Mini-Project: Marketing Campaign

Data Set: Customer Personality Analysis (marketing_campaign.xlsx)

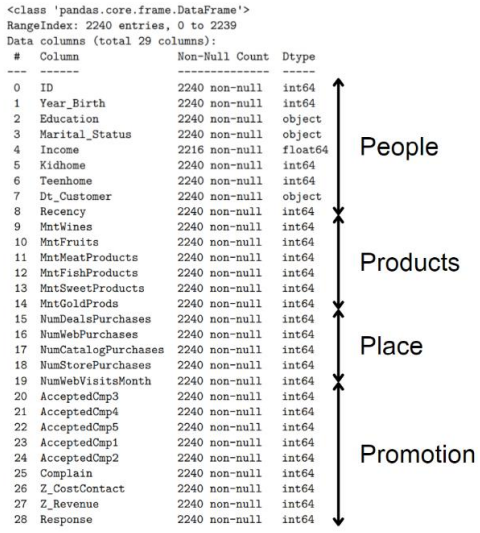
https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis

จุดประสงค์

1. ศึกษาการนำความรู้ทางด้านสถิติมาประยุกต์ใช้กับงานทางด้านการตลาด เช่น การวิเคราะห์กลุ่มลูกค้า
2. กำหนดกลยุทธ์ทางด้านการตลาดที่เหมาะสมจากข้อมูลที่ผ่านการวิเคราะห์แล้ว
3. เปรียบเทียบผลลัพธ์ที่ได้จากการแบ่งกลุ่มด้วย Euclidean distance และ Correlation-based distance

ขั้นตอนการทำงาน

1. Data Preprocessing



รูปที่ 1 แสดงรายละเอียดของข้อมูลที่ใช้

ข้อมูลชุดนี้ประกอบข้อมูลทั้งหมด 2240 ตัว โดยประกอบด้วยทั้งหมด 4 หมวด รวม 29 ฟีเจอร์ คือ

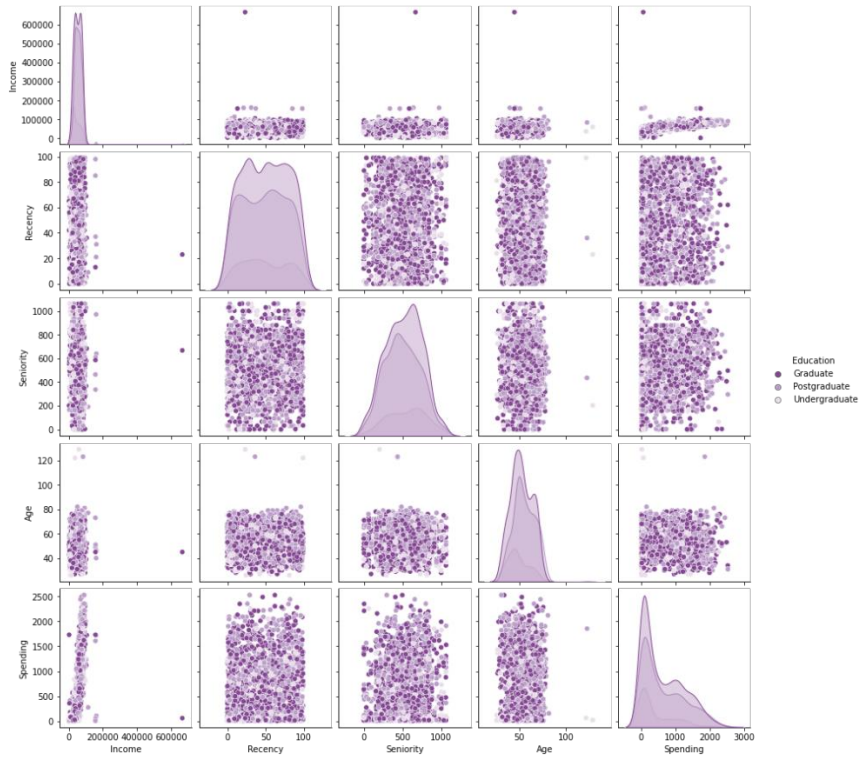
- People (ข้อมูลส่วนตัวของลูกค้า)
- Products (ข้อมูลการซื้อสินค้าของลูกค้าในแต่ละหมวด)
- Place (ข้อมูลการซื้อสินค้าของลูกค้าในแต่ละช่องทาง)
- Promotion (ข้อมูลการตอบรับโปรโมชันของลูกค้า)

แต่จากรูปที่ 1 จะเห็นได้ว่าข้อมูล Income มีอยู่เพียง 2216 ตัวเท่านั้น จึงต้องลบข้อมูลในแถวที่เป็น missing value ออกเพื่อให้ทุกฟีเจอร์มีจำนวนเท่ากัน

