

Coffee Quality

ชื่อชุดข้อมูล ชื่อคอลัมน์ข้อมูลที่น่าสนใจ

- 1.altitude_mean_meters
- 2.Species
- 3.Processing.Method
- 4.Country.of.Origin
- 5.Total.Cup.Points

Why is it interesting?

กาแฟเป็นเครื่องดื่มยอดนิยมสำหรับคนที่ต้องการให้ตนเองนั้นสดชื่น เพื่อเตรียมพร้อมสำหรับวันใหม่กาแฟมีรสชาติที่ซับซ้อน และกลิ่นที่หอม การได้ดื่มกาแฟที่ดั้นด้นยอมทำให้วันนั้นมีแรงกายในการดำเนินชีวิต กาแฟที่ดีย่อมมาจากเมล็ดพันธุ์ที่ดีเช่นกัน ฉะนั้นพวกเราจึงสนใจที่จะนำข้อมูลชุดนี้เพื่อดูความสัมพันธ์ระหว่าง เมล็ดกาแฟที่มีคุณภาพเกี่ยวข้องกับการจัดเก็บ และพื้นที่ปลูกอย่างไร

แหล่งที่มาของข้อมูล

<https://www.kaggle.com/volpatto/coffee-quality-database-from-cqi>

คำอธิบายชื่อคอลัมน์ข้อมูลทีเลือก และวิธีการรวบรวมข้อมูล (Data collection)

- | | |
|------------------------|---|
| 1.altitude_mean_meters | - ค่าเฉลี่ยของความสูงที่ใช้ในการเพาะปลูกเมล็ดกาแฟ |
| 2.Species | - ประเภทของเมล็ดกาแฟ |
| 3.Processing.Method | - รูปแบบของการบ่มเพาะกาแฟ |
| 4.Country.of.Origin | - ประเทศที่แหล่งปลูกเมล็ดกาแฟ |
| 5.Total.Cup.Points | - คะแนนโดยรวมของกาแฟ |

วิธีการรวบรวมข้อมูล

เป็นข้อมูลที่รวบรวมมาจาก www.coffeeinstitute.org ซึ่งเป็นองค์กรที่รวบรวมข้อมูลและวัดคุณภาพของกาแฟโดยเฉพาะ

คอลัมน์ที่ 1 altitude_mean_meters ต่อ Total.Cup.Point

-altitude_mean_meters - ค่าเฉลี่ยของความสูงที่ใช้ในการเพาะปลูก

-Total.Cup.Points - คะแนนโดยรวมของกาแฟ

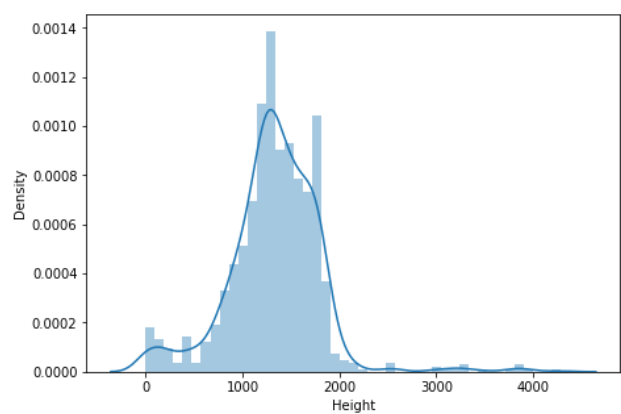
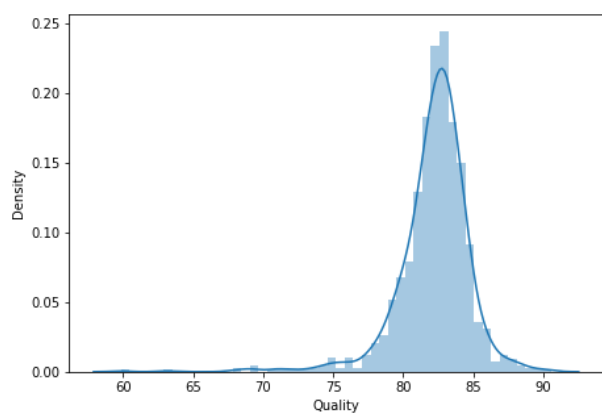
ค่าทางสถิติพื้นฐานของความสูงที่ใช้เพาะปลูก (trimmed เรียบร้อยแล้ว)

หน่วย m (เมตร)

| | |
|--------------------|---------|
| Mean | 1300.9 |
| Max | 2560 |
| Min | 1 |
| Standard Deviation | 419.29 |
| Median | 1310.64 |

จำนวน Rows = 1,075 rows in total

Distribution Graph และ Histogram



จะเห็นได้จากกราฟ ยังมี Outliers ที่อยู่มากจำเป็นที่ต้องลบทิ้งออกไป

การกำจัด Outlier

คำนวณขอบเขตความสูงที่อนุญาตได้จากค่าสิ่ง จากสมการที่บอกถึง extreme outliers

```
print("Highest height allowed",altitude_series.mean() + 3*altitude_series.std())
print("Lowest height allowed",altitude_series.mean() - 3*altitude_series.std())
```

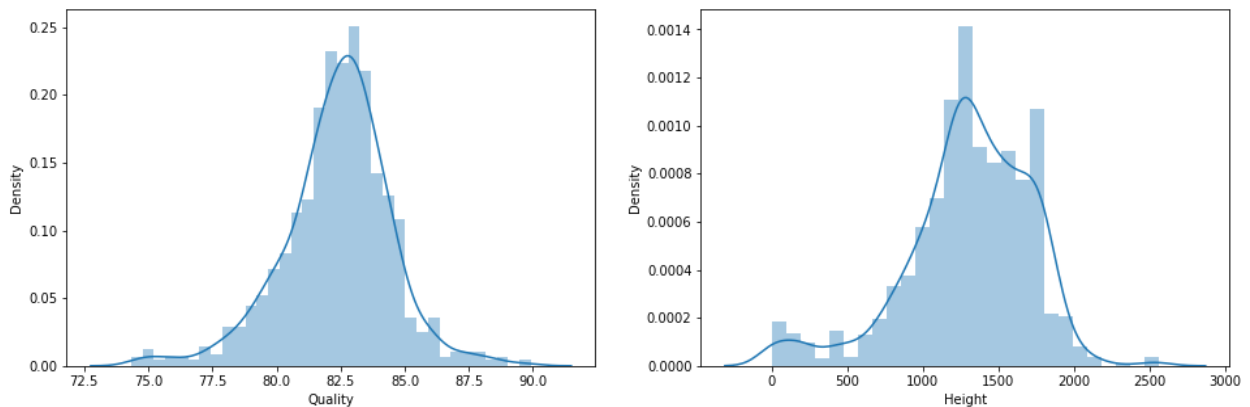
จะได้ผลลัพธ์ออกมาเป็นดังนี้

```
Highest height allowed 2787.474144089573
Lowest height allowed -131.447897434962
```

เมื่อได้ขอบเขตความสูงของข้อมูลแล้ว ถ้าหากเกินกว่า 2787.474 เมตร จะถือว่าเป็น Outliers
จากนั้นให้คำนวณหา Outliers ของ Total.Cup.Point จะได้ออกมาดังนี้

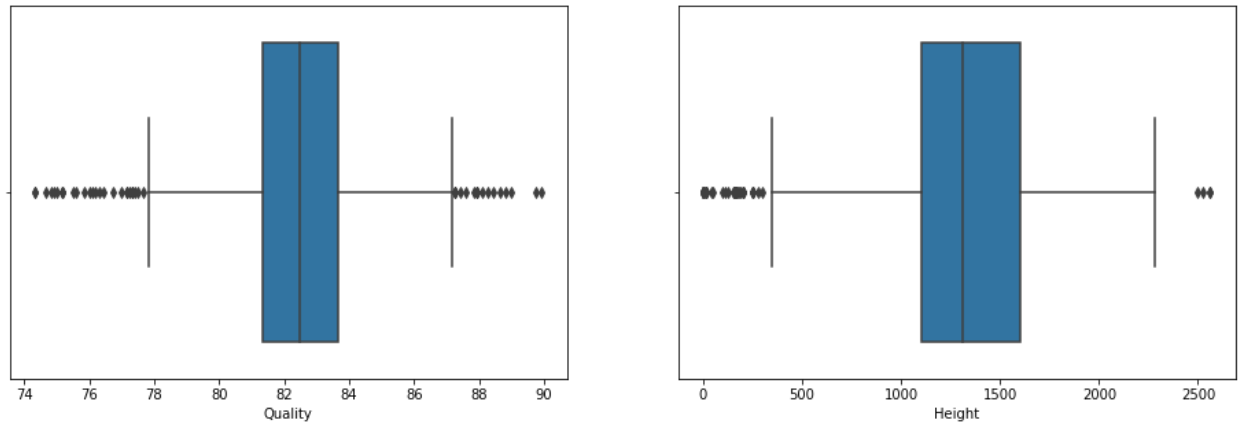
```
Highest Quality allowed 90.06682065614615
Lowest Quality allowed 74.30034028417936
```

เมื่อได้ค่าขอบเขตของทั้งสองค่าแล้ว จึงนำมาจัดทำ DataFrame ใหม่ที่ตัดเอาส่วนที่เป็น Outlier ออกไปเรียบร้อยแล้ว เพื่อทำ Distribution และ Histogram graph ใหม่อีกรอบ



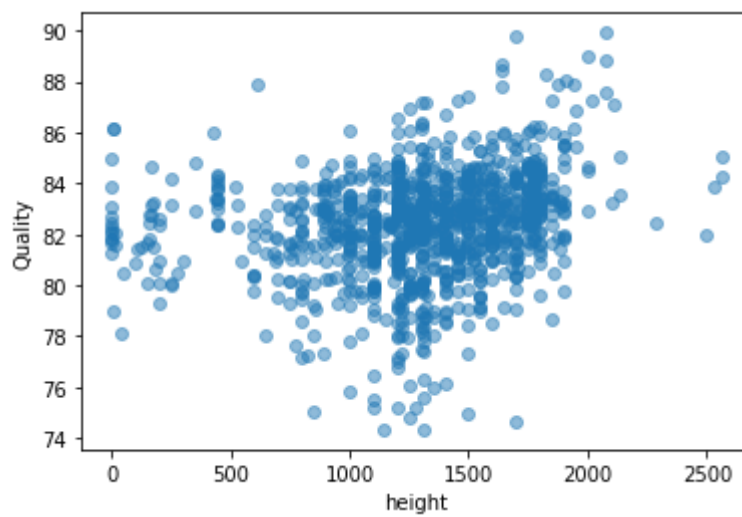
จากข้อมูลทั้งหมด 1,106 rows จะพบว่ามี Outlier ทั้งหมด 31 rows ซึ่งคิดเป็น 2.80% ของทั้งหมด เหลือทั้งหมด 1,075 rows

Boxplot หลัง trimmed ข้อมูล



Scatter plot

ดูความสัมพันธ์ระหว่าง Total.Cup.Point กับ altitude_mean_meters



-เป็นการดูแนวโน้มของความสูงที่ใช้ในการปลูกต่อคุณภาพของกาแฟ

[illegible][illegible]

