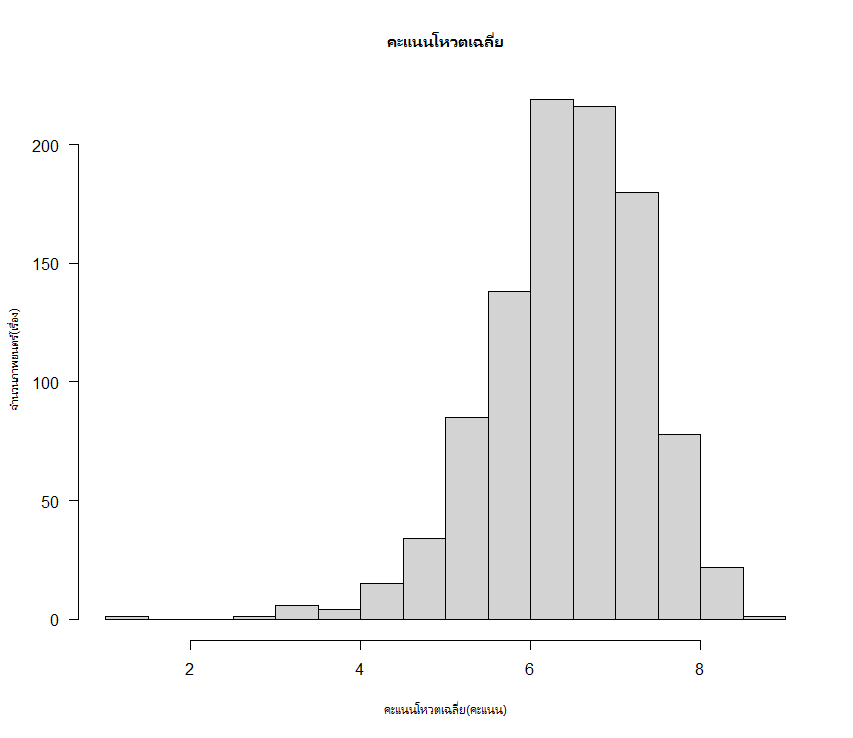
S.D. = 276.0558 ล้านดอลลาร์สหรัฐ

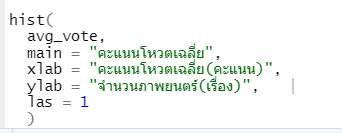
Min = 1.01 ล้านดอลลาร์สหรัฐ

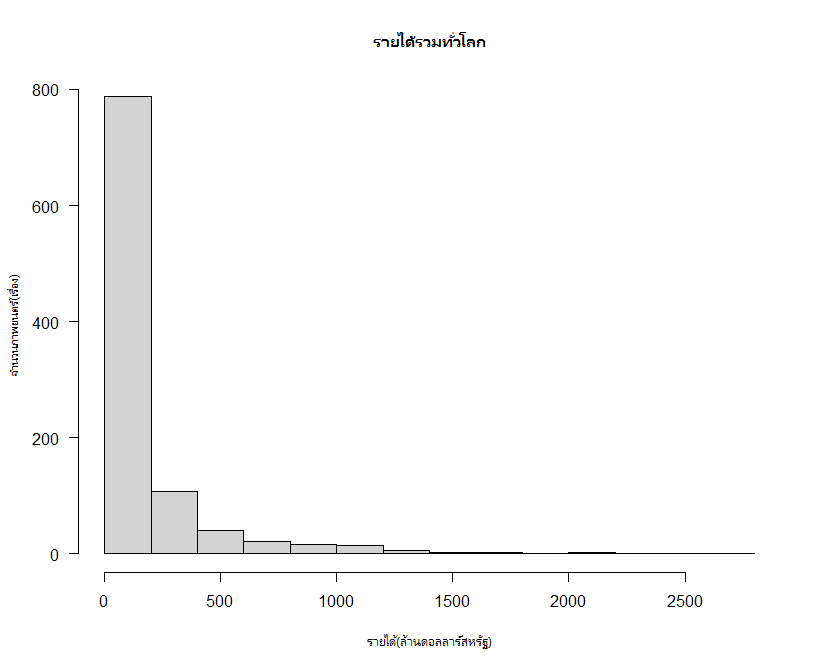
Max = 2797.80 ล้านดอลลาร์สหรัฐ

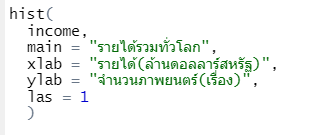
**วาดกราฟ**

- Histogram

คอลัมน์ avg\_vote (คะแนนโหวตเฉลี่ย)

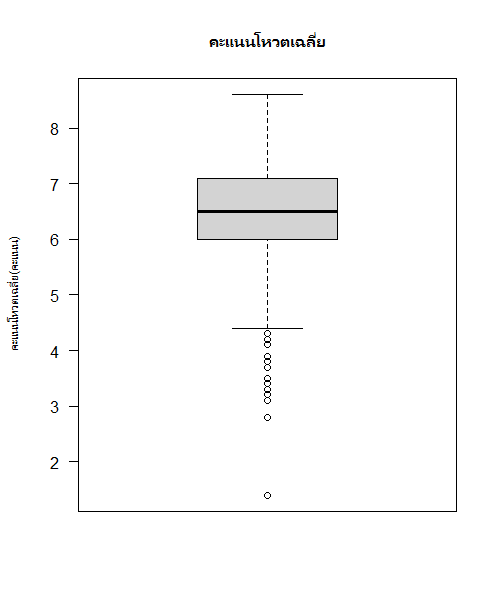


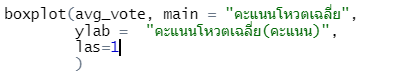
คอลัมน์ income(รายได้รวมทั่วโลก)



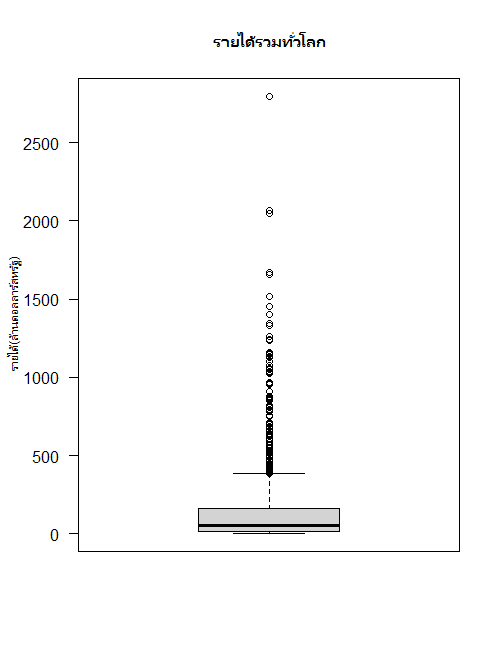
- Box Plot

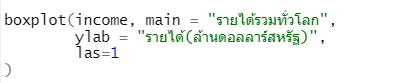
คอลัมน์ avg\_vote (คะแนนโหวตเฉลี่ย)





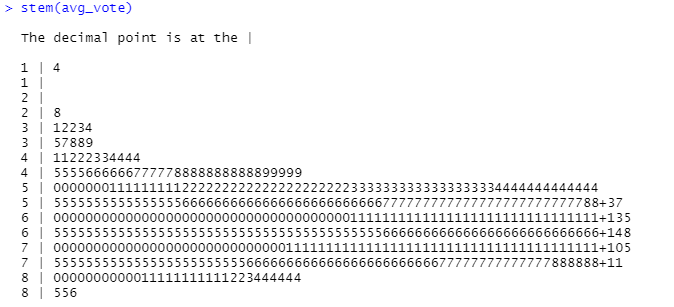
คอลัมน์ income(รายได้รวมทั่วโลก)





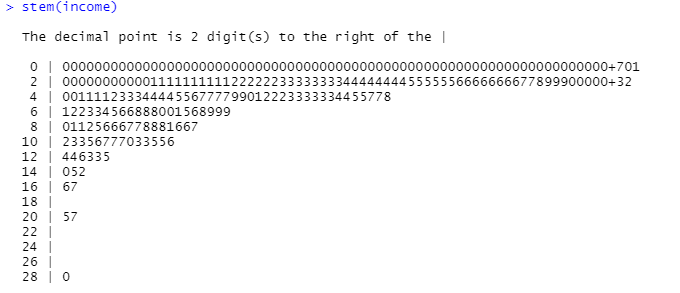
-Stem and Leave

คอลัมน์ avg\_vote (คะแนนโหวตเฉลี่ย)

