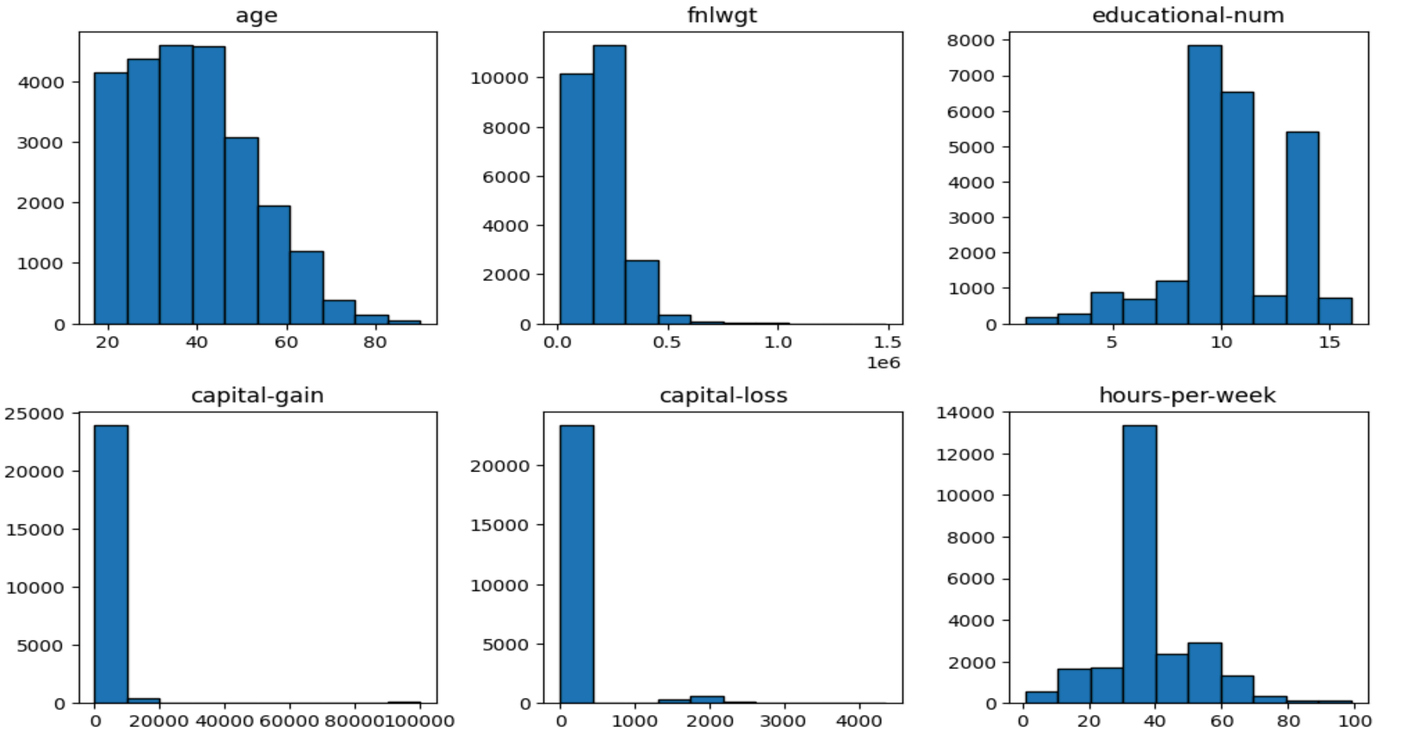
# REPORT

## Tools Used

- **Pandas**: For data manipulation and analysis.
- **Descriptive Statistics**: To compute summary statistics for numerical and categorical columns.
- **Transformers (Hugging Face)**:
    1. **Summarization**: Using the sshleifer/distilbart-cnn-12-6 model to summarize text insights.
    2. **Text Generation**: Using GPT-2 model for generating detailed explanations

## Generated Insights

1. **Numerical Columns:**



These histograms help visualize the distribution and skewness of each numerical feature in the dataset, providing insight into the typical values and the range of the data.

- **Age:**
    - The distribution of age is right-skewed.
    - Values range between 17 and 90. Most individuals fall between the ages of 20 and 50, with the highest concentration around the age of 30-40.
    - Average (mean): 38.64.
    - Median: 37.00.
    - Mode: 36 (occurs 1348 times).
    - Std: 13.64
    - Implications: Age could be a significant factor in analysis related to income, work hours, and educational attainment.

