

การมองภาพข้อมูล

การมองภาพข้อมูล เป็นอีกแนวทางหนึ่งในการสำรวจและอธิบายข้อมูล นอกเหนือจากการใช้สถิติเชิงพรรณนา ซึ่งได้กล่าวมาในบทที่ 3 การมองภาพข้อมูลทำให้เราเห็นภาพรวมข้อมูลทั้งหมดในมุมมองที่สนใจด้วยภาพเพียงไม่กี่ภาพ และยังสามารถแปลความหมายของสิ่งที่ปรากฏในข้อมูลได้ง่ายกว่าค่าทางสถิติมาก จึงทำให้การมองภาพข้อมูลถูกนำไปใช้สำหรับการนำเสนอรายงานการวิเคราะห์อีกด้วย ในบทนี้จะกล่าวถึงวิธีการมองภาพข้อมูลแบบต่างๆ ที่นำเสนอข้อมูลในมุมมองที่ต่างกัน สำหรับข้อมูลแต่ละประเภท อย่างไรก็ตาม ปัจจุบันมีวิธีการมองภาพข้อมูลถูกนำเสนอขึ้นเป็นจำนวนมาก จึงจะขอนำมากล่าวเฉพาะวิธีการพื้นฐานที่ควรทราบและเป็นที่แพร่หลายเท่านั้น

วัตถุประสงค์

1. นักศึกษารู้รูปแบบการนำเสนอภาพข้อมูลพื้นฐานแบบต่างๆ
2. นักศึกษาสามารถอธิบายลักษณะของข้อมูลจากภาพข้อมูลรูปแบบต่างๆ ได้
3. นักศึกษาสามารถเลือกใช้รูปแบบการนำเสนอภาพข้อมูลได้อย่างเหมาะสม

6.1 เกี่ยวกับการมองภาพข้อมูล

การมองภาพข้อมูล (Data Visualization) คือ การใช้แผนภาพหรือกราฟในการแสดงข้อมูล ถือเป็นแนวทางการวิเคราะห์เชิงพรรณนาอีกแนวทางหนึ่ง และยังเป็นการสร้างการสื่อสารผลการวิเคราะห์ข้อมูลระหว่างนักวิเคราะห์ข้อมูลกับบุคคลอื่นด้วย การนำข้อมูลมาสร้างเป็นภาพในศาสตร์ของวิทยาการข้อมูลถือเป็นทั้งวิทยาศาสตร์และศิลปะ โดยจะต้องนำเอาความรู้ทางศิลปะมาใช้สร้างภาพจากข้อมูลโดยไม่ทำให้ความจริง และเหตุผลในข้อมูลผิดเพี้ยนไป นั่นหมายความว่า เมื่อสร้างภาพจากข้อมูลแล้ว ภาพนั้นจะต้องยังแสดงข้อมูลได้อย่างถูกต้อง ครบถ้วน และไม่ก่อให้เกิดการตีความข้อมูลจากภาพผิดไปจากความเป็นจริง

การมองภาพข้อมูล มีบทบาทสำคัญในกระบวนการวิเคราะห์ข้อมูล 2 ประการ คือ

- ▶ เป็นการทำให้เห็นข้อมูลปริมาณมากในมุมมองต่างๆ อันจะทำให้เกิดความเข้าใจในข้อมูลมากยิ่งขึ้น และนำไปสู่การตั้งสมมติฐานสำหรับการวิเคราะห์ข้อมูลต่อไป
- ▶ เป็นการสร้างการสื่อสารผลลัพธ์ของการวิเคราะห์ข้อมูลต่อบุคคลทั่วไปได้อย่างชัดเจน

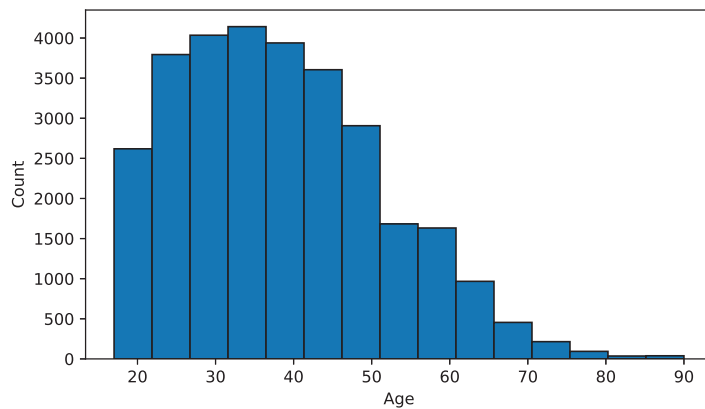
เพื่อให้การมองภาพข้อมูลเป็นไปอย่างมีประสิทธิภาพและสัมฤทธิ์ผล มีข้อแนะนำในการสร้างภาพจากข้อมูล ดังนี้

1. ตั้งเป้าหมายหรือวัตถุประสงค์ของการสร้างภาพข้อมูล บางครั้งอาจเป็นในรูปของคำถามที่ต้องการทราบหรือชี้ชวนให้ผู้ฟังติดตาม ผู้สร้างภาพข้อมูลมีหน้าที่ที่ต้องนำเสนอข้อมูลในมุมมองต่างๆ เพื่อนำไปสู่การบรรลุเป้าหมายหรือตอบคำถามที่ตั้งไว้
2. ให้ความสำคัญกับผู้ฟังและความต้องการของผู้ฟัง ผู้สร้างภาพข้อมูลจะต้องคำนึงถึงระดับความรู้หรือความเข้าใจที่ผู้ฟังมีอยู่ วางแผนนำเสนอข้อมูลที่ผู้ฟังสามารถติดตามและทำความเข้าใจได้โดยง่าย และตรงกับความต้องการ หลีกเลี่ยงการใช้ศัพท์ทางเทคนิคหรืออักษรย่อที่ผู้ฟังไม่คุ้นเคย
3. เลือกใช้รูปแบบการแสดงผลภาพข้อมูลให้เหมาะสม ผู้สร้างภาพข้อมูลจะต้องพิจารณาเลือกรูปแบบแผนภาพให้เหมาะสมกับชนิดข้อมูล สามารถแสดงสิ่งที่ต้องการนำเสนอและนำไปสู่เป้าหมายได้ นั่นหมายความว่าผู้สร้างภาพข้อมูลจะต้องรู้คุณลักษณะของข้อมูลที่จะนำมาสร้างภาพเป็นอย่างไร รู้จักและเข้าใจการแสดงผลภาพข้อมูลรูปแบบต่างๆ เพื่อจะสามารถเลือกรูปแบบแผนภูมิได้อย่างเหมาะสม
4. สร้างภาพข้อมูลให้มีความชัดเจน สะอาด และเลือกใส่องค์ประกอบของแผนภูมิ เช่น ชื่อแกน และคำบรรยาย ให้เหมาะสม หลีกเลี่ยงคำบรรยายที่มากเกินไป ควรให้ภาพแสดงสิ่งที่ต้องการนำเสนอได้ด้วยตัวของมันเอง อีกทั้งนำเสนอข้อมูลได้อย่างตรงประเด็น ขจัดองค์ประกอบหรือการตกแต่งที่ไม่จำเป็นออก และเน้นเฉพาะในสิ่งที่ต้องการนำเสนอเท่านั้น
5. เลือกแบบอักษร ขนาดและสีขององค์ประกอบในแผนภาพให้เหมาะสม สามารถอ่านและสังเกตได้ง่าย มีความคมชัด และแตกต่างจากสีพื้นหลังที่วางอยู่ชัดเจน

6.2 การมองภาพข้อมูลสำหรับข้อมูลเชิงตัวเลข

ฮิสโทแกรม

ฮิสโทแกรม (Histogram) เป็นรูปแบบการนำเสนอข้อมูลเชิงตัวเลขสำหรับแสดงลักษณะการกระจายของค่าข้อมูล โดยการแบ่งข้อมูลเชิงตัวเลขออกเป็นช่วงๆ โดยแต่ละช่วงมีค่าที่ไม่ซ้อนทับกัน และมักจะใช้ขนาดของช่วงที่เท่ากัน แล้วจึงสร้างแท่งกราฟ โดย ความสูง ของแต่ละแท่งกราฟแสดงจำนวนข้อมูลในแต่ละช่วงข้อมูล หรือแสดงอัตราส่วนข้อมูลในแต่ละช่วงค่าข้อมูลต่อข้อมูลทั้งหมด มักให้แกนนอนแทนค่าข้อมูล ซึ่งเรียงลำดับจากค่าน้อยไปมาก หรือค่ามากไปน้อย และแกนตั้งแทนจำนวนข้อมูล ฮิสโทแกรมนี้ทำให้เห็นถึงลักษณะการแจกแจงของค่าข้อมูล ค่าข้อมูลส่วนใหญ่ ความแปรปรวนของค่าข้อมูล ลักษณะความเบ้และความโด่งของข้อมูล นอกจากนี้ยังทำให้เห็นค่าข้อมูลที่มาก



ภาพ 6.1: ฮิสโทแกรมแสดงลักษณะ การกระจายของค่าตัวแปรอายุในชุดข้อมูล Adult Census Income

ที่สุดและน้อยที่สุดอีกด้วย แสดงตัวอย่างฮิสโทแกรมของตัวแปรอายุจากชุดข้อมูล Adult Census Income [38] ซึ่งค่าตัวแปรอายุถูกแบ่งออกเป็น 15 ช่วงข้อมูล ดังภาพ 6.1

ตัวอย่าง 6.1

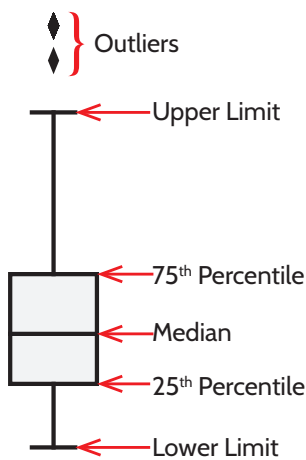
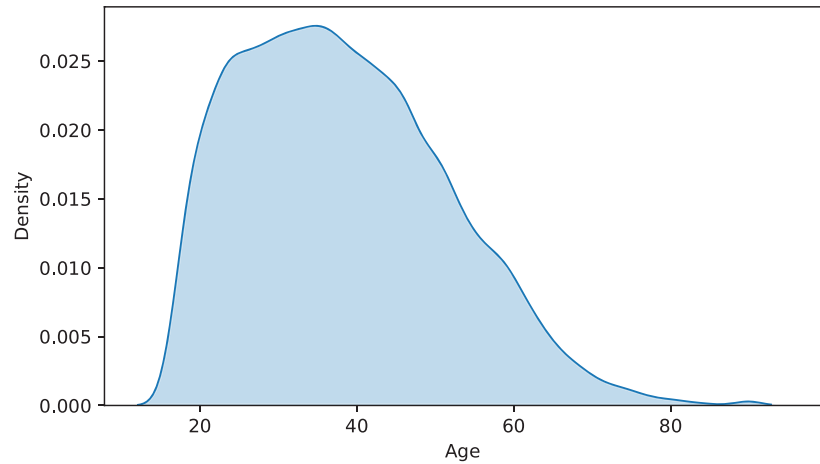
จากฮิสโทแกรมแสดงลักษณะการกระจายของค่าตัวแปรอายุในชุดข้อมูล Adult Census Income ดังภาพ 6.1 ทำให้ทราบว่า ตัวอย่างที่ได้จากการสำรวจส่วนใหญ่มีอายุอยู่ในช่วงประมาณ 32 - 36 ปี โดยตัวอย่างที่มีอายุน้อยที่สุดมีอายุ 17 ปี และตัวอย่างที่มีอายุมากที่สุดมีอายุ 90 ปี ค่าตัวแปรอายุมีลักษณะการแจกแจงแบบเบ้ขวา

[38] Ronny Kohavi and Barry Becker (n.d.). *Adult Census Income*. Accessed on 15 November 2020. url: <https://www.kaggle.com/c/adult-census-income>

แผนภาพความหนาแน่น

หากเรากำหนดช่วงข้อมูลในฮิสโทแกรมให้มีขนาดเล็กมาก และทำการปรับกราฟให้เรียบโดยใช้การประมาณค่าความหนาแน่น เราจะได้ **แผนภาพความหนาแน่น (Density Graph)** ซึ่งแท่งฮิสโทแกรมจะถูกปรับให้มีความต่อเนื่องเป็นเส้นโค้งความหนาแน่น แกนนอนแสดงค่าข้อมูล และแกนตั้งแสดงค่าความหนาแน่นของค่าข้อมูล ซึ่งเป็นอัตราส่วนของจำนวนข้อมูลในแต่ละค่าในแกนนอนต่อจำนวนข้อมูลทั้งหมด พื้นที่ใต้เส้นโค้งความหนาแน่นจะมีค่าเท่ากับ 1 เสมอ แผนภาพความหนาแน่นนี้ สามารถแสดงลักษณะการแจกแจงข้อมูลได้เช่นเดียวกับฮิสโทแกรม ภาพ 6.2 แสดงลักษณะการแจกแจงของข้อมูลอายุจากชุดข้อมูล Adult Census Income โดยใช้แผนภาพความหนาแน่น เมื่อเปรียบเทียบกับภาพ 6.1 จะสังเกตได้ว่าลักษณะการแจกแจงข้อมูลอายุจากแผนภาพทั้งสองมีความคล้ายคลึงกัน

ภาพ 6.2: แผนภาพ ความหนาแน่น แสดงลักษณะการกระจายของค่าตัวแปรอายุในชุดข้อมูล Adult Census Income



