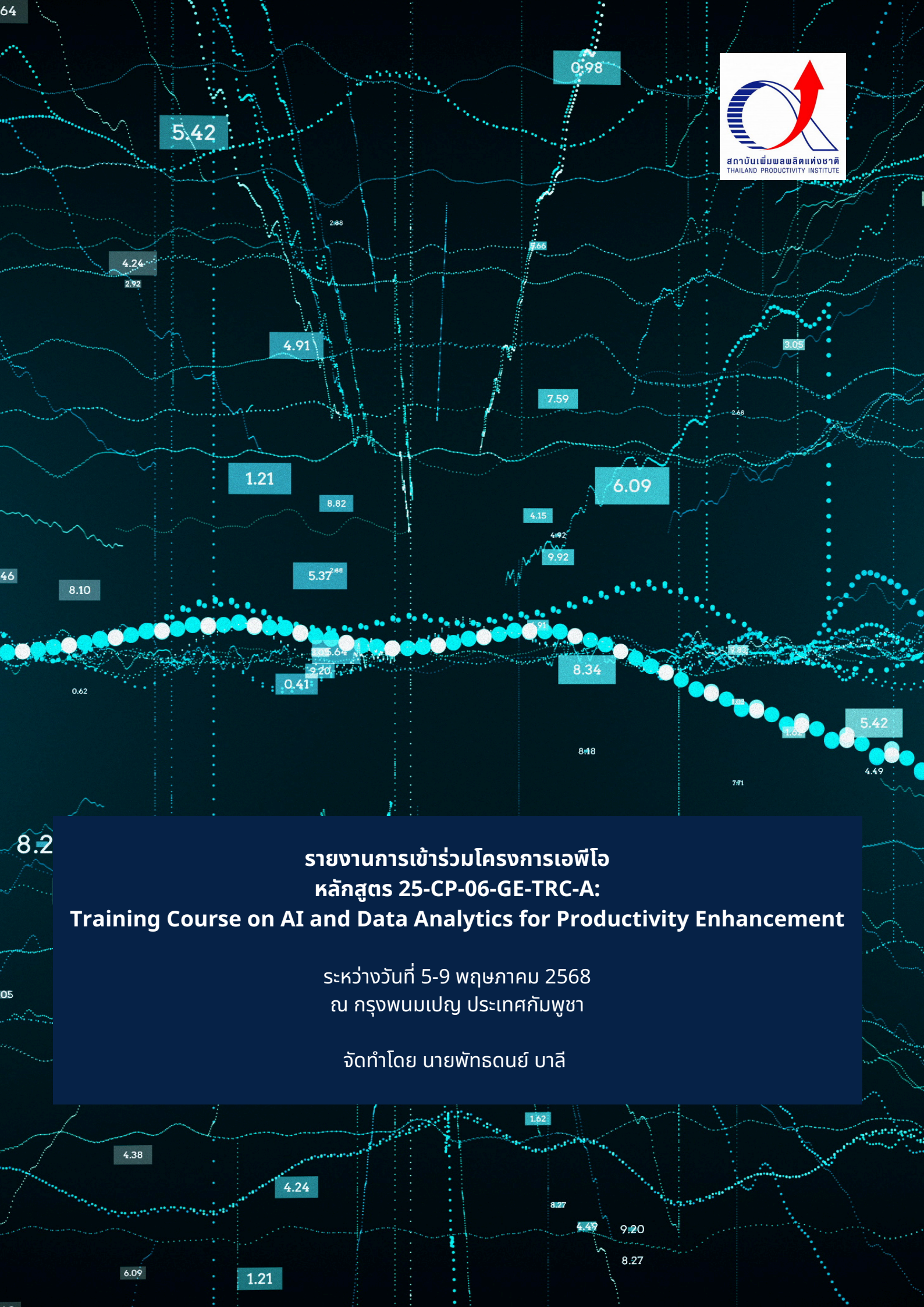


สถาบันเพิ่มผลผลิตแห่งชาติ
THAILAND PRODUCTIVITY INSTITUTE

รายงานการเข้าร่วมโครงการเอพีโอ
หลักสูตร 25-CP-06-GE-TRC-A:
Training Course on AI and Data Analytics for Productivity Enhancement

ระหว่างวันที่ 5-9 พฤษภาคม 2568
ณ กรุงพนมเปญ ประเทศกัมพูชา

จัดทำโดย นายพัทธดนย์ บาลี

รายงานการเข้าร่วมโครงการเอพีโอ
หลักสูตร 25-CP-06-GE-TRC-A:
Training Course on AI and Data Analytics for Productivity Enhancement
ระหว่างวันที่ 5-9 พฤษภาคม 2568
ณ กรุงพนมเปญ ประเทศกัมพูชา

จัดทำโดย นายพัทธดนย์ บาลี
รองกรรมการผู้จัดการ
บริษัท เค.ที.พี. (ประเทศไทย) จำกัด

ข้าพเจ้าได้มีโอกาสเข้าร่วมการฝึกอบรมหลักสูตร Training Course on AI and Data Analytics for Productivity Enhancement ระหว่างวันที่ 5 – 9 พฤษภาคม 2568 ณ กรุงพนมเปญ ประเทศกัมพูชา หลักสูตรดังกล่าวจัดโดย Asian Productivity Organization (APO) ร่วมกับ National Productivity Center of Cambodia (NPCC) โดยมีผู้เข้าร่วมอบรมทั้งหมด 25 ท่าน จากจำนวน 15 ประเทศ วัตถุประสงค์หลักของการจัดหลักสูตรครั้งนี้ เพื่อต้องการส่งเสริมให้ผู้เข้าอบรมในฐานะตัวแทนของประเทศสมาชิก APO มีความเข้าใจเกี่ยวกับทฤษฎีและการประยุกต์ใช้ AI และ Data Analytics ในการเพิ่มผลิตภาพขององค์กรและเสริมสร้างยกระดับขีดความสามารถในการแข่งขัน เนื่องจาก AI เป็นเทคโนโลยีที่สามารถนำมาเพิ่มประสิทธิภาพและคุณภาพของผลผลิตในการทำงาน ขณะที่ Data Analytics สามารถช่วยให้ผู้ใช้งานได้เรียนรู้ถึง Insight จากข้อมูล เพื่อเพิ่มคุณภาพการตัดสินใจ จึงถือได้ว่าเป็นหลักสูตรที่มีประโยชน์อย่างยิ่ง

สำหรับการจัดฝึกอบรมหลักสูตรในครั้งนี้ จัดขึ้นที่โรงแรม Poulo Wai Hotel & Apartment โดยได้รับการสนับสนุนจาก APO และ NPCC หลักสูตรมีการเชิญวิทยากรผู้ทรงคุณวุฒิเฉพาะด้านเกี่ยวกับ AI และ Data Analytics จำนวน 4 ท่าน เพื่อบรรยายให้ผู้เข้าร่วมอบรมมีความเข้าใจเกี่ยวกับศาสตร์ AI และ Data Analytics ประกอบด้วย

1. Dr. Chintan Amrit, Senior Associate Professor, University of Amsterdam, Netherlands
2. Dr. Murphy Choy, CEO, Alionova Education, Singapore
3. Quinn Phan, Director of Solution Consulting, Meiro Pte. Ltd., Singapore
4. Dr. Chutiporn Anutariya, Associate Professor, Information Management, Department of Information and Communication Technologies, Asian Institute of Technology, Thailand

นอกจากนี้ ในวันที่ 8 พฤษภาคม 2568 ทาง NPCC ได้มีการพาผู้เข้าอบรมเยี่ยมชมศึกษาดูงานสถานที่ 2 แห่ง ได้แก่ AI Farm Factory และ Cambodia Academy Digital Technology ซึ่งถือเป็นตัวอย่างขององค์กรด้านนวัตกรรมที่มีการใช้ AI และ Digital Technology เข้ามาประยุกต์ใช้ เพื่อสนับสนุนการขับเคลื่อนพัฒนาประเทศกัมพูชาให้มีขีดความสามารถในการแข่งขันเพิ่มมากขึ้น

การฝึกอบรมหลักสูตรในครั้งนี้ถือว่าประสบความสำเร็จเป็นอย่างดี เนื่องจากข้าพเจ้าและผู้เข้ารับการอบรมต่างได้รับความรู้และมีความเข้าใจเกี่ยวกับประเด็นด้านเทคโนโลยี AI และ Data Analytics เพิ่มมากขึ้น ทั้งในภาคทฤษฎีและภาคปฏิบัติ โดยวิทยากรผู้ทรงคุณวุฒิได้มีการจัดบรรยายแบบมีส่วนร่วม และลงมือปฏิบัติจริง ผ่านการทำ Workshop และ

Case Studies ข้าพเจ้าสามารถนำความรู้ที่ได้รับมาพัฒนาต่อยอดกับองค์กรที่ข้าพเจ้าสังกัดและสามารถถ่ายทอดความรู้ส่วนหนึ่งให้แก่ทีมงาน เพื่อร่วมกันขับเคลื่อนองค์กรสู่ยุคแห่ง AI และ Data Analytics

รูปภาพบรรยากาศบางส่วนของการอบรมหลักสูตร

Training Course on AI and Data Analytics for Productivity Enhancement

องค์ความรู้ที่ได้จากการฝึกอบรมหลักสูตร 25-CP-06-GE-TRC-A

Training Course on AI and Data Analytics for Productivity Enhancement

Session 1: Introduction to AI and Data Analytics

โดย Dr. Chintan Amrit, Associate Professor, Business Analytics Department, University of Amsterdam, Netherlands

การเรียนการสอนในชั้นเรียนนี้ ผู้สอนได้แนะนำเกี่ยวกับประเด็นพื้นฐานที่เกี่ยวกับ Artificial Intelligence (AI) และ Data Analytics โดยปูพื้นฐานตั้งแต่ความเข้าใจเกี่ยวกับศาสตร์ของ AI และ Data Analytics โดยเฉพาะในเนื้อหาส่วน Data Analytics มิใช่เพียงแค่การวิเคราะห์ข้อมูลเพียงอย่างเดียว แต่เป็นการบูรณาการกันของศาสตร์ 3 ด้าน ได้แก่ (1) ศาสตร์ด้าน Analytics และสถิติ Statistics (2) ความรู้ในเชิงธุรกิจ และ (3) ศาสตร์ด้าน Computer Science เข้าด้วยกัน เพื่อแปลงข้อมูลพื้นฐาน (Data) ให้ได้มาซึ่งองค์ความรู้ (Knowledge) ที่สามารถนำไปใช้ประโยชน์ในเชิงธุรกิจหรือสนับสนุนการตัดสินใจขององค์กรได้ การขาด Foundation ในศาสตร์ใดศาสตร์หนึ่ง จะทำให้ Data Analytics ขาดความสมบูรณ์ และส่งผลต่อการนำไปใช้งานจริง ยกตัวอย่างเช่น การมุ่งเน้นไปที่การวิเคราะห์ Data โดยปราศจากโจทย์ปัญหาทางธุรกิจอย่างแท้จริง หรือขาดความเข้าใจในประเด็นที่ธุรกิจต้องการ เช่น การโยนโจทย์ข้อมูลให้กับนักวิเคราะห์ข้อมูล ให้ค้นหา Pattern ข้อมูลบางอย่าง โดยปราศจาก Clear Objectives ทางธุรกิจ จะทำให้ผลลัพธ์ที่ได้อาจไม่เป็นไปดั่งที่คาดหวัง เมื่อเทียบกับกรณีที่มีโจทย์ทางธุรกิจที่ชัดเจน เช่นเดียวกันกับการวิเคราะห์ข้อมูลโดยใช้เทคนิคทางสถิติ เพื่อแก้ปัญหาธุรกิจบางอย่าง แต่ปราศจากการสนับสนุนของ Computer Science ที่มีศักยภาพในการประมวลข้อมูลจำนวนมาก จะทำให้การวิเคราะห์หรือค้นหาองค์ความรู้ดังกล่าวเกิดข้อจำกัด เนื่องจากไม่สามารถประมวลผลข้อมูลจำนวนมากได้โดยปราศจาก Computer ดังนั้น ทั้ง 3 ศาสตร์จึงถือเป็นพื้นฐานสำคัญอย่างยิ่งต่อการประสบความสำเร็จด้าน Data Analytics

ตัวอย่างปัญหาทางธุรกิจที่สามารถใช้ศาสตร์ด้าน Data Analytics เข้ามาช่วยประมวล วิเคราะห์ และแก้ไขปัญหาได้มีมากมาย โดยผู้สอนได้ยกตัวอย่างกรณีของการพยากรณ์การมาถึงของรถบรรทุก (Truck Arrival Prediction) ในสถานที่จัดเก็บสินค้าหรือคลังสินค้า เพื่อให้สามารถบริหารจัดการ Traffic ของคลังสินค้าได้อย่างมีประสิทธิภาพสูงสุด หรือการพยากรณ์เพื่อค้นหาเด็กและเยาวชนที่มีแนวโน้มถูกทำร้าย (Abuse) โดยใช้เทคนิคด้าน Text Mining วิเคราะห์ลักษณะข้อความหรือประโยคที่แพทย์เขียนลงในรายงานเกี่ยวกับเด็กและเยาวชนที่สามารถบ่งชี้ถึงผู้ที่มีแนวโน้มตกเป็นเหยื่อการถูกทำร้ายได้ เพื่อที่จะสามารถหาทางป้องกันและแก้ไขก่อนที่เหยื่อจะเกิดความเสียหายทางร่างกายและจิตใจมากขึ้น นอกจากนี้ Data Analytics ยังสามารถช่วยให้มองเห็น Pattern ของการโน้มน้าวการตัดสินใจทางการเมืองของผู้คนบนโลก Internet ตลอดจนช่วยทำนายประเทศในแถบแอฟริกาที่มีแนวโน้มจะเกิดภาวะทุพพลภาพทางโภชนาการของเด็ก เพื่อให้เกิดความช่วยเหลือได้ทันเวลามากขึ้น จึงเห็นได้ว่า ศาสตร์ Data Analytics สามารถสนับสนุนการแก้ไขปัญหาทั้งระดับองค์กร สังคม ชุมชน และประเทศชาติได้ ตลอดจนสร้างมูลค่าเพิ่มทางเศรษฐกิจและสังคม โดยอาศัยการแปลงข้อมูลทั่วไปให้กลายเป็นองค์ความรู้ที่เกิดประโยชน์อย่างชัดเจน

ผู้สอนได้แนะนำ Concept พื้นฐานที่เกี่ยวข้องกับศาสตร์ Data Analytics โดยแนะนำคำว่า สถิติ (Statistics), Machine Learning, Data Mining, Data Warehouse / Storage, Querying / Reporting, OLAP รวมถึงคำอื่น เพื่อให้ผู้เรียนเกิดความเข้าใจในเนื้อหา เมื่อมีการกล่าวถึงคำเหล่านี้ในบทเรียนถัดไป แต่ละคำมีความหมายและจุดประสงค์ที่แตกต่างกันออกไป เช่น

- สถิติ (Statistics) มุ่งเน้นไปที่การเรียนรู้เกี่ยวกับทฤษฎี โมเดล และการทดสอบทางสถิติ (Hypothesis Testing)
- Machine Learning จะมุ่งเน้นไปที่ Heuristics (ทางลัดในการตัดสินใจ) โดย Computer หรือเครื่องจักร จะทำหน้าที่ในการเรียนรู้และสกัด Pattern บางอย่างออกมา เพื่อให้สามารถทำนายหรือตัดสินใจได้เองโดยอัตโนมัติ ดังนั้น จึงมุ่งเน้นไปที่การเรียนรู้ด้วยตัวเอง เพื่อพัฒนาประสิทธิภาพการทำงาน โดย Machine Learning มี 3 รูปแบบใหญ่
    1. Supervised Learning คือ การเรียนรู้ของ Computer หรือเครื่องจักร โดยมีมนุษย์ทำหน้าที่ช่วยสอนและให้ Feedback เพื่อให้เกิดความเข้าใจที่ถูกต้องของ Computer หรือเครื่องจักร เช่น เราป้อนข้อมูล Raw Data เป็นรูปภาพสัตว์ที่แตกต่างกัน โดยเราต้องการให้ Computer จำแนกได้ว่า รูปที่ป้อนเข้าไปคือ รูปสัตว์ประเภทใด เช่น แมวหรือสุนัข โดยมนุษย์จะทำหน้าที่สอนว่า รูปลักษณะนี้บ่งบอกถึงแมวหรือสุนัข จากนั้น Machine จะทำหน้าที่เรียนรู้และ Classify ได้ เมื่อได้รับการสอนและเรียนรู้ Pattern จากข้อมูลจำนวนมาก
    2. Unsupervised Learning คือ การเรียนรู้ของ Computer หรือเครื่องจักร โดยไม่มีมนุษย์ช่วยทำหน้าที่สอนและให้ Feedback กล่าวคือ Machine ต้องค้นหา Pattern ด้วยตัวเองจากข้อมูลที่มีอยู่ เช่น การทำนายความสำเร็จของภาพยนตร์ (อ้างอิงจากรายได้ที่ภาพยนตร์ทำได้) โดยใช้ตัวแปรต้นจากข้อมูลแหล่งต่างๆ เช่น Rating ภาพยนตร์ ประเภทของภาพยนตร์ (Genre) จำนวนโรงภาพยนตร์ที่เข้าฉาย (Number of Screens) เป็นต้น ซึ่งในศาสตร์ Data Analytics จะมีโมเดลหลากหลายรูปแบบที่ใช้สำหรับการทำ Prediction Model เช่น SVM, ANN, C&RT, Random Forest, Boosted Tree หรือ Fusion เป็นต้น แต่ละโมเดลจะสามารถทำนายความสำเร็จของภาพยนตร์ โดยมีความแม่นยำที่แตกต่างกัน (Accuracy) ดังนั้น Computer หรือ Machine จะต้องเรียนรู้ว่า ตัวแปรต้นชุดใด และโมเดลใดที่เหมาะสมกับการทำนายในแต่ละบริบทที่แตกต่างกันออกไป โดยทั้งหมดนี้ สามารถทำได้โดยที่ไม่มีมนุษย์เข้ามาทำหน้าที่ช่วยสอนหรือให้ Feedback
    3. Reinforcement Learning หมายถึง การเรียนรู้ของ Computer หรือ Machine ผ่านการลองผิดลองถูก โดยที่มีรางวัลหรือบทลงโทษ (Reward/Penalty) เป็นตัวไกด์ให้ Machine ทราบว่า มาถูกแนวทางแล้วหรือไม่ ในกรณีที่ Machine ทำได้สอดคล้องกับเป้าหมาย จะได้รับรางวัล ในขณะที่ Machine ทำไม่ได้สอดคล้องกับเป้าหมาย จะได้รับการลงโทษ ในที่นี้ การลงโทษไม่ได้หมายถึงการดุหรือตำหนิ แต่หมายถึงการให้คะแนนติดลบกับ Machine กรณีที่ทำได้ไม่สอดคล้องกับเป้าหมาย ยกตัวอย่างเช่น หากต้องการให้หุ่นยนต์ค้นหาทางออกในเขาวงกตโดยอัตโนมัติ กรณีที่เดินไปถึงประตูทางออก จะได้รับคะแนน +100 คะแนน แต่ในขณะที่หากเดินชนกำแพง จะได้รับการลงโทษ -10 คะแนน เมื่อหุ่นยนต์เดินชนกำแพงจะได้รับคะแนนลดลง จึงทำให้เกิดการเรียนรู้ด้วยตัวเองว่า หากชนกำแพงจะถูกลงโทษ (กรณีนี้คือการโดนหักคะแนน) ทำให้เมื่อเรียนรู้อย่างต่อเนื่อง หุ่นยนต์จะทำหน้าที่ได้ดีขึ้น จนกระทั่งหาทางออกได้ โดยหลีกเลี่ยงการชนกำแพง
- Data Mining มุ่งเน้นการผสมผสานกันทั้งทฤษฎีและ Heuristics เพื่อค้นหาองค์ความรู้ที่เกิดประโยชน์ (Knowledge Discovery) โดยเริ่มต้นจากคลังข้อมูล (Data Warehouse) ที่ประกอบด้วยข้อมูลจำนวนมาก จากนั้นผู้ปฏิบัติงานต้องเลือกข้อมูลเฉพาะชุดที่เกี่ยวข้อง (Selection) และ Clean ข้อมูลบางส่วน (Cleaning) เพื่อนำไป

วิเคราะห์หา Patterns และ Rules ที่อยู่เบื้องหลังของชุดข้อมูลเหล่านั้น เมื่อนำ Patterns/Rules มาตีความและประเมินผล จะนำไปสู่ Knowledge Discovery ที่สามารถนำมาใช้ประโยชน์ได้จริง

- Data Warehouse / Storage หรือคลังข้อมูล คือ การรวมเอาข้อมูลจากหลายส่วนในองค์กรเข้ามาไว้ในชุดข้อมูลกลุ่มเดียวกัน บางครั้งข้อมูลเหล่านี้อาจมาจาก Transaction ที่เกิดจากภายในหรือภายนอกองค์กร
- Querying / Reporting โดยมากจะดำเนินการในโปรแกรม เช่น SQL, Excel, QBE โดยโปรแกรมเหล่านี้จะช่วยทำให้การประมวล คัดเลือก และนำเสนอข้อมูลเป็นไปได้อย่างสะดวกและมีประสิทธิภาพ โดยการนำเสนออาจอยู่ในรูปแบบ Visualization หรือแผนภาพ เพื่อให้เกิดความเข้าใจง่ายขึ้น
- OLAP (Online Analytical Processing) เป็นระบบประมวลผลข้อมูลขนาดใหญ่ (Big Data) ที่ข้อมูลมีมิติที่หลากหลาย ซับซ้อน ช่วยทำให้การวิเคราะห์สะดวก มีประสิทธิภาพ และยืดหยุ่นมากขึ้น เช่น ผู้ปฏิบัติงานสามารถวิเคราะห์ยอดขายสินค้า ได้จากหลายมิติ เช่น ช่วงเวลา สถานที่ และประเภทสินค้าหรือแบรนด์ที่แตกต่างกัน เหมาะสำหรับการตัดสินใจระดับบริหาร

ผู้สอนยังได้นำเสนอพื้นฐานเกี่ยวกับรูปแบบข้อมูลที่ใช้ในการทำ Data Mining หรือการค้นหาองค์ความรู้ที่ก่อให้เกิดประโยชน์ โดยจำแนกข้อมูลออกเป็น (1) Nominal คือ การจำแนกข้อมูลประเภทที่ไม่ใช่ตัวเลข โดยไม่มีการเรียงลำดับ เช่น เพศ (ชาย หรือ หญิง) ประเทศ แผนกในบริษัท (2) Ordinal คือ การจำแนกข้อมูลประเภทที่ไม่ใช่ตัวเลข โดยมีการเรียงลำดับ เช่น ระดับการศึกษา (มัธยม < ปริญญาตรี < ปริญญาโท < ปริญญาเอก) (3) Interval คือ การจำแนกข้อมูลประเภทตัวเลข ที่ไม่มีศูนย์แท้ (หมายถึง การไม่มีสิ่งนั้นอย่างแท้จริง) เช่น อุณหภูมิ (0 องศา Celsius ไม่ได้แปลว่า ไม่มีอุณหภูมิ) (4) Ratio การจำแนกข้อมูลประเภทตัวเลข ที่มีศูนย์แท้ (หมายถึง การไม่มีสิ่งนั้นอย่างแท้จริง) เช่น รายได้ 0 บาท หมายถึง การไม่มีรายได้เลย หรือ น้ำหนัก 0 กิโลกรัม หมายถึง การไม่มีน้ำหนักเลย เป็นต้น

การนำข้อมูลมาใช้ค้นหา Pattern เพื่อสร้างองค์ความรู้ โดยใช้เทคนิค Data Mining มักจะใช้การค้นหา Patterns ผ่าน Association, Prediction, Cluster (Segmentation) หรือ Time Series Relationship โดยแต่ละรูปแบบจะมี Algorithm ที่สามารถช่วยให้ค้นหา Pattern ได้ เช่น การจำแนกข้อมูลเป็นกลุ่ม (Clustering) สามารถใช้วิธี K-Means หรือ EM (Expectation Maximization) หรือการทำ Prediction สามารถใช้วิธี Linear / Non-linear Regression, ANN, MLP, SVM

ทั้งนี้ การค้นหาองค์ความรู้ไม่ควรเริ่มต้นจากสถิติหรือโมเดล แต่ควรเริ่มต้นจากความเข้าใจปัญหาทางธุรกิจหรือองค์กรที่ต้องการแก้ไขเสียก่อน จึงต้องเริ่มจาก Business Understanding จากนั้น จึงทำความเข้าใจว่า Data ที่เรามีนั้นสามารถตอบโจทย์หรือนำไปสู่การแก้ปัญหาที่มีอยู่หรือไม่ เมื่อ Data นั้นสามารถนำไปสู่การแก้ปัญหาได้ จึงเข้าสู่กระบวนการ Data Analytics ได้แก่ การเตรียมข้อมูล (Data Preparation) การเตรียม Model เพื่อวิเคราะห์ข้อมูลและค้นหา Pattern ข้อมูล การทดสอบและประเมินผลลัพธ์จาก Model และการนำ Model ไปใช้จริง ทั้งหมดนี้ หลักการค้นหาองค์ความรู้ (Knowledge Discovery Process) อยู่ภายใต้โมเดลที่เรียกว่า CRISP-DM ซึ่งให้ความสำคัญกับการเข้าใจปัญหาทางธุรกิจ การเข้าใจข้อมูล และการเตรียมข้อมูล 80% ส่วนอีก 20% เป็นเรื่องของเทคนิคและกระบวนการด้านสถิติและโมเดล ท้ายที่สุด ผู้สอนได้แนะนำพื้นฐานของคำว่า Deep Learning ซึ่งเป็นสาขาย่อยของ Machine Learning แต่ใช้อัลกอริทึมขั้นสูง เช่น Neural Network ซึ่งเป็นพื้นฐานของ Artificial Intelligence (AI) สามารถเรียนรู้และเก่งขึ้นได้ด้วยตัวเอง โดยใช้การเรียนรู้จากข้อมูลจำนวนมาก

Session 2: Installation and Configuration of Basic Tools

โดย Dr. Murphy Choy, CEO, Alionova Education, Singapore

การเรียนการสอนในชั้นเรียนนี้ ผู้สอนได้แนะนำให้ผู้เรียนรู้จักและติดตั้งโปรแกรมและเครื่องมือพื้นฐานที่นักวิจัยหรือนักวิทยาศาสตร์ด้านข้อมูล (Data Scientists) สามารถใช้งาน เพื่อจัดทำ Data Analytics ประกอบด้วย 3 โปรแกรม ได้แก่ (1) Doubao หรือเครื่องมือ AI ที่พัฒนาโดยประเทศจีน (2) Orange Data Mining หรือเครื่องมือที่ทำให้สามารถ Visualize การทำ Data Analysis ได้ง่ายขึ้น (3) Tableau Public หรือเครื่องมือสำหรับการแสดงผลข้อมูล (Data Visualization Tool) ในรูปแบบที่เข้าใจง่าย และสามารถมี Interactive กับชุดข้อมูลได้ด้วย โดยเครื่องมือเหล่านี้ ผู้สอนคัดเลือกมาเพื่อสาธิตให้ผู้เรียนเข้าใจได้ง่าย รวมถึงโปรแกรมเหล่านี้มี Advanced Features ที่ทรงพลังและมีประสิทธิภาพหากผู้ใช้งานต้องการทำงานที่มีความซับซ้อนมากขึ้น

ในลำดับแรก ผู้สอนได้แนะนำให้ชั้นเรียนรู้จักเครื่องมือ AI ที่มีชื่อว่า Doubao (ต้นฉบับในเวอร์ชั่นภาษาจีน) หรือ Cici (เวอร์ชั่นภาษาอังกฤษ) โดยผู้ใช้งานสามารถเข้าได้ที่ https://www.doubao.com/ หรือ https://www.cici.com/ โดยลักษณะของ Doubao / Cici มีความคล้ายคลึงกับ AI ที่เป็นที่รู้จักในประเทศไทย เช่น ChatGPT, Google Gemini แต่ Doubao จะทรงพลังมาก หากใช้ในเวอร์ชั่นต้นฉบับภาษาจีน โดยมีความสามารถในการทำ Deep Research ในหัวข้อต่างๆ ภายในเวลารวดเร็ว ซึ่งในกรณีที่ต้องการ Source ข้อมูลที่มีภาษาจีนเป็นพื้นฐาน AI ตัวอื่นที่ไม่ได้สร้างในประเทศจีน จะมีข้อจำกัดในการเข้าถึงข้อมูลเหล่านี้ จึงทำให้ผลลัพธ์อาจไม่เป็นที่น่าพอใจเท่าไหร่นัก ภายในชั้นเรียน ผู้สอนได้เปิดโอกาสให้ผู้เรียนได้ทดลองใช้ AI ตัวนี้ (Doubao / Cici) ด้วยตนเอง และหารือผลลัพธ์ที่ได้ร่วมกับเพื่อนร่วมชั้นเรียน เป็นระยะเวลาประมาณ 20 นาที

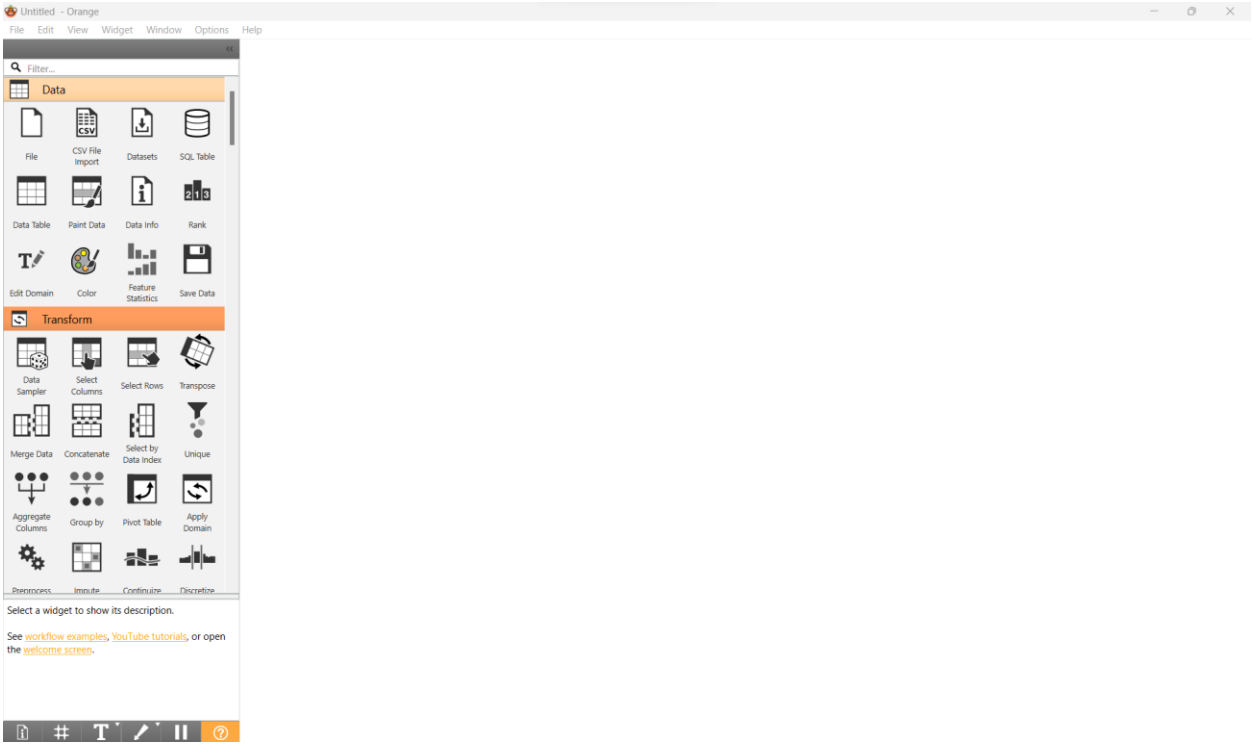
ในลำดับถัดไป ผู้สอนได้แนะนำให้ผู้เรียนได้ดาวน์โหลดและติดตั้งโปรแกรม Orange Data Mining ลงในเครื่องคอมพิวเตอร์ Laptop ของผู้เรียนทุกคน โดยโปรแกรมนี้สามารถติดตั้งได้ทั้งระบบ Windows, MacOS และ Linux โดยไม่มีค่าใช้จ่ายแต่อย่างใด ทั้งนี้ ผู้ใช้งานสามารถเข้าดาวน์โหลดโปรแกรมได้ที่ https://orangedatamining.com/download/ การติดตั้งทำได้ง่าย ไม่แตกต่างจาก Software ทั่วไป โดยกดเข้าไฟล์ที่ดาวน์โหลด .exe เพื่อติดตั้ง จากนั้นกดตาม Setup Wizard เพื่อติดตั้งโปรแกรม ผู้สอนได้เปิดโอกาสให้ผู้เรียนได้ทดลองดาวน์โหลด ติดตั้ง และทดลองใช้งานเมนูภายในโปรแกรมด้วยตนเอง ทั้งหมดประมาณ 20 นาที ภายในโปรแกรมมี Feature ในการจัดการข้อมูลหลายอย่าง เช่น การ Clean ข้อมูลประเภท Missing Data หรือ การ Merge ข้อมูลตั้งแต่ 2 ชุดขึ้นไปเข้าด้วยกัน โดย Interface ของโปรแกรมในภาพรวมจัดว่ามีความ User Friendly สูง จึงทำให้สามารถเข้าใจและศึกษาได้ง่าย

ในลำดับท้ายสุดของชั้นเรียนนี้ ผู้สอนได้แนะนำให้รู้จักและติดตั้งโปรแกรม Tableau Public ซึ่งเป็นโปรแกรมสำหรับการนำเสนอข้อมูลในรูปแบบที่เข้าใจง่าย (Data Visualization) ซึ่งเป็น Version ที่ไม่เสียค่าใช้จ่าย แต่ต้องแลกด้วยการเปิดเผยข้อมูลในรูปแบบสาธารณะ (ในกรณีที่ต้องการใช้งานแบบที่ไม่แสดงข้อมูลแบบสาธารณะ หรือ Private Use จะต้องใช้ Software อีกเวอร์ชั่นที่เสียค่าใช้จ่าย) โดยผู้ใช้งานสามารถดาวน์โหลดโปรแกรมได้ที่ https://public.tableau.com/ เช่นเดียวกันกับ Orange Data Mining โปรแกรมนี้ สามารถติดตั้งได้ง่าย เมื่อดาวน์โหลดเสร็จสิ้น ผู้ใช้งานสามารถกดตาม Setup Wizard เพื่อติดตั้งโปรแกรม หลังจากนั้น ผู้สอนได้เปิดโอกาสให้สอบถาม (Q&A) ก่อนสิ้นสุดชั้นเรียน

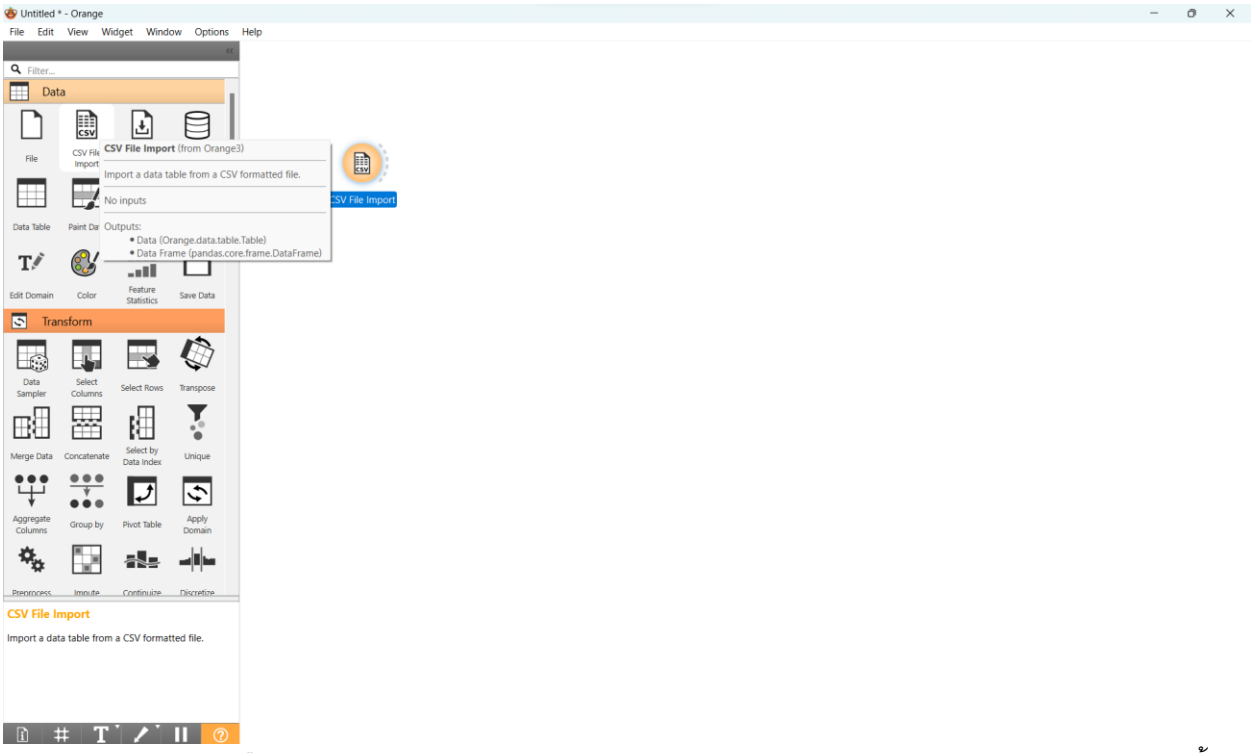
Session 3: Hands-on Exercises with Introductory Datasets

โดย Dr. Murphy Choy, CEO, Alionova Education, Singapore

ในชั้นเรียนนี้ ผู้สอนได้สาธิตและแนะนำวิธีการใช้งานโปรแกรม Orange Data Mining ที่ติดตั้งแล้วใน Session ที่ 2 โดยมีชุดข้อมูลตัวอย่างจำลองให้ผู้เรียน เพื่อใช้ประกอบการฝึกปฏิบัติจริง สำหรับชุดข้อมูลที่ใช้ในการสาธิต ได้แก่ ชุดข้อมูล Customer Purchase History โดยเป็นตัวอย่างของข้อมูลขนาดใหญ่ (Big Data) ที่มีรายการข้อมูลของร้านค้า 2,469 สาขา ตลอดระยะเวลา 1 ปี โดยผู้สอนจะสาธิตวิธีการใช้งานโปรแกรมพื้นฐาน Interface โปรแกรม และตัวอย่างการใช้งานโปรแกรมจริง เช่น การ Cleaning ข้อมูลที่มีปัญหา อาทิ รูปแบบข้อมูลสูญหาย (Missing Value) หรือข้อมูลที่มีลักษณะซ้ำกัน (Data Duplicates) เพื่อใช้ในการวิเคราะห์ในขั้นถัดไป (อ้างอิงตัวอย่างรูปภาพที่นำเสนอในหน้าถัดไป) โดยข้อดีของโปรแกรม Orange Data Mining คือ การที่โปรแกรมมี User Interface ที่เข้าใจและเรียนรู้ได้ง่าย ทำให้ผู้เรียนสามารถปฏิบัติตามผู้สอนได้ง่าย ผู้สอนยังแนะนำเกี่ยวกับหลักการและเทคนิคในการทำ Data Analytics ประกอบในแต่ละขั้นตอนด้วย



รูปภาพหน้าจอโปรแกรม Orange Data Mining หลังจากเปิดโปรแกรมใหม่



รูปภาพหน้าจอการเลือก Feature สำหรับนำเข้าข้อมูล CSV File บริเวณแถบด้านซ้ายบน ภายหลังจากกดปุ่มเสร็จสิ้น บริเวณ Workspace ตรงกลางจะปรากฏ Icon CSV File Import ซึ่งใช้ในการ Import CSV File

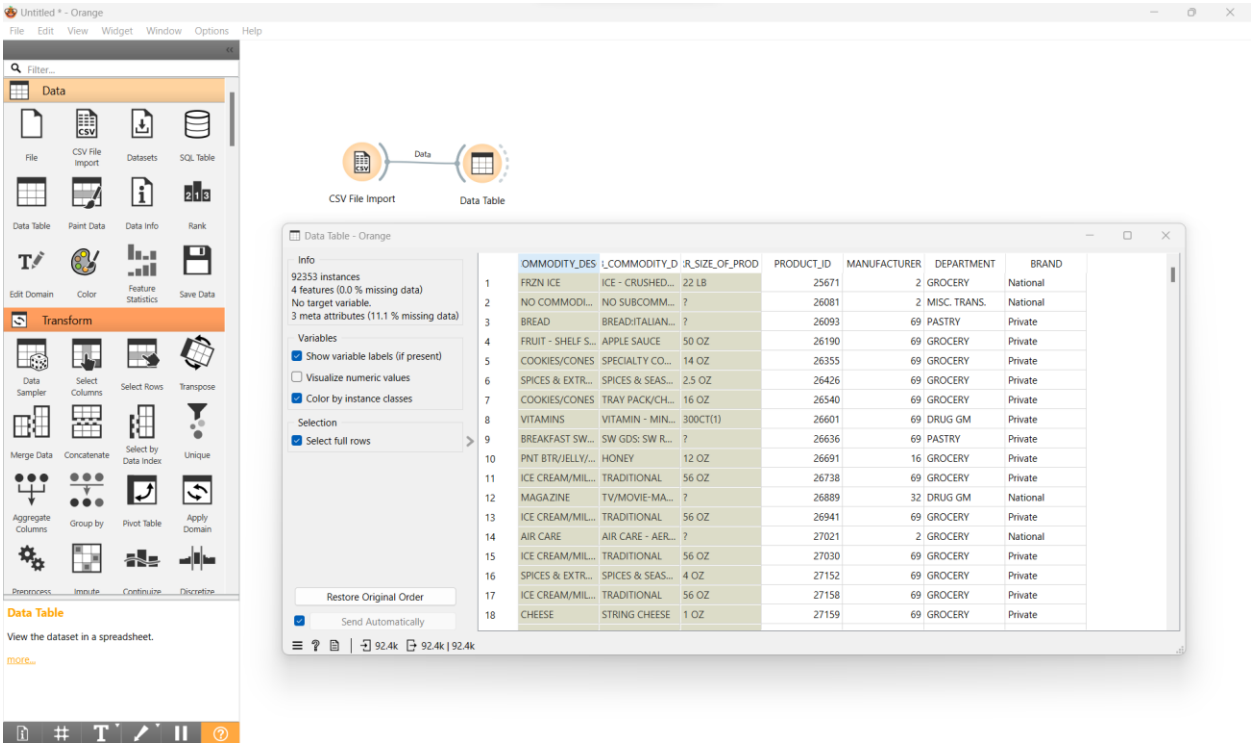
ในช่วงแรกของการฝึกอบรมช่วงนี้ ผู้สอนได้แนะนำวิธีการนำเข้า CSV File เพื่อใช้สำหรับการวิเคราะห์ข้อมูลในขั้นถัดไป โดยในโปรแกรมจะมี Workspace บริเวณตรงกลาง ซึ่งหากเปิดโปรแกรมในตอนแรก จะเป็นพื้นหลังสีขาวว่างเปล่า แต่เมื่อผู้ใช้งานได้เลือกคลิก Icon บริเวณแถบด้านซ้าย จะปรากฏ Icon ของฟังก์ชั่นที่ผู้ใช้งานเลือกบริเวณ Workspace ในที่นี้ การ Import CSV File ให้เลือกแถบ CSV File Import จากนั้นจะปรากฏ Icon ของฟังก์ชั่น โดยผู้ใช้งานจะต้อง Double Click ที่ Icon บริเวณ Workspace เพื่อนำเข้า CSV File เพื่อใช้ในการวิเคราะห์ข้อมูลต่อไป

ขั้นต่อไปหลังจากที่นำเข้าข้อมูล CSV File ผู้สอนได้แนะนำให้ผู้เรียนทำการตรวจสอบชุดข้อมูลที่นำเข้า เพื่อตรวจสอบความถูกต้องของข้อมูล ซึ่งจะส่งผลต่อความแม่นยำในการวิเคราะห์ในขั้นถัดไป จึงต้องศึกษาเทคนิคการทำ Data Cleaning โดยผู้สอนได้นำเสนอว่า ปัญหาส่วนใหญ่ (Common Issues) ที่มักพบในประเด็นข้อมูล ได้แก่ ข้อมูลสูญหายไม่ครบถ้วน (Missing Values) ข้อมูลบางส่วนแตกต่างจากค่าส่วนใหญ่มาก (Outliers) หรือชุดข้อมูลที่มีการบันทึกซ้ำกัน (Duplicate Records) ซึ่งเหล่านี้ต้องได้รับการจัดการอย่างเหมาะสมก่อนด้วยเทคนิคที่เรียกว่า Data Cleaning จึงจะนำไปวิเคราะห์ต่อในขั้นถัดไป

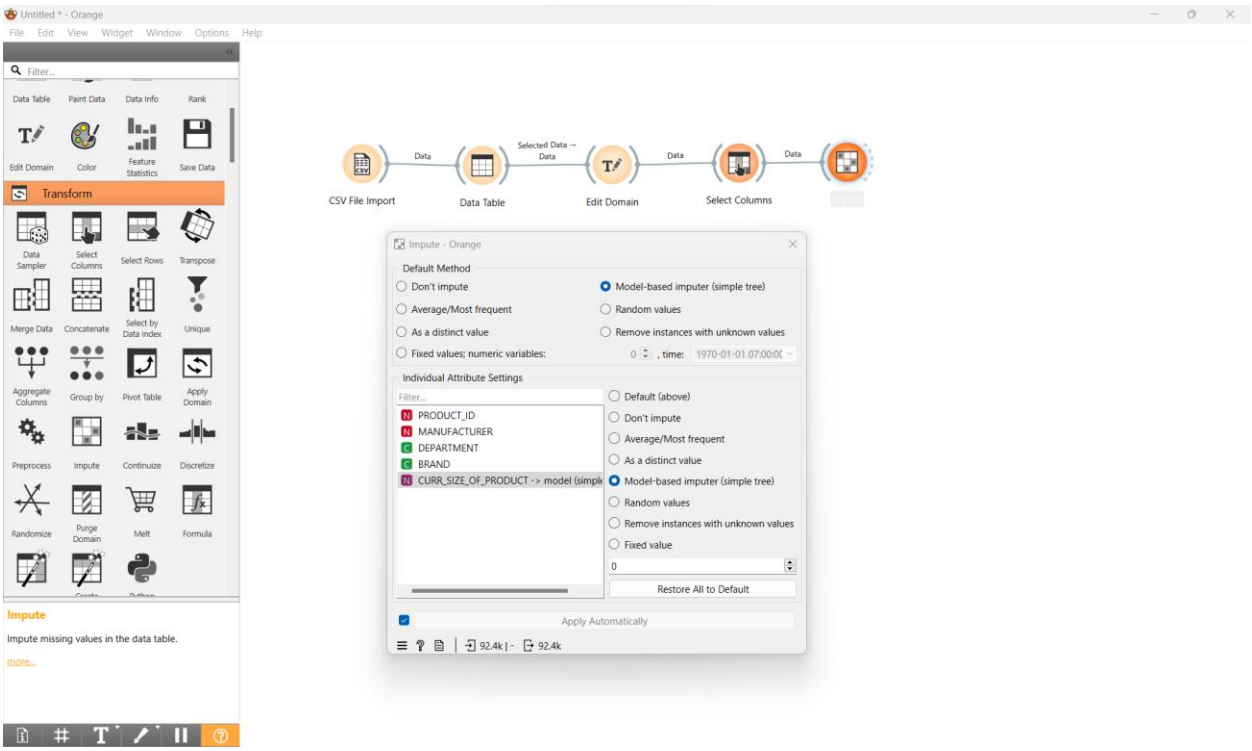
ลำดับแรก สิ่งที่ผู้ปฏิบัติงานทุกคนควรตรวจสอบก่อนดำเนินการในขั้นถัดไปคือ การค้นหาข้อมูลสูญหายไม่ครบถ้วน (Missing Values) โดยในโปรแกรม Orange Data Mining หลังจากที่ผู้ใช้งานนำเข้าข้อมูลจาก CSV File เรียบร้อยแล้ว สามารถตรวจสอบข้อมูลโดยแสดงผลไฟล์ข้อมูล Import อยู่ในรูปของ Data Table โดยเลือกคลิกที่แถบด้านซ้ายมือ คำว่า Data Table จากนั้น Icon Data Table จะปรากฏขึ้นใน Workspace หลังจากนั้น ลากเส้นจาก Icon CSV File Import ไปสู่ Icon Data Table โปรแกรมจะทำการสร้างเส้นเชื่อมระหว่าง 2 Icons เข้ากันโดยอัตโนมัติ แสดงว่า ขณะนี้ชุดข้อมูล

CSV File ที่นำเข้าสู่โปรแกรม สามารถแสดงผลในรูปแบบ Data Table ได้แล้ว จากนั้น กด Double Click ที่ Icon Data Table จะปรากฏรายการข้อมูลทั้งหมด พร้อมด้วย % ร้อยละของข้อมูล Missing Values จากตัวอย่างข้อมูลของผู้สอน ปรากฏว่า มีข้อมูลประมาณร้อยละ 11.1 ที่มี Missing Values (ตัวอย่างรูปภาพในหน้าถัดไป)

เมื่อค้นพบว่า ข้อมูลของเรามี Missing Values จึงต้องทำการ Clean ข้อมูลก่อนดำเนินการในขั้นถัดไป โดยผู้สอนได้แนะนำเทคนิคสำหรับจัดการ Missing Values ด้วย 3 เทคนิค ประกอบด้วย (1) Imputation หรือการแทนที่ข้อมูลที่สูญหายไปโดยใช้ค่าประมาณ อาทิ ค่าเฉลี่ย (Mean) ค่ามัธยฐาน (Median) ค่าฐานนิยม (Mode) หรือวิธีการอื่น ทั้งนี้ แต่ละรูปแบบจะเหมาะกับข้อมูลที่แตกต่างกันออกไป (2) Deletion หมายถึง การลบข้อมูลทั้งแถวที่ปรากฏ Missing Values ทำให้ไม่ต้องนำข้อมูลที่มีปัญหามาพิจารณา แต่มีข้อจำกัดคือ หากชุดข้อมูลมี Missing Values จำนวนมาก จะทำให้ข้อมูลถูกลบทิ้งมากเกินไป จนอาจส่งผลกระทบต่อการวิเคราะห์ ผู้สอนจึงแนะนำว่า ใช้ได้เฉพาะในกรณีที่มีค่าสูญหายไม่เกิน 5% ของชุดข้อมูลทั้งหมดเท่านั้น และ (3) Interpolation หมายถึง การแทนที่ Missing Values ด้วยค่าประมาณระหว่างจุดสองจุด เหมาะกับข้อมูลที่มีลักษณะ Time Series ทั้งนี้ ในโปรแกรม Orange Data Mining เราสามารถ Impute ข้อมูลได้ โดยกดคลิกเลือก Impute ที่บริเวณแถบด้านซ้ายมือ จากนั้นลากเส้นเชื่อมระหว่างชุดข้อมูลที่ต้องการ Impute (Data Table) กับ Icon Impute บริเวณ Workspace จากนั้น เลือกวิธีการ Impute ที่ต้องการ เพื่อ Clean Data



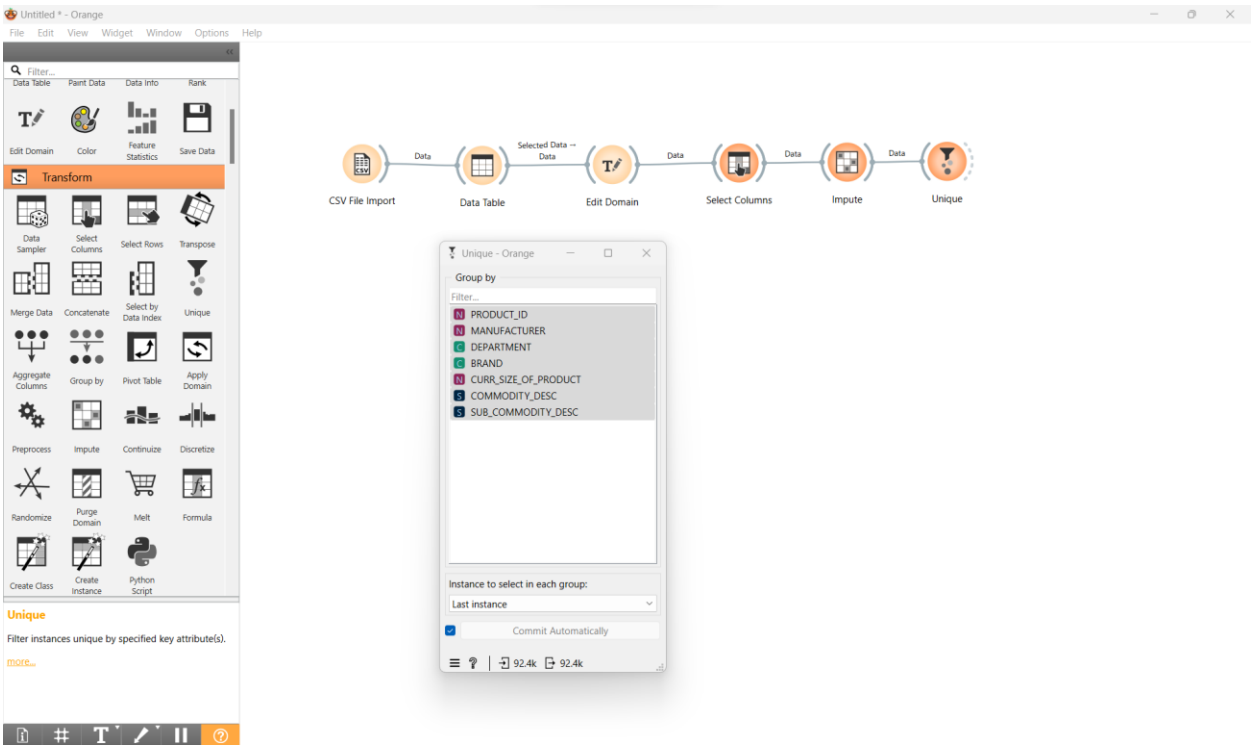
รูปภาพหน้าจอ Data Table แสดงผลไฟล์ข้อมูล CSV ที่นำเข้าสู่โปรแกรม ซึ่งระบุร้อยละของ Missing Data



รูปภาพหน้าจอตัวอย่างการ Impute Missing Values โดยกดเลือก Impute บริเวณแถบด้านซ้ายมือ จากนั้น เชื่อมข้อมูลที่ต้องการเข้ากับ Icon Impute แล้วจึงเลือกวิธีการ Impute (ในที่นี้ ผู้สอนให้เลือก Simple Tree)

นอกจากนี้ การ Clean Data ยังต้องจัดการกับกรณีที่มีชุดข้อมูลบันทึกซ้ำกัน เพื่อป้องกันความคลาดเคลื่อนในการวิเคราะห์ข้อมูล โดยในโปรแกรม Orange Data Mining มีฟังก์ชั่น Unique เพื่อใช้สำหรับการจัดการชุดข้อมูลที่มีการบันทึกซ้ำกัน โดยกดคลิกที่ปุ่ม Unique บริเวณแถบด้านซ้ายมือ จากนั้น จะปรากฏ Icon Unique บริเวณ Workspace และให้ผู้ใช้งานเชื่อมเส้นระหว่างข้อมูลที่ผ่านการ Impute แล้ว โดยเลือกที่ Icon Impute และเชื่อมเส้นเข้ากับ Icon Unique จากนั้น ให้ Double Click ที่ Icon Unique เพื่อทำการเลือก Clean Data ทั้งหมดที่มีการบันทึกซ้ำกันออกไป (ตัวอย่างรูปภาพบริเวณท้ายหน้า)

ในส่วนของการจัดการข้อมูลที่มีความแตกต่างจากค่าส่วนมาก (Outliers) ผู้สอนได้แนะนำให้ผู้เรียนตรวจสอบข้อมูลดังกล่าวผ่านการ Visualize ก่อนว่า ลักษณะของข้อมูล Outliers อยู่ในตำแหน่งใด เมื่อเทียบกับข้อมูลทั้งหมด เป็นข้อมูลสำคัญหรือค่าผิดพลาดจากการกรอกข้อมูลหรือไม่ จากนั้น จึงตัดสินใจว่าจะดำเนินการอย่างไรกับข้อมูลดังกล่าว อาทิ เงินเดือนของนายกรัฐมนตรีประเทศสิงคโปร์ เมื่อเทียบกับเจ้าหน้าที่ทั่วไป ถือว่าสูงมาก กล่าวได้ว่า เป็นข้อมูล Outlier แต่ถือว่าเป็นข้อมูลที่สำคัญ ไม่ได้เกิดจากการกรอกผิดพลาด โดยวิธีการ Clean Data ทำได้หลายวิธี เช่น การลบชุดข้อมูลที่มี Outlier ออกทั้งแถว การเปลี่ยนสเกลข้อมูล เช่น Log Transform มักใช้ในกรณีที่ข้อมูลมีความ Skew สูง หรืออาจใช้วิธีการ Impute เช่นเดียวกัน เช่น แทนที่ค่า Outliers ด้วย Mean, Median หรือ Mode อย่างความเหมาะสม โดยสรุป ในชั้นเรียนนี้ ผู้สอนสาธิตการใช้งานโปรแกรม Orange Data Mining พื้นฐาน และเทคนิควิธีการ Clean Data หรือการจัดการกับข้อมูลที่มีปัญหาในเบื้องต้น ก่อนนำข้อมูลที่จัดการแล้ว ไปวิเคราะห์ต่อไป



รูปภาพหน้าจอตัวอย่างการกำจัดชุดข้อมูลที่มีการบันทึกซ้ำกัน (Duplicate Records) โดยใช้ฟังก์ชั่น Unique

Session 4: Case Studies of AI and Data Analytics in Various Sectors

โดย Quinn Pham, Director of Solution Consulting, Meiro Pte. Ltd., Singapore

ในคลาสเรียนนี้ ผู้สอนได้แนะนำให้ผู้เรียนรู้จักกับองค์ประกอบของ AI และกรณีศึกษาของการนำเทคโนโลยี AI มาใช้แก้ไขปัญหาที่เกิดขึ้นจริง โดย AI ประกอบด้วย 8 คุณสมบัติสำคัญ ได้แก่ (1) Reasoning หรือความสามารถในการให้เหตุผล (2) Decision Making หรือความสามารถในการตัดสินใจบนพื้นฐานของเป้าหมาย (3) Knowledge Representation หรือความสามารถในการจัดเก็บและแสดงผลข้อมูลและองค์ความรู้ (4) Robotics หมายถึง การมีคุณสมบัติในการช่วยทำภารกิจบางอย่างให้สำเร็จแบบอัตโนมัติ (5) Computer Vision หรือการที่สามารถประมวลข้อมูลในรูปแบบภาพได้ ใกล้เคียงกับรูปแบบการมองเห็นของมนุษย์ (6) Machine Learning หรือ ความสามารถในการเรียนรู้ได้ด้วยตนเอง (7) Natural Language Processing หมายถึง ความสามารถในการเข้าใจและแสดงผลลัพธ์ในรูปแบบภาษามนุษย์ และ (8) Natural Interaction หมายถึง การที่ Machine สามารถมีปฏิสัมพันธ์กับมนุษย์ได้อย่างเป็นธรรมชาติ ซึ่งคุณสมบัติทั้งหมดนี้ถือเป็นพื้นฐานของ AI ในการแก้ไขปัญหาที่เกิดขึ้นจริง

ผู้สอนได้ยกตัวอย่างกรณีศึกษาของการใช้ AI ในหลายอุตสาหกรรม อาทิ การใช้ AI ในภาคเกษตรกรรม อย่างเช่นตัวอย่าง Platform CropX (ศึกษาข้อมูลเพิ่มเติมได้ที่ https://cropx.com/) ที่เป็น Platform สำหรับสนับสนุนการทำเกษตรกรรมให้มีผลผลิตมากขึ้น โดยมี AI ช่วยประมวลผลอยู่เบื้องหลัง หลักการคือการวาง Sensor ที่ตรวจจับคุณภาพดินและปุ๋ย รวมถึงการใช้ภาพถ่ายจากดาวเทียมร่วมกันวิเคราะห์ เพื่อค้นหารูปแบบการรดน้ำและการใส่ปุ๋ยให้เกิด Crop Yield

สูงที่สุด และลดการใช้น้ำอย่างสิ้นเปลือง นอกจากนี้ ยังมีกรณีศึกษาของ FarmWise (ศึกษาข้อมูลเพิ่มเติมได้ที่ https://farmwise.io/) ซึ่งเป็นการพัฒนาหุ่นยนต์ที่สามารถทำงานได้ด้วยตนเอง (Autonomous Robots) โดยใช้หลักการของ Computer Vision และ Machine Learning เพื่อตรวจจับและกำจัดวัชพืช เพื่อลดการใช้ปุ๋ยเคมีและเพิ่มผลผลิตทางการเกษตร

พร้อมกันนี้ ผู้สอนยังได้แนะนำกรณีศึกษาการประยุกต์ใช้งาน AI ในภาคเกษตรกรรมของ Blue River Technology (ศึกษาข้อมูลเพิ่มเติมได้ที่ https://www.bluerivertechnology.com/) ที่มีชื่อว่า See & Spray Technology โดยมีหลักการทำงานคือ การใช้ Computer Vision ในการแยกแยะและตรวจจับวัชพืชและทำการพ่นยากำจัดวัชพืชลงบนวัชพืชเท่านั้น แตกต่างจากวิธีดั้งเดิมคือ การพ่นยาครอบคลุมทั้งอาณาบริเวณ ทำให้เกิดความสิ้นเปลืองและเกิดสารเคมีตกค้างบนผลผลิต ทำให้วิธีนี้ช่วยทำให้เกิดความยั่งยืนในการทำการเกษตรมากขึ้น อีกกรณีศึกษาหนึ่งที่ถูกยกตัวอย่างในชั้นเรียนนี้ คือ AgriWebb หรือ Software ที่ทำงานโดยใช้เทคโนโลยี AI บริหารจัดการฟาร์มปศุสัตว์ (ศึกษาข้อมูลเพิ่มเติมได้ที่ https://www.agriwebb.com/) โดยผู้ใช้งานสามารถติดตามสัตว์ทุกตัว เพื่อติดตามสถานะ น้ำหนัก สุขภาพของสัตว์แต่ละตัว ทำให้บริหารจัดการปศุสัตว์ได้อย่างมีประสิทธิภาพมากขึ้น

กรณีศึกษาการประยุกต์ใช้งาน AI ในมิติด้านการศึกษาที่น่าสนใจได้แก่ Duolingo ซึ่งเป็น Platform การเรียนรู้ด้านภาษาที่เป็นที่นิยม (ศึกษาข้อมูลเพิ่มเติมได้ที่ https://www.duolingo.com/) ซึ่งภายใน Application มีการประยุกต์ใช้ AI เพื่อออกแบบบทเรียนให้สอดคล้องกับระดับของผู้เรียน โดยประเมินจากทักษะและความสามารถด้านภาษาของผู้เรียนแต่ละคน นอกจากนี้ ตัว Application สามารถกำหนด Personalized Learning Path สำหรับผู้เรียนแต่ละคน โดยขึ้นอยู่กับจุดอ่อนทางภาษา ลักษณะข้อผิดพลาด และลักษณะการใช้ภาษาของแต่ละคน รวมทั้งมี Feature ในการ Interactive กับผู้เรียนผ่านบทสนทนาแบบ Role Play ทำให้ผู้เรียนเกิดความรู้สึกสนุกสนานและเพลิดเพลินไปกับการเรียนรู้มากขึ้น

