Final Assignment

Region : Central
Province 1 : Chon Buri
Province 2 : Prachin Buri
Province 3 : Samut Songkhram
Classification :Human Development

โดย
นาย ปพนธ์ ชุณหคล้าย 6210503691

เสนอ
ผศ.ดร.สุภาพร เอื้อจงมานี

รายงานฉบับนี้เป็นส่วนหนึ่งของรายวิชา
สถิติสำหรับการประยุกต์ทางวิศวกรรมคอมพิวเตอร์(01204314)
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
มหาวิทยาลัยเกษตรศาสตร์
ปีการศึกษา 2564 ภาคปลาย

สารบัญ

Part 1: Feature Selection and Dimensionality Reduction

<u>วัตถุประสงค์</u>
- วิเคราะห์ว่า ปัจจัยใดส่งผลต่อ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita) ของจังหวัดที่ได้รับมอบหมาย

โดยปัจจัยที่ได้เพิ่มเข้ามา 11 ปัจจัยมีดังนี้
ปัจจัยระดับจังหวัด 6 ปัจจัย

1.จำนวนตำแหน่งงานว่าง (ความต้องการแรงงาน) ของแต่ล่ะจังหวัด โดยเหตุผลเนื่องจากคิดว่าหากปีไหนที่มีผลผลิตภัณฑ์จังหวัดต่อคน (GPP per capita) สูงอาจจะตีความได้ว่า จำนวนตำแหน่งงานว่างน่าจะน้อยลงเนื่องจากทุกคนมีงานทำและทำให้ผลิตภัณฑ์สามารถผลิตได้มากขึ้นนั้นเอง จึงคิดว่าปัจจัยนี้น่าจะส่งผลต่อ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita)
แหล่งที่มา : http://statbbi.nso.go.th/staticreport/Page/sector/TH/report/sector_02_26_TH_.xlsx

2.จำนวนผู้ที่มาเยี่ยมเยือนในแต่ล่ะจังหวัด โดยเหตุผลเนื่องจากคิดว่าหากปีไหนที่มีผู้มาเยี่ยมเยือนในจังหวัดสูงอาจตีความได้ว่ามีการเข้ามาเที่ยว,ซื้อสินค้า,บริการของจังหวัดนั้นๆส่งผลให้จังหวัดนั้นๆมีการเติบโตทางด้านเศรษฐกิจซึ่งจะนำไปสู่ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita) ที่มากขึ้นจึงคิดว่าปัจจัยนี้น่าจะส่งผลต่อ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita)
แหล่งที่มา :http://statbbi.nso.go.th/staticreport/Page/sector/TH/report/sector_17_12101_TH_.xlsx

3.ดัชนีราคาผู้บริโภคทั่วไป(CPI)ที่จำแนกเป็นรายจังหวัด โดย CPI คือเครื่องมือทางสถิติที่ใช้วัดการเปลี่ยนแปลงราคาขายปลีกของสินค้าและ บริการโดยเฉลี่ยที่ผู้บริโภคจ่ายเพื่อซื้อสินค้าและบริการจ านวนหนึ่ง ณ เวลาหนึ่งเทียบกับแต่ล่ะปี โดยเหตุผลเนื่องจากคิดว่าหากปีไหนมีค่า CPI สูงอาจหมายถึงราคาสินค้ามีราคาแพงอาจส่งผลให้มี ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita) ที่น้อยลง จึงคิดว่าปัจจัยนี้น่าจะส่งผลต่อ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita)
แหล่งที่มา : http://statbbi.nso.go.th/staticreport/Page/sector/TH/report/sector_14_24_TH_.xlsx

4.จำนวนคนจน (ด้านรายจ่าย) จำแนกเป็นรายจังหวัด โดยเหตุผลเนื่องจากคิดว่าหาก มีจำนวนคนจนเยอะอาจหมายถึงเศรษฐกิจของจังหวัดนั้นๆแย่ซึ่งน่าจะส่งผลให้ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita) น้อยลง จึงคิดว่าปัจจัยนี้น่าจะส่งผลต่อ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita)
แหล่งที่มา : http://statbbi.nso.go.th/staticreport/Page/sector/TH/report/sector_08_11_TH_.xlsx

5.สถิติโรงงานอุตสาหกรรมที่จดทะเบียนกับกระทรวงอุตสาหกรรม และได้รับอนุญาตให้ประกอบกิจการใหม่ (ตามพระราชบัญญัติโรงงาน พ.ศ. 2535) จำแนกตามจังหวัด โดยเหตุผลเนื่องจากคิดว่าหากมีโรงงานอุตสาหกรรมที่เปิดใหม่เยอะน่าจะส่งผลให้คนส่วนใหญ่มีงานทำเพิ่มมากขึ้นและช่วยกระตุ้นเศรษฐกิจของจังหวัดนั้นๆส่งผลให้น่าจะมี ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita) ที่สูงขึ้นจึงคิดว่าปัจจัยนี้น่าจะส่งผลต่อ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita)
แหล่งที่มา : http://statbbi.nso.go.th/staticreport/Page/sector/TH/report/sector_12_3_TH_.xls

6.จำนวนเงินให้สินเชื่อคงเหลือของธนาคารออมสิน จำแนกตามจังหวัด โดยเหตุผลเนื่องจากคิดว่าหากปล่อยให้สินเชื่อเยอะอาจสามารถตีความได้ว่ามีการกู้ยืมเพื่อนำไปทำอะไรสักอย่าง เช่น สร้างธุรกิจหรือกิจการใหม่ๆ หรือเป็นการนำไปซื้อบ้าน แต่ก็ถือเป็นการกระตุ้นเศรษฐกิจในทางหนึ่ง ซึ่งอาจจะส่งผลทางอ้อมให้ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita) สูงขึ้น จึงคิดว่าปัจจัยนี้น่าจะส่งผลต่อ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita)

แหล่งที่มา : http://statbbi.nso.go.th/staticreport/Page/sector/TH/report/sector_18_14_TH_.xlsx

ปัจจัยระดับประเทศ 5 ปัจจัย

1.Households and NPISHs final consumption expenditure (% of GDP) หรือก็คือรายจ่ายครัวเรือนโดยอิงจากเปอร์เซ็นต์ของ GDP โดยเหตุผลเนื่องจากคิดว่าหาก มีค่าใช้จ่ายครัวเรือนที่สูงน่าจะตีความได้ว่ามีการใช้จ่ายที่มากขึ้นส่งผลให้เศรษฐกิจดีขึ้นซึ่งอาจจะส่งผลให้ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita) สูงขึ้น จึงคิดว่าปัจจัยนี้น่าจะส่งผลต่อ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita)
แหล่งที่มา : World Development Indicators | Data Catalog (worldbank.org)

2.Central government debt, total (% of GDP) หรือก็คือหนี้รัฐบาล โดยอิงจากเปอร์เซ็นต์ของ GDP ซึ่งเป็นส่วนหนึ่งของหนี้สาธารณะ รัฐบาลจึงอาจจะต้องออกนโยบายบางอย่างเพื่อนำเงินไปชดใช้หนี้สาธารณะ ซึ่งอาจจะเป็นกระตุ้นเศรษฐกิจหรือกระตุ้นการค้าขายให้แต่ละจังหวัด จึงอาจจะส่งผลให้ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita) สูงขึ้นได้ จึงคิดว่าปัจจัยนี้น่าจะส่งผลต่อ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita)
แหล่งที่มา : World Development Indicators | Data Catalog (worldbank.org)

3.Exports of goods and services (% of GDP) หรือก็คือ การส่งออกสินค้าและบริการโดยอิงจากเปอร์เซ็นต์ของ GDP โดยเหตุผลเนื่องจากคิดว่าหากปีไหนมีการส่งออกสินค้าหรือบริการเยอะแสดงว่าประเทศไทยมีสินค้าหรือบริการที่ต่างชาติต้องการจึงอาจจะมีการเพิ่มกำลังผลิตซึ่งอาจจะส่งผลให้ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita) สูงขึ้นจึงคิดว่าปัจจัยนี้น่าจะส่งผลต่อ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita)
แหล่งที่มา : World Development Indicators | Data Catalog (worldbank.org)

4.Broad money (% of GDP) หรือก็คือเงินในวงกว้างโดยอิงจากเปอร์เซ็นต์ของ GDPที่ถือเป็นวิธีที่ครอบคลุมมากที่สุดในการประเมินสถานะของปริมาณเงินในประเทศหรือตลาดโลกเป็นสิ่งสำคัญในการระบุโอกาสในการสร้างผลกำไรจากการลงทุน ด้วยการทำหน้าที่เป็นตัวบ่งชี้สถานะทางการเงินซึ่งหากปีไหนมีเงินในวงกว้างเยอะก็อาจจะส่งผลให้ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita) สูงขึ้น จึงคิดว่าปัจจัยนี้น่าจะส่งผลต่อ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita)
แหล่งที่มา : World Development Indicators | Data Catalog (worldbank.org)

5.Manufacturing, value added (% of GDP) หรือก็คือค่าใช้จ่ายการผลิตที่เพิ่มเข้าไปในสินค้าโดยอิงจากเปอร์เซ็นต์ของ GDP ซึ่งอาจตีความได้ว่าหากปีไหนที่มีค่าใช้จ่ายในการผลิตสูงอาจส่งผลให้ความสามารถในการผลิตน้อยลงและอาจจะส่งผลให้ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita)น้อยลงจึงคิดว่าปัจจัยนี้น่าจะส่งผลต่อ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita)
แหล่งที่มา : World Development Indicators | Data Catalog (worldbank.org)

โดยจังหวัดที่ได้รับมอบหมายมีดังนี้ :
Province 1 : Chon Buri
Province 2 : Prachin Buri
Province 3 : Samut Songkhram

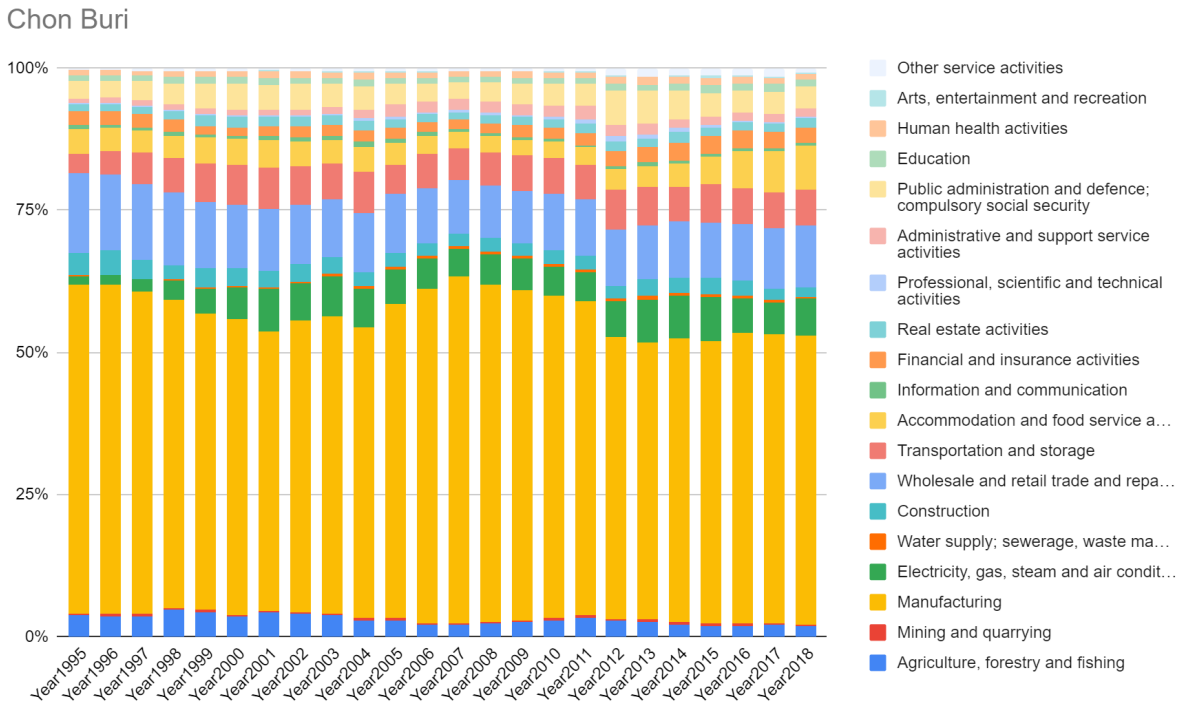
จังหวัด ชลบุรี

1.1 ระบุปัจจัยที่เพิ่มขึ้นมา แหล่งข้อมูล และเหตุผลในการเลือกปัจจัยนั้น

ปัจจัยที่เพิ่ม	ระดับจังหวัดหรือระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018) หรือเป็นข้อมูลที่ทำ regression	หากทำ regression ระบุอัตราส่วนระหว่างข้อมูลที่มีกับ ข้อมูลที่ทำ regression
จำนวนตำแหน่งงานว่าง (ความต้องการแรงงาน) ของแต่ละจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (8:16)
จำนวนผู้ที่มาเยี่ยมเยือนในแต่ละจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (9:15)
ดัชนีราคาผู้บริโภคทั่วไป(CPI)ที่จำแนกเป็นรายจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (9:15)
จำนวนคนจน (ด้านรายจ่าย) จำแนกเป็นรายจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (8:16)
สถิติโรงงานอุตสาหกรรมที่จดทะเบียนกับกระทรวงอุตสาหกรรม และได้รับอนุญาตให้ประกอบกิจการใหม่ (ตามพระราชบัญญัติโรงงาน พ.ศ. 2535) จำแนกตามจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (8:16)
จำนวนเงินให้สินเชื่อคงเหลือของธนาคารออมสิน จำแนกตามจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (8:16)
Households and NPISHs final consumption	ระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018)	-

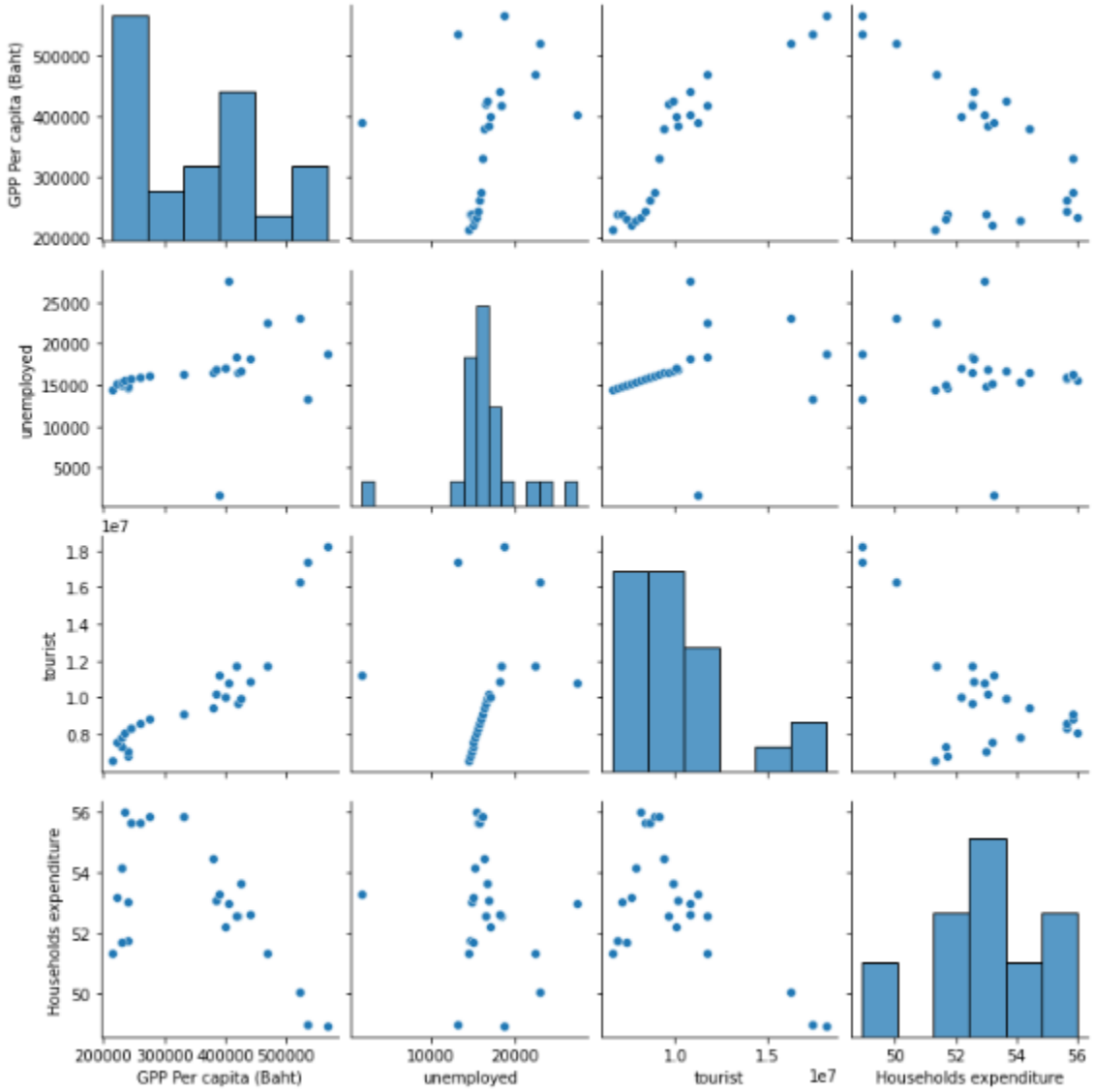
expenditure (% of GDP)			
Central government debt, total (% of GDP)	ระดับประเทศ	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (21:3)
Exports of goods and services (% of GDP)	ระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018)	-
Broad money (% of GDP)	ระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018)	-
Manufacturing, value added (% of GDP)	ระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018)	-

1.2 ทำ data exploration แสดงและบรรยายประกอบ
-visualization ของสัดส่วน 19 หมวดย่อย ที่มีต่อ GPP per capita
-visualization ของความสัมพันธ์ของ ปัจจัยที่เลือกมา และ GPP per capita เลือกทำ 3 ปัจจัย



ในvisualization ของสัดส่วน 19 หมวดย่อย ที่มีต่อ GPP per capitaของจังหวัดชลบุรี (กำหนดให้ใน 1 กราฟแท่งคือค่าของแต่ล่ะปี)จะสังเกตุได้ว่าปัจจัยที่มีสัดส่วนมากที่สุดอันดับ 1 คือ Manufacturing (สีเหลือง) , อันดับที่ 2 คือ Wholesale and retail trade and repair of motor vehicles (สีฟ้า) และอันดับที่ 3

คือ Transportation and storage (สีแดงอ่อน) โดยในสัดส่วนในช่วงปี ค.ศ.2012-2018 จะมีค่าสัดส่วนที่ใกล้เคียงกัน



ในส่วนของvisualization ของความสัมพันธ์ของ ปัจจัยที่เลือกมา และ GPP per capita กับ 3 ปัจจัย ได้ทำ pair plot ของข้อมูล 4 ตัวคือ Households and NPISHs final consumption expenditure (% of GDP) , tourist (จำนวนผู้เยี่ยมเยือน โดยรวม) , unemployed (จำนวนตำแหน่งงานว่าง) และ GPP Per capita (Baht) จะเห็นได้ว่ามีเพียงคู่ GPP Per capita (Baht) และ tourist ที่ดูเหมือนจะมีความสัมพันธ์ไปในทางบวก

1.3 สำหรับข้อมูลของจังหวัดชลบุรี

เมื่อนำปัจจัยทุกตัวมาทำ regression แล้ว ได้ผลลัพธ์เป็นดังนี้

```
df = dummy_df.dropna()
X = df[input_vars]
y = df.GPP_Per_capita
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=99)
lm = LinearRegression()
lm.fit(X_train, y_train)
y_pred = lm.predict(X_test)
[np.sqrt(metrics.mean_squared_error(y_test,y_pred)),metrics.r2_score(y_test,y_pred)]
```

```
[48417.53897385943, 0.6600169623641962]
```

ได้ Regression model ที่มีค่า RMSE(root mean square error) = 48417.53897385943 และมีค่า R-Square = 0.6600169623641962 ซึ่งถือว่าเป็น model ที่ดีเนื่องจาก |r| = 0.8124 หมายความว่ามีค่าความสัมพันธ์สูง

เมื่อนำปัจจัยไปทำ feature selection ด้วยวิธี

- Filter 2 วิธี : ใช้วิธี Low Variance Filtering และวิธี High Correlation Filtering
- Wrapper 2 วิธี : ใช้วิธี Forward Feature Selection และวิธี Recursive Feature Elimination
- Embedded 1 วิธี : ใช้วิธี Random Forest

จากนั้น สรุปต่อไปนี้
มีปัจจัยใดบ้างที่ส่งผลต่อ GPP per capita

		Regression	
วิธี	จำนวนปัจจัยที่ส่งผล	R-Squared	RMSE
Low Variance Filtering	21 ปัจจัย	0.6618214839453294	48288.87558969716
High Correlation Filtering	27 ปัจจัย	0.6599497191696188	48422.32684364398
Forward Feature Selection	26 ปัจจัย	0.6441596817120553	49533.80189481629
Recursive Feature Elimination	9 ปัจจัย	0.9162697094168951	24027.876169616535
Random Forest	11 ปัจจัย	0.8306469079419364	34172.01993926983

- ปัจจัยที่ส่งผลจาก Feature selection นั้น มีกี่ปัจจัยที่เป็นระดับจังหวัด มีกี่ปัจจัยที่เป็นระดับประเทศ มีกี่ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) และมีกี่ปัจจัยที่เป็นปัจจัยได้จากการทำ regression/ interpolation

1.Low Variance Filtering ปัจจัยที่ส่งผลมีทั้งหมด 21 ปัจจัย ดังนี้

'Agriculture, forestry and fishing',
'Mining and quarrying',
'Manufacturing',
'Electricity, gas, steam and air conditioning supply',
'Water supply; sewerage, waste management and remediation activities',
'Construction',
'Wholesale and retail trade and repair of motor vehicles',
'Transportation and storage',
'Accommodation and food service activities',
'Information and communication',
'Financial and insurance activities',
'Real estate activities',
'Professional, scientific and technical activities',
'Administrative and support service activities',
'Public administration and defence; compulsory social security',
'Education',
'Human health activities',
'Other service activities',
'unemployed',
'tourist',
'Credit'

โดยมีปัจจัยระดับจังหวัด 3 ปัจจัย คือ 'unemployed'(จำนวนตำแหน่งงานว่าง) , 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'Credit'(จำนวนเงินให้สินเชื่อ) และไม่มีปัจจัยระดับประเทศที่ส่งผล ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 18 ปัจจัย และ ปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 3 ปัจจัยคือ 'unemployed' (จำนวนตำแหน่งงานว่าง) , 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'Credit'(จำนวนเงินให้สินเชื่อ)

2. High Correlation Filtering ปัจจัยที่ส่งผลมีทั้งหมด 27 ปัจจัย ดังนี้

'Agriculture, forestry and fishing',
'Mining and quarrying',
'Manufacturing',
'Electricity, gas, steam and air conditioning supply',
'Water supply; sewerage, waste management and remediation activities',
'Construction',
'Wholesale and retail trade and repair of motor vehicles',
'Transportation and storage',
'Accommodation and food service activities',

'Information and communication', 'Financial and insurance activities',
'Real estate activities',
'Professional, scientific and technical activities',
'Administrative and support service activities',
'Public administration and defence; compulsory social security',
'Education', 'Human health activities',
'Arts, entertainment and recreation', 'Other service activities',
'unemployed',
'tourist',
'CPI',
'factory',
'credit',
'Central government debt, total (% of GDP)',
'Exports of goods and services (% of GDP)',
'Broad money (% of GDP)'

โดยมีปัจจัยระดับจังหวัด 5 ปัจจัย คือ 'unemployed'(จำนวนตำแหน่งงานว่าง) , 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'Credit'(จำนวนเงินให้สินเชื่อ) ,'CPI'(ดัชนีราคาผู้บริโภคทั่วไป),'factory'(จำนวนโรงงาน), และปัจจัยระดับประเทศ 3 ปัจจัย คือ 'Central government debt total (% of GDP)', 'Exports of goods and services (% of GDP)', 'Broad money (% of GDP)'ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 21 ปัจจัย และ ปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 6 ปัจจัยคือ 'unemployed' (จำนวนตำแหน่งงานว่าง) , 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'Credit'(จำนวนเงินให้สินเชื่อ),'CPI'(ดัชนีราคาผู้บริโภคทั่วไป),'factory'(จำนวนโรงงาน), 'Central government debt total (% of GDP)'

3.Forward Feature Selection **ปัจจัยที่ส่งผลมีทั้งหมด 26 ปัจจัย ดังนี้**

'Agriculture, forestry and fishing',
'Mining and quarrying',
'Manufacturing',
'Electricity, gas, steam and air conditioning supply',
'Water supply; sewerage, waste management and remediation activities',
'Construction',
'Wholesale and retail trade and repair of motor vehicles',
'Transportation and storage',
'Accommodation and food service activities',
'Information and communication',
'Financial and insurance activities',
'Real estate activities',
'Professional, scientific and technical activities',

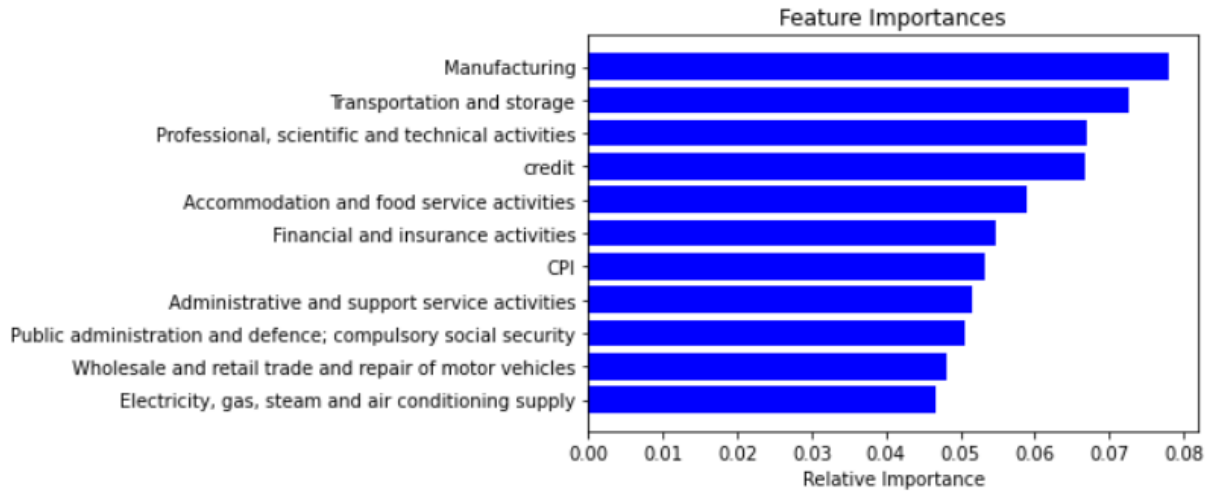
'Administrative and support service activities',
'Public administration and defence; compulsory social security',
'Education',
'Human health activities',
'Arts, entertainment and recreation',
'Other service activities',
'tourist',
'CPI',
'factory',
'credit',
'Households and NPISHs final consumption expenditure (% of GDP)',
'Central government debt, total (% of GDP)',
'Exports of goods and services (% of GDP)'

โดยมีปัจจัยระดับจังหวัด 4 ปัจจัย คือ 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'Credit'(จำนวนเงินให้สินเชื่อ) ,'CPI'(ดัชนีราคาผู้บริโภคทั่วไป),'factory'(จำนวนโรงงาน), และปัจจัยระดับประเทศ 3 ปัจจัย คือ 'Households and NPISHs final consumption expenditure (% of GDP)','Central government debt total (% of GDP)', 'Exports of goods and services (% of GDP)' ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 21 ปัจจัย และ ปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 5 ปัจจัยคือ 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'Credit'(จำนวนเงินให้สินเชื่อ),'CPI'(ดัชนีราคาผู้บริโภคทั่วไป),'factory' (จำนวนโรงงาน), 'Central government debt total (% of GDP)'

