

**นำเสนอ  
ผศ.ดร.สุรินทร์ กิตติธรกุล**

**จัดทำโดย**

**62010889 นายศุภกฤต โลห์แก้ว**

**62011019 นายอภิรักษ์ อุลิศ**

**รายงานฉบับนี้เป็นส่วนหนึ่งของรายวิชา  
01076253 PROBABILITY AND STATISTICS   
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง  
ภาคเรียนที่ 2 ปีการศึกษา 2563**

**สารบัญ**

[PROBABILITY AND STATISTICS HW 1 3](#_Toc71487274)

[**ชื่อชุดข้อมูล** Travel Review Rating Dataset 3](#_Toc71487275)

[ชื่อคอลัมน์ 3](#_Toc71487276)

[ทำไมถึงสนใจ 3](#_Toc71487277)

[คำอธิบายชื่อคอลัมน์ข้อมูล และวิธีการรวบรวมข้อมูล 3](#_Toc71487278)

[PROBABILITY AND STATISTICS HW 2 4](#_Toc71487279)

[ข้อมูลการคำนวณข้อมูลเชิงสถิติพื้นฐานแต่ละคอลัมน์ 5](#_Toc71487280)

[บทวิเคราะห์ข้อมูล 13](#_Toc71487281)

[PROBABILITY AND STATISTICS HW 3 14](#_Toc71487282)

[Average ratings on beaches 14](#_Toc71487283)

[Average ratings on parks 16](#_Toc71487284)

[Average ratings on malls 18](#_Toc71487285)

[Average ratings on beauty & spas 20](#_Toc71487286)

[Average ratings on cafes 22](#_Toc71487287)

[บทวิเคราะห์ข้อมูล 24](#_Toc71487288)

[PROBABILITY AND STATISTICS HW 4 25](#_Toc71487289)

[ค่า Confidence Interval ที่ Confidence Level 90% 26](#_Toc71487290)

[ค่า Confidence Interval ที่ Confidence Level 95% 28](#_Toc71487291)

[ค่า Confidence Interval ที่ Confidence Level 99% 30](#_Toc71487292)

[บทวิเคราะห์ข้อมูล 32](#_Toc71487293)

[PROBABILITY AND STATISTICS HW 5 33](#_Toc71487294)

[หา Linear Regression ทำคู่กับ นายศุภกฤต โลห์แก้ว 62010889 33](#_Toc71487295)

[Graph 33](#_Toc71487296)

[Coefficients 34](#_Toc71487297)

[R-Square 34](#_Toc71487298)

[บทวิเคราะห์ข้อมูล 34](#_Toc71487299)

# PROBABILITY AND STATISTICS HW 1

## **ชื่อชุดข้อมูล** Travel Review Rating Dataset

## ชื่อคอลัมน์

Numeric -> ประเภทการรีวิวแยก 5 ประเภท

Category -> หมายเลข User ผู้รีวิวสถานที่เหล่านั้น

## ทำไมถึงสนใจ

ผมเป็นคนที่รัก และชอบในการเดินทางท่องเที่ยวไปยังสถานที่ต่าง ๆ แต่บางครั้งผมก็มีปัญหาอยู่ในใจมากมาย เช่น ผมควรจะไปเยี่ยมชมสถานที่ไหนดี ? มีสถานที่ที่น่าสนใจที่ตรงกับไลฟ์สไตล์ของผม บ่อยครั้งที่ผมใช้เวลาหลายชั่วโมงเพื่อค้นหาสถานที่ที่น่าสนใจที่จะออกไปข้างนอก

จะเกิดอะไรขึ้น ถ้าเราสามารถสร้างระบบแนะนํา หรือการรีวิว ซึ่งสามารถแนะนําสถานที่ที่น่าสนใจหลายแห่งตามความต้องการของแต่ละคนได้ ด้วยข้อมูลจากการตรวจสอบของ Google ซึ่งผมจะพยายามแบ่งผู้ใช้รีวิว Google ออกเป็นกลุ่มที่สนใจคล้ายกัน ดังตัวอย่างข้อมูลที่นำมานี้ เพื่อที่จะได้เป็นทางเลือกในการตัดสินใจต่อใครหลาย ๆ คนที่ประสบพบเจอปัญหาแบบเดียวกับผม

**แหล่งที่มาของชุดข้อมูล** [Travel Review Rating Dataset | Kaggle](https://www.kaggle.com/wirachleelakiatiwong/travel-review-rating-dataset)

## คำอธิบายชื่อคอลัมน์ข้อมูล และวิธีการรวบรวมข้อมูล

**คำอธิบายชื่อคอลัมน์ข้อมูล**

Numeric (เลือกมาจาก 5 ประเภท ใน 24 ประเภทของข้อมูลจริง)

Attribute 1 : Unique user id Attribute 2 : Average ratings on beaches

Attribute 3 : Average ratings on parks

Attribute 4 : Average ratings on cafes

Attribute 5 : Average ratings on malls

Attribute 6 : Average ratings on beauty & spas

Category

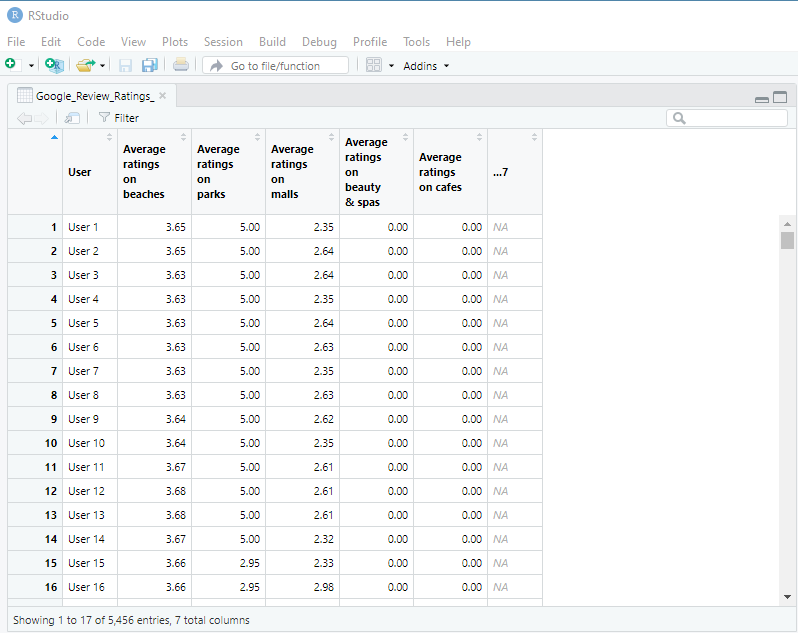
จำนวนผู้ใช้งานที่เข้ามาให้การแนะนำ หรือรีวิวในประเภทต่าง ๆ

**วิธีการรวบรวมข้อมูล**

ชุดข้อมูลนี้มาจากที่เก็บแมชชีนเลิร์นนิงของมหาวิทยาลัยแคลิฟอร์เนีย, เออร์ไวน์ (UC Irvine) : ข้อมูลการจัดอันดับรีวิวการเดินทาง ชุดข้อมูลนี้จะถูกเติมโดยการจับคะแนนของผู้ใช้จากรีวิวของ Google รีวิวเกี่ยวกับสถานที่ท่องเที่ยวจาก 24 หมวดหมู่ทั่วยุโรปได้รับการพิจารณา คะแนนผู้ใช้ Google มีตั้งแต่ 1 ถึง 5 และมีการคํานวณคะแนนผู้ใช้เฉลี่ยต่อหมวดหมู่

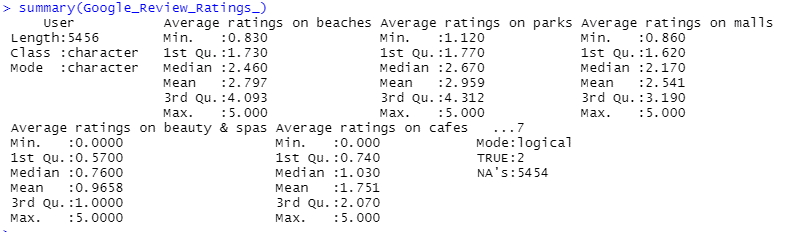
# PROBABILITY AND STATISTICS HW 2

สืบเนื่องมาจาก HW 1 ที่ทางผู้จัดทำได้หยิบยกชุดข้อมูลที่เกี่ยวกับ Travel Review Rating Dataset ( From the Machine Learning Repository of University of California ) โดยประกอบด้วยคอลัมน์ Numeric (ประเภทการรีวิวแยก 5 ประเภท), และ Category (หมายเลข User ผู้รีวิวสถานที่เหล่านั้น) และเพื่อที่จะได้คําานวณหาค่าสถิติพื้นฐาน และวิเคราะห์ข้อมูลจากกราฟได้ง่ายดายมากขึ้น ทางผู้จัดทำจึงได้ใช้โปรแกรม RStudio ในการวิเคราะห์ข้อมูลด้วย ภาษา R โดยมีรายละเอียดดังต่อไปนี้

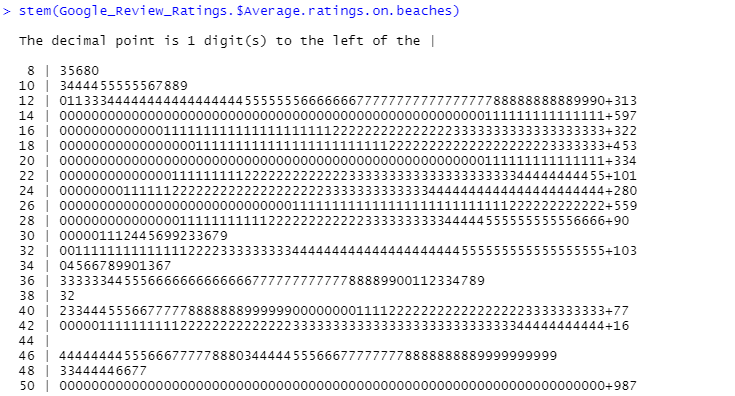


*ภาพแสดงตารางข้อมูลเมื่อ import file .csv เข้ามาในโปรแกรม RStudio*

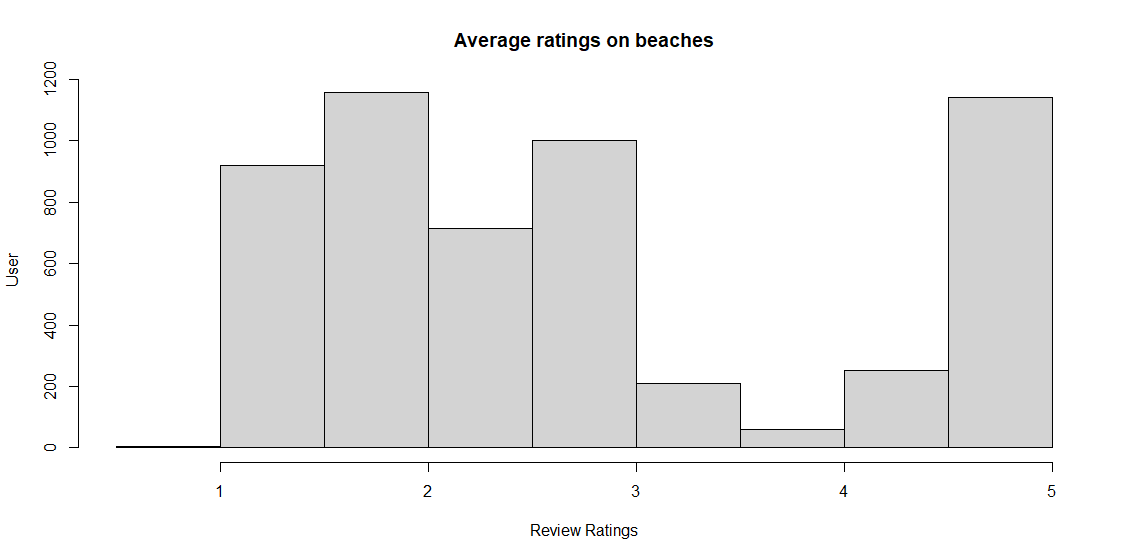
## ข้อมูลการคำนวณข้อมูลเชิงสถิติพื้นฐานแต่ละคอลัมน์