ผู้สอนยังได้นำเสนอกรณีศึกษาของการใช้งาน AI ในบริบทของ Healthcare โดยระบุว่า ในอนาคต โลกของเรามีแนวโน้มที่จะนำเอาเทคโนโลยี AI และ Robot เข้ามาใช้ในด้านการแพทย์มากขึ้น เช่น การใช้ AI ในการสนับสนุนการตัดกรองและวินิจฉัยโรค หุ่นยนต์ช่วยผ่าตัด รวมถึงการวิเคราะห์ภาพถ่ายของผู้ป่วยเพื่อช่วยวินิจฉัยโรค ซึ่งบางส่วนได้มีกรณีศึกษาเช่นนี้เกิดขึ้นจริงแล้ว โดยทีมนักวิจัยของ MIT ได้มีการพัฒนาเครื่องมือ AI ที่สามารถ Detect ลักษณะจุดบนผิวหนัง (Melanoma) เพื่อทำนายอัตราความเสี่ยงของการเกิดโรคมะเร็ง ทำให้สามารถจัดการผู้ป่วยได้มีประสิทธิภาพมากขึ้น นอกจากนี้ยังมีกรณีของการใช้ Application เพื่อช่วยตรวจสอบข้อมูลของผู้ป่วยในรูปแบบ Real Time ทำให้สามารถสนับสนุนการดูแลรักษาผู้ป่วยได้ดียิ่งขึ้น

นอกจากนี้ ยังมีกรณีศึกษาของการใช้งาน AI ในภาคอุตสาหกรรมการเงินและธนาคาร โดยยกตัวอย่างของธนาคาร DBS ซึ่งเป็นธนาคารที่ใหญ่ที่สุดในภูมิภาค South East Asia ที่มีการใช้งาน AI เพื่อสนับสนุนการทำงานหลายด้าน เช่น การใช้ AI ทำหน้าที่ Virtual Assistant เพื่อสนับสนุนการทำหน้าที่ของผู้ให้บริการดูแลลูกค้าของธนาคาร ช่วยบันทึกการสนทนาและแปลงให้อยู่ในรูปของ Text เพื่อนำไปวิเคราะห์ต่อยอด และช่วยลดระยะเวลาในการจัดการลูกค้าแต่ละรายลงได้เฉลี่ย 20% นอกจากนี้ ยังใช้ AI ในการช่วยผลักดัน (Nudge) โดยวิเคราะห์จากลักษณะโปรไฟล์ของลูกค้า ทำให้ลูกค้าของธนาคารมีการออมมากขึ้น ซื้อบริการผลิตภัณฑ์ด้านประกันความเสี่ยงมากขึ้น รวมถึงมีการลงทุนมากขึ้น รวมถึง มีการใช้ AI เพื่อช่วยตรวจจับลักษณะของหนี้สินที่เกิดความเสี่ยงที่จะชำระคืนไม่ได้ (NPL) ก่อนล่วงหน้าได้ถึง 3 เดือน ทำให้ช่วยเหลือลูกหนี้ได้ก่อนผิดนัดชำระ รวมถึงมีการใช้ AI เพื่อสนับสนุนการเติบโตในเส้นทางอาชีพของพนักงาน DBS โดยแนะนำ Roles,

Training และรูปแบบของงาน เพื่อสนับสนุนการเติบโตและการทำหน้าที่ของพนักงานแต่ละคนได้อย่างมีประสิทธิภาพสูงสุด (Put the right man on the right job)

Session 5: Data Collection Techniques

โดย Dr. Chutiporn Anutariya, Associate Professor, Information Management, Department of Information and Communication Technologies, Asian Institute of Technology, Thailand

ในการเรียนการสอนของชั้นเรียนนี้ ผู้สอนได้แนะนำถึงความสำคัญของการเก็บข้อมูล เนื่องจากเป็นขั้นตอนแรกของการทำ Data Analytics หากปราศจากข้อมูลที่มีคุณภาพ การดำเนินการในขั้นต่อไปจะไม่สามารถทำได้ ดังนั้น ผู้สอนจึงปูพื้นฐานแนวคิดและเทคนิคเกี่ยวกับการเก็บข้อมูลให้กับชั้นเรียน

การเก็บข้อมูล (Data Collection) จะต้องมุ่งเน้นไปที่การตอบคำถามหรือการแก้ไขปัญหาทางธุรกิจ องค์กร ชุมชน หรือสังคมที่เรากำลังตั้งเป้าหมายไว้ ดังนั้น จึงจำเป็นต้องเลือกข้อมูลที่เหมาะสม แหล่งที่มาของข้อมูลดังกล่าว ตลอดจนวิธีได้มาซึ่งข้อมูลที่เราต้องการ อาทิ หากต้องการข้อมูลที่เกี่ยวข้องกับยอดขายในร้านค้า เราสามารถทำได้โดยดึงข้อมูลจากระบบ Point of Sales (POS) หากต้องการข้อมูลที่ใช้ประมวลผลเกี่ยวกับความเสี่ยงทางการเงินของลูกค้าธนาคาร เราสามารถเลือกเก็บข้อมูลพฤติกรรมการใช้งานบัตรเครดิตของลูกค้าแต่ละราย หรือหากต้องการข้อมูลที่ใช้ประเมินความเสี่ยงด้านสาธารณสุขในภาพรวม เราสามารถเลือกใช้ชุดข้อมูลประวัติการลงทะเบียนและการรักษาของผู้ป่วยที่อยู่ในระบบข้อมูลของภาครัฐ เป็นต้น ทั้งนี้ ในการเก็บข้อมูล เราสามารถแบ่งได้เป็น 2 รูปแบบ ได้แก่ (1) การเก็บข้อมูลแบบปฐมภูมิ (Primary Data Collection) คือ การเก็บข้อมูลจากแหล่งข้อมูลที่ต้องการโดยตรง เช่น หากต้องการข้อมูลความพึงพอใจของผู้บริโภค เราสามารถเลือกสัมภาษณ์ลูกค้าที่ใช้บริการสินค้าเราโดยตรง และ (2) การเก็บข้อมูลแบบทุติยภูมิ (Secondary Data Collection) หมายถึง การรวมรวมข้อมูลจากแหล่งที่มาที่ถูกจัดเก็บโดยผู้อื่นมาแล้ว เช่น บริษัทวิจัย รัฐบาล สถาบันการศึกษา เป็นต้น

หลักการที่ถือว่าเป็นวิธีปฏิบัติที่เป็นเลิศ (Best Practices) สำหรับการเก็บข้อมูล ได้แก่ (1) การกำหนดวัตถุประสงค์ของการเก็บข้อมูลที่ชัดเจนว่า ต้องการเก็บข้อมูลไปเพื่ออะไร อะไรคือจุดมุ่งหมายหลักในการเก็บข้อมูลครั้งนี้ (2) การเลือกเก็บข้อมูลที่สอดคล้องกับวัตถุประสงค์ของการเก็บข้อมูล (3) การเลือกเครื่องมือและเทคนิคในการเก็บข้อมูลอย่างเหมาะสม (4) การออกแบบชุดคำถาม (Questionnaires) ให้เหมาะสม ในกรณีที่จัดเก็บข้อมูลที่มีลักษณะของข้อคำถาม และ (5) การตรวจสอบและยืนยันความถูกต้องของข้อมูลที่จัดเก็บได้ ทั้งนี้ ในการเก็บข้อมูลควรต้องมีการระมัดระวังในประเด็นด้านจริยธรรม (Ethics) เช่น การเก็บข้อมูลควรมีความโปร่งใส ระมัดระวังในการใช้ข้อมูลที่จัดเก็บได้ไปใช้ในทางที่ไม่เหมาะสม ไม่เปิดเผยข้อมูลของผู้อื่นโดยที่ไม่ได้รับความยินยอม เป็นต้น

รูปแบบข้อมูลที่จัดเก็บสามารถจำแนกออกได้เป็น 3 ประเภท ได้แก่ (1) ข้อมูลแบบมีโครงสร้าง (Structured Data) หมายถึง ข้อมูลที่สามารถจัดเก็บในรูปแบบของตาราง มีโครงสร้างที่ชัดเจน ง่ายต่อการค้นหาและบริหารจัดการ เช่น ข้อมูลประเภทชื่อ อายุ เพศ ยอดขาย เป็นต้น (2) ข้อมูลแบบกึ่งมีโครงสร้าง (Semi-structured Data) หมายถึง ข้อมูลที่มีโครงสร้างบางส่วน แต่ไม่ได้มีลักษณะรูปแบบที่ชัดเจนมาก ข้อมูลมีความยืดหยุ่นสูง เช่น อีเมล ไฟล์ JSON, XML เป็นต้น (3) ข้อมูล

แบบไม่มีโครงสร้าง (Unstructured Data) หมายถึง ข้อมูลที่ไม่มีรูปแบบตายตัวชัดเจน จึงไม่สามารถจัดเก็บในรูปแบบตารางได้ การวิเคราะห์ข้อมูลลักษณะนี้จึงมีความซับซ้อนสูง เช่น ไฟล์ภาพ ไฟล์วีดีโอ ไฟล์เสียง เป็นต้น

ความท้าทายของการเก็บข้อมูลมีหลายมิติ เช่น ข้อมูลที่ต้องการอาจไม่สามารถเข้าถึงได้ (Inaccessible) ข้อมูลที่จัดเก็บอาจสูญหายหรือไม่ครบถ้วน ลักษณะของข้อมูลอาจมีความไม่สอดคล้องกันในรูปแบบ Format (Data Inconsistency) ข้อมูลที่ได้รับมีลักษณะแบบไม่มีโครงสร้าง ทำให้บริหารจัดการได้ยาก ประเด็นด้านความถูกต้องของข้อมูลที่อาจมีปัญหาเรื่องของ Error หรือ Outliers นอกจากนี้ ยังมีประเด็นด้านต้นทุนการจัดเก็บข้อมูลบางประเภทที่อาจสูงเกินไป รวมถึงประเด็นด้านความเสี่ยงของการรั่วไหลของข้อมูลที่จัดเก็บ

เมื่อเราต้องการเริ่มต้นการจัดเก็บข้อมูล เราต้องสามารถระบุแหล่งข้อมูลเป้าหมายที่เราต้องการได้ ไม่ว่าจะเป็นแหล่งข้อมูลภายในองค์กร (Internal Sources) หรือแหล่งข้อมูลภายนอกองค์กร (External Sources) ที่มีความหลากหลาย รวมถึงแหล่งข้อมูลประเภท Open Source ที่เปิดให้ทุกคนเข้าถึงได้ อาทิ แหล่งข้อมูลของรัฐบาล Website ของสำนักข่าวและสื่อ รวมถึงองค์กรระหว่างประเทศ เช่น World Bank, WHO ทั้งนี้ ขึ้นอยู่กับลักษณะข้อมูลที่เราต้องการ นอกจากนี้ ผู้สอนยังได้นำเสนอเทคนิคของการทำ Web Scraping หรือการให้ Robot จัดเก็บข้อมูลในหน้า Website เป้าหมาย โดยที่มนุษย์ไม่จำเป็นต้องเข้าไปทำการเก็บข้อมูลแบบ Manual ทำให้สามารถรวบรวมข้อมูลจาก Website ที่หลากหลายได้อย่างรวดเร็ว มีประสิทธิภาพ และลดต้นทุน โดยข้อมูลที่ได้รับสามารถแปลงให้อยู่ในรูปของ File JSON หรือ XLS ขึ้นอยู่กับ Software และการตั้งค่าที่เลือกใช้

ผู้สอนได้นำเสนอเกี่ยวกับเทคนิคการเก็บข้อมูลโดยวิธีการสำรวจ (Survey & Questionnaires) เนื่องจากข้อมูลบางประเภทจำเป็นต้องได้รับจากผู้ที่เกี่ยวข้องโดยตรง เช่น ลูกค้า ผู้ใช้งาน (Users) หรือพนักงาน ดังนั้น การออกแบบข้อคำถามจึงต้องมีหลักการดังนี้ (1) คำถามต้องมีความชัดเจนและเข้าใจง่าย (2) คำถามต้องไม่ถูกออกแบบให้ชี้นำหรือมี Bias (3) ลำดับของข้อคำถามอาจส่งผลต่อคำตอบ จึงต้องพิจารณาเรียบเรียงให้เหมาะสม (4) ทดสอบข้อคำถามในกลุ่มย่อย (Pilot Test) ก่อนนำไปใช้งานจริง และ (5) ระมัดระวังการประมวลผลและการนำเสนอข้อมูล ควรคำนึงถึงสิทธิและความปลอดภัยส่วนบุคคล โดยเครื่องมือที่สนับสนุนการทำ Survey มีหลากหลาย อาทิ Google Forms, Qualtrics, SurveyMonkey เป็นต้น โดยแต่ละเครื่องมือจะมี Feature ที่แตกต่างกันออกไป ขึ้นอยู่กับความต้องการใช้งานและงบประมาณที่มี

นอกเหนือจากการเก็บข้อมูลโดย Web Scraping และ Survey ผู้สอนยังได้แนะนำถึงรูปแบบการเก็บข้อมูลโดยใช้ Sensor (IoT) เพื่อรวบรวมข้อมูลหลายประเภท เช่น คุณภาพอากาศ คุณภาพน้ำ ระดับเสียง การตรวจจับจำนวนช่องจอดรถที่เหลืออยู่ การตรวจจับสัญญาณไฟจราจร ไปจนถึง Sensor ใน Device ส่วนบุคคล เช่น การตรวจจับชีพจร (Heart Rate) การตรวจจับก้าวเดิน เป็นต้น และข้อมูลอีกประเภทอาจได้มาโดยการใช้ Application Programming Interface (API) โดยการเชื่อมระบบเข้ากับแหล่งข้อมูลต้นทางที่ต้องการเข้าถึง (ต้องได้รับอนุญาตจากแหล่งต้นทาง) เช่น การเก็บข้อมูล Post บน Social Media ผ่านการเชื่อม API เข้ากับ Facebook, X (Twitter) หรือ Social Media อื่น

เมื่อได้รับข้อมูลที่ต้องการแล้ว ผู้ปฏิบัติงานจะต้องนำข้อมูลมาดำเนินการต่อ ได้แก่ การตรวจสอบความถูกต้องของข้อมูลที่ได้รับ (Data Cleaning & Preprocessing) การทำ Data Validation การจัดการข้อมูลอย่างเป็นระบบ การวิเคราะห์ข้อมูลขั้นต้น (Exploratory Data Analysis) เพื่อสำรวจ Trends และ Patterns การแปลงข้อมูล (Data Transformation) เฉพาะกรณีที่จำเป็น เช่น ข้อมูลมีความ Skew สูงมาก หลังจากนั้น จึงเข้าสู่การวิเคราะห์และการทำโมเดลข้อมูล การรายงานผล และการติดตามผล

ทั้งหมดนี้ การดำเนินการยังคงขึ้นอยู่กับหลักการพื้นฐาน ได้แก่ การมีเป้าหมายในการเก็บข้อมูลที่ชัดเจน การระบุแหล่งที่มาของข้อมูลให้ชัดเจนว่า ต้องรวบรวมแหล่งข้อมูลจากที่ใดบ้าง การเลือกใช้เทคนิคสำหรับการวิเคราะห์ข้อมูลที่ถูกต้องและเหมาะสม ตลอดจนการระมัดระวังเกี่ยวกับประเด็นด้านจริยธรรมข้อมูล

Session 6: Data Collection Exercises

โดย Dr. Chutiporn Anutariya, Associate Professor, Information Management, Department of Information and Communication Technologies, Asian Institute of Technology, Thailand

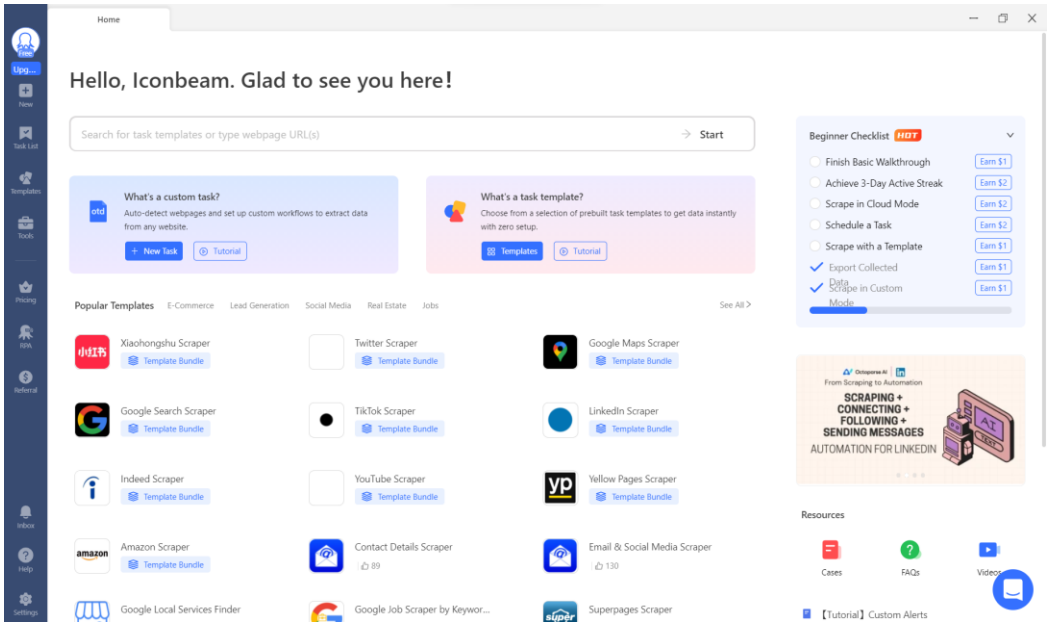
การเรียนการสอนในชั้นเรียนนี้ สืบเนื่องต่อจากหัวข้อใน Session 5 โดยผู้สอนได้สาธิตวิธีการใช้งานโปรแกรมและการเข้าถึงชุดข้อมูลประเภท Open Datasets เพื่อให้ผู้เรียนเกิดความเข้าใจเชิงรูปธรรมมากขึ้น โดยสมมติให้ชั้นเรียนเป็น Data Analyst ที่ต้องการศึกษากลยุทธ์ในการเพิ่มยอดขาย จึงต้องการรวบรวมข้อมูลที่เกี่ยวข้องเพื่อทำ Data Analytics

ในลำดับแรก ผู้สอนได้แนะนำเกี่ยวกับแหล่งข้อมูลภายนอกประเภท Open Datasets ที่มีความน่าเชื่อถือ เช่น ข้อมูลจากหน่วยงานของภาครัฐ (Government Portal) World Bank หรือ Kaggle โดยชุดข้อมูลที่ต้องเข้าไปสืบค้นจะต้องเป็นข้อมูลที่เกี่ยวข้องกับการเพิ่มยอดขาย เช่น รูปแบบการใช้จ่าย Spending Trends ของผู้บริโภคในแต่ละพื้นที่ ข้อมูลที่เกี่ยวข้องกับระดับรายได้หรือขนาดของครอบครัว (Socio-economic Data) ข้อมูลเกี่ยวกับราคาสินค้าคู่แข่งและการจัดโปรโมชั่น ข้อมูลเกี่ยวกับพฤติกรรมการใช้งาน E-Commerce หรือรูปแบบการสั่งซื้อสินค้าของคนในแต่ละพื้นที่ โดยมีหลักในการเลือกข้อมูลที่นำมาใช้คือ (1) แหล่งข้อมูลนั้นต้องมีความน่าเชื่อถือ (2) ข้อมูลต้องมีความเกี่ยวข้องกับเป้าหมายหรือจุดประสงค์ของเรา (3) คุณภาพของข้อมูลต้องถูกต้อง แม่นยำ ครบถ้วน และมีความทันสมัย ในกรณีที่ต้องการข้อมูลเป็นปัจจุบัน

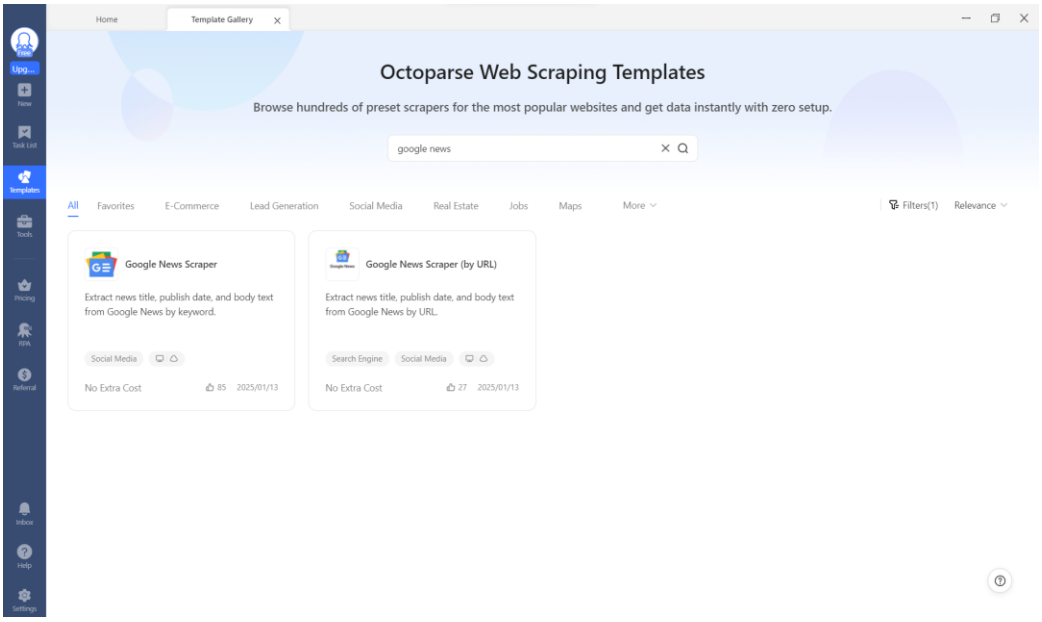
หลังจากนั้น ผู้สอนได้แนะนำโปรแกรม Octoparse ซึ่งเป็นโปรแกรมสำหรับการทำ Web Scraping เพื่อดึงข้อมูลจาก Website เป้าหมายโดยอัตโนมัติ (ดาวน์โหลดได้ที่ https://www.octoparse.com/) โดยตัวโปรแกรมมีข้อดี คือ การออกแบบหน้าจอ Interface ให้สามารถเข้าใจได้ง่าย เหมาะสำหรับผู้ใช้งานที่ไม่ใช่ Developers โดยมีทั้งเวอร์ชั่นฟรี (มีข้อจำกัดในการใช้งานหลายอย่าง) และเวอร์ชั่นเสียเงิน ทั้งนี้ แม้แต่ในโปรแกรมเวอร์ชั่นฟรี มี Template จำนวนมากที่เปิดผู้ใช้งานสามารถใช้ได้โดยไม่มีค่าใช้จ่าย ในที่นี้ ผู้สอนได้สาธิตการใช้ Template Google News Scraper เพื่อใช้ในการค้นหาข่าวที่เกี่ยวข้องกับคู่แข่งธุรกิจ (อ้างอิงรูปภาพในหน้าถัดไป) ซึ่งแสดงผลลัพธ์ของข้อมูลที่เกี่ยวข้องกับคู่แข่งธุรกิจจำนวนมากใน Website ที่แตกต่างกัน โดยที่ผู้ใช้งานไม่จำเป็นต้องเข้าไปค้นหา Manual ทีละ Website ส่งผลทำให้ประหยัดเวลาลงได้มาก

นอกจากนี้ โปรแกรมยังมีความสามารถในการทำ Web Scraping ใน Website เป้าหมายเฉพาะที่เราต้องการค้นหาได้ด้วย ซึ่งสามารถประยุกต์ใช้สำหรับการเปรียบเทียบราคา Competitor Prices (ในที่นี้ ผู้สอนได้นำเสนอตัวอย่างการทำ Web Scraping ใน Website https://www.makro.pro/en/c/search?q=rice แทนที่ผู้ปฏิบัติการเก็บข้อมูลจะต้องเข้า Website นี้ทุกวันและเก็บข้อมูลโดยการบันทึกแบบ Manual การใช้โปรแกรม Octoparse จะช่วยทำให้สามารถเก็บข้อมูลเหล่านี้ได้อัตโนมัติ ทำให้ลดระยะเวลาดำเนินการ ลดต้นทุนจากการที่ต้องใช้แรงงานคนเพื่อเก็บข้อมูล อีกทั้งข้อมูลที่ได้จาก

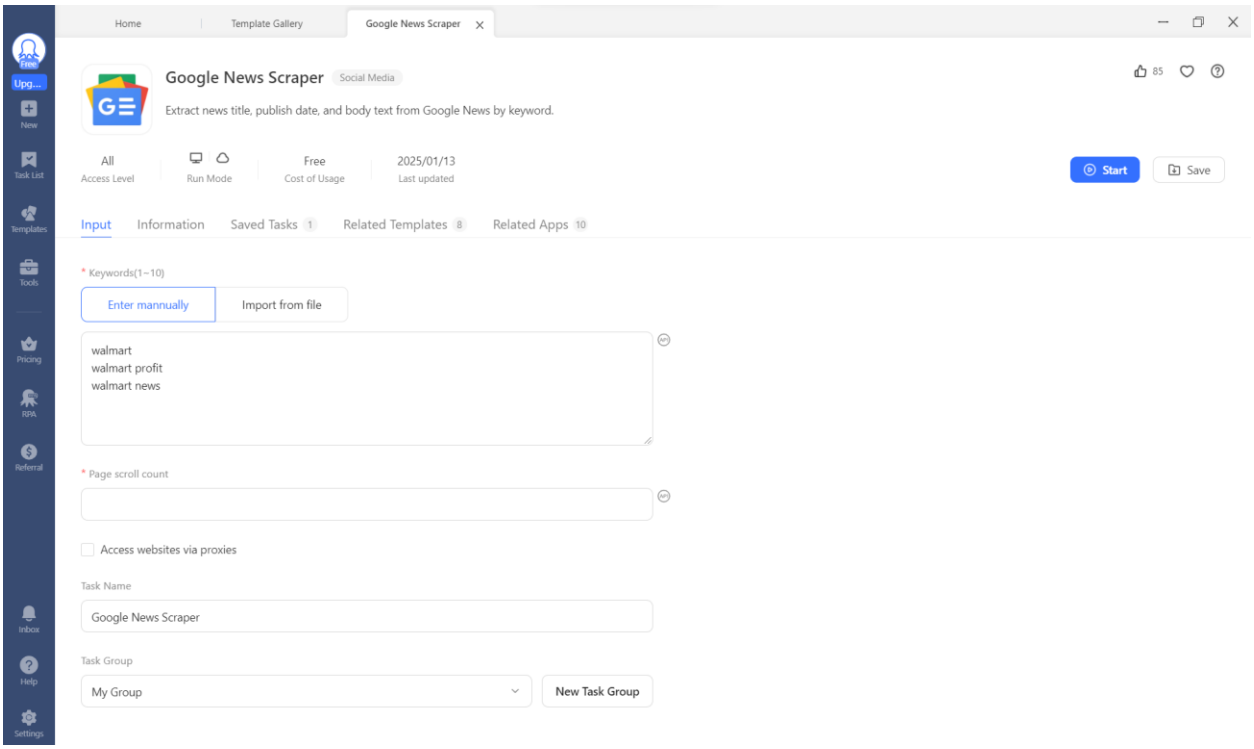
Web Scraping จะมีการจัดโครงสร้างเป็นระบบ อยู่ในรูปแบบของตาราง (ชุดข้อมูลแบบมีโครงสร้าง) ผู้ใช้งานจึงสามารถนำไปวิเคราะห์ต่อยอดได้สะดวกมากขึ้น



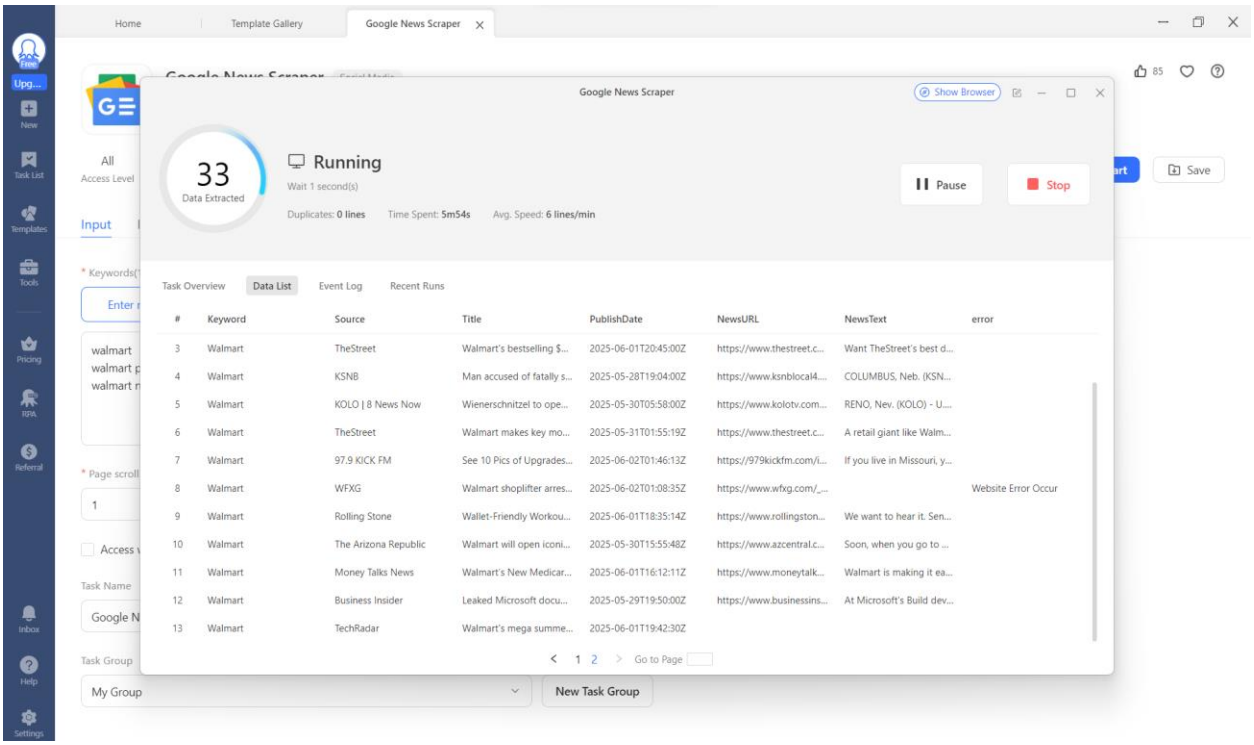
รูปภาพหน้าจอโปรแกรม Octoparse หลังจากเปิดโปรแกรมใหม่



