



รายงานเชิงวิชาการ  
เรื่อง อัตราส่วนผสม ที่ส่งผลต่อคุณภาพของไวน์แดง

เสนอ

ผศ.ดร.สุรินทร์ กิตติรักล

จัดทำโดย

นายวงศ์วริศ พันธ์เจริญ

รหัสนักศึกษา 62010787

นายสิริวิชญ์ สุขวัฒนาวิทย์

รหัสนักศึกษา 62010948

รายงานนี้เป็นส่วนหนึ่งของรายวิชา (01076253) PROBABILITY AND STATISTICS  
ภาคเรียนที่ 2/2563 ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์  
สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

## คำนำ

รายงานนี้เป็นส่วนหนึ่งของวิชา (01076253) PROBABILITY AND STATISTICS โดยมีจุดประสงค์เพื่อศึกษาเกี่ยวกับอัตราส่วนผสมต่าง ๆ ที่ส่งผลต่อคุณภาพของไวน์แดง

โดยจะนำหลักคำนวณทางสถิติ ได้แก่ การคำนวณค่าพื้นฐานทางสถิติ, ค่าเฉลี่ย, ค่าสูงสุด, ค่าต่ำสุด, ส่วนเบี่ยงเบนมาตรฐาน, ค่าความแปรปรวน, Histogram, Box Plot, Stem and Leaf, Scatter Plot, การคำนวณหา Outlier, Probability Density Function, Cumulative Probability Function, Confidence Interval (CI) of Mean และ Linear Regression นำมาวิเคราะห์ หาความหมายและความสัมพันธ์ของอัตราส่วนผสมต่างๆ ที่มีต่อคุณภาพของไวน์แดง

คณะกรรมการได้เลือกหัวข้อนี้ในการทำรายงานนี้ เนื่องจากมีความสนใจในเรื่องไวน์แดง จึงนำหลักวิธีการคิดทางสถิติที่ได้ศึกษาเล่าเรียน มาประยุกต์ใช้กับข้อมูลจริง เพื่อให้ทราบว่า ไวน์ที่มีลักษณะทางเคมีที่แตกต่างกัน จะทำให้มีคุณภาพที่แตกต่างกันอย่างไร และลักษณะแบบใด ที่จะทำให้ไวน์แดงที่มีคุณภาพที่ดีมากยิ่งขึ้น เหมาะสมสำหรับการนำไปใช้ตรวจสอบไวน์แดง และทำการผลิตไวน์แดงที่มีคุณภาพต่อไป

คณะกรรมการทำหัวข่าว่ารายงานเล่มนี้จะเป็นประโยชน์แก่ผู้อ่าน นักเรียนหรือนักศึกษาที่กำลังศึกษาเรื่องนี้อยู่ หากมีข้อเสนอแนะหรือข้อผิดพลาดประการใด ทางผู้จัดทำขออนุญาตไว้และขอภัยมา ณ ที่นี่

คณะกรรมการ

## สารบัญ

|  |           |
|--|-----------|
| คำนำ   | ก         |
| สารบัญ   | ข         |
| <b>บทที่ 1 ที่มาและความสำคัญ</b>                                 | <b>1</b>  |
| 1.1 ที่มาและความสำคัญ  | 1         |
| 1.2 วัตถุประสงค์   | 1         |
| 1.3 ขอบเขตของโครงการ   | 1         |
| 1.4 ประโยชน์ที่คาดว่าจะได้รับ                                    | 2         |
| <b>บทที่ 2 เอกสารที่เกี่ยวข้อง</b>                               | <b>3</b>  |
| 2.1 การคำนวณสถิติพื้นฐาน   | 3         |
| 2.2 การสร้างกราฟทางสถิติ   | 4         |
| <b>บทที่ 3 ขั้นตอนและวิธีการดำเนินงาน</b>                        | <b>11</b> |
| 3.1 คอลัมน์ที่ 1 Alcohol และ คอลัมน์ที่ 4 Quality                | 11        |
| 3.2 คอลัมน์ที่ 2 Fixed Acidity (ปริมาณกรด หน่วยคือ กรัม / ลิตร ) | 18        |
| 3.3 คอลัมน์ที่ 3 ชัลเฟอร์ไดออกไซด์                               | 23        |
| <b>บทที่ 4 ผลการดำเนินงาน</b>                                    | <b>28</b> |
| 4.1 คอลัมน์ที่ 1 Alcohol และ คอลัมน์ที่ 4 Quality                | 28        |
| 4.2 คอลัมน์ที่ 2 ปริมาณความเป็นกรด และ คอลัมน์ที่ Quality        | 35        |
| 4.3 คอลัมน์ที่ 3 ชัลเฟอร์ไดออกไซด์ และ คอลัมน์ที่ Quality        | 40        |
| <b>บทที่ 5 สรุปผลการดำเนินงาน</b>                                | <b>45</b> |
| 5.1 สรุปผลปริมาณแอลกอฮอล์ที่ส่งผลต่อคุณภาพของไวน์แดง             | 45        |
| 5.2 สรุปผลปริมาณของกรดที่ส่งผลต่อคุณภาพของไวน์แดง                | 46        |
| 5.3 สรุปผลปริมาณชัลเฟอร์ไดออกไซด์ที่ส่งผลต่อคุณภาพของไวน์แดง     | 47        |
| ข้อเสนอแนะแนวทางการศึกษาเพิ่มเติม                                | 48        |
| <b>บรรณานุกรม</b>  | <b>49</b> |
| <b>ภาคผนวก</b>   | <b>50</b> |

# บทที่ 1

## บทนำ

### 1.1 ที่มาและความสำคัญ

ไวน์ เป็นเครื่องดื่มประเภทหนึ่งที่ผลิตจากองุ่นหรือผลไม้อื่นๆ ใช้สำหรับงานสังสรรค์เพื่อความสนุกสนาน เพลิดเพลินในงานกิจกรรมหรือเนื่องในโอกาสพิเศษต่างๆ อีกทั้งยังมีวิธีการทำที่ค่อนข้างพิถีพิถัน ผ่านกรรมวิธีหลายขั้นตอนกว่าจะได้ไวน์ที่มีคุณภาพมาตรฐาน คณะกรรมการดูแลห้ามนำเข้าประเทศไทย ไม่ได้因为ไวน์ในเรื่องขององค์ประกอบต่างๆ ที่ทำให้คุณภาพของไวน์ดีขึ้น

คณะกรรมการดูแลห้ามนำเข้าประเทศไทย ไม่ได้因为ไวน์มีองค์ประกอบในการทำอยู่มากมาย โดยทางคณะกรรมการได้ทำการเลือกส่วนผสมที่น่าสนใจมาทั้งหมด 3 อย่าง ได้แก่ ปริมาณแอลกอฮอล์, ปริมาณกรด และ ปริมาณซัลเฟอร์ไดออกไซด์ในไวน์ เพื่อดูความสัมพันธ์ว่ามีผลต่อคุณภาพของไวน์แดงหรือไม่ หากน้อยเพียงใด เพื่อที่จะนำข้อมูลจากการวิเคราะห์ครั้งนี้ นำมาพัฒนาคุณภาพของไวน์แดงต่อไป

### 1.2 วัตถุประสงค์

- เพื่อศึกษาความสัมพันธ์ของส่วนผสมต่าง ๆ และคุณภาพของไวน์แดง ว่ามีความสัมพันธ์กันอย่างไร
- เพื่อศึกษาระบวนการคิดทางสถิติ และนำความรู้มาประยุกต์ใช้กับข้อมูลจริง
- เพื่อฝึกการคิดวิเคราะห์ สังเคราะห์ อ่านค่าทางสถิติ เพื่อนำมาใช้เป็นประโยชน์ในการทำให้คุณภาพของไวน์แดงดียิ่งขึ้น

### 1.3 ขอบเขตของโครงการ

ผู้ศึกษาเลือกที่จะเก็บข้อมูลจากการทดสอบด้านเคมีภysisของไวน์ (Physicochemical Test) ในโรงงานผลิตไวน์ จังหวัด Minho ทางตอนเหนือของประเทศโปรตุเกส โดยจะมีข้อมูลดังนี้

1.3.1) ชื่อชุดข้อมูล : Winequality-red.csv

1.3.2) ชื่อคอลัมน์ข้อมูลที่เลือกศึกษา :

1. Alcohol      2. Fixed Acidity      3. ซัลเฟอร์ไดออกไซด์      4. Quality

1.3.3) แหล่งที่มาของชุดข้อมูล :

ที่มาของชุดข้อมูล [Winequality-red.csv](#)

1.3.4) Reference / URL :

<https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009>

### 1.3.5) คำอธิบายชื่อคอลัมน์ข้อมูลที่เลือก :

#### 1) Alcohol

ปริมาณแอลกอฮอล์ที่บรรจุอยู่ในไวน์ มีหน่วยเป็น %/ปริมาตร (%/volume)

#### 2) Fixed Acidity

ปริมาณกรด ( acidity ) เป็นตัวแปรสำคัญที่จะส่งผลต่อรสชาติของไวน์ หรือหมายถึงค่าความเปรี้ยวและเมาด์ในไวน์ มีหน่วยเป็น กรัม / ลิตร

#### 3) ซัลเฟอร์ไดออกไซด์ ( $\text{SO}_2$ )

ปริมาณซัลเฟอร์ไดออกไซด์ในไวน์ ส่งผลต่อความ สดใหม่ และกลิ่นของไวน์ ซึ่งในทางกลับกัน อาจจะส่งผลเสียต่อร่างกายที่มีปริมาณมากจนเกินไป มีหน่วยเป็น กรัม / ลิตร

#### 4) Quality

คุณภาพของไวน์ ระดับ 1-10 おิจากโรงงานผลิตไวน์ จังหวัด Minho ทางตอนเหนือของประเทศโปรตุเกส

## 1.4 ประโยชน์ที่คาดว่าจะได้รับ

1. ได้ศึกษาความสัมพันธ์ของส่วนผสมต่าง ๆ และคุณภาพของไวน์แดง ว่ามีความสัมพันธ์กันอย่างไร
2. ได้ศึกษาระบบการคิดทางสถิติ และนำความรู้มาประยุกต์ใช้กับข้อมูลจริงได้
3. สามารถฝึกการคิดวิเคราะห์ สังเคราะห์ อ่านค่าทางสถิติ เพื่อนำมาใช้เป็นประโยชน์ในการทำให้คุณภาพของไวน์แดงดียิ่งขึ้น



## บทที่ 2

### เอกสารที่เกี่ยวข้อง

#### 2.1 การคำนวณสถิติพื้นฐาน

สิ่งที่จะใช้ในการคำนวณ ต่อข้อมูล 1 คอลัมน์ โดยทั้งหมดเป็นกลุ่มข้อมูลตัวอย่าง (Sample) มีดังนี้

##### 2.1.1) จำนวนของข้อมูล ( $n$ )

จำนวนข้อมูลทั้งหมด ต่อ 1 คอลัมน์

##### 2.1.2) ค่าต่ำสุด (Min)

ค่าที่ต่ำที่สุดของข้อมูลนั้น ๆ

##### 2.1.3) ค่าสูงสุด (Max)

ค่าที่สูงที่สุดของข้อมูลนั้น ๆ

##### 2.1.4) ค่าเฉลี่ย (Mean)

$$\text{Sample mean: } \bar{x} = \frac{\sum x}{n}$$

##### 2.1.5) ค่ามัธยฐาน (Median)

$n$  is odd,

**Median**  $\xrightarrow{\quad}$   $\text{Median} = \left(\frac{n+1}{2}\right)^{\text{th}} \text{ observation}$

$n$  is even,

$\text{Median} = \frac{\left(\frac{n}{2}\right)^{\text{th}} + \left(\frac{n}{2}+1\right)^{\text{th}} \text{ observation}}{2}$

##### 2.1.6) ค่าฐานนิยม (Mode)

ค่าที่มีจำนวนซ้ำกันมากที่สุด อาจมีได้มากกว่า 2 ค่า หรือ อาจไม่มีฐานนิยมก็ได้

##### 2.1.7) ส่วนเบี่ยงเบนมาตรฐาน (Standard deviation)

$$\text{Sample standard deviation: } s = \sqrt{\frac{\sum(x - \bar{x})^2}{n-1}}$$

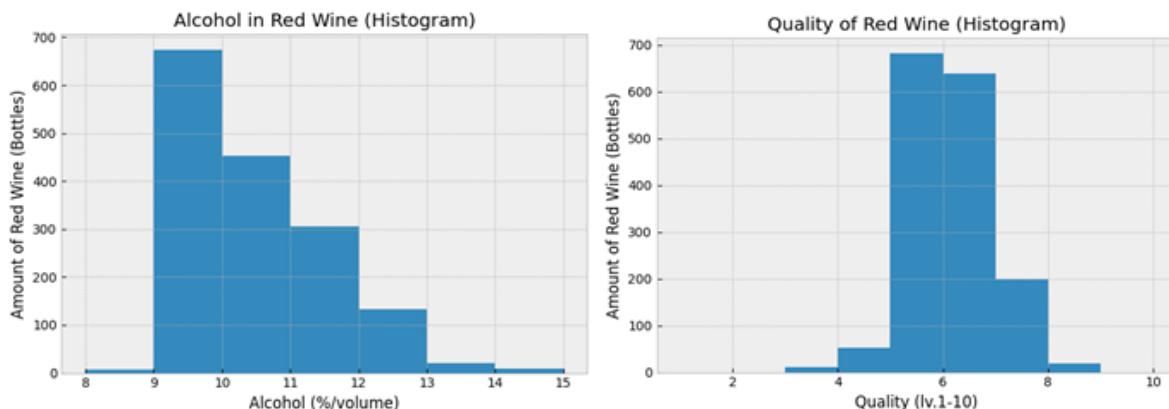
##### 2.1.8) ค่าความแปรปรวน (Variance)

$$\text{Sample variance: } s^2$$

## 2.2 การสร้างกราฟทางสถิติ

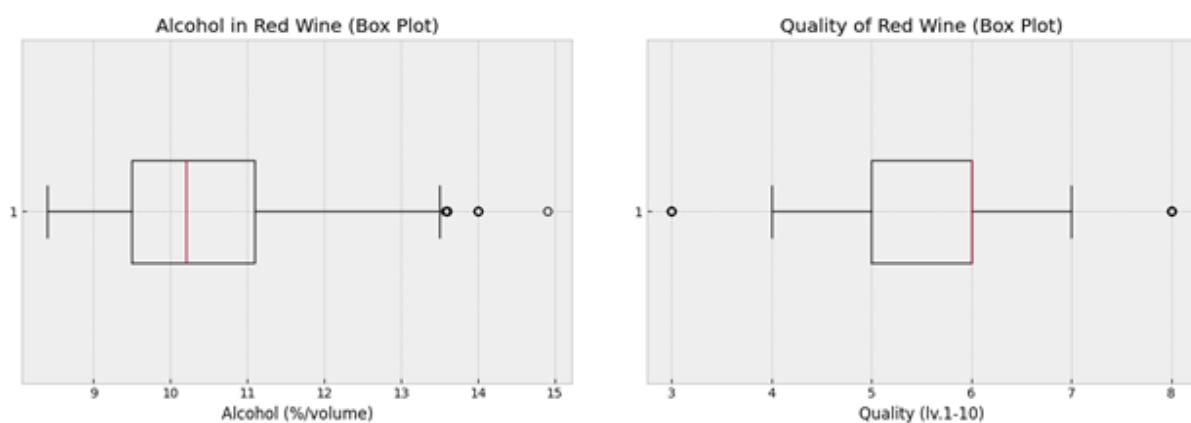
กราฟทางสถิติ มีหลายประเภท ดังนี้

### 2.2.1) Histogram



ภาพตัวอย่าง Histogram

### 2.2.2) Box Plot



ภาพตัวอย่าง Box Plot

### การคำนวณหา Outliers

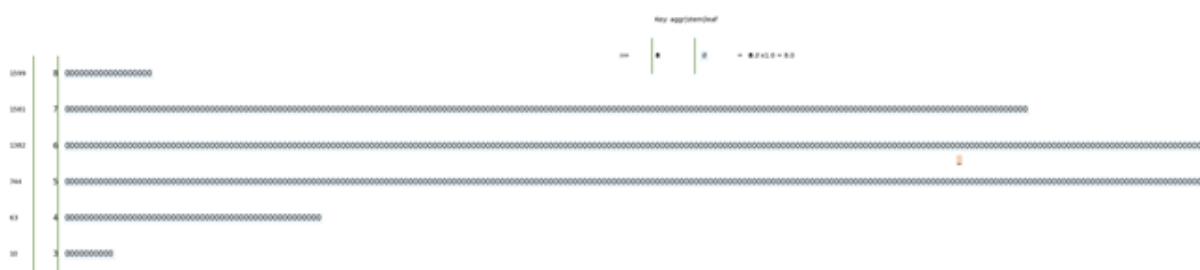
สามารถดูได้จากการ Box Plot หรือจากการคำนวณหา Outliers โดยใช้สูตรดังนี้

- 1)  $Q1 = (N+1) * 1 / 4$  (คิดแบบสูตรการไม่แจกแจงความถี่ / เห็นข้อมูลเป็นตัว ไม่ใช่ช่วง)
- 2)  $Q3 = (N+1) * 3 / 4$  (คิดแบบสูตรการไม่แจกแจงความถี่ / เห็นข้อมูลเป็นตัว ไม่ใช่ช่วง)
- 3)  $IQR = Q3 - Q1$  (คิดแบบ Inclusive เพื่อทำให้ IQR มีค่าแคบลงเพื่อตัด Outliers ได้มากขึ้น)
- 4) Extreme Outlier Lower Boundary =  $Q1 - IQR * 3$
- 5) Mild Outlier Lower Boundary =  $Q1 - IQR * 1.5$
- 6) Mild Outlier Upper Boundary =  $Q3 + IQR * 1.5$
- 7) Extreme Outlier Upper Boundary =  $Q3 + IQR * 3$
- 8) Extreme Outlier (Lower) = if (value < Ext. Outlier Lower B.)
- 9) Mild Outlier (Lower) = if (Ext. Outlier Lower B.  $\leq$  value < Mild Outlier Lower B.)
- 10) Mild Outlier (Upper) = if (Mild Outlier Upper B.  $<$  value  $\leq$  Ext. Outlier Upper B.)
- 11) Extreme Outlier (Upper) = if (value  $\geq$  Ext. Outlier Upper B.)

### 2.2.3) Stem and Leaf

## Alcohol in Red Wine Stem-And-Leaf

## Quality of Red Wine Stem-And-Leaf



## ภาพตัวอย่าง Stem and Leaf

#### 2.2.4) Scatter Plot

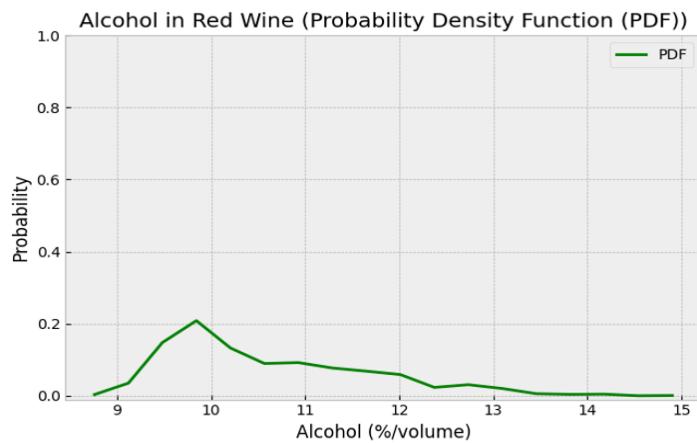
ความสัมพันธ์ระหว่างปริมาณแอลกอฮอล์และคุณภาพของไวน์แดง



ภาพตัวอย่าง Scatter Plot

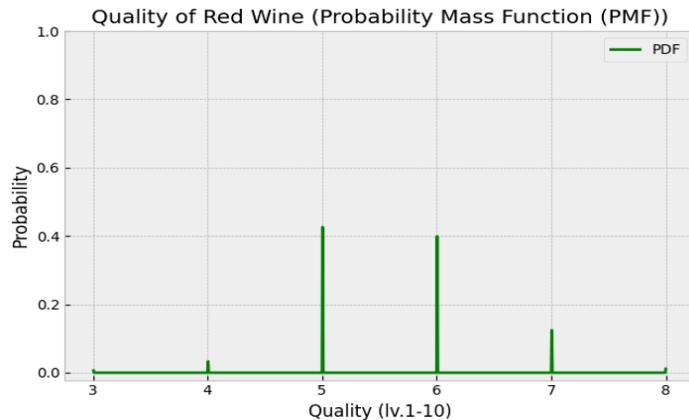
- กำหนดให้ แนวแกน x เป็นตัวแปรต้น (Independent Variable)
- กำหนดให้ แนวแกน y เป็นตัวแปรตาม (Dependent Variable)

## 2.2.5) Probability Density Function (PDF)



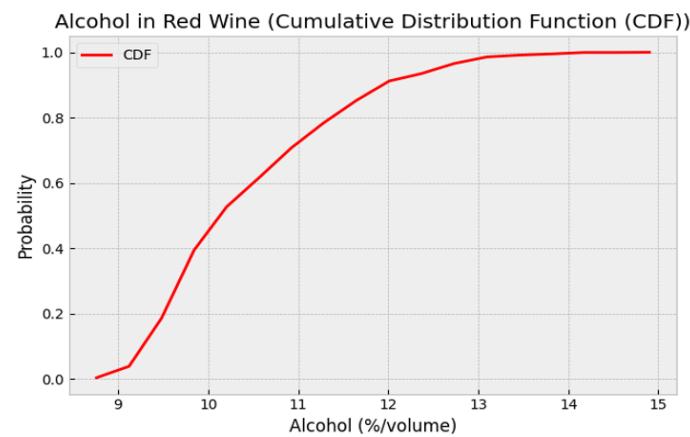
ກາພຕ້ວຍ່າງ Probability Density Function (PDF)

## 2.2.6) Probability Mass Function (PMF)



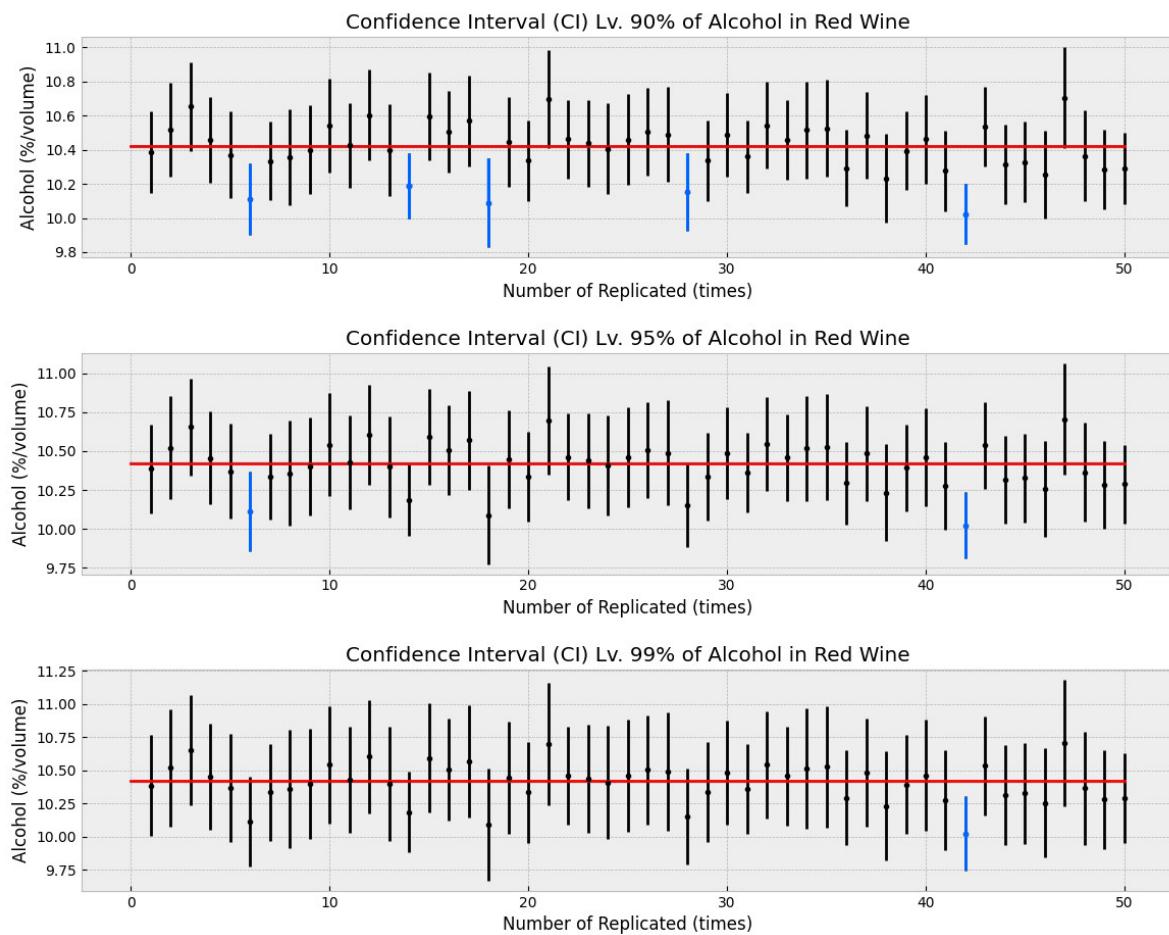
ກາພຕ້ວຍ່າງ Probability Mass Function (PMF)

## 2.2.7) Cumulative Distribution Function (CDF)



ກາພຕ້ວຍ່າງ Cumulative Distribution Function (CDF)