4.Recursive Feature Elimination ปัจจัยที่ส่งผลมีทั้งหมด 9 ปัจจัย ดังนี้
'Education',
'Arts, entertainment and recreation',
'CPI',
'poor',
'factory',
'Households and NPISHs final consumption expenditure (% of GDP)',
'Central government debt, total (% of GDP)',
'Exports of goods and services (% of GDP)',
'Broad money (% of GDP)'

โดยมีปัจจัยระดับจังหวัด 3 ปัจจัย คือ 'poor'(จำนวนคนจน) ,'CPI'(ดัชนีราคาผู้บริโภคทั่วไป), 'factory'(จำนวนโรงงาน), และปัจจัยระดับประเทศ 4 ปัจจัย คือ 'Households and NPISHs final consumption expenditure (% of GDP)','Central government debt total (% of GDP)', 'Exports of goods and services (% of GDP)','Broad money (% of GDP)'ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 5 ปัจจัย และ ปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 4 ปัจจัยคือ 'poor'(จำนวนคนจน) , 'CPI'(ดัชนีราคาผู้บริโภคทั่วไป),'factory'(จำนวนโรงงาน), 'Central government debt total (% of GDP)'

5.Random Forest ปัจจัยที่ส่งผลมีทั้งหมด 11 ปัจจัย ดังนี้



โดยมีปัจจัยระดับจังหวัด 2 ปัจจัย คือ 'Credit'(จำนวนเงินให้สินเชื่อ) ,'CPI'(ดัชนีราคาผู้บริโภคทั่วไป) และไม่มีปัจจัยระดับประเทศ ที่ส่งผล และ ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 9 ปัจจัย และปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 2 ปัจจัยคือ 'Credit'(จำนวนเงินให้สินเชื่อ) ,'CPI'(ดัชนีราคาผู้บริโภคทั่วไป)

1.4 เมื่อนำปัจจัยไปทำ dimensionality reduction ด้วยวิธี PCA

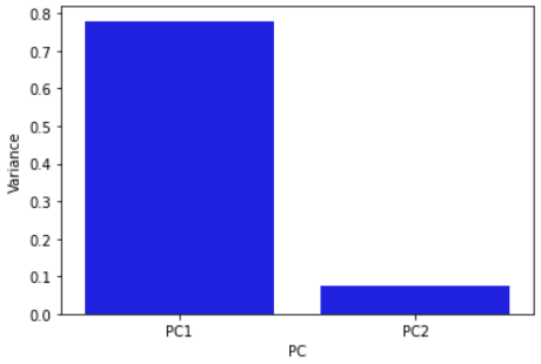
- มี component กี่ตัว (หรือสามารถลด dimension ได้เหลือเท่าใด) เหตุใดจึงเลือก component เท่านี้ ใช้ค่าอะไรประกอบการพิจารณา
-

```
pca.explained_variance_ratio_
```

```
array([0.77944952, 0.07515276])
```

```
temp_df = pd.DataFrame({'Variance':pca.explained_variance_ratio_,
                        'PC':['PC1','PC2']})
sns.barplot(x='PC',y="Variance", data=temp_df, color="blue")
```

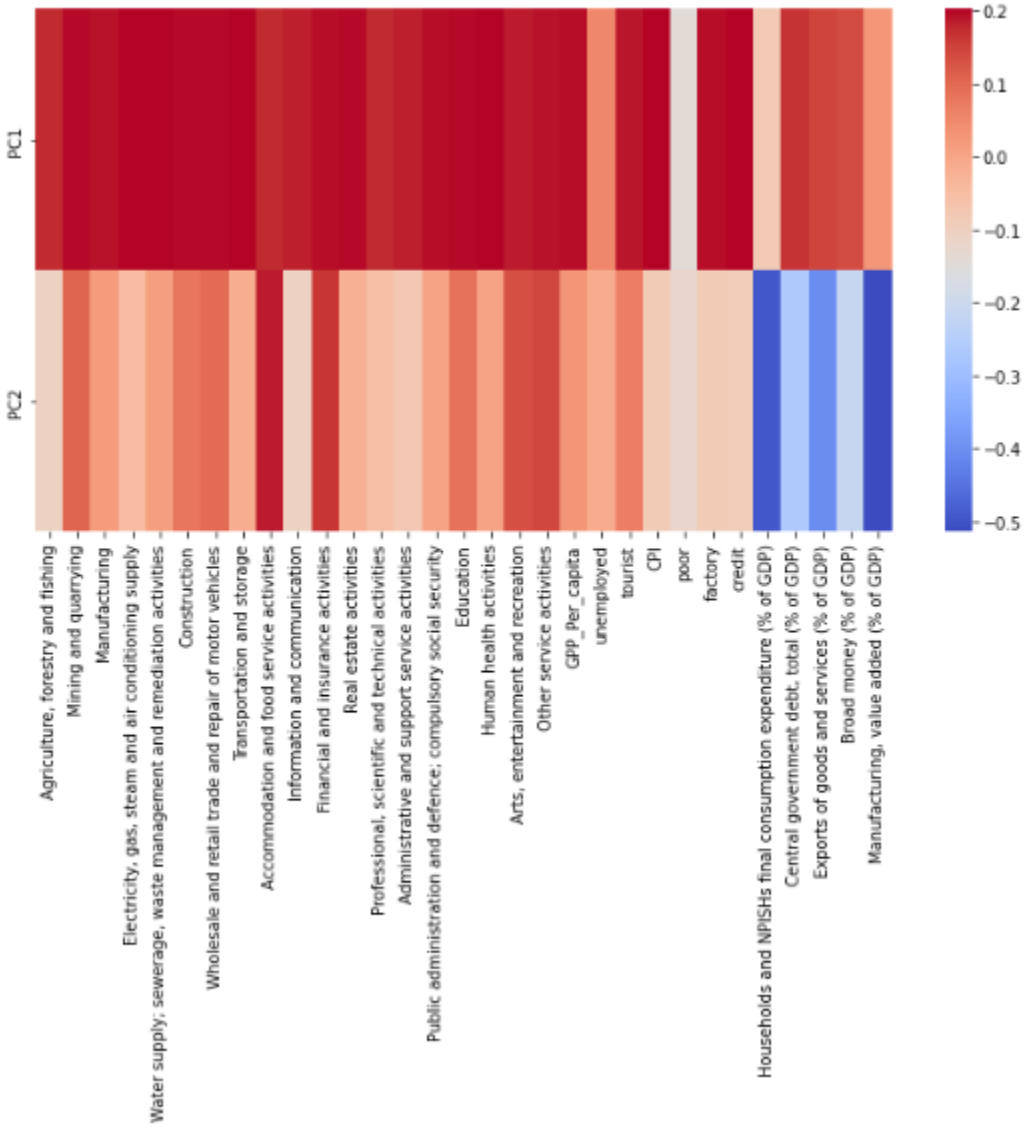
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2bc2da1410>
```



มี component 2 ตัว สาเหตุที่เลือก component 2 ตัว ใช้ค่า explained variance ratio ประกอบการพิจารณาเนื่องจาก Component แรก มีค่า explained variance ratio = 0.77944952 และ

Component ที่สอง มีค่า explained variance ratio = 0.07515276 ได้ผลรวมของ 2 Component = 0.85460228 แสดงว่าการเลือก 2 Component นี้สามารถอธิบายข้อมูลได้ประมาณ 85 เปอร์เซ็นต์

- Component ที่ได้ มีความสัมพันธ์กับ ปัจจัยเดิมอย่างใด



ในส่วนของ Component แรกจะเป็นได้ว่ามีค่า correlation สูงในแทบจะทุกปัจจัย แต่ในส่วนของ Component ที่สองจะมีค่า correlation ของปัจจัย Accommodation and food service activities สูงกว่าของ Component แรก จึงได้นำสอง Component นี้มาใช้เพื่อการอธิบายข้อมูลที่ครบถ้วนมากขึ้น

1.5 อภิปราย สำหรับข้อมูลของจังหวัดชลบุรี

- อภิปรายว่า จาก ตารางของปัจจัยที่ส่งผล จาก Feature selection ใน วิธีทำ สมเหตุสมผลหรือไม่ ปัจจัยใดน่าจะมีผลจริง ปัจจัยใดไม่น่าจะมีผล (ถึงแม้จะถูกเลือกมาด้วย feature selection ก็ตาม) หากมีผลจริง เรียงลำดับปัจจัยที่ส่งผลมากไปน้อย (อาจจะไม่ต้องเรียงลำดับปัจจัยที่ส่งผลทุกตัว แต่ควรบอกได้ว่า ปัจจัยลำดับต้นๆที่ส่งผลมากที่สุดมีอะไรบ้าง)

ใน Feature selection ส่วนใหญ่ให้ผลลัพธ์ที่สมเหตุสมผล คือมีปัจจัยที่มีสัดส่วนเยอะเป็นอันดับ1 คือ Manufacturing , อันดับที่ 2 คือ Wholesale and retail trade and repair of motor vehicles และ อันดับที่ 3 คือ Transportation and storage เกี่ยวข้องอยู่ในผลลัพธ์ของการทำ Feature selection ด้วย (อ้างอิงจากข้อ 1.2) และปัจจัยระดับจังหวัดที่มีแนวโน้มว่าจะส่งผลคือ 'Credit'(จำนวนเงินให้สินเชื่อ),'CPI' (ดัชนีราคาผู้บริโภคทั่วไป),factory'(จำนวนโรงงาน),'tourist'(จำนวนผู้เยี่ยมเยือน) เนื่องจากโผล่ในผลลัพธ์ของการทำ Feature selection 4 ครั้ง , 4 ครั้ง , 3 ครั้ง , 3 ครั้ง ตามลำดับ และปัจจัยระดับประเทศที่มีแนวโน้มว่าจะส่งผลคือ 'Central government debt, total (% of GDP)', 'Exports of goods and services (% of GDP)', เนื่องจากโผล่ในผลลัพธ์ของการทำ Feature selection 3 ครั้ง , 3 ครั้ง ตามลำดับ

- อภิปรายว่า จาก component ที่ได้จาก dimensionality reduction สมเหตุสมผลหรือไม่ มีความสัมพันธ์กับปัจจัยที่สรุปได้จาก feature selection ว่าส่งผล หรือไม่ สัมพันธ์อย่างไร

component ที่ได้จาก dimensionality reduction มีความสมเหตุสมผลเพราะเนื่องจาก GPP per capita เป็นค่าที่ได้จากการนำ 19 ปัจจัย บวกกันและหารจำนวนประชากร ดังนั้นจึงไม่แปลกที่ใน 1 component จะมีค่า correlation สูงในแทบจะทุกปัจจัย และในส่วนของ component ที่สองจะเป็นได้ว่าในส่วนของปัจจัยระดับประเทศ มีค่า correlation ต่ำ ซึ่งอาจบอกได้ว่าข้อมูลระดับประเทศไม่เจาะจงเท่ากับข้อมูลระดับจังหวัดดังนั้นจึงมีค่าความสัมพันธ์ที่ต่ำตามในรูปจากข้อ 1.4และในปัจจัยที่ feature selection เห็นว่าส่งผล ก็จะเห็นได้ว่าใน Component มีค่าความสัมพันธ์สูงอยู่ด้วย

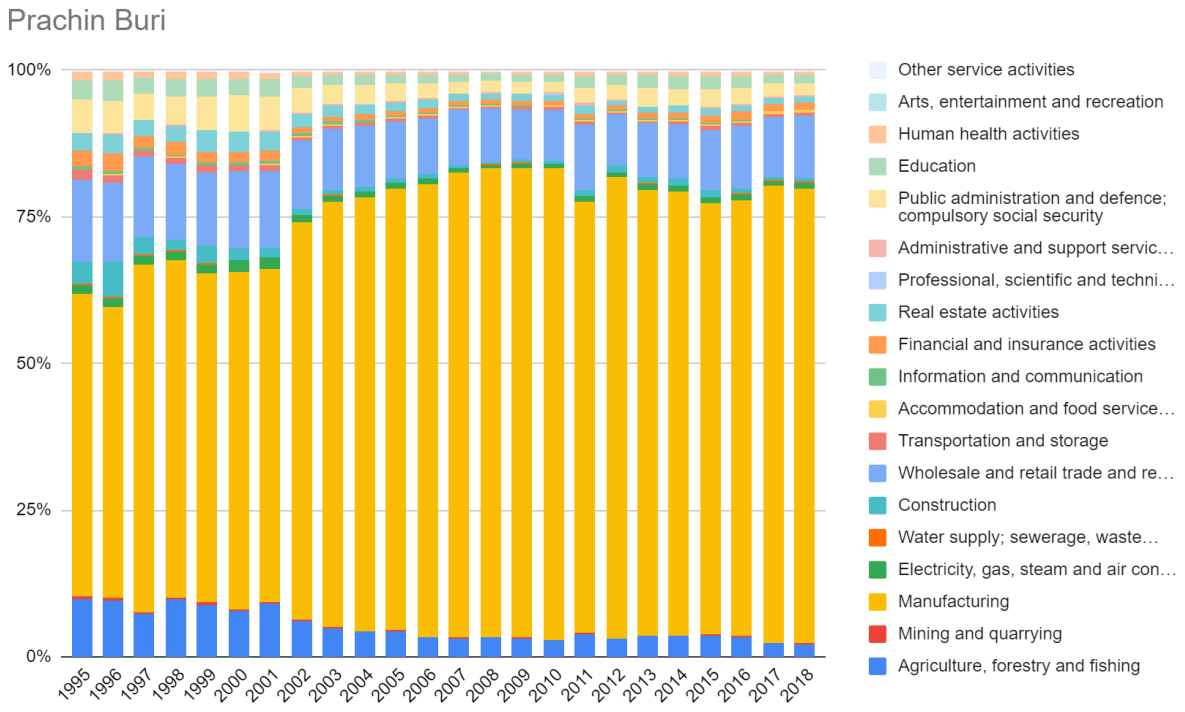
จังหวัด ปราจีนบุรี

1.1 ระบุปัจจัยที่เพิ่มขึ้นมา แหล่งข้อมูล และเหตุผลในการเลือกปัจจัยนั้น

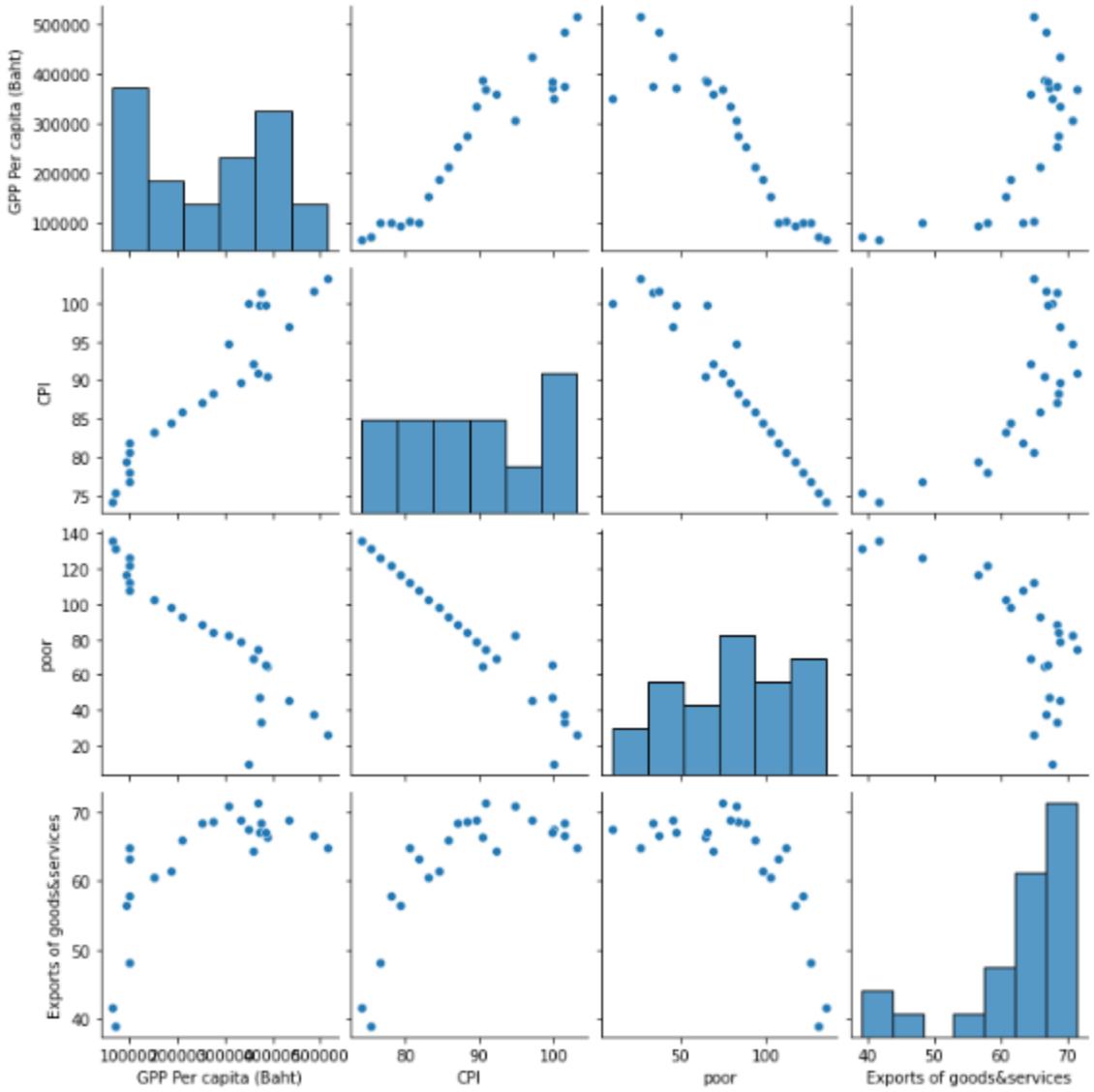
ปัจจัยที่เพิ่ม	ระดับจังหวัดหรือระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018) หรือเป็นข้อมูลที่ทำ regression	หากทำ regression ระบุอัตราส่วนระหว่างข้อมูลที่มีกับ ข้อมูลที่ทำ regression
จำนวนตำแหน่งงานว่าง (ความต้องการแรงงาน) ของแต่ล่ะจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (8:16)
จำนวนผู้ที่มาเยี่ยมเยือนในแต่ล่ะจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (9:15)
ดัชนีราคาผู้บริโภคทั่วไป(CPI)ที่จำแนกเป็นรายจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (9:15)
จำนวนคนจน (ด้าน	ระดับ	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี :

รายจ่าย) จำแนกเป็นรายจังหวัด	จังหวัด		ข้อมูลที่ทำ regression) = (8:16)
สถิติโรงงานอุตสาหกรรมที่จดทะเบียนกับกระทรวงอุตสาหกรรม และได้รับอนุญาตให้ประกอบกิจการใหม่ (ตามพระราชบัญญัติโรงงาน พ.ศ. 2535) จำแนกตามจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (8:16)
จำนวนเงินให้สินเชื่อคงเหลือของธนาคารออมสิน จำแนกตามจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (8:16)
Households and NPISHs final consumption expenditure (% of GDP)	ระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018)	-
Central government debt, total (% of GDP)	ระดับประเทศ	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (21:3)
Exports of goods and services (% of GDP)	ระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018)	-
Broad money (% of GDP)	ระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018)	-
Manufacturing, value added (% of GDP)	ระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018)	-

1.2 ทำ data exploration แสดงและบรรยายประกอบ
-visualization ของสัดส่วน 19 หมวดย่อย ที่มีต่อ GPP per capita
-visualization ของความสัมพันธ์ของ ปัจจัยที่เลือกมา และ GPP per capita เลือกทำ 3 ปัจจัย



ในvisualization ของสัดส่วน 19 หมวดย่อย ที่มีต่อ GPP per capitaของจังหวัดปราจีนบุรี (กำหนดให้ใน 1 กราฟแท่งคือค่าของแต่ล่ะปี)จะสังเกตุได้ว่าปัจจัยที่มีสัดส่วนมากที่สุดอันดับ 1 คือ Manufacturing (สีเหลือง) , อันดับที่ 2 คือ Wholesale and retail trade and repair of motor vehicles (สีฟ้า) และอันดับที่ 3 คือ Agriculture, forestry and fishing (สีน้ำเงิน)โดยในสัดส่วนในช่วงปี ค.ศ.2006-2018 จะมีค่า Manufacturing ที่เพิ่มมากขึ้นอย่างเห็นได้ชัด



ในส่วนของvisualization ของความสัมพันธ์ของ ปัจจัยที่เลือกมา และ GPP per capita กับ 3 ปัจจัย ได้ทำ pair plot ของข้อมูล 4 ตัวคือ Exports of goods and services (% of GDP) , CPI (ดัชนีราคาผู้บริโภคทั่วไป) , poor (จำนวนคนจน) และ GPP Per capita (Baht) จะเห็นได้ว่ามีคู่ GPP Per capita (Baht) และ CPI ที่ดูเหมือนจะมีความสัมพันธ์ไปในทางบวก และ คู่ GPP Per capita (Baht) และ poor ที่ดูเหมือนจะมีความสัมพันธ์ไปในทางลบ

1.3 สำหรับข้อมูลของจังหวัดปราจีนบุรี

เมื่อนำปัจจัยทุกตัวมาทำ regression แล้ว ได้ผลลัพธ์เป็นดังนี้

```
df = dummy_df.dropna()
X = df[input_vars]
y = df.GPP_Per_capita
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=99)
lm = LinearRegression()
lm.fit(X_train, y_train)
y_pred = lm.predict(X_test)
[np.sqrt(metrics.mean_squared_error(y_test,y_pred)),metrics.r2_score(y_test,y_pred)]
```

```
[23319.26753866343, 0.953487648877027]
```

ได้ Regression model ที่มีค่า RMSE(root mean square error) = 23319.26753866343 และมีค่า R-Square = 0.953487648877027 ซึ่งถือว่าเป็น model ที่ดีเนื่องจาก |r| = 0.9765 หมายความว่ามี ค่าความสัมพันธ์สูง

เมื่อนำปัจจัยไปทำ feature selection ด้วยวิธี

- Filter 2 วิธี : ใช้วิธี Low Variance Filtering และวิธี High Correlation Filtering
- Wrapper 2 วิธี : ใช้วิธี Forward Feature Selection และวิธี Recursive Feature Elimination
- Embedded 1 วิธี : ใช้วิธี Random Forest

จากนั้น สรุปต่อไปนี้
มีปัจจัยใดบ้างที่ส่งผลต่อ GPP per capita

		Regression	
วิธี	จำนวนปัจจัยที่ส่งผล	R-Squared	RMSE
Low Variance Filtering	8 ปัจจัย	0.9802648240008693	15189.772188098626
High Correlation Filtering	26 ปัจจัย	0.9855828819708079	12982.848389674482
Forward Feature Selection	26 ปัจจัย	0.9856297661959359	12961.72119716623
Recursive Feature Elimination	3 ปัจจัย	0.65669200621565	63353.80689810831
Random Forest	11 ปัจจัย	0.9896792267086372	10984.666542845212

- ปัจจัยที่ส่งผลจาก Feature selection นั้น มีกี่ปัจจัยที่เป็นระดับจังหวัด มีกี่ปัจจัยที่เป็นระดับประเทศ มีกี่ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) และมีกี่ปัจจัยที่เป็นปัจจัยได้จากการทำ regression/ interpolation

1.Low Variance Filtering ปัจจัยที่ส่งผลมีทั้งหมด 8 ปัจจัย ดังนี้

'Agriculture, forestry and fishing',
'Manufacturing',
'Wholesale and retail trade and repair of motor vehicles',
'Public administration and defence; compulsory social security',
'Education',
'unemployed',
'tourist',
'credit'

โดยมีปัจจัยระดับจังหวัด 3 ปัจจัย คือ 'unemployed'(จำนวนตำแหน่งงานว่าง) , 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'Credit'(จำนวนเงินให้สินเชื่อ) และไม่มีปัจจัยระดับประเทศที่ส่งผล ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 5 ปัจจัย และ ปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 3 ปัจจัยคือ 'unemployed' (จำนวนตำแหน่งงานว่าง) , 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'Credit'(จำนวนเงินให้สินเชื่อ)

2. High Correlation Filtering ปัจจัยที่ส่งผลมีทั้งหมด 26 ปัจจัย ดังนี้

'Agriculture, forestry and fishing',
'Mining and quarrying',
'Manufacturing',
'Electricity, gas, steam and air conditioning supply',
'Water supply; sewerage, waste management and remediation activities',
'Construction',
'Wholesale and retail trade and repair of motor vehicles',
'Transportation and storage',
'Accommodation and food service activities',
'Information and communication', 'Financial and insurance activities',
'Real estate activities',
'Professional, scientific and technical activities',
'Administrative and support service activities',
'Public administration and defence; compulsory social security',
'Education', 'Human health activities',
'Arts, entertainment and recreation', 'Other service activities',
'tourist',
'CPI',
'factory',
'credit',
'Central government debt, total (% of GDP)',

'Exports of goods and services (% of GDP)',
'Broad money (% of GDP)'

โดยมีปัจจัยระดับจังหวัด 4 ปัจจัย คือ 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'Credit'(จำนวนเงินให้สินเชื่อ) ,'CPI'(ดัชนีราคาผู้บริโภคทั่วไป),'factory'(จำนวนโรงงาน), และปัจจัยระดับประเทศ 3 ปัจจัย คือ 'Central government debt total (% of GDP)','Exports of goods and services (% of GDP)', 'Broad money (% of GDP)'ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 21 ปัจจัย และ ปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 5 ปัจจัยคือ 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'Credit'(จำนวนเงินให้สินเชื่อ),'CPI'(ดัชนีราคาผู้บริโภคทั่วไป),'factory'(จำนวนโรงงาน), 'Central government debt total (% of GDP)'

3.Forward Feature Selection **ปัจจัยที่ส่งผลมีทั้งหมด 26 ปัจจัย ดังนี้**

'Agriculture, forestry and fishing',
'Mining and quarrying',
'Manufacturing',
'Electricity, gas, steam and air conditioning supply',
'Water supply; sewerage, waste management and remediation activities',
'Construction',
'Wholesale and retail trade and repair of motor vehicles',
'Transportation and storage',
'Accommodation and food service activities',
'Information and communication',
'Financial and insurance activities',
'Real estate activities',
'Professional, scientific and technical activities',
'Administrative and support service activities',
'Public administration and defence; compulsory social security',
'Education',
'Human health activities',
'Arts, entertainment and recreation',
'Other service activities',
'tourist',
'CPI',
'poor',
'factory',
'credit',
'Central government debt, total (% of GDP)',
'Exports of goods and services (% of GDP)'

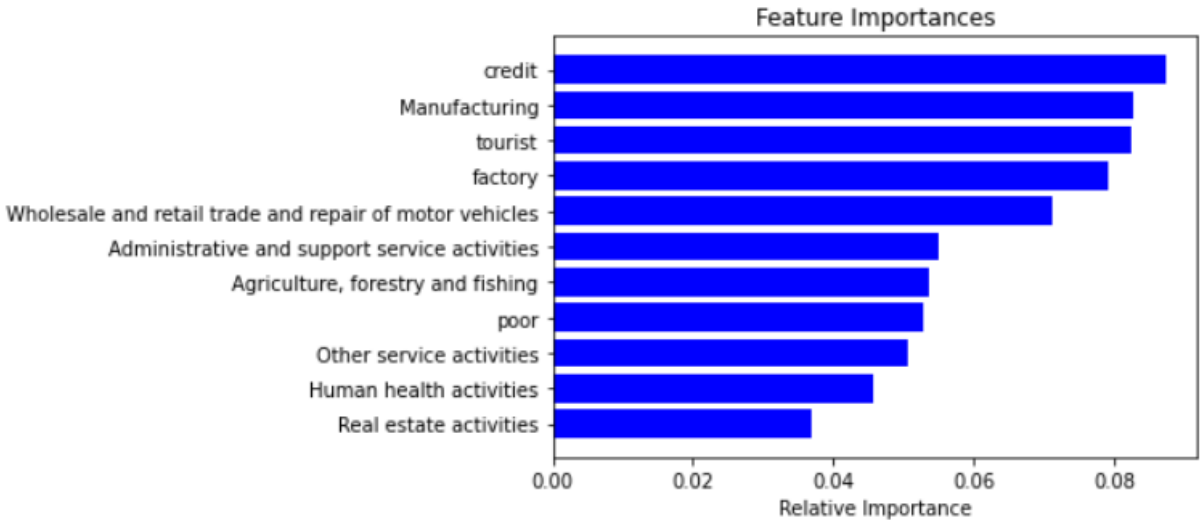
โดยมีปัจจัยระดับจังหวัด 5 ปัจจัย คือ 'poor'(จำนวนคนจน), 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'Credit' (จำนวนเงินให้สินเชื่อ) ,'CPI'(ดัชนีราคาผู้บริโภคทั่วไป),'factory'(จำนวนโรงงาน), และปัจจัยระดับประเทศ 2 ปัจจัย คือ 'Central government debt total (% of GDP)', 'Exports of goods and services (% of GDP)' ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 20 ปัจจัย และ ปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 6 ปัจจัยคือ 'poor'(จำนวนคนจน), 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'Credit'(จำนวนเงินให้สินเชื่อ),'CPI'(ดัชนีราคาผู้บริโภคทั่วไป),'factory'(จำนวนโรงงาน), 'Central government debt total (% of GDP)'

4.Recursive Feature Elimination **ปัจจัยที่ส่งผลมีทั้งหมด 3 ปัจจัย ดังนี้**

'CPI',
'Central government debt, total (% of GDP)',
'Exports of goods and services (% of GDP)'

โดยมีปัจจัยระดับจังหวัด 1 ปัจจัย คือ 'CPI'(ดัชนีราคาผู้บริโภคทั่วไป) และปัจจัยระดับประเทศ 2 ปัจจัย คือ 'Central government debt total (% of GDP)', 'Exports of goods and services (% of GDP)' ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 1 ปัจจัย และ ปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 2 ปัจจัยคือ 'CPI'(ดัชนีราคาผู้บริโภคทั่วไป), 'Central government debt total (% of GDP)'

5.Random Forest **ปัจจัยที่ส่งผลมีทั้งหมด 11 ปัจจัย ดังนี้**



โดยมีปัจจัยระดับจังหวัด 4 ปัจจัย คือ 'Credit'(จำนวนเงินให้สินเชื่อ) ,'tourist'(จำนวนผู้เยี่ยมเยือน) , 'factory'(จำนวนโรงงาน), 'poor'(จำนวนคนจน) และไม่มีปัจจัยระดับประเทศ ที่ส่งผล และ ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 7 ปัจจัย และ ปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 4 ปัจจัยคือ 'Credit'(จำนวนเงินให้สินเชื่อ) ,'tourist'(จำนวนผู้เยี่ยมเยือน) ,'factory'(จำนวนโรงงาน), 'poor'(จำนวนคนจน)

1.4 เมื่อนำปัจจัยไปทำ dimensionality reduction ด้วยวิธี PCA

- มี component กี่ตัว (หรือสามารถลด dimension ได้เหลือเท่าใด) เหตุใดจึงเลือก component เท่านี้ ใช้ค่าอะไรประกอบการพิจารณา

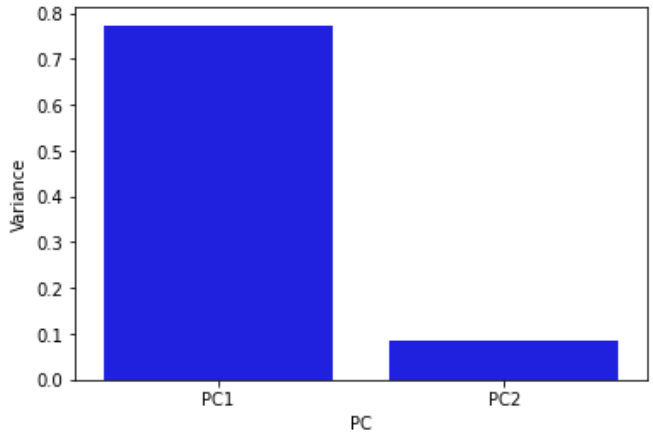
-