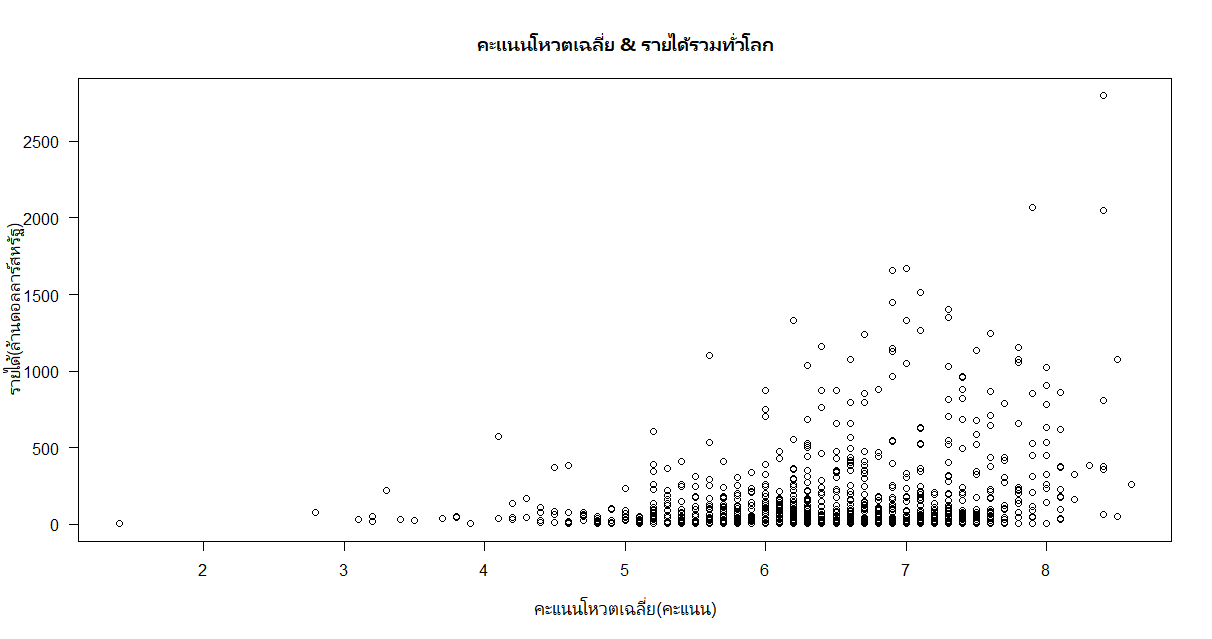


คะแนนโหวตเฉลี่ย(คะแนน)

คอลัมน์ income(รายได้รวมทั่วโลก)



รายได้(ล้านดอลลาร์สหรัฐ)

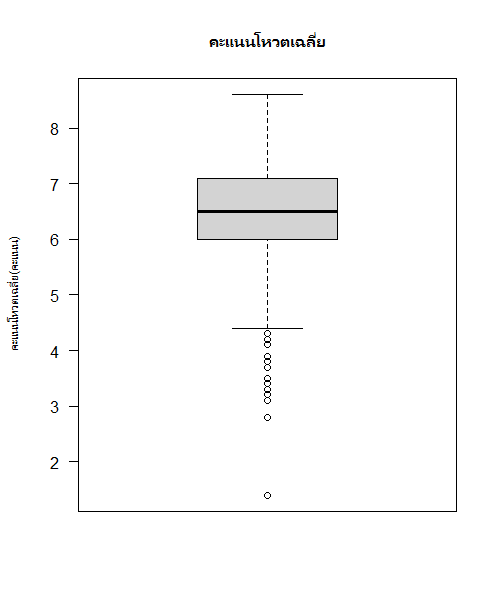
 -XY (Scatter) Plot

plot(avg\_vote,income)

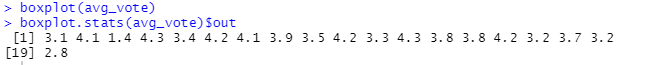
สาเหตุที่เลือก ตัวแปรต้นเป็น คะแนนโหวตเฉลี่ย และ ตัวแปรตามเป็น รายได้รวมทั่วโลก เพราะผมอยากรู้ว่าคะแนนโหวตจะส่งผลอย่างไรกับรายได้ของภาพยนตร์

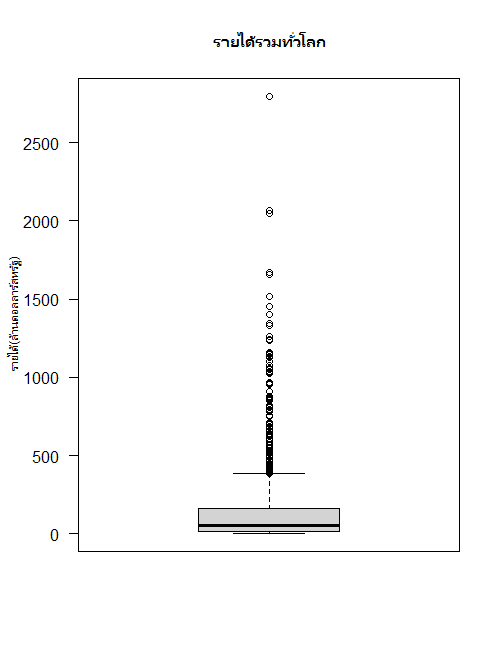
-Outlier

คอลัมน์ avg\_vote (คะแนนโหวตเฉลี่ย)

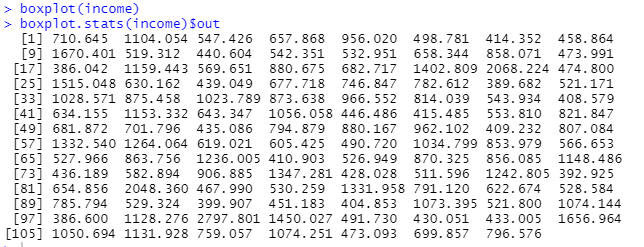


ค่าที่มีแนวโน้มจะเป็น outlier ได้แก่



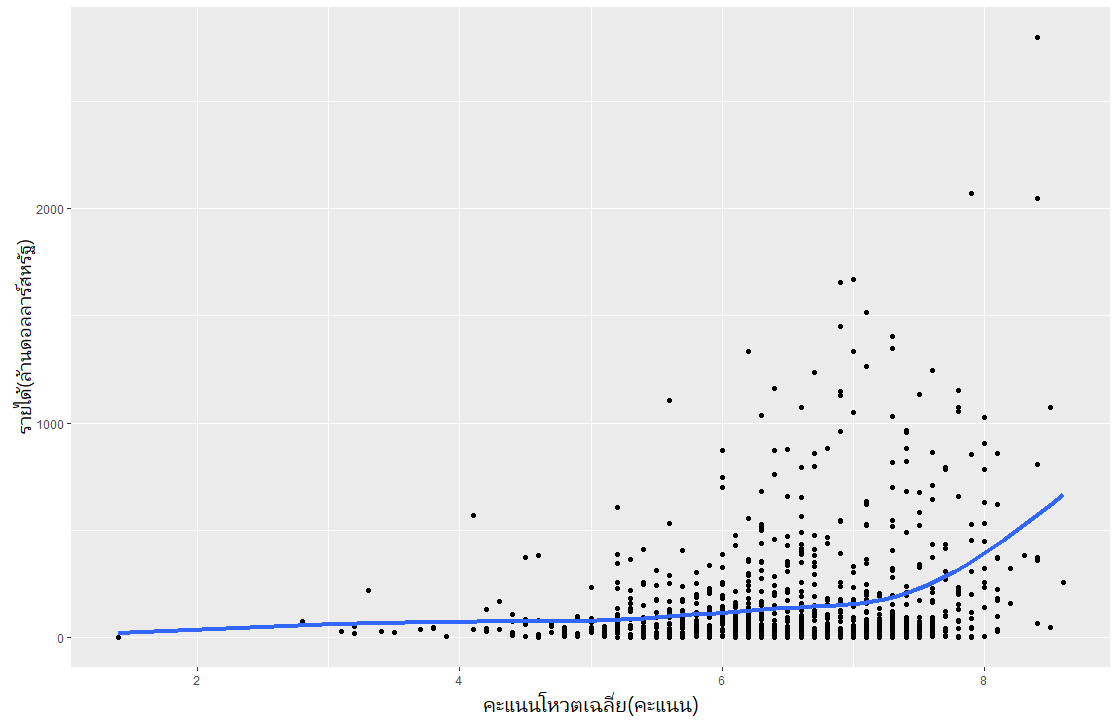
คอลัมน์ income(รายได้รวมทั่วโลก)

