บทที่ 4 การสำรวจข้อมูล

หัวข้อหลัก

- การสำรวจข้อมูล (data exploration หรือ exploratory data analysis) คือการศึกษาเกี่ยวกับคุณลักษณะพื้นฐานของชุดข้อมูลที่จะศึกษา เพื่อทำความเข้าใจเกี่ยวกับโครงสร้าง, ความสัมพันธ์ระหว่างแอทริบิวต์, และการกระจายตัวของข้อมูล
- เครื่องมือหลักสำหรับการสำรวจข้อมูล มีสองอย่างคือ (1) สถิติเชิงพรรณนา (descriptive statistics) และ (2) การแสดงข้อมูลด้วยภาพ (data visualization)

ก่อนเริ่มการวิเคราะห์ข้อมูลเชิงลึกด้วยเทคนิคขั้นสูง นักวิทยาศาสตร์ข้อมูลต้องทำความเข้าใจคุณลักษณะทั่วไปของชุดข้อมูลที่ทำการศึกษาก่อน เพื่อให้สามารถเลือกวิธีการเตรียมข้อมูลและเทคนิคการวิเคราะห์ข้อมูลที่เหมาะสมได้ กระบวนการทำความเข้าใจคุณสมบัติเบื้องต้นของข้อมูลดังกล่าว เรียกว่า **การสำรวจข้อมูล (data exploration)**

การสำรวจข้อมูลสามารถแบ่งออกได้เป็นสองประเภทตามชนิดของเครื่องมือที่นำมาใช้ ได้แก่ วิธีแรกคือการสำรวจข้อมูลด้วย**สถิติเชิงพรรณนา (descriptive statistics)** และ วิธีการที่สองคือการสำรวจข้อมูลด้วย**การทำให้เห็นภาพ (data visualization)**

1. การสำรวจข้อมูลใช้เมื่อใด

การสำรวจข้อมูล ถูกนำไปใช้ในขั้นตอนต่าง ๆ ในกระบวนการทางวิทยาศาสตร์ข้อมูล ดังนี้

- *การทำความเข้าใจข้อมูล* ในขั้นตอนนี้ เราใช้การสำรวจข้อมูลเพื่อให้เห็นความสัมพันธ์ระหว่างแอทริบิวต์ต่าง ๆ ในชุดข้อมูล รวมไปถึงช่วงของค่าของข้อมูล
- *การเตรียมข้อมูล* ขั้นตอนการเตรียมข้อมูลถือได้ว่าเป็นขั้นตอนที่ใช้เวลามากและมีความสำคัญที่สุดขั้นตอนหนึ่งของการบวนการทางวิทยาศาสตร์ข้อมูล เป้าหมายหลักคือการเตรียมข้อมูลให้อยู่ในรูปแบบที่เหมาะกับอัลกอริทึมการเรียนรู้ที่จะใช้ และการกำจัดค่าที่ผิดพลาดจากสาเหตุต่าง ๆ ออกไปจากชุดข้อมูล ตัวอย่างเช่น อัลกอริทึมการเรียนรู้บางชนิดจะมีประสิทธิภาพไม่ดีหากชุดข้อมูลสำหรับเรียนรู้มีแอทริบิวต์ที่มีความสัมพันธ์กัน (correlated attributes) เราสามารถใช้เทคนิคการสำรวจข้อมูลเพื่อตรวจจับ correlated attributes เหล่านี้ได้
- *การสร้างโมเดล (หรือการเรียนรู้รูปแบบที่ฝังตัวในชุดข้อมูล)* ในบางครั้งการสำรวจข้อมูลขั้นพื้นฐาน ก็สามารถค้นพบรูปแบบแฝงในชุดข้อมูลได้ตามความต้องการของงานทางวิทยาศาสตร์ข้อมูล เช่น การแบ่งกลุ่ม (clustering) เป็นต้น
- *การตีความผลลัพธ์* เทคนิคการสำรวจข้อมูลสามารถนำไปใช้ทำความเข้าใจผลลัพธ์ที่ได้จากกระบวนการทางวิทยาศาสตร์ข้อมูลได้ เช่น การใช้ตารางแจกแจงความถี่ (histogram) เพื่อทำความเข้าใจคุณสมบัติของสมาชิกของแต่ละคลาสที่ได้จากการทำนายของตัวจำแนกประเภท (classifier)

2. สถิติเชิงพรรณนา (Descriptive Statistics)

