Adult Income Dataset

Osama Aldarabseh

Agenda

Description of the Dataset

Preprocessing

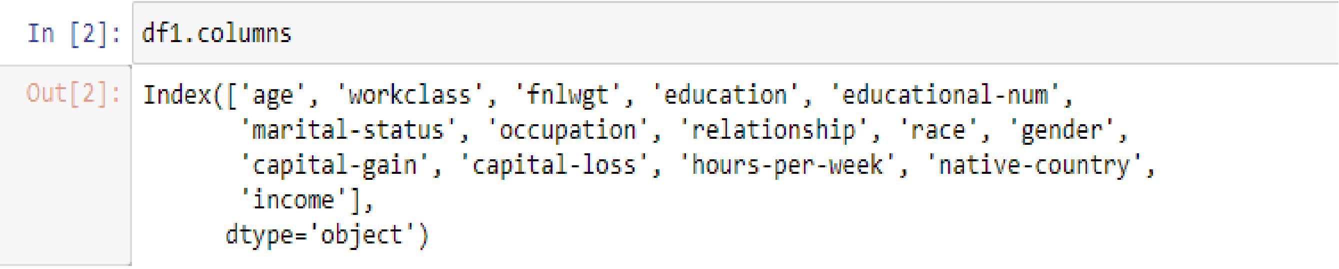
Analysis

Adult Income Dataset

The Adult Income Dataset is a popular dataset used in machine learning for binary classification tasks. The prediction task is to determine whether a person makes over $50K a year and many other perspectives. The dataset contains 15 input variables that are a mixture of categorical, ordinal, and numerical data types.

Dataset Elements

The dataset are 14 columns as follows:



```
In [2]: df1.columns

Out[2]: 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')
```

Pre-processing

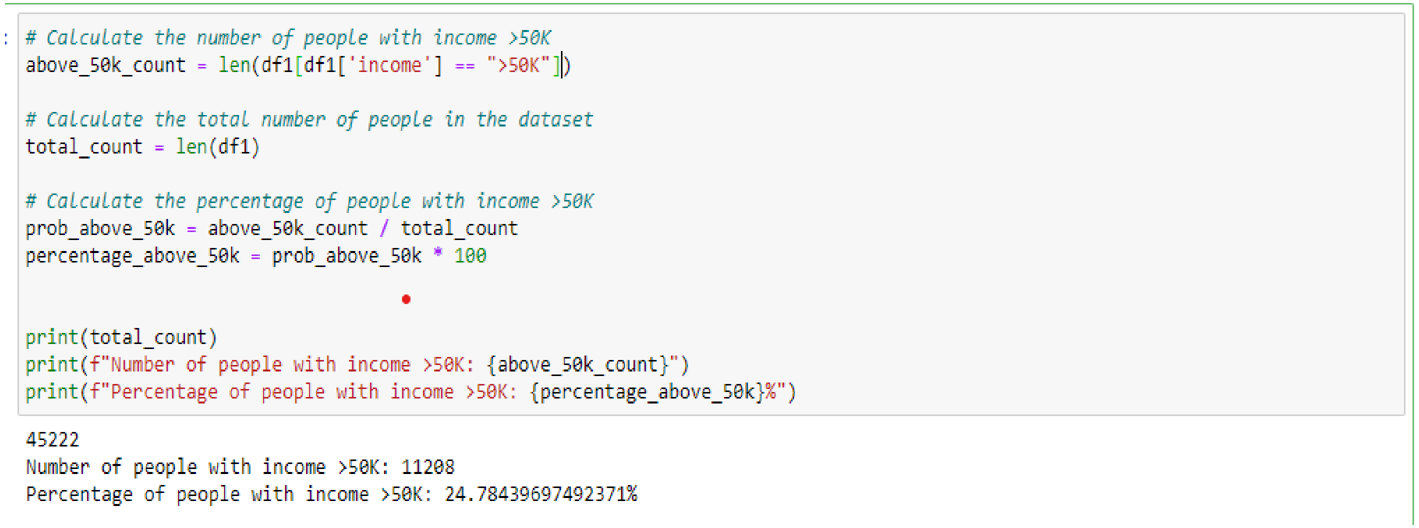
- In the pre-processing phase, operations like cleaning, handling missing data, drop rows or columns and remove duplicates if found will be done.
- In this dataset, ‘capital-gain’ and ‘capital-loss’ columns were dropped due to the inaccurate or zero entries. And in some columns inaccurate entries like ‘?’ were spotted in many rows, all the rows like that, were dropped.
- The shape of the dataset before the operations was (48842 ,15) , and then after few operations became (45222 , 13).

