

The Educational Statistics and Data Sciences Series

1

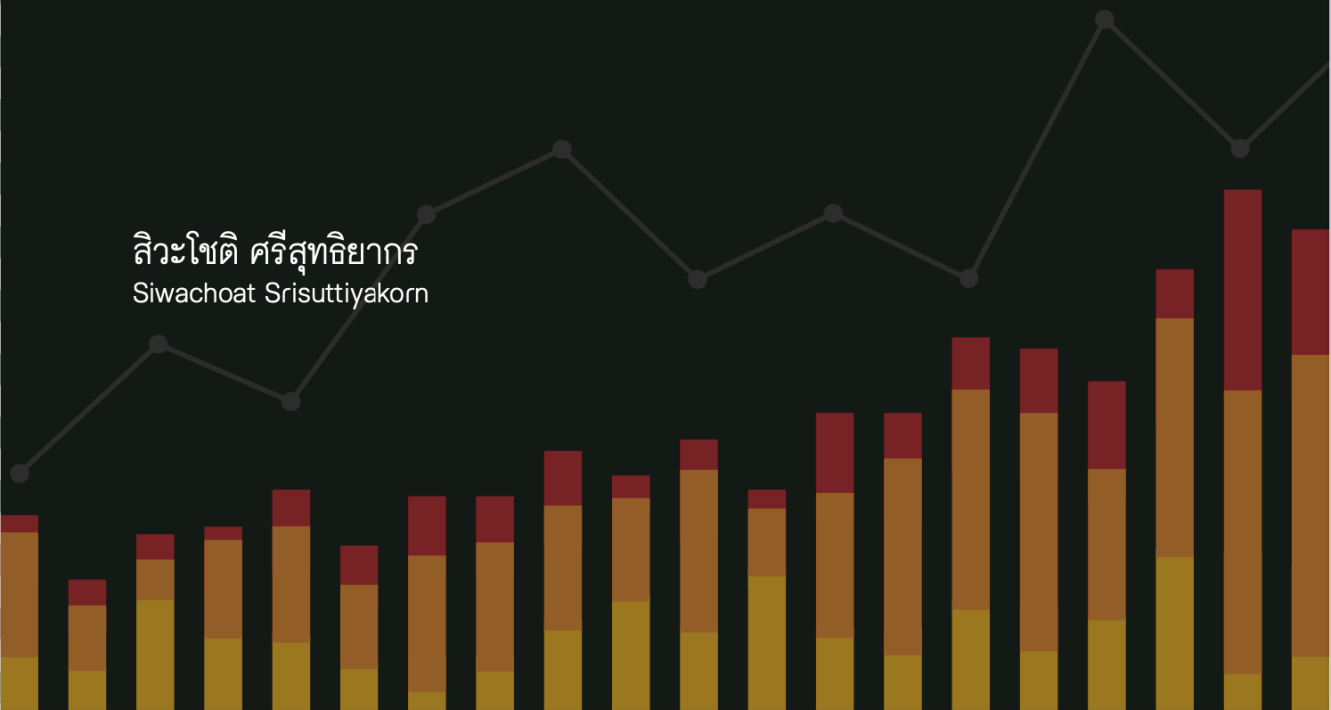
สถิติและวิทยาการ ข้อมูลทางการศึกษา

R สำหรับการจัดระเบียบและจัดกระทำข้อมูล

— — — — —
Educational Statistics and Data Sciences

R for tidying and manipulating data

สิวะโชติ ศรีสุทธียากร
Siwachoat Srisuttiyakorn





Chula
Chulalongkorn University

The Educational Statistics and Data Sciences Series

1

สถิติและวิทยาการ ข้อมูลทางการศึกษา

R สำหรับการจัดระเบียบและจัดการทำข้อมูล

Educational Statistics and Data Sciences:

R for tidying and manipulating data

สิวะโชติ ศรีสุทธียากร

Siwachoat Srisuttiyakorn

สถิติและวิทยาการข้อมูลทางการศึกษา : R สำหรับจัดระเบียบและจัดกระทำข้อมูล

โดย สิวะโชติ ศรีสุทธียากร

ลิขสิทธิ์ของคณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

การผลิตและการลอกเลียนหนังสือเล่มนี้ไม่ว่ารูปแบบใดทั้งสิ้น

ต้องได้รับอนุญาตเป็นลายลักษณ์อักษรจากคณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

ข้อมูลทางบรรณานุกรมของหอสมุดแห่งชาติ

National Library of Thailand Catalog in Publication Data

สิวะโชติ ศรีสุทธียากร.

สถิติและวิทยาการข้อมูลทางการศึกษา : R สำหรับจัดระเบียบและจัดกระทำข้อมูล. --

กรุงเทพฯ : คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย, 2564.

285 หน้า.— (สถิติและวิทยาการข้อมูลทางการศึกษา).

1. สถิติวิเคราะห์. 2. สถิติวิเคราะห์ -- โปรแกรมคอมพิวเตอร์. I. ชื่อเรื่อง.

005.55

ISBN 978-616-407-594-8

จัดทำโดย

คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย

254 ถนนพญาไท แขวงวังใหม่ เขตปทุมวัน กรุงเทพฯ 10330

โทร. 0-2218-2565 โทรสาร. 0-2218-2563

ออกแบบปก : สิวะโชติ ศรีสุทธียากร

ออกแบบรูปเล่ม : สิวะโชติ ศรีสุทธียากร

วาดภาพประกอบ : สิวะโชติ ศรีสุทธียากร

สารบัญ

สารบัญ	v
สารบัญรูป	vii
สารบัญตาราง	xv
คำนำ	xvii
บทที่ 1 บทนำ	1
1.1 ขอบเขตเนื้อหาของหนังสือ	3
1.2 ความรู้เบื้องต้นที่จำเป็นสำหรับผู้อ่าน	4
1.3 การดาวน์โหลดและติดตั้งโปรแกรม R และ package	5
1.4 ตัวอย่างคำสั่ง และชุดข้อมูลที่ใช้ในหนังสือ	5
บทที่ 2 พื้นฐานโปรแกรม R	7
2.1 สภาพแวดล้อมของโปรแกรม R	7
2.3 การคำนวณทางคณิตศาสตร์พื้นฐาน	9
2.4 ฟังก์ชัน (function)	9
2.4 ตัวแปร (Variable)	11
2.5 เวกเตอร์ (vectors)	15
2.6 เมทริกซ์ (matrices)	19
2.7 ตัวแปรลิสต์ (Lists)	24
2.8 ชุดข้อมูล (data frame)	27
2.9 ตัวแปรจัดประเภท	31
บทที่ 3 การนำข้อมูลเข้าสู่โปรแกรม R	37
3.1 การกำหนด working directory	37
3.2 การนำเข้าและส่งออกไฟล์ข้อมูลแบบ flat file	38
3.3 การนำเข้าและส่งออกไฟล์ข้อมูล .xlsx	43
3.4 การนำเข้าข้อมูลจาก clipboard	47
3.5 การนำเข้าไฟล์ข้อมูลนามสกุล .sav	48

