COS3302 LAB#4

การวิเคราะห์ทางสถิติและการนำเสนอภาพข้อมูล

จากชุดข้อมูล

Adult Dataset

จัดทำโดย

นายศุภณัฐ แซ่เตีย

ID: 6505000270

OVERVIEW

- ลักษณะของชุดข้อมูล
- คุณสมบัติของแต่ละ Feature
  - Numerical
  - Categorical
- การทำ Data Preprocess
  - Missing Data Handle
  - Feature Engineering
  - Handling Outliers

ลักษณะของชุดข้อมูล

ชุดข้อมูลนี้เป็นชุดข้อมูลที่มาจากการสำรวจประชากรสหรัฐอเมริกาในปี ค.ศ. 1994 โดยชุดข้อมูลนี้ประกอบด้วยข้อมูลจำนวน 32,561 ตัวอย่าง และประกอบด้วย 15 Feature ได้แก่ อายุ, ประเภทงาน, Final Weight, ระดับการศึกษา, สถานภาพสมรส, สาขาอาชีพ, สถานะในครอบครัว, เชื้อชาติ, เพศ, ผลกำไรจากการลงทุน, ผลขาดทุนจากการลงทุน, ชั่วโมงทำงาน, ถิ่นกำเนิด และรายได้

	age	work-class	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

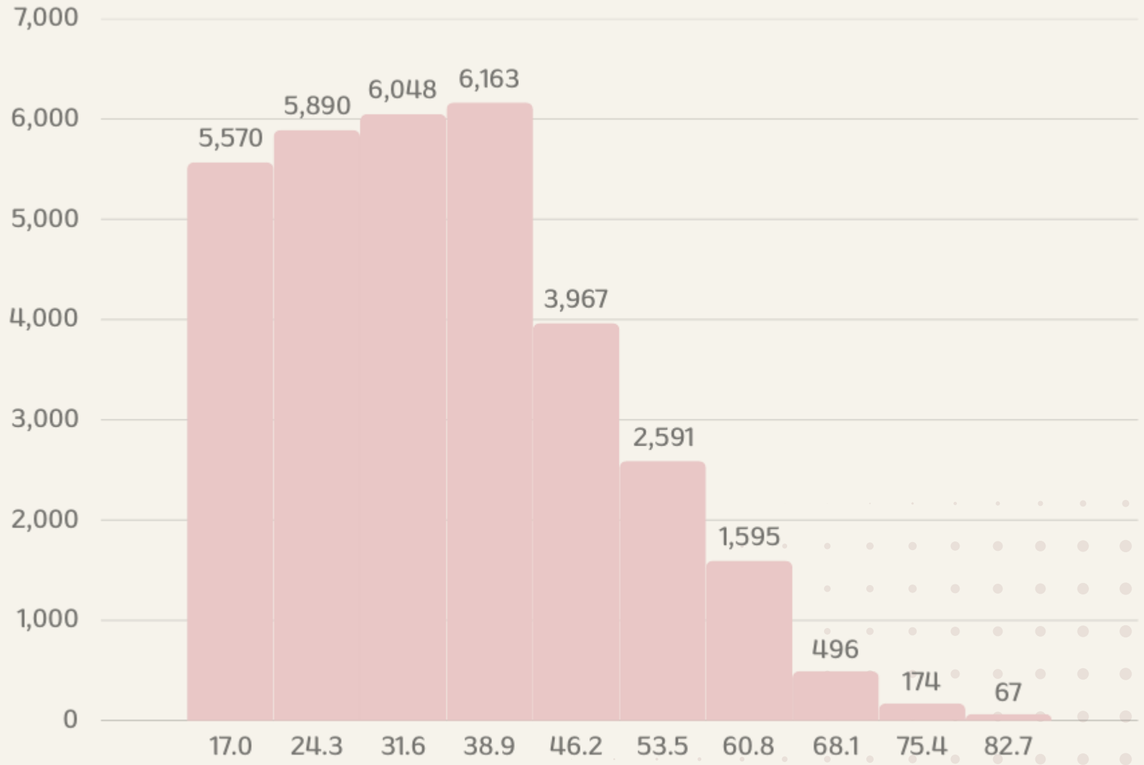
คุณสมบัติของแต่ละ FEATURE

ในประเภทเชิงปริมาณ (NUMERICAL)

อายุ (AGE)

เป็น Feature ที่แสดงเกี่ยวกับอายุของแต่ละตัวอย่าง โดย Feature นี้เป็นประเภทเชิงปริมาณ (Numerical) แบบอัตราส่วน (Ratio) โดยข้อมูลมีลักษณะเบ้ไปทางขวา (Right-skewed) ในระดับปานกลาง

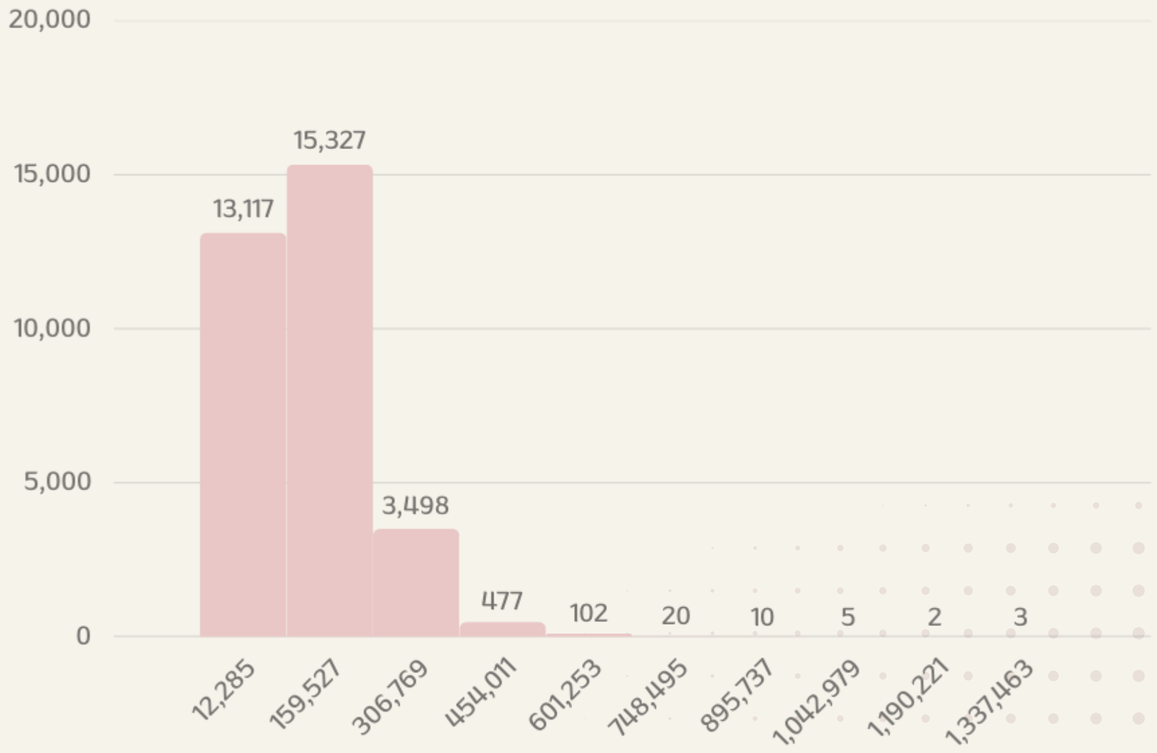
Mean	38.5816	Min	17
Median	37	Q_1	28
Mode	36	Q_2	37
SD.	13.6404	Q_3	48
Var.	186.061	Max	90



FINAL WEIGHT

เป็น Feature ที่แสดงว่าตัวอย่างนั้น ๆ เป็นตัวแทนของประชากรจริงจำนวนเท่าใด โดย Feature นี้เป็นประเภทเชิงปริมาณ (Numerical) แบบอัตราส่วน (Ratio) โดยข้อมูลมีลักษณะเบ้ไปทางขวา (Right-skewed) ในระดับสูง

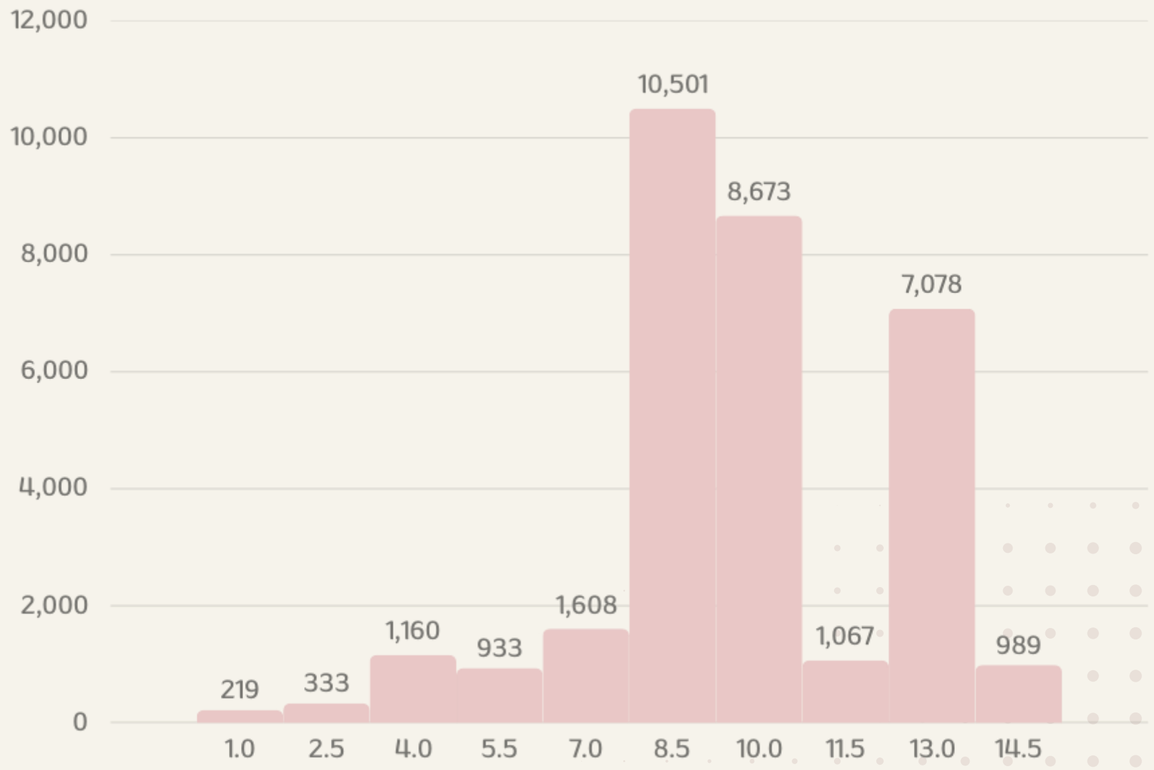
Mean	189,778.3665	Min	12,285
Median	178,356	Q_1	117,827
Mode	123,011	Q_2	178,356
SD.	105,549.9777	Q_3	237,051
Var.	11,140,797,791.8419	Max	1,484,705



ระดับการศึกษา (EDUCATION-NUM)

เป็น Feature ที่แสดงเกี่ยวกับระดับการศึกษาสูงสุดของแต่ละตัวอย่างในรูปแบบตัวเลข โดย Feature นี้เป็นประเภทเชิงปริมาณ (Numerical) แบบอัตราส่วน (Ratio) โดยข้อมูลมีลักษณะเบ้ไปทางซ้าย (Left-skewed) เล็กน้อย

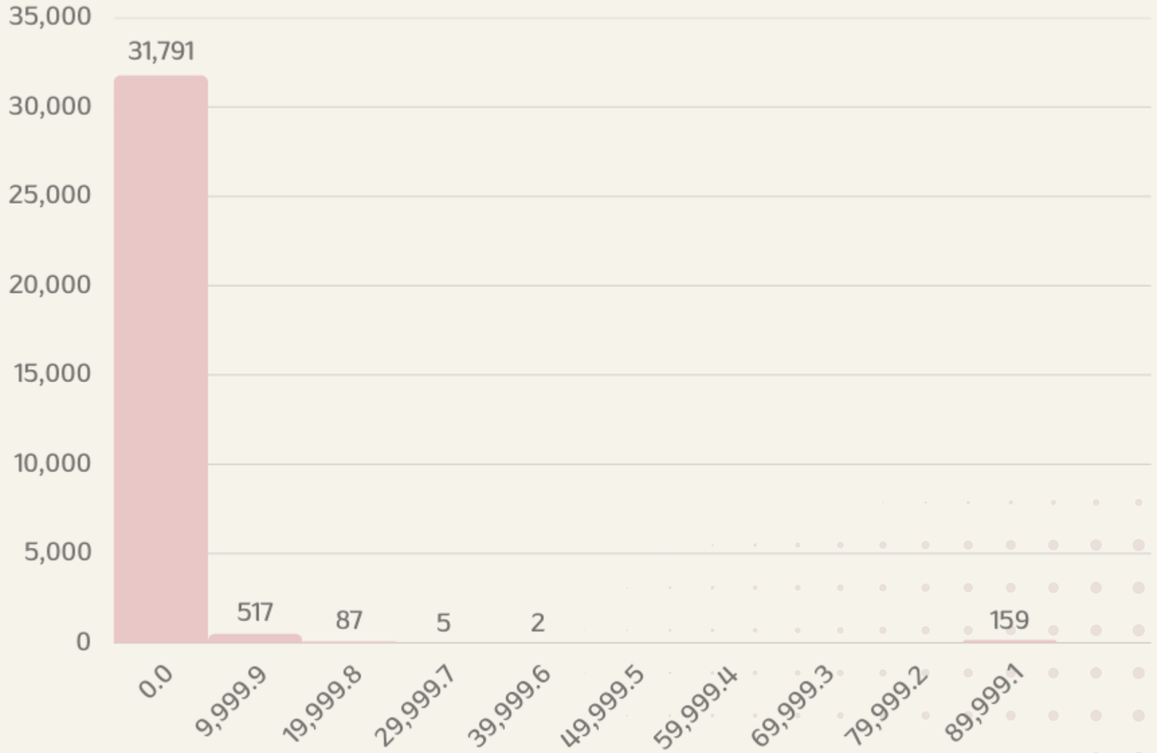
Mean	10.0807	Min	1
Median	10	Q_1	9
Mode	9	Q_2	10
SD.	2.5727	Q_3	12
Var.	6.6189	Max	16



ผลกำไรจากการลงทุน (CAPITAL-GAIN)

เป็น Feature ที่แสดงเกี่ยวกับกำไรหรือรายได้จากแหล่งการลงทุนที่ไม่ใช่เงินเดือน โดย Feature นี้เป็นประเภทเชิงปริมาณ (Numerical) แบบอัตราส่วน (Ratio) โดยข้อมูลมีลักษณะเบ้ไปทางขวา (Right-skewed) ในระดับสูง

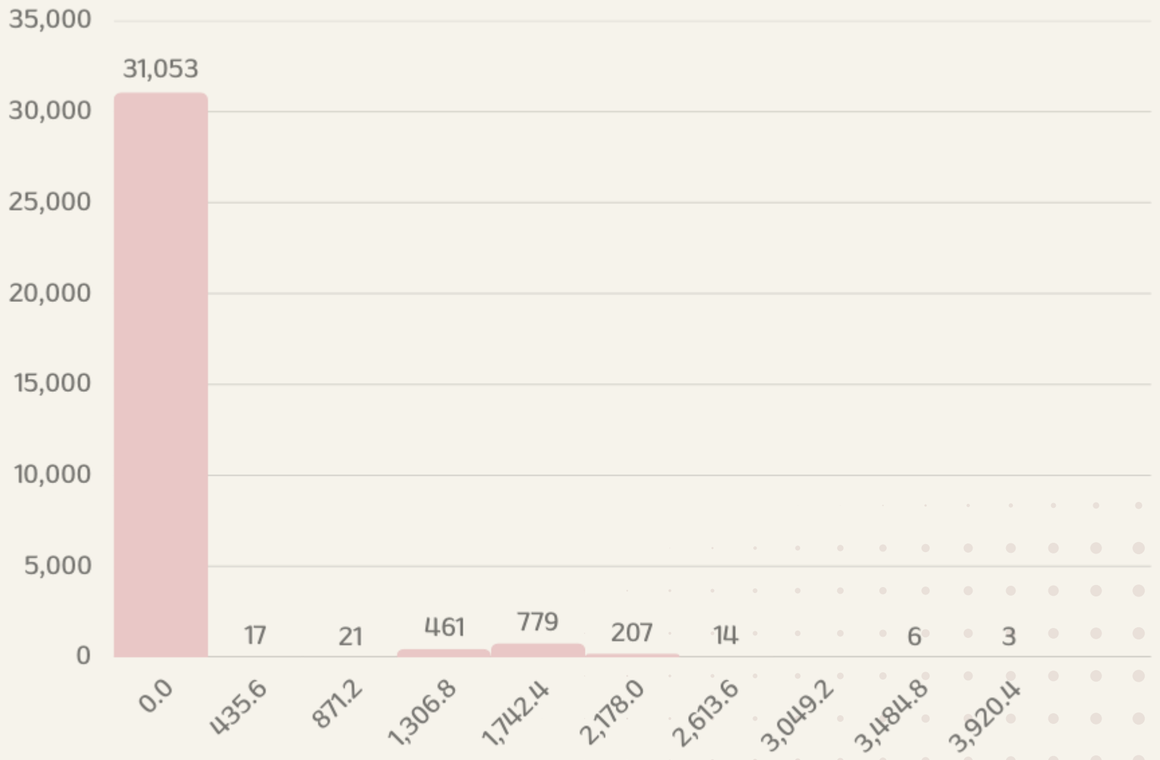
Mean	1,077.6488	Min	0
Median	0	Q_1	0
Mode	0	Q_2	0
SD.	7,385.2921	Q_3	0
Var.	54,542,539.1784	Max	99,999



ผลขาดทุนจากการลงทุน (CAPITAL-LOSS)

เป็น Feature ที่แสดงเกี่ยวกับการสูญเสียเงินจากการลงทุน โดย Feature นี้เป็นประเภทเชิงปริมาณ (Numerical) แบบอัตราส่วน (Ratio) โดยข้อมูลมีลักษณะเบ้ไปทางขวา (Right-skewed) ในระดับสูง

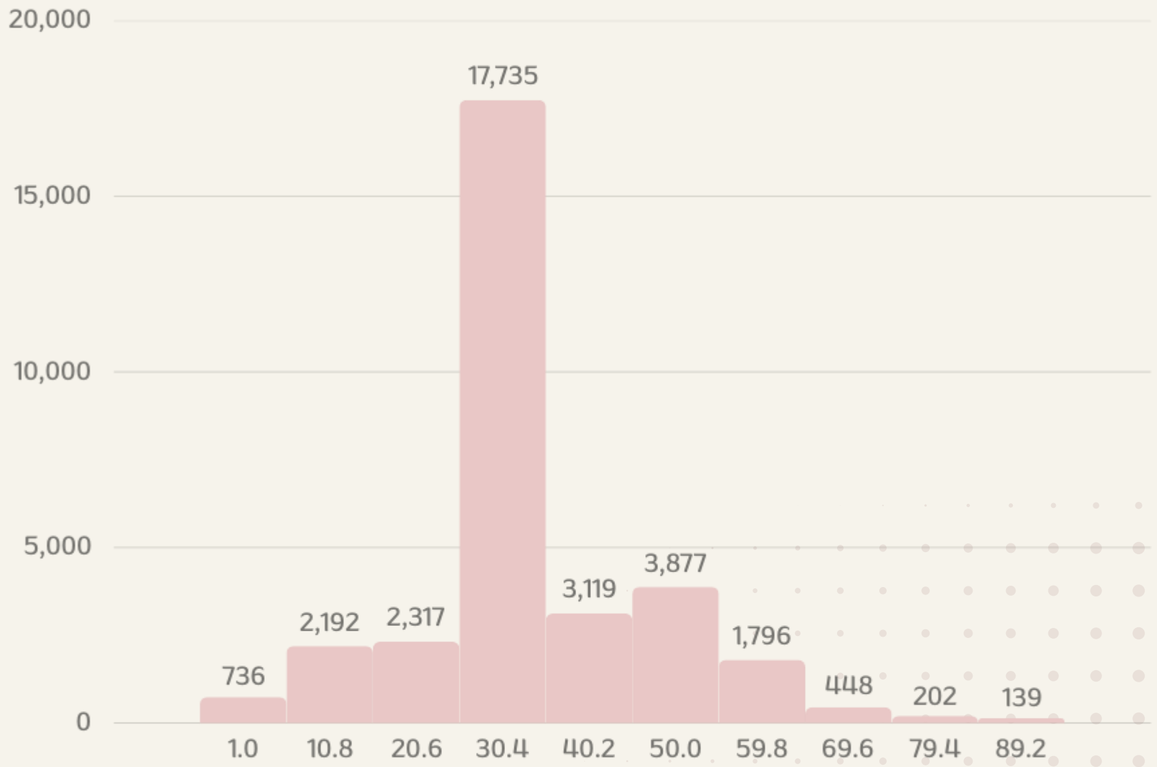
Mean	87.3038	Min	0
Median	0	Q_1	0
Mode	0	Q_2	0
SD.	402.9602	Q_3	0
Var.	162,376.9378	Max	4,356



