Exploratory Data Analysis

Data Preview:

The dataset given has 48842 rows and 15 columns in total in which 6 are continuous and remaining are categorical attributes.

Columns and their data types:

```
df_dataset.dtypes
```

```
age                int64
workplace         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: object
```

1.Age(17-90)

2.Hours per week (1-99) etc.

Here we can see the number of categories in categorical attributes.

```
df = df_dataset.groupby('relationship').nunique()
for column in df_dataset.select_dtypes('object'):
    print('Number of categories in ', column,' are:', len(df_dataset.groupby(column).nunique()))
```

```
Number of categories in  workplace  are: 7
Number of categories in  education  are: 16
Number of categories in  marital_status  are: 7
Number of categories in  occupation  are: 14
Number of categories in  relationship  are: 6
Number of categories in  race  are: 5
Number of categories in  sex  are: 2
Number of categories in  native_country  are: 41
Number of categories in  income  are: 2
```

There are no duplicate rows in the dataset and some missing values in columns. Here we can see column wise missing values.

```
for i in df_dataset.columns:
    print('missing values in ',i,' column:', df_dataset.loc[df_dataset[i]=='?', i].size)
```

```
missing values in  age  column: 0
missing values in  workplace  column: 2799
missing values in  fnlwgt  column: 0
missing values in  education  column: 0
missing values in  education_num  column: 0
missing values in  marital_status  column: 0
missing values in  occupation  column: 2809
missing values in  relationship  column: 0
missing values in  race  column: 0
missing values in  sex  column: 0
missing values in  capital_gain  column: 0
missing values in  capital_loss  column: 0
missing values in  hours_per_week  column: 0
missing values in  native_country  column: 857
missing values in  income  column: 0
```

After removing the rows with missing values we left with 45222 entries in the dataset.

Description of the original dataset given :

```
df_dataset.describe()
```

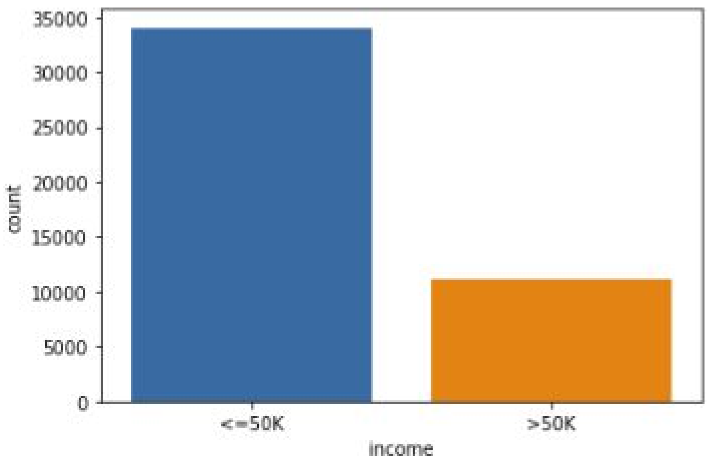
	age	fnlwgt	education_num	capital_gain	capital_loss	hours_per_week
count	48842.000000	4.884200e+04	48842.000000	48842.000000	48842.000000	48842.000000
mean	38.643585	1.896641e+05	10.078089	1079.067626	87.502314	40.422382
std	13.710510	1.056040e+05	2.570973	7452.019058	403.004552	12.391444
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.175505e+05	9.000000	0.000000	0.000000	40.000000
50%	37.000000	1.781445e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.376420e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.490400e+06	16.000000	99999.000000	4356.000000	99.000000

Plot Distributions:

- Population proportion based on income.

```
sns.countplot(df_dataset.income)
printmd('## Income count')
```

Income count

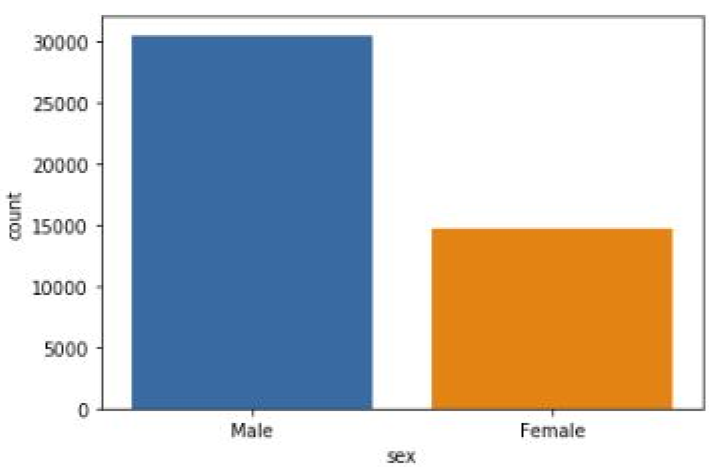


Only one fourth of people are earning more than 50K.

- Population proportion based on Gender.

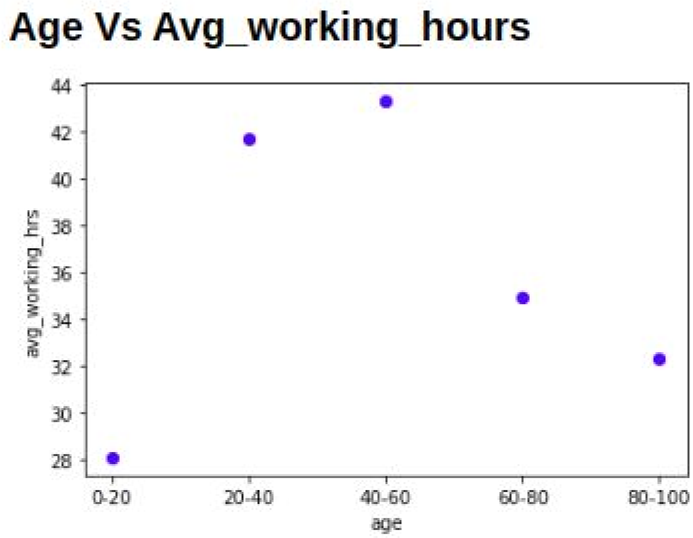
```
sns.countplot(df_dataset.sex)
printmd('## Gender count')
```

Gender count



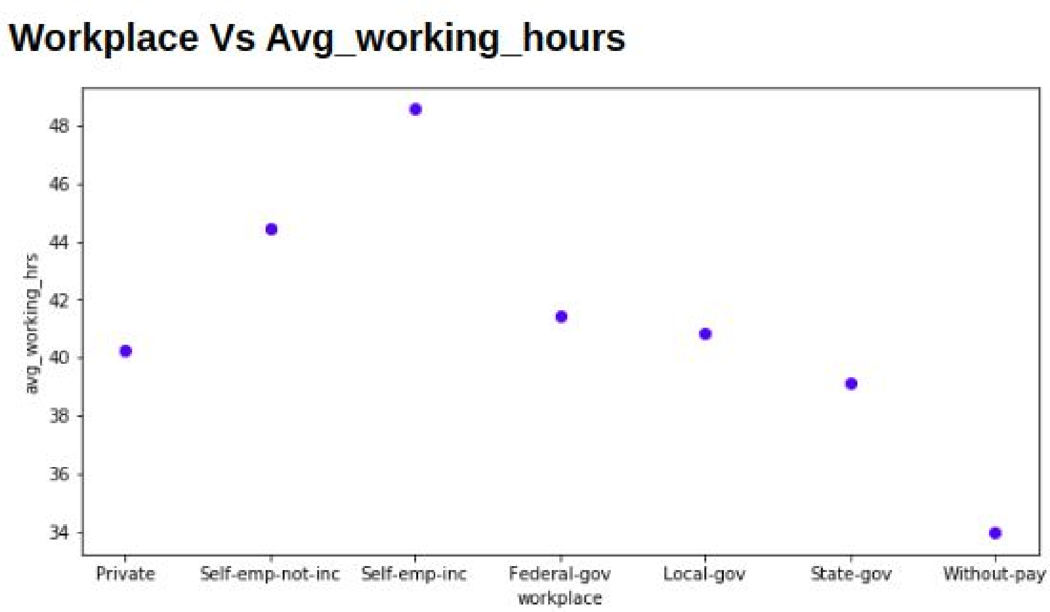
In the given dataset one third of females are earning and remaining.

