# การศึกษาวิธีการแบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อโดยใช้เทคนิคเคมีน

# A STUDY OF CUSTOMER SEGMENTATION

# BASED ON PURCHASING BEHAVIOUR USING THE K-MEANS TECHNIQUE

กมลทิพย์ มนตรีสา

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2566

# การศึกษาวิธีการแบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อโดยใช้เทคนิคเคมีน

กมลทิพย์ มนตรีสา

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2566

# A STUDY OF CUSTOMER SEGMENTATION BASED ON PURCHASING BEHAVIOUR USING THE K-MEANS TECHNIQUE

KAMONTIP MONTRISA

A Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Data Science)

Faculty of Science, Srinakharinwirot University

2023

สารนิพนธ์

เรื่อง

การศึกษาวิธีการแบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อโดยใช้เทคนิคเคมีน

ของ

กมลทิพย์ มนตรีสา

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาวิทยาการข้อมูล

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

---

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

---

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(อาจารย์ ดร.เรืองศักดิ์ ตระกูลพุทธิรักษ์)

..... ประธาน

(อาจารย์ ดร.สุทธิพงศ์ ธัชยพงษ์)

..... กรรมการ

(ผู้ช่วยศาสตราจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ)

| | |
|---|---|
| ชื่อเรื่อง | การศึกษาวิธีการแบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อโดยใช้เทคนิคเคมีน |
| ผู้วิจัย | กมลทิพย์ มนตรีสา |
| ปริญญา | วิทยาศาสตรมหาบัณฑิต |
| ปีการศึกษา | 2566 |
| อาจารย์ที่ปรึกษา | อาจารย์ ดร. เรืองศักดิ์ ตระกูลพุทธิรักษ์ |

การทำธุรกิจและการตลาดในยุคปัจจุบัน ควรคำนึงถึงความหลากหลายและความแตกต่างในพฤติกรรมของลูกค้าการเข้าใจลักษณะและพฤติกรรมการซื้อสินค้าของลูกค้าช่วยสร้างความเข้าใจในกลุ่มลูกค้าและนำข้อมูลเหล่านี้มาใช้ในการตัดสินใจทางธุรกิจและกำหนดกลยุทธ์การตลาดที่มีประสิทธิภาพ ตอบโจทย์ต่อความต้องการของลูกค้าอย่างเหมาะสม งานวิจัยนี้มีความมุ่งหมายในการศึกษาวิธีการแบ่งกลุ่มลูกค้า (Clustering) โดยพิจารณาจากพฤติกรรมการซื้อของลูกค้า ด้วยเทคนิคการเรียนรู้ของเครื่องแบบไม่มีผู้สอนโดยใช้เทคนิค K-Means และแบบจำลอง RFM (การซื้อครั้งล่าสุด, ความถี่ในการซื้อ, ยอดค่าใช้จ่ายรวม) และเพิ่มตัวแปร Basket Size (ยอดการซื้อแต่ละครั้ง) และตัวแปร Day Type (วันที่มาใช้บริการ) ผู้วิจัยเลือกศึกษาธุรกิจประเภทร้านค้าปลีก ใช้ชุดข้อมูลสาธารณะจากเว็บไซต์ www.kaggle.com จัดเก็บข้อมูลในปี ค.ศ. 2019 ถึงต้นปี ค.ศ. 2023 มีจำนวนข้อมูลทั้งหมด 29,103 รายการ มี 7 คุณลักษณะ ใช้เทคนิค K-Means ในการแบ่งกลุ่มลูกค้า กำหนดค่า K ที่เหมาะสมด้วย Elbow Method และวัดประสิทธิภาพการแบ่งกลุ่มด้วย Silhouette Score และ Davies-Bouldin Index ผลการวิเคราะห์แบ่งกลุ่มลูกค้าสามารถแบ่งกลุ่มได้ 4 กลุ่ม และผลจากการศึกษาลักษณะเฉพาะของกลุ่มนิยามได้ดังนี้ กลุ่มที่ 1 มีจำนวน 259 ราย ให้คำนิยามเป็น “กลุ่มลูกค้ามาน้อย จ่ายน้อย ช้อปวันธรรมดา” กลุ่มที่ 2 มีจำนวน 140 ราย ให้คำนิยามเป็น “กลุ่มลูกค้ามาบ่อย จ่ายหนัก ช้อปวันธรรมดา” กลุ่มที่ 3 มีจำนวน 66 ราย ให้คำนิยามเป็น “กลุ่มลูกค้ามาบ่อย จ่ายหนัก ช้อปวันหยุด” กลุ่มที่ 4 มีจำนวน 42 ราย ให้คำนิยามเป็น “กลุ่มลูกค้ามาน้อย จ่ายน้อย ช้อปวันหยุด” การวิเคราะห์คุณลักษณะนี้สามารถช่วยให้ธุรกิจเข้าใจและจัดการกับกลุ่มลูกค้าได้อย่างมีประสิทธิภาพและเหมาะสม เช่น การพัฒนาและการนำเสนอโปรโมชั่นหรือแนวทางการตลาดที่เหมาะสมสำหรับกลุ่มลูกค้าที่มีลักษณะและพฤติกรรมทางการซื้อเหล่านี้ และสามารถใช้ข้อมูลเพิ่มเติมเพื่อวิเคราะห์พฤติกรรมการซื้อของลูกค้าได้ละเอียดมากขึ้น

คำสำคัญ : การแบ่งกลุ่มลูกค้า, พฤติกรรมการซื้อ, การจัดกลุ่มแบบเคมีน, แบบจำลองอาร์เอฟเอ็ม

| | |
|---|---|
| Title | A STUDY OF CUSTOMER SEGMENTATION BASED ON PURCHASING BEHAVIOUR USING THE K-MEANS TECHNIQUE |
| Author | KAMONTIP MONTRISA |
| Degree | MASTER OF SCIENCE |
| Academic Year | 2023 |
| Thesis Advisor | Lecturer Ruangsak Trakunphutthirak , Ph.D. |

In contemporary business and marketing, understanding the diversity and differences in customer behaviors is crucial. Analyzing and comprehending customers' purchasing patterns helps grasp customer groups and utilize this data. This research aims to study customer segmentation using clustering techniques, considering customers' purchasing behaviors and employing K-Means and RFM models (Recency, Frequency, Monetary), along with additional variables, like Basket Size and Day Type. The study focused on retail businesses, utilizing a dataset from www.kaggle.com from 2019-2023, comprising 29,103 entries with seven features. K-Means clustering is employed to determine the appropriate K value using the Elbow Method and evaluate the clustering performance using the Silhouette Score and Davies-Bouldin Index. The analysis results in four customer groups: Group One included 259 customers, defined as Low-Volume, Low-Spending, Regular Shopping on Weekdays; Group Two includes 140 customers, defined as Frequent Shoppers, Heavy Spenders, Shopping on Weekdays; Group Three includes 66 customers defined as Frequent Shoppers, Heavy Spenders, Shopping on Weekends and Group Four includes 42 customers defined as Low-Volume, Low-Spending, Shopping on Weekends. The analysis of these characteristics can aid businesses in managing and catering to customer groups. For instance, developing and presenting promotions or marketing strategies tailored to the specific characteristics and purchasing behaviors of these customer groups. The further data analysis can provide more detailed insights into customer purchasing behaviors, enabling businesses to better understand and serve their customers.

Keyword : Customer Segmentation, Purchasing Behavior, K-Means Clustering, RFM Model

# กิตติกรรมประกาศ

รายงานสารนิพนธ์ฉบับนี้ ได้รับความช่วยเหลือและความเอาใจใส่เป็นอย่างดียิ่งจากอาจารย์ที่ปรึกษาหลัก อ.ดร. เรืองศักดิ์ ตระกูลพุทธิรักษ์ ที่ชี้แนวทางในสิ่งที่เป็นประโยชน์ต่อการศึกษาและการทำสารนิพนธ์ ตลอดจนช่วยแนะนำและแก้ไขข้อบกพร่องต่างๆ ในการทำงานวิจัยครั้งนี้ ขอกราบขอบพระคุณอย่างสูงมา ณ ที่นี้

ขอขอบคุณคณาจารย์ทุกท่านที่ให้ความรู้ตามแผนการเรียนรู้ในหลักสูตรของสาขาวิทยาการข้อมูล คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ และให้คำแนะนำต่างๆ ในการทำงานที่เกี่ยวข้องกับงานวิจัยครั้งนี้

ขอขอบคุณเพื่อนๆ และทุกท่านที่คอยสนับสนุน ให้ความช่วยเหลือและให้กำลังใจผู้วิจัยตลอดมา หวังเป็นอย่างยิ่งว่างานวิจัยฉบับนี้จะเป็นประโยชน์แก่ผู้ที่เกี่ยวข้องและสนใจต่อไปไม่มากก็น้อย หากมีข้อผิดพลาดประการใด ผู้วิจัยขอน้อมรับและขออภัยมา ณ ที่นี้ด้วย

กมลทิพย์ มนตรีสา

# สารบัญ

หน้า

# สารบัญตาราง

หน้า

# สารบัญรูปภาพ

หน้า

# บทที่ 1
# บทนำ

## 1.1 ที่มาและความสำคัญของงานวิจัย

การดำเนินธุรกิจและการตลาดในปัจจุบัน ควรที่จะคำนึงถึงพฤติกรรมของลูกค้าที่มีความแตกต่างกันและมีความหลากหลายของแต่ละบุคคล การพิจารณารับฟังปัญหา ข้อเสนอแนะและตอบสนองต่อปัญหาและความต้องการของลูกค้านั้นเป็นปัจจัยสำคัญในการทำธุรกิจ หากเจ้าของธุรกิจต่าง ๆ ไม่สามารถทำความเข้าใจเกี่ยวกับพฤติกรรมที่แตกต่างกันของลูกค้าอาจทำให้ลูกค้าหันไปใช้บริการกับคู่แข่งทางธุรกิจเพิ่มสูงขึ้น ในสถานการณ์ปัจจุบันลูกค้าเป็นบุคคลสำคัญที่เป็นตัวแปรกำหนดทิศทางการใช้สินค้าและบริการ ลูกค้าจะเลือกเฉพาะสินค้าหรือบริการที่ทำให้ตนเองพอใจ หากว่าธุรกิจไหนมีบริการที่ทำให้ลูกค้าพึงพอใจ ลูกค้าถึงจะเลือกซื้อหรือใช้บริการกับธุรกิจนั้น ถ้าหากว่าตัวสินค้าหรือบริการไหนที่ลูกค้าใช้อยู่มีปัญหาหรือสร้างประสบการณ์ที่ไม่ดี ตอบสนองความต้องการลูกค้าไม่ได้ ก็มีสิทธิ์ที่จะเปลี่ยนใจไปใช้สินค้าหรือบริการจากผู้ให้บริการรายอื่นที่มีสินค้าหรือบริการใกล้เคียงกัน

การแบ่งลูกค้าออกเป็นกลุ่ม นับว่าเป็นกระบวนการสำคัญที่จะวิเคราะห์กลุ่มลูกค้าที่มีพฤติกรรมการซื้อสินค้าที่คล้ายคลึงกัน ข้อมูลที่เกี่ยวข้องหรือเป็นตัวแทนของลักษณะการซื้อสินค้าของลูกค้ามีความสำคัญในกระบวนการนี้ ทั้งนี้การเข้าใจลักษณะและพฤติกรรมการซื้อสินค้าของลูกค้ายังช่วยสร้างความเข้าใจในกลุ่มลูกค้า และนำข้อมูลนี้มาใช้ในการตัดสินใจทางธุรกิจและกำหนดกลยุทธ์การตลาดที่มีประสิทธิผล ตอบโจทย์ได้ตรงความต้องการของลูกค้าอย่างเหมาะสม สร้างความเชื่อมั่นและพึงพอใจให้กับลูกค้าอย่างยั่งยืน นอกจากนี้การแบ่งกลุ่มยังช่วยในการสื่อสารกับกลุ่มลูกค้าผ่านช่องทางหรือแพลตฟอร์มที่เหมาะสมและช่วยในการค้นพบโอกาสใหม่ ๆ

จากที่กล่าวมา ผู้วิจัยจึงสนใจที่จะศึกษาวิธีการและขั้นตอนแบ่งกลุ่มลูกค้า เพื่อทำความเข้าใจถึงพฤติกรรมของลูกค้า ได้เลือกทำการศึกษาวิธีการแบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อสินค้า โดยใช้ชุดข้อมูลสาธารณะจากเว็บไซต์ www.kaggle.com เกี่ยวกับรายการขายปลีกทั้งการขายจากสาขาและผู้ค้าปลีก จัดเก็บข้อมูลในปี 2019 ถึงต้นปี 2023 มีข้อมูล 29,103 รายการ มี 7 คุณลักษณะ เพื่อที่จะสามารถนำวิธีการนี้ไปใช้ประโยชน์ทางด้านธุรกิจและวางแผนทำการตลาดได้

## 1.2 วัตถุประสงค์ของงานวิจัย

1.เพื่อศึกษากระบวนการและวิธีแบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อ โดยใช้เทคนิค K-Means ด้วยการใช้แบบจำลอง RFM และเพิ่มตัวแปร Basket Size และตัวแปร Day Type

2. เพื่อศึกษาวิธีการเลือกค่าจำนวนกลุ่มที่เหมาะสมในการแบ่งกลุ่มด้วย Elbow Method

3. เพื่อศึกษาวิธีการประเมินผลลัพธ์การแบ่งกลุ่มที่ได้จากการทำ K-Means clustering เพื่อประเมินว่าจำนวนกลุ่มที่คำนวณขึ้นมาเหมาะสมหรือไม่ ด้วยวิธี Silhouette Score และ Davies-Bouldin Index

## 1.3 ขอบเขตของงานวิจัย

**ประชากร**

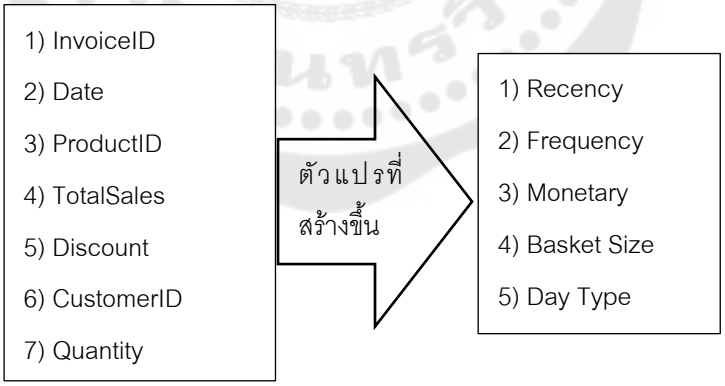
ข้อมูลรายการขายปลีก ที่เป็นชุดข้อมูลสาธารณะ ชื่อว่า Retail Data Set ที่มีการจัดเก็บข้อมูลในปี 2019 ถึงต้นปี 2023

**กลุ่มตัวอย่าง**

ข้อมูลรายการขายปลีก จัดเก็บข้อมูลในปี 2019 ถึงต้นปี 2023 มีข้อมูล 29,103 รายการ 7 คุณลักษณะ

**ตัวแปรที่ศึกษา**

ตัวแปรอิสระ แบ่งเป็นดังนี้




ตัวแปรตาม ได้แก่ จำนวนกลุ่มของลูกค้าร้านขายปลีกที่แบ่งได้ตามพฤติกรรมการซื้อสินค้า ที่ถูกสร้างโดย K-Means

**1.4 กรอบแนวคิดของงานวิจัย**

ศึกษาขั้นตอนวิธีการแบ่งกลุ่มลูกค้าเพื่อทำความเข้าใจถึงพฤติกรรมของลูกค้า โดยใช้เทคนิค Unsupervised Learning การเรียนรู้ของเครื่องแบบไม่มีผู้สอน ประเภท Clustering ทำการแบ่งลูกค้าออกเป็นกลุ่มตามพฤติกรรมการซื้อสินค้าของร้านขายปลีกแห่งหนึ่งที่เป็นชุดข้อมูลสาธารณะเกี่ยวกับรายการขายปลีกทั้งการขายจากสาขาและผู้ค้าปลีก จัดเก็บข้อมูลในปี 2019 ถึงต้นปี 2023 มีจำนวนข้อมูลทั้งหมด 29,103 รายการ มี 7 คุณลักษณะ ทำการแปลงข้อมูลให้มีความเหมาะสมสำหรับการนำไปใช้ในการวิเคราะห์และแบ่งกลุ่ม โดยการใช้เทคนิคการสร้างฟีเจอร์ (Feature Engineering) ด้วยแบบจำลอง RFM และเพิ่มตัวแปรยอดการซื้อสินค้าแต่ละครั้ง วันที่มาใช้บริการ ใช้เทคนิค K-Means ในการทำ Clustering เพื่อแบ่งกลุ่มลูกค้า กำหนดค่า K (จำนวนกลุ่มที่ต้องการ) โดยใช้เทคนิคการหา K ที่เหมาะสม เช่น Elbow Method และประเมินประสิทธิภาพว่าการจัดกลุ่มที่คำนวณขึ้นมาเหมาะสมหรือไม่ ด้วย Silhouette Score และ Davies-Bouldin Index จากนั้นวิเคราะห์ผลลัพธ์ที่ได้ สรุป และอภิปรายผล

ในบทที่ 1 นี้ กล่าวถึงภาพรวมทั้งหมดของงานวิจัยที่ศึกษา ประกอบด้วย ที่มาและความสำคัญ วัตถุประสงค์ ขอบเขต และกรอบแนวคิดของงานวิจัย โดยจะทำการทบทวนวรรณกรรมที่เกี่ยวข้อง ในบทที่ 2 ต่อไป

# บทที่ 2
# วรรณกรรมและงานวิจัยที่เกี่ยวข้อง

ในงานวิจัยนี้ ผู้วิจัยได้ทำการศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง และได้นำเสนอดังต่อไปนี้

1. แนวความคิดของพฤติกรรมผู้บริโภค
2. การแบ่งกลุ่มลูกค้า (Customer Segmentation)
3. การจัดกลุ่ม (Clustering)
4. การทำเหมืองข้อมูล (Data Mining)
5. RFM Model
6. การประเมินประสิทธิภาพการแบ่งกลุ่ม (Evaluation of Clustering)
7. งานวิจัยที่เกี่ยวข้อง

## 2.1 แนวความคิดของพฤติกรรมผู้บริโภค

พฤติกรรมผู้บริโภค คือ กิจกรรมที่เกี่ยวข้องกับการซื้อของและแสดงออกของบุคคลที่ถูกกระตุ้นให้เกิดพฤติกรรมตอบสนอง มี 2 ประเภท คือ

1. พฤติกรรมภายใน (Convert behavior) เป็นพฤติกรรมที่ไม่แสดงออกมาภายนอก แต่จะแสดงผลทางความรู้สึกและความคิดของบุคคล เช่น ความพึงพอใจในส่วนประกอบของสินค้าหรือบริการ การประเมินและวิเคราะห์ข้อมูลเพื่อตัดสินใจ