ค่าที่มีแนวโน้มจะเป็น outlier ได้แก่



**บทวิเคราะห์ข้อมูลจากกราฟ**

คะแนนโหวตเฉลี่ยและรายได้รวมทั่วโลกของภาพยนตร์



จากข้อมูลจากกราฟความสัมพันธ์ระหว่างคะแนนโหวตเฉลี่ย รายได้และสาเหตุที่เลือก ตัวแปรต้นเป็น คะแนนโหวตเฉลี่ย และ ตัวแปรตามเป็น รายได้รวมทั่วโลก เพราะผมอยากรู้ว่าคะแนนโหวตจะส่งผลอย่างไรกับรายได้ของภาพยนตร์ สามารถวิเคราะห์ได้ว่าเมื่อคะแนนโหวตสูง รายได้ของภาพยนตร์ก็จะสูงด้วย ซึ่งผมคิดว่าอาจจะเกิดจากคนดูภาพยนตร์จะดูคะแนนโหวตของภาพยนตร์ก่อนไปดูภาพยนตร์เรื่องนั้น ถ้าคะแนนเยอะก็จะไปดู ทำให้รายได้ของภาพยนตร์เรื่องนั้นๆสูง ถ้าคะแนนน้อยก็จะไม่ดู ทำให้รายได้ของภาพยนตร์เรื่องนั้นๆต่ำ

**Source Code**



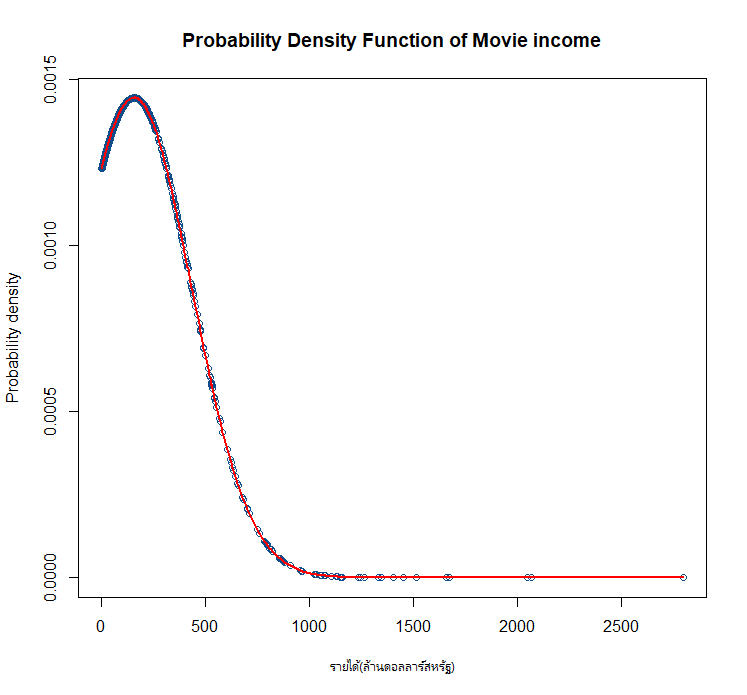


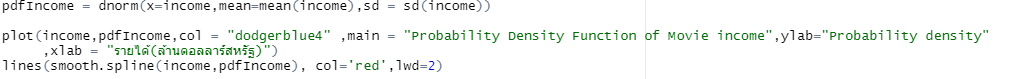
**HW#3 - Probability Density Function/Cumulative Prob Function**

**IMDb Movies**

**Probability Density Function**

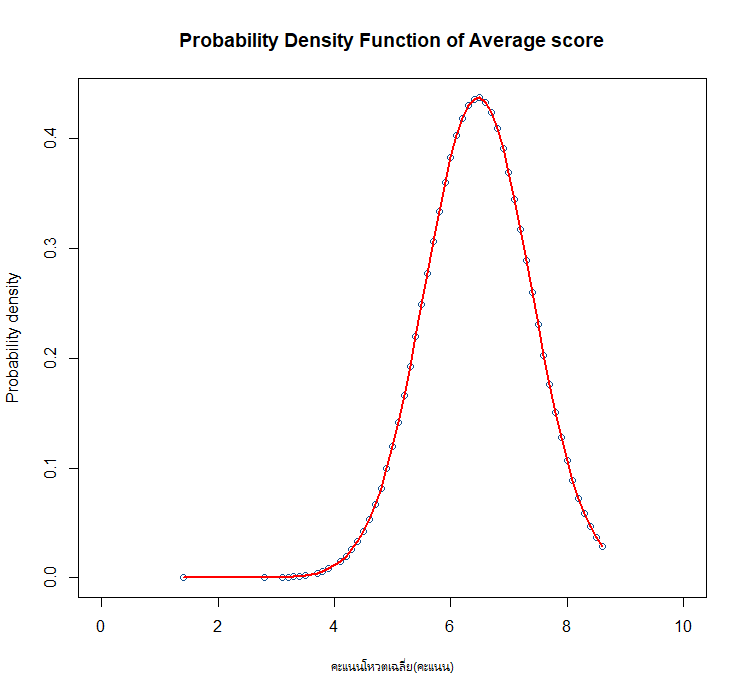
-รายได้ภาพยนตร์รวมทั่วโลก

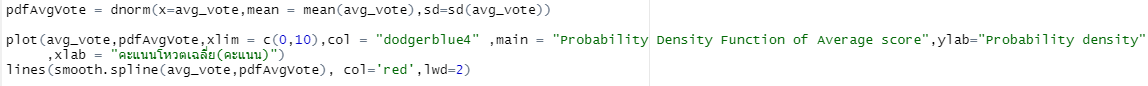




แกน x เป็นรายได้(ล้านดอลลาร์สหรัฐ) แกน y เป็น ค่าความหนาแน่นที่สอดคล้องกับ mean และ sd ของรายได้(ล้านดอลลาร์สหรัฐ)

-คะแนนโหวตเฉลี่ย

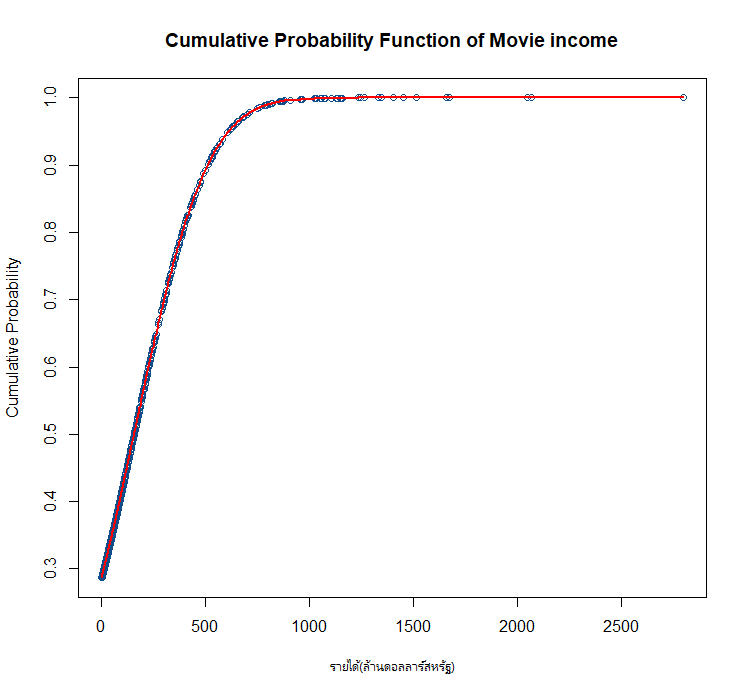


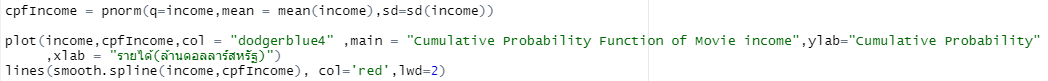


แกน x เป็นคะแนนโหวตเฉลี่ย(คะแนน) แกน y เป็น ค่าความหนาแน่นที่สอดคล้องกับ mean และ sd ของคะแนนโหวตเฉลี่ย(คะแนน)

