

DATA WAREHOUS AND DATA MINING

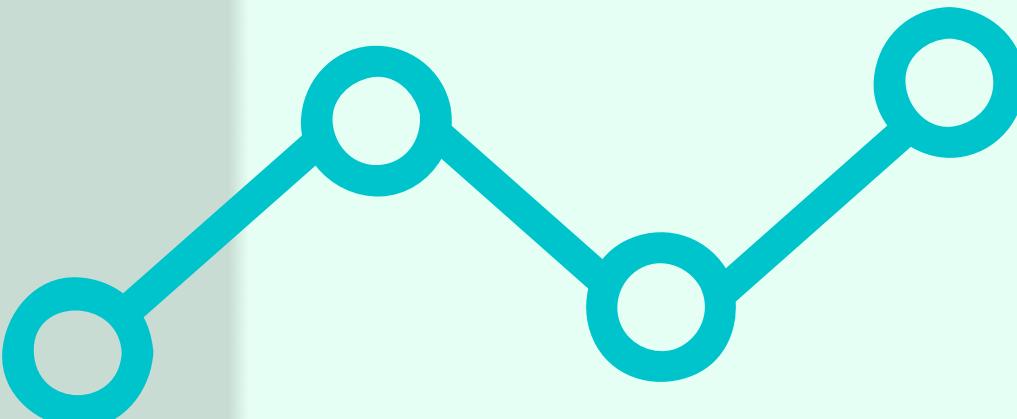
GROUP : מינהרכ



กลุ่มวิชาการได้ใช้ข้อมูลทั้งหมด 3 ชุดที่แตกต่างกันได้แก่

- 1.) ข้อมูลระบบกำจัดของเสีย แหล่งที่มา องค์กร : สำนักงานนโยบายและแผนทรัพยากรธรรมชาติและสิ่งแวดล้อม
- 2.) ข้อมูลค่าเฉลี่ยประชากร แหล่งที่มา องค์กร : สำนักงานเลขานุการสภาการศึกษา
- 3.) ข้อมูลรายได้ประชากร แหล่งที่มา องค์กร : สำนักงานปลัดกระทรวงการท่องเที่ยวและกีฬา

ข้อมูลทุกบุคคลอ้างอิงจาก : [HTTPS://DATA.GO.TH/](https://data.go.th/)





การนำข้อมูลเข้าเบื้องต้น



```
import pandas as pd #ทำงานกับข้อมูลลักษณะตาราง  
  
from google.colab import drive #เชื่อม Google drive  
drive.mount('/content/drive')  
  
import os # os ไว้จัดการไฟล์และค่าทาง อาทิคือที่อยู่ของไฟล์  
  
path = '/content/drive/My Drive/DWDM2021_DATA' #ชื่อ path นำไปไฟล์ที่เราต้องการ  
data_BB = pd.read_csv(os.path.join(path, 'กำจัดของเสีย.csv'))  
data_BB #ข้อมูลหลักชุดที่ 1
```



ข้อมูลหลักชุดที่ 1

ข้อมูลระบบกำจัดของเสีย

องค์กร : สำนักงานนโยบายและ แผนกรพยากรณ์ธรรมชาติและ สิ่งแวดล้อม

In []: path = '/content/drive/My Drive/DWDM2021_DATA' #ชื่อ path ชี้ไปไฟล์ที่เราต้องการ
data_BB = pd.read_csv(os.path.join(path,'ก้าจัดของเสีย.csv'))
data_BB #ข้อมูลหลักชุดที่ 1

Out[]:

	ที่	ภาค	จังหวัด	หน่วยงานที่รับ ผิดชอบ	ที่ดัง	ชื่อย่อ ของ ระบบ	ประเภท ของระบบ	พ.ศ. ที่ได้ รับงบ ประมาณ	พ.ศ. ที่ เปิดใช้ งาน	สถานภาพ ปัจจุบัน	ปริมาณน้ำ เสียที่รองรับ ได้ (ลบ.ม./วัน)	แหล่งงบประมาณ
0	1	เหนือ	เชียงราย	ทม.เชียงราย	ข้างสนามกีฬากลาง จังหวัด ต.สหมิด ต.รอบเวียง อ....	AL	ระบบสรช เดิมอาคต	2540	2544.0	เปิดเดินระบบ	22,700	กระทรวงมหาดไทย กระทรวง ทรัพยากรธรรมชาติและสิ่ง แวดล้อม
1	2	เหนือ	เชียงราย	ทต.จันจว�	เลขที่ 49 ม.2 ต.จัน จว. อ.แม่จัน จ.เชียงราย	WL	ระบบบึง ประดิษฐ์	2548	2550.0	เปิดเดินระบบ	50	กระทรวง ทรัพยากรธรรมชาติและสิ่ง แวดล้อม
2	3	เหนือ	เชียงราย	ทต.แม่สาย	บ้านเหมืองแดง ม.2 ต.แม่สาย อ.แม่สาย จ.เชียงราย	AL	ระบบสรช เดิมอาคต	2556	NaN	อยู่ระหว่าง การก่อสร้าง	6,500	กระทรวง ทรัพยากรธรรมชาติและสิ่ง แวดล้อม
3	4	เหนือ	เชียงใหม่	ทม.เชียงใหม่	หลังวัดท่าใหม่ อ.สัน ปิงหวาน อ.ทางดง จ.เชียงใหม่	AL	ระบบสรช เดิมอาคต	2545	2547.0	เปิดเดินระบบ	55,000	กระทรวงมหาดไทย กระทรวง ทรัพยากรธรรมชาติและสิ่ง แวดล้อม
4	5	เหนือ	เชียงใหม่	ทม.เชียงใหม่	อ.พากอง ต.ในเวียง อ.เมืองเชียงใหม่ จ.เชียงใหม่	SP	ระบบบ่อ ปรับเสถียร	2540	2544.0	เปิดเดินระบบ	8,259	กระทรวง ทรัพยากรธรรมชาติและสิ่ง แวดล้อม
...
89	90	ตะวัน ออก เชียง เหนือ	อุดรธานี	ทม.หนองส่า烘ง	เทศบาลนครอุดรธานี อ.เมืองอุดรธานี จ.อุดรธานี	WCS	ระบบ รวมรวมน้ำ เสีย	2559	2563.0	เปิดเดินระบบ	800	กระทรวง ทรัพยากรธรรมชาติและสิ่ง แวดล้อม
90	91	ตะวัน ออก เชียง เหนือ	อุบลราชธานี	ทม.อุบลราชธานี	ต.โพลังฤทธิ์ อ.เมือง อุบลราชธานี จ.อุบลราชธานี	OD	ระบบ คลองวัน เวียน	2546	2546.0	เปิดเดินระบบ	22,000	กระทรวงมหาดไทย กระทรวง ทรัพยากรธรรมชาติและสิ่ง แวดล้อม
91	92	ตะวัน ออก เชียง เหนือ	อุบลราชธานี	ทม.วารินชำราบ	บริเวณชุมชนเกตเเก้ว อ.วารินชำราบ จ.อุบลราชธานี	SP	ระบบบ่อ ปรับเสถียร	2541	2545.0	เปิดเดินระบบ	18,000	กระทรวง ทรัพยากรธรรมชาติและสิ่ง แวดล้อม
92	93	ตะวัน ออก เชียง เหนือ	ศรีสะเกษ	ทม.ศรีสะเกษ	สวนสักการะ ต.บุญนร ต.หนองครก อ.เมือง ศรีสะเกษ...	OD	ระบบ คลองวัน เวียน	2560	NaN	อยู่ระหว่าง การก่อสร้าง	12,600	กระทรวง ทรัพยากรธรรมชาติและสิ่ง แวดล้อม
93	94	ตะวัน ออก เชียง เหนือ	อุบลราชธานี	ทม.อุบลราชธานี	หนองพิพิพย์ ม. 4 ต.บุญนร อ.เมือง อุบลราชธานี จ.อุบล...	SP	ระบบบ่อ ปรับเสถียร	2545	2549.0	เปิดเดินระบบ	12,819	กระทรวง ทรัพยากรธรรมชาติและสิ่ง แวดล้อม

94 rows × 12 columns

ข้อมูลหลักชุดที่ 2

ข้อมูลค่าเฉลี่ยประชากร

องค์กร : สำนักงานเลขาริการสภาพการศึกษา

In []: `data_PP = pd.read_csv(os.path.join(path,'ค่าเฉลี่ยประชากร.csv'))
data_PP.dropna() #ข้อมูลรองชุดที่ 2`

Out[]:

	จังหวัด	15 - 39 ปี	40 - 59 ปี	15 - 59 ปี	15 ปีขึ้นไป	60 ปีขึ้นไป	ผลเฉลี่ย
0	กรุงเทพมหานคร	12.30	10.87	11.66	11.11	8.13	54.07
1	สมุทรปราการ	11.17	9.38	10.34	9.88	7.29	48.06
2	นนทบุรี	12.12	10.87	11.49	10.86	8.06	53.40
3	ปทุมธานี	11.33	10.20	10.84	10.31	7.08	49.76
4	อุบลราชธานี	11.11	9.04	10.26	9.32	5.47	45.20
...
72	ตรัง	10.16	8.31	9.29	8.45	5.06	41.27
73	พัทลุง	10.83	8.67	9.75	8.51	4.83	42.59
74	ปัตตานี	9.77	7.03	8.68	7.70	3.67	36.85
75	ยะลา	10.69	8.05	9.61	8.63	3.56	40.54
76	นราธิวาส	9.22	6.51	8.12	7.23	2.97	34.05

77 rows × 7 columns

ข้อมูลหลักชุดที่ 3

ข้อมูลรายได้ประชากรเฉลี่ย

องค์กร : สำนักงานปลัดกระทรวงการท่องเที่ยวและกีฬา

```
In [ ]: data_PC = pd.read_csv(os.path.join(path,'รายได้เฉลี่ย.csv'))  
data_PC.dropna() #ข้อมูลรองชุดที่ 3
```