บทที่ 4 การจัดระเบียบชุดข้อมูล	51
4.1 การสำรวจโครงสร้างของชุดข้อมูล	52
4.2 การรวมชุดข้อมูลตามแถว (row combined)	58
4.3 การรวมชุดข้อมูลตามคอลัมน์ (column combined)	64
4.4 การปรับเปลี่ยนรูปแบบชุดข้อมูล	69
4.5 การแยกคอลัมน์	74
4.6 การยุบรวมคอลัมน์	80
4.7 การจัดการกับหน่วยข้อมูลที่ซ้ำซ้อน	82
4.8 การยุบรวมหน่วยข้อมูล	85
บทที่ 5 การสำรวจข้อมูล	87
5.1 มโนทัศน์พื้นฐานของทัศนภาพข้อมูล	90
5.2 ฟังก์ชันกราฟิกพื้นฐาน	99
5.3 การสำรวจการแจกแจงของตัวแปร	106
5.4 การสำรวจสภาพตัวแปรด้วยสถิติพื้นฐาน	129
5.5 การสำรวจความสัมพันธ์ระหว่างตัวแปร	141
บทที่ 6 การจัดกระทำข้อมูล	169
6.1 การคัดเลือกตัวแปร และการคัดกรองหน่วยข้อมูล	170
6.2 การแปลงข้อมูล	181
6.3 การสรุปข้อมูลด้วยค่าสถิติพื้นฐาน	192
บทที่ 7 การวิเคราะห์ค่าสูญหาย	197
7.1 การสำรวจและการแปลงรหัสค่าสูญหาย	198
7.2 กลไกการเกิดค่าสูญหาย	220
7.3 การวินิจฉัยกลไกการสูญหายของข้อมูล	228
7.4 การแก้ปัญหาเมื่อเกิดค่าสูญหายด้วยวิธีแบบดั้งเดิม	238
7.5 การทดแทนค่าสูญหายแบบหลายค่า	267

สารบัญรูป

รูป 2.1	สภาพแวดล้อมในโปรแกรม R	8
รูป 2.2	คะแนนสอบวิชาสถิติของนิสิต	15
รูป 2.3	การสร้างเมทริกซ์ด้วยฟังก์ชัน <code>cbind()</code> และ <code>rbind()</code>	20
รูป 2.4	ตัวอย่างโครงสร้างข้อมูลภายในตัวแปรแบบลิสต์	24
รูป 2.5	อัลกอริทึมของฟังก์ชัน <code>ifelse()</code>	32
รูป 3.1	ลักษณะ flat file แบบ CSV และ TSV	36
รูป 3.2	ไฟล์ข้อมูล <code>score2.txt</code> ที่ใช้ตัวคั่นเป็น <code>/</code>	42
รูป 3.3	ไฟล์ข้อมูล <code>KAS.xlsx</code>	45
รูป 3.4	หน้าต่างแสดงไฟล์ข้อมูล <code>score2.txt</code> บน google drive	47
รูป 4.1	ลักษณะของชุดข้อมูลจัดระเบียบ	52
รูป 4.2	การยุบรวมข้อมูลตามแถว (row combined)	58
รูป 4.3	การทำงานของฟังก์ชันการทวนซ้ำแบบ <code>for</code>	62
รูป 4.4	ลักษณะการยุบรวมตารางข้อมูลตามคอลัมน์	64
รูป 4.5	ลักษณะการยุบรวมตารางข้อมูลตามคอลัมน์ด้วยฟังก์ชัน <code>merge</code> ใน R	65
รูป 4.6	การทำงานของฟังก์ชัน <code>join</code> ทั้ง 4 ลักษณะ	66
รูป 5.1	แผนภาพ Alluvial แสดงการแจกแจงความถี่หลายทางของจำนวนนักเรียน	90
รูป 5.2	แผนภาพการกระจายของตัวแปร <code>y</code> และ <code>x</code>	91
รูป 5.3	ตัวอย่างภาพที่ใช้พิกัดแบบคาร์ทีเซียน	92
รูป 5.4	ตัวอย่างพิกัดแบบเชิงขั้ว	93
รูป 5.5	qualitative palettes	96
รูป 5.6	ตัวอย่างแผนภาพวงกลมที่ประเภทของข้อมูลมีมากเกินไป	96
รูป 5.7	sequential palettes	97
รูป 5.8	ร้อยละของนักเรียนในแต่ละจังหวัดที่มีโอกาสทางการศึกษาอยู่ในระดับสูงมาก	97

รูป 5.9	diverging palettes	98
รูป 5.10	ลักษณะของจานสีแบบไม่ต่อเนื่องและต่อเนื่อง	98
รูป 5.11	ระดับปฏิสัมพันธ์ของนักเรียนในการเรียน online	99
รูป 5.12	ตัวอย่างการใช้สีเพื่อเน้นข้อมูล	99
รูป 5.13	ฮิสโทแกรมที่สร้างจากฟังก์ชัน <code>hist()</code>	100
รูป 5.14	สีพื้นฐาน 8 สีสำหรับฟังก์ชันกราฟิกในโปรแกรม R	102
รูป 5.15	ตัวอย่างฮิสโทแกรมที่กำหนดสีด้วยระบบสี RGB	104
รูป 5.16	แผนภูมิแท่งแสดงการแจกแจงความถี่ของระดับชั้นของผู้เรียน	116
รูป 5.17	แผนภูมิแท่งแจกแจงความถี่สโตร์การเรียนรู้ที่ชอบของผู้เรียน	117
รูป 5.18	แผนภูมิแท่งเปรียบเทียบสโตร์การเรียนรู้ที่ชอบของผู้เรียนจำแนกตามระดับชั้น	118
รูป 5.19	ฮิสโทแกรมของผลสัมฤทธิ์โดยรวมของนักเรียนจำแนกตามประเภทรายวิชา	119
รูป 5.20	ฮิสโทแกรมของผลสัมฤทธิ์โดยรวมของนักเรียนในรายวิชาที่สอนแบบ online	120
รูป 5.21	ฮิสโทแกรมที่มีการกำหนดให้อันตรภาคชั้นมีจำนวนเท่ากับ 5, 10 และ 30	121
รูป 5.22	ฮิสโทแกรมสร้างจากฟังก์ชัน <code>hist()</code> โดยมีการกำหนดอาร์กิวเมนต์ลักษณะต่าง ๆ	122
รูป 5.23	ตัวอย่างฮิสโทแกรมที่มีการซ้อนแผนภาพโค้งความหนาแน่น และ rug plot	123
รูป 5.24	ฮิสโทแกรมของผลสัมฤทธิ์โดยรวมของนักเรียนในรายวิชา online ที่จำแนกตาม ลักษณะความคิดเห็นต่อการเรียนของผู้เรียน	124
รูป 5.25	แผนภูมิวงกลมของภูมิสำเนาของผู้เรียน	125
รูป 5.26	แผนภูมิวงกลมของภูมิสำเนาของผู้เรียนสร้างจากฟังก์ชัน <code>pie()</code> โดยมีการกำหนดอาร์กิวเมนต์ในลักษณะต่าง ๆ	126
รูป 5.27	แผนภูมิแท่งเชิงซ้อนแสดงจำนวนนักเรียนจำแนกตามระดับชั้นปี และสโตร์การเรียนรู้ที่ชอบของนักเรียนที่สร้างจากฟังก์ชัน <code>barplot()</code>	127
รูป 5.28	แผนภูมิแท่งเชิงซ้อนแสดงจำนวนนักเรียนจำแนกตามระดับชั้นปี และสโตร์การเรียนรู้ที่ชอบของผู้เรียนที่สร้างจากฟังก์ชัน <code>barplot()</code> และ มีการกำหนดอาร์กิวเมนต์ของฟังก์ชันในลักษณะต่าง ๆ	127
รูป 5.29	แผนภาพโมเสกแสดงการแจกแจงร่วมของผู้เรียนจำแนกตามเพศ และระดับชั้น	129
รูป 5.30	ลักษณะของแผนภาพกล่อง	130
รูป 5.31	ตัวอย่างการใช้ฟังก์ชัน <code>boxplot()</code> และมีการกำหนดอาร์กิวเมนต์ลักษณะต่าง ๆ	140

