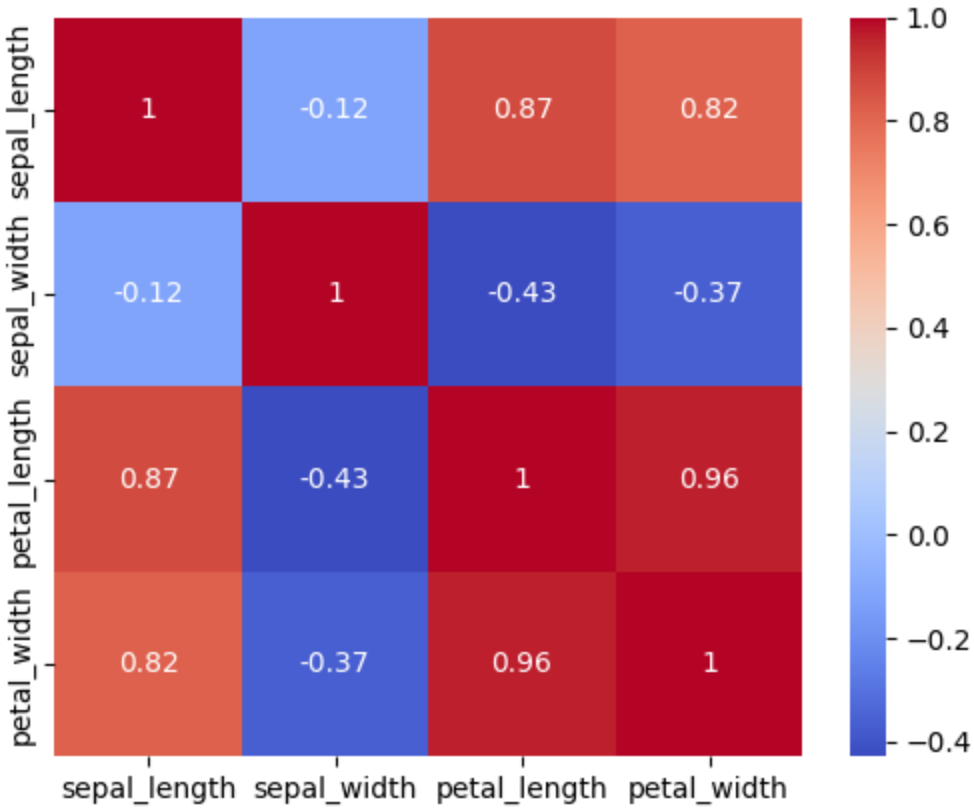
B20AI013 DV lab 5&6 - Exploratory Data Analysis

1. Iris Data Analysis

The Iris Dataset is a multivariate data set introduced by Sir Ronald Aylmer Fisher in 1936. It contains 50 samples from each of three species of Iris (Iris setosa, Iris virginica, and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters.

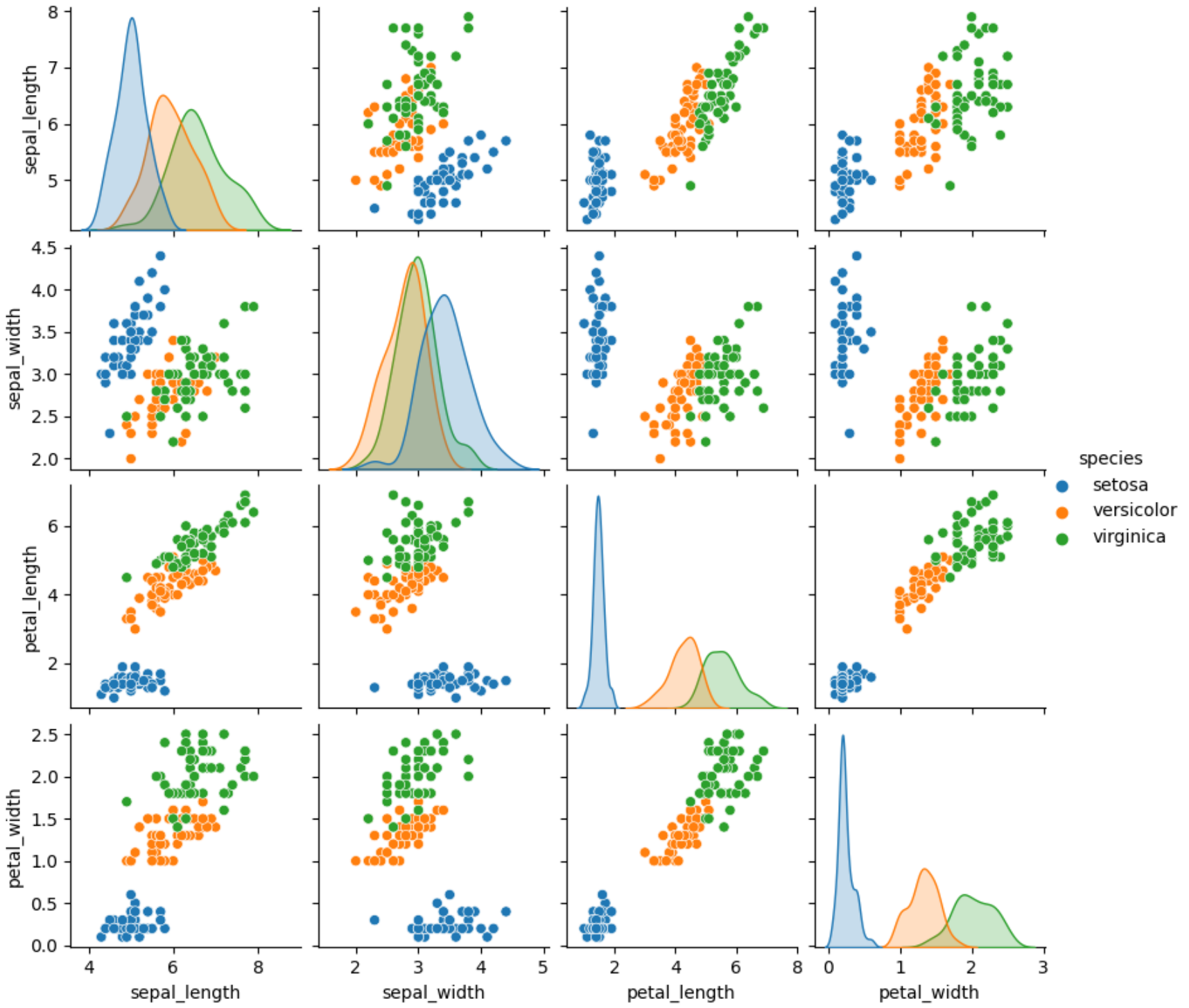
To begin analyzing the relationships among different variables, correlation heatmaps are a valuable tool that can provide insight into relations between variables.