2. พฤติกรรมภายนอก (Overt behavior) เป็นพฤติกรรมที่แสดงออกภายนอกต่อสังคม ได้แก่ การเลือกซื้อสินค้า การพิจารณาสินค้า การพูดคุยเกี่ยวกับสินค้า การสร้างและแสดงออกเกี่ยวกับการเลือกซื้อและการใช้สินค้า

ลักษณะพฤติกรรมของผู้บริโภคว่ามี 7 ประการ ดังนี้

1. พฤติกรรมผู้บริโภคเป็นการกระทำที่จูงใจโดยมุ่งหวังให้บรรลุวัตถุประสงค์ต่าง ๆ ผู้บริโภคทำกิจกรรมเพื่อความต้องการของตน

2. พฤติกรรมผู้บริโภคมีหลายขั้นตอน เช่น วางแผนการซื้อ ดูโฆษณา ถามคำแนะนำ ตัดสินใจซื้อ ใช้สินค้า บำรุงรักษา และมีผลพลอยได้

3. พฤติกรรมผู้บริโภคเกิดขึ้นจากกระบวนการทางความนึกคิดและอารมณ์ เช่น การตัดสินใจในการซื้อสินค้า

4. พฤติกรรมผู้บริโภคมีความซับซ้อนต่อการตัดสินใจ เช่น "จังหวะเวลา" และ "ความสลับซับซ้อน" ที่มีบทบาทในกระบวนการตัดสินใจ

5. พฤติกรรมผู้บริโภคเกี่ยวข้องกับหลายบทบาท เช่น ผู้ซื้อ ผู้ใช้ ผู้ตัดสินใจ และผู้ที่มีอิทธิพลต่อการตัดสินใจ

6. พฤติกรรมผู้บริโภคมักได้รับอิทธิพลจากปัจจัยภายนอก เช่น จิตวิทยา สังคมวิทยา และเศรษฐศาสตร์

7. พฤติกรรมผู้บริโภคแตกต่างกันตามบุคคล ด้วยบุคลิกภาพ สถานภาพ รสนิยม วิสัยทัศน์ และกลุ่มเป้าหมายที่แตกต่างกัน

การศึกษาพฤติกรรมของผู้บริโภคเป็นสิ่งสำคัญที่ช่วยให้เราเข้าใจว่าผู้บริโภคมีพฤติกรรมและทัศนคติอย่างไรต่อการตัดสินใจในการซื้อสินค้าและการใช้บริการ ประโยชน์หลัก ๆ มีดังนี้

1. ประโยชน์ต่อผู้บริโภค การเข้าใจพฤติกรรมช่วยให้ผู้บริโภคมีชีวิตที่มีคุณภาพมากขึ้น เนื่องจากพวกเขาสามารถทำการตัดสินใจเลือกซื้อสินค้าและบริการที่เหมาะสมกับความต้องการและความพึงพอใจของตนเองได้ดีขึ้น

2. ประโยชน์ต่อผู้ผลิต ผู้ประกอบการ หรือผู้ให้บริการ การทราบถึงแนวโน้มและความพึงพอใจของผู้บริโภคช่วยในการวางแผนการตลาดและพัฒนาผลิตภัณฑ์หรือบริการที่ตอบสนองต่อความต้องการของตลาดได้อย่างเหมาะสม

3. ประโยชน์ต่อเศรษฐกิจและสังคม สามารถที่จะใช้เป็นแนวทางปรับนโยบายสาธารณะและวางแผนการประกอบธุรกิจ เป็นข้อมูลที่ใช้ในการควบคุมพฤติกรรมเพื่อป้องกันปัญหาสังคม (สิรินี ว่องวิไลรัตน์, 2560)

## 2.2 **การแบ่งกลุ่มลูกค้า** (Customer Segmentation)

ความหมายของ Customer Segmentation คือกระบวนการแบ่งกลุ่มลูกค้าเป็นกลุ่มย่อยที่มีคุณลักษณะคล้ายคลึงกัน เพื่อใช้ในการระบุความต้องการของกลุ่มลูกค้า ที่ยังไม่ได้รับการบริการ หรือการตอบสนอง โดยการใช้ข้อมูลเหล่านี้เป็นเครื่องมือที่มีความสำคัญในการปรับปรุงสินค้าและบริการให้สนองต่อความต้องการของเฉพาะกลุ่มลูกค้าได้ โดยวิธีการ Customer Segmentation ที่มีความนิยมประกอบไปด้วย

1. การใช้สถิติประชากร (Demographic) เช่น เพศ (gender) อายุ (age) สถานะสมรส (marital status) รายได้ (income) การศึกษา (education) อาชีพ (occupation) เพื่อแบ่งกลุ่มตามลักษณะทางสังคม

2. การวิเคราะห์ข้อมูลภูมิศาสตร์ (Geographical Data) เพื่อแบ่งกลุ่มตามพื้นที่หรือทวีปที่แตกต่างกันตามความต้องการของธุรกิจ

3. การใช้ลักษณะจิตวิทยา (Psychographics) เพื่อแบ่งกลุ่มตามฐานะทางสังคม (social class) วิถีทางการดำเนินชีวิต (lifestyle) และลักษณะบุคลิกภาพ (personality traits) เพื่อเข้าใจเพิ่มเติมเกี่ยวกับพฤติกรรมและความคิดเห็นของลูกค้า

4. การใช้ข้อมูลพฤติกรรม (Behavioral data) เพื่อแบ่งกลุ่มตามพฤติกรรมการซื้อขาย เช่น นิสัยการใช้จ่าย (spending habits) พฤติกรรมการใช้งานสินค้าและบริการ (usage behavior) และประโยชน์ที่ต้องการ (desired benefits) เพื่อให้ผลิตภัณฑ์เหมาะสมกับกลุ่มลูกค้าแต่ละกลุ่ม

การทำ Customer Segmentation ช่วยให้ธุรกิจสามารถเข้าใจลูกค้าได้ลึกซึ้งและปรับเปลี่ยนกลยุทธ์การตลาดเพื่อตอบสนองต่อความต้องการของลูกค้าได้อย่างมีประสิทธิภาพและเหมาะสม (วิทยา พรพัชรพงศ์, 2549)

**2.3 การจัดกลุ่ม (Clustering)**

การวิเคราะห์จัดกลุ่มข้อมูล หรือ Cluster Analysis เป็นหนึ่งในเทคนิคของการเรียนรู้แบบไม่มีผู้สอน (Unsupervised Learning) ที่นำมาใช้ในการแบ่งข้อมูลออกเป็นกลุ่มตามลักษณะที่คล้ายคลึงกัน โดยไม่ต้องมีการให้ข้อมูลตัวอย่างล่วงหน้า โดยจะจัดกลุ่มข้อมูลตามความคล้ายคลึงในลักษณะต่างๆ ซึ่งจะทำให้ข้อมูลที่คล้ายกันมีโอกาสถูกจัดเข้ากลุ่มเดียวกัน ในขณะที่ข้อมูลที่แตกต่างกันจะถูกจัดให้อยู่ในกลุ่มที่แตกต่างกัน เทคนิคนี้ไม่ได้หาผลลัพธ์ที่ต้องการวัดค่าความแม่นยำ เป็นเพียงการแบ่งกลุ่มข้อมูลตามลักษณะคล้ายกัน การใช้เทคนิคนี้มีจุดประสงค์เพื่อค้นหาความสัมพันธ์หรือโครงสร้างในข้อมูลที่อาจจะไม่เป็นที่รู้จักและนำข้อมูลเหล่านี้มาเรียนรู้เพื่อทำความเข้าใจและค้นหาความสัมพันธ์ที่เป็นประโยชน์ในการประยุกต์ใช้งานต่อไป

การสร้างโมเดลแบบไม่มีผู้สอนมีความแตกต่างจากการสร้างโมเดลแบบมีผู้สอนตรงที่ไม่ต้องการแบ่งชุดข้อมูลออกเป็นชุดสำหรับการฝึกสอนและทดสอบ เนื่องจากข้อมูลไม่มีคลาสคำตอบที่จะนำมาฝึกสอน ดังนั้น เราสามารถใช้ข้อมูลทั้งหมดในขั้นตอนการสร้างโมเดลได้เลย โดยมีขั้นตอนสำหรับการสร้างโมเดลแบบไม่มีผู้สอนทั้งหมด 4 ขั้นตอน ดังนี้

1. กำหนดวิธีวัดความเหมือนหรือความต่างของข้อมูล เช่น Euclidean Distance Cosine Similarity และ Manhattan Distance เป็นต้น โดยแต่ละวิธีจะมีลักษณะและการคำนวณที่แตกต่างกัน ในการกำหนดวิธีการวัดเพื่อบ่งชี้ความคล้ายคลึงระหว่างข้อมูล ได้อธิบายวิธีการ

ยูคลิเดียน (Euclidean Distance) เป็นวิธีการที่พบบ่อยในการวัดระยะห่างระหว่างจุดข้อมูล โดยสมการดังนี้

$$D = \sqrt{\sum_{i=1}^{d}(x_1^i - x_2^i)^2} \quad (1)$$

โดยที่ D คือ ระยะห่างระหว่างข้อมูลที่ 1 และ 2 ซึ่งถ้าค่าเท่ากับ 0 แสดงว่าข้อมูลมีความเหมือนกัน

$x_1^i$ คือ ค่าของแอทริบิวต์ที่ i ของข้อมูลที่ 1

$x_2^i$ คือ ค่าของแอทริบิวต์ที่ i ของข้อมูลที่ 2

2. การเลือกอัลกอริทึมที่ใช้ในการจัดกลุ่มข้อมูลมีการแบ่งออกเป็น 2 ประเภทหลัก คือการแบ่งกลุ่มอย่างชัดเจน (Hard clustering) และการแบ่งกลุ่มแบบไม่ชัดเจน (Soft clustering)

3. การกำหนดจำนวนกลุ่มที่ต้องการจะขึ้นอยู่กับลักษณะของข้อมูลและวัตถุประสงค์ของการวิเคราะห์ ในกรณีที่มีการแบ่งกลุ่มอย่างชัดเจนจำเป็นต้องระบุจำนวนกลุ่ม โดยอาจใช้เทคนิคเช่น Elbow method หรือ Silhouette method เพื่อช่วยในการกำหนดจำนวนกลุ่มที่เหมาะสม สำหรับการแบ่งกลุ่มแบบไม่ชัดเจน ไม่จำเป็นต้องกำหนดจำนวนกลุ่มล่วงหน้า เนื่องจากอัลกอริทึมจะหาจำนวนกลุ่มที่เหมาะสมตามข้อมูลที่กำลังทำการวิเคราะห์

4. ในการประเมินผลในกรณีของการแบ่งกลุ่มแบบไม่มีผู้สอน เราสามารถใช้เครื่องมือหรือวิธีการที่ไม่ใช้การวัดค่าแม่นยำ เช่น การวัดความพึงพอใจในแบบจำลองที่ได้ หรือการใช้การวิเคราะห์เพื่อปรับปรุงแบบจำลองในการหาค่าที่เหมาะสม (Optimization Model) เช่น การใช้แบบจำลองการจัดกลุ่มในการประยุกต์ใช้ในการจัดส่งสินค้าเพื่อลดค่าใช้จ่ายในการขนส่ง การประเมินผลด้วยวิธีการเช่นนี้จะช่วยให้เราทราบถึงความเหมาะสมและประสิทธิภาพของการจัดกลุ่มข้อมูลที่ได้จากอัลกอริทึมที่เราใช้ (ไกรศักดิ์ เกษร, 2564, pp. น. 286-287)

K-means เป็นหนึ่งในอัลกอริทึมที่ได้รับความนิยมมากในการแบ่งกลุ่มข้อมูล (Clustering Algorithm) เนื่องจากมีความเรียบง่ายและมีประสิทธิภาพในการใช้งาน โดยในขั้นตอนการทำงานของ K-means จะทำการกำหนดจำนวน ของกลุ่มที่ต้องการ (K) และจุดเริ่มต้นของกลุ่ม (Centroids) จากนั้นจะทำการกำหนดข้อมูลให้เข้ากับกลุ่มที่ใกล้ที่สุด ส่วนระยะห่างที่ใช้คำนวณสามารถปรับเปลี่ยนได้ตามความเหมาะสมของงาน โดยทั่วไปจะเป็น

Euclidean distance ด้วยสมการที่ให้ค่าระหว่างข้อมูล และ centroid ถ้าค่าเป็น 0 แสดงว่าข้อมูลมีความคล้ายคลึงกัน โดยระยะห่างนั้นสามารถคำนวณโดยใช้ Euclidean distance (ภคภูมิ สารพัฒน์, 2563)

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

โดยที่ x และ y เป็นข้อมูลแต่ละแถวที่มีความยาว n หรือมี n ค่า ซึ่งสามารถเข้าถึงแต่ละค่าได้จาก index I ตัวอย่างเช่น การคำนวณ Euclidean distance สำหรับข้อมูลที่มีความยาว
n = 2 ของ x = [-1,3] และ y = [2,7] ทำได้ดังนี้

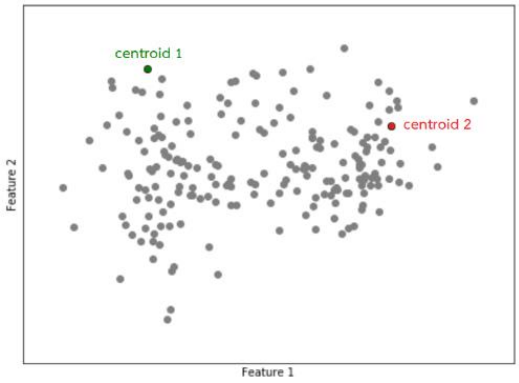
$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (3)$$

$$= \sqrt{(-1-2)^2 + (3-7)^2} \quad (4)$$

$$= 5 \quad (5)$$

เมื่อจัดกลุ่มครบถ้วนในทุกแถวแล้วจะทำการการปรับตำแหน่งของ centroid แต่ละกลุ่มใหม่ และจะทำวนซ้ำไปเรื่อย ๆ จนกว่าจะตรงตามเงื่อนไข สรุปเป็นขั้นตอนได้ดังนี้

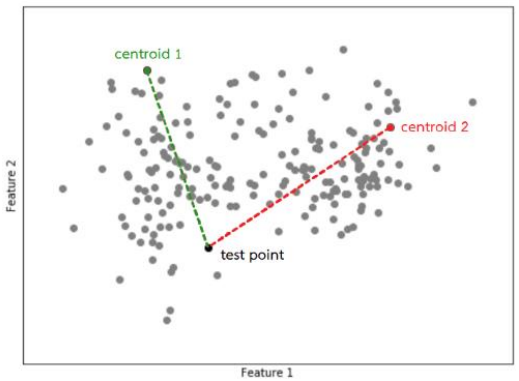
1. เลือกจุดเริ่มต้นสำหรับ centroid ของแต่ละกลุ่ม ซึ่งสามารถเลือกโดยสุ่มหรือใช้วิธีการเลือกที่มีประสิทธิภาพเพื่อให้สอดคล้องกับโครงสร้างข้อมูล ดังภาพประกอบ 1




ภาพประกอบ 1 แสดงการสุ่มค่า k เพื่อใช้เป็น centroid

ที่มา : https://bigdata.go.th/big-data-101/k-means-algorithm-for-clustering-large-data-sets-with-categorical-values/

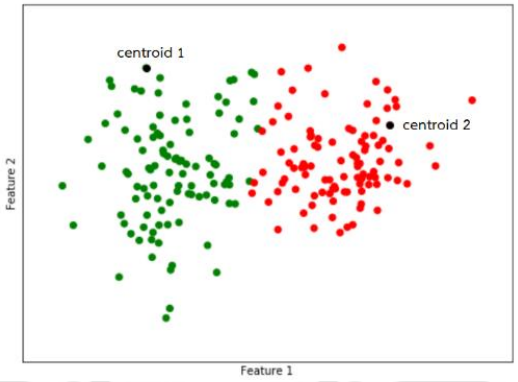
2. กำหนดข้อมูลให้เข้ากลุ่มโดยใช้ centroid ที่กำหนดไว้ โดยการคำนวณระยะห่างระหว่างข้อมูลกับ centroid และกำหนดให้ข้อมูลอยู่ในกลุ่มที่มี centroid ที่ใกล้ที่สุด ดังภาพประกอบ 2




ภาพประกอบ 2 แสดงระยะห่างระหว่างข้อมูลแต่ละแถวกับ centroid

ที่มา : https://bigdata.go.th/big-data-101/k-means-algorithm-for-clustering-large-data-sets-with-categorical-values/

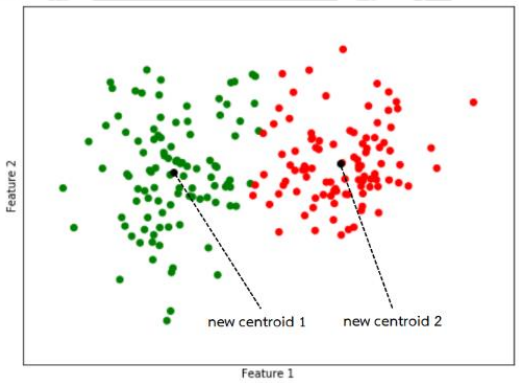
3. คำนวณตำแหน่งใหม่ของ centroid ในแต่ละกลุ่มโดยใช้ข้อมูลที่อยู่ในกลุ่มนั้น ๆ โดยหาค่าเฉลี่ยของตำแหน่งของข้อมูลทั้งหมดในกลุ่มและกำหนดตำแหน่งใหม่ของ centroid ให้เป็นค่าเฉลี่ยนั้นดังภาพประกอบ 3




ภาพประกอบ 3 แสดงการจัดกลุ่มของข้อมูลกลุ่มเดียวกัน

ที่มา : https://bigdata.go.th/big-data-101/k-means-algorithm-for-clustering-large-data-sets-with-categorical-values/

4. ปรับตำแหน่งของ centroid ในแต่ละกลุ่มไปเป็นค่าเฉลี่ยของข้อมูลทั้งหมดที่อยู่ในกลุ่มนั้น ดังภาพประกอบ 4




ภาพประกอบ 4 แสดงการปรับตำแหน่งของ centroid

ที่มา : https://bigdata.go.th/big-data-101/k-means-algorithm-for-clustering-large-data-sets-with-categorical-values/

5. ทำซ้ำขั้นตอนที่ 2 และ 4 ไปจนกว่าจะไม่มีการเปลี่ยนแปลงตำแหน่งของ centroid หรือมีการเปลี่ยนแปลงเพียงเล็กน้อยเท่านั้น

สรุป K-means Clustering เป็นเทคนิคที่มีการใช้งานและปรับใช้กันอย่างแพร่หลายในการแบ่งกลุ่มข้อมูลเนื่องจากมีความเร็วและความง่ายในการทำงาน และเป็นเทคนิคที่มีประสิทธิภาพในการจัดกลุ่มข้อมูลที่มีการกระจายแบบกลุ่มในรูปแบบที่ชัดเจน เป็นวิธีการแบ่งกลุ่มที่ทำงานโดยการกำหนดจุดเริ่มต้นศูนย์กลางและปรับปรุง centroid ในแต่ละรอบเพื่อให้ข้อมูลในกลุ่มมีระยะห่างน้อยที่สุดในแต่ละกลุ่ม (ภคภูมิ สารพัฒน์, 2020) เป็นเทคนิคการจัดกลุ่มแบบไม่เป็นขั้นตอนหรือ Nonhierarchical Cluster Analysis หรือการแบ่งส่วน (Partioning) เป็นวิธีการเรียนรู้โดยไม่มีผู้สอนที่แก้ปัญหาการจัดกลุ่มที่รู้จักกันทั่วไป (Sirilak Ketchaya, ม.ป.ป.)