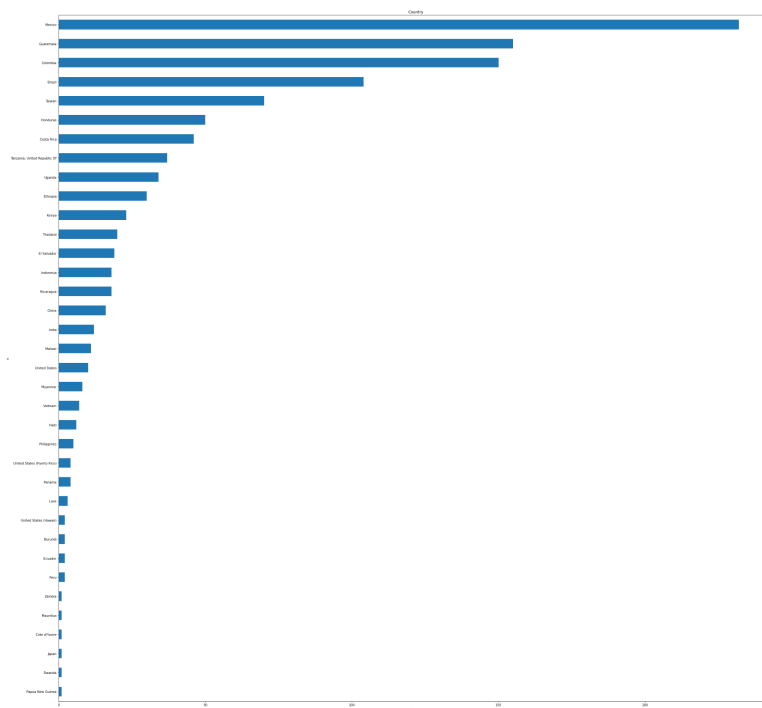
คอลัมน์ที่ 2 Country.of.Origin ต่อ Total.Cup.Point

-ประเทศที่ผลิตมีผลต่อ Total.Cup.Points - คะแนนโดยรวมของกาแฟ อย่างไร

ค่าสถิติพื้นฐานของ คะแนนรวมของกาแฟ

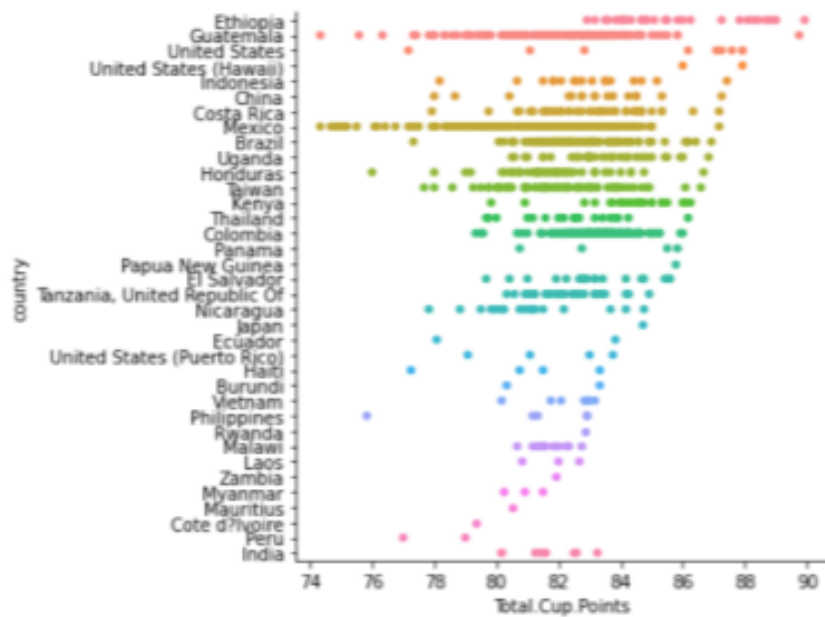
| | |
|--------------------|-------|
| Min | 74.33 |
| Max | 89.92 |
| Standard Deviation | 2.12 |
| Mean | 82.37 |
| Median | 82.50 |

จำนวนการปลูกเมล็ดกาแฟในแต่ละประเทศ



จากที่สังเกตประเทศ Mexico มีการปลูกเมล็ดกาแฟมากที่สุด ตามด้วย Guatemala, Colombia และ Brazil

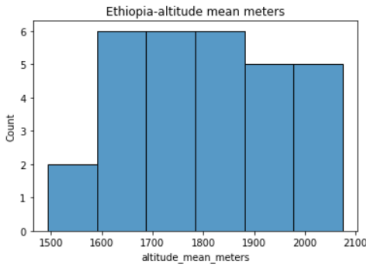
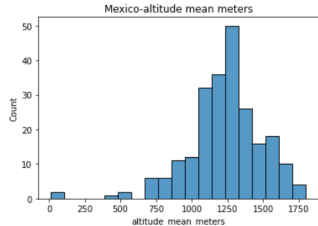
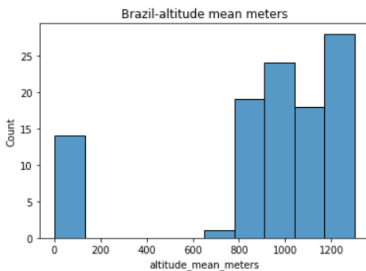
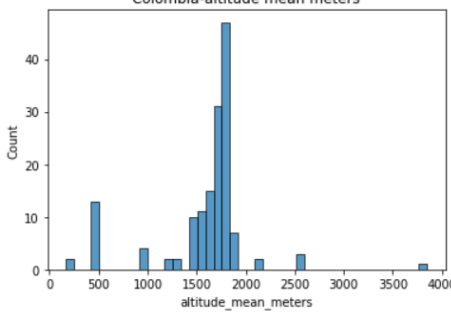
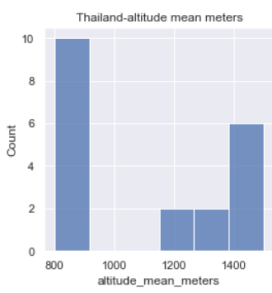
Plot Graph เทียบคะแนนคุณภาพของกาแฟกับประเทศที่ปลูกกาแฟ



จะเห็นว่าเมล็ดกาแฟจาก Ethiopia มีคะแนนคุณภาพที่สูงและเกาะกลุ่มกันพอสมควร
Guatemala และ Mexico มีคะแนนคุณภาพที่กระจัดกระจาย ตั้งแต่คะแนนสูงไปถึงต่ำมาก
และ Brazil และ Colombia มีคะแนนคุณภาพที่เกาะกลุ่มกันซึ่งคะแนนอยู่ช่วง 80 - 86 คะแนน
ซึ่งสายพันธุ์ของเมล็ดกาแฟจากประเทศเหล่านี้เป็น Arabica ทั้งหมด ต่อไปเราจะหาค่าเฉลี่ยของคะแนนคุณภาพกาแฟของประเทศที่สนใจก่อน
ดังนี้

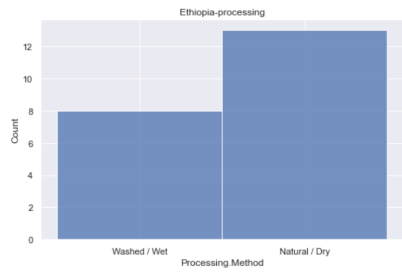
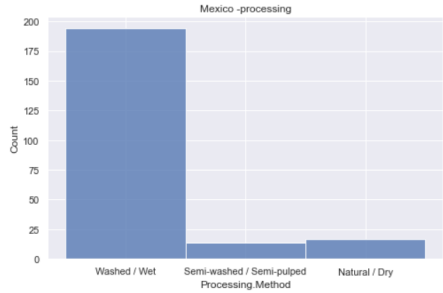
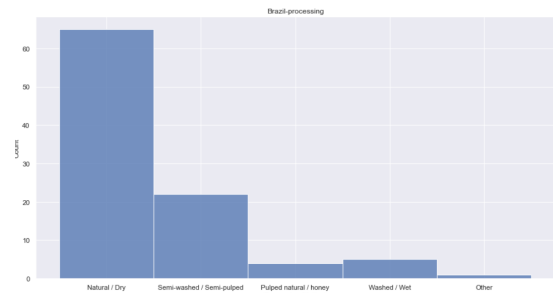
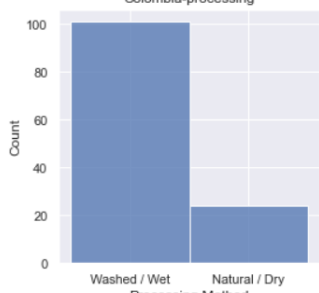
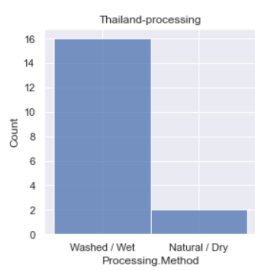
| ประเทศที่สนใจ | คะแนนเฉลี่ย |
|---------------|-------------|
| Ethiopia | 85.9 คะแนน |
| Mexico | 80.8 คะแนน |
| Brazil | 82.7 คะแนน |
| Colombia | 83.2 คะแนน |
| Thailand | 82.4 คะแนน |

ซึ่งจะนำข้อมูลส่วนอื่นมาวิเคราะห์เสริมได้ เราจะดูกันว่าในแต่ละประเทศมีการปลูกกาแฟที่ความสูงเท่าไร โดยเรียงจากคะแนน
คุณภาพและการเกาะกลุ่มกันเป็นหลัก โดยหยิบประเทศที่น่าสนใจมาวิเคราะห์เพิ่มเติม ตามตารางด้านล่าง

| ประเทศ | histogram | คำอธิบาย |
|----------|---|--|
| Ethiopia |  | ประเทศเอธิโอเปียปลูกที่ความสูง 1,500-2,100 เมตร ซึ่งเป็นความสูงที่ปลูกเฉลี่ยกันพอสมควร |
| Mexico |  | ประเทศเม็กซิโกปลูกที่ความสูง 0-1,710 เมตร ซึ่งส่วนมากปลูกกันที่ความสูงราว ๆ 1,500 เมตร |
| Brazil |  | ประเทศบราซิลปลูกที่ความสูง 0-1,200 เมตร ซึ่งเป็นความสูงที่ไม่เยอะมากเมื่อเทียบกับประเทศอื่น |
| Colombia |  | ประเทศโคลอมเบียปลูกที่ความสูง 0-2,500 เมตร และมีความสูงที่โดดไปถึง 3,850 เมตร ความสูงที่ใช้ในการปลูกเกาะกลุ่มกันอยู่ที่ 1,800 เมตร |
| Thailand |  | ประเทศไทยปลูกที่ความสูงราว ๆ 0-1,400 เมตร ส่วนมากจะปลูกกันที่ระนาบกับภาคพื้นดินส่วนใหญ่ |

จาก category plot และ histogram plot ที่แสดงด้านบน จะเห็นได้ว่า ความสูงที่ใช้ในการปลูกกาแฟมีผลต่อคะแนนคุณภาพอย่างเห็นได้ชัด ซึ่งปลูกกันอยู่ที่ความสูงระหว่าง 1,000 -2,000 เป็นส่วนใหญ่ แต่บราซิลซึ่งปลูกกันที่ความสูงน้อยกว่าประเทศอื่นอาจเป็นเพราะปัจจัยเสริมอื่น ๆ เราจะไปดูกระบวนการผลิตของประเทศต่าง ๆ กันต่อไป