- **fnlwgt:**
    - The distribution is heavily skewed to the right, with most values concentrated below 200,000.
    - Values range between 12285 and 1490400.
    - Average (mean): 189664.13.
    - Median: 178144.50.
    - Mode: 203488 (occurs 21 times).
    - Std: 105604.03
    - Implications: When using weighted analysis, ensure proper handling of this feature to avoid bias.

- **educational-num:**
    - The distribution shows that most individuals have between 9 and 13 years of education, peaking around 10 years.
    - Values range between 1 and 16.
    - Average (mean): 10.08.
    - Median: 10.00.
    - Mode: 9 (occurs 15784 times).
    - Std: 2.57
    - Implications: This feature can be crucial in understanding income levels, job types, and other socio-economic factors.

- **capital-gain:**
    - The capital gain distribution is highly left-skewed,with most data points under 20,000.
    - Values range between 0 and 99999.
    - Average (mean): 1079.07.
    - Median: 0.00.
    - Mode: 0 (occurs 44807 times).

- Std: 7452.02
- Most individuals have a capital gain of 0, with very few having higher values
- Implications: Capital gain might be an indicator of additional income sources, investment behavior, or socio-economic status.

- **capital-loss:**
    - Similar to capital gain, the capital loss distribution is highly right-skewed.
    - The distribution of capital loss indicates that most observations fall below $1,000, with some between $1,000 and $2,000, and one observation between $3,000 and $4,000.
    - Values range between 0 and 4356.
    - Average (mean): 87.50.
    - Median: 0.00.
    - Mode: 0 (occurs 46560 times).
    - Std: 402.96
    - Implications: This could be used to study financial behavior, investment outcomes, and economic resilience.

- **hours-per-week:**
    - The distribution of hours worked per week shows a peak around 40 hours, which is common for full-time employment.
    - Values range between 1 and 99.
    - Average (mean): 40.42.
    - Median: 40.00.
    - Mode: 40 (occurs 22803 times).
    - Std: 12.35
    - Implications: This is a critical feature for analyzing labor patterns, job satisfaction, and correlating income with work hours

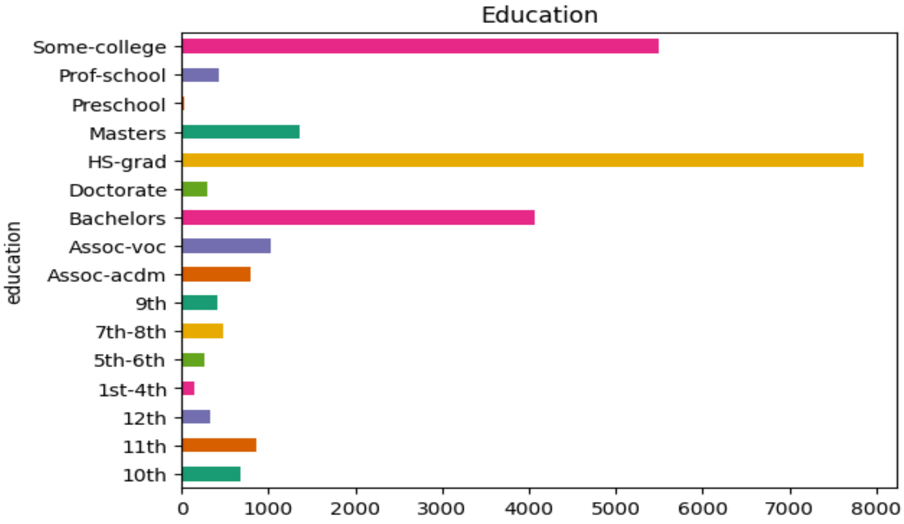
2. **Categorical Columns:**

    - **Workclass:**
        1. Most common value is 'Private' with 33906 occurrences.
        2. Insight: This column has the following distribution:

           Private: 33906, Self-emp-not-inc: 3862, Local-gov: 3136, ? : 2799, State-gov: 1981, Self-emp-inc: 1695, Federal-gov: 1432, Without-pay: 21, Never-worked: 10.

3. Explanation: The column's distribution shows that most individuals are employed in the private sector, with 33,906 entries. The second largest group is self-employed but not incorporated, with 3,862 entries. Local government employment ranks third with 3,136 entries. There are also 2,799 entries with missing or unspecified workclass information. The remaining categories each have fewer than 2,000 entries, with the smallest categories being 'Without-pay' and 'Never-worked'.