```
pca.explained_variance_ratio_
```

```
array([0.77353517, 0.08583812])
```

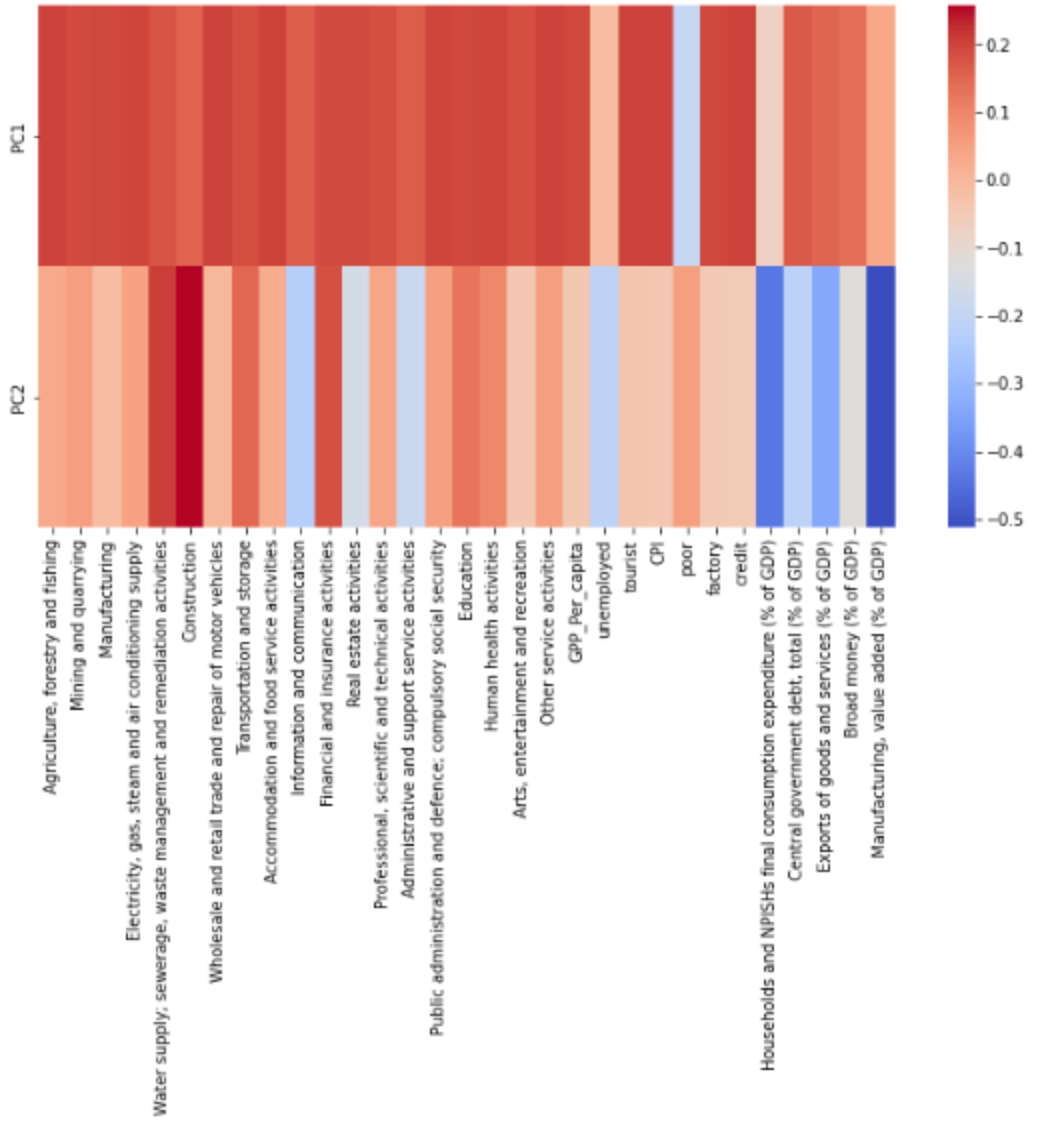
```
temp_df = pd.DataFrame({'Variance':pca.explained_variance_ratio_,
                        'PC':['PC1','PC2']})
sns.barplot(x='PC',y="Variance", data=temp_df, color="blue")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2bc217c1d0>
```



มี component 2 ตัว สาเหตุที่เลือก component 2 ตัว ใช้ค่า explained variance ratio ประกอบการพิจารณาเนื่องจาก Component แรก มีค่า explained variance ratio = 0.77353517 และ Component ที่สอง มีค่า explained variance ratio = 0.08583812 ได้ผลรวมของ 2 Component = 0.85937329 แสดงว่าการเลือก 2 Component นี้สามารถอธิบายข้อมูลได้ประมาณ 85 เปอร์เซ็นต์

- Component ที่ได้ มีความสัมพันธ์กับ ปัจจัยเดิมอย่างใด



ในส่วนของ Component แรกจะเป็นได้ว่ามีค่า correlation สูงในแทบจะทุกปัจจัย แต่ในส่วนของ Component ที่สองจะมีค่า correlation ของปัจจัย Construction , Water supply; sewerage, waste management and remediation activities , poor(จำนวนคนจน) สูงกว่าของ Component แรก จึงได้นำสอง Component นี้มาใช้เพื่อการอธิบายข้อมูลที่ครบถ้วนมากขึ้น

1.5 อภิปราย สำหรับข้อมูลของจังหวัดปราจีนบุรี

- อภิปรายว่า จาก ตารางของปัจจัยที่ส่งผล จาก Feature selection ใน วิธีทำ สมเหตุสมผลหรือไม่ ปัจจัยใดน่าจะมีผลจริง ปัจจัยใดไม่น่าจะมีผล (ถึงแม้จะถูกเลือกมาด้วย feature selection ก็ตาม) หากมีผลจริง เรียงลำดับปัจจัยที่ส่งผลมากไปน้อย (อาจจะไม่ต้องเรียงลำดับปัจจัยที่ส่งผลทุกตัว แต่ควรบอกได้ว่า ปัจจัยลำดับต้นๆที่ส่งผลมากที่สุดมีอะไรบ้าง)

ใน Feature selection ส่วนใหญ่ให้ผลลัพธ์ที่สมเหตุสมผล คือมีปัจจัยที่มีสัดส่วนเยอะเป็นอันดับ1 คือ Manufacturing , อันดับที่ 2 คือ Wholesale and retail trade and repair of motor vehicles และอันดับที่ 3 คือ Agriculture, forestry and fishing เกี่ยวข้องอยู่ในผลลัพธ์ของการทำ Feature selection ด้วย (อ้างอิงจากข้อ 1.2) และปัจจัยระดับจังหวัดที่มีแนวโน้มว่าจะส่งผลคือ 'Credit'(จำนวนเงินให้สินเชื่อ),

'CPI'(ดัชนีราคาผู้บริโภคทั่วไป),'tourist'(จำนวนผู้เยี่ยมเยือน) เนื่องจากโผล่ในผลลัพธ์ของการทำ Feature selection 4 ครั้ง , 3 ครั้ง , 4 ครั้ง ตามลำดับ และปัจจัยระดับประเทศที่มีแนวโน้มว่าจะส่งผลคือ 'Central government debt, total (% of GDP)', 'Exports of goods and services (% of GDP)', เนื่องจากโผล่ในผลลัพธ์ของการทำ Feature selection 3 ครั้ง , 3 ครั้ง ตามลำดับ

- อภิปรายว่า จาก component ที่ได้จาก dimensionality reduction สมเหตุสมผลหรือไม่ มีความสัมพันธ์กับปัจจัยที่สรุปได้จาก feature selection ว่าส่งผล หรือไม่ สัมพันธ์อย่างไร

component ที่ได้จาก dimensionality reduction มีความสมเหตุสมผลเพราะเนื่องจาก GPP per capita เป็นค่าที่ได้จากการนำ 19 ปัจจัย บวกกันและหารจำนวนประชากร ดังนั้นจึงไม่แปลกที่ใน 1 component จะมีค่า correlation สูงในแทบจะทุกปัจจัย และในส่วนของ component ที่สองจะเป็นได้ว่าในส่วนของปัจจัยระดับประเทศ มีค่า correlation ต่ำ ซึ่งอาจบอกได้ว่าข้อมูลระดับประเทศไม่เจาะจงเท่ากับข้อมูลระดับจังหวัดดังนั้นจึงมีค่าความสัมพันธ์ที่ต่ำตามในรูปจากข้อ 1.4 และในปัจจัยที่ feature selection เห็นว่าส่งผล ก็จะเห็นได้ว่าใน Component มีค่าความสัมพันธ์สูงอยู่ด้วย

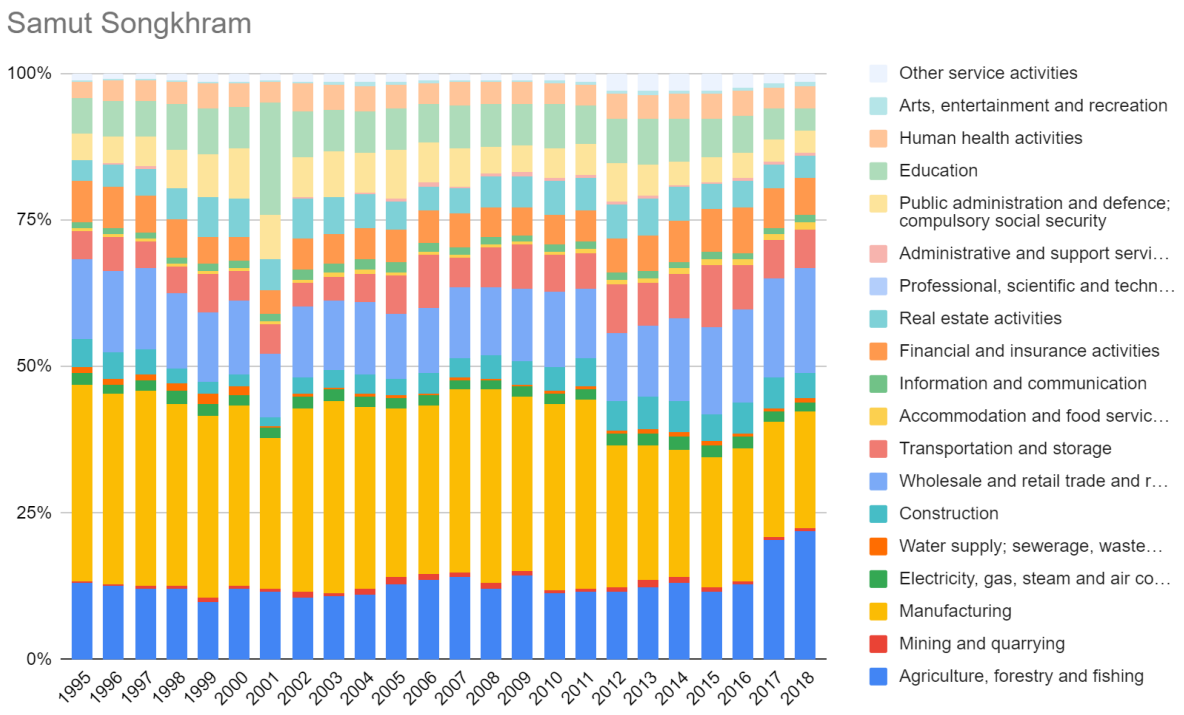
จังหวัด สมุทรสงคราม

1.1 ระบุปัจจัยที่เพิ่มขึ้นมา แหล่งข้อมูล และเหตุผลในการเลือกปัจจัยนั้น

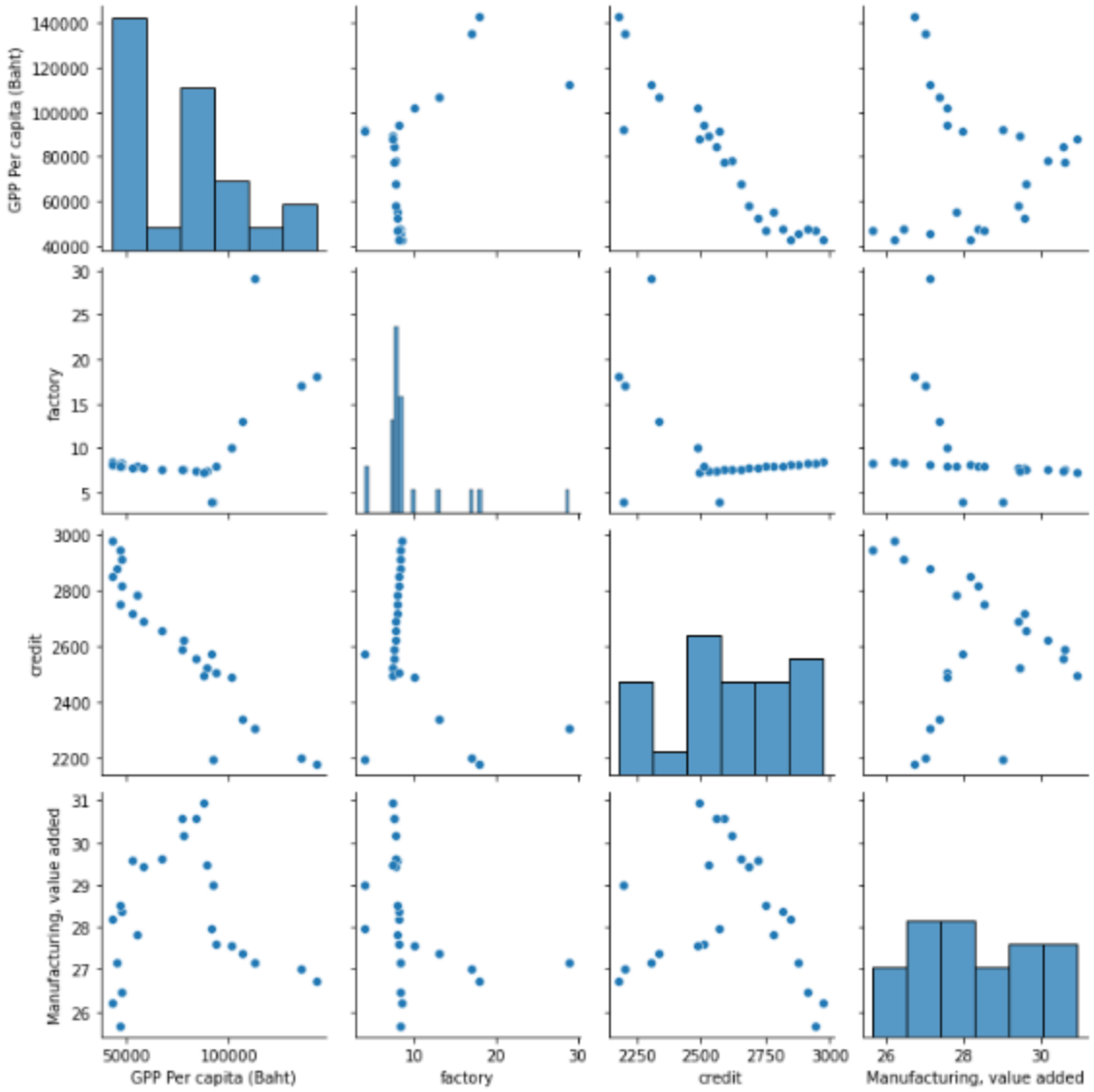
ปัจจัยที่เพิ่ม	ระดับจังหวัดหรือระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018) หรือเป็นข้อมูลที่ทำ regression	หากทำ regression ระบุอัตราส่วนระหว่างข้อมูลที่มีกับ ข้อมูลที่ทำ regression
จำนวนตำแหน่งงานว่าง (ความต้องการแรงงาน) ของแต่ล่ะจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (8:16)
จำนวนผู้ที่มาเยี่ยมเยือนในแต่ล่ะจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (9:15)
ดัชนีราคาผู้บริโภคทั่วไป(CPI)ที่จำแนกเป็นรายจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (9:15)
จำนวนคนจน (ด้านรายจ่าย) จำแนกเป็นรายจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (8:16)
สถิติโรงงานอุตสาหกรรมที่จดทะเบียนกับ	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (8:16)

กระทรวงอุตสาหกรรม และได้รับอนุญาตให้ประกอบกิจการใหม่ (ตามพระราชบัญญัติโรงงาน พ.ศ. 2535) จำแนกตามจังหวัด			
จำนวนเงินให้สินเชื่อคงเหลือของธนาคารออมสิน จำแนกตามจังหวัด	ระดับจังหวัด	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (8:16)
Households and NPISHs final consumption expenditure (% of GDP)	ระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018)	-
Central government debt, total (% of GDP)	ระดับประเทศ	เป็นข้อมูลที่ทำ regression	(อัตราส่วนระหว่างข้อมูลที่มี : ข้อมูลที่ทำ regression) = (21:3)
Exports of goods and services (% of GDP)	ระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018)	-
Broad money (% of GDP)	ระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018)	-
Manufacturing, value added (% of GDP)	ระดับประเทศ	มีข้อมูลครบทุกปี (1995-2018)	-

1.2 ทำ data exploration แสดงและบรรยายประกอบ
-visualization ของสัดส่วน 19 หมวดย่อย ที่มีต่อ GPP per capita
-visualization ของความสัมพันธ์ของ ปัจจัยที่เลือกมา และ GPP per capita เลือกทำ 3 ปัจจัย



ในvisualization ของสัดส่วน 19 หมวดย่อย ที่มีต่อ GPP per capitaของจังหวัดสมุทรสงคราม (กำหนดให้ใน 1 กราฟแท่งคือค่าของแต่ล่ะปี)จะสังเกตุได้ว่าปัจจัยที่มีสัดส่วนมากที่สุดอันดับ 1 คือ Manufacturing (สีเหลือง) , อันดับที่ 2 คือ Wholesale and retail trade and repair of motor vehicles (สีฟ้า) และอันดับที่ 3 คือ Agriculture, forestry and fishing (สีน้ำเงิน) โดยในสัดส่วนในช่วงปี ค.ศ. 2002-2011 จะมีค่าสัดส่วนที่ใกล้เคียงกัน และในช่วงปี ค.ศ.2017-2018 จะมี ค่า Agriculture, forestry and fishing เพิ่มขึ้นอย่างเห็นได้ชัด



ในส่วนของvisualization ของความสัมพันธ์ของ ปัจจัยที่เลือกมา และ GPP per capita กับ 3 ปัจจัย ได้ทำ pair plot ของข้อมูล 4 ตัวคือ Manufacturing, value added (% of GDP) , factory (จำนวนโรงงาน) , Credit (จำนวนเงินให้สินเชื่อ) และ GPP Per capita (Baht) จะเห็นได้ว่ามีเพียงคู่ GPP Per capita (Baht) และ Credit ที่ดูเหมือนจะมีความสัมพันธ์ไปในทางลบ

1.3 สำหรับข้อมูลของจังหวัดสมุทรสงคราม

เมื่อนำปัจจัยทุกตัวมาทำ regression แล้ว ได้ผลลัพธ์เป็นดังนี้

```
df = dummy_df.dropna()
X = df[input_vars]
y = df.GPP_Per_capita
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=99)
lm = LinearRegression()
lm.fit(X_train, y_train)
y_pred = lm.predict(X_test)
[np.sqrt(metrics.mean_squared_error(y_test,y_pred)),metrics.r2_score(y_test,y_pred)]
```