ภาพ 6.3: องค์ประกอบของแผนภาพกล่อง 1 กล่อง

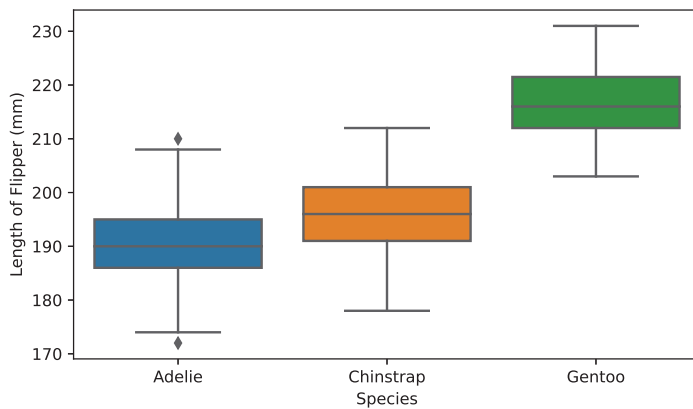
แผนภาพกล่อง

แผนภาพกล่อง (Boxplots) เป็นรูปแบบการแสดงลักษณะการกระจายของข้อมูลอีกรูปแบบหนึ่ง โดยนำเสนอค่าทางสถิติด้วยกล่อง 1 กล่อง ดังแสดงในภาพ 6.3 องค์ประกอบของกล่อง 1 กล่องซึ่งแสดงลักษณะการกระจายของค่าข้อมูล 1 ตัวแปร ประกอบด้วย ค่ามัธยฐาน ค่าเปอร์เซ็นต์ไทล์ที่ 25 และ 75 ค่าขีดจำกัดล่าง (Lower Limit) ค่าขีดจำกัดบน (Upper Limit) และค่าผิดปกติ นอกจากนี้ยังสามารถสังเกตเห็นความเบ้จากตำแหน่งค่ามัธยฐาน และค่าเปอร์เซ็นต์ไทล์ที่ 25 และ 75 ได้อีกด้วย โดยหากเส้นแสดงมัธยฐานโน้มเอียงไปใกล้เส้นแสดงค่าเปอร์เซ็นต์ไทล์ที่ 25 แสดงว่าค่าข้อมูลมีการแจกแจงแบบเบ้ขวา หากค่าเส้นแสดงมัธยฐานโน้มเอียงเข้าใกล้เส้นแสดงค่าเปอร์เซ็นต์ไทล์ที่ 75 แสดงว่าค่าข้อมูลมีการแจกแจงแบบเบ้ซ้าย และหากค่าเส้นแสดงมัธยฐานมีตำแหน่งอยู่กึ่งกลางระหว่างเส้นแสดงค่าเปอร์เซ็นต์ไทล์ที่ 25 และ 75 แสดงว่าค่าข้อมูลมีการแจกแจงปกติ ยิ่งไปกว่านั้น เรายังสามารถวางกล่องมากกว่า 1 กล่องบนแผนภาพเดียวกัน เพื่อเปรียบเทียบลักษณะการกระจายของค่าตัวแปรหนึ่งในแต่ละกลุ่มข้อมูลย่อยได้อีกด้วย ภาพ 6.4 แสดงแผนภาพกล่องเปรียบเทียบความยาวครีบทองของเพนกวิน 3 ชนิดพันธุ์ ได้แก่ Adelie Chinstrap และ Gentoo จากชุดข้อมูล palmerpenguins [39]

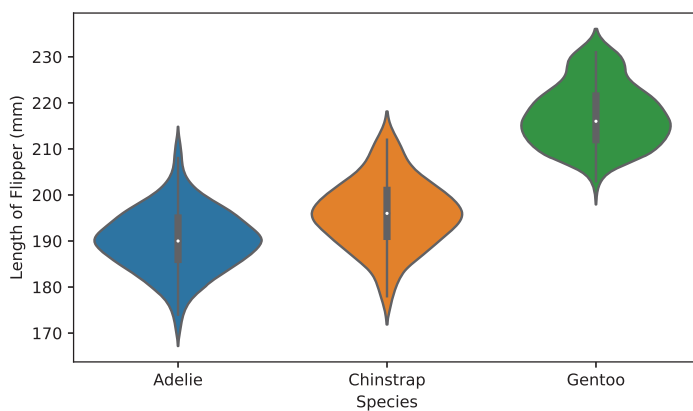
[39] Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman (2020). *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. R package version 0.1.0. url: <https://allisonhorst.github.io/palmerpenguins/>

ตัวอย่าง 6.2

จากแผนภาพกล่องเปรียบเทียบความยาวครีบทองของเพนกวิน 3 ชนิดพันธุ์ ได้แก่ Adelie Chinstrap และ Gentoo ในชุดข้อมูล palmerpenguins ดังภาพ 6.4 ทำให้ทราบว่า เพนกวินสายพันธุ์ Adelie ส่วนใหญ่มีความยาวครีบทองประมาณ 190 มิลลิเมตร และมีเพนกวินบางตัวมีความยาวครีบทองต่างจากกลุ่มอย่างมาก ในขณะที่เพนกวินสายพันธุ์ Chinstrap ส่วนมากมีความยาวครีบทองประมาณ 197 มิลลิเมตร และเพนกวินสายพันธุ์ Gentoo ส่วนใหญ่มีความยาวครีบทองโดยประมาณ 214 มิลลิเมตร เมื่อเปรียบเทียบความยาวครีบทอง



ภาพ 6.4: แผนภาพกล่องเปรียบเทียบความยาว ครีบ ของ เพนกวิน 3 ชนิด พันธุ์ ได้แก่ Adelie Chinstrap และ Gentoo จาก ชุดข้อมูล palmerpenguins



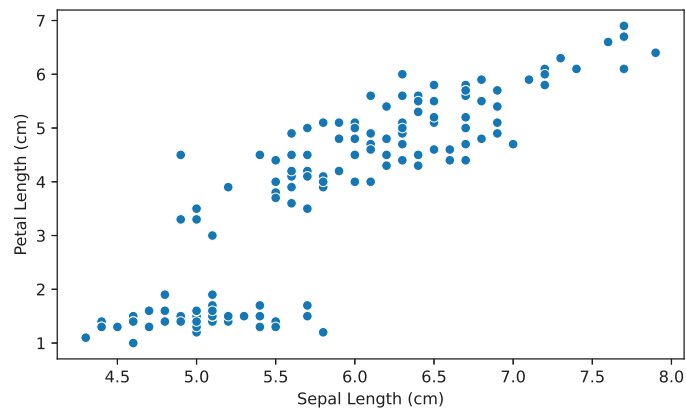
ภาพ 6.5: แผนภาพรูปไวโอลินเปรียบเทียบความยาวครีบของเพนกวิน 3 ชนิดพันธุ์ ได้แก่ Adelie Chinstrap และ Gentoo จาก ชุดข้อมูล palmerpenguins

เพนกวินสายพันธุ์ Adelie และ Chinstrap พบว่าเพนกวินสายพันธุ์ Chinstrap มีครีวยาวกว่า สายพันธุ์ Adelie เพียงเล็กน้อย ในขณะที่เพนกวินสายพันธุ์ Gentoo มีครีบที่ยาวกว่าเพนกวินสายพันธุ์อื่นมาก

แผนภาพรูปไวโอลิน

แผนภาพรูปไวโอลิน (Violin Plot) เป็นแผนภาพที่เพิ่มเติมแผนภาพกล่องเพื่อให้ข้อมูลลักษณะการกระจายของค่าตัวแปรที่มากขึ้น โดยแทนที่กล่องด้วยโค้งความหนาแน่นทั้งฝั่งซ้ายและฝั่งขวาแบบสะท้อนกัน ทำให้เกิดลักษณะคล้ายรูปไวโอลิน และลดการแสดงค่าผิดปกติ ซึ่งถูกแทนด้วยจุดในแผนภาพกล่อง แสดงตัวอย่างแผนภาพไวโอลิน ดังภาพ 6.5 ซึ่งเปรียบเทียบความยาวครีบของเพนกวิน 3 ชนิดพันธุ์ สามารถอธิบายได้เช่นเดียวกับแผนภาพกล่องในภาพ 6.4 และมองเห็นลักษณะการกระจายของค่าความยาวครีบของเพนกวินแต่ละชนิดได้ครบถ้วน

ภาพ 6.6: แผนภาพ การกระจาย ของ ค่าตัวแปรความยาวของกลีบเลี้ยง (Sepal Length) และความยาวของกลีบดอก (Petal Length) ของดอกไอริส ในชุดข้อมูล Fisher Iris



แผนภาพการกระจาย

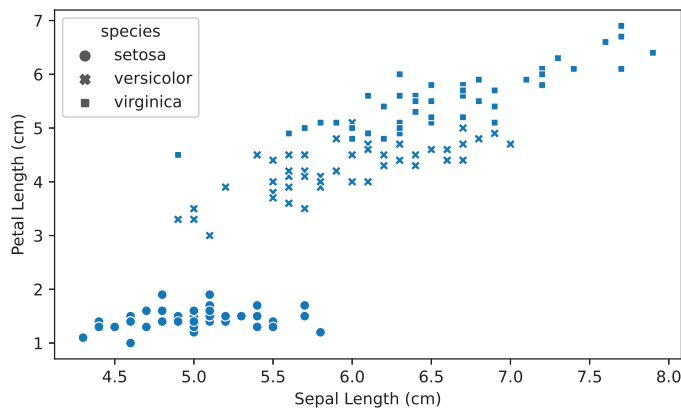
แผนภาพการกระจาย (Scatterplot) เป็นวิธีการมองภาพข้อมูลที่ถูกนำมาใช้อย่างกว้างขวางในการวิเคราะห์ข้อมูล โดยเฉพาะ การสังเกตความสัมพันธ์ระหว่างตัวแปร 2 ตัวแปร เป็นการวางค่าของคู่ตัวแปรของจุดข้อมูลแต่ละข้อมูลลงบนระนาบ 2 มิติ โดยให้แกนหนึ่งแทนค่าตัวแปรหนึ่งและอีกแกนหนึ่งแทนค่าของอีกตัวแปรหนึ่ง ทั้งนี้สามารถขยายได้ถึง 3 มิติ โดยการเพิ่มแกนอีก 1 แกนสำหรับค่าของอีกตัวแปร แต่มีข้อจำกัดของการแสดงผลเมื่อไม่สามารถหมุนแกนเพื่อดูข้อมูลแต่ละด้านได้ แสดงตัวอย่างแผนภาพการกระจายของตัวแปรความยาวของกลีบเลี้ยง (Sepal Length) และความยาวของกลีบดอก (Petal Length) ของดอกไอริส ในชุดข้อมูล Fisher Iris [7] ในภาพ 6.6

[7] Penny Analytics (n.d.). *Fisher Iris Data*. Accessed on 5 October 2019. url: <https://pennyanalytics.com/free-trial/>

ตัวอย่าง 6.3

จากแผนภาพการกระจายของค่าตัวแปรความยาวของกลีบเลี้ยง และความยาวของกลีบดอก ของดอกไอริส ดังภาพ 6.6 สังเกตได้ว่าเมื่อพิจารณาความยาวของกลีบเลี้ยงและความยาวของกลีบดอก สามารถมองเห็นกลุ่มของดอกไอริสต่างกันชัดเจน 2 กลุ่ม โดยกลุ่มแรก (บริเวณมุมซ้ายล่างของแผนภาพ) ความยาวของกลีบเลี้ยงและกลีบดอกไม่มีความสัมพันธ์กัน และกลุ่มของดอกไอริสอีกกลุ่มหนึ่ง (บริเวณกลางจนกระทั่งมุมขวาบนของแผนภาพ) ความยาวของกลีบเลี้ยงและกลีบดอกค่อนข้างมีความสัมพันธ์กัน โดยดอกไอริสที่มีกลีบเลี้ยงยาวมีแนวโน้มที่จะมีกลีบดอกยาวด้วย ในขณะที่ดอกที่มีกลีบเลี้ยงสั้นมีแนวโน้มที่จะมีกลีบดอกสั้นด้วยเช่นกัน

นอกจากนี้เรายังสามารถแสดงค่าตัวแปรเชิงกลุ่มของข้อมูลบนแผนภาพการกระจายได้ โดยเลือกใช้สัญลักษณ์หรือสีของจุดที่แตกต่างกันอย่างชัดเจนสำหรับแต่ละค่าตัวแปรเชิงกลุ่มที่สนใจ ภาพ 6.7 แสดงแผนภาพการกระจายของตัวแปรความยาวของกลีบเลี้ยง (Sepal Length) และความยาวของกลีบ



ภาพ 6.7: แผนภาพ การกระจาย ของ ค่าตัวแปรความยาวของกลีบเลี้ยง (Sepal Length) และความยาวของกลีบดอก (Petal Length) ของดอกไอริส 3 ชนิดพันธุ์ ได้แก่ setosa versicolor และ virginica

ดอก (Petal Length) ของดอกไอริส 3 ชนิดพันธุ์ ได้แก่ setosa versicolor และ virginica โดยใช้สัญลักษณ์จุดที่แตกต่างกัน 3 ลักษณะ สำหรับแทนค่าแต่ละชนิดพันธุ์

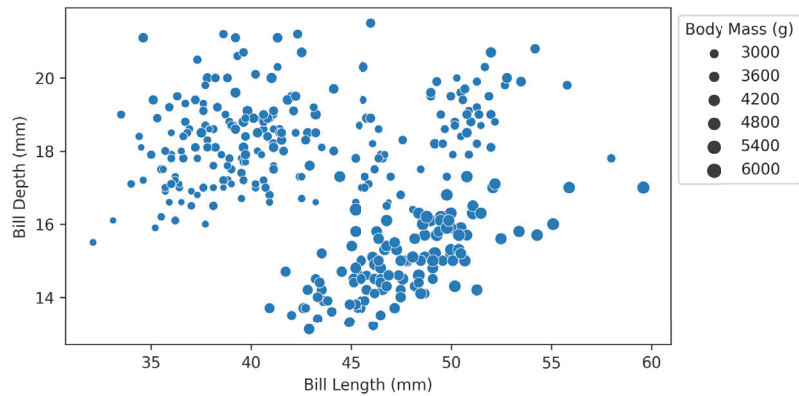
ตัวอย่าง 6.4

จากแผนภาพการกระจายของค่าตัวแปรความยาวของกลีบเลี้ยง และความยาวของกลีบดอก ของดอกไอริส 3 ชนิดพันธุ์ ดังภาพ 6.7 พบว่า ความยาวของกลีบเลี้ยงและกลีบดอกของดอกไอริสสายพันธุ์ setosa ไม่มีความสัมพันธ์กัน ในขณะที่ความยาวของกลีบเลี้ยงและกลีบดอกของสายพันธุ์ versicolor และ virginica ค่อนข้างมีความสัมพันธ์กัน

นอกจากนี้ยังสังเกตได้ว่าความยาวของกลีบดอกของสายพันธุ์ setosa แตกต่างอย่างชัดเจนจากสายพันธุ์อื่น ทำให้สามารถแบ่งแยกดอกไอริสสายพันธุ์ setosa จากสายพันธุ์อื่นได้อย่างง่ายโดยพิจารณาความยาวของกลีบดอก ส่วนดอกไอริสสายพันธุ์ versicolor และ virginica นั้น ความยาวของกลีบดอกไม่แตกต่างกันแบ่งแยกกันไม่ชัดเจน อย่างไรก็ตามดอกไอริสสายพันธุ์ versicolor มีความยาวของกลีบดอกค่อนข้างสั้นกว่ากลีบดอกของสายพันธุ์ virginica

เมื่อพิจารณาความยาวของกลีบเลี้ยงของดอกไอริสทั้ง 3 ชนิดพันธุ์ พบว่า ไม่มีความแตกต่างกันอย่างชัดเจน อย่างไรก็ตามความยาวของกลีบเลี้ยงของสายพันธุ์ setosa มีแนวโน้มสั้นกว่าสายพันธุ์อื่นๆ และความยาวของกลีบเลี้ยงของสายพันธุ์ virginica ค่อนข้างยาวกว่ากลีบเลี้ยงของสายพันธุ์อื่นๆ

ภาพ 6.8: แผนภาพฟองของค่าตัวแปรความยาวของปาก (Bill Length) ความสูงของปาก (Bill Depth) และ น้ำหนักตัว (Body Mass) ของเพนกวิน ทั้ง 3 ชนิด พันธุ์ในชุดข้อมูล palmerpenguins

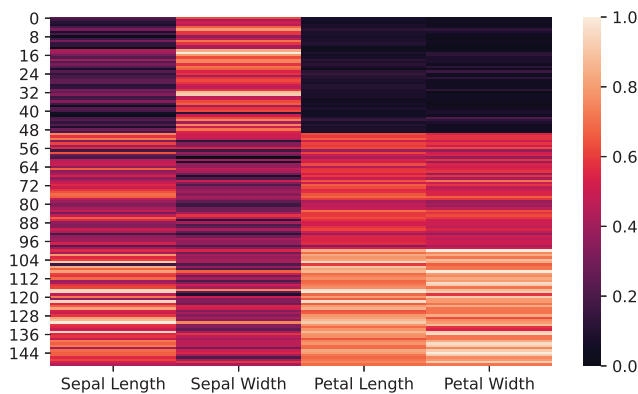


แผนภาพฟอง

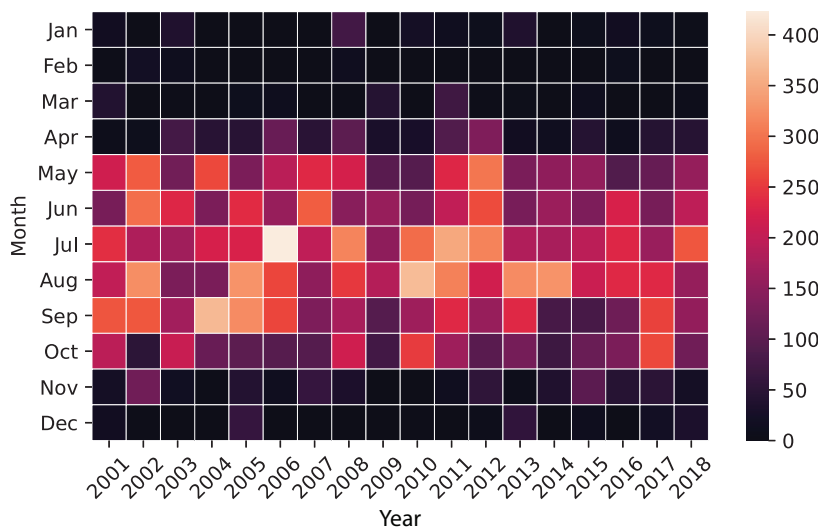
แผนภาพฟอง (Bubble Plot) เป็นการนำแผนภาพการกระจายมาเพิ่มเติมสารสนเทศ โดยนำเสนอค่าตัวแปรเชิงตัวเลขอีกตัวแปรหนึ่งด้วยขนาดของจุด ข้อมูลบนแผนภาพการกระจาย แม้ว่าแผนภาพฟองจะสามารถแสดงค่าตัวแปรเชิงปริมาณได้จำนวนมากขึ้น ด้วยมาตราส่วน ตำแหน่ง และขนาด แต่ก็ทำให้ยากต่อการตรวจสอบความสัมพันธ์ระหว่างตัวแปรต่างๆ ด้วยสายตา อีกทั้งการใช้ขนาดของจุดข้อมูลในการแทนค่าตัวแปรที่ค่ามากที่สุดและค่าน้อยที่สุดมีความแตกต่างกันอย่างมาก ทำให้การสังเกตความแตกต่างระหว่างค่าข้อมูลทำได้ยาก ภาพ 6.8 แสดงตัวอย่างแผนภาพฟองของค่าตัวแปรความยาวของปาก (Bill Length) ความสูงของปาก (Bill Depth) และน้ำหนักตัว (Body Mass) ของเพนกวินทั้ง 3 ชนิดพันธุ์ในชุดข้อมูล palmerpenguins [39] โดย ค่าตัวแปรความยาวและความสูงของปากแสดงโดยใช้ตำแหน่งของจุดข้อมูลบนระนาบที่มีแกนตั้งแทนค่าตัวแปรความสูงของปาก และแกนนอนแทนค่าตัวแปรความยาวของปาก ค่าตัวแปรน้ำหนักตัวถูกแทนด้วยขนาดของจุดข้อมูล

ฮีตแมป

ฮีตแมป (Heatmap) เป็นรูปแบบการแสดงความหนาแน่นของค่าตัวแปรเชิงปริมาณในเมทริกซ์ 2 มิติหรือตารางข้อมูลด้วยระดับสีหรือความเข้มของแสงที่สามารถสังเกตความแตกต่างได้ เหมาะสำหรับการแสดงมุมมองทั่วไปของตัวแปรเชิงตัวเลขโดยไม่ชี้เฉพาะจุดข้อมูล เนื่องจากฮีตแมปใช้ระดับสีหรือความเข้มของแสงเพียง 1 มาตราส่วน ดังนั้น ข้อมูลที่นำมาแสดงควรต้องอยู่ในช่วงข้อมูลเดียวกันทั้งหมด หรืออาจทำให้เป็นบรรทัดฐานเดียวกันก่อน ในกรณีที่เป็นการนำเสนอค่าตัวแปรหลายตัวแปรที่มีช่วงข้อมูลต่างกัน แสดงตัวอย่างการแสดงผลข้อมูลความยาวของกลีบเลี้ยง (Sepal Length) ความกว้างของกลีบเลี้ยง (Sepal Width) ความยาวของกลีบดอก (Petal Length) และความกว้างของกลีบดอก (Petal Width) ของดอกไอริสจากชุดข้อมูล Fisher Iris [7] ใน



ภาพ 6.9: ฮีตแมปแสดงข้อมูลความยาวของกลีบเลี้ยง (Sepal Length) ความกว้างของกลีบเลี้ยง (Sepal Width) ความยาวของกลีบดอก (Petal Length) และความกว้างของกลีบดอก (Petal Width) ของชุดข้อมูล Fisher Iris



ภาพ 6.10: ฮีตแมปแสดง ปริมาณ น้ำ ฝน สะสม ราย เดือน ระหว่าง ปี ค.ศ. 2001 ถึง 2018 ตรวจวัดที่สถานีตรวจวัดอากาศจังหวัดแม่ฮ่องสอน โดยกรมอุตุนิยมวิทยา

ภาพ 6.9 โดยค่าตัวแปรทั้ง 4 ตัวแปรถูกทำให้เป็นบรรทัดฐานก่อนนำมาแสดงด้วยฮีตแมป และภาพ 6.10 แสดงตัวอย่างฮีตแมปของปริมาณน้ำฝนสะสมรายเดือนระหว่างปี ค.ศ. 2001 ถึง 2018 ตรวจวัดที่สถานีตรวจวัดอากาศจังหวัดแม่ฮ่องสอน โดยกรมอุตุนิยมวิทยา

ตัวอย่าง 6.5

จากฮีตแมปในภาพ 6.9 ซึ่งแสดงค่าตัวแปรความยาวของกลีบเลี้ยง (Sepal Length) ความกว้างของกลีบเลี้ยง (Sepal Width) ความยาวของกลีบดอก (Petal Length) และความกว้างของกลีบดอก (Petal Width) ของดอกไอริสทั้งหมดจากชุดข้อมูล Fisher Iris เมื่อพิจารณาข้อมูลลำดับที่ 0 ถึง 49 พบว่า มีค่าตัวแปรความยาวและความกว้างของกลีบดอกน้อยกว่าข้อมูลในลำดับที่ 50 ถึง 149 อย่างชัดเจน อีกทั้งยังมีค่าตัวแปรความยาวของกลีบเลี้ยงค่อนข้างน้อย แต่มีค่าตัวแปรความกว้างของกลีบเลี้ยงค่อนข้างมากกว่าข้อมูลในลำดับอื่นๆ

สำหรับข้อมูลในลำดับที่ 50 ถึง 99 โดยส่วนใหญ่ ค่าตัวแปรความยาวและความกว้างของกลีบดอก มีค่าค่อนข้างน้อยกว่าข้อมูลในลำดับที่ 100 ถึง 149 ส่วนค่าตัวแปรความยาวของกลีบเลี้ยงยังสามารถสังเกตเห็นความแตกต่างเล็กน้อยเมื่อเทียบกับข้อมูลในลำดับที่ 100 ถึง 149 แต่ไม่สามารถแยกความแตกต่างของค่าความกว้างของกลีบเลี้ยงระหว่างข้อมูลลำดับที่ 50 ถึง 99 และข้อมูลลำดับที่ 100 ถึง 149 ได้

จากการสังเกตฮิสโตแกรมโดยรวม เราสามารถจัดกลุ่มข้อมูลได้เป็น 3 กลุ่มใหญ่ เมื่อพิจารณาตัวแปรทั้งหมด 4 ตัวแปร คือ กลุ่มที่ 1 ประกอบด้วยข้อมูลแถวที่ 0 ถึง 49 กลุ่มที่ 2 ประกอบด้วยข้อมูลแถวที่ 50 ถึง 99 และกลุ่มที่ 3 เป็นกลุ่มของข้อมูลแถวที่เหลือทั้งหมด

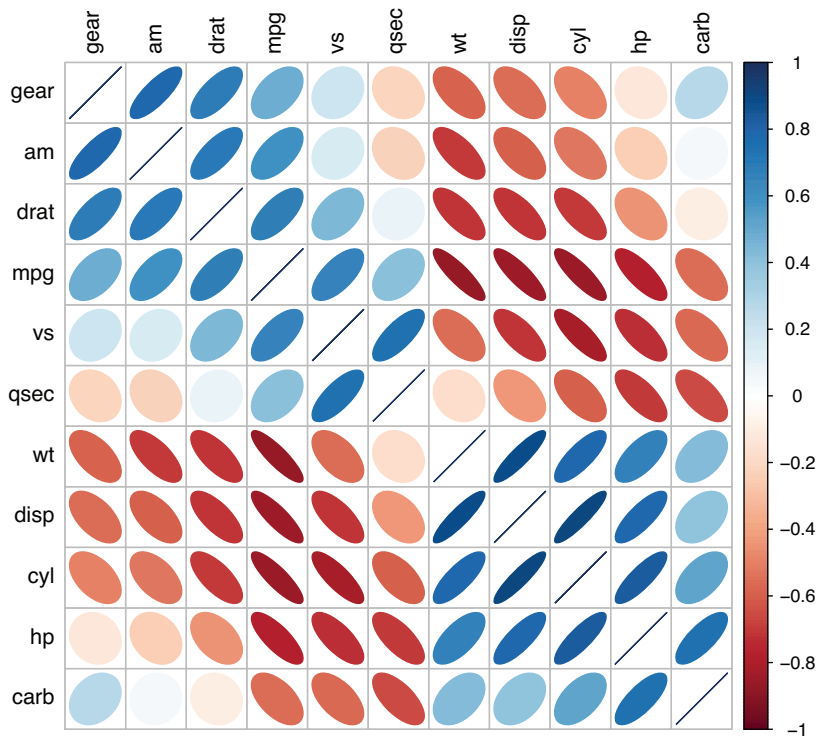
ตัวอย่าง 6.6

จากข้อมูลปริมาณน้ำฝนสะสมรายเดือน (หน่วย: มิลลิเมตร) ระหว่างปี ค.ศ. 2001 ถึง 2018 ตรวจวัดจากสถานีตรวจวัดอากาศจังหวัดแม่ฮ่องสอน โดยกรมอุตุนิยมวิทยา ดังแสดงด้วยฮิสโตแกรมในภาพ 6.10 ทำให้เห็นว่าโดยภาพรวมแล้วฝนจะตกในบริเวณดังกล่าวในช่วงเดือนพฤษภาคมถึงเดือนตุลาคม และฝนตกในปริมาณมากในช่วงเดือนกรกฎาคมถึงเดือนสิงหาคม โดยในปี ค.ศ. 2016 มีปริมาณน้ำฝนสะสมมากกว่า 400 มิลลิเมตร ในเดือนกรกฎาคม ปี ค.ศ. 2006 นอกจากนี้ในปี ค.ศ. 2005 2008 และ 2013 มีฝนตกในช่วงเดือนธันวาคมและมกราคม ซึ่งเป็นช่วงกลางฤดูหนาว ในปริมาณสะสมเกินกว่า 50 มิลลิเมตร อย่างไรก็ตามในช่วงปี ค.ศ. 2015 ถึง 2018 ปริมาณน้ำฝนสะสมรายเดือนมากที่สุด ในพื้นที่จังหวัดแม่ฮ่องสอนลดลงจากช่วง 6 ปีก่อนหน้า (ค.ศ. 2010 ถึง 2014) อย่างเห็นได้ชัด