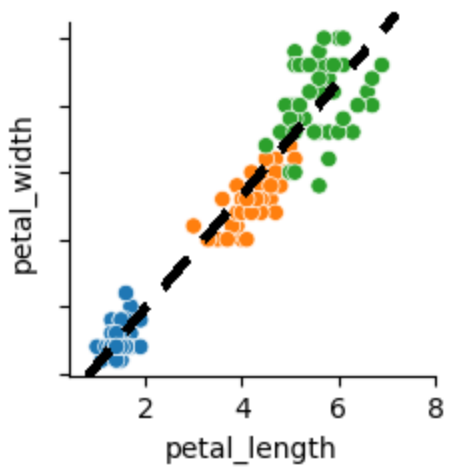
From the above visualization, we can see:

1. `petal_length` and `petal_width` are highly correlated, while `sepal_length` and `sepal_width` are moderately correlated.

To visualize the relationships among different variables, we can create a scatter plot matrix.



From the scatter plot matrix, we can see that there is a clear linear relationship between petal length and petal width.

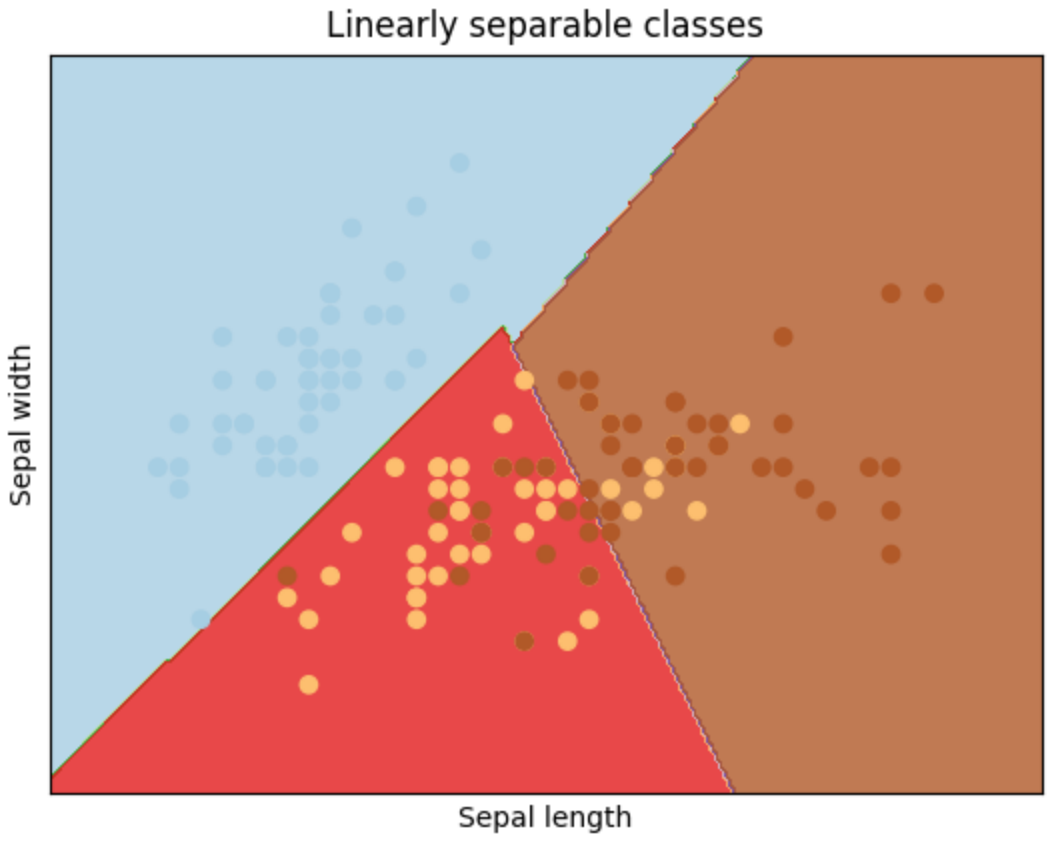


This relationship is evident in all three species of Iris. There is also some separation between the species based on petal length and width, but it is not as clear.

We can also see that `petal_length` and `petal_width` have a stronger correlation with the species than `sepal_length` and `sepal_width`.

2. Linear Separability

To determine which class is linearly separable, we can use a support vector machine (SVM) classifier. SVMs are commonly used for binary classification problems and can determine whether the classes are linearly separable or not.