ในส่วนนี้เราจะดูที่กระบวนการผลิตซึ่งมีการบ่มเพาะกาแฟ เพื่อนำข้อมูลไว้วิเคราะห์ต่อ

| ประเทศ | histogram | คำอธิบาย | | | | | | | | | | | | |
|---------------------------|--|-------------------|-------|---------------|-----|---------------------------|----|--|----|--|----|-------|---|---|
| Ethiopia |  <table border="1"><caption>Ethiopia-processing</caption><thead><tr><th>Processing Method</th><th>Count</th></tr></thead><tbody><tr><td>Washed / Wet</td><td>8</td></tr><tr><td>Natural / Dry</td><td>12</td></tr></tbody></table> | Processing Method | Count | Washed / Wet | 8 | Natural / Dry | 12 | ประเทศเอธิโอเปียมีการบ่มแบบ Washed/Wet และ Natural/Dry เป็นจำนวนเกือบเท่า ๆ กัน | | | | | | |
| Processing Method | Count | | | | | | | | | | | | | |
| Washed / Wet | 8 | | | | | | | | | | | | | |
| Natural / Dry | 12 | | | | | | | | | | | | | |
| Mexico |  <table border="1"><caption>Mexico-processing</caption><thead><tr><th>Processing Method</th><th>Count</th></tr></thead><tbody><tr><td>Washed / Wet</td><td>190</td></tr><tr><td>Semi-washed / Semi-pulped</td><td>10</td></tr><tr><td>Natural / Dry</td><td>20</td></tr></tbody></table> | Processing Method | Count | Washed / Wet | 190 | Semi-washed / Semi-pulped | 10 | Natural / Dry | 20 | ประเทศเม็กซิโกมีการบ่มแบบ Washed/Wet เป็นส่วนใหญ่ ส่วน semi wash / semi dry และ Natural/Dry มีแต่ไม่เยอะมาก | | | | |
| Processing Method | Count | | | | | | | | | | | | | |
| Washed / Wet | 190 | | | | | | | | | | | | | |
| Semi-washed / Semi-pulped | 10 | | | | | | | | | | | | | |
| Natural / Dry | 20 | | | | | | | | | | | | | |
| Brazil |  <table border="1"><caption>Brazil-processing</caption><thead><tr><th>Processing Method</th><th>Count</th></tr></thead><tbody><tr><td>Natural / Dry</td><td>65</td></tr><tr><td>Semi-washed / Semi-pulped</td><td>30</td></tr><tr><td>Pulped natural / honey</td><td>5</td></tr><tr><td>Washed / Wet</td><td>10</td></tr><tr><td>Other</td><td>1</td></tr></tbody></table> | Processing Method | Count | Natural / Dry | 65 | Semi-washed / Semi-pulped | 30 | Pulped natural / honey | 5 | Washed / Wet | 10 | Other | 1 | ประเทศบราซิลมีการบ่มแบบ Natural/Dry เป็นส่วนใหญ่ ส่วน semi wash / semi dry ใช้รองลงมา และมีการใช้วิธีหมักด้วย (Pulped natural / honey) |
| Processing Method | Count | | | | | | | | | | | | | |
| Natural / Dry | 65 | | | | | | | | | | | | | |
| Semi-washed / Semi-pulped | 30 | | | | | | | | | | | | | |
| Pulped natural / honey | 5 | | | | | | | | | | | | | |
| Washed / Wet | 10 | | | | | | | | | | | | | |
| Other | 1 | | | | | | | | | | | | | |
| Colombia |  <table border="1"><caption>Colombia-processing</caption><thead><tr><th>Processing Method</th><th>Count</th></tr></thead><tbody><tr><td>Washed / Wet</td><td>100</td></tr><tr><td>Natural / Dry</td><td>25</td></tr></tbody></table> | Processing Method | Count | Washed / Wet | 100 | Natural / Dry | 25 | ประเทศโคลอมเบียมีการบ่มแบบ Washed/Wet เป็นส่วนมากและมีการบ่มแบบ Natural/Dry บางส่วน | | | | | | |
| Processing Method | Count | | | | | | | | | | | | | |
| Washed / Wet | 100 | | | | | | | | | | | | | |
| Natural / Dry | 25 | | | | | | | | | | | | | |
| Thailand |  <table border="1"><caption>Thailand-processing</caption><thead><tr><th>Processing Method</th><th>Count</th></tr></thead><tbody><tr><td>Washed / Wet</td><td>15</td></tr><tr><td>Natural / Dry</td><td>2</td></tr></tbody></table> | Processing Method | Count | Washed / Wet | 15 | Natural / Dry | 2 | ประเทศไทยมีการบ่มแบบ Washed/Wet เป็นส่วนมากและมีการบ่มแบบ Natural/Dry เป็นส่วนน้อย | | | | | | |
| Processing Method | Count | | | | | | | | | | | | | |
| Washed / Wet | 15 | | | | | | | | | | | | | |
| Natural / Dry | 2 | | | | | | | | | | | | | |

จะเห็นว่ากระบวนการบ่มเพาะของแต่ละประเทศต่างก็มีวิธีการเป็นของตัวเอง

วิเคราะห์เบื้องต้น

-Total.Cup.Point (คะแนนคุณภาพของกาแฟโดยรวม)

จากข้อมูลเบื้องต้นจะเห็นว่าคะแนนคุณภาพของกาแฟอยู่ในช่วง 74-89 คะแนน ซึ่งมีค่าเฉลี่ยอยู่ที่ 82.37 คะแนน

ซึ่งจะกำหนดให้ 85-89 ถือเป็นกาแฟคุณภาพเยี่ยม

80-85 ถือเป็นกาแฟที่โอเคในระดับหนึ่ง

70-79 ถือเป็นกาแฟปกติ ไม่มีความพิเศษ

จากค่าเฉลี่ยอาจสรุปได้ว่ากาแฟในกลุ่มตัวอย่างเป็นกาแฟที่มีคุณภาพในระดับหนึ่ง

-altitude_mean_meters (ความสูงที่ใช้ในการเพาะปลูก)

จากกราฟ Scatter plot จะเห็นว่าช่วงของค่าของ altitude_mean_meters ที่มีผลทำให้ เมล็ดกาแฟ

- ที่ความสูง 1,500 - 2,000 เมตร มีแนวโน้มของคะแนนคุณภาพกาแฟที่สูงและกระจุกกันมากขึ้นอยู่ที่ 84 - 90 คะแนน
 - ที่ความสูง 1,000 - 1,500 เมตร มีแนวโน้มของคะแนนคุณภาพกาแฟที่ปานกลางและกระจายกันพอสมควรอยู่ที่ 86 - 73 คะแนน
 - ที่ความสูง 0 - 1,000 เมตร มีแนวโน้มของคะแนนคุณภาพกาแฟที่ปานกลางและกระจายกันมากขึ้นอยู่ที่ 86-78 คะแนน
 - ที่ความสูง 2,000 เมตรขึ้นไป มีข้อมูลที่น้อยจนเกินไป
- ข้อมูลของคะแนนอยู่ระดับ 82-85 คะแนน

ทำให้คาดการณ์ได้ว่า เมื่อปลูกที่ความสูง 1,500-2,000 เมตร กาแฟจะมีคุณภาพที่โอเคกว่าความสูงอื่น ๆ ที่ต่ำกว่าลงมา สังเกตได้จากความกระจายของข้อมูลซึ่งกระจุกกันมากที่ความสูง 1,500-2,000 เมตร และมีคะแนนที่สูง ส่วนความสูงที่มากและน้อยจนเกินไปอาจคาดการณ์ได้ว่าเป็นความสูงที่ไม่นิยมปลูกกาแฟ เพราะให้กาแฟที่มีคุณภาพไม่สูงมากนัก

ตรงส่วนนี้จะนำไปวิเคราะห์เพิ่มในส่วนของ Linear Regression เพื่อค้นหาแนวโน้มของความสูงในการปลูกเฉลี่ยที่แม่นยำมากยิ่งขึ้น

-Country.of.Origin (ประเทศที่ผลิตกาแฟ)

จากกราฟของข้อมูลตัวอย่างจำนวนของประเทศที่ปลูกเมล็ดกาแฟ เห็นได้ว่าประเทศที่ปลูกมากที่สุดก็คือประเทศ Mexico มีการปลูกเมล็ดกาแฟมากที่สุด ตามด้วย guatemala, colombia และ brazil

ซึ่งเราจะสำรวจประเทศที่น่าสนใจ 5 ประเทศ คือ

- 1.ประเทศเม็กซิโก เพราะในชุดข้อมูลมีการปลูกกาแฟเยอะที่สุด
- 2.ประเทศเอธิโอเปีย เพราะในชุดข้อมูลมีคะแนนคุณภาพเฉลี่ยของเมล็ดกาแฟที่เยอะที่สุด
- 3.ประเทศบราซิล เพราะเป็นประเทศที่มีการส่งออกเมล็ดกาแฟเยอะที่สุดในโลก
- 4.ประเทศโคลอมเบีย เพราะเป็นประเทศที่มีการส่งออกเมล็ดกาแฟเยอะเป็นอันดับสาม
- 5.ประเทศไทย

ส่วนประเทศที่น่าสนใจไม่แพ้กันคือประเทศเวียดนาม ซึ่งเป็นประเทศที่ส่งออกกาแฟเยอะที่สุดเป็นอันดับสองของโลก และเป็นตลาดนำเข้าที่สำคัญมากของประเทศไทยโดยประเทศไทยนำเข้าเป็นจำนวนทั้งหมด 81.05% ของเมล็ดกาแฟที่นำเข้าทั้งหมด แต่เนื่องจากชุดข้อมูลของเมล็ดกาแฟจากเวียดนามมีน้อยจนเกินไป (8 ข้อมูล)

จากกราฟ ความสัมพันธ์ระหว่างคะแนนและประเทศที่ใช้ปลูก สังเกตได้ว่ากาแฟจากประเทศเอธิโอเปียมีการกระจุกกันของคะแนนที่สูงมาก ซึ่งในจุดนี้เราจะดูข้อมูลอื่นประกอบไปด้วย โดยวิเคราะห์ข้อมูลจากประเทศที่สนใจ

-ความสูงที่ใช้ปลูกภายในประเทศนั้น ๆ และกระบวนการในการบ่มเพาะของประเทศ

จากตาราง จะเห็นว่าประเทศเอธิโอเปียที่มีคะแนนคุณภาพของกาแฟสูงนั้นปลูกที่ความสูงในช่วง 1,500-2,100 เมตร และกระบวนการบ่มเพาะที่ไม่มีวิธีไหนโดดเด่น แต่เนื่องจากข้อมูลหลายแห่งบอกว่า ประเทศเอธิโอเปียเป็นต้นกำเนิดของการปลูกกาแฟ และเป็นต้นกำเนิดของสายพันธุ์อาราบิก้า และได้มีสายพันธุ์ต่าง ๆ เกิดขึ้นมากมายจนนำไปสู่การปลูกที่ประเทศอื่น และมีการแปรรูปกาแฟที่ส่งสมมานานมากนี้อาจเป็นเหตุผลที่เมล็ดกาแฟจากประเทศเอธิโอเปียมีคุณภาพที่ดีที่สุด ข้อมูลจาก [ลัดเลาะแหล่งกำเนิดผลไม้ปศิจา ‘เอธิโอเปีย’ เมืองสวรรค์ของคอกาแฟ \(adaymagazine.com\)](#)

ส่วนกาแฟจากประเทศเม็กซิโกมีการกระจายตัวที่เยอะและมีคะแนนคุณภาพที่น้อยกว่าประเทศอื่น ๆ แต่ยังมีการผลิตที่เยอะกว่าประเทศอื่น ๆ และใช้วิธีการบ่มเพาะแบบเปียก (Wash process)

กาแฟจากบราซิลใช้วิธีบ่มเพาะแบบแห้ง (Dry process) ความสูงที่ใช้ในการปลูกไม่เยอะมาก ตั้งแต่ปลูกที่ระดับพื้นดิน ไปถึงช่วง 800 - 1,200 เมตร อาจไม่สูงมากเทียบกับประเทศอื่น ๆ แต่เนื่องจากเป็นประเทศที่ส่งออกมากที่สุด คะแนนของคุณภาพกาแฟถือว่าระดับหนึ่ง ทั้งนี้เป็นเพราะกระบวนการในการบ่มเพาะแบบแห้งที่ไม่เหมือนประเทศอื่น

กาแฟจากโคลอมเบียปลูกที่ความสูงราว ๆ 1,400 - 1,800 เมตร ทั้งนี้อาจเป็นเพราะความสูงที่ใช้ในการปลูกเลยมีผลต่อคะแนนคุณภาพ และใช้วิธีบ่มเพาะแบบเปียก ซึ่งเหมือนกับประเทศไทย แต่ประเทศไทยปลูกที่ความสูงราว ๆ พื้นดินส่วนใหญ่ปลูกไม่สูงมาก

สรุปเบื้องต้น

จากข้อมูลเบื้องต้นบอกได้ว่าความสูงมีผลต่อการปลูกจริง ๆ ซึ่งจากการหาข้อมูลต่าง ๆ พบว่ากระบวนการบ่มเพาะกาแฟมีผลต่อกาแฟเหมือนกัน ยกตัวอย่างเช่น

