```
import sklearn as sk
import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import OrdinalEncoder
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
```

```
pd.set_option('display.max_columns', 50)
```

```
/usr/local/lib/python3.10/dist-packages/jupyter_client/jsonutil.py:67: DeprecationWarning: Ipykernel 7.0 has removed support for Python 3.8. Please upgrade your Python to at least Python 3.9 or downgrade your ipykernel to 6.29.5.
  pd.set_option('display.max_columns', 50)
```

```
df = pd.read_csv("Data for Clustering & Modeling.csv")
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             32561 non-null  int64
 1   workclass       32561 non-null  object
 2   fnlwgt          32561 non-null  int64
 3   education       32561 non-null  object
 4   education-num   32561 non-null  int64
 5   marital-status  32561 non-null  object
 6   occupation      32561 non-null  object
 7   relationship    32561 non-null  object
 8   race            32561 non-null  object
 9   sex             32561 non-null  object
 10  capital-gain    32561 non-null  int64
 11  capital-loss    32561 non-null  int64
 12  hours-per-week  32561 non-null  int64
 13  native-country  32561 non-null  object
 14  income          32561 non-null  object
dtypes: int64(6), object(9)
memory usage: 3.7 MB
```

```
df.shape
```

```
(32561, 15)
```

```
df.isnull().sum()
```

```
age               0
workclass         0
fnlwgt            0
education         0
education-num     0
marital-status    0
occupation        0
relationship      0
race              0
sex               0
capital-gain      0
capital-loss      0
hours-per-week    0
native-country    0
income            0
dtype: int64
```

```
df.dtypes
```

```
age                int64
workclass         object
fnlwgt            int64
education         object
education-num     int64
marital-status    object
occupation        object
relationship      object
race              object
sex               object
capital-gain      int64
capital-loss      int64
hours-per-week    int64
native-country    object
income            object
dtype: int64
```

```
df.nunique()
```

```
age                 73
workclass            9
fnlwgt           21648
education           16
education-num       16
marital-status       7
occupation          15
relationship         6
race                 5
sex                  2
capital-gain       119
capital-loss        92
hours-per-week      94
native-country      42
income               2
dtype: int64
```

```
df.sample(5)
```

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
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
5	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States	<=50K

```
df.describe()
```

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
count	32561.000000	32561.000000	32561.000000	32561.000000	32561.000000	32561.000000
mean	38.581647	189778.366512	10.080679	1077.648844	87.303830	40.437456
std	13.640433	105549.977697	2.572720	7385.292085	402.960219	12.347429
min	17.000000	12285.000000	1.000000	0.000000	0.000000	1.000000
25%	28.000000	117827.000000	9.000000	0.000000	0.000000	40.000000
50%	37.000000	178356.000000	10.000000	0.000000	0.000000	40.000000
75%	48.000000	237051.000000	12.000000	0.000000	0.000000	45.000000
max	90.000000	1484705.000000	16.000000	99999.000000	4356.000000	99.000000

```
df.describe(include='object')
```

	workclass	education	marital-status	occupation	relationship	race	sex	native-country	income
count	32561	32561	32561	32561	32561	32561	32561	32561	32561
unique	9	16	7	15	6	5	2	42	2
top	Private	HS-grad	Married-civ-spouse	Prof-specialty	Husband	White	Male	United-States	<=50K
freq	22696	10501	14976	4140	13193	27816	21790	29170	24720

```
df.drop(columns=['income'], inplace=True)
```

```
df.shape
```

```
(32561, 14)
```

```
df.sample(5)
```

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba
5	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States

```
df.dtypes
```

```
age                int64
workclass         object
fnlwgt            int64
education         object
education-num     int64
marital-status    object
occupation        object
relationship      object
race              object
sex               object
capital-gain      int64
capital-loss      int64
hours-per-week    int64
native-country    object
dtype: int64
```

```
df.nunique()
```

```
age                 73
workclass            9
fnlwgt           21648
education           16
education-num       16
marital-status       7
occupation          15
relationship         6
race                 5
sex                  2
capital-gain       119
capital-loss        92
hours-per-week      94
native-country      42
dtype: int64
```

```
df.isnull().sum
```

```
<bound_method> NDFrame.sum of age               0
workclass         0
fnlwgt            0
education         0
education-num     0
marital-status    0
occupation        0
relationship      0
race              0
sex               0
capital-gain      0
capital-loss      0
hours-per-week    0
native-country    0
dtype: int64>
```

```
df.describe
```

```
<bound_method> NDFrame.describe of       age         workclass  fnlwgt  education  education-num  \
0      39         State-gov   77516  Bachelors             13
1      50  Self-emp-not-inc   83311  Bachelors             13
2      38           Private  215646    HS-grad              9
3      53           Private  234721       11th              7
4      28           Private  338409  Bachelors             13
...   ...               ...     ...        ...            ...
32556  27           Private  257302 Assoc-acdm             12
32557  40           Private  154374    HS-grad              9
32558  58           Private  151910    HS-grad              9
32559  22           Private  201490    HS-grad              9
32560  52      Self-emp-inc  287927    HS-grad              9

       marital-status         occupation   relationship   race   sex  \
0       Never-married       Adm-clerical  Not-in-family  White  Male
1  Married-civ-spouse    Exec-managerial        Husband  White  Male
2            Divorced  Handlers-cleaners  Not-in-family  White  Male
3  Married-civ-spouse  Handlers-cleaners        Husband  Black  Male
4  Married-civ-spouse     Prof-specialty           Wife  Black  Male
...               ...                ...            ...    ...    ...
32556  Married-civ-spouse       Tech-support           Wife  White  Female
32557  Married-civ-spouse  Machine-op-inspct        Husband  White    Male
32558             Widowed       Adm-clerical      Unmarried  White  Female
32559       Never-married       Adm-clerical      Unmarried  White    Male
32560  Married-civ-spouse    Exec-managerial           Wife  White  Female

       capital-gain  capital-loss  hours-per-week  native-country  income
0          2174             0              40   United-States   <=50K
1             0             0              13   United-States   <=50K
2             0             0              40   United-States   <=50K
3             0             0              40   United-States   <=50K
4             0             0              40            Cuba   >50K
...         ...           ...             ...             ...     ...
32556         0             0              38   United-States   <=50K
32557         0             0              40   United-States   >50K
32558         0             0              40   United-States   <=50K
32559         0             0              20   United-States   <=50K
32560     15024             0              40   United-States   >50K

[32561 rows x 15 columns]>
```

```
df.describe()
```

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
count	32561.000000	32561.000000	32561.000000	32561.000000	32561.000000	32561.000000
mean	38.581647	189778.366512	10.080679	1077.648844	87.303830	40.437456
std	13.640433	105549.977697	2.572720	7385.292085	402.960219	12.347429
min	17.000000	12285.000000	1.000000	0.000000	0.000000	1.000000
25%	28.000000	117827.000000	9.000000	0.000000	0.000000	40.000000
50%	37.000000	178356.000000	10.000000	0.000000	0.000000	40.000000
75%	48.000000	237051.000000	12.000000	0.000000	0.000000	45.000000
max	90.000000	1484705.000000	16.000000	99999.000000	4356.000000	99.000000

```
df.describe(include='object')
```

	workclass	education	marital-status	occupation	relationship	race	sex	native-country
count	32561	32561	32561	32561	32561	32561	32561	32561
unique	9	16	7	15	6	5	2	42
top	Private	HS-grad	Married-civ-spouse	Prof-specialty	Husband	White	Male	United-States
freq	22696	10501	14976	4140	13193	27816	21790	29170

```
df.drop(columns=['fnlwgt'], inplace=True)
```

```
df.shape
```

```
(32561, 13)
```

```
df.sample(5)
```