- **Education:**

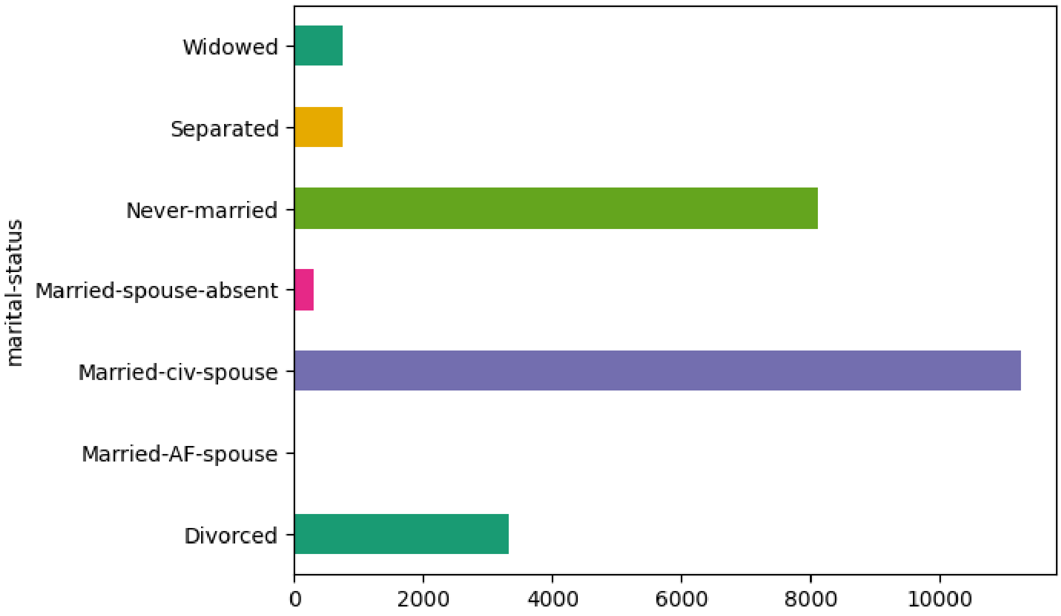


1. Most common value is 'HS-grad' with 15784 occurrences.
2. Insight: This column has the following distribution:

   HS-grad: 15784, Some-college: 10878, Bachelors: 8025, Masters: 2657, Assoc-voc: 2061, 11th: 2002, Assoc-acdm: 1601, 10th: 1389, 7th-8th: 1174, Prof-school: 1067, 9th: 946, 12th: 644, Doctorate: 594, 5th-6th: 509, 1st-4th: 247, Preschool: 83.

3. Explanation: The column shows that the highest number of individuals have a high school graduation level, with 15,784 entries. This is followed by those who have completed some college and those with a bachelor's degree. Master's degree holders number 2,657. Other categories such as associate degrees, 11th grade, and professional school have fewer entries. The least common educational levels are 'Preschool' with 83 entries and '1st-4th grade' with 247 entries.

- **Marital-status:**

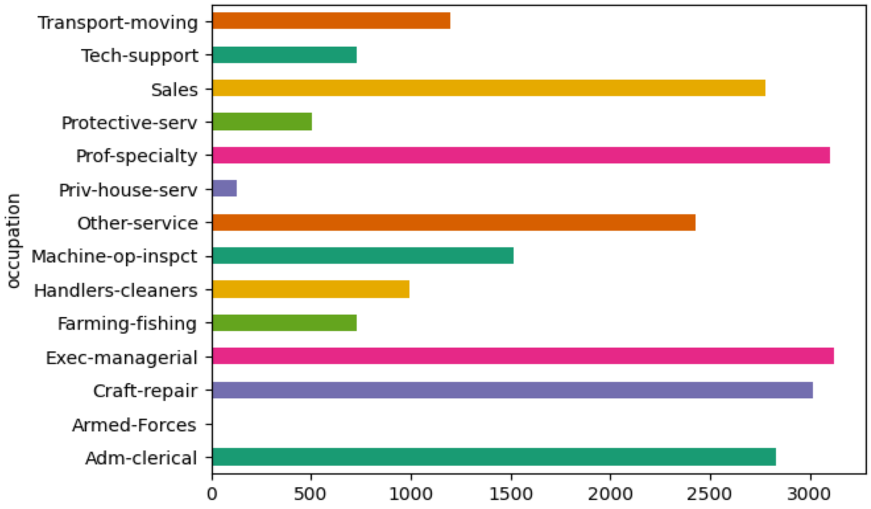


1. Most common value is 'Married-civ-spouse' with 22379 occurrences.
2. Insight: The column has the following distribution:

   Married-civ-spouse: 22379, Never-married: 16117, Divorced: 6297, Separated: 1530, Widowed: 1518, Married-spouse-absent: 628, Married-AF-spouse: 23."