## 2.2.8) Confidence Interval (CI) of Mean

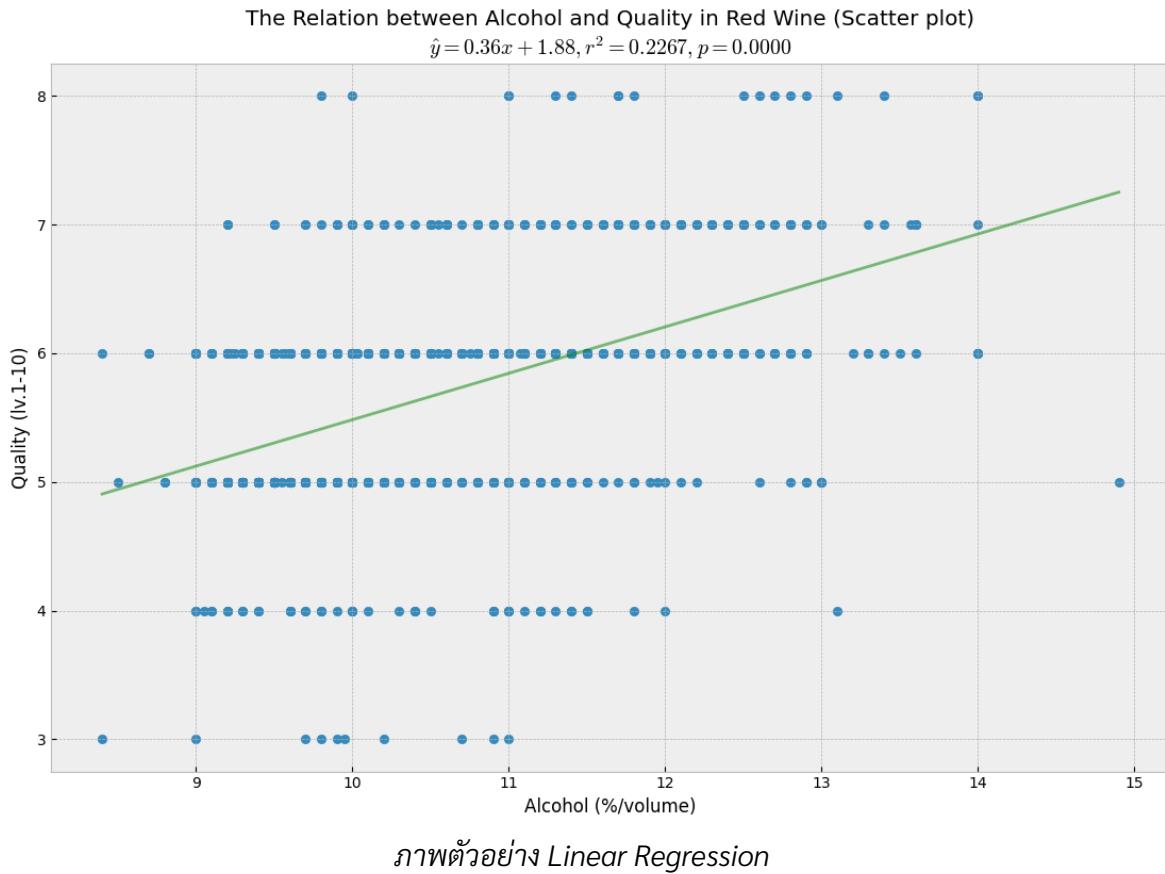


ภาพตัวอย่าง Confidence Interval (CI) of Mean

Confidence Interval (CI) of Mean เป็นช่วงค่าเฉลี่ยที่บอกระดับความมั่นใจของข้อมูล โดยสามารถจากกลุ่มตัวอย่าง (sample) และสามารถไปถึงกลุ่มข้อมูลจริง (population) ทั้งหมดได้

|  |  |  |
|--|--|--|
| <p><b>สูตรการคำนวณหา</b></p> $\text{Confidence Interval (CI)} = \bar{x} \pm t_{n-1, \alpha/2} \left( \frac{s}{\sqrt{n}} \right)$ | <p><b>Margin of error (ME)</b></p> <div style="border: 1px solid black; padding: 10px; display: inline-block;"> <math>\bar{x} \pm t_{n-1, \alpha/2} \left( \frac{s}{\sqrt{n}} \right)</math> </div> <div style="display: flex; justify-content: space-around; margin-top: 10px;"> <div style="width: 45%;"> <p>t-score at the given df and alpha level</p> </div> <div style="width: 45%;"> <p>Standard error (SE) of the mean</p> </div> </div> | <p><b>Confidence Interval (CI) of Mean</b></p> <p><math>s</math> = Standard Deviation<br/> <math>n</math> = จำนวน Samples<br/> <math>t</math> = t-score หรือ z-score<br/> <math>\bar{x}</math> = ค่าเฉลี่ย</p> |
|--|--|--|

## 2.2.9) Linear Regression



ภาพตัวอย่าง Linear Regression

1) สมการทดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression)

$$y = A + Bx$$

Constant term or  $y$ -intercept      Slope  
 ↓    ↓  
 Dependent variable                      Independent variable

2) สูตรหาค่า Slope ( $b$ ) และ Y-Intercept ( $a$ )

$$b = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

where

$$\text{ss}_{xy} = \sum xy - \frac{(\sum x)(\sum y)}{n} \quad \text{and} \quad \text{ss}_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

3) สูตรหาค่า Sum of Square Error ( $S_e$ )

$$S_e = \sqrt{\frac{SS_{yy} - bSS_{xy}}{n-2}}$$

where

$$SS_{yy} = \sum y^2 - \frac{(\sum y)^2}{n}$$

4) สูตรหาค่า Coefficient of Determination (CD หรือ  $r^2$ )

$$r^2 = \frac{b SS_{xy}}{SS_{yy}}$$

5) สูตรหาค่า Correlation Coefficient ( $r$ )

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

โดยกราฟ Linear Regression จะสามารถบอกได้ถึงแนวโน้มค่าของข้อมูลได้ว่ามีความสัมพันธ์ของข้อมูลระหว่างตัวแปรต้นและตาม เป็นไปในทิศทางไหน หากหรืออน้อยเพียงใด โดยสามารถดูได้จากค่า Coefficient of Determination (CD หรือ  $r^2$ ) โดยจะมีค่าอยู่ระหว่าง  $0.0 \leq r^2 \leq 1.0$

ซึ่งถ้าเกิดเข้าใกล้ค่า 1.0 มากเท่าไหร่ ก็จะถือว่า กราฟมีความสัมพันธ์ไปในทิศทางที่เส้นเคลื่อนที่ไปมาก และถ้าเกิดเข้าใกล้ 0.0 มากเท่าไหร่ แสดงว่าตัวแปรต้น และตัวแปรตาม มีความสัมพันธ์ที่น้อย หรือไม่มีความสัมพันธ์กันเลย

อย่างไรก็ตาม การจะวิเคราะห์ข้อมูลเหล่านี้ได้ ขึ้นอยู่กับสภาพของข้อมูล ผู้ศึกษาควรจะวิเคราะห์โดยรวมก่อนว่า สามารถใช้โมเดลของ Linear Regression ได้หรือไม่ เพราะบางข้อมูลอาจไม่ได้มีความสัมพันธ์เป็นแนวเส้นตรงเสมอไป ซึ่งผู้ศึกษาเก้าครรภจะเลือกใช้โมเดลอื่นในการมาคิดวิเคราะห์แทน เพื่อให้ข้อมูลและโมเดลที่ใช้มีความเหมาะสม และค่าเข้าใกล้ความจริงมากขึ้น

### บทที่ 3 ขั้นตอนและวิธีการดำเนินงาน

รวบรวมกราฟและการคำนวณทางสถิติทั้งหมด ดังนี้

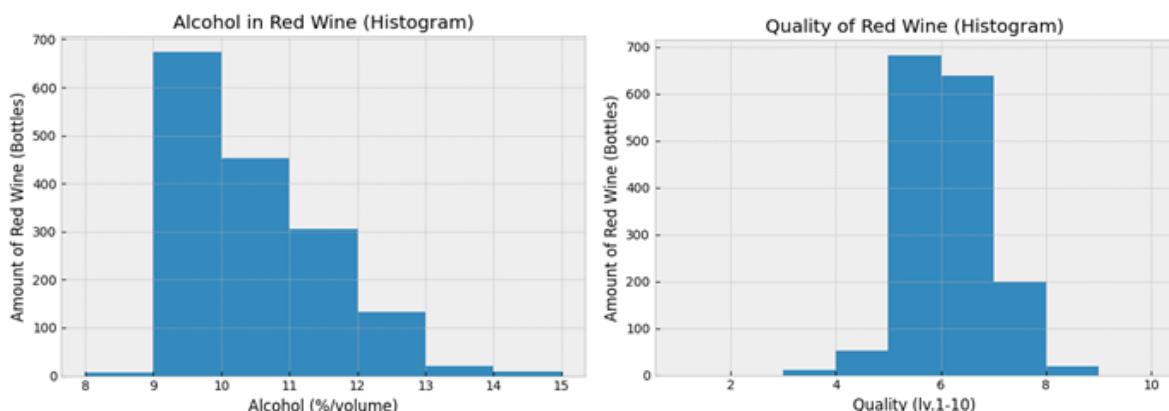
#### 3.1 คอลัมน์ที่ 1 Alcohol และ คอลัมน์ที่ 4 Quality

##### 3.1.1) ค่าทางสถิติ ได้ผลได้นี้

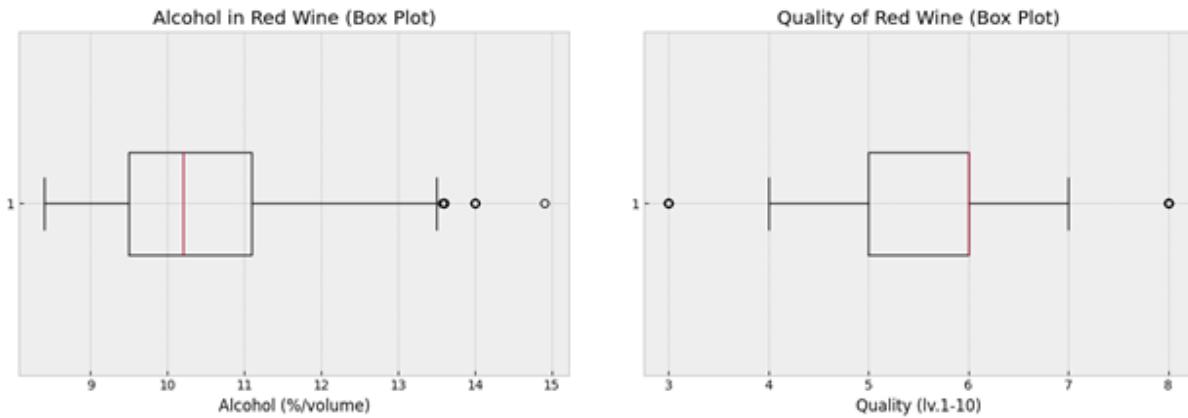
|                         | Min | Mean  | Median | Mode | Max  | Sample Standard Deviation | Sample Variation |
|-------------------------|-----|-------|--------|------|------|---------------------------|------------------|
| 1.Alcohol<br>(%/volume) | 8.4 | 10.42 | 10.20  | 9.5  | 14.9 | 1.0656                    | 1.1356           |
| 2.Quality<br>(lv.1-10)  | 3   | 5.6   | 6.0    | 5    | 8    | 0.8075                    | 0.6521           |

##### 3.1.2) กราฟทางสถิติ

###### 1) Histogram



## 2) Box Plot



2.1 Alcohol in Red Wine : มีจำนวนรวมจุด Outliers **ทั้งหมด 13 ค่า** ดังนี้ (หน่วย: %/Volume)

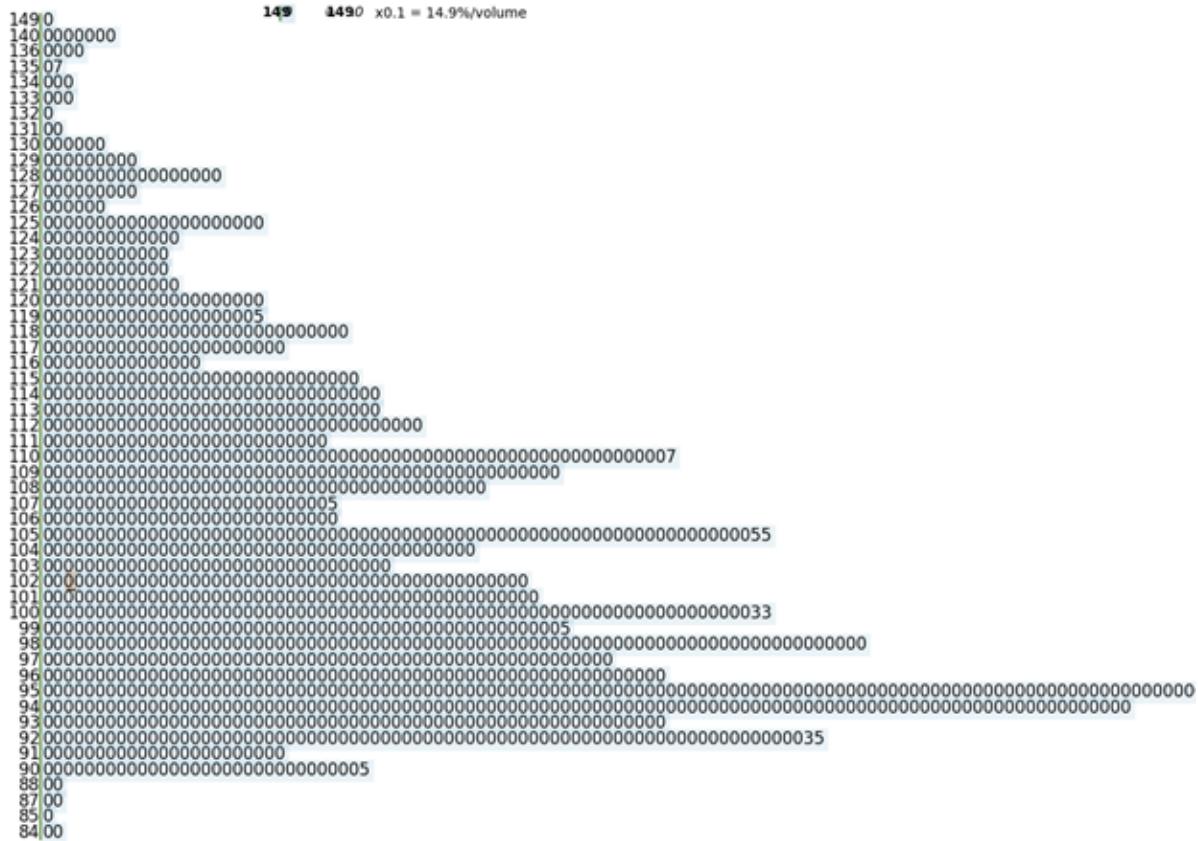
- 1) Extreme Outlier (Lower) = []
- 2) Mild Outlier (Lower) = []
- 3) Mild Outlier (Upper) = [13.57, 13.6, 13.6, 13.6, 13.6, 13.6, 14.0, 14.0, 14.0, 14.0, 14.0, 14.0, 14.9]
- 4) Extreme Outlier (Upper) = []

2.2 Quality of Red Wine : มีจำนวนรวมจุด Outliers **ทั้งหมด 28 ค่า** ดังนี้ (หน่วย: - (lv.1-10))

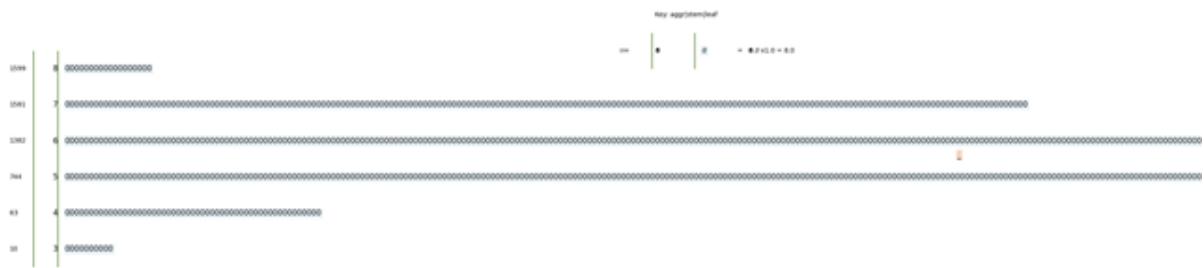
- 1) Extreme Outlier (Lower) = []
- 2) Mild Outlier (Lower) = [3, 3, 3, 3, 3, 3, 3, 3, 3, 3]
- 3) Mild Outlier (Upper) = [8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8]
- 4) Extreme Outlier (Upper) = []

### 3) Stem-And-Leaf

Stem And Leaf Graph  
Alcohol in Red Wine Stem-And-Leaf

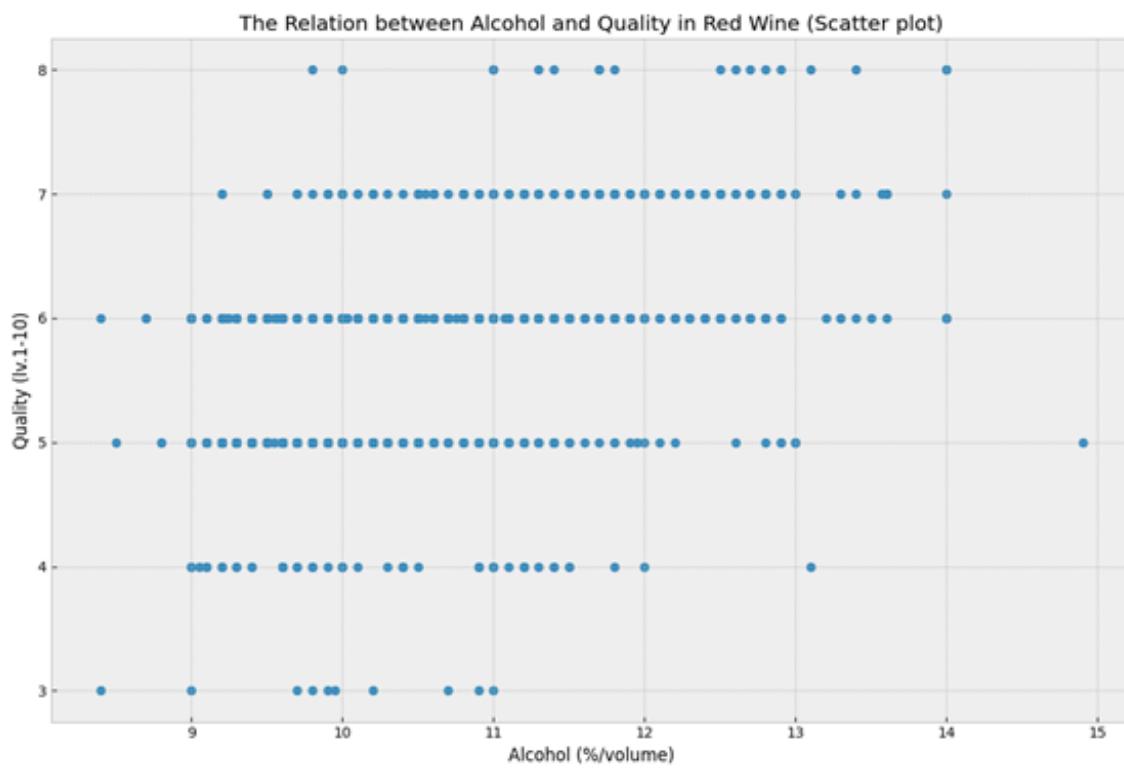


Quality of Red Wine Stem-And-Leaf



\*หมายเหตุ กราฟ Quality of Red Wine Stem-And-Leaf สามารถรวมเพื่อ Sum (เก็บแบบ Aggregation) จากด้านหน้า Stem ได้ จะเป็นการรวมกันเป็นลำดับขึ้นไปเรื่อยๆ จากล่างขึ้นบน รวมทั้งหมด 1599 ແຕງข้อมูล

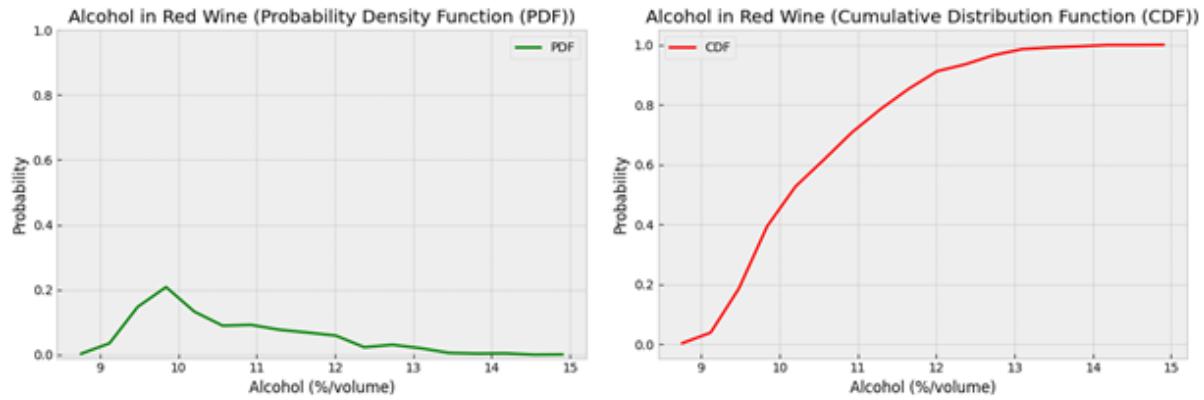
#### 4) Scatter Plot



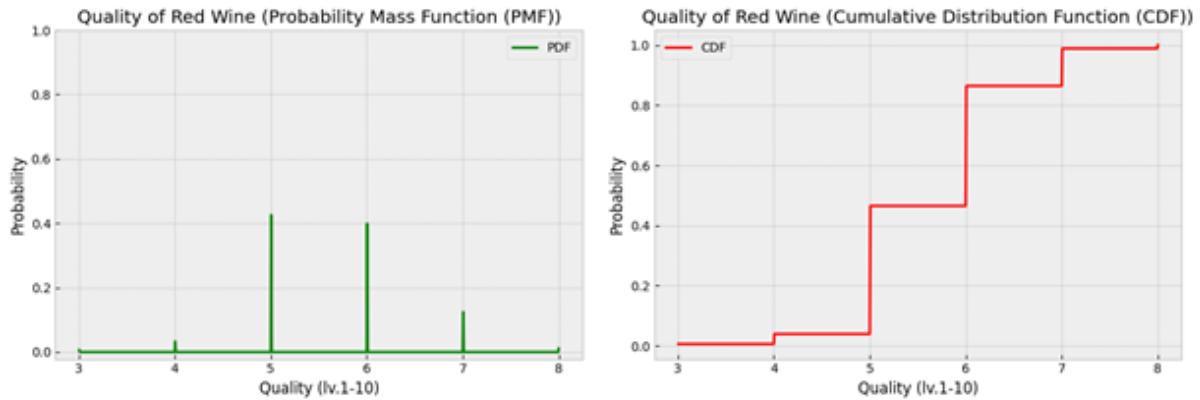
กำหนดให้แนวแกนx เป็นตัวแปรต้น(Independent Variable) = ปริมาณแอลกอฮอล์(%/ปริมาตร)  
กำหนดให้แนวแกนy เป็นตัวแปรตาม(Dependent Variable) = คุณภาพของไวน์(ระดับ 1-10)

เหตุผลที่ใช้ปริมาณแอลกอฮอล์เป็นตัวแปรต้น เพราะว่าต้องการที่จะศึกษาเกี่ยวกับ  
ปริมาณแอลกอฮอล์ในไวน์แดง ว่ามีผลต่อคุณภาพของไวน์แดงมากแค่ไหน เช่น หากเรามีปริมาณ  
แอลกอฮอล์จำนวน %/ปริมาตร ในไวน์แดงที่มาก อาจทำให้คุณภาพของไวน์แดงมากขึ้นตามด้วย  
จึงกำหนดให้ปริมาณแอลกอฮอล์เป็นตัวแปรต้น และคุณภาพของไวน์เป็นตัวแปรตาม

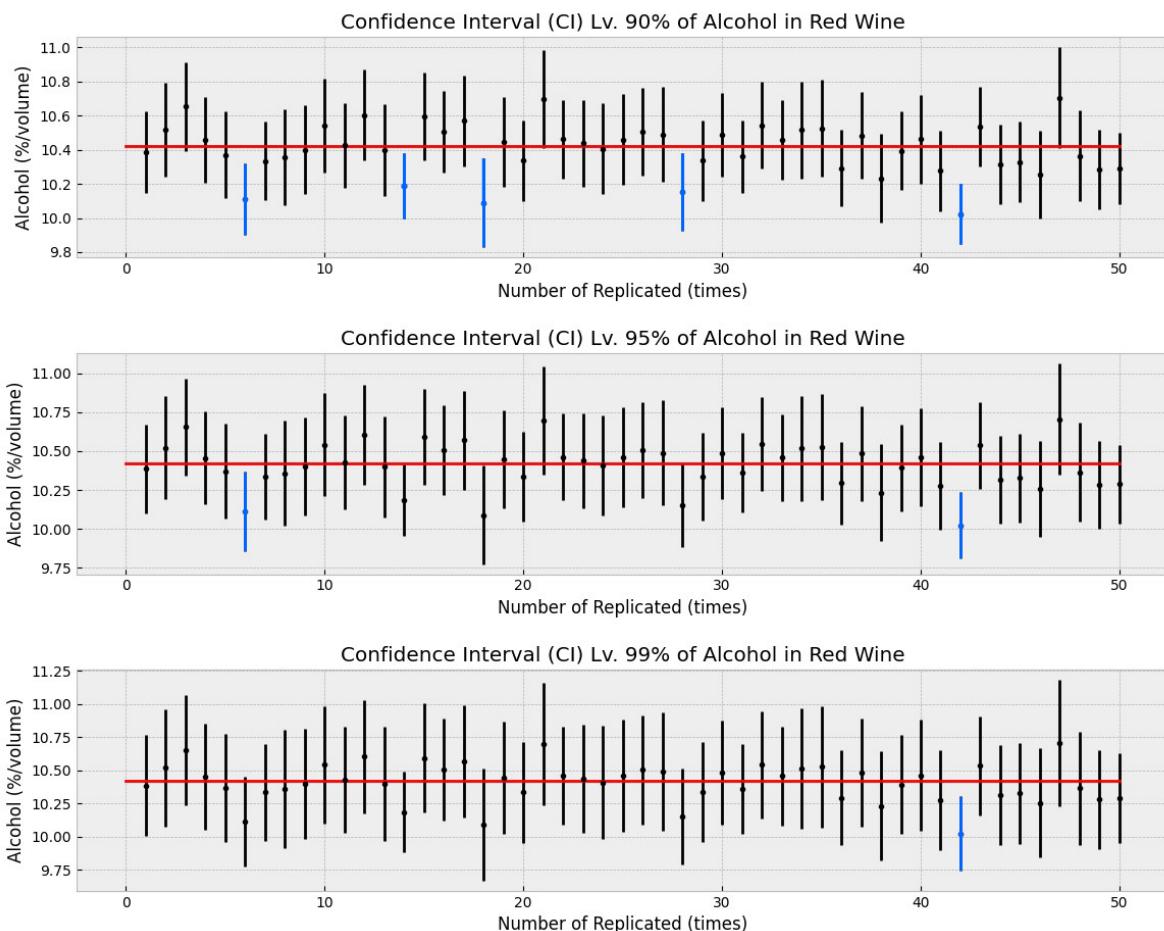
5) Probability Density Function (PDF) และ Cumulative Distribution Function (CDF) ของปริมาณแอลกอฮอล์ในไวน์แดง



6) กราฟ Probability Mass Function (PMF) และ Cumulative Distribution Function (CDF) ของคุณภาพของไวน์แดง



7) Confidence Interval (CI) of Mean ปริมาณแอลกอฮอล์ของไวน์แดง  
ที่ Level 90%, 95% และ 99%