	age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country
1	50	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States
2	38	Private	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States
3	53	Private	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States
4	28	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba
5	37	Private	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States

```
df.dtypes
```

```
age                int64
workclass         object
education         object
education-num     int64
marital-status    object
occupation        object
relationship      object
race              object
sex               object
capital-gain      int64
capital-loss      int64
hours-per-week    int64
native-country    object
dtype: int64
```

```
df.nunique()
```

```
age                 73
workclass            9
education           16
education-num       16
marital-status       7
occupation          15
relationship         6
race                 5
sex                  2
capital-gain       119
capital-loss        92
hours-per-week      94
native-country      42
dtype: int64
```

```
df['workclass'].value_counts
```

```
workclass
Private             22696
Self-emp-not-inc     2541
Local-gov            2093
?                    1836
State-gov            1298
Self-emp-inc         1116
Federal-gov           960
Without-pay            14
Never-worked            7
Name: count, dtype: int64
```

```
df.replace('?', np.nan, inplace=True)
```

```
df['workclass'].value_counts
```

```
workclass
Private             22696
Self-emp-not-inc     2541
Local-gov            2093
NaN                  1836
State-gov            1298
Self-emp-inc         1116
Federal-gov           960
Without-pay            14
Never-worked            7
Name: count, dtype: int64
```

```
df.isnull().sum()
```

```
age               0
workclass      1836
education         0
education-num     0
marital-status    0
occupation        0
relationship      0
race              0
sex               0
capital-gain      0
capital-loss      0
hours-per-week    0
native-country    0
dtype: int64
```

```
df.dtypes
```

```
age                int64
workclass         object
education         object
education-num     int64
marital-status    object
occupation        object
relationship      object
race              object
sex               object
capital-gain      int64
capital-loss      int64
hours-per-week    int64
native-country    object
dtype: int64
```

```
df.nunique()
```

```
age                 73
workclass            8
education           16
education-num       16
marital-status       7
occupation          15
relationship         6
race                 5
sex                  2
capital-gain       119
capital-loss        92
hours-per-week      94
native-country      42
dtype: int64
```

```
df['workclass'].value_counts
```

```
workclass
Private             22696
Self-emp-not-inc     2541
Local-gov            2093
NaN                  1836
State-gov            1298
Self-emp-inc         1116
Federal-gov           960
Without-pay            14
Never-worked            7
Name: count, dtype: int64
```

```
df['workclass'].value_counts(normalize=1)
```

```
workclass
Private             0.697030
Self-emp-not-inc    0.078048
Local-gov           0.064279
NaN                 0.056391
State-gov           0.039864
Self-emp-inc        0.034274
Federal-gov         0.029483
Without-pay         0.000430
Never-worked        0.000215
Name: proportion, dtype: int64
```

```
df['workclass'].value_counts(normalize=1).sum()
```

```
1.0
```

```
df['workclass'].fillna('Unknown', inplace=True)
```

```
df['workclass'].value_counts(normalize=1)
```

```
workclass
Private             0.697030
Self-emp-not-inc    0.078048
Local-gov           0.064279
Unknown             0.056391
State-gov           0.039864
Self-emp-inc        0.034274
Federal-gov         0.029483
Without-pay         0.000430
Never-worked        0.000215
Name: proportion, dtype: int64
```

```
df['workclass'].value_counts(normalize=1).sum()
```

```
1.0
```

```
df.isnull().sum()
```

```
age               0
workclass         0
education         0
education-num     0
marital-status    0
occupation        0
relationship      0
race              0
sex               0
capital-gain      0
capital-loss      0
hours-per-week    0
native-country    0
dtype: int64
```

```
df.dtypes
```

```
age                int64
workclass         object
education         object
education-num     int64
marital-status    object
occupation        object
relationship      object
race              object
sex               object
capital-gain      int64
capital-loss      int64
hours-per-week    int64
native-country    object
dtype: int64
```

```
df.nunique()
```

```
age                 73
workclass            9
education           16
education-num       16
marital-status       7
occupation          15
relationship         6
race                 5
sex                  2
capital-gain       119
capital-loss        92
hours-per-week      94
native-country      42
dtype: int64
```

```
df['workclass'].value_counts
```

```
workclass
Private             22696
Self-emp-not-inc     2541
Local-gov            2093
Unknown              1836
State-gov            1298
Self-emp-inc         1116
Federal-gov           960
Without-pay            14
Never-worked            7
Name: count, dtype: int64
```

```
df['education'].value_counts
```

```
education
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: count, dtype: int64
```

```
df['education'].value_counts(normalize=1)
```

```
education
HS-grad      0.322502
Some-college 0.223918
Bachelors    0.164461
Masters      0.052916
Assoc-voc    0.042443
11th         0.036086
Assoc-acdm   0.032769
10th         0.028654
7th-8th      0.019840
Prof-school  0.017690
9th          0.015786
12th         0.013298
Doctorate    0.012684
5th-6th      0.010227
1st-4th      0.005160
Preschool    0.001567
Name: proportion, dtype: int64
```

```
df['education'].value_counts(normalize=1).sum()
```

```
1.0
```

```
df['education-num'].value_counts
```

```
education-num
9     10501
10     7291
13     5355
14     1723
11     1382
7      1175
12     1067
6       933
8       646
15      576
5       514
4       433
16      413
