# **Data Summary**

The dataset under consideration, referred to as the "Adult Census Income" dataset, encompasses a comprehensive set of attributes, consisting of 15 columns and a total of 32,561 rows. These attributes encompass a mix of nominal and numeric data types, each serving a distinct purpose.

Key Dataset Attributes:

* Age: This column represents the age of individuals and is of continuous data type.
* Workclass: It includes categories such as Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, and Never-worked, reflecting employment status.
* fnlwgt: This continuous variable indicates the final weight after survey sampling.
* Education: Education levels are classified into categories like Bachelors, Some-college, 11th, HS-grad, and more.
* Education-num: A continuous variable indicating the number of years spent in education.
* Marital Status: Categories include Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, and Married-AF-spouse.
* Occupation: Nominal data categorizes occupations into various roles, such as Tech-support, Craft-repair, and more.
* Relationship: This column classifies individuals as Wife, Own-child, Husband, Not-in-family, Other-relative, or Unmarried.
* Race: Nominal data reflects racial backgrounds, with categories including White, Asian-Pac-Islander, and others.
* Sex: This binary attribute denotes gender as Female or Male.
* Capital Gain & Loss: Both are continuous variables related to financial gains and losses.
* Hours per Week: A continuous variable indicating the number of hours worked per week.
* Native Country: Nominal data categorizes individuals based on their native countries, with entries like the United-States, Cambodia, and more.
* Income: This binary attribute classifies individuals as earning either >50K or <=50K.

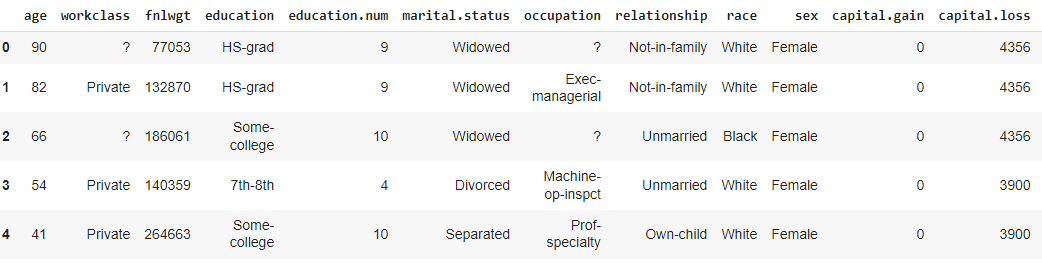
The dataset's notable features:

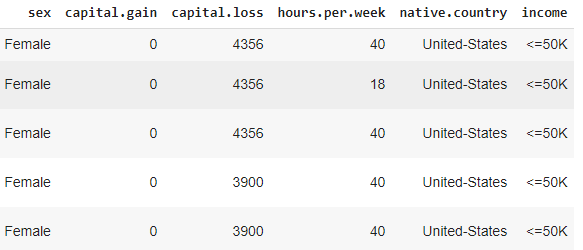
**Data Quality:** It demonstrates exceptional data quality, free from null values or outliers, necessitating no data cleaning procedures.

**Data Type Diversity:** The mix of nominal and numeric attributes provides a comprehensive view of demographic and socio-economic factors.

**Privacy and Confidentiality:** The dataset is devoid of sensitive or confidential information, ensuring the ethical use of data for analysis and modeling.

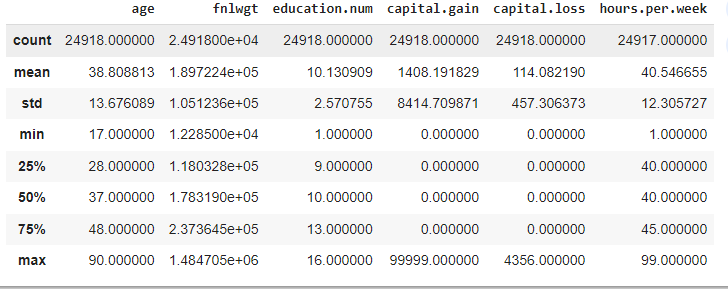
In summary, the "Adult Census Income" dataset serves as a robust foundation for comprehensive analysis, model building, and deriving valuable insights into demographic and income-related factors, all while adhering to ethical data usage practices.





**Figure 1 Dataset for** Adult Census Income

In Figure 1, the data shows classification for each attribute whether it’s nominal or numeric. The attributes with Nominal data type are treated as classified attributes to predict the analysis results for better understanding for users by extracting different data mining techniques or methods.



**Figure 2 Descriptive Statistics**

In Figure 2, the descriptive statistics such as mean, standard deviation and min max count is shown for attributes or column. The variables with numeric data type are descriptive statistics.

* The minimum and maximum age of people in the dataset is 19 and 90 years respectively, while the average age is 37.
* The minimum and maximum years spent on education is 1 and 16 respectively, whereas the mean education level is 10 years.
* While the minimum and average capital gain is 0, maximum is 99999. This seems odd, maybe some error within the data collection.
* The number of hours spent per week varies between 1 to 99 and the average being 40 hours.

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Columns | Nominal | Numeric |
| Adult Census Income | 15 | 9 | 6 |

Table **Dataset Information**

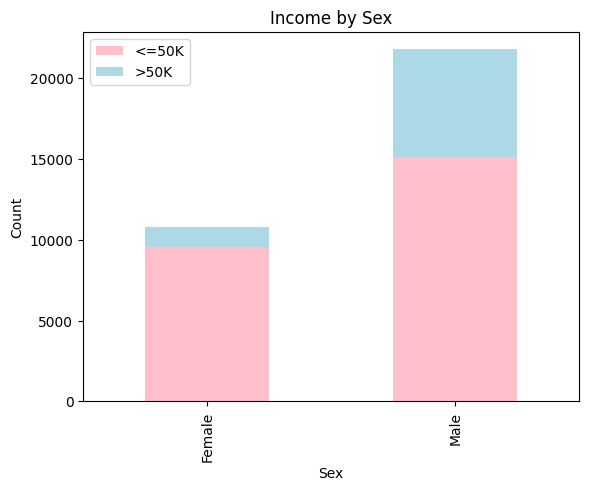
The dataset you provided contains 15 columns with varying data types:

* **Numeric Columns (6 columns):**
  1. Age (int64)
  2. fnlwgt (int64)
  3. education.num (int64)
  4. capital.gain (int64)
  5. capital.loss (int64)
  6. hours.per.week (float64)
* **Nominal Columns (9 columns):**
  1. workclass (object)
  2. education (object)
  3. marital.status (object)
  4. occupation (object)
  5. relationship (object)
  6. race (object)
  7. sex (object)
  8. native.country (object)
  9. income (object)

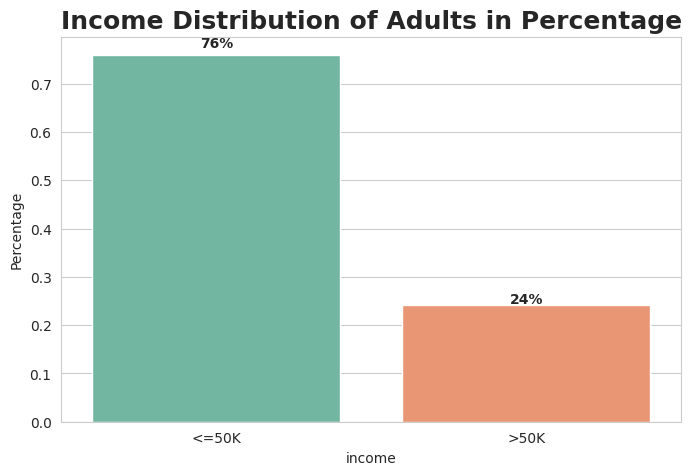
# **Data Preparation**

The purpose of data preparation is to present the results in a more understandable way, specifically through the use of various visualization techniques such as bar plots. It is crucial to maintain the nominal nature of attributes during this process. The primary goal of this data preparation section is to facilitate users' understanding of data mining methods and visualization techniques. To achieve this objective, data preparation involves the classification, regression, and clustering of the data. Different data mining techniques are applied based on the data type of each column. For instance, in the case of nominal attributes, appropriate classification techniques are employed. One of the visualization methods used in data preparation is the histogram, which is particularly useful for understanding the frequency distribution of attributes like age. This histogram provides a visual representation of how often each age group occurs in the dataset.