ชั่วโมงการทำงาน (HOURS-PER-WEEK)

เป็น Feature ที่แสดงเกี่ยวกับจำนวนชั่วโมงที่บุคคลทำงานต่อสัปดาห์ โดย Feature นี้เป็นประเภทเชิงปริมาณ (Numerical) แบบอัตราส่วน (Ratio) โดยข้อมูลมีลักษณะเบ้ไปทางขวา (Right-skewed) เล็กน้อย

Mean	40.4375	Min	1
Median	40	Q_1	40
Mode	40	Q_2	40
SD.	12.3474	Q_3	45
Var.	152.459	Max	99

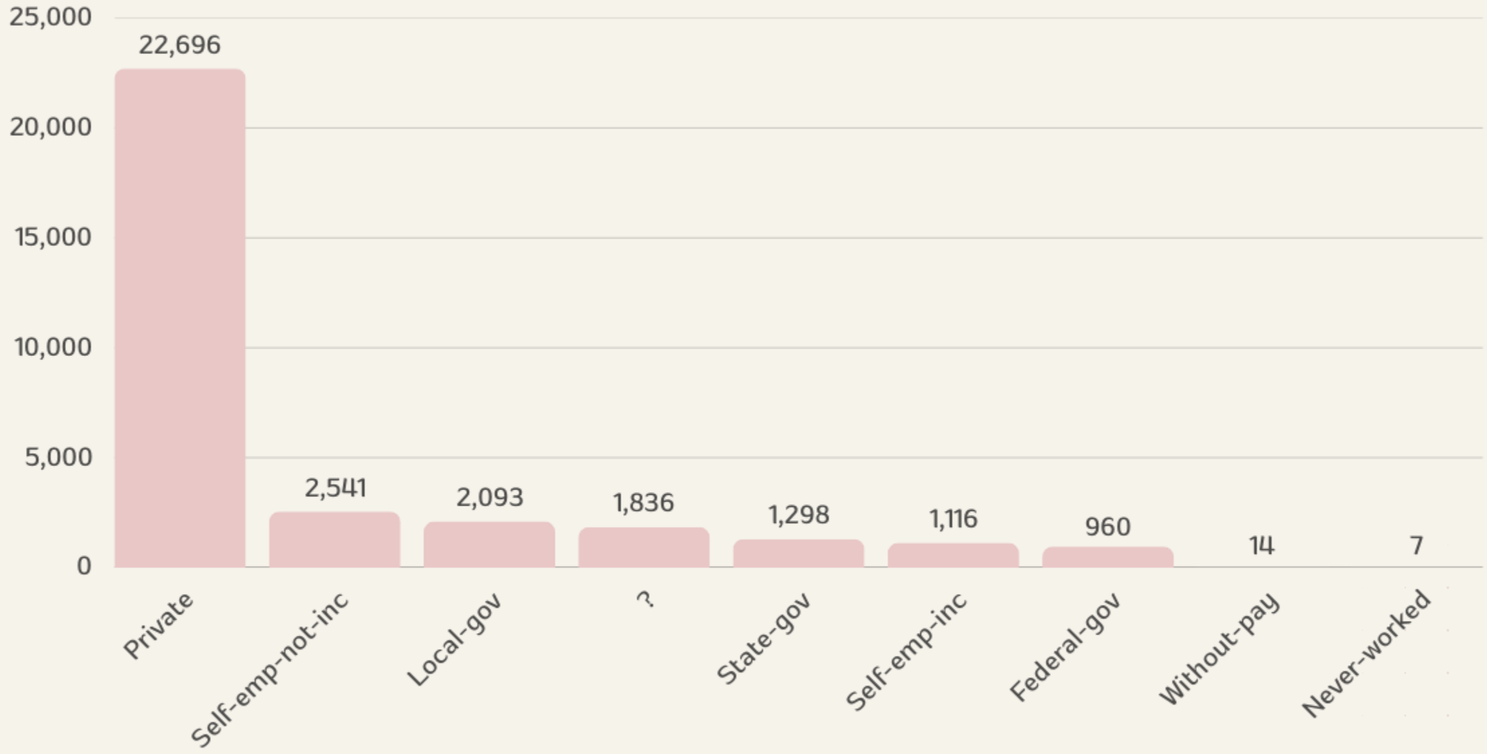


คุณสมบัติของแต่ละ FEATURE

ในประเภทเชิงคุณภาพ (CATEGORICAL)

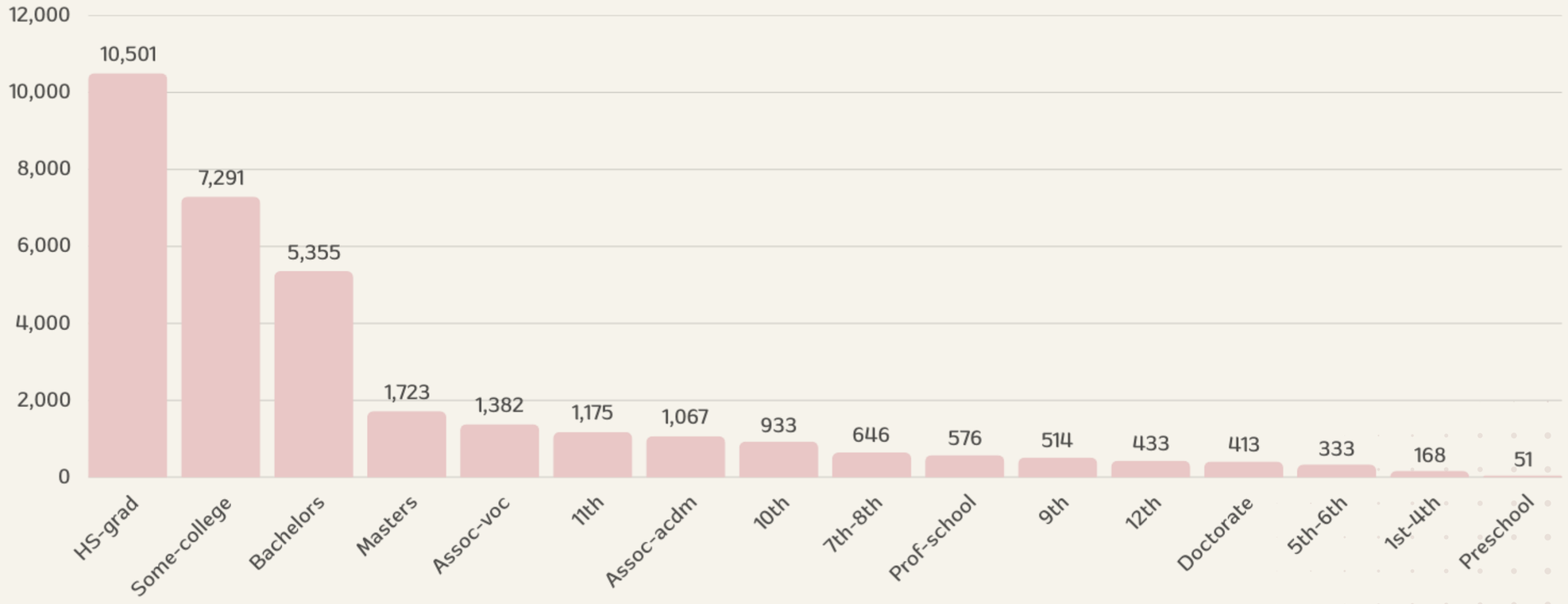
ประเภทงาน (WORK CLASS)

เป็น Feature ที่แสดงเกี่ยวกับประเภทขององค์กรที่ตัวอย่างของข้อมูลนั้น ๆ สังกัดอยู่ โดย Feature นี้เป็นประเภทเชิงคุณภาพ (Categorical) แบบไม่มีลำดับ (Nominal) โดยประกอบด้วย 9 หมวดหมู่ เช่น เอกชน (Private), ธุรกิจส่วนตัวแบบมีบริษัท (Self-emp-inc) เป็นต้น



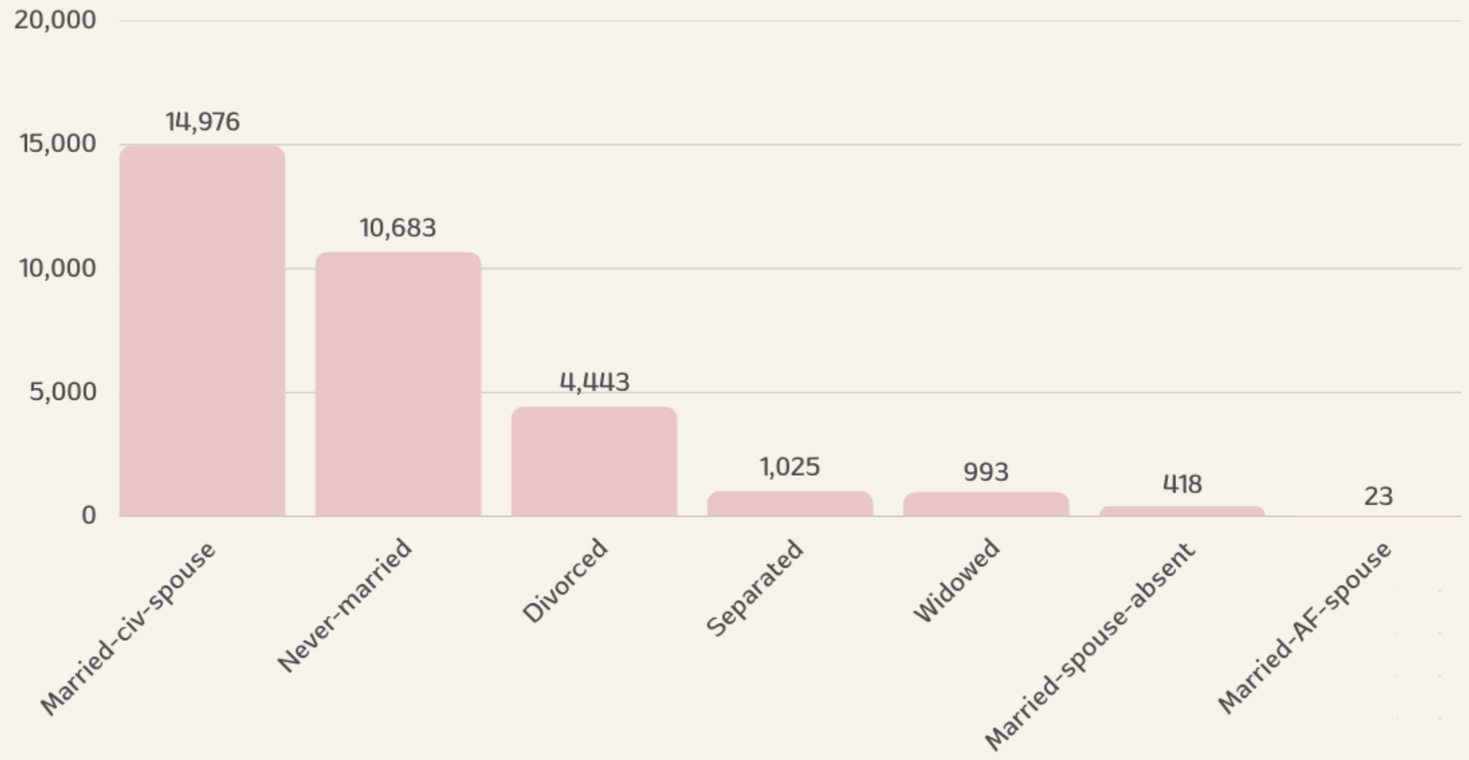
ระดับการศึกษา (EDUCATION)

เป็น Feature ที่แสดงเกี่ยวกับระดับการศึกษาสูงสุดของตัวอย่างข้อมูลนั้น ๆ โดย Feature นี้เป็นประเภทเชิงคุณภาพ (Categorical) แบบมีลำดับ (Ordinal) โดยประกอบด้วย 16 หมวดหมู่ เช่น ก่อนวัยเรียน (Preschool), มัธยมปลาย (HS-grad), ปริญญาตรี (Bachelors) เป็นต้น



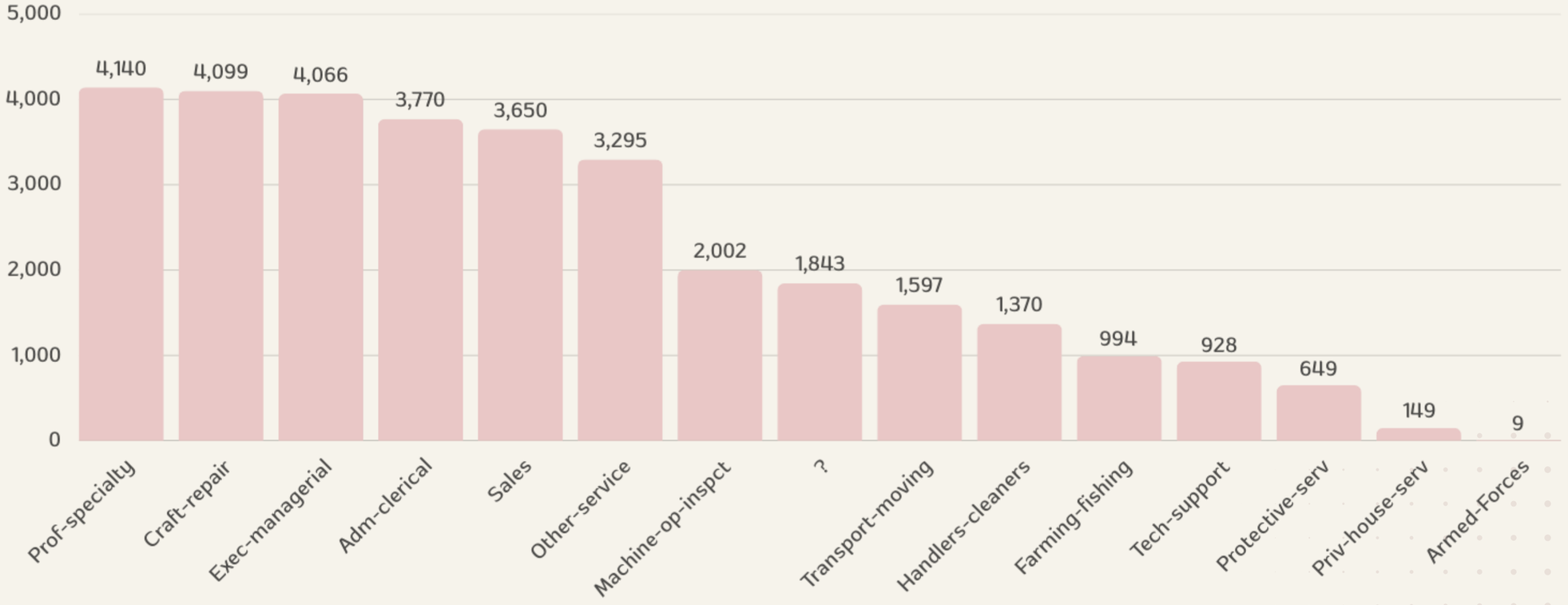
สถานภาพสมรส (MARITAL STATUS)

เป็น Feature ที่แสดงเกี่ยวกับสถานภาพสมรสของตัวอย่างข้อมูลนั้น ๆ โดย Feature นี้เป็นประเภทเชิงคุณภาพ (Categorical) แบบไม่มีลำดับ (Nominal) โดยประกอบด้วย 7 หมวดหมู่ เช่น โสด (Never-married), หย่าร้าง (Divorced), แยกกันอยู่ (Separated) เป็นต้น



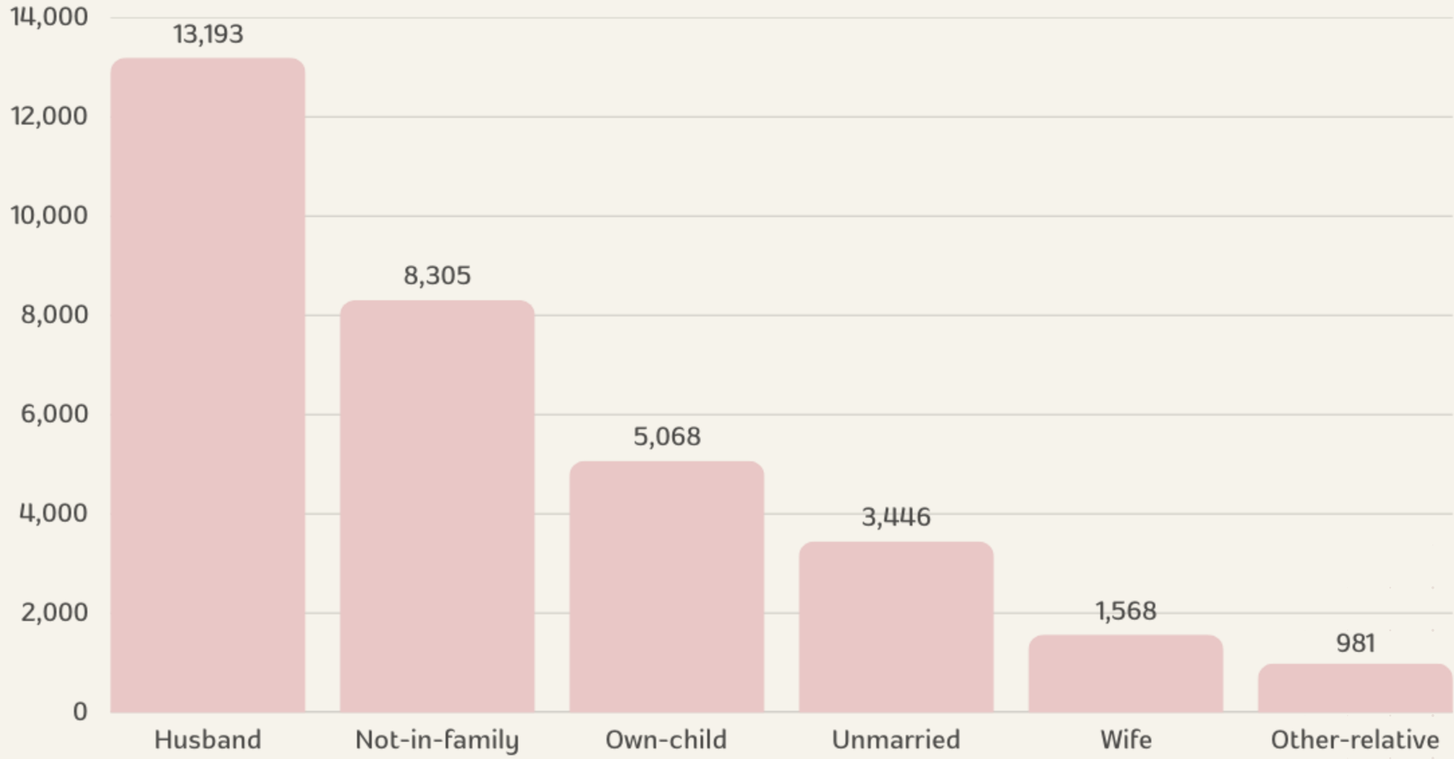
สาขาอาชีพ (OCCUPATION)