3       333
2       168
1        51
Name: count, dtype: int64
```

```
df['education-num'].value_counts(normalize=1)
```

```
education-num
9     0.322502
10    0.223918
13    0.164461
14    0.052916
11    0.042443
7     0.036086
12    0.032769
6     0.028654
8     0.019840
15    0.017690
5     0.015786
4     0.013298
16    0.012684
3     0.010227
2     0.005160
1     0.001567
Name: proportion, dtype: int64
```

```
df['education-num'].value_counts(normalize=1).sum()
```

```
1.0
```

```
df['marital-status'].value_counts
```

```
marital-status
Married-civ-spouse    14976
Never-married         10683
Divorced               4443
Separated              1025
Widowed                993
Married-spouse-absent  418
Married-AF-spouse       23
Name: count, dtype: int64
```

```
df['marital-status'].value_counts(normalize=1)
```

```
marital-status
Married-civ-spouse    0.459937
Never-married         0.328062
Divorced              0.136452
Separated             0.031479
Widowed               0.030497
Married-spouse-absent 0.012837
Married-AF-spouse     0.000706
Name: proportion, dtype: int64
```

```
df['marital-status'].value_counts(normalize=1).sum()
```

```
1.0
```

```
df['occupation'].value_counts
```

```
occupation
Prof-specialty       4140
Exec-managerial      4066
Craft-repair         4099
Adm-clerical         3770
Sales                3650
Other-service        3295
Machine-op-inspct    2002
Transport-moving     1597
Handlers-cleaners    1370
Unknown              1843
Farming-fishing      994
Tech-support         928
Protective-serv      649
Priv-house-serv      149
Armed-Forces           9
Name: count, dtype: int64
```

```
df['occupation'].value_counts(normalize=1)
```

```
occupation
Prof-specialty       0.127146
Exec-managerial      0.124873
Craft-repair         0.125891
Adm-clerical         0.115783
Sales                0.112092
Other-service        0.101200
Machine-op-inspct    0.061483
Transport-moving     0.049050
Handlers-cleaners    0.042075
Unknown              0.056605
Farming-fishing      0.030526
Tech-support         0.028497
Protective-serv      0.019929
Priv-house-serv      0.004576
Armed-Forces         0.000276
Name: proportion, dtype: int64
```

```
df['occupation'].value_counts(normalize=1).sum()
```

```
1.0
```

```
df['relationship'].value_counts
```

```
relationship
Husband        13193
Not-in-family   8305
Own-child       5068
Unmarried       3446
Wife            1568
Other-relative   981
Name: count, dtype: int64
```

```
df['relationship'].value_counts(normalize=1)
```

```
relationship
Husband        0.405178
Not-in-family  0.255060
Own-child      0.155646
Unmarried      0.105835
Wife           0.048156
Other-relative 0.030135
Name: proportion, dtype: int64
```

```
df['relationship'].value_counts(normalize=1).sum()
```

```
1.0
```

```
df['race'].value_counts
```

```
race
White                 27816
Black                  3124
Asian-Pac-Islander     1039
Amer-Indian-Eskimo      311
Other                   271
Name: count, dtype: int64
```

```
df['race'].value_counts(normalize=1)
```

```
race
White                 0.854274
Black                 0.095943
Asian-Pac-Islander    0.031909
Amer-Indian-Eskimo    0.009551
Other                 0.008323
Name: proportion, dtype: int64
```

```
df['race'].value_counts(normalize=1).sum()
```

```
1.0
```

```
df['sex'].value_counts
```

```
sex
Male    21790
Female  10771
Name: count, dtype: int64
```

```
df['sex'].value_counts(normalize=1)
```

```
sex
Male    0.669205
Female  0.330795
Name: proportion, dtype: int64
```

```
df['sex'].value_counts(normalize=1).sum()
```

```
1.0
```

```
df['native-country'].value_counts
```

```
native-country
United-States                 29170
Mexico                          643
NaN                             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
Greece                           29
France                           27
Ecuador                          28
Ireland                          24
Hong                             20
Cambodia                         19
Trinidad&Tobago                  19
Laos                             18
Thailand                         18
Yugoslavia                       16
Outlying-US(Guam-USVI-etc)       14
Honduras                         13
Hungary                          13
Scotland                         12
Holand-Netherlands                1
Name: count, dtype: int64
```

```
df['native-country'].value_counts(normalize=1)
```

```
native-country
United-States                 0.895857
Mexico                        0.019748
NaN                           0.017905
Philippines                   0.006081
Germany                       0.004207
Canada                        0.003716
Puerto-Rico                   0.003502
El-Salvador                   0.003255
India                         0.003071
Cuba                          0.002918
England                       0.002764
Jamaica                       0.002488
South                         0.002457
China                         0.002303
Italy                         0.002242
Dominican-Republic            0.002150
Vietnam                       0.002058
Guatemala                     0.001966
Japan                         0.001904
Poland                        0.001843
Columbia                      0.001813
Taiwan                        0.001566
Haiti                         0.001351
Iran                          0.001321
Portugal                      0.001137
Nicaragua                     0.001044
Peru                          0.000952
Greece                        0.000891
France                        0.000829
Ecuador                       0.000860
Ireland                       0.000737
Hong                          0.000614
Cambodia                      0.000583
Trinidad&Tobago               0.000583
Laos                          0.000553
Thailand                      0.000553
Yugoslavia                    0.000491
Outlying-US(Guam-USVI-etc)    0.000430
Honduras                      0.000399
Hungary                       0.000399
Scotland                      0.000368
Holand-Netherlands            0.000031
Name: proportion, dtype: int64
```

```
df['native-country'].value_counts(normalize=1).sum()
```

```
1.0
```

```
df.sample(5)
```

