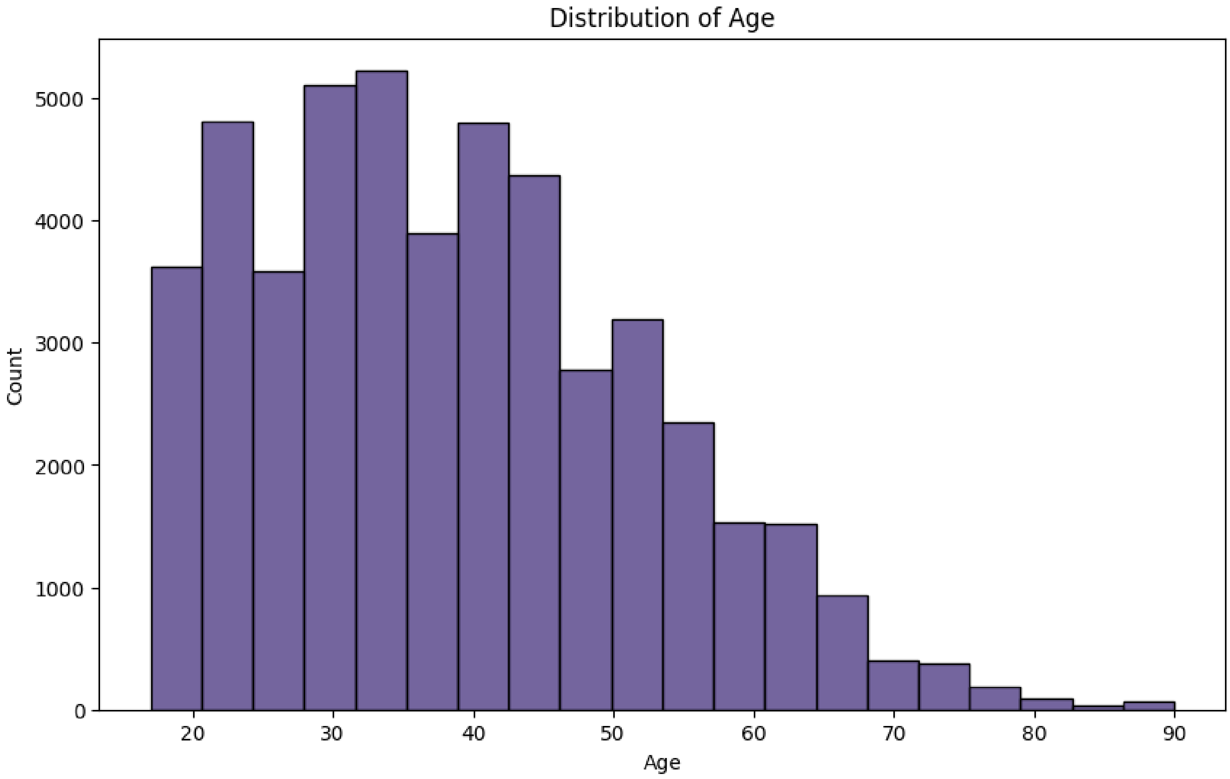
Please refer to the data set explanation above document

Variable Name: Age
Variable Type: Numerical
Scales of Measurement: Ratio Scale Data
The "Age" attribute represents the age of individuals in the dataset.It is a numerical variable because it represents a measurable quantity that can take on a range of real values.
The "Age" variable has a ratio scale of measurement, as it has a true zero point (i.e., an age of 0 represents the absence of age), and the ratios between values are meaningful (e.g., someone who is 30 years old is twice as old as someone who is 15 years old).

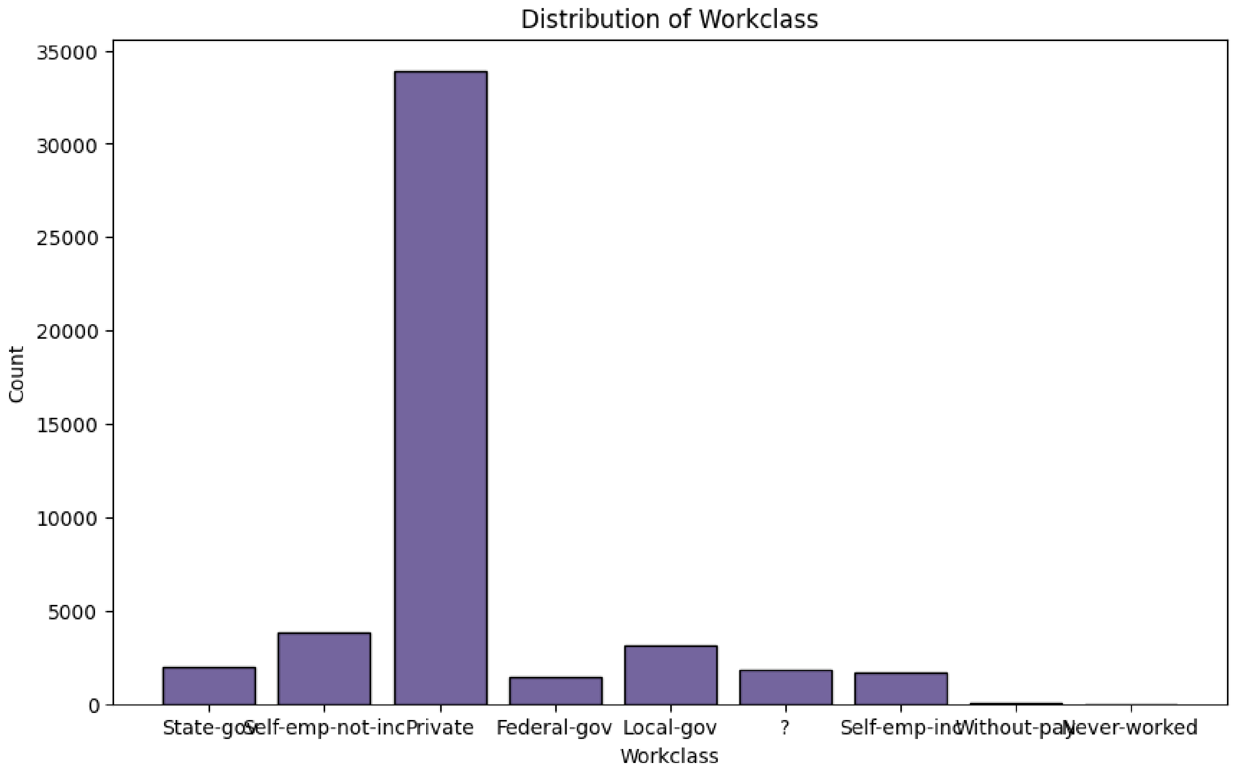
Type of Plot: Histogram

A histogram is an effective way to visualize the distribution of numerical data, such as age.
It provides insights into the frequency or count of individuals in different age ranges.
Helps in identifying patterns in the age distribution, such as the central tendency and spread of ages.Useful for understanding the density of observations at different points in the age range**.**

