CSE 578 Data Visualization - Project Milind Parab

XYZ Corporation has been tasked with creating marketing profiles for UVW College

Project

XYZ Corporation uses data to develop marketing profiles on people. These profiles are then sold to numerous companies for marketing purposes. You work at XYZ as a data analyst. You have just been given a new project working with UVW College, a local college looking to bolster enrollment. UVW has chosen a salary as a key demographic to determine criteria for marketing its degree programs. You must develop marketing profiles using data supplied by the United States Census Bureau, and you will be focusing on 50,000 as a key number for salary. There are many key variables that must be assessed for individuals making less than and more than $50,000, including age, gender, education status, marital status, occupation, etc. For example, if the data show that the majority of individuals making less than 50,000 is under 34 years old, male, single, and has a high school diploma, the college can market to this demographic with tuition amounts, program concentrations, and even ground or online programs appropriate to this demographic. To achieve its enrollment target, the marketing team at UVW would like to develop an application to find the factors that determine the individual's income. One way to accomplish this is to use the United States Census Bureau data provided by the XYZ company. The marketing team wants to group the factors that can be used in the development of their proposed model/application. They also want the application to predict the income of an individual based on different values of the input parameters so that they can tailor their marketing efforts when reaching out to the individuals.

Problem Statement:

XYZ Corporation has been tasked with creating marketing profiles for UVW College, focusing on using demographic data to assess and predict individuals' incomes, particularly focusing on the salary threshold of 50,000. The objective is to analyze various demographic factors (such as age, gender, education, marital status, and occupation) from the data provided by the United States Census Bureau. The goal is to segment individuals based on their income level (under or over $50,000) and use these insights to help UVW College target potential students more effectively. The marketing team seeks to develop an application that can predict income based on demographic variables, and group the relevant factors to enhance their marketing strategies.

Objective:

- **Segment individuals** into two groups based on income: those earning less than 50,000 and those earning more than 50,000.
- **Identify key demographic factors** (e.g., age, gender, education, marital status, occupation) that influence income.
- **Develop a predictive model/application** that forecasts an individual’s income based on selected input parameters.
- **Group related factors** that are critical to UVW College’s marketing efforts, allowing them to tailor their outreach strategies for increased enrollment.

Approach:

1. **Data Preprocessing**:
   - Clean and preprocess the data from the United States Census Bureau, ensuring it is structured and ready for analysis.
   - Handle missing values, outliers, and data normalization where needed.
2. **Exploratory Data Analysis (EDA)**:
   - Conduct exploratory analysis to understand the relationships between income and key demographic factors (age, gender, education, marital status, occupation).
   - Visualize data distributions and correlations to identify patterns.
3. **Segmentation**:
   - Divide the dataset into two income groups (less than 50,000 and greater than $50,000).
   - Perform statistical analysis to identify significant differences between these two groups in terms of demographic variables.

1. **Factor Grouping**:

- Identify which demographic variables are most strongly correlated with income using feature importance techniques (e.g., decision trees, random forests).
- Group related factors to ensure the marketing team can tailor outreach to specific demographics.

1. **Marketing Strategy Recommendations**:

- Based on the segmentation and predictive model, provide recommendations for UVW College’s marketing campaigns (e.g., focusing on tuition, program types, delivery formats).

By following this approach, the objective is to deliver an application that can predict income based on demographic factors, thereby allowing UVW College to strategically target potential students for its programs.

Dataset

age: continuous.

workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

education-num: continuous.

marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.

relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.

sex: Female, Male.

capital-gain: continuous.

capital-loss: continuous.

hours-per-week: continuous.

native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

In [172]:

```
import pandas as pd
import numpy as np
import matplotlib_inline
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

In [173]:

```
from google.colab import drive
drive.mount("/content/drive")
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount
("/content/drive", force_remount=True).
```

In [174]:

```
# Specify the path to the folder after mounting
Data = pd.read_csv("/content/drive/MyDrive/ASU_CSE578_DV/dataset.csv")
```

In [175]:

```
df = Data.copy()
```

Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is an essential first step in understanding the structure of your data, uncovering patterns, and identifying potential problems. This process involves summarizing the data visually and numerically, checking for missing values, understanding the relationships between variables, and detecting any anomalies or outliers.

Dataset

• Numerical Columns (6): o age: Age of the individual. o education-num: Number of years of education. o capital-gain: Capital gains from investments. o capital-loss: Capital losses from investments. o hours-per-week: Number of hours worked per week. • Categorical Columns (8): o workclass: Employment type o education:

Highest level of education attained (e.g., Bachelors, High School, etc.). o marital-status: Marital status o occupation: Job occupation o relationship: Relationship status o race: Race of the individual. o sex: Gender of the individual. o native-country: Country of origin (e.g., United States, Mexico). • Target Column (1): o income: Whether the individual's income is above or below $50,000 (<=50K or >50K).

In [176]:

```
df.tail(2)
```

Out[176]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	h
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	
32560	52	Self-emp-inc	287927	HS-grad	9	Married-civ-spouse	Exec-managerial	Wife	White	Female	15024	0	

In [177]:

```
df.head(2)
```

Out[177]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13

In [178]:

```
df.sample()
```

Out[178]:

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	h
26824	25	Private	122489	Bachelors	13	Never-married	Exec-managerial	Own-child	White	Female	0	1726	

In [179]:

```
df.describe()
```

Out[179]:

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
count	32561.000000	3.256100e+04	32561.000000	32561.000000	32561.000000	32561.000000
mean	38.581647	1.897784e+05	10.080679	1077.648844	87.303830	40.437456
std	13.640433	1.055500e+05	2.572720	7385.292085	402.960219	12.347429
min	17.000000	1.228500e+04	1.000000	0.000000	0.000000	1.000000
25%	28.000000	1.178270e+05	9.000000	0.000000	0.000000	40.000000

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
50%	37.000000	1.783560e+05	10.000000	0.000000	0.000000	40.000000
75%	48.000000	2.370510e+05	12.000000	0.000000	0.000000	45.000000
max	90.000000	1.484705e+06	16.000000	99999.000000	4356.000000	99.000000

In [180]:

```
df.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   education-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   sex             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 [181]:

```
df.shape
```

Out[181]:

```
(32561, 15)
```

In [182]:

```
df.education.unique()
```

Out[182]:

```
array([' Bachelors', ' HS-grad', ' 11th', ' Masters', ' 9th',
       ' Some-college', ' Assoc-acdm', ' Assoc-voc', ' 7th-8th',
       ' Doctorate', ' Prof-school', ' 5th-6th', ' 10th', ' 1st-4th',
       ' Preschool', ' 12th'], dtype=object)
```

In [183]:

```
df.isna().any().any()
```

Out[183]:

```
False
```

In [184]:

```
df.isnull().sum()
```

Out[184]:

	0
age	0
workclass	0
fnlwgt	0

education	0
education-num	0
marital-status	0
occupation	0
relationship	0
race	0
sex	0
capital-gain	0
capital-loss	0
hours-per-week	0
native-country	0
income	0

dtype: int64

Data Cleaning

The code lines provided are part of a data cleaning process, specifically for replacing certain unwanted values (in this case, the string ? or its variations) with NaN (Not a Number, a placeholder for missing or undefined data).

In [185]:

```
df.replace(' ?', np.nan, inplace=True)
df.replace('?', np.nan, inplace=True)
df.replace('? ', np.nan, inplace=True)
```

Univariate and Bivariate Analysis

Univariate and bivariate analysis are essential steps in the exploratory data analysis (EDA) process. These analyses help understand the distribution and relationships in your dataset, providing insights that can guide your decision-making for feature engineering and modeling.

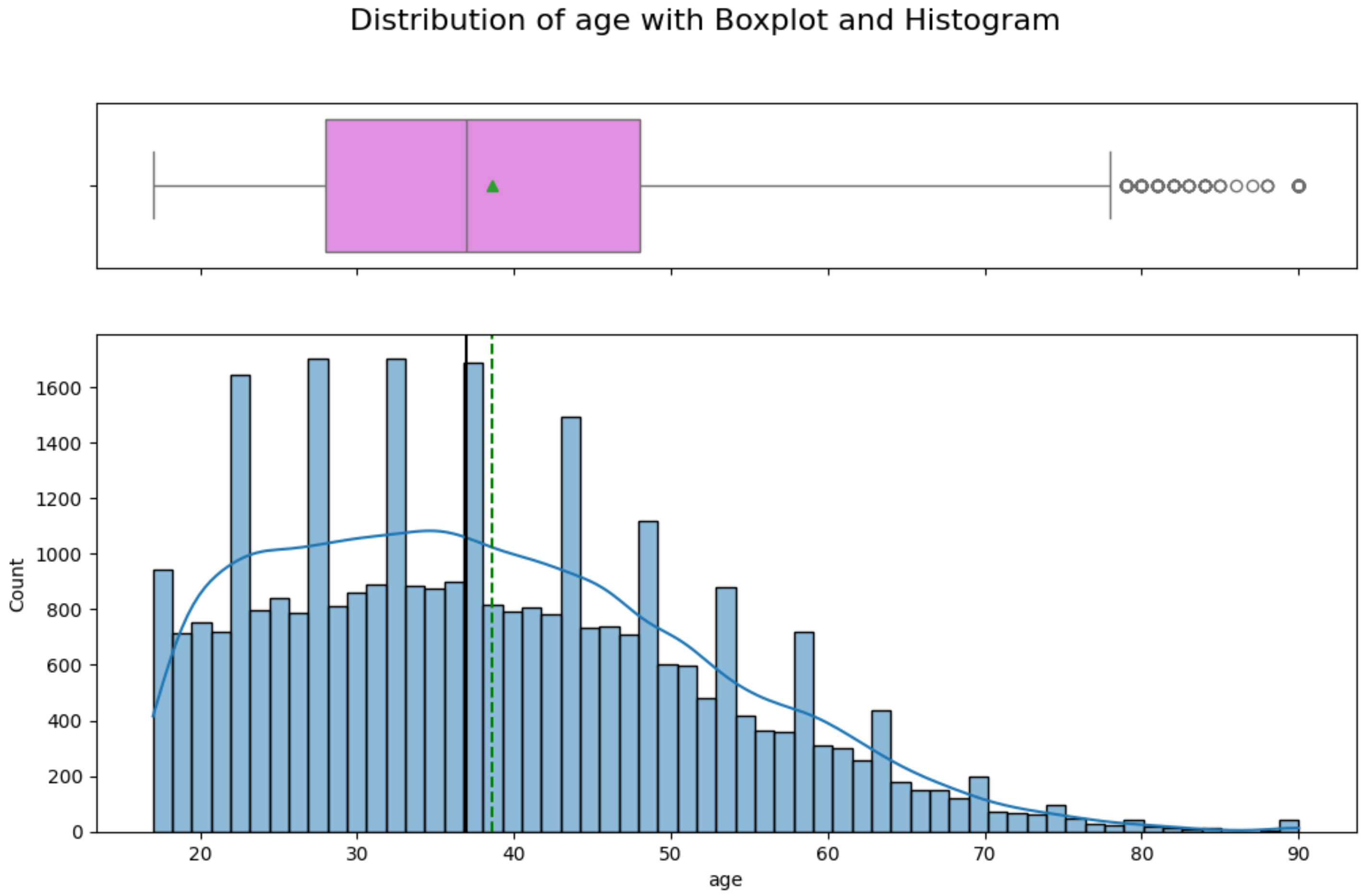
CASE 1 - "Examining the Relationship Between Age and Income"

In [215]:

```
# function to plot a boxplot and a histogram along the same scale.


def histogram_boxplot(data, feature, figsize=(12, 7), kde=False, bins=None):
    """
    Boxplot and histogram combined

    data: dataframe
    feature: dataframe column
    figsize: size of figure (default (12,7))
    kde: whether to the show density curve (default False)
    bins: number of bins for histogram (default None)
    """
    f2, (ax_box2, ax_hist2) = plt.subplots(
        nrows=2,  # Number of rows of the subplot grid= 2
        sharex=True,  # x-axis will be shared among all subplots
        gridspec_kw={"height_ratios": (0.25, 0.75)},
        figsize=figsize,
    )  # creating the 2 subplots
    sns.boxplot(
        data=data, x=feature, ax=ax_box2, showmeans=True, color="violet"
    )  # boxplot will be created and a triangle will indicate the mean value of the colum
n
    sns.histplot(
        data=data, x=feature, kde=kde, ax=ax_hist2, bins=bins, palette="winter"
    ) if bins else sns.histplot(
        data=data, x=feature, kde=kde, ax=ax_hist2
    )  # For histogram
    ax_hist2.axvline(
        data[feature].mean(), color="green", linestyle="--"
    )  # Add mean to the histogram
    ax_hist2.axvline(
        data[feature].median(), color="black", linestyle="-"
    )  # Add median to the histogram
    f2.suptitle(f'Distribution of {feature} with Boxplot and Histogram', fontsize=16)
```