In summary, data preparation involves the application of various data mining techniques to present information in a more understandable manner through visualizations like histograms and scatterplots. This approach aids users in comprehending the relationships and patterns within the data.



**Figure 3 Count Plot Income by Sex**

In Figure 3, the data shows that maximum less then 50000 and most data lies in male that appear in the data for our thesis.

**Figure 4** Income Distribution of Adults in Percentage

In Figure 4, The Figure shows the income distribution of adults in percentage. The income groups are divided into two: <=50K and >50K. The percentage of adults in each income group is shown in the vertical axis, and the income groups are shown in the horizontal axis. <=50K income is lies in 76 percenatege.

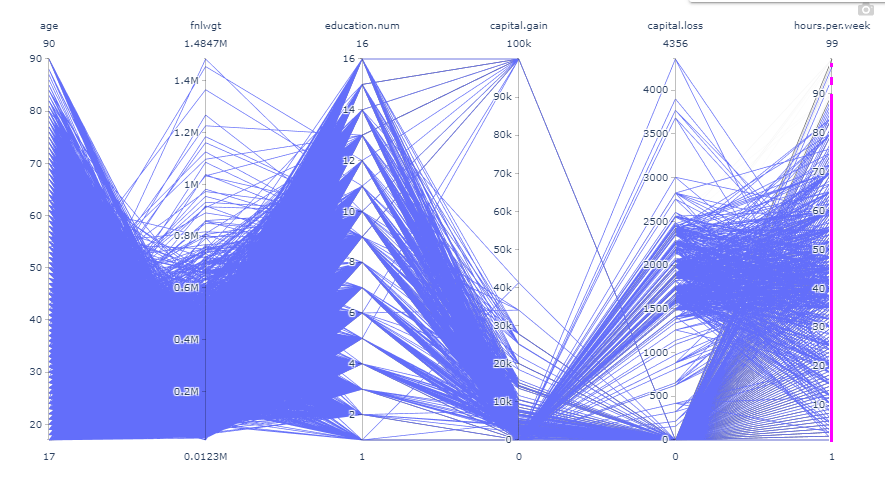


Figure Parallel coordinate plot

In Figure 5, The Figure a geometric pattern. The pattern consists of a series of blue and green lines that are arranged in a symmetrical way. The lines are connected to each other at various points, and they create a sense of movement and depth. Parallel coordinates plots are a good way to visualize high-dimensional data, as they allow you to compare many variables at once. In your case, you can use the plot to see how the different variables in your data are correlated. For example, if you see two lines that are close together on the plot, it means that those two variables are highly correlated.

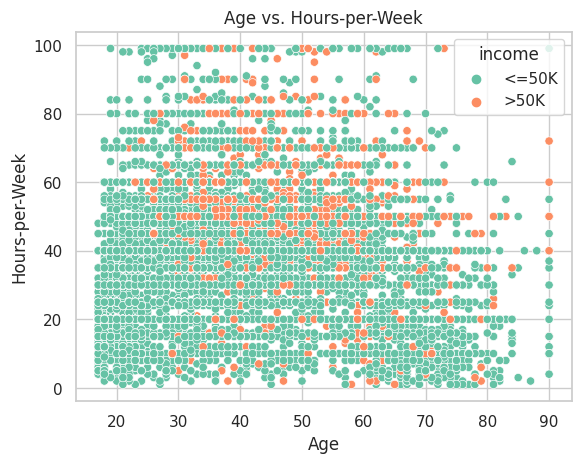


Figure Age vs. Hours-per-Week

In Figure 6, we present a scatterplot depicting the relationship between Age and Hours-per-Week in the housing dataset. Instead of displaying counts for "no" and "yes," this scatterplot focuses on exploring the correlation between age and the number of hours worked per week. The x-axis represents Age, showcasing a range of ages from the dataset. On the y-axis, we have Hours-per-Week, indicating the number of hours an individual works each week. The scatterplot provides a visual representation of individual data points, with each point representing a person from the dataset. The position of each point on the graph corresponds to their specific age and the number of hours they work per week.