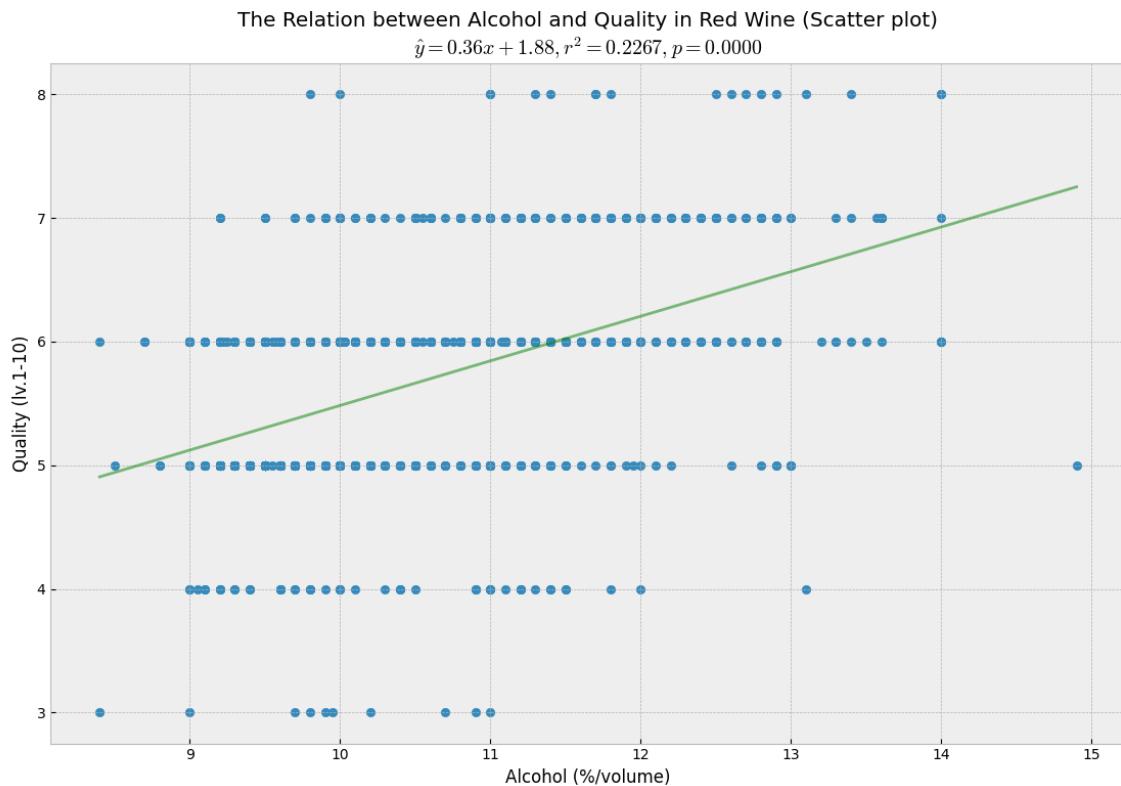


โดยจากการ มีกลุ่มตัวอย่างปริมาณแอลกอฮอล์อยู่ทั้งหมด 1599 ข้อมูล เราจะสามารถสรุปได้ดังนี้  
**Population Mean (mu) : 10.42 %/volume**  
**มีการสุ่มทั้งหมด 50 ครั้ง ครั้งละ 50 Sample จาก 1599 ข้อมูล**

- ช่วงระดับความเชื่อมั่น 90% จะมี **45/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 90 %
- ช่วงระดับความเชื่อมั่น 95% จะมี **48/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 96 %
- ช่วงระดับความเชื่อมั่น 99% จะมี **49/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 98 %

## 8) Linear Regression

ความสัมพันธ์ระหว่างปริมาณแอลกอฮอล์และคุณภาพของไวน์แดง  
ด้วยสมการทดถอยเชิงเส้นอย่างง่าย (Simple Linear Regression)



ได้สมการทดถอยเชิงเส้นอย่างง่าย และค่าอื่นๆอ กมาดังนี้

$$y = 0.36x + 1.88, r^2 = 0.2267, \text{Standard Error} = 0.7104$$

โดย  $r^2 = 0.2267$  แสดงว่า ชุดข้อมูลปริมาณแอลกอฮอล์ และ คุณภาพของไวน์แดง มี ความสัมพันธ์กันในแนวโน้มที่ค่อนข้างน้อย อยู่ที่ประมาณ 22.67% เมื่อปริมาณแอลกอฮอล์ เพิ่ม ขึ้น 1%/volume จะทำให้ระดับคุณภาพของไวน์แดง เพิ่มขึ้น 0.36 (สังเกตได้จากค่าอัตราการเปลี่ยนแปลงหรือ Slope)

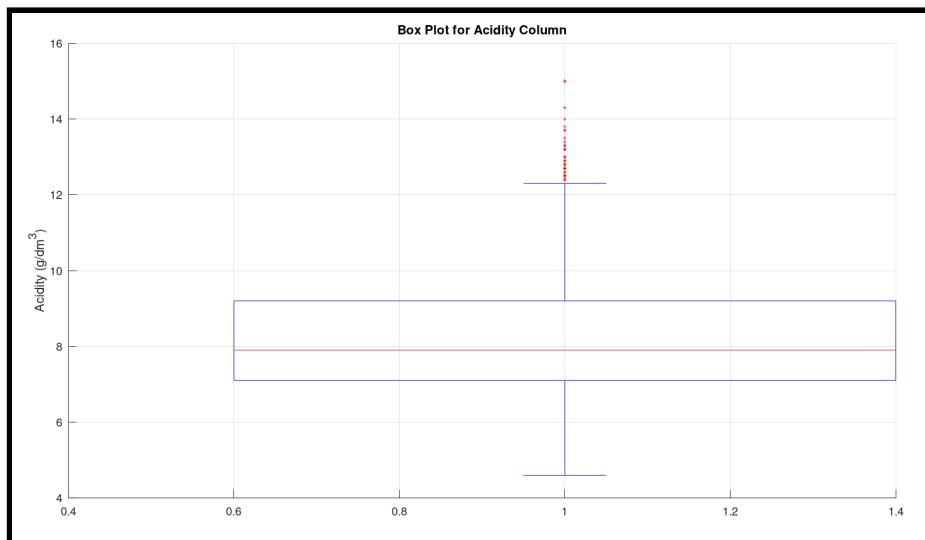
### 3.2 คอลัมน์ที่ 2 Fixed Acidity (ปริมาณกรด หน่วยคือ กรัม / ลิตร )

#### 3.2.1) ค่าทางสถิติ ได้ผลได้นี้

|                      |      |          |       |
|----------------------|------|----------|-------|
| จำนวนชุดข้อมูล (ชุด) | 1599 |          |       |
| Mean                 | 8.13 | Max      | 15.90 |
| Median               | 7.90 | Min      | 4.60  |
| Mode                 | 7.20 | Range    | 11.30 |
| Standard Deviation   | 1.75 | Variance | 3.030 |

#### 3.2.2) กราฟทางสถิติ

##### 1) Box Plot

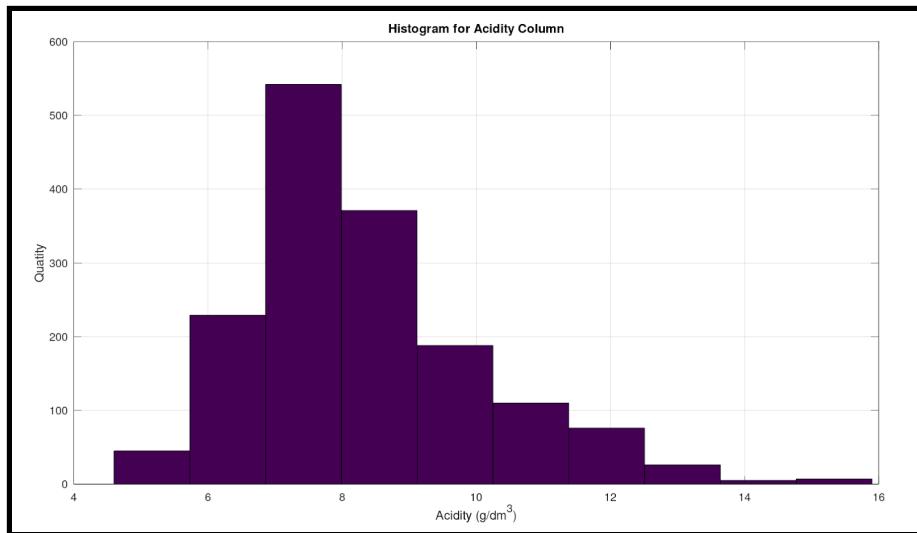


จากชุดข้อมูลจะได้ Outlier ดังนี้

(0),(12.4),(12.4),(12.4),(12.4),(12.5),(12.5),(12.5),(12.5),(12.5),(12.5),(12.5),(12.6),(12.6),  
 (12.6),(12.6),(12.7),(12.7),(12.7),(12.7),(12.8),(12.8),(12.8),(12.8),(12.8),(12.8),(12.9),(12.9),  
 (13),(13),(13),(13.2),(13.2),(13.2),(13.3),(13.3),(13.3),(13.4),(13.4),(13.5),(13.5),(13.7),(13.7),(13.8),  
 (14),(14.3),(15),(15),(15.5),(15.5),(15.6),(15.6),(15.9)

เป็นจำนวน 50 คู่ คิดเป็น 3.125% ของข้อมูลทั้งหมด

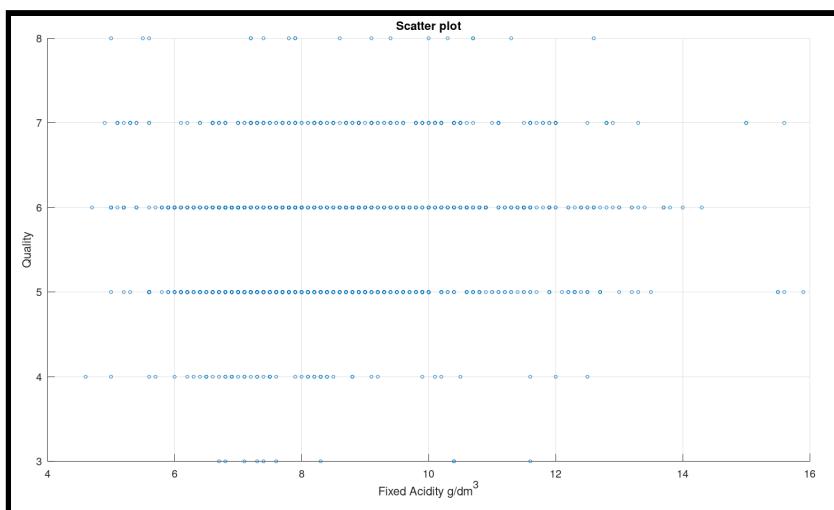
## 2) Histogram



### 3) Stem-And-Leaf

|  |  |  |
|--|--|--|
| 0  |  | 777777777767578888778778667777668657785778786478658668777787777769877789886686777668797788754677887887788876687778667777569867779887788777688789765886877788777896777876787768778997886776687767877798779999999999998677878798997788999899979695997889669877596997888789886878999876688878778899789999778869 |
| 99888899999999998677878798997788999899979695997889669877596997888789886878999876688878778899789999778869     |  |  |
| 99867887788787786889989998777997776687769679987877899978578888776979746677775758766667966766998997986757777  |  |  |
| 888888789986878889896887787997977977767556879968899887888887787876777787877887689667799785888878988876897    |  |  |
| 856888998986797977667588966767688967676777789777976666987889897957777677867998577767777767766869887765575778 |  |  |
| 977776677988777686887876785777688877766687868877777787868668967777768877775667776667766776677667766666676    |  |  |
| 768676777676766677777766675687686667666677566656566  |  |  |
| 1  |  | 1001221125501021101012110300100000010300122001303101212202004111322112120201020120210215021101011  |
| 3031110000502021010031001100002020001010100002000000220011010110110000000000000002001111000001011            |  |  |

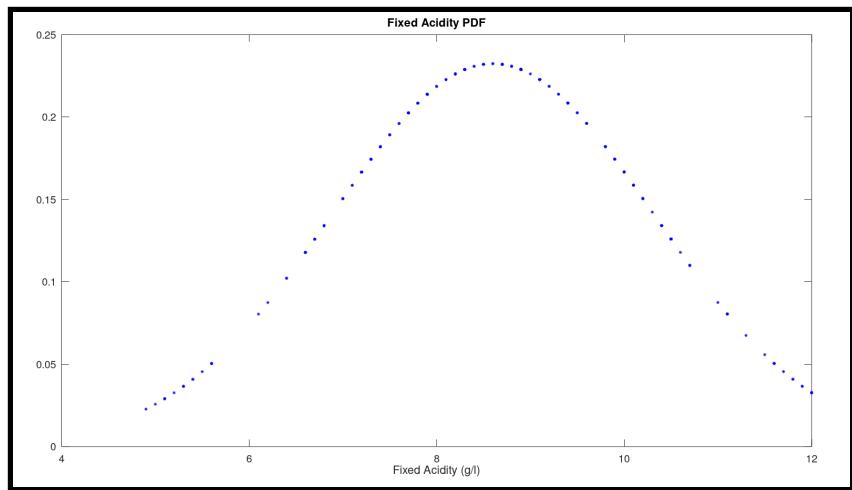
#### 4) Scatter Plot



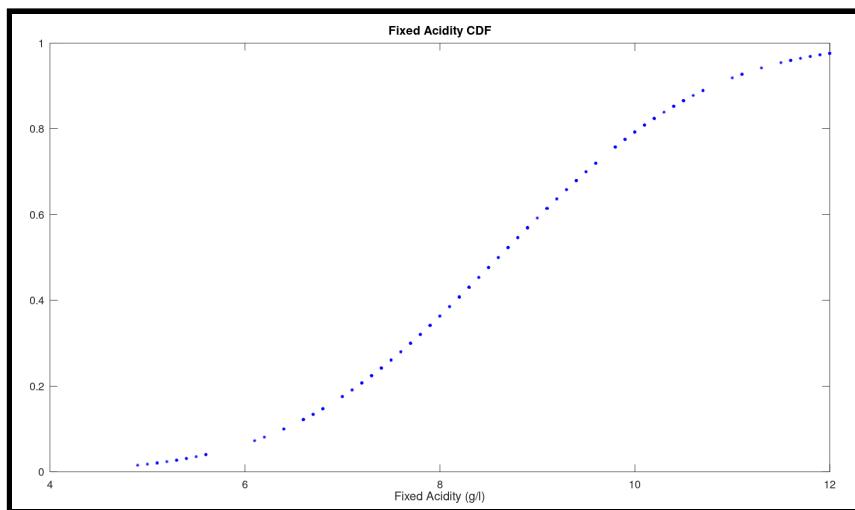
กำหนดให้แนวแกน x เป็นตัวแปรต้น(Independent Variable) = ปริมาณกรด ( กรัม / ลิตร )

กำหนดให้แนวแกน y เป็นตัวแปรตาม(Dependent Variable) = คุณภาพของไวน์(ระดับ 1-10)

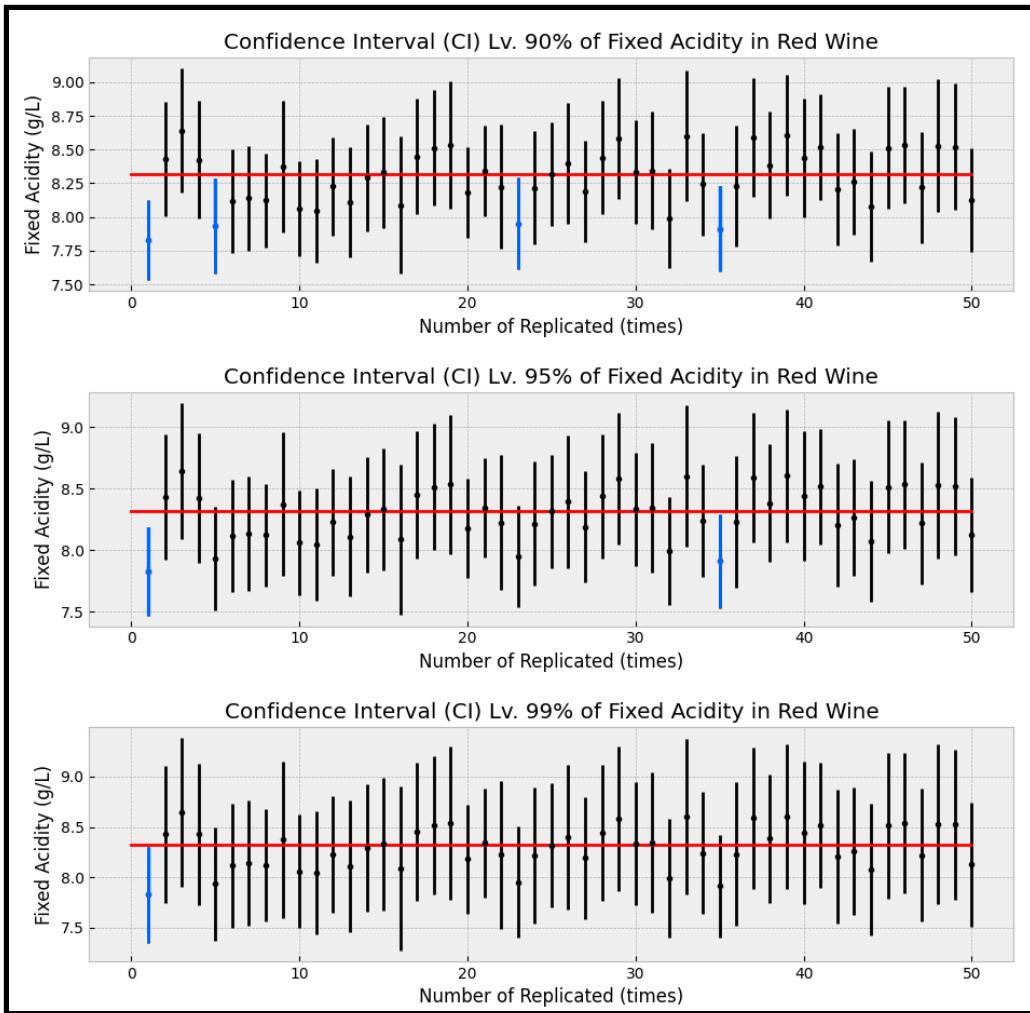
## 5) Probability Density Function (PDF)



## 6) Cumulative Distribution Function (CDF)



### 7) Confidence Interval graph ที่ Level 90%, 95% และ 99%

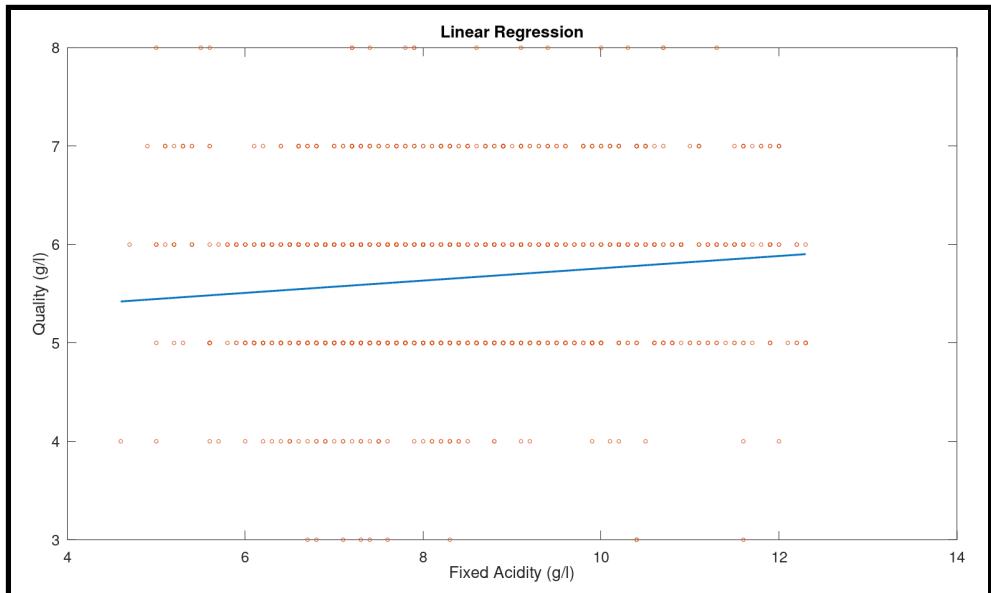


โดยจากราฟ มีกลุ่มตัวอย่างปริมาณของค่าความเป็นกรดอยู่ทั้งหมด 1599 ข้อมูล เราจะสามารถสรุปได้ดังนี้

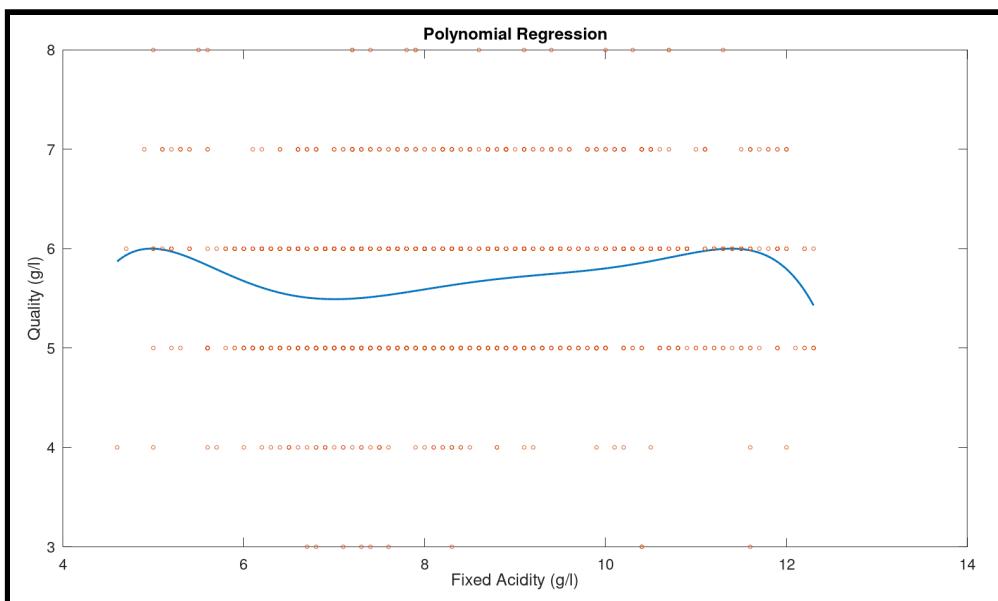
**Population Mean ( $\mu$ ) : 8.13 g/L**  
**มีการสุ่มทั้งหมด 50 ครั้ง ครั้งละ 50 Sample จาก 1599 ข้อมูล**

- ช่วงระดับความเชื่อมั่น 90% จะมี **46/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 92 %
- ช่วงระดับความเชื่อมั่น 95% จะมี **48/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 96 %
- ช่วงระดับความเชื่อมั่น 99% จะมี **49/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 98 %

## 8) Regression



ได้สมการถดถอยเชิงเส้นอย่างง่าย และค่าอื่นๆอ กมาดังนี้  
 $y = 0.062x + 5.13$ ,  $r^2 = 0.0320$ , Standard Error = 0.6392



ได้สมการถดถอยพหุนาม และค่าอื่นๆอ กมาดังนี้  
 ดีกรี = 6 สมการคือ  $y = c + m_1x + m_2x^2 + m_3x^3 + m_4x^4 + m_5x^5 + m_6x^6$   
 โดยได้  $m_i$  ดังนี้ = [ -229.78733226, 186.56840158, -59.90086736, 10.00862157, -0.92088352,  
 0.04436100, -0.00087589 ] ตามลำดับ  
 $r^2 = 0.0320$ , Standard Error = 0.6392

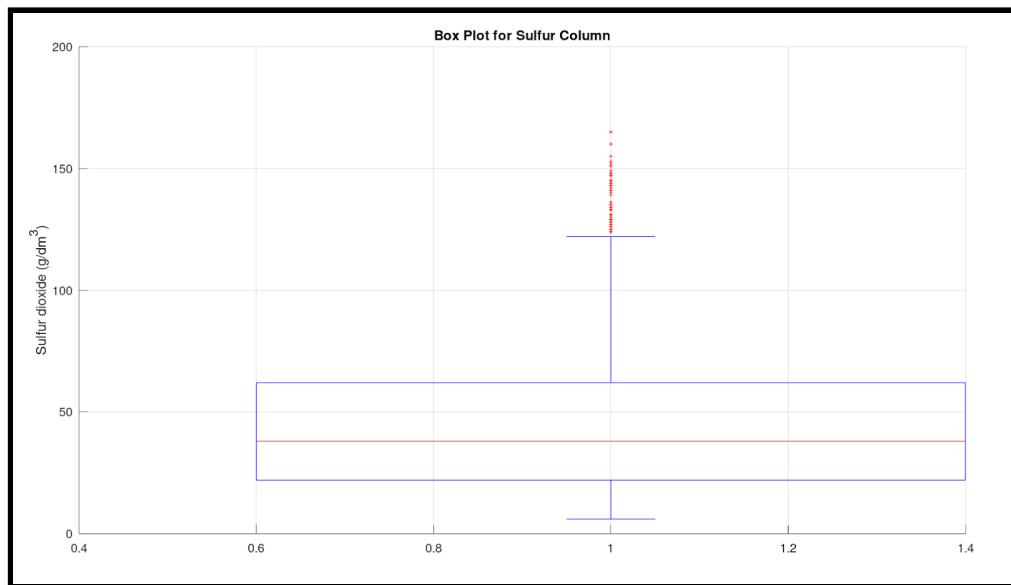
### 3.3 คอลัมน์ที่ 3 ซัลเฟอร์ไดออกไซด์

#### 3.3.1) ค่าทางสถิติ ได้ผลได้นี้

|                      |       |          |         |
|----------------------|-------|----------|---------|
| จำนวนชุดข้อมูล (ชุด) | 1599  |          |         |
| Mean                 | 46.47 | Max      | 289.00  |
| Median               | 38.00 | Min      | 6.00    |
| Mode                 | 28.00 | Range    | 283.00  |
| Standard Deviation   | 32.89 | Variance | 1082.10 |

#### 3.3.2) กราฟทางสถิติ

##### 1) Box Plot

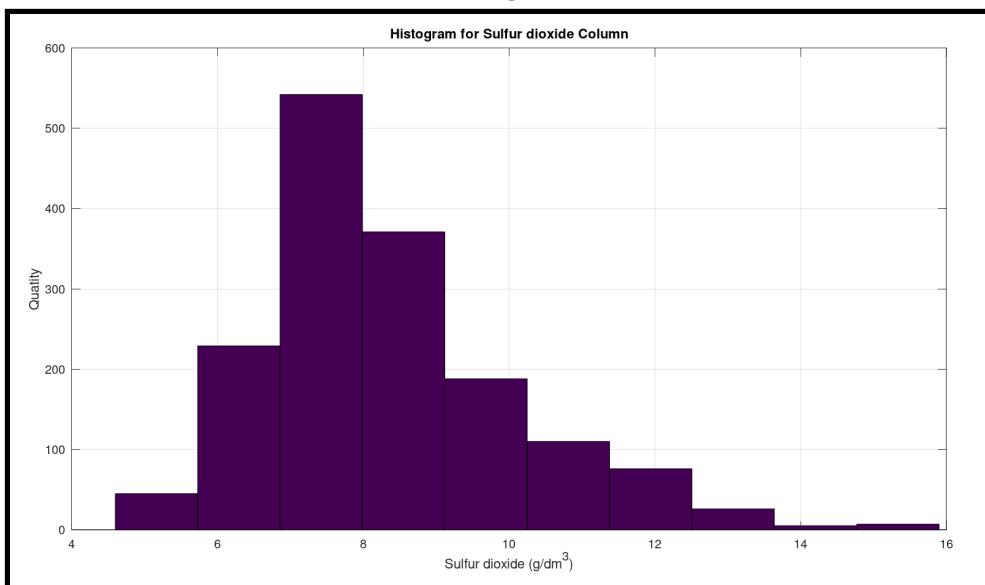


จากชุดข้อมูลจะได้ Outlier ดังนี้

(122) ,(122) ,(122) ,(124) ,(124) ,(124) ,(125) ,(125) ,(126) ,(127) ,(127) ,(127) ,(128) ,(128) ,(129) ,(129) ,(129) ,(130) ,(131) ,(131) ,(131) ,(133) ,(133) ,(133) ,(134) ,(134) ,(134) ,(135,5) ,(135) ,(136) ,(136) ,(139) ,(140) ,(141) ,(141) ,(141) ,(142) ,(143) ,(143) ,(144) ,(144) ,(144) ,(144) ,(145) ,(145) ,(145) ,(147) ,(147) ,(147) ,(147) ,(148) ,(148) ,(148) ,(149) ,(149) ,(151) ,(151) ,(152) ,(153) ,(155) ,(160) ,(165) ,(278) ,(289)