Analysis

Data analysis is often performed to discover patterns and relationships within the data. These patterns can provide valuable insights that can be used for decision making, predictions, strategy development, and many other applications.

In the Adult Dataset, I have found many insights that can used for multiple purposes, for example the average working hours for every job or even the effect of your education level on your income.

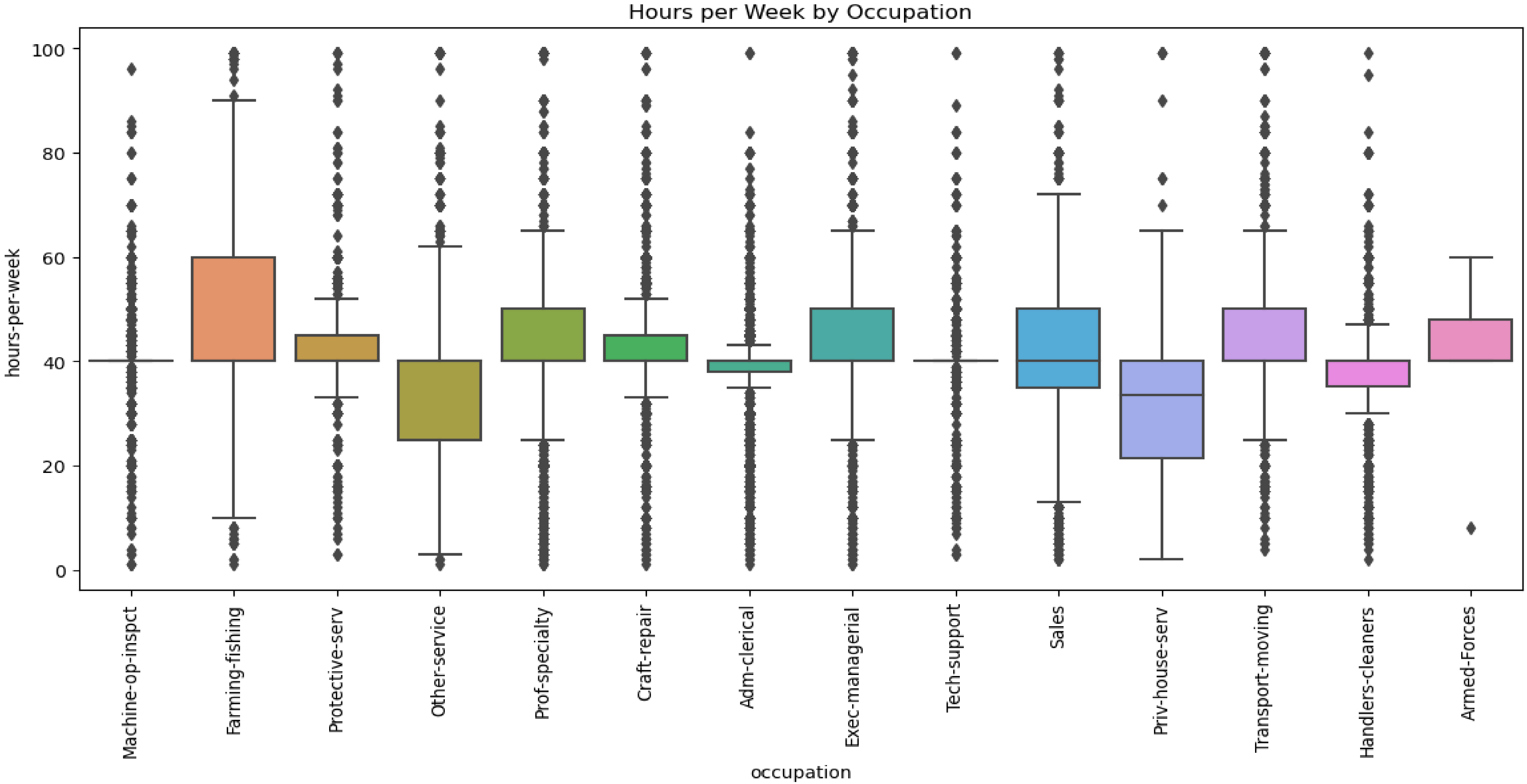
The first analysis as it should be is the number of people who have income ">50K" and their percentage:



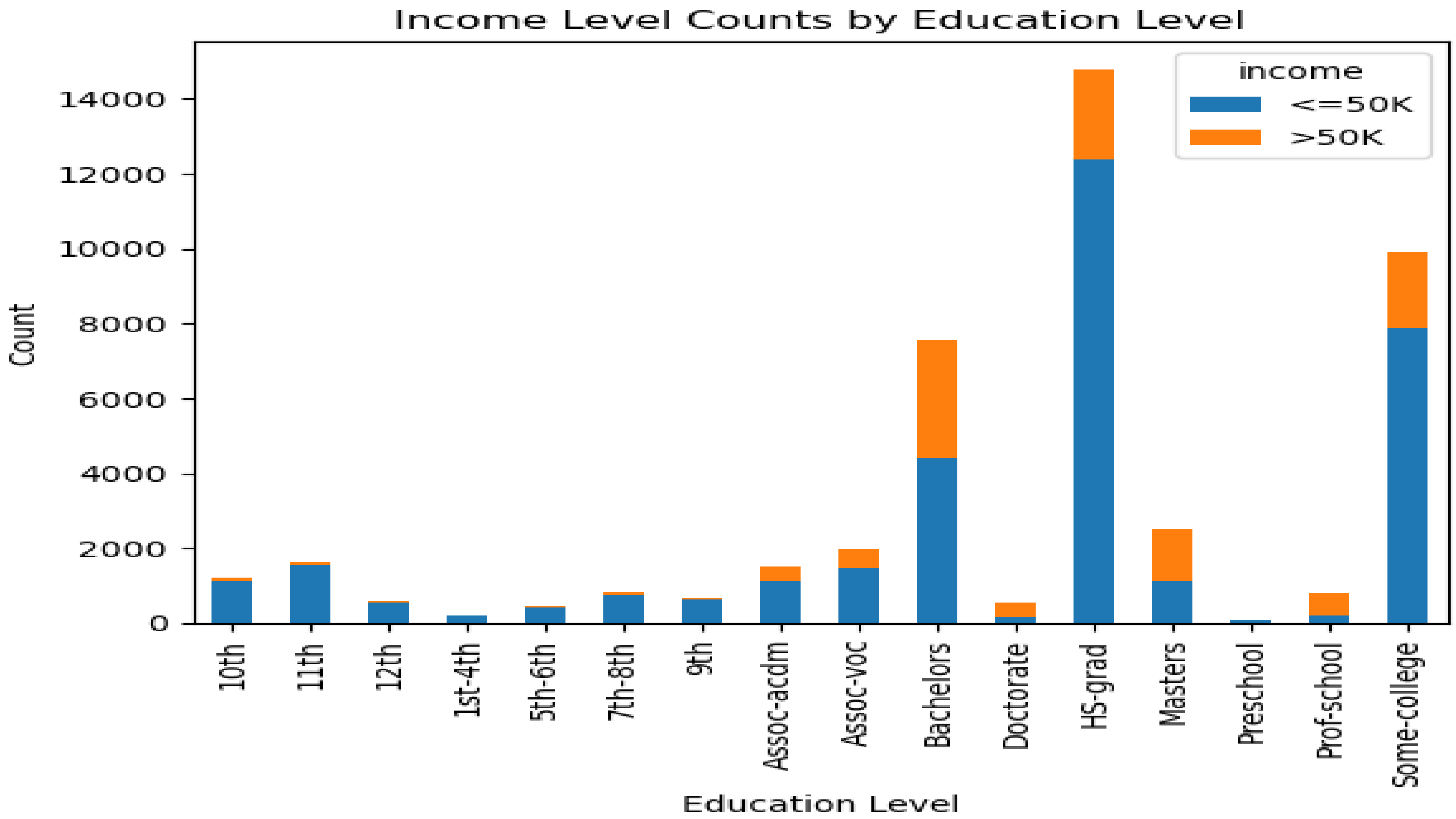
```python
# Calculate the number of people with income >50K
above_50k_count = len(df1[df1['income'] == ">50K"])

# Calculate the total number of people in the dataset
total_count = len(df1)

# Calculate the percentage of people with income >50K
prob_above_50k = above_50k_count / total_count
percentage_above_50k = prob_above_50k * 100


print(total_count)
print(f"Number of people with income >50K: {above_50k_count}")
print(f"Percentage of people with income >50K: {percentage_above_50k}%")
```

```
45222
Number of people with income >50K: 11208
Percentage of people with income >50K: 24.78439697492371%
```

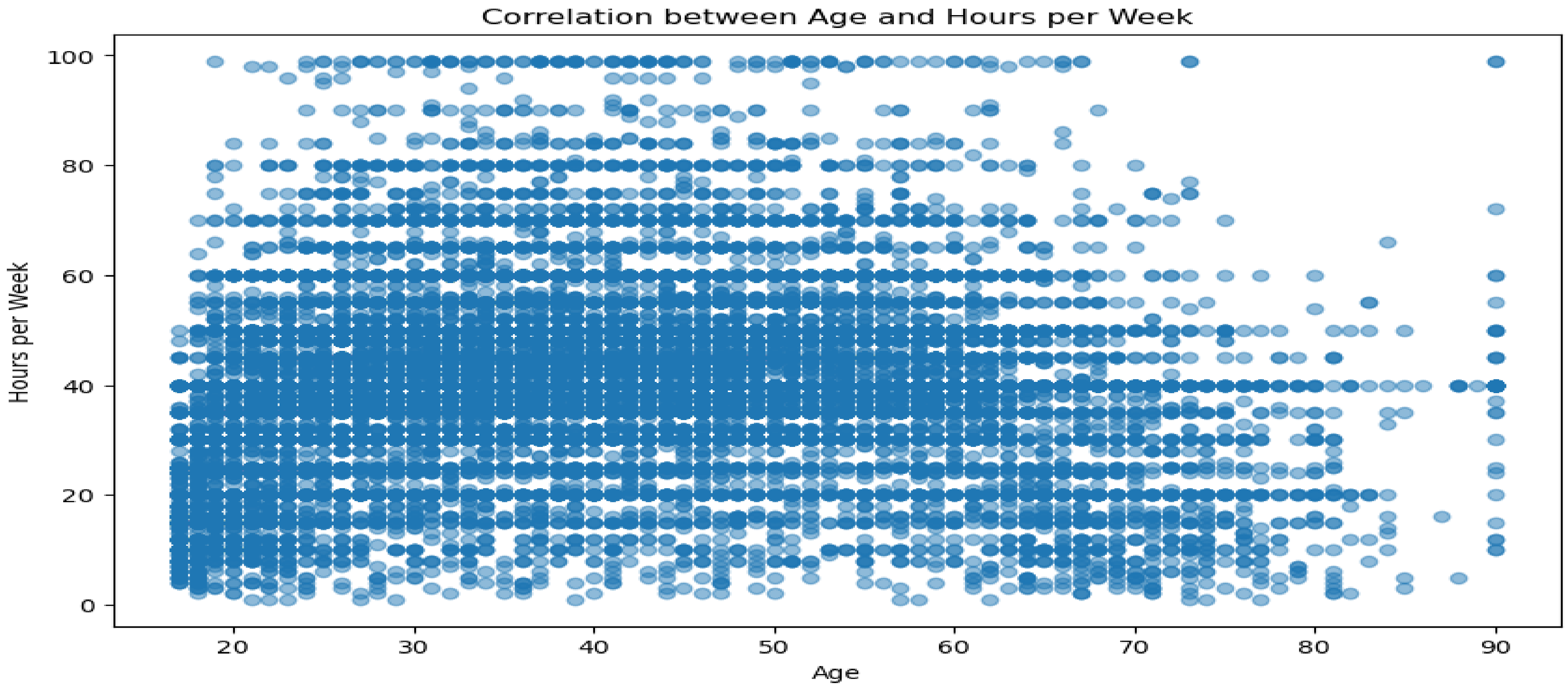
Hours-per-week by occupation



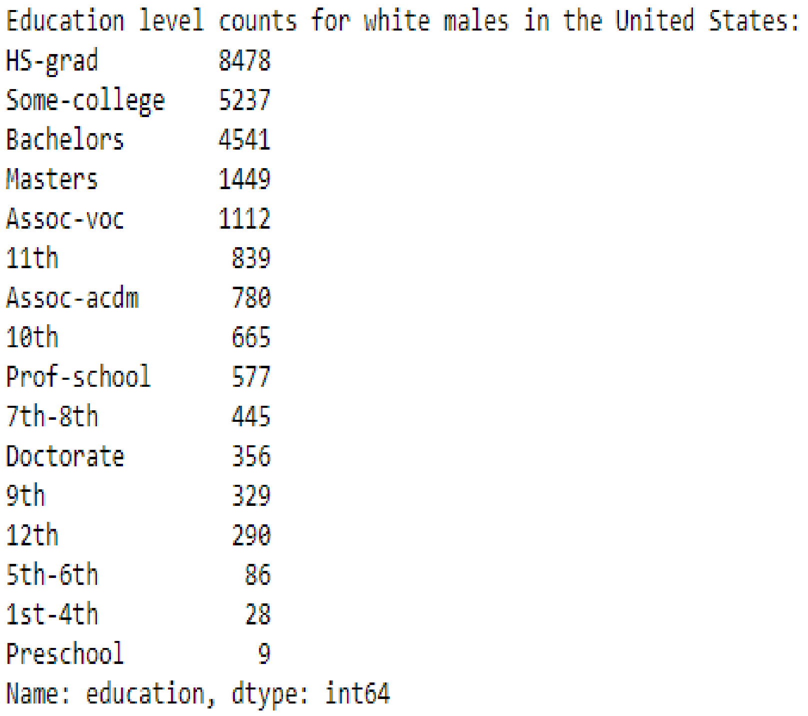
Income level counts by Education level