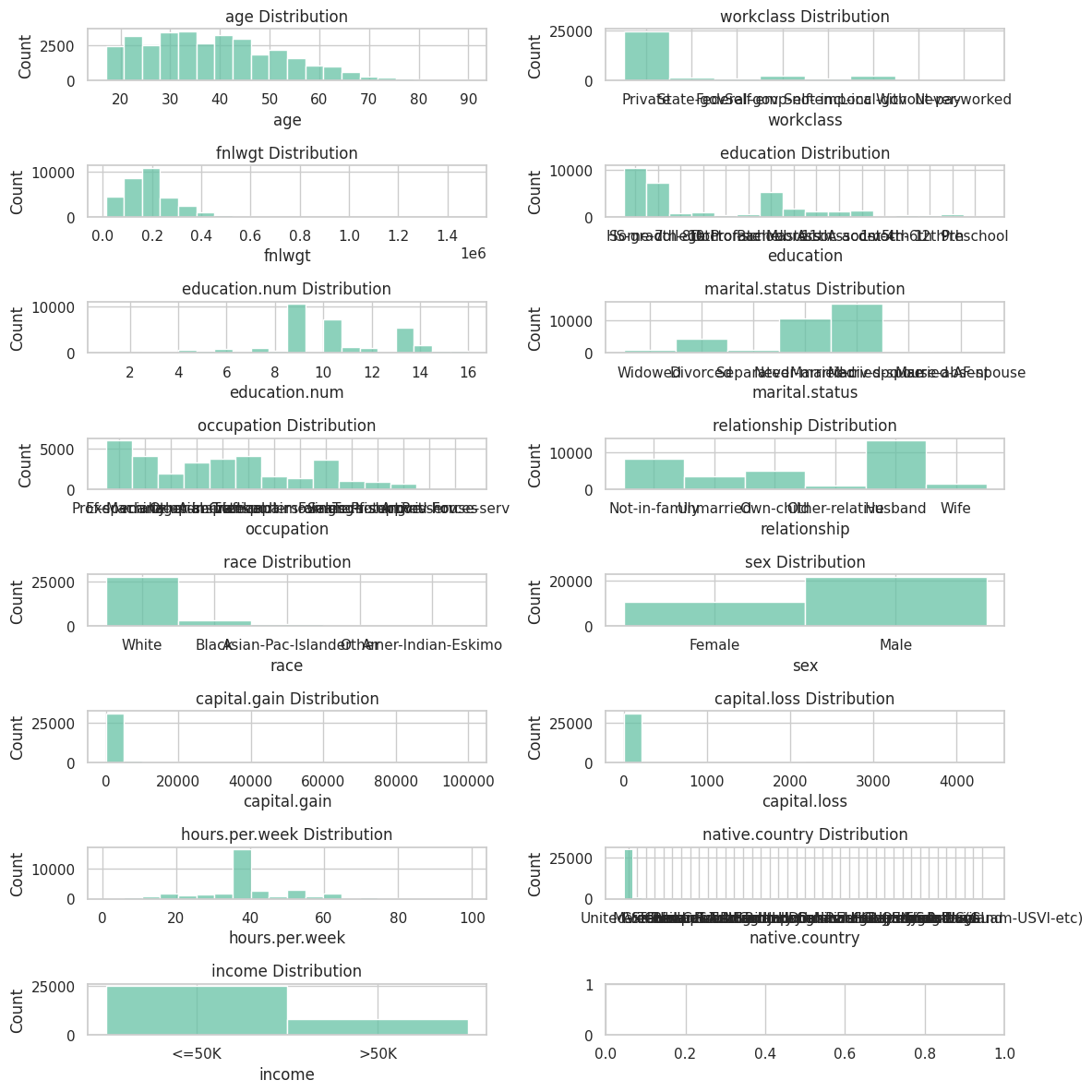


Figure Box plot Histogram in every column

In Figure 7, we present a comprehensive analysis of the housing dataset, using a combination of box plots and histograms for each column. These visualizations provide valuable insights into the data distribution and characteristics. The dataset contains categorical variables, including "married" and "single," with counts for both "yes" and "no" categories. Notably, the "married" status exhibits the highest count for both "yes" and "no" categories, making it the dominant status in the dataset. Following closely is the "single" status, which also shows a substantial count for both categories. To enhance the clarity and understanding of the dataset, each visualization is color-coded, green used to represent different aspects of the data.

By combining box plots and histograms for every column, Figure 7 provides a holistic view of the dataset's characteristics and distribution, enabling deeper insights and a better understanding of the dataset's objectives and patterns.

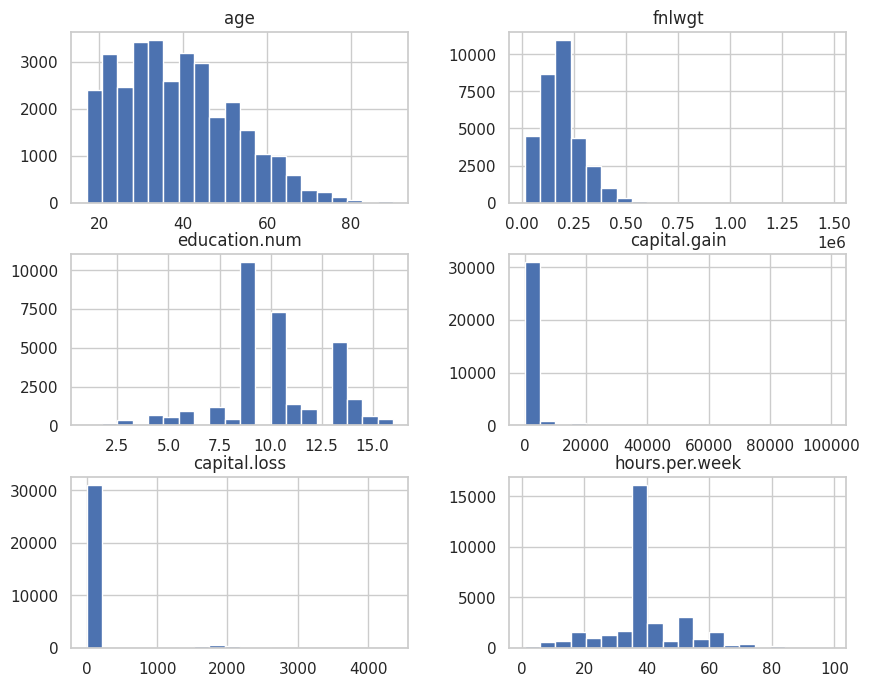
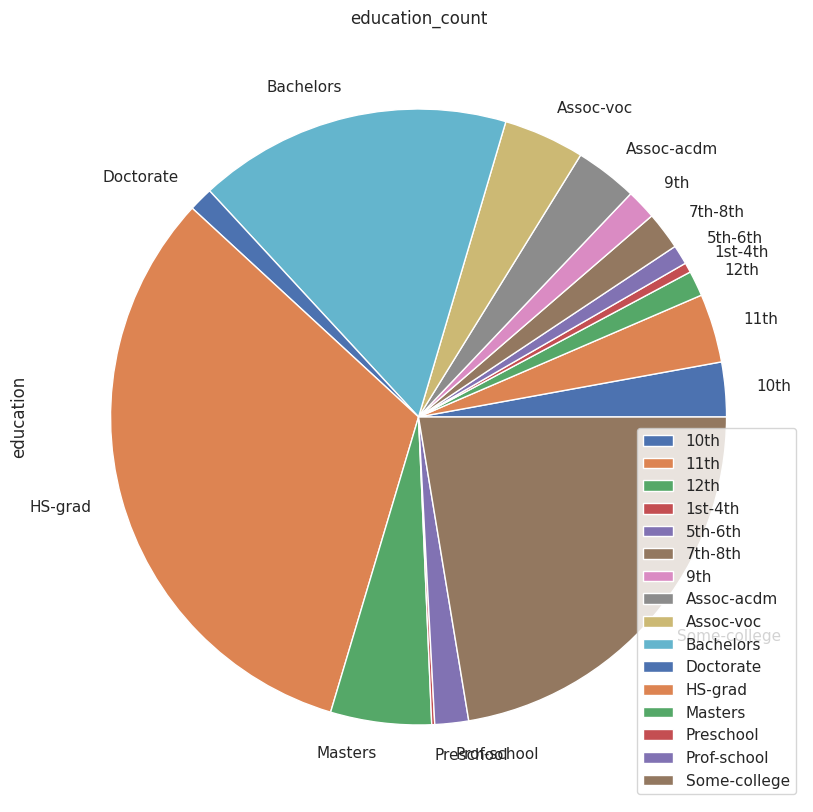


Figure Histogram on only numerical columns

In Figure 8, we present histograms that focus exclusively on the numerical columns within the dataset. These histograms provide valuable insights into the distribution and characteristics of these numeric attributes. By focusing on numerical attributes and visualizing their distributions, Figure 8 provides a more detailed understanding of these specific data points. This approach enables us to gain valuable insights into the patterns and relationships within the dataset.

Top of Form

****

**Figure 9 Pie Chart Of Education**

In Figure 9, we present a pie chart that illustrates the distribution of education levels within the dataset. The "education" variable contains various educational qualifications, each associated with a specific count. This visualization provides a clear and concise overview of the educational attainment of individuals within the dataset. The pie chart displays different education categories, such as "Assoc-acdm," "Assoc-voc," "Bachelors," "Doctorate," "HS-grad," "Masters," "Preschool," "Prof-school," and "Some-college." Each slice of the pie corresponds to one of these education categories, and the size of each slice is proportional to the count of individuals who hold that particular educational qualification.For example, the "HS-grad" category, which represents individuals with a high school diploma or equivalent, has the largest slice in the pie, indicating that it is the most common educational level within the dataset with a count of 10,501. On the other hand, the "Preschool" category has the smallest slice, representing individuals with preschool-level education, and it has a count of 51.This pie chart offers a quick and visually intuitive way to grasp the distribution of education levels among the dataset's individuals. It provides essential insights into the educational diversity and composition of the dataset, aiding in better understanding the demographics and educational backgrounds of the included individuals.



