Problem Statement:

In today's data-driven world, effectively utilizing structured datasets is crucial for gaining a competitive edge. Knowledge Representation involves organizing raw data into a structured form that highlights relationships and meanings, making it easier to work with. Insights Generation is the process of analyzing this organized data to uncover patterns and trends that can inform decision-making

Unique Idea Brief (Solution):

We provide Solution for : Policy Implications

Policy Implications: Insights derived from the analysis can inform policy decisions aimed at addressing income inequality. For instance, enhancing educational opportunities and vocational training programs could significantly impact income distribution.

Features Offered:

Age: The age of the individual

Workclass: The type of employment, e.g., Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.

fnlwgt: The final weight, representing the number of people the census believes the entry represents.

Education: The highest level of education attained by the individual, e.g., Bachelors, HS-grad, 11th, Masters, 9th, Some-college, Assoc-acdm, Assoc-voc, 7th-8th, Doctorate, Prof-school, 5th-6th, 10th, 1st-4th, Preschool, 12th.

Education-Num: The number of years of education completed.

Marital Status: The marital status of the individual, e.g., Married-civ-spouse, Divorced,

Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.

Occupation: The type of occupation, e.g., 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: The individual's relationship to the household, e.g., Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.

Race: The race of the individual, e.g., White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
Sex: The gender of the individual, e.g., Male, Female.

Capital Gain: Income from capital gains
.
Capital Loss: Losses from capital.

Hours per Week: The number of hours worked per week.

Native Country: The country of origin, e.g., 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.

Income: The income class, divided into '>50K' and '<=50K' categories.

Processflow :

1. Data Collection

- **Source**: Obtain the Adult Dataset from the UCI Machine Learning Repository or other reliable sources.
- **Format:** Ensure the dataset is in a readable format (e.g., CSV)**.**

2. Data Preprocessing

- **Load Data:**
  - Load the dataset into a data processing tool (e.g., Python with pandas).
- **Data Cleaning:**
  - Handle missing values (e.g., impute, drop).
  - Remove or correct any anomalies or inconsistencies.
- **Data Transformation:**
  - Convert categorical variables to numerical format (e.g., one-hot encoding).

    - Normalize/standardize numerical features if necessary.
- Feature Engineering:
    - Create new features from existing ones (e.g., binning age into age groups, creating income brackets).

3. Data Exploration and Visualization

- **Descriptive Statistics:**
    - Calculate basic statistics (mean, median, mode) for numerical features.
    - Determine the distribution of categorical features.
- **Visual Exploration:**
    - Use histograms, bar charts, and box plots to visualize feature distributions.
    - Create correlation matrices to understand relationships between features.

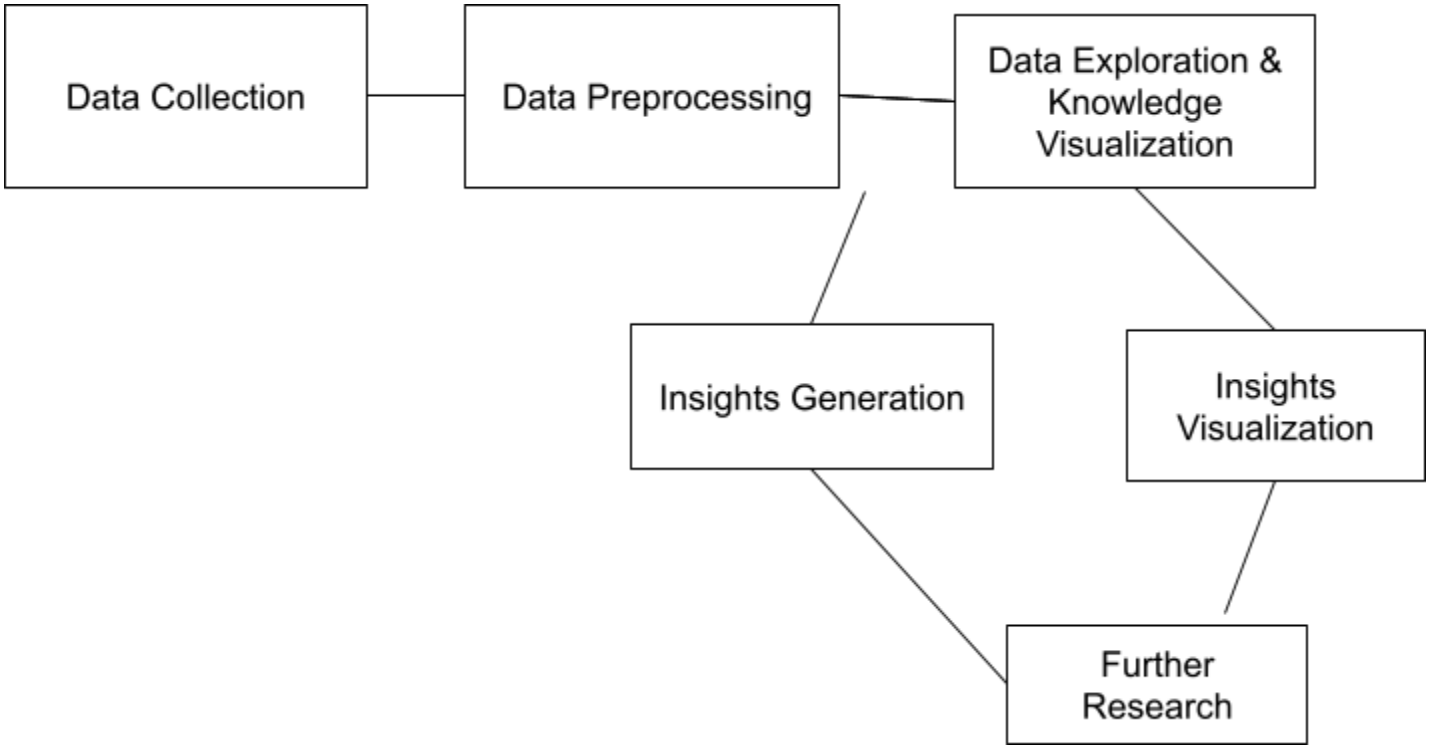
4. Knowledge Representation

In this we draw frequency and distributed graphs.for target column and other features.