*ข้อมูลการคำนวณข้อมูลเชิงสถิติของคอลัมน์ Average ratings on beaches, parks, malls, beauty & spas, cafes*

**

ข้อมูลแสดงกราฟ Stem and leaf เลือกมา 1 คอลัมน์ คือคอลัมน์ Average ratings on beaches (ข้อมูลแสดงจำนวนคะแนนเฉลี่ยของการรีวิว จากคะแนนเต็ม 5 คะแนน ของ User ผู้รีวิวสถานที่เหล่านั้นจำนวน 5,456 คน)



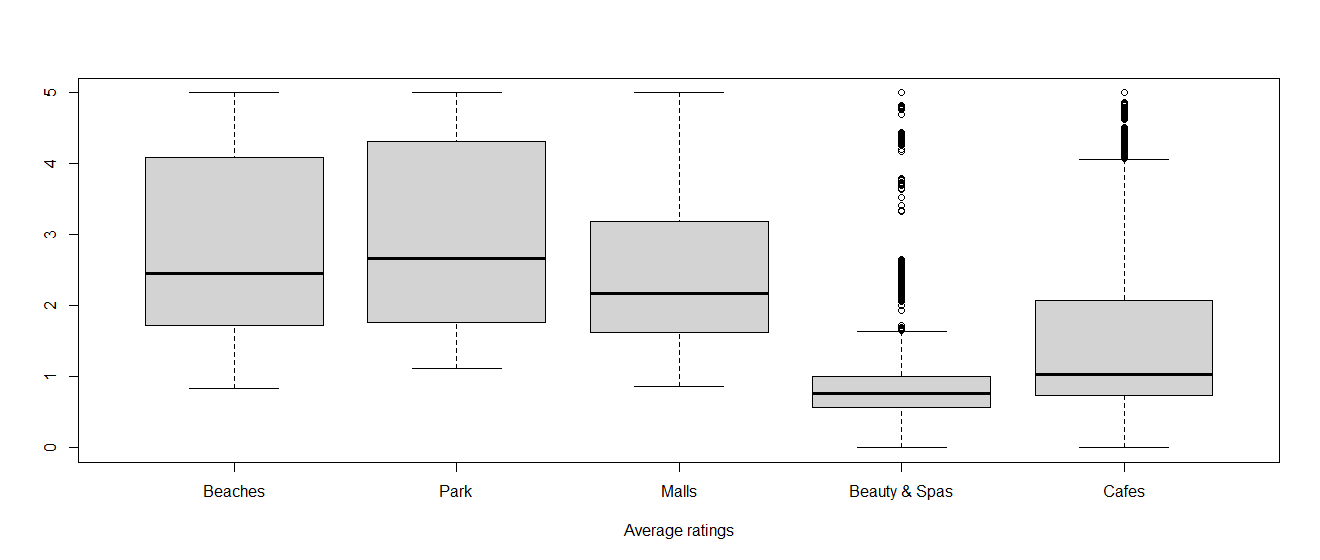
(ID)

(Stars)

*Histogram เลือกมา 1 คอลัมน์ คือคอลัมน์ Average ratings on beaches (ข้อมูลแสดงจำนวนคะแนนเฉลี่ยของการรีวิว จากคะแนนเต็ม 5 คะแนน* ของ User ผู้รีวิวสถานที่เหล่านั้น*จำนวน 5,456 คน)*

*Code สำหรับสร้าง Histogram*

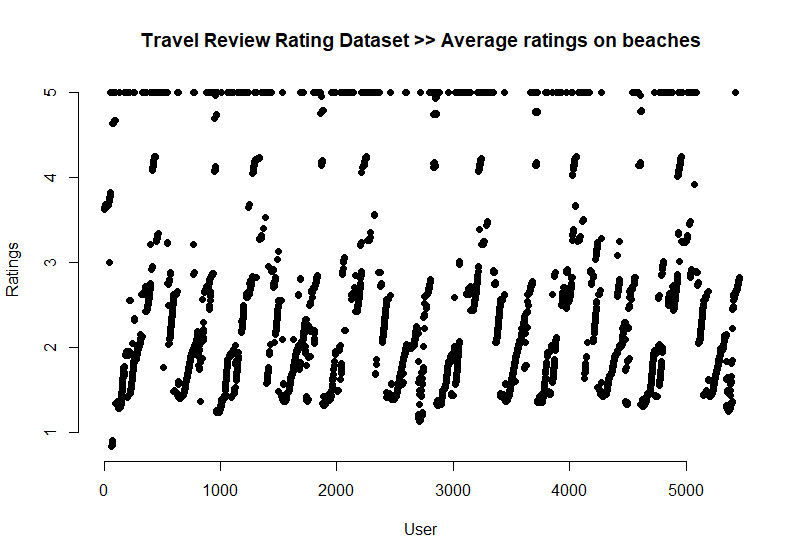
> hist(Google\_Review\_Ratings\_$'Average ratings on beaches', main = "Average ratings on beaches", xlab = "Review Ratings", ylab = "User" )



*Boxplot แสดงแต่ละประเภทของการรีวิวของ User ผู้รีวิวสถานที่เหล่านั้นจำนวน 5,456 คน*

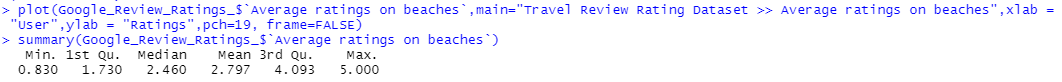
*Code สำหรับสร้าง Boxplot*

> boxplot(Google\_Review\_Ratings\_$'Average ratings on beaches', Google\_Review\_Ratings\_$'Average ratings on parks', Google\_Review\_Ratings\_$'Average ratings on malls', Google\_Review\_Ratings\_$'Average ratings on beauty & spas', Google\_Review\_Ratings\_$'Average ratings on cafes', names = c('Beaches','Park','Malls','Beauty & Spas','Cafes'), xlab="Average ratings" )



*XY Scatter plot (Average ratings on beaches, User ผู้รีวิวสถานที่เหล่านั้นจำนวน 5,456 คน)*

*Code สำหรับการ Plot XY Scatter Graph*

**

ข้อมูลสถิติเบื้องต้นของ Average ratings on beaches จาก User ทั้งหมด พบว่า

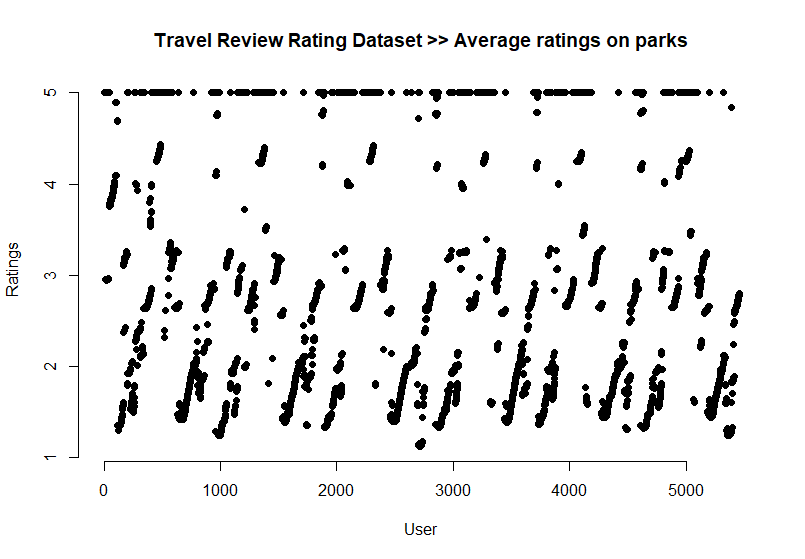
- มี Min Ratings คือ 0.830

- มีค่า Mean คือ 2.797

- มี Max Ratings คือ 5.000

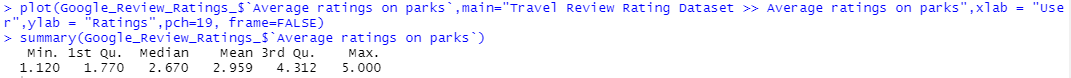
- มีค่าที่ตกใน Quartile ที่ 1 คือ 1.730

- มีค่าที่ตกใน Quartile ที่ 3 คือ 4.093



*XY Scatter plot (Average ratings on parks, User ผู้รีวิวสถานที่เหล่านั้นจำนวน 5,456 คน)*

