รายงาน

Fundamentals of Data Science

กิตติพงศ์ พวงสินธ์ 66070016

ชัชชัย แสวงนิล 66070044

ยศกร ชูวงษ์ 66070168

ศุภณัฐ จันทรสาขา 66070196

สิรภพ สรรค์ศิลา 66070204

วรวุฒิ มหาทอง 66070307

เสนอ

อ.เฉลิมพล ศิริกายน

รายงานเล่มนี้เป็นส่วนหนึ่งของรายวิชา

Fundamentals of Data Science

รหัสวิชา 06026208

คณะเทคโนโลยีสารสนเทศ

สาขาวิทยาการข้อมูลและการวิเคราะห์เชิงธุรกิจ

สถาบันเทคโนโลยีพระจอมเกล้าเจ้าคุณทหารลาดกระบัง

ภาคเรียนที่ 2 ปีการศึกษา 2567

สารบัญ

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ.....

คำนำ

รายงานฉบับนี้มีวัตถุประสงค์เพื่อวิเคราะห์ข้อมูลของผู้กระทำความรุนแรงในครอบครัว โดยมุ่งเน้นไปที่การทำความเข้าใจปัจจัยที่มีผลต่อสุขภาพจิตของบุคคลเหล่านี้ และการแบ่งกลุ่มประชากรออกเป็น Cluster ตามลักษณะพฤติกรรมและความเสี่ยงที่แตกต่างกัน เพื่อนำไปสู่การพัฒนาแนวทางแก้ไขปัญหาที่มีประสิทธิภาพมากยิ่งขึ้น

ความรุนแรงในครอบครัวเป็นปัญหาที่ซับซ้อนและส่งผลกระทบอย่างรุนแรงต่อผู้ถูกกระทำ รวมถึงผู้กระทำเองด้วย การทำความเข้าใจปัจจัยที่เกี่ยวข้องกับสุขภาพจิตของผู้กระทำความรุนแรงจึงเป็นสิ่งสำคัญในการวางแผนการป้องกันและแก้ไขปัญหาอย่างยั่งยืน โดยรายงานนี้จะตอบคำถามสำคัญดังต่อไปนี้:

- ปัจจัยใดบ้างที่มีผลกระทบต่อสุขภาพจิตของผู้กระทำความรุนแรงมากที่สุด?
- สามารถแบ่งประชากรออกเป็นกลุ่มที่มีความเสี่ยงแตกต่างกันได้อย่างไร?
- Cluster ใดมีแนวโน้มสุขภาพจิตที่แย่ที่สุด และแต่ละ Cluster มีลักษณะพฤติกรรมที่แตกต่างกันอย่างไร?

ข้อมูลที่ใช้ในการวิเคราะห์ประกอบด้วยข้อมูล 564 แถว และ 19 คอลัมน์ ซึ่งครอบคลุมทั้งข้อมูลเชิงหมวดหมู่ เช่น ภูมิภาค จังหวัด เพศ และสถานภาพ และข้อมูลเชิงตัวเลข เช่น อายุ ระดับการดื่มแอลกอฮอล์ และระดับความโกรธ การวิเคราะห์ข้อมูลจะใช้เทคนิคทางสถิติและการเรียนรู้ของเครื่อง (Machine Learning) เพื่อระบุรูปแบบและความสัมพันธ์ที่ซ่อนอยู่ในข้อมูล

ผลการวิเคราะห์ที่ได้จากรายงานนี้ จะเป็นประโยชน์อย่างยิ่งต่อผู้ที่เกี่ยวข้องในการพัฒนาแนวทางแก้ไขปัญหาความรุนแรงในครอบครัว เช่น นักจิตวิทยา นักสังคมสงเคราะห์ ผู้กำหนดนโยบาย และองค์กรที่ทำงานด้านการช่วยเหลือผู้ที่ได้รับผลกระทบจากความรุนแรงในครอบครัว โดยสามารถนำไปใช้ในการพัฒนาโปรแกรมให้คำปรึกษา นโยบายสนับสนุน และโปรแกรมช่วยเหลือกลุ่มเสี่ยงต่างๆ ได้อย่างตรงจุดและมีประสิทธิภาพมากยิ่งขึ้น

หวังเป็นอย่างยิ่งว่ารายงานฉบับนี้จะเป็นประโยชน์ต่อการทำความเข้าใจและแก้ไขปัญหาความรุนแรงในครอบครัวต่อไป

1. Business Understanding

1.1 วัตถุประสงค์

เพื่อศึกษาแนวโน้มการเกิดปัญหาความรุนแรงในครอบครัว โดยพิจารณาจากปัจจัยต่างๆ เช่น การใช้สารเสพติด การดื่มเครื่องดื่มมึนเมา ความเครียดทางการเงิน ปัญหาทางสุขภาพจิต และความสัมพันธ์นั้น เป็นเหตุผลที่สำคัญและเกี่ยวข้องกับปัญหาทางสังคมที่ซับซ้อน

1.2 สามารถนำไปใช้ทำอะไรได้บ้าง

- **ระบุกลุ่มเสี่ยงและช่วยเหลือเชิงรุก:** ข้อมูลจะช่วยให้สามารถระบุกลุ่มบุคคลที่มีความเสี่ยงต่อการกระทำความรุนแรงในครอบครัวได้แม่นยำยิ่งขึ้น ทำให้สามารถวางแผนและดำเนินการช่วยเหลือเชิงรุกได้อย่างมีประสิทธิภาพ เช่น การจัดโปรแกรมให้คำปรึกษาหรือการสนับสนุนเฉพาะกลุ่ม
- **พัฒนาแนวทางป้องกันและแก้ไขปัญหา:** ข้อมูลเชิงลึกเกี่ยวกับปัจจัยที่เกี่ยวข้องกับความรุนแรงในครอบครัวจะช่วยในการพัฒนาแนวทางป้องกันและแก้ไขปัญหาที่มีประสิทธิภาพมากยิ่งขึ้น เช่น การปรับปรุงนโยบาย การสร้างโปรแกรมการศึกษา หรือการฝึกอบรม
- **นำเสนอข้อมูลเชิงลึกเกี่ยวกับปัจจัยที่เกี่ยวข้องกับความรุนแรงในครอบครัว:** การวิเคราะห์ข้อมูลจะช่วยให้เข้าใจถึงปัจจัยที่ส่งผลต่อการเกิดความรุนแรงในครอบครัวได้อย่างลึกซึ้ง เช่น ปัจจัยทางเศรษฐกิจ สังคม จิตวิทยา หรือความสัมพันธ์ ทำให้สามารถวางแผนการแก้ไขปัญหาได้อย่างตรงจุด
- **สร้างความตระหนักรู้และกระตุ้นให้สังคมร่วมมือกันแก้ไขปัญหา:** การเผยแพร่ข้อมูลและผลการวิเคราะห์จะช่วยสร้างความตระหนักรู้เกี่ยวกับปัญหาความรุนแรงในครอบครัวในสังคม กระตุ้นให้เกิดการเปลี่ยนแปลงทัศนคติและพฤติกรรม และส่งเสริมให้สังคมร่วมมือกันแก้ไขปัญหาอย่างจริงจัง