3. Explanation: The column indicates that most individuals are 'Married-civ-spouse', followed by 'Never-married' individuals. 'Divorced' individuals number 6,297, while 'Separated' and 'Widowed' categories have 1,530 and 1,518 entries respectively. The least common categories are 'Married-spouse-absent' and 'Married-AF-spouse'.

- **Occupation:**



1. Most common value is 'Prof-specialty' with 6172 occurrences.
2. Insight: The column has the following distribution:

   Prof-specialty: 6172, Craft-repair: 6112, Exec-managerial: 6080, Adm-clerical: 5648, Sales: 5504, Other-service: 4808, Machine-op-inspct: 3022, ? : 2809, Transport-moving: 2355, Handlers-cleaners: 2072, Farming-fishing: 1480, Tech-support: 1457, Protective-serv: 982, Priv-house-serv: 232, Armed-Forces: 15.

3. Explanation: The 'occupation' column's distribution highlights that the 'Prof-specialty' category is the most common occupation, with 6,172 entries. 'Craft-repair' and 'Exec-managerial' follow closely with 6,112 and 6,080 entries respectively. Administrative and clerical positions, sales, and other service jobs are also prevalent. Occupations like 'Priv-house-serv' and 'Armed-Forces' are the least common, with only 232 and 15 entries respectively.

- **Relationship:**

1. Most common value is 'Husband' with 19716 occurrences.
2. Insight: The column has the following distribution:

   Husband: 19716, Not-in-family: 12583, Own-child: 7581, Unmarried: 5163, Wife: 2331, Other-relative: 1506."

3. Explanation: The column indicates that the 'Husband' category is the most common and the 'Not-in-family' category is the second most common. 'Own-child' and 'Unmarried' categories have 7,581 and 5,163 entries respectively.

The 'Wife' category has 2,331 entries, and the least common category is 'Other-relative' with 1,506 entries.

- **Race:**
  1. Most common value is 'White' with 41762 occurrences.
  2. Insight: The  column has the following distribution:

     White: 41762, Black: 4685, Asian-Pac-Islander: 1519, Amer-Indian-Eskimo: 470, Other: 406.

  3. Explanation: The column shows that the majority of individuals are 'White', with 41,762 entries. The second largest group is 'Black' with 4,685 entries. 'Asian-Pac-Islander', 'Amer-Indian-Eskimo', and 'Other' categories have significantly fewer entries, with 1,519, 470, and 406 entries respectively.

- **Gender:**
  1. Most common value is 'Male' with 32650 occurrences.
  2. Insight: The 'gender' column has the following distribution:

     Male: 32650, Female: 16192.

  3. Explanation: The column reveals that there are more 'Male' individuals compared to 'Female' individuals in the dataset.

- **Native-country:**
  1. Most common value is 'United-States' with 43832 occurrences.
  2. Insight: The column has the following distribution:

     United-States: 43832, Mexico: 951, Philippines: 295, Germany: 206, Canada: 206, Puerto-Rico: 184, El-Salvador: 182, India: 151, Cuba: 138, England: 127, Jamaica: 106, South: 103, China: 99, Italy: 97, Dominican-Republic: 93, Vietnam: 86, Guatemala: 86, Japan: 81, Poland: 78, Columbia: 73, Taiwan: 67, Haiti: 67, Iran: 59, Portugal: 56, Nicaragua: 49, Peru: 46, Greece: 46, Ecuador: 45, France: 38, Ireland: 37, Hong: 30, Trinidad Tobago: 28, Cambodia: 28, Thailand: 23, Laos: 23, Yugoslavia: 16, Outlying-US(Guam-USVI-etc): 14, Hungary: 13, Honduras: 13, Scotland: 12, Holland-Netherlands: 1

  3. Explanation: The column indicates that the majority of individuals are from the 'United States', with 43,832 entries. The second most common country is 'Mexico' with 951 entries. Other countries such as the 'Philippines', 'Germany',

and 'Canada' also appear, but with significantly fewer entries. The least common country of origin in the dataset is 'Holland-Netherlands' with just 1 entry.

