การจัดกลุ่มผู้บริโภคเพื่อวางแผนกลยุทธ์ทางการตลาดด้วยวิธีการแบ่งกลุ่มข้อมูลแบบเคมีน

Customer clustering for a marketing strategy with K-Means Clustering on the customer's records

ธรณ์ธันย์ ยะสุคำ

วิทยาลัยนวัตกรรม มหาวิทยาลัยธรรมศาสตร์

Digital Business Transformation, College of Innovation, Thammasat University

บทคัดย่อ

รายงานฉบับนี้นำเสนอการจัดกลุ่มข้อมูลผ่านกระบวนการทำเหมืองข้อมูลโดยใช้อัลกอริทึมการเรียนรู้โดยไม่มีผู้สอน Unsupervised Learning แบบ Clustering Algorithm ประกอบด้วย 4 ขั้นตอนหลัก เริ่มต้นด้วยการสำรวจข้อมูลเพื่อการทำความเข้าใจข้อมูลและทำความเข้าใจปัญหา จากนั้นเข้าสู่กระบวนการเตรียมข้อมูลซึ่งประกอบด้วยขั้นตอนย่อยได้แก่ การลบข้อมูลที่ไม่สมบูรณ์ การจัดการกับ feature การจัดการค่าผิดปกติ การเปลี่ยนชนิดข้อมูลและปรับข้อมูลให้เป็นการแจกแจงปกติมาตรฐาน ตลอดจนการลดมิติของข้อมูล (Dimensionality Reduction) จากนั้นจึงนำผลลัพธ์ที่ได้จากการการจัดกลุ่มผู้บริโภคไปใช้ศึกษาหาความต้องการเพื่อนำไปสู่สร้างกลยุทธ์ทางการตลาด

การลดขนาดของข้อมูลในรายงานฉบับนี้จะกล่าวถึงเฉพาะวิธี Principal component analysis หรือ PCA เท่านั้น ซึ่งเป็นกระบวนการลดมิติของข้อมูลหนึ่งที่จะช่วยลดขนาดของข้อมูลที่มีความสำคัญน้อยให้กลายเป็นกลุ่มของข้อมูลใหม่เพื่อนำเสนอผลลัพธ์ในรูปแบบสามมิติ

คำสำคัญ: การจัดกลุ่มข้อมูล (Clustering), กลยุทธ์ทางการตลาด(marketing strategy), การลดมิติ (Dimensionality Reduction)

ABSTRACT

This project represented data mining techniques in the Unsupervised Learning (Descriptive Modeling) part of the Clustering Algorithm. This project has 4 main parts, Data exploration, Data preparation including clear missing value, Feature Engineering, Label encoding, Scaling, and Dimensionality Reduction(Principal component analysis (PCA) on customer record for representing cluster on 3-dimension), Modeling, Evaluation, and final is the conclusion. The result of the project is to cluster customers for make marketing strategy.

1.บทนำ

ในสังคมปัจจุบันที่มีการแข่งขันสูงในทุกด้านโดยเฉพาะด้านธุรกิจที่ต้องการความรวดเร็วและความแม่นยำในการหาความต้องการของผู้บริโภคที่ถูกต้อง หลายองค์กรจึงพยายามเข้าถึงลูกค้าทุกกลุ่ม ดังนั้น คุณลักษณะของกลุ่มลูกค้าจึงเป็นข้อมูลสำคัญที่จะนำมาซึ่งกลยุทธ์ทางการตลาด ทำให้หลายองค์กรเริ่มนำข้อมูลเข้ามาตัดสินใจ เพราะทุกองค์กรล้วนเข้าใจว่าข้อมูลเป็นสิ่งที่สำคัญในยุคปัจจุบันแต่องค์กรส่วนใหญ่กลับไม่ได้ใช้ประโยชน์จากข้อมูลได้อย่างเต็มที่จึงเกิดเทคนิคการทำเหมืองข้อมูลหรือดาต้าไมนนิง (Data Mining Techniques) ที่เป็นกระบวนการช่วยให้ข้อมูลเกิดประโยชน์มากที่สุด

ดาต้าไมนนิง (Data mining) (พนิดา และพยุง , 2547) คือ กระบวนการที่จะสกัดความรู้ที่มีประโยชน์จากข้อมูลดิบที่มีความสัมพันธของขอมูลซ่อนเร้นอยู่ ที่เราไม่ทราบมาก่อน ทำให้เกิดศักยภาพในการใช้ข้อมูลในฐานข้อมูล ดาต้าไมนนิง มีอยู่ทั้งหมด 5 รูปแบบ

1. Association Rule (ความเชื่อมโยงหรือความสัมพันธ์ของขอมูลตั้งแต่สองชุดขึ้นไป)

2.Classification and Prediction (การจำแนกประเภทและทำนายค่าบางอย่างที่ไม่รู้)

3.Cluster analysis (การวิเคราะห์กลุ่ม),

4.Outlier analysis (การค้นหาค่าผิดปกติที่เกิดขึ้นในข้อมูล)

5.Trend and evolution analysis (การวิเคราะห์แนวโน้ม)

จากทั้งหมดที่กล่าวมา การไมนิ่งข้อมูลมีวิธีการและกระบวนการหลากหลาย ดังนั้นในการใช้งานควรเลือกมาใช้ให้เหมาะสมกับลักษณะของข้อมูลซึ่งรายงานฉบับนี้มุ่งเน้นไปที่การวิเคราะห์กลุ่มกับข้อมูลลูกค้า ด้วยความคาดหวังว่ารูปแบบข้อมูลที่ได้จะสามารถนำมาสร้างกลยุทธ์ทางการตลาด

1belief Company Thailand (2560) ให้ความหมายกลยุทธ์การตลาดว่า เป็นแบบแผนพื้นฐานหรือแนวทางที่ถูกกำหนดขึ้นสำหรับสร้างผลิตภัณฑ์เพื่อตอบสนองความต้องการของกลุ่มเป้าหมายและตลาดเป้าหมาย ถูกใช้เป็นเครื่องมือในการสำหรับการต่อสู้แข่งขันกันในการเข้าถึงกลุ่มเป้าหมายของบริษัทและตลาดคู่แข่ง โดยเบื้องต้นผู้ผลิตจะต้องมีคือการดำเนินงานที่มีขั้นตอน การตัดสินใจในเกี่ยวกับงบประมาณค่าใช้จ่ายทางการตลาดที่เหมาะสม รวมทั้งจะต้องสามารถจัดสรรทรัพยากรที่มีอยู่มาใช้ให้เกิดประโยชน์และได้ผลงานที่มีคุณภาพสูงสุด

การใช้การวิเคราะห์กลุ่มในการตลาด (Use of cluster analysis in marketing) (Girish and David , 1983) หลักในการวิเคราะห์กลุ่มในการตลาดประกอบไปด้วย ข้อที่หนึ่ง คือเพื่อแบ่งส่วนตลาด(Marketing Segmentation) หมายถึงวิธีที่จะระบุกลุ่มผู้บริโภคที่คล้ายกันโดยมีเกณฑ์กำหนด เช่น ความชอบ, แนวโน้มที่จะซื้อสินค้าและ ทัศนคติที่ส่งผลต่อตัดสินใจในการซื้อสินค้า เป็นต้น ข้อที่สอง คือการเข้าถึงพฤติกรรมของผู้บริโภคที่อยู่ในกลุ่มเดียวกัน ทำให้การแบ่งกลุ่มถูกนำมาใช้เพื่อพัฒนาสินค้าในแต่ละกลุ่ม ทำให้องค์กรสามารถกำหนดได้ว่าควรจะเสนอสินค้า

แบบใดและให้กับผู้บริโภคกลุ่มไหนจึงจะมีโอกาสได้รับการตอบรับมากที่สุด ดังนั้น ในรายงานเล่มนี้จะกล่าวถึงดาต้าไมน์นิงในส่วนของการวิเคราะห์กลุ่ม (Cluster analysis) ผ่านวิธีการจัดกลุ่มในแบบ k-Means algorithm

