

จัดการและวิเคราะห์ข้อมูลด้วย Data Science และพื้นฐาน Machine Learning โดยใช้ภาษา Python



การใช้ VS Code
ร่วมกับ Miniconda
สำหรับงาน Data Science



จัดการและวิเคราะห์ข้อมูลแบบ
Data Science โดยใช้
Pandas DataFrame



เปรียบเทียบข้อมูลโดยพล็อตกราฟ
แบบ Visualization ด้วย Matplotlib



ใช้ Machine Learning ทำนายผล
ด้วยโมเดลต่างๆ โดยใช้ Scikit-learn



DOWNLOAD

ดาวน์โหลดโค้ดหนังสือเล่มนี้ได้ที่
<http://www.developerthai.com>

บัญชา ปะสีละเตสัง

ผู้เขียนหนังสือขายดีระดับ Best Seller
ด้าน Programming

จัดการและวิเคราะห์ข้อมูลด้วย Data Science และพื้นฐาน Machine Learning โดยใช้ภาษา Python

โดย บัญชา ปะสีละเตลัง

สงวนลิขสิทธิ์ตามกฎหมาย โดย บัญชา ปะสีละเตลัง © พ.ศ. 2569

ห้ามคัดลอก ลอกเลียน ดัดแปลง ทำซ้ำ จัดพิมพ์ หรือกระทำการอื่นใด โดยวิธีการใดๆ ในรูปแบบใดๆ
ไม่ว่าส่วนหนึ่งส่วนใดของหนังสือเล่มนี้ เพื่อเผยแพร่ในสื่อทุกประเภท หรือเพื่อวัตถุประสงค์ใดๆ
นอกจากจะได้รับอนุญาต

ข้อมูลทางบรรณานุกรมของหอสมุดแห่งชาติ

บัญชา ปะสีละเตลัง.

จัดการและวิเคราะห์ข้อมูลด้วย Data Science และพื้นฐาน Machine Learning โดยใช้ภาษา Python.--

กรุงเทพฯ : ซีเอ็ดดูเคชั่น, 2569.

288 หน้า.

1. ไพธอน (ภาษาคอมพิวเตอร์). I. ชื่อเรื่อง.

005.133

Barcode (e-book) : 9786160857012

ผลิตและจัดจำหน่ายโดย



บริษัท ซีเอ็ดดูเคชั่น จำกัด (มหาชน)
SE-EDUCATION PUBLIC COMPANY LIMITED

เลขที่ 1858/87-90 ถนนเทพรัตน แขวงบางนาใต้ เขตบางนา กรุงเทพฯ 10260

โทรศัพท์ 0-2826-8000

[หากมีคำแนะนำหรือติชม สามารถติดต่อได้ที่ comment@se-ed.com]

.. คำนำ ..

Data Science ถือเป็นวิทยาศาสตร์อีกแขนงหนึ่งที่ว่าด้วยการวิเคราะห์และจัดการข้อมูลในรูปแบบต่างๆ นอกจากนี้ยังเป็นพื้นฐานไปสู่การเรียนรู้เทคโนโลยีแขนงอื่นๆ อีกหลายอย่าง ไม่ว่าจะเป็น Data Mining, Deep Learning, Machine Learning และ Artificial Intelligence (AI) นอกจากนี้ยังสามารถนำ Data Science ไปประยุกต์ใช้กับงานหลายๆ ด้าน เช่น ไอที อุตสาหกรรม และโดยเฉพาะอย่างยิ่งทางด้านธุรกิจ ก็ได้นำ Data Science มาเป็นเครื่องมือสำหรับการวิเคราะห์ข้อมูลต่างๆ กันอย่างแพร่หลาย เพื่อสร้างโอกาสและความได้เปรียบในการแข่งขัน และด้วยเหตุผลดังที่กล่าวมา ในปัจจุบันจึงมีผู้สนใจการเรียนรู้เกี่ยวกับ Data Science เป็นจำนวนมาก

