Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Информатика и системы управления

КАФЕДРА Системы обработки информации и управления

Отчёт по лабораторной работе № 2

По дисциплине:
«Технологии машинного обучения»

По теме:
«Изучение библиотек обработки данных»

Выполнил:

Студент группы ИУ5-63 _____ **Труфанов В.А.**

(Подпись, дата) (Фамилия И.О.)

Проверил:

_____ **Гапанюк Ю.Е.**

(Подпись, дата) (Фамилия И.О.)

Москва, 2020

Цель рабораторной работы

Изучение библиотеки обработки данных Pandas.

Задание

Выполнить набор заданий по датасету Adult, содержащему следующие признаки:

- `age` : число.
- `workclass` : Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- `fnlwgt` : число.
- `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` : число.
- `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` : число.
- `capital-loss` : число.
- `hours-per-week` : число.
- `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

Решение заданий

In [50]:
```
# импорт данных и библиотек
import pandas as pd
```

In [51]:
```
data = pd.read_csv('../input/adult.data.csv')
data.head()
```

Out[51]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	cap
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	

1. Сколько мужчин и женщин (признак *sex*) представлено в датасете?

In [52]:
```
data.groupby('sex').size()
```

Out[52]:
```
sex
Female    10771
Male      21790
dtype: int64
```

2. Какой средний возраст (признак *age*) женщин?

In [53]:
```
data[data['sex']=='Female']['age'].mean()
```

Out[53]:
```
36.85823043357163
```

3. Какой процент граждан Германии (признак *native-country*)?

In [54]:
```
data['native-country'].value_counts(normalize=True)['Germany'].round(4)*
100
```

Out[54]:
```
0.42
```

4-5. Какое среднее и стандартное отклонение возраста тех, кто кто зарабатывает больше и меньше 50K в год (признак *salary*)?

In [55]:
```
sal_50_less = data.loc[data['salary'] == '<=50K']
sal_50_more = data.loc[data['salary'] == '>50K']
print("Средний возраст богатых: {0} +- {1} лет, бедных - {2} +- {3} ле
т.".format(
   round(sal_50_more['age'].mean()), round(sal_50_more['age'].std(), 1),
   round(sal_50_less['age'].mean()), round(sal_50_less['age'].std(),
1)))
```

```
Средний возраст богатых: 44 +- 10.5 лет, бедных - 37 +- 14.0 лет.
```

6. Правда ли, что люди, зарабатывающие больше 50K, имеют хотя бы среднее образование? (признак *education – Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters* или *Doctorate*)

In [56]:
```
sal_50_more[sal_50_more['education'].isin(['Bachelors','Prof-school','Assoc-acdm', 'Assoc-voc', 'Masters','Doctorate'])].shape[0]/sal_50_more.shape[0] == 1
```

Out[56]:
```
False
```

7. Вывести возрастную статистику для каждой расы (признак *race*) и каждого гендера (признак *sex*). Использовать *groupby()* и *describe()*. Найти старшего и младшего мужчину *Amer-Indian-Eskimo* расы.

In [57]:
```
data.groupby(['race','sex'])['age'].describe()
```

Out[57]:

		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

In [58]:
```
data[(data['race']=='Amer-Indian-Eskimo')& (data['sex']=='Male')].groupby(['race','sex'])['age'].max()
```

Out[58]:
```
race                sex
Amer-Indian-Eskimo  Male    82
Name: age, dtype: int64
```

8. Среди кого больше тех, кто зарабатывает много (>50K): женатых или холостых (признак *marital-status*)? Считать женатыми тех, у кого *marital-status* начинается с *Married* (Married-civ-spouse, Married-spouse-absent or Married-AF-spouse), остальные считаются холостыми.

In [59]:
```
data[(data['sex']=='Male') & ~((data['marital-status'].str.startswith('Married')))]['salary'].value_counts()
```

Out[59]:
```
<=50K    7552
>50K      697
Name: salary, dtype: int64
```

In [60]:
```
data[(data['sex']=='Male') & (data['marital-status'].str.startswith('Married'))]['salary'].value_counts()
```

Out[60]:
```
<=50K    7576
>50K     5965
Name: salary, dtype: int64
```

9. Какое набольшое количество часов человек рабоет в неделю (признак *hours-per-week*)? Сколько человек работает такое количество часов и каков процент тех, кто зарабатывает много (>50K) среди них?

In [61]:
```
max_hours = data['hours-per-week'].max()
print(max_hours)
```

```
99
```

In [62]:
```
hard_workers = data[data['hours-per-week']==max_hours].shape[0]
print(hard_workers)
```

```
85
```

In [63]:
```
percent = data[(data['hours-per-week']==max_hours) & (data['salary']=='>
50K')].shape[0]/ hard_workers
print ("{0:.0%}".format(percent))
```

```
29%
```

10. Посчитать среднее рабочее время (*hours-per-week*) для тех, кто зарабатывает мало и много (*salary*) для каждой страны (*native-country*). Какие значения получатся для Японии?

In [64]:
```
data.groupby(['native-country','salary'])['hours-per-week'].mean().round
(2)
```

Out[64]:

native-country	salary	
?	<=50K	40.16
	>50K	45.55
Cambodia	<=50K	41.42
	>50K	40.00
Canada	<=50K	37.91
	>50K	45.64
China	<=50K	37.38
	>50K	38.90
Columbia	<=50K	38.68
	>50K	50.00
Cuba	<=50K	37.99
	>50K	42.44
Dominican-Republic	<=50K	42.34
	>50K	47.00
Ecuador	<=50K	38.04
	>50K	48.75
El-Salvador	<=50K	36.03
	>50K	45.00
England	<=50K	40.48
	>50K	44.53
France	<=50K	41.06
	>50K	50.75
Germany	<=50K	39.14
	>50K	44.98
Greece	<=50K	41.81
	>50K	50.62
Guatemala	<=50K	39.36
	>50K	36.67
Haiti	<=50K	36.33
	>50K	42.75
		...
Mexico	>50K	46.58
Nicaragua	<=50K	36.09
	>50K	37.50
Outlying-US(Guam-USVI-etc)	<=50K	41.86
Peru	<=50K	35.07
	>50K	40.00
Philippines	<=50K	38.07
	>50K	43.03
Poland	<=50K	38.17
	>50K	39.00
Portugal	<=50K	41.94
	>50K	41.50
Puerto-Rico	<=50K	38.47
	>50K	39.42
Scotland	<=50K	39.44
	>50K	46.67
South	<=50K	40.16
	>50K	51.44
Taiwan	<=50K	33.77
	>50K	46.80
Thailand	<=50K	42.87
	>50K	58.33
Trinadad&Tobago	<=50K	37.06
	>50K	40.00
United-States	<=50K	38.80
	>50K	45.51
Vietnam	<=50K	37.19
	>50K	39.20
Yugoslavia	<=50K	41.60
	>50K	49.50

Name: hours-per-week, Length: 82, dtype: float64

In [65]:
```
data[data['native-country']=='Japan'].groupby(['native-country','salary
'])['hours-per-week'].mean().round(2)
```

Out[65]:
```
native-country  salary
Japan           <=50K     41.00
                >50K      47.96
Name: hours-per-week, dtype: float64
```

Выводы

Были получены навыки фильтрации и обработки данных, представленных в виде датасетов, используя библиотеку Pandas.