*Code สำหรับการ Plot XY Scatter Graph*

**

ข้อมูลสถิติเบื้องต้นของ Average ratings on parks จาก User ทั้งหมด พบว่า

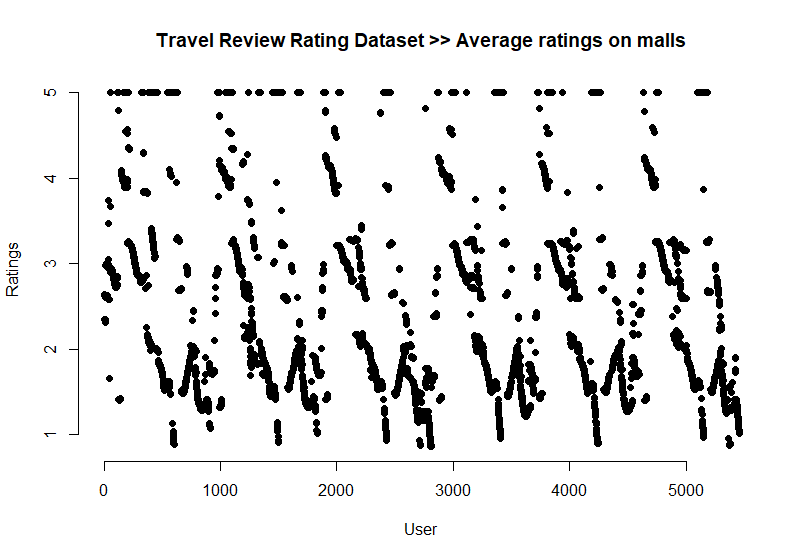
- มี Min Ratings คือ 1.120

- มีค่า Mean คือ 2.959

- มี Max Ratings คือ 5.000

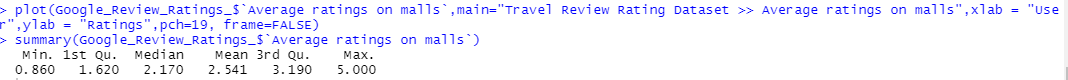
- มีค่าที่ตกใน Quartile ที่ 1 คือ 1.770

- มีค่าที่ตกใน Quartile ที่ 3 คือ 4.312



*XY Scatter plot (Average ratings on malls, User ผู้รีวิวสถานที่เหล่านั้นจำนวน 5,456 คน)*

*Code สำหรับการ Plot XY Scatter Graph*

**

ข้อมูลสถิติเบื้องต้นของ Average ratings on malls จาก User ทั้งหมด พบว่า

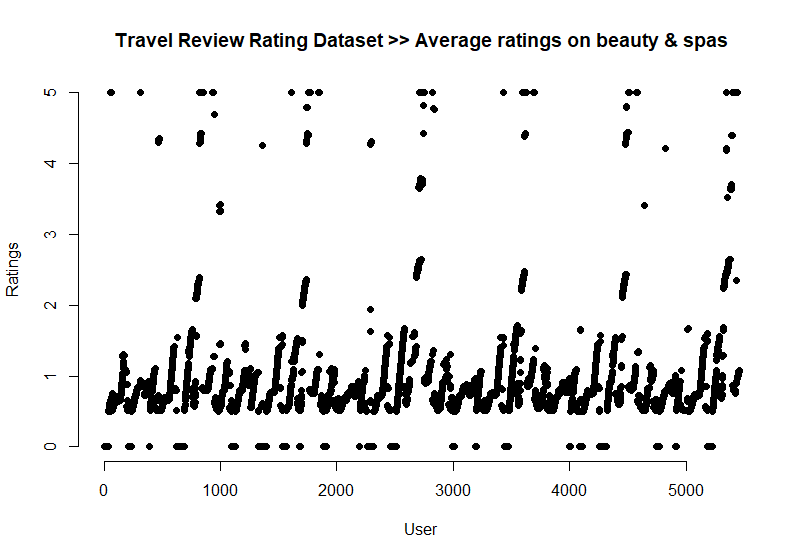
- มี Min Ratings คือ 0.860

- มีค่า Mean คือ 2.541

- มี Max Ratings คือ 5.000

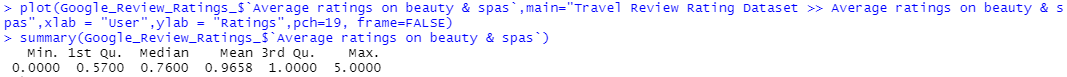
- มีค่าที่ตกใน Quartile ที่ 1 คือ 1.620

- มีค่าที่ตกใน Quartile ที่ 3 คือ 3.190



*XY Scatter plot (Average ratings on beauty & spas, User ผู้รีวิวสถานที่เหล่านั้นจำนวน 5,456 คน)*

*Code สำหรับการ Plot XY Scatter Graph*

**

ข้อมูลสถิติเบื้องต้นของ Average ratings on beauty & spas จาก User ทั้งหมด พบว่า

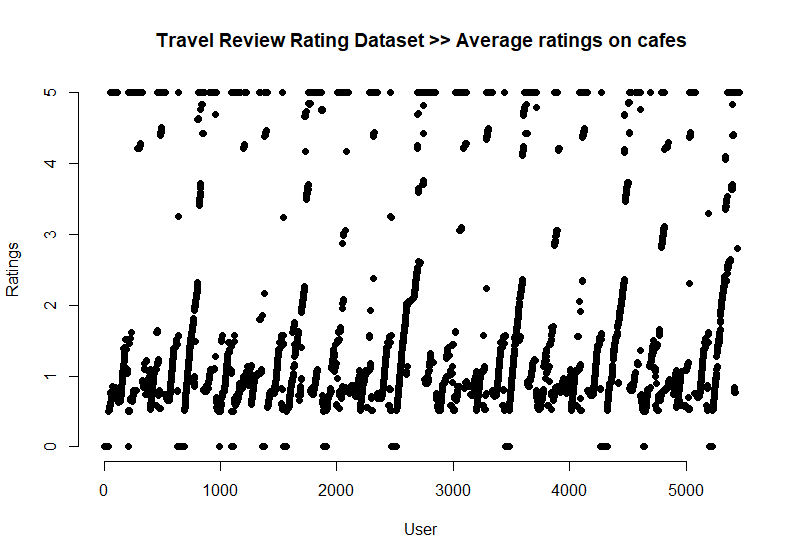
- มี Min Ratings คือ 0.000

- มีค่า Mean คือ 0.9658

- มี Max Ratings คือ 5.000

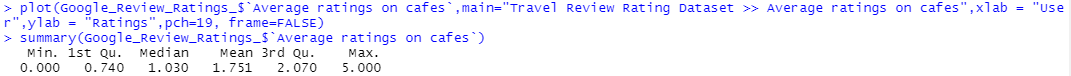
- มีค่าที่ตกใน Quartile ที่ 1 คือ 0.570

- มีค่าที่ตกใน Quartile ที่ 3 คือ 1.000



*XY Scatter plot (Average ratings on cafes, User ผู้รีวิวสถานที่เหล่านั้นจำนวน 5,456 คน)*

*Code สำหรับการ Plot XY Scatter Graph*

**

ข้อมูลสถิติเบื้องต้นของ Average ratings on cafes จาก User ทั้งหมด พบว่า

- มี Min Ratings คือ 0.000

- มีค่า Mean คือ 1.751

- มี Max Ratings คือ 5.000

- มีค่าที่ตกใน Quartile ที่ 1 คือ 0.740

- มีค่าที่ตกใน Quartile ที่ 3 คือ 2.070

## บทวิเคราะห์ข้อมูล

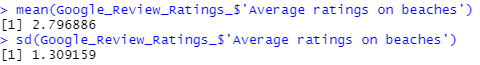
ชุดข้อมูลนี้มาจากที่เก็บแมชชีนเลิร์นนิงของมหาวิทยาลัยแคลิฟอร์เนีย, เออร์ไวน์ (UC Irvine) : ข้อมูลการจัดอันดับรีวิวการเดินทาง ชุดข้อมูลนี้จะถูกเติมโดยการจับคะแนนของผู้ใช้จากรีวิวของ Google รีวิวเกี่ยวกับสถานที่ท่องเที่ยวจาก 24 หมวดหมู่ทั่วยุโรปได้รับการพิจารณา ทางผู้จัดทำได้หยิบยกคะแนนผู้ใช้ Google มีตั้งแต่ 1 ถึง 5 และมีการคํานวณคะแนนผู้ใช้เฉลี่ยต่อหมวดหมู่มาคิดวิเคราะห์

จากการวิเคราะห์กราฟ *XY Scatter plot ทางผู้จัดทำจะแบ่งเป็น 5 ส่วนตามแต่ละประเภทของการรีวิว เพราะเนื่องจากเป็นข้อมูลที่ไม่มีความซับซ้อนมาก และนำจำนวน User ผู้รีวิวสถานที่เหล่านั้นจำนวน 5,456 คน* มาเทียบกับข้อมูลสถิติเบื้องต้นของ Average ratings ทั้ง 5 ประเภท พบว่า User ส่วนมากที่เข้ามารีวิวนั้น มีการให้คะแนนเป็นไปในทางที่ดีเกี่ยวกับประเภท parks เห็นได้จากคะแนนเฉลี่ยนั้นมีค่ามากที่สุดคือ 2.959 ส่วนรองลงมาจะเป็นประเภท beaches ที่มีคะแนน 2.797 คะแนน และประเภท Ratings on malls ที่มีคะแนน 2.541 คะแนน แต่กลับกันจากกราฟยังพบอีกว่าค่าคะแนน Average ratings ของ beauty & spas กับ cafes ที่มีคะแนนการรีวิวเฉลี่ยต่ำมากคือ 0.965 คะแนน และ 1.751 คะแนน ตามลำดับ เนื่องจากความเป็นจริงสถานที่ที่ผู้คนส่วนใหญ่มักจะไปกันจะเป็นที่ที่เกี่ยวเนื่องกับธรรมชาติ อาจเป็นเพราะความลี้ลับและความสวยงามของสถานที่เที่ยวทางธรรมชาติ ที่มีความน่าดึงดูดผู้คนมากกว่าสถานที่ที่ถูกสร้างขึ้นมาเองโดยมนุษย์ ซึ่งอาจมีข้อดีได้หลายปัจจัย อาทิ ได้รู้จักการวางแผน ได้ฝึกไหวพริบและรู้จักการแก้ปัญหาเฉพาะหน้า ได้ใกล้ชิดกับธรรมชาติและเข้าใจวิธีชีวิต ส่วนสถานที่เที่ยวรองลงมาคงหนีไม่พ้นกับ พวกประเภทห้างสรรพสินค้าต่าง ๆ เนื่องจากผู้คนต่างก็มีความอยากได้สิ่งของที่จำเป็น อาทิ ยารักษาโรค และของอุปโภคบริโภค รวมถึงสิ้นค้าฟุ่มเฟือยต่าง ๆ หรืออาจเป็นการผ่อนคลายทางด้านร่างกายจิตใจ และอารมภ์ เช่น การเข้าฟิตเนส การดูภาพยนต์ การเล่นดนตรี เป็นต้น และประเภทสุดท้ายจากกราฟคือ สถานที่ที่พักผ่อนหย่อนใจ อาทิ ร้านนวดแผนไทย ร้านสปา และคาเฟ่ต่าง ๆ เพราะมนุษย์นั้นมีความความรักสวย รักงามและมีเส้นทางที่จะเดินในแบบของตัวเอง

