นำเสนอ

ผศ.ดร. สุรินทร์ กิตตธรกุล

จัดทำโดย

นาย วัชระ วิริยะกุล 63010871

นาย อัศราวุธ สานทอง 63011078

รายงานฉบับนี้เป็นส่วนหนึ่งของรายวิชา

01076253 PROBABILITY AND STATISTICS

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 2 ปีการศึกษา 2564

สารบัญ

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

แหล่งที่มาของข้อมูล : **https://www.kaggle.com/adityakadiwal/water-potability**

Why is it interesting?

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

Data frame Info :

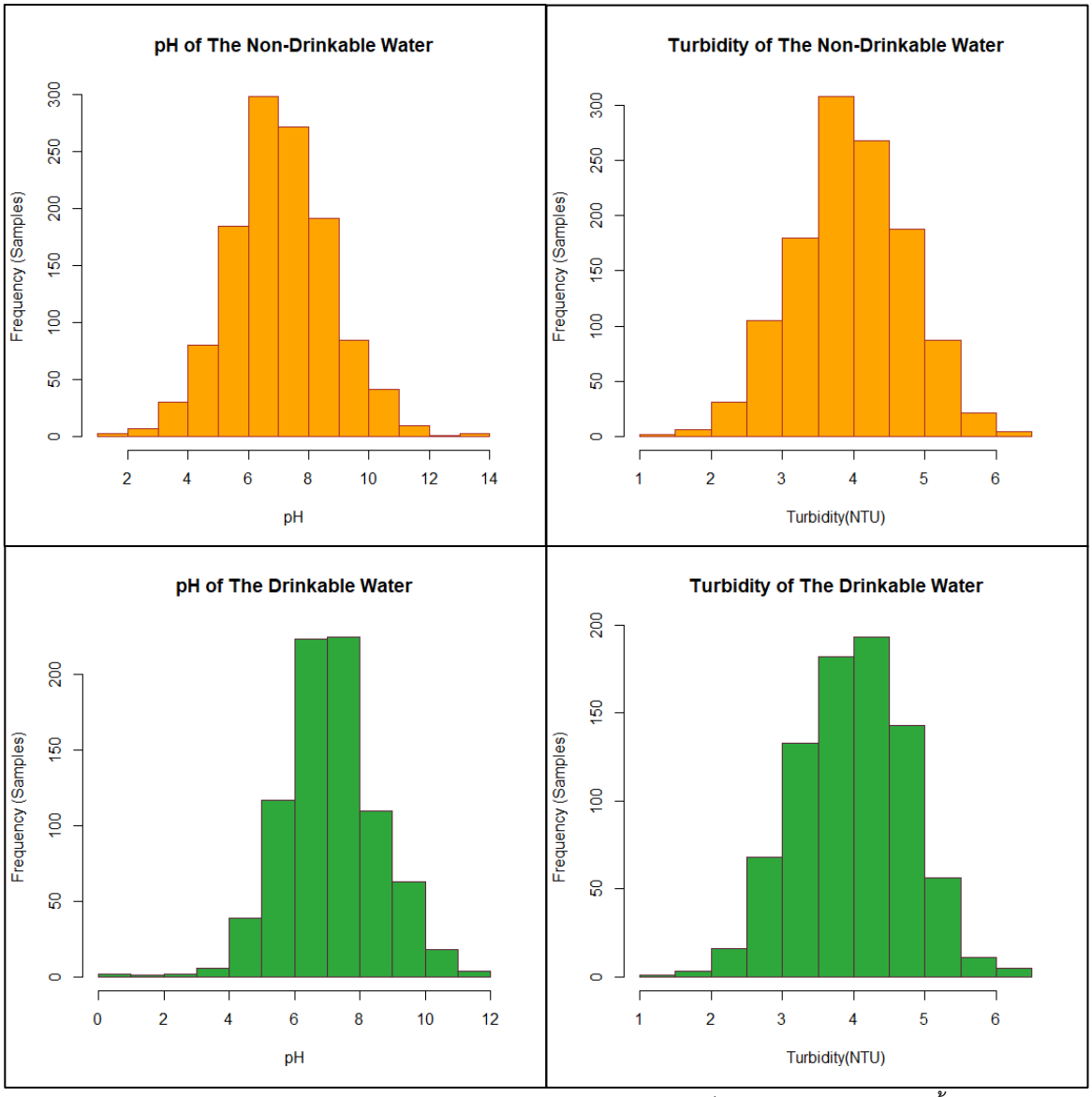
	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
4	8.316766	214.37339	22018.417	8.059332	356.8861	363.2665	18.436524	100.34167	4.628771	0
5	9.092223	181.10151	17978.986	6.546600	310.1357	398.4108	11.558279	31.99799	4.075075	0
6	5.584087	188.31332	28748.688	7.544869	326.6784	280.4679	8.399735	54.91786	2.559708	0
7	10.223862	248.07174	28749.717	7.513408	393.6634	283.6516	13.789695	84.60356	2.672989	0
8	8.635849	203.36152	13672.092	4.563009	303.3098	474.6076	12.363817	62.79831	4.401425	0
10	11.180284	227.23147	25484.508	9.077200	404.0416	563.8855	17.927806	71.97660	4.370562	0
11	7.360640	165.52080	32452.614	7.550701	326.6244	425.3834	15.586810	78.74002	3.662292	0
13	7.119824	156.70499	18730.814	3.606036	282.3441	347.7150	15.929536	79.50078	3.445756	0
16	6.347272	186.73288	41065.235	9.629596	364.4877	516.7433	11.539781	75.07162	4.376348	0
18	9.181560	273.81381	24041.326	6.904990	398.3505	477.9746	13.387341	71.45736	4.503661	0
20	7.371050	214.49661	25630.320	4.432669	335.7544	469.9146	12.509164	62.79728	2.560299	0
22	6.660212	168.28375	30944.364	5.858769	310.9309	523.6713	17.884235	77.04232	3.749701	0
25	5.400302	140.73906	17266.593	10.056852	328.3582	472.8741	11.256381	56.93191	4.824786	0
26	6.514415	198.76735	21218.703	8.670937	323.5963	413.2905	14.900000	79.84784	5.200885	0
27	3.445062	207.92626	33424.769	8.782147	384.0070	441.7859	13.805902	30.28460	4.184397	0
31	7.181449	209.62560	15196.230	5.994679	338.3364	342.1113	7.922598	71.53795	5.088860	0
33	10.433291	117.79123	22326.892	8.161505	307.7075	412.9868	12.890709	65.73348	5.057311	0
34	7.414148	235.04453	32555.853	6.845952	387.1753	411.9834	10.244815	44.48930	3.160624	0
36	5.115817	191.95274	19620.545	6.060713	323.8364	441.7484	10.966486	49.23823	3.902089	0
37	3.641630	183.90872	24752.072	5.538314	286.0596	456.8601	9.034067	73.59466	3.464353	0
40	9.267188	198.61439	24683.724	6.110612	328.0775	396.8769	16.471969	30.38331	4.324005	0
42	5.331940	194.87407	16658.877	7.993830	316.6752	335.1204	10.180514	59.57271	4.434820	0
43	7.145772	238.68993	28780.340	6.814029	385.9757	332.0327	11.093163	66.13804	5.182591	0

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

Histogram :

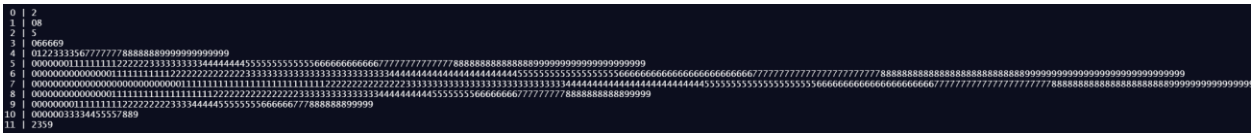


หมายเหตุ : NTU ย่อมาจาก Nephelometric turbidity unit โดยเป็นหน่วยวัดความขุ่นของน้ำโดยวัดจากสารแขวนลอยในน้ำในหน่วย mg/l หรือ ppm ซึ่งสารแขวนลอย 1 mg/l เท่ากับ 3 NTU องค์การอนามัยโลกกำหนดไว้ว่าน้ำที่สามารถนำมาบริโภคได้ไม่ควรจะมีค่าของ Turbidity เกิน 5 NTU

```r
#Histrogram
#pH
hist(ph,
     main = "pH of The Water",
     xlab = "pH",
     ylab = "Frequency (Samples)",
     col = "orange",
     border = "brown"
     )
#Turbidity
hist(turbidity,
     main = "Turbidity of The Water",
     xlab = "Turbidity(NTU)",
     ylab = "Frequency (Samples)",
     col = rgb(.182,.66,.23),
     border = rgb(.37,.119,.232)
)
```

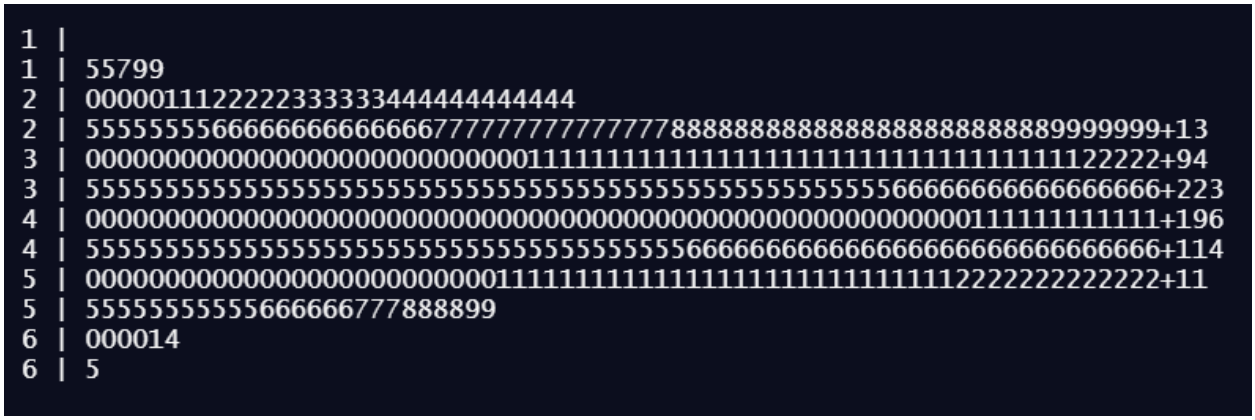
Stem and Leaf :

- **pH**
    - **Non-Drinkable Water**

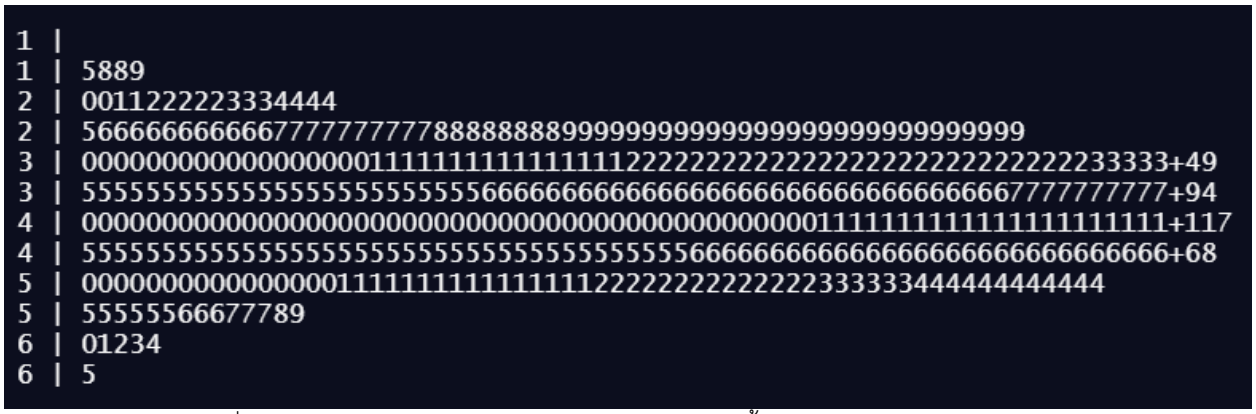


    - **Drinkable Water**

- **Turbidity**
    - **Non-Drinkable Water**

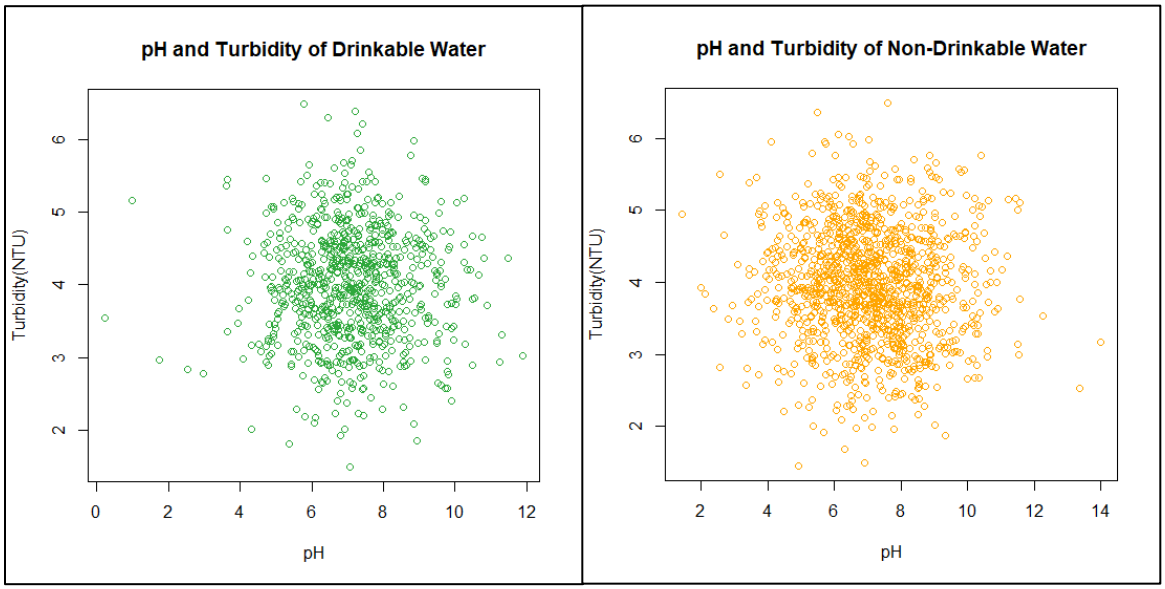


    - **Drinkable Water**



หมายเหตุ : เนื่องจากข้อมูลมีจำนวนมากจึงไม่สามารถแสดงออกมาทั้งหมดได้ เพราะจะทำให้เกิดความไม่สะอาดในการรายงานผู้จัดทำจึงกำหนดความกว้างของ leaf ไว้เพียง 80 หน่วยเท่านั้น

