

**นำเสนอ**

**ผศ.ดร. สุรินทร์ กิตตธรกุล**

**จัดทำโดย**

**นาย วัชระ วิริยะกุล 63010871**

**นายอัศราวุธ สานทอง 63011078**

**รายงานฉบับนี้เป็นส่วนหนึ่งของรายวิชา**

**01076253 PROBABILITY AND STATISTICS**

**สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง**

**ภาคเรียนที่ 2 ปีการศึกษา 2564**

**สารบัญ**

Water Quality (Drinking water potability)……………………………………………………………………………….…………3

Why is it interesting?……………………………………………………………………………………………………………….………….3

Fundamental Statistical Value……………………………………………………………………………………………….………….4

Histogram………………………………………………………………………………………………………………….………….…5

Stem&Leaf…………………………………………………………………………………………………………………….…….….6

Scatterplot………………………………………………………………………………………………………………..………….…7

BoxPlot…………………………………………………………………………………………………………………….…………..…8

Probability Density Function………………………………………………………………………………………………...10

Cumulative Distribution Function…………………………………………………………………………………………11

บทวิเคราะห์จากกราฟ………………………………………….………….…………………………………………………………12

Confidence Interval (CI) of Mean………………………………………………………………………………..…………………….13

pH Column…………………………………………………………………………………………………………………………….13

Turbidity Column…………………………………………………………………………………………………………………..19

บทวิเคราะห์ข้อมูลจากกราฟ……………………………………………………………………………………………………….25

Linear Regression……………………………………………………………………………………………………………………………….28

บทวิเคราะห์ข้อมูลจากกราฟ……………………………………………………………………………………………………….31

ชื่อชุดข้อมูล : **Water Quality (Drinking water potability)**

แหล่งที่มาของข้อมูล : [**https://www.kaggle.com/adityakadiwal/water-potability**](https://www.kaggle.com/adityakadiwal/water-potability)

**Why is it interesting?**

น้ำเป็นสิ่งที่จำเป็นต่อการดำรงชีวิตของมนุษย์เป็นอย่างมาก ซึ่งคุณภาพของน้ำเป็น สิ่งจำเป็นที่เราต้องรู้ก่อนที่จะนำเข้าสู่ร่างกาย ไม่ว่าจะเป็นความเป็นกรด-เบส ปริมาณสาร แขวนลอยต่างๆ และ Organic Carbon ที่ถูกสร้างจากแบคทีเรียที่อยู่ในแหล่งน้ำ สิ่งเหล่านี้ล้วนส่งผลต่อคุณภาพน้ำและส่งผลต่อผู้บริโภคเป็นอย่างมาก

**Graphical user interface, text

Description automatically generated with medium confidenceData frame Info :**

**หมายเหตุ :** จะเห็นว่าลำดับของข้อมูลไม่ได้ถูกเรียงตามลำดับเนื่องจากมีการตัดแถวของข้อมูลที่ไม่ทราบค่าออกไป [ na.omit(data) : in R programing ] เพื่อให้ได้การวิเคราะห์ผลที่แม่นยำที่สุดจากข้อมูลทั้งหมด 3276 ตัวอย่างเมื่อตัดข้อมูลที่ไม่ทราบค่าทั้งหมดออกไปจะเหลือข้อมูลทั้งหมด 1980(pH), 2011(Turbidity) ตัวอย่าง และ **เลือกมาวิเคราะห์เฉพาะคอลัมน์ที่สนใจคือ pH, Turbidity และ Potability**

**Fundamental Statistical Value :**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Statistical value** | **Non-Drinkable Water** | | **Drinkable Water** | |
| **pH** | **Turbidity (NTU)** | **pH** | **Turbidity (NTU)** |
| **Mean** | 7.0672 | 3.9552 | 7.1138 | 3.9913 |
| **Median** | 6.9920 | 3.9441 | 7.0465 | 4.0073 |
| **Mode** | 8.3168 | 4.6288 | 9.4451 | 3.8752 |
| **1st Quartile** | 5.9829 | 3.4447 | 6.2560 | 3.4406 |
| **3rd Quartile** | 8.1420 | 4.4975 | 7.9552 | 4.5275 |
| **Interquartile** | 2.1591 | 1.0528 | 1.6991 | 1.0868 |
| **Min** | 1.4318 | 1.4500 | 0.2275 | 1.4922 |
| **Max** | 14.0000 | 6.4947 | 11.8981 | 6.4942 |
| **Range** | 12.5682 | 5.0447 | 11.6706 | 5.0020 |
| **STD.** | 1.6591 | 0.7829 | 1.4376 | 0.7764 |
| **Variance** | 2.7526 | 0.6131 | 2.0668 | 0.6028 |