อย่างไรก็ตาม ทางเลือกสำหรับการศึกษา Data Science นั้นมีหลายแนวทาง เช่น การใช้เครื่องมือสำเร็จรูป หรือการเขียนโค้ดด้วยภาษาคอมพิวเตอร์ แต่แหล่งข้อมูลเกี่ยวกับ Data Science ที่เป็นภาษาไทยยังมีค่อนข้างน้อย ดังนั้นผู้เขียนจึงได้จัดทำหนังสือเล่มนี้ขึ้นมา เพื่อใช้ประกอบการเรียนรู้เกี่ยวกับ Data Science โดยใช้ภาษา Python ร่วมกับไลบรารีต่างๆ เช่น NumPy, Pandas, Matplotlib และ Scikit-learn ซึ่งนอกจาก Data Science แล้ว ก็ยังมีกล่าวถึง Machine Learning ขึ้นพื้นฐานเอาไว้ด้วย เพื่อให้ผู้อ่านสามารถนำไปจัดการและวิเคราะห์ข้อมูลในเบื้องต้นได้ รวมทั้งเป็นแนวทางสำหรับการเรียนรู้เพิ่มเติมในระดับที่สูงขึ้นต่อไป

บัญชา ปะสิละเตสัง

banchar.pa@gmail.com

facebook.com/DeveloperThai

<http://www.developerthai.com>



.. สารบัญ ..

บทที่ 1 : การจัดเตรียมเครื่องมือ	11
การติดตั้ง Visual Studio Code.....	11
การติดตั้ง Anaconda/Miniconda	12
สร้าง Environment ของ Anaconda	13
ติดตั้งส่วนเสริม Jupyter บน VS Code.....	14
สร้างไฟล์ Jupyter (.ipynb).....	15
เลือก Kernel สำหรับไฟล์ Jupyter Notebook.....	16
แนะนำการใช้ Jupyter Notebook.....	16
การนำโค้ดของหนังสือมาใช้งาน	19
 บทที่ 2 : nungu Python สำหรับ Data Science	 20
การกำหนดตัวแปร.....	20
การแสดงผลข้อมูล	21
การแทรกค่าตัวแปรในสตริง.....	23
การจัดรูปแบบตัวเลข.....	24
เครื่องหมายชนิดต่าง ๆ.....	25
การกำหนดเงื่อนไข	26
การวนลูป	28
การสร้างและใช้งานฟังก์ชัน	29
ฟังก์ชันเกี่ยวกับตัวเลข	31
ฟังก์ชันเกี่ยวกับสตริง	33

รายการข้อมูลแบบ List	34
รายการข้อมูลแบบ Tuple.....	38
รายการข้อมูลแบบ Dictionary.....	40

บทที่ 3: การใช้ไลบรารี NumPy 44

เกี่ยวกับ NumPy	44
การสร้างอาร์เรย์.....	45
การเข้าถึงสมาชิกของอาร์เรย์.....	50
จัดการอาร์เรย์ด้วยโอเปอเรเตอร์คณิตศาสตร์	54
การสร้างอาร์เรย์แบบ 2 มิติ.....	56
การเข้าถึงสมาชิกของอาร์เรย์ 2 มิติ	58
การจัดการอาร์เรย์ 2 มิติด้วยโอเปอเรเตอร์.....	60
Universal Function ของ NumPy	60

บทที่ 4: เริ่มต้นการใช้ Series และ DataFrame 69

เกี่ยวกับไลบรารี Pandas	69
คลาส Series.....	70
การกำหนดข้อมูลให้แก่ Series.....	71
การกำหนด Index ของ Series	72
คลาส DataFrame	73
การกำหนดข้อมูลให้กับ DataFrame.....	75
ลำดับแถวและชื่อคอลัมน์	81
การกำหนดรูปแบบของตัวเลข	83

บทที่ 5: การเลือกช่วงข้อมูล..... 88

การเลือกข้อมูลโดยระบุแถวและคอลัมน์.....	88
การเลือกข้อมูลโดยใช้พีธรีอปเพอร์ตี at และ iat.....	90
การเลือกข้อมูลโดยวิธี Index Slicing.....	91
การเลือกข้อมูลโดยวิธี Fancy Index และ Index Slicing	92
การใช้ loc และ iloc ร่วมกับ Fancy Index.....	93
การใช้ loc และ iloc ร่วมกับ Index Slicing	96

การใช้ loc และ iloc ร่วมกับ Index Slicing + Fancy Index	99
การเข้าถึงข้อมูลโดยวิธีวนรอบ (Iteration)	100

บทที่ 6 : การคัดกรองด้วยเงื่อนไขและการคำนวณ..... 102

การเลือกข้อมูลแบบ Boolean Index (Masking)	102
การกำหนดเงื่อนไขด้วย where()	106
การเพิ่มคอลัมน์ใหม่.....	108
การคำนวณด้วยเครื่องหมายทางคณิตศาสตร์	110
สร้างฟังก์ชันเพื่อกำหนดเงื่อนไขด้วย apply()	113

