Урок 1. Домашнее задание.

Автор материала: Юрий Кашницкий, программист-исследователь Mail.Ru Group

Выполнил: Сергей Цветков группа **PY_gr2**

In [53]:
```
import pandas as pd
import numpy as np
import math
data = pd.read_csv('adult.data.csv')
data.head(5)
```

Out[53]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

1.Сколько мужчин и женщин (признак sex) представлено в этом наборе данных?

1. Каков средний возраст (признак age) женщин?
2. Какова доля граждан Германии (признак native-country)? 4-5. Каковы средние значения и среднеквадратичные отклонения возраста тех, кто получа ет более 50K в год (признак salary) и тех, кто получает менее 50K в год?
3. Правда ли, что люди, которые получают больше 50k, имеют как минимум высшее образование? (признак education – Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters или Doctorate)
4. Выведите статистику возраста для каждой расы (признак race) и каждого пола. Используйте groupby и describe. Найдите таким образом максимальный возраст мужчин расы Amer-Indian-Eskimo.
5. Среди кого больше доля зарабатывающих много (>50K): среди женатых или холостых мужчин (признак marital-status)? Женатыми считаем тех, у кого marital-status начинается с Married (Married-civ-spouse, Married-spouse-absent или Married-AF-spouse), остальных считаем холостыми.
6. Какое максимальное число часов человек работает в неделю (признак hours-per-week)? Сколько людей работают такое количество часов и каков среди них процент зарабатывающих много?
7. Посчитайте среднее время работы (hours-per-week) зарабатывающих мало и много (salary) для каждой страны (native-country).

1. Сколько мужчин и женщин (признак sex) представлено в этом наборе данных?

Предварительный анализ DataFrame

In [3]:
```
Data_Shape = data.shape # Получаем количество строк и столбцов в DataFrame
print("Строк:",Data_Shape[0], "Столбцов:", Data_Shape[1])
```

```
Строк: 32561 Столбцов: 15
```

In [5]:
```
All_group = data.groupby('sex') # Груперовка по полю sex (пол)
Male = All_group.get_group('Male') # Обращаемся к конкретной группе
Female = All_group.get_group('Female') # Обращаемся к конкретной группе
# Выводим на печать количество мужчин и женщин
print ("Количество мужчин:", len(Male))
print ("Количество женщин:", len(Female))
```

```
Количество мужчин: 21790
Количество женщин: 10771
```

2. Каков средний возраст (признак age) женщин?

In [21]:
```
mean_age = data[(data['sex'] == 'Female')]['age'].mean() # Получайм средний возраст
print("Средний возраст женщин:", round(mean_age, 2)) # Выводим средний возраст, округленный до 2-х символов после запя
той
```

```
Средний возраст женщин: 36.86
```

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

In [15]:
```
germany_count = data[data['native-country'] == 'Germany']['native-country'].count() # Подсчет количества записей
germany_percentage = germany_count / Data_Shape[0]
print("Количество граждан Германии:",germany_count,", что составляет", str(round(germany_percentage * 100, 2)) + "%")
```

```
Количество граждан Германии: 137 , что составляет 0.42%
```

4-5. Каковы средние значения и среднеквадратичные отклонения возраста тех, кто получа ет более 50K в год (признак salary) и тех, кто получает менее 50K в год?

In [26]:
```
rich_mean = data[data['salary'] == '>50K']['age'].mean()
poor_mean = data[data['salary'] == '<=50K']['age'].mean()

rich_std = data[data['salary'] == '>50K']['age'].std()
poor_std = data[data['salary'] == '<=50K']['age'].std()

print('Средний возраст получающих более 50K:\t', round(rich_mean, 2))
print('Среднеквадратичное отклонение:\t\t', round(rich_std, 2))
print('Средний возраст получающих менее 50K:\t', round(poor_mean, 2))
print('Среднеквадратичное отклонение:\t\t', round(poor_std, 2))
```

```
Средний возраст получающих более 50K:    44.25
Среднеквадратичное отклонение:           10.52
Средний возраст получающих менее 50K:    36.78
Среднеквадратичное отклонение:           14.02
```

4. Правда ли, что люди, которые получают больше 50k, имеют как минимум высшее образование? (признак education – Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters или Doctorate)

In [123]:
```
# Задаем список признаков высшего образования
higher_education_sign = ['Bachelors', 'Prof-school', 'Assoc-acdm', 'Masters', 'Doctorate']

# Получаем количество ВСЕХ людей, имеющих высшее образование
higher_edu = data[data['salary'] == '>50K']['education'].count()

# Получаем количество людей, имеющих высшее образование и получающие больше 50K
higher_edu_rich = data[data['education'].isin(higher_education_sign) & (data['salary'] == '>50K')]['education'].count
()

print("Общее количество получающих > 50K:\t", higher_edu)
print("Из них имеют высшее образование:\t",higher_edu_rich)
print("Из них НЕ имеют высшего образования:\t",higher_edu - higher_edu_rich)
print("*"*50 )
if higher_edu_rich > higher_edu - higher_edu_rich:
#     print ("Утверждение ВЕРНО -", str(round((higher_edu_rich * 100)/ higher_edu, 2))+"%", "имеют высшее образовани
e")
    print ("Утверждение ВЕРНО -", str(round((higher_edu_rich * 100)/ higher_edu, 2))+"%", "имеют высшее образование")
else:
    print ("Утверждение НЕ ВЕРНО")
```

```
Общее количество получающих > 50K:      7841
Из них имеют высшее образование:        4174
Из них НЕ имеют высшего образования:    3667
**************************************************
Утверждение ВЕРНО - 53.23% имеют высшее образование
```

5. Выведите статистику возраста для каждой расы (признак race) и каждого пола. Используйте groupby и describe. Найдите таким образом максимальный возраст мужчин расы Amer-Indian-Eskimo.

In [6]:
```
# Группировка по полю Раса
race_stat = data.groupby('race')
# Получаем уникальные значения по полю Расса - необходимо для перебора
u_race = data['race'].unique()

print("Присутствие расс в выборке")

for one_race in u_race:
    race_group = race_stat.get_group(one_race)
    print(one_race, " - " , race_group['race'].count())
print("*"*50 )
print("Средний возраст по рассам")

for age_race in u_race:
    race_group = race_stat.get_group(age_race)
    mean_age = data[data['race'] == age_race]['age'].mean()
    print (age_race, " - " ,round(mean_age, 2))
print("*"*50 )
print("Соотношение полов и среднего возраста по полам в рассах")

for sex_race in u_race:
    race_group = race_stat.get_group(sex_race)
    rase_male = data[(data['race'] == sex_race) & (data['sex'] == 'Male')]['sex'].count()
    rase_female = data[(data['race'] == sex_race) & (data['sex'] == 'Female')]['sex'].count()

    age_male = data[(data['race'] == sex_race) & (data['sex'] == 'Male')]['age'].mean()
    age_female = data[(data['race'] == sex_race) & (data['sex'] == 'Female')]['age'].mean()

    print("{} - Мужчины:{} ({}) / Женщины:{} ({})".format(sex_race,rase_male,round(age_male, 2),rase_female,round(age_
female, 2)))
print("*"*50 )

max_eskimo = data[(data['race'] == 'Amer-Indian-Eskimo') & (data['sex'] == 'Male')]['age'].max()
print("Максимальный возраст мужчин расы Amer-Indian-Eskimo:", max_eskimo)
```

```
Присутствие расс в выборке
White  -  27816
Black  -  3124
Asian-Pac-Islander  -  1039
Amer-Indian-Eskimo  -  311
Other  -  271
**************************************************
Средний возраст по рассам
White  -  38.77
Black  -  37.77
Asian-Pac-Islander  -  37.75
Amer-Indian-Eskimo  -  37.17
Other  -  33.46
**************************************************
Соотношение полов и среднего возраста по полам в рассах
White - Мужчины:19174 (39.65) / Женщины:8642 (36.81)
Black - Мужчины:1569 (37.68) / Женщины:1555 (37.85)
Asian-Pac-Islander - Мужчины:693 (39.07) / Женщины:346 (35.09)
Amer-Indian-Eskimo - Мужчины:192 (37.21) / Женщины:119 (37.12)
Other - Мужчины:162 (34.65) / Женщины:109 (31.68)
**************************************************
Максимальный возраст мужчин расы Amer-Indian-Eskimo: 82
```

6. Среди кого больше доля зарабатывающих много (>50K): среди женатых или холостых мужчин (признак marital-status)? Женатыми считаем тех, у кого marital-status начинается с Married (Married-civ-spouse, Married-spouse-absent или Married-AF-spouse), остальных считаем холостыми.

In [67]:
```
# Получаем уникальные значения по полю семейное положение
m_plus = [] # Коллекция значений, соответствующих женатым
m_status = data['marital-status'].unique()
# Заполняем коллекцию
for k in m_status:
    if k[:4] == 'Marr':
        m_plus.append(k)
# Считаем всех богательников
all_rich = data[data['salary'] == '>50K']['sex'].count()
# Считаем богатеньких + женатых
married_rich = data[(data['salary'] == '>50K') & (data['marital-status'].isin(m_plus))]['sex'].count()

print("Доля зарабатывающих много среди женатых: {:.2%}".format((married_rich)/all_rich))
print("Доля зарабатывающих много среди холостых: {:.2%}".format((all_rich-married_rich)/all_rich))
```

```
Доля зарабатывающих много среди женатых: 85.91%
Доля зарабатывающих много среди холостых: 14.09%
```

7. Какое максимальное число часов человек работает в неделю (признак hours-per-week)? Сколько людей работают такое количество часов и каков среди них процент зарабатывающих много?

In [18]:
```
# Получаем максимальное количество рабочих часов
max_hours_per_week = data['hours-per-week'].max()
print("Максимальное число рабочих часов в неделю:", max_hours_per_week)
# Получаем количество людей, которые столько работают
work_top_counter = data[data['hours-per-week'] == max_hours_per_week]['hours-per-week'].count()
print(work_top_counter, 'человек работают столько часов в неделю')
# Получаем процент богатых людей из работающих много
work_top_counter_rich = data[(data['hours-per-week'] == max_hours_per_week) & (data['salary'] == '>50K')]['hours-per-w
eek'].count()
print("{:.2%} из этих людей зарабатывают больше 50K".format(work_top_counter_rich/work_top_counter))
```

```
Максимальное число рабочих часов в неделю: 99
85 человек работают столько часов в неделю
29.41% из этих людей зарабатывают больше 50K
```

8. Посчитайте среднее время работы (hours-per-week) зарабатывающих мало и много (salary) для каждой страны (native-country).