K-Means Clustering (วันเพ็ญ, นีลวัสน์, และ ณรงค์ศักดิ์ , 2559) เป็นเทคนิคการแบ่งกลุ่มข้อมูลแบบไม่เป็นขั้นตอน (Nonhierarchical Cluster Analysis) ผู้ใช้จะต้องกำหนดเองว่าต้องการแบ่งเป็นกี่กลุ่ม กระบวนการการทำงานเป็นแบบวนซ้ำหลายรอบ แต่ละรอบจะเกิดการรวมกลุ่มที่พิจารณาจากระยะห่างค่ากลางของกลุ่ม จากนั้นทำการคำนวณค่ากลางของกลุ่มใหม่เพื่อเปรียบเทียบ ทำซ้ำกระบวนการจนกระทั่งค่ากลางไม่เปลี่ยนแปลงหรือครบรอบที่กำหนดไว้

การลดขนาดข้อมูล (Data Reduction) (ธรรมศักดิ์, 2548) ในการทำเหมืองข้อมูลสำหรับทำอัลกอริทึมจัดกลุ่มข้อมูลแต่ละอัลกอริทึมมีวัตถุประสงค์และข้อกำหนดสำหรับข้อมูลที่ใช้แตกต่างกัน อัลกอริทึมจัดกลุ่มข้อมูลบางอัลกอริทึมอาจมีข้อจำกัดเรื่องความสามารถในการรองรับข้อมูลที่มีขนาดใหญ่ การใช้ข้อมูลที่มีขนาดเหมาะสมจึงเป็นสิ่งที่ต้องคำนึงถึงด้วยเช่นกันในการทำเหมืองข้อมูลจากข้อมูลที่ไม่ได้ผ่านการเตรียมข้อมูลอาจเกิดปัญหาขึ้นในระหว่างกระบวนการได้ เช่น หน่วยความจำไม่เพียงพอหรือใช้เวลานานจนไม่สามารถนำผลการวิเคราะห์มาใช้ประโยชน์ การลดขนาดข้อมูลเป็นกระบวนการหนึ่งในขั้นตอนการเตรียมข้อมูล นั่นคือการทำให้ข้อมูลตั้งต้นมีขนาดลดลงโดยสูญเสียลักษณะสำคัญของข้อมูลและสูญเสียความถูกต้องของผลลัพธ์น้อยที่สุดเพื่อสามารถใช้เป็นตัวแทนของข้อมูลส่วนใหญ่ได้

2.วัตถุประสงค์

2.1 เพื่อแบ่งส่วนตลาดโดยอาศัยเกณฑ์ด้านพฤติกรรม จิตวิทยา และข้อมูลส่วนตัวของผู้บริโภคหลังจากการแบ่งกลุ่ม เพื่อสร้างกลยุทธ์ทางการตลาดที่ทำให้บรรลุวัตถุประสงค์ตามที่วางแผนไว้ เช่น การทำโปรโมชั่น การรณรงค์ทางการตลาด การผลิตสินค้าที่ตรงกับความต้องการ

2.2 เสนอวิธีการทำ K-means ซึ่งหนึ่งในการทำเหมืองข้อมูลเพื่อการอธิบาย (descriptive data mining) ซึ่งเป็นการค้นหารูปแบบของกลุ่มข้อมูลที่มีความสัมพันธ์กันโดยไม่ได้เจาะจงเพื่อหารูปแบบหรือโมเดลของข้อมูลอย่างหนึ่งอย่างใดเพื่อการทำนายเท่านั้นแต่เป็นการค้นหาทุกรูปแบบที่น่าสนใจและเนื่องจากข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลเปิด ดังนั้น จำนวนในการแบ่งกลุ่มจึงไม่สามารถพิจารณาจากองค์กรโดยตรงได้ จึงใช้ silhouette score ในการค้นหาจำนวนกลุ่ม

2.3 เพื่อศึกษาการลดมิติของข้อมูลด้วยเทคนิค PCA (Principal Components Analysis) เนื่องจากการแสดงผลผลลัพธ์ของข้อมูลด้วยแผนภาพ (Data visualization) ผ่านไลบรารี่ Matplotlib ในรูปแบบหลายมิติถือเป็นเรื่องที่ยากและไม่เหมาะสม จึงใช้วิธีการลดมิติดังกล่าวเพื่อลดมิติของข้อมูลให้เหลือเพียง 3 มิติ สำหรับใช้นำเสนอผลลัพธ์ให้อยู่ในรูปแบบของสามมิติ

3.การทบทวนวรรณกรรม (Literature review)

3.1 แนวคิดการจัดกลุ่ม(Clustering) การจัดกลุ่มเป็นการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) หมายถึง การเรียนรู้จากข้อมูลที่ไม่มีการกำหนดเป้าหมาย (Target) แต่ใช้ข้อมูลในการแบ่งกลุ่มเพื่อการวิเคราะห์

3.2 การแบ่งส่วนตลาด (Marketing Segmentation) มีแนวคิดที่คล้ายกับวิเคราะห์การจัดกลุ่มแต่เปลี่ยนจากจุดข้อมูลเป็นผู้บริโภค ดังนั้น ผู้บริโภคที่คล้ายกันไว้จะถูกจัดอยู่ในกลุ่มเดียวกัน

3.3 เกณฑ์สำหรับการแบ่งส่วนตลาด brandingchamp (2563) กล่าวถึงการแบ่งส่วนตลาดได้ ดังนี้

3.3.1 แบ่งตามประชากรศาสตร์ เช่น เพศ การศึกษา หรือขนาดครอบครัว

3.3.2 แบ่งตามพฤติกรรม เช่น ความพร้อมผู้ซื้อ โอกาสในการซื้อ

3.3.3 แบ่งตามภูมิศาสตร์ เช่น ลูกค้าจากจังหวัดในภาคเหนือ

3.3.4 แบ่งตามจิตวิทยา เช่น รสนิยม

3.4 หลักการทำกลยุทธ์ทางการตลาดให้ตรงกับเป้าหมาย 1belief Company Thailand (2560) เสนอกลยุทธ์ทางการตลาดที่นักธุรกิจควรรู้ คือ การทำตลาดให้ตรงกับกลุ่มเป้าหมายและการรักษาลูกค้าปัจจุบัน ด้วยการประชาสัมพันธ์หรือการจัดโปรโมชั่น เพื่อการสื่อสารให้ตรงกับกลุ่มเป้าหมายอย่างชัดเจนที่สุด ซึ่งอาจส่งผลให้เกิดการจัดทำการตลาดใหม่ ๆ ที่ตอบโจทย์ผู้บริโภคในปัจจุบันและตอบสนองความต้องการได้ตรงกับผู้บริโภครายใหม่ ดังนั้น ในการทำจัดกลุ่มผู้บริโภคจะทำให้นักธุรกิจเห็นภาพของกลุ่มหมายที่มีอยู่และเห็นภาพกลุ่มเป้าหมายใหม่ในการดึงกลุ่มเป้าหมายเหล่านี้ให้ผูกติดกับตลาด

3.5 การจัดทำเหมืองข้อมูล เป็นการค้นความรู้ที่ซ้อนอยู่ในข้อมูลดิบ ประกอบด้วย 5 ขั้นตอนหลัก ได้แก่

3.5.1 กำหนดวัตถุประสงค์

3.5.2 การเตรียมข้อมูล ซึ่งประกอบด้วย การคัดเลือกเลือกข้อมูลที่ต้องการจากนั้น ตรวจสอบข้อมูลทางสถิติเพื่อลบข้อมูลที่ไม่จำเป็นออกและทำปรับข้อมูลให้เหมาะสมกับอัลกอรึทึม

3.5.3 การทำดาต้าไมนิง หมายถึง การประมวลข้อมูลตามอัลกอรึทึมที่ใช้

3.5.4 วิเคราะห์ผลลัพธ์

3.5.5 การนำไปประยุกต์ใช้

3.6 แนวคิดการจัดกลุ่มด้วย K-MEANS ในการจัดกลุ่มด้วยวิธีนี้จำเป็นต้องกำหนดจำนวนกลุ่มก่อนเริ่มการทำงาน โดยผลลัพธ์ของการจัดกลุ่มจะขึ้นอยู่กับระยะห่างของข้อมูลแต่ละจุดกับจุดกึ่งกลางของกลุ่ม (Centroid) ถ้าข้อมูลจุดใดอยู่ใกล้กับจุดกึ่งกลางของกลุ่มจะถือว่าเป็นสมาชิกของกลุ่มนั้น หลังจากนั้นจึงวนซ้ำการทำงานคำนวนจุดกึ่งกลางใหม่และเปรียบเทียบว่ามีค่าเปลี่ยนแปลงหรือไม่ ถ้าไม่มีจะถือว่าจบการทำงาน