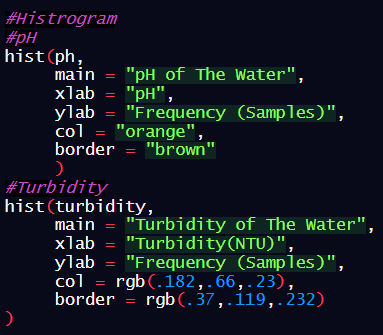
**Chart, histogram

Description automatically generatedChart, histogram

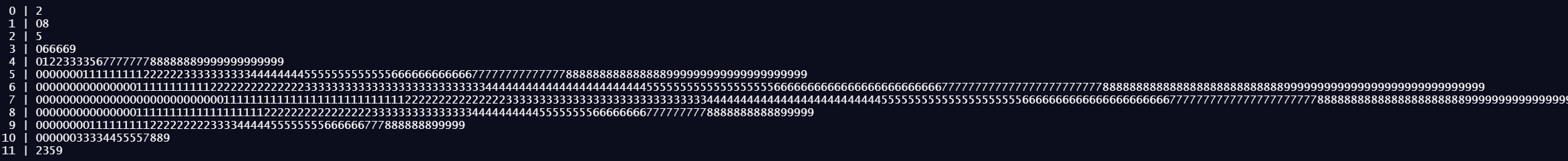
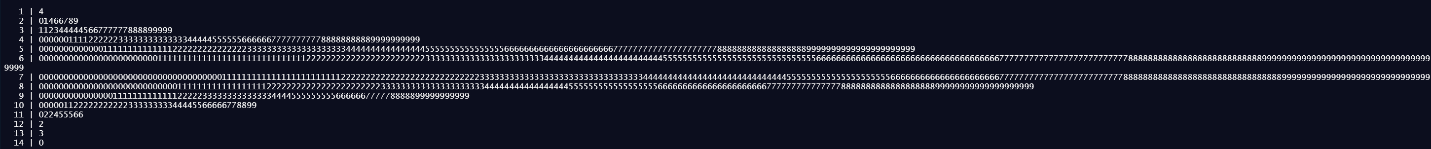
Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedHistogram :**

** หมายเหตุ :** NTU ย่อมาจาก Nephelometric turbidity unit โดยเป็นหน่วยวัดความขุ่นของน้ำโดยวัดจากสารแขวนลอยในน้ำในหน่วย mg/l หรือ ppm ซึ่งสารแขวนลอย 1 mg/l เท่ากับ 3 NTU องค์การอนามัยโลกกำหนดไว้ว่าน้ำที่สามารถนำมาบริโภคได้ไม่ควรจะมีค่าของ Turbidity เกิน 5 NTU

**Stem and Leaf :**

* **pH**
  + **Non-Drinkable Water**
  + **Drinkable Water**
* **Turbidity**
  + **A black screen with white text

    Description automatically generated with low confidenceNon-Drinkable Water**
  + **A black screen with white text

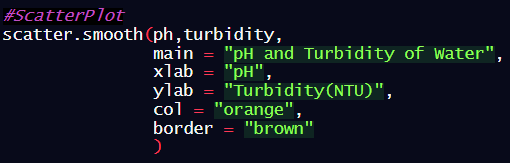
    Description automatically generated with low confidenceDrinkable Water**

**หมายเหตุ :** เนื่องจากข้อมูลมีจำนวนมากจึงไม่สามารถแสดงออกมาทั้งหมดได้ เพราะจะทำให้เกิดความไม่สะอาดในการรายงานผู้จัดทำจึงกำหนดความกว้างของ leaf ไว้เพียง 80 หน่วยเท่านั้น

**Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generatedScatter Plot :**

****

ตัวแปรต้น : pH

ตัวแปรตาม : Turbidity

**Boxplot :**

**Chart, box and whisker chart

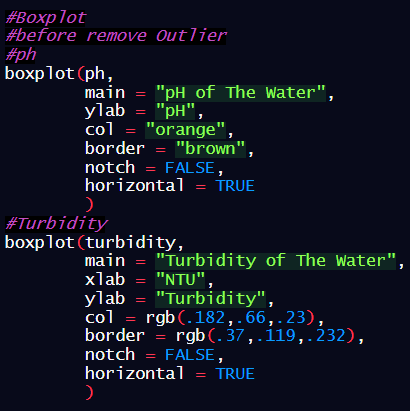
Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated Before remove outlier :**

outlier

****

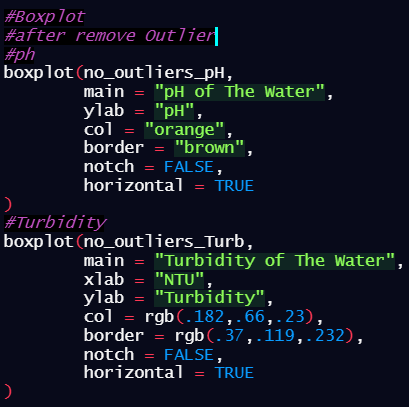
**Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