แผนภาพสหสัมพันธ์

แผนภาพสหสัมพันธ์ (Correlogram) เป็นการแสดงเมทริกซ์สหสัมพันธ์ ซึ่งใช้ประโยชน์ในการวิเคราะห์ความสัมพันธ์ของแต่ละคู่ตัวแปรเชิงปริมาณ (ดังกล่าวรายละเอียดมาแล้วในบทที่ 3) โดยการแสดงค่าสหสัมพันธ์ระหว่างคู่ตัวแปรเชิงปริมาณด้วยระดับสี ขนาดหรือรูปร่างของสัญลักษณ์ในเมทริกซ์สหสัมพันธ์ แทนการแสดงด้วยค่าสหสัมพันธ์โดยตรง ภาพ 6.11 แสดงแผนภาพสหสัมพันธ์ของคุณสมบัติ 11 คุณสมบัติของรถยนต์ 32 รุ่นในชุดข้อมูล Gasoline Mileage [40] ซึ่งรวบรวมได้จากวารสาร 1974 Motor Trend US magazine โดยขนาดของค่าสหสัมพันธ์ถูกแสดงด้วยสีและความรีของรูปวงรี และทิศทางของค่าสหสัมพันธ์ถูกแทนด้วยทิศทางการเอียงของรูปวงรี

[40] Harold V. Henderson and Paul F. Velleman (1981). "Building Multiple Regression Models Interactively." In: *Biometrics* 37.2



ภาพ 6.11: แผนภาพสหสัมพันธ์ของ
คุณสมบัติ 11 คุณสมบัติ ของ รถยนต์ 32
รุ่น จาก วารสาร 1974 Motor Trend US
magazine

6.3 การมองภาพข้อมูลสำหรับข้อมูลเชิงข้อมูลเชิงกลุ่ม

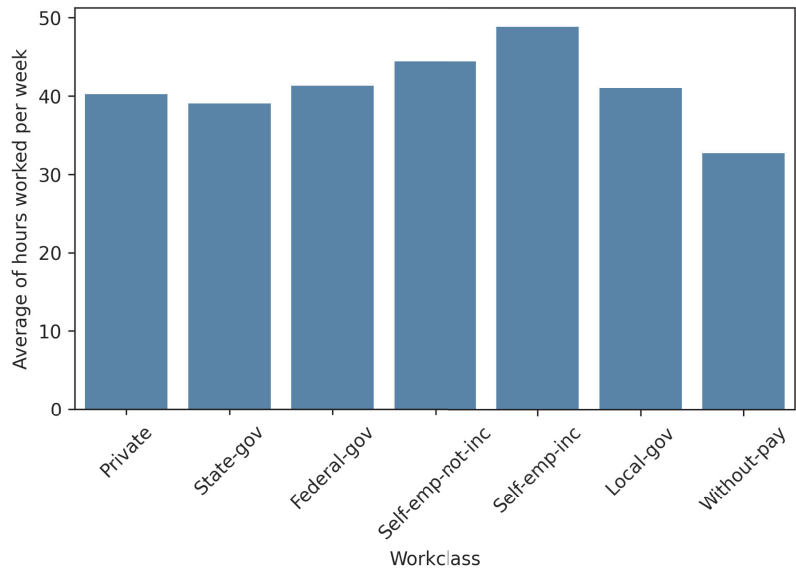
แผนภูมิแท่ง

แผนภูมิแท่ง (Bar Chart) เป็นรูปแบบการนำเสนอข้อมูลที่เป็นที่รู้จักอย่างแพร่หลาย ซึ่งแสดงข้อมูลเชิงกลุ่มด้วยแท่งสี่เหลี่ยมที่มีความสูงหรือความยาวตามสัดส่วนกับค่าเชิงปริมาณที่นำเสนอสำหรับแต่ละกลุ่มข้อมูล มักถูกใช้ประโยชน์ในการเปรียบเทียบค่าเชิงปริมาณของแต่ละกลุ่มข้อมูล แท่งกราฟสามารถถูกวางในแนวนตั้งหรือแนวนอนก็ได้ตามความเหมาะสม กรณีที่แผนภูมิแท่งถูกวางในแนวนตั้ง บางครั้งเรียกว่า แผนภูมิคอลัมน์ (Column Chart) ภาพ 6.12 แสดงตัวอย่างแผนภูมิแท่งเปรียบเทียบจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ของแต่ละกลุ่มอาชีพในชุดข้อมูล Adult Census Income

ตัวอย่าง 6.7

จากแผนภูมิแท่งเปรียบเทียบจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ของ 7 กลุ่มอาชีพในชุดข้อมูล Adult Census Income แสดงดังภาพ 6.12 จะเห็นได้ว่าชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ในแต่ละกลุ่มอาชีพมีการกระจายอยู่ในช่วง 30 - 50 ชั่วโมง โดยกลุ่มอาชีพอิสระในรูปแบบบริษัทมีชั่วโมงทำงานเฉลี่ยต่อสัปดาห์มากที่สุดอย่างเห็นได้ชัด รองลงมาได้แก่กลุ่มอาชีพอิสระ

ภาพ 6.12: แผนภูมิแท่งเปรียบเทียบจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ของ 7 กลุ่มอาชีพในชุดข้อมูล Adult Census Income ได้แก่ ธุรกิจส่วนตัว (Private) เจ้าหน้าที่รัฐบาลท้องถิ่น (State-gov) เจ้าหน้าที่รัฐบาลกลาง (Federal-gov) อาชีพอิสระที่ไม่ใช่บริษัท (Self-emp-not-inc) อาชีพอิสระในรูปแบบบริษัท (Self-emp-inc) เจ้าหน้าที่องค์กรท้องถิ่น (Local-gov) และอาชีพที่ปราศจากรายได้ (Without-pay)

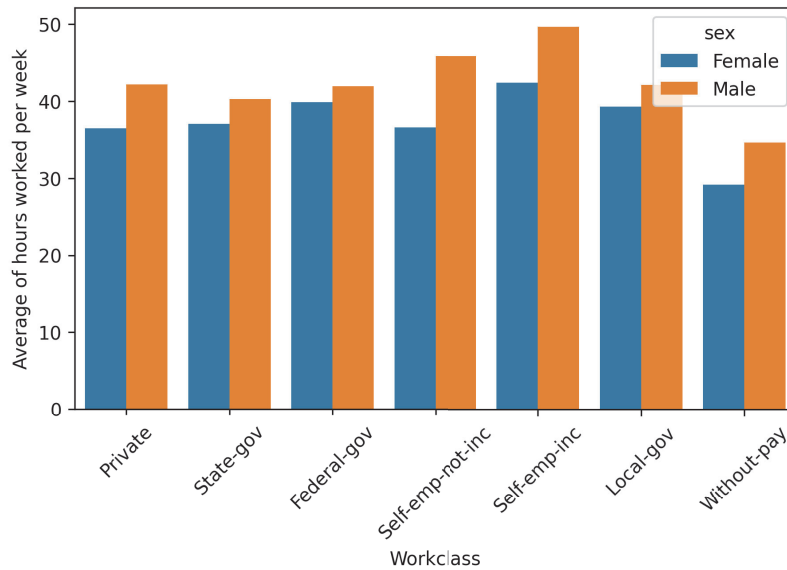


ที่ไม่ใช่บริษัท ในขณะที่คนในกลุ่มอาชีพที่ปราศจากรายได้มีจำนวนชั่วโมงทำงานต่อสัปดาห์น้อยที่สุด ส่วนกลุ่มอาชีพอื่นๆ ได้แก่ ธุรกิจส่วนตัว เจ้าหน้าที่รัฐบาลท้องถิ่น เจ้าหน้าที่รัฐบาลกลาง และเจ้าหน้าที่องค์กรท้องถิ่น นั้น มีชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ไม่ต่างกันมากนัก

นอกจากนี้ แผนภูมิแท่งยังสามารถนำเสนอข้อมูลเชิงกลุ่มในระดับที่ 2 ซึ่งเป็นกลุ่มย่อยในแต่ละกลุ่มข้อมูลหลัก โดยการแยกแท่งกราฟในแต่ละกลุ่มข้อมูลหลักให้เป็นแท่งย่อยๆ แท่งกราฟแต่ละแท่งแทนกลุ่มย่อยแต่ละกลุ่ม และจัดวางแท่งกราฟเป็นชุดๆ ตามกลุ่มหลัก ทั้งนี้สามารถใช้สีหรือลวดลายที่แตกต่างกันสำหรับกลุ่มย่อยแต่ละกลุ่มได้ โดยกลุ่มย่อยประเภทเดียวกันใช้สีของแท่งกราฟเหมือนกัน แสดงตัวอย่างการเปรียบเทียบจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ระหว่างเพศชายและหญิงในแต่ละกลุ่มอาชีพจากชุดข้อมูล Adult Census Income ดังภาพ 6.13 โดยค่าตัวแปรกลุ่มอาชีพ (Work Class) ถือเป็นข้อมูลเชิงกลุ่มหลัก และค่าตัวแปรเพศ ถือเป็นข้อมูลเชิงกลุ่มย่อย หรือ ข้อมูลเชิงกลุ่มรอง แผนภูมิแท่งลักษณะนี้ เรียกว่า **แผนภูมิแท่งแบบกลุ่ม (Clustered Bar Chart)**

ตัวอย่าง 6.8

จากแผนภูมิแท่งเปรียบเทียบจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ระหว่างเพศชายและหญิงในแต่ละกลุ่มอาชีพในจากข้อมูล Adult Census Income ในภาพ 6.13 พบว่า เพศชายในทุกกลุ่มอาชีพมีจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์มากกว่าเพศหญิง โดยกลุ่มอาชีพอิสระที่ไม่ใช่บริษัทมีความแตกต่างของชั่วโมงทำงานต่อสัปดาห์ของเพศชายและหญิงมากที่สุด ในขณะที่กลุ่มเจ้าหน้าที่รัฐบาลกลางมีความแตกต่างของชั่วโมงทำงานต่อสัปดาห์ของเพศ



ภาพ 6.13: แผนภูมิแท่งเปรียบเทียบจำนวนชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ระหว่างเพศชายและหญิงใน 7 กลุ่มอาชีพจากชุดข้อมูล Adult Census Income ได้แก่ ธุรกิจส่วนตัว (Private) เจ้าหน้าที่รัฐบาลท้องถิ่น (State-gov) เจ้าหน้าที่รัฐบาลกลาง (Federal-gov) อาชีพอิสระที่ไม่ใช่บริษัท (Self-emp-not-inc) อาชีพอิสระในรูปแบบบริษัท (Self-emp-inc) เจ้าหน้าที่องค์กรท้องถิ่น (Local-gov) และอาชีพที่ปราศจากรายได้ (Without-pay) ในจากข้อมูล Adult Census Income

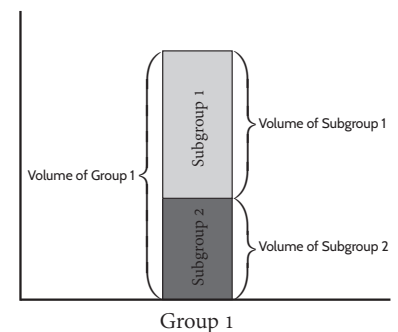
ชายและหญิงน้อยที่สุด

เมื่อเปรียบเทียบชั่วโมงทำงานเฉลี่ยต่อสัปดาห์ของเพศชายและหญิงระหว่างกลุ่มอาชีพ พบว่า ทั้งเพศชายและหญิงในกลุ่มอาชีพอิสระในรูปแบบบริษัทมีชั่วโมงทำงานมากที่สุด ในขณะที่กลุ่มอาชีพที่ปราศจากรายได้มีชั่วโมงทำงานน้อยที่สุดทั้งเพศชายและหญิง

อย่างไรก็ตามแผนภูมิแท่งแบบกลุ่มทำให้สูญเสียการนำเสนอค่าปริมาณรวมของกลุ่มหลัก แผนภูมิแท่งแบบวางซ้อน (Stacked Bar Chart) จึงถูกนำเสนอขึ้นโดยการนำแท่งกราฟสำหรับกลุ่มย่อยในกลุ่มหลักเดียวกันวางซ้อนกันและให้สีแท่งกราฟที่แตกต่างกันสำหรับแต่ละกลุ่มย่อย ความสูงรวมของแท่งกราฟที่วางซ้อนกันในกลุ่มหลักนั้น แทนค่าเชิงปริมาณโดยรวมสำหรับกลุ่มหลัก ในขณะที่ความสูงของแต่ละแท่งกราฟย่อยซึ่งแทนค่าเชิงปริมาณสำหรับแต่ละกลุ่มย่อยยังสามารถสังเกตได้ด้วยสีที่แตกต่างกัน แสดงสัญญาณของแท่งกราฟบนแผนภูมิแท่งแบบวางซ้อนดังภาพ 6.14 ภาพ 6.15 เปรียบเทียบรายได้จากค่าอาหารรวมของร้านอาหารแห่งหนึ่งในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ จากชุดข้อมูล Tips [41] โดยในแต่ละวันยังสามารถสังเกตเห็นค่าอาหารรวมที่ได้รับจากผู้ให้บริการเพศชายและหญิงด้วย

ตัวอย่าง 6.9

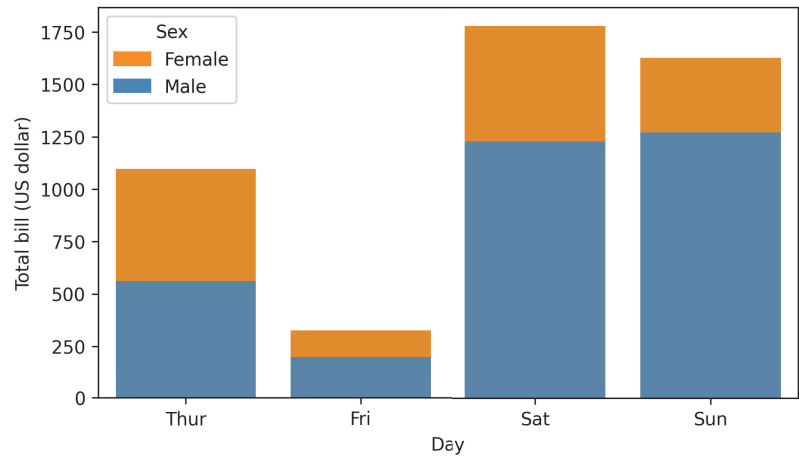
จากแผนภูมิแท่งแบบวางซ้อนเปรียบเทียบรายได้รวมจากค่าอาหารของร้านอาหารแห่งหนึ่งในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ จากชุดข้อมูล Tips ในภาพ 6.15 สังเกตได้ว่าร้านอาหารมีรายได้จากค่าอาหารมากที่สุดในวันเสาร์ สูงเกือบ 1,750 เหรียญดอลลาร์สหรัฐ และมีรายได้จากค่าอาหารน้อย



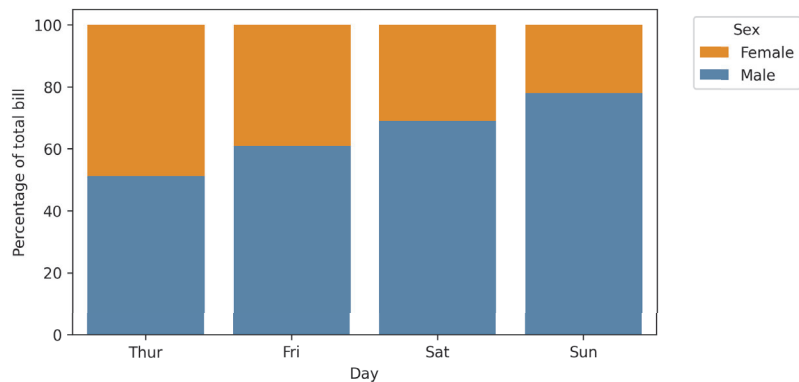
ภาพ 6.14: สัญญาณของแท่งกราฟบนแผนภูมิแท่งแบบวางซ้อน

[41] Peter G. Bryant and Marlene A. Smith (1994). *Practical Data Analysis: Case Studies in Business Statistics*. McGraw-Hill Professional. isbn: 0256158282

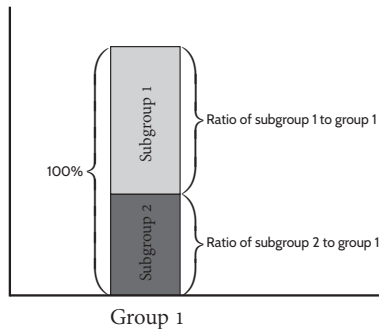
ภาพ 6.15: แผนภูมิแท่งแบบวางซ้อนเปรียบเทียบรายได้รวมจากค่าอาหารของร้านอาหารแห่งหนึ่ง จากลูกค้าเพศชายและหญิง ในวันพฤหัสบดี ศุกร์ เสาร์ และ อาทิตย์ จากชุดข้อมูล Tips



ภาพ 6.17: แผนภูมิแท่งร้อยละแบบวางซ้อนเปรียบเทียบสัดส่วนรายได้จากค่าอาหารของผู้ใช้บริการเพศชายและหญิง ในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ จากชุดข้อมูล Tips



ที่สุดในวันศุกร์ มูลค่าประมาณ 300 ถึง 350 เหรียญดอลลาร์สหรัฐ เมื่อพิจารณารายได้จากค่าอาหารในแต่ละวันแล้ว พบว่า ร้านอาหารมีรายได้จากลูกค้าเพศชายมากกว่าเพศหญิง



ภาพ 6.16: สันฐาน ของแท่งกราฟบนแผนภูมิแท่งร้อยละแบบวางซ้อน

หากต้องการเปรียบเทียบสัดส่วนค่าเชิงปริมาณของกลุ่มย่อยในแต่ละกลุ่มหลัก ยังสามารถทำได้ด้วยการสร้าง **แผนภูมิแท่งร้อยละแบบวางซ้อน (100% Stacked Bar Chart)** ซึ่งเป็นการนำแผนภูมิแท่งแบบวางซ้อนมาขยายให้ความสูงรวมของแท่งกราฟในแต่ละกลุ่มหลัก สูงเท่าๆ กัน โดยคิดเป็นสัดส่วน 100 เปอร์เซนต์ และแต่ละแท่งกราฟสำหรับกลุ่มย่อยถูกขยายให้สูงตามสัดส่วนค่าเชิงปริมาณต่อค่าเชิงปริมาณรวมของกลุ่มหลัก แสดงดังรูป 6.16 วิธีการนำเสนอรูปแบบนี้ ทำให้เห็นสัดส่วนของค่าเชิงปริมาณของแต่ละกลุ่มย่อยในกลุ่มหลักได้ แสดงตัวอย่างแผนภูมิแท่งร้อยละแบบวางซ้อนเปรียบเทียบสัดส่วนรายได้จากค่าอาหารของผู้ใช้บริการเพศชายและหญิง ในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ ดังภาพ 6.17

ตัวอย่าง 6.10

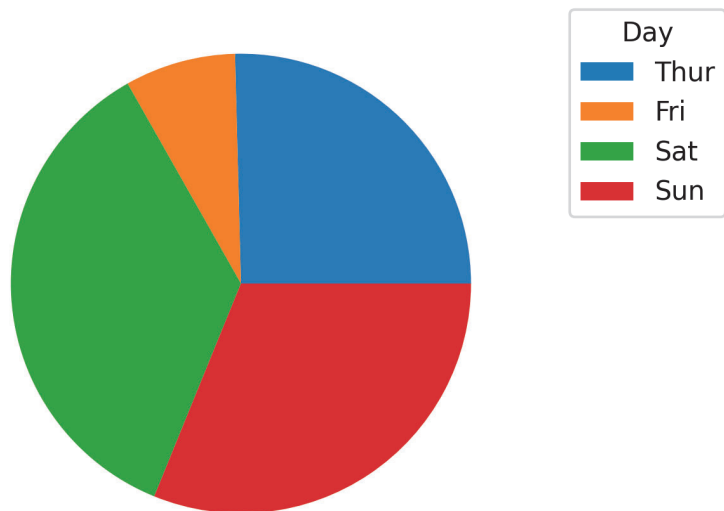
จากแผนภูมิแท่งร้อยละแบบวางซ้อนเปรียบเทียบสัดส่วนรายได้รวมจากค่าอาหารของผู้ใช้บริการเพศชายและหญิงของร้านอาหารแห่งหนึ่งในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ จากชุดข้อมูล Tips ในภาพ 6.17 ในวันพฤหัสบดีรายได้รวมจากค่าอาหารมาจากผู้ใช้บริการเพศชายและเพศหญิงมีค่าใกล้เคียงกัน ในวันถัดๆ มาร้านอาหารมีสัดส่วนรายได้จากผู้ใช้บริการเพศชายเพิ่มขึ้นเป็นลำดับ จนกระทั่งวันอาทิตย์ รายได้จากผู้ใช้บริการเพศชายเพิ่มขึ้นจนมีสัดส่วนเกือบร้อยละ 80 ของรายได้รวมจากค่าอาหาร ส่วนรายได้จากผู้ใช้บริการเพศหญิงลดลงอย่างต่อเนื่องจากประมาณ 50 เปอร์เซ็นต์ของรายได้ทั้งหมดในวันพฤหัสบดี จนเหลือประมาณร้อยละ 22 ของรายได้ทั้งหมด

แผนภูมिवงกลม

แผนภูมिवงกลม (Pie Chart) เป็นแผนภูมิที่แสดงด้วยรูปวงกลม โดยแบ่งวงกลมออกเป็นส่วนๆ จากจุดศูนย์กลางของวงกลม พื้นที่แต่ละส่วนแทนสัดส่วนปริมาณของกลุ่มข้อมูลหนึ่งๆ ต่อปริมาณข้อมูลทั้งหมด นิยมใช้ในการเปรียบเทียบสัดส่วนปริมาณของค่าตัวแปรเชิงกลุ่ม ทั้งการเปรียบเทียบปริมาณของแต่ละกลุ่มข้อมูลกับจำนวนข้อมูลทั้งหมด และการเปรียบเทียบปริมาณระหว่างกลุ่มข้อมูล อย่างไรก็ตามเมื่อมีกลุ่มข้อมูลจำนวนมาก จะต้องแบ่งส่วนวงกลมออกเป็นส่วนจำนวนมากขึ้น ทำให้กลุ่มข้อมูลที่มีขนาดเล็กสังเกตเห็นได้ยาก นอกจากนี้หากความแตกต่างของสัดส่วนระหว่างกลุ่มมีเพียงเล็กน้อย จะทำให้การเปรียบเทียบขนาดของแต่ละชิ้นส่วนทำได้ยาก ภาพ 6.18 แสดงตัวอย่างแผนภูมिवงกลมที่นำเสนอสัดส่วนจำนวนผู้ใช้บริการในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ ของร้านอาหารแห่งหนึ่ง จากชุดข้อมูล Tips

ตัวอย่าง 6.11

จากแผนภูมिवงกลมในภาพ 6.18 ซึ่งแสดงสัดส่วนจำนวนผู้ใช้บริการในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ ของร้านอาหารแห่งหนึ่ง จากชุดข้อมูล Tips พบว่าผู้ใช้บริการร้านอาหารแห่งนี้มีจำนวนน้อยที่สุดในวันศุกร์ นอกจากนี้ผู้มาใช้บริการในวันเสาร์ และอาทิตย์มีปริมาณไม่ต่างกันมากนัก และมากกว่าจำนวนผู้มาใช้บริการในวันพฤหัสบดี



ภาพ 6.18: แผนภูมิ วงกลม แสดง สัดส่วน จำนวนผู้ใช้บริการในวันพฤหัสบดี ศุกร์ เสาร์ และอาทิตย์ ของร้านอาหารแห่งหนึ่ง จากชุด ข้อมูล Tips

แผนภาพทรีแม็ป

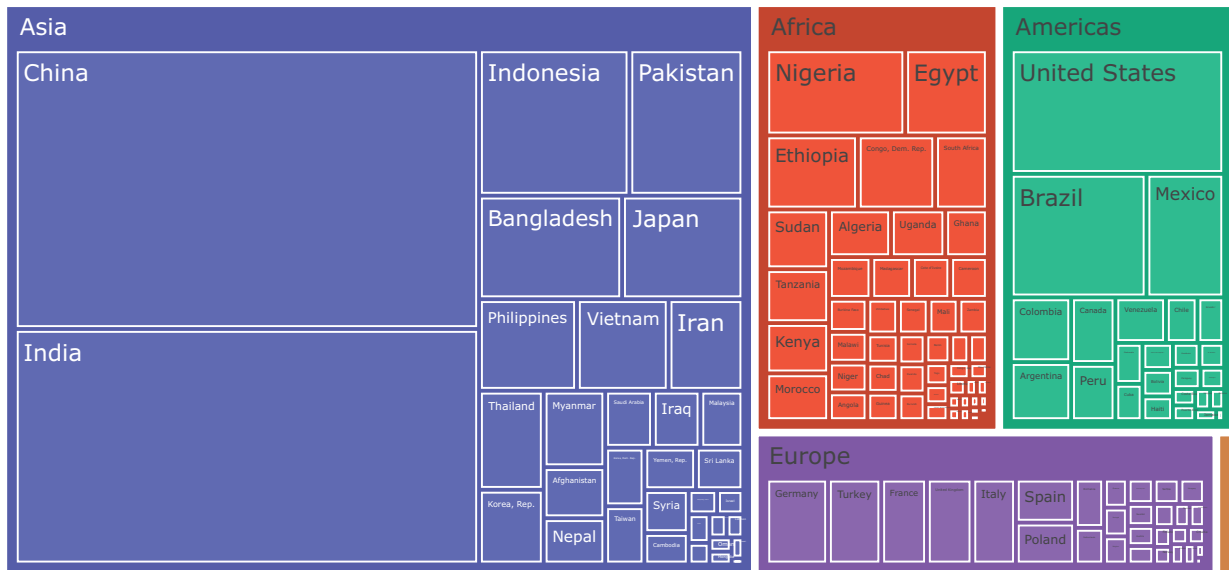
แผนภาพทรีแม็ป (Treemap) เป็นแผนภาพที่นำเสนอค่าเชิงปริมาณของข้อมูลที่มีการจัดกลุ่มแบบลำดับชั้น โดยการแบ่งพื้นที่ภายในรูปสี่เหลี่ยมใหญ่ที่แทนกลุ่มข้อมูลที่น่าสนใจ เป็นรูปสี่เหลี่ยมขนาดเล็กหลายรูปซ้อนอยู่ภายในสำหรับแทนค่าเชิงปริมาณของแต่ละกลุ่มข้อมูลย่อยในกลุ่มข้อมูลที่น่าสนใจนั้น ขนาดพื้นที่รูปสี่เหลี่ยมแสดงสัดส่วนของค่าเชิงปริมาณของกลุ่มข้อมูลที่แทนด้วยรูปสี่เหลี่ยม นั้นต่อค่าเชิงปริมาณของข้อมูลทั้งหมดที่น่าสนใจ พื้นที่ของรูปสี่เหลี่ยมใหญ่มีค่าเท่ากับผลรวมพื้นที่ของรูปสี่เหลี่ยมย่อยภายใน ภาพ 6.19 แสดงตัวอย่างแผนภาพทรีแม็ปของจำนวนประชากรในแต่ละประเทศในปี ค.ศ. 2007 จากชุดข้อมูล gapminder [42] สี่เหลี่ยมใหญ่ทั้งแผนภาพแทนจำนวนประชากรโลกทั้งหมด โดยคิดเป็น 100 เปอร์เซ็นต์ สี่เหลี่ยมใหญ่นี้ถูกแบ่งย่อยตามทวีปทั้งหมด 5 ทวีป พื้นที่สี่เหลี่ยมย่อยในระดับนี้ แทน สัดส่วนของประชากรในแต่ละทวีปต่อประชากรโลกทั้งหมด ยิ่งไปกว่านั้นสี่เหลี่ยมที่แทนสัดส่วนจำนวนประชากรแต่ละทวีป ยังถูกแบ่งย่อยลงไปอีกระดับหนึ่งตามประเทศในทวีปนั้นๆ พื้นที่สี่เหลี่ยมย่อยในระดับนี้ แทน สัดส่วนจำนวนประชากรแต่ละประเทศต่อประชากรในทวีป (เมื่อเปรียบเทียบกับพื้นที่สี่เหลี่ยมที่แทนทวีป) และสัดส่วนจำนวน ประชากร แต่ละ ประเทศ ต่อ ประชากร โลก (เมื่อเปรียบเทียบกับ พื้นที่สี่เหลี่ยมใหญ่ทั้งแผนภาพ)

[42] Jennifer Bryan (n.d.). *gapminder: Data from Gapminder*. <https://github.com/jennybc/gapminder>

ตัวอย่าง 6.12

จากแผนภาพทรีแม็ปแสดงจำนวนประชากรโลก ปี ค.ศ.2007 ในแต่ละทวีป และประเทศ จากชุดข้อมูล gapminder แสดงดังภาพ 6.19 สังเกตได้ว่า ทวีปเอเชียมีจำนวนประชากรมากที่สุดเกินครึ่งหนึ่งของประชากรโลก ส่วน

World



ภาพ 6.19: แผนภาพทรีแม็พจำนวนประชากรโลก ปี ค.ศ.2007 ในแต่ละทวีปและประเทศ จากชุดข้อมูล gapminder

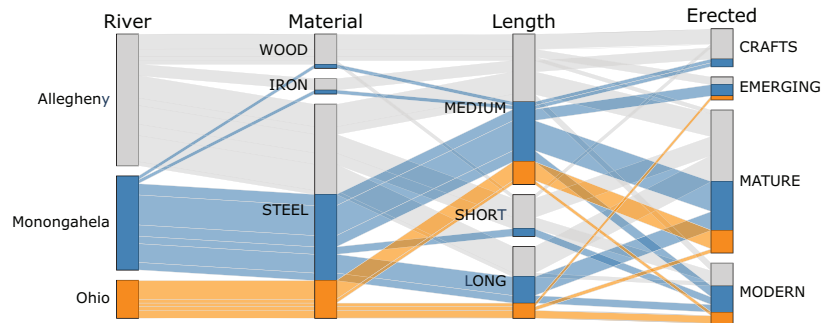
ทวีปแอฟริกาและทวีปอเมริกามีจำนวนประชากรใกล้เคียงกัน ทวีปยุโรปมีจำนวนประชากรมากเป็นอันดับที่ 4 ของโลก และทวีปโอเชียเนียมีจำนวนประชากรน้อยที่สุด

เมื่อพิจารณาจำนวนประชากรในแต่ละทวีป พบว่าประเทศจีนมีประชากรมากที่สุดทั้งในทวีปเอเชีย และมากที่สุดในโลกด้วย ตามมาด้วยประเทศอินเดีย ประเทศที่มีประชากรมากที่สุดในทวีปแอฟริกา คือ ประเทศไนจีเรีย ส่วนทวีปอเมริกานั้น ประชากรจำนวนมากอาศัยอยู่ในประเทศสหรัฐอเมริกา ประเทศเยอรมนีเป็นประเทศที่มีประชากรอาศัยอยู่มากที่สุดในประเทศยุโรป

แผนภาพเซตคู่ขนาน

แผนภาพเซตคู่ขนาน (Parallel Set Diagram) คือ การแสดงภาพชุดข้อมูลตัวแปรเชิงกลุ่มหลายตัวแปร ตัวแปรแต่ละตัวแปรในชุดข้อมูลถูกแสดงด้วยคอลัมน์ของรูปสี่เหลี่ยมผืนผ้า โดยที่แต่ละรูปสี่เหลี่ยมผืนผ้าจะสอดคล้องกับค่าที่เป็นไปได้ทั้งหมดของตัวแปรนั้น ความสูงของรูปสี่เหลี่ยมแสดงความถี่สัมพัทธ์ของการเกิดขึ้นของค่าตัวแปร รูปสี่เหลี่ยมตามค่าตัวแปรต่างๆ ที่เกิดขึ้นร่วมกันจะเชื่อมต่อกันด้วยแถบสายพาน ความสูงของแถบสายพานแทนความถี่สัมพัทธ์ของการเกิดขึ้นร่วมกันของค่าตัวแปรทั้งสองที่เชื่อมกัน มักจะสร้างแถบสายพานนี้เฉพาะคอลัมน์ที่วางข้างกันเท่านั้น ในบางครั้งสีของแถบสายพานที่แตกต่างกันยังใช้แทนค่าตัวแปรเชิงกลุ่มตัวแปรใดตัวแปรหนึ่งด้วย

ภาพ 6.20: แผนภาพเซตคู่ขนานของข้อมูลสะพานในเมืองพิตต์สเบิร์ก (Pittsburgh) ตามวัสดุก่อสร้าง (Material) ความยาว (Length) ยุคของการก่อสร้าง (Erected) และแม่น้ำที่ทอดข้าม (River) จากชุดข้อมูล Pittsburgh Bridges โดยสีที่แตกต่างกันของแถบสายพานแทนค่าตัวแปรแม่น้ำ 3 สาย ได้แก่ Allegheny Monongahela และ Ohio



ภาพ 6.20 แสดงแผนภาพเซตคู่ขนานของข้อมูลสะพานในเมืองพิตต์สเบิร์ก (Pittsburgh) ตามวัสดุก่อสร้าง (Material) ความยาว (Length) ยุคของการก่อสร้าง (Erected) และแม่น้ำที่ทอดข้าม (River) จากชุดข้อมูล Pittsburgh Bridges [43] โดยสีที่แตกต่างกันของแถบสายพานแทนค่าตัวแปรแม่น้ำ 3 สาย ได้แก่ Allegheny Monongahela และ Ohio

[43] Yoram Reich and Steven J. Fennes (n.d.). *Pittsburgh Bridges Data Set*. Accessed on 15 November 2020. url: <https://archive.ics.uci.edu/ml/datasets/Pittsburgh+Bridges>

ตัวอย่าง 6.13

แผนภาพเซตคู่ขนานของข้อมูลสะพานในเมืองพิตต์สเบิร์ก (Pittsburgh) ตามวัสดุก่อสร้าง (Material) ความยาว (Length) ยุคของการก่อสร้าง (Erected) และแม่น้ำที่ทอดข้าม (River) จากชุดข้อมูล Pittsburgh Bridges ดังแสดงในภาพ 6.20 เห็นได้ว่า แม่น้ำ Allegheny มีจำนวนสะพานที่ทอดข้ามมากที่สุด และแม่น้ำ Ohio มีจำนวนสะพานที่ทอดข้ามน้อยที่สุด สะพานในเมืองพิตต์สเบิร์กส่วนใหญ่สร้างจากเหล็กกล้าและมีความยาวปานกลาง มีการสร้างสะพานจำนวนมากในยุคที่เมืองมีความเจริญอย่างเต็มที่

สะพานที่ทอดข้ามแม่น้ำ Allegheny ส่วนมากสร้างมาจากเหล็กกล้า มีความยาวปานกลาง และถูกสร้างขึ้นในยุคที่เมืองพิตต์สเบิร์กมีความเจริญอย่างเต็มที่ ในขณะที่สะพานที่ทอดข้ามแม่น้ำ Monongahela ส่วนมากถูกสร้างมาจากเหล็กกล้าและมีความยาวปานกลาง อีกทั้งยังถูกสร้างในยุคที่เมืองมีความเจริญอย่างเต็มที่ ส่วนสะพานที่ทอดข้ามแม่น้ำ Ohio นั้นทั้งหมดถูกสร้างมาจากเหล็กกล้า ส่วนมากมีความยาวปานกลางและถูกสร้างขึ้นในยุคที่เมืองมีความเจริญอย่างเต็มที่เช่นกัน

เมื่อพิจารณาวัสดุก่อสร้าง พบว่าสะพานที่ถูกสร้างจากไม้และเหล็กทอดข้ามเฉพาะแม่น้ำ Allegheny และ Monongahela และโดยส่วนมากเป็นสะพานที่ทอดข้ามแม่น้ำ Allegheny สะพานไม้ส่วนมากมีความยาวปานกลาง ในขณะที่สะพานที่ถูกสร้างจากเหล็กมีความยาวปานกลางทั้งหมด ส่วนสะพานที่สร้างจากเหล็กกล้าสามารถพบได้ที่แม่น้ำทั้ง 3 สาย มีจำนวนสะพานเหล็กกล้าที่ทอดข้ามแม่น้ำ Allegheny และ Monongahela ใกล้เคียงกัน และเป็นสะพานเหล็กกล้าที่ทอดข้ามแม่น้ำ Ohio น้อยที่สุด สะพานที่สร้างจากเหล็กมีความยาวทั้งขนาดสั้น ปานกลาง และยาวมาก

หากพิจารณา ความยาว ของ สะพาน แล้ว พบว่า เราสามารถ พบเห็น สะพานไม้ เหล็ก และเหล็กกล้า ที่มีความยาวปานกลางทอดข้ามแม่น้ำ ในเมืองพิตส์เบิร์กได้ อีกทั้งสะพานที่มีความยาวปานกลางเริ่มมีการสร้าง มาตั้งแต่ยุคที่มีการสร้างด้วยมือจนถึงยุคสมัยใหม่ ส่วนสะพานที่มีความยาว สั้นนั้น ไม่มีสะพานใดเลยถูกสร้างจากเหล็กและสร้างในยุคของการพัฒนา ส่วนสะพานที่มีความยาวมากนั้นทุกสะพานถูกสร้างมาจากเหล็กกล้าและ เริ่มมีการสร้างตั้งแต่ยุคของการพัฒนาเป็นต้นมา

เมื่อพิจารณาตามยุคสมัยแล้ว สังเกตได้ว่าสะพานส่วนมากที่สร้างในยุค การสร้างด้วยมือมีความยาวปานกลาง และไม่มีสะพานใดเลยมีความยาว มาก อีกทั้งยังไม่มีการสร้างสะพานข้ามแม่น้ำ Ohio เลย ต่อมาในยุคของการ พัฒนา สะพานส่วนใหญ่มีความยาวปานกลาง และเริ่มมีการสร้างสะพานที่ มีความยาวมาก ในขณะที่สะพานที่มีความยาวสั้นไม่ได้มีการสร้างในยุคนี้ นอกจากนี้ยุคของการพัฒนายังเป็นยุคของการเริ่มต้นสร้างสะพานสำหรับ ข้ามแม่น้ำ Ohio ส่วนในยุคที่เมืองมีความเจริญอย่างเต็มที่ มีการก่อสร้าง สะพานจำนวนมากทั้งที่มีความยาวสั้น ความยาวปานกลาง และความยาว มาก จนกระทั่งในยุคสมัยใหม่ยังคงมีการสร้างสะพานทอดข้ามแม่น้ำทั้ง 3 สายและเป็นสะพานที่มีความยาวทั้ง 3 ขนาด แต่จำนวนสะพานที่ถูกสร้าง ลดลงอย่างเห็นได้ชัดจากยุคที่เมืองมีความเจริญอย่างเต็มที่

6.4 การมองภาพข้อมูลสำหรับข้อมูลอนุกรมเวลา

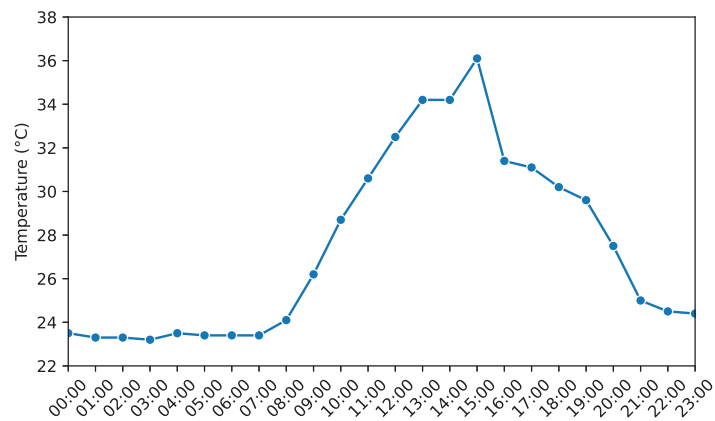
แผนภูมิเส้น

แผนภูมิเส้น (Line Chart) เป็นรูปแบบการนำเสนอข้อมูล โดยใช้จุดและส่วน ของเส้นตรงที่ลากเชื่อมต่อจุด ซึ่งแต่ละจุดแสดงค่าเชิงปริมาณของข้อมูลแต่ละ รายการ นิยมใช้แผนภูมิเส้นกับข้อมูลที่แสดงการเปลี่ยนแปลงอย่างต่อเนื่อง ตาม กาลเวลา โดยใช้แกนนอนแทนลำดับเวลาที่เปลี่ยนแปลงไป และแกนตั้งแทน ค่าตัวแปรเชิงปริมาณตัวแปรหนึ่งที่สนใจ ภาพ 6.21 แสดงตัวอย่างแผนภูมิเส้น การเปลี่ยนแปลงอุณหภูมิในแต่ละชั่วโมง (หน่วย: องศาเซลเซียส) ที่ตรวจวัดจาก ระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่ ในวันที่ 1 พฤษภาคม พ.ศ. 2563

ตัวอย่าง 6.14

จากภาพ 6.21 ซึ่งแสดงการเปลี่ยนแปลงอุณหภูมิในแต่ละชั่วโมง ที่ ตรวจวัดจากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่ ในวันที่ 1 พฤษภาคม พ.ศ.2563 พบว่าในช่วงก่อน 7

ภาพ 6.21: การเปลี่ยนแปลงอุณหภูมิในแต่ละชั่วโมง (หน่วย: องศาเซลเซียส) ที่ตรวจวัดจากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่ ในวันที่ 1 พฤษภาคม พ.ศ.2563

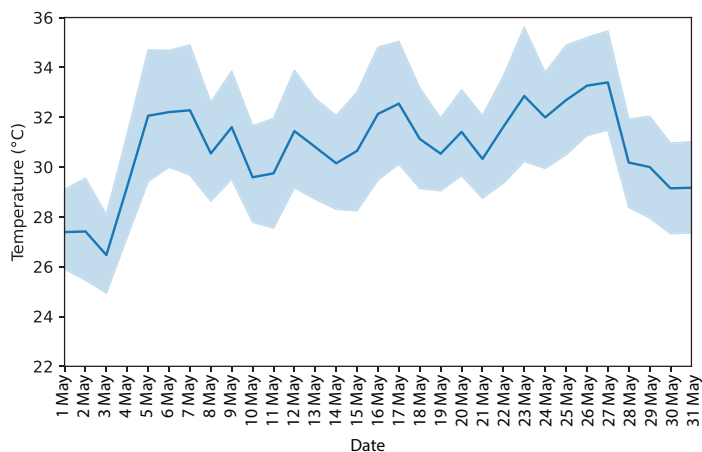


นาฬิกา อุณหภูมิมีค่าต่ำกว่า 24 องศาเซลเซียสและค่อยข้างคงที่ จนกระทั่งเวลา 8 นาฬิกา อุณหภูมิเริ่มเพิ่มสูงขึ้น และในเวลา 9 นาฬิกาถึง 15 นาฬิกา อุณหภูมิเพิ่มสูงขึ้นอย่างรวดเร็วจนมีค่าสูงสุดประมาณ 36 องศาเซลเซียส จากนั้นอุณหภูมิจึงลดลงอย่างต่อเนื่องและรวดเร็ว จนกระทั่งเวลา 22 นาฬิกา อุณหภูมิมีค่าประมาณ 25 องศาเซลเซียส และเริ่มคงที่อีกครั้ง

