Exploratory Report : UCI adult census data

Dataset: [UCI adult census dataset]

Initial Highlights:

Income Distribution

- **Income Classification:** The dataset classifies income into two categories: `<=50K` and `>50K`. A significant portion of the population falls into the `<=50K` category, indicating income inequality.

Demographic Breakdown

- **Sex:** Males tend to have a higher representation in the `>50K` income category compared to females, indicating a gender income disparity.
- **Race:** The dataset reveals racial disparities in income. For instance, White individuals are more likely to be in the `>50K` income category compared to other races.

Age

- **Age vs. Income:** There's a positive correlation between age and income up to a certain point, after which income levels tend to plateau or decrease. This trend suggests that income generally increases with experience and age until retirement approaches.

Education:

- **Education Level:**Higher education levels are strongly correlated with higher income. Individuals with advanced degrees (e.g., Doctorate, Masters) are more likely to fall into the `>50K` income category.
- **Education vs. Sex:** Males generally earn more than females at each education level, highlighting a persistent gender pay gap even among highly educated individuals.

Work Hours

- **Hours per Week:** Those who work more hours per week tend to earn higher incomes. However, there's a threshold beyond which additional hours do not significantly increase the probability of earning `>50K`.
- **Sex vs. Work Hours:** Males tend to work more hours per week than females on average, which could contribute to the observed income disparities.

Occupation and Workclass

- **Occupation:** Certain occupations, such as those in executive or managerial roles, have a higher likelihood of earning `>50K`.

- **Workclass:** Individuals working in the private sector or self-employed tend to earn more compared to those in government jobs or other sectors.

Marital Status

- **Marital Status vs. Income**:Married individuals (especially those married with a spouse present) are more likely to be in the `>50K` income category. This could be due to combined household incomes or the economic stability that marriage can provide.

Capital Gains

- **Capital Gains:** Higher capital gains are strongly associated with higher income levels. Individuals with significant capital gains are more likely to earn `>50K`.

Geographic Distribution

- **Native Country:** Individuals born in certain countries (e.g., United States) are more likely to earn higher incomes compared to immigrants from other countries, reflecting potential economic opportunities and systemic inequalities.

Sex and Relationship

- **Relationship Status:** The relationship status (e.g., Husband, Wife) impacts income. For example, husbands are more likely to earn `>50K`, which may reflect traditional gender roles and income distribution within households.

About the Dataset

- Has a total of 32561 data points and 15 features
    - Features are: "age", "workclass", "fnlwgt", "education", "education-num", "marital-status", "occupation", "relationship", "race", "sex", "capital-gain", "capital-loss", "hours-per-week", "native-country", "income".
    - Workclass, Occupation and native-country have null or missing values.
    - Of these missing values 1836 values are missing in the row in both Workclass as well as occupation columns. And 27 rows have all 3 of these values missing.
    - A total 2399 rows have missing values.
- Handling missing values, Substituting categorical values with most frequent occurring variables had a negative impact on model ~-6%, not handling the missing values gave an accuracy of ~83% and post handling dropped to ~76%.
- Highest accuracy achieved by removing all rows with missing values and training the model, accuracy ~85%. Hence this option was opted for.
- 4 features in this dataset are natively of the numerical type, Capital-gain, capital-loss, hours-per-week, education-num, age. The following are descriptive statistics corresponding to each feature.

- **Age:**
  - count 32561.000000
  - mean 38.581647
  - std 13.640433
  - min 17.000000
  - 25% 28.000000
  - 50% 37.000000
  - 75% 48.000000
  - max 90.000000

- **Capital Gain:**
  - count 32561.000000
  - mean 1077.648844
  - std 7385.292085
  - min 0.000000
  - 25% 0.000000
  - 50% 0.000000
  - 75% 0.000000
  - max 99999.000000

- **Capital Loss:**
  - count 32561.000
  - mean 87.303830
  - std 402.96021
  - min 0.000000
  - 25% 0.000000
  - 50% 0.000000
  - 75% 0.000000
  - max 4356.0000

- **Hours per week:**
  - count 32561.000
  - mean 40.437456
  - std 12.347429
  - min 1.000000
  - 25% 40.000000
  - 50% 40.000000
  - 75% 45.000000
  - max 99.000000

- **Education Num:**
  - count 32561.000
  - mean 10.080679
  - std 2.572720
  - min 1.000000
  - 25% 9.000000
  - 50% 10.000000
  - 75% 12.000000
  - max 16.000000

- The following is the frequency Distribution for the categorical features:
    - Frequency distribution for **workclass**:
        - Private 22696
        - Self-emp-not-inc 2541
        - Local-gov 2093
        - State-gov 1298
        - Self-emp-inc 1116
        - Federal-gov 960
        - Without-pay 14
        - Never-worked 7

    - Frequency distribution for **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

    - Frequency distribution for **marital-status**:
        - Married-civ-spouse 14976
        - Never-married 10683
        - Divorced 4443
        - Separated 1025
        - Widowed 993
        - Married-spouse-absent 418
        - Married-AF-spouse 23
        - Frequency distribution for **occupation**:
        - Prof-specialty 4140
        - Craft-repair 4099
        - Exec-managerial 4066
        - Adm-clerical 3770
        - Sales 3650
        - Other-service 3295
        - Machine-op-inspct 2002
        - Transport-moving 1597
        - Handlers-cleaners 1370
        - Farming-fishing 994
        - Tech-support 928

        - Protective-serv 649
        - Priv-house-serv 149
        - Armed-Forces 9

    - Frequency distribution for **Relationship**:
        - Husband 13193
        - Not-in-family 8305
        - Own-child 5068
        - Unmarried 3446
        - Wife 1568
        - Other-relative 981

    - Frequency distribution for **Race**:
        - White 27816
        - Black 3124
        - Asian-Pac-Islander 1039
        - Amer-Indian-Eskimo 311
        - Other 271

    - Frequency distribution for **Sex**
        - Male 21790
        - Female 10771

    - Frequency distribution for **Native-country**
        - United-States 29170
        - Mexico 643
        - 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
- Trinadad&Tobago 19
- Cambodia 19
- Thailand 18
- Laos 18
- Yugoslavia 16
- Outlying-US 14
- Honduras 13
- Hungary 13
- Scotland 12
- Holand-Netherlands 1

Key Relationships and Patterns

In this section, we delve into the relationships between key variables in the dataset and uncover significant patterns. These analyses are crucial for understanding the factors that influence income levels and other demographic characteristics.

Income vs. Demographics

Income vs. Age:

- Observation: There is a positive correlation between age and income up to a certain age, after which income levels tend to plateau or decrease.
- Explanation: This trend suggests that income generally increases with experience and age, peaking around middle age, and then stabilizes or declines as individuals approach retirement.

Income vs. Sex:

- Observation: Males are more likely to earn `>50K` compared to females.
- Explanation: This indicates a gender income disparity, potentially due to differences in employment opportunities, work hours, and societal roles.

Income vs. Race:

- Observation: White individuals have a higher likelihood of earning `>50K` compared to other races.
- Explanation: This suggests racial disparities in income, possibly reflecting systemic inequalities and differences in access to opportunities.

Income vs. Education:

- Observation: Higher education levels are strongly correlated with higher income.
- Explanation: Individuals with advanced degrees are more likely to earn `>50K`, highlighting the importance of education in achieving higher income levels.

Income vs. Marital Status:
Observation: Married individuals, especially those married with a spouse present, are more likely to earn `>50K`.
Explanation: This could be due to combined household incomes or the economic stability that marriage can provide.

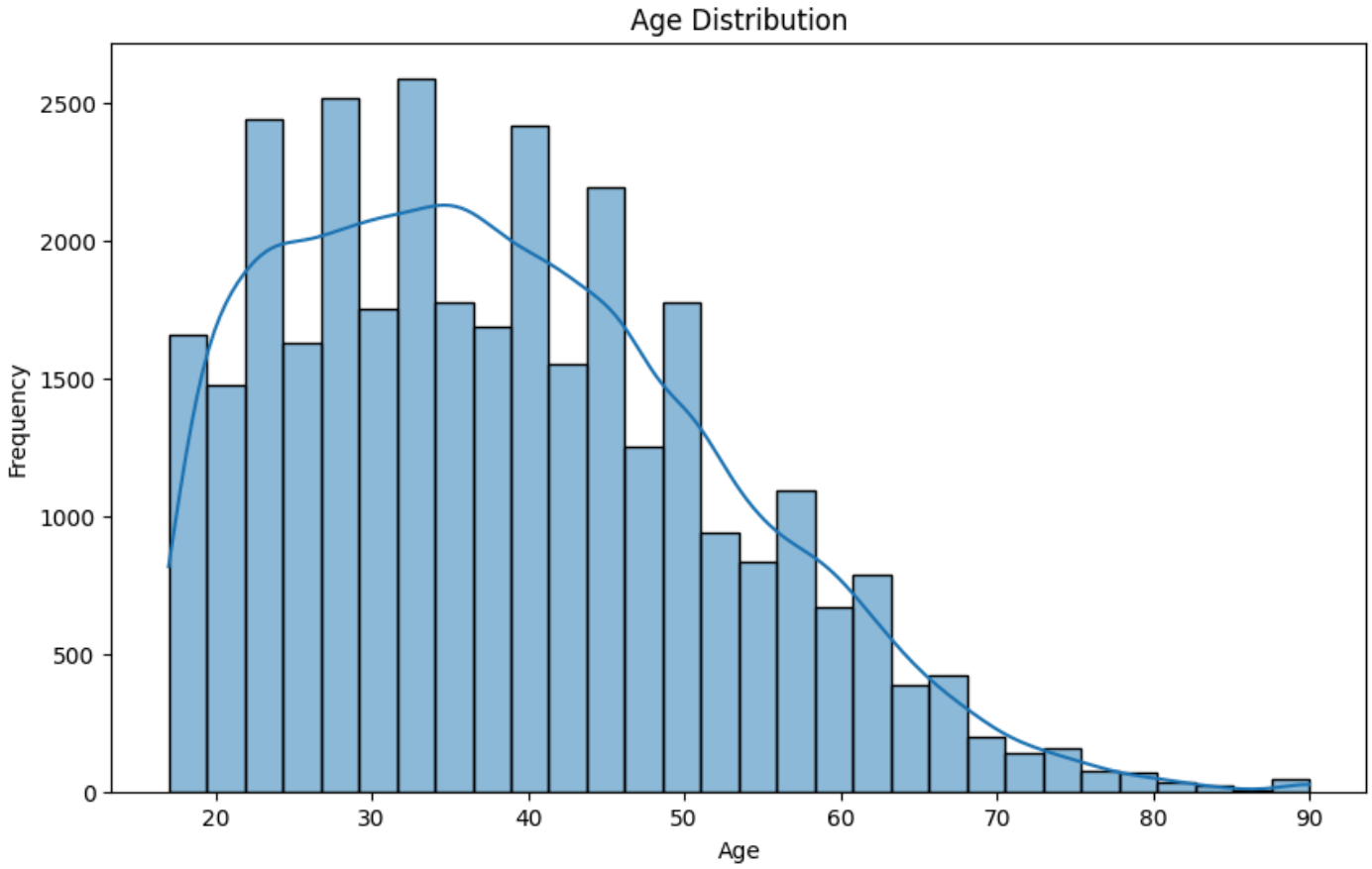
Income vs. Occupation

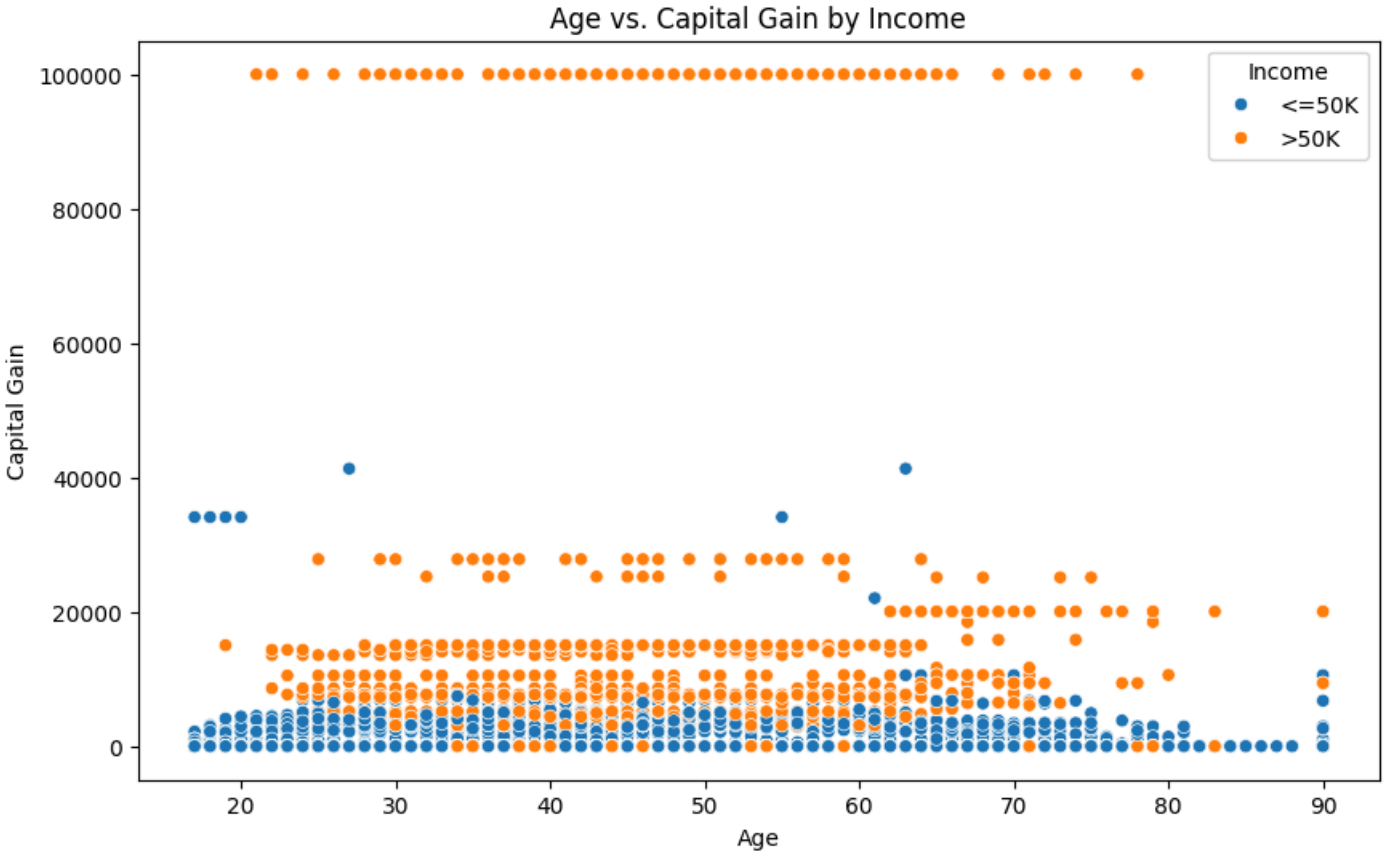
Observation: Certain occupations, such as those in executive or managerial roles, have a higher likelihood of earning `>50K`.
Explanation: These roles often come with higher responsibilities and compensation, reflecting the income disparity across different job types.

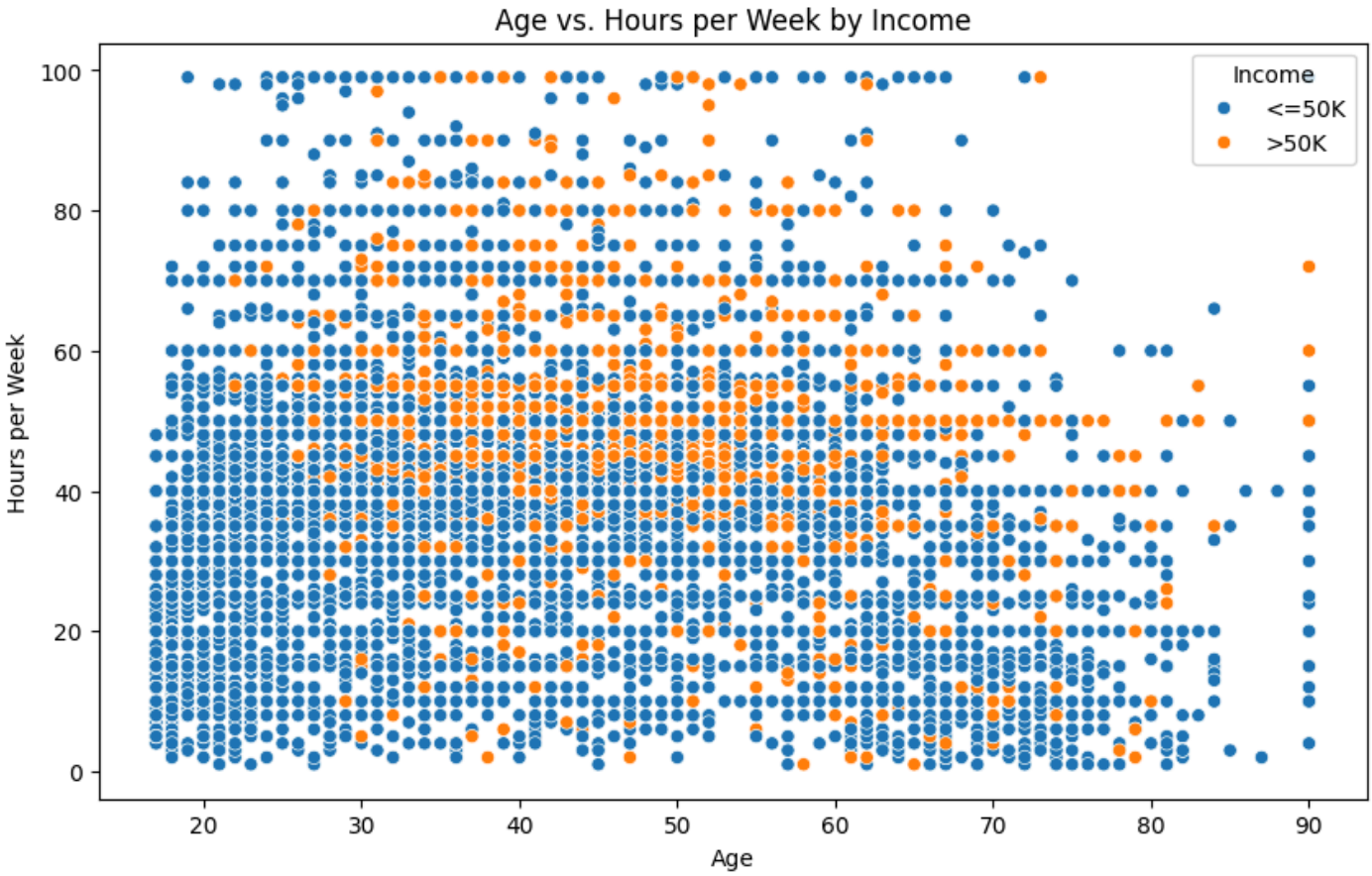
Work Hours vs. Income

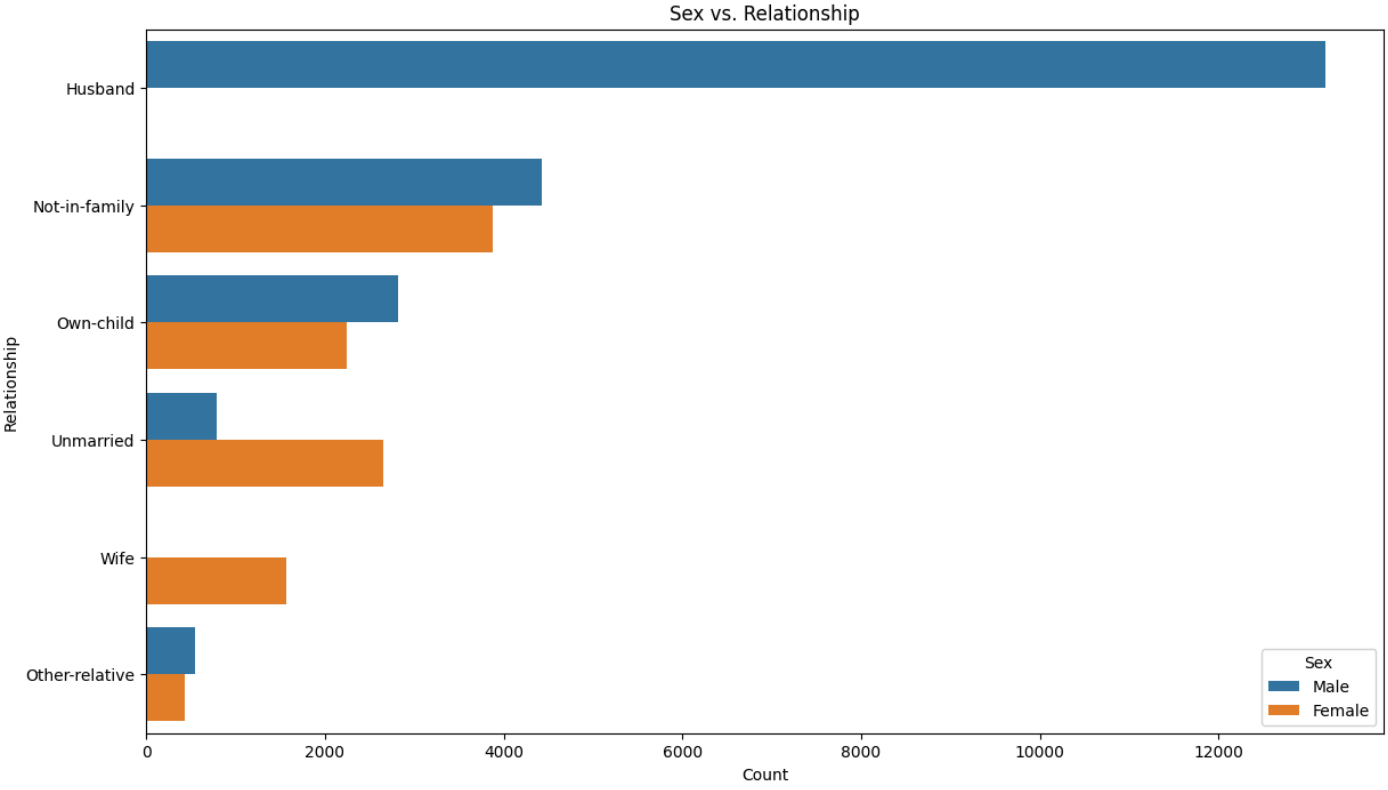
Observation: Individuals who work more hours per week tend to earn higher incomes.
Explanation: There is a positive correlation between work hours and income, indicating that working longer hours can lead to higher earnings. However, there is a threshold beyond which additional hours do not significantly increase the probability of earning `>50K`.

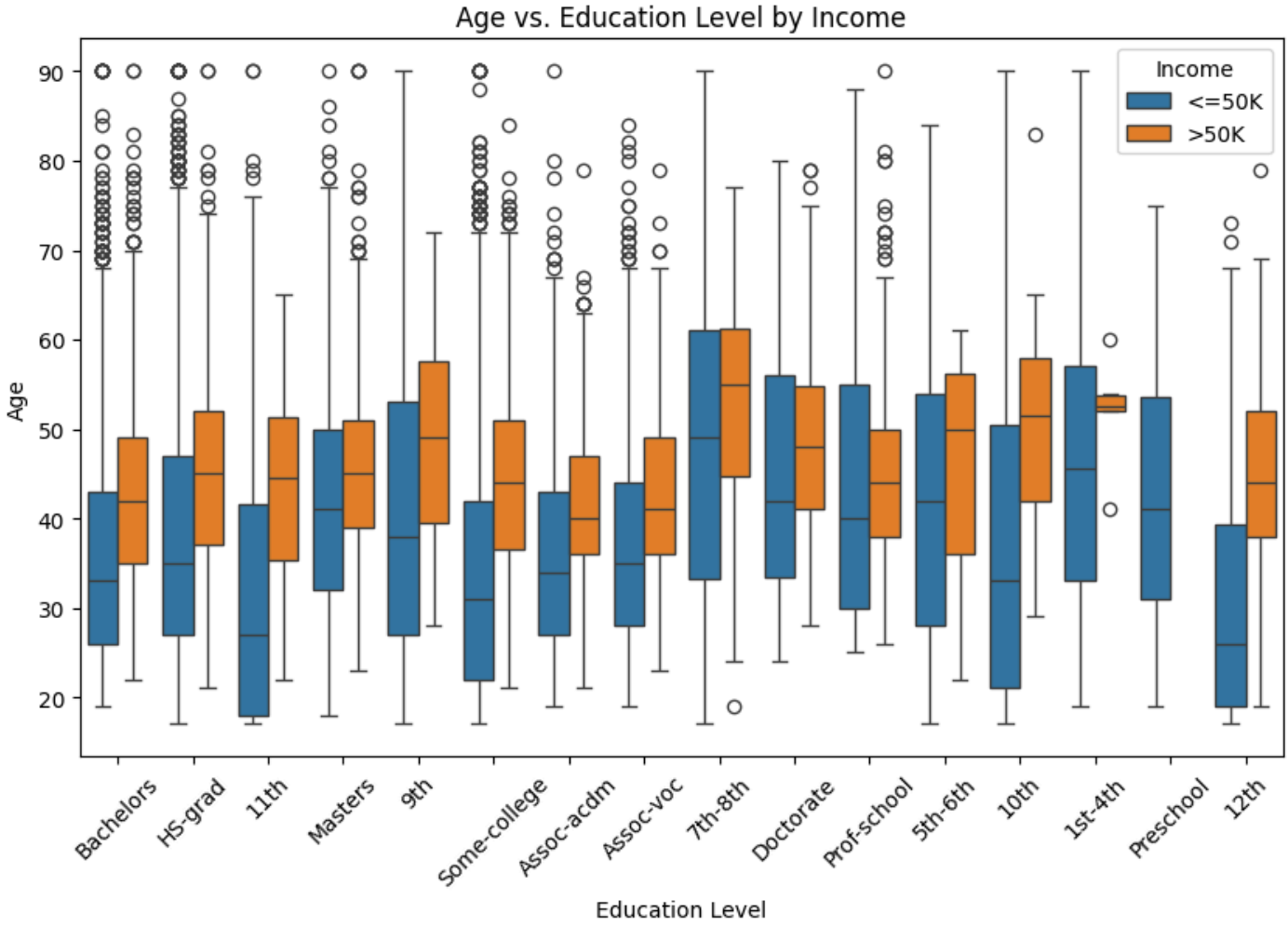
Visualizations for further understanding:

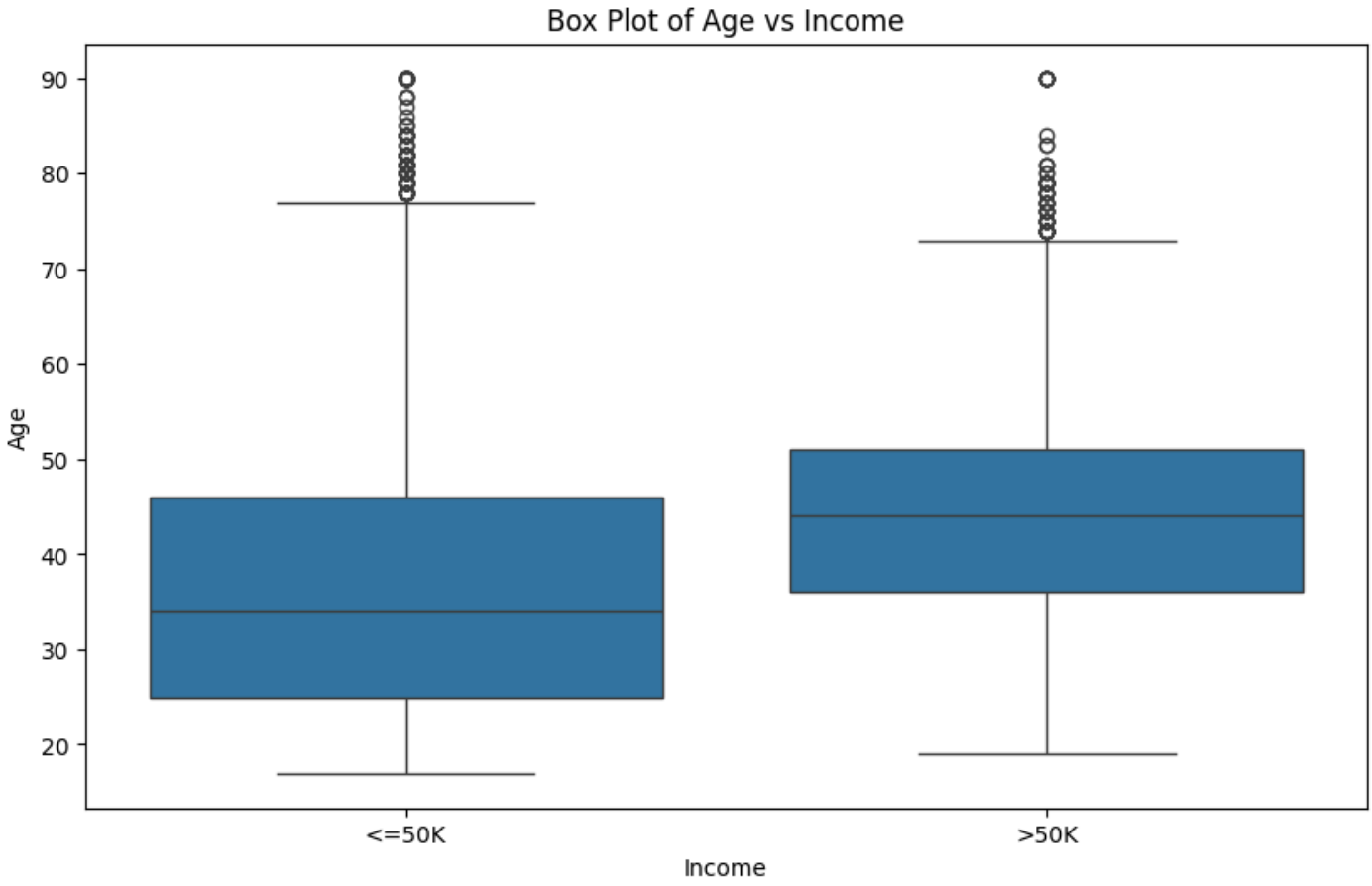


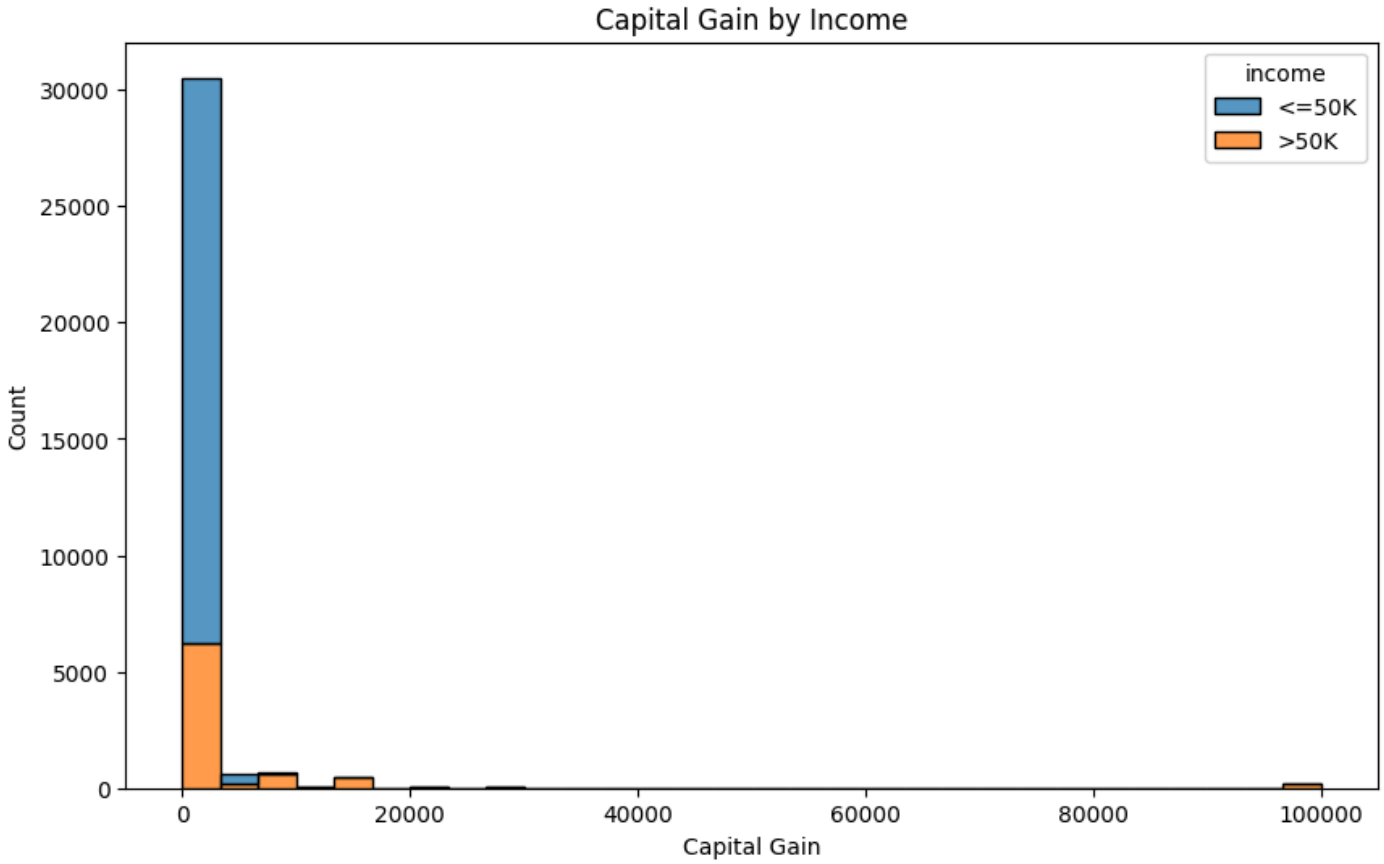


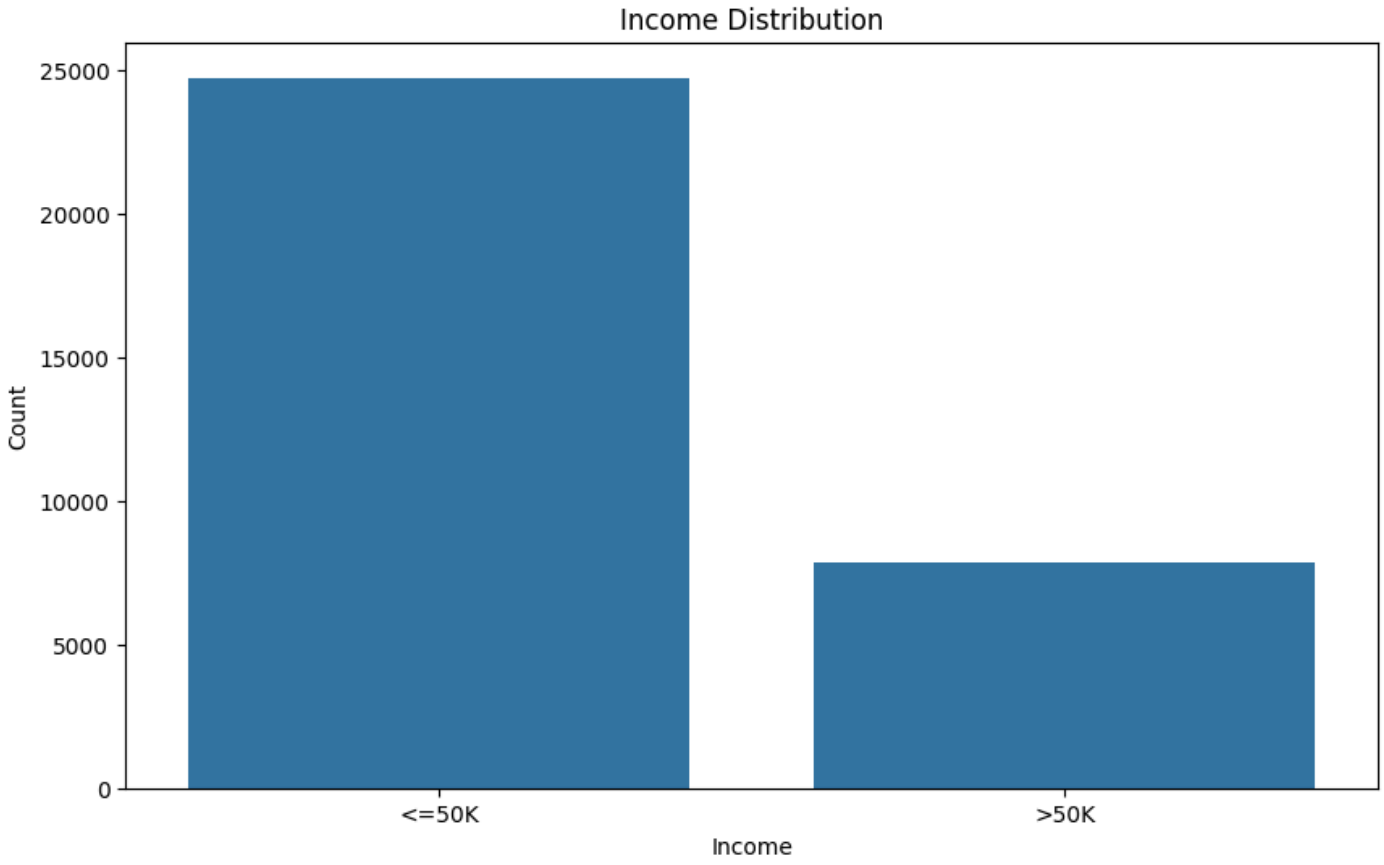

Age vs. Hours per Week by Income
Income
<=50K
>50K
Hours per Week
Age
0
20
40
60
80
100
20
30
40
50
60
70
80
90

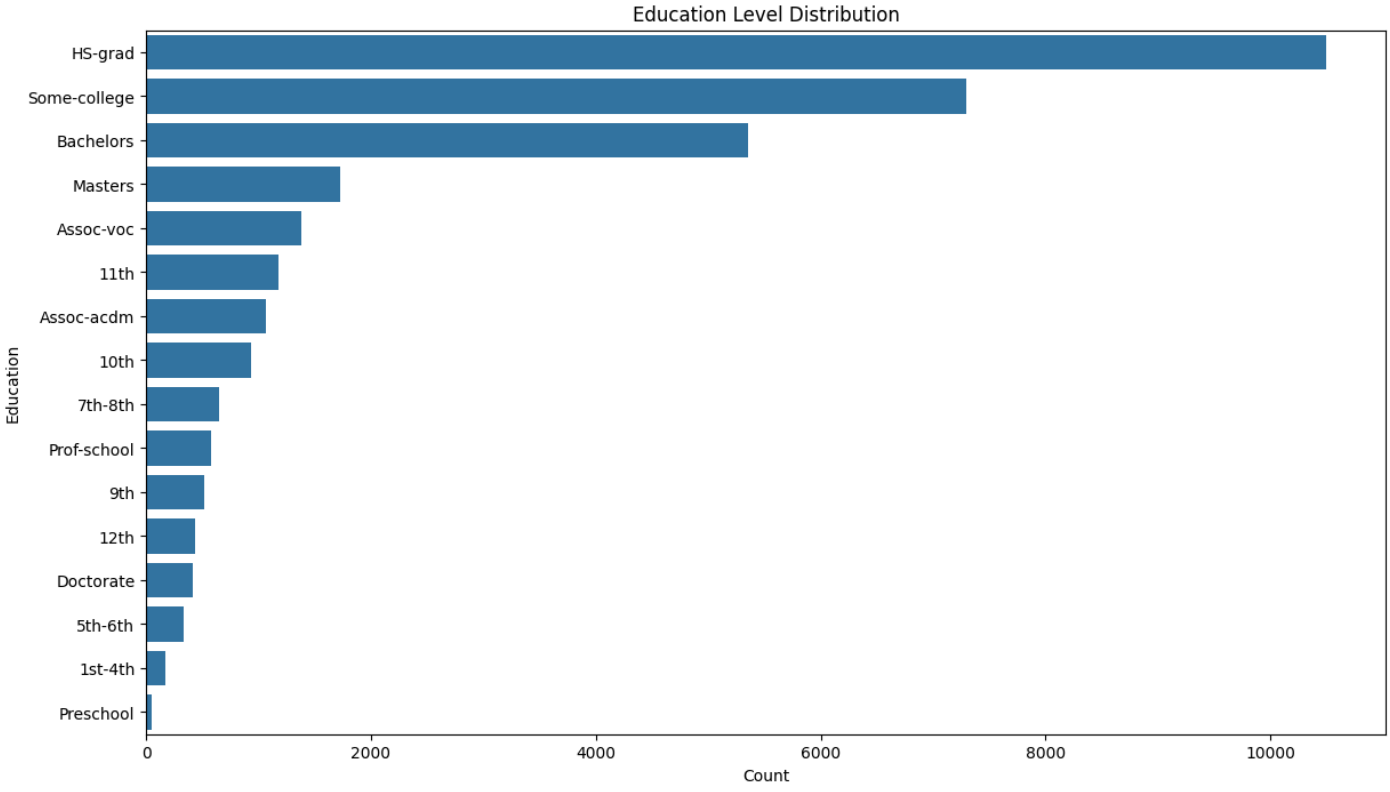

Sex vs. Relationship
Husband
Not-in-family
Own-child
Unmarried
Wife
Other-relative
Relationship
Sex
Male
Female
0
2000
4000
6000
8000
10000
12000
Count

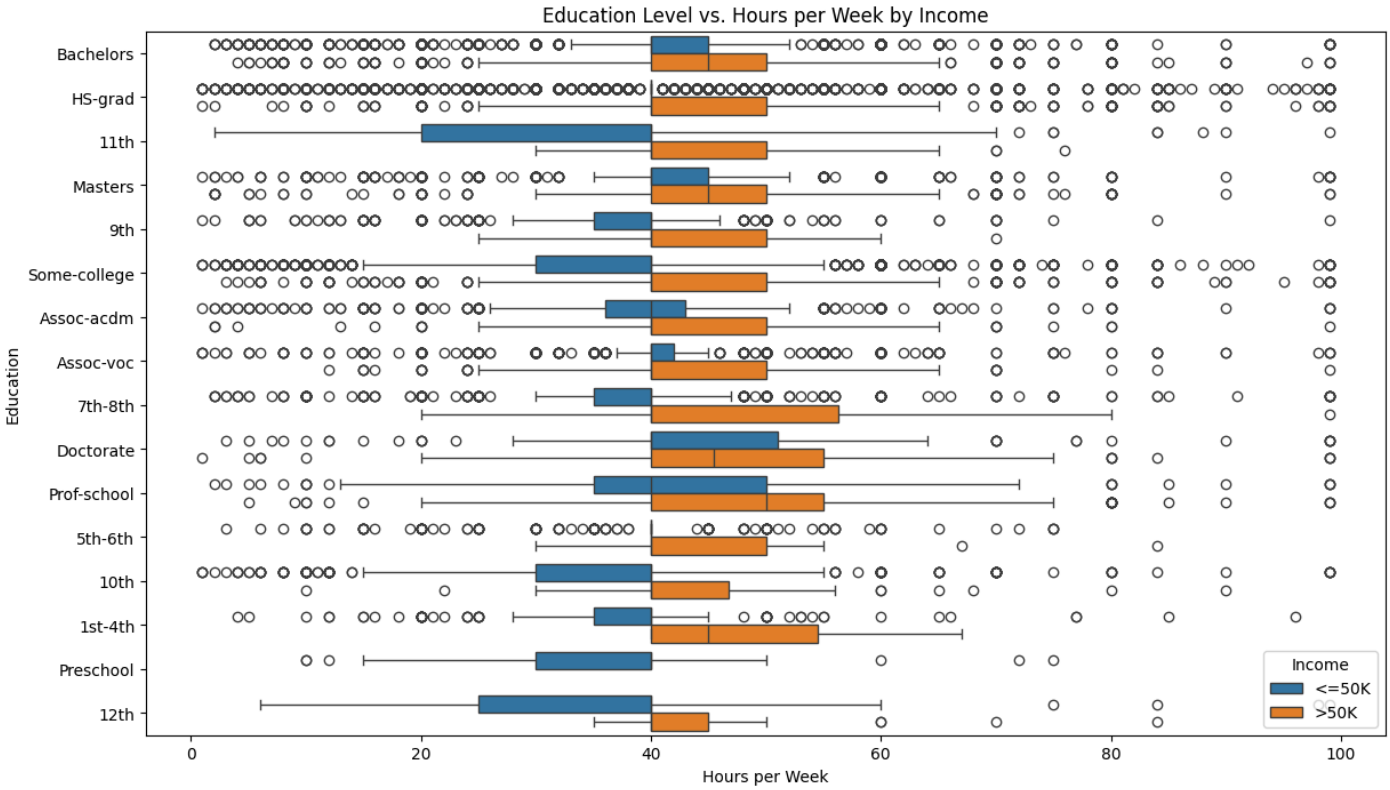

Age vs. Education Level by Income
Income
<=50K
>50K
Age
90
80
70
60
50
40
30
20
Bachelors
HS-grad
11th
Masters
9th
Some-college
Assoc-acdm
Assoc-voc
7th-8th
Doctorate
Prof-school
5th-6th
10th
1st-4th
Preschool
12th
Education Level

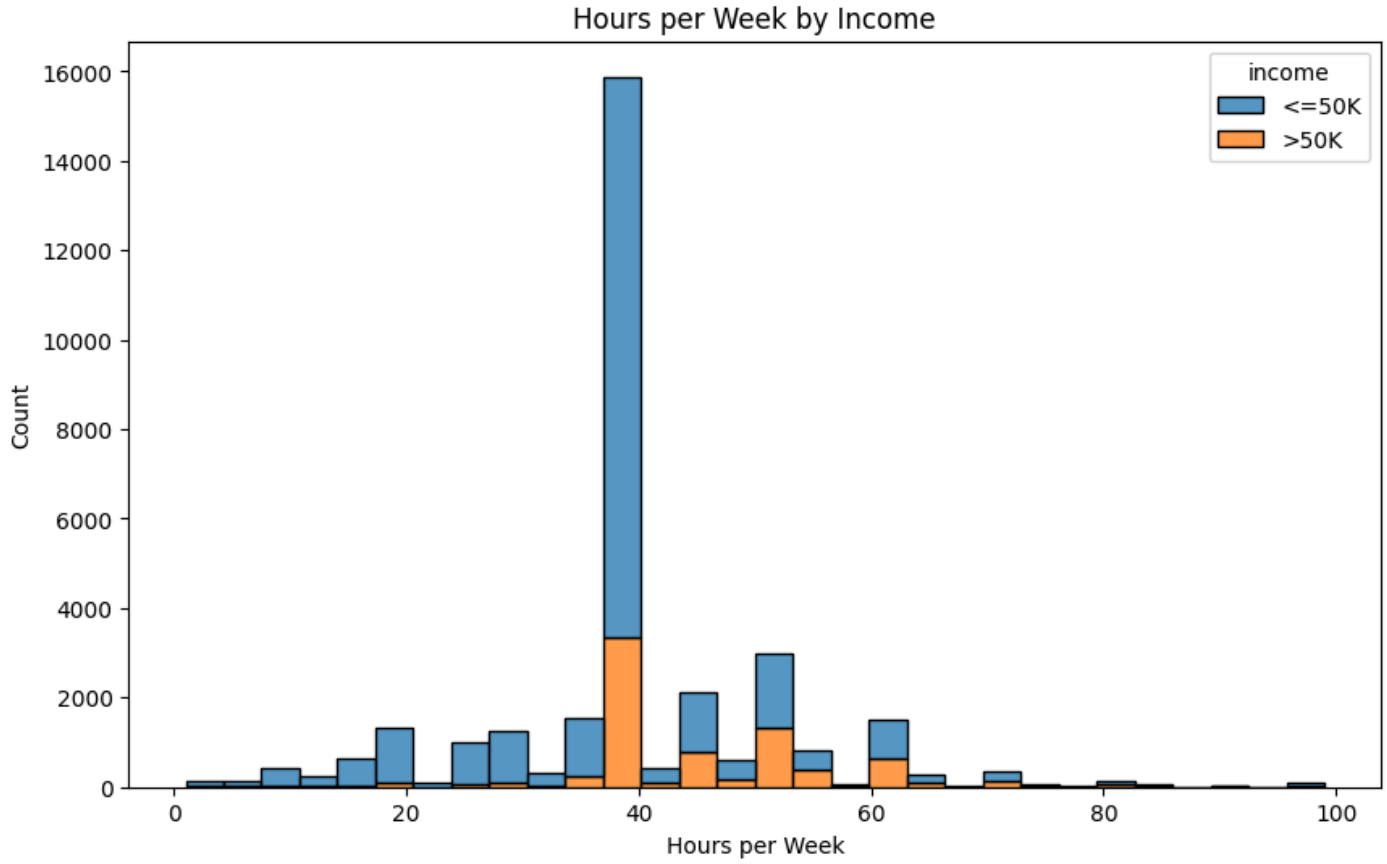

Box Plot of Age vs Income
Age
Income
<=50K
>50K
90
80
70
60
50
40
30
20

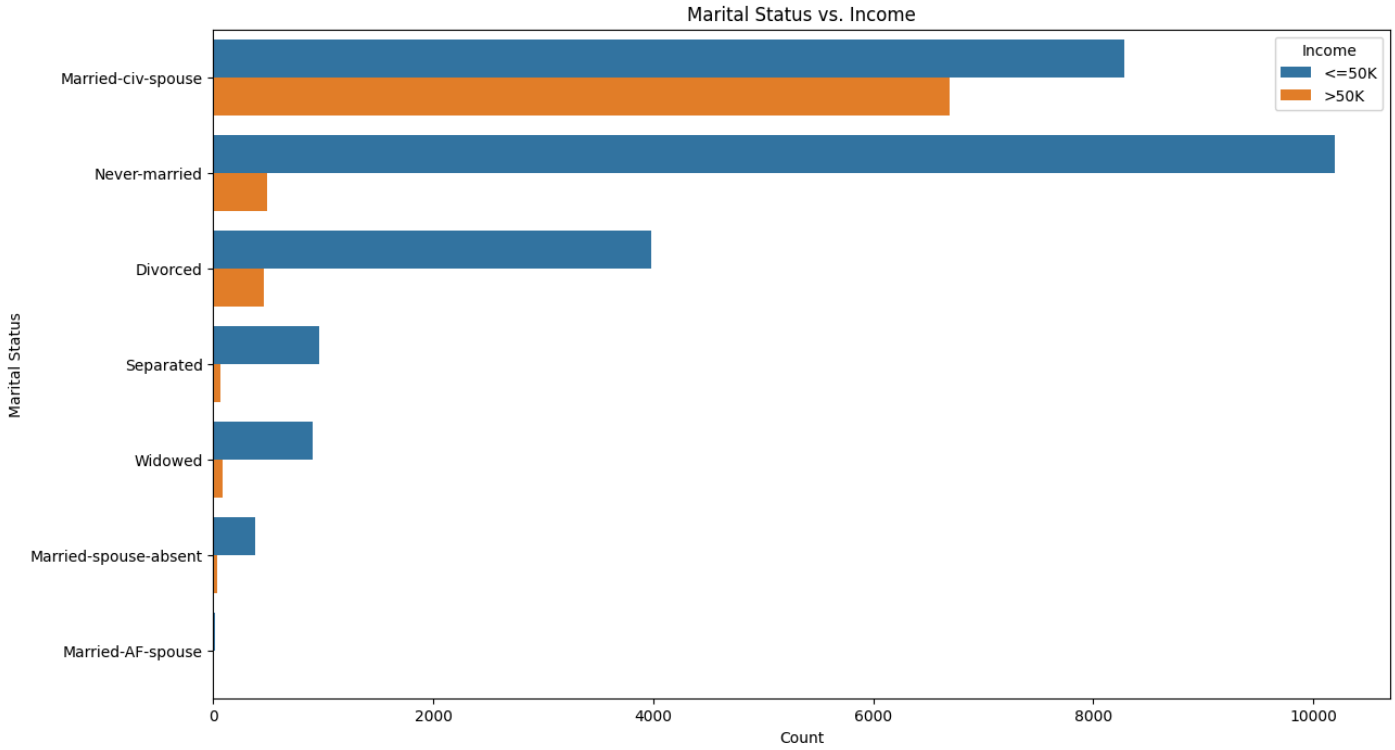

Capital Gain by Income
income
<=50K
>50K
Count
Capital Gain
30000
25000
20000
15000
10000
5000
0
0
20000
40000
60000
80000
100000

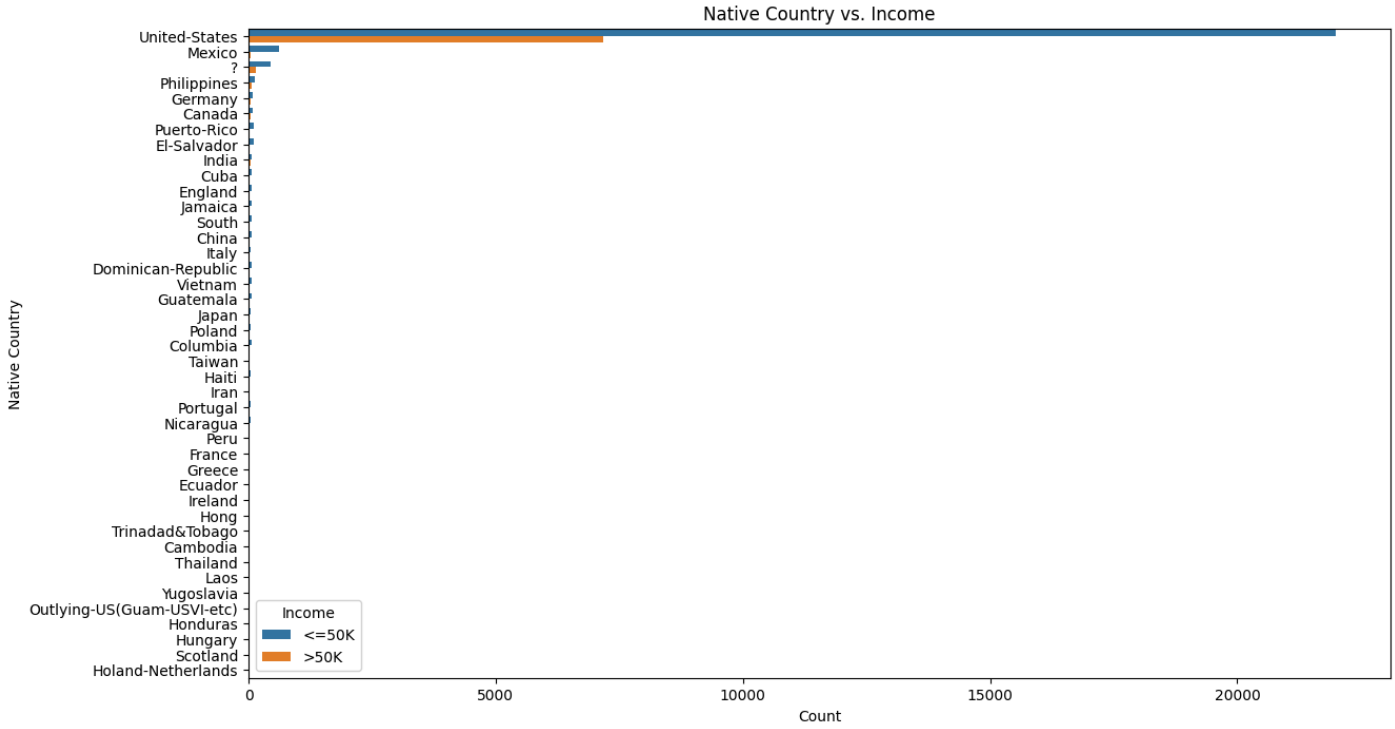

Income Distribution
25000
20000
15000
10000
5000
0
Count
<=50K
>50K
Income

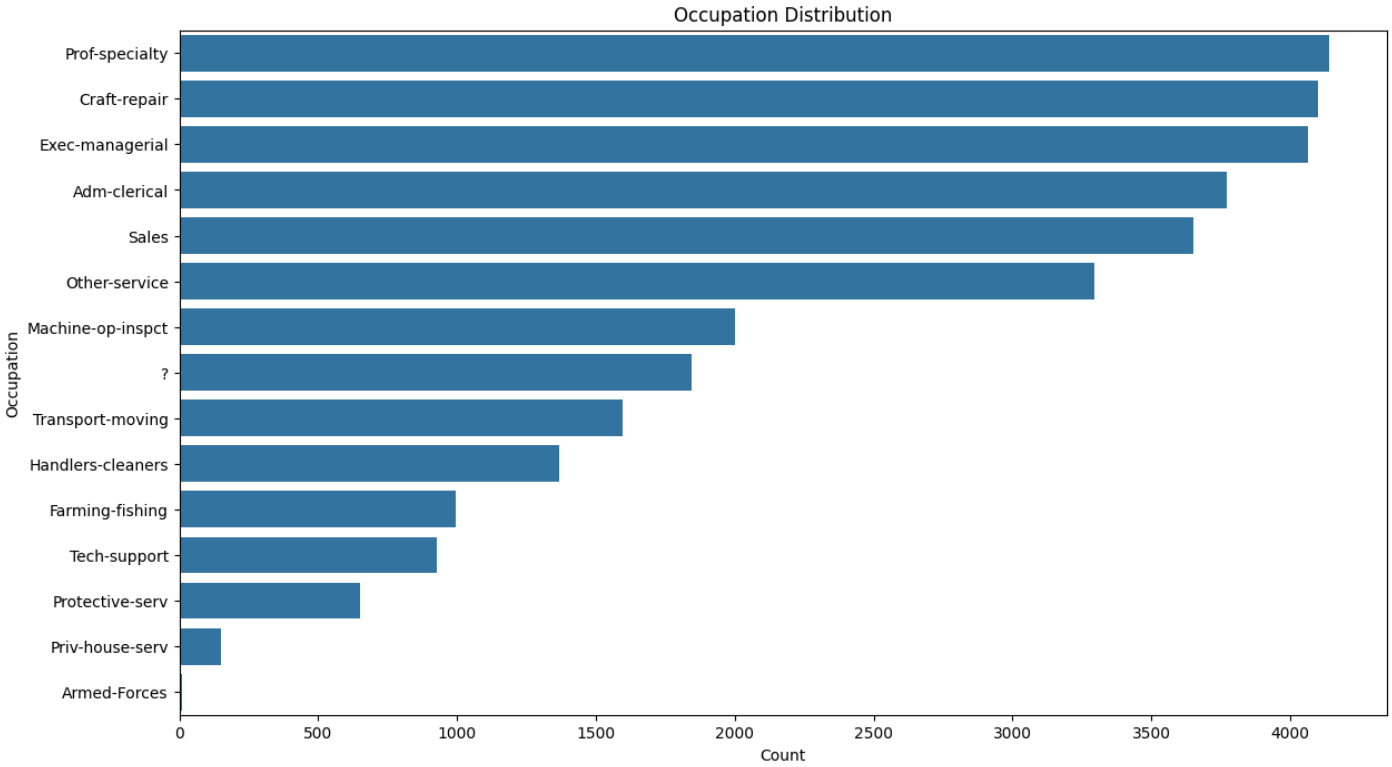

Education Level Distribution
HS-grad
Some-college
Bachelors
Masters
Assoc-voc
11th
Assoc-acdm
10th
7th-8th
Prof-school
9th
12th
Doctorate
5th-6th
1st-4th
Preschool
Education
0
2000
4000
6000
8000
10000
Count

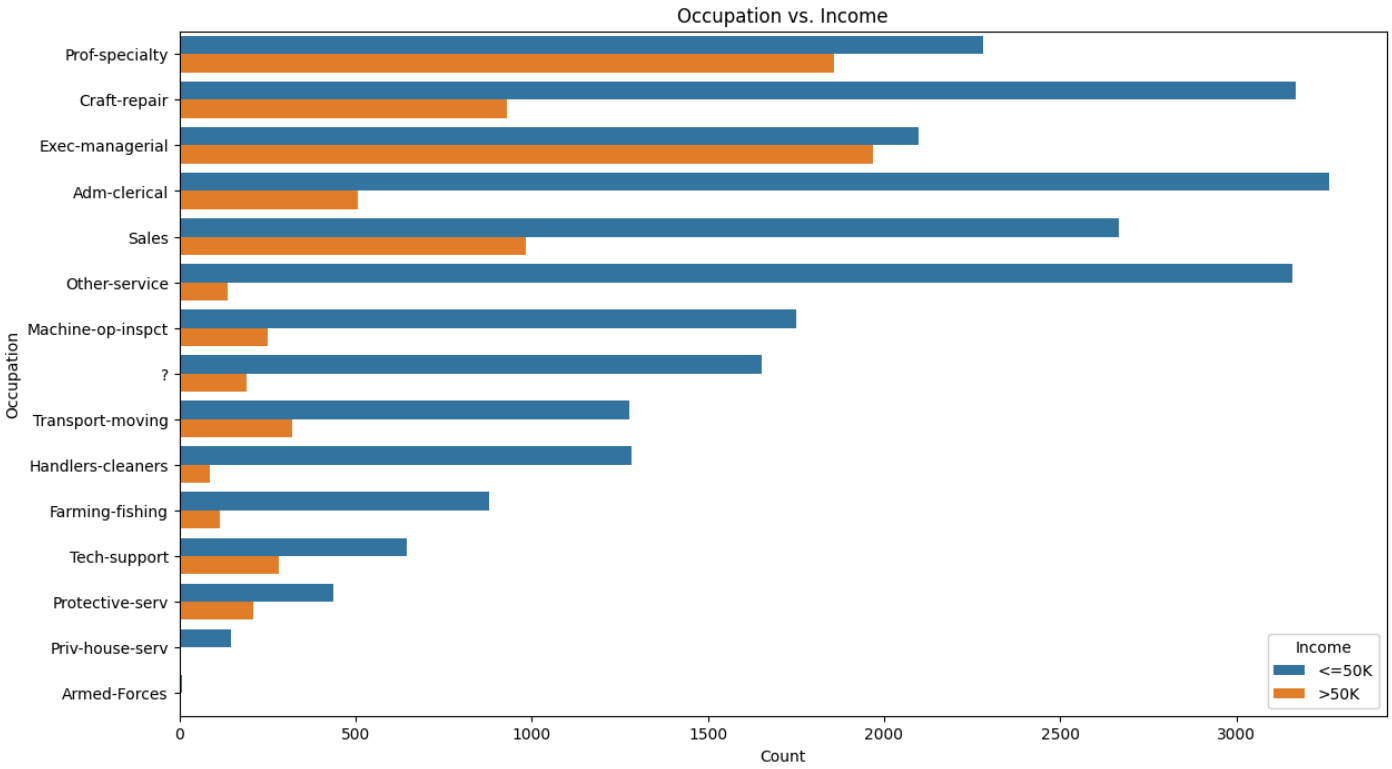

Education Level vs. Hours per Week by Income
Bachelors
HS-grad
11th
Masters
9th
Some-college
Assoc-acdm
Assoc-voc
7th-8th
Doctorate
Prof-school
5th-6th
10th
1st-4th
Preschool
12th
Education
0
20
40
60
80
100
Hours per Week
Income
<=50K
>50K

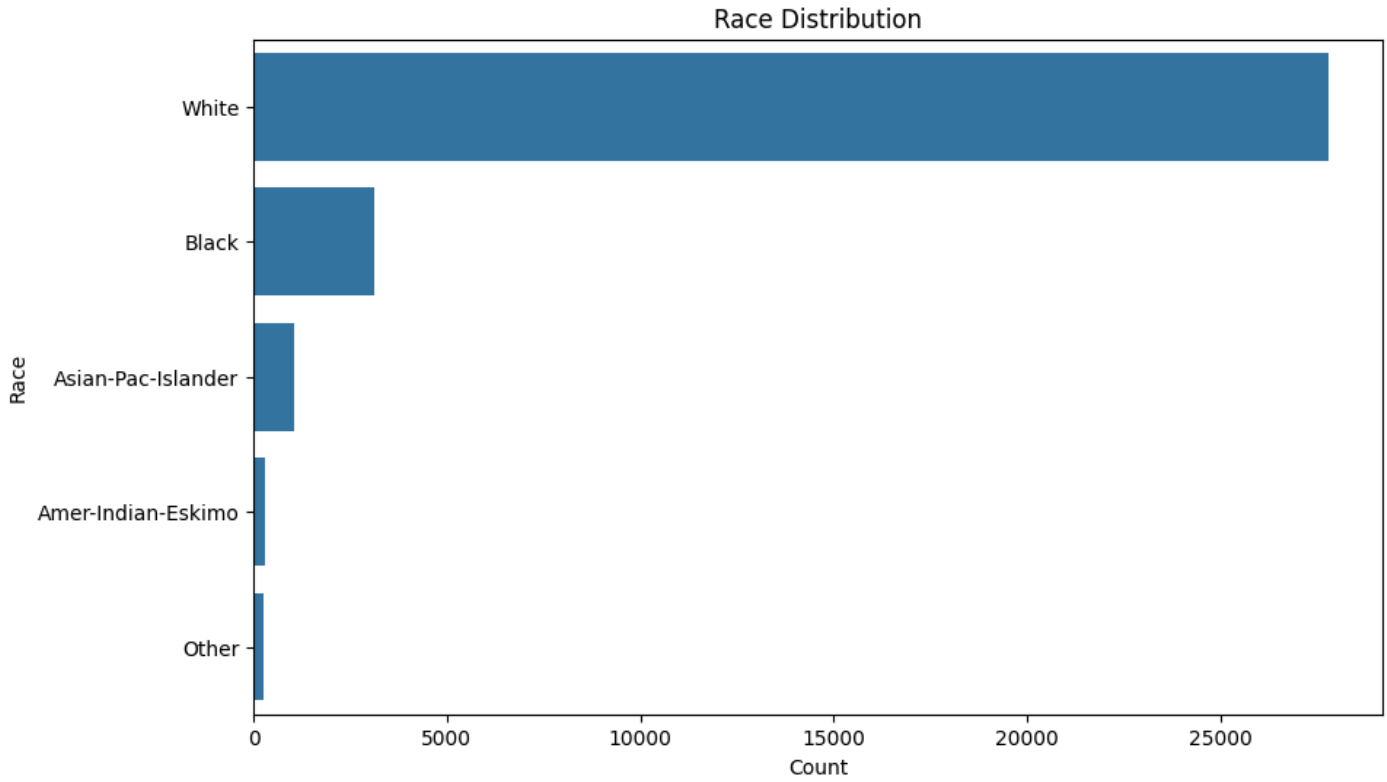

Hours per Week by Income
income
<=50K
>50K
Count
16000
14000
12000
10000
8000
6000
4000
2000
0
0
20
40
60
80
100
Hours per Week

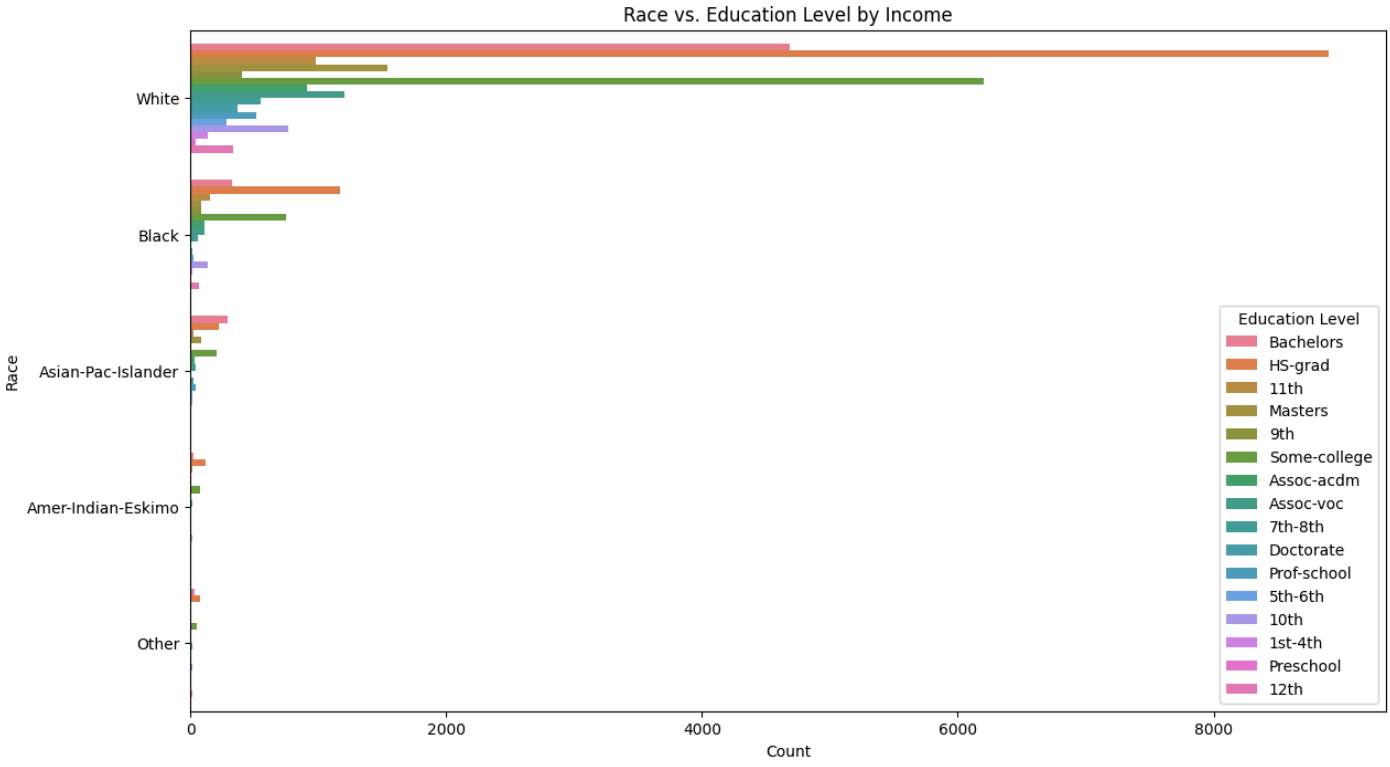

Marital Status vs. Income
Income
<=50K
>50K
Married-civ-spouse
Never-married
Divorced
Separated
Widowed
Married-spouse-absent
Married-AF-spouse
Marital Status
0
2000
4000
6000
8000
10000
Count

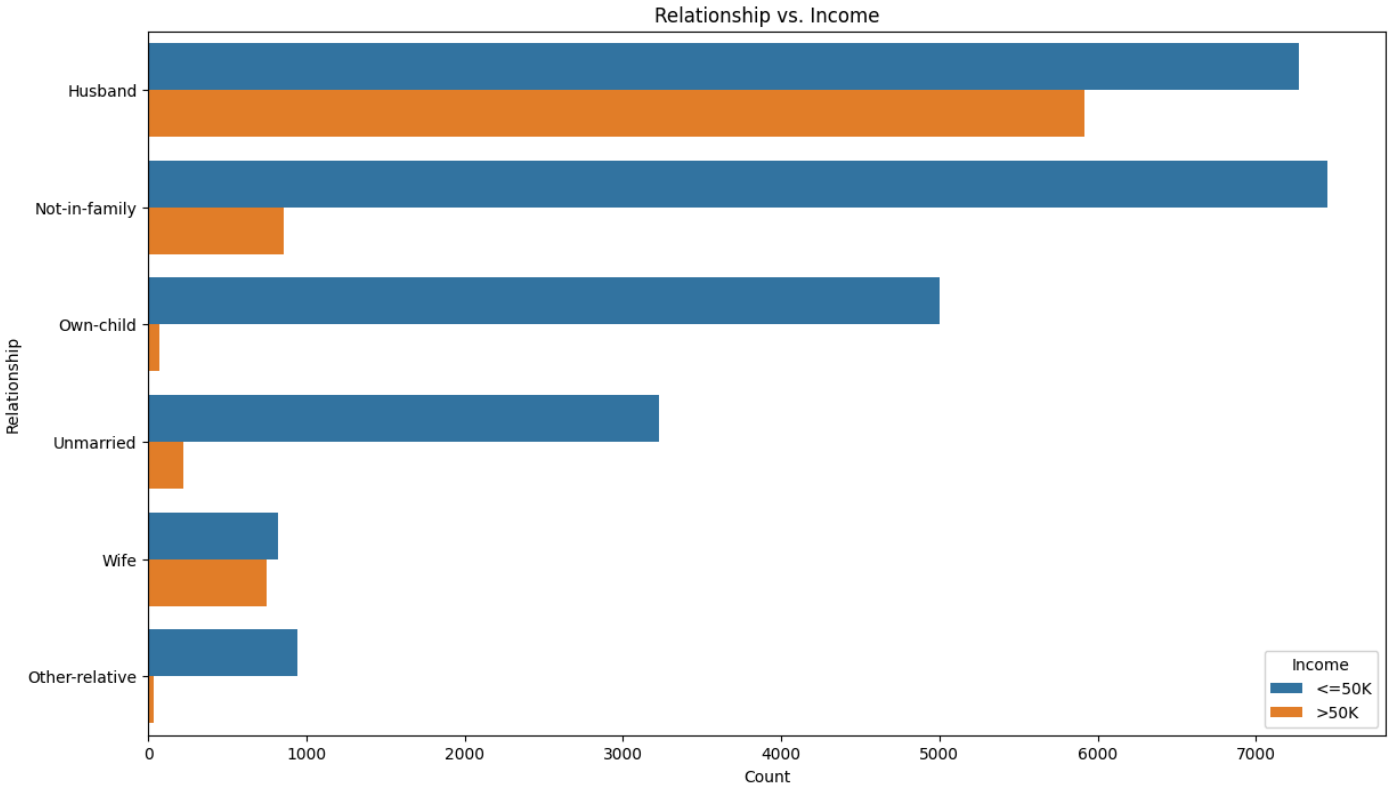

Native Country vs. Income
United-States
Mexico
?
Philippines
Germany
Canada
Puerto-Rico
El-Salvador
India
Cuba
England
Jamaica
South
China
Italy
Dominican-Republic
Vietnam
Guatemala
Japan
Poland
Columbia
Taiwan
Haiti
Iran
Portugal
Nicaragua
Peru
France
Greece
Ecuador
Ireland
Hong
Trinadad&Tobago
Cambodia
Thailand
Laos
Yugoslavia
Outlying-US(Guam-USVI-etc)
Honduras
Hungary
Scotland
Holand-Netherlands
Income
<=50K
>50K
Native Country
0
5000
10000
15000
20000
Count

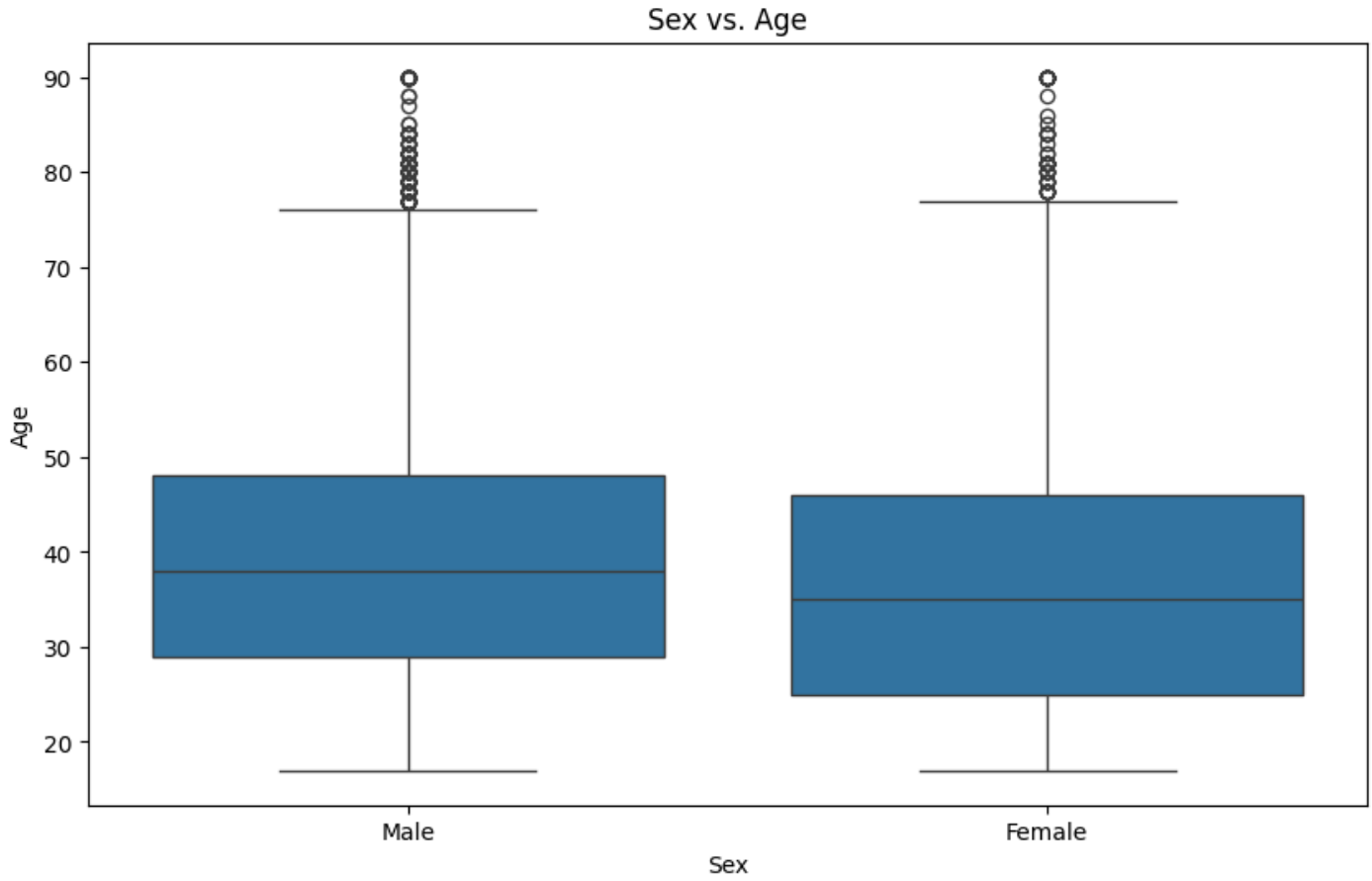

Occupation Distribution
Prof-specialty
Craft-repair
Exec-managerial
Adm-clerical
Sales
Other-service
Machine-op-inspct
?
Transport-moving
Handlers-cleaners
Farming-fishing
Tech-support
Protective-serv
Priv-house-serv
Armed-Forces
Occupation
0
500
1000
1500
2000
2500
3000
3500
4000
Count

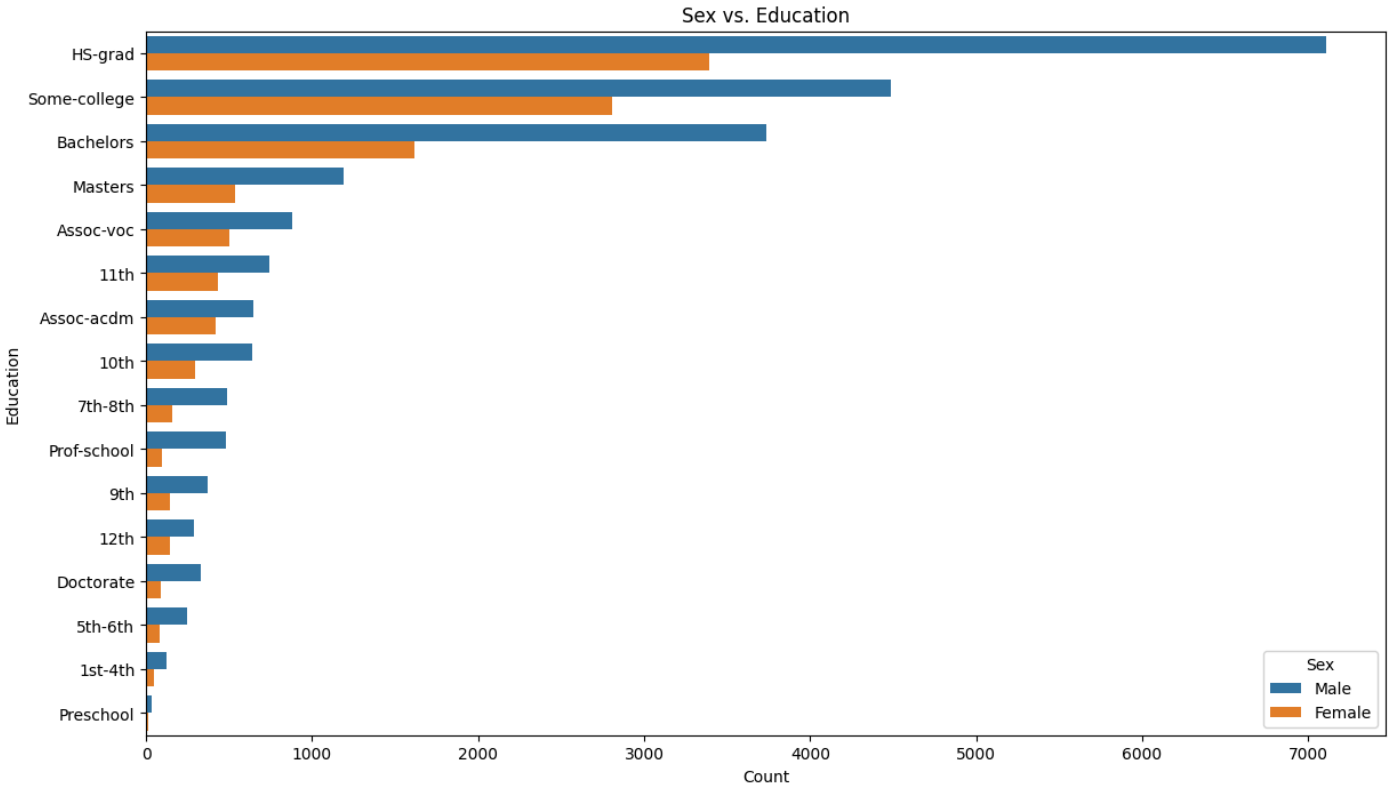

Occupation vs. Income
Prof-specialty
Craft-repair
Exec-managerial
Adm-clerical
Sales
Other-service
Machine-op-inspct
?
Transport-moving
Handlers-cleaners
Farming-fishing
Tech-support
Protective-serv
Priv-house-serv
Armed-Forces
Occupation
Income
<=50K
>50K
0
500
1000
1500
2000
2500
3000
Count

