Assignment 2

Assigned Aspect : ด้านการศาสนา วัฒนธรรมและนันนทาการ
First Ministry : กระทรวงศึกษาธิการ
Second Ministry : สำนักนายกรัฐมนตรี
Third Ministry : กระทรวงวัฒนธรรม

โดย
นาย ปพนธ์ ชุณหคล้าย 6210503691
นาย จิณณเจตน์ อจลพงศ์ 6210505163

เสนอ
ผศ.ดร.สุภาพร เอื้อจงมานี

รายงานฉบับนี้เป็นส่วนหนึ่งของรายวิชา
สถิติสำหรับการประยุกต์ทางวิศวกรรมคอมพิวเตอร์(01204314)
ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์
มหาวิทยาลัยเกษตรศาสตร์
ปีการศึกษา 2564 ภาคปลาย

สารบัญ

Clustering Part

วัตถุประสงค์เพื่อจัดกลุ่มกระทรวงตามงบประมาณรายจ่าย

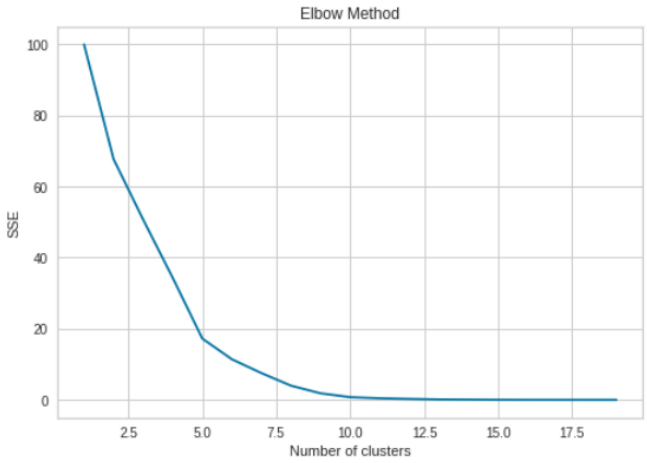
Part 1

ใช้ชุดข้อมูล งบประมาณรายจ่ายส่วนย่อย 5 ด้าน ในปี พ.ศ.2564 (ที่ผ่านการทำ normalized แล้ว) เพื่อจัดกลุ่มกระทรวงตามงบประมาณรายจ่าย(การทำ normalized คือการทำให้ข้อมูลอยู่ในrangeเดียวกัน)

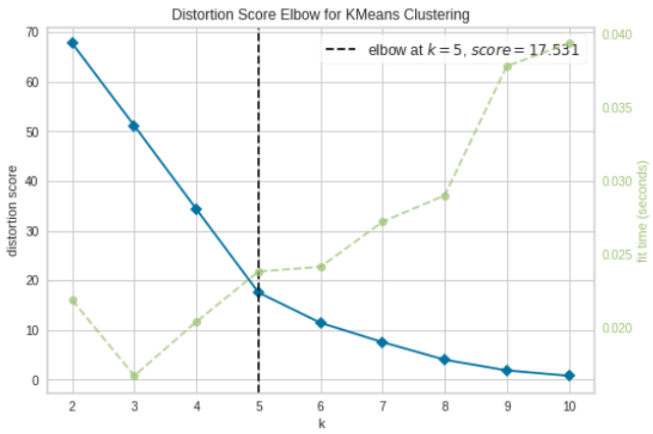
1.Clustering ด้วยวิธี K-mean, Hierarchical และ DBScan ในแต่ละวิธี ให้เหตุผลในการเลือก ค่า parameter ในการทำ clustering และ ระบุว่า วิธีใด ให้ผลลัพธ์ Clustering ที่ดีที่สุด อภิปรายเหตุผล ว่าเพราะเหตุใดจึงเป็นเช่นนั้น

Clustering ด้วยวิธี K-mean

นำข้อมูลที่ผ่านการทำ 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
ผลลัพธ์แสดงออกมาได้ดังนี้

Cluster_id	จำนวนข้อมูลที่อยู่ใน cluster นั้นๆ
0	2
1	13
2	3
3	1
4	1

โดยค่าเฉลี่ยของงบในแต่ละ Cluster มีดังนี้

cluster_id	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น
0	17040.650000	6966.000000	124650.950000	1029.100000	658.200000
1	4285.000000	2419.238462	2954.292308	15838.707692	2189.023077
2	140591.666667	19514.600000	16962.866667	33353.600000	28504.800000
3	12988.200000	6567.900000	3421.200000	599.400000	244863.500000
4	19360.200000	12718.100000	38441.400000	235503.600000	28464.400000

Performance measurement ได้วัดผ่านตัวชี้วัด 2 ตัวคือ

1.Silhouette index จากสูตรการคำนวณ $\frac{b-a}{\max(a,b)}$ (a คือค่าเฉลี่ยของระยะทางของ intra-cluster และ b คือค่าเฉลี่ยระยะทางของ inter-cluster)โดย ค่าจะดีที่สุดที่ 1 และแย่ที่สุดที่ -1

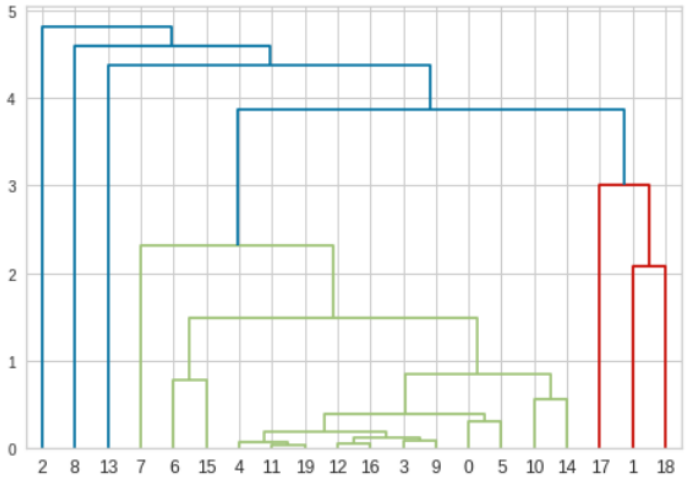
ค่า Silhouette index ของการ clustering ด้วยวิธี k-mean จะอยู่ที่ 0.5172476961349493

2.Davies-Bouldin index จากสูตรการคำนวณ $\frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j}\{\frac{\Delta x_i+\Delta x_j}{\delta(x_i,x_j)}\}$ โดยค่าที่ได้นั้นยิ่งได้ค่าน้อยจะบ่งบอกถึงโมเดลที่ดี โดยค่าที่น้อยที่สุดจะอยู่ที่ 0

ค่า Davies-Bouldin index ของการ clustering ด้วยวิธี k-mean จะอยู่ที่ 0.5723699449579567

Clustering ด้วยวิธี Hierarchical

นำข้อมูลที่ผ่านการทำ normalized มาทำ Agglomerative clustering เพื่อนำไป plot ใน dendrogram และทางผู้จัดทำได้พิจารณาจาก dendrogram ได้กำหนดให้ threshold value = 3 และนำไปทำ Agglomerative clustering อีกรอบเพื่อให้ได้ผลลัพธ์



ผลลัพธ์แสดงออกมาได้ดังนี้

Cluster_id	จำนวนข้อมูลที่อยู่ใน cluster นั้นๆ
0	14
1	2
2	1
3	1
4	1
5	1

โดยค่าเฉลี่ยของงบในแต่ล่ะ Cluster มีดังนี้

cluster_id	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น
0	5720.114286	3067.564286	7907.042857	14850.307143	2073.007143
1	99287.700000	19257.150000	15725.800000	5715.950000	40179.850000
2	223199.600000	20029.500000	19437.000000	88628.900000	5154.700000
3	12988.200000	6567.900000	3421.200000	599.400000	244863.500000
4	19360.200000	12718.100000	38441.400000	235503.600000	28464.400000
5	9704.700000	2436.200000	177009.100000	57.100000	751.600000

Performance measurement ได้วัดผ่านตัวชี้วัด 2 ตัวคือ

1.Silhouette index จากสูตรการคำนวณ $\frac{b-a}{\max(a,b)}$ (a คือค่าเฉลี่ยของระยะทางของ intra-cluster และ b คือค่าเฉลี่ยระยะทางของ inter-cluster)โดย ค่าจะดีที่สุดที่ 1 และแย่ที่สุดที่ -1

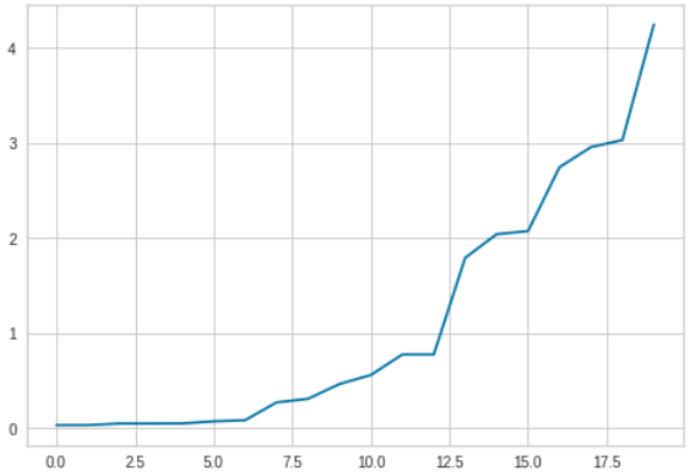
ค่า Silhouette index ของการ clustering ด้วยวิธี Hierarchical จะอยู่ที่ 0.5043462069124952

2.Davies-Bouldin index จากสูตรการคำนวณ $\frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j}\{\frac{\Delta x_i+\Delta x_j}{\delta(x_i,x_j)}\}$ โดยค่าที่ได้นั้นยิ่งได้ค่าน้อยจะบ่งบอกถึงโมเดลที่ดี โดยค่าที่น้อยที่สุดจะอยู่ที่ 0

ค่า Davies-Bouldin index ของการ clustering ด้วยวิธี Hierarchical จะอยู่ที่ 0.359135032441

Clustering ด้วยวิธี DBSCAN

นำข้อมูลที่ผ่านการทำ normalized มาทำการหาระยะทางของ point ตั้งต้นและอีก 2 point ข้างๆ ระยะทางเป็นเท่าไหร่ (ให้ค่า parameter n_neighbors=2) ก่อนจะนำมา sort ระยะทางและนำไป plot graph



โดยได้พิจารณาจากใน graph และได้ระบุค่า epsilon = 2.1 และค่า Minimum sample value = 10 (อ้างอิงจาก (Sander et al., 1998) ได้พูดถึงวิธีการเลือก Minimum sample value ให้เท่ากับ 2 เท่าของ dimensions ของ data set ซึ่ง data set ที่ใช้มี dimensions = 5) และกำหนดให้ค่า noise = -1

ผลลัพธ์แสดงออกมาได้ดังนี้

Cluster_id	จำนวนข้อมูลที่อยู่ใน cluster นั้นๆ
-1	5
0	15

โดยค่าเฉลี่ยของงบในแต่ล่ะ Cluster มีดังนี้

cluster_id	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น
-1	71256.02	13170.92	50546.96	65916.680000	71882.8
0	12508.64	3823.82	8514.94	14302.786667	1946.8

Performance measurement ได้วัดผ่านตัวชี้วัด 2 ตัวคือ

1.Silhouette index จากสูตรการคำนวณ $\frac{b-a}{\max(a,b)}$ (a คือค่าเฉลี่ยของระยะทางของ intra-cluster และ b คือค่าเฉลี่ยระยะทางของ inter-cluster)โดย ค่าจะดีที่สุดที่ 1 และแย่ที่สุดที่ -1

ค่า Silhouette index ของการ clustering ด้วยวิธี DBSCAN จะอยู่ที่ 0.5111494849290473

2.Davies-Bouldin index จากสูตรการคำนวณ $\frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j}\{\frac{\Delta x_i+\Delta x_j}{\delta(x_i,x_j)}\}$ โดยค่าที่ได้นั้นยิ่งได้ค่าน้อยจะบ่งบอกถึงโมเดลที่ดี โดยค่าที่น้อยที่สุดจะอยู่ที่ 0

ค่า Davies-Bouldin index ของการ clustering ด้วยวิธี DBSCAN จะอยู่ที่ 1.69892648889161

สรุป

วิธีที่ให้ผลลัพธ์ดีที่สุดคือวิธี Hierarchical โดยพิจารณาจากการดู index ทั้ง 2 ตัว ดังนี้

	Silhouette index	Davies-Bouldin index
K-mean	0.5172476961349493	0.5723699449579567
Hierarchical	0.5043462069124952	0.359135032441
DBSCAN	0.5111494849290473	1.69892648889161

พิจารณาที่ Silhouette index จากสูตรการคำนวณ $\frac{b-a}{\max(a,b)}$ (a คือค่าเฉลี่ยของระยะทางของ intra-cluster และ b คือค่าเฉลี่ยระยะทางของ inter-cluster)โดย ค่าจะดีที่สุดที่ 1 และแย่ที่สุดที่ -1 โดยค่าที่ได้ในแต่ล่ะรูปแบบของการทำ clustering มีค่าใกล้เคียงกัน จึงต้องพิจารณาที่ Davies-Bouldin index จากสูตรการคำนวณ $\frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j}\{\frac{\Delta x_i+\Delta x_j}{\delta(x_i,x_j)}\}$ โดยค่าที่ได้นั้นยิ่งได้ค่าน้อยจะบ่งบอกถึงโมเดลที่ดี โดยค่าที่น้อยที่สุดจะอยู่ที่ 0 จะเห็นได้ว่า รูปแบบ Hierarchical ได้ค่า Davies-Bouldin index ดีที่สุด กล่าวคือสัดส่วนระยะทางของ intra-cluster และ inter-cluster มีความเหมาะสมจึงพิจารณาได้ว่าวิธี Hierarchical ได้ให้ผลลัพธ์ที่ดีที่สุดนั้นเอง

2.จากการจัดกลุ่มทีได้ของแต่ละวิธี อภิปรายผลลัพธ์ลักษณะรวมของกลุ่มแต่ละกลุ่มที่ได้ เหตุใดการจัดกลุ่มจึงออกมาเป็นเช่นนั้น

ผลลัพธ์ของวิธี K-mean ได้ทำการแบ่ง cluster ออกเป็น 5 กลุ่ม

กลุ่มที่ 1

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
7	กระทรวงเกษตรและสหกรณ์	24376.6	11495.8	72292.8	2001.1	564.8	110731.1	0
8	กระทรวงคมนาคม	9704.7	2436.2	177009.1	57.1	751.6	189958.7	0

กลุ่มที่ 2

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
0	สำนักนายกรัฐมนตรี	4104.5	3078.3	6104.3	12397.0	13606.8	39290.9	1
3	กระทรวงการต่างประเทศ	2697.3	2509.6	397.5	1348.7	1203.3	8156.4	1
4	กระทรวงการท่องเที่ยวและกีฬา	1572.6	892.1	1564.4	735.9	1327.7	6092.7	1
5	กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์	2746.6	2519.3	397.8	16598.3	79.8	22341.8	1
6	กระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัยและนวัตกรรม	11248.4	939.9	9956.4	104384.6	1597.7	128127.0	1
9	กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม	1311.6	2009.9	1243.4	1945.8	1791.0	8301.7	1
10	กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	10992.6	5199.8	10408.8	1103.6	1630.9	29335.7	1
11	กระทรวงพลังงาน	822.0	531.2	707.8	30.7	187.5	2279.2	1
12	กระทรวงพาณิชย์	2216.4	1585.4	845.4	479.5	1698.8	6825.5	1
14	กระทรวงยุติธรรม	10363.6	8775.0	3333.1	1937.6	2418.8	26828.1	1
15	กระทรวงแรงงาน	3671.6	1375.7	375.7	63712.9	584.4	69720.3	1
16	กระทรวงวัฒนธรรม	2321.9	1430.5	2325.4	1070.7	814.4	7962.9	1
19	กระทรวงอุตสาหกรรม	1635.9	603.4	745.8	157.9	1516.2	4659.2	1

กลุ่มที่ 3

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
1	กระทรวงกลาโหม	91027.4	24102.9	14426.1	4794.4	80179.8	214530.6	2
17	กระทรวงศึกษาธิการ	223199.6	20029.5	19437.0	88628.9	5154.7	356449.7	2
18	กระทรวงสาธารณสุข	107548.0	14411.4	17025.5	6637.5	179.9	145802.3	2

กลุ่มที่ 4

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
2	กระทรวงการคลัง	12988.2	6567.9	3421.2	599.4	244863.5	268440.2	3

กลุ่มที่ 5

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
13	กระทรวงมหาดไทย	19360.2	12718.1	38441.4	235503.6	28464.4	334487.7	4

โดยข้อมูลรายจ่ายโดยเฉลี่ยของแต่ละกลุ่มมีดังนี้

cluster_id	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น
0	17040.650000	6966.000000	124650.950000	1029.100000	658.200000
1	4285.000000	2419.238462	2954.292308	15838.707692	2189.023077
2	140591.666667	19514.600000	16962.866667	33353.600000	28504.800000
3	12988.200000	6567.900000	3421.200000	599.400000	244863.500000
4	19360.200000	12718.100000	38441.400000	235503.600000	28464.400000

สรุปผลลัพธ์การทำ clustering ด้วยวิธี k-mean

เนื่องจากค่าที่ได้ในแต่ละกลุ่มไม่สามารถเจาะจงอธิบาลรายละเอียดได้อย่างชัดเจนจึงอธิบายผ่านข้อมูลรายจ่ายโดยเฉลี่ยของแต่ละกลุ่มแทน โดยแต่ละกลุ่มจะมีลักษณะคร่าวๆดังนี้

กลุ่มที่ 1 จะมีงบลงทุนสูง และ มีงบบุคลากรอยู่ในช่วงปานกลางถึงสูง และ มีงบดำเนินงาน,งบเงินอุดหนุน อยู่ในช่วงปานกลางถึงต่ำ และ มีงบรายจ่ายอื่นต่ำ

กลุ่มที่ 2 จะมีงบเงินอุดหนุน อยู่ในช่วงปานกลางถึงสูง และ มีงบดำเนินงาน,งบรายจ่ายอื่น,งบบุคลากร,งบลงทุน อยู่ในช่วงปานกลางถึงต่ำ

กลุ่มที่ 3 จะมีงบบุคลากรสูง และ มีงบลงทุน,งบดำเนินงาน,งบเงินอุดหนุน,งบรายจ่ายอื่น อยู่ในช่วงปานกลางถึงสูง

กลุ่มที่ 4 จะมีงบรายจ่ายอื่นสูง และมีงบบุคลากร อยู่ในช่วงปานกลางถึงสูง และ มีงบดำเนินงาน,งบลงทุน อยู่ในช่วงปานกลางถึงต่ำ และมีงบเงินอุดหนุนต่ำ

กลุ่มที่ 5 จะมีงบเงินอุดหนุนสูง และ มีงบลงทุน,งบดำเนินงาน,งบบุคลากร,งบรายจ่ายอื่น อยู่ในช่วงปานกลางถึงสูง

โดยสาเหตุที่ผลการจัดกลุ่มออกมาเป็นเช่นนี้อาจเป็นเพราะค่ารายจ่ายของแต่ละชุดข้อมูลมีความใกล้เคียงกับค่าเฉลี่ยของแต่ละกลุ่มเลยถูกจัดกลุ่มออกมาในรูปแบบนี้

ผลลัพธ์ของวิธี Hierarchical ได้ทำการแบ่ง cluster ออกเป็น 6 กลุ่ม

กลุ่มที่ 1

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
0	สำนักนายกรัฐมนตรี	4104.5	3078.3	6104.3	12397.0	13606.8	39290.9	0
3	กระทรวงการต่างประเทศ	2697.3	2509.6	397.5	1348.7	1203.3	8156.4	0
4	กระทรวงการท่องเที่ยวและกีฬา	1572.6	892.1	1564.4	735.9	1327.7	6092.7	0
5	กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์	2746.6	2519.3	397.8	16598.3	79.8	22341.8	0
6	กระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัยและนวัตกรรม	11248.4	939.9	9956.4	104384.6	1597.7	128127.0	0
7	กระทรวงเกษตรและสหกรณ์	24376.6	11495.8	72292.8	2001.1	564.8	110731.1	0
9	กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม	1311.6	2009.9	1243.4	1945.8	1791.0	8301.7	0
10	กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	10992.6	5199.8	10408.8	1103.6	1630.9	29335.7	0
11	กระทรวงพลังงาน	822.0	531.2	707.8	30.7	187.5	2279.2	0
12	กระทรวงพาณิชย์	2216.4	1585.4	845.4	479.5	1698.8	6825.5	0
14	กระทรวงยุติธรรม	10363.6	8775.0	3333.1	1937.6	2418.8	26828.1	0
15	กระทรวงแรงงาน	3671.6	1375.7	375.7	63712.9	584.4	69720.3	0
16	กระทรวงวัฒนธรรม	2321.9	1430.5	2325.4	1070.7	814.4	7962.9	0
19	กระทรวงอุตสาหกรรม	1635.9	603.4	745.8	157.9	1516.2	4659.2	0

กลุ่มที่ 2

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
1	กระทรวงกลาโหม	91027.4	24102.9	14426.1	4794.4	80179.8	214530.6	1
18	กระทรวงสาธารณสุข	107548.0	14411.4	17025.5	6637.5	179.9	145802.3	1

กลุ่มที่ 3

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
17	กระทรวงศึกษาธิการ	223199.6	20029.5	19437.0	88628.9	5154.7	356449.7	2

กลุ่มที่ 4

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
2	กระทรวงการคลัง	12988.2	6567.9	3421.2	599.4	244863.5	268440.2	3

กลุ่มที่ 5

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
13	กระทรวงมหาดไทย	19360.2	12718.1	38441.4	235503.6	28464.4	334487.7	4

กลุ่มที่ 6

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
8	กระทรวงคมนาคม	9704.7	2436.2	177009.1	57.1	751.6	189958.7	5

โดยข้อมูลรายจ่ายโดยเฉลี่ยของแต่ละกลุ่มมีดังนี้

cluster_id	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น
0	5720.114286	3067.564286	7907.042857	14850.307143	2073.007143
1	99287.700000	19257.150000	15725.800000	5715.950000	40179.850000
2	223199.600000	20029.500000	19437.000000	88628.900000	5154.700000
3	12988.200000	6567.900000	3421.200000	599.400000	244863.500000
4	19360.200000	12718.100000	38441.400000	235503.600000	28464.400000
5	9704.700000	2436.200000	177009.100000	57.100000	751.600000

สรุปผลลัพธ์การทำ clustering ด้วยวิธี Hierarchical

เนื่องจากค่าที่ได้ในแต่ละกลุ่มไม่สามารถเจาะจงอธิบาลรายละเอียดได้อย่างชัดเจนจึงอธิบายผ่านข้อมูลรายจ่ายโดยเฉลี่ยของแต่ละกลุ่มแทน โดยแต่ละกลุ่มจะมีลักษณะคร่าวๆดังนี้

กลุ่มที่ 1 จะมีงบเงินอุดหนุนอยู่ในช่วงปานกลางถึงสูง และ มีงบดำเนินงาน,งบบุคลากร,งบรายจ่ายอื่น,งบลงทุน อยู่ในช่วงปานกลางถึงต่ำ

กลุ่มที่ 2 จะมี งบดำเนินงาน,งบรายจ่ายอื่น,งบบุคลากร,งบลงทุนอยู่ในช่วงปานกลางถึงสูง และ มีงบเงินอุดหนุนอยู่ในช่วงปานกลางถึงต่ำ

กลุ่มที่ 3 จะมีงบบุคลากรสูง และ มีงบลงทุน,งบดำเนินงาน,งบเงินอุดหนุน อยู่ในช่วงปานกลางถึงสูง และมีงบรายจ่ายอื่น อยู่ในช่วงปานกลางถึงต่ำ

กลุ่มที่ 4 จะมีงบรายจ่ายอื่นสูง และมีงบบุคลากร อยู่ในช่วงปานกลางถึงสูง และ มีงบดำเนินงาน,งบลงทุน อยู่ในช่วงปานกลางถึงต่ำ และมีงบเงินอุดหนุนต่ำ

กลุ่มที่ 5 จะมีงบเงินอุดหนุนสูง และ มีงบลงทุน,งบดำเนินงาน,งบบุคลากร,งบรายจ่ายอื่น อยู่ในช่วงปานกลางถึงสูง

กลุ่มที่ 6 จะมีงบลงทุนสูง และ มีงบดำเนินงาน,งบบุคลากร อยู่ในช่วงปานกลางถึงต่ำ และมีงบรายจ่ายอื่นต่ำ และมีงบเงินอุดหนุนต่ำมาก

โดยสาเหตุที่ผลการจัดกลุ่มออกมาเป็นเช่นนี้อาจเป็นเพราะค่ารายจ่ายของแต่ละชุดข้อมูลมีความใกล้เคียงกับค่าเฉลี่ยของแต่ละกลุ่มเลยถูกจัดกลุ่มออกมาในรูปแบบนี้

ผลลัพธ์ของวิธี DBSCAN ได้ทำการแบ่ง cluster ออกเป็น 2 กลุ่ม

กลุ่มที่ 1

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
1	กระทรวงกลาโหม	91027.4	24102.9	14426.1	4794.4	80179.8	214530.6	-1
2	กระทรวงการคลัง	12988.2	6567.9	3421.2	599.4	244863.5	268440.2	-1
8	กระทรวงคมนาคม	9704.7	2436.2	177009.1	57.1	751.6	189958.7	-1
13	กระทรวงมหาดไทย	19360.2	12718.1	38441.4	235503.6	28464.4	334487.7	-1
17	กระทรวงศึกษาธิการ	223199.6	20029.5	19437.0	88628.9	5154.7	356449.7	-1

กลุ่มที่ 2

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
0	สำนักนายกรัฐมนตรี	4104.5	3078.3	6104.3	12397.0	13606.8	39290.9	0
3	กระทรวงการต่างประเทศ	2697.3	2509.6	397.5	1348.7	1203.3	8156.4	0
4	กระทรวงการท่องเที่ยวและกีฬา	1572.6	892.1	1564.4	735.9	1327.7	6092.7	0
5	กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์	2746.6	2519.3	397.8	16598.3	79.8	22341.8	0
6	กระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัยและนวัตกรรม	11248.4	939.9	9956.4	104384.6	1597.7	128127.0	0
7	กระทรวงเกษตรและสหกรณ์	24376.6	11495.8	72292.8	2001.1	564.8	110731.1	0
9	กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม	1311.6	2009.9	1243.4	1945.8	1791.0	8301.7	0
10	กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	10992.6	5199.8	10408.8	1103.6	1630.9	29335.7	0
11	กระทรวงพลังงาน	822.0	531.2	707.8	30.7	187.5	2279.2	0
12	กระทรวงพาณิชย์	2216.4	1585.4	845.4	479.5	1698.8	6825.5	0
14	กระทรวงยุติธรรม	10363.6	8775.0	3333.1	1937.6	2418.8	26828.1	0
15	กระทรวงแรงงาน	3671.6	1375.7	375.7	63712.9	584.4	69720.3	0
16	กระทรวงวัฒนธรรม	2321.9	1430.5	2325.4	1070.7	814.4	7962.9	0
18	กระทรวงสาธารณสุข	107548.0	14411.4	17025.5	6637.5	179.9	145802.3	0
19	กระทรวงอุตสาหกรรม	1635.9	603.4	745.8	157.9	1516.2	4659.2	0

โดยข้อมูลรายจ่ายโดยเฉลี่ยของแต่ละกลุ่มมีดังนี้

cluster_id	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น
-1	71256.02	13170.92	50546.96	65916.680000	71882.8
0	12508.64	3823.82	8514.94	14302.786667	1946.8

สรุปผลลัพธ์การทำ clustering ด้วยวิธี DBSCAN

เนื่องจากค่าที่ได้ในแต่ล่ะกลุ่มไม่สามารถเจาะจงอธิบาลรายละเอียดได้อย่างชัดเจนจึงอธิบายผ่านข้อมูลรายจ่ายโดยเฉลี่ยของแต่ล่ะกลุ่มแทน โดยแต่ล่ะกลุ่มจะมีลักษณะคร่าวๆดังนี้

กลุ่มที่ 1 จะมีงบเงินอุดหนุน,งบดำเนินงาน,งบบุคลากร,งบรายจ่ายอื่น,งบลงทุนอยู่ในช่วงปานกลางถึงสูง และ ถูกจัดกลุ่มว่าเป็น noise

กลุ่มที่ 2 จะมี งบบุคลากร,งบเงินอุดหนุนอยู่ในช่วงปานกลางถึงสูง และ มีงบลงทุน,งบดำเนินงาน,งบรายจ่ายอื่น อยู่ในช่วงปานกลางถึงต่ำ

โดยสาเหตุที่ผลการจัดกลุ่มออกมาเป็นเช่นนี้อาจเป็นเพราะค่ารายจ่ายของแต่ล่ะชุดข้อมูลมีความใกล้เคียงกับค่าเฉลี่ยของแต่ล่ะกลุ่มเลยถูกจัดกลุ่มออกมาในรูปแบบนี้

3.เปรียบเทียบกลุ่มที่ได้จาก K-mean, Hierarchical และ DBScan ว่า วิธีใดจัดกลุ่มแล้ว ให้กลุ่มที่มีความเหมือนหรือความคล้ายกัน วิธีใดให้ผลลัพธ์ที่ต่างกันออกไป

จากการเปรียบเทียบพบว่า กลุ่มที่ได้จาก K-mean และกลุ่มที่ได้จาก Hierarchical ได้กลุ่มที่มีรายจ่ายใกล้เคียงกันและมีจำนวนกลุ่มใกล้เคียงกัน แต่กลุ่มที่ได้จาก DBScan พบว่าได้ผลลัพธ์แตกต่างจากสองกลุ่มแรกโดยสิ้นเชิง

Part 2

ใช้งบประมาณรายจ่ายส่วนย่อย 5 ด้าน ในปีพ.ศ. 2560-2564 จัดกลุ่มกระทรวงตามงบประมาณรายจ่าย แยกตามปี ด้วยวิธี K-mean

4.จากผลลัพธ์ที่ได้ อภิปรายผลลัพธ์ของการจัดกลุ่ม เมื่อปีเปลียนไป ลักษณะรวมของกลุ่มที่ได้ เหมือนหรือมีความแตกต่างอย่างไร เมื่อไร กระทรวงใดมีความเปลี่ยนแปลงบ้าง เปลี่ยนแปลงอย่างไร