- **Age vs Average working hours per week**



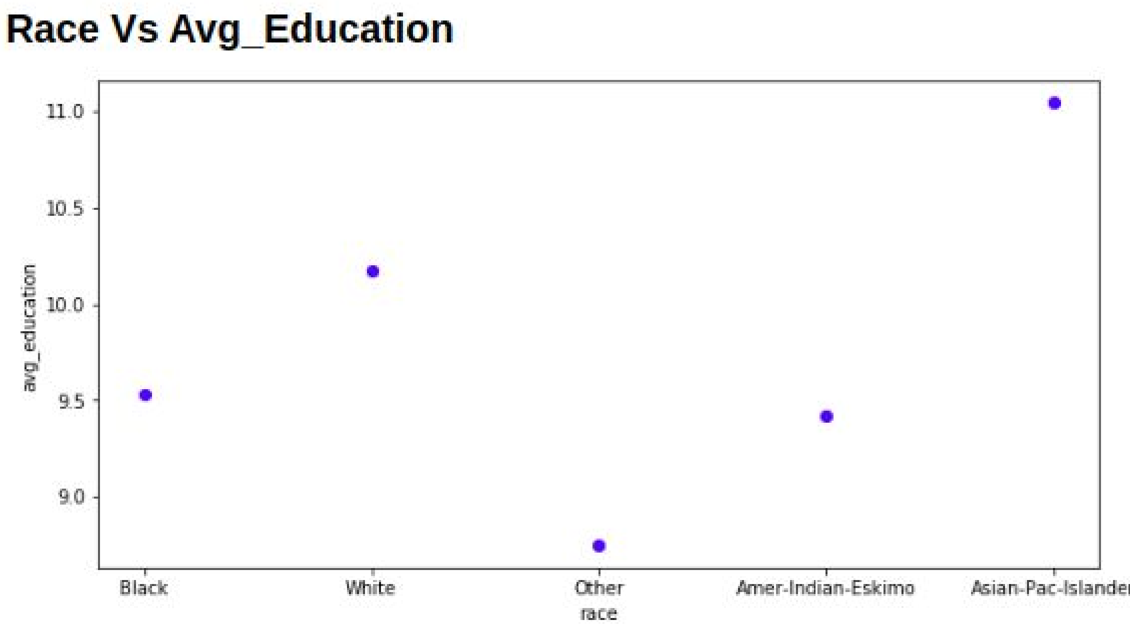
From this we can observe that the age group between 20-40 and 40-60 are working more number of hours per week than the remaining age groups and 0-20 age groups average working hours per week are much less than the remaining.

- **Workplace and average working hours per week**

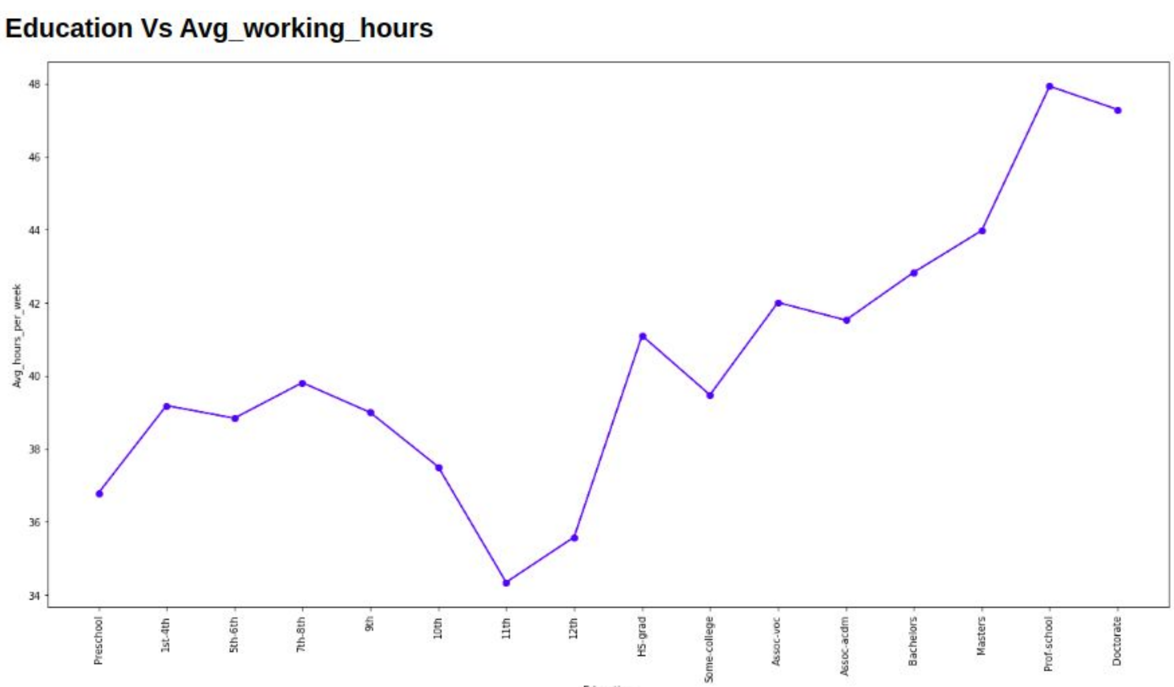


From this we can observe that **self-emp-not-inc** category's are working for more hours per week than other and The one's in the **without-pay** category are working less.