สถิติเชิงพรรณนา คือการศึกษาชุดข้อมูลในเชิงปริมาณ ด้วยการวัดค่าสรุปในมิติต่าง ๆ เกี่ยวกับข้อมูล เราจะแบ่งประเภทของการศึกษานี้ออกเป็นสองประเภทหลักคือ การสำรวจข้อมูลทีละหนึ่งแอทริบิวต์ (univariate exploration) และ การสำรวจข้อมูลทีละหลายแอทริบิวต์ (multivariate exploration)

2.1 Univariate Exploration

การสำรวจข้อมูลทีละแอทริบิวต์โดยใช้เครื่องมือทางสถิติ ตามตัวอย่างดังแสดงในตารางที่ 1

ตารางที่ 1 ค่าวัดทางสถิติ และ คุณลักษณะของข้อมูล

คุณลักษณะของชุดข้อมูล	เทคนิคการวัดทางสถิติเชิงพรรณนา
ค่ากลางของข้อมูล (center of dataset)	ค่าเฉลี่ย (mean), มัธยฐาน (median), ฐานนิยม (mode)
การแผ่กระจายของข้อมูล (spread of dataset)	พิสัย (range), ความแปรปรวน (variance), และส่วนเบี่ยงเบนมาตรฐาน (standard deviation)
รูปร่างของการกระจายตัวของข้อมูล (shape of the distribution of the dataset)	สมมาตร, เบนซ้ายหรือขวา (left skewed, right skewed)

ตัววัดค่าแนวโน้มสู่ส่วนกลาง (central tendency)

- ค่าเฉลี่ย คำนวณโดยการหารผลรวมของค่าแอทริบิวต์ที่ได้จากการสังเกตด้วยจำนวนข้อมูลทั้งหมด
- มัธยฐาน คือจุดกึ่งกลางของการกระจายของข้อมูล ค่ามัธยฐานหาได้โดยเรียงลำดับข้อมูลทั้งหมดจากน้อยไปหามาก และเลือกค่าที่อยู่กึ่งกลางของชุดข้อมูลที่เรียงลำดับแล้ว หากจำนวนข้อมูลเป็นจำนวนคู่ค่ามัธยฐานจะเท่ากับค่าเฉลี่ยของจุดข้อมูลสองจุดที่อยู่กึ่งกลาง
- ฐานนิยม คือค่าของข้อมูลที่พบบ่อยที่สุดในชุดข้อมูล

ความใกล้เคียงกันหรือความต่างกันของ ค่าเฉลี่ย, ค่ามัธยฐาน และค่าฐานนิยม เป็นตัวบ่งชี้รูปร่างของการกระจายของข้อมูล กล่าวคือ (1) หากค่าทั้งสามมีค่าเท่ากันแสดงว่าชุดข้อมูลมีการกระจายตัวแบบปกติ (normal distribution) (2) หากชุดข้อมูลมีค่าผิดปกติ (outliers) อยู่จะมีผลกระทบทำให้ค่าเฉลี่ย น้อยหรือมากกว่า ค่าอีกสองตัวที่เหลือ (3) หากชุดข้อมูลประกอบด้วยการกระจายตัวแบบปกติมากกว่าหนึ่งตัว จะมีผลกระทบทำให้ค่าฐานนิยม มีค่าต่างไปจากค่าเฉลี่ยและมัธยฐาน และ (4) หากชุดข้อมูลมีความเบี่ยงเบนในการกระจายตัว จะส่งผลให้ค่ากลางทั้งสามตัวมีค่าไม่เท่ากัน

ตัววัดการแผ่กระจายของข้อมูล (spread)

- พิสัย คือความแตกต่างระหว่างค่าที่มากที่สุดและค่าที่น้อยที่สุดภายในชุดข้อมูล
- ความแปรปรวน คือผลรวมของกำลังสองของความแตกต่างระหว่างค่าของข้อมูลแต่ละจุดกับค่าเฉลี่ยหารด้วยจำนวนจุดข้อมูล
- ส่วนเบี่ยงเบนมาตรฐาน คือรากที่สองของค่าความแปรปรวน

หากค่าของตัววัดการแผ่กระจาย (พิสัย, ความแปรปรวน, ส่วนเบี่ยงเบนมาตรฐาน) มีค่าสูง แสดงว่าชุดข้อมูลมีการกระจายตัวแบบแผ่กว้างรอบจุดกึ่งกลาง ถ้าชุดข้อมูลมีการกระจายตัวแบบปกติ (normal distribution) 68% ของจุดข้อมูลจะอยู่ห่างจากค่าเฉลี่ยไม่เกินหนึ่งเท่าของค่าส่วนเบี่ยงเบนมาตรฐาน