รูป 5.32 แผนภาพกล่องแสดงผลสัมฤทธิ์โดยรวมของผู้เรียนในรายวิชาที่จัดการสอน	
แบบ onine จำแนกตามลักษณะความคิดเห็นของผู้เรียนที่มีต่อการเรียน online	141
รูป 5.33 ตัวแปรที่มีความสัมพันธ์กันเชิงเส้นทางบวกและทางลบ	141
รูป 5.34 ความสัมพันธ์กันแบบไม่ใช่เส้นตรงระหว่างตัวแปร	142
รูป 5.35 แผนภาพการกระจายของคะแนนสอบปลายภาควิชาสถิติกับ	
พื้นฐานความรู้ทางคณิตศาสตร์ของนิสิต	143
รูป 5.36 ตัวอย่างความสัมพันธ์ระหว่างตัวแปรที่มีทิศทางและค่าสหสัมพันธ์ในระดับต่าง ๆ	145
รูป 5.37 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร Y กับ X ภายใต้สถานการณ์ที่	
ตัวแปรทั้งสองมีความสัมพันธ์เชิงเส้นตรงโดยสมบูรณ์	
และมีความชันในลักษณะต่าง ๆ	146
รูป 5.38 ค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างตัวแปร Y กับ X ภายใต้สถานการณ์ที่	
ตัวแปรทั้งสองเป็นอิสระซึ่งกันและกัน และตัวแปรทั้งสอง	
มีความสัมพันธ์กันแบบไม่ใช่เชิงเส้น	147
รูป 5.39 เมทริกซ์ของแผนภาพการกระจายระหว่าง	
คะแนนสอบย่อยของนักเรียนในสัปดาห์ที่ 1 ถึงสัปดาห์ที่ 4	149
รูป 5.40 เมทริกซ์ของแผนภาพการกระจายระหว่างคะแนนสอบย่อยของนักเรียน	
ในสัปดาห์ที่ 1 ถึงสัปดาห์ที่ 4 ที่มีการปรับลักษณะของ	
สารสนเทศในสามเหลี่ยมบนและสามเหลี่ยมล่างของแผนภาพด้วย	
อาร์กิวเมนต์ <code>lower.panel</code> และ <code>upper.panel</code>	151
รูป 5.41 แผนภาพการกระจายของผลสัมฤทธิ์โดยรวมในรายวิชาที่สอนแบบ online	
กับระดับความจดจ่อกับเนื้อหาและกิจกรรมในชั้นเรียนของผู้เรียน	152
รูป 5.42 แผนภาพการกระจายของผลสัมฤทธิ์โดยรวมในรายวิชาที่สอนแบบ online	
กับระดับความจดจ่อกับเนื้อหาและกิจกรรมในชั้นเรียนของผู้เรียน	
ที่มีการใส่ตัวรบกวนด้วยฟังก์ชัน <code>jitter()</code>	153
รูป 5.43 แผนภาพกล่องแสดงความสัมพันธ์ระหว่างผลสัมฤทธิ์โดยรวม	
กับความจดจ่อกับเนื้อหา และกิจกรรมในชั้นเรียนของผู้เรียน	154

รูป 5.44 แผนภาพกล่องที่มีการซ้อนแผนภาพการกระจายเพื่อแสดง ความสัมพันธ์ระหว่างผลสัมฤทธิ์โดยรวมกับความจดจ่อกับเนื้อหา และกิจกรรมในชั้นเรียนของผู้เรียน	155
รูป 5.45 แผนภาพการกระจายระหว่างความเข้าใจในเนื้อหาของการเรียนแต่ละครั้ง กับความจดจ่อกับเนื้อหาและกิจกรรมในชั้นเรียน	156
รูป 5.46 แผนภาพโมเสกระหว่างความเข้าใจในเนื้อหาของการเรียนแต่ละครั้ง กับความจดจ่อกับเนื้อหา และ กิจกรรมในชั้นเรียน	157
รูป 5.47 แผนภาพการกระจายระหว่างอายุเฉลี่ยของผู้สอนกับเพศของผู้เรียน	158
รูป 5.48 เปรียบเทียบสัมประสิทธิ์สหสัมพันธ์ของเพียร์สันและสเปียร์แมน ภายใต้สถานการณ์ที่ตัวแปรสองตัวมีความสัมพันธ์แบบทางเดียว	160
รูป 5.49 เมทริกซ์แผนภาพการกระจายและสหสัมพันธ์ของการสร้างบรรยากาศการเรียนรู้ ของผู้สอนในรายวิชาที่จัดการสอนแบบ online	162
รูป 6.1 การจัดกระทำข้อมูลลักษณะต่าง ๆ ด้วยฟังก์ชันของ <code>pacakge-dplyr</code>	170
รูป 6.2 การทำงานของตัวดำเนินการ <code>piping (%>%)</code>	176
รูป 6.3 การแจกแจงความถี่ของตัวแปร Gender จากชุดข้อมูล <code>cleaning.csv</code>	183
รูป 6.4 การแจกแจงความถี่ของตัวแปรเพศเมื่อมีการแปลงรหัสข้อมูลแล้ว	184
รูป 6.5 ฮิสโทแกรมของคะแนนผลสัมฤทธิ์โดยรวมของผู้เรียนในรายวิชา online	187
รูป 6.6 แผนภาพกล่องของตัวแปรผลที่เกิดขึ้นกับผู้เรียนในรายวิชาที่สอนแบบ onsite ที่สร้างจากตัวแปรย่อย 4 ตัวได้แก่ <code>outcome.onsite1-outcome.onsite4</code>	189
รูป 6.7 แสดงการเปรียบเทียบผลลัพธ์จากการรวมคะแนนข้อคำถามทั้ง 5	192
รูป 7.1 ฮิสโทแกรมและแผนภาพกล่องแสดงร้อยละของค่าสูญหาย ในตัวแปรของชุดข้อมูล <code>onlinelearning_miss.csv</code>	201
รูป 7.2 แผนภาพแสดงจำนวนค่าสูญหายในตัวแปร	203
รูป 7.3 แผนภาพแสดงจำนวนค่าสูญหายในตัวแปร จำแนกตามประเภทสถานศึกษา	206
รูป 7.4 แผนภาพแสดงรูปแบบและปริมาณของค่าสูญหายในหน่วยข้อมูล	210
รูป 7.5 เมทริกซ์ค่าสูญหายที่สร้างจากฟังก์ชัน <code>vis_miss()</code>	213
รูป 7.6 แผนภาพ MVP ของชุดข้อมูล <code>cleaning.csv</code>	214
รูป 7.7 แผนภาพ MVP ของชุดข้อมูล <code>onlinelearning_miss.csv</code>	215
รูป 7.8 แผนภาพ <code>upset</code> ของค่าสูญหายที่สร้างจากฟังก์ชัน <code>gg_miss_upset()</code>	216

