Age Vs Salary Classification either obove 50k or less 50k through logistic regression classification

Reference of data set: https://www.kaggle.com/wenruliu/adult-income-dataset

```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
%matplotlib inline
```

```
df = pd.read_csv(r'https://github.com/kaopanboonyuen/Python-Data-Science/raw/master/Dataset/a
df.head()
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relat
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	
[illegible]	[illegible]	[illegible]	[illegible]	Assoc-	[illegible]	Married-	Protective-	

```
df.columns
```

```
Index(['age', 'workclass', 'fnlwgt', 'education', 'educational-num',
       'marital-status', 'occupation', 'relationship', 'race', 'gender',
       'capital-gain', 'capital-loss', 'hours-per-week', 'native-country',
       'income'],
      dtype='object')
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48842 entries, 0 to 48841
Data columns (total 15 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   age              48842 non-null  int64 
 1   workclass        48842 non-null  object
 2   fnlwgt           48842 non-null  int64 
 3   education        48842 non-null  object
 4   educational-num  48842 non-null  int64 
 5   marital-status   48842 non-null  object
 6   occupation       48842 non-null  object
 7   relationship     48842 non-null  object
 8   race             48842 non-null  object
 9   gender           48842 non-null  object
 10  capital-gain     48842 non-null  int64 
 11  capital-loss     48842 non-null  int64 
 12  hours-per-week   48842 non-null  int64 
 13  native-country   48842 non-null  object
 14  income           48842 non-null  object
dtypes: int64(6), object(9)
memory usage: 5.6+ MB
```

Analysing data

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

```
<=50K    37155
>50K     11687
Name: income, dtype: int64
```

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

```
36    1348
35    1337
33    1335
23    1329
31    1325
      ... 
88       6
85       5
87       3
89       2
86       1
Name: age, Length: 74, dtype: int64
```

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

```
Private             33906
Self-emp-not-inc     3862
Local-gov            3136
?                    2799
State-gov            1981
Self-emp-inc         1695
Federal-gov          1432
Without-pay            21
Never-worked           10
Name: workclass, dtype: int64
```

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

```
HS-grad         15784
Some-college    10878
Bachelors        8025
Masters          2657
Assoc-voc        2061
11th             1812
Assoc-acdm       1601
10th             1389
7th-8th           955
Prof-school       834
9th               756
12th              657
Doctorate         594
5th-6th           509
1st-4th           247
Preschool          83
Name: education, dtype: int64
```

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

```

```
Prof-specialty       6172
Craft-repair         6112
Exec-managerial      6086
Adm-clerical         5611
Sales                5504
Other-service        4923
Machine-op-inspct    3022
?                    2809
Transport-moving     2355
Handlers-cleaners    2072
Farming-fishing      1490
Tech-support         1446
Protective-serv       983
Priv-house-serv       242
Armed-Forces           15
Name: occupation, dtype: int64
```

```
df['capital-gain'].value_counts()
#ตัดทิ้ง
```

```
0        44807
15024      513
7688       410
7298       364
99999      244
         ... 
```

```
1111         1
7262         1
22040        1
1639         1
2387         1
Name: capital-gain, Length: 123, dtype: int64
```

```
df.groupby('income')['educational-num'].value_counts()
```

```
income  educational-num
<=50K   9                  13281
        10                  8815
        13                  4712
        7                   1720
        11                  1539
        6                   1302
        14                  1198
        12                  1188
        4                    893
        5                    715
        8                    609
        3                    482
        2                    239
        15                   217
        16                   163
        1                     82
>50K    13                  3313
        9                   2503
        10                  2063
        14                  1459
        15                   617
        11                   522
        16                   431
        12                   413
        7                     92
        6                     87
        4                     62
        8                     48
        5                     41
        3                     27
        2                      8
        1                      1
Name: educational-num, dtype: int64
```