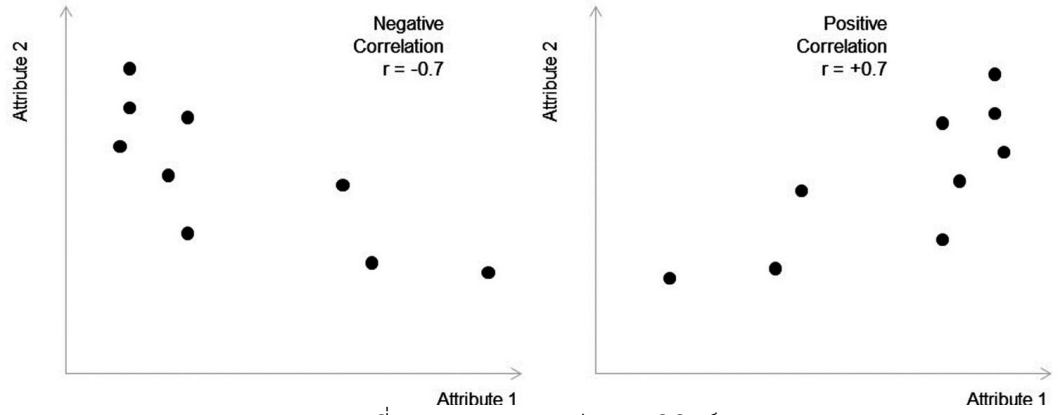
2.2 Multivariate Exploration

การสำรวจข้อมูลทีละหลายแอทริบิวต์ คือการศึกษาความสัมพันธ์ระหว่างแอทริบิวต์ (หลายตัว) ในชุดข้อมูล ซึ่งทำได้โดยใช้เครื่องมือทางสถิติเชิงพรรณนา (เช่น การวัดค่า correlation) หรือการใช้เทคนิค data visualization ในการทำให้เห็นแนวโน้มของความสัมพันธ์ระหว่างแอทริบิวต์แต่ละตัว

ค่า correlation

เป็นการวัดความสัมพันธ์ในเชิงสถิติระหว่างแอทริบิวต์สองตัว เพื่อตรวจสอบ dependency ของแอทริบิวต์ หากแอทริบิวต์สองตัวมี correlation สูง แสดงว่าแอทริบิวต์คู่ดังกล่าวจะมีอัตราการเปลี่ยนแปลงในอัตราพอ ๆ กันในทิศทางเดียวกันหรือตรงกันข้าม ตัวอย่างเช่น แอทริบิวต์ อุณหภูมิ กับ แอทริบิวต์ ยอดขายไอศกรีม มี correlation สูงแบบทิศทางเดียวกัน เนื่องจากเมื่อค่าอุณหภูมิสูงขึ้น ยอดขายไอศกรีมก็มักจะสูงขึ้นตามด้วย

ค่า correlation ของแอทริบิวต์สองตัวมักถูกวัดโดยใช้ *Pearson correlation coefficient* ซึ่งใช้วัดดีกรีของ ความแปรผันเชิงเส้น (linear dependence) ดังรูปตัวอย่างที่ 1



