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 :

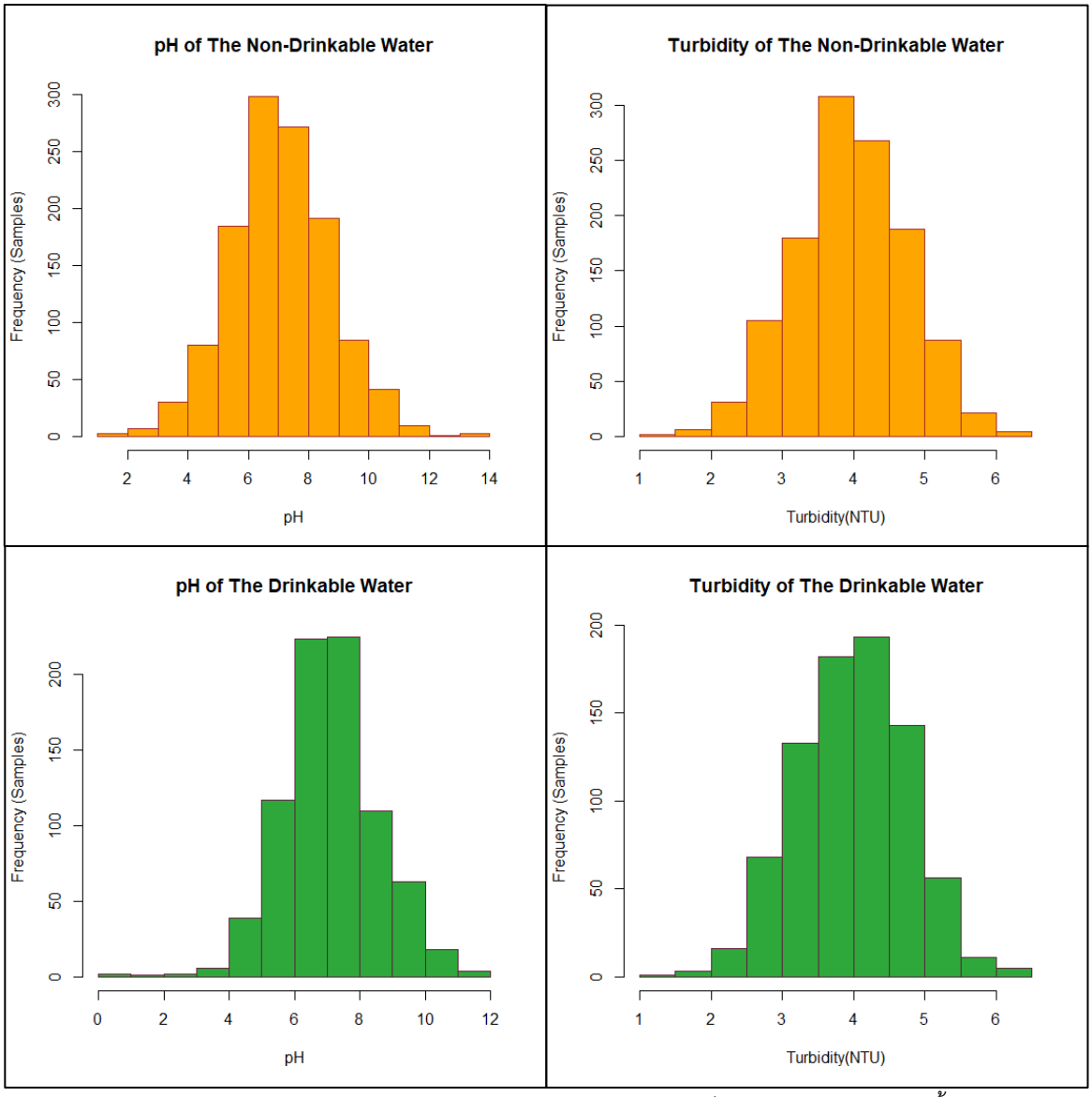
	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
4	8.316766	214.37339	22018.417	8.059332	356.8861	363.2665	18.436524	100.34167	4.628771	0
5	9.092223	181.10151	17978.986	6.546600	310.1357	398.4108	11.558279	31.99799	4.075075	0
6	5.584087	188.31332	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