```
[2432.074922451968, 0.9846460447973302]
```

ได้ Regression model ที่มีค่า RMSE(root mean square error) = 2432.074922451968 และมีค่า R-Square = 0.9846460447973302 ซึ่งถือว่าเป็น model ที่ดีเนื่องจาก |r| = 0.9923 หมายความว่ามีค่าความสัมพันธ์สูง

เมื่อนำปัจจัยไปทำ feature selection ด้วยวิธี

- Filter 2 วิธี : ใช้วิธี Low Variance Filtering และวิธี High Correlation Filtering
- Wrapper 2 วิธี : ใช้วิธี Forward Feature Selection และวิธี Recursive Feature Elimination
- Embedded 1 วิธี : ใช้วิธี Random Forest

จากนั้น สรุปต่อไปนี้
มีปัจจัยใดบ้างที่ส่งผลต่อ GPP per capita

		Regression	
วิธี	จำนวนปัจจัยที่ส่งผล	R-Squared	RMSE
Low Variance Filtering	9 ปัจจัย	0.8975816176853071	6281.3901774446995
High Correlation Filtering	27 ปัจจัย	0.9833323308050322	2533.9862343352684
Forward Feature Selection	24 ปัจจัย	0.97621239669641	3027.2072773646305
Recursive Feature Elimination	9 ปัจจัย	0.9111585294705711	5850.253489823966
Random Forest	5 ปัจจัย	0.9412506793212084	4757.38435137459

- ปัจจัยที่ส่งผลจาก Feature selection นั้น มีกี่ปัจจัยที่เป็นระดับจังหวัด มีกี่ปัจจัยที่เป็นระดับประเทศ มีกี่ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) และมีกี่ปัจจัยที่เป็นปัจจัยได้จากการทำ regression/ interpolation

1.Low Variance Filtering ปัจจัยที่ส่งผลมีทั้งหมด 9 ปัจจัย ดังนี้

'Agriculture, forestry and fishing',
'Manufacturing',
'Construction',
'Wholesale and retail trade and repair of motor vehicles',
'Transportation and storage',
'Financial and insurance activities',
'Education',
'unemployed',
'tourist'

โดยมีปัจจัยระดับจังหวัด 2 ปัจจัย คือ 'unemployed'(จำนวนตำแหน่งงานว่าง) , 'tourist'(จำนวนผู้เยี่ยมเยือน) และไม่มีปัจจัยระดับประเทศที่ส่งผล ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 7 ปัจจัย และปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 2 ปัจจัยคือ 'unemployed'(จำนวนตำแหน่งงานว่าง) , 'tourist'(จำนวนผู้เยี่ยมเยือน)

2. High Correlation Filtering ปัจจัยที่ส่งผลมีทั้งหมด 27 ปัจจัย ดังนี้

'Agriculture, forestry and fishing',
'Mining and quarrying',
'Manufacturing',
'Electricity, gas, steam and air conditioning supply',
'Water supply; sewerage, waste management and remediation activities',
'Construction',
'Wholesale and retail trade and repair of motor vehicles',
'Transportation and storage',
'Accommodation and food service activities',
'Information and communication', 'Financial and insurance activities',
'Real estate activities',
'Professional, scientific and technical activities',
'Administrative and support service activities',
'Public administration and defence; compulsory social security',
'Education', 'Human health activities',
'Arts, entertainment and recreation', 'Other service activities',
'unemployed',
'tourist',
'CPI',
'factory',

'poor',
'Central government debt, total (% of GDP)',
'Exports of goods and services (% of GDP)',
'Broad money (% of GDP)'

โดยมีปัจจัยระดับจังหวัด 5 ปัจจัย คือ 'unemployed'(จำนวนตำแหน่งงานว่าง) , 'tourist'(จำนวนผู้เยี่ยมเยือน) ,'poor'(จำนวนคนจน) , 'CPI'(ดัชนีราคาผู้บริโภคทั่วไป),'factory'(จำนวนโรงงาน), และปัจจัยระดับประเทศ 3 ปัจจัย คือ 'Central government debt total (% of GDP)',
'Exports of goods and services (% of GDP)', 'Broad money (% of GDP)'ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 21 ปัจจัย และ ปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 6 ปัจจัยคือ 'unemployed' (จำนวนตำแหน่งงานว่าง) , 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'poor'(จำนวนคนจน), 'CPI'(ดัชนีราคาผู้บริโภคทั่วไป),'factory'(จำนวนโรงงาน), 'Central government debt total (% of GDP)'

3.Forward Feature Selection ปัจจัยที่ส่งผลมีทั้งหมด 24 ปัจจัย ดังนี้
'Agriculture, forestry and fishing',
'Mining and quarrying',
'Manufacturing',
'Electricity, gas, steam and air conditioning supply',
'Construction',
'Wholesale and retail trade and repair of motor vehicles',
'Transportation and storage',
'Accommodation and food service activities',
'Information and communication',
'Financial and insurance activities',
'Real estate activities',
'Professional, scientific and technical activities',
'Administrative and support service activities',
'Public administration and defence; compulsory social security',
'Human health activities',
'Arts, entertainment and recreation',
'Other service activities',
'tourist',
'CPI',
'credit',
'Households and NPISHs final consumption expenditure (% of GDP)',
'Central government debt, total (% of GDP)',
'Exports of goods and services (% of GDP)',
'Broad money (% of GDP)'

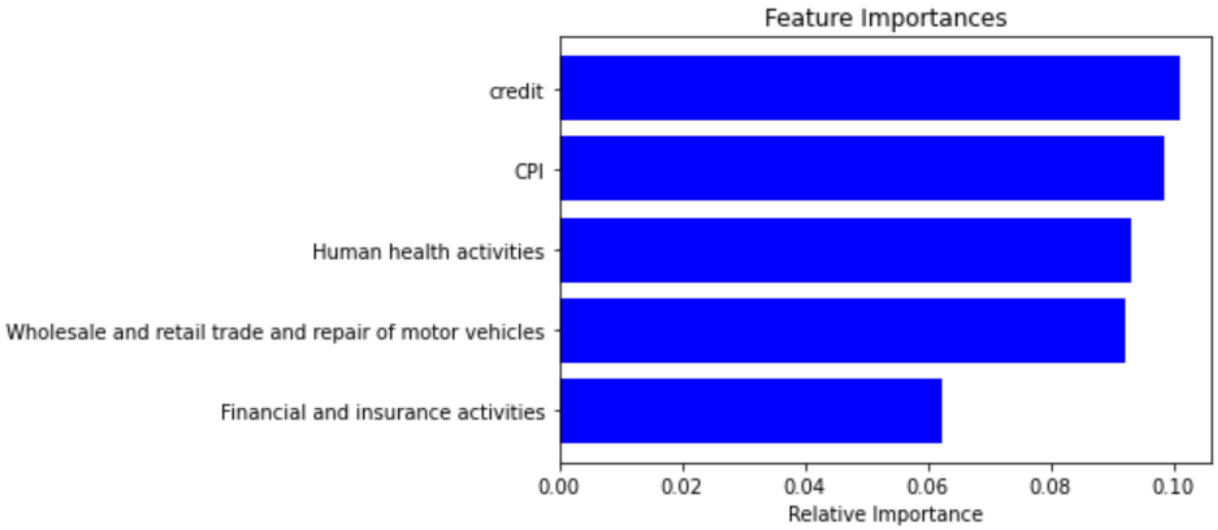
โดยมีปัจจัยระดับจังหวัด 3 ปัจจัย คือ 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'Credit'(จำนวนเงินให้สินเชื่อ) ,'CPI'(ดัชนีราคาผู้บริโภคทั่วไป) และปัจจัยระดับประเทศ 4 ปัจจัย คือ 'Households and NPISHs final consumption expenditure (% of GDP)','Central government debt total (% of GDP)', 'Exports of goods and services (% of GDP)' , 'Broad money (% of GDP)' ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 20 ปัจจัย และ ปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 4 ปัจจัยคือ 'tourist'(จำนวนผู้เยี่ยมเยือน) , 'Credit'(จำนวนเงินให้สินเชื่อ),'CPI'(ดัชนีราคาผู้บริโภคทั่วไป), 'Central government debt total (% of GDP)'

4.Recursive Feature Elimination ปัจจัยที่ส่งผลมีทั้งหมด 9 ปัจจัย ดังนี้

'Accommodation and food service activities',
'Administrative and support service activities',
'Other service activities',
'CPI',
'poor',
'factory',
'Households and NPISHs final consumption expenditure (% of GDP)',
'Exports of goods and services (% of GDP)',
'Broad money (% of GDP)'

โดยมีปัจจัยระดับจังหวัด 3 ปัจจัย คือ 'poor'(จำนวนคนจน) ,'CPI'(ดัชนีราคาผู้บริโภคทั่วไป), 'factory'(จำนวนโรงงาน), และปัจจัยระดับประเทศ 3 ปัจจัย คือ 'Households and NPISHs final consumption expenditure (% of GDP)' , 'Exports of goods and services (% of GDP)','Broad money (% of GDP)'ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 6 ปัจจัย และ ปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 3 ปัจจัยคือ 'poor'(จำนวนคนจน) , 'CPI'(ดัชนีราคาผู้บริโภคทั่วไป),'factory'(จำนวนโรงงาน)

5.Random Forest ปัจจัยที่ส่งผลมีทั้งหมด 5 ปัจจัย ดังนี้



โดยมีปัจจัยระดับจังหวัด 2 ปัจจัย คือ 'Credit'(จำนวนเงินให้สินเชื่อ) ,'CPI'(ดัชนีราคาผู้บริโภคทั่วไป) และไม่มีปัจจัยระดับประเทศ ที่ส่งผล และ ปัจจัยที่เป็นข้อมูลครบทุกปี (1995-2018) มี 3 ปัจจัย และปัจจัยที่เป็นปัจจัยได้จากการทำ regression มี 2 ปัจจัยคือ 'Credit'(จำนวนเงินให้สินเชื่อ) ,'CPI'(ดัชนีราคาผู้บริโภคทั่วไป)

1.4 เมื่อนำปัจจัยไปทำ dimensionality reduction ด้วยวิธี PCA

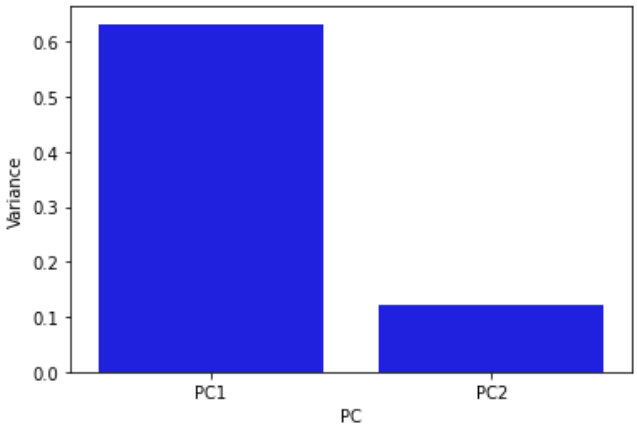
- มี component กี่ตัว (หรือสามารถลด dimension ได้เหลือเท่าใด) เหตุใดจึงเลือก component เท่านี้ ใช้ค่าอะไรประกอบการพิจารณา

