การจัดกลุ่มลูกค้าด้วยเทคนิค K-mean clustering

1. Introduction

แหล่งที่มาของข้อมูล เป็นข้อมูลทางการตลาด ที่เก็บจากข้อมูลทั่วไปของลูกค้า นำมาใช้จัดกลุ่มลูกค้า เพื่อช่วยให้ทราบถึงลักษณะและพฤติกรรมของลูกค้าที่คล้ายกัน เพื่อการส่งข้อเสนอพิเศษหรือโปรโมชั่นที่เหมาะสมกับลูกค้าแต่ละกลุ่ม ประกอบไปด้วยตัวแปร ทั้งหมด 16 ตัวแปร ดังนี้

ลำดับ	ชื่อตัวแปร	ความหมาย
1	ID	รหัสลูกค้า
2	Year_Birth	ปีเกิดของลูกค้า
3	Education	ระดับการศึกษาของลูกค้า
4	Marital	สถานภาพสมรสของลูกค้า
5	Income	รายได้ประจำปีของครัวเรือนของลูกค้า
6	Kidhome	จำนวนเด็กในครัวเรือนของลูกค้า
7	Teenhome	จำนวนวัยรุ่นในครัวเรือนของลูกค้า
8	DtCustomer	วันที่ลูกค้าลงทะเบียนกับบริษัท
9	Recency	จำนวนวันตั้งแต่การซื้อของครั้งล่าสุด
10	MntWines	จำนวนเงินที่ใช้ในการซื้อไวน์ในช่วง 2 ปีที่ผ่านมา
11	MntFruits	จำนวนเงินที่ใช้ในการซื้อผลไม้ในช่วง 2 ปีที่ผ่านมา
12	MntMeatProducts	จำนวนเงินที่ใช้ในการซื้อผลิตภัณฑ์จากเนื้อสัตว์ในช่วง 2 ปีผ่านมา
13	MntFishProducts	จำนวนเงินที่ใช้ในการซื้อผลิตภัณฑ์จากปลาในช่วง 2 ปีผ่านมา
14	MntSweetProducts	จำนวนเงินที่ใช้ในการซื้อผลิตภัณฑ์ของหวานในช่วง 2 ที่ผ่านมา
15	MntGoldProds	จำนวนเงินที่ใช้ในการซื้อผลิตภัณฑ์ทองในช่วง 2 ปีที่ผ่านมา
16	Response	1 หากลูกค้ายอมรับข้อเสนอในแคมเปญล่าสุด, 0 หากไม่ยอมรับ

2. Method วิธีการที่ใช้จัดการกับข้อมูล

2.1 Data Preparation

การเตรียมข้อมูลสำหรับการวิเคราะห์ข้อมูล โดยทำการตรวจสอบข้อมูลเพื่อทราบถึงลักษณะของข้อมูล ตรวจสอบความสมบูรณ์ของข้อมูลให้พร้อมสำหรับการวิเคราะห์ จากการจัดรูปแบบข้อมูลให้ถูกต้อง การตรวจสอบข้อมูลที่ซ้ำกัน การตรวจสอบข้อมูลที่เป็นช่องว่าง

2.2 Preprocess

เป็นกระบวนการตรวจสอบและการแก้ไข รายการข้อมูลที่ไม่ถูกต้องจากชุดข้อมูล การทำ Data Cleaning เช่น การปรับรูปแบบของวันที่ การเติมข้อมูลลงในช่องว่าง หรือการสร้างคุณลักษณะใหม่ให้เหมาะสมกับการวิเคราะห์ข้อมูล การจัดการกับค่าที่เป็น Outlier คือค่าที่อยู่นอกขอบเขตของข้อมูล มีแนวทางการแก้ไข โดยลบข้อมูลชุดนั้นออก

