k-Anonymity Library Demo with k=3

In [1]:
```
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
import kAnonymityLib as daio_dpt
import pandas as pd
```

In [3]:
```
dai_anonymization = daio_dpt.kAnonymity()
print(dai_anonymization)
```

```
k-Anonymity Class Library with k=3
```

In [4]:
```
names = dai_anonymization.file_to_list("headers.txt")
dai_anonymization.set_headers(names)
dai_anonymization.read_datafile("adult-all.txt")
```

In [5]:
```
df = dai_anonymization.dataframe
```

In [6]:
```
df.workclass.unique()
```

Out[6]:
```
array([' State-gov', ' Self-emp-not-inc', ' Private', ' Federal-gov',
       ' Local-gov', ' ?', ' Self-emp-inc', ' Without-pay',
       ' Never-worked'], dtype=object)
```

In [7]:
```
df.age.describe()
```

Out[7]:
```
count    48842.000000
mean        38.643585
std         13.710510
min         17.000000
25%         28.000000
50%         37.000000
75%         48.000000
max         90.000000
Name: age, dtype: float64
```

In [8]:
```
age_range = lambda age: ("<= 20" if age <= 20
    else ("21 - 30" if age <= 30
    else ("31 - 40" if age <= 40
    else ("41 - 50" if age <= 50
    else ("51 - 60" if age <= 60
    else ("61 - 70" if age <= 70 else "> 70"))))))
```

In [9]:
```
df["age_group"] = df.apply(lambda x: age_range(x.age), axis=1)
```

In [10]:
```
df.age_group.value_counts()
```

Out[10]:
```
age_group
31 - 40    12838
21 - 30    12170
41 - 50    10403
51 - 60     6202
<= 20       3623
61 - 70     2738
> 70         868
Name: count, dtype: int64
```

In [11]:
```
del df["age"]
```

In [12]:
```
categorical = dai_anonymization.file_to_list("categorical.txt")
dai_anonymization.set_categorial(categorical)
```

In [13]:
```
dai_anonymization.set_sensitive_column("income")
```

In [14]:
```
feature_columns = dai_anonymization.file_to_list("features.txt")
dai_anonymization.set_feature_columns(feature_columns)
print(feature_columns)
```

```
['race', 'sex', 'age_group']
```

In [15]:
```
df = dai_anonymization.dataframe
dd = pd.Series({c: df[c].unique() for c in df})
print(dd)
```

```
workclass         [' State-gov', ' Self-emp-not-inc', ' Private'...
fnlwgt            [77516, 83311, 215646, 234721, 338409, 284582,...
education         [' Bachelors', ' HS-grad', ' 11th', ' Masters'...
education-num     [13, 9, 7, 14, 5, ..., 3, 6, 2, 1, 8]
Length: ...
marital-status    [' Never-married', ' Married-civ-spouse', ' Di...
occupation        [' Adm-clerical', ' Exec-managerial', ' Handle...
relationship      [' Not-in-family', ' Husband', ' Wife', ' Own-...
race              [' White', ' Black', ' Asian-Pac-Islander', ' ...
sex               [' Male', ' Female']
Categories (2, object): [...
capital-gain      [2174, 0, 14084, 5178, 5013, 2407, 14344, 1502...
capital-loss      [0, 2042, 1408, 1902, 1573, 1887, 1719, 1762, ...
hours-per-week    [40, 13, 16, 45, 50, 80, 30, 35, 60, 20, 52, 4...
native-country    [' United-States', ' Cuba', ' Jamaica', ' Indi...
income            [' <=50k', ' >50k']
Categories (2, object): ['...
age_group         ['31 - 40', '41 - 50', '51 - 60', '21 - 30', '...
dtype: object
```

In [16]:
```
dai_anonymization.partition_dataset()
print( len(dai_anonymization.finished_partitions) )
```

```
67
```

In [17]:
```
dai_anonymization.build_anonymized_dataset()
```

```
build_anonymized_dataset
```

In [18]:
```
dai_anonymization.result_df.head(10)
```

Out[18]:

	race	sex	age_group	income	count
0	White	Male	31 - 40	<=50k	5187
1	White	Male	31 - 40	>50k	2711
2	Black	Male	31 - 40	<=50k	504
3	Black	Male	31 - 40	>50k	127
4	Black	Female	21 - 30	<=50k	607
5	Black	Female	21 - 30	>50k	13
6	White	Female	21 - 30	<=50k	3550
7	White	Female	21 - 30	>50k	189
8	Asian-Pac-Islander	Male	21 - 30	<=50k	232
9	Asian-Pac-Islander	Male	21 - 30	>50k	38

In [19]:
```
dai_anonymization.result_df.describe()
```

Out[19]:

	count
count	111.000000
mean	439.882883
std	996.691561
min	3.000000
25%	16.500000
50%	55.000000
75%	244.000000
max	5752.000000

In [20]:
```
dai_anonymization.result_df.to_csv("results.csv")
```

In [ ]:
```

```

In [ ]:
```

```