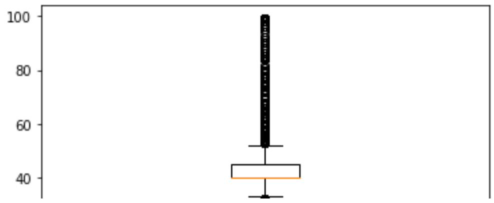
```
df['capital-loss'].value_counts()
#ตัดทิ้ง
```

```
0       46560
1902      304
1977      253
1887      233
2415       72
        ...
2465        1
2080        1
155         1
1911        1
2201        1
Name: capital-loss, Length: 99, dtype: int64
```

```
df['hours-per-week'].value_counts()
df['hours-per-week'].describe()
```

```
count    48842.000000
mean        40.422382
std         12.391444
min          1.000000
25%         40.000000
50%         40.000000
75%         45.000000
max         99.000000
Name: hours-per-week, dtype: float64
```

```
plt.boxplot(df['hours-per-week'])
fig = plt.figure(figsize =(10, 7))
plt.show()
```



```
q1 = np.quantile(df['hours-per-week'], 0.25)
q3 = np.quantile(df['hours-per-week'], 0.75)
print(q1,q3)
```

```
40.0 45.0
```

```
print(len(df[(df['hours-per-week']<=45)&(df['hours-per-week']>=40)]))
print(len(df[df['hours-per-week']<40]))
print(len(df[df['hours-per-week']>45]))
# แบ่งเปนสาม class few general hard
```

```
26454
11687
10701
```

```
df.groupby('income')['age', 'workclass', 'fnlwgt', 'education', 'educational-num','marital-st
                       'capital-gain', 'capital-loss', 'hours-per-week', 'native-country'].des
```

```
<ipython-input-16-6e6c01bdb606>:1: FutureWarning: Indexing with multiple keys
  df.groupby('income')['age', 'workclass', 'fnlwgt', 'education', 'educationa
```

	age								fnlwgt	
	count	mean	std	min	25%	50%	75%	max	count	mean
income										
<=50K	37155.0	36.872184	14.104118	17.0	25.0	34.0	46.0	90.0	37155.0	190039.5655
>50K	11687.0	44.275178	10.558983	19.0	36.0	43.0	51.0	90.0	11687.0	188470.5745

```
df.groupby('income')['age', 'workclass', 'fnlwgt', 'education', 'educational-num','marital-st
                       'capital-gain', 'capital-loss', 'hours-per-week', 'native-country'].mea
```

```
<ipython-input-21-01d05853866b>:1: FutureWarning: Indexing with multiple keys
  df.groupby('income')['age', 'workclass', 'fnlwgt', 'education', 'educationa
```

	age	fnlwgt	educational-num	capital-gain	capital-loss	hours-per-week
income						
<=50K	36.872184	190039.565523	9.598493	147.010308	54.151931	38.840048

ดับเบิลคลิก (หรือกด Enter) เพื่อแก้ไข

```
df.groupby('income')['age', 'workclass', 'fnlwgt', 'education', 'educational-num','marital-st
                       'capital-gain', 'capital-loss', 'hours-per-week', 'native-country'].med
```

```
<ipython-input-28-bdd548662ebf>:1: FutureWarning: Indexing with multiple keys
  df.groupby('income')['age', 'workclass', 'fnlwgt', 'education', 'educationa
```

	age	fnlwgt	educational-num	capital-gain	capital-loss	hours-per-week
income						
<=50K	34.0	178811.0	9.0	0.0	0.0	40.0

```
df.head()
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relat
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	
				Assoc-		Married-	Protective-	

```
#Categories variable : race , gender
```

```
# df.loc[df['income'] == '>50K','age'].value_counts()
# df.loc[df['income'] == '>50K','race'].plot(kind='bar')
```

```
df['income_class'] = df.loc[:, 'income']
```

```
df
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	r
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	
4	18	?	103497	Some-college	10	Never-married	?	
...	...	...	...	...	...	...	...	
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	

```
#เปลี่ยนเป็น income เป็น num
df.loc[df['income'] == '>50K','income_class'] = 0
df.loc[df['income'] == '<=50K','income_class'] = 1
df
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	r
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	
4	18	?	103497	Some-college	10	Never-married	?	
...	...	...	...	...	...	...	...	
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	

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

```
Husband           19716
Not-in-family     12583
Own-child          7581
Unmarried          5125
Wife               2331
```

