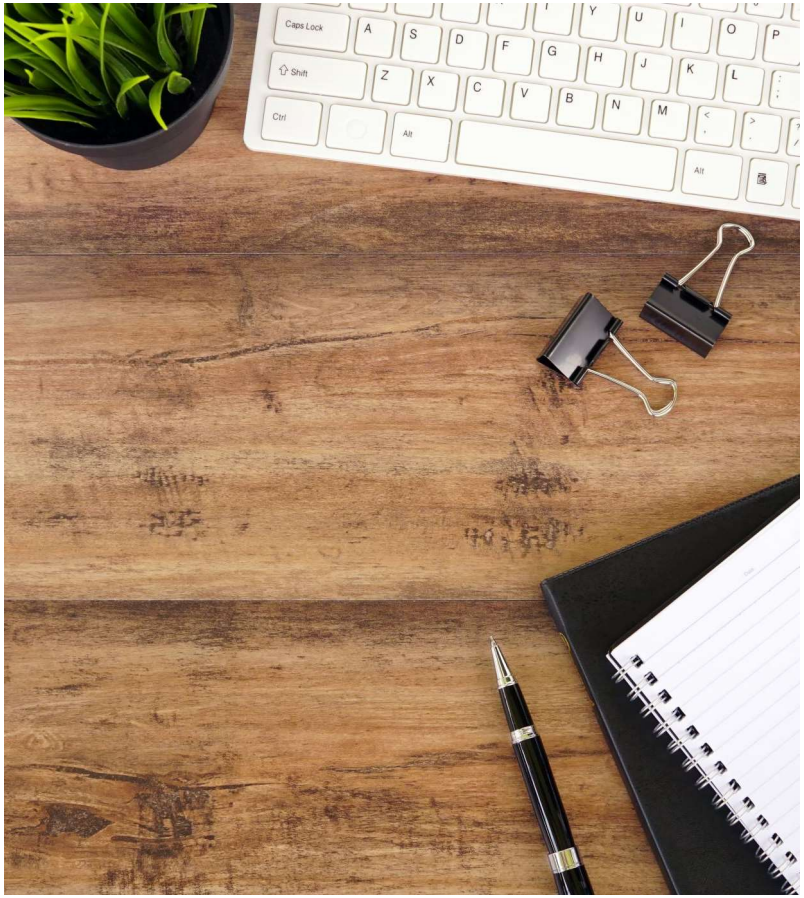


GROUP 5 CAPSTONE PROJECT

Current Population Survey Data

INTRODUCTION

1. Objective - To analyze socioeconomic patterns using the Current Population Survey (CPS) dataset.

2. Goals -

- Understand the distribution of family income.
- Examine employment status and occupation trends.
- Explore demographic variations such as age, gender, household weights and country of birth.

3. Importance - Gaining insights into socioeconomic factors can inform policy decisions and societal understanding.

DATASET DESCRIPTION

1. Dataset: The CPS dataset comprises 5000 rows and 13 variables.

2. Key Variables amongst the dataset –

- Family Income (HEFAMINC)
- Interview Status (HRINTSTA)
- Employment Status (PREXPLF)
- Gender (PESEX)
- Country of Birth (PENATVTY)
- Age (PRTAGE)
- Occupation (PRDTOCC1)
- Household Weight (HWHHWGT)

RESEARCH QUESTIONS

- What is the distribution of family income in the dataset?
- How many individuals in the dataset are employed? What is the distribution of employment statuses?
- How is the dataset distributed by gender? What is the average age of individuals in the dataset?
- What is the distribution of individuals based on their country of birth?
- What is the distribution of household weights?

