In [16]:

```
import pandas as pd
import numpy as np
from google.colab import drive
drive.mount('/content/drive')
from google.colab import drive
import os
import pandas as pd
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_rem
ount=True).
```

In [70]:

```
path = "/content/drive/MyDrive/Jntu/Assignment4/"
fnames = os.listdir(path)
print(fnames)

df = pd.read_csv(os.path.join(path, fnames[1]))
df
```

```
['Adult.ipynb', 'adult.data', 'adult.names', 'Assignment 4.txt']
```

Out[70]:

	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
0	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
1	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
2	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
3	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
4	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32555	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
32556	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
32557	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
32558	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
32559	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

32560 rows × 15 columns

In [133]:

```
features = ["Age", "Workclass", "fnlwgt", "Education", "Education-Num", "Martial Status", "Occupation", "Relationshi
p",
             "Race", "Sex", "Capital Gain", "Capital Loss", "Hours per week", "Country", "Target"]

df = pd.read_csv(os.path.join(path, fnames[1]), names=features)
#/content/drive/MyDrive/my_data/data.csv
df
```

Out[133]:

	Age	Workclass	fnlwgt	Education	Education-Num	Martial Status	Occupation	Relationship	Race	Sex	Capital Gain	Capital Loss	Hours per week	Country	Target
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
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	40	United-States	>50K

32561 rows × 15 columns

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

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

In [30]:

```
df["Sex"].value_counts()
```

Out[30]:

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

In [39]:

```
df.loc[df['Sex'] == 'Female', 'Age'].mean()

#What is the average age (age feature) of women? 38.58164675532078
```

Out[39]:

```
38.58164675532078
```

In [69]:

```
df['Country'].value_counts()
get_count =  137/32561
print(get_count)
# What is the proportion of German citizens (native-country feature)?  0.004207487485028101
```

```
0.004207487485028101
```

In [104]:

```
   age1 = df.loc[df['Target'] ==  '>50K','Age'].std()

print(age1)
```

```
13.640432553581341
```

In [131]:

```
df['Target']  ==  '<=50K'
```

Out[131]:

```
0        False
1        False
2        False
3        False
4        False
         ...
32556    False
32557    False
32558    False
32559    False
32560    False
Name: Target, Length: 32561, dtype: bool
```

In [156]:

```
df.loc[df['Target'] == '>50K', 'Education'].unique()

 #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)
 #No
```

Out[156]:

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