



นำเสนอ
ผศ.ดร.สุรินทร์ กิตติธรรมกุล

จัดทำโดย
62010889 นายศุภกฤต โล่ห์แก้ว
62011019 นายอภิรักษ์ อุลิศ

รายงานฉบับนี้เป็นส่วนหนึ่งของรายวิชา
01076253 PROBABILITY AND STATISTICS
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง
ภาคเรียนที่ 2 ปีการศึกษา 2563

สารบัญ

PROBABILITY AND STATISTICS HW 13

 ชื่อชุดข้อมูล Travel Review Rating Dataset.....3

 ชื่อคอลัมน์.....3

 ทำไมถึงสนใจ3

 คำอธิบายชื่อคอลัมน์ข้อมูล และวิธีการรวบรวมข้อมูล.....3

PROBABILITY AND STATISTICS HW 24

 ข้อมูลการคำนวณข้อมูลเชิงสถิติพื้นฐานแต่ละคอลัมน์.....5

 บทวิเคราะห์ข้อมูล..... 13

PROBABILITY AND STATISTICS HW 3 14

 Average ratings on beaches 14

 Average ratings on parks..... 16

 Average ratings on malls..... 18

 Average ratings on beauty & spas..... 20

 Average ratings on cafes..... 22

 บทวิเคราะห์ข้อมูล..... 24

PROBABILITY AND STATISTICS HW 4 25

 ค่า Confidence Interval ที่ Confidence Level 90%..... 26

 ค่า Confidence Interval ที่ Confidence Level 95%..... 28

 ค่า Confidence Interval ที่ Confidence Level 99%..... 30

 บทวิเคราะห์ข้อมูล..... 32

PROBABILITY AND STATISTICS HW 5 33

 หา Linear Regression ทำคู่กับ นายศุภกฤต โลห์แก้ว 62010889 33

 Graph..... 33

 Coefficients..... 34

 R-Square..... 34

 บทวิเคราะห์ข้อมูล..... 34

PROBABILITY AND STATISTICS HW 1

ชื่อชุดข้อมูล Travel Review Rating Dataset

ชื่อคอลัมน์

Numeric -> ประเภทการรีวิวแยก 5 ประเภท

Category -> หมายเลข User ผู้รีวิวสถานที่เหล่านั้น

ทำไมถึงสนใจ

ผมเป็นคนที่ชอบ และชอบในการเดินทางท่องเที่ยวไปยังสถานที่ต่าง ๆ แต่บางครั้งผมก็มีปัญหาอยู่ในใจมากมาย เช่น ผมควรจะไปเยี่ยมชมสถานที่ไหนดี ? มีสถานที่ที่น่าสนใจที่ตรงกับไลฟ์สไตล์ของผม บ่อยครั้งที่ผมใช้เวลาหลายชั่วโมงเพื่อค้นหาสถานที่ที่น่าสนใจที่จะออกไปข้างนอก

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

แหล่งที่มาของชุดข้อมูล [Travel Review Rating Dataset | Kaggle](#)

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

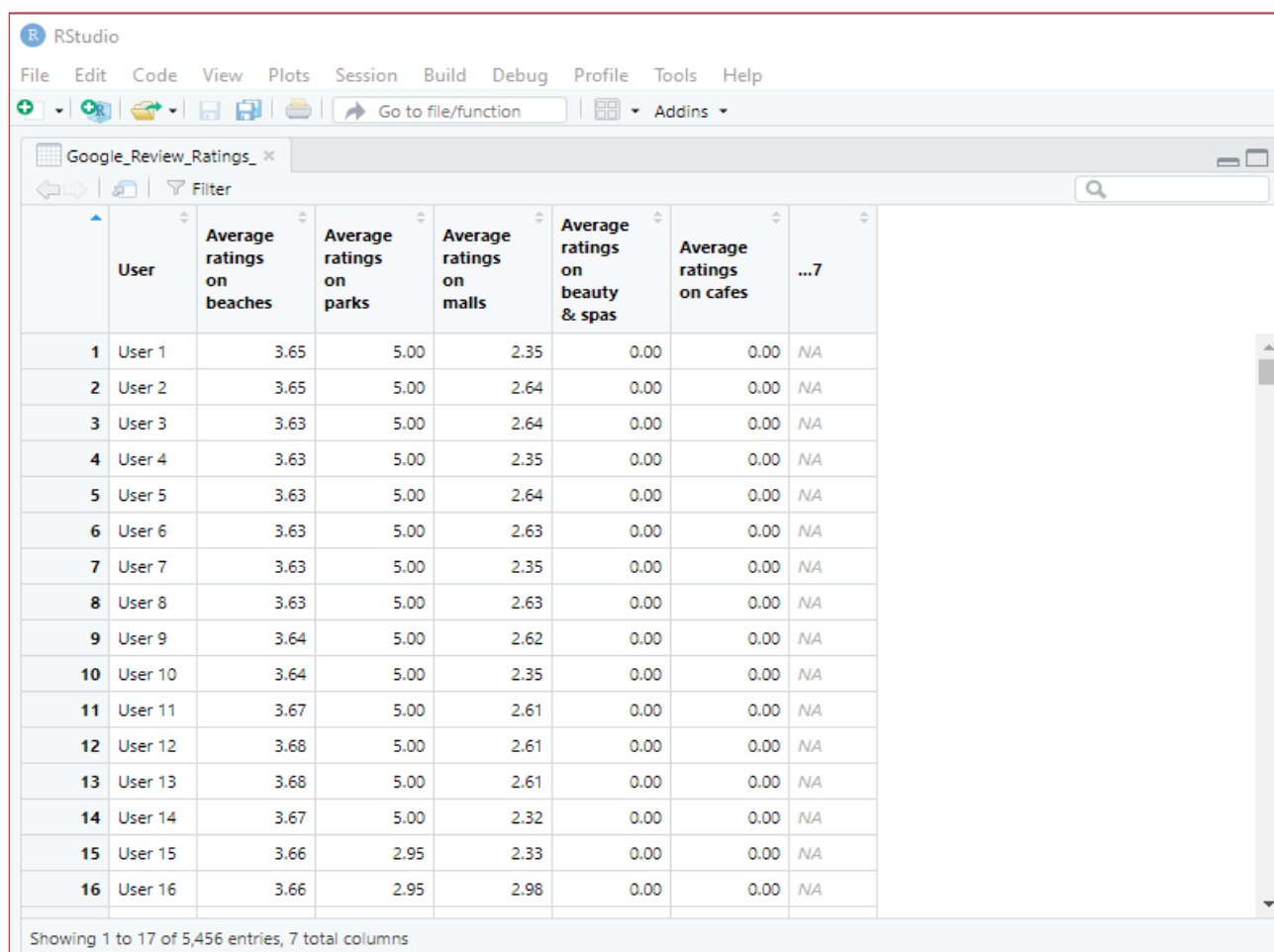
จำนวนผู้ใช้งานที่เข้ามาให้การแนะนำ หรือรีวิวในประเภทต่าง ๆ

วิธีการรวบรวมข้อมูล

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

PROBABILITY AND STATISTICS HW 2

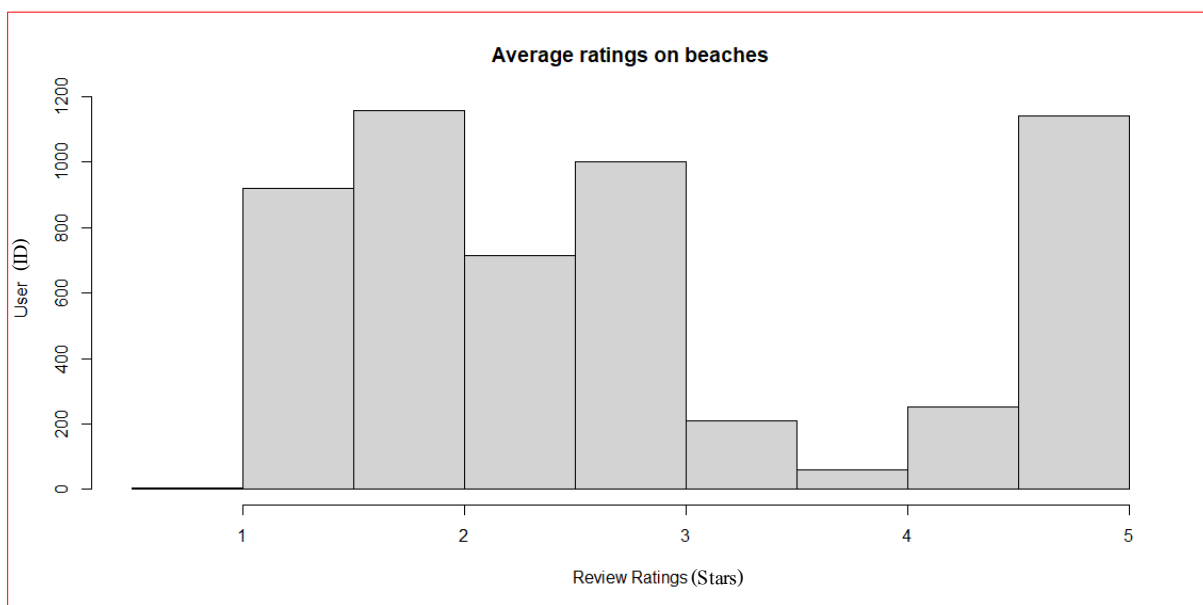
สืบเนื่องมาจาก HW 1 ที่ทางผู้จัดทำได้หยิบยกชุดข้อมูลเกี่ยวกับ Travel Review Rating Dataset (From the Machine Learning Repository of University of California) โดยประกอบด้วยคอลัมน์ Numeric (ประเภทการรีวิวแยก 5 ประเภท), และ Category (หมายเลข User ผู้รีวิวสถานที่เหล่านั้น) และเพื่อที่จะได้คำนวณหาค่าสถิติพื้นฐาน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



	User	Average ratings on beaches	Average ratings on parks	Average ratings on malls	Average ratings on beauty & spas	Average ratings on cafes	...7
1	User 1	3.65	5.00	2.35	0.00	0.00	NA
2	User 2	3.65	5.00	2.64	0.00	0.00	NA
3	User 3	3.63	5.00	2.64	0.00	0.00	NA
4	User 4	3.63	5.00	2.35	0.00	0.00	NA
5	User 5	3.63	5.00	2.64	0.00	0.00	NA
6	User 6	3.63	5.00	2.63	0.00	0.00	NA
7	User 7	3.63	5.00	2.35	0.00	0.00	NA
8	User 8	3.63	5.00	2.63	0.00	0.00	NA
9	User 9	3.64	5.00	2.62	0.00	0.00	NA
10	User 10	3.64	5.00	2.35	0.00	0.00	NA
11	User 11	3.67	5.00	2.61	0.00	0.00	NA
12	User 12	3.68	5.00	2.61	0.00	0.00	NA
13	User 13	3.68	5.00	2.61	0.00	0.00	NA
14	User 14	3.67	5.00	2.32	0.00	0.00	NA
15	User 15	3.66	2.95	2.33	0.00	0.00	NA
16	User 16	3.66	2.95	2.98	0.00	0.00	NA

Showing 1 to 17 of 5,456 entries, 7 total columns

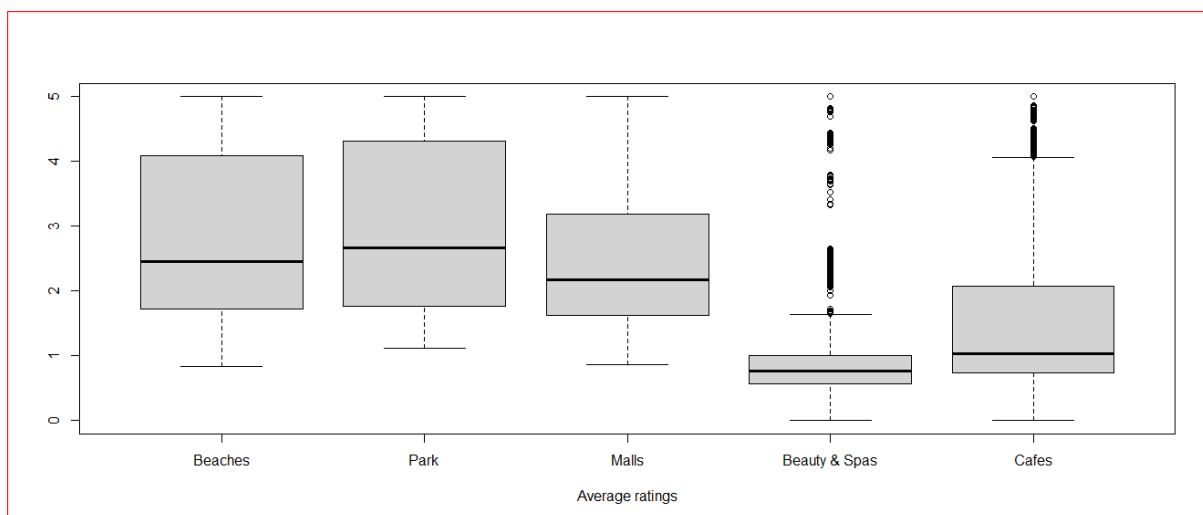
ภาพแสดงตารางข้อมูลเมื่อ import file .csv เข้ามาในโปรแกรม RStudio



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

Code สำหรับสร้าง Histogram

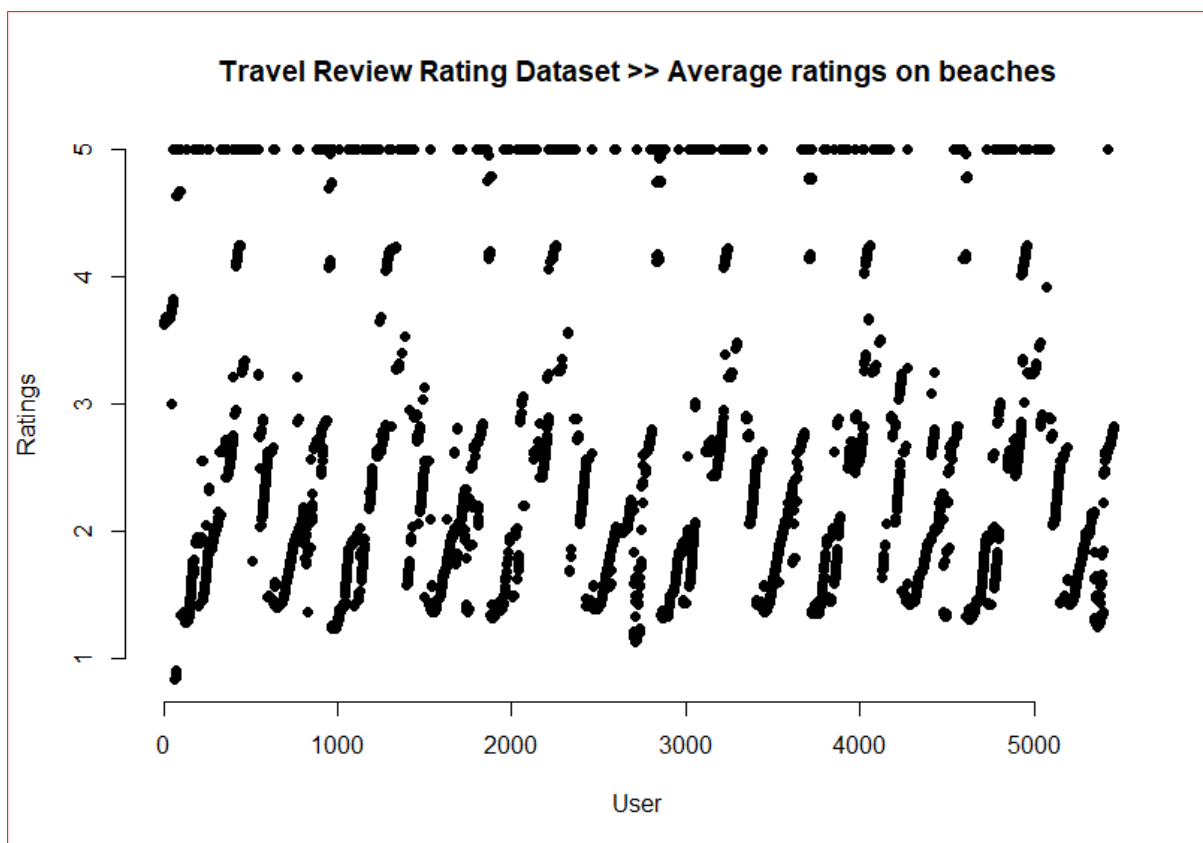
```
> hist(Google_Review_Ratings_.$'Average ratings on beaches', main = "Average ratings on beaches", xlab = "Review Ratings", ylab = "User" )
```



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" )
```



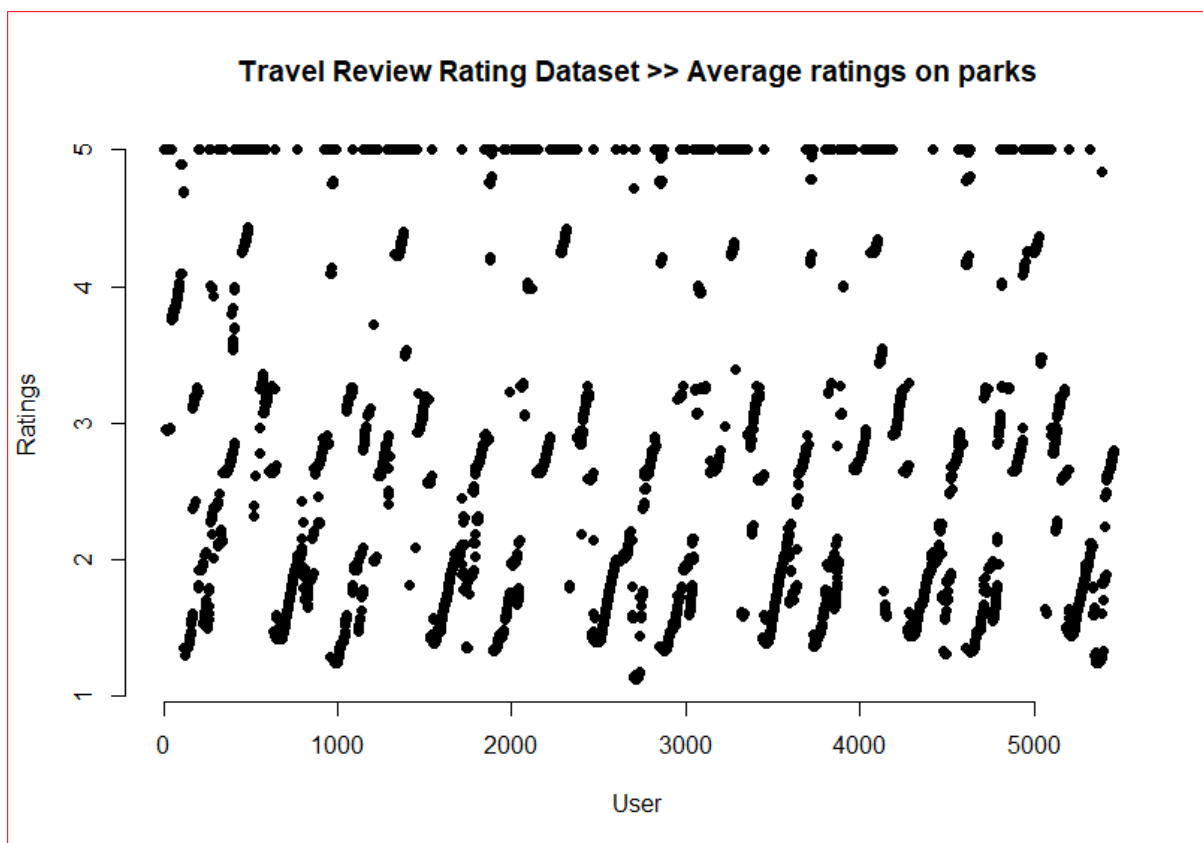
XY Scatter plot (Average ratings on beaches, User ผู้รีวิวสถานที่เหล่านั้นจำนวน 5,456 คน)

Code สำหรับการ Plot XY Scatter Graph

```
> plot(Google_Review_Ratings_$'Average ratings on beaches',main="Travel Review Rating Dataset >> Average ratings on beaches",xlab = "user",ylab = "Ratings",pch=19, frame=FALSE)
> summary(Google_Review_Ratings_$'Average ratings on beaches')
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.830  1.730   2.460   2.797  4.093   5.000
```

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



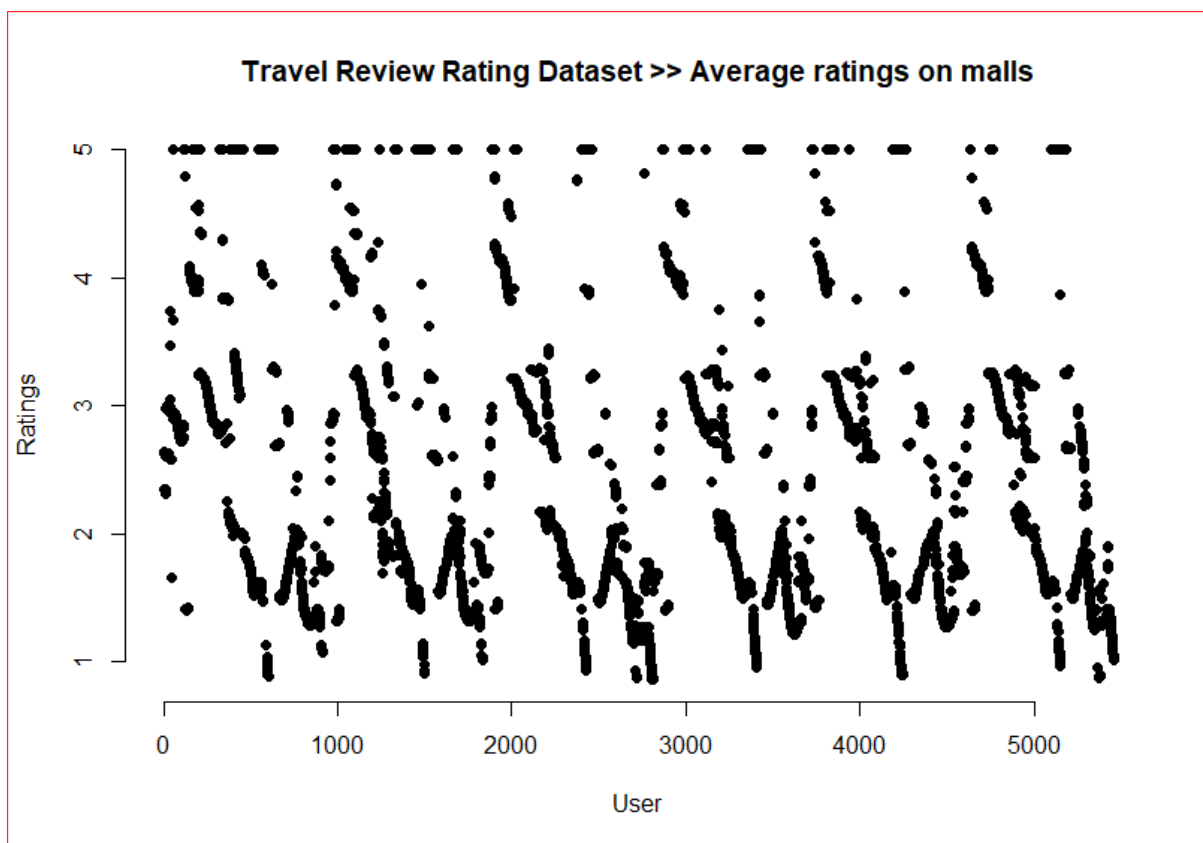
XY Scatter plot (Average ratings on parks, User ผู้รีวิวสถานที่เหล่านั้นจำนวน 5,456 คน)

Code สำหรับการ Plot XY Scatter Graph

```
> plot(Google_Review_Ratings_$'Average ratings on parks',main="Travel Review Rating Dataset >> Average ratings on parks",xlab = "User",ylab = "Ratings",pch=19, frame=FALSE)
> summary(Google_Review_Ratings_$'Average ratings on parks')
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1.120  1.770   2.670   2.959  4.312   5.000
```

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



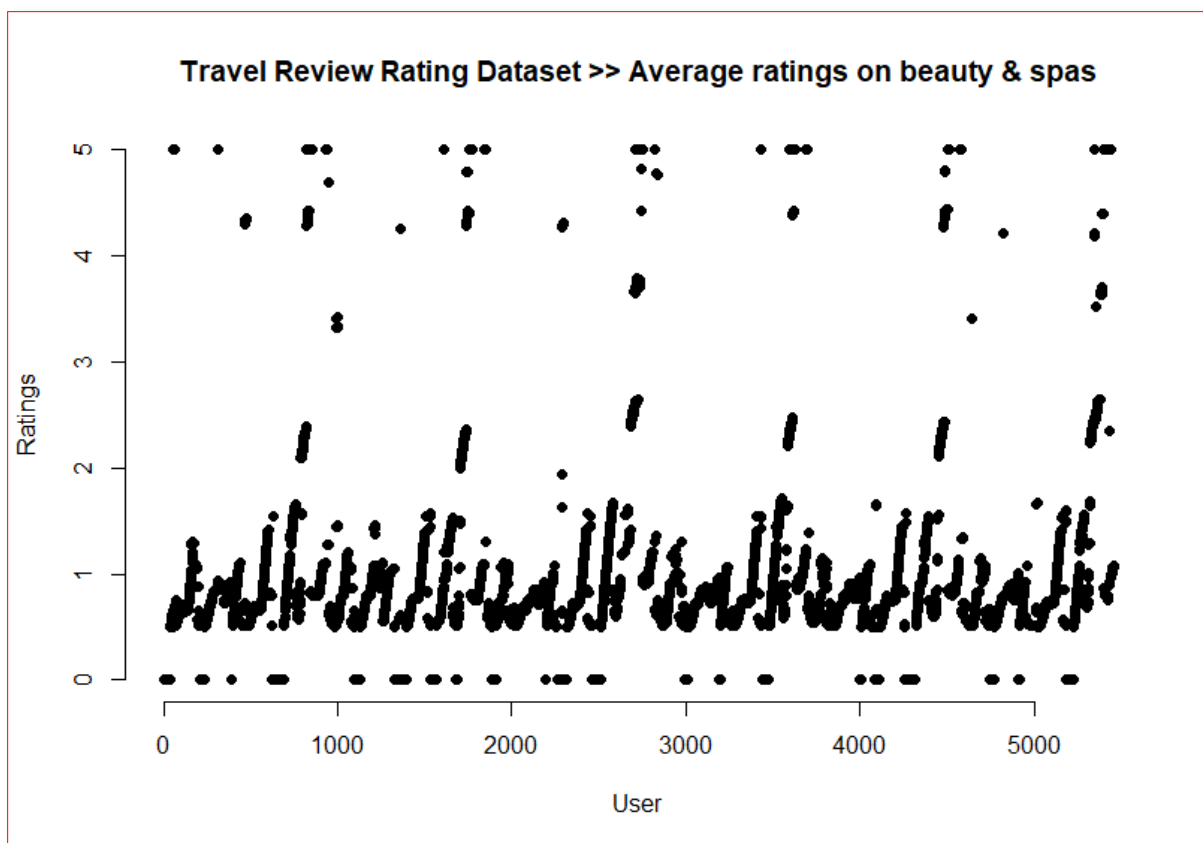
XY Scatter plot (Average ratings on malls, User ผู้รีวิวสถานที่เหล่านั้นจำนวน 5,456 คน)

Code สำหรับการ Plot XY Scatter Graph

```
> plot(Google_Review_Ratings_$'Average ratings on malls',main="Travel Review Rating Dataset >> Average ratings on malls",xlab = "User",ylab = "Ratings",pch=19, frame=FALSE)
> summary(Google_Review_Ratings_$'Average ratings on malls')
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.860  1.620   2.170   2.541   3.190   5.000
```

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



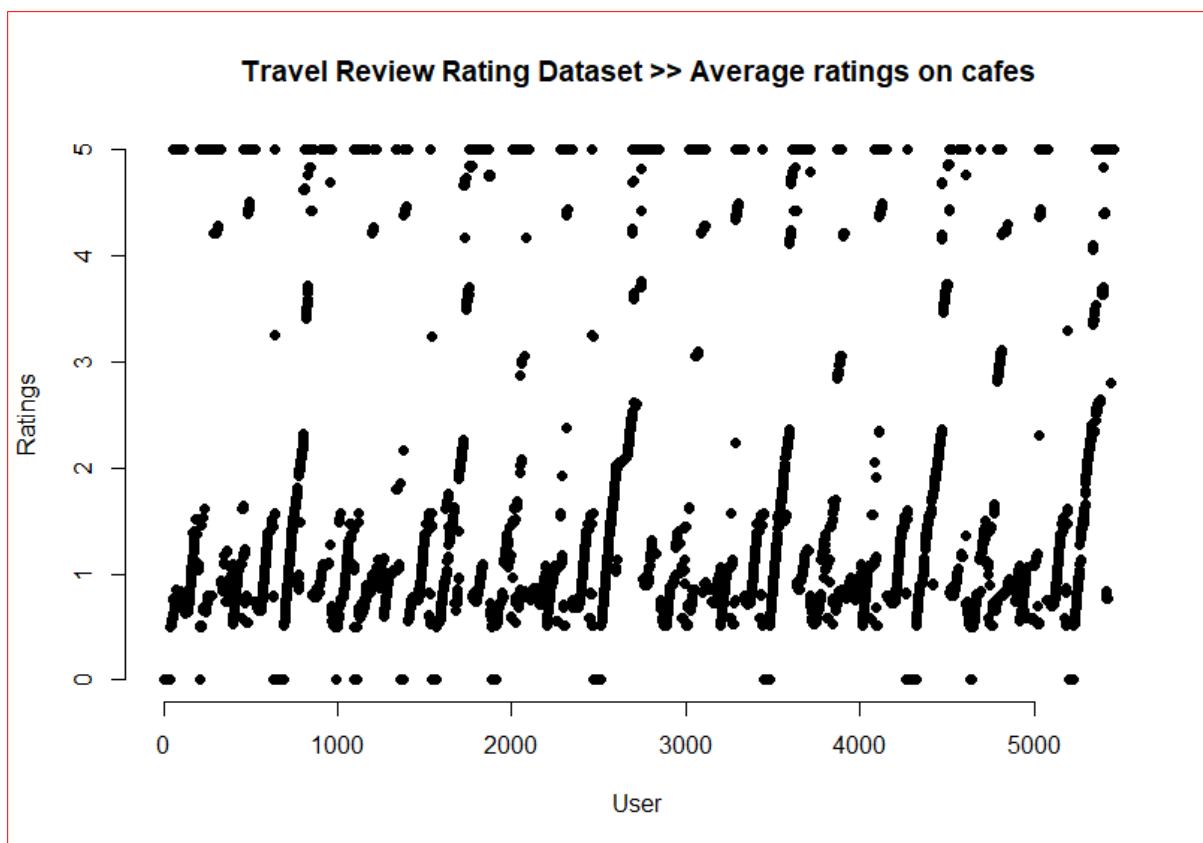
XY Scatter plot (Average ratings on beauty & spas, User ผู้รีวิวสถานที่เหล่านั้นจำนวน 5,456 คน)

Code สำหรับการ Plot XY Scatter Graph

```
> plot(Google_Review_Ratings_.$`Average ratings on beauty & spas`,main="Travel Review Rating Dataset >> Average ratings on beauty & s
pas",xlab = "User",ylab = "Ratings",pch=19, frame=FALSE)
> summary(Google_Review_Ratings_.$`Average ratings on beauty & spas`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.0000  0.5700  0.7600  0.9658  1.0000  5.0000
```

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

```
> plot(Google_Review_Ratings$`Average ratings on cafes`,main="Travel Review Rating Dataset >> Average ratings on cafes",xlab = "User",ylab = "Ratings",pch=19, frame=FALSE)
> summary(Google_Review_Ratings$`Average ratings on cafes`)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.000  0.740   1.030   1.751   2.070   5.000
```

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 ทางผู้จัดทำได้หดยกคะแนนผู้ใช้ Google มีตั้งแต่ 1 ถึง 5 และมีการคำนวณคะแนนเฉลี่ยต่อหมวดหมู่มาคิดวิเคราะห์

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 ที่มีคะแนนการรีวิวเฉลี่ยต่ำมากที่สุดคือ 0.965 คะแนน และ 1.751 คะแนน ตามลำดับ เนื่องจากความเป็นจริงสถานที่ที่ผู้คนส่วนใหญ่มักจะไปกันจะเป็นที่ที่เกี่ยวเนื่องกับธรรมชาติ อาจเป็นเพราะความลึกลับและความสวยงามของสถานที่ที่เกี่ยวทางธรรมชาติ ที่มีความน่าดึงดูดผู้คนมากกว่าสถานที่ที่ถูกสร้างขึ้นมาเองโดยมนุษย์ ซึ่งอาจมีข้อดีได้หลายปัจจัย อาทิ ได้รู้จักการวางแผน ได้ฝึกไหวพริบและรู้จักการแก้ปัญหาเฉพาะหน้า ได้ใกล้ชิดกับธรรมชาติและเข้าใจวิถีชีวิต ส่วนสถานที่ที่เี่ยวรองลงมาคงหนีไม่พ้นกับ พวกประเภทห้างสรรพสินค้าต่าง ๆ เนื่องจากผู้คนต่างก็มีความอยากได้สิ่งของที่จำเป็น อาทิ ยารักษาโรค และของอุปโภคบริโภค รวมถึงสินค้าฟุ่มเฟือยต่าง ๆ หรืออาจเป็นการผ่อนคลายทางด้านร่างกายจิตใจ และอารมณ์ เช่น การเข้าฟิตเนส การดูภาพยนตร์ การเล่นดนตรี เป็นต้น และประเภทสุดท้ายจากกราฟคือ สถานที่ที่พักผ่อนหย่อนใจ อาทิ ร้านนวดแผนไทย ร้านสปา และคาเฟ่ต่าง ๆ เพราะมนุษย์นั้นมีความความรักสวย รักงามและมีเส้นทางที่จะเดินในแบบของตัวเอง

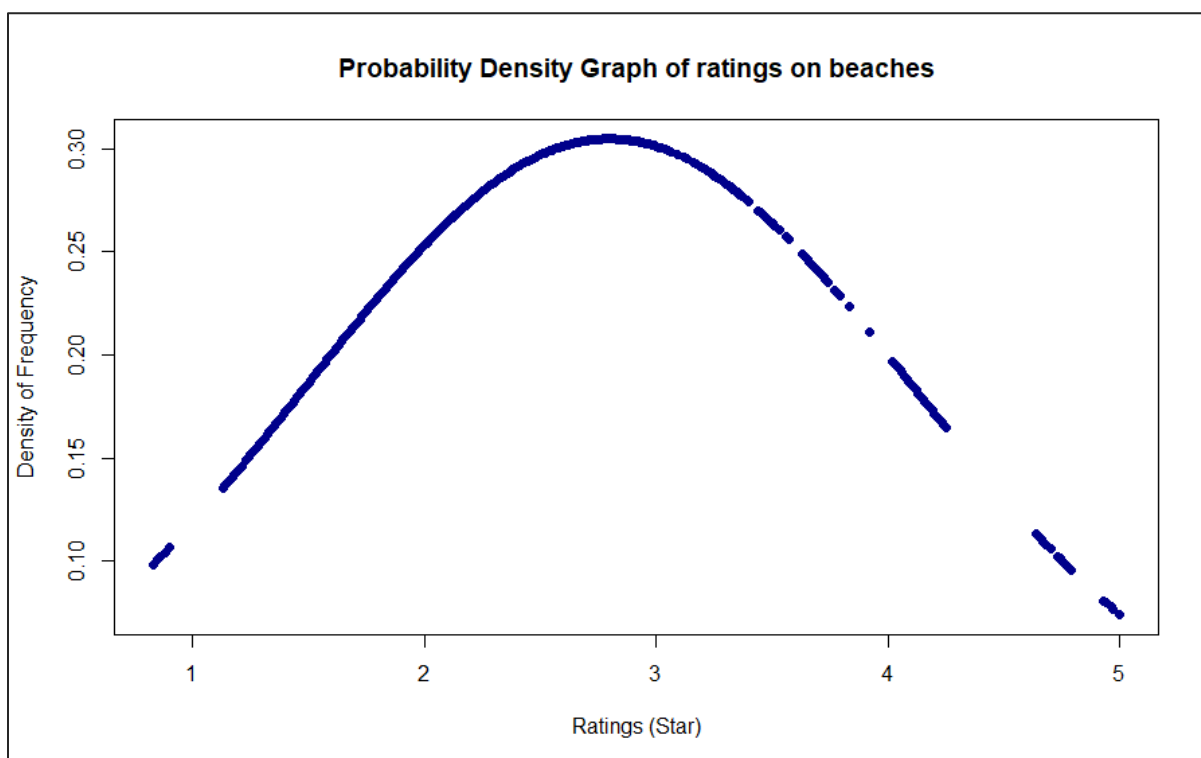
PROBABILITY AND STATISTICS HW 3

สืบเนื่องมาจาก HW 2 ที่ทางผู้จัดทำได้หยิบยกชุดข้อมูลที่เกี่ยวข้องกับ 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 เพื่อหาความสนใจและความต้องการของผู้ที่มา รีวิวสถานที่ของแต่ละประเภทได้ง่ายขึ้น และโดย HW 3 นี้ จะแสดงถึง Probability Density Graph และ Cumulative Distribution Graph ของแต่ละสถานที่เพื่อให้เห็นการกระจายตัวของข้อมูลให้ชัดเจนมากขึ้น

Average ratings on beaches

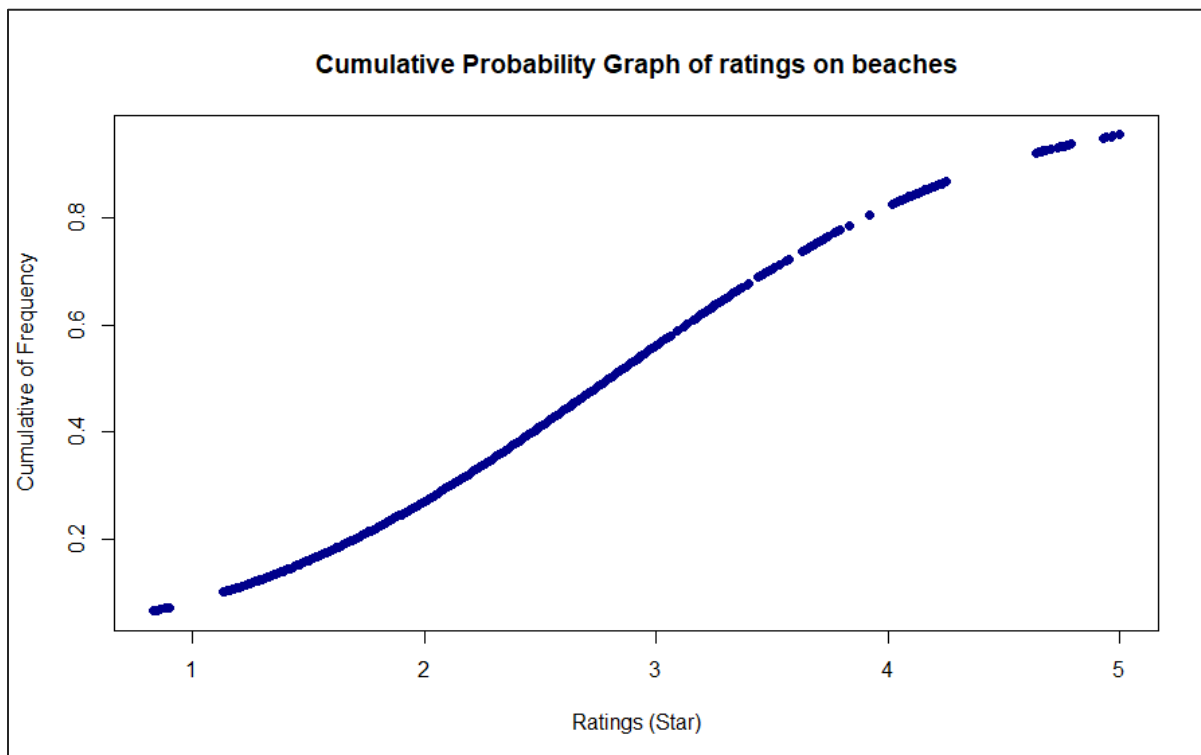
```
> mean(Google_Review_Ratings_$'Average ratings on beaches')
[1] 2.796886
> sd(Google_Review_Ratings_$'Average ratings on beaches')
[1] 1.309159
```

```
> DENS <- dnorm(Google_Review_Ratings_$'Average ratings on beaches', mean=2.796886, sd=1.309159)
> plot(Google_Review_Ratings_$'Average ratings on beaches', DENS, type = "p", col="dark blue", pch=19, main = "Probability Density Graph of ratings on beaches", xlab="Ratings (Star)", ylab="Density of Frequency")
```



Probability Density Graph of ratings on beaches

```
> CUMS <- pnorm(Google_Review_Ratings_$'Average ratings on beaches', mean=2.796886, sd=1.309159)
> plot(Google_Review_Ratings_$'Average ratings on beaches',CUMS, type = "p",col="dark blue",pch=19 , main = "Cumulative Probability Graph of ratings on beaches", xlab="Ratings (Star)", ylab="Cumulative of Frequency")
```



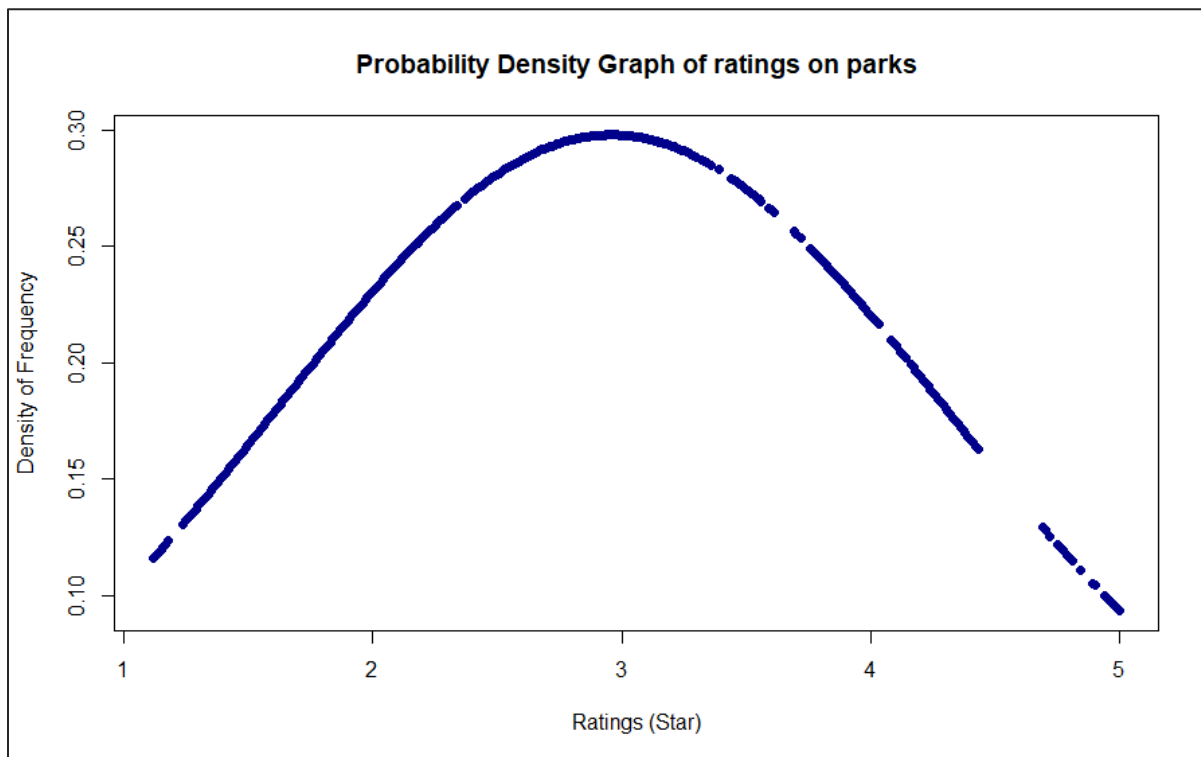
Cumulative Distribution Graph of ratings on beaches

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

```
> mean(Google_Review_Ratings_.$'Average ratings on parks')
[1] 2.958941
> sd(Google_Review_Ratings_.$'Average ratings on parks')
[1] 1.339056
```

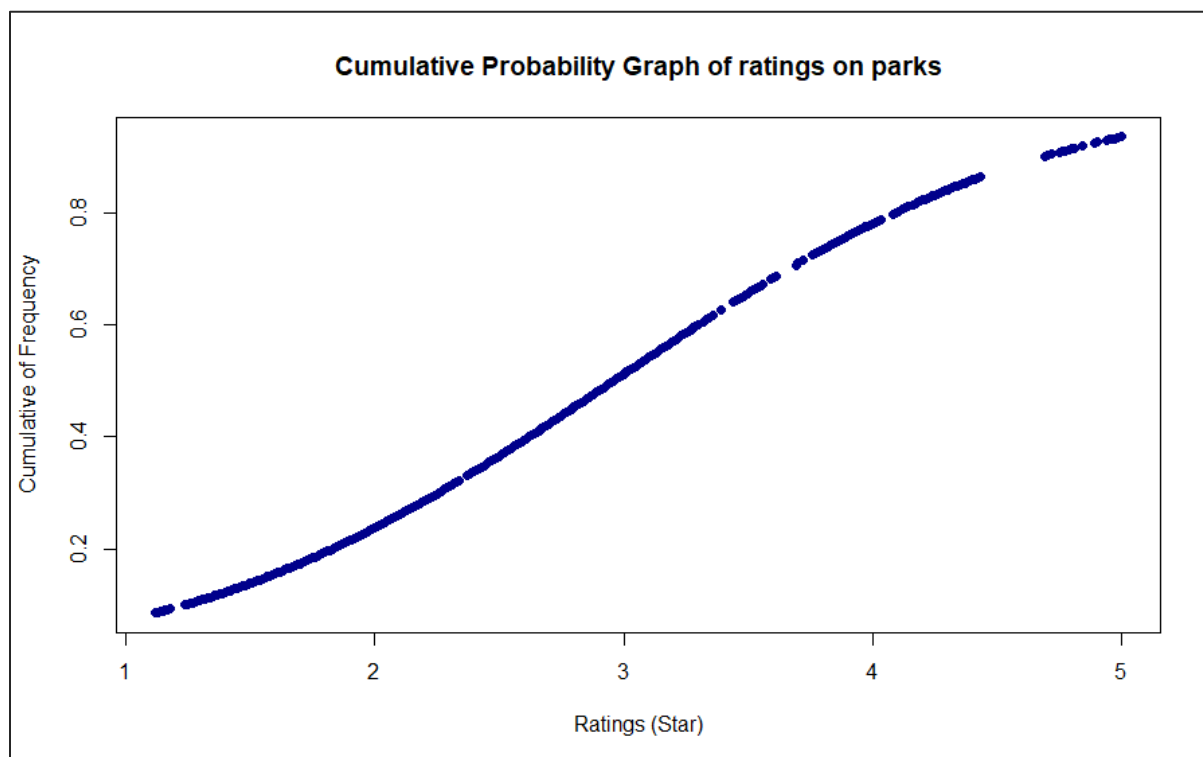
```
> DENS <- dnorm(Google_Review_Ratings_.$'Average ratings on parks', mean=2.958941, sd=1.339056)
> plot(Google_Review_Ratings_.$'Average ratings on parks',DENS, type = "p",col="dark blue",pch=19 , main = "Probability Density Graph of ratings on parks", xlab="Ratings (Star)", ylab="Density of Frequency")
```



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 เป็น คะแนนเต็ม 5 ดาว แกน y เป็นค่าการสะสมที่สอดคล้องกับค่า mean และ sd
3. จะได้กราฟของ Probability Density Graph of ratings on parks มีหน่วยเป็น Star