Figure Violin plot Hours-per-Week vs. Income by Sex

In Figure 10, we present a Violin plot that visualizes the relationship between Hours-per-Week and Income, with a focus on gender as the stratifying variable. This visualization offers valuable insights into the distribution of working hours and income levels among different income groups. The plot showcases violin-shaped distributions, with each violin representing a distinct income category. The width of each violin corresponds to the density of data points, providing information about the concentration of values at different Hours-per-Week levels.

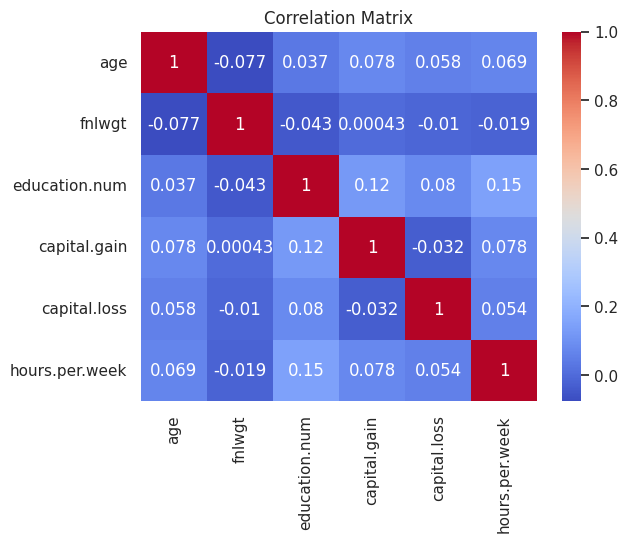


Figure heatmap of the correlation matrix

In Figure 11, we present a heatmap that visually represents the correlation matrix of the dataset. This heatmap provides valuable insights into the relationships and associations between different variables within the dataset. The heatmap uses a color scale to encode the strength and direction of correlations. Positive correlations are typically depicted in one color (e.g., shades of blue), while negative correlations may be shown in another color (e.g., shades of red or orange). The intensity of the color corresponds to the strength of the correlation, with darker shades indicating stronger relationships. By examining the heatmap, we can identify patterns of association between variables. Variables that exhibit strong positive correlations tend to move in the same direction, while those with strong negative correlations move in opposite directions. In contrast, weak or near-zero correlations are represented by lighter colors.The heatmap is a powerful tool for understanding the interdependencies and connections between variables in the dataset. It allows us to pinpoint which variables have significant relationships and can provide critical insights for data analysis and decision-making.



Figure Pairplot of Age, Education-num, and Hours-per-week by Income Level

Figure 12 The pairplot graph visualizes the relationships and distributions of three numerical variables ('age,' 'education.num,' and 'hours.per.week') by income level ('greater than $50,000' and 'less than or equal to $50,000') using color. The graph consists of scatter plots, histograms, and density plots arranged in a matrix.

Scatter Plots:

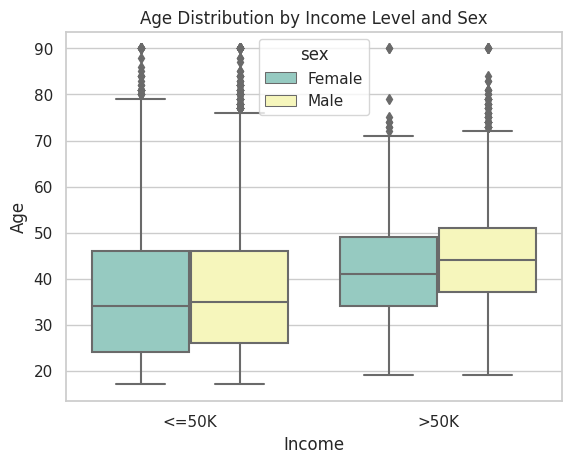
* The scatter plots in the graph show how pairs of numerical variables ('age,' 'education.num,' and 'hours.per.week') relate to each other.
* Each point on a scatter plot represents an individual in the dataset.
* The color of each point distinguishes between two income levels: red dots for incomes greater than $50,000 and green dots for incomes less than or equal to $50,000.
* Scatter plots help visualize the relationship and distribution of data points between these variables.

Histograms:

* The diagonal elements of the matrix display histograms for each numerical variable.
* Histograms show the distribution of values for each variable.
* They help understand the frequency or density of data points at different values of the variable.

Density Plots:

* The off-diagonal elements (non-diagonal squares) of the matrix show density plots.
* Density plots illustrate the distribution of data points as smooth curves, providing insights into data concentration and spread.