1.3 คำถามที่ต้องตอบ

- ปัจจัยอะไรที่ส่งผลกระทบต่อสุขภาพจิตมากที่สุด
- เราสามารถแบ่งประชากรออกเป็นกลุ่มที่มีความเสี่ยงต่างกันได้หรือไม่?
- Cluster ไหนมีแนวโน้มสุขภาพจิตแย่ที่สุด และแต่ละ Cluster มีลักษณะพฤติกรรมที่แตกต่างกันอย่างไร?

2. data Understanding

2.1 โครงสร้างของข้อมูล

- ข้อมูลมีทั้งหมด 564 แถว และ 19 คอลัมน์
- คอลัมน์ที่มีข้อมูลเชิงหมวดหมู่ (เช่น ภูมิภาค จังหวัด เพศ สถานภาพ ฯลฯ)
- คอลัมน์ที่เป็นตัวเลข (เช่น อายุ แอลกอฮอล์ ความโกรธ ฯลฯ)

2.2 ข้อมูลเชิงหมวดหมู่

- Regional (ภูมิภาค) แบ่งออกเป็น 4 ภูมิภาค (ภาคกลางเป็นสัดส่วนมากที่สุด)
- Province (จังหวัด) มี 69 จังหวัด (กรุงเทพฯ เป็นจังหวัดที่พบมากที่สุด)
- Gender (เพศ) พบว่า ส่วนใหญ่เป็นเพศชาย (461 คน) รองลงมาคือเพศหญิงและไม่ระบุเพศ
- Age Range (ช่วงอายุ)
    - วัยกลางคน (36-59 ปี) เป็นกลุ่มที่ถูกกระทำมากที่สุด (287 คน)
    - รองลงมาคือวัยผู้ใหญ่ตอนต้น (19-35 ปี)

2.3 ข้อมูลเชิงตัวเลข

- Age (อายุ)
    - ค่าเฉลี่ย 40.2 ปี
    - ช่วงอายุที่พบมากสุดอยู่ในวัยกลางคน
- Alcohol, Drug, Rage และปัจจัยอื่นๆ มีค่าเฉลี่ยต่ำ แต่มีค่าผิดปกติสูงมากในบางแถว

2.4 ตารางข้อมูล

	Regional	Province	District	Sub-District	Gender	Age	Age Range	Relation	Mariage Registration	Alcohol	Drug	Authoritative	Rage	Jealous	Divorce	Health Problem	Mental Problem	Gambling Addict	Economics Stress
0	ภาคกลาง	กรุงเทพมหานคร	เขตหนองจอก	สำนักขี	หญิง	61	วัยสูงอายุ 60 ปีขึ้นไป	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่
1	ภาคกลาง	กรุงเทพมหานคร	เขตบางแค	บางแค	หญิง	26	วัยผู้ใหญ่ตอนต้น 19 - 35 ปี	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่
2	ภาคกลาง	กรุงเทพมหานคร	เขตหวัวัฒนา	ศาลาธรรมสพน์	หญิง	55	วัยกลางคน 36 - 59 ปี	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่
3	ภาคกลาง	กรุงเทพมหานคร	เขตบางขื่อ	วงศ์สว่าง	หญิง	64	วัยสูงอายุ 60 ปีขึ้นไป	แยกกันอยู่	จดทะเบียนหย่า	ไม่	ไม่	ใช่	ใช่	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่
4	ภาคกลาง	กรุงเทพมหานคร	เขตประเวศ	ดอกไม้	ชาย	33	วัยผู้ใหญ่ตอนต้น 19 - 35 ปี	สมรส	ยังไม่ได้จดทะเบียนสมรส	ใช่	ไม่	ไม่	ใช่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่

2.5 จำนวนแถวและคอลัมน์

```
df.shape
✓ 0.0s
(564, 19)
```

มีทั้งหมด 564 แถว 19 คอลัมน์

2.6 สำรวจว่ามีค่าว่างในข้อมูลหรือไม่

```
df.info()
✓ 0.0s
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 564 entries, 0 to 563
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Regional              564 non-null    object
 1   Province              564 non-null    object
 2   District              564 non-null    object
 3   Sub-District          564 non-null    object
 4   Gender                564 non-null    object
 5   Age                   564 non-null    int64
 6   Age Range             564 non-null    object
 7   Relation              564 non-null    object
 8   Mariage Registration  564 non-null    object
 9   Alcohol               564 non-null    int64
 10  Drug                  564 non-null    int64
 11  Authoritative         564 non-null    int64
 12  Rage                  564 non-null    int64
 13  Jealous               564 non-null    int64
 14  Divorce               564 non-null    int64
 15  Health Problem        564 non-null    int64
 16  Mental Problem        564 non-null    int64
 17  Gambling Addict       564 non-null    int64
 18  Economics Stress      564 non-null    int64
dtypes: int64(11), object(8)
memory usage: 83.8+ KB
```

```
df.isna().sum()
✓ 0.0s
Regional              0
Province              0
District              0
Sub-District          0
Gender                0
Age                   0
Age Range             0
Relation              0
Mariage Registration  0
Alcohol               0
Drug                  0
Authoritative         0
Rage                  0
Jealous               0
Divorce               0
Health Problem        0
Mental Problem        0
Gambling Addict       0
Economics Stress      0
dtype: int64
```