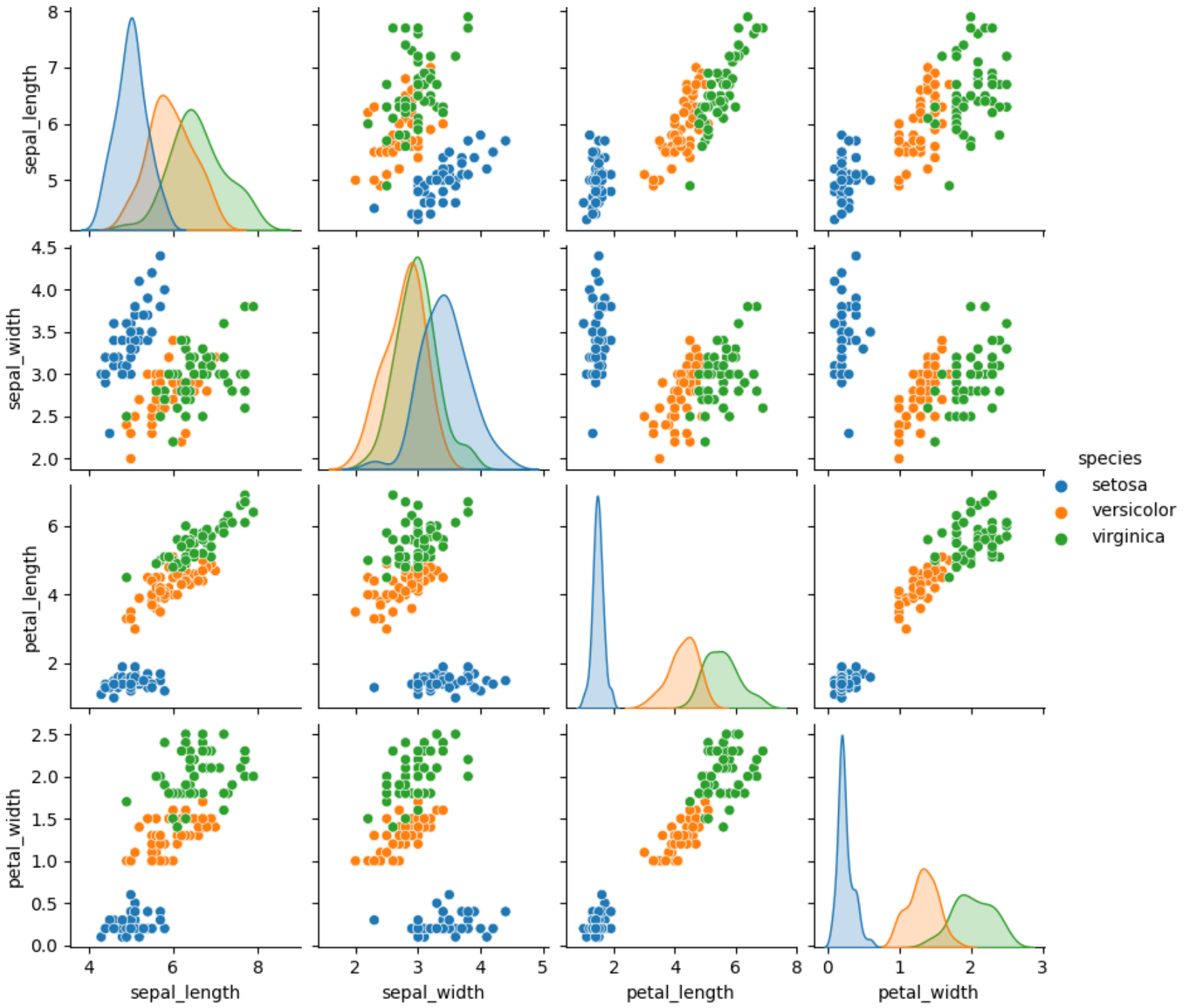
From the above visualization, we can see that the classes are more or less linearly separable in the Iris dataset. The Setosa (blue) samples are clearly linearly separated from Versicolor and Virginica, however there is some confusion between Versicolor (red) and Virginica (brown).

3. EDA: UCI Adult Dataset

We will use EDA to analyse the dataset. First, we load the dataset by downloading it from the website link given. There are 14 columns: `["age", "workclass", "fnlwgt", "education", education-num", "marital-status", "occupation", "relationship", "race", "sex", "capital-gain", "capital-loss", "hours-per-week", "native-country", "income"]`

These columns include information about individuals such as age, workclass, education, marital status, occupation, relationship, race, sex, capital gain, capital loss, hours per week, and native country. The final column, `"income"`, indicates whether the individual makes more or less than $50,000 per year.