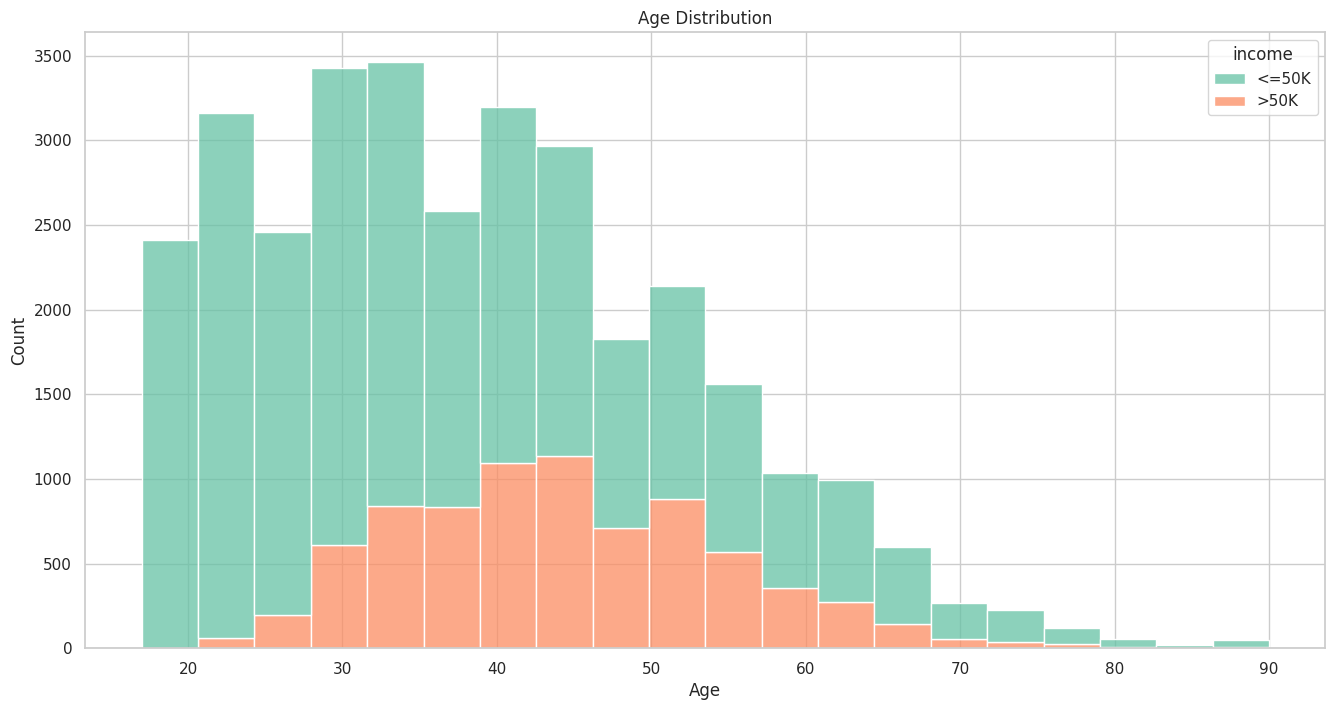


**Figure 13 Age Distribution by Income Level and Sex**

In Figure 13, For each income level ('greater than $50,000' and 'less than or equal to $50,000'), there are two boxplots, one for each gender ('Male' and 'Female'). These boxplots are arranged side by side for easy comparison.Each boxplot displays the distribution of ages for a specific group of individuals.The box in each boxplot represents the interquartile range (IQR), which contains the middle 50% of the ages in the group. The top and bottom edges of the box represent the 75th percentile (Q3) and the 25th percentile (Q1) of the age distribution, respectively.The line inside the box represents the median age (Q2 or the 50th percentile).The "whiskers" extend from the edges of the box to the minimum and maximum ages within a defined range (typically 1.5 times the IQR). Any data points beyond the whiskers are considered outliers and are plotted individually as dots.

The boxplots are color-coded based on gender (sex). For example, 'Male' = Yellow and 'Female.'= Blue This color distinction helps differentiate between age distributions for each gender within the income groups.This figure allows you to compare the age distributions of individuals with different income levels (greater than $50,000 and less than or equal to $50,000) and how those distributions vary between genders (Male and Female). It provides insights into factors such as income disparity and age-related patterns based on gender.

Top of Form



**Figure 15 Age Distribution**

In Figure 15, The histogram is a bar chart that shows the frequency or count of individuals falling into different age groups.The data is divided into 20 bins (specified by the bins=20 parameter), and the height of each bar represents the count of individuals within that age range.The histogram is divided into two colors based on the 'Income' level. It uses a stacked bar format, where one color represents individuals with 'Income' greater than $50,000, and the other color represents individuals with 'Income' less than or equal to $50,000.This stacking allows you to see the distribution of ages for both income groups in the same chart, making it easy to compare the two.The colors used in the histogram are based on the "Set2" color palette, which provides a distinct color for each category greater than $50,000 equal to orange and other is fgreen color.The x-axis is labeled as 'Age,' indicating the variable being plotted.The y-axis is labeled as 'Count,' indicating the number of individuals in each age group.The x-axis and y-axis labels help provide context to the graph. This histogram allows you to visually explore the distribution of ages in your dataset, with a particular focus on how it varies between individuals with different income levels (greater than $50,000 and less than or equal to $50,000). It helps you understand the age composition of the dataset and how it relates to income.

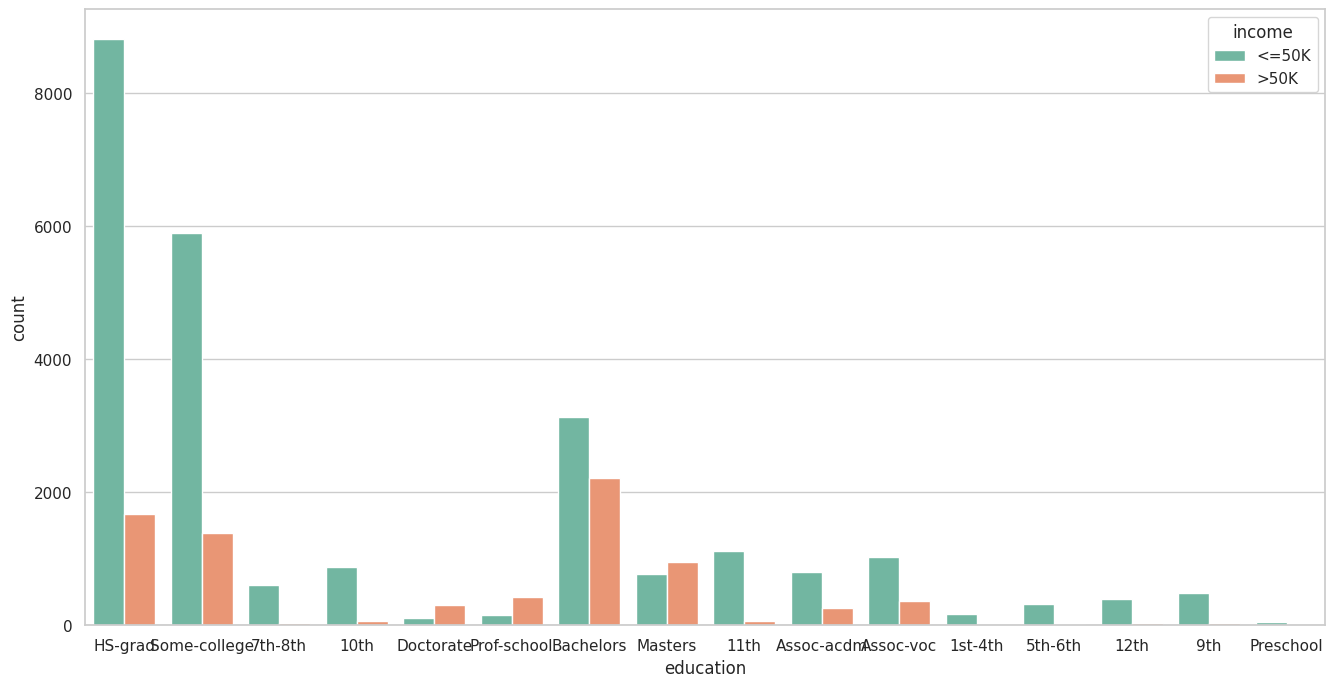


Figure 16 distribution of education levels

In Figure 16, A countplot is a bar chart that displays the frequency or count of each category within a categorical variable.Each bar on the chart represents a different education level, and the height of the bar indicates how many individuals in the dataset have that level of education.The bars are divided into two colors based on 'Income' level. The color distinction helps differentiate the count of individuals with 'Income' greater than $50,000 and 'Income' less than or equal to $50,000 for each education level.the red for greater than 50000 and green is for less than 50000

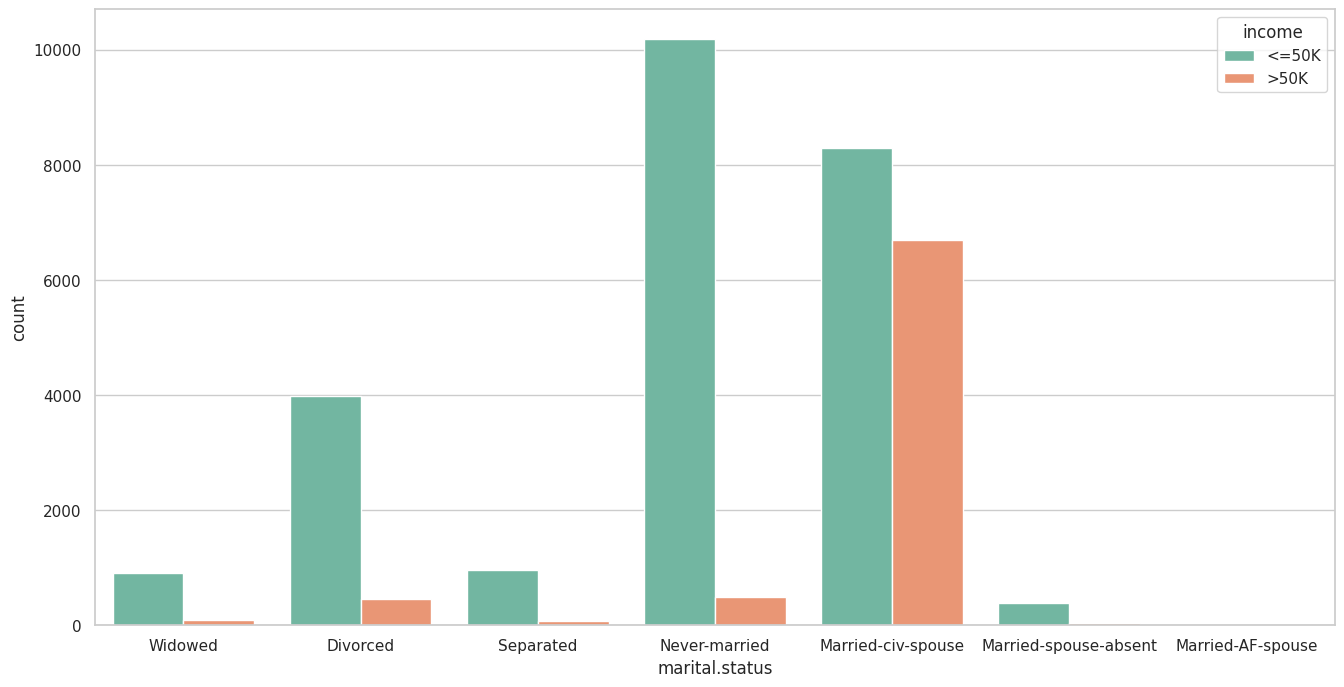


Figure 17 distribution of education levels

In Figure 17, generates a countplot to visualize the distribution of marital statuses in your dataset, with different colors representing income levels The countplot is a bar chart that visualizes the distribution of individuals among different marital status categories.The x-axis represents 'Marital Status,' indicating the various marital status categories.The y-axis represents the count or number of individuals in each category .For each marital status category, you can see the distribution of individuals with different income levels.By looking at the bars, you can visually compare the number of individuals in each category based on income: For example, you can compare the number of individuals with 'Married-civ-spouse' status who have income greater than $50,000 to those with income less than or equal to $50,000.

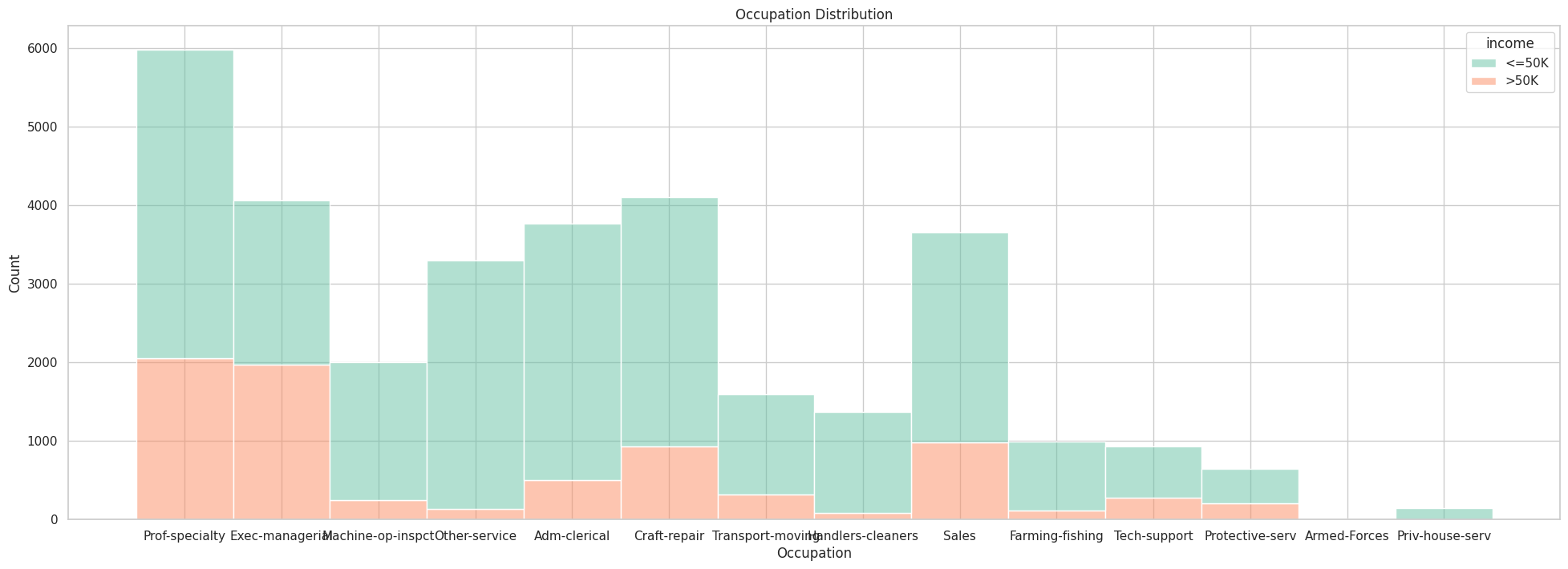


Figure 18 distribution of occupation

In Figure 18, generates The stacked histogram represents the distribution of individuals among different occupation categories.Each bar on the chart represents a different occupation, and the height of the bar indicates how many individuals in the dataset have that occupation.The bars are divided into two colors based on 'Income' level:One orange color represents individuals with 'Income' greater than $50,000.The blue other color represents individuals with 'Income' less than or equal to $50,000.For example, you can see that the 'Exec-managerial' occupation has a significant number of individuals with income greater than $50,000, and a smaller number with income less than or equal to $50,000.The stacked histogram helps you understand the distribution of income within each occupation category.It allows you to visually compare the number of individuals in each occupation category based on income levels.For instance, you can observe that 'Prof-specialty' and 'Exec-managerial' occupations have a substantial number of individuals with income greater than $50,000, while 'Priv-house-serv' has a much smaller number in that income category.