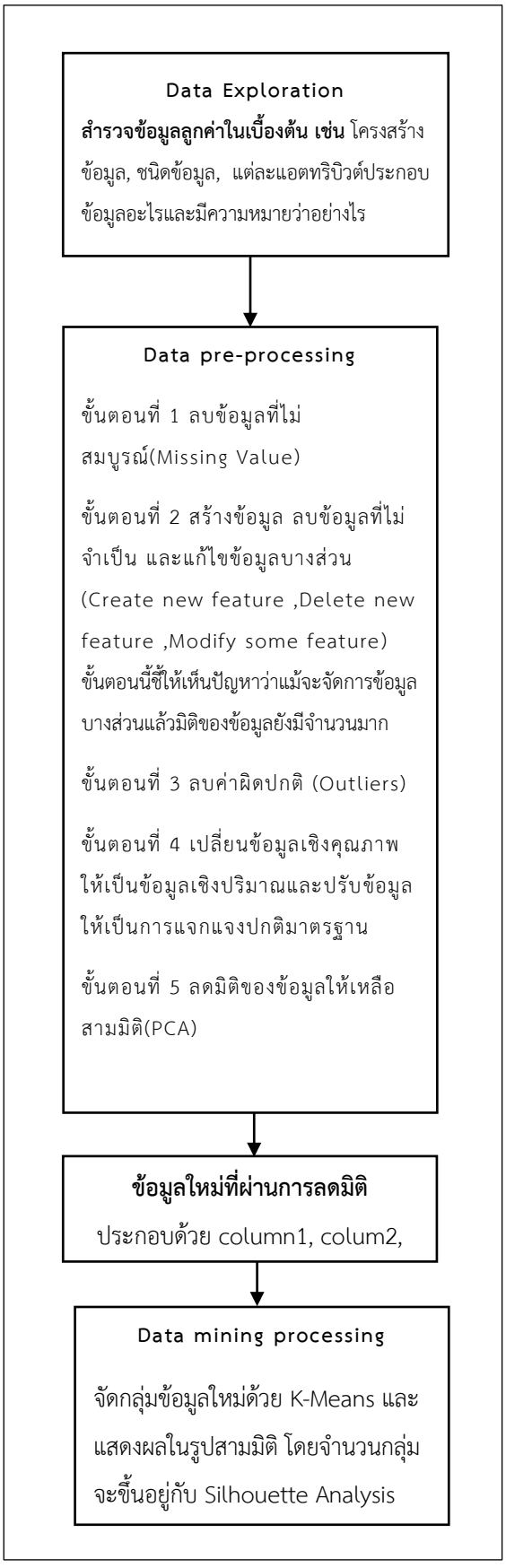
3.7 การเลือกจำนวนกลุ่มด้วยวิธี silhouette analysis (Weerasak , 2017) เป็นค่าใช้สำหรับวัดความสัมพันธ์ของ Cluster กับระยะทางระหว่าง Cluster และระยะทางภายใน Cluster เพื่อเปรียบเทียบว่ามีความเหมือนกับกลุ่ม

ของตัวเองมากแค่ไหนเมื่อเทียบกับกลุ่มอื่นๆ ผลลัพธ์ที่ได้จะเป็นค่า Silhouette score (Coefficient [-1,1]) หมายความว่า ถ้ามีค่าความคล้ายของกลุ่มของตัวเองมากและมีความค่าคล้ายกับกลุ่มอื่นน้อย

3.8 การลดขนาดของข้อมูล (ธรรมศักดิ์, 2548) คือ ตัดแอททริบิวต์เพื่อให้มิติของข้อมูลที่ไม่เกี่ยวข้องหรือเกี่ยวข้องกันน้อยออกไปแต่ไม่ส่งผลกับข้อมูลหรืออาจจะส่งผลเพียงแค่เล็กน้อย เพื่อให้ได้มาซึ่งรูปแบบของข้อมูลใหม่ที่มีคุณลักษณะโดยรวมเหมือนเดิมแต่มีความกระชับและเหมาะสมแก่การนำไปวิเคราะห์ในขั้นตอนต่อไป

3.9 การวิเคราะห์องค์ประกอบหลัก (Principal Components Analysis) (ปริญญา ,2562) คือการสร้างปริภูมิใหม่ที่มีความแปรปรวนสูงที่สุดด้วยการหมุนแกน แต่ยังคงไว้ซึ่งทิศทางเดิม เปรียบเทียบได้กับการแยกลูกอมหลายสีที่กระจัดกระจายออกจากถาด ถ้าหากถาดเล็กคงเป็นเรื่องยากที่จะคัดแยกลูกอมได้แต่เมื่อถาดมีขนาดใหญ่ขึ้นจะทำให้มีพื้นที่ในการคัดแยกและคัดแยกได้ง่ายขึ้น ขนาดถาดเปรียบเทียบกับความแปรปรวน แต่ทั้งหมดเปลี่ยนแปลงแค่ขนาดของถาดเท่านั้น รูปแบบการกระจายของลูกอมยังคงไว้ทิศทางเดิม ทั้งหมดนี้คือแนวคิดพื้นฐานของ PCA

4.กรอบแนวคิดในการวิจัย (Conceptual Model)



5.การสำรวจข้อมูล (Data Exploration)

ก่อนเริ่มการทำความสะอาดข้อมูลที่ไม่สมบูรณ์จำเป็นต้องตรวจสอบพื้นฐานข้อมูลก่อน ข้อมูลที่นำมาวิเคราะห์เป็นข้อมูลผู้บริโภคของห้างสรรพสินค้าแห่งหนึ่ง จากการตรวจสอบเบื้องต้นพบว่าโครงสร้างของข้อมูลประกอบไปด้วย 2240 แถว 29 คอลัมน์ และชนิดของข้อมูลประกอบด้วย ข้อมูลเชิงปริมาณ และ ข้อมูลเชิงคุณภาพ จากการสำรวจโครงสร้างสามารถวางแผนเบื้องต้นได้ว่า มิติข้อมูลในปัจจุบันมีค่อนข้างมากและมีข้อมูลที่ไม่ใช่ตัวเลขประกอบอยู่ เมื่อทำความสะอาดข้อมูลเสร็จแล้วจำเป็นต้องเปลี่ยนชนิดของข้อมูลและลดมิติของข้อมูลให้น้อยลง

```
#df.shape
print("Data shape: " + str(df.shape[0]) + " row " + str(df.shape[1]) + " Column")
```

```
Data shape: 2240 row 29 Column
```