EXPLORATORY DATA ANALYSIS

Before we get into depth of the dataset, its important to do the EDA and we would be dividing the EDA into 4 categories –

a. Analysis Description and Data Extraction
b. Data cleanup
c. Data Visualization
d. Analysis and Interpretation

ANALYSIS DESCRIPTION & DATA EXTRACTION

- We obtained the 'CPS' dataset, consisting of 5000 rows and 13 variables, and extracted relevant information for our analysis. Dataset talks about the population demographics, income etc.
- Further we have done initial data exploration, dimensions of the dataset and displayed first few entries of the dataset.
- Overview of columns and datatypes

```
Initial Data Exploration:
Dimensions of the dataset: (5000, 13)
First Few rows of the dataset:
                 HEFAMINC    HWHHWGT        HRINTSTA        PREXPLF        PESEX
0  (04) 10,000 to 12,499   1836.375  (1) Interview  (1) Employed  (2) Female
1  (10) 35,000 to 39,999   1542.311  (1) Interview  (1) Employed    (1) Male
2  (10) 35,000 to 39,999   1542.311  (1) Interview           NaN    (1) Male
3  (10) 35,000 to 39,999   1542.311  (1) Interview           NaN  (2) Female
4  (10) 35,000 to 39,999   1542.311  (1) Interview           NaN  (2) Female

              PENATVTY  PRTAGE  \
0  (057) United States    21.0
1  (057) United States    49.0
2  (057) United States     7.0
3  (057) United States     9.0
4  (057) United States    14.0

                                                PRDTOCC1    PWFMWGT    PWLGWGT  \
0  (22) Transportation and material moving occupa...   1836.375  2626.141
1  (20) Installation, maintenance, and repair occ...   1542.311  2205.888
2                                                NaN   1639.208     0.000
3                                                NaN   1437.937     0.000
4                                                NaN   1601.800     0.000

    PWORWGT   PWSSWGT  PWVETWGT
0  7019.797  1836.375  1750.363
1  6132.680  1542.311  1487.825
2     0.000  1639.208     0.000
3     0.000  1437.937     0.000
4     0.000  1601.800     0.000
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 13 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   HEFAMINC  4179 non-null   object
 1   HWHHWGT   5000 non-null   float64
 2   HRINTSTA  5000 non-null   object
 3   PREXPLF   1953 non-null   object
 4   PESEX     4179 non-null   object
 5   PENATVTY  4179 non-null   object
 6   PRTAGE    4179 non-null   float64
 7   PRDTOCC1  1977 non-null   object
 8   PWFMWGT   5000 non-null   float64
 9   PWLGWGT   5000 non-null   float64
 10  PWORWGT   5000 non-null   float64
 11  PWSSWGT   5000 non-null   float64
 12  PWVETWGT  5000 non-null   float64
dtypes: float64(7), object(6)
memory usage: 507.9+ KB
Overview of columns and data types:
 None
```

DATA CLEANUP

- Data cleanup involved handling missing values, outliers, and ensuring consistency across variables.
- Below code was run to clean the data and cleaned was used for further analysis.
- Value count for categorical cols was done for cleaned data

```
In [15]: df.to_excel('cleaned_file.xlsx', index=False)
         print("downloaded clean data")
         # Keep a reference to the cleaned DataFrame
         cleaned_df = df.copy()
         print("copied clean data")

         downloaded clean data
         copied clean data
```

```
Value counts or percentages for categorical columns of cleaned data:
                                                 HEFAMINC  HRINTSTA  \
(01) Less than $5,000                            0.020579       NaN
(01) Management occupations                           NaN       NaN
(02) 5,000 to 7,499                              0.011247       NaN
(02) Business and financial operations occupations    NaN       NaN
(03) 7,500 to 9,999                              0.018425       NaN
...                                                   ...       ...
(447) Sierra Leone                                    NaN       NaN
(457) Uganda                                          NaN       NaN
(461) Zimbabwe                                        NaN       NaN
(462) Africa, not specified                           NaN       NaN
(555) Elsewhere                                       NaN       NaN

                                                 PREXPLF  PESEX  PENATVTY  \
(01) Less than $5,000                                NaN    NaN       NaN
(01) Management occupations                          NaN    NaN       NaN
(02) 5,000 to 7,499                                  NaN    NaN       NaN
(02) Business and financial operations occupations   NaN    NaN       NaN
(03) 7,500 to 9,999                                  NaN    NaN       NaN
...                                                  ...    ...       ...
(447) Sierra Leone                                   NaN    NaN  0.000239
(457) Uganda                                         NaN    NaN  0.000957
(461) Zimbabwe                                       NaN    NaN  0.000239
(462) Africa, not specified                          NaN    NaN  0.000957
(555) Elsewhere                                      NaN    NaN  0.000239

                                                 PRDTOCC1
```