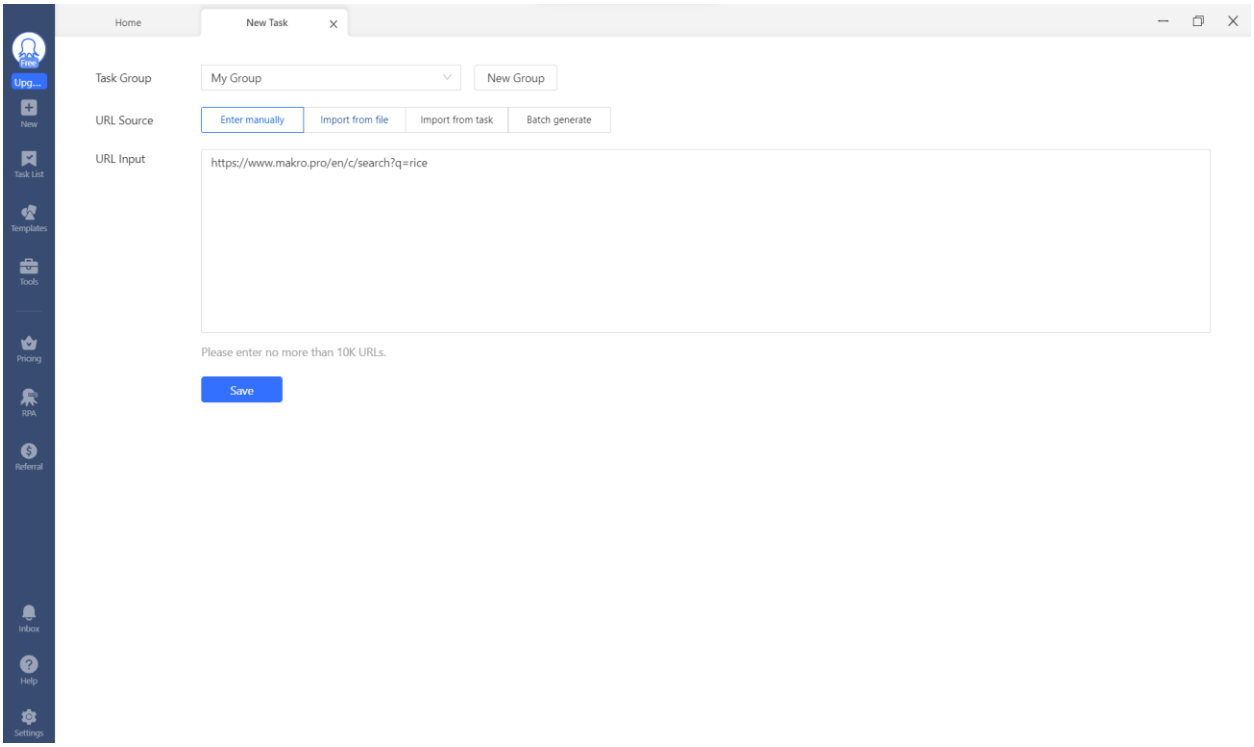
รูปภาพหน้าจอโปรแกรม Octoparse หลังจากที่เราพิมพ์คำว่า "Google News" ลงในกล่องค้นหาและกดปุ่ม Start หน้าจอจะแสดงผล Template ของ Google News Scraper ที่สามารถใช้งานได้ฟรี



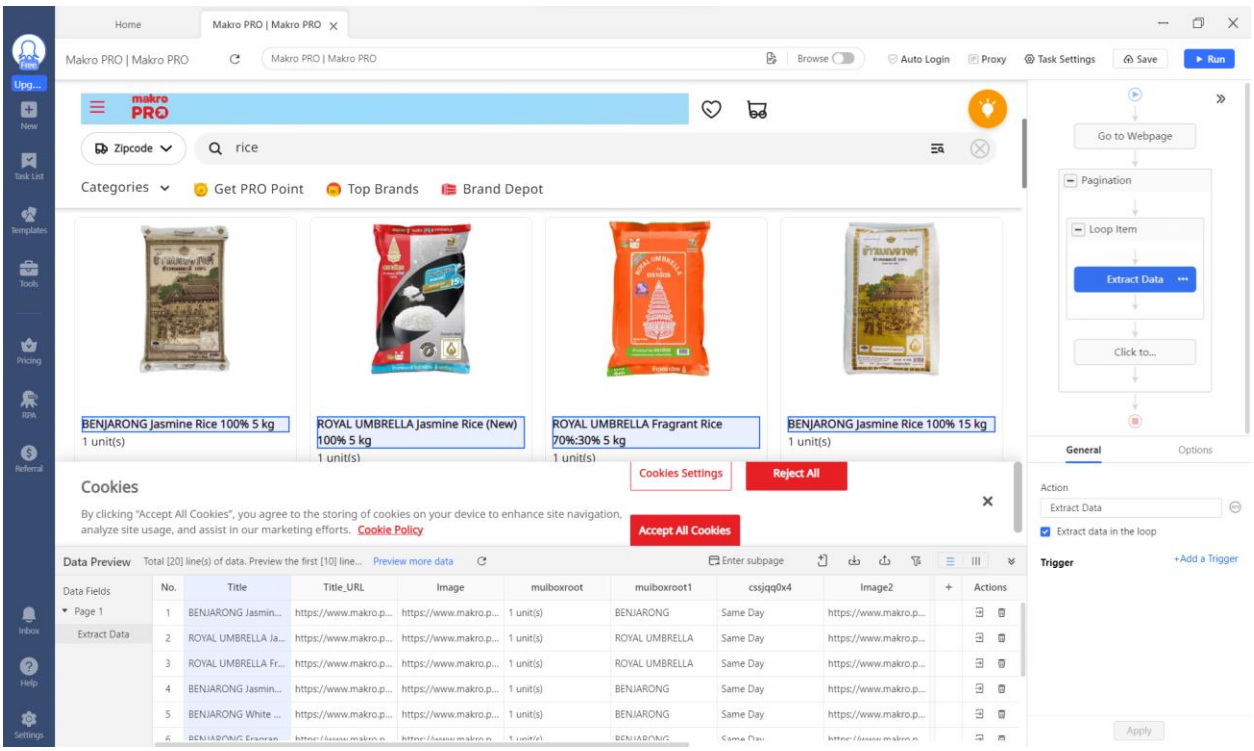
รูปภาพหน้าจอแสดงวิธีการใช้งาน Google News Scraper Template ในโปรแกรม Octoparse โดยพิมพ์ข้อความที่ต้องการให้โปรแกรมค้นหาอัตโนมัติลงในช่องกลาง โดยในที่นี้ ค้นหาคำที่เกี่ยวข้องกับ Walmart



รูปภาพหน้าจอแสดงผลการรันโปรแกรม Octoparse เพื่อค้นหา Source ทั้งหมดที่มีการระบุถึง Walmart



รูปภาพหน้าจอโปรแกรม Octoparse ในการสร้าง Custom Task เพื่อเปรียบเทียบราคากับคู่แข่ง



รูปภาพหน้าจอภายหลังจากระบุ Website เป้าหมาย เพื่อการทำ Web Scraping ในที่นี้ โปรแกรม Octoparse จะทำการระบุโครงสร้างข้อมูลแบบอัตโนมัติ ผู้ใช้งานสามารถกด Run ที่แถบมุมบนขวา เพื่อทำ Web Scraping

Session 7: Data Preprocessing

โดย Dr. Murphy Choy, CEO, Alionova Education, Singapore

การเรียนการสอนในชั้นเรียนนี้ ผู้สอนได้ปูพื้นฐานทฤษฎีที่เกี่ยวข้องกับการทำ Data Preprocessing หรือการจัดการกับข้อมูลภายหลังจากที่จัดเก็บและรวบรวมมาได้ ให้อยู่ในรูปแบบที่สามารถนำไปวิเคราะห์ในขั้นถัดไปได้ เนื่องจากข้อมูลดิบ (Raw Data) อาจมีปัญหาเช่น ข้อมูลไม่ครบถ้วนสมบูรณ์ ข้อมูลบางชุดมีค่าที่แตกต่างจากค่าส่วนใหญ่มาก (Outliers) หรือข้อมูลที่มีการกระจายตัวที่ไม่เป็นปกติ อาทิ ข้อมูลมี Skewness สูงมาก ทำให้ข้อมูลเหล่านี้ต้องได้รับการจัดการอย่างเหมาะสมก่อนที่จะนำไปวิเคราะห์ต่อไป

ในทางทฤษฎี Data Preprocessing หมายถึง การ Clean หรือขจัดข้อมูลที่ไม่เหมาะสม การแปลงข้อมูล และการบริหารจัดการข้อมูลดิบก่อนที่จะนำไปวิเคราะห์ เพราะข้อมูลที่ไม่มีคุณภาพจะนำไปสู่ผลลัพธ์จากการวิเคราะห์ที่ผิดเพี้ยนไป โดยขั้นตอนการทำ Data Preprocessing มีอยู่ด้วยกัน 4 ขั้นตอนหลัก ประกอบด้วย (1) Data Collection หรือการเก็บข้อมูลดิบ (2) Data Cleaning คือ การระบุข้อมูลที่มีความผิดพลาด ไม่ถูกต้อง ไม่ครบถ้วนสมบูรณ์ หรือข้อมูล Outliers และจัดการข้อมูลเหล่านี้อย่างเหมาะสม (3) Data Transformation หมายถึง การแปลงข้อมูลให้อยู่ในรูปแบบการกระจายปกติ หรือแปลงข้อมูลให้พร้อมต่อการวิเคราะห์ (4) Data Reduction หรือการลดทอนมิติของข้อมูลบางส่วน ในกรณีที่ข้อมูลมีมิติจำนวนมาก ทั้งนี้ กระบวนการดังกล่าวยังคงรักษาไว้ซึ่งแกนหลักสำคัญของข้อมูลทั้งหมด

ตัวอย่างปัญหาคุณภาพด้านข้อมูลที่ได้จากการจัดเก็บที่ต้องจัดการก่อน อาทิ ข้อมูลสูญหาย ไม่ครบถ้วน (Missing Values) ข้อมูล Outliers หรือค่าที่มีความแตกต่างไปจากข้อมูลอื่นส่วนใหญ่อย่างมาก ข้อมูลในรูปแบบที่มี Inconsistent Format โดยส่วนใหญ่มักพบกรณีที่รวบรวมข้อมูลจากหลายแหล่ง การจัดเก็บข้อมูลอาจไม่อยู่ในรูปแบบเดียวกัน นอกจากนี้ยังรวมถึง Duplicate Records หรือข้อมูลที่มีค่าซ้ำกันทั้ง Entries ด้วย

หลักการจัดการ Missing Values มีหลายรูปแบบ อาทิ การลบชุดข้อมูลที่มีความไม่สมบูรณ์นั้นทิ้งไป (Deletion Methods) การทำ Impute (Imputation Methods) หรือหมายถึง การแทนที่ค่า Missing Values ด้วยค่าใดค่าหนึ่ง อาทิ ค่าเฉลี่ย ค่ามัธยฐาน เป็นต้น หรืออาจใช้วิธีการ Advanced Method ซึ่งมีหลายวิธี เช่น Machine Learning, Multiple Imputation และ Maximum Likelihood Estimation ทั้งหมดนี้ สามารถจำแนกรายละเอียดปลีกย่อยได้ดังนี้

- Deletion Methods สามารถแบ่งได้ออกเป็น 2 รูปแบบ ได้แก่
    - Listwise Deletion หมายถึง การลบทั้งแถบข้อมูลออกไป ในกรณีที่มีข้อมูลส่วนใดส่วนหนึ่งหายแม้แต่ช่องเดียว ข้อดีคือ ทำได้ง่าย สะดวก แต่จะทำให้ขนาดตัวอย่างลดลงได้มาก หากมีข้อมูลขาดเยอะ
    - Pairwise Deletion หมายถึง การใช้ข้อมูลเฉพาะคู่ตัวแปรที่มีครบ ไม่ได้ลบข้อมูลทั้งแถว เช่น ข้อมูลชุดหนึ่งมี A, B, C แต่ข้อมูล C หายไปเท่านั้น ในกรณีที่ต้องการวิเคราะห์ค่าความสัมพันธ์ระหว่าง A กับ B ก็สามารถใช้ข้อมูลทั้งหมดได้ โดยที่ไม่ต้องลบแถวที่ C หายไปออก
- Imputation Methods สามารถทำได้ทั้งในรูปแบบทั่วไป และรูปแบบขั้นสูง
    - การใช้ค่าเฉลี่ยแทนที่ Missing Value
    - การใช้ค่ามัธยฐานแทนที่ Missing Value
    - การใช้ค่าฐานนิยมแทนที่ Missing Value

- การใช้ค่าคงที่ เช่น 0, 1 แทนที่ Missing Value
- Multiple Imputation หมายถึง การทดลองแทนค่าหลายรูปแบบ แล้ววิเคราะห์ผลลัพธ์ในแต่ละกรณีที่แตกต่างกัน จากนั้นนำผลลัพธ์ทั้งหมดที่ได้ในแต่ละกรณีมารวมกัน เพื่อให้ได้ค่าเฉลี่ยของสถิติที่มีความเสถียร (รูปแบบ Imputation ขั้นสูง)
- K-Nearest Neighbors Imputation หมายถึง การแทนที่ด้วยค่าจากข้อมูลชุดอื่นที่มีความใกล้เคียงกับชุดข้อมูลที่มี Missing Value โดยใช้หลักการทางสถิติขั้นสูง (รูปแบบ Imputation ขั้นสูง)
- Machine Learning Imputation หมายถึง การใช้อัลกอริทึมในการทำนายค่าที่หายไป เช่น Decision Tree, Random Forest, Neural Network มีความแม่นยำสูง แต่ต้องอาศัยความรู้ Data Science ขั้นสูงเช่นเดียวกัน

นอกเหนือจากประเด็นที่เกี่ยวข้องกับ Missing Value แล้ว ข้อมูลที่มีลักษณะ Outliers ต้องได้รับการจัดการอย่างเหมาะสมเช่นกัน โดย Outliers หมายถึงชุดข้อมูลที่มีค่าทางสถิติแตกต่างออกไปจากค่าส่วนใหญ่ในชุดข้อมูลเดียวกันเป็นอย่างมาก ซึ่งอาจเกิดได้จากการวัดค่าที่ผิดพลาด หรือเป็นปรากฏการณ์เฉพาะ ข้อมูลเหล่านี้จะส่งผลทำให้เกิดความผิดเพี้ยนของผลลัพธ์จากการวิเคราะห์ได้ เพราะค่า Outliers จะไปทำการดึงค่ากลางของข้อมูลทั้งหมด รวมถึงทำให้การกระจายตัวข้อมูลเพี้ยนไป และทำให้เกิดข้อสรุปจากการวิเคราะห์ที่ไม่ถูกต้อง

การตรวจสอบ Outliers สามารถทำได้โดยใช้ Visual Methods หรือการสร้างแผนภาพเพื่อดูค่าที่มีความแตกต่างจากค่าอื่นอย่างมาก หรือการใช้เทคนิคทางสถิติเพื่อตรวจสอบได้เช่นกัน โดยเทคนิครูปแบบ Visual Methods ผู้ปฏิบัติงานสามารถใช้เครื่องมือ อาทิ Box Plots, Scatter Plots หรือ Histograms เพื่อดูการกระจายตัวของข้อมูลว่า มีค่าใดที่แตกต่างออกไปจากค่าอื่นอย่างมีสาระสำคัญหรือไม่ ในส่วนของเทคนิคทางสถิติ โดยส่วนใหญ่จะใช้วิธีการ Z-Score หรือ IQR โดยวิธีแรกทำได้ สามารถตรวจสอบ Outliers ได้โดยการดูว่า ค่าใดที่มีความแตกต่างไปจากค่าเฉลี่ยทั้งทางบวกและลบ เกินกว่า 3 Standard Deviation (3 SD) จะถือว่าเป็นค่า Outliers ส่วนวิธี IQR สามารถตรวจสอบได้ โดยแปลงข้อมูลออกเป็น 4 Quartiles และคำนวณ Interquartile Range (IQR) โดยสูตรทางสถิติ และหากพบว่า ค่าใดที่มีค่าน้อยกว่า Quartile 1 – (1.5 x IQR) หรือ ค่าใดที่มีค่าสูงกว่า Quartile 3 + (1.5 x IQR) จะถือว่า ค่าดังกล่าวเป็น Outliers

สำหรับการจัดการ Outliers สามารถทำได้หลายวิธี เช่น (1) Removal Strategy คือ การกำจัดค่า Outliers ทิ้งออกไป โดยถือว่า Outliers เป็นค่าที่ไม่ถูกต้อง (Error) (2) Data Transformation หรือการแปลงข้อมูลให้อยู่ในรูปแบบ Logarithmic หรือ Square Root เพื่อลดความ Skew ของข้อมูล (3) Value Capping หรือการกำหนดค่าสูงสุด/ต่ำสุด เมื่อค่า Outliers เกินกว่าค่า Cap ที่กำหนด จะถือว่า มีค่าเท่ากับ Cap

นอกจากนี้ ผู้สอนได้นำเสนอวิธีการแปลงข้อมูล (Data Transformation) ด้วยวิธีการที่หลากหลาย เช่น Min-Max Scaling หรือ การแปลงข้อมูลทุกค่าให้อยู่ในช่วง 0 – 1 โดยใช้สูตร ค่าที่แปลงแล้ว = (ค่าเดิม – ค่าต่ำสุดในชุดข้อมูล)/(ค่าสูงสุดในชุดข้อมูล - ค่าต่ำสุดในชุดข้อมูล) การใช้ Z-Score Normalization คือ การแปลงข้อมูลให้อยู่ในรูปแบบที่ค่าเฉลี่ย = 0 และค่า Standard Deviation = 1 การแปลงข้อมูลโดยใช้ Logarithmic Transformation โดยใช้ Log Function เพื่อแปลงข้อมูล ทำให้ค่าที่มีขนาดใหญ่มากลดลง ช่วยลด Skewness ของข้อมูลที่มีความเบ้ไปทางขวาสูง

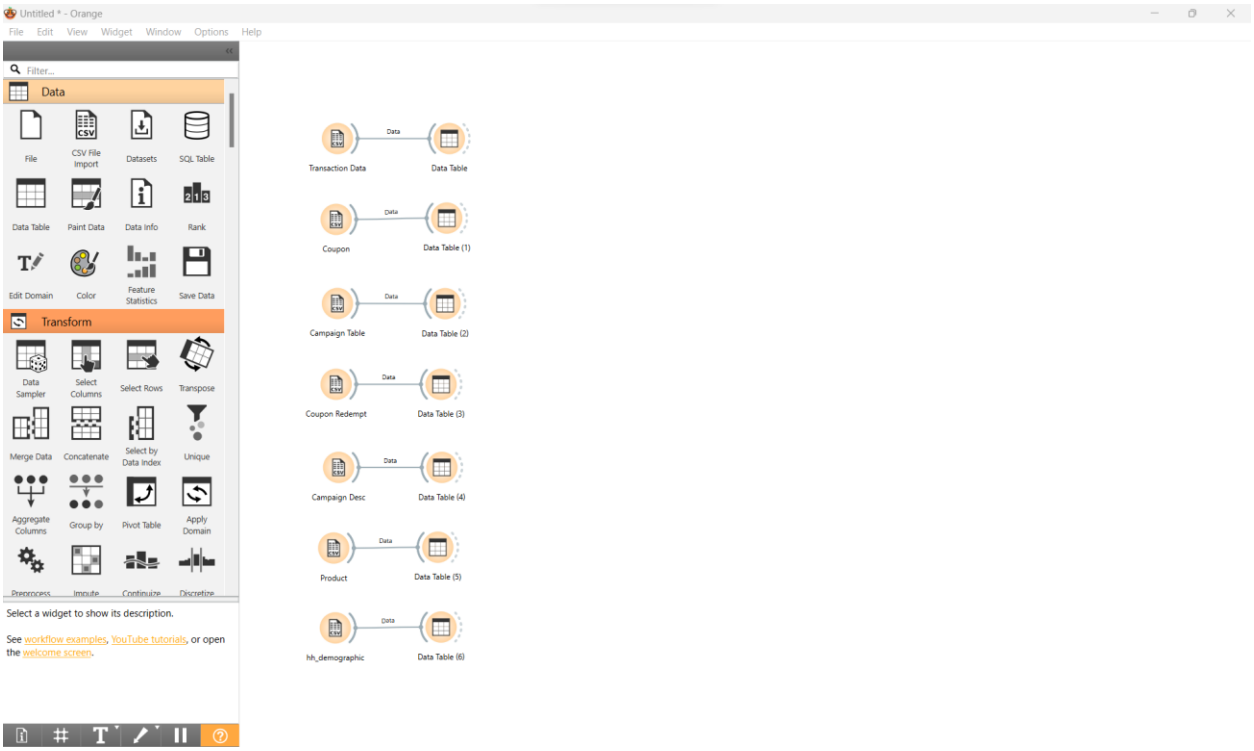
Session 8: Data Preprocessing Exercises

โดย Dr. Murphy Choy, CEO, Alionova Education, Singapore

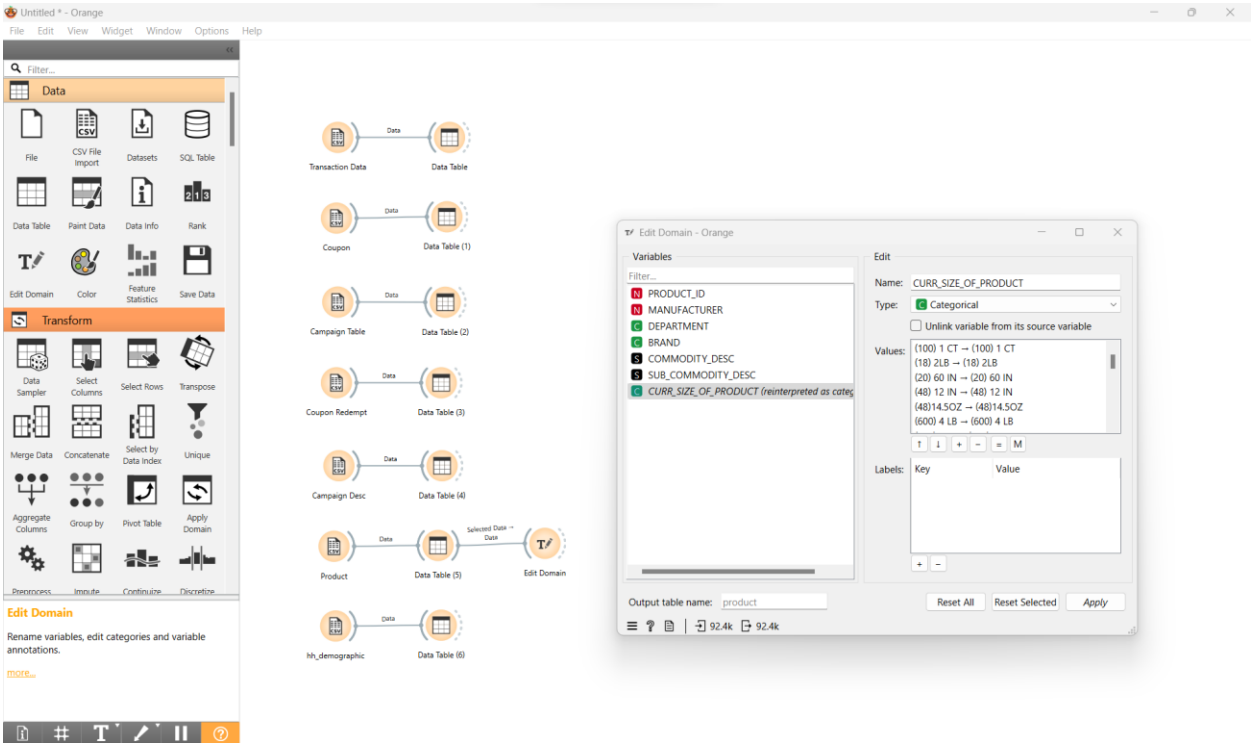
ในการเรียนการสอนช่วงนี้ ผู้สอนได้สาธิตวิธีการใช้งานโปรแกรม Orange Data Mining สืบเนื่องต่อจาก Session 3 โดยเริ่มต้นจากกระบวนการนำเข้าข้อมูล (Import CSV File) เช่นเดียวกับที่เคยทำไปเมื่อช่วง Session 3 แต่มีการเปลี่ยนแปลงการนำเข้าข้อมูลที่มีจำนวนไฟล์มากขึ้น (อ้างอิงรูปภาพในหน้าถัดไป) จากนั้น จะต้องดำเนินการตรวจสอบไฟล์ข้อมูลที่นำเข้าแต่ละชุดว่า มีความถูกต้องสมบูรณ์หรือไม่ เพื่อจัดการกับ Missing Values (ขั้นตอนการ Clean Data) โดยพบว่า มีไฟล์ข้อมูลบางชุด (ในตัวอย่างที่ผู้สอนสาธิตคือ File ข้อมูลเกี่ยวกับ Product) ที่มี Missing Value จึงต้องทำการ Impute ข้อมูลก่อนนำไปวิเคราะห์ต่อ อย่างไรก็ตาม ในบางกรณี ข้อมูลที่อยู่ในไฟล์อาจอยู่ในรูปของ Metatext ทำให้ไม่สามารถนำไปใช้งานต่อได้ทันที จึงต้องทำการแปลงข้อมูล Metatext ให้เป็น Categorial Data โดยใช้ฟังก์ชั่น Edit Domain โดยเลือกจากแถบด้านซ้าย จากนั้นเลือกชุดข้อมูลที่มี Missing Data และต้องการแปลงจาก Metatext เป็น Categorical Data จากนั้นกด Apply เพื่อเสร็จสิ้นขั้นตอนการแปลงประเภทข้อมูล หลังจากนั้น ทำตามขั้นตอน Impute ที่ผู้สอนเคยได้สาธิตแล้วใน Session 3 เพื่อแทนค่า Missing Values ก่อนนำไปวิเคราะห์ต่อ

จากนั้นผู้สอนได้แนะนำการเลือกใช้ฟังก์ชั่น Select Column โดยเลือกจากแถบเมนูด้ายซ้าย จะปรากฏ Icon Select Column ขึ้นบริเวณ Workspace จากนั้นให้เชื่อมต่อเส้นระหว่าง Data Table กับ Select Column Icon เมื่อกด Double Click จะปรากฏหน้าต่างที่ให้เลือก Column ชุดข้อมูลที่ต้องการ (ในขั้นตอนนี้ ผู้สอนได้สาธิตกับตัวอย่างข้อมูล Transaction Data) หลังจากที่เลือกชุดข้อมูลบางส่วนแล้ว ผู้สอนได้แนะนำวิธีการใช้งานฟังก์ชั่น Group by เพื่อประมวลข้อมูลในรูปแบบ Aggregate (Mean, Sum) และได้แสดงผลลัพธ์ให้ดูว่า หลังจากดำเนินการแล้ว ตารางข้อมูลมีการแสดงผลอย่างไร (อ้างอิงรูปภาพในหน้าถัดไป)

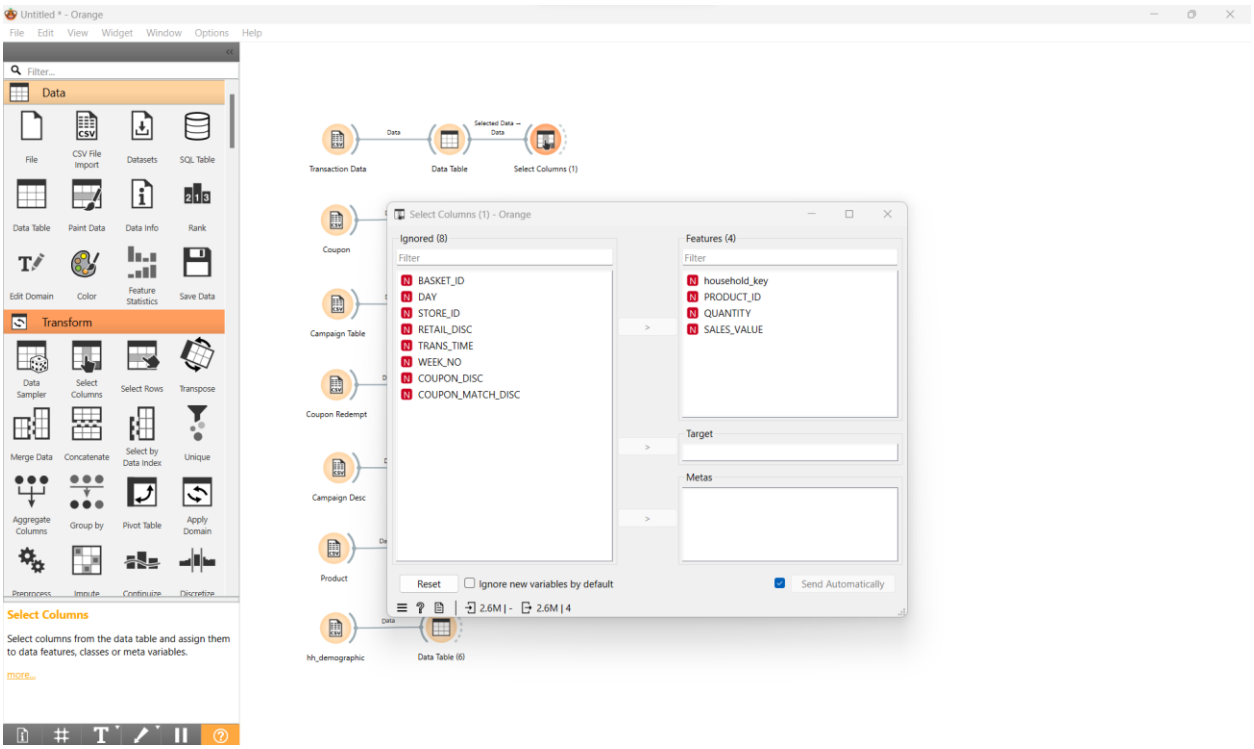
ระหว่างการเรียนการสอน ผู้สอนได้แนะนำให้ผู้เรียนทุกคนทดลองปฏิบัติจริงร่วมด้วย เพื่อความเข้าใจในการใช้งานโปรแกรมมากขึ้น โดยผู้สอนได้สาธิตวิธีการควบรวมตาราง (Merge Data) ระหว่างข้อมูลตาราง 2 ชุด เพื่อสร้างเป็นตารางข้อมูลใหม่ ในที่นี้ ผู้สอนได้ยกตัวอย่างการใช้ household_id (คอลัมน์ข้อมูลในชุดข้อมูลตัวอย่าง) เป็นตัวเชื่อมต่อข้อมูลระหว่างตาราง 2 ชุด ทำให้สามารถสร้างตารางใหม่ผ่านการ Merge



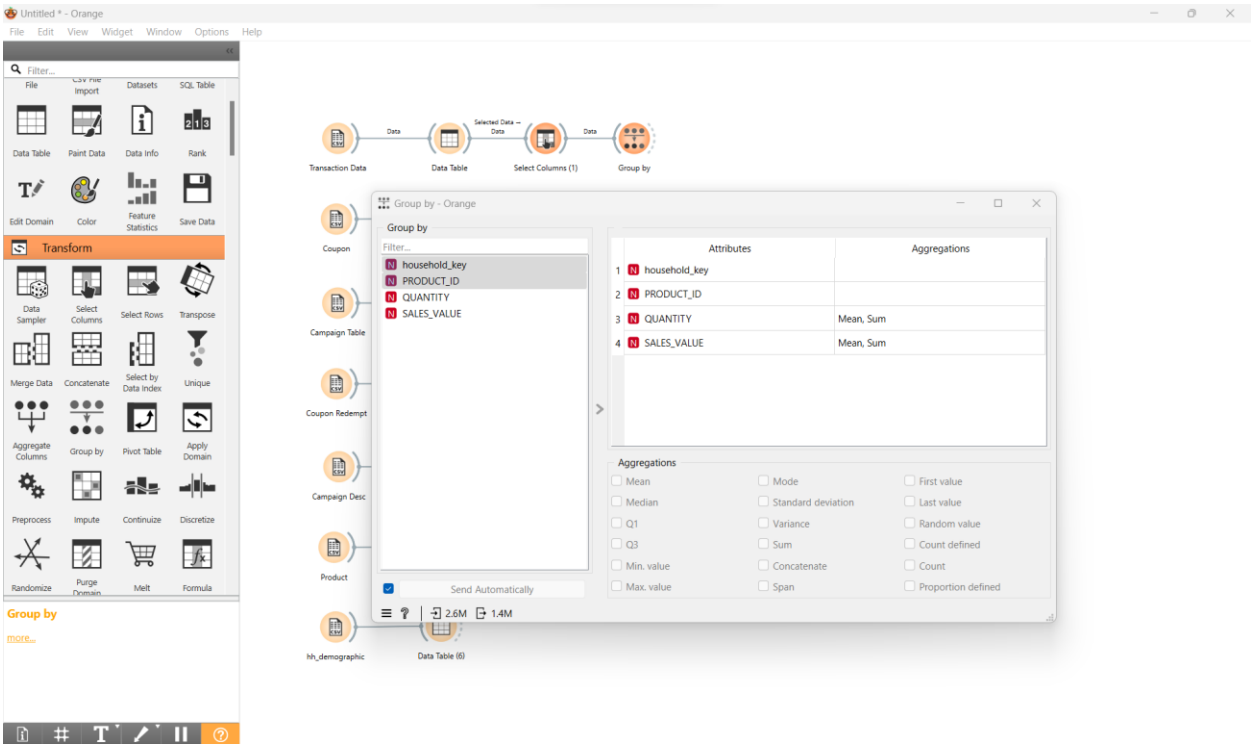
รูปภาพหน้าจอโปรแกรม Orange Data Mining ที่มีการนำเข้าข้อมูล CSV File จำนวน 7 รายการ และเชื่อมต่อข้อมูลกับ Data Table เพื่อแสดงผลลัพธ์ สามารถเปลี่ยนชื่อ Icon ด้วยการคลิกขวาที่ Icon และกด Rename



