

Homework #2

วิเคราะห์สถิติและกราฟ คุณภาพของไวน์แดง

ผู้จัดทำเลือกทั้งหมด 2 จาก 3 คอลัมน์ แต่ละคอลัมน์มีทั้งหมด 1599 แถว:

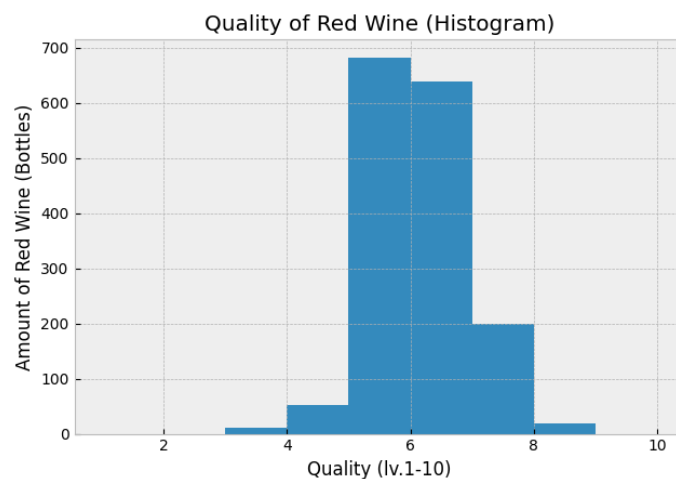
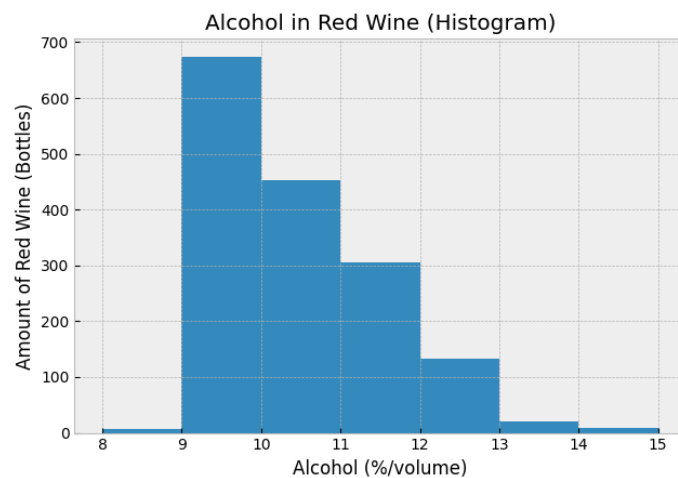
1. Alcohol (g/liter)
2. Quality (lv. 1-10)

คำนวณค่าสถิติพื้นฐาน

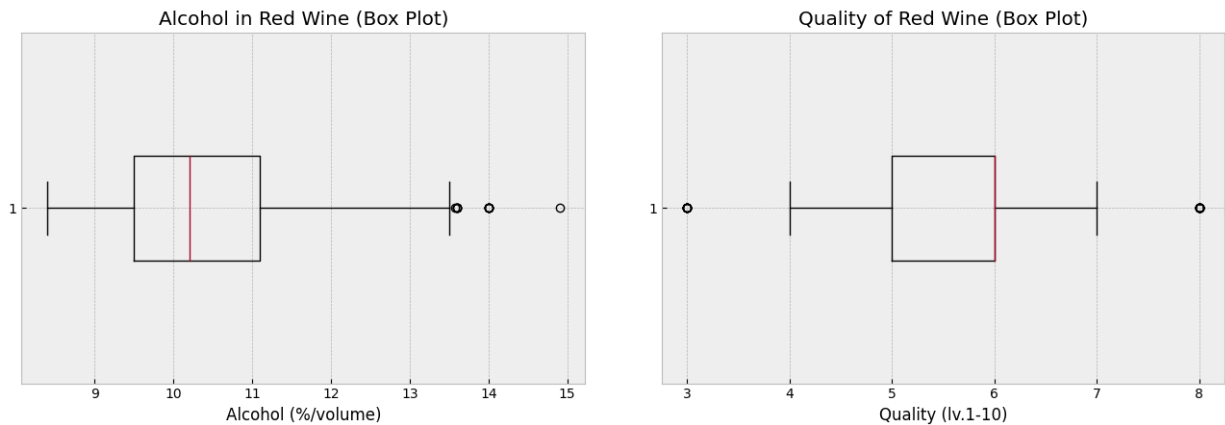
	Min	Mean	Median	Max	Mode	Sample Standard Deviation	Sample Variation
1.Alcohol (%/volume)	8.4	10.42	10.20	14.9	9.5	1.0656	1.1356
2.Quality (lv.1-10)	3	5.6	6.0	8	5	0.8075	0.6521