Scatter Plot :



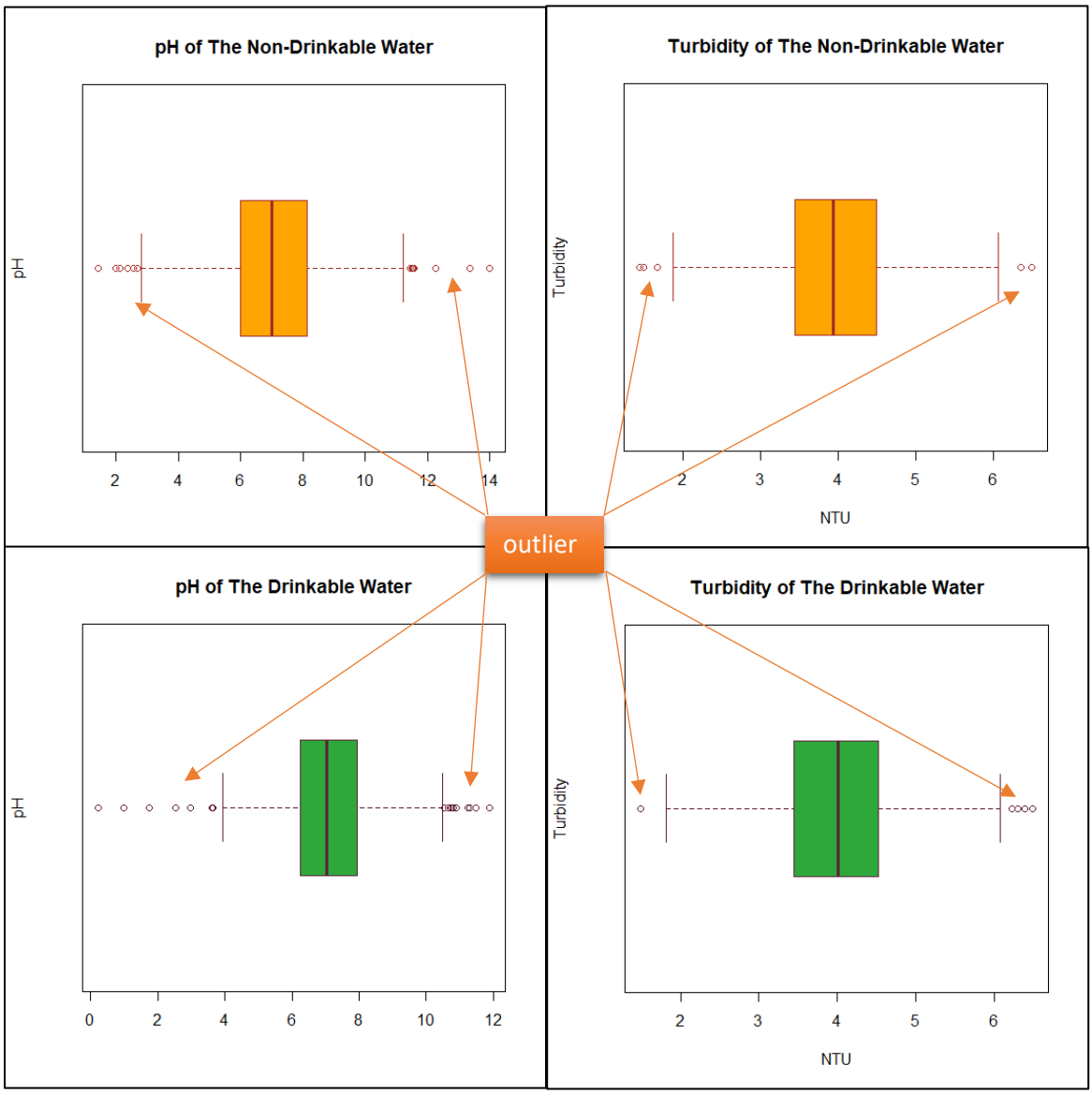
ตัวแปรต้น : pH

ตัวแปรตาม : Turbidity

```r
#ScatterPlot
scatter.smooth(ph,turbidity,
               main = "pH and Turbidity of Water",
               xlab = "pH",
               ylab = "Turbidity(NTU)",
               col = "orange",
               border = "brown"
               )
```

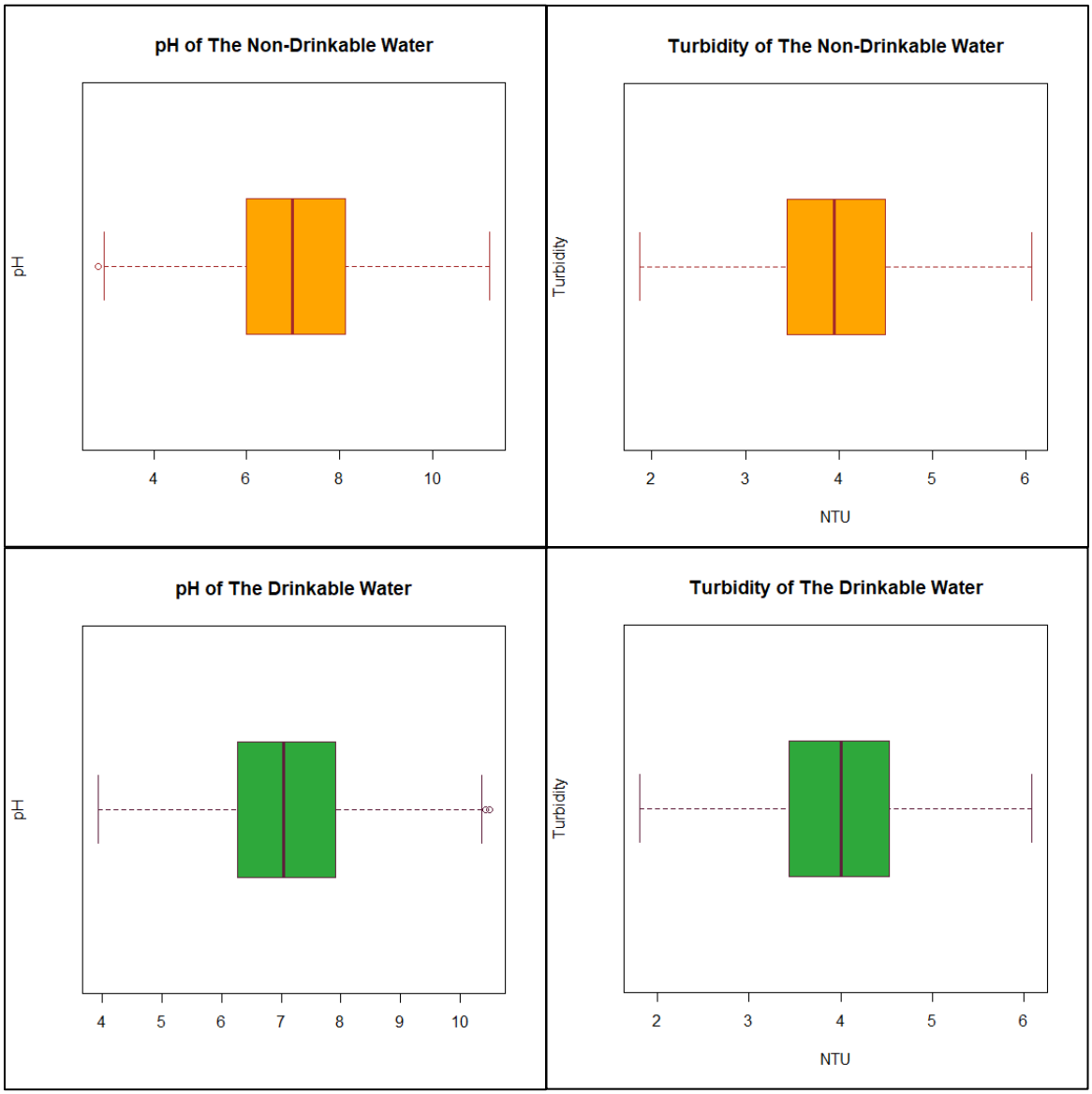
Boxplot :

Before remove outlier :




```r
#Boxplot
#before remove Outlier
#ph
boxplot(ph,
        main = "pH of The Water",
        ylab = "pH",
        col = "orange",
        border = "brown",
        notch = FALSE,
        horizontal = TRUE
        )
#Turbidity
boxplot(turbidity,
        main = "Turbidity of The Water",
        xlab = "NTU",
        ylab = "Turbidity",
        col = rgb(.182,.66,.23),
        border = rgb(.37,.119,.232),
        notch = FALSE,
        horizontal = TRUE
        )
```

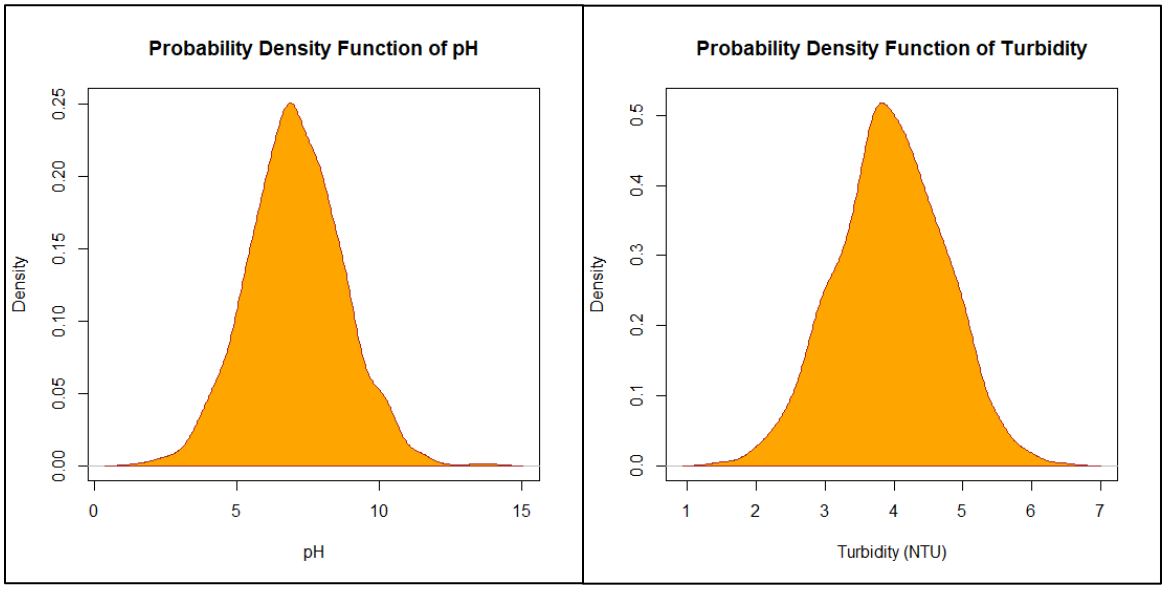
After remove outlier :



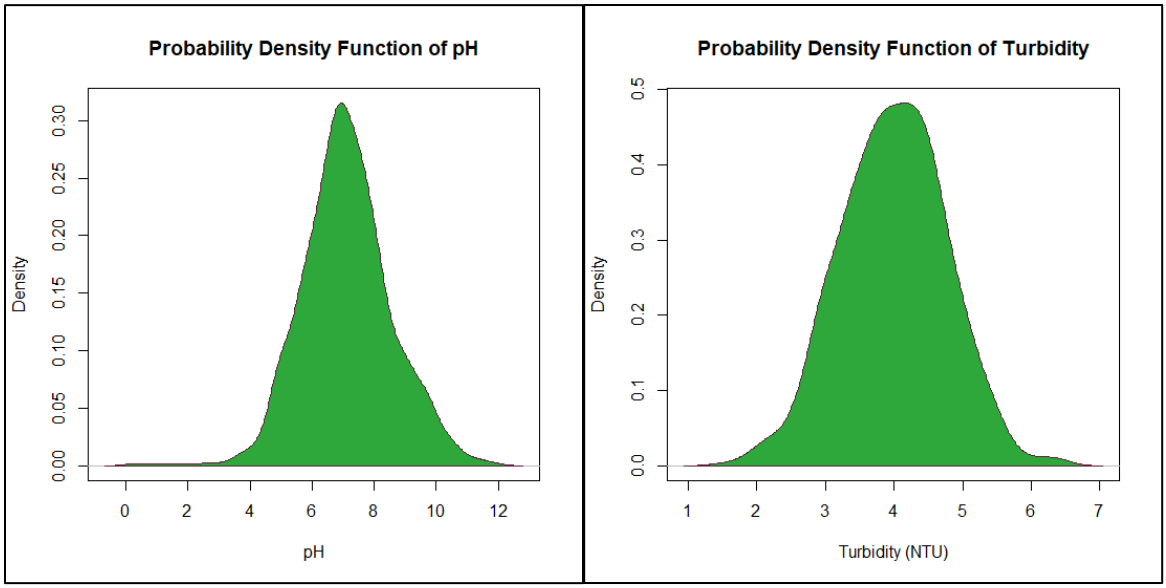

```r
#Boxplot
#after remove Outlier
#ph
boxplot(no_outliers_pH,
        main = "pH of The Water",
        ylab = "pH",
        col = "orange",
        border = "brown",
        notch = FALSE,
        horizontal = TRUE
)
#Turbidity
boxplot(no_outliers_Turb,
        main = "Turbidity of The Water",
        xlab = "NTU",
        ylab = "Turbidity",
        col = rgb(.182,.66,.23),
        border = rgb(.37,.119,.232),
        notch = FALSE,
        horizontal = TRUE
)
```

Probability Density Function :

Non-Drinkable Water



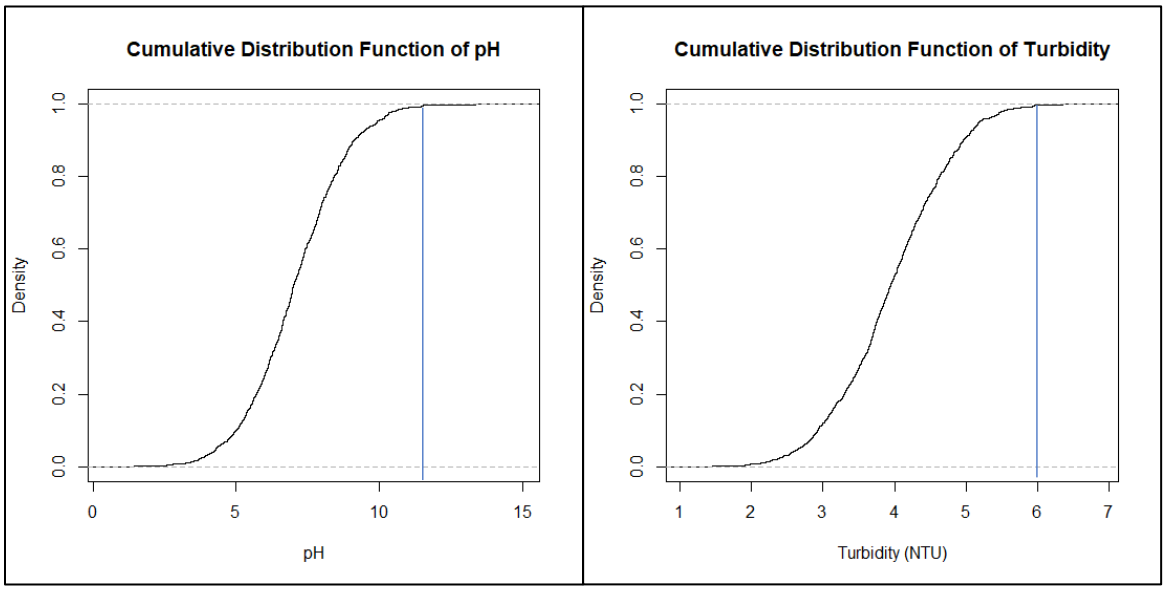
Drinkable Water



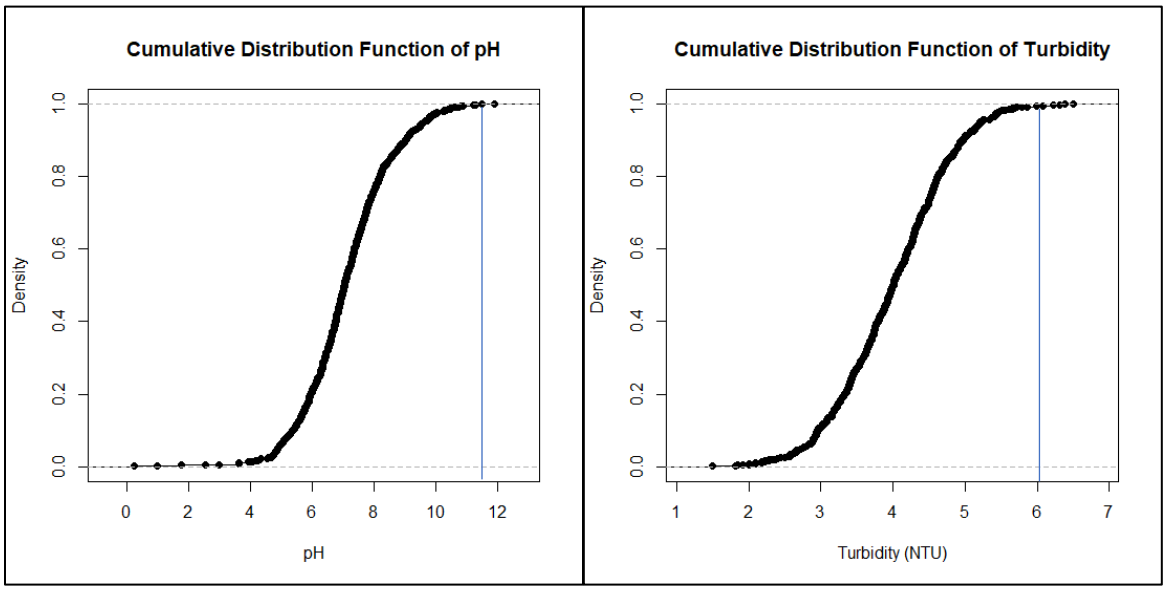

```r
#Prob density func.
dens_pH <- density(ph)
plot(dens_pH,
     main = "Probability Density Function of pH",
     xlab = "pH",
     ylab = "Density",
)
polygon(dens_pH, col = "orange", border = "brown")
#Turbidity
dens_Turb <- density(turbidity)
plot(dens_Turb,
     main = "Probability Density Function of Turbidity",
     xlab = "Turbidity (NTU)",
     ylab = "Density",
)
polygon(dens_Turb, col = rgb(.182,.66,.23), border = rgb(.37,.119,.232))
```

Cumulative Distribution Function :

Non-Drinkable Water



Drinkable Water




```r
#Cumulative distribution function
#pH
cdf_pH <- ecdf(ph)
plot(cdf_pH,
     main = "Cumulative Distribution Function of pH",
     xlab = "pH",
     ylab = "Density",
)
#Turbidity
cdf_Turb <- ecdf(turbidity)
plot(cdf_Turb,
     main = "Cumulative Distribution Function of Turbidity",
     xlab = "Turbidity (NTU)",
     ylab = "Density",
)
```

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) อยู่ในช่วง 3-5 NTU

สรุปผล

น้ำที่สามารถบริโภคได้นั้นจากกลุ่มตัวอย่างจะมีค่า pH เป็นกลางคือช่วง 6-8 และในส่วนของค่าความขุ่นจะอยู่ในช่วง 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

- **Non-Drinkable Water**

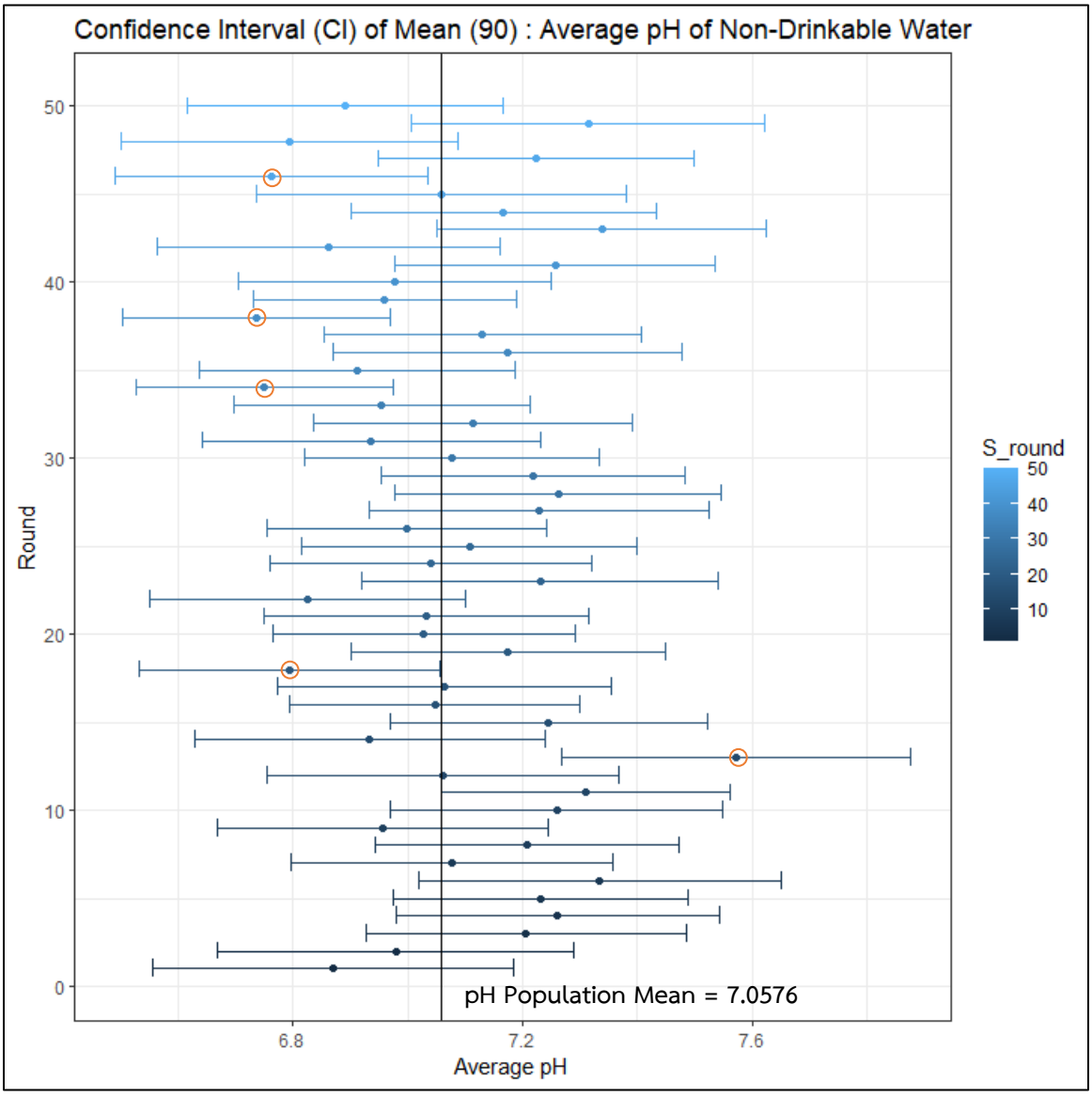


Figure 1: Confidence Interval of pH Mean (Confidence level: 90%) from non-drinkable water samples round.