Out[]:

	จังหวัด	รหัสจังหวัด	2555	2556	2557	2558	2559	2560	2561	รวมเฉลี่ย
0	กรุงเทพมหานคร	th-bm	570327.58	626535.05	634803.16	730318.31	822454.21	947946.30000	1051118.51	5.383503e+06
1	กาญจนบุรี	th-kn	11910.00	13417.00	14576.27	17659.54	20404.73	24440.77288	27138.35	1.295467e+05
2	ฉะเชิงเทรา	th-cc	4214.00	4569.00	4673.31	5239.98	5630.03	7628.47000	8469.45	4.042424e+04
3	ชลบุรี	th-cb	3571.69	3735.06	3781.62	3984.33	4240.41	4655.55000	5007.28	2.897594e+04
4
72	ระนอง	th-rn	2780.00	3250.00	3272.34	3623.68	3954.75	4311.07000	4706.13	2.589797e+04
73	สงขลา	th-sg	26702.00	37276.00	41304.80	47260.67	53493.02	59831.77000	68097.42	3.339657e+05
74	สตูล	th-sa	5332.00	5905.00	6162.46	6846.72	7438.85	8168.70000	9101.47	4.895520e+04
75	สุราษฎร์ธานี	th-st	35312.00	51550.00	56952.27	65420.63	76546.39	92228.01000	104503.71	4.825130e+05
76	นีงกาฟ	th-bk	0.00	0.00	685.02	806.10	900.50	976.61000	1047.13	4.415360e+03

77 rows × 10 columns

สรุปข้อมูลเป็นรายคอลลัมมีว่ามี MISSING

```
In [62]: data_BB.isnull().any() #any สรุปข้อมูลเป็นรายคอลลัมน์ว่ามี missing ไหม
```

Out[62]:	ที่	False
	ภาค	False
	จังหวัด	False
	หน่วยงานที่รับผิดชอบ	False
	ที่ตั้ง	False
	ชื่อย่อของระบบ	False
	ประเภทของระบบ	False
	พ.ศ. ที่ได้รับงบประมาณ	False
	พ.ศ. ที่เปิดใช้งาน	True
	สถานภาพปัจจุบัน	False
	ปริมาณน้ำเสียที่รองรับได้ (ลบ.ม./วัน)	True
	แหล่งงบประมาณ	False
	dtype: bool	

จากการตรวจสอบค่า MISSING VALUE ของทั้ง 3 ชุดข้อมูลพบว่ามีค่าเป็น TRUE อยู่ 2 COLUMN ในข้อมูลชุดที่ 1 คือ พ.ศ. ที่เปิดใช้งาน และ ปริมาณน้ำเสียที่รองรับได้ (ลบ.ม./วัน)

DROP MISSING VALUE : เป็นการลบข้อมูลที่มีค่า MISSING ออกไปจากตาราง

```
data_BB = data_BB.dropna()  
data_BB
```

ตรวจสอบเมื่อลบออกไปแล้วตรวจสอบว่ายังไม่ข้อมูลที่หายไปหรือไม่

```
In [66]: data_BB.isnull().any()
```

```
Out[66]: ที่  
ภาค  
จังหวัด  
หน่วยงานที่รับผิดชอบ  
ที่ตั้ง  
ชื่อย่อของระบบ  
ประเภทของระบบ  
พ.ศ. ที่ได้รับงบประมาณ  
พ.ศ. ที่เปิดใช้งาน  
สถานภาพปัจจุบัน  
ปริมาณเน้าเสียที่รองรับได้ (ลบ.ม./วัน)  
แหล่งงบประมาณ  
dtype: bool
```

ที่	False
ภาค	False
จังหวัด	False
หน่วยงานที่รับผิดชอบ	False
ที่ตั้ง	False
ชื่อย่อของระบบ	False
ประเภทของระบบ	False
พ.ศ. ที่ได้รับงบประมาณ	False
พ.ศ. ที่เปิดใช้งาน	False
สถานภาพปัจจุบัน	False
ปริมาณเน้าเสียที่รองรับได้ (ลบ.ม./วัน)	False
แหล่งงบประมาณ	False
dtype: bool	

ตรวจสอบแล้วพบว่าไม่มีข้อมูลที่หายไป

PERCENT OF MISSING DATA FROM DROPNA

เป็นการตรวจสอบว่าเมื่อลบข้อมูลทิ้งไปแล้วเราสูญเสียข้อมูลไปกี่เปอร์เซ็นต์

```
In [67]: before = 94  
        after = 85  
        print(f'before = {before} \nafter = {after}\nwhich is {100*((before-after)/before)}')  
  
before = 94  
after = 85  
which is 9.574468085106384
```

จะเห็นได้ว่าค่าของข้อมูลที่สูญหายไปคิดเป็น 9.574468085106384 %



รวมตาราง

นำตารางข้อมูลชุดที่ 1 และ 2 รวมเข้าด้วยกัน โดยใช้ตารางใหม่ที่ได้มาซึ่ง TABLE

```
In [16]: table = pd.merge(data_BB,data_PP,on='จังหวัด') #รวมตารางข้อมูลชุดที่ 1 และ 2
```

```
In [17]: table
```

```
Out[17]:
```

	ที่	ภาค	จังหวัด	หน่วยงานที่รับผิดชอบ	ที่ตั้ง	ชื่อย่อของระบบ	ประเภทของระบบ	พ.ศ. ที่ได้รับงบประมาณ	พ.ศ. ที่เบิกใช้งาน	สถานภาพปัจจุบัน	ปริมาณน้ำเสียที่รองรับได้(ลบ.ม./วัน)	แหล่งงบประมาณ	15 - 39 ปี	40 - 59 ปี	15 - 59 ปี	15 ปีขึ้นไป	60 ปีขึ้นไป	ผลเฉลี่ย
0	1	เหนือ	เชียงราย	ทน.เชียงราย	ข้างสนามกีฬากลางจังหวัด ถ.สหเมือง ต.รอบเวียง อ....	AL	ระบบสารเติมอากาศ	2540	2544.0	เบิกเดินระบบ	22,700	กระทรวงมหาดไทย ทรัพยากรธรรมชาติและสิ่งแ...	10.43	6.39	8.40	7.24	3.90	36.36
1	2	เหนือ	เชียงราย	ทต.สันจว้า	เลขที่ 49 ม.2 ต.สันจว้า อ.แม่สัน จ.เชียงราย	WL	ระบบบึงประดิษฐ์	2548	2550.0	เบิกเดินระบบ	50	กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	10.43	6.39	8.40	7.24	3.90	36.36
2	4	เหนือ	เชียงใหม่	ทน.เชียงใหม่	หลังวัดท่าใหม่ อ.ต.สันศักดิ์ หวาน อ.ทางดง จ.เชียงใหม่	AL	ระบบสารเติมอากาศ	2545	2547.0	เบิกเดินระบบ	55,000	กระทรวงมหาดไทย กระทรวงทรัพยากรธรรมชาติและสิ่งแ...	10.28	7.34	8.99	7.96	4.49	39.06
3	5	เหนือ	น่าน	ทม.น่าน	ถ.พากอง ต.ในเวียง อ.เมือง น่าน จ.น่าน	SP	ระบบบ่อปรับเลดี้ร	2540	2544.0	เบิกเดินระบบ	8,259	กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	11.08	7.58	9.13	7.77	4.40	39.96
4	6	เหนือ	พะเยา	ทม.พะ夷า	ซอยข้างสินธนา ต.แม่ต่า อ.เมืองพะ夷า จ.พะ夷า	SP	ระบบบ่อปรับเลดี้ร	2538	2541.0	เบิกเดินระบบ	9,700	กระทรวงมหาดไทย กระทรวงทรัพยากรธรรมชาติและสิ่งแ...	11.26	7.56	9.17	7.84	4.45	40.28

รวมตาราง

และนำตารางซึ่ง TABLE ที่ได้มา
จากการรวมตารางข้อมูลที่ 1,2
มารวมเข้ากับตารางข้อมูลชุด
ที่ 3 โดยให้ชื่อว่า TABLE_3

	ที่	ภาค	จังหวัด	หน่วยงานที่รับผิดชอบ	ที่ตั้ง	ชื่อย่อของระบบ	ประเภทของระบบ	พ.ศ. ที่ได้รับงบประมาณ	พ.ศ. ที่เบิกใช้งาน	สถานภาพปัจจุบัน	ปริมาณน้ำเสียที่รองรับได้ (ลบ.ม./วัน)	แหล่งงบประมาณ	15 - 39 ปี	40 - 59 ปี	15 - 59 ปี	15 ปีขึ้นไป	60 ปีขึ้นไป	ผลเฉลี่ย	
0	1	เหนือ	เชียงราย	ทน.เชียงราย	ข้างสนามกีฬากลางจังหวัด ต.สหเมือง ต.รอบเวียง อ....	AL	ระบบสารเติมอากาศ	2540	2544.0	เบิกเดินระบบ	22,700	กระทรวงมหาดไทย ทรัพยากรธรรมชาติและสิ่งแวดล้อม	10.43	6.39	8.40	7.24	3.90	36.36	t
1	2	เหนือ	เชียงราย	ทค.จันจว�	เลขที่ 49 ม.2 ต.จันจว้า อ.แม่จัน จ.เชียงราย	WL	ระบบบึงประดิษฐ์	2548	2550.0	เบิกเดินระบบ	50	กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	10.43	6.39	8.40	7.24	3.90	36.36	t
2	4	เหนือ	เชียงใหม่	ทน.เชียงใหม่	หลังวัดท่าใหม่ อ.ต.สันถก หวาน อ.ทางดง จ.เชียงใหม่	AL	ระบบสารเติมอากาศ	2545	2547.0	เบิกเดินระบบ	55,000	กระทรวงมหาดไทย กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	10.28	7.34	8.99	7.96	4.49	39.06	t
3	5	เหนือ	น่าน	ทม.น่าน	ต.ปากกอง ต.ในเวียง อ.เมืองน่าน จ.น่าน	SP	ระบบบ่อปรับเสถียร	2540	2544.0	เบิกเดินระบบ	8,259	กระทรวงทรัพยากรธรรมชาติและสิ่งแวดล้อม	11.08	7.58	9.13	7.77	4.40	39.96	t