2.3 Learning Techniques

เลือกใช้เทคนิคการเรียนรู้ด้วย K-mean clustering เพื่อการจัดกลุ่มลูกค้าให้เหมาะสมกับลักษณะและพฤติกรรมของลูกค้าที่คล้ายกัน โดยใช้หลักการของ Elbow Method เป็นวิธีที่ใช้ประเมินค่า K ที่เหมาะสมที่สุด โดยการรัน K-means clustering ด้วยค่า K ตั้งแต่ 1 ถึง K ที่สนใจ จากนั้นวัดความคลาดเคลื่อนของข้อมูลจาก centroids ของกลุ่มแต่ละกลุ่ม ซึ่งความคลาดเคลื่อนจะลดลงเรื่อย ๆ ดังนั้นจำนวนกลุ่มที่ดีที่สุดจะขึ้นอยู่กับค่าความคลาดเคลื่อนที่มีการเปลี่ยนแปลงสูง ๆ

3. Experimental Results

3.1 ขนาดของข้อมูล

จำนวนชุดข้อมูลทั้งหมดที่ใช้ในการวิเคราะห์ คือ 2,240 ชุดข้อมูล และมีจำนวนคุณสมบัติทั้งหมด 5 คุณสมบัติ ได้แก่ ระดับการศึกษา (Education) สถานภาพสมรส (Marital) ระดับรายได้ประจำปีของครัวเรือน (Income_level) ผลรวมของจำนวนเด็กและจำนวนวัยรุ่นในครัวเรือน (Total number of children) และสถานการณ์ซื้อ (Recency)

3.2 การออกแบบการทดลอง

การออกแบบการทดลอง K-Means Clustering ทำให้สามารถทดสอบและวิเคราะห์การจัดกลุ่มข้อมูลได้อย่างมีประสิทธิภาพ ดังนี้:

1. Collection การคัดเลือกข้อมูลให้เหมาะสมกับการวิเคราะห์ ในที่นี้เลือกใช้ข้อมูลทางการตลาด ที่เก็บจากข้อมูลทั่วไปของลูกค้า เพื่อการจัดกลุ่มลูกค้าให้เหมาะสมกับข้อเสนอทางธุรกิจ
2. Understanding การเข้าใจข้อมูล โดยตรวจสอบประเภทข้อมูลเบื้องต้นและความสมบูรณ์ของข้อมูล

3. Cleaning ทำการจัดการกับข้อมูลที่หายไปหรือขาดหาย เพิ่ม Features ที่เหมาะสมกับการวิเคราะห์ข้อมูล การตรวจสอบและจัดการกับข้อมูลที่ออกนอกขอบเขตปกติ และทำการตรวจสอบความสัมพันธ์ระหว่างตัวแปร
4. Modeling การสร้างแบบจำลองจากการเลือกตัวแปรที่เหมาะสมสำหรับ K-Means Clustering จากหลักการ Elbow method เพื่อเลือกจำนวน cluster ที่เหมาะสม
5. Conclusion การสรุปผล

3.3 ผลการทดลอง

1. ทำความเข้าใจข้อมูล จากข้อมูลทั้งหมด พบว่า มีจำนวนข้อมูล 2,240 ชุดข้อมูล แต่มีเพียงตัวแปร Response ที่มี 2,216 ข้อมูล ซึ่งอาจหมายการพบข้อมูลว่างที่ตัวแปรดังกล่าว มีตัวแปรทั้งหมด 16 ตัวแปร ข้อมูลลักษณะ Numeric Data 13 ตัวแปร และข้อมูลลักษณะ categorical data 3 ตัวแปร

```
Data columns (total 16 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           2240 non-null   int64
 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  Response         2216 non-null   float64
dtypes: float64(1), int64(12), object(3)
```

2. ตรวจสอบค่าที่เป็นช่องว่าง พบว่าส่วนใหญ่ไม่พบข้อมูลที่เป็นว่าง เว้นแต่ตัวแปร Response ที่พบข้อมูลที่เป็นช่องว่างทั้งหมด 24 ข้อมูล

```
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
Response            24
dtype: int64
```

3. ตรวจสอบชุดข้อมูลที่อาจจะเหมือนกัน พบว่า ไม่มีชุดข้อมูลใดเลยที่มีการเก็บข้อมูลเหมือนกันทุกคอลัมน์

4. Data Cleaning

- ปรับรูปแบบของคอลัมน์วันเกิด จาก 4/9/2012 เป็น 2012-04-09
- กรอกค่าที่หายไปของข้อมูล Response ด้วยค่าที่อยู่ในแถวถัดไปของข้อมูล
- เพิ่ม Features ที่เหมาะสมกับการวิเคราะห์ข้อมูล

  - อายุ สามารถคำนวณได้จาก วันเกิดของลูกค้า เทียบกับ ปี 2023

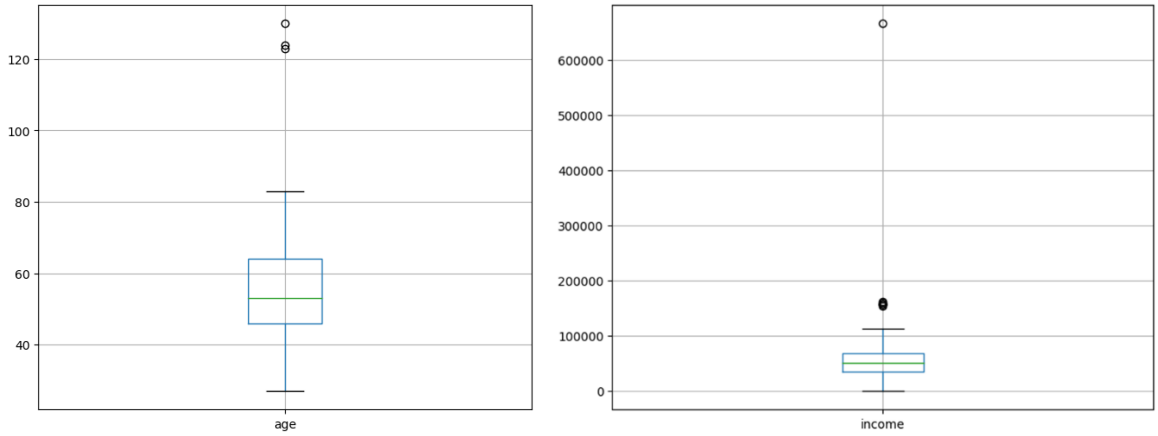
  - ระดับรายได้ แบ่งออกเป็น 3 ระดับ คือ รายได้ต่ำ มีค่าน้อยกว่า 50,000 รายได้ปานกลาง มีค่าอยู่ระหว่าง 50,001 – 150,000 และรายได้สูง มีค่ามากกว่า 150,001

  - จำนวณเด็กรวมในครัวเรือน คำนวณจากจำนวนเด็กรวมกับจำนวนวัยรุ่นในครัวเรือน

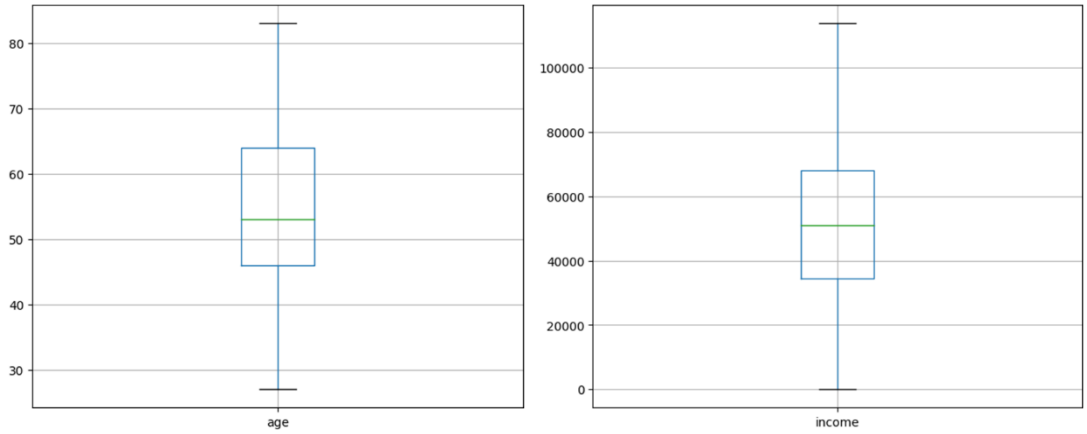
  - ยอดซื้อรวม เป็นผลรวมของจำนวนเงินในการซื้อสินค้าต่าง ๆ ได้แก่ ไวน์ ผลไม้ เนื้อสัตว์ ปลา ของหวาน และทองคำ