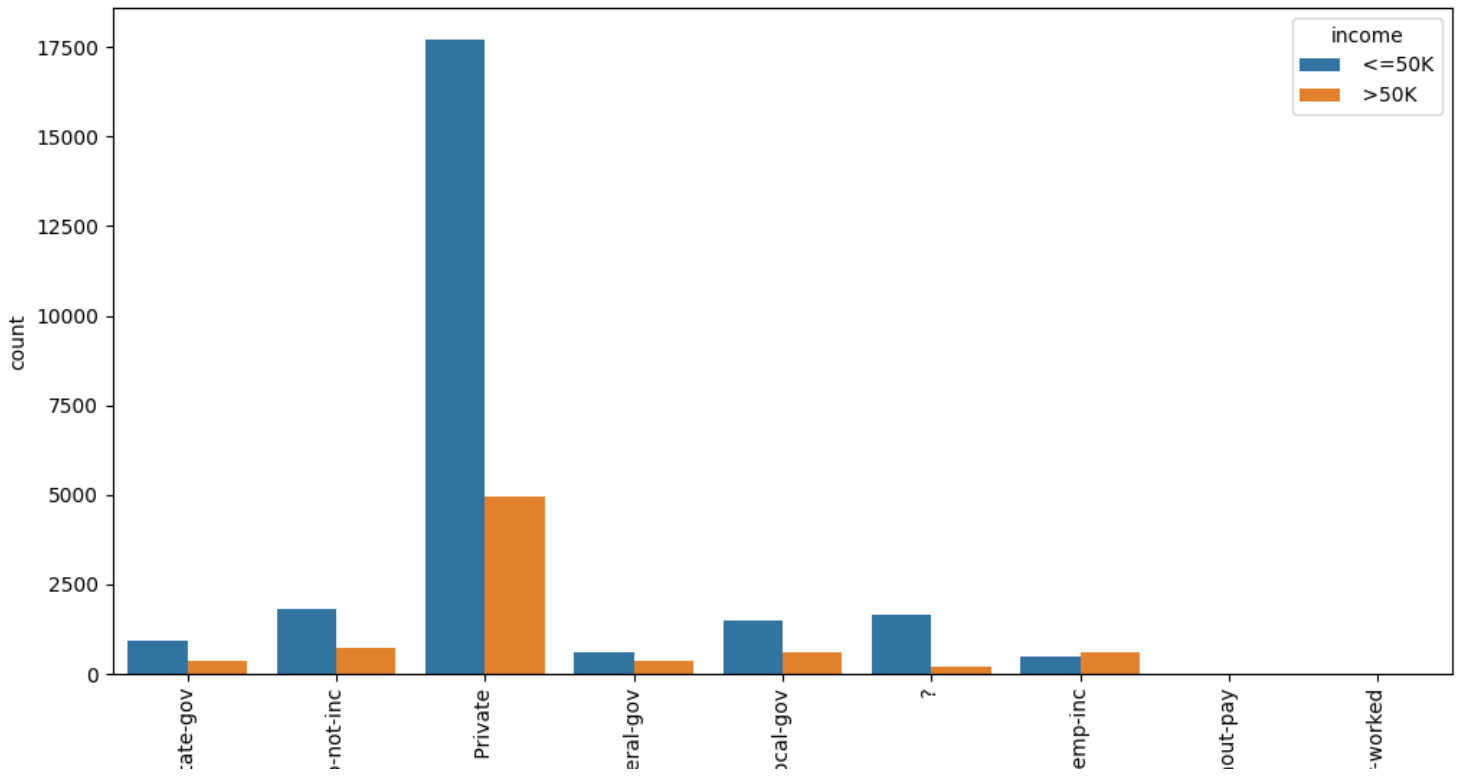
Looking at the distribution of the target variable, `"income"` :