In [72]: `table_3.isnull().any()`

```
Out[72]: ที           False
ภาค          False
จังหวัด      False
หน่วยงานที่รับผิดชอบ  False
ทีตั้ง        False
ชื่อย่อของระบบ  False
ประเภทของระบบ  False
พ.ศ. ที่ได้รับงบประมาณ  False
พ.ศ. ที่เบิกใช้งาน  False
สถานภาพปัจจุบัน  False
  ปริมาณน้ำเสียที่รองรับได้ (ลบ.ม/วัน)  False
แหล่งงบประมาณ  False
15 - 39 ปี  False
40 - 59 ปี  False
15 - 59 ปี  False
15 ปีขึ้นไป  False
60 ปีขึ้นไป  False
ผลเฉลี่ย  False
รหัสจังหวัด  False
2555  False
2556  False
2557  False
2558  False
2559  False
2560  False
2561  False
  รวมเฉลี่ย  False
dtype: bool
```

ตรวจสอบ TABLE_3 แล้วพบว่าไม่มีข้อมูลที่หายไป
ด้วยคำสั่ง `TABLE_3.ISNULL().ANY()`

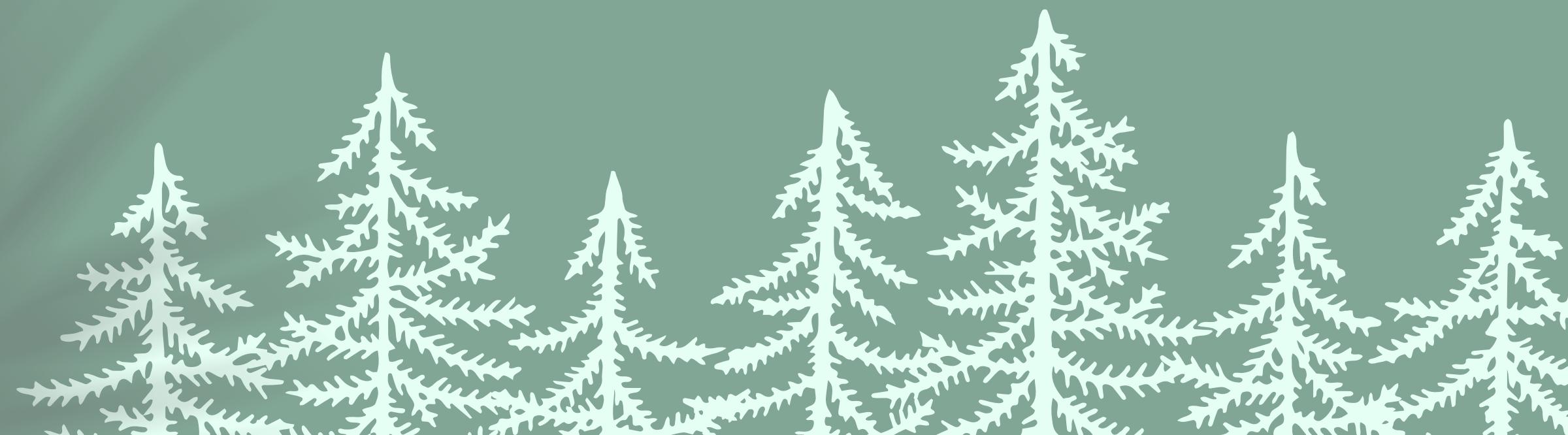
In [21]: `table_3.shape #ตาราง table_3 มี 74 และ 27 column`

Out[21]: (74, 27)

ตาราง TABLE_3 มี 74 และ 27 คอลัมน์

ปัญหา

ต้องการท่านนายว่าระบบกำจัดของเสียประเภทใดมีการใช้งานมากสุดของแต่ละภาค



CLASSIFICATION

ทำการดูว่าประเภทของระบบใดในแต่ละภาคมีการใช้งานมากที่สุด

```
South = Southdata.groupby('ประเภทของระบบ').count().sort_values(by=['ภาค'], ascending=False) ##นำข้อมูล จังหวัด_y จากตาราง Data โดยพิจารณาจาก StFinal
South ##เรียกคุณภาพ
```

	โซน	ภาค	จังหวัด	ผลเฉลี่ย	รวมเฉลี่ย
ประเภทของระบบ					
ระบบคลองน้ำเสีย	3	3	3	3	3
ระบบสระเติมอากาศ	3	3	3	3	3
ระบบบ่อปรับเสถียร	2	2	2	2	2
ระบบกลุ่มอาคาร	1	1	1	1	1
ระบบงานหมุนชี้ภาพ	1	1	1	1	1
ระบบตะกอนเร่ง+ปั๊งประดิษฐ์	1	1	1	1	1
ระบบบ่อปรับเสถียร+ปั๊งประดิษฐ์	1	1	1	1	1

จะเห็นได้ว่าภาคใต้มีการใช้ประเภทของระบบกำจัดของเสียที่มากที่สุด คือ ระบบคลองน้ำเสีย

```
Nort = Nortdata.groupby('ประเภทของระบบ').count().sort_values(by=['ภาค'], ascending=False) ##นำข้อมูล จังหวัด_y จากตาราง Data โดยพิจารณาจาก StFinal
Nort ##เรียกคุณภาพ
```

	โซน	ภาค	จังหวัด	ผลเฉลี่ย	รวมเฉลี่ย
ประเภทของระบบ					
ระบบบ่อปรับเสถียร	3	3	3	3	3
ระบบสระเติมอากาศ	2	2	2	2	2
ระบบกลุ่มอาคาร	1	1	1	1	1
ระบบปั๊งประดิษฐ์	1	1	1	1	1
ระบบแอลฟ์เนอร์	1	1	1	1	1

จะเห็นได้ว่าภาคเหนือมีการใช้ประเภทของระบบกำจัดของเสียที่มากที่สุด คือ ระบบบ่อปรับเสถียร

เป็นการคำนวณภาคใต้

เป็นการคำนวณภาคเหนือ

```
central = centraldata.groupby('ประเภทของระบบ').count().sort_values(by=['ภาค'], ascending=False) ##นำข้อมูล จังหวัด_y จากตาราง Data โดยพิจารณาจาก StFinal
central ##เรียกดูตาราง
```

	โซน	ภาค	จังหวัด	ผลเฉลี่ย	รวมเฉลี่ย
ประเภทของระบบ					
ระบบบ่อปรับเสถียร	9	9	9	9	9
ระบบกลุ่มอาคาร	5	5	5	5	5
ระบบสระเต็มอากาศ	4	4	4	4	4
ระบบตอกอนเร่ง	2	2	2	2	2
ระบบบึงประดิษฐ์	2	2	2	2	2
ระบบตอกอนเร่งแบบชีวนوال	1	1	1	1	1

จะเห็นได้ว่าภาคกลางมีการใช้ประเภทของระบบกำจัดของเสียงมากที่สุด คือ ระบบบ่อปรับเสถียร

เป็นการท่านายของภาคกลาง

```
East = Eastdata.groupby('ประเภทของระบบ').count().sort_values(by=['ภาค'], ascending=False) ##นำข้อมูล จังหวัด_y จากตาราง Data โดยพิจารณาจาก StFinal
East ##เรียกดูตาราง
```

	โซน	ภาค	จังหวัด	ผลเฉลี่ย	รวมเฉลี่ย
ประเภทของระบบ					
ระบบบ่อปรับเสถียร	3	3	3	3	3
ระบบคลองวันเวียน	2	2	2	2	2
ระบบสระเต็มอากาศ	2	2	2	2	2
ระบบบึงประดิษฐ์	1	1	1	1	1

จะเห็นได้ว่าภาคตะวันออกมีการใช้ประเภทของระบบกำจัดของเสียงมากที่สุด คือ ระบบบ่อปรับเสถียร

เป็นการท่านายของภาคตะวันออก

```
West = Westdata.groupby('ประเภทของระบบ').count().sort_values(by=[ 'ภาค'],ascending=False) ##นำข้อมูล จังหวัด_y จากตาราง Data โดยพิจารณาจาก StFinal
West ##เรียกดูตาราง
```

	โซน	ภาค	จังหวัด	ผลเฉลี่ย	รวมเฉลี่ย
ประเภทของระบบ					
ระบบบ่อปรับเสถียร	5	5	5	5	5
ระบบคลองวันเวียน	1	1	1	1	1
ระบบรวมรวมน้ำเสีย	1	1	1	1	1
ระบบสระเติมอากาศ	1	1	1	1	1

จะเห็นได้ว่าภาคตะวันตกมีการใช้ประเภทของระบบกำจัดของเสียที่มากที่สุด คือ ระบบบ่อปรับเสถียร

เป็นการคำนวณภาคตะวันตก

```
Northeast = Northeastdata.groupby('ประเภทของระบบ').count().sort_values(by=[ 'ภาค'],ascending=False) ##นำข้อมูล จังหวัด_y จากตาราง Data โดยพิจารณาจาก StFinal
Northeast ##เรียกดูตาราง
```

	โซน	ภาค	จังหวัด	ผลเฉลี่ย	รวมเฉลี่ย
ประเภทของระบบ					
ระบบบ่อปรับเสถียร	7	7	7	7	7
ระบบสระเติมอากาศ	3	3	3	3	3
ระบบบึงประดิษฐ์	2	2	2	2	2
ระบบคลองวันเวียน	1	1	1	1	1
ระบบบ่อปรับเสถียร+บึงประดิษฐ์	1	1	1	1	1
ระบบรวมรวมน้ำเสีย	1	1	1	1	1

จะเห็นได้ว่าภาคตะวันออกเฉียงเหนือมีการใช้ประเภทของระบบกำจัดของเสียที่มากที่สุด คือ ระบบบ่อปรับเสถียร

เป็นการคำนวณภาค ตะวันออกเฉียงเหนือ

SPLIE DATE

```
: X = data_1[['โซน', 'ผลเฉลี่ย', 'รวมเฉลี่ย']] #กำหนดค่าของ X และ y เพื่อทำ decision tree ต่อความลึกทั้งหมด
```

```
: y = data_1['ประเภทของระบบ']
```

```
: X
```

	โซน	ผลเฉลี่ย	รวมเฉลี่ย
0	1	36.36	160035.50
1	1	36.36	160035.50
2	1	39.06	557866.57
3	1	39.96	13634.01
4	1	40.28	7862.51
...
69	3	39.80	58165.80
70	3	39.80	58165.80
71	3	36.25	42377.14
72	3	36.25	42377.14
73	3	38.66	2585.07

74 rows × 3 columns

กำหนดค่า X และ Y

โซน คือ เลขกำหนดภาค

ผลรวมเฉลี่ย คือ ค่าผลรวมเฉลี่ยของประชากร
จำแนกตามอายุ

รวมเฉลี่ย คือ ค่าผลรวมเฉลี่ยจากการแยกด้วย
การท่องเที่ยว

SPLIE DATE

```
In [219]: from sklearn.model_selection import train_test_split  
  
In [220]: # train-test  
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.5, random_state=8)  
  
In [221]: # train-validation  
         X_train2, X_val, y_train2, y_val = train_test_split(X_train, y_train, test_size=0.5, random_state=8)  
  
In [222]: X.shape #ເນື້ອຂອນກາຕີ column ສອງ X.shape  
Out[222]: (74, 3)  
  
In [223]: X_train.shape #ເນື້ອຂອນກາຕີ column ສອງ X_train.shape  
Out[223]: (37, 3)  
  
In [224]: X_test.shape #ເນື້ອຂອນກາຕີ column ສອງ X_test.shape  
Out[224]: (37, 3)
```

DECISION TREE

โดยค่าเฉลี่ยของมาที่มีค่าร้อยละมากที่สุดเท่ากับ 42%

```
In [225]: from sklearn.tree import DecisionTreeClassifier #Import  
from sklearn.metrics import accuracy_score
```

```
ให้ก่อนที่ splitter='random',random_state=13
```

```
In [226]: #define  
mytree = DecisionTreeClassifier(splitter="random",random_state=13)  
#train  
mytree.fit(X_train2,y_train2)  
#test  
mytree_result2 = mytree.predict(X_val)  
#accuracy_score  
accuracy_score(y_val, mytree_result2)
```

```
Out[226]: 0.2631578947368421
```

```
ให้ก่อนที่ min_samples_leaf=3,splitter='random',random_state=13
```

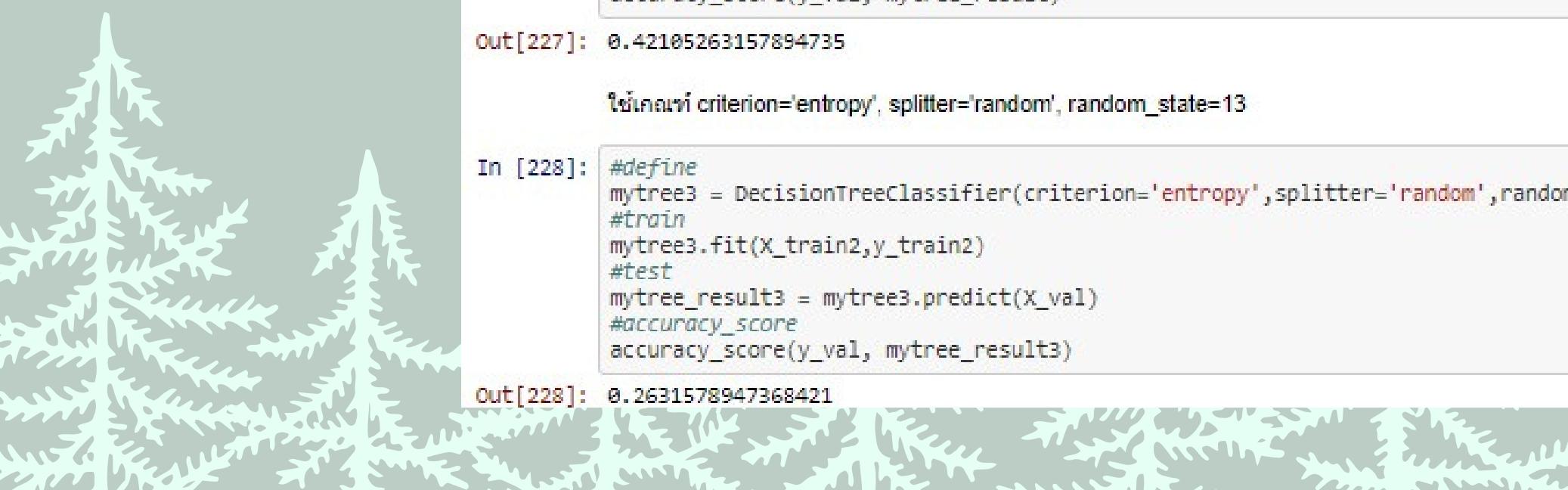
```
In [227]: # Define  
mytree2 = DecisionTreeClassifier(min_samples_leaf=3,splitter='random',random_state=13)  
#train  
mytree2.fit(X_train2,y_train2)  
#test  
mytree_result = mytree2.predict(X_val)  
#accuracy_score  
accuracy_score(y_val, mytree_result)
```

```
Out[227]: 0.42105263157894735
```

```
ให้ก่อนที่ criterion='entropy', splitter='random', random_state=13
```

```
In [228]: #define  
mytree3 = DecisionTreeClassifier(criterion='entropy',splitter='random',random_state=13)  
#train  
mytree3.fit(X_train2,y_train2)  
#test  
mytree_result3 = mytree3.predict(X_val)  
#accuracy_score  
accuracy_score(y_val, mytree_result3)
```

```
Out[228]: 0.2631578947368421
```



KNN

มีค่าความแม่นยำอยู่ที่ร้อยละเท่ากับ 36% ของทุกตาราง

Train Model

```
In [235]: from sklearn.neighbors import KNeighborsClassifier  
from sklearn.metrics import accuracy_score
```

Knn1

```
In [236]: # Define  
neigh1 = KNeighborsClassifier(n_neighbors=3, weights='uniform') # กำหนดจำนวนที่ใกล้สุด 3 คน และเข้าทุกคนเท่ากัน  
# Train  
neigh1.fit(X_train2,y_train2)  
# Test  
knn1_result = neigh1.predict(X_val)  
accuracy_score(y_val, knn1_result)
```

```
Out[236]: 0.3684210526315789
```

Knn2

```
In [237]: # Define  
neigh2 = KNeighborsClassifier(n_neighbors=10, weights='distance') # กำหนดจำนวนที่ใกล้สุด 10 คน และเข้าคนที่อยู่ใกล้มากกว่าใกล้  
# Train  
neigh2.fit(X_train2,y_train2)  
# Test  
knn2_result = neigh2.predict(X_val)  
accuracy_score(y_val, knn2_result)
```

```
Out[237]: 0.3684210526315789
```

Knn3

```
In [238]: # Define  
neigh3 = KNeighborsClassifier(n_neighbors=1) # เข้าเพื่อนบ้านที่ใกล้ที่สุดคนเดียว  
# Train  
neigh3.fit(X_train2,y_train2)  
# Test  
knn3_result = neigh3.predict(X_val)  
accuracy_score(y_val, knn3_result)
```

```
Out[238]: 0.3684210526315789
```

NEURAL NETWORK

จากการทำ NEURAL NETWORK พบร่วมค่า KNN อยู่ที่ 42 %

```
In [242]: from sklearn.neural_network import MLPClassifier
```

```
In [243]: perceptron1 = MLPClassifier(random_state=13,max_iter=2000,hidden_layer_sizes=1,learning_rate_init=0.05)
```

Train - Test

```
In [244]: # Train  
perceptron1.fit(X_train2,y_train2)  
# Test  
ann1_result = perceptron1.predict(X_val)  
accuracy_score(y_val, ann1_result)
```

```
Out[244]: 0.42105263157894735
```

Ann2

```
In [245]: perceptron2 = MLPClassifier(random_state=13,max_iter=2000,hidden_layer_sizes=10,learning_rate_init=0.05)  
# Train  
perceptron2.fit(X_train2,y_train2)  
# Test  
ann2_result = perceptron2.predict(X_val)  
accuracy_score(y_val, ann2_result)
```