เป็นจำนวน 58 คู่ คิดเป็น 3.625% ของข้อมูลทั้งหมด

## 2) Histogram



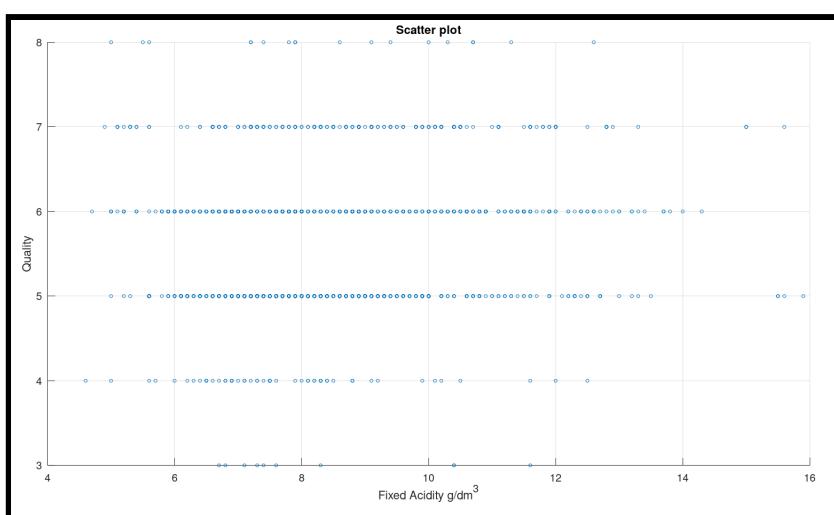
### 3) Stem-And-Leaf

```

0 | 777777776757888778778667776686577857786478658668777787777769877789886686777668797788754677887887
7888766877786677756986777988778877768878976588687778877877778967778786787768778997886776878779877999
998888999999998867787897997789998999789695997889698698877596997988788986878999876888878778997889999778869
98967887788778776889989998777977766877696799878789997857888777697974667777575876666796676699899798675777
888888789688768898968877899797797767755867996869989878778888778767777878778887689667799785888879898876897
85688998986797977667588966767689676767777789777976666987889897957777677867998577677777677668698876555778
9777766779887776868787685777688877666878688777777886866689677777688777775667776677667776678766666676
76867677767676667777766676578666676666757665665656
1 | 100122112501021101012110300100000010300012200013031012122020041113322112120201020120210215021101011
303111000050202101000310011000020200101010000200000022001101011010000000000000000200111100001011

```

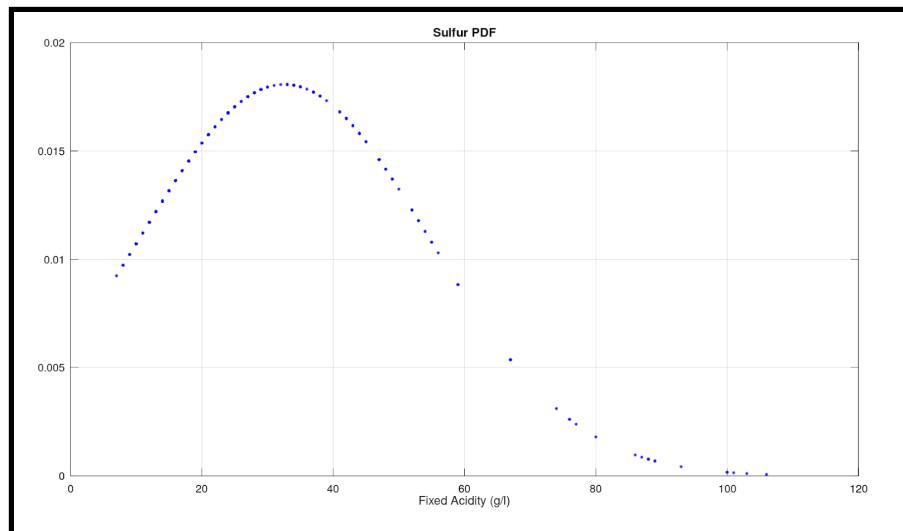
#### 4) Scatter Plot



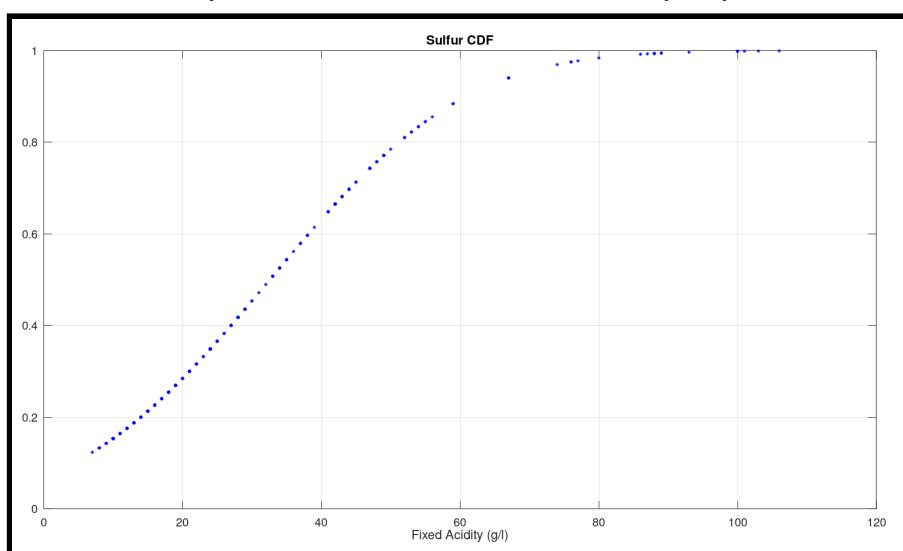
กำหนดให้แนวแกน x เป็นตัวแปรต้น(Independent Variable) = ปริมาณ ชัลเฟอร์ไดออกไซด์ (กรัม / ลิตร )

กำหนดให้แนวแกน y เป็นตัวแปรตาม(Dependent Variable) = คุณภาพของไวน์(ระดับ 1-10)

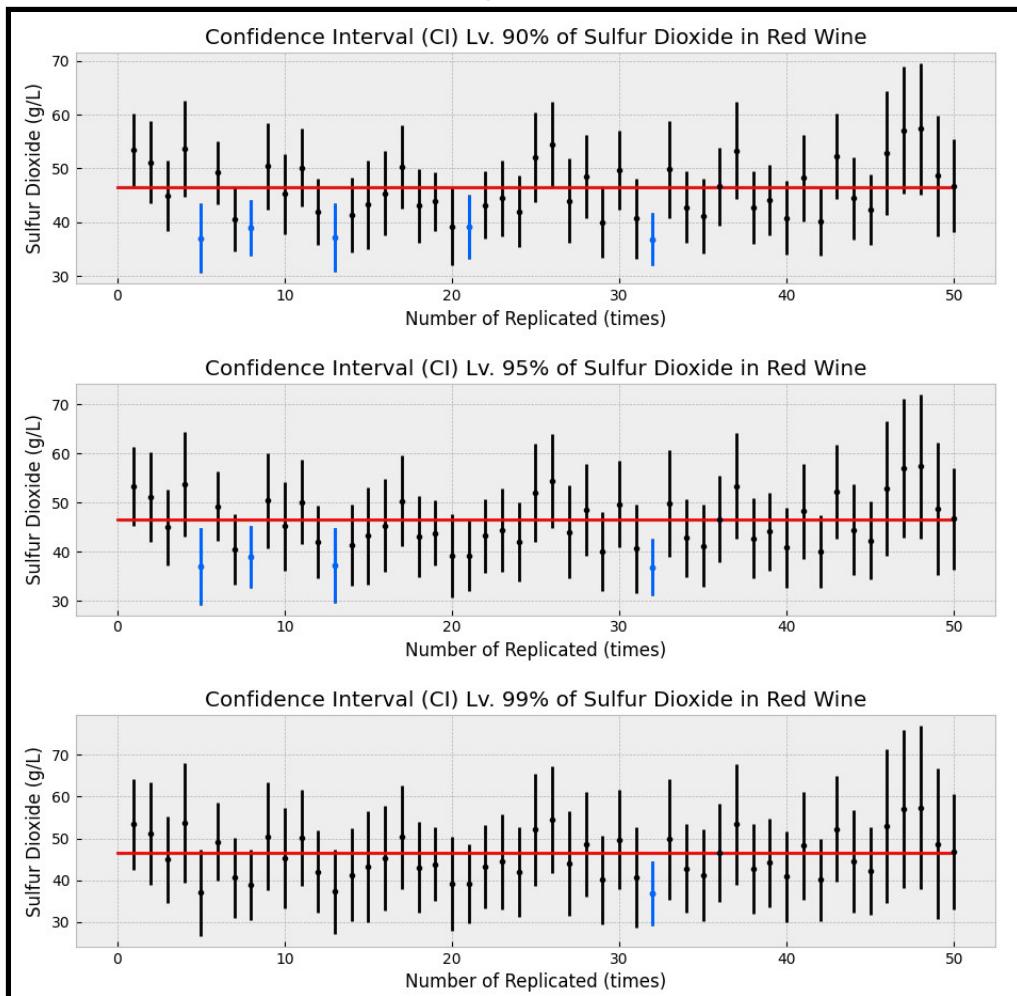
## 5) Probability Density Function (PDF)



## 6) Cumulative Distribution Function (CDF)



### 7) Confidence Interval graph ที่ Level 90%, 95% และ 99%

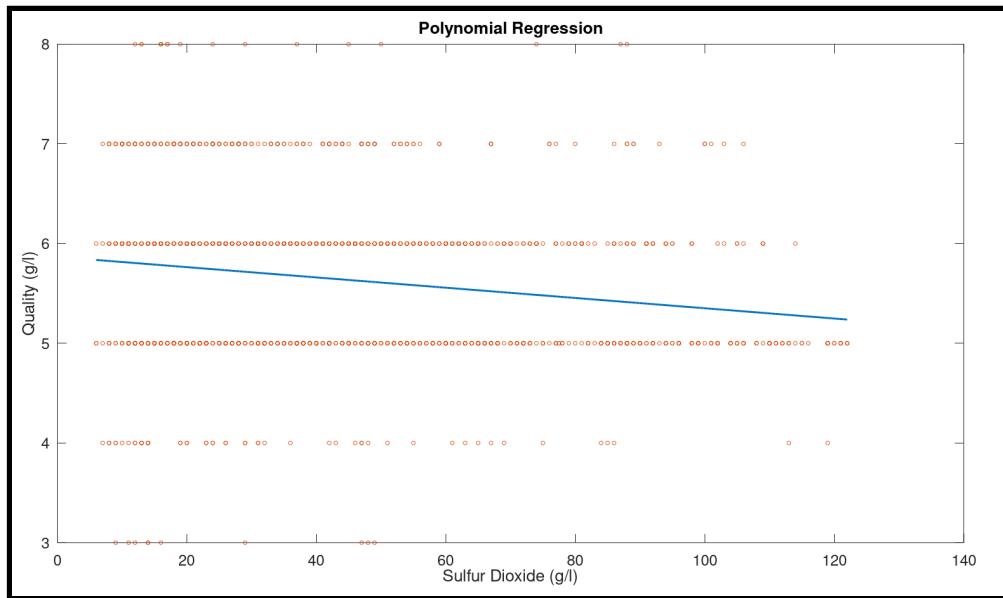


โดยจากราฟ มีกลุ่มตัวอย่างปริมาณของซัลเฟอร์ไดออกไซด์อยู่ทั้งหมด 1599 ข้อมูล เราจะสามารถสรุปได้ดังนี้

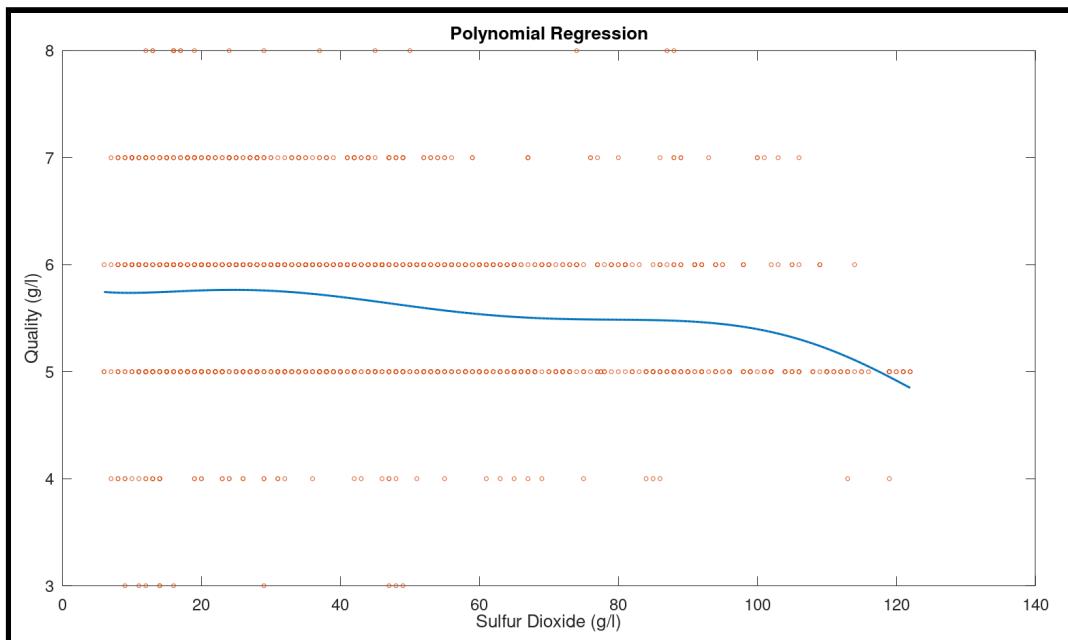
**Population Mean ( $\mu$ ) : 46.47 g/L**  
**มีการสุ่มทั้งหมด 50 ครั้ง ครั้งละ 50 Sample จาก 1599 ข้อมูล**

- ช่วงระดับความเชื่อมั่น 90% จะมี **45/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 90 %
- ช่วงระดับความเชื่อมั่น 95% จะมี **46/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 92 %
- ช่วงระดับความเชื่อมั่น 99% จะมี **49/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 98 %

## 8) Regression



ได้สมการถดถอยเชิงเส้นอย่างง่าย และค่าอื่นๆอุปกรณ์ดังนี้  
 $y = -0.005x + 5.86$ ,  $r^2 = 0.0304$ , Standard Error = 0.6382



ได้สมการถดถอยพหุนาม และค่าอื่นๆอุปกรณ์ดังนี้  
 ดีกรี = 6 สมการคือ  $y = c + m_1x + m_2x^2 + m_3x^3 + m_4x^4 + m_5x^5 + m_6x^6$   
 โดยได้  $m_i$  ดังนี้ = [ 5.820751576973668, -0.022111074060547, 0.001953325170093  
 $-0.000067953148009, 0.000001029067351 -0.000000007036056 0.000000000017662 ]$   
 ตามลำดับ  
 $r^2 = 0.03497$ , Standard Error = 0.6373

## บทที่ 4 ผลการดำเนินงาน

บทวิเคราะห์ข้อมูลจากการและข้อมูลทางสถิติ มีทั้งหมด ดังนี้

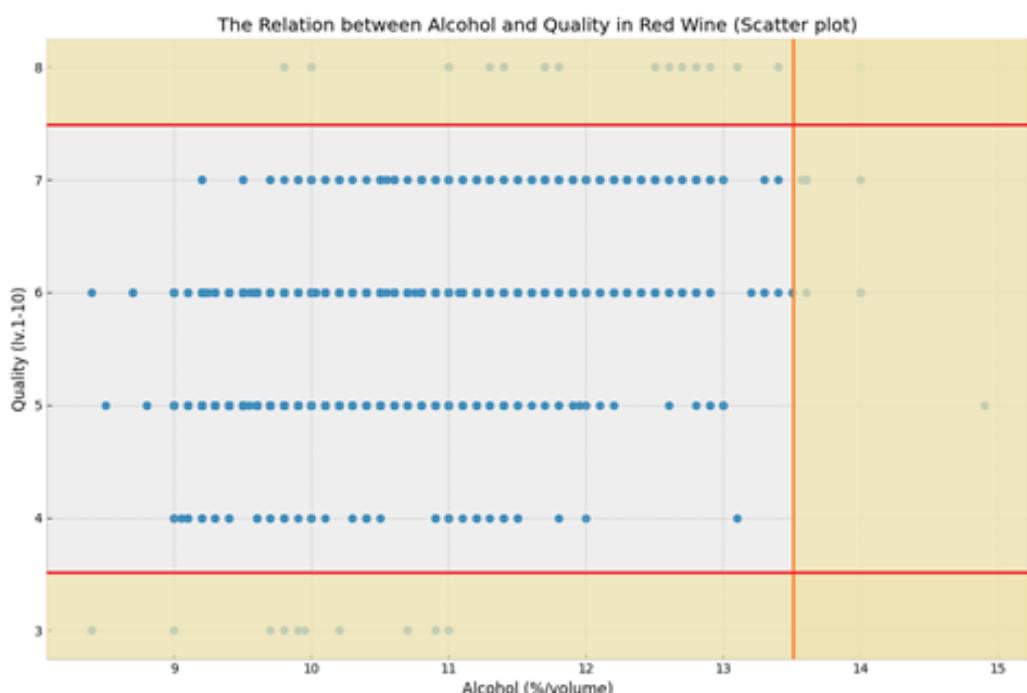
### 4.1 คอลัมน์ที่ 1 Alcohol และ คอลัมน์ที่ 4 Quality

4.1.1) บทวิเคราะห์กราฟ Scatter Plot ระหว่างความสัมพันธ์ของปริมาณแอลกอฮอล์ และคุณภาพของไวน์แดง

เนื่องจากกราฟ Scatter ความสัมพันธ์ระหว่างปริมาณแอลกอฮอล์และคุณภาพของไวน์แดงในปัจจุบัน มีจำนวน Outliers อยู่จำนวนหนึ่ง ผู้จึงทำการลง plot กราฟ Scatter ใหม่ โดยการตีเส้นและไม่นำ Outliers มาคิดในการวิเคราะห์ จะได้ออกมาเป็นกราฟดังนี้ (อ้างอิงจากการคำนวณในหน้าที่ 12)

Outlier Alcohol [ value < 7.1, value > 13.5] (เส้นสีส้ม)

Outlier Quality [ value < 3.5, value > 7.5] (เส้นสีแดง)



จากการที่ได้ เราจะสนใจแค่เพียงในส่วนสีขาวเท่านั้น (ที่ไม่ใช่แรเงาสีเหลือง) จะเห็นได้ว่า ข้อมูลที่ได้ค่อนข้างมีความกระจุกตัวอยู่บริเวณตรงกลางเป็นส่วนมาก และมีแนวโน้มเอียงขึ้นไปทางบนขวาเล็กน้อย

ซึ่งถ้าหากเราสร้างเกตกราฟโดยละเอียด จะพบว่า

- ช่วงปริมาณแอลกอฮอล์ตั้งแต่ 8.0-9.5 %/volume คุณภาพของไวน์ส่วนใหญ่จะอยู่ที่ระดับ 4-6

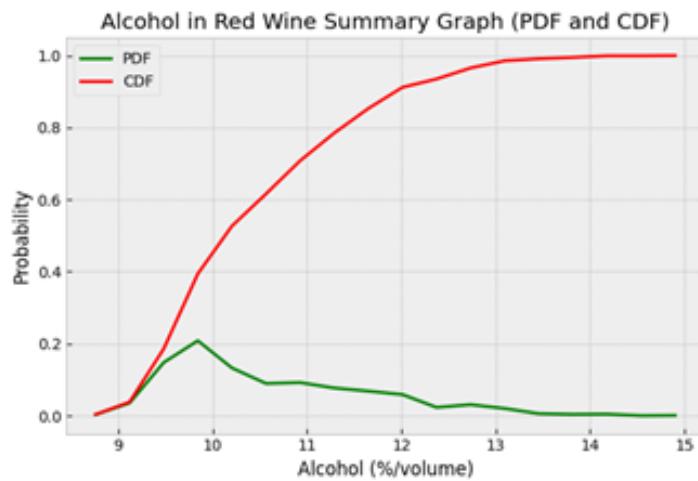
2. และช่วงปริมาณแอลกอฮอล์ตั้งแต่ 12.0-13.0 %/volume ขึ้นไป จะมีคุณภาพของไวน์ตั้งแต่ระดับ 6-7 เป็นส่วนมาก

ซึ่งจากการวิเคราะห์ที่ได้ หากเรามีปริมาณแอลกอฮอล์ในไวน์แดงในปริมาณน้อย คุณภาพของไวน์แดงจะมีแนวโน้มที่จะน้อยตามไปด้วย และถ้าหากเรามีปริมาณแอลกอฮอล์ในไวน์แดงที่มาก คุณภาพของไวน์แดงก็จะมีแนวโน้มมากขึ้นตามไปด้วย โดยข้อมูลทั้งหมดนี้ถูกเก็บรวบรวมและอ้างอิงมาจากโรงงานผลิตไวน์ จังหวัด Minho ทางตอนเหนือของประเทศโปรตุเกสฯ

กล่าวโดยสรุปคือ ปริมาณแอลกอฮอล์ในไวน์แดงที่มากขึ้น อาจมีแนวโน้มที่จะทำให้คุณภาพของไวน์แดงเพิ่มขึ้นตามไปด้วย โดยความสัมพันธ์ระหว่างปริมาณแอลกอฮอล์ในไวน์แดง และคุณภาพของไวน์แดงเป็นความสัมพันธ์แบบแปรผันตรงหรือคล้ายตามกัน

อย่างไรก็ตาม ทั้งหมดนี้ยังไม่สามารถกล่าวได้อย่างชัดเจน 100% เป็นเพียงแค่แนวโน้มเท่านั้น เนื่องจากในการผลิตไวน์จริง จะมีส่วนผสมอื่นๆ และมีอีกหลายปัจจัยในการกำหนดคุณภาพของไวน์แดง เช่น ปริมาณกรด, น้ำตาลคงค้างที่เหลือในไวน์แดง, ระยะเวลาการผลิตไวน์แดง, คุณภาพของอุ่นที่นำมาใช้ในการผลิต และ เกณฑ์การวัดคุณภาพของไวน์แดง เป็นต้น ซึ่ง เกณฑ์การวัดคุณภาพของไวน์แดง (ระดับ 1-10) ในครั้งนี้ อ้างอิงมาจากโรงงานผลิตไวน์ ในจังหวัด Minho ทางตอนเหนือของประเทศโปรตุเกสเท่านั้น

#### 4.1.2) บทวิเคราะห์กราฟ Probability Density Function (PDF)/ Cumulative Distribution Function (CDF) ของปริมาณแอลกอฮอล์



##### ปริมาณแอลกอฮอล์ในไวน์แดง

ในกราฟปริมาณแอลกอฮอล์ เส้นสีเขียว (PDF) สังเกตได้ ดังนี้

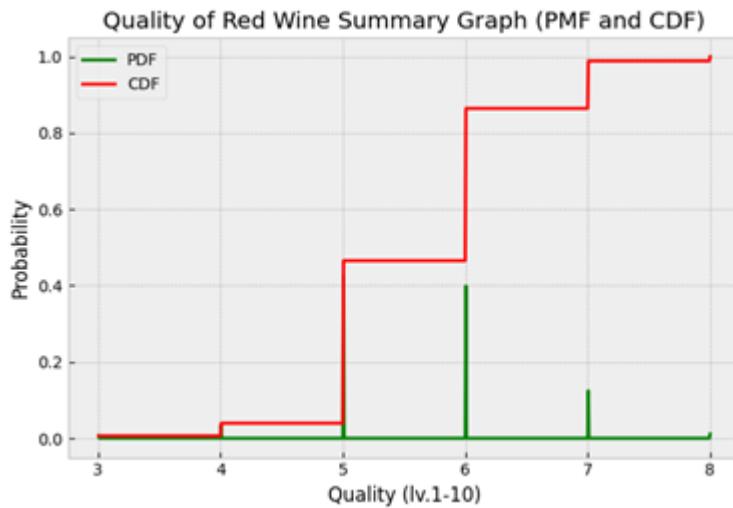
- จะมีลักษณะเป็น ภูเขาสูง ระหว่างช่วงปริมาณแอลกอฮอล์ตั้งแต่ 9.2 – 10.5 %/volume
- จะมีลักษณะเป็น ที่รบлен ระหว่างช่วงปริมาณแอลกอฮอล์ตั้งแต่ 8.5 – 9.1 %/volume และ 10.6 – 15.0 %/volume

ในกราฟปริมาณแอลกอฮอล์ เส้นสีแดง (CDF) สังเกตได้ ดังนี้

- เป็นการรวมค่าความน่าจะเป็น(ความถี่สัมพัทธ์) จากเส้นกราฟสีเขียว เพิ่มขึ้นเรื่อยๆ สะสมจนครบ 1.0 เมื่อสิ้นสุดกราฟนั้น ๆ ซึ่งเป็นความถี่สะสมสัมพัทธ์
- จะมีค่าความน่าจะเป็น = 0.5 ที่ประมาณ 10.2 %/volume ซึ่งเป็นจุดกึ่งกลางของกราฟ พอดี (Median) ทำให้ทางซ้ายและทางขวาของพื้นที่ใต้กราฟสีเขียวแบ่งออกเป็น 50% เท่า ๆ กัน
- อัตราการเปลี่ยนแปลงของกราฟช่วงประมาณ 9.2 - 9.7 %/volume มีอัตราการเปลี่ยนแปลงโดยรวมมาก และจะค่อย ๆ ลดหลั่นลงไปเรื่อยๆ เพื่อมีปริมาณแอลกอฮอล์มากขึ้น

กล่าวโดยสรุป จากข้อมูลที่ได้ ไวน์จะมีปริมาณแอลกอฮอล์ประมาณ 9.2 – 10.5 %/volume เป็นส่วนมาก เนื่องจากมีพื้นที่ใต้กราฟเส้นสีเขียวเยอะ (Density หนาแน่น) และไวน์จะมีปริมาณแอลกอฮอล์ประมาณ 8.5 – 9.1 %/volume และ 10.6 – 13.0 %/volume เป็นส่วนที่น้อย เนื่องจากพื้นที่ใต้กราฟเส้นสีเขียนน้อย (Density เบาบาง) และถ้ายิ่งมีปริมาณแอลกอฮอล์ที่มากขึ้นไปกว่า 13.0 %/volume ก็จะมีจำนวนไวน์ที่จะลดลงตามลำดับลงไป (Density เบาบางมาก ๆ) ซึ่งข้อมูลข้างต้นสามารถนำมาดูเปรียบเทียบได้กับเส้นสีแดงควบคู่กัน ถ้าเส้นสีแดงมีอัตราการเปลี่ยนแปลงมาก ก็จะมีการเพิ่มจำนวนไวน์ในปริมาณแอลกอฮอล์ช่วงนั้นที่มากขึ้น ส่วนถ้ามีอัตราการเปลี่ยนแปลงน้อย ก็จะมีการเพิ่มจำนวนไวน์ในปริมาณแอลกอฮอล์ช่วงนั้นที่น้อยลง