```
pca.explained_variance_ratio_
```

```
array([0.63249386, 0.12308292])
```

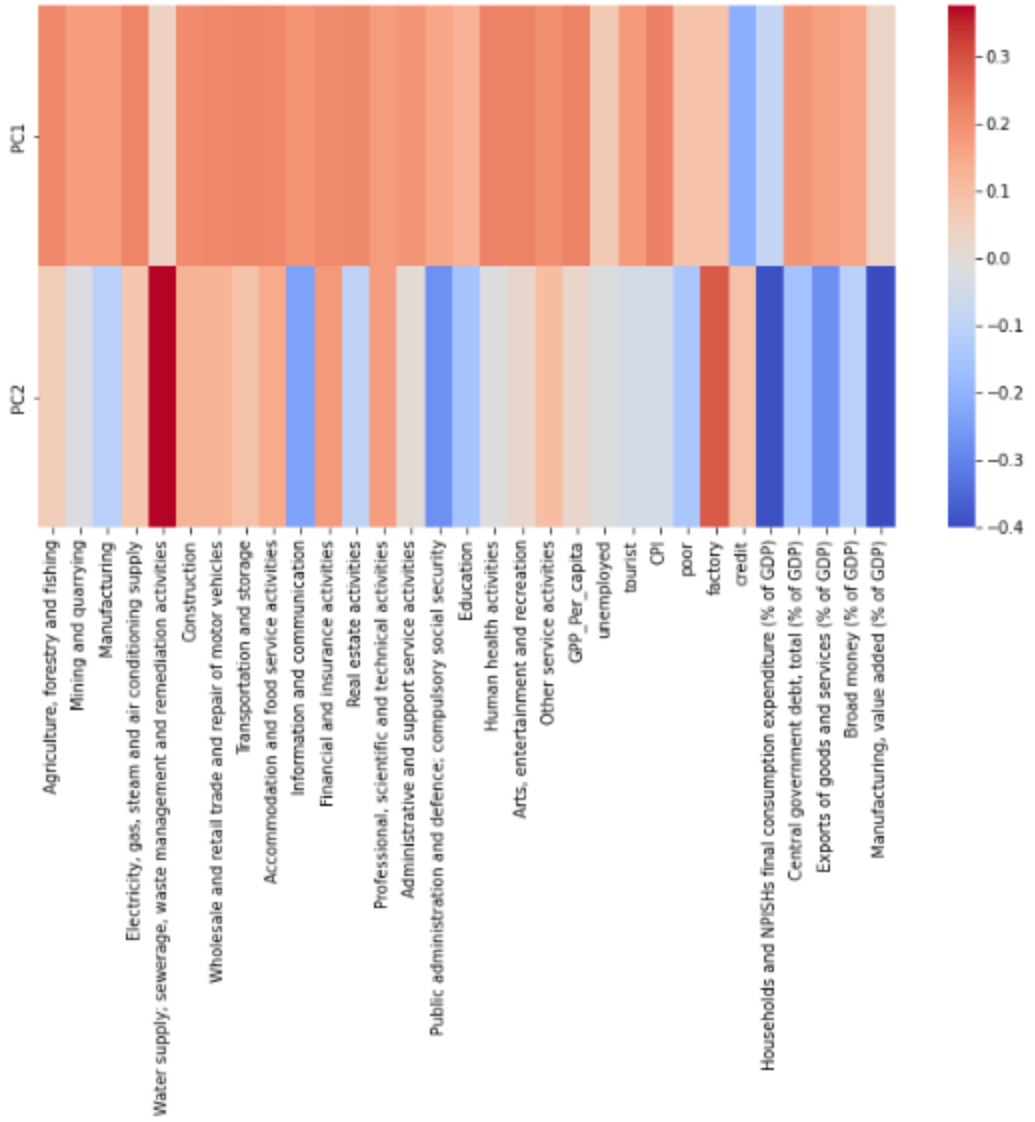
```
temp_df = pd.DataFrame({'Variance':pca.explained_variance_ratio_,
                        'PC':['PC1','PC2']})
sns.barplot(x='PC',y="Variance", data=temp_df, color="blue")
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f2bc23cf990>
```



มี component 2 ตัว สาเหตุที่เลือก component 2 ตัว ใช้ค่า explained variance ratio ประกอบการพิจารณาเนื่องจาก Component แรก มีค่า explained variance ratio = 0.63249386 และ Component ที่สอง มีค่า explained variance ratio = 0.12308292 ได้ผลรวมของ 2 Component = 0.75557678 แสดงว่าการเลือก 2 Component นี้สามารถอธิบายข้อมูลได้ประมาณ 75 เปอร์เซ็นต์

- Component ที่ได้ มีความสัมพันธ์กับ ปัจจัยเดิมอย่างใด



ในส่วนของ Component แรกจะเป็นได้ว่ามีค่า correlation สูงในแทบจะทุกปัจจัย แต่ในส่วนของ Component ที่สองจะมีค่า correlation ของปัจจัย Water supply; sewerage, waste management and remediation activities , factory (จำนวนโรงงาน) , Credit (จำนวนเงินให้สินเชื่อ) สูงกว่าของ Component แรก จึงได้นำสอง Component นี้มาใช้เพื่อการอธิบายข้อมูลที่ครบถ้วนมากขึ้น และ Component ของการทำ PCA ของข้อมูลจังหวัดสมุทรสงครามค่อนข้างเห็นการตั้งฉากของ Component ที่ 1 และ Component ที่ 2 ได้ชัดเจน

1.5 อภิปราย สำหรับข้อมูลของจังหวัดชลบุรี

- อภิปรายว่า จาก ตารางของปัจจัยที่ส่งผล จาก Feature selection ใน วิธีทำ สมเหตุสมผลหรือไม่ ปัจจัยใดน่าจะมีผลจริง ปัจจัยใดไม่น่าจะมีผล (ถึงแม้จะถูกเลือกมาด้วย feature selection ก็ตาม) หากมีผลจริง เรียงลำดับปัจจัยที่ส่งผลมากไปน้อย (อาจจะไม่ต้องเรียงลำดับปัจจัยที่ส่งผลทุกตัว แต่ควรบอกได้ว่า ปัจจัยลำดับต้นๆที่ส่งผลมากที่สุดมีอะไรบ้าง)

ใน Feature selection ส่วนใหญ่ให้ผลลัพธ์ที่สมเหตุสมผล คือมีปัจจัยที่มีสัดส่วนเยอะเป็นอันดับ1 คือ Manufacturing , อันดับที่ 2 คือ Wholesale and retail trade and repair of motor vehicles และ

อันดับที่ 3 คือ Agriculture, forestry and fishing เกี่ยวข้องอยู่ในผลลัพธ์ของการทำ Feature selection ด้วย (อ้างอิงจากข้อ 1.2) และปัจจัยระดับจังหวัดที่มีแนวโน้มว่าจะส่งผลคือ 'CPI'(ดัชนีราคาผู้บริโภคทั่วไป), 'tourist'(จำนวนผู้เยี่ยมเยือน) เนื่องจากโผล่ในผลลัพธ์ของการทำ Feature selection 4 ครั้ง , 3 ครั้ง ตามลำดับ และปัจจัยระดับประเทศที่มีแนวโน้มว่าจะส่งผลคือ 'Broad money (% of GDP)', 'Exports of goods and services (% of GDP)', เนื่องจากโผล่ในผลลัพธ์ของการทำ Feature selection 3 ครั้ง , 3 ครั้ง ตามลำดับ

- อภิปรายว่า จาก component ที่ได้จาก dimensionality reduction สมเหตุสมผลหรือไม่ มีความสัมพันธ์กับปัจจัยที่สรุปได้จาก feature selection ว่าส่งผล หรือไม่ สัมพันธ์อย่างไร

component ที่ได้จาก dimensionality reduction มีความสมเหตุสมผลเพราะเนื่องจาก GPP per capita เป็นค่าที่ได้จากการนำ 19 ปัจจัย บวกกันและหารจำนวนประชากร ดังนั้นจึงไม่แปลกที่ใน 1 component จะมีค่า correlation สูงในแทบจะทุกปัจจัย และในส่วนของ component ที่สองจะเป็นได้ว่าในส่วนของปัจจัยระดับประเทศ มีค่า correlation ต่ำ ซึ่งอาจบอกได้ว่าข้อมูลระดับประเทศไม่เจาะจงเท่ากับข้อมูลระดับจังหวัดดังนั้นจึงมีค่าความสัมพันธ์ที่ต่ำ โดยที่ Component ของการทำ PCA ของข้อมูลจังหวัดสมุทรสงครามค่อนข้างเห็นการตั้งฉากของ Component ที่ 1 และ Component ที่ 2 ได้ชัดเจน ตามในรูปจากข้อ 1.4และในปัจจัยที่ feature selection เห็นว่าส่งผล ก็จะเห็นได้ว่าใน Component มีค่าความสัมพันธ์สูงอยู่ด้วย

1.6 โดยรวม ปัจจัยที่ส่งผลต่อ GPP per capita ของ 3 จังหวัด มีความเหมือนหรือต่างกัน ให้เหตุผลว่าเหมือนหรือต่างกันเพราะเหตุใด

จังหวัดชลบุรี	จังหวัดปราจีนบุรี	จังหวัดสมุทรสงคราม
Manufacturing	Manufacturing	Manufacturing
Wholesale and retail trade and repair of motor vehicles	Wholesale and retail trade and repair of motor vehicles	Wholesale and retail trade and repair of motor vehicles
	Agriculture, forestry and fishing	Agriculture, forestry and fishing
Transportation and storage		
CPI(ดัชนีราคาผู้บริโภคทั่วไป)	CPI(ดัชนีราคาผู้บริโภคทั่วไป)	CPI(ดัชนีราคาผู้บริโภคทั่วไป)
tourist(จำนวนผู้เยี่ยมเยือน)	tourist(จำนวนผู้เยี่ยมเยือน)	tourist(จำนวนผู้เยี่ยมเยือน)
Credit(จำนวนเงินให้สินเชื่อ)	Credit(จำนวนเงินให้สินเชื่อ)	
factory(จำนวนโรงงาน)		
Exports of goods and services (% of GDP)	Exports of goods and services (% of GDP)	Exports of goods and services (% of GDP)
Central government debt, total (% of GDP)	Central government debt, total (% of GDP)	
		Broad money (% of GDP)

ปัจจัยที่ส่งผลต่อ GPP per capita ของ 3 จังหวัด มีความต่างกันบางปัจจัย อาจเพราะสภาพภูมิศาสตร์ที่ต่างกัน เช่น ชลบุรีเป็นจังหวัดที่ติดทะเล และยังอยู่ใกล้กับพื้นที่เศรษฐกิจ คือ พัทยา ดังนั้นพื้นที่ของชลบุรีจึงมีมูลค่าสูงกว่าอีกสองจังหวัด จึงไม่เหมาะกับการนำมาใช้พื้นที่ในการทำเกษตร เหมาะแก่การใช้ขนส่งเพราะเป็นพื้นที่ติดทะเล และ ใช้เป็นทางผ่านในการไปพัทยา เป็นต้น

Part 2: Clustering

<u>วัตถุประสงค์</u>

- วิเคราะห์การจัดกลุ่มของจังหวัดในประเทศไทย

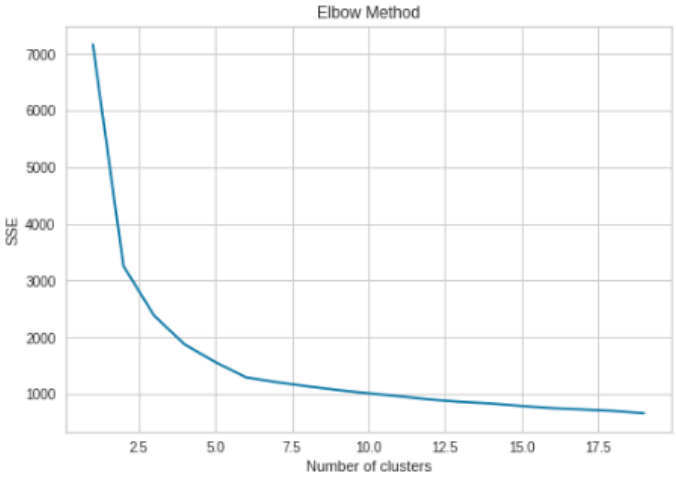
2.1 จัดกลุ่ม (Clustering) จังหวัด โดยใช้ทุก (19+11ปัจจัย และ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita))

จากการจัดข้อมูลได้ตัวอย่างของตารางข้อมูลมาดังนี้

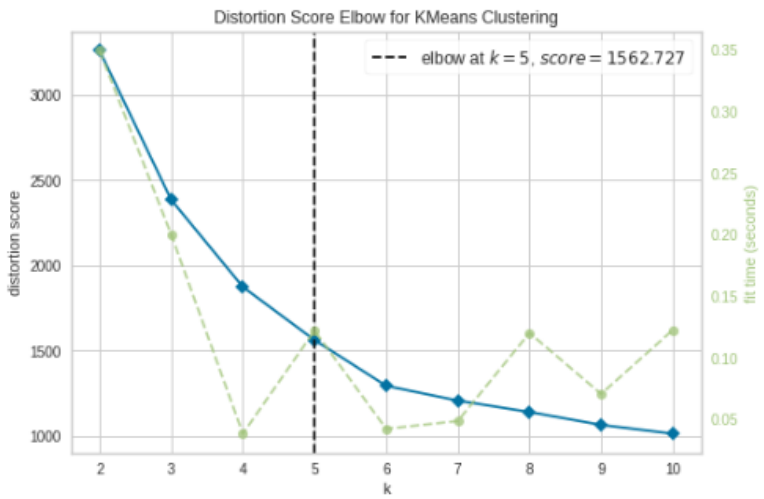
province	Year	Agriculture, forestry and fishing	Mining and quarrying	Manufacturing	Electricity, gas, steam and air conditioning supply	Water supply; sewerage, waste management and remediation activities	Construction	Wholesale and retail trade and repair of motor vehicles	Transportation and storage	Accommodation and food service activities	Information and communication	...	tourist	CPI	poor	factory	credit	Households and NPISHs final consumption expenditure (% of GDP)	Central government debt, total (% of GDP)	Exports of goods and services (% of GDP)	Broad money (% of GDP)	Manufacturing, value added (% of GDP)
Bangkok	2016	2479	0	601000	38798	21202	102327	1028711	362345	375719	261626	...	59196331	100.9	118.0	193	369297.90	50.009455	38.131933	67.070884	125.396343	27.144239
	2017	2414	0	594681	38225	22898	109938	1104473	383936	447780	275680	...	63575737	103.3	98.3	189	590251.58	48.930863	39.491027	66.672832	124.044740	27.014815
	2018	2601	0	622672	36361	25574	111787	1213150	398968	501608	290159	...	65534281	105.3	123.6	164	731024.54	48.902763	40.850120	64.856468	122.854209	26.718436
Samut Prakan	2016	2992	8	291877	15972	2537	10409	97576	163458	9021	4547	...	2900180	100.6	6.0	236	18788.41	50.009455	38.131933	67.070884	125.396343	27.144239
	2017	4238	7	287135	16028	2696	9631	108377	174780	10784	5933	...	3218934	101.0	0.8	315	20382.93	48.930863	39.491027	66.672832	124.044740	27.014815
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Amnat Charoen	2017	5085	11	983	220	46	892	2107	273	16	123	...	306856	102.1	39.0	13	5879.46	48.930863	39.491027	66.672832	124.044740	27.014815
	2018	5457	10	1083	198	50	927	2324	308	18	134	...	314595	103.2	64.8	12	5659.87	48.902763	40.850120	64.856468	122.854209	26.718436
Bungkan	2016	8772	320	4201	265	26	600	2384	378	53	99	...	566753	102.2	26.6	22	10115.97	50.009455	38.131933	67.070884	125.396343	27.144239
	2017	10467	289	5427	283	38	897	2854	406	57	118	...	581664	101.2	19.5	9	9680.27	48.930863	39.491027	66.672832	124.044740	27.014815
	2018	8304	449	4466	273	39	931	2263	423	64	130	...	602972	102.2	41.4	37	9387.88	48.902763	40.850120	64.856468	122.854209	26.718436

231 rows × 33 columns

Clustering ด้วยวิธี K-mean โดยนำข้อมูลที่ผ่านการทำ normalized มาทำ Elbow Method โดยได้ทำการรัน parameter ค่า k ตั้งแต่ 1 ถึง 20 และวัดระยะทาง intra-cluster จะได้ผลลัพธ์ออกมาดังนี้



ทางผู้จัดทำได้เลือก parameter ค่า k = 5 เนื่องจาก มีค่า Sum of intra-cluster measures ที่ไม่มากหรือน้อยจนเกินไป บ่งบอกถึงมีการแบ่ง cluster ได้อย่างเหมาะสม



โดยทางผู้จัดทำยังได้ทำการใช้ yellowbrick framework เพื่อช่วยยืนยัน ค่า parameter k=5 ว่าเหมาะสมที่จะนำไปใช้ในการแบ่งจำนวน cluster ด้วยวิธี k-mean

2.2 อภิปราย ผลลัพธ์ที่เกิดขึ้น

- เปรียบเทียบ กลุ่มที่เกิดขึ้น มีกี่กลุ่ม แต่ละกลุ่ม รวมจังหวัดที่มีลักษณะเหมือนกันอย่างไร
- จังหวัดที่ได้รับมอบหมาย อยู่กลุ่มใด และสมควรถูกจัดในกลุ่มนั้นหรือไม่
- จังหวัดในภาคที่ได้รับมอบหมายกระจายไปอยู่กลุ่มที่มีลักษณะแบบใดบ้าง มีกี่จังหวัดในกลุ่มนั้น จากการกระจายตัวของจังหวัดไปกลุ่มต่างๆนั้น สามารถสรุปลักษณะของภาคที่ได้รับมอบหมายได้อย่างไร

ผลลัพธ์ของวิธี K-mean ได้ทำการแบ่ง cluster ออกเป็น 5 กลุ่ม ดังนี้

Cluster_id	จำนวนข้อมูลที่อยู่ใน cluster นั้นๆ
0	4
1	1
2	24
3	24
4	24

กลุ่มที่ 1

	province	Year	Agriculture, forestry and fishing	Mining and quarrying	Manufacturing	Electricity, gas, steam and air conditioning supply	Water supply; sewerage, waste management and remediation activities	Construction	Wholesale and retail trade and repair of motor vehicles	Transportation and storage	...	CPI	poor	factory	credit	Households and NPISHs final consumption expenditure (% of GDP)	Central government debt, total (% of GDP)	Exports of goods and services (% of GDP)	Broad money (% of GDP)	Manufacturing, value added (% of GDP)	cluster_id
6	Pathum Thani	2016	5789	133	185025	9546	2427	11901	53144	8398	...	100.5	1.6	270	21947.55	50.009455	38.131933	67.070884	125.396343	27.144239	0
12	Nakhon Pathom	2016	21313	2119	167758	4636	1294	10547	37786	6780	...	99.8	16.6	107	17947.76	50.009455	38.131933	67.070884	125.396343	27.144239	0
15	Nonthaburi	2016	5535	53	47301	5548	1995	15053	52403	9908	...	101.1	10.6	91	22605.25	50.009455	38.131933	67.070884	125.396343	27.144239	0
18	Saraburi	2016	12164	17027	113403	30463	622	4261	21233	12705	...	100.6	25.5	58	15633.18	50.009455	38.131933	67.070884	125.396343	27.144239	0
21	Sing Buri	2016	2456	50	7880	759	108	784	2937	678	...	100.6	20.3	11	5309.82	50.009455	38.131933	67.070884	125.396343	27.144239	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
216	Maha Sarakham	2016	12059	56	6690	754	223	2900	5711	1355	...	100.0	121.6	20	25239.94	50.009455	38.131933	67.070884	125.396343	27.144239	0
219	Si Sa Ket	2016	19113	184	4238	818	127	2729	7983	977	...	100.6	112.0	37	32860.34	50.009455	38.131933	67.070884	125.396343	27.144239	0
222	Nong Bua Lam Phu	2016	6014	335	5409	291	71	1157	3422	609	...	101.7	47.8	9	10904.21	50.009455	38.131933	67.070884	125.396343	27.144239	0
225	Amnat Charoen	2016	4758	13	961	213	39	948	1919	228	...	100.8	66.6	14	6037.43	50.009455	38.131933	67.070884	125.396343	27.144239	0
228	Bungkan	2016	8772	320	4201	265	26	600	2384	378	...	102.2	26.6	22	10115.97	50.009455	38.131933	67.070884	125.396343	27.144239	0

72 rows × 36 columns

กลุ่มที่ 2

	province	Year	Agriculture, forestry and fishing	Mining and quarrying	Manufacturing	Electricity, gas, steam and air conditioning supply	Water supply; sewerage, waste management and remediation activities	Construction	Wholesale and retail trade and repair of motor vehicles	Transportation and storage	...	CPI	poor	factory	credit	Households and NPISHs final consumption expenditure (% of GDP)	Central government debt, total (% of GDP)	Exports of goods and services (% of GDP)	Broad money (% of GDP)	Manufacturing, value added (% of GDP)	cluster_id
0	Bangkok	2016	2479	0	601000	38798	21202	102327	1028711	362345	...	100.9	118.0	193	369297.90	50.009455	38.131933	67.070884	125.396343	27.144239	1
1	Bangkok	2017	2414	0	594681	38225	22898	109938	1104473	383936	...	103.3	98.3	189	590251.58	48.930863	39.491027	66.672832	124.044740	27.014815	1
2	Bangkok	2018	2601	0	622672	36361	25574	111787	1213150	398968	...	105.3	123.6	164	731024.54	48.902763	40.850120	64.856468	122.854209	26.718436	1

3 rows × 36 columns

กลุ่มที่ 3

	province	Year	Agriculture, forestry and fishing	Mining and quarrying	Manufacturing	Electricity, gas, steam and air conditioning supply	Water supply; sewerage, waste management and remediation activities	Construction	Wholesale and retail trade and repair of motor vehicles	Transportation and storage	...	CPI	poor	factory	credit	Households and NPISHs final consumption expenditure (% of GDP)	Central government debt, total (% of GDP)	Exports of goods and services (% of GDP)	Broad money (% of GDP)	Manufacturing, value added (% of GDP)	cluster_id
7	Pathum Thani	2017	6207	112	202558	11415	2754	11801	65417	9288	...	101.3	0.0	202	22721.03	48.930863	39.491027	66.672832	124.04474	27.014815	2
13	Nakhon Pathom	2017	20317	1968	175766	4811	1496	9598	40659	7451	...	101.1	17.3	138	17419.12	48.930863	39.491027	66.672832	124.04474	27.014815	2
16	Nonthaburi	2017	5430	53	53499	5454	2260	14256	58351	10549	...	102.0	11.5	79	22930.64	48.930863	39.491027	66.672832	124.04474	27.014815	2
19	Saraburi	2017	12601	15331	112466	29052	575	2929	21983	12629	...	100.4	11.2	79	15410.01	48.930863	39.491027	66.672832	124.04474	27.014815	2
22	Sing Buri	2017	3402	45	7512	764	138	786	3334	754	...	102.3	20.6	6	5221.91	48.930863	39.491027	66.672832	124.04474	27.014815	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
217	Maha Sarakham	2017	12065	47	7030	780	233	2960	6100	1627	...	103.0	92.9	18	24334.35	48.930863	39.491027	66.672832	124.04474	27.014815	2
220	Si Sa Ket	2017	19830	197	4356	852	203	2872	8492	1128	...	100.1	86.4	27	32053.82	48.930863	39.491027	66.672832	124.04474	27.014815	2
223	Nong Bua Lam Phu	2017	6602	360	5182	304	102	1045	3943	710	...	103.9	55.8	8	10563.79	48.930863	39.491027	66.672832	124.04474	27.014815	2
226	Amnat Charoen	2017	5085	11	983	220	46	892	2107	273	...	102.1	39.0	13	5879.46	48.930863	39.491027	66.672832	124.04474	27.014815	2
229	Bungkan	2017	10467	289	5427	283	38	897	2854	406	...	101.2	19.5	9	9680.27	48.930863	39.491027	66.672832	124.04474	27.014815	2

72 rows × 36 columns

กลุ่มที่ 4

	province	Year	Agriculture, forestry and fishing	Mining and quarrying	Manufacturing	Electricity, gas, steam and air conditioning supply	Water supply; sewerage, waste management and remediation activities	Construction	Wholesale and retail trade and repair of motor vehicles	Transportation and storage	...	CPI	poor	factory	credit	Households and NPISHs final consumption expenditure (% of GDP)	Central government debt, total (% of GDP)	Exports of goods and services (% of GDP)	Broad money (% of GDP)	Manufacturing, value added (% of GDP)	cluster_id
8	Pathum Thani	2018	6689	118	203401	9683	2592	11482	72533	10045	...	102.7	23.6	197	22703.60	48.902763	40.85012	64.856468	122.854209	26.718436	3
14	Nakhon Pathom	2018	21188	2138	190441	4383	1548	10368	45203	7890	...	102.0	24.6	175	17153.25	48.902763	40.85012	64.856468	122.854209	26.718436	3
17	Nonthaburi	2018	5861	54	50295	5473	2367	14167	62328	11171	...	103.2	11.8	92	23354.51	48.902763	40.85012	64.856468	122.854209	26.718436	3
20	Saraburi	2018	13484	13698	121481	29074	559	2566	24668	11053	...	101.4	50.5	51	15205.95	48.902763	40.85012	64.856468	122.854209	26.718436	3
23	Sing Buri	2018	4314	46	7184	625	200	816	3727	773	...	103.0	38.6	5	5193.54	48.902763	40.85012	64.856468	122.854209	26.718436	3
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
218	Maha Sarakham	2018	13000	67	8021	704	261	3343	6580	1740	...	102.4	98.4	20	23362.16	48.902763	40.85012	64.856468	122.854209	26.718436	3
221	Si Sa Ket	2018	20472	209	4596	771	175	3053	9190	1296	...	102.7	225.5	26	31353.32	48.902763	40.85012	64.856468	122.854209	26.718436	3
224	Nong Bua Lam Phu	2018	6912	360	5162	275	100	1326	4148	720	...	104.5	88.7	13	10098.10	48.902763	40.85012	64.856468	122.854209	26.718436	3
227	Amnat Charoen	2018	5457	10	1083	198	50	927	2324	308	...	103.2	64.8	12	5659.87	48.902763	40.85012	64.856468	122.854209	26.718436	3
230	Bungkan	2018	8304	449	4466	273	39	931	2263	423	...	102.2	41.4	37	9387.88	48.902763	40.85012	64.856468	122.854209	26.718436	3

72 rows × 36 columns

กลุ่มที่ 5

	province	Year	Agriculture, forestry and fishing	Mining and quarrying	Manufacturing	Electricity, gas, steam and air conditioning supply	Water supply; sewerage, waste management and remediation activities	Construction	Wholesale and retail trade and repair of motor vehicles	Transportation and storage	...	CPI	poor	factory	credit	Households and NPISHs final consumption expenditure (% of GDP)	Central government debt, total (% of GDP)	Exports of goods and services (% of GDP)	Broad money (% of GDP)	Manufacturing, value added (% of GDP)	cluster_id
3	Samut Prakan	2016	2992	8	291877	15972	2537	10409	97576	163458	...	100.6	6.0	236	18788.41	50.009455	38.131933	67.070884	125.396343	27.144239	4
4	Samut Prakan	2017	4238	7	287135	16028	2696	9631	108377	174780	...	101.0	0.8	315	20382.93	48.930863	39.491027	66.672832	124.044740	27.014815	4
5	Samut Prakan	2018	4386	7	331085	16133	2718	11138	124765	171883	...	102.2	7.0	296	20983.56	48.902763	40.850120	64.856468	122.854209	26.718436	4
9	Samut Sakhon	2016	14104	764	236941	8386	2703	7267	47602	5137	...	100.2	19.2	408	6844.73	50.009455	38.131933	67.070884	125.396343	27.144239	4
10	Samut Sakhon	2017	7585	847	268143	8597	2702	6979	53271	5473	...	101.3	4.9	451	6688.63	48.930863	39.491027	66.672832	124.044740	27.014815	4
11	Samut Sakhon	2018	7718	557	280167	7827	2374	7559	59175	5701	...	102.1	52.0	450	6653.83	48.902763	40.850120	64.856468	122.854209	26.718436	4
54	Chon Buri	2016	17512	3997	460858	56565	4444	22782	90290	57498	...	101.0	0.5	234	36124.57	50.009455	38.131933	67.070884	125.396343	27.144239	4
55	Chon Buri	2017	19580	3365	484636	54055	4460	17813	101209	60028	...	102.5	35.5	243	37378.56	48.930863	39.491027	66.672832	124.044740	27.014815	4
56	Chon Buri	2018	19730	3318	522108	66966	4724	17555	111368	65041	...	104.1	5.7	250	38681.35	48.902763	40.850120	64.856468	122.854209	26.718436	4
60	Rayong	2016	18392	248718	413495	56536	2366	7614	66064	17843	...	99.7	19.3	142	18455.00	50.009455	38.131933	67.070884	125.396343	27.144239	4
61	Rayong	2017	22342	255652	464253	73696	2482	6910	75628	19075	...	100.1	16.9	121	17784.43	48.930863	39.491027	66.672832	124.044740	27.014815	4
62	Rayong	2018	18967	281800	473613	83220	2519	7555	80646	22765	...	101.6	39.7	139	16857.54	48.902763	40.850120	64.856468	122.854209	26.718436	4

12 rows × 36 columns

โดยค่าเฉลี่ยของปัจจัยในแต่ละ Cluster มีดังนี้

cluster_id	Agriculture, forestry and fishing	Mining and quarrying	Manufacturing	Electricity, gas, steam and air conditioning supply	Water supply; sewerage, waste management and remediation activities	Construction	Wholesale and retail trade and repair of motor vehicles	Transportation and storage	Accommodation and food service activities	Information and communication	...	tourist	CPI	poor	factory	credit	Households and NPISHs final consumption expenditure (% of GDP)	Central government debt, total (% of GDP)	Exports of goods and services (% of GDP)	Broad money (% of GDP)	Manufacturing, value added (% of GDP)
0	13128.833333	66598.666667	378209.250000	38865.083333	3080.416667	11101.000000	8.466425e+04	84056.833333	21026.166667	3432.750000	...	7.323581e+06	101.366667	17.291667	273.750000	20468.628333	49.281027	39.491027	66.200062	124.098431	26.959164
1	2498.000000	0.000000	606117.666667	37794.666667	23224.666667	108017.333333	1.115445e+06	381749.666667	441702.333333	275821.666667	...	6.276878e+07	103.166667	113.300000	182.000000	563524.673333	49.281027	39.491027	66.200062	124.098431	26.959164
2	16410.083333	2031.569444	27176.236111	2734.569444	352.458333	3472.513889	1.190597e+04	3587.777778	3468.583333	887.805556	...	2.518387e+06	100.633333	78.427778	42.222222	15903.230278	50.009455	38.131933	67.070884	125.396343	27.144239
3	17730.625000	1823.777778	29806.208333	2939.180556	413.541667	3544.569444	1.397689e+04	4031.597222	4477.583333	1122.583333	...	2.911409e+06	102.233333	89.644444	39.111111	15399.018056	48.902763	40.850120	64.856468	122.854209	26.718436
4	17289.402778	1753.966333	28996.819444	2861.125000	402.805556	3381.277778	1.360293e+04	3925.388889	3954.402778	1010.902778	...	2.772827e+06	101.405556	71.779167	38.305556	15617.326389	48.930863	39.491027	66.672832	124.044740	27.014815

5 rows × 31 columns

เนื่องจากค่าที่ได้ในแต่ละกลุ่มไม่สามารถเจาะจงอธิบายรายละเอียดได้อย่างชัดเจนจึงอธิบายผ่านข้อมูลปัจจัยโดยเฉลี่ยของแต่ละกลุ่มแทน โดยแต่ละกลุ่มจะมีลักษณะคร่าวๆดังนี้

กลุ่มที่ 1

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Agriculture, forestry and fishing	16410.0833333333	สูง
Mining and quarrying	2031.56944444444	ต่ำ
Manufacturing	27176.2361111111	กลาง
Electricity, gas, steam and air conditioning supply	2734.569	กลาง
Water supply; sewerage, waste management and remediation activities	352.4583	ต่ำ
Construction	3472.514	ต่ำ
Wholesale and retail trade	11905.97	ต่ำ

and repair of motor vehicles		
Transportation and storage	3587.778	ต่ำ
Accommodation and food service activities	3458.583	ต่ำ
Information and communication'	887.8056	ต่ำ
Financial and insurance activities	5230.25	ต่ำ
Real estate activities	3043.319	ต่ำ
Professional, scientific and technical activities	336.125	ต่ำ
Administrative and support service activities	742.8333	ต่ำ
Public administration and defence; compulsory social security	4456.653	ต่ำ
Education	7128.75	กลาง
Human health activities	2786.944	ต่ำ
Arts, entertainment and recreation	336.9583	ต่ำ
Other service activities	846.7639	ต่ำ
GPP Per capita (Baht)	127042.1	ต่ำ

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
unemployed(จำนวนตำแหน่งงานว่าง)	3474.889	ต่ำ
tourist(จำนวนผู้เยี่ยมเยือน)	2518387	ต่ำ
CPI(ดัชนีราคาผู้บริโภคทั่วไป)	100.6333	กลาง
poor(จำนวนคนจน)	78.42778	กลาง
factory(จำนวนโรงงาน)	42.22222	ต่ำ

Credit(จำนวนเงินให้สินเชื่อ)	15903.23	กลาง

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Households and NPISHs final consumption expenditure (% of GDP)	50.00945	กลาง
Central government debt, total (% of GDP)	38.13193	กลาง
Exports of goods and services (% of GDP)	67.07088	กลาง
Broad money (% of GDP)	125.3963	กลาง
Manufacturing, value added (% of GDP)	27.14424	กลาง

กลุ่มที่ 2

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Agriculture, forestry and fishing	2498	ต่ำ
Mining and quarrying	0	ต่ำ
Manufacturing	606117.7	สูง
Electricity, gas, steam and air conditioning supply	37794.67	สูง
Water supply; sewerage, waste management and remediation activities	23224.67	สูง
Construction	108017.3	สูง
Wholesale and retail trade and repair of motor vehicles	1115445	สูง
Transportation and storage	381749.7	สูง
Accommodation and food service activities	441702.3	สูง

Information and communication'	275821.7	สูง
Financial and insurance activities	713617.3	สูง
Real estate activities	100784.7	สูง
Professional, scientific and technical activities	212400.7	สูง
Administrative and support service activities	148614.7	สูง
Public administration and defence; compulsory social security	520194.7	สูง
Education	89610.33	สูง
Human health activities	84743.67	สูง
Arts, entertainment and recreation	59606	สูง
Other service activities	140995	สูง
GPP Per capita (Baht)	573019.3	สูง

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
unemployed(จำนวนตำแหน่งงานว่าง)	80552.67	สูง
tourist(จำนวนผู้เยี่ยมเยือน)	62768783	สูง
CPI(ดัชนีราคาผู้บริโภคทั่วไป)	103.1667	สูง
poor(จำนวนคนจน)	113.3	สูง
factory(จำนวนโรงงาน)	182	สูง
Credit(จำนวนเงินให้สินเชื่อ)	563524.7	สูง

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง

Households and NPISHs final consumption expenditure (% of GDP)	49.28103	กลาง
Central government debt, total (% of GDP)	39.49103	กลาง
Exports of goods and services (% of GDP)	66.20006	กลาง
Broad money (% of GDP)	124.0984	กลาง
Manufacturing, value added (% of GDP)	26.95916	กลาง

กลุ่มที่ 3

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Agriculture, forestry and fishing	17289.4	สูง
Mining and quarrying	1753.958	ต่ำ
Manufacturing	28996.82	กลาง
Electricity, gas, steam and air conditioning supply	2861.125	กลาง
Water supply; sewerage, waste management and remediation activities	402.8056	ต่ำ
Construction	3381.278	ต่ำ
Wholesale and retail trade and repair of motor vehicles	13062.93	ต่ำ
Transportation and storage	3825.389	ต่ำ
Accommodation and food service activities	3954.403	ต่ำ
Information and communication'	1010.903	ต่ำ
Financial and insurance activities	5377.819	ต่ำ
Real estate activities	3262.722	ต่ำ

Professional, scientific and technical activities	344.2778	ต่ำ
Administrative and support service activities	824.9861	ต่ำ
Public administration and defence; compulsory social security	4642.403	ต่ำ
Education	7291.931	กลาง
Human health activities	2944.472	ต่ำ
Arts, entertainment and recreation	390.3889	ต่ำ
Other service activities	874.5139	ต่ำ
GPP Per capita (Baht)	134612.9	ต่ำ

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
unemployed(จำนวนตำแหน่งงานว่าง)	3665.694	ต่ำ
tourist(จำนวนผู้เยี่ยมเยือน)	2772827	ต่ำ
CPI(ดัชนีราคาผู้บริโภคทั่วไป)	101.4056	กลาง
poor(จำนวนคนจน)	71.77917	กลาง
factory(จำนวนโรงงาน)	38.30556	ต่ำ
Credit(จำนวนเงินให้สินเชื่อ)	15617.33	กลาง

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Households and NPISHs final consumption expenditure (% of GDP)	48.93086	กลาง
Central government debt, total (% of GDP)	39.49103	
Exports of goods and	66.67283	กลาง

services (% of GDP)		
Broad money (% of GDP)	124.0447	กลาง
Manufacturing, value added (% of GDP)	27.01482	กลาง

กลุ่มที่ 4

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Agriculture, forestry and fishing	17730.63	สูง
Mining and quarrying	1823.778	ต่ำ
Manufacturing	29806.21	กลาง
Electricity, gas, steam and air conditioning supply	2939.181	กลาง
Water supply; sewerage, waste management and remediation activities	413.5417	ต่ำ
Construction	3544.569	ต่ำ
Wholesale and retail trade and repair of motor vehicles	13976.89	ต่ำ
Transportation and storage	4031.597	ต่ำ
Accommodation and food service activities	4477.583	ต่ำ
Information and communication'	1122.583	ต่ำ
Financial and insurance activities	5610.681	ต่ำ
Real estate activities	3481.333	ต่ำ
Professional, scientific and technical activities	340.8194	ต่ำ
Administrative and support service activities	895.7361	ต่ำ
Public administration and	4966.111	ต่ำ

defence; compulsory social security		
Education	7245.597	กลาง
Human health activities	3181.597	ต่ำ
Arts, entertainment and recreation	449.6389	ต่ำ
Other service activities	902.875	ต่ำ
GPP Per capita (Baht)	139678.9	ต่ำ

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
unemployed(จำนวนตำแหน่งงานว่าง)	3066.917	ต่ำ
tourist(จำนวนผู้เยี่ยมเยือน)	2911409	ต่ำ
CPI(ดัชนีราคาผู้บริโภคทั่วไป)	102.2333	สูง
poor(จำนวนคนจน)	89.64444	สูง
factory(จำนวนโรงงาน)	39.11111	ต่ำ
Credit(จำนวนเงินให้สินเชื่อ)	15399.02	กลาง

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Households and NPISHs final consumption expenditure (% of GDP)	48.90276	กลาง
Central government debt, total (% of GDP)	40.85012	กลาง
Exports of goods and services (% of GDP)	64.85647	กลาง
Broad money (% of GDP)	122.8542	กลาง
Manufacturing, value added (% of GDP)	26.71844	กลาง

กลุ่มที่ 5

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Agriculture, forestry and fishing	13128.83	กลาง
Mining and quarrying	66586.67	สูง
Manufacturing	376209.3	สูง
Electricity, gas, steam and air conditioning supply	38665.08	สูง
Water supply; sewerage, waste management and remediation activities	3060.417	กลาง
Construction	11101	กลาง
Wholesale and retail trade and repair of motor vehicles	84664.25	กลาง
Transportation and storage	64056.83	กลาง
Accommodation and food service activities	21026.17	กลาง
Information and communication'	3432.75	กลาง
Financial and insurance activities	18747.92	กลาง