```
Other-relative     1506
Name: relationship, dtype: int64
```

```
# df.groupby('income_class')['age'].plot(kind='bar')
```

```
# df.groupby('income').plot(kind='bar')
```

```
cols = ['age', 'workclass', 'fnlwgt', 'education', 'educational-num','marital-status', 'occup
                      'capital-gain', 'capital-loss', 'hours-per-week', 'native-country']
for i in cols:
    print(df.groupby('income')[i].value_counts())
# age :class1 >80 , class0 <80
# fnlwgt :class1 >900k , class0 <900k
# race : White , other
# ตัด work class, education, educational-num, marital-status, occupation
```

```
income  age
<=50K   23     1307
        24     1162
        22     1161
        25     1119
        27     1117
               ...
>50K    83        2
        20        1
        84        1
        85        1
        88        1
Name: age, Length: 142, dtype: int64
income  workclass
<=50K   Private             26519
        Self-emp-not-inc     2785
        ?                    2534
        Local-gov            2209
        State-gov            1451
        Federal-gov           871
        Self-emp-inc          757
        Without-pay            19
        Never-worked           10
>50K    Private              7387
        Self-emp-not-inc     1077
        Self-emp-inc          938
        Local-gov             927
        Federal-gov           561
        State-gov             530
        ?                     265
        Without-pay             2
Name: workclass, dtype: int64
income  fnlwgt
<=50K   113364     14
        190290     14
        203488     14
        117789     13
        194630     13
                   ..
>50K    914061      1
        953588      1
        1033222     1
        1097453     1
        1226583     1
Name: fnlwgt, Length: 32732, dtype: int64
income  education
<=50K   HS-grad         13281
        Some-college     8815
        Bachelors        4712
        11th             1720
        Assoc-voc        1539
        10th             1302
        Masters          1198
        Assoc-acdm       1188
        7th-8th           893
        9th               715
        12th              609
        5th-6th           482
```

Observe

```python
for i in cols:
    print(df.loc[df['income'] == '>50K',i].value_counts())
    print(df.loc[df['income'] == '<=50K',i].value_counts())

```

```
46    439
47    429
41    427
39    423
37    422
     ...
83      2
85      1
20      1
88      1
84      1
Name: age, Length: 68, dtype: int64
23    1307
24    1162
22    1161
25    1119
27    1117
      ...
88       5
85       4
87       3
89       2
86       1
Name: age, Length: 74, dtype: int64
Private             7387
Self-emp-not-inc    1077
Self-emp-inc         938
Local-gov            927
Federal-gov          561
State-gov            530
?                    265
Without-pay            2
Name: workclass, dtype: int64
Private             26519
Self-emp-not-inc     2785
?                    2534
Local-gov            2209
State-gov            1451
Federal-gov           871
Self-emp-inc          757
Without-pay            19
Never-worked           10
Name: workclass, dtype: int64
121124    12
125892    12
148995    12
123011    12
132879    11
          ..
138022     1
87418      1
177307     1
270335     1
287927     1
Name: fnlwgt, Length: 8172, dtype: int64
113364    14
190290    14
203488    14
```

```python
len(df[(df['income'] == '>50K')&(df['hours-per-week'] <= 30)])

```

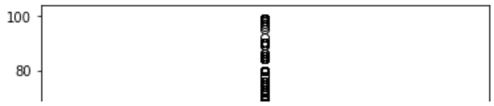
```
526
```

```python
len(df[(df['income'] == '<=50K')&(df['hours-per-week'] < 30)])
```

```
5741
```

```python
df.loc[df['income'] == '>50K','hours-per-week'].plot(kind='box')
```

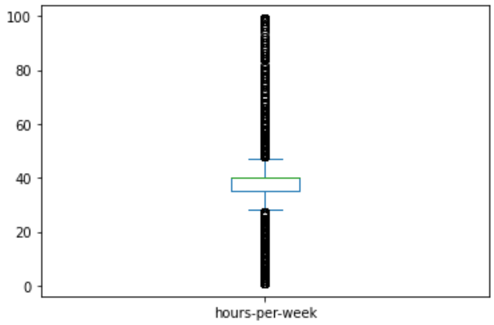
```
<AxesSubplot:>
```