จากนั้นจึงทำการสร้างฟีเจอร์ใหม่ขึ้นมาเพื่อเป็นการลดจำนวนฟีเจอร์ที่ใช้งานและเป็นการเปลี่ยนฟีเจอร์ที่เป็นข้อมูลเชิงคุณภาพบางฟีเจอร์ให้เป็นข้อมูลเชิงคุณภาพให้สามารถใช้งานในโมเดลได้ ดังนี้

- Age: อายุของลูกค้าซึ่งหาจากปีปัจจุบัน (2022) ลบด้วยปีเกิดของลูกค้า
- Spending: ยอดรวมการซื้อสินค้าของลูกค้าซึ่งหาจากผลรวมของฟีเจอร์ในหมวด Products
- Seniority: อายุการใช้งานของลูกค้าซึ่งหาจากจำนวนวันนับจากวันแรกที่ลูกค้าสมัครเป็นสมาชิก
- Family Members: จำนวนสมาชิกของครอบครัวของลูกค้าซึ่งหาจากสถานภาพสมรสบวกกับจำนวนเด็กเล็กและจำนวนเด็กวัยรุ่น

เมื่อได้ฟีเจอร์ตามที่ต้องการแล้วจึงทำการพล็อตความสัมพันธ์ของฟีเจอร์บางคู่ที่เป็นตัวเลขและมีความต่อเนื่อง ดู จะได้ดังรูปที่ 2

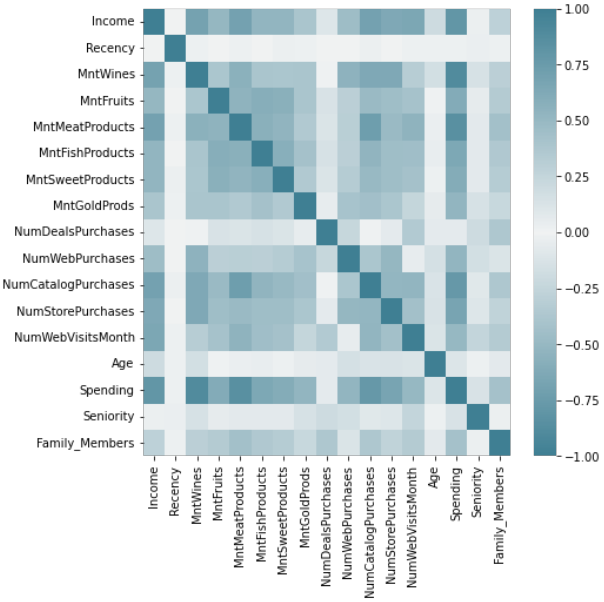


รูปที่ 2 แสดงความสัมพันธ์ของฟีเจอร์ที่เป็นตัวเลขแบบต่อเนื่อง

จะเห็นว่ามี Outliers อยู่ใน Income และ Age จึงต้องการทำลบออก โดยจะใช้ข้อมูล Income ที่ไม่เกิน 200000 หน่วยต่อปี (ไม่ทราบหน่วยเงินที่แน่ชัด) และข้อมูล Age ที่ไม่เกิน 90 ปีเท่านั้น

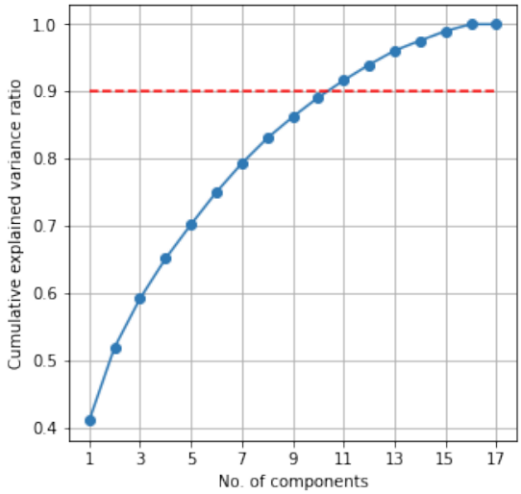
2. Clustering Analysis

ก่อนที่จะเริ่มทำการแบ่งกลุ่ม เมื่อลองตรวจสอบดูจะพบว่ามีหลาย ๆ ฟีเจอร์ที่มี Correlation กับฟีเจอร์อื่น ๆ สูงมาก ดังรูปที่ 3 จึงทำการใช้ Principal Components Analysis (PCA) ในการลดมิติของข้อมูล