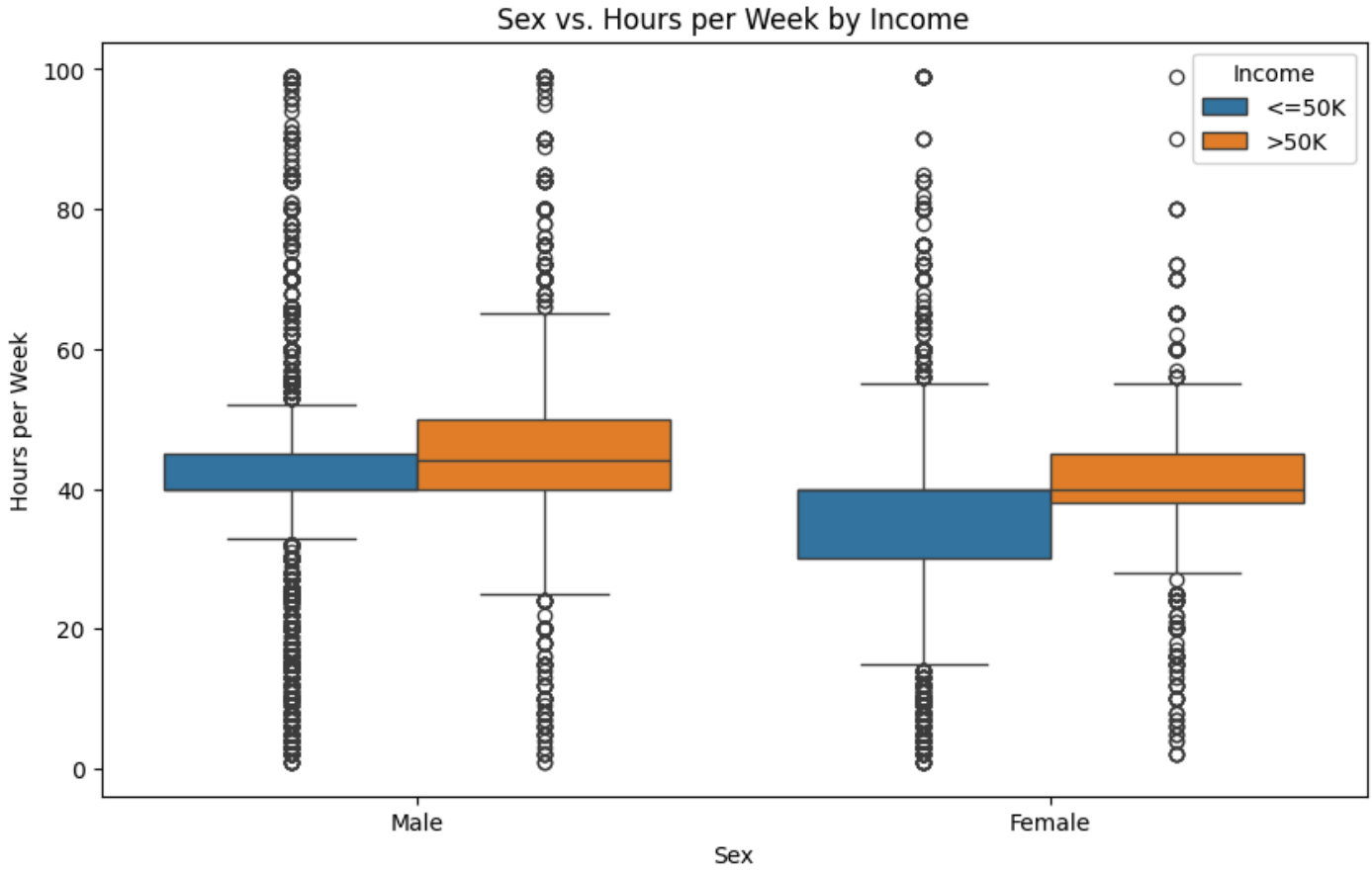

Race Distribution
White
Black
Asian-Pac-Islander
Amer-Indian-Eskimo
Other
Race
0
5000
10000
15000
20000
25000
Count

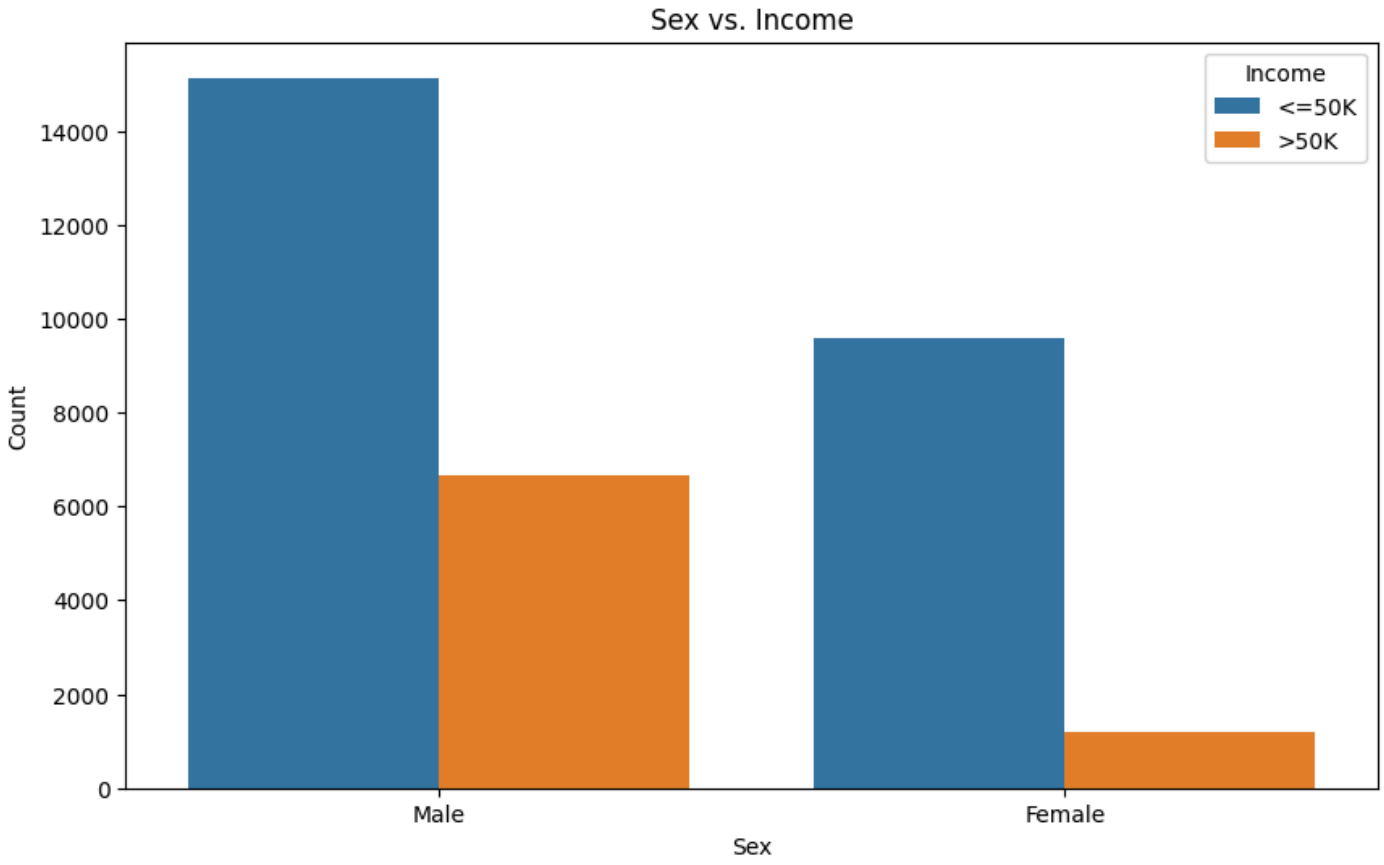

Race vs. Education Level by Income
White
Black
Asian-Pac-Islander
Amer-Indian-Eskimo
Other
Race
Education Level
Bachelors
HS-grad
11th
Masters
9th
Some-college
Assoc-acdm
Assoc-voc
7th-8th
Doctorate
Prof-school
5th-6th
10th
1st-4th
Preschool
12th
0
2000
4000
6000
8000
Count

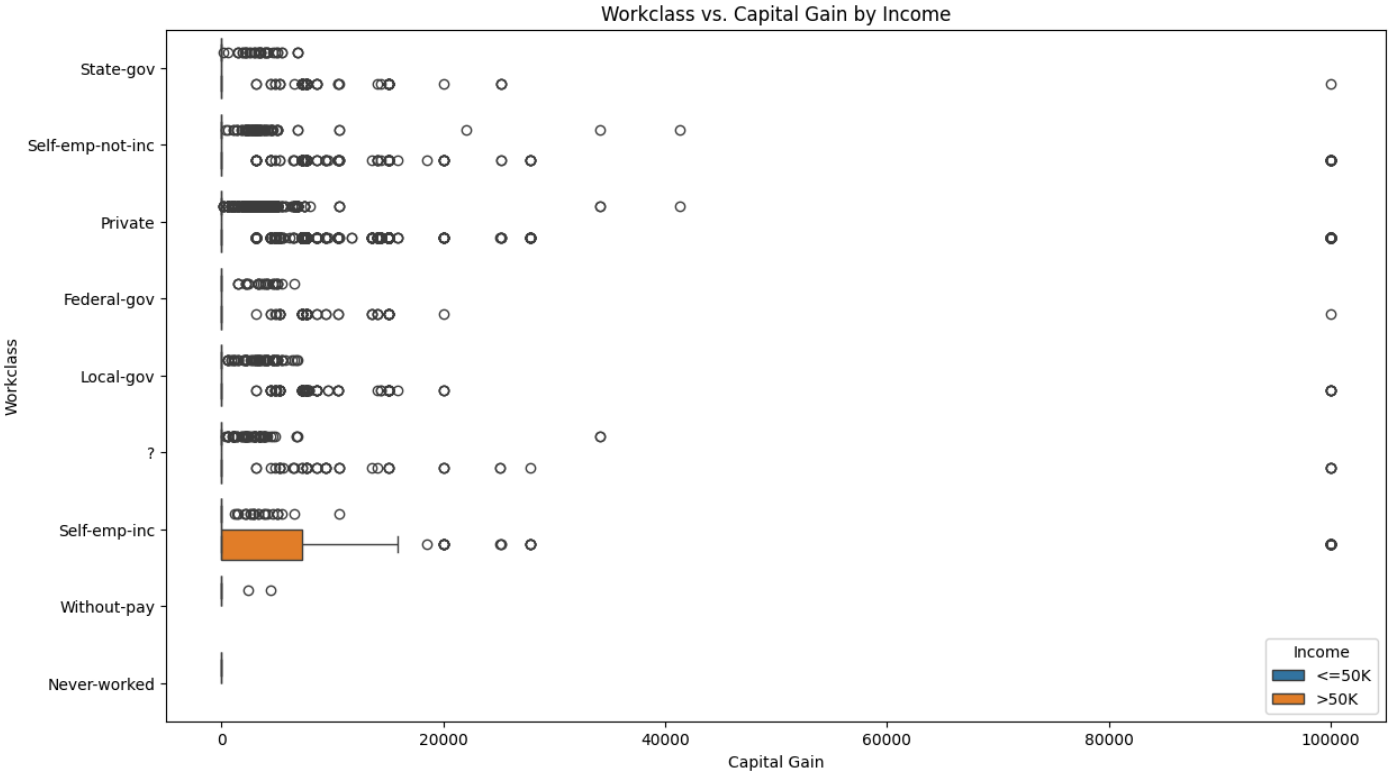

Relationship vs. Income
Husband
Not-in-family
Own-child
Unmarried
Wife
Other-relative
Relationship
0
1000
2000
3000
4000
5000
6000
7000
Count
Income
<=50K
>50K

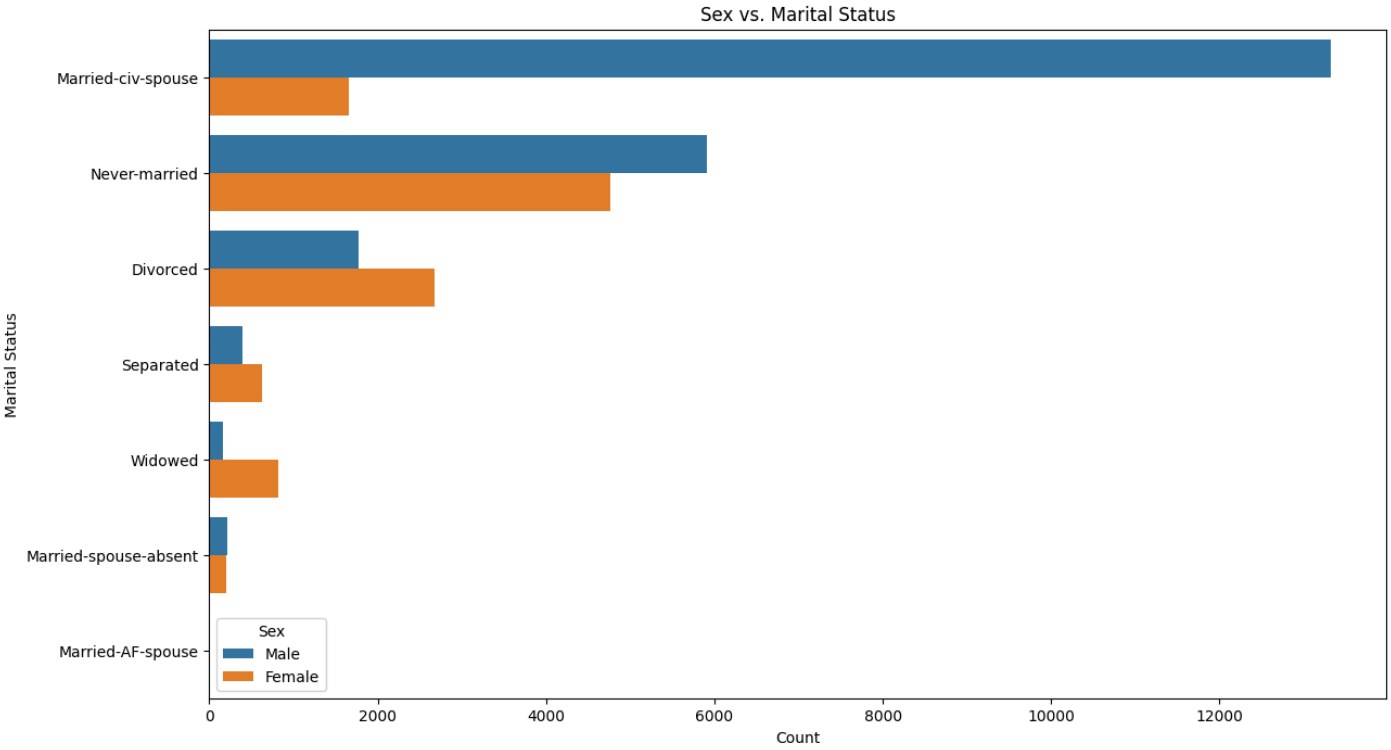

Sex vs. Age
90
80
70
60
50
40
30
20
Age
Male
Female
Sex

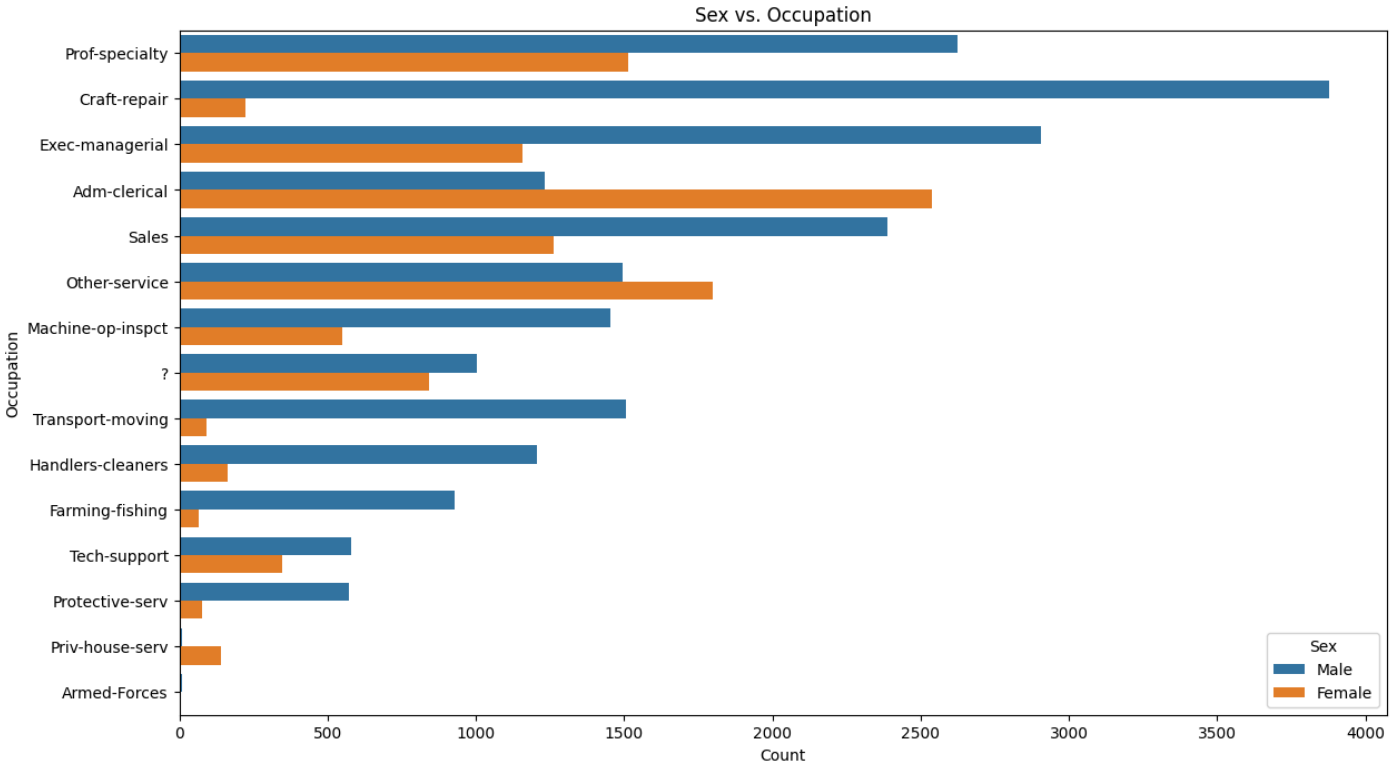

Sex vs. Education
HS-grad
Some-college
Bachelors
Masters
Assoc-voc
11th
Assoc-acdm
10th
7th-8th
Prof-school
9th
12th
Doctorate
5th-6th
1st-4th
Preschool
Education
0
1000
2000
3000
4000
5000
6000
7000
Count
Sex
Male
Female

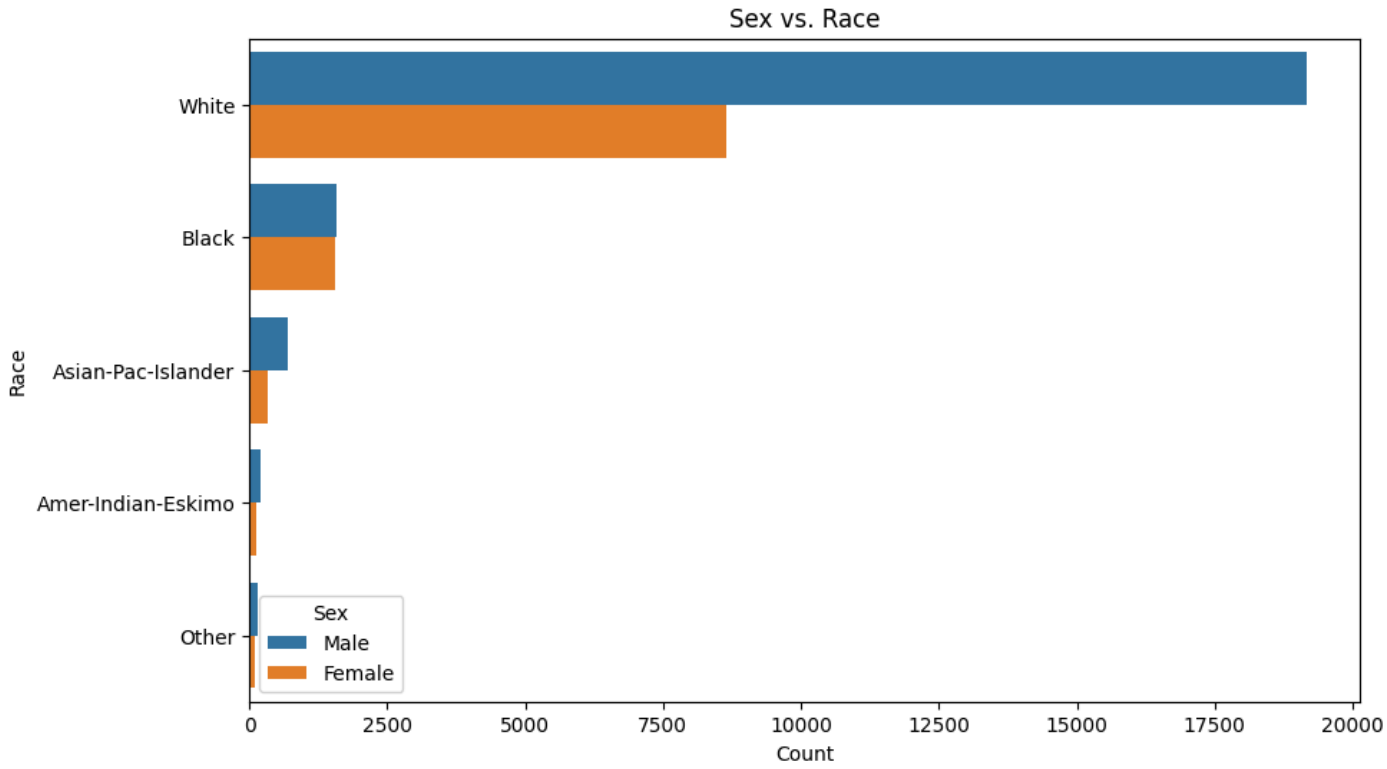

Sex vs. Hours per Week by Income
100
80
60
40
20
0
Hours per Week
Male
Female
Sex
Income
<=50K
>50K

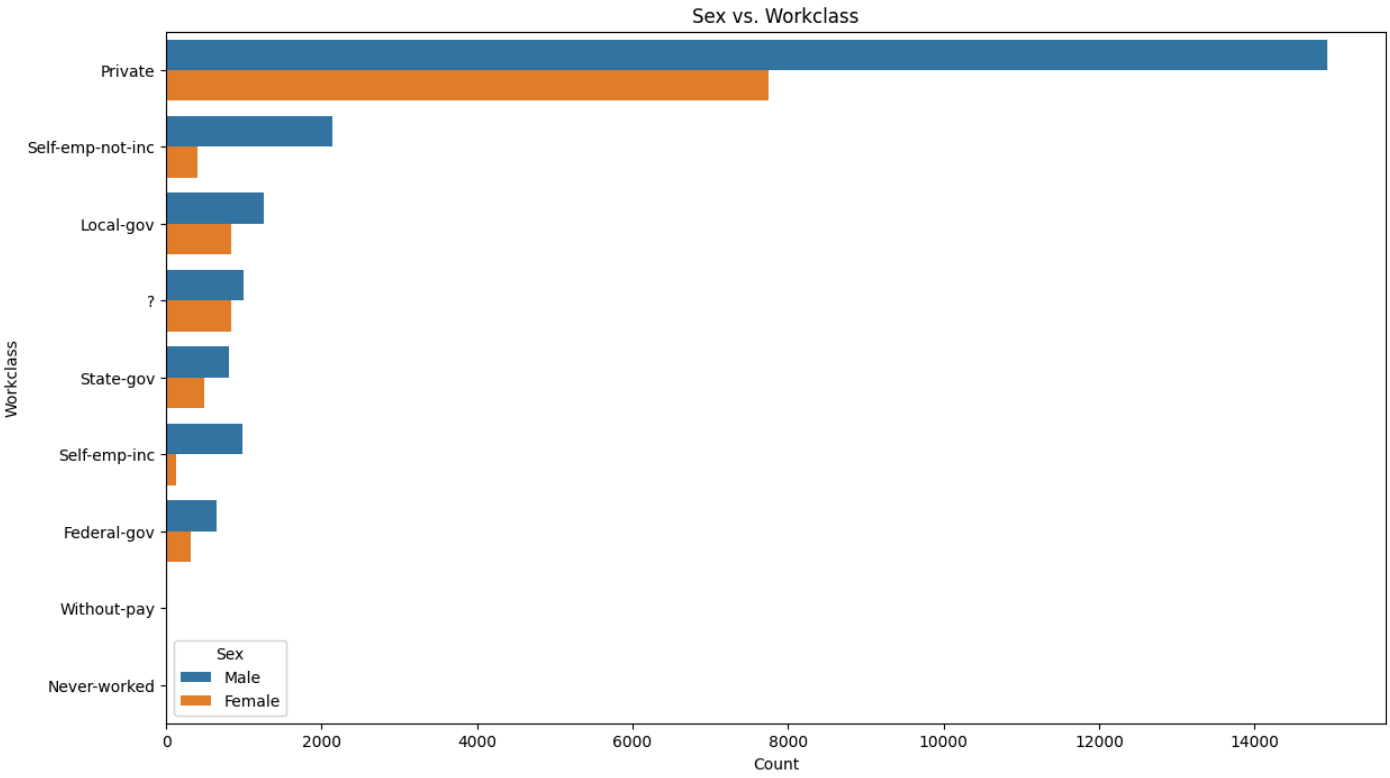

Sex vs. Income
Income
<=50K
>50K
Count
14000
12000
10000
8000
6000
4000
2000
0
Male
Female
Sex


Workclass vs. Capital Gain by Income
State-gov
Self-emp-not-inc
Private
Federal-gov
Local-gov
?
Self-emp-inc
Without-pay
Never-worked
Workclass
Income
<=50K
>50K
0
20000
40000
60000
80000
100000
Capital Gain


Sex vs. Marital Status
Married-civ-spouse
Never-married
Divorced
Separated
Widowed
Married-spouse-absent
Married-AF-spouse
Marital Status
Sex
Male
Female
0
2000
4000
6000
8000
10000
12000
Count


Sex vs. Occupation
Prof-specialty
Craft-repair
Exec-managerial
Adm-clerical
Sales
Other-service
Machine-op-inspct
?
Transport-moving
Handlers-cleaners
Farming-fishing
Tech-support
Protective-serv
Priv-house-serv
Armed-Forces
Occupation
Sex
Male
Female
0
500
1000
1500
2000
2500
3000
3500
4000
Count


Sex vs. Race
White
Black
Asian-Pac-Islander
Amer-Indian-Eskimo
Other
Race
Sex
Male
Female
0
2500
5000
7500
10000
12500
15000
17500
20000
Count


Sex vs. Workclass
Private
Self-emp-not-inc
Local-gov
?
State-gov
Self-emp-inc
Federal-gov
Without-pay
Never-worked
Workclass
Sex
Male
Female
0
2000
4000
6000
8000
10000
12000
14000
Count

Final Thoughts:

The UCI 1996 Adult Census dataset provides valuable insights into the factors influencing income levels. Here are the key findings:

1. Income Disparity:

- Gender: Males are more likely to earn `>50K` than females, indicating a gender income gap.
- Race: White individuals are more likely to earn `>50K` compared to other races, highlighting racial income disparities.
- Education: Higher education levels correlate strongly with higher income, emphasizing the importance of education.

2. Demographic Influences:

- Age: Income generally increases with age until middle age, then stabilizes or decreases near retirement.
- Marital Status: Married individuals, especially those with a spouse present, are more likely to earn higher incomes.

3. Occupational and Work Characteristics:

- Occupation: High-income earners are predominantly in executive, managerial, and professional roles.
- Work Hours: More hours worked per week generally lead to higher incomes, though the effect plateaus at higher hours.
- Workclass: Private sector employees and the self-employed tend to have higher incomes compared to those in government jobs.

4. Additional Patterns:

- Capital Gains: Significant capital gains are associated with higher incomes, highlighting the impact of investments.
- Geographic Factors: Individuals born in the United States are more likely to earn higher incomes compared to immigrants, reflecting potential systemic inequalities.

This analysis reveals how demographics, education, and occupation impact income levels. Gender and racial disparities are evident, underscoring the need for policies promoting equality. Education is a key factor in achieving higher income, and work characteristics also play a significant role.

Also an important feature to keep in mind is the simple fact that this data is collected from 1996, things were a lot different back then than they are now.