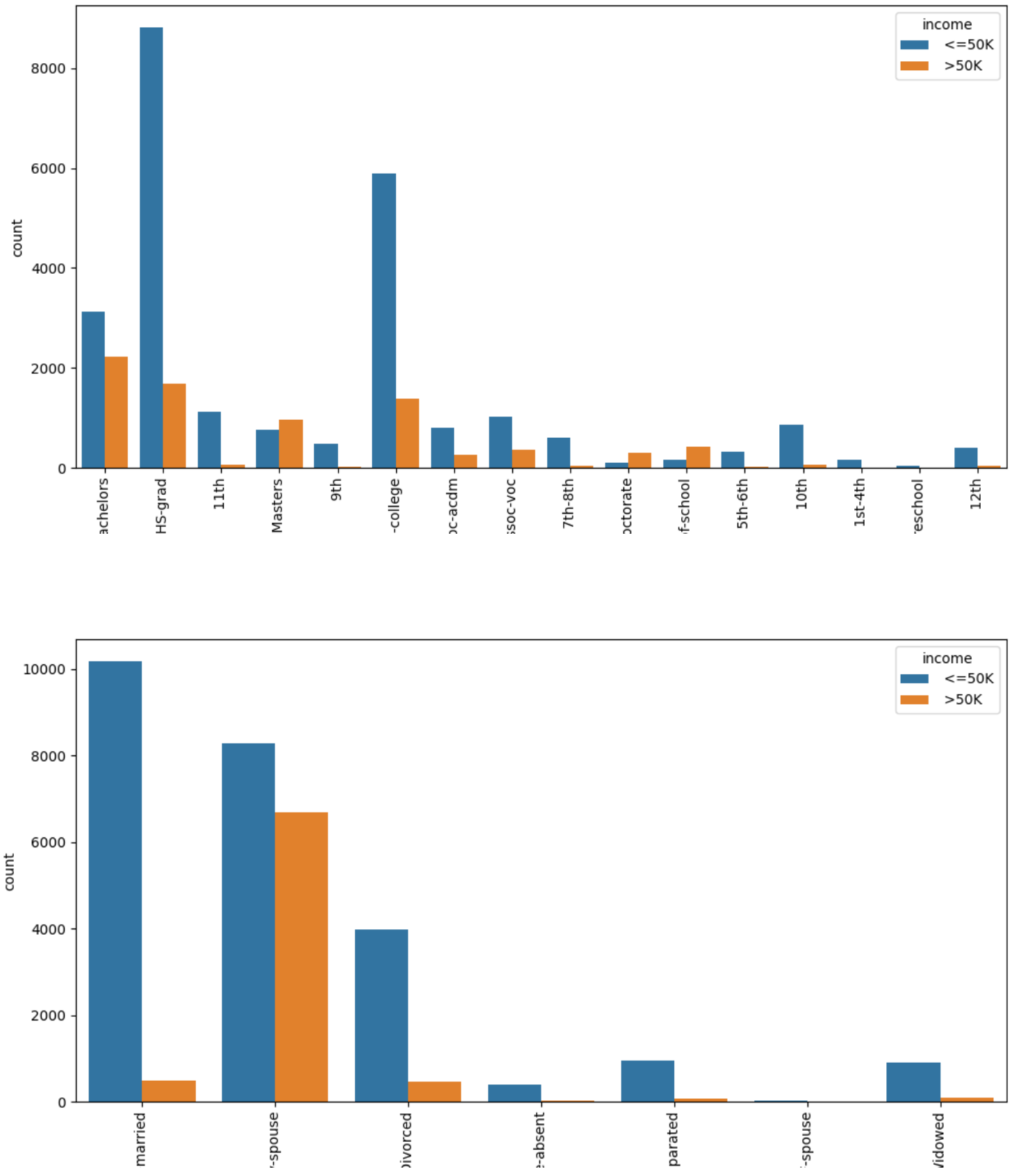


From the plot, we can see that the dataset is imbalanced with a majority of individuals having an income of less than or equal to $50,000 per year.

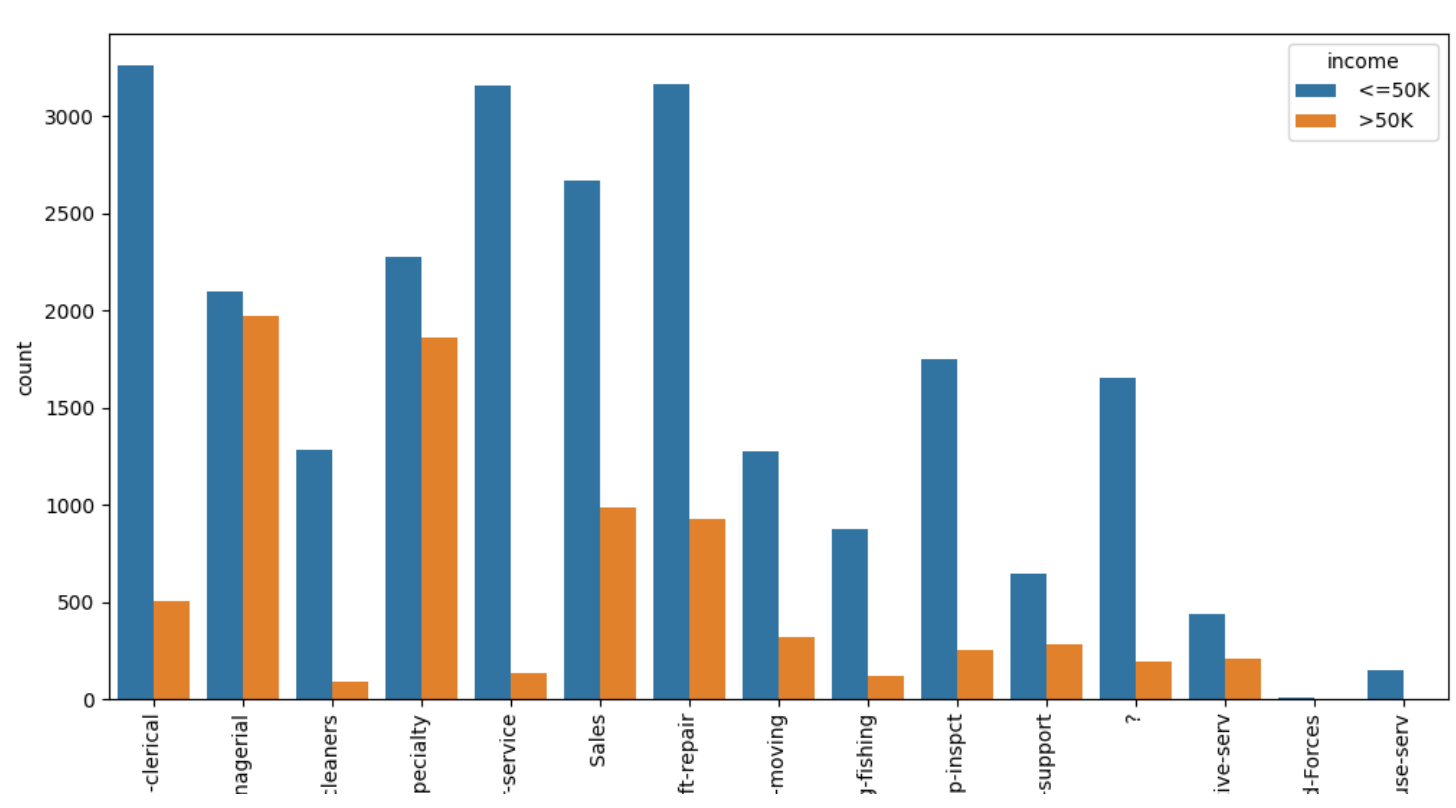
Numerical Features

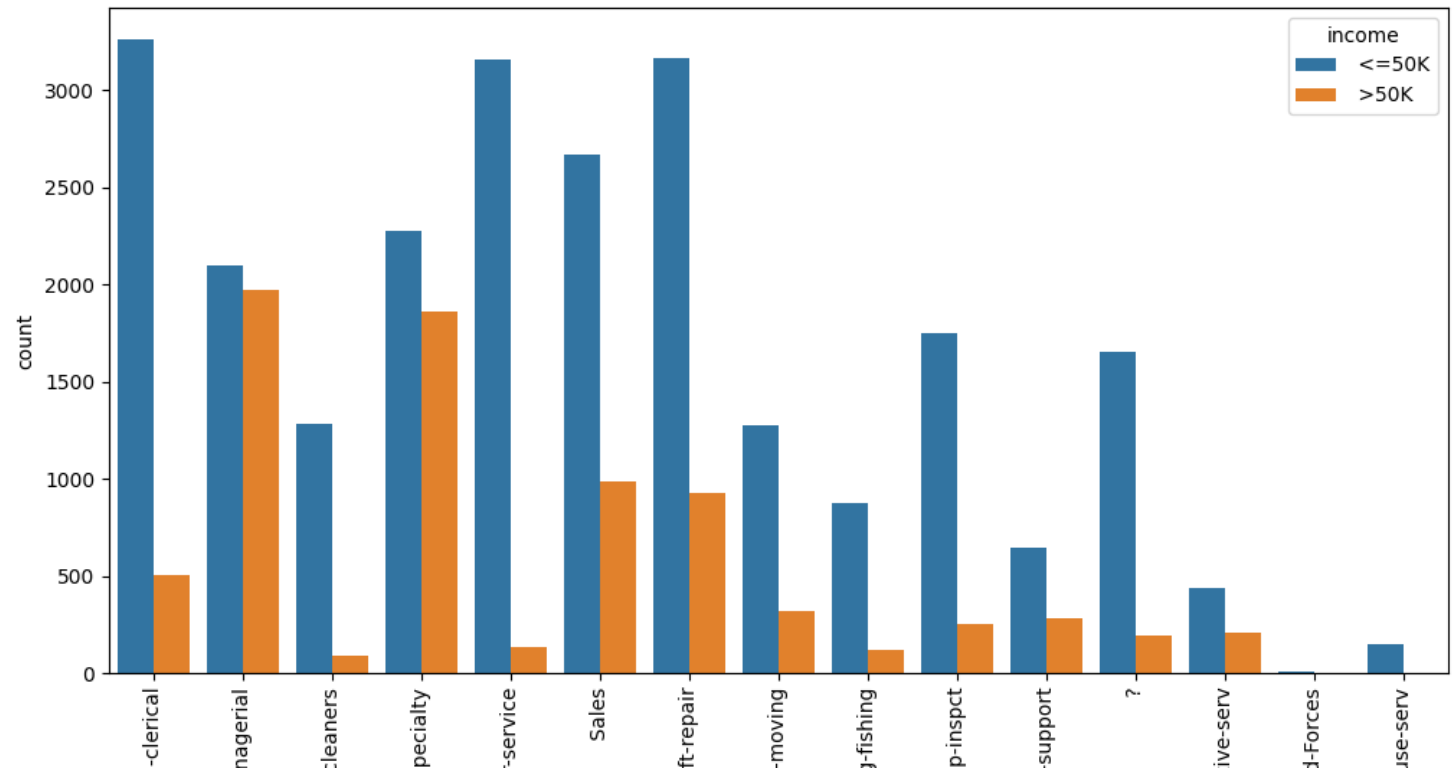
Let's create a pairplot to visualize the relationships between the numerical features and the target variable.

