X	Y	Z
13	1927	cat
NaN	2300	dog
15	NaN	bird
16	5959	cat
16	AB	cat
NaN	4594	dog
19	1927	cat
20	2879	birdหมา
21	NaN	NaN
0	4096	cat
A	6730	cat
25	0	bird
0	2792	dog
33	2575	dogนก
1000	4959	bird
19	1927	cat
36	4580	dog
40	5869	NaN
NaN	4178	dog
45	NaN	cat

ตารางจากการทำ Data Cleaning

X	Y	Z
13	1927	cat
13	2300	dog
15	2300	bird
16	5959	cat
16	5959	cat
16	4594	dog
19	1927	cat
20	2879	cat
21	2879	cat
0	4096	cat
0	6730	cat
25	0	bird
25	2792	dog
33	2575	dog
1000	4959	bird
36	4580	dog
40	5869	dog
40	4178	dog
45	4178	cat

ตารางจากการทำ Data Cleaning และ Sort column X&Y เพื่อหา Outliers

X	Y	Z
0	0	cat
0	1927	dog
13	1927	bird
13	2300	cat
15	2300	cat
16	2575	dog
16	2792	cat
16	2879	cat
19	2879	cat
20	4096	cat
21	4178	cat
25	4178	bird
25	4580	dog
33	4594	dog
36	4959	bird
40	5869	dog
40	5959	dog
45	5959	dog
1000	6730	cat

: แทนที่ Missing Values & Invalid Values ด้วยการแทนที่ copy ค่าบนลงค่าล่าง และตัด Duplicated Samples ทิ้ง

1. ระบุจำนวน variables (features, dimension), Samples, Missing values, Duplicated Samples, Invalid values, Outliers (if any)

Variable	3
Samples	20
Missing Values	8
Duplicated Samples	1
Invalid Values	4

แถวที่ 9 & 18

Outliers : แทนที่ Missing Values & Invalid Values ด้วยการแทนที่ copy ค่าบนลงค่าล่าง และ ตัด Duplicated Samples ทิ้ง

X	Q1	15.5		
	Q2	20		
	Q3	34.5		
	IQR	19		Amout
	Maximum Outliers	63	> Maximum Outliers	1 ; (1000)
	Minimum Outliers	-13	< Minimum Outliers	0

Y	Q1	2437.5		
	Q2	4096		
	Q3	4776.5		
	IQR	2339		Amout
	Maximum Outliers	8285	> Maximum Outliers	0
	Minimum Outliers	-1071	< Minimum Outliers	0

2. ทำการปรับแก้ข้อมูล (Data Cleaning) ปรับช่วงของข้อมูล (Max-Min norm และ Standardize)

X		Y	
MAX-X	1000	MAX-Y	6730
MIN-X	0	MIN-Y	0
MEDIAN-X	20	MEDIAN-Y	4096
SD-X	224.7546941	SD-Y	1751.058615

Max-Min norm	Standardize	Max-Min norm	Standardize
0.013	-0.03114506697	0.2863298663	-1.238679266
0.013	-0.03114506697	0.3417533432	-1.025665266
0.015	-0.02224647641	0.3417533432	-1.025665266
0.016	-0.01779718113	0.8854383358	1.063927834
0.016	-0.01779718113	0.8854383358	1.063927834
0.016	-0.01779718113	0.682615156	0.2843993888
0.019	-0.004449295282	0.2863298663	-1.238679266
0.02	0	0.4277860327	-0.6950081449
0.021	0.004449295282	0.4277860327	-0.6950081449
0	-0.08898590564	0.6086181278	0
0	-0.08898590564	1	1.504232912
0.025	0.02224647641	0	-2.339156419
0.025	0.02224647641	0.414858841	-0.7446923755
0.033	0.05784083867	0.382615156	-0.8686174104
1	4.360309376	0.7368499257	0.492844724
0.036	0.07118872451	0.6805349183	0.2764042253
0.04	0.08898590564	0.8720653789	1.012530354
0.04	0.08898590564	0.6208023774	0.04682881502
0.045	0.111232382	0.6208023774	0.04682881502

Max-Min norm

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardize (Z-Norm / Z-score norm)

$$x' = z = \frac{x - \mu}{\sigma}$$

3. คำนวนค่าสถิติของข้อมูลตัวเลข (mean, variance, standard deviation (std))

		Max-Min norm	Standardize
X	mean	0.029	0.04004365754
	variance	0.000512	0.010135669
	standard deviation	0.022627417	0.1006760597

		Max-Min norm	Standardize
Y	mean	0.4535661218	-0.5959252253
	variance	0.05593593036	0.8262655128
	standard deviation	0.2365077808	0.9089914811

4. เปลี่ยนข้อมูลตัวอักษร (String) ให้เป็นตัวเลข คำนวนค่า histogram และ normalized histogram

กำหนดให้ ;

name	value
bird	0
cat	1
dog	2

Z	Z	Normalized Z
cat	1	NaN
dog	2	NaN
bird	0	NaN
cat	1	NaN
cat	1	1
dog	2	1.2
cat	1	1
cat	1	1.2
cat	1	1.2
cat	1	1.2
cat	1	1
bird	0	0.8
dog	2	1
dog	2	1.2
bird	0	1
dog	2	1.2
dog	2	1.6
dog	2	1.6
cat	1	1.4