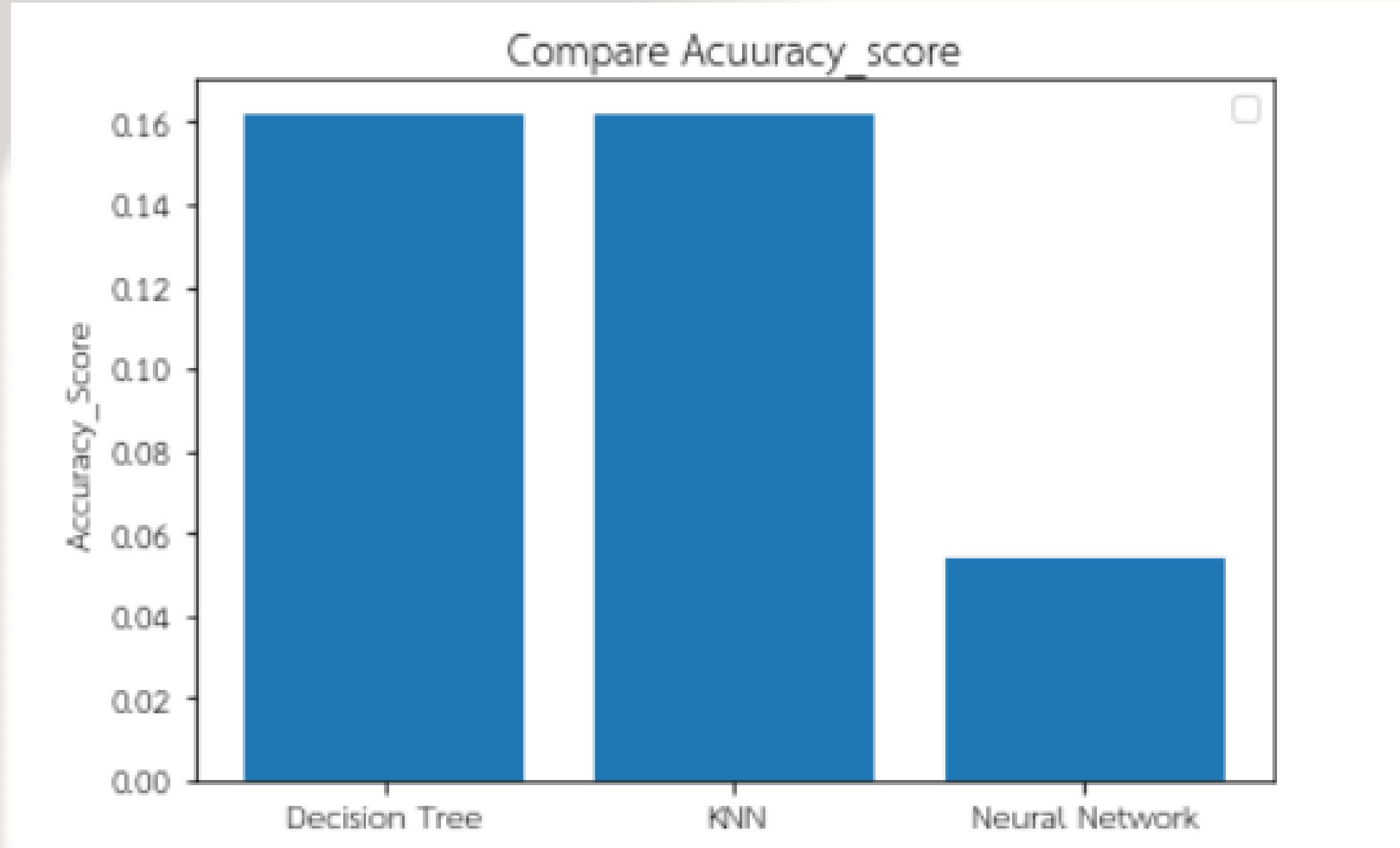
```
Out[245]: 0.42105263157894735
```

ANN 3

```
In [246]: perceptron3 = MLPClassifier(random_state=13,max_iter=20000,hidden_layer_sizes=10,learning_rate_init=0.01)  
# Train  
perceptron3.fit(X_train2,y_train2)  
# Test  
ann3_result = perceptron3.predict(X_val)  
accuracy_score(y_val, ann3_result)
```

```
Out[246]: 0.3157894736842105
```

การวัดผล

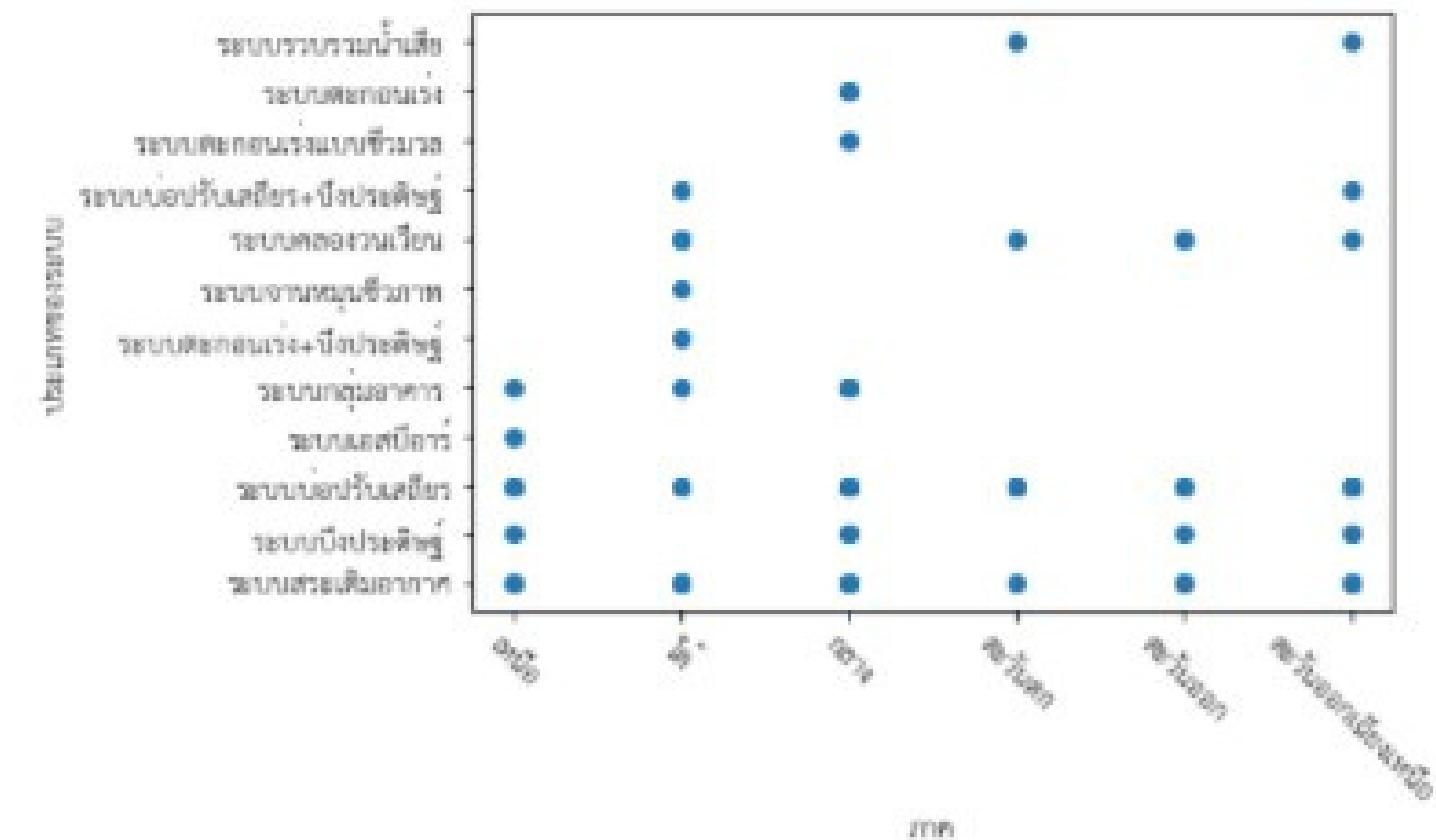


จะเห็นได้ว่าจากการเปรียบเทียบ
และวัดผลของทั้ง 3 MODEL จะเห็นได้ว่า
ค่า DECISION TREE และ KNN นั้นมีค่า
แม่นยำสุด

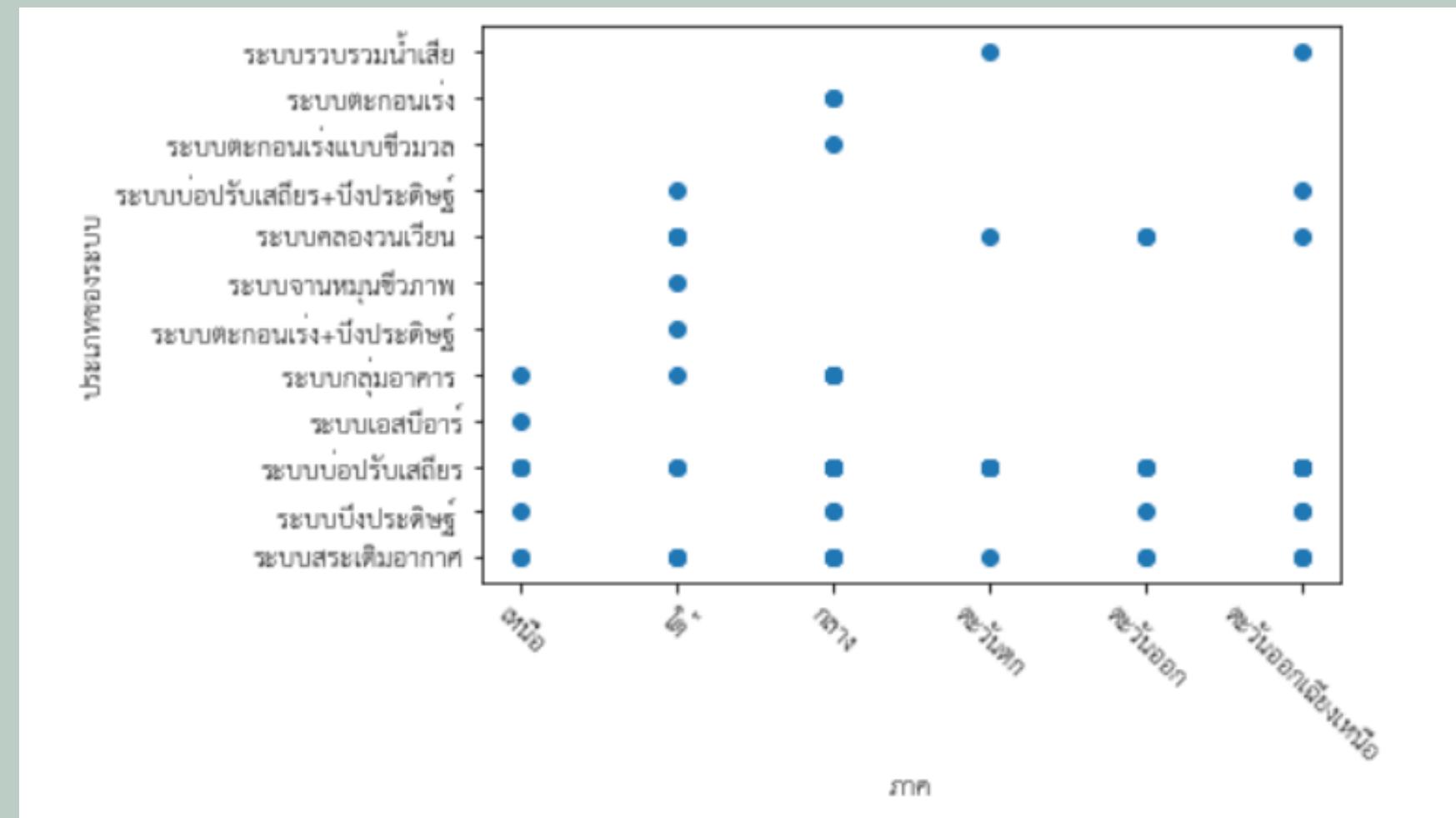
VISUALIZATION

สำหรับการทำ VISUALIZATION กลุ่มวิชาการเมืองได้ทำเป็น 4 ตารางได้แก่

```
In [270]: plt.scatter(data=data_1,y='ประเภทของระบบ',x='ภาค') ##plot กราฟการกระจายของช่องทาง km และ จำนวนผู้โดยสารเดินทาง (คน)_x  
plt.ylabel('ประเภทของระบบ')  
plt.xlabel('ภาค')  
plt.xticks(rotation=-45)  
plt.show()
```



1. ตารางแสดงผลว่าแต่ละภาคมีประเภทระบบกำจัดน้ำเสียอะไรบ้าง โดยเราใช้ x เป็นภาคซึ่งประกอบไปด้วย ๒ ภาค และ Y คือประเภทของระบบทั้งหมด จะเห็นได้ว่าแต่ละภาคมีประเภทระบบที่แตกต่างกัน



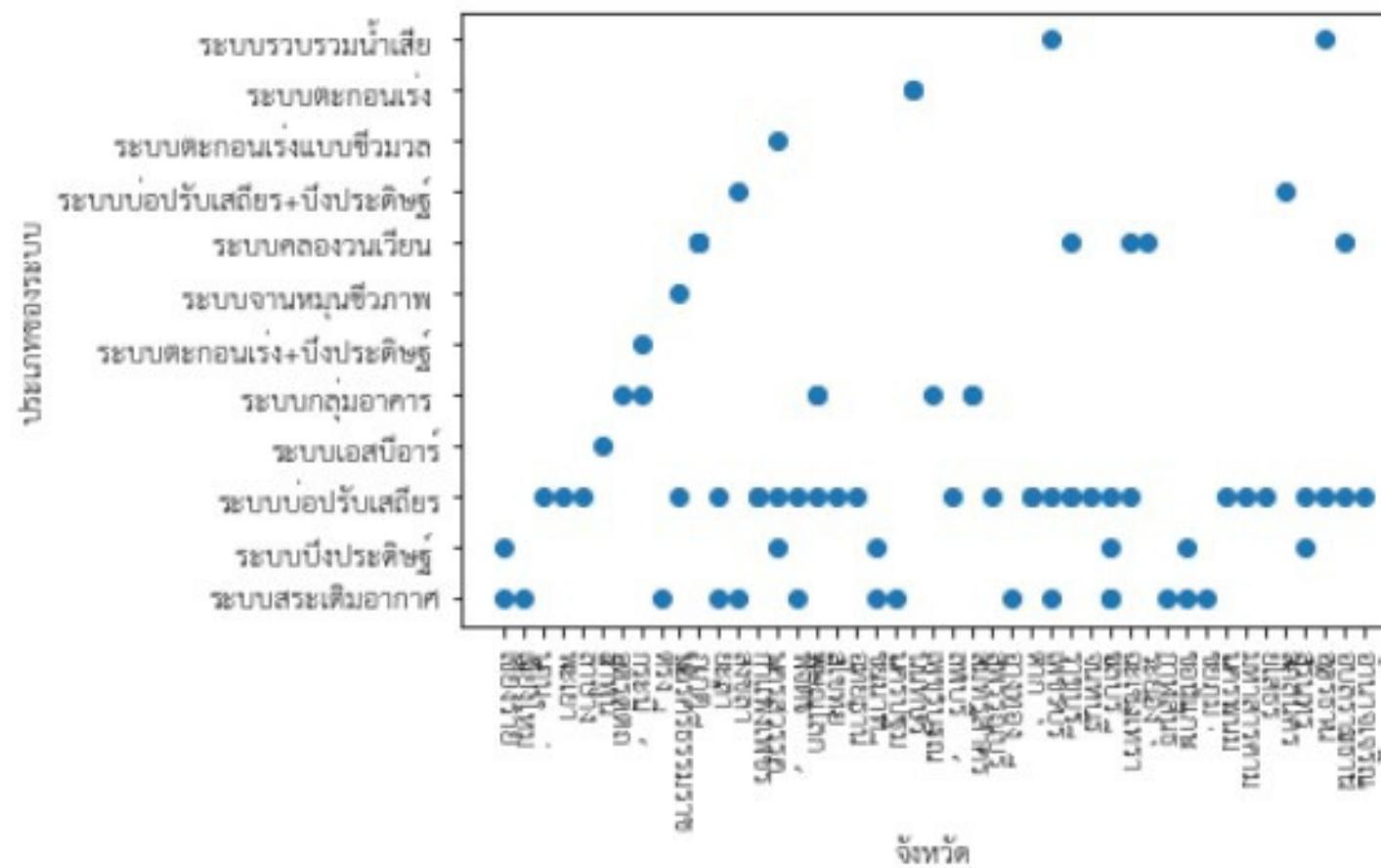
ชั่งความสามารถสรุปผลได้ 2 แบบคือ

1. ภาคที่มีประเภทของระบบกำจัดของเสียที่มากที่สุดคือภาคใต้ และภาคที่มีประเภทของระบบกำจัดของเสียที่น้อยที่สุดคือภาคตะวันตก
2. ประเภทของระบบกำจัดของเสียที่ถูกใช้งานที่สุดคือระบบบ่อปรับเสถียรและระบบสารเติมอากาศ ประเภทของระบบกำจัดของเสียที่ถูกใช้งานน้อยที่สุดคือระบบตักกอนเรงและระบบตักกอนเรงแบบชั่วโมง

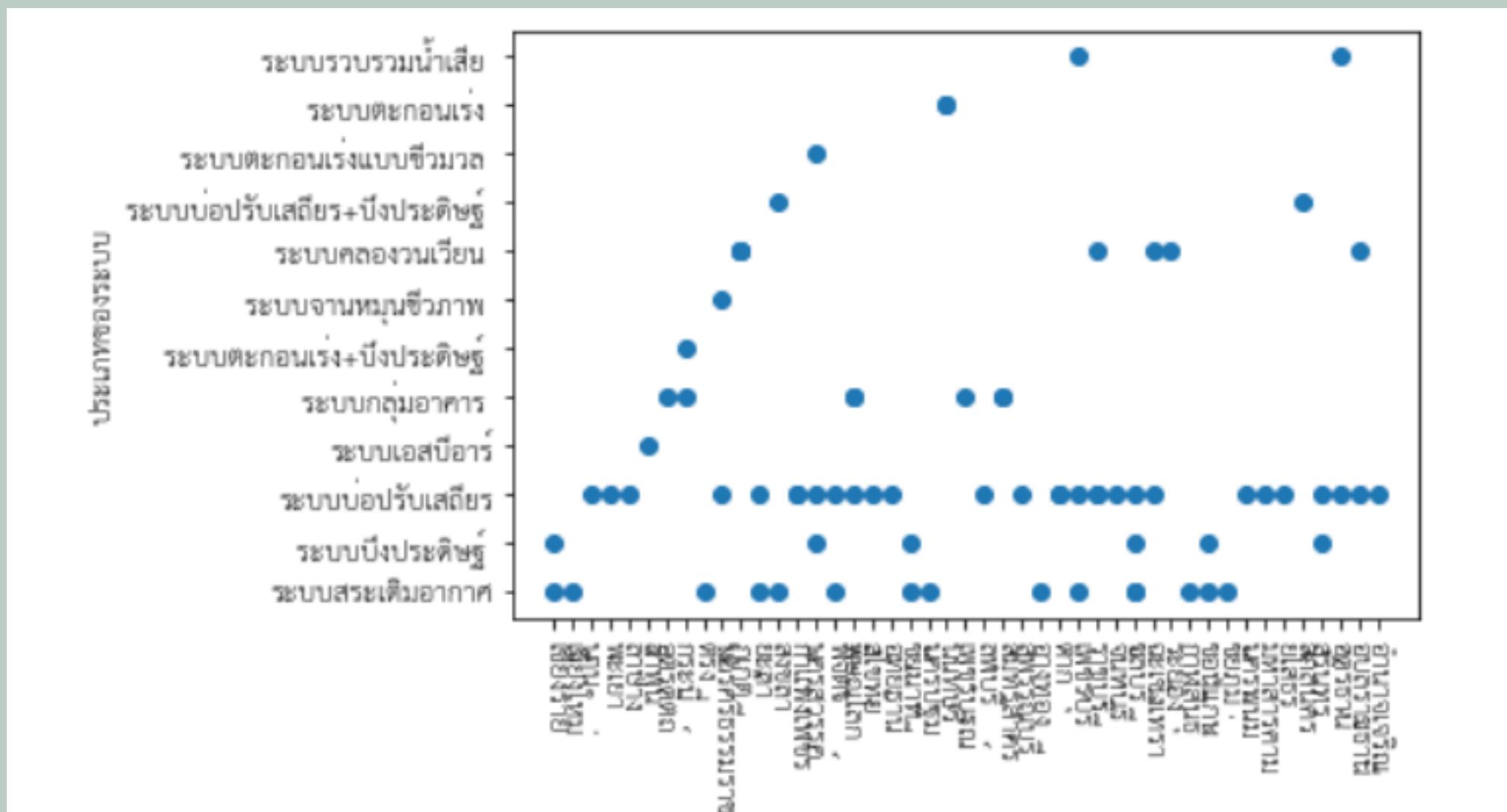
VISUALIZATION

In [121...]

```
plt.scatter(data=data_1,y='ประเภทของระบบ',x='จังหวัด') ##plot กราฟดูการกระจายของประเภทของระบบ และ จังหวัด  
plt.ylabel('ประเภทของระบบ')  
plt.xlabel('จังหวัด')  
plt.xticks(rotation=-90)  
plt.show()
```



