Московский государственный технический университет им. Н.Э. Баумана
Кафедра «Системы обработки информации и управления»

Лабораторная работа №2
по дисциплине
«Методы машинного обучения»
на тему
«Изучение библиотек обработки данных»

Выполнил:
студент группы ИУ5-64Б
Береговая Д. в.

_____

Москва — 2019 г.

1. Цель лабораторной работы

Изучить библиотеки обработки данных Pandas и PandaSQL.

2. Задание

Задание состоит из двух частей.

2.1. Часть 1

Требуется выполнить первое демонстрационное задание под названием «Exploratory data analysis with Pandas» со страницы курса mlcourse.ai.

2.2. Часть 2

Требуется выполнить следующие запросы с использованием двух различных библиотек — Pandas и PandaSQL:

один произвольный запрос на соединение двух наборов данных, один произвольный запрос на группировку набора данных с использованием функций агрегирования. Также требуется сравнить время выполнения каждого запроса в Pandas и PandaSQL.

2.3. Ход выполнения работы

[7]:
```
import numpy as np
import pandas as pd
```

[8]:
```
data = pd.read_csv('adult.data.csv')
data.head()
```

[8]:

	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

2.3.1. Количество мужчин и женщин

[9]:
```
data['sex'].value_counts()
```

[9]:
```
Male      21790
Female    10771
Name: sex, dtype: int64
```

2.3.2. Средний возраст женщин

[10]:
```
data.loc[data['sex'] == 'Female', 'age'].mean()
```

[10]:
```
36.85823043357163
```

2.3.3. Доля граждан Германии

[11]:
```
float((data['native-country'] == 'Germany').sum()) / data.shape[0]
```

[11]:
```
0.004207487485028101
```

2.3.4. Среднее значение и стандартное отклонение возраста людей следующих категорий:

- кто получал более 50 тысяч в год
- кто получал менее 50 тысяч в год

[12]:
```
ages1 = data.loc[data['salary'] == '>50K', 'age']
ages2 = data.loc[data['salary'] == '<=50K', 'age']
print("                    50    : {0} +- {1}   ,           50    : - {2} +-␣
 ↪{3} years.".format(
    round(ages1.mean()), round(ages1.std(), 1),
    round(ages2.mean()), round(ages2.std(), 1)))
```

```
                    50    : 44 +- 10.5   ,           50
  : - 37 +- 14.0 years.
```

2.4. Оценка образования людей, получающих больше 50к в год

[13]:
```
data.loc[data['salary'] == '>50K', 'education'].unique()
```

[13]:
```
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)
```

2.4.1. Статистика возрастов для каждой расы и пола

2.4.2. максимальный возраст мужчин расы Amer-Indian-Eskimo

[14]:
```
for (race, sex), sub_df in data.groupby(['race', 'sex']):
    print("  : {0},   : {1}".format(race, sex))
    print(sub_df['age'].describe())
```

```
  : Amer-Indian-Eskimo,   : 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
  : Amer-Indian-Eskimo,   : 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
  : Asian-Pac-Islander,   : 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
  : Asian-Pac-Islander,   : 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
  : Black,   : 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
  : Black,   : 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
  : Other,   : 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
  : Other,   : 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
  : White,   : 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
  : White,   : 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
```

2.5. Среди кого больше доля тех, кто зарабатывает больше 50 тыс в год: среди женатых мужчин или одиноких? (Женатые - те, у кого атрибут martial-status начинается с “Married”)

[15]:
```
data.loc[(data['sex'] == 'Male') &
     (data['marital-status'].isin(['Never-married',
                                   'Separated',
                                   'Divorced',
                                   'Widowed'])), 'salary'].value_counts()
```

[15]:
```
<=50K    7552
>50K      697
Name: salary, dtype: int64
```

[16]:
```
data.loc[(data['sex'] == 'Male') &
     (data['marital-status'].str.startswith('Married')), 'salary'].
 ↪value_counts()
```

[16]:
```
<=50K    7576
>50K     5965
Name: salary, dtype: int64
```

В среднем женатые мужчины зарабатывают больше

2.5.1. Максимальное количество часов, которые человек работает в неделю

2.5.2. Количество людей, работающих такое количество часов

2.5.3. Процент тех, кто много зарабатывает среди них

[17]:
```
max_load = data['hours-per-week'].max()
print("                        - {0}  ".format(max_load))

num_workaholics = data[data['hours-per-week'] == max_load].shape[0]
print("                  - {0}".format(num_workaholics))

rich_share = float(data[(data['hours-per-week'] == max_load)
                 & (data['salary'] == '>50K')].shape[0]) / num_workaholics
print("                     - {0}%".format(int(100 * rich_share)))
```

```
                        - 99
                  - 85
                     - 29%
```

2.6. Среднее время работы тех, кто зарабатывает мало и много для каждой страны

[18]:
```
pd.crosstab(data['native-country'], data['salary'],
            values=data['hours-per-week'], aggfunc=np.mean).T
```

[18]:
```
native-country          ?   Cambodia     Canada      China   Columbia  \
salary
<=50K           40.164760  41.416667  37.914634  37.381818  38.684211
>50K            45.547945  40.000000  45.641026  38.900000  50.000000

native-country       Cuba  Dominican-Republic    Ecuador  El-Salvador  \
salary
<=50K           37.985714           42.338235  38.041667    36.030928
>50K            42.440000           47.000000  48.750000    45.000000

native-country    England  …   Portugal  Puerto-Rico   Scotland     South  \
salary                     …
<=50K           40.483333  …  41.939394    38.470588  39.444444  40.15625
>50K            44.533333  …  41.500000    39.416667  46.666667  51.43750

native-country     Taiwan   Thailand  Trinadad&Tobago  United-States  \
salary
<=50K           33.774194  42.866667        37.058824      38.799127
>50K            46.800000  58.333333        40.000000      45.505369

native-country    Vietnam  Yugoslavia
salary
<=50K           37.193548        41.6
>50K            39.200000        49.5

[2 rows x 42 columns]
```

Таиланд бьёт все рекорды. Лучше всего там просто отдыхать)

[ ]:

[ ]:

[ ]: