Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

*Дисциплина «Технологии машинного обучения»*

**Отчёт**

по лабораторной работе №2

«Изучение библиотек обработки данных»

*Вариант 12*

Студент:

Крюков Г. М.

Группа ИУ5-61Б

Преподаватель:

Гапанюк Ю. Е.

Москва, 2020 г.

**Цель лабораторной работы:**

Изучение библиотеки обработки данных Pandas.

**Задание:**

* Условие задания - <https://nbviewer.jupyter.org/github/Yorko/mlcourse_open/blob/master/jupyter_english/assignments_demo/assignment01_pandas_uci_adult.ipynb?flush_cache=true>
* Официальный датасет находится здесь, но данные и заголовки хранятся отдельно, что неудобно для анализа - <https://archive.ics.uci.edu/ml/datasets/Adult>
* Поэтому готовый набор данных для лабораторной работы удобнее скачать здесь - <https://raw.githubusercontent.com/Yorko/mlcourse.ai/master/data/adult.data.csv> (удобнее всего нажать на данной ссылке правую кнопку мыши и выбрать в контекстном меню пункт "сохранить ссылку", будет предложено сохранить файл в формате CSV)

**Текст программы:**

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('/content/adult.data.csv')

data.head()

data['sex'].value\_counts()

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

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

ages1 = data.loc[data['salary'] == '>50K', 'age']

ages2 = data.loc[data['salary'] == '<=50K', 'age']

print("Average age of those, who recieve more than 50K per year : {0} +- {1} years, less than 50K per year : {2} +- {3} years.".format(

    round(ages1.mean()), round(ages1.std(), 1),

    round(ages2.mean()), round(ages2.std(), 1)))

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

for (race, sex), s in data.groupby(['race', 'sex']):

    print("Race: {0}, sex: {1}".format(race, sex))

    print(s['age'].describe())

data.loc[(data['sex'] == 'Male') &

     (data['marital-status'].isin(['Never-married',

                                   'Separated',

                                   'Divorced',

                                   'Widowed'])), 'salary'].value\_counts()

data.loc[(data['sex'] == 'Male') &

     (data['marital-status'].str.startswith('Married')), 'salary'].value\_counts()

data['marital-status'].value\_counts()

max\_load = data['hours-per-week'].max()

print("Maximum time = {0} hours./week.".format(max\_load))

num\_workers = data[data['hours-per-week'] == max\_load].shape[0]

print("Number of workers, who work such a number of hours: {0}".format(num\_workers))

rich\_share = float(data[(data['hours-per-week'] == max\_load)

                 & (data['salary'] == '>50K')].shape[0]) / num\_workaholics

print("Percentage of those who earn a lot (>50K) among them: {0}%".format(int(100 \* rich\_share)))

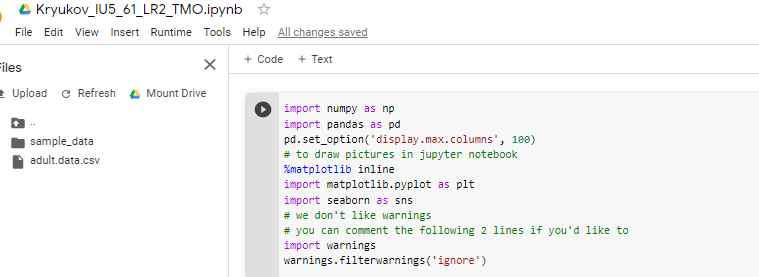
for (country, salary), sub\_df in data.groupby(['native-country', 'salary']):

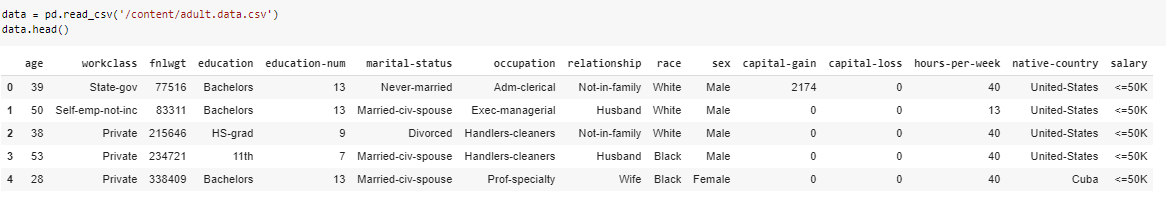
    print(country, salary, round(sub\_df['hours-per-week'].mean(), 2))

pd.crosstab(data['native-country'], data['salary'],

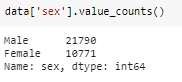
           values=data['hours-per-week'], aggfunc=np.mean).T

