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 ИУ5-23М
Иванников А. В.

Москва — 2019 г.

1. Цель лабораторной работы

Изучение библиотек обработки данных Pandas и PandaSQL

2. Задание

2.1. Часть 1.

Выполните первое демонстрационное задание “demo assignment” под названием “Exploratory data analysis with Pandas”

2.2. Часть 2.

Выполните 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

Сравните время выполнения каждого запроса в Pandas и PandaSQL.

2.3. Сформировать отчет и разместить его в своем репозитории на GitHub.

3. Ход выполнения работы

3.1. Исследовательский анализ данных с помощью Pandas

Используется набор данных о взрослых жителях США (data.adult.csv) - 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

[1]:
```
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')
```

[2]:
```
data = pd.read_csv('adult.data.csv')
data.head()
```

[2]:
```
   age         workclass  fnlwgt  education  education-num  \
0   39         State-gov   77516  Bachelors             13
1   50  Self-emp-not-inc   83311  Bachelors             13
2   38           Private  215646    HS-grad              9
3   53           Private  234721       11th              7
4   28           Private  338409  Bachelors             13

       marital-status         occupation   relationship   race     sex  \
0       Never-married       Adm-clerical  Not-in-family  White    Male
1  Married-civ-spouse    Exec-managerial        Husband  White    Male
2            Divorced  Handlers-cleaners  Not-in-family  White    Male
3  Married-civ-spouse  Handlers-cleaners        Husband  Black    Male
4  Married-civ-spouse     Prof-specialty           Wife  Black  Female

   capital-gain  capital-loss  hours-per-week native-country salary
0          2174             0              40  United-States  <=50K
1             0             0              13  United-States  <=50K
2             0             0              40  United-States  <=50K
3             0             0              40  United-States  <=50K
4             0             0              40           Cuba  <=50K
```

3.1.1. Количество женщин и мужчин в наборе данных

[3]:
```
data['sex'].value_counts()
```

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

3.2. Средний возраст женщин

[4]:
```
data.loc[data['sex'] == 'Female', 'age'].mean()
```

[4]:
```
36.85823043357163
```

3.2.1. Доля родившихся в Германии

[5]:
```
float((data['native-country'] == 'Germany').sum()) / data.shape[0]
```

[5]:
```
0.004207487485028101
```

3.2.2. Средний возраст людей, а также отклонение от среднего, с заработком меньше 50k и больше 50k

[6]:
```
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: 44.0 +- 10.5 years, poor - 37.0 +- 14.0 years.
```

3.2.3. Правда ли, что ли, зарабатывающие больше 50к - с высшим образованием?

[7]:
```
data.loc[data['salary'] == '>50K', 'education'].value_counts()
### Нет, люди со средним и средне-специальным образованием также получают
 ↪больше 50к, хотя их и немного
```

[7]:
```
Bachelors       2221
HS-grad         1675
Some-college    1387
Masters          959
Prof-school      423
Assoc-voc        361
Doctorate        306
Assoc-acdm       265
10th              62
11th              60
7th-8th           40
12th              33
9th               27
5th-6th           16
1st-4th            6
Name: education, dtype: int64
```

3.2.4. Статистика по каждой расе и полу. Использование groupby() и describe(). Нахождение наиболее возрастного человека расы Amer-Indian-Eskimo

[8]:
```
for (race, sex), sub_df in data.groupby(['race', 'sex']):
    print("Race: {0}, sex: {1}".format(race, sex))
    print(sub_df['age'].describe())
```

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

3.2.5. Доля мужчик с заработком больше 50к выше среди мужчин в браке или холостых?

[9]:
```
data.loc[(data['sex'] == 'Male') &
     (data['marital-status'].isin(['Never-married',
                                   'Separated',
                                   'Divorced',
```

```
                                'Widowed'])), 'salary'].value_counts()
```

[9]:
```
<=50K    7552
>50K      697
Name: salary, dtype: int64
```

[10]:
```
data.loc[(data['sex'] == 'Male') &
     (data['marital-status'].str.startswith('Married')), 'salary'].
 ↪value_counts()
```

[10]:
```
<=50K    7576
>50K     5965
Name: salary, dtype: int64
```

[11]:
```
### Среди женатых людей доля обеспеченных выше
```

3.2.6. Наибольше число рабочих часов в неделю, количество людей с таким количеством часов, процент обеспеченных среди них

[12]:
```
max_load = data['hours-per-week'].max()
print("Max time - {0} hours./week.".format(max_load))

num_workaholics = data[data['hours-per-week'] == max_load].shape[0]
print("Total number of such hard workers {0}".format(num_workaholics))

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

```
Max time - 99 hours./week.
Total number of such hard workers 85
Percentage of rich among them 29%
```

3.2.7. Среднее число рабочих часов для людей с разным заработком для каждый страны

[13]:
```
pd.crosstab(data['native-country'], data['salary'],
           values=data['hours-per-week'], aggfunc=np.mean).T
```

[13]:
```
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     France    Germany     Greece  Guatemala  ␣
 ↪Haiti  \
```

```
salary
<=50K          40.483333  41.058824  39.139785  41.809524  39.360656  36.
 ↪325
>50K           44.533333  50.750000  44.977273  50.625000  36.666667  42.
 ↪750

native-country  Holand-Netherlands   Honduras       Hong  Hungary
 ↪India  \
salary
<=50K                        40.0  34.333333  39.142857     31.3  38.
 ↪233333
>50K                          NaN  60.000000  45.000000     50.0  46.
 ↪475000

native-country   Iran    Ireland   Italy    Jamaica      Japan    Laos  \
salary
<=50K          41.44  40.947368  39.625  38.239437  41.000000  40.375
>50K           47.50  48.000000  45.400  41.100000  47.958333  40.000

native-country    Mexico  Nicaragua  Outlying-US(Guam-USVI-etc)
 ↪Peru  \
salary
<=50K          40.003279   36.09375                    41.857143  35.068966
>50K           46.575758   37.50000                          NaN  40.000000

native-country  Philippines     Poland   Portugal  Puerto-Rico    Scotland
 ↪\
salary
<=50K             38.065693  38.166667  41.939394    38.470588  39.444444
>50K              43.032787  39.000000  41.500000    39.416667  46.666667

native-country     South     Taiwan   Thailand  Trinadad&Tobago  \
salary
<=50K          40.15625  33.774194  42.866667        37.058824
>50K           51.43750  46.800000  58.333333        40.000000

native-country  United-States    Vietnam  Yugoslavia
salary
<=50K              38.799127  37.193548        41.6
>50K               45.505369  39.200000        49.5
```