ในบางครั้งข้อมูลที่เกิดขึ้น ณ จุดเวลาหนึ่งมีจำนวนมากกว่า 1 ข้อมูล การลากเส้นตรงเชื่อมจุดข้อมูลแต่ละจุด อาจไม่ก่อให้เกิดความหมายใดๆ อีกทั้งยังทำให้การแปลความหมายจากแผนภูมิเส้นทำได้ยาก จึงต้องทำการสรุปรวมข้อมูลที่เกิดขึ้น ณ เวลาเดียวกัน อาจเป็นการใช้ค่ากลางข้อมูลเพื่อเป็นตัวแทนข้อมูลทั้งหมด ณ จุดเวลานั้น แล้วจึงทำการลากเส้นเชื่อมค่ากลางข้อมูลของแต่ละจุดเวลาเข้าด้วยกัน เส้นตรงเชื่อมค่ากลางข้อมูลนี้แสดงแนวโน้มของค่าข้อมูลโดยรวมที่เปลี่ยนแปลงไปตามกาลเวลา อาจใช้แถบสีแสดงช่วงความเชื่อมั่นหรือช่วงความแปรปรวนของข้อมูลขนานไปกับเส้นตรงที่เชื่อมระหว่างค่ากลางข้อมูล ภาพ 6.22 แสดงตัวอย่างแผนภูมิเส้นที่เกิดจากการนำข้อมูลอุณหภูมิในแต่ละวัน ในช่วงเดือนพฤษภาคม พ.ศ.2563 จากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่ ซึ่งทำการตรวจวัดทุกๆ ชั่วโมง มานำเสนอในรูปแบบแผนภูมิเส้นแสดงการเปลี่ยนแปลงอุณหภูมิรายวัน โดยเส้นกราฟแสดงค่าเฉลี่ยอุณหภูมิในแต่ละวัน และความกว้างของแถบสีที่ขนานกับเส้นกราฟแสดงความแปรปรวนของข้อมูลอุณหภูมิที่เกิดขึ้นในวันหนึ่งๆ

ตัวอย่าง 6.15

จากแผนภูมิเส้นแสดงการเปลี่ยนแปลงอุณหภูมิรายวัน ในช่วงเดือนพฤษภาคม พ.ศ.2563 ณ สถานีตรวจวัดเขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่ ซึ่งแสดงในภาพ 6.21 พบว่าในช่วงต้นเดือน อุณหภูมิเฉลี่ยรายวันมีค่าต่ำกว่า



ภาพ 6.22: การเปลี่ยนแปลง อุณหภูมิ ในแต่ละวัน (หน่วย:องศาเซลเซียส) ในช่วงเดือน พฤษภาคม พ.ศ.2563 ที่ตรวจวัดจากระบบ โทรมาตร ณ สถานีตรวจวัด เขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่

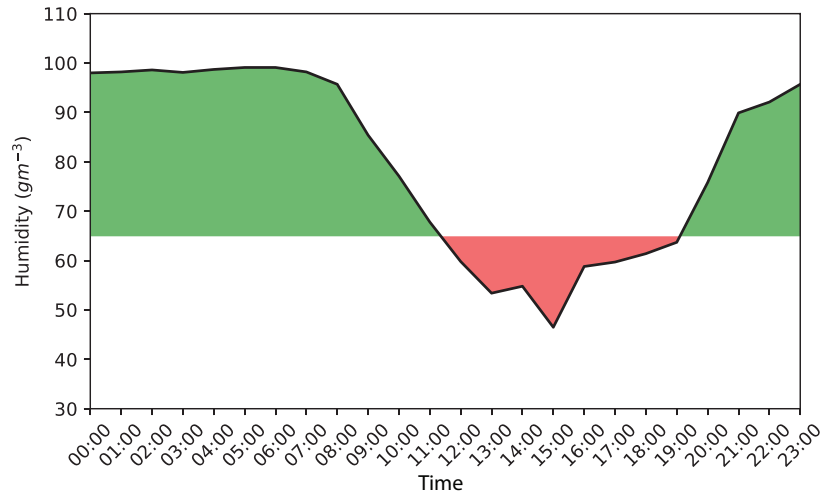
28 องศาเซลเซียส ต่อมา ในช่วงวันที่ 4 ถึงวันที่ 27 อุณหภูมิเฉลี่ยรายวัน เพิ่มขึ้นจากช่วงก่อน อย่างเห็นได้ชัดและมีการเปลี่ยนแปลงค่อนข้างคงที่ในช่วง 29 ถึง 33 องศาเซลเซียส จนกระทั่งวันที่ 28 อุณหภูมิเฉลี่ยรายวันลดลงอย่างมีนัยสำคัญอยู่ที่ประมาณ 29 องศาเซลเซียส และลดลงอย่างต่อเนื่องจนกระทั่งสิ้นสุดเดือนพฤษภาคม

เมื่อพิจารณาการเปลี่ยนแปลงอุณหภูมิในระหว่างวันของแต่ละวันแล้ว พบว่าอุณหภูมิระหว่างวันในช่วงเดือนพฤษภาคม มีการเปลี่ยนแปลงมากที่สุดในช่วง ± 2.5 องศาเซลเซียส จากค่าเฉลี่ย และมีการเปลี่ยนแปลงน้อยที่สุดในช่วง ± 1.5 องศาเซลเซียสจากค่าเฉลี่ย

แผนภูมิพื้นที่

แผนภูมิพื้นที่ (Area Chart) เป็นรูปแบบการนำเสนอข้อมูลที่มีความคล้ายคลึงกับแผนภูมิเส้น เว้นแต่พื้นที่ระหว่างแกนนอนและเส้นแสดงค่าข้อมูลถูกระบายด้วยสีหรือแรเงา ทั้งนี้แกนนอนนั้นอาจเป็นแกนนอนที่ค่าตัวแปรในแกนตั้งเท่ากับ 0 หรือเท่ากับค่าอื่นๆ ขึ้นอยู่กับสารสนเทศที่ต้องการนำเสนอ แผนภูมิพื้นที่นิยมใช้ในการนำเสนอการเปลี่ยนแปลงของค่าตัวแปรเชิงตัวเลขตามกาลเวลาเช่นเดียวกับแผนภูมิเส้น นอกจากนี้ยังสามารถระบายสีที่แตกต่างกันสำหรับนำเสนอค่าตัวแปรเชิงกลุ่มของข้อมูลในแต่ละจุดเวลาได้อีกด้วย ดังแสดงตัวอย่างในภาพ 6.23 ซึ่งเป็นแผนภูมิพื้นที่ที่แสดงการเปลี่ยนแปลงความชื้นในแต่ละชั่วโมง (หน่วย: กรัมของไอน้ำต่อลูกบาศก์เมตร) ที่ตรวจวัดจากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จัดสมบูรณ์ชล จังหวัดเชียงใหม่ ในวันที่ 1 พฤษภาคม พ.ศ. 2563 สีที่ระบายลงบนพื้นที่แสดงทิศทางระหว่างค่าความชื้นและค่าความชื้นเฉลี่ยตลอดเดือนพฤษภาคม โดย สีเขียว หมายถึง ค่าความชื้น

ภาพ 6.23: การเปลี่ยนแปลง ค่า ความชื้น ในแต่ละ ชั่วโมง (หน่วย: กรัมของไอน้ำต่อลูกบาศก์เมตร) ในวันที่ 1 พฤษภาคม พ.ศ. 2563 ที่ตรวจวัดจากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จันทน์สมบูรณ์ชล จังหวัดเชียงใหม่ เปรียบเทียบกับ ค่า ความชื้นเฉลี่ยตลอดเดือนพฤษภาคม



สูงกว่าค่าความชื้นเฉลี่ย และสีแดง หมายถึง ค่าความชื้นต่ำกว่าค่าความชื้นเฉลี่ย

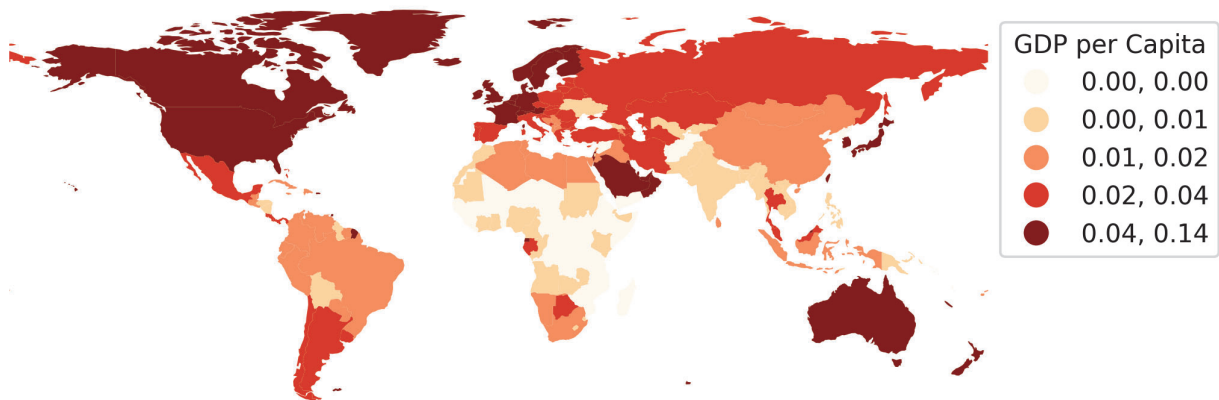
ตัวอย่าง 6.16

จากแผนภูมิพื้นที่แสดงการเปลี่ยนแปลงค่าความชื้นในแต่ละชั่วโมง ในวันที่ 1 พฤษภาคม พ.ศ.2563 เปรียบเทียบกับค่าความชื้นเฉลี่ยตลอดเดือนพฤษภาคม ที่ตรวจวัดจากระบบโทรมาตร ณ สถานีตรวจวัดเขื่อนแม่จันทน์สมบูรณ์ชล จังหวัดเชียงใหม่ ในภาพ 6.23 สังเกตได้ว่าความชื้นในช่วงก่อน 8 นาฬิกามีค่าค่อนข้างคงที่อยู่ที่ประมาณ 97 ถึง 99 กรัมของไอน้ำต่อลูกบาศก์เมตร หลังจากนั้นความชื้นมีค่าลดลงอย่างรวดเร็วจนมีค่าต่ำสุดประมาณ 46 กรัมของไอน้ำต่อลูกบาศก์เมตร เมื่อเวลา 15 นาฬิกา จากนั้นค่าความชื้นเพิ่มสูงขึ้นอย่างต่อเนื่อง กระทั่งเวลา 23 นาฬิกา ค่าความชื้นอยู่ที่ระดับประมาณ 97 กรัมของไอน้ำต่อลูกบาศก์เมตร นอกจากนี้ยังพบว่า ช่วงเวลาประมาณ 11 นาฬิกา ถึง 19 นาฬิกา ค่าความชื้นต่ำกว่าค่าความชื้นเฉลี่ยตลอดเดือนพฤษภาคม

6.5 การมองภาพข้อมูลสำหรับแผนที่และเครือข่าย

แผนที่โคโรเพลท

แผนที่โคโรเพลท (Choropleth Map) เป็นการแสดงพื้นที่ทางภูมิศาสตร์ที่ถูกแบ่งออกเป็นเขต เมือง ประเทศ หรือ ภูมิภาค โดยระบายสีพื้นที่ด้วยระดับสี หรือ ความเข้มที่สัมพันธ์กับตัวแปรเชิงตัวเลขของพื้นที่นั้น ทำให้เห็นค่าตัวแปรเชิงตัวเลขที่เปลี่ยนแปลงไปตามพื้นที่ อย่างไรก็ตาม ภูมิภาคพื้นที่ขนาดใหญ่ มักถูกให้น้ำหนักในการตีความที่มากกว่าภูมิภาคพื้นที่ขนาดเล็ก



ภาพ 6.24: แผนที่โครเพลทแสดงค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรของแต่ละประเทศ จากชุดข้อมูล Natural Earth

เล็กกว่า ซึ่งทำให้เกิดอคติในการแปลความหมายจากภาพ ภาพ 6.24 เป็นตัวอย่างการนำเสนอข้อมูลค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากร (GDP per Capita) จากชุดข้อมูล Natural Earth [44] ของแต่ละประเทศในรูปแบบแผนที่โครเพลท โดยการแบ่งค่าผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรออกเป็นช่วง 5 ช่วงค่าข้อมูล และใช้ระดับความเข้มของสีแทนค่าข้อมูลแต่ละช่วงค่าข้อมูล โดยข้อมูลที่มีค่าน้อยแทนด้วยระดับความเข้มของสีที่ต่ำ และระดับความเข้มของสีจะเพิ่มขึ้น เมื่อข้อมูลมีค่ามากขึ้น

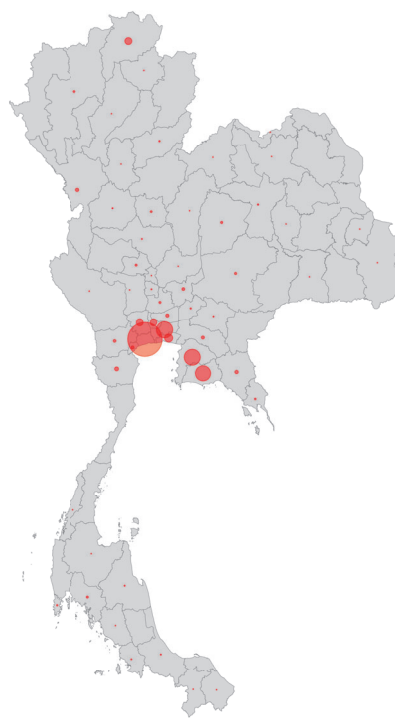
[44] Nathaniel Vaughn Kelso and Tom Patterson (2012). *Natural Earth v2.0.0*. Accessed on 15 November 2020. url: <https://www.naturalearthdata.com/downloads/>