รูปที่ 3 แสดง Correlation ของฟีเจอร์ที่ใช้ในโมเดล

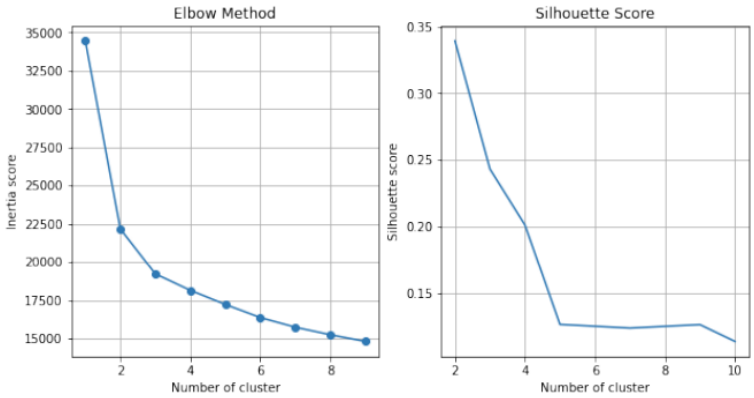
โดยจะเลือกใช้จำนวน components ที่เหมาะสมทั้งหมด 11 components เนื่องจากมีค่า explained variance ratio รวมกันอยู่ที่ 0.9 ซึ่งถือว่าเป็นค่าที่ยอมรับได้ ดังรูปที่ 4



รูปที่ 4 แสดงความสัมพันธ์ของ explained variance ratio และจำนวน components

K-mean (Euclidean Distance)

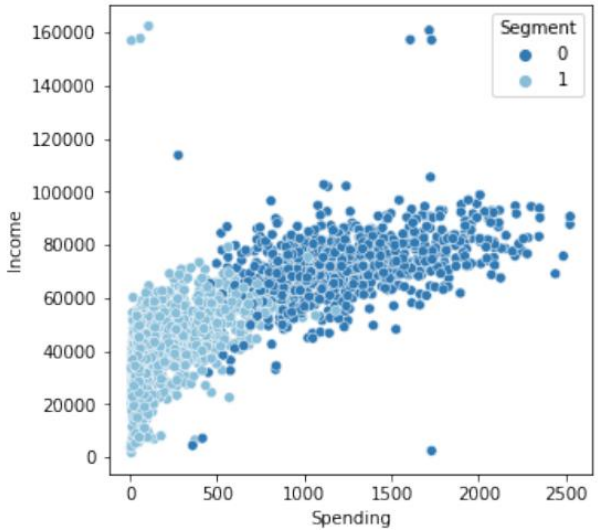
จากนั้นจึงทำการแบ่งกลุ่มด้วย K-mean โดยใช้ Elbow method และ Silhouette Score เป็นเกณฑ์ในการเลือกจำนวนกลุ่มที่เหมาะสมซึ่งทั้งสองวิธีได้ผลลัพธ์เท่ากัน คือ ค่า K = 2 ดังรูปที่ 5



รูปที่ 5 แสดงความสัมพันธ์ของ Elbow method (ซ้าย)/Silhouette Score (ขวา) และจำนวนกลุ่ม

เมื่อแบ่งกลุ่มลูกค้าออกเป็น 2 กลุ่มแล้วพิจารณาลักษณะเฉพาะของแต่ละกลุ่มจะพบว่าลูกค้าทั้งสองกลุ่มมีรายได้ (Income) และรายจ่าย (Spending) ที่แตกต่างกัน ดังรูปที่ 6 คือ

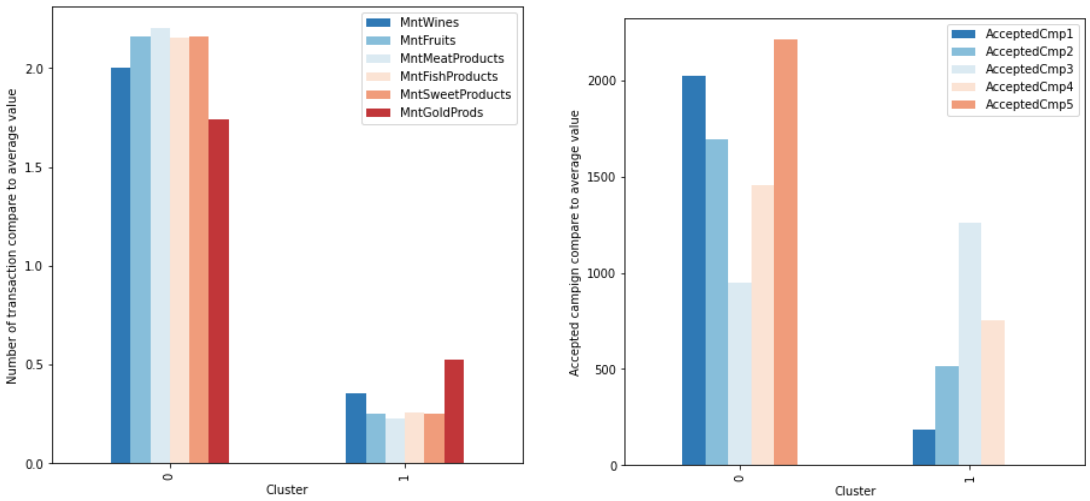
- กลุ่มที่ 0 - จะมี Income ที่มากและไปในทิศทางเดียวกับ Spending ที่มากเช่นกัน
- กลุ่มที่ 1 - จะมี Income ที่น้อยและไปทิศทางเดียวกับ Spending ที่น้อยเช่นกัน



รูปที่ 6 แสดงความสัมพันธ์ของ Income และ Spending ของลูกค้าแต่ละกลุ่ม

เมื่อเรามาศึกษาข้อมูลพฤติกรรมของลูกค้าทั้งในกลุ่มที่ 0 และกลุ่มที่ 1 ดังรูปที่ 7 จะเห็นว่าปริมาณการซื้อสินค้าของทั้ง 2 กลุ่มแบ่งอย่างชัดเจนซึ่งเห็นได้ว่า กลุ่มที่ 0 ที่มีรายได้มากกว่านั้นจะมีจำนวนการซื้อสินค้าที่มากกว่ากลุ่มที่ 1 ในทุกชนิดสินค้าแล้วยังมีการซื้อ Wines เยอะมากเป็นพิเศษซึ่งสอดคล้องกับความเป็นจริงที่ว่าคนรวยมักจะกินไวน์นั่นเอง

ส่วนการตอบรับโปรโมชันที่ทางบริษัทเสนอให้ซึ่งมีตั้งแต่แคมเปญที่ 1-5 จะเห็นว่าทั้งสองกลุ่มมีการรับข้อเสนอของแคมเปญที่แตกต่างกันอย่างชัดเจนโดยกลุ่มที่ 0 จะมีแนวโน้มในการรับข้อเสนอของแคมเปญมากกว่ากลุ่มที่ 1 (ยกเว้นแคมเปญที่ 3) เนื่องจากกลุ่มที่ 0 เป็นกลุ่มที่มีรายได้ที่สูงกว่าย่อมมีความสามารถในการใช้จ่ายมากกว่า ทำให้สามารถเข้าร่วมโปรโมชันต่าง ๆ ที่ทางบริษัทเสนอให้ได้มากกว่า ดังรูปที่ 7 ซึ่งอาจตีความได้ว่าแคมเปญที่ 3 อาจเป็นแคมเปญที่เกี่ยวข้องกับของราคาถูกที่ทำให้กลุ่มที่ 1 สามารถเข้าร่วมโปรโมชันได้ เช่น ขนมคบเคี้ยวราคาพิเศษ เป็นต้น



รูปที่ 7 แสดงจำนวนการซื้อสินค้า (ซ้าย) และจำนวนโปรโมชั่นที่ตอบรับ (ขวา) ของลูกค้าแต่ละกลุ่ม

Hierarchical clustering (Correlation-Based Distance)

เนื่องจากข้อมูลที่ใช้มีข้อมูลส่วนที่เป็นจำนวนการซื้อสินค้าของลูกค้าในแต่ละหมวดซึ่งไม่เหมาะสมที่จะใช้ผลต่างของลูกค้าแต่ละคนมาเป็นตัววัด เพราะบางครั้งลูกค้าที่มีรสนิยมเหมือนกันหรือซื้อสินค้าหมวดเดียวกันอาจจะซื้อมากน้อยไม่เท่ากันก็ได้จึงควรใช้ Correlation-Based Distance เป็นตัววัดระยะห่างแทน Euclidean Distance เพื่อลองสังเกตดูผลลัพธ์ที่อาจจะแตกต่างไปจากเดิมนั่นเอง โดยจะใช้ฟีเจอร์ในหมวด Product ทั้งหมด 6 ฟีเจอร์เท่านั้น ดังรูปที่ 8

	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
0	635	88	546	172	88	88
1	11	1	6	2	1	6
2	426	49	127	111	21	42
3	11	4	20	10	3	5
4	173	43	118	46	27	15

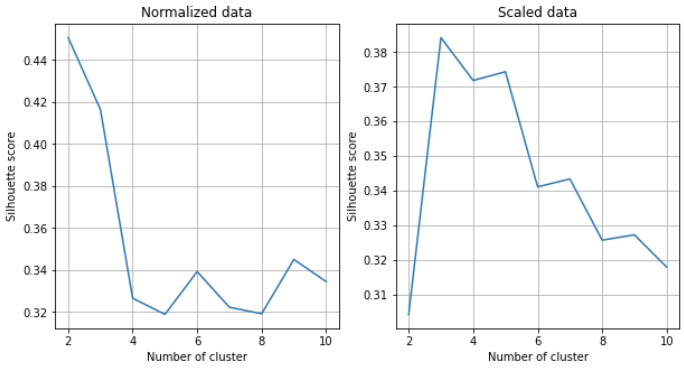
รูปที่ 8 แสดงข้อมูลเฉพาะฟีเจอร์ในหมวด Product