5. Distribution

- ตรวจสอบการกระจายตัวของข้อมูลด้วย Boxplots ของตัวแปรอายุ และตัวแปรรายได้ พบว่า มีข้อมูลที่อยู่นอกขอบเขตอยู่ ซึ่งตัวแปรรายได้มีค่าอยู่นอกขอบเขตมากกว่าตัวแปรอื่น

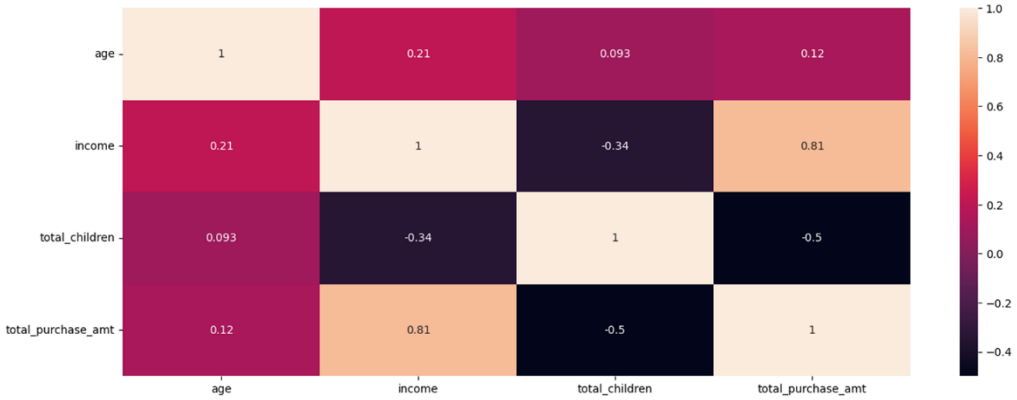


- ลบชุดข้อมูลที่เป็น Outlier



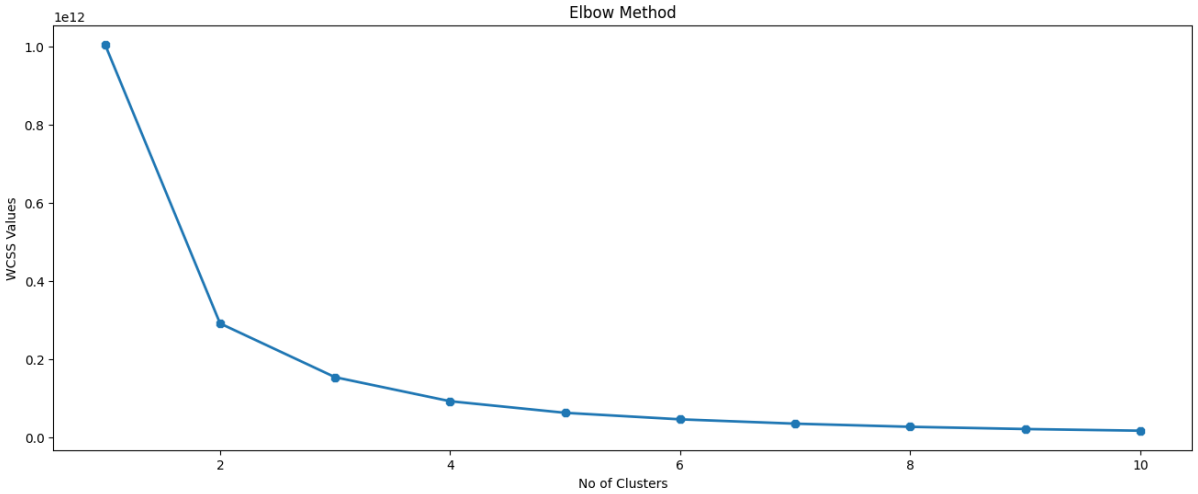
6. Correlation

ค่าสหสัมพันธ์ (Correlation) เป็นตัววัดที่บ่งบอกถึงความสัมพันธ์ระหว่างตัวแปรที่เป็นตัวเลข มีค่าอยู่ระหว่าง -1 ถึง 1 ค่า +1 หมายถึง มีความสัมพันธ์กันอย่างแข็งแรงทางบวก ค่า -1 หมายถึง มีความสัมพันธ์กันอย่างแข็งแรงทางลบ ค่า 0 หมายถึง ไม่มีความสัมพันธ์กัน จากตารางค่าสหสัมพันธ์ พบว่า ความสัมพันธ์ระหว่างรายได้ และ ยอดซื้อรวม มีค่าเท่ากับ 0.81 แสดงถึงมีความสัมพันธ์กันอย่างแข็งแรงทางบวก ในทางกลับกัน ระดับความสัมพันธ์ระหว่างตัวแปรอื่น ๆ อยู่ในระดับต่ำ คือมีค่าความสัมพันธ์น้อยกว่า 0.60



7. Model Development