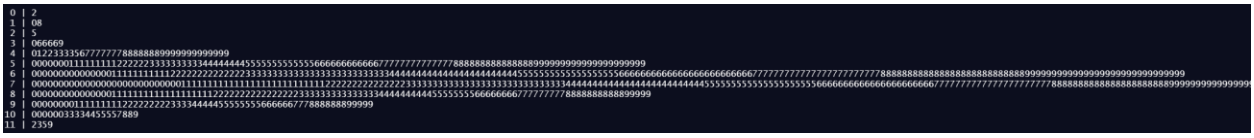
Histogram :



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

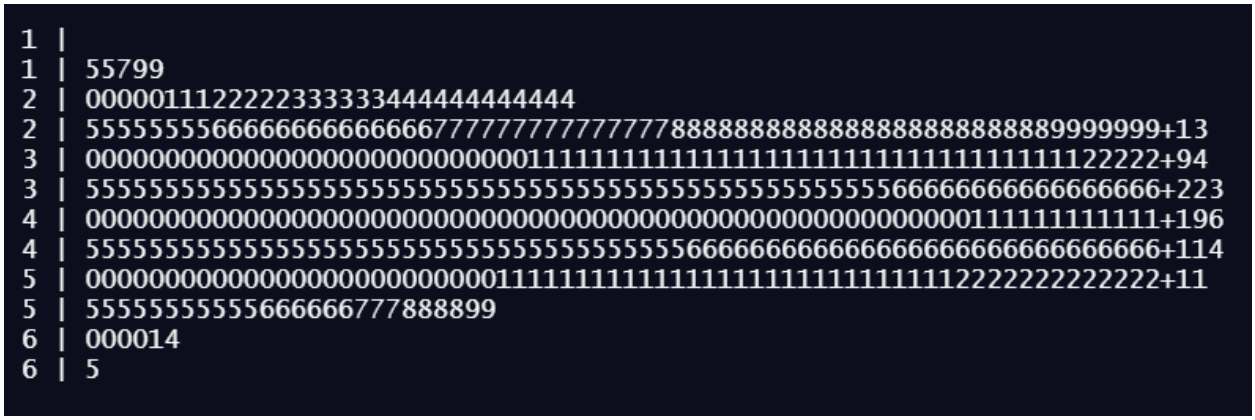
Stem and Leaf :

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

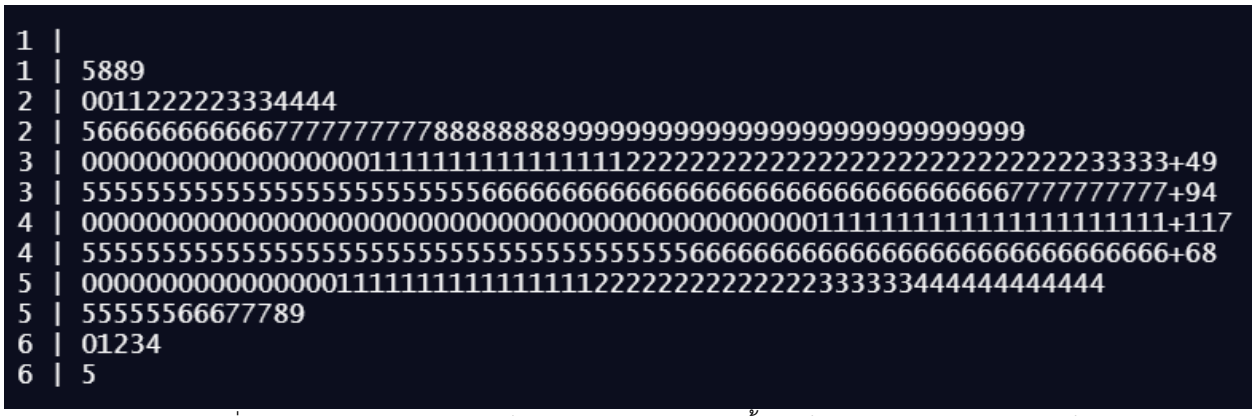


    - **Drinkable Water**

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

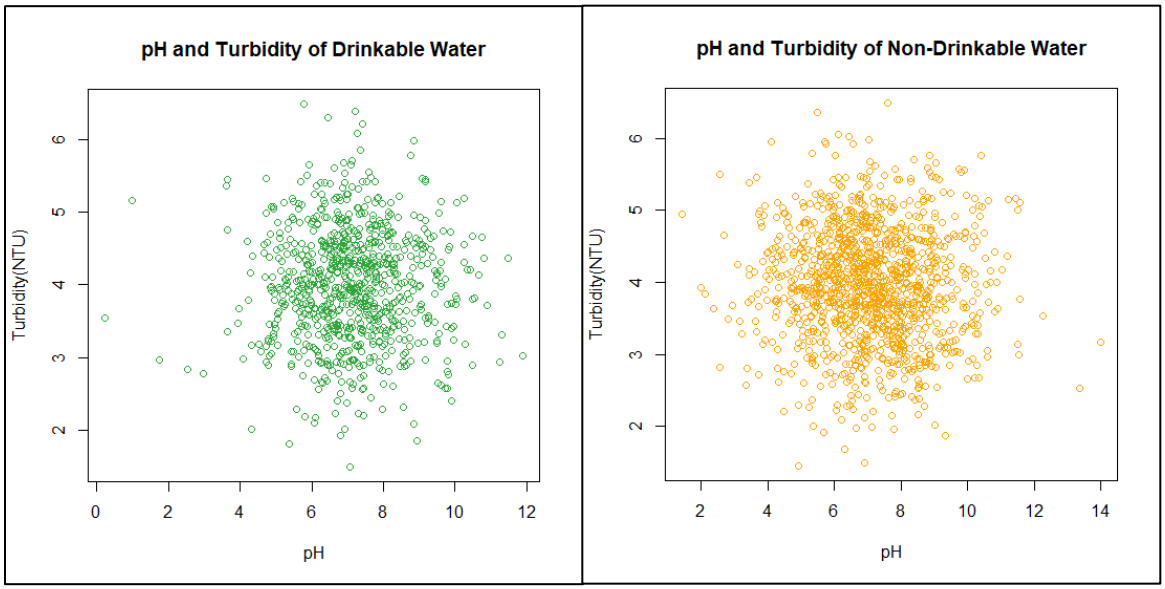


    - **Drinkable Water**



หมายเหตุ : เนื่องจากข้อมูลมีจำนวนมากจึงไม่สามารถแสดงออกมาทั้งหมดได้ เพราะจะทำให้เกิดความไม่สะอาดในการรายงานผู้จัดทำจึงกำหนดความกว้างของ leaf ไว้เพียง 80 หน่วยเท่านั้น

Scatter Plot :