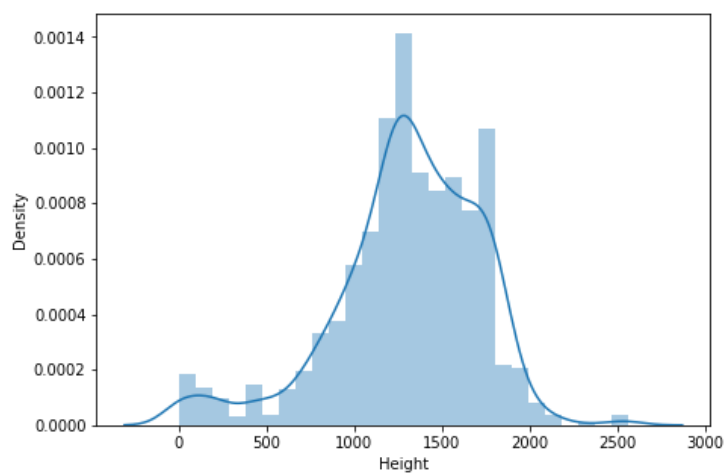
Washed process เป็นการบ่มกาแฟแบบเปียก ซึ่งเหมาะกับกาแฟเกรดดี

ส่วน Natural process เป็นการบ่มกาแฟแบบแห้ง เป็นวิธีธรรมชาติในการบ่มเพาะกาแฟ ซึ่งสามารถทำให้กาแฟมีคุณภาพดีที่สุด จะเห็นได้ว่าประเทศบราซิลและเอธิโอเปียที่ใช้วิธีนี้จะเป็นตัวท็อปในด้านของกาแฟทั้งนั้น

ส่วนประเทศที่ใช้วิธีบ่มเพาะแบบเปียกส่วนมากจะปลูกกันที่ความสูงที่สูงพอสมควรหรือไม่ก็ใช้เมล็ดกาแฟคุณภาพดีปลูกที่ระดับพื้นดิน โดยรวมยังให้คะแนนคุณภาพกาแฟเฉลี่ยที่สูงอยู่ แต่ก็มีกระจายของคะแนนที่เยอะพอสมควร

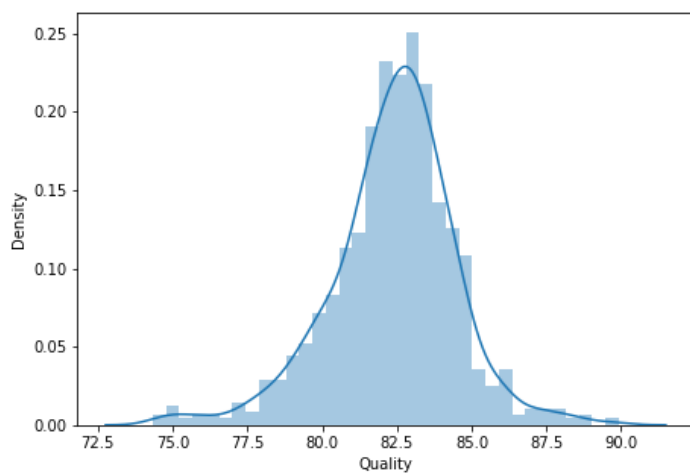
Probability Density Function

-ความสูงเฉลี่ยที่ใช้ในการปลูก



แกน x เป็นความสูงในการปลูก และแกน y เป็นความหนาแน่นของความสูงที่ใช้ในการปลูกเมล็ดกาแฟ

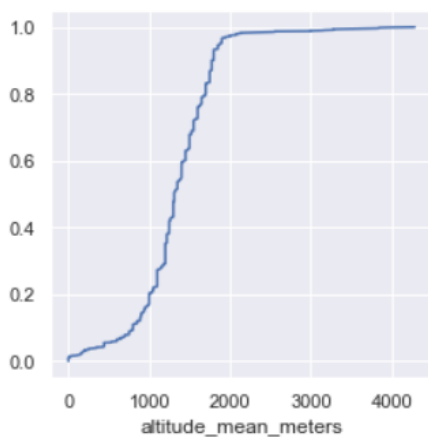
-คะแนนคุณภาพของกาแฟ



แกน x เป็นคะแนนคุณภาพของกาแฟ และแกน y เป็นความหนาแน่นของความสูงที่ใช้ในการปลูกเมล็ดกาแฟ

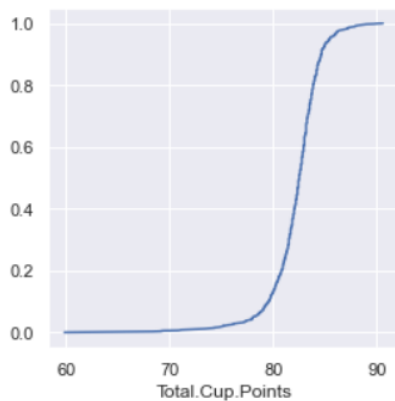
Cumulative Probability Function

-ความสูงเฉลี่ยที่ใช้ในการปลูก



แกน x เป็นความสูงในการปลูก และแกน y เป็นความหนาแน่นของความสูงที่ใช้ในการปลูกเมล็ดกาแฟ

-คะแนนคุณภาพของกาแฟ



แกน x เป็นคะแนนคุณภาพของกาแฟ และแกน y เป็นความหนาแน่นของความสูงที่ใช้ในการปลูกเมล็ดกาแฟ

บทวิเคราะห์ข้อมูลจากกราฟ

Probability Density Function

- จากกราฟความสูงที่ใช้ในการเพาะปลูกเมล็ดกาแฟ จะเห็นได้ว่าในช่วงความสูงประมาณ 1,000-1,800 เมตรจะมีค่าความหนาแน่นมากที่สุด และในช่วงความสูงต่ำกว่า 1,000 เมตรและมากกว่า 1,800 เมตรเป็นต้นไปก็ จะค่อย ๆ ลดลงเรื่อย ๆ วิเคราะห์ได้ว่า ประเทศส่วนมากนิยมปลูกที่ความสูง 1,000-1,800 เมตร

- จากกราฟคะแนนคุณภาพของกาแฟ จะเห็นได้ว่าในคะแนนจะหนาแน่นมากในช่วง 82.5 คะแนน แล้วลดลงเรื่อยๆตั้งแต่ 80 และ 85 คะแนน ซึ่งคะแนนคุณภาพกาแฟเฉลี่ยจะอยู่ที่ราว ๆ 82.5 คะแนน

Cumulative Probability Function

- จากกราฟความสูงที่ใช้ในการเพาะปลูกเมล็ดกาแฟ จะเห็นได้ว่าในช่วงความสูงประมาณ 0-1,000 เมตร ค่าความน่าจะเป็นสะสมจะเพิ่มขึ้นอย่างช้า ๆ ส่วนที่ความสูงระหว่าง 1,000 - 2,000 เมตร ค่าความน่าจะเป็นสะสมจะเพิ่มขึ้นอย่างรวดเร็ว และตั้งแต่ 2,000 เมตรเป็นต้นไปค่าความน่าจะเป็นสะสมจะเพิ่มขึ้นน้อยมาก วิเคราะห์ได้ว่า รายได้ของความสูงที่ใช้ในการปลูกส่วนใหญ่อยู่ระหว่าง 1,000 - 2,000 เมตร

- จากกราฟคะแนนคุณภาพของกาแฟ จะเห็นได้ว่าตั้งแต่คะแนน 60 - 75 คะแนน ค่าความน่าจะเป็นสะสมจะเพิ่มขึ้นน้อยมาก ส่วนที่คะแนนประมาณ 75 - 83 คะแนน

ค่าความน่าจะเป็นสะสมจะเพิ่มขึ้นอย่างรวดเร็วแล้วเพิ่มขึ้นน้อยลงถึงน้อยมากในช่วง 85 คะแนนขึ้นไป วิเคราะห์ได้ว่าคะแนนมีการกระจุกตัวมาก อยู่ระหว่าง 75 - 83 คะแนน

Confidence Interval (CI) of Mean

หา Confidence Interval (CI) ของคอลัมน์ Total.Cup.Points ซึ่งเป็นคะแนนรวมของคุณภาพเมล็ดกาแฟที่ถูกเฉลี่ยมาแล้ว

ช่วง Confidence levels 90%

```
LCL90 = myMean + ME[0]
UCL90 = myMean - ME[0]
print(LCL90,UCL90)
```

```
82.26520850227269 82.47767521865757
```

รูปที่ 1

ช่วงของ Confidence Interval จากช่วง Confidence levels 90% มีค่าอยู่ในช่วงของ 82.265 ถึง 82.478

ช่วง Confidence levels 95%

```
LCL95 = myMean + ME[1]
UCL95 = myMean - ME[1]

print(LCL95,UCL95)
```

```
82.24482350156332 82.49806021936693
```

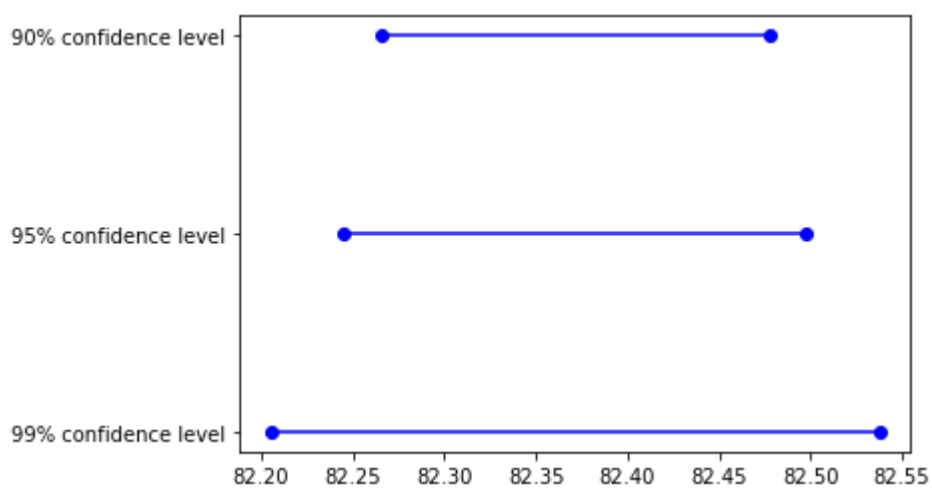
รูปที่ 2

ช่วงของ Confidence Interval จากช่วง Confidence levels 95% มีค่าอยู่ในช่วง ของ 82.245 ถึง 82.498

ช่วง Confidence levels 99%

```
LCL99 = myMean + ME[2]  
UCL99 = myMean - ME[2]  
  
print(LCL99,UCL99)  
  
82.2049287611408 82.53795495978945
```

ช่วงของ Confidence Interval จากช่วง Confidence levels 99% มีค่าอยู่ในช่วง ของ 82.205 ถึง 82.538



รูปที่ 4

จากกราฟนี้จะเห็นได้ว่ายิ่งค่าของ confidence levels มีค่าเปอร์เซ็นต์มากขึ้นช่วงก็จะยาวมากขึ้น

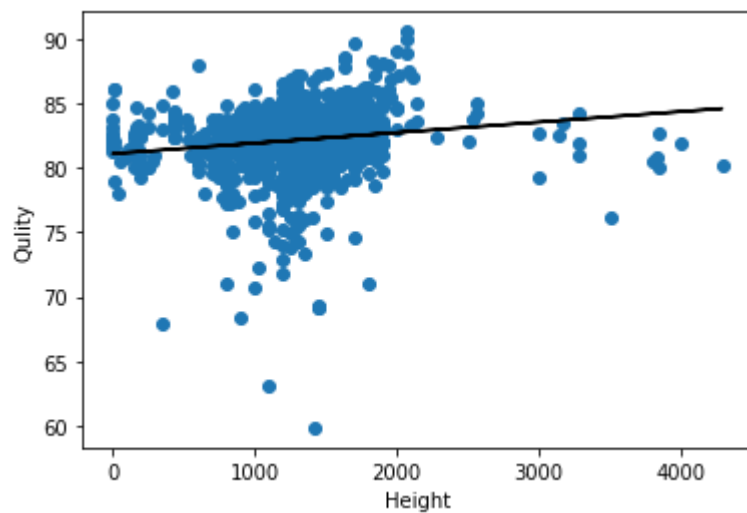
บทวิเคราะห์ของ Confidence Interval

จากค่า Confidence Interval ของ Total.Cup.Points ซึ่งเป็นค่าเฉลี่ยของคุณภาพเมล็ดกาแฟ จากรูปที่ 1 ค่า Confidence Interval จะอยู่ที่ประมาณ 82.265 ถึง 82.478 รูปที่ 2 อยู่ในช่วง ของ 82.245 ถึง 82.498 รูปที่ 3 มีค่าอยู่ในช่วง ของ 82.205 ถึง 82.538

จะสังเกตเห็นได้จากรูปที่ 4 ว่า ยิ่งค่าเปอร์เซ็นต์ Confidence levels ช่วงของค่า Confidence Interval จะยิ่งมากขึ้นทำให้ค่าของ Confidence Interval มีโอกาสครอบคลุมค่าเฉลี่ยของคุณภาพเมล็ดกาแฟที่มากขึ้นด้วย โดยจาก confidence levels x % หมายถึง มีโอกาสตาม x % ที่จะมีความผิดพลาดเกิดขึ้น

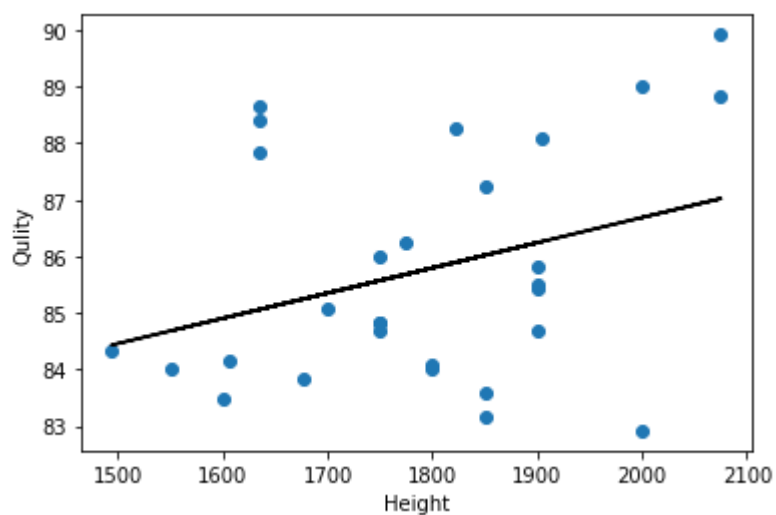
Linear Regression

เปรียบเทียบกราฟถดถอยเชิงเส้นกับกราฟ XY(Scatter) Plot ของข้อมูล ความสูงของพื้นที่ปลูก และ คุณภาพของกาแฟ



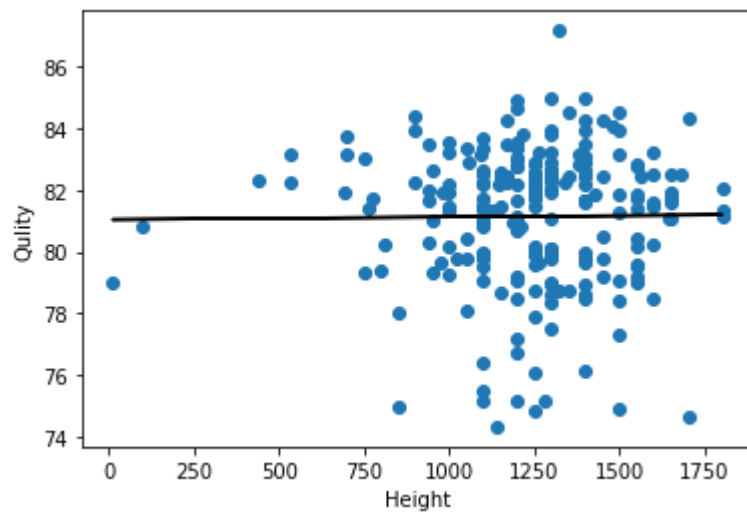
รูปที่ 1 ความสูงของข้อมูลทั้งหมด

เป็นกราฟรวมของทุกประเทศที่มีในชุดข้อมูล โดยที่แกน x เป็นระดับความสูงของสถานที่ปลูกกาแฟ และ แกน y เป็น คะแนนคุณภาพของเมล็ดกาแฟ



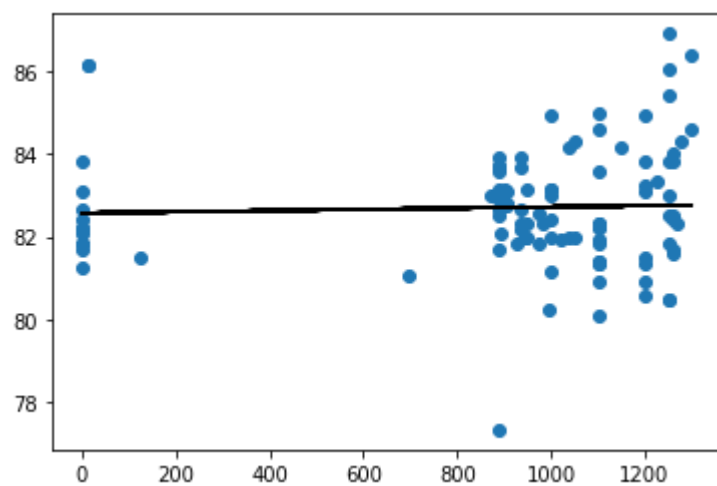
รูปที่ 2 เอธิโอเปีย

เป็นกราฟเฉพาะของประเทศเอธิโอเปีย โดยที่แกน x เป็นระดับความสูงของสถานที่ปลูกกาแฟ และ แกน y เป็น คะแนนคุณภาพของเมล็ดกาแฟ



รูปที่ 3 เม็กซิโก

เป็นกราฟเฉพาะของประเทศเอธิโอเปีย โดยที่แกน x เป็นระดับความสูงของสถานที่ปลูกกาแฟ และ แกน y เป็น คะแนนคุณภาพของเมล็ดกาแฟ



รูปที่ 4 บราซิล

เป็นกราฟเฉพาะของประเทศบราซิล โดยที่แกน x เป็นระดับความสูงของสถานที่ปลูกกาแฟ และ แกน y เป็น คะแนนคุณภาพของเมล็ดกาแฟ

บทวิเคราะห์กราฟของ Linear Regression

จากข้อมูลจากกราฟ Linear Regression - (คะแนนคุณภาพภาพและความสูงที่ใช้ปลูก)

โดยคิดจากหลาย ๆ ประเทศพบว่าข้อมูลมีค่า Correlation Coefficient ที่น้อยมาก ทำให้กราฟในส่วนนี้ไม่เหมาะกับการวิเคราะห์ด้วย Linear Regression มากนัก

จะวิเคราะห์ได้ว่าความสูงมีผลในช่วงนี้มากกว่าตามสรุปเบื้องต้นก่อนหน้านี้ที่กล่าวไป และมีปัจจัยอื่นก็มีผลเช่นกัน เช่น กระบวนการผลิต เป็นต้น

สรุปผลการศึกษาและเสนอแนะแนวทางการศึกษาเพิ่มเติม

จากการศึกษาข้อมูล บอกได้ว่าความสูงมีผลต่อการปลูกจริง ๆ แต่อยู่ในความสูงช่วงใดช่วงหนึ่ง ซึ่งจากการหาข้อมูลต่าง ๆ พบว่า กระบวนการบ่มเพาะกาแฟมีผลต่อกาแฟเหมือนกัน ยกตัวอย่างเช่น Washed process เป็นการบ่มกาแฟแบบเปียก ซึ่งเหมาะกับกาแฟเกรดดี ส่วน Natural process เป็นการบ่มกาแฟแบบแห้ง เป็นวิถีธรรมชาติในการบ่มเพาะกาแฟ ซึ่งสามารถทำให้กาแฟมีคุณภาพดีที่สุด จะเห็นได้ว่า ประเทศบราซิลและเอธิโอเปียที่ใช้วิธีนี้จะเป็นตัวท็อปในด้านของกาแฟทั้งนั้น

ในส่วนของประเทศไทยก็มีการนำเข้าเมล็ดกาแฟจากเวียดนาม ซึ่งติดอันดับส่งออกเป็นอันดับสองของโลก และเป็นตลาดนำเข้าที่สำคัญของประเทศไทย ข้อมูลบอกว่าเวียดนามใช้วิธีแบบแห้งซึ่งเป็นวิธีทำให้กาแฟได้คุณภาพที่ดีพอสมควร ซึ่งร้านกาแฟในไทยส่วนมากก็ใช้เมล็ดกาแฟจากเวียดนาม จะสังเกตได้ว่ากระบวนการผลิตแบบแห้ง กาแฟจะมีรสชาติที่ติดออกเปรี้ยวไม่มาก ต่างกับแบบเปียกที่จะรสชาติจะออกเปรี้ยวและกลิ่นกาแฟชัดกว่า

1. ควรจะศึกษาปัจจัยคะแนนด้านอื่น ๆ ของกาแฟเพิ่มเติม เช่น สายพันธุ์ที่หลากหลายขึ้น คะแนนรสชาติ คะแนนความเป็นกรด คะแนนกลิ่น คะแนนเนื้อสัมผัส เพื่อให้สามารถวิเคราะห์ข้อมูลได้ละเอียดยิ่งขึ้น ไม่ว่าจะเป็นกระบวนการผลิตมีผลต่อกลิ่น รสชาติ หรือ เนื้อสัมผัสอย่างไร สายพันธุ์มีผลต่อคะแนนต่าง ๆ หรือไม่ เป็นต้น
2. ควรศึกษาข้อมูลความเป็นมาของการผลิตกาแฟแต่ละประเทศ เพราะแต่ละประเทศที่ขึ้นชื่อเรื่องการส่งออกจะมีกระบวนการที่ต่างกับที่อื่น ส่งผลให้มีผลต่อคะแนนคุณภาพหรือทำให้เกิดการส่งออกที่มากขึ้นก็ได้

สามารถ Download Project ได้ตามลิงค์ต่อไปนี้

Source code: [arrin/coffee-analysis \(github.com\)](https://github.com/arrin/coffee-analysis) and [arrin/coffee-analysis at Kla \(github.com\)](https://github.com/arrin/coffee-analysis)

```
In [183...  
import pandas as pd  
import matplotlib.pyplot as plt  
import seaborn as sns  
import numpy as np  
import stemgraphic
```

```
In [184...  
df = pd.read_csv('workspace\coffee analysis\merged_data_cleaned.csv')  
# 1.altitude_mean_meters  
# 2.Species  
# 3.Processing.Method  
# 4.Region  
# 5.Total.Cup.Points  
df.dropna(subset = ["altitude_mean_meters"], inplace=True)  
df['Country.of.Origin'][df['Country.of.Origin'].isna()] = 'Unknown'
```

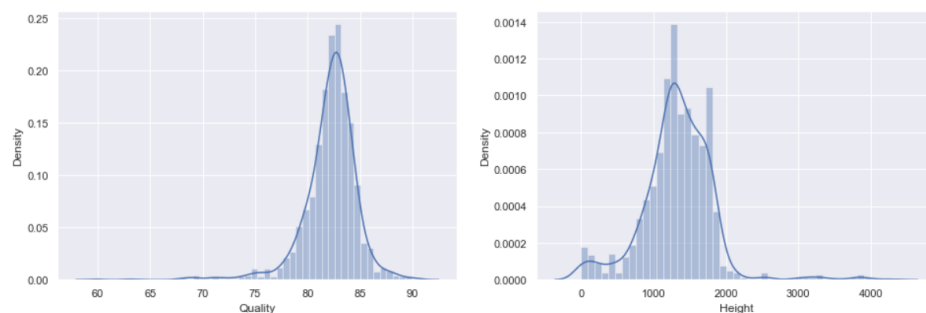
```
In [185...  
df.describe()
```

```
Out[185...  
      Unnamed: 0  altitude_mean_meters  Aroma  Flavor  Aftertaste  Acidity  Body  Balance  Uniformity  Clean.Cup  Sweetness  Cupper.Points  Tota  
count  1106.000000      1106.000000      1106.000000      1106.000000      1106.000000      1106.000000      1106.000000      1106.000000      1106.000000      1106.000000      1106.000000      1106.000000  
mean      669.899638      1328.013123      7.578002      7.527604      7.401311      7.535714      7.513382      7.512559      9.876555      9.860072      9.882450      7.495678  
std       381.077935      486.487007      0.308928      0.330967      0.339747      0.312954      0.289186      0.353711      0.429018      0.725699      0.523281      0.413252  
min         0.000000         1.000000      5.080000      6.170000      6.170000      5.250000      5.170000      5.250000      6.000000      0.000000      1.330000      5.170000  
25%      340.250000      1100.000000      7.420000      7.330000      7.250000      7.330000      7.330000      7.330000      10.000000      10.000000      10.000000      7.250000  
50%      674.500000      1310.640000      7.580000      7.580000      7.420000      7.500000      7.500000      7.500000      10.000000      10.000000      10.000000      7.500000  
75%      991.750000      1600.000000      7.750000      7.750000      7.580000      7.750000      7.670000      7.750000      10.000000      10.000000      10.000000      7.750000  
max     1336.000000      4287.000000      8.750000      8.830000      8.670000      8.750000      8.580000      8.750000      10.000000      10.000000      10.000000      10.000000
```

Mean ของความสูงที่ปลูกกาแฟ = 1775.030545 เมตร Mean ของคะแนนเฉลี่ย = 82.109594 คะแนน

```
In [186...  
altitude_series = df.loc[:, 'altitude_mean_meters']  
point = df.loc[:, 'Total.Cup.Points']  
plt.figure(figsize=(16,5))  
plt.subplot(1,2,1)  
sns.distplot(a=point)  
plt.xlabel("Quality")  
plt.subplot(1,2,2)  
sns.distplot(a=altitude_series)  
plt.xlabel("Height")
```

```
Out[186...  
Text(0.5, 0, 'Height')
```



จะเห็นได้ว่ากราฟ distribution ยังมี Outlier ที่เยอะมากอยู่ จึงต้องลบทิ้ง

```
In [187... print("Highest height allowed",altitude_series.mean() + 3*altitude_series.std())
print("Lowest height allowed",altitude_series.mean() - 3*altitude_series.std())

print("Highest Quality allowed",point.mean() + 3*point.std())
print("Lowest Quality allowed",point.mean() - 3*point.std())

Highest height allowed 2787.474144089573
Lowest height allowed -131.447897434962
Highest Quality allowed 90.06682065614615
Lowest Quality allowed 74.30034028417936

นี่คือขอบเขตของข้อมูลของความสูง ถ้ามากกว่า 2787.474144089573 จะถือว่าเป็น outliers
```

Find Outliers

```
In [188... df[(altitude_series> 2787.474144089573) | (altitude_series < -131.447897434962)]
```

Out[188...

| Unnamed: 0 | Species | Country.of.Origin | Altitude | altitude_mean_meters | Processing.Method | Aroma | Flavor | Aftertaste | Acidity | ... | Balance | Uniformity | Clean.Cup | Sweetness | |
|------------|---------|-------------------|---------------|----------------------|-------------------|---------------|--------|------------|---------|------|---------|------------|-----------|-----------|-------|
| 215 | 215 | Arabica | Guatemala | 3280 | 3280.0 | Washed / Wet | 7.58 | 7.83 | 7.58 | 7.83 | ... | 7.67 | 10.00 | 10.0 | 10.00 |
| 627 | 628 | Arabica | Colombia | 1800 meters (5900) | 3850.0 | NaN | 7.33 | 7.58 | 7.42 | 7.42 | ... | 7.67 | 10.00 | 10.0 | 10.00 |
| 836 | 837 | Arabica | Guatemala | 3280 | 3280.0 | Washed / Wet | 7.58 | 7.50 | 7.33 | 7.42 | ... | 7.25 | 10.00 | 10.0 | 10.00 |
| 839 | 840 | Arabica | Myanmar | 4001 | 4001.0 | Washed / Wet | 7.33 | 7.58 | 7.50 | 7.42 | ... | 7.33 | 10.00 | 10.0 | 10.00 |
| 999 | 1001 | Arabica | Guatemala | 3280 | 3280.0 | Washed / Wet | 7.42 | 7.25 | 7.17 | 7.50 | ... | 7.17 | 10.00 | 10.0 | 10.00 |
| 1036 | 1038 | Arabica | Myanmar | 3825 | 3825.0 | Washed / Wet | 6.92 | 7.50 | 7.00 | 7.58 | ... | 7.08 | 10.00 | 10.0 | 10.00 |
| 1071 | 1073 | Arabica | Myanmar | 3800 | 3800.0 | Washed / Wet | 7.17 | 7.33 | 7.17 | 7.42 | ... | 7.08 | 10.00 | 10.0 | 10.00 |
| 1096 | 1098 | Arabica | Myanmar | 4287 | 4287.0 | Natural / Dry | 7.42 | 7.00 | 7.08 | 7.00 | ... | 7.33 | 10.00 | 10.0 | 10.00 |
| 1121 | 1123 | Arabica | Myanmar | 3845 | 3845.0 | Washed / Wet | 7.17 | 7.33 | 7.00 | 7.42 | ... | 7.00 | 10.00 | 10.0 | 10.00 |
| 1267 | 1269 | Arabica | Indonesia | 3500 | 3500.0 | Natural / Dry | 7.33 | 7.00 | 6.50 | 6.08 | ... | 6.33 | 9.33 | 10.0 | 9.33 |
| 1309 | 1312 | Robusta | India | 3170 | 3170.0 | Washed / Wet | 8.00 | 7.75 | 7.92 | 8.00 | ... | 7.92 | 10.00 | 10.0 | 8.00 |
| 1313 | 1316 | Robusta | India | 3000' | 3000.0 | NaN | 8.00 | 7.92 | 7.67 | 8.00 | ... | 7.92 | 10.00 | 10.0 | 7.75 |
| 1315 | 1318 | Robusta | India | 3140 | 3140.0 | Washed / Wet | 7.67 | 7.75 | 7.83 | 7.67 | ... | 7.83 | 10.00 | 10.0 | 7.92 |
| 1330 | 1333 | Robusta | United States | 3000' | 3000.0 | Natural / Dry | 7.92 | 7.50 | 7.42 | 7.42 | ... | 7.42 | 9.33 | 10.0 | 7.58 |

14 rows × 21 columns

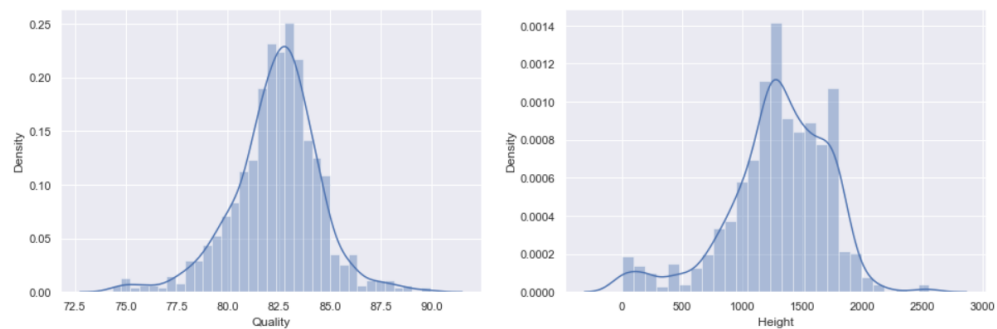


สร้างชุดข้อมูลที่กำจัด Outliers ออกไปแล้ว

```
In [189... new_df = df[(altitude_series < 2787.474144089573) & (altitude_series > -131.447897434962)
              & (point < 90.09512695261628) & (point > 74.29642982394051)]
new_df.describe()
```

```
Out[189... Unnamed: 0  altitude_mean_meters  Aroma  Flavor  Aftertaste  Acidity  Body  Balance  Uniformity  Clean.Cup  Sweetness  Cupper.Points  Tota
count  1075.000000      1075.000000  1075.000000  1075.000000  1075.000000  1075.000000  1075.000000  1075.000000  1075.000000  1075.000000  1075.000000
mean    656.469767      1300.941576    7.587209    7.539265    7.413712    7.543767    7.519814    7.526540    9.895330    9.921767    9.913870    7.509814
std     374.164429      419.288961    0.297584    0.314020    0.320668    0.300123    0.280413    0.334983    0.370103    0.416600    0.387479    0.397782
min      1.000000        1.000000    5.080000    6.330000    6.330000    5.250000    5.170000    5.250000    6.670000    5.330000    6.670000    5.170000
25%     332.500000      1100.000000    7.420000    7.330000    7.250000    7.330000    7.330000    7.330000    10.000000    10.000000    10.000000    7.330000
50%     659.000000      1310.640000    7.580000    7.580000    7.420000    7.500000    7.500000    7.500000    10.000000    10.000000    10.000000    7.500000
75%     970.500000      1600.000000    7.750000    7.750000    7.580000    7.750000    7.670000    7.750000    10.000000    10.000000    10.000000    7.750000
max    1336.000000      2560.000000    8.750000    8.670000    8.580000    8.580000    8.580000    8.750000    10.000000    10.000000    10.000000    8.580000
```

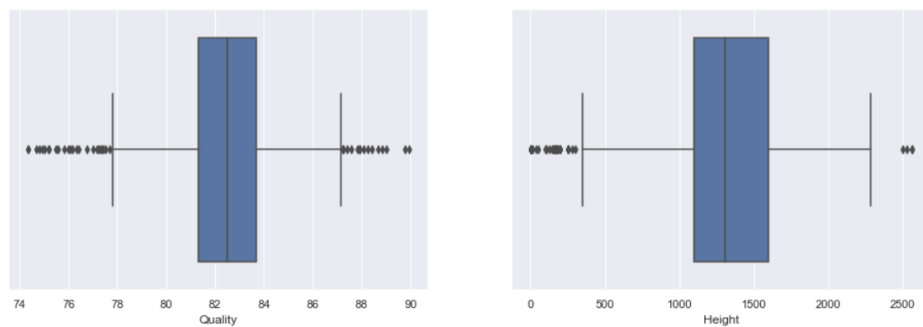
```
In [208... altitude_series = new_df.loc[:, 'altitude_mean_meters']
point = new_df.loc[:, 'Total.Cup.Points']
plt.figure(figsize=(16,5))
plt.subplot(1,2,1)
sns.distplot(a=point)
plt.xlabel("Quality")
plt.subplot(1,2,2)
sns.distplot(a=altitude_series)
plt.xlabel("Height")
```



Boxplot ของ Quality และ Height

```
In [191... plt.figure(figsize=(16,5))
plt.subplot(1,2,1)
sns.boxplot(x=point)
plt.xlabel("Quality")
plt.subplot(1,2,2)
sns.boxplot(x=altitude_series)
plt.xlabel("Height")
```

```
Out[191... Text(0.5, 0, 'Height')
```