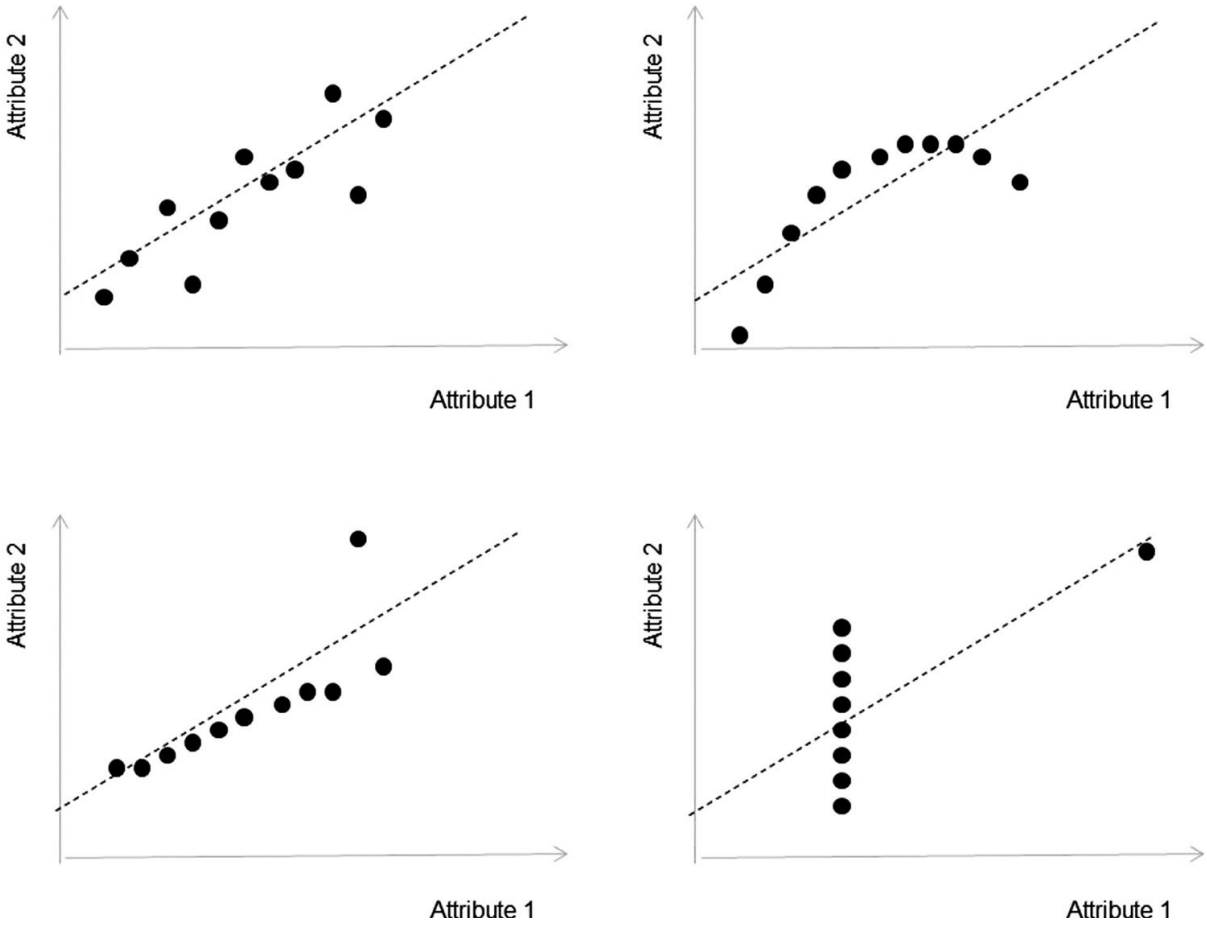
รูปที่ 1 correlation ระหว่างแอทริบิวต์

ค่า Pearson correlation coefficients จะมีค่าอยู่ระหว่าง -1 ถึง 1 ค่าที่ใกล้เคียง -1 หรือ 1 บ่งชี้ว่าแอทริบิวต์คู่นั้นมีความสัมพันธ์กันแบบแปรผกผัน หรือ แปรผันตามกันสูง ตามลำดับ หากค่า coefficient ที่ได้มีค่าเท่ากับ 0 หมายความว่าแอทริบิวต์คู่ดังกล่าวไม่มีความสัมพันธ์เชิงเส้นระหว่างกัน สูตรคำนวณค่า Pearson correlation coefficient ระหว่างแอทริบิวต์ *x* และ *y* คือ

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{N\times s_x\times s_y}$$

เมื่อ $\bar{x}$ คือ ค่าเฉลี่ยของแอทริบิวต์ *x*, $\bar{y}$ คือ ค่าเฉลี่ยของแอทริบิวต์ *y*, s_x คือค่าเบี่ยงเบนมาตรฐานของ x และ s_y คือค่าเบี่ยงเบนมาตรฐานของ *y*

Pearson correlation coefficient สามารถวัดได้เฉพาะความสัมพันธ์ระหว่างตัวแปรแบบเชิงเส้นเท่านั้น หากชุดข้อมูลมีความสัมพันธ์ที่ซับซ้อน ดังเช่นตัวอย่างในรูปที่ 2 เราจำเป็นต้องใช้วิธีอื่น (เช่น data visualization) ค้นหาความสัมพันธ์ดังกล่าวแทน