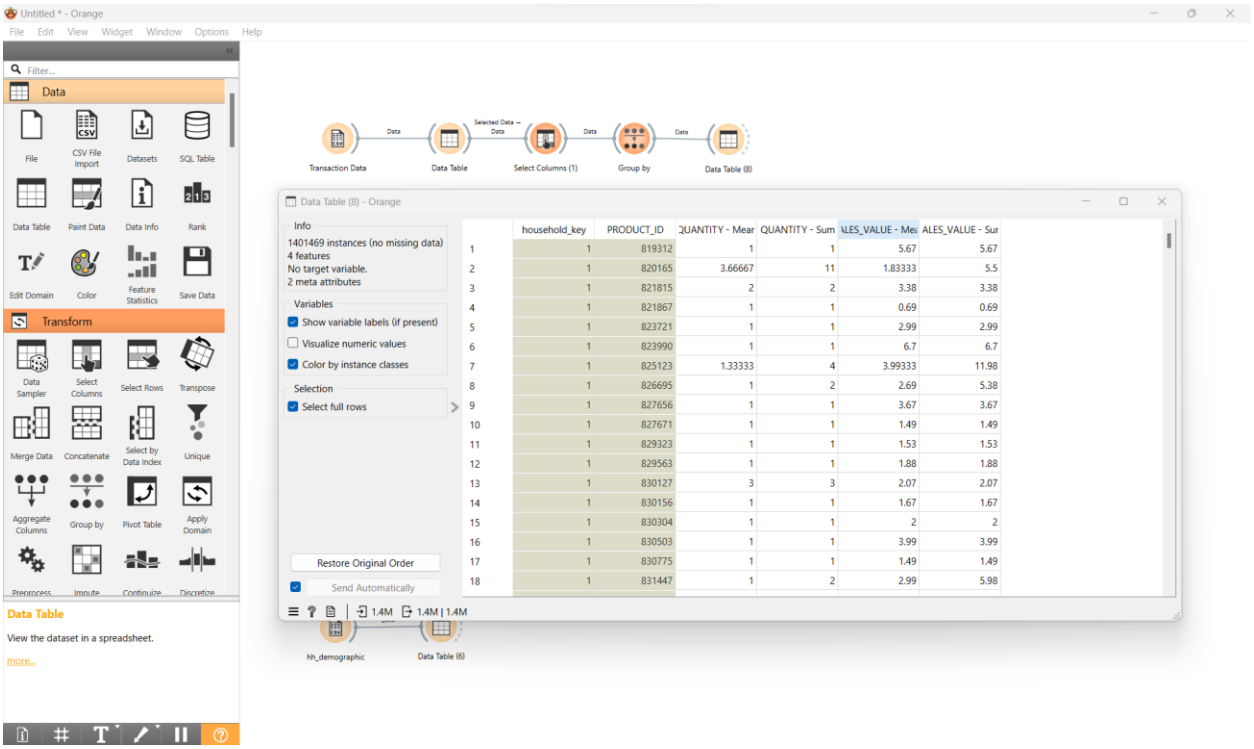
รูปภาพหน้าจอการแปลงลักษณะข้อมูลจาก Metatext เป็น Categorical Data เพื่อให้จัดการข้อมูลต่อได้



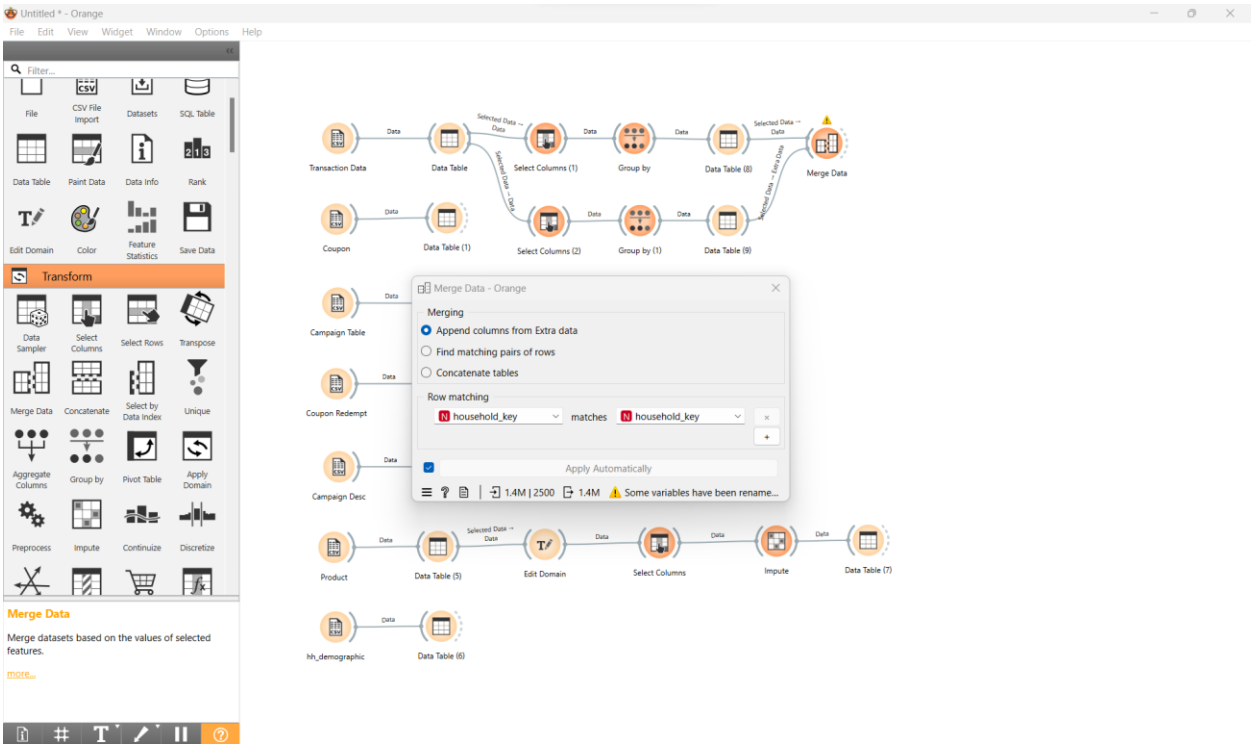
รูปภาพการใช้ฟังก์ชั่น Select Column เพื่อเลือกเฉพาะ Column ข้อมูลบางรายการเพื่อเตรียมประมวลผล



รูปภาพการใช้ฟังก์ชั่น Group by เพื่อประมวลข้อมูล Household (ครัวเรือน) และ Product_ID (สินค้า) ว่ามีจำนวน Quantity และ Sales Value ทั้งในรูปแบบค่าเฉลี่ย (Mean) และรวมทั้งหมด (Sum) เท่าไหร่



รูปภาพตัวอย่างการแสดงผล Data Table หลังจากที่มีการใช้ฟังก์ชั่น Group by เพื่อประมวลข้อมูลแล้ว

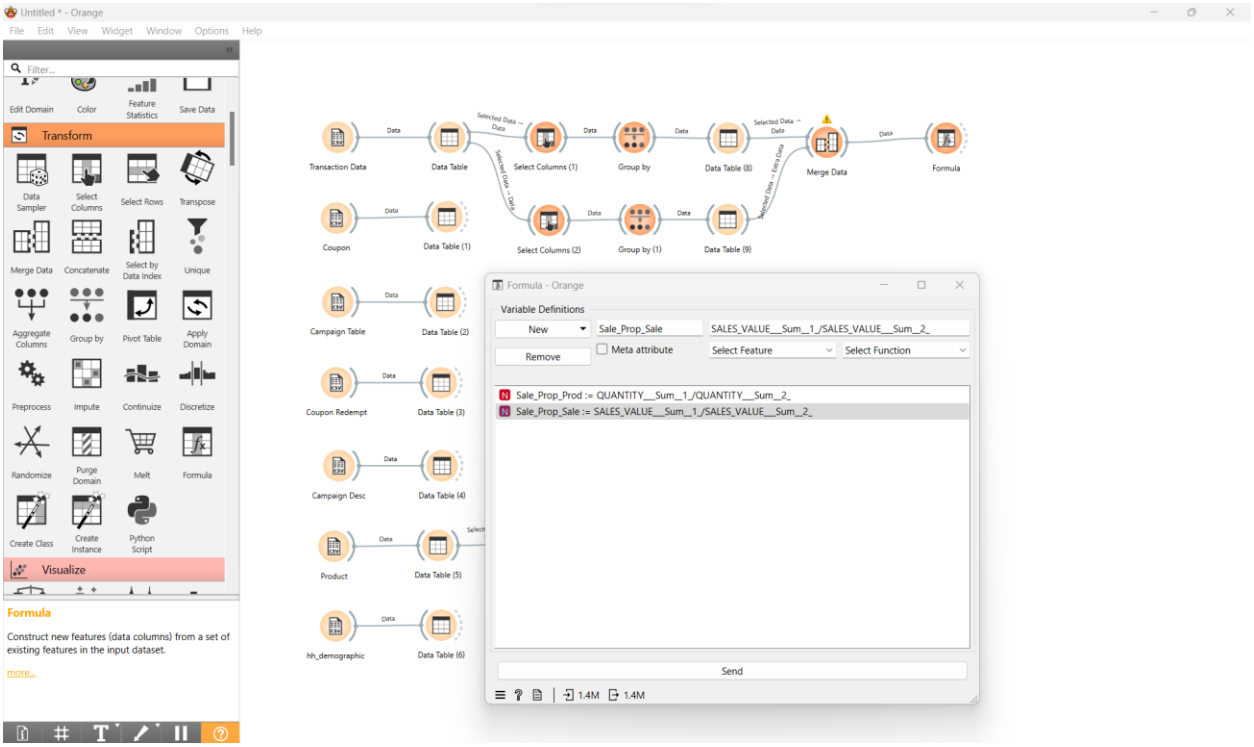


รูปภาพการใช้ฟังก์ชั่น Merge Data โดยมีหน้าที่ในการควบรวมตารางข้อมูล 2 ชุดเข้าด้วยกัน
ในที่นี้ ผู้สอนใช้ household_key เป็นตัวเชื่อมข้อมูลระหว่างตารางเข้าด้วยกัน เพื่อสร้างเป็นตารางใหม่

นอกจากนี้ ผู้สอนยังสาธิตวิธีการ Manipulate ข้อมูล ผ่านการใช้ฟังก์ชั่น Formula โดยเลือกที่แถบด้านซ้ายมือ จากนั้นเชื่อมโยงตารางข้อมูลที่ผ่านการ Merge Data เข้ากับ Icon Formula และ Double Click จะปรากฏหน้าต่างสำหรับทำรายการ ในที่นี้ ผู้สอนได้ให้ผู้เรียนทดลองสร้าง Column ข้อมูลใหม่ ซึ่งเป็นสัดส่วนของปริมาณยอดขาย (Quantity) และมูลค่าการขาย (Sales Value) ของสินค้าแต่ละชนิด สำหรับแต่ละครัวเรือน เมื่อเทียบกับปริมาณปริมาณยอดขายและมูลค่าการขายของสินค้าทุกรายการ โดยสามารถทำได้โดยไปที่ Variable Definitions จากนั้นเลือก Click ที่แถบ New และเลือก Numeric จากนั้น ให้ทำการกด Select Feature เพื่อระบุตัวแปรที่ต้องการใช้คำนวณ (ในที่นี้คือ QUANTITY_SUM_1 และ QUANTITY_SUM_2 สำหรับตัวแปรใหม่ Sale_Prop_Prod และ SALES_VALUE_SUM_1 และ SALES_VALUE_SUM_2 สำหรับตัวแปรใหม่ Sale_Prop_Sale) และพิมพ์เครื่องหมาย / (หาร) ลงไประหว่างตัวแปรทั้งสอง จากนั้นกดปุ่ม Send เพื่อสร้าง Column ข้อมูล Sale_Prop_Prod และ Sale_Prop_Sale เพิ่มในตารางข้อมูลที่มีอยู่แล้ว (อ้างอิงรูปภาพด้านล่าง)

ทั้งนี้ หากผู้ใช้งานต้องการตรวจสอบตารางข้อมูลที่ผ่านการแปลงข้อมูลในขั้นตอนต่างๆ อาทิ Merge Data, Formula สามารถทำได้โดยการเลือกฟังก์ชั่น Data Table และลากเส้นเชื่อมจาก Icon ที่ต้องการไปยัง Icon Data Table จะปรากฏตารางข้อมูลให้ผู้ใช้งานตรวจสอบได้

ในการเรียนการสอนคลาสนี้ ผู้สอนมุ่งเน้นการปฏิบัติจริงของผู้เรียน โดยให้ผู้เรียนทุกคนทดลองปฏิบัติตามเนื้อหา เมื่อเกิดความไม่เข้าใจส่วนใด สามารถสอบถามได้ ทำให้เกิดความกระตือรือร้นในการทดลองทำจริง



รูปภาพการใช้ฟังก์ชั่น Formula เพื่อสร้าง Column ข้อมูลใหม่ โดยคำนวณจากข้อมูลเดิมที่มีอยู่แล้ว

Session 9: Data Analysis Techniques

โดย Dr. Murphy Choy, CEO, Alionova Education, Singapore

สำหรับการเรียนการสอนในคลาสนี้ ผู้สอนได้แนะนำให้รู้จักเกี่ยวกับเทคนิคของการวิเคราะห์ข้อมูล ซึ่งถือเป็นขั้นตอนที่ต้องดำเนินการหลังจากการเก็บข้อมูล และการทำ Data Preprocessing เรียบร้อยแล้ว โดยขั้นตอนนี้ถือเป็นขั้นตอนที่มีความสำคัญ เนื่องจากผู้ปฏิบัติงานจะสามารถค้นพบ Insight จากข้อมูลได้ ซึ่งสามารถนำมาซึ่งข้อสรุปและสนับสนุนการตัดสินใจขององค์กร/ผู้บริหารได้ ทั้งนี้ เนื่องจากลักษณะของข้อมูลมีหลากหลายแตกต่างกันไป ทั้งแบบที่มีโครงสร้าง (Structured Data) และแบบที่ไม่มีโครงสร้าง (Unstructured Data) การวิเคราะห์ข้อมูลทั้งสองรูปแบบจึงต้องใช้เทคนิควิธีการที่แตกต่างกัน ดังนั้น ก่อนดำเนินการ เราจึงควรศึกษาลักษณะรูปแบบข้อมูลที่เรามี เพื่อให้สามารถเลือกใช้วิธีการวิเคราะห์ข้อมูลได้อย่างเหมาะสม

รูปแบบของการวิเคราะห์ข้อมูลมีหลายแบบ อาทิ (1) Descriptive หมายถึง การศึกษาข้อมูลเพื่อค้นหา Pattern และสรุปผลว่า ภายในชุดข้อมูลเกิดอะไรขึ้น (2) Diagnostic หมายถึง การตรวจสอบรูปแบบของ Trend ข้อมูลเพื่อศึกษาหาความสัมพันธ์ระหว่างชุดข้อมูล อาทิ Correlation (สหสัมพันธ์) และ Causal Relationship (ความสัมพันธ์เชิงสาเหตุ) (3) Predictive หมายถึง การใช้เทคนิคทางสถิติและโมเดลสถิติ รวมถึง Machine Learning เพื่อพยากรณ์ผลลัพธ์ในอนาคต (4) Prescriptive หมายถึง การนำผลลัพธ์จากการวิเคราะห์มานำเสนอ Optimal Actions (สิ่งที่ควรทำมากที่สุด) เพื่อให้ได้ผลลัพธ์ตามเป้าหมายสูงสุด

ผู้สอนได้นำเสนอแนวคิดด้านสถิติพื้นฐาน เพื่อให้ผู้เรียนเกิดความเข้าใจ อาทิ Concept เกี่ยวกับค่ากลาง ได้แก่ ค่าเฉลี่ย (การนำค่าข้อมูลทั้งหมดมารวมกันหารด้วยจำนวนข้อมูลทั้งหมด) ค่ามัธยฐาน (ค่ากึ่งกลางของข้อมูลที่แบ่งครึ่งระหว่างกลุ่มชุดข้อมูลที่มีค่าสูง 50% บน และกลุ่มชุดข้อมูลที่มีค่าต่ำ 50% ล่าง) ค่าฐานนิยม (ค่าที่ปรากฏบ่อยมากที่สุดในชุดข้อมูล) นอกจากนี้ ยังนำเสนอ Concept เกี่ยวกับการกระจายตัวของข้อมูล (Dispersion) ซึ่งสามารถวัดค่าได้หลายรูปแบบ อาทิ Range หรือพิสัย หมายถึง ค่าความแตกต่างระหว่างค่าที่มากที่สุดในชุดข้อมูลและค่าที่น้อยที่สุดในชุดข้อมูล (Max – Min) ค่าความแปรปรวน (Variance) หมายถึง ค่าเฉลี่ยของความห่างของข้อมูลจากค่าเฉลี่ย และค่าส่วนเบี่ยงเบนมาตรฐาน (Standard Deviation) ที่เป็นมาตรวัดที่นิยมใช้มากที่สุด โดยการถอดรากที่สองของค่าความแปรปรวน

Concept ทางสถิติอื่นที่น่าสนใจ ได้แก่ (1) การกระจายตัวของชุดข้อมูล ซึ่งมีหลายรูปแบบ เช่น Normal Distribution (การกระจายตัวแบบมาตรฐาน) ในรูปแบบทรงระฆังคว่ำ นอกจากนี้ ยังมี Binomial Distribution ที่เหมาะสำหรับการตรวจสอบเหตุการณ์ที่มีความเป็นไปได้ 2 รูปแบบ เช่น การโยนเหรียญหัวก้อย และ Poisson Distribution ที่ใช้กับการหาความน่าจะเป็นของเหตุการณ์ที่มีลักษณะเฉพาะและเหตุการณ์เป็นอิสระจากกัน เช่น โอกาสที่รถชนในแต่ละสี่แยกไฟแดง เป็นต้น (2) การทดสอบสมมติฐาน (Hypothesis Testing) มักจะใช้กับงานวิจัยเพื่อระบุว่า ความเชื่อใดความเชื่อหนึ่งมีหลักฐานสนับสนุนมากพอหรือไม่ ประกอบด้วย 1. Null Hypothesis คือ สมมติฐานตั้งต้นว่า “ไม่มีความแตกต่าง” หรือ “ไม่มีความสัมพันธ์เกิดขึ้น” เช่น ยาตัวใหม่ที่ทดสอบไม่ได้ช่วยให้ผู้ป่วยอาการดีขึ้น 2. Alternative Hypothesis คือ สมมติฐานที่ผู้วิจัยต้องการพิสูจน์ว่า มีความสัมพันธ์ หรือมีความแตกต่างเกิดขึ้นจริง เช่น ยาตัวใหม่ที่ทดสอบช่วยให้ผู้ป่วยอาการดีขึ้น และ Concept เกี่ยวกับ P-Value หมายถึง ความน่าจะเป็นที่จะเกิดผลลัพธ์ หากสมมติว่า Null Hypothesis เป็นจริง ในกรณีที่ P-Value มีค่าน้อย เช่น น้อยกว่า .05 หมายถึง โอกาสที่ผลลัพธ์นี้จะเกิดขึ้นได้มีน้อยมาก หากสมมติว่า

Null Hypothesis เป็นจริง นั่นหมายความว่า มีแนวโน้มที่ Null Hypothesis ไม่เป็นจริงสูง ดังนั้นในทางสถิติ ถือว่า มีนัยสำคัญ จึงมักปฏิเสธ Null Hypothesis

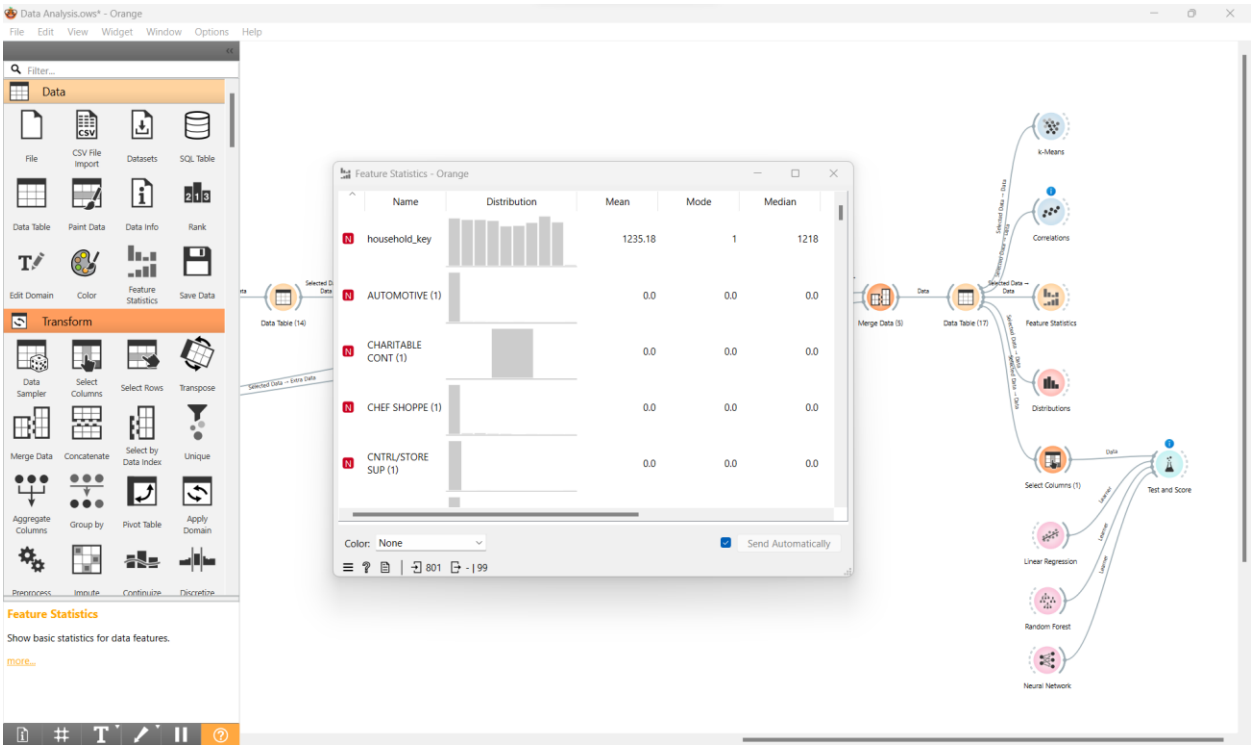
นอกจากนี้ ผู้สอนยังได้นำเสนอเกี่ยวกับเทคนิคทางสถิติ อาทิ Correlation Analysis (Pearson, Spearman), Regression Analysis (Simple, Multiple, Logistic), ANOVA (Analysis of Variance), Clustering Technique (K-Means Clustering, Hierarchical Clustering) รวมถึงอัลกอริทึมขั้นสูง เช่น Random Forests และ Decision Trees ซึ่งถือเป็นเทคนิคในการวิเคราะห์ข้อมูลที่ผู้ปฏิบัติงาน Data Analytics ทุกคนควรรู้ แต่ผู้สอนไม่ได้เจาะลึกลงในรายละเอียดแต่ละหัวข้อ เนื่องจากมีเวลาจำกัด โดยแนะนำให้ผู้เรียนไปศึกษาความรู้เพิ่มเติมเกี่ยวกับหัวข้อเหล่านี้ โดยเฉพาะเมื่อมีความจำเป็นต้องใช้เทคนิคเหล่านี้ เพื่อวิเคราะห์ข้อมูล

Session 10: Data Analysis Exercises

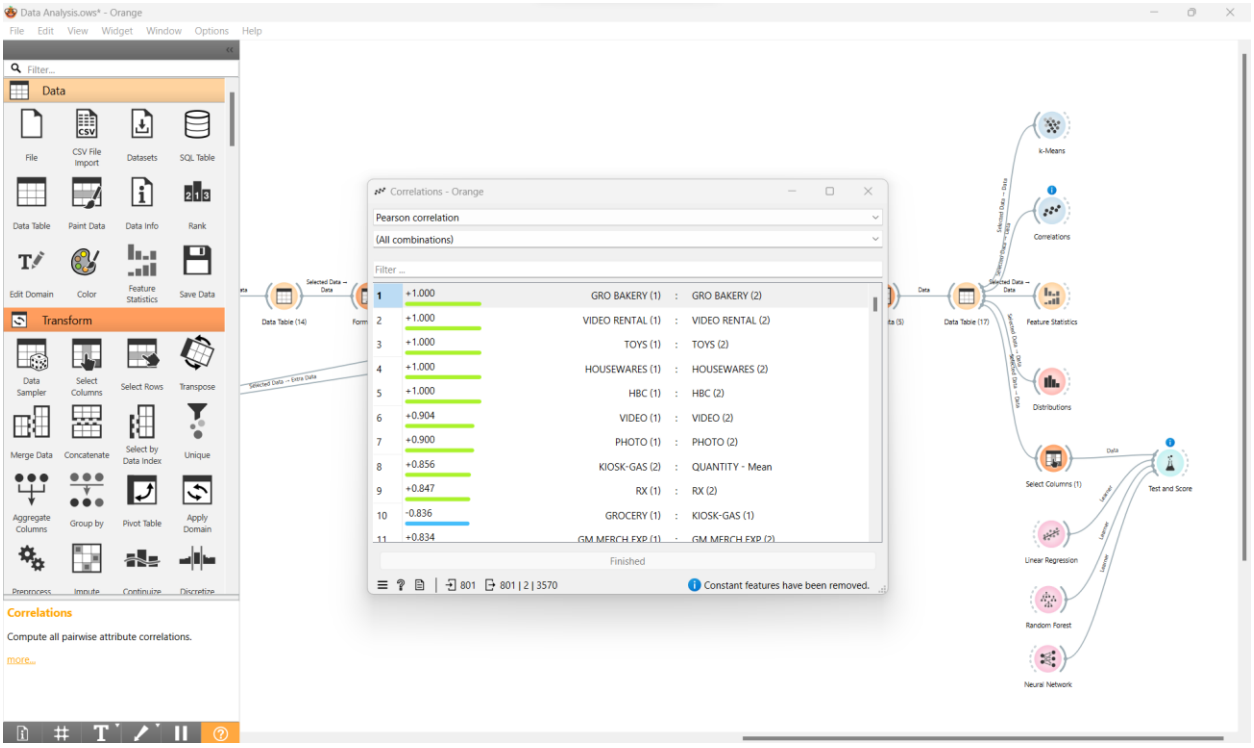
โดย Dr. Murphy Choy, CEO, Alionova Education, Singapore

ในชั้นเรียนคลาสนี้ ผู้สอนได้สาธิตให้ผู้เรียนทำการเชื่อมโยงข้อมูล และ Clean ข้อมูลเพิ่มเติม รวมถึงสร้างตัวแปรใหม่ เช่นเดียวกับใน Session 8 แต่มุ่งเน้นไปที่การใช้ Data Analysis Tools ในโปรแกรม Orange Data Mining ประกอบด้วย 4 ฟังก์ชั่นหลัก ได้แก่ (1) Feature Statistics เพื่อตรวจสอบการกระจายตัวของชุดข้อมูล รวมถึงค่าสถิติพื้นฐาน เช่น Mean, Median, Mode, Min, Max (2) Correlation เพื่อตรวจสอบค่าสัมประสิทธิ์สหสัมพันธ์ (Correlation Coefficient) ระหว่างชุดข้อมูลสองชุด (3) k-Means หรือการทดสอบการลดทอนมิติข้อมูลออกเป็น Clusters โดยดูจาก Silhouette Score ว่า ควรจำแนกออกเป็นกี่ Clusters ที่มีคุณสมบัติของข้อมูลใกล้เคียงกันระหว่างชุดข้อมูลในกลุ่ม Cluster เดียวกัน (ค่ายิ่งมาก ยิ่งดี) และ (4) Test and Score เพื่อใช้ประเมินประสิทธิภาพของโมเดลทำนาย อาทิ Regression, Random Forest, Neural Network เพื่อทำนายค่าของ Dependent Variables (Set Up ค่าตัวแปรดังกล่าวในฟังก์ชั่น Regression, Random Forest, Neural Network) ในที่นี้ R-Square (R2) บ่งบอกถึงประสิทธิภาพการทำนาย (ค่ายิ่งเข้าใกล้ 1 มาก ยิ่งดี) ผู้สอนได้ให้เวลาผู้เรียนได้ทดลองปฏิบัติจริงกับชุดข้อมูล เนื่องจากมีความซับซ้อนสูง

