



**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 ИУ5 «Системы обработки информации и управления»**

**«Технологии разведочного анализа и обработки данных»
по курсу «Технологии машинного обучения»
Лабораторная работа №2
Вариант 5**

Выполнил:
студент группы ИУ5 – 62Б
Гринин О.Е.
подпись, дата

Проверил:
преподаватель кафедры ИУ5
Гапанюк Ю.Е.
подпись, дата

2020 г.

```

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 [25]:

```

data = pd.read_csv('data/adult.data', sep = ',')

data.head()

```

Out[25]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
039		State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
150		Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
238		Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
353		Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
428		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 [26]:

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

Out[26]:

```

Male      21790
Female    10771
Name: sex, dtype: int64

```

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

In [33]:

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

Out[33]:

36.85823043357163

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

In [29]:

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

Out[29]:

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 [38]:

```
ages1 = data.loc[data['salary'] == '>50K', 'age']
ages2 = data.loc[data['salary'] == '<=50K', 'age']
print("The average age of the rich: {0} +- {1} years, poor - {2} +- {3} years.".format(
    round(ages1.mean()), round(ages1.std(), 1),
    round(ages2.mean()), round(ages2.std(), 1)))
```

The average age of the rich: 44 +- 10.5 years, poor - 37 +- 14.0 years.

1. 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 [40]:

```
data.loc[data['salary'] == '>50K', 'education'].unique()
```

Out[40]:

```
array(['HS-grad', 'Masters', 'Bachelors', 'Some-college',
       'Assoc-voc', 'Doctorate', 'Prof-school', 'Assoc-acdm',
       '7th-8th', '12th', '10th', '11th', '9th', '5th-6th',
       '1st-4th'], dtype=object)
```

1. 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 [47]:

```
data.loc[data['race'] == 'Amer-Indian-Eskimo', 'age'].max()
```

Out[47]:

82

In [51]:

```
data.groupby('race')['age'].describe()
```

Out[51]:

	count	mean	std	min	25%	50%	75%	max
race								
Amer-Indian-Eskimo	311.0	37.173633	12.447130	17.0	28.0	35.0	45.5	82.0
Asian-Pac-Islander	1039.0	37.746872	12.825133	17.0	28.0	36.0	45.0	90.0

	count	mean	std	min	25%	50%	75%	max
race								
Black	3124.0	37.767926	12.759290	17.0	28.0	36.0	46.0	90.0
Other	271.0	33.457565	11.538865	17.0	25.0	31.0	41.0	77.0
White	27816.0	38.769881	13.782306	17.0	28.0	37.0	48.0	90.0

1. 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 [87]:

```
otv = data.loc[(data['salary'] == ' >50K') & (data['sex'] == ' Male') & (data['marital-
-status'].isin([' Never-married',
                ' Separated',
                ' Divorced',
                ' Widowed']))][ 'sex']

otv1 = data.loc[(data['salary'] == ' >50K') & (data['sex'] == ' Male')][ 'sex']
print('>50K Family', otv.count(), 'All', otv1.count())
```

>50K Family 697 All 6662

1. 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 [99]:

```
q = data['hours-per-week'].max()
t = data.loc[(data['salary'] == ' >50K') & (data['hours-per-week'] == q)][ 'salary'].co
unt()
e = data.loc[data['hours-per-week'] == q][ 'salary'].count()
print('All: ', e, " Rich: ", t, " RICH/ALL: ",int(t/e*100),"%")
```

All: 85 Rich: 25 RICH/ALL: 29 %

1. 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 [112]:

```
for (country, salary), sub_df in data.groupby(['native-country', 'salary']):
    print(country, salary, round(sub_df['hours-per-week'].mean(), 2))
```

```
? <=50K 40.16
? >50K 45.55
Cambodia <=50K 41.42
Cambodia >50K 40.0
Canada <=50K 37.91
Canada >50K 45.64
China <=50K 37.38
China >50K 38.9
Columbia <=50K 38.68
Columbia >50K 50.0
Cuba <=50K 37.99
```

Cuba >50K 42.44
Dominican-Republic <=50K 42.34
Dominican-Republic >50K 47.0
Ecuador <=50K 38.04
Ecuador >50K 48.75
El-Salvador <=50K 36.03
El-Salvador >50K 45.0
England <=50K 40.48
England >50K 44.53
France <=50K 41.06
France >50K 50.75
Germany <=50K 39.14
Germany >50K 44.98
Greece <=50K 41.81
Greece >50K 50.62
Guatemala <=50K 39.36
Guatemala >50K 36.67
Haiti <=50K 36.33
Haiti >50K 42.75
Holand-Netherlands <=50K 40.0
Honduras <=50K 34.33
Honduras >50K 60.0
Hong <=50K 39.14
Hong >50K 45.0
Hungary <=50K 31.3
Hungary >50K 50.0
India <=50K 38.23
India >50K 46.48
Iran <=50K 41.44
Iran >50K 47.5
Ireland <=50K 40.95
Ireland >50K 48.0
Italy <=50K 39.62
Italy >50K 45.4
Jamaica <=50K 38.24
Jamaica >50K 41.1
Japan <=50K 41.0
Japan >50K 47.96
Laos <=50K 40.38
Laos >50K 40.0
Mexico <=50K 40.0
Mexico >50K 46.58
Nicaragua <=50K 36.09
Nicaragua >50K 37.5
Outlying-US (Guam-USVI-etc) <=50K 41.86
Peru <=50K 35.07
Peru >50K 40.0
Philippines <=50K 38.07
Philippines >50K 43.03
Poland <=50K 38.17
Poland >50K 39.0
Portugal <=50K 41.94
Portugal >50K 41.5
Puerto-Rico <=50K 38.47
Puerto-Rico >50K 39.42
Scotland <=50K 39.44
Scotland >50K 46.67
South <=50K 40.16
South >50K 51.44
Taiwan <=50K 33.77
Taiwan >50K 46.8
Thailand <=50K 42.87
Thailand >50K 58.33
Trinidad&Tobago <=50K 37.06
Trinidad&Tobago >50K 40.0
United-States <=50K 38.8
United-States >50K 45.51
Vietnam <=50K 37.19
Vietnam >50K 39.2
Yugoslavia <=50K 41.6
Yugoslavia >50K 49.5