	age	workclass	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country
1	50	Self-emp-not-inc	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States
2	38	Private	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States
3	53	Private	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States
4	28	Private	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba
5	37	Private	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40	United-States

```
df.to_csv('/content/drive/MyDrive/Data/Modified/Modified.csv')
```

```
df.describe()
```

	age	education-num	capital-gain	capital-loss	hours-per-week
count	32561.000000	32561.000000	32561.000000	32561.000000	32561.000000
mean	38.581647	10.080679	1077.648844	87.303830	40.437456
std	13.640433	2.572720	7385.292085	402.960219	12.347429
min	17.000000	1.000000	0.000000	0.000000	1.000000
25%	28.000000	9.000000	0.000000	0.000000	40.000000
50%	37.000000	10.000000	0.000000	0.000000	40.000000
75%	48.000000	12.000000	0.000000	0.000000	45.000000
max	90.000000	16.000000	99999.000000	4356.000000	99.000000

```
df.describe(include='object')
```

	workclass	education	marital-status	occupation	relationship	race	sex	native-country
count	32561	32561	32561	32561	32561	32561	32561	32561
unique	9	16	7	15	6	5	2	42
top	Private	HS-grad	Married-civ-spouse	Prof-specialty	Husband	White	Male	United-States
freq	22696	10501	14976	4140	13193	27816	21790	29170