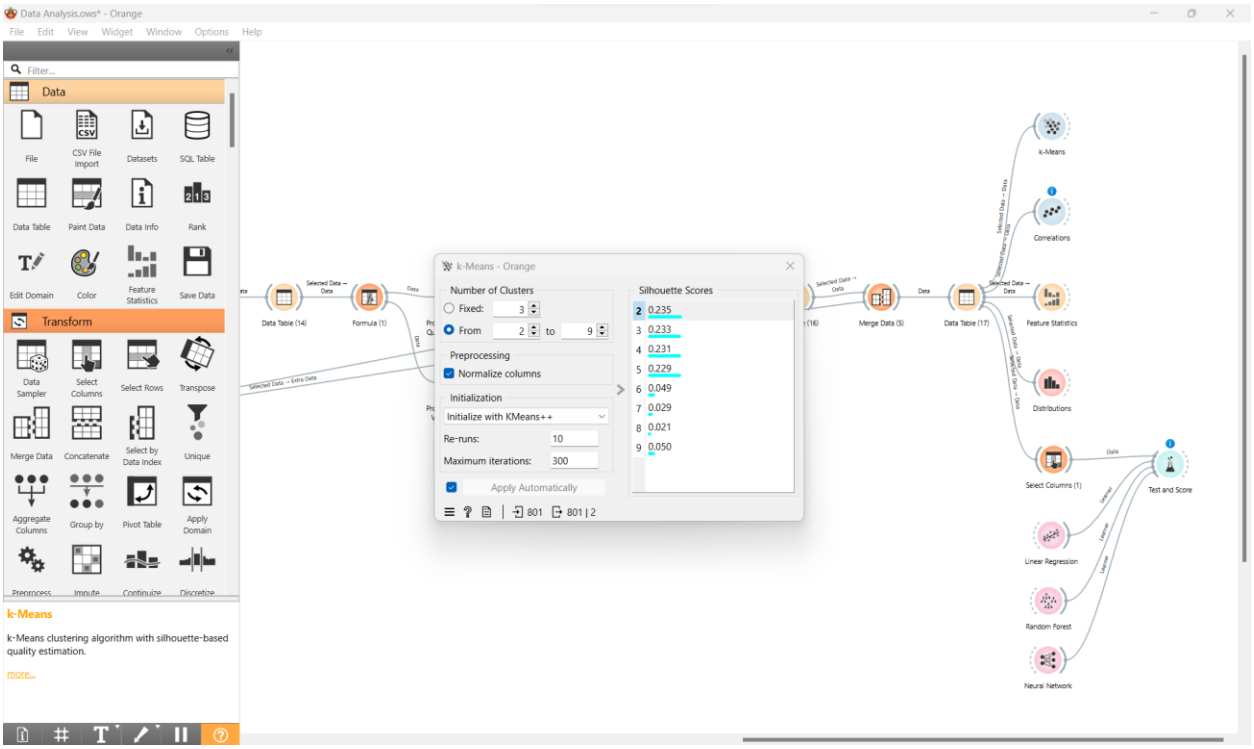
เนื่องด้วยเวลาที่มีค่อนข้างจำกัด ผู้สอนจึงแนะนำให้ผู้เรียนไปทดลองศึกษาและใช้งานโปรแกรมจริงเพิ่มเติม เนื่องจากโปรแกรมดังกล่าวมี Feature จำนวนมาก และมีประสิทธิภาพในการนำไปใช้ทำ Data Analysis นอกจากนี้ ยังแนะนำให้ไปศึกษาหลักการเพิ่มเติมเกี่ยวกับสถิติและโมเดลขั้นสูง เช่น Random Forest, Neural Network โดยเฉพาะผู้ที่ปฏิบัติงานในสาย Data ทั้งนี้ การศึกษาเนื้อหาดังกล่าว อาจใช้เวลาค่อนข้างมาก เนื่องจากเนื้อหามีความซับซ้อนสูง แต่หากมีความเข้าใจแล้ว จะเป็นประโยชน์ต่อการทำงานอย่างมาก



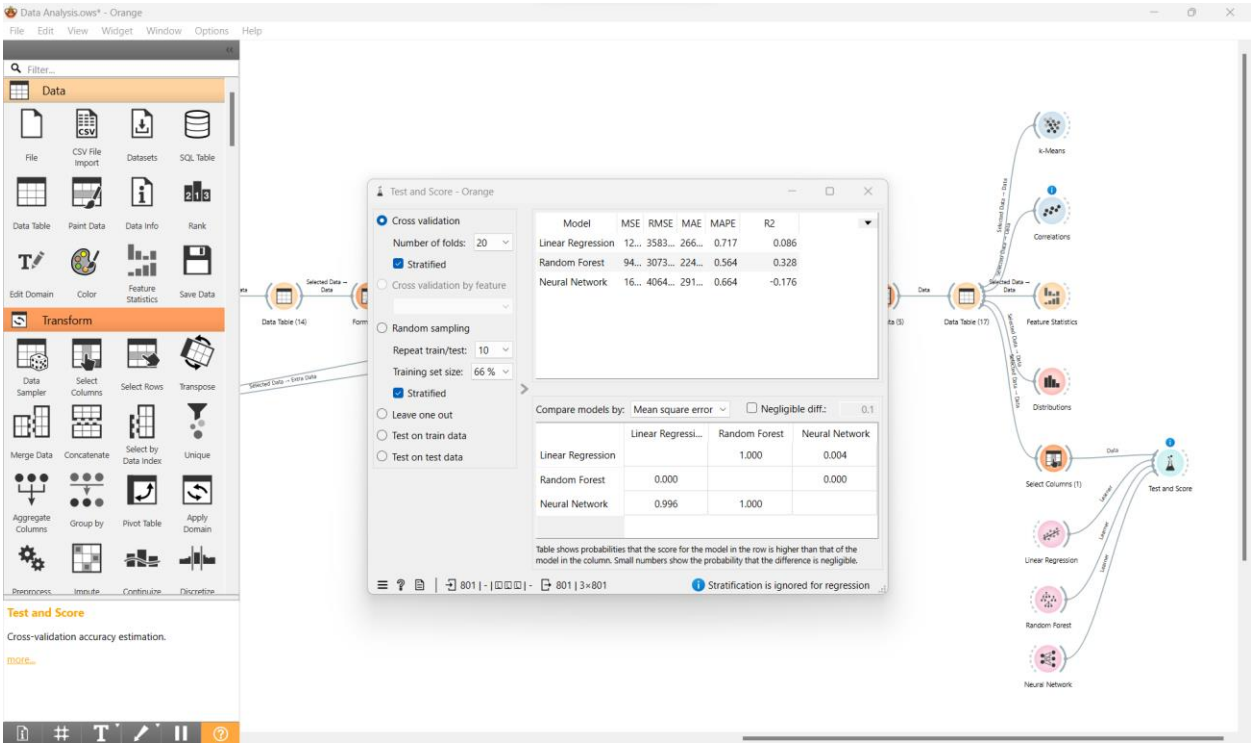
รูปภาพการใช้ฟังก์ชั่น Feature Statistics เพื่อตรวจสอบการกระจายตัวของข้อมูล รวมถึงสถิติพื้นฐาน อาทิ ค่าเฉลี่ย ค่ามัธยฐาน ค่าฐานนิยม Min Max รวมถึงจำนวน Missing Values ในแต่ละชุดข้อมูล



รูปภาพการใช้ฟังก์ชั่น Correlations เพื่อตรวจสอบค่าสัมประสิทธิ์สหสัมพันธ์ระหว่างชุดข้อมูล 2 ชุด



รูปภาพการใช้ฟังก์ชั่น k-Means เพื่อทดสอบการจำแนกข้อมูลออกเป็น Clusters โดย Silhouette Score บ่งบอกถึง แนวโน้มของคุณภาพ Clusters กรณีที่มี 2, 3, 4 จนถึง 9 Clusters โดยในที่นี้ 2 Clusters ดีที่สุด



รูปภาพการใช้ฟังก์ชั่น Test and Score เพื่อประเมินประสิทธิภาพโมเดลทำนาย ในที่นี้ Random Forest ดีที่สุด

Session 11: Data Visualization Principles and Tools

โดย Dr. Chintan Amrit, Associate Professor, Business Analytics Department, University of Amsterdam, Netherlands

การเรียนการสอนในชั้นเรียนนี้ ผู้สอนได้บรรยายถึงความสำคัญของการนำเสนอข้อมูลในรูปแบบที่เข้าใจง่าย หรือการใช้ Data Visualization ในฐานะเครื่องมือนำเสนอผลที่ได้จากการวิเคราะห์ข้อมูล โดยเริ่มต้น ผู้สอนได้หยิบยกกรณีศึกษาในสมัยก่อนในช่วงปี 1900 กว่า ในยุคนั้นเกิดโรคท้องร่วงระบาดของกรุงลอนดอน ในขณะนั้นไม่มีผู้ใดระบุสาเหตุได้ว่า เกิดจากเหตุอะไร จึงมีข้อสันนิษฐานแตกต่างกันออกไป อาทิ โรคดังกล่าวอาจสามารถติดต่อกันได้ผ่านทางอากาศ เป็นต้น จนกระทั่งมีนาย John Snow ได้เกิดข้อสงสัย จึงได้ทำการระบุตำแหน่งข้อมูลของบ้านที่มีผู้ป่วยโรคท้องร่วงลงในแผนที่ของเมือง แล้วพบว่า ผู้ป่วยส่วนใหญ่อาศัยอยู่ในถนนแห่งหนึ่ง รวมถึงอาณาบริเวณใกล้เคียง และได้สืบหาสาเหตุจนกระทั่งพบว่า เกิดจากท่อน้ำเสียในย่านนี้เกิดการรั่วไหลจนกระทั่งน้ำเสียไปปะปนกับน้ำที่ใช้อุปโภคบริโภค จนทำให้เกิดโรคท้องร่วงระบาดขึ้น กรณีศึกษานี้เองเป็นตัวอย่างของการนำ Data Visualization มาใช้ เพื่อนำเสนอข้อมูลอันนำไปสู่ข้อสรุปที่เป็นประโยชน์

ผู้สอนได้ย้อนกลับไปทบทวนถึง CRISP-DM Model ที่ได้ระบุว่า กระบวนการด้าน Data Analytics ควรเริ่มต้นจากความเข้าใจทางธุรกิจ (Business Understanding) ความเข้าใจเกี่ยวกับข้อมูล (Data Understanding) และการเตรียมข้อมูล (Data Preparation) ประมาณร้อยละ 80 อีกร้อยละ 20 เป็นกระบวนการทางสถิติและโมเดล รวมถึงการนำเสนอผลลัพธ์ของข้อมูล ดังนั้น การทำ Data Visualization จึงควรเริ่มต้นจากความเข้าใจปัญหาธุรกิจและข้อมูลที่เรามีอยู่เสียก่อน

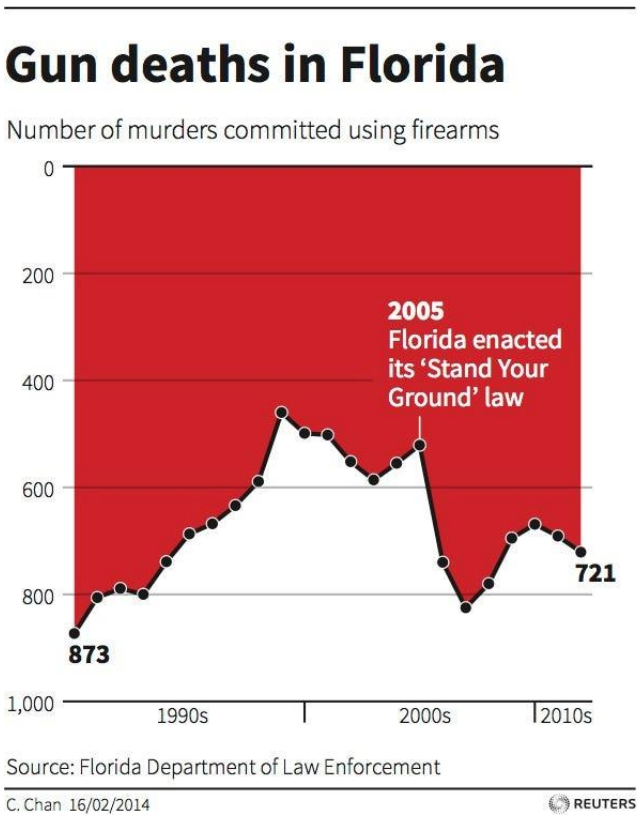
ตัวอย่างของการทำ Data Visualization ที่มีประสิทธิภาพ คือ การนำเสนอข้อมูลอย่างเป็นกลาง และทำให้ผู้อ่านเข้าใจง่าย ไม่เกิดความเข้าใจคลาดเคลื่อน กรณีตัวอย่างที่ไม่ดี อาทิ การนำเสนอข้อมูลความถี่ของการเข้าโบสถ์ของประชาชนที่สำรวจ แต่ไม่ได้แบ่งช่วงความถี่ของข้อมูลให้เท่ากัน เช่น แบ่งเป็นช่วง 1-4 ครั้งต่อปี 5-8 ครั้งต่อไป 9-11 ครั้งต่อไป 12-24 ครั้งต่อปี (สังเกตว่า ระยะช่วงห่างของ 12-24 ครั้ง แตกต่างจาก ระยะช่วงห่างของ 1-4 ครั้ง) หรือการที่นำเสนอข้อมูลโดยให้ค่า 0 อยู่ด้านบนสุดของแกน Y (โดยปกติ ค่า 0 จะอยู่ด้านล่างสุดของแกน Y) ทำให้เกิดความเข้าใจผิดและตีความได้ยาก

หลักการหนึ่งของการนำเสนอ Data Visualization คือ การใช้ Chart ที่ทำให้ผู้อ่านเข้าใจได้ง่ายที่สุด ไม่ใช่ Chart ที่ดูดีหรือดูเท่มีระดับ เช่น การใช้การนำเสนอในรูปแบบตาราง หรือ Bar Chart มักทำให้เกิดความเข้าใจได้ง่าย สำหรับข้อมูลส่วนใหญ่ หลีกเลี่ยงการใช้กราฟแสดงผล 3 มิติ หากไม่มีความจำเป็น เพราะผู้อ่านมักเกิดความสับสนได้ นอกจากนี้ ควรระบุให้ชัดเจนว่า สิ่งที่ต้องการสื่อสารจาก Visualization ไปยังผู้อ่านคืออะไร อย่าให้ผู้อ่านต้องตีความเอง เพราะอาจเกิดความเข้าใจคลาดเคลื่อนได้ เช่น หากมีการนำเสนอกราฟเส้นของยอดขายและกำไรในรูปแบบ Time Series ให้มีการระบุหัวข้อให้ชัดเจนว่า “ยอดขายและกำไรปี XX มีแนวโน้มเติบโตอย่างเห็นได้ชัด” แทนที่จะไม่ระบุหัวข้อ และทำให้ผู้อ่านต้องใช้ดุลยพินิจในการตีความ ซึ่งแต่ละคนอาจเข้าใจได้ไม่เหมือนกัน

หลักการอีกข้อหนึ่งที่ควรนำไปใช้คือ ระมัดระวังอย่าให้ผู้อ่านต้องทำการเปรียบเทียบ “พื้นที่” (Area) เพราะมนุษย์เราไม่สามารถบ่งบอกความแตกต่างของขนาดพื้นที่ได้ดีนัก เว้นแต่จะมีความแตกต่างกันอย่างเห็นได้ชัด หากต้องมีการเปรียบเทียบ ผู้สอนแนะนำให้เปรียบเทียบระหว่างความยาว (Length) เพราะสายตาของมนุษย์สามารถจำแนกความยาวได้

ดีกว่า (แต่มีข้อยกเว้น ในกรณีที่ความแตกต่างของค่ามีน้อยมาก เช่น 1500 กับ 1501 หากใช้ Bar Chart ในการนำเสนออาจทำให้ผู้อ่านไม่เห็นถึงความแตกต่างระหว่าง 2 ค่า ในที่นี้ จึงแนะนำให้ระบุค่ากำกับลงไปใน Chart ร่วมด้วย) หลีกเลี่ยงการใช้ Pie Chart เพราะการเปรียบเทียบขนาดของ Pie ทำให้ได้ยากมาก โดยเฉพาะหากไม่มีการระบุค่าข้อมูลกำกับไว้

การนำเสนอข้อมูลด้วยสี ควรระมัดระวังอย่าให้พื้นหลังมีสี ควรเน้นพื้นหลังสีขาว นอกจากนี้ การนำเสนอผลลัพธ์ต้องคำนึงถึง Common Sense ของมนุษย์ เช่น เส้นที่อยู่ในตำแหน่งที่สูงขึ้น มักหมายถึง การเพิ่มขึ้นของค่าบางอย่าง กรณีศึกษาของการใช้ Visualization อย่างไม่มีประสิทธิภาพคือ กรณีของ Gun Death ใน Florida ที่เส้นในตำแหน่งสูงขึ้นกลับหมายถึงค่า Number of Murder โดย Firearms ที่ลดลง (อ้างอิงรูปภาพด้านล่าง) กรณีดังกล่าวควรหลีกเลี่ยง เพราะจะทำให้เกิดความสับสนของผู้อ่าน

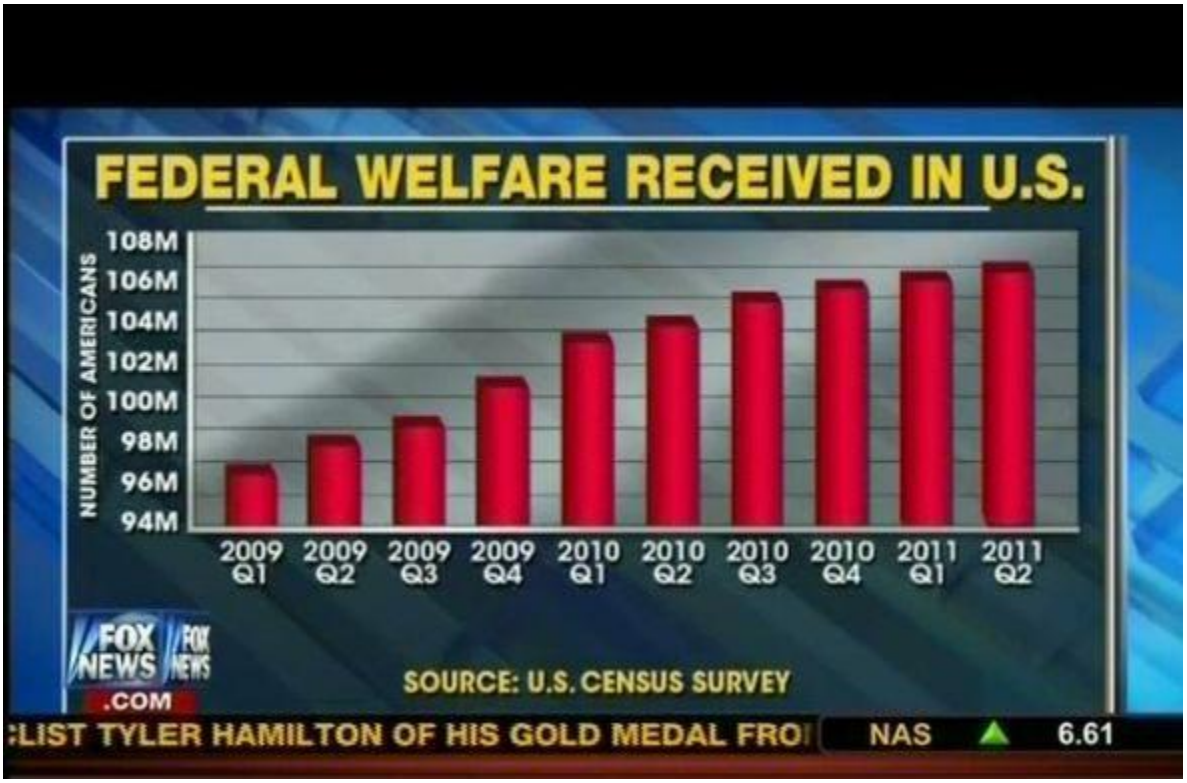


รูปภาพตัวอย่างการนำเสนอข้อมูลอย่างไม่มีประสิทธิภาพ ในที่นี้ เส้นที่เพิ่มขึ้น กลับหมายถึง ค่าที่ลดลง ไม่สอดคล้องกับ Common Sense การรับรู้ของมนุษย์

(ที่มารูปภาพ Gun Deaths in Florida Increased With 'Stand Your Ground' - Business Insider)

นอกจากนี้ การนำเสนอข้อมูลโดย Visualization ควรดำเนินการอย่างตรงไปตรงมา และไม่มีเจตนาสร้างความเข้าใจผิด หรือบิดเบือนการรับรู้ของผู้อ่าน ซึ่งมักเกิดขึ้นอยู่บ่อยครั้งในการนำเสนอข่าวบางรายการ ข้อมูลแบบ Chart ซึ่งเส้นแกน Y ไม่ได้เริ่มต้นจาก 0 ทำให้เกิดความเข้าใจผิดในระดับ Scale (อ้างอิงตัวอย่างรูปภาพด้านล่าง) หรือการแบ่งระยะของ Scale ไม่เท่ากัน เช่น จาก 0 ขึ้นเป็น 30 แต่จาก 30 ขึ้นเป็น 80 ทำให้เกิดการรับรู้ของผู้อ่านที่ไม่ถูกต้อง

นอกจากนี้ ผู้สอนยังได้แนะนำให้ผู้เรียนรู้จักกับ Data Visualization Technique หลากหลาย อาทิ Simple Tables, Summary Tables, Contingency Tables, Super Tables, Scatterplot Matrix, Parallel Coordinates, Star Glyphs, Face Glyphs, Dimensional Stacking, Chord Diagram รวมถึงแนะนำให้ศึกษาความรู้เพิ่มเติมเกี่ยวกับเทคนิคการใช้ Chart แต่ละประเภท เช่น Bar, Column, Line, Area, Bubble Chart, Spider and Rader Chart, Scatter, Stacked Bar Chart โดยแนะนำหนังสือ The Visual Display of Quantitative Information เขียนโดย Edward R. Tufte (Data Visualization Books by Edward Tufte | Order Online) และ Information Dashboard Design: Displaying Data for At-a-glance Monitoring เขียนโดย Stephen Few (Perceptual Edge - Library)



รูปภาพตัวอย่างการนำเสนอข้อมูลอย่างไม่มีประสิทธิภาพ ในที่นี้ เส้นแกน Y ไม่ได้เริ่มต้นจาก 0 ทำให้ลักษณะการเพิ่มขึ้นของข้อมูลแต่ละช่วงดูกว้างมากกว่าที่ควรเป็น

(ที่มารูปภาพ The Tricky Ways Fox News Uses Data - Business Insider)

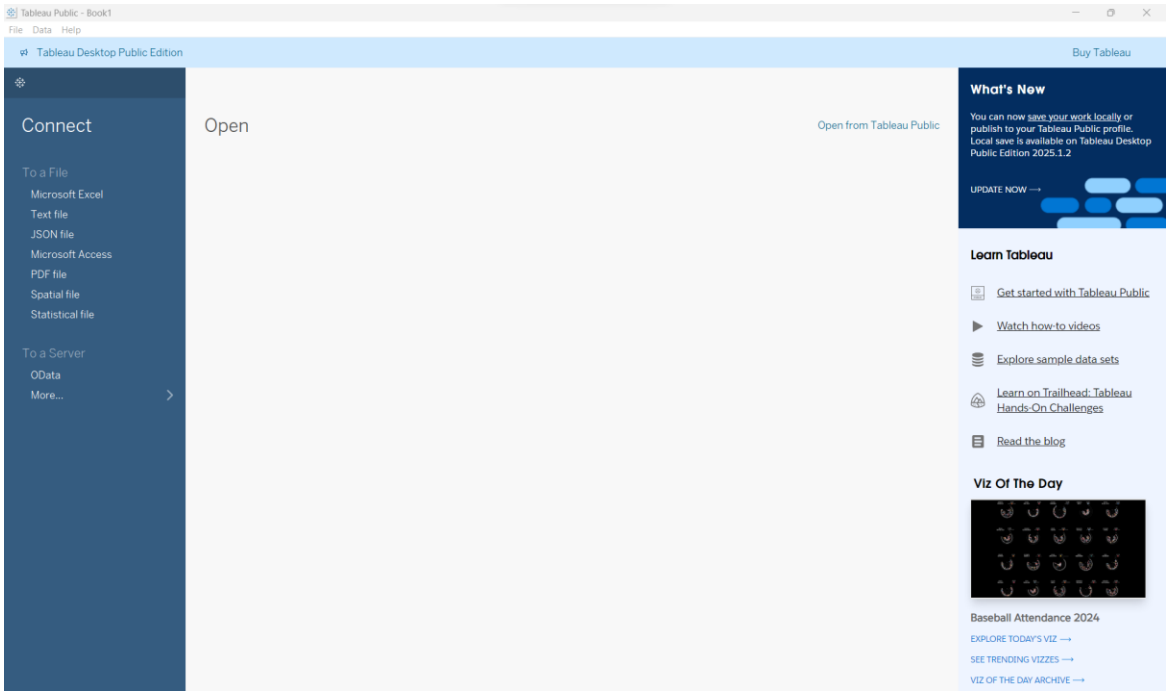
Session 12: Data Visualization Exercises

โดย Dr. Chintan Amrit, Associate Professor, Business Analytics Department, University of Amsterdam, Netherlands

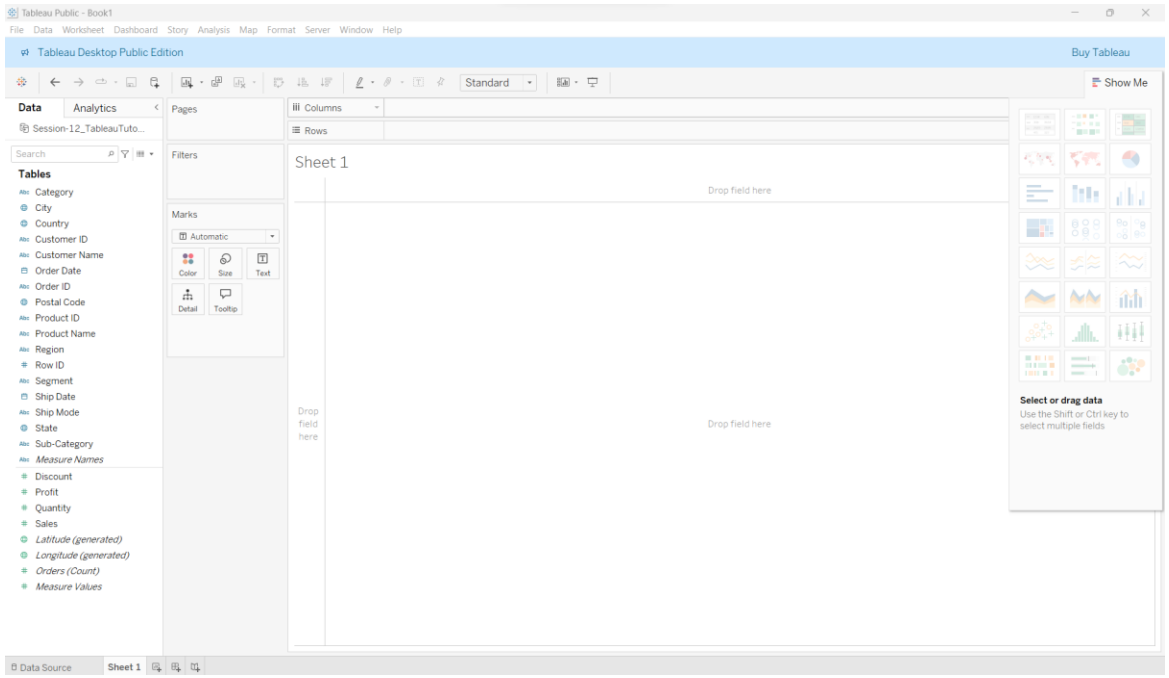
ในการเรียนการสอนคลาสนี้ ผู้บรรยายได้แนะนำให้ผู้เข้ารับการอบรมรู้จักกับโปรแกรม Tableau Public สำหรับใช้ในการทำ Data Visualization ซึ่งถือเป็นโปรแกรมที่ใช้งานได้ฟรี โดยไม่มีค่าใช้จ่าย แต่ต้องแลกมาด้วยการเปิดเผยข้อมูลในที่สาธารณะ (ในกรณีที่ต้องการ Privacy ของข้อมูล ต้องใช้เวอร์ชั่นแบบเสียค่าใช้จ่าย) โดยเริ่มต้นตั้งแต่กระบวนการดาวน์โหลดและติดตั้ง Software โดยผู้ปฏิบัติงานสามารถเข้าไปที่ Website https://public.tableau.com/ เพื่อดาวน์โหลดโปรแกรม จากนั้นกดปุ่ม Install และลงทะเบียนใช้งานโปรแกรม เพื่อเข้าสู่การใช้งาน จากนั้นผู้ใช้งานจะต้องทำการเปิด File เพื่อเชื่อมต่อ (Connect) กับข้อมูล (ในที่นี้ ผู้สอนได้จัดทำ File ตัวอย่าง สำหรับใช้ในการสาธิตเพื่อประกอบการบรรยาย: ดูตัวอย่างรูปภาพหน้าถัดไป) เมื่อดำเนินการแล้วเสร็จ ผู้ใช้งานจะเห็นหน้าจอสำหรับทำ Data Visualization โดยมีแถบด้านซ้ายคือ Column ชุดข้อมูลที่สามารถเลือกได้ เพื่อจัดทำ Visualization

ผู้บรรยายได้สาธิตวิธีการสร้าง Data Visualization ด้วยโปรแกรม ซึ่งสามารถทำได้ง่าย เนื่องจากโปรแกรมมีลักษณะการทำงานแบบ Drag & Drop คือ ลากข้อมูลที่ต้องการจากแถบด้านซ้ายมือของหน้าจอ ลงในกล่อง Column หรือกล่อง Row ที่แสดงอยู่บริเวณตรงกลางของหน้าจอ ในที่นี้ ผู้สอนได้ให้ผู้เรียนลากข้อมูล YEAR (Order Date) ลงใน Column และข้อมูล Sales ลงใน Row จะปรากฏ Line Chart ที่แสดงยอดขายของแต่ละปี (อ้างอิงรูปภาพด้านล่าง) และหากต้องการเจาะลึกมิติของข้อมูลลงไป ผู้ใช้งานสามารถทำได้ง่าย โดยการลากข้อมูลเพิ่มเติมที่ต้องการเจาะลึกลงใน Column หรือ Row โดยบทเรียนที่นี้ ผู้สอนได้ให้ผู้เรียนลาก Category และ Sub-Category ลงในช่อง Column เพิ่มเติม จะปรากฏ Bar Chart ที่แสดงยอดขายของหมวดสินค้าย่อยทุกหมวดในแต่ละปี (อ้างอิงรูปภาพด้านล่าง) หากไม่ต้องการข้อมูลส่วนใดแล้ว สามารถแก้ไขได้ง่าย โดยการ Drag ข้อมูลออกจากกล่อง Column หรือ Row