In [216]:

```
histogram_boxplot(df, "age", kde=True)
```



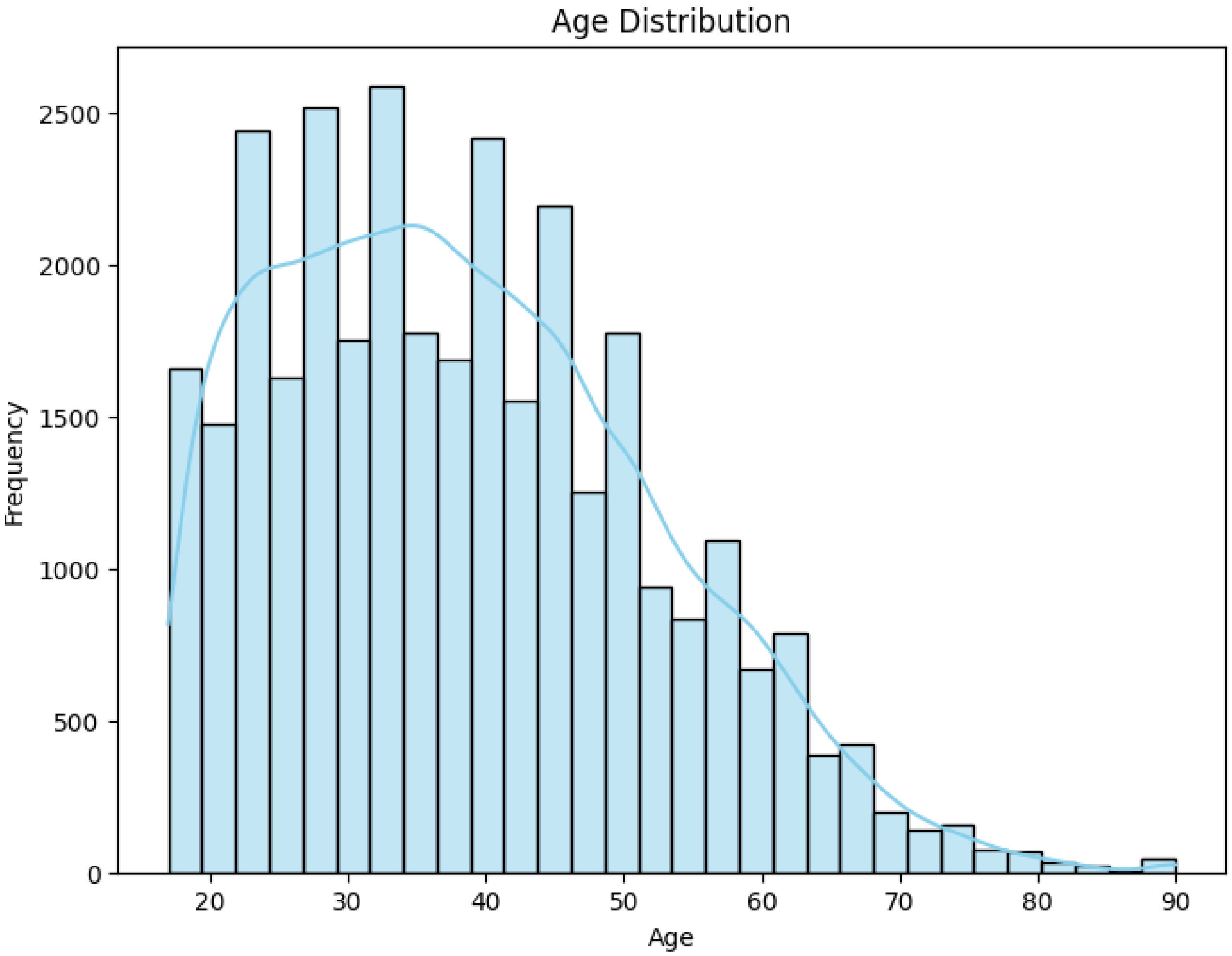
This code defines a function stacked_barplot that creates a stacked bar chart to visualize the relationship between two categorical variables (one independent and one target variable) in a given dataset.

In [217]:

```
import pandas as pd
import matplotlib.pyplot as plt

# Function to plot stacked bar chart
def stacked_barplot(data, predictor, target):
    """
    Print the category counts and plot a stacked bar chart

    data: dataframe
    predictor: independent variable
    target: target variable
    """
    count = data[predictor].nunique()
    sorter = data[target].value_counts().index[-1]
    tab1 = pd.crosstab(data[predictor], data[target], margins=True).sort_values(
        by=sorter, ascending=False
    )
    print(tab1)
    print("-" * 120)
    tab = pd.crosstab(data[predictor], data[target], normalize="index").sort_values(
        by=sorter, ascending=False
    )
    tab.plot(kind="bar", stacked=True, figsize=(count + 1, 5))
    plt.legend(
        loc="lower left", frameon=False,
    )
    plt.legend(loc="upper left", bbox_to_anchor=(1, 1))

    # Add a title to the chart
    plt.title(f'Stacked Bar Chart of {predictor} vs {target}', fontsize=16)

    # Show the plot
    plt.show()
```

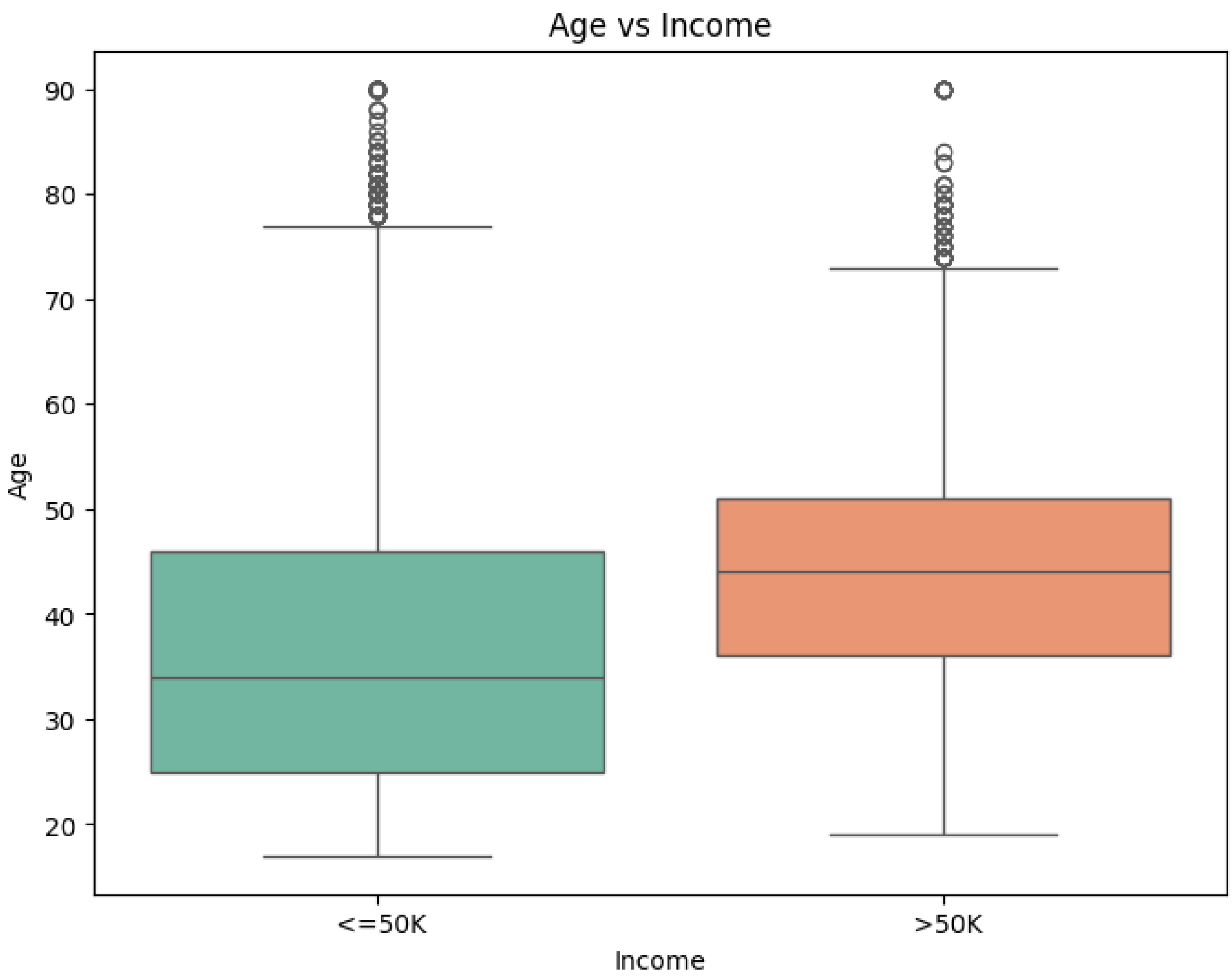
In [189]:

```
# Plotting Age Distribution
plt.figure(figsize=(8,6))
sns.histplot(df['age'], kde=True, bins=30, color='skyblue')
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



In [190]:

```
# Plotting Age vs Income
plt.figure(figsize=(8,6))
sns.boxplot(x='income', y='age', data=df, palette='Set2')
plt.title('Age vs Income')
plt.xlabel('Income')
plt.ylabel('Age')
plt.show()
```



CASE 1 - "Examining the Relationship Between Age and Income"

Analysis of the Box Plot reveals a discernible relationship between age and income, indicating a tendency for older individuals to have a higher likelihood of earning over $50K.

1. **Visualization: Box Plot - Age vs. Income (<=50K vs. >50K)**

This box plot demonstrates the distribution of age across two income categories: <=50K and >50K. It visually compares the median, quartiles, and range of ages within each income group, highlighting potential differences in age distribution between those earning less than or equal to 50,000 and those earning more than $50,000.

1. **Design Process: • Data Preparation: The dataset was cleaned and organized to include age and income information for each individual. The income variable was categorized into two groups: <=50K and >50K. • Chart Selection: A box plot was selected as the most appropriate visualization to represent the data. This choice was based on the need to show the distribution of age across two income categories and to highlight potential differences in age distribution.**

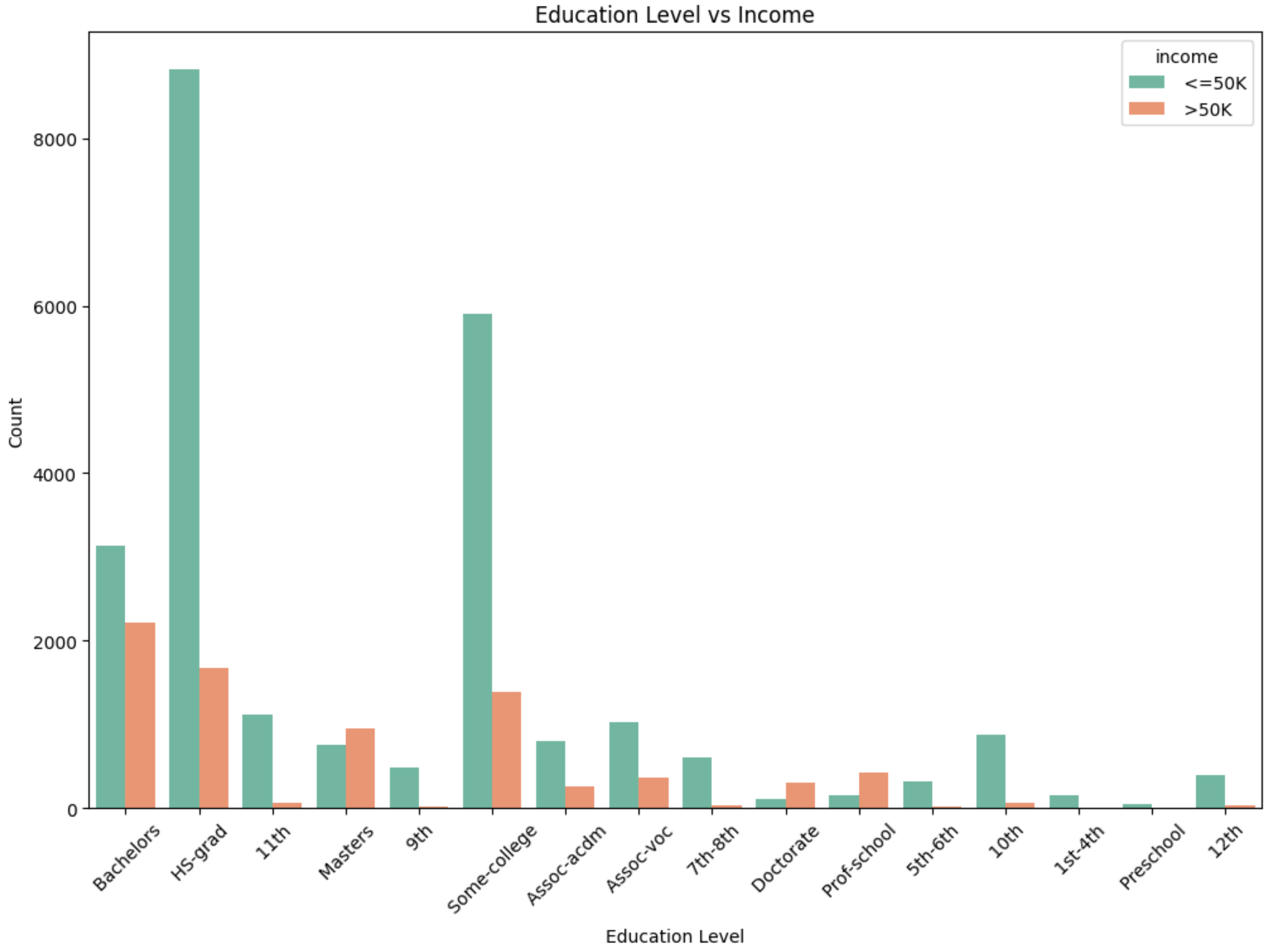
1. **Conclusion: • The box plot effectively demonstrates a relationship between age and income. The median age for individuals earning >50K is noticeably higher than the median age for those earning <=50K, suggesting that older individuals are more likely to earn more.**

CASE 2 - "The Impact of Education on Earning Potential"

In [191]:

```
plt.figure(figsize=(12, 8))
sns.countplot(data=df, x='education', hue='income', palette='Set2')
```

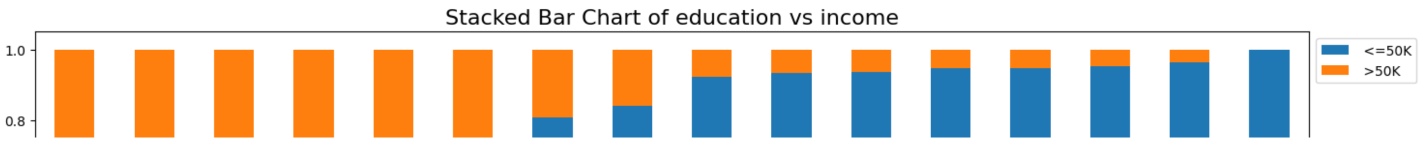
```
plt.title('Education Level vs Income')
plt.xlabel('Education Level')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```

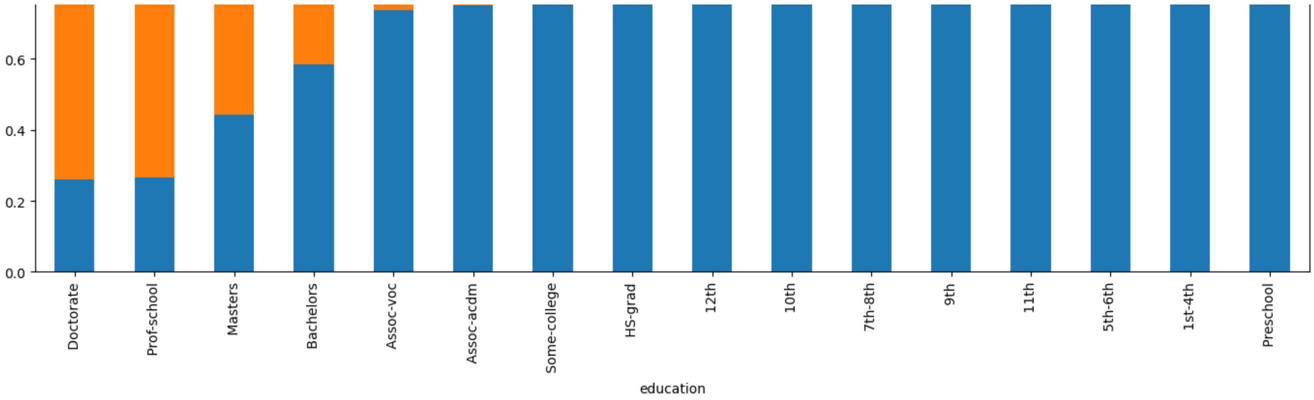


In [218]:

```
stacked_barplot(df,"education", "income")
```

```
income          <=50K   >50K    All
education
All             24720   7841  32561
 Bachelors       3134   2221   5355
 HS-grad         8826   1675  10501
 Some-college    5904   1387   7291
 Masters          764    959   1723
 Prof-school      153    423    576
 Assoc-voc       1021    361   1382
 Doctorate        107    306    413
 Assoc-acdm       802    265   1067
 10th             871     62    933
 11th            1115     60   1175
 7th-8th          606     40    646
 12th             400     33    433
 9th              487     27    514
 5th-6th          317     16    333
 1st-4th          162      6    168
 Preschool         51      0     51
----------------------------------------------------------------------------------------
--------------------------------
```





CASE 2 Summary - "The Impact of Education on Earning Potential"

Visualization: Stacked Bar Chart - Education vs. Income (<=50K vs. >50K)

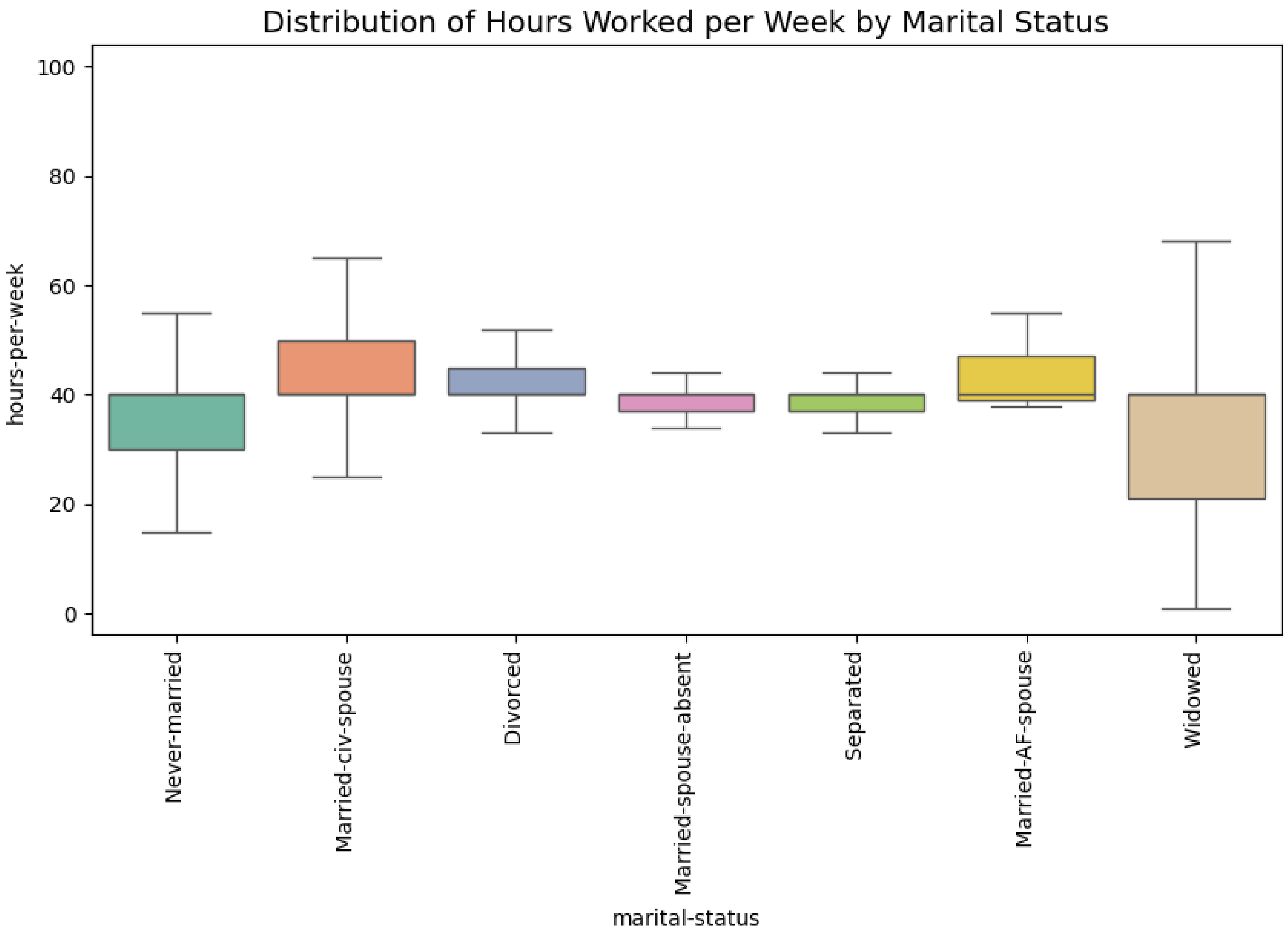
1. **Explanation of the Visualization and Rationale: • Demonstration: This stacked bar chart demonstrates the relationship between education level and income, specifically the proportion of individuals earning <=50K and >50K within each education category. It visually compares the distribution of income across different education levels. • Rationale: The stacked bar chart was chosen to effectively show the composition of income groups within each education category. It allows for a direct comparison of the relative proportions of <=50K and >50K earners for each level of educational attainment. This visualization is ideal for showcasing how the distribution of income changes as education level increases. It's clear and intuitive, making it easy to understand the correlation between education and income.**
2. **Design Process: • Data Preparation: The initial step involved organizing and cleaning the data to group individuals by their education level and income bracket (<=50K and >50K). The data was then aggregated to calculate the proportion of each income group within each education category. • Chart Selection: A stacked bar chart was selected as the most appropriate visualization to represent the data. This choice was based on the need to show the composition of income groups within each education category and to facilitate easy comparisons across different education levels.**
3. **Conclusion: The stacked bar chart effectively demonstrates the strong positive correlation between education level and earning potential. The visualization clearly shows that individuals with higher education levels are significantly more likely to earn over $50K. • The chart highlights the importance of education as a pathway to economic opportunity and higher income levels. • The "inversion point" around the "Some-college" and "Assoc-acdm" categories provides a clear visual indication of the threshold where education begins to significantly impact income. • The high proportion of >50K earners in the Doctorate and Prof-school categories underscores the economic benefits of pursuing advanced degrees. In conclusion, this stacked bar chart provides a clear and compelling visual representation of the relationship between education and income, making it a valuable tool for understanding the economic impact of educational attainment.**

CASE 3 "Exploring the Impact of Marital Status and Work Hours on Income"**

In [214]:

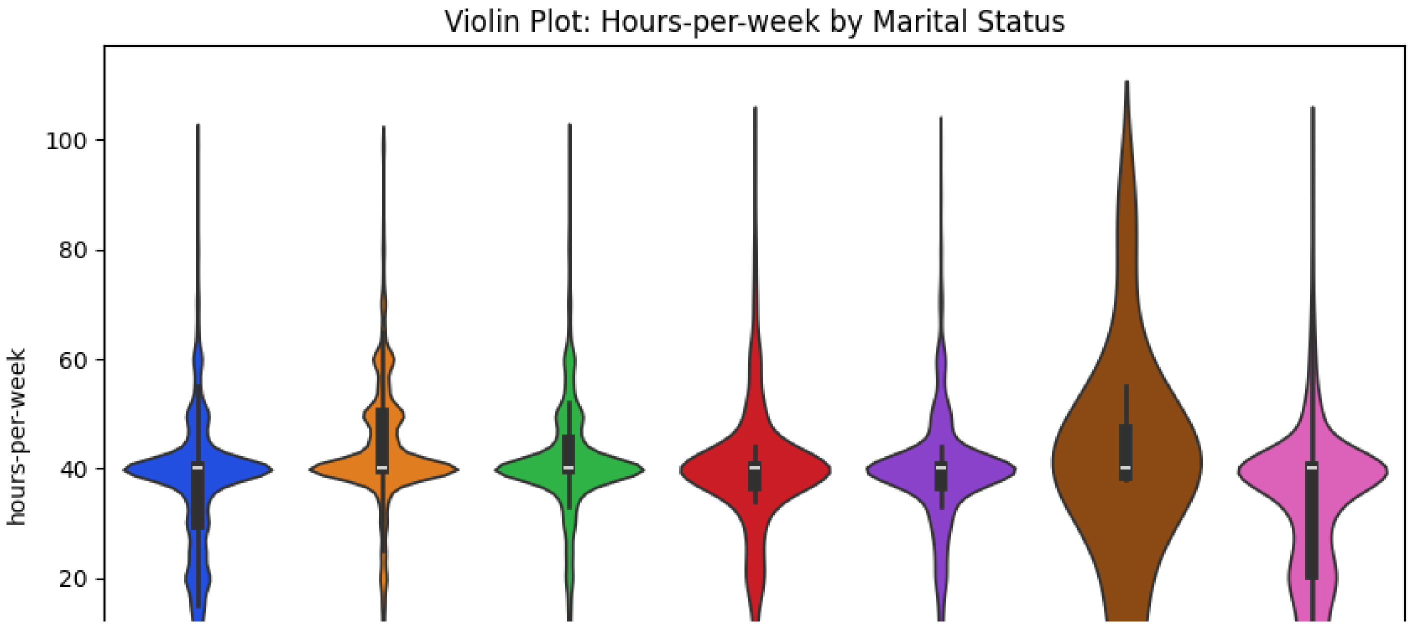
```
import seaborn as sns
import matplotlib.pyplot as plt

# Create the boxplot without fliers
plt.figure(figsize=(10,5))
sns.boxplot(data=df, x='marital-status', y='hours-per-week', fliersize=0, palette='Set2'
)

# Add a title to the plot
plt.title('Distribution of Hours Worked per Week by Marital Status', fontsize=14)

# Rotate the x-axis labels
plt.xticks(rotation=90)
```

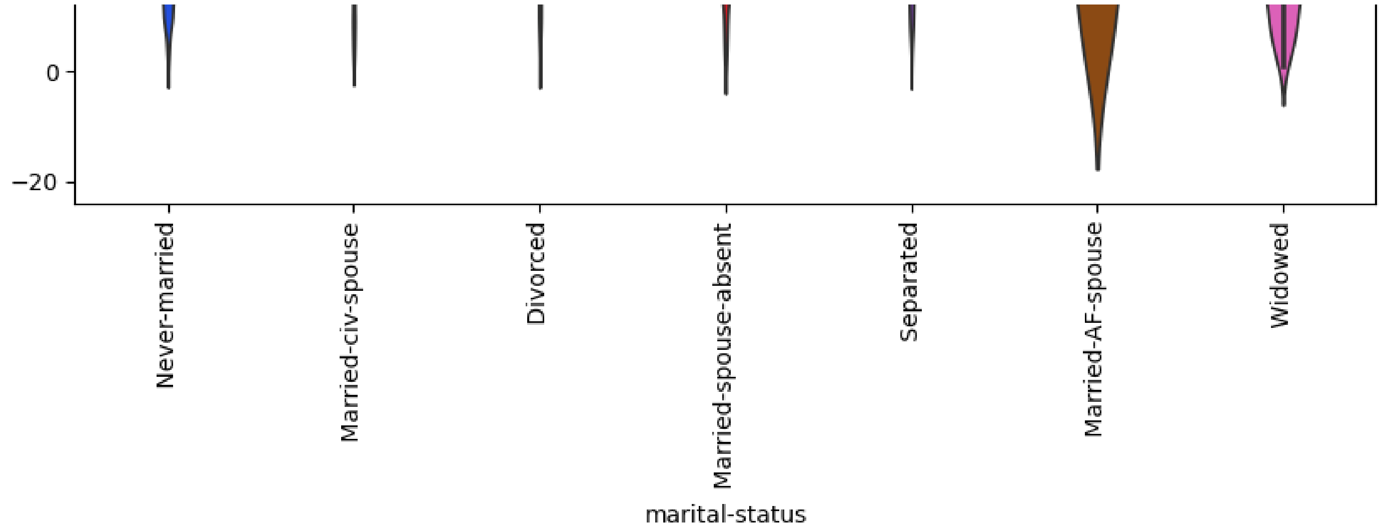
```
# Show the plot
plt.show()
```



In [194]:

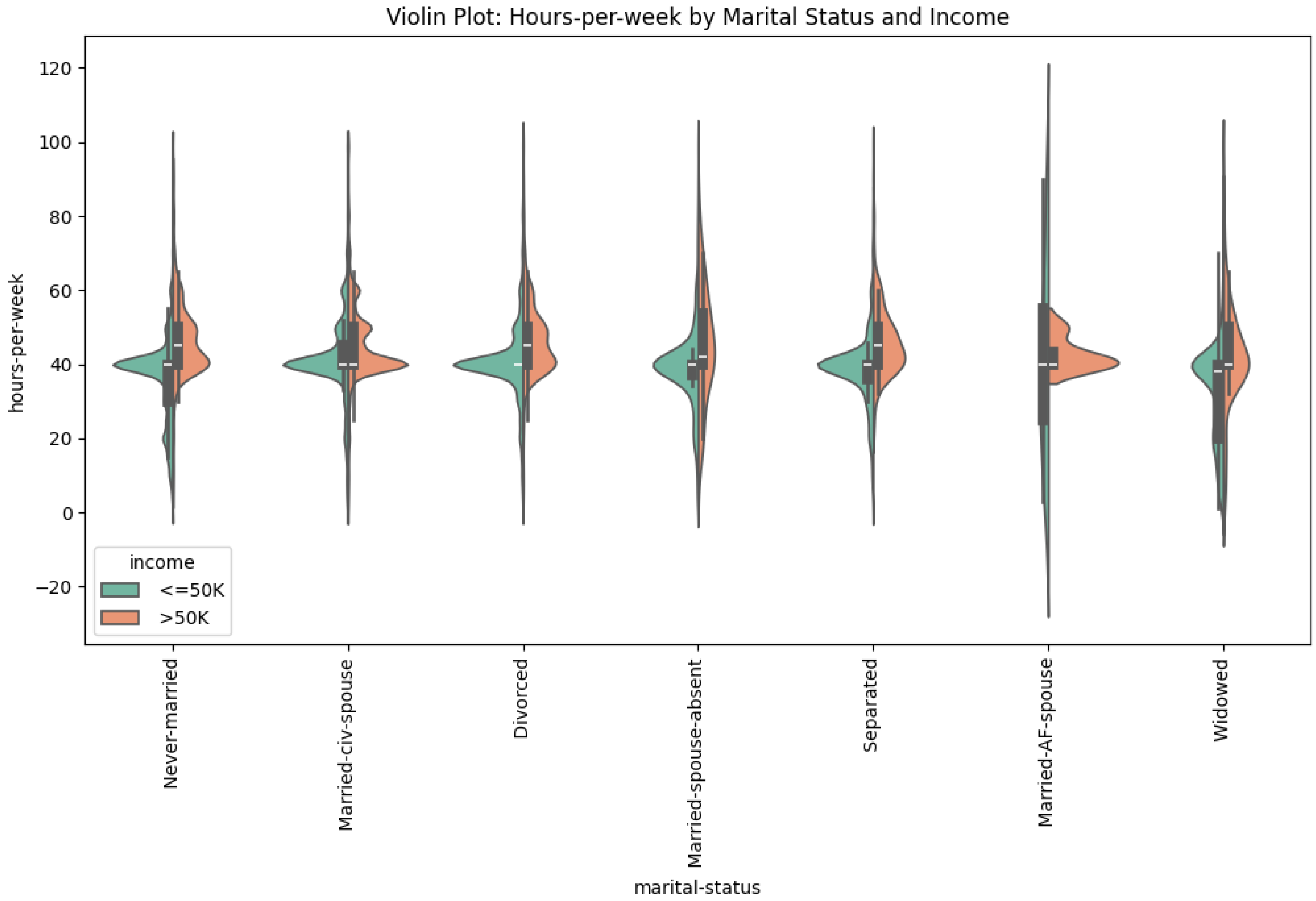
```
import seaborn as sns
import matplotlib.pyplot as plt

# Create a Violin Plot for Hours-per-week by Marital Status with a bright color palette
plt.figure(figsize=(10, 6))
sns.violinplot(x='marital-status', y='hours-per-week', data=df, palette='bright')  # Using 'bright' palette
plt.title('Violin Plot: Hours-per-week by Marital Status')
plt.xticks(rotation=90)
plt.show()
```





In [195]:

```python
import seaborn as sns
import matplotlib.pyplot as plt

# Create a Violin Plot for Hours-per-week by Marital Status and Income with a bright color palette
plt.figure(figsize=(12, 6))
sns.violinplot(x='marital-status', y='hours-per-week', hue='income', data=df, palette='Set2', split=True)
plt.title('Violin Plot: Hours-per-week by Marital Status and Income')
plt.xticks(rotation=90)
plt.show()
```



CASE 3 "Exploring the Impact of Marital Status and Work Hours on Income"**

1. **Explanation of the Visualization and Rationale: • Demonstration: This violin plot demonstrates the**

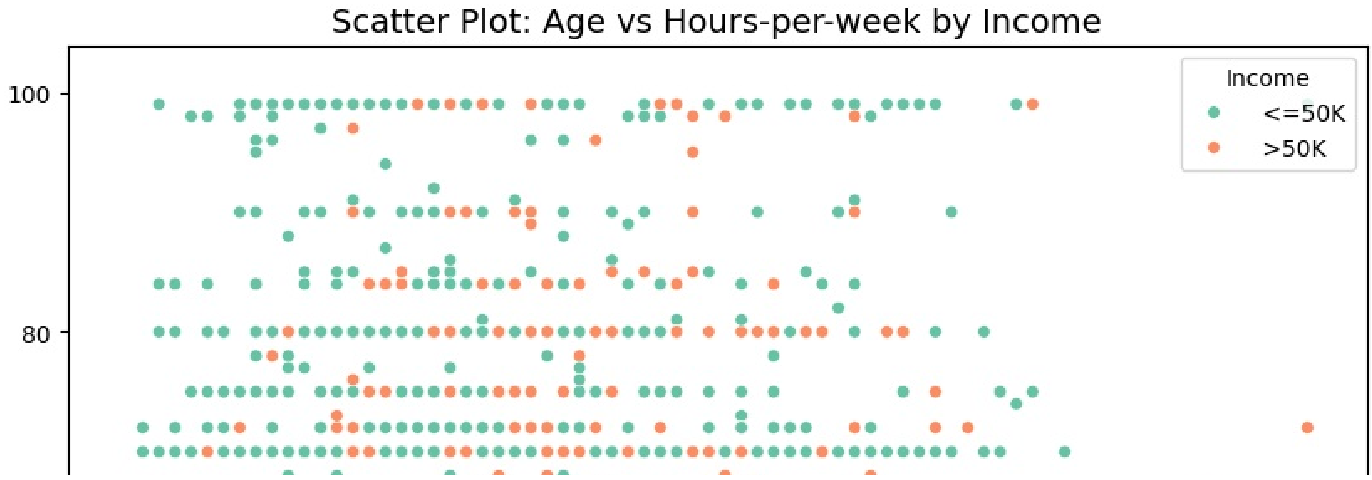
distribution of hours worked per week across different marital status categories, further segmented by income levels (<=50K and >50K). It visually compares the shape and spread of the data, highlighting differences in typical hours worked and the range of hours worked for each group.

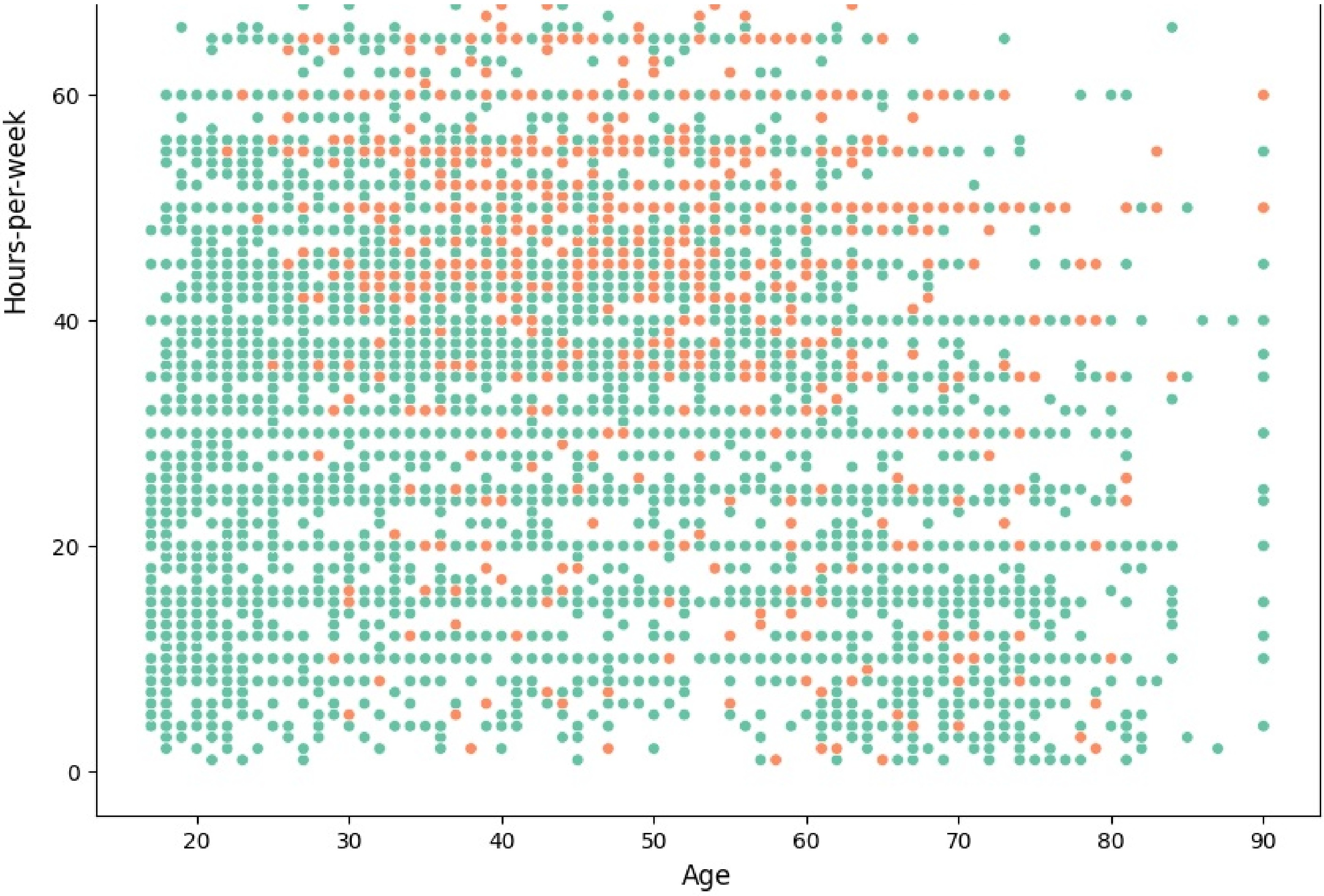
1. **Design Process: • Data Preparation: The initial step involved organizing and cleaning the data to group individuals by their marital status and income bracket (<=50K and >50K). The data was then aggregated to calculate the distribution of hours worked per week for each group. • Chart Selection: A violin plot was selected as the most appropriate visualization to represent the data. This choice was based on the need to show the distribution of a continuous variable (hours-per-week) across multiple categorical variables (marital status and income). • Axis Definition: The x-axis was defined to represent the marital status categories, arranged for clear comparison. The y-axis was defined to represent the hours worked per week, ranging from -20 to 120 to encompass the entire range of values. • Violin Creation: Violin plots were created for each marital status category, split by income level (<=50K and >50K). The width of each violin represents the density of data points at different values of hours worked per week.**
2. **Conclusion: The violin plot effectively demonstrates the complex relationship between marital status, income, and hours worked per week. The visualization clearly shows that: • Married-civ-spouse Stands Out: Individuals who are "Married-civ-spouse" and earn >50K tend to work significantly longer hours compared to other groups. This suggests a strong correlation between working longer hours and higher income for married individuals with civilian spouses. • Marital Status Influences Work Patterns: The distribution of hours worked varies across different marital status categories, indicating that marital status plays a role in influencing work patterns. In conclusion, this violin plot provides a comprehensive and insightful view into the interplay between marital status, income, and hours worked per week. It highlights the importance of considering these factors together to understand work patterns and income disparities.**

CASE 4 - Interesting Case: The "Workaholic Senior" Phenomenon

In [196]:

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


# Create the scatter plot with 'age' vs 'hours-per-week' and use 'income' as hue
plt.figure(figsize=(10, 10))
sns.scatterplot(data=df, x='age', y='hours-per-week', hue='income', palette='Set2', s=30
)

# Add title and labels
plt.title('Scatter Plot: Age vs Hours-per-week by Income', fontsize=14)
plt.xlabel('Age', fontsize=12)
plt.ylabel('Hours-per-week', fontsize=12)

# Show plot
plt.legend(title='Income')
plt.show()
```





CASE 4 - Interesting Case: The "Workaholic Senior" Phenomenon

We're interested in exploring the relationship between age, hours worked per week, and income, specifically focusing on the potential for older individuals to work longer hours and earn more. We're particularly curious to see if there's evidence of a "workaholic senior" phenomenon, where older individuals continue to work long hours, potentially to maintain income or pursue personal fulfillment. Visualization: Scatter Plot - Age vs. Hours-per-week by Income

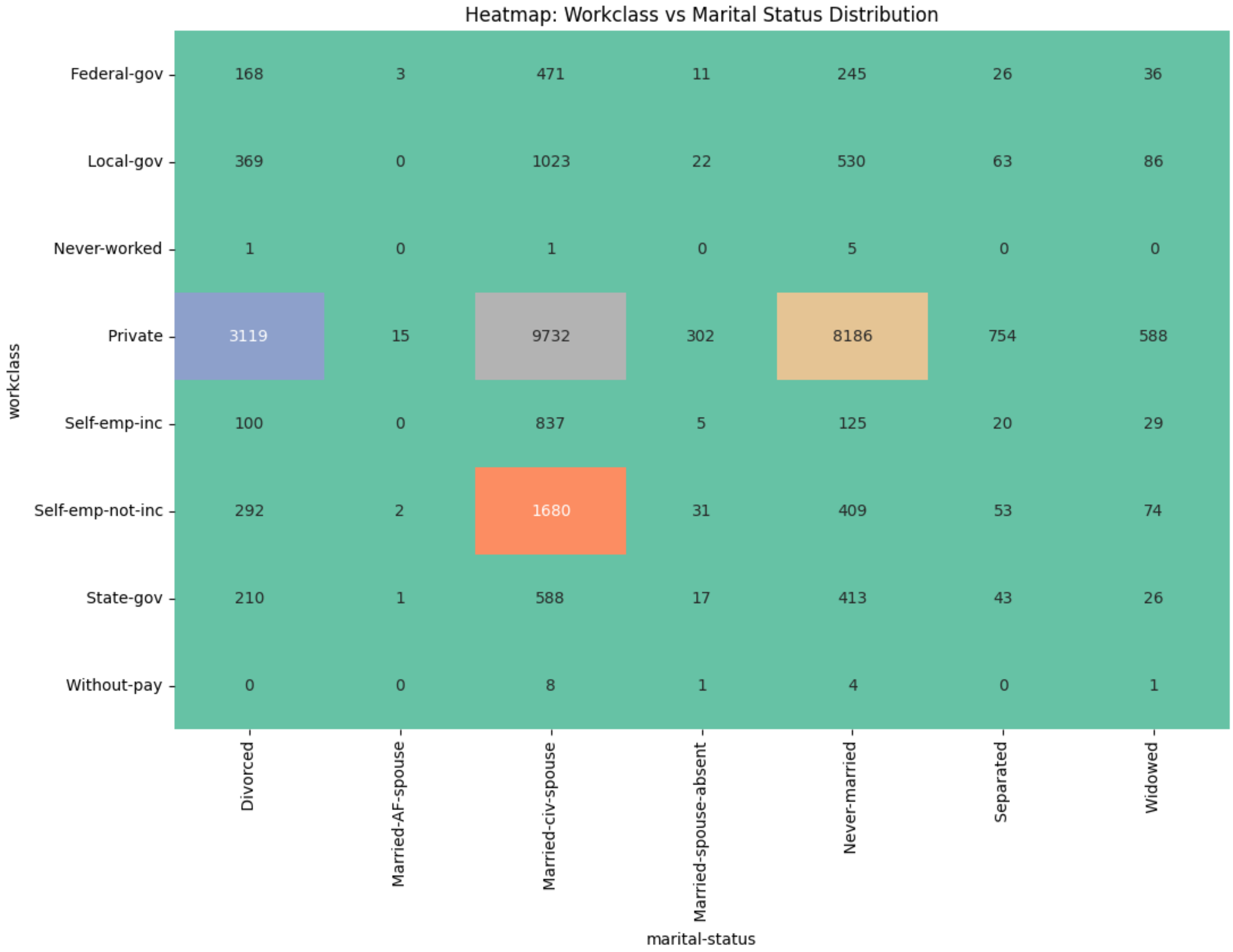
1. Explanation of the Visualization and Rationale: • Demonstration: This scatter plot demonstrates the relationship between age, hours worked per week, and income (<=50K and >50K). Each point represents an individual, with the x-coordinate representing age, the y-coordinate representing hours worked per week, and the color representing income. It allows for a visual exploration of potential trends and patterns between these three variables. • Rationale: A scatter plot was chosen to effectively visualize the relationship between two continuous variables (age and hours-per-week) while also incorporating a categorical variable (income) through color coding. This type of plot is ideal for identifying potential correlations, clusters, and outliers. It allows us to observe whether there's a trend of older individuals working longer hours, and whether this trend is associated with higher income.
2. Design Process: • Data Preparation: The initial step involved cleaning and organizing the data to include age, hours-per-week, and income information for each individual. The income variable was categorized into <=50K and >50K. • Chart Selection: A scatter plot was selected as the most appropriate visualization to represent the data. This choice was based on the need to show the relationship between two continuous variables and to incorporate a categorical variable.
3. Conclusion: The scatter plot provides valuable insights into the relationship between age, hours worked per week, and income. • Higher Concentration of >50K at Older Ages: There's a higher concentration of individuals earning >50K in the older age ranges, particularly above 60. This aligns with the understanding that older individuals may have accumulated more experience and skills, leading to higher-paying positions. • Potential "Workaholic Seniors": While not a dominant trend, there are individuals in the older age ranges (70+) who work significantly longer hours (60+ per week) and earn >50K. This provides some evidence for the "workaholic senior" phenomenon. • Spread of Hours Worked: The hours worked per week are spread across a wide range for all age groups, indicating that various factors influence work patterns, regardless of age. • Income Distribution: The distribution of income is not uniform across age groups. There's a tendency

for the >50K group to be more prevalent in the older age ranges, but there are also individuals earning <=50K in these ranges.

CASE 5 - "Analyzing the Relationship Between Workclass and Marital Status"

In [212]:

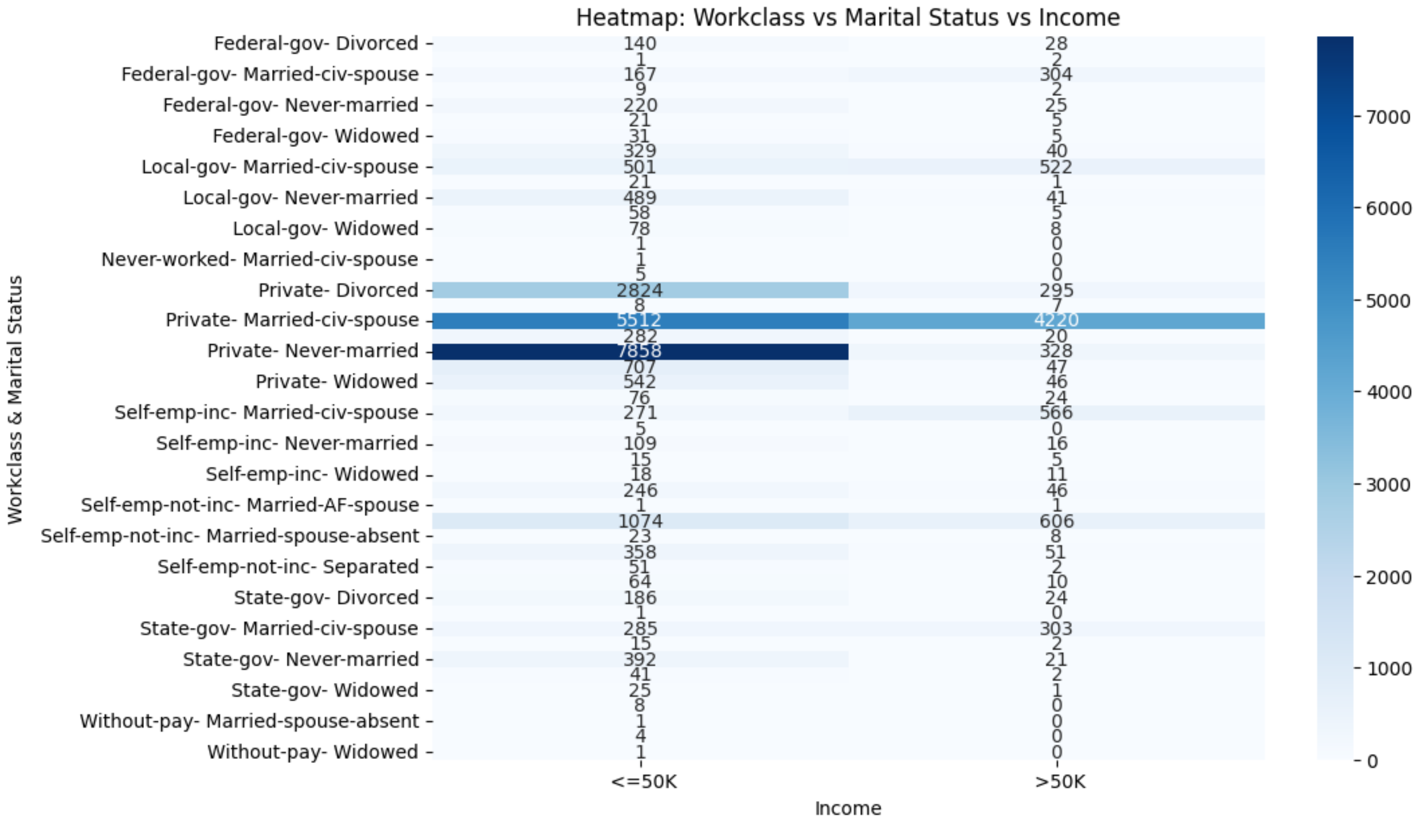
```
# Create a crosstab between workclass and education
workclass_education = pd.crosstab(df['workclass'], df['marital-status'])

# Plot the Heatmap for categorical data
plt.figure(figsize=(12, 8))
sns.heatmap(workclass_education, annot=True, cmap='Set2', fmt='d', cbar=False)
plt.title('Heatmap: Workclass vs Marital Status Distribution')
plt.show()
```



workclass	Divorced	Married-AF-spouse	Married-civ-spouse	Married-spouse-absent	Never-married	Separated	Widowed
Federal-gov	168	3	471	11	245	26	36
Local-gov	369	0	1023	22	530	63	86
Never-worked	1	0	1	0	5	0	0
Private	3119	15	9732	302	8186	754	588
Self-emp-inc	100	0	837	5	125	20	29
Self-emp-not-inc	292	2	1680	31	409	53	74
State-gov	210	1	588	17	413	43	26
Without-pay	0	0	8	1	4	0	1

In [208]:

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



# Step 1: Create a crosstab of 'workclass', 'marital-status', and 'income'
crosstab = pd.crosstab([df['workclass'], df['marital-status']], df['income'])

# Step 2: Visualize the crosstab using a heatmap
plt.figure(figsize=(10, 7))
sns.heatmap(crosstab, annot=True, fmt='d', cmap='Blues', cbar=True)
plt.title('Heatmap: Workclass vs Marital Status vs Income')
plt.xlabel('Income')
plt.ylabel('Workclass & Marital Status')
plt.show()
```



CASE 5 - "Analyzing the Relationship Between Workclass and Marital Status"

This heatmap reveals the distribution of individuals across different workclasses and marital statuses. The "Private" workclass dominates, with the highest counts, particularly for "Married-civ-spouse" and "Never-married" individuals. "Local-gov" and "Federal-gov" show significant numbers, with "Married-civ-spouse" being the most common. Self-employed categories have lower counts, while "Never-worked" and "Without-pay" show minimal representation. The heatmap highlights the prevalence of the "Private" workclass and the strong association between marital status and workclass distribution.

Visualization: Heatmap - Workclass vs. Marital Status vs. Income

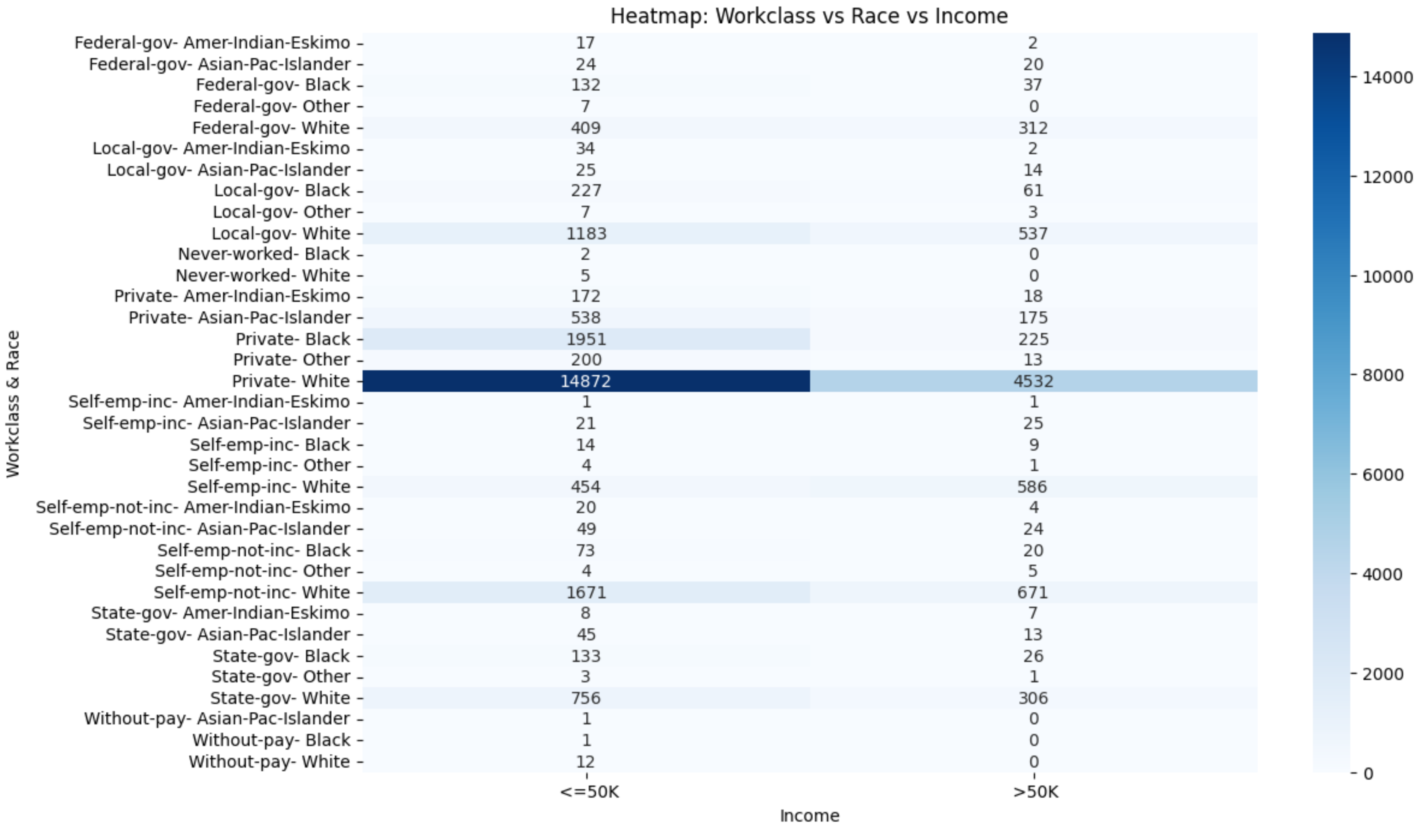
1. **Explanation of the Visualization and Rationale: • Demonstration: This heatmap visualizes the distribution of individuals across different combinations of workclass, marital status, and income levels (<=50K and >50K). It displays the count of individuals in each specific category, with color intensity representing the magnitude of these counts. • Rationale: A heatmap was chosen to effectively represent the relationship between three categorical variables. It allows for a clear and concise visual representation of the data, highlighting patterns and concentrations of individuals across different categories. The color gradient enables quick identification of categories with higher or lower counts, making it easy to compare and contrast different combinations.**
2. **Design Process: • Data Preparation: The dataset was cleaned and organized to create a table with counts of individuals for each unique combination of workclass, marital status, and income.The counts were calculated for each category, which would be the numerical data represented by the heatmap. • Chart Selection: A heatmap was selected as the most appropriate visualization due to its ability to effectively display the relationships between multiple categorical variables and the magnitude of counts.**
3. **Conclusion: • The heatmap effectively highlights the dominant role of the "Private" workclass, particularly for "Married-civ-spouse" and "Never-married" individuals. This indicates a significant concentration of individuals in these categories. • The visualization also reveals a clear pattern of income distribution across different workclasses and marital statuses. For example, within the "Private" workclass, there's a significant number of individuals in both income brackets, but the distribution varies depending on marital status.**