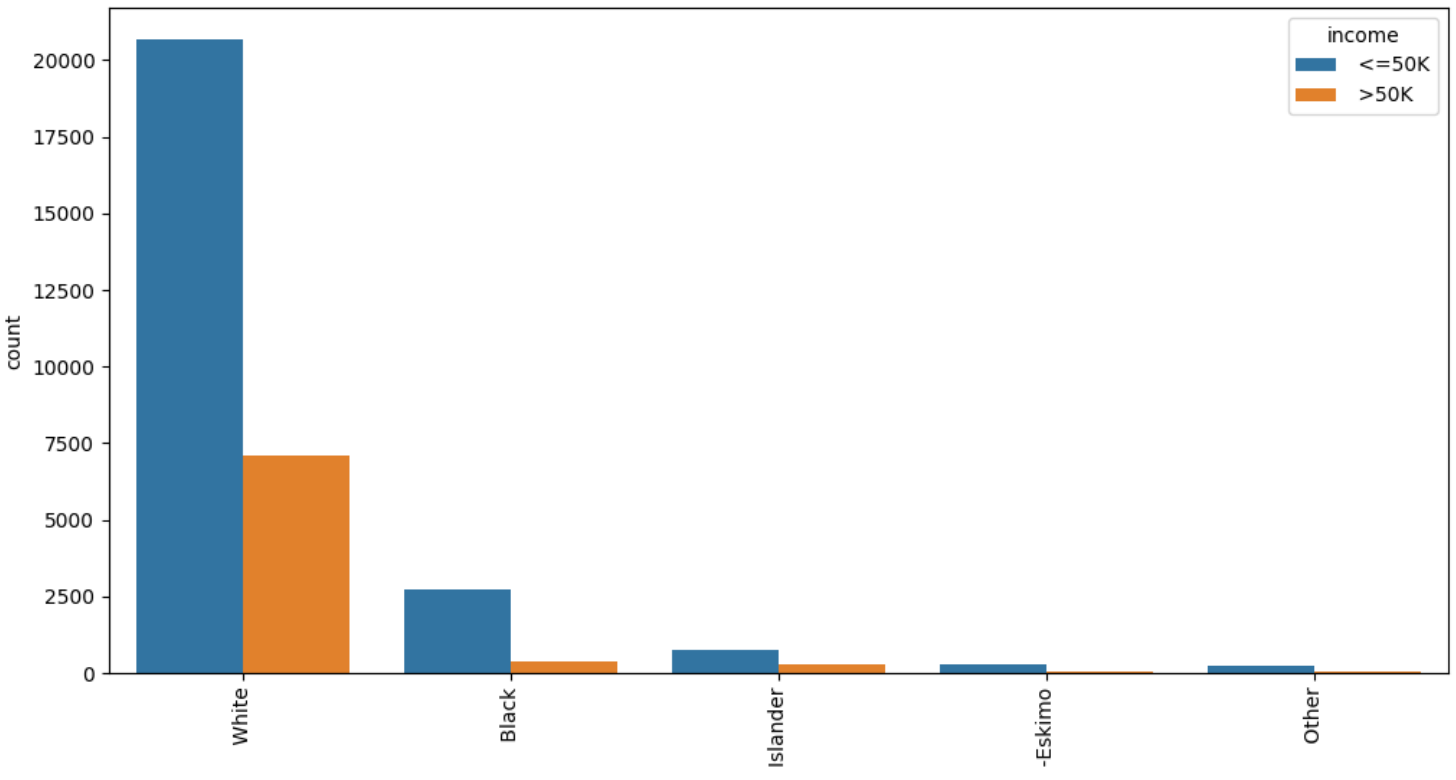


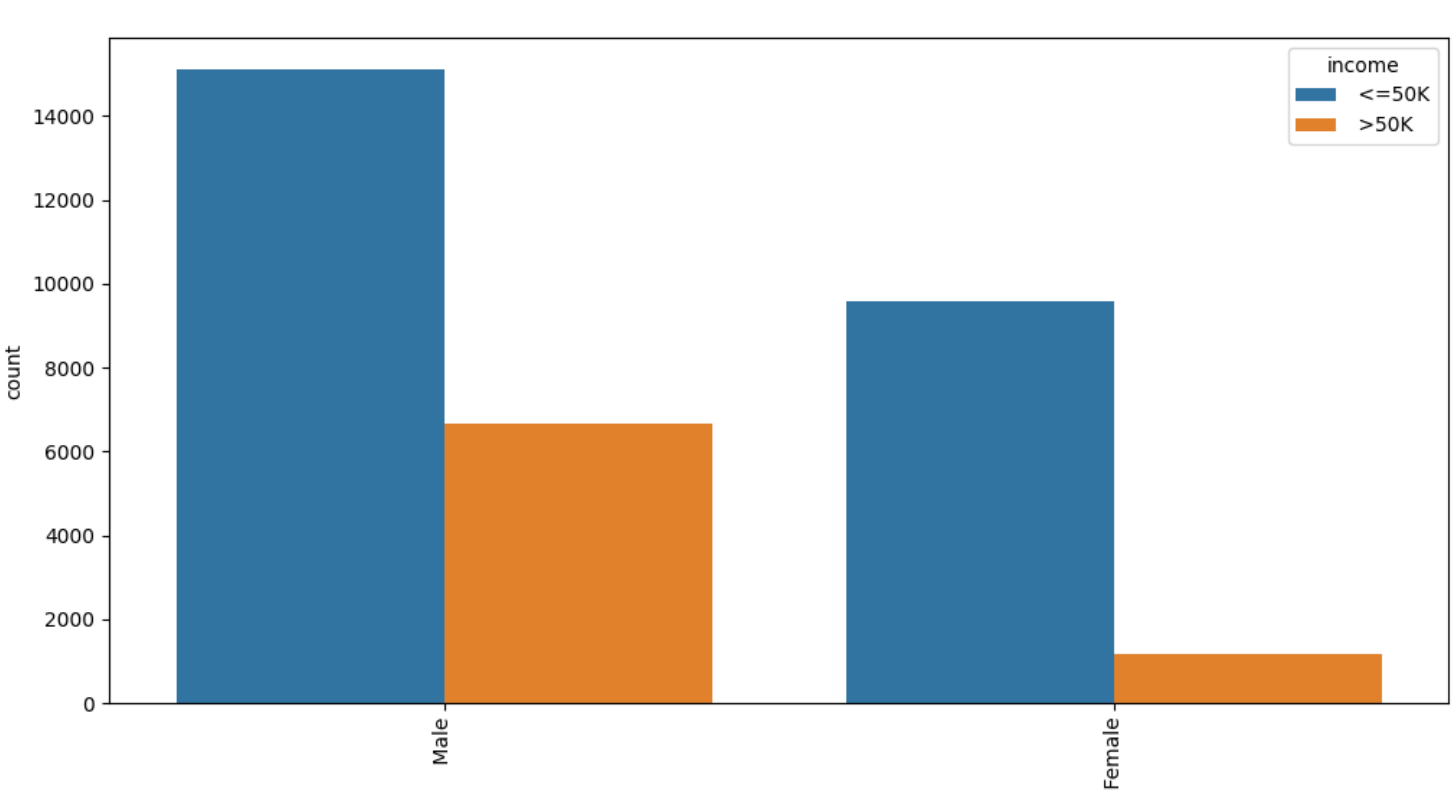

income
<=50K
>50K
count
8000
6000
4000
2000
0
achelors
HS-grad
11th
Masters
9th
-college
oc-acdm
soc-voc
7th-8th
octorate
f-school
5th-6th
10th
1st-4th
reschool
12th