ข้อดีของการทำ K-Means Clustering

1. ความเร็วในการประมวลผล เมื่อมีจำนวนข้อมูลมากและจำนวนกลุ่มน้อย การหาค่าเฉลี่ยแบบ K-means สามารถคำนวณได้เร็วกว่าการจัดกลุ่มแบบอื่น

2. การจัดกลุ่มที่เป็นรูปวงกลม ขั้นตอนการหาค่าเฉลี่ยแบบ K-means มักจะสร้างกลุ่มที่มีสมาชิกภายในกลุ่มหนาแน่นกว่าการจัดกลุ่มแบบ Hierarchical Clustering โดยเฉพาะเมื่อกลุ่มเป็นรูปวงกลม

ข้อด้อยของการทำ K-Means Clustering

1. การเลือกจำนวนกลุ่มที่เหมาะสม การหาค่า K ที่เหมาะสมสามารถคาดเดาได้ยาก เนื่องจากไม่มีวิธีที่ชัดเจนในการเลือกค่า K

2. ปัญหาการทำงานบนข้อมูลที่มีรูปร่างที่ซับซ้อน การทำงานของ K-Means อาจทำงานได้ไม่ดีถ้ากลุ่มข้อมูลไม่เป็นรูปวงกลม หรือมีความหนาแน่นและขนาดที่แตกต่างกันในแต่ละกลุ่ม

3. ข้อจำกัดในเรื่องขนาด ความหนาแน่น และรูปร่าง K-Means มีข้อจำกัดในการจัดกลุ่มข้อมูลที่มีขนาด ความหนาแน่น และรูปร่างที่แตกต่างกันในแต่ละกลุ่ม ซึ่งอาจทำให้ผลลัพธ์ไม่แม่นยำถ้าข้อมูลมีคุณสมบัติเหล่านี้

**2.4 การทำเหมืองข้อมูล** (Data Mining)

กระบวนการการทำเหมืองข้อมูล (data mining) มุ่งเน้นการวิเคราะห์และการค้นหาในข้อมูลสารสนเทศที่มีมากมายมหาศาล เพื่อสร้างความรู้ใหม่ ๆ โดยใช้เทคนิคที่ซับซ้อนกว่าการวิเคราะห์ทางสถิติหรือการสืบค้นแบบ SQL โดยมีเครื่องมือและเทคนิคต่าง ๆ ที่มีอยู่ เช่น ต้นไม้การตัดสินใจ (decision tree) การจัดกลุ่ม (clustering) การจัดความสัมพันธ์ (association) ข้อมูลอนุกรมเวลา (time series) การวิเคราะห์ลำดับการเกิดข้อมูล (sequence analysis) และการวิเคราะห์การเบี่ยงเบนของข้อมูล (deviation analysis) เทคนิคเหมืองข้อมูลสามารถประยุกต์ใช้ได้ในหลายแขนง เช่น การแพทย์ การทหาร และการดำเนินธุรกิจ เพื่อให้ได้ข้อมูลและแนวทางการปฏิบัติที่มีคุณค่าสูงในแต่ละสาขา ยกตัวอย่างเช่น สำหรับธุรกิจค้าปลีกสามารถใช้เหมืองข้อมูลในการวิเคราะห์และกำหนดกลยุทธ์สินค้าที่วางบนชั้นวางเพื่อเพิ่มยอดขาย สร้างกลุ่มลูกค้าและวิเคราะห์พฤติกรรมของลูกค้าเพื่อเสนอสินค้าที่เข้ากับกลุ่มเป้าหมายของแต่ละกลุ่ม (สุจิรา ไชยกุสินธุ์, 2559)

กระบวนการ CRISP-DM (Cross-Industry Standard Process for Data Mining) เป็นมาตรฐานสำหรับการทำเหมืองข้อมูลในหลายอุตสาหกรรม ที่มีแนวคิดในการจัดการข้อมูลที่มีมากๆ เพื่อสร้างความรู้ และเข้าใจแนวโน้มและความสัมพันธ์ที่อาจซ่อนอยู่ในข้อมูล เพื่อให้เกิดประโยชน์แก่ธุรกิจและองค์กรที่เป็นผู้ใช้งาน มีขั้นตอนทั้งหมด 6 ขั้นตอน ดังนี้

1. การเข้าใจธุรกิจ (Business Understanding) เน้นการเข้าใจธุรกิจและระบุโอกาสหรือปัญหาที่ต้องการแก้ไข กำหนดขอบเขตข้อมูลและผลลัพธ์ที่คาดหวังจากการทำเหมืองข้อมูล

2. การเข้าใจข้อมูล (Data Understanding) ทำความเข้าใจข้อมูลโดยรวบรวมข้อมูลที่เกี่ยวข้อง และคัดเลือกข้อมูลที่ถูกต้องและสำคัญสำหรับการวิเคราะห์

3. การเตรียมข้อมูล (Data Preparation) แปลงข้อมูลให้เป็นรูปแบบที่เหมาะสมสำหรับการวิเคราะห์ เช่น การคัดเลือกและแปลงรูปแบบข้อมูล

4. การสร้างแบบจำลอง (Modeling) สร้างแบบจำลองเพื่อวิเคราะห์ข้อมูลและทดสอบแบบจำลองเพื่อเลือกแบบจำลองที่ดีที่สุด

5. การประเมินผล (Evaluation) ประเมินผลลัพธ์แบบจำลองว่าตรงกับวัตถุประสงค์หรือไม่ และปรับปรุงแบบจำลองที่จำเป็น

6. การนำไปใช้ (Deployment) นำผลลัพธ์จากการวิเคราะห์มาประยุกต์ใช้ในองค์กรและติดตามผลลัพธ์เพื่อปรับปรุงต่อไป

การทำเหมืองข้อมูลในงานระบบทางธุรกิจเป็นกระบวนการที่เน้นการจัดการข้อมูลที่มีจำนวนมากในหลากหลายรูปแบบ โดยมุ่งเน้นการคัดเลือกข้อมูลที่สำคัญและจำเป็นต้องนำมาใช้งาน และทำการกำหนดรูปแบบการแบ่งกลุ่มและลำดับความสำคัญของข้อมูลก่อนจะเริ่มค้นหารูปแบบและแนวทางการประยุกต์ใช้ข้อมูลนั้นในธุรกิจ ขั้นตอนแต่ละขั้นจะเกี่ยวข้องกันและใช้ผลลัพธ์จากขั้นตอนก่อนหน้าเป็นข้อมูลนำเข้าในขั้นตอนต่อไปผ่านกระบวนการนี้ การทำเหมืองข้อมูลช่วยเปลี่ยนข้อมูลดิบ (raw data) เป็นสารสนเทศที่มีประโยชน์และสามารถนำไปใช้งานได้อย่างมีประสิทธิภาพ การระบุแหล่งข้อมูลที่ถูกต้องเป็นสิ่งสำคัญต่อผลลัพธ์ที่ได้ จากการวิเคราะห์เนื่องจากข้อมูลที่ถูกต้องและครบถ้วน จะช่วยให้การวิเคราะห์มีความเชื่อถือได้และสามารถสร้างความเข้าใจที่ถูกต้องเกี่ยวกับธุรกิจ หรือองค์กรได้ดียิ่งขึ้น (Rattanatat, 2562)

### 2.5. RFM Model

ลูกค้าเป็นบุคคลที่มีบทบาทสำคัญในความอยู่รอดของธุรกิจและองค์กร การสร้างโอกาสทางการขายกับกลุ่มลูกค้าใหม่นั้นเป็นสิ่งจำเป็น เช่น ผ่านกิจกรรมทางการตลาด การพัฒนาแบรนด์ และสร้างสินค้าหรือบริการตอบที่สามารถสนองความต้องการของลูกค้า อย่างไรก็ตาม การบริหารความสัมพันธ์กับลูกค้าเก่าเป็นสิ่งสำคัญที่สุด เนื่องจากมีผลมากต่อความยั่งยืนของธุรกิจ การรักษาความพึงพอใจและความเชื่อมั่นของลูกค้าเก่าอาจช่วยในการสร้างลูกค้าที่เชื่อถือได้และพร้อมที่จะแนะนำผลิตภัณฑ์หรือบริการของธุรกิจให้กับผู้อื่นที่รู้จักด้วย หนึ่งในวิธีการเพิ่มประสิทธิผลการทำการตลาดคือการหากลุ่มลูกค้าชั้นดี (Best Customers) และสนองตอบกลับพวกเขาด้วยประสบการณ์ที่ดี โดยใช้ข้อมูลเกี่ยวกับ RFM เป็นตัวชี้วัด โดยแนวคิดของพาเรโตระบุว่า 80% ของยอดขายมักมาจากลูกค้าชั้นดีเพียงแค่ 20% ของลูกค้าทั้งหมด การใช้ข้อมูล RFM เพื่อระบุลูกค้าชั้นดีจะช่วยให้มองเห็นโอกาสในการเพิ่มยอดขายได้อย่างมีประสิทธิภาพ เพื่อสร้างความเชื่อมั่นและความสัมพันธ์กับลูกค้าชั้นดีเหล่านี้ (ปรีดี นุกุลสมปรารถนา, 2563) โดย RFM Model มีรายละเอียดดังนี้

Recency คือ การซื้อสินค้าหรือบริการล่าสุดของลูกค้า จะเสมือนเป็นสัญญาณที่ว่าพวกเขามีโอกาสกลับมาซื้ออีกครั้ง เนื่องจากความทรงจำและความปรารถนาที่เกิดจากการซื้อครั้งล่าสุดยังคงอยู่ในจิตใจของลูกค้ามากขึ้นเมื่อเปรียบเทียบกับการซื้อในอดีต ดังนั้นในด้านการตลาด การสร้างกิจกรรมหรือย้ำเตือนความทรงจำของลูกค้าให้กลายเป็นกลุ่มผู้ซื้อซ้ำ (Repeat Customer) เป็นสิ่งสำคัญ เพื่อคงความสนใจและเพิ่มโอกาสในการซื้ออีกครั้งจากลูกค้า

Frequency คือ จำนวนครั้งที่ลูกค้าซื้อสินค้าหรือบริการ ความถี่นี้สามารถมาจากหลายปัจจัย เช่น ประเภทของสินค้า ราคาสินค้า สถานที่ซื้อ และความต้องการในด้านต่าง ๆ ของลูกค้าลูกค้า เพื่อให้สินค้าหรือบริการที่พวกเขาซื้อ เป็นไปตามความต้องการและความสะดวก การเข้าใจว่าลูกค้าซื้อสินค้าหรือบริการอย่างไรจะช่วยให้ธุรกิจปรับการตลาดและการบริการเพื่อตอบสนองต่อความต้องการของพวกเขาได้อย่างเหมาะสม นักการตลาดสามารถวิเคราะห์วงจรการซื้อ (Purchase Cycle) เพื่อให้เข้าใจความถี่ในการซื้อและทำการตลาดให้เหมาะสมกับลูกค้า

Monetary เป็นปัจจัยสุดท้ายที่สำคัญ คือ จำนวนเงินที่ลูกค้าใช้ในการซื้อสินค้าหรือบริการ ยิ่งลูกค้าใช้จ่ายเงินมูลค่ามากกับการซื้อสินค้าหรือบริการเท่าไหร่ ก็ยิ่งช่วยบ่งบอกถึงความสนใจและความพึงพอใจในสินค้าหรือบริการนั้นๆ นักการตลาดสามารถใช้ข้อมูลจำนวนเงินที่ใช้ซื้อเพื่อพัฒนาสินค้าหรือบริการให้เหมาะสมและเพิ่มโอกาสในการขายในอนาคต

การวิเคราะห์แบ่งกลุ่มลูกค้าโดยพิจารณาความสำคัญของปัจจัยทั้งสามพร้อมกัน ในทางปฏิบัติ ข้อมูลเหล่านี้ไม่สามารถดูแยกจากกันได้ ต้องพิจารณาพร้อมกันเพื่อทราบถึงลักษณะการซื้อของลูกค้าแต่ละคน เช่น ลูกค้าคนไหนซื้อบ่อยแค่ไหน ซื้อจำนวนเงินเยอะแค่ไหน และเข้าร้านล่าสุดสุดเมื่อไหร่ เพื่อให้เราสามารถรับรู้และวิเคราะห์ได้ว่าเราควรจะจัดการกับกลุ่มลูกค้านี้หรือไม่ เพื่อสร้างการตลาดที่เหมาะสม วัตถุประสงค์หลักของธุรกิจคือการสร้างรายได้และทำให้ลูกค้าพอใจด้วยการจัดการตลาดที่ดีขึ้น ถึงอย่างนั้นก็ไม่ได้แปลว่าต้องทำการตลาดกับทุกกลุ่ม ธุรกิจสามารถเลือกกลุ่มลูกค้าที่สำคัญต่อธุรกิจได้ และนำข้อมูลเหล่านี้ปรับแผน การตลาดหรือแคมเปญและกลยุทธ์ในทางที่เหมาะสม การเก็บข้อมูลและการวิเคราะห์ด้วย RFM Analysis จะช่วยให้นักการตลาดมีข้อมูลและเป็นเครื่องมือที่สำคัญช่วยในการตัดสินใจได้อย่างมีเหตุผล เพื่อให้จัดทรัพยากรและกลยุทธ์การตลาดให้เหมาะสมและมีประสิทธิภาพมากยิ่งขึ้น

## 2.6 **การประเมินประสิทธิภาพของการจัดกลุ่ม** (Evaluation of Clustering)

มีหลายวิธีที่ใช้ในการประเมินประสิทธิภาพของการใช้ K-Means ในการจัดกลุ่ม (clustering) ซึ่งบางวิธีอาจเหมาะสำหรับบางประเภทของข้อมูลหรือการใช้งานมากกว่า ในงานวิจัยนี้จะยกมาบางส่วน ดังนี้

1. Elbow Method

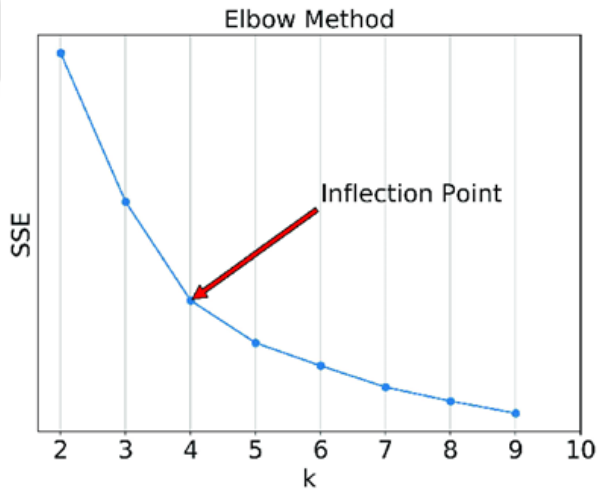
Elbow Method ใช้ในการประเมินประสิทธิภาพของการแบ่งกลุ่มด้วย K-Means Clustering โดยวัดค่าความคลาดเคลื่อนรวม (Sum of Square Error - SSE) ซึ่งเกิดจากระยะห่างระหว่างจุดข้อมูลและจุดศูนย์กลางของกลุ่ม (Centroid) ในแต่ละกลุ่ม จากนั้นจะวาดกราฟของค่า SSE จากการจัดกลุ่มที่แตกต่างกัน ถ้ากราฟมีลักษณะคล้ายข้อศอก จุดที่มีความชันเริ่มเรียบลงคือจุดที่เหมาะสมที่สุดในการเลือกจำนวนกลุ่ม (k) ที่เหมาะสม โดยค่า k นี้ จะมีค่า SSE ที่น้อยที่สุดในชุดข้อมูล การใช้ Elbow Method ช่วยในการจัดกลุ่มลูกค้าโดยคำนวณจากค่า SSE ที่เบี่ยงเบนน้อยที่สุด ซึ่งช่วยให้มีการตัดสินใจเลือกกลุ่มที่เหมาะสมและมีประสิทธิภาพในการวิเคราะห์ข้อมูลลูกค้าด้วย K-Means Clustering อย่างมีความเหมาะสม (ภัครพล อาจอาษา, 2564) แสดงจุดหักศอกของ Elbow Method ดังภาพประกอบ 5

$$SSE = \sum_{i=1}^{n} (y_i - y')^2 \quad (6)$$

โดยที่ n คือ จำนวนอินสแตนซ์ ,

$y_i$ คือ ค่าของอินสแตนซ์ที่ i

$y'$ คือ ค่าเฉลี่ยของอินสแตนซ์




ภาพประกอบ 5 แสดงจุดหักศอกของ Elbow Method

ที่มา : https://www.researchgate.net/figure/Elbow-method-for-choosing-the-number-of-cluster-centers_fig5_346716551

2. Silhouette Score

Silhouette Score เป็นอัลกอริทึมที่ใช้ในการประเมินประสิทธิภาพในการจัดกลุ่มด้วย K-Means Clustering โดยการวัดความคล้ายคลึงของข้อมูลภายในกลุ่มและความคล้ายคลึงกับกลุ่มอื่น ๆ ผ่านค่า Silhouette Score ที่มีค่าอยู่ระหว่าง -1 ถึง 1 การใช้ Silhouette Score ช่วยกำหนดจำนวนกลุ่มของลูกค้าร่วมกับ Elbow Method ในการค้นหาจำนวนกลุ่มที่เหมาะสม ค่า Silhouette Score สูงหมายความว่าการจัดกลุ่มที่ดี และมีความคล้ายคลึงภายในกลุ่มแสดงประสิทธิภาพการจัดกลุ่ม จากค่า Silhouette Score ดังภาพประกอบ 6 อีกทั้ง Silhouette Score ยังเป็นมาตรวัดที่ช่วยในการแยกกลุ่มข้อมูล โดยการใช้สมการคำนวณหาค่า Silhouette Score จะต้องใช้ระยะห่างระหว่างข้อมูลโดยใช้ Distance Metric เช่น Euclidean หรือ Manhattan distance และคำนวณค่าความคล้ายคลึงของข้อมูลภายในกลุ่มและกลุ่มอื่น ๆ ด้วยสูตรที่กำหนดไว้ (ภัครพล อาจอาษา, 2564)

$$s(i) = \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \tag{7}$$

$$b(i) = k_{\neq i}^{min} \frac{1}{|c_k|} \sum_{j \in C_k} d(i,j) \tag{8}$$

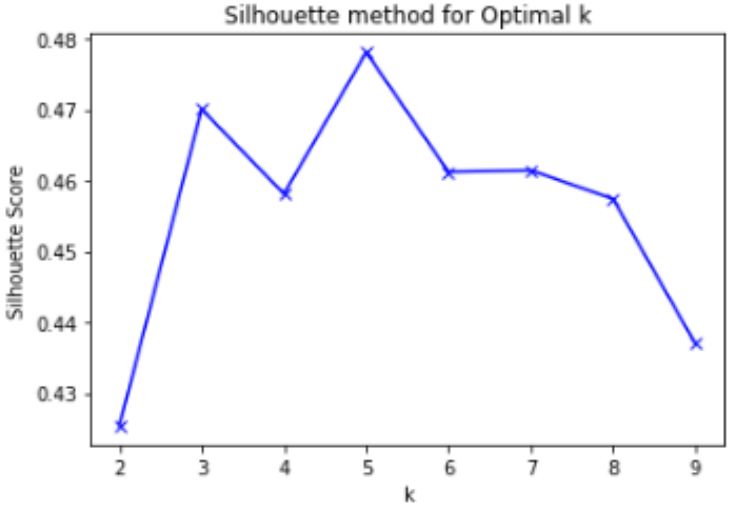
$$a(i) = \frac{1}{|c_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j) \tag{9}$$