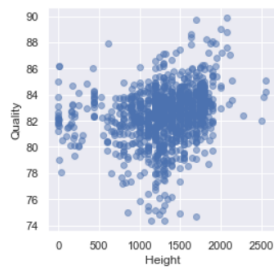


ทำ Scatter plot หาค่าความสัมพันธ์ระหว่าง Quality กับ height

ทำ Scatter plot หาความสัมพันธ์ระหว่าง Quality กับ height

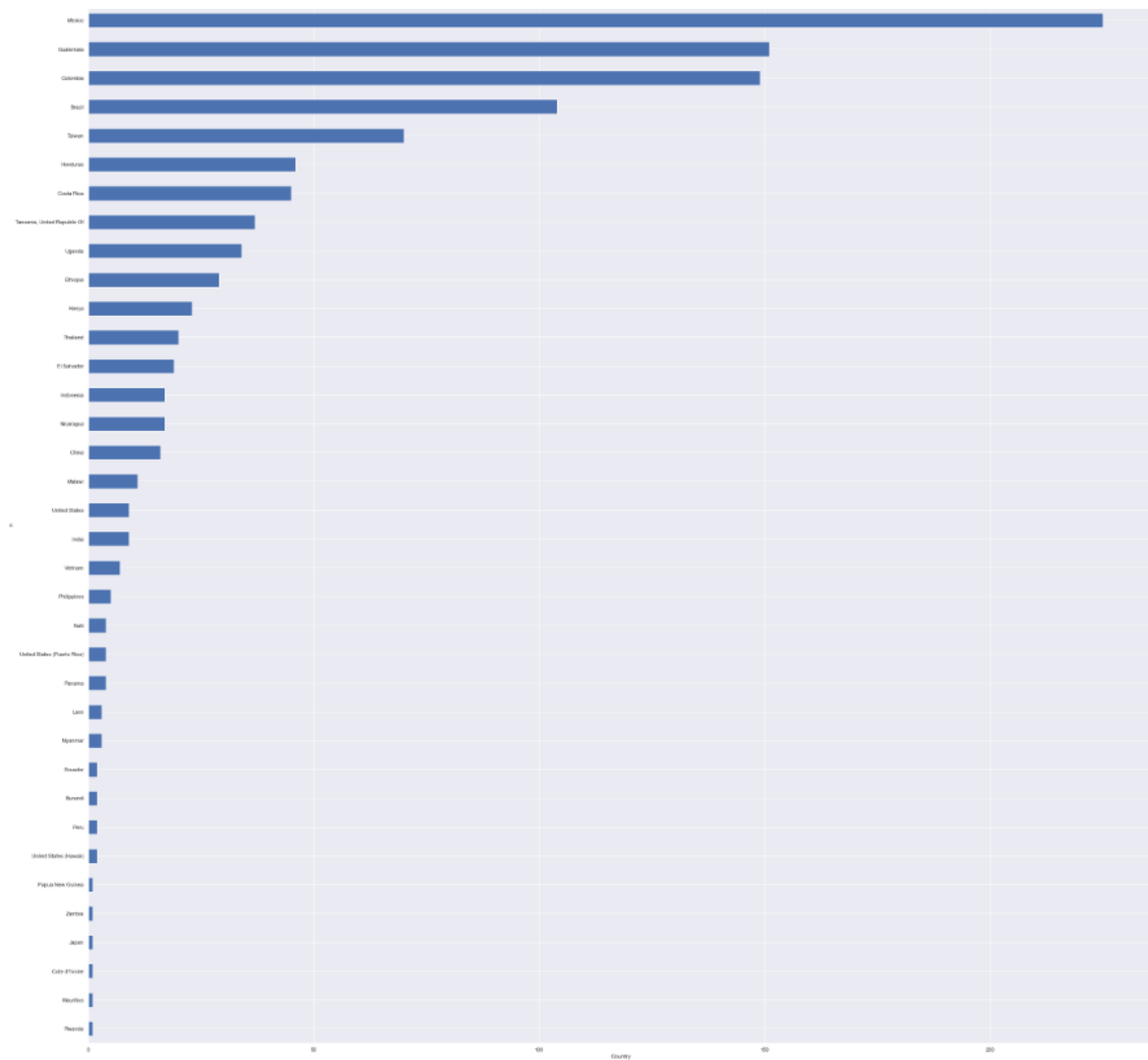
In [214..

```
plt.scatter(altitude_series, point, alpha=0.5)
plt.xlabel("Height")
plt.ylabel("Quality")
plt.show()
```



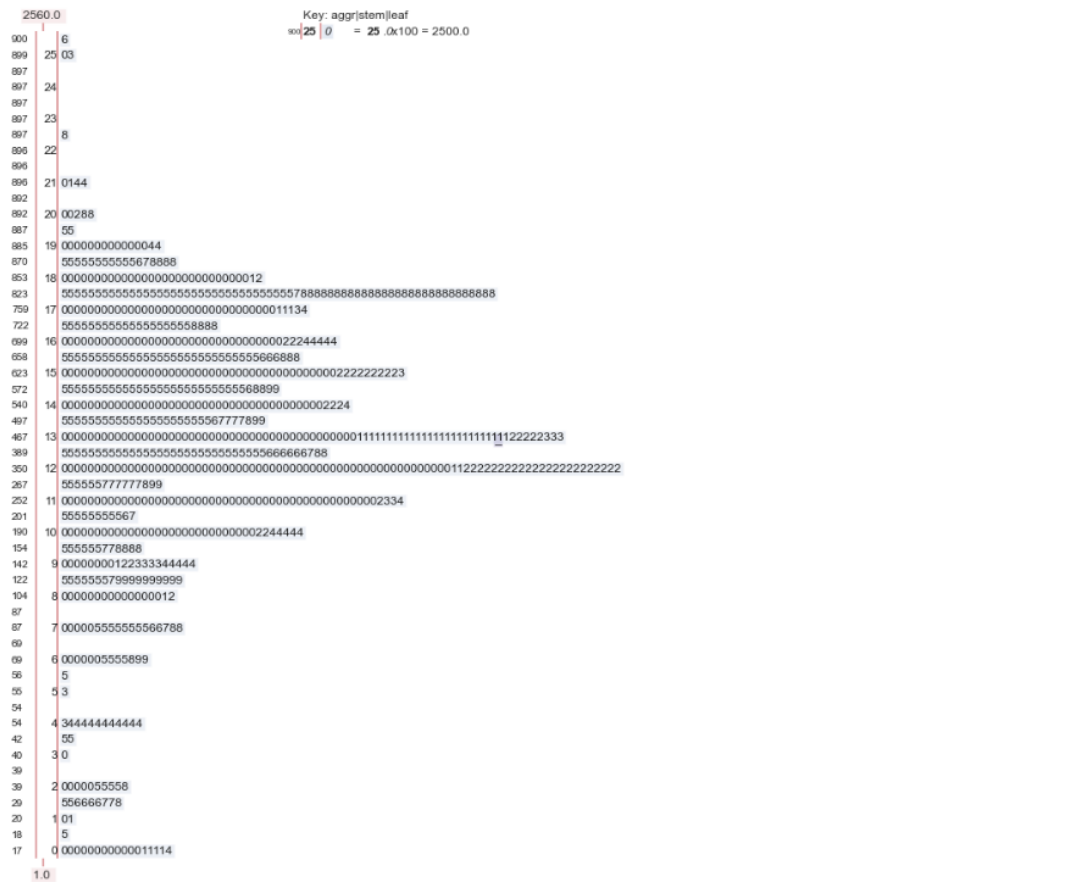
In [193..

```
new_df['country'] = new_df.loc[:, 'Country.of.Origin']
plt.figure(figsize=(40,40))
new_df.country.value_counts().sort_values().plot(kind = 'barh')
plt.xlabel("Country")
plt.ylabel("n")
plt.title("Country")
plt.show()
```



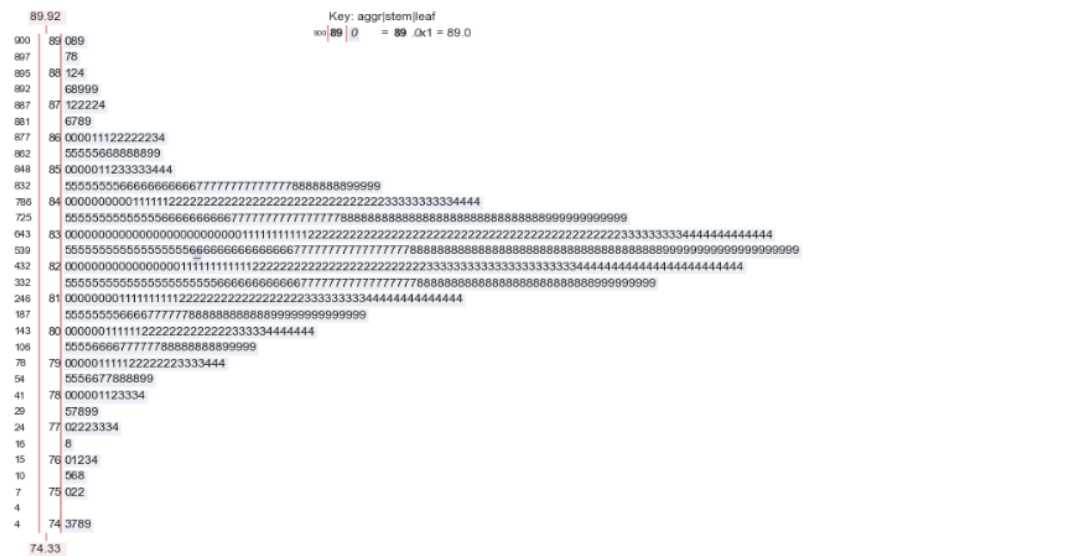
```
In [194]: stemgraphic.stem_graphic(new_df['altitude_mean_meters'])
```

```
Out[194_... (<Figure size 540x990 with 1 Axes>, <Axes:>)
```



```
In [195_ stemgraphic.stem_graphic(new_df['Total.Cup.Points'])
```

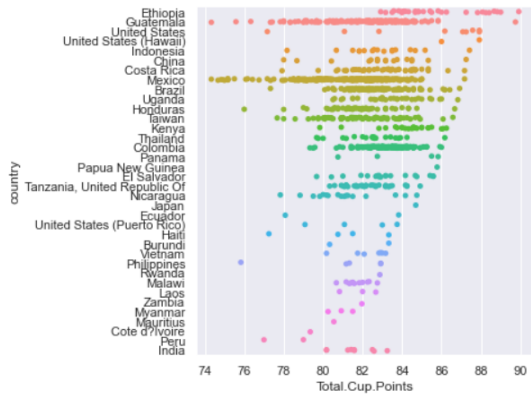
```
Out[195_]: (<Figure size 540x612 with 1 Axes>, <Axes:>)
```



ดูว่าแต่ละประเทศมีคะแนนคุณภาพแบบไหน

In [196..

```
try:
    sns.catplot(point,new_df['country'],data=new_df)
except ValueError:
    plt.show()
```



หาค่าเฉลี่ยคะแนนของประเทศที่สนใจ

In [210..

```
print('average points')
print('Ethiopia = '+str(df.loc[df['Country.of.Origin'] == 'Ethiopia']['Total.Cup.Points'].mean()) + ' points')
print('Mexico = '+str(df.loc[df['Country.of.Origin'] == 'Mexico']['Total.Cup.Points'].mean()) + ' points')
print('Brazil = '+str(df.loc[df['Country.of.Origin'] == 'Brazil']['Total.Cup.Points'].mean()) + ' points')
print('Colombia = '+str(df.loc[df['Country.of.Origin'] == 'Colombia']['Total.Cup.Points'].mean()) + ' points')
print('Guatemala = '+str(df.loc[df['Country.of.Origin'] == 'Guatemala']['Total.Cup.Points'].mean()) + ' points')
print('Thailand = '+str(df.loc[df['Country.of.Origin'] == 'Thailand']['Total.Cup.Points'].mean()) + ' points')
```

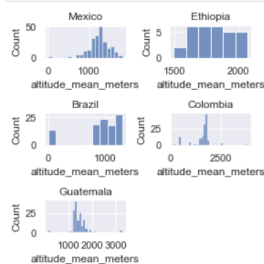
```
average points
Ethiopia = 85.91633333333333 points
Mexico = 80.86306034482759 points
Brazil = 82.71144230769232 points
Colombia = 83.22666666666667 points
Guatemala = 82.00954838709677 points
Thailand = 82.42999999999999 points
```