- Race vs Average education



- **Education vs Average working hours**

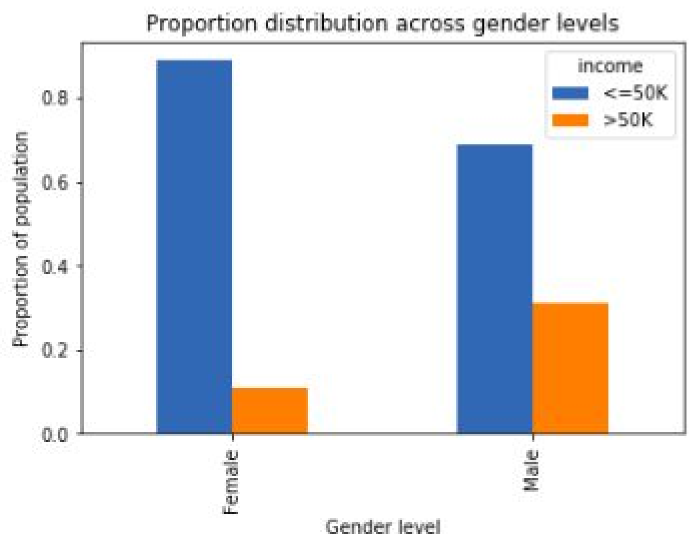


From the above graph we can see that after the 11th the average working hours per week are increasing.

- **Gender Proportion based on Income**

Gender Propotion Based on Income

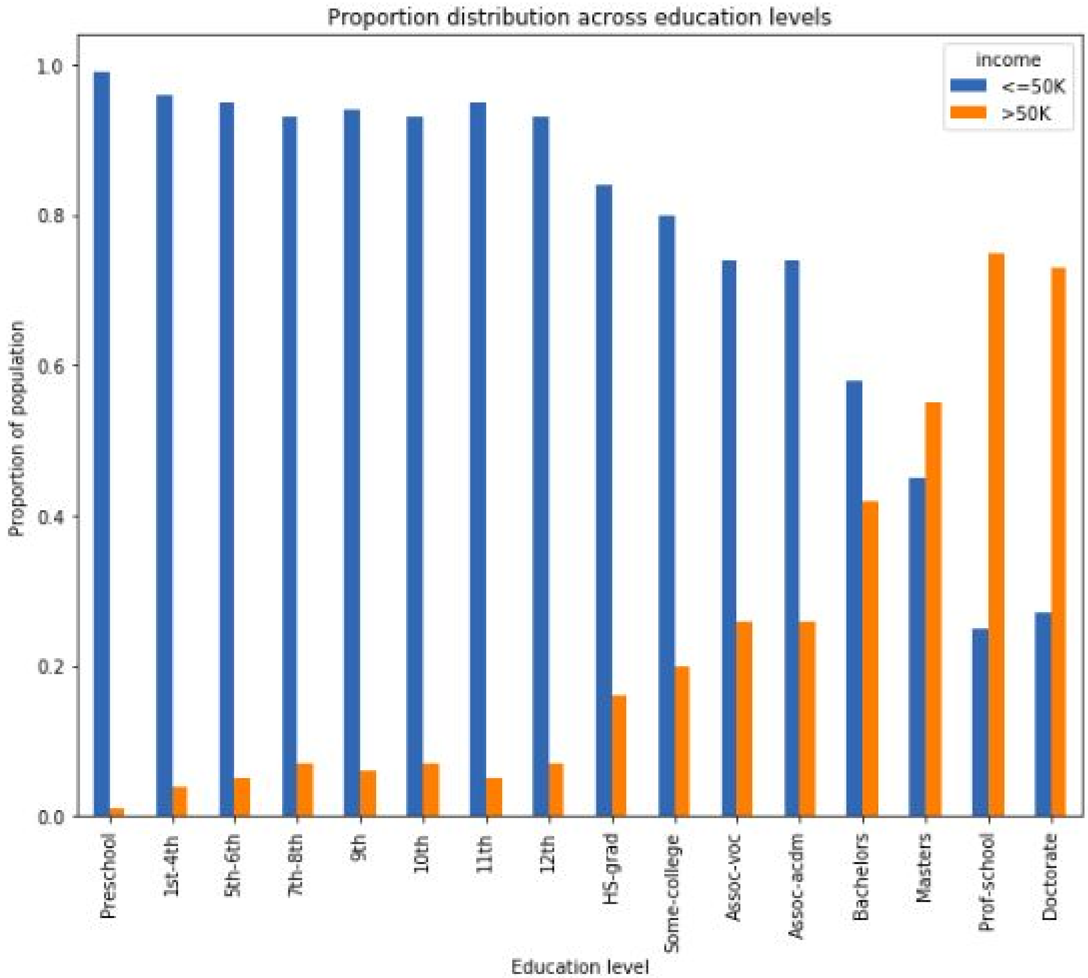
```
]: Text(0, 0.5, 'Proportion of population')
```