เป็น Feature ที่แสดงเกี่ยวกับประเภทหรือสาขาอาชีพที่ตัวอย่างข้อมูลนั้น ๆ ทำอยู่ โดย Feature นี้เป็นประเภทเชิงคุณภาพ (Categorical) แบบไม่มีลำดับ (Nominal) โดยประกอบด้วย 15 หมวดหมู่ เช่น ผู้เชี่ยวชาญเฉพาะทาง (Prof-specialty) เป็นต้น



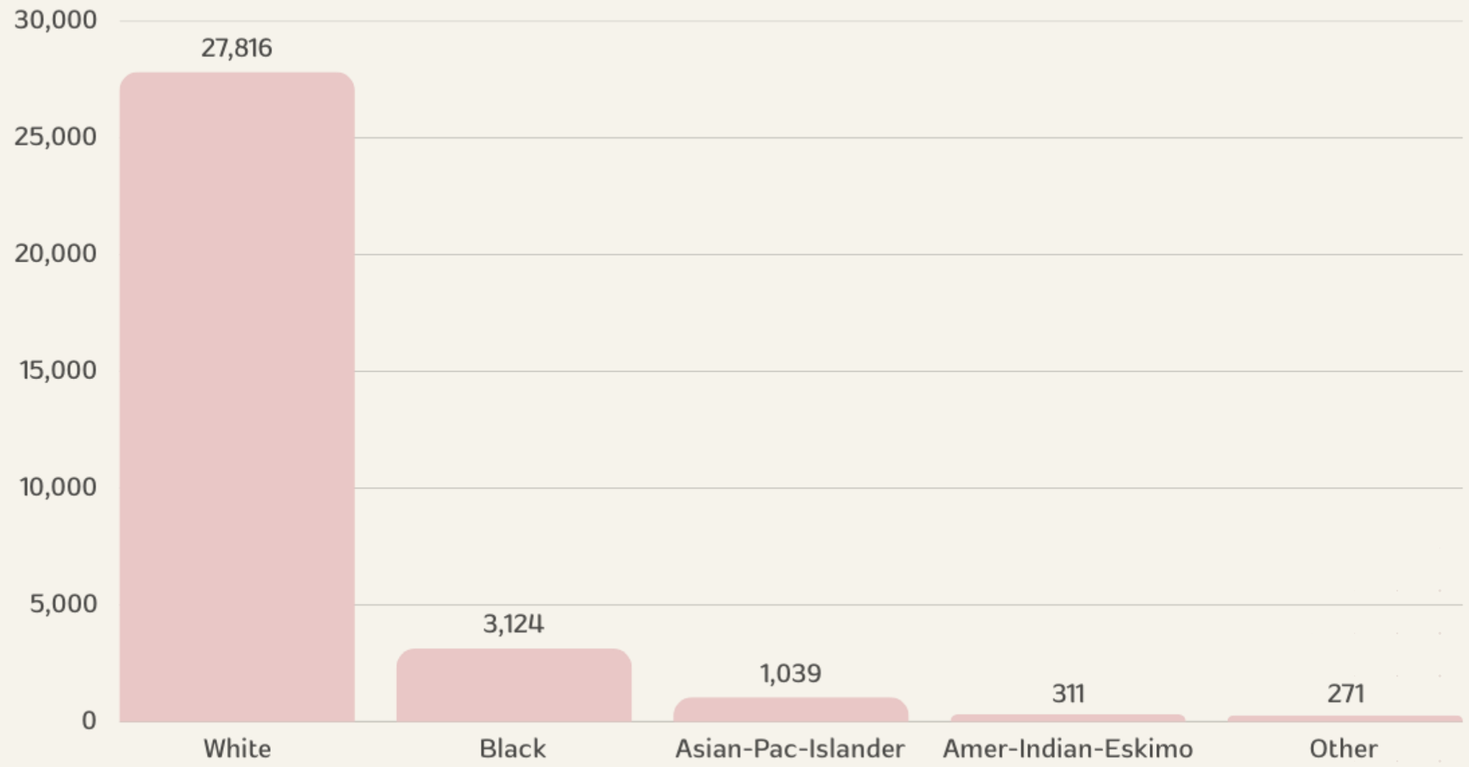
สถานะในครอบครัว (RELATIONSHIP)

เป็น Feature ที่แสดงเกี่ยวกับสถานะในครอบครัวของตัวอย่างข้อมูลนั้น ๆ โดย Feature นี้เป็นประเภทเชิงคุณภาพ (Categorical) แบบไม่มีลำดับ (Nominal) โดยประกอบด้วย 6 หมวดหมู่ เช่น สามี (Husband), ภรรยา (Wife) เป็นต้น



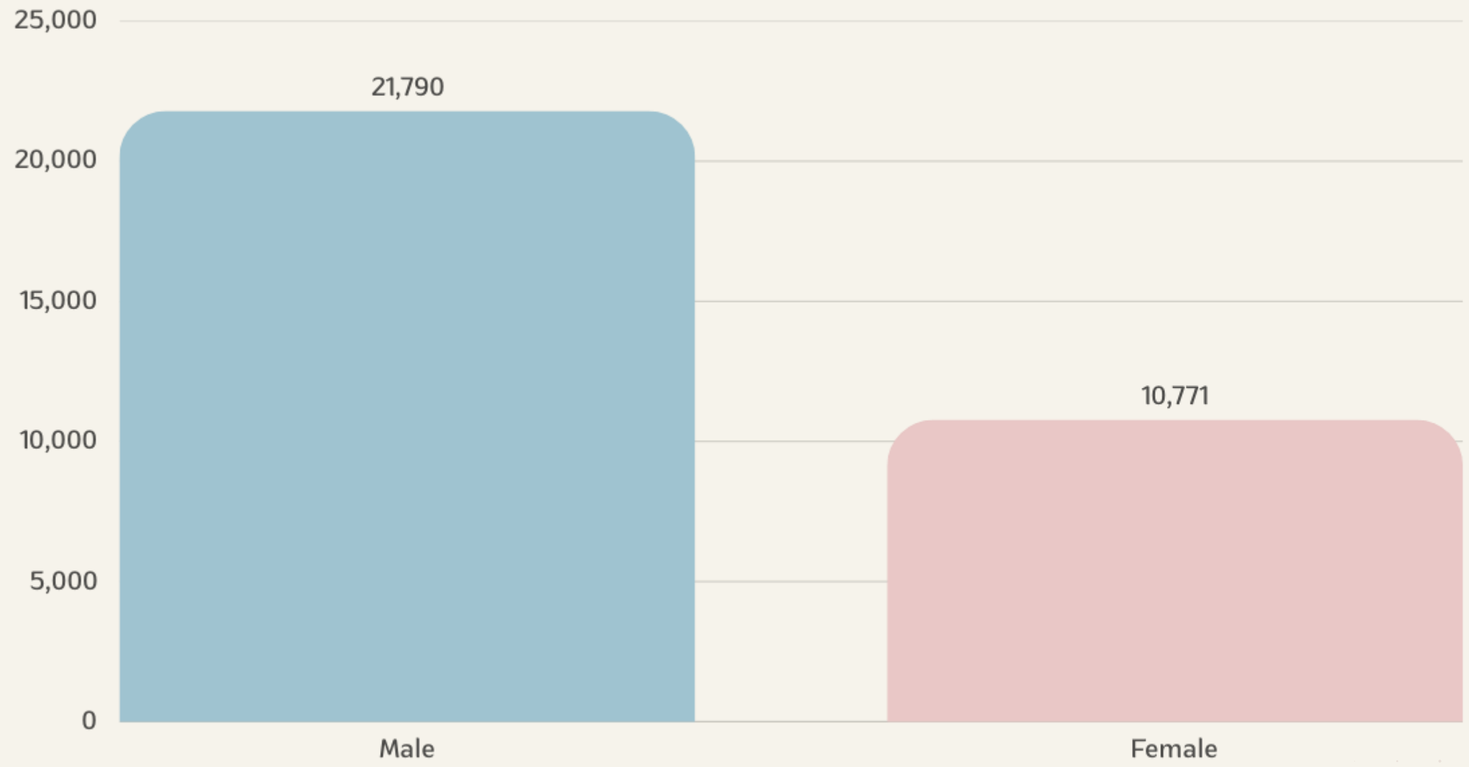
เชื้อชาติ (RACE)

เป็น Feature ที่แสดงเกี่ยวกับเชื้อชาติของตัวอย่างข้อมูลนั้น ๆ โดย Feature นี้เป็นประเภทเชิงคุณภาพ (Categorical) แบบไม่มีลำดับ (Nominal) โดยประกอบด้วย 5 หมวดหมู่ เช่น ผิวขาว (White), ผิวดำ (Black), อินเดียแดงและเอสกิโม (Amer-Indian-Eskimo) เป็นต้น



เพศ (SEX)

เป็น Feature ที่แสดงเกี่ยวกับเพศของตัวอย่างข้อมูลนั้น ๆ โดย Feature นี้เป็นประเภทเชิงคุณภาพ (Categorical) แบบไม่มีลำดับ (Nominal) โดยประกอบด้วย 2 หมวดหมู่ ได้แก่ ชาย (Male) และ หญิง (Female)

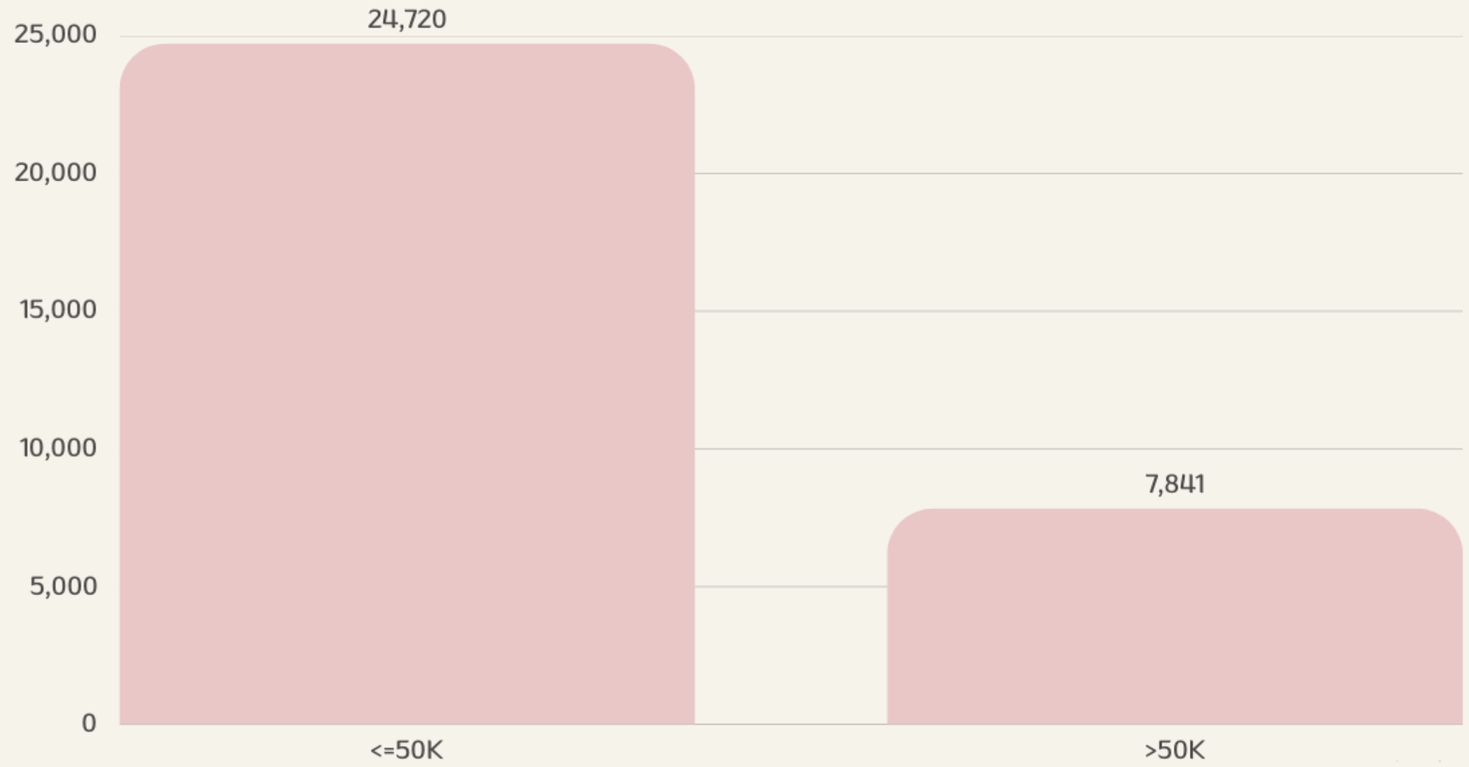


ถิ่นกำเนิด (NATIVE COUNTRY)

เป็น Feature ที่แสดงเกี่ยวกับถิ่นกำเนิดของตัวอย่างข้อมูลนั้น ๆ โดย Feature นี้เป็นประเภทเชิงคุณภาพ (Categorical) แบบไม่มีลำดับ (Nominal) โดยประกอบด้วย 42 หมวดหมู่ เช่น สหรัฐอเมริกา (United-States), เม็กซิโก (Mexico) เป็นต้น

รายได้ (INCOME)

เป็น Feature ที่แสดงเกี่ยวกับรายได้ของตัวอย่างข้อมูลนั้น ๆ โดย Feature นี้เป็นประเภทเชิงคุณภาพ (Categorical) แบบมีลำดับ (Ordinal) โดยประกอบด้วย 2 หมวดหมู่ ได้แก่ ไม่เกิน $50,000 (<=50K) และมากกว่า $50,000 (>50K)

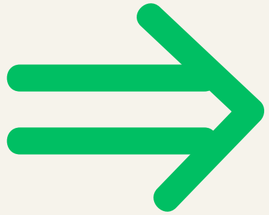


DATA PREPROCESS

MISSING DATA HANDLE

แทนที่ข้อมูลที่ว่าง (?) ของ Attribute “Work class” ด้วยค่า “Private”

Private	22,696
Self-emp-not-inc	2,541
Local-gov	2,093
?	1,836
State-gov	1,298
Self-emp-inc	1,116
Federal-gov	960
Without-pay	14
Never-worked	7



Private	24,532
Self-emp-not-inc	2,541
Local-gov	2,093
State-gov	1,298
Self-emp-inc	1,116
Federal-gov	960
Without-pay	14
Never-worked	7

MISSING DATA HANDLE

แทนที่ข้อมูลที่ว่าง (?) ของ Attribute “Occupation” ด้วยค่า “Other-service”

Prof-specialty	4,140
Craft-repair	4,099
Exec-managerial	4,066
Adm-clerical	3,770
Sales	3,650
Other-service	3,295
Machine-op-inspct	2,002
?	1,843

Transport-moving	1,597
Handlers-cleaners	1,370
Farming-fishing	994
Tech-support	928
Protective-serv	649
Priv-house-serv	149
Armed-Forces	9

Other-service	5,138	Transport-moving	1,597
Prof-specialty	4,140	Handlers-cleaners	1,370
Craft-repair	4,099	Farming-fishing	994
Exec-managerial	4,066	Tech-support	928
Adm-clerical	3,770	Protective-serv	649
Sales	3,650	Priv-house-serv	149
Machine-op-inspct	2,002	Armed-Forces	9

FEATURE ENGINEERING

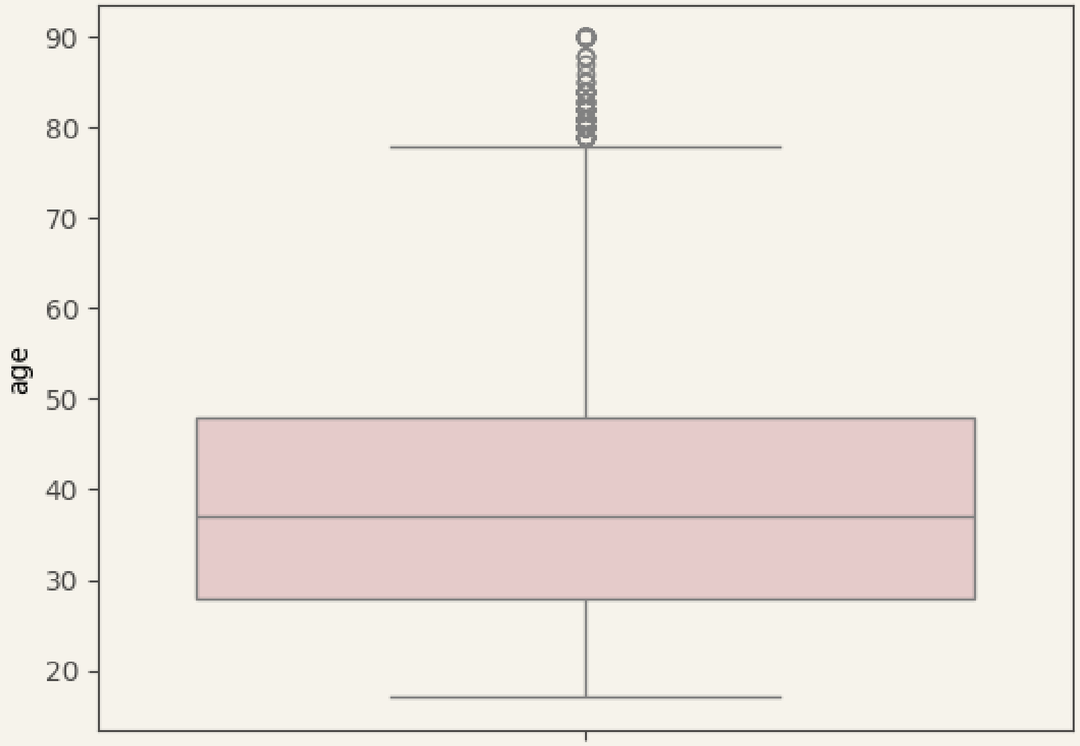
รวมข้อมูลระดับชั้นต่าง ๆ ของ Feature “Education” ให้เข้าเป็นกลุ่มเดียวกัน

HS-grad	10,501	7th - 8th	646
Some-college	7,291	Prof-school	576
Bachelors	5,355	9th	514
Masters	1,723	12th	433
Assoc-voc	1,382	Doctorate	413
11th	1,175	5th-6th	333
Assoc-acdm	1,067	1st-4th	168
10th	933	Preschool	51

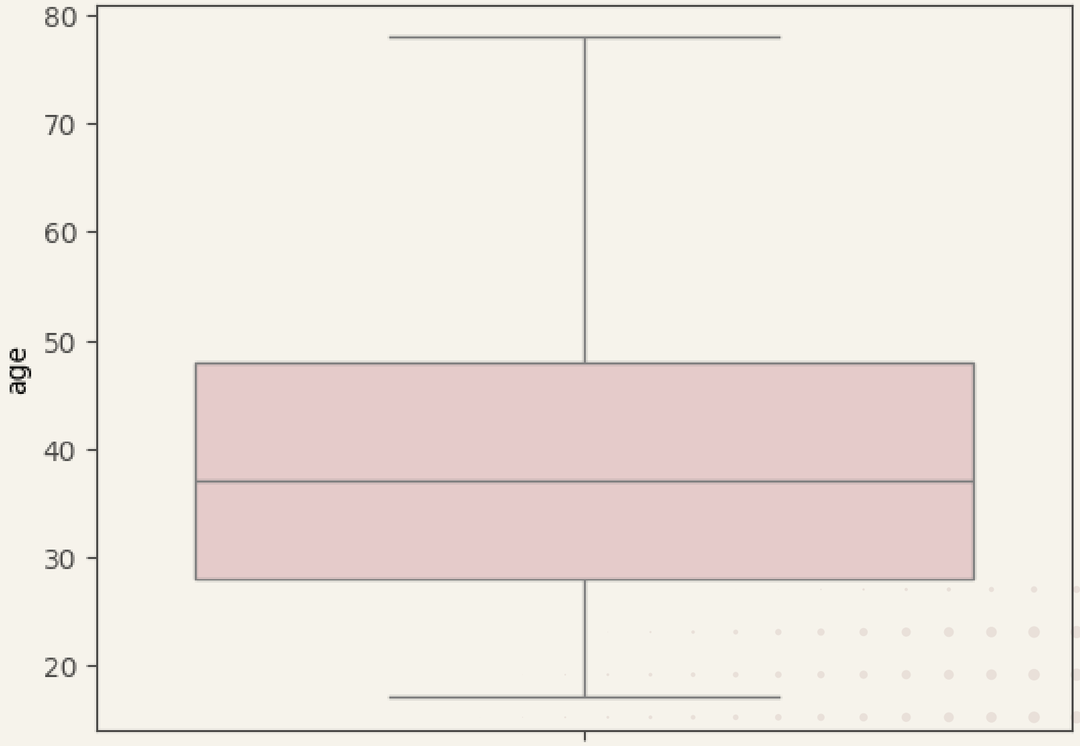
HS-grad	10,501
Bachelors	5,355
Assoc-Degree	1,382
Masters	1,723
JS-grad	10,501
Prof-school	576
Elementary	576
Doctorate	413
Preschool	51