ภาพที่ 1 จำนวนแถวและหลักของข้อมูล

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
 #   Column               Non-Null Count  Dtype
---  ------               --------------  -----
 0   ID                   2240 non-null   int64
 1   Year_Birth           2240 non-null   int64
 2   Education            2240 non-null   object
 3   Marital_Status       2240 non-null   object
 4   Income               2216 non-null   float64
 5   Kidhome              2240 non-null   int64
 6   Teenhome             2240 non-null   int64
 7   Dt_Customer          2240 non-null   object
 8   Recency              2240 non-null   int64
 9   MntWines             2240 non-null   int64
 10  MntFruits            2240 non-null   int64
 11  MntMeatProducts      2240 non-null   int64
 12  MntFishProducts      2240 non-null   int64
 13  MntSweetProducts     2240 non-null   int64
 14  MntGoldProds         2240 non-null   int64
 15  NumDealsPurchases    2240 non-null   int64
 16  NumWebPurchases      2240 non-null   int64
 17  NumCatalogPurchases  2240 non-null   int64
 18  NumStorePurchases    2240 non-null   int64
 19  NumWebVisitsMonth    2240 non-null   int64
 20  AcceptedCmp3         2240 non-null   int64
 21  AcceptedCmp4         2240 non-null   int64
 22  AcceptedCmp5         2240 non-null   int64
 23  AcceptedCmp1         2240 non-null   int64
 24  AcceptedCmp2         2240 non-null   int64
 25  Complain             2240 non-null   int64
 26  Z_CostContact        2240 non-null   int64
 27  Z_Revenue            2240 non-null   int64
 28  Response             2240 non-null   int64
dtypes: float64(1), int64(25), object(3)
memory usage: 507.6+ KB
```

ภาพที่ 2 ชนิดของข้อมูลแค่ละแอททริบิวต์

ทั้ง 29 แอตทริบิวต์ของข้อมูลแสดงรายละเอียดข้อมูลเบื้องต้นของผู้บริโภค ดังนี้

AcceptedCmp1 - 1 if customer accepted the offer in the 1st campaign, 0 otherwise

AcceptedCmp2 - 1 if customer accepted the offer in the 2nd campaign, 0 otherwise

AcceptedCmp3 - 1 if customer accepted the offer in the 3rd campaign, 0 otherwise

AcceptedCmp4 - 1 if customer accepted the offer in the 4th campaign, 0 otherwise

AcceptedCmp5 - 1 if customer accepted the offer in the 5th campaign, 0 otherwise

Response (target) - 1 if customer accepted the offer in the last campaign, 0 otherwise

Complain - 1 if customer complained in the last 2 years

DtCustomer - date of customer's enrolment with the company

Education - customer's level of education

Marital - customer's marital status

Kidhome - number of small children in customer's household

Teenhome - number of teenagers in customer's household

Income - customer's yearly household income

MntFishProducts - amount spent on fish products in the last 2 years

MntMeatProducts - amount spent on meat products in the last 2 years

MntFruits - amount spent on fruits products in the last 2 years

MntSweetProducts - amount spent on sweet products in the last 2 years

MntWines - amount spent on wine products in the last 2 years

MntGoldProds - amount spent on gold products in the last 2 years

NumDealsPurchases - number of purchases made with discount

NumCatalogPurchases - number of purchases made using catalogue

NumStorePurchases - number of purchases made directly in stores

NumWebPurchases - number of purchases made through company's web site

NumWebVisitsMonth - number of visits to company's web site in the last month

Recency - number of days since the last purchase

จากนั้นทำการสำรวจเพิ่มเติม ผู้จัดทำสังเกตว่าในส่วน Education, Marital_Status, Dt_Customer เป็นข้อมูลเชิงคุณภาพและทุกข้อมูลไม่ได้เป็นข้อมูลแบบ Binary data แต่เป็น Ordinal data

```
# count value
for col in object_cols:
    print("Total categories in the feature ",col  ,"total", data[col].value_counts(), "\n")
```

```
Total categories in the feature  Education total Graduation    1116
PhD            481
Master         365
2n Cycle       200
Basic           54
Name: Education, dtype: int64

Total categories in the feature  Marital_Status total Married     857
Together    573
Single      471
Divorced    232
Widow        76
Alone         3
Absurd        2
YOLO          2
Name: Marital_Status, dtype: int64

Total categories in the feature  Dt_Customer total 2012-08-31    12
2012-09-12    11
2013-02-14    11
2014-05-12    11
2014-05-22    10
              ..
2013-12-28     1
2012-09-19     1
2012-09-30     1
2014-01-12     1
2012-11-27     1
Name: Dt_Customer, Length: 662, dtype: int64
```

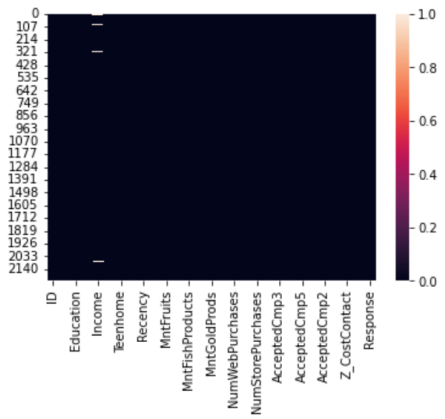
6.การเตรียมขอมูลสำหรับทำดาตาไมนนิง (Data preparation)

หลังจากสำรวจข้อมูลผู้จัดทำได้แบ่งขั้นตอนการเตรียมข้อมูล ออกเป็น 4 ขั้นตอน ดังนี้

6.1 ทำความสะอาดข้อมูลที่ไม่สมบูรณ์(Cleaning missing data)

```
cols = df.columns
sns.heatmap(df[cols].isnull())
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f6d3bd3dc50>
```



ภาพที่ 3 Heatmap of missing value

จากภาพที่ 3 เมื่อใช้ความสามารถของ Heatmap จะพบ Missing value อยู่ในคอลัมน์ Income ซึ่งจำเป็นที่จะต้องลบค่าว่างนี้ออก

```
df = df.dropna()
print("total row after removing" , len(df))
```

```
total row after removing 2216
```

ภาพที่ 4 delete missing value

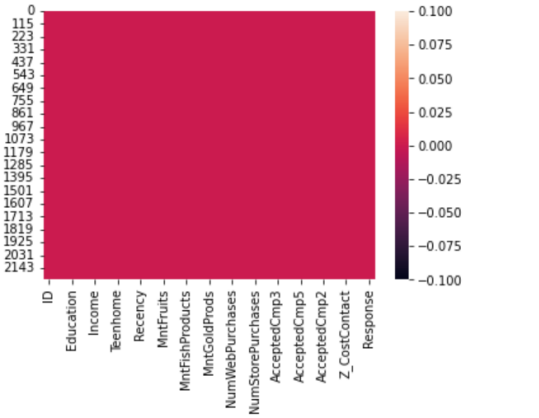
หลังจากลบ Missing value ด้วย drop.na() จะทำให้ขณะนี้ข้อมูลเหลือ 2216 แถว (24 ข้อมูลที่หายไปเป็น Missing value) เมื่อตรวจสอบข้อมูลอีกครั้งพบว่าขณะนี้ข้อมูลไม่มี Missing value เหลืออยู่ในข้อมูลแล้ว

```
df.isna().sum()
```

```
ID                     0
Year_Birth             0
Education              0
Marital_Status         0
Income                 0
Kidhome                0
Teenhome               0
Dt_Customer            0
Recency                0
MntWines               0
MntFruits              0
MntMeatProducts        0
MntFishProducts        0
MntSweetProducts       0
MntGoldProds           0
NumDealsPurchases      0
NumWebPurchases        0
NumCatalogPurchases    0
NumStorePurchases      0
NumWebVisitsMonth      0
AcceptedCmp3           0
AcceptedCmp4           0
AcceptedCmp5           0
AcceptedCmp1           0
AcceptedCmp2           0
Complain               0
Z_CostContact          0
Z_Revenue              0
Response               0
dtype: int64
```

ภาพที่ 5 check missing value

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fcfab83b450>
```



ภาพที่ 6 Recheck Heatmap of missing value

6.2 สร้างข้อมูล ลบข้อมูลที่ไม่จำเป็นและแก้ไขข้อมูลบางส่วน (Create new feature, Delete new feature ,Modify some feature)

6.2.1 แอตทริบิวต์ที่ถูกสร้างขึ้นใหม่ ประกอบด้วย

- **Customer_member_time**: สร้างแอตทริบิวต์บันทึกระยะเวลาในการเป็นสมาชิกของลูกค้าโดยนับตั้งแต่วันที่ลูกค้าลงทะเบียน (DT_Customer) จนถึงลูกค้ารายล่าสุดที่ลงทะเบียน (max(DT_Customer))
- **Age**: สร้างแอตทริบิวต์เก็บอายุลูกค้า
- **Total_spent**: รวมค่าใช้จ่ายให้เป็นหนึ่งส่วน
- **Relationship**: เปลี่ยนความสัมพันธ์เป็น Binary data ประกอบด้วย Alone(โสด) และ Lover(มีคู่รัก)
- **Children**: รวมจำนวนสมาชิกทายาทให้ทายาทที่เป็นเด็กและทายาทที่เป็นวัยรุ่นรวมเป็นหนึ่งส่วน
- **Family_size**: จำนวนสมาชิกในครอบครัวจะเกิดจากความสัมพันธ์(Relationship) ถ้าโสดแทนค่าด้วยเลข 1 ถ้ามีคู่จะแทนค่าด้วยเลข 2 รวมกับจำนวนรวมจำนวนสมาชิกทายาท
- **Is_parent**: ถ้าหากมีบุตรหลายจะมีค่าเท่ากับ 1 แต่ถ้าหากไม่มีจะมีค่าเท่ากับ 0
- **Education**: เปลี่ยนความสัมพันธ์เป็น Binary data ประกอบด้วย graduate (สำเร็จการศึกษาปริญญาตรีเป็นขั้นพื้นฐาน) และ Undergraduate