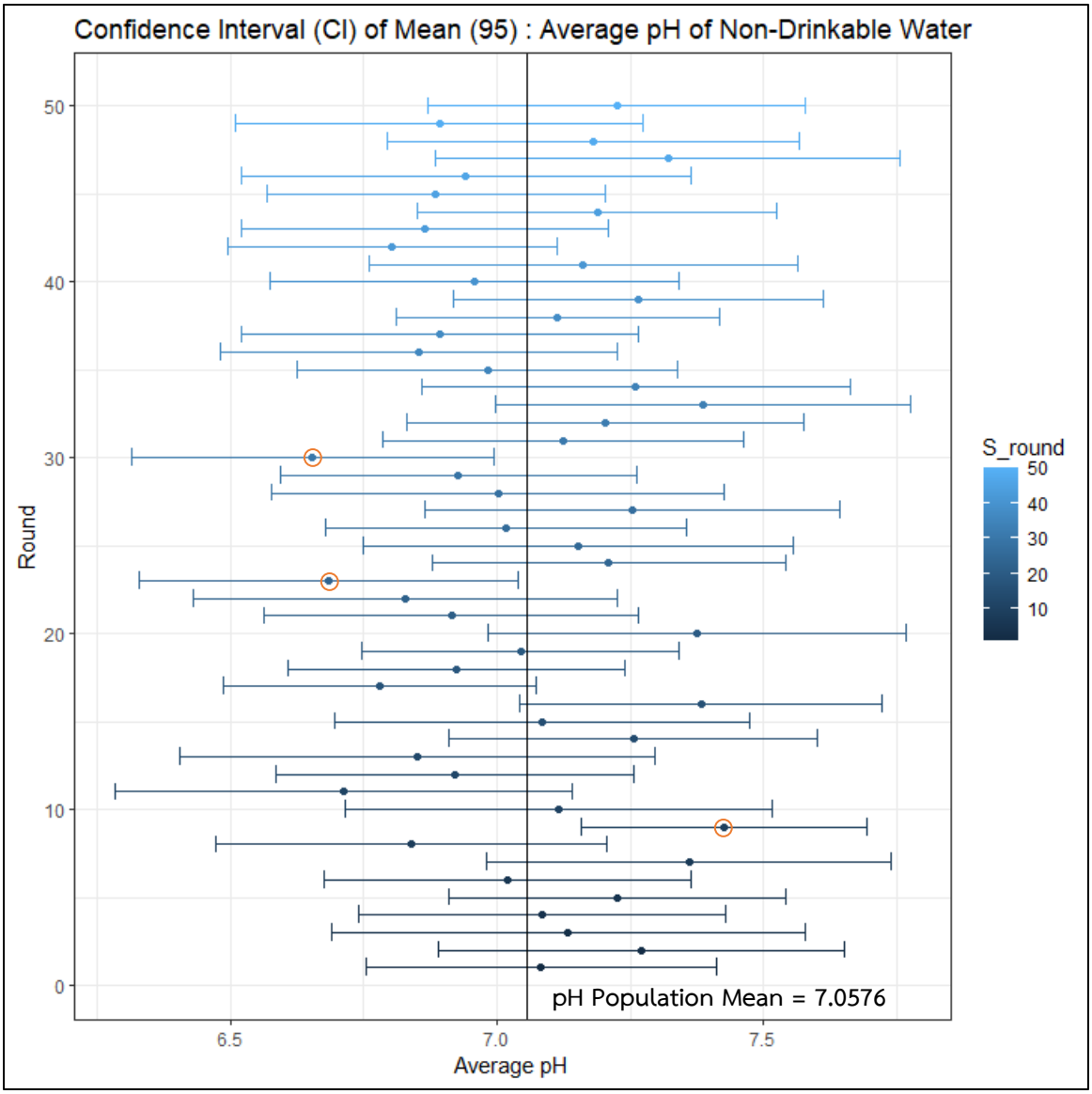


Figure 2: Confidence Interval of pH Mean (Confidence level: 95%) from 50 non-drinkable water samples/round.

```r
#### Confidence Interval
getCI <- function(cl,n,x){
  m <- mean(x)
  s <-sd(x)
  se<- s/sqrt(n)
  z <- qnorm(cl)
  me <- se*z
  ci <- c(m-me,m+me)
  return(ci)
}
rounds = 50
S_round = c(1:rounds)
nsamples = 50
pop_pot0_ph_mean = mean(data_Pot0_no_out$pH)
pop_pot0_ph_sd = sd(data_Pot0_no_out$pH)
pop_pot1_ph_mean = mean(data_Pot1_no_out$pH)
pop_pot1_ph_sd = sd(data_Pot1_no_out$pH)
pop_pot0_turb_mean = mean(data_Pot0_no_out$Turbidity)
pop_pot0_turb_sd = sd(data_Pot0_no_out$Turbidity)
pop_pot1_turb_mean = mean(data_Pot1_no_out$Turbidity)
pop_pot1_turb_sd = sd(data_Pot1_no_out$Turbidity)
cl90 = 0.9
cl95 = 0.95
cl99 = 0.99
```

```r
## Confidence Interval 90 pH : Potability = 0 ##
CI_pH_Pot0_90_mean = c()
CI_pH_Pot0_90_lower = c()
CI_pH_Pot0_90_upper = c()
Arr_sample_AVG = c()
Arr_sample_Mean = c()
Arr_sample_SD = c()
## random sample
for(a in 1:rounds){
  sample_pH = sample(data_Pot0_no_out$pH,nsamples)
  Arr_sample_AVG[a] = c(data.frame(sample_pH))
  Arr_sample_SD[a] = c(sd(sample_pH))
  Arr_sample_Mean[a] = c(mean(sample_pH))
}
## assign array lower upper mean
for(b in 1:rounds){
  CI_pH_Pot0_90_lower[b]= getCI(cl90,nsamples,Arr_sample_AVG[[b]])[1]
  CI_pH_Pot0_90_upper[b]= getCI(cl90,nsamples,Arr_sample_AVG[[b]])[2]
  CI_pH_Pot0_90_mean[b] = mean(Arr_sample_Mean[b])
}

CI_pH_Pot0_90 = data.frame(S_round,CI_pH_Pot0_90_mean,CI_pH_Pot0_90_lower,CI_pH_Pot0_90_upper)

qplot(
  x = CI_pH_Pot0_90_mean,
  y = S_round,
  color = S_round,
  data = CI_pH_Pot0_90,main = "Confidence Interval (CI) of Mean (90) : Average pH of Non-Drinkable Water",
  xlab = "Average pH",
  ylab = "Round"
)+geom_errorbar(aes(xmin = CI_pH_Pot0_90_lower, xmax = CI_pH_Pot0_90_upper, width = 1)) + geom_vline(xintercept = pop_pot0_ph_mean)
```

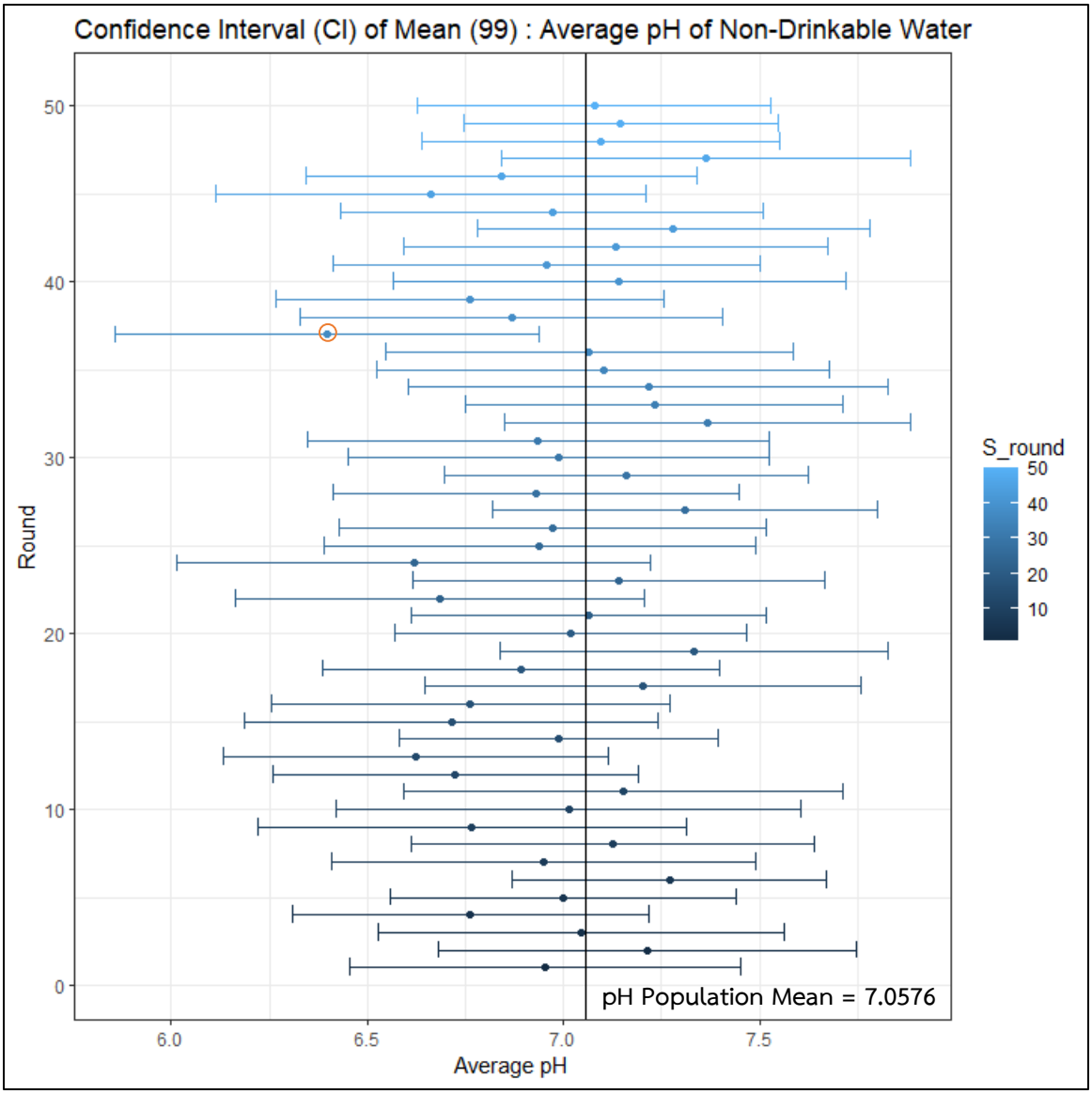


Figure 3: Confidence Interval of pH Mean (Confidence level: 99%) from 50 non-drinkable water samples/round.

- Drinkable Water

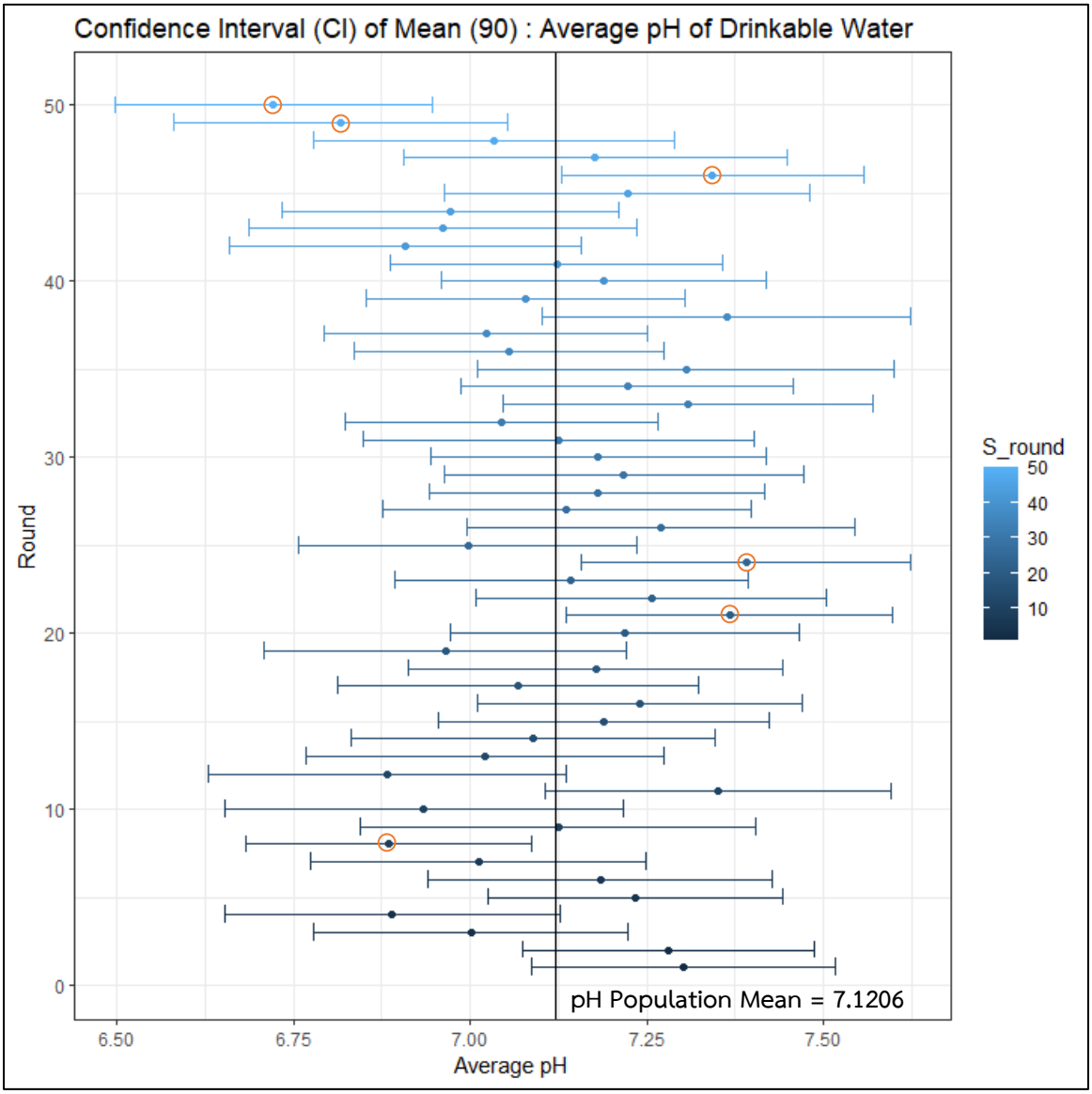


Figure 4: Confidence Interval of pH Mean (Confidence level: 90%) from 50 drinkable water samples/round.