รูป 7.9 ตัวอย่างของชุดข้อมูล D กับเมทริกซ์ M	221
รูป 7.10 ฮิสโทแกรมของ MathAtt เปรียบเทียบระหว่างการแจกแจงของ ข้อมูลสมบูรณ์ (สีอ่อน) และการแจกแจงของข้อมูลสูญหายแบบ MCAR (สีเข้ม)	223
รูป 7.11 เปรียบเทียบค่าสถิติของตัวแปร MathAtt ระหว่างสถานการณ์ที่มีข้อมูลสมบูรณ์ กับการสูญหายแบบ MCAR จำนวนร้อยละ 30	224
รูป 7.12 แผนภาพการกระจายระหว่างตัวแปร MathAtt ที่มีค่าสูญหายแบบ MCAR กับ ตัวแปรอื่น ๆ ในชุดข้อมูล เปรียบเทียบระหว่างข้อมูลที่สังเกตได้กับข้อมูลที่สูญหาย	224
รูป 7.13 แผนภาพการกระจายเปรียบเทียบความสัมพันธ์ระหว่างตัวแปร ในชุดข้อมูลที่สังเกตได้กับข้อมูลสูญหายของตัวแปร MathAtt โดยค่าสูญหายที่เกิดขึ้นเป็นแบบ MAR คิดเป็นจำนวนร้อยละ 30	225
รูป 7.14 ฮิสโทแกรมของ MathAtt เปรียบเทียบระหว่างการแจกแจงของ ข้อมูลสมบูรณ์ (สีอ่อน) และการแจกแจงของข้อมูลสูญหายแบบ MNAR (สีเข้ม)	226
รูป 7.15 แผนภาพการกระจายระหว่างตัวแปร MathAtt ที่มีค่าสูญหายแบบ MNAR จำนวนร้อยละ 30 กับตัวแปรอื่น ๆ ในชุดข้อมูล เปรียบเทียบ ระหว่างข้อมูลที่สังเกตได้ กับข้อมูลที่สูญหาย	227
รูป 7.16 แผนภาพ margin ระหว่างเจตคติทางการเรียนคณิตศาสตร์ และ คะแนนสอบวิชาคณิตศาสตร์ ของชุดข้อมูล cleaning.csv	235
รูป 7.17 เมทริกซ์ของแผนภาพ margin ระหว่าง การสร้างบรรยากาศการเรียนรู้ของผู้สอน พฤติกรรมการณ์เรียนของผู้เรียน และผลที่เกิดขึ้นจากการเรียนของผู้เรียน ในชุดข้อมูล onlinelearning_miss.csv	236
รูป 7.18 แผนที่ความร้อนของค่าสูญหาย สำหรับวิเคราะห์ความสัมพันธ์ระหว่างค่าสูญหาย ของตัวแปรในชุดข้อมูลกับที่ตั้งสถานศึกษา	237
รูป 7.19 แผนที่ความร้อนของค่าสูญหาย สำหรับวิเคราะห์ความสัมพันธ์ระหว่างค่าสูญหาย ของตัวแปรในชุดข้อมูลกับระดับของการสร้างบรรยากาศการเรียนรู้ของผู้เรียน ในรายวิชาที่สอนแบบ online	238
รูป 7.20 ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานของคะแนนเจตคติ ในการเรียนวิชาคณิตศาสตร์ (MathAtt) ระหว่างสถานการณ์จำลอง ที่มีข้อมูลสมบูรณ์ (รูปซ้าย) กับที่มีการสูญหายแบบ MCAR โดยแก้ปัญหา ด้วยการตัดหน่วยข้อมูล (รูปกลาง) และที่ทดแทนค่าสูญหายด้วยค่าเฉลี่ย (รูปขวา)	246

รูป 7.21 แผนภาพการกระจายแสดงความสัมพันธ์ระหว่างคะแนนเจตคติ ทางการเรียนคณิตศาสตร์ (Att) และผลสัมฤทธิ์ทางการเรียนวิชา คณิตศาสตร์ (Math) ในชุดข้อมูล cleaning.csv เมื่อมีการทดแทนค่าสูญหายในตัวแปร Att ด้วยค่าเฉลี่ย	247
รูป 7.22 แผนภาพการกระจายแสดงความสัมพันธ์ระหว่างคะแนนเจตคติ ทางการเรียนคณิตศาสตร์ (Att) และผลสัมฤทธิ์ทางการเรียน วิชาคณิตศาสตร์ (Math) ในชุดข้อมูล cleaning.csv เมื่อมีการทดแทน ค่าสูญหายในตัวแปร Att ด้วยการวิเคราะห์ความถดถอย	249
รูป 7.23 แผนภาพจุด (dotplot) เปรียบเทียบการแจกแจงของค่าสังเกตในตัวแปร Att ของชุดข้อมูล cleaning.csv ระหว่างข้อมูลที่สังเกตค่าได้ (!NA) และ ข้อมูลสูญหายที่ได้รับการทดแทนด้วยการวิเคราะห์ความถดถอย (NA)	251
รูป 7.24 อัลกอริทึม random forest	254
รูป 7.26 แผนภาพจุด และแผนภาพการกระจาย เปรียบเทียบการแจกแจงของตัวแปร Att ระหว่างค่าที่สังเกตได้และค่าทดแทนค่าสูญหายของตัวแปร Att ที่ทดแทนด้วย การวิเคราะห์ความถดถอยแบบสเตคสติก	256
รูป 7.27 แผนภาพจุด และแผนภาพการกระจาย เปรียบเทียบการแจกแจงของตัวแปร Att ระหว่างค่าที่สังเกตได้และค่าทดแทนค่าสูญหายของตัวแปร Att ที่ทดแทนด้วย predictive mean matching	258
รูป 7.28 แผนภาพการกระจาย เปรียบเทียบการแจกแจงของตัวแปร Att ระหว่าง ค่าที่สังเกตได้และค่าทดแทนค่าสูญหายของตัวแปร Att ที่ทดแทนด้วย predictive mean matching โดยกำหนดจำนวน candidate donors (k) เท่ากับ 1 และ 10 หน่วย	262
รูป 7.29 ตัวอย่างรูปแบบการสูญหายของข้อมูล	265
รูป 7.30 แผนภาพจุด และแผนภาพการกระจายเปรียบเทียบการแจกแจงของตัวแปร Att ระหว่างค่าที่สังเกตได้และค่าทดแทนค่าสูญหาย ภายหลังจากการทดแทน ค่าสูญหายแบบหลายตัวแปรด้วยอัลกอริทึม FCS และ โมเดล predictive mean matching	267

รูป 7.30 มโนทัศน์ของการทดแทนค่าสูญหายแบบหลายค่า	268
รูป 7.31 trace plot ของค่าเฉลี่ย และส่วนเบี่ยงเบนมาตรฐานของค่าทดแทน ค่าสูญหายแบบหลายค่า ของตัวแปร Math และ Att ที่มีการทดแทน จำนวน 20 ครั้ง ในชุดข้อมูล cleaning.csv ด้วยฟังก์ชัน mice()	271
รูป 7.32 ตัวอย่าง trace plot ที่บ่งชี้ถึงปัญหาในการทดแทนค่าสูญหาย	272
รูป 7.33 โค้งความหนาแน่นของข้อมูลที่เกิดขึ้นค่าได้ และค่าทดแทน ค่าสูญหายจำนวน 20 ชุด ของตัวแปร Math และ Att ที่สร้างจากฟังก์ชัน densityplot()	273
รูป 7.34 โค้งความหนาแน่นของข้อมูลที่เกิดขึ้นค่าได้ และค่าทดแทน ค่าสูญหายจำนวน 20 ชุด ของตัวแปร Math และ Att จำแนกตามเพศของนักเรียน และภูมิฐานะของนักเรียน	274
รูป 7.35 แผนภาพจุดที่สร้างจากฟังก์ชัน stripplot() แสดงการเปรียบเทียบ การแจกแจงของข้อมูลที่เกิดขึ้นค่าได้กับค่าทดแทนค่าสูญหายของตัวแปร Math	275
รูป 7.36 แผนภาพจุดที่สร้างจากฟังก์ชัน stripplot() แสดงการเปรียบเทียบ การแจกแจงของข้อมูลที่เกิดขึ้นค่าได้กับค่าทดแทนค่าสูญหายของตัวแปร Math จำแนกตามภูมิฐานะของนักเรียน	276
รูป 7.37 แผนภาพกล่องเปรียบเทียบการแจกแจงของข้อมูลที่เกิดขึ้นค่าได้ (สีน้ำเงิน) กับการแจกแจงของค่าทดแทนค่าสูญหาย (สีแดง) ของตัวแปร Math และ Att ที่ได้จากการทดแทนค่าสูญหาย จำนวน 20 ครั้ง	277
รูป 7.38 แผนภาพกล่องเปรียบเทียบการแจกแจงของข้อมูลที่เกิดขึ้นค่าได้ กับการแจกแจงของค่าทดแทนค่าสูญหายในตัวแปร Math จำแนกตามภูมิฐานะของนักเรียน	278

สารบัญตาราง

ตาราง 2.1	คะแนนสอบกลางภาคและปลายภาคของนิสิต	18
ตาราง 2.2	ข้อมูลส่วนตัวของนักเรียน	22
ตาราง 4.1	ฟังก์ชันพื้นฐานสำหรับสำรวจโครงสร้างของชุดข้อมูล	54
ตาราง 4.2	เปรียบเทียบความสามารถระหว่างฟังก์ชัน <code>join</code> กับฟังก์ชัน <code>merge()</code>	69
ตาราง 4.3	ตัวอย่างชุดข้อมูลรูปแบบยาวของจำนวนนักศึกษาในปีการศึกษา 2560 จำแนกตามระดับการศึกษา และ เพศ	71
ตาราง 4.3	ชุดข้อมูลรูปแบบกว้างของจำนวนนักศึกษาในปีการศึกษา 2560 จำแนกตามระดับการศึกษา และ เพศ	71
ตาราง 4.4	คะแนนเข้าวิปัญาของหน่วยข้อมูล	74
ตาราง 4.5	สัญลักษณ์และความหมายของนิพจน์ปกติพื้นฐาน	77
ตาราง 4.6	ข้อมูลการยืมหนังสือของห้องสมุด	81
ตาราง 4.7	รายได้เฉลี่ยต่อหัวของประชากรและอันดับผลการสอบ O-NET ของโรงเรียนในจังหวัด	85
ตาราง 4.8	รายได้เฉลี่ยต่อหัวต่อปี และอันดับของโรงเรียนโดยเฉลี่ย จำแนกตามจังหวัดของประเทศไทย	86
ตาราง 5.1	ตัวอย่างอาร์กิวเมนต์ทั่วไปที่สามารถกำหนดได้ในฟังก์ชันกราฟิกระดับสูง	101
ตาราง 5.2	ฟังก์ชันกราฟิกระดับล่างที่สำคัญและคำอธิบาย	105
ตาราง 5.3	ฟังก์ชันสำหรับวิเคราะห์คำนวณค่าสถิติพื้นฐานของตัวแปรในโปรแกรม R	128
ตาราง 7.1	วิธีทดแทนค่าสูญหายที่สามารถเลือกใช้ได้กับฟังก์ชัน <code>mice()</code>	263

คำนำ

โปรแกรม R เป็นภาษาคอมพิวเตอร์ยุคใหม่ที่ถูกพัฒนาขึ้นให้มีความสามารถอย่างหลากหลาย ซึ่งกล่าวได้ว่าครอบคลุมการดำเนินงานทั้งหมดทางด้านสถิติและวิทยาการข้อมูล ทั้งความสามารถในการเก็บเกี่ยวและนำเข้าข้อมูลได้หลากหลายประเภท การจัดระเบียบและจัดกระทำข้อมูล มีฟังก์ชันสำเร็จรูปจำนวนมากที่ช่วยสำหรับการควบคุมการทำงาน การประมวลผลทางคณิตศาสตร์และสถิติ ทั้งฟังก์ชันที่ถูกติดตั้งมาพร้อมการติดตั้งโปรแกรมในครั้งแรก และฟังก์ชันที่ผู้ใช้สามารถดาวน์โหลดได้เพิ่มเติมจาก package เสริมที่มีให้เลือกใช้จำนวนมาก สามารถเชื่อมต่อและทำงานร่วมกับโปรแกรมวิเคราะห์ข้อมูลเฉพาะทางอื่น ๆ ได้หลายโปรแกรม และยังเป็นโปรแกรมที่มีประสิทธิภาพสูงมากสำหรับการสร้างทัศนภาพข้อมูล R เป็นโปรแกรมภาษาที่มีความยืดหยุ่นในการทำงานสูง แต่มีหลักไวยากรณ์ที่ไม่ซับซ้อนง่ายต่อการเรียนรู้ ปัจจุบัน R จึงเป็นโปรแกรมที่ได้รับความนิยมและได้รับการยอมรับอย่างกว้างขวางว่าเหมาะสำหรับการนำไปใช้สำหรับการดำเนินงานทางด้านสถิติและวิทยาการข้อมูลในทุกระดับ