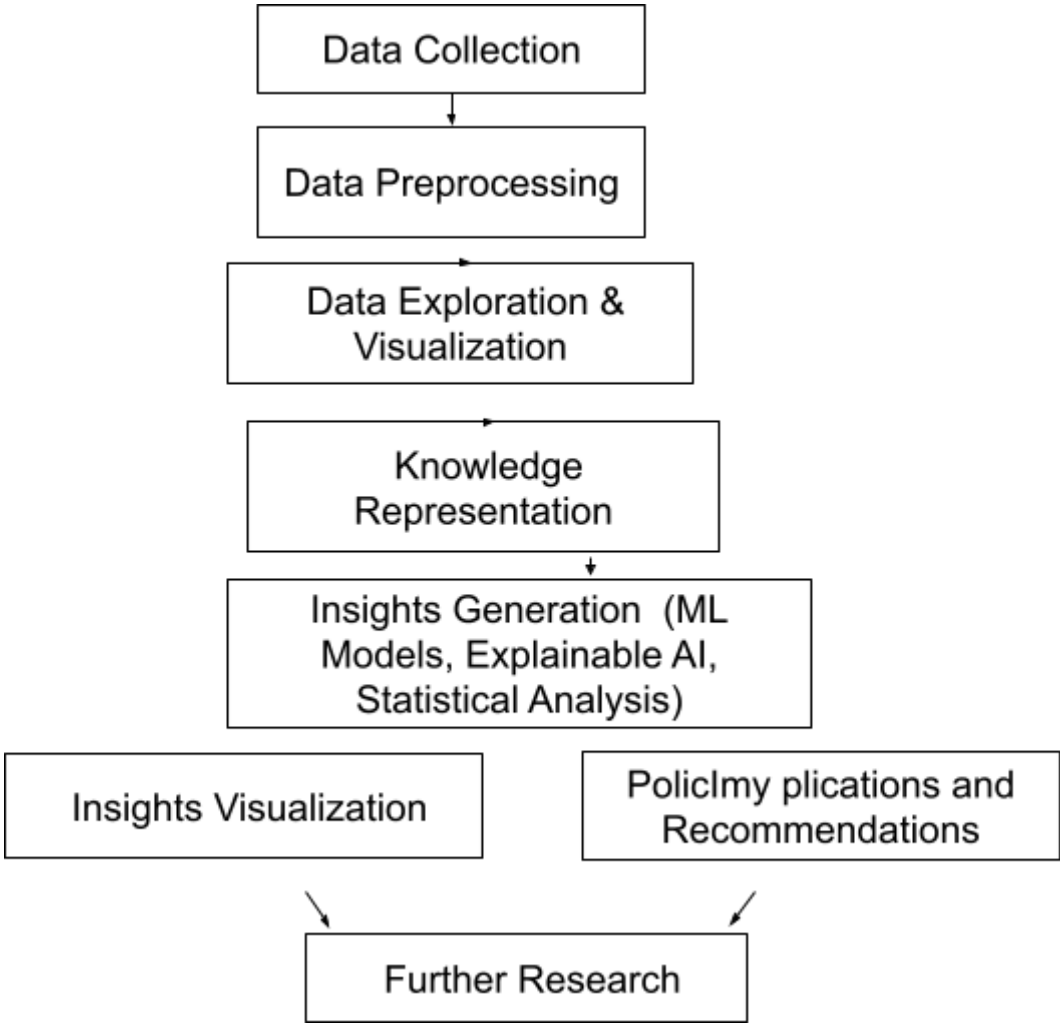
5. Insights Generation

- **Machine Learning Models:**
  - **Classification:**
    - Apply algorithms (e.g., decision trees, random forests) to predict income class.
  - **Clustering:**
    - Use clustering techniques (e.g., K-means) to identify distinct demographic groups.

Diagram of the Process Flow:



Architecture Diagram :->


Data Collection
Data Preprocessing
Data Exploration & Visualization
Knowledge Representation
Insights Generation (ML Models, Explainable AI, Statistical Analysis)
Insights Visualization
PolicImy plications and Recommendations
Further Research

Technologiesused:

Programming Languages:

Python: Widely used for its extensive libraries in data analysis and machine learning.

Libraries and Frameworks:

- **Pandas:** Data manipulation and analysis in Python.
- **Scikit-learn:** Machine learning library in Python for classification, regression, clustering, and more.
- **Matplotlib:** Plotting library in Python for creating static, animated, and interactive visualizations.
- **Seaborn:** Statistical data visualization library based on Matplotlib, providing a high-level interface for drawing attractive and informative statistical graphics.

Statistical Analysis Tools:

- **Statsmodels:** Python library for estimating and testing statistical models.
- **SciPy:** Python library for scientific and technical computing, including statistical tests.

Development Environment:

- **Jupyter Notebook / JupyterLab:** Interactive computing environments for creating and sharing documents containing live code, equations, visualizations, and narrative text.

Team members and contribution:

While Making Project Each Team Member Make Efforts For Sucess Completion of project .

Team Members :

Harmandeep Kaur
Harmandeep Singh
Jashandeep Singh
Vansh
Vikramjeet Singh

Contribution:

Harmandeep kaur and Harmandeep Singh: Making Report and Presentation.

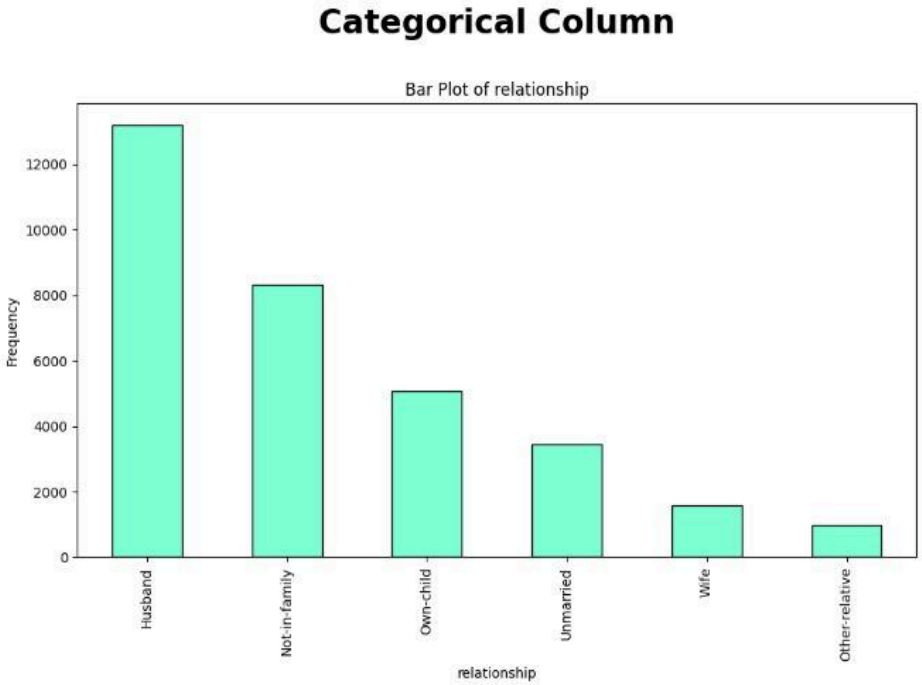
Jashandeep Singh: Manage Frontend.

Vansh : Handle Knowledge Representation.

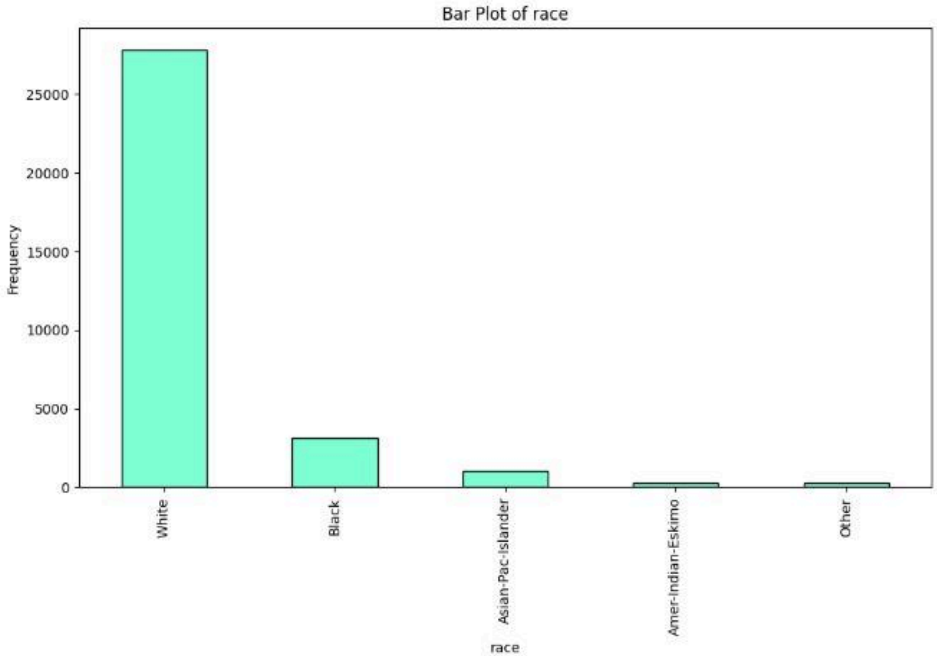
Vikramjeet Singh: Handle important part of Project manage all project code.