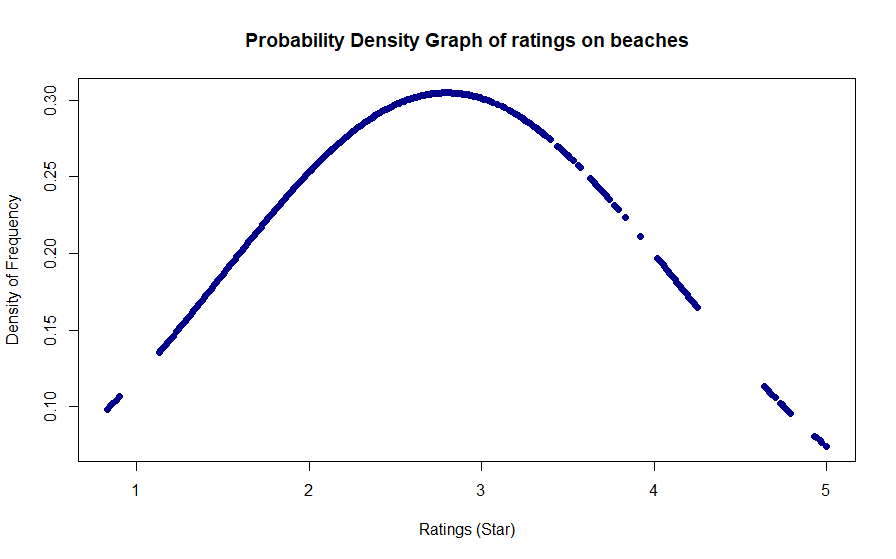
# PROBABILITY AND STATISTICS HW 3

สืบเนื่องมาจาก HW 2 ที่ทางผู้จัดทำได้หยิบยกชุดข้อมูลที่เกี่ยวกับ Travel Review Rating Dataset ( From the Machine Learning Repository of University of California ) ซึ่งมี ข้อมูลการคำนวณข้อมูลเชิงสถิติพื้นฐานแต่ละคอลัมน์ ข้อมูลแสดงกราฟ Stem and leaf แผนภาพ Histogram Boxplot และ XY Scatter plot ของแต่ละคอลัมน์ ซึ่งทางผู้จัดทำได้เห็นข้อมูลที่เกิดจากการเรียงตัวของข้อมูลมากขึ้น ทำให้สามารถหาค่าสถิติพื้นฐานได้มากขึ้น ไม่ว่าจะเป็น ค่า Min Mean Max และ Quartile ของกราฟ จากข้อมูลเหล่านี้สามารถนำมาวิเคราะห์ เพื่อหาความสนใจและความต้องการของผู้ที่มารีวิวสถานที่ของแต่ละประเภทได้ง่ายขึ้น และโดย HW 3 นี้ จะแสดงถึง Probability Density Graph และ Cumulative Distribution Graph ของแต่ละสถานที่เพื่อได้เห็นการกระจายตัวของข้อมูลให้ชัดเจนมากขึ้น

## Average ratings on beaches

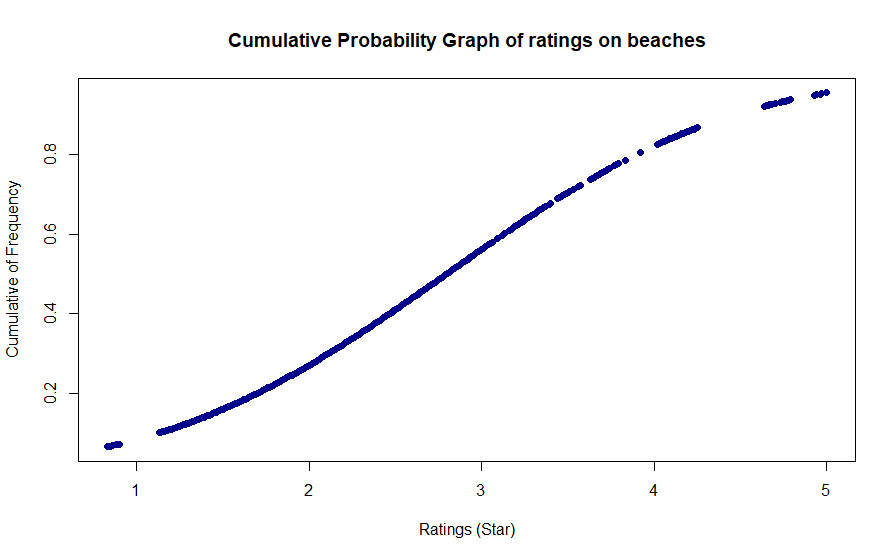


****



Probability Density Graph of ratings on beaches

****



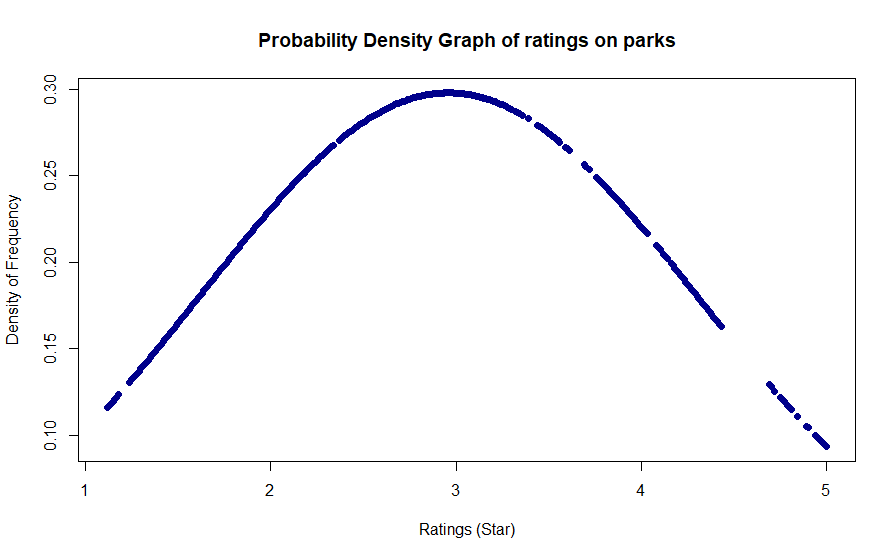
Cumulative Distribution Graph of ratings on beaches

1. หาค่า mean และ sd ของคอลัมน์ที่ต้องการจะทำการหา Probability Density
2. หลังจากนั้นใช้ตัวแปร “DENS” รับค่าคำสั่ง dnorm ซึ่งเป็นคำสั่งในการหา Probability Density โดยต้องใส่ค่า ข้อมูลของคอลัมน์ mean และ sd และสุดท้ายทำการ plot() โดยค่าในแกน x เป็นคะแนนเต็ม 5 ดาว แกน y เป็นค่าการสะสมที่สอดคล้องกับค่า mean และ sd
3. จะได้กราฟของ Probability Density Graph of ratings on beaches มีหน่วยเป็น Star
4. ใช้ตัวแปร “CUM” รับค่าคำสั่ง pnorm ซึ่งเป็นคำสั่งในการหา Cumulative Distribution โดยต้องใส่ค่า ข้อมูลของคอลัมน์, mean และ sd และสุดท้ายทำการ plot() โดยค่าในแกน x เป็นคะแนนเต็ม 5 ดาว แกน y เป็นค่าการสะสมที่สอดคล้องกับค่า mean และ sd
5. จะได้กราฟของ Cumulative Distribution Graph of ratings on beaches มีหน่วยเป็น Star

## Average ratings on parks

****

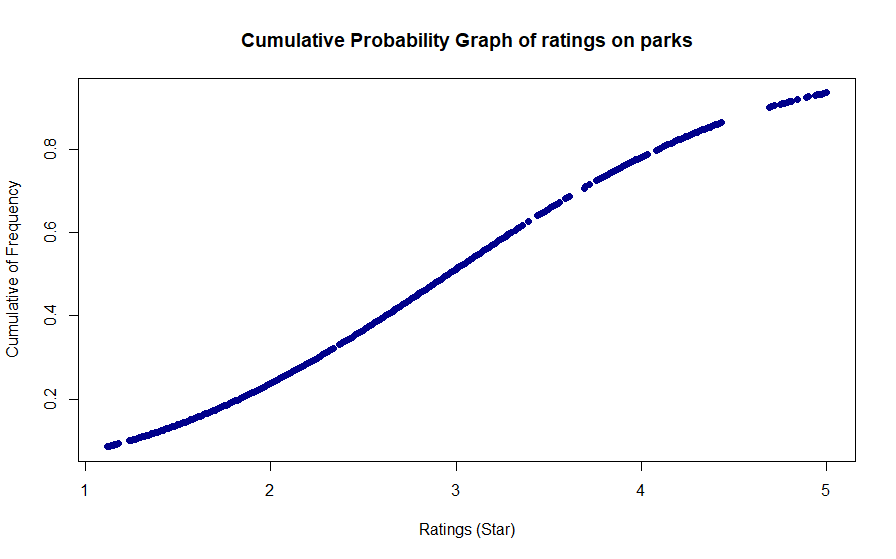
****



Probability Density Graph of ratings on parks

1. หาค่า mean และ sd ของคอลัมน์ที่ต้องการจะทำการหา Probability Density
2. หลังจากนั้นใช้ตัวแปร “DENS” รับค่าคำสั่ง dnorm ซึ่งเป็นคำสั่งในการหา Probability Density โดยต้องใส่ค่า ข้อมูลของคอลัมน์ mean และ sd และสุดท้ายทำการ plot() โดยค่าในแกน x เป็นคะแนนเต็ม 5 ดาว แกน y เป็นค่าการสะสมที่สอดคล้องกับค่า mean และ sd
3. จะได้กราฟของ Probability Density Graph of ratings on parks มีหน่วยเป็น Star

****



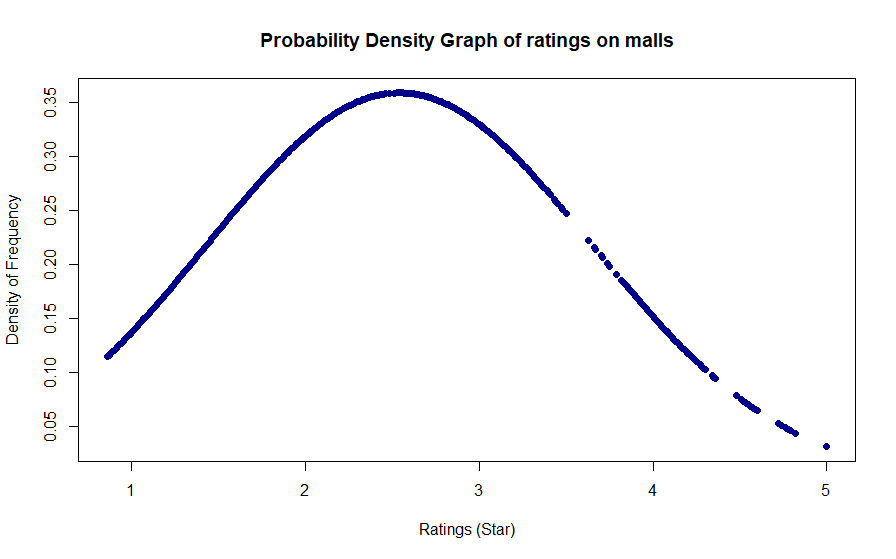
Cumulative Distribution Graph of ratings on parks

