# assignment-4-2

June 30, 2023

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

[2]:
```
features = ["Age", "Workclass", "fnlwgt", "Education", "Education-Num",␣
 ↪"Martial Status", "Occupation", "Relationship",
             "Race", "Sex", "Capital Gain", "Capital Loss", "Hours per week",␣
 ↪"Native-Country", "Target"]

df = pd.read_csv('adult.data', names=features)
df
```

[2]:

| | Age | Workclass | fnlwgt | Education | Education-Num | \ |
|---|---|---|---|---|---|---|
| 0 | 39 | State-gov | 77516 | Bachelors | 13 | |
| 1 | 50 | Self-emp-not-inc | 83311 | Bachelors | 13 | |
| 2 | 38 | Private | 215646 | HS-grad | 9 | |
| 3 | 53 | Private | 234721 | 11th | 7 | |
| 4 | 28 | Private | 338409 | Bachelors | 13 | |
| … | … | … | … | … | … | |
| 32556 | 27 | Private | 257302 | Assoc-acdm | 12 | |
| 32557 | 40 | Private | 154374 | HS-grad | 9 | |
| 32558 | 58 | Private | 151910 | HS-grad | 9 | |
| 32559 | 22 | Private | 201490 | HS-grad | 9 | |
| 32560 | 52 | Self-emp-inc | 287927 | HS-grad | 9 | |

| | Martial Status | Occupation | Relationship | Race | \ |
|---|---|---|---|---|---|
| 0 | Never-married | Adm-clerical | Not-in-family | White | |
| 1 | Married-civ-spouse | Exec-managerial | Husband | White | |
| 2 | Divorced | Handlers-cleaners | Not-in-family | White | |
| 3 | Married-civ-spouse | Handlers-cleaners | Husband | Black | |
| 4 | Married-civ-spouse | Prof-specialty | Wife | Black | |
| … | … | … | … | … | |
| 32556 | Married-civ-spouse | Tech-support | Wife | White | |
| 32557 | Married-civ-spouse | Machine-op-inspct | Husband | White | |
| 32558 | Widowed | Adm-clerical | Unmarried | White | |
| 32559 | Never-married | Adm-clerical | Own-child | White | |
| 32560 | Married-civ-spouse | Exec-managerial | Wife | White | |

```
            Sex  Capital Gain  Capital Loss  Hours per week  Native-Country  \
0          Male          2174             0              40   United-States
1          Male             0             0              13   United-States
2          Male             0             0              40   United-States
3          Male             0             0              40   United-States
4        Female             0             0              40            Cuba
…          …           …           …              …              …
32556    Female             0             0              38   United-States
32557      Male             0             0              40   United-States
32558    Female             0             0              40   United-States
32559      Male             0             0              20   United-States
32560    Female         15024             0              40   United-States

       Target
0       <=50K
1       <=50K
2       <=50K
3       <=50K
4       <=50K
…       …
32556   <=50K
32557    >50K
32558   <=50K
32559   <=50K
32560    >50K

[32561 rows x 15 columns]
```

https://rstudio-pubs-static.s3.amazonaws.com/538563_85cb2b4cd06b4dc48d33de73fa97a297.html

https://archive.ics.uci.edu/dataset/2/adult

### 0.0.1 Question: Do data analysis using Pandas and answer following questions?

1. How many men and women (sex feature) are represented in this dataset?
2. What is the average age (age feature) of women?
3. What is the proportion of German citizens (native-country feature)?

4-5. What are mean value and standard deviation of the age of those who recieve more than 50K per year (salary feature) and those who receive less than 50K per year?

6. Is it true that people who receive more than 50k have at least high school education? (education - Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters or Doctorate feature)

[3]:
```
# 1.How many men and women (sex feature) are represented in this dataset?
print(df['Sex'].unique())
sex_counts = df['Sex'].value_counts()
print(sex_counts)
```

```
[' Male' ' Female']
 Male      21790
 Female    10771
Name: Sex, dtype: int64
```

[4]:
```
# 2.What is the average age (age feature) of women?

average_age_women = df.loc[df['Sex'] == 'Female', 'Age'].mean()
print("Average age of women:", average_age_women)
```

```
Average age of women: nan
```

[9]:
```
# 3.What is the proportion of German citizens (native-country feature)?

german_citizens_prop_1 = (df['Native-Country'] == 'Germany').mean() * 100
print("Proportion of German citizens:", german_citizens_prop_1)
```

```
Proportion of German citizens: 0.0
```

[10]:
```
# 4-5. What are mean value and standard deviation of the age of those who␣
 ↪recieve more than 50K per year (salary feature) and those who receive less␣
 ↪than 50K per year?
high_salary_age_mean = df[df['Target'] == '>50K']['Age'].mean()
high_salary_age_std = df[df['Target'] == '>50K']['Age'].std()
print("Mean age of those who receive >50K:", high_salary_age_mean)
print("Standard deviation of age of those who receive >50K:",␣
 ↪high_salary_age_std)
low_salary_age_mean = df[df['Target'] == '>50K']['Age'].mean()
low_salary_age_std = df[df['Target'] == '>50K']['Age'].std()
print("Mean age of those who receive <50K:", low_salary_age_mean)
print("Standard deviation of age of those who receive <50K:",␣
 ↪low_salary_age_std)
```

```
Mean age of those who receive >50K: nan
Standard deviation of age of those who receive >50K: nan
Mean age of those who receive <50K: nan
Standard deviation of age of those who receive <50K: nan
```

[11]:
```
# 6. Is it true that people who receive more than 50k have at least high school␣
 ↪education? (education - Bachelors, Prof-school, Assoc-acdm, Assoc-voc,␣
 ↪Masters or Doctorate feature)

high_salary_education = ['Bachelors', 'Prof-school', 'Assoc-acdm', 'Assoc-voc',␣
 ↪'Masters', 'Doctorate']
high_salary_education_check = df[df['Target'] == '>50K']['Education'].
 ↪isin(high_salary_education).all()
```

```
print("People who receive >50K have at least a high school education:",␣
  ↪high_salary_education_check)
```

```
People who receive >50K have at least a high school education: True
```