เมื่อนำข้อมูลฟีเจอร์ดังกล่าวไปใช้ในโมเดลแล้วพบว่าผลลัพธ์ที่ได้ออกมาไม่ดี เนื่องจากสินค้าในแต่ละหมวดมีการซื้อมากน้อยที่แตกต่างกัน เช่น ไวน์มีราคาแพงกว่าผลไม้มาก จึงต้องทำการ Normalize และ Scale ข้อมูลแต่ละฟีเจอร์เสียก่อน ดังนี้

$$x_{norm} = \frac{x_i}{S_x}$$

$$x_{scaled} = \frac{x_i - \bar{x}}{S_x}$$

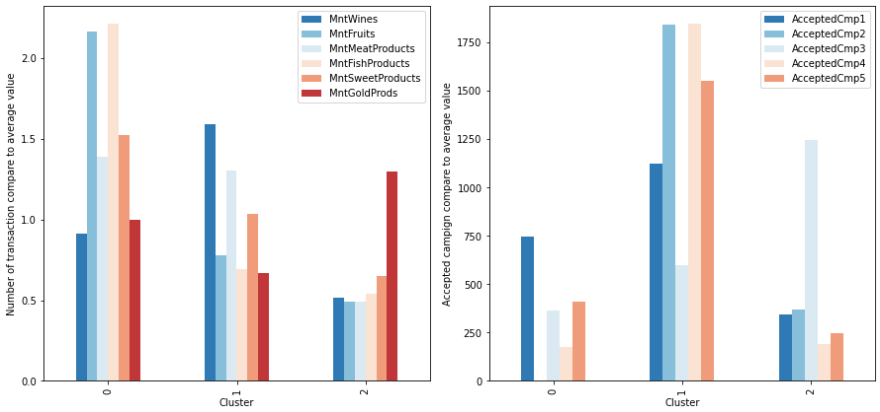
จากนั้นจึงใช้ Silhouette Score ในการหาจำนวนกลุ่มที่เหมาะสมซึ่งจะได้ผลลัพธ์ที่แตกต่างกันระหว่างข้อมูล 2 แบบ ดังรูปที่ 9 โดยข้อมูลที่ถูก Normalize จะได้ค่า K = 2 และข้อมูลที่ถูก Scale จะได้ค่า K = 3 เราจึงเลือกใช้ ข้อมูลที่ถูก Scale เพราะจะได้จำนวนกลุ่มที่แตกต่างจาก K-mean



รูปที่ 9 แสดง Silhouette Score ของข้อมูลที่ถูก Normalize (ซ้าย) และข้อมูลที่ถูก Scale (ขวา)

เมื่อเรามาศึกษาข้อมูลพฤติกรรมของลูกค้าทั้งในกลุ่มที่ 0, กลุ่มที่ 1 และกลุ่มที่ 2 ดังรูปที่ 10 จะเห็นว่าปริมาณการซื้อสินค้าของทั้ง 3 กลุ่มมีความแตกต่างกัน โดยกลุ่มที่ 0 ซื้อสินค้าในหมวดผลไม้, ปลา และของหวานมากที่สุด ส่วนกลุ่มที่ 1 ซื้อสินค้าในหมวดไวน์มากที่สุด ซึ่งสมเหตุสมผลเพราะกลุ่มนี้มีรายได้เฉลี่ยมากที่สุด (60365 หน่วยต่อปี) และกลุ่มที่ 2 ซื้อสินค้าในหมวดทองคำมากที่สุดซึ่งถือว่าเป็นสิ่งที่แปลกมาก เพราะกลุ่มนี้เป็นกลุ่มที่มีรายได้เฉลี่ยน้อยที่สุด (41947 หน่วยต่อปี) ทั้งหมดนี้อาจเป็นเพราะโมเดลที่ใช้ Correlation-Based Distance นี้พิจารณาเฉพาะฟีเจอร์ในหมวด Product และข้อมูลที่ใช้มีเพียง 2240 ตัวเท่านั้น จึงอาจทำให้ความบังเอิญส่งผลต่อผลลัพธ์ได้ เช่น คนรวยซื้อทองเยอะ คนจนซื้อทองน้อย แต่ถ้ารายการสินค้าที่ซื้อมีความคล้ายกันก็จะถูกจับกลุ่มรวมกันได้

ส่วนการตอบรับโปรโมชันที่ทางบริษัทเสนอให้ซึ่งมีตั้งแต่แคมเปญที่ 1-5 จะเห็นว่าทั้งสองกลุ่มมีการรับข้อเสนอของแคมเปญที่แตกต่างกันอย่างชัดเจนโดยกลุ่มที่ 0 จะมีแนวโน้มในการรับข้อเสนอของแคมเปญมากกว่ากลุ่มที่ 1 (ยกเว้นแคมเปญที่ 3) ซึ่งสามารถอธิบายได้ว่ากลุ่มที่มีรายได้ที่สูงกว่าย่อมมีความสามารถในการใช้จ่ายมากกว่า ทำให้สามารถเข้าร่วมโปรโมชันต่าง ๆ ที่ทางบริษัทเสนอให้ได้มากกว่า ดังรูปที่ 7



รูปที่ 10 แสดงจำนวนการซื้อสินค้าแต่ละหมวด (ซ้าย) และจำนวนโปรโมชันที่ตอบรับ (ขวา) เทียบกับค่าเฉลี่ยของลูกค้าแต่ละกลุ่ม

3. Association Rules

ก่อนที่จะเริ่มใช้ Association rules ในขั้นตอนแรกจำเป็นต้องเปลี่ยนแปลง ข้อมูลเชิงปริมาณทั้งหมดให้เป็นข้อมูลเชิงคุณภาพก่อนเนื่องจาก Association rules ไม่สามารถใช้กับข้อมูลเชิงปริมาณได้ โดยข้อมูลที่ใช้กับ Association Rules แสดงดังรูปที่ 11

Education	cluster_scaled	Seniority_group	Wines_segment	Fruits_segment	Meat_segment	Fish_segment	Sweets_segment	Gold_segment
Graduate	0	Old customers	Biggest consumer	Biggest consumer	Biggest consumer	Biggest consumer	Biggest consumer	Biggest consumer
Graduate	2	New customers	Low consumer	Low consumer	Low consumer	Low consumer	Low consumer	Low consumer
Graduate	0	Discovering customers	Frequent consumer	Biggest consumer	Frequent consumer	Biggest consumer	Frequent consumer	Frequent consumer
Graduate	0	New customers	Low consumer	Low consumer	Frequent consumer	Frequent consumer	Low consumer	Low consumer
Postgraduate	0	New customers	Frequent consumer	Frequent consumer	Frequent consumer	Frequent consumer	Frequent consumer	Frequent consumer
...	...	...	...	...	...	...	...	...
Graduate	2	Experienced customers	Biggest consumer	Frequent consumer	Frequent consumer	Frequent consumer	Biggest consumer	Biggest consumer
Postgraduate	1	New customers	Frequent consumer	Non consumer	Frequent consumer	Non consumer	Non consumer	Low consumer
Graduate	1	New customers	Biggest consumer	Biggest consumer	Frequent consumer	Frequent consumer	Frequent consumer	Frequent consumer
Postgraduate	0	New customers	Frequent consumer	Frequent consumer	Frequent consumer	Biggest consumer	Frequent consumer	Biggest consumer
Postgraduate	2	Old customers	Frequent consumer	Low consumer	Frequent consumer	Low consumer	Low consumer	Frequent consumer

รูปที่ 11 แสดงตารางข้อมูลที่ใช้กับ Association rules

เมื่อใช้ Association rules โดยสังเกตุจากค่า Lift เป็นหลักจะได้ผลดังรูปที่ 12

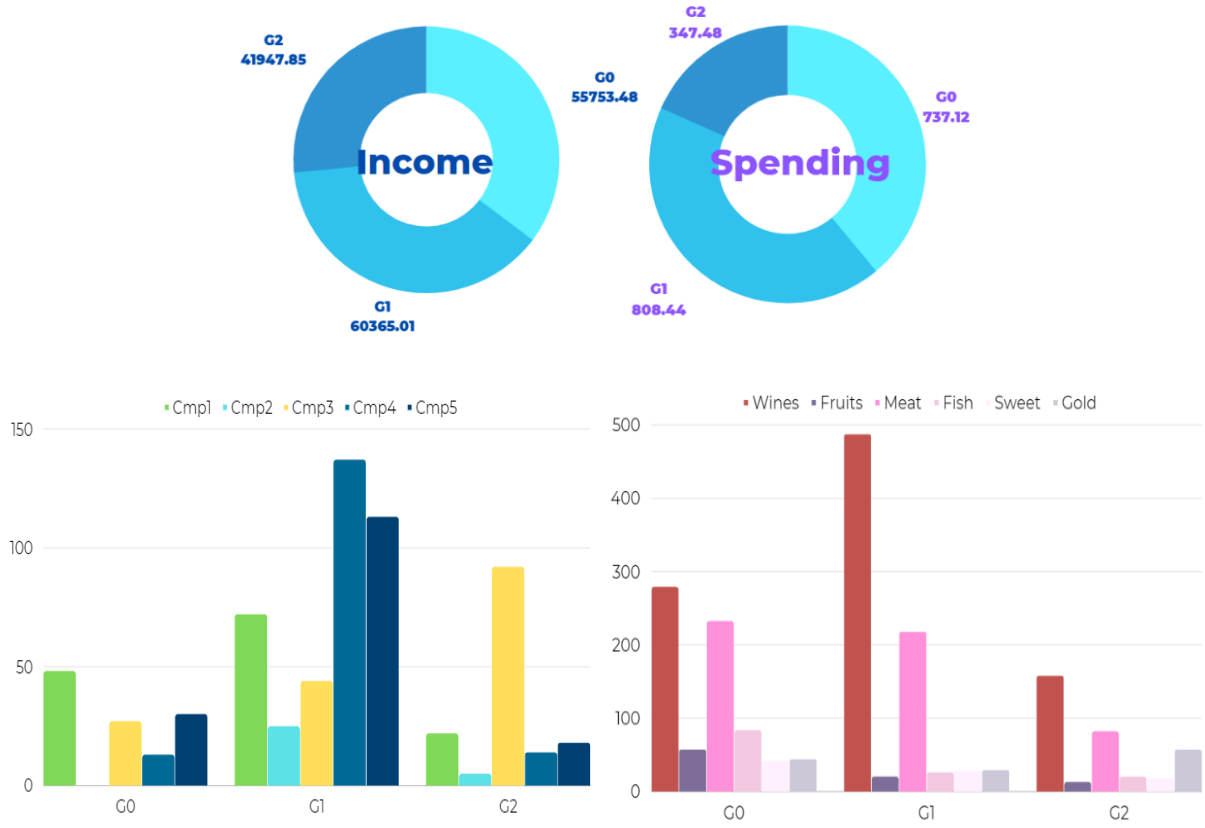
antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
(cluster_scaled_0, Meat_segment_Biggest consumer)	(Fish_segment_Biggest consumer)	0.094075	0.207146	0.083673	0.889423	4.293700	0.064185	7.170157
(cluster_scaled_0, Fruits_segment_Biggest cons...	(Fish_segment_Biggest consumer)	0.109905	0.207146	0.085029	0.773663	3.734864	0.062263	3.502973
(cluster_scaled_0, Fish_segment_Biggest consumer)	(Fruits_segment_Biggest consumer)	0.121664	0.203076	0.085029	0.698885	3.441502	0.060322	2.646576
(cluster_scaled_2, Meat_segment_Low consumer)	(Wines_segment_Low consumer)	0.151515	0.253279	0.123021	0.811940	3.205714	0.084646	3.970659
(cluster_scaled_1, Meat_segment_Biggest consumer)	(Wines_segment_Biggest consumer)	0.114428	0.248756	0.087743	0.766798	3.082530	0.059278	3.221435
...	...	...	...	...	...	...	...	...
(cluster_scaled_1)	(Sweets_segment_Frequent consumer, Fish_segmen...	0.360018	0.237901	0.086839	0.241206	1.013891	0.001190	1.004355
(Education_Graduate, cluster_scaled_2)	(Fruits_segment_Frequent consumer)	0.219810	0.395749	0.088195	0.401235	1.013862	0.001206	1.009162
(cluster_scaled_2, Wines_segment_Frequent cons...	(Gold_segment_Frequent consumer)	0.192221	0.471280	0.091361	0.475294	1.008518	0.000772	1.007650
(cluster_scaled_1)	(Meat_segment_Frequent consumer, Fish_segment_...	0.360018	0.254184	0.091814	0.255025	1.003311	0.000303	1.001130
(cluster_scaled_2)	(Fish_segment_Frequent consumer, Fruits_segmen...	0.398462	0.238806	0.095432	0.239501	1.002909	0.000277	1.000913

รูปที่ 12 ตารางแสดงผลของ Association rules เมื่อเรียงจากค่า lift

เมื่อวิเคราะห์ผลที่ได้จากการทำ Association rules เปรียบเทียบกับผลของการทำ Clustering Analysis พบว่าผลที่ได้มีความคล้ายคลึงกัน แต่ตัวของ Association rules นั้นมีความยากในการวิเคราะห์มากกว่าเนื่องจากกลุ่มที่สามารถแบ่งได้นั้นมีความชัดเจนและมีน้อยกลุ่มมากจนสามารถวิเคราะห์ได้เอง ทำให้มีความยุ่งยากในการดูความสัมพันธ์ของกลุ่มลูกค้าและสินค้า จึงสามารถสรุปได้ว่าการใช้ Association rules ไม่เหมาะสมกับข้อมูลชุดนี้

4. Conclusion

เนื่องจากการแบ่งกลุ่มด้วยวิธี Correlation-Based Distance นอกจากจะสามารถแบ่งกลุ่มรายรับและการใช้จ่ายของลูกค้าได้แล้วยังสามารถแบ่งกลุ่มการซื้อสินค้าต่าง ๆ ได้ชัดเจนกว่าการใช้ Euclidian Distance ดังนั้นเราจึงเลือกใช้ผลที่ได้จากการจัดกลุ่มด้วยวิธี Correlation-Based Distance



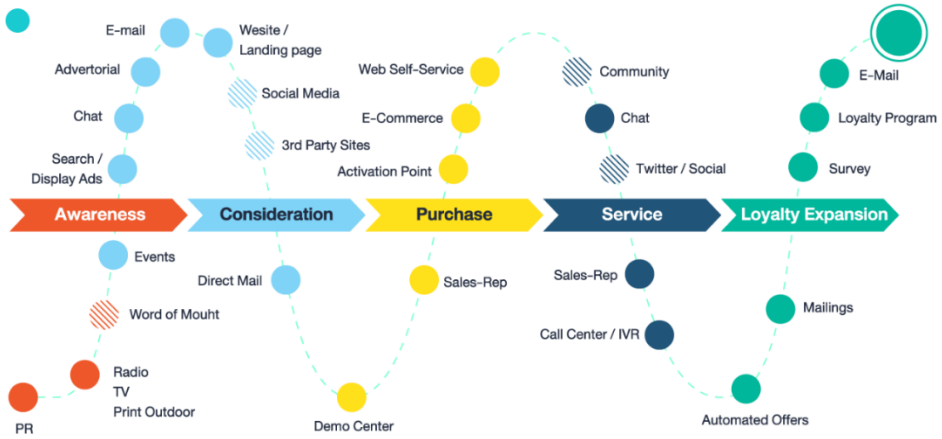
รูปที่ 13 กราฟแสดงผลของการแบ่งกลุ่มโดยใช้ Correlation-Based Distance

จากการวิเคราะห์ผลได้เราสามารถเสนอกลยุทธ์ได้ดังนี้

1. เติมไวน์และผลิตภัณฑ์จากเนื้อสัตว์ประเภทต่าง ๆ และเตรียมเพื่อไม่ให้สินค้าหมด เนื่องจากลูกค้าทุกกลุ่มบริโภคไวน์และเนื้อเป็นจำนวนมากหากสินค้าไม่หมดเราก็สามารถมีกำไรจากส่วนนี้ได้มากขึ้น
2. หากต้องการใช้เงินน้อยลงในการทำการตลาดไวน์ใหม่ที่ยังไม่เป็นที่รู้จักในขณะที่ยังทำเงินได้มากควรทำการตลาดที่กลุ่ม G1 เป็นหลักเนื่องจากเป็นกลุ่มที่บริโภคไวน์มากที่สุดและยังเป็นกลุ่มที่มีการใช้จ่ายโดยรวมมากที่สุดด้วย

3. เสนอทีมการตลาดเพื่อประเมินและปรับปรุงแคมเปญที่ 2 แคมเปญ 4 และแคมเปญ 5 เนื่องจากจะเห็นได้ว่าแคมเปญที่ 2 ลูกค้าทุกกลุ่มรับน้อยมากจึงอาจสรุปได้ว่าแคมเปญนี้ไม่ตอบโจทย์กลุ่มลูกค้าของเราและในส่วนของแคมเปญที่ 4 และ 5 จะเห็นได้ว่ากลุ่ม G0 นั้นรับแคมเปญค่อนข้างน้อยมากทั้ง ๆ ที่เป็นกลุ่มที่มีรายรับและรายจ่ายใกล้เคียงกับกลุ่ม G1 จึงอยากให้ปรับปรุงแคมเปญที่ 4 และ 5 เนื่องจากลูกค้ากลุ่ม G0 นั้นมีการใช้จ่ายที่สูงไม่แพ้กลุ่ม G1 หากเราสามารถทำให้ 2 แคมเปญนี้ตอบโจทย์ได้ทั้งกลุ่ม G0 และ G1 เราจะได้กำไรมากขึ้นจากการที่กลุ่ม G0 อยากใช้จ่ายมากขึ้น

5. Future work suggestion



รูปด้านบนแสดงถึง Customer journey ของลูกค้าแต่ละคนเนื่องจากข้อมูลที่เราใช้ในการศึกษาถูกบันทึกในปี 2012-2014 ซึ่งในช่วงนั้นยังไม่มีการเก็บข้อมูลที่มากและการค้าขายผ่านทางออนไลน์ยังไม่เป็นที่นิยมจึงทำให้การเก็บข้อมูลทำได้ยาก ซึ่งหากเรานำ Model Statistical learning มาประยุกต์ใช้กับ Customer journey ข้างต้นกับจำนวนลูกค้าที่มากจะทำให้เราสามารถบอกได้ว่าลูกค้ากลุ่มไหนชอบซื้อสินค้าแบบไหน แล้วลูกค้าติดปัญหาในขั้นตอนไหนก็จะสามารถเข้าไปแก้ไขได้ถูกจุดหรือสามารถเลือกลงทุนเงินในแต่ละขั้นตอนได้ถูกต้องมากยิ่งขึ้น เนื่องจากเราสามารถวิเคราะห์ได้ว่าจุดไหนทำให้ลูกค้ารู้จักหรือติดใจในบริการและสินค้าของเรา