#### 4.1.3) บทวิเคราะห์กราฟ Probability Mass Function (PMF)/ Cumulative Distribution Function (CDF) ของคุณภาพของไวน์แดง



#### คุณภาพของไวน์แดง

ในกราฟคุณภาพไวน์แดง เส้นสีเขียว (PMF) สังเกตได้ ดังนี้

- คุณภาพระดับ 5 และ 6 จะมีปริมาณมาก เมื่อเทียบกับคุณภาพไวน์อื่นๆ เนื่องจากมีแท่งสีเขียวที่ค่อนข้างสูง得多เด่นเป็นพิเศษ และมีค่าความน่าจะเป็น(ความถี่สัมพัทธ์) โดยประมาณ = 0.4
- คุณภาพระดับ 7 จะมี ความน่าจะเป็น(ความถี่สัมพัทธ์) โดยประมาณ = 0.1
- คุณภาพระดับ 3 4 และ 8 จะมี ความน่าจะเป็น(ความถี่สัมพัทธ์) โดยประมาณน้อยกว่า 0.05ซึ่งเป็นค่าที่น้อยมากๆ

ในกราฟคุณภาพไวน์แดง เส้นสีแดง (CDF) สังเกตได้ ดังนี้

- เป็นการรวมค่าความน่าจะเป็น(ความถี่สัมพัทธ์) จากเส้นกราฟสีเขียว เพิ่มขึ้นเรื่อยๆ สะสมจนครบ 1.0 เมื่อสิ้นสุดกราฟนั้น ๆ ซึ่งเป็นความถี่สะสมสัมพัทธ์
- จะมีค่าความน่าจะเป็น = 0.5 ที่คุณภาพระดับ 6 ซึ่งเป็นจุดกึ่งกลางของกราฟพอดี (Median)
- อัตราการเปลี่ยนแปลงของกราฟคุณภาพระดับที่ 5 และ 6 จะลากยาว และมีความสูงเป็นพิเศษกว่าระดับคุณภาพอื่นๆ เนื่องจากมีจำนวนคุณภาพระดับ 5 และ 6 ที่มีเยอะมากกว่าระดับอื่นๆ

กล่าวโดยสรุป จากข้อมูลที่ได้จากเส้นสีเขียว ไวน์จะมีระดับคุณภาพอยู่ที่ประมาณ 5 และ 6 เป็นส่วนใหญ่ (คิดเป็นประมาณ 80% ของไวน์ทั้งหมด) รองลงมาไวน์จะมีคุณภาพอยู่ในระดับ 7 มีปริมาณปานกลาง(คิดเป็นประมาณ 12% ของไวน์ทั้งหมด) และ คุณภาพระดับ 3 4 และ 8 มีปริมาณที่ค่อนข้างน้อย (คิดเป็นประมาณ 8% ของไวน์ทั้งหมด) ซึ่งสามารถนำข้อมูลข้างต้นมาดูเปรียบเทียบได้กับเส้นสีแดงควบคู่กัน ถ้าเส้นสีแดงมีอัตราการเปลี่ยนแปลงมาก ก็จะมีการจำนวนคุณภาพของไวน์ในระดับนั้นมาก ถ้ามีอัตราการเปลี่ยนแปลงน้อย ก็จะมีจำนวนคุณภาพของไวน์ในระดับนั้นน้อย

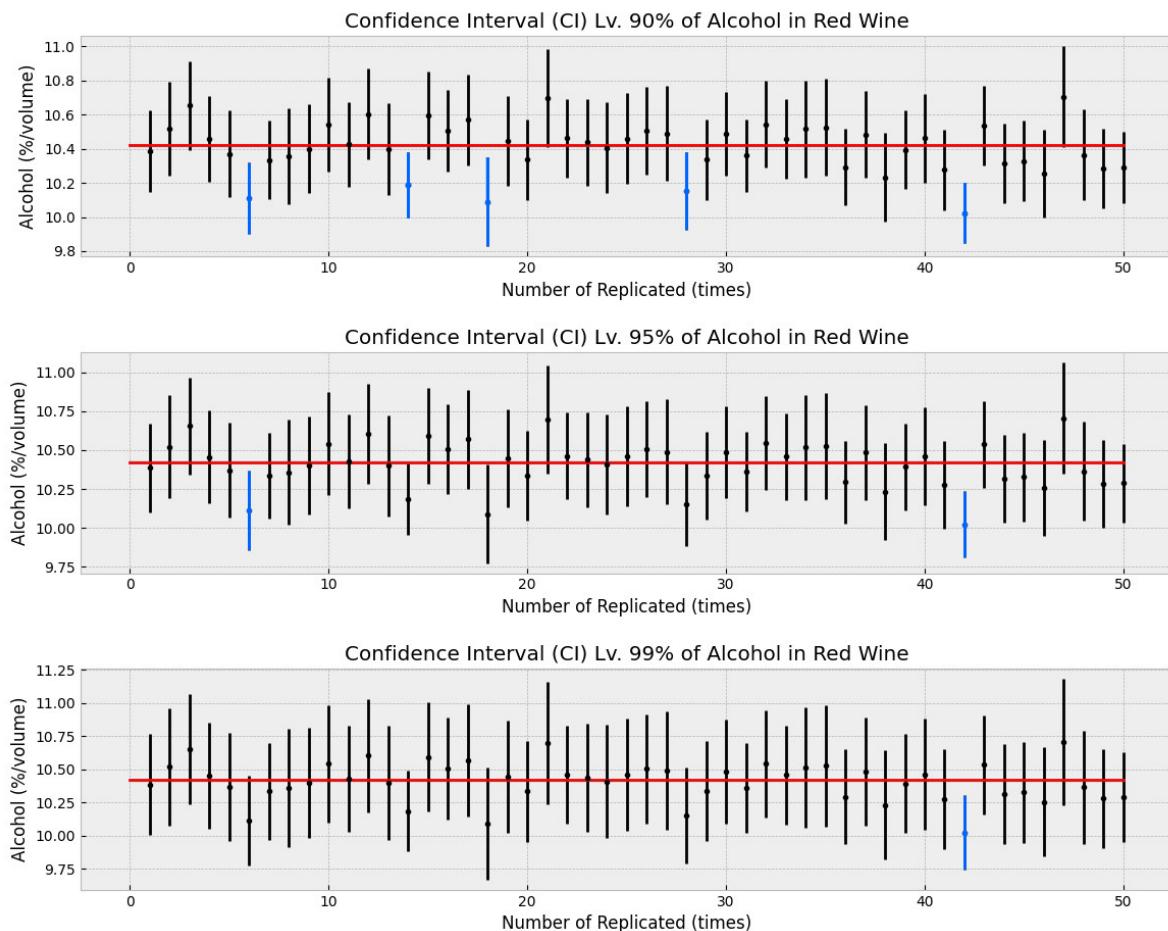
#### 4.1.4) บทวิเคราะห์ภาพรวมของกราฟในข้อที่ 4.1.2 และ 4.1.3

จากการวิเคราะห์ทั้งหมด จะมีข้อสรุปได้ว่า ไวน์ส่วนใหญ่ที่ได้จากการผลิตในความสัมพันธ์ระหว่างปริมาณแอลกอฮอล์ในไวน์และคุณภาพของไวน์ ปริมาณแอลกอฮอล์ในไวน์ช่วงตั้งแต่ประมาณ 9.2 – 10.5 %/volume จะมีไวน์ปริมาณมากที่อยู่ในช่วงนี้ ซึ่งจะมีคุณภาพของไวน์เฉลี่ยอยู่ในระดับ 5 และ 6 เป็นส่วนใหญ่ด้วยเช่นกัน ถือว่าไวน์ส่วนใหญ่ที่มีระดับแอลกอฮอล์ 9.2 – 10.5 %/volume จะมีระดับคุณภาพที่พอใช้-ดี ในขณะที่ปริมาณแอลกอฮอล์ในไวน์ในช่วงตั้งแต่ประมาณ 8.5-9.1 %/volume และ 10.6 – 13.0 %/volume เป็นส่วนที่มีจำนวนไวน์ที่หาได้ยาก หรือมีจำนวนที่ลดลง จะมีคุณภาพของไวน์เฉลี่ยอยู่ในระดับ 7 จึงอาจตีความได้ว่า เมื่อปริมาณแอลกอฮอล์ในไวน์ยิ่งสูงขึ้น จำนวนไวน์จะยิ่งหาได้ยากขึ้นหรือมีจำนวนที่ลดลง คุณภาพของไวน์อาจจะเพิ่มขึ้นไปอยู่ที่ระดับ 7 ซึ่งเป็นคุณภาพในระดับที่ดี-ดีมากส่วนคุณภาพในระดับ 3 4 และ 8 จะไม่นำมาคิด เพราะมีค่าความนำจะเป็นที่น้อยมากๆ จากค่า outlier ของข้อมูล รวมถึงไม่คิดในส่วนปริมาณแอลกอฮอล์ที่มากเกินกว่า 13.0 %/volume ด้วย

ดังนั้น ปริมาณแอลกอฮอล์ในไวน์แดงที่มากขึ้น อาจมีแนวโน้มที่จะทำให้คุณภาพของไวน์แดงเพิ่มขึ้นตามไปด้วย โดยความสัมพันธ์ระหว่างปริมาณแอลกอฮอล์ในไวน์แดง และคุณภาพของไวน์แดงเป็นความสัมพันธ์แบบแปรผันตรงหรือคล้อยตามกัน

อย่างไรก็ตาม ทั้งหมดนี้ยังไม่สามารถกล่าวได้อย่างชัดเจน 100% เป็นเพียงแค่แนวโน้มเท่านั้น เนื่องจากในการผลิตไวน์จริง จะมีส่วนผสมอื่นๆ และมีอีกหลายปัจจัยในการกำหนดคุณภาพของไวน์แดง เช่น ปริมาณกรด, น้ำตาลคงค้างที่เหลือในไวน์แดง, ระยะเวลาการผลิตไวน์แดง, คุณภาพของอุ่นที่นำมาใช้ในการผลิต และ เกณฑ์การวัดคุณภาพของไวน์แดง เป็นต้น ซึ่งเกณฑ์การวัดคุณภาพของไวน์แดง(ระดับ 1-10)ในครั้งนี้ อ้างอิงมาจากโรงงานผลิตไวน์ ในจังหวัด Minho ทางตอนเหนือของประเทศโปรตุเกสเท่านั้น

#### 4.1.5) บทวิเคราะห์กราฟ Confidence Interval (CI) of Mean ปริมาณแอลกอฮอล์ของไวน์แดง



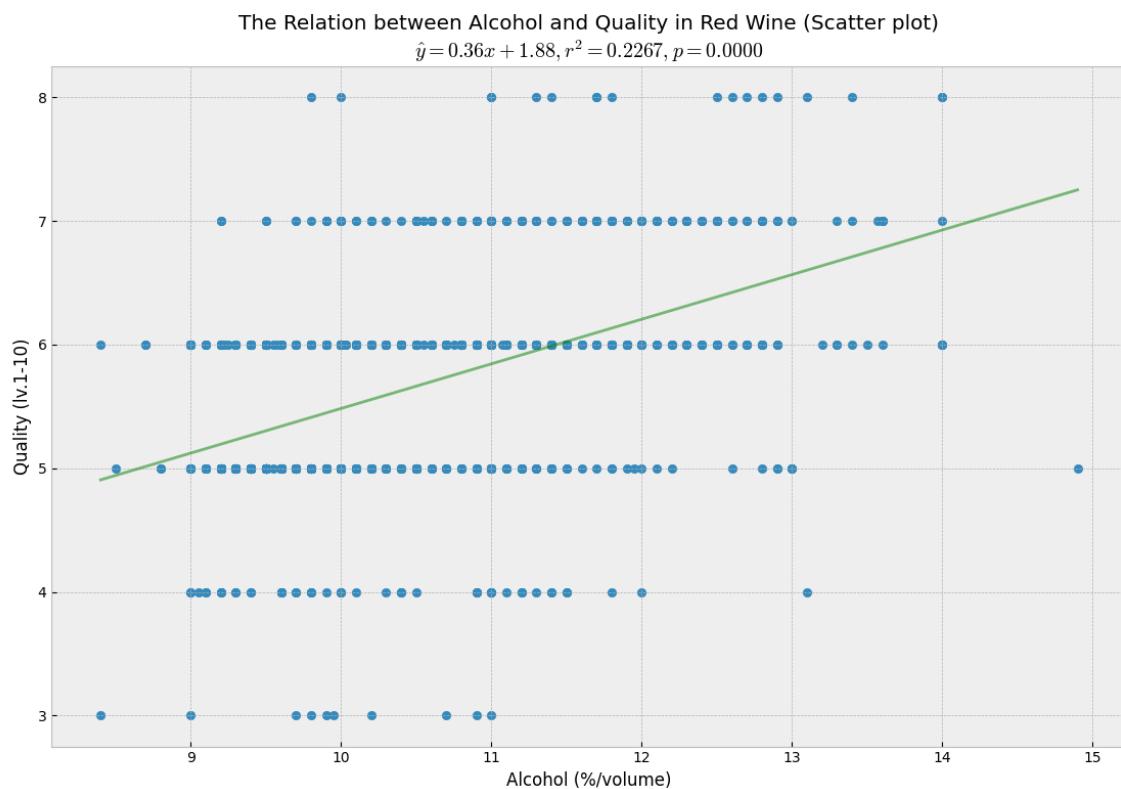
**Population Mean ( $\mu$ ) : 10.42 %/volume**  
**มีการสุ่มทั้งหมด 50 ครั้ง ครั้งละ 50 Sample จาก 1599 ข้อมูล**

- ช่วงระดับความเชื่อมั่น 90% จะมี **45/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 90 %
- ช่วงระดับความเชื่อมั่น 95% จะมี **48/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 96 %
- ช่วงระดับความเชื่อมั่น 99% จะมี **49/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 98 %

เราจึงสามารถสรุปได้ว่า ช่วงระดับความเชื่อมั่นในระดับต่างๆ สามารถอิงไปถึงกลุ่มข้อมูลจริง (Population) ได้ โดยมีระดับความเชื่อมั่น ตามเปอร์เซ็นต์ที่กำหนดไว้ โดยค่าที่ได้ อาจมีความคลาดเคลื่อนจากความเป็นจริงไปบ้างเล็กน้อย ขึ้นอยู่กับจำนวนข้อมูลที่นำมา Sample และ จำนวนรอบการทดลอง ยิ่งมีมาก ก็ยิ่งอ้างอิงถึงประชากรได้ถูกต้องมากขึ้น แต่ยิ่งมีมาก ก็ไม่สะดวกสำหรับการสุ่มตัวอย่างขนาดมาก ๆ เพราะจะเหนื่อยสำหรับผู้เก็บกลุ่มตัวอย่างเอง

การที่มี CI หรือช่วงระดับความเชื่อมั่น ก็เป็นหนึ่งวิธีการคำนวณทางสถิติที่สามารถอ้างอิงถึงจำนวนประชากรหมุ่มากได้โดยสะดวก เมื่อมีจำนวนประชากรที่เยอะมาก ๆ นั้นเอง

#### 4.1.6) บทวิเคราะห์กราฟ Linear Regression



$$y = 0.36x + 1.88, r^2 = 0.2267, \text{ Standard Error} = 0.7104$$

โดย  $r^2 = 0.2267$  แสดงว่า ชุดข้อมูลปริมาณแอลกอฮอล์ และ คุณภาพของไวน์แดง มีความสัมพันธ์กันในแนวโน้มที่ค่อนข้างน้อย อยู่ที่ประมาณ 22.67% เมื่อปริมาณแอลกอฮอล์ เพิ่มขึ้น 1 %/volume จะทำให้ระดับคุณภาพของไวน์แดง เพิ่มขึ้น 0.36 (สังเกตได้จากค่าอัตราการเปลี่ยนแปลงหรือ Slope)

จากการวิเคราะห์กราฟ จะสังเกตเห็นว่า ข้อมูลปริมาณแอลกอฮอล์มีการกระจายตัวอยู่ทั่วไปตามพื้นที่โดยรอบ โดยที่แนวแกน x เป็นตัวแปรต้น หรือปริมาณแอลกอฮอล์ (%/volume) และมีแนวแกน y เป็นตัวแปรตาม หรือคุณภาพของไวน์แดง (ระดับ 1-10) โดยมีเส้นสมการถดถอยเชิงเส้นอย่างง่าย คือเส้นสีเขียว ลากผ่านจากซ้ายไปขวา โดยมีแนวโน้มที่เพิ่มขึ้นเพียงเล็กน้อยเท่านั้น สังเกตได้จากค่าอัตราการเปลี่ยนแปลงเท่ากับ 0.36 และมีค่า  $r^2$  เท่ากับ 0.2267 หรือ 22.67 % ที่เป็นตัวชี้วัดว่า โมเดล Simple Linear Regression สามารถยอมรับได้สำหรับข้อมูลนี้อยู่เพียง 22.67 % เท่านั้น (จดอยู่ในระดับ weak)

กล่าวโดยสรุป ความสัมพันธ์ของปริมาณแอลกอฮอล์ที่ผสมอยู่ในไวน์แดง มีผลต่อคุณภาพของไวน์แดง โดยใช้โมเดล Simple Linear Regression หรือ สมการถดถอยเชิงเส้นอย่างง่าย มีแนวโน้มที่มีความสัมพันธ์ไปในทางที่เล็กน้อย โดยค่าเฉลี่ยของระดับคุณภาพของไวน์แดง อาจเพิ่มขึ้น 0.36 ในทุกๆ 1 %/volume ของปริมาณแอลกอฮอล์ที่เพิ่มขึ้น โดยการประมาณนี้มีความน่าเชื่อถือ หรือยอมรับได้ อยู่ที่ 22.67 %

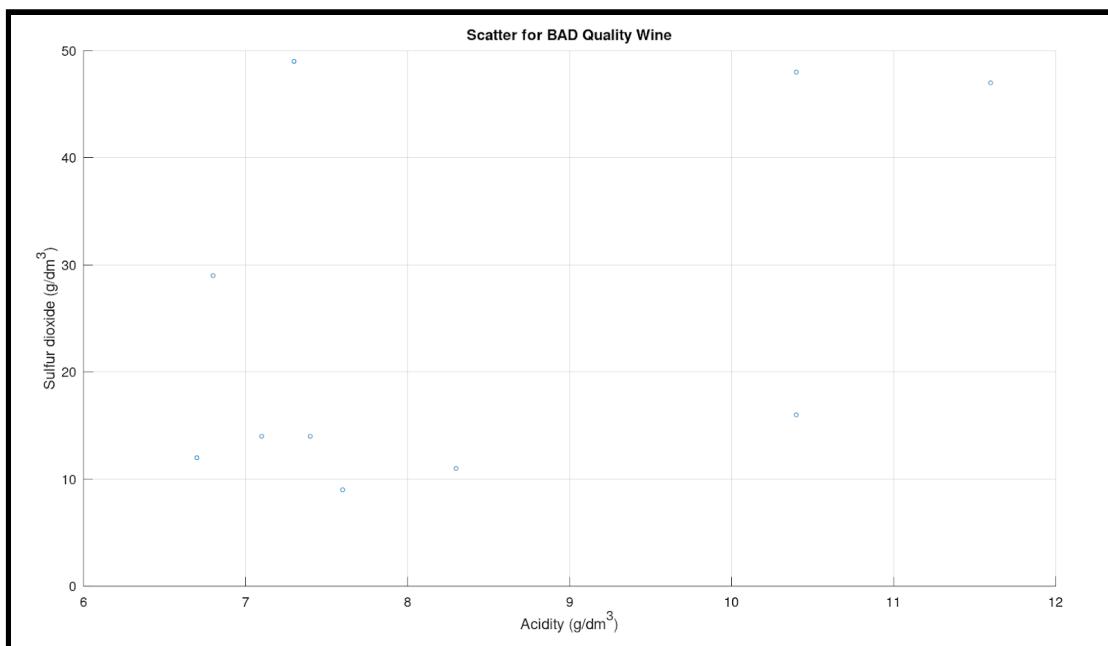
## 4.2 คอลัมน์ที่ 2 ปริมาณความเป็นกรด และ คอลัมน์ที่ Quality

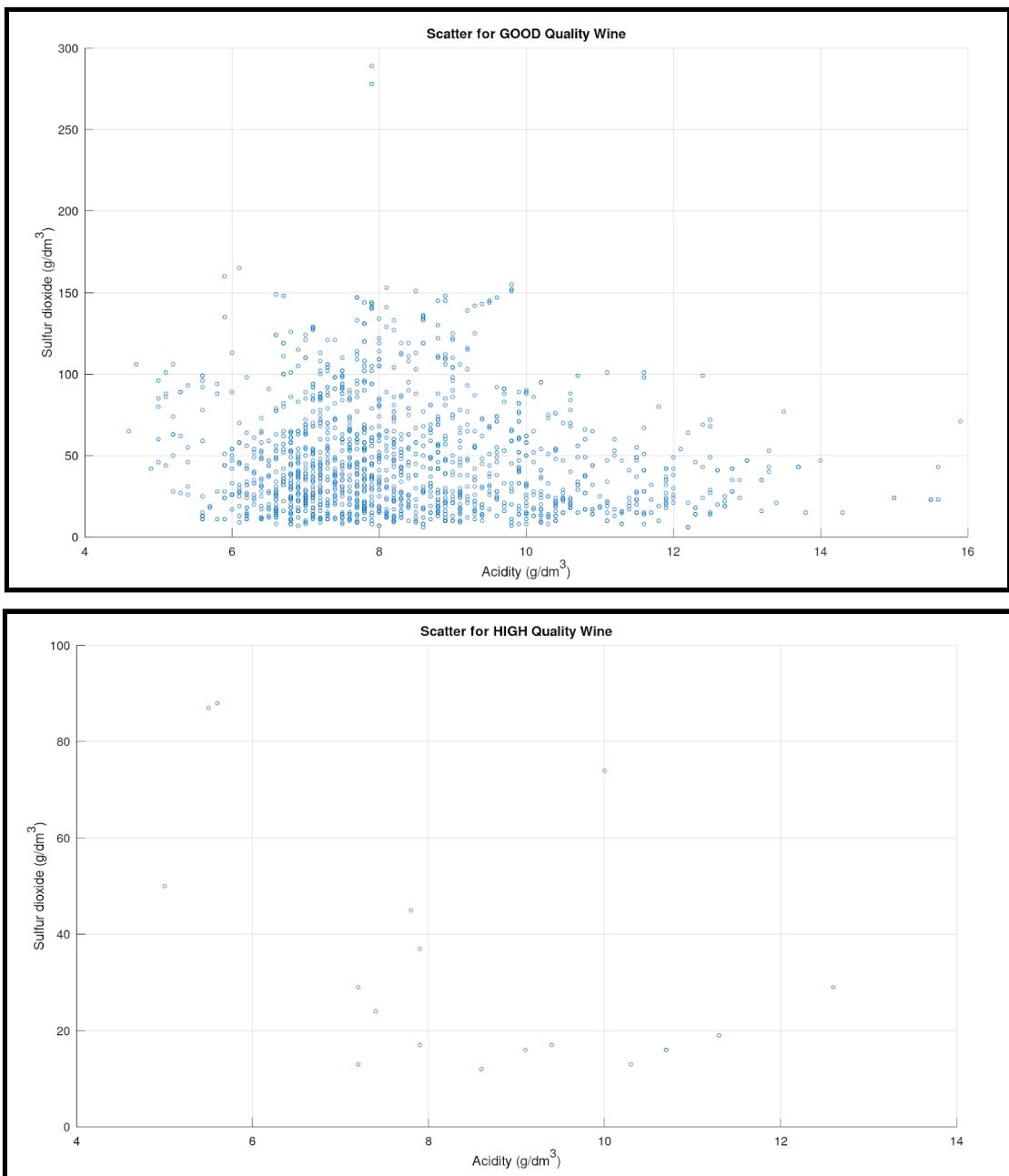
### 4.2.1) บทวิเคราะห์กราฟ Scatter Plot ระหว่างความสัมพันธ์ของปริมาณความเป็นกรด และคุณภาพของไวน์แดง

เพื่อวัดระดับความสัมพันธ์ของปริมาณความเป็นกรดที่มีต่อคุณภาพของไวน์ เราได้แบ่ง ระดับตามความเหมาะสม 3 ระดับดังนี้

| ช่วงค่าของคุณภาพ | ระดับ   |
|------------------|---------|
| 1 - 3            | แย่     |
| 4 - 6            | ปานกลาง |
| 7 - 10           | ดี      |

จากนั้นเราจะแยก Scatter plot ออกเป็น 3 กราฟ โดยแต่ละกราฟบ่งบอกถึง ช่วงค่าของปริมาณ ความเป็นกรด ในไวน์ระดับต่างๆ



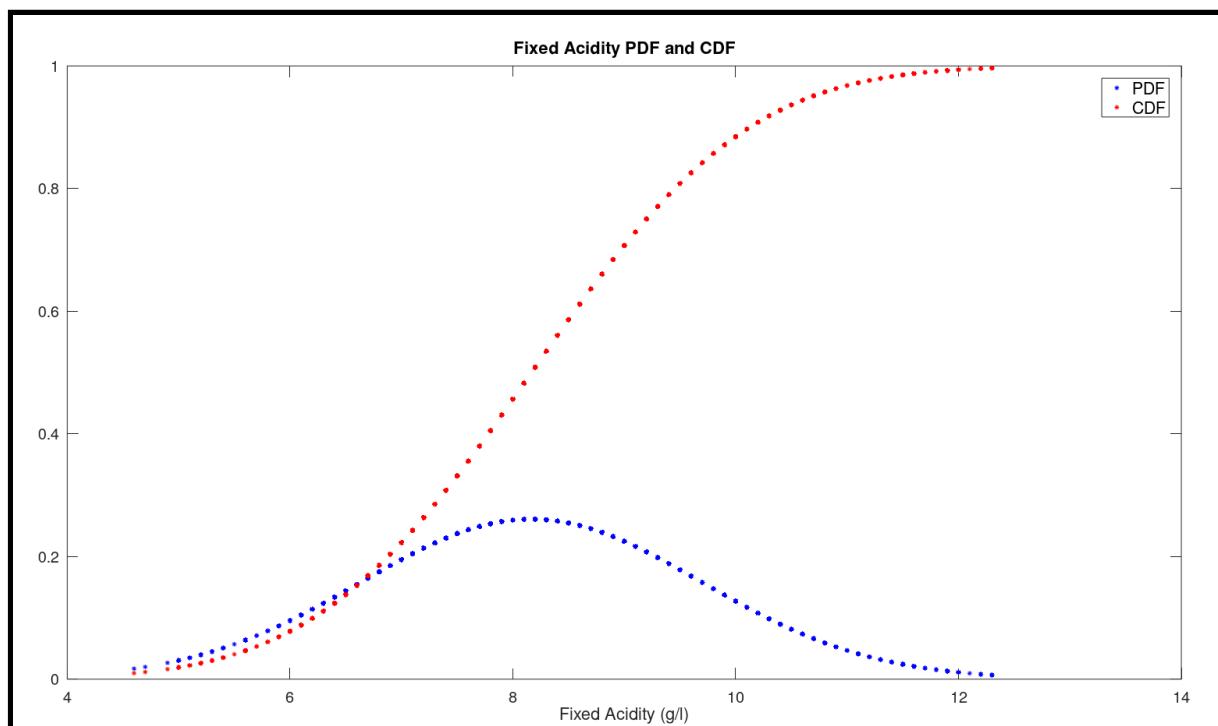


จาก Scatter Plot จะแสดงให้เห็นว่า ช่วงค่าของ Fixed Acidity ที่มีผลทำให้ Wine

- มีคุณภาพที่แข็งคือช่วง 6 - 11 (g / L) โดย Median มีค่า 8.36
- มีคุณภาพปานกลางคือช่วง 4 - 12 (g/L) โดย Median มีค่า 8.1
- มีคุณภาพดีคือช่วง 5 - 12 (g / L) โดย Median มีค่า 8.6