HANDLING OUTLIERS

- กำจัด Outliers ใน Feature “Age”
- โดยช่วง Outliers จะอยู่ที่น้อยกว่า -2.0 (Lower Bound) และมากกว่า 78 (Upper Bound)



Box plot ก่อนการกำจัด Outliers



Box plot หลังการกำจัด Outliers

HANDLING OUTLIERS

- กำจัด Outliers ใน Feature “Age”
- โดยช่วง Outliers จะอยู่ที่น้อยกว่า -2.0 (Lower Bound) และมากกว่า 78 (Upper Bound)

Mean	38.5816	Min	17
Median	37	Q_1	28
Mode	36	Q_2	37
SD.	13.6404	Q_3	48
Var.	186.061	Max	90

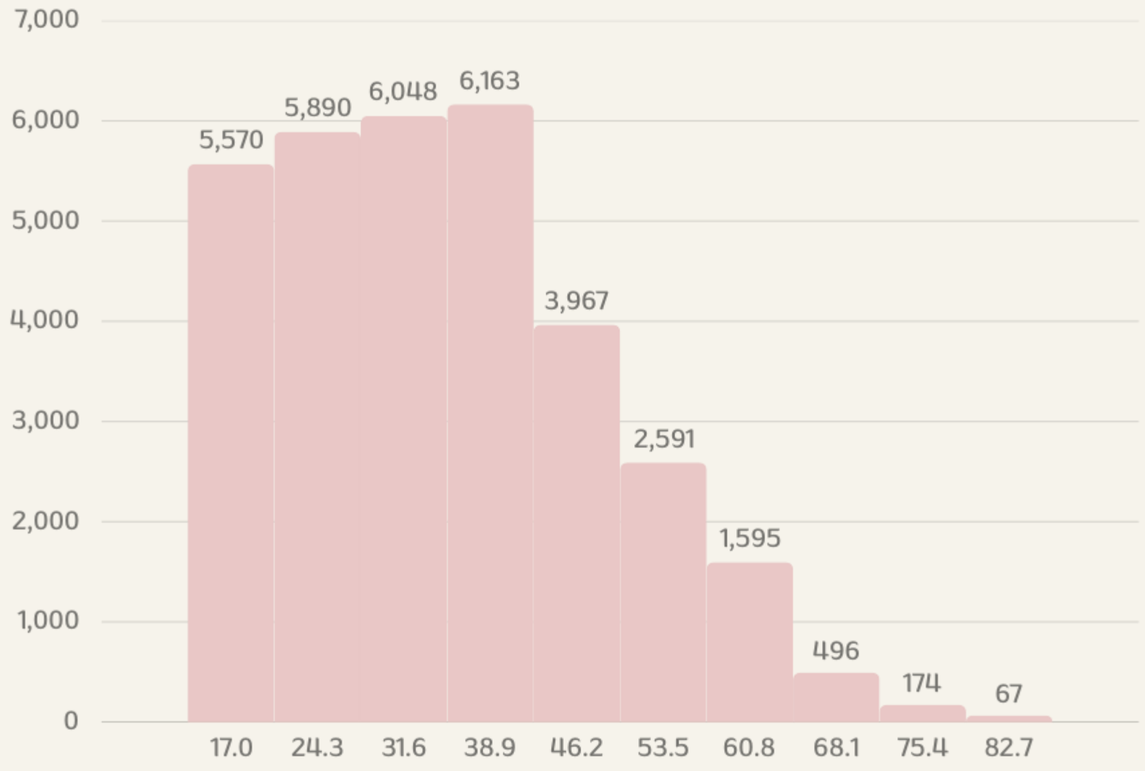
ค่าทางสถิติก่อนการกำจัด Outliers

Mean	38.5556	Min	17
Median	37	Q_1	28
Mode	36	Q_2	37
SD.	13.5561	Q_3	48
Var.	183.769	Max	78

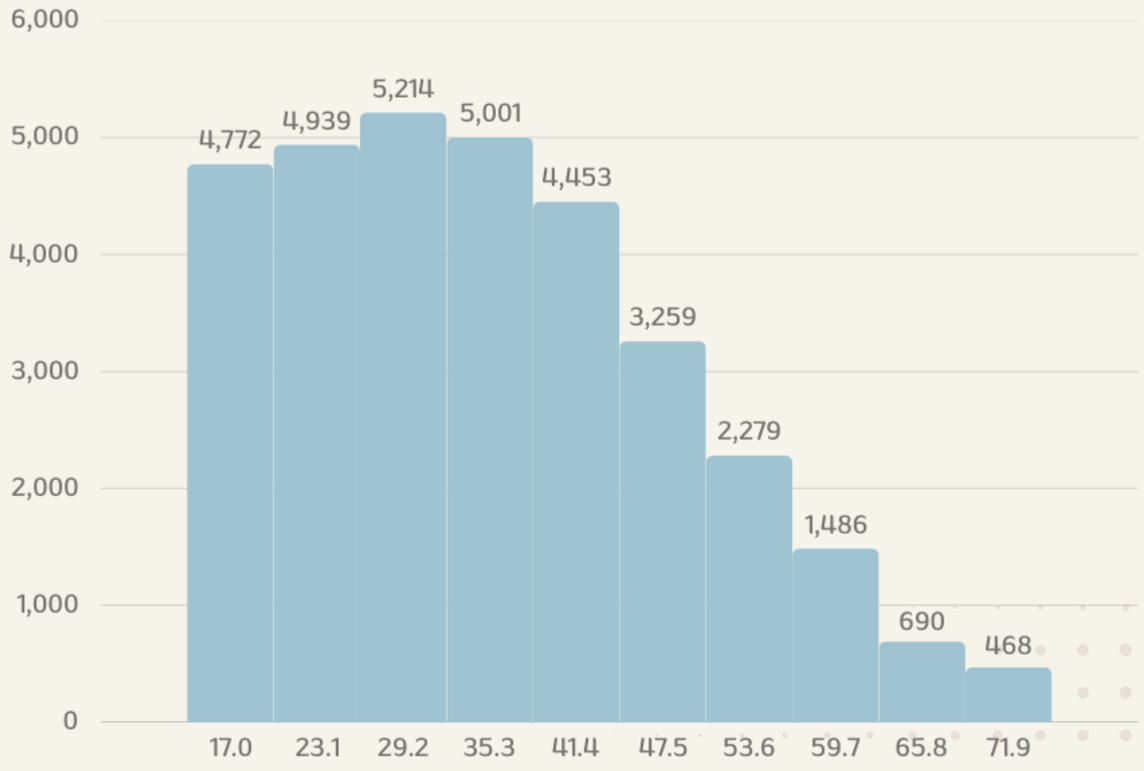
ค่าทางสถิติหลังการกำจัด Outliers

HANDLING OUTLIERS

- กำจัด Outliers ใน Feature “Age”
- โดยช่วง Outliers จะอยู่ที่น้อยกว่า -2.0 (Lower Bound) และมากกว่า 78 (Upper Bound)



