2306 AML117 MAHESH BABU M ASSIGNMENT4

In [1]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
features = ["Age", "Workclass", "fnlwgt", "Education", "Education-Num", "Martial Status", "Occupation", "Relationship",
            "Race", "Sex", "Capital Gain", "Capital Loss", "Hours per week", "Country", "Target"]

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

Out[2]:

	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 [3]:
```
new_df=df
new_df.head()
```

Out[3]:

	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

In [4]:
```
new_df.columns
```

Out[4]:
```
Index(['Age', 'Workclass', 'fnlwgt', 'Education', 'Education-Num',
       'Martial Status', 'Occupation', 'Relationship', 'Race', 'Sex',
       'Capital Gain', 'Capital Loss', 'Hours per week', 'Country', 'Target'],
      dtype='object')
```

In [5]:
```
new_df['Sex'].value_counts()
```

Out[5]:
```
 Male      21790
 Female    10771
Name: Sex, dtype: int64
```

Q1 ; ANS; data includes males 21790 and female 10771

In [6]:
```
new_df.pivot_table(['Age'],['Sex'])
```

Out[6]:

Sex	Age
Female	36.858230
Male	39.433547

In [7]:
```
group_data = new_df.groupby('Sex')
```

In [8]:
```
group_data.describe()
# Provide the mean for each numeric column by sex
group_data.mean(numeric_only=True)
```

Out[8]:

Sex	Age	fnlwgt	Education-Num	Capital Gain	Capital Loss	Hours per week
Female	36.858230	185746.311206	10.035744	568.410547	61.187633	36.410361
Male	39.433547	191771.449013	10.102891	1329.370078	100.213309	42.428086

Q2 average age of female is 36.85

In [58]:
```
new_df.columns
```

Out[58]:
```
Index(['Age', 'Workclass', 'fnlwgt', 'Education', 'Education-Num',
       'Martial Status', 'Occupation', 'Relationship', 'Race', 'Sex',
       'Capital Gain', 'Capital Loss', 'Hours per week', 'Country', 'Target'],
      dtype='object')
```

In [59]:
```
new_df['Country'].value_counts()
```

Out[59]:
```
 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
```

Germany=137 ratio of total to german is

In [11]:
```
137/32561
```

Out[11]:
```
0.004207487485028101
```

Q3 0.004207487485028101 is proportion of german citizens

In [ ]:

In [12]:
```
new_df.pivot_table(['Age'],['Target'])
```

Out[12]:

Target	Age
<=50K	36.783738
>50K	44.249841

In [13]:
```
salary=new_df.groupby(['Target'])
```

In [14]:
```
salary.describe()
```

Out[14]:

	Age								fnlwgt		...	Capital Loss		Hours per week							
	count	mean	std	min	25%	50%	75%	max	count	mean	...	75%	max	count	mean	std	min	25%	50%	75%	max
Target																					
<=50K	24720.0	36.783738	14.020088	17.0	25.0	34.0	46.0	90.0	24720.0	190340.86517	...	0.0	4356.0	24720.0	38.840210	12.318995	1.0	35.0	40.0	40.0	99.0
>50K	7841.0	44.249841	10.519028	19.0	36.0	44.0	51.0	90.0	7841.0	188005.00000	...	0.0	3683.0	7841.0	45.473026	11.012971	1.0	40.0	40.0	50.0	99.0

2 rows × 48 columns

Q 4-5 THE MEAN =36.78, sd=14 for who's salary is <= 50K, mean=44.2 and sd=10.5 for whose salary is >= 50K

In [15]:
```
education=new_df.groupby('Education')
```

In [16]:
```
education.describe()
```

Out[16]:

	Age								fnlwgt		...	Capital Loss		Hours per week							
	count	mean	std	min	25%	50%	75%	max	count	mean	...	75%	max	count	mean	std	min	25%	50%	75%	max
Education																					
10th	933.0	37.429796	16.720713	17.0	22.00	34.0	52.0	90.0	933.0	196832.465166	...	0.0	3770.0	933.0	37.052519	13.788112	1.0	30.0	40.0	40.0	99.0
11th	1175.0	32.355745	15.545485	17.0	18.00	28.0	43.0	90.0	1175.0	194928.077447	...	0.0	2824.0	1175.0	33.925957	13.965416	2.0	20.0	40.0	40.0	99.0
12th	433.0	32.000000	14.334625	17.0	19.00	28.0	41.0	79.0	433.0	199097.508083	...	0.0	2258.0	433.0	35.780600	12.626412	6.0	30.0	40.0	40.0	99.0
1st-4th	168.0	46.142857	15.615625	19.0	33.00	46.0	57.0	90.0	168.0	239303.000000	...	0.0	2603.0	168.0	38.255952	12.848727	4.0	35.0	40.0	40.0	96.0
5th-6th	333.0	42.885886	15.557285	17.0	29.00	42.0	54.0	84.0	333.0	232448.333333	...	0.0	2603.0	333.0	38.897898	10.551727	3.0	40.0	40.0	40.0	84.0
7th-8th	646.0	48.445820	16.092350	17.0	34.25	50.0	61.0	90.0	646.0	188079.171827	...	0.0	3900.0	646.0	39.366873	14.201870	2.0	35.0	40.0	40.0	99.0
9th	514.0	41.060311	15.946862	17.0	28.00	39.0	54.0	90.0	514.0	202485.066148	...	0.0	2231.0	514.0	38.044747	11.064402	1.0	36.0	40.0	40.0	99.0
Assoc-acdm	1067.0	37.381443	11.095177	19.0	29.00	36.0	44.0	90.0	1067.0	193424.093721	...	0.0	2824.0	1067.0	40.504217	12.196666	1.0	40.0	40.0	45.0	99.0
Assoc-voc	1382.0	38.553546	11.631300	19.0	30.00	37.0	46.0	84.0	1382.0	181936.016643	...	0.0	2603.0	1382.0	41.610709	10.793384	1.0	40.0	40.0	45.0	99.0
Bachelors	5355.0	38.904949	11.912210	19.0	29.00	37.0	46.0	90.0	5355.0	188055.914846	...	0.0	2824.0	5355.0	42.614006	11.446185	2.0	40.0	40.0	50.0	99.0
Doctorate	413.0	47.702179	11.784716	24.0	39.00	47.0	55.0	80.0	413.0	186698.760291	...	0.0	3683.0	413.0	46.973366	15.084447	1.0	40.0	45.0	55.0	99.0
HS-grad	10501.0	38.974479	13.541524	17.0	28.00	37.0	48.0	90.0	10501.0	189538.739739	...	0.0	4356.0	10501.0	40.575374	11.333757	1.0	40.0	40.0	42.0	99.0
Masters	1723.0	44.049913	11.068935	18.0	36.00	43.0	51.0	90.0	1723.0	179852.362739	...	0.0	2824.0	1723.0	43.836332	12.277801	1.0	40.0	40.0	50.0	99.0
Preschool	51.0	42.764706	15.126914	19.0	31.00	41.0	53.5	75.0	51.0	235889.372549	...	0.0	1719.0	51.0	36.647059	12.555196	10.0	30.0	40.0	40.0	75.0
Prof-school	576.0	44.746528	11.962477	25.0	36.00	43.0	51.0	90.0	576.0	185663.706597	...	0.0	2824.0	576.0	47.425347	14.806038	2.0	40.0	48.0	55.0	99.0
Some-college	7291.0	35.756275	13.474051	17.0	24.00	34.0	45.0	90.0	7291.0	188742.922370	...	0.0	4356.0	7291.0	38.852284	12.761901	1.0	35.0	40.0	43.0	99.0

16 rows × 48 columns

In [42]:
```
new_df[new_df['Target'] == '>50K'].value_counts()
```

Out[42]:
```
Series([], dtype: int64)
```

In [63]:
```
data=new_df['Education'].value_counts()
print(data)
```

```
 HS-grad         10501
 Some-college     7291
 Bachelors        5355
 Masters          1723
 Assoc-voc        1382
 11th             1175
 Assoc-acdm       1067
 10th              933
 7th-8th           646
 Prof-school       576
 9th               514
 12th              433
 Doctorate         413
 5th-6th           333
 1st-4th           168
 Preschool          51
Name: Education, dtype: int64
```