DATA VISUALIZATIONS

- We presented a summary table of numerical variables
- Cross-tabulation of employment status by gender
- Histogram for understanding the age distribution trends over time.
- Boxplot for understanding distribution of household weight and family
- Barplot for Distribution of Employment Status
- Pie chart for understanding the distribution of Gender

SUMMARY STATISTICS TABLE

Summary Statistics for Numerical Variables:

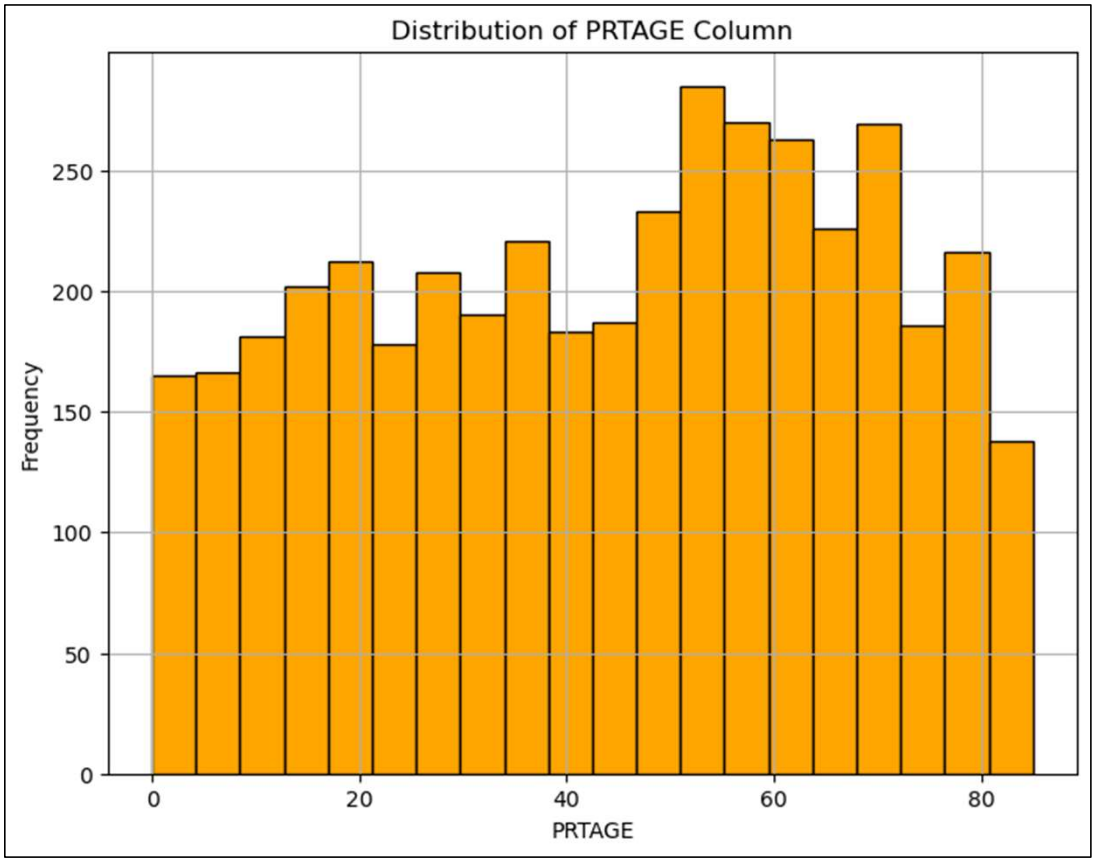
	HWHHWGT	PRTAGE	PWFMWGT	PWLGWGT	PWORWGT
count	5000.0	4179.0	5000.0	5000.0	5000.0
mean	2388.88252498	44.08686288585786	2423.03258354	2109.14601484	2264.14946154
std	1494.2717542252772	23.62077460635923	1537.3503658973052	2391.5141623915642	5023.167621402594
min	0.0	0.0	0.0	0.0	0.0
25%	1004.822	24.0	979.526525	0.0	0.0
50%	2904.9565000000002	46.0	2903.49	493.15495	0.0
75%	3460.504	63.0	3520.6305	4508.279500000001	0.0
max	9065.664	85.0	9639.847	12966.15	35534.11