อ้างอิงจากการคำนวณในโปรแกรม WineGraph.py

Histogram Graphs



Box Plot Graphs



การคำนวณหา Outliers

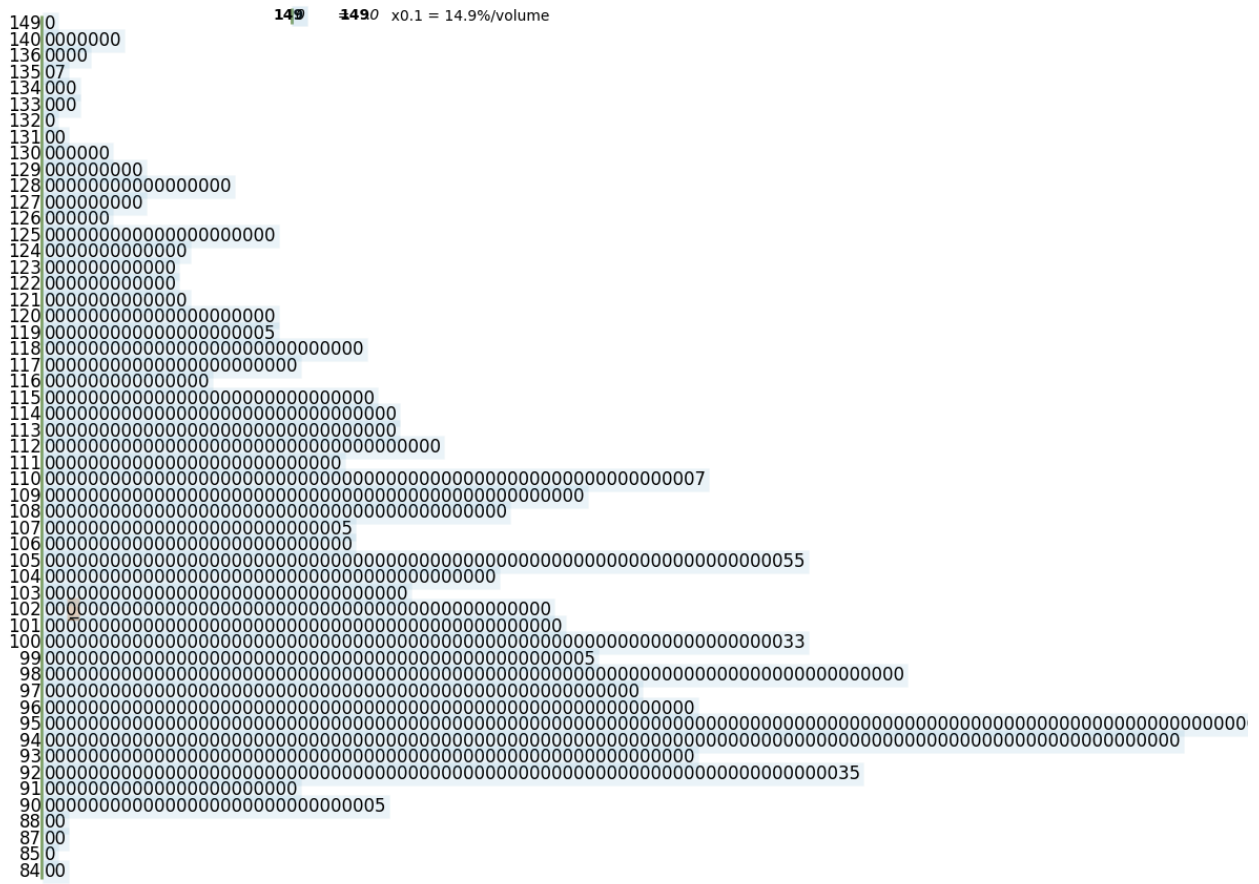
สามารถดูได้จากกราฟ Box Plot หรือจากการคำนวณหา Outliers โดยใช้สูตรดังนี้

- 1) $Q1 = (N+1) * 1/4$ (คิดแบบสูตรการไม่แจกแจงความถี่ / เห็นข้อมูลเป็นตัว ไม่ใช่ช่วง)
 - 2) $Q3 = (N+1) * 3/4$ (คิดแบบสูตรการไม่แจกแจงความถี่ / เห็นข้อมูลเป็นตัว ไม่ใช่ช่วง)
 - 3) $IQR = Q3 - Q1$ (คิดแบบ Inclusive เพื่อให้ 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. <= value < Mild Outlier Lower B.)
 - 10) Mild Outlier (Upper) = if (Mild Outlier Upper B. < value <= Ext. Outlier Upper B.)
 - 11) Extreme Outlier (Upper) = if (value >= Ext. Outlier Upper B.)
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, 14.0, 14.0, 14.0, 14.0, 14.0, 14.0, 14.0, 14.9]
 - 4) Extreme Outlier (Upper) = []
 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) Mild Outlier (Upper) = [8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8]
 - 4) Extreme Outlier (Upper) = []

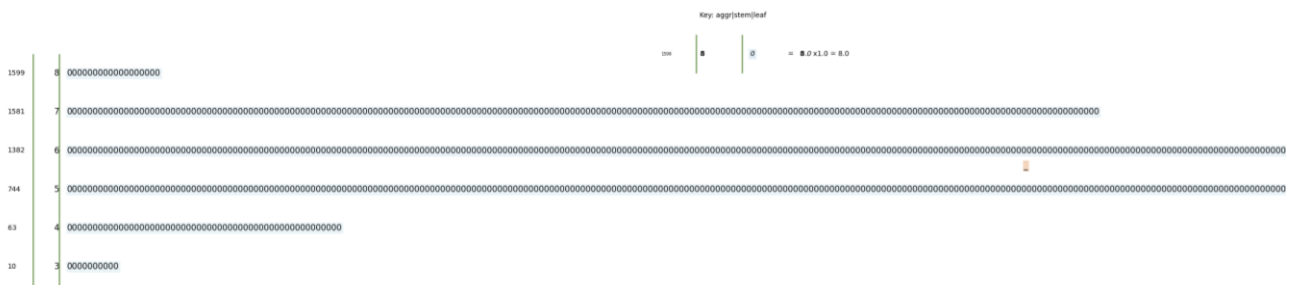
(รายละเอียดในแต่ละค่าเพิ่มเติม สามารถดูได้ใน OUTPUT ของโปรแกรม WineGraph.py ในหน้าที่ 10)