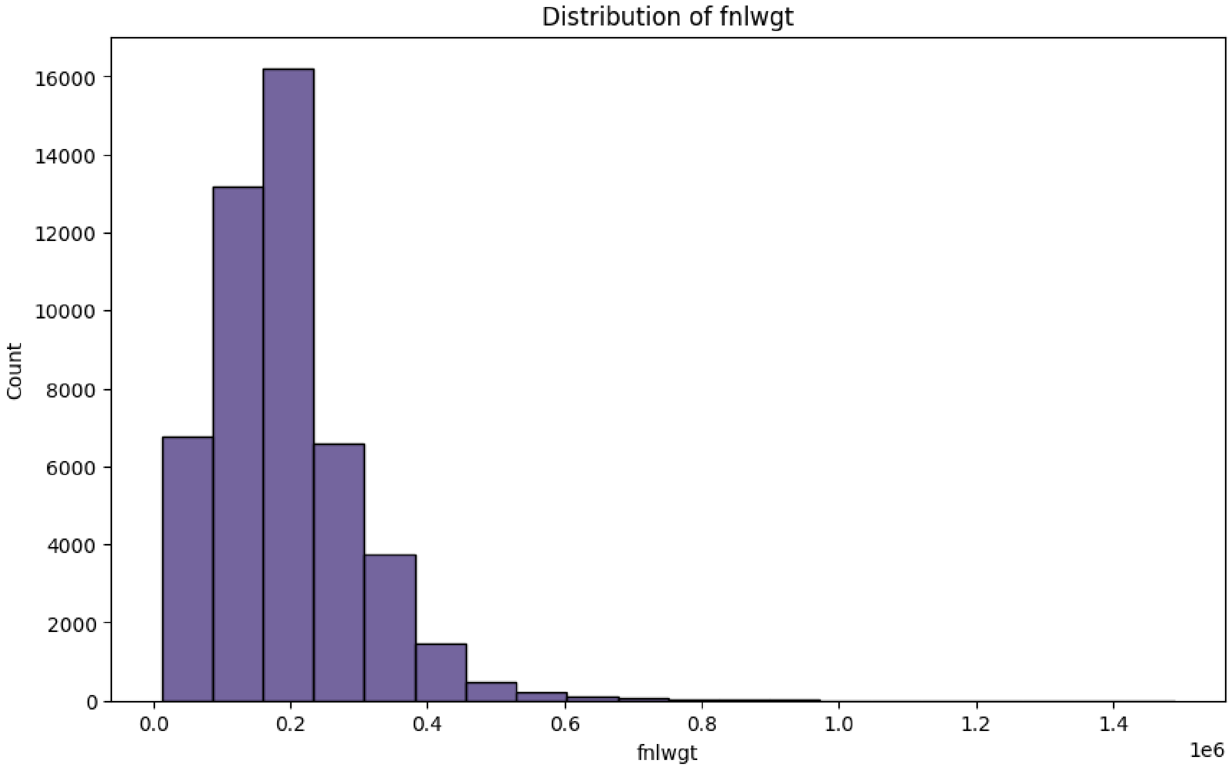


Variable Name: Workclass
Variable Type: Categorical
Scales of Measurement: Nominal Scale Data
The "Workclass" attribute represents the type of employment or workclass of individuals in the dataset.It is a categorical variable because it falls into distinct categories without a specific order

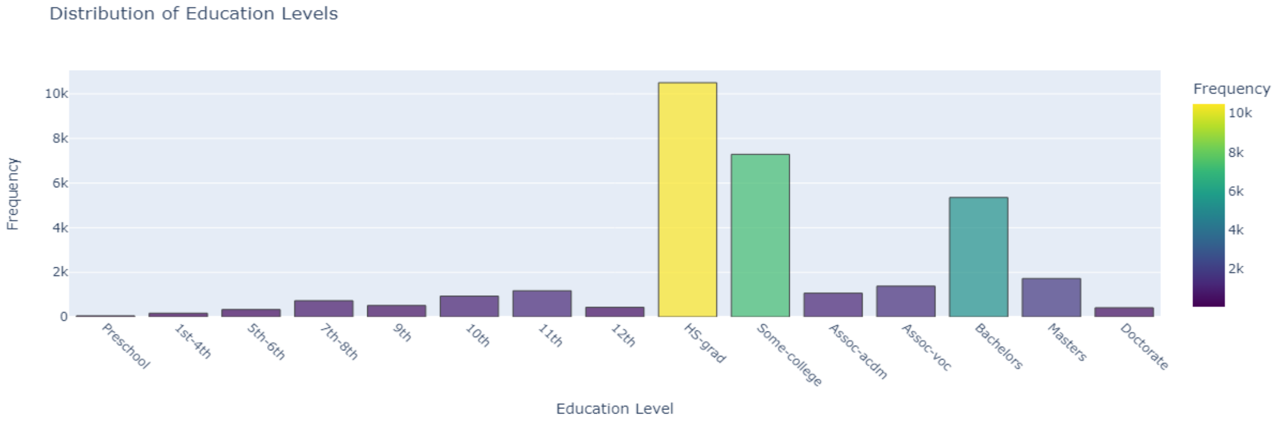
or hierarchy.The "Workclass" variable has a nominal scale of measurement, as the categories (e.g., Private, Self-emp-not-inc, Federal-gov, etc.) are labels without inherent order.
Type of Plot: Histogram A Histogram is suitable for visualizing the distribution of categorical data, such as workclass categories.It displays the count or frequency of observations in each category.
Effective for comparing the number of individuals in different workclass categories.