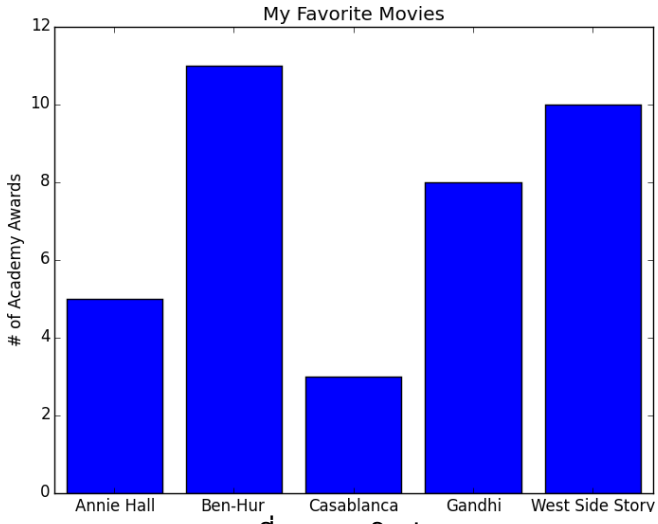
รูปที่ 2 Anscombe's Quartet: ชุดข้อมูลทั้งสี่ชุดมีค่าวัดเชิงสถิติเท่ากัน (ค่าเฉลี่ย ความแปรปรวน และ correlation) แต่กลับมีลักษณะแตกต่างกันอย่างสิ้นเชิงเมื่อพล็อตบนแผนภูมิ

3. การแสดงข้อมูลด้วยภาพ (Data Visualization)

Data visualization ถูกนำไปใช้งานในสองด้าน คือ เพื่อการสำรวจข้อมูล และ เพื่อสื่อสารข้อมูลหรือผลลัพธ์ที่ได้จากกระบวนการทางวิทยาศาสตร์ข้อมูล รูปแบบของ data visualization ที่ใช้งานบ่อย ได้แก่ แผนภูมิแท่ง (bar charts), แผนภูมิเส้น (line charts), แผนภูมิแบบกระจาย (scatterplots), ฮิสโตแกรม (histograms), บ๊อกซ์วิสก์เกอร์ (box whisker plot)

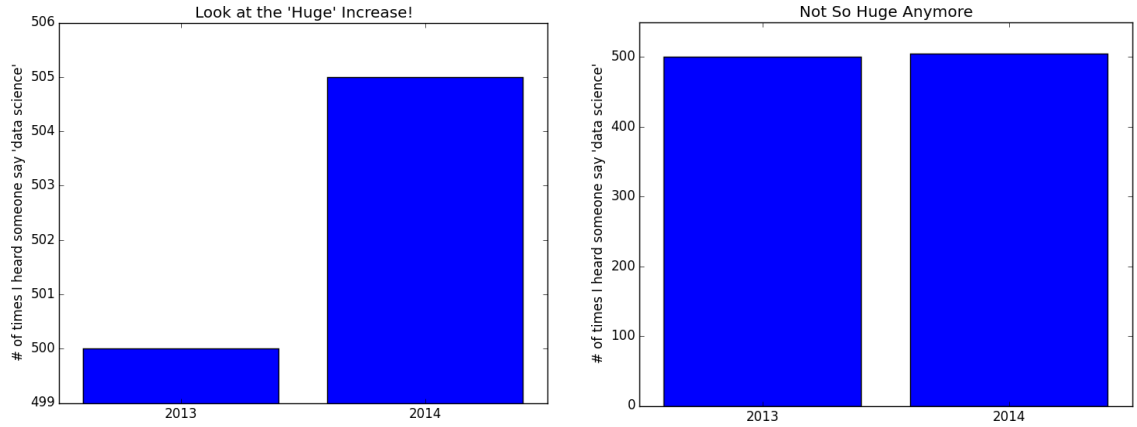
3.1 แผนภูมิแท่ง (Bar Charts)

แผนภูมิแท่งเหมาะสำหรับใช้แสดงให้เห็นถึงการเปลี่ยนแปลงค่าของ items แต่ละตัวในชุดข้อมูล เช่น แผนภูมิแท่งในรูปที่ 3 แสดงจำนวนรางวัลออสการ์ที่ภาพยนตร์แต่ละเรื่องได้รับ



รูปที่ 3 แผนภูมิแท่ง

ข้อควรระวังในการใช้แผนภูมิแท่งคือ แกน y ของแผนภูมิจะต้องมีค่าตั้งต้นจากศูนย์ ไม่เช่นนั้นอาจทำให้การตีความหมายข้อมูลผิดพลาดได้ ดังตัวอย่างในรูปที่ 4

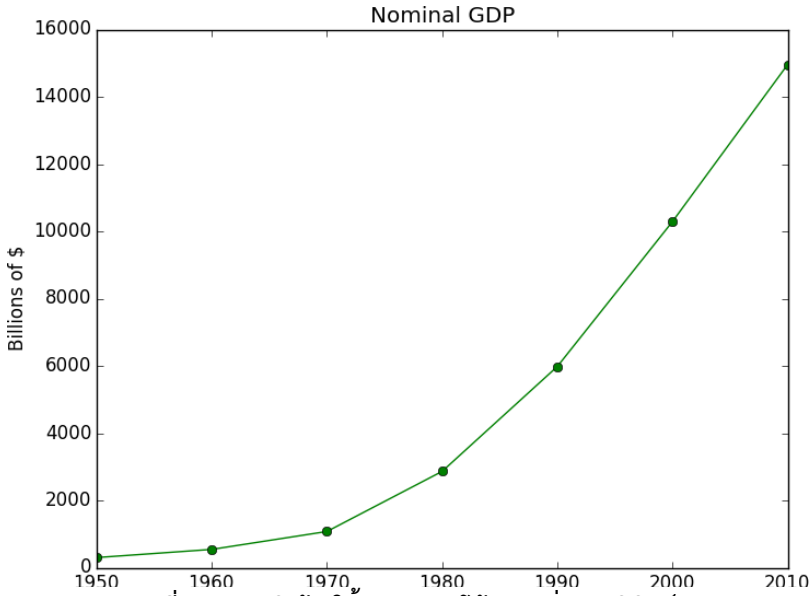


(ก) (ข)

รูปที่ 4 (ก) แผนภูมิแท่งที่แกน y ไม่ได้เริ่มจากศูนย์ ทำให้ดูเหมือนค่าของปี 2013 กับ 2014 มีความต่างกันมาก (ข) เมื่อพล็อตแผนภูมิแท่งโดยให้แกน y เริ่มต้นจากศูนย์ จะเห็นได้ชัดว่าค่าของปี 2013 กับ 2014 ต่างกันน้อยมาก

3.2 แผนภูมิเส้น (Line Charts)

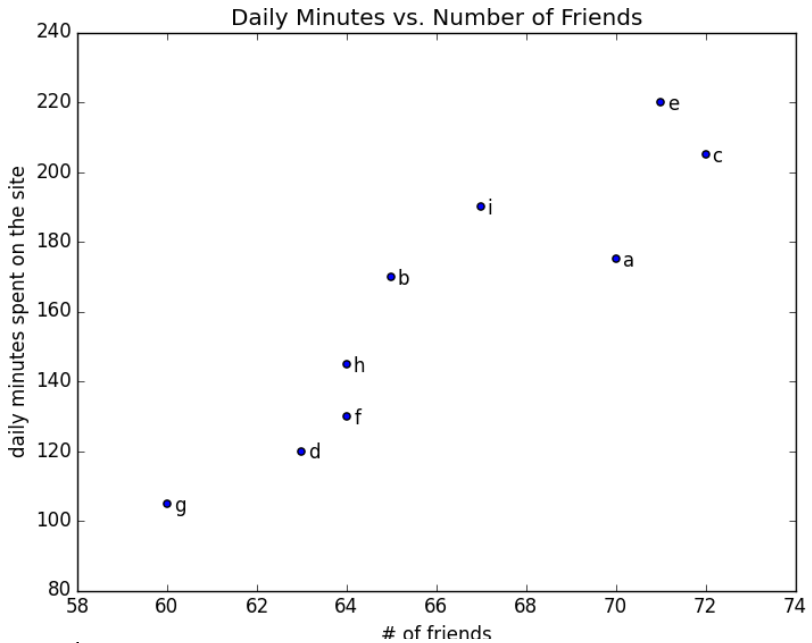
แผนภูมิเส้นเหมาะสำหรับ การแสดงแนวโน้ม ภายในชุดข้อมูล เช่น แผนภูมิเส้นในรูปที่ 5 แสดงให้เห็นว่าค่า Nominal GDP มีแนวโน้มเพิ่มสูงขึ้นเรื่อย ๆ ในช่วงปี 1950 ถึงปี 2010



