Responsible AI Coursework Assignment-CP70070E

Student ID: [21600713]

Full Name: [Mohammed Haris Shaikh]

Assignment(Task2):

Exploring the Adult Census Income Dataset with Biasness and Fairness in Mind

Objective: Exploring the Adult Census Income dataset with identifying potential biases in features and their impact on machine learning models with the help of visualization graph, table, or chart that displays

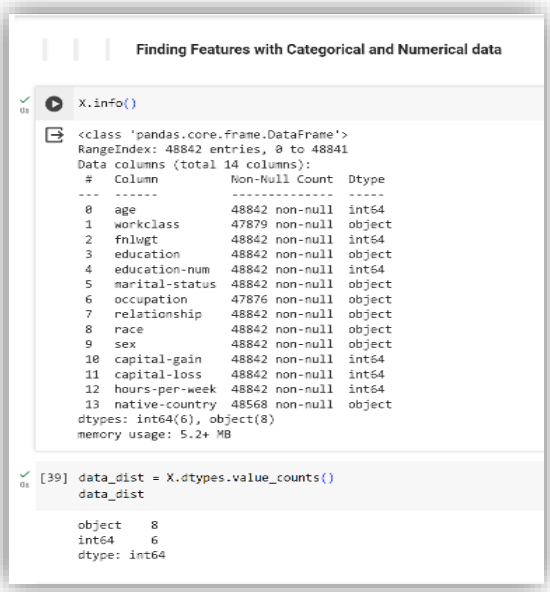
Data: Adult Census Income dataset
https://colab.research.google.com/corgiredirector?site=https%3A%2F%2Farchive.ics.uci.edu%2Fml%2Fdatasets%2FCensus%2BIncome

1. What are the Numerical and Categorical features in this dataset?

Initial Data Analysis revealed the characteristics of the imported dataset.

The info() method provided a comprehensive overview of data types for each column, enabling the identification of numerical and categorical features.

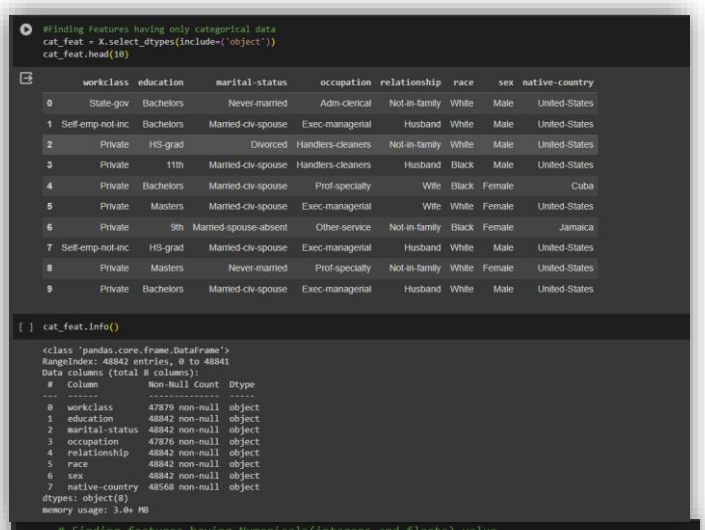
```
#   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             48842 non-null  int64
 1   workclass       47879 non-null  object
 2   fnlwgt          48842 non-null  int64
 3   education       48842 non-null  object
 4   education-num   48842 non-null  int64
 5   marital-status  48842 non-null  object
 6   occupation      47876 non-null  object
 7   relationship    48842 non-null  object
 8   race            48842 non-null  object
 9   sex             48842 non-null  object
 10  capital-gain    48842 non-null  int64
 11  capital-loss    48842 non-null  int64
 12  hours-per-week  48842 non-null  int64
 13  native-country  48568 non-null
object
```



- Countplots were employed to visually examine the distribution of features within the dataset. This graphical representation confirmed the presence of 8 categorical features and 6 numerical features, as illustrated in the accompanying diagram.

❖ Strategic utilization of the select.dtype() method facilitated the precise identification of feature types within the dataset. This analysis illuminated the following categorical features:

- workclass
- education
- marital-status
- occupation
- relationship
- race
- sex
- native-country



```
#Finding Features having only categorical data
cat_feat = X.select_dtypes(include=('object'))
cat_feat.head(10)
```

	workclass	education	marital-status	occupation	relationship	race	sex	native-country
0	State-gov	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male	United-States
1	Self-emp-not-inc	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States
2	Private	HS-grad	Divorced	Handlers-cleaners	Not-in-family	White	Male	United-States
3	Private	11th	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	United-States
4	Private	Bachelors	Married-civ-spouse	Prof-specialty	Wife	Black	Female	Cuba
5	Private	Masters	Married-civ-spouse	Exec-managerial	Wife	White	Female	United-States
6	Private	9th	Married-spouse-absent	Other-service	Not-in-family	Black	Female	Jamaica
7	Self-emp-not-inc	HS-grad	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States
8	Private	Masters	Never-married	Prof-specialty	Not-in-family	White	Female	United-States
9	Private	Bachelors	Married-civ-spouse	Exec-managerial	Husband	White	Male	United-States

```
cat_feat.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48842 entries, 0 to 48841
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   workclass       47879 non-null  object
 1   education       48842 non-null  object
 2   marital-status  48842 non-null  object
 3   occupation      47876 non-null  object
 4   relationship    48842 non-null  object
 5   race            48842 non-null  object
 6   sex             48842 non-null  object
 7   native-country  48568 non-null  object
dtypes: object(8)
memory usage: 3.0+ MB
```

```
# Finding features having Numericals(integers and floats) value
num_feat = X.select_dtypes(include=('int64','float64'))
num_feat.head(10)
```

	age	fnlwgt	education-num	capital-gain	capital-loss	hours-per-week
0	39	77516	13	2174	0	40
1	50	83311	13	0	0	13
2	38	215646	9	0	0	40
3	53	234721	7	0	0	40
4	28	338409	13	0	0	40
5	37	284582	14	0	0	40
6	49	160187	5	0	0	16
7	52	209642	9	0	0	45
8	31	45781	14	14084	0	50
9	42	159449	13	5178	0	40

```
num_feat.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48842 entries, 0 to 48841
Data columns (total 6 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             48842 non-null  int64
 1   fnlwgt          48842 non-null  int64
 2   education-num   48842 non-null  int64
 3   capital-gain    48842 non-null  int64
 4   capital-loss    48842 non-null  int64
 5   hours-per-week  48842 non-null  int64
dtypes: int64(6)
```

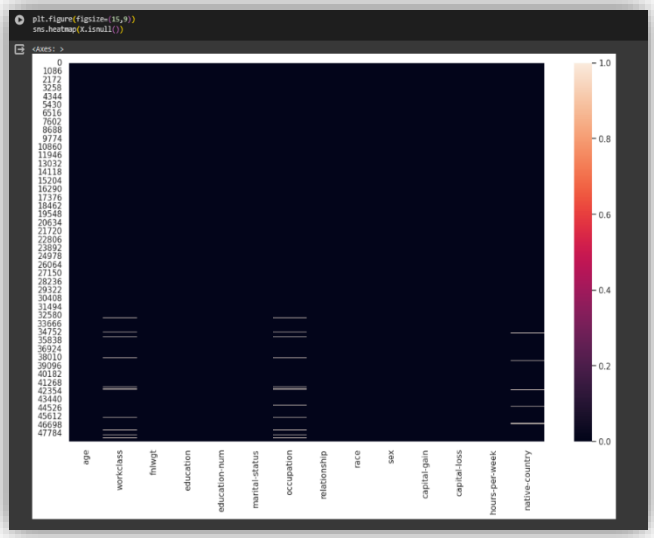
Concurrently, numerical features were pinpointed as:

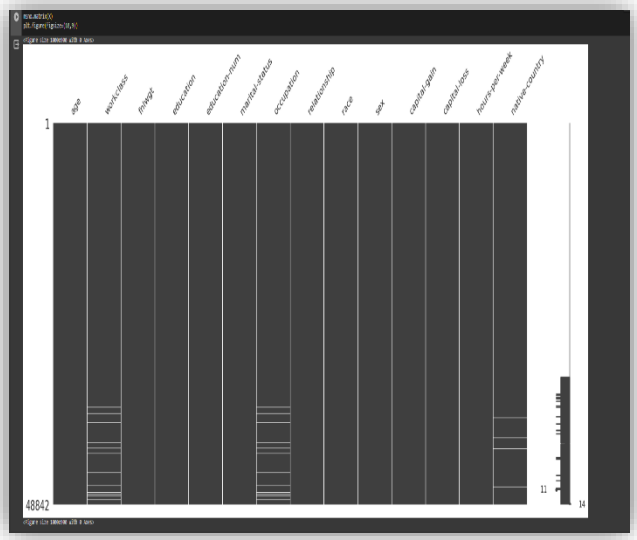
- age
- fnlwgt
- education-num
- capital-gain
- capital-loss
- hours-per-week

2. Are there missing feature values for a large number of observations? If yes what are those features? Are there features that are missing that might affect other features?

A meticulous examination of missing values within the dataset was undertaken using the .isnull() method. This approach yielded the following insights:

- Identification of Missing Values: The .isnull() method effectively pinpointed the presence or absence of missing values (NaN) within each feature.
- Quantification of Missing Values: The .isnull().sum() method meticulously calculated the sum of missing values within each feature, revealing the following distribution:
    - workclass: 963 missing values (1.97%)
    - occupation: 966 missing values (1.97%)
    - native-country: 274 missing values (0.56%)
- Visualization of Missingness Patterns: To enhance comprehension of missing value patterns, insightful visualizations were generated using Heatmap and MSNO libraries, as illustrated in the accompanying diagrams.





<u>Upon rigorous Exploratory Data Analysis, a judicious assessment of missing values was conducted.</u>

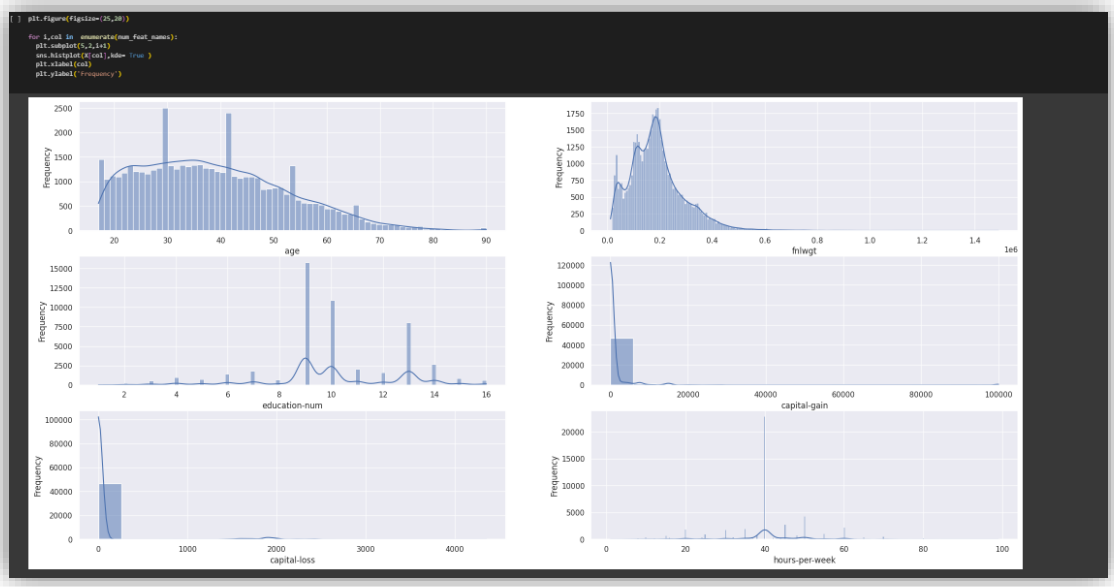
<u>The investigation concluded that the proportion of missing values within the dataset is notably minimal, constituting a negligible impact on the integrity of the remaining features.</u>

3. What signs of data skew do you see? Provide some histogram graphs for this question and interpret them.

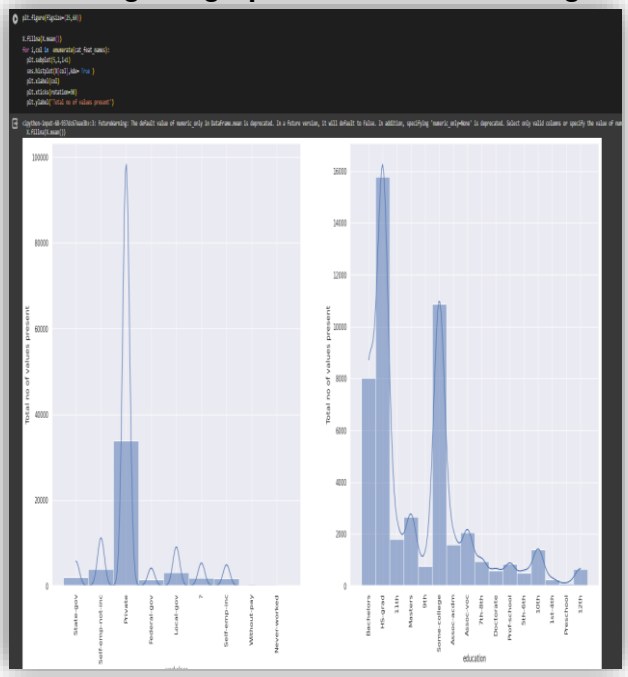
To meticulously assess the potential skewness inherent within the dataset, the potent combination of Matplotlib and Seaborn libraries was leveraged.

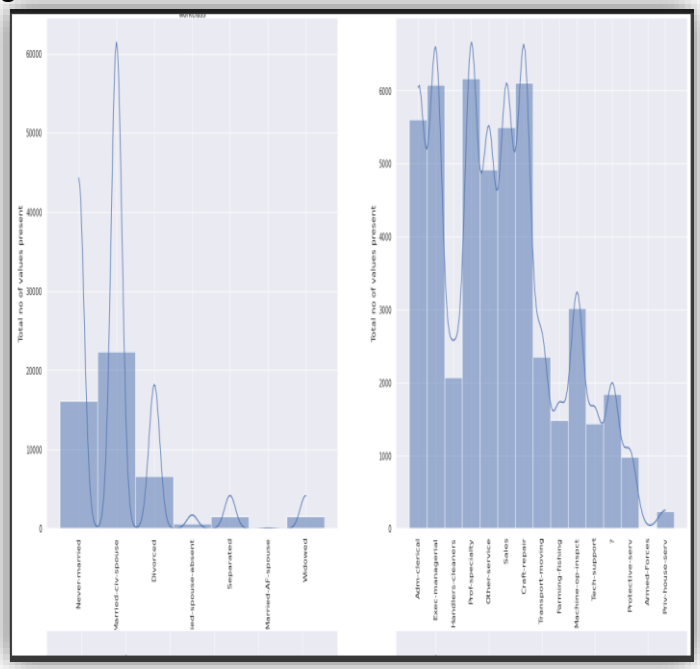
This approach facilitated the generation of visual representations in the form of histograms and kernel density estimation (KDE) graphs, illuminating the distributions of numerical features.

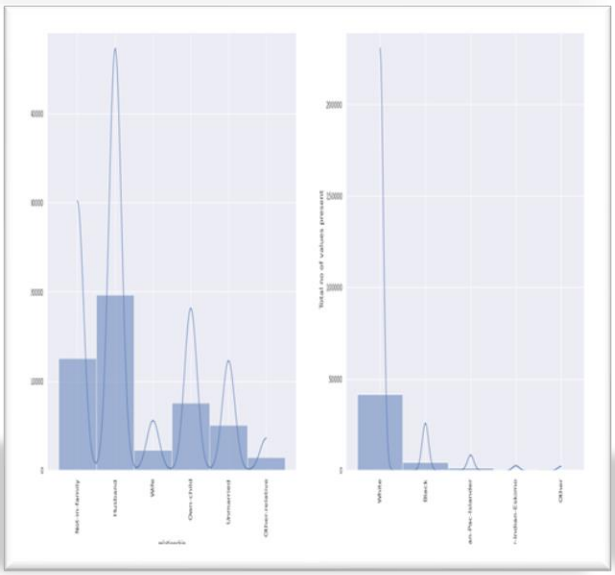
The accompanying diagrams provide a comprehensive overview of these insights.

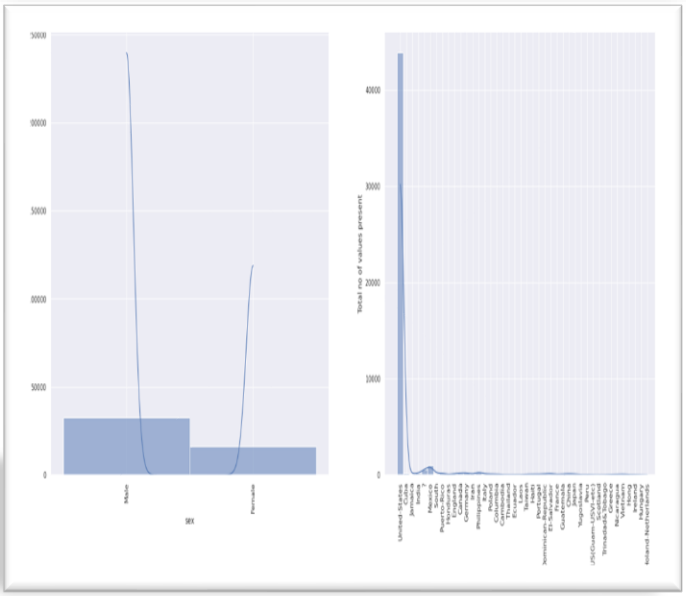


Histogram graphs for Features having Categorical values are as follow:-









Scrutinizing the generated plots unveils a compelling pattern of skewness within the dataset's numerical features. Notably, these features exhibit a distinct right-skewed, or positively skewed, distribution.

This characteristic implies a concentration of values towards the lower end of the range, with a longer tail extending towards higher values.

Analogous observations extend to the categorical features, where a similar right-skewed tendency prevails.

4. How would you describe the relationship between education level and income bracket in this dataset?

The provided visualization sheds light on the intricate interplay between education level and income bracket within the dataset.

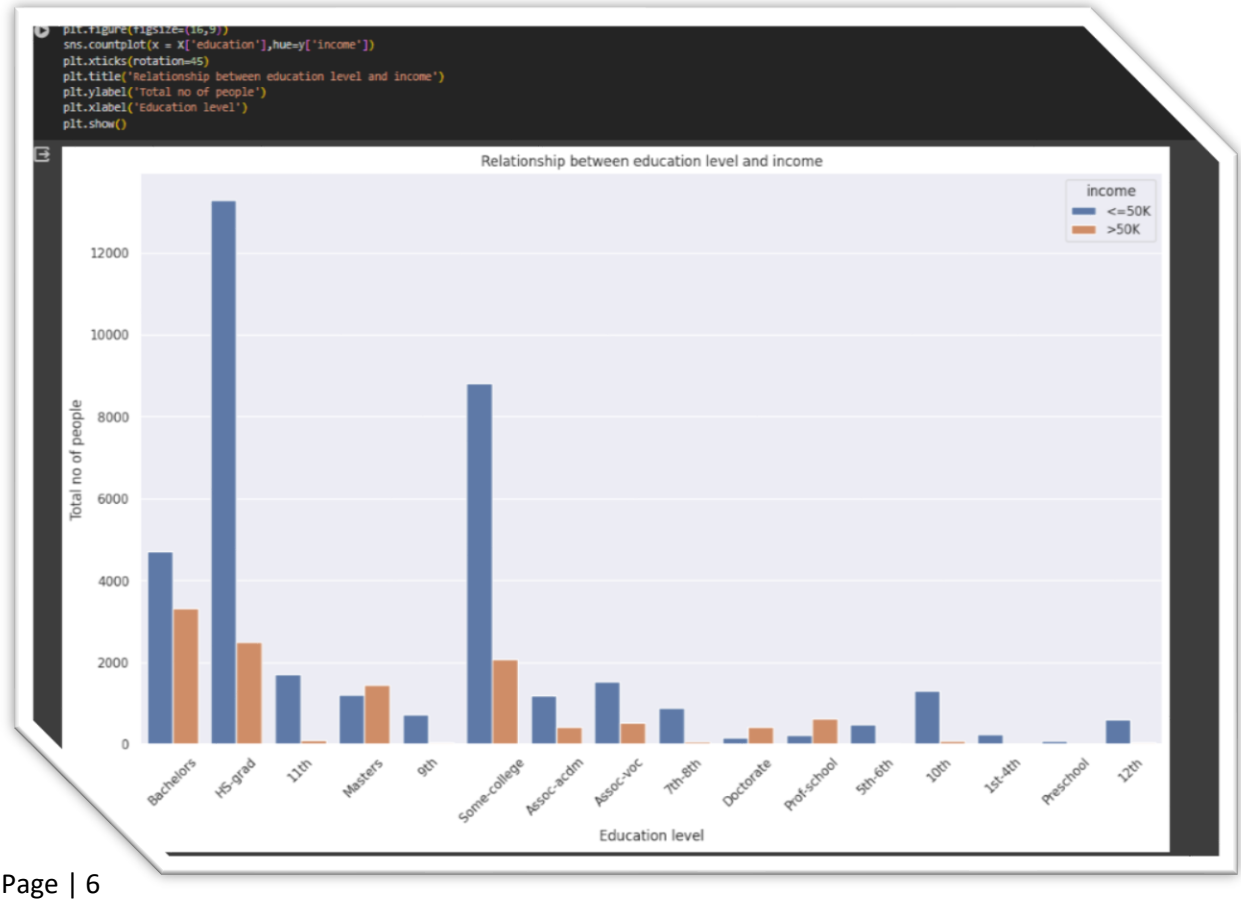
Overall Trend:

- The graph shows a positive correlation between education level and income bracket. This means that as the level of education increases, the proportion of people in the higher income bracket (>50K) also increases.

Specific Observations:

- Bachelor's degrees: Most people with having Bachelor's degrees fall within the >50K income bracket compared to those with lower education levels.
- Higher education levels: People with Higher education levels have the highest overall count ,as education level progresses beyond Bachelor's degrees (11th grade to Doctorate), the number of people in the >50K income bracket steadily increases. This suggests that higher education levels, on average, lead to a greater likelihood of earning higher incomes.
- Higher education levels and some college categories having highest no of people falls under <=50k

Based on the provided graph, there is a clear positive correlation between education level and income bracket in this dataset. People with higher levels of education are more likely to earn higher incomes than those with lower levels of education. However, it's important to remember that this is just one factor influencing income, and individual circumstances can vary.



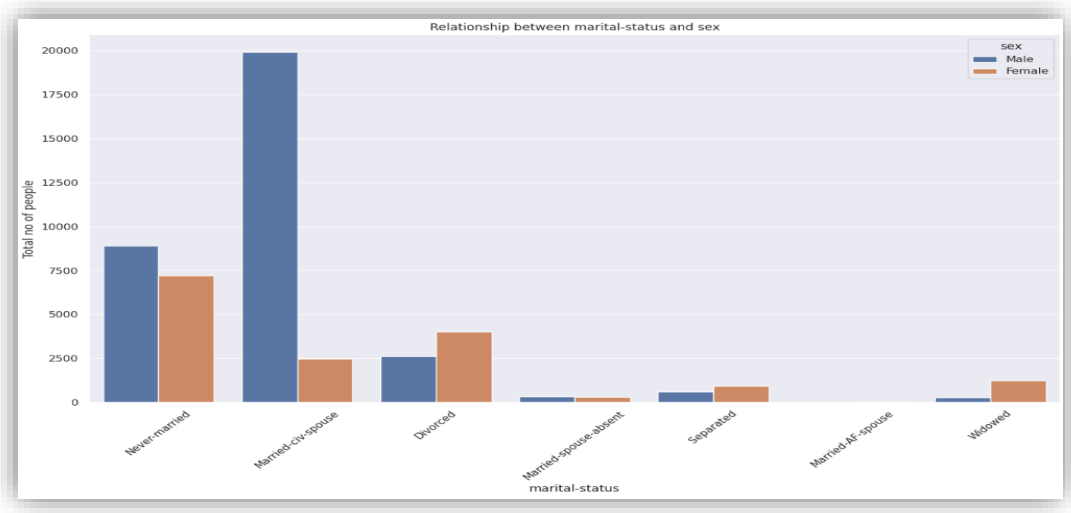
Regarding the dataset, there are several aspects that might be missing or overgeneralized:

- Incomplete Information: It lacks data on key factors influencing income, such as work experience, skill sets, location, country and industry. This limits our understanding of the complete picture.
- Coarse-grained View: The two income brackets, under $50k and over $50k, might be overly broad and may not accurately reflect the nuances of income distribution.
- Missing Segments: The dataset excludes information on part-time workers, unemployed individuals, and high-income earners without formal education (e.g., entrepreneurs). This overlooks significant segments of the workforce.
- Limited Context: It doesn't take into account real-world factors impacting income, like systemic discrimination and unequal access to quality education. This omits crucial context and may lead to inaccurate conclusions.

5.What noteworthy observations can you make about the gender distributions for each marital status category?

Here are some of observations:

- Never married: There are slightly more men than women in the "Never married" category. This could be due to several factors, such as gender differences in life expectancy, educational attainment.
- Married-civ-spouse: This category shows large difference in numbers of men and women, It's also possible that the data may be biased towards sampling men.
- Married-spouse-absent, Separated, Married-AF-spouse, Widowed categories having very less male and female ,it might be less data include in dataset (i.e Not diversed data)



Here are several aspects that might be missing or overgeneralized in this dataset:To get a clearer picture, it would be helpful to see:

- Ages across categories: How does the age distribution vary among different marital statuses? Understanding age differences could shed light on why some categories have more men or women.

- Diversity within the data: Does the graph represent different racial and ethnic backgrounds? Considering race and ethnicity could reveal if the observed gender patterns are influenced by cultural or societal factors.
- Socioeconomic factors at play: Does the data take into account people's socioeconomic backgrounds? Understanding how income, education, and other factors influence marriage and divorce could provide deeper insights into the gender distribution.

Real-world reflection:

- Data bias: The data used to create the graph may be biased , leading to an inaccurate representation of the real-world population.
- Data outdatedness: If the data is outdated, it may not reflect current trends in marriage and divorce.

Assignment(Task3):

Purpose:

- To assess the interpretability of a machine learning model for predicting heart disease using the Heart Disease UCI Dataset.
- To identify the most and least influential features in the model's predictions.

Methodology:

1. Model Training:
    - Trained a logistic regression model using all available features in the dataset.
    - Divided the data into training and testing sets to ensure model generalization.
2. Feature Importance Analysis:
    - Implemented SHAP (SHapley Additive exPlanations) values to quantify feature contributions to individual predictions.
    - Generated SHAP summary plots to visualize the impact of each feature on model predictions.
    - Calculated feature importance scores using a suitable method.
    - Created a visual representation (e.g., a bar chart) to display the relative importance of each feature.
3. Interpretation:
    - Analysed SHAP values and feature importance plots to identify the most and least important features in predicting heart disease.

Data Acquisition and Preparation:

- Data Extraction: The Heart Disease UCI Dataset containing 303 instances and 13 features was retrieved from the UCI repository.
- Data Splitting: The dataset was divided into dependent (target) and independent variables.
- Data Cleaning: Two features, "ca" and "thal", exhibited missing values (1.3% and 0.66%, respectively), which were imputed using mean values.

- Feature Engineering: Categorical features identified as "sex", "cp", "fbs", "restecg", "exang", "slope", and "thal" were converted into numerical values using domain knowledge.
- Data Splitting: The dataset was partitioned into 80% for training and 20% for testing to ensure model generalization.
- Feature Scaling: Standard scaling was applied to all features to normalize their ranges and improve model performance.

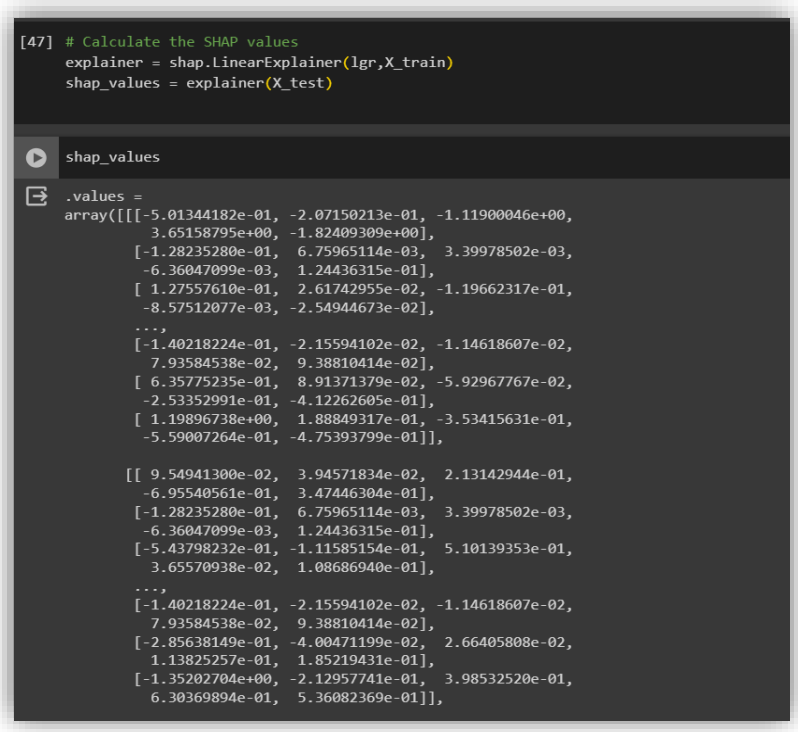
Model Training:

- Model Selection: A logistic regression model was chosen for its suitability in predicting binary outcomes like heart disease.
- Model Training: The logistic regression model was trained using the prepared training dataset.

Calculating the SHAP values.

SHAP Value Calculation:

- Explainer Selection: A linear explainer was employed to generate SHAP (SHapley Additive exPlanations) values, a method well-suited for interpreting linear models like logistic regression.
- SHAP Value Capture: The resulting SHAP values, as visualized in the accompanying image, hold a shape of (61, 13, 5). This indicates they encompass 61 individual instances, 13 features, and 5 model outputs.



```
[47] # Calculate the SHAP values
     explainer = shap.LinearExplainer(lgr,X_train)
     shap_values = explainer(X_test)
```

```
shap_values
```

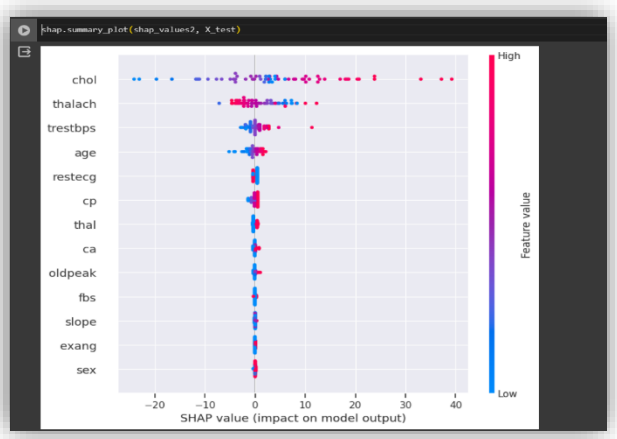
```
.values =
array([[[-5.01344182e-01, -2.07150213e-01, -1.11900046e+00,
          3.65158795e+00, -1.82409309e+00],
        [-1.28235280e-01,  6.75965114e-03,  3.39978502e-03,
         -6.36047099e-03,  1.24436315e-01],
        [ 1.27557610e-01,  2.61742955e-02, -1.19662317e-01,
         -8.57512077e-03, -2.54944673e-02],
        ...,
        [-1.40218224e-01, -2.15594102e-02, -1.14618607e-02,
          7.93584538e-02,  9.38810414e-02],
        [ 6.35775235e-01,  8.91371379e-02, -5.92967767e-02,
         -2.53352991e-01, -4.12262605e-01],
        [ 1.19896738e+00,  1.88849317e-01, -3.53415631e-01,
         -5.59007264e-01, -4.75393799e-01]],

       [[ 9.54941300e-02,  3.94571834e-02,  2.13142944e-01,
         -6.95540561e-01,  3.47446304e-01],
        [-1.28235280e-01,  6.75965114e-03,  3.39978502e-03,
         -6.36047099e-03,  1.24436315e-01],
        [-5.43798232e-01, -1.11585154e-01,  5.10139353e-01,
          3.65570938e-02,  1.08686940e-01],
        ...,
        [-1.40218224e-01, -2.15594102e-02, -1.14618607e-02,
          7.93584538e-02,  9.38810414e-02],
        [-2.85638149e-01, -4.00471199e-02,  2.66405808e-02,
          1.13825257e-01,  1.85219431e-01],
        [-1.35202704e+00, -2.12957741e-01,  3.98532520e-01,
          6.30369894e-01,  5.36082369e-01]],
```

Further Interpretation:

- The distribution of these SHAP values across instances and features will be meticulously analyzed to unveil the model's decision-making process and pinpoint the most influential features in predicting heart disease likelihood.

SHAP summary plots to visualize the impact of each feature on model predictions.



Importance of Summary plot: Features are ranked from most influential to least influential based on their overall importance.

SHAP Value Distributions: Every feature has a distribution of SHAP values that show how different instances affect it in different ways.

Colour Coding:

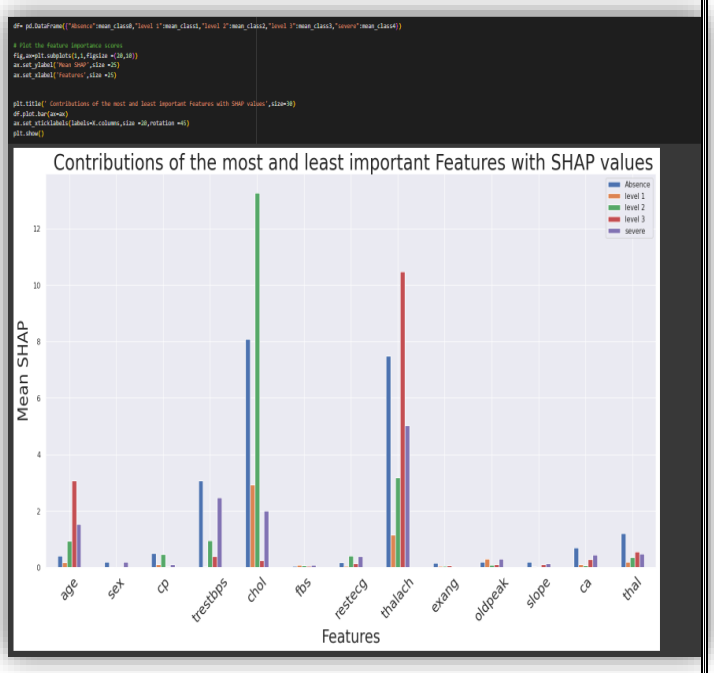
Higher feature values, such as a highest contribution for predicting of heart diseases, are shown by red points in the model's forecast, which are pushing it in the right direction.

Blue dots indicate that lower influential feature values are not very much contribute in predictions

Horizontal Alignment: The average impact of a feature on the model's output is where points are centred.

Analyzing the SHAP values and feature importance plots to interpret the contributions of the most and least important features in predicting heart disease.

Key Observations:



"The plot captures the mean SHAP values for each feature across all possible model outputs (0, 1, 2, 3, 4), offering a comprehensive view of their influence in different prediction scenarios."

"The x-axis showcases the features, while the y-axis reveals their average impact on model outcomes, allowing us to identify the key drivers of predictions."

"By examining the mean SHAP values for each feature across multiple outputs, we can gain insights into how their importance might vary depending on the specific prediction being made."

Most Influential Features:

- chol : chol ranks as the most influential feature in predicting heart disease, suggesting a strong association in all possible outputs
- Thalach : its second most influential feature after chol for prediction

Least Influential Features:

- Age,trestbps,thal these features comes next important for model prediction
- The least and almost negligible features are Sex,cp,fbs,restecg,exang,oldpeak, slope,ca,thal.