In [4]:

```
import pandas as pd
import numpy as np
```

In [6]:

```
df = pd.read_csv("adult.data.csv")
```

In [8]:

```
df.head()
```

Out[8]:

	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	217
0	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	
1	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	
2	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	
3	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	
4	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	

In [10]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32560 entries, 0 to 32559
Data columns (total 15 columns):
 #   Column          Non-Null Count  Dtype 
---  ------          --------------  ----- 
 0   39              32560 non-null  int64 
 1    State-gov      32560 non-null  object
 2    77516          32560 non-null  int64 
 3    Bachelors      32560 non-null  object
 4    13             32560 non-null  int64 
 5    Never-married  32560 non-null  object
 6    Adm-clerical   32560 non-null  object
 7    Not-in-family  32560 non-null  object
 8    White          32560 non-null  object
 9    Male           32560 non-null  object
 10   2174           32560 non-null  int64 
 11   0              32560 non-null  int64 
 12   40             32560 non-null  int64 
 13   United-States  32560 non-null  object
 14   <=50K          32560 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7+ MB
```

In [12]:

```
df.shape
```

Out[12]:

```
(32560, 15)
```

In [14]:

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

df = pd.read_csv(
    "adult.data.csv",
    names=columns,
    na_values="?",
    skipinitialspace=True
)
```

In [16]:

```
df.head()
```

Out[16]:

	age	workclass	fnlwgt	education	education_num	marital_status	occupation	rela
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	

In [28]:

```
df_small = df[["age", "hours_per_week", "income"]]
df_small.head()
```

Out[28]:

	age	hours_per_week	income
0	39	40	<=50K
1	50	13	<=50K
2	38	40	<=50K
3	53	40	<=50K
4	28	40	<=50K

In [30]:

```
df_small = df_small.dropna()
```

In [32]:

```
age_stats = df_small.groupby("income")["age"].agg(
    Mean="mean",
    Median="median",
    Minimum="min",
    Maximum="max",
    Standard_Deviation="std"
)

age_stats
```

Out[32]:

	Mean	Median	Minimum	Maximum	Standard_Deviation
income					
<=50K	36.783738	34.0	17	90	14.020088
>50K	44.249841	44.0	19	90	10.519028

In [34]:

```
hours_stats = df_small.groupby("income")["hours_per_week"].agg(
    Mean="mean",
    Median="median",
    Minimum="min",
    Maximum="max",
    Standard_Deviation="std"
)

hours_stats
```

Out[34]:

	Mean	Median	Minimum	Maximum	Standard_Deviation
income					
<=50K	38.840210	40.0	1	99	12.318995
>50K	45.473026	40.0	1	99	11.012971

In [36]:

```
age_list = df_small.groupby("income")["age"].apply(list)
age_list
```

Out[36]:

```
income
<=50K    [39, 50, 38, 53, 28, 37, 49, 23, 32, 34, 25, 3...
>50K     [52, 31, 42, 37, 30, 40, 43, 40, 56, 54, 31, 5...
Name: age, dtype: object
```

In [38]:

```
df_small.shape
```

Out[38]:

```
(32561, 3)
```

In [40]:

```
hours_list = df_small.groupby("income")["hours_per_week"].apply(list)
hours_list
```

Out[40]:

```
income
<=50K    [40, 13, 40, 40, 40, 40, 16, 30, 50, 45, 35, 4...
>50K     [45, 50, 40, 80, 40, 40, 45, 60, 40, 60, 38, 4...
Name: hours_per_week, dtype: object
```

In [ ]:

```

```