Министерство образования и науки Российской Федерации

МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ им. Н.Э. БАУМАНА

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления» (ИУ5)

ДИСЦИПЛИНА: «Технологии машинного обучения»

Отчет по лабораторной работе №2
«Изучение библиотек обработки данных»

Выполнил:

Студент группы ИУ5-61Б

Кочетков М.Д.

Преподаватель:

Гапанюк Ю.Е.

Москва, 2020 г.

Цель лабораторной работы: изучение библиотеки обработки данных Pandas.

Задание:

Выполните первое демонстрационное задание "demo assignment" под названием "Exploratory data analysis with Pandas" со страницы курса https://mlcourse.ai/assignments

Выполнение ЛР:

In [1]:
```
import numpy as np
import pandas as pd
pd.set_option('display.max.columns', 100)
# to draw pictures in jupyter notebook
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
# we don't like warnings
# you can comment the following 2 lines if you'd like to
import warnings
warnings.filterwarnings('ignore')
```

In [2]:
```
data = pd.read_csv('adult.data.csv')
data.head()
```

Out[2]:

	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. How many men and women (sex feature) are represented in this dataset?

In [3]:
```
data['sex'].value_counts()
```

Out[3]:
```
Male      21790
Female    10771
Name: sex, dtype: int64
```

2. What is the average age (age feature) of women?

In [4]:
```
(data[data['sex'] == 'Female']['age']).mean()
```

Out[4]:
```
36.85823043357163
```

3. What is the percentage of German citizens (native-country feature)?

In [5]:
```
(data['native-country'] == 'Germany').sum()/data.shape[0]
```

Out[5]:
```
0.004207487485028101
```

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

In [6]:
```
salary1 = data[data['salary'] == '>50K']['age']
salary2 = data[data['salary'] == '<=50K']['age']
mean1 = salary1.mean()
mean2 = salary2.mean()
stdev1 = salary1.std()
stdev2 = salary2.std()
print(f'''Средний возраст людей с зарплатой более 50к: {mean1}''')
print(f'''Стандартное отклонение возраста людей с зарплатой более 50к: {stdev1}''')
print(f'''Средний возраст людей с зарплатой равной или менее 50к: {mean2}''')
print(f'''Стандартное отклонение возраста с зарплатой равной или менее 50к: {stdev2}''')
```

```
Средний возраст людей с зарплатой более 50к: 44.24984058155847
Стандартное отклонение возраста людей с зарплатой более 50к: 10.519027719851826
Средний возраст людей с зарплатой равной или менее 50к: 36.78373786407767
Стандартное отклонение возраста с зарплатой равной или менее 50к: 14.02008849082488
```

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

In [7]:
```
list((data[data['salary'] == '>50K']['education']).unique())
#ответ - нет, т.к. в списке присутствуют и другие школы
```

Out[7]:
```
['HS-grad',
 'Masters',
 'Bachelors',
 'Some-college',
 'Assoc-voc',
 'Doctorate',
 'Prof-school',
 'Assoc-acdm',
 '7th-8th',
 '12th',
 '10th',
 '11th',
 '9th',
 '5th-6th',
 '1st-4th']
```

7. Display age statistics for each race (race feature) and each gender (sex feature). Use groupby() and describe(). Find the maximum age of men of Amer-Indian-Eskimo race.

In [8]:
```
data['race'].unique()
for (race, sex), new_data in data.groupby(['race', 'sex']):
    print(f'''Race: {race}, sex: {sex}''')
    print(new_data['age'].describe())
    print('')
```

```
Race: Amer-Indian-Eskimo, sex: Female
count    119.000000
mean      37.117647
std       13.114991
min       17.000000
25%       27.000000
50%       36.000000
75%       46.000000
max       80.000000
Name: age, dtype: float64

Race: Amer-Indian-Eskimo, sex: Male
count    192.000000
mean      37.208333
std       12.049563
min       17.000000
25%       28.000000
50%       35.000000
75%       45.000000
max       82.000000
Name: age, dtype: float64

Race: Asian-Pac-Islander, sex: Female
count    346.000000
mean      35.089595
std       12.300845
min       17.000000
25%       25.000000
50%       33.000000
75%       43.750000
max       75.000000
Name: age, dtype: float64

Race: Asian-Pac-Islander, sex: Male
count    693.000000
mean      39.073593
std       12.883944
min       18.000000
25%       29.000000
50%       37.000000
75%       46.000000
max       90.000000
Name: age, dtype: float64

Race: Black, sex: Female
count    1555.000000
mean       37.854019
std        12.637197
min        17.000000
25%        28.000000
50%        37.000000
75%        46.000000
max        90.000000
Name: age, dtype: float64

Race: Black, sex: Male
count    1569.000000
mean       37.682600
std        12.882612
min        17.000000
25%        27.000000
50%        36.000000
75%        46.000000
max        90.000000
Name: age, dtype: float64

Race: Other, sex: Female
count    109.000000
mean      31.678899
std       11.631599
min       17.000000
25%       23.000000
50%       29.000000
75%       39.000000
max       74.000000
Name: age, dtype: float64

Race: Other, sex: Male
count    162.000000
mean      34.654321
std       11.355531
min       17.000000
25%       26.000000
50%       32.000000
75%       42.000000
max       77.000000
Name: age, dtype: float64

Race: White, sex: Female
count    8642.000000
mean       36.811618
std        14.329093
min        17.000000
25%        25.000000
50%        35.000000
75%        46.000000
max        90.000000
Name: age, dtype: float64

Race: White, sex: Male
count    19174.000000
mean        39.652498
std         13.436029
min         17.000000
25%         29.000000
50%         38.000000
75%         49.000000
max         90.000000
Name: age, dtype: float64
```

8. Among whom is the proportion of those who earn a lot (>50K) greater: married or single men (marital-status feature)? Consider as married those who have a marital-status starting with Married (Married-civ-spouse, Married-spouse-absent or Married-AF-spouse), the rest are considered bachelors.

In [9]:
```
data[(data['sex'] == 'Male') & (data['marital-status'].str.startswith('Married'))]['salary'].value_counts()
```

Out[9]:
```
<=50K    7576
>50K     5965
Name: salary, dtype: int64
```

In [10]:
```
data['marital-status'].unique()
data[(data['sex'] == 'Male') & (data['marital-status'].isin(['Never-married','Divorced','Separated','Widowed']))]['salary'].valu
e_counts()
# Ответ: зарабатывают больше женатые мужчины :)
```

Out[10]:
```
<=50K    7552
>50K      697
Name: salary, dtype: int64
```

9. What is the maximum number of hours a person works per week (hours-per-week feature)? How many people work such a number of hours, and what is the percentage of those who earn a lot (>50K) among them?

In [11]:
```
#максимальное количество рабочих часов в неделю
max_hours = data['hours-per-week'].max()

#количество людей, работающих максимальное число часов в неделю
max_hours_workers = (data['hours-per-week'] == max_hours).sum()

#процент тех, кто зарабатывает больше 50к и работает много
pers_hard = data[(data['hours-per-week'] == max_hours) & (data['salary'] == '>50K')].shape[0] / max_hours_workers

print(f'''Максимальное количество рабочих часов в неделю: {max_hours}''')
print(f'''Число людей, работающих максимальное число часов в неделю: {max_hours_workers}''')
print(f'''Процент людей, работающих много и получающих много: {pers_hard}''')
```

```
Максимальное количество рабочих часов в неделю: 99
Число людей, работающих максимальное число часов в неделю: 85
Процент людей, работающих много и получающих много: 0.29411764705882354
```

10.Count the average time of work (hours-per-week) for those who earn a little and a lot (salary) for each country (native-country). What will these be for Japan?