1. ใช้ตัวแปร “CUM” รับค่าคำสั่ง pnorm ซึ่งเป็นคำสั่งในการหา Cumulative Distribution โดยต้องใส่ค่า ข้อมูลของคอลัมน์, mean และ sd และสุดท้ายทำการ plot() โดยค่าในแกน x เป็นคะแนนเต็ม 5 ดาว แกน y เป็นค่าการสะสมที่สอดคล้องกับค่า mean และ sd
2. จะได้กราฟของ Cumulative Distribution Graph of ratings on beaches มีหน่วยเป็น Star

## Average ratings on malls

****

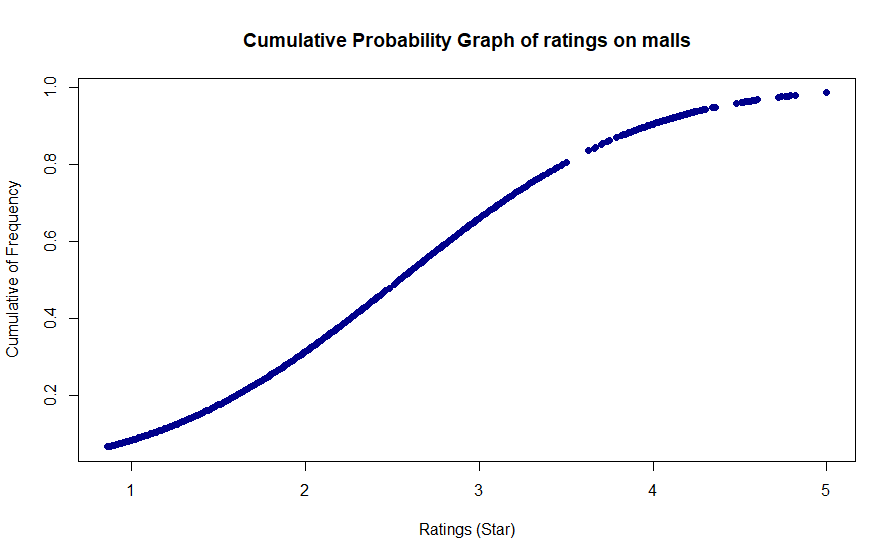
****



Probability Density Graph of ratings on malls

1. หาค่า mean และ sd ของคอลัมน์ที่ต้องการจะทำการหา Probability Density
2. หลังจากนั้นใช้ตัวแปร “DENS” รับค่าคำสั่ง dnorm ซึ่งเป็นคำสั่งในการหา Probability Density โดยต้องใส่ค่า ข้อมูลของคอลัมน์ mean และ sd และสุดท้ายทำการ plot() โดยค่าในแกน x เป็นคะแนนเต็ม 5 ดาว แกน y เป็นค่าการสะสมที่สอดคล้องกับค่า mean และ sd
3. จะได้กราฟของ Probability Density Graph of ratings on malls มีหน่วยเป็น Star

****



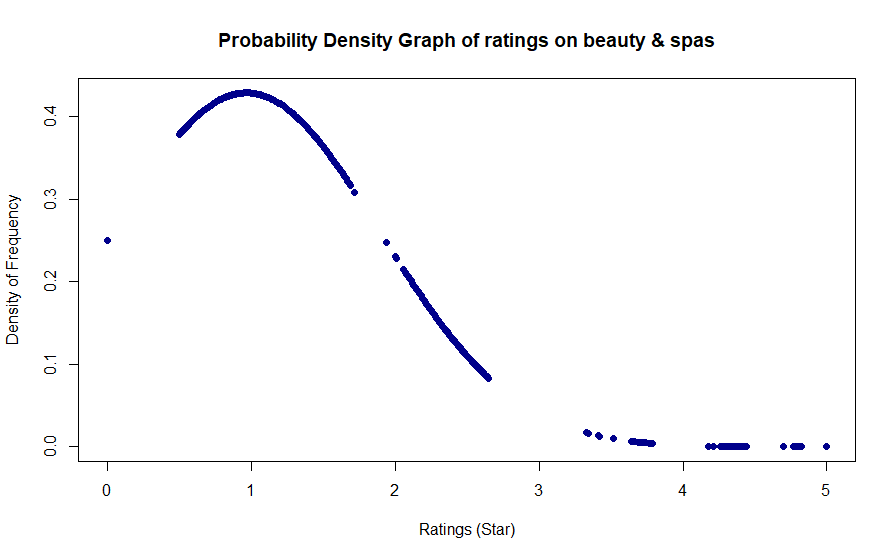
Cumulative Distribution Graph of ratings on malls

1. ใช้ตัวแปร “CUM” รับค่าคำสั่ง pnorm ซึ่งเป็นคำสั่งในการหา Cumulative Distribution โดยต้องใส่ค่า ข้อมูลของคอลัมน์, mean และ sd และสุดท้ายทำการ plot() โดยค่าในแกน x เป็นคะแนนเต็ม 5 ดาว แกน y เป็นค่าการสะสมที่สอดคล้องกับค่า mean และ sd
2. จะได้กราฟของ Cumulative Distribution Graph of ratings on malls มีหน่วยเป็น Star

## Average ratings on beauty & spas

****

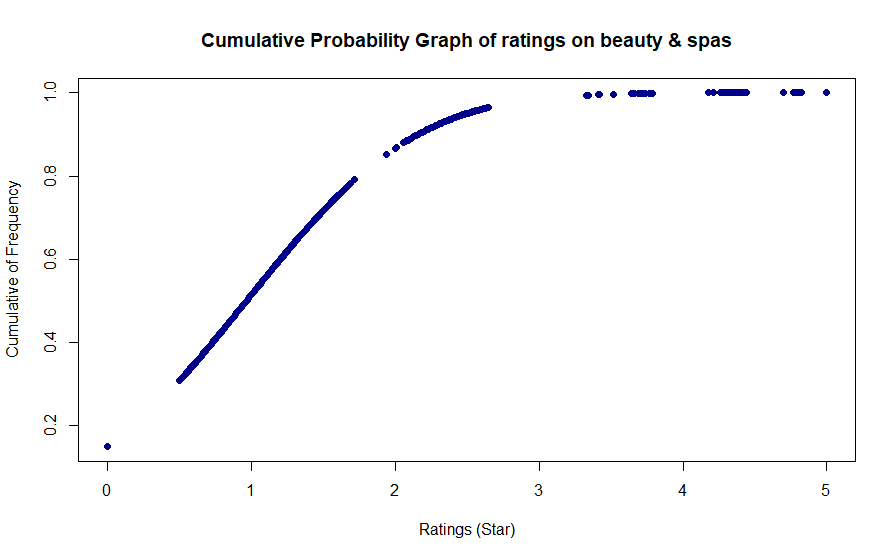
****



Probability Density Graph of ratings on beauty & spas

1. หาค่า mean และ sd ของคอลัมน์ที่ต้องการจะทำการหา Probability Density
2. หลังจากนั้นใช้ตัวแปร “DENS” รับค่าคำสั่ง dnorm ซึ่งเป็นคำสั่งในการหา Probability Density โดยต้องใส่ค่า ข้อมูลของคอลัมน์ mean และ sd และสุดท้ายทำการ plot() โดยค่าในแกน x เป็นคะแนนเต็ม 5 ดาว แกน y เป็นค่าการสะสมที่สอดคล้องกับค่า mean และ sd
3. จะได้กราฟของ Probability Density Graph of ratings on beauty & spas มีหน่วยเป็น Star

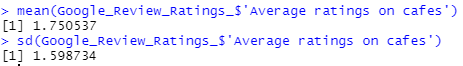
****



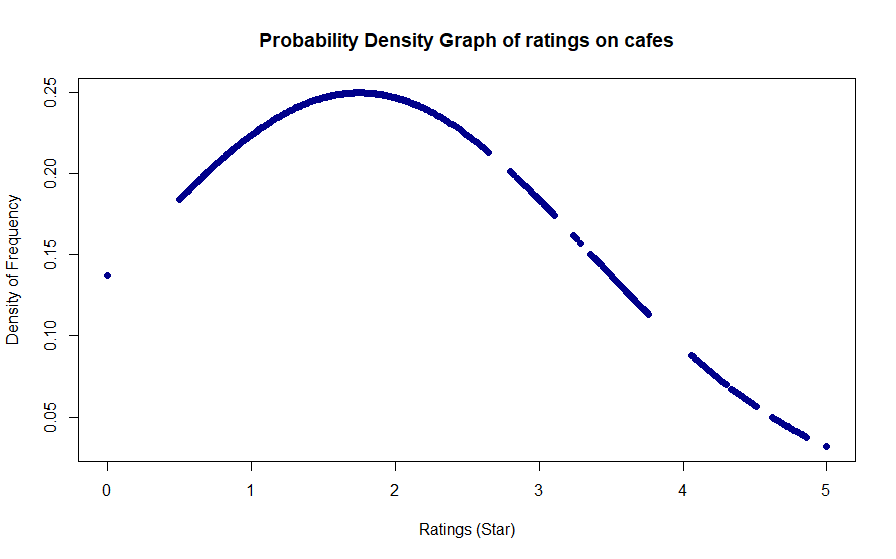
Cumulative Distribution Graph of ratings on beauty & spas

1. ใช้ตัวแปร “CUM” รับค่าคำสั่ง pnorm ซึ่งเป็นคำสั่งในการหา Cumulative Distribution โดยต้องใส่ค่า ข้อมูลของคอลัมน์, mean และ sd และสุดท้ายทำการ plot() โดยค่าในแกน x เป็นคะแนนเต็ม 5 ดาว แกน y เป็นค่าการสะสมที่สอดคล้องกับค่า mean และ sd
2. จะได้กราฟของ Cumulative Distribution Graph of ratings on beauty & spas มีหน่วยเป็น Star