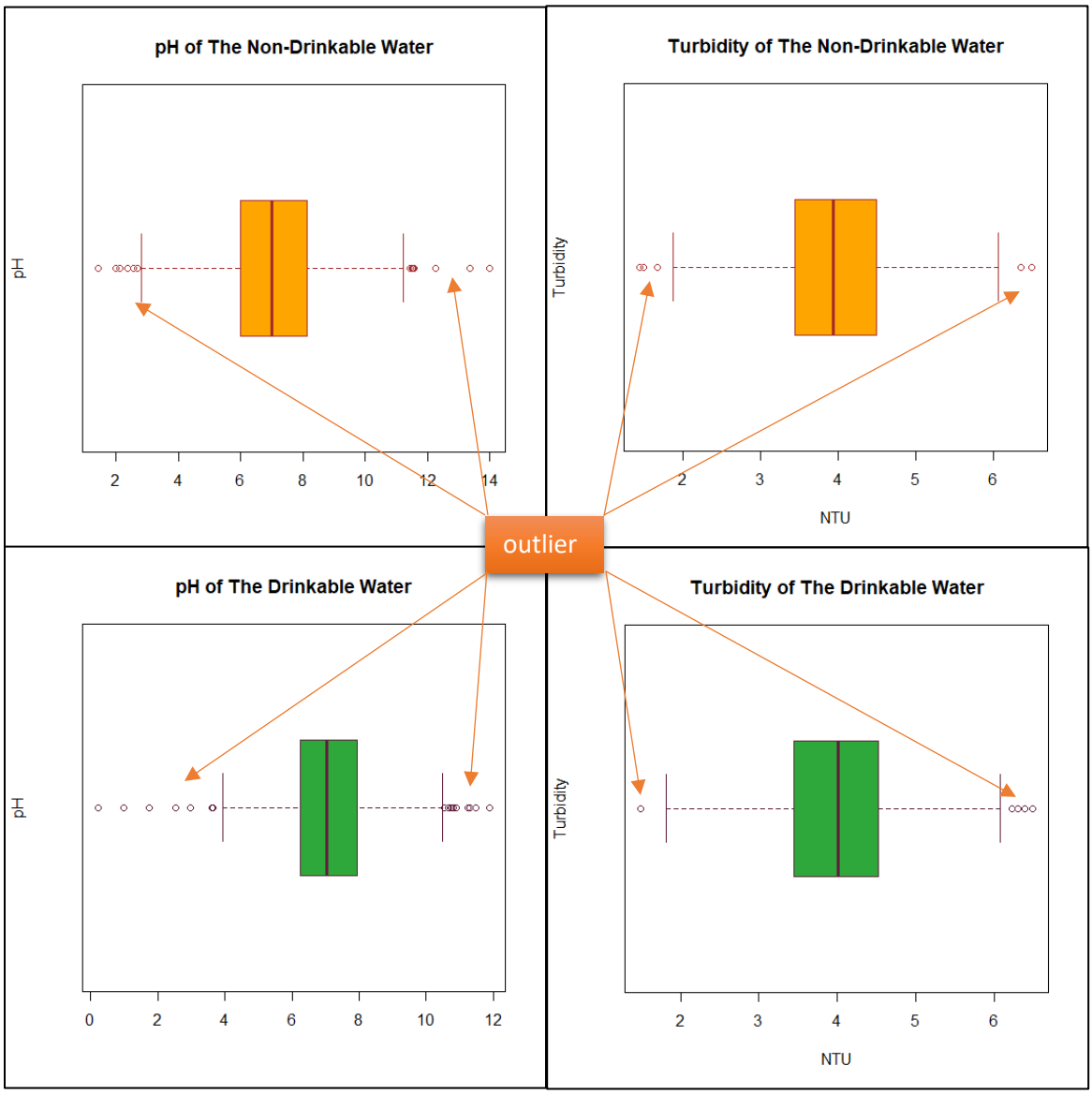


ตัวแปรต้น : pH

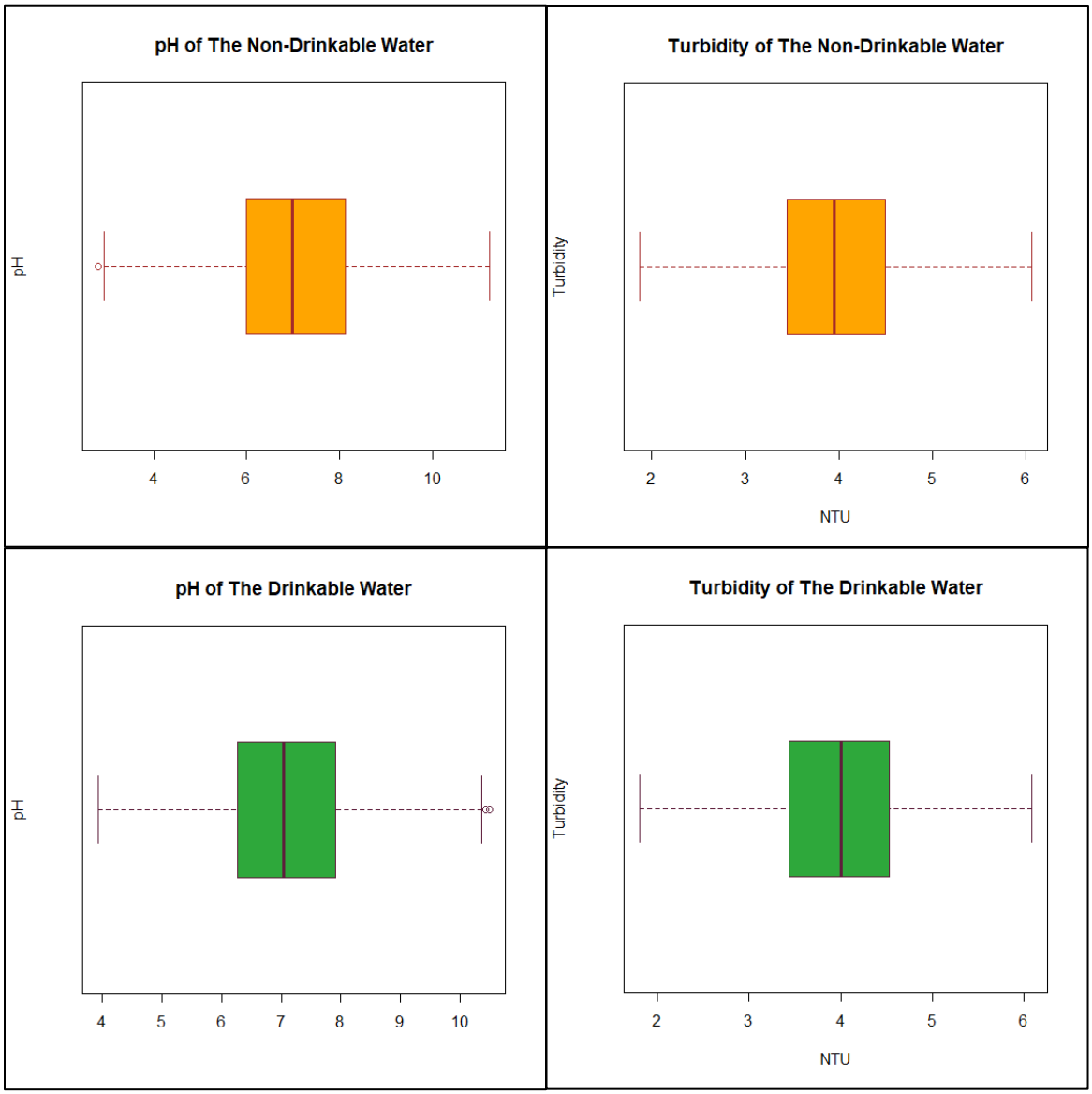
ตัวแปรตาม : Turbidity

Boxplot :

Before remove outlier :