ตัวอย่าง 6.17

จากแผนที่โครเพลทแสดงค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรของแต่ละประเทศ จากชุดข้อมูล Natural Earth แสดงดังรูป 6.24 ทำให้เห็นว่าประเทศในทวีปอเมริกาเหนือ ยุโรป ส่วนใหญ่มีค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรค่อนข้างสูง ส่วนประเทศในทวีปเอเซียนั้นส่วนมากมีค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรในระดับปานกลาง ยกเว้นประเทศญี่ปุ่นที่มีค่าประมาณอยู่ในระดับสูง ในทวีปแอฟริกา ประเทศส่วนใหญ่มีค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรค่อนข้างน้อย โดยเฉพาะประเทศในแถบตอนกลางของทวีป นอกจากนี้ยังพบว่าประเทศออสเตรเลียมีค่าประมาณผลิตภัณฑ์มวลรวมภายในประเทศต่อหัวประชากรสูง

แผนที่ฟอง

แผนที่ฟอง (Bubble Map) เป็นรูปแบบการนำเสนอข้อมูลที่คล้ายคลึงกับแผนที่โครเพลท แตกต่างกันที่แผนที่ฟองใช้การวางรูปร่างกลมบนพิกัดแผนที่แทนการระบายสีลงบนพื้นที่ และใช้ขนาดของรูปร่างกลมแสดงขนาดของ



ภาพ 6.25: แผนที่ฟอง แสดง จำนวน ผู้ติดเชื้อไวรัสโคโรนา 2019 ในแต่ละจังหวัดของประเทศไทย ในช่วงเดือนธันวาคม ค.ศ.2020 รายงาน โควิด-19 ประจำวัน ข้อมูลประจำประเทศไทย โดยกรมควบคุมโรค

ค่าตัวแปรเชิงตัวเลขแทนการใช้ระดับสี แผนที่ฟองช่วยหลีกเลี่ยงอคติที่เกิดจากขนาดพื้นที่ภูมิภาคในแผนที่โครเพลท อย่างไรก็ตามหากพื้นที่ในแผนที่มีความหนาแน่นมากจนทำให้การวางรูปวงกลมมีความแออัดกัน จะทำให้เกิดการบดบังกันของรูปวงกลมจนไม่สามารถมองเห็นวงกลมที่มีขนาดเล็กได้ แสดงตัวอย่างแผนที่ฟองดังภาพ 6.25 ซึ่งนำเสนอข้อมูลจำนวนผู้ติดเชื้อไวรัสโคโรนา 2019 ในแต่ละจังหวัดของประเทศไทย ในช่วงเดือนธันวาคม ค.ศ.2020 จากรายงาน โควิด-19 ประจำวัน ข้อมูลประจำประเทศไทย โดยกรมควบคุมโรค [45] ขนาดของรูปวงกลมถูกใช้แทนจำนวนผู้ป่วยโควิด-19 ในจังหวัดที่จุดศูนย์กลางของรูปวงกลมนั้นวางอยู่

[45] กรม ควบคุม โรค (2021). รายงาน COVID-19 ประจำ วัน ข้อมูล ประจำ ประเทศไทย. Accessed on 1 February 2021. url: <https://data.go.th/dataset/covid-19-daily>

ตัวอย่าง 6.18

จากแผนที่ฟองแสดงจำนวนผู้ติดเชื้อไวรัสโคโรนา 2019 ในแต่ละจังหวัดของประเทศไทย ในช่วงเดือนธันวาคม ค.ศ.2020 แสดงในภาพ 6.25 ให้เห็นว่าช่วงเดือนธันวาคม ค.ศ.2020 ผู้ติดเชื้อไวรัสโคโรนา 2019 ส่วนใหญ่อยู่ในพื้นที่ปริมณฑล จังหวัดชลบุรี และจังหวัดระยอง โดยจังหวัดที่พบผู้ป่วยโควิด-19 มากที่สุด คือ จังหวัดสมุทรสาคร ส่วนจังหวัดในพื้นที่อื่นๆ มีจำนวนผู้ติดเชื้อเล็กน้อยหรือไม่พบการติดเชื้อในพื้นที่ นอกจากนี้ยังสังเกตได้ว่า ที่จังหวัดเชียงราย มีการจำนวนผู้ติดเชื้อมากที่สุดในภาคเหนือ ในช่วงเดือนธันวาคมนี้

ตัวอย่าง 6.19

แผนภาพเครือข่ายแสดงความสัมพันธ์ของผู้เขียนร่วมของ Vincent Ranwez ในภาพ 6.26 ทำให้เห็นว่าผู้เขียนร่วมของ Vincent Ranwez เคยเขียนบทความร่วมกันอย่างน้อย 2 คน กล่าวคือ ไม่มีผู้เขียนร่วมคนใดเลยที่ไม่เคยร่วมเขียนงานกับผู้เขียนร่วมคนอื่นๆ นอกจากนี้ยังสังเกตได้ว่าผู้เขียนร่วมแบ่งออกเป็นกลุ่ม 11 กลุ่มที่เคยเขียนบทความร่วมกัน กลุ่มที่เล็กที่สุดมีผู้เขียนร่วมจำนวน 2 คน และมีกลุ่มเครือข่ายผู้เขียนร่วมขนาดใหญ่ 1 กลุ่ม

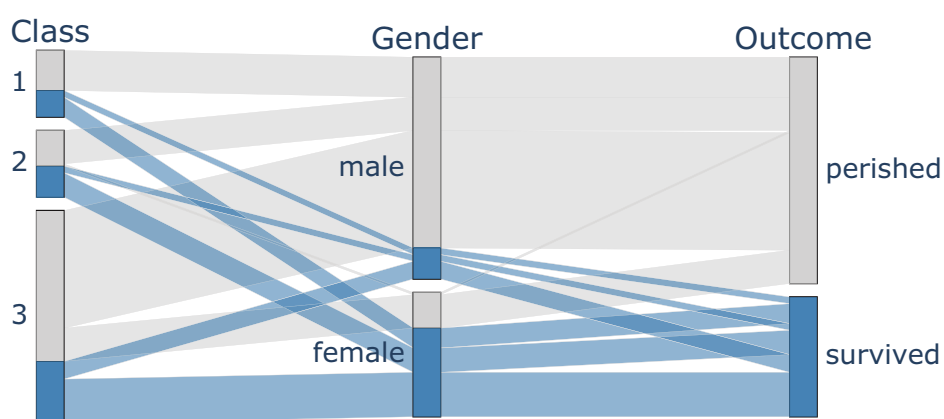
6.6 เอกสารอ้างอิงและแหล่งศึกษาเพิ่มเติม

1. Ronny Kohavi and Barry Becker (n.d.). *Adult Census Income*. Accessed on 15 November 2020. url: <https://www.kaggle.com/c/adult-census-income>
2. Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman (2020). *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. R package version 0.1.0. url: <https://allisonhorst.github.io/palmerpenguins/>
3. Penny Analytics (n.d.). *Fisher Iris Data*. Accessed on 5 October 2019. url: <https://pennyanalytics.com/free-trial/>
4. Harold V. Henderson and Paul F. Velleman (1981). “Building Multiple Regression Models Interactively.” In: *Biometrics* 37.2
5. Peter G. Bryant and Marlene A. Smith (1994). *Practical Data Analysis: Case Studies in Business Statistics*. McGraw-Hill Professional. isbn: 0256158282
6. Jennifer Bryan (n.d.). *gapminder: Data from Gapminder*. <https://github.com/jennybc/gapminder>
7. Yoram Reich and Steven J. Fenves (n.d.). *Pittsburgh Bridges Data Set*. Accessed on 15 November 2020. url: <https://archive.ics.uci.edu/ml/datasets/Pittsburgh+Bridges>
8. Nathaniel Vaughn Kelso and Tom Patterson (2012). *Natural Earth v2.0.0*. Accessed on 15 November 2020. url: <https://www.naturalearthdata.com/downloads/>
9. กรมควบคุมโรค (2021). *รายงาน COVID-19 ประจำวัน ข้อมูลประจำประเทศไทย*. Accessed on 1 February 2021. url: <https://data.go.th/dataset/covid-19-daily>
10. Claus O. Wilke (2019). *Fundamentals of Data Visualization*. USA: O’Reilly Media, Inc.
11. Wendy L. Martinez, Angel R. Martinez, and Jeffrey L. Solka (2017). *Exploratory Data Analysis with MATLAB*. USA: CRC Press, Inc.
12. *From Data to Viz* (2018). Accessed on 31 March 2020. url: <https://www.data-to-viz.com/>

6.7 แบบฝึกหัดท้ายบท

1. จากรูปแบบการมองภาพข้อมูลที่กล่าวมาในบทที่ 6 รูปแบบการการมองภาพข้อมูลใดบ้างที่เหมาะสมสำหรับนำมาใช้นำเสนอข้อมูลเพื่อวัตถุประสงค์ต่อไปนี้
 - a) เพื่อค้นหาความสัมพันธ์ระหว่างค่าตัวแปรเชิงตัวเลข 2 ตัวแปรใดๆ
 - b) เพื่ออธิบายลักษณะการกระจายของค่าตัวแปรหนึ่ง สำหรับแต่ละกลุ่มข้อมูล
 - c) เพื่อค้นหาค่าผิดปกติในตัวแปรเชิงตัวเลข
 - d) เพื่อเปรียบเทียบจำนวนหรือปริมาณในแต่ละกลุ่มข้อมูล
 - e) เพื่อศึกษาการเปลี่ยนแปลงของค่าข้อมูลเชิงตัวเลขที่แปรเปลี่ยนไปตามกาลเวลา
 - f) เพื่อนำเสนอความสัมพันธ์ระหว่างข้อมูล 2 ข้อมูลใดๆ ในชุดข้อมูล
 - g) เพื่อศึกษาลักษณะการกระจายของค่าข้อมูล 2 ตัวแปรใดๆ ในแต่ละกลุ่มข้อมูล

2. จงอธิบายความเหมือน และความแตกต่างระหว่างรูปแบบการมองภาพข้อมูลด้วยฮีตแมป และแผนที่โคโรเพลท
3. หากนักศึกษาต้องการนำเสนอสัดส่วนของจำนวนผู้ป่วยโรคโควิด-19 ในแต่ละเพศ โดยแยกตามช่วงอายุแล้ว วิธีการมองภาพข้อมูลรูปแบบใดเหมาะสมในการนำมาใช้ พร้อมอธิบายแนวทางการนำรูปแบบการมองภาพข้อมูลที่เหมาะสมมาใช้นำเสนอข้อมูล
4. จากชุดข้อมูล Fisher Iris หากนักศึกษาต้องการเลือกตัวแปรเชิงตัวเลขชุดหนึ่งสำหรับใช้เป็นตัวแปรจำแนกชนิดพันธุ์ของดอกไอริส วิธีการมองภาพข้อมูลรูปแบบใดที่นักศึกษานำมาช่วยในการเลือกตัวแปรที่เหมาะสมสำหรับจำแนกชนิดพันธุ์ของดอกไอริสได้ จงอธิบายแนวทางการนำรูปแบบการมองภาพข้อมูลที่เหมาะสมมาใช้ประโยชน์
5. หากนักศึกษาต้องการนำเสนอเมทริกซ์ความสับสน (Confusion Matrix) ซึ่งเป็นผลลัพธ์จากการวัดประสิทธิภาพตัวแบบทำนายข้อมูลชุดหนึ่งที่มีตัวแปรเป้าหมายมีค่าที่เป็นไปได้หลายค่า รูปแบบการมองภาพข้อมูลที่เหมาะสมสำหรับนำเสนอเมทริกซ์ความสับสนนี้ พร้อมทั้งอธิบายแนวทางการนำเสนอ
6. จากข้อมูลราย COVID-19 ประจำวัน ข้อมูลประจำประเทศไทย ของกรมควบคุมโรค (ดูข้อมูลที่ <https://data.go.th/dataset/covid-19-daily>) จงสร้างแผนภาพนำเสนอข้อมูล ต่อไปนี้ พร้อมทั้งอธิบายแผนภาพ
 - a) แผนภาพเปรียบเทียบอัตราความเสี่ยง (Risk) ของการติดเชื้อไวรัสโคโรนา-2019 ในผู้ป่วยแต่ละเพศ
 - b) แผนภาพเปรียบเทียบอัตราของผู้ป่วยชายและหญิงในแต่ละจังหวัดที่ผู้ป่วยได้รับการรักษา (province of isolation)
 - c) แผนภาพการกระจายของค่าตัวแปรอายุผู้ป่วย
 - d) แผนภาพจำนวนผู้ติดเชื้อที่ได้รับการรักษาในแต่ละจังหวัด
 - e) แผนภาพแสดงการเปลี่ยนแปลงจำนวนผู้ติดเชื้อรวมทั้งประเทศในแต่ละวัน
7. จากแผนภาพชุดคู่ขนานแสดงประเภทชั้นผู้โดยสาร (Class) เพศ (Sex) และ ผลลัพธ์ (Outcome) ของผู้โดยสารเรือไททานิค จากชุดข้อมูล Titanic [49] ด้านล่าง



ข้อสรุปต่อไปนี กล่าวถูกต้อง ไม่ถูกต้อง หรือไม่สามารถสรุปได้

- a) ผู้โดยสารเรือไททานิคส่วนใหญ่เสียชีวิตจากโศกนาฏกรรมเรือไททานิคอับปาง
- b) ผู้โดยสารเรือไททานิคที่รอดชีวิตส่วนใหญ่เป็นเพศหญิง
- c) ผู้โดยสารเพศชายและหญิงในชั้นโดยสารทั้ง 3 ชั้นมีอัตราส่วนใกล้เคียงกัน

- d) ผู้โดยสารเพศหญิงในชั้นโดยสารที่ 1 ทั้งหมดเสียชีวิต
- e) ผู้โดยสารเพศชายที่เสียชีวิตทั้งหมดเป็นผู้โดยสารในชั้นโดยสารที่ 3
- f) ผู้โดยสารเพศหญิงทั้งหมดที่เสียชีวิตมีสาเหตุมาจากการจมน้ำ