## Average ratings on cafes

****

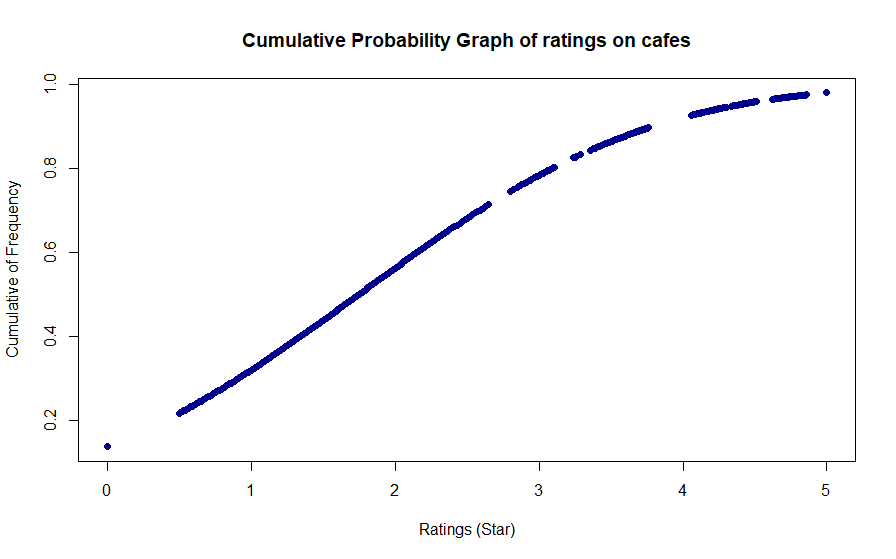
****



Probability Density Graph of ratings on cafes

1. หาค่า mean และ sd ของคอลัมน์ที่ต้องการจะทำการหา Probability Density
2. หลังจากนั้นใช้ตัวแปร “DENS” รับค่าคำสั่ง dnorm ซึ่งเป็นคำสั่งในการหา Probability Density โดยต้องใส่ค่า ข้อมูลของคอลัมน์ mean และ sd และสุดท้ายทำการ plot() โดยค่าในแกน x เป็นคะแนนเต็ม 5 ดาว แกน y เป็นค่าการสะสมที่สอดคล้องกับค่า mean และ sd
3. จะได้กราฟของ Probability Density Graph of ratings on cafes มีหน่วยเป็น Star

****



Cumulative Distribution Graph of ratings on cafes

1. ใช้ตัวแปร “CUM” รับค่าคำสั่ง pnorm ซึ่งเป็นคำสั่งในการหา Cumulative Distribution โดยต้องใส่ค่า ข้อมูลของคอลัมน์, mean และ sd และสุดท้ายทำการ plot() โดยค่าในแกน x เป็นคะแนนเต็ม 5 ดาว แกน y เป็นค่าการสะสมที่สอดคล้องกับค่า mean และ sd
2. จะได้กราฟของ Cumulative Distribution Graph of ratings on cafes มีหน่วยเป็น Star

## บทวิเคราะห์ข้อมูล

ชุดข้อมูลนี้มาจากที่เก็บแมชชีนเลิร์นนิงของมหาวิทยาลัยแคลิฟอร์เนีย, เออร์ไวน์ (UC Irvine) : ข้อมูลการจัดอันดับรีวิวการเดินทาง ชุดข้อมูลนี้จะถูกเติมโดยการจับคะแนนของผู้ใช้จากรีวิวของ Google รีวิวเกี่ยวกับสถานที่ท่องเที่ยวจาก 24 หมวดหมู่ทั่วยุโรปได้รับการพิจารณา ทางผุ้จัดทำได้หยิบยกคะแนนผู้ใช้ Google มีตั้งแต่ 1 ถึง 5 และมีการคํานวณคะแนนผู้ใช้เฉลี่ยต่อหมวดหมู่มาคิดวิเคราะห์

จากกราฟ Probability Density Graph ที่ทางผู้จัดทำได้แยกจัดทำแบ่งเป็น 5 ประเภทได้แก่ส่วนของ beaches, parks, malls, beauty & spas และ cafes วิเคราะห์ได้ว่า จากความหนาแน่นในแต่ละช่วงของคะแนนโดยรวมของทั้ง 5 ประเภท นั้นสามารถบ่งบอกได้ว่าจำนวน User ผู้รีวิวสถานที่แต่ละประเภท มีการให้คะแนนในแต่ละประเภทไปทางที่สูงมาก และมีคะแนนการรีวิวเฉลี่ยสูงเกินกว่าครึ่ง จาก User ผู้รีวิวสถานที่ทั้งหมดของ จากภาพโดยรวมจะเห็นว่าแนวโน้มในช่วง 0 – 3 ดาว มีความหนาแน่นสูง ทำให้เราสามารถวิเคราะห์ได้ว่า User ผู้รีวิวสถานที่ ส่วนใหญ่อยู่ในช่วง 0 – 3 ดาว ในประเภท beaches, parks, malls แต่ในประเภท beauty & spas และ cafes แนวโน้มในช่วง 0 – 1 และ 1 - 2 ดาว ตามลำดับ ซึ่ง User ผู้รีวิวสถานที่ การให้คะแนนมีความหนาแน่นสูงใน 2 ตามประเภทที่กล่าวมา

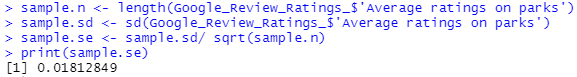
จากกราฟ Cumulative Distribution Graph ที่ทางผู้จัดทำได้แยกจัดทำแบ่งเป็น 5 ประเภทได้แก่ส่วนของ beaches, parks, malls, beauty & spas และ cafes วิเคราะห์ได้ว่าทั้ง 5 ประเภทมีความชันที่ไม่คงที่ และจะมีการกระจุกตัวของข้อมูลของช่วงต้นของกราฟมากกว่า กราฟในช่วงหลัง ซึ่งแสดงถึงการสะสมของ คะแนนเฉลี่ยของ User ผู้รีวิวสถานที่ ข้อมูลที่ได้จะสอดคล้องกับกราฟ Probability Density Graph ที่ได้ ยิ่ง Cumulative Distribution Graph มีความชันน้อย กราฟจะมีการกระจุกตัวของข้อมูลก็ยิ่งมีค่าใกล้เคียง หรืออาจจะเท่ากัน อยู่มาก ถ้าความชันมาก และค่าต่างกันมาก ถึงเล็กน้อย กราฟจะมีความชันมากแบบต่อเนื่อง

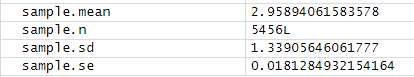
สามารถสรุปได้ว่าการใช้กราฟ Probability Density และ Cumulative Distribution สามารถนำมาใช้วิเคราะห์ ข้อมูลจากแนวโน้มความหนาแน่นหรือการสะสมของข้อมูลที่สนใจได้

# PROBABILITY AND STATISTICS HW 4

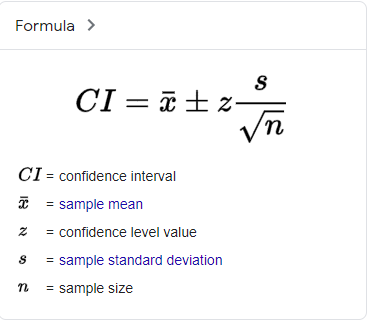
**ในการหา Confidence Interval (CI) ทางผู้จัดทำได้เลือก คอลัมน์ Average ratings on parks**

****





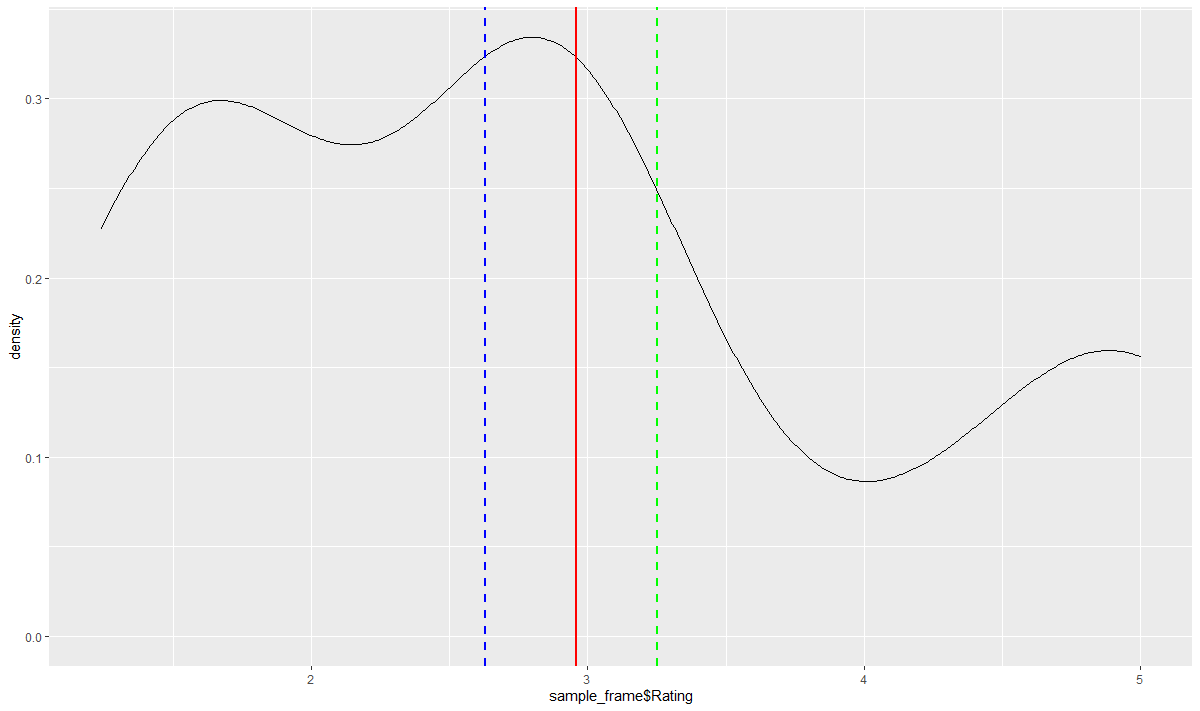
จากรูปจะเห็นค่าของ Mean, Sample Size, Standard Deviation และ Standard error



นี่คือสูตรในการใช้คำนวนหาค่า Confidence Interval โดยในที่นี้จะใช้จากเว็บไซต์

[Easy Confidence Interval Calculator (socscistatistics.com)](https://www.socscistatistics.com/confidenceinterval/default3.aspx)

## ค่า Confidence Interval ที่ Confidence Level 90%



Star(s)

M = 2.958941  
Z = 1.64

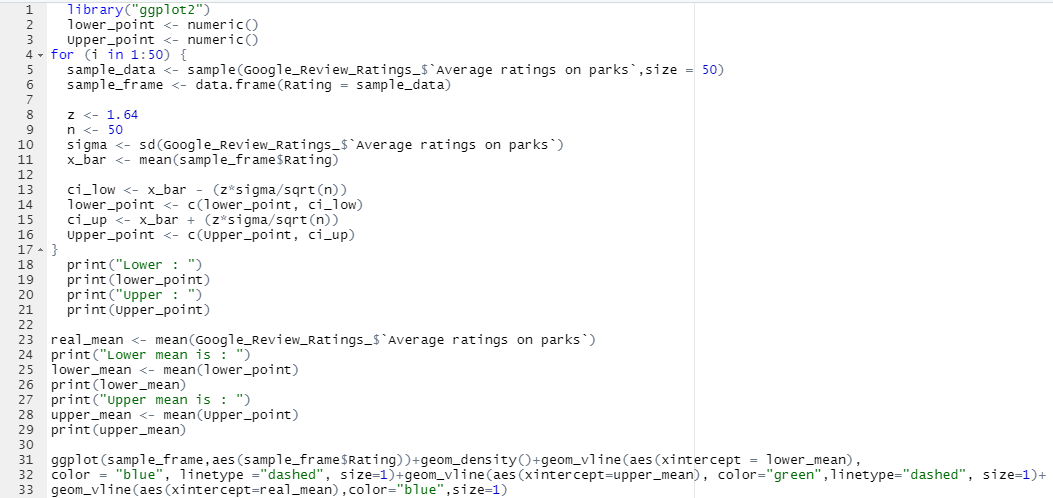
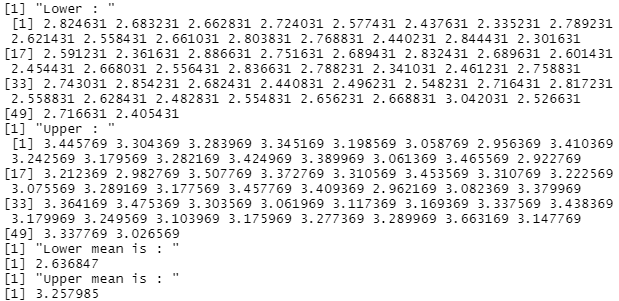
ค่า Confidence Interval ของคะแนนการรีวิว คอลัมน์ Average ratings on parks

ที่ Confidence Level 90% คือระหว่างช่วง [2.63687, 3.257985] คะแนน

**จากกราฟจะเห็นได้ว่าช่วง Confidence Level 90%**

**คือระหว่างช่วง [2.63687, 3.257985] คะแนน**

สูตรที่ใช้หา **Confidence Level 90%**



ในที่นี้เลือกใช้ข้อมูลจาก Rating on parks แบ่งข้อมูลออกเป็น 50 ชุดข้อมูล ชุดละ 50 ข้อมูล ให้แต่ละข้อมูล ทำการหาค่า CI ที่ 90% นำข้อมูลทั้งหมดมาเฉลี่ยหา Upper และ Lower CI และนำมา plot ลงกราฟ เส้นประสีเขียว และ เส้นประสีน้ำเงิน และเส้นสีแดงคือ mean ของข้อมูลทั้งหมด

Lower คือ ข้อมูล Lower CI แต่ละชุดข้อมูลที่สุ่มออกมา

Upper คือ ข้อมูล Upper CI แต่ละชุดข้อมูลที่สุ่มออกมา

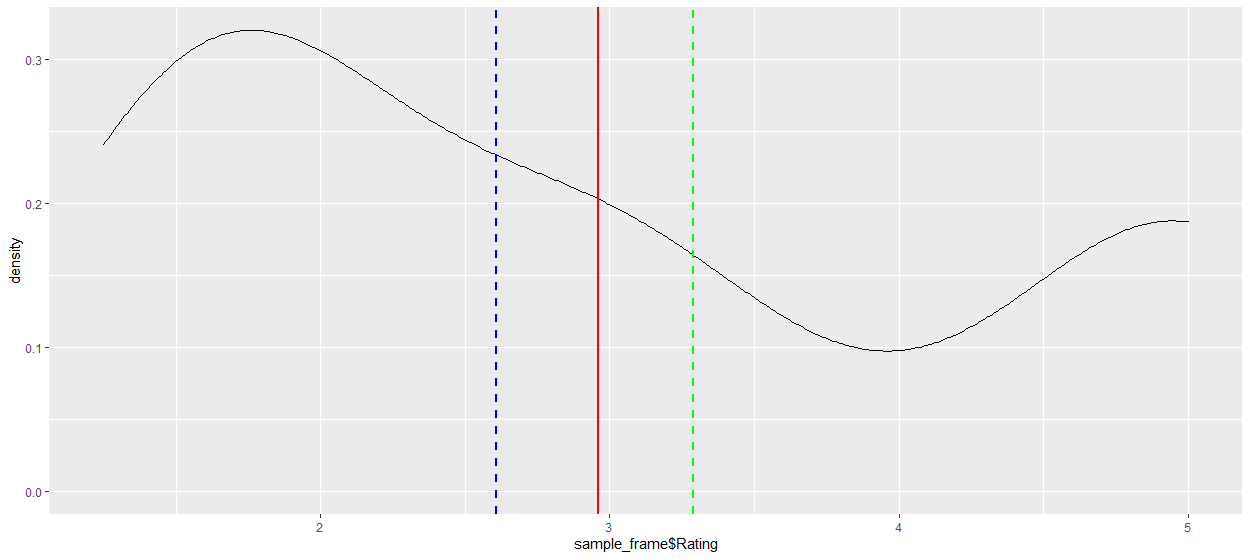
Lower mean คือ ค่าเฉลี่ยข้อมูลของ Lower CI ทั้งหมด มีค่า 2.63687

Upper mean คือ ค่าเฉลี่ยข้อมูลของ Upper CI ทั้งหมด มีค่า 3.257985

population mean มีค่า 43.933

โดยมีข้อมูลจำนวน 4 ชุดข้อมูลที่ไม่คลอบคลุม population mean

## ค่า Confidence Interval ที่ Confidence Level 95%



Star(s)

M = 2.958941  
Z = 1.96

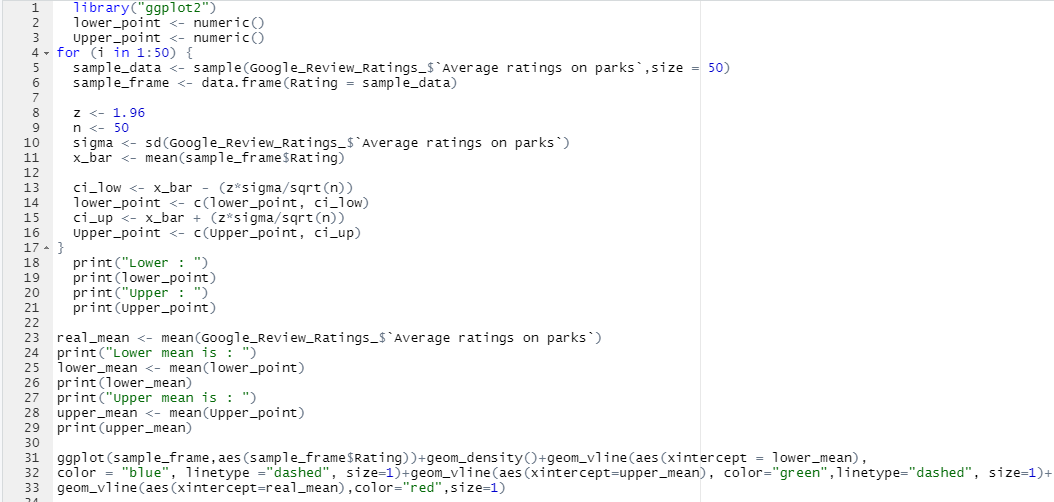
ค่า Confidence Interval ของคะแนนการรีวิว คอลัมน์ Average ratings on parks

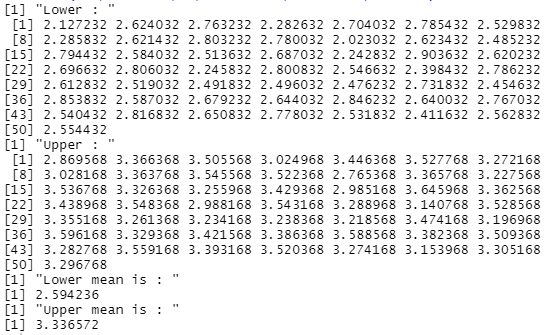
ที่ Confidence Level 95% คือระหว่างช่วง [2.594236, 3.336572] คะแนน

**จากกราฟที่ขยายจะเห็นได้ว่าช่วง Confidence Level 95%**

**คือระหว่างช่วง [2.594236, 3.336572] คะแนน**

สูตรที่ใช้หา **Confidence Level 95%**





ในที่นี้เลือกใช้ข้อมูลจาก Rating on parks แบ่งข้อมูลออกเป็น 50 ชุดข้อมูล ชุดละ 50 ข้อมูล ให้แต่ละข้อมูล ทำการหาค่า CI ที่ 95% นำข้อมูลทั้งหมดมาเฉลี่ยหา Upper และ Lower CI และนำมา plot ลงกราฟ เส้นประสีเขียว และ เส้นประสีน้ำเงิน และเส้นสีแดงคือ mean ของข้อมูลทั้งหมด

Lower คือ ข้อมูล Lower CI แต่ละชุดข้อมูลที่สุ่มออกมา

Upper คือ ข้อมูล Upper CI แต่ละชุดข้อมูลที่สุ่มออกมา

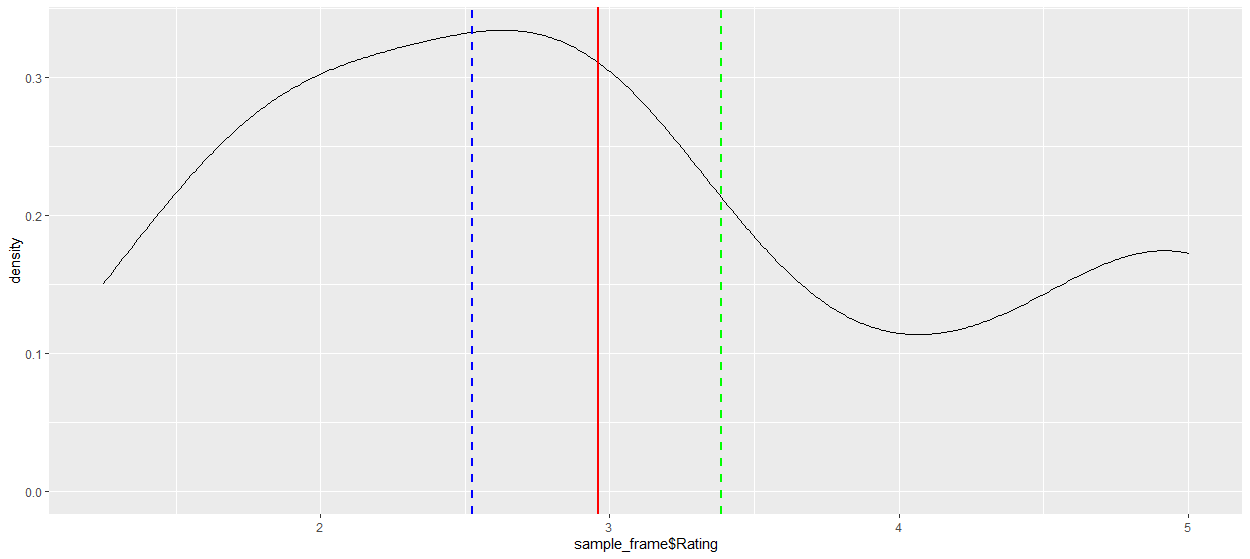
Lower mean คือ ค่าเฉลี่ยข้อมูลของ Lower CI ทั้งหมด มีค่า 2.594236