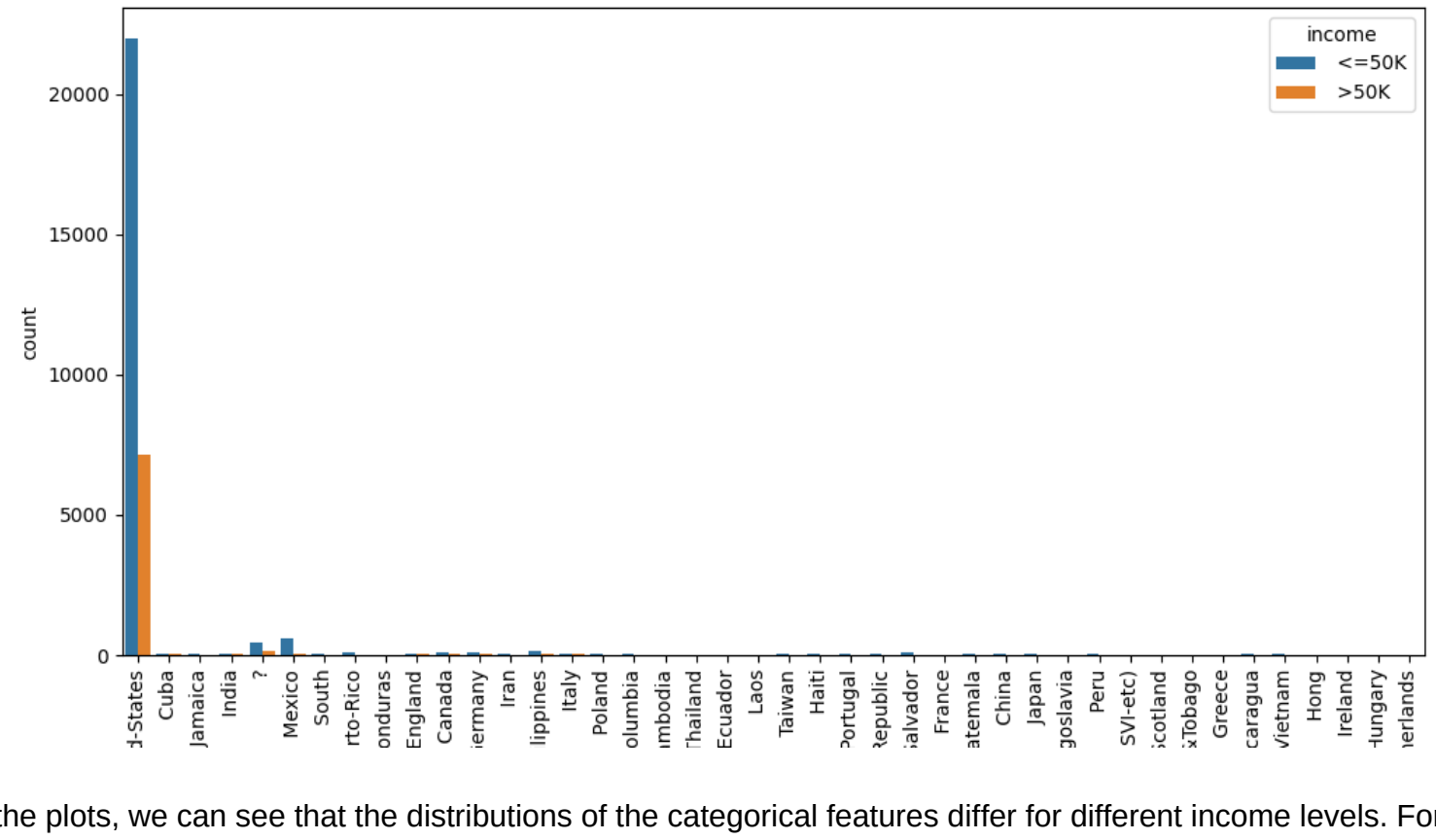
income
<=50K
>50K
count
10000
8000
6000
4000
2000
0
married
-spouse
ivorced
-absent
parated
-spouse
idowed


income
<=50K
>50K
count
3000
2500
2000
1500
1000
500
0
-clerical
nagerial
leaners
pecialty
-service
Sales
ft-repair
-moving
g-fishing
p-inspct
-support
?
ive-serv
d-Forces
se-serv


income
<=50K
>50K
count
3000
2500
2000
1500
1000
500
0
-clerical
nagerial
cleaners
pecialty
-service
Sales
ft-repair
-moving
g-fishing
p-inspct
-support
?
ive-serv
d-Forces
se-serv


income
<=50K
>50K
count
20000
17500
15000
12500
10000
7500
5000
2500
0
White
Black
Islander
-Eskimo
Other


income
<=50K
>50K
count
14000
12000
10000
8000
6000
4000
2000
0
Male
Female



From the plots, we can see that the distributions of the categorical features differ for different income levels. For example, those who work in the private sector are more likely to have an income of less than or equal to $50,000 per year, while those who work in executive/managerial positions are more likely to have an income of greater than $50,000 per year. Additionally, those who are married and have a spouse are more likely to have an income of greater than $50,000 per year.

Feature Engineering

We can create a new feature by combining `"capital-gain"` and `"capital-loss"` into a single feature called `"net-capital-gain"`.

```
df["net-capital-gain"] = df["capital-gain"] - df["capital-loss"]
df.drop(["capital-gain", "capital-loss"], axis=1, inplace=True)
```

Correlation

Correlation measures how strong the linear interdependence between two variables is. An absolute value of correlation closer to 1 is indicative of a string correlation. Printing the correlation of the numerical features with the target variable, we get the output:

```
income             1.000000
education-num      0.335154
net-capital-gain   0.223013
age                0.234714
hours-per-week     0.229306
fnlwgt            -0.008957
Name: income, dtype: float64
```

From the output, we can see that `"education-num"`, `"net-capital-gain"`, `"age"`, and `"hours-per-week"` have a moderate correlation with income.