บทที่ 7 : ค่าที่สูญหายและฟังก์ชันด้านสถิติ..... 117

ลักษณะแนวแกนของ DataFrame	117
ข้อมูลที่มีค่าว่าง (Missing Value)	118
การตรวจสอบค่าว่าง.....	119
การลบรายการที่เป็นค่าว่าง	121
การเติมค่าแทนที่ค่าสูญหาย	124
การคำนวณด้านสถิติ	127

บทที่ 8 : การใช้ฟังก์ชันอื่นๆ ของ DataFrame..... 132

ฟังก์ชันเกี่ยวกับ Index.....	132
ฟังก์ชันเกี่ยวกับการเปลี่ยนแปลงข้อมูล.....	135
ฟังก์ชันการรวม DataFrame	139
การเรียงลำดับและฟังก์ชันเพิ่มเติมอื่นๆ	146
การคัดเลือกข้อมูลเพื่อใช้กับฟังก์ชัน.....	149

บทที่ 9 : ข้อมูลสตริง วันเวลา และจากไฟล์..... 151

วิธีดำเนินการกับข้อมูลชนิดสตริง	151
วิธีดำเนินการกับข้อมูลชนิดวันเวลา.....	153
การใช้ข้อมูลจากไฟล์ CSV	156
การใช้ข้อมูลจากไฟล์ Excel	163
การใช้ข้อมูลจากไฟล์ฐานข้อมูล.....	167

unที่ 10 : กลุ่มข้อมูลและ Pivot Table..... 170

หลักการ Split-Appl-Combine	170
การจัดกลุ่ม DataFrame โดยใช้คอลัมน์เดียว	171
การจัดกลุ่ม DataFrame โดยใช้หลายคอลัมน์.....	178
จัดการกลุ่มข้อมูลแบบ Aggregate Function.....	182
ลักษณะที่ควรรู้เพิ่มเติมเกี่ยวกับจัดกลุ่ม	184
การสร้าง Pivot Table.....	189

unที่ 11 : การพล็อตกราฟแบบ Visualization 192

ไลบรารี Matplotlib.....	192
การพล็อตกราฟเส้น.....	193
การพล็อตกราฟแท่ง	208
การพล็อตกราฟวงกลม	213
การพล็อตกราฟ Histogram	215
การพล็อตกราฟแบบ Scatter.....	217

unที่ 12 : ML : พื้นฐานและการจัดเตรียมข้อมูล 219

ลักษณะของ Machine Learning	219
หลักการทำนายผลของ Machine Learning	221
โมเดลของแมชชีนเลิร์นนิง (Machine Learning)	225
การใช้ไลบรารี Scikit-learn สำหรับ ML	229
ลักษณะข้อมูลสำหรับ ML.....	231
การกำหนดแหล่งข้อมูลสำหรับ ML.....	234
การตรวจสอบค่าที่สูญหาย (Missing Value).....	235
การเข้ารหัสสตริงให้เป็นตัวเลข (Encoding)	237
การปรับสเกลของข้อมูล.....	242
การแบ่งข้อมูลเป็นชุด Train และ Test.....	246
ความสัมพันธ์ระหว่างคอลัมน์.....	249

