**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)

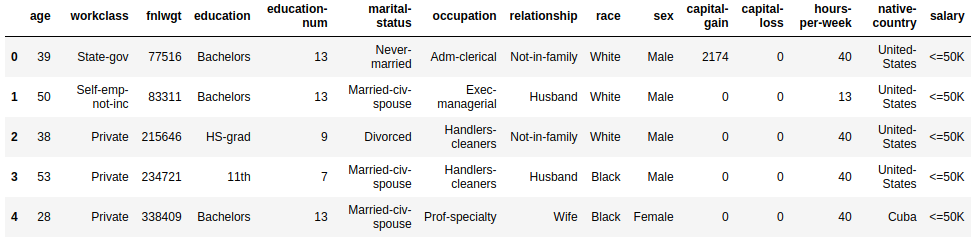
%matplotlib inline

**import** matplotlib.pyplot **as** plt

**import** seaborn **as** sns

data = pd.read\_csv('data/adult.data.csv')

data.head()



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

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



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

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



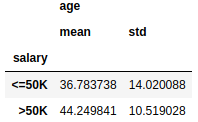
### **3. What is the percentage of German citizens (native-country feature)?**

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



### **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']})



### **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



high\_salary\_all = data[data['salary'] == '>50K'].shape[0]

high\_salary\_all

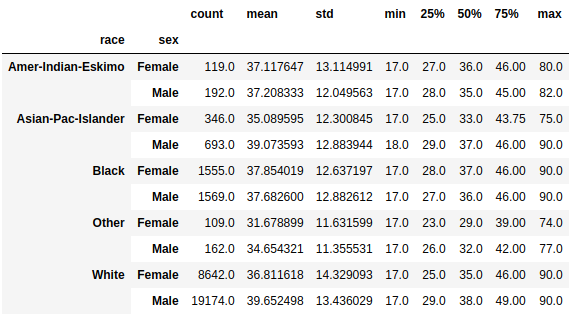


high\_salary\_high\_education == high\_salary\_all



### **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()

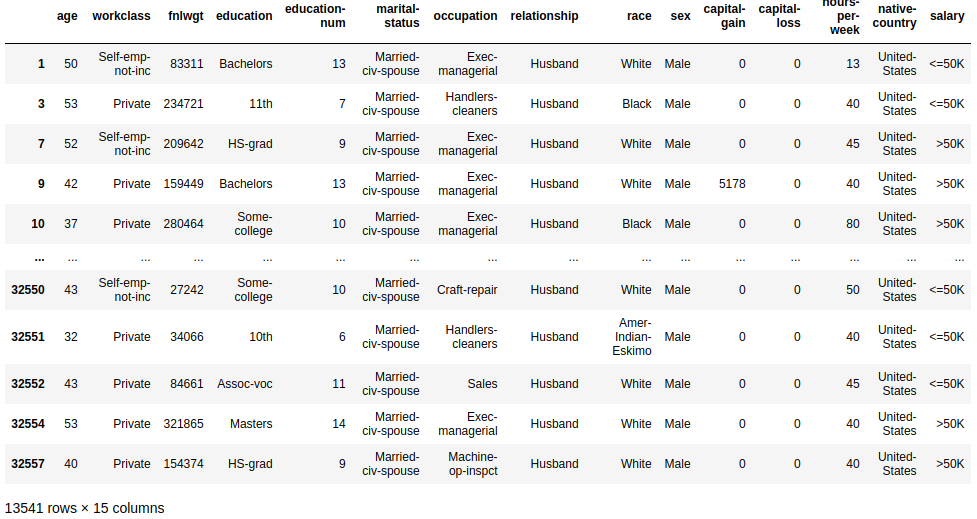


Answer: 82

### **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



married\_men\_propotion = married\_men[married\_men['salary'] == '>50K'].shape[0] / married\_men.shape[0]

married\_men\_propotion



single\_men = data[(data['sex'] == 'Male') & (data['marital-status'].isin(['Divorced', 'Never-married',

'Separated', 'Widowed']))]

single\_men



single\_men\_propotion = single\_men[single\_men['salary'] == '>50K'].shape[0] / single\_men.shape[0]

single\_men\_propotion



single\_men\_propotion > married\_men\_propotion



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



people\_working\_hard = data[data['hours-per-week'] == max\_hours\_number]

people\_working\_hard.shape[0]



people\_working\_hard[people\_working\_hard['salary'] == '>50K'].shape[0] /

people\_working\_hard.shape[0] \* 100.0



### **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()





Answer: 41 and 48