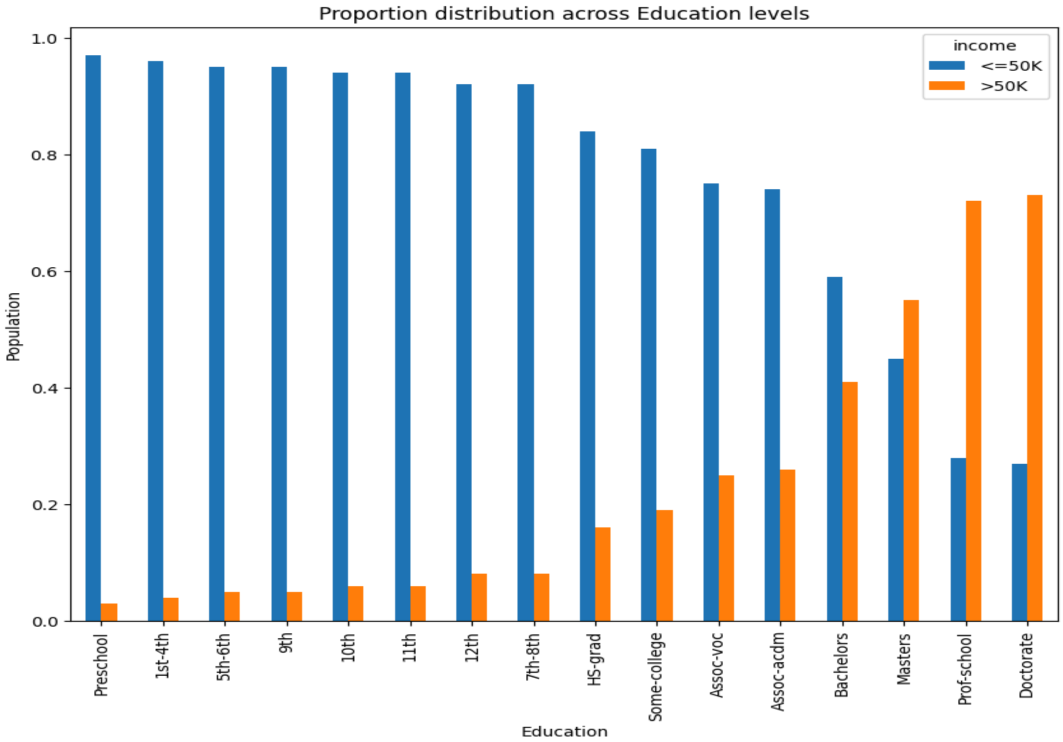
- **Income:**
  1. Most common value is '<=50K' with 37155 occurrences.
  2. Insight: The column has the following distribution:

     <=50K: 37155, >50K: 11687.

  3. Explanation: The column indicates that a majority of individuals have an income of less than or equal to 50K, while a smaller group have an income greater than 50K.

**Comparative plots:**

- **Education vs Income:**

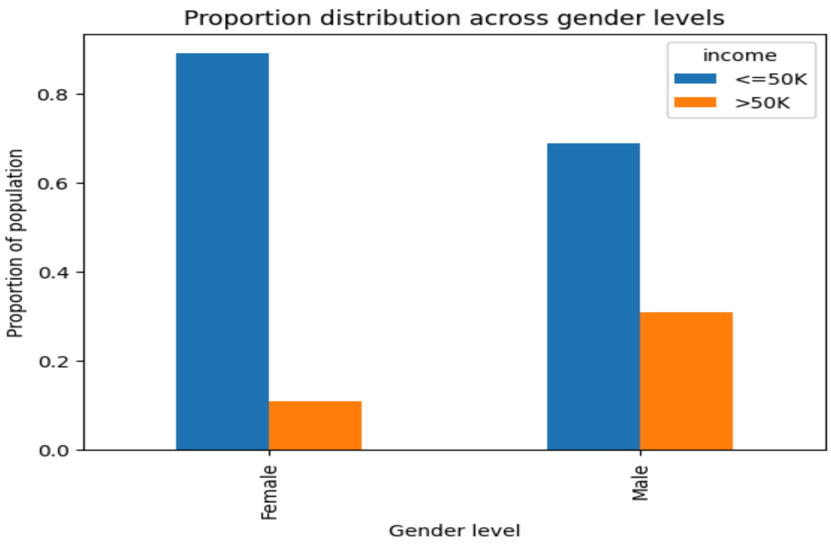


The graph illustrates the relationship between education level and income, specifically focusing on two income groups: those earning less than or equal to 50K (<=50K) and those earning more than 50K (>50K). There is a clear positive correlation between education level and income. As the education level increases, the proportion of individuals earning more than 50K also increases.