```
> CUMS <- pnorm(Google_Review_Ratings_$'Average ratings on parks', mean=2.958941, sd=1.339056)
> plot(Google_Review_Ratings_$'Average ratings on parks',CUMS, type = "p",col="dark blue",pch=19 , main = "Cumulative Probability Graph of ratings on parks", xlab="Ratings (Star)", ylab="Cumulative of Frequency")
```



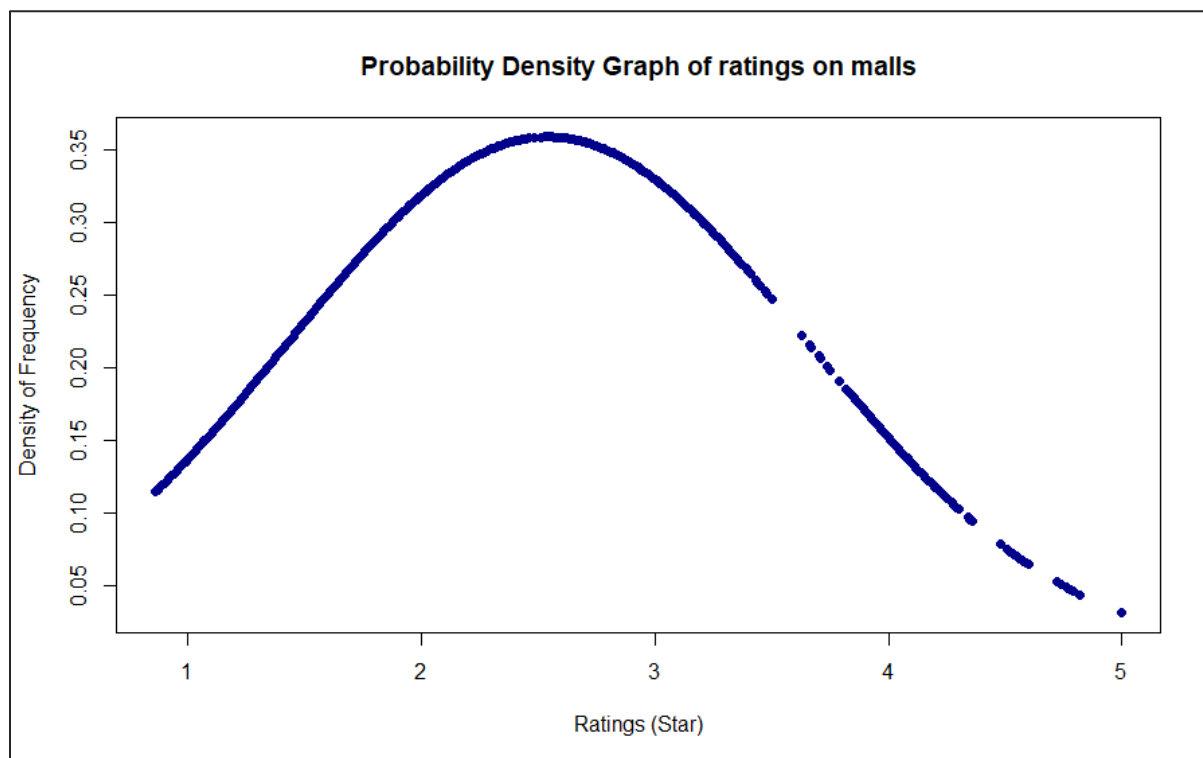
Cumulative Distribution Graph of ratings on parks

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

```
> mean(Google_Review_Ratings_.$'Average ratings on malls')
[1] 2.540795
> sd(Google_Review_Ratings_.$'Average ratings on malls')
[1] 1.111391
```

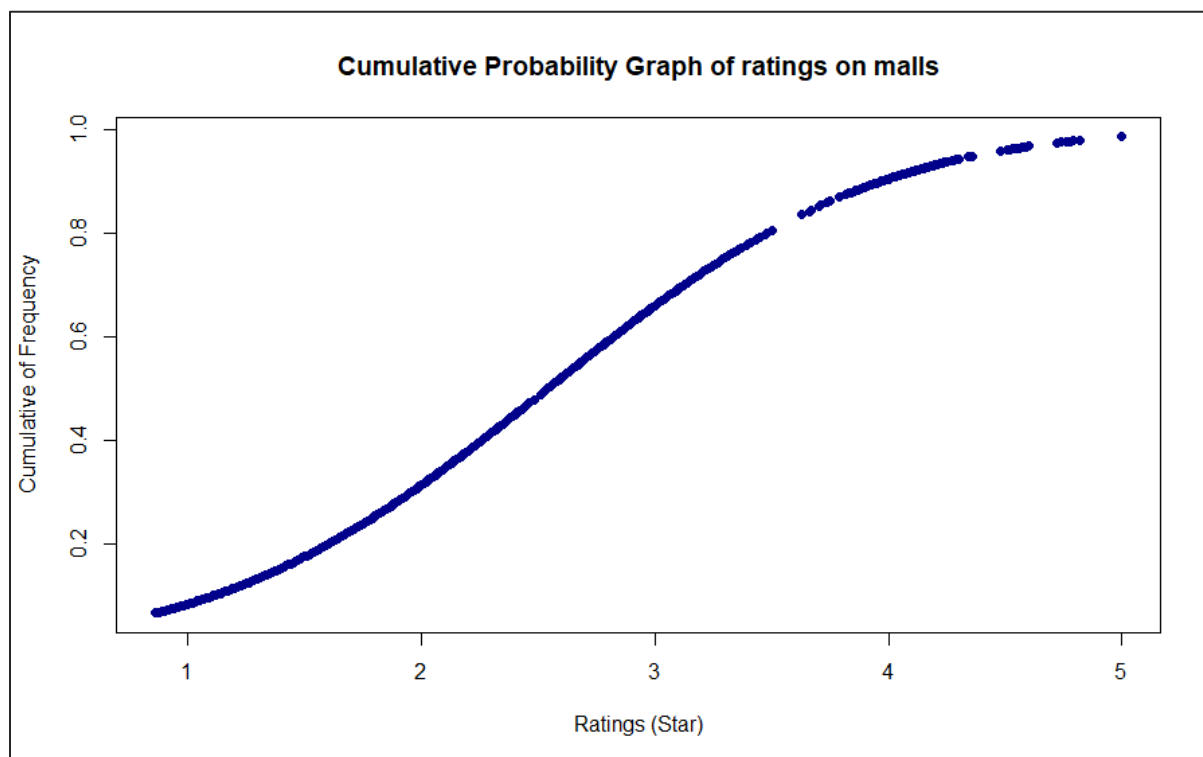
```
> DENS <- dnorm(Google_Review_Ratings_.$'Average ratings on malls', mean=2.540795, sd=1.111391)
> plot(Google_Review_Ratings_.$'Average ratings on malls', DENS, type = "p", col="dark blue", pch=19, main = "Probability Density Graph of ratings on malls", xlab="Ratings (Star)", ylab="Density of Frequency")
```



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 เป็น คะแนนเต็ม 5 ดาว แกน y เป็นค่าการสะสมที่สอดคล้องกับค่า mean และ sd
3. จะได้กราฟของ Probability Density Graph of ratings on malls มีหน่วยเป็น Star

```
> CUMS <- pnorm(Google_Review_Ratings_$'Average ratings on malls', mean=2.540795, sd=1.111391)
> plot(Google_Review_Ratings_$'Average ratings on malls',CUMS, type = "p",col="dark blue",pch=19 , main = "Cumulative Probability Graph of ratings on malls", xlab="Ratings (Star)", ylab="Cumulative of Frequency")
```



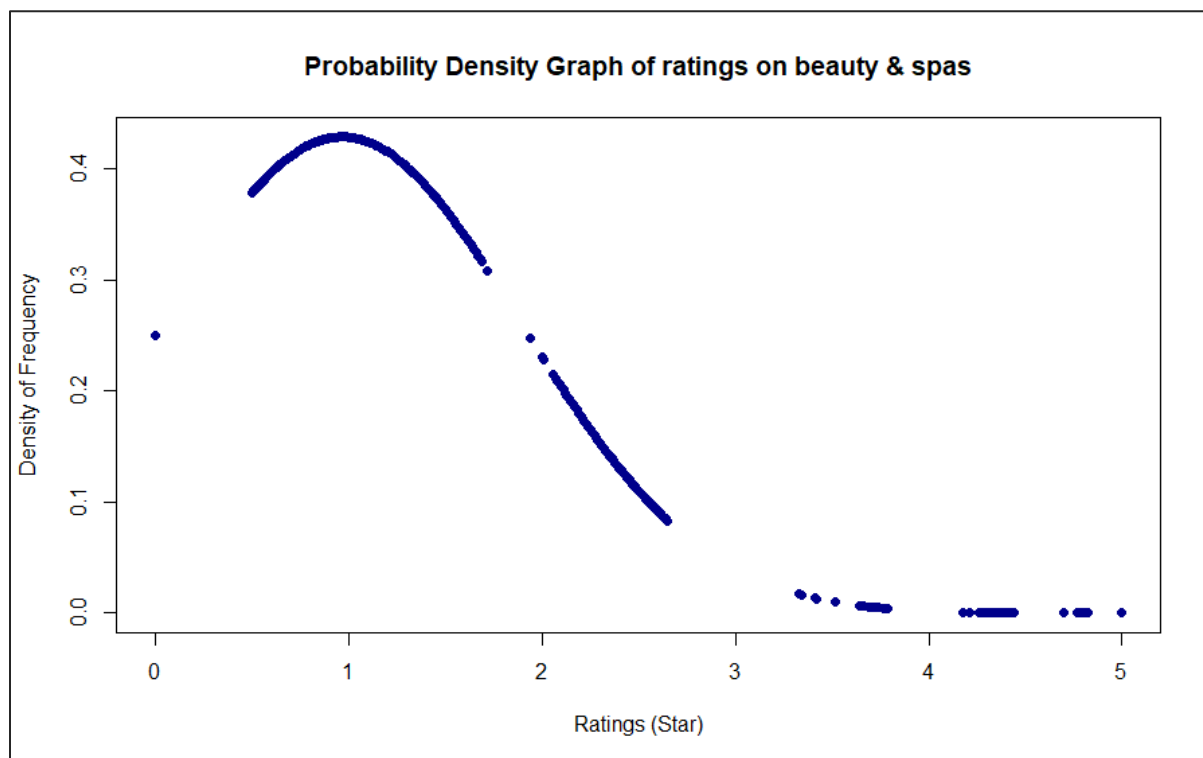
Cumulative Distribution Graph of ratings on malls

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