บทที่ 13 : ML : การทำนายผลด้วยโมเดลแบบต่าง ๆ (1) 251

โมเดล Simple Regression..... 251

โมเดล Multiple Regression..... 259

โมเดล Polynomial Regression..... 264

บทที่ 14 : ML : การทำนายผลด้วยโมเดลแบบต่าง ๆ (2) 271

โมเดล Logistic Regression..... 271

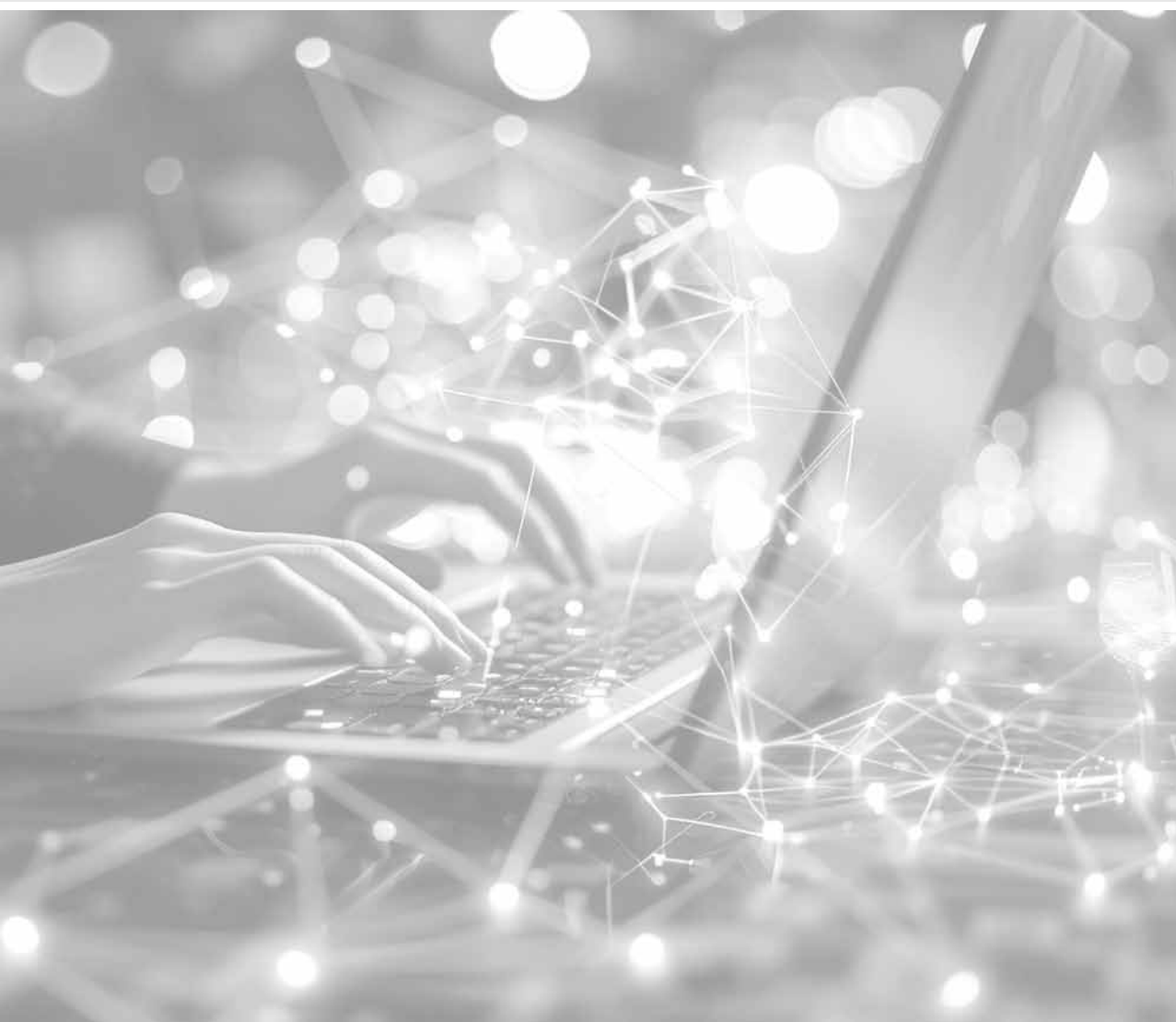
โมเดล K-Nearest Neighbors แบบ Classification..... 275

โมเดล K-Nearest Neighbors แบบ Regression..... 280

โมเดล Decision Tree แบบ Classification..... 282

โมเดล Decision Tree แบบ Regression..... 286

* * *



1

การจัดเตรียมเครื่องมือ

ในการใช้งานเกี่ยวกับ Data Science จำเป็นต้องนำใช้เครื่องมือหลายอย่างมาใช้ร่วมกัน เพื่อการเขียนและทดสอบการใช้งาน ดังนั้นในบทนี้จึงจะเริ่มต้นด้วยการติดตั้งเครื่องมือที่จำเป็นสำหรับการใช้งานในบทต่อไป ได้แก่ Visual Studio Code, Miniconda และ Jupyter Notebook พร้อมแนะนำหลักการใช้งานที่สำคัญในเบื้องต้น