**Cumulative Probability Function**

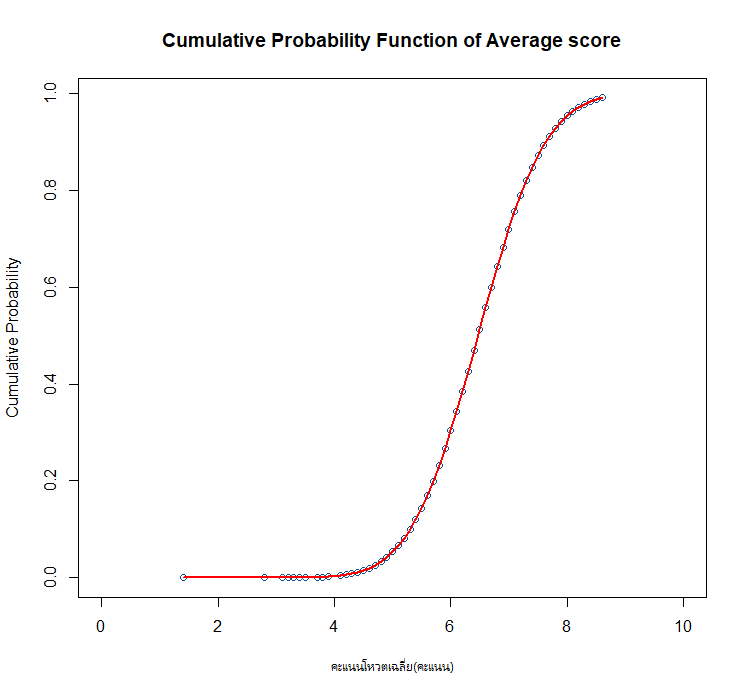
-รายได้ภาพยนตร์รวมทั่วโลก

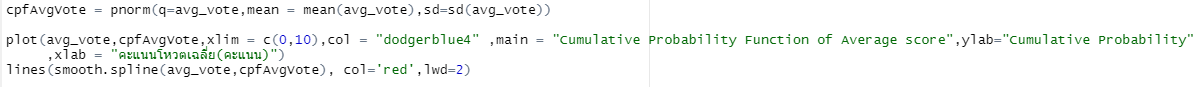




แกน x เป็นรายได้(ล้านดอลลาร์สหรัฐ) แกน y เป็น ค่าสะสมที่สอดคล้องกับ mean และ sd ของรายได้(ล้านดอลลาร์สหรัฐ)

-คะแนนโหวตเฉลี่ย





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

**บทวิเคราะห์ข้อมูลจากกราฟ**

Probability Density Function

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

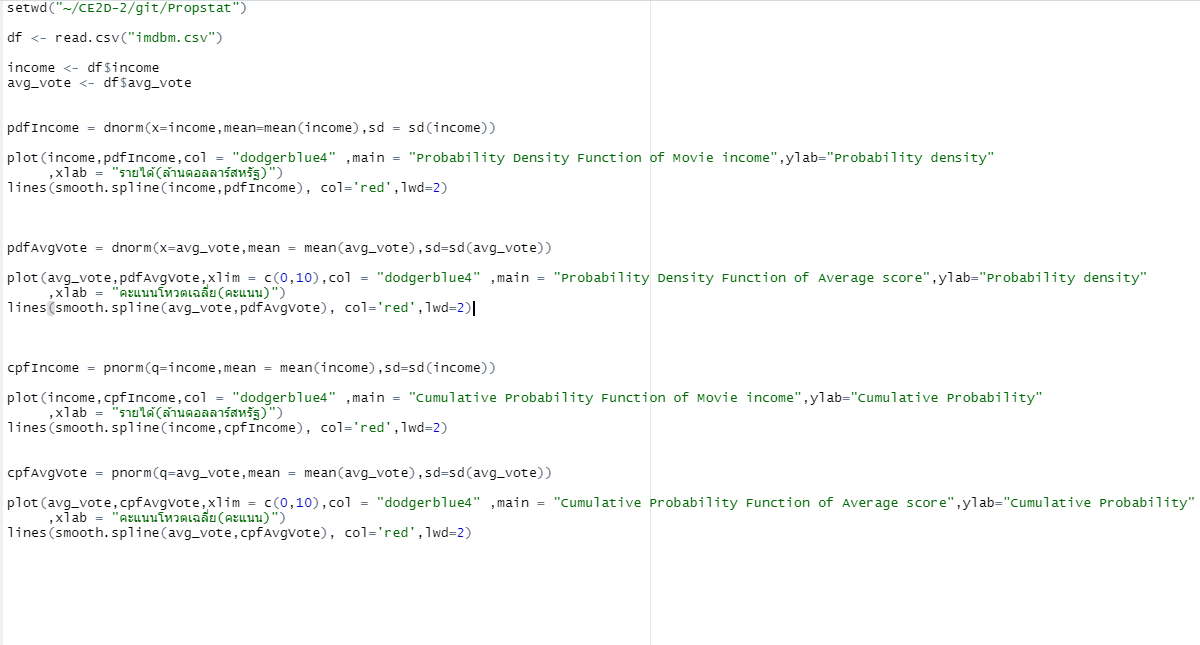
-จากกราฟคะแนนโหวตเฉลี่ย จะเห็นได้ว่าในช่วงคะแนนโหวตเฉลี่ย 0-4 คะแนนค่าความหนาแน่นเพิ่มขึ้นน้อยมาก และเพิ่มขึ้นอย่างรวดเร็วในช่วงคะแนนโหวตเฉลี่ย 4-6 คะแนน และในช่วงคะแนนโหวตเฉลี่ย 6-7 คะแนนจะมีค่าความหนาแน่นจะมากที่สุด และในช่วงคะแนนโหวตเฉลี่ย 7 คะแนนเป็นต้นไปค่าความหนาแน่นก็จะลดลงเรื่อย ๆ วิเคราะห์ได้ว่า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

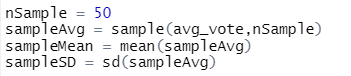
**Source Code**



**HW4# - Confidence Interval (CI) of Mean**

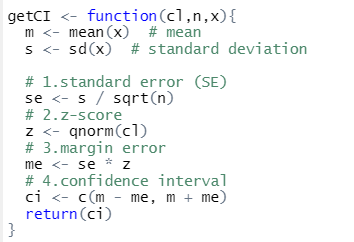
**IMDb Movies**

หา Confidence Interval (CI) ของคอลัมน์ คะแนนโหวตเฉลี่ย IMDb Movies

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

ทำการสุ่มภาพยนตร์ตัวอย่างมา 50 เรื่อง จะได้ sample mean = 6.552 คะแนน และ

sd = 0.8179716 คะแนน

-หา Confidence Interval (CI) ของแต่ละ Confidence Level

Confidence Level = 90%

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

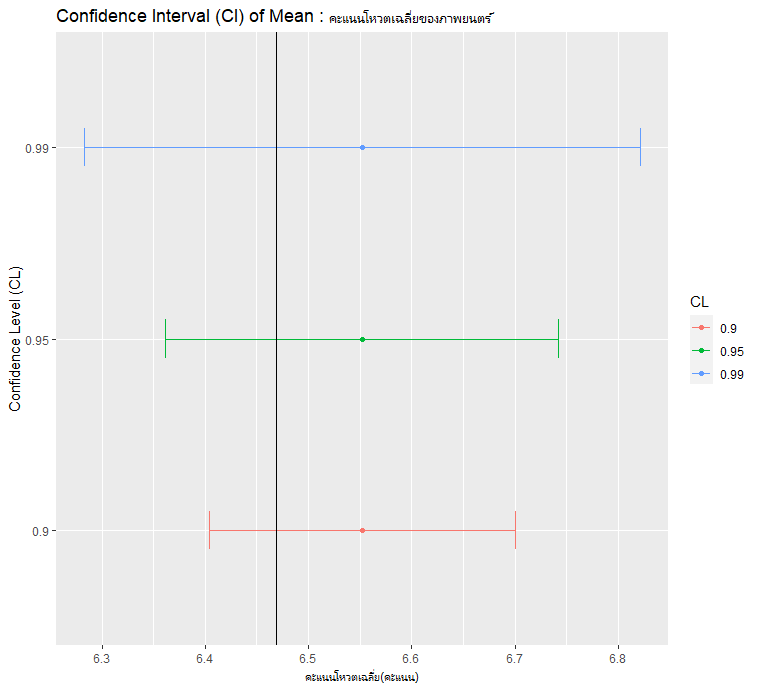
Confidence Level = 95%

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

Confidence Level = 99%

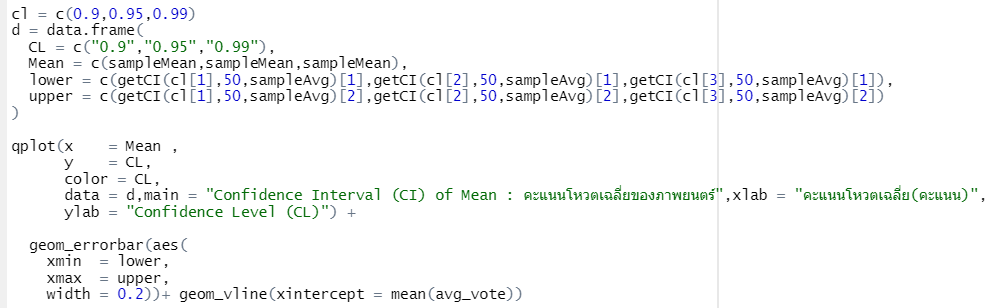
-99% confidence interval = [6.282891, 6.821109] คะแนน