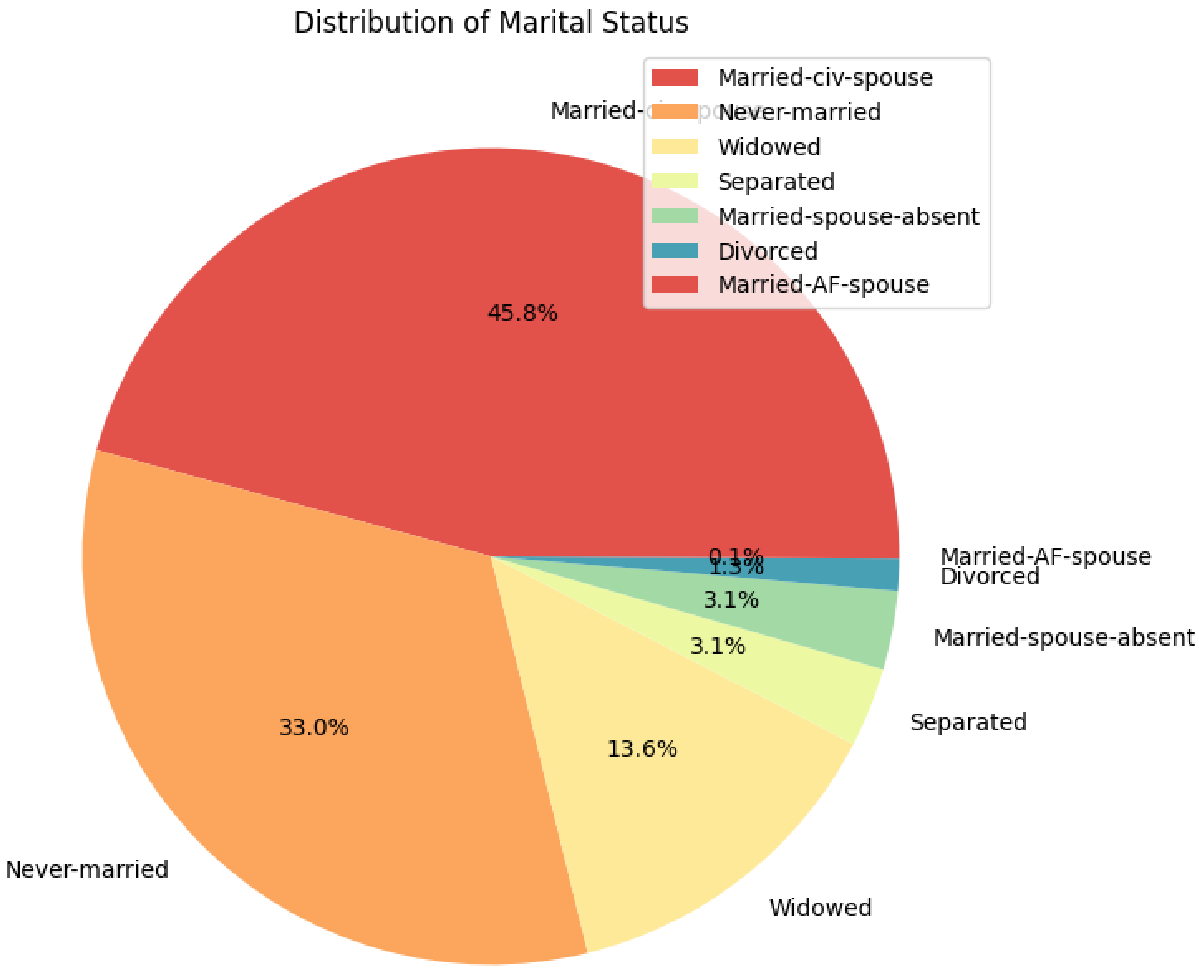
Variable Name: fnlwgt
Variable Type: Numerical
Scales of Measurement: Ratio Scale Data
The "fnlwgt" attribute represents the final weight assigned to an entry, indicating the number of people the census believes that entry represents.
It is a numerical variable because it represents a measurable quantity that can take on a range of real values. The "fnlwgt" variable has a ratio scale of measurement, as it has a true zero point (i.e., a value of 0 represents the absence of the variable), and ratios between values are meaningful.
Type of Plot: Histogram
A histogram is an effective way to visualize the distribution of numerical data, such as "fnlwgt." It provides insights into the frequency or count of observations within different ranges of "fnlwgt." Helps in identifying patterns in the distribution, such as the central tendency and spread of values.



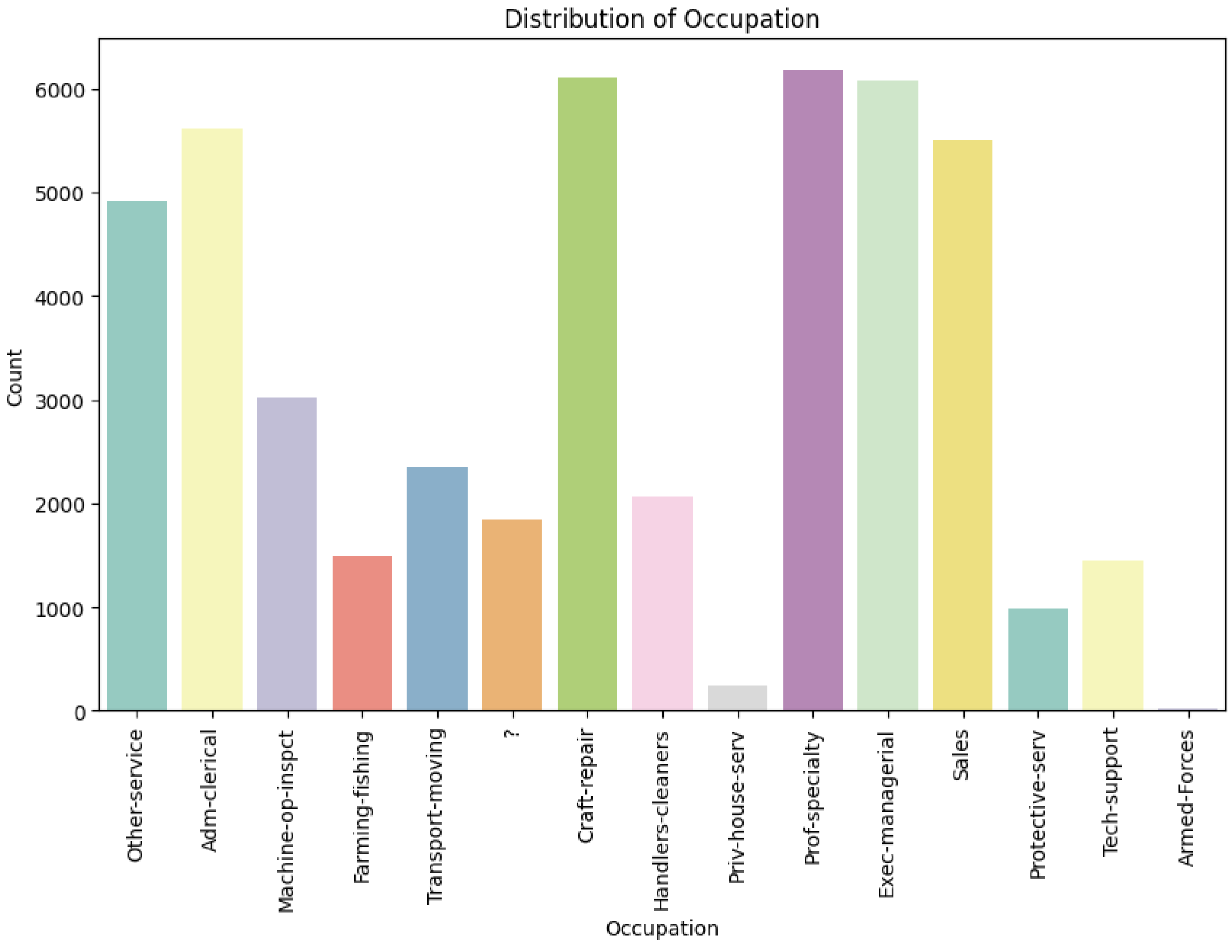
Variable Name: Education
Variable Type: Categorical
Scales of Measurement: Ordinal Scale Data
The "Education" attribute represents the highest level of education attained by individuals in the dataset.It is a categorical variable because it falls into distinct categories without a specific order or hierarchy.The "Education" variable has an ordinal scale of measurement, as the education levels have a clear order or hierarchy. For example, "Preschool" is considered lower than "1st-4th," and "Masters" is higher than "Bachelors."
Type of Plot: Bar chart
An ordered bar chart is a suitable way to visualize ordinal data, where the order or hierarchy among categories is important.
It emphasizes the ordinal relationship between education levels and allows for a clear comparison of frequencies.
Each bar is ordered based on the hierarchy of education levels, providing an intuitive representation of the distribution