```
df.loc[df['income'] == '>50K','hours-per-week'].plot(kind='box')
```

```
df.loc[df['income'] == '<=50K','hours-per-week'].plot(kind='box')
```

```
<AxesSubplot:>
```



```
df.loc[df['income'] == '<=50K','hours-per-week'].describe()
```

```
count    37155.000000
mean        38.840048
std         12.356849
min          1.000000
25%         35.000000
50%         40.000000
75%         40.000000
max         99.000000
Name: hours-per-week, dtype: float64
```

```
len(df[df['fnlwgt']>900000])
```

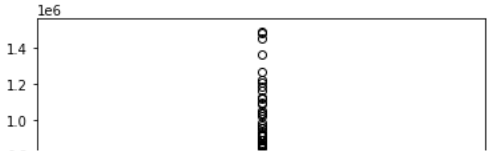
```
18
```

```
df[df['income']=='>50K']
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	r
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	
7	63	Self-emp-not-inc	104626	Prof-school	15	Married-civ-spouse	Prof-specialty	
10	65	Private	184454	HS-grad	9	Married-civ-spouse	Machine-op-inspct	
14	48	Private	279724	HS-grad	9	Married-civ-spouse	Machine-op-inspct	
...	...	...	...	...	...	...	...	
48820	71	?	287372	Doctorate	16	Married-civ-spouse	?	

```
df['fnlwgt'].plot(kind='box')
```

```
<AxesSubplot:>
```



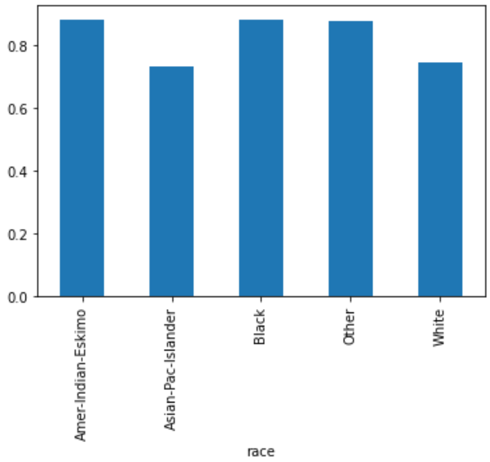
```
df.loc[df['income'] == '<=50K','capital-gain'].value_counts()
```

```
0        35611
5013       117
3325        81
2174        74
4650        63
         ...
1731         1
1111         1
22040        1
1639         1
2387         1
Name: capital-gain, Length: 92, dtype: int64
```

```
df.loc[df['income'] == '<=50K','capital-gain'].value_counts()
```

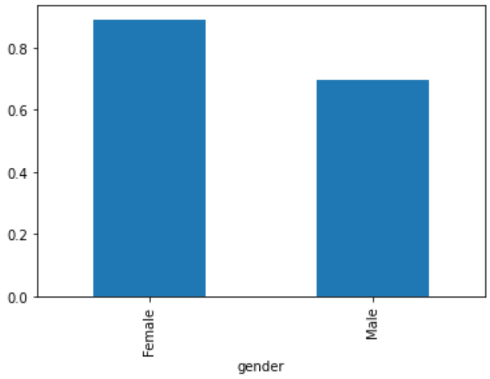
```
# n_col = ['race','gender','relationship']
df.groupby('race')['income_class'].mean().plot(kind='bar')
```

```
<AxesSubplot:xlabel='race'>
```



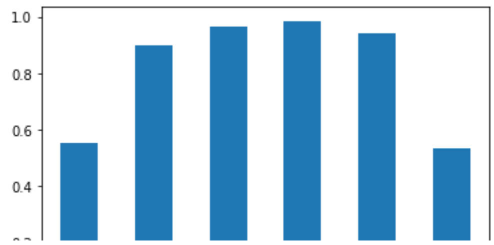
```
df.groupby('gender')['income_class'].mean().plot(kind='bar')