2. ตารางแสดงผลว่าแต่ละจังหวัดมีประเภทระบบกำจัดน้ำเสียอะไรบ้าง โดยใช้ x เป็นจังหวัดซึ่งประกอบไปด้วย 77 จังหวัด และ y คือประเภทของระบบทั้งหมด จะเห็นได้ว่าแต่ละจังหวัดมีประเภทระบบที่แตกต่างกัน

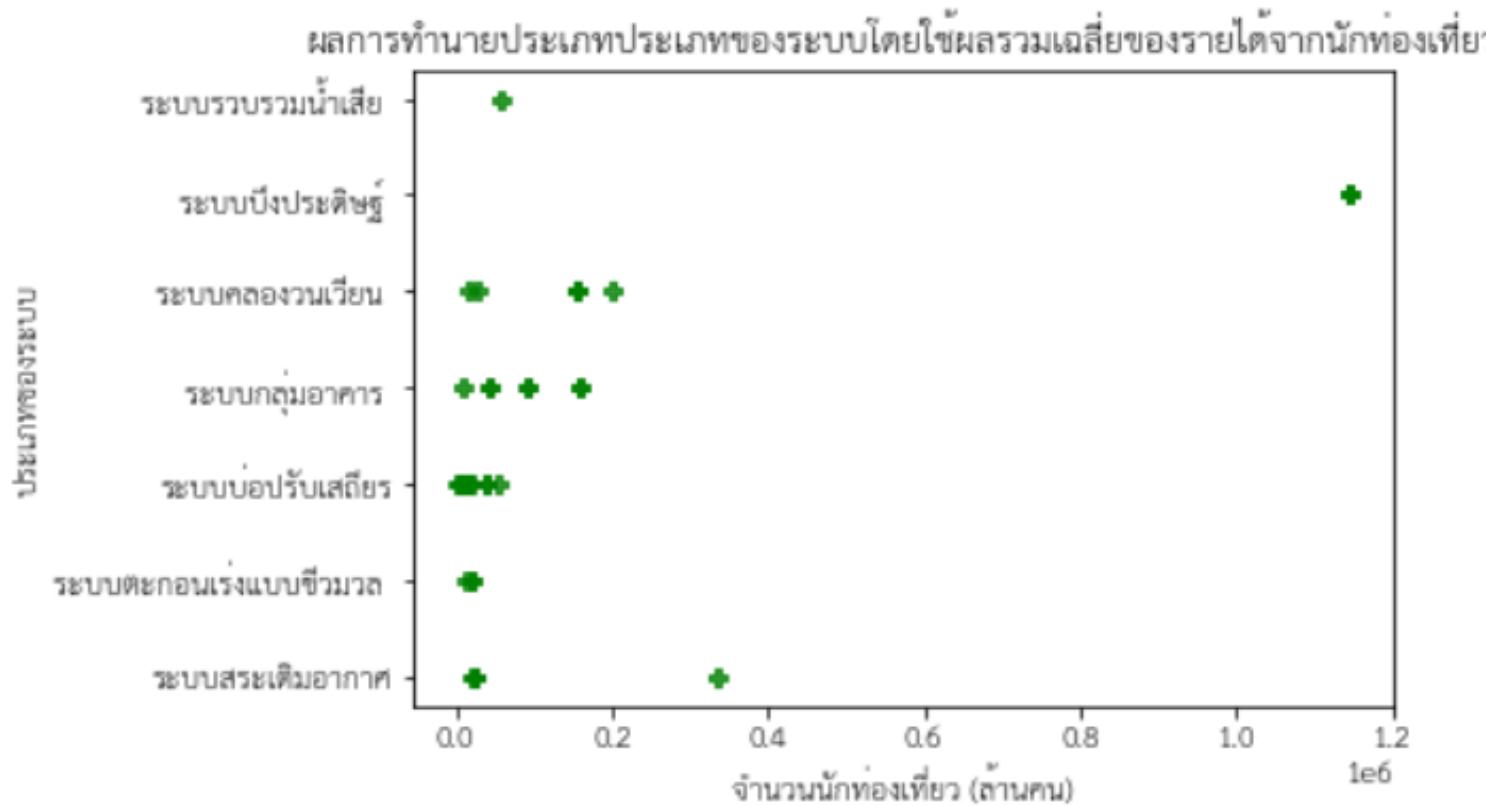


ซึ่งเราสามารถสรุปผลได้ดังนี้

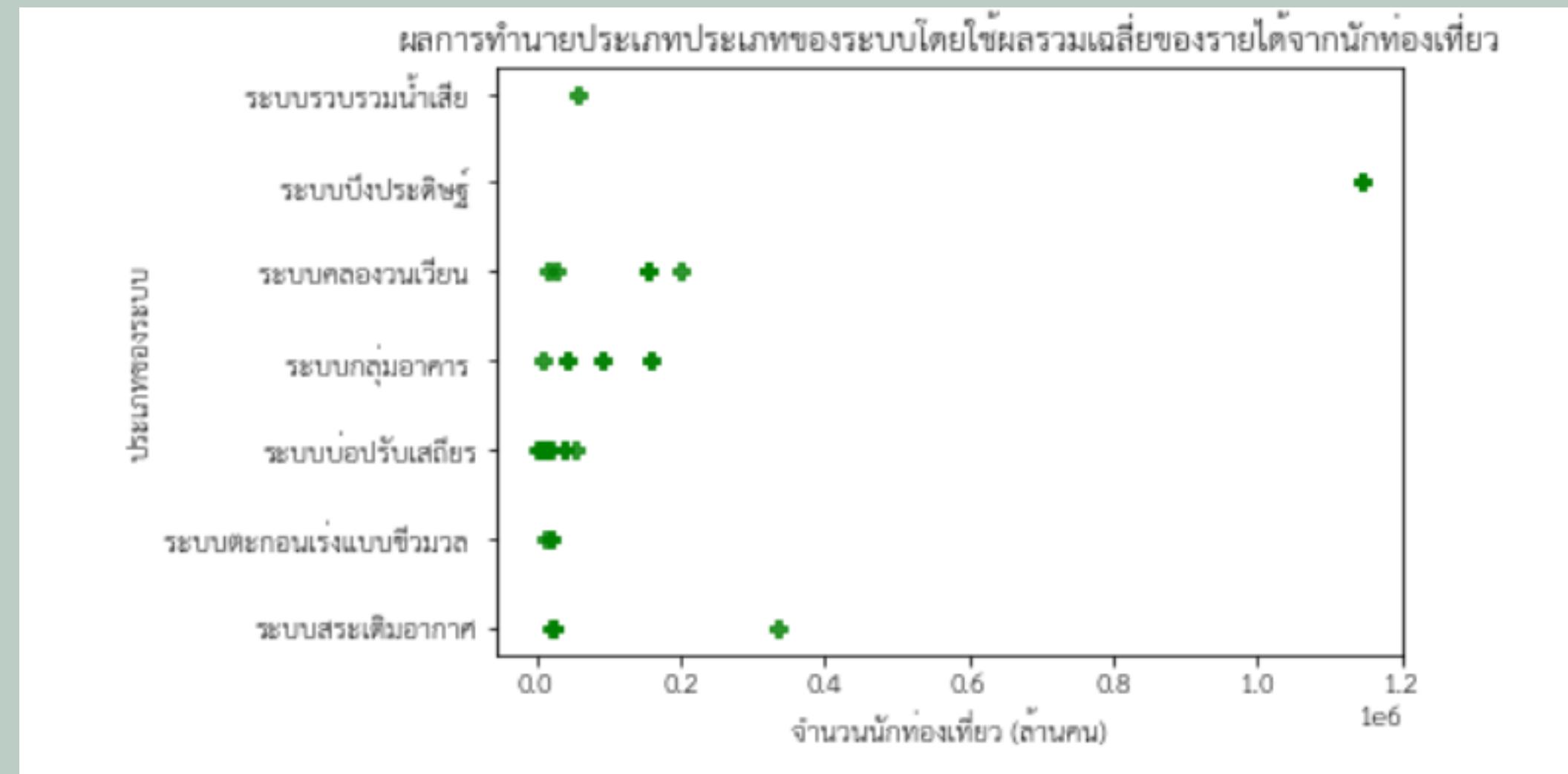
ประเภทของระบบกำจัดของเสียที่ถูกใช้งานที่สุดคือระบบบ่อปรับเสถียร และประเภทของระบบกำจัดของเสียที่ถูกใช้น้อยที่สุดคือระบบตะกอนเร่งและระบบตะกอนเร่งแบบชีวมวล

VISUALIZATION

```
In [272]: plt.plot(X_test['รวมเฉลี่ย'], mytree3_final_result, 'Pg', alpha =0.8)
plt.xlabel('จำนวนนักท่องเที่ยว (ล้านคน)')
plt.ylabel('ประมาณการ')
plt.title('ผลการทำนายประมาณการของระบบโดยใช้ผลรวมเฉลี่ยของรายได้จากนักท่องเที่ยว')
plt.show()
```



3. ตารางแสดงผลการทำนายว่าหากรายได้ของนักท่องเที่ยวอยู่ประมาณเท่าใด จะใช้ระบบประมาณการจำจัดน้ำเสียแบบบีด