การติดตั้ง Visual Studio Code

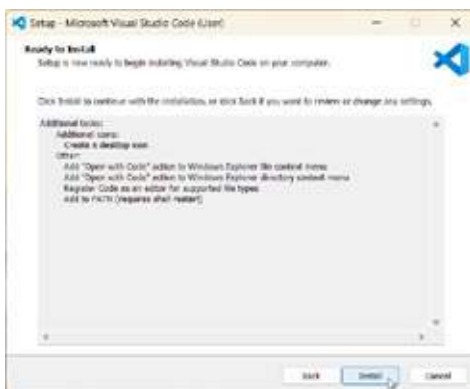
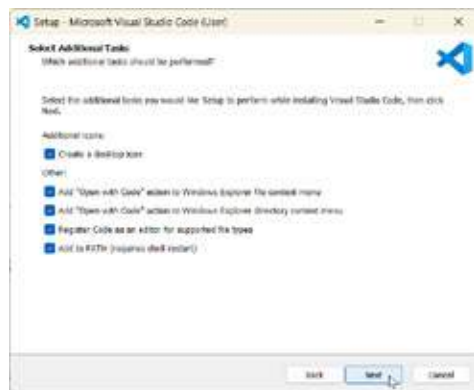
การเขียนและทดสอบโค้ด Data Science แม้จะมีหลายทางเลือก แต่สำหรับในหนังสือเล่มนี้ ผู้เขียนเลือกใช้ Visual Studio Code (ต่อไปจะเรียกสั้นๆ ว่า VS Code) ด้วยเหตุผลดังนี้

- เป็นโปรแกรมประเภท IDE (Integrated Development Environment) สำหรับการเขียนโค้ดที่มีขนาดไม่ใหญ่มาก และได้รับความนิยมสูงสุดในปัจจุบัน
- รองรับการเขียนโค้ดได้หลายภาษา
- ทำงานได้อย่างชาญฉลาด พร้อมตัวช่วยในการเขียนโค้ดอย่างครบครัน รวมถึงยืดหยุ่นต่อการปรับแต่งให้สามารถใช้ร่วมกับภาษาคอมพิวเตอร์อื่นๆ ได้เป็นอย่างดี
- มีส่วนเสริมการทำงานได้เลือกใช่มากมาย และสามารถนำมาใช้งานได้ฟรี

สำหรับการติดตั้ง VS Code ให้สามารถดาวน์โหลดติดตั้งได้จากเว็บไซต์ <https://code.visualstudio.com/download>



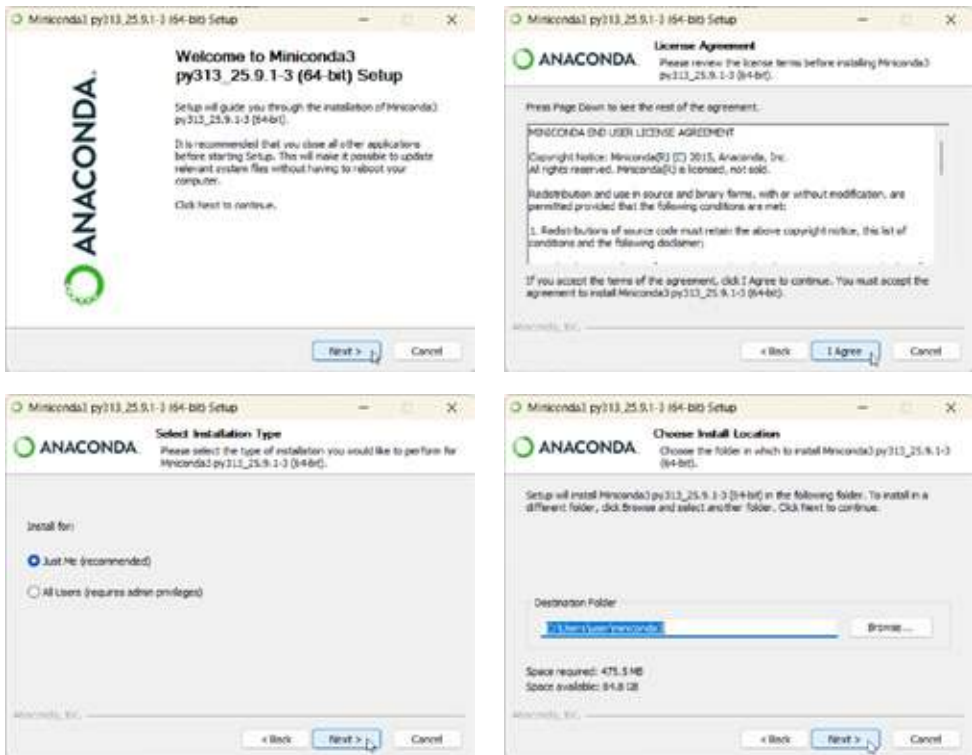
หลังดาวน์โหลดเสร็จ ก็ดับเบิลคลิกไฟล์ติดตั้ง ซึ่งขั้นตอนบางส่วนที่น่าสนใจมีดังนี้