```

```
<AxesSubplot:xlabel='gender'>
```



```
df.groupby('relationship')['income_class'].mean().plot(kind='bar')
#แบ่งเป็น married(Husband,Wife) กับ other
```

```
<AxesSubplot:xlabel='relationship'>
```



```
df['relationship_class'] = df.loc[:,'relationship']
```

```
df['relationship_class'] = np.where((df['relationship_class'] == 'Husband')|(df['relationship
```

```
df
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	r
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	
4	18	?	103497	Some-college	10	Never-married	?	
...	...	...	...	...	...	...	...	
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	

```
df['Adm-clerical'] = df.loc[:,'occupation']
df['Adm-clerical'] = np.where(df['occupation'] == 'Adm-clerical',1,0)
df
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	r
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	
4	18	?	103497	Some-college	10	Never-married	?	
...	...	...	...	...	...	...	...	
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	

```
df['Mexico'] = df.loc[:,'native-country']
df['Mexico'] = np.where(df['native-country'] == 'Mexico',1,0)
df
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital los
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	
4	18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	
...	...	...	...	...	...	...	...	...	...	...	...	
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	
48838	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	

```
df['Own-child'] = df.loc[:,'relationship']
df['Own-child'] = np.where(df['relationship'] == 'Own-child',1,0)
df
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capital los
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	
4	18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	0	
...	...	...	...	...	...	...	...	...	...	...	...	
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	
48838	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	
48839	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	
48840	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	
		Self-emp-				Married-	Exec-					

```
# utlying-US(Guam-USVI-etc)
df['Outlying-US(Guam-USVI-etc)'] = df.loc[:,'native-country']
df['Outlying-US(Guam-USVI-etc)'] = np.where(df['native-country'] == 'Outlying-US(Guam-USVI-etc
df
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	...	relationship
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	...	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	...	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	...	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	...	

```
df['work_hard'] = df.loc[:,'hours-per-week']
df['work_hard'] = np.where(df['hours-per-week'] < 30,1,0)
df
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relationship	race	gender	...	relationship
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	...	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	...	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	...	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	...	
4	18	?	103497	Some-college	10	Never-married	?	Own-child	White	Female	...	
...	...	...	...	...	...	...	...	...	...	...	...	
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	...	
48838	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	...	
48839	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	...	
48840	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	...	
48841	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	...	

```
df['capg_h'] = df.loc[:,'capital-gain']
df['capg_h'] = np.where(df['capital-gain'] > 5100,1,0)
df
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	r
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	
4	18	?	103497	Some-college	10	Never-married	?	

```
df['Never-married'] = df.loc[:,'marital-status']
df['Never-married'] = np.where(df['marital-status'] == 'Never-married',1,0)
df
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	r
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	
4	18	?	103497	Some-college	10	Never-married	?	
...	...	...	...	...	...	...	...	
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	
48838	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	
48839	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	
48840	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	
48841	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	

```
df['pre_school'] = df.loc[:,'education']
df['pre_school'] = np.where(df['education'] == 'Preschool',1,0)
df
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	r
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	
4	18	?	103497	Some-college	10	Never-married	?	
...	...	...	...	...	...	...	...	
48837	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	
48838	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	
48839	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	
48840	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	
48841	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48842 entries, 0 to 48841
Data columns (total 26 columns):
 #   Column                      Non-Null Count  Dtype 
---  ------                      --------------  ----- 
 0   age                         48842 non-null  int64 
 1   workclass                   48842 non-null  object
 2   fnlwgt                      48842 non-null  int64 
 3   education                   48842 non-null  object
 4   educational-num             48842 non-null  int64 
 5   marital-status              48842 non-null  object
 6   occupation                  48842 non-null  object
 7   relationship                48842 non-null  object
 8   race                        48842 non-null  object
 9   gender                      48842 non-null  object
 10  capital-gain                48842 non-null  int64 
 11  capital-loss                48842 non-null  int64 
 12  hours-per-week              48842 non-null  int64 
 13  native-country              48842 non-null  object
 14  income                      48842 non-null  object
 15  income_class                48842 non-null  int64 
 16  relationship_class          48842 non-null  object
 17  Adm-clerical                48842 non-null  int64 
 18  Mexico                      48842 non-null  int64 
 19  Own-child                   48842 non-null  int64 
 20  utlying-US(Guam-USVI-etc)   48842 non-null  int64 
 21  work_hard                   48842 non-null  int64 
 22  capg_h                      48842 non-null  int64 
 23  Never-married               48842 non-null  int64 
 24  pre_school                  48842 non-null  int64 
 25  Outlying-US(Guam-USVI-etc)  48842 non-null  int64 
dtypes: int64(16), object(10)
memory usage: 9.7+ MB
```

```
df['income_class'] = df['income_class'].astype(int)
```

```
df.corr()
```