Real estate activities	10493.08	กลาง
Professional, scientific and technical activities	8782.833	กลาง
Administrative and support service activities	10868.42	กลาง
Public administration and defence; compulsory social security	15212.25	กลาง
Education	5397.5	ต่ำ
Human health activities	7003.667	กลาง
Arts, entertainment and	922.8333	ต่ำ

recreation		
Other service activities	7141.5	กลาง
GPP Per capita (Baht)	572967.8	สูง

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
unemployed(จำนวนตำแหน่งงานว่าง)	15846.5	กลาง
tourist(จำนวนผู้เยี่ยมเยือน)	7323581	กลาง
CPI(ดัชนีราคาผู้บริโภคทั่วไป)	101.3667	กลาง
poor(จำนวนคนจน)	17.29167	ต่ำ
factory(จำนวนโรงงาน)	273.75	สูง
Credit(จำนวนเงินให้สินเชื่อ)	20468.63	กลาง

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Households and NPISHs final consumption expenditure (% of GDP)	49.28103	กลาง
Central government debt, total (% of GDP)	39.49103	กลาง
Exports of goods and services (% of GDP)	66.20006	กลาง
Broad money (% of GDP)	124.0984	กลาง
Manufacturing, value added (% of GDP)	26.95916	กลาง

โดยสาเหตุที่ผลการจัดกลุ่มออกมาเป็นเช่นนี้อาจเป็นเพราะปัจจัยต่างๆของแต่ล่ะชุดข้อมูลมีความใกล้เคียงกับค่าเฉลี่ยของแต่ล่ะกลุ่มเลยถูกจัดกลุ่มออกมาในรูปแบบนี้

จังหวัดที่ได้รับมอบหมายคือ
-ชลบุรี อยู่ในกลุ่ม 5 โดยมี 24 จังหวัดอยู่ในกลุ่มนี้
-ปราจีนบุรี อยู่ในกลุ่ม 3 โดยมี 24 จังหวัดอยู่ในกลุ่มนี้
-สมุทรสงคราม อยู่ในกลุ่ม 3 โดยมี 24 จังหวัดอยู่ในกลุ่มนี้

จังหวัดชลบุรี มีรายได้หลักๆคือนิคมอุสาหกรรม เช่น อมตะนคร , นิคมอุสาหกรรมรจนะ นอกจากนี้ยังมีท่าเรือน้ำลึก ท่าเรือแหลมฉบังสำหรับส่งสินค้าเข้าและออกจากประเทศไทย และเป็นจังหวัดการท่องเที่ยวโดยเฉพาะเมืองพัทยาที่ทำรายได้ มีนักท่องเที่ยวมาเป็นจำนวนมาก มีฐานทัพเรือสัทหีบและมีสนามบินนานาชาติอู่ตะเพา และมีแหล่งท่องเที่ยวอื่ๆอีกมากมายเช่น บางแสน เกาะสีชัง เกาล้าน และตัวจังหวัดชลบุรีมีพื้นที่กว้างใหญ่และสามารถทำเกษตรได้จึงมีการปลูกพืนเช่น ข้าว อ้อย มันสำปะหลังและยางพารานอกจากนี้พื้นที่ยังติดชายฝั่งทะเลจึงมีการทำประมง บริเวณพื้นที่ที่ติดชายฝั่ง และมีจำนวนประชากรอาศัยอยู่เป็นจำนวนมาก โดยมีประชากรในทะเบียนกว่า 1.5 ล้านคน และมีประชากรแรงงานแฝงที่เป็นคนไทยประมาณ 1 ล้านคนและแรงงานต่างชาติ 1 ล้านคน มีที่ตั้งไม่ห่างจากกรุงเทพ มีเส้นทางรถยนต์เชื่อมต่อกับกรุงเทพหลายเส้นทาง ทั้งมอเตอร์เวย์ ถนนสุขุมวิท และทางด่วนบูรพาวิถี และมีเส้นทางรถไฟเชื่อมต่ออีกทั้งจะมีการสร้างทางรถไฟความเร็วสูงเชื่อมต่อ 3 นามบินไปยังชลบุรีในอีกไม่นาน

จังหวัดปราจีนบุรี มีพื้นที่เล็กกว่าจังหวัดชลบุรี และมีประชากร ราว 5 แสนคน เริ่มมีนิคมอุตสาหกรรมในตัวจังหวัด เช่น นิคมอุตสาหกรรม โรจนะ นิคมอุตสาหกรรมกระบินบุรี เป็นรายได้หลักของจังหวัด ส่วนรายได้รองลงมาคือการทำเกษตรกรรมและการเลี้ยงสัตว์ โดยมีเส้นทางสำหรับส่งสินค้าอุสาหรรมที่ท่าเรือน้ำลึกแหลมฉบังจังหวัดชลบุรี

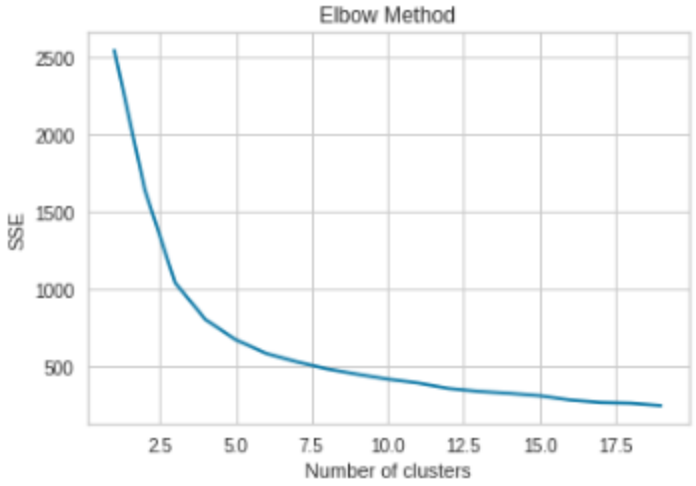
จังหวัดสมุทรสงครามเป็นจังหวัดที่มีขนาดเล็ก ประมาณ 416 ตารางกิโลเมตร (มีเพียง 3 อำเภอ) และมีประชากรเพียง 1 แสน 9 หมื่นคนโดยรายได้หลักมาจากการเกษตร กรารประมง และการท่องเที่ยว มีการทำนาเกลือ

จากการกระจายตัวของจังหวัดไปกลุ่มต่างๆนั้น สามารถสรุปลักษณะของภาคที่ได้รับมอบหมายได้ดังนี้ เนื่องด้วยเหตุผลดังกล่าวจึงจัดกลุ่มได้ว่า ชลบุรี อยู่ในกลุ่ม 5 โดยมีลักษณะของค่าของปัจจัยต่างๆอยู่ในช่วงกลางถึงสูงเนื่องจากมีประชากรเยอะ และ มีรายได้จากหลายปัจจัยทำให้จังหวัด มีค่าของปัจจัยในแต่ล่ะด้านสูง ต่างจากจังหวัด ปราจีนบุรี และสมุทรสงคราม ที่มีประชากรน้อยและเป็นจังหวัดขนาดเล็ก ดังนั้นจึงถูกจัดอยู่ในกลุ่มที่ 3 ซึ่งมีค่าของปัจจัยอย่ที่กลางถึงต่ำ

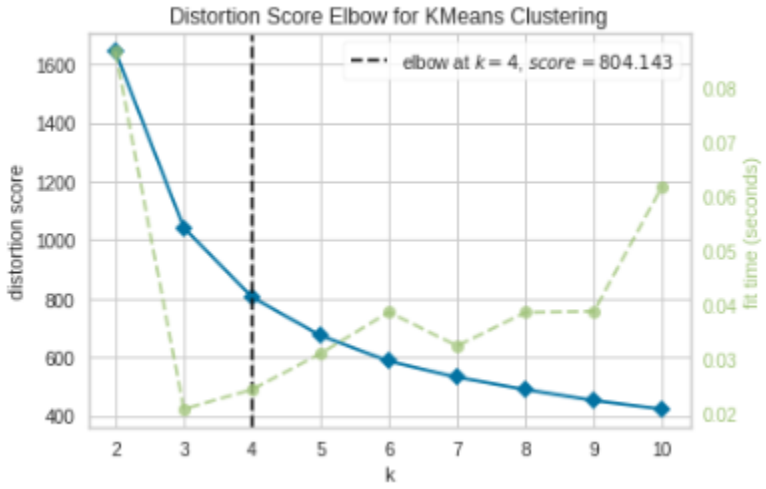
2.3 ทำซ้ำ 2.1 แต่ใช้เพียงบางปัจจัยที่ส่งผลต่อ GPP per capita และ ผลิตภัณฑ์จังหวัดต่อคน (GPP per capita))

คัดเลือกปัจจัยที่ส่งผลจากใน Part 1 โดยจะมีปัจจัยที่ส่งผล ดังนี้ 'Agriculture, forestry and fishing','Manufacturing','Wholesale and retail trade and repair of motor vehicles','Transportation and storage','tourist', 'CPI','factory', 'credit','Central government debt, total (% of GDP)','Exports of goods and services (% of GDP)', 'Broad money (% of GDP)'

Clustering ด้วยวิธี K-mean โดยนำข้อมูลปัจจัยที่ส่งผลที่ผ่านการทำ normalized มาทำ Elbow Method โดยได้ทำการรัน parameter ค่า k ตั้งแต่ 1 ถึง 20 และวัดระยะทาง intra-cluster จะได้ผลลัพธ์ออกมาดังนี้



ทางผู้จัดทำได้เลือก parameter ค่า k = 4 เนื่องจาก มีค่า Sum of intra-cluster measures ที่ไม่มากหรือน้อยจนเกินไป บ่งบอกถึงมีการแบ่ง cluster ได้อย่างเหมาะสม



โดยทางผู้จัดทำยังได้ทำการใช้ yellowbrick framework เพื่อช่วยยืนยัน ค่า parameter k=4 ว่าเหมาะสมที่จะนำไปใช้ในการแบ่งจำนวน cluster ด้วยวิธี k-mean

2.2 อภิปราย ผลลัพธ์ที่เกิดขึ้น

- เปรียบเทียบ กลุ่มที่เกิดขึ้น มีกี่กลุ่ม แต่ละกลุ่ม รวมจังหวัดที่มีลักษณะเหมือนกันอย่างไร
- จังหวัดที่ได้รับมอบหมาย อยู่กลุ่มใด และสมควรถูกจัดในกลุ่มนั้นหรือไม่
- จังหวัดในภาคที่ได้รับมอบหมายกระจายไปอยู่กลุ่มที่มีลักษณะแบบใดบ้าง มีกี่จังหวัดในกลุ่มนั้น จากการกระจายตัวของจังหวัดไปกลุ่มต่างๆนั้น สามารถสรุปลักษณะของภาคที่ได้รับมอบหมายได้อย่างไร

ผลลัพธ์ของวิธี K-mean ได้ทำการแบ่ง cluster ออกเป็น 4 กลุ่ม ดังนี้

Cluster_id	จำนวนข้อมูลที่อยู่ใน cluster นั้นๆ
0	24
1	47
2	1
3	5

กลุ่มที่ 1

	province	Year	Agriculture, forestry and fishing	Mining and quarrying	Manufacturing	Electricity, gas, steam and air conditioning supply	Water supply; sewerage, waste management and remediation activities	Construction	Wholesale and retail trade and repair of motor vehicles	Transportation and storage	...	CPI	poor	factory	credit	Households and NPISHs final consumption expenditure (% of GDP)	Central government debt, total (% of GDP)	Exports of goods and services (% of GDP)	Broad money (% of GDP)	Manufacturing, value added (% of GDP)	cluster_id
8	Pathum Thani	2018	6689	118	203401	9683	2592	11482	72533	10045	...	102.7	23.6	197	22703.60	48.902763	40.85012	64.856468	122.854209	26.718436	0
14	Nakhon Pathom	2018	21188	2138	190441	4383	1548	10368	45203	7890	...	102.0	24.6	175	17153.25	48.902763	40.85012	64.856468	122.854209	26.718436	0
17	Nonthaburi	2018	5861	54	50295	5473	2367	14167	62328	11171	...	103.2	11.8	92	23354.51	48.902763	40.85012	64.856468	122.854209	26.718436	0
20	Saraburi	2018	13484	13698	121481	29074	559	2566	24668	11053	...	101.4	50.5	51	15205.95	48.902763	40.85012	64.856468	122.854209	26.718436	0
23	Sing Buri	2018	4314	46	7184	625	200	816	3727	773	...	103.0	38.6	5	5193.54	48.902763	40.85012	64.856468	122.854209	26.718436	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
218	Maha Sarakham	2018	13000	67	8021	704	261	3343	6580	1740	...	102.4	98.4	20	23362.16	48.902763	40.85012	64.856468	122.854209	26.718436	0
221	Si Sa Ket	2018	20472	209	4596	771	175	3053	9190	1296	...	102.7	225.5	26	31353.32	48.902763	40.85012	64.856468	122.854209	26.718436	0
224	Nong Bua Lam Phu	2018	6912	360	5162	275	100	1326	4148	720	...	104.5	88.7	13	10098.10	48.902763	40.85012	64.856468	122.854209	26.718436	0
227	Amnat Charoen	2018	5457	10	1083	198	50	927	2324	308	...	103.2	64.8	12	5659.87	48.902763	40.85012	64.856468	122.854209	26.718436	0
230	Bungkan	2018	8304	449	4466	273	39	931	2263	423	...	102.2	41.4	37	9387.88	48.902763	40.85012	64.856468	122.854209	26.718436	0

73 rows × 36 columns

กลุ่มที่ 2

	province	Year	Agriculture, forestry and fishing	Mining and quarrying	Manufacturing	Electricity, gas, steam and air conditioning supply	Water supply; sewerage, waste management and remediation activities	Construction	Wholesale and retail trade and repair of motor vehicles	Transportation and storage	...	CPI	poor	factory	credit	Households and NPISHs final consumption expenditure (% of GDP)	Central government debt, total (% of GDP)	Exports of goods and services (% of GDP)	Broad money (% of GDP)	Manufacturing, value added (% of GDP)	cluster_id
12	Nakhon Pathom	2016	21313	2119	167758	4636	1294	10547	37786	6780	...	99.8	16.6	107	17947.76	50.009455	38.131933	67.070884	125.396343	27.144239	1
13	Nakhon Pathom	2017	20317	1968	175766	4811	1496	9596	40659	7451	...	101.1	17.3	138	17419.12	48.930863	39.491027	66.672832	124.044740	27.014815	1
15	Nonthaburi	2016	5535	53	47301	5548	1995	15053	52403	9908	...	101.1	10.6	91	22605.25	50.009455	38.131933	67.070884	125.396343	27.144239	1
16	Nonthaburi	2017	5430	53	53499	5454	2260	14256	58351	10549	...	102.0	11.5	79	22930.64	48.930863	39.491027	66.672832	124.044740	27.014815	1
18	Saraburi	2016	12164	17027	113403	30463	622	4261	21233	12705	...	100.6	25.5	58	15633.18	50.009455	38.131933	67.070884	125.396343	27.144239	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
223	Nong Bua Lam Phu	2017	6602	360	5182	304	102	1045	3943	710	...	103.9	55.8	8	10563.79	48.930863	39.491027	66.672832	124.044740	27.014815	1
225	Amnat Charoen	2016	4758	13	961	213	39	948	1919	228	...	100.8	66.6	14	6037.43	50.009455	38.131933	67.070884	125.396343	27.144239	1
226	Amnat Charoen	2017	5085	11	983	220	46	892	2107	273	...	102.1	39.0	13	5879.46	48.930863	39.491027	66.672832	124.044740	27.014815	1
228	Bungkan	2016	8772	320	4201	265	26	600	2384	378	...	102.2	26.6	22	10115.97	50.009455	38.131933	67.070884	125.396343	27.144239	1
229	Bungkan	2017	10467	289	5427	283	38	897	2654	406	...	101.2	19.5	9	9680.27	48.930863	39.491027	66.672832	124.044740	27.014815	1

140 rows × 36 columns

กลุ่มที่ 3

	province	Year	Agriculture, forestry and fishing	Mining and quarrying	Manufacturing	Electricity, gas, steam and air conditioning supply	Water supply; sewerage, waste management and remediation activities	Construction	Wholesale and retail trade and repair of motor vehicles	Transportation and storage	...	CPI	poor	factory	credit	Households and NPISHs final consumption expenditure (% of GDP)	Central government debt, total (% of GDP)	Exports of goods and services (% of GDP)	Broad money (% of GDP)	Manufacturing, value added (% of GDP)	cluster_id
0	Bangkok	2016	2479	0	601000	38798	21202	102327	1028711	362345	...	100.9	118.0	193	369297.90	50.009455	38.131933	67.070884	125.396343	27.144239	2
1	Bangkok	2017	2414	0	594681	38225	22898	109938	1104473	383936	...	103.3	98.3	189	590251.58	48.930863	39.491027	66.672832	124.044740	27.014815	2
2	Bangkok	2018	2601	0	622672	36361	25574	111787	1213150	398968	...	105.3	123.6	164	731024.54	48.902763	40.850120	64.856468	122.854209	26.718436	2

3 rows × 36 columns

กลุ่มที่ 4

	province	Year	Agriculture, forestry and fishing	Mining and quarrying	Manufacturing	Electricity, gas, steam and air conditioning supply	Water supply; sewerage, waste management and remediation activities	Construction	Wholesale and retail trade and repair of motor vehicles	Transportation and storage	...	CPI	poor	factory	credit	Households and NPISHS final consumption expenditure (% of GDP)	Central government debt, total (% of GDP)	Exports of goods and services (% of GDP)	Broad money (% of GDP)	Manufacturing, value added (% of GDP)	cluster_id
3	Samut Prakan	2016	2992	8	291877	15972	2537	10409	97576	163458	...	100.6	6.0	236	18788.41	50.009455	38.131933	67.070884	125.396343	27.144239	3
4	Samut Prakan	2017	4238	7	287135	16028	2696	9631	108377	174780	...	101.0	0.8	315	20382.93	48.930863	39.491027	66.672832	124.044740	27.014815	3
5	Samut Prakan	2018	4386	7	331085	16133	2718	11138	124765	171883	...	102.2	7.0	296	20983.56	48.902763	40.850120	64.856468	122.854209	26.718436	3
6	Pathum Thani	2016	5789	133	185025	9546	2427	11901	53144	8398	...	100.5	1.6	270	21947.55	50.009455	38.131933	67.070884	125.396343	27.144239	3
7	Pathum Thani	2017	6207	112	202558	11415	2754	11801	65417	9288	...	101.3	0.0	202	22721.03	48.930863	39.491027	66.672832	124.044740	27.014815	3
9	Samut Sakhon	2016	14104	764	236941	8386	2703	7267	47602	5137	...	100.2	19.2	408	6844.73	50.009455	38.131933	67.070884	125.396343	27.144239	3
10	Samut Sakhon	2017	7585	847	268143	8597	2702	6979	53271	5473	...	101.3	4.9	451	6688.63	48.930863	39.491027	66.672832	124.044740	27.014815	3
11	Samut Sakhon	2018	7718	557	280167	7827	2374	7559	59175	5701	...	102.1	52.0	450	6653.83	48.902763	40.850120	64.856468	122.854209	26.718436	3
34	Phra Nakhon Si Ayutthaya	2017	10334	2190	268841	9169	801	4736	46668	14165	...	101.2	21.6	117	27267.56	48.930863	39.491027	66.672832	124.044740	27.014815	3
54	Chon Buri	2016	17512	3997	460858	56565	4444	22782	90290	57498	...	101.0	0.5	234	36124.57	50.009455	38.131933	67.070884	125.396343	27.144239	3
55	Chon Buri	2017	19580	3365	484836	54055	4460	17813	101209	60028	...	102.5	35.5	243	37378.56	48.930863	39.491027	66.672832	124.044740	27.014815	3
56	Chon Buri	2018	19730	3318	522108	66966	4724	17555	111368	65041	...	104.1	5.7	250	38681.35	48.902763	40.850120	64.856468	122.854209	26.718436	3
60	Rayong	2016	18392	248718	413495	56536	2366	7614	66064	17843	...	99.7	19.3	142	18455.00	50.009455	38.131933	67.070884	125.396343	27.144239	3
61	Rayong	2017	22342	255652	464253	73696	2482	6910	75628	19075	...	100.1	16.9	121	17784.43	48.930863	39.491027	66.672832	124.044740	27.014815	3
62	Rayong	2018	18967	281800	473613	83220	2519	7555	80646	22765	...	101.6	39.7	139	16857.54	48.902763	40.850120	64.856468	122.854209	26.718436	3

โดยค่าเฉลี่ยของปัจจัยในแต่ล่ะ Cluster มีดังนี้

cluster_id	Agriculture, forestry and fishing	Manufacturing	Wholesale and retail trade and repair of motor vehicles	Transportation and storage	tourist	CPI	factory	credit	Central government debt, total (% of GDP)	Exports of goods and services (% of GDP)	Broad money (% of GDP)
0	17629.972603	29467.123288	1.383553e+04	3982.931507	2.882041e+06	102.280822	38.589041	15253.505342	40.831502	64.881350	122.870517
1	17097.500000	24164.164286	1.163481e+04	3581.557143	2.633268e+06	100.986429	37.200000	15662.623929	38.801772	66.874701	124.730196
2	2498.000000	606117.666667	1.115445e+06	381749.666667	6.276878e+07	103.166667	182.000000	563524.673333	39.491027	66.200062	124.098431
3	11991.733333	344729.000000	7.874667e+04	53368.866667	6.628357e+06	101.293333	258.266667	21170.645333	39.400420	66.321153	124.177799

เนื่องจากค่าที่ได้ในแต่ล่ะกลุ่มไม่สามารถเจาะจงอธิบาลรายละเอียดได้อย่างชัดเจนจึงอธิบายผ่านข้อมูลปัจจัยโดยเฉลี่ยของแต่ล่ะกลุ่มแทน โดยแต่ล่ะกลุ่มจะมีลักษณะคร่าวๆดังนี้

กลุ่มที่ 1

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Agriculture, forestry and fishing	17629.9726	สูง
Manufacturing	29467.12329	ต่ำ
Wholesale and retail trade and repair of motor vehicles	13835.53425	ต่ำ
Transportation and storage	3982.931507	ต่ำ

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
tourist(จำนวนผู้เยี่ยมเยือน)	2882041.219	ต่ำ

CPI(ดัชนีราคาผู้บริโภคทั่วไป)	102.2808219	สูง
factory(จำนวนโรงงาน)	38.5890411	ต่ำ
Credit(จำนวนเงินให้สินเชื่อ)	15253.50534	กลาง

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Central government debt, total (% of GDP)	40.83150225	กลาง
Exports of goods and services (% of GDP)'	64.88135019	กลาง
Broad money (% of GDP)	122.8705173	กลาง

กลุ่มที่ 2

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Agriculture, forestry and fishing	17097.5	สูง
Manufacturing	24164.16429	ต่ำ
Wholesale and retail trade and repair of motor vehicles	11634.81429	ต่ำ
Transportation and storage	3581.557143	ต่ำ

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
tourist(จำนวนผู้เยี่ยมเยือน)	2633267.664	ต่ำ
CPI(ดัชนีราคาผู้บริโภคทั่วไป)	100.9864286	กลาง
factory(จำนวนโรงงาน)	37.2	ต่ำ
Credit(จำนวนเงินให้สินเชื่อ)	15662.62393	กลาง

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Central government debt, total (% of GDP)	38.80177225	กลาง
Exports of goods and services (% of GDP)	66.87470145	กลาง
Broad money (% of GDP)	124.7301958	กลาง

กลุ่มที่ 3

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Agriculture, forestry and fishing	2498	ต่ำ
Manufacturing	606117.6667	สูง
Wholesale and retail trade and repair of motor vehicles	1115444.667	สูง
Transportation and storage	381749.6667	สูง

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
tourist(จำนวนผู้เยี่ยมเยือน)	62768783	สูง
CPI(ดัชนีราคาผู้บริโภคทั่วไป)	103.1666667	สูง
factory(จำนวนโรงงาน)	182	กลาง
Credit(จำนวนเงินให้สินเชื่อ)	563524.6733	สูง

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Central government debt, total (% of GDP)	39.4910267	กลาง
Exports of goods and services (% of GDP)	66.20006164	กลาง
Broad money (% of GDP)	124.0984305	กลาง

กลุ่มที่ 4

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Agriculture, forestry and fishing	11991.73333	กลาง
Manufacturing	344729	กลาง
Wholesale and retail trade and repair of motor vehicles	78746.66667	กลาง
Transportation and storage	53368.86667	กลาง

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
tourist(จำนวนผู้เยี่ยมเยือน)	6628357.333	กลาง
CPI(ดัชนีราคาผู้บริโภคทั่วไป)	101.2933333	กลาง
factory(จำนวนโรงงาน)	258.2666667	สูง
Credit(จำนวนเงินให้สินเชื่อ)	21170.64533	กลาง

ปัจจัย	ค่าเฉลี่ย	อยู่ในช่วง
Central government debt, total (% of GDP)	39.40042048	กลาง
Exports of goods and services (% of GDP)	66.32115258	กลาง
Broad money (% of GDP)	124.1777993	กลาง

โดยสาเหตุที่ผลการจัดกลุ่มออกมาเป็นเช่นนี้อาจเป็นเพราะปัจจัยต่างๆของแต่ล่ะชุดข้อมูลมีความใกล้เคียงกับค่าเฉลี่ยของแต่ล่ะกลุ่มเลยถูกจัดกลุ่มออกมาในรูปแบบนี้

จังหวัดที่ได้รับมอบหมายคือ
-ชลบุรี อยู่ในกลุ่ม 4 โดยมี 5 จังหวัดอยู่ในกลุ่มนี้
-ปราจีนบุรี อยู่ในกลุ่ม 2 โดยมี 47 จังหวัดอยู่ในกลุ่มนี้
-สมุทรสงคราม อยู่ในกลุ่ม 2 โดยมี 47 จังหวัดอยู่ในกลุ่มนี้

จากการกระจายตัวของจังหวัดไปกลุ่มต่างๆนั้น สามารถสรุปลักษณะของภาคที่ได้รับมอบหมายได้ดังนี้ เนื่องด้วยเหตุผลทางด้าน ประชากร รายได้ และ ขนาดพื้นที่ของจังหวัด จึงจัดกลุ่มได้ว่า ชลบุรี อยู่ใน

กลุ่ม 4 โดยมีลักษณะของค่าของปัจจัยต่างๆอยู่ในช่วงกลางถึงสูงเนื่องจากมีประชากรเยอะ และ มีรายได้จากหลายปัจจัยทำให้จังหวัด มีค่าของปัจจัยในแต่ล่ะด้านสูง ต่างจากจังหวัด ปราจีนบุรี และสมุทรสงคราม ที่มีประชากรน้อยและเป็นจังหวัดขนาดเล็ก ดังนั้น จึงถูกจัดอยู่ในกลุ่มที่ 2 ซึ่งมีค่าของปัจจัยอย่ที่กลางถึงต่ำ

2.4 อภิปราย ผลลัพธ์ที่เกิดขึ้น ตาม 2.2 และ เปรียบเทียบผลลัพธ์ เมื่อใช้บางปัจจัย (2.3) เมื่อเทียบกับใช้ทุกปัจจัย (2.1) ว่า ผลลัพธ์ของ cluster ที่ได้ เมื่อใช้บางปัจจัยและทุกปัจจัย มีความเหมือน ความต่างอย่างไร

เปรียบเมื่อใช้ทุกปัจจัยกับใช้บางปัจจัยพบ ว่า ผลลัพธ์ของ cluster ต่างกันคือมี 5 กลุ่ม และ 4 กลุ่ม ตามลำดับ กล่าวคือการใช้ปัจจัยที่ส่งผลจะช่วยให้เห็นกลุ่มได้ชัดเจนขึ้น โดย เมื่อใช้ทุกปัจจัย จะเห็นได้ว่ามี 3 กลุ่ม ที่มีจำนวน 24 จังหวัด โดยจะมีค่าของปัจจัยอยู่ประมาณ กลางถึงต่ำ เมื่อเทียบกับการใช้ บางปัจจัย จะเห็นได้ว่า มี 1 กลุ่มที่มีจำนวน 47 จังหวัด ซึ่งมีค่าของปัจจัยอยู่ประมาณ กลางถึงต่ำ เช่นเดียวกัน แต่ในทั้งสองครั้งของการทำ Clustering จะเห็นได้ว่ามี 1 กลุ่มที่แยกออกมาและมีจำนวน 1 จังหวัด คือ กรุงเทพมหานคร ซึ่งมีค่าของปัจจัยสูง อาจสามารถกล่าวได้ว่ามีความเหลื่อมล้ำ ของจังหวัดเกิดขึ้นจึงได้เกิดการแบ่งกลุ่มเช่นนี้ และยังมีเหตุผลสนับสนุนอีกคือ ในทั้งสองครั้งของการทำ Clustering จะเห็นได้ว่ามีอีก 1 กลุ่มที่จำแนกออกมาและมีสามาชิกอยู่ประมาณ 4 - 5 จังหวัด ซึ่งมีค่าของปัจจัยอยู่ประมาณกลางถึงสูง อาจกล่าวได้ว่ามีจังหวัดที่มีความเจริญใกล้เคียงกับกรุงเทพมหานคร แต่มีจำนวนน้อย เช่น จังหวัด ชลบุรี เป็นต้น

Part 3: Classification

<u>วัตถุประสงค์</u>

- จำแนกประเทศ ตามด้านที่สนใจ

3.1 Decision Tree

โดยข้อมูลที่ได้รับมอบหมายคือ Human Development โดยจะเป็นการจำแนกประเทศตามดัชนีการพัฒนามนุษย์ และ ในส่วนนี้ใช้ข้อมูลปี 2014-2016 เป็น train data และ ปี 2017 เป็น test data

3.1.1 จากปัจจัยที่ให้มา ใช้ Feature selection ช่วยในการหาว่า ปัจจัยใด ควรถูกเลือกมา สำหรับการจำแนกแต่ละแบบ และ แสดงผลลัพธ์ และ วิเคราะห์ผลลัพธ์ที่ได้

Feature selection ใช้วิธี Low variance Filtering ได้ทำการคัดเลือก ปัจจัยที่ min_var_threshold = 100 ได้ผลลัพธ์ปัจจัยที่ผ่านการคัดเลือกดังนี้

'pop',
'gdp',
'primary',
'secondary',
'tertiary',
'mortality_infant_rate',
'basic_drinking_water',

'forest_land',
'broadband',
'mobile',
'Birth_rate'

โดยผลลัพธ์ที่ได้มีความใกล้เคียงกับปัจจัยจริง คือมีปัจจัยจริงอยู่ในปัจจัยที่ผ่านการคัดเลือกด้วย ก็คือปัจจัย 'primary', 'secondary', 'tertiary', หรือก็คือ Education และ ยังมี GDP ซึ่งก็คือค่าที่ใกล้เคียงกับ GNI อยู่ด้วยซึ่งถือว่าผลลัพธ์สมเหตุสมผล

3.1.2 สร้างโมเดลจำแนกประเทศตามด้านที่ได้รับมอบหมาย โดยใช้ train data และ ใช้

- ทุกปัจจัย (19/18 ปัจจัย) ในการจำแนก
- ปัจจัยที่เลือกมาจาก feature selection 3.1.1 ระบุปัจจัยที่เลือกมา
- ปัจจัยจริง

เปรียบเทียบผลลัพธ์ทั้ง 2 แบบ รายงานผลของ test data ในตารางด้านล่าง

ส่วนบางปัจจัยที่ใช้จำแนกตามการพัฒนามนุษย์ คือ Gross National Income (GNI) , life expectancy และ Education

ปัจจัยที่เลือกมาจาก feature selection มีดังนี้

'pop',
'gdp',
'primary',
'secondary',
'tertiary',
'mortality_infant_rate',
'basic_drinking_water',
'forest_land',
'broadband',
'mobile',
'Birth_rate'

ปัจจัยจริง ได้ใช้ปัจจัยดังนี้

'Gdp', (ถ้าดูตามความหมายของ GNI จะใกล้เคียงกับ GDP มากที่สุด)
'Secondary', (ใช้ปัจจัยนี้เพราะมีค่าการกระจายตัวสูงที่สุดระหว่าง Education ทั้ง 3 ตัว)
'life expectancy',

	โมเดลที่ใช้ทุกปัจจัย	โมเดลที่ใช้ปัจจัยจาก feature selection	โมเดลที่ใช้ปัจจัยจริง
Precision Class 0	0.83	0.86	0.91

Precision Class 1	0.80	0.76	0.82
Precision Class 2	0.77	0.85	0.76
Precision Class 3	0.96	0.91	0.85
Recall Class 0	0.50	0.79	0.83
Recall Class 1	0.92	0.89	0.79
Recall Class 2	0.91	0.77	0.85
Recall Class 3	0.96	0.94	0.88
F-score Class 0	0.62	0.83	0.87
F-score Class 1	0.86	0.82	0.81
F-score Class 2	0.83	0.81	0.80
F-score Class 3	0.96	0.92	0.87
Accuracy	0.87	0.85	0.84
(macro) Precision	0.84	0.85	0.84
(macro) Recall	0.82	0.85	0.84
(macro) F-score	0.82	0.84	0.84

3.1.3 เลือกค่า Precision, Recall, Accuracy, F1-score มา 1 คลาส ของโมเดลที่ใช้ทุกปัจจัย แสดงวิธีทำตัวอย่างการคำนวณค่า Precision,Recall,Accuracy และ F1-score ของ Class 3 จากโมเดลที่ใช้ทุกปัจจัย

Confusion Matrix ที่ได้จากโมเดลที่ใช้ทุกปัจจัย

		ข้อมูลที่ได้จากการทำนาย			
		Class 0	Class 1	Class 2	Class 3
ข้อมูลจริง	Class 0	5	1	3	1
	Class 1	1	12	0	0
	Class 2	0	1	10	0
	Class 3	0	1	0	25

True positive(TP) : ข้อมูลจริงที่อยู่ใน Class 3 และโมเดลทำนายได้ถูกต้องว่าอยู่ Class 3
True negative (TN) : ข้อมูลจริงที่ไม่ได้อยู่ใน Class 3 และโมเดลทำนายได้ถูกต้องตรงตาม Class ข้อมูลจริง
False positive(FP) : ข้อมูลจริงที่ไม่ได้อยู่ใน Class 3 แต่โมเดลทำนายว่าอยู่ Class 3
False negative (FN) : ข้อมูลจริงที่อยู่ใน Class 3 แต่โมเดลไม่ได้ทำนายว่าอยู่ Class 3

แสดงวิธีทำ : TP = 25 TN = 10+12+5 = 27
FP = 1+0+0 = 1 FN = 0+1+0 = 1

Precision ของ class 3 = $\frac{TP}{TP+FP}$ = $\frac{25}{25+1}$ = 0.9615

Recall ของ Class 3 = $\frac{TP}{TP+FN}$ = $\frac{25}{25+1}$ = 0.9615

F1-score ของ Class 3 = 2 × $\frac{Precision \times Recall}{Precision + Recall}$ = 2 × $\frac{0.9615 \times 0.9615}{0.9615 + 0.9615}$ = 0.9615

Accuracy รวม = $\frac{TP + TN}{All}$ = $\frac{25+27}{60}$ = 0.8667

3.1.4 อภิปรายผลลัพธ์ที่เกิดขึ้น ของการจำแนกทั้ง 2 ด้าน

- เปรียบเทียบ โมเดลที่ใช้ทุกปัจจัย ใช้ปัจจัยจาก feature selection และใช้ปัจจัยจริง
- ในแต่ละโมเดล เปรียบเทียบการจำแนกประเทศที่ได้จาก test data ประเทศที่มีการจำแนกผิดพลาด จำแนกผิดไปคลาสใด อภิปรายเหตุผลที่ทำให้เกิดการจำแนกผิดดังกล่าว
- วิเคราะห์คลาสแต่ละคลาส และให้คำนิยาม ว่า คลาส 0,1,2(,3)… แทนกลุ่มประเทศที่มีลักษณะใด (สังเกตค่าปัจจัยของประเทศต่างๆในคลาสเดียวกัน ว่า มีค่าใดเหมือนหรือคล้ายกัน) นอกจากนี้ ประเทศไทย อยู่ในคลาสใด (จากคลาสจริง และคลาสทำนาย)

โมเดลที่ใช้ทุกปัจจัย และ ใช้ปัจจัยจาก feature selection และใช้ปัจจัยจริงมีค่า Accuracy , (macro) Precision , (macro) Recall , (macro) F-score ที่ใกล้เคียงกัน โดย ประเทศที่มีการจำแนกผิดพลาด สาเหตุเนื่องมาจาก วิเคราะผลลัพธ์ของโมเดลที่ได้ ดูจากค่า precision , recall , f1-score เป็นค่าเอาไว้ประเมินโมเดล (precision เอาไว้บอกว่าค่าที่ทำนายมาแล้วค่าที่ทำนายถูก ตรงกับข้อมูลจริง , recall เป็นค่าที่เอาไว้บอกว่าค่าที่ทำนายมาเป็น positive เป็นค่าข้อมูลที่ positive จริงๆ เท่าไหร่ , f1-score เป็นค่าการถ่วงน้ำหนักระหว่าง precision และ recall) เมื่อดูจากค่า f1-score จะเห็นได้ว่าข้อมูลเป็น imbalance data จึงมีการตอบข้อมูลที่เป็น class 3 มามากกว่า class อื่นๆ โดยในแต่ละคลาส class 0,1,2,3 คือ class ที่มีลักษณะ human development index จำแนกตาม Class นั้นๆ และประเทศไทย มี Class จริงอยู่ที่ Class 0 และได้ Class ทำนายว่าอยู่ใน Class 0 เช่นเดียวกัน อาจบ่งบอกได้ว่าประเทศไทย มี human development index ต่ำนั้นเอง

3.2 Logistics Regression

ใช้ข้อมูลการจำแนกกลุ่มตามประเทศที่พัฒนา (1) และไม่พัฒนา (0) เพื่อแสดงการจำแนกกลุ่มด้วยการใช้วิธี Logistics Regression

3.2.1 สร้างโมเดลจำแนกประเทศ โดยใช้ปัจจัยที่ให้มา แบ่งข้อมูล train:test 70:30 วิเคราะผลลัพธ์ของโมเดลที่ได้

สร้างโมเดลออกมาได้ผลลัพธ์ดังนี้

Logit Regression Results

Dep. Variable:	dev	**No. Observations:**	166
Model:	Logit	**Df Residuals:**	153
Method:	MLE	**Df Model:**	12
Date:	Wed, 06 Apr 2022	**Pseudo R-squ.:**	0.8405
Time:	09:51:46	**Log-Likelihood:**	-14.976
converged:	True	**LL-Null:**	-93.893
Covariance Type:	nonrobust	**LLR p-value:**	1.450e-27

	coef	std err	z	P>|z|	[0.025	0.975]
pop	-1.409e-08	2.97e-08	-0.475	0.635	-7.22e-08	4.41e-08
gdp	3.849e-13	9.12e-13	0.422	0.673	-1.4e-12	2.17e-12
gni	-6.866e-14	5.55e-13	-0.124	0.902	-1.16e-12	1.02e-12
mortality_infant_rate	-0.6401	0.332	-1.930	0.054	-1.290	0.010
life_expectancy	-0.3237	0.270	-1.198	0.231	-0.853	0.206
basic_drinking_water	-0.4301	0.241	-1.782	0.075	-0.903	0.043
co2	0.1336	0.197	0.677	0.499	-0.253	0.521
forest_land	-0.0412	0.036	-1.154	0.249	-0.111	0.029
broadband	0.0262	0.081	0.322	0.747	-0.133	0.185
mobile	0.0128	0.025	0.521	0.602	-0.035	0.061
birth_rate	0.1538	0.549	0.280	0.779	-0.922	1.229
age_15_64	0.1240	0.457	0.271	0.786	-0.772	1.020
senior	0.8412	0.562	1.497	0.134	-0.260	1.943
const	47.7797	55.254	0.865	0.387	-60.517	156.076

แล้วทำการแบ่งข้อมูล train:test 70:30 เพื่อดูผลลัพธ์ที่ได้ ดังนี้

	precision	recall	f1-score	support
0	0.88	0.72	0.79	39
1	0.39	0.64	0.48	11
accuracy			0.70	50
macro avg	0.63	0.68	0.64	50
weighted avg	0.77	0.70	0.72	50

วิเคราะผลลัพธ์ของโมเดลที่ได้ ดูจากค่า precision , recall , f1-score เป็นค่าเอาไว้ประเมินโมเดล (precision เอาไว้บอกว่าค่าที่ทำนายมาแล้วค่าที่ทำนายถูก ตรงกับข้อมูลจริง , recall เป็นค่าที่เอาไว้บอกว่าค่าที่ทำนายมาเป็น positive เป็นค่าข้อมูลที่ positive จริงๆ เท่าไหร่ , f1-score เป็นค่าการถ่วงน้ำหนักระหว่าง precision และ recall) เมื่อดูจากค่า f1-score จะเห็นได้ว่าข้อมูลเป็น imbalance data จึงมีการตอบข้อมูลที่เป็น class 0 มามากกว่า class 1 แต่หากดูจากค่าของแต่ล่ะ class แล้ว (Class 0 มี precision = 0.88 , recall = 0.72 , f1-score = 0.79) , (Class 1 มี precision = 0.39 , recall = 0.64 , f1-score = 0.48) จึงถือเป็นโมเดลที่ไม่ค่อยดีนัก

3.2.2 ปัจจัยใดบ้าง ที่คาดว่า จะส่งผลต่อการจำแนกกลุ่มตามการพัฒนาของประเทศมากกว่าปัจจัยอื่นๆ (ในกรณีนี้ พิจารณาค่า p-value ที่ต่ำหรือสูง (ไม่จำเป็นต้องต่ำกว่า 0.1 ในการพิจารณา) และ จากปัจจัยกลุ่มนี้สะท้อนให้เห็นคุณสมบัติของประเทศที่พัฒนาแล้วอย่างไรบ้าง

พิจารณาที่ค่า p-value ต่ำกว่า 0.3 ปัจจัยที่คาดว่าจะส่งผลมีดังนี้ :

- Mortality_infant_rate (อัตราการตายของทารก)
- Life_expectancy (ช่วงอายุของชีวิต)
- Basic_drinking_water (คุณภาพการเข้าถึงน้ำสะอาด)
- Forest_land (พื้นที่ป่าไม้)
- Senior (ประชากรผู้สูงอายุ)

จากปัจจัยกลุ่มนี้ สะท้อนให้เห็นคุณสมบัติของประเทศที่พัฒนาแล้ว เป็นเหมือนการมองภาพรวมที่เมื่ออยู่ในประเทศที่พัฒนาควรจะมีความสามารถด้านการแพทย์ให้เด็กทารกสามารถคลอดออกมาแล้วมีชีวิตได้ปกติ ไม่พิการ และได้เติบโต , ใช้ชีวิตผ่านพื้นที่ธรรมชาติเพื่อสภาวะแวดล้อมที่น่าอยู่และสามารถเข้าถึงน้ำสะอาดได้โดยง่าย โดยการใช้ชีวิตก็จะต้องใช้ชีวิตที่ยืนยาว ดังนั้นประชากรผู้สูงอายุจึงเป็นตัวบ่งบอกว่าประเทศนี้เหมาะแก่การใช้ชีวิต ไม่เจ็บป่วยหรือตายก่อนที่จะถึงวัยสูงอายุนั้นเอง

3.2.3 นำโมเดลที่ได้ ไปทำนายกลุ่มของประเทศไทย และอีก 1 ประเทศ วิเคราะห์ผลลัพธ์ที่ได้

ทำนายกลุ่มของประเทศไทย ได้ผลลัพธ์ ดังนี้