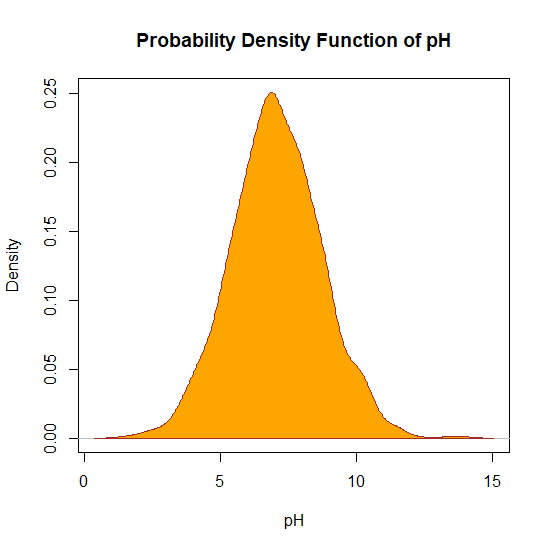
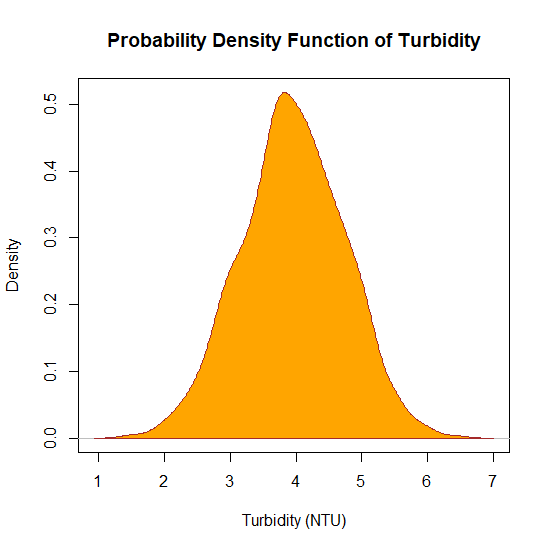
Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated After remove outlier :**

****

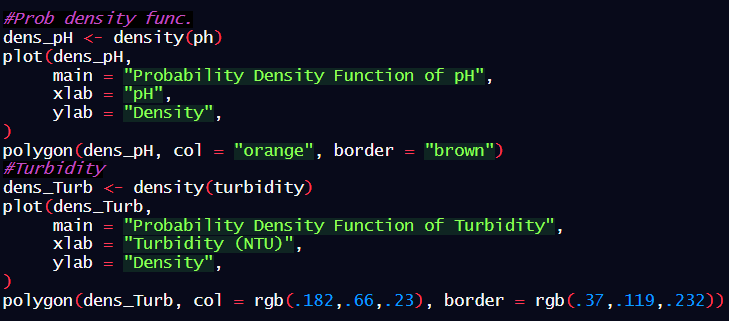
**Probability Density Function :**

** Non-Drinkable Water**

**Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated Drinkable Water**

****

**Cumulative Distribution Function :**

**Chart, histogram

Description automatically generated Non-Drinkable Water**

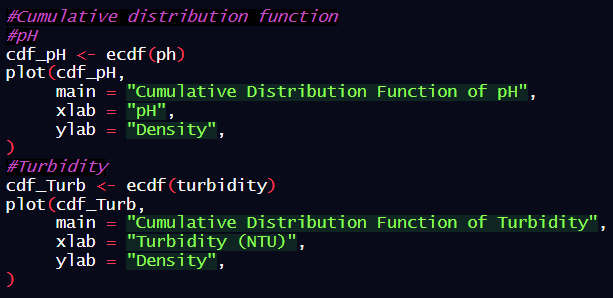
**Chart, histogram

Description automatically generated**

**Chart, line chart, histogram

Description automatically generatedChart

Description automatically generated Drinkable Water**

****

**บทวิเคราะห์**

จากการวิเคราะห์ Histogram ทั้งในส่วนของน้ำที่บริโภคได้และบริโภคไม่ได้ค่าของ pH มีช่วงที่ใกล้เคียงกันคือ 4-8 pH ซึ่งไม่สามารถสรุปได้ว่า pH เท่าใดควรเป็นค่าที่เหมาะสมสำหรับน้ำที่นำมาบริโภคจึงสามารถสรุปได้ว่า pH ไม่ใช่พารามิเตอร์ที่เหมาะสมสำหรับข้อมูลชุดนี้ในการสร้างโมเดลทำนายน้ำที่บริโภคได้หรือไม่ได้

จากการวิเคราะห์ในส่วนของ Scatter Plot ของทั้งน้ำที่สามารถนำมาบริโภคได้และบริโภคไม่ได้ ค่าความขุ่น(Turbidity)จะอยู่ในช่วงเดียวกันแต่จะแตกต่างกันค่อนข้างชัดเจนคือค่าของ pH ที่ในส่วนของน้ำที่บริโภคได้นั้นจะเกาะกลุ่มในช่วง pH 6-8 และน้ำบริโภคไม่ได้จะเกาะกลุ่มในช่วง pH 4-8 อีกทั้งค่าของ pH ส่งผลต่อความขุ่นของน้ำโดย pH แปรผันตรงกับความขุ่น

จากการวิเคราะห์ในส่วนของ PDF และ CDF จะเป็นแนวโน้มเดียวกับ Scatter Plot คือน้ำที่สามารถบริโภคได้จะมี pH 6-9 และน้ำบริโภคไม่ได้จะเกาะกลุ่มในช่วง pH 4-10 และค่าความขุ่น(Turbidity) อยู่ใมนช่วง 3-5 NTU

**สรุปผล**

น้ำที่สามารถบริโภคได้นั้นจากกลุ่มตัวอย่างจะมีค่า pH เป็นกลางคือช่วง 6-8 และในส่วนของค่าความขุ่นจะอยู่ใน ช่วง 3-5 NTU โดยยังมีค่าที่ทับซ้อนกันอยู่บางส่วนซึ่งสามารถคำนวณเป็น Error ของการคำนวณได้แต่ก็ยากที่จะสรุปให้ลงตัวได้ ยกตัวอย่างเช่น ผลกระทบของ pH ต่อความขุ่นของน้ำ ทั้งนี้หากต้องการความชัดเจนที่มากขึ้นควรจะใช้คอลัมน์อื่นๆเพื่อมาประกอบการพิจารณาด้วย เช่น จำนวนของคาร์บอนที่ถูกสร้างจากแบคทีเรียในน้ำเป็นต้น

**Confidence Interval (CI) of Mean**

pH Population Mean (µ) of non-drinkable water = 7.0576

pH Population Mean (µ) of drinkable water = 7.1206

Turbidity Population Mean (µ) of non-drinkable water = 3.956

Turbidity Population Mean (µ) of drinkable water = 3.9794

**pH Column**

* **Chart, box and whisker chart

  Description automatically generatedNon-Drinkable Water**

**pH Population Mean = 7.0576**

Figure 1: Confidence Interval of pH Mean (Confidence level: 90%) from non-drinkable water samples round.

Chart, box and whisker chart

Description automatically generated

**pH Population Mean = 7.0576**

Figure 2: Confidence Interval of pH Mean (Confidence level: 95%) from 50 non-drinkable water samples/round.

**Text

Description automatically generatedText

Description automatically generated**

**Chart, box and whisker chart

Description automatically generated**

**pH Population Mean = 7.0576**

Figure 3: Confidence Interval of pH Mean (Confidence level: 99%) from 50 non-drinkable water samples/round.

* **Chart, box and whisker chart

  Description automatically generatedDrinkable Water**

**pH Population Mean = 7.1206**

Figure 4: Confidence Interval of pH Mean (Confidence level: 90%) from 50 drinkable water samples/round.

**Chart, box and whisker chart

Description automatically generated**

**pH Population Mean = 7.1206**

Figure 5: Confidence Interval of pH Mean (Confidence level: 95%) from 50 drinkable water samples/round.

**Chart, box and whisker chart

Description automatically generated**

**pH Population Mean = 7.1206**

Figure 6: Confidence Interval of pH Mean (Confidence level: 99%) from 50 drinkable water samples/round.

**Turbidity Column**

* **Chart, box and whisker chart

  Description automatically generatedNon-Drinkable Water**

**Turbidity Population Mean = 3.956**

Figure 7: Confidence Interval of Turbidity Mean (Confidence level: 90%) from 50 non-drinkable water samples/round.

**Chart, box and whisker chart

Description automatically generated**

**Turbidity Population Mean = 3.956**

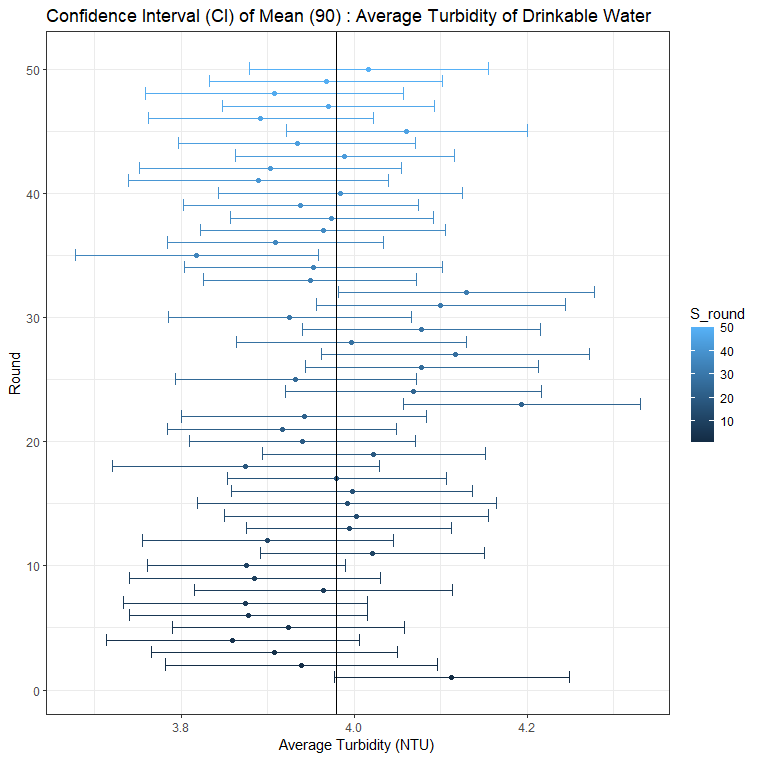
Figure 8: Confidence Interval of Turbidity Mean (Confidence level: 95%) from 50 non-drinkable water samples/round.

**Chart, box and whisker chart

Description automatically generated**

**Turbidity Population Mean = 3.956**

Figure 9: Confidence Interval of Turbidity Mean (Confidence level: 99%) from 50 non-drinkable water samples/round.

* **Drinkable Water**

**Turbidity Population Mean = 3.9794**

Figure 10: Confidence Interval of Turbidity Mean (Confidence level: 90%) from 50 drinkable water samples/round.

**Chart, box and whisker chart

Description automatically generated**

**Turbidity Population Mean = 3.9794**

Figure 11: Confidence Interval of Turbidity Mean (Confidence level: 95%) from 50 drinkable water samples/round.

**Chart, box and whisker chart

Description automatically generated**

**Turbidity Population Mean = 3.9794**

Figure 12: Confidence Interval of Turbidity Mean (Confidence level: 99%) from 50 drinkable water samples/round.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Mean With CL | Non-Drinkable Water | | | | | | Drinkable Water | | | | | |
| pH | | | Turbidity | | | pH | | | Turbidity | | |
| Lower bound | Upper bound | Range | Lower bound | Upper bound | Range | Lower bound | Upper bound | Range | Lower bound | Upper bound | Range |
| 90% | 6.4893 | 7.8756 | 1.3863 | 3.5725 | 4.3288 | 0.7563 | 6.5004 | 7.7151 | 1.2147 | 3.6774 | 4.3310 | 0.6536 |
| 95% | 6.2824 | 7.7758 | 1.4934 | 3.5576 | 4.3392 | 0.7816 | 6.4268 | 7.7559 | 1.3291 | 3.6189 | 4.3661 | 0.7472 |
| 99% | 5.8556 | 7.8850 | 2.0294 | 3.4676 | 4.4050 | 0.9374 | 6.3693 | 7.9054 | 1.5361 | 3.5610 | 4.4882 | 0.9272 |
|  | | | | | | | | | | | | |