หนังสือเล่มนี้เป็นเล่มที่หนึ่งในชุดหนังสือสถิติและวิทยาการข้อมูลทางการศึกษา โดยเป็นการถ่ายทอดความรู้จากประสบการณ์ตรงที่ได้จากการเรียนรู้และทำงานที่เกี่ยวข้องกับสถิติและวิทยาการข้อมูลมาเป็นระยะเวลากว่า 15 ปี ของผู้เขียนเอง โดยได้ผ่านการตรวจสอบแนะนำจากผู้ทรงคุณวุฒิที่แต่งตั้งโดยศูนย์นวัตกรรมทางการศึกษาสื่อสิ่งพิมพ์และสื่อออนไลน์ คณะครุศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย เนื้อหาภายในเล่มนี้เป็นจุดเริ่มต้นของการทำงานทางด้านสถิติและวิทยาการข้อมูลทางการศึกษาโดยเน้นเสนอแนะทัศนคติและวิธีการใช้โปรแกรม R ที่เกี่ยวข้องกับการจัดเตรียมและแก้ปัญหาต่าง ๆ ที่เกิดขึ้นในข้อมูล เพื่อให้ได้ชุดข้อมูลที่มีความพร้อมและเหมาะสมสำหรับการนำไปวิเคราะห์ตามวัตถุประสงค์ต่าง ๆ ของงาน หนังสือเล่มนี้ ผู้เขียนออกแบบการดำเนินเรื่องในหนังสือให้เป็นลำดับจากง่ายไปยาก มีการอธิบายมโนทัศน์หรือทฤษฎีที่จำเป็นควบคู่กับการแสดงตัวอย่างผู้อ่านทั้งที่มีและไม่มีประสบการณ์ในการใช้โปรแกรม R สามารถใช้ประกอบการศึกษาและฝึกหัดเพื่อให้ความรู้และทักษะปฏิบัติได้โดยง่าย โดยผู้อ่านที่ยังไม่เคยมีประสบการณ์ในการใช้โปรแกรม R มาก่อน แนะนำให้ศึกษาเนื้อหาตั้งแต่บทที่ 1 ส่วนผู้อ่านที่มีประสบการณ์ในการใช้โปรแกรม R มาแล้วสามารถศึกษาขอบเขตและคำแนะนำต่าง ๆ ในการใช้หนังสือได้จากบทที่ 1 จากนั้นสามารถข้ามไปศึกษาดังแต่บทที่ 3 เป็นต้นไป

ผู้เขียนขอกราบขอบพระคุณคุณครูและอาจารย์ทุกท่านในทุกระดับที่ได้ช่วยกันประสิทธิ์
ประสาทวิชาความรู้ให้ผู้เขียนมีความรู้และทักษะจำเป็นที่สามารถนำมาใช้ทำงานได้ในปัจจุบัน
ขอบพระคุณกัลยาณมิตรทุกท่านในภาควิชาวิจัยและจิตวิทยาการศึกษา คณะครุศาสตร์ จุฬาลงกรณ์
มหาวิทยาลัย ที่ให้สนับสนุนทางด้านการแก่ผู้เขียนมาโดยตลอด กราบขอบพระคุณคุณป้า รอง
ศาสตราจารย์ แพทย์หญิง ศิริรีน และครอบครัววสินธุ์ศักดิ์ ที่ให้ความเอาใจใส่ ช่วยเหลือดูแลด้วยความ
เมตตา และสนับสนุนผู้เขียนในทุก ๆ ด้านจนถึงปัจจุบัน และสุดท้ายขอกราบขอบพระคุณคุณย่าเอ็ง
คุณแม่ศุภศรี คุณพ่อสมศักดิ์ พี่สาว พันโท แพทย์หญิง ชูติกา และครอบครัวศรีสุทธียากร ที่ได้เลี้ยงดู
อบรม สันับสนุน เป็นกำลังใจ และเป็นแบบอย่างที่ดีในการดำเนินชีวิตให้แก่ผู้เขียนตลอดมา ซึ่ง
ทั้งหมดล้วนส่วนสำคัญยิ่งที่ทำให้หนังสือเล่มนี้สามารถสำเร็จลุล่วงลงได้ ความดีทั้งหมดที่มีในหนังสือ
เล่มนี้ผู้เขียนขอยกให้แก่ผู้มีพระคุณทุกท่าน ส่วนความผิดพลาดใดที่อาจเกิดขึ้นในหนังสือเล่มนี้ ผู้
เขียนขอนอมนับและจะนำไปแก้ไขปรับปรุงให้สมบูรณ์ยิ่งขึ้นในโอกาสต่อไป

สิระโชติ ศรีสุทธียากร
พฤษภาคม 2564



บทนำ

เล่มนี้มีอะไรบ้าง?

R เป็นภาษาคอมพิวเตอร์ยุคใหม่ที่ถูกพัฒนาขึ้นให้มีความสามารถอย่างหลากหลาย และมีประสิทธิภาพสูงและดีมากสำหรับการทำงานทางด้านสถิติและวิทยาการข้อมูล โปรแกรม R ได้รับการริเริ่มพัฒนาขึ้นโดยผู้พัฒนาที่เป็นนักสถิติสองท่าน ได้แก่ Ross Ihaka และ Robert Gentleman จาก University of Auckland ประเทศนิวซีแลนด์ โดยพัฒนาต่อยอดมาจากภาษา S และ S+ และได้ทำการเผยแพร่ให้บุคคลทั่วไปได้ใช้งานตั้งแต่ปี ค.ศ. 1997 ภายใต้สัญญาอนุญาตสาธารณะทั่วไปของกนู (GNU General Public License) โปรแกรม R จัดเป็นโปรแกรมประเภท open source ซึ่งมีลักษณะเป็นโปรแกรมที่เผยแพร่ให้บุคคลทั่วไปมีสิทธิในการเข้าใช้งาน และพัฒนาโปรแกรมอย่างอิสระตามความต้องการโดยไม่เสียค่าใช้จ่ายไม่ว่าจะเป็นการใช้งานทั่วไป การแก้ไขปรับปรุง หรือพัฒนาต่อยอดโปรแกรม (Maindonale and Bruan, 2010; Field, Mild and Field, 2012; Schumacker, 2012; Brundon and Comber, 2013)

R เป็นโปรแกรมที่ดีมากและมีประสิทธิภาพสูงสำหรับการทำงานทางด้านสถิติและวิทยาการข้อมูล โดยเป็นโปรแกรมแบบหลายแพลตฟอร์ม (multiple platform) กล่าวคือสามารถติดตั้งและทำงานบนคอมพิวเตอร์ที่มีระบบปฏิบัติการที่สำคัญได้ทุกระบบ ได้แก่ ระบบปฏิบัติการ Windows, Mac OS, Linux, Unix รวมทั้ง Chrome OS จุดเด่นนี้ทำให้โปรแกรม R สามารถเข้าถึงผู้ใช้งานได้อย่างทั่วถึง นอกจากนี้โปรแกรม R ยังมีความสามารถที่หลากหลายครอบคลุมการดำเนินงานโดยเฉพาะด้านสถิติและวิทยาการข้อมูล จึงทำให้ปัจจุบันโปรแกรม R เป็นโปรแกรมหนึ่งที่ได้รับคามนิยมเป็นอย่างสูงจากผู้ใช้งานทั่วโลก

ผู้เขียนได้ลองจัดกลุ่มสามารถหลักของโปรแกรมพบว่า อาจจำแนกการนำเข้าข้อมูล (importing data) การจัดระเบียบและจัดกระทำข้อมูล (tidying and manipulating data) การคำนวณทางคณิตศาสตร์ (mathematical computations) การวิเคราะห์ข้อมูลและพัฒนาโมเดลทางสถิติ (data analysis and statistical modelling) และการนำเสนอข้อมูลทางสถิติ (data presentation) โดยเมื่อพิจารณาความสามารถในแต่ละด้านข้างต้น พบว่ามีจุดเด่นหลายประการ ดังนี้

1. สามารถนำเข้าข้อมูลได้หลายประเภท ด้วยวิธีการที่หลากหลาย โดยสามารถดำเนินการได้ทั้งการป้อนข้อมูลโดยตรง การนำเข้าจากไฟล์ข้อมูลประเภทต่าง ๆ ซึ่งครอบคลุมทุกประเภทของไฟล์ข้อมูล การนำเข้าข้อมูลจากฐานข้อมูล ไปจนถึงการดาวน์โหลดและเก็บเกี่ยวข้อมูลจากเว็บไซต์ (web scrapping)
2. เป็นโปรแกรมที่มีความสามารถสูงในการจัดระเบียบและจัดกระทำข้อมูล โดยมีเครื่องมือที่มีประสิทธิภาพสูงมากหลายตัวที่ช่วยจัดระเบียบตารางข้อมูลในอยู่ในรูปแบบที่เหมาะสมสำหรับการวิเคราะห์ข้อมูลในงานต่าง ๆ และช่วยจัดกระทำข้อมูล เช่น การแบ่งส่วนย่อยของชุดข้อมูล การแปลงรหัสหรือแปลงค่าของตัวแปร หรือการจัดลำดับของข้อมูลตามค่าของตัวแปรที่กำหนด เป็นต้น
3. โปรแกรม R มีฟังก์ชันสำเร็จรูปทั้งทางคณิตศาสตร์และสถิติจำนวนมากที่ถูกติดตั้งมาพร้อมกับการติดตั้งโปรแกรมในครั้งแรก ซึ่งสามารถเรียกใช้เพื่อช่วยให้การดำเนินงานต่าง ๆ สามารถทำได้โดยง่ายและมีประสิทธิภาพ และนอกจากนี้ยังมีฟังก์ชันสำเร็จรูปอีกจำนวนมากจาก package เสริมต่าง ๆ บน CRAN server ที่ถูกพัฒนาขึ้นโดยนักวิชาการ หรือนักพัฒนาจากทั่วโลก โดยปัจจุบันมี package จำนวนมากกว่า 10,000 ตัว บน server ดังกล่าวที่ผู้ใช้สามารถดาวน์โหลดและติดตั้งมาใช้เพื่อเสริมความสามารถในการดำเนินงานของโปรแกรม คุณสมบัตินี้ทำให้การดำเนินงานทางด้านสถิติและวิทยาการข้อมูลด้วยโปรแกรม R สามารถดำเนินงานได้ยืดหยุ่นมาก สามารถปรับแต่งและเลือกวิธีการดำเนินงานให้มีความเหมาะสม ทันสมัย และมีประสิทธิภาพสูงที่สุด โดยมีข้อจำกัดในการดำเนินงานที่น้อย
4. สามารถเชื่อมต่อและทำงานร่วมกับโปรแกรมวิเคราะห์ข้อมูลเฉพาะทางอื่น ๆ ได้หลายโปรแกรม เช่น Mplus, MLWins, OpenBUGS, JAGS หรือ Stan ซึ่งช่วยให้สามารถใช้ความสามารถของโปรแกรกดังกล่าวได้ และยกระดับความสามารถของโปรแกรม R ให้เทียบเท่าและในบางกรณีอาจเหนือกว่าการใช้โปรแกรมวิเคราะห์ข้อมูลดังกล่าวโดยตรง
5. มีความประสิทธิภาพสูงมากสำหรับการทำงานด้านกราฟิกหรือการสร้างทัศนภาพข้อมูล (data visualization) โดยเป็นโปรแกรมที่มี package หลายตัวที่ถูกพัฒนาขึ้นสำหรับสร้างทัศนภาพข้อมูลโดยเฉพาะ เช่น graphics, ggplot2, lattice, boken, rmarkdown, flexdashboard และ shiny ซึ่งช่วยให้ผู้ใช้สามารถสร้างทัศนภาพข้อมูลได้อย่างหลากหลาย

ทั้งในรูปแบบทัศนภาพข้อมูลเชิงสถิต (static data visualization) ทัศนภาพข้อมูลเชิงพลวัต (dynamic data visualization) และทัศนภาพข้อมูลเชิงปฏิสัมพันธ์ (interactive data visualization)

6. เป็นโปรแกรมภาษาที่ง่ายต่อการเรียนรู้และใช้งาน นอกจากนี้ยังมีชุมชนผู้ใช้โปรแกรม R และแหล่งการเรียนรู้ออนไลน์ที่สามารถให้คำตอบแก่ผู้ใช้ได้อย่างกว้างขวางและตรงประเด็น ดังนั้นผู้ใช้โปรแกรม R หรือผู้ที่ต้องการศึกษา R ที่ไม่ได้มีพื้นฐานการเขียนโปรแกรมมาก่อนจึงสามารถเรียนรู้ภาษา R เป็นภาษาแรกได้โดยง่าย

1.1 ขอบเขตเนื้อหาของหนังสือ

เนื้อหาของหนังสือเล่มนี้มีขอบเขตที่เกี่ยวข้องกับการใช้โปรแกรม R สำหรับการดำเนินงานพื้นฐานทางด้านสถิติและวิทยาการข้อมูล โดยเนื้อหาส่วนใหญ่เน้นไปที่มินิทัศน์ ทฤษฎี และวิธีการใช้โปรแกรม R สำหรับการเตรียมและแก้ปัญหาที่เกิดขึ้นในชุดข้อมูล เพื่อให้ได้ชุดข้อมูลที่มีความพร้อมที่จะนำไปวิเคราะห์ตามวัตถุประสงค์หลักของการดำเนินงานได้อย่างเหมาะสม เนื้อหาอีกส่วนจะกล่าวถึงการดำเนินการวิเคราะห์เพื่อสำรวจข้อมูลด้วยโปรแกรม R โดยจะกล่าวถึงวิธีการวิเคราะห์ทั้งที่ใช้สถิติวิเคราะห์ และใช้ทัศนภาพข้อมูล ผู้อ่านที่เคยใช้ R มาก่อนสามารถใช้หนังสือเล่มนี้ได้โดยในบทที่ 2 จะกล่าวถึงความรู้พื้นฐานที่จำเป็นเกี่ยวกับการใช้โปรแกรม R โดยเริ่มตั้งแต่การดาวน์โหลด ติดตั้ง แนะนำสภาพแวดล้อมของโปรแกรม และอธิบายมินิทัศน์ที่จำเป็นต้องทราบสำหรับการดำเนินงานทางด้านสถิติและวิทยาการข้อมูลภายใต้สภาพแวดล้อมของโปรแกรม R ส่วนผู้อ่านที่มีพื้นฐานการใช้ R มาแล้วสามารถข้ามไปศึกษาเนื้อหาตั้งแต่บทที่ 3 เป็นต้นไป

