Mondrian *k*-Anonymity Testing

Expected Output Tests

Test Case 1

Non-anonymised data:

attrib1	attrib2	attrib3	attrib4
10	a	x	blue
20	b	y	red
20	b	z	green
30	a	z	yellow
10	b	x	orange
10	a	y	red

Test:

Create Strict 2-Anonymisation using QIDs {attrib1, attrib2}, removing headers in process

Input:

anon.py test1.csv 0,1 2 strict_test1_2_2.csv 1 1

Pass Conditions:

- Headers removed
- All records in actual output match expected output, disregarding order

Expected output:

10	[a,b]	x	blue
10	[a,b]	y	red
10	[a,b]	x	orange
[20,30]	[a,b]	y	red
[20,30]	[a,b]	z	green
[20,30]	[a,b]	z	yellow

Actual output:

10	['a', 'b']	x	blue
10	['a', 'b']	x	orange
10	['a', 'b']	y	red
[20, 30]	['a', 'b']	y	red
[20, 30]	['a', 'b']	z	green
[20, 30]	['a', 'b']	z	yellow

Passed: Yes

Test Case 2

Non-anonymised data:

10	a	x	blue
20	b	y	red
20	b	z	green
30	a	z	yellow
10	b	x	orange
10	a	y	red
30	b	z	orange
30	b	y	red
30	a	z	blue
30	b	x	green
30	a	x	red

Test:

Create Relaxed 2-Anonymisation using QIDs {attrib1, attrib2}

Input:

anon.py test2.csv 0,1 2 relaxed_test2_2_2.csv 1 0

Pass Conditions:

- *k*-Anonymity tool should produce an output due to generalisation value intersections – noting strict output is not possible for this dataset and parameters
- All records in actual output should match expected output, disregarding order

Expected output:

10	[a,b]	x	blue
10	[a,b]	x	orange
10	[a,b]	y	red
[20,30]	[a,b]	y	red
[20,30]	[a,b]	z	green
[20,30]	[a,b]	z	yellow
30	[a,b]	z	orange
30	[a,b]	z	blue
30	[a,b]	x	red
30	b	y	red
30	b	x	green

Actual output:

10	['a', 'b']	x	blue
10	['a', 'b']	x	orange
10	['a', 'b']	y	red
[20, 30]	['a', 'b']	y	red
[20, 30]	['a', 'b']	z	green
[20, 30]	['a', 'b']	z	yellow
30	['a', 'b']	z	blue
30	['a', 'b']	x	red
30	['a', 'b']	z	orange
30	b	y	red
30	b	x	green

Passed: Yes

Test Case 3

Non-anonymised data:

attrib1	attrib2	attrib3	attrib4
10	a	x	blue
20	b	y	red
20	b	z	green
30	a	z	yellow
10	b	x	orange
10	a	y	red

Test:

Create Strict 2-Anonymisation using QIDs {attrib1, attrib2, attrib3}, removing headers in process

Input:

anon.py test1.csv 0,1,2 2 strict_test1_3_2.csv 1 1

Pass Conditions:

- Headers removed
- All records in actual output should match expected output, disregarding order

Expected output:

10	[a,b]	[x,y]	blue
10	[a,b]	[x,y]	orange
10	[a,b]	[x,y]	red
[20, 30]	[a,b]	[y,z]	red
[20, 30]	[a,b]	[y,z]	yellow
[20, 30]	[a,b]	[y,z]	green

Actual output:

10	['a', 'b']	['x', 'y']	blue
10	['a', 'b']	['x', 'y']	orange
10	['a', 'b']	['x', 'y']	red
[20, 30]	['a', 'b']	['y', 'z']	red
[20, 30]	['a', 'b']	['y', 'z']	green
[20, 30]	['a', 'b']	['y', 'z']	yellow

Passed: Yes

Test Case 4

Non-anonymised data:

10	a	x	blue
20	b	y	red
20	b	z	green
30	a	z	yellow
10	b	x	orange
10	a	y	red
30	b	z	orange
30	b	y	red
30	a	z	blue
30	b	x	green
30	a	x	red

Test:

Create Relaxed 5-Anonymisation using QIDs {attrib1, attrib2}

Input:

anon.py test2.csv 0,1 5 strict_test2_2_5.csv 0 0

Pass Conditions:

- *k*-Anonymity tool should produce an output due to generalisation value intersections – noting strict output is not possible for this dataset and parameters
- All records in actual output should match expected output, disregarding order

Expected output:

[10,20,30]	[a,b]	x	blue
[10,20,30]	[a,b]	x	orange
[10,20,30]	[a,b]	y	red
[10,20,30]	[a,b]	y	red
[10,20,30]	[a,b]	z	green
[10,20,30]	[a,b]	z	yellow
30	[a,b]	z	orange
30	[a,b]	y	red
30	[a,b]	z	blue
30	[a,b]	x	green
30	[a,b]	x	red

Actual output:

[10, 20, 30]	['a', 'b']	x	blue
[10, 20, 30]	['a', 'b']	y	red
[10, 20, 30]	['a', 'b']	z	green
[10, 20, 30]	['a', 'b']	x	orange

[10, 20, 30]	['a', 'b']	y	red
[10, 20, 30]	['a', 'b']	z	yellow
30	['a', 'b']	z	orange
30	['a', 'b']	y	red
30	['a', 'b']	z	blue
30	['a', 'b']	x	green
30	['a', 'b']	x	red

Passed: Yes

Test Case 5

Non-anonymised data:

attrib1	attrib2	attrib3	attrib4
10	a	x	blue
20	b	y	red
20	b	z	green
30	a	z	yellow
10	b	x	orange
10	a	y	red

Test:

Create Strict 100-Anonymisation using QIDs {attrib1, attrib2}. Sanity test – impossible, should produce no output

Input:

anon.py test1.csv 0,1 100 strict_test1_2_100.csv 1 1

Pass Conditions:

- *k*-Anonymity tool should produce error stating that anonymisation is impossible for selected parameters

Output:

No anonymisation and following printed to terminal:

```
STRICT
Cannot be anonymised for K =  100
```

Passed: Yes

Test Case 6

Non-anonymised data:

attrib1	attrib2	attrib3	attrib4
10	a	x	blue
20	b	y	red
20	b	z	green
30	a	z	yellow
10	b	x	orange
10	a	y	red

Test:

Create Strict 1-Anonymisation using QIDs {attrib1, attrib2}. 1-Anonymisation is not a valid anonymisation, essentially resulting in original data. Headers should be removed regardless. Note: Mondrian algorithm should produce smallest possible partitions – therefore, whilst *k* is a minimum size for equivalence classes, a 1-Anonymisation should never have equivalence classes larger than 1. This case essentially tests optimal output of algorithm.

Input:

anon.py test1.csv 0,1 1 strict_test1_2_1.csv 1 1

Pass Conditions:

- Headers removed
- *k*-Anonymity tool should output original data, re-ordered to suggest equivalence classes were made and generalisation was performed on partitions of size = 1

Output:

10	a	x	blue
10	a	y	red
10	b	x	orange
20	b	y	red
20	b	z	green
30	a	z	yellow

Passed: Yes