โดยไวน์ตัวอย่าง ส่วนใหญ่จะมีค่าของ Fixed Acidity อยู่ที่ค่าของ 7.20 (g / L) และ ค่าเฉลี่ย Fixed Acidity ที่อยู่ในไวน์จะอยู่ที่ 8.13 ( g / L ) ซึ่งมีค่าอยู่ช่วงปานกลางถึงดี

#### 4.2.2) บทวิเคราะห์กราฟ Probability Density Function (PDF)/ Cumulative Distribution Function (CDF)



#### วิเคราะห์กราฟ

PDF : กราฟ PDF จะมีความเหมือนกับ Bell distribution ซึ่งเป็นสิ่งที่ดี และง่ายต่อการ Normalize เพื่อวิเคราะห์ต่อไป

CDF : กราฟ CDF มี Curve เป็นตัว S สวยงาม สื่อถึงการที่ข้อมูลที่ความ Continuous

#### วิเคราะห์ข้อมูล

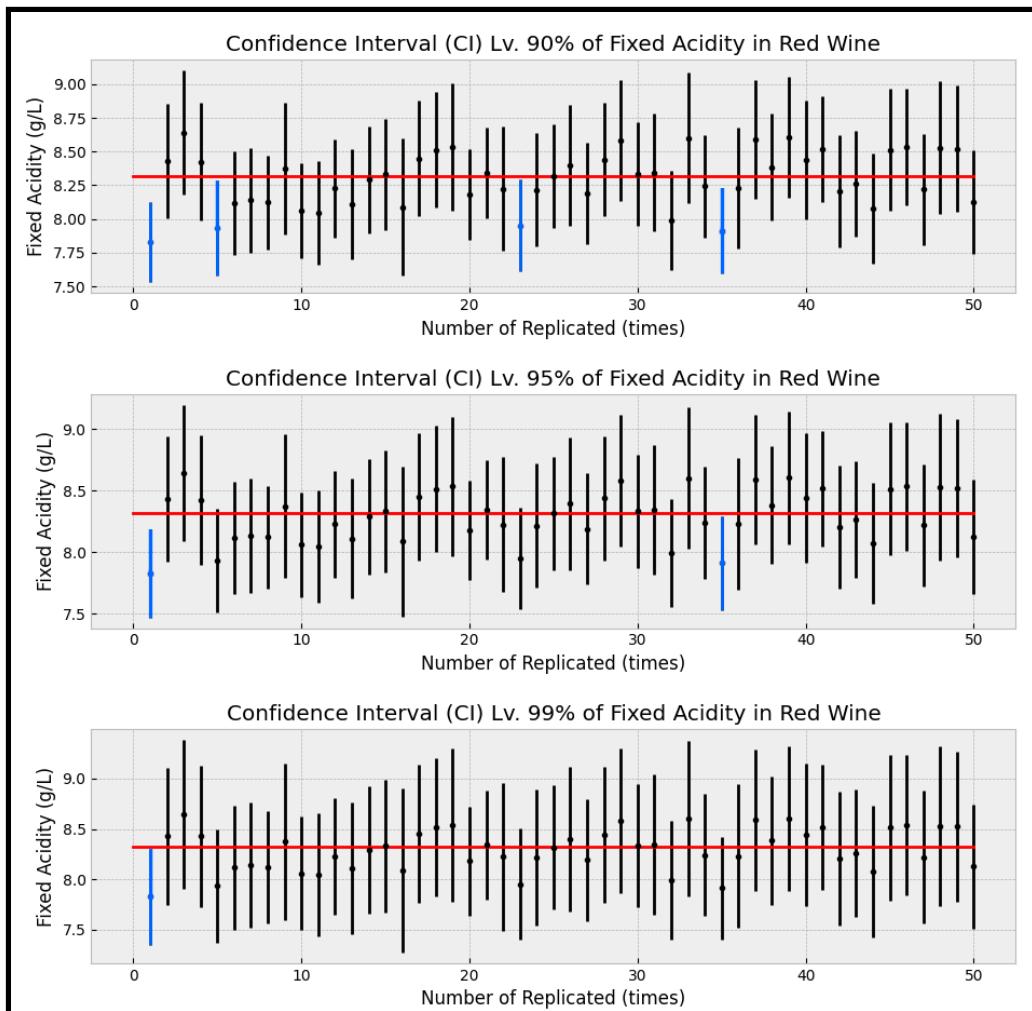
จากราฟ PDF จะหา Tolerance Interval จากจุด Mean มาใช้ในการวิเคราะห์ข้อมูลต่อ แต่เพื่อความง่ายจะใช้ช่วงประมาณ  $\bar{x} - s \leq x \leq \bar{x} + s$  หรือจากราฟก็คือช่วง 6.6 - 10.5 (g / L)

จากราฟ CDF จะเห็นได้ว่ากราฟเริ่มมี Curve ตั้งแต่ช่วงประมาณ 6 - 11 (g / L)

#### สรุป

จากราฟ PDF และ CDF ช่วงค่า 6.6 - 10.5 (g / L) เป็นช่วงค่าที่ส่วนใหญ่ของปริมาณ กรณ์ที่ไวน์แดงมี อ้างอิงจากตารางในข้อ 4.2.1) จะได้ว่าไวน์แดงส่วนใหญ่มีปริมาณกรดอยู่ในช่วง ของไวน์แดงคุณภาพปานกลาง ไม่เยี่ยม และไม่ได้ดี

#### 4.2.3) บทวิเคราะห์กราฟ Confidence Interval (CI) of Mean



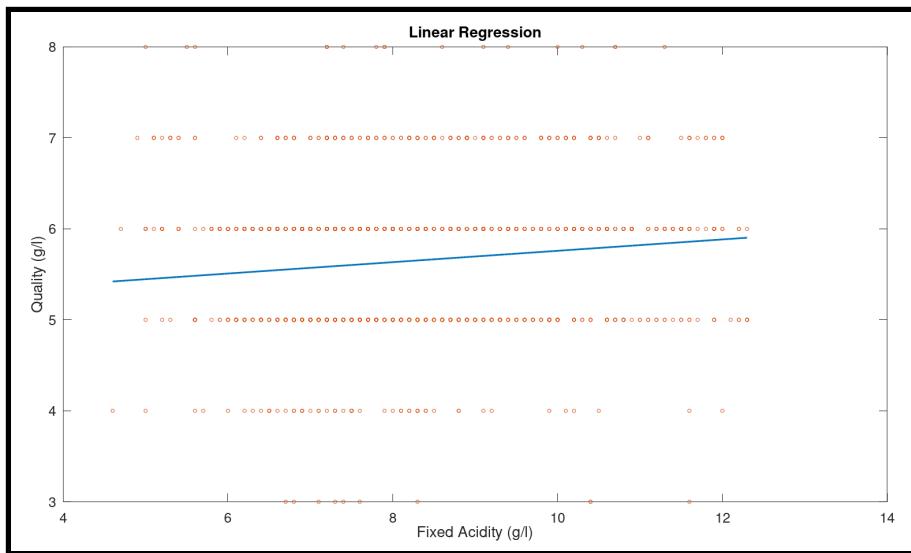
โดยจากกราฟ มีกลุ่มตัวอย่างปริมาณของค่าความเป็นกรดอยู่ทั้งหมด 1599 ข้อมูล เราจะสามารถสรุปได้ดังนี้

**Population Mean ( $\mu$ ) : 8.13 g/L**  
**มีการสุ่มทั้งหมด 50 ครั้ง ครั้งละ 50 Sample จาก 1599 ข้อมูล**

- ช่วงระดับความเชื่อมั่น 90% จะมี **46/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 92 %
- ช่วงระดับความเชื่อมั่น 95% จะมี **48/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 96 %
- ช่วงระดับความเชื่อมั่น 99% จะมี **49/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 98 %

จากผลที่ได้จากการคำนวณ Confidence interval ยิ่ง Confidence Level เพิ่มขึ้นช่วงของ CI จะยิ่งมาก เนื่องจากยิ่งช่วงของค่ามากขึ้นยิ่งเพิ่มโอกาสในการที่ค่า Mean จะอยู่ในช่วงนั้นๆ

#### 4.2.4) บทวิเคราะห์กราฟ Regression (สมการทดถอย)



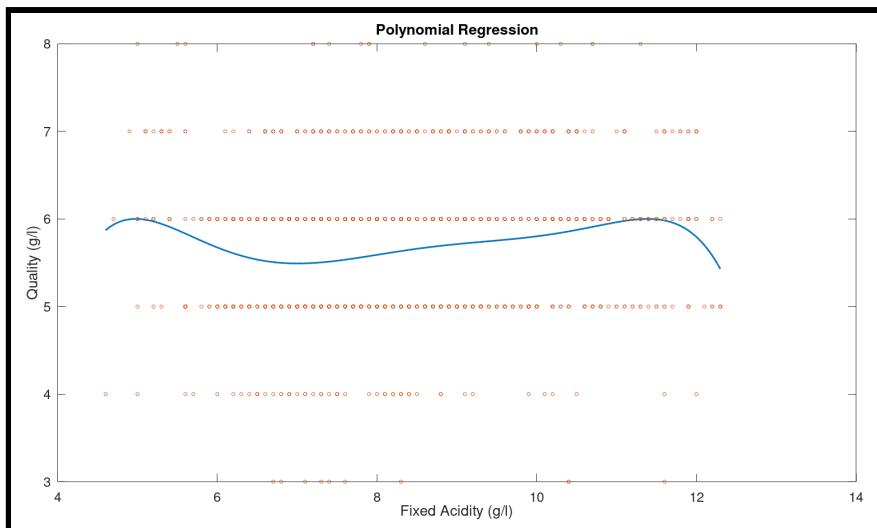
สมการทดถอยเชิงเส้นของกราฟนี้คือ  $y = 0.062x + 5.13$

โดยมี  $r^2 = 0.0320$  และ Standard Error = 0.6392

วิเคราะห์จาก  $r^2$  ที่มีค่า 0.032 หรือ 3% นั้น บ่งบอกว่ากราฟนี้ Weak เป็นอย่างมาก แต่เนื่องจากข้อมูลที่กระจาย และเกากรุ่มเป็นกลุ่มเดียวจึงส่งผลทำให้สมการทดถอยเชิงเส้นนี้มีความ Weak เป็นอย่างมาก แต่ เรา秧งสามารถเห็นแนวโน้มของปริมาณกรดที่ส่งผลต่อคุณภาพของไวน์แดงได้ ถึง Slope จะน้อย แต่บ่งบอกว่าเมื่อความเป็นกรดเพิ่มมากขึ้นจะทำให้คุณภาพของไวน์เพิ่มขึ้น ทีละ 0.062 หน่วย

เพิ่มเติม

เนื่องจากผู้จัดทำได้ทำการ Plot สมการทดถอยพหุนามดูแล้วพบว่า ที่ พหุนามดีกรีเท่ากับ 6 หรือมากกว่า ให้ผลดังกราฟด้านล่าง



จากราฟเราจะเห็น Local Maximum อยู่สองช่วงคือช่วงประมาณ 5 g/l และ 11.5 g/l ซึ่งหมายถึงเป็นช่วงที่เหมาะสมที่สุดที่จะทำให้ Wine มีคุณภาพที่ดีที่สุด

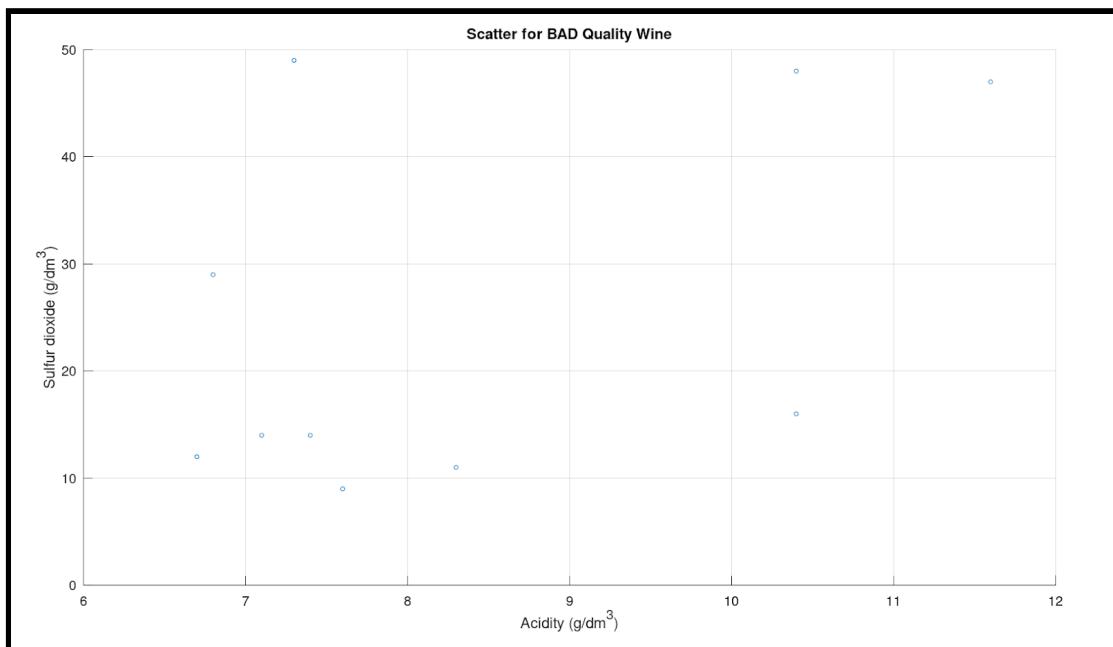
#### 4.3 คอลัมน์ที่ 3 ชัลเพอร์ไ/do開啟ไซด์ และ คอลัมน์ที่ Quality

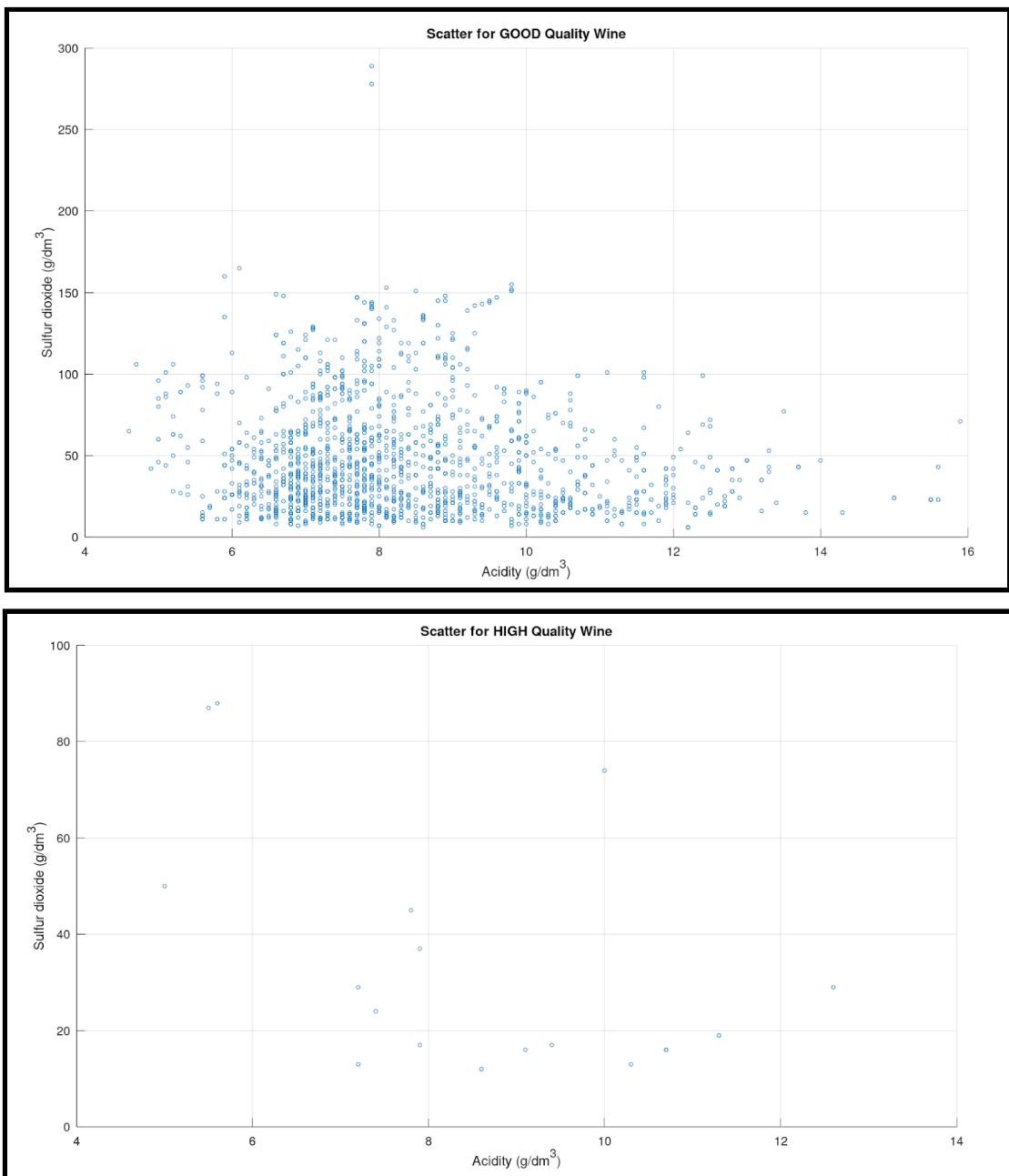
##### 4.3.1) บทวิเคราะห์กราฟ Scatter Plot ระหว่างความสัมพันธ์ของปริมาณ ชัลเพอร์ไ/do開啟ไซด์ และคุณภาพของไวน์แดง

เพื่อวัดระดับความสัมพันธ์ของปริมาณความเป็นกรดที่มีต่อคุณภาพของไวน์ เราได้แบ่งระดับตามความเหมาะสม 3 ระดับดังนี้

| ช่วงค่าของคุณภาพ | ระดับ   |
|------------------|---------|
| 1 - 3            | แย่     |
| 4 - 6            | ปานกลาง |
| 7 - 10           | ดี      |

จากนั้นเราจะแยก Scatter plot ออกเป็น 3 กราฟ โดยแต่ละกราฟบ่งบอกถึง ช่วงค่าของปริมาณความเป็นกรด ในไวน์ระดับต่างๆ



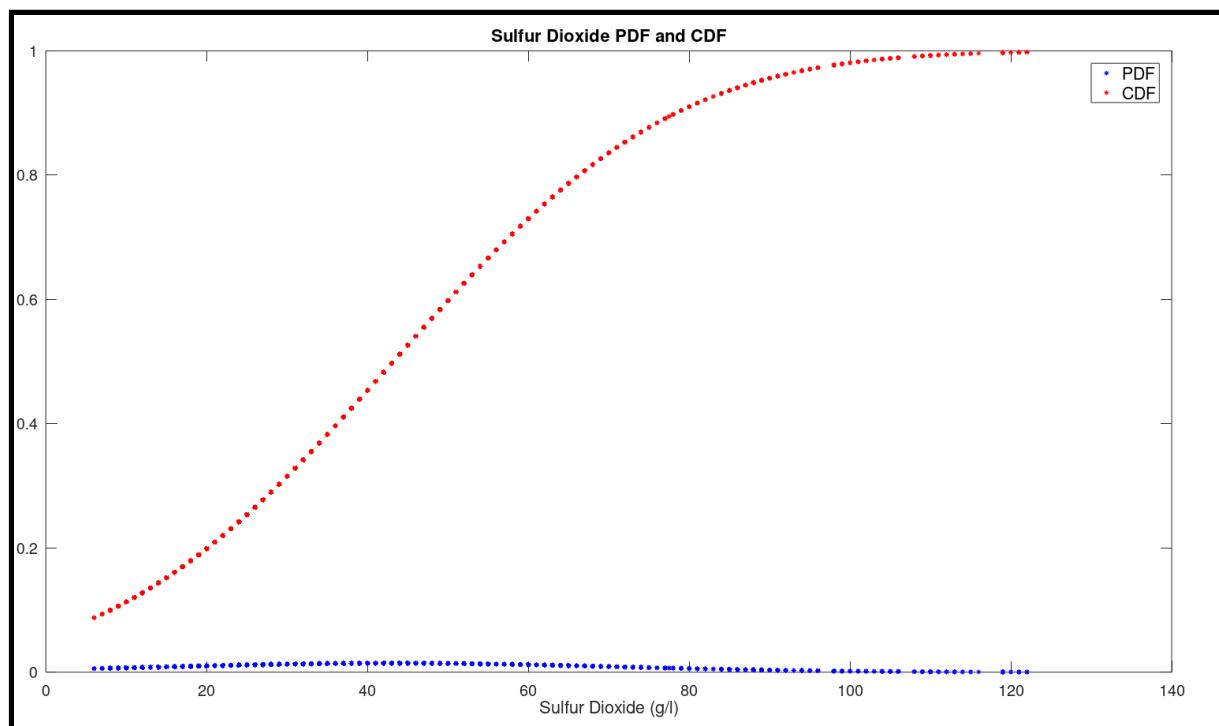


จาก Scatter Plot จะแสดงให้เห็นว่าช่วงค่าของ ซัลเฟอร์ไดออกไซด์ ที่มีผลทำให้ Wine

- มีคุณภาพที่แย่คือช่วง 9 - 49 (g/L) โดย Median มีค่า 24.9
- มีคุณภาพปานกลางคือช่วง 6 - 121 (g/L) โดย Median มีค่า 44.7
- มีคุณภาพดีคือช่วง 7 - 106 (g/L) โดย Median มีค่า 32.5

โดยไวน์ตัวอย่างส่วนใหญ่จะมีค่าของ ซัลเฟอร์ไดออกไซด์ อยู่ที่ค่าของ 28 (g/L) และ ค่าเฉลี่ย Fixed Acidity ที่อยู่ในไวน์จะอยู่ที่ 46.44 ( g/L ) ซึ่งมีค่าอยู่ช่วงแย่

#### 4.3.2) บทวิเคราะห์กราฟ Probability Density Function (PDF)/ Cumulative Distribution Function (CDF)



วิเคราะห์กราฟ :

PDF : กราฟ PDF จะมีการเบี้ปทางขวาเนื่องจาก Mode < Median

CDF : กราฟ CDF มี Curve เกือบเป็นตัว S ที่สวยงาม เพราะข้อมูลแจกแจงหนักไปทางซ้ายมากกว่าทางขวา

วิเคราะห์ข้อมูล :

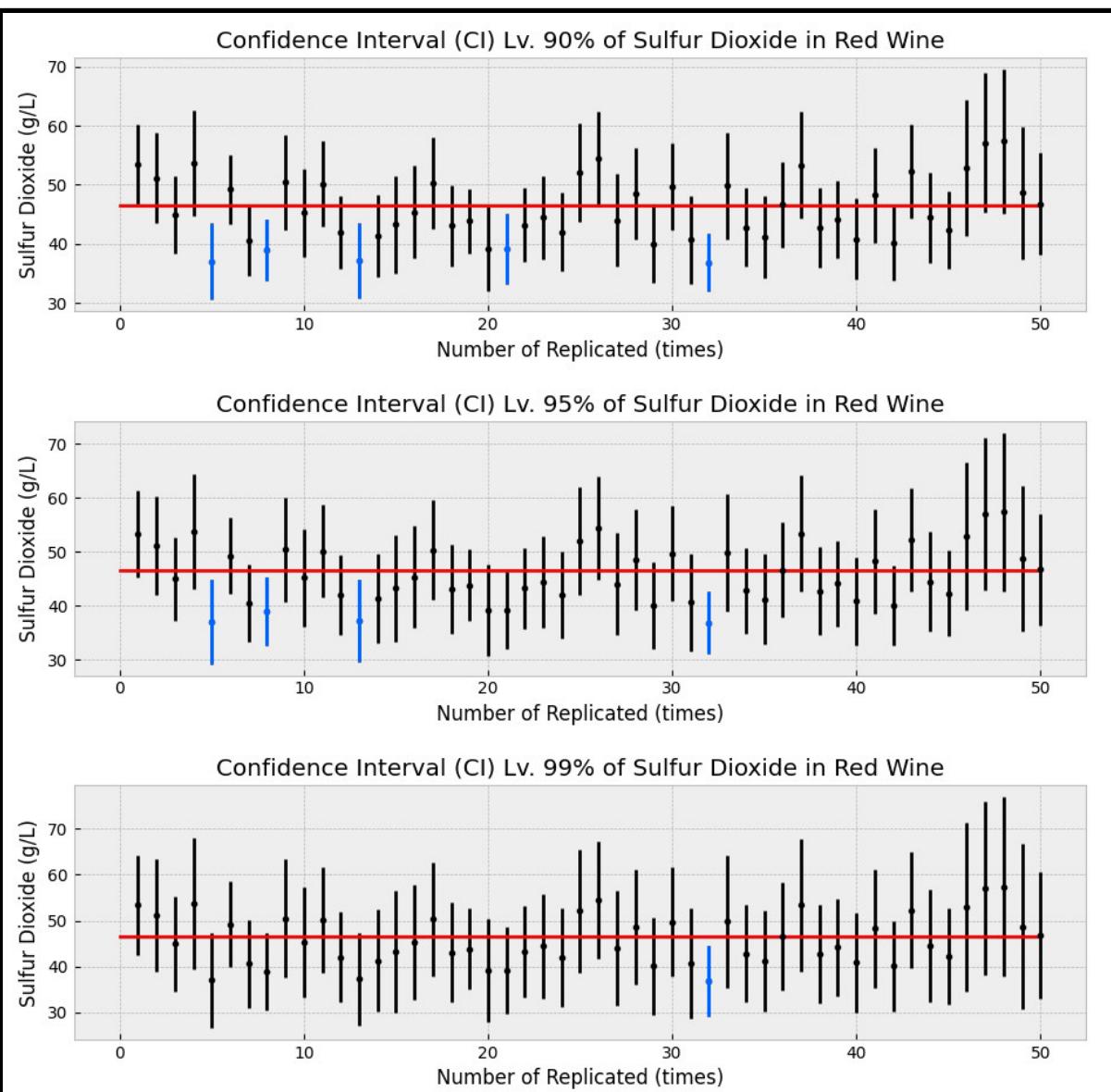
จากราฟ PDF จะหา Tolerance Interval จากจุด Mean มาใช้ในการวิเคราะห์ข้อมูลต่อแต่เพื่อความง่ายจะใช้ช่วงประมาณ  $\bar{x} - s \leq x \leq \bar{x} + s$  หรือจากราฟก็คือช่วง 10 - 50 (g/L)

จากราฟ CDF จะเห็นได้ว่ากราฟเริ่มมี Curve ตั้งแต่ช่วงประมาณ 10 - 80 (g/L)

สรุป

จากราฟ PDF และ CDF ช่วงค่า 10 - 50 (g/L) เป็นช่วงค่าที่ส่วนใหญ่ของปริมาณกรดที่ไวน์แดงมี อ้างอิงจากตารางในข้อ 4.3.1) จะได้ว่าไวน์แดงส่วนใหญ่มีปริมาณ ชัลเฟอร์ไดออกไซด์อยู่ในช่วงของไวน์แดงคุณภาพปานกลาง ไม่เย่ และไม่ได้ดี