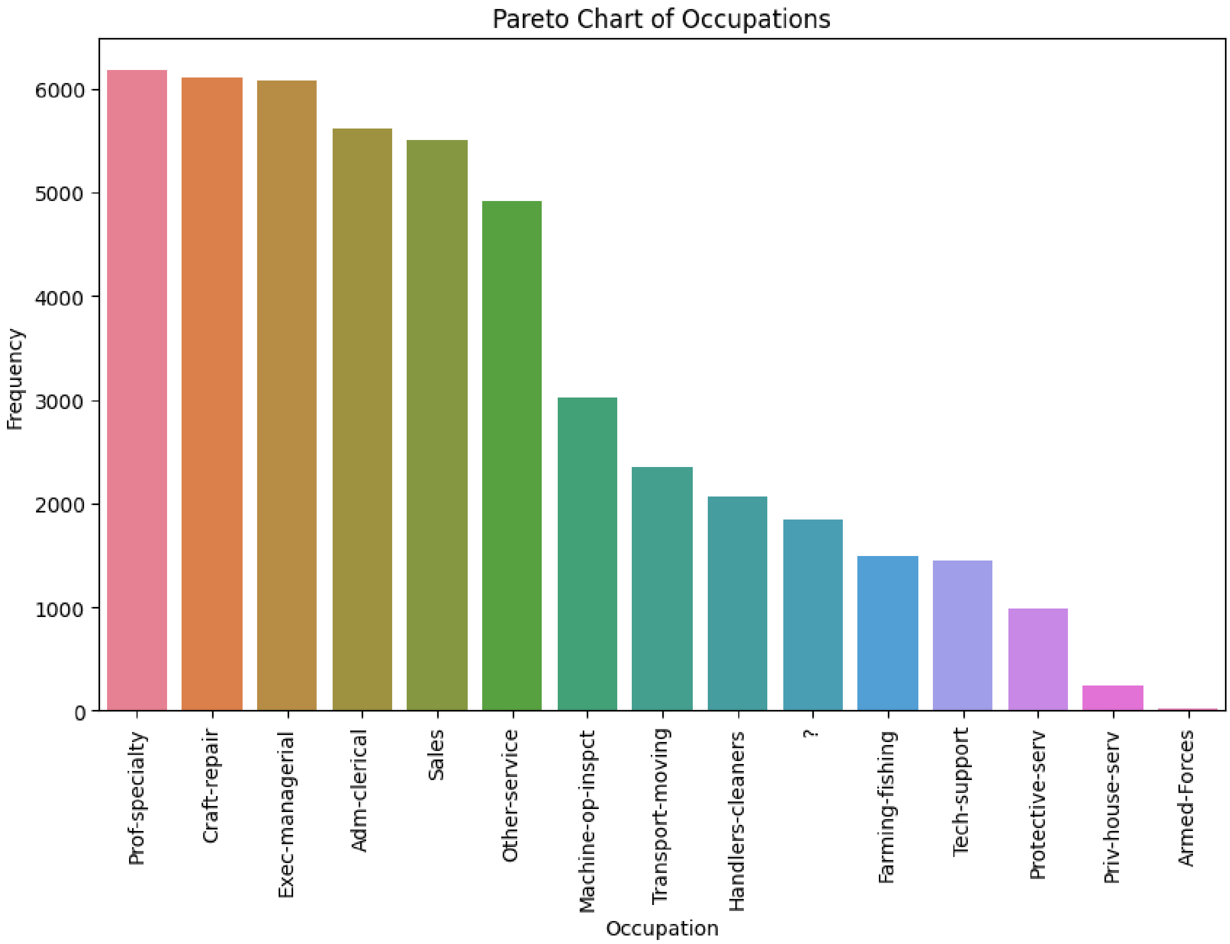


Variable Name: Marital-status
Variable Type: Categorical
Scales of Measurement: Nominal Scale Data
The "Marital-status" attribute represents the marital status of individuals in the dataset.
It is a categorical variable because it falls into distinct categories without a specific order or hierarchy.The "Marital-status" variable has a nominal scale of measurement, as the categories (e.g., Married, Divorced, Never-married) are labels without inherent order.
Type of Plot: Pie Chart
A pie chart is suitable for visualizing the distribution of categorical data when you have a small number of categories.It provides a clear representation of the proportion of each marital status category in the overall dataset.

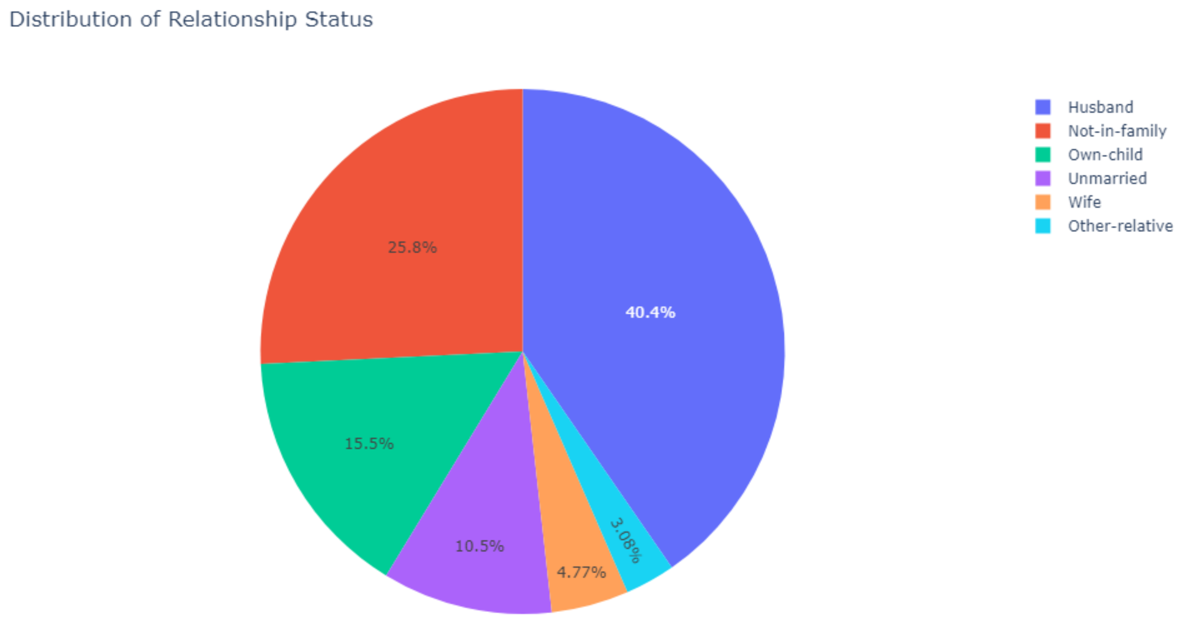


Variable Name: Occupation
Variable Type: Categorical
Scales of Measurement: Nominal Scale Data
The "Occupation" attribute represents the occupation of individuals in the dataset.
It is a categorical variable because it falls into distinct categories without a specific order or hierarchy.The "Occupation" variable has a nominal scale of measurement, as the categories (e.g., Tech-support, Craft-repair, Sales) are labels without inherent order.
Type of Plot: Bar Chart and Pareto chart
A bar chart is suitable for visualizing the distribution of nominal data, where the order or hierarchy among categories is not important.It provides a clear comparison of frequencies for each occupation category.Each bar represents a different occupation category, and the height of the bar corresponds to the count or frequency of that category.
A Pareto chart is typically used to prioritize and visualize the most significant factors based on their cumulative impact. For a Pareto chart, you would typically need a variable that represents different categories or factors and another variable that measures the frequency, count, or impact of each category.


Distribution of Occupation
Count
0
1000
2000
3000
4000
5000
6000
Occupation
Other-service
Adm-clerical
Machine-op-inspct
Farming-fishing
Transport-moving
?
Craft-repair
Handlers-cleaners
Priv-house-serv
Prof-specialty
Exec-managerial
Sales
Protective-serv
Tech-support
Armed-Forces

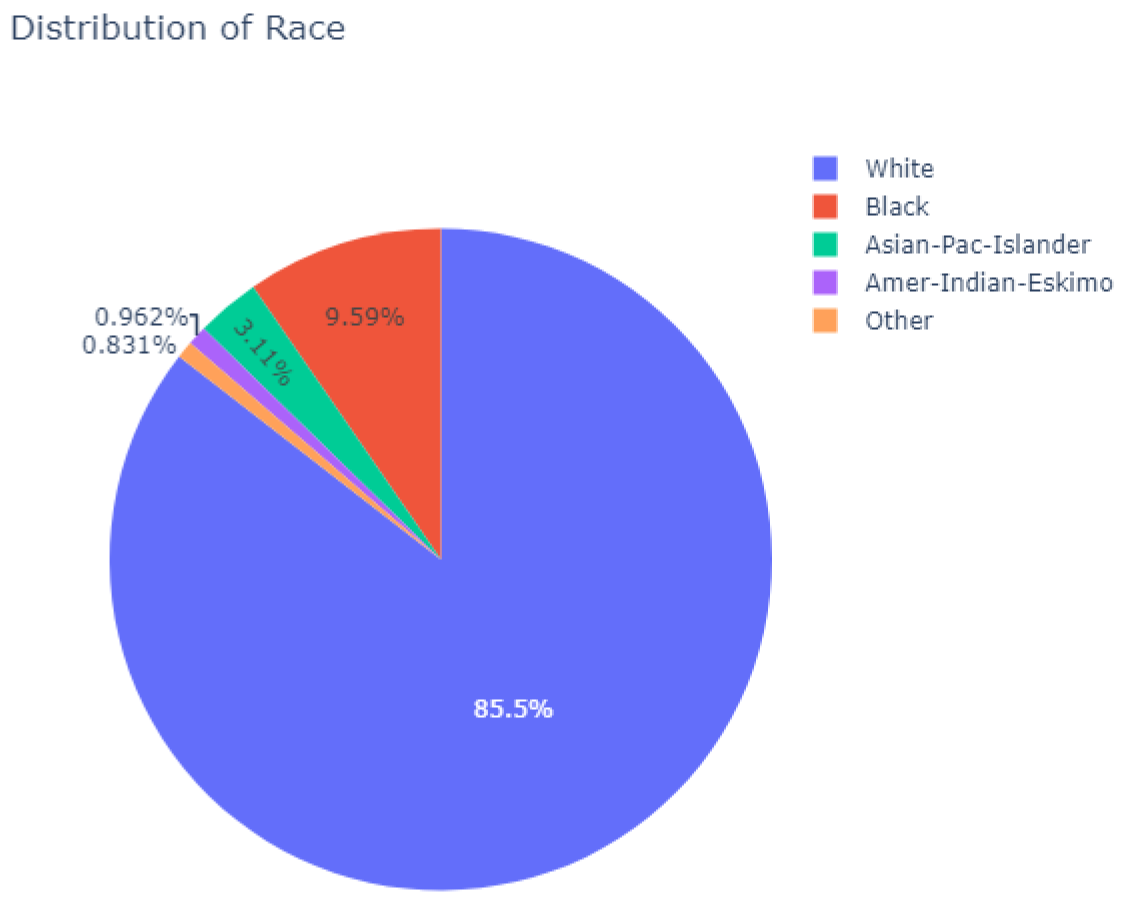


Variable Name: Relationship
Variable Type: Categorical
Scales of Measurement: Nominal Scale Data
The "Relationship" attribute represents the type of relationship an individual has in the dataset.
It is a categorical variable because it falls into distinct categories without a specific order or hierarchy.The "Relationship" variable has a nominal scale of measurement, as the categories (e.g., Wife, Own-child, Husband) are labels without inherent order.
Type of Plot: Pie Chart
A pie chart is suitable for visualizing the distribution of categorical data with a small number of categories.It provides a clear representation of the proportion of each relationship category in the overall dataset.