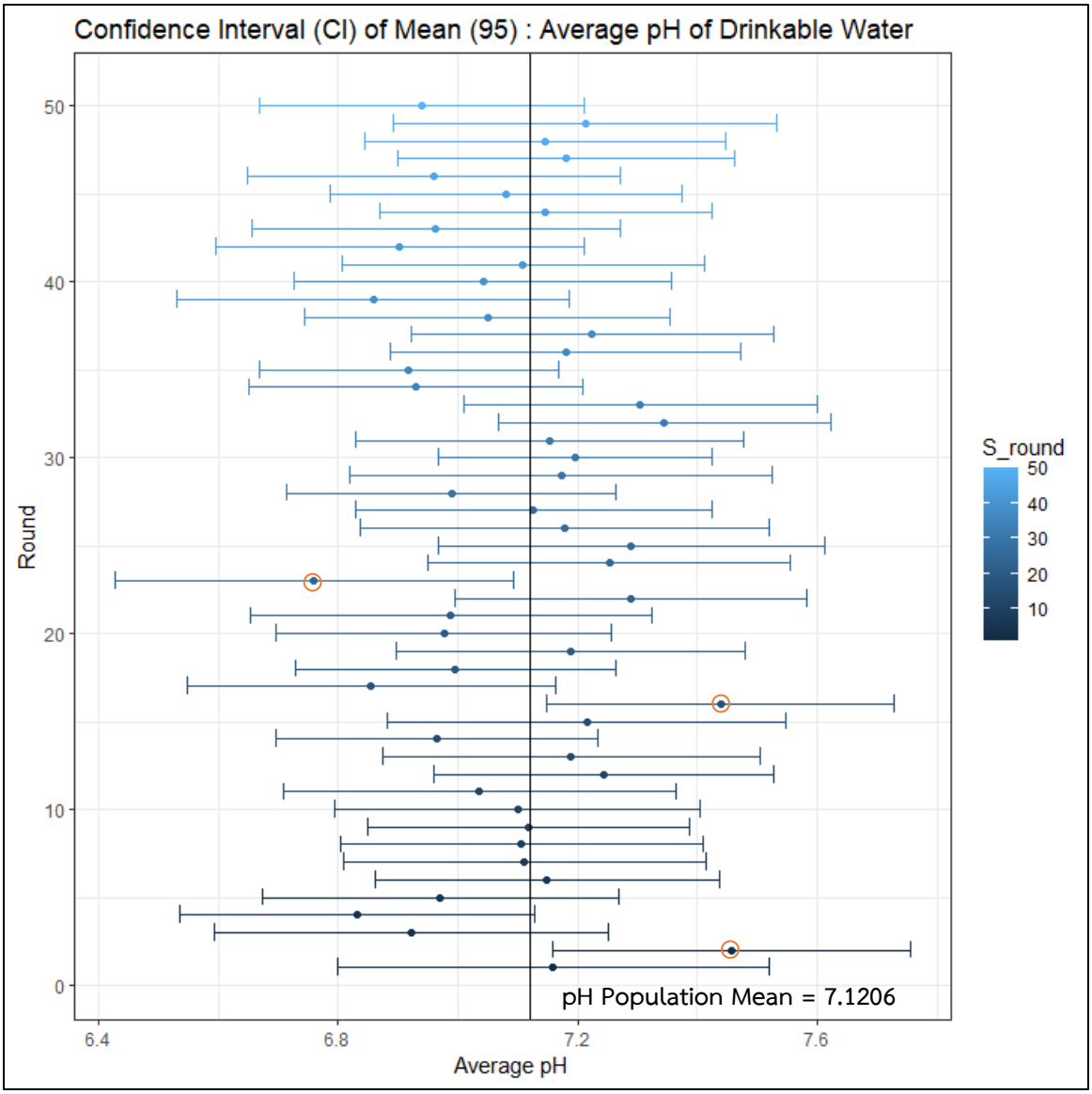


Figure 5: Confidence Interval of pH Mean (Confidence level: 95%) from 50 drinkable water samples/round.

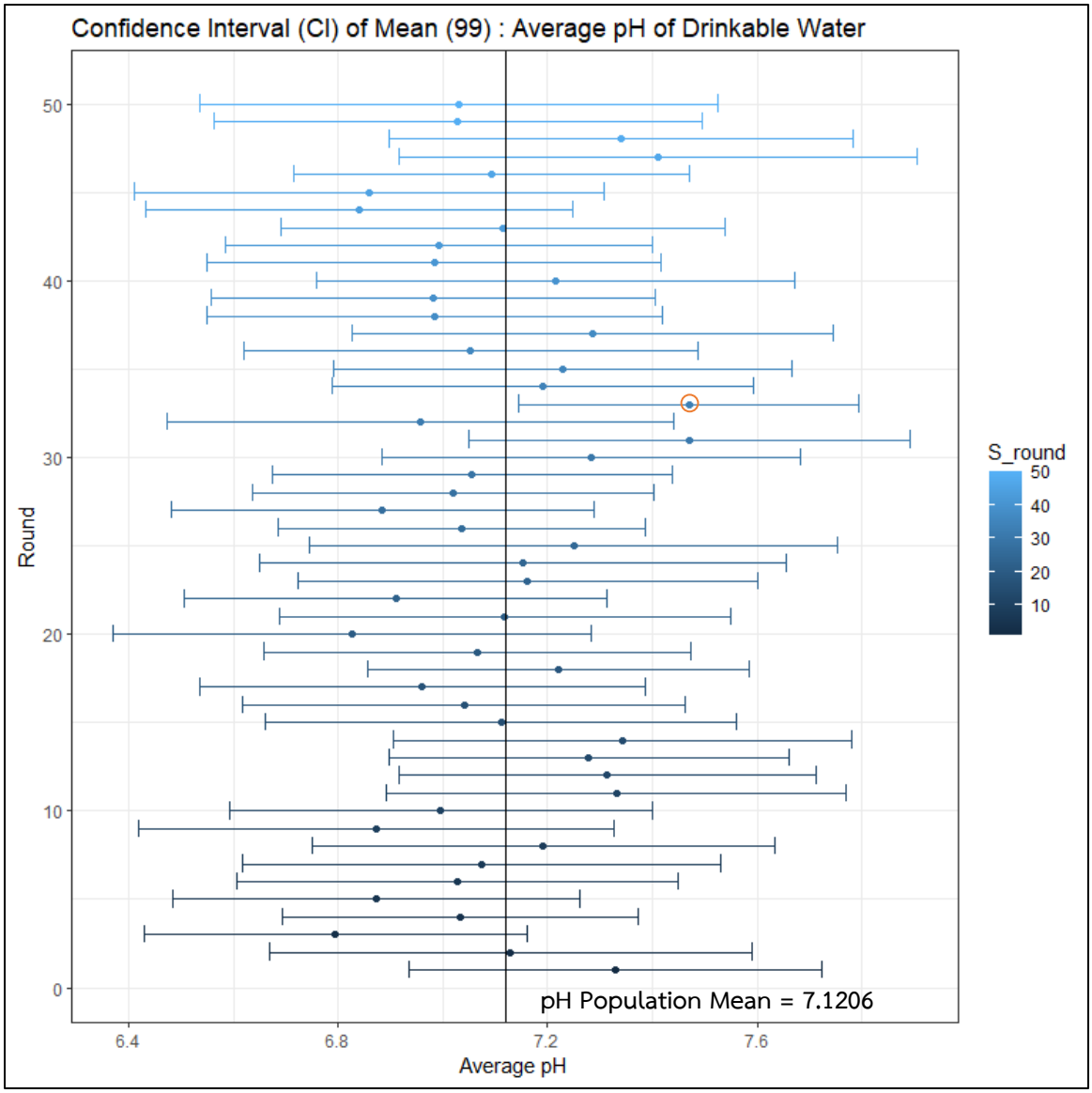


Figure 6: Confidence Interval of pH Mean (Confidence level: 99%) from 50 drinkable water samples/round.

Turbidity Column

- Non-Drinkable Water

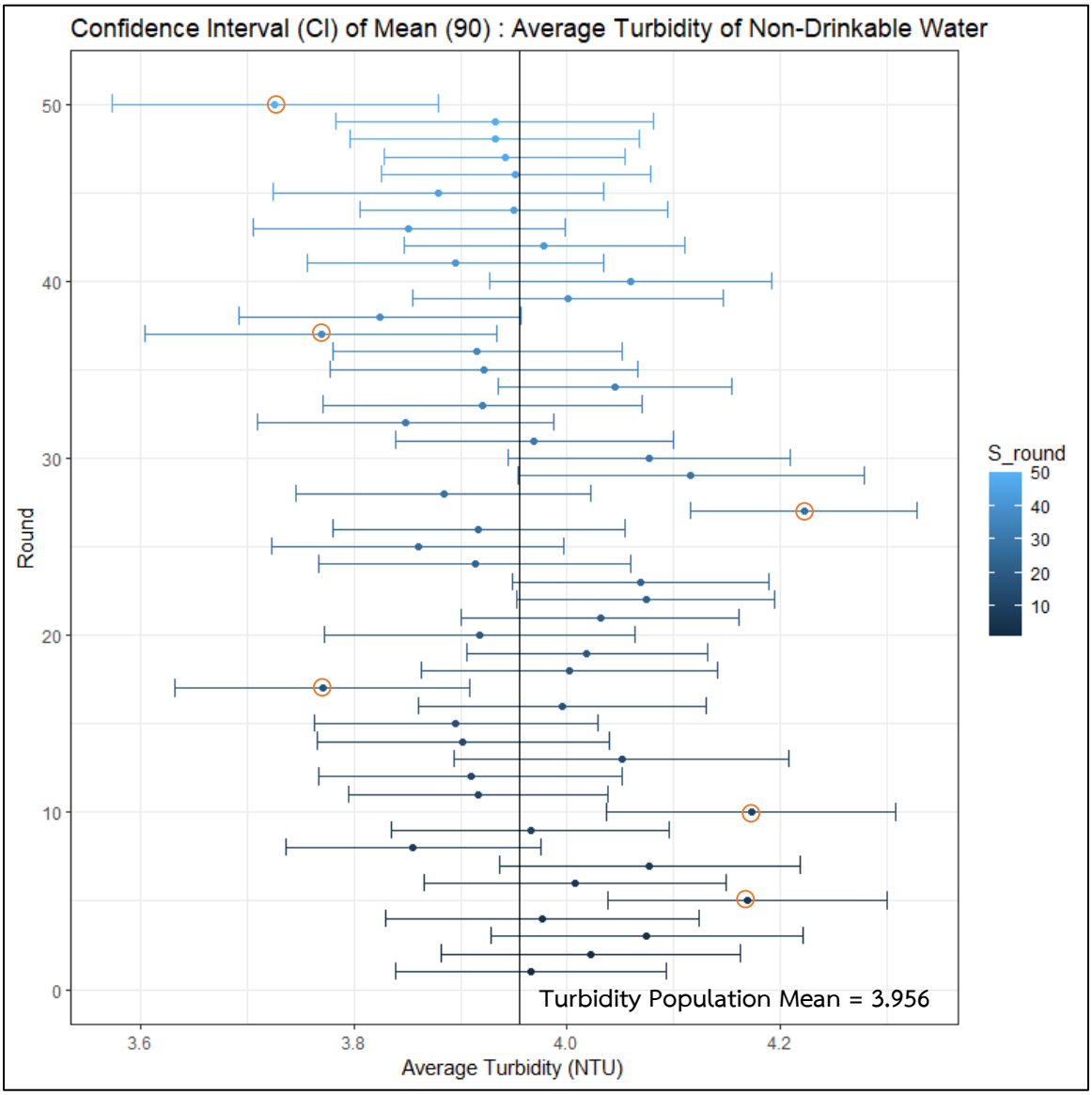


Figure 7: Confidence Interval of Turbidity Mean (Confidence level: 90%) from 50 non-drinkable water samples/round.

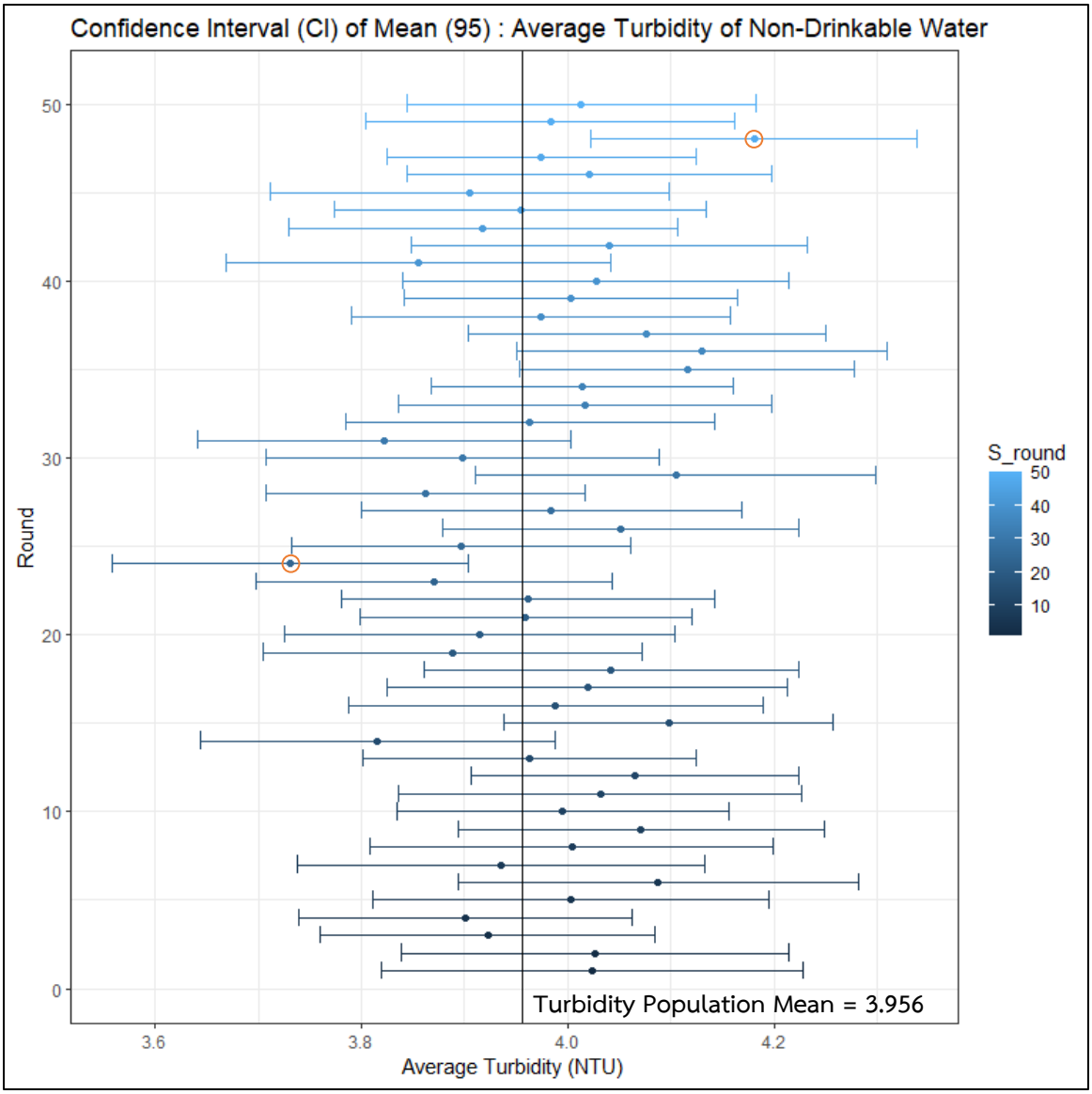


Figure 8: Confidence Interval of Turbidity Mean (Confidence level: 95%) from 50 non-drinkable water samples/round.