หลักการ Elbow method เป็นเทคนิคที่นิยมใช้ในการเลือกจำนวน cluster ที่เหมาะสมในการทำ K-Means clustering หรือวิธีการแบ่งกลุ่มข้อมูลอื่น ๆ ที่ต้องระบุจำนวนกลุ่ม (หรือ K) ล่วงหน้า โดยใช้กราฟของค่าส่วนเบี่ยงเบนคำนวณ (Inertia) หรือค่า SSE (Sum of Squared Errors) ของแต่ละจำนวน cluster ที่เทียบกัน จุดที่เหมาะสมของจำนวน cluster ที่ให้ Inertia ลดลงอย่างมีนัยสำคัญ และต่ำที่สุดโดยไม่สร้างจำนวน cluster มากเกินไปที่ไม่จำเป็น การเลือกจำนวน cluster ที่เหมาะสมจะช่วยให้ได้โมเดลที่มีประสิทธิภาพและการจัดกลุ่มที่มีความหมาย จุดที่เหมาะสมสำหรับจำนวน cluster ที่สังเกตได้จากกราฟ คือ จำนวน 3 cluster เนื่องจากเปรียบเทียบจากอัตราการเปลี่ยนแปลงค่า SSE ระหว่างจำนวน 2 cluster กับจำนวน 3 cluster มีค่าสูง เมื่อเทียบกับจำนวน cluster ตั้งแต่ 4 cluster ขึ้นไปมีอัตราการเปลี่ยนแปลงค่า SSE เพียงเล็กน้อย ดังนั้น การเลือกใช้จำนวน cluster เป็น 3 cluster เป็นจุดที่เหมาะสม