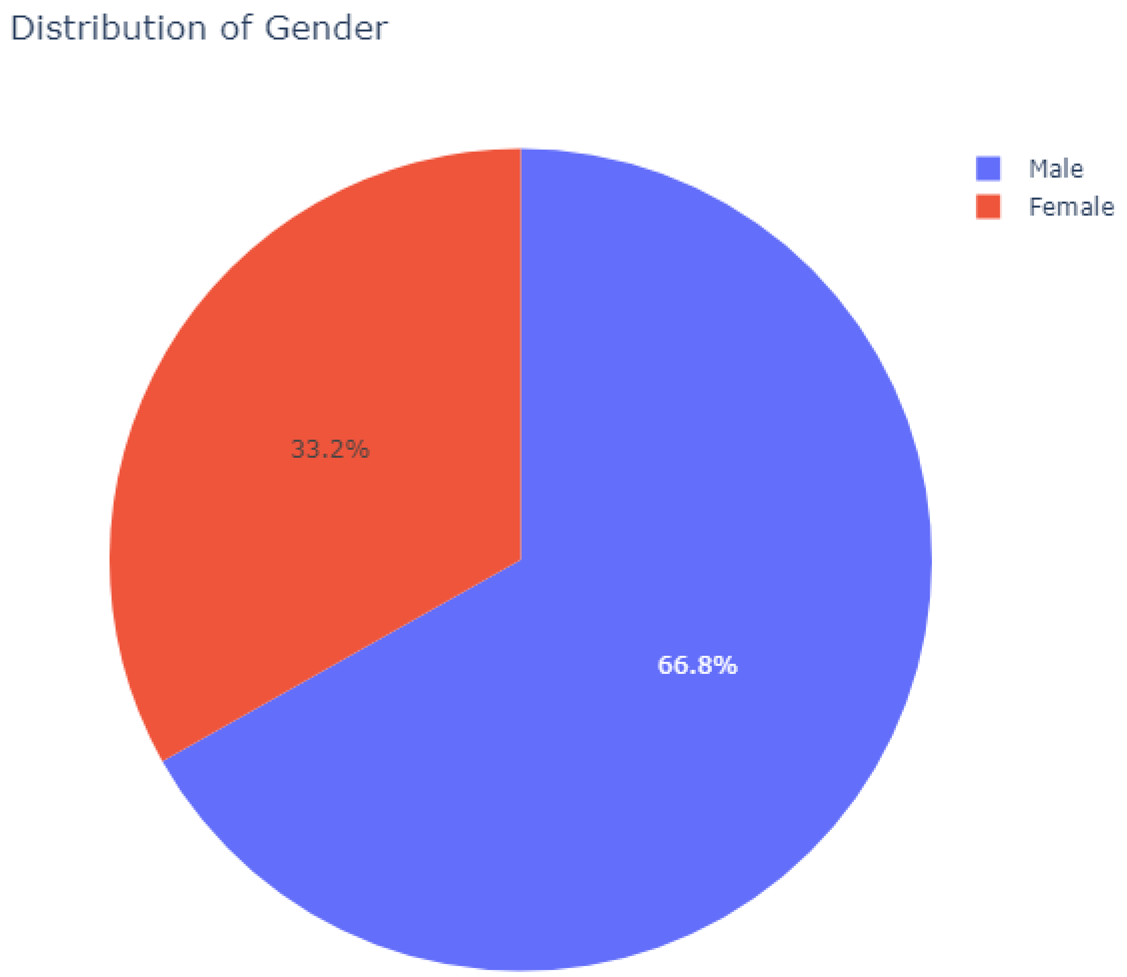


Variable Name: Race
Variable Type: Categorical
Scales of Measurement: Nominal Scale Data

Description: The "Race" attribute in the dataset represents the racial background of individuals. It falls under the category of nominal scale data because the different races do not have a meaningful order or hierarchy. Each race category is distinct, and there is no inherent ranking among them.
For example, the categories within the "Race" variable include White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, and Black.
Type of plot : a bar chart or a pie chart. These types of plots effectively display the distribution of categories without implying any order or magnitude.



Variable Name: Sex
Variable Type: Categorical
Scales of Measurement: Nominal Scale Data
The "Sex" attribute represents the gender of individuals in the dataset.
It is a categorical variable because it falls into distinct categories without a specific order or hierarchy.The "Sex" variable has a nominal scale of measurement, as the categories (e.g., Female, Male) are labels without inherent order.
Type of Plot: Pie Chart:
A pie chart is suitable for visualizing the distribution of categorical data with a small number of categories.It provides a clear representation of the proportion of each gender category in the overall dataset.

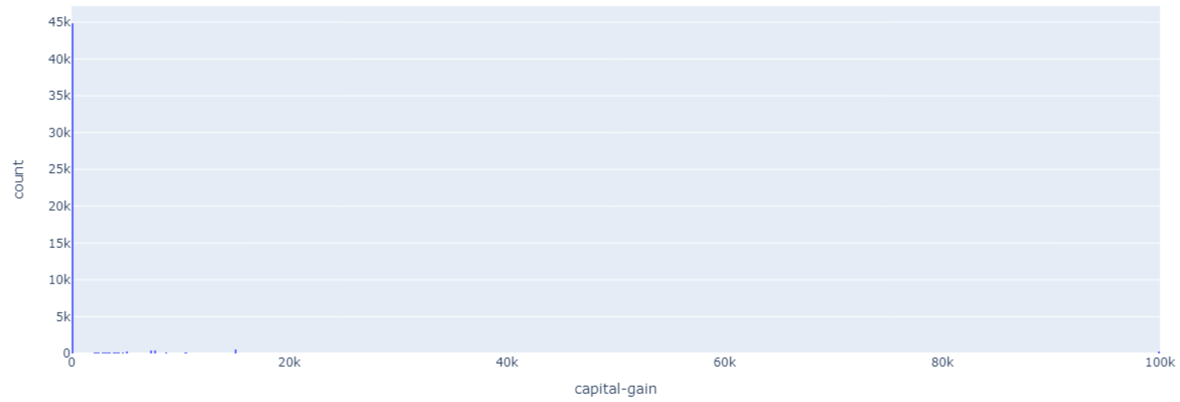


Variable Name: Capital Gain
Variable Type: Numerical
Scales of Measurement: Ratio Scale Data
The "Capital Gain" attribute represents the financial gain an individual experiences from the sale of an investment or property. It is a numerical variable because it can take on any real value within a range, and it has a true zero point, making it a ratio scale variable. The presence of a true zero means that a value of zero indicates the absence of the variable, and ratios between values are meaningful.

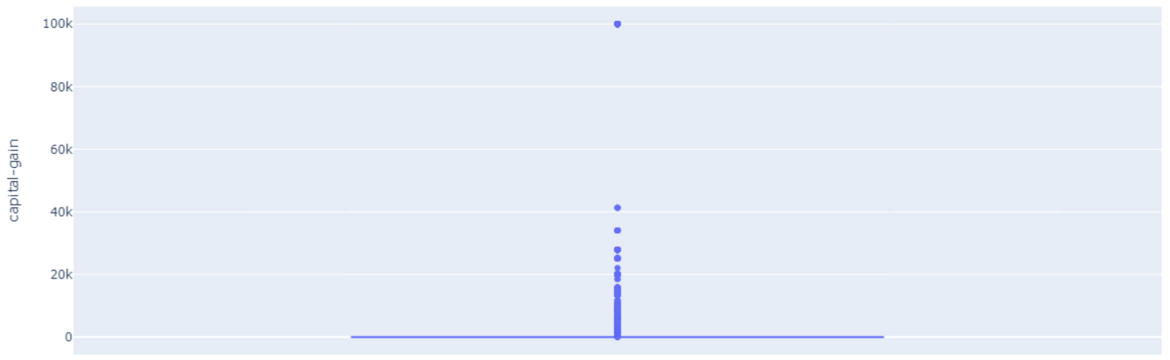
Type of plot : histogram or a box plot. These types of plots effectively display the distribution and central tendency of the numerical data.