```
> mean(Google_Review_Ratings_.$'Average ratings on beauty & spas')
[1] 0.9658376
> sd(Google_Review_Ratings_.$'Average ratings on beauty & spas')
[1] 0.9298533
```

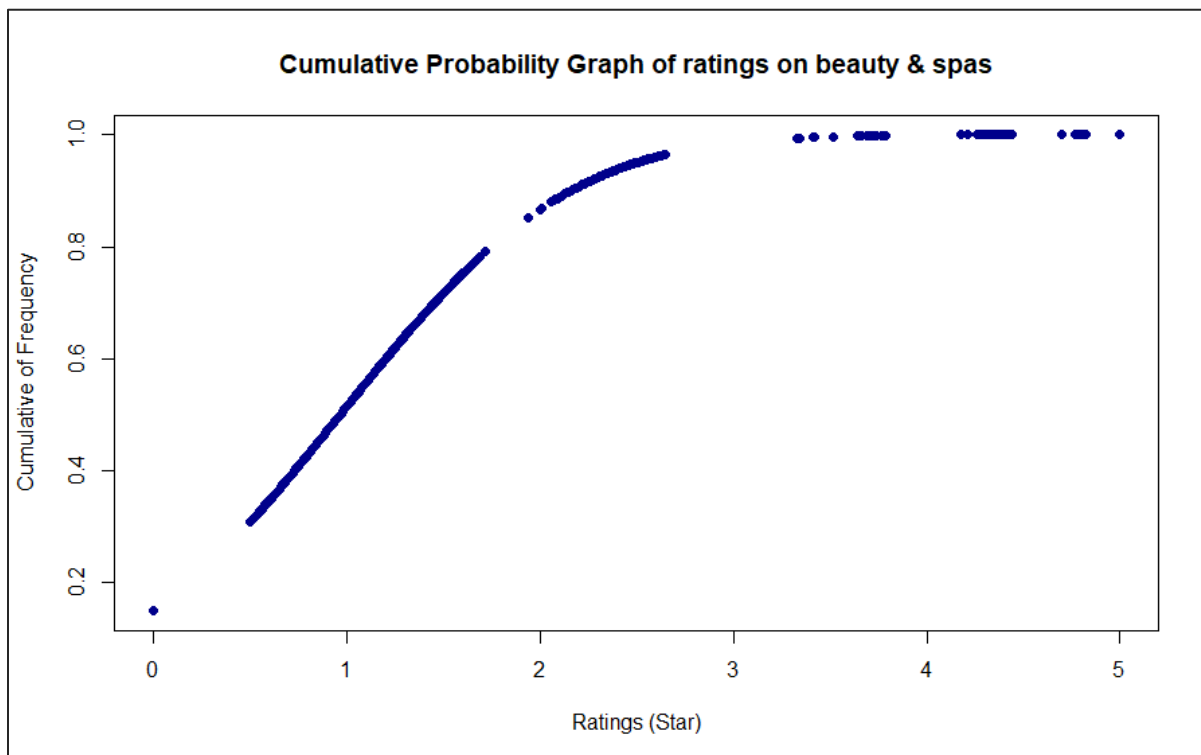
```
> DENS <- dnorm(Google_Review_Ratings_.$'Average ratings on beauty & spas', mean= 0.9658376, sd=0.9298533)
> plot(Google_Review_Ratings_.$'Average ratings on beauty & spas', DENS, type = "p", col="dark blue", pch=19 , main = "Probability Density Graph of ratings on beauty & spas", xlab="Ratings (Star)", ylab="Density of Frequency")
```



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 เป็น คะแนนเต็ม 5 ดาว แกน y เป็นค่าการสะสมที่สอดคล้องกับค่า mean และ sd
3. จะได้กราฟของ Probability Density Graph of ratings on beauty & spas มีหน่วยเป็น Star

```
> CUMS <- pnorm(Google_Review_Ratings_$'Average ratings on beauty & spas', mean= 0.9658376, sd=0.9298533)
> plot(Google_Review_Ratings_$'Average ratings on beauty & spas',CUMS, type = "p",col="dark blue",pch=19 , main = "Cumulative Probability Graph of ratings on beauty & spas", xlab="Ratings (Star)", ylab="Cumulative of Frequency")
```



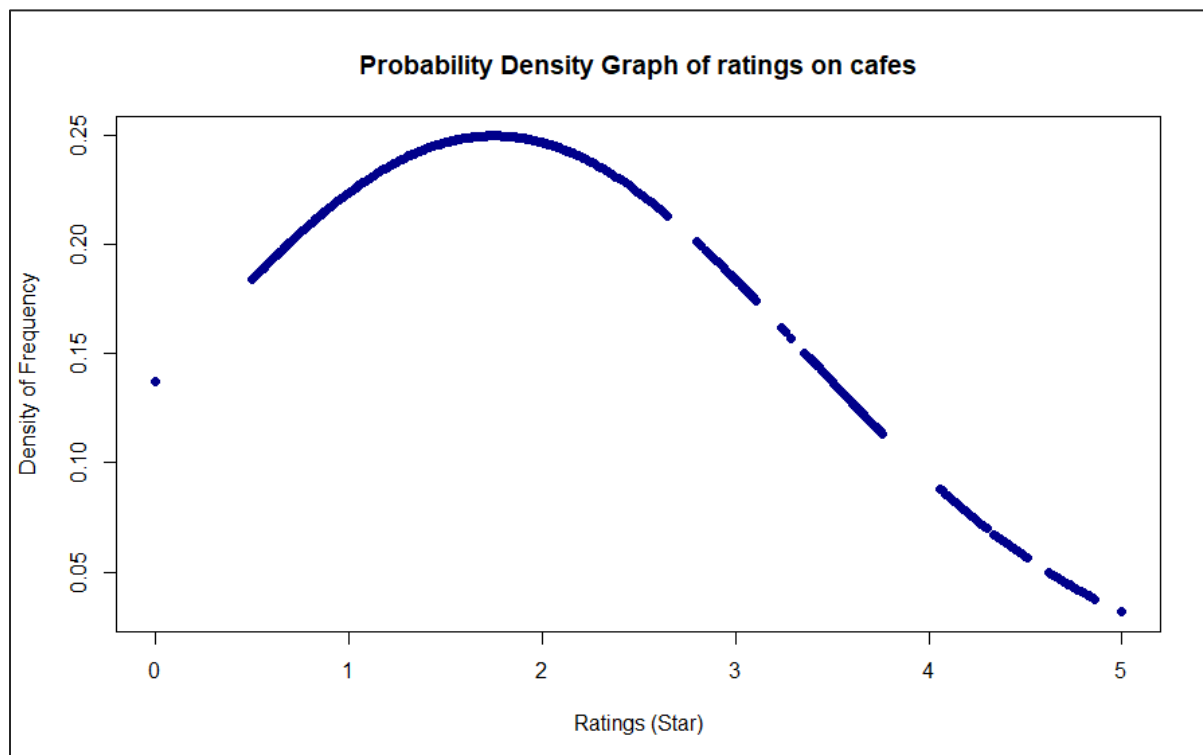
Cumulative Distribution Graph of ratings on beauty & spas

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

```
> mean(Google_Review_Ratings_$'Average ratings on cafes')
[1] 1.750537
> sd(Google_Review_Ratings_$'Average ratings on cafes')
[1] 1.598734
```

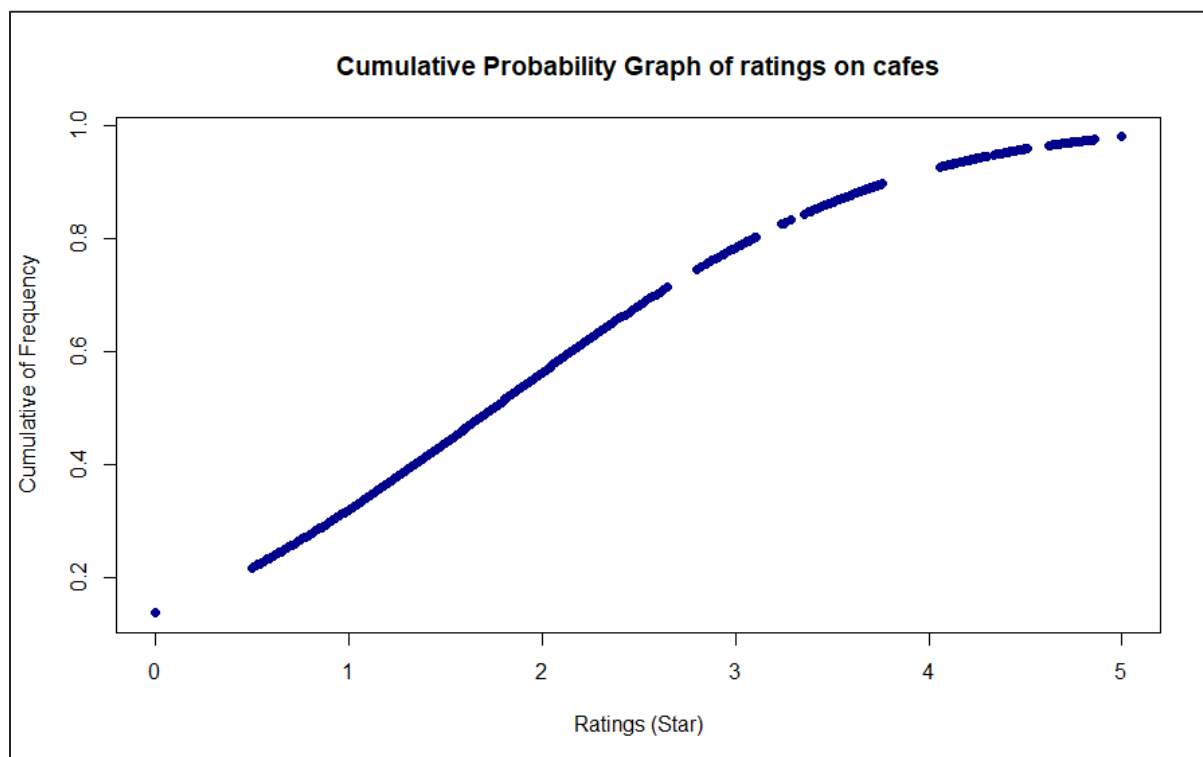
```
> DENS <- dnorm(Google_Review_Ratings_$'Average ratings on cafes', mean= 1.750537, sd= 1.598734)
> plot(Google_Review_Ratings_$'Average ratings on cafes', DENS, type = "p", col="dark blue", pch=19, main = "Probability Density Graph of ratings on cafes", xlab="Ratings (Star)", ylab="Density of Frequency")
```



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

```
> CUM5 <- pnorm(Google_Review_Ratings_$'Average ratings on cafes', mean= 1.750537, sd= 1.598734)
> plot(Google_Review_Ratings_$'Average ratings on cafes',CUM5, type = "p",col="dark blue",pch=19 , main = "Cumulative Probability Graph of ratings on cafes", xlab="Ratings (star)", ylab="Cumulative of Frequency")
```



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 ทางผู้จัดทำได้หียบกคะแนนผู้ใช้ Google มีตั้งแต่ 1 ถึง 5 และมีการคำนวณคะแนนเฉลี่ยต่อหมวดหมู่มาคิดวิเคราะห์

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 ข้อมูลจากแนวโน้มความหนาแน่นหรือการสะสมของข้อมูลที่น่าสนใจได้

PROBABILITY AND STATISTICS HW 4

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

```
> sample.mean <- mean(Google_Review_Ratings_.$'Average ratings on parks')
> print(sample.mean)
[1] 2.958941
```

```
> sample.n <- length(Google_Review_Ratings_.$'Average ratings on parks')
> sample.sd <- sd(Google_Review_Ratings_.$'Average ratings on parks')
> sample.se <- sample.sd/ sqrt(sample.n)
> print(sample.se)
[1] 0.01812849
```

sample.mean	2.95894061583578
sample.n	54561
sample.sd	1.33905646061777
sample.se	0.0181284932154164

จากรูปจะเห็นค่าของ Mean, Sample Size, Standard Deviation และ Standard error

Formula >

$$CI = \bar{x} \pm z \frac{s}{\sqrt{n}}$$

CI = confidence interval

\bar{x} = sample mean

z = confidence level value

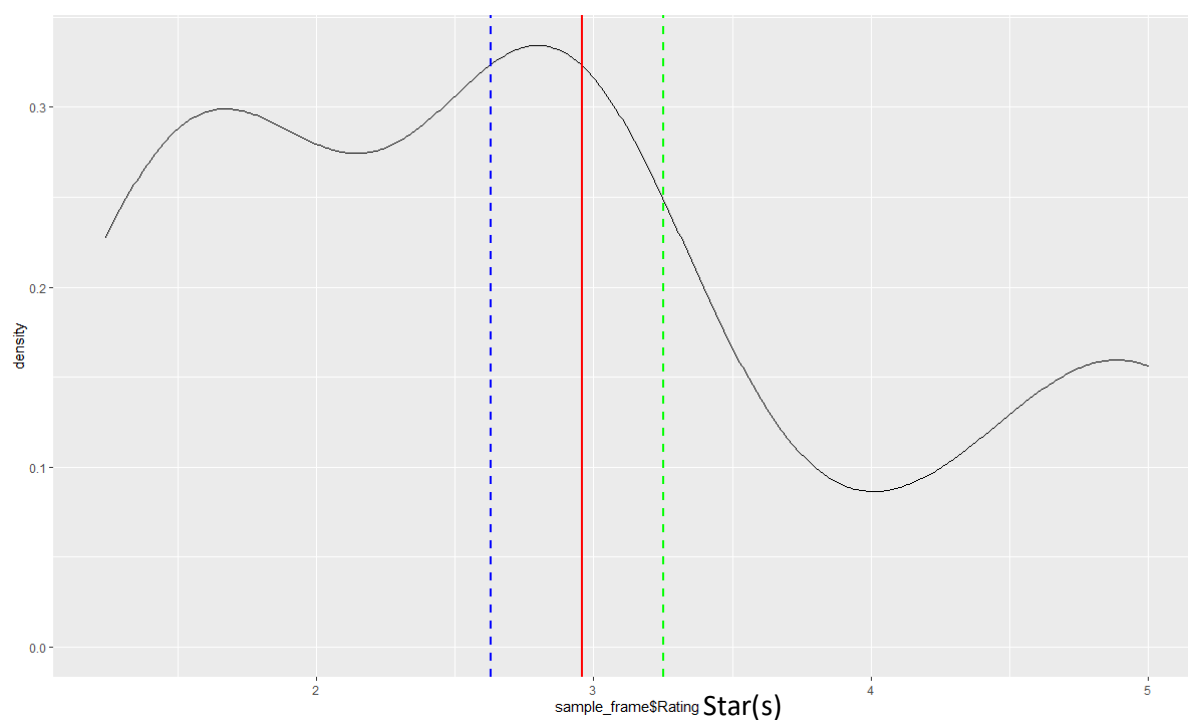
s = sample standard deviation

n = sample size

นี่คือสูตรในการใช้คำนวณหาค่า Confidence Interval โดยในที่นี้จะใช้จากเว็บไซต์

[Easy Confidence Interval Calculator \(socscistatistics.com\)](https://www.socscistatistics.com/easy-confidence-interval-calculator/)

ค่า Confidence Interval ที่ Confidence Level 90%



$M = 2.958941$

$Z = 1.64$

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%