6.2.2 แอตทริบิวต์ที่ถูกแก้ไขชื่อเพื่อความเข้าใจ โดยข้อมูลที่ถูกแก้ไขชื่อทั้งหมดประกอบด้วย

แอตทริบิวต์เดิม	แอตทริบิวต์ใหม่
MntWines	Wines
MntFruits	Fruits
MntMeatProducts	Meat
MntFishProducts	Fish
MntSweetProducts	Sweets
MntGoldProds	Gold

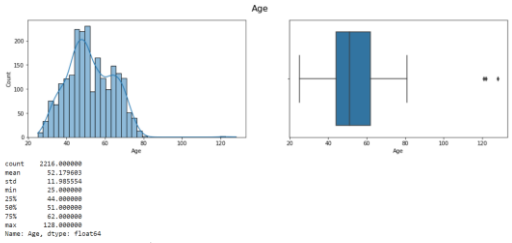
6.2.3 แอตทริบิวต์ที่ถูกลบ ประกอบด้วย

- Marital_Status Dt_Customer
- Z_CostContact
- Z_Revenue
- Year_Birth
- ID

6.3 ลบค่าผิดปกติ (Outliers)

ในการลบค่าผิดปกติ ผู้จัดทำได้สร้างฟังชั่นในการเพื่อดูการกระจายและค่าผิดปกติจากกราฟ Boxplot และรายละเอียดจาก describe()

```
def check_outlier_numeric(feature,dataset):
    fig, ax = plt.subplots(1,2)
    fig.set_size_inches(16,4)
    fig.suptitle(feature, fontsize=16)
    sns.histplot(data=dataset, x=feature,
                 kde=True, ax=ax[0])
    sns.boxplot(data=dataset, x=feature
                ,ax=ax[1])
    plt.show()
        print(dataset[feature].describe())
```



ภาพที่ 7 Example output from function

เมื่อสร้างฟังชั่นสำหรับลบค่าผิดปกติแล้ว เมื่อวิเคราะห์จาก Boxplot ผู้จัดทำจะไม่ลบค่าผิดปกติทั้งชุดข้อมูลแต่ลบเฉพาะบางส่วนของข้อมูลประกอบด้วย

1. Age ที่มากกว่า 90 ปี
2. Income ที่มากกว่า 600000
3. Total_Spent ที่มากกว่า 2500

6.4 เปลี่ยนข้อมูลเชิงคุณภาพให้เป็นข้อมูลเชิงปริมาณและปรับข้อมูลให้เป็นการแจกแจงปกติมาตรฐาน

6.4.1 ตรวจสอบข้อมูลคุณภาพ

```
[99] #Get list of categorical variables
     object_item = (data.dtypes == 'object')
     object_cols = list(object_item[object_item].index)

     print("Categorical variables in the dataset:", object_cols)

     Categorical variables in the dataset: ['Education', 'Relationship']
```

ภาพที่ 8 Show output Qualitative data

หลังจากที่จัดการข้อมูลไปในขั้นตอนที่แล้วปรากฏว่าในขณะนี้ข้อมูลเชิงคุณภาพมีเพียงสองจำนวน

6.4.2 เปลี่ยนแปลงชนิดของข้อมูลใช้โมดูล LabelEncodeing จากไลบรารี่ scikit-learn

```
[100] #Label Encoding the object dtypes.
      labeler = LabelEncoder()
      for i in object_cols:
          data[i] = data[[i]].apply(labeler.fit_transform)

      print("All features are now numerical")

      All features are now numerical
```

ภาพที่ 9 Check numerical

6.4.3 ปรับข้อมูลให้เป็นการแจกแจงปกติมาตรฐานโดยการใช้โมดูล StandardScaler จากไลบรารี่ scikit-learn

```
[104] scaler = StandardScaler()
      #Scaling
      scaler.fit(data_numeric)
      #create dataframe for modeling
      scaled_data = pd.DataFrame(scaler.transform(data_numeric), columns=data_numeric.columns)
```

ภาพที่ 9 Scaling

Dataframe to be used for further modelling:

	Education	Income	Kidhome	Teenhome	Recency
0	0.358843	0.289892	-0.823691	-0.930876	0.309801
1	0.358843	-0.258869	1.038677	0.906752	-0.381078
2	0.358843	0.916868	-0.823691	-0.930876	-0.795605
3	0.358843	-1.175393	1.038677	-0.930876	-0.795605
4	0.358843	0.297104	1.038677	-0.930876	1.553383

ภาพที่ 10 Example some Z-Score Normalization of data

เมื่อเสร็จสิ้นกระบวนการทั้งสามจะพบว่าข้อมูลในปัจจุบันได้เป็นข้อมูลเชิงปริมาณที่ผ่านการปรับค่าให้อยู่ในรูปการแจกแจงปกติมาตรฐาน (standard normal distribution) ดังรูปภาพที่ 10

6.5 การลดขนาดของข้อมูลด้วยวิเคราะห์องค์ประกอบหลัก (Principle Component Analysis (PCA))

ในการทำ PCA สามารถใช้โมดูล PCA จากไลบรารี่ scikit-learn โดยอัลกอริธึม PCA สามารถสรุปได้ในรูปภาพที่ 11 จากนั้นกำหนดเป้าหมายให้เหลือเพียง 3 component หรือ 3 มิติสุดท้ายจะได้ข้อมูลในรูปแบบใหม่ ดังรูปภาพที่ 13 และทำ data visualization ใน 3 มิติได้ ดังรูปภาพที่ 14

1. Compute the mean feature vector
   $\mu = \frac{1}{p}\sum_{k=1}^{p} x_k$, where, x_k is a pattern ($k = 1$ to p), p = number of patterns, x is the feature matrix
2. Find the covariance matrix
   $C = \frac{1}{p}\sum_{k=1}^{p}\{x_k - \mu\}\{x_k - \mu\}^T$ where, T represents matrix transposition
3. Compute Eigen values λ_i and Eigen vectors v_i of covariance matrix
   $Cv_i = \lambda_i v_i$ $(i = 1, 2, 3, \ldots q)$, q = number of features
4. Estimating high-valued Eigen vectors
   (i) Arrange all the Eigen values (λ_i) in descending order
   (ii) Choose a threshold value, θ
   (iii) Number of high-valued λ_i can be chosen so as to satisfy the relationship
   $\left(\sum_{i=1}^{s}\lambda_i\right)\left(\sum_{i=1}^{q}\lambda_i\right)^{-1} \geq \theta$, where, s = number of high valued λ_i chosen
   (iv) Select Eigen vectors corresponding to selected high valued λ_i
5. Extract low dimensional feature vectors (principal components) from raw feature matrix.
   $P = V^T x$, where, V is the matrix of principal components and x is the feature matrix

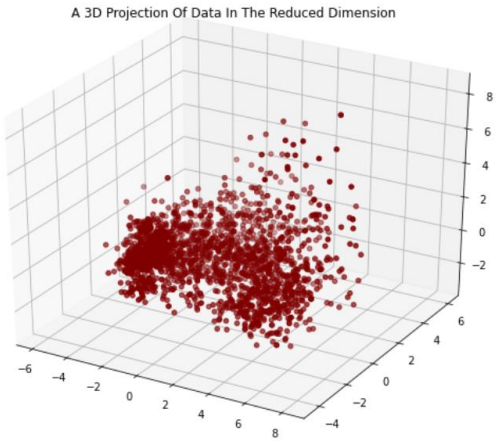
ภาพที่ 11 The PCA algorithm (Vijay ,Sirnicasa, Sriram, 2013, p.646)

```
[169] #Initiating PCA to reduce dimentions aka features to 3
      pca = PCA(n_components=3)
      pca.fit(scaled_data)
      PCA_data = pd.DataFrame(pca.transform(scaled_data), columns=(["col1","col2","col3"]))
```

ภาพที่ 12 PCA implementation

	col1	col2	col3
0	4.944775	-0.322338	0.310618
1	-2.942899	0.061989	-0.360208
2	2.368151	-0.759181	-1.148926
3	-2.734283	-1.444043	0.011422