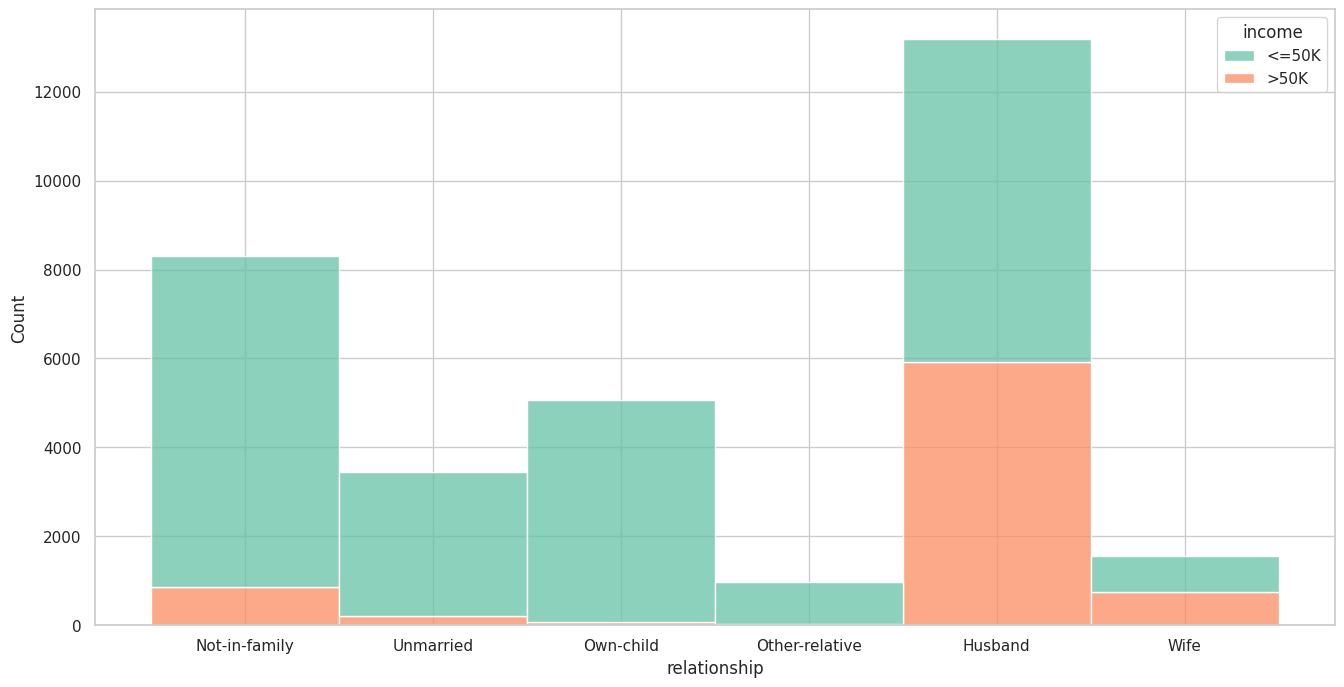


Figure 19 distribution of relationship

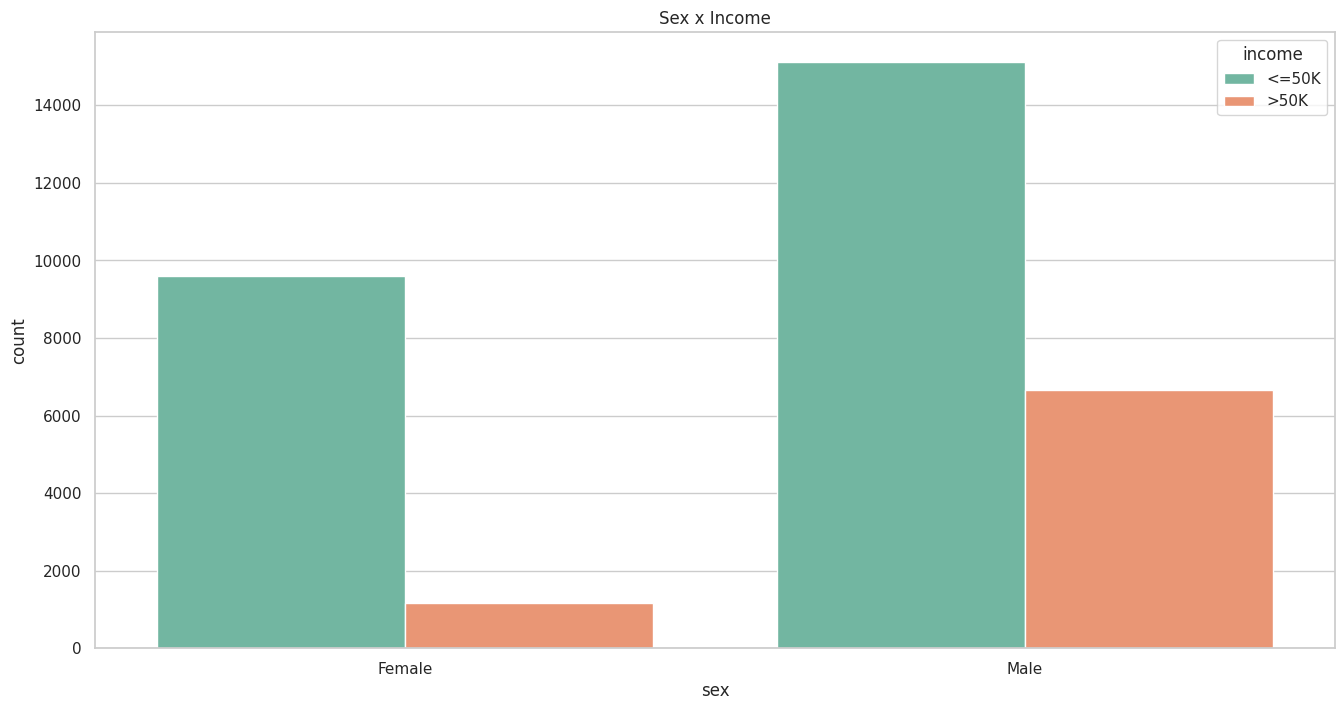
In Figure 19, the The stacked histogram is a bar chart that displays the frequency or count of each relationship category.Each bar on the chart represents a different relationship type, and the height of the bar indicates how many individuals in the dataset have that type of relationship.The bars are divided into two colors based on 'Income' level:One orange color represents individuals with 'Income' greater than $50,000.The other blue color represents individuals with 'Income' less than or equal to $50,000.For example, you can see that the 'Husband' relationship category has a significant number of individuals with income greater than $50,000, while 'Own-child' has a larger number of individuals with income less than or equal to $50,000.The stacked histogram helps you understand the distribution of income within each relationship category.It allows you to visually compare the number of individuals in each relationship category based on income levels.For instance, you can observe that individuals categorized as 'Husband' tend to have a substantial number with income greater than $50,000, while 'Own-child' has a larger number with income less than or equal to $50,000.

**A graph with a bar graph

Description automatically generated with medium confidence**

**Figure 20 Race x income countplot**

In Figure 20, The countplot allows you to compare the income distribution across different racial categories.It indicates that income levels vary among racial groups.While some racial categories have a higher proportion of individuals with income greater than $50,000 (e.g., White'), others have a higher proportion with income less than or equal to $50,000 (e.g., 'white ').The visualization helps highlight disparities in income levels across racial groups and provides insights into how income is distributed within each racial category in your dataset.

******

**Figure 21 Sex x income countplot**

In Figure 21, The countplot allows you to compare the income distribution between males and females.It indicates that income levels vary between genders.Males appear to have a higher proportion of individuals with income greater than $50,000 compared to females.The visualization helps highlight differences in income levels between genders in your dataset, showing that a larger proportion of males have higher incomes compared to females. Female: There are 10,771 females in your dataset. The countplot shows that a significant portion of females has income levels less than or equal to $50,000, while a smaller portion has income levels greater than $50,000.Male: The dataset includes 21,790 males. The countplot suggests that a considerable number of males have income levels greater than $50,000, with a smaller number having income levels less than or equal to $50,000.