#### 4.3.3) บทวิเคราะห์กราฟ Confidence Interval (CI) of Mean



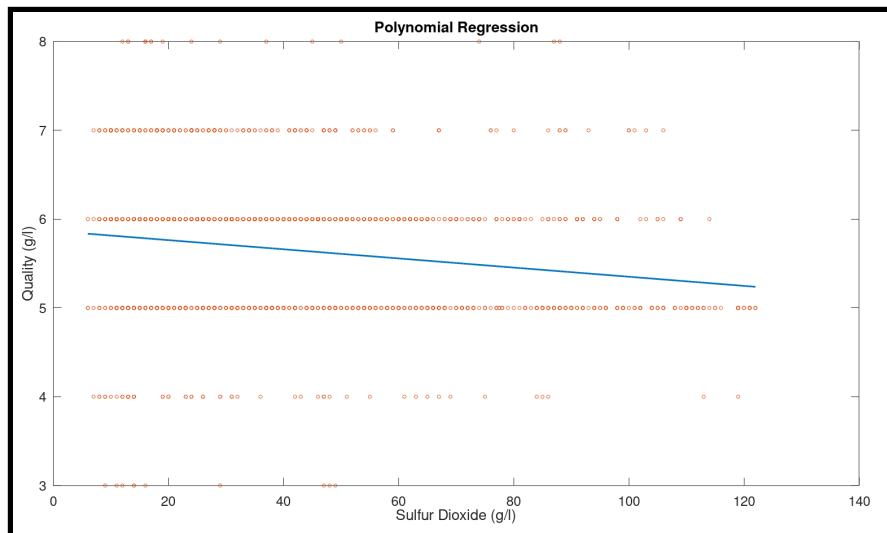
โดยจากการ มีกลุ่มตัวอย่างปริมาณของซัลเฟอร์ไดออกไซด์อยู่ทั้งหมด 1599 ข้อมูล เราจะสามารถสรุปได้ดังนี้

**Population Mean ( $\mu$ ) : 46.47 g/L**  
**มีการสุ่มทั้งหมด 50 ครั้ง ครั้งละ 50 Sample จาก 1599 ข้อมูล**

- ช่วงระดับความเชื่อมั่น 90% จะมี **45/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 90 %
- ช่วงระดับความเชื่อมั่น 95% จะมี **46/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 92 %
- ช่วงระดับความเชื่อมั่น 99% จะมี **49/50** ช่วงความเชื่อมั่นที่เก็บค่า  $\mu$  ไว้ หรือคิดเป็น 98 %

จากผลที่ได้จากการคำนวณ Confidence interval ยิ่ง Confidence Level เพิ่มขึ้นช่วงของ CI จะยิ่งมาก เนื่องจากยิ่งช่วงของค่ามากขึ้นยิ่งเพิ่มโอกาสในการที่ค่า Mean จะอยู่ในช่วงนั้นๆ

#### 4.3.4) บทวิเคราะห์กราฟ Regression (สมการทดถอย)



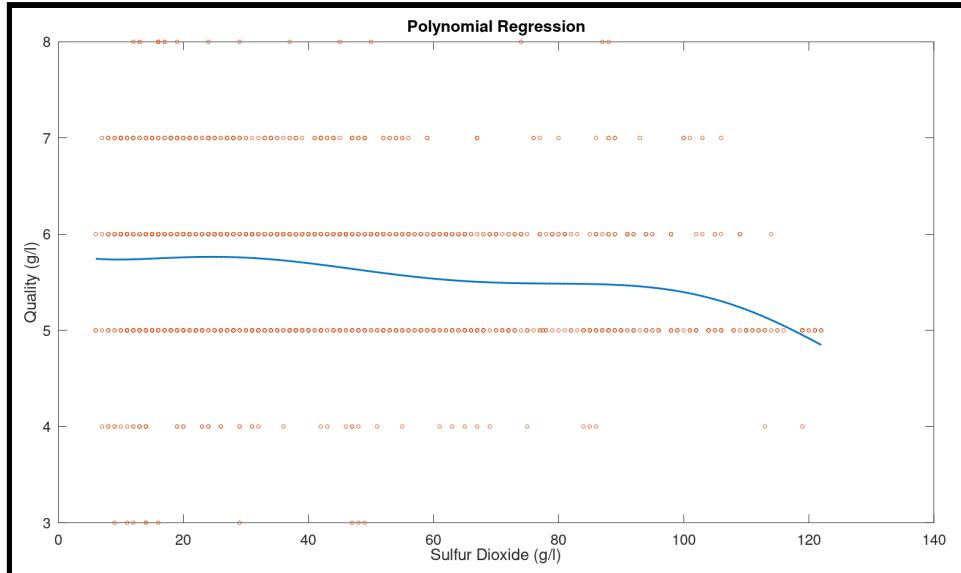
สมการทดถอยเชิงเส้นของกราฟนี้คือ  $y = -0.005x + 5.86$

โดยมี  $r^2 = 0.0304$  และ Standard Error = 0.6382

วิเคราะห์จาก  $r^2$  ที่มีค่า 0.0304 หรือ 3% นั้น บ่งบอกว่ากราฟนี้ Weak เป็นอย่างมาก แต่เนื่องจากข้อมูลที่กระจาย และเกากลุ่มเป็นกลุ่มเดียวจึงส่งผลทำให้สมการทดถอยเชิงเส้นนี้มีความ Weak เป็นอย่างมาก แต่ เรายังสามารถเห็นแนวโน้มของปริมาณกรดที่ส่งผลต่อคุณภาพของไวน์แดงได้ ถึง Slope จะน้อย แต่บ่งบอกว่าเมื่อปริมาณ ซัลเฟอร์ไดออกไซด์ เพิ่มมากขึ้นจะทำให้คุณภาพของไวน์ ลดลง ที่ละ 0.062 หน่วย

เพิ่มเติม

เนื่องจากผู้จัดทำได้ทำการ Plot สมการทดถอยพหุนามดูแล้วพบว่า ที่ พหุนามดีกรีเท่ากับ 6 หรือมากกว่า ให้ผลดังกราฟด้านล่าง



จากการเราจะเห็นเห็นว่ากราฟนั้นมีแนวโน้มที่คุณภาพของไวน์แดงจะลดลงเมื่อปริมาณของ ซัลเฟอร์ไดออกไซด์ ในไวน์แดงนั้นมากขึ้น ซึ่งสอดคล้องกับ สมการทดถอยเชิงเส้น

## บทที่ 5

### สรุปผลการดำเนินงาน

#### สรุปผลการดำเนินงาน

##### 5.1 สรุปผลปริมาณแอลกอฮอล์ที่ส่งผลต่อกลุ่มคนภาพของไวน์แดง

ไวน์ เป็นเครื่องดื่มแอลกอฮอล์ชนิดหนึ่ง ที่ได้จากการหมักผลไม้ชนิดต่างๆ โดยส่วนมากใช้เป็นองุ่นนำมาหมักกับบีสต์ เพื่อทำให้น้ำตาลของผลไม้ และเปลี่ยนแปลงกลไกเป็นแอลกอฮอล์ และกีชาซาร์บอนไดออกไซด์ และจะระเหยเหลือแต่น้ำผลไม้ที่หมักและแอลกอฮอล์ผสมกันอยู่ ซึ่งมีสมออยู่เล็กน้อยเท่านั้น

โดยปกติแล้วปริมาณแอลกอฮอล์ในไวน์จะอยู่ที่ประมาณ 8 -15 % (อ้างอิงจากบทวิเคราะห์ [บทที่ 4.1](#)) และไม่มากเท่าสุรา แต่ได้รับประโยชน์ของผลไม้ และวิตามินต่างๆที่ผลไม้ชนิดนั้นมี หากดื่มน้ำในปริมาณที่จำกัดจะเป็นการบำรุงร่างกาย และเหมาะสมสำหรับงานเลี้ยงทางสังคม งานรื่นเริงต่าง ๆ

ทางผู้จัดทำ ได้นำข้อมูลของปริมาณแอลกอฮอล์ที่มีอยู่ในไวน์แดงนั้น นำมาวิเคราะห์ข้อมูลทางสถิติ เพื่อที่จะเทียบระหว่าง ปริมาณแอลกอฮอล์ และคุณภาพของไวน์แดง ว่าปริมาณแอลกอฮอล์ ควรจะต้องมีมากหรือน้อย เพื่อทำให้คุณภาพของไวน์แดงดียิ่งขึ้น

จากการคำนวณทางสถิติ ในรูปแบบต่างๆ ได้พบวิเคราะห์สรุปได้ดังนี้

“ ปริมาณแอลกอฮอล์ที่มากขึ้น (หน่วยเป็น %/volume) จะมีแนวโน้มที่ทำให้คุณภาพของไวน์แดง เพิ่มขึ้น เพียงเล็กน้อยเท่านั้น ” (อ้างอิงจากบทวิเคราะห์ [บทที่ 4.1](#))

ทางผู้จัดทำจึงขอสรุปว่า ไวน์โดยทั่วไป จะมีปริมาณแอลกอฮอล์อยู่ที่ประมาณ 8 - 15 % โดยจะเกิดระหว่างการหมักของผลไม้ โดยยิ่งมีปริมาณแอลกอฮอล์ที่เพิ่มมากขึ้น อาจมีแนวโน้มที่ทำให้คุณภาพของไวน์แดงเพิ่มขึ้นเล็กน้อย

## 5.2 สรุปผลปริมาณของกรดที่ส่งผลต่อคุณภาพของไวน์แดง

ปริมาณกรด ( acidity ) เป็นตัวแปรสำคัญที่จะส่งผลต่อรสชาติของไวน์ หรือหมายถึงค่าความเปรี้ยวและฝาดในไวน์ มีหน่วยเป็น กรัม / ลิตร หากความสามารถหาช่วงค่าที่เหมาะสมของปริมาณกรดของไวน์แดง เราจะสามารถสรุปได้ว่ารสชาติไวน์แดงแบบใดคือรสชาติของไวน์คุณภาพดี

ในการสรุปผลครั้งนี้เราจะอ้างอิงจากผลการวิเคราะห์จากบทที่ [4.2.4](#) ) บทวิเคราะห์กราฟ Regression (สมการถดถอย) เป็นหลัก โดย ผลการวิเคราะห์จากข้อดังกล่าวได้ผลดังนี้

จากราฟการถดถอยเชิงเส้น จะสังเกตเห็นแนวโน้มของปริมาณกรดที่ส่งผลต่อคุณภาพของไวน์แดงได้ ถึง Slope จะน้อย แต่บ่งบอกว่าเมื่อความเป็นกรดเพิ่มมากขึ้นจะทำให้คุณภาพของไวน์ เพิ่มขึ้น ทีละ 0.062 หน่วย

อ้างอิงจากการฟิล์มการถดถอยพหุนามจากบทที่ [4.2.4](#)) ที่ได้ผลวิเคราะห์ว่า มี Local Maximum อยู่สองช่วงคือช่วงประมาณ 5 g/l และ 11.5 g/l ซึ่งหมายถึงเป็นช่วงที่เหมาะสมที่สุดที่จะทำให้ Wine มีคุณภาพที่ดีที่สุด และอ้างอิงจากผลการดำเนินการบทที่ [4.2.1](#)) ที่สรุปช่วงค่าของความเป็นกรดที่ส่งผลให้ไวน์มีคุณภาพที่มีคุณภาพดีคือช่วง 5 - 12 (g / L) โดย Median มีค่า 8.6 ได้ข้อสรุปว่าปริมาณกรดที่เหมาะสมที่จะนำไปรับใช้ในการหมักไวน์แดงคือ 5 g/l และ 11.5 g/l

### 5.3 สรุปผลปริมาณซัลเฟอร์ไดออกไซด์ที่ส่งผลต่อคุณภาพของไวน์แดง

ปริมาณซัลเฟอร์ไดออกไซด์ในไวน์ ส่งผลต่อความ สดใหม่ และกลิ่นของไวน์ ซึ่งในทางกลับกัน อาจจะส่งผลเสียต่อร่างกายที่มีปริมาณมากจนเกินไป มีหน่วยเป็น กรัม / ลิตร หากเราสามารถหาช่วงค่าที่เหมาะสมของปริมาณ ซัลเฟอร์ไดออกไซด์ เราจะสามารถสรุปได้ว่าความ สดใหม่ และ กลิ่นของไวน์แดงแบบใดถึงจะทำให้ไวน์คุณภาพดี

ในการสรุปผลครั้งนี้เราจะอ้างอิงจากผลการวิเคราะห์จากบทที่ [4.3.4](#) ) บทวิเคราะห์กราฟ Regression (สมการถดถอย) เป็นหลัก โดย ผลการวิเคราะห์จากข้อดังกล่าวได้ผลดังนี้

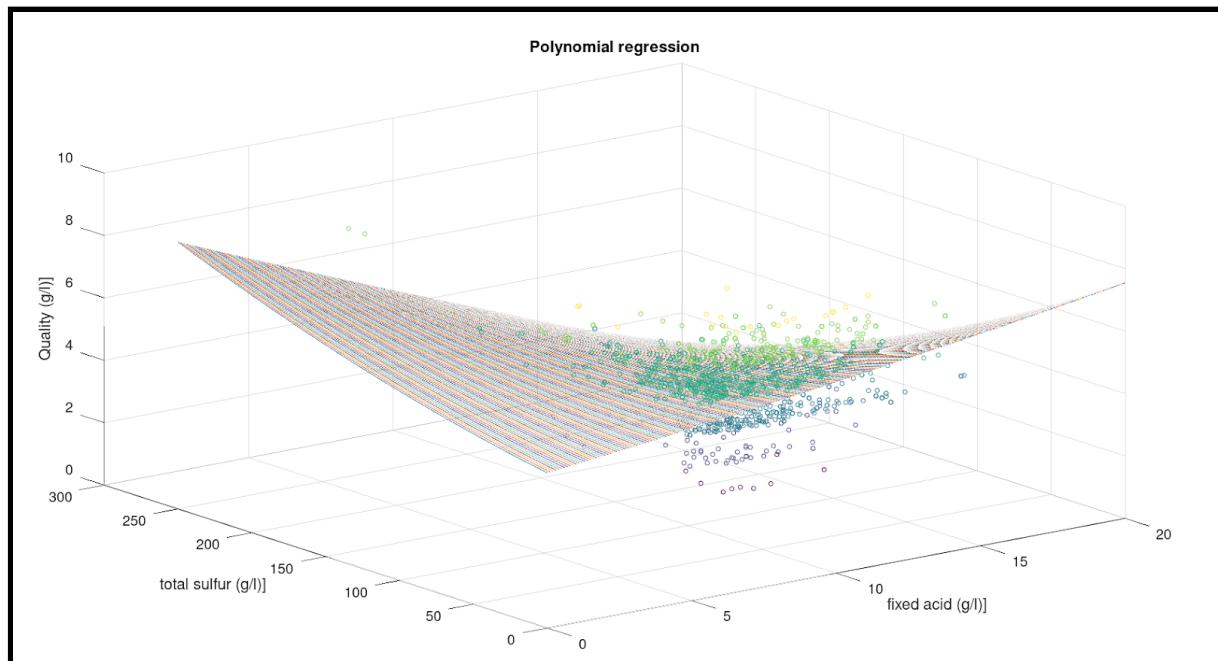
จากราฟการถดถอยเชิงเส้น จะสังเกตเห็นแนวโน้มของปริมาณกรดที่ส่งผลต่อคุณภาพ ของไวน์แดงได้ ถึง Slope จะน้อย แต่บวกกว่าเมื่อปริมาณ ซัลเฟอร์ไดออกไซด์ เพิ่มมากขึ้นจะ ทำให้คุณภาพของไวน์ ลดลง ที่ละ 0.062 หน่วย

อ้างอิงจากการสมการถดถอยพหุนามจากบทที่ [4.3.4](#)) ที่ให้ผลสอดคล้องกับสมการถดถอยเชิงเส้น และอ้างอิงจากผลการดำเนินการบทที่ [4.3.1](#)) ที่สรุปช่วงค่าของปริมาณซัลเฟอร์ไดออกไซด์ ที่ส่งผลให้ไวน์มีคุณภาพที่มีคุณภาพดีคือช่วง 7 - 106 (g/ L) โดย Median มีค่า 32.5 อ้างอิงจากได้ข้อสรุปว่าปริมาณซัลเฟอร์ไดออกไซด์ที่เหมาะสมที่จะนำไปใช้ในการหมักไวน์ แดงคือช่วง 7 - 20 g / L

อย่างไรก็ตาม ข้อมูลต่างๆที่เรานำมาวิเคราะห์นั้นเป็นเพียงปัจจัยหนึ่งของการกำหนด คุณภาพของไวน์ ซึ่ง ในการผลิตไวน์จริง จะมีส่วนผสมอื่นๆ และมีอีกหลายปัจจัยในการกำหนด คุณภาพของไวน์แดง เช่น น้ำตาลคงค้างที่เหลือในไวน์แดง, ระยะเวลาการผลิตไวน์แดง, คุณภาพขององุ่นที่นำมาใช้ในการผลิต และ เกณฑ์การวัดคุณภาพของไวน์แดง และอื่นๆอีก มากมาย ซึ่งเกณฑ์การวัดคุณภาพของไวน์แดง (ระดับ 1-10) ในครั้งนี้ อ้างอิงมาจากโรงงานผลิต ไวน์ ในจังหวัด Minho ทางตอนเหนือของประเทศโปรตุเกสเท่านั้น

## ข้อเสนอแนะแนวทางการศึกษาเพิ่มเติม

- เพิ่มเติมการวิเคราะห์จากสมการถดถอยเชิงเส้น เป็นหลายมิติ ตัวอย่างเช่น สมการถดถอยพหุนาม 3 มิติ ของคอลัมน์ปริมาณกรด และ คอลัมน์ปริมาณซัลเฟอร์ไดออกไซด์ ทำให้เห็นภาพได้ชัดเจนขึ้น



ถ้าสังเกตจากราฟจะสามารถเห็นความสัมพันธ์ของปริมาณกรด และ ปริมาณซัลเฟอร์ไดออกไซด์ ณ ช่วงค่าต่างๆที่ส่งผลต่อคุณภาพของไวน์ ที่ลึกและมีมิติมากขึ้น ส่งผลให้การวิเคราะห์ข้อมูลต่างๆสามารถทำได้ลึก และระเอียดขึ้น เช่นเดียวกัน

## บรรณานุกรม

- ที่มาของชุดข้อมูล Winequality-red.csv

[Red Wine Quality | Kaggle](#)

- ที่มาคำอธิบายแต่ละส่วนประกอบของไวน์

[Wine Quality Data Set](#)

[Red and White Wine Quality](#)

(TH) รู้ไหมว่า...ระดับปริมาณแอลกอฮอล์ในไวน์มีกี่เปอร์เซ็นต์ ?

- วิธีการทำไวน์

[How Wine Is Made](#)

- ประเพณีของไวน์

[UNLOCK WINE 101: รวมทุกเรื่อง “ไวน์” เข้าใจง่ายที่ผู้ชายต้องรู้เพื่อความคุ้ลและมีระดับ](#)  
[Wine101 สอนดื่มไวน์ ซื้อไวน์ดีๆ เลือกไวน์อย่างไร](#)

- รายละเอียดอื่นๆ เกี่ยวกับคุณภาพและวิธีรับรสที่ดีของไวน์

[\[Aftertaste\] ไวน์ดราught เป็นอย่างไร](#)

[The Special Technique for Tasting Wine - dummies](#)

[4 Ways to Know if Your Wine Is Good](#)

## ภาคผนวก

## รายละเอียด Source Code และ OUTPUT ของโปรแกรมทั้งหมด

### 1. wineGraph.py

#### 1.1. Source Code wineGraph.py

```
◆ wineGraph.py X ◆ wineGraph.py (Working Tree)
◆ wineGraph.py > ...
You, 3 hours ago | 1 author (You)
1 import matplotlib.pyplot as plt      # plot graphs
2 import pandas                      # collection for data
3 import stemgraphic as stm          # stem-leaf graphs
4 import statistics as stc           # statistics
5
6 # Init Style of Graph and Insert table of data in form of columns
7 plt.style.use('bmh')
8 columns = pandas.read_csv('testgraphredwine.csv')
9
10 # All columns
11 x = columns['alcohol']            # x (independent variable) = alcohol
12 y = columns['quality']           # y (dependent variable) = quality
13
14 ...
15     # Calculations
16     1. Mean
17     2. Median
18     3. Mode
19     4. Sample Standard Deviation
20     5. Variance
21
22     Independent
23         1. Alcohol
24             2. Residual sugar
25     Dependent
26         1. Quality
27 ...
28
29 print("All Statistics")
30 print('-----Alcohol in Wines-----')
31
32 # Alcohol calculations
33 alcoholMin = min(x)
34 alcoholMean = stc.mean(x)
35 alcoholMed = stc.median(x)
36 alcoholMax = max(x)
37 alcoholMode = stc.mode(x)
38 alcoholSampleSD = stc.stdev(x)
39 alcoholSampleV = stc.variance(x)
40
41 print("alcohol Unit: %/volume")
42 print("alcohol Min", alcoholMin)
43 print("alcohol Mean :", alcoholMean)
44 print("alcohol Median :", alcoholMed)
45 print("alcohol Max", alcoholMax)
46 print("alcohol Mode :", alcoholMode)
47 print("alcohol Sample Standard Deviation :", alcoholSampleSD)
48 print("alcohol Sample Variance :", alcoholSampleV)
49
50 # Alcohol Outlier
51 aQt = stc.quantiles(x, method='inclusive')
52 print("Alcohol[Q1, Q2, Q3]= ", aQt)
53 al_q1 = aQt[0]
54 al_q3 = aQt[2]
55 al_iqr = al_q3 - al_q1
56 print("IQR = ", al_iqr)
57 al_mild_low_bound = al_q1 - al_iqr*1.5
58 al_extreme_low_bound = al_q1 - al_iqr*3
59 al_mild_up_bound = al_q3 + al_iqr*1.5
60 al_extreme_up_bound = al_q3 + al_iqr*3
```

```

61   print('\n-----All Alcohol Outliers Boundaries-----')
62   print("al_extreme_low_bound = ", al_extreme_low_bound)
63   print("al_mild_low_bound = ", al_mild_low_bound)
64   print("al_mild_up_bound = ", al_mild_up_bound)
65   print("al_extreme_up_bound = ", al_extreme_up_bound)
66
67
68   al_extreme_low = []
69   al_mild_low = []
70   al_mild_up = []
71   al_extreme_up = []
72
73   for i in x:
74       if i < al_extreme_low_bound:
75           al_extreme_low.append(i)
76       elif al_extreme_low_bound <= i < al_mild_low_bound:
77           al_mild_low.append(i)
78       elif al_mild_up_bound < i <= al_extreme_up_bound:
79           al_mild_up.append(i)
80       elif i >= al_extreme_up_bound:
81           al_extreme_up.append(i)
82
83   print('\n-----All Alcohol Outliers-----')
84   print("Extreme Outlier(Lower) = ", al_extreme_low)
85   print("Mild Outlier(Lower) = ", al_mild_low)
86   print("Mild Outlier(Upper) = ", al_mild_up)
87   print("Extreme Outlier(Upper) = ", al_extreme_up)
88
89
90
91
92   print('\n\n-----Quality of Wines-----')
93   # Quality calculations
94   qualityMin = min(y)
95   qualityMean = stc.mean(y)
96   qualityMed = stc.median(y)
97   qualityMax = max(y)
98   qualityMode = stc.mode(y)
99   qualitySampleSD = stc.stdev(y)
100  qualitySampleV = stc.variance(y)
101
102
103  print("\nquality Unit: None (lv.1-10)")
104  print("quality Min",qualityMin)
105  print("quality Mean :", qualityMean)
106  print("quality Median :", qualityMed)
107  print("quality Max",qualityMax)
108  print("quality Mode :", qualityMode)
109  print("quality Sample Standard Deviation :", qualitySampleSD)
110  print("quality Sample Variance :", qualitySampleV)
111
112
113  # Quality Outlier
114  qQt = stc.quantiles(y, method='inclusive')
115  print("Quality[Q1, Q2, Q3]= ",qQt)
116  qu_q1 = qQt[0]
117  qu_q3 = qQt[2]
118  qu_iqr = qu_q3 - qu_q1
119  print("IQR = ", qu_iqr)
120  qu_mild_low_bound = qu_q1 - qu_iqr*1.5
121  qu_extreme_low_bound = qu_q1 - qu_iqr*3
122  qu_mild_up_bound = qu_q3 + qu_iqr*1.5
123  qu_extreme_up_bound = qu_q3 + qu_iqr*3
124
125  print('\n-----All Quality Outliers Boundaries-----')
126  print("qu_extreme_low_bound = ", qu_extreme_low_bound)
127  print("qu_mild_low_bound = ", qu_mild_low_bound)
128  print("qu_mild_up_bound = ", qu_mild_up_bound)
129  print("qu_extreme_up_bound = ", qu_extreme_up_bound)

```

```

130     qu_extreme_low = []
131     qu_mild_low = []
132     qu_mild_up = []
133     qu_extreme_up = []
134
135     for i in y:
136         if i < qu_extreme_low_bound:
137             qu_extreme_low.append(i)
138         elif qu_extreme_low_bound <= i < qu_mild_low_bound:
139             qu_mild_low.append(i)
140         elif qu_mild_up_bound < i <= qu_extreme_up_bound:
141             qu_mild_up.append(i)
142         elif i >= qu_extreme_up_bound:
143             qu_extreme_up.append(i)
144
145     print('\n-----All Quality Outliers-----')
146     print("Extreme Outlier(Lower) = ", qu_extreme_low)
147     print("Mild Outlier(Lower) = ", qu_mild_low)
148     print("Mild Outlier(Upper) = ", qu_mild_up)
149     print("Extreme Outlier(Upper) = ", qu_extreme_up)
150
151
152
153     ...
154     #Graphs
155     1. Histogram
156     2. Box Plot
157     3. Stem and Leave
158     4. XY (Scatter) Plot (suitable variable)(describe more)
159
160     Detail
161     1. Name of Graph
162     2. Name of Axis
163     3. Suitable variable
164     4. Identify Outlier
165     ...
166
167
168     # Scatter Plot
169     figure, scat = plt.subplots(figsize=(12, 8))
170     plt.tight_layout(pad=4)
171     scat.set_title('The Relation between Alcohol and Quality in Red Wine (Scatter plot)')
172     scat.set_xlabel('Alcohol (%/volume)') #independetnd
173     scat.set_ylabel('Quality (lv.1-10)') #dependent
174     scat.scatter(x, y)
175
176     # Histogram
177     figure, his = plt.subplots(1,2, figsize=(14, 5))
178     plt.tight_layout(pad=4, w_pad=6, h_pad=1.0)
179     his[0].set_title('Alcohol in Red Wine (Histogram)')
180     his[0].set_xlabel("Alcohol (%/volume)")
181     his[0].set_ylabel("Amount of Red Wine (Bottles)")
182     his[0].hist(x,range(8,16)) [You, seconds ago * Uncommitted changes]
183     his[1].set_title('Quality of Red Wine (Histogram)')
184     his[1].set_xlabel("Quality (lv.1-10)")
185     his[1].set_ylabel("Amount of Red Wine (Bottles)")
186     his[1].hist(y,range(1,11))
187
188     # Box Plot
189     figure, box = plt.subplots(1, 2, figsize=(14, 5))
190     plt.tight_layout(pad=4, w_pad=3, h_pad=1.0)
191     box[0].set_title('Alcohol in Red Wine (Box Plot)')
192     box[0].set_xlabel("Alcohol (%/volume)")
193     box[0].boxplot(x, vert=False, widths=0.3)
194     box[1].set_title('Quality of Red Wine (Box Plot)')
195     box[1].set_xlabel("Quality (lv.1-10)")
196     box[1].boxplot(y, vert=False, widths=0.3)
197
198     # Stem and Leaf
199     figure, stem = stm.graphic.stem_graphic(x, scale=0.1, leaf_order=1, aggregation=False, unit='%/volume', display=3000, compact=True)
200     stem.set_title("Alcohol in Red Wine Stem-And-Leaf")
201
202     figure, stem = stm.graphic.stem_graphic(y, scale=1.0, leaf_order=1, aggregation=True, display=3000)
203     stem.set_title("Quality of Red Wine Stem-And-Leaf")
204
205     # Show
206     plt.show()

