

**National Institute of Technology
Warangal**

Lab 1 Assignment

Name: Ayush Rana

Roll Number: 24CSM2R05

Course Title: Data Privacy

**Department: Computer Science and
Engineering**

**Program: M. Tech in Computer Science
and**

Information Security

Semester: 2

1. Implement K-Anonymization for the Dataset given in below link. (Assume K =2,3,4 etc)

CODE:

```
[1]: import pandas as pd

# Define column names from the dataset documentation
columns = ["age", "workclass", "fnlwgt", "education", "education_num", "marital_status",
           "occupation", "relationship", "race", "sex", "capital_gain", "capital_loss",
           "hours_per_week", "native_country", "income"]

# Load the dataset into a Pandas DataFrame
df = pd.read_csv("adult.data", header=None, names=columns, na_values="?", skipinitialspace=True)

# Save it as a CSV file
df.to_csv("adult.csv", index=False)

print("CSV file saved successfully!")
```

CSV file saved successfully!

[2]: df

```
[2]:
```

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_co
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_co
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	
...
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-

32561 rows × 15 columns

◀ ▶

```
[3]: import pandas as pd
import numpy as np

# Define column names
columns = ["age", "workclass", "fnlwgt", "education", "education_num", "marital_status",
           "occupation", "relationship", "race", "sex", "capital_gain", "capital_loss",
           "hours_per_week", "native_country", "income"]

# Load dataset
df = pd.read_csv("adult.csv")

# Select Quasi-Identifiers (QIDs) for anonymization
quasi_identifiers = ["age", "workclass", "education", "marital_status", "race", "sex"]
```

```
[4]: # Create age groups for anonymization
bins = [0, 25, 40, 60, 100] # Age ranges
labels = ["0-25", "26-40", "41-60", "61-100"]
df["age"] = pd.cut(df["age"], bins=bins, labels=labels)
```

```
[5]: north_america = ["United-States", "Canada", "Mexico"]
south_america = ["Columbia", "Ecuador", "Peru", "Guatemala"]
asia = ["India", "China", "Japan", "Philippines", "Vietnam"]
europe = ["Germany", "England", "Italy", "France", "Greece"]

df["native_country"] = df["native_country"].apply(
    lambda x: "North America" if x in north_america else
              "South America" if x in south_america else
              "Asia" if x in asia else
              "Europe" if x in europe else "Other")
```

[1]:	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_1
0	26-40	Government	77516	Higher	13	Single	Adm-clerical	Not-in-family	White	Male	2174	0	40	North
1	41-60	Self-Employed	83311	Higher	13	Married	Exec-managerial	Husband	White	Male	0	0	13	North
2	26-40	Employed	215646	Middle	9	Separated	Handlers-cleaners	Not-in-family	White	Male	0	0	40	North
3	41-60	Employed	234721	Middle	7	Married	Handlers-cleaners	Husband	Black	Male	0	0	40	North

[11]: df

[11]:

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_
0	26-40	Government	77516	Higher	13	Single	Adm-clerical	Not-in-family	White	Male	2174	0	40	North.
1	41-60	Self-Employed	83311	Higher	13	Married	Exec-managerial	Husband	White	Male	0	0	13	North.
2	26-40	Employed	215646	Middle	9	Separated	Handlers-cleaners	Not-in-family	White	Male	0	0	40	North.
3	41-60	Employed	234721	Middle	7	Married	Handlers-cleaners	Husband	Black	Male	0	0	40	North.
4	26-40	Employed	338409	Higher	13	Married	Prof-specialty	Wife	Black	Female	0	0	40	
...
32556	26-40	Employed	257302	High	12	Married	Tech-support	Wife	White	Female	0	0	38	North.
32557	26-40	Employed	154374	Middle	9	Married	Machine-op-inspct	Husband	White	Male	0	0	40	North.
32558	41-60	Employed	151910	Middle	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	North.
32559	0-25	Employed	201490	Middle	9	Single	Adm-clerical	Own-child	White	Male	0	0	20	North.
32560	41-60	Self-Employed	287927	Middle	9	Married	Exec-managerial	Wife	White	Female	15024	0	40	North.

32561 rows × 15 columns

2. Implement the K-Anonmization technique using Full Domain Generalization mechanism.

Code:

```
[1]: import pandas as pd
import numpy as np

# Define column names
columns = ["age", "workclass", "fnlwgt", "education", "education_num", "marital_status",
           "occupation", "relationship", "race", "sex", "capital_gain", "capital_loss",
           "hours_per_week", "native_country", "income"]

# Load the dataset
df = pd.read_csv("adult.csv")

[2]: df

[2]:
```

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_co
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	
...
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-

```
[2]: df

[2]:
```

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_co
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	
...
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-

32561 rows × 15 columns

```
[3]: # Select Quasi-Identifiers (QIDs) for anonymization
quasi_identifiers = ["age", "workclass", "education", "marital_status", "occupation", "race", "sex", "native_country", "hours_per_week"]
bins = [0, 25, 40, 60, 100]
labels = ["0-25", "26-40", "41-60", "61-100"]
df["age"] = pd.cut(df["age"], bins=bins, labels=labels)
df["education"] = df["education"].replace({
    "Preschool": "Low",
    "1st-4th": "Low", "5th-6th": "Low", "7th-8th": "Low",
    "9th": "Middle", "10th": "Middle", "11th": "Middle", "12th": "Middle",
    "HS-grad": "Middle",
    "Some-college": "High", "Assoc-voc": "High", "Assoc-acdm": "High",
    "Bachelors": "Higher", "Masters": "Higher", "Doctorate": "Higher", "Prof-school": "Higher"
})
```

```
[4]: df["workclass"] = df["workclass"].replace({
    "Private": "Employed",
    "Self-emp-not-inc": "Self-Employed",
    "Self-emp-inc": "Self-Employed",
    "Federal-gov": "Government",
    "Local-gov": "Government",
    "State-gov": "Government",
    "Without-pay": "Unemployed",
    "Never-worked": "Unemployed"
})
```

```
[5]: df["marital_status"] = df["marital_status"].replace({
    "Married-civ-spouse": "Married",
    "Married-AF-spouse": "Married",
    "Divorced": "Separated",
    "Separated": "Separated",
    "Widowed": "Widowed",
    "Never-married": "Single"
})
```

```
[6]: north_america = ["United-States", "Canada", "Mexico"]
south_america = ["Columbia", "Ecuador", "Peru", "Guatemala"]
asia = ["India", "China", "Japan", "Philippines", "Vietnam"]
europe = ["Germany", "England", "Italy", "France", "Greece"]
df["native_country"] = df["native_country"].apply(
    lambda x: "North America" if x in north_america else
              "South America" if x in south_america else
              "Asia" if x in asia else
              "Europe" if x in europe else "Other")
bins = [0, 20, 40, 60, 100]
labels = ["0-20", "21-40", "41-60", "61-100"]
df["hours_per_week"] = pd.cut(df["hours_per_week"], bins=bins, labels=labels)
```

```
[7]: k = 5 # Minimum number of records per group
grouped = df.groupby(quasi_identifiers, observed=False) # Apply groupby with observed=False
df_anonymized = grouped.filter(lambda x: len(x) >= k)

print(f"Original dataset size: {len(df)}, After anonymization: {len(df_anonymized)}")

Original dataset size: 32561, After anonymization: 23006
```

```
[8]: df_anonymized.to_csv("adult_anonymized_fulldomain.csv", index=False)
print("K-Anonymized dataset saved as 'adult_anonymized_fulldomain.csv'")

K-Anonymized dataset saved as 'adult_anonymized_fulldomain.csv'
```

[10]: df_anonymized

🔍 ↑ ↓ 📄 📌 🗑️

[10]:

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	relationship	race	sex	capital_gain	capital_loss	hours_per_week	native_u
0	26-40	Government	77516	Higher	13	Single	Adm-clerical	Not-in-family	White	Male	2174	0	21-40	North
2	26-40	Employed	215646	Middle	9	Separated	Handlers-cleaners	Not-in-family	White	Male	0	0	21-40	North
3	41-60	Employed	234721	Middle	7	Married	Handlers-cleaners	Husband	Black	Male	0	0	21-40	North
5	26-40	Employed	284582	Higher	14	Married	Exec-managerial	Wife	White	Female	0	0	21-40	North
7	41-60	Self-Employed	209642	Middle	9	Married	Exec-managerial	Husband	White	Male	0	0	41-60	North
...
32555	0-25	Employed	310152	High	10	Single	Protective-serv	Not-in-family	White	Male	0	0	21-40	North
32556	26-40	Employed	257302	High	12	Married	Tech-support	Wife	White	Female	0	0	21-40	North
32557	26-40	Employed	154374	Middle	9	Married	Machine-op-inspct	Husband	White	Male	0	0	21-40	North
32558	41-60	Employed	151910	Middle	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	21-40	North
32559	0-25	Employed	201490	Middle	9	Single	Adm-clerical	Own-child	White	Male	0	0	0-20	North

23006 rows × 15 columns