Московский государственный технический университет им. Н. Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

Курс «Технологии машинного обучения»

Отчет по лабораторной работе №2

Изучение библиотек обработки данных

Группа: ИУ5-62Б

Студент: Селедкина А.С.

Преподаватель: Гапанюк Ю.Е.

Москва, 2020 г.

Цель лабораторной работы: изучение библиотеки обработки данных Pandas.

Описание задания

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

Описание набора 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

Текст программы и примеры выполнения

```
import numpy as np
import pandas as pd
pd.set_option('display.max.columns', 100)
pd.set_option('display.max.rows', 100)
%matplotlib inline
import matplotlib.pyplot as plt
import seaborn as sns
data = pd.read_csv('data/adult.data.csv')
data.head()
```

	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?

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

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

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

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

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

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

```
data[data['native-country'] == 'Germany'].shape[0] / data.shape[0] * 100.0
```

```
Out[5]: 0.42074874850281013
```

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?

```
data.groupby('salary').agg({'age': ['mean', 'std']})
```

	age	
	mean	std
salary		
<=50K	36.783738	14.020088
>50K	44.249841	10.519028

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)

```
high_salary_high_education = data[(data['salary'] == '>50K') &
      (data['education'].isin(['Bachelors', 'Prof-school', 'Assoc-acdm',
      'Assoc-voc', 'Masters', 'Doctorate']))].shape[0]
high_salary_high_education
```

```
Out[7]: 4535
```

```
high_salary_all = data[data['salary'] == '>50K'].shape[0]
high_salary_all
```

```
Out[8]: 7841
```

```
high_salary_high_education == high_salary_all
```

```
Out[9]: False
```

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.

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

		count	mean	std	min	25%	50%	75%	max
race	sex								
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

Answer: 82

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.

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

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
7	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
9	42	Private	159449	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	5178	0	40	United-States	>50K
10	37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32550	43	Self-emp-not-inc	27242	Some-college	10	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0	50	United-States	<=50K
32551	32	Private	34066	10th	6	Married-civ-spouse	Handlers-cleaners	Husband	Amer-Indian-Eskimo	Male	0	0	40	United-States	<=50K
32552	43	Private	84661	Assoc-voc	11	Married-civ-spouse	Sales	Husband	White	Male	0	0	45	United-States	<=50K
32554	53	Private	321865	Masters	14	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	40	United-States	>50K
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K

13541 rows × 15 columns

```
married_men_propotion = married_men[married_men['salary'] ==
'>50K'].shape[0] / married_men.shape[0]
married_men_propotion
```

```
Out[12]: 0.44051399453511156
```

```
single_men = data[(data['sex'] == 'Male') &
(data['marital-status'].isin(['Divorced', 'Never-married',
      'Separated', 'Widowed']))]
single_men
```

	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
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
13	32	Private	205019	Assoc-acdm	12	Never-married	Sales	Not-in-family	Black	Male	0	0	50	United-States	<=50K
16	25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35	United-States	<=50K
17	32	Private	186824	HS-grad	9	Never-married	Machine-op-inspct	Unmarried	White	Male	0	0	40	United-States	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32537	30	Private	345898	HS-grad	9	Never-married	Craft-repair	Not-in-family	Black	Male	0	0	46	United-States	<=50K
32548	65	Self-emp-not-inc	99359	Prof-school	15	Never-married	Prof-specialty	Not-in-family	White	Male	1086	0	60	United-States	<=50K
32553	32	Private	116138	Masters	14	Never-married	Tech-support	Not-in-family	Asian-Pac-Islander	Male	0	0	11	Taiwan	<=50K
32555	22	Private	310152	Some-college	10	Never-married	Protective-serv	Not-in-family	White	Male	0	0	40	United-States	<=50K
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K

8249 rows × 15 columns

```
single_men_propotion = single_men[single_men['salary'] == '>50K'].shape[0]
/ single_men.shape[0]
single_men_propotion
```

Out[14]: 0.08449509031397745

```
single_men_propotion > married_men_propotion
```

Out[15]: False

Answer: among married

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?

```
max_hours_number = data[['hours-per-week']].max()[0]
max_hours_number
```

Out[16]: 99

```
people_working_hard = data[data['hours-per-week'] == max_hours_number]
people_working_hard.shape[0]
```

Out[17]: 85

```
people_working_hard[people_working_hard['salary'] == '>50K'].shape[0] /
      people_working_hard.shape[0] * 100.0
```

Out[18]: 29.411764705882355

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?

```
data.groupby(['salary', 'native-country'])['hours-per-week'].mean()
```

```
salary  native-country
<=50K   ?                             40.164760
        Cambodia                      41.416667
        Canada                        37.914634
        China                         37.381818
        Columbia                      38.684211
        Cuba                          37.985714
        Dominican-Republic            42.338235
        Ecuador                       38.041667
        El-Salvador                   36.030928
        England                       40.483333
        France                        41.058824
        Germany                       39.139785
        Greece                        41.809524
        Guatemala                     39.360656
        Haiti                         36.325000
        Holand-Netherlands            40.000000
        Honduras                      34.333333
        Hong                          39.142857
        Hungary                       31.300000
        India                         38.233333
        Iran                          41.440000
        Ireland                       40.947368
        Italy                         39.625000
        Jamaica                       38.239437
        Japan                         41.000000
        Laos                          40.375000
        Mexico                        40.003279
        Nicaragua                     36.093750
        Outlying-US(Guam-USVI-etc)    41.857143
        Peru                          35.068966
        Philippines                   38.065693
        Poland                        38.166667
        Portugal                      41.939394
        Puerto-Rico                   38.470588
        Scotland                      39.444444
        South                         40.156250
        Taiwan                        33.774194
        Thailand                      42.866667
        Trinadad&Tobago               37.058824
        United-States                 38.799127
        Vietnam                       37.193548
        Yugoslavia                    41.600000
```

```
>50K    ?                             45.547945
        Cambodia                      40.000000
        Canada                        45.641026
        China                         38.900000
        Columbia                      50.000000
        Cuba                          42.440000
        Dominican-Republic            47.000000
        Ecuador                       48.750000
        El-Salvador                   45.000000
        England                       44.533333
        France                        50.750000
        Germany                       44.977273
        Greece                        50.625000
        Guatemala                     36.666667
        Haiti                         42.750000
        Honduras                      60.000000
        Hong                          45.000000
        Hungary                       50.000000
        India                         46.475000
        Iran                          47.500000
        Ireland                       48.000000
        Italy                         45.400000
        Jamaica                       41.100000
        Japan                         47.958333
        Laos                          40.000000
        Mexico                        46.575758
        Nicaragua                     37.500000
        Peru                          40.000000
        Philippines                   43.032787
        Poland                        39.000000
        Portugal                      41.500000
        Puerto-Rico                   39.416667
        Scotland                      46.666667
        South                         51.437500
        Taiwan                        46.800000
        Thailand                      58.333333
        Trinadad&Tobago               40.000000
        United-States                 45.505369
        Vietnam                       39.200000
        Yugoslavia                    49.500000
Name: hours-per-week, dtype: float64
```

Answer: 41 and 48