- **Education and Income level**

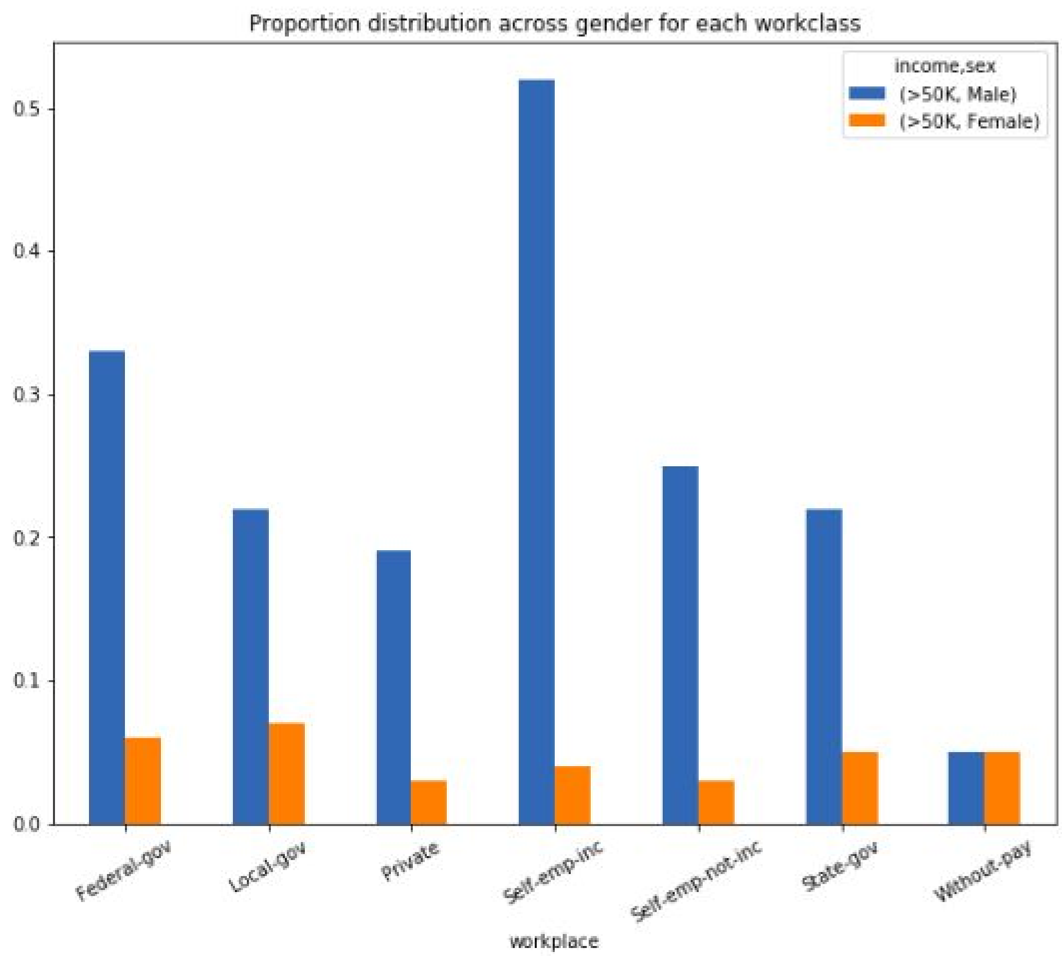
Education and Income level

```
Text(0, 0.5, 'Proportion of population')
```

