# Exploratoryt Data Analysis Of Adult Income Dataset

In [1]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:

```
Data = pd.read_csv('adult.csv-Dataset/adult.data',header = None)
```

In [3]:

```
Data.rename(columns={0:'age', 1:'workclass', 2:'fnlwgt', 3:'education', 4:'educational-num', 5:'marital-status'
```

In [4]:

```
Data.head(n = 5)
```

Out[4]:

| | age | workclass | fnlwgt | education | educational-num | marital-status | occupation | relationship | race | gender | capital-gain | capital-loss | hours-per-week | n co |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 39 | State-gov | 77516 | Bachelors | 13 | Never-married | Adm-clerical | Not-in-family | White | Male | 2174 | 0 | 40 | U |
| **1** | 50 | Self-emp-not-inc | 83311 | Bachelors | 13 | Married-civ-spouse | Exec-managerial | Husband | White | Male | 0 | 0 | 13 | U |
| **2** | 38 | Private | 215646 | HS-grad | 9 | Divorced | Handlers-cleaners | Not-in-family | White | Male | 0 | 0 | 40 | U |
| **3** | 53 | Private | 234721 | 11th | 7 | Married-civ-spouse | Handlers-cleaners | Husband | Black | Male | 0 | 0 | 40 | U |
| **4** | 28 | Private | 338409 | Bachelors | 13 | Married-civ-spouse | Prof-specialty | Wife | Black | Female | 0 | 0 | 40 | |

In [5]:

```
Data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32561 entries, 0 to 32560
Data columns (total 15 columns):
 #   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   educational-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   gender           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
```

In [6]:

```
Data.describe
```

Out[6]:

```
<bound method NDFrame.describe of        age         workclass  fnlwgt    education  educational-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  \
0            Never-married        Adm-clerical   Not-in-family   White
1       Married-civ-spouse     Exec-managerial         Husband   White
2                 Divorced   Handlers-cleaners   Not-in-family   White
3       Married-civ-spouse   Handlers-cleaners         Husband   Black
4       Married-civ-spouse      Prof-specialty            Wife   Black
...                    ...                 ...             ...     ...
32556   Married-civ-spouse        Tech-support            Wife   White
32557   Married-civ-spouse   Machine-op-inspct         Husband   White
32558              Widowed        Adm-clerical       Unmarried   White
32559        Never-married        Adm-clerical       Own-child   White
32560   Married-civ-spouse     Exec-managerial            Wife   White

        gender  capital-gain  capital-loss  hours-per-week  native-country  \
0         Male          2174             0              40   United-States
1         Male             0             0              13   United-States
2         Male             0             0              40   United-States
3         Male             0             0              40   United-States
4       Female             0             0              40            Cuba
...        ...           ...           ...             ...             ...
32556   Female             0             0              38   United-States
32557     Male             0             0              40   United-States
32558   Female             0             0              40   United-States
32559     Male             0             0              20   United-States
32560   Female         15024             0              40   United-States

       income
0       <=50K
1       <=50K
2       <=50K
3       <=50K
4       <=50K
...       ...
32556   <=50K
32557    >50K
32558   <=50K
32559   <=50K
32560    >50K

[32561 rows x 15 columns]>
```

In [7]:

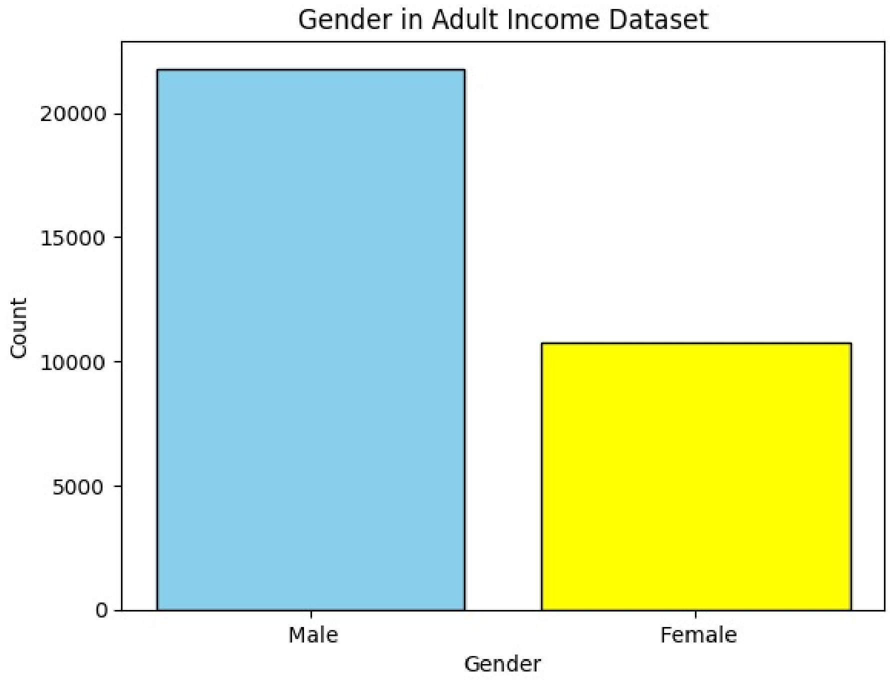
```
Data.isna().sum()
```

Out[7]:

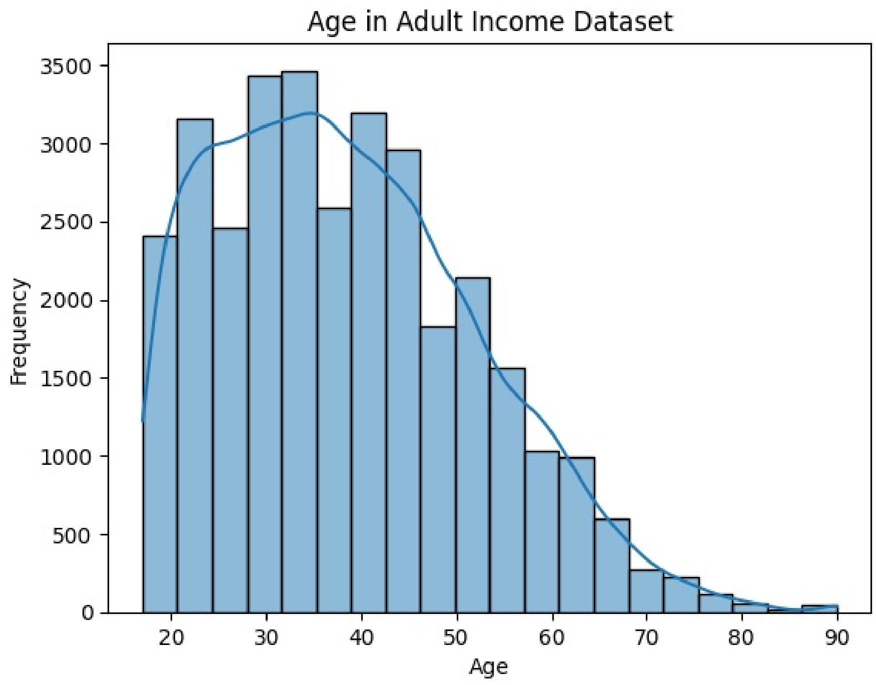
```
age                0
workclass          0
fnlwgt             0
education          0
educational-num    0
marital-status     0
occupation         0
relationship       0
race               0
gender             0
capital-gain       0
capital-loss       0
hours-per-week     0
native-country     0
income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

In [8]:

```
g_counts = Data['gender'].value_counts()
plt.bar(g_counts.index, g_counts, color=['skyblue', 'yellow'], edgecolor="black")
plt.xlabel('Gender')
plt.ylabel('Count')
plt.title('Gender in Adult Income Dataset')
plt.show()
```

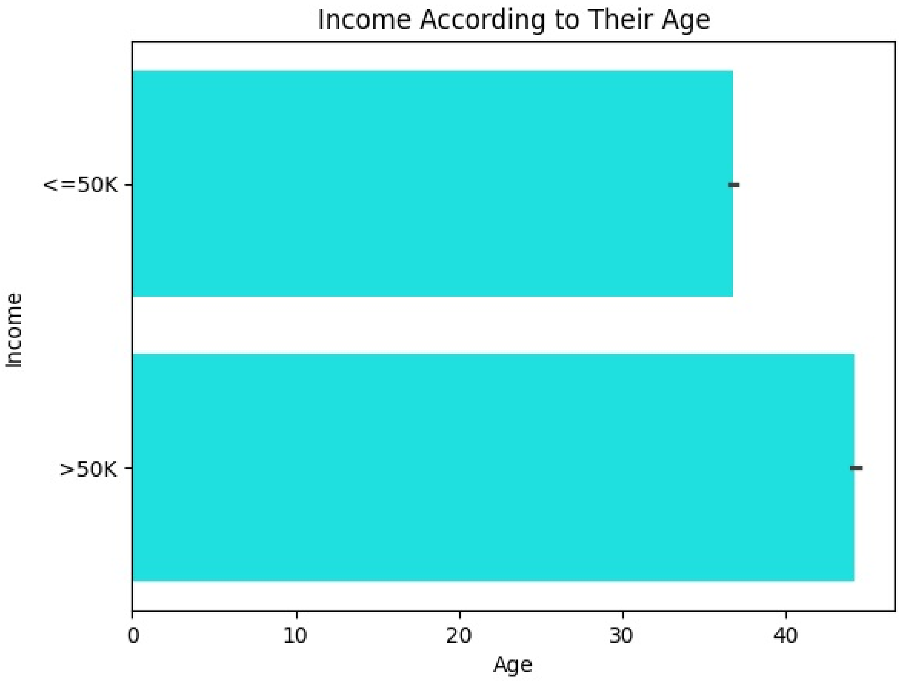


In [9]:
```
sns.histplot(Data['age'], bins=20, kde=True)
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.title('Age in Adult Income Dataset')
plt.show()
```



**Bar Plot for Age vs Income**

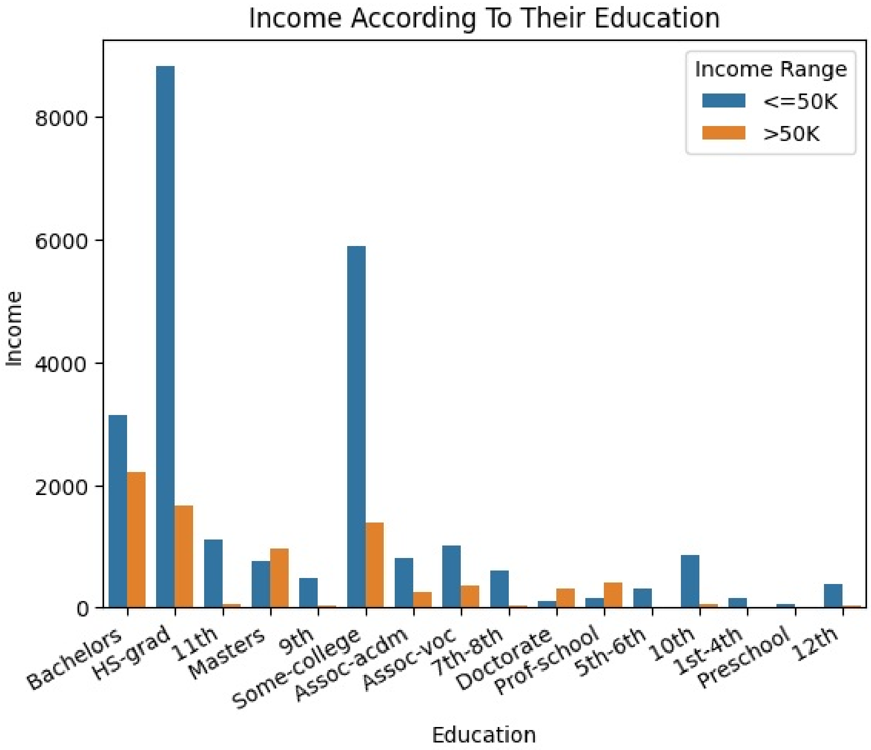
In [10]:
```
sns.barplot(x='age', y='income', data=Data, color="Cyan")
plt.xlabel('Age')
plt.ylabel('Income')
plt.title('Income According to Their Age')
plt.show()
```



**Bar Plot for Education vs Income**

In [11]:

```
sns.countplot(x='education', hue='income', data=Data)
plt.xticks(rotation=30, ha='right')
plt.xlabel('Education')
plt.ylabel('Income')
plt.title('Income According To Their Education')
plt.legend(title='Income Range', loc='upper right', labels=['<=50K', '>50K'])
plt.show()
```

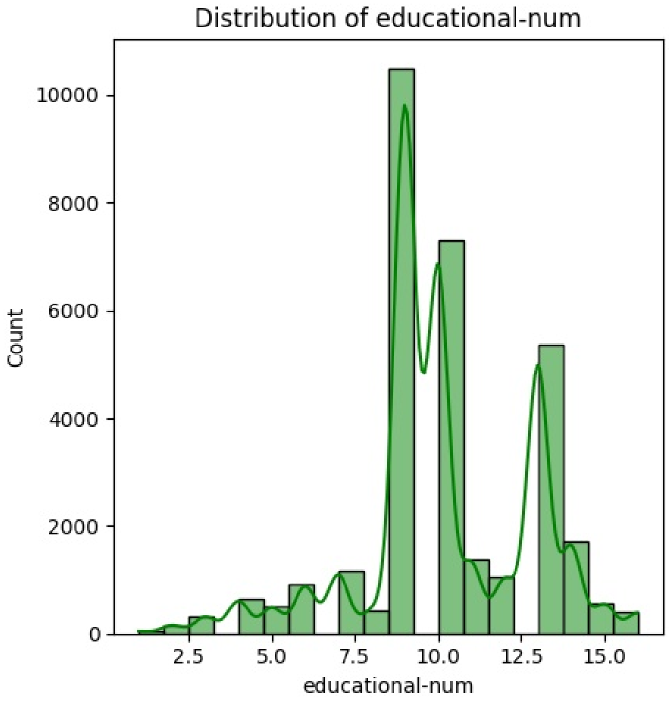


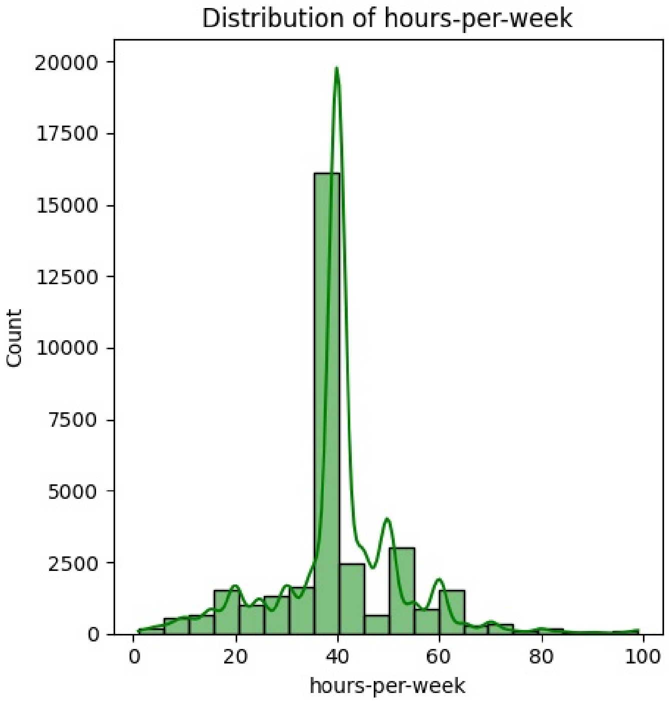
**Histogram for distribution of education-num and hours-per-week**

In [12]:

```
features = ['educational-num', 'hours-per-week']
plt.figure(figsize=(14, 5))
for i, feature in enumerate(features, 1):
    plt.subplot(1, 3, i)
    sns.histplot(Data[feature].dropna(), bins=20, kde=True, color='green')
    plt.title(f'Distribution of {feature}')

plt.tight_layout()
plt.show()
```





Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js