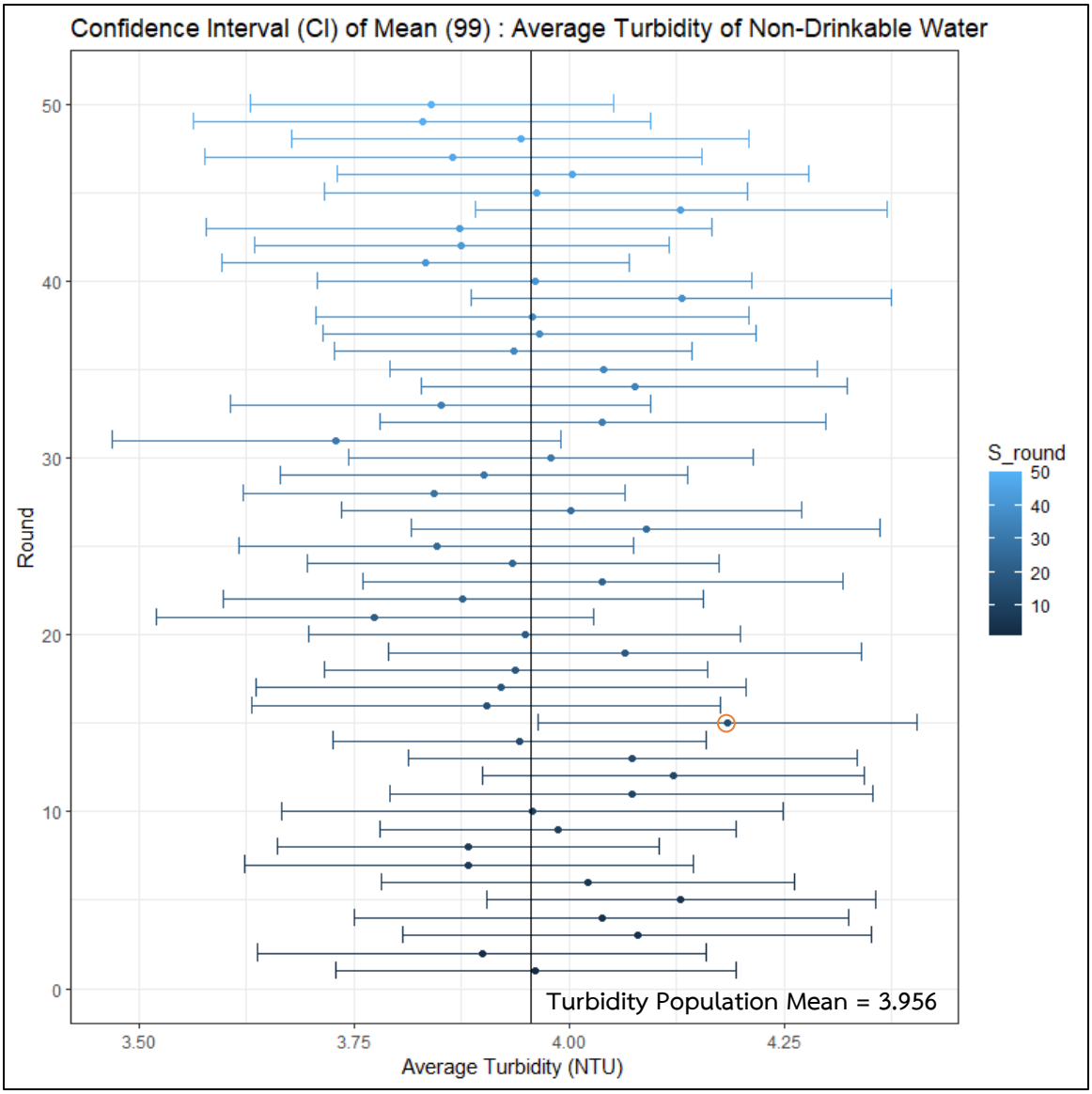


Figure 9: Confidence Interval of Turbidity Mean (Confidence level: 99%) from 50 non-drinkable water samples/round.

- Drinkable Water

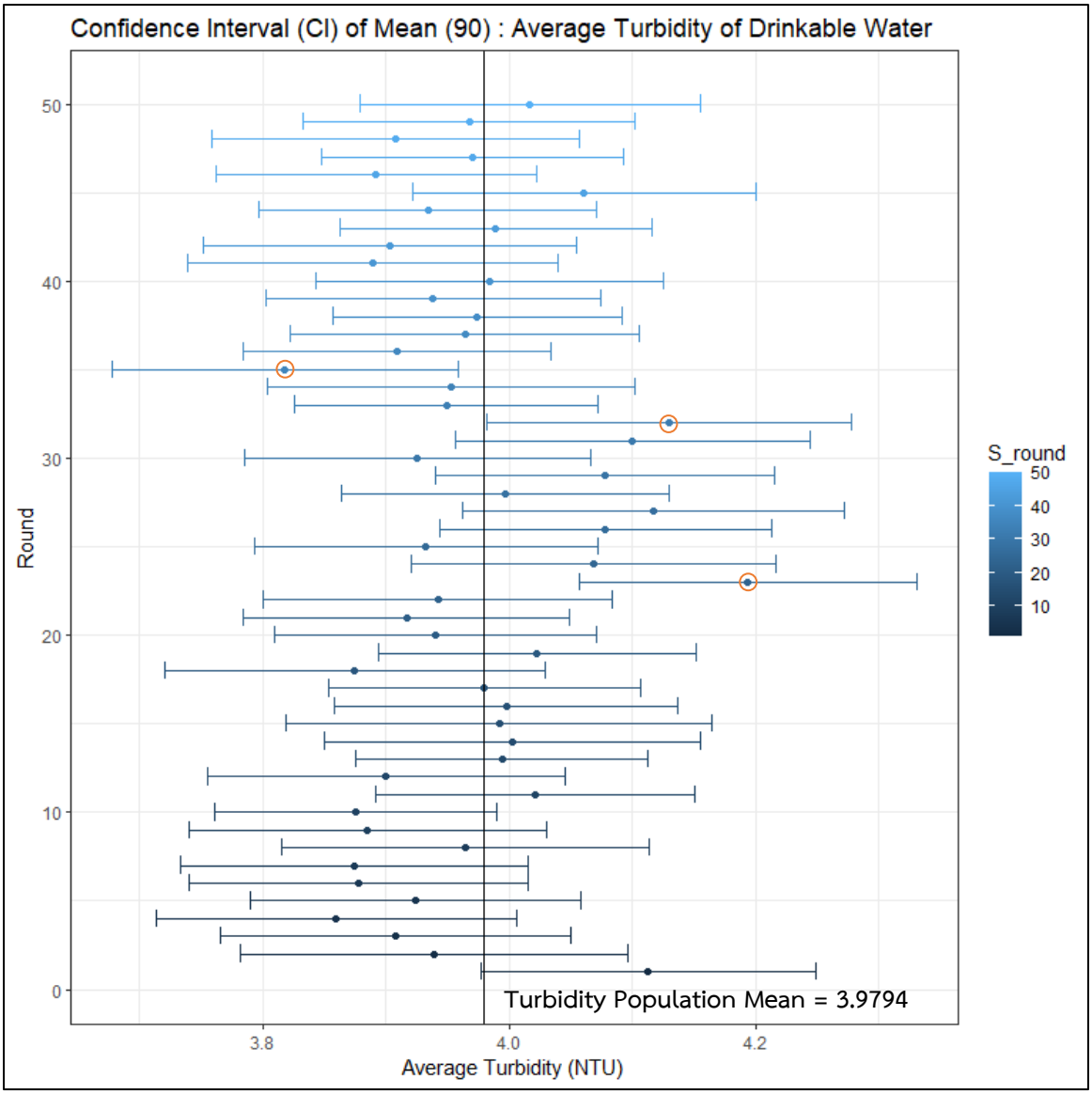


Figure 10: Confidence Interval of Turbidity Mean (Confidence level: 90%) from 50 drinkable water samples/round.

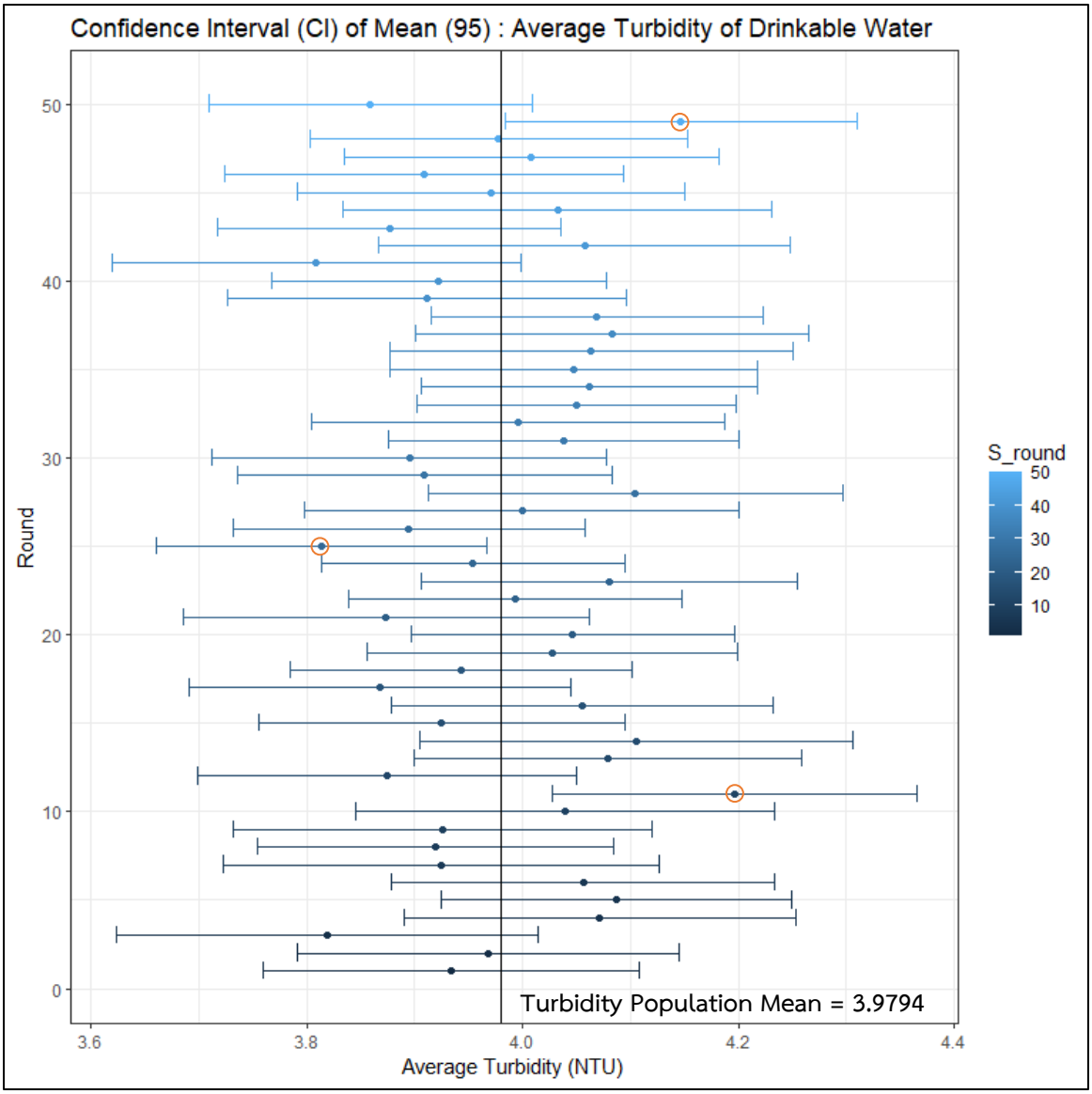


Figure 11: Confidence Interval of Turbidity Mean (Confidence level: 95%) from 50 drinkable water samples/round.

