Dataset :

http://archive.ics.uci.edu/dataset/2/adult

Question: Do data analysis using Pandas and answer following questions?

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

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

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

4-5. What are mean value and standard deviation of the age of those who recieve more than 50K per year (salary feature) and those who receive less than 50K per year?

3.Is it true that people who receive more than 50k have at least high school education? (education - Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters or Doctorate feature

```
import pandas as pd
import numpy as np


features = ["Age", "Workclass", "fnlwgt", "Education", "Education-Num", "Martial Status",
             "Race", "Sex", "Capital Gain", "Capital Loss", "Hours per week", "Country", "

df = pd.read_csv('adult.data', names=features)
df
```

	Age	Workclass	fnlwgt	Education	Education-Num	Martial Status	Occu
0	39	State-gov	77516	Bachelors	13	Never-married	Adr
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	ma
2	38	Private	215646	HS-grad	9	Divorced	H
3	53	Private	234721	11th	7	Married-civ-spouse	H

https://rstudio-pubs-static.s3.amazonaws.com/538563_85cb2b4cd06b4dc48d33de73fa97a297.html

https://archive.ics.uci.edu/dataset/2/adult

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

```
#1 Question answer
df['Sex'].value_counts()
```

```
 Male      21790
 Female    10771
Name: Sex, dtype: int64
```

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

```
#2 Question answer
average_age_women = df.loc[df['Sex'].str.contains('Female'), 'Age'].mean()
print(average_age_women)
```

```
36.917076598735065
```

```
df[["Sex", "Age"]].groupby("Sex").mean()
```

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

```
df['Country'].value_counts()
```

```
 United-States                 29170
 Mexico                          643
 ?                               583
 Philippines                     198
 Germany                         137
 Canada                          121
 Puerto-Rico                     114
 El-Salvador                     106
 India                           100
 Cuba                             95
 England                          90
 Jamaica                          81
 South                            80
 China                            75
 Italy                            73
 Dominican-Republic               70
 Vietnam                          67
 Guatemala                        64
 Japan                            62
 Poland                           60
 Columbia                         59
 Taiwan                           51
 Haiti                            44
 Iran                             43
 Portugal                         37
 Nicaragua                        34
 Peru                             31
 France                           29
 Greece                           29
 Ecuador                          28
 Ireland                          24
 Hong                             20
 Cambodia                         19
 Trinadad&Tobago                  19
 Laos                             18
 Thailand                         18
 Yugoslavia                       16
 Outlying-US(Guam-USVI-etc)       14
 Honduras                         13
 Hungary                          13
 Scotland                         12
 Holand-Netherlands                1
Name: Country, dtype: int64
```

```
#3 Question answer
df[df['Country'].str.contains('Germany')] ['Country'].value_counts()/len(df)*100
```

```
 Germany    0.420749
Name: Country, dtype: float64
```

```
country_germany = df[df['Country'].str.contains('Germany')]
```

```
country_germany.describe()
```

	Age	fnlwgt	Education-Num	Capital Gain	Capital Loss
count	137.000000	137.000000	137.000000	137.000000	137.000000
mean	39.255474	189325.313869	10.985401	887.094891	77.978102
std	12.962065	100809.067728	2.370112	3627.385181	371.502899
min	18.000000	21306.000000	4.000000	0.000000	0.000000
25%	29.000000	116391.000000	9.000000	0.000000	0.000000
50%	36.000000	178322.000000	10.000000	0.000000	0.000000
75%	47.000000	231604.000000	13.000000	0.000000	0.000000
max	74.000000	606111.000000	16.000000	27828.000000	1977.000000

4-5. What are mean value and standard deviation of the age of those who recieve more than 50K per year (salary feature) and those who receive less than 50K per year?

```
#4-5 Question answer
age_more50k= df[df['Target'].str.contains('>50K')]['Age'].mean()
print("Mean value of Age who is having Target >50K:",age_more50k)
```

```
Mean value of Age who is having Target >50K: 44.24984058155847
```

```
age_more50k=df[df['Target'].str.contains('>50K')]['Age'].std()
print("Std value of Age who is having Target >50K:",age_more50k)
```

```
Std value of Age who is having Target >50K: 10.519027719851826
```

```
age_less50k=df[df['Target'].str.contains('<=50K')]['Age'].mean()
print("Mean value of Age who is having Target <=50K:",age_less50k)
```

```
Mean value of Age who is having Target <=50K: 36.78373786407767
```

```
age_less50k=df[df['Target'].str.contains('<=50K')]['Age'].std()
print("Std value of Age who is having Target <=50K:",age_less50k)
```

```
Std value of Age who is having Target <=50K: 14.02008849082488
```

6.Is it true that people who receive more than 50k have at least high school education? (education - Bachelors, Prof-school, Assoc-acdm, Assoc-voc, Masters or Doctorate feature

```
#6 Question answer
df[df['Target'].str.contains('>50K')] ['Education'].unique()
```

```
array([' HS-grad', ' Masters', ' Bachelors', ' Some-college',
       ' Assoc-voc', ' Doctorate', ' Prof-school', ' Assoc-acdm',
       ' 7th-8th', ' 12th', ' 10th', ' 11th', ' 9th', ' 5th-6th',
       ' 1st-4th'], dtype=object)
```