Upper mean คือ ค่าเฉลี่ยข้อมูลของ Upper CI ทั้งหมด มีค่า 3.336572

population mean มีค่า 43.933

โดยมีข้อมูลจำนวน 4 ชุดข้อมูลที่ไม่คลอบคลุม population mean

## ค่า Confidence Interval ที่ Confidence Level 99%



Star(s)

M = 2.958941  
Z = 2.58

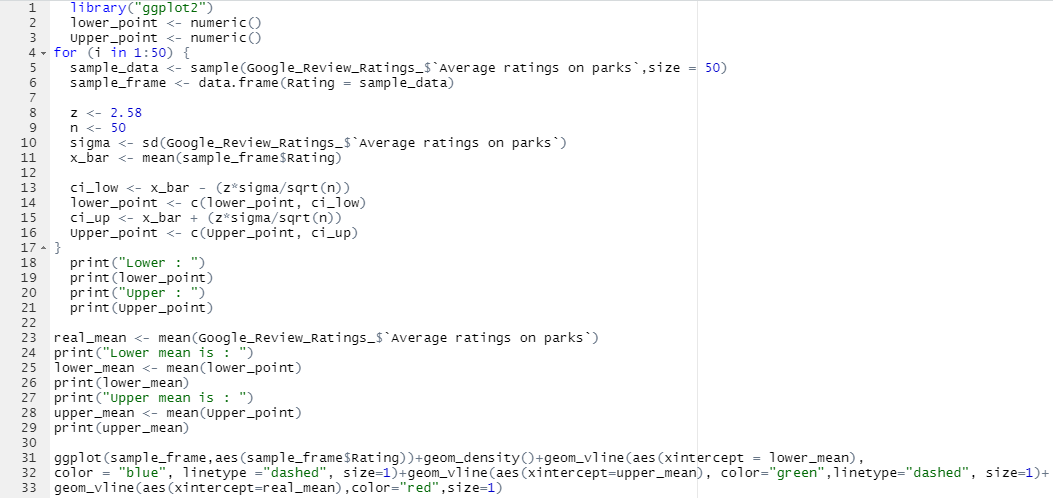
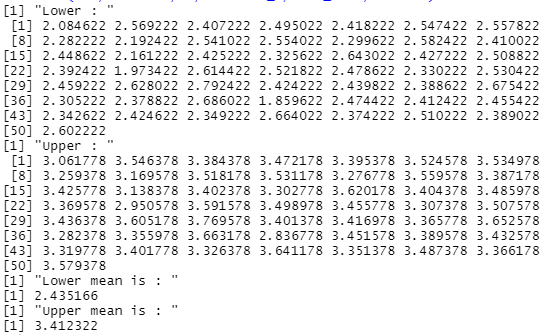
ค่า Confidence Interval ของคะแนนการรีวิว คอลัมน์ Average ratings on parks

ที่ Confidence Level 99% คือระหว่างช่วง [2.435166, 3.412322] คะแนน

**จากกราฟที่ขยายจะเห็นได้ว่าช่วง Confidence Level 99%**

**คือระหว่างช่วง [2.435166, 3.412322] คะแนน**

สูตรที่ใช้หา **Confidence Level 99%**



ในที่นี้เลือกใช้ข้อมูลจาก Rating on parks แบ่งข้อมูลออกเป็น 50 ชุดข้อมูล ชุดละ 50 ข้อมูล ให้แต่ละข้อมูล ทำการหาค่า CI ที่ 99% นำข้อมูลทั้งหมดมาเฉลี่ยหา Upper และ Lower CI และนำมา plot ลงกราฟ เส้นประสีเขียว และ เส้นประสีน้ำเงิน และเส้นสีแดงคือ mean ของข้อมูลทั้งหมด

Lower คือ ข้อมูล Lower CI แต่ละชุดข้อมูลที่สุ่มออกมา

Upper คือ ข้อมูล Upper CI แต่ละชุดข้อมูลที่สุ่มออกมา

Lower mean คือ ค่าเฉลี่ยข้อมูลของ Lower CI ทั้งหมด มีค่า 2.435166

Upper mean คือ ค่าเฉลี่ยข้อมูลของ Upper CI ทั้งหมด มี 3.412322

population mean มีค่า 43.933

โดยมีข้อมูลจำนวน 4 ชุดข้อมูลที่ไม่คลอบคลุม population mean

## บทวิเคราะห์ข้อมูล

**จากการทำการหาค่า Confidence Interval ได้ผลลัพธ์ดังนี้**

ค่า Confidence Interval ของคะแนนการรีวิว คอลัมน์ Average ratings on parks

ที่ Confidence Level 90% คือระหว่างช่วง [2.63687, 3.257985] คะแนน

ค่า Confidence Interval ของคะแนนการรีวิว คอลัมน์ Average ratings on parks

ที่ Confidence Level 95% คือระหว่างช่วง [2.594236, 3.336572] คะแนน

ค่า Confidence Interval ของคะแนนการรีวิว คอลัมน์ Average ratings on parks

ที่ Confidence Level 99% คือระหว่างช่วง [2.435166, 3.412322] คะแนน

**เราสามารถวิเคราะห์ข้อมูลจากค่า Confidence Interval ได้ว่า**

คะแนนการรีวิว คอลัมน์ Average ratings on parks ของ User กว่า 90% นั้นมีการให้คะแนนเฉลี่ยอยู่คือระหว่างช่วง [2.63687, 3.257985] คะแนน

คะแนนการรีวิว คอลัมน์ Average ratings on parks ของ User กว่า 95% นั้นมีการให้คะแนนเฉลี่ยอยู่คือระหว่างช่วง [2.594236, 3.336572] คะแนน

คะแนนการรีวิว คอลัมน์ Average ratings on parks ของ User กว่า 99% นั้นมีการให้คะแนนเฉลี่ยอยู่คือระหว่างช่วง [2.435166, 3.412322] คะแนน

**และสรุปจากคะแนนทั้งหมดได้ว่า**

ค่าคะแนนการรีวิวเกี่ยวกับ Travel Review Rating Dataset ( From the Machine Learning Repository of University of California ) นั้นมีข้อสังเกตุ และข้อสงสัยที่ว่าคะแนนที่ได้มานั้น ค่าเฉลี่ยไปกองรวมกันที่ 2.9 โดยประมาณ และในการโหวตคะแนนสังเกตุว่า มีการให้คะแนนเต็ม (5 ดาว) กันหลาย User ซึ่งบ่งบอกได้หลายกรณีเช่น อาจมีการรับจ้างรีวิว, อาจจะมีการใช้บอททำหน้าที่โหวตแทน, อาจจะรีวิวโดยให้คะแนนสุ่ม, อาจจะรีวิวเพื่อแค่แลกของรางวัลที่ทางสถานที่นั้น ๆ จัดขึ้น, หรืออาจจะรีวิวจากการได้รับส่วนลดค่าที่พัก เป็นต้น ซึ่งข้อสันนิษฐานของผู้จัดทำข้างต้นนี้มีความเป็นไปได้สูงที่จะสามารถเกิดขึ้น เพราะด้วยสังคมโลกยุคปัจจุบันการเอาตัวรอดในสถานการณ์ต่าง ๆ เป็นสิ่งสำคัญ สถานที่หรือผู้ประกอบการบริษัทนั้น ๆ สามารถลอกลวงผู้บริโภคเพื่อประโยชน์ส่วนตนได้ ดั่งเช่นตัวอย่างจากข่าวนี้ [ร้อง สคบ. เอาผิดเว็บจองที่พักชื่อดัง หลอกลวงผู้บริโภค | ข่าวช่อง 8 (thaich8.com)](https://www.thaich8.com/news_detail/54898) ซึ่งทางผู้จัดทำหวังเป็นอย่างยิ่งว่าผู้ชมที่เข้ามาศึกษางานนี้จะมีวิธีรับมือ และไม่หลงเชื่ออะไรง่าย ๆ ครับ

# PROBABILITY AND STATISTICS HW 5

## หา Linear Regression ทำคู่กับ นายศุภกฤต โลห์แก้ว 62010889

## Graph

Chart, scatter chart

Description automatically generated

กราฟ Linear Regression แกน y เป็นราคาเปิดของ Bitcoin แกน x เป็น Volume ของ Bitcoin ของราคาเปิดในวันนั้นๆ

Graphical user interface, text, application

Description automatically generated

ส่วนของโปรแกรมที่ใช้ในการ Plot graph โดยใช้คำสั่ง lm ในการหาสมการที่ใช้คำนวณ Linear Regression

## Coefficients

Text

Description automatically generated

การหา Linear Regression สามารถหาได้จากสมการ y = mx + c ผ่านตัวแปร linmod

## R-Square



ส่วนของโปรแกรมในการหาค่า r-square หรือความคลาดเคลื่อนของข้อมูลว่าห่างจากเส้น Linear Regression ที่คำนวณไว้ซึ่งยิ่งค่า r-square เข้าใกล้ 1 มากเท่าไหร่แสดงว่าข้อมูลมีความแม่นยำมากขึ้น

## บทวิเคราะห์ข้อมูล

จากกราฟที่ได้ทำขึ้นเป็นกราฟความสัมพันธ์ของราคาเปิด และวอลุ่ม ของ Bitcoin ในแต่ละวัน โดยมีเส้น linear regression ซึ่งมีสมการคือ y = mx + c โดย m แสดงถึงความชันของกราฟ และ c แสดงถึงจุดตัดแกน y โดยมี x, y เป็นตัวแปรตามแนวแกนของกราฟ ซึ่งก็คือ ราคาเปิดและวอลุ่ม โดยความชันของกราฟนี้ ได้แก่ค่า Coefficients ตามข้อมูลด้านบน และ ค่า c สามารถหาได้จากกราฟ ซึ่งเมื่อเรารู้ราคาเปิดของวันนั้น ก็จะคาดการณ์ได้ถึงวอลุ่มวันนั้นด้วย ส่วนค่า R-Square มีค่าอยู่ที่ 0.838 เมื่อคิดเป็นเปอร์เซ็นแล้ว จะมีค่าเท่ากับ 83.8% แสดงให้เห็นว่า วอลุ่ม สามารถอธิบายได้ด้วยราคาเปิด โดยที่มีความแม่นยำ 83.8%