```

## 1.2. Output wineGraph.py

```
[Running] python -u "c:\Users\ASUS\Desktop\Prob-stat\wineGraph.py"
All Statistics
-----Alcohol in Wines-----
alcohol Unit: %volume
alcohol Min 8.4
alcohol Mean : 10.422983114446529
alcohol Median : 10.2
alcohol Max 14.9
alcohol Mode : 9.5
alcohol Sample Standard Deviation : 1.0656771926520383
alcohol Sample Variance : 1.1356678789387298
Alcohol[Q1, Q2, Q3]= [9.5, 10.2, 11.1]
IQR = 1.5999999999999996

-----All Alcohol Outliers Boundaries-----
al_extreme_low_bound = 4.700000000000001
al_mild_low_bound = 7.100000000000005
al_mild_up_bound = 13.5
al_extreme_up_bound = 15.899999999999999

-----All Alcohol Outliers-----
Extreme Outlier(Lower) = []
Mild Outlier(lower) = []
Mild Outlier(Upper) = [13.57, 13.6, 13.6, 13.6, 13.6, 14.0, 14.0, 14.0, 14.0, 14.0, 14.0, 14.0, 14.9]
Extreme Outlier(Upper) = []

-----Quality of Wines-----
quality Unit: None (lv.1-10)
quality Min 3
quality Mean : 5.6360225140712945
quality Median : 6
quality Max 8
quality Mode : 5
quality Sample Standard Deviation : 0.8075694397347849
quality Sample Variance : 0.6521683999934251
Quality[Q1, Q2, Q3]= [5.0, 6.0, 6.0]
IQR = 1.0

-----All Quality Outliers Boundaries-----
qu_extreme_low_bound = 2.0
qu_mild_low_bound = 3.5
qu_mild_up_bound = 7.5
qu_extreme_up_bound = 9.0

-----All Quality Outliers-----
Extreme Outlier(Lower) = []
Mild Outlier(lower) = [3, 3, 3, 3, 3, 3, 3, 3, 3]
Mild Outlier(Upper) = [8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8]
Extreme Outlier(Upper) = []

[Done] exited with code=0 in 12.689 seconds
```

## 2. wineGraph2.py

### 2.1. Source Code wineGraph2.py

```
◆ wineGraph2.py X
◆ wineGraph2.py > ...
    You, 14 minutes ago | 1 author (You)
1 # defining the libraries
2
3 import numpy as np
4 import matplotlib.pyplot as plt
5 import pandas           # collection for data
6
7
8 # Init Style of Graph and Insert table of data in form of columns
9 plt.style.use('bmh')
10 columns = pandas.read_csv('testgraphredwine.csv')
11
12 # All columns
13 x = columns['alcohol']      # x (independent variable) = alcohol
14 y = columns['quality']     # y (dependent variable) = quality
15
16
17 # getting data of the histogram
18 al_count, al_bins_count = np.histogram(x, bins=18)
19 qu_count, qu_bins_count = np.histogram(y, bins=1000) #6
20
21 # finding the PDF of the histogram using count values
22 al_pdf = al_count / sum(al_count)
23 al_cdf = np.cumsum(al_pdf)
24
25 qu_pdf = qu_count / sum(qu_count)
26 qu_cdf = np.cumsum(qu_pdf)
27
28
29 figure, al_func = plt.subplots(1, 2, figsize=(14, 5))
30 plt.tight_layout(pad=4, w_pad=3, h_pad=1.0)

◆ wineGraph2.py X
◆ wineGraph2.py > ...
    You, 14 minutes ago | 1 author (You)
30 plt.tight_layout(pad=4, w_pad=3, h_pad=1.0)
31
32 al_func[0].set_title('Alcohol in Red Wine (Probability Density Function (PDF))')
33 al_func[0].set_xlabel("Alcohol (%/volume)")
34 al_func[0].set_ylabel("Probability")
35 al_func[0].plot(al_bins_count[1:], al_pdf, color="green", label="PDF", )
36 al_func[0].legend()
37 al_func[0].axis(ymax=1)
38
39 al_func[1].set_title('Alcohol in Red Wine (Cumulative Distribution Function (CDF))')
40 al_func[1].set_xlabel("Alcohol (%/volume)")
41 al_func[1].set_ylabel("Probability")
42 al_func[1].plot(al_bins_count[1:], al_cdf, color="red", label="CDF")
43 al_func[1].legend()
44
45
46 figure, qu_func = plt.subplots(1, 2, figsize=(14, 5))
47 plt.tight_layout(pad=4, w_pad=3, h_pad=1.0)
48
49 qu_func[0].set_title('Quality of Red Wine (Probability Mass Function (PMF))')
50 qu_func[0].set_xlabel("Quality (lv.1-10)")
51 qu_func[0].set_ylabel("Probability")
52 qu_func[0].plot(qu_bins_count[1:], qu_pdf, color="green", label="PDF", )
53 qu_func[0].legend()
54 qu_func[0].axis(ymax=1)
55
56 qu_func[1].set_title('Quality of Red Wine (Cumulative Distribution Function (CDF))')
57 qu_func[1].set_xlabel("Quality (lv.1-10)")
58 qu_func[1].set_ylabel("Probability")
59 qu_func[1].plot(qu_bins_count[1:], qu_cdf, color="red", label="CDF")
60 qu_func[1].legend()
```

```
63     figure, total = plt.subplots(1, 2, figsize=(14, 5))
64     plt.tight_layout(pad=4, w_pad=3, h_pad=1.0)
65
66     total[0].set_title('Alcohol in Red Wine Summary Graph (PDF and CDF)')
67     total[0].set_xlabel("Alcohol (%/volume)")
68     total[0].set_ylabel("Probability")
69     total[0].plot(al_bins_count[1:], al_pdf, color="green", label="PDF")
70     total[0].plot(al_bins_count[1:], al_cdf, color="red", label="CDF")
71     total[0].legend()
72
73     total[1].set_title('Quality of Red Wine Summary Graph (PMF and CDF)')
74     total[1].set_xlabel("Quality (v.1-10)")
75     total[1].set_ylabel("Probability")
76     total[1].plot(qu_bins_count[1:], qu_pdf, color="green", label="PDF")
77     total[1].plot(qu_bins_count[1:], qu_cdf, color="red", label="CDF")
78     total[1].legend()
79
80
81
82     plt.show()
83
```

### 3. wineGraph3.py

#### 3.1. Source Code wineGraph3.py

```
◆ wineGraph3.py ◆ testcode.py
◆ wineGraph3.py > ...
You, seconds ago | 1 author (You)
1 # defining the libraries
2
3 import numpy as np
4 import matplotlib.pyplot as plt
5 import pandas # collection for data
6
7 import scipy.stats
8
9 # Init Style of Graph and Insert table of data in form of columns
10 plt.style.use('bmh')
11 columns = pandas.read_csv('testgraphredwine.csv')
12
13 # All columns
14 x = columns['alcohol'] # x (independent variable) = alcohol
15
16 # initial value
17 dataArray = 1.0 * np.array(x)
18 print('dataArray :', dataArray)
19 number = len(dataArray)
20 mean = np.mean(dataArray)
21 standardError = scipy.stats.sem(dataArray)
22 # standardError = standard deviation / samples 1.0656771926520383/ 39.98749804626440991456385162254 # from HW 1
23 print('standardError(hw1) :', 1.0656771926520383 / 39.98749804626440991456385162254 ) # from HW 1
24 print('standardError(hw4) :', standardError, '\n')
25

◆ wineGraph3.py ◆ testcode.py
◆ wineGraph3.py > ...
26
27 def confidence_interval(confidence):
28     con = format(confidence, '.2f')
29     print(f'**** confidence of {con} % ****')
30
31     z_score = scipy.stats.t.ppf( (1 + confidence) / 2.0, number-1 )
32     print('z-score :', z_score)
33
34     marginError = standardError * z_score
35     # marginError = standardError * z-score
36
37     print('margin error :', marginError, '\n')
38
39     print('****SUMMARY****')
40     print('mean :', mean)
41     print('lower-upper boundary:', mean - marginError, mean + marginError)
42
43     print('*****\n\n')
44     return mean, mean - marginError, mean + marginError
45
46 mean1, lowerB1, upperB1 = confidence_interval(0.90)
47 mean2, lowerB2, upperB2 = confidence_interval(0.95)
48 mean3, lowerB3, upperB3 = confidence_interval(0.99)
49
50
51
52 # getting data of the histogram
53 al_count, al_bins_count = np.histogram(x, bins=18) # y (quantity) and x (value)
54
55 # finding the PDF of the histogram using count values
56 al_pdf = al_count / sum(al_count)

◆ wineGraph3.py ◆ testcode.py
◆ wineGraph3.py > ...
58
59
60 figure, al_func = plt.subplots(3, 1, figsize=(8, 10))
61 plt.tight_layout(pad=5, h_pad=5.0)
62
63 y = np.linspace(0,1)
64
65 al_func[0].set_title('Confidence Interval (CI) Lv. 90% of Alcohol in Red Wine (PDF Graph)')
66 al_func[0].set_xlabel("Alcohol (%/volume)")
67 al_func[0].set_ylabel("Probability")
68 al_func[0].plot(al_bins_count[1:], al_pdf, color="green", label="PDF" )
69 x1 = np.linspace(lowerB1,lowerB1)
70 x2 = np.linspace(upperB1,upperB1)
71 al_func[0].plot(x1,y, label="Lower Boundary = {:.4f}".format(lowerB1))
72 al_func[0].plot(x2,y, label="Upper Boundary = {:.4f}".format(upperB1))
73 al_func[0].legend()
74 al_func[0].axis(ymax=1)
75
76 al_func[1].set_title('Confidence Interval (CI) Lv. 95% of Alcohol in Red Wine (PDF Graph)')
77 al_func[1].set_xlabel("Alcohol (%/volume)")
78 al_func[1].set_ylabel("Probability")
79 al_func[1].plot(al_bins_count[1:], al_pdf, color="green", label="PDF" )
80 x1 = np.linspace(lowerB2,lowerB2)
81 x2 = np.linspace(upperB2,upperB2)
82 al_func[1].plot(x1,y, label="Lower Boundary = {:.4f}".format(lowerB2))
83 al_func[1].plot(x2,y, label="Upper Boundary = {:.4f}".format(upperB2))
84 al_func[1].legend()
85 al_func[1].axis(ymax=1)
86
```

```

86
87 al_func[2].set_title('Confidence Interval (CI) Lv. 99% of Alcohol in Red Wine (PDF Graph)')
88 al_func[2].set_xlabel("Alcohol (%volume)")
89 al_func[2].set_ylabel("Probability")
90 al_func[2].plot(al_bins_count[1:], al_pdf, color="green", label="PDF" )
91 x1 = np.linspace(lowerB3,lowerB3)
92 x2 = np.linspace(upperB3,upperB3)
93 al_func[2].plot(x1,y, label="Lower Boundary = {:.4f}".format(lowerB3))
94 al_func[2].plot(x2,y, label="Upper Boundary = {:.4f}".format(upperB3))
95 al_func[2].legend()
96 al_func[2].axis(ymax=1)
97
98
99 plt.show()
100

```

### 3.2. Output wineGraph3.py

```

[Running] python -u "c:\Users\ASUS\Desktop\Prob-stat\wineGraph3.py"
dataArray : [ 8.4  8.4  8.5 ... 14.  14.  14.9]
standardError(hw1) : 0.026650259324028716
standardError(hw4) : 0.026650259324028723

*** confidence of 0.99 % ***
z-score : 1.6458077310000542
margin error : 0.04386120282865575

*****SUMMARY*****
mean : 10.422983114446529
lower-upper boundary: 10.379121911617874 10.466844317275184
*****SUMMARY*****

*** confidence of 0.95 % ***
z-score : 1.9614496156420889
margin error : 0.05227314890787792

*****SUMMARY*****
mean : 10.422983114446529
lower-upper boundary: 10.37070997353865 10.475256255354408
*****SUMMARY*****

*** confidence of 0.99 % ***
z-score : 2.5789094543589206
margin error : 0.06872860573185464

*****SUMMARY*****
mean : 10.422983114446529
lower-upper boundary: 10.354254508714675 10.491711720178383
*****SUMMARY*****

```

## 4. wineGraph4.py

### 4.1. Source Code wineGraph4.py

```
wineGraph4.py X
wineGraph4.py
You, 19 hours ago | 1 author (You)
1 import matplotlib.pyplot as plt      # plot graphs
2 import pandas                      # collection for data
3 from scipy import stats
4 import math
5 import statistics as stc
6
7
8 # Init Style of Graph and Insert table of data in form of columns
9 plt.style.use('bmh')
10 columns = pandas.read_csv('testgraphredwine.csv')
11
12 # All columns
13 x = columns['alcohol']           # x (independent variable) = alcohol
14 y = columns['quality']          # y (dependent variable) = quality
15
16 ...
17     #Graphs
18     1. Histogram
19     2. Box Plot
20     3. Stem and Leave
21     4. XY (Scatter) Plot (suitable variable)(describe more)
22
23     Detail
24     1. Name of Graph
25     2. Name of Axis
26     3. Suitable variable
27     4. Identify Outlier
28
29 ...
30 ...
31     y = mx + c
32
33     m = SSxy/SSxx
34
35     SSxy = sum(x*y) + sum(x)*sum(y)/n
36     SSxx = sum(x*x) + sum(x)*sum(x)/n
37     SSyy = sum(y*y) + sum(y)*sum(y)/n
38
39     r = SSxy / sqrt(SSxx*SSyy)
40
41 ...
42     #y = m*x + c
43
44     n = len(x)
45     y_bar = stc.mean(y)
46     x_bar = stc.mean(x)
47
48     # method 1 (easy to calculate)
49     SSxy = sum(x*y) - sum(x)*sum(y)/n
50     SSxx = sum(x*x) - sum(x)*sum(x)/n
51     SSyy = sum(y*y) - sum(y)*sum(y)/n
52
53     # method 2 (easy to understand)
54     SSxy_mean = sum((x - x_bar) * (y - y_bar))
55     SSxx_mean = sum((x - x_bar) ** 2)
56     SSyy_mean = sum((y - y_bar) ** 2)
57
58     # find slope (m)
59     m = SSxy / SSxx
60     # find intercept (c)
61     c = y_bar - m * x_bar
62     # find correlation coefficient (r)
63     r = SSxy / math.sqrt(SSxx*SSyy)
64
```

```

wineGraph4.py
wineGraph4.py
55 y_estimate = m * x + c
56 SSError = sum((y - estimate) ** 2) # ( actual value - estimate value ) ^ 2 # graph -> ('/.) # value is + or - change to ^2
57 Rsquare = 1 - SSError / SSyy
58 SSErrorInvert = sum((y - estimate - y_bar) ** 2) # ( estimate value - mean value ) ^ 2 # graph -> ('/-) # sum is = 0 if not ^2
59 RsquareV2 = SSErrorInvert / SSyy
60
61 # conclude: SSErrorInvert = (m * SSxy) = sum((y - estimate - y_bar) ** 2)
62 # conclude: SSyy = SSyy - (m * SSxy) = sum((y - y_estimate) ** 2)
63
64 RsquareV3 = (m * SSxy) / SSyy
65
66 stdERR = math.sqrt(SSError / (n - 2))
67 stdERRV2 = math.sqrt((SSyy - (m * SSxy)) / (n - 2))
68
69
70 print('----- My Own Calculation -----')
71
72 print("slope (m) = {:.2f}\n".format(m))
73 print("intercept (c) = {:.2f}\n".format(c))
74 print("n = {:.2f}\n".format(n))
75 print("y_bar = {:.2f}\n".format(y_bar))
76 print("x_bar = {:.2f}\n".format(x_bar))
77 print("SSxy_calc = {:.4f}\n".format(SSxy))
78 print("SSyy_mean = {:.4f}\n".format(SSyy_mean))
79 print("SSxx_calc = {:.4f}\n".format(SSxx))
80 print("SSyy_mean = {:.4f}\n".format(SSyy_mean))
81 print("SSxx_mean = {:.4f}\n".format(SSxx_mean))
82 print("SSyy_mean = {:.4f}\n".format(SSyy_mean))
83 print("r_value (r) = {:.4f}\n".format(r))
84 print("r_value**2 (r**2) = {:.4f}\n".format(r ** 2))
85 #print("y_estimate = (y - estimate)\n")
86 print("SSerror = {:.4f}\n".format(SSError))
87 print("SSerrorInvert = {:.4f}\n".format(SSErrorInvert))
88 print("Rsquare = {:.4f}\n".format(Rsquare))
89 print("RsquareV2 = {:.4f}\n".format(RsquareV2))

wineGraph4.py
wineGraph4.py
100 print("RsquareV3 = {:.4f}\n".format(RsquareV3))
101 print("stdERR = {:.4f}\n".format(stdERR))
102 print("stdERRV2 = {:.4f}\n".format(stdERRV2))
103
104 print('----- Use Library Function -----')
105
106 slope, intercept, r_value, p_value, std_err = stats.linregress(x,y)
107 print("slope (m) = {:.2f}\n".format(slope))
108 print("intercept (c) = {:.2f}\n".format(intercept))
109 print("r_value (r) = {:.4f}\n".format(r_value))
110 print("r_value**2 (r**2) = {:.4f}\n".format(r_value ** 2))
111 print("p_value (p) = {:.4f} # Not use ".format(p_value))
112 print("std_err = {:.4f} # this is wrong, Don't use this one, IDK why ? use 0.7104 instead\n".format(std_err))
113 print("n = {:.2f}x + {:.2f}, r**2 = {:.4f}, p = {:.4f}\n".format("n", slope, intercept, r_value ** 2, p_value))
114
115 # for title text
116 lineEquation = "n = {:.2f}x + {:.2f}, r**2 = {:.4f}, p = {:.4f}\n".format("n", slope, intercept, r_value ** 2, p_value)
117
118 # plot graph
119 xMin_xMax = [min(x),max(x)]
120 yMin_yMax = [slope*min(x) + intercept, slope*max(x) + intercept] # y = mx + c
121
122 # Scatter Plot
123 figure, scat = plt.subplots(figsize=(12, 8))
124 plt.tight_layout(pad=4)
125 scat.set_title('The Relation between Alcohol and Quality in Red Wine (Scatter plot)\n' + lineEquation)
126 scat.set_xlabel('Alcohol (%/volume)') #independent
127 scat.set_ylabel('Quality (1v.1-10)') #dependent
128 scat.scatter(x, y)
129 scat.plot(xMin_xMax, yMin_yMax, alpha=.5, color="green")
130
131 # Show
132 plt.show()

```

## 4.2. Output wineGraph4.py

```
TERMINAL PROBLEMS OUTPUT DEBUG CONSOLE
----- My Own Calculation -----
slope (m) = 0.36
intercept (c) = 1.88

n = 1599.00
y_bar = 5.64
x_bar = 10.42

SSxy_calc = 654.8462
SSxy_mean = 654.8462
SSxx_calc = 1814.7973
SSxx_mean = 1814.7973
SSyy_calc = 1042.1651
SSyy_mean = 1042.1651

r_value (r) = 0.4762
r_value^2 (r^2) = 0.2267
SSerror = 805.8723
SSerrorInvert = 236.2928
Rsquare = 0.2267
RsquareV2 = 0.2267
RsquareV3 = 0.2267

stdERR = 0.7104
stdERRV2 = 0.7104

----- Use Library Function -----
slope (m) = 0.36
intercept (c) = 1.88

r_value (r) = 0.4762
r_value^2 (r^2) = 0.2267

p_value (p) = 0.0000 # Not use
std_err = 0.0167 # this is wrong, Don't use this one, IDK why ? use 0.7104 instead

y = 0.36x + 1.88, r^2 = 0.2267, p = 0.0000
```

Source Code สำหรับการสร้างกราฟ และค่าทางสถิติต่างๆ ของคอลัมน์ Fixed Acidity และ Sulfur Dioxide

1) stats.m

```
raw = csvread('wine.csv');

x = raw(:,1);
y = raw(:,2);
z = raw(:,3);

for i = 1:2
    x = raw(:,i);
    Mean = mean(x)
    Median = median(x)
    Mode = mode(x)
    STD = std(x)
    Max = max(x)
    Min = min(x)
    Range = Max - Min
    Var = var(x)
    printf("-----\n");
endfor
```

2) Outlier.m

```
raw = csvread('wine.csv');

col_sel = 2;

raw = sortrows(raw,col_sel);

x = raw(:,col_sel);
z = raw(:,3);

n = size(raw)(1);
q1_term = 1/4 *( n + 1 );
q3_term = 3/4 *( n + 1 );

q1_t_floor = floor(q1_term);
q3_t_floor = floor(q3_term);

q1 = x(q1_t_floor) + (q1_term - q1_t_floor)*( x(q1_t_floor + 1) -
x(q1_t_floor) )
q3 = x(q3_t_floor) + (q3_term - q3_t_floor)*( x(q3_t_floor + 1) -
x(q3_t_floor) )
```

```

IQR = q3 - q1

Lower = q1 - 1.5*IQR;
Higher= q3 + 1.5*IQR;

printf("Lower : %d\n",Lower);
printf("Higher: %d\n\n",Higher);

Lower = max([0,Lower]);

bad = [];
average = [];
good = [];

for i = 1:1599
    if ( x(i) > Lower && x(i) < Higher )
        if(z(i) <= 3)
            bad = [bad,x(i)];
        elseif (z(i) <= 6)
            average = [average,x(i)];
        else
            good = [good,x(i)];
        endif
    endif
endfor

```

### 3) Scatter.m

```

[raw,x,y,z] = load();

grade = [0,3,7,10];
label = [];

for j = 1:3
    count = 0;
    figure(j);
    tX = [];
    tY = [];
    for i = 1:size(z)(1)
        if (z(i) <= grade(j+1) && (z(i) > grade(j)) )
            tX = [tX,x(i)];
            tY = [tY,y(i)];
            count++;
        endif
    endfor

```

```

scatter(tX,tY);
if j == 1
    label = "BAD";
elseif j == 2
    label = "GOOD";
else
    label = "HIGH";
endif
printf("Scatter for grade less than %d\n",grade(j+1));
count
title(["Scatter for " label " Quality Wine"]);
xlabel("Acidity (g/dm^3)");
ylabel("Sulfur dioxide (g/dm^3)");
set(gca,'fontsize',20);
endfor

figure(4)
boxplot(x,'o',1)
title(["Box Plot for Acidity Column"]);
ylabel("Acidity (g/dm^3)");
xlabel("");
set(gca,'fontsize',20);

figure(5)
boxplot(y,'o',1)
title(["Box Plot for Sulfur Column"]);
xlabel("");
ylabel("Sulfur dioxide (g/dm^3)");
set(gca,'fontsize',20);

figure(6)
hist(x);
title(["Histogram for Acidity Column"]);
xlabel("Acidity (g/dm^3)");
ylabel("Quatity");
set(gca,'fontsize',20);

figure(7)
hist(x);

xlabel("Sulfur dioxide (g/dm^3)");
ylabel("Quatity");
set(gca,'fontsize',20);

```

#### 4) trimed.m

```
raw = csvread('wine.csv');
x = raw(:,1);
y = raw(:,2);
z = raw(:,3);

function [hi,low] = triming(data)
    data = sort(data);
    n = 1599;
    q1_term = 1/4 *( n + 1 );
    q3_term = 3/4 *( n + 1 );

    q1_t_floor = floor(q1_term);
    q3_t_floor = floor(q3_term);

    q1 = data(q1_t_floor) + (q1_term - q1_t_floor)*( data(q1_t_floor +
1) - data(q1_t_floor) );
    q3 = data(q3_t_floor) + (q3_term - q3_t_floor)*( data(q3_t_floor +
1) - data(q3_t_floor) );

    IQR = q3 - q1;

    low = q1 - 1.5*IQR
    hi = q3 + 1.5*IQR

end

[hx,lx] = triming(x);
[hy,ly] = triming(y);
nx = [];
ny = [];
nz = [];
for i = 1:1599
    if ( x(i) >= lx && x(i) <= hx ) && ( y(i) >= ly && y(i) <= hy )
        nx = [nx;x(i)];
        ny = [ny;y(i)];
        nz = [nz;z(i)];
    endif
endfor
```

## 5) dist\_Plot.m

```
function [] = distPlot(x,label ="Fixed Acidity")

# PDF
figure(1);
p = normpdf(x,mean(x),std(x));
plot(x,p,"b*");
title([label," PDF and CDF"]);
xlabel([label," (g/l)"]);
set(gca,'fontsize',24);
hold on;

# CDF
c = normcdf(x,mean(x),std(x));
plot(x,c,"r*");
legend("PDF","CDF");
xlabel([label," (g/l)"]);
set(gca,'fontsize',20);

hold off;
```

## 6) Poly\_regression.m

```
function [b,stats] = poly_regress(x ,y , n = 2 , label = "")

X = [ones(size(x),1)];
for i = 1 : n;
    X = [X, x .^ i ];
endfor

[b,bint,r,rint,stats] = regress(y,X);
x1 = linspace(min(x),max(x),200);

one_array = ones(200,1);
p0 = b(1) .* one_array;
p1 = b(2) .* x1';

zz = p0 + p1 ;
for i = 2 : n;
    zz = zz + b(i + 1).*(x1' .^ i);
endfor

plot(x1,zz, "linewidth", 3);
title("Polynomial Regression");
```

```

xlabel([label," (g/l)"]);
ylabel("Quality (g/l)");
set(gca, 'fontsize',24);
hold;
scatter(x,y);
hold off;

end

```

### 7) poly\_regression\_multi\_dimension.m

```

function [b,stats] = poly_regression_multi_dimension(x,y,z, n = 2)

f0 = x .* y;
#f1 = y .* ( x.^ 2 );
#f1 = x.^ 2;
#f2 = x .* ( y.^ 2 );
#f3 = ( x.^ 2 ) .* ( y.^ 2 );
X = [ones(size(x),1),x,y];
for i = 2 : n
    X = [X,x.^ i , y.^ i];
endfor

[b,bint,r,rint,stats] = regress(z,X);
x1 = linspace(0,20,200);
y1 = linspace(0,250,200);

[xx,yy] = meshgrid(x1,y1);
one_array = [ones(size(xx))];

p0 = b(1) .* one_array;
p1 = b(2) .* xx;
p2 = b(3) .* yy;

zz = p0 + p1 + p2 ;

for i = 4 : size(b);
    d = floor(i/2);
    if( mod(i,2) == 0 )
        zz = zz + (b(i).* (xx .^ d)) ;
    else
        zz = zz + (b(i).* (yy .^ d ));
    endif
endfor

```

```

endfor
#p3 = b(4) .* ( xx .* yy );
#p3 = b(4) .* ( xx .* yy );
#p4 = b(5) .* ( (xx .^ 2) .* yy);
#p5 = b(6) .* ( (yy .^ 2) .* xx);
#p6 = b(7) .* ( (xx .^ 2) .* (yy .^ 2));

#zz = p0 + p1 + p2 + p3 + p4 + p5 ;
#zz = [ones(size(xx)),xx,yy] * b;
plot3(x1 ,y1,zz);
title("Polynomial regression 3D");
xlabel("fixed acid (g/l)");
ylabel("total sulfur (g/l)");
zlabel("Quality (g/l)");
set(gca, 'fontsize',20);
hold;
scatter3(x,y,z);
hold off;

end

```