ผลลัพธ์ของวิธี K-mean ได้ทำการแบ่งข้อมูลงบประมาณรายจ่ายส่วนย่อยปี พ.ศ. 2560 ออกเป็น 6 กลุ่ม

Cluster_id	จำนวนข้อมูลที่อยู่ใน cluster นั้นๆ
0	13
1	1
2	1
3	1
4	1
5	3

กลุ่ม 1

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
0	สำนักนายกรัฐมนตรี	3894.0	2853.3	4777.3	9458.4	14429.3	35412.3	0
3	กระทรวงการต่างประเทศ	2997.4	2590.2	604.2	1147.2	1370.6	8709.6	0
4	กระทรวงการท่องเที่ยวและกีฬา	1424.4	1113.7	1658.1	405.3	1946.9	6548.4	0
5	กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์	2667.9	3401.0	952.7	4419.3	219.1	11660.0	0
8	ระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	10468.6	6712.3	12880.3	2224.8	1435.0	33721.0	0
9	ระทรวงเทคโนโลยีสารสนเทศและการสื่อสาร	1164.5	809.1	978.4	2263.6	2468.6	7684.2	0
10	กระทรวงพลังงาน	783.7	542.2	451.3	24.7	251.1	2053.0	0
11	กระทรวงพาณิชย์	2215.6	1969.0	411.8	623.0	1815.3	7034.7	0
13	กระทรวงยุติธรรม	8618.7	8988.1	3128.0	931.0	1885.0	23550.8	0
14	กระทรวงแรงงาน	3454.8	1700.3	461.9	40695.9	878.0	47190.9	0
15	กระทรวงวัฒนธรรม	2131.1	775.0	1816.6	1495.5	1168.7	7386.9	0
16	กระทรวงวิทยาศาสตร์และเทคโนโลยี	402.3	338.4	620.9	8922.8	1218.4	11502.8	0
19	กระทรวงอุตสาหกรรม	1622.3	638.9	586.5	575.7	2243.4	5666.8	0

กลุ่ม 2

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
17	กระทรวงศึกษาธิการ	258020.8	36824.8	33810.3	175509.7	9244.4	513410.0	1

กลุ่ม 3

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
2	กระทรวงการคลัง	12014.6	6188.9	3343.7	238.8	195927.6	217713.6	2

กลุ่ม 4

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
7	กระทรวงคมนาคม	9248.9	2344.5	137892.7	462.3	801.6	150750.0	3

กลุ่ม 5

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
12	กระทรวงมหาดไทย	17293.9	12809.2	29055.5	246083.2	27566.7	332808.5	4

กลุ่ม 6

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
1	กระทรวงกลาโหม	90837.3	25772.7	8398.4	3536.4	84905.1	213449.9	5
6	กระทรวงเกษตรและสหกรณ์	25027.8	14022.7	46346.7	1952.3	917.9	88267.4	5
18	กระทรวงสาธารณสุข	89008.3	14573.1	20126.3	6793.7	262.9	130764.3	5

โดยข้อมูลรายจ่ายโดยเฉลี่ยของแต่ละกลุ่มมีดังนี้

cluster_id	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น
0	3218.869231	2494.730769	2256.000000	5629.784615	2409.953846
1	258020.800000	36824.800000	33810.300000	175509.700000	9244.400000
2	12014.600000	6188.900000	3343.700000	238.800000	195927.600000
3	9248.900000	2344.500000	137892.700000	462.300000	801.600000
4	17293.900000	12809.200000	29055.500000	246083.200000	27566.700000
5	68291.133333	18122.833333	24957.133333	4094.133333	28695.300000

ผลลัพธ์ของวิธี K-mean ได้ทำการแบ่งข้อมูลงบประมาณรายจ่ายส่วนย่อยปี 2561 ออกเป็น 6 กลุ่ม

Cluster_id	จำนวนข้อมูลที่อยู่ใน cluster นั้นๆ
0	3
1	1
2	13
3	1
4	1
5	1

กลุ่ม 1

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
1	กระทรวงกลาโหม	91153.1	25039.5	9544.1	6610.7	88176.3	220523.7	0
6	กระทรวงเกษตรและสหกรณ์	25019.6	15538.5	55276.8	3139.3	856.5	99830.7	0
18	กระทรวงสาธารณสุข	91525.1	14110.2	22035.6	6978.5	244.6	134894.0	0

กลุ่ม 2

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
7	กระทรวงคมนาคม	9492.6	2308.9	156059.3	268.4	638.8	168768.0	1

กลุ่ม 3

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
0	สำนักนายกรัฐมนตรี	3941.3	2770.8	3682.2	9474.7	14005.7	33874.7	2
3	กระทรวงการต่างประเทศ	3045.3	2534.2	574.0	1182.7	1425.3	8761.5	2
4	กระทรวงการท่องเที่ยวและกีฬา	1342.8	1107.3	1911.2	340.1	1921.5	6622.9	2
5	กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์	2652.9	3542.9	1383.2	5951.5	187.0	13717.5	2
8	กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม	1163.7	1163.4	1379.2	2587.1	277.8	6571.2	2
9	กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	10632.0	6284.7	12447.3	2493.2	1787.5	33644.7	2
10	กระทรวงพลังงาน	785.1	542.7	544.0	24.5	324.8	2221.1	2
11	กระทรวงพาณิชย์	2208.7	1716.8	396.2	637.0	2115.0	7073.7	2
13	กระทรวงยุติธรรม	8775.7	8648.1	4094.6	1014.6	1947.8	24480.8	2
14	กระทรวงแรงงาน	3491.5	1582.6	511.9	42776.4	1197.3	49559.7	2
15	กระทรวงวัฒนธรรม	2115.6	733.9	2733.9	1338.6	1180.3	8102.3	2
16	กระทรวงวิทยาศาสตร์และเทคโนโลยี	406.6	337.8	682.5	12124.0	724.5	14275.4	2
19	กระทรวงอุตสาหกรรม	1586.3	609.1	800.8	516.3	1735.7	5248.2	2

กลุ่ม 4

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
2	กระทรวงการคลัง	12235.8	6177.6	3177.1	89.2	216561.9	238241.6	3

กลุ่ม 5

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
12	กระทรวงมหาดไทย	17815.0	14050.7	29648.7	264311.4	28477.8	354303.6	4

กลุ่ม 6

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
17	กระทรวงศึกษาธิการ	250754.5	33148.8	34287.6	180797.9	8958.9	507947.7	5

โดยข้อมูลรายจ่ายโดยเฉลี่ยของแต่ละกลุ่มมีดังนี้

cluster_id	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น
0	69232.600000	18229.400000	28952.166667	5576.166667	29759.133333
1	9492.600000	2308.900000	156059.300000	268.400000	638.800000
2	3242.115385	2428.792308	2395.461538	6189.284615	2217.707692
3	12235.800000	6177.600000	3177.100000	89.200000	216561.900000
4	17815.000000	14050.700000	29648.700000	264311.400000	28477.800000
5	250754.500000	33148.800000	34287.600000	180797.900000	8958.900000

ผลลัพธ์ของวิธี K-mean ได้ทำการแบ่งข้อมูลงบประมาณรายจ่ายส่วนย่อยปี **พ.ศ.** 2562 ออกเป็น 6 กลุ่ม

Cluster_id	จำนวนข้อมูลที่อยู่ใน cluster นั้นๆ
0	13
1	3
2	1
3	1
4	1
5	1

กลุ่ม 1

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
0	สำนักนายกรัฐมนตรี	3954.7	2962.6	6772.5	11813.6	15327.0	40830.4	0
3	กระทรวงการต่างประเทศ	2951.1	2577.8	419.5	1409.3	1839.6	9197.3	0
4	กระทรวงการท่องเที่ยวและกีฬา	1509.6	1111.2	1441.6	369.5	1643.1	6075.0	0
5	กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์	2741.8	2998.5	915.5	6510.5	176.3	13342.6	0
8	กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม	1233.6	896.1	1243.2	1773.2	267.3	5413.4	0
9	กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	10875.0	6189.6	11026.6	948.3	1653.1	30692.6	0
10	กระทรวงพลังงาน	788.6	528.1	641.0	34.3	309.0	2301.0	0
11	กระทรวงพาณิชย์	2219.5	1609.1	453.2	650.9	1956.5	6889.2	0
13	กระทรวงยุติธรรม	9734.9	9147.3	3123.0	950.8	2187.9	25143.9	0
14	กระทรวงแรงงาน	3589.5	1565.1	461.2	46233.3	745.2	52594.3	0
15	กระทรวงวัฒนธรรม	2131.6	736.4	2628.9	1382.8	1235.6	8115.3	0
16	กระทรวงวิทยาศาสตร์และเทคโนโลยี	415.0	324.5	781.1	12947.3	259.7	14727.6	0
19	กระทรวงอุตสาหกรรม	1620.3	620.3	642.3	378.0	1970.3	5231.2	0

กลุ่ม 2

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
1	กระทรวงกลาโหม	93590.8	24735.3	11188.3	5377.5	92234.7	227126.6	1
6	กระทรวงเกษตรและสหกรณ์	24739.8	14596.1	65676.5	2979.3	1005.2	108996.9	1
18	กระทรวงสาธารณสุข	95815.7	14808.5	17895.0	6653.7	215.8	135388.7	1

กลุ่ม 3

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
7	กระทรวงคมนาคม	9472.2	2468.7	166867.5	45.9	744.3	179598.6	2

กลุ่ม 4

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
2	กระทรวงการคลัง	12679.7	6433.3	3326.1	753.0	219755.9	242948.0	3

กลุ่ม 5

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
17	กระทรวงศึกษาธิการ	237718.1	28233.1	32548.7	179311.0	9835.5	487646.4	4

กลุ่ม 6

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
12	กระทรวงมหาดไทย	18779.3	14070.3	31700.4	278770.1	28481.6	371801.7	5

โดยข้อมูลรายจ่ายโดยเฉลี่ยของแต่ละกลุ่มมีดังนี้

cluster_id	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น
0	3366.553846	2405.123077	2349.969231	6569.369231	2274.661538
1	71382.100000	18046.633333	31586.600000	5003.500000	31151.900000
2	9472.200000	2468.700000	166867.500000	45.900000	744.300000
3	12679.700000	6433.300000	3326.100000	753.000000	219755.900000
4	237718.100000	28233.100000	32548.700000	179311.000000	9835.500000
5	18779.300000	14070.300000	31700.400000	278770.100000	28481.600000

ผลลัพธ์ของวิธี K-mean ได้ทำการแบ่งข้อมูลงบประมาณรายจ่ายส่วนย่อยปี 2563 ออกเป็น 5 กลุ่ม

Cluster_id	จำนวนข้อมูลที่อยู่ใน cluster นั้นๆ
0	13
1	2
2	3
3	1
4	1

กลุ่ม 1

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
0	สำนักนายกรัฐมนตรี	4076.1	3087.5	6567.2	10321.9	15056.2	39108.9	0
3	กระทรวงการต่างประเทศ	2754.0	2558.6	468.3	1483.9	1662.8	8927.6	0
4	กระทรวงการท่องเที่ยวและกีฬา	1552.5	1068.9	1315.9	740.8	1393.2	6071.3	0
5	กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์	2745.1	2812.2	645.1	14924.3	154.9	21281.6	0
6	กระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัย และนวัตกรรม	11700.1	1012.7	9091.6	104146.9	1944.2	127895.5	0
9	กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม	1273.4	1517.0	1385.6	1927.9	794.0	6897.9	0
10	กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	11012.2	5924.0	10285.0	1216.6	1932.2	30370.0	0
11	กระทรวงพลังงาน	808.6	536.7	489.1	31.2	292.4	2158.0	0
12	กระทรวงพาณิชย์	2249.4	1636.9	792.3	529.6	2344.9	7553.1	0
14	กระทรวงยุติธรรม	10254.9	9090.7	4336.6	1112.4	2154.5	26949.1	0
15	กระทรวงแรงงาน	3654.6	1516.4	489.6	54593.0	624.8	60878.4	0
16	กระทรวงวัฒนธรรม	2352.7	1430.2	2391.5	1273.7	1121.6	8569.7	0
19	กระทรวงอุตสาหกรรม	1633.6	620.1	844.3	218.6	2047.2	5363.8	0

กลุ่ม 2

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
7	กระทรวงเกษตรและสหกรณ์	24706.3	13751.9	67815.5	2700.8	859.2	109833.7	1
8	กระทรวงคมนาคม	9684.5	2500.4	165887.2	59.1	708.9	178840.1	1

กลุ่ม 3

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
1	กระทรวงกลาโหม	92346.6	24777.8	15433.6	5162.5	95632.9	233353.4	2
17	กระทรวงศึกษาธิการ	230196.3	19847.4	22527.3	89316.7	6772.6	368660.3	2
18	กระทรวงสาธารณสุข	99248.1	14860.6	17465.5	6945.7	209.9	138729.8	2

กลุ่ม 4

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
2	กระทรวงการคลัง	13093.2	6734.3	3108.5	873.7	225866.3	249676.0	3

กลุ่ม 5

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
13	กระทรวงมหาดไทย	19158.1	13538.0	34694.8	256160.3	29456.2	353007.4	4

โดยข้อมูลรายจ่ายโดยเฉลี่ยของแต่ละกลุ่มมีดังนี้

cluster_id	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น
0	4312.861538	2523.992308	3007.853846	14809.292308	2424.838462
1	17195.400000	8126.150000	116851.350000	1379.950000	784.050000
2	140597.000000	19828.600000	18475.466667	33808.300000	34205.133333
3	13093.200000	6734.300000	3108.500000	873.700000	225866.300000
4	19158.100000	13538.000000	34694.800000	256160.300000	29456.200000

ผลลัพธ์ของวิธี K-mean ได้ทำการแบ่งข้อมูลงบประมาณรายจ่ายส่วนย่อยปี **พ.ศ.** 2564 ออกเป็น 5 กลุ่ม

Cluster_id	จำนวนข้อมูลที่อยู่ใน cluster นั้นๆ
0	13
1	3
2	2
3	1
4	1

กลุ่ม 1

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
0	สำนักนายกรัฐมนตรี	4104.5	3078.3	6104.3	12397.0	13606.8	39290.9	0
3	กระทรวงการต่างประเทศ	2697.3	2509.6	397.5	1348.7	1203.3	8156.4	0
4	กระทรวงการท่องเที่ยวและกีฬา	1572.6	892.1	1564.4	735.9	1327.7	6092.7	0
5	กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์	2746.6	2519.3	397.8	16598.3	79.8	22341.8	0
6	กระทรวงการอุดมศึกษา วิทยาศาสตร์ วิจัยและนวัตกรรม	11248.4	939.9	9956.4	104384.6	1597.7	128127.0	0
9	กระทรวงดิจิทัลเพื่อเศรษฐกิจและสังคม	1311.6	2009.9	1243.4	1945.8	1791.0	8301.7	0
10	กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	10992.6	5199.8	10408.8	1103.6	1630.9	29335.7	0
11	กระทรวงพลังงาน	822.0	531.2	707.8	30.7	187.5	2279.2	0
12	กระทรวงพาณิชย์	2216.4	1585.4	845.4	479.5	1698.8	6825.5	0
14	กระทรวงยุติธรรม	10363.6	8775.0	3333.1	1937.6	2418.8	26828.1	0
15	กระทรวงแรงงาน	3671.6	1375.7	375.7	63712.9	584.4	69720.3	0
16	กระทรวงวัฒนธรรม	2321.9	1430.5	2325.4	1070.7	814.4	7962.9	0
19	กระทรวงอุตสาหกรรม	1635.9	603.4	745.8	157.9	1516.2	4659.2	0

