Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ _____ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ_____

КАФЕДРА _____СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ (ИУ5)____

О Т Ч Е Т

Лабораторная работа №2

по дисциплине: Машинное обучение _____

на тему: Изучение библиотек обработки данных.

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

Студент __ИУ5-63Б_____ _____ **Садыков М.Р.**
(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель _____ _____
(Подпись, дата) (И.О.Фамилия)

2020 г.

Лабораторная работа №2

Информация о наборе данных: age: continuous. workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked. fnlwgt: continuous. education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool. education-num: continuous. marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse. occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces. relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried. race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black. sex: Female, Male. capital-gain: continuous. capital-loss: continuous. hours-per-week: continuous. native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands. salary: >50K,<=50K

In [3]:

```
import numpy as np
import pandas as pd
pd.set_option('display.max.columns', 100)
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

In [4]:

```
#Загрузим данные в data и посмотрим вид наборы данных
data = pd.read_csv('adult.data.csv')
data.head()
```

Out[4]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black

1. Сколько мужчин и женщин представлено в этом наборе данных?

In [5]:

```
data['sex'].value_counts()
```

Out[5]:

```
Male      21790
Female    10771
Name: sex, dtype: int64
```

2. Каков средний возраст женщины?

In [6]:

```
data.loc[data['sex'] == 'Female', 'age'].mean()
```

Out[6]:

```
36.85823043357163
```

3. Каков процент граждан Германии (характеристика родной страны)?

In [7]:

```
(float((data['native-country'] == 'Germany').sum()) / data.shape[0])*100
```

Out[7]:

```
0.42074874850281013
```

4-5. Каково среднее значение и стандартное отклонение возраста для тех, кто зарабатывает более 50 тысяч в год (функция заработной платы) и тех, кто зарабатывает менее 50 тысяч в год?

In [8]:

```
list_of_ages_more_50k = data.loc[data['salary'] == '>50K', 'age']
list_of_ages_less_50k = data.loc[data['salary'] == '<=50K', 'age']

print('Средний возраст для тех, кто зарабатывает более 50 тысяч', round(list_of_
ages_more_50k.mean()))
print('Стандартное отклонение возраста для тех, кто зарабатывает более 50 тысяч'
, round(list_of_ages_more_50k.std(), 1))
print('Средний возраст для тех, кто зарабатывает менее 50 тысяч', round(list_of_
ages_less_50k.mean()))
print('Стандартное отклонение возраста для тех, кто зарабатывает менее 50 тысяч'
, round(list_of_ages_less_50k.std(), 1))
```

```
Средний возраст для тех, кто зарабатывает более 50 тысяч 44.0
Стандартное отклонение возраста для тех, кто зарабатывает более 50 т
ысяч 10.5
Средний возраст для тех, кто зарабатывает менее 50 тысяч 37.0
Стандартное отклонение возраста для тех, кто зарабатывает менее 50 т
ысяч 14.0
```

6. Правда ли, что люди, которые зарабатывают более 50 тысяч, имеют хотя бы среднее образование? (education – Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters or Doctorate feature)

In [19]:

```
sample = data.loc[data['salary'] == '>50K', 'education'].unique()
check = ['Bachelors', 'Prof-school', 'Assoc-acdm', 'Assoc-voc', 'Masters', 'Doct
orate feature']
if(sample == check):
    print('Правда')
else:
    print('Не правда')
```

```
Не правда
```

7. Отобразите возрастную статистику для каждой расы и каждого пола. Используйте groupby() и describe(). Найти максимальный возраст среди мужчин американо-индийско-эскимосской расы (Amer-Indian-Eskimo race).

In [30]:

```
for (race, sex), item in data.groupby(['race', 'sex']):
    print('Расса:', race, ', пол:', sex)
    print(item['age'].describe())
    print()

print('Максимальный возраст среди мужчин американо-индийско-эскимосской расы:',
max(data.loc[data['race'] == 'Amer-Indian-Eskimo', 'age']))
```

```
Расса: Amer-Indian-Eskimo , пол: 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

Расса: Amer-Indian-Eskimo , пол: 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

Расса: Asian-Pac-Islander , пол: 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

Расса: Asian-Pac-Islander , пол: 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

Расса: Black , пол: 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

Расса: Black , пол: 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

Расса: Other , пол: 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

Расса: Other , пол: 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

Расса: White , пол: 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

Расса: White , пол: 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

Максимальный возраст среди мужчин американо-индийско-эскимосской рас
ы: 82
```

8. Среди кого больше доля тех, кто много зарабатывает (> 50 тыс.): в браке или одинокие мужчины? Считается, что в браке находятся те, кто имеет семейное положение, начиная с женатых Married-civ-spouse, Married-spouse-absent or Married-AF-spouse), остальные считаются холостяками.

In [55]:

```
men = data.loc[(data['sex'] == 'Male')]
un_married_men = men.loc[(data['marital-status'].isin(['Never-married', 'Separat
ed', 'Divorced', 'Widowed']))]
married_men = men.loc[(data['marital-status'].isin(['Married-civ-spouse', 'Marri
ed-spouse-absent', 'Married-AF-spouse']))]

un_married_men_more50k = un_married_men.loc[(data['salary'] == '>50K')]
un_married_men_less50k = un_married_men.loc[(data['salary'] == '<=50K')]
married_men_more50k = married_men .loc[(data['salary'] == '>50K')]
married_men_less50k = married_men .loc[(data['salary'] == '<=50K')]

if(len(un_married_men_more50k) > len(married_men_more50k)):
    print('Среди неженатых:', len(un_married_men_more50k))
elif(len(un_married_men_more50k) < len(married_men_more50k)):
    print('Среди женатых:', len(married_men_more50k))
else:
    print('Одинаковое количество:', len(married_men_more50k))
```

```
Среди женатых: 5965
```

9. Какое максимальное количество часов работает человек в неделю ? Сколько человек работает такое количество часов, и каков процент тех, кто зарабатывает (> 50 тыс.) среди них?

In [71]:

```
max_working_time = max(data['hours-per-week'])
max_working_time_list = data.loc[(data['hours-per-week'] == max_working_time)]
max_working_time_more50_list = max_working_time_list.loc[(data['salary'] == '>50
K')]

print('Максимальное количество часов работает человек в неделю:', max_working_ti
me)
print('Количество человек, работающее такое количество часов:', len(max_working_
time_list))
print('Каков процент тех, кто зарабатывает (> 50 тыс.) среди них:', 100 * float
((len(max_working_time_more50_list))/(len(max_working_time_list))))
```

```
Максимальное количество часов работает человек в неделю: 99
Количество человек, работающее такое количество часов: 85
Каков процент тех, кто зарабатывает (> 50 тыс.) среди них: 29.411764
705882355
```

10. Посчитайте среднее время работы для тех, кто мало и много зарабатывает для каждой страны (родной страны). Что это будет для Японии?

In [82]:

```
for (country, salary), item in data.groupby(['native-country', 'salary']):
    if ((country == 'Japan') and (salary == '>50K')):
        japan_more50 = item ['hours-per-week'].mean()
    if ((country == 'Japan') and (salary == '<=50K')):
        japan_less50 = item ['hours-per-week'].mean()
    print('Страна:', country,', Запралата:', salary,', Среднее время работы:', i
tem ['hours-per-week'].mean())

print()
print('Япония, запралата >50K:', japan_more50)
print('Япония, запралата <=50K:', japan_less50)
```

```
Страна: ? , Запралата: <=50K , Среднее время работы: 40.164759725400
46
Страна: ? , Запралата: >50K , Среднее время работы: 45.5479452054794
5
Страна: Cambodia , Запралата: <=50K , Среднее время работы: 41.41666
6666666664
Страна: Cambodia , Запралата: >50K , Среднее время работы: 40.0
Страна: Canada , Запралата: <=50K , Среднее время работы: 37.9146341
4634146
Страна: Canada , Запралата: >50K , Среднее время работы: 45.64102564
102564
Страна: China , Запралата: <=50K , Среднее время работы: 37.38181818
181818
Страна: China , Запралата: >50K , Среднее время работы: 38.9
Страна: Columbia , Запралата: <=50K , Среднее время работы: 38.68421
052631579
Страна: Columbia , Запралата: >50K , Среднее время работы: 50.0
Страна: Cuba , Запралата: <=50K , Среднее время работы: 37.985714285
71429
Страна: Cuba , Запралата: >50K , Среднее время работы: 42.44
Страна: Dominican-Republic , Запралата: <=50K , Среднее время работ
ы: 42.338235294117645
Страна: Dominican-Republic , Запралата: >50K , Среднее время работы:
47.0
Страна: Ecuador , Запралата: <=50K , Среднее время работы: 38.041666
666666664
Страна: Ecuador , Запралата: >50K , Среднее время работы: 48.75
Страна: El-Salvador , Запралата: <=50K , Среднее время работы: 36.03
092783505155
Страна: El-Salvador , Запралата: >50K , Среднее время работы: 45.0
Страна: England , Запралата: <=50K , Среднее время работы: 40.483333
333333334
Страна: England , Запралата: >50K , Среднее время работы: 44.5333333
3333333
Страна: France , Запралата: <=50K , Среднее время работы: 41.0588235
2941177
Страна: France , Запралата: >50K , Среднее время работы: 50.75
Страна: Germany , Запралата: <=50K , Среднее время работы: 39.139784
94623656
Страна: Germany , Запралата: >50K , Среднее время работы: 44.9772727
2727273
Страна: Greece , Запралата: <=50K , Среднее время работы: 41.8095238
0952381
Страна: Greece , Запралата: >50K , Среднее время работы: 50.625
Страна: Guatemala , Запралата: <=50K , Среднее время работы: 39.3606
5573770492
Страна: Guatemala , Запралата: >50K , Среднее время работы: 36.66666
6666666664
Страна: Haiti , Запралата: <=50K , Среднее время работы: 36.325
Страна: Haiti , Запралата: >50K , Среднее время работы: 42.75
Страна: Holand-Netherlands , Запралата: <=50K , Среднее время работ
ы: 40.0
Страна: Honduras , Запралата: <=50K , Среднее время работы: 34.33333
3333333336
Страна: Honduras , Запралата: >50K , Среднее время работы: 60.0
Страна: Hong , Запралата: <=50K , Среднее время работы: 39.142857142
857146
Страна: Hong , Запралата: >50K , Среднее время работы: 45.0
Страна: Hungary , Запралата: <=50K , Среднее время работы: 31.3
Страна: Hungary , Запралата: >50K , Среднее время работы: 50.0
Страна: India , Запралата: <=50K , Среднее время работы: 38.23333333
```

```
3333334
Страна: India , Запралата: >50K , Среднее время работы: 46.475
Страна: Iran , Запралата: <=50K , Среднее время работы: 41.44
Страна: Iran , Запралата: >50K , Среднее время работы: 47.5
Страна: Ireland , Запралата: <=50K , Среднее время работы: 40.947368
42105263
Страна: Ireland , Запралата: >50K , Среднее время работы: 48.0
Страна: Italy , Запралата: <=50K , Среднее время работы: 39.625
Страна: Italy , Запралата: >50K , Среднее время работы: 45.4
Страна: Jamaica , Запралата: <=50K , Среднее время работы: 38.239436
61971831
Страна: Jamaica , Запралата: >50K , Среднее время работы: 41.1
Страна: Japan , Запралата: <=50K , Среднее время работы: 41.0
Страна: Japan , Запралата: >50K , Среднее время работы: 47.958333333
333336
Страна: Laos , Запралата: <=50K , Среднее время работы: 40.375
Страна: Laos , Запралата: >50K , Среднее время работы: 40.0
Страна: Mexico , Запралата: <=50K , Среднее время работы: 40.0032786
8852459
Страна: Mexico , Запралата: >50K , Среднее время работы: 46.57575757
575758
Страна: Nicaragua , Запралата: <=50K , Среднее время работы: 36.0937
5
Страна: Nicaragua , Запралата: >50K , Среднее время работы: 37.5
Страна: Outlying-US(Guam-USVI-etc) , Запралата: <=50K , Среднее врем
я работы: 41.857142857142854
Страна: Peru , Запралата: <=50K , Среднее время работы: 35.068965517
24138
Страна: Peru , Запралата: >50K , Среднее время работы: 40.0
Страна: Philippines , Запралата: <=50K , Среднее время работы: 38.06
5693430656935
Страна: Philippines , Запралата: >50K , Среднее время работы: 43.032
786885245905
Страна: Poland , Запралата: <=50K , Среднее время работы: 38.1666666
66666664
Страна: Poland , Запралата: >50K , Среднее время работы: 39.0
Страна: Portugal , Запралата: <=50K , Среднее время работы: 41.93939
393939394
Страна: Portugal , Запралата: >50K , Среднее время работы: 41.5
Страна: Puerto-Rico , Запралата: <=50K , Среднее время работы: 38.47
0588235294116
Страна: Puerto-Rico , Запралата: >50K , Среднее время работы: 39.416
666666666664
Страна: Scotland , Запралата: <=50K , Среднее время работы: 39.44444
444444444
Страна: Scotland , Запралата: >50K , Среднее время работы: 46.666666
666666664
Страна: South , Запралата: <=50K , Среднее время работы: 40.15625
Страна: South , Запралата: >50K , Среднее время работы: 51.4375
Страна: Taiwan , Запралата: <=50K , Среднее время работы: 33.7741935
48387096
Страна: Taiwan , Запралата: >50K , Среднее время работы: 46.8
Страна: Thailand , Запралата: <=50K , Среднее время работы: 42.86666
666666667
Страна: Thailand , Запралата: >50K , Среднее время работы: 58.333333
333333336
Страна: Trinadad&Tobago , Запралата: <=50K , Среднее время работы: 3
7.05882352941177
Страна: Trinadad&Tobago , Запралата: >50K , Среднее время работы: 4
0.0
Страна: United-States , Запралата: <=50K , Среднее время работы: 38.
```

```
79912723305605
Страна: United-States , Запралата: >50K , Среднее время работы: 45.5
0536884674383
Страна: Vietnam , Запралата: <=50K , Среднее время работы: 37.193548
387096776
Страна: Vietnam , Запралата: >50K , Среднее время работы: 39.2
Страна: Yugoslavia , Запралата: <=50K , Среднее время работы: 41.6
Страна: Yugoslavia , Запралата: >50K , Среднее время работы: 49.5

Япония, запралата >50K: 47.958333333333336
Япония, запралата <=50K: 41.0
```