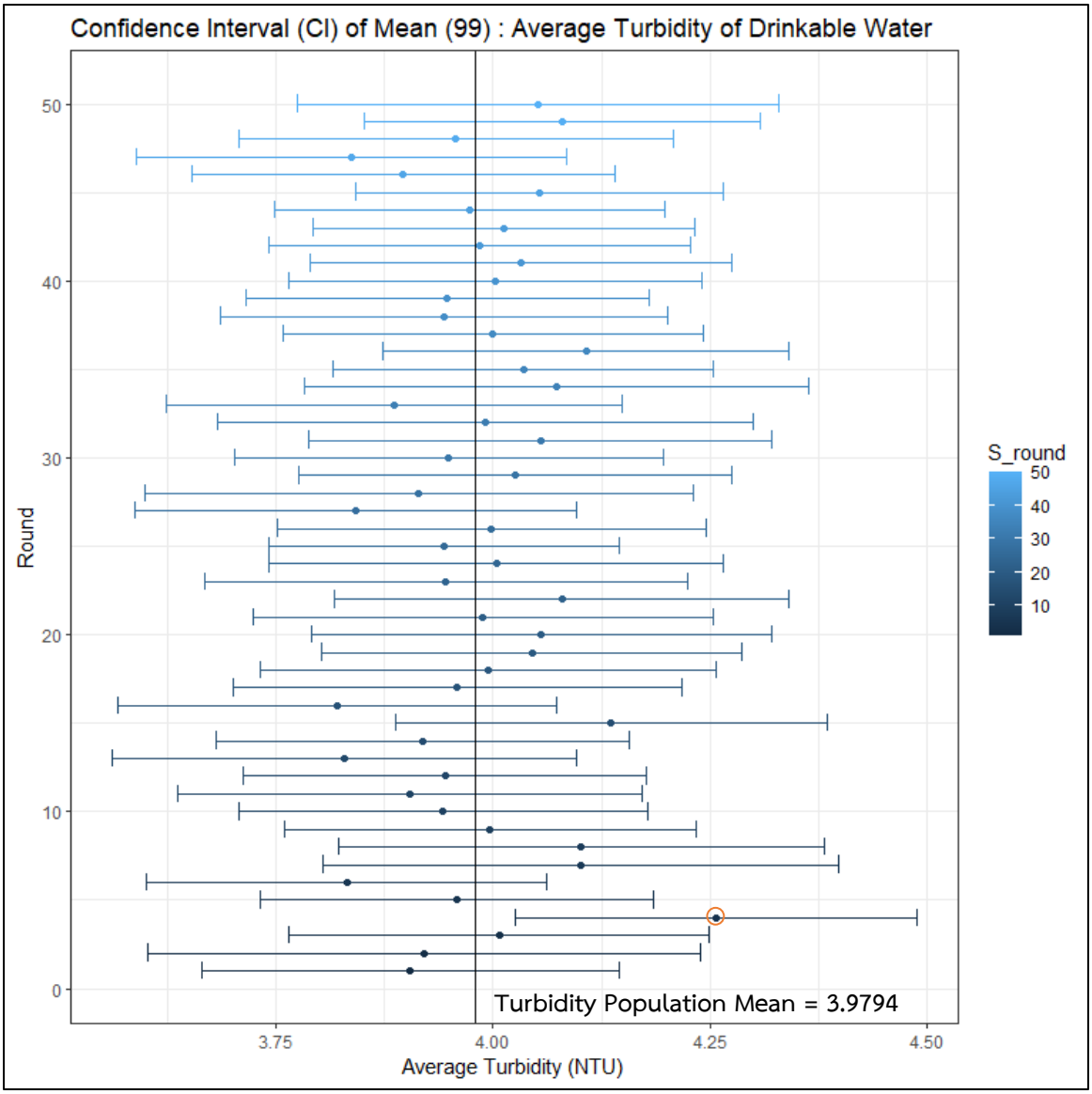


Figure 12: Confidence Interval of Turbidity Mean (Confidence level: 99%) from 50 drinkable water samples/round.

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
- ตัวแปรตามคือ Turbidity(ความขุ่น)

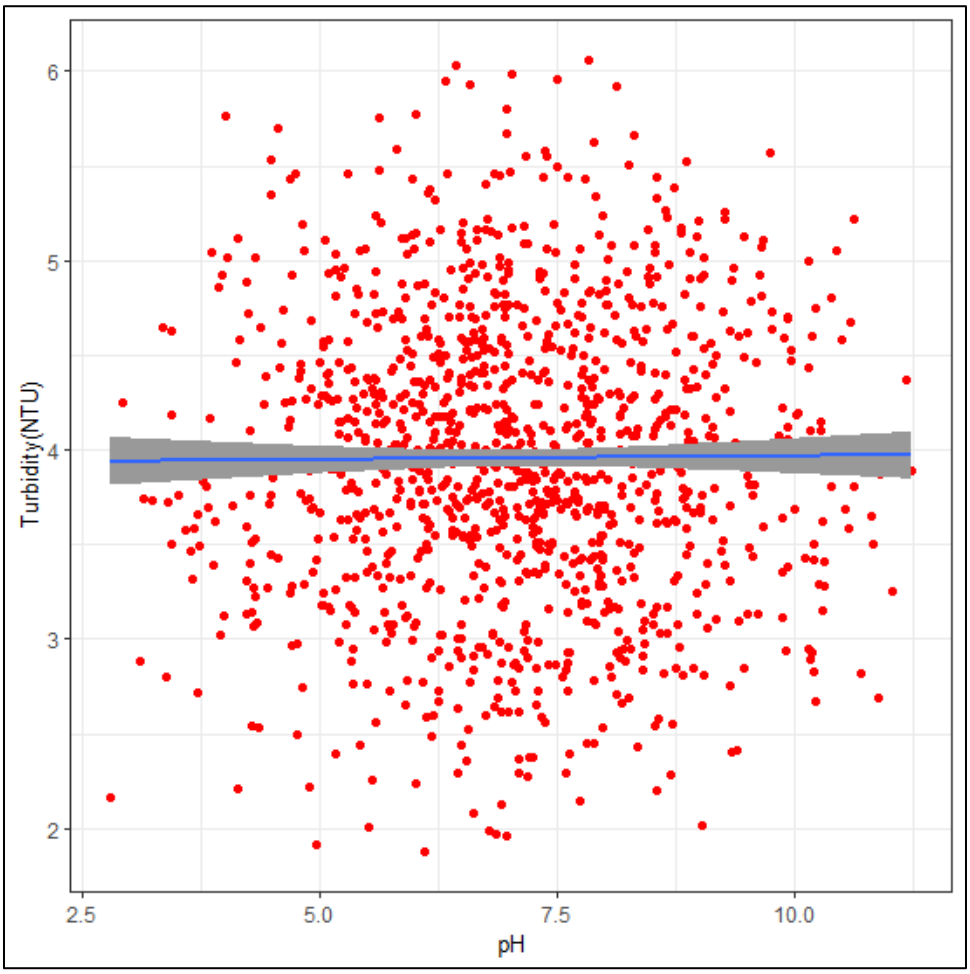


Figure 14: Linear Regression Line with Scatterplot XY: X(pH), Y(Turbidity)

จะได้สมการคือ $y = 0.004x + 3.9277$ โดยมีค่า $r = 0.0081$ ซึ่งเป็นค่า r ที่น้อยมากหรือเรียกว่า Weak or No Correlation บ่งบอกได้ถึงข้อมูลทั้งสองคอลัมน์นั้นไม่มีความสัมพันธ์เชิงเส้นตรงต่อกันและกันทาง ผู้จัดทำจึงลองปรับเปลี่ยนการจับคู่ของคอลัมน์ดังผลลัพธ์ถัดไป

```
> x<-data_Pot0_no_out_1$pH
> y<-data_Pot0_no_out_1$Turbidity
> xbar <- mean(x)
> ybar <- mean(y)
> n <- length(y)
> SSxy <- sum(x*y) - n*xbar*ybar
> SSxx <- sum(x^2) - n*xbar^2
> SSyy <- sum(y^2) - n*ybar^2
> r <- SSxy/(sqrt(SSxx)*sqrt(SSyy))
> r
[1] 0.008141539
```

```
> model

Call:
lm(formula = Turbidity ~ pH, data = data_Pot0_no_out_1)

Coefficients:
(Intercept)           pH
   3.927741     0.003998

> tidy(model)
# A tibble: 2 x 5
  term        estimate std.error statistic   p.value
  <chr>          <dbl>     <dbl>     <dbl>     <dbl>
1 (Intercept)  3.93       0.103      38.0   2.16e-207
2 pH           0.00400    0.0143      0.280 7.80e-  1
```

```
ggplot(data_Pot0_no_out_1,aes(pH,Turbidity))+
  geom_point(color = "red")+
  geom_smooth(method = "lm",size = 1, alpha = 1)+
  xlab("pH")+ylab("Turbidity(NTU)")+
  theme(axis.title = element_text(size=10))

model <- lm(Turbidity~pH,data = data_Pot0_no_out_1)
model
tidy(model)
x<-data_Pot0_no_out_1$pH
y<-data_Pot0_no_out_1$Turbidity
xbar <- mean(x)
ybar <- mean(y)
n <- length(y)
# y = 0.307

SSxy <- sum(x*y) - n*xbar*ybar
SSxx <- sum(x^2) - n*xbar^2
SSyy <- sum(y^2) - n*ybar^2
r <- SSxy/(sqrt(SSxx)*sqrt(SSyy))
```

- ตัวแปรต้นคือ pH
- ตัวแปรตามคือ Potability(สามารถบริโภคได้)

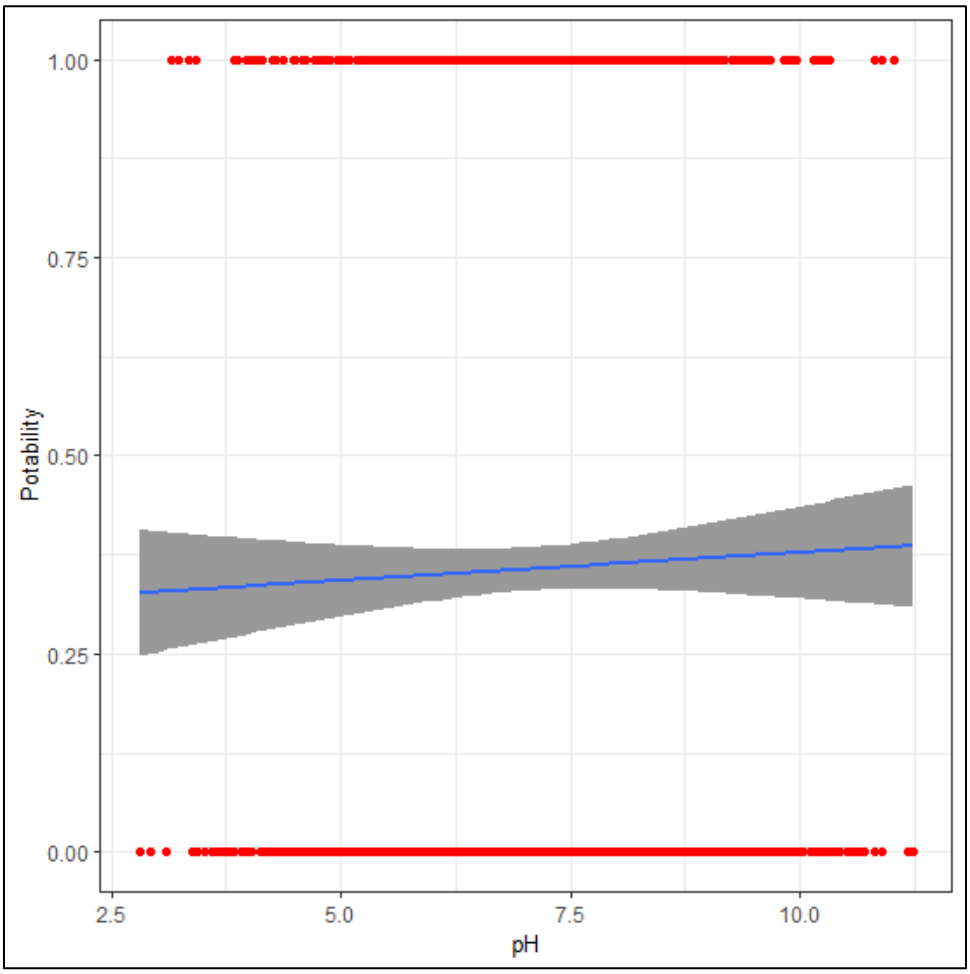


Figure 15: Linear Regression Line with Scatterplot XY: X(pH), Y(Potability)