กลุ่ม 2

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
1	กระทรวงกลาโหม	91027.4	24102.9	14426.1	4794.4	80179.8	214530.6	1
17	กระทรวงศึกษาธิการ	223199.6	20029.5	19437.0	88628.9	5154.7	356449.7	1
18	กระทรวงสาธารณสุข	107548.0	14411.4	17025.5	6637.5	179.9	145802.3	1

กลุ่ม 3

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
7	กระทรวงเกษตรและสหกรณ์	24376.6	11495.8	72292.8	2001.1	564.8	110731.1	2
8	กระทรวงคมนาคม	9704.7	2436.2	177009.1	57.1	751.6	189958.7	2

กลุ่ม 4

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
2	กระทรวงการคลัง	12988.2	6567.9	3421.2	599.4	244863.5	268440.2	3

กลุ่ม 5

	กระทรวง-งบรายจ่าย	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น	รวม	cluster_id
13	กระทรวงมหาดไทย	19360.2	12718.1	38441.4	235503.6	28464.4	334487.7	4

โดยข้อมูลรายจ่ายโดยเฉลี่ยของแต่ละกลุ่มมีดังนี้

cluster_id	งบบุคลากร	งบดำเนินงาน	งบลงทุน	งบเงินอุดหนุน	งบรายจ่ายอื่น
0	4285.000000	2419.238462	2954.292308	15838.707692	2189.023077
1	140591.666667	19514.600000	16962.866667	33353.600000	28504.800000
2	17040.650000	6966.000000	124650.950000	1029.100000	658.200000
3	12988.200000	6567.900000	3421.200000	599.400000	244863.500000
4	19360.200000	12718.100000	38441.400000	235503.600000	28464.400000

สรุปคำตอบข้อ 4

การจัดกลุ่มกระทรวงตามงบประมาณรายจ่าย แยกตามปี ด้วยวิธี k-mean ในปี พ.ศ. 2560 , 2561, 2562 จะได้ผลลัพธ์ในรูปแบบคล้ายกัน คือได้จำนวน 6 กลุ่ม และสมาชิกกระทรวงในแต่ล่ะกลุ่มเหมือนกัน คือ กลุ่ม1 (กระทรวงคมนาคม),กลุ่ม2 (กระทรวงการคลัง),กลุ่ม3 (กระทรวงศึกษาธิการ),กลุ่ม4 (กระทรวงมหาดไทย),กลุ่ม5 (กระทรวงกลาโหม , กระทรวงเกษตรและสหกรณ์ , กระทรวงสาธารณสุข),กลุ่ม6 (กระทรวงอื่ๆที่เหลืออีก 13 กระทรวง) แต่ในการจัดกลุ่มกระทรวงตามงบประมาณรายจ่าย แยกตามปี ด้วยวิธี k-mean ในปี พ.ศ. 2563 , 2564 จะได้ผลลัพธ์ในรูปแบบแตกต่างจาก 3 ปีแรก คือได้จำนวน 5 กลุ่ม โดยกระทรวงที่มีความเปลี่ยนแปลงคือกระทรวงศึกษาธิการ ,กระทรวงคมนาคม,กระทรวงกลาโหม , กระทรวงเกษตรและสหกรณ์ , กระทรวงสาธารณสุข โดย 5 กระทรวงนี้ได้ถูกจัดกลุ่มใหม่ตามนี้ กลุ่ม1(กระทรวงกลาโหม ,กระทรวงศึกษาธิการ , กระทรวงสาธารณสุข),กลุ่ม2(กระทรวงเกษตรและสหกรณ์ ,กระทรวงคมนาคม) ส่วนกลุ่มที่เหลือจะถูกจัดกลุ่มตามเดิมดังนี้ กลุ่ม3(กระทรวงการคลัง) ,กลุ่ม4(กระทรวงมหาดไทย) ,กลุ่ม5(กระทรวงอื่ๆที่เหลืออีก 13 กระทรวง)

5.จากกระทรวงที่ 1 และ 2 ที่ได้รับมอบหมาย วิเคราะห์ว่า ทั้ง 2 กระทรวงอยู่ในกลุ่มใด ลักษณะของกลุ่มมีอะไรบ้าง

กระทรวงที่ได้รับมอบหมายกระทรวงที่ 1 คือ กระทรวงศึกษาธิการ และกระทรวงที่ 2 คือสำนักนายกรัฐมนตรี โดยในปีพ.ศ. 2560 ถึง 2562 กระทรวงศึกษาธิการอยู่ในกลุ่มที่มีลักษณะคือ จะมีงบบุคลากร, งบเงินอุดหนุนสูง และ มีงบดำเนินงาน,งบลงทุน,งบรายจ่ายอื่นอยู่ในช่วงปานกลางถึงสูง โดยในส่วนของสำนักนายกรัฐมนตรีอยู่ในกลุ่มที่มีลักษณะคือ จะมีงบบุคลากร,งบเงินอุดหนุน,งบดำเนินงาน,งบลงทุน,งบรายจ่ายอื่นอยู่ในช่วงปานกลางถึงต่ำ ต่อมาในส่วนของปีพ.ศ. 2563 ถึง 2564 กระทรวงศึกษาธิการอยู่ในกลุ่มที่มีลักษณะคือ จะมีงบบุคลากรสูง และ มีงบดำเนินงาน,งบเงินอุดหนุน,งบลงทุน,งบรายจ่ายอื่นอยู่ในช่วงปานกลางถึงสูง
โดยในส่วนของสำนักนายกรัฐมนตรีอยู่ในกลุ่มที่มีลักษณะคือ จะมีงบเงินอุดหนุนอยู่ในช่วงปานกลางถึงสูง และมีงบบุคลากร,งบดำเนินงาน,งบลงทุน,งบรายจ่ายอื่นอยู่ในช่วงปานกลางถึงต่ำ

Regression Part

วัตถุประสงค์เพื่อทำนายค่าใช้จ่ายรัฐบาลตามลักษณะงาน โดยใช้ข้อมูลงบประมาณรายจ่ายกระทรวงต่างๆ และ เพื่อวิเคราะห์ประสิทธิภาพของโมเดลที่ได้

โดยจะ มีงบประมาณรายจ่ายกระทรวงต่างๆกระทรวง 20 กระทรวง และมีค่าใช้จ่ายรัฐบาลตามลักษณะงาน 7 ด้าน ดังนี้

- ด้านการป้องกันประเทศ
- ด้านการรักษาความสงบภายใน
- ด้านการเศรษฐกิจ
- ด้านการสาธารณสุข
- ด้านการศาสนา วัฒนธรรมและนันนทาการ
- ด้านการศึกษา
- ด้านการสังคมสงเคราะห์

Part 1

สร้างโมเดล multiple linear regression ทำนายค่าใช้จ่ายรัฐบาลด้านการศาสนา วัฒนธรรมและนันนทาการ โดยใช้ข้อมูลงบประมาณรายจ่ายกระทรวงต่างๆ

Part 1.1

ลักษณะงานที่กลุ่มได้รับมอบหมายคือ ด้านการศาสนา วัฒนธรรมและนันนทาการ โดยใน part 1.1 จะทำการสร้างโมเดลทำนายค่าใช้จ่ายรัฐบาลตาม ด้านการศาสนา วัฒนธรรมและนันนทาการ โดยใช้กระทรวงตามค่า correlation

1.เลือกงบประมาณกระทรวงที่มีค่า correlation สูง 5 ลำดับแรก มาเป็นตัวแปรต้นหรือ input ของโมเดล

นำข้อมูลมาหาค่า correlation กับทุกกระทรวงและทำการ sort เพื่อหา 5 อันดับสูงสุดที่มีความสัมพันธ์กับด้านการศาสนา วัฒนธรรมและนันนทาการ โดยได้ออกมา 5 อันดับ คือ กระทรวงมหาดไทย , กระทรวงศึกษาธิการ , กระทรวงพาณิชย์ , กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม , กระทรวงวัฒนธรรม ตามลำดับ ดังนี้

	กระทรวงมหาดไทย	กระทรวงศึกษาธิการ	กระทรวงพาณิชย์	กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	กระทรวงวัฒนธรรม
0	102793.9181	190262.8362	4620.8538	14883.2082	2239.6428
1	139800.1791	203752.3867	4658.0021	16150.1673	2480.8597
2	160643.4755	225592.2688	4829.7984	16237.9477	2937.3920
3	179115.7526	282254.3414	5826.1381	17515.4592	4281.0641
4	190930.1453	301085.1582	6152.7260	18819.2856	4371.0958
5	195269.3223	332298.6160	6416.5642	20603.2571	4914.6578
6	187998.7071	346713.0933	6251.6728	20108.5233	4347.8166
7	230781.2686	391131.8796	6899.2261	22936.7930	5119.5841
8	285255.0000	420490.0000	6592.7000	25585.0000	5468.0000
9	308835.0000	460411.7000	7430.3000	30657.8000	5957.1000
10	333145.2000	482788.6000	9171.6000	31487.3000	6624.7000
11	340171.6000	501326.1000	7247.2000	30245.4000	7047.4000
12	341820.8000	517076.7000	7192.6000	35877.9000	7742.3000
13	332808.5000	513410.0000	7034.7000	33721.0000	7386.9000
14	354303.6000	507947.7000	7073.7000	33644.7000	8102.3000
15	371801.7000	487646.4000	6889.2000	30692.6000	8115.3000
16	353007.4000	368660.3000	7553.1000	30370.0000	8569.7000

A.สร้างโมเดล multiple linear regression ทำนายค่าใช้จ่ายรัฐบาลด้านการศาสนา วัฒนธรรมและนันนทาการ

ได้ผลลัพธ์โมเดล multiple linear regression ทำนายค่าใช้จ่ายรัฐบาลด้านการศาสนา วัฒนธรรมและนันนทาการ ดังนี้

OLS Regression Results

Dep. Variable:	การศาสนา วัฒนธรรม และนันทนาการ	R-squared:	0.986
Model:	OLS	Adj. R-squared:	0.980
Method:	Least Squares	F-statistic:	157.8
Date:	Fri, 04 Mar 2022	Prob (F-statistic):	7.46e-10
Time:	13:57:48	Log-Likelihood:	-133.66
No. Observations:	17	AIC:	279.3
Df Residuals:	11	BIC:	284.3
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
กระทรวงมหาดไทย	0.0366	0.012	2.985	0.012	0.010	0.064
กระทรวงศึกษาธิการ	0.0097	0.006	1.652	0.127	-0.003	0.023
กระทรวงพาณิชย์	2.0873	0.319	6.543	0.000	1.385	2.789
กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	-0.0319	0.124	-0.257	0.802	-0.305	0.242
กระทรวงวัฒนธรรม	-0.4075	0.397	-1.025	0.327	-1.282	0.467
const	-7049.5027	1371.239	-5.141	0.000	-1.01e+04	-4031.426

Omnibus:	0.470	Durbin-Watson:	1.912
Prob(Omnibus):	0.790	Jarque-Bera (JB):	0.538
Skew:	-0.313	Prob(JB):	0.764
Kurtosis:	2.394	Cond. No.	3.50e+06

B.จากโมเดลที่ได้ พิจารณาว่า

a.เป็น linear model ที่ดีหรือไม่

b.งบประมาณกระทรวงใด ส่งผลหรือไม่ส่งผล ต่อค่าใช้จ่ายรัฐบาลด้านการศาสนา วัฒนธรรมและนันนทาการ

ตอบคำถามข้อ a. : จะสังเกตุได้ว่าค่า R-squared ของ model จะอยู่ที่ 0.986 หรือสามารถบอกได้ว่า 98.6 เปอร์เซ็นต์ของข้อมูลเราสามารถอธิบายได้ด้วย linear model นี้และสามารถดูได้ว่าค่า |r| = 0.993 ซึ่งอยู่ในช่วง 0.8 ≤ |r| ≤ 1 บ่งบอกว่าเป็น strong correlation แสดงว่าเป็น linear model ที่ดี

ตอบคำถามข้อ b. : จะได้ว่า กระทรวงมหาดไทยและกระทรวงพาณิชย์ ส่งผลต่อค่าใช้จ่ายรัฐบาลด้านการศาสนา วัฒนธรรมและนันนทาการ แต่ กระทรวงศึกษาธิการ , กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม , กระทรวงวัฒนธรรม ไม่ส่งผลต่อ ค่าใช้จ่ายรัฐบาลด้านการศาสนา วัฒนธรรมและนันนทาการ โดยได้พิจารณาจากการทำ Hypothesis testing ดังนี้

กำหนดให้ null hypothesis(H0) คือไม่ขึ้นกับค่านั้น(slope b=0) และ alternative hypothesis(Ha) คือขึ้นกับค่านั้น(slope b!=0) โดยจะกำหนดให้ significance level ที่ 0.05 ดังนี้ถ้า p value น้อยกว่า 0.05 จะถือว่า reject null โดยค่า p value ของแต่ละกระทรวงมีดังนี้ (ตัดที่ทศนิยม 3 ตำแหน่ง)

กระทรวงมหาดไทย มีค่า p value เท่ากับ 0.012
กระทรวงศึกษาธิการ มีค่า p value เท่ากับ 0.127
กระทรวงพาณิชย์ มีค่า p value เท่ากับ 0.000
กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม มีค่า p value เท่ากับ 0.802
กระทรวงวัฒนธรรม มีค่า p value เท่ากับ 0.327

ดังนั้นจะได้ว่า กระทรวงมหาดไทยและกระทรวงพาณิชย์ reject null ที่ significance level ที่ 0.05 แสดงว่าสองกระทรวงนี้ส่งผลต่อด้านการศาสนา วัฒนธรรมและนันนทาการ

C.หากงบประมาณกระทรวงใดไม่ส่งผล สามารถนำออกจากตัวแปรต้น และนำกระทรวงที่เหลือไปสร้างโมเดลใหม่ หรือทำขั้นตอน A-C .ซ้ำไปเรื่อยๆ จนสุดท้าย ได้กระทรวงที่ส่งผลต่อค่าใช้จ่ายรัฐบาลตามลักษณะงานนี้ หรือ อาจจะไม่เหลือกระทรวงที่ส่งผล

ได้นำกระทรวงที่ไม่ส่งผลคือ กระทรวงศึกษาธิการ , กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม , กระทรวงวัฒนธรรม ออกจากตัวแปรต้น และได้นำ กระทรวงที่ส่งผลไปสร้าง model ใหม่ ได้ผลลัพธ์ออกมาดังนี้

OLS Regression Results

Dep. Variable:	การศาสนา วัฒนธรรม และนันทนาการ	R-squared:	0.980
Model:	OLS	Adj. R-squared:	0.977
Method:	Least Squares	F-statistic:	339.2
Date:	Fri, 04 Mar 2022	Prob (F-statistic):	1.38e-12
Time:	13:58:39	Log-Likelihood:	-136.94
No. Observations:	17	AIC:	279.9
Df Residuals:	14	BIC:	282.4
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
กระทรวงมหาดไทย	0.0346	0.004	8.202	0.000	0.026	0.044
กระทรวงพาณิชย์	2.2780	0.326	6.992	0.000	1.579	2.977
const	-7152.1137	1393.509	-5.132	0.000	-1.01e+04	-4163.335

Omnibus:	0.218	Durbin-Watson:	1.273
Prob(Omnibus):	0.897	Jarque-Bera (JB):	0.410
Skew:	-0.112	Prob(JB):	0.815
Kurtosis:	2.273	Cond. No.	1.87e+06

2. เมื่อนำกระทรวงที่ไม่ส่งผลออก เปรียบเทียบค่า r-square รวมทั้งค่า p-value ของแต่ละกระทรวง ที่เปลี่ยนไป ว่าค่ามีการเปลี่ยนแปลงอย่างไร ค่า r-square และค่า p-value มีการเปลี่ยนแปลงที่สัมพันธ์กันหรือไม่

Model A คือ linear model ที่รวมเอากระทรวงที่ไม่ส่งผลนำไปคำนวณด้วย
Model B คือ linear model ที่นำกระทรวงที่ไม่ส่งผลออกก่อนนำไปคำนวณ

	r-square
Model A	0.986
Model B	0.980

ในส่วนของ Model B ได้ทดสอบ ค่า index เพิ่มเติม ดังนี้
MSE(Mean Square Error) เท่ากับ 580936.6
RMSE(Root Mean Square Error) เท่ากับ 762.192

โดยในส่วนของค่า p-value ของแต่ล่ะกระทรวงใน Model A มีดังนี้(ตัดที่ทศนิยม 3 ตำแหน่ง)
กระทรวงมหาดไทย มีค่า p value เท่ากับ 0.012
กระทรวงศึกษาธิการ มีค่า p value เท่ากับ 0.127
กระทรวงพาณิชย์ มีค่า p value เท่ากับ 0.000
กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม มีค่า p value เท่ากับ 0.802
กระทรวงวัฒนธรรม มีค่า p value เท่ากับ 0.327

ในส่วนของค่า p-value ของแต่ล่ะกระทรวงใน Model B มีดังนี้(ตัดที่ทศนิยม 3 ตำแหน่ง)
กระทรวงมหาดไทย มีค่า p value เท่ากับ 0.000
กระทรวงพาณิชย์ มีค่า p value เท่ากับ 0.000

จะสังเกตุได้ว่า ค่า p-value ของกระทรวงมหาดไทยและกระทรวงพาณิชย์ มีค่าน้อยลง โดยจะหมายความว่ามีหลักฐานทางสถิติที่ให้สนับสนุนให้ปฏิเสธสมมติฐานหลักมากขึ้น(หรือก็คือกระทรางนั้นมีความสัมพันธ์กับ output) กล่าวได้ว่า ค่า r-square และค่า p-value มีการเปลี่ยนแปลงที่ไม่สัมพันธ์กัน เพราะค่า r-square ที่น้อยลงบ่งบอกถึงความสัมพันธ์ของข้อมูลที่ต่ำลง (อาจเนื่องมาจากการที่ตัวแปรต้นหายไป ทำให้ความแม่นยำลดน้อยลง) ต่างกับ ค่า p-value ที่น้อยลงบ่งบอกถึง มีหลักฐานทางสถิติที่ให้สนับสนุนให้ปฏิเสธสมมติฐานหลักมากขึ้น ซึ่งขัดแย้งกัน

Part 1.2

ลักษณะงานที่กลุ่มได้รับมอบหมายคือ ด้านการศาสนา วัฒนธรรมและนันนทาการ โดยใน part 1.1 จะทำการสร้างโมเดลทำนายค่าใช้จ่ายรัฐบาลตาม ด้านการศาสนา วัฒนธรรมและนันนทาการ โดยใช้กระทรวงที่ทางกลุ่มเลือก

1.เลือกงบประมาณกระทรวง 5 กระทรวงที่ทางกลุ่มเห็นว่า งบประมาณกระทรวงเหล่านี้ คาดว่ามีผลต่อค่าใช้จ่ายรัฐบาลตาม ด้านการศาสนา วัฒนธรรมและนันนทาการ มากที่สุด 5 ลำดับแรก ให้เหตุผลในการเลือกกระทรวงทั้งห้า

ได้ทำการเลือกมา 5 กระทรวงดังนี้
1.กระทรวงการท่องเที่ยวและกีฬา เหตุผลคือ หากการท่องเที่ยวมีการพัฒนา ด้านการศาสนา วัฒนธรรมและนันนทาการ น่าจะมีการพัฒนาด้วยเนื่องจากคิดว่าชาวต่างชาติที่มาเที่ยวในเมืองไทยชอบเกี่ยวกับการเที่ยววัด ศาสนา และสัมผัสถึงวัฒนธรรมประเพณีของไทยทำให้คิดว่ากระทรวงการท่องเที่ยวและกีฬาน่าจะเกี่ยวข้องกัน

2.กระทรวงวัฒนธรรม เหตุผลคือ กระทรวงวัฒนธรรมมีหน้าที่เกี่ยวกับการคุ้มครอง ป้องกัน อนุรักษ์ บำรุงรักษา ฟื้นฟู ส่งเสริม สร้างสรรค์ เผยแพร่ ค้นคว้า วิจัย พัฒนา สืบทอดศิลปะและทรัพย์สินมรดกทางศิลปวัฒนธรรมของชาติ เพื่อธำรงคุณค่าและเอกลักษณ์ของความเป็นชาติ ซึ่งน่าจะเกี่ยวกับ ด้านการศาสนา วัฒนธรรมและนันนทาการ โดยตรง

3.กระทรวงศึกษาธิการ เหตุผลคือ ในกิจกรระมในโรงเรียนน่าจะมีการส่งเสริมเรื่องนันทนาการและศาสนาและวัฒนา ในการเรียนการสอนภายในชั้นเรียน จึงคิดว่ามีการเกี่ยวข้องกับด้าน การศาสนา วัฒนธรรมและนันนทาการ

4.กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์ เหตุผลคือ นันทนาการมีความเกี่ยวข้องกับชีวิตมนุษย์ตั้งแต่เกิดจนกระทั้งเสียชีวิต และนันทนาการก็มีบทบาทสำคัญในการพัฒนาคุณภาพชีวิตให้มีประสิทธิภาพมากยิ่งขึ้น จึงทำให้คิดว่าน่าจะเกี่ยวกับ ด้านการศาสนา วัฒนธรรมและนันนทาการ โดยตรง

5.กระทรวงการต่างประเทศ เหตุผลคือ กระทรวงการต่างประเทศมีความเกี่ยวข้องกับวัฒนธรรม เช่น การเผยแพร่วัฒนธรรม หรือ ประชาสัมพันธ์งานต่างๆก็จะเกี่ยวกับวัฒนธรรมของประเทศ จึงคิดว่าน่าจะเกี่ยวกับ ด้านการศาสนา วัฒนธรรมและนันนทาการ

A.สร้างโมเดล multiple linear regression ทำนายค่าใช้จ่ายรัฐบาลด้านการศาสนา วัฒนธรรมและนันนทาการ จากกระทรวงทั้งห้า

ได้ผลลัพธ์โมเดล multiple linear regression ทำนายค่าใช้จ่ายรัฐบาลด้านการศาสนา วัฒนธรรมและนันนทาการ ดังนี้

OLS Regression Results

Dep. Variable:	การศาสนา วัฒนธรรม และนันทนาการ	R-squared:	0.987
Model:	OLS	Adj. R-squared:	0.984
Method:	Least Squares	F-statistic:	327.6
Date:	Fri, 04 Mar 2022	Prob (F-statistic):	1.71e-12
Time:	13:59:34	Log-Likelihood:	-133.22
No. Observations:	17	AIC:	274.4
Df Residuals:	13	BIC:	277.8
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
กระทรวงการท่องเที่ยวและกีฬา	0.7572	0.065	11.687	0.000	0.617	0.897
กระทรวงวัฒนธรรม	1.2668	0.264	4.792	0.000	0.696	1.838
กระทรวงการต่างประเทศ	1.0600	0.491	2.158	0.050	-0.001	2.121
const	-3247.5256	2385.548	-1.361	0.197	-8401.188	1906.137

Omnibus:	0.312	Durbin-Watson:	1.122
Prob(Omnibus):	0.855	Jarque-Bera (JB):	0.465
Skew:	-0.040	Prob(JB):	0.793
Kurtosis:	2.194	Cond. No.	1.66e+05

B.จากโมเดลที่ได้ พิจารณาว่า

a.เป็น linear model ที่ดีหรือไม่

b.งบประมาณกระทรวงใด ส่งผลหรือไม่ส่งผล ต่อค่าใช้จ่ายรัฐบาลด้านการศาสนา วัฒนธรรมและนันนทาการ

ตอบคำถามข้อ a. : จะสังเกตุได้ว่าค่า R-squared ของ model จะอยู่ที่ 0.987 หรือสามารถบอกได้ว่า 98.7 เปอร์เซ็นต์ของข้อมูลเราสามารถอธิบายได้ด้วย linear model นี้และสามารถดูได้ว่าค่า |r| = 0.993 ซึ่งอยู่ในช่วง 0.8 ≤ |r| ≤ 1 บ่งบอกว่าเป็น strong correlation แสดงว่าเป็น linear model ที่ดี

ตอบคำถามข้อ b. : จะได้ว่า กระทรวงการท่องเที่ยวและกีฬา,กระทรวงวัฒนธรรม,กระทรวงการต่างประเทศ ส่งผลต่อค่าใช้จ่ายรัฐบาลด้านการศาสนา วัฒนธรรมและนันนทาการ แต่ กระทรวงศึกษาธิการ , กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์ ไม่ส่งผลต่อ ค่าใช้จ่ายรัฐบาลด้านการศาสนา วัฒนธรรมและนันนทาการ โดยได้พิจารณาจากการทำ Hypothesis testing ดังนี้

กำหนดให้ null hypothesis(H0) คือไม่ขึ้นกับค่านั้น(slope b=0) และ alternative hypothesis(Ha) คือขึ้นกับค่านั้น(slope b!=0) โดยจะกำหนดให้ significance level ที่ 0.05 ดังนี้ถ้า p value น้อยกว่า 0.05 จะถือว่า reject null โดยค่า p value ของแต่ล่ะกระทรวงมีดังนี้ (ตัดที่ทศนิยม 3 ตำแหน่ง)

กระทรวงการท่องเที่ยวและกีฬา มีค่า p value เท่ากับ 0.000

กระทรวงวัฒนธรรม มีค่า p value เท่ากับ 0.071

กระทรวงศึกษาธิการ มีค่า p value เท่ากับ 0.359

กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์ มีค่า p value เท่ากับ 0.928

กระทรวงการต่างประเทศ มีค่า p value เท่ากับ 0.053

ดังนั้นจะได้ว่า กระทรวงการท่องเที่ยวและกีฬา,กระทรวงวัฒนธรรม,กระทรวงการต่างประเทศ reject null ที่ significance level ที่ 0.05 แสดงว่าสองกระทรวงนี้ส่งผลต่อด้านการศาสนา วัฒนธรรมและนันนทาการ

C.หากงบประมาณกระทรวงใดไม่ส่งผล สามารถตัดออกจากตัวแปรต้น และทำขั้นตอน A-C.ซ้ำได้

ได้นำกระทรวงที่ไม่ส่งผลคือ กระทรวงศึกษาธิการ , กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์ ออกจากตัวแปรต้น และได้นำ กระทรวงที่ส่งผลไปสร้าง model ใหม่ ได้ผลลัพธ์ออกมาดังนี้

OLS Regression Results

Dep. Variable:	การศาสนา วัฒนธรรม และนันทนาการ	R-squared:	0.987
Model:	OLS	Adj. R-squared:	0.984
Method:	Least Squares	F-statistic:	327.6
Date:	Fri, 04 Mar 2022	Prob (F-statistic):	1.71e-12
Time:	13:59:34	Log-Likelihood:	-133.22
No. Observations:	17	AIC:	274.4
Df Residuals:	13	BIC:	277.8
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
กระทรวงการท่องเที่ยวและกีฬา	0.7572	0.065	11.687	0.000	0.617	0.897
กระทรวงวัฒนธรรม	1.2668	0.264	4.792	0.000	0.696	1.838
กระทรวงการต่างประเทศ	1.0600	0.491	2.158	0.050	-0.001	2.121
const	-3247.5256	2385.548	-1.361	0.197	-8401.188	1906.137

Omnibus:	0.312	Durbin-Watson:	1.122
Prob(Omnibus):	0.855	Jarque-Bera (JB):	0.465
Skew:	-0.040	Prob(JB):	0.793
Kurtosis:	2.194	Cond. No.	1.66e+05

2.เปรียบเทียบ ค่า r-square รวมทั้งค่า p-value ของตัวแปรต้นที่เปลี่ยนไป ระหว่าง Part 1.1 และ Part 1.2 ว่าค่ามีการเปลี่ยนแปลงอย่างไร

Model Part 1.1 คือ linear model ที่คัดเลือกกระทรวงตัวแปรต้นมาด้วย correlation
Model Part 1.2 คือ linear model ที่คัดเลือกกระทรวงตัวแปรต้นมาโดยทางกลุ่มเป็นคนเลือก

	r-square
Model Part 1.1 (นำกระทรวงที่ไม่เกี่ยวข้องออก)	0.980
Model Part 1.2 (นำกระทรวงที่ไม่เกี่ยวข้องออก)	0.987

ในส่วนของ Model Part 1.1 ได้ทดสอบ ค่า index เพิ่มเติม ดังนี้
MSE(Mean Square Error) เท่ากับ 580936.6
RMSE(Root Mean Square Error) เท่ากับ 762.192

ในส่วนของ Model Part 1.2 ได้ทดสอบ ค่า index เพิ่มเติม ดังนี้
MSE(Mean Square Error) เท่ากับ 375069.72
RMSE(Root Mean Square Error) เท่ากับ 612.43

ค่า r-square ของ Model Part 1.2 มีค่ามากกว่า ค่า r-square ของ Model Part 1.1 บ่งบอกได้ว่า Model Part 1.2 เป็น Model ที่สามารถแทนข้อมูลได้ถูกต้อง มากกว่า Model Part 1.1

โดยในส่วนของค่า p-value ของแต่ละกระทรวงใน Model Part 1.1 ก่อนเอากระทรวงที่ไม่เกี่ยวข้องออก มีดังนี้(ตัดที่ทศนิยม3ตำแหน่ง)
กระทรวงมหาดไทย มีค่า p value เท่ากับ 0.012
กระทรวงศึกษาธิการ มีค่า p value เท่ากับ 0.127
กระทรวงพาณิชย์ มีค่า p value เท่ากับ 0.000
กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม มีค่า p value เท่ากับ 0.802
กระทรวงวัฒนธรรม มีค่า p value เท่ากับ 0.327

ในส่วนของค่า p-value ของแต่ละกระทรวงใน Model Part 1.1 โดยเอากระทรวงที่ไม่เกี่ยวข้องออกแล้ว มีดังนี้(ตัดที่ทศนิยม 3 ตำแหน่ง)
กระทรวงมหาดไทย มีค่า p value เท่ากับ 0.000
กระทรวงพาณิชย์ มีค่า p value เท่ากับ 0.000

โดยในส่วนของค่า p-value ของแต่ละกระทรวงใน Model Part 1.2 ก่อนเอากระทรวงที่ไม่เกี่ยวข้องออก มีดังนี้(ตัดที่ทศนิยม3ตำแหน่ง)
กระทรวงการท่องเที่ยวและกีฬา มีค่า p value เท่ากับ 0.000
กระทรวงวัฒนธรรม มีค่า p value เท่ากับ 0.071
กระทรวงศึกษาธิการ มีค่า p value เท่ากับ 0.359
กระทรวงการพัฒนาสังคมและความมั่นคงของมนุษย์ มีค่า p value เท่ากับ 0.928
กระทรวงการต่างประเทศ มีค่า p value เท่ากับ 0.053

ในส่วนของค่า p-value ของแต่ละกระทรวงใน Model Part 1.2 โดยเอากระทรวงที่ไม่เกี่ยวข้องออกแล้ว มีดังนี้(ตัดที่ทศนิยม 3 ตำแหน่ง)
กระทรวงการท่องเที่ยวและกีฬา มีค่า p value เท่ากับ 0.000
กระทรวงวัฒนธรรม มีค่า p value เท่ากับ 0.000
กระทรวงการต่างประเทศ มีค่า p value เท่ากับ 0.050

จะสังเกตุได้ว่า ค่า p-value ของกระทรวงตัวแปรต้นหลังเอากระทรวงที่ไม่เกี่ยวข้องออก มีค่าน้อยลง โดยจะหมายความว่ามีหลักฐานทางสถิติที่ให้สนับสนุนให้ปฏิเสธสมมติฐานหลักมากขึ้น(หรือก็คือกระทรางนั้นมีความสัมพันธ์กับ output)

3.เปรียบเทียบกระทรวงที่ส่งผลจาก Part 1.1 และ 1.2 ว่า อันใดสมเหตุสมผลมากกว่ากัน

กระทรงที่ส่งผลจาก Part 1.1 คือ กระทรวงมหาดไทยและกระทรวงพาณิชย์ ส่วนกระทรวงที่ส่งผลจาก Part 1.2 คือ กระทรวงการท่องเที่ยวและกีฬา,กระทรวงวัฒนธรรม,กระทรวงการต่างประเทศ โดยคิดว่ากระทรวงจาก Part 1.2 สมเหตุสมผลกว่า เนื่องจาก ทางผู้จัดทำได้คัดเลือกมาจากความน่าจะเป็นที่จะเกี่ยวข้องกับด้านการศาสนา วัฒนธรรมและนันนทาการ โดยดูจากหน้าที่และลักษณะงานของกระทรวงต่างๆ ต่อจากของ Part 1.1 ที่ดูแค่ correlation ซึ่งอาจเป็นไปได้ว่าการที่ได้ Model ที่ค่า r-square มาก และตัวแปรต้นได้ค่า p-value น้อยเพราะแค่ข้อมูลมีความคล้ายกัน แต่ในความเป็นจริงไม่ได้เกี่ยวข้องกันก็ได้

Part 2

กระทรวงที่ 1 ที่ทางกลุ่มได้รับมอบหมาย เป็นตัวแปรต้นที่ส่งผล ของโมเดลทำนายค่าใช้จ่ายรัฐบาลตามลักษณะงาน อย่างน้อย 2 ด้าน โดยที่ค่า r-square ของทั้ง 2 ด้าน ไม่ต่ำกว่า 0.9 ระบุด้านดังกล่าว มา 2 ด้าน (แสดงวิธีทำการได้มาของด้านดังกล่าวในรายงานด้วย)

กระทรวงที่ได้รับมอบหมายกระทรวงที่ 1 คือกระทรวงศึกษาธิการ

โดยเริ่มต้นได้ทำการหาค่า correlation ของทุกข้อมูลกับกระทรวงศึกษาธิกาแล้วทำการ sort เพื่อดูผลลัพธ์ ได้ผลลัพธ์ออกมาดังนี้

	กระทรวงศึกษาธิการ	การศึกษา	การศาสนา วัฒนธรรม และนันทนาการ	การสาธารณสุข	การป้องกันประเทศ	การรักษาความสงบภายใน	ปี	การเศรษฐกิจ	การสังคมสงเคราะห์
ปี	0.866009	0.919007	0.883614	0.980123	0.964086	0.995657	1.000000	0.943751	0.953363
การป้องกันประเทศ	0.893821	0.950833	0.899601	0.973354	1.000000	0.972224	0.964086	0.862016	0.879568
การรักษาความสงบภายใน	0.875281	0.931177	0.890448	0.981021	0.972224	1.000000	0.995657	0.942879	0.949667
การเศรษฐกิจ	0.783286	0.834743	0.837611	0.909736	0.862016	0.942879	0.943751	1.000000	0.943968
การสาธารณสุข	0.936495	0.965081	0.935280	1.000000	0.973354	0.981021	0.980123	0.909736	0.892718
การศาสนา วัฒนธรรม และนันทนาการ	0.940386	0.966511	1.000000	0.935280	0.899601	0.890448	0.883614	0.837611	0.746810
การศึกษา	0.973864	1.000000	0.966511	0.965081	0.950833	0.931177	0.919007	0.834743	0.781584
การสังคมสงเคราะห์	0.703469	0.781584	0.746810	0.892718	0.879568	0.949667	0.953363	0.943968	1.000000
กระทรวงศึกษาธิการ	1.000000	0.973864	0.940386	0.936495	0.893821	0.875281	0.866009	0.783286	0.703469

จะสังเกตุได้ว่า ด้านที่มี correlation กับกระทรวงศึกษาธิการสูงสุด 3 อันดับคือ
1.ด้านการศึกษา
2.ด้านการศาสนา วัฒนธรรมและนันนทาการ
3.ด้านการสาธารณสุข

จากนั้นจึงลองเอา กระทรวงศึกษาเป็นตัวแปรต้น ใน Model ที่จะทำนายค่าใช้จ่ายรัฐบาลด้านการศึกษาได้ผลลัพธ์ตามนี้

OLS Regression Results

Dep. Variable:	การศึกษา	R-squared:	0.948
Model:	OLS	Adj. R-squared:	0.945
Method:	Least Squares	F-statistic:	275.8
Date:	Fri, 04 Mar 2022	Prob (F-statistic):	4.58e-11
Time:	14:09:51	Log-Likelihood:	-193.92
No. Observations:	17	AIC:	391.8
Df Residuals:	15	BIC:	393.5
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
กระทรวงศึกษาธิการ	0.8437	0.051	16.606	0.000	0.735	0.952
const	1.097e+05	2.03e+04	5.398	0.000	6.64e+04	1.53e+05

Omnibus:	23.775	Durbin-Watson:	1.234
Prob(Omnibus):	0.000	Jarque-Bera (JB):	30.341
Skew:	2.204	Prob(JB):	2.58e-07
Kurtosis:	7.837	Cond. No.	1.45e+06

Model นี้ได้ค่า R-square ได้เท่ากับ 0.948 ดังนั้นจึงได้ด้านแรกคือด้านการศึกษาเป็นที่เรียบร้อย จากนั้นจึงได้ลองนำ ด้านการศาสนา วัฒนธรรมและนันนทาการ , ด้านการสาธารณสุข มาลองสร้าง linear model ได้ผลลัพธ์ว่าค่า R-square ออกมาประมาณ 0.8 กว่าๆ ทั้งสอง Model โดยที่ ด้านสาธารณสุข มีค่า R-square มากกว่าด้านการศาสนา วัฒนธรรมและนันนทาการ จึงได้ลองนำด้านที่เหลือทั้งหมดมาลองทำ train test split 10 รอบ เพื่อทดลองหาค่า Average R-square แต่ก็ไม่มีด้านไหนมีค่า Average R-square ถึง 0.9 (ตัวอย่างของการหาค่า Average R-square ของแต่ล่ะด้าน)

การป้องกันประเทศ
r2_score round1: 0.9029520227507707
r2_score round2: 0.20041461612551337
r2_score round3: 0.9080695996477743
r2_score round4: 0.08950820869084519
r2_score round5: 0.8918636709500172
r2_score round6: 0.5877133003959973
r2_score round7: 0.9433205732380772
r2_score round8: -6.40756722536356
r2_score round9: 0.789427330351752
r2_score round10: 0.7614377643312862
Avg r2_score: -0.033286013888152687
การรักษาความสงบภายใน
r2_score round1: 0.36907622508586746
r2_score round2: 0.38167589798973933
r2_score round3: 0.5247853346708278
r2_score round4: -0.18013998294197564
r2_score round5: -0.9546553614827415
r2_score round6: 0.9698303720376452
r2_score round7: -0.8277874303045172
r2_score round8: 0.7802279907053775
r2_score round9: 0.9324133349494027
r2_score round10: 0.4699491385238235
Avg r2_score: 0.2465375519233449

การเศรษฐกิจ
r2_score round1: 0.06338157024115232
r2_score round2: -1.0583077666586722
r2_score round3: 0.9018490528370994
r2_score round4: 0.6748645310669101
r2_score round5: 0.7756019708523727
r2_score round6: 0.937806760046148
r2_score round7: 0.7638439508671222
r2_score round8: -3.539151389804865
r2_score round9: 0.8027962727691693
r2_score round10: 0.7699871299373273
Avg r2_score: 0.10926720821537639
การสาธารณสุข
r2_score round1: 0.9655294829177867
r2_score round2: 0.892780555833889
r2_score round3: 0.9453350181840113
r2_score round4: 0.679879053966032
r2_score round5: 0.9457141842347734
r2_score round6: 0.9440254257147301
r2_score round7: 0.2294412354662192
r2_score round8: 0.5707209497634791
r2_score round9: 0.8710198091220676
r2_score round10: 0.9655753885385402
Avg r2_score: 0.8010021103741529

จึงได้นำด้านการสาธารณสุข ซึ่งได้ R-square เป็นอันดับสองรองจากการศึกษามาสร้าง linear Model โดยรอบนี้ได้ใช้ตัวแปรต้น คือ กระทรวงศึกษา และ กระทรวงสาธารณสุข มาสร้าง Model ที่จะทำนายค่าใช้จ่ายรัฐบาลด้านการสาธารณสุขได้ผลลัพธ์ตามนี้

OLS Regression Results

Dep. Variable:	การสาธารณสุข	R-squared:	0.988
Model:	OLS	Adj. R-squared:	0.986
Method:	Least Squares	F-statistic:	556.4
Date:	Sat, 05 Mar 2022	Prob (F-statistic):	4.57e-14
Time:	10:35:45	Log-Likelihood:	-177.63
No. Observations:	17	AIC:	361.3
Df Residuals:	14	BIC:	363.8
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
กระทรวงศึกษาธิการ	0.2078	0.043	4.811	0.000	0.115	0.300
กระทรวงสาธารณสุข	1.6666	0.149	11.161	0.000	1.346	1.987
const	-2.226e+04	8113.480	-2.743	0.016	-3.97e+04	-4853.822

Omnibus:	0.509	Durbin-Watson:	1.178
Prob(Omnibus):	0.775	Jarque-Bera (JB):	0.587
Skew:	-0.204	Prob(JB):	0.746
Kurtosis:	2.186	Cond. No.	1.50e+06

Model นี้ได้ค่า R-square ได้เท่ากับ 0.988 ดังนั้นจึงได้ด้านที่สองคือด้านการสาธารณสุขเป็นที่เรียบร้อย

สรุป ได้ด้านที่เป็นคำตอบ คือด้านการศึกษาและด้านสาธารณสุข โดย Model ที่จะทำนายค่าใช้จ่ายรัฐบาลด้านการศึกษาได้ค่า R-square เท่ากับ 0.948 และ Model ที่จะทำนายค่าใช้จ่ายรัฐบาลด้านการสาธารณสุขได้ค่า R-square เท่ากับ 0.988

Part 3

จากโมเดลของด้านที่ได้รับมอบหมาย สร้าง Nonlinear regression 3 โมเดล ได้แก่ Logarithm Function, Exponential function และ Power function โดยเป็นโมเดลที่ทางกลุ่มเลือกจาก Part 1.1-1.2 ที่ทำนายค่าใช้จ่ายรัฐบาลด้านการศาสนา วัฒนธรรมและนันนทาการ โดยที่จำนวนตัวแปรต้นไม่ต่ำกว่า 3 ตัว

<u>คำแนะนำ</u> ใช้ log ของ numpy library ช่วยในการใส่ค่า log ใน dataframe และอาจต้องปรับใช้ normalization ที่ให้ค่าไม่เป็นศูนย์ เพราะ log 0 = undefined.

1. เปรียบเทียบ Error ของ 3 Nonlinear models และ Linear model ตามตารางด้านล่าง อภิปราย (discuss) ลักษณะของ Error ทั้ง 4 ค่า ว่าเป็นอย่างไร ค่ามากหรือน้อย สัมพันธ์กับ R^2 หรือไม่ แต่ละค่าช่วยในการดูประสิทธิภาพของโมเดลอย่างไร หากต้องเลือกมา 1 ค่า จะเลือกค่าใด

	R^2	MSE	RMSE	MAD	MAPE
Linear	0.94385	1089369.93	1029.07	870.92	0.06112
Logarithm	0.95125	1127207.00	1043.59	852.13	0.07459
Exponential	0.54061	12237503.20	3067.65	2258.38	0.12981
Power	0.91099	1480164.23	1170.85	1013.06	0.07252

แต่ละค่าช่วยในการดูประสิทธิภาพของโมเดล ในรูปแบบที่แตกต่างกัน ดังนี้

R^2 บ่งบอกว่า model ของเราสามารถอธิบายข้อมูลได้ถูกต้องมากน้อยแค่ไหน และค่า |r| ยังสามารถบ่งบอกถึงความสัมพันธ์ในระดับ strong correlation หรือ weak correlation ได้อีกด้วย

MSE(Mean Square Error) มาจากสูตร $\frac{1}{n}\sum_{i=0}^{n}(y_{t_i} - \hat{y}_{t_i})^2$ กล่าวคือการเอาผลรวมของ ค่า y ที่เกิดขึ้นจริงลบกับค่า y ที่ model ทำนายออกมาแล้วยกกำลังสอง และหารด้วยจะนวนข้อมูล เอาไว้บ่งบอกค่า error ที่เกิดขึ้นจากการเทียบข้อมูลจริงกับข้อมูลที่ทำนายออกมาโดยไม่ค่อยนิยมใช้กันนักเนื่องจากค่า error ยกกำลัง 2 จึงจะนิยมใช้ค่า RMSE มากกว่า แต่หากมีข้อมูลที่ error ค่าเยอะอยู่ในข้อมูล MSE จะขยายค่าของความผิดพลาดนั้น ด้วยกำลังสองจึงทำให้ ดูเหมือนกับว่า model มีจำนวน error เยอะแต่จริงๆ แล้วอาจจะมีแค่ไม่กี่ตัวที่เกิด error

RMSE(Root Mean Square Error) มาจากสูตร $\sqrt{MSE}$ กล่าวคือการเอาค่า MSE มาถอดรูทจะนิยมใช้กันมากกว่าเนื่องจากสามารถพิจารณาค่าได้ง่ายกว่า สามารถมองเป็น จำนวนของสิ่งนั้นๆ ได้ตรงตัว

มากกว่า เช่น มองเป็นจำนวนคน

MAD (MAE) หรือเรียกว่า Mean Absolute Deviation (Error) มาจากสูตร $\frac{1}{n}\sum_{i=0}^{n}|y_{t_i}-\hat{y}_{t_i}|$ หากเกิดปัญหาที่มีข้อมูลที่ error ค่าเยอะอยู่ในข้อมูลแต่จริงๆแล้วอาจจะมีแค่ไม่กี่ตัวที่เกิด error MAD จะสามารถให้ค่าที่สมเหตุสมผลกับข้อมูลได้มากกว่า MSE โดยข้อดีคือสามารถอธิบายได้ง่าย

MAPE (Mean Absolute Percent Error) มาจากสูตร $\frac{1}{n}\sum_{i=0}^{n}\frac{|y_{t_i}-\hat{y}_{t_i}|}{y_{t_i}}\times 100$ บ่งบอกว่า model ทำนายไกลไปแค่ไหนจากค่าจริง โดยการทำเป็นเปอร์เซ็นต์ เป็นค่าทางสถิติที่เหมาะกับการอธิบายให้คนที่ไม่มีพื้นฐานด้านสถิติเข้าใจได้ง่าย เพราะเป็นการมองค่า error ในรูปแบบเปอร์เซ็นต์ แต่มีข้อเสียคือ ถ้าค่าจริงที่เอามาหารมีค่าเท่ากับ 0 จะไม่สามารถหา MAPE ได้

ส่วนมากนิยมใช้ค่า MSE , RMSE แต่หากมี outlier หรือ มี error ที่มีค่าสูง จะทำให้ค่า MSE , RMSE สูงตามไปด้วย

โดยในการอธิบาย Nonlinear models จะอธิบายโดยการเปรียบเทียบกับ linear model ดังนี้

Logarithm model เมื่อเทียบกับ linear model มีค่า R-square เพิ่มขึ้นซึ่งบ่งบอกว่าสามารถอธิบายข้อมูลได้มากขึ้น (ส่งผลดี)แต่ในส่วนของค่า MSE , RMSE นั้นมีค่าเพิ่มขึ้นบ่งบอกว่ามี error ที่เพิ่มขึ้นด้วย (ส่งผลแย่) เช่นเดียวกับค่า MAPE ที่มีค่าเพิ่มขึ้นบ่งบอกว่ามีเปอร์เซ็นต์ความผิดพลาดที่เพิ่มขึ้นด้วย (ส่งผลแย่) แต่ ค่า MAD มีค่าลดลงบ่งบอกว่ามีข้อมูลส่วนน้อยที่เกิด error หรือ error ลดลง (ส่งผลดี) ดังนั้นจึงสรุปว่าใน Logarithm model ค่าที่สัมพันธ์กับค่า R-square มีแค่ ค่า MAD ที่ส่งผลลัพธ์ไปในทางเดียวกัน

Exponential model และ Power model เมื่อเทียบกับ linear model มีค่า R-square ลดลงซึ่งบ่งบอกว่าสามารถอธิบายข้อมูลได้น้อยลง (ส่งผลแย่)และในส่วนของค่า MSE , RMSE,MAD นั้นมีค่าเพิ่มขึ้นบ่งบอกว่ามี error ที่เพิ่มขึ้นด้วย (ส่งผลแย่) เช่นเดียวกับค่า MAPE ที่มีค่าเพิ่มขึ้นบ่งบอกว่ามีเปอร์เซ็นต์ความผิดพลาดที่เพิ่มขึ้นด้วย (ส่งผลแย่) ดังนั้นจึงสรุปว่าใน Exponential model และ Power model ค่าที่สัมพันธ์กับค่า R-square คือ ค่า MSE,RMSE,MAD และ MAPE ที่ส่งผลลัพธ์ไปในทางเดียวกัน

โดยหากต้องเลือกใช้เพียง 1 ค่าจะเลือกค่า MAD เพราะมีความสมเหตุสมผลกับค่า R-square และค่า error ต่างๆที่เกิดขึ้น และยังสามารถอธิบายให้ผู้อื่นเข้าใจได้ง่ายอีกด้วย

2.อภิปราย ประสิทธิภาพของโมเดลว่า โมเดลใดดีที่สุด ให้ค่าความผิดพลาดน้อยที่สุด หากต้องเลือกโมเดลมาแทน Linear จะใช้โมเดลใด เพราะเหตุใด

จากการดูค่า R-square,MSE,RMSE,MAD และ MAPE คิดว่า linear model ดีที่สุด และให้ค่าความผิดพลาดน้อยที่สุด โดยหากต้องเลือกโมเดลมาแทน Linear Model จะเลือกใช้ Logarithm Model เนื่องจาก ค่า R-square,MSE,RMSE,MAD และ MAPE มีความใกล้เคียงกับ Linear Model มากกว่า Exponential model และ Power model และให้ค่าความผิดพลาดที่ต่างจาก Linear Model เพียงเล็กน้อย

Part 4

สร้างโมเดล โดยใช้ข้อมูลตัวแปรต้นหรือ input 3 ตัว จากสำนักงานสถิติแห่งชาติ (http://statbbi.nso.go.th/staticreport/page/sector/th/index.aspx) เพื่อนำไปทำนายงบประมาณรายจ่ายกระทรวงที่ 3

หมายเหตุ ไม่จำเป็นต้องใช้ข้อมูลงบประมาณรายจ่ายกระทรวงทุกปีที่ให้ไป เนื่องจากข้อมูลตัวแปรต้นอาจมีไม่ครบทุกปี

ให้เหตุผลในการเลือกตัวแปรต้นดังกล่าว

- วิเคราะห์โมเดลที่ได้ ประเมินว่า โมเดลนี้เป็นโมเดลที่ดีหรือไม่
- หากมีข้อมูลตัวแปรต้นที่ส่งผลต่องบประมาณรายจ่าย วิเคราะห์ว่า เหตุใดตัวแปรต้น จึงส่งผล และส่งผลต่อ งบประมาณรายจ่ายรวมของกระทรวง มากหรือน้อย
- หากมีข้อมูลตัวแปรต้นที่ไม่ส่งผลต่องบประมาณรายจ่าย วิเคราะห์ว่า เหตุใดตัวแปรต้น จึงไม่ส่งผล

คำแนะนำ ไม่ควรดูแต่สัมประสิทธิ์ ควรหาเหตุผลมาประกอบการอภิปราย เช่น ดูรายละเอียดของงบประมาณรายจ่ายกระทรวง และโครงสร้างของกระทรวงนี้เพิ่มเติม

ตัวแปรต้นที่เลือกมามี 4 ตัวคือ จำนวนพระภิกษุ , จำนวนสามเณร , จำนวนวัด , งบสำนักพุทธ โดยเป็นข้อมูลตั้งแต่ปี พ.ศ. 2552 ถึง ปี พ.ศ. 2561 ดังนี้

	ปี	จำนวนพระภิกษุ	จำนวนสามเณร	จำนวนวัด	งบสำนักพุทธ
0	2552	267939.0	65937.0	35616.0	3449.27
1	2553	291116.0	70408.0	37100.0	3588.75
2	2554	290331.0	62478.0	37084.0	3788.64
3	2555	293879.0	61416.0	37322.0	4329.51
4	2556	289131.0	60528.0	37734.0	4821.77
5	2557	290015.0	58418.0	39277.0	5362.03
6	2558	298580.0	59587.0	39883.0	5121.30
7	2559	289334.0	58426.0	40772.0	5360.19
8	2560	267848.0	49152.0	41142.0	5054.93
9	2561	281058.0	44430.0	41252.0	5020.29

โดยเหตุผลในการเลือกตัวแปรต้นมีดังนี้

1.จำนวนพระภิกษุ เหตุผลคือ กระทรวงวัฒนธรรมมีภารกิจเกี่ยวกับการดำเนินงานของรัฐด้านศาสนา โดยการทำนุบำรุง ส่งเสริมและให้ความอุปถัมภ์คุ้มครองกิจการด้านพุทธศาสนาและศาสนาอื่นๆ ที่ทางราชการรับรองดังนั้นจำนวนของพระภิกษุน่าจะส่งผลต่อการใช้งบในกิจของสงฆ์ ต่างๆ ดังนั้น ยิ่งจำนวนพระภิกษุมากน่าจะมีการใช้จ่ายงบในกระทรวงวัฒนธรรมเพิ่มขึ้นมากตามไปด้วย

2.จำนวนสามเณร เหตุผลคือ กระทรวงวัฒนธรรมมีภารกิจเกี่ยวกับการดำเนินงานของรัฐด้านศาสนา โดยการทำนุบำรุง ส่งเสริมและให้ความอุปถัมภ์คุ้มครองกิจการด้านพุทธศาสนาและศาสนาอื่นๆ ที่ทางราชการรับรองดังนั้นจำนวนสามเณรน่าจะส่งผลต่อการใช้งบในกิจของสงฆ์ ต่างๆ ดังนั้น ยิ่งจำนวนสามเณรมากน่าจะมีการใช้จ่ายงบในกระทรวงวัฒนธรรมเพิ่มขึ้นมากตามไปด้วย

3.จำนวนวัด เหตุผลคือ กระทรวงวัฒนธรรมมีภารกิจเกี่ยวกับการดำเนินงานของรัฐด้านศาสนา โดยการทำนุบำรุง ส่งเสริมและให้ความอุปถัมภ์คุ้มครองกิจการด้านพุทธศาสนาและศาสนาอื่นๆ ที่ทางราชการรับรองดังนั้นจำนวนวัดน่าจะส่งผลต่อการใช้งบในการบูรณะวัดหากมีการเสื่อมโทรมเกิดขึ้นหรือหากมีการ

ดำเนินกิจการภายในวัดเกิดขึ้น ดังนั้น ยิ่งจำนวนวัดมากน่าจะมีการใช้จ่ายงบในกระทรวงวัฒนธรรมเพิ่มขึ้นมากตามไปด้วย

4. งบสำนักพุทธ เหตุผลคือ สำนักพุทธเป็นหน่วยงานที่ทำงานเกี่ยวกับศาสนา เช่นเดียวกับกรมการศาสนาที่สังกัดอยู่ใน กระทรวงวัฒนธรรมดังนั้นคิดว่างบของสำนักพุทธน่าจะมีความเกี่ยวข้องกับกระทรวงวัฒนธรรม

หลังจากได้ลองนำตัวแปรต้นทั้ง 4 ตัวมาสร้าง linear model เพื่อดูค่า P-value ของตัวแปรแต่ละตัว ได้ผลลัพธ์ออกมาดังนี้

OLS Regression Results

Dep. Variable:	y	R-squared:	0.948
Model:	OLS	Adj. R-squared:	0.907
Method:	Least Squares	F-statistic:	22.94
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.00205
Time:	13:23:18	Log-Likelihood:	-70.513
No. Observations:	10	AIC:	151.0
Df Residuals:	5	BIC:	152.5
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
จำนวนพระภิกษุ	-0.0074	0.016	-0.458	0.666	-0.049	0.034
จำนวนสามเณร	-0.0429	0.037	-1.164	0.297	-0.138	0.052
จำนวนวัด	0.2791	0.171	1.633	0.163	-0.160	0.718
งบสำนักพุทธ	0.6276	0.381	1.646	0.161	-0.353	1.608
const	-2769.1283	7473.436	-0.371	0.726	-2.2e+04	1.64e+04

Omnibus:	1.576	Durbin-Watson:	2.527
Prob(Omnibus):	0.455	Jarque-Bera (JB):	0.757
Skew:	0.090	Prob(JB):	0.685
Kurtosis:	1.664	Cond. No.	1.76e+07

กำหนดให้ null hypothesis(H0) คือไม่ขึ้นกับค่านั้น(slope b=0) และ alternative hypothesis(Ha) คือขึ้นกับค่านั้น(slope b!=0) โดยจะกำหนดให้ significance level ที่ 0.1 ดังนี้ถ้า p value น้อยกว่า 0.1 จะถือว่า reject null โดยค่า p value ของแต่ละตัวแปรต้นมีดังนี้ (ตัดที่ทศนิยม 3 ตำแหน่ง)

จำนวนพระภิกษุ มีค่า p value เท่ากับ 0.666

จำนวนสามเณร มีค่า p value เท่ากับ 0.297

จำนวนวัด มีค่า p value เท่ากับ 0.163

งบสำนักพุทธ มีค่า p value เท่ากับ 0.161

ดังนั้นจะได้ว่า ตัวแปรต้นทุกตัวไม่มีตัวไหนเลย reject null ที่ significance level ที่ 0.1 แสดงว่าตัวแปรต้นที่เลือกมาทั้ง 4 ตัวนี้ไม่เกี่ยวข้องกับรายจ่ายของกระทรวงวัฒนธรรมเลย จึงได้ลองนำตัวแปรต้นที่มีค่า P-value มากที่สุดออก ซึ่งก็คือ จำนวนพระภิกษุ และได้ทำการสร้าง linear model เพื่อดูผลลัพธ์ใหม่อีกครั้งได้ผลลัพธ์ดังนี้

OLS Regression Results

Dep. Variable:	y	R-squared:	0.946
Model:	OLS	Adj. R-squared:	0.919
Method:	Least Squares	F-statistic:	35.15
Date:	Tue, 08 Mar 2022	Prob (F-statistic):	0.000334
Time:	13:23:40	Log-Likelihood:	-70.718
No. Observations:	10	AIC:	149.4
Df Residuals:	6	BIC:	150.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
จำนวนสามเณร	-0.0517	0.029	-1.759	0.129	-0.124	0.020
จำนวนวัด	0.2753	0.159	1.731	0.134	-0.114	0.664
งบสำนักพุทธ	0.5580	0.326	1.713	0.138	-0.239	1.355
const	-3893.7770	6576.895	-0.592	0.575	-2e+04	1.22e+04

Omnibus:	1.325	Durbin-Watson:	2.270
Prob(Omnibus):	0.515	Jarque-Bera (JB):	0.927
Skew:	0.505	Prob(JB):	0.629
Kurtosis:	1.902	Cond. No.	4.01e+06

โดยรอบนี้กำหนดให้ null hypothesis(H0) คือไม่ขึ้นกับค่านั้น(slope b=0) และ alternative hypothesis(Ha) คือขึ้นกับค่านั้น(slope b!=0) โดยจะกำหนดให้ significance level ที่ 0.1 ดังนี้ถ้า p value น้อยกว่า 0.1 จะถือว่า reject null โดยค่า p value ของแต่ละตัวแปรต้นมีดังนี้ (ตัดที่ทศนิยม 3 ตำแหน่ง)