- Impact of Higher Education:
  Individuals with a higher education level, such as a Bachelor's, Master's, or Doctorate degree, have a significantly higher proportion of individuals earning above 50K compared to those with lower education levels. There is a noticeable jump in the proportion of individuals earning more than 50K for those who have completed high school (HS-grad) compared to those who have not.
- Potential Outliers:
  While the overall trend is clear, there are slight variations in specific categories. For example, the proportion of individuals with some college education earning above 50K is relatively lower compared to the general trend. This could be due to various factors like the field of study or economic conditions.

Overall, this graph highlights the importance of education in achieving higher income levels. It emphasizes the need for accessible and quality education to reduce income inequality and provide opportunities for upward mobility.

- **Gender vs Income:**



We observed a graph depicting gender proportions based on income levels. Females dominate both income categories.

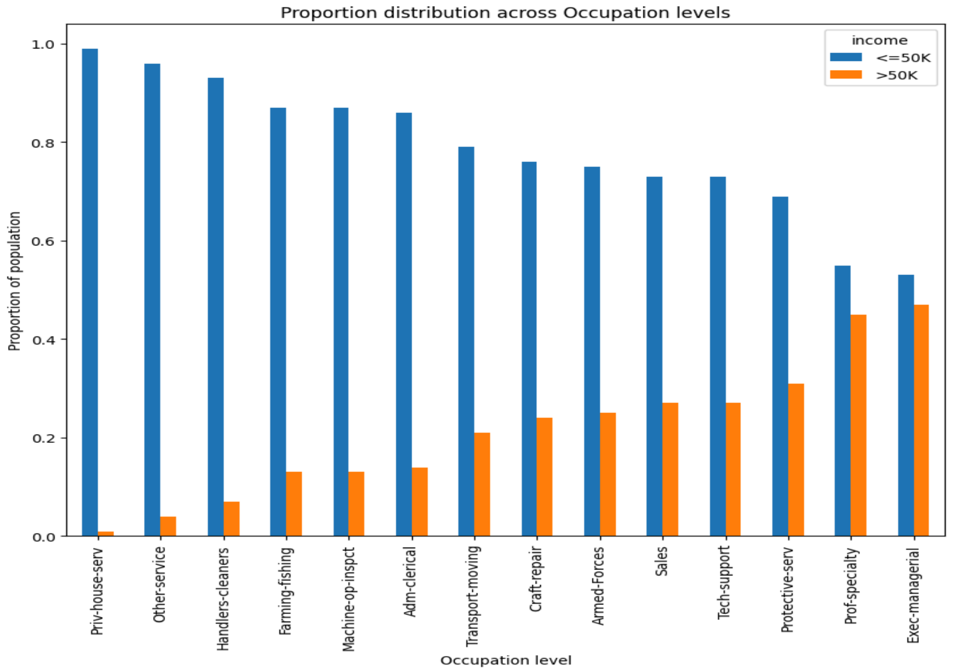
  - Females with income ≤ 50k form the highest proportion.
  - Males with income > 50k rank second.
  - Females with income > 50k follow.
  - Males with income ≤ 50k are the minority.

- **Occupation Level vs Income:**

  The graph presents the proportion of the population in various occupation categories, further divided by income levels: <=50K and >50K.The blue bars denote ≤ $50k, while orange bars denote > $50k.

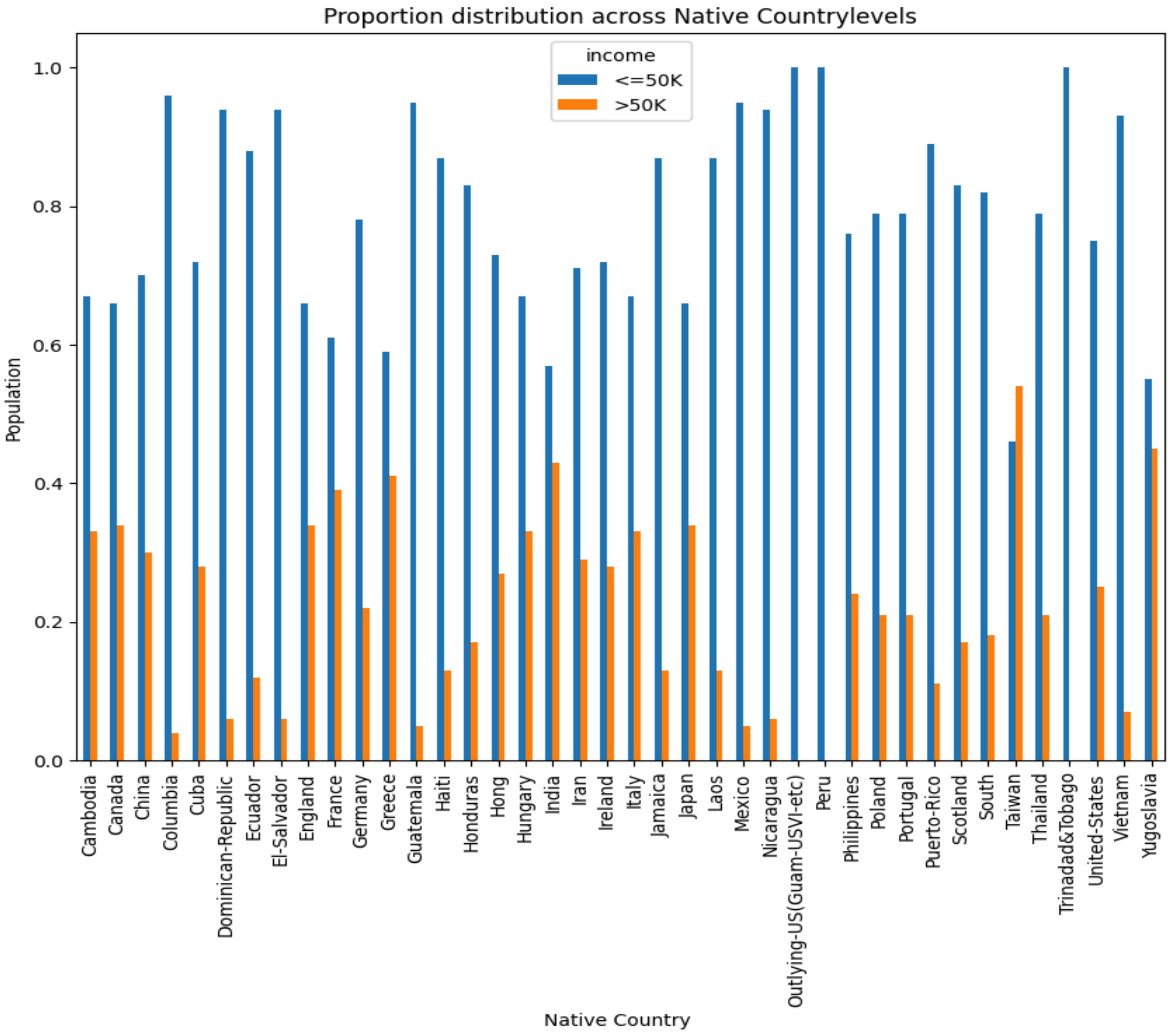
  Income disparity across occupations:
  There's a clear income gap between different occupations. For example, a larger proportion of people in Exec-managerial positions earn >50K compared to those in Handlers-cleaners roles.

  - High-income occupations:
    Occupations like Exec-managerial and Prof-specialty have a higher proportion of individuals earning >50K, suggesting these are typically higher-paying fields.
  - Low-income occupations:
    Occupations such as Farming-fishing, Handlers-cleaners, and Other-service have a larger proportion of individuals earning <=50K, indicating these are typically lower-paying fields.
  - Middle-income occupations:
    Occupations like Sales, Adm-clerical, and Craft-repair show a more even distribution between income levels, suggesting a mix of income potential within these fields.



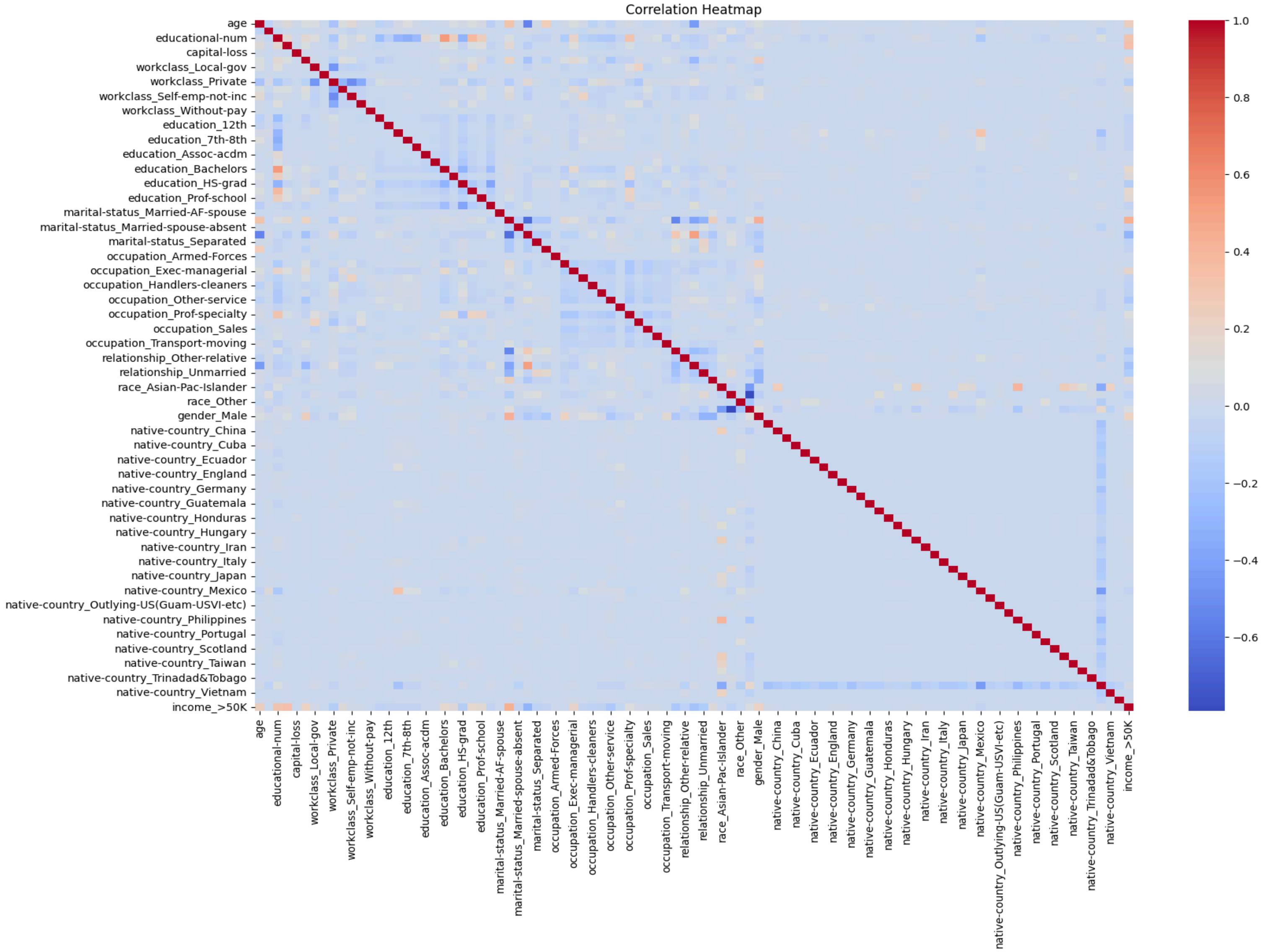
- **Native-country vs Income:**

The graph illustrates income proportions across countries, with most having a higher percentage of earners making ≤ $50k. Cambodia, Laos, and Vietnam have the highest proportions in this category, while the United States, Canada, and Switzerland have the lowest. Developed nations generally have more earners making > $50k, with many countries showing an upward trend in this bracket. Governments need policies to tackle inequality and social issues, with options like tax reforms, education/training investments, and strengthening social safety nets.



## Correlation Analysis:

Correlation heatmaps were generated before and after dropping columns with low correlation to the income variable. The initial heatmap provided a comprehensive view of relationships between all features and the target variable.



Strong Positive Correlations:

- There are strong positive correlations among certain features related to education and work, which can be seen by the dark red patches off the diagonal.
- Education level "Bachelors" and educational-num (correlation = 0.50).
- Age and marital status "Never married" (correlation = -0.54).

Strong Negative Correlations:

- Some strong negative correlations are visible (dark blue patches). For instance, native-country_X categories might have strong negative correlations with each other due to the nature of one-hot encoding.
- Age and marital-status[Never-married] (-0.538): Younger individuals are more likely to be never married.

Interesting Feature Interactions:

- Income:

  Income_>50K has some visible positive and negative correlations with features like education_Bachelors, occupation_Exec-managerial, and marital-status_Married-civ-spouse, suggesting these features are influential in predicting income.

- Education and Work:

  Education-related features (education_Bachelors, education_Masters, etc.) show correlations with occupation categories, indicating the influence of education level on job type.

## Outcomes

This summary report provides a detailed overview of the dataset, highlighting key statistics and common values for each column. Such insights can be useful for understanding the data distribution, identifying patterns, and making informed decisions for further analysis or modeling.