ภาพที่ 13 Example data after PCA



ภาพที่ 11 A 3D Projection Of Data In The Reduced Dimension

7.กระบวนการทำดาตาไมนนิง (Data mining processing)

7.1 Silhouette Analysis

หลังจากเสร็จสิ้นกระบวนการเตรียมข้อมูลขณะนี้ ข้อมูลทั้งหมดพร้อมสำหรับการจัดกลุ่มด้วย K-Means ในขั้นตอนแรกของการดำเนินการจำเป็นต้องเลือกจำนวนกลุ่มก่อน โดยผู้จัดทำใช้ Silhouette Analysis ในการเลือกจำนวนกลุ่มที่ดีที่สุด

```
[172] #K-mean find num of cluster
      avgs = []
      min_k = 2
      x = np.array(list(zip(PCA_data['col1'],PCA_data['col2'],PCA_data['col3'])))

      for k in range(min_k,10):
        km = KMeans(n_clusters=k).fit(x)
        s = metrics.silhouette_score(x , km.labels_)
        print("Silhouette Coefficients of k =", k ,"is", s)
      avgs.append(s)

      suitableK = avgs.index(max(avgs)) + min_k
      print("Optimal K is ", suitableK)

      Silhouette Coefficients of k = 2 is 0.4545277236070744
      Silhouette Coefficients of k = 3 is 0.415611164169147
      Silhouette Coefficients of k = 4 is 0.41322773530015183
      Silhouette Coefficients of k = 5 is 0.36566808816825913
      Silhouette Coefficients of k = 6 is 0.34451444399167047
      Silhouette Coefficients of k = 7 is 0.3534764528867503
      Silhouette Coefficients of k = 8 is 0.36891787004804766
      Silhouette Coefficients of k = 9 is 0.34830446342844384
      Optimal K is  2
```

ภาพที่ 14 Best component of group

หลังจากการทำ Silhouette Analysis พบว่าในการเลือกจำนวนกลุ่มที่ดีที่สุด คือ การแบ่งกลุ่มข้อมูลให้มีเพียงแค่ 2 กลุ่มเท่านั้น

7.2 K-MEANS

Algorithm: *k*-means. The *k*-means algorithm for partitioning based on the mean value of the object in the cluster.
Input: The number of cluster *k* and a database containing *n* objects.
Output: A set of *k* clusters that minimizes the squared-error criterion.
Method:
(1) arbitrarily choose *k* objects as the initial cluster center;
(2) repeat
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, i.e., calculate the mean value of the objects for each cluster;
(5) until no change;

ภาพที่ 11 อัลกอริธึม k-means (ธรรมศักดิ์ , 2548 , p.8)

เมื่อกำหนดจำนวนกลุ่มในการแบ่งได้แล้ว ในการทำ K-MEANS สามารถใช้โมดูล KMeans จากไลบรารี่ scikit-learn

```
In [ ]: model=KMeans(n_clusters=suitableK)
        model

Out[ ]: KMeans(n_clusters=2)

In [ ]: x = PCA_data[['col1','col2','col3']]
        model.fit(x)

Out[ ]: KMeans(n_clusters=2)

In [ ]: x = np.array(list(zip(PCA_data['col1'],PCA_data['col2'],PCA_data['col3'])))
        labels = model.predict(x)
        centroids = model.cluster_centers_

        /usr/local/lib/python3.7/dist-packages/sklearn/base.py:446: UserWarning: X do
        s
          "X does not have valid feature names, but"

Centroids after clustering

In [ ]: centroids

Out[ ]: array([[-2.01309213, -0.12358633,  0.09088627],
               [ 3.30619079,  0.20297133, -0.14926657]])
```

ภาพที่ 15 K-mean

จากภาพที่ 15 จะได้จุดกึ่งกลางของแต่ละกลุ่มโดยกลุ่มที่ 1 อยู่ตำแหน่ง -2.01309213, 0.12358633, 0.09088627 กลุ่มที่ 2 อยู่ตำแหน่ง 3.30619079, 0.20297133, -0.14926657 หลังจากนั้นนำข้อมูลกลุ่มที่ได้ไปจัดอยู่ในข้อมูลที่ลดมิติแล้วจะได้ข้อมูลที่บอกสมาชิกของกลุ่ม

```
PCA_data['cluster'] = model.labels_
data['cluster'] = model.labels_
```

```
PCA_data
```

	col1	col2	col3	cluster
0	4.944778	-0.321905	0.297819	1
1	-2.942899	0.062128	-0.357598	0
2	2.368152	-0.759214	-1.148842	1
3	-2.734282	-1.444265	0.008384	0
4	-0.764421	0.212691	-0.776983	0
...	...	...	...	...

ภาพที่ 16 PCA data with cluster

สุดท้าย เมื่อได้ข้อมูลมที่บอกถึงสมาชิกกลุ่มแล้วนำข้อมูลที่บอกถึงจำนวนกลุ่มนั้นไปรวมกับข้อมูลตั้งต้นจะได้ผลลัพธ์ที่ต้องการนั้นคือข้อมูลที่บอกสมาชิกกลุ่มกับข้อมูลตั้งต้นจากนั้นทำ Data visualization จะได้ภาพสามมิติที่บอกถึงการแบ่งกลุ่มออกเป็นสองกลุ่ม

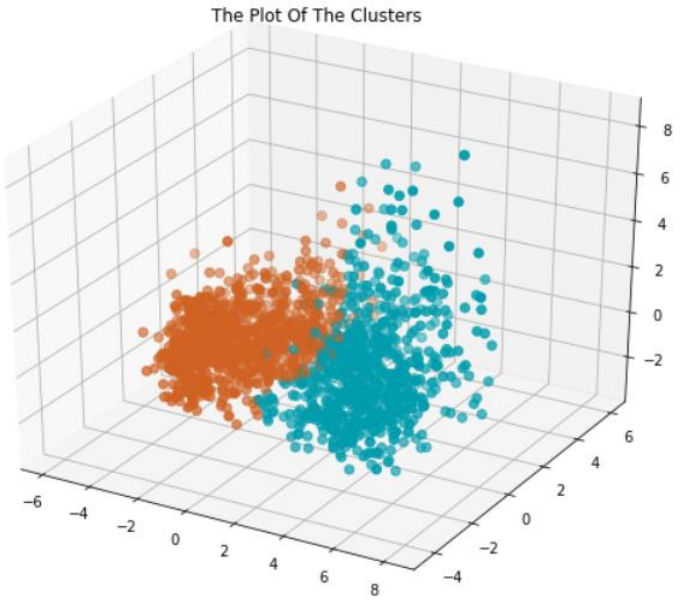
Customer_member_time	Age	Total_Spent	Relationship	Children	Family_Size	Is_Parent	cluster
57283200000000000	64	1617	0	0	1	0	1
9763200000000000	67	27	0	2	3	1	0
26956800000000000	56	776	1	0	2	0	1
12009600000000000	37	53	1	1	3	1	0

ภาพที่ 17 Example Data with Cluster

```
x =PCA_data["col1"]
y =PCA_data["col2"]
z =PCA_data["col3"]

color_plot = colors.ListedColormap(["#D06224","#009DAE"])

#Plotting the clusters
fig = plt.figure(figsize=(10,8))
ax = plt.subplot(111, projection='3d', label="bla")
ax.scatter(x, y, z, s=40, c=PCA_data["cluster"], marker='o', cmap = color_plot )
# ax.scatter(centroids[0][0], centroids[0][1],centroids[0][2], marker='*', s = 100, c = 'yellow')
ax.set_title("The Plot Of The Clusters")
plt.show()
print(centroids)
```



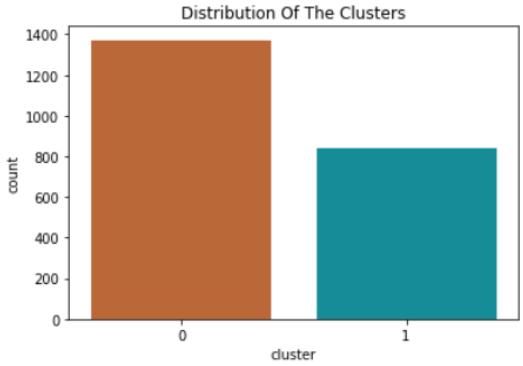

```
[[-2.01309213 -0.12358633  0.09088627]
 [ 3.30619079  0.20297133 -0.14926657]]
```