Conclusion:

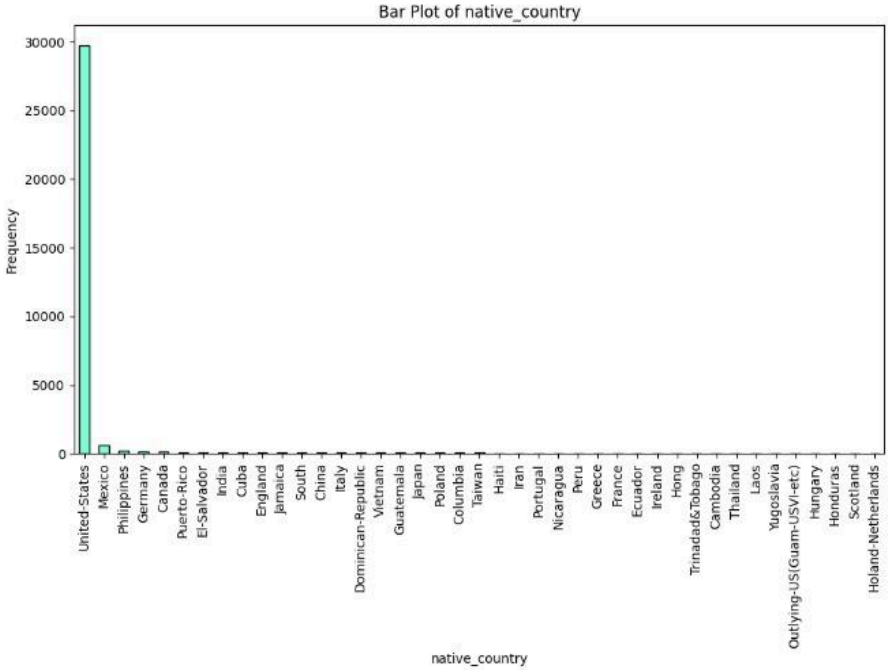
We make conclusion in the form of Graphs.