AS THE DATA HAS MANY OUTLIER THE FIGURE IS NOT REPRESENTATIVE

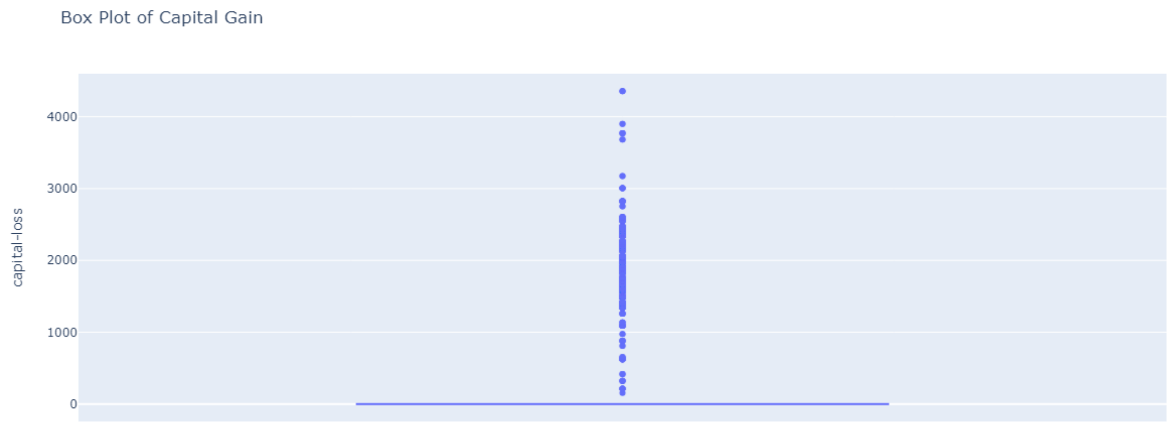
Distribution of Capital Gain



Box Plot of Capital Gain



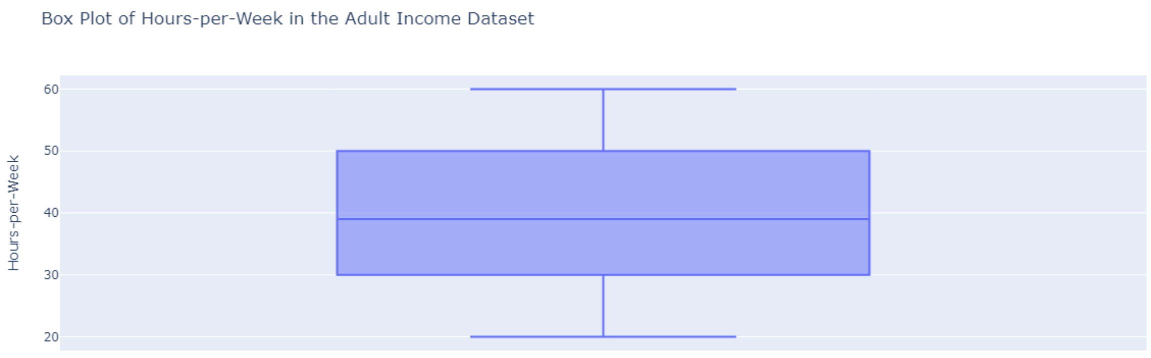
Variable Name: Capital Loss
Variable Type: Numerical
Scales of Measurement: Ratio Scale Data
The "Capital Loss" attribute represents the financial loss an individual incurs. It is a numerical variable because it represents a measurable quantity. The "Capital Loss" variable has a ratio scale of measurement, as it has a true zero point (0 represents no loss).
Type of Plot: Box Plot
A box plot is suitable for visualizing the distribution, central tendency, and potential outliers of numerical data.It provides insights into the spread and statistical summary of the "Capital Loss" variable.



Variable Name: Hours-per-Week
Variable Type: Numerical
Scales of Measurement: Interval Scale Data

Description: The "Hours-per-Week" attribute in the dataset represents the number of hours an individual works per week. It is a numerical variable with an interval scale, where the intervals between consecutive values are consistent and meaningful. However, it lacks a true zero point, and the value of zero does not imply the absence of working.

Type of plot : a histogram or a box plot. These types of plots effectively display the distribution and central tendency of the numerical data.

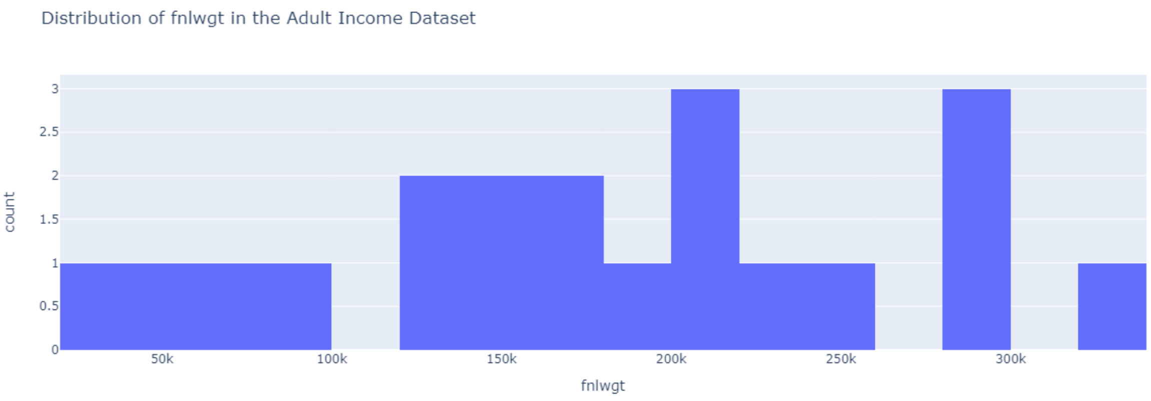


Variable Name: fnlwgt
Variable Type: Numerical
Scales of Measurement: Ratio Scale Data

Description: The "fnlwgt" attribute in the dataset represents the final weight assigned to an entry, indicating the number of people the census believes that entry represents. It is a numerical variable with a ratio scale, as it has a true zero point, and ratios between values are meaningful.

Type of plot : histogram is a suitable choice for visualizing the distribution of a numerical variable, such as "fnlwgt" in the "Adult Income" dataset, for several reasons:

- Distribution Overview
- Range Exploration.
- Visualizing Density.
- Easy Comparison.