In [13]:
```
for (country, salary), new_data in data.groupby(['native-country', "salary"]):
    print(f'''Страна: {country} Зарплата: {salary} Среднее время работы в неделю: {new_data['hours-per-week'].mean()}''')

# Японцы в среднем работают по 47 часов в неделю
```

```
Страна: ? Зарплата: <=50K Среднее время работы в неделю: 40.16475972540046
Страна: ? Зарплата: >50K Среднее время работы в неделю: 45.54794520547945
Страна: Cambodia Зарплата: <=50K Среднее время работы в неделю: 41.416666666666664
Страна: Cambodia Зарплата: >50K Среднее время работы в неделю: 40.0
Страна: Canada Зарплата: <=50K Среднее время работы в неделю: 37.91463414634146
Страна: Canada Зарплата: >50K Среднее время работы в неделю: 45.64102564102564
Страна: China Зарплата: <=50K Среднее время работы в неделю: 37.38181818181818
Страна: China Зарплата: >50K Среднее время работы в неделю: 38.9
Страна: Columbia Зарплата: <=50K Среднее время работы в неделю: 38.68421052631579
Страна: Columbia Зарплата: >50K Среднее время работы в неделю: 50.0
Страна: Cuba Зарплата: <=50K Среднее время работы в неделю: 37.98571428571429
Страна: Cuba Зарплата: >50K Среднее время работы в неделю: 42.44
Страна: Dominican-Republic Зарплата: <=50K Среднее время работы в неделю: 42.338235294117645
Страна: Dominican-Republic Зарплата: >50K Среднее время работы в неделю: 47.0
Страна: Ecuador Зарплата: <=50K Среднее время работы в неделю: 38.041666666666664
Страна: Ecuador Зарплата: >50K Среднее время работы в неделю: 48.75
Страна: El-Salvador Зарплата: <=50K Среднее время работы в неделю: 36.03092783505155
Страна: El-Salvador Зарплата: >50K Среднее время работы в неделю: 45.0
Страна: England Зарплата: <=50K Среднее время работы в неделю: 40.483333333333334
Страна: England Зарплата: >50K Среднее время работы в неделю: 44.53333333333333
Страна: France Зарплата: <=50K Среднее время работы в неделю: 41.05882352941177
Страна: France Зарплата: >50K Среднее время работы в неделю: 50.75
Страна: Germany Зарплата: <=50K Среднее время работы в неделю: 39.13978494623656
Страна: Germany Зарплата: >50K Среднее время работы в неделю: 44.97727272727273
Страна: Greece Зарплата: <=50K Среднее время работы в неделю: 41.80952380952381
Страна: Greece Зарплата: >50K Среднее время работы в неделю: 50.625
```

Страна: Guatemala Зарплата: <=50K Среднее время работы в неделю: 39.36065573770492
Страна: Guatemala Зарплата: >50K Среднее время работы в неделю: 36.666666666666664
Страна: Haiti Зарплата: <=50K Среднее время работы в неделю: 36.325
Страна: Haiti Зарплата: >50K Среднее время работы в неделю: 42.75
Страна: Holand-Netherlands Зарплата: <=50K Среднее время работы в неделю: 40.0
Страна: Honduras Зарплата: <=50K Среднее время работы в неделю: 34.333333333333336
Страна: Honduras Зарплата: >50K Среднее время работы в неделю: 60.0
Страна: Hong Зарплата: <=50K Среднее время работы в неделю: 39.142857142857146
Страна: Hong Зарплата: >50K Среднее время работы в неделю: 45.0
Страна: Hungary Зарплата: <=50K Среднее время работы в неделю: 31.3
Страна: Hungary Зарплата: >50K Среднее время работы в неделю: 50.0
Страна: India Зарплата: <=50K Среднее время работы в неделю: 38.233333333333334
Страна: India Зарплата: >50K Среднее время работы в неделю: 46.475
Страна: Iran Зарплата: <=50K Среднее время работы в неделю: 41.44
Страна: Iran Зарплата: >50K Среднее время работы в неделю: 47.5
Страна: Ireland Зарплата: <=50K Среднее время работы в неделю: 40.94736842105263
Страна: Ireland Зарплата: >50K Среднее время работы в неделю: 48.0
Страна: Italy Зарплата: <=50K Среднее время работы в неделю: 39.625
Страна: Italy Зарплата: >50K Среднее время работы в неделю: 45.4
Страна: Jamaica Зарплата: <=50K Среднее время работы в неделю: 38.23943661971831
Страна: Jamaica Зарплата: >50K Среднее время работы в неделю: 41.1
Страна: Japan Зарплата: <=50K Среднее время работы в неделю: 41.0
Страна: Japan Зарплата: >50K Среднее время работы в неделю: 47.958333333333336
Страна: Laos Зарплата: <=50K Среднее время работы в неделю: 40.375
Страна: Laos Зарплата: >50K Среднее время работы в неделю: 40.0
Страна: Mexico Зарплата: <=50K Среднее время работы в неделю: 40.00327868852459
Страна: Mexico Зарплата: >50K Среднее время работы в неделю: 46.57575757575758
Страна: Nicaragua Зарплата: <=50K Среднее время работы в неделю: 36.09375
Страна: Nicaragua Зарплата: >50K Среднее время работы в неделю: 37.5
Страна: Outlying-US(Guam-USVI-etc) Зарплата: <=50K Среднее время работы в неделю: 41.857142857142854
Страна: Peru Зарплата: <=50K Среднее время работы в неделю: 35.06896551724138
Страна: Peru Зарплата: >50K Среднее время работы в неделю: 40.0
Страна: Philippines Зарплата: <=50K Среднее время работы в неделю: 38.065693430656935
Страна: Philippines Зарплата: >50K Среднее время работы в неделю: 43.032786885245905
Страна: Poland Зарплата: <=50K Среднее время работы в неделю: 38.166666666666664
Страна: Poland Зарплата: >50K Среднее время работы в неделю: 39.0
Страна: Portugal Зарплата: <=50K Среднее время работы в неделю: 41.93939393939394
Страна: Portugal Зарплата: >50K Среднее время работы в неделю: 41.5

Страна: Puerto-Rico Зарплата: <=50K Среднее время работы в неделю: 38.470588235294116
Страна: Puerto-Rico Зарплата: >50K Среднее время работы в неделю: 39.416666666666664
Страна: Scotland Зарплата: <=50K Среднее время работы в неделю: 39.44444444444444
Страна: Scotland Зарплата: >50K Среднее время работы в неделю: 46.666666666666664
Страна: South Зарплата: <=50K Среднее время работы в неделю: 40.15625
Страна: South Зарплата: >50K Среднее время работы в неделю: 51.4375
Страна: Taiwan Зарплата: <=50K Среднее время работы в неделю: 33.774193548387096
Страна: Taiwan Зарплата: >50K Среднее время работы в неделю: 46.8
Страна: Thailand Зарплата: <=50K Среднее время работы в неделю: 42.86666666666667
Страна: Thailand Зарплата: >50K Среднее время работы в неделю: 58.333333333333336
Страна: Trinadad&Tobago Зарплата: <=50K Среднее время работы в неделю: 37.05882352941177
Страна: Trinadad&Tobago Зарплата: >50K Среднее время работы в неделю: 40.0
Страна: United-States Зарплата: <=50K Среднее время работы в неделю: 38.79912723305605
Страна: United-States Зарплата: >50K Среднее время работы в неделю: 45.50536884674383
Страна: Vietnam Зарплата: <=50K Среднее время работы в неделю: 37.193548387096776
Страна: Vietnam Зарплата: >50K Среднее время работы в неделю: 39.2
Страна: Yugoslavia Зарплата: <=50K Среднее время работы в неделю: 41.6
Страна: Yugoslavia Зарплата: >50K Среднее время работы в неделю: 49.5