```
df[df["country"] == "Thailand"]

predictions = model.predict(df[df["country"] == "Thailand"].drop(['country','unemployment','primary', 'secondary','tertiary','phy
predictions
```

```
array([0])
```

กล่าวคือโมเดลทำนายว่าจากปัจจัยทั้งหมดประเทศไทยยังเป็นประเทศที่ยังไม่พัฒนา

ทำนายกลุ่มของประเทศนิวซีแลนด์ ได้ผลลัพธ์ ดังนี้

```
df[df["country"] == "New Zealand"]

predictions = model.predict(df[df["country"] == "New Zealand"].drop(['country','unemployment','primary', 'secondary','tertiary','
predictions
```

```
array([1])
```

กล่าวคือโมเดลทำนายว่าจากปัจจัยทั้งหมดประเทศนิวซีแลนด์ยังเป็นประเทศที่พัฒนาแล้ว

3.3 Bayes Classification

แสดงการจำแนกกลุ่มด้วยการใช้วิธี Bayes Classification เพื่อทำนายกลุ่มของ Happiness (ในการทำ Bayes Classification ได้ใช้ของมูลของ Happiness ปี 2014 และได้ใช้ 4 ปัจจัย คือ

1.GDP
2. Unemployment
3.life_expectancy
4.Social support

3.3.1 ระบุจำนวนประเทศใน กลุ่ม 1,2 และ 3 เพื่อนำไปหาความน่าจะเป็นของกลุ่ม 0, 1 และ 2 (เทียบเท่ากับการหา Prior distribution P(θ))

	จำนวนประเทศในแต่ละ Class	P(Happiness)
กลุ่ม 1	31 ประเทศ	0.2897196262
กลุ่ม 2	43 ประเทศ	0.4018691589
กลุ่ม 3	33 ประเทศ	0.308411215

3.3.2 เลือกปัจจัยมา 3 ตัว ที่ส่งผลต่อ Happiness และใช้ Social support เป็นปัจจัยตัวที่ 4 ระบุความน่าจะเป็นของปัจจัยต่างๆ เมื่อทราบกลุ่ม (เทียบได้กับการหา Likelihood function $P(X \mid \theta)$

ทำการเลือกมา 4 ปัจจัย คือ

1.GDP
2. Unemployment
3.life_expectancy
4.Social support

เนื่องจากปัจจัยเป็นข้อมูลตัวเลข ไม่ใช่ categorical data ในการระบุความน่าจะเป็น ต้องแบ่งค่าของปัจจัยแต่ละตัวเป็นช่วงๆก่อน แล้วจึงคำนวนความน่าจะเป็น Likelihood function $P(X \mid \theta)$

ได้ทำการแบ่งช่วงตามค่าต่อไปนี้

	Low	Mid	High
GDP	ต่ำกว่า 99×10^9	อยู่ระหว่าง 100×10^9 ถึง 999×10^9	สูงกว่า $1{,}000 \times 10^9$
Unemployment	ต่ำกว่า 5.99	อยู่ระหว่าง 6.00 ถึง 10.99	สูงกว่า 11.00
life_expectancy	ต่ำกว่า 70	อยู่ระหว่าง 71 ถึง 79	สูงกว่า 80
Social support	ต่ำกว่า 0.69	อยู่ระหว่าง 0.70 ถึง 0.89	สูงกว่า 0.90

ได้ทำการหา Likelihood function $P(X \mid \theta)$ ได้ผลลัพธ์ดังนี้

1.GDP

	X1=GDP		
	P(X1=Low|Happiness)	P(X1=Mid|Happiness)	P(X1=High|Happiness)
class 1	0.5483870968	0.4193548387	0.03225806452
class 2	0.5813953488	0.2558139535	0.1627906977

class 3	0.2727272727	0.5151515152	0.2121212121

2. Unemployment

	X2=unemployment		
	P(X2=Low|Happiness)	P(X2=Mid|Happiness)	P(X2=High|Happiness)
class 1	0.3548387097	0.3870967742	0.2580645161
class 2	0.511627907	0.2325581395	0.2558139535
class 3	0.4545454545	0.4545454545	0.09090909091

3.life_expectancy

	X3=life_expectancy		
	P(X3=Low|Happiness)	P(X3=Mid|Happiness)	P(X3=High|Happiness)
class 1	0.3870967742	0.5161290323	0.09677419355
class 2	0.1627906977	0.6046511628	0.2325581395
class 3	0.1212121212	0.3939393939	0.4848484848

4.Social support

	X4=Social support		
	P(X4=Low|Happiness)	P(X4=Mid|Happiness)	P(X4=High|Happiness)
class 1	0.3548387097	0.6451612903	0
class 2	0.06976744186	0.6976744186	0.2325581395
class 3	0	0.3333333333	0.6666666667

3.3.3 ทำนาย ว่า ประเทศไทย หากใช้วิธีการคำนวณแบบ Bayes Classification จะอยู่ในกลุ่ม Happiness กลุ่มใด แสดงวิธีทำประกอบ

ประเทศไทยมีช่วงของแต่ล่ะปัจจัย ดังนี้ :

1.GDP : Mid
2. Unemployment : Low
3.life_expectancy : Mid
4.Social support : High

จึงได้นำค่าความน่าจะเป็นของแต่ล่ะปัจจัยมาคูณกันเพื่อดูผลลัพธ์

Thailand	P(X1=Mid|Happiness)	P(X2=Low|Happiness)	P(X3=Mid|Happiness)	P(X4=High|Happiness)	P(Happiness|X1=Mid,X2=Low,X3=Mid,X4=High)
class 1	0.4193548387	0.3548387097	0.5161290323	0	0
class 2	0.2558139535	0.511627907	0.6046511628	0.2325581395	0.01840411302
class 3	0.5151515152	0.4545454545	0.3939393939	0.6666666667	0.06149650778

ค่า P(Happiness|X1=Mid,X2=Low,X3=Mid,X4=High) เป็นผลลัพธ์จากการคูณกันของแต่ล่ะ row ซึ่งจะเป็นได้ว่า row ของ Class 3 ได้ค่า P(Happiness|X1=Mid,X2=Low,X3=Mid,X4=High) ออกมาเยอะที่สุดดังนั้นประเทศไทยจึงถูกจัดอยู่ใน กลุ่ม 3 ตามการคำนวณแบบ Bayes Classification

3.3.4 เลือกประเทศเพิ่มอีก 1 ประเทศ ที่กลุ่ม Happiness จริงของประเทศนั้น ต่างจากกลุ่มจริงของประเทศไทย และใช้ Bayes Classification เพื่อทำนายกลุ่มของประเทศที่เลือกมา แสดงวิธีทำประกอบ

ประเทศที่เลือกมาคือประเทศจีนซึ่งกลุ่ม Happiness จริงของประเทศจีน คือ กลุ่ม 1 ส่วนประเทศไทย อยู่ในกลุ่ม Happiness กลุ่มที่ 2

มีช่วงของแต่ล่ะปัจจัย ดังนี้ :

1.GDP : High
2. Unemployment : Low
3.life_expectancy : Mid
4.Social support : Mid

จึงได้นำค่าความน่าจะเป็นของแต่ล่ะปัจจัยมาคูณกันเพื่อดูผลลัพธ์

China	P(X1=High|Happiness)	P(X2=Low|Happiness)	P(X3=Mid|Happiness)	P(X4=Mid|Happiness)	P(Happiness|X1=High,X2=Low,X3=Mid,X4=Mid)
class 1	0.03225806452	0.3548387097	0.5161290323	0.6451612903	0.003811499684
class 2	0.1627906977	0.511627907	0.6046511628	0.6976744186	0.03513512486
class 3	0.2121212121	0.4545454545	0.3939393939	0.3333333333	0.01266104572

ค่า P(Happiness|X1=High,X2=Low,X3=Mid,X4=Mid) เป็นผลลัพธ์จากการคูณกันของแต่ล่ะ row ซึ่งจะเป็นได้ว่า row ของ Class 2 ได้ค่า P(Happiness|X1=High,X2=Low,X3=Mid,X4=Mid) ออกมาเยอะที่สุดดังนั้นประเทศจีนจึงถูกจัดอยู่ใน กลุ่ม 2 ตามการคำนวณแบบ Bayes Classification

3.3.5 อภิปรายว่า ค่าความน่าจะเป็นของกลุ่ม P(θ) และ ค่าความน่าจะเป็นของปัจจัยเมื่อทราบกลุ่ม (P(X | θ)) ส่งผลต่อการจำแนกกลุ่มอย่างไร

ค่าความน่าจะเป็นของกลุ่ม เมื่อทราบปัจจัย P($\theta|X$) เป็นค่าที่ใช้ในการเปรียบเทียบความน่าจะเป็นในแต่ละกลุ่มว่าประเทศต่างๆควรจะอยู่ในกลุ่มไหนซึ่งเป็นวิธีการของ Bays Classification โดยที่ค่าที่ส่งผลต่อค่าความน่าจะเป็นของกลุ่มเมื่อทราบปัจจัย P($\theta|X$) จะประกอบไปด้วยค่าความน่าจะเป็นของแต่ละปัจจัยเมื่อทราบกลุ่มความสุข P($X|\theta$) (likelihood function) และค่าความน่าจะเป็นของกลุ่ม P(θ) (Prior distribution) โดยที่ค่าความน่าจะเป็นของกลุ่ม P(θ) (Prior distribution) ได้จากปริมาณข้อมูลที่จะนำมาสร้างโมเดล Bays Classification ดังนั้นยิ่งมีข้อมูลกลุ่มใดกลุ่มหนึ่งอยู่มากเกินไปจะส่งผลให้กลุ่มนั้นมีความน่าจะเป็นของกลุ่มเมื่อทราบปัจจัย P($\theta|X$) สูง ดังนั้นในตัวข้อมูลที่จะนำมาสร้างโมเดลควรมีตัวอย่างข้อมูลแต่ละประเทศที่อยู่ในแต่ละกลุ่มในปริมาณที่ใกล้เคียงกันและค่าความน่าจะเป็นของแต่ละปัจจัยเมื่อทราบกลุ่มความสุข P($X|\theta$) (likelihood function) ก็เช่นกันหากมีการให้ตัวอย่างข้อมูลฝั่งไหนมากเกินไป ในแต่ละปัจจัยหรือในแต่ละกลุ่ม อาจะทำให้ค่าความน่าจะเป็นของแต่ละปัจจัย P($X|\theta$) (likelihood function) อยู่ในฝั่งๆนึงมากเกินไป โดยอาจขึ้นอยู่กับการแบ่งช่วงของข้อมูลด้วยว่าการแบ่งช่วงข้อมูลสมเหตุสมผลหรือไม่ หรือมีการจัดช่วงให้แต่ละปัจจัยข้อมูลพอๆกันหรือไม่ จึงสมารถกล่าวได้ว่า ค่าความน่าจะเป็นของแต่ละปัจจัยเมื่อทราบกลุ่มความสุข P($X|\theta$) (likelihood function) และค่าความน่าจะเป็นของแต่ละกลุ่ม P(θ) (Prior distribution) ส่งผลต่อการทำนายของโมเดล Bays Classification ในรูปของค่าความน่าจะเป็นของกลุ่มเมื่อทราบปัจจัย P($\theta|X$)

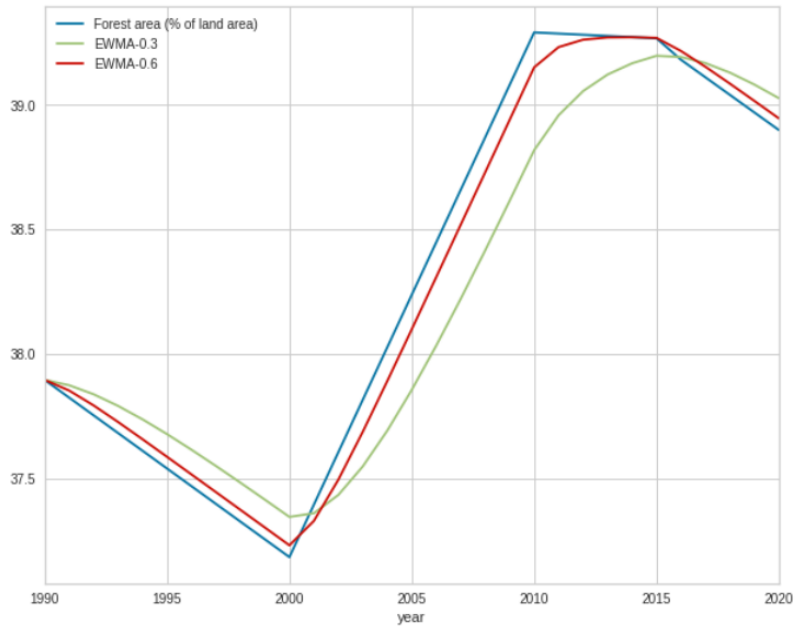
Part 4: Time-Series Analysis

4.1 จาก Assignment 1 เลือกข้อมูล time series มา 2 ตัว โดย 1 ในนั้นเป็นของประเทศไทย (จะเลือกด้านเดียวกัน 2 ประเทศ หรือ เลือก 2 ด้านแต่เป็นข้อมูลประเทศไทยทั้งหมดก็ได้) นำมาวิเคราะห์องค์ประกอบ (decomposition) ของ time series ที่เลือกมา พยายามเลือกข้อมูลตัวที่มีจำนวนปีข้อมูลต่อเนื่อง
วิเคราะห์ผลลัพธ์ และแสดงกราฟที่แสดงข้อมูล time series ที่เลือกมาด้วย

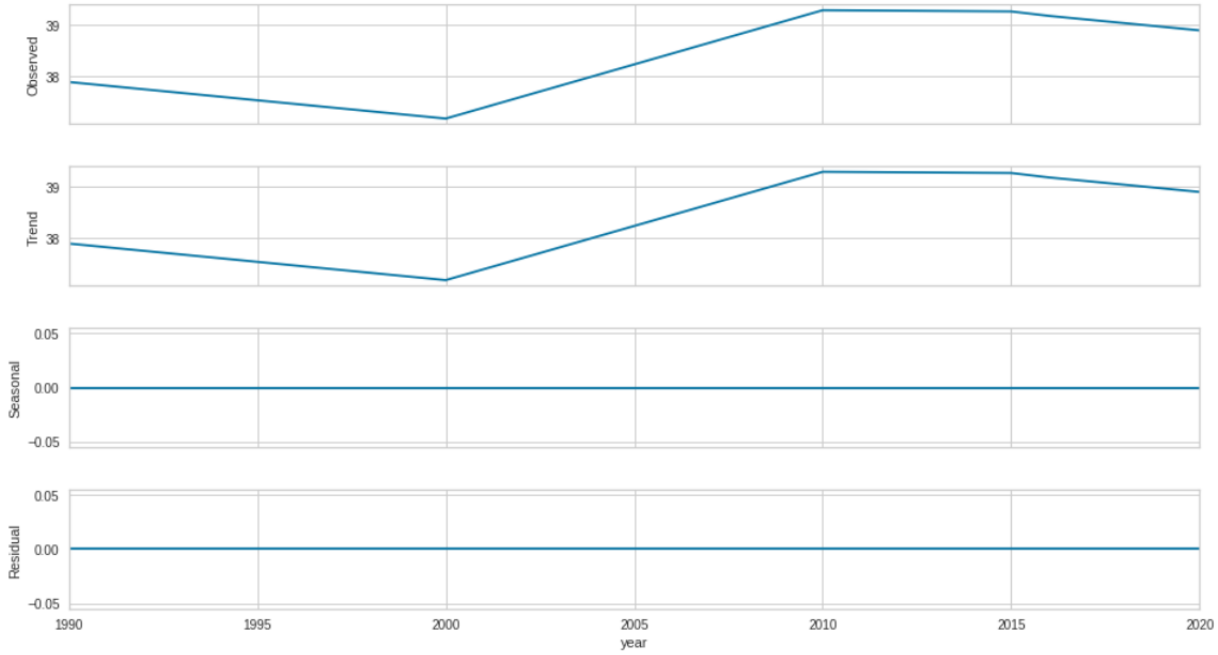
จากข้อมูลของ assignment 1 ได้ทำการเลือกข้อมูล forest land (พื้นที่ป่าไม้) ของประเทศไทยและประเทศสิงคโปร์ โดยข้อมูลตั้งแต่ปี ค.ศ. 1990-2020 (ข้อมูลต่อเนื่อง)

ประเทศไทย

แสดงกราฟที่แสดงข้อมูล forest land (พื้นที่ป่าไม้) ของประเทศไทยได้ผลลัพธ์ time series ที่ผ่านการทำ Exponential Smoothing (ทำการ smooth ผ่าน function ewm)ดังนี้



วิเคราะห์องค์ประกอบ (decomposition) ได้ผลลัพธ์ดังนี้



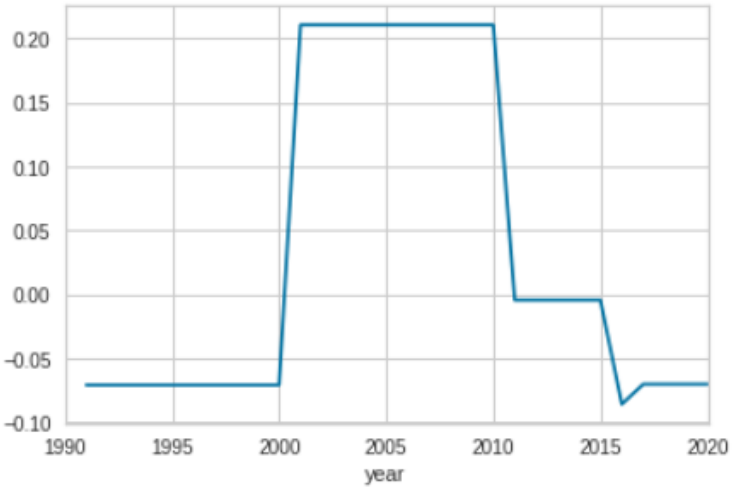
จะเห็นได้ว่าข้อมูล forest land (พื้นที่ป่าไม้) ของประเทศไทย มี trend ที่เพิ่มขึ้นเรื่อยๆ และ ไม่มี Season และ noise ในข้อมูล

4.2 นำข้อมูล timeseries จาก 4.1 ไปทำการทำนาย Arima Model และวิเคราะห์ผลลัพธ์ที่ได้

นำข้อมูล forest land (พื้นที่ป่าไม้) ของประเทศไทย มาทำการ Check stationary ผลลัพธ์ในรอบแรกว่า ไม่เป็น stationary

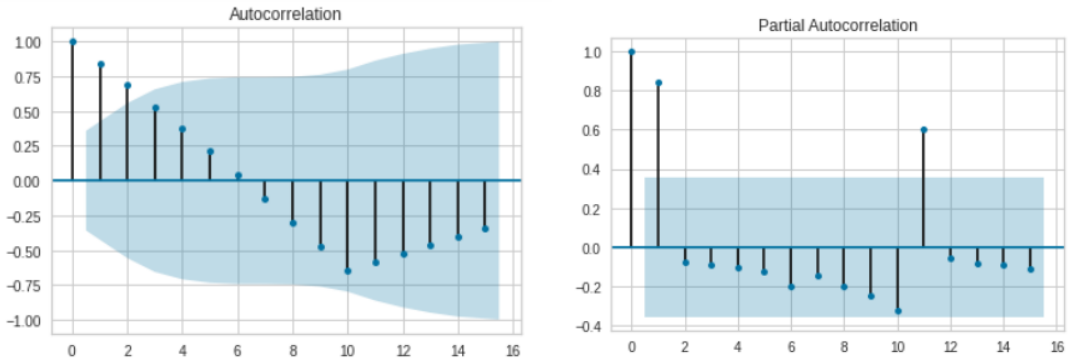
```
Augmented Dickey-Fuller Test:
ADF Test Statistic : -2.024548331747382
p-value : 0.27591430459333255
#Lags Used : 1
Number of Observations Used : 29
Do not reject the null hypothesis. Data is not stationary
```

จึงได้ทำการเอา time series มาลบกับ time series ที่ shift ไป 1 ได้ผลลัพธ์ออกมาเป็น first difference แล้วจึงนำมา Check stationary อีกรอบ ได้ผลลัพธ์ว่าเป็น stationary