จาก df.isnull().sum() แล้วทุกคอลัมน์เป็น 0 แสดงว่า ข้อมูลไม่มีค่าว่าง

2.7 การกระจายตัวของอายุ

จากกราฟ สามารถอธิบายการกระจายตัวของอายุได้ดังนี้

- ช่วงอายุที่มีความถี่สูงสุด: ช่วงอายุที่มีความถี่สูงสุด (mode) อยู่ในช่วงประมาณ 30-40 ปี ซึ่งแสดงว่ากลุ่มตัวอย่างส่วนใหญ่มีอายุอยู่ในช่วงนี้
- ช่วงอายุที่มีความถี่ต่ำ: ช่วงอายุที่มีความถี่ต่ำจะอยู่บริเวณส่วนปลายของกราฟ คือช่วงอายุ 10-20 ปี และช่วงอายุ 60-80 ปี
- การกระจายตัวโดยรวม: กราฟแสดงให้เห็นว่าการกระจายตัวของอายุมีความสมมาตร (symmetric) โดยประมาณ ซึ่งบ่งชี้ว่าข้อมูลมีการกระจายตัวอย่างสม่ำเสมอในแต่ละช่วงอายุ

2.8 ข้อมูลสถิติพื้นฐาน เฉพาะชุดข้อมูลที่ไม่ใช่ตัวเลข

```
df.describe(include=['O'])
```
✓ 0.0s

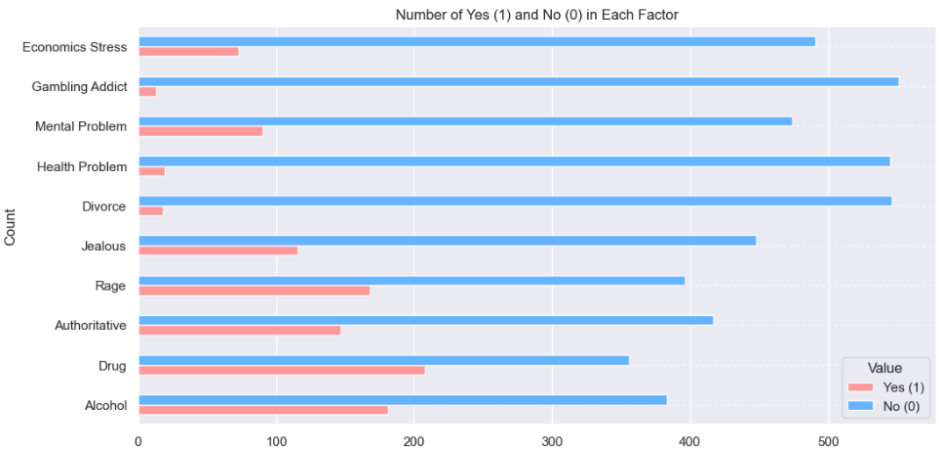
	Regional	Province	District	Sub-District	Gender	Age Range	Relation	Mariage Registration
count	564	564	564	564	564	564	564	564
unique	4	69	303	453	3	5	4	5
top	ภาคกลาง	กรุงเทพมหานคร	เมือง ราชบุรี	ในเมือง	ชาย	วัยกลางคน 36 - 59 ปี	สมรส	ไม่ระบุ
freq	257	59	14	10	461	287	259	216

จากภาพแสดงข้อมูลให้เห็นว่า

- ภาคที่มากที่สุดคือ ภาคกลาง
- จังหวัดที่มากที่สุดคือ กรุงเทพมหานคร
- เพศที่มากที่สุด คือ เพศชาย
- ช่วงอายุที่มากที่สุดคือ ช่วงวัยกลางคน 36-59 ปี
- สถานะที่มากที่สุดคือ สมรส

2.9 ข้อมูลสถิติของปัจจัยที่ส่งผลต่อกระทำความรุนแรงในครอบครัว

	ใช่ (1)	ไม่ (0)
Alcohol	181	383
Drug	208	356
Authoritative	147	417
Rage	168	396
Jealous	116	448
Divorce	18	546
Health Problem	19	545
Mental Problem	90	474
Gambling Addict	13	551
Economics Stress	73	491



3. Data Preparation

3.1 นำปัจจัยที่ลักษณะคล้ายกันมารวมกัน

```
# สร้างฟีเจอร์ใหม่โดยรวมคอลัมน์ที่เกี่ยวข้อง
pref_data["Substance Use"] = pref_data["Alcohol"] + pref_data["Drug"]
pref_data["Aggression"] = pref_data["Authoritative"] + pref_data["Rage"]
pref_data["Relationship Issues"] = pref_data["Jealous"] + pref_data["Divorce"]
pref_data["Health Issues"] = pref_data["Health Problem"] + pref_data["Mental Problem"]
pref_data["Financial Issues"] = pref_data["Gambling Addict"] + pref_data["Economics Stress"]

# ลบคอลัมน์เดิมที่ถูกใช้ไปรวมแล้ว
pref_data.drop(columns=["Alcohol", "Drug", "Authoritative", "Rage", "Jealous", "Divorce",
                        "Health Problem", "Mental Problem", "Gambling Addict", "Economics Stress"], inplace=True)
pref_data
```