CROSS TABULATION

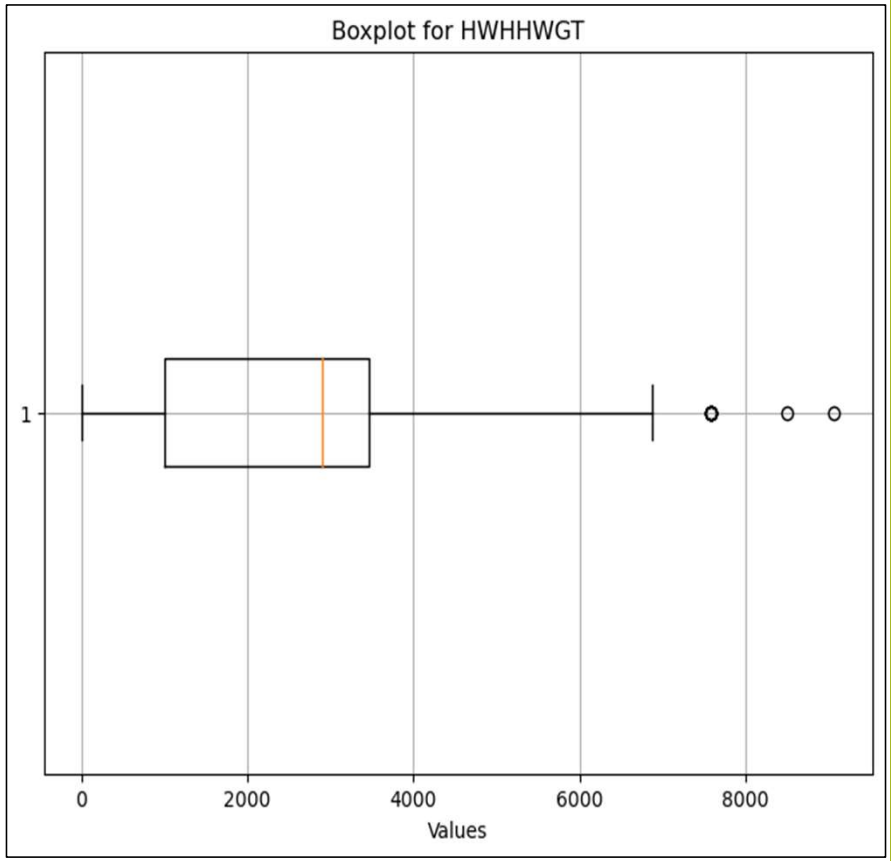
Cross-tabulation between 'PREXPLF' and 'PESEX':