Figure 13: Conclusion Table of Confidence Interval

**บทวิเคราะห์ข้อมูลจากกราฟ**

* **วิเคราะห์ข้อมูลจากกราฟกลุ่มของ Confidence Level 90 %**
  + Non-Drinkable Water
    - pH: มี 45 ค่าจาก 50 ค่าที่อยู่ในช่วงของ Confidence Interval ที่สร้างขึ้นมา หรือคิดเป็น 90% ของข้อมูลทั้งหมดและมี 10% ที่ค่า Population mean ไม่ได้อยู่ในช่วง Confidence Interval
    - Turbidity: มี 44 ค่าจาก 50 ค่าที่อยู่ในช่วงของ Confidence Interval ที่สร้างขึ้นมา หรือคิดเป็น 88% ของข้อมูลทั้งหมดและมี 12% ที่ค่า Population mean ไม่ได้อยู่ในช่วง Confidence Interval
  + Drinkable Water
    - pH: มี 44 ค่าจาก 50 ค่าที่อยู่ในช่วงของ Confidence Interval ที่สร้างขึ้นมา หรือคิดเป็น 88% ของข้อมูลทั้งหมดและมี 12% ที่ค่า Population mean ไม่ได้อยู่ในช่วง Confidence Interval
    - Turbidity: มี 47 ค่าจาก 50 ค่าที่อยู่ในช่วงของ Confidence Interval ที่สร้างขึ้นมา หรือคิดเป็น 94% ของข้อมูลทั้งหมดและมี 6% ที่ค่า Population mean ไม่ได้อยู่ในช่วง Confidence Interval
* **วิเคราะห์ข้อมูลจากกราฟกลุ่มของ Confidence Level 95 %**
  + Non-Drinkable Water
    - pH: มี 47 ค่าจาก 50 ค่าที่อยู่ในช่วงของ Confidence Interval ที่สร้างขึ้นมา หรือคิดเป็น 94% ของข้อมูลทั้งหมดและมี 6% ที่ค่า Population mean ไม่ได้อยู่ในช่วง Confidence Interval
    - Turbidity: มี 48 ค่าจาก 50 ค่าที่อยู่ในช่วงของ Confidence Interval ที่สร้างขึ้นมา หรือคิดเป็น 96% ของข้อมูลทั้งหมดและมี 4% ที่ค่า Population mean ไม่ได้อยู่ในช่วง Confidence Interval
  + Drinkable Water
    - pH: มี 47 ค่าจาก 50 ค่าที่อยู่ในช่วงของ Confidence Interval ที่สร้างขึ้นมา หรือคิดเป็น 94% ของข้อมูลทั้งหมดและมี 6% ที่ค่า Population mean ไม่ได้อยู่ในช่วง Confidence Interval
    - Turbidity: 47 ค่าจาก 50 ค่าที่อยู่ในช่วงของ Confidence Interval ที่สร้างขึ้นมา หรือคิดเป็น 94% ของข้อมูลทั้งหมดและมี 6% ที่ค่า Population mean ไม่ได้อยู่ในช่วง Confidence Interval
* **วิเคราะห์ข้อมูลจากกราฟกลุ่มของ Confidence Level 99 %**
  + Non-Drinkable Water
    - pH: มี 49 ค่าจาก 50 ค่าที่อยู่ในช่วงของ Confidence Interval ที่สร้างขึ้นมา หรือคิดเป็น 98% ของข้อมูลทั้งหมดและมี 2% ที่ค่า Population mean ไม่ได้อยู่ในช่วง Confidence Interval
    - Turbidity: มี 49 ค่าจาก 50 ค่าที่อยู่ในช่วงของ Confidence Interval ที่สร้างขึ้นมา หรือคิดเป็น 98% ของข้อมูลทั้งหมดและมี 2% ที่ค่า Population mean ไม่ได้อยู่ในช่วง Confidence Interval
  + Drinkable Water
    - pH: มี 49 ค่าจาก 50 ค่าที่อยู่ในช่วงของ Confidence Interval ที่สร้างขึ้นมา หรือคิดเป็น 98% ของข้อมูลทั้งหมดและมี 2% ที่ค่า Population mean ไม่ได้อยู่ในช่วง Confidence Interval
    - Turbidity: มี 49 ค่าจาก 50 ค่าที่อยู่ในช่วงของ Confidence Interval ที่สร้างขึ้นมา หรือคิดเป็น 98% ของข้อมูลทั้งหมดและมี 2% ที่ค่า Population mean ไม่ได้อยู่ในช่วง Confidence Interval

ซึ่งค่าเหล่านี้สามารถบ่งบอกได้ว่า มีโอกาส k% (Confidence level = k) โดยประมาณที่ Confidence interval ที่ถูกสร้างขึ้นมาจะคลอบคลุมค่าของ Population mean

**การนำไปใช้จริง**

* Non-Drinkable Water
  + pH
    - มีโอกาสประมาณ 90% ที่น้ำจะมีค่า pH อยู่ระหว่าง 6.4893 - 7.8756
    - มีโอกาสประมาณ 95% ที่น้ำจะมีค่า pH อยู่ระหว่าง 6.2824 - 7.7758
    - มีโอกาสประมาณ 99% ที่น้ำจะมีค่า pH อยู่ระหว่าง 5.8556 - 7.8850
  + Turbidity (NTU)
    - มีโอกาสประมาณ 90% ที่น้ำจะมีค่า Turbidity อยู่ระหว่าง 3.5725 - 4.3288 NTU
    - มีโอกาสประมาณ 95% ที่น้ำจะมีค่า Turbidity อยู่ระหว่าง 3.5576 - 4.3392 NTU
    - มีโอกาสประมาณ 99% ที่น้ำจะมีค่า Turbidity อยู่ระหว่าง 3.4676 - 4.405 NTU
* Drinkable Water
  + pH
    - มีโอกาสประมาณ 90% ที่น้ำจะมีค่า pH อยู่ระหว่าง 6.5004 - 7.7151
    - มีโอกาสประมาณ 95% ที่น้ำจะมีค่า pH อยู่ระหว่าง 6.4268 - 7.7559
    - มีโอกาสประมาณ 99% ที่น้ำจะมีค่า pH อยู่ระหว่าง 6.3693 - 7.9054
  + Turbidity (NTU)
    - มีโอกาสประมาณ 90% ที่น้ำจะมีค่า Turbidity อยู่ระหว่าง 3.6774 - 4.331 NTU
    - มีโอกาสประมาณ 95% ที่น้ำจะมีค่า Turbidity อยู่ระหว่าง 3.6189 - 4.3661 NTU
    - มีโอกาสประมาณ 99% ที่น้ำจะมีค่า Turbidity อยู่ระหว่าง 3.561 - 4.4882 NTU

**Linear Regression**

**Graph**

* ตัวแปรต้น คือ pH
* Chart, scatter chart

  Description automatically generatedตัวแปรตามคือ Turbidity(ความขุ่น)

Figure 14: Linear Regression Line with Scatterplot XY: X(pH), Y(Turbidity)