Stem And Leaf Graph

Alcohol in Red Wine Stem-And-Leaf



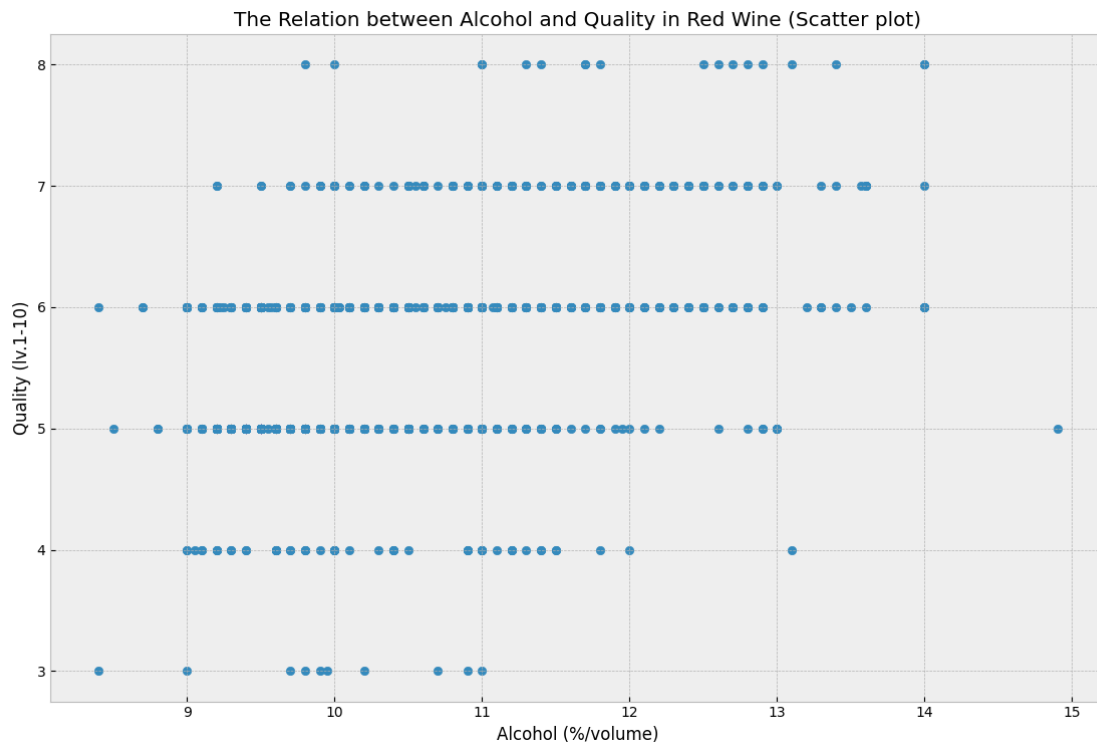
Quality of Red Wine Stem-And-Leaf



*หมายเหตุ กราฟ Quality of Red Wine Stem-And-Leaf สามารถดูเพื่อดู Sum (เก็บแบบ Aggregation) จาก
ด้านหน้า Stem ได้ จะเป็นการรวมกันเป็นลำดับขึ้นไปเรื่อยๆ จากล่างขึ้นบน รวมทั้งหมด 1599 แถวข้อมูล

Scatter Plot Graph

ความสัมพันธ์ระหว่างปริมาณแอลกอฮอล์และคุณภาพของไวน์แดง



กำหนดให้ แนวแกน x เป็นตัวแปรต้น (Independent Variable) = ปริมาณแอลกอฮอล์ (%/ปริมาตร)

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

เหตุผลที่ใช้ปริมาณแอลกอฮอล์เป็นตัวแปรต้น เพราะว่าต้องการที่จะศึกษาเกี่ยวกับปริมาณแอลกอฮอล์ในไวน์แดง ว่ามีผลต่อคุณภาพของไวน์แดงมากแค่ไหน เช่น หากเรามีปริมาณแอลกอฮอล์จำนวน %/ปริมาตร ในไวน์แดงที่มาก อาจทำให้คุณภาพของไวน์แดงมากขึ้นตามด้วย จึงกำหนดให้ปริมาณแอลกอฮอล์เป็นตัวแปรต้น และคุณภาพของไวน์เป็นตัวแปรตาม

บทวิเคราะห์ข้อมูลจากกราฟทั้งหมด

เนื่องจากกราฟ Scatter ความสัมพันธ์ระหว่างปริมาณแอลกอฮอล์และคุณภาพของไวน์แดงในปัจจุบัน มีจำนวน Outliers อยู่จำนวนหนึ่ง ผมจึงทำการลอง plot กราฟ Scatter ใหม่ โดยการตีเส้นและไม่นำ Outliers มาคิดในการวิเคราะห์ จะได้ออกมาเป็นกราฟดังนี้ (อ้างอิงจากการคำนวณในหน้าที่ 2)

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

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



จากกราฟที่ได้ เราจะสนใจแค่เพียงในส่วนสีขาวเท่านั้น (ที่ไม่ใช่แรเงาสีเหลือง) จะเห็นได้ว่า ข้อมูลที่ได้ค่อนข้างมีความกระจุกตัวอยู่บริเวณตรงกลางเป็นส่วนมาก และมีแนวโน้มเอียงขึ้นไปทางบนขวาเล็กน้อย

ซึ่งถ้าหากเราสังเกตกราฟโดยละเอียด จะพบว่า

1. 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 ทางตอนเหนือของประเทศโปรตุเกส

กล่าวโดยสรุปคือ ปริมาณแอลกอฮอล์ในไวน์แดงที่มากขึ้น อาจมีแนวโน้มที่จะทำให้คุณภาพของไวน์แดงเพิ่มขึ้นตามไปด้วย โดยความสัมพันธ์ระหว่างปริมาณแอลกอฮอล์ในไวน์แดง และคุณภาพของไวน์แดงเป็นความสัมพันธ์แบบแปรผันตรงหรือคล้ายตามกัน

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



รายละเอียด Source Code และ OUTPUT ของโปรแกรม 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
62 print('\n-----All Alcohol Outliers Boundaries-----')
63 print("al_extreme_low_bound = ", al_extreme_low_bound)
64 print("al_mild_low_bound = ", al_mild_low_bound)
65 print("al_mild_up_bound = ", al_mild_up_bound)
66 print("al_extreme_up_bound = ", al_extreme_up_bound)
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_low_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
92
93 print('\n\n-----Quality of Wines-----')
94 # Quality calculations
95 qualityMin = min(y)
96 qualityMean = stc.mean(y)
97 qualityMed = stc.median(y)
98 qualityMax = max(y)
99 qualityMode = stc.mode(y)
100 qualitySampleSD = stc.stdev(y)
101 qualitySampleV = stc.variance(y)
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 # Quality Outlier
113 qQt = stc.quantiles(y, method='inclusive')
114 print("Quality[Q1, Q2, Q3]= ",qQt)
115 qu_q1 = qQt[0]
116 qu_q3 = qQt[2]
117 qu_iqr = qu_q3 - qu_q1
118 print("IQR = ", qu_iqr)
119 qu_mild_low_bound = qu_q1 - qu_iqr*1.5
120 qu_extreme_low_bound = qu_q1 - qu_iqr*3
121 qu_mild_up_bound = qu_q3 + qu_iqr*1.5
122 qu_extreme_up_bound = qu_q3 + qu_iqr*3
123
124 print('\n-----All Quality Outliers Boundaries-----')
125 print("qu_extreme_low_bound = ", qu_extreme_low_bound)
126 print("qu_mild_low_bound = ", qu_mild_low_bound)
127 print("qu_mild_up_bound = ", qu_mild_up_bound)
128 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)') #independent
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))
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()

```

OUTPUT ของโปรแกรม 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 = 7.1000000000000005
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 : 0.8075694397347049
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, 3]
Mild Outlier(Upper) = [8, 8, 8, 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

```

แหล่งที่มาของชุดข้อมูล (Reference/URL) :

- ที่มาของชุดข้อมูล Winequality-red.csv

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

- ที่มาคำอธิบายแต่ละส่วนประกอบของไวน์

<https://archive.ics.uci.edu/ml/datasets/wine+quality>

https://rstudio-pubs-static.s3.amazonaws.com/57835_c4ace81da9dc45438ad0c286bcbb4224.html

<https://waterlibrary.com/th-รู้ไหมว่า-ระดับปริมาณแ:/#:~:text=ปัจจุบันมีแอลกอฮอล์อยู่ใน,สูงขึ้นด้วยเช่นกัน>

- วิธีการทำไวน์

https://www.youtube.com/watch?v=7gguYRxLMFI&ab_channel=Insider

- ประเภทของไวน์

<https://www.unlockmen.com/terrazas-unlock-wine-101-1/>

<https://thewinelist.shop/blog/news/wine-101>

- รายละเอียดอื่นๆ เกี่ยวกับคุณภาพและวิธีรับรสที่ดีของไวน์

<https://www.blockdit.com/posts/5e5f68d77b00780ed6462939>

<https://www.dummies.com/food-drink/drinks/wine/the-special-technique-for-tasting-wine/>

<https://www.quickanddirtytips.com/house-home/entertaining/wine/4-ways-to-know-if-your-wine-is-good>