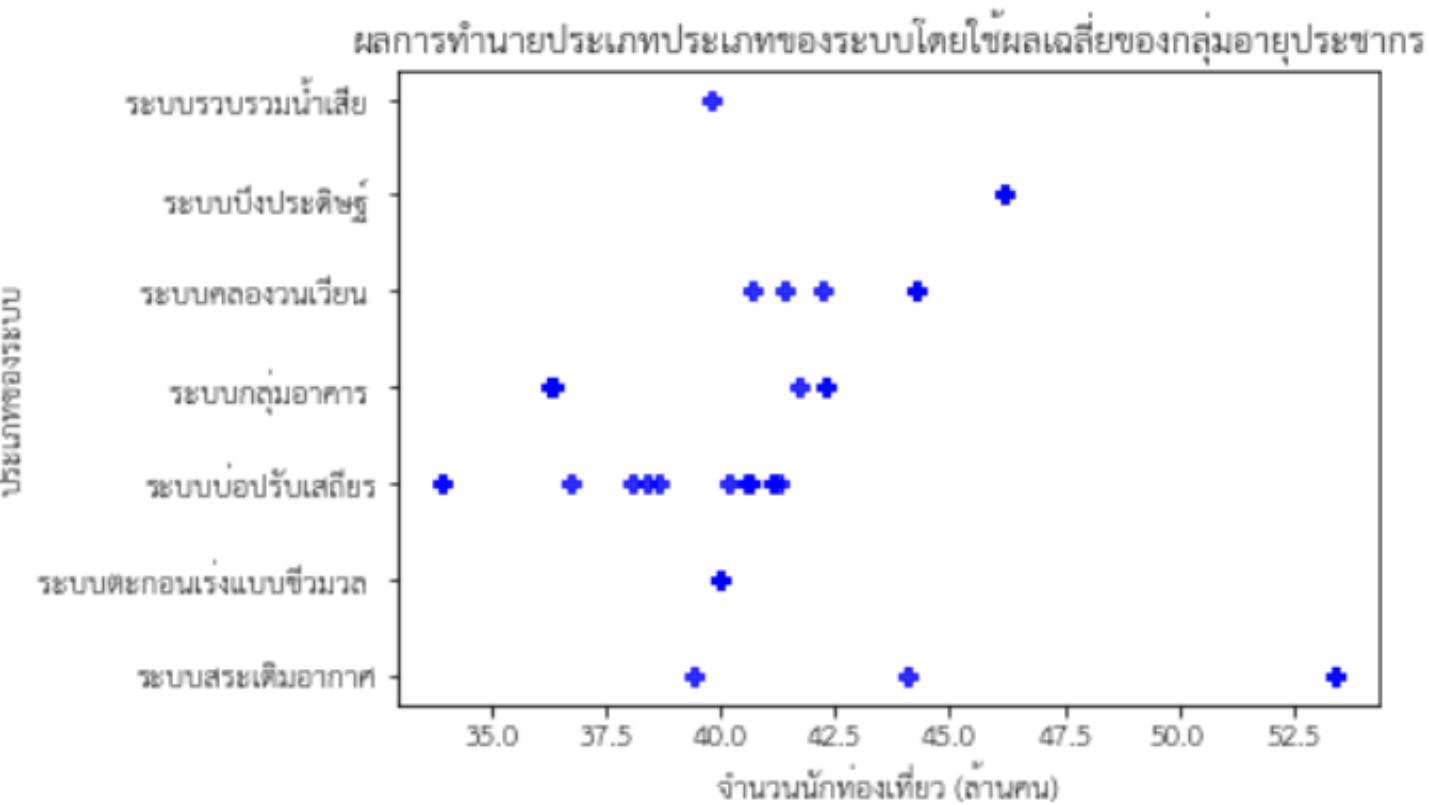


ซึ่งเราสามารถสรุปผลได้ดังนี้

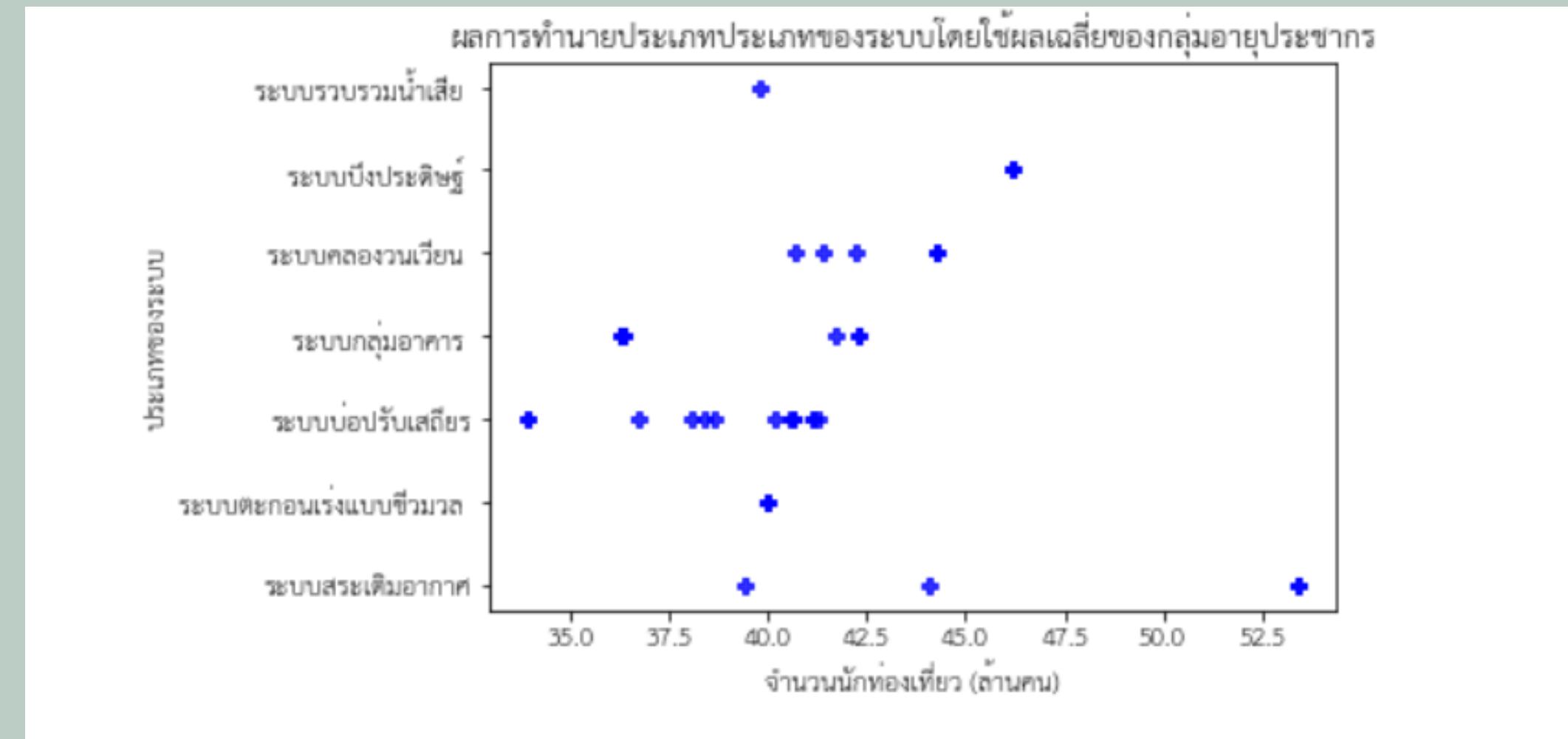
หากรายได้ของนักท่องเที่ยวอยู่ที่ 0 – 200000 บาทนักจะใช้ระบบรวมน้ำเสีย, คลองวันเวียน, กลุ่มอาคาร บ่อปรับเสถียร, ตากอนเร่งแบบชีวนิวาลหรือสระเติมอากาศ แต่หากรายได้นักท่องเที่ยวอยู่ที่ 1000000 – 1200000 มักใช้ระบบบีงประดิษฐ์แทน

VISUALIZATION

```
In [273]: plt.plot(X_test['ผลเฉลี่ย'], mytree3_final_result,'Pb',alpha = 0.8)
plt.xlabel('จำนวนนักท่องเที่ยว (ล้านคน)')
plt.ylabel('ประเภทของระบบ')
plt.title('ผลการทำนายประเภทของระบบโดยใช้ผลเฉลี่ยของกลุ่มอายุประชากร')
plt.show()
```



4. ตารางแสดงผลการทำนายว่าหากอายุของประชากรอยู่ประมาณเท่าใด จะใช้ระบบประเภทกำจัดน้ำเสียแบบใด



ซึ่งเราสามารถสรุปผลได้ดังนี้

หากอายุของประชากรอยู่ช่วงระหว่าง 35 – 37.5 ปี มักใช้ระบบบกกลุ่มอาคาร หากอายุประชากรอยู่ช่วง 37.5 – 40 ปี มักใช้ระบบรวมน้ำเสีย หากอายุประชากรอยู่ช่วง 40 – 42.5 ปี มักใช้ระบบคลองน้ำเวียน, บ่อปรับเสถียรหรือตะกอนเร่งแบบชีวนวลด หากอายุประชากรอยู่ช่วง 45 – 47.5 ปี มักใช้ระบบบึงประดิษฐ์ และหากอายุประชากรอยู่ที่ 52.5 ขึ้นไปมักใช้ระบบสร้างเติมอาคาร



ສນະບກໃນກລຸ່ມ

"ວາກາເມນ"

- นายปริชญา วงศ์ทองคำ 623020528-4
- นายมันนี่ พิทักษ์ 623020532-3
- นายสิงห์ตักษะ จรัสแสง 623020541-2
- นายชนะชัย อิสรະกุล 623021045-9
- นางสาวกอฟัน พงษ์พิเดช 623021048-3

THANK
YOU