In [55]:
```
# Групируем людей по странам
country_worker = data.groupby('native-country')
# Получаем уникальные значения по полю родная страна - необходимо для перебора
countries = data['native-country'].unique()
for country in countries:
    if country != '?':
        rich = data[(data['native-country'] == country) & (data['salary'] == '>50K')]['hours-per-week'].mean()
        poor = data[(data['native-country'] == country) & (data['salary'] == '<=50K')]['hours-per-week'].mean()
        print(">>>"+country+"<<<")
        # Здесь проверяем - является ли выборка числом а не NaN
        if not math.isnan(rich):
            print("Богатые работают в среднем", round(rich),"часов")
        if not math.isnan(poor):
            print("Бедные работают в среднем", round(poor),"часов")
        print("*"*50 )
    else:
        continue
```

```
>>>United-States<<<
Богатые работают в среднем 46.0 часов
Бедные работают в среднем 39.0 часов
**************************************************
>>>Cuba<<<
Богатые работают в среднем 42.0 часов
Бедные работают в среднем 38.0 часов
**************************************************
>>>Jamaica<<<
Богатые работают в среднем 41.0 часов
Бедные работают в среднем 38.0 часов
**************************************************
>>>India<<<
Богатые работают в среднем 46.0 часов
Бедные работают в среднем 38.0 часов
**************************************************
>>>Mexico<<<
Богатые работают в среднем 47.0 часов
Бедные работают в среднем 40.0 часов
**************************************************
>>>South<<<
Богатые работают в среднем 51.0 часов
Бедные работают в среднем 40.0 часов
**************************************************
>>>Puerto-Rico<<<
Богатые работают в среднем 39.0 часов
Бедные работают в среднем 38.0 часов
**************************************************
>>>Honduras<<<
Богатые работают в среднем 60.0 часов
Бедные работают в среднем 34.0 часов
**************************************************
>>>England<<<
Богатые работают в среднем 45.0 часов
Бедные работают в среднем 40.0 часов
**************************************************
>>>Canada<<<
Богатые работают в среднем 46.0 часов
Бедные работают в среднем 38.0 часов
**************************************************
>>>Germany<<<
Богатые работают в среднем 45.0 часов
Бедные работают в среднем 39.0 часов
**************************************************
>>>Iran<<<
Богатые работают в среднем 48.0 часов
Бедные работают в среднем 41.0 часов
**************************************************
>>>Philippines<<<
Богатые работают в среднем 43.0 часов
Бедные работают в среднем 38.0 часов
**************************************************
>>>Italy<<<
Богатые работают в среднем 45.0 часов
Бедные работают в среднем 40.0 часов
**************************************************
>>>Poland<<<
Богатые работают в среднем 39.0 часов
Бедные работают в среднем 38.0 часов
**************************************************
>>>Columbia<<<
Богатые работают в среднем 50.0 часов
Бедные работают в среднем 39.0 часов
**************************************************
>>>Cambodia<<<
Богатые работают в среднем 40.0 часов
Бедные работают в среднем 41.0 часов
**************************************************
>>>Thailand<<<
Богатые работают в среднем 58.0 часов
Бедные работают в среднем 43.0 часов
**************************************************
>>>Ecuador<<<
Богатые работают в среднем 49.0 часов
Бедные работают в среднем 38.0 часов
**************************************************
>>>Laos<<<
Богатые работают в среднем 40.0 часов
Бедные работают в среднем 40.0 часов
**************************************************
>>>Taiwan<<<
Богатые работают в среднем 47.0 часов
Бедные работают в среднем 34.0 часов
**************************************************
>>>Haiti<<<
Богатые работают в среднем 43.0 часов
Бедные работают в среднем 36.0 часов
**************************************************
```

```
>>>Portugal<<<
Богатые работают в среднем 42.0 часов
Бедные работают в среднем 42.0 часов
**************************************************
>>>Dominican-Republic<<<
Богатые работают в среднем 47.0 часов
Бедные работают в среднем 42.0 часов
**************************************************
>>>El-Salvador<<<
Богатые работают в среднем 45.0 часов
Бедные работают в среднем 36.0 часов
**************************************************
>>>France<<<
Богатые работают в среднем 51.0 часов
Бедные работают в среднем 41.0 часов
**************************************************
>>>Guatemala<<<
Богатые работают в среднем 37.0 часов
Бедные работают в среднем 39.0 часов
**************************************************
>>>China<<<
Богатые работают в среднем 39.0 часов
Бедные работают в среднем 37.0 часов
**************************************************
>>>Japan<<<
Богатые работают в среднем 48.0 часов
Бедные работают в среднем 41.0 часов
**************************************************
>>>Yugoslavia<<<
Богатые работают в среднем 50.0 часов
Бедные работают в среднем 42.0 часов
**************************************************
>>>Peru<<<
Богатые работают в среднем 40.0 часов
Бедные работают в среднем 35.0 часов
**************************************************
>>>Outlying-US(Guam-USVI-etc)<<<
Бедные работают в среднем 42.0 часов
**************************************************
>>>Scotland<<<
Богатые работают в среднем 47.0 часов
Бедные работают в среднем 39.0 часов
**************************************************
>>>Trinadad&Tobago<<<
Богатые работают в среднем 40.0 часов
Бедные работают в среднем 37.0 часов
**************************************************
>>>Greece<<<
Богатые работают в среднем 51.0 часов
Бедные работают в среднем 42.0 часов
**************************************************
>>>Nicaragua<<<
Богатые работают в среднем 38.0 часов
Бедные работают в среднем 36.0 часов
**************************************************
>>>Vietnam<<<
Богатые работают в среднем 39.0 часов
Бедные работают в среднем 37.0 часов
**************************************************
>>>Hong<<<
Богатые работают в среднем 45.0 часов
Бедные работают в среднем 39.0 часов
**************************************************
>>>Ireland<<<
Богатые работают в среднем 48.0 часов
Бедные работают в среднем 41.0 часов
**************************************************
>>>Hungary<<<
Богатые работают в среднем 50.0 часов
Бедные работают в среднем 31.0 часов
**************************************************
>>>Holand-Netherlands<<<
Бедные работают в среднем 40.0 часов
**************************************************
```