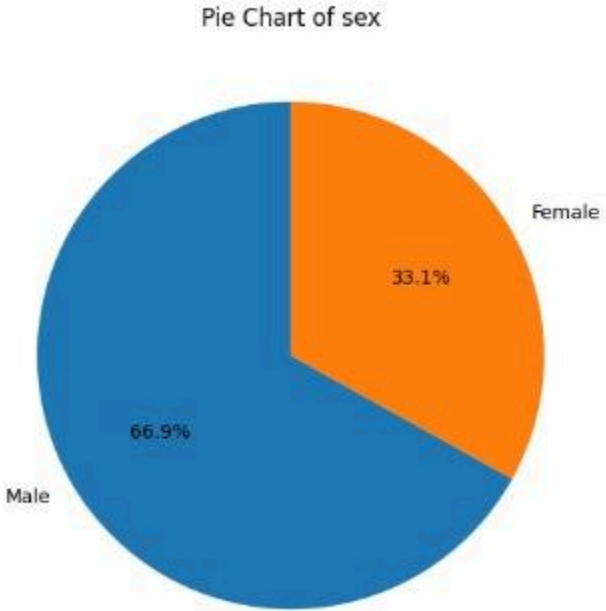
Categorical Column



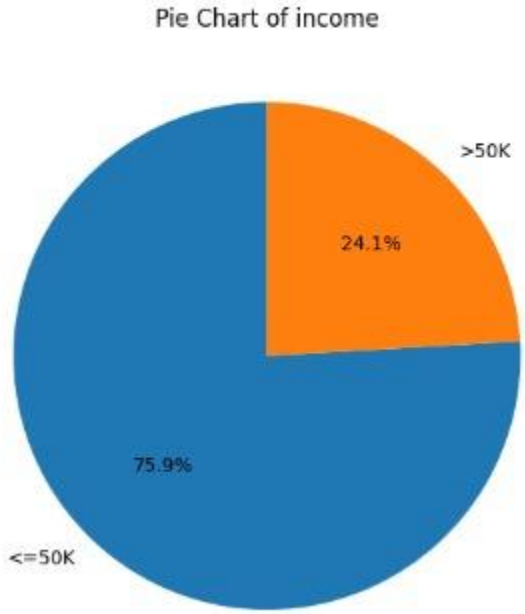
Categorical Column



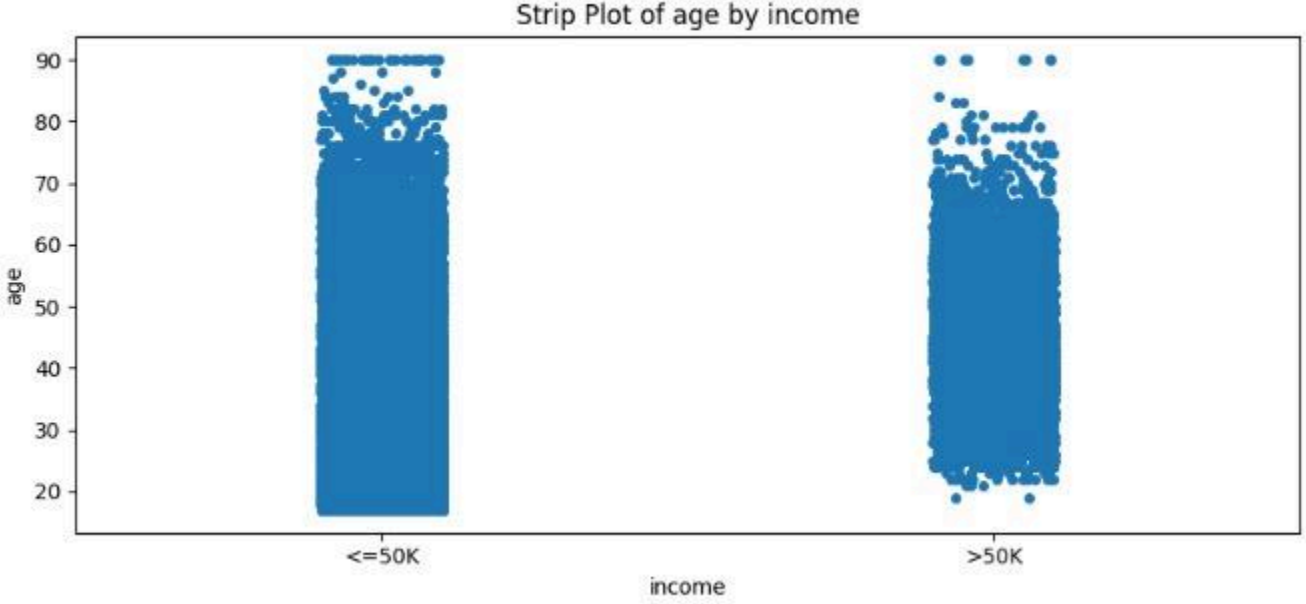
Boolean Column



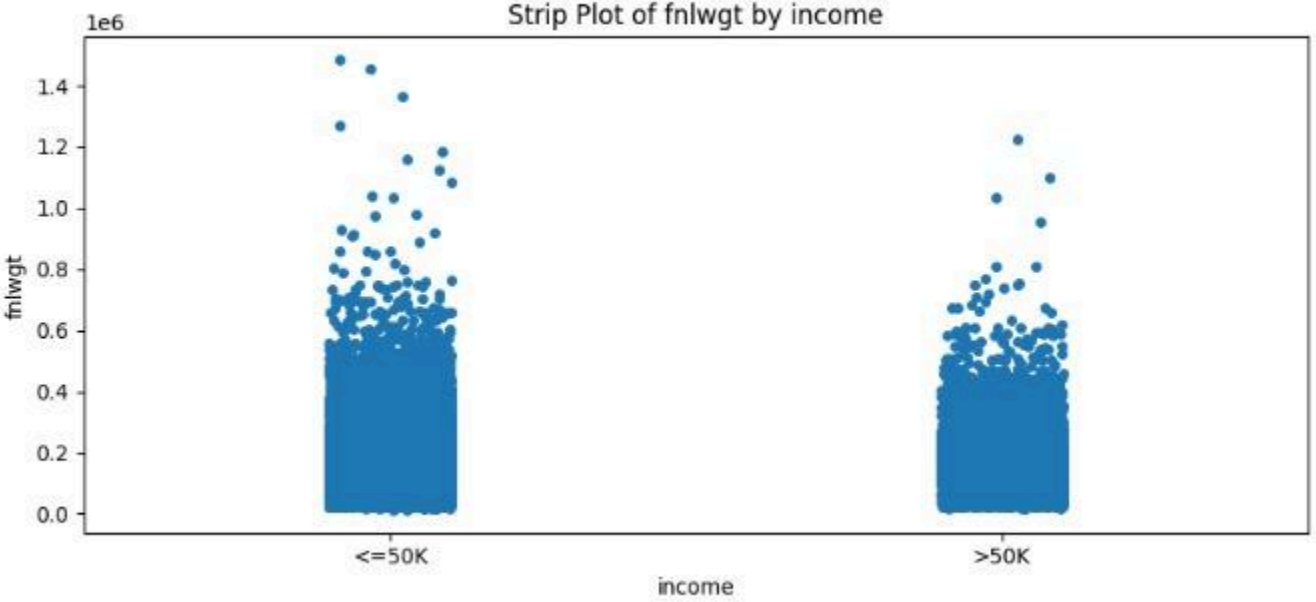
Target Column



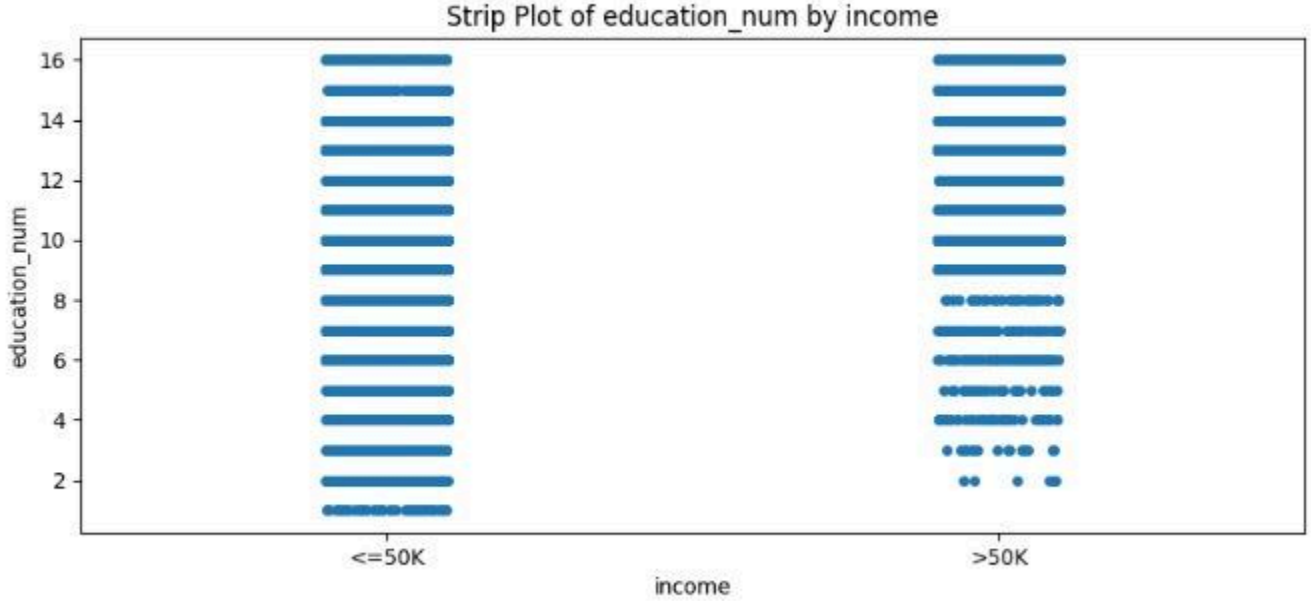
Numeric Column



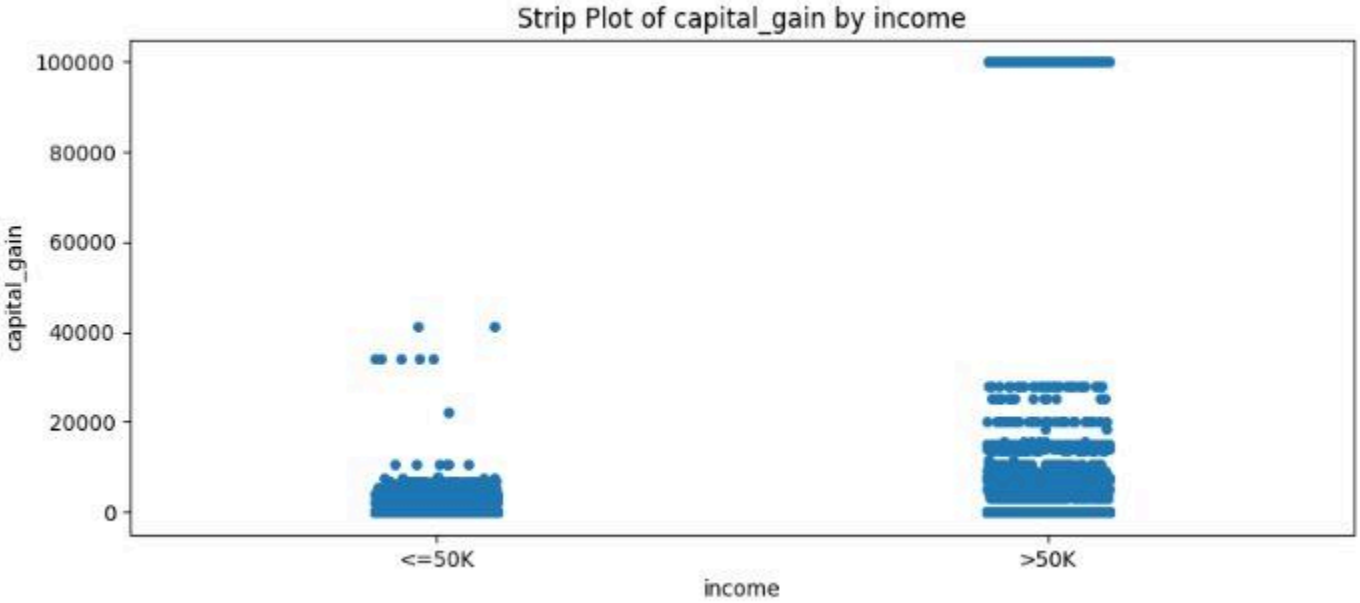
Numeric Column



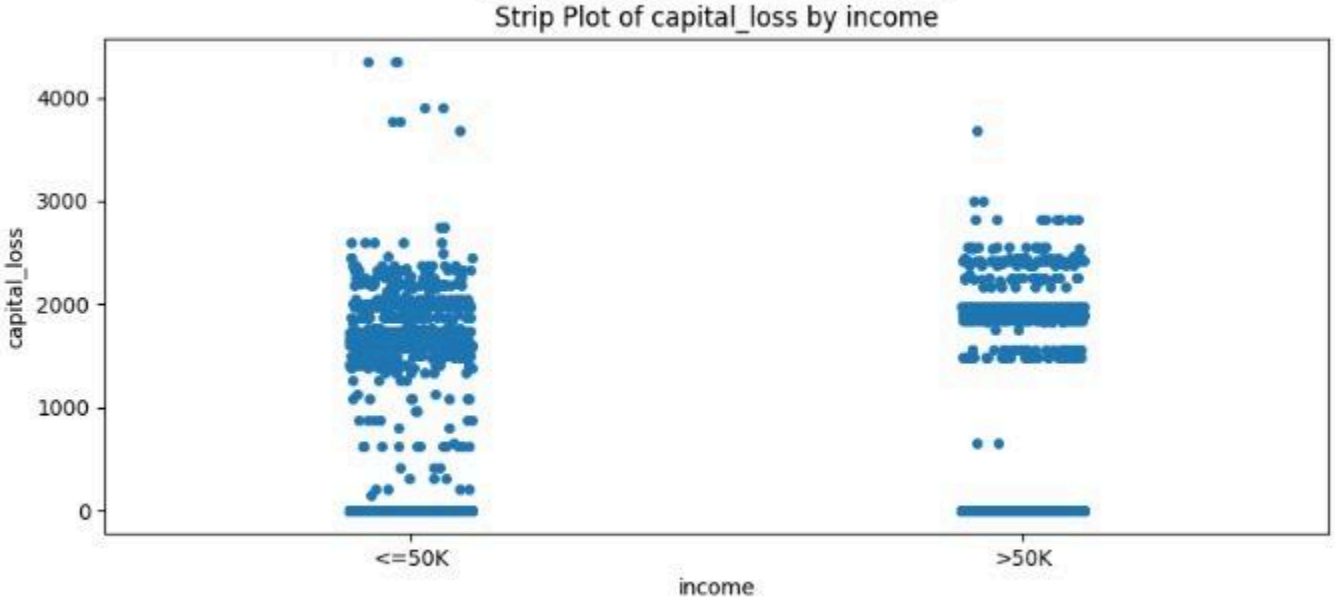
Numeric Column



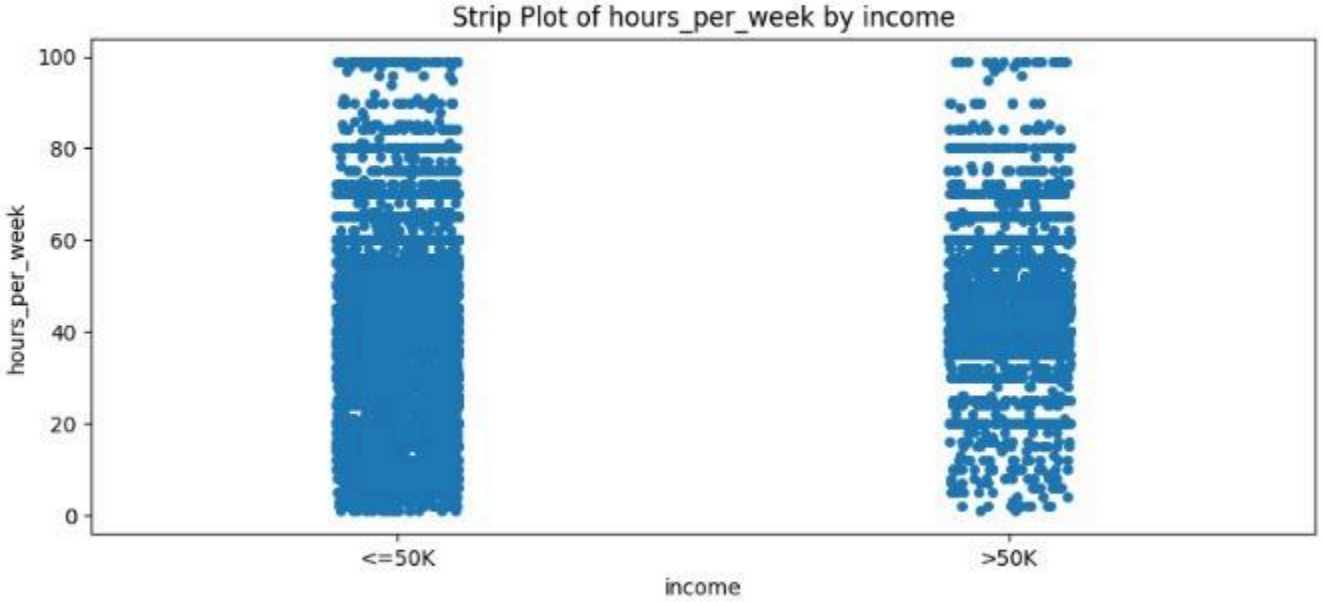
Numeric Column



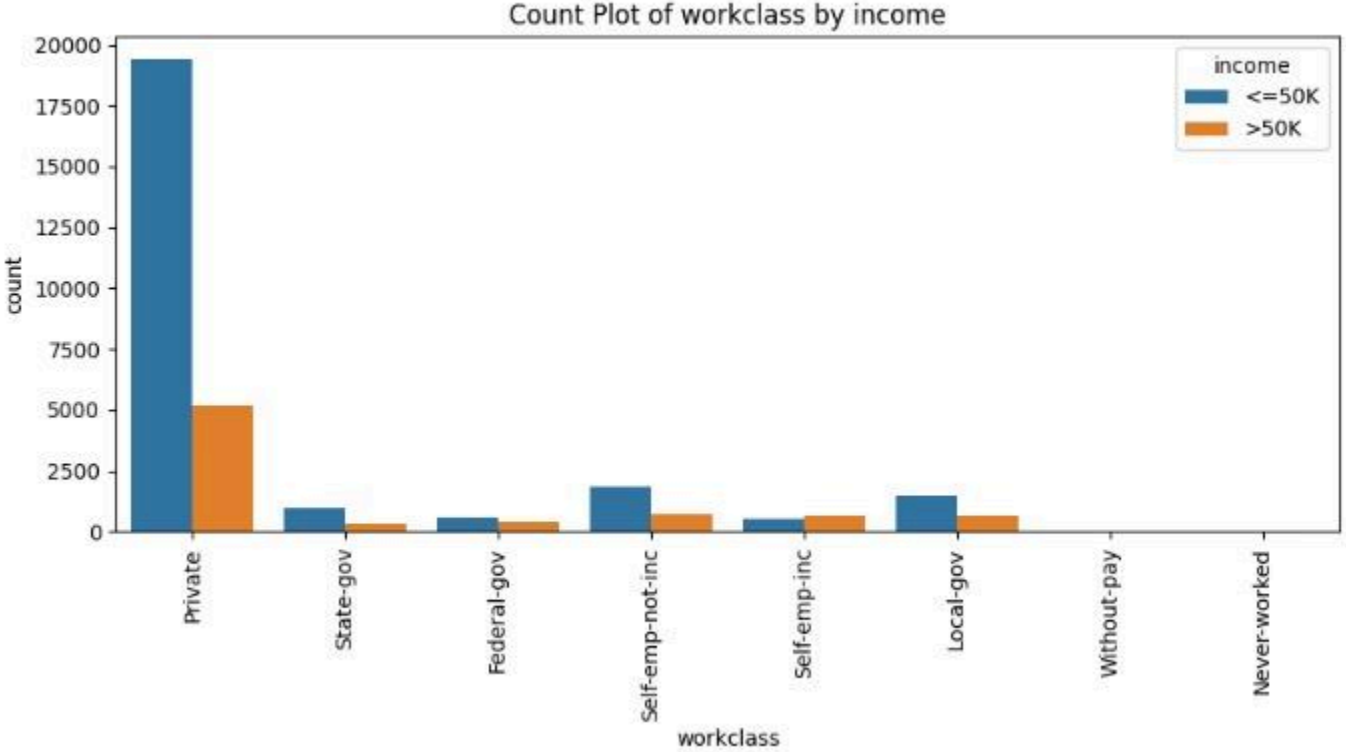
Numeric Column



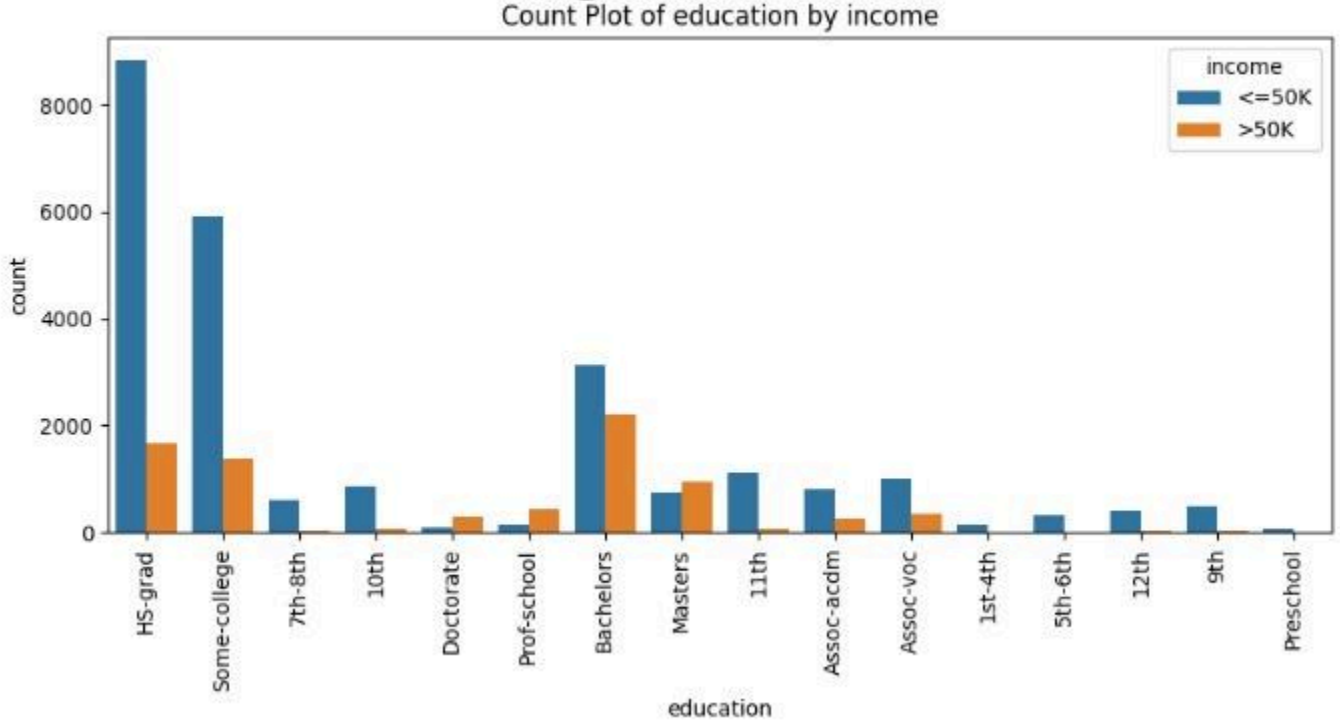
Numeric Column



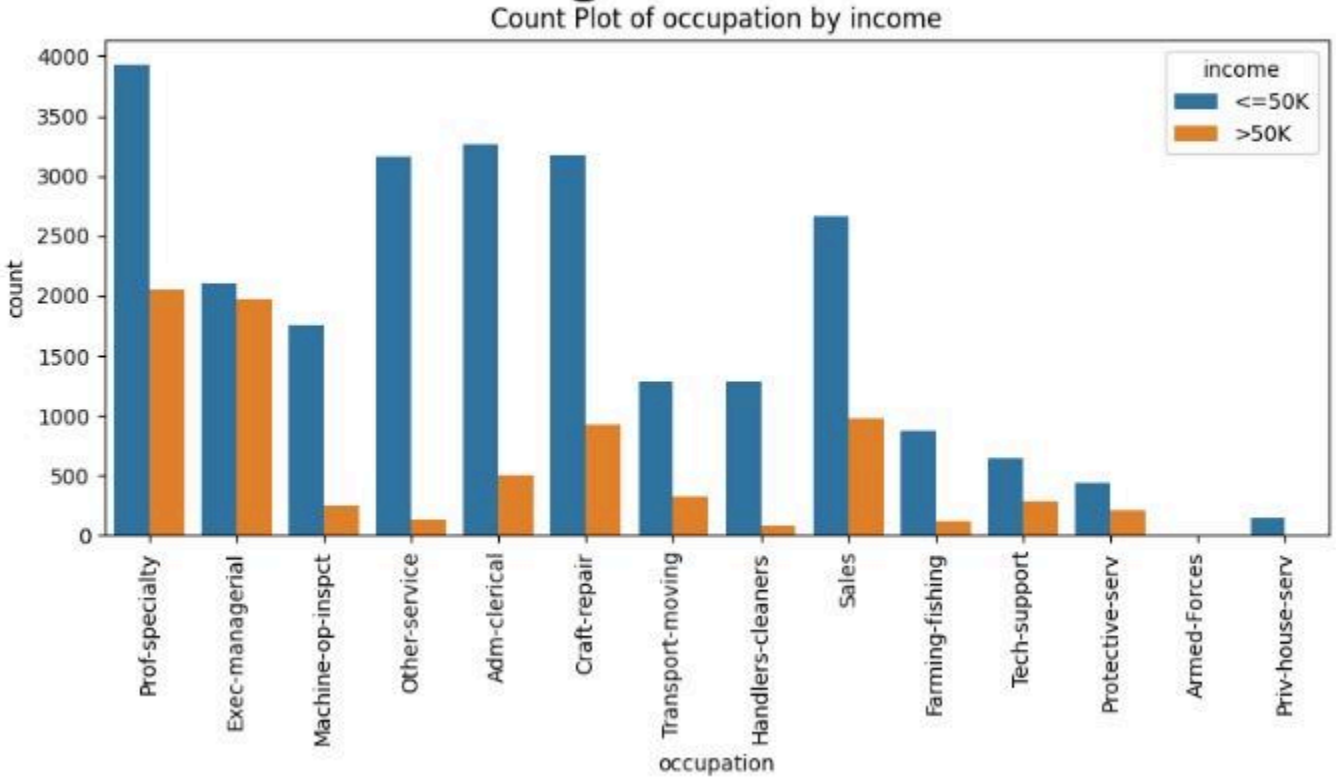
Categorical Column



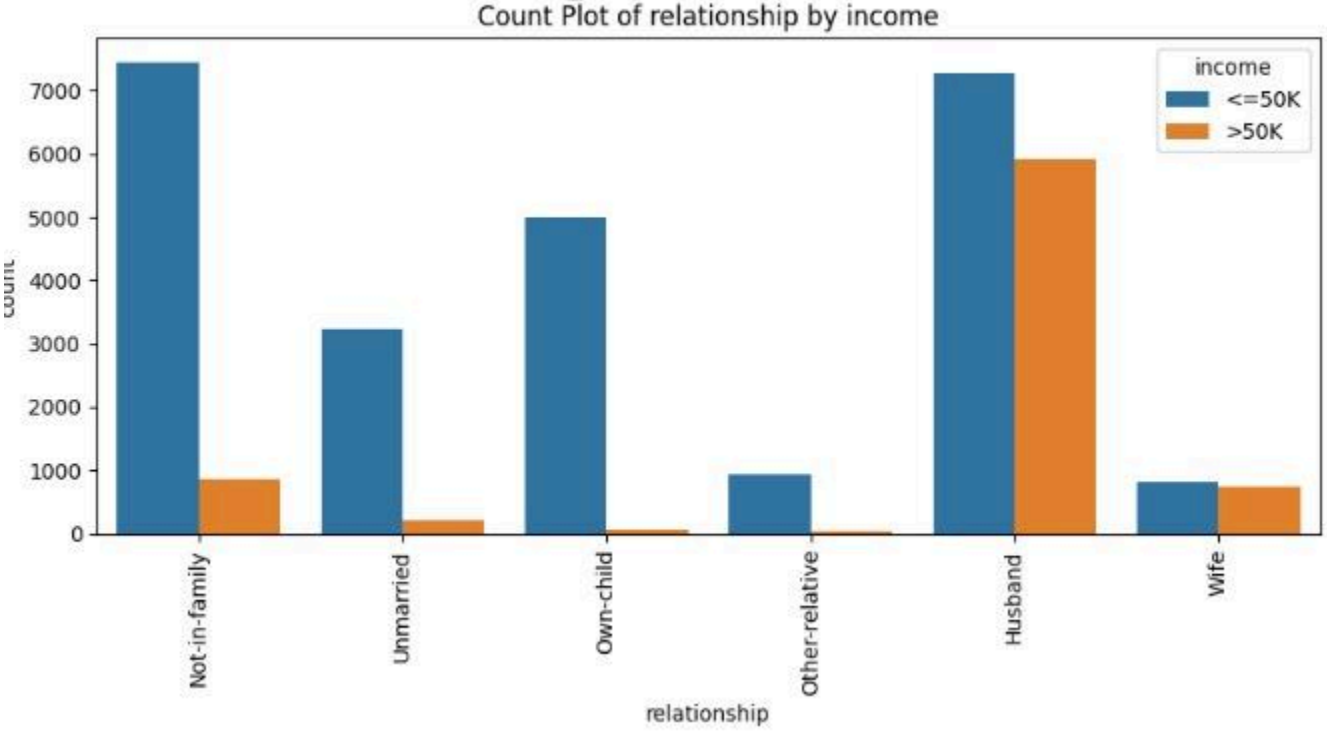
Categorical Column



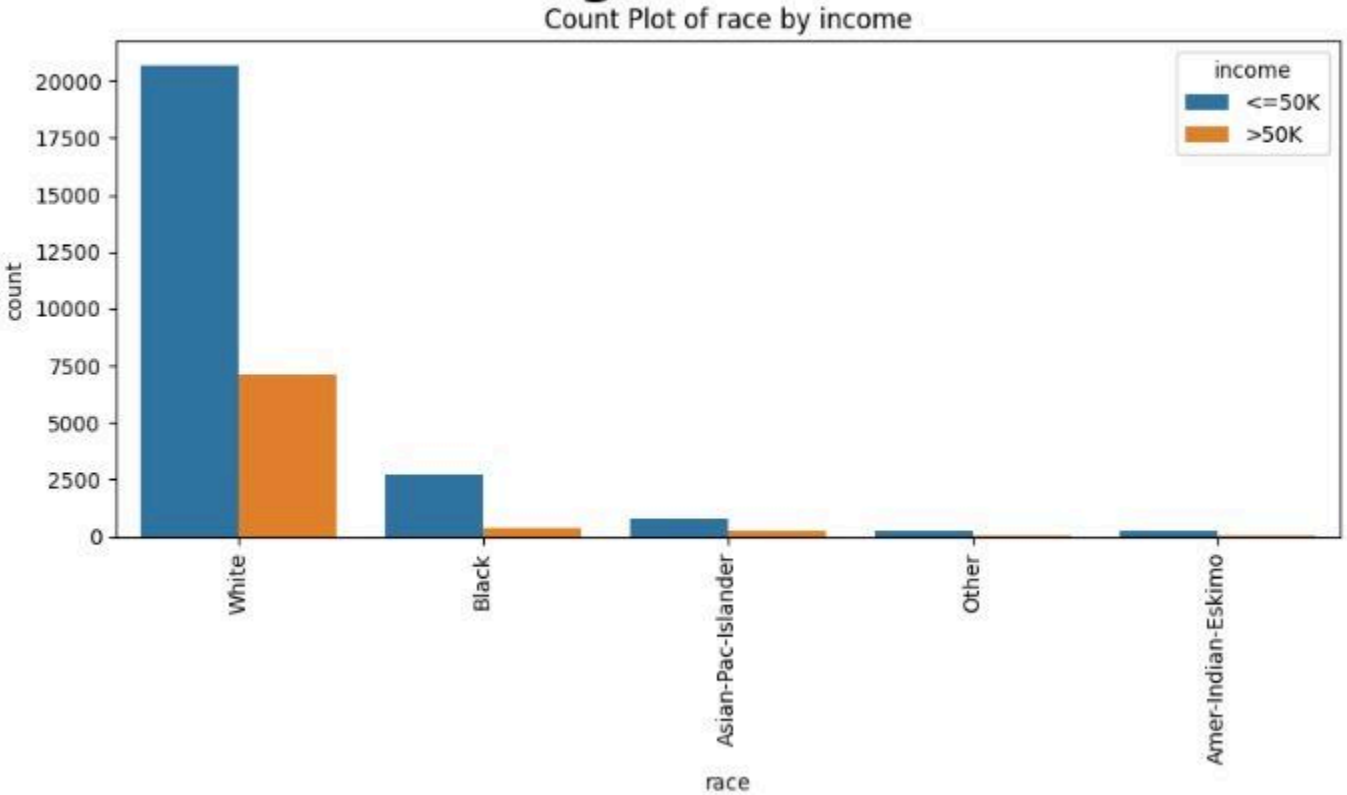
Categorical Column



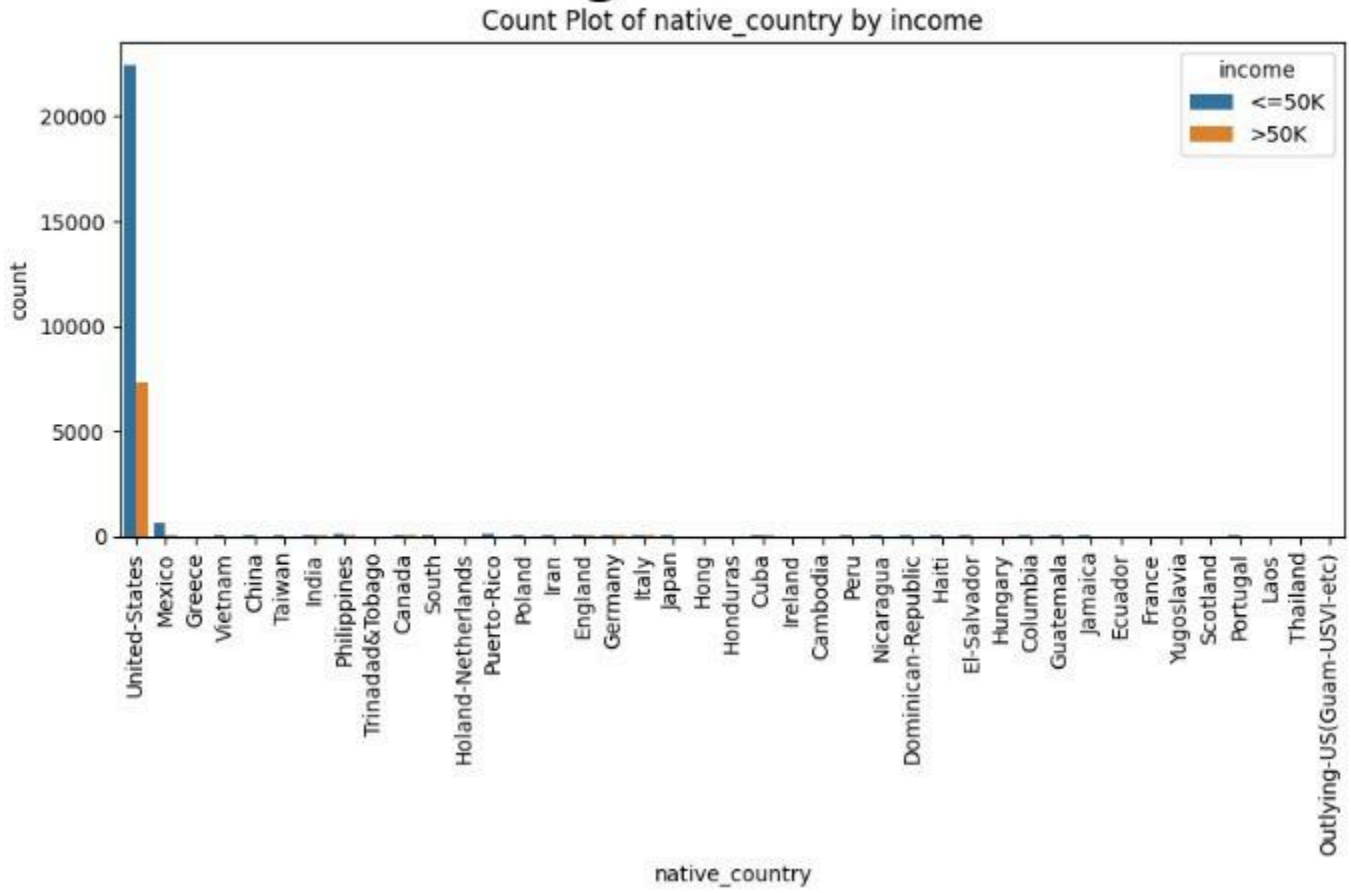
Categorical Column

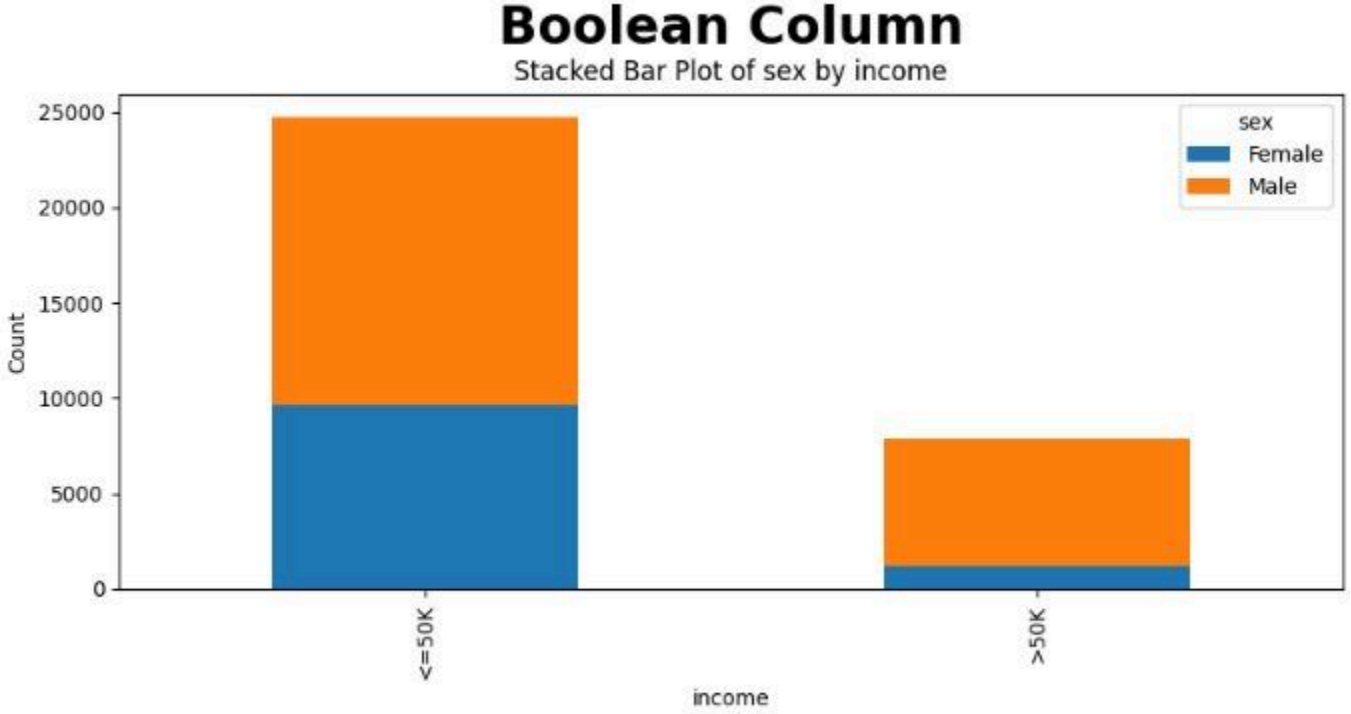


Categorical Column



Categorical Column




Boolean Column
Stacked Bar Plot of sex by income
sex
Female
Male
Count
25000
20000
15000
10000
5000
0
<=50K
>50K
income