CASE 6 "Disparities in WorkClass Distribution Across Racial

Groups"

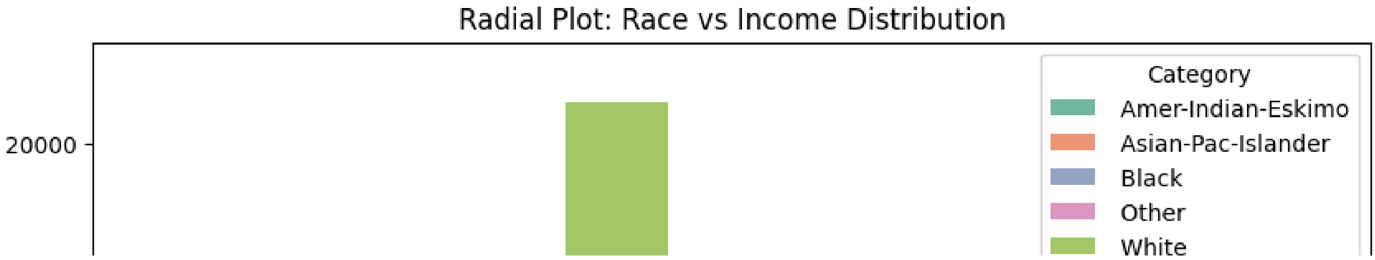
In [209]:

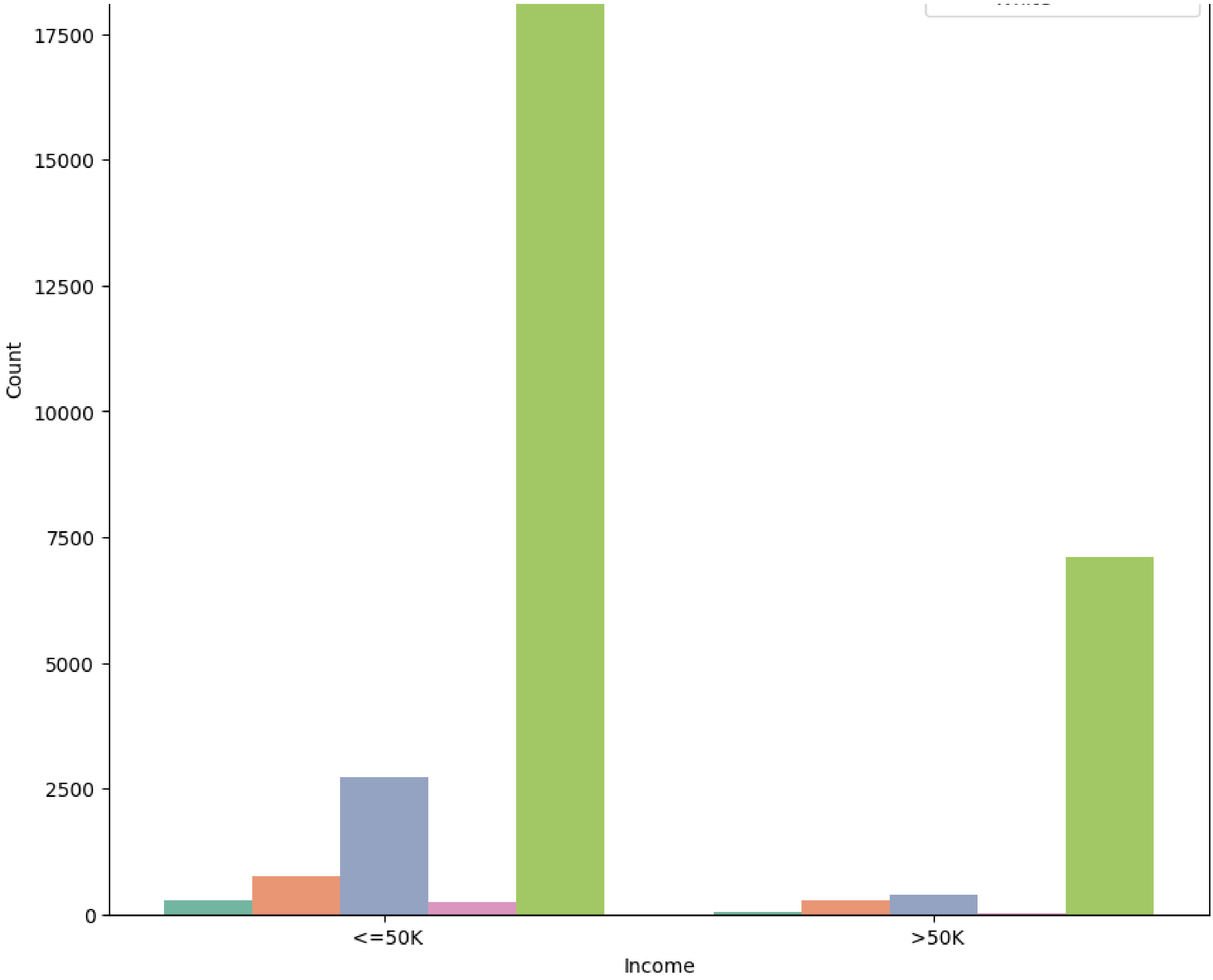
```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Step 1: Create a crosstab of 'workclass', 'race', and 'income'
crosstab = pd.crosstab([df['workclass'], df['race']], df['income'])

# Step 2: Visualize the crosstab using a heatmap
plt.figure(figsize=(12, 8))  # Adjust figure size for better readability
sns.heatmap(crosstab, annot=True, fmt='d', cmap='Blues', cbar=True)
plt.title('Heatmap: Workclass vs Race vs Income')
plt.xlabel('Income')
plt.ylabel('Workclass & Race')
plt.show()
```



In [213]:

```
# Create a circular plot for Race and Income
race_income_counts = pd.crosstab(df['race'], df['income'])

# Plot the Radial Plot (Circular plot)
race_income_counts = race_income_counts.reset_index().melt(id_vars=['race'], value_vars=
race_income_counts.columns)
race_income_counts.columns = ['Category', 'Income', 'Count']

plt.figure(figsize=(10, 10))
sns.barplot(x='Income', y='Count', hue='Category', data=race_income_counts, palette='Set
2')
plt.title('Radial Plot: Race vs Income Distribution')
plt.show()
```





CASE 6 - "Disparities in WorkClass Distribution Across Racial Groups"

This heatmap illustrates the distribution of individuals across different workclasses, races, and income brackets (<=50K and >50K). The "Private" workclass, particularly for White individuals, dominates with the highest counts, especially in the <=50K category. Significant numbers of White individuals are also seen in the >50K bracket within the "Private" sector. Government sectors ("Federal-gov" and "Local-gov") show a higher concentration of White individuals, with a notable portion earning >50K. Black individuals are predominantly in the "Private" and "Local-gov" sectors, largely in the <=50K bracket. Other racial groups show smaller counts across all categories. The heatmap highlights the disparity in income and representation across different workclasses and racial groups, with a clear concentration of White individuals in the higher-income bracket, particularly within the "Private" sector.

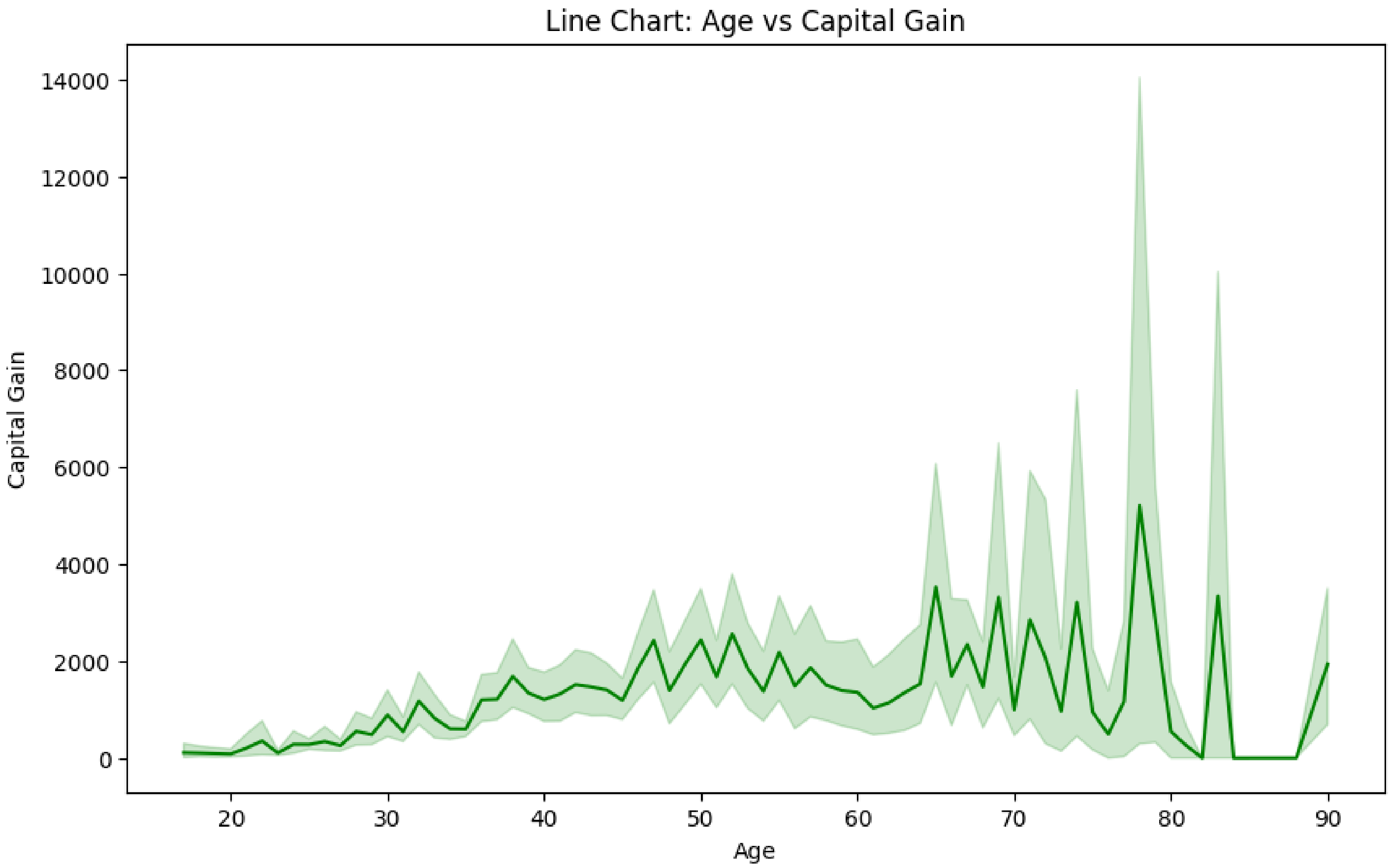
This radial plot, more accurately described as a grouped bar chart, illustrates the income distribution (<=50K and >50K) across different racial categories. White individuals dominate both income brackets, particularly in the <=50K range. While still significant, the number of White individuals earning >50K is notably lower. Black individuals show a substantial presence in the <=50K category but are significantly underrepresented in the >50K category. Other racial groups (Amer-Indian-Eskimo, Asian-Pac-Islander, and Other) show relatively low counts across both income brackets. The chart highlights a clear disparity in income distribution, with White individuals having a significantly larger presence in both income brackets, and Black individuals being disproportionately represented in the lower income bracket.

Additional Plots to Check patterns in the data

Capital Gain/Loss Vs. Age

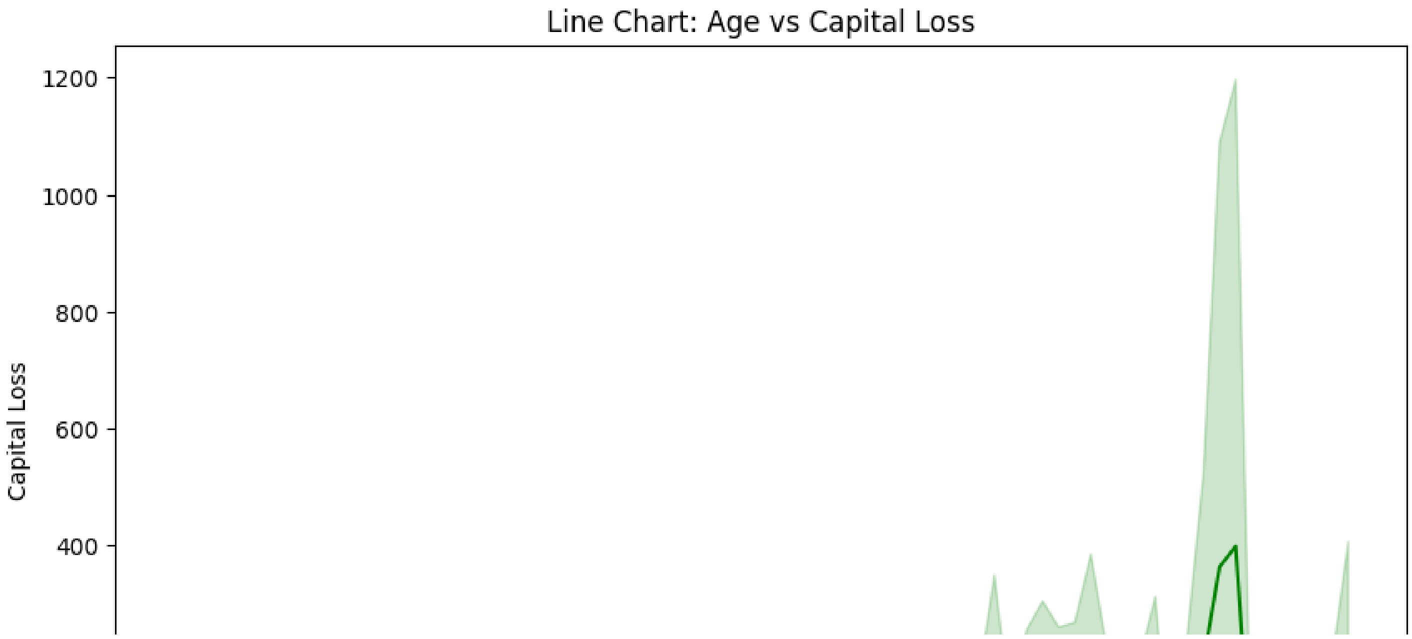
In [197]:

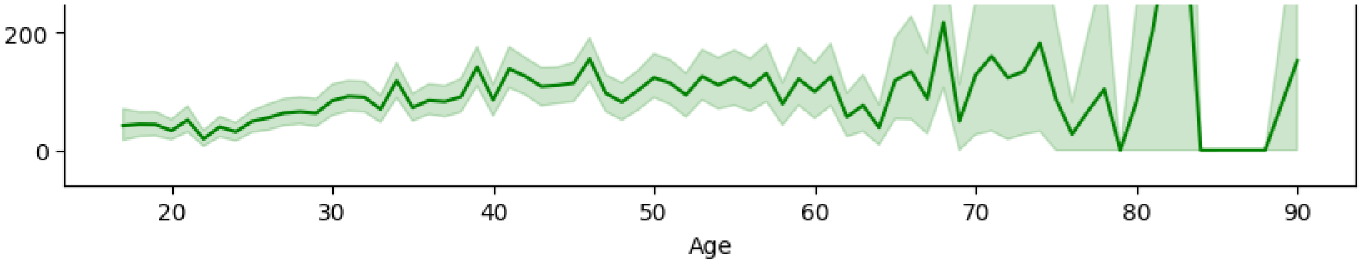
```
# Plotting Line chart for Age vs Capital-gain
plt.figure(figsize=(10, 6))
sns.lineplot(x='age', y='capital-gain', data=df, color='green')
plt.title('Line Chart: Age vs Capital Gain')
plt.xlabel('Age')
plt.ylabel('Capital Gain')
plt.show()
```



In [198]:

```
# Plotting Line chart for Age vs Capital-gain
plt.figure(figsize=(10, 6))
sns.lineplot(x='age', y='capital-loss', data=df, color='green')
plt.title('Line Chart: Age vs Capital Loss')
plt.xlabel('Age')
plt.ylabel('Capital Loss')
plt.show()
```

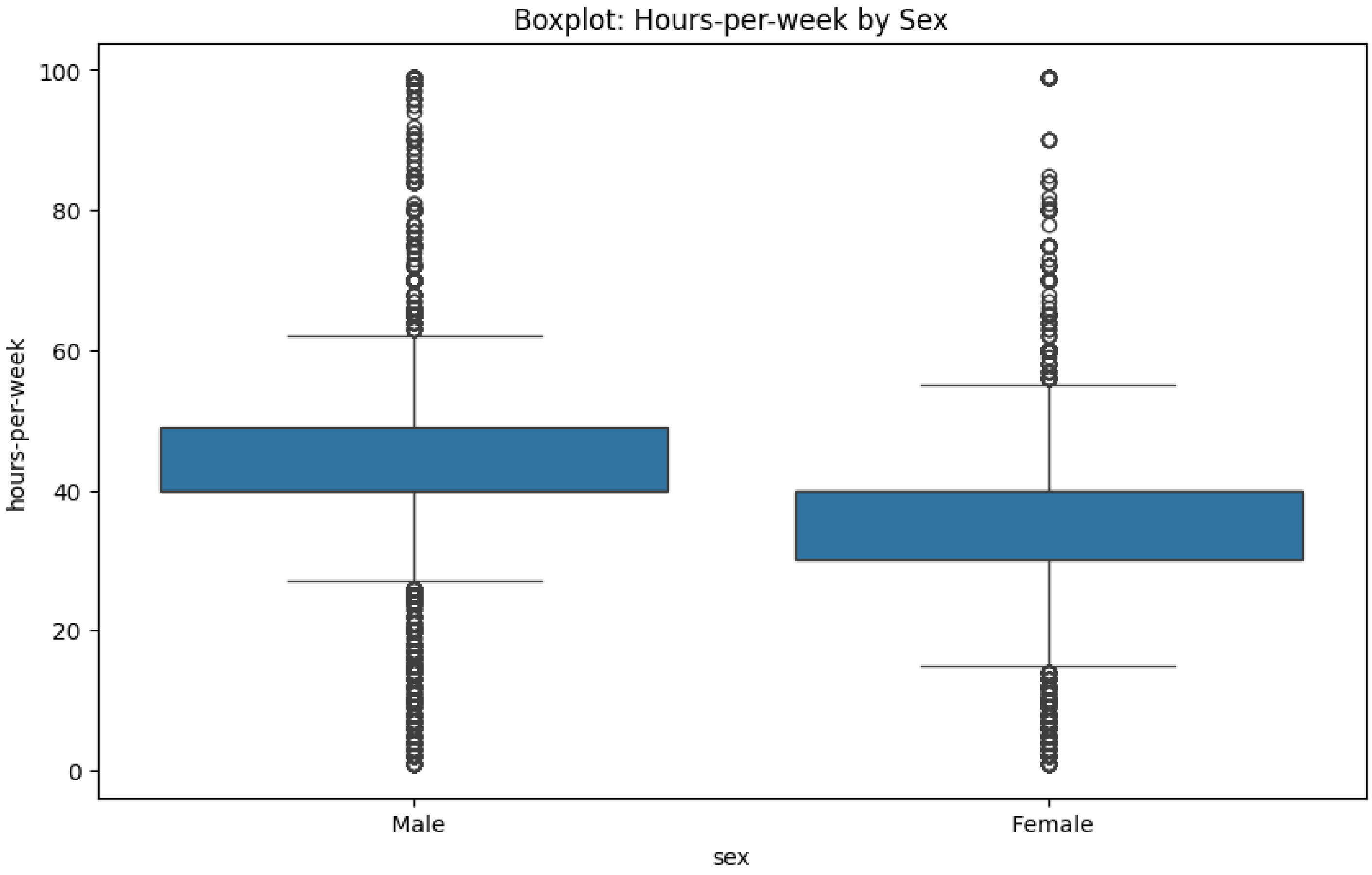




Hours Per Week by Sex

In [202]:

```
# Create a Boxplot for Hours-per-week by Sex
plt.figure(figsize=(10, 6))
sns.boxplot(x='sex', y='hours-per-week', data=df)
plt.title('Boxplot: Hours-per-week by Sex')
plt.show()
```

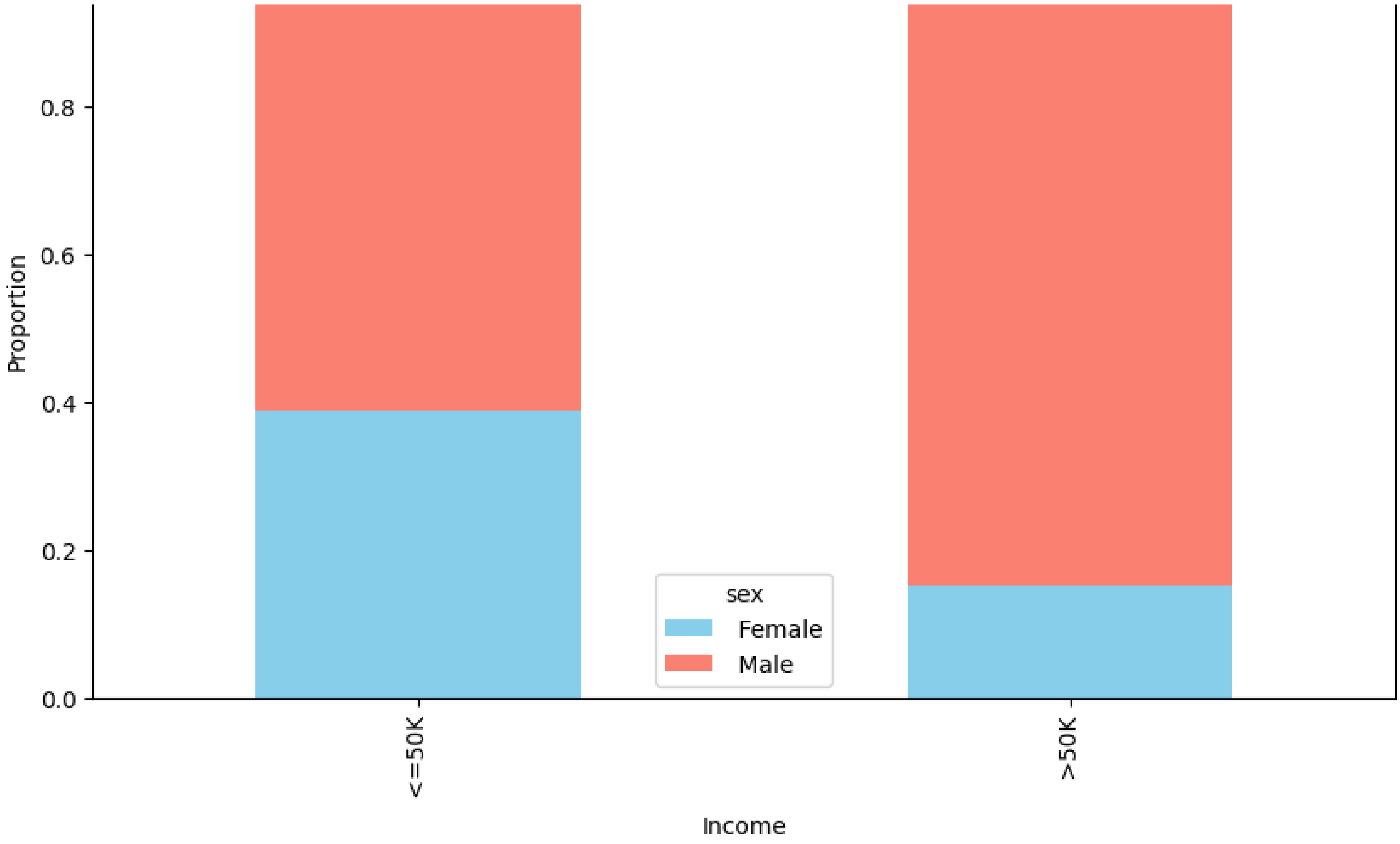


Stacked Bar Chart: Proportions of Sex within Income Categories

In [203]:

```
# Create a Stacked Bar Chart for Sex within Income
income_sex = pd.crosstab(df['income'], df['sex'], normalize='index')

# Plot the Stacked Bar Chart
income_sex.plot(kind='bar', stacked=True, figsize=(10, 6), color=['skyblue', 'salmon'])
plt.title('Stacked Bar Chart: Proportions of Sex within Income Categories')
plt.xlabel('Income')
plt.ylabel('Proportion')
plt.show()
```

Stacked Bar Chart: Proportions of Sex within Income Categories
1.0

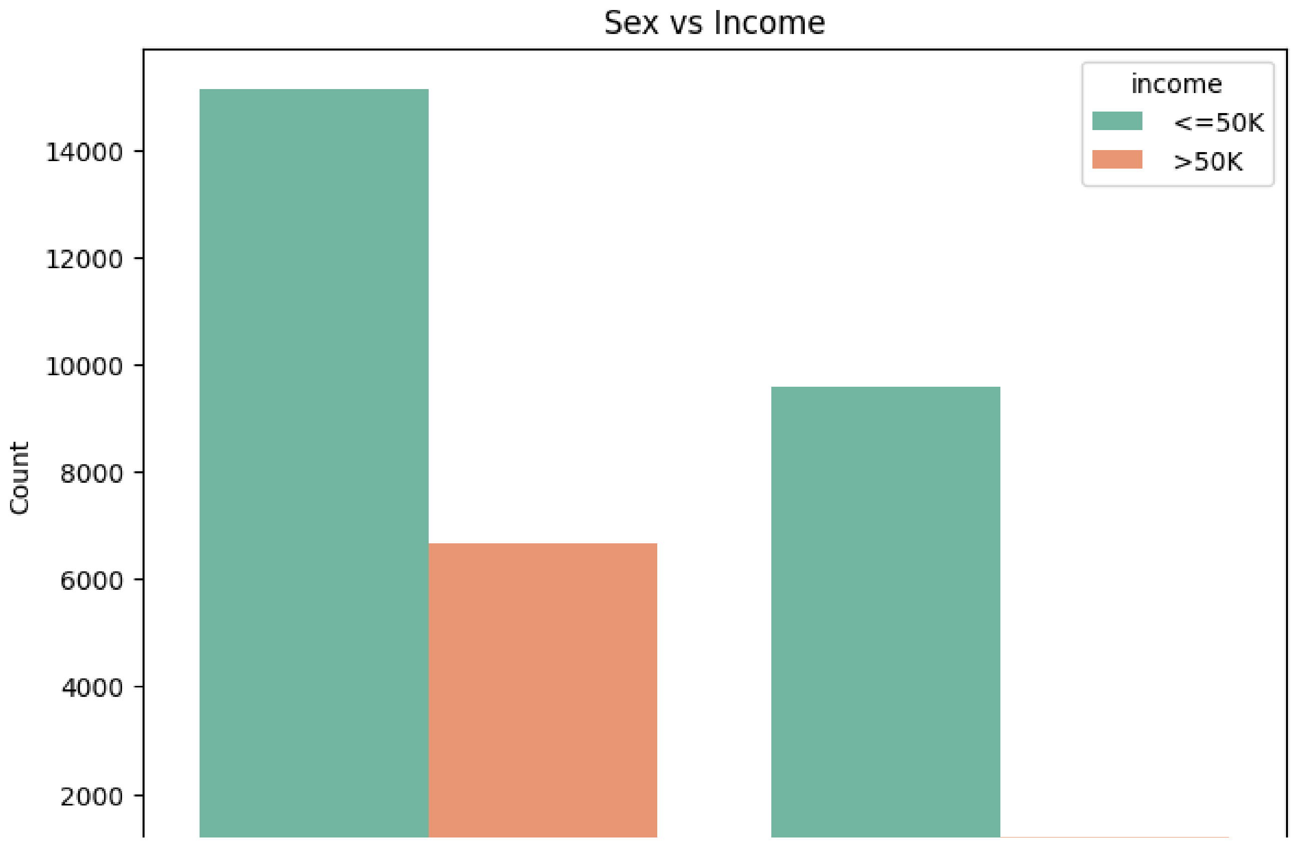


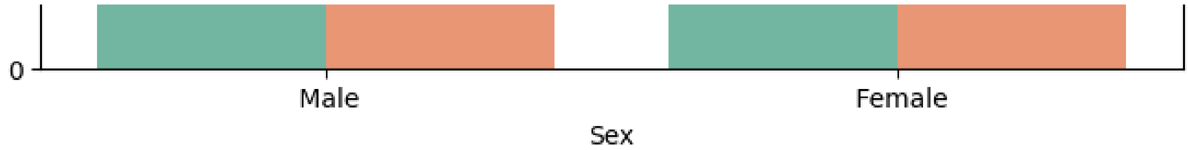
Sex vs Income

Sex is a categorical variable, and we can examine how it influences income by using a count plot.