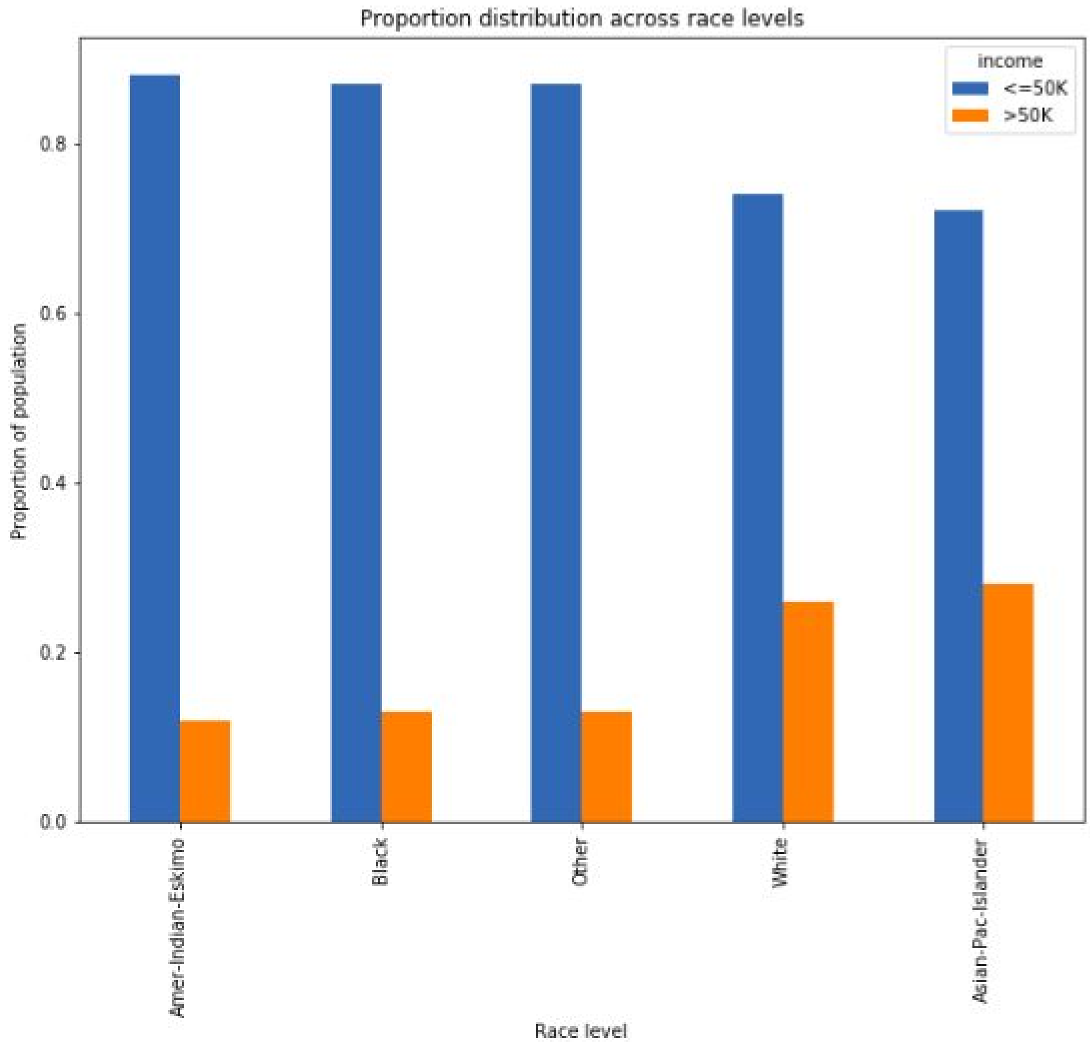


As Education level increases the number proportion of the people earning more than 50K is increasing.

- **Workplace and Income**



- Race and Income level



Building Classifier:
Used Logistic regression classifier to classify the given data.
Splitted the dataset in to train_data and test_data.
The metrics of the model are:

	accuracy	precision	recall	f_measure	sensitivity	specificity	error_rate
logistic_reg	0.8472	0.7244	0.6157	0.6656	0.6157	0.9232	0.1528