	Regional	Province	District	Sub-District	Gender	Age	Age Range	Relation	Mariage Registration	Substance Use	Aggression	Relationship Issues	Health Issues	Financial Issues
0	1	2	207	247	2	61.0	วัยสูงอายุ 60 ปีขึ้นไป	โสด	ไม่ระบุ	0	1	0	0	0
1	1	2	193	155	2	26.0	วัยผู้ใหญ่ตอนต้น 19 - 35 ปี	โสด	ไม่ระบุ	0	1	0	0	0
2	1	2	184	275	2	55.0	วัยกลางคน 36 - 59 ปี	โสด	ไม่ระบุ	0	0	0	1	0
3	1	2	189	249	2	64.0	วัยสูงอายุ 60 ปีขึ้นไป	แยกกันอยู่	จดทะเบียนหย่า	0	2	0	1	0
4	1	2	194	46	1	33.0	วัยผู้ใหญ่ตอนต้น 19 - 35 ปี	สมรส	ยังไม่ได้จดทะเบียนสมรส	1	1	1	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
560	4	48	255	225	1	47.0	วัยกลางคน 36 - 59 ปี	สมรส	ยังไม่ได้จดทะเบียนสมรส	1	0	0	0	0
561	4	56	135	69	1	39.0	วัยกลางคน 36 - 59 ปี	แยกกันอยู่	ยังไม่ได้จดทะเบียนหย่า	1	0	1	0	0
562	4	56	263	230	1	40.0	วัยกลางคน 36 - 59 ปี	แยกกันอยู่	จดทะเบียนหย่า	2	0	1	0	0
563	4	56	3	95	1	64.0	วัยสูงอายุ 60 ปีขึ้นไป	สมรส	ยังไม่ได้จดทะเบียนสมรส	2	1	0	0	0
564	0	0	0	0	0	NaN	NaN	NaN	NaN	389	315	134	109	86

565 rows × 14 columns

ปัจจัยที่มีความหมายใกล้เคียงกันหรือเกี่ยวข้องซึ่งกันและกัน ได้ถูกจัดกลุ่มเข้าด้วยกันเป็นประเภทเดียวกัน วิธีนี้ช่วยลดจำนวนตัวแปรที่ต้องวิเคราะห์ และทำให้เห็นความสัมพันธ์ระหว่างปัจจัยต่างๆ ได้ชัดเจนยิ่งขึ้น

จากภาพแสดงข้อมูลให้เห็นว่า

- column Drug กับ column Alcohol รวมกันเป็น column Substance Use
- column authoritative กับ column Rage รวมกันเป็น column Agression
- column Jealous กับ column Divorce รวมกันเป็น column Relationship Issue
- column Health กับ column Mental รวมกันเป็น column Health Issue
- column Gambling กับ column Economic stress รวมกันเป็น column Financial Issue

3.2 แบ่งช่วงอายุของข้อมูล

ข้อมูลอายุเดิมที่อาจเป็นค่าต่อเนื่อง ได้ถูกแบ่งออกเป็นช่วงอายุที่เหมาะสม เพื่อให้เห็นภาพรวมของกลุ่มอายุต่างๆ ได้ชัดเจนยิ่งขึ้น การแบ่งช่วงอายุนี้ช่วยลดความซับซ้อนของข้อมูลและทำให้การวิเคราะห์ง่ายขึ้น

- วัยเด็กตอนกลาง 7 - 12 ปี
- วัยรุ่น 13 - 18 ปี
- วัยผู้ใหญ่ตอนต้น 19 - 35 ปี
- วัยกลางคน 36 - 59 ปี
- วัยสูงอายุตั้งแต่ 60 ปีขึ้นไป

3.3 นำปัจจัยที่ไม่ได้ใช้ในการวิเคราะห์ออกจากข้อมูล

ประกอบด้วย Province, District, Sub-District

3.4 การแปลงข้อมูลเชิงคุณภาพเป็นเชิงปริมาณ

ข้อมูลที่อยู่ในรูปแบบ "ใช่" หรือ "ไม่ใช่" ได้ถูกแปลงเป็นเลข 0 และ 1 เพื่อให้สามารถนำไปวิเคราะห์ทางสถิติได้ง่ายขึ้น

Regional	Province	District	Sub-District	Gender	Age	Age Range	Relation	Mariage Re	Alcohol	Drug	Authoritati	Rage	Jealous	Divorce	Health Prol	Mental Pro	Gambling A	Economics	Stress
ภาคกลาง	กรุงเทพมห	เขตหนองจ	สำพักขี้	หญิง	61	วัยสูงอายุ 6	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตบางแค	บางแค	หญิง	26	วัยผู้ใหญ่ตอ	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตทวีวัฒน	ศาลาธรรมส	หญิง	55	วัยกลางคน	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตบางขื่อ	วงศ์สว่าง	หญิง	64	วัยสูงอายุ 6	แยกกันอยู่	จดทะเบียนา	ไม่	ไม่	ใช่	ใช่	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตประเวศ	ดอกไม้	ชาย	33	วัยผู้ใหญ่ตอ	สมรส	ยังไม่ได้จด	ใช่	ไม่	ไม่	ใช่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตคลองเต	พระโขนงเห	ชาย	36	วัยกลางคน	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตประเวศ	ประเวศ	ชาย	40	วัยกลางคน	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตคลองส	บางชัน	ชาย	42	วัยกลางคน	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตประเวศ	ประเวศ	ชาย	18	วัยรุ่น 13 -	โสด	ไม่ระบุ	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตบางเขน	อนุสาวรีย์	ชาย	33	วัยผู้ใหญ่ตอ	สมรส	จดทะเบียนา	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตทุ่งครุ	บางมด	ชาย	22	วัยผู้ใหญ่ตอ	แยกกันอยู่	ยังไม่ได้จด	ไม่	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตดุสิต	ดุสิต	หญิง	68	วัยสูงอายุ 6	สมรส	จดทะเบียนา	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตมีนบุรี	มีนบุรี	ชาย	36	วัยกลางคน	โสด	ไม่ระบุ	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตทุ่งครุ	บางมด	ชาย	40	วัยกลางคน	สมรส	ยังไม่ได้จด	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตภาษีเจริ	บางหว้า	ชาย	48	วัยกลางคน	สมรส	ยังไม่ได้จด	ใช่	ไม่	ใช่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตสะพานส	สะพานสูง	ชาย	27	วัยผู้ใหญ่ตอ	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตสายไหม	คลองถนน	ชาย	39	วัยกลางคน	สมรส	จดทะเบียนา	ไม่	ใช่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตจอมทอ	บางค้อ	ชาย	73	วัยสูงอายุ 6	แยกกันอยู่	จดทะเบียนา	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตสวนหล	ออนนุช	ชาย	30	วัยผู้ใหญ่ตอ	สมรส	ยังไม่ได้จด	ใช่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตลาดพร้	ลาดพร้าว	ชาย	52	วัยกลางคน	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ใช่	ไม่	ใช่	
ภาคกลาง	กรุงเทพมห	เขตพระนคร	บ้านพานถม	ชาย	56	วัยกลางคน	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตบางพลั	บางพลัด	ชาย	39	วัยกลางคน	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตจอมทอ	บางขุนเทีย	หญิง	45	วัยกลางคน	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ใช่	
ภาคกลาง	กรุงเทพมห	เขตบางขุนเ	แสมดำ	ชาย	45	วัยกลางคน	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตหนองแ	หนองค้างพ	ชาย	44	วัยกลางคน	สมรส	จดทะเบียนา	ไม่	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตวังทองห	วังทองหลา	ชาย	38	วัยกลางคน	สมรส	จดทะเบียนา	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตหนองแ	หนองแขม	ชาย	50	วัยกลางคน	สมรส	ยังไม่ได้จด	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตลาดกระ	คลองสามป	หญิง	58	วัยกลางคน	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตภาษีเจริ	บางหว้า	ชาย	48	วัยกลางคน	สมรส	ยังไม่ได้จด	ใช่	ไม่	ใช่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตบางแค	บางแค	ชาย	30	วัยผู้ใหญ่ตอ	โสด	ไม่ระบุ	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตบางแค	บางแค	หญิง	26	วัยผู้ใหญ่ตอ	โสด	ไม่ระบุ	ไม่	ไม่	ใช่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตลาดกระ	คลองสองต	ชาย	55	วัยกลางคน	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตคลองส	บางชัน	ชาย	60	วัยสูงอายุ 6	สมรส	จดทะเบียนา	ไม่	ไม่	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตบางขุนเ	แสมดำ	ชาย	44	วัยกลางคน	สมรส	ยังไม่ได้จด	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตหลักสี่	ตลาดบางเข	ชาย	33	วัยผู้ใหญ่ตอ	สมรส	ยังไม่ได้จด	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตบางแค	หลักสอง	หญิง	32	วัยผู้ใหญ่ตอ	โสด	ไม่ระบุ	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตประเวศ	ประเวศ	ชาย	68	วัยสูงอายุ 6	โสด	ไม่ระบุ	ไม่	ไม่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	ไม่	
ภาคกลาง	กรุงเทพมห	เขตราษฎร์บ	บางปะกอก	ชาย	25	วัยผู้ใหญ่ตอ	โสด	ไม่ระบุ	ไม่	ไม่	ไม่	ใช่	ใช่	ไม่	ไม่	ไม่	ไม่	ไม่	