จำนวนสามเณร มีค่า p value เท่ากับ 0.129

จำนวนวัด มีค่า p value เท่ากับ 0.134

งบสำนักพุทธ มีค่า p value เท่ากับ 0.138

ดังนั้นจะได้ว่า ตัวแปรต้นทุกตัวไม่มีตัวไหนเลย reject null ที่ significance level ที่ 0.1 แสดงว่าตัวแปรต้นที่เลือกมาทั้ง 3 ตัวนี้ไม่เกี่ยวข้องกับรายจ่ายของกระทรวงวัฒนธรรมเลย **จึงได้ลองนำข้อมูลมาทำ** train test split 5 **รอบ เพื่อทดลองหาค่า** Average R-square **ได้ผลลัพธ์ออกมาดังนี้**

```
----- Round 1 -----
r2: -0.3918951622590785
MSE: 135305.25956200363
RMSE: 367.8386325034439
MAD: 323.7275345436589
MAPE: 0.04607146317457855
----- Round 2 -----
r2: 0.8436038514533466
MSE: 317766.09788354597
RMSE: 563.7074577150332
MAD: 551.0326682986997
MAPE: 0.08777506617746268
----- Round 3 -----
r2: 0.8620483216853311
MSE: 320122.3052089909
RMSE: 565.7935181751297
MAD: 535.4189459224754
MAPE: 0.09663647348899242
----- Round 4 -----
r2: 0.5276826604247895
MSE: 1012499.0238712133
RMSE: 1006.2301048324947
MAD: 906.5619605250835
MAPE: 0.1330676055980736
----- Round 5 -----
r2: 0.5191086707359571
MSE: 1137458.763955423
RMSE: 1066.5171184539997
MAD: 915.0353858475115
MAPE: 0.17099231825314365
----- AVG -----
avg r2: 0.47210966840806917
avg MSE: 584630.2900962352
avg RMSE: 714.0173663360204
avg MAD: 646.3552990274858
avg MAPE: 0.10690858533845018
```

	R^2	MSE	RMSE	MAD	MAPE
Linear	0.47211	584630.29	714.02	646.36	0.10691

จะสังเกตุได้ว่าค่า R-squared ของ model จะอยู่ที่ 0.47211 หรือสามารถบอกได้ว่า 47.211 เปอร์เซ็นต์ของข้อมูลเราสามารถอธิบายได้ด้วย linear model นี้และสามารถดูได้ว่าค่า |r| = 0.687 ซึ่งอยู่ในช่วง 0.5 < |r| < 0.8 บ่งบอกว่ามีความสัมพันธ์ในระดับปานกลาง แสดงว่าเป็น linear model ที่ไม่ดีเท่าไหร่นักแต่ก็ไม่ได้แย่มากเช่นกัน

เนื่องจากตัวแปรต้นทั้งสามตัวไม่ได้เกี่ยวข้องกับงบประมาณรายจ่ายของกระทรวงวัฒนธรรมเลยจึงมาวิเคราะห์ว่าเหตุใดตัวแปรทั้งสามตัวจึงไม่เกี่ยวข้องกับงบประมาณรายจ่ายของกระทรวงวัฒนธรรมดังนี้

1.จำนวนสามเณร จากการวิเคราะห์ สามเณรไม่น่าส่งผลต่อรายจ่ายของกระทรวงวัฒนธรรมมากขนาดนั้นเพราะว่า การใช้ชีวิตส่วนใหญ่ ไม่ได้ใช้เงินจากกระทรวงวัฒนธรรม เช่น การบิณฑบาตรเพื่อรับประทานอาหารมือเช้า , มีเจ้าภาพที่ศรัทธาในพุทธศาสนาจัดมื้ออาหารกลางวันให้ ,ส่วนมื้อเย็นก็ไม่ได้มีการรับประทานอาหาร และกิจกรรมในวัดส่วนใหญ่ก็ไม่ได้มีการใช้เงิน จึงน่าจะเป็นเหตุผลให้ จำนวนสามเณรไม่ส่งผลต่องบประมาณรายจ่ายของกระทรวงวัฒนธรรม

2.จำนวนวัด จากการวิเคราะห์ คือหากมองค่าใช้จ่ายของวัด เช่น ค่าน้ำ , ค่าไฟ น่าจะมีเงินส่วนหนึ่งจากผู้ที่ศรัทธาในพุทธศาสนา ได้บริจาคมาให้วัดจำนวนหนึ่ง อีกทั้งยังมีรายได้จากผู้ที่ใช้พื้นที่ของวัดทำกิจการ หรือจะเป็นการซ่อมบูรณวัด , ต่อเติมวัด ก็จะมีเงินมาจากผู้ที่ศรัทธา หรือ ได้รับเงินบางส่วนมาจากสำนักพุทธ ดังนั้น จำนวนวัดจึงไม่ส่งผลต่องบประมาณรายจ่ายของกระทรวงวัฒนธรรม

3.งบสำนักพุทธ จากการวิเคราะห์ว่าถึงแม้งานหลักๆจะทำงานเกี่ยวกับศาสนาเหมือนกันแต่ในความเป็นจริงแล้ว ในปี พ.ศ. 2545 ได้มีการปรับปรุงโครงสร้างกระทรวง มีการแบ่งส่วนราชการของกรมการศาสนาเดิม ออกเป็น 2 ส่วน คือ กรมการศาสนา สังกัดกระทรวงวัฒนธรรม และสำนักงานพระพุทธศาสนาแห่งชาติ เป็นหน่วยงานขึ้นตรงต่อนายกรัฐมนตรี กล่าวได้ว่า สำนักพุทธไม่ได้อยู่ภายใต้กระทรวงวัฒนธรรม แต่แยกออกมาเป็นหน่วยงานขึ้นตรงต่อนายกรัฐมนตรี ดังนั้น งบสำนักพุทธจึงไม่ส่งผลต่องบประมาณรายจ่ายของกระทรวงวัฒนธรรม (อ้างอิงจาก https://th.wikipedia.org/wiki/สำนักงานพระพุทธศาสนาแห่งชาติ)

Summary Part

จากการวิเคราะห์ผ่าน 2 วิธีการเรียนรู้ สรุปสิ่งที่ได้จากการศึกษางบประมาณในงานนี้

สรุปสิ่งที่ได้เรียนรู้จาก Clustering Part มีดังนี้

โดยใน Part 1 ได้เรียนรู้เกี่ยวกับรูปแบบการทำ clustering ต่างๆ คือ K-mean , Hierarchical , DBSCAN และเปรียบเทียบถึงผลลัพธ์ที่เกิดขึ้นจากการแบ่งกลุ่มด้วยวิธีต่างๆ ได้ลองพิจารณาถึงข้อดีข้อเสียของวิธีต่างๆ โดยได้ผลลัพธ์ออกมาว่า วิธี K-mean และ Hierarchical ให้ผลลัพธ์การแบ่งกลุ่มที่ใกล้เคียงกัน แต่ DBSCAN ให้ผลลัพธ์การแบ่งกลุ่มที่ต่างจากสองวิธีแรกโดยเมื่ออิงจากค่า Silhouette index จะได้ว่าทั้งสามวิธีมีค่าใกล้เคียงกัน แต่หากพิจารณา Davies-Bouldin index จะได้ว่าวิธี Hierarchical ให้ผลลัพธ์ดีที่สุด และวิธี DBSCAN ให้ผลลัพธ์แย่ที่สุด โดยต่อมาในส่วนของ

ในส่วนของ Part 2 ได้ลองจัดกลุ่มข้อมูลในช่วงเวลาที่ต่างกันคือตั้งแต่ปี พ.ศ. 2560 - 2564 โดยใช้วิธี K-mean ในการแบ่งกลุ่ม และจากการวิเคราะห์พบว่า เมื่อได้ลองจัดกลุ่มข้อมูลในแต่ล่ะปีะจะเห็นรูปแบบการแบ่งกลุ่มที่คล้ายๆกัน หรือ บางปีที่มีการจัดกลุ่มที่แตกต่างออกไป โดยได้ผลลัพธ์ว่าการจัดกลุ่มกระทรวงตามงบประมาณรายจ่าย แยกตามปี ด้วยวิธี k-mean ในปี พ.ศ. 2560 , 2561, 2562 จะได้ผลลัพธ์ในรูปแบบคล้ายกัน คือได้จำนวน 6 กลุ่ม และสมาชิกกระทรวงในแต่ล่ะกลุ่มเหมือนกัน คือ กลุ่ม1 (กระทรวงคมนาคม), กลุ่ม2 (กระทรวงการคลัง),กลุ่ม3 (กระทรวงศึกษาธิการ),กลุ่ม4 (กระทรวงมหาดไทย),กลุ่ม5 (กระทรวงกลาโหม , กระทรวงเกษตรและสหกรณ์ , กระทรวงสาธารณสุข),กลุ่ม6(กระทรวงอื่ๆที่เหลืออีก 13 กระทรวง) แต่ในการจัดกลุ่มกระทรวงตามงบประมาณรายจ่าย แยกตามปี ด้วยวิธี k-mean ในปี พ.ศ. 2563 , 2564 จะได้ผลลัพธ์ในรูปแบบแตกต่างจาก 3 ปีแรก คือได้จำนวน 5 กลุ่ม โดยกระทรวงที่มีความเปลี่ยนแปลงคือ กระทรวงศึกษาธิการ ,กระทรวงคมนาคม,กระทรวงกลาโหม , กระทรวงเกษตรและสหกรณ์ , กระทรวงสาธารณสุข โดย 5 กระทรวงนี้ได้ถูกจัดกลุ่มใหม่ตามนี้ กลุ่ม1(กระทรวงกลาโหม ,กระทรวงศึกษาธิการ , กระทรวงสาธารณสุข),กลุ่ม2(กระทรวงเกษตรและสหกรณ์ ,กระทรวงคมนาคม) ส่วนกลุ่มที่เหลือจะถูกจัดกลุ่มตามเดิมดังนี้ กลุ่ม3(กระทรวงการคลัง) ,กลุ่ม4(กระทรวงมหาดไทย) ,กลุ่ม5(กระทรวงอื่ๆที่เหลืออีก 13 กระทรวง)

สรุปสิ่งที่ได้เรียนรู้จาก Regression Part มีดังนี้

โดยใน Part 1.1 ได้ลองสร้าง linear model โดยให้หากระทรวงตัวแปรต้นจากการทำการหาจาก correlation จากนั้นได้ลองหาค่า p-value เพื่อทำ Hypothesis testing หาว่ากระทรวงไหนเกี่ยวข้องหรือไม่เกี่ยวข้อง และ ได้นำกระทรวงที่เกี่ยวข้องไปทำการสร้าง model เพื่อจะนำไปเปรียบเทียบ Part 1.2 โดยใน Part 1.2 ได้ลองสร้าง linear model โดยให้หากระทรวงตัวแปรต้นจากการคัดเลือกของผู้จัดทำว่ากระทรวงใดมีแนวโน้มที่จะเกี่ยวข้องโดยดูจากหน้าที่ของกระทรวงนั้นๆ จากนั้นได้ลองหาค่า p-value เพื่อทำ Hypothesis testing หาว่ากระทรวงไหนเกี่ยวข้องหรือไม่เกี่ยวข้อง และ ได้นำกระทรวงที่เกี่ยวข้องไปทำการสร้าง model เพื่อมาเปรียบเทียบกับ model ของ Part 1.1 เพื่อดูผลลัพธ์ และจากการวิเคราะห์ได้ข้อสรุปออกมาว่ากระทรงที่ส่งผลจาก Part 1.1 คือ กระทรวงมหาดไทยและกระทรวงพาณิชย์ ส่วนกระทรวงที่ส่งผลจาก Part 1.2 คือ กระทรวงการท่องเที่ยวและกีฬา,กระทรวงวัฒนธรรม,กระทรวงการต่างประเทศ โดยคิดว่ากระทรวงจาก Part 1.2 สมเหตุสมผลกว่า เนื่องจาก ทางผู้จัดทำได้คัดเลือกมาจากความน่าจะเป็นที่จะเกี่ยวข้องกับด้านการศาสนา วัฒนธรรมและนันนทาการ โดยดูจากหน้าที่และลักษณะงานของกระทรวงต่างๆ ต่อจากของ Part 1.1 ที่ดูแค่ correlation ซึ่งอาจเป็นไปได้ว่าการที่ได้ Model ที่ค่า r-square มาก และตัวแปรต้นได้ค่า p-value น้อยเพราะแค่ข้อมูลมีความคล้ายกัน แต่ในความเป็นจริงไม่ได้เกี่ยวข้องกันก็ได้

ต่อมาในส่วนของ Part 2 ได้ลองนำกระทรวงที่ 1 ที่ได้รับมอบหมาย คือ กระทรวงศึกษา มาทำการสร้าง linear model เพื่อหาด้านที่เกี่ยวข้อง 2 ด้านโดยที่แต่ละด้านต้องมีค่า R-square มากกว่า 0.9 โดยใน Part นี้ได้เรียนรู้เกี่ยวกับวิธีการเลือกข้อมูลที่คาดว่าจะส่งผลให้ model มีค่า R-squar ที่ดีโดยลองวิธีต่างๆที่ไม่ได้มีแบบแผนแน่นอนในการหาเนื่องจากเราไม่มีทางรู้ว่าข้อมูลไหนจะส่งผลให้ค่า r-square เยอะ จึงต้องลองไปเรื่อยๆ ดังนั้น Part 2 จึงเหมือนการฝึกทัศนคติของผู้จัดทำว่าควรจะเลือกข้อมูลตัวไหนเพื่อให้ได้คำตอบออกมา โดยผลลัพธ์ที่ได้ออกมา 2 ด้านคือ ด้านการศึกษาและด้านการสาธารณสุข

ในส่วนของ Part 3 ได้ลองสร้าง Nonlinear regression 3 โมเดล ได้แก่ Logarithm Function, Exponential function และ Power function และพิจารณาถึงลักษณะของ Error ทั้ง 4 ค่า (MSE,RMSE, MAD และ MAPE)ว่าเป็นอย่างไร ค่าแต่ละตัวมีข้อดีข้อเสียแตกต่างกันอย่างไรและความสัมพันธ์ของแต่ละค่ากับ R-Square เป็นอย่างไร โดยจะนำค่าเหล่านี้มาพิจารณา Nonlinear regression 3 โมเดล ได้แก่ Logarithm Function, Exponential function และ Power function ว่าให้ผลลัพธ์อย่างไรเมื่อเทียบกับ linear model โดยจากการลองทำ Part 3 ได้ข้อสรุปว่า Logarithm Model ได้ค่า R-square , MSE,RMSE, MAD และ MAPE ใกล้เคียงกับ linear model มากที่สุด

และส่วนสุดท้ายใน Part 4 ได้ลองสร้าง linear model จากข้อมูลตัวแปรต้นที่ทางผู้จัดทำได้ไปเลือกข้อมูลมาเองจากความน่าจะเป็นที่ว่าข้อมูลตัวแปรต้นทั้ง 4 (จำนวนพระภิกษุ ,จำนวนสามเณร ,จำนวนวัด ,งบสำนักพุทธ) ที่ไปหามาน่าจะมีความเกี่ยวข้องกับรายจ่ายของกระทรวงวัฒนธรรม แต่ผลลัพธ์จากการสร้าง linear model เพื่อ หาค่า p-value และทำ Hypothesis testing หาว่าตัวแปรตัวไหนเกี่ยวข้องหรือไม่เกี่ยวข้องได้ผลลัพธ์ออกมาว่าตัวแปรทั้ง 4 ไม่ได้เกี่ยวข้องกับ รายจ่ายของกระทรวงวัฒนธรรม เลยดังนั้นสิ่งที่ได้เรียนรู้จาก Part 4 คือ เราควรจะดูสิ่งที่เกิดขึ้นจริงของตัวแปรแต่ละตัวให้ละเอียดก่อนนำมาใช้เป็นตัวแปรต้น เพราะบางทีตัวแปรที่เราได้เลือกและคิดว่ามีความเกี่ยวข้อง อาจไม่ได้ส่งผลต่อ output (รายจ่ายของกระทรวงวัฒนธรรม) มากขนาดนั้น