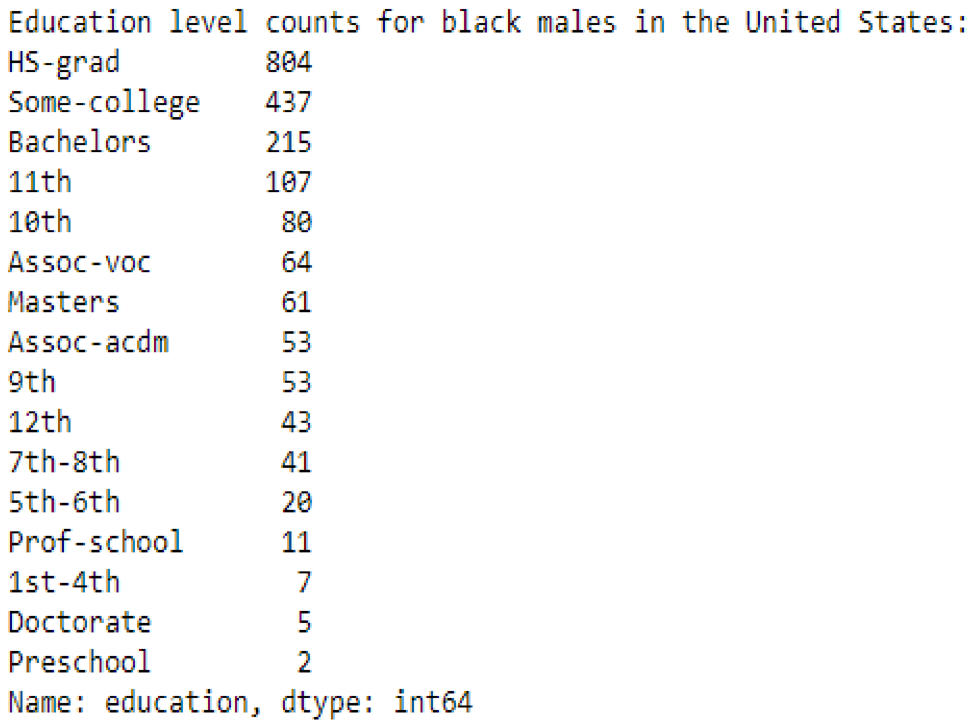
Also I found a correlation between the age and the working hours per week