```
Augmented Dickey-Fuller Test:
ADF Test Statistic : -12.261008331649672
p-value : 9.07940110652488e-23
#Lags Used : 9
Number of Observations Used : 20
Reject the null hypothesis. Data is stationary
```

เนื่องจากทำ difference ไป 1 รอบแล้วได้ข้อมูลที่เป็น stationary จึงแทนค่า d=1



ต่อมาทำการหาค่า p ผ่านการทำ Partial Autocorrelation และหาค่า q ผ่านการทำ Autocorrelation ได้ว่า ค่า p = 1 และ ค่า q = 1 และเนื่องจากข้อมูลนี้ไม่มี seasonal จึงไม่ได้ทำการเช็ค seasonal และนำข้อมูลมา ทำนาย Arima Model ผ่าน stat model ได้ผลลัพธ์ ดังนี้

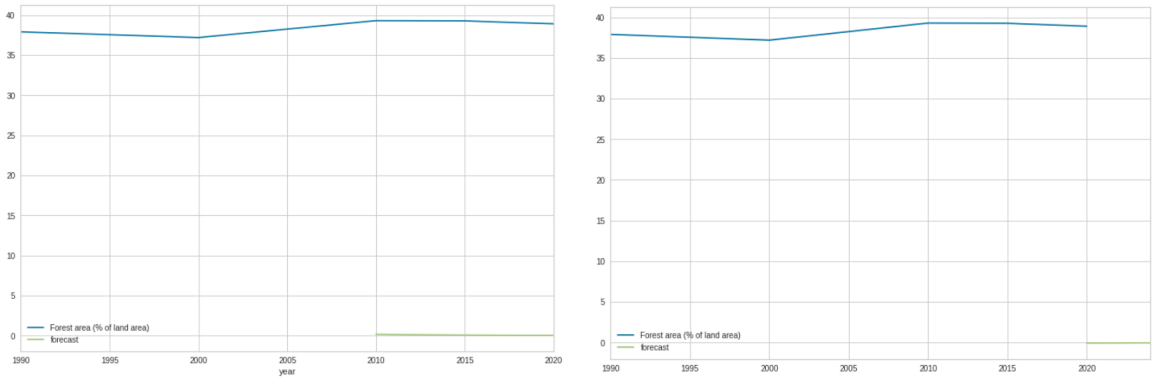
```
                             ARIMA Model Results
==============================================================================================
Dep. Variable:     D.Forest area (% of land area)   No. Observations:                   30
Model:                         ARIMA(1, 1, 1)   Log Likelihood                  38.986
Method:                               css-mle   S.D. of innovations              0.065
Date:                        Thu, 07 Apr 2022   AIC                            -69.972
Time:                                10:43:56   BIC                            -64.367
Sample:                              01-01-1991   HQIC                           -68.179
                                   - 01-01-2020
==============================================================================================
                                        coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------------
const                                 0.0065      0.068      0.095      0.925      -0.127       0.140
ar.L1.D.Forest area (% of land area)     0.8380      0.105      7.961      0.000       0.632       1.044
ma.L1.D.Forest area (% of land area)     0.0580      0.197      0.294      0.771      -0.328       0.444
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.1933           +0.0000j            1.1933            0.0000
MA.1          -17.2408           +0.0000j           17.2408            0.5000
-----------------------------------------------------------------------------
```

และได้ลองใส่ค่า p,d,q หลายๆค่าเพื่อทดสอบค่า AIC หาผลลัพธ์ของ model ที่ให้ค่า AIC น้อยที่สุดมาทำการทำนาย

```
from statsmodels.tsa.arima_model import ARIMAResults
model = ARIMA(df['Forest area (% of land area)'].dropna(),order=(p,d,q))
result = model.fit()
print(result2.aic)
model2 = ARIMA(df['Forest area (% of land area)'].dropna(),order=(0,d,0))
result2 = model2.fit()
print(result2.aic)
model3 = ARIMA(df['Forest area (% of land area)'].dropna(),order=(1,d,1))
result3  =model3.fit()
print(result3.aic)
model4 = ARIMA(df['Forest area (% of land area)'].dropna(),order=(0,d,1))
result4  =model4.fit()
print(result4.aic)
model5 = ARIMA(df['Forest area (% of land area)'].dropna(),order=(1,d,0))
result5  =model5.fit()
print(result5.aic)
```

```
-69.97221301725813
-34.47516846281063
-69.97221301725813
-54.39848186793586
-71.88535947825716
```

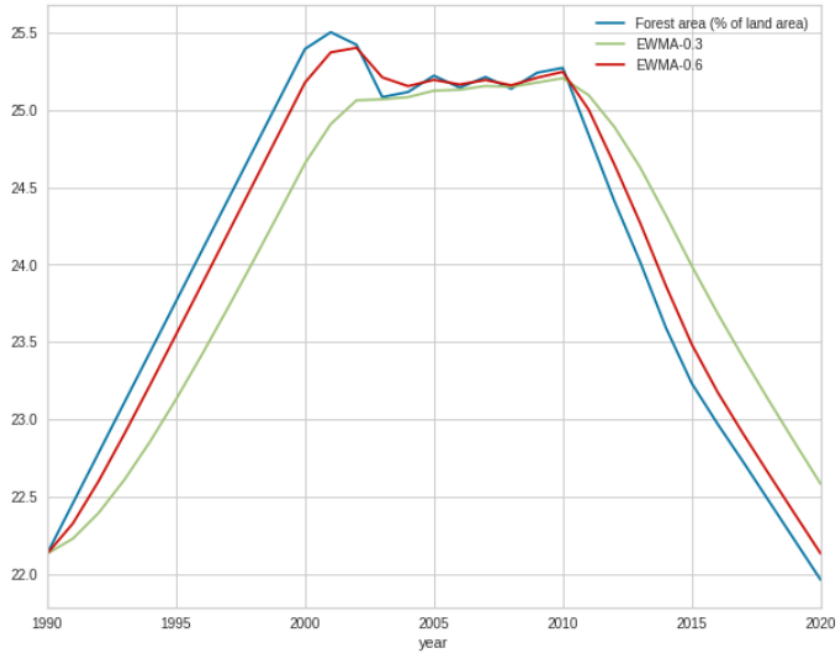
ได้ผลลัพธ์ว่าโมเดล 5 ที่ให้ค่า AIC น้อยที่สุดคือ q=0 และ p=1 จึงได้เอา โมเดล 5 มาทำนายข้อมูล ได้ผลลัพธ์ดังนี้ (สีเขียว)



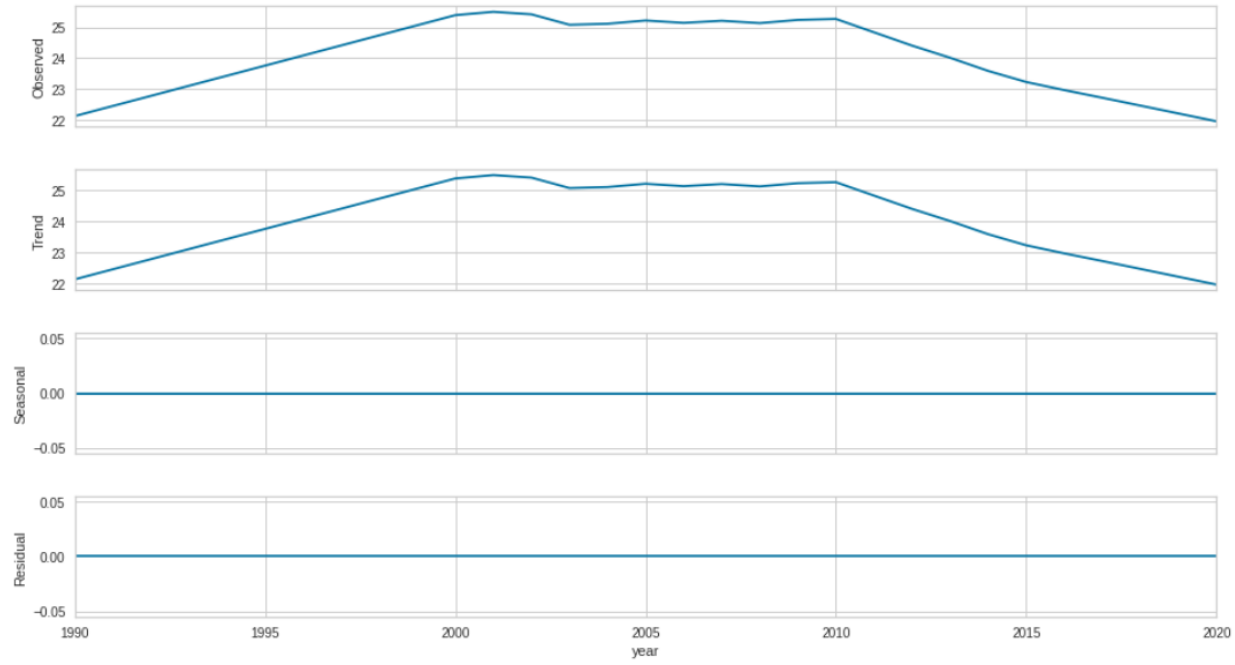
จะเห็นได้ว่าผลลัพธ์ที่ได้ไม่ตรงกับความเป็นจริงเท่าไหร่นักเนื่องจากเป็นการเอาข้อมูลมาทำนายข้อมูลด้วยข้อมูลเดิมจึงไม่ส่งผลที่ดีเท่าไหร่นัก

ประเทศสิงคโปร์

แสดงกราฟที่แสดงข้อมูล forest land (พื้นที่ป่าไม้) ของประเทศสิงคโปร์ได้ผลลัพธ์ time series ที่ผ่านการทำ Exponential Smoothing (ทำการ smooth ผ่าน function ewm) ดังนี้



วิเคราะห์องค์ประกอบ (decomposition) ได้ผลลัพธ์ดังนี้



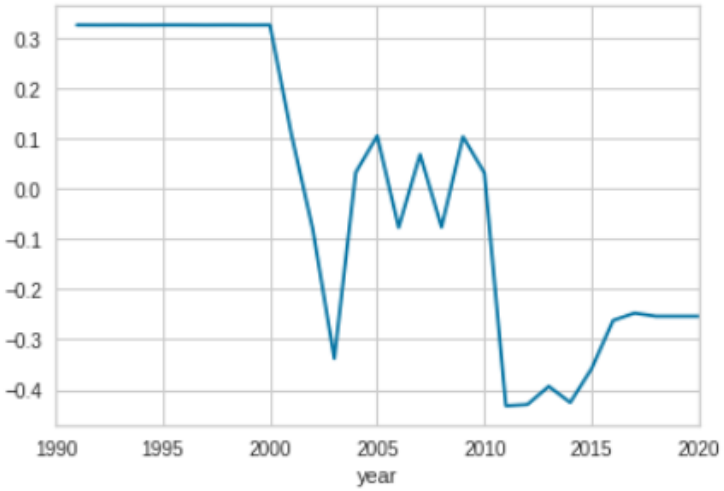
จะเห็นได้ว่าข้อมูล forest land (พื้นที่ป่าไม้) ของประเทศสิงคโปร์ มี trend ทั้งช่วงขาขึ้นและขาลง และ ไม่มี Season และ noise ในข้อมูล

4.2 นำข้อมูล timeseries จาก 4.1 ไปทำการทำนาย Arima Model และวิเคราะห์ผลลัพธ์ที่ได้

นำข้อมูล forest land (พื้นที่ป่าไม้) ของประเทศสิงคโปร์ มาทำการ Check stationary ผลลัพธ์ในรอบแรกว่า ไม่เป็น stationary

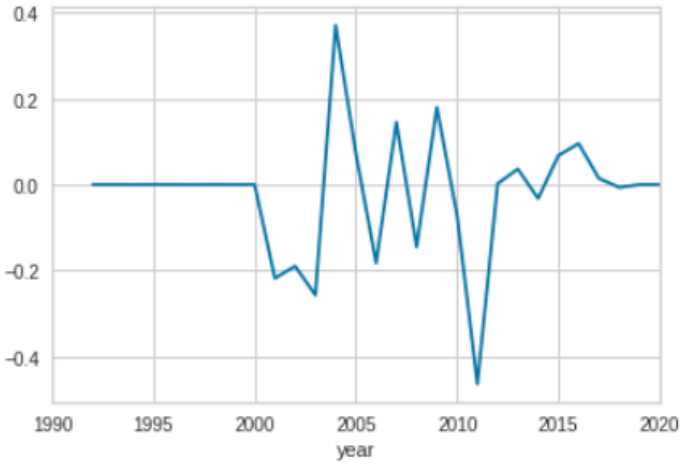
```
Augmented Dickey-Fuller Test:
ADF Test Statistic : -1.2179047873480868
p-value : 0.6658464922644651
#Lags Used : 1
Number of Observations Used : 29
Do not reject the null hypothesis. Data is not stationary
```

จึงได้ทำการเอา time series มาลบกับ time series ที่ shift ไป 1 ได้ผลลัพธ์ออกมาเป็น first difference แล้วจึงนำมา Check stationary อีกรอบ แต่ยัง ได้ผลลัพธ์ว่าไม่เป็น stationary



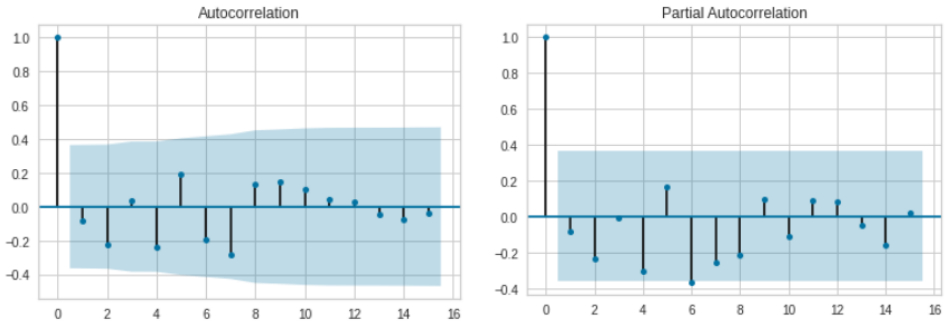

```
Augmented Dickey-Fuller Test:
ADF Test Statistic : -1.5138473655667612
p-value : 0.526650462906363
#Lags Used : 0
Number of Observations Used : 29
Do not reject the null hypothesis. Data is not stationary
```

จึงได้ทำการเอา First Difference มาลบกับ First Difference ที่ shift ไป 1 ได้ผลลัพธ์ออกมาเป็น Second Difference แล้วจึงนำมา Check stationary อีกรอบ ซึ่งรอบนี้ได้ผลลัพธ์ว่าเป็น stationary




```
Augmented Dickey-Fuller Test:
ADF Test Statistic : -3.7398321930623193
p-value : 0.0035865132998067653
#Lags Used : 9
Number of Observations Used : 19
Reject the null hypothesis. Data is stationary
```

เนื่องจากทำ difference ไป 2 รอบแล้วได้ข้อมูลที่เป็น stationary จึงแทนค่า d=2



ต่อมาทำการหาค่า p ผ่านการทำ Partial Autocorrelation และหาค่า q ผ่านการทำ Autocorrelation ได้ว่า ค่า p = 0 และ ค่า q = 0 และเนื่องจากข้อมูลนี้ไม่มี seasonal จึงไม่ได้ทำการเช็ค seasonal และนำข้อมูลมา ทำนาย Arima Model ผ่าน stat model ได้ผลลัพธ์ ดังนี้

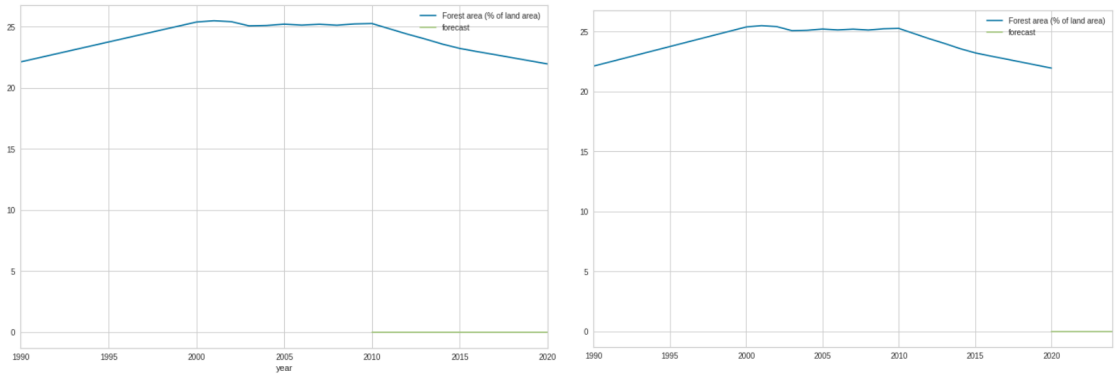
```
                                ARIMA Model Results
==============================================================================================
Dep. Variable:     D2.Forest area (% of land area)   No. Observations:                   29
Model:                             ARIMA(0, 2, 0)   Log Likelihood                14.483
Method:                                       css   S.D. of innovations             0.147
Date:                             Thu, 07 Apr 2022   AIC                          -24.967
Time:                                     10:51:23   BIC                          -22.232
Sample:                                 01-01-1992   HQIC                         -24.110
                                      - 01-01-2020
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0200      0.027     -0.734      0.469      -0.073       0.033
==============================================================================
```

และได้ลองใส่ค่า p,d,q หลายๆค่าเพื่อทดสอบค่า AIC หาผลลัพธ์ของ model ที่ให้ค่า AIC น้อยที่สุดมาทำการทำนาย

```
from statsmodels.tsa.arima_model import ARIMAResults
model = ARIMA(df['Forest area (% of land area)'].dropna(),order=(p,d,q))
result = model.fit()
print(result2.aic)
model2 = ARIMA(df['Forest area (% of land area)'].dropna(),order=(0,d,0))
result2 = model2.fit()
print(result2.aic)
model3 = ARIMA(df['Forest area (% of land area)'].dropna(),order=(0,d,1))
result3  =model3.fit()
print(result3.aic)
model4 = ARIMA(df['Forest area (% of land area)'].dropna(),order=(0,d,1))
result4  =model4.fit()
print(result4.aic)
model5 = ARIMA(df['Forest area (% of land area)'].dropna(),order=(1,d,0))
result5  =model5.fit()
print(result5.aic)
```

```
-23.337762416657355
-24.96687789779282
-23.337762416657355
-23.337762416657355
-23.171309854652534
```

ได้ผลลัพธ์ว่าโมเดล 2 ที่ให้ค่า AIC น้อยที่สุดคือ q=0 และ p=0 จึงได้เอา โมเดล 2 มาทำนายข้อมูลได้ผลลัพธ์ดังนี้ (สีเขียว)



จะเห็นได้ว่าผลลัพธ์ที่ได้ไม่ตรงกับความเป็นจริงเท่าไหร่นักเนื่องจากเป็นการเอาข้อมูลมาทำนายข้อมูลด้วยข้อมูลเดิมจึงไม่ส่งผลที่ดีเท่าไหร่นัก

Part 5: Summary

นำผลที่ได้จาก 1-4 มา สรุป โครงสร้างจังหวัดในประเทศไทย การจัดคลาสประเทศในโลก และการวิเคราะห์ข้อมูล timeseries

Part 1 มีการทำ feature selection หลายๆ วิธีผ่านข้อมูล 30 ปัจจัย โดยเป็นการหาปัจจัยที่ทำผู้จัดทำคิดว่าเกี่ยวข้องมา 11 ปัจจัย โดย 5 ปัจจัยเป็นข้อมูลระดับประเทศ และอีก 6 ข้อมูลเป็นข้อมูลระดับจังหวัด เพื่อหาว่ามีปัจจัยใดบ้างที่เกี่ยวข้องกับ GPP per capita ในสามจังหวัดที่ได้รับมอบหมายคือ จังหวัด ชลบุรี , ปราจีนบุรี และ สมุทรสงคราม ซึ่งในผลลัพธ์ได้สะท้อนให้เห็นว่าแต่ละจังหวัดก็มีปัจจัยที่ส่งผลต่อ GPP per capita ต่างกันออกไปอาจเพราะเหตุผลขึ้นอยู่กับจำนวนประชากร , ขนาดจังหวัด , รายได้ของจังหวัดนั้นๆ ด้วย

Part 2 มีการทำ Clustering โดยเปรียบเมื่อใช้ทุกปัจจัยกับใช้บางปัจจัยที่ส่งผล(จาก Part 1)พบ ว่าผลลัพธ์ของ cluster ต่างกันคือมี 5 กลุ่ม และ 4 กลุ่มตามลำดับ กล่าวคือการใช้ปัจจัยที่ส่งผลจะช่วยให้เห็นกลุ่มได้ชัดเจนขึ้น โดย เมื่อใช้ทุกปัจจัย จะเห็นได้ว่ามี 3 กลุ่ม ที่มีจำนวน 24 จังหวัด โดยจะมีค่าของปัจจัยอยู่ประมาณ กลางถึงต่ำ เมื่อเทียบกับการใช้ บางปัจจัย จะเห็นได้ว่า มี 1 กลุ่มที่มีจำนวน 47 จังหวัด ซึ่งมีค่าของปัจจัยอยู่ประมาณ กลางถึงต่ำ เช่นเดียวกัน แต่ในทั้งสองครั้งของการทำ Clustering จะเห็นได้ว่ามี 1 กลุ่มที่แยกออกมาและมีจำนวน 1 จังหวัด คือ กรุงเทพมหานคร ซึ่งมีค่าของปัจจัยสูง อาจสามารถกล่าวได้ว่ามีความเหลื่อมล้ำ ของจังหวัดเกิดขึ้นจึงได้เกิดการแบ่งกลุ่มเช่นนี้ และยังมีเหตุผลสนับสนุนอีกคือ ในทั้งสองครั้งของการทำ Clustering จะเห็นได้ว่ามีอีก 1 กลุ่มที่จำแนกออกมาและมีสามาชิกอยู่ประมาณ 4 - 5 จังหวัด ซึ่งมีค่าของปัจจัยอยู่ประมาณกลางถึงสูง อาจกล่าวได้ว่ามีจังหวัดที่มีความเจริญใกล้เคียงกับกรุงเทพมหานคร แต่มีจำนวนน้อย เช่น จังหวัด ชลบุรี เป็นต้น

ใน Part 3.1 มีการทำ Classification โดยใช้ Decision tree โดยโมเดลที่ใช้ทุกปัจจัย และ ใช้ปัจจัยจาก feature selection และใช้ปัจจัยจริงมีค่า Accuracy , (macro) Precision , (macro) Recall , (macro) F-score ที่ใกล้เคียงกัน โดย ประเทศที่มีการจำแนกผิดพลาด สาเหตุเนื่องมาจากวิเคราะผลลัพธ์ของโมเดลที่ได้ ดูจากค่า precision , recall , f1-score เป็นค่าเอาไว้ประเมินโมเดล (precision เอาไว้บอกว่าค่าที่ทำนายมาแล้วค่าที่ทำนายถูก ตรงกับข้อมูลจริง , recall เป็นค่าที่เอาไว้บอกว่าค่าที่ทำนายมาเป็น positive เป็นค่าข้อมูลที่ positive จริงๆ เท่าไหร่ , f1-score เป็นค่าการถ่วงน้ำหนักระหว่าง precision และ recall) เมื่อดูจากค่า f1-score จะเห็นได้ว่าข้อมูลเป็น imbalance data จึงมีการตอบข้อมูลที่เป็น class 3 มามากกว่า class อื่นๆ โดยในแต่ละคลาส class 0,1,2,3 คือ class ที่มี ลักษณะ human development index จำแนกตาม Class นั้นๆ และประเทศไทย มี Class จริงอยู่ที่ Class 0 และได้ Class ทำนายว่าอยู่ใน Class 0 เช่นเดียวกัน อาจบ่งบอกได้ว่าประเทศไทย มี human development index ต่ำนั้นเอง

ใน Part 3.2 มีการทำ Logistics Regression โดยหลังจากพิจารณาผลลัพธ์ที่ค่า p-value ต่ำกว่า 0.3 ปัจจัยที่คาดว่าจะส่งผลมีดังนี้ :

- Mortality_infant_rate (อัตราการตายของทารก)
- Life_expectancy (ช่วงอายุของชีวิต)

- Basic_drinking_water (คุณภาพการเข้าถึงน้ำสะอาด)
- Forest_land (พื้นที่ป่าไม้)
- Senior (ประชากรผู้สูงอายุ)

จากปัจจัยกลุ่มนี้ สะท้อนให้เห็นคุณสมบัติของประเทศที่พัฒนาแล้ว เป็นเหมือนการมองภาพรวมที่เมื่ออยู่ในประเทศที่พัฒนาควรจะมีความสามารถด้านการแพทย์ให้เด็กทารกสามารถคลอดออกมาแล้วมีชีวิตได้ปกติ ไม่พิการ และได้เติบโต , ใช้ชีวิตผ่านพื้นที่ธรรมชาติเพื่อสภาวะแวดล้อมที่น่าอยู่และสามารถเข้าถึงน้ำสะอาดได้โดยง่าย โดยการใช้ชีวิตก็จะต้องใช้ชีวิตที่ยืนยาว ดังนั้นประชากรผู้สูงอายุจึงเป็นตัวบ่งบอกว่าประเทศนี้เหมาะแก่การใช้ชีวิต ไม่เจ็บป่วยหรือตายก่อนที่จะถึงวัยสูงอายุนั้นเอง

ใน Part 3.3 มีการทำ Bayes Classification ได้ผลลัพธ์ของการทำนายประเทศไทยออกมาว่า ค่า P(Happiness|X1=Mid,X2=Low,X3=Mid,X4=High) เป็นผลลัพธ์จากการคูณกันของแต่ล่ะ row ซึ่งจะเป็นได้ว่า row ของ Class 3 ได้ค่า P(Happiness|X1=Mid,X2=Low,X3=Mid,X4=High) ออกมาเยอะที่สุด ดังนั้นประเทศไทยจึงถูกจัดอยู่ใน กลุ่ม 3 ตามการคำนวณแบบ Bayes Classification แต่ในความเป็นจริงประเทศไทยอยู่ใน กลุ่ม 2 ดังนั้น จึงสมารถกล่าวได้ว่า ค่าความน่าจะเป็นของแต่ล่ะปัจจัยเมื่อทราบกลุ่มความสุข P($X|\theta$) (likelihood function) และค่าความน่าจะเป็นของแต่ล่ะกลุ่ม P(θ) (Prior distribution) ส่งผลต่อการทำนายของโมเดล Bays Classification ในรูปของค่าความน่าจะเป็นของกลุ่มเมื่อทราบปัจจัย P($\theta|X$) จึงอาจสามารถทำนายผิดได้หากมีข้อมูลของปัจจัยไหนส่งผลต่อค่าความน่าจะเป็นเยอะจนเกินไป

ใน Part 4 มีการทำ Time-Series Analysis โดยจะสามารถมองเห็นแนวโน้มผ่านข้อมูลนั้นๆ เช่น การดูข้อมูลของประเทศไทยจาก trend และคาดเดาอนาคตว่าจะมีทิศทางไปไหนทางไหน โดยใน part 4 จะเป็นการทำเปรียบเทียบข้อมูล forest land (พื้นที่ป่าไม้) ของประเทศไทยและประเทศสิงคโปร์ โดยในผลลัพธ์ของการทำนายจะเห็นได้ว่าได้ค่าการทำนายที่ไม่ค่อยสมจริงเท่าไหร่นักเนื่องจากเป็นการเอาข้อมูลมาทำนายข้อมูลด้วยข้อมูลเดิมจึงไม่ส่งผลที่ดีเท่าไหร่นัก แต่ก็ยังสามารถเห็น trend forest land (พื้นที่ป่าไม้) ของประเทศไทยและประเทศสิงคโปร์ ว่าประเทศไทยมีแนวโม้ม ในขาขึ้นหรือก็คือมีพื้นที่ป่าไม้เพิ่มขึ้นเรื่อยๆ แต่ ประเทศสิงคโปร์เริ่มจะมีแนวโม้ม ในขาลง หรือก็คือเริ่มมีพื้นที่ป่าไม้ลดลงในอนาคตนั้นเอง