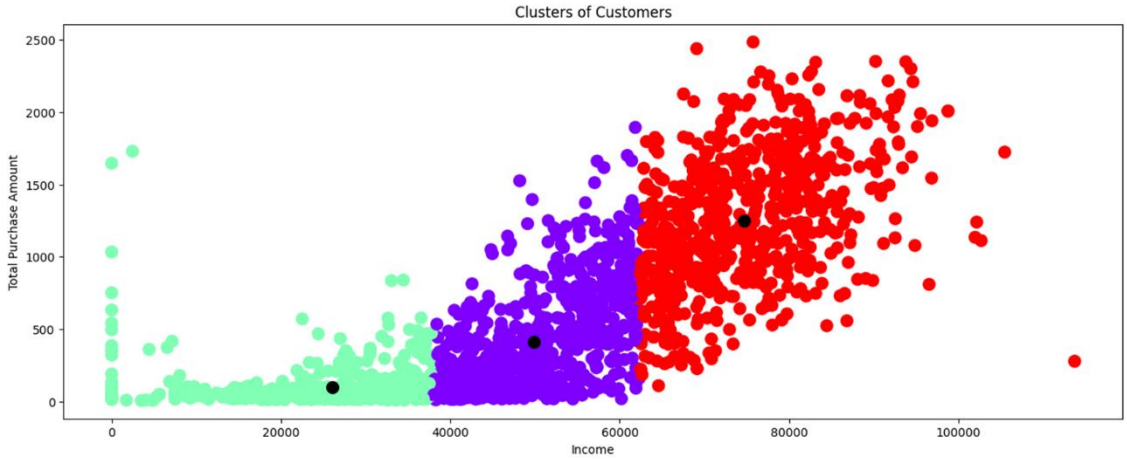


4. Conclusion

การใช้ K-Means Clustering เป็นเทคนิคที่นิยมในการแบ่งกลุ่มข้อมูลเพื่อทำนายหรือทำความเข้าใจลักษณะของกลุ่มต่าง ๆ ภายในข้อมูล จากกราฟสามารถสรุปได้ดังนี้

Light Blue Cluster	แทนลูกค้าที่มีระดับรายได้ต่ำและยอดซื้อน้อย
Red Cluster	แทนลูกค้าที่มีระดับรายได้ปานกลางและยอดซื้อปานกลางซึ่งเป็นกลุ่มที่เป็นเป้าหมายและต้องการให้มีการโฆษณาหรือการส่งเสริมการขาย.
Violet Cluster	แทนลูกค้าที่มีระดับรายได้สูงและยอดซื้อสูง

การจัดกลุ่มลูกค้าด้วย K-Means Clustering ช่วยให้ธุรกิจสามารถเข้าใจลักษณะของกลุ่มลูกค้าและวางแผนกลยุทธ์การตลาดที่เหมาะสม เช่นการปรับเป้าหมายการโฆษณาหรือการส่งเสริมการขายเพื่อเข้าถึงลูกค้าในกลุ่มที่มีภาพรวมและความต้องการที่คล้ายกัน


Clusters of Customers
Total Purchase Amount
Income
2500
2000
1500
1000
500
0
0
20000
40000
60000
80000
100000

สรุปผล

ผลลัพธ์จากการทดลอง K-means Clustering สามารถนำมาสรุปได้ดังนี้:

การแบ่งกลุ่มลูกค้า:

K-means Clustering ได้ทำการแบ่งกลุ่มลูกค้าออกเป็นกลุ่มต่าง ๆ โดยใช้ข้อมูลทางทะเบียนและพฤติกรรมการซื้อของลูกค้า
ลักษณะของ Cluster Centers:

ตำแหน่งของ Cluster Centers ช่วยในการแบ่งแยกกลุ่มและทำความเข้าใจลักษณะของกลุ่มลูกค้าแต่ละกลุ่ม
ค่า Silhouette Score หรือ Inertia:

การใช้ Silhouette Score หรือ Inertia เป็นตัววัดประสิทธิภาพของ Clustering Algorithm ช่วยในการประเมินความเหมาะสมของจำนวน Cluster ที่ถูกต้อง
การวิเคราะห์และอภิปรายผล:

การวิเคราะห์ Cluster Centers และผลลัพธ์ของ Clustering Algorithm ช่วยในการอธิบายลักษณะของกลุ่มลูกค้าแต่ละกลุ่ม
การวิเคราะห์ Silhouette Score หรือ Inertia ช่วยในการทำความเข้าใจความถูกต้องของการแบ่งกลุ่ม
การใช้ประโยชน์ในการตลาด:

ผลลัพธ์ที่ได้จากการทดลองนี้สามารถนำไปใช้ในการปรับปรุงกลยุทธ์การตลาด, การทำกิจกรรมตลาด, และการสร้างข้อเสนอที่เหมาะสมกับแต่ละกลุ่มลูกค้า
สรุปท้าย:
ผลลัพธ์จากการทดลอง K-means Clustering จะช่วยให้เข้าใจลักษณะของกลุ่มลูกค้า, ทำนายพฤติกรรมการซื้อ, และทำให้สามารถปรับปรุงกลยุทธ์การตลาดให้มีประสิทธิภาพมากยิ่งขึ้นในอนาคต. การใช้เทคนิคการแบ่งกลุ่มนี้จะช่วยในการทำธุรกรรมในทางธุรกิจและเพิ่มประสิทธิภาพในการบริหารจัดการลูกค้า.