โดยที่ $d(i,j)$ คือ Distance ระหว่าง $i$ and $j$ ใน ($C$)

$k$ คือ จำนวนของ Clusters

$a(i)$ คือ ค่าเฉลี่ยของ ระยะห่างระหว่างภายในกลุ่มข้อมูล

$b(i)$ คือ ค่าเฉลี่ยของ ระยะห่างระหว่างแต่ละกลุ่มของข้อมูล




ภาพประกอบ 6 แสดงประสิทธิภาพการจัดกลุ่ม จากค่า Silhouette Score

ที่มา : https://www.linkedin.com/pulse/selecting-optimal-number-clusters-kmeans-score-jyoti-yadav

3. Davies-Bouldin index

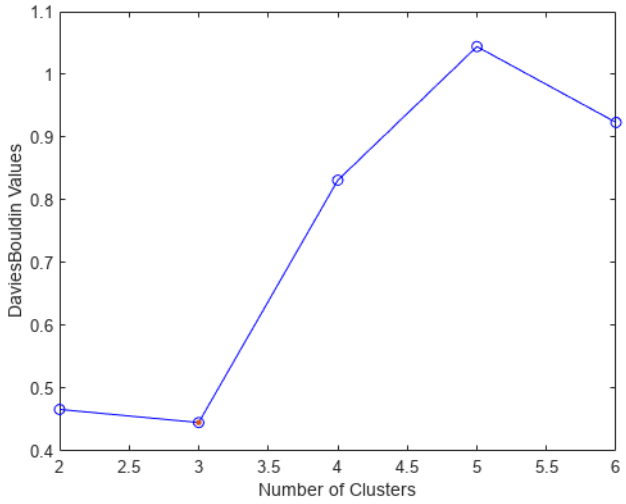
Davies-Bouldin index เป็นการประเมินคุณภาพของการจัดกลุ่มข้อมูล วิธีการนี้วัดความคล้ายคลึงระหว่างกลุ่มโดยใช้ค่าเฉลี่ยของความคล้ายคลึงภายในกลุ่มและระหว่างกลุ่ม โดยที่มีการกระจายตัวของข้อมูลในกลุ่มน้อยลงและมีระยะห่างระหว่างกลุ่มมากขึ้น การคำนวณด้วยวิธีนี้ค่าที่น้อยเป็นการแบ่งกลุ่มที่ดีมาก ซึ่งมักใช้เพื่อเปรียบเทียบความคล้ายคลึงระหว่างกลุ่มในการจัดกลุ่มที่ต่างกัน สมการคำนวณค่าของ Davies-Bouldin index ดังต่อไปนี้ (สุทธิพงศ์ ผ่องแผ้ว, 2555)

$$DB = \frac{1}{K}\sum_{k=1}^{K} \max\left\{\frac{S(U_k) + S(U_l)}{d(U_k, U_l)}\right\} \qquad (10)$$

โดย ค่า $S(U_k)+S(U_l)$ เป็นระยะห่างของข้อมูลภายในกลุ่ม k และกลุ่ม l

ค่า $d(U_k, U_l)$ เป็นระยะห่างระหว่างจุดกึ่งกลางกลุ่ม k กับกลุ่ม l

ดังนั้น หากค่าของ Davies-Bouldin index มีค่าน้อยที่สุดจะทำให้ได้การแบ่งแยกของกลุ่มดีที่สุด ดังภาพประกอบ 7




ภาพประกอบ 7 แสดงประสิทธิภาพการจัดกลุ่ม จากค่า Davies-Bouldin index

ที่มา :

https://www.mathworks.com/help/stats/clustering.evaluation.daviesbouldinevaluation.html

## 2.7 งานวิจัยที่เกี่ยวข้อง

งานวิจัยนี้ได้ทำการศึกษาค้นคว้างานวิจัยที่เกี่ยวข้อง โดยมีรายละเอียดดังนี้

1. **บทความวิจัยเรื่อง** Clustering Customers Using Their In-Depth Buying Behavior: A Pet Food Manufacturing Company Case Study

บทความกล่าวถึงเรื่อง การศึกษาวิเคราะห์แบ่งกลุ่มลูกค้าโดยใช้ข้อมูลการซื้อขายอาหารสัตว์เลี้ยง ได้ทำการศึกษาพฤติกรรมลักษณะของลูกค้าในแต่ละกลุ่ม โดยใช้เทคนิค K-means clustering รวบรวมข้อมูลการซื้อขายผลิตภัณฑ์อาหารสัตว์ของบริษัท ในช่วงปี ค.ศ. 2018-2020 มีลูกค้าทั้งหมด พบว่าวิธี Elbow สามารถกำหนดจำนวนกลุ่มได้ 8 กลุ่ม นิยามกลุ่มได้ดังนี้ กลุ่มลูกค้าทั่วไป คืนสินค้าน้อย, กลุ่มลูกค้าทั่วไปใจไม่นิ่ง, กลุ่มลูกค้าห่างไกล, กลุ่มลูกค้าขยันคืน, กลุ่มลูกค้าซื้อง่ายขายคล่อง, กลุ่มลูกค้ากระเป๋าหนัก, กลุ่มลูกค้าที่มีแนวโน้มควรรักษาไว้ และ กลุ่มลูกค้าขาจร (ณรรฐคุณ วิรุฬห์ศรี และคณะ, 2565)

2. **บทความวิจัยเรื่อง** THE STUDY OF CUSTOMER SEGMENTATION BY USING RFM MODEL AND TEXT ANALYTICS

บทความนี้กล่าวถึงเรื่อง การศึกษาวิธีการแบ่งกลุ่มลูกค้าด้วย 2 วิธีหลัก 1) การแบ่งกลุ่มลูกค้าด้วยเทคนิคอาร์เอฟเอ็ม 2) การแบ่งกลุ่มด้วยเทคนิคอาร์เอฟเอ็มและการวิเคราะห์ข้อความด้วย Natural Language Toolkit เปรียบเทียบกับการจัดกลุ่มด้วยค่า Adjusted Rand Index และค่า Normalized Mutual Information ซึ่งให้ค่า ARI เท่ากับ 0.5116 และ NMI เท่ากับ 0.3646 จากการเปรียบเทียบทั้ง 2 วิธี ผลที่ได้ไม่สอดคล้องกัน การแบ่งกลุ่มด้วยเทคนิคอาร์เอฟเอ็มร่วมกับการวิเคราะห์ข้อความ มีประสิทธิภาพดีกว่าวิธีที่ใช้เทคนิคอาร์เอฟเอ็มเพียงเทคนิคเดียว ที่ให้ข้อมูลเชิงลึกของการสั่งซื้อร่วมกับข้อมูลในส่วนของศักยภาพของลูกค้าจากข้อมูลอาร์เอฟเอ็ม (เอกปรียา ใบสนิ, 2563)

3. **บทความวิจัยเรื่อง** Customer Segmentation using K-means Clustering

บทความนี้กล่าวถึงเรื่อง การจัดกลุ่มด้วยอัลกอริทึมที่แตกต่างกัน 3 แบบ คือ K-Means, Agglomerative และ Meanshift เพื่อแบ่งกลุ่มลูกค้าและเปรียบเทียบผลลัพธ์ของกลุ่มที่ได้ ชุดข้อมูลมีคุณสมบัติ 2 ประการ จำนวน 200 ตัวอย่าง ที่นำมาจากร้านค้าปลีกในพื้นที่ ทั้ง 2 คุณสมบัติ คือ ค่าเฉลี่ยของปริมาณการซื้อโดยลูกค้า และค่าเฉลี่ยการเข้าเยี่ยมชมของลูกค้าในร้านค้าเป็นประจำทุกปี ข้อมูลได้รับการปรับขนาดโดยการใช้ตัวปรับขนาดมาตรฐาน Standard Scaler ข้อมูลจะมีศูนย์กลางประมาณ 0 โดยมีค่าเบี่ยงเบนมาตรฐานเป็น 1 clusters ของ K-means และ Agglomerative มีเท่ากัน โดยให้รูปแบบเดียวกัน ประยุกต์การจัดกลุ่มลูกค้าได้ 5 กลุ่ม ที่แบ่งออกเป็น Careless, Careful, Standard, Target และ Sensible การแบ่งกลุ่มด้วย mean shift clustering แบ่งได้ 2 กลุ่ม คือ High buyers and frequent visitors และ High buyers and occasional visitors ผลลัพธ์ของกลุ่มจากทั้ง 3 อัลกอริทึม ไม่มีความแตกต่างอย่างมีนัยสำคัญ ใน K-means และ Agglomerative clustering ดังนั้นอัลกอริทึมทั้งสองนี้สามารถจัดกลุ่มของข้อมูลได้ดีกว่า Mean shift วัดผลการประเมินด้วย silhouette score (Kansal et al., 2018)

4. **บทความวิจัยเรื่อง** Customer Segmentation Based on RFM Value Using K-Means Algorithm

บทความกล่าวถึงเรื่อง การแบ่งกลุ่มลูกค้าและวัดค่าเพื่อให้เจ้าของธุรกิจสามารถกำหนดว่าลูกค้ากลุ่มไหนที่ให้ประโยชน์และกลุ่มไหนที่ไม่ให้ประโยชน์ การระบุเกณฑ์ของลูกค้าในการสร้างกลุ่มตามค่า RFM (Recency, Frequency, Monetary) วิธีการจัดกลุ่มนี้ใช้ K-Means Clustering ผลลัพธ์จากวิธี Elbow Method แบ่งกลุ่มได้ทั้งหมด 3 กลุ่ม พร้อมลักษณะของแต่ละกลุ่ม กลุ่มที่ 1 ประกอบด้วยลูกค้า 69 คน จัดให้เป็น “ลูกค้าที่มาช้อปทุกวัน”, กลุ่มที่ 2 ประกอบด้วยลูกค้า 95 คน จัดให้เป็น “ลูกค้าที่ไม่มีความเคลื่อนไหว”, กลุ่มที่ 3 มีลูกค้า 11 คน จัดให้เป็น “ลูกค้าชั้นดี” ค่าของความแตกต่างของ SSE คือ 2.7630 และค่าเฉลี่ยของ Silhouette Index คือ 0.7210 บริษัทสามารถใช้เป็นสื่อส่งเสริมการขาย โปรโมชั่น ให้กับลูกค้าชั้นดีได้ (Dedi et al., 2019)

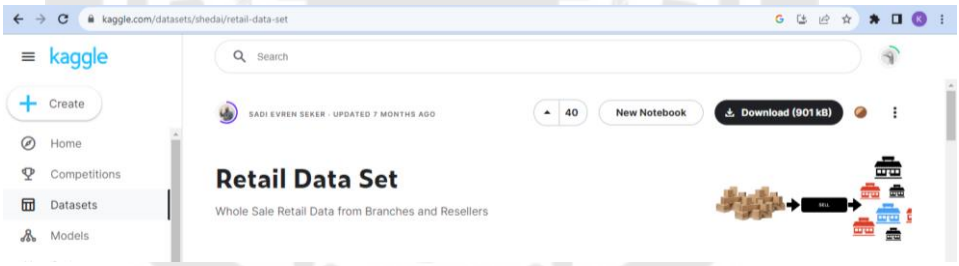
ในบทที่ 2 นี้ กล่าวถึงการทบทวนวรรณกรรมและงานวิจัยที่เกี่ยวข้องต่าง ๆ เพื่อนำมาประยุกต์ใช้และต่อยอด สำหรับวิธีดำเนินการวิจัย จะกล่าวถึงในบทที่ 3 ต่อไป

# บทที่ 3
# ขั้นตอนการดำเนินการวิจัย


ภาพประกอบ 8 แสดงขั้นตอนการวิจัย

## 3.1 การทำความเข้าใจปัญหา (Business Understanding)

ศึกษาเกี่ยวกับธุรกิจประเภทร้านขายปลีก ซึ่งใช้เป็นชุดข้อมูลสาธารณะจากเว็บไซต์ Kaggle.com ดังภาพประกอบ 9 โดยมีความต้องการแบ่งกลุ่มลูกค้า (Customer Segmentation) เพื่อให้สามารถแบ่งกลุ่มลูกค้าที่คล้ายคลึงกัน เพื่อนำไปสู่การพัฒนากลยุทธ์และวางแผนการตลาดให้ตอบโจทย์และเหมาะสมกับลูกค้าในทุก ๆ กลุ่ม




ภาพประกอบ 9 แสดงภาพบันทึกหน้าจอเว็บไซต์ Kaggle

ที่มา : https://www.kaggle.com/datasets/shedai/retail-data-set

## 3.2 การทำความเข้าใจข้อมูล (Data Understanding)

รวบรวมข้อมูลที่เกี่ยวข้องและเชื่อถือได้ ตรวจสอบคุณภาพของข้อมูลเพื่อให้เข้าใจลักษณะของข้อมูลและสามารถจัดการกับข้อมูลให้เหมาะสำหรับนำมาใช้วิเคราะห์ รวบรวมข้อมูลการขายของร้านขายปลีกทั้งการขายจากสาขาและผู้ค้าปลีก จัดเก็บข้อมูลในปี ค.ศ. 2019 ถึงต้นปี ค.ศ. 2023 มีข้อมูลทั้งหมด 29,103 รายการ มี 7 คุณลักษณะ แสดงถึงข้อมูลของแต่ละบุคคล ดังตาราง 1 แสดงตัวอย่างข้อมูลดังภาพประกอบ 10 และชนิดข้อมูลดังภาพประกอบ 11

ตาราง 1 ชื่อคอลัมน์ (Column Name) และคำอธิบาย (Description)

| Column Name | Description |
|---|---|
| InvoiceID | รหัสใบเสร็จ |
| Date | วันที่ทำรายการ |
| ProductID | รหัสสินค้าเฉพาะแต่ละรายการ |
| TotalSales | ยอดขายรวมต่อใบเสร็จ |
| Discount | จำนวนส่วนลด |
| CustomerID | รหัสลูกค้า ที่ไม่ซ้ำกัน |
| Quantity | จำนวนสินค้าที่ขายต่อใบเสร็จ |

```
df.head()
```

| | InvoiceID | Date | ProductID | TotalSales | Discount | CustomerID | Quantity |
|---|---|---|---|---|---|---|---|
| 0 | 328 | 2019-12-27 | 1684 | 796.61 | 143.39 | 185 | 4 |
| 1 | 329 | 2019-12-27 | 524 | 355.93 | 64.07 | 185 | 2 |
| 2 | 330 | 2019-12-27 | 192 | 901.69 | 162.31 | 230 | 4 |
| 3 | 330 | 2019-12-27 | 218 | 182.75 | 32.90 | 230 | 1 |
| 4 | 330 | 2019-12-27 | 247 | 780.10 | 140.42 | 230 | 4 |

```
# knowing its shape
df.shape
```

```
(29103, 7)
```

ภาพประกอบ 10 แสดงตัวอย่างของข้อมูล

```
#ดูชนิดข้อมูล
df.dtypes
```

```
InvoiceID             int64
Date          datetime64[ns]
ProductID             int64
TotalSales          float64
Discount            float64
CustomerID            int64
Quantity              int64
dtype: object
```

```
#เปลี่ยน InvoiceID , ProductID , CustomerID จาก int เป็น object
categorical_variables = ['InvoiceID', 'ProductID', 'CustomerID']

for i in categorical_variables:
    df[i] = df[i].astype(object)
```

ภาพประกอบ 11 แสดงชนิดของข้อมูล

### 3.3 **การเตรียมข้อมูล** (Data Preparation)

1. ทำความสะอาดข้อมูล

ตรวจสอบและแก้ไขข้อมูล เพื่อให้ข้อมูลอยู่ในรูปแบบที่ถูกต้องสมบูรณ์ เช่น ข้อมูลที่ซ้ำซ้อน (Duplicate data) ข้อมูลไม่ถูกต้อง (Incorrectly data) การสูญหายของข้อมูลบางส่วน (Missing Value) ค่าความผิดปกติหรือแตกต่างจากข้อมูลในกลุ่ม (Outliers) เป็นต้น ดังภาพประกอบ 12 ถึง 15

```
# ตรวจสอบความซ้ำซ้อน
df.duplicated().sum()
```

```
293
```

```
# ลดความซ้ำซ้อน
df.drop_duplicates(keep= 'first', inplace= True)
```

```
#หา missing value ข้อมูลที่ขาดหายไป
df.isnull().sum()
```

```
InvoiceID     0
Date          0
ProductID     0
TotalSales    0
Discount      0
CustomerID    0
Quantity      0
dtype: int64
```

ภาพประกอบ 12 แสดงการตรวจสอบและแก้ไขข้อมูล

```
# ตรวจสอบว่า TotalSales และ Quantity มีศูนย์อยู่หรือไม่
(df[['TotalSales', 'Quantity']] == 0).sum()
```

```
TotalSales    176
Quantity      158
dtype: int64
```

```
# เรียกดู TotalSales และ Quantity ที่มีศูนย์
df[(df.TotalSales == 0) & (df.Quantity == 0)]
```

| | InvoiceID | Date | ProductID | TotalSales | Discount | CustomerID | Quantity |
|---|---|---|---|---|---|---|---|
| **2328** | 1930 | 2019-10-05 | 885 | 0.00 | 0.00 | 17 | 0 |
| **6533** | 4718 | 2019-04-13 | 751 | 0.00 | 0.00 | 404 | 0 |
| **8706** | 289 | 2019-03-23 | 1830 | 0.00 | 0.00 | 430 | 0 |
| **8707** | 291 | 2019-10-21 | 224 | 0.00 | 0.00 | 276 | 0 |
| **8708** | 262 | 2019-11-18 | 925 | 0.00 | 0.00 | 364 | 0 |
| **12153** | 5899 | 2020-02-22 | 1793 | 0.00 | 0.00 | 171 | 0 |
| **14770** | 5900 | 2020-05-08 | 221 | 0.00 | 0.00 | 114 | 0 |
| **21488** | 290 | 2021-11-05 | 503 | 0.00 | 0.00 | 382 | 0 |
| **22081** | 5903 | 2022-02-19 | 1743 | 0.00 | 0.00 | 51 | 0 |

ภาพประกอบ 13 แสดงการตรวจสอบยอดซื้อและจำนวนสินค้าที่เป็นศูนย์

```
# เรียกดู ProductID ที่ไม่ซ้ำใคร
unique_product_id = df['ProductID'].unique()
unique_product_id
```

```
array([1684, 524, 192, ..., 1201, 1261, 371], dtype=object)
```

ณหา Unit Price กับ Quantity Mode

```
# คำนวณราคาต่อหน่วยโดยเฉลี่ยของผลิตภัณฑ์แต่ละรายการ และปริมาณที่นิยม

list_unit_price = [] # empty list to store avg unit price
list_qty = [] # empty list to store mode qty

for i in unique_product_id:
    temp_df = df[df.ProductID == i] # temp dataframe to store each product data

    temp_unit_price = temp_df['TotalSales'].sum() / temp_df['Quantity'].sum() # calculating avg unit price of particular product
    list_unit_price.append(temp_unit_price) # appending to list_unit_price

    temp_qty = round(temp_df['Quantity'].mode().iloc[0]) # calculating mode quantity of the particular product
    list_qty.append(temp_qty) # appending to list_qty
```

```
# สร้าง dataframe ที่มี ProductID พร้อมด้วย Avg_Unit_Price และ Mode_Quantity

product_dict = {'ProductID' : unique_product_id, 'Avg_Unit_Price' : list_unit_price, 'Mode_Quantity' : list_qty} # columns

product_df = pd.DataFrame(data = product_dict) # creating dataframe
product_df.head()
```

| | ProductID | Avg_Unit_Price | Mode_Quantity |
|---|---|---|---|
| 0 | 1684 | 198.35 | 4 |
| 1 | 524 | 209.74 | 4 |
| 2 | 192 | 301.05 | 4 |
| 3 | 218 | 225.84 | 4 |
| 4 | 247 | 234.05 | 4 |

```
# สร้าง dataframe ชั่วคราวที่มี TotalSales เป็นศูนย์จากชุดข้อมูลดั้งเดิม
df_temp = df[df['TotalSales'] == 0]
df_temp.head()
```

| | InvoiceID | Date | ProductID | TotalSales | Discount | CustomerID | Quantity |
|---|---|---|---|---|---|---|---|
| 2328 | 1930 | 2019-10-05 | 885 | 0.00 | 0.00 | 17 | 0 |
| 6533 | 4718 | 2019-04-13 | 751 | 0.00 | 0.00 | 404 | 0 |
| 8663 | 242 | 2019-01-15 | 1790 | 0.00 | 0.00 | 122 | 4 |
| 8665 | 244 | 2019-02-15 | 1470 | 0.00 | 0.00 | 129 | 4 |
| 8667 | 246 | 2019-02-28 | 1499 | 0.00 | 0.00 | 129 | 4 |

```
# แทนที่ศูนย์ของ TotalSales ด้วย Avg_Unit_Price และศูนย์ของ Quantity ด้วย Mode_Quantity ของผลิตภัณฑ์นั้น ๆ

k = 0
for i in product_df['ProductID']: # getting ProductID from product_df

    for j in df_temp.index: # getting index number from df_temp

        if df_temp['ProductID'][j] == i and df_temp['Quantity'][j] == 0: # validating condition

            df_temp['Quantity'][j] = product_df['Mode_Quantity'][k] # updating quantity
            df_temp['TotalSales'][j] = product_df['Avg_Unit_Price'][k] * product_df['Mode_Quantity'][k] # updating totalsales

        elif df_temp['ProductID'][j] == i and df_temp['Quantity'][j] != 0: # validating condition

            df_temp['TotalSales'][j] = product_df['Avg_Unit_Price'][k] * df_temp['Quantity'][j] # updating totalsales

        else:
            pass

    k+=1
```



ภาพประกอบ 14 แสดงการแทนที่ด้วยการแก้ไขเป็นค่าที่ถูกต้องหรือแทนด้วยค่าใหม่

```
# ตรวจสอบอีกครั้งหลังจากอัปเดตว่ามีเลขศูนย์อยู่หรือไม่
(df[['TotalSales', 'Quantity']] == 0).sum()
```

```
TotalSales    16
Quantity      16
dtype: int64
```

```
# ลบข้อมูลที่ยังคงเป็นศูนย์

drop_list = list(df[df['TotalSales'] == 0].index) # getting the index of those records
df.drop(drop_list, inplace= True) # dropping
```

ภาพประกอบ 15 แสดงการตรวจสอบและลบข้อมูลที่ยังไม่ถูกต้อง

2. สร้างตัวแปรใหม่จากข้อมูลที่มีอยู่

การสร้างตัวแปรใหม่หรือ Feature Engineering จากข้อมูลการซื้อสินค้า (Transaction) ที่เก็บมา โดยปกติแล้วข้อมูลที่มีไม่สามารถนำมาใช้ได้โดยตรง เนื่องจากการซื้อสินค้าหนึ่งครั้งมากกว่า 1 แถว และลูกค้า 1 คน มีการซื้อมากกว่า 1 ครั้งได้ วัตถุประสงค์ของการวิเคราะห์ข้อมูลในกรณีนี้ต้องการแบ่งกลุ่มลูกค้าแต่ในข้อมูลที่มีจะเป็นรายละเอียดของการซื้อสินค้าแต่ละครั้ง ดังนั้น จึงต้องการสร้างเป็น profile ของลูกค้าแต่ละรายเสียก่อน โดยการสร้าง Profile นี้จะพิจารณาตัวแปรต่างๆ ดังนี้

1) Recency (R) คือ จำนวนวันที่ซื้อครั้งล่าสุดห่างจากวันคำนวณกี่วัน ดังภาพประกอบ 16

```
# การคำนวณ recency ของลูกค้าแต่ละราย

recency = df_part.groupby('CustomerID').agg({'Date': lambda x :
                                              (df_part_latest_date - x.max()).days}).reset_index() # calculating recency

recency.rename(columns= {'Date':'Recency'}, inplace= True) # renaming columns

recency.head()
```

| | CustomerID | Recency |
|---|---|---|
| 0 | 0 | 7 |
| 1 | 1 | 735 |
| 2 | 2 | 52 |
| 3 | 3 | 625 |
| 4 | 4 | 92 |

ภาพประกอบ 16 แสดงการสร้างตัวแปร Recency

2) Frequency (F) คือ จำนวนครั้งในการซื้อสินค้า พิจารณาจากหมายเลขคำสั่งซื้อที่ต่างกันของลูกค้าแต่ละราย ดังภาพประกอบ 17

```
# การคำนวณ frequency ของลูกค้าแต่ละราย

frequency = df_part.groupby('CustomerID').agg({'CustomerID':'count'}) # calculating frequency

frequency.rename(columns= {'CustomerID':'Frequency'}, inplace= True) # renaming columns
frequency.reset_index(inplace= True) # resetting index

frequency.head()
```

| | CustomerID | Frequency |
|---|---|---|
| 0 | 0 | 50 |
| 1 | 1 | 13 |
| 2 | 2 | 36 |
| 3 | 3 | 1 |
| 4 | 4 | 38 |

ภาพประกอบ 17 แสดงการสร้างตัวแปร Frequency

3) Monetary (M) คือ ยอดค่าใช้จ่ายรวมจากทุกตาราง ดังภาพประกอบ 18

```
# การคำนวณ monetary ของลูกค้าแต่ละราย

monetary = df_part.groupby('CustomerID').agg({'TotalSales':'sum'}).reset_index() # calculting monetary

monetary.rename(columns= {'TotalSales':'Monetary'}, inplace= True) # renaming columns

monetary.head()
```

| | CustomerID | Monetary |
|---|---|---|
| 0 | 0 | 303317.93 |
| 1 | 1 | 20476.73 |
| 2 | 2 | 51963.71 |
| 3 | 3 | 300.85 |
| 4 | 4 | 60977.72 |

ภาพประกอบ 18 แสดงการสร้างตัวแปร Monetary

4) Basket Size (B) คือ ยอดค่าใช้จ่ายแต่ละครั้ง ดังภาพประกอบ 19

```
# การคำนวณ basket size ของลูกค้าแต่ละราย

basketsize = df_part.groupby('CustomerID').agg({'TotalSales':'mean'}).reset_index() # calculting basket size

basketsize.rename(columns= {'TotalSales':'Basket Size'}, inplace= True) # renaming columns

basketsize.head()
```

| | CustomerID | Basket Size |
|---|---|---|
| 0 | 0 | 6066.36 |
| 1 | 1 | 1575.13 |
| 2 | 2 | 1443.44 |
| 3 | 3 | 300.85 |
| 4 | 4 | 1604.68 |

ภาพประกอบ 19 แสดงการสร้างตัวแปร Basket Size

5) Day_Type (D) คือ วันที่ลูกค้ามาใช้บริการ ดังภาพประกอบ 20

```
# สร้างฟังก์ชันในการแบ่งกลุ่มวันที่เป็นวันหยุดหรือวันธรรมดา
def classify_weekday(day_of_week):
    if day_of_week < 5:
        return "Weekday"
    else:
        return "Weekend"

# ใช้ฟังก์ชันในการสร้างคอลัมน์ "Day_Type" ที่แบ่งกลุ่มวันที่
df_part['Day_Type'] = df_part['Day_of_Week'].apply(classify_weekday)
```

```
# แสดงผล DataFrame ที่แบ่งกลุ่มวันที่เป็นวันหยุดและวันธรรมดา
df_part.head()
```

| | Date | InvoiceID | ProductID | TotalSales | Discount | CustomerID | Quantity | Year | Month | Day | Day_of_Week | Day_Type |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 2019-12-27 | 328 | 1684 | 796.61 | 143.39 | 185 | 4 | 2019 | December | Friday | 4 | Weekday |
| 1 | 2019-12-27 | 329 | 524 | 355.93 | 64.07 | 185 | 2 | 2019 | December | Friday | 4 | Weekday |
| 2 | 2019-12-27 | 330 | 192 | 901.69 | 162.31 | 230 | 4 | 2019 | December | Friday | 4 | Weekday |
| 3 | 2019-12-27 | 330 | 218 | 182.75 | 32.90 | 230 | 1 | 2019 | December | Friday | 4 | Weekday |
| 4 | 2019-12-27 | 330 | 247 | 780.10 | 140.42 | 230 | 4 | 2019 | December | Friday | 4 | Weekday |

ภาพประกอบ 20 แสดงการสร้างคอลัมน์ Day_Type

### 3.4 **การสร้างแบบจำลอง** (Modeling)

นำเอาข้อมูลที่ได้ทำการวิเคราะห์และเตรียมข้อมูลเสร็จแล้ว มาเข้าสู่การสร้างแบบจำลอง เพื่อแบ่งกลุ่มลูกค้าตามที่ต้องการ ด้วยวิธีการและเทคนิคต่าง ๆ ซึ่งหากการทำ Customer Segmentation หรือ Clustering ด้วย RFM หรือ K-means พบว่ามีข้อมูลผิดพลาด สามารถกลับไปทำการเตรียมข้อมูลใหม่อีกครั้งได้เช่นกัน

ในการศึกษานี้ จะสร้างแบบจำลอง 2 แบบ โดยแบบจำลองที่ 1 จะเป็นการแบ่งกลุ่มด้วยข้อมูล RFM แบบจำลองที่ 2 จะเป็นการแบ่งกลุ่มด้วยข้อมูล RFMBD เพิ่มตัวแปรจากยอดใช้จ่ายต่อครั้ง ร่วมกับการจัดกลุ่มตามวันที่มาใช้บริการ โดยแบบจำลองแต่ละแบบมีรายละเอียดดังนี้

**แบบจำลองที่** 1 การแบ่งกลุ่มด้วยการใช้ข้อมูล RFM

| | Recency | Frequency | Monetary |
|---|---|---|---|
| 0 | 7 | 50 | 303317.93 |
| 1 | 735 | 13 | 20476.73 |
| 2 | 52 | 36 | 51963.71 |
| 3 | 625 | 1 | 300.85 |
| 4 | 92 | 38 | 60977.72 |

**แบบจำลองที่** 2 การแบ่งกลุ่มด้วยการใช้ข้อมูล RFMBD

| | Recency | Frequency | Monetary | Basket Size | Day_Type |
|---|---|---|---|---|---|
| **0** | 7.00 | 50.00 | 303317.93 | 6066.36 | 0 |
| **1** | 735.00 | 13.00 | 20476.73 | 1575.13 | 0 |
| **2** | 52.00 | 36.00 | 51963.71 | 1443.44 | 0 |
| **3** | 625.00 | 1.00 | 300.85 | 300.85 | 0 |
| **4** | 92.00 | 38.00 | 60977.72 | 1604.68 | 0 |

### 3.5 **การประเมินประสิทธิภาพ** (Evaluation)

ในการศึกษานี้ ใช้วิธีวิเคราะห์หาจำนวนกลุ่มที่เหมาะสมด้วยค่า Silhouette Score และค่า Davies-Bouldin Index ทำการประเมินประสิทธิภาพของผลลัพธ์เพื่อเปรียบเทียบกลุ่มข้อมูลจากทั้ง 2 แบบ ด้วยตัวประเมิน Adjusted Rand Index (ARI) และ Normalized Mutual Information (NMI) เพื่อวัดความคล้ายคลึงระหว่างการแบ่งกลุ่มของทั้งสองแบบจำลอง

ในบทที่ 3 นี้ ได้นำเสนอขั้นตอนการดำเนินงานวิจัย ได้กล่าวถึงชุดข้อมูล การดำเนินงานทำความเข้าใจปัญหาและใจข้อมูล การเตรียมข้อมูล สร้างแบบจำลอง และการประเมินประสิทธิภาพของการแบ่งกลุ่ม โดยในบทที่ 4 จะมีการกล่าวถึงรายละเอียดและผลการศึกษาของการแบ่งกลุ่มของลูกค้าต่อไป

# บทที่ 4
# ผลการศึกษา

การศึกษาวิธีการแบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อโดยใช้เทคนิคเคมีน ดำเนินการตามขั้นตอนต่างๆ เพื่อให้เป็นไปตามวัตถุประสงค์ที่กำหนดไว้ ดังนี้

1. ผลการแบ่งกลุ่ม ทำ K-Means Clustering ด้วย RFM
2. ผลการแบ่งกลุ่ม ทำ K-Means Clustering ด้วย RFMBD
3. ผลการเปรียบเทียบระหว่างการแบ่งกลุ่มด้วย RFM และ RFMBD

## 4.1 ผลการแบ่งกลุ่ม ทำ K-Means Clustering ด้วย RFM

แบบจำลองการแบ่งกลุ่มด้วย RFM ใช้ข้อมูลลูกค้า ดังภาพที่ 21 การทำ Clustering จะเป็นการแบ่งกลุ่มข้อมูลที่เลือกค่า K ที่ดีที่สุด โดยใช้ Elbow Method ดังภาพที่ 22 และพิจารณาจากค่า Silhouette Score ดังภาพที่ 23 และค่า Davies-Bouldin Index ดังภาพที่ 24 ผลลัพธ์จากการจัดกลุ่มแสดงได้ดังภาพที่ 25

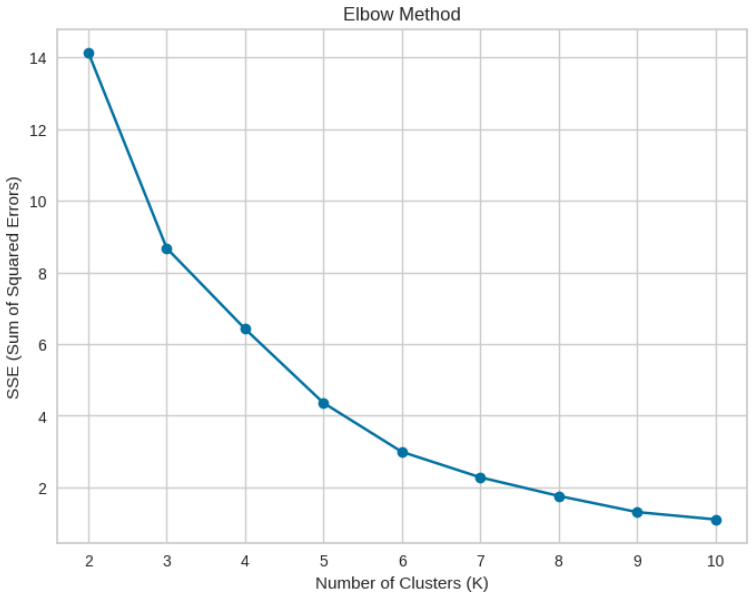
| | Recency | Frequency | Monetary |
|---|---|---|---|
| **0** | 7.0 | 50.0 | 303317.932 |
| **1** | 735.0 | 13.0 | 20476.729 |
| **2** | 52.0 | 36.0 | 51963.712 |
| **3** | 625.0 | 1.0 | 300.847 |
| **4** | 92.0 | 38.0 | 60977.720 |
| **...** | ... | ... | ... |
| **502** | 7.0 | 130.0 | 275282.339 |
| **503** | 229.0 | 3.0 | 19930.508 |
| **504** | 131.0 | 5.0 | 16088.983 |
| **505** | 12.0 | 67.0 | 286124.865 |
| **506** | 814.0 | 298.0 | 356847.178 |

507 rows × 3 columns

ภาพประกอบ 21 แสดงข้อมูล RFM จากขั้นตอนการวิเคราะห์ข้อมูลลูกค้า

1. **ผลลัพธ์จำนวนกลุ่มที่เหมาะสมด้วยวิธี** Elbow

จำนวนกลุ่มที่เหมาะสม เมื่อพิจารณาจากการหักศอก ได้จำนวนเท่ากับ 4 กลุ่ม ดังภาพประกอบ 22 และรายละเอียดดังตาราง 2



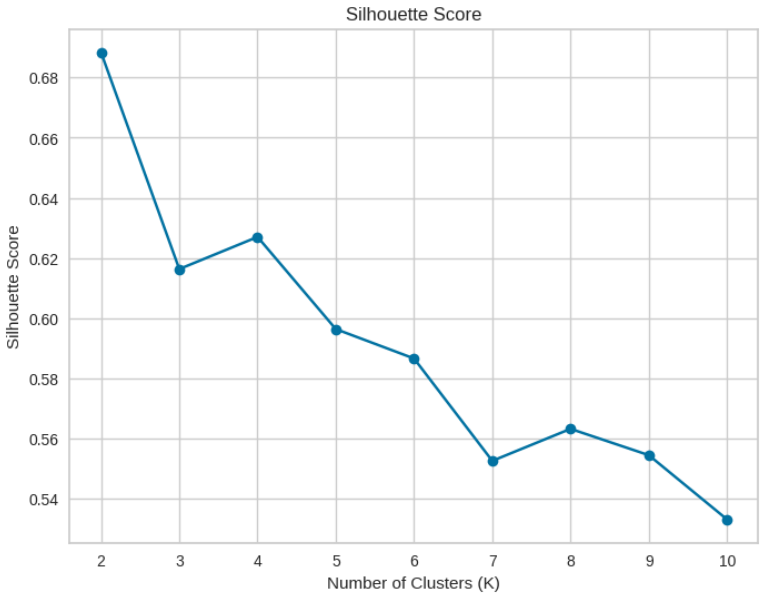

ภาพประกอบ 22 แสดงผลลัพธ์จำนวนกลุ่ม RFM ที่เหมาะสมด้วยวิธี Elbow

ตาราง 2 แสดงค่า Sum of squared errors (SSE) ของข้อมูล RFM

| | |
|---|---|
| SSE for K = 2 | 14.1414 |
| SSE for K = 3 | 8.6759 |
| SSE for K = 4 | 6.4226 |
| SSE for K = 5 | 4.3557 |
| SSE for K = 6 | 2.9904 |
| SSE for K = 7 | 2.2785 |
| SSE for K = 8 | 1.7598 |
| SSE for K = 9 | 1.3088 |
| SSE for K = 10 | 1.1037 |