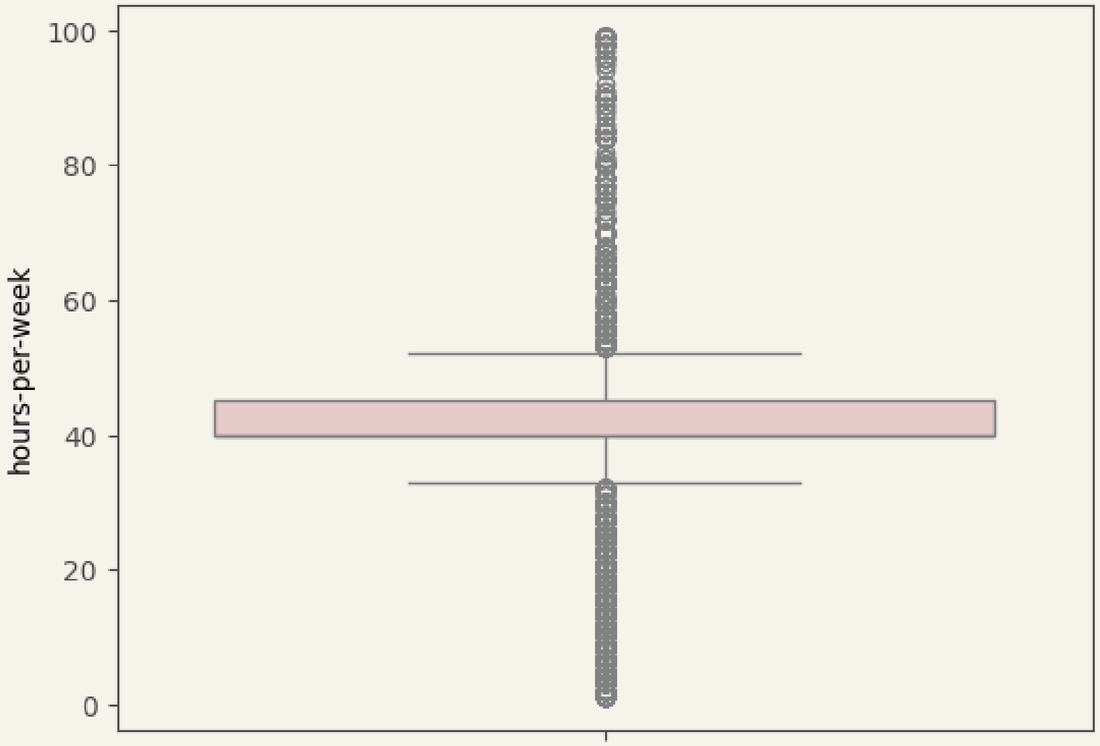
ฮิสโตแกรมก่อนการกำจัด Outliers



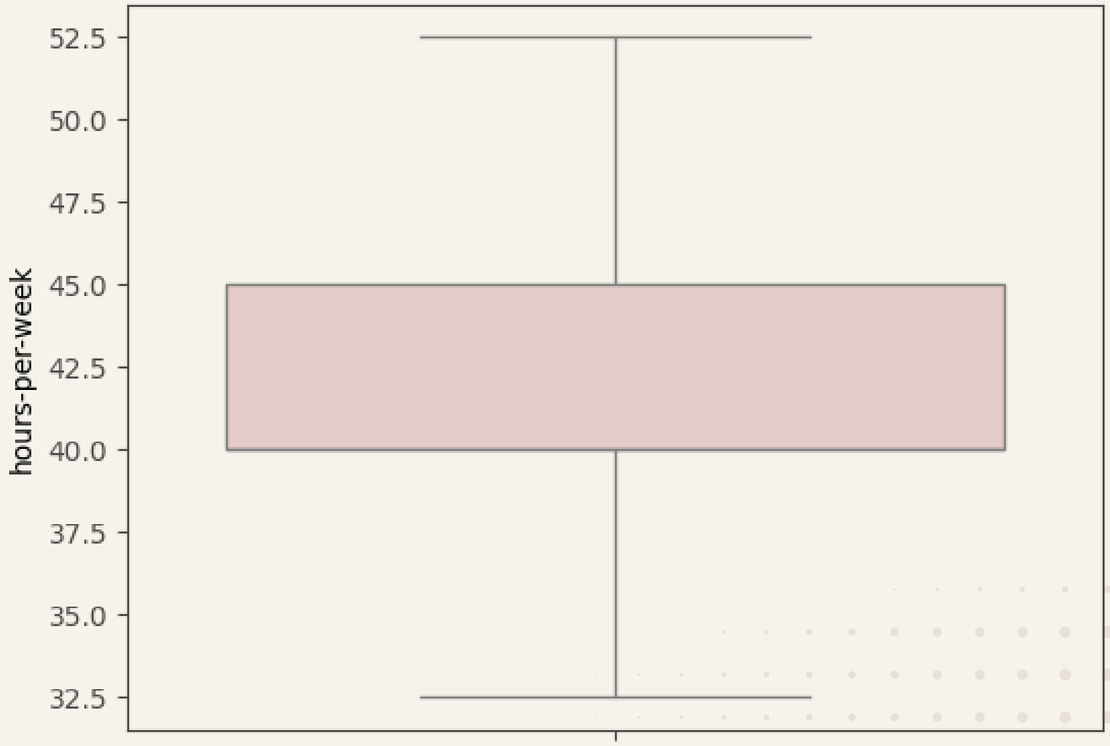
ฮิสโตแกรมหลังการกำจัด Outliers

HANDLING OUTLIERS

- กำจัด Outliers ใน Feature “Hours-per-week”
- โดยช่วง Outliers จะอยู่ที่น้อยกว่า 32.5 (Lower Bound) และมากกว่า 52.5 (Upper Bound)



Box plot ก่อนการกำจัด Outliers



Box plot หลังการกำจัด Outliers

HANDLING OUTLIERS

- กำจัด Outliers ใน Feature “Hours-per-week”
- โดยช่วง Outliers จะอยู่ที่น้อยกว่า 32.5 (Lower Bound) และมากกว่า 52.5 (Upper Bound)

Mean	40.4375	Min	1
Median	40	Q_1	40
Mode	40	Q_2	40
SD.	12.3474	Q_3	45
Var.	152.459	Max	99

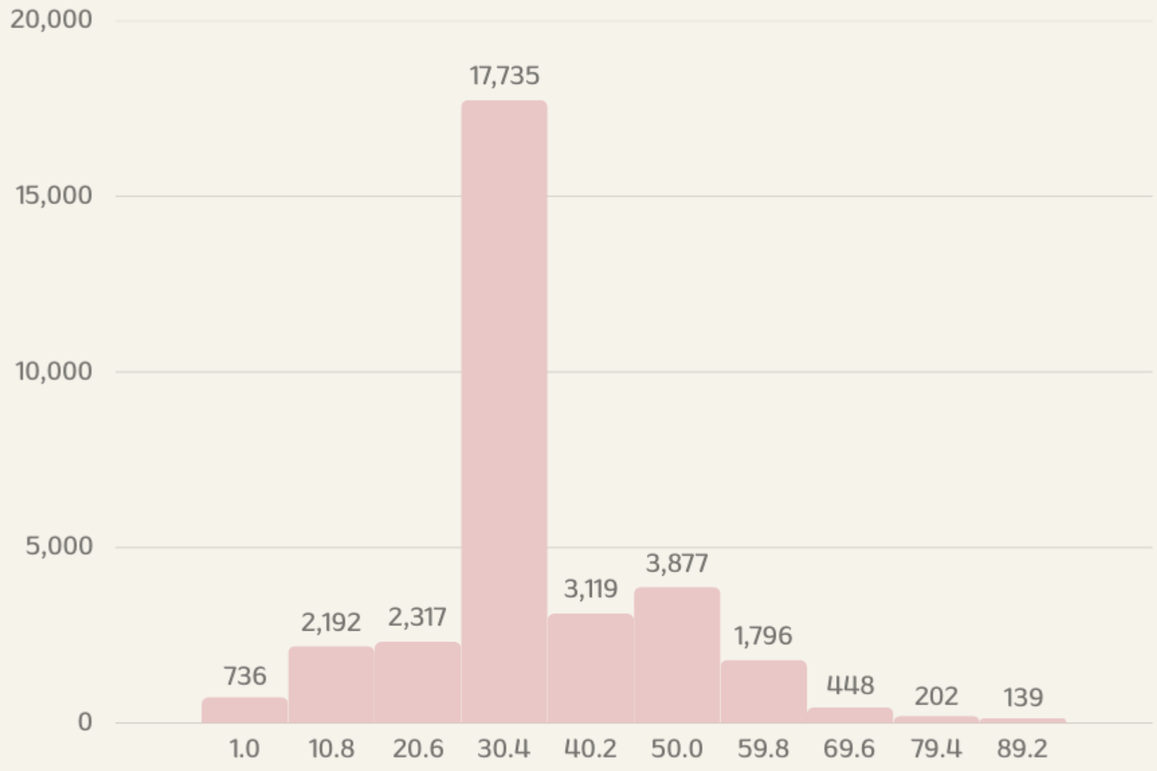
ค่าทางสถิติก่อนการกำจัด Outliers

Mean	41.2025	Min	32.5
Median	40	Q_1	40
Mode	40	Q_2	40
SD.	6.187	Q_3	45
Var.	38.279	Max	52.5

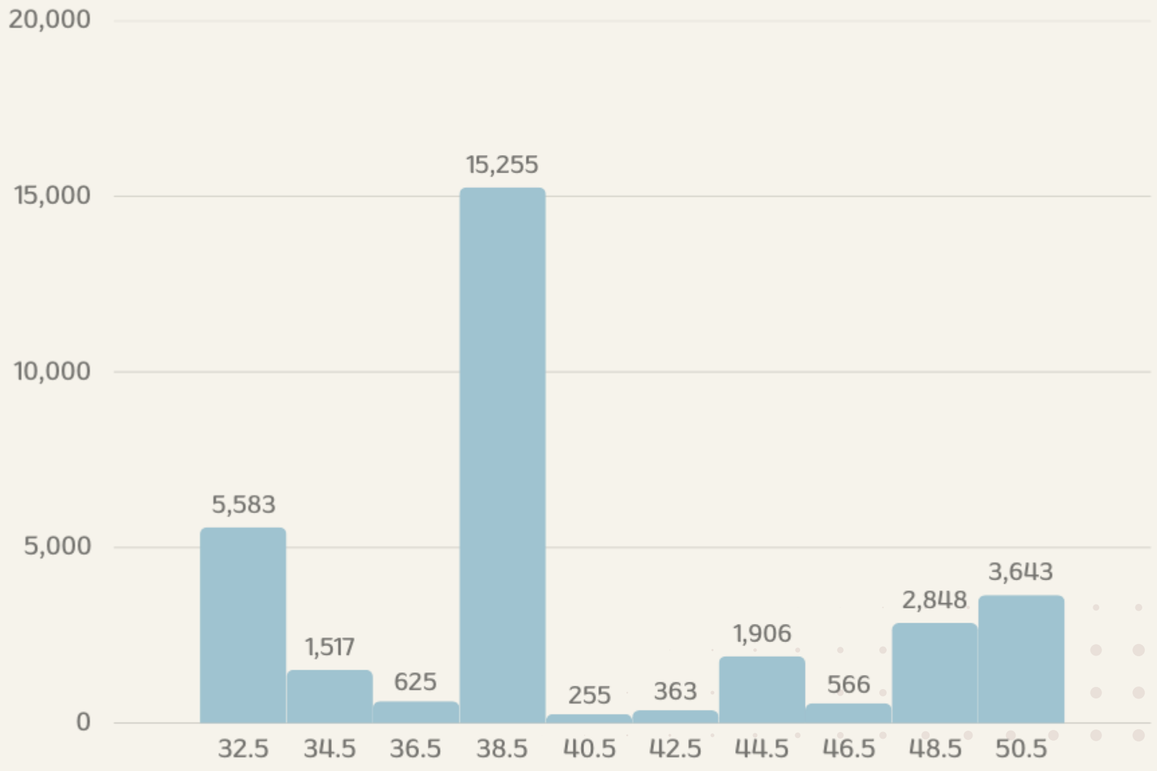
ค่าทางสถิติหลังการกำจัด Outliers

HANDLING OUTLIERS

- กำจัด Outliers ใน Feature “Hours-per-week”
- โดยช่วง Outliers จะอยู่ที่น้อยกว่า 32.5 (Lower Bound) และมากกว่า 52.5 (Upper Bound)



ฮิสโตแกรมก่อนการกำจัด Outliers



ฮิสโตแกรมหลังการกำจัด Outliers

ขอบคุณครับ

6505000270 นายศุภณัฐ แซ่เตีย