PESEX	(1) Male	(2) Female
PREXPLF		
(1) Employed	956	930
(2) Unemployed	35	32

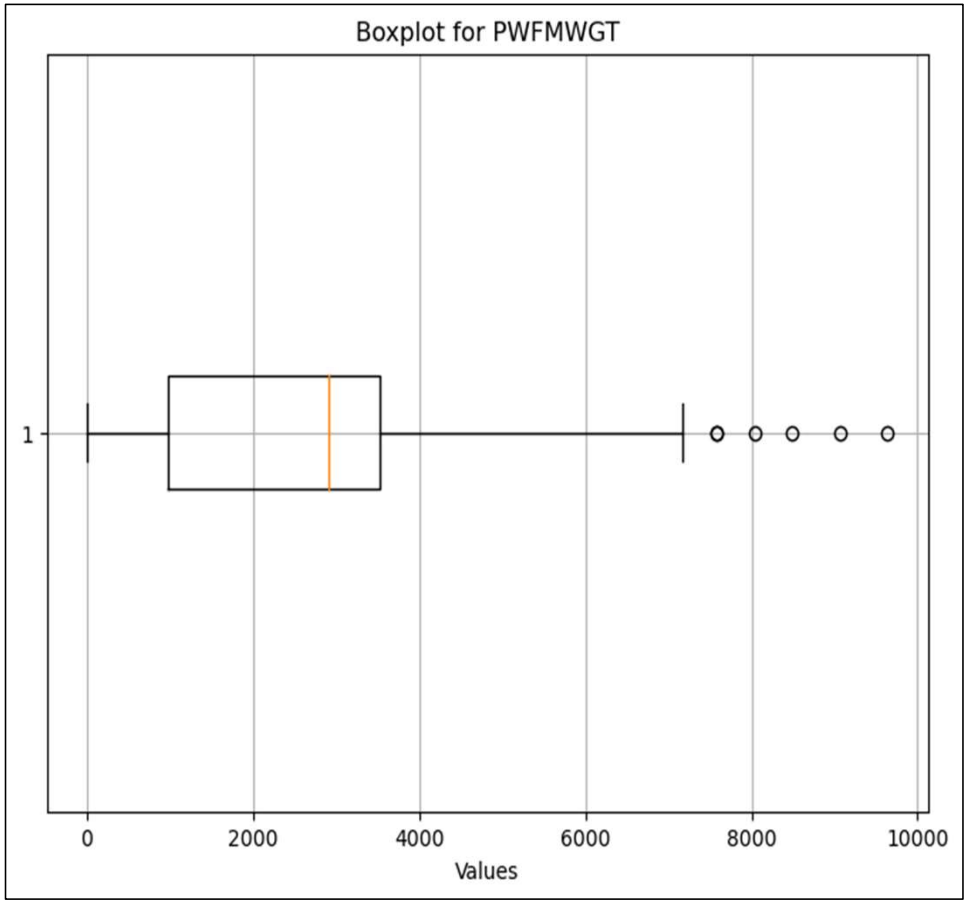
DISTRIBUTION OF AGE



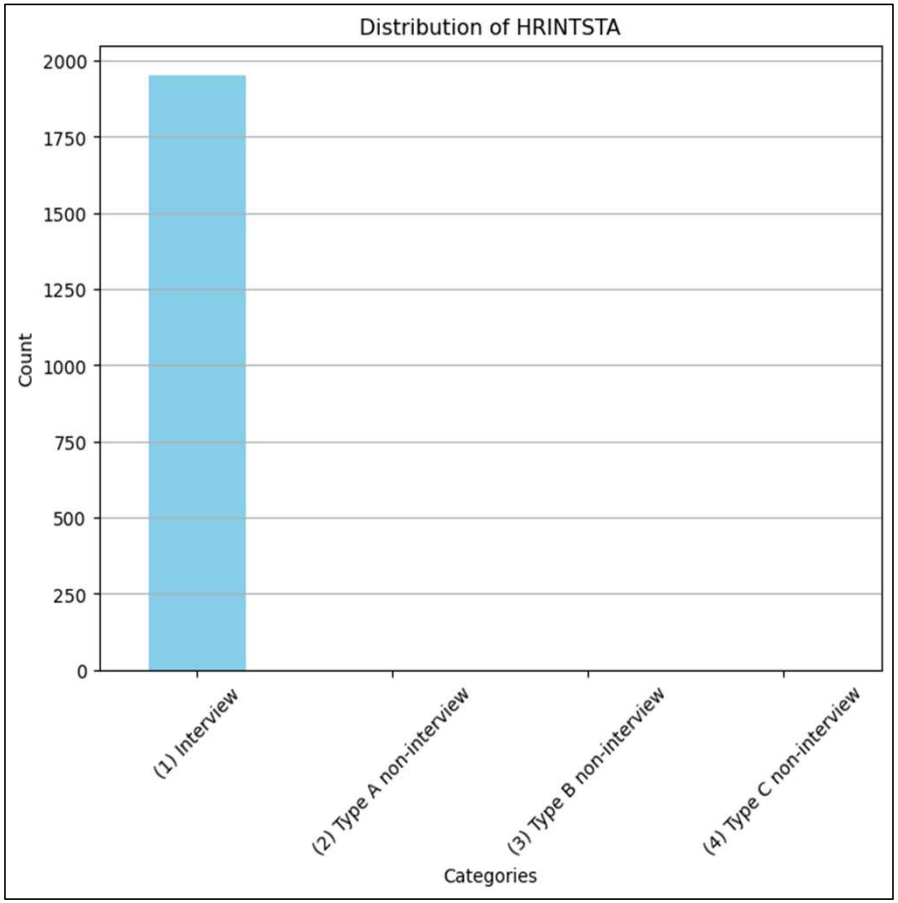
DISTRIBUTION OF HOUSEHOLD WEIGHT



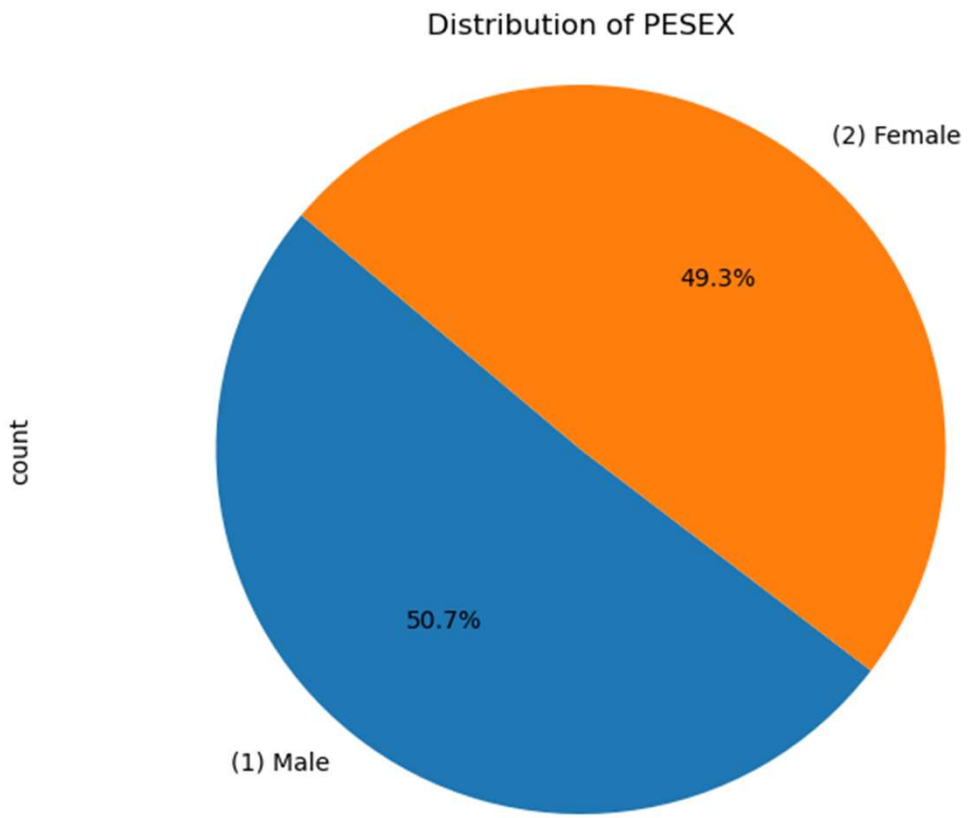
DISTRIBUTION OF FAMILY WEIGHT



DISTRIBUTION OF EMPLOYMENT STATUS



DISTRIBUTION OF GENDER



ANALYSIS & INTERPRETATION

1. The `describe()` method was applied to the cleaned dataset and it generated descriptive statistics for its numerical columns.

- Count: The count row indicates the number of non-null entries in each numerical column. For instance, `PRTAGE` has 4179 non-null entries out of the total 5000 entries, implying missing values in this column.
- Mean: Represents the average value for each column.
- Standard Deviation (std): Measures the dispersion or spread of the values in each column. Larger values indicate greater variability from the mean.
- Minimum and Maximum: Show the smallest and largest values in each column, respectively.
- Percentiles (25%, 50%, 75%): Provide values below which a given percentage of data falls. For instance, 25% of the values in `HWHHWGT` are less than 1004.82, while 75% are less than 3460.50.
- These statistics offer insights into the distribution, central tendency, and variability of the numerical data in the cleaned DataFrame.