	age	fnlwgt	educational-num	capital-gain	capital-loss	hours-per-week	inc
age	1.000000	-0.076628	0.030940	0.077229	0.056944	0.071558	
fnlwgt	-0.076628	1.000000	-0.038761	-0.003706	-0.004366	-0.013519	
educational-num	0.030940	-0.038761	1.000000	0.125146	0.080972	0.143689	
capital-gain	0.077229	-0.003706	0.125146	1.000000	-0.031441	0.082157	
capital-loss	0.056944	-0.004366	0.080972	-0.031441	1.000000	0.054467	
hours-per-week	0.071558	-0.013519	0.143689	0.082157	0.054467	1.000000	
income_class	-0.230369	0.006339	-0.332613	-0.223013	-0.147554	-0.227687	
Adm-clerical	-0.038116	0.007480	0.004142	-0.029105	-0.021457	-0.078916	
Mexico	-0.051478	0.126589	-0.222085	-0.012540	-0.019178	-0.002376	
Own-child	-0.432990	0.016716	-0.097316	-0.052038	-0.049167	-0.251827	
utlying-US(Guam-USVI-etc)	NaN	NaN	NaN	NaN	NaN	NaN	
work_hard	-0.053907	-0.010005	-0.091675	-0.038148	-0.034280	-0.692304	
capg_h	0.115607	-0.003924	0.162125	0.576914	-0.048619	0.098459	

Test 1

Logistic Regression model

```
sex = pd.get_dummies(df['gender'])
```

```
sex.head(3)
```

	Female	Male
0	0	1
1	0	1
2	0	1

```
df_new = df.drop(['gender'],axis=1)
```

```
df_new = pd.concat([df_new,sex],axis=1)
```

```
df_new.head()
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relat
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	
4	18	?	103497	Some-college	10	Never-married	?	

```
df_new.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48842 entries, 0 to 48841
Data columns (total 27 columns):
 #   Column                      Non-Null Count  Dtype 
---  ------                      --------------  ----- 
 0   age                         48842 non-null  int64 
 1   workclass                   48842 non-null  object
 2   fnlwgt                      48842 non-null  int64 
 3   education                   48842 non-null  object
 4   educational-num             48842 non-null  int64 
 5   marital-status              48842 non-null  object
 6   occupation                  48842 non-null  object
 7   relationship                48842 non-null  object
 8   race                        48842 non-null  object
 9   capital-gain                48842 non-null  int64 
 10  capital-loss                48842 non-null  int64 
 11  hours-per-week              48842 non-null  int64 
 12  native-country              48842 non-null  object
 13  income                      48842 non-null  object
 14  income_class                48842 non-null  int64 
 15  relationship_class          48842 non-null  object
 16  Adm-clerical                48842 non-null  int64 
 17  Mexico                      48842 non-null  int64 
 18  Own-child                   48842 non-null  int64 
 19  utlying-US(Guam-USVI-etc)   48842 non-null  int64 
 20  work_hard                   48842 non-null  int64 
 21  capg_h                      48842 non-null  int64 
 22  Never-married               48842 non-null  int64 
 23  pre_school                  48842 non-null  int64 
 24  Outlying-US(Guam-USVI-etc)  48842 non-null  int64 
 25  Female                      48842 non-null  uint8 
 26  Male                        48842 non-null  uint8 
dtypes: int64(16), object(9), uint8(2)
memory usage: 9.4+ MB
```