ดูกันว่าในแต่ละประเทศมีการปลูกกาแฟที่ความสูงเท่าไร

ดูกันว่าในแต่ละประเทศมีการปลูกกาแฟที่ความสูงเท่าไร

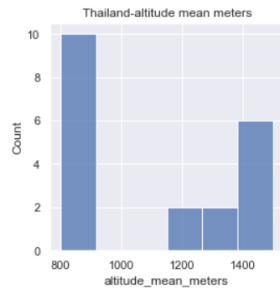
In [198..

```
h1 = plt.subplot(3,2,1);
plt.title("Mexico");
plt.tight_layout()
sns.histplot(df.loc[df['Country.of.Origin'] == 'Mexico']['altitude_mean_meters'])
h2 = plt.subplot(3,2,2);
plt.title("Ethiopia");
plt.tight_layout()
sns.histplot(df.loc[df['Country.of.Origin'] == 'Ethiopia']['altitude_mean_meters']);
h3 = plt.subplot(3,2,3);
plt.title("Brazil");
plt.tight_layout()
sns.histplot(df.loc[df['Country.of.Origin'] == 'Brazil']['altitude_mean_meters']);
h4 = plt.subplot(3,2,4);
plt.title("Colombia");
plt.tight_layout()
sns.histplot(df.loc[df['Country.of.Origin'] == 'Colombia']['altitude_mean_meters']);
h5 = plt.subplot(3,2,5);
plt.title("Guatemala");
plt.tight_layout()
sns.histplot(df.loc[df['Country.of.Origin'] == 'Guatemala']['altitude_mean_meters']);
```



```
In [199...
# plt.title("Colombia-altitude mean meters")
# sns.histplot(df.loc[df['Country.of.Origin'] == 'Colombia']['altitude_mean_meters']);
```

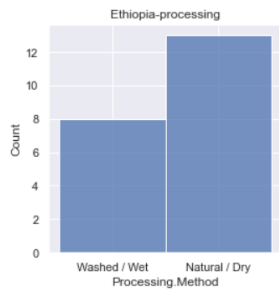
```
In [212...
plt.title("Thailand-altitude mean meters")
sns.histplot(df.loc[df['Country.of.Origin'] == 'Thailand']['altitude_mean_meters']);
```



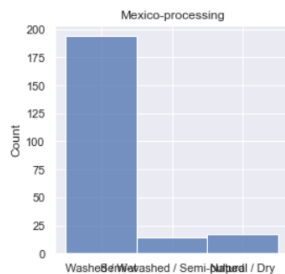
มาดูกันอีกว่าแต่ละประเทศใช้กระบวนการผลิตกาแฟแบบไหน

```
In [201...
# h1 = plt.subplot(3,2,1);
# plt.title("Ethiopia-processing")
# plt.tight_layout()
# sns.histplot(df.loc[df['Country.of.Origin'] == 'Ethiopia']['Processing.Method']);
# h2 = plt.subplot(3,2,2);
# plt.title("Mexico-processing")
# plt.tight_layout()
# sns.histplot(df.loc[df['Country.of.Origin'] == 'Mexico']['Processing.Method']);
# h3 = plt.subplot(3,2,3);
# plt.title("Brazil-processing")
# plt.tight_layout()
# sns.set(rc = {'figure.figsize':(15,8)})
# sns.histplot(df.loc[df['Country.of.Origin'] == 'Brazil']['Processing.Method']);
# h4 = plt.subplot(3,2,4);
# plt.title("Colombia-processing")
# plt.tight_layout()
# sns.histplot(df.loc[df['Country.of.Origin'] == 'Colombia']['Processing.Method']);
# h5 = plt.subplot(3,2,5);
# plt.title("Guatemala-processing")
# plt.tight_layout()
# sns.set(rc = {'figure.figsize':(4,4)})
# sns.histplot(df.loc[df['Country.of.Origin'] == 'Guatemala']['Processing.Method']);
```

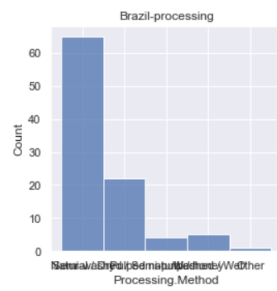
```
In [202...
plt.title("Ethiopia-processing")
sns.histplot(df.loc[df['Country.of.Origin'] == 'Ethiopia']['Processing.Method']);
```



```
In [203...
plt.title("Mexico-processing")
sns.histplot(df.loc[df['Country.of.Origin'] == 'Mexico']['Processing.Method']);
```



```
In [204... plt.title("Brazil-processing")
sns.set(rc = {'figure.figsize':(15,8)})
sns.histplot(df.loc[df['Country.of.Origin'] == 'Brazil']['Processing.Method']);
```

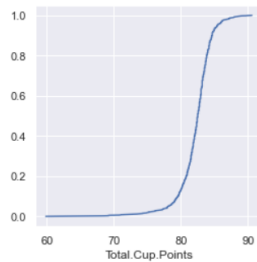


```
In [ ]: plt.title("Colombia-processing")
sns.histplot(df.loc[df['Country.of.Origin'] == 'Colombia']['Processing.Method']);
```

```
In [ ]: plt.title("Thailand-processing")
sns.set(rc = {'figure.figsize':(4,4)})
sns.histplot(df.loc[df['Country.of.Origin'] == 'Thailand']['Processing.Method']);
```

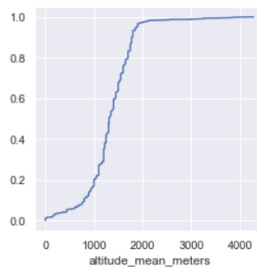
```
In [218... point = df.loc[:, 'Total.Cup.Points']
x = np.sort(point)
y = 1. * np.arange(len(point)) / (len(point) - 1)
plt.plot(x, y)
plt.xlabel('Total.Cup.Points')
```

Out[218... Text(0.5, 0, 'Total.Cup.Points')



```
In [219... altitude_series = df.loc[:, 'altitude_mean_meters']
x = np.sort(altitude_series)
y = 1. * np.arange(len(altitude_series)) / (len(altitude_series) - 1)
plt.plot(x, y)
plt.xlabel('altitude_mean_meters')
```

Out[219... Text(0.5, 0, 'altitude_mean_meters')



In [167...

```
point = pd.to_numeric(point)
print(point)

# Mean of point
myMean = point.mean()
print(myMean)

# degrees of freedom (dof) = samplesize - 1
n = len(point)
print(n)
dof = n-1

# Standard error (SE)
SE = sem(point)
print(SE)

#Confidence Interval (CI) 90% 95% 99%
CIlist = [0.90 , 0.95 , 0.99]
print(t.interval(CIlist[0],dof,loc=myMean,scale= SE ))
print(t.interval(CIlist[1],dof,loc=myMean,scale= SE ))
print(t.interval(CIlist[2],dof,loc=myMean,scale= SE ))

def getCI(cl,dof,mean,se) :
    return t.interval(cl,dof,loc=mean,scale=se)
```

```
alpha = [1-CIlist[0],1-CIlist[1],1-CIlist[2] ]
print(alpha)

myT = [t.ppf(alpha[0]/2,dof,loc=0, scale = 1),t.ppf(alpha[1]/2,dof,loc=0, scale = 1),t.ppf(alpha[2]/2,dof,loc=0, scale = 1)]
print(myT)

#margin of error
ME = [myT[0] * SE, myT[1] * SE,myT[2] * SE]

print(ME)
```

```
LCL90 = myMean + ME[0]
UCL90 = myMean - ME[0]
print(LCL90,UCL90)
```

82.26520850227269 82.47767521865757

```
LCL95 = myMean + ME[1]
UCL95 = myMean - ME[1]

print(LCL95,UCL95)
```

82.24482350156332 82.49806021936693

```
LCL99 = myMean + ME[2]
UCL99 = myMean - ME[2]

print(LCL99,UCL99)
```

82.2049287611408 82.53795495978945

```
data_dict = {}
data_dict['category'] = ['99% confidence level','95% confidence level','90% confidence level']
data_dict['lower'] = [LCL99,LCL95,LCL90]
data_dict['upper'] = [UCL99,UCL95,UCL90]
dataset = pd.DataFrame(data_dict)
for lower,upper,y in zip(dataset['lower'],dataset['upper'],range(len(dataset))):
    plt.plot((lower,upper),(y,y),'ro-',color='blue')
plt.yticks(range(len(dataset)),list(dataset['category']))
```

```

# CI
arr = []
for i in range(50):
    df_point = df.sample(n=50)
    point2 = df_point.loc[:, 'Total.Cup.Points']
    myMean1 = point2
    size = len(point2)
    dgf = size-1
    SE1 = sem(point2)
    arr.append(getCI(CIlist[0], dgf, myMean, SE1))

print(arr)

```

```

import sklearn
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures

#df_height = df[['altitude_mean_meters']]
df_height = df['altitude_mean_meters']
df_point = df['Total.Cup.Points']
mean = df_point.mean().sum()
plt.scatter(df_height, df_point)
plt.xlabel("Height")
plt.ylabel("Quality")

print(df_height)

```

```

In [168... model = LinearRegression() # creat object
x=df[['altitude_mean_meters']] # expect 2D array not Series (1D)
y=df['Total.Cup.Points']

model.fit(x, y)
model.score(x, y)

```

Out[168... 0.022922923589222655

```

In [168... model.intercept_

```

Out[168... 81.09753094876508

```

In [168... model.coef_

```

Out[168... array([0.0008178])

```

In [168... predictions = model.predict(x)
predictions

```

Out[168... array([82.79446656, 82.79446656, 82.48779145, ..., 83.55093183,
81.13024296, 81.74768218])

```

In [168... plt.scatter(x,y)
plt.plot(x,predictions,color='black')
plt.xlabel("Height")
plt.ylabel("Quality")

```

```

df_countrys = new_df.groupby(df['Country.of.Origin'])
df_ethiopia = df_countrys.get_group('Ethiopia')
df_mexico = df_countrys.get_group('Mexico')
df_brazil = df_countrys.get_group('Brazil')
df_colombia = df_countrys.get_group('Colombia')
df_thailand = df_countrys.get_group('Thailand')

```

```

#Ethiopia
df_ethiopia = df_countrys.get_group('Ethiopia')
x = df_ethiopia[['altitude_mean_meters']]
y = df_ethiopia['Total.Cup.Points']
model.fit(x, y)
model.intercept_
model.coef_
predictions = model.predict(x)
predictions
plt.scatter(x,y)
plt.plot(x,predictions,color='black')
plt.xlabel("Height")
plt.ylabel("Quality")
model.score(x,y)

```

#Mexico

```
x = df_mexico[['altitude_mean_meters']]
y = df_mexico['Total.Cup.Points']
model.fit(x, y)
model.intercept_
model.coef_
predictions = model.predict(x)
predictions
plt.scatter(x,y)
plt.plot(x,predictions,color='black')
plt.xlabel("Height")
plt.ylabel("Quality")
model.score(x,y)
```

#Brazil

```
x = df_brazil[['altitude_mean_meters']]
y = df_brazil['Total.Cup.Points']
model.fit(x, y)
model.intercept_
model.coef_
predictions = model.predict(x)
predictions
plt.scatter(x,y)
plt.plot(x,predictions,color='black')
plt.xlabel("Height")
plt.ylabel("Quality")
model.score(x,y)
```

#Colombia

```
x = df_colombia[['altitude_mean_meters']]
y = df_colombia['Total.Cup.Points']
model.fit(x, y)
model.intercept_
model.coef_
predictions = model.predict(x)
predictions
plt.scatter(x,y)
plt.plot(x,predictions,color='black')
```

#Thailand

```
x = df_thailand[['altitude_mean_meters']]
y = df_thailand['Total.Cup.Points']
model.fit(x, y)
model.intercept_
model.coef_
predictions = model.predict(x)
predictions
plt.scatter(x,y)
plt.plot(x,predictions,color='black')
```


เอกสารอ้างอิง

<https://bit.ly/3Liuq9>

<https://api.dtn.go.th/files/v3/5e8712f2ef4140204c3022ce/download>

[ไล่เกาะแหล่งกำเนิดผลไม้ป่าจ ‘เอธิโอเปีย’ เมืองสวรรค์ของคอกาแฟ \(adaymagazine.com\)](http://adaymagazine.com)

[เมล็ดกาแฟ Washed , Natural หรือ Honey Process เรื่องที่คนรักกาแฟทุกคนควรรู้ - Coffee Press \(coffeepressthailand.com\)](http://coffeepressthailand.com)

[บทความที่น่าสนใจ - BKMS.com \(bluekoff.com\)](http://bluekoff.com)