In [204]:

```
# Plotting Sex vs Income
plt.figure(figsize=(8,6))
sns.countplot(data=df, x='sex', hue='income', palette='Set2')
plt.title('Sex vs Income')
plt.xlabel('Sex')
plt.ylabel('Count')
plt.show()
```

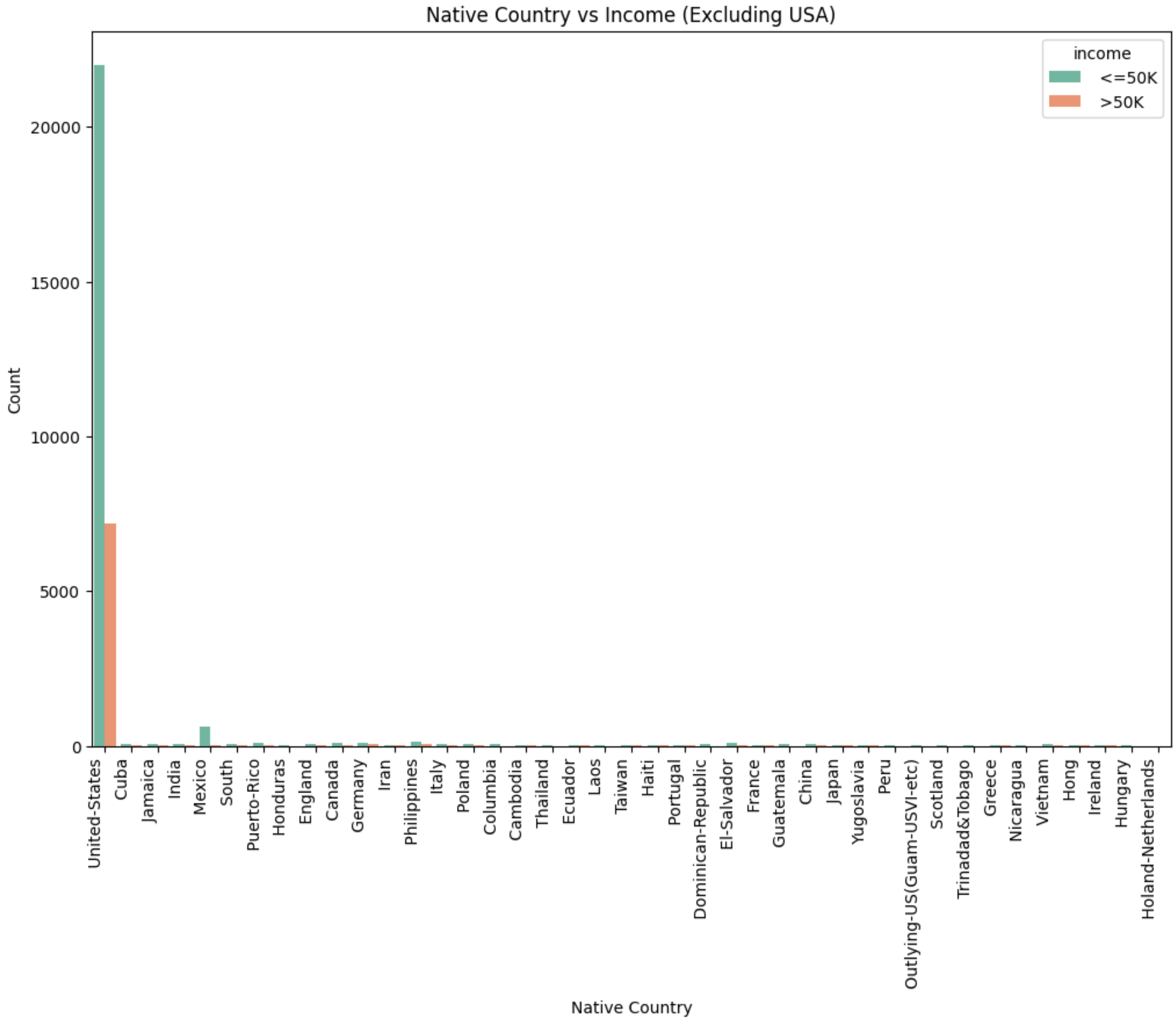




Native Country Vs. Income

In [205]:

```python
import matplotlib.pyplot as plt
import seaborn as sns

# Exclude rows where the 'native-country' is 'United States' (or 'USA' if it appears that
way in your data)
df_filtered = df[df['native-country'] != 'United-States']

# Plotting Native Country vs Income (excluding USA)
plt.figure(figsize=(12,8))
sns.countplot(data=df_filtered, x='native-country', hue='income', palette='Set2')
plt.title('Native Country vs Income (Excluding USA)')
plt.xlabel('Native Country')
plt.ylabel('Count')
plt.xticks(rotation=90, ha='right')
plt.show()
```

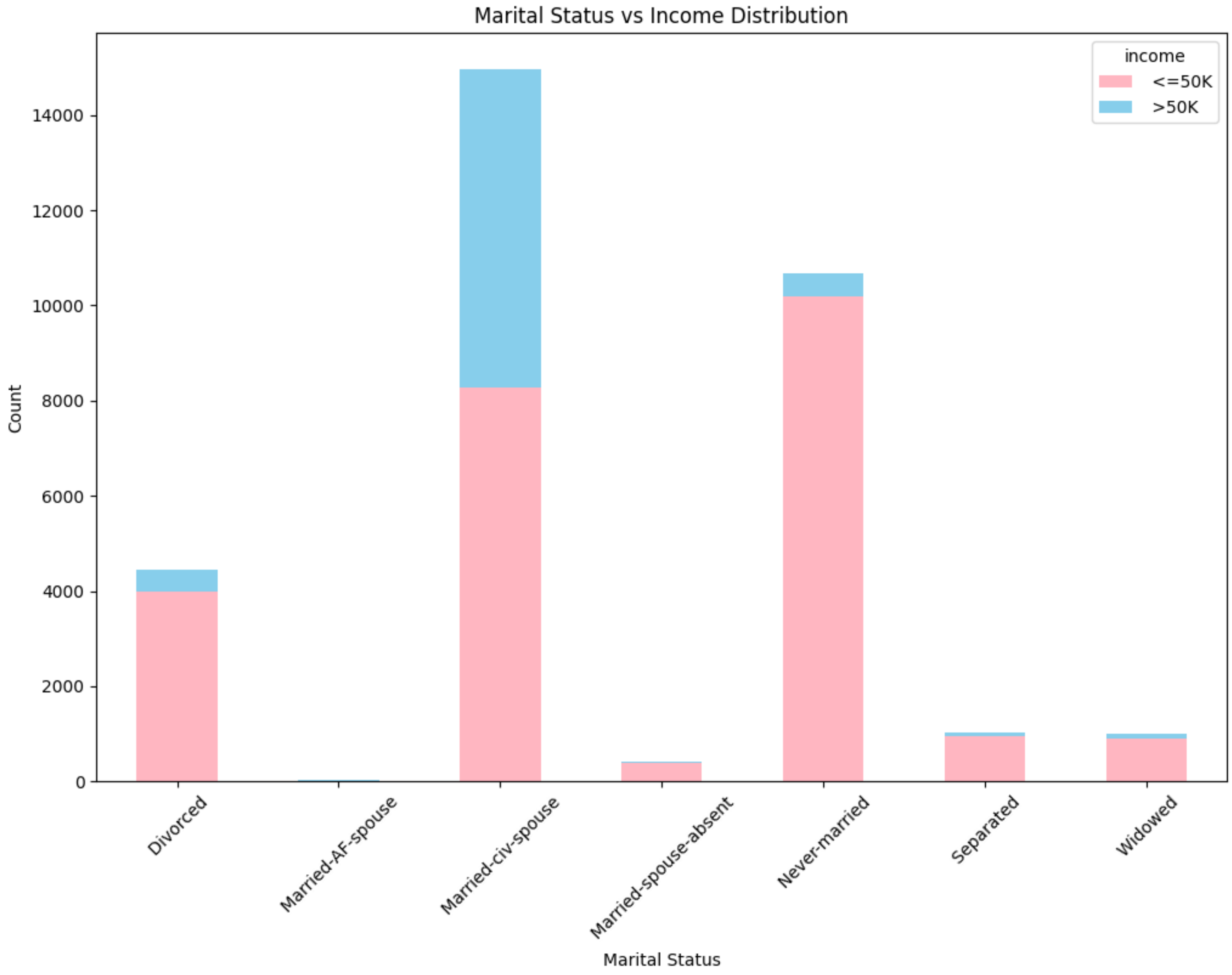


Marital Status vs Income Distribution

In [206]:

```
plt.figure(figsize=(12, 8))
df_marital_income = pd.crosstab(df['marital-status'], df['income'])
df_marital_income.plot(kind='bar', stacked=True, color=['#FFB6C1', '#87CEEB'], figsize=(
12, 8))
plt.title('Marital Status vs Income Distribution')
plt.xlabel('Marital Status')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.show()
```

```
<Figure size 1200x800 with 0 Axes>
```



Marketing Strategies:

1. Target Yonger Individuals for Higher Enrollment

- **Insight:** The box plot analysis reveals that younger individuals are more likely to earn under $50K.
- **Recommendation:** Focus marketing efforts on older individuals, particularly those below the age of 40. UVW College could design specialized programs (e.g., part-time degrees, certifications) that appeal to this demographic, offering flexible scheduling and online learning options to accommodate their busy lives.

2. Highlight the Economic Benefits of Higher Education

- **Insight:** The stacked bar chart clearly demonstrates a strong correlation between higher education and income levels. Individuals with higher educational attainment, particularly those with Doctorate and Prof-school degrees, are more likely to earn over $50K.
- **Recommendation:** UVW College should emphasize the long-term financial benefits of obtaining a higher

education. This messaging should be incorporated into marketing materials, highlighting how education leads to higher earning potential. Scholarships or financial aid targeted towards students pursuing higher education should also be promoted.

3. Appeal to Married Individuals with Higher Work Hours

- **Insight:** The violin plot shows that married individuals, especially those with civilian spouses, tend to work longer hours and are more likely to earn above $50K.
- **Recommendation:** Create marketing campaigns tailored to married individuals looking to balance work, family, and education. UVW College can market flexible and online programs that accommodate full-time working professionals, emphasizing the ability to balance work and family responsibilities while achieving higher earnings through education.

4. Cater to Older Adults Who Continue Working

- **Insight:** The scatter plot suggests that older adults, especially those over 60, tend to work longer hours and are more likely to earn more. This may indicate a desire or necessity to continue working longer for financial stability or personal fulfillment.
- **Recommendation:** UVW College should target older professionals, particularly those over 60, who may be interested in upskilling or changing careers. The marketing team can create outreach programs that promote lifelong learning and professional development, emphasizing the value of education for career growth at any age.

5. Focus on the Private Sector for Targeted Outreach

- **Insight:** Heatmaps reveal that the "Private" workclass dominates in terms of both marital status and income level. Additionally, the distribution of individuals in the Private sector shows significant representation in both income brackets.
- **Recommendation:** UVW College should specifically target individuals working in the private sector, especially those with higher-income potential (over $50K). Customized programs that align with the skills and qualifications needed in private sector jobs (such as business management, tech skills, etc.) should be highlighted in marketing efforts.

6. Focus on Racial Equity and Inclusion in Marketing Campaigns

- **Insight:** The heatmap and grouped bar chart show disparities in income distribution across racial categories, with White individuals being more represented in higher-income brackets, while Black individuals tend to be more concentrated in lower-income categories.
- **Recommendation:** UVW College should be proactive in addressing racial disparities by promoting inclusivity and providing opportunities for underrepresented groups. Targeted scholarships, mentorship programs, and recruitment strategies should be used to increase enrollment from diverse racial backgrounds. Highlighting success stories of individuals from underrepresented groups who have succeeded after obtaining an education from UVW College would be beneficial.

7. Tailor Programs to Work-Life Balance Needs

- **Insight:** The heatmap showing marital status and workclass, along with the income distribution across these categories, highlights the need for flexibility in education for individuals working in various sectors and balancing family responsibilities.
- **Recommendation:** UVW College should design and promote flexible educational programs for working adults, especially those in the private and government sectors. Programs that offer evening classes, weekend workshops, and online learning should be advertised as a way to achieve work-life balance while continuing education.

8. Use Data to Refine Targeted Advertising

- **Insight:** The data reveals significant patterns based on age, education, marital status, and work class that influence income levels.
- **Recommendation:** XYZ Corporation should use the demographic insights from this analysis to develop more granular and targeted advertising strategies. For example, digital ads can be personalized based on age and educational background, while social media campaigns can be adjusted to target individuals based on

educational background, while social media campaigns can be adjusted to target individuals based on marital status and work class. This approach will help UVW College reach the right audience more effectively.

9. Address Income Inequities in Marketing Strategies

- **Insight:** The income disparities across racial groups, especially the underrepresentation of Black individuals in the >$50K category, suggest that marketing strategies need to address socio-economic challenges and barriers.
- **Recommendation:** UVW College should create outreach campaigns that specifically address the financial barriers to higher education faced by underrepresented racial groups. Offering financial aid, emphasizing affordable tuition options, and providing access to resources for navigating the application process will help to reduce these barriers.

Conclusion:

By using these insights and recommendations, XYZ Corporation can help UVW College create targeted marketing strategies that not only increase enrollment but also promote inclusivity and diversity. Focusing on key demographics—such as older professionals, individuals with higher education, married individuals, and those in the private sector—while addressing income and racial disparities will help UVW College appeal to a broad range of prospective students and meet their educational goals.