Then I have shown some statistics:

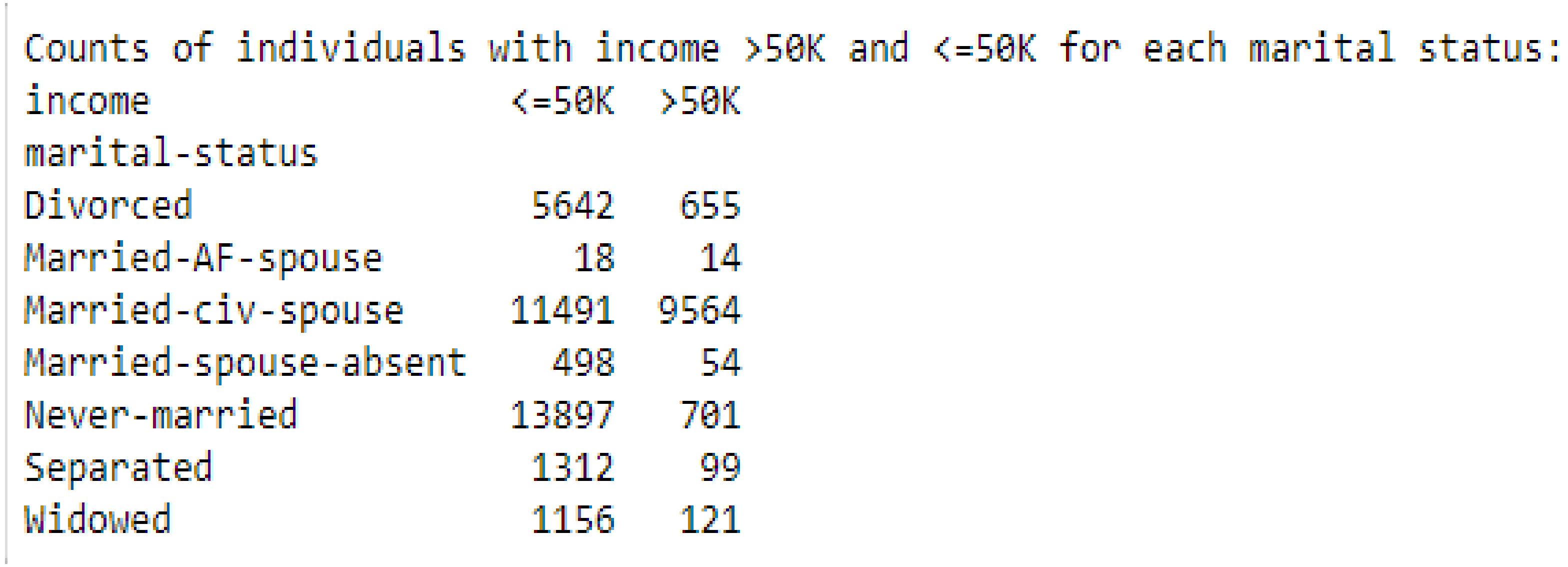


```
Education level counts for white males in the United States:
HS-grad         8478
Some-college    5237
Bachelors       4541
Masters         1449
Assoc-voc       1112
11th             839
Assoc-acdm       780
10th             665
Prof-school      577
7th-8th          445
Doctorate        356
9th              329
12th             290
5th-6th           86
1st-4th           28
Preschool          9
Name: education, dtype: int64
```