(ภาพชุดข้อมูล Before Data Preparation)

ID	Regional	Age	Age Range_วัยเด็กตอนกลาง 7 - 12 ปี	Age Range_วัยรุ่น 13 - 18 ปี	Age Range_วัยผู้ใหญ่ตอนต้น 19 - 35 ปี	Age Range_วัยกลางคน 36 - 59 ปี	Age Range_วัยสูงอายุ 60 ปีขึ้นไป	Relation_แยกกันอยู่	Relation_โสด	Substance Use	Aggression	Relationship Issues	Health Issues	Financial Issues
0	ภาคกลาง	61	0	0	0	0	1	0	1	0	1	0	0	0
1	ภาคกลาง	26	0	0	1	0	0	0	1	0	1	0	0	0
2	ภาคกลาง	55	0	0	0	1	0	0	1	0	0	0	1	0
3	ภาคกลาง	64	0	0	0	0	1	1	0	0	2	0	1	0
4	ภาคกลาง	33	0	0	1	0	0	0	0	1	1	1	0	0
5	ภาคกลาง	36	0	0	0	1	0	0	1	0	0	1	0	0
6	ภาคกลาง	40	0	0	0	1	0	0	1	0	1	0	0	0
7	ภาคกลาง	42	0	0	0	1	0	0	1	0	0	1	0	0
8	ภาคกลาง	18	0	1	0	0	0	0	1	1	0	0	0	0
9	ภาคกลาง	33	0	0	1	0	0	0	0	0	1	0	0	0
10	ภาคกลาง	22	0	0	1	0	0	1	0	0	0	1	0	0
11	ภาคกลาง	68	0	0	0	0	1	0	0	0	0	0	1	0
12	ภาคกลาง	36	0	0	0	1	0	0	1	1	0	0	0	0
13	ภาคกลาง	40	0	0	0	1	0	0	0	1	0	0	0	0
14	ภาคกลาง	48	0	0	0	1	0	0	0	1	1	1	0	0
15	ภาคกลาง	27	0	0	1	0	0	0	1	0	1	0	0	0
16	ภาคกลาง	39	0	0	0	1	0	0	0	1	1	0	0	0
17	ภาคกลาง	73	0	0	0	0	1	1	0	0	1	0	0	0
18	ภาคกลาง	30	0	0	1	0	0	0	0	2	0	1	0	0
19	ภาคกลาง	52	0	0	0	1	0	0	1	0	0	0	1	1
20	ภาคกลาง	56	0	0	0	1	0	0	1	0	1	0	0	0
21	ภาคกลาง	39	0	0	0	1	0	0	1	0	1	0	0	0
22	ภาคกลาง	45	0	0	0	1	0	0	1	0	0	0	0	1
23	ภาคกลาง	45	0	0	0	1	0	0	1	0	1	0	0	0
24	ภาคกลาง	44	0	0	0	1	0	0	0	0	0	1	0	0
25	ภาคกลาง	38	0	0	0	1	0	0	0	0	1	0	1	0
26	ภาคกลาง	50	0	0	0	1	0	0	0	1	0	0	0	0
27	ภาคกลาง	58	0	0	0	1	0	0	1	0	0	0	1	0
28	ภาคกลาง	48	0	0	0	1	0	0	0	1	1	1	0	0
29	ภาคกลาง	30	0	0	1	0	0	0	1	1	0	0	0	0
30	ภาคกลาง	26	0	0	1	0	0	0	1	0	2	0	0	0
31	ภาคกลาง	55	0	0	0	1	0	0	1	0	0	1	0	0
32	ภาคกลาง	60	0	0	0	0	1	0	0	0	0	1	0	0
33	ภาคกลาง	44	0	0	0	1	0	0	0	1	0	0	0	0
34	ภาคกลาง	33	0	0	1	0	0	0	0	1	0	0	0	0
35	ภาคกลาง	32	0	0	1	0	0	0	1	0	1	0	0	0
36	ภาคกลาง	68	0	0	0	0	1	0	1	0	1	0	0	0
37	ภาคกลาง	25	0	0	1	0	0	0	1	0	1	1	0	0
38	ภาคกลาง	61	0	0	0	0	1	0	0	0	1	0	0	0
39	ภาคกลาง	30	0	0	1	0	0	0	1	0	0	0	0	1
40	ภาคกลาง	53	0	0	0	1	0	0	1	1	0	0	0	0
41	ภาคกลาง	24	0	0	1	0	0	0	1	0	0	0	0	1
42	ภาคกลาง	37	0	0	0	1	0	0	0	0	0	0	0	1
43	ภาคกลาง	43	0	0	0	1	0	0	0	1	0	0	0	0
44	ภาคกลาง	31	0	0	1	0	0	0	1	1	2	0	0	0
45	ภาคกลาง	29	0	0	1	0	0	0	1	0	1	0	0	0