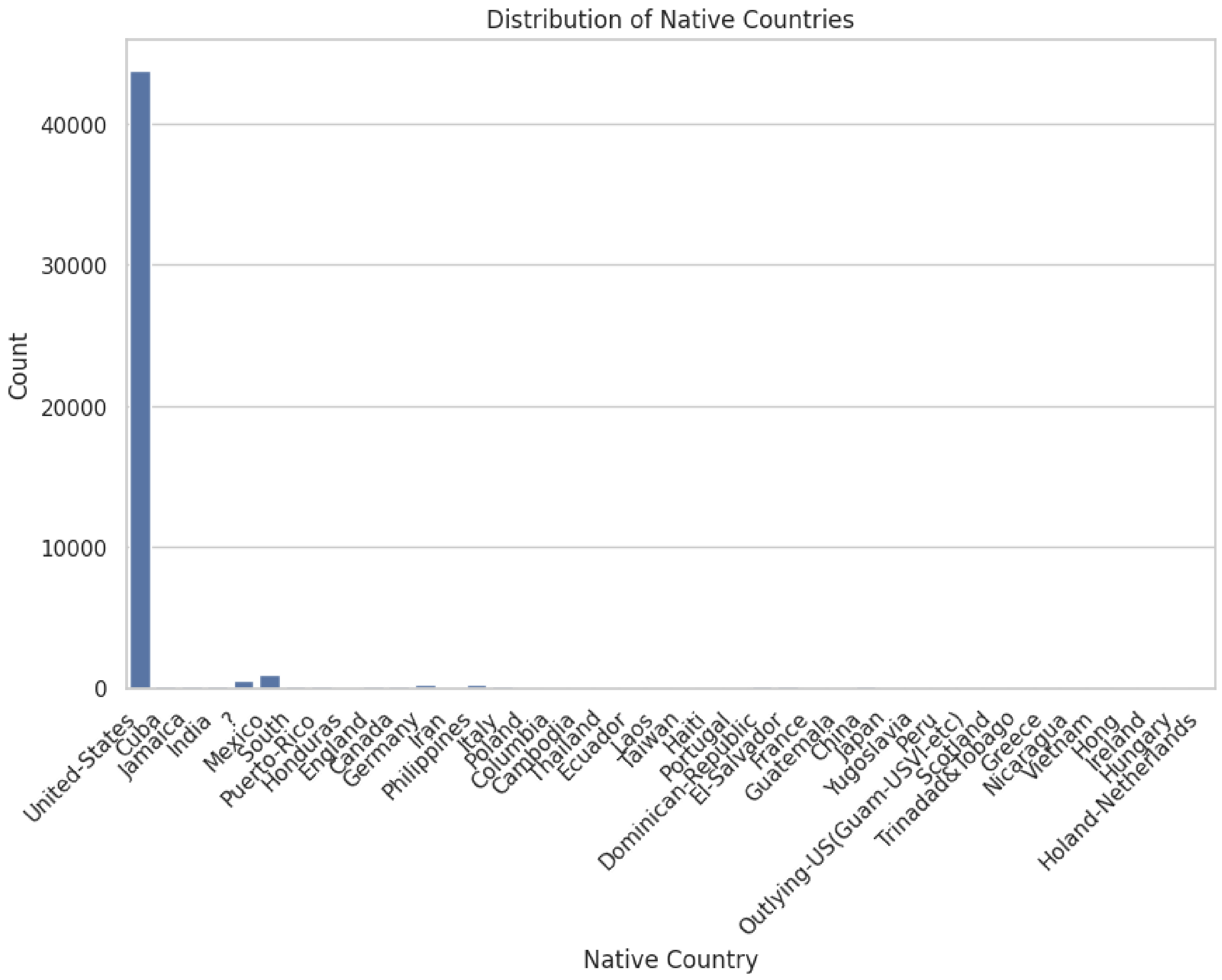


Variable Name: Native Country
Variable Type: Categorical
Scales of Measurement: Nominal Scale Data

The "Native Country" attribute represents the country of origin or citizenship of individuals in the dataset. It is a categorical variable because it falls into distinct categories without a specific order or hierarchy. The "Native Country" variable has a nominal scale of measurement, as the categories (e.g., United-States, India, China) are labels without inherent order.

Type of Plot: Bar Chart

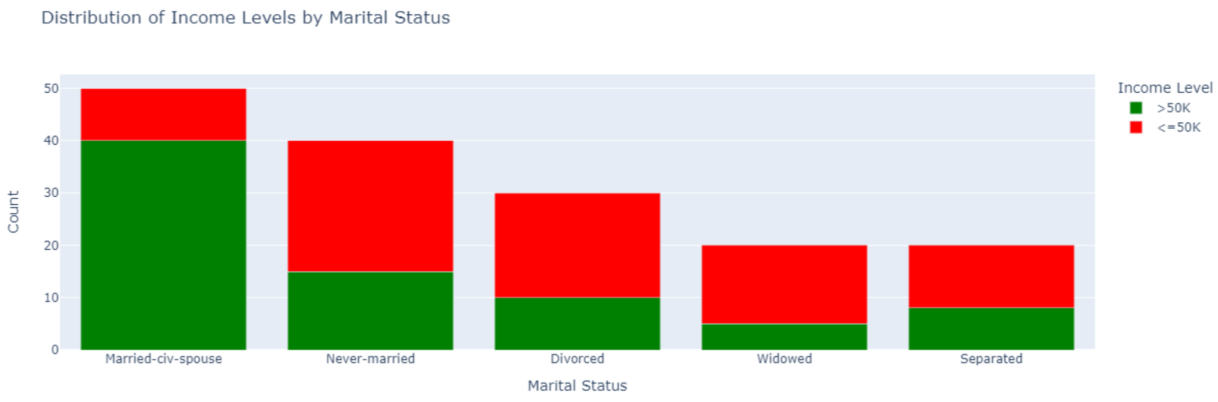
A bar chart is suitable for visualizing the distribution of categorical data. It provides a clear comparison of the frequency of each native country category in the overall dataset.



Variable Name: Marital Status
Variable Type: Categorical
Scales of Measurement: Nominal Scale Data

Description: The "Marital Status" attribute in the dataset represents the marital status of individuals. It is a categorical variable with a nominal scale, indicating distinct categories without an inherent order.
Type of Chart : stacked bar chart is a suitable choice when you want to visualize the composition of a total variable (in this case, the total count of individuals) across different categories (marital statuses) and show the contribution of subcategories (income levels) within each category. Here are some reasons why a stacked bar chart is appropriate:

- Comparison of Parts to Whole: Stacked bar charts allow you to easily compare the distribution of income levels (>50K and <=50K) within each marital status category. The length of each bar represents the total count of individuals, and the different colors represent the proportions contributed by each income level.
-
- Total and Subcategory Relationships: Stacked bar charts emphasize the relationship between the total and its subcategories. Viewers can quickly discern how the total (marital status) is divided into its components (income levels) for each category.
-
- Visualization of Patterns: Stacked bars help visualize patterns, such as which marital status categories have a higher concentration of individuals with income >50K or <=50K. It provides insights into the relationships between two categorical variables.
-
- Easy Interpretation: Stacked bar charts are intuitive and easy to interpret, making them accessible for a wide audience. The colors make it clear which portion of each bar corresponds to each income level.


Distribution of Income Levels by Marital Status
Count
Married-civ-spouse
Never-married
Divorced
Widowed
Separated
Marital Status
Income Level
🟩 >50K
🟥 <=50K