**2. ผลลัพธ์จำนวนกลุ่มที่เหมาะสมด้วยวิธี Silhouette Score**

จำนวนกลุ่มที่เหมาะสม เมื่อพิจารณาจากค่าเฉลี่ยของ Silhouette Score เท่ากับ 0.6884 มีค่าเข้าใกล้ 1 มากที่สุด ได้จำนวนเท่ากับ 2 กลุ่ม ดังภาพประกอบ 23 และรายละเอียดดังตาราง 3



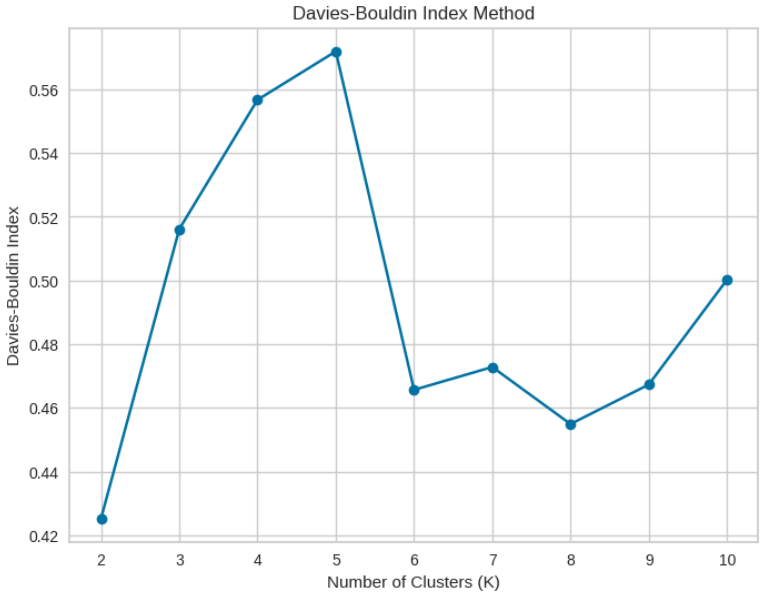

ภาพประกอบ 23 แสดงผลลัพธ์จำนวนกลุ่ม RFM ที่เหมาะสมด้วยวิธี Silhouette Score

ตาราง 3 แสดงค่า Silhouette Score ของข้อมูล RFM

| | |
|---|---|
| SC for K = 2 | 0.6884 |
| SC for K = 3 | 0.6163 |
| SC for K = 4 | 0.6270 |
| SC for K = 5 | 0.5964 |
| SC for K = 6 | 0.5866 |
| SC for K = 7 | 0.5526 |
| SC for K = 8 | 0.5632 |
| SC for K = 9 | 0.5545 |
| SC for K = 10 | 0.5331 |

**3. ผลลัพธ์จำนวนกลุ่มที่เหมาะสมด้วยวิธี Davies-Bouldin Index**

จำนวนกลุ่มที่เหมาะสมของลูกค้า เมื่อพิจารณาจากค่าเฉลี่ยของ Davies-Bouldin Index อยู่ที่ 0.4253 ซึ่งมีค่าน้อยที่สุด ได้จำนวนเท่ากับ 2 กลุ่ม ดังภาพประกอบ 24 และรายละเอียดดังตาราง 4




ภาพประกอบ 24 แสดงผลลัพธ์จำนวนกลุ่ม RFM ที่เหมาะสมด้วยวิธี Davies-Bouldin Index

ตาราง 4 แสดงค่า Davies-Bouldin Index ของข้อมูล RFM

| | |
|---|---|
| DBI for K = 2 : | 0.4253 |
| DBI for K = 3 : | 0.5161 |
| DBI for K = 4 : | 0.5567 |
| DBI for K = 5 : | 0.5719 |
| DBI for K = 6 : | 0.4657 |
| DBI for K = 7 : | 0.4728 |
| DBI for K = 8 : | 0.4549 |
| DBI for K = 9 : | 0.4673 |
| DBI for K = 10 : | 0.5002 |

| | Recency | Frequency | Monetary | Cluster_RFM |
|---|---|---|---|---|
| **0** | 7.0 | 50.0 | 303317.932 | 0 |
| **1** | 735.0 | 13.0 | 20476.729 | 1 |
| **2** | 52.0 | 36.0 | 51963.712 | 0 |
| **3** | 625.0 | 1.0 | 300.847 | 0 |
| **4** | 92.0 | 38.0 | 60977.720 | 0 |
| **...** | ... | ... | ... | ... |
| **502** | 7.0 | 130.0 | 275282.339 | 0 |
| **503** | 229.0 | 3.0 | 19930.508 | 0 |
| **504** | 131.0 | 5.0 | 16088.983 | 0 |
| **505** | 12.0 | 67.0 | 286124.865 | 0 |
| **506** | 814.0 | 298.0 | 356847.178 | 1 |

507 rows × 4 columns

ภาพประกอบ 25 แสดงตัวอย่างผลลัพธ์การจัดกลุ่มด้วยเทคนิค RFM

จากผลลัพธ์ของ Elbow Method และ Silhouette Score และ Davies-Bouldin Index ผู้วิจัยเลือกจัดกลุ่มข้อมูลเป็น 2 กลุ่ม โดยได้ผลลัพธ์ภาพประกอบ 26

| Cluster_RFM | Recency | Frequency | Monetary |
|---|---|---|---|
| **0** | 157.571 | 82.724 | 219442.884 |
| **1** | 1130.420 | 7.167 | 8698.319 |

ภาพประกอบ 26 แสดงจำนวนผลลัพธ์การจัดกลุ่มด้วยเทคนิค RFM

### 4.2 ผลการแบ่งกลุ่ม ทำ K-Means Clustering RFMBD

สร้างแบบจำลองการจัดกลุ่มด้วย RFM และเพิ่มตัวแปร B กับ D โดยใช้ข้อมูลลูกค้าดังตัวอย่างในภาพที่ 27 การทำ Clustering จะเป็นการแบ่งกลุ่มข้อมูลที่เลือกค่า K ที่ดีที่สุดโดยใช้ Elbow Method ดังภาพที่ 28 และพิจารณาจากค่า Silhouette Score ดังภาพที่ 29 และค่า Davies-Bouldin Index ดังภาพที่ 30 ผลลัพธ์จากการจัดกลุ่มแสดงได้ดังภาพที่ 31

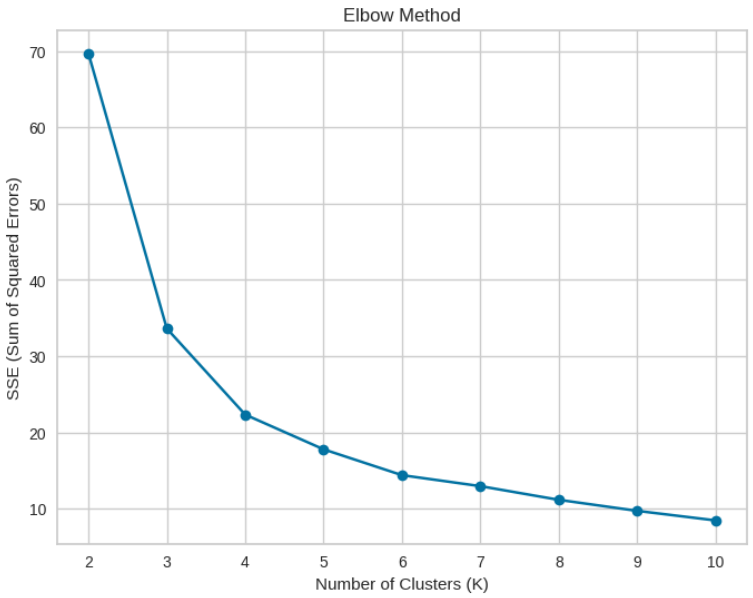
| | Recency | Frequency | Monetary | Basket Size | Day_Type |
|---|---|---|---|---|---|
| **0** | 7.0 | 50.0 | 303317.932 | 6066.359 | 0 |
| **1** | 735.0 | 13.0 | 20476.729 | 1575.133 | 0 |
| **2** | 52.0 | 36.0 | 51963.712 | 1443.436 | 0 |
| **3** | 625.0 | 1.0 | 300.847 | 300.847 | 0 |
| **4** | 92.0 | 38.0 | 60977.720 | 1604.677 | 0 |
| **...** | ... | ... | ... | ... | ... |
| **502** | 7.0 | 130.0 | 275282.339 | 2117.556 | 0 |
| **503** | 229.0 | 3.0 | 19930.508 | 6643.503 | 0 |
| **504** | 131.0 | 5.0 | 16088.983 | 3217.797 | 0 |
| **505** | 12.0 | 67.0 | 286124.865 | 4270.520 | 0 |
| **506** | 814.0 | 298.0 | 356847.178 | 1197.474 | 0 |

507 rows × 5 columns

ภาพประกอบ 27 แสดงข้อมูล RFMBD ที่ได้จากขั้นตอนการวิเคราะห์ข้อมูลลูกค้า

1. **ผลลัพธ์จำนวนกลุ่มที่เหมาะสมด้วยวิธี** Elbow

จำนวนกลุ่มที่เหมาะสมของลูกค้า เมื่อพิจารณาจากการหักศอก ได้จำนวนเท่ากับ 4 กลุ่ม ดังภาพประกอบ 28 และรายละเอียด ดังตาราง 5



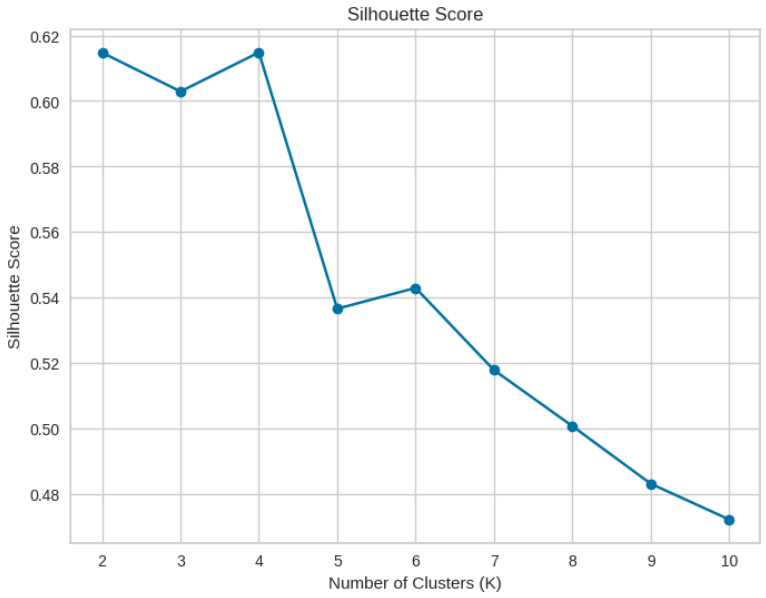

ภาพประกอบ 28 แสดงผลลัพธ์จำนวนกลุ่ม RFMBD ที่เหมาะสมด้วยวิธี Elbow

ตาราง 5 แสดงค่า Sum of squared errors (SSE) ของข้อมูล RFMBD

| | |
|---|---|
| SSE for K = 2 : | 69.7246 |
| SSE for K = 3 : | 33.6266 |
| SSE for K = 4 : | 22.3355 |
| SSE for K = 5 : | 17.7845 |
| SSE for K = 6 : | 14.3939 |
| SSE for K = 7 : | 12.9532 |
| SSE for K = 8 : | 11.1594 |
| SSE for K = 9 : | 9.71310 |
| SSE for K = 10 : | 8.4590 |

**2. ผลลัพธ์จำนวนกลุ่มที่เหมาะสมด้วยวิธี** Silhouette Score

จำนวนกลุ่มที่เหมาะสมของลูกค้า เมื่อพิจารณาจากค่าเฉลี่ยของ Silhouette Score อยู่ที่ 0.6148 ซึ่งมีค่าเข้าใกล้ 1 มากที่สุด ได้จำนวนเท่ากับ 4 กลุ่ม ดังภาพประกอบ 29 และรายละเอียด ดังตาราง 6



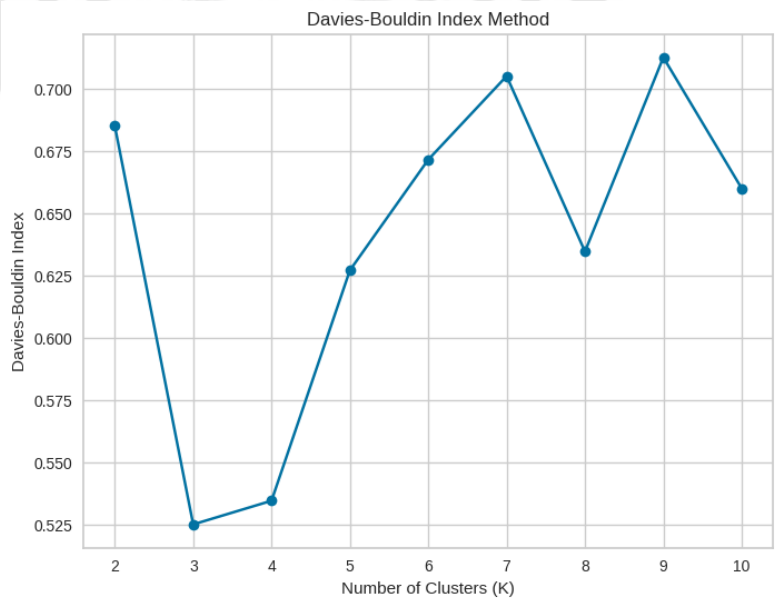

ภาพประกอบ 29 แสดงผลลัพธ์จำนวนกลุ่ม RFMBD ที่เหมาะสมด้วยวิธี Silhouette Score

ตาราง 6 แสดงค่า Silhouette Score ของข้อมูล RFMBD

| | |
|---|---|
| SC for K = 2 : | 0.6147 |
| SC for K = 3 : | 0.6029 |
| SC for K = 4 : | 0.6148 |
| SC for K = 5 : | 0.5366 |
| SC for K = 6 : | 0.5430 |
| SC for K = 7 : | 0.5179 |
| SC for K = 8 : | 0.5008 |
| SC for K = 9 : | 0.4832 |
| SC for K = 10 : | 0.4723 |

**3. ผลลัพธ์จำนวนกลุ่มที่เหมาะสมด้วยวิธี Davies-Bouldin Index**

จำนวนกลุ่มที่เหมาะสมของลูกค้า เมื่อพิจารณาจากค่าเฉลี่ยของ Davies-Bouldin Index อยู่ที่ 0.5249 ซึ่งมีค่าน้อยที่สุด ได้จำนวนเท่ากับ 3 กลุ่ม ดังภาพประกอบ 30 และรายละเอียด ดังตาราง 7




ภาพประกอบ 30 แสดงผลลัพธ์จำนวนกลุ่ม RFMBD ที่เหมาะสมด้วยวิธี Davies-Bouldin Index

ตาราง 7 แสดงค่า Davies-Bouldin Index ของข้อมูล RFMBD

| | |
|---|---|
| DBI for K = 2 : | 0.6850 |
| DBI for K = 3 : | 0.5249 |
| DBI for K = 4 : | 0.5344 |
| DBI for K = 5 : | 0.6270 |
| DBI for K = 6 : | 0.6715 |
| DBI for K = 7 : | 0.7051 |
| DBI for K = 8 : | 0.6346 |
| DBI for K = 9 : | 0.7126 |
| DBI for K = 10 : | 0.6599 |

| | Recency | Frequency | Monetary | Basket Size | Day_Type | Cluster_RFMBD |
|---|---|---|---|---|---|---|
| **0** | 7.0 | 50.0 | 303317.932 | 6066.359 | 0 | 1 |
| **1** | 735.0 | 13.0 | 20476.729 | 1575.133 | 0 | 0 |
| **2** | 52.0 | 36.0 | 51963.712 | 1443.436 | 0 | 1 |
| **3** | 625.0 | 1.0 | 300.847 | 300.847 | 0 | 0 |
| **4** | 92.0 | 38.0 | 60977.720 | 1604.677 | 0 | 1 |
| **...** | ... | ... | ... | ... | ... | ... |
| **502** | 7.0 | 130.0 | 275282.339 | 2117.556 | 0 | 1 |
| **503** | 229.0 | 3.0 | 19930.508 | 6643.503 | 0 | 1 |
| **504** | 131.0 | 5.0 | 16088.983 | 3217.797 | 0 | 1 |
| **505** | 12.0 | 67.0 | 286124.865 | 4270.520 | 0 | 1 |
| **506** | 814.0 | 298.0 | 356847.178 | 1197.474 | 0 | 0 |

507 rows × 6 columns

ภาพประกอบ 31 แสดงตัวอย่างผลลัพธ์การจัดกลุ่มด้วยเทคนิค RFMBD

จากผลลัพธ์ของ Elbow Method และ Silhouette Score และ Davies-Bouldin Index ผู้วิจัยเลือกจัดกลุ่มข้อมูลเป็น 4 กลุ่ม ได้ผลลัพธ์ดังภาพประกอบ 32

| Cluster_RFMBD | Recency | Frequency | Monetary | Basket Size | Day_Type |
|---|---|---|---|---|---|
| **0** | 1110.229 | 6.771 | 8560.058 | 1450.213 | 0.0 |
| **1** | 155.610 | 67.027 | 191683.148 | 3261.989 | 0.0 |
| **2** | 109.364 | 154.106 | 354585.785 | 3480.616 | 1.0 |
| **3** | 1100.262 | 7.500 | 8118.628 | 1160.119 | 1.0 |

ภาพประกอบ 32 แสดงจำนวนผลลัพธ์การจัดกลุ่มด้วยเทคนิค RFMBD

### 4.3 ผลการเปรียบเทียบระหว่างการแบ่งกลุ่มด้วย RFM และการแบ่งกลุ่มด้วย RFMBD

ผลการประเมินของการแบ่งกลุ่มที่ได้จากการทดลองแบ่งกลุ่มด้วย RFM ดังตาราง 8

ตาราง 8 แสดงผลการประเมินของการแบ่งกลุ่มของข้อมูล RFM

| Cluster RFM | Silhouette Score | Davies-Bouldin Index |
|---|---|---|
| 2 | 0.6884 | 0.4253 |
| 3 | 0.6163 | 0.5161 |
| 4 | 0.6270 | 0.5567 |
| 5 | 0.5964 | 0.5712 |
| 6 | 0.5866 | 0.4657 |
| 7 | 0.5526 | 0.4728 |
| 8 | 0.5632 | 0.4549 |
| 9 | 0.5545 | 0.4673 |
| 10 | 0.5331 | 0.5002 |

จากตารางที่ 8 จะเห็นว่า Cluster ที่มี Silhouette Score สูงและ Davies-Bouldin Index ต่ำสุด จะถือว่าเป็นการแบ่งกลุ่มที่ดีที่สุด โดย Cluster 2 มีผลการประเมินที่ดีที่สุดในทั้งสองดัชนีนี้

จากผลลัพธ์ของ Elbow Method และ Silhouette Score และ Davies-Bouldin Index ผู้วิจัยเลือกแบ่งกลุ่มข้อมูลเป็น 2 กลุ่ม ได้ผลลัพธ์ค่าเฉลี่ยของ Cluster ดังตาราง 9

ตาราง 9 แสดงค่าเฉลี่ยของ Cluster RFM

| Cluster RFM | Recency | Frequency | Monetary |
|---|---|---|---|
| Cluster 0 | 157.5706 | 82.7237 | 219,442.8836 |
| Cluster 1 | 1,130.4195 | 7.1667 | 8,698.3188 |

ผลการประเมินของการแบ่งกลุ่มที่ได้จากการทดลองแบ่งกลุ่มด้วย RFMBD ดังตาราง 10

ตาราง 10 แสดงผลการประเมินของการแบ่งกลุ่มของข้อมูล RFMBD

| Cluster RFMBD | Silhouette Score | Davies-Bouldin Index |
|---|---|---|
| 2 | 0.6147 | 0.6850 |
| 3 | 0.6029 | 0.5249 |
| 4 | 0.6148 | 0.5344 |
| 5 | 0.5366 | 0.6270 |
| 6 | 0.5430 | 0.6715 |
| 7 | 0.5179 | 0.7051 |
| 8 | 0.5008 | 0.6346 |
| 9 | 0.4832 | 0.7126 |
| 10 | 0.4723 | 0.6599 |

จากตารางที่ 10 จะเห็นว่า Cluster ที่มี Silhouette Score สูงสุด คือ Cluster 4 และ Davies-Bouldin Index ต่ำสุด คือ Cluster 3 การเลือกค่าที่เหมาะสมจึงต้องพิจารณาความเหมาะสมตามวัตถุประสงค์และลักษณะของข้อมูล

จากผลลัพธ์ของ Elbow Method และ Silhouette Score และ Davies-Bouldin Index ผู้วิจัยเลือกแบ่งกลุ่มข้อมูลเป็น 4 กลุ่ม ได้ผลลัพธ์ค่าเฉลี่ยของ Cluster ดังตาราง 11

ตาราง 11 แสดงค่าเฉลี่ยของ Cluster RFMBD

| Cluster RFMBD | Recency | Frequency | Monetary | Basket Size | Day_Type |
|---|---|---|---|---|---|
| Cluster 0 | 1,110.2286 | 6.7714 | 8,560.0576 | 1,450.2134 | 0 |
| Cluster 1 | 155.6100 | 67.0270 | 191,683.1485 | 3,261.9892 | 0 |
| Cluster 2 | 109.3636 | 154.1061 | 354,585.7849 | 3,480.6156 | 1 |
| Cluster 3 | 1,100.2619 | 7.5000 | 8,118.6285 | 1,160.1186 | 1 |

แบบจำลองที่ 1 ใช้วิธี K-Means clustering แบ่งได้เป็น 2 กลุ่ม โดยกลุ่มที่ 1 มีสมาชิก 333 คน กลุ่มที่ 2 มีสมาชิก 174 คน

แบบจำลองที่ 2 ใช้วิธี K-Means clustering แบ่งได้เป็น 4 กลุ่ม โดยกลุ่มที่ 1 มีสมาชิก 259 คน กลุ่มที่ 2 มีสมาชิก 140 คน กลุ่มที่ 3 มีสมาชิก 66 คน กลุ่มที่ 4 มีสมาชิก 42 คน

**ผลการเปรียบเทียบการแบ่งกลุ่มของแบบจำลอง**

ค่าผลลัพธ์ของตัวประเมิน Adjusted Rand Index (ARI) และ Normalized Mutual Information (NMI) เปรียบเทียบทั้งสองแบบจำลอง ได้ค่าดังตารางที่ 12

ตาราง 12 ค่า Adjusted Rand index และค่า Normalized Mutual Information

| ตัวประเมิน | ค่า |
|---|---|
| Adjusted Rand Index | 0.4846 |
| Normalized Mutual Information | 0.4987 |

ในบทที่ 4 นี้ ได้นำเสนอผลการศึกษา ผลการแบ่งกลุ่ม ทำ K-Means Clustering ด้วย RFM และ RFMBD และผลการเปรียบเทียบการแบ่งกลุ่ม โดยในบทที่ 5 จะมีการกล่าวถึงการสรุปผล อภิปรายผล และข้อเสนอแนะต่อไป

# บทที่ 5
# สรุปผลการวิจัย อภิปรายผลการวิจัย และข้อเสนอแนะ

## 5.1. สรุปผลการวิจัย

การวิเคราะห์แบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อสินค้าของร้านขายปลีกแห่งหนึ่งจากชุดข้อมูลลูกค้า 507 คน จัดเก็บข้อมูลในปี ค.ศ. 2019 ถึงต้นปี ค.ศ. 2023 ด้วยแบบจำลอง RFM และเพิ่มตัวแปรยอดการซื้อสินค้าแต่ละครั้ง (B) วันที่มาใช้บริการ (D) โดยใช้เทคนิค K-Means ในการทำ Clustering เพื่อแบ่งกลุ่มลูกค้า กำหนดจำนวนกลุ่มที่ต้องการโดยพิจารณาจำนวนกลุ่มที่เหมาะสมด้วยวิธี Elbow ค่าเฉลี่ยของ Silhouette Score และค่าเฉลี่ยของ Davies-Bouldin ได้ผลลัพธ์ดังนี้

1. ผลลัพธ์ที่ได้จากการแบ่งกลุ่มลูกค้าโดยใช้วิธี K-Means clustering ด้วยคุณลักษณะ Recency, Frequency, และ Monetary (RFM) แบ่งได้เป็น 2 กลุ่ม คือ

<u>กลุ่มที่ 1</u> มีสมาชิก 333 คน ลูกค้าในกลุ่มนี้มีค่า Recency (จำนวนวันที่ลูกค้าทำการซื้อครั้งล่าสุด) เฉลี่ยที่ 157.5706 มีค่า Frequency (ความถี่ในการซื้อสินค้า) เฉลี่ยที่ 82.7237 และค่า Monetary (ยอดค่าใช้จ่ายรวม) เฉลี่ยที่ 219,442.8836

<u>กลุ่มที่ 2</u> มีสมาชิก 174 คน ลูกค้าในกลุ่มนี้มีค่า Recency (จำนวนวันที่ลูกค้าทำการซื้อครั้งล่าสุด) เฉลี่ยที่ 1,130.4195 มีค่า Frequency (ความถี่ในการซื้อสินค้า) เฉลี่ยที่ 7.1667 และค่า Monetary (ยอดค่าใช้จ่ายรวม) เฉลี่ยที่ 8,698.3188

2. ผลลัพธ์ที่ได้จากการแบ่งกลุ่มลูกค้าโดยใช้วิธี K-Means clustering ด้วยคุณลักษณะ Recency, Frequency, และ Monetary (RFM) และการเพิ่มคุณลักษณะ Basket Size และ Day Type (RFMBD) แบ่งได้เป็น 4 กลุ่ม คือ

<u>กลุ่มที่ 1</u> มีสมาชิก 259 คน ลูกค้าในกลุ่มนี้มีค่า Recency (จำนวนวันที่ลูกค้าทำการซื้อครั้งล่าสุด) เฉลี่ยที่ 1,110.2286 มีค่า Frequency (ความถี่ในการซื้อสินค้า) เฉลี่ยที่ 6.7714 และค่า Monetary (ยอดค่าใช้จ่ายรวม) เฉลี่ยที่ 8,560.0576 มีค่า Basket Size (ยอดค่าใช้จ่ายต่อครั้ง) เฉลี่ยที่ 1,450.2134 และค่า Day Type (วันที่มาใช้บริการ) เป็นวันธรรมดา

กลุ่มที่ 2 มีสมาชิก 140 คน ลูกค้าในกลุ่มนี้มีค่า Recency (จำนวนวันที่ลูกค้าทำการซื้อครั้งล่าสุด) เฉลี่ยที่ 155.6100 มีค่า Frequency (ความถี่ในการซื้อสินค้า) เฉลี่ยที่ 67.0270 ค่า Monetary (ยอดค่าใช้จ่ายรวม) เฉลี่ยที่ 191,683.1485 มีค่า Basket Size (ยอดค่าใช้จ่ายต่อครั้ง) เฉลี่ยที่ 3,261.9892 และค่า Day Type (วันที่มาใช้บริการ) เป็นวันธรรมดา

กลุ่มที่ 3 มีสมาชิก 66 คน ลูกค้าในกลุ่มนี้มีค่า Recency (จำนวนวันที่ลูกค้าทำการซื้อครั้งล่าสุด) เฉลี่ยที่ 109.3636 มีค่า Frequency (ความถี่ในการซื้อสินค้า) เฉลี่ยที่ 154.1061 ค่า Monetary (ยอดค่าใช้จ่ายรวม) เฉลี่ยที่ 354,585.7849 มีค่า Basket Size (ยอดค่าใช้จ่ายต่อครั้ง) เฉลี่ยที่ 3,480.6156 และค่า Day Type (วันที่มาใช้บริการ) เป็นวันหยุด

กลุ่มที่ 4 มีสมาชิก 42 คน ลูกค้าในกลุ่มนี้มีค่า Recency (จำนวนวันที่ลูกค้าทำการซื้อครั้งล่าสุด) เฉลี่ยที่ 1,100.2619 มีค่า Frequency (ความถี่ในการซื้อสินค้า) เฉลี่ยที่ 7.5000 ค่า Monetary (ยอดค่าใช้จ่ายรวม) เฉลี่ยที่ 8,118.6285 มีค่า Basket Size (ยอดค่าใช้จ่ายต่อครั้ง) เฉลี่ยที่ 1,160.1186 และค่า Day Type (วันที่มาใช้บริการ) เป็นวันนหยุด

### 5.2. **อภิปรายผลการวิจัย**

งานวิจัยนี้ ได้มุ่งเน้นที่การศึกษาวิธีการแบ่งกลุ่มลูกค้าตามพฤติกรรมการซื้อของร้านขายปลีกแห่งหนึ่ง ด้วยแบบจำลอง RFM และมีการเพิ่มตัวแปรยอดการซื้อสินค้าแต่ละครั้ง วันที่มาใช้บริการ โดยใช้เทคนิค K-Means ในการทำ Clustering เพื่อแบ่งกลุ่มลูกค้า ได้แบ่งการทดลองเป็นแบบจำลอง 2 แบบ โดยแบบที่ 1 คือ การใช้ข้อมูล RFM และแบบที่ 2 คือการใช้ข้อมูล RFMBD พิจารณาจำนวนกลุ่มที่เหมาะสมด้วยวิธี Elbow ดูค่าของ Silhouette Score และค่าของ Davies-Bouldin Index และนำผลมาวิเคราะห์เปรียบเทียบระหว่างผลลัพธ์ของทั้ง 2 แบบจำลอง พบว่า

แบบจำลองที่ 1 การใช้ข้อมูล RFM ผลลัพธ์ของ Elbow Method ได้จำนวนเท่ากับ 4 กลุ่ม ค่าของ Silhouette Score อยู่ที่ 0.6884 ได้จำนวนเท่ากับ 2 กลุ่ม และค่าของ Davies-Bouldin Index อยู่ที่ 0.4253 ได้จำนวนเท่ากับ 2 กลุ่ม ผู้วิจัยเลือกแบ่งกลุ่มลูกค้าออกเป็น 2 กลุ่ม โดยแต่ละกลุ่มมีลักษณะดังนี้

กลุ่มที่ 1 มีจำนวน 333 คน ค่า Recency เฉลี่ยที่ 157.5706 หมายถึงลูกค้าในกลุ่มนี้มีแนวโน้มที่จะทำการซื้อสินค้าใหม่อีกครั้งในระยะเวลาไม่นานหลังจากการทำซื้อครั้งก่อน ค่า Frequency เฉลี่ยที่ 82.7237 แสดงให้เห็นว่าลูกค้าในกลุ่มนี้มักมีความถี่ในการทำการซื้อสินค้า

บ่อยมาก ส่วนค่า Monetary เฉลี่ยที่ 219,442.8836 บ่งบอกถึงมูลค่าการใช้จ่ายของลูกค้าในกลุ่มนี้ที่มีค่าสูง ซึ่งอาจแสดงถึงความสำคัญและการให้ความสำคัญต่อธุรกิจหรือสินค้าที่เสนอให้กับกลุ่มลูกค้านี้ จึงนิยามกลุ่มที่ 1 เป็น “กลุ่มลูกค้าซื้อเยอะ จ่ายหนัก”

กลุ่มที่ 2 มีจำนวน 174 คน ค่า Recency เฉลี่ยที่ 1,130.4195 หมายถึงลูกค้าในกลุ่มนี้มักจะทำการซื้อสินค้าน้อยลงหรือไม่ทำการซื้อเลยในระยะเวลาที่ผ่านมา ค่า Frequency เฉลี่ยที่ 7.1667 แสดงให้เห็นว่าลูกค้าในกลุ่มนี้มีแนวโน้มที่จะทำการซื้อสินค้าน้อยเมื่อเปรียบเทียบกับกลุ่มอื่น ค่า Monetary เฉลี่ยที่ 8,698.3188 แสดงให้เห็นว่าลูกค้าในกลุ่มนี้มักมีความสนใจในการซื้อสินค้ามูลค่าต่ำหรือจำนวนน้อย จึงนิยามกลุ่ม 2 เป็น “กลุ่มลูกค้าซื้อน้อย คิดนาน”

แบบจำลอง 2 การใช้ข้อมูล RFMBD ผลลัพธ์ของ Elbow Method ได้จำนวนเท่ากับ 4 กลุ่ม ค่าของ Silhouette Score อยู่ที่ 0.6148 ได้จำนวนเท่ากับ 4 กลุ่ม และค่าของ Davies-Bouldin Index อยู่ที่ 0.5249 ได้จำนวนเท่ากับ 3 กลุ่ม ผู้วิจัยเลือกแบ่งกลุ่มลูกค้าออกเป็น 4 กลุ่ม โดยแต่ละกลุ่มมีลักษณะดังนี้

กลุ่มที่ 1 มีจำนวน 259 ราย ค่า Recency เฉลี่ยที่ 1,110.2286 ค่าสูงมากแสดงถึงลูกค้าในกลุ่มนี้มักจะไม่ทำการซื้อสินค้าบ่อยๆ หรืออาจจะทำการซื้อสินค้าอย่างไม่สม่ำเสมอ ค่า Frequency เฉลี่ยที่ 6.7714 แสดงให้เห็นว่าลูกค้าในกลุ่มนี้มักจะทำการซื้อสินค้าไม่บ่อย ค่า Monetary เฉลี่ยที่ 8,560.0576 แสดงให้เห็นว่าลูกค้าในกลุ่มนี้มักมีมูลค่าการใช้จ่ายในการซื้อสินค้าน้อย ค่า Basket Size เฉลี่ยที่ 1,450.2134 แสดงถึงมูลค่าการใช้จ่ายเฉลี่ยที่ลูกค้าในกลุ่มนี้ใช้จ่ายในแต่ละครั้งที่ซื้อสินค้ามีค่าน้อย และค่า Day Type เป็นวันธรรมดา บ่งบอกถึงว่าลูกค้าในกลุ่มนี้มักมาใช้บริการในวันธรรมดา จึงนิยามกลุ่ม 1 เป็น “กลุ่มลูกค้ามาน้อย จ่ายน้อย ช้อปวันธรรมดา”

กลุ่มที่ 2 มีจำนวน 140 ราย ค่า Recency เฉลี่ยที่ 155.6100 แสดงให้เห็นว่าลูกค้า ในกลุ่มนี้มักทำการซื้อสินค้าอย่างสม่ำเสมอหรือบ่อยๆ ค่า Frequency เฉลี่ยที่ 67.0270 แสดงให้เห็นว่าลูกค้าในกลุ่มนี้มีมักมาใช้บริการซื้อสินค้าบ่อยมาก ค่า Monetary เฉลี่ยที่ 191,683.1485 แสดงให้เห็นว่าลูกค้าในกลุ่มนี้มีมูลค่าการใช้จ่ายในการซื้อสินค้าสูง ค่า Basket Size เฉลี่ยที่ 3,261.9892 แสดงถึงมูลค่าการใช้จ่ายเฉลี่ยที่ลูกค้าในกลุ่มนี้ใช้จ่ายในแต่ละครั้งที่ซื้อสินค้ามีค่ามาก และค่า Day Type เป็นวันธรรมดา บ่งบอกถึงว่าลูกค้าในกลุ่มนี้มักมาใช้บริการในวันธรรมดา จึงนิยามกลุ่มที่ 2 เป็น “กลุ่มลูกค้ามาบ่อย จ่ายหนัก ช้อปวันธรรมดา ”

กลุ่มที่ 3 มีจำนวน 66 ราย ค่า Recency เฉลี่ยที่ 109.3636 ค่า Recency แสดงให้เห็นว่าลูกค้าในกลุ่มนี้มักทำการซื้อสินค้าเป็นระยะเวลาสม่ำเสมอหรือบ่อยๆ ค่า Frequency เฉลี่ยที่ 154.1061 แสดงให้เห็นว่าลูกค้าในกลุ่มนี้มีมักมาใช้บริการซื้อสินค้าบ่อยมาก ค่า Monetary เฉลี่ยที่ 354,585.7849 แสดงให้เห็นว่าลูกค้าในกลุ่มนี้มีมูลค่าการใช้จ่ายในการซื้อสินค้าสูงมาก ค่า Basket Size เฉลี่ยที่ 3,480.6156 แสดงถึงมูลค่าค่าการใช้จ่ายเฉลี่ยที่ลูกค้าในกลุ่มนี้ใช้จ่ายในแต่ละครั้งที่ซื้อสินค้ามีค่ามาก และค่า Day Type เป็นวันหยุด บ่งบอกถึงว่าลูกค้าในกลุ่มนี้มักมาใช้บริการในวันที่เป็นวันหยุด จึงนิยามกลุ่มที่ 3 เป็น “กลุ่มลูกค้ามาบ่อย จ่ายหนัก ช้อปวันหยุด”

กลุ่มที่ 4 มีจำนวน 42 ราย ค่า Recency เฉลี่ยที่ 1,100.2619 แสดงถึงลูกค้าในกลุ่มนี้ มักไม่ทำการซื้อสินค้าอย่างสม่ำเสมอ ค่า Frequency เฉลี่ยที่ 7.500 แสดงให้เห็นว่าลูกค้าในกลุ่มนี้มักมาใช้บริการซื้อสินค้าไม่บ่อย ค่า Monetary เฉลี่ยที่ 8,118.6258 ค่า แสดงให้เห็นว่าลูกค้าในกลุ่มนี้มีมูลค่าการใช้จ่ายในการซื้อสินค้าน้อย ค่า Basket Size เฉลี่ยที่ 1,160.1186 ค่า Basket Size แสดงถึงมูลค่าการใช้จ่ายเฉลี่ยที่ลูกค้าในกลุ่มนี้ใช้จ่ายในแต่ละครั้งที่ซื้อสินค้ามีค่าน้อยและค่า Day Type เป็นวันหยุด บ่งบอกถึงว่าลูกค้าในกลุ่มนี้มักมาใช้บริการในวันที่เป็นวันหยุด จึงนิยามกลุ่มที่ 4 เป็น “กลุ่มลูกค้ามาน้อย จ่ายน้อย ช้อปวันหยุด”

การวิเคราะห์เปรียบเทียบผลของการแบ่งกลุ่มระหว่างแบบจำลองที่ 1 กับแบบจำลองที่ 2 ด้วยตัวประเมิน Adjusted Rand Index (ARI) และ Normalized Mutual Information (NMI) เป็นการวัดความคล้ายคลึงระหว่างการแบ่งกลุ่มของแบบจำลองทั้ง 2 แบบ มีค่าอยู่ในช่วงระหว่าง 0 ถึง 1 โดยค่าที่มีค่ามากกว่าหรือเข้าใกล้ 1 จะแสดงถึงความคล้ายคลึงระหว่างการแบ่งกลุ่มของทั้งสองแบบจำลองในระดับที่สูงขึ้น ดังนั้น ค่า Adjusted Rand Index เท่ากับ 0.4846 หมายถึงมีความสัมพันธ์ระหว่างการแบ่งกลุ่มของทั้งสองแบบจำลองในระดับที่พอใช้ได้ แต่ไม่สามารถบอกได้แน่ชัดว่าการแบ่งกลุ่มของทั้งสองแบบจำลองมีความคล้ายคลึงกันมากหรือน้อยกว่านี้ และค่า Normalized Mutual Information เท่ากับ 0.4987 หมายถึงการแบ่งกลุ่มของทั้งสองแบบจำลอง มีความคล้ายคลึงกันปานกลาง ซึ่งหมายความว่ามีการแบ่งกลุ่มที่สอดคล้องกันในระดับที่มีความสัมพันธ์กันเป็นไปได้

สำหรับการตัดสินใจที่จะเลือกใช้งานแบบจำลองใดที่มีเหมาะสม ขึ้นอยู่กับวัตถุประสงค์และความต้องการหรือปัญหาของธุรกิจ นอกจากนี้ยังขึ้นอยู่กับผลลัพธ์ที่ได้จากการวิเคราะห์และความเหมาะสมของข้อมูล พิจารณาดังนี้

1. พิจารณาจากความเรียบง่ายและความชัดเจนในการทำนายเป็นหลัก แบบจำลองที่ 1 อาจจะเป็นตัวเลือกที่ดี เนื่องจากมีจำนวนกลุ่มน้อยและสมดุลกัน ซึ่งทำให้ง่ายต่อการทำนายและตีความผลลัพธ์

2. พิจารณาจากความหลากหลายของกลุ่ม แบบจำลองที่ 2 มีจำนวนกลุ่มมากกว่าและมีความแตกแตกต่างหลากหลายของกลุ่มมากขึ้น อาจจะช่วยให้เราเข้าใจลักษณะและพฤติกรรมของกลุ่มลูกค้าได้ดีขึ้น

3. พิจารณาการใช้คุณลักษณะเพิ่มเติม ถ้าต้องการความละเอียดสำหรับการวิเคราะห์มากขึ้น แบบจำลองที่ 2 อาจเหมาะกว่า เนื่องจากมีการแบ่งกลุ่มลูกค้าเป็นจำนวนมากกว่า และมีข้อมูลเพิ่มเติม เช่น Basket Size และ Day Type ซึ่งอาจช่วยให้เข้าใจลึกซึ้งยิ่งขึ้นเกี่ยวกับพฤติกรรมและลักษณะของกลุ่มลูกค้าได้

## 5.3 **ข้อเสนอแนะ**

เพื่อเป็นแนวทางในการพัฒนางานวิจัยในอนาคต อาจประยุกต์ใช้วิธีการนี้ และสามารถปรับปรุงแบบจำลองในการสำรวจเพิ่มเติมได้โดยเพิ่มตัวแปรที่มีผลต่อพฤติกรรมการซื้อของลูกค้าด้านอื่น ๆ เช่น ข้อมูลของลูกค้า ไลฟ์สไตล์ ความสนใจ และงานอดิเรกของลูกค้า ศึกษาทัศนคติแรงจูงใจและปัจจัยที่มีอิทธิพลต่อพฤติกรรมการซื้อของลูกค้า และการใช้ตัวประเมินประสิทธิภาพเพิ่มเติมเพื่อเปรียบเทียบการเลือกจำนวนกลุ่มที่เหมาะสม

# บรรณานุกรม


Amrulloh, K., Pudjiantoro, T. H., Sabrina, P. N., & Hadiana, A. I. (2022). Comparison Between Davies-Bouldin Index and Silhouette Coefficient Evaluation Methods in Retail Store Sales Transaction Data Clusterization Using K-Medoids Algorithm. Proceedings of the 3rd South American International Industrial Engineering and Operations Management Conference, Asuncion, Paraguay.

Dedi, Dzulhaq, M. I., Sari, K. W., Ramdhan, S., Tullah, R., & Sutarman. (2019, 16-17 Oct. 2019). Customer Segmentation Based on RFM Value Using K-Means Algorithm. 2019 Fourth International Conference on Informatics and Computing (ICIC),

Gustriansyah, R., Suhandi, N., & Antony, F. (2020). Clustering optimization in RFM analysis based on k-means. *Indonesian Journal of Electrical Engineering and Computer Science*, *18*, 470-477. https://doi.org/10.11591/ijeecs.v18.i1.pp470-477

Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018, 21-22 Dec. 2018). Customer Segmentation using K-means Clustering. 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS),

Kodinariya, T. M., & Makwana, D. P. R. (2013). Review on determining number of Cluster in K-Means Clustering. *International Journal of Advance Research in Computer Science and Management Studies 1*(6). https://www.researchgate.net/publication/313554124

Malini, P. A., & Patil, M. (2022). RFM model for customer purchase behavior using K-Means algorithm. *Journal of King Saud University – Computer and Information Sciences*, *34*(5), 1785-1792. https://www.sciencedirect.com/science/article/pii/S1319157819309802?via%3Dihub

Rattanatat. (2562, 15 มีนาคม 2562). การทำเหมืองข้อมูล (*Data Mining)*. https://dol.dip.go.th/th/category/2019-02-08-08-57-30/2019-03-15-08-49-57

Sirilak Ketchaya. (ม.ป.ป.). การทำเหมืองข้อมูลด้วยเทคนิค Clustering : K-means. https://elsci.ssru.ac.th/sirilak_ke/pluginfile.php/58/mod_resource/content/1/DM_Ch

apter5.pdf

ไกรศักดิ์ เกษร. (2564). วิทยาศาสตร์ข้อมูล (*Data Science)*. ภาควิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ. https://csit.nu.ac.th/kraisak/ds/ds/chapter08/Chapter08.pdf

ณรรฐคุณ วิรุฬห์ศรี และคณะ. (2565). Clustering Customers Using Their In-Depth Buying Behavior: A Pet Food Manufacturing Company Case Study. วารสารวิทยาศาสตร์ลาดกระบัง. https://li01.tci-thaijo.org/index.php/science_kmitl/article/download/250784/173839/918760

ปรีดี นุกูลสมปรารถนา. (2563). ค้นหาลูกค้าชั้นดีด้วย *RFM Framework*. https://www.popticles.com/business/find-the-best-customer-with-rfm-framework/

ภคภูมิ สารพัฒน์. (2563, 8 มีนาคม 2563). อีกขั้นของ *k-means algorithm* ที่สามารถแบ่งกลุ่มข้อมูลได้ทุกประเภท. https://bigdata.go.th/big-data-101/k-means-algorithm-for-clustering-large-data-sets-with-categorical-values/

ภรนนท์ แดงขาว และคณะ. (2562). การแบ่งกลุ่มผู้ค้าปลีกที่ซื้อกางเกงแฟชั่นผู้หญิงในตึกโบ๊เบ๊ทาวเวอร์ โดยใช้พฤติกรรมการซื้อและด้านความสำคัญของส่วนประสมทางการตลาด. การประชุมนำเสนอผลงานวิจัยระดับบัณฑิตศึกษา, *12*, 639-646 https://rsujournals.rsu.ac.th/index.php/rgrc/article/download/700/468/

ภัครพล อาจอาษา. (2564). การวิเคราะห์คุณภาพน้ำด้วยเทคนิคการจัดกลุ่มข้อมูล. http://olarik.it.msu.ac.th/wp-content/uploads/2021/10/complete_62011284504.pdf

วิทยา พรพัชรพงศ์. (2549, 2012). การจัดกลุ่มลูกค้าและการทำเหมืองข้อมูล (*Customer Segmentation and Data Mining)*. GotoKnow. https://www.gotoknow.org/posts/56616

สิรินี ว่องวิไลรัตน์. (2560). พฤติกรรมผู้บริโภค. ศูนย์หนังสือนอร์ทเทิร์น. http://online.northern.ac.th/moodle/pluginfile.php/16269/mod_resource/content/2/%E0%B8%9E%E0%B8%A4%E0%B8%95%E0%B8%B4%E0%B8%81%E0%B8%A3%E0%B8%A3%E0%B8%A1%E0%B8%9C%E0%B8%B9%E0%B9%89%E0%B8%9A%E0%B8%A3%E0%B8%B4%E0%B9%82%E0%B8%A0%E0%B8%84.pdf

สุจิรา ไชยกุสินธุ์. (2559). การทำเหมืองข้อมูลเพื่อการพยากรณ์ธุรกิจ. แบบฟอร์มแนวปฏิบัติที่ดี (*Good Practice)*. https://bus.rmutp.ac.th/km/wp-content/uploads/2016/02/datamining.pdf

สุทธิพงศ์ ผ่องแผ้ว. (2555). STUDY OF THE SEED CLUSTERING PROCESS USING STRUCTURAL DATA. http://ir-ithesis.swu.ac.th/dspace/handle/123456789/49

อังคาร ปริญญาชัยศักดิ์ และคณะ. (2022). การจัดกลุ่มนักศึกษาสำหรับวางแผนอบรมเสริมความรู้ก่อนจบการศึกษา สาขาคอมพิวเตอร์ศึกษา มหาวิทยาลัยราชภัฏบ้านสมเด็จเจ้าพระยา ด้วยเทคนิคการเรียนรู้ของเครื่อง. การประชุมวิชาการนำเสนอผลงานวิจัยระดับชาติและนานาชาติ, *15*. http://journalgrad.ssru.ac.th/index.php/8thconference/article/view/2689

เอกปรียา ใบสนิ. (2563). THE STUDY OF CUSTOMER SEGMENTATION BY USING RFM MODEL AND TEXT ANALYTICS. . http://ir-ithesis.swu.ac.th/dspace/handle/123456789/1083

# ภาคผนวก

รายละเอียดการดูวันที่และเลือกใช้ในชุดข้อมูล

```
[133] # วันที่เริ่มต้นใบแจ้งหนี้ในชุดข้อมูล
      start_invoice_date = df['Date'].min()
      start_invoice_date

      Timestamp('2019-01-02 00:00:00')

[134] # วันที่ใบแจ้งหนี้ล่าสุดในชุดข้อมูล
      latest_invoice_date = df['Date'].max()
      latest_invoice_date

      Timestamp('2023-03-25 00:00:00')
```

วันสุดท้ายคือ 2023-03-25 สำหรับการทำธุรกรรม ดังนั้นจะใช้วันที่ล่าสุดสำหรับการวัดประสิทธิภาพเป็น 2023-03-25 ในการคำนวณ

```
[135] df_part = df.copy() # creating copy
      df_part.set_index('Date', inplace= True) # setting Date as index
      df_part = df_part.loc[:'2023-03-25'] # slicing the data

[136] df_part.reset_index(inplace= True)

[137] # ดูข้อมูลของ df_part
      df_part.head()
```

รายละเอียดการทำ Scale ข้อมูลโดยใช้ MinMaxScaler

```
from sklearn.preprocessing import MinMaxScaler

# ทำการ Scale ข้อมูลใช้ Min-Max Scaling
scaler = MinMaxScaler()
RFM_scaled = scaler.fit_transform(RFM)

# แสดง DataFrame ที่ Scale แล้ว แสดงผลลัพธ์เป็นทศนิยม 4 ตำแหน่ง
RFM_scaled.round(4)
```

```
array([[0.0045, 0.0127, 0.0683],
       [0.477 , 0.0031, 0.0046],
       [0.0337, 0.0091, 0.0117],
       ...,
       [0.085 , 0.001 , 0.0036],
       [0.0078, 0.0171, 0.0645],
       [0.5282, 0.077 , 0.0804]])
```

```
from sklearn.preprocessing import MinMaxScaler

# ทำการ Scale ข้อมูลใช้ Min-Max Scaling
scaler = MinMaxScaler()
RFMBD_scaled = scaler.fit_transform(RFMBD)

# แสดง DataFrame ที่ Scale แล้ว แสดงผลลัพธ์เป็นทศนิยม 4 ตำแหน่ง
RFMBD_scaled.round(4)
```

```
array([[0.0045, 0.0127, 0.0683, 0.3455, 0.    ],
       [0.477 , 0.0031, 0.0046, 0.0795, 0.    ],
       [0.0337, 0.0091, 0.0117, 0.0717, 0.    ],
       ...,
       [0.085 , 0.001 , 0.0036, 0.1768, 0.    ],
       [0.0078, 0.0171, 0.0645, 0.2391, 0.    ],
       [0.5282, 0.077 , 0.0804, 0.0571, 0.    ]])
```

รายละเอียดการวิเคราะห์หาจำนวนกลุ่มที่เหมาะสม (K)

```python
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
from yellowbrick.cluster import KElbowVisualizer

# สร้างรายการเปล่าสำหรับเก็บค่า SSE
sse = []

# ลูปผ่านจำนวนคลัสเตอร์ต่าง ๆ
for k in range(2, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    kmeans.fit(scaled_data1)
    # คำนวณค่า SSE และเก็บไว้ในรายการ
    sse.append(kmeans.inertia_)

# พล็อตกราฟ Elbow Method
plt.figure(figsize=(8, 6))
plt.plot(range(2, 11), sse, marker='o')
plt.xlabel('Number of Clusters (K)')
plt.ylabel('SSE (Sum of Squared Errors)')
plt.title('Elbow Method')
plt.xticks(range(2, 11))
plt.grid(True)
plt.show()
```

```python
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score
import matplotlib.pyplot as plt

# สร้างรายการเปล่าสำหรับเก็บค่า Silhouette Score
silhouette_scores = []

# ลูปผ่านจำนวนคลัสเตอร์ต่าง ๆ
for k in range(2, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    labels = kmeans.fit_predict(scaled_data1)
    silhouette_scores.append(silhouette_score(scaled_data1, labels))

# พล็อตกราฟ Silhouette Score ต่อจำนวนคลัสเตอร์
plt.figure(figsize=(8, 6))
plt.plot(range(2, 11), silhouette_scores, marker='o')
plt.xlabel('Number of Clusters (K)')
plt.ylabel('Silhouette Score')
plt.title('Silhouette Score')
plt.xticks(range(2, 11))
plt.grid(True)
plt.show()
```

```python
from sklearn.cluster import KMeans
from sklearn.metrics import davies_bouldin_score
import matplotlib.pyplot as plt

# สร้างรายการเปล่าสำหรับเก็บค่า Davies-Bouldin Index
davies_bouldin_scores = []

# ลูปผ่านจำนวนคลัสเตอร์ต่าง ๆ
for k in range(2, 11):
    kmeans = KMeans(n_clusters=k, random_state=42)
    labels = kmeans.fit_predict(scaled_data1)
    davies_bouldin_scores.append(davies_bouldin_score(scaled_data1, labels))

# พล็อตกราฟ Davies-Bouldin Index ต่อจำนวนคลัสเตอร์
plt.figure(figsize=(8, 6))
plt.plot(range(2, 11), davies_bouldin_scores, marker='o')
plt.xlabel('Number of Clusters (K)')
plt.ylabel('Davies-Bouldin Index')
plt.title('Davies-Bouldin Index Method')
plt.xticks(range(2, 11))
plt.grid(True)
plt.show()
```

รายละเอียดจำนวนผลลัพธ์ของแต่ละคลัสเตอร์

```
# นับจำนวนของแต่ละ cluster
cluster_counts = RFMBD['Cluster_RFMBD'].value_counts()

# แสดงผลลัพธ์
cluster_counts
```

```
Cluster_RFMBD
1    259
0    140
2     66
3     42
Name: count, dtype: int64
```

```
# แสดงค่าเฉลี่ยของแต่ละ cluster
cluster_means_rfmbd = RFMBD.groupby('Cluster_RFMBD').mean()

# แสดงผลลัพธ์
cluster_means_rfmbd.round(4)
```

| Cluster_RFMBD | Recency | Frequency | Monetary | Basket Size | Day_Type |
|---|---|---|---|---|---|
| 0 | 1110.2286 | 6.7714 | 8560.0576 | 1450.2134 | 0.0 |
| 1 | 155.6100 | 67.0270 | 191683.1485 | 3261.9892 | 0.0 |
| 2 | 109.3636 | 154.1061 | 354585.7849 | 3480.6156 | 1.0 |
| 3 | 1100.2619 | 7.5000 | 8118.6285 | 1160.1186 | 1.0 |

```
from sklearn.metrics import adjusted_rand_score

# ข้อมูลการจัดกลุ่มของแบบจำลองที่ 1
labels_modelRFM = [0] * 333 + [1] * 174

# ข้อมูลการจัดกลุ่มของแบบจำลองที่ 2
labels_modelRFMBD = [0] * 259 + [1] * 140 + [2] * 66 + [3] * 42

# คำนวณ Adjusted Rand Index (ARI)
ari = adjusted_rand_score(labels_modelRFM, labels_modelRFMBD)
print("Adjusted Rand Index (ARI):", ari)
```

```
Adjusted Rand Index (ARI): 0.4846114477549461
```

```
from sklearn.metrics.cluster import normalized_mutual_info_score

# สร้างรายการของการจัดกลุ่มสำหรับแบบจำลองที่ 1
labels_modelRFM = [0] * 333 + [1] * 174

# สร้างรายการของการจัดกลุ่มสำหรับแบบจำลองที่ 2
labels_modelRFMBD = [0] * 259 + [1] * 140 + [2] * 66 + [3] * 42

# คำนวณ Normalized Mutual Information (NMI)
nmi = normalized_mutual_info_score(labels_modelRFM, labels_modelRFMBD)
print("Normalized Mutual Information (NMI):", nmi)
```

```
Normalized Mutual Information (NMI): 0.4987214421458666
```

# ประวัติผู้เขียน