จะได้สมการคือ $y = 0.0071x + 0.3073$ โดยมีค่า $r = 0.0231$ ซึ่งเป็นค่า r ที่ดีกว่าการจับคู่ระหว่าง pH กับ Turbidity เล็กน้อย คือจาก 0.8% มาสู่ 2.31% แต่ก็ยังคงเป็น Weak or No Correlation อยู่ดี ผู้จัดทำจึงลองปรับเปลี่ยนการจับคู่ของคอลัมน์ดังผลลัพธ์ถัดไป

```
> model

Call:
lm(formula = Potability ~ pH, data = data_Pot0_no_out_1)

Coefficients:
(Intercept)           pH
   0.307341     0.007074

> tidy(model)
# A tibble: 2 x 5
  term        estimate std.error statistic    p.value
  <chr>          <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept)  0.307     0.0643       4.78  0.00000200
2 pH           0.00707   0.00890      0.795 0.427
```

```
> x<-data_Pot0_no_out_1$pH
> y<-data_Pot0_no_out_1$Potability
> xbar <- mean(x)
> ybar <- mean(y)
> n <- length(y)
> SSxy <- sum(x*y) - n*xbar*ybar
> SSxx <- sum(x^2) - n*xbar^2
> SSyy <- sum(y^2) - n*ybar^2
> r <- SSxy/(sqrt(SSxx)*sqrt(SSyy))
> r
[1] 0.02311571
```

```
ggplot(data_Pot0_no_out_1,aes(pH,Potability))+
  geom_point(color = "red")+
  geom_smooth(method = "lm",size = 1, alpha = 1)+
  xlab("pH")+ylab("Potability")+
  theme(axis.title = element_text(size=10))

model <- lm(Potability~pH,data = data_Pot0_no_out_1)
model
tidy(model)
x<-data_Pot0_no_out_1$pH
y<-data_Pot0_no_out_1$Potability
xbar <- mean(x)
ybar <- mean(y)
n <- length(y)
# y = 0.307

SSxy <- sum(x*y) - n*xbar*ybar
SSxx <- sum(x^2) - n*xbar^2
SSyy <- sum(y^2) - n*ybar^2
r <- SSxy/(sqrt(SSxx)*sqrt(SSyy))
```

- ตัวแปรต้นคือ Turbidity(ความขุ่น)
- ตัวแปรตามคือ Potability(สามารถบริโภคได้)

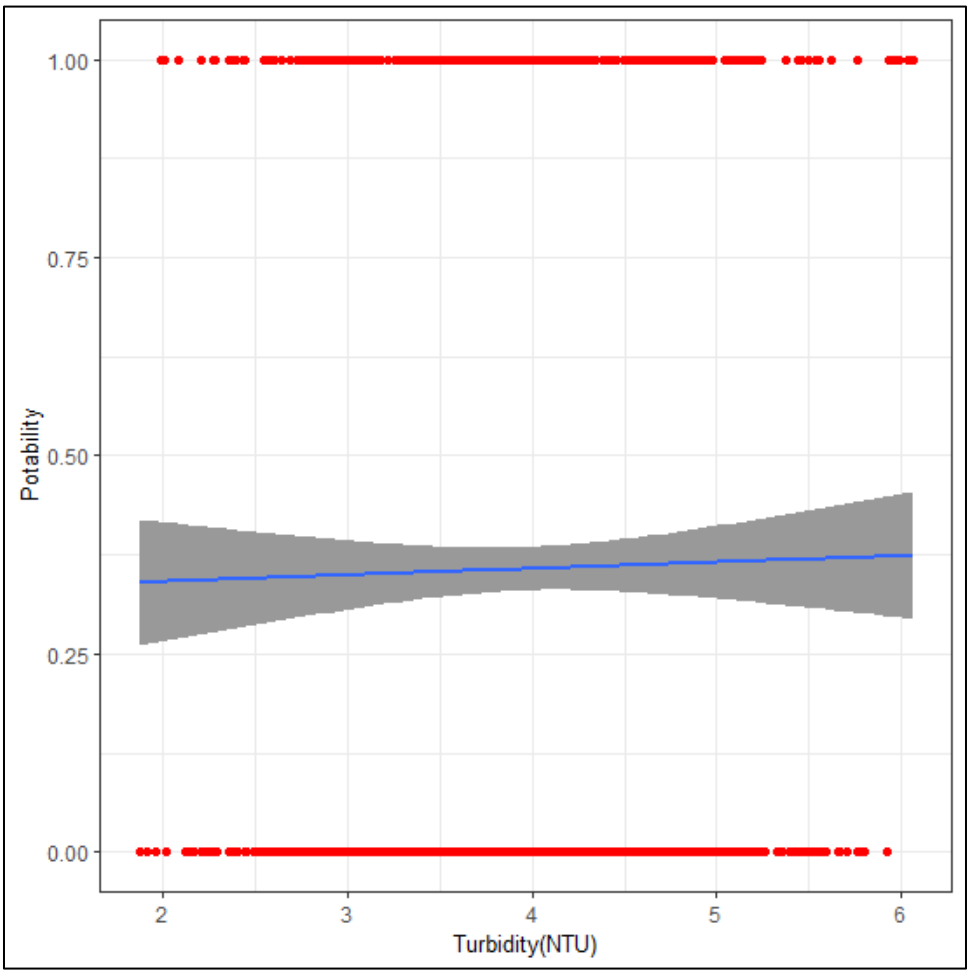


Figure 16: Linear Regression Line with Scatterplot XY: X(pH), Y(Potability)

จะได้สมการคือ $y = 0.0081x + 0.325$ โดยมีค่า $r = 0.0129$ ซึ่งเป็นค่า r ที่ด้อยกว่าการจับคู่ระหว่าง pH กับ Portability เสียอีก

```
> model

Call:
lm(formula = Potability ~ Turbidity, data = data_Pot0_no_out_1)

Coefficients:
(Intercept)    Turbidity
    0.325387     0.008058

> tidy(model)
# A tibble: 2 x 5
  term        estimate std.error statistic    p.value
  <chr>          <dbl>     <dbl>     <dbl>      <dbl>
1 (Intercept)  0.325      0.0730     4.46  0.0000917
2 Turbidity    0.00806    0.0181     0.445 0.657
> x<-data_Pot0_no_out_1$Turbidity
> y<-data_Pot0_no_out_1$Potability
> xbar <- mean(x)
> ybar <- mean(y)
> n <- length(y)
> SSxy <- sum(x*y) - n*xbar*ybar
> SSxx <- sum(x^2) - n*xbar^2
> SSyy <- sum(y^2) - n*ybar^2
> r <- SSxy/(sqrt(SSxx)*sqrt(SSyy))
> r
[1] 0.0129315
```

```
ggplot(data_Pot0_no_out_1,aes(Turbidity,Potability))+
  geom_point(color = "red")+
  geom_smooth(method = "lm",size = 1, alpha = 1)+
  xlab("Turbidity")+ylab("Potability")+
  theme(axis.title = element_text(size=10))

model <- lm(Potability~pH,data = data_Pot0_no_out_1)
model
tidy(model)
x<-data_Pot0_no_out_1$Turbidity
y<-data_Pot0_no_out_1$Potability
xbar <- mean(x)
ybar <- mean(y)
n <- length(y)
# y = 0.307

SSxy <- sum(x*y) - n*xbar*ybar
SSxx <- sum(x^2) - n*xbar^2
SSyy <- sum(y^2) - n*ybar^2
r <- SSxy/(sqrt(SSxx)*sqrt(SSyy))
```

<u>วิเคราะห์ข้อมูลจากกราฟ</u>

บทวิเคราะห์ตามหลักคณิตศาสตร์

จากกราฟ Linear Regression ทั้งสามกราฟที่ได้ ดังนี้

$(\text{pH, Turbidity}) : y = 0.004x + 3.9277 \; ; r = 0.0081$

$(\text{pH, Potability}) : y = 0.0071x + 0.3073 \; ; r = 0.0231$

$(\text{Turbidity, Potability}) : y = 0.0081x + 0.325 \; ; r = 0.0129$

จากกราฟ Linear Regression ทั้งสามกราฟที่ได้จัดทำขึ้นซึ่งเป็นการจับคู่กันของคอลัมน์ข้อมูลที่สนใจทั้ง 3 คอลัมน์ พบว่าความชันของแต่ละกราฟมีค่าเป็นบวก บ่งบอกถึงว่าข้อมูลของแต่ละกราฟมีแนวโน้มที่จะแปรผันตามกัน อย่างเช่น หากน้ำที่มีค่า 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

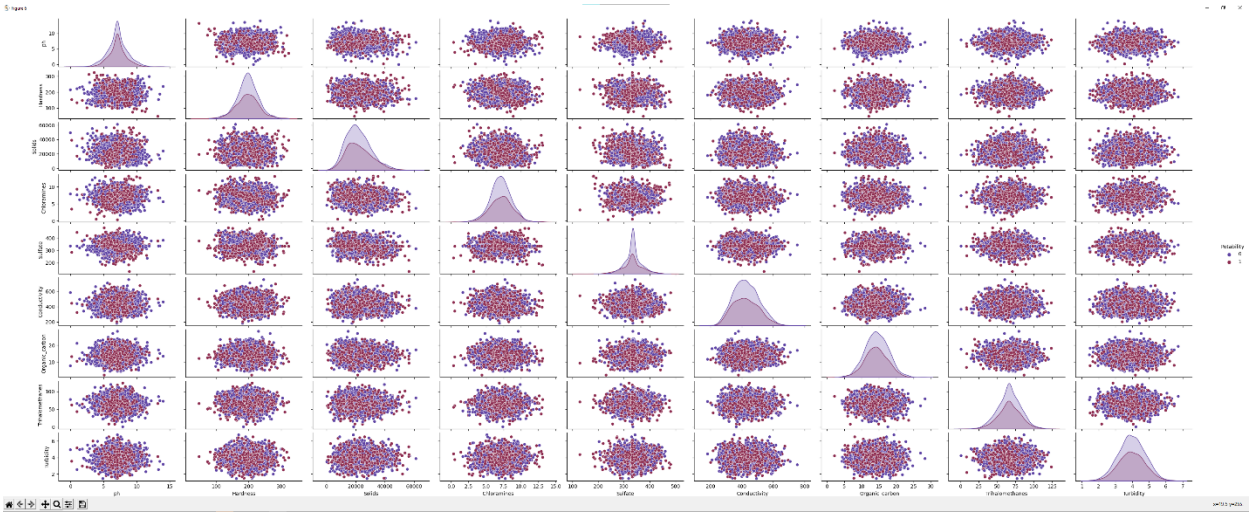


Figure 17: Scatterplot for all column

สรุปผลการศึกษาและเสนอแนะแนวทางการศึกษาเพิ่มเติม

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

```
In [40]:
#Logistic Regression
logreg = LogisticRegression()
logreg.fit(X_train, y_train)

acc_log = round(logreg.score(X_train, y_train) * 100, 2)
acc_log

Out[40]:
62.64
```

```
In [41]:
#Support Vector Classifier
svc = SVC()
svc.fit(X_train, y_train)

acc_svc = round(svc.score(X_train, y_train) * 100, 2)
acc_svc

Out[41]:
73.5
```

```
In [42]:
#KNeighbors Classifier
knn = KNeighborsClassifier(n_neighbors = 3)
knn.fit(X_train, y_train)

acc_knn = round(knn.score(X_train, y_train) * 100, 2)
acc_knn

Out[42]:
78.27
```

```
In [43]:
#GaussianNB
gaussian = GaussianNB()
gaussian.fit(X_train, y_train)

acc_gaussian = round(gaussian.score(X_train, y_train) * 100, 2)
acc_gaussian

Out[43]:
61.54
```

```
In [45]:
#Random Forest
random_forest = RandomForestClassifier(n_estimators=100)
random_forest.fit(X_train, y_train)

random_forest.score(X_train, y_train)
acc_random_forest = round(random_forest.score(X_train, y_train) * 100, 2)
acc_random_forest

Out[45]:
100.0
```

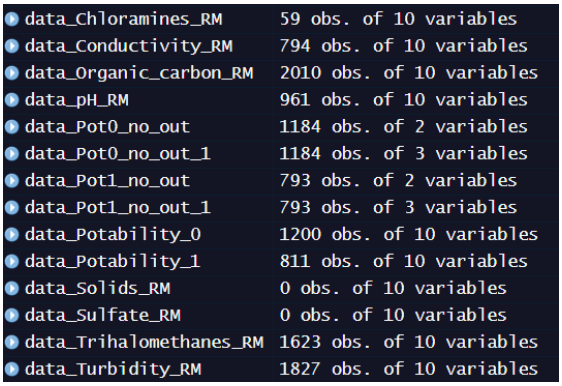
```
In [46]:
models = pd.DataFrame({
    'Model': ['Support Vector Machines', 'KNN', 'Logistic Regression',
              'Random Forest', 'Naive Bayes',
              'Decision Tree'],
    'Score': [acc_svc, acc_knn, acc_log,
              acc_random_forest, acc_gaussian, acc_decision_tree]})
models.sort_values(by='Score', ascending=False)

Out[46]:
```

	Model	Score
3	Random Forest	100.00
5	Decision Tree	100.00
1	KNN	78.27
0	Support Vector Machines	73.50
2	Logistic Regression	62.64
4	Naive Bayes	61.54

จะเห็นได้ว่าการจัดการข้อมูลแบบอื่นๆ เหมาะสมและสามารถทำนายได้ดีกว่าการทำแบบ Linear Regression

ดังนั้นแนวทางการศึกษาเพิ่มเติมคือ

1. ควรจะใช้การจัดการข้อมูลแบบอื่น ๆ ในการสร้างโมเดลทำนายแนวโน้ม
2. ควรศึกษาพารามิเตอร์ที่สำคัญเกี่ยวกับคุณภาพน้ำให้ละเอียดเพื่อให้สามารถตัดสินใจได้ว่าควรตัดสินใจเลือกพารามิเตอร์ใดเป็นหลัก 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