ภาพที่ 18 3-Dimention of Customer

7.3 ข้อสรุปจากโมเดล

7.3.1 ข้อสรุปประชากรศาสตร์

1. ข้อสรุป เรื่องจำนวน



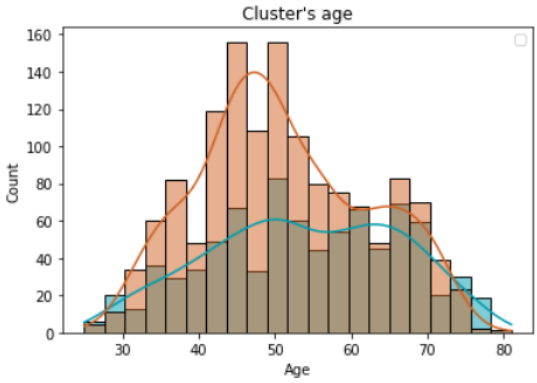

```
Count clusters in the feature cluster total
 0    1372
1     837
Name: cluster, dtype: int64
```

ภาพที่ 19 จำนวนสมาชิกในแต่ละกลุ่ม

<u>กลุ่มที่ 0</u> มีทั้งสิ้น 1372 คน

<u>กลุ่มที่ 1</u> มีทั้งสิ้น 837 คน

2. ข้อสรุป เรื่องอายุ

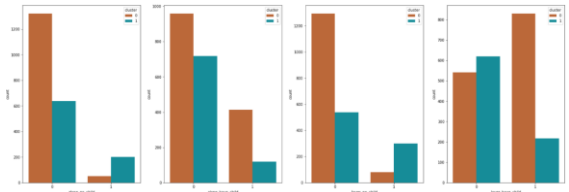


ภาพที่ 20 อายุแต่ละกลุ่ม

<u>กลุ่มที่ 0</u> มีการกระจายตัวอยู่ในช่วงวัยผู้ใหญ่เป็นส่วนใหญ่ตั้งแต่ 35 – 60 ปี

<u>กลุ่มที่ 1</u> มีการกระจายตัวคล้ายกับกลุ่มที่หนึ่งแต่มีจำนวนน้อยกว่าแต่ทั้งสองกลุ่มมีผู้บริโภควัยชราช่วง 70 ปีเป็นจำนวนมากเช่นเดียวกัน

3. ข้อสรุป ความสัมพันธ์



ภาพที่ 21 จำนวนความสัมพันธ์ทั้ง 4 รูป

ความสัมพันธ์มีทั้งหมด 4 รูปแบบ

- โสดแต่ไม่มีทายาท
- โสดแต่มีทายาท
- มีคู่แต่ไม่มีทายาท
- มีคู่และมีทายาท

จากภาพที่ 21 เรียงจากซ้ายไปขวาโดยแต่ละกราฟจะมีสองด้านคือซ้ายและขวา กราฟด้านซ้ายแทนคำว่าไม่มีความสัมพันธ์(No) กราฟด้านขวามีความสัมพันธ์(Yes)

<u>โสดแต่ไม่มีทายาท:</u> กลุ่มที่ 1 มีจำนวนมากกว่ากลุ่มที่ 0 เป็นเท่าตัว

<u>โสดแต่มีทายาท:</u> กลุ่มที่ 0 มีความสัมพันธ์เช่นนี้เป็นจำนวนมากและกลุ่มที่ 1 มีจำนวนใกล้เคียงกับความสัมพันธ์แบบโสดแต่ไม่มีทายาท

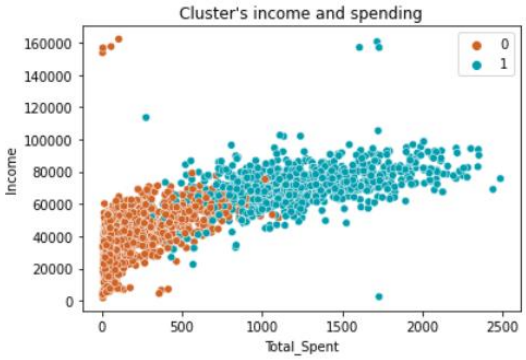
<u>มีคู่แต่ไม่มีทายาท:</u> กลุ่มที่ 0 ไม่ค่อยมีความสัมพันธ์เช่นนี้เมื่อเทียบกับกลุ่มที่ 1

<u>มีคู่และมีทายาท:</u> กลุ่มที่ 0 มีความสัมพันธ์ในลักษณะนี้สูงมากเมื่อเทียบกับความสัมพันธ์อื่นและกลุ่มที่ 1 มีจำนวนเทียบเท่ากับมีคู่แต่ไม่มีทายาท

จากผลลัพธ์พอที่จะคาดเดาได้ว่าลูกค้าส่วนใหญ่คือกลุ่มวัยผู้ใหญ่ที่มีครอบครัวโดยเฉพาะกลุ่มที่ 0 ที่มีความเป็นกลุ่มครอบครัวสูงมาก และกลุ่มที่ 1 มีความสัมพันธ์กระจัดกระจายมีจำนวนเท่าๆ กันในทุกความสัมพันธ์

7.3.2 ข้อสรุปด้านจิตวิทยา

1. ข้อสรุป เรื่องการนิสัยการใช้จ่าย

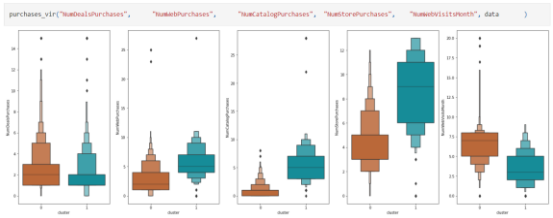


ภาพที่ 22 ความสัมพันธ์รายรับและจำนวนครั้งที่ใช้จ่าย

<u>กลุ่มที่ 0</u> มีรายได้ที่ค่อนข้างสูงแต่จำนวนครั้งในการใช้จ่ายซื้อสินค้ากลับมีน้อยบางส่วนแทบไม่เคยซื้อสินค้าเลย อาจเป็นกลุ่มคนประหยัดหรือเป็นกลุ่มที่ใช้จ่ายน้อยครั้งแต่ในการใช้จ่ายแต่ละครั้งมีจำนวนเงินที่สูง

<u>กลุ่มที่ 1</u> มีรายรับที่แปรผันตรงกับจำนวนครั้ง ยิ่งมีรายได้มากยิ่งเกิดการใช้จ่ายที่มาก ลูกค้ากลุ่มนี้อาจชื่นชอบในโปรโมชั่นของแถมหรือเป็นกลุ่มที่ชอบซื้อสินค้าเป็นพื้นฐานอยู่แล้ว

2. ข้อสรุป รูปแบบในการซื้อสินค้า



ภาพที่ 23 boxenplot about the shape of the distribution

รูปแบบในการซื้อสินค้ามีทั้งหมด 4 รูปแบบ

- จำนวนการซื้อที่มีส่วนลด
- จำนวนการซื้อผ่านเว็บไซต์
- จำนวนการซื้อที่ใช้แคตตาล็อก
- จำนวนการซื้อจากร้านค้าโดยตรง
- จำนวนครั้งที่ชมเว็บไซต์ในเดือนที่ผ่านมา

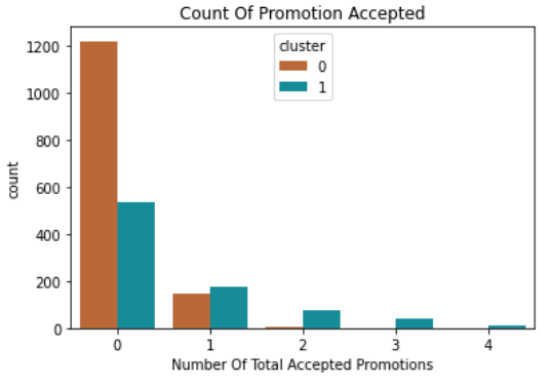
<u>กลุ่มที่ 0</u>

- ไม่ค่อยสนใจในส่วนลดมากนักอีกทั้งไม่สนใจสินค้าที่ทำแคตตาล็อกเลย
- สืบค้นข้อมูลสินค้าจากเว็บไซน์แต่ไม่นิยมซื้อทางออนไลน์มากเท่ากับซื้อโดยตรงจากร้านค้า

<u>กลุ่มที่ 1</u>

- ไม่ค่อยสนใจในส่วนลดมากแต่กลับสนใจสินค้าที่ทำแคตตาล็อกเลย
- มีโอกาสในการสืบค้นข้อมูลสินค้าจากเว็บไซน์และซื้อสินค้าทางออนไลน์เลย
- มีความสะดวกในการซื้อสินค้าจากทางร้านค้าโดยตรง

7.3.3 ข้อสรุปด้านพฤติกรรม



ภาพที่ 24 Number of accepted Promotions

จากการทำโปรโมชั่นที่ผ่านมาทั้งหมด 5 โปรโมชั่น ผู้บริโภคส่วนใหญ่มีพฤติกรรมที่ไม่ตอบรับโปรโมชั่นที่ได้จัดทำขึ้นและไม่มีกลุ่มใดเลยที่ตอบรับทุกโปรโมชั่น โดยเฉพาะกลุ่มที่ 0 ไม่เคยตอบรับโปรโมชั่นมากกว่า 2 โปรโมชั่นเลย แต่เมื่อเทียบพฤติกรรมกับกลุ่มที่ 1 ที่แม้ว่าจะไม่ค่อยตอบรับแต่ก็พอมีโอกาสตอบรับโปรโมชั่นในทุกครั้งและในกลุ่มนี้ได้มีการตอบรับมากที่สุดถึง 4 โปรโมชั่น

8. การประเมินผล (Evaluation Model)

Imad (2018) อธิบายว่าการประเมินผลของการแบ่งกลุ่ม (Clustering analysis)จะมีความแตกต่างจากการประเมินผลแบบ supervised learning ที่ไม่มีการวัดผลด้วยค่าสถิติ มากไปกว่านั้นในส่วนของ K-Means จำเป็นต้องกำหนดจำนวนกลุ่มก่อน ดังนั้นโมเดลที่ถูกต้องอาจขึ้นอยู่กับจำนวนกลุ่ม ในรายงานเล่มนี้ผู้จัดทำได้ทราบถึงปัญหาการวัดผลนี้ก่อนแล้วจึงได้ใช้ Silhouette analysis ในการวัดค่าความคลายระหว่างภายในกลุ่มและนอกกลุ่มเมื่อจัดกลุ่มด้วยจำนวนที่แตกต่างกัน ผลลัพธ์คือจำนวนกลุ่มที่ดีที่สุดคือทั้งหมด 2 กลุ่มตามที่ได้นำเสนอ

9. สรุปผล (conclusion)

จากการนำเสนอการจัดกลุ่มของข้อมูลร่วมกับเทคนิคการลดมิติพบว่าผลลัพธ์ที่ได้หลังการวิเคราะห์สามารถแบ่งส่วนตลาดได้ แม้ว่าจำนวนกลุ่มในการจำแนกที่มีจำนวนน้อยอาจไม่เพียงพอต่อการนำไปพัฒนาเป็นกลยุทธ์ทางการตลาดแต่จำนวนกลุ่มนี้คือจำนวนกลุ่มที่ดีที่สุดในการจำแนกเมื่ออ้างอิง Silhouette analysis อย่างไรก็ตามการจัดกลุ่มแสดงให้เห็นว่าในขณะนี้กลุ่มเป้าหมายที่มีอยู่เป็นใครและมีพฤติกรรมแบบใด ซึ่งทั้งหมดสามารถนำไปใช้ในการการรณรงค์ทางการตลาดหรือจัดทำแคมเปญการตลาดได้และในส่วนการลดมิติของข้อมูลที่ผู้จัดทำได้ไปศึกษาเพิ่มเติมทำให้ลดความซับซ้อนได้จริง ส่งผลให้สังเคราะหผลลัพธ์ที่ต้องการได้อย่างรวดเร็วนับว่าเป็นความรู้ที่มีประโยชน์อย่างมากต่อการนำไปต่อยอดในอนาคต

ในการศึกษาครั้งต่อไปผู้จัดทำมุ่งศึกษาการเตรียมข้อมูลอย่างถูกต้องเพื่อใช้ในการเพิ่มประสิทธิภาพของโมเดลหรือการจัดกลุ่มโดยเฉพาะการวิเคราะห์ข้อมูลแบบตาราง หรือ Feature Engineering และศึกษาการลดมิติในรูปแบบอื่นนอกเหนือจาก PCA สุดท้ายในการประเมินผลโมผู้จัดทำควรเปรียบเทียบการแบ่งกลุ่มของ Silhouette analysis กับ elbow method ว่าได้ผลลัพธ์เท่ากันหรือไม่ สุดท้ายผู้จัดทำควรเพิ่มความรู้ในด้านธุรกิจให้มากขึ้นเพื่อที่จะสามารถมองผลลัพธ์ได้อย่างถูกต้องและนำผลจากการวิเคราะห์ไปประยุกต์ใช้ได้อย่างมีประสิทธิภาพ

10. เอกสารอ้างอิง (References)

พนิดา ยืนยงสวัสดิ์, และพยุง มีสัจ. (2549). การพยากรณปริมาณการใช่ยาโดยใช่โครงขายประสาทเทียม. *วรสารเทคโนโลยีสารสนเทศ*, 2(3), 1-9.

Girish Punji and David W. Stewart. (1983). Cluster analysis in marketing research. *Journal of Marketing Research*, 134-148. https://www.researchgate.net/

วันเพ็ญ ผลิศร, นีลวัสน์ ดิษฐสวรรค์, และณรงค์ศักดิ์ แสงป้อม. (2559). *การพัฒนาระบบสารสนเทศเพื่อการจัดการโครงงานและผลงานวิจัยด้วยเทคนิค Cluster Analysis กรณีศึกษาสาขาวิชาวิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ.* (รายงานการวิจัย).นนทบุรี:มหาวิทยาลัยเทคโนโลยีราชมงคลสุวรรณภูมิ.

ธรรมศักดิ์ เธียรนิเวศน์. (2548). *การลดขนาดข้อมูลด้วยน้ำหนักความหนาแน่นเพื่อการจัดกลุ่มข้อมูลขนาดใหญ่*. (ปริญญาวิศวกรรมศาสตร์มหาบัณฑิตช สาขาวิชาวิศวกรรมคอมพิวเตอร์). บัณฑิตวิทยาลัย: มหาวิทยาลัยเทคโนโลยีสุรนารี.

Vijay Gs, P. Srinivasa Pai, & N. S. Sriram. (2013). Radial basis function neural network based comparison of dimensionality reduction techniques for effective bearing diagnostics. *Journal of Engineering Tribology*, 227(6), 640–653. doi:10.1177/1350650112464927

บริษัท วันบีลีฟ จำกัด. (2560). *ทำความรู้จักกลยุทธ์ทางการตลาด สมัยใหม่ที่น่าสนใจบางอย่างคุณอาจคาดไม่ถึง*. สืบค้น 18 ธันวาคม 2564. จาก https://www.1belief.com/article/marketing-strategy/

ที่ปรึกษาการตลาดออนไลน์. (2563). *การแบ่งส่วนตลาด ผู้บริโภค มีหลักเกณฑ์อะไรบ้าง*. สืบค้น 18 ธันวาคม 2564. จาก https://www.brandingchamp.com/การแบ่งส่วนตลาด/

Weerasak Thachai. (2017). *การหาจำนวน k ที่เหมาะสมที่สุดด้วยวิธี Silhouette*. สืบค้น 18 ธันวาคม 2564. จาก https://medium.com/espressofx-notebook/การหาจำนวน-k-ที่เหมาะสมที่สุดด้วยวิธี-silhouette-b367fdae24d4

ปริญญา สงวนสัตย์. (2563). *Artificial intelligence with Machine learning, AI สร้างได้ด้วยแมชชีนเลิร์นนิ่ง*. นนทบุรี: ไอดีซี ฟรีเมียร์

Appendix

ในส่วนโค้ดภาษาไพธอน(Python notebook) และข้อมูลที่ใช้พัฒนา(Datasets)ของรายงานฉบับนี้จะถูกจัดเก็บอยู่ใน Github Repository ของผู้จัดทำ สามารถเข้าถึงได้ที่ ***https://github.com/PleumjaiOfficial/Clustering-Marketing-Campaign***