**กราฟ Confidence Interval (CI) of Mean**



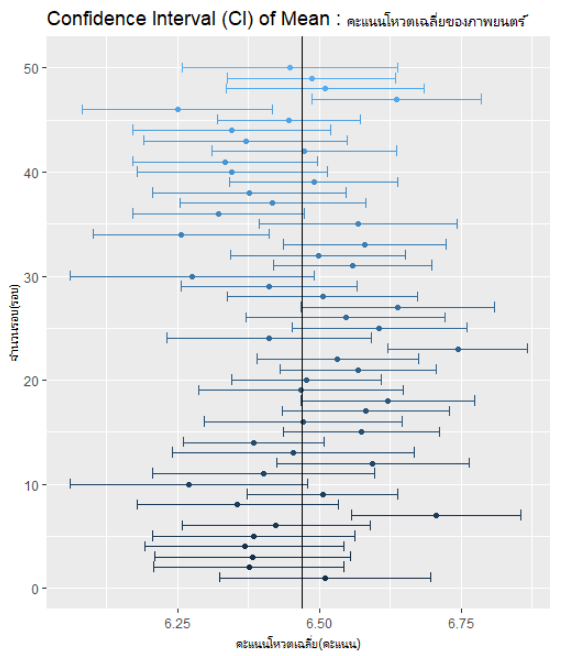
Population Mean = 6.4693 คะแนน

รูปที่ 1



**หา Confidence Interval**

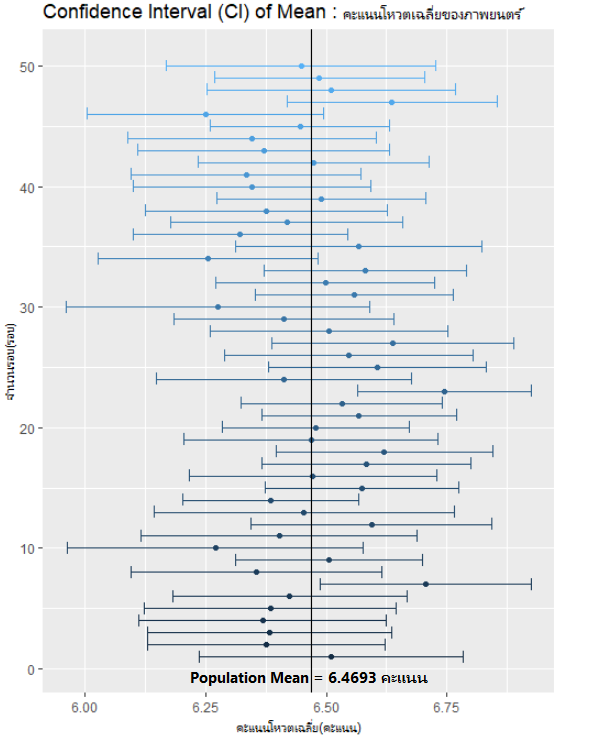
จากข้อมูลประชากร ทำการสุ่มข้อมูลมาจำนวน 50 รอบๆ ละ 50 ตัวอย่าง จะได้

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

รูปที่ 2

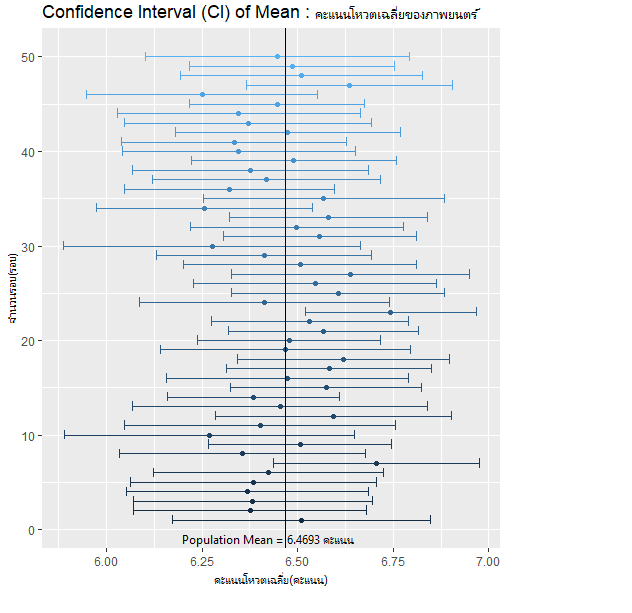
**Population Mean = 6.4693 คะแนน**

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



รูปที่ 3

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



รูปที่ 4

**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

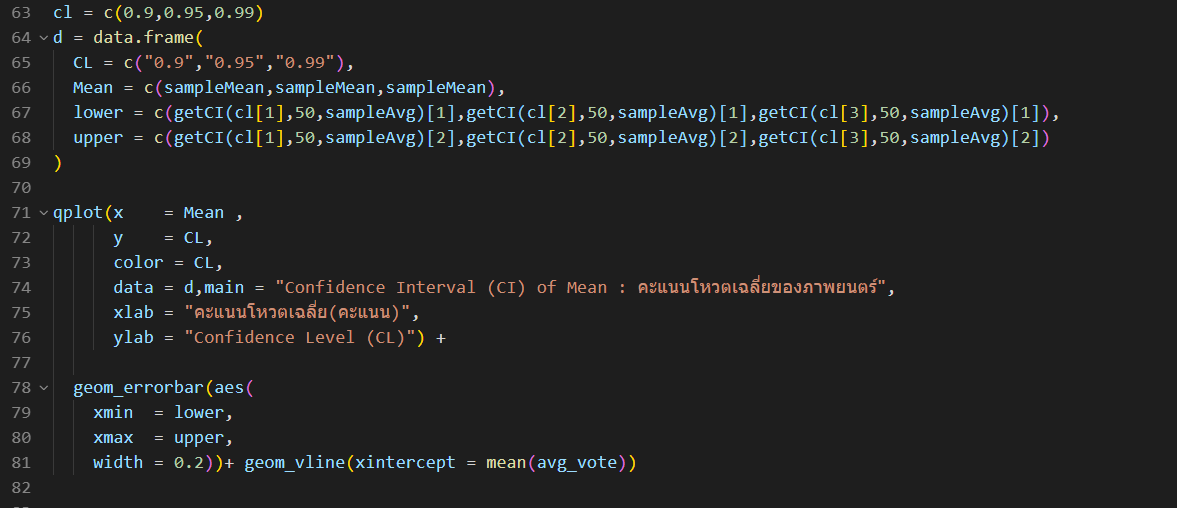
confidence interval ที่สร้างขึ้นมา หรือคิดเป็น 98 % และมี 2 % ที่ค่า population mean ไม่ได้อยู่ในช่วง Confidence Level

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

**Source Code**





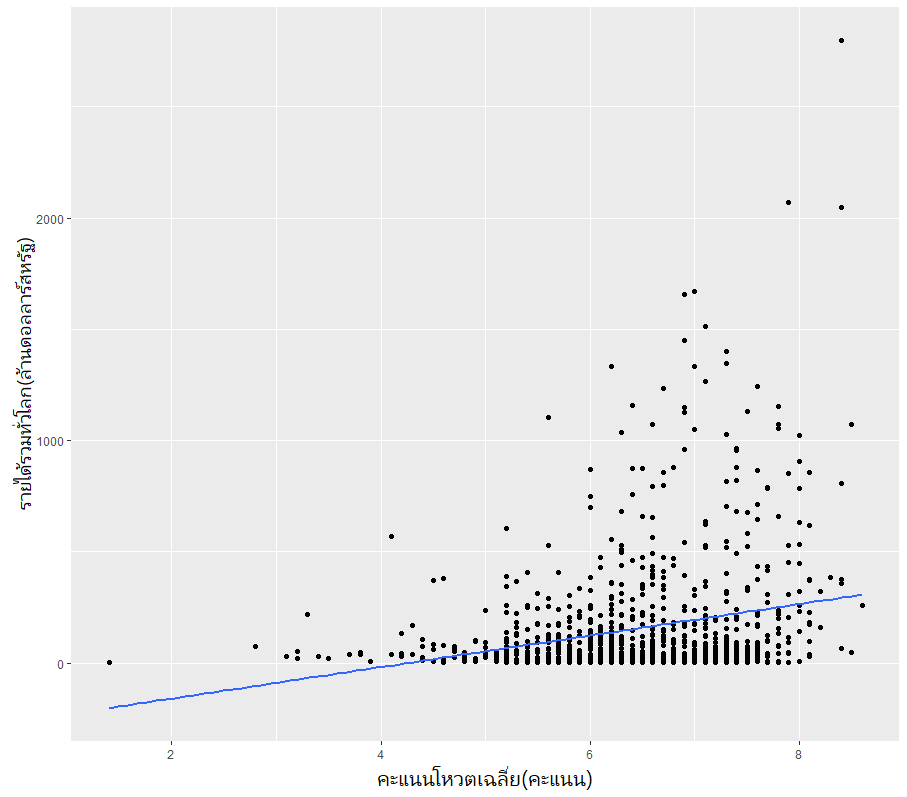


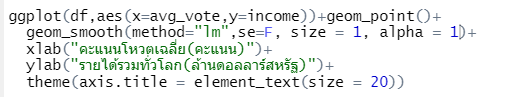
**HW5# - Linear Regression**

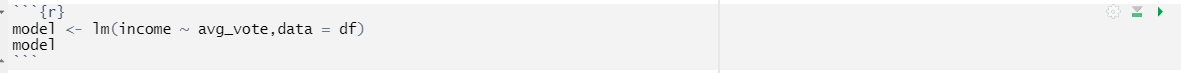
**IMDb Movies**

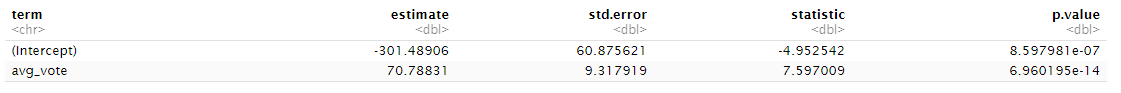
เปรียบเทียบกราฟถดถอยเชิงเส้นกับกราฟ XY(Scatter) Plot ของข้อมูล คะแนนโหวตเฉลี่ย(independent) และ รายได้รวมทั่วโลก(dependent)

**Linear Regression-(คะแนนโหวตเฉลี่ย และ รายได้รวมทั่วโลก) ของภาพยนตร์**









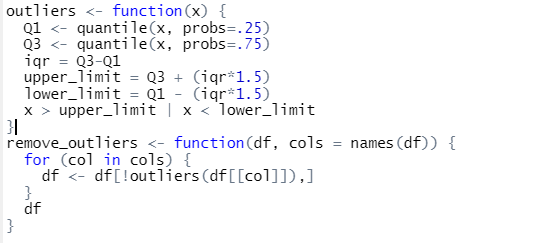
จะได้สมการ y = -301.49 + 70.79 \* x

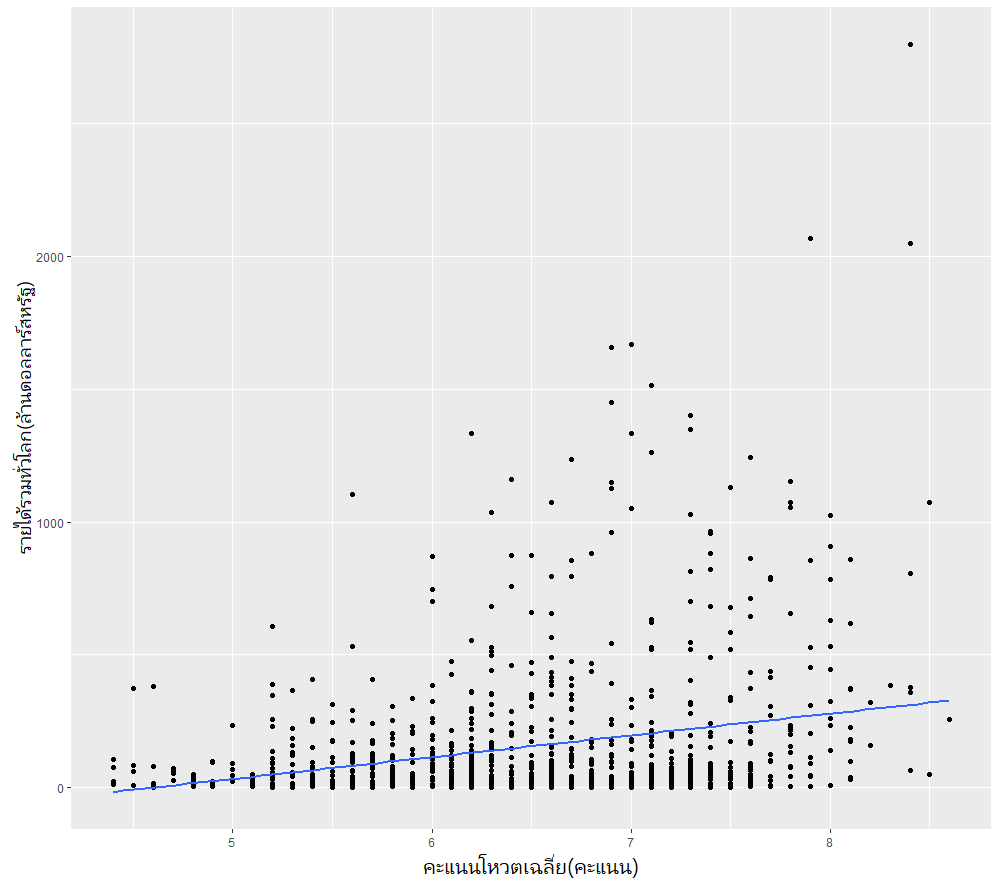
y คือ income(ล้านดอลลาร์สหรัฐ) , x คือ avg\_vote(คะแนน)

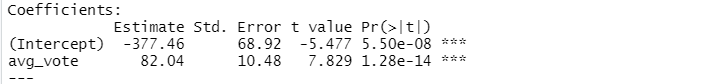
Correlation Coefficient



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

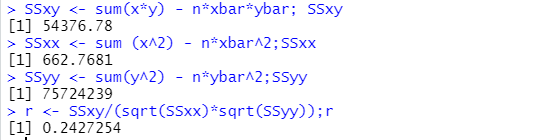
เนื่องจากค่า r มีค่าเข้าใกล้ศูนย์ซึ่งไม่ค่อยมีความสัมพันธ์ในแนวเส้นตรง ผมจึงเอา outlier ของ คะแนนโหวตเฉลี่ยออกได้ดังนี้





จะได้สมการ y = -377.46 + 82.04 \* x

y คือ income(ล้านดอลลาร์สหรัฐ) , x คือ avg\_vote(คะแนน)

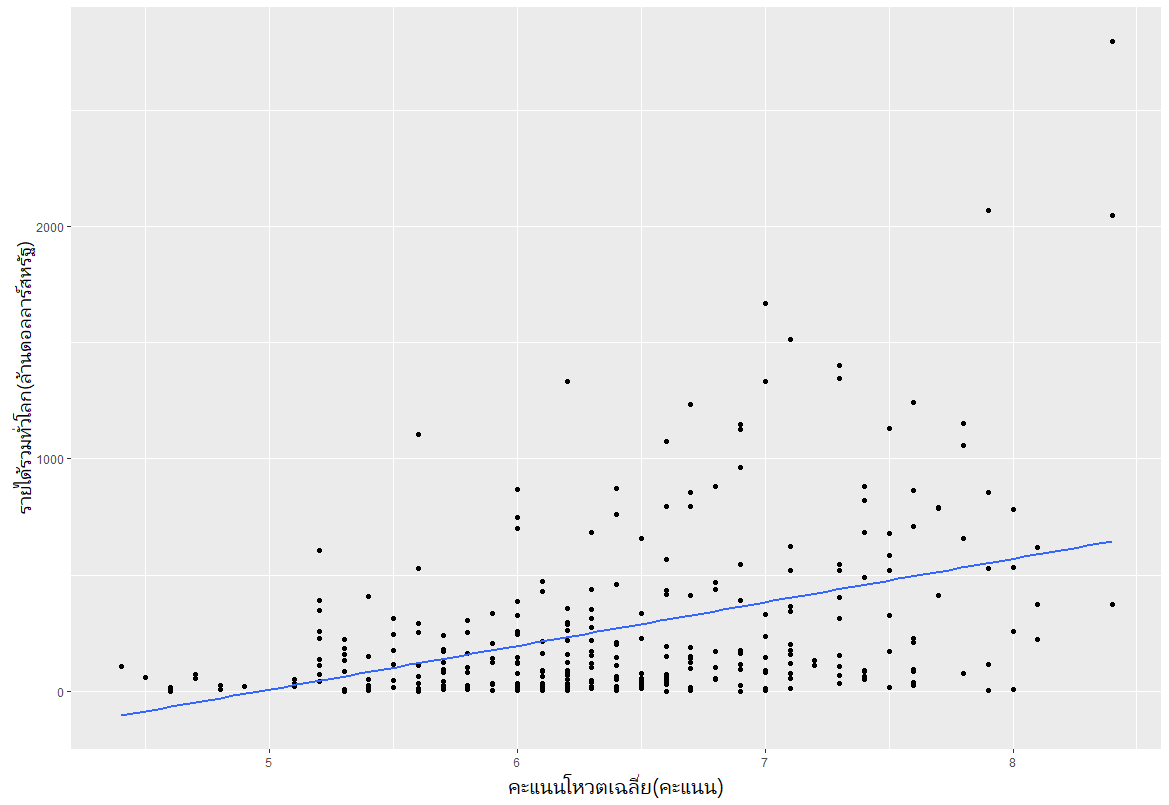


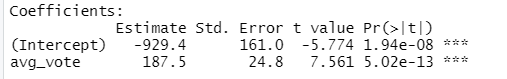
ได้ r = 0.243 ซึ่งดีกว่าเดิมนิดนึงแต่ก็ยังไม่เป็นที่น่าพอใจ ผมจึงคิดว่าควรจะหาวิธีใหม่

แบ่งตามประเภทของภาพยนตร์ได้ดังนี้



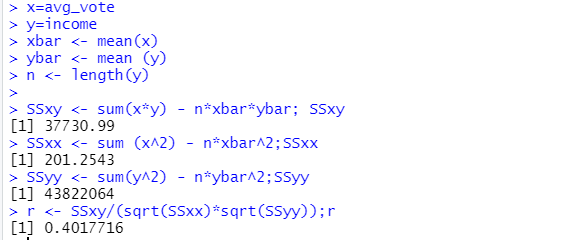
1.Action





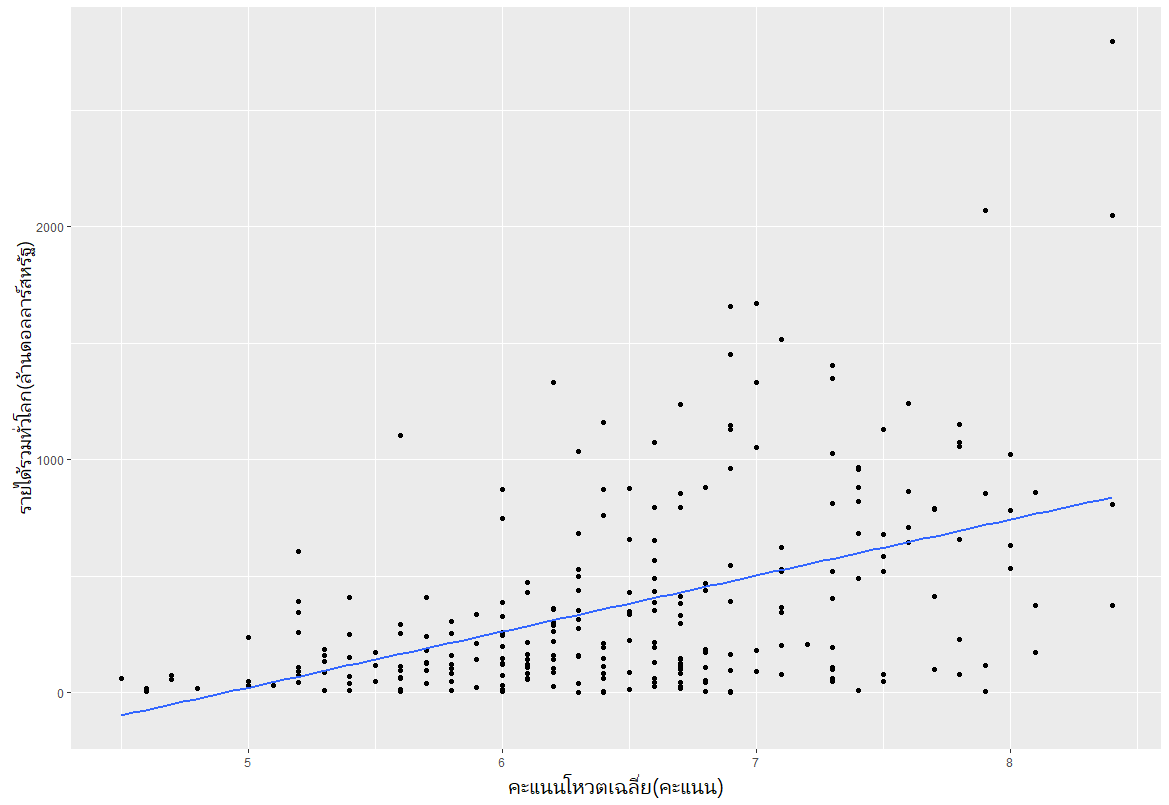
จะได้สมการ y = -929.4 + 187.5 \* x

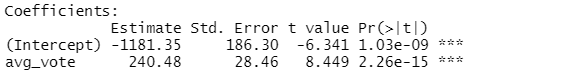
y คือ income(ล้านดอลลาร์สหรัฐ) , x คือ avg\_vote(คะแนน)



ได้ r = 0.402

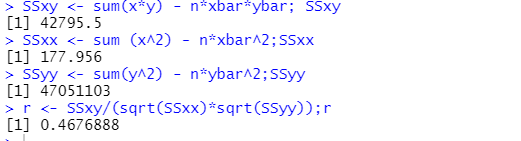
2. Adventure





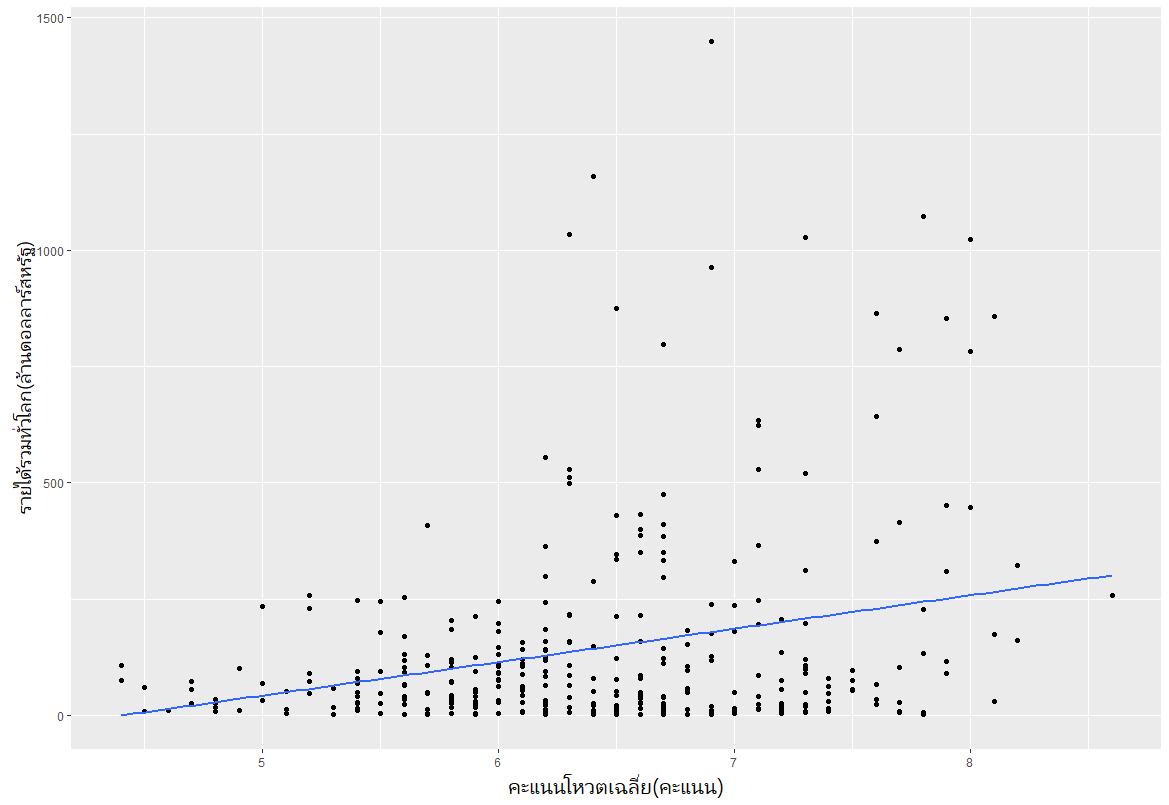
จะได้สมการ y = -1181.35 + 240.48 \* x

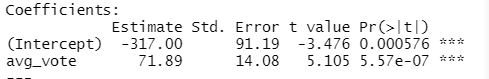
y คือ income(ล้านดอลลาร์สหรัฐ) , x คือ avg\_vote(คะแนน)