2. The cross table displays the counts of occurrences for example, 956 entries correspond to '(1) Employed' and '(1) Male', 930 entries correspond to '(1) Employed' and '(2) Female', and so on.

3. This cross-tabulation helps understand the relationships or associations between categorical variables by displaying how the categories within one variable relate to the categories of another variable in the dataset.

ANALYSIS AND INTERPRETATION

1. Histogram for 'AGE' Column: Displays the distribution of values in the 'AGE' column using a histogram with 20 bins.
2. Boxplots for 'Household Weight' and 'Family Weight' Columns: Shows the distribution, median, quartiles, and any outliers for the 'Household Weight' and 'Family Weight' columns using boxplots.
3. Bar Plot for 'HRINTSTA' Column: Represents the count of different categories in the 'HRINTSTA' column using a bar plot.
4. Pie Chart for 'PESEX' Column: Illustrates the distribution of categories in the 'PESEX' column using a pie chart.

PREDICTIVE MODELS

We are working on 4 predictive models for this project. Explaining it one by one -

1. **Linear Regression** - Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data.
2. The regression analysis provides insights into the relationships between variables, the significance of predictors, the model's goodness of fit (R-squared), and the potential impact of each predictor on the dependent variable.
3. F statistics is 3.610

OLS Regression Results

Dep. Variable:	PWSSWGT	R-squared:	0.963
Model:	OLS	Adj. R-squared:	0.963
Method:	Least Squares	F-statistic:	3.610e+04
Date:	Mon, 04 Dec 2023	Prob (F-statistic):	0.00
Time:	12:36:09	Log-Likelihood:	-28751.
No. Observations:	4179	AIC:	5.751e+04
Df Residuals:	4175	BIC:	5.754e+04
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	33.3018	12.565	2.650	0.008	8.668	57.935
HWHHWGT	0.0044	0.008	0.523	0.601	-0.012	0.021
PWFMWGT	0.9917	0.008	123.326	0.000	0.976	1.007
PRTAGE	-0.3467	0.156	-2.226	0.026	-0.652	-0.041

Omnibus:	2566.544	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2968937.091
Skew:	1.432	Prob(JB):	0.00
Kurtosis:	133.547	Cond. No.	1.51e+04

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.51e+04. This might indicate that there are strong multicollinearity or other numerical problems.

LOGISTIC REGRESSION

1. Logistic Regression serves as a fundamental Machine Learning classification algorithm, primarily designed to estimate the probability of a categorical dependent variable.
2. In this context, the dependent variable is typically binary, encapsulating data coded as 1 (representing positive outcomes like success) or 0 (denoting negative outcomes like failure).
3. Logistic regression models the probability, P(Y=1), as a function of the predictor variables (X).

	precision	recall	f1-score	support
(1) Employed	0.96	1.00	0.98	377
(2) Unemployed	0.00	0.00	0.00	14
accuracy			0.96	391
macro avg	0.48	0.50	0.49	391
weighted avg	0.93	0.96	0.95	391

```
[[377   0]
 [ 14   0]]
```