(ภาพชุดข้อมูล After Data Preparation)

4. Modeling

4.1

```
pref_models = setup(
    data=pref_data,
    session_id=123,
    ignore_features=['Province', 'District', 'Sub-District', 'Gender', 'Age Range', 'Relation', 'Mariage Registration']
)
```

	Description	Value
0	Session id	123
1	Original data shape	(564, 14)
2	Transformed data shape	(564, 7)
3	Ignore features	7
4	Numeric features	7
5	Preprocess	True
6	Imputation type	simple
7	Numeric imputation	mean
8	Categorical imputation	mode
9	CPU Jobs	-1
10	Use GPU	False
11	Log Experiment	False
12	Experiment Name	cluster-default-name
13	USI	a2c2

4.2

```
models = {
    "KMeans": create_model('kmeans', num_clusters=3),
    "DBSCAN": create_model('dbscan'),
    "Hierarchical": create_model('hclust')
}
```

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.4982	1203.8465	0.6161	0	0	0

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0	0	0	0	0	0

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.4244	1096.5299	0.6385	0	0	0

5. Evaluation

5.1

```
selected_model = models["KMeans"]
pref_clustered = assign_model(selected_model)
pref_clustered
```

	Regional	Age	Substance Use	Aggression	Relationship Issues	Health Issues	Financial Issues	Cluster
0	0	61	0	1	0	0	0	Cluster 1
1	0	26	0	1	0	0	0	Cluster 0
2	0	55	0	0	0	1	0	Cluster 1
3	0	64	0	2	0	1	0	Cluster 1
4	0	33	1	1	1	0	0	Cluster 0
...	...	...	...	...	...	...	...	...
559	3	25	2	0	0	1	1	Cluster 0
560	3	47	1	0	0	0	0	Cluster 2
561	3	39	1	0	1	0	0	Cluster 2
562	3	40	2	0	1	0	0	Cluster 2
563	3	64	2	1	0	0	0	Cluster 1

564 rows × 8 columns

6. Deployment

6.1 บันทึกโมเดลที่ต้องการนำไปใช้

```
save_model(final_model, 'final_cluster_model')
```

```
Transformation Pipeline and Model Successfully Saved
(Pipeline(memory=Memory(location=None),
          steps=[('numerical_imputer',
                  TransformerWrapper(include=['Regional', 'Age', 'Substance Use',
                                              'Aggression',
                                              'Relationship Issues',
                                              'Health Issues',
                                              'Financial Issues'],
                                     transformer=SimpleImputer())),
                 ('categorical_imputer',
                  TransformerWrapper(include=[],
                                     transformer=SimpleImputer(strategy='most_frequent'))),
                 ('trained_model',
                      Regional  Age  Substance Use  Aggression  Relationship Issues  \
0           0   61             0             1                   0
1           0   26             0             1                   0
2           0   55             0             0                   0
3           0   64             0             2                   0
4           0   33             1             1                   1
..        ...  ...           ...           ...                 ...
559         3   25             2             0                   0
560         3   47             1             0                   0
561         3   39             1             0                   1
562         3   40             2             0                   1
563         3   64             2             1                   0

     Health Issues  Financial Issues    Cluster
0                0                 0  Cluster 1
1                0                 0  Cluster 0
2                1                 0  Cluster 1
3                1                 0  Cluster 1
4                0                 0  Cluster 0
..             ...               ...        ...
559              1                 1  Cluster 0
560              0                 0  Cluster 2
561              0                 0  Cluster 2
562              0                 0  Cluster 2
563              0                 0  Cluster 1

[564 rows x 8 columns])]),
'final_cluster_model.pkl')
```

6.2 นำเข้า library ที่จำเป็นในการใช้งาน

```
%%writefile test.py
import streamlit as st
import pandas as pd
import numpy as np
import json
import joblib
from sklearn.cluster import AgglomerativeClustering
from sklearn.metrics.pairwise import euclidean_distances
```

- Streamlit เป็น library ของ Python ที่ใช้สำหรับสร้างเว็บแอปพลิเคชันสำหรับ Machine Learning และ Data Science ได้อย่างรวดเร็วและง่ายดาย
- Pandas เป็น library ที่ใช้สำหรับการจัดการและวิเคราะห์ข้อมูลที่มีโครงสร้าง เช่น ตาราง (DataFrames)
- NumPy เป็น library ที่ใช้สำหรับการคำนวณทางคณิตศาสตร์และวิทยาศาสตร์ โดยเฉพาะการจัดการกับอาร์เรย์ (Arrays)

- json เป็น library ที่ใช้ในการทำงานกับข้อมูลในรูปแบบ JSON (JavaScript Object Notation) ซึ่งเป็นรูปแบบข้อมูลที่นิยมใช้ในการแลกเปลี่ยนข้อมูลระหว่างแอปพลิเคชัน
- joblib เป็น library ที่ใช้สำหรับการบันทึกและโหลดโมเดล Machine Learning เพื่อนำไปใช้งานในภายหลัง