```

1 library("ggplot2")
2 lower_point <- numeric()
3 upper_point <- numeric()
4 for (i in 1:50) {
5   sample_data <- sample(Google_Review_Ratings_.$`Average ratings on parks`,size = 50)
6   sample_frame <- data.frame(Rating = sample_data)
7
8   z <- 1.64
9   n <- 50
10  sigma <- sd(Google_Review_Ratings_.$`Average ratings on parks`)
11  x_bar <- mean(sample_frame$Rating)
12
13  ci_low <- x_bar - (z*sigma/sqrt(n))
14  lower_point <- c(lower_point, ci_low)
15  ci_up <- x_bar + (z*sigma/sqrt(n))
16  upper_point <- c(upper_point, ci_up)
17 }
18 print("Lower : ")
19 print(lower_point)
20 print("Upper : ")
21 print(upper_point)
22
23 real_mean <- mean(Google_Review_Ratings_.$`Average ratings on parks`)
24 print("Lower mean is : ")
25 lower_mean <- mean(lower_point)
26 print(lower_mean)
27 print("Upper mean is : ")
28 upper_mean <- mean(upper_point)
29 print(upper_mean)
30
31 ggplot(sample_frame,aes(sample_frame$Rating))+geom_density()+geom_vline(aes(xintercept = lower_mean),
32 color = "blue", linetype="dashed", size=1)+geom_vline(aes(xintercept=upper_mean), color="green",linetype="dashed", size=1)+
33 geom_vline(aes(xintercept=real_mean),color="blue",size=1)

```

```

[1] "Lower : "
[1] 2.824631 2.683231 2.662831 2.724031 2.577431 2.437631 2.335231 2.789231
2.621431 2.558431 2.661031 2.803831 2.768831 2.440231 2.844431 2.301631
[17] 2.591231 2.361631 2.886631 2.751631 2.689431 2.832431 2.689631 2.601431
2.454431 2.668031 2.556431 2.836631 2.788231 2.341031 2.461231 2.758831
[33] 2.743031 2.854231 2.682431 2.440831 2.496231 2.548231 2.716431 2.817231
2.558831 2.628431 2.482831 2.554831 2.656231 2.668831 3.042031 2.526631
[49] 2.716631 2.405431
[1] "Upper : "
[1] 3.445769 3.304369 3.283969 3.345169 3.198569 3.058769 2.956369 3.410369
3.242569 3.179569 3.282169 3.424969 3.389969 3.061369 3.465569 2.922769
[17] 3.212369 2.982769 3.507769 3.372769 3.310569 3.453569 3.310769 3.222569
3.075569 3.289169 3.177569 3.457769 3.409369 2.962169 3.082369 3.379969
[33] 3.364169 3.475369 3.303569 3.061969 3.117369 3.169369 3.337569 3.438369
3.179969 3.249569 3.103969 3.175969 3.277369 3.289969 3.663169 3.147769
[49] 3.337769 3.026569
[1] "Lower mean is : "
[1] 2.636847
[1] "Upper mean is : "
[1] 3.257985

```

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

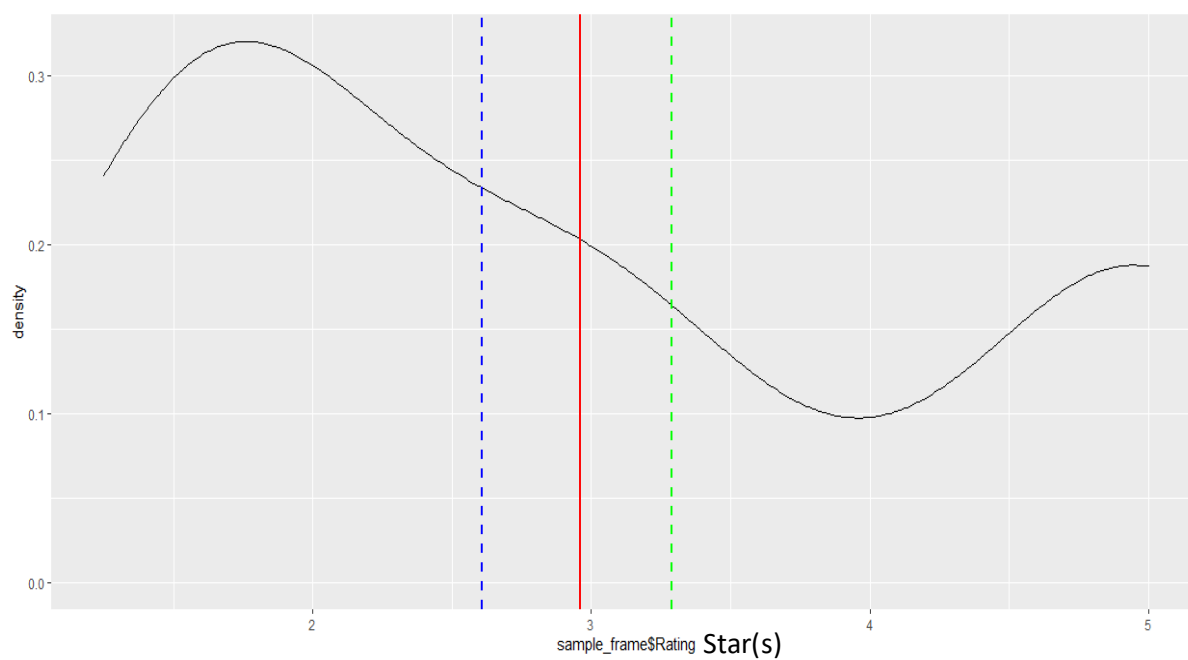
Lower คือ ข้อมูล Lower CI แต่ละชุดข้อมูลที่สุ่มออกมา

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

ค่า Confidence Interval ที่ Confidence Level 95%



$M = 2.958941$

$Z = 1.96$

ค่า Confidence Interval ของคะแนนการรีวิว คอลัมน์ Average ratings on parks

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

จากกราฟที่ขยายจะเห็นว่าช่วง *Confidence Level 95%*

คือระหว่างช่วง $[2.594236, 3.336572]$ คะแนน

สูตรที่ใช้หา Confidence Level 95%

```

1 library("ggplot2")
2 lower_point <- numeric()
3 upper_point <- numeric()
4 for (i in 1:50) {
5   sample_data <- sample(Google_Review_Ratings$`Average ratings on parks`,size = 50)
6   sample_frame <- data.frame(Rating = sample_data)
7
8   z <- 1.96
9   n <- 50
10  sigma <- sd(Google_Review_Ratings$`Average ratings on parks`)
11  x_bar <- mean(sample_frame$Rating)
12
13  ci_low <- x_bar - (z*sigma/sqrt(n))
14  lower_point <- c(lower_point, ci_low)
15  ci_up <- x_bar + (z*sigma/sqrt(n))
16  upper_point <- c(upper_point, ci_up)
17 }
18 print("Lower : ")
19 print(lower_point)
20 print("Upper : ")
21 print(upper_point)
22
23 real_mean <- mean(Google_Review_Ratings$`Average ratings on parks`)
24 print("Lower mean is : ")
25 lower_mean <- mean(lower_point)
26 print(lower_mean)
27 print("Upper mean is : ")
28 upper_mean <- mean(upper_point)
29 print(upper_mean)
30
31 ggplot(sample_frame,aes(sample_frame$Rating))+geom_density()+geom_vline(aes(xintercept = lower_mean),
32 color = "blue", linetype = "dashed", size=1)+geom_vline(aes(xintercept=upper_mean), color="green",linetype="dashed", size=1)+
33 geom_vline(aes(xintercept=real_mean),color="red",size=1)
34

```

```

[1] "Lower : "
[1] 2.127232 2.624032 2.763232 2.282632 2.704032 2.785432 2.529832
[8] 2.285832 2.621432 2.803232 2.780032 2.023032 2.623432 2.485232
[15] 2.794432 2.584032 2.513632 2.687032 2.242832 2.903632 2.620232
[22] 2.696632 2.806032 2.245832 2.800832 2.546632 2.398432 2.786232
[29] 2.612832 2.519032 2.491832 2.496032 2.476232 2.731832 2.454632
[36] 2.853832 2.587032 2.679232 2.644032 2.846232 2.640032 2.767032
[43] 2.540432 2.816832 2.650832 2.778032 2.531832 2.411632 2.562832
[50] 2.554432
[1] "Upper : "
[1] 2.869568 3.366368 3.505568 3.024968 3.446368 3.527768 3.272168
[8] 3.028168 3.363768 3.545568 3.522368 2.765368 3.365768 3.227568
[15] 3.536768 3.326368 3.255968 3.429368 2.985168 3.645968 3.362568
[22] 3.438968 3.548368 2.988168 3.543168 3.288968 3.140768 3.528568
[29] 3.355168 3.261368 3.234168 3.238368 3.218568 3.474168 3.196968
[36] 3.596168 3.329368 3.421568 3.386368 3.588568 3.382368 3.509368
[43] 3.282768 3.559168 3.393168 3.520368 3.274168 3.153968 3.305168
[50] 3.296768
[1] "Lower mean is : "
[1] 2.594236
[1] "Upper mean is : "
[1] 3.336572

```

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

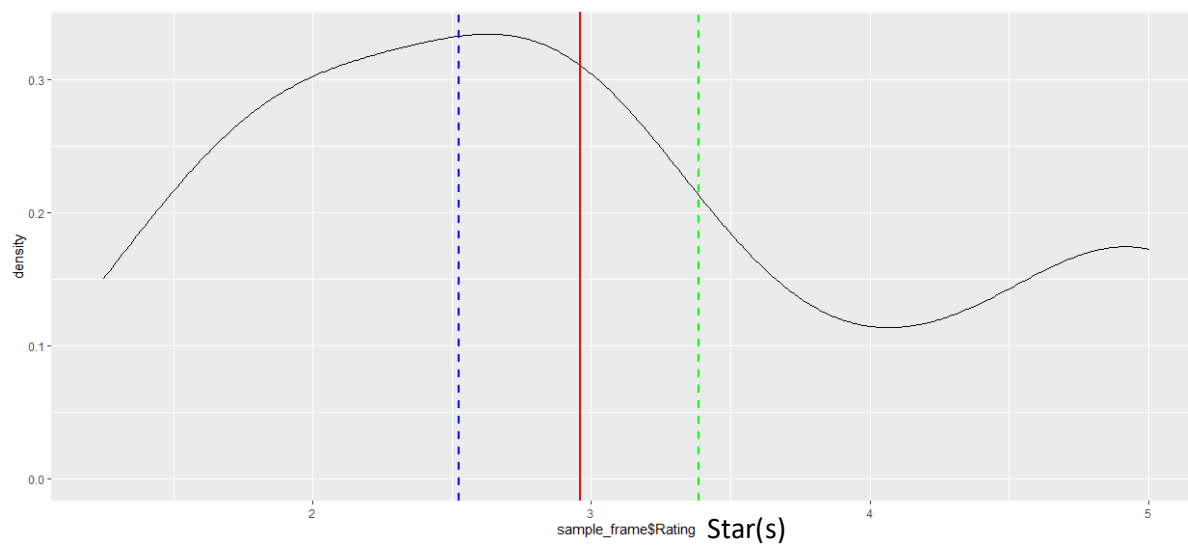
Lower คือ ข้อมูล Lower CI แต่ละชุดข้อมูลที่สุ่มออกมา

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

ค่า Confidence Interval ที่ Confidence Level 99%



$M = 2.958941$

$Z = 2.58$

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%