**Выполнение работы:**





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



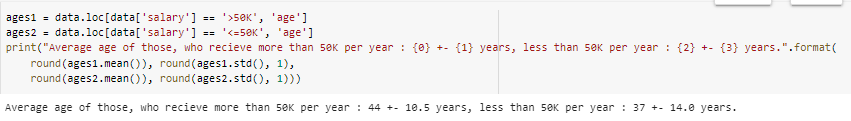
1. What is the average age (age feature) of women?



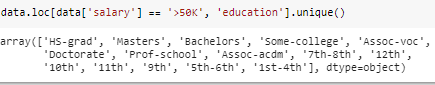
1. What is the percentage of German citizens (native-country feature)?



1. 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?

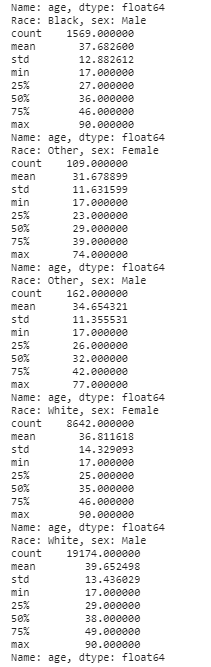
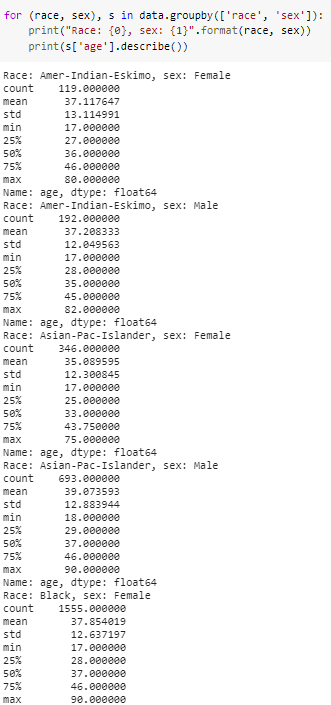
m

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)



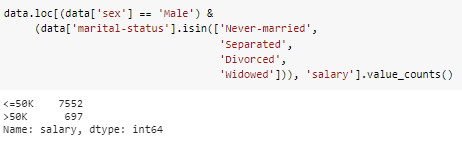
Answer: No, it’s not true

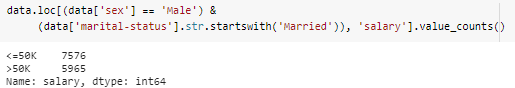
1. Display age statistics for each race (race feature) and each gender (sex feature). Use groupby() and describe(). Find the maximum age of men of Amer-Indian-Eskimo race.

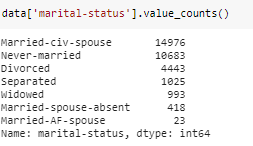


Answer: maximum age of men of Amer-Indian-Eskimo race = 80

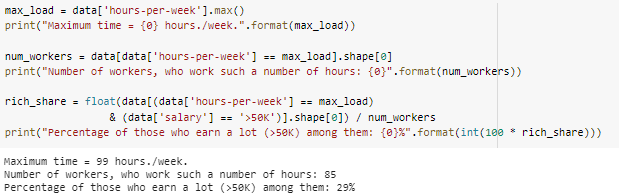
1. Among whom is the proportion of those who earn a lot (>50K) greater: married or single men (marital-status feature)? Consider as married those who have a marital-status starting with Married (Married-civ-spouse, Married-spouse-absent or Married-AF-spouse), the rest are considered bachelors.







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?



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?