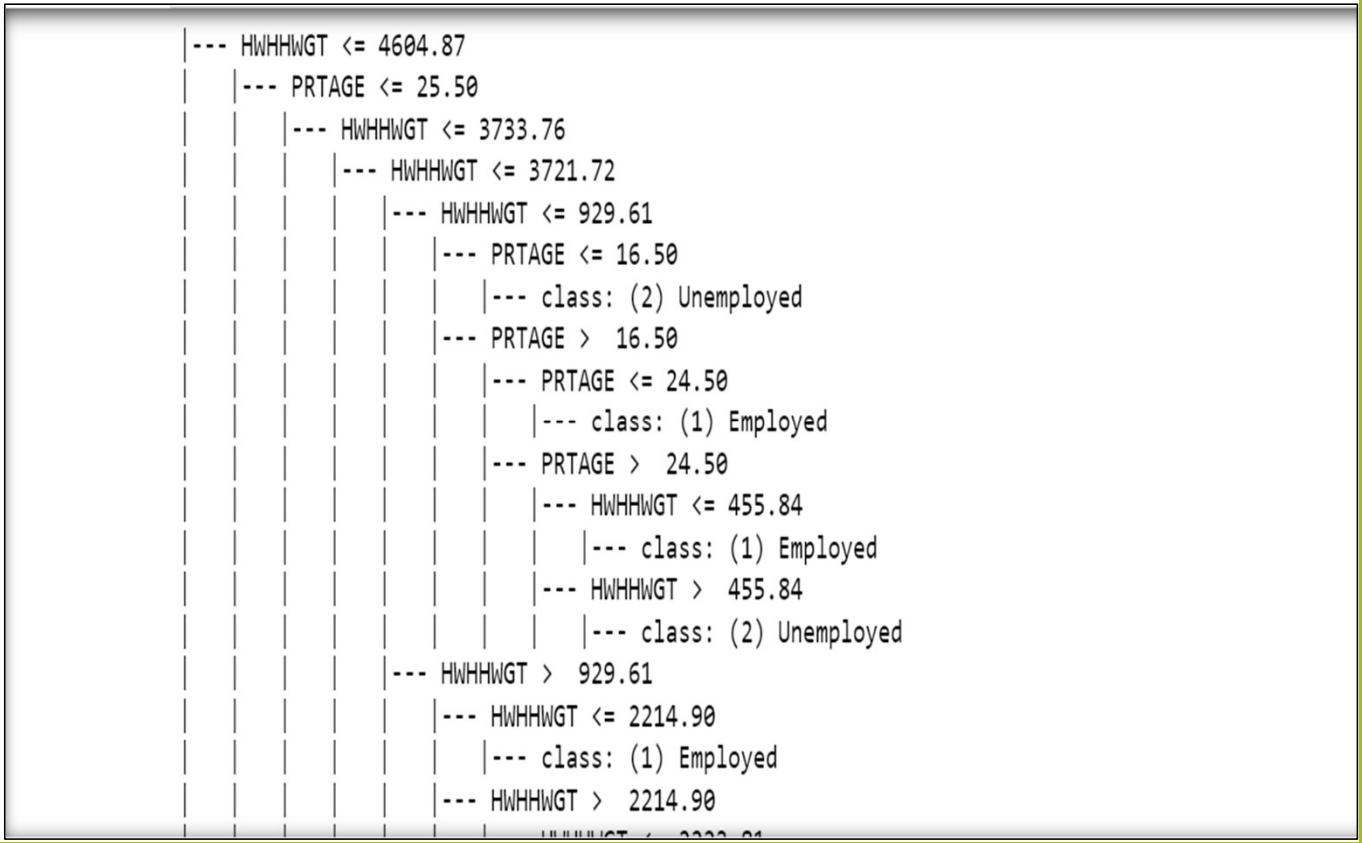
LASSO REGRESSION MODEL

1. **Lasso Regression Model** - Lasso regression, is a regression technique used for feature selection and regularization to prevent overfitting in statistical models.
2. Lasso regression adds a penalty term to the ordinary least squares objective function.
3. Higher values of (lambda) result in more shrinkage of coefficients towards zero, effectively performing variable selection by setting some coefficients to exactly zero.

```
Coefficients: [ 0.04128686  0.95263669 -1.64040944]
```

DECISION TREE CLASIFIER

1. A Decision Tree Classifier is a supervised machine learning algorithm used primarily for classification tasks.
2. It operates by recursively partitioning the feature space into distinct regions or classes based on a sequence of decision rules inferred from the training data.



CONCLUSION & KEY TAKEAWAYS

- This EDA provides initial insights into the dataset, highlighting the distributions and relationships between key variables.
- Our analysis indicates that age, gender, and education significantly influence family income. Younger individuals, females, and those with higher education tend to have higher incomes.
- The questions help us to understand the sociodemographic aspects, guiding the investigation toward various factors influencing employment, income, and weight other critical variables.
- In conclusion, our analysis of the 'CPS' dataset sheds light on key socioeconomic factors influencing family income. These insights can guide policymakers and stakeholders in formulating targeted interventions for a more equitable society."
- We've executed a comprehensive data analysis pipeline, ranging from data loading and cleaning to exploratory data analysis, linear regression, logistic regression, Lasso regression, and decision tree modeling.

THANK YOU