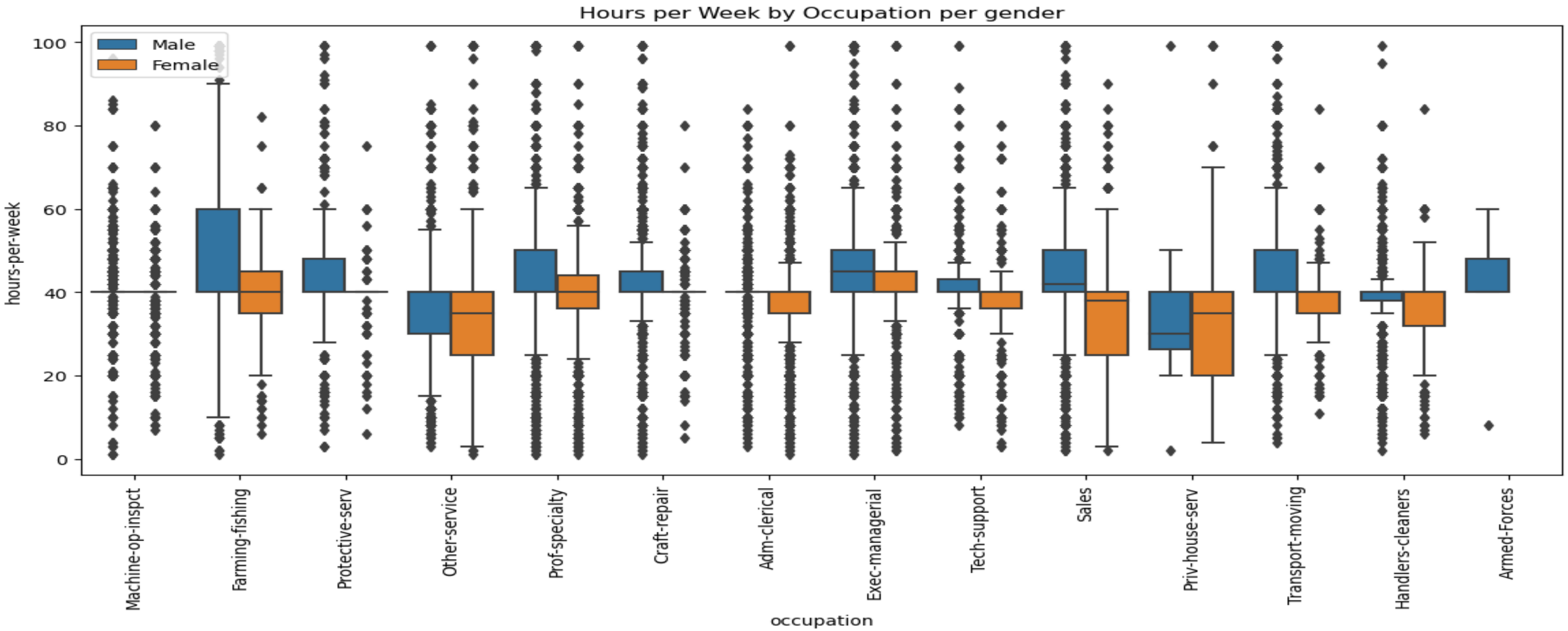
```
Education level counts for black males in the United States:
HS-grad         804
Some-college    437
Bachelors       215
11th            107
10th             80
Assoc-voc        64
Masters          61
Assoc-acdm       53
9th              53
12th             43
7th-8th          41
5th-6th          20
Prof-school      11
1st-4th           7
Doctorate         5
Preschool         2
Name: education, dtype: int64
```

The income according to the marital status



```
Counts of individuals with income >50K and <=50K for each marital status:
income                 <=50K  >50K
marital-status
Divorced                5642   655
Married-AF-spouse         18    14
Married-civ-spouse     11491  9564
Married-spouse-absent    498    54
Never-married          13897   701
Separated               1312    99
Widowed                 1156   121
```

The working hours per week by gender



Thank you

👤

Osama Aldarabseh

✉

osamaqasim32@outlook.com