- from sklearn.cluster import AgglomerativeClustering เป็นการนำเข้าคลาส AgglomerativeClustering จาก module cluster ของ sklearn ซึ่งเป็นอัลกอริทึมที่ใช้สำหรับการจัดกลุ่มข้อมูลแบบ Hierarchical Clustering.
- from sklearn.metrics.pairwise import euclidean_distances เป็นการนำเข้าฟังก์ชัน euclidean_distances ซึ่งใช้สำหรับคำนวณระยะห่างแบบยุคลิด (Euclidean distance) ระหว่างจุดข้อมูล

6.3

```
# โหลดข้อมูลจาก CSV แล้วลบคอลัมน์ที่ไม่ต้องการ ("dtype", "Regional")
try:
    final_cluster_data = pd.read_csv("/content/sample_data/pref_data.csv")
    final_cluster_data = final_cluster_data.drop(columns=["dtype", "Regional"], errors="ignore")
    st.sidebar.success("Successfully loaded clustering data from /content/sample_data/pref_data.csv")

    st.sidebar.info(f"Data shape: {final_cluster_data.shape}")
    if 'Cluster' in final_cluster_data.columns:
        st.sidebar.info(f"Found {final_cluster_data['Cluster'].nunique()} clusters in the data")
except Exception as e:
    st.sidebar.error(f"Error loading CSV data: {e}")
    st.stop()
```

6.4

```
# โหลดข้อมูลภูมิภาค
try:
    with open("/content/sample_data/region_province_district_subdistrict.json", "r", encoding="utf-8") as f:
        region_data = json.load(f)
    st.sidebar.success("Successfully loaded region data")
except Exception as e:
    st.sidebar.error(f"Error loading region data: {e}")
    st.stop()
```

6.5

```python
# โหลด label encoders (หากใช้งาน)
try:
    les = joblib.load("/content/les.pkl")
    st.sidebar.success("Successfully loaded label encoders")
except Exception as e:
    st.sidebar.error(f"Error loading label encoders: {e}")
    st.stop()
```

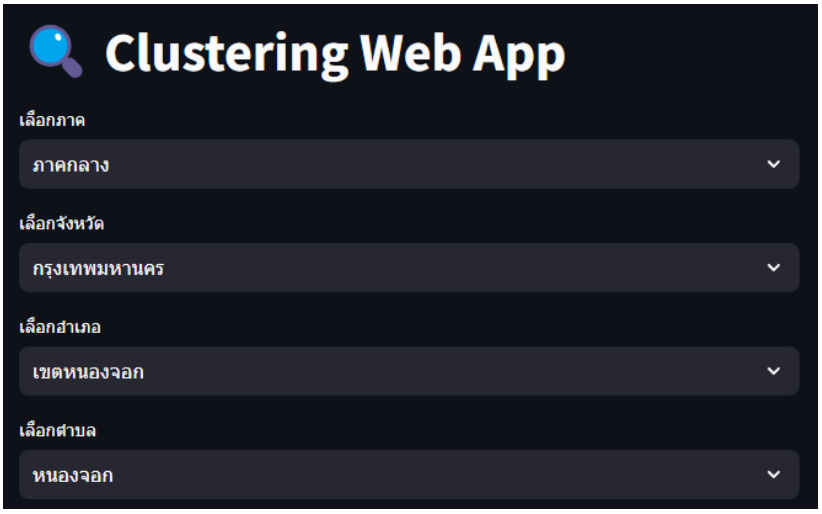
6.6 ออกแบบฟอร์มการนำเข้าข้อมูลและประมวลผลคำตอบ

```python
try:
    # ระบุคอลัมน์ที่ไม่มีผลต่อการ clustering
    irrelevant_cols = ['Province', 'District', 'Sub-District', 'Gender', 'Age Range', 'Relation', 'Mariage Registration']

    # เตรียมข้อมูลสำหรับการวิเคราะห์ โดยไม่รวมคอลัมน์ "Cluster", "Regional" และคอลัมน์ที่ไม่มีผลต่อ clustering
    input_for_clustering = pd.DataFrame()
    for col in ref_cols:
        if col in ['Cluster', 'Regional'] or col in irrelevant_cols:
            continue
        if col in new_df.columns:
            input_for_clustering[col] = new_df[col]
        else:
            input_for_clustering[col] = 0

    st.write("### 🔍 ข้อมูลที่ใช้ในการวิเคราะห์ (เฉพาะฟีเจอร์ที่มีผล):")
    st.dataframe(input_for_clustering)

    # กำหนดคอลัมน์ที่ใช้สำหรับวิเคราะห์
    feature_cols = input_for_clustering.columns.tolist()
    if not feature_cols:
        st.error("ไม่พบคอลัมน์ที่สามารถใช้วิเคราะห์ได้")
        st.stop()

    reference_features = final_cluster_data[feature_cols]
    input_features = input_for_clustering[feature_cols]

    # st.write("### 📊 คอลัมน์ที่ใช้สำหรับการวิเคราะห์:")
    # st.write(", ".join(feature_cols))

    # คำนวณระยะห่างแบบ Euclidean
    distances = euclidean_distances(input_features, reference_features)
    most_similar_idx = np.argmin(distances[0])

    # ดึงค่า Cluster จากข้อมูลอ้างอิง (ถ้ามีคอลัมน์ Cluster)
    if 'Cluster' in final_cluster_data.columns:
        assigned_cluster = final_cluster_data.iloc[most_similar_idx]['Cluster']
        # หาก assigned_cluster เป็นสตริงที่มี "Cluster " นำหน้า ให้ตัดออก
        if isinstance(assigned_cluster, str) and assigned_cluster.startswith("Cluster"):
            try:
                assigned_cluster = int(assigned_cluster.replace("Cluster", "").strip())
            except Exception as e:
                st.error(f"ไม่สามารถแปลงค่า cluster ได้: {e}")
                st.stop()
    else:
        combined_data = pd.concat([reference_features, input_features], ignore_index=True)
        model = AgglomerativeClustering(n_clusters=3)
        clusters = model.fit_predict(combined_data)
        assigned_cluster = clusters[-1]

    st.subheader(f"📌 คุณถูกจัดอยู่ในกลุ่มที่: **Cluster {assigned_cluster}**")
    st.write(f"ระยะห่างจากจุดอ้างอิงใกล้เคียงที่สุด: {distances[0][most_similar_idx]:.2f}")
```

