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)

Отчет
Лабораторная работа №2

«Изучение библиотек обработки данных»

По курсу: «Технологии машинного обучения»

Выполнил: студент группы ИУ5-64Б _____ Корыткина А.Н.
(Подпись, дата) (Ф.И.О.)

Проверил: _____ Гапанюк Ю.Е.
(Подпись, дата) (Ф.И.О.)

2020 г.

Цель лабораторной работы: изучение библиотеки обработки данных Pandas.

Описание: выполнение первого демонстрационного задания "demo assignment" под названием "Exploratory data analysis with Pandas" со страницы курса https://mlcourse.ai/assignments.

Текст программы и экранные формы с примерами выполнения программы:

In this task you should use Pandas to answer a few questions about the Adult dataset. (You don't the repository). Choose the answers in the web-form.

```
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')
```

```
data = pd.read_csv('data/adult.data.csv') data.head()
```

	age	workclass	fnlwgt	education	education-num	marital-status	occ
0	39	State-gov	77516	Bachelors	13	Never-married	Ad
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-m
2	38	Private	215646	HS-grad	9	Divorced	Handlers
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof

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

```
data['sex'].value_counts() # data.groupby('sex').count()
```

```
Male      21790
Female    10771
```

Name: sex, dtype: int64

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

```
data.groupby(['sex'])['age'].mean()
```

```
sex
Female    36.858230
Male      39.433547
Name: age, dtype: float64
```

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

```
print(round((data['native-country'] == 'Germany').sum() / data.shape[0] * 100, 2),
```

0.42 %

4-5) What are the mean and standard deviation of age for those who earn more than 50K per year

```
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.".fo round(ages1.mean()),
      round(ages1.std(), 1),
      round(ages2.mean()), round(ages2.std(), 1)))
```

The average age of the rich: 44.0 +- 10.5 years, poor - 37.0 +- 14.0 years.

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

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

```
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)
```

No, it isn't true

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

```
data.groupby(['race', 'sex'])['age'].describe() # the maximum age of men of Amer-I
```

race	sex	count	mean	std	min	25%	50%	75%	max
Amer-Indian-Eskimo	Female	119.0	37.117647	13.114991	17.0	27.0	36.0	46.00	80.0
	Male	192.0	37.208333	12.049563	17.0	28.0	35.0	45.00	82.0
Asian-Pac-Islander	Female	346.0	35.089595	12.300845	17.0	25.0	33.0	43.75	75.0
	Male	693.0	39.073593	12.883944	18.0	29.0	37.0	46.00	90.0
Black	Female	1555.0	37.854019	12.637197	17.0	28.0	37.0	46.00	90.0
	Male	1569.0	37.682600	12.882612	17.0	27.0	36.0	46.00	90.0
Other	Female	109.0	31.678899	11.631599	17.0	23.0	29.0	39.00	74.0
	Male	162.0	34.654321	11.355531	17.0	26.0	32.0	42.00	77.0
White	Female	8642.0	36.811618	14.329093	17.0	25.0	35.0	46.00	90.0
	Male	19174.0	39.652498	13.436029	17.0	29.0	38.0	49.00	90.0

8) Among whom is the proportion of those who earn a lot (>50K) greater: married or single men (those who have a marital-status starting with Married (Married-civ-spouse, Married-spouse-absconsidered bachelors.

```
data.loc[(data['sex'] == 'Male') & (~data['marital-status'].str.startswith('Marrie
```

```
<=50K    7552
>50K      697
Name: salary, dtype: int64
```

```
data.loc[(data['sex'] == 'Male') & (data['marital-status'].str.startswith('Married
```

```
<=50K    7576
>50K     5965
Name: salary, dtype: int64
```

married > singe men (earn >50K)

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

```
max_num = data['hours-per-week'].max()
quantity = data.loc[data['hours-per-week'] == max_num, 'age'].count() per=data[(data['hours-per-week'] ==
max_num) & (data['salary'] == '>50K')].shape[0 print('maximum number of hours a person works per week^ ',
max_num)
print('people work such a number of hours: ', quantity)
print('the percentage of those who earn a lot (>50K): ', round(per, 2), "%")
```

```
maximum number of hours a person works per week^                99
people work such a number of hours:             85
the percentage of those who earn a lot (>50K):                  29.41 %
```

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

```
pd.options.display.max_rows = 999
data.groupby(['native-country', 'salary'])['hours-per-week'].mean()
```

native-country	salary	
?	<=50K	40.164760
	>50K	45.547945
Cambodia	<=50K	41.416667
	>50K	40.000000
Canada	<=50K	37.914634
	>50K	45.641026
China	<=50K	37.381818
	>50K	38.900000
Columbia	<=50K	38.684211
	>50K	50.000000
Cuba	<=50K	37.985714
	>50K	42.440000
Dominican-Republic	<=50K	42.338235
	>50K	47.000000
Ecuador	<=50K	38.041667
	>50K	48.750000
El-Salvador	<=50K	36.030928
	>50K	45.000000
England	<=50K	40.483333
	>50K	44.533333
France	<=50K	41.058824
	>50K	50.750000
Germany	<=50K	39.139785
	>50K	44.977273
Greece	<=50K	41.809524
	>50K	50.625000
Guatemala	<=50K	39.360656
	>50K	36.666667
Haiti	<=50K	36.325000
	>50K	42.750000
Holand-Netherlands	<=50K	40.000000
Honduras	<=50K	34.333333
	>50K	60.000000
Hong	<=50K	39.142857
	>50K	45.000000
Hungary	<=50K	31.300000
	>50K	50.000000
India	<=50K	38.233333
	>50K	46.475000

Iran	<=50K	41.440000
	>50K	47.500000
Ireland	<=50K	40.947368
	>50K	48.000000
Italy	<=50K	39.625000
	>50K	45.400000
Jamaica	<=50K	38.239437
	>50K	41.100000
Japan	<=50K	41.000000
	>50K	47.958333
Laos	<=50K	40.375000
	>50K	40.000000
Mexico	<=50K	40.003279
	>50K	46.575758
Nicaragua	<=50K	36.093750
	>50K	37.500000
Outlying-US(Guam-USVI-etc)	<=50K	41.857143
Peru	<=50K	35.068966

	>50K	40.000000
Philippines	<=50K	38.065693
	>50K	43.032787
Poland	<=50K	38.166667
	>50K	39.000000
Portugal	<=50K	41.939394

Japan <=50K 41.000000 >50K 47.958333

Scotland	<=50K	39.444444
	>50K	46.666667
South	<=50K	40.156250
	>50K	51.437500
Taiwan	<=50K	33.774194
	>50K	46.800000
Thailand	<=50K	42.866667
	>50K	58.333333
Trinadad&Tobago	<=50K	37.058824
	>50K	40.000000
United-States	<=50K	38.799127
	>50K	45.505369
Vietnam	<=50K	37.193548