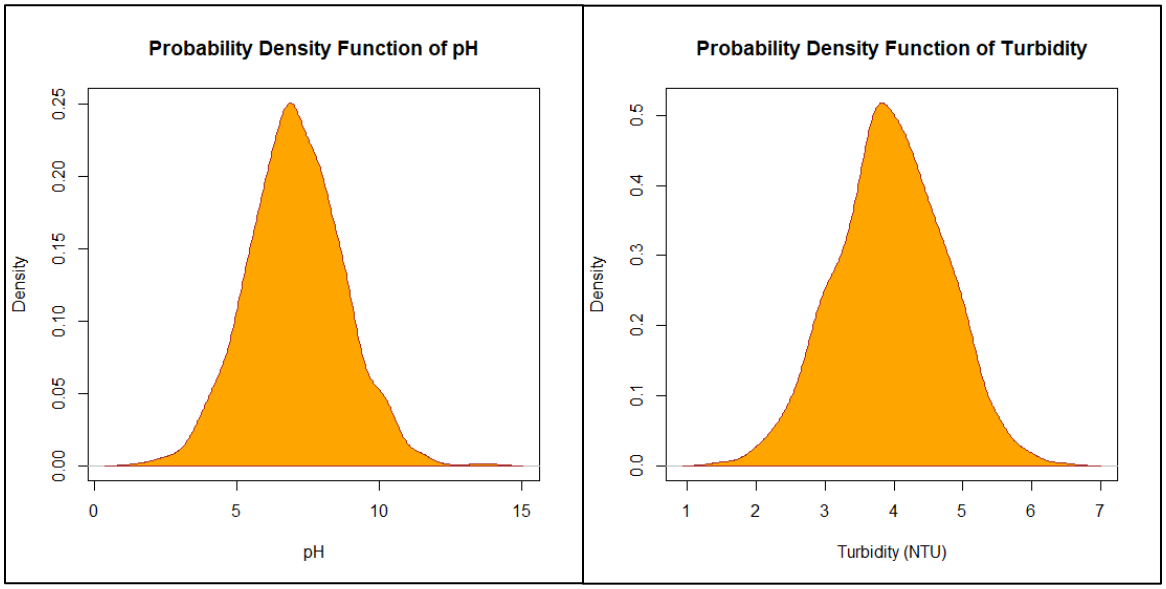


After remove outlier :

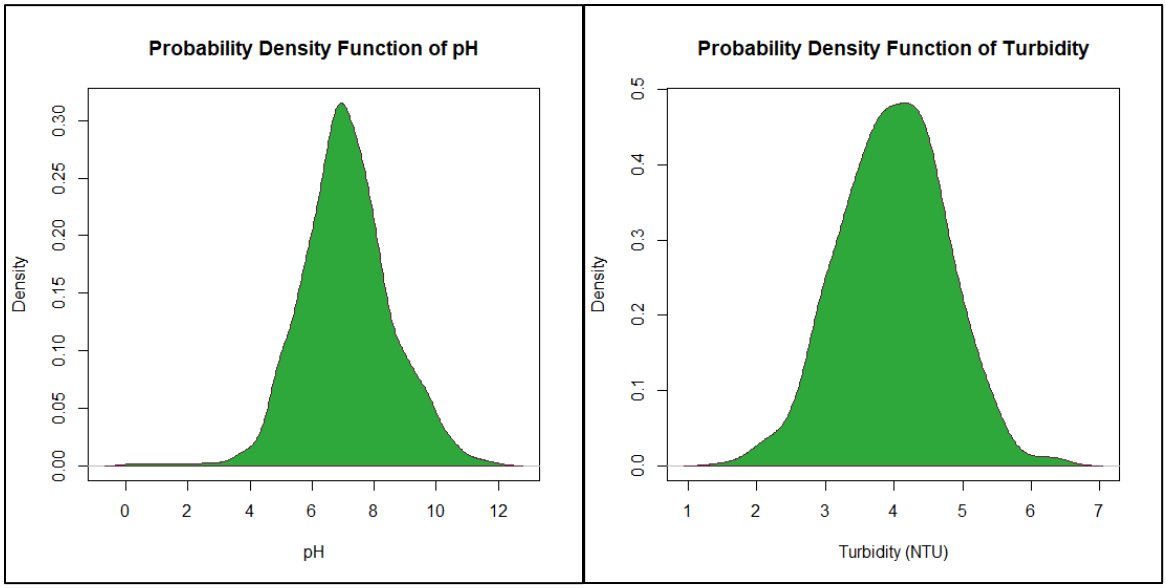


Probability Density Function :

Non-Drinkable Water

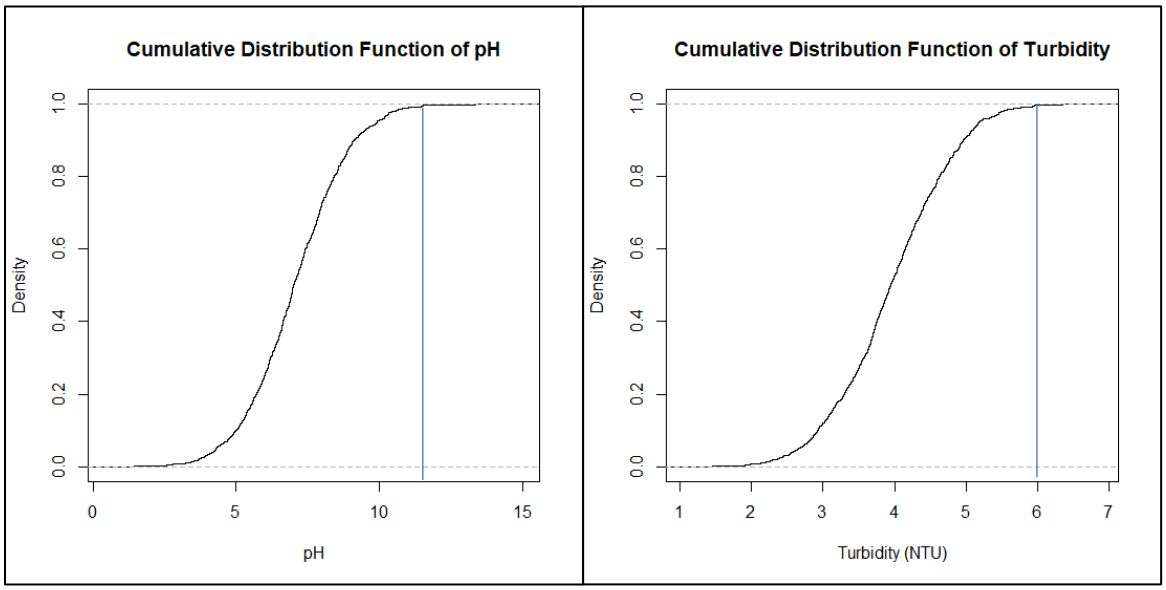


Drinkable Water

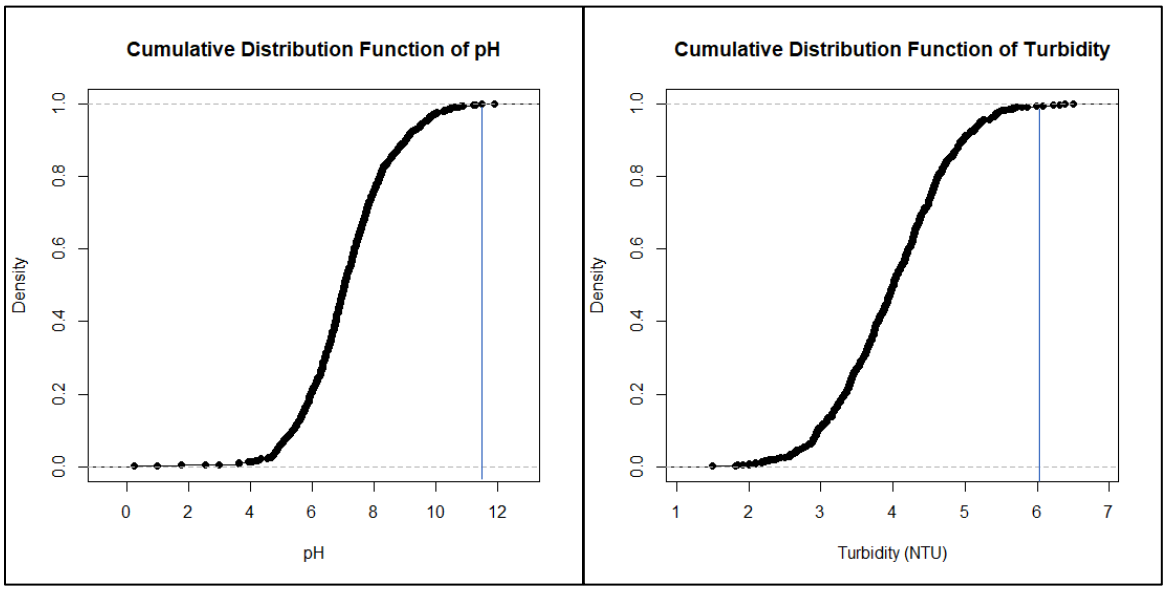


Cumulative Distribution Function :

Non-Drinkable Water



Drinkable Water



บทวิเคราะห์

จากการวิเคราะห์ Histogram ของน้ำที่สามารถนำมาบริโภคได้ pH จะอยู่ในช่วง 5-9 ส่วนของน้ำที่บริโภคไม่ได้นั้นจะอยู่ในช่วง 4-10 จะเห็นว่าน้ำที่นำมาบริโภคได้นั้นโดยมากแล้วมีค่าเป็นกลางส่วนน้ำที่บริโภคไม่ได้มีค่าคววามกว้างของ pH มากกว่าคือเป็นกรดแก่และเบสแก่ ซึ่งไม่สามารถนำมาบริโภคได้ โดยอ้างอิงจากน้ำที่นำมาบริโภคได้ต้องมีค่า pH ระหว่าง 6.5-8.5 ตามเกณฑ์ขององค์การอนามัยโลก

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) อยู่ไมนช่วง 3-5 NTU

สรุปผล

น้ำที่สามารถบริโภคได้นั้นจากกลุ่มตัวอย่างจะมีค่า pH เป็นกลางคือช่วง 6-8 และในส่วนของค่าความขุ่นจะอยู่ในช่วง 3-5 NTU โดยยังมีค่าที่ทับซ้อนกันอยู่บางส่วนซึ่งสามารถคำนวณเป็น Error ของการคำนวณได้แต่ก็ยากที่จะสรุปให้ลงตัวได้ ยกตัวอย่างเช่น ผลกระทบของ pH ต่อความขุ่นของน้ำ ทั้งนี้หากต้องการความชัดเจนที่มากขึ้นควรจะใช้คอลลัมน์อื่นๆเพื่อมาประกอบการพิจารณาด้วย เช่น จำนนวนของคาร์บอนที่ถูกสร้างจากแบคทีเรียในน้ำเป็นต้น

Code :

```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)
)
```

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

```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))
```

```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",
)
```

```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
        )
```

```r
#Remove outlier method
#pH
q1_pH <- quantile(ph, .25)
q3_pH <- quantile(ph, .75)
iqr_pH <- IQR(ph)
no_outliers_pH <- subset(dataRm_Na$ph, ph > (q1_pH - 1.5*iqr_pH) & ph < (q3_pH + 1.5*iqr_pH))
#Turbidity
q1_Turb <- quantile(turbidity, .25)
q3_Turb <- quantile(turbidity, .75)
iqr_Turb <- IQR(turbidity)
no_outliers_Turb <- subset(dataRm_Na$Turbidity, turbidity > (q1_Turb - 1.5*iqr_Turb) & turbidity < (q3_Turb + 1.5*iqr_Turb))
```

```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
)
```

Source Code file :

https://drive.google.com/file/d/182M48Lo26oLYOgRzpCp3cKMKcw1kes_I/view?usp=sharing