Московский государственный технический университет

им. Н.Э. Баумана

(МГТУ им. Н.Э. Баумана)

Радиотехнический факультет (РТ)

Отчёт по лабораторной работе №2

По дисциплине

«Технологии машинного обучения»

Проверил:

Преподаватель кафедры ИУ-5

Гапанюк Ю.Е.

Подпись: _____

«___» _____ 2020 г.

Выполнил:

студент группы РТ5-61Б

Ануров Н.С.

Подпись: _____

«___» _____ 2020 г.

Москва, 2020

Задание:

Выполните первое демонстрационное задание "demo assignment" под названием "Exploratory data analysis with Pandas" со страницы курса https://mlcourse.ai/assignments

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

In [3]:
```
df=pd.read_csv("adult.data.csv")
df.head()
```

Out[3]:

	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

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

In [17]:
```
df['sex'].value_counts()
```

Out[17]:
```
Male      21790
Female    10771
Name: sex, dtype: int64
```

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

In [22]:
```
df.loc[df['sex']=='Female','age'].mean()
```

Out[22]:
```
36.85823043357163
```

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

In [28]:
```
round(len(df[df['native-country']=='Germany'])/len(df),4)
```

Out[28]:
```
0.0042
```

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?

In [50]:
```
round(df[df['salary']=='>50K'].age.mean(),3),round(df[df['salary']=='>50K'].age.std(),3)
```

Out[50]:
```
(44.25, 10.519)
```

In [51]:
```
round(df[df['salary']=='<=50K'].age.mean(),3),round(df[df['salary']=='<=50K'].age.std(),3)
```

Out[51]:
```
(36.784, 14.02)
```

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)

In [12]:
```
df[(df.salary=='>50K')&(df.education!='Bachelors')&(df.education!='Prof-school')&(df.education!='Assoc-acdm')&(df.education!='Ass
```

Out[12]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	salary
7	52	Self-emp-not-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	45	United-States	>50K
10	37	Private	280464	Some-college	10	Married-civ-spouse	Exec-managerial	Husband	Black	Male	0	0	80	United-States	>50K
27	54	?	180211	Some-college	10	Married-civ-spouse	?	Husband	Asian-Pac-Islander	Male	0	0	60	South	>50K
38	31	Private	84154	Some-college	10	Married-civ-spouse	Sales	Husband	White	Male	0	0	38	?	>50K
55	43	Private	237993	Some-college	10	Married-civ-spouse	Tech-support	Husband	White	Male	0	0	40	United-States	>50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32510	39	Private	107302	HS-grad	9	Married-	Prof-	Husband	White	Male	0	0	45	?	>50K

It is false

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.

In [71]:
```
df.groupby(['race','sex'])['age'].describe()
```

Out[71]:

race	sex	count	mean	std	min	25%	50%	75%	max
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

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.

In [24]:
```
len(df[(df['sex']=='Male') & (df['salary']=='>50K')&((df['marital-status']=='Married-civ-spouse')|(df['marital-status']=='Marriec
```

Out[24]: 5965

In [27]:
```
len(df[(df['sex']=='Male') & (df['salary']=='>50K')&((df['marital-status']!='Married-civ-spouse')&(df['marital-status']!='Marriec
```

Out[27]: 697

Married men are greater than single men

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?

In [78]:
```
df['hours-per-week'].max()
```

Out[78]: 99

In [91]:
```
df1=df[df['hours-per-week']==99]
l1=len(df1)
```

Out[91]: 85

In [93]:
```
l=len(df1[df1['salary']=='<=50K'])
round(l/l1,3)
```

Out[93]: 0.706

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?

In [96]:
```
df2=df.groupby(['native-country','salary'])['hours-per-week'].mean().reset_index()
df2.pivot(columns='salary',index='native-country',values='hours-per-week')
```

Out[96]:

salary	<=50K	>50K
native-country		
?	40.164760	45.547945
Cambodia	41.416667	40.000000
Canada	37.914634	45.641026
China	37.381818	38.900000
Columbia	38.684211	50.000000
Cuba	37.985714	42.440000
Dominican-Republic	42.338235	47.000000
Ecuador	38.041667	48.750000
El-Salvador	36.030928	45.000000
England	40.483333	44.533333
France	41.058824	50.750000
Germany	39.139785	44.977273
Greece	41.809524	50.625000
Guatemala	39.360656	36.666667