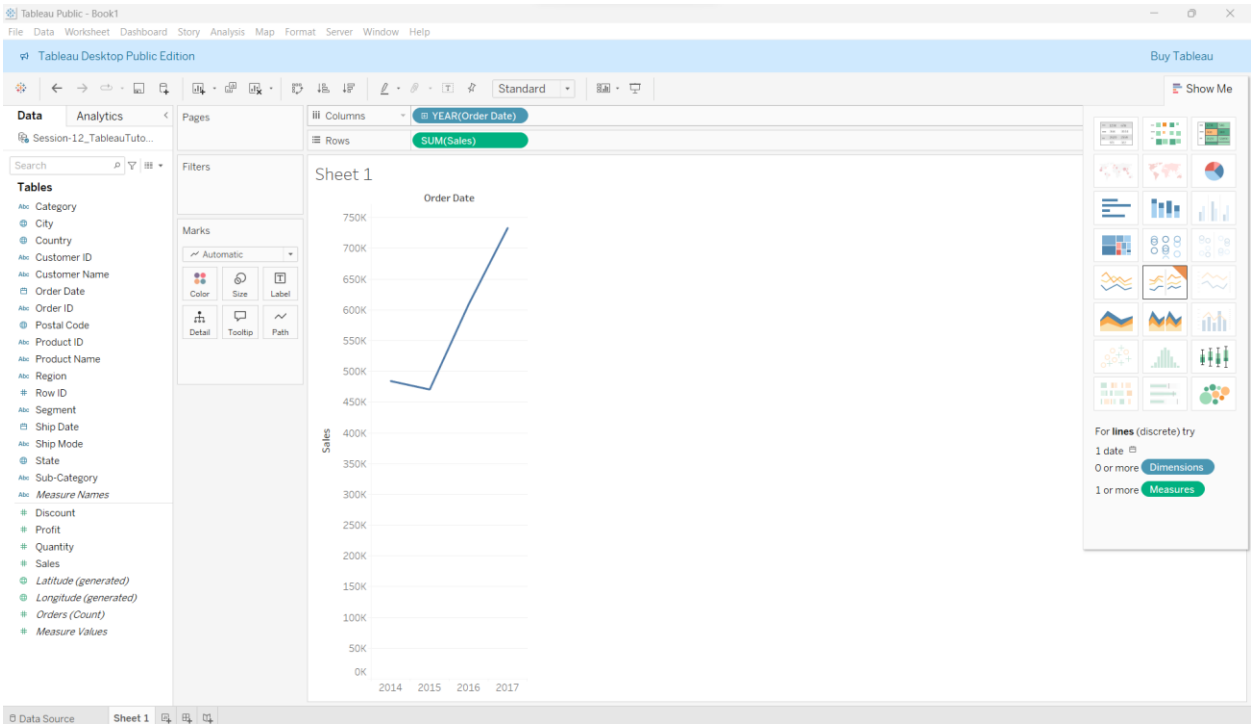
นอกจากนี้ ผู้บรรยายยังได้สาธิตวิธีการใส่ Filter ลงใน Visualization โดยการคลิกขวาที่ข้อมูลที่ต้องการสร้าง Filter (กรณีตัวอย่างที่สาธิต คือ YEAR และ Sub-Category) จากนั้น กดคำว่า Show Filter จะปรากฏ Filter ข้อมูลบริเวณด้านขวาของโปรแกรม ซึ่งผู้ใช้งานสามารถคลิกเพื่อแสดงข้อมูลบางรายการเท่านั้น นอกจากนี้ ผู้ใช้งานยังสามารถเพิ่มสีลงใน Visualization เพื่อให้ข้อมูลเพิ่มเติมได้ด้วย โดยในที่นี้ ผู้สอนต้องการแสดง Profit ของรายการสินค้าย่อยแต่ละประเภท จึงทำการลากข้อมูล Profit ลงใน Color Marks บริเวณแถบด้านซ้าย ติดกับแถบที่แสดงชุดข้อมูลทั้งหมด จากนั้น Chart จะเปลี่ยนสีโดยอัตโนมัติ (ในที่นี้ กำไรน้อย จะแสดงด้วยสีแดง กำไรมาก จะแสดงด้วยสีน้ำเงิน) เมื่อรวมกันทั้งหมดแล้ว จะทำให้ผู้ใช้งานทราบได้ว่า ยอดขายสินค้าในแต่ละประเภทย่อยในแต่ละปี มียอดมากน้อยเท่าไหร่ และภายในแต่ละประเภทนั้นมีกำไรมากหรือน้อย หรือขาดทุน โดยผู้สอนได้ให้ผู้เรียนเพิ่มข้อมูล Regions เข้าไปใน Row เพิ่มเติม เพื่อจำแนกยอดขายในแต่ละพื้นที่ภูมิภาค และให้เลือกแสดง Filter เฉพาะ Bookcases, Machines, Tables เมื่อลาก Profit เข้าสู่ Color Marks ทำให้ทราบผลกำไร/ขาดทุนของประเภทสินค้าย่อยแต่ละรายการ ดังที่แสดงในรูปภาพหน้าถัดไป



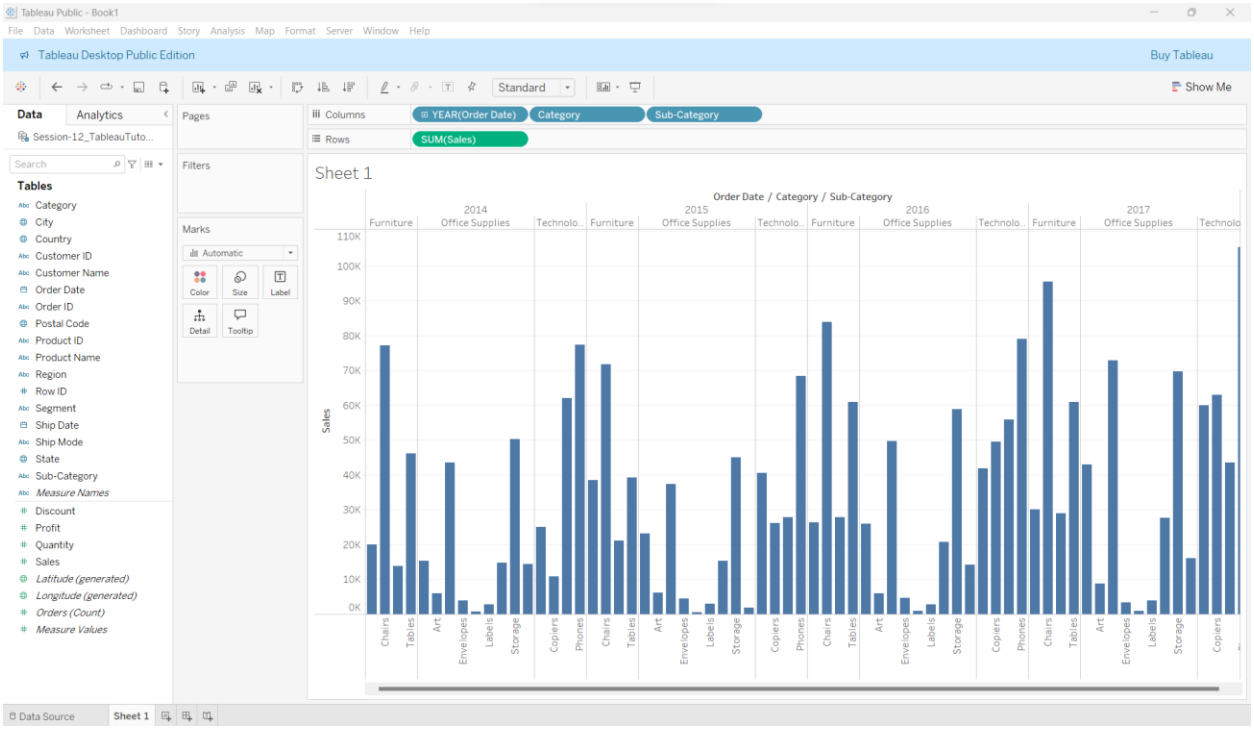
รูปภาพหน้าจอโปรแกรม Tableau Public หลังจากเปิดโปรแกรมใหม่



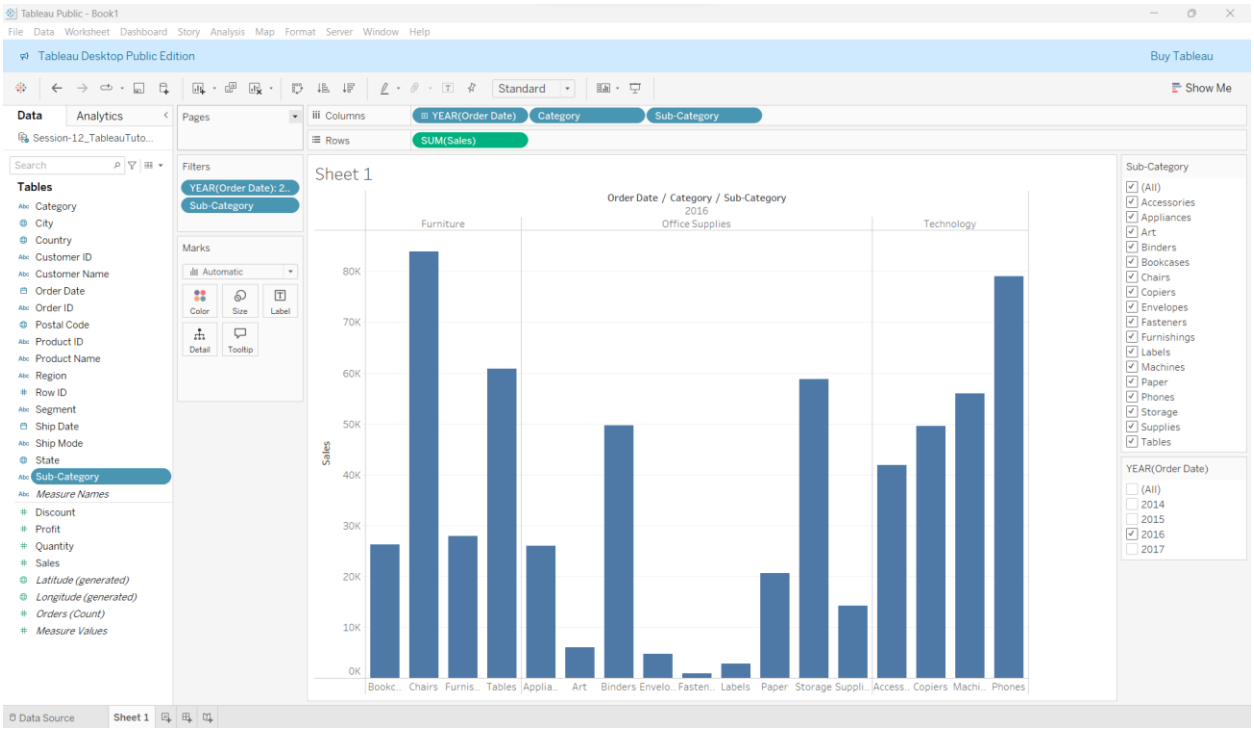
รูปภาพหน้าจอโปรแกรม Tableau Public หลังจากที่มีการเชื่อมต่อกับ File ตัวอย่างข้อมูลที่ผู้บรรยายใช้สาธิตในการใช้งานโปรแกรม



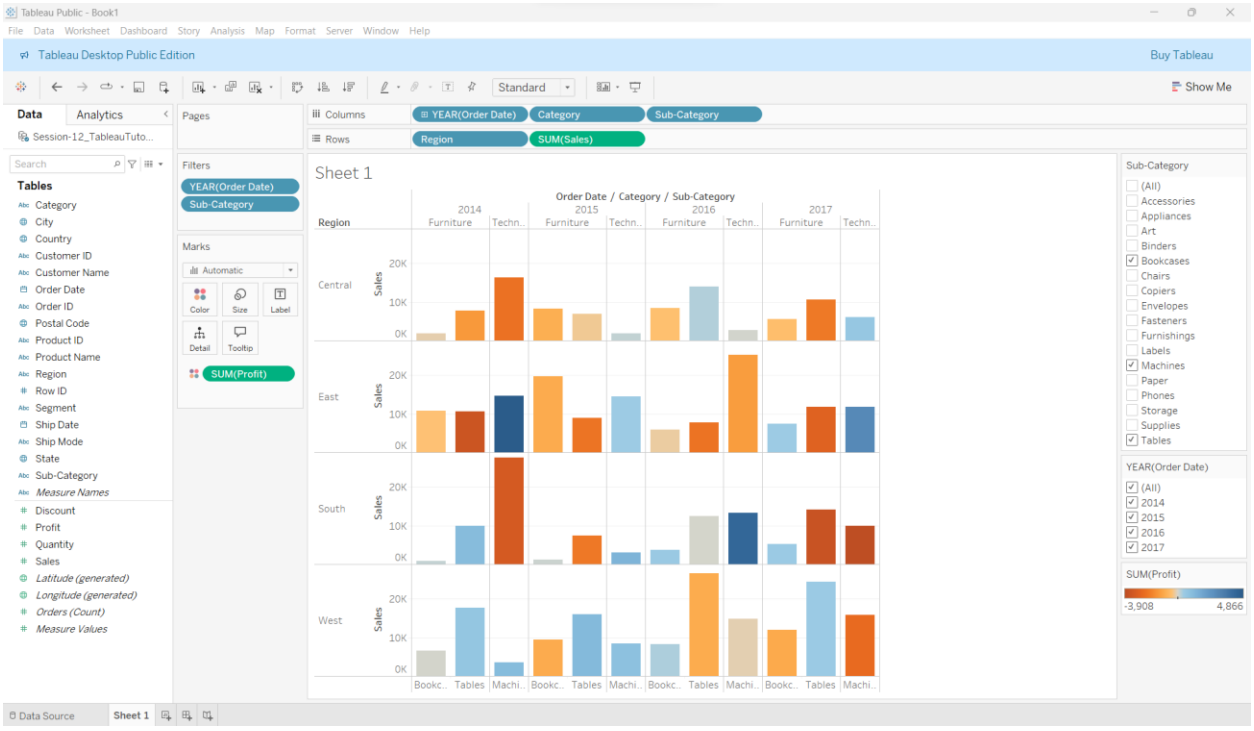
รูปภาพตัวอย่างการทำ Visualization อย่างง่าย โดยเลือก Columns = ปี และ Rows = ยอดขาย



รูปภาพตัวอย่างการทำ Visualization โดยเลือก Columns = ปี, ประเภทของสินค้า, ประเภทย่อยของสินค้า และ Rows = ยอดขาย โปรแกรมจะทำการสร้าง Chart อัตโนมัติ โดยแบ่งยอดขายแต่ละปีตามประเภทย่อยสินค้า



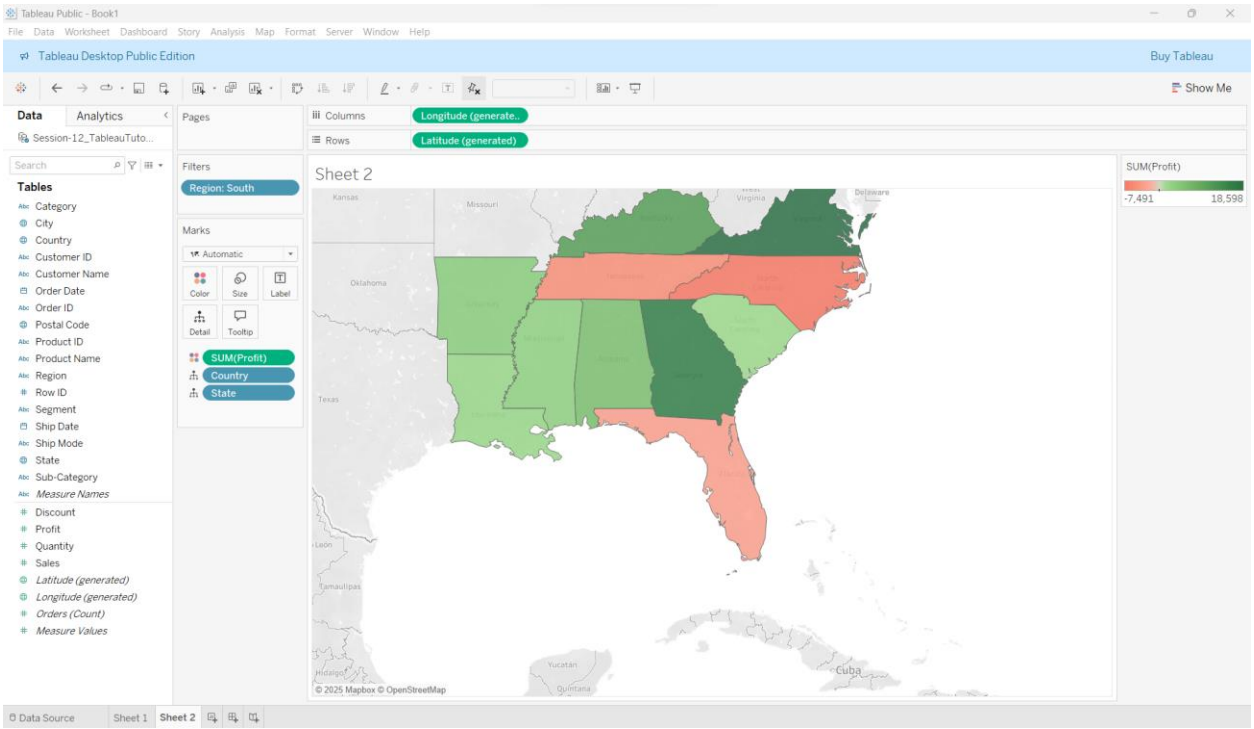
รูปภาพการใส่ Filter ข้อมูล (YEAR, Sub-Category) และเลือกเฉพาะปี 2016 จะแสดงเฉพาะชุดข้อมูลบางส่วน



รูปภาพการเพิ่มข้อมูล Regions เข้าไปในแถบ Rows และเลือก Filter เฉพาะ Bookcases, Machines, Tables และเพิ่มข้อมูล Profit ลงใน Color Marks เพื่อระบุข้อมูลผลกำไร ซึ่งแทนด้วยสี (สีแดง หมายถึง กำไรต่ำ สีน้ำเงิน หมายถึง กำไรสูง)

นอกเหนือจากการแสดง Visualization ด้วย Line Chart และ Bar Chart แล้ว โปรแกรม Tableau ยังมีความสามารถในการแสดงผลลัพธ์ในรูปแบบ Geographical Map โดยผู้สอนได้สาธิตวิธีการแสดงผลข้อมูลวิเคราะห์ผลกำไรของร้านค้า จำแนกตาม Geographical Territory โดยเริ่มต้นจากการสร้าง Worksheet ใหม่ โดยเลือกคลิกแถบบริเวณด้านล่างโปรแกรม แล้วกด New Worksheet โปรแกรมจะสร้างหน้าต่างว่างเปล่า เพื่อให้ทำ Visualization ได้ เช่นเดียวกับตอนแรกสุด จากนั้น กดเลือก Double Click ชุดข้อมูล State จะปรากฏข้อมูล Longitude และ Latitude โดยอัตโนมัติ ในช่อง Column และ Row จากนั้นให้เลือกใส่ Filter Region (กดคลิกขวาที่แถบ Region แล้วกด Show Filter) จากนั้นเลือกเฉพาะ South จากนั้นให้ลากชุดข้อมูล Profit ไปใส่ที่ Color Marks จะปรากฏ Visualization แบบสีบริเวณ South Area แต่สีที่ได้จะเป็นสีแดง-น้ำเงิน หากต้องการปรับให้แสดงผลดังรูปภาพตัวอย่างด้านล่าง ให้กดคลิกที่ Color Marks แล้วกดปุ่ม Edit Color จากนั้นเลือก Red-Green Diverging จะได้สีดังที่แสดงผลลัพธ์ด้านล่างนี้

โดยภาพรวมการเรียนการสอนในคลาสนี้ ผู้บรรยายได้เน้นการปฏิบัติจริงของผู้เรียนทุกคน โดยแนะนำตั้งแต่พื้นฐานการดาวน์โหลดและติดตั้งโปรแกรม และอธิบายวิธีการใช้งานในแต่ละขั้นตอน ดังที่ข้าพเจ้าได้อธิบายทั้งหมดข้างต้น ผู้สอนได้แนะนำให้ผู้เรียนไปศึกษาการใช้งานโปรแกรมนี้ต่อด้วยตนเอง เนื่องจากระยะเวลาการสอนมีจำกัด โดยเสนอให้ผู้เรียนทุกคนเข้าไปศึกษาต่อที่ https://help.tableau.com/current/guides/get-started-tutorial/en-us/get-started-tutorial-connect.htm/ เพื่อเพิ่มระดับความเข้าใจโปรแกรมต่อไป



รูปภาพการแสดง Visualization ในรูปแบบแผนที่ โดยกด Double Click ที่ชุดข้อมูล State และเลือก Filter Region เฉพาะ South และเพิ่มข้อมูล Profit ลงใน Color Marks เพื่อแสดงกำไรแทนด้วยสี จากนั้นปรับ Edit Color ใน Color Marks เพื่อให้สีแดง หมายถึง กำไรติดลบ สีเขียว หมายถึง กำไรสูง

Site Visit 1: AI Farm Factory

โดย National Productivity Center of Cambodia (NPCC), Cambodia

การศึกษาดูงานในประเทศกัมพูชาที่แรกของวันนี้ คือ AI Farm Factory ซึ่งถือเป็นโรงงาน Robotics แห่งแรกของประเทศกัมพูชา ก่อตั้งขึ้นในปี 2017 โดยนาย Kim Khorn Long ผู้ก่อตั้งบริษัท (Founder) โดยมีวิสัยทัศน์ต้องการส่งเสริมการยกระดับประเทศกัมพูชาสู่การขับเคลื่อนด้วย Robot เนื่องจากตระหนักถึงความสำคัญของพลังจากเทคโนโลยี Robotics และระบบ Automation ในการสร้างโอกาสทางเศรษฐกิจ สถานที่ตั้งของบริษัทอยู่ที่ชานเมืองของกรุงพนมเปญ เมืองหลวงของประเทศกัมพูชา (ศึกษารายละเอียดเพิ่มเติมได้ที่ https://www.facebook.com/aifarm.dev/)

เริ่มแรกเมื่อผู้อบรมมาถึงสถานที่ เจ้าหน้าที่ได้แนะนำเกี่ยวกับประวัติและวิสัยทัศน์ของบริษัท พร้อมทั้งพาเดินเยี่ยมชมบรรยากาศภายในบริษัท ซึ่งแบ่งออกเป็นห้องต่างๆ อาทิ ห้องที่พัฒนาระบบโดรนสำหรับการเรียนรู้ของเด็กและเยาวชน และโดรนสำหรับภาคอุตสาหกรรม ห้องที่พัฒนาหุ่นยนต์และระบบขับเคลื่อนหุ่นยนต์สำหรับการเรียนรู้ของเด็กและเยาวชน ห้องพัฒนาหุ่นยนต์สำหรับใช้ในภาคอุตสาหกรรม (หุ่นยนต์ขนาดใหญ่) ห้องจัดแสดงตัวอย่างของการทำงานหุ่นยนต์ที่ใช้ขนสินค้า ห้องจัดประกวดการแข่งขันหุ่นยนต์ต่อสู้ เป็นต้น จากนั้น เจ้าหน้าที่ได้นำเสนอตัวอย่างผลงานที่กำลังอยู่ในระหว่างการ Develop จำนวน 3 โครงการ ประกอบด้วย (1) ระบบรักษาความปลอดภัย สำหรับตรวจจับใบหน้าของผู้บุกรุกในยามวิกาล (2) ระบบสแกนใบหน้าเพื่อบันทึกเวลาเข้าทำงาน และ (3) ระบบตรวจจับและอ่านป้ายทะเบียนรถแบบอัตโนมัติที่เป็นภาษาเขมร หลังจากที่นำเสนอทั้งหมด เจ้าหน้าที่ได้เปิดโอกาสให้ผู้อบรมได้สอบถามในข้อสงสัย จากนั้น จึงเดินทางต่อไปยัง Site Visit 2

รูปภาพบรรยากาศด้านหน้าและบริเวณภายในบริษัท AI Farm Factory ตั้งอยู่ในเขตชานเมือง กรุงพนมเปญ

Site Visit 2: Cambodia Academy Digital Technology

โดย National Productivity Center of Cambodia (NPCC), Cambodia

หลังจากที่ได้ศึกษาดูงานที่ AI Farm Factory แล้ว คณะได้เดินทางต่อไปยัง Cambodia Academy Digital Technology (CADT) โดยสถานที่แห่งนี้เป็นสถาบันการศึกษาระดับอุดมศึกษาของประเทศกัมพูชา มุ่งเน้นการสอนและงานวิจัยเฉพาะด้านเกี่ยวกับเทคโนโลยีดิจิทัล โดยรูปแบบมีความใกล้เคียงกับสถาบันเทคโนโลยีระดับอุดมศึกษาในประเทศไทย CADT ก่อตั้งขึ้นในปี 2014 ภายใต้การกำกับดูแลของ Ministry of Post and Telecommunications มีวิสัยทัศน์มุ่งสู่การศูนย์กลางของประเทศด้านการวิจัยและการศึกษาทางด้านเทคโนโลยีดิจิทัล ปัจจุบันเปิดการศึกษาในระดับปริญญาตรีและปริญญาโทในด้านที่เกี่ยวข้องกับ Computer Science, Data Science และ AI โดยมีสถาบันการศึกษาย่อย 3 แห่ง ที่อยู่ในการกำกับดูแล (เทียบได้กับคณะ) ได้แก่ Institute of Digital Technology, Institute of Digital Governance และ Institute of Digital Research & Innovation (ศึกษารายละเอียดเพิ่มเติมได้ที่ https://cadt.edu.kh/)

เจ้าหน้าที่ได้บรรยายเกี่ยวกับประวัติและความเป็นมาของสถาบัน CADT จากนั้นได้พาผู้อบรมเข้าเยี่ยมชม บริเวณโดยรอบสถาบัน ประกอบด้วย ห้องอบรม (Lecture Room) ห้องจัดแสดงผลงานของนักศึกษา รวมถึงห้อง Workshop ส่วนกลางที่เปิดโอกาสให้นักศึกษาได้เข้ามาใช้งาน เพื่อสร้างหุ่นยนต์หรืองานประดิษฐ์ โดยไม่มีค่าใช้จ่าย จากนั้นได้แนะนำงานวิจัยของสถาบันที่ได้รับการยอมรับ อาทิ การพัฒนาเทคโนโลยีการพิมพ์อักษรเบลล์สำหรับคนตาบอด โดยแปลงจากภาษาเขมร โดยใช้ Computer Vision การพัฒนาแผนระดับประเทศ Digital Skills Development Roadmap ประจำปี 2024-2035 ของประเทศกัมพูชา รวมถึงแนะนำสิ่งที่สถาบันดำเนินการไปแล้ว เช่น การจัดฝึกอบรม Digital Transformation สำหรับผู้นำภาครัฐ จากนั้น เจ้าหน้าที่สถาบันได้เปิดโอกาสให้ผู้อบรมสอบถาม จากนั้น จึงเดินทางกลับสู่โรงแรมที่พัก

รูปภาพห้องบรรยายของ Cambodia Academy Digital Technology และตัวอย่างผลงานนักศึกษา

Session 13: Implementing AI Solutions & Session 15: Future Trends in AI and Data Analytics

โดย Dr. Chintan Amrit, Associate Professor, Business Analytics Department, University of Amsterdam, Netherlands และ Dr. Murphy Choi, CEO, Alionova Education, Singapore

ในการเรียนการสอนคลาสนี้ ได้มีการควบรวมเนื้อหากันของ 2 Sessions โดยผู้บรรยายทั้งสองท่านได้ร่วมกันบรรยายเนื้อหา โดยแนะนำเกี่ยวกับ Advanced Concept และ Trends ที่เกี่ยวข้องกับ AI ในอนาคต โดยเริ่มต้นจากการแนะนำ Deep Learning หรือหมายถึง สาขาหนึ่งของรูปแบบการเรียนรู้ Machine Learning แต่มีความสลับซับซ้อนมากกว่า โดย Machine จะพยายามศึกษาข้อมูลด้วยตนเอง โดยที่ไม่จำเป็นต้องได้รับการกำกับหรือสอนโดยมนุษย์ (แต่ยังสามารถกำกับหรือสอนโดยมนุษย์ได้ด้วยเช่นเดียวกัน) โดยใช้หลักการ Hierarchy of Multiple Layers โดยประโยชน์ของ Deep Learning คือ การเรียนรู้ได้ด้วยตนเองของ Machine ที่มีความรวดเร็วและมีประสิทธิภาพ แต่ต้องอาศัยการมี Training Data จำนวนมาก

นอกจากนี้ ผู้สอนยังได้แนะนำเกี่ยวกับโครงสร้างพื้นฐานของอัลกอริทึม Neural Network โดยประกอบด้วย Input Layer (ข้อมูลนำเข้า) Hidden Layer (ขั้นตอนประมวลผลข้อมูล ก่อนส่งไปยัง Output Layer) และ Output Layer (ผลลัพธ์ของโมเดล) ในที่นี้ มีสูตรในการคำนวณจุดที่ต้องเรียนรู้ Learnable Parameter คือ น้ำหนัก (เกิดจาก Input Layer x Hidden Layer + Hidden Layer x Output Layer) + Bias (เกิดจากจำนวน Hidden Layer + Output Later) สมมติว่า มี Input Layer = 3, Hidden Layer = 4, Output Layer = 2 จะได้ว่า Learning Parameter = [(3 x 4) + (4 x 2)] + (4 +2) = 26 หมายถึง จำนวน Parameter ที่ต้องเรียนรู้ใน Neural Network มีทั้งหมด 26 จุดที่ต้องเรียนรู้และปรับค่า เพื่อให้สามารถทำนายข้อมูลได้ดีที่สุด ในกรณีที่มี Layers มากขึ้น จำนวนพารามิเตอร์มากขึ้น ทำให้โมเดลเรียนรู้ได้ซับซ้อนขึ้น

หลักการในการฝึก Neural Network จะมีอยู่ 4 ขั้นตอนหลัก โดยมีเป้าหมายคือ การทำให้โมเดลสามารถทำนายได้แม่นยำขึ้น ผ่านการปรับน้ำหนักให้เหมาะสม เริ่มแรกคือ ขั้นตอนของ Sample Labeled Data คือ การให้ข้อมูลกับโมเดลว่า A คืออะไร B คืออะไร เช่น ให้ข้อมูลชุดรูปภาพแมวจำนวนหนึ่ง โดยระบุว่า นี้คือรูปภาพแมว จากนั้นจะเข้าสู่ขั้นตอนของการ Forward through the Network คือ การส่งข้อมูล Input เข้าสู่ Neural Network เพื่อให้โมเดลทำนายผลลัพธ์ เช่น ระบุว่า ภาพนี้มีโอกาสที่เป็นแมว 70% หลังจากนั้น จึงเข้าสู่ขั้นตอนของการ Backpropagate the Errors หรือการประเมินความผิดพลาดระหว่างค่าที่โมเดลทำนายและค่าจริง จากนั้นส่งย้อนกลับไปยัง Neural Network เพื่อระบุว่า น้ำหนักของโมเดลควรมีการปรับจุดไหน อย่างไร เพื่อให้โมเดลเรียนรู้ได้ดีขึ้นในรอบถัดไป

ผู้สอนยังได้นำเสนอหลักการทำงานของ Large Language Models (LLMs) เช่น ChatGPT ว่าสามารถทำงานได้อย่างไร โดยระบุว่า LLM มีลักษณะของระบบที่ใช้หลักการทางสถิติ ควบคู่กับ Machine Learning เพื่อคาดเดาคำถัดไป จากสิ่งที่ใช้ผู้ใช้งานป้อนข้อมูลเข้ามา โดยเริ่มแรก LLMs จะเรียนรู้ข้อมูลจำนวนมากที่เกี่ยวกับภาษา เช่น หนังสือ Website และจะเรียนรู้ว่า คำนี้มักจะตามมาด้วยคำไหน จากนั้น พอเวลาที่ผู้ใช้งานพิมพ์คำเข้าไป ระบบจะทำนายว่าคำใดควรจะมาเป็นคำต่อไปมากที่สุด เช่น ป้อนคำว่า “ประเทศไทย” โมเดลจะคำนวณว่า คำไหนที่ควรตามหลังคำดังกล่าว เช่น “สวยงาม”, “ตั้งอยู่ในทวีปเอเชีย” เป็นต้น อย่างไรก็ตาม LLMs ไม่ได้มีความเข้าใจใน Output ที่ตัวเองสื่อออกมาแต่อย่างใด เพียงแต่การนำเสนอ Output เกิดจากการคำนวณทางสถิติว่า คำไหนควรตามหลังด้วยคำไหน โดยใช้หลักของความคุ้นชินจากรูป

ประโยคที่ LLMs ได้เรียนรู้จากข้อมูลขนาดใหญ่มาก โดยปัจจุบัน Model Size ของ LLMs ได้เติบโตขึ้นอย่างมาก จากเดิมในปี 2018 ที่มีเพียงไม่ถึง 1 ร้อยล้าน Parameter ปัจจุบันคาดว่าจะมีเกินกว่า 1 ล้านล้าน Parameter แล้ว

ท้ายสุด ผู้สอนได้แนะนำเกี่ยวกับการเข้ามาของ Deep Seek ที่เป็น AI จากประเทศจีนที่มีต้นทุนถูกกว่า ChatGPT อย่างมาก เนื่องจาก Development Cost ที่ต่ำกว่า แต่แลกมาด้วย Computational Efficiency ที่ลดลง (แต่ใช้งานได้ฟรี ไม่มีค่าใช้จ่าย) อย่างไรก็ตาม สิ่งที่ควรระวัง คือ การรั่วไหลของข้อมูลที่นำเข้าไปใน AI เพราะข้อมูลเหล่านี้สามารถมีสิทธิถูกนำไปเป็นข้อมูล Train สำหรับ AI โดยเฉพาะอย่างยิ่ง ข้อมูลที่มีความ Sensitive สูง ในหน่วยงานของรัฐบาลหรือหน่วยความมั่นคง ไม่ควรนำเอาข้อมูลนำเข้าไปใน AI ทางออกหนึ่งใน คือ การใช้งาน AI ในรูปแบบ Local Toolkit หรือการติดตั้ง AI ลงในคอมพิวเตอร์ของตัวเอง ไม่ได้ไปใช้ Server บน Cloud เช่นเดียวกับ AI ทั่วไป โดยผู้สอนแนะนำโปรแกรม LM Studio (สามารถดาวน์โหลดได้ที่ https://lmstudio.ai/) ซึ่งทำให้ป้องกันเรื่องของ Privacy และ Confidentiality ได้ดีขึ้น แต่ต้องแลกมาด้วยศักยภาพในการทำงานที่ลดลง เนื่องจาก LLMs ที่ถูกติดตั้งลงบนเครื่อง มาพร้อมกับชุดข้อมูลการ Train ที่มีจำกัด และไม่ Real Time