```

1 library("ggplot2")
2 lower_point <- numeric()
3 upper_point <- numeric()
4 for (i in 1:50) {
5   sample_data <- sample(Google_Review_Ratings$`Average ratings on parks`,size = 50)
6   sample_frame <- data.frame(Rating = sample_data)
7
8   z <- 2.58
9   n <- 50
10  sigma <- sd(Google_Review_Ratings$`Average ratings on parks`)
11  x_bar <- mean(sample_frame$Rating)
12
13  ci_low <- x_bar - (z*sigma/sqrt(n))
14  lower_point <- c(lower_point, ci_low)
15  ci_up <- x_bar + (z*sigma/sqrt(n))
16  upper_point <- c(upper_point, ci_up)
17 }
18 print("Lower : ")
19 print(lower_point)
20 print("Upper : ")
21 print(upper_point)
22
23 real_mean <- mean(Google_Review_Ratings$`Average ratings on parks`)
24 print("Lower mean is : ")
25 lower_mean <- mean(lower_point)
26 print(lower_mean)
27 print("Upper mean is : ")
28 upper_mean <- mean(upper_point)
29 print(upper_mean)
30
31 ggplot(sample_frame,aes(sample_frame$Rating))+geom_density()+geom_vline(aes(xintercept = lower_mean),
32 color = "blue", linetype = "dashed", size=1)+geom_vline(aes(xintercept=upper_mean), color="green",linetype="dashed", size=1)+
33 geom_vline(aes(xintercept=real_mean),color="red",size=1)

```

```

[1] "Lower : "
[1] 2.084622 2.569222 2.407222 2.495022 2.418222 2.547422 2.557822
[8] 2.282222 2.192422 2.541022 2.554022 2.299622 2.582422 2.410022
[15] 2.448622 2.161222 2.425222 2.325622 2.643022 2.427222 2.508822
[22] 2.392422 1.973422 2.614422 2.521822 2.478622 2.330222 2.530422
[29] 2.459222 2.628022 2.792422 2.424222 2.439822 2.388622 2.675422
[36] 2.305222 2.378822 2.686022 1.859622 2.474422 2.412422 2.455422
[43] 2.342622 2.424622 2.349222 2.664022 2.374222 2.510222 2.389022
[50] 2.602222
[1] "Upper : "
[1] 3.061778 3.546378 3.384378 3.472178 3.395378 3.524578 3.534978
[8] 3.259378 3.169578 3.518178 3.531178 3.276778 3.559578 3.387178
[15] 3.425778 3.138378 3.402378 3.302778 3.620178 3.404378 3.485978
[22] 3.369578 2.950578 3.591578 3.498978 3.455778 3.307378 3.507578
[29] 3.436378 3.605178 3.769578 3.401378 3.416978 3.365778 3.652578
[36] 3.282378 3.355978 3.663178 2.836778 3.451578 3.389578 3.432578
[43] 3.319778 3.401778 3.326378 3.641178 3.351378 3.487378 3.366178
[50] 3.579378
[1] "Lower mean is : "
[1] 2.435166
[1] "Upper mean is : "
[1] 3.412322

```

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

Lower คือ ข้อมูล LowerCI แต่ละชุดข้อมูลที่สุ่มออกมา

Upper คือ ข้อมูล UpperCI แต่ละชุดข้อมูลที่สุ่มออกมา

Lower mean คือ ค่าเฉลี่ยข้อมูลของ LowerCI ทั้งหมด มีค่า 2.435166

Upper mean คือ ค่าเฉลี่ยข้อมูลของ UpperCI ทั้งหมด มี 3.412322

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% นั้นมีการให้คะแนนเฉลี่ย
อยู่คือระหว่างช่วง [2.63687, 3.257985] คะแนน

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

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

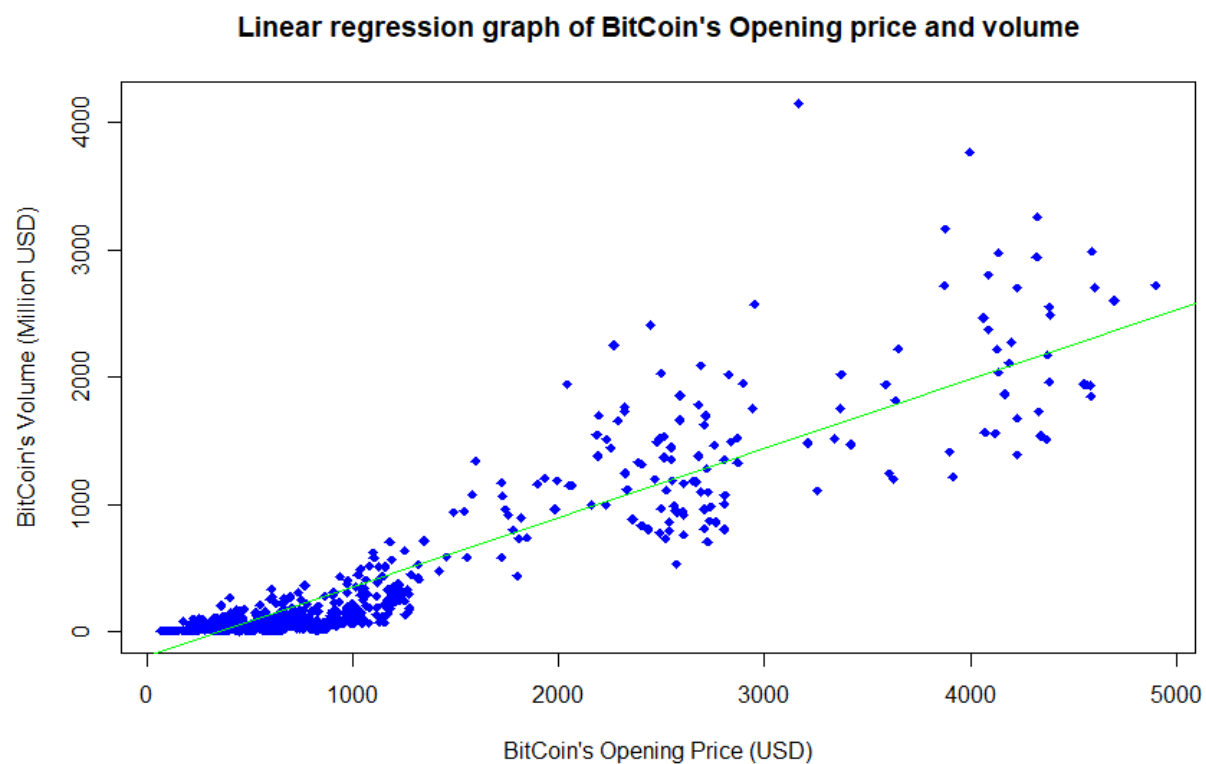
และสรุปจากคะแนนทั้งหมดได้ว่า

ค่าคะแนนการรีวิวเกี่ยวกับ Travel Review Rating Dataset (From the Machine Learning Repository of University of California) นั้นมีข้อสังเกต และข้อสงสัยที่ว่าคะแนนที่ได้มานั้น ค่าเฉลี่ยไป
กองรวมกันที่ 2.9 โดยประมาณ และในการโหวตคะแนนสังเกตว่า มีการให้คะแนนเต็ม (5 ดาว) กันหลาย
User ซึ่งบ่งบอกได้หลายกรณีเช่น อาจมีการรับจ้างรีวิว, อาจจะมีการใช้บอททำหน้าที่โหวตแทน, อาจจะมี
โดยให้คะแนนสูง, อาจจะมีรีวิวเพื่อแค่แลกของรางวัลที่ทางสถานที่นั้น ๆ จัดขึ้น, หรืออาจจะรีวิวจากการได้รับ
ส่วนลดค่าที่พัก เป็นต้น ซึ่งข้อสันนิษฐานของผู้จัดทำข้างต้นนี้มีความเป็นไปได้สูงที่จะสามารถเกิดขึ้น เพราะ
ด้วยสังคมโลกยุคปัจจุบันการเอาตัวรอดในสถานการณ์ต่าง ๆ เป็นสิ่งสำคัญ สถานที่หรือผู้ประกอบการบริษัท
นั้น ๆ สามารถล่อลวงผู้บริโภคเพื่อประโยชน์ส่วนตนได้ ดังเช่นตัวอย่างจากข่าวนี้ [ร้อง สคบ.เอาผิดเว็บจองที่พัก](#)
[ชื่อดัง หลอกหลวงผู้บริโภค | ข่าวช่อง 8 \(thaich8.com\)](#) ซึ่งทางผู้จัดทำหวังเป็นอย่างยิ่งว่าผู้ชมที่เข้ามาศึกษางานนี้
จะมีวิธีรับมือ และไม่หลงเชื่ออะไรง่าย ๆ ครับ

PROBABILITY AND STATISTICS HW 5

ทำ Linear Regression ทำคู่กับ นายศุภกฤต โล่ห์แก้ว 62010889

Graph



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

```
linmod = lm(Bitcoin_Copy$New_volume ~ Bitcoin_Copy$Open)
plot(Bitcoin_Copy$Open,
     Bitcoin_Copy$New_volume,
     main = "Linear regression graph of BitCoin's opening price and volume",
     xlab = "BitCoin's Opening Price (USD)",
     ylab = "BitCoin's volume (Million USD)",
     pch = 18,
     col = 'Blue',)
abline(linmod, col="green")
```

ส่วนของโปรแกรมที่ใช้ในการ Plot graph โดยใช้คำสั่ง lm ในการหาสมการที่ใช้คำนวณ Linear Regression

Coefficients

```
> linmod
```

Call:

```
lm(formula = Bitcoin_Copy$New_Volume ~ Bitcoin_Copy$Open)
```

Coefficients:

```
(Intercept) Bitcoin_Copy$Open
-185.4798      0.5429
```

การหา Linear Regression สามารถหาได้จากสมการ $y = mx + c$ ผ่านตัวแปร linmod

R-Square

```
> summary(lm(Bitcoin_Copy$New_Volume ~ Bitcoin_Copy$Open))$r.square
[1] 0.8383111
```

ส่วนของโปรแกรมในการหาค่า r-square หรือความคลาดเคลื่อนของข้อมูลว่าห่างจากเส้น Linear Regression ที่คำนวณไว้ซึ่ง
ยิ่งค่า r-square เข้าใกล้ 1 มากเท่าไรแสดงว่าข้อมูลมีความแม่นยำมากขึ้น

บทวิเคราะห์ข้อมูล

จากกราฟที่ได้ทำขึ้นเป็นกราฟความสัมพันธ์ของราคาเปิด และวอลุ่ม ของ Bitcoin ในแต่ละวัน โดยมี
เส้น linear regression ซึ่งมีสมการคือ $y = mx + c$ โดย m แสดงถึงความชันของกราฟ และ c แสดงถึง
จุดตัดแกน y โดยมี x, y เป็นตัวแปรตามแนวแกนของกราฟ ซึ่งก็คือ ราคาเปิดและวอลุ่ม โดยความชันของ
กราฟนี้ ได้แก่ค่า Coefficients ตามข้อมูลด้านบน และ ค่า c สามารถหาได้จากกราฟ ซึ่งเมื่อเรารู้ราคาเปิด
ของวันนั้น ก็จะคาดการณ์ได้ถึงวอลุ่มวันนั้นด้วย ส่วนค่า R-Square มีค่าอยู่ที่ 0.838 เมื่อคิดเป็นเปอร์เซ็นต์แล้ว
จะมีค่าเท่ากับ 83.8% แสดงให้เห็นว่า วอลุ่ม สามารถอธิบายได้ด้วยราคาเปิด โดยที่มีความแม่นยำ 83.8%