รูปที่ 5 แผนภูมิเส้น ใช้แสดงแนวโน้มของค่าแอทริบิวต์

3.3 แผนภูมิแบบกระจาย (scatterplots)

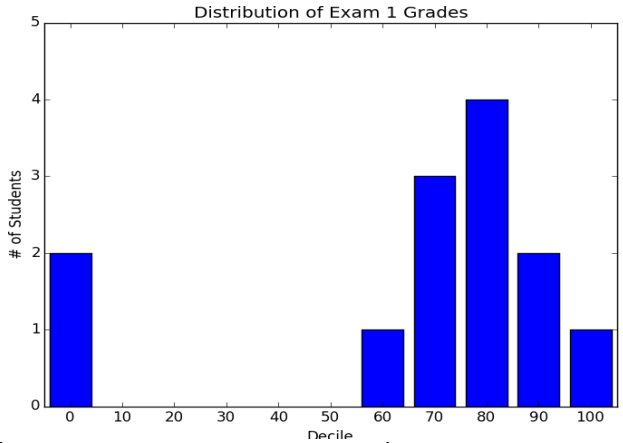
แผนภูมิแบบกระจาย เหมาะสำหรับใช้แสดงความสัมพันธ์ระหว่างแอทริบิวต์สองตัวในชุดข้อมูล ตัวอย่างเช่น รูปที่ 6 แสดงความสัมพันธ์ระหว่างจำนวนเพื่อนและจำนวนนาทีการใช้งานเว็บไซต์ของผู้ใช้งานในชุดข้อมูล (a, b, c, d, e, f, g, h, i)



รูปที่ 6 แผนภูมิแบบกระจาย แสดงความสัมพันธ์ระหว่างแอทริบิวต์สองตัว

3.4 ฮิสโตแกรม (Histogram)

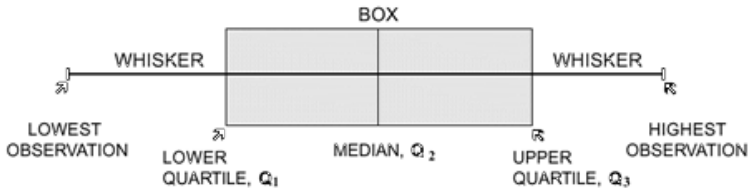
ฮิสโตแกรม เป็นเทคนิคพื้นฐานของการแสดงข้อมูลให้เป็นภาพ ซึ่งใช้สำหรับทำความเข้าใจความถี่ของการเกิดขึ้นของค่าแต่ละค่าของแอทริบิวต์ ตัวอย่างเช่น ฮิสโตแกรมในรูปที่ 7 แสดงการกระจายของคะแนนสอบของนักศึกษากลุ่มหนึ่ง



รูปที่ 7 ฮิสโตแกรม แสดงจำนวนนักศึกษาที่ได้คะแนนสอบในแต่ละช่วง

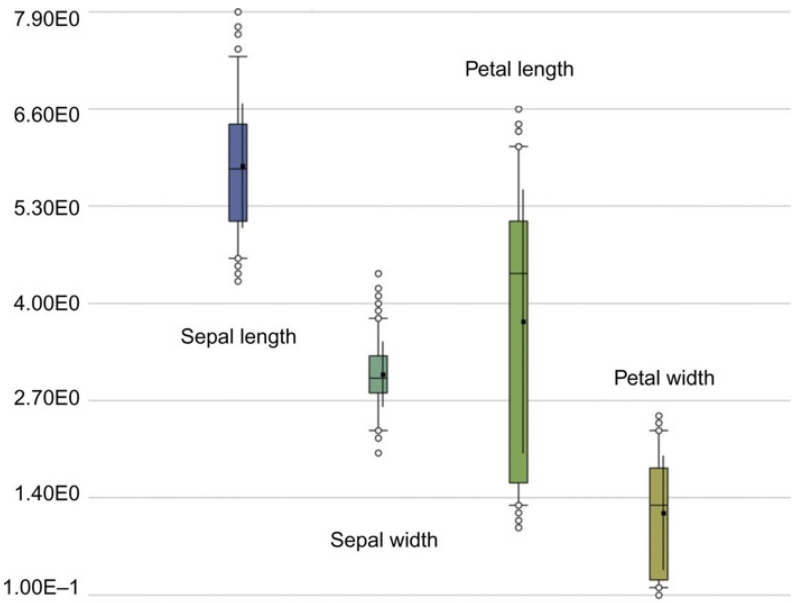
3.5 แผนภูมิบ๊อกซ์วิสก์เกอร์ (box whisker plot)