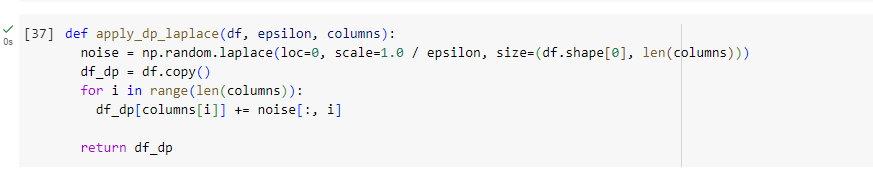
### **Apply DP on Dataset**

**DP** can be applied to any column of a dataset, but it is important to consider the following factors when choosing which columns to apply it to:

**Privacy sensitivity:** Some columns, such as age, race, and income, are more sensitive than others. It is important to apply DP to these columns with a higher privacy budget, which will add more noise to the data and make it more difficult to identify individuals.

**Accuracy:** DP can reduce the accuracy of the data, so it is important to apply it to columns where accuracy is less important. For example, it may be less important to know the exact age of a person than it is to know their general age range.

**Utility:** DP can also reduce the utility of the data, so it is important to apply it to columns where the data will still be useful even after noise has been added. For example, it may still be useful to know the age range of a person even if the exact age is not known.

****

**Figure 22 laplace**

Figure 22, shows jupyter cell apply\_dp\_laplace(df, epsilon, columns) is a function that takes three parameters:

* **df:** The DataFrame to which DP will be applied.
* **epsilon:** A privacy parameter that controls the amount of noise added to the data. Smaller values of epsilon provide stronger privacy guarantees but may result in noisier data.
* **columns:** A list of column names in the DataFrame df to which DP will be applied.

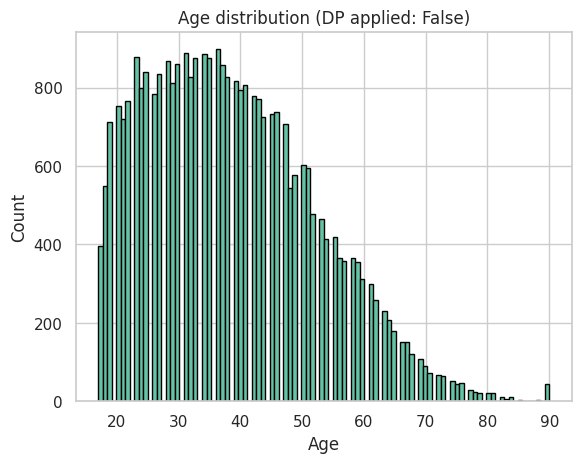
**noise = np.random.laplace(loc=0, scale=1.0 / epsilon, size=(df.shape[0], len(columns))):** This line generates Laplace noise. Laplace noise is a type of statistical noise that is often added to data to preserve privacy. It has a probability density function (PDF) that resembles a double-exponential distribution. The parameters are as follows:

* **loc=0:** The mean of the Laplace distribution (usually set to 0).
* **scale=1.0 / epsilon:** The scale parameter, which determines the spread of the noise. Smaller values of epsilon result in larger scale values and more noise.
* **size**=(df.shape[0], len(columns)): The size of the noise array, which matches the dimensions of the DataFrame. It generates noise for each row and each specified column.

**df\_dp = df.copy():** This line creates a copy of the original DataFrame df. The privacy-preserving operations will be applied to this copy to avoid modifying the original data.The loop for i in range(len(columns)): iterates through the specified columns.

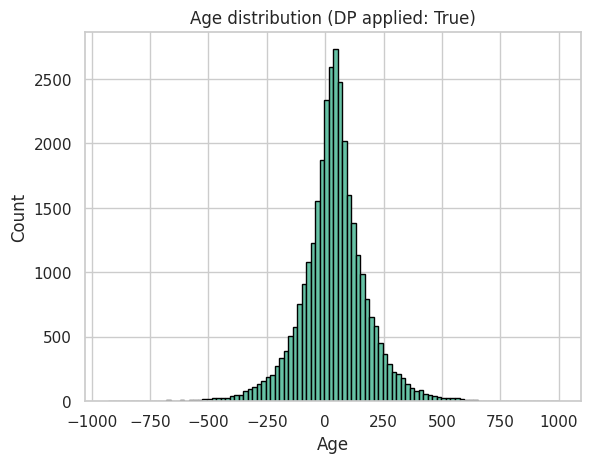
**df\_dp[columns[i]] += noise[:, i]:** For each column specified in columns, this line adds the corresponding Laplace noise to every row in the DataFrame. This step injects noise into the selected columns, thereby ensuring that the data becomes differentially private.

The result is a new DataFrame df\_dp that contains the original data with Laplace noise added to the specified columns. This process helps protect the privacy of individual data points while allowing for certain statistical analyses to be performed on the noisy data with privacy guarantees. The amount of noise added is controlled by the epsilon parameter, where smaller values of epsilon provide stronger privacy protection but may result in less accurate data analysis.



**Figure 22 Without DP applied**

Figure 22 The graph shows a bell-shaped curve with a peak at around age 36. This means that the median age of the population is 36 years old. The curve is slightly skewed to the right, meaning that there are less people in older age groups than in younger age groups.



**Figure 23 DP applied**

Figure 23 The graph shows a similar bell-shaped curve, but the peak is shifted to the right, meaning that the median age of the population is projected to increaseand decreased . The curve is also more skewed to the right Overall, the graphs show that the population is aging.

**Comparing the visualizations of the original data with the DP-protected data:**

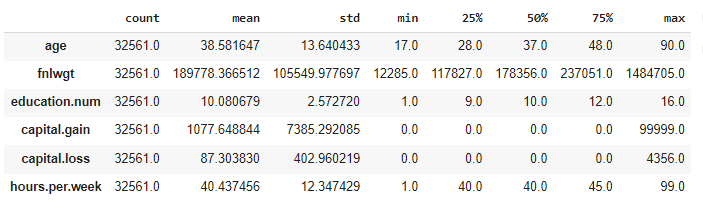
The two visualizations are very different overall, but there are some key differences. The original data shows a more detailed and nuanced picture of the age distribution of the population. For example, the original data shows that the 24-36 age group is slightly larger than the 36-80 age group, while the DP-protected data shows that the 24-36 age group is slightly larger than the 36-80 age group. Additionally, the original data shows that the -250-500 age group is slightly larger than the 500-700 and -750--500 age group.

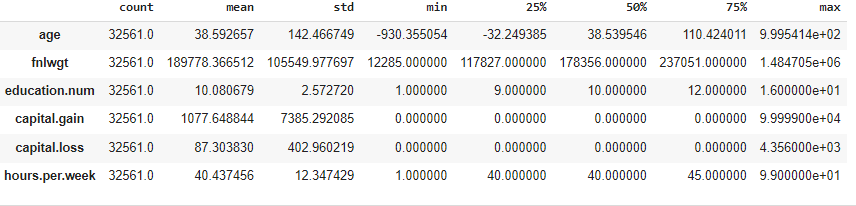
**How DP affects the clarity and interpretability of visualizations:**

DP can reduce the clarity and interpretability of visualizations by adding noise to the data. This is necessary to protect the privacy of individuals in the dataset. However, the amount of noise added should be minimized to the extent possible, so that the visualizations remain useful and informative.

**Whether the DP-protected data still provides meaningful insights for your analysis:**

No, DP-protected data may not show meaningful insights in plots. This is because DP adds noise to the data, which can obscure the underlying patterns and trends

**Figure 24 without DP Matrices**

** Figure 25 Applied DP Matrices**