```
race = pd.get_dummies(df['race'])
```

```
df_new = df_new.drop(['race'],axis=1)
df_new = pd.concat([df_new,race],axis=1)
df_new.head()
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relat
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	
4	18	?	103497	Some-college	10	Never-married	?	

```
rela = pd.get_dummies(df['relationship_class'])
```

```
df_new = df_new.drop(['relationship_class'],axis=1)
df_new = pd.concat([df_new,rela],axis=1)
df_new.head()
```

	age	workclass	fnlwgt	education	educational-num	marital-status	occupation	relat
0	25	Private	226802	11th	7	Never-married	Machine-op-inspct	
1	38	Private	89814	HS-grad	9	Married-civ-spouse	Farming-fishing	
2	28	Local-gov	336951	Assoc-acdm	12	Married-civ-spouse	Protective-serv	
3	44	Private	160323	Some-college	10	Married-civ-spouse	Machine-op-inspct	
4	18	?	103497	Some-college	10	Never-married	?	

Train

```
df.corr()
```

	age	fnlwgt	educational-num	capital-gain	capital-loss	hours-per-week	inc
age	1.000000	-0.076628	0.030940	0.077229	0.056944	0.071558	
fnlwgt	-0.076628	1.000000	-0.038761	-0.003706	-0.004366	-0.013519	
educational-num	0.030940	-0.038761	1.000000	0.125146	0.080972	0.143689	
capital-gain	0.077229	-0.003706	0.125146	1.000000	-0.031441	0.082157	
capital-loss	0.056944	-0.004366	0.080972	-0.031441	1.000000	0.054467	
hours-per-week	0.071558	-0.013519	0.143689	0.082157	0.054467	1.000000	
income_class	-0.230369	0.006339	-0.332613	-0.223013	-0.147554	-0.227687	
Adm-clerical	-0.038116	0.007480	0.004142	-0.029105	-0.021457	-0.078916	
Mexico	-0.051478	0.126589	-0.222085	-0.012540	-0.019178	-0.002376	
Own-child	-0.432990	0.016716	-0.097316	-0.052038	-0.049167	-0.251827	
utlying-US(Guam-USVI-etc)	NaN	NaN	NaN	NaN	NaN	NaN	
work_hard	-0.053907	-0.010005	-0.091675	-0.038148	-0.034280	-0.692304	
capg_h	0.115607	-0.003924	0.162125	0.576914	-0.048619	0.098459	
Never-							

```
cor = df.corr()
cor_target = abs(cor['income_class'])

#Selecting highly correlated features
relevant_features = cor_target[cor_target>0.2]
relevant_features
```

```
age                0.230369
educational-num    0.332613
capital-gain       0.223013
```

```
hours-per-week     0.227687
income_class       1.000000
Own-child          0.225691
capg_h             0.371346
Never-married      0.318782
Name: income_class, dtype: float64
```

```
from sklearn.model_selection import train_test_split
```

```
# col_pre = ['Female','Male','Black','White']
# col_pre = ['capital-gain','capital-loss','Adm-clerical','Mexico','Own-child',]
# col_pre = ['educational-num','capg_h','age','Never-married']
col_pre = ['educational-num','capg_h','Never-married']

# col_pre = ['capital-gain','capital-loss','Never-married']

```

```
X = df_new[col_pre]
y = df_new['income_class']
y = y.astype('int')
X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y, test_size=0.30, random_
```

```
from sklearn.linear_model import LogisticRegression
```

```
model = LogisticRegression()
model.fit(X_train,y_train)
```

```
LogisticRegression()
```

```
y_pred = model.predict(X_test)
```

```
print(list(y_test[:5]))
print(y_pred[:5])
```

```
[1, 1, 1, 1, 1]
[1 1 1 1 0]
```

Evaluate Model

```
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
```

```
confusion_matrix(y_test, y_pred)
```

```
array([[ 1908,  1598],
       [ 1000, 10147]])
```

```
print(classification_report(y_test,y_pred))
```

```
              precision    recall  f1-score   support

           0       0.66      0.54      0.59      3506
           1       0.86      0.91      0.89     11147

    accuracy                           0.82     14653
   macro avg       0.76      0.73      0.74     14653
weighted avg       0.81      0.82      0.82     14653
```

Compute odd ratio

```
print(X.columns)
print(model.intercept_)
print(model.coef_)
```

```
Index(['educational-num', 'capg_h', 'Never-married'], dtype='object')
[4.75837117]
[[-0.36937834 -4.14552532  2.57254605]]
```