Session 14: Implementing AI Solutions

โดย Quinn Pham, Director of Solution Consulting, Meiro Pte. Ltd., Singapore

สำหรับการเรียนการสอนในคลาสนี้ ผู้บรรยายได้ทบทวน Concept ของ CRISP-DM Model (The Cross-Industry Standard Process for Data Mining) โดยอธิบายการประยุกต์ใช้โมเดลดังกล่าวในยุค AI เริ่มต้นจากการทำความเข้าใจเกี่ยวกับธุรกิจ (Business Understanding) การทำความเข้าใจเกี่ยวกับข้อมูล (Data Understanding) การเตรียมข้อมูล การพัฒนาโมเดล การประเมินโมเดล การใช้โมเดลและการติดตามผล ไปจนถึงการให้ Feedback และ Retraining โมเดล ทั้งหมดนี้ มีวัตถุประสงค์เพื่อให้การทำ Data Analytics ในยุค AI สามารถแก้ไขปัญหาของธุรกิจ องค์กร ชุมชน และสังคมได้อย่างมีประสิทธิภาพ

ผู้สอนได้หยิบยกกรณีศึกษาของประเทศสิงคโปร์ที่มีการพัฒนา Passport-less Immigration Clearance หรือการตรวจคนเข้าเมืองโดยไม่ต้องใช้ Passport ที่สนามบิน Changi โดยเริ่มต้นจากการเข้าใจเป้าหมายองค์กรและการกำหนดกรอบปัญหาที่ต้องการแก้ไข ในที่นี้ รัฐบาลสิงคโปร์ต้องการให้กระบวนการตรวจคนเข้าเมืองมีประสิทธิภาพสูงขึ้น นั่นคือ ทำได้รวดเร็วขึ้น ปลอดภัยขึ้น ทดแทนการใช้วิธีตรวจสอบ Passport แบบดั้งเดิมที่ใช้การจัดเก็บ Biometric ทั้งนี้ มีการกำหนด Use Case เริ่มต้นจากการใช้งานที่ Terminal 4 และเริ่มต้นจากประชากรที่เป็นคนสิงคโปร์และบุคคลที่เข้าออกประเทศสิงคโปร์บ่อย โดยมีการลงทะเบียน Biometric Data และจำกัดไม่ใช้งานกับนักเดินทางที่เข้าประเทศสิงคโปร์เป็นครั้งแรก การกำหนด Use Case ในลักษณะนี้ จะทำให้การบริหาร Project เป็นไปอย่างมีประสิทธิภาพและรวดเร็วขึ้น เพราะเริ่มต้นทดสอบเฉพาะคนบางกลุ่มและใช้เฉพาะในบางพื้นที่เท่านั้น นอกจากนี้ รัฐบาลยังมีการกำหนด Success Metrics ซึ่งถือเป็นหัวใจอีกอย่างหนึ่ง ของการดำเนิน Project ครั้งนี้ ได้แก่ การลดระยะเวลาในการตรวจคนเข้าเมืองลงอย่างน้อย 50% ความถูกต้องของการตรวจสอบด้วยระบบนี้ มีความถูกต้องสอดคล้องกับระบบ Biometric 98% เป็นอย่างน้อย ลดอัตราการตรวจคนเข้าเมืองด้วยวิธีแบบดั้งเดิมให้เหลือเพียง 2% และนักเดินทางมีความพึงพอใจในกระบวนการตรวจคนเข้าเมืองอย่างน้อย 90%

หลังจากที่กำหนดเป้าหมายและขอบเขต รวมถึง Success Metric ของโครงการเรียบร้อยแล้ว ขั้นตอนต่อไปคือการทำความเข้าใจเกี่ยวกับข้อมูลและการประเมินความพร้อมของข้อมูลที่จะใช้ในการ Develop Project ครั้งนี้ โดยข้อมูลที่ต้องการได้แก่ รูปภาพใบหน้าและดวงตาของผู้เดินทางแบบ High Resolution ข้อมูลใน Passport ข้อมูลการเดินทาง (Travel Logs) และข้อมูล Timestamp ในการเดินทาง ทั้งนี้ ข้อมูลดังกล่าวสามารถดึงได้จาก Biometric Data จากระบบทะเบียน National ID ของประเทศสิงคโปร์ รวมถึง SG Arrival Card และระบบกล้องติดตามที่ Immigration Gates อย่างไรก็ตาม ประเด็นที่ต้องจัดการคือ รูปภาพใบหน้าบางครั้งอาจมีความแตกต่างกัน ด้วยเรื่องของแสง มุม อายุที่เปลี่ยนไป หรือบางครั้งอาจมีการสวมใส่แว่นตาดำ หรือ Mask ทำให้ระบุคนได้ไม่ชัด ดังนั้น เป้าหมายในขั้นนี้คือ การที่กำหนดว่า ระบบใหม่จะต้องทำงานได้ดี ไม่ผิดพลาดมากเกินกว่า Success Metrics ภายใต้ข้อมูลที่มีอย่างจำกัด

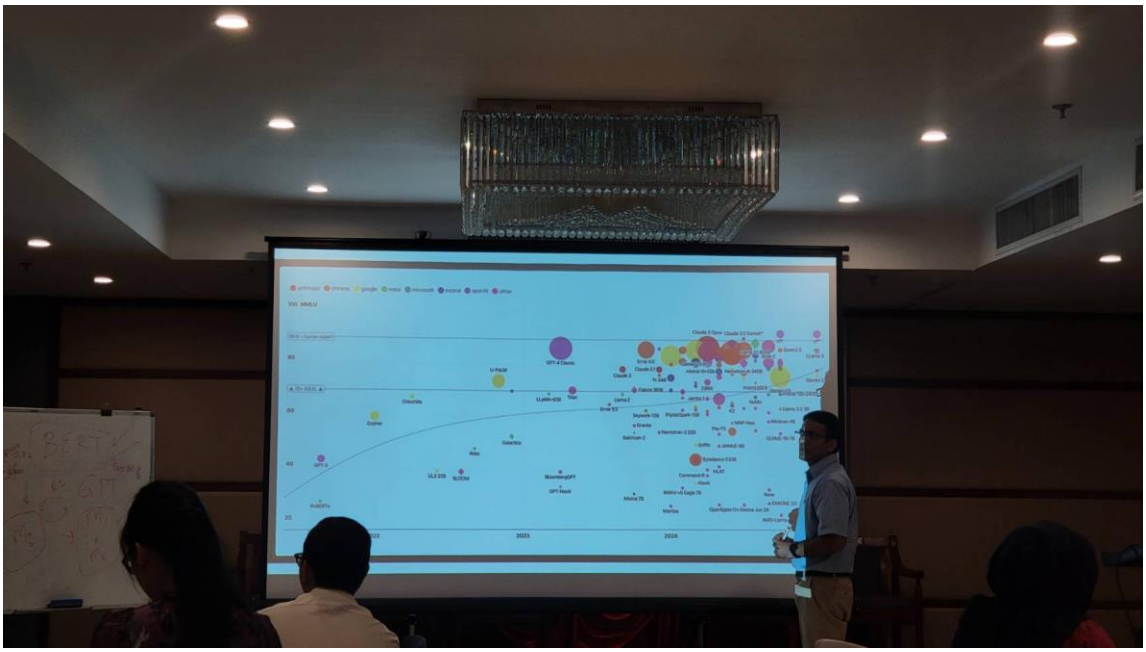
เมื่อประเมินข้อมูลเรียบร้อยแล้ว จึงต้องมีการเตรียมข้อมูล โดยรวบรวมข้อมูล Feature ที่จำเป็นได้แก่ รูปภาพใบหน้า รูปภาพดวงตา ข้อมูลเกี่ยวกับสัญชาติ Timestamp Gate ID และข้อมูล Match Result ทั้งหมดนี้ สามารถรวมรวมได้จากหลายแห่ง อาทิ กล้องจับภาพบริเวณสนามบิน หรือข้อมูลการเดินทางจาก Arrival Cards System โดยมีการ Clean ข้อมูลที่มีปัญหา เช่น Duplicated Traveler Records หรือ Inconsistent Image Formats ก่อนดำเนินการ หลังจากนั้น รัฐบาลสิงคโปร์จึงได้ตัดสินใจทำ Project โดยใช้การสร้าง Biometric Model ขึ้นมาเอง ไม่ใช้งาน Third-party Software เพราะมีความเสี่ยงต่อการรั่วไหลของข้อมูล มีการพัฒนาโมเดลดังกล่าว โดยใช้ Convolutional Neural Networks (CNN) สำหรับการตรวจจับใบหน้าและดวงตา มีการทดสอบโมเดลในหลายรูปแบบ ทำ A/B Testing เพื่อปรับโมเดลจนกระทั่งสามารถใช้งานได้ดี

หลังจากที่มีการทดลองนำไปใช้ขั้นแรก พบว่า โมเดลดังกล่าวมักจะพบปัญหา False Rejection (ปฏิเสธคนที่สมควรได้เข้าประเทศ) ในกรณีที่มีแสงสว่างไม่เพียงพอ หรือบุคคลมีการสวมใส่หมวก แว่นตา หรือหน้ากาก ในกรณีนี้ถือว่ายอมรับได้ เพราะมีเพียงบางคนเท่านั้นที่ได้รับผลกระทบ แต่แก้ไขได้โดยให้บุคคลที่ติดปัญหา เข้าช่องตรวจคนเข้าเมืองในรูปแบบปกติ การทดสอบสามารถใช้การได้ดีในเกือบทุกสภาวะ (ร้อยละ 99) ทำให้ท้ายสุด รัฐบาลสิงคโปร์ได้มีการนำโมเดลดังกล่าวไปใช้จริงที่สนามบิน และมีการทำ Retraining โมเดลทุก 3-6 เดือน (เนื่องจากผู้เดินทางอาจมีอายุที่เพิ่มขึ้น หรือมีการเปลี่ยนแปลงโครงสร้างใบหน้า ทำให้ต้องอัพเดทข้อมูลตลอด)

Session 16: Group Discussion

โดย Dr. Chutiporn Anutariya, Associate Professor, Information Management, Department of Information and Communication Technologies, Asian Institute of Technology, Thailand

ผู้บรรยายได้เปิดโอกาสให้ผู้เข้าร่วมทุกคนได้มีโอกาสทำ Reflection ว่า ตลอดระยะเวลาฝึกอบรมได้มีการเรียนรู้อะไรบ้าง โดยมีการทบทวน (Recap) เนื้อหาทั้งหมดตั้งแต่การอบรมวันแรกจนถึงวันสุดท้าย ผู้เข้าอบรมทุกคนต่างมีส่วนร่วมในการแลกเปลี่ยนสิ่งที่ตนเองได้เรียนรู้ระหว่างการฝึกอบรมในครั้งนี้ โดยส่วนใหญ่ ต่างได้รับความรู้เกี่ยวกับกระบวนการทำ Data Analytics การใช้โปรแกรม Software ที่เกี่ยวข้อง และความรู้ทั้งในด้านสถิติ การทำ Data Visualization รวมถึงหลักการเกี่ยวกับ AI และ Data Mining ที่เป็นประโยชน์ต่อการแก้ปัญหาขององค์กร



รูปภาพบรรยากาศการฝึกอบรมหลักสูตร 25-CP-06-GE-TRC-A

Training Course on AI and Data Analytics for Productivity Enhancement

ประโยชน์ที่ได้รับจากหลักสูตร 25-CP-06-GE-TRC-A
Training Course on AI and Data Analytics for Productivity Enhancement
และการขยายผลจากการเข้าร่วมโครงการ

การเข้าร่วมหลักสูตรในครั้งนี้ ทำให้ข้าพเจ้าหลักสูตรมีความเข้าใจพื้นฐานที่ชัดเจนเกี่ยวกับการวิเคราะห์ข้อมูล (Data Analytics) ในลำดับขั้นตอนที่ครอบคลุมทั้งหมด ได้แก่

1. การทำความเข้าใจธุรกิจและปัญหา ซึ่งเป็นขั้นตอนที่สำคัญมาก เพราะหากเราไม่เข้าใจปัญหาที่ต้องการแก้ไขอย่างถ่องแท้ การวิเคราะห์ข้อมูลจะไม่ก่อให้เกิดประโยชน์แต่อย่างใด
2. การเก็บรวบรวมข้อมูล ต้องสอดคล้องกับปัญหาที่ต้องการแก้ไข โดยสามารถใช้เครื่องมือที่ศึกษาในชั้นเรียน อาทิ Web Scraping เพื่อช่วยให้ได้ข้อมูลที่ครบถ้วน
3. การเตรียมข้อมูล (Data Preprocessing) เป็นขั้นตอนที่สำคัญมากเช่นกัน ต้องทำการ Clean ข้อมูล เช่น การจัดการ Missing Value, Duplicate Values, Outlier เพื่อให้ผลการวิเคราะห์แม่นยำมากขึ้น
4. การวิเคราะห์ข้อมูล (Data Analysis) โดยผู้สอนแนะนำหลากหลายเทคนิค เช่น Linear Regression, K-Means, Random Forest ซึ่งแม้จะเป็นเนื้อหาที่ลึก แต่ช่วยให้ข้าพเจ้ามีคุ้นเคยกับคำศัพท์เหล่านี้ และสามารถศึกษาต่อด้วยตนเองได้ในอนาคต
5. การนำเสนอผลการวิเคราะห์ (Data Visualization) โดยเน้นว่าการเลือกเครื่องมือให้เหมาะสมเป็นสิ่งสำคัญ เช่น Pie Chart อาจไม่เหมาะกับการเปรียบเทียบเพราะข้อจำกัดด้านการรับรู้ของมนุษย์ ขณะที่ Table ในรูปแบบธรรมดา อาจมีประโยชน์มากกว่าในหลายบริบท

นอกจากนี้ ข้าพเจ้ายังได้เรียนรู้การใช้เครื่องมือพื้นฐาน เช่น Orange, Tableau และ Octoparse ซึ่งเป็นจุดเริ่มต้นที่ดีในการทำ Data Analytics ในองค์กรต่อไป

<u>แนวทางการนำความรู้ไปประยุกต์ใช้ในองค์กร</u>

ข้าพเจ้าต้องการนำเทคโนโลยีด้าน Data Analytics และ AI มาประยุกต์ใช้เพื่อเพิ่มประสิทธิภาพในกระบวนการทำงาน โดยเฉพาะในฝ่ายปฏิบัติการและทรัพยากรบุคคล ซึ่งในปัจจุบันองค์กรของข้าพเจ้ายังประสบปัญหา เช่น ผลิตภาพมีความผันผวนสูง การใช้กำลังคนไม่เหมาะสม และการรับคนที่ไม่เหมาะกับงาน โดยแนวทางที่ข้าพเจ้าจะดำเนินการ ได้แก่

- เก็บข้อมูลเกี่ยวกับประสิทธิภาพของพนักงานรายบุคคล (เช่น อัตราการผลิตต่อชั่วโมง)
- เก็บข้อมูลคุณลักษณะของพนักงาน เช่น คะแนนด้านสติปัญญา ระดับความอดทนทางกายภาพ และการประสานระหว่างมือและตา (Hand-eye Co-ordination)
- วิเคราะห์และสร้างโมเดลเพื่อพยากรณ์ประสิทธิภาพของการทำงานบุคลากรแต่ละคน

ข้าพเจ้าประเมินว่า สิ่งที่ดำเนินการข้างต้นจะนำไปสู่เป้าหมายหลัก 3 ข้อ ได้แก่ การจัดพนักงานให้เหมาะสมกับลักษณะงาน การคัดกรองผู้สมัครที่ไม่เหมาะสมตั้งแต่ต้น การคาดการณ์ได้ว่าพนักงานรายใดมีความเสี่ยงต่อผลการปฏิบัติงานต่ำ เพื่อป้องกันและแก้ไขได้ทันเวลา ทั้งหมดนี้ จะช่วยยกระดับประสิทธิภาพขององค์กรได้อย่างยั่งยืน

<u>กิจกรรมขยายผลที่จะดำเนินการภายใน 6 เดือน</u>

หลังจากจบหลักสูตร ข้าพเจ้าจะดำเนินโครงการนำร่องเพื่อพัฒนาเครื่องมือช่วยคัดกรองผู้สมัครงานในองค์กร โดยนำความรู้ด้าน Data Analytics และเครื่องมือที่ได้เรียนรู้จากหลักสูตรมาใช้จริง โดยเฉพาะการประยุกต์ใช้ Orange ในการวิเคราะห์ข้อมูล และ Tableau ในการสร้าง Dashboard สำหรับการตัดสินใจ แนวคิดหลักคือ การเก็บข้อมูลจากผู้สมัครในกระบวนการคัดเลือก เช่น คะแนนจากแบบทดสอบความสามารถด้านสติปัญญา แบบทดสอบเฉพาะทาง และข้อมูลเบื้องต้นที่เกี่ยวข้อง เช่น อายุ วุฒิการศึกษา หรือประสบการณ์ จากนั้นนำข้อมูลเหล่านี้มาวิเคราะห์ด้วย Orange โดยใช้เทคนิคเช่น K-Means หรือ Decision Tree เพื่อจำแนกกลุ่มผู้สมัครตามระดับความเหมาะสม

การดำเนินการดังกล่าวจะช่วยให้ฝ่ายทรัพยากรบุคคลและหัวหน้าที่เกี่ยวข้อง สามารถพิจารณาข้อมูลได้อย่างรวดเร็วและมีประสิทธิภาพในวันสัมภาษณ์จริง โดยระบบจะแสดงผลคะแนนของแต่ละผู้สมัคร เปรียบเทียบกับค่าเฉลี่ยของกลุ่ม และจัดอันดับตามความเหมาะสมกับตำแหน่งงานที่ต้องการ ซึ่งช่วยเพิ่มประสิทธิภาพและลดเวลาในการคัดกรอง

ข้าพเจ้าคาดว่าจะสามารถดำเนินการนำระบบเบื้องต้นนี้มาใช้ทดสอบจริงภายในเดือนธันวาคม 2568 ซึ่งจะเป็นการวางรากฐานสำคัญสำหรับการนำ Data Analytics มาใช้ในองค์กรอย่างเป็นระบบต่อไปในอนาคต นอกจากนี้ ข้าพเจ้ายังมีแผนจะขอคำปรึกษาเพิ่มเติมจากผู้บรรยายผ่านช่องทาง Discord ของ APO หากพบว่า มีปัญหาและอุปสรรคเกิดขึ้น

<u>อุปสรรคหรือข้อจำกัดที่อาจเผชิญ</u>

สิ่งที่ท้าทายคือองค์กรของข้าพเจ้า คือ การที่ปัจจุบันยังไม่มีผู้เชี่ยวชาญด้าน Data Science หรือ AI และอาจไม่สามารถลงทุนจ้างบุคลากรเฉพาะทางได้ในระยะสั้น เนื่องจากการขาดแคลนทรัพยากรทางการเงิน แนวทางที่เป็นไปได้คือการสรรหาคนรุ่นใหม่ที่มีพื้นฐานด้านสถิติหรือเทคโนโลยีที่ดี (แต่ไม่จำเป็นต้องมีพื้นฐานด้าน Data Science และ AI) และส่งเสริมให้พนักงานกลุ่มนี้เรียนรู้และเติบโตไปพร้อมกับองค์กร โดยข้าพเจ้าสามารถทำหน้าที่เป็นโค้ชสำหรับพนักงานรุ่นใหม่เหล่านี้

อีกหนึ่งประเด็นที่มีความท้าทาย คือ ปัญหาทางเทคนิคที่อาจเกิดขึ้นระหว่างการทดลองใช้เครื่องมือที่ได้รับการฝึกสอนในคลาสเรียน ซึ่งอาจต้องใช้เวลาและความพยายามในการเรียนรู้พอสมควร ข้าพเจ้าจึงตั้งใจจะศึกษาหาความรู้เพิ่มเติมอย่างต่อเนื่อง และเชื่อมต่อกับผู้เชี่ยวชาญผ่านช่องทางเช่น Discord ของ APO เพื่อทำให้สามารถเรียนรู้สิ่งเหล่านี้ได้อย่างมีประสิทธิภาพและรวดเร็วยิ่งขึ้น

<u>แนวคิดเกี่ยวกับการเรียนรู้และการพัฒนาต่อยอดในอนาคต</u>

ข้าพเจ้ามีความสนใจในการเข้าร่วมการฝึกอบรมหรือ Workshop ที่เน้นการนำเสนอกรณีศึกษา (Use Cases) เกี่ยวกับการประยุกต์ใช้ AI และ Data Analytics ในภาคการผลิต รวมถึงการศึกษาดูงานจากบริษัทต้นแบบที่มีความก้าวหน้าในการใช้เทคโนโลยีเหล่านี้เพื่อเพิ่มขีดความสามารถในการแข่งขันขององค์กร โดยในอนาคตอันสั้น ข้าพเจ้ามุ่งหวังที่จะยกระดับองค์กรของข้าพเจ้าให้สามารถขับเคลื่อนไปข้างหน้า โดยใช้ Data Analytics และ AI เป็นเครื่องมือในการยกระดับผลิตภาพอย่างเต็มรูปแบบ

การอบรมในหลักสูตรครั้งนี้ ได้เปิดโอกาสให้ข้าพเจ้าได้รับความรู้ของทั้งกระบวนการในการทำ Data Analytics รวมถึงประเด็นสำคัญเกี่ยวกับ AI อีกทั้งยังจุดประกายแนวคิดในการพัฒนาองค์กรผ่านการใช้ข้อมูลอย่างมีประสิทธิภาพ ในโอกาสนี้ ข้าพเจ้าจึงขอขอบคุณ APO และสถาบันเพิ่มผลผลิตแห่งชาติ เป็นอย่างยิ่งที่ได้ให้โอกาสข้าพเจ้าได้เข้าร่วมหลักสูตรในครั้งนี้ ข้าพเจ้าตระหนักดีว่า การเปลี่ยนแปลงเชิงระบบต้องอาศัยทั้งความรู้ ความตั้งใจ และความร่วมมือจากทีมงาน ข้าพเจ้าจึงมุ่งมั่นที่จะนำแนวทางที่ได้เรียนรู้ไปปรับใช้จริงในองค์กร เพื่อยกระดับการบริหารจัดการให้สอดคล้องกับเป้าหมายด้านประสิทธิภาพ ผลิตภาพ และเพิ่มระดับความสามารถในการแข่งขันขององค์กรได้อย่างยั่งยืน

รายละเอียดหลักสูตร 25-CP-06-GE-TRC-A

Training Course on AI and Data Analytics for Productivity Enhancement

25-CP-06-GE-TRC-A
Training Course on AI and Data Analytics for Productivity Enhancement
5 - 9 May 2025

Implementing Organizations: National Productivity Center of Cambodia (NPCC), Ministry of Industry, Science, Technology and Innovation

Time (Local Time)	Agenda	Facilitator/Speaker
Day 0: Sunday, 4 May 2025		
Participants Arrival Venue: POULO WAI Hotel & Apartment Address: #71-73, Oknha Ket Street (St. 174), Sangkat Phsar Thmey 3, Khan Daun Penh, Phnom Penh, Cambodia Tel. No.: +855 23 211 666 Website: www.poulowaihotel.com		
Day 1: Monday, 5 May 2025		
08:30-09:00	Registration of Participants Venue: POULO WAI Hotel & Apartment	NPCC
09:00-09:30	**Opening Session** Welcome Remarks by NPCC Opening Remarks by APO Introduction of Resource Persons and Participants Group Photo	NPCC and APO
09:30-09:50	Coffee/Tea break	
09:50-10:10	**Overview of the project**	APO
10:10-11:10	**Session 1: Introduction to AI and Data Analytics** This session provides a foundational overview of AI and its role in transforming data analytics. Participants will explore key concepts such as machine learning, transformers, generative AI, and recent breakthroughs and industry applications.	**Dr. Chintan Amrit** Senior Associate Professor, University of Amsterdam, Netherlands
11:10-12:10	**Session 2: Installation and configuration of basic tools** This session covers setting up essential tools for AI and data analytics. Participants will configure their environments and learn best practices for a smooth workflow, preparing for hands-on activities in subsequent sessions.	**Dr. Murphy Choy** CEO, Alionova Education, Singapore
12:10-13:40	Lunch Break	

13:40-15:00	**Session 3: Hands-on exercises with introductory datasets** This session offers practical experience in working with foundational datasets. Participants will apply data cleaning, manipulation, and visualization techniques using and analytics tools. Through guided exercises, they will gain insights into identifying patterns, trends, and relationships with data.	**Dr. Murphy Choy** Support by all RPs Dr. Chintan Amrit Dr. Chutiporn Anutariya Quinn Pham
15:00-15:20	Coffee/Tea break	
15:20-16:30	**Session 4: Case studies of AI and Data Analytics in various sectors** This session explores real-world AI and data analytics applications across diverse industries, including finance, healthcare, retail, and marketing. Participants will analyze case studies to understand how organizations leverage AI to gain insights, optimize operations, and drive innovation.	**Quinn Pham** Director of Solution Consulting, Meiro Pte. Ltd.
18:30-20:00	Welcome Dinner Venue: (TBC)	APO
End of Day 1		

Time (Local Time)	Agenda	Facilitator/Speaker
Day 2: Tuesday, 6 July 2025		
08:45-09:00	Registration of Participants	NPCC
09:00-10:20	**Session 5: Data Collection Techniques** This session provides an overview of effective data collection methods for AI and data analytics projects. Participants will learn about structured and unstructured data sources, web scraping, APIs, and ethical considerations in data gathering. Practical tips for ensuring data quality and relevance will also be covered.	**Dr. Chutiporn Anutariya** Associate Professor, Information Management, Department of Information and Communication Technologies, Asian Institute of Technology, Thailand
10:20-10:40	Coffee/Tea Break	
10:40-12:00	**Session 6: Data Collection Exercises** In this session, participants will engage in hands-on activities to practice collecting data from various sources. They will explore methods such as manual data entry, accessing public datasets, and fetching data, and will learn how to identify reliable data sources, structure the data for analysis, and overcome common challenges during data collection. Through the exercises, they will have experience gathering datasets that are ready for cleaning and analysis in subsequent steps.	**Dr. Chtiporn Anutariya** Support by all RPs Dr. Chintan Amrit Dr. Murphy Choy Quinn Pham

12:00-13:30	Lunch	
13:30-15:00	**Session 7: Data Preprocessing** In this session, Participants will learn essential techniques for preparing data for analysis. This includes handling missing values, outlier detection, data normalization, and feature engineering to improve data quality and model performance.	**Dr. Murphy Choy**
15:00-15:20	Coffee/Tea break	
15:20-17:00	**Session 8: Data Preprocessing Exercises** This hands-on session provides practical experience in cleaning and transforming datasets. Participants will apply preprocessing techniques using tools such as Python's Pandas library to ensure the data is ready for advanced analytics.	**Dr. Murphy Choy** Support by all RPs Dr. Chintan Amrit Dr. Chutiporn Anutariya Quinn Pham
End of Day 2		

Time (Local Time)	Agenda	Facilitator/Speaker
Day 3: Wednesday, 7 May 2025		
08:45-09:00	Registration of Participants	NPCC
09:00-10:00	**Session 9: Data Analysis Techniques** Participants will explore key methods for analyzing data, including statistical analysis, pattern identification, and predictive modeling. Emphasis will be placed on actionable insights and informed decision-making.	**Dr. Murphy Choy**
10:00-10:20	Coffee/Tea Break	
10:20-12:00	**Session 10: Data Analysis Exercises** Through guided exercises, participants will apply data analysis techniques to real datasets. They will learn how to interpret results, identify correlations, and generate insights for business applications.	**Dr. Murphy Choy** Support by all RPs Dr. Chintan Amrit Dr. Chutiporn Anutariya 3. Quinn Pham
12:00-13:30	Lunch break	NPCC
13:00-15:30	**Session 11: Data Visualization Principles and Tools** This session covers the principles and tools for visualizing data effectively. Participants will learn to select appropriate chart types, design meaningful dashboards, and communicate findings visually.	**Dr. Chintan Amrit**
15:30-15:50	Coffee/Tea Break	
15:50-17:00	**Session 12: Data Visualization Exercises** Participants will practice creating impactful visualizations using Power BI or Looker Studio tools.	**Quinn Pham** Support by all RPs Dr. Murphy Choy

	They will focus on translating data insights into easy-to-understand graphics and dashboards.	Dr. Chutiporn Anutariya Quinn Pham
End of Day 3		

Time (Local Time)	Agenda	Facilitator/Speaker
Day 4: Thursday, 8 May 2025		
08:45-09:00	Registration of Participants	NPCC
09:00-10:00	Travel for site visit	NPCC
10:00-11:00	[Site visit 1] AI Farm Factory AI Farm Robotics is leading a technological transformation in Phnom Penh, Cambodia. Founded in 2017, the company is dedicated to advancing robotics and automation, aiming to position Cambodia as a prominent player in the global robotics industry. With a mission to make Cambodia a technologically independent and sovereign nation, AI Farm Robotics focuses on creating autonomous systems capable of producing other robots.	NPCC All RPs
11:30-13:30	Lunch	NPCC
13:30-14:00	Travel for site visit	
14:00-16:00	[Site visit 2] Cambodia Academy Digital Technology The Cambodia Academy of Digital Technology (CADT) is a prominent institution dedicated to advancing digital technology and innovation in Cambodia. Established in 2014, CADT aims to nurture digital talent and drive the country toward a digital society.	NPCC All RPs
16:00–17:00	Return to hotel	
18:30–20:00	Farewell Dinner Venue: (TBC)	NPCC
End of Day 4		

Time (Local Time)	Agenda	Facilitator/Speaker
Day 5: Friday, 9 May 2025		
08:45-09:00	Registration of Participants	
09:00-10:00	**Session 13: AI-Driven Decision Making** This session explores how AI models and data analytics inform strategic decisions. Participants will understand how businesses leverage AI for forecasting, customer behavior analysis, and optimization.	**Dr. Murphy Choy**
10:00-10:20	Coffee/Tea Break	

10:20-11:30	**Session 14: Implementing AI Solutions** In this session, participants will learn the practical steps for implementing AI solutions in organizational settings. This includes best practices for deployment, integration, and scaling AI-driven projects.	**Quinn Pham**
11:30-13:00	Lunch break	
13:00-14:00	**Session 15: Future Trends in AI and Data Analytics** This session discusses emerging technologies and innovations shaping AI and data analytics. Participants will explore trends such as generative AI, ethical AI, and advancements in automation, preparing them for future developments in this field.	**Dr. Chintan Amrit**
14:00-14:20	Coffee/Tea Break	
14:20-15:00	**Session 16: Group Discussion** In this session, participants will engage in a discussion reflecting on key takeaways from the training course. They will share insights, discuss possible challenges in implementing AI solutions, and exchange ideas for collaboration.	**Dr. Chutiporn Anutariya**
15:00-15:30	Closing Session Closing Remarks by NPCC Program Evaluation Group Photo	NPCC and APO
End of the Program		

Time (Local Time)	Agenda	Facilitator/Speaker
Day 6: Saturday, 6 May, 2025		
Participants Departure		