บทที่ 3 ของหนังสือจะกล่าวถึงรายละเอียดเกี่ยวกับการนำข้อมูลลักษณะต่าง ๆ เข้าสู่โปรแกรม อย่างไรก็ตามข้อมูลที่นำเข้าสู่โปรแกรมนั้นไม่จำเป็นที่จะต้องอยู่ในรูปแบบที่มีความพร้อมหรือเหมาะสำหรับการนำไปวิเคราะห์ตามวัตถุประสงค์หลักของการดำเนินงานในทันที เนื้อหาในบทที่ 4 จะกล่าวถึงมินิทัศน์ของตารางข้อมูลจัดระเบียบ (tidy data) ที่เป็นลักษณะของตารางข้อมูลที่เหมาะสำหรับการดำเนินงานวิเคราะห์ข้อมูล และจะกล่าวถึงการจัดระเบียบตารางข้อมูลให้นำเข้า เพื่อให้อยู่ในรูปแบบตารางข้อมูลจัดระเบียบดังกล่าว

ตารางข้อมูลจัดระเบียบเป็นเงื่อนไขจำเป็น แต่ยังไม่เพียงพอที่จะทำให้การดำเนินงานวิเคราะห์ข้อมูลสามารถทำได้ถูกต้องและมีประสิทธิภาพ เพราะถึงแม้ว่ารูปแบบตารางข้อมูลจะมีความพร้อม แต่ถ้าข้อมูลภายในตารางยังไม่มี ความเหมาะสม เช่น ยังไม่ใช่ข้อมูลของหน่วยข้อมูล หรือตัวแปรที่อยู่ภายใต้ขอบเขตของการวิเคราะห์ หรือข้อมูลภายในตารางมีปัญหาเชิงเทคนิคเกิดขึ้น โดยส่วนใหญ่ปัญหาที่พบคือ ปัญหาค่าสูญหาย (missing values) ค่าผิดปกติ (outliers) การแจกแจงของข้อมูลไม่สอดคล้องกับข้อตกลงเบื้องต้นของสถิติวิเคราะห์ที่ใช้เป็นหลักในการดำเนินงาน หรืออาจมีหน่วยข้อมูล/ตัวแปรที่ไม่เกี่ยวข้องในการดำเนินงาน เมื่อทำการสำรวจข้อมูลแล้วพบว่าปัญหาดังกล่าว ผู้วิเคราะห์จำเป็นต้องดำเนินการแก้ไขปัญหาก็เพื่อให้สามารถรับ

ประกันความแม่นยำ และความน่าเชื่อถือของผลการวิเคราะห์ที่จะได้รับ เนื้อหาในบทที่ 6 จะกล่าวถึงโมเดลและวิธีการที่จำเป็นสำหรับการจัดกระทำข้อมูลด้วยโปรแกรม R เพื่อแก้ปัญหาต่างหรือปรับเปลี่ยน/ดัดแปลงข้อมูลภายในตาราง ให้สอดคล้องกับการดำเนินงานของผู้วิเคราะห์ เช่น การคัดเลือกตัวแปร การคัดกรองหน่วยข้อมูล การยุบหรือกระจายหน่วยข้อมูล การสร้างตัวแปรใหม่ด้วยการรวมคะแนน หรือแปลงค่าสังเกตของตัวแปร เป็นต้น

การดำเนินการจัดกระทำข้อมูลในข้างต้นสามารถใช้เพื่อแก้ปัญหาที่เกิดขึ้นในชุดข้อมูลได้เป็นส่วนใหญ่ แต่ยังไม่สามารถใช้เพื่อแก้ปัญหาค่าสูญหายที่เกิดขึ้นในชุดข้อมูลด้วย ปัญหาค่าสูญหายเป็นปัญหาที่มักเกิดขึ้นอย่างสม่ำเสมอ และหลีกเลี่ยงได้ค่อนข้างยาก จากประสบการณ์ของผู้เขียนพบว่า ผู้วิเคราะห์มือใหม่ส่วนใหญ่มักจะเลือกใช้วิธีที่ไม่เหมาะสม ซึ่งทำให้ผลการวิเคราะห์ที่ได้จากข้อมูลดังกล่าวมีความลำเอียง และไม่มีประสิทธิภาพ เนื้อหาในบทที่ 7 จะกล่าวถึง โมเดลของวิธีการวิเคราะห์และทดแทนค่าสูญหาย โดยจะกล่าวถึงโมเดล และประเภทของค่าสูญหายที่จำแนกตามลักษณะของการเกิด รวมทั้งกล่าวถึงผลกระทบของการเกิดค่าสูญหายแต่ละประเภทที่มีต่อผลการวิเคราะห์ นอกจากนี้จะให้โมเดลเกี่ยวกับการวิเคราะห์เพื่อสำรวจและค้นหารูปแบบของการเกิดค่าสูญหายภายในชุดข้อมูล และหัวข้อสุดท้ายจะกล่าวถึงวิธีการแก้ปัญหาค่าสูญหายด้วยการทดแทนค่าสูญหายแบบหลายค่า (multiple imputation) ที่เป็นวิธีการสมัยใหม่ที่มีประสิทธิภาพสูงสำหรับการแก้ปัญหาค่าสูญหายดังกล่าว ทั้งนี้เพื่อให้ผลการวิเคราะห์ที่ได้ภายหลังจากการทดแทนค่าสูญหายดังกล่าวมีความแม่นยำและน่าเชื่อถือมากที่สุดเท่าที่จะเป็นไปได้

1.2 ความรู้เบื้องต้นที่จำเป็นสำหรับผู้อ่าน

ผู้อ่านไม่จำเป็นต้องมีความรู้พื้นฐานเกี่ยวกับโปรแกรม R มาก่อน แต่ควรมีพื้นฐานความรู้เกี่ยวกับสถิติพื้นฐานหรือเคยเรียนรายวิชาสถิติพื้นฐานในระดับปริญญาบัณฑิตมาอย่างน้อย 1 รายวิชา นอกจากนี้การมีพื้นฐานทางคณิตศาสตร์ในระดับมัธยมศึกษาตอนปลายจะช่วยให้สามารถทำความเข้าใจเนื้อหาบางส่วนของหนังสือเล่มนี้ได้ดียิ่งขึ้น

1.3 การดาวน์โหลดและติดตั้งโปรแกรม R และ package

ตัวอย่างในหนังสือเล่มนี้ที่มีการประมวลผลด้วยโปรแกรม R ประมวลผลบนโปรแกรม R version 4.0.2 (Taking Off Again) สำหรับผู้อ่านที่ยังไม่มีโปรแกรมสามารถดาวน์โหลดโปรแกรมได้จาก <http://www.r-project.org/> โดยเมื่อเข้าสู่ website ให้คลิกที่คำว่า “download R” เพื่อดาวน์โหลดซอฟต์แวร์จาก CRAN (Comprehensive R Archive Network) โดยให้ดาวน์โหลดตัว base distribution ที่เหมาะสมกับระบบปฏิบัติการของตนเอง แล้วดำเนินการติดตั้งโปรแกรมตามขั้นตอนที่ตัวช่วยการติดตั้งแนะนำ