ได้ r= 0.468

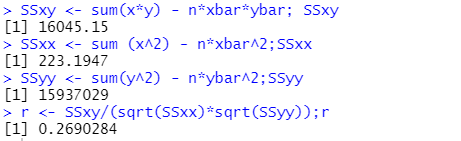
3. Comedy





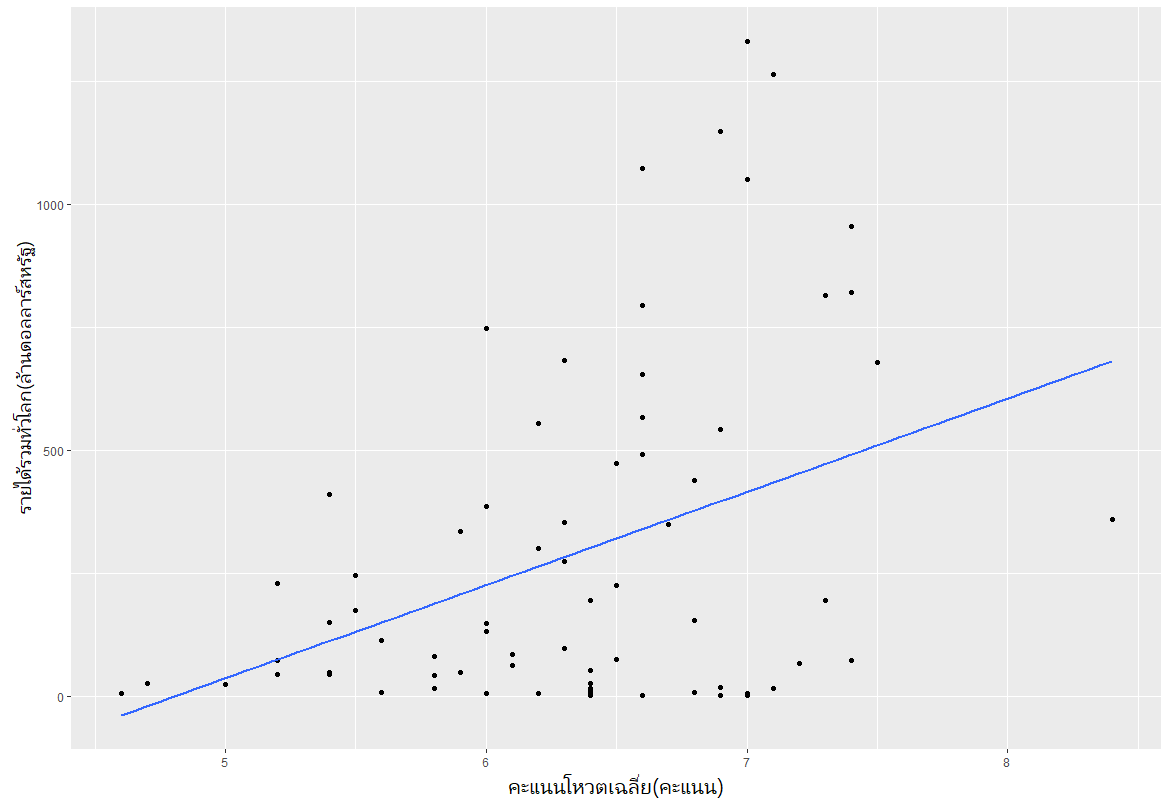
จะได้สมการ y = -1181.35 + 240.48 \* x

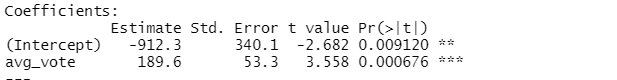
y คือ income(ล้านดอลลาร์สหรัฐ) , x คือ avg\_vote(คะแนน)



ได้ r = 0.269

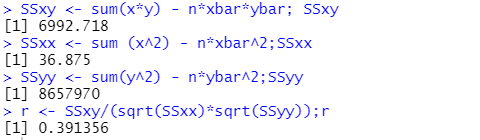
4.Fantasy





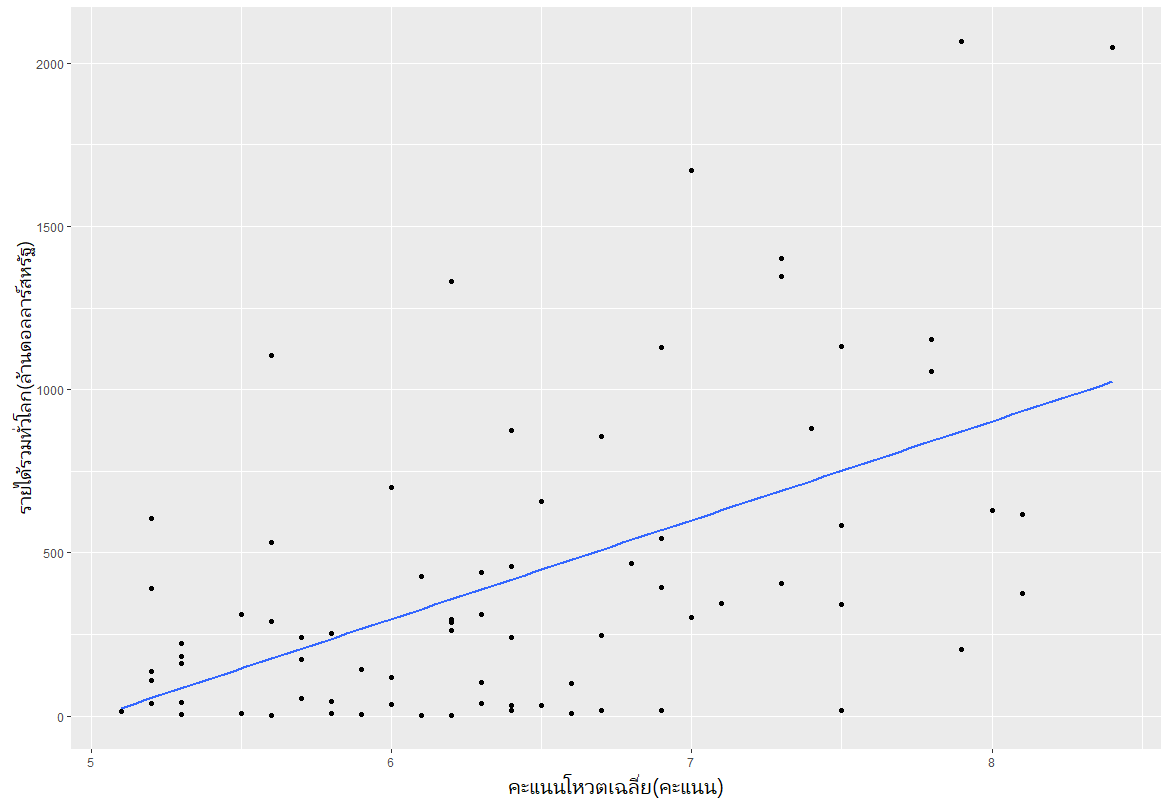
จะได้สมการ y = -912.3 + 189.6 \* x

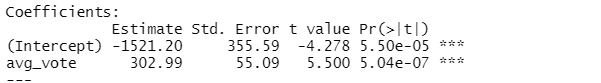
y คือ income(ล้านดอลลาร์สหรัฐ) , x คือ avg\_vote(คะแนน)



ได้ r= 0.391

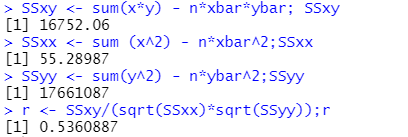
5.Sci-Fi





จะได้สมการ y = -1521.20 + 302.99 \* x

y คือ income(ล้านดอลลาร์สหรัฐ) , x คือ avg\_vote(คะแนน)



ได้ r = 0.536

จากที่ได้ลองแบ่งตามประเภทของภาพยนตร์ ค่า r ที่ได้ดีขึ้นกว่าเดิมพอสมควร

**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

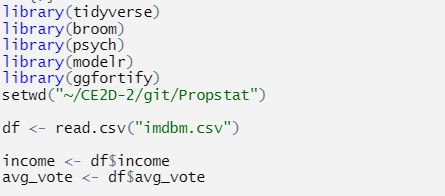
4.Fantasy ได้ r = 0.391

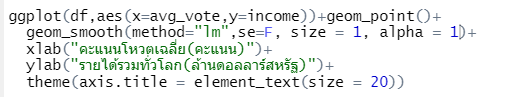
5.Sci-Fi ได้ r = 0.536

ซึ่งถือว่าค่า r ดีกว่าเดิมมาก ทำให้ linear correlation ดีขึ้นกว่าเดิม

ผมคิดว่าเวลาจะวิเคราะห์ข้อมูลต่างๆเราควรมองหลายๆปัจจัย เพื่อช่วยให้เราวิเคราะห์ข้อมูลได้ง่ายและดีขึ้น

**Source Code**











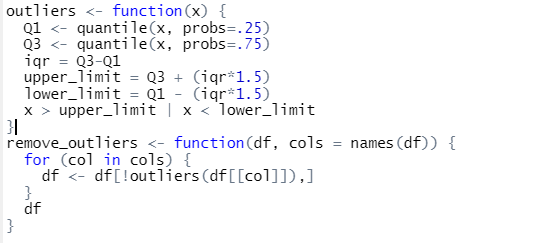














**สรุปผลการศึกษาและเสนอแนะแนวทางการศึกษาเพิ่มเติม**

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