Text

Description automatically generatedText

Description automatically generatedจะได้สมการคือ โดยมีค่า ซึ่งเป็นค่า ที่น้อยมากหรือเรียกว่า Weak or No Correlation บ่งบอกได้ถึงข้อมูลทั้งสองคอลัมน์นั้นไม่มีความสัมพันธ์เชิงเส้นตรงต่อกันและกันทางผู้จัดทำจึงลองปรับเปลี่ยนการจับคู่ของคอลัมน์ดังผลลัพธ์ถัดไป

Text

Description automatically generated

* ตัวแปรต้นคือ pH
* Chart, line chart

  Description automatically generatedตัวแปรตามคือ Potability(สามารถบริโภคได้)

Figure 15: Linear Regression Line with Scatterplot XY: X(pH), Y(Potability)

Text

Description automatically generatedจะได้สมการคือ โดยมีค่า ซึ่งเป็นค่า ที่ดีกว่าการจับคู่ระหว่าง pH กับ Turbidity เล็กน้อย คือจาก 0.8% มาสู่ 2.31% แต่ก็ยังคงเป็น Weak or No Correlation อยู่ดีผู้จัดทำจึงลองปรับเปลี่ยนการจับคู่ของคอลัมน์ดังผลลัพธ์ถัดไป

Text

Description automatically generated

Text

Description automatically generated

* ตัวแปรต้นคือ Turbidity(ความขุ่น)
* Chart

  Description automatically generatedตัวแปรตามคือ Potability(สามารถบริโภคได้)

Figure 16: Linear Regression Line with Scatterplot XY: X(pH), Y(Potability)

จะได้สมการคือ โดยมีค่า ซึ่งเป็นค่า ที่ด้อยกว่าการจับคู่ระหว่าง pH กับ Portability เสียอีก

Text

Description automatically generatedText

Description automatically generated

**วิเคราะห์ข้อมูลจากกราฟ**

**บทวิเคราะห์ตามหลักคณิตศาสตร์**

จากกราฟ Linear Regression ทั้งสามกราฟที่ได้ ดังนี้

จากกราฟ Linear Regression ทั้งสามกราฟที่ได้จัดทำขึ้นซึ่งเป็นการจับคู่กันของคอลัมน์ข้อมูลที่สนใจทั้ง 3 คอลัมน์พบว่าความชันของแต่ละกราฟมีค่าเป็นบวก บ่งบอกถึงว่าข้อมูลของแต่ละกราฟมีแนวโน้มที่จะแปรผันตามกัน อย่างเช่นหากน้ำที่มีค่า pH สูงขึ้นก็ยังมีแนวโน้มที่จะมีความขุ่นเพิ่มมากขึ้นด้วยหรือน้ำที่สามารถบริโภคได้ส่วนใหญ่ค่า pH ก็จะอยู่ในช่วงที่มีค่าสูงเช่นกัน แต่ถึงอย่างไรความแม่นยำของกราฟนั้นมีไม่ถึง 5% จึงไม่สามารถสรุปได้ว่าข้อมูลที่เหลือทั้งหมดนอกจากกลุ่มตัวอย่างที่ตัดข้อมูลที่เป็น Outlier หรือ NA ออกนั้นจะมีแนวโน้มดังเช่นกราฟ การสรุปภาพรวมนั้นจึงต้องอาศัยข้อมูลคอลัมน์อื่นประกอบด้วย ยกตัวอย่างคอลัมน์ที่สำคัญสำหรับพิจารณาคุณภาพของน้ำคือ Trihalomethanes เป็นสารพิษที่ถูกสร้างมาจากการเติมคอรีนเพื่อฆ่าเชื้อภายในน้ำเป็นต้น

กล่าวโดยสรุปคือ การทำ Linear Regression ไม่เหมาะกับชุดข้อมูลที่เลือกมาทำให้ผลการทำนายโดยใช้ Linear Regression มีความแม่นยำต่ำ ไม่สามารถนำไปใช้ต่อยอดได้