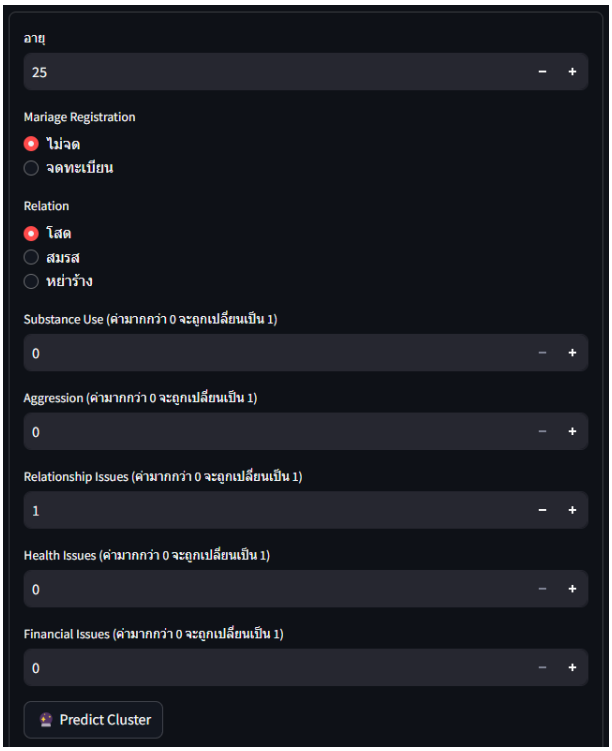
```
    st.subheader(f"📌 คุณถูกจัดอยู่ในกลุ่มที่: **Cluster {assigned_cluster}**")
    st.write(f"ระยะทางจากจุดอ้างอิงใกล้เคียงที่สุด: {distances[0][most_similar_idx]:.2f}")

    st.write("### 📊 ลักษณะของกลุ่ม:")
    if int(assigned_cluster) == 0:
        st.info("Cluster 0: กลุ่มที่มีความเสี่ยงต่ำ")
    elif int(assigned_cluster) == 1:
        st.warning("Cluster 1: กลุ่มที่มีความเสี่ยงปานกลาง")
    elif int(assigned_cluster) == 2:
        st.error("Cluster 2: กลุ่มที่มีความเสี่ยงสูง")
    else:
        st.info(f"Cluster {assigned_cluster}: ไม่มีข้อมูลเพิ่มเติมสำหรับกลุ่มนี้")

    # แสดงข้อมูลอ้างอิงที่ใกล้เคียงที่สุด (รวม Cluster ด้วย)
    st.write("### 📊 ข้อมูลอ้างอิงที่ใกล้เคียงที่สุด:")
    similar_data = final_cluster_data.iloc[most_similar_idx].to_frame().T
    st.dataframe(similar_data)

except Exception as e:
    st.error(f"เกิดข้อผิดพลาดในการทำนาย: {e}")
    st.info("รายละเอียดเพิ่มเติม:")
    import traceback
    st.code(traceback.format_exc())
    st.write("### 📝 ตัวอย่างข้อมูลอ้างอิง:")
    st.dataframe(final_cluster_data.head())
```

6.7 ตัวอย่างแบบฟอร์ม

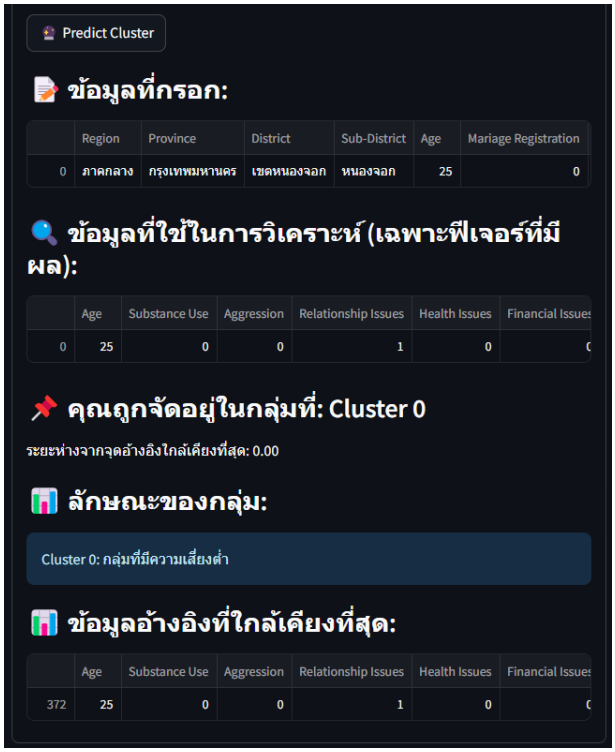
6.7.1 กรอกข้อมูล ภูมิภาค/จังหวัด/อำเภอ/ตำบล



6.7.2 กรอกข้อมูล อายุ/ข้อมูลจดการทะเบียน/ปัจจัยที่เกิดขึ้น



6.7.3 แสดงข้อมูลที่กรอก/ข้อมูลที่ใช้วิเคราะห์/ลักษณะของกลุ่ม/ข้อมูลอ้างอิง



เอกสารอ้างอิง

Dataset : https://data.go.th/dataset/gdpublish-dwf-pb-dmv01-050507-04

Business Understanding: https://rsucon.rsu.ac.th/files/proceedings/nation2019/NA19-136.pdf

https://www.m-society.go.th/ewtadmin/ewt/mso_web/download/article/article_20150206144414.pdf

Modeling: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

เอกสารประกอบการเรียน:
https://colab.research.google.com/drive/1H_xg12m0jeJDgqEe99QqTu26MaQ5g-52?authuser=1&usp=classroom_web#scrollTo=7JLDYx_mu7Ss