การติดตั้ง Anaconda/Miniconda

Anaconda เป็นชุดแพ็คเกจที่รวบรวมไลบรารีเกี่ยวกับ Data Science และ Machine Learning อย่างครบครัน เราจึงไม่ต้องเสียเวลาแยกติดตั้งทีละตัว ซึ่งมีแนวทางดังนี้

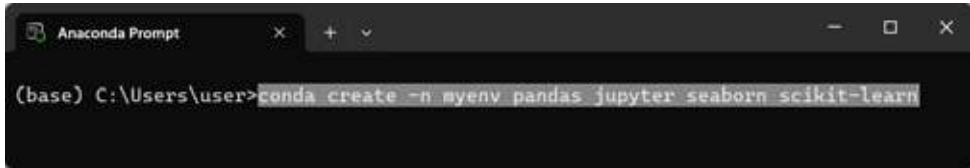
1. ดาวน์โหลด Anaconda ได้ที่เว็บไซต์ <https://www.anaconda.com/download> หรือจะใช้ Miniconda (<https://docs.anaconda.com/miniconda/>) ก็เพียงพอต่อการใช้งานแล้ว (Anaconda ต้องการพื้นที่ประมาณ 4.4 GB ขณะที่ Miniconda ต้องการเพียง 480 MB)
2. หลังจากดาวน์โหลดเสร็จ ก็ติดตั้งลงในเครื่องเหมือนโปรแกรมทั่ว ๆ ไป



สร้าง Environment ของ Anaconda

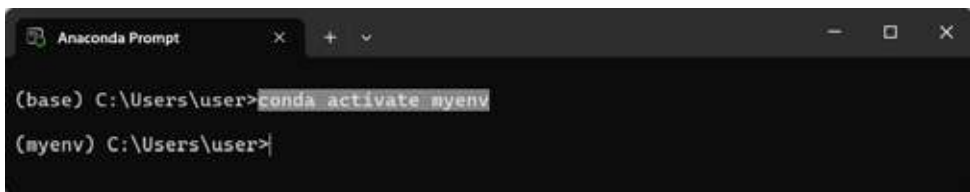
เราต้องสร้าง Environment ซึ่งเป็นเสมือนเซิร์ฟเวอร์เพื่อใช้เชื่อมต่อระหว่าง Anaconda กับ VS Code โดยหลังการติดตั้ง Anaconda เสร็จ ให้เปิด Anaconda Prompt (**คลิกที่ปุ่ม Start > โปรแกรม Anaconda > เลือก Anaconda Prompt**) ซึ่งในที่นี้ผู้เขียนจะสร้าง Env. ชื่อ myenv (จะใช้ชื่ออื่นก็ได้แต่ต้องจำให้ได้ว่าตั้งชื่อเป็นอะไร เพราะต้องอ้างอิงชื่อนี้ในภายหลัง) และก็จะติดตั้งไลบรารีต่อไปนี้เป็น pandas jupyter seaborn scikit-learn หรือเราจะติดตั้งไลบรารีอื่นๆ เพิ่มเติมก็ได้ โดยพิมพ์คำสั่งลงบน Anaconda Prompt ดังนี้

```
conda create -n myenv pandas jupyter seaborn scikit-learn
```



ขณะติดตั้งมีขั้นตอนที่เราต้องตอบ y(es) เพื่อยืนยันด้วย หลังจากสร้าง Environment เสร็จแล้วก็ให้เรา Activate การทำงาน โดยพิมพ์คำสั่งดังนี้ (myenv คือชื่อ Environment ที่เราสร้างขึ้นในคำสั่งที่ผ่านมา)

```
conda activate myenv
```



ในส่วนการติดตั้ง Anaconda ก็ถือว่าเสร็จสิ้นเพียงเท่านี้

ติดตั้งส่วนเสริม Jupyter บน VS Code

การเขียนโค้ด Data Science มักเป็นเพียงลักษณะปลีกย่อยไม่ยืดเยื้อนัก ดังนั้นจึงไม่นิยมเขียนโค้ดเป็นไฟล์ภาษาไพธอน (.py) แต่จะแบ่งออกเป็นส่วนย่อยๆ เพื่อการทดสอบในแต่ละกรณี ซึ่งการเขียนโค้ดแบบนี้เรานิยมใช้เครื่องมือที่เรียกว่า Jupyter Notebook และสามารถใช้งานบน VS Code ได้ โดยการติดตั้งเพิ่มเข้ามาในแบบส่วนเสริม ดังนี้ (ถ้าผู้อ่านเคยศึกษาหนังสือ Python ฉบับพื้นฐานของผู้เขียน ซึ่งได้เขียนโค้ดด้วย Jupyter Notebook บน VS Code เช่นกัน และถ้ามี Jupyter Notebook ติดตั้งเอาไว้แล้ว ก็ข้ามขั้นตอนนี้ไปได้เลย)



1. เปิดเข้าสู่ VS Code แล้วค้นหาส่วนเสริมที่ชื่อ Jupyter
2. คลิกปุ่ม install เพื่อติดตั้งส่วนเสริมนี้ลงใน VS Code

จัดการและวิเคราะห์ข้อมูลด้วย Data Science และพื้นฐาน Machine Learning โดยใช้ภาษา Python

Data Science ถือเป็นวิทยาศาสตร์อีกแขนงหนึ่งที่ว่าด้วยการวิเคราะห์และจัดการข้อมูลในรูปแบบต่างๆ นอกจากนี้ยังเป็นพื้นฐานไปสู่การเรียนรู้เทคโนโลยีแขนงอื่นๆ อีกหลายอย่าง ไม่ว่าจะเป็น Data Mining, Deep Learning, Machine Learning และ Artificial Intelligence (AI) ดังนั้นในปัจจุบันจึงมีผู้สนใจที่จะศึกษา Data Science สำหรับนำไปประยุกต์ใช้กับงานด้านต่างๆ เป็นจำนวนมาก ซึ่งหนังสือเล่มนี้ได้จัดทำขึ้นเพื่อใช้ประกอบการเรียนรู้เกี่ยวกับ Data Science โดยใช้ภาษา Python ร่วมกับไลบรารีต่างๆ เช่น NumPy, Pandas, Matplotlib และ Scikit-learn ซึ่งนอกจาก Data Science แล้ว ก็ยังกล่าวถึง Machine Learning ขึ้นพื้นฐานเอาไว้ด้วย เพื่อให้ผู้อ่านสามารถนำไปจัดการ วิเคราะห์ และทำนายผลข้อมูลในเบื้องต้นได้ รวมทั้งเป็นแนวทางสำหรับการเรียนรู้เพิ่มเติมในระดับที่สูงขึ้นต่อไป

เนื้อหาในหนังสือประกอบด้วย :

- บทที่ 1 : การจัดเตรียมเครื่องมือ
- บทที่ 2 : ทบทวน Python สำหรับ Data Science
- บทที่ 3 : การใช้ไลบรารี NumPy
- บทที่ 4 : เริ่มต้นการใช้ Series และ DataFrame
- บทที่ 5 : การเลือกช่วงข้อมูล
- บทที่ 6 : การคัดกรองด้วยเงื่อนไขและการคำนวณ
- บทที่ 7 : ค่าที่สูญหายและฟังก์ชันด้านสถิติ
- บทที่ 8 : การใช้ฟังก์ชันอื่น ๆ ของ DataFrame
- บทที่ 9 : ข้อมูลสตริง วันเวลา และจากไฟล์
- บทที่ 10 : กลุ่มข้อมูลและ Pivot Table
- บทที่ 11 : การพล็อตกราฟแบบ Visualization
- บทที่ 12 : ML : พื้นฐานและการจัดเตรียมข้อมูล
- บทที่ 13 : ML : การทำนายผลด้วยโมเดลแบบต่าง ๆ (1)
- บทที่ 14 : ML : การทำนายผลด้วยโมเดลแบบต่าง ๆ (2)

พร้อมจำหน่ายในรูปแบบ

- ☒ e-book (PDF) ☐ e-book (EPUB) ☐ audiobooks
☒ ปกอ่อน ☐ LARGE PRINT (ตัวอักษรขนาดใหญ่)



www.se-ed.com



sbc.fans



SE-ED Publisher



ISBN 978-616-08-5701-2



9

786160

857012

259 บาท

คอมพิวเตอร์-การเขียนโปรแกรม