In [ ]:
```
import pandas as pd
```

In [ ]:
```
data = pd.read_csv('G:\study\documents_study\materials\\adult.csv')
```

In [ ]:
```
print(data)
```

```
       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

           marital-status         occupation   relationship   race     sex  \
0           Never-married       Adm-clerical  Not-in-family  White    Male
1      Married-civ-spouse    Exec-managerial        Husband  White    Male
2                Divorced  Handlers-cleaners  Not-in-family  White    Male
3      Married-civ-spouse  Handlers-cleaners        Husband  Black    Male
4      Married-civ-spouse     Prof-specialty           Wife  Black  Female
...                   ...                ...            ...    ...     ...
32556  Married-civ-spouse       Tech-support           Wife  White  Female
32557  Married-civ-spouse  Machine-op-inspct        Husband  White    Male
32558             Widowed       Adm-clerical      Unmarried  White  Female
32559       Never-married       Adm-clerical      Own-child  White    Male
32560  Married-civ-spouse    Exec-managerial           Wife  White  Female

       capitalGain  capitalLoss  hoursPerWeek  nativeCountry income
0             2174            0            40  United-States  <=50K
1                0            0            13  United-States  <=50K
2                0            0            40  United-States  <=50K
3                0            0            40  United-States  <=50K
4                0            0            40           Cuba  <=50K
...            ...          ...           ...            ...    ...
32556            0            0            38  United-States  <=50K
32557            0            0            40  United-States   >50K
32558            0            0            40  United-States  <=50K
32559            0            0            20  United-States  <=50K
32560        15024            0            40  United-States   >50K

[32561 rows x 15 columns]
```

1. Сколько мужчин и женщин (признак sex) представлено в этом датасете Adult

In [ ]:
```
women = data[data['sex'] == 'Female']
men = data[data['sex'] == 'Male']

print('Female', len(women))
print('Male', len(men))
```

```
Female 10771
Male 21790
```

2. Каков средний возраст (признак age) женщин?

In [ ]:
```
print(women['age'].mean())
```

```
36.85823043357163
```

3. Какова доля граждан Германии (признак native-country)?

In [ ]:
```
print(data['nativeCountry'].value_counts(normalize=True)['Germany'])
```

```
0.004284195384326724
```

4-5. Каковы средние значения и среднеквадратичные отклонения возраста тех, кто получает более 50K в год и тех, кто получает менее 50K в год?

In [ ]:
```
print(data[data['income'] == '>50K']['age'].describe())
print(data[data['income'] == '<=50K']['age'].describe())
```

```
count    7841.000000
mean       44.249841
std        10.519028
min        19.000000
25%        36.000000
50%        44.000000
75%        51.000000
max        90.000000
Name: age, dtype: float64
count    24720.000000
mean        36.783738
std         14.020088
min         17.000000
25%         25.000000
50%         34.000000
75%         46.000000
max         90.000000
Name: age, dtype: float64
```

In [ ]:
```
pd.pivot_table(data,columns='income',values='age',aggfunc=['mean','std'])
```

Out[ ]:

	mean		std	
income	**<=50K**	**>50K**	**<=50K**	**>50K**
age	36.783738	44.249841	14.020088	10.519028

6. Правда ли, что люди, которые получают больше 50k, имеют как минимум высшее образование?

In [ ]:
```
tab = pd.crosstab(data['income'], data['education'])
print(tab.loc['>50K'])
print((tab.loc['>50K', 'Assoc-acdm'] + tab.loc['>50K', 'Assoc-voc'] + tab.loc['>50K
      + tab.loc['>50K', 'Doctorate'] + tab.loc['>50K', 'Masters'] + tab.loc['>50K',
```

```
education
10th              62
11th              60
12th              33
1st-4th            6
5th-6th           16
7th-8th           40
9th               27
Assoc-acdm       265
Assoc-voc        361
Bachelors       2221
Doctorate        306
HS-grad         1675
Masters          959
Preschool          0
Prof-school      423
Some-college    1387
Name: >50K, dtype: int64
False
```

7. Выведите статистику возраста для каждой расы (признак race) и каждого пола. Используйте groupby и describe.

In [ ]:
```
adult_groups = data.groupby(['race', 'sex'])['age'].describe()
print(adult_groups)

#

print('Men - Amer-Indian-Eskimo:')
adult_groups['max']['Amer-Indian-Eskimo','Male']
```

```
                            count       mean        std   min   25%   50%  \
race               sex
Amer-Indian-Eskimo Female    119.0  37.117647  13.114991  17.0  27.0  36.0
                   Male      192.0  37.208333  12.049563  17.0  28.0  35.0
Asian-Pac-Islander Female    346.0  35.089595  12.300845  17.0  25.0  33.0
                   Male      693.0  39.073593  12.883944  18.0  29.0  37.0
Black              Female   1555.0  37.854019  12.637197  17.0  28.0  37.0
                   Male     1569.0  37.682600  12.882612  17.0  27.0  36.0
Other              Female    109.0  31.678899  11.631599  17.0  23.0  29.0
                   Male      162.0  34.654321  11.355531  17.0  26.0  32.0
White              Female   8642.0  36.811618  14.329093  17.0  25.0  35.0
                   Male    19174.0  39.652498  13.436029  17.0  29.0  38.0

                            75%   max
race               sex
Amer-Indian-Eskimo Female  46.00  80.0
                   Male    45.00  82.0
Asian-Pac-Islander Female  43.75  75.0
                   Male    46.00  90.0
Black              Female  46.00  90.0
                   Male    46.00  90.0
Other              Female  39.00  74.0
                   Male    42.00  77.0
White              Female  46.00  90.0
                   Male    49.00  90.0
Men - Amer-Indian-Eskimo:
```

Out[ ]:
```
82.0
```

8. Среди кого больше доля зарабатывающих много (>50K): среди женатых или холостых мужчин (признак marital-status)?

In [ ]:
```
men1 = men[men['income'] == '>50K']
tabl = pd.crosstab(men1['income'], men1['marital-status'], normalize=True)
print((tabl.loc['>50K', 'Never-married'] + tabl.loc['>50K', 'Married-civ-spouse'] +
```

```
True
```

9. Какое максимальное число часов человек работает в неделю (признак hours-per-week)?

In [ ]:
```
print(data['hoursPerWeek'].describe()['max'])
h = data[data['hoursPerWeek'] == 99]
print(len(h))
print(h['income'].value_counts(normalize=True).loc['>50K'])
```

```
99.0
85
0.29411764705882354
```

10. Посчитайте среднее время работы (hours-per-week) зарабатывающих мало и много для каждой страны

In [ ]:
```
table = data.groupby(['nativeCountry', 'income'])['hoursPerWeek'].std()
```

In [ ]:
```
data.pivot_table( ["hoursPerWeek"], ['nativeCountry'], ["income"], aggfunc="mean",
```

Out[ ]:

	hoursPerWeek	
income	<=50K	>50K
nativeCountry		
Cambodia	41.416667	40.000000
Canada	37.914634	45.641026
China	37.381818	38.900000
Columbia	38.684211	50.000000
Cuba	37.985714	42.440000

In [ ]:
```
def work():
    dm = data.groupby(['nativeCountry', 'income'])['hoursPerWeek'].mean()
    ds = data.groupby(['nativeCountry', 'income'])['hoursPerWeek'].std()

    d = pd.concat([dm,ds],axis=1)
    d.columns = ['mean','std']
    print(d)

work()
```

nativeCountry	income	mean	std
Cambodia	<=50K	41.416667	3.088346
	>50K	40.000000	0.000000
Canada	<=50K	37.914634	13.012056
	>50K	45.641026	12.066673
China	<=50K	37.381818	11.439844
...		...	...
United-States	>50K	45.505369	11.025092
Vietnam	<=50K	37.193548	12.422664
	>50K	39.200000	1.788854
Yugoslavia	<=50K	41.600000	11.305849
	>50K	49.500000	11.202678

[80 rows x 2 columns]