**การปรับใช้จริง**

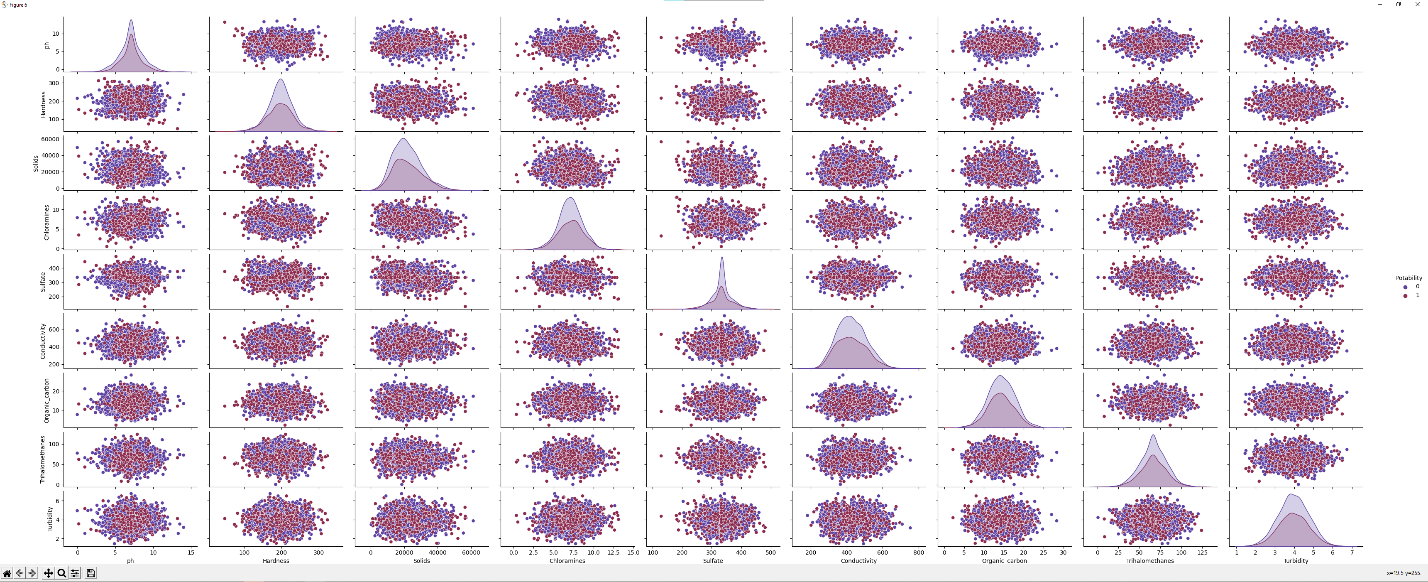
จากกราฟ Linear Regression ทำให้เห็นถึงความสำคัญของ Parameters ต่างๆที่มีผลต่อคุณภาพน้ำ จากที่ผู้จัดทำได้ทดลองกับทุกๆคอลัมน์พบว่า ความชันของกราฟหรือค่า m (y = mx + c) โดยส่วนใหญ่มีค่าเป็นบวกกับแทบทุกๆ parameter ดังภาพที่แนบมา

Figure 17: Scatterplot for all column

**สรุปผลการศึกษาและเสนอแนะแนวทางการศึกษาเพิ่มเติม**

จากการศึกษาพบว่าชุดข้อมูลชุดนี้ไม่เหมาะกับการวิเคราะห์แบบ Linear Regression เพราะไม่สามารถทำนายผลแนวโน้มที่จะเกิดขึ้นได้ ผู้จัดทำจึงได้ทำการลองการจัดการข้อมูลแบบอื่น ได้แก่ Random, ForestDecision, TreeKNNSupport, Vector, MachinesLogistic, RegressionNaive Bayes ดังผลลัพธ์ต่อไปนี้

Text

Description automatically generatedGraphical user interface, text

Description automatically generated

Text

Description automatically generated

Text

Description automatically generated

A screenshot of a computer

Description automatically generated with medium confidenceText

Description automatically generated

จะเห็นได้ว่าการจัดการข้อมูลแบบอื่นๆ เหมาะสมและสามารถทำนายได้ดีกว่าการทำแบบ Linear Regression

ดังนั้นแนวทางการศึกษาเพิ่มเติมคือ

1. ควรจะใช้การจัดการข้อมูลแบบอื่น ๆ ในการสร้างโมเดลทำนายแนวโน้ม
2. A screenshot of a computer

   Description automatically generated with low confidenceควรศึกษาพารามิเตอร์ที่สำคัญเกี่ยวกับคุณภาพน้ำให้ละเอียดเพื่อให้สามารถตัดสินใจได้ว่าควรตัดสินใจเลือกพารามิเตอร์ใดเป็นหลัก ให้เหมาะกับการคัดเลือกคุณภาพน้ำ และเพื่อลดต้นทุนการทดสอบ ยกตัวอย่างเช่นหากเลือกกำจัดน้ำที่ไม่ได้คุณภาพโดยใช้ค่าของ Conductive หรือค่าความนำไฟฟ้าของน้ำ จะเห็นได้ว่าเราสามารถตัดกลุ่มตัวอย่างออกไปได้จำนวนมาก จะทำให้สามารถลดค่าใช้จ่ายในการตรวจสอบข้อมูลได้
3. หากยังมีการใช้งานข้อมูลทางสถิติในอนาคตอย่างต่อเนื่องผู้ศึกษาควรจะศึกษาเครื่องมือต่างๆให้ดีมากกว่านี้ เพื่อความเหมาะสมในการคัดเลือกแนวทางการวิเคราะห์ข้อมูล

SourceCode :

R : https://drive.google.com/drive/folders/1q8bIktKrqvKYfphIOVWFxkhdK9-gMuHT?usp=sharing

Python (Pandas,Seaborn) : https://drive.google.com/drive/folders/1oqYMAHUXm0QICeCvAVRLJjvzljpjJ\_f6?usp=sharing