Box whisker plot เป็นแผนภูมิที่ใช้แสดงและเปรียบเทียบค่าทางสถิติห้าตัวคือ ค่าต่ำสุด ค่าควอร์ไทล์ลำดับที่หนึ่ง ค่ามัธยฐาน ค่าควอร์ไทล์ลำดับที่สอง และค่าสูงสุด ของแอทริบิวต์ ดังภาพประกอบในรูปที่ 8

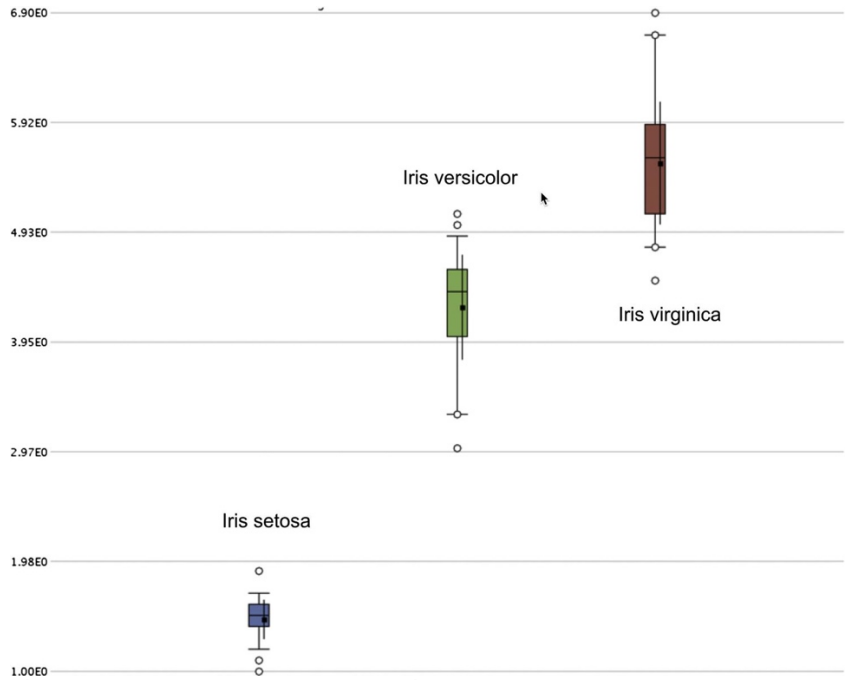


รูปที่ 8 Box whisker plot

Box whisker plot มักถูกใช้สำหรับเปรียบเทียบการกระจายตัวของข้อมูลระหว่างแอทริบิวต์หลาย ๆ ตัว หรือระหว่างแอทริบิวต์ตัวเดียวกันจากแต่ละประเภทในชุดข้อมูล ดังตัวอย่างในรูปที่ 9 และรูปที่ 10 ตามลำดับ



รูปที่ 9 Box whisker plot เปรียบเทียบการกระจายตัวของแอทริบิวต์สี่ตัวคือ sepal length, sepal width, petal length, petal width



รูปที่ 10 Box whisker plot เปรียบเทียบค่า petal length ของดอกไอริสสามสายพันธุ์

แบบฝึกหัด

1. กำหนดค่า GDP (Gross Domestic Product) ในช่วงแปดปี (พ.ศ. 2541-2548) ดังนี้คือ 300.2, 543.3, 1075.9, 2862.5, 5976.6, 10289.7, 14958.3, 543.3 จงคำนวณค่าเฉลี่ย, มัธยฐาน, ฐานนิยม, พิสัย, ความแปรปรวน, และ ค่าเบี่ยงเบนมาตรฐาน
2. หากต้องการแสดงแนวโน้มของค่า GDP ควรใช้แผนภูมิชนิดใด
3. เมื่อใดควรใช้แผนภูมิการกระจาย (scatterplots) ในการแสดงข้อมูลด้วยภาพ จงยกตัวอย่างประกอบ

เอกสารอ้างอิง

[1] Bala Deshpande, Vijay Kotu. *Data Science*. 2nd Edition, Morgan Kaufmann, 2018.

[2] Joel Grus. *Data Science from Scratch*, 2nd Edition, O'Reilly Media, Inc., 2019.

[3] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining Concepts and Techniques*, 3rd Edition, Morgan Kaufmann, 2012.