In [68]:
```
ntr=new_df.pivot_table(index='Education', columns='Target')
print(ntr)
```

```
                   Age            Capital Gain               Capital Loss  \
Target           <=50K       >50K        <=50K          >50K        <=50K
Education
 10th        36.510907  50.338710   135.154994   4189.483871    37.167623
 11th        31.765022  43.333333    85.787444   2618.116667    41.613453
 12th        30.952500  44.696970    79.075000   2769.090909    30.385000
 1st-4th     45.925926  52.000000    83.080247   1281.333333    50.117284
 5th-6th     42.690852  46.750000    99.504732   1692.000000    65.744479
 7th-8th     48.163366  52.725000   173.485149   1149.825000    60.488449
 9th         40.630390  48.814815   134.453799   4087.222222    22.229979
 Assoc-acdm  35.962594  41.675472   110.389027   2244.430189    64.000000
 Assoc-voc   37.142018  42.545706   196.825661   2180.725762    47.919687
 Bachelors   36.273452  42.618190   162.901723   4004.705088    59.712189
 Doctorate   45.158879  48.591503   220.485981   6361.039216    46.065421
 HS-grad     37.826762  45.022090   153.878541   2805.281194    53.642987
 Masters     42.291885  45.450469   285.774869   4376.397289    79.301047
 Preschool   42.764706        NaN   898.392157           NaN    66.490196
 Prof-school 43.718954  45.118203   186.954248  14113.711584    93.385621
 Some-college 33.778794  44.173756   125.684621   2612.822639    51.200203

                         Education-Num       Hours per week             \
Target             >50K          <=50K  >50K          <=50K       >50K
Education
 10th        333.290323           6.0   6.0      36.574053  43.774194
 11th        207.400000           7.0   7.0      33.322870  45.133333
 12th         56.000000           8.0   8.0      35.035000  44.818182
 1st-4th       0.000000           2.0   2.0      37.864198  48.833333
 5th-6th     117.937500           3.0   3.0      38.539432  46.000000
 7th-8th     144.150000           4.0   4.0      38.830033  47.500000
 9th         151.074074           5.0   5.0      37.667351  44.851852
 Assoc-acdm  182.452830          12.0  12.0      39.264339  44.256604
 Assoc-voc   142.994460          11.0  11.0      40.817826  43.853186
 Bachelors   201.093201          13.0  13.0      40.586152  45.475462
 Doctorate   338.647059          16.0  16.0      45.429907  47.513072
 HS-grad     159.114627           9.0   9.0      39.727510  45.042985
 Masters     236.362878          14.0  14.0      41.223822  45.917623
 Preschool          NaN           1.0   NaN      36.647059        NaN
 Prof-school 281.052009          15.0  15.0      42.816993  49.092199
 Some-college 158.630137          10.0  10.0      37.450203  44.820476

                   fnlwgt
Target              <=50K           >50K
Education
 10th        196755.117107  197919.080645
 11th        196349.593722  168511.566667
 12th        198153.130000  210544.515152
 1st-4th     239657.265432  229737.833333
 5th-6th     235683.451104  168352.562500
 7th-8th     188927.740924  175223.350000
 9th         204506.583162  166022.888889
 Assoc-acdm  194091.038653  191405.641509
 Assoc-voc   181821.534770  182259.800554
 Bachelors   186724.088066  189935.223773
 Doctorate   179209.542056  189317.539216
 HS-grad     190402.221391  184988.835821
 Masters     177683.837696  181579.946820
 Preschool   235889.372549            NaN
 Prof-school 187208.300654  185105.023641
 Some-college 187238.682927  195145.971882
```

```
C:\Users\sys\AppData\Local\Temp\ipykernel_7008\925973283.py:1: FutureWarning: pivot_table dropped a column because it failed to aggregate. This behavior is de
precated and will raise in a future version of pandas. Select only the columns that can be aggregated.
  ntr=new_df.pivot_table(index='Education', columns='Target')
```

In [69]:
```
ntr.loc['Assoc-acdm':'Some-college']
```

Out[69]:

	Age		Capital Gain		Capital Loss		Education-Num		Hours per week		fnlwgt	
Target	<=50K	>50K	<=50K	>50K	<=50K	>50K	<=50K	>50K	<=50K	>50K	<=50K	>50K
Education												

In [70]:
```
ntr.describe()
```

Out[70]:

	Age		Capital Gain		Capital Loss		Education-Num		Hours per week		fnlwgt	
Target	<=50K	>50K	<=50K	>50K	<=50K	>50K	<=50K	>50K	<=50K	>50K	<=50K	>50K
count	16.000000	15.000000	16.000000	15.000000	16.000000	15.000000	16.000000	15.000000	16.000000	15.000000	16.000000	15.000000
mean	39.472313	46.256948	195.739262	3765.745664	54.341476	180.679940	8.500000	9.000000	38.862297	45.792167	198768.843763	187736.652027
std	5.176286	3.489900	195.467968	3176.176920	17.860382	92.296814	4.760952	4.472136	2.967110	1.693233	20282.085740	16732.060937
min	30.952500	41.675472	79.075000	1149.825000	22.229979	0.000000	1.000000	2.000000	33.322870	43.774194	177683.837696	166022.888889
25%	36.195738	43.753545	107.667954	2212.577975	44.952429	143.572230	4.750000	5.500000	37.249417	44.819329	187087.247507	178401.648410
50%	39.228576	45.118203	144.516767	2769.090909	52.421595	159.114627	8.500000	9.000000	38.684733	45.133333	192246.630022	185105.023641
75%	43.003268	48.703159	189.422102	4138.353047	64.436120	221.881439	12.250000	12.500000	40.644070	46.750000	199741.493291	193275.806696
max	48.163366	52.725000	898.392157	14113.711584	93.385621	338.647059	16.000000	16.000000	45.429907	49.092199	239657.265432	229737.833333

Q6

it is not true that who are education lessthan high school are takingthe more than 50k . at least high school criteria is not good enough to take >50k

In [40]:

```
Empty DataFrame
Columns: [Age, Workclass, fnlwgt, Education, Education-Num, Martial Status, Occupation, Relationship, Race, Sex, Capital Gain, Capital Loss, Hours per week, C
ountry, Target]
Index: []
```

In [ ]: