Adult Census Income Analytical Dashboard

Flask-Dash-Interactive-Dashboard

ABSTARCT

The objective of this project is the data handling, data cleaning, data preprocessing, and EDA on a dashboard based on Flask framework. We choose the dataset of adult income which gives us information about a population sample from United States with salary/income less than or greater than 50000. We chose the data set from UCI machine learning repository. After selecting the data, we performed the data cleaning, preprocessing, and visualization to analyze the determinants of salary classification.

TABLE OF CONTENTS

1 System Study / Domain Analysis 4

1.1 Domain Analysis 4

1.1.1 Business Process related to Dataset 5

1.2 Describing the Dataset 6

1.2.1 Classification and Regression 6

1.2.2 Balanced or Imbalanced Dataset? 6

1.2.3 Data Composition 7

2 Data Cleaning 9

3 Exploratory Data Analysis 10

3.1 Univariate Analysis: 11

3.2 Bivariate Analysis 17

4 Data Preprocessing 22

4.1 Discretization 22

4.2 Normalization 22

4.3 Label Encoding (Feature Encoding) 23

4.4 Train-Test Split 23

5 Dash 24

1 SYSTEM STUDY / DOMAIN ANALYSIS

1.1 DOMAIN ANALYSIS

1.1.1 Business Process related to Dataset

The Adult Income dataset is related to demographic and employment information of a united states population sample, with features such as age, workclass, education, marital status, occupation, and more. To describe the business process related to this dataset, we can infer that it might be associated with human resources, employment, or social research. The dataset appears to capture information about individuals, their education, work details, and income levels.

There are many different types of business processes that can be derived from the analysis on this particular dataset like 'Employee Recruitment and Retention' in which analyzing the distribution of education levels, work experience, and other demographic factors can help in understanding the characteristics of the workforce. Such a business process and its associated analytics can be useful for HR departments to tailor recruitment strategies, identify skill gaps, and enhance employee retention programs.

Similarly, the process can also be used for diversity and inclusion purposes by examining the distribution of race, gender, and other demographic factors which can provide insights into the diversity of the workforce.

Occupational Analysis can also be done by studying the distribution of occupations and their corresponding income levels which can provide insights into the organization's skill set and structure. This could help in optimizing job roles, identifying skill gaps, and ensuring the right talent is in the right positions.

However, for this project we have decided on a business process that does 'salary analysis'.
By exploring the distribution of income levels, capital gains, and losses we can provide insights into salary structures and financial well-being which could in turn help in setting competitive salary structures and financial wellbeing of employees. This process is crucial for ensuring that the company's salary offerings are competitive, fair, and aligned with both industry standards and the organization's budget constraints.

Step by Step Process:

The salary analysis and budgeting business process involves examining and managing the compensation structure within an organization. This process is crucial for ensuring that the company's salary offerings are competitive, fair, and aligned with both industry standards and the organization's budget constraints. Here's a detailed breakdown of the steps involved in the salary analysis and budgeting process:

1. Data Collection:

Objective: Gather relevant data on employee salaries and related factors.
Data Sources: Human Resources Information System (HRIS), payroll data, employee records.
Details: Collect information on individual salaries, bonuses, benefits, and any other monetary compensation components.

2. Demographic Analysis:

Objective: Understand the distribution of salaries across different employee demographics.
Analytics: Analyze salary data by factors such as job role, education level, experience, gender, and race.
Details: Identify any disparities or inequities in compensation and assess the diversity and inclusion aspects of salary distribution.

3. Benchmarking:

Objective: Compare the organization's salary structure with industry benchmarks.
Analytics: Utilize external salary surveys and market research to understand compensation trends in the industry.
Details: Assess how the organization's salaries compare to similar roles in the market to ensure competitiveness.

4. Financial Assessment:

Objective: Evaluate the financial feasibility of current salary levels and potential adjustments.
Analytics: Analyze the organization's budget constraints and financial health to determine the affordability of salary changes.
Details: Consider factors such as revenue, profit margins, and overall financial goals when planning salary adjustments.

5. Performance Evaluation:

Objective: Link salary decisions to employee performance.
Analytics: Incorporate performance data, such as employee reviews and key performance indicators (KPIs), to determine merit-based salary adjustments.
Details: Reward high performers with competitive salary increases and align compensation with individual contributions.

6. Communication and Transparency:

Objective: Communicate salary decisions transparently to employees.
Details: Clearly communicate any changes to the salary structure, criteria for salary adjustments, and the rationale behind these decisions.
Benefits: Fosters trust among employees and promotes a positive work environment.

7. Continuous Monitoring and Adjustment:

Objective: Regularly review and adjust the salary structure as needed.
Details: Stay informed about market trends, organizational changes, and employee feedback to make timely adjustments.
Benefits: Ensures that the organization's salary offerings remain competitive and aligned with its goals.

By following these steps, organizations can effectively manage their salary analysis and budgeting processes, ensuring that they attract and retain top talent while maintaining financial sustainability and legal compliance.

1.2 Describing the Dataset

1.2.1 Classification and Regression

The adult income dataset is a Classification dataset. Classification is a type of supervised learning where the goal is to categorize or classify instances into predefined classes or labels. In this dataset, the task is to be analyze the income levels of individuals based on various features such as age, workclass, education, marital status, occupation, and more.

The target variable for classification is to be "income," which has two classes: ">50K" (indicating an income level greater than $50,000) and "<=50K" (indicating an income level less than or equal to $50,000). This binary classification problem aims to predict whether an individual's income falls above or below the $50,000 threshold.

However, in this project we would only be performing data cleaning, preprocessing, exploratory data analysis and visualization. The predictive part of the data science process is out of scope of this course.

1.2.2 Balanced or Imbalanced Dataset?

The data we chose is imbalanced as the number of instances in both classes is not equal. The classes in the target variable are ">50K" (indicating an income level greater than $50,000) and "<=50K" (indicating an income level less than or equal to $50,000).

The possible problems associated with having an imbalanced dataset could be 'Bias in Model Performance' and 'Poor Generalization'.

Machine learning models trained on imbalanced datasets may exhibit bias towards the majority class. The model might be more inclined to predict the majority class, resulting in a high accuracy that is misleading.

Imbalanced datasets can lead to models that do not generalize well to new, unseen data, especially for the minority class. The model might struggle to make accurate predictions for instances in the underrepresented class.

These are just a few examples of the problems faced in using an imbalanced dataset.

1.2.3 Data Composition

The travel dataset comprises of 48842 rows and 15 columns. The dataset revolves around the income status of adults in United States categorized by the attributes mentioned in the table below.

The attribute names and data types are as below:

Attribute Names	Numerical or Categorical	Data Types
age	Numerical	int64
workclass	Categorical	object
fnlwgt	Numerical	int64
education	Categorical	object
educational-num	Numerical	int64
marital-status	Categorical	object
occupation	Categorical	object
relationship	Categorical	object
race	Categorical	object
gender	Categorical	object
capital-gain	Numerical	int64
capital-loss	Numerical	int64
hours-per-week	Numerical	int64
native-country	Categorical	object
income	Categorical	object

Around 7.4 % of the total records have missing values with most of them being in workclass and occupation field.

Feature importance:

Before delving into the detailed exploratory data analysis, it seems that age, workclass and education seem to be the features having the highest impact on income level being less than or greater than 50000.

2 DATA CLEANING

The dataset includes values that are incorrect, duplicate or erroneous. For better interpretation and analysis of data, the data scientist ensures data is cleanedand handled properly such that it is ready for the next phase of processing and can train a model.

Using the data.describe() command, the statistical description of entire dataset is displayed. Through scanning of the statistical values, we can see the maximum and minimum values and other useful statistical measures such as mean,hence identifying any wrong or incorrect values.

Data Type Conversion:

Using data.dtypes() command, all the data types for the dataset are checked. Some features/fields are not in suitable data types so they are converted into proper data type. Some of the conversions are as follows:

- Appropriate data type for educational-num is integer as it represents the education ranking based on the categorical order described in education variable.
- Appropriate data type for marital-status is object and there's no categorical order.
- Appropriate data type for occupation is object and there's no categorical order.
- Appropriate data type for relationship is object and there's no categorical order.
- Appropriate data type for race is object and there's no categorical order.
- Appropriate data type for gender is object and there's no categorical order.
- Appropriate data type for capital-gain is float since wealth gained is not necessarily a discrete sum.

Handling of the missing values in categorical fields of the dataset:

The adult income dataset does not contain null values but instead contains '?' so first we convert all the data with '?' to null values.

Now using the data.isnull().sum() command, the dataset is checked for any null or missing values.

Our strategy was to drop the missing values as after superficial analysis of the data it was seen that Workclass and occupation has simultaneous missing values so imputing with mode might lead to erroneous results as modal value for each variable would be different and the values of these 2 variables are related to each other for example protective service is part of local government etc. and so imputation without this consideration would lead to incorrect data. Native-country has very few values missing so it's safe to remove them instead of imputation which might lead to unintended biasness.

Looking at missing value records, it seems they're missing at random (no specific pattern found) and percentage of missing values in relation to other attribute classes follow the same pattern as class distribution within a variable and individual column missing value are 5% so it's safe to remove them given these two reasons as well.

Total 7% data missing out of 48k records which is relatively low so again it's safe to drop them.

Objective is to EDA (Exploratory Data Analysis) and not applying some ML algorithm so prediction accuracy and efficiency is not a concern but meaningful data insights are of concern and imputation might lead to incorrect findings so again it's better to drop missing values.

Duplicate values removal:

Duplicate were removed from the dataset.

Deletion of redundant data:

The column/field ‘fnlwgt’ was removed from the dataset as it appears to be some sort of sampling weight (number of people in population represented by a particular sample or record) which is irrelevant to when analyzing a subset of large dataset and for the purpose of EDA.

3 EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) stands as a pivotal data analysis approach, employing visual methods to distill key characteristics of a dataset. It is a systematic process of uncovering essential data features, utilizing visualization techniques to glean insights.

Exploratory Data Analysis is majorly performed using the following methods:

- Uni-variate analysis

Provides summary statistics for each field in the raw data set (or) summary only on one variable.

Ex:- histogram, Box plot, Pie charts, Horizontal and vertical bars

- Bivariate analysis

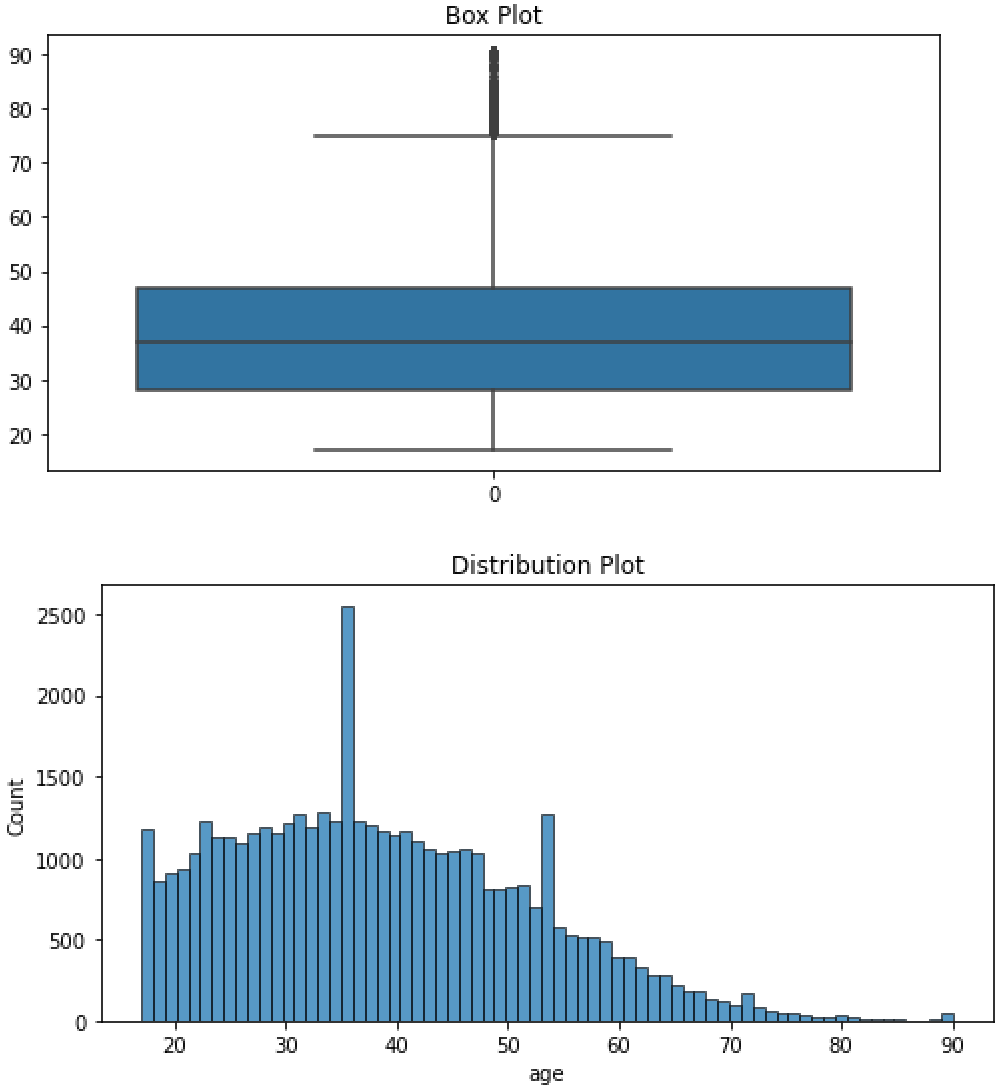
This kind of analysis is performed to find the relationship between each variable in the dataset and the target variable of interest (or) using 2 variables and finding the relationship between them.

Ex:-Box plot, Line plot, Horizontal and vertical bars.

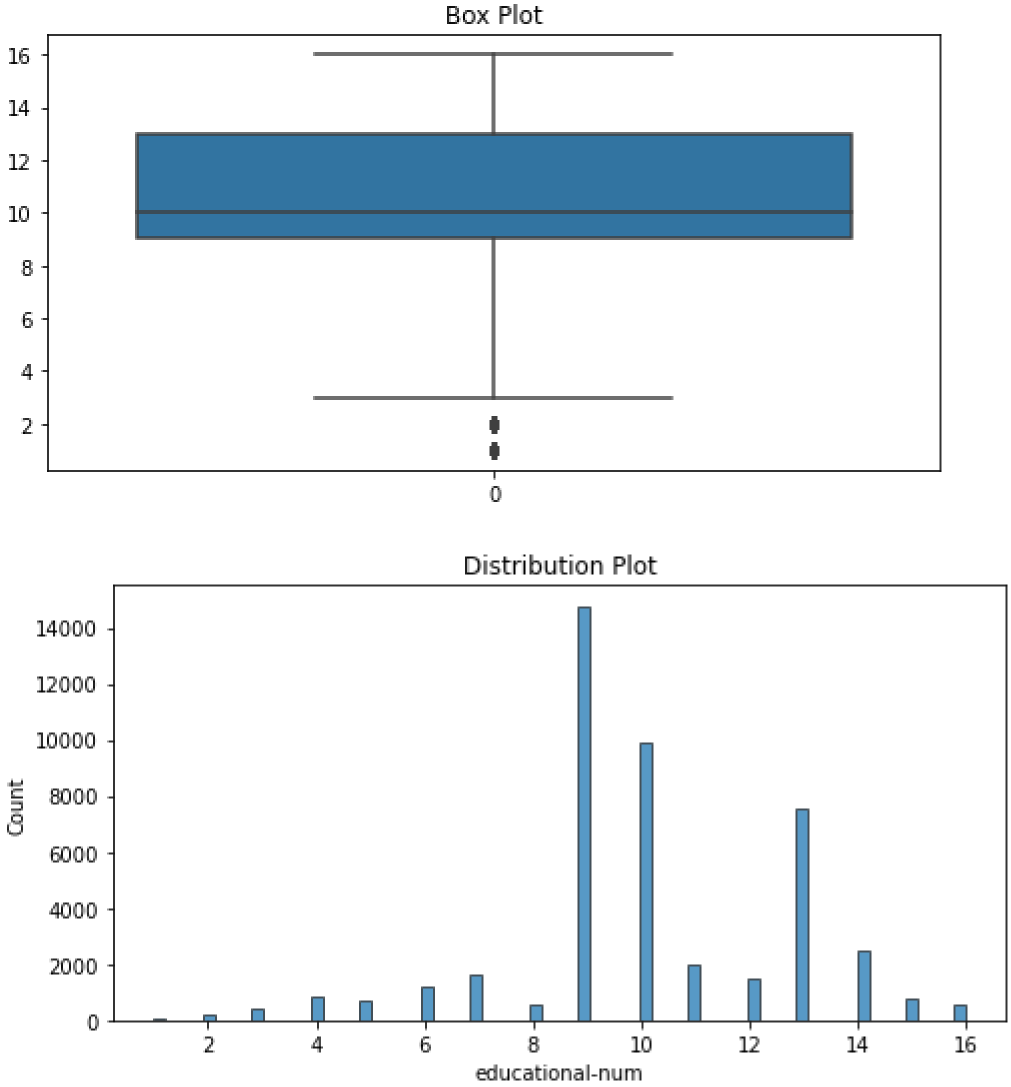
3.1 Univariate Analysis of Project:

Plots for Numerical Variables

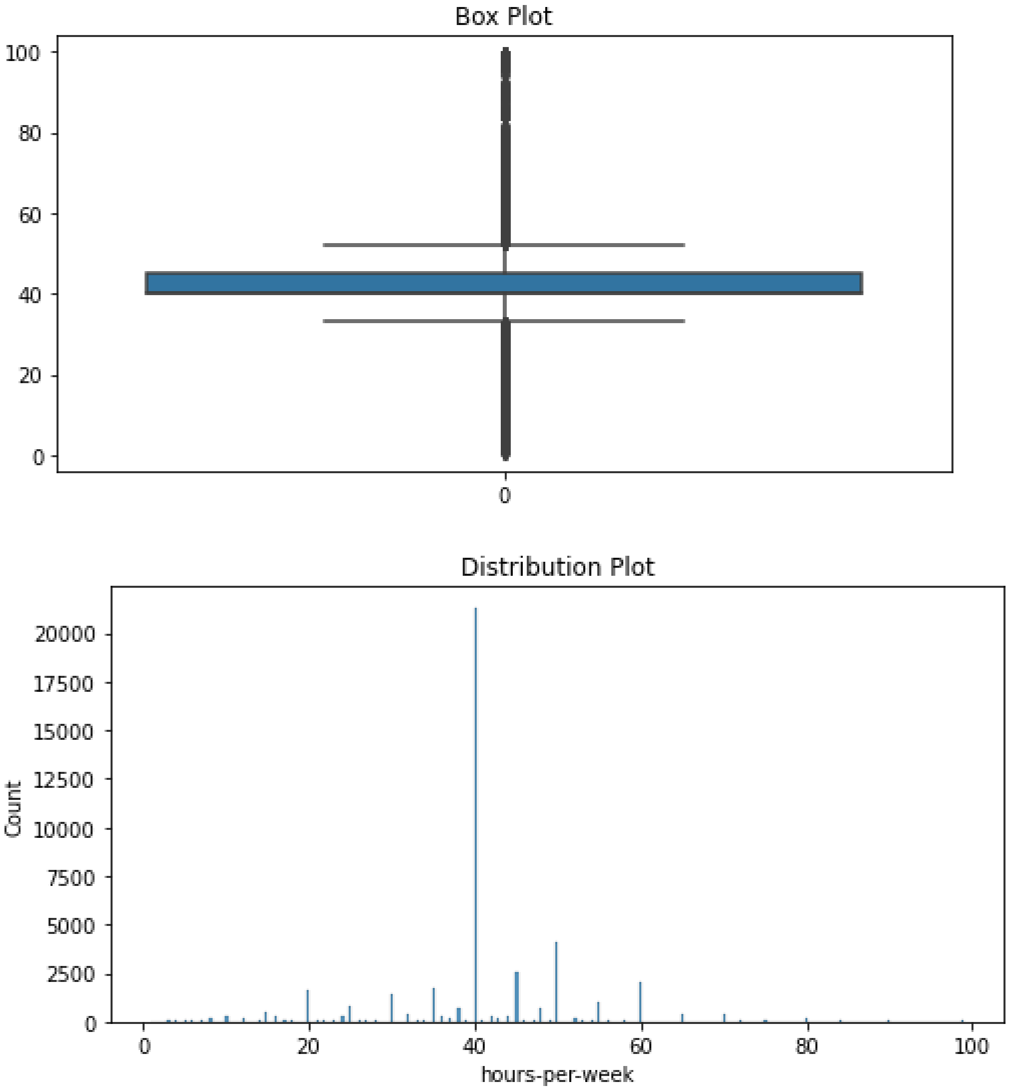
Box and Distribution Plots (Age)



Box and Distribution Plots (Educational Num)



Box and Distribution Plots (Hours per Week)



Valuable Insights:

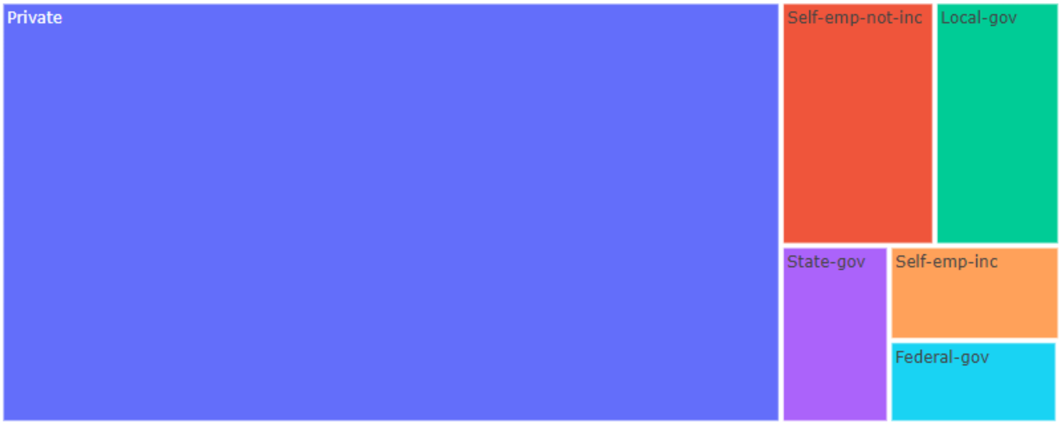
From these plots it can be seen that the majority of individuals in this dataset are between ages 27 and 47 and that the age distribution is skewed towards right. Besides this, the JB Test for normality tells us that the distribution is not normal.

Similarly, the educational num i.e the number of years spent in educations plots indicate that in the dataset   most of the individuals have 9 years of education.

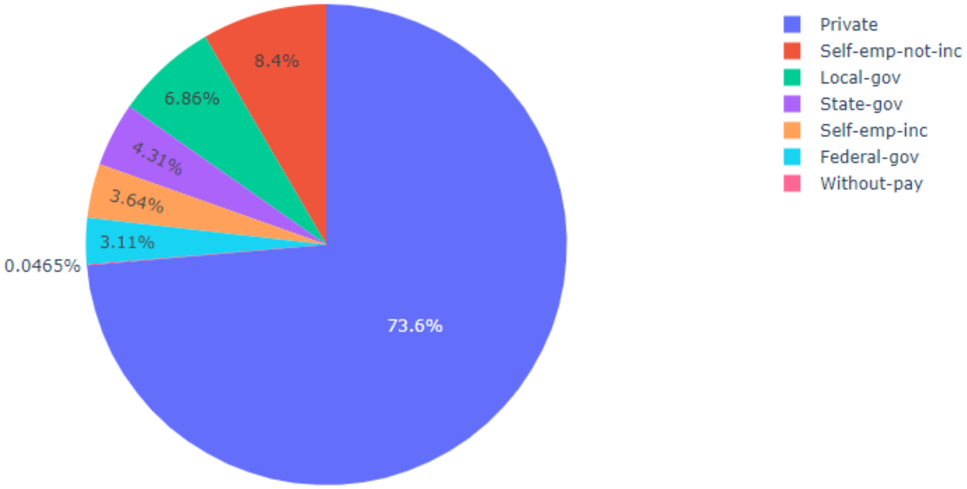
Moreover, most individuals in the dataset work for about 40 hours per week.

Plots for Categorical Variables

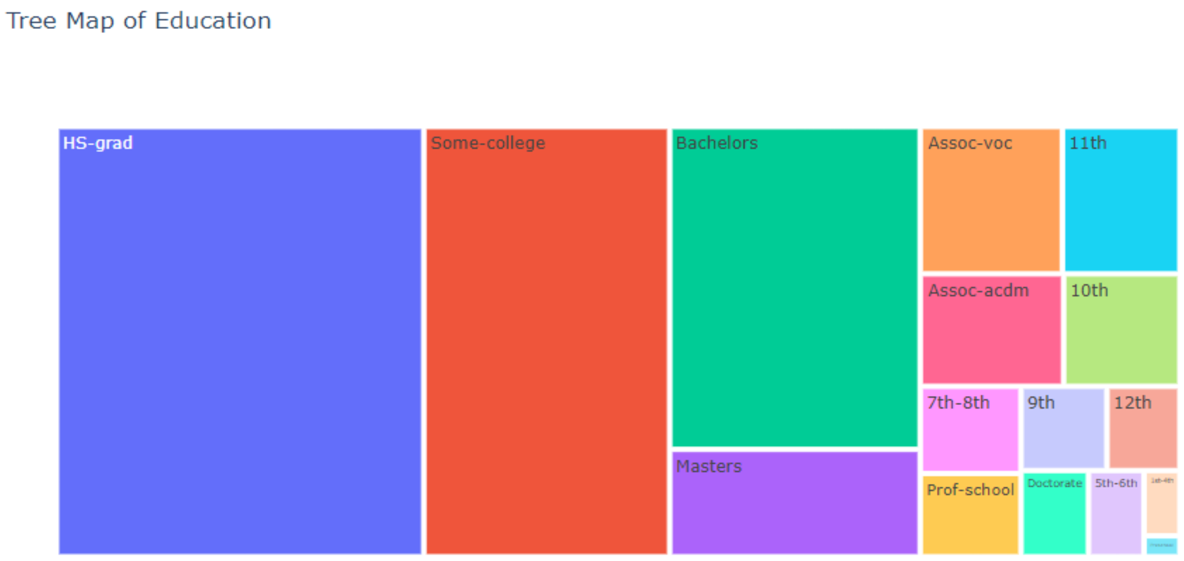
Tree Map and Pie Chart of Work Class:



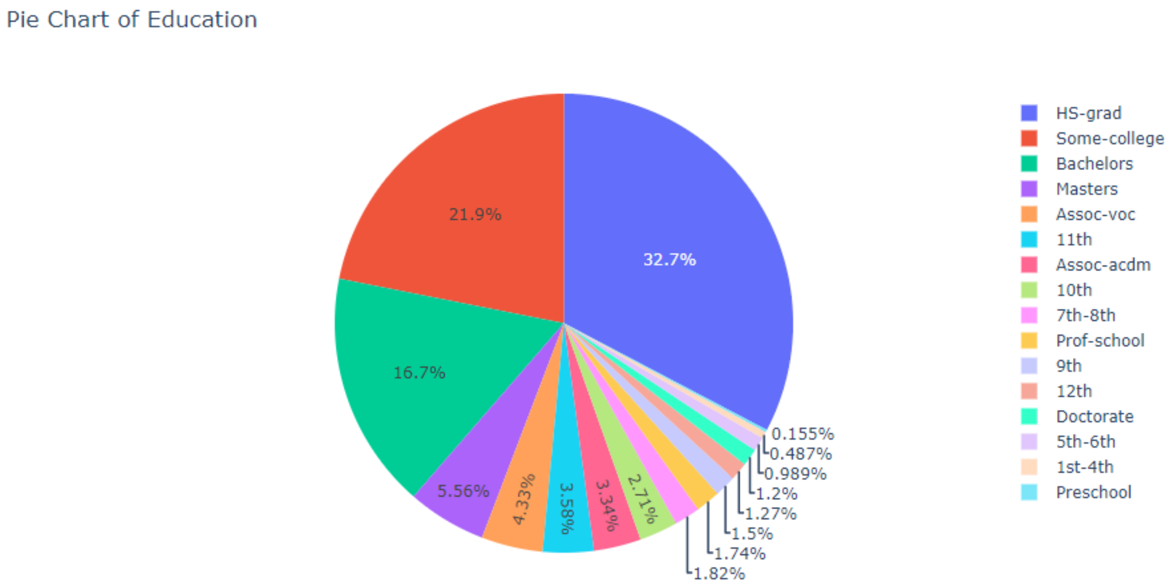
The above tree map shows that the most individuals in the dataset belong to 'private' workclass and the below pie chart is another way to represent the share of each workclass in the dataset.



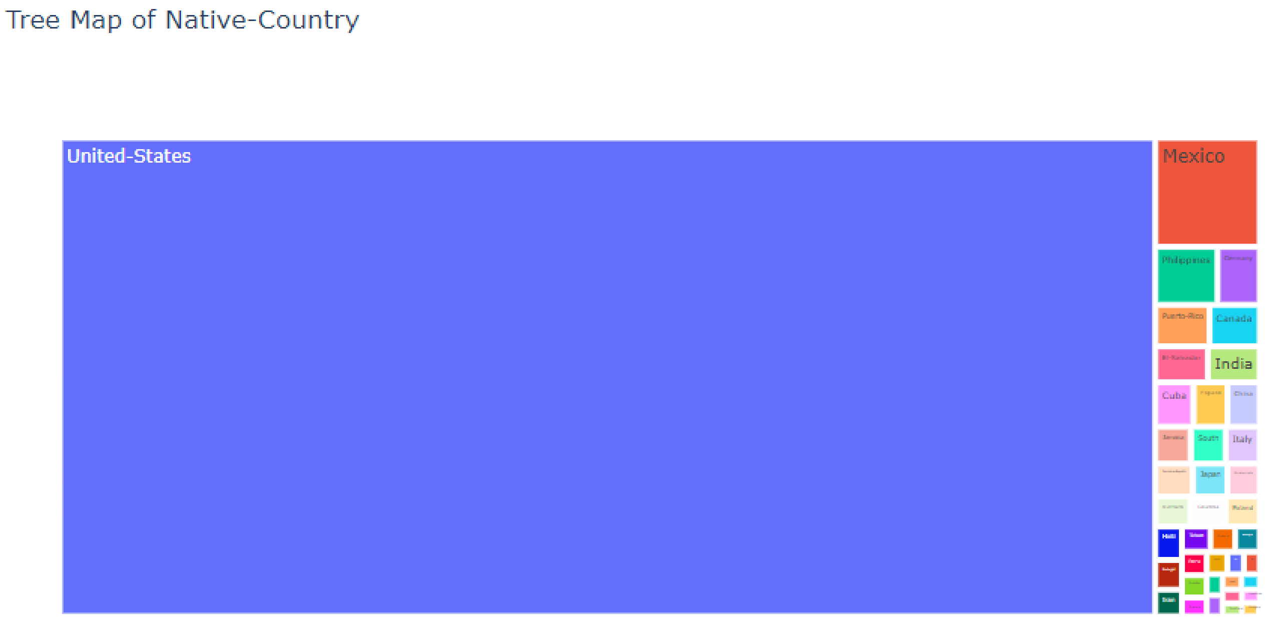
Tree Map and Pie Chart of Education:



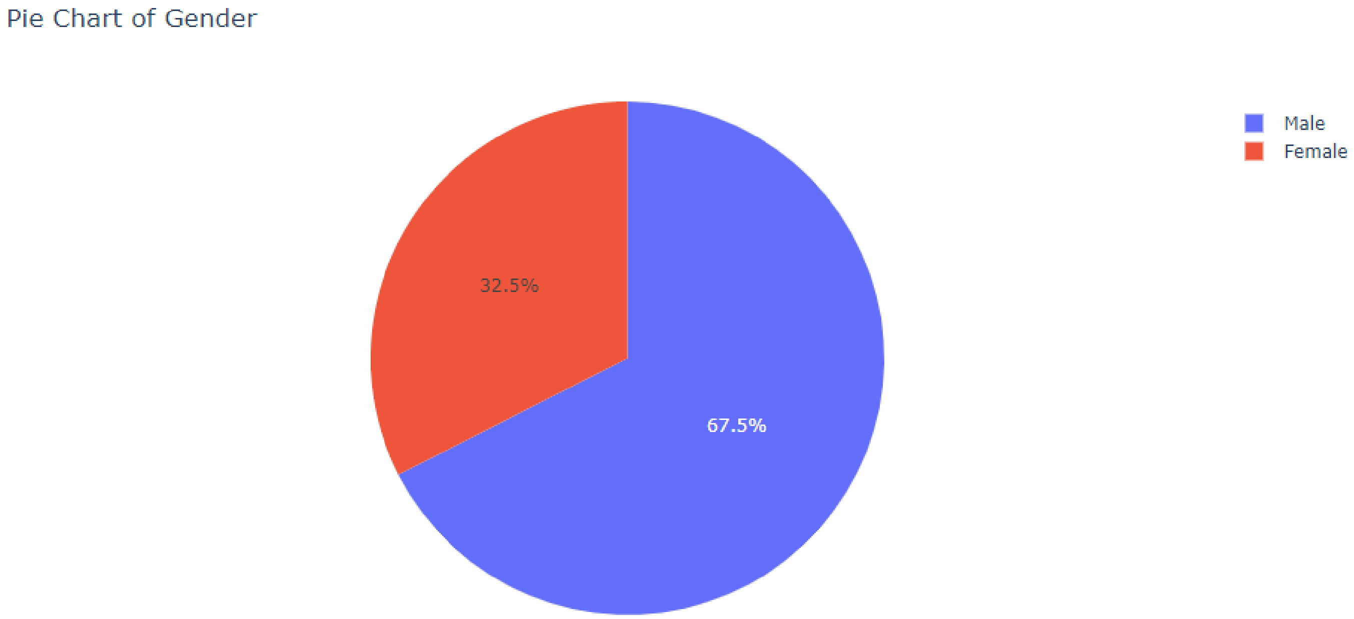
The Tree map of education shows that High School Graduates compose the highest number of data points in the dataset. Similarly, the below pie chart also shows the share of each level of education in the whole dataset.



Tree Map of Native Country:



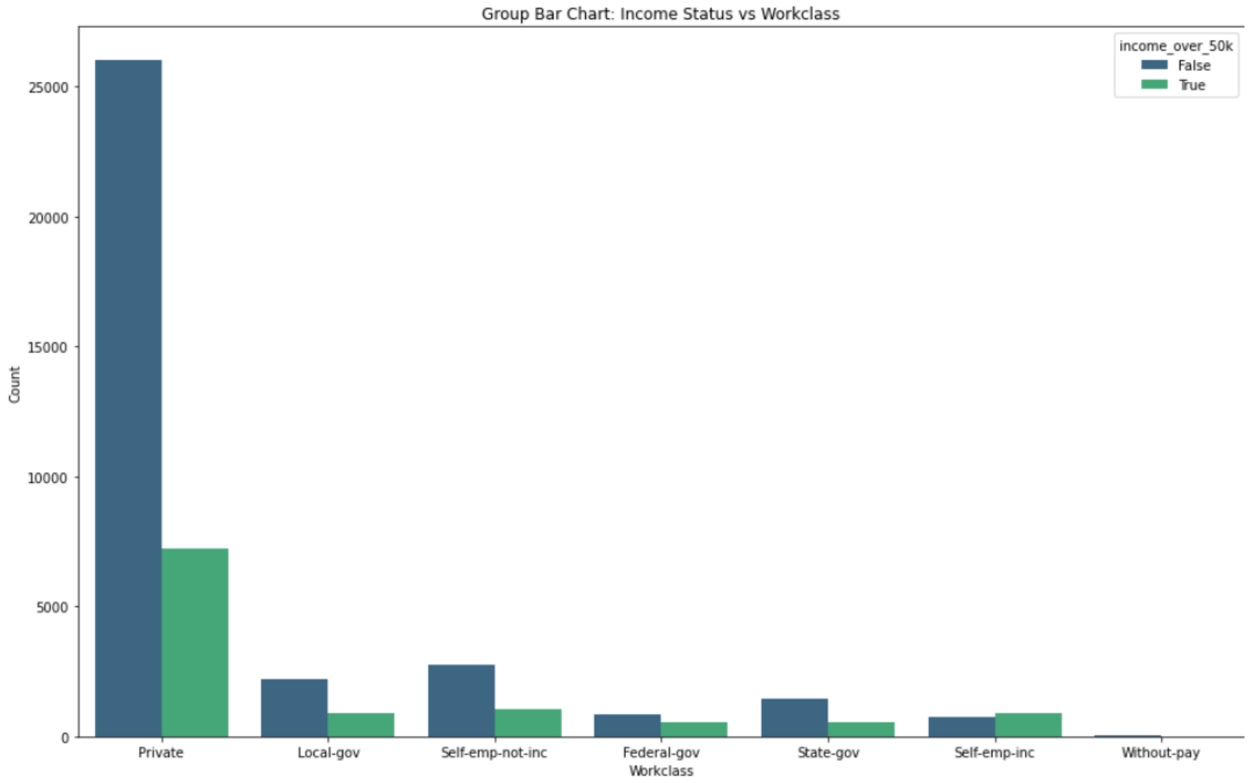
The Tree Map of Native Country shows us that in our dataset the highest number of individuals are those whose native country is United States.



The Pie Chart of Gender gives us a clear picture of the percentage of individuals in the dataset who are male and female.

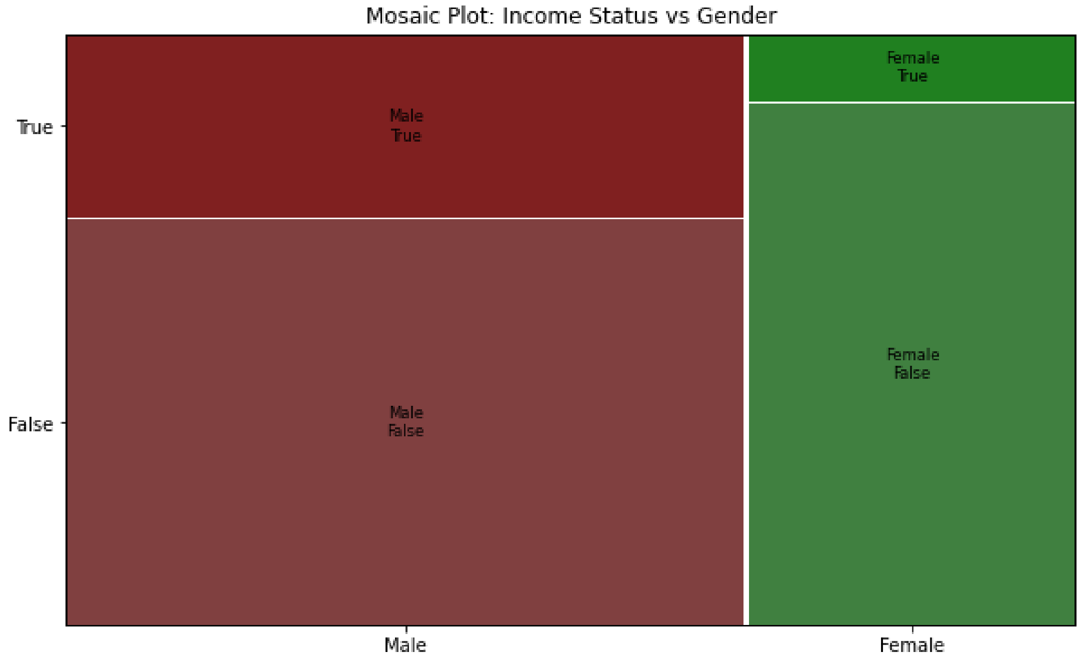
3.2 Bivariate Analysis

Bar Chart For Income Status Vs Work Class



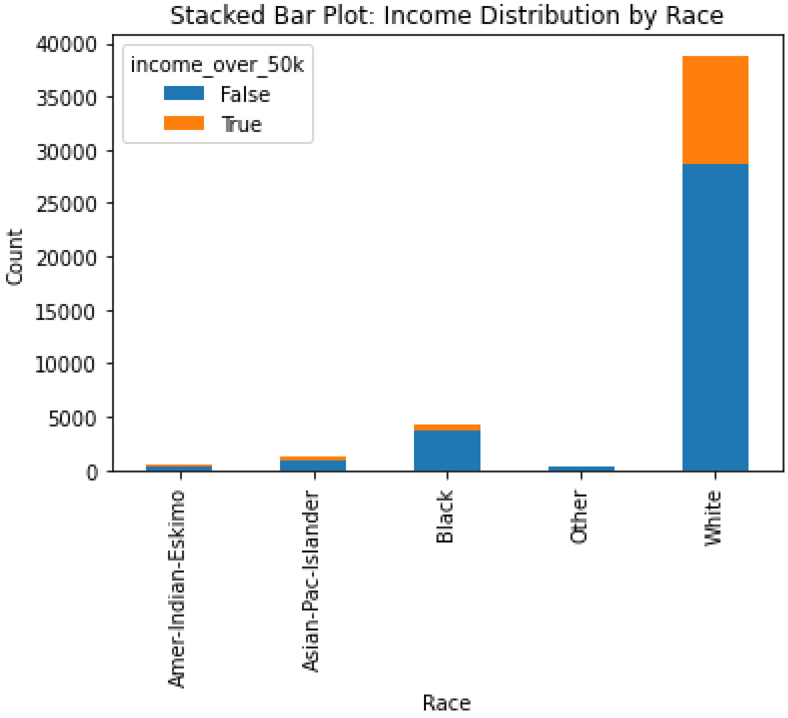
The bar chart shows relationship between 2 variables i.e. work class and income status where false means the income is less than 50000 and true means greater than 50000.

Mosaic Chart for Income Status Vs Gender:



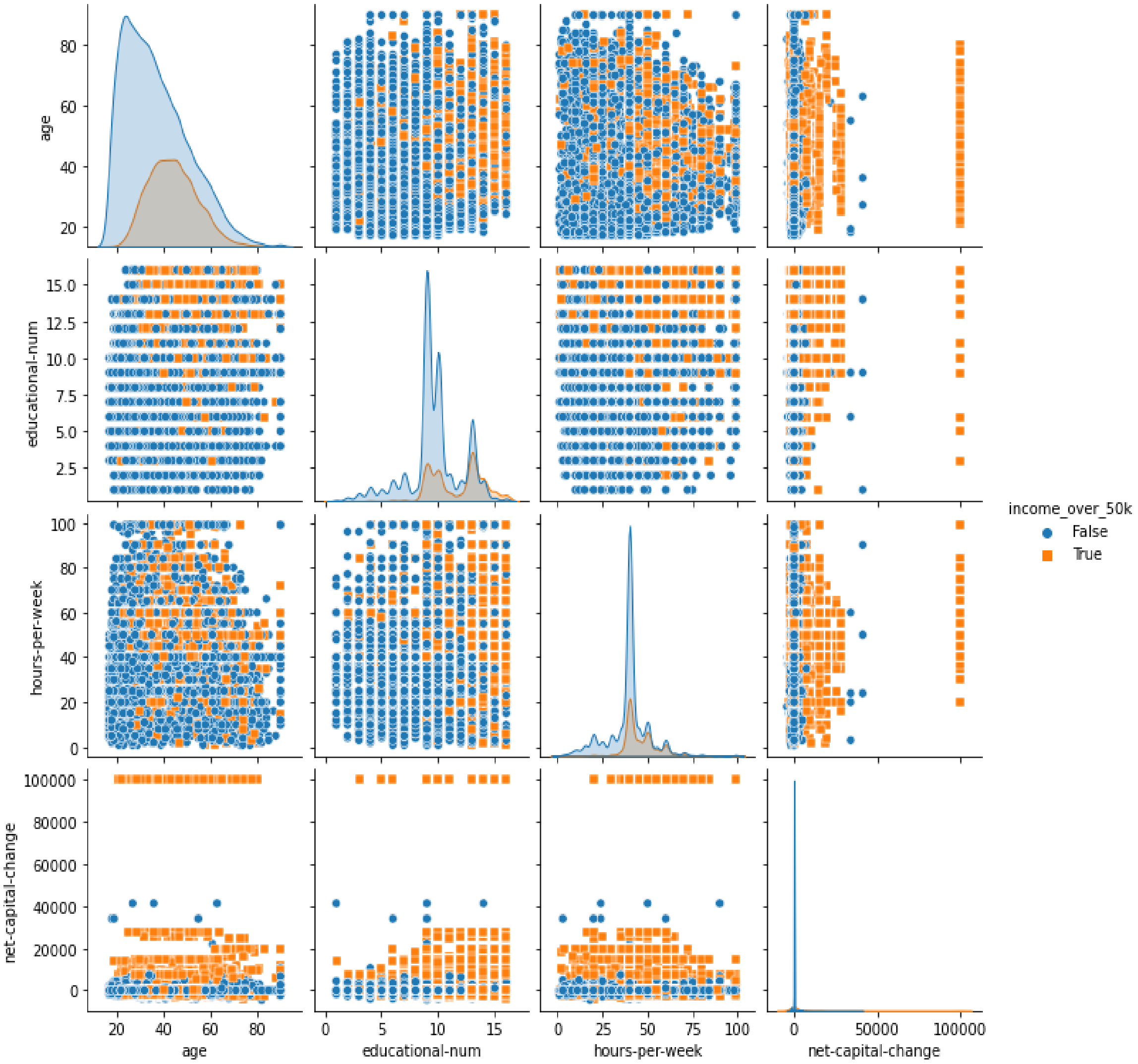
The above Mosaic chart shows that there are more males making more than 50000

Stacked Bar Chart for Income Status Vs Race:



The stacked bar chart gives us the insight that in our dataset the white race individuals have the highest count of thos with income above 50000.

Importance of each feature based on the bivariate analysis (Scatter Plots)



Scatter Plot 1: Net Capital Change vs. Age

This scatter plot shows a positive correlation between net capital change and age. This means that, in general, older people tend to have accumulated more wealth than younger people. There are a few possible explanations for this:

- Older people have had more time to work and save money.
- Older people may be more likely to own assets such as homes and businesses.
- Older people may be more risk-averse and therefore invest their money more conservatively, which can lead to higher returns over time.

Scatter Plot 2: Net Capital Change vs. Educational Attainment

This scatter plot also shows a positive correlation, but it is not as strong as the correlation between net capital change and age. This suggests that education does play a role in wealth accumulation, but it is not the only factor.

Scatter Plot 3: Net Capital Change vs. Hours Worked per Week

This scatter plot shows a very weak correlation between net capital change and hours worked per week. This suggests that working more hours does not necessarily lead to greater wealth accumulation. There are a few possible explanations for this:

- People who work more hours may have less time to save money.
- People who work more hours may be more likely to be paid lower wages.
- People who work more hours may be more likely to spend more money on other expenses, such as childcare and transportation.

It is important to note that these scatter plots only show correlations, not causation. This means that we cannot say for sure that one variable causes the other. For example, it is possible that the positive correlation between net capital change and age is due to other factors, such as the fact that older people are more likely to be homeowners.

Overall, the scatter plots in the image suggest that age and educational attainment are both important factors in wealth accumulation. However, the relationship between net capital change and hours worked per week is less clear.

4 DATA PREPROCESSING

Data preprocessing is a crucial step in any Machine Learning task, as it involves transforming raw data into a clean and usable format for model training. This essential process ensures that both numerical and categorical data are appropriately prepared for analytical operations, predictions, and modeling. By employing various techniques, preprocessing aims to enhance the quality and suitability of the data, setting the foundation for effective machine learning analyses.

4.1 DISCRETIZATION

Feature discretization is an essential part of data preprocessing. Through discretization, continuous data variable can be transformed into discrete form. This is executed by making a set of contiguous intervals that are applied to the entire range of continuous data variable/column, making it discrete.

In our Project we applied discretization to age and hours per week.

4.2 NORMALIZATION

Normalization is the process of scaling numeric variables to a standard range, often between 0 and 1. This ensures that variables with different scales contribute equally to the analysis and prevents certain features from dominating others. Standard normalization techniques include Min-Max scaling and Z-score normalization. Min-Max scaling scales values based on the range of the variable, while Z-score normalization transforms values to have a mean of 0 and a standard deviation of 1.

However, z-score is preferred more when trying to handle outliers but min-max normalization technique has proven more useful in bringing features on the same scale.

In our project min-max scaling was used so as to preserve the shape of data, and since distribution is not normal and there's no requirement of normal distribution. Also, it doesn't eliminate the impact of outliers which sometimes provide useful information

4.3 Feature Encoding

Feature encoding is essential for transforming categorical variables into a format suitable for machine learning algorithms. This includes encoding nominal and ordinal variables. Nominal variables, without inherent order, are often one-hot encoded. Ordinal variables, with a meaningful order, can be label encoded or assigned numerical values based on their order. Feature encoding ensures that algorithms can effectively interpret and utilize categorical information. There are three techniques of encoding labels:

- Manual Encoding
  In this technique, the codes are assigned manually by the programmer. It is mainly used for ordinal categorical attributes.
- Label Encoding
  Applied mainly on nominal categorical attributes.
- One-Hot Encoding
  This technique allows encoding of categorical data by making a separate column.

In our project we did manual encoding of class column to preserve the class attribute while all the categorical columns were one-hot encoded.

4.4 Train-Test Split

In train and test split, the data set is divided into a ratio of 70:30. It is used as an estimation technique for the performance of machine learning algorithms, ensuring accurate results by providing the training dataset and comparing the machine results dataset with the own learning model.

5 DASH

Dash relies on three key technologies for its core functionality:

Flask:
Flask provides essential functionality to the web server, forming the backbone of the Dash framework.

React.js:
React.js is a fundamental technology used within Dash for building interactive user interfaces.

Plotly.js:
Plotly.js is employed in Dash for generating charts, enabling the implementation of various graphical elements within the application.

To establish a foundation for your Dash application, it's crucial to import important libraries, including dash, dash core components, dash html components, and pandas. Dash takes on the role of initializing the application, while dcc facilitates the creation of interactive components such as graphs, drop-downs, and date ranges. Dash html components provide access to HTML tags, and pandas is utilized for data comprehension and organization.

The layout of the application is defined through Dash, employing dash html components. For instance, html.div is used to structure the page layout into parent and children components. Additionally, html.h and html.p are utilized for creating headings and paragraphs within the children body, incorporating all the main content of the web application, such as texts and graphs.

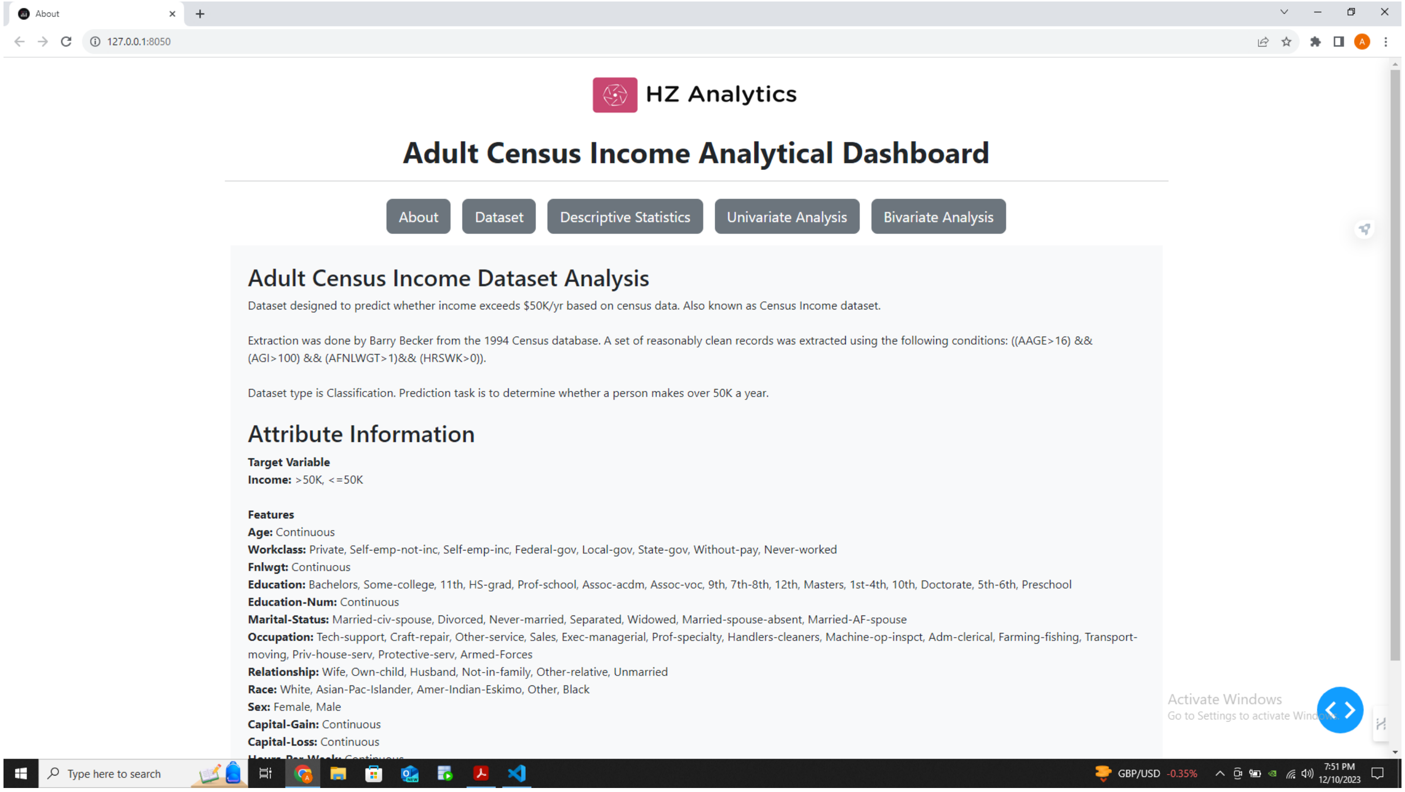
After defining the application's layout, the following Flask and Dash commands are typically used:

```
if __name__ == "__main__":
    app.run_server(debug=True)
```

This allows us to run the Dash application locally using the built-in Flask server.

The following are the screenshots from the dash application used in our project :

<u>The About Page</u>



HZ Analytics

Adult Census Income Analytical Dashboard

Adult Census Income Dataset Analysis

Dataset designed to predict whether income exceeds $50K/yr based on census data. Also known as Census Income dataset.

Extraction was done by Barry Becker from the 1994 Census database. A set of reasonably clean records was extracted using the following conditions: ((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0)).

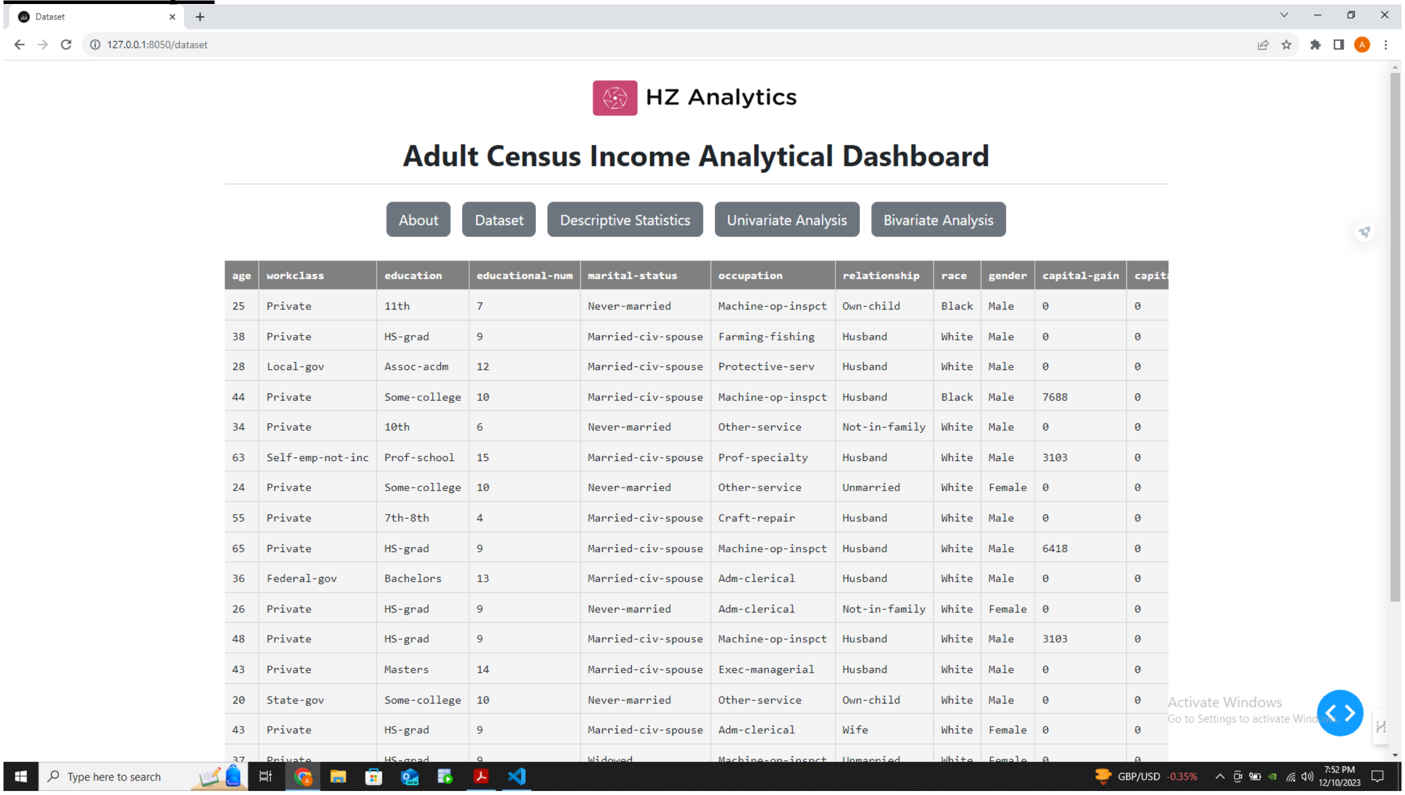
Dataset type is Classification. Prediction task is to determine whether a person makes over 50K a year.

Attribute Information

Target Variable
Income: >50K, <=50K

Features
Age: Continuous
Workclass: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
Fnlwgt: Continuous
Education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool
Education-Num: Continuous
Marital-Status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse
Occupation: 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: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
Race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
Sex: Female, Male
Capital-Gain: Continuous
Capital-Loss: Continuous
Hours-Per-Week: Continuous

Dataset Output

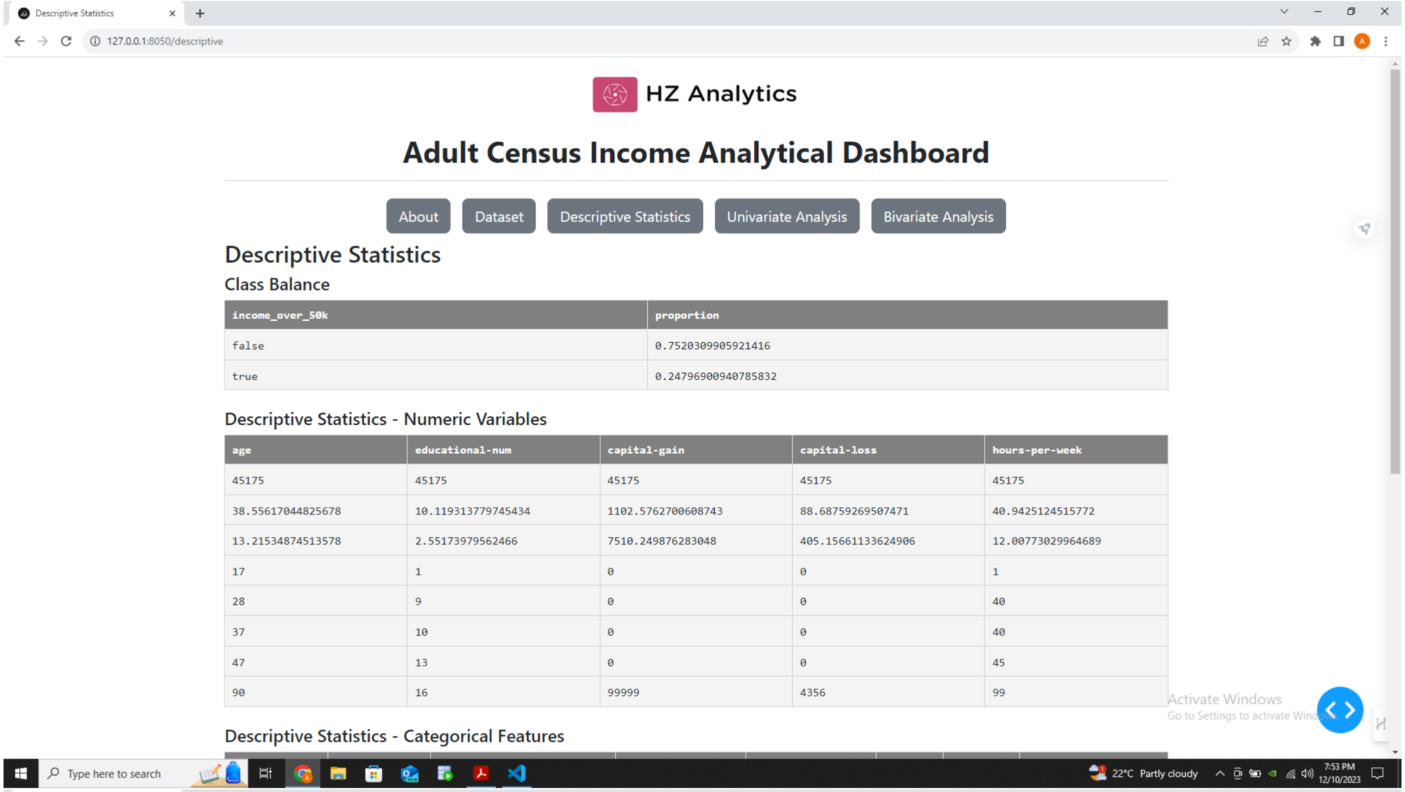


HZ Analytics

Adult Census Income Analytical Dashboard

age	workclass	education	educational-num	marital-status	occupation	relationship	race	gender	capital-gain	capit
25	Private	11th	7	Never-married	Machine-op-inspct	Own-child	Black	Male	0	0
38	Private	HS-grad	9	Married-civ-spouse	Farming-fishing	Husband	White	Male	0	0
28	Local-gov	Assoc-acdm	12	Married-civ-spouse	Protective-serv	Husband	White	Male	0	0
44	Private	Some-college	10	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688	0
34	Private	10th	6	Never-married	Other-service	Not-in-family	White	Male	0	0
63	Self-emp-not-inc	Prof-school	15	Married-civ-spouse	Prof-specialty	Husband	White	Male	3103	0
24	Private	Some-college	10	Never-married	Other-service	Unmarried	White	Female	0	0
55	Private	7th-8th	4	Married-civ-spouse	Craft-repair	Husband	White	Male	0	0
65	Private	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	6418	0
36	Federal-gov	Bachelors	13	Married-civ-spouse	Adm-clerical	Husband	White	Male	0	0
26	Private	HS-grad	9	Never-married	Adm-clerical	Not-in-family	White	Female	0	0
48	Private	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	3103	0
43	Private	Masters	14	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0
20	State-gov	Some-college	10	Never-married	Other-service	Own-child	White	Male	0	0
43	Private	HS-grad	9	Married-civ-spouse	Adm-clerical	Wife	White	Female	0	0
37	Private	HS-grad	9	Widowed	Machine-op-inspct	Unmarried	White	Female	0	0

Descriptive Outputs



HZ Analytics

Adult Census Income Analytical Dashboard

About | Dataset | Descriptive Statistics | Univariate Analysis | Bivariate Analysis

Descriptive Statistics

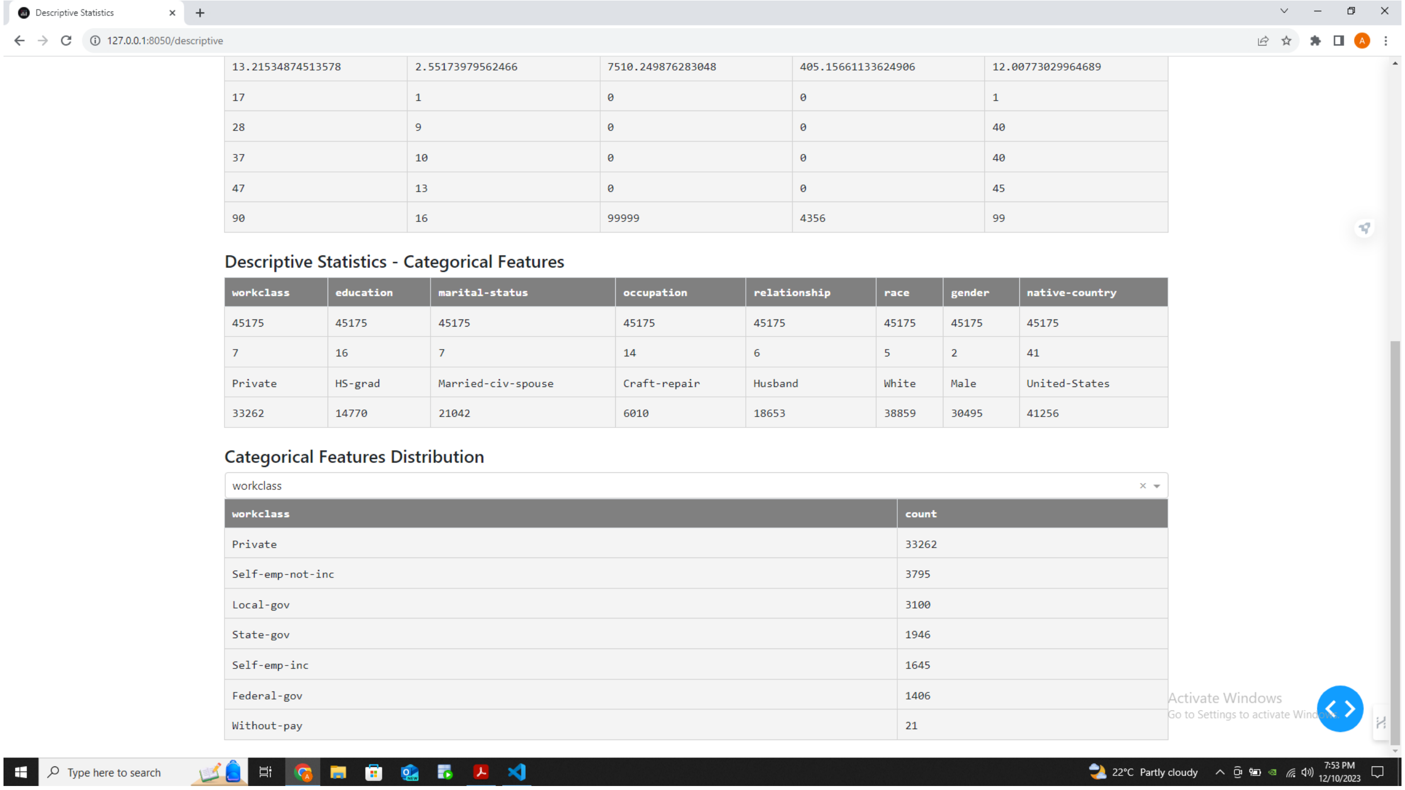
Class Balance

income_over_50k	proportion
false	0.7520309905921416
true	0.24796900940785832

Descriptive Statistics - Numeric Variables

age	educational-num	capital-gain	capital-loss	hours-per-week
45175	45175	45175	45175	45175
38.55617044825678	10.119313779745434	1102.5762700608743	88.68759269507471	40.9425124515772
13.21534874513578	2.55173979562466	7510.249876283048	405.15661133624906	12.00773029964689
17	1	0	0	1
28	9	0	0	40
37	10	0	0	40
47	13	0	0	45
90	16	99999	4356	99

Descriptive Statistics - Categorical Features



13.21534874513578	2.55173979562466	7510.249876283048	405.15661133624906	12.00773029964689
17	1	0	0	1
28	9	0	0	40
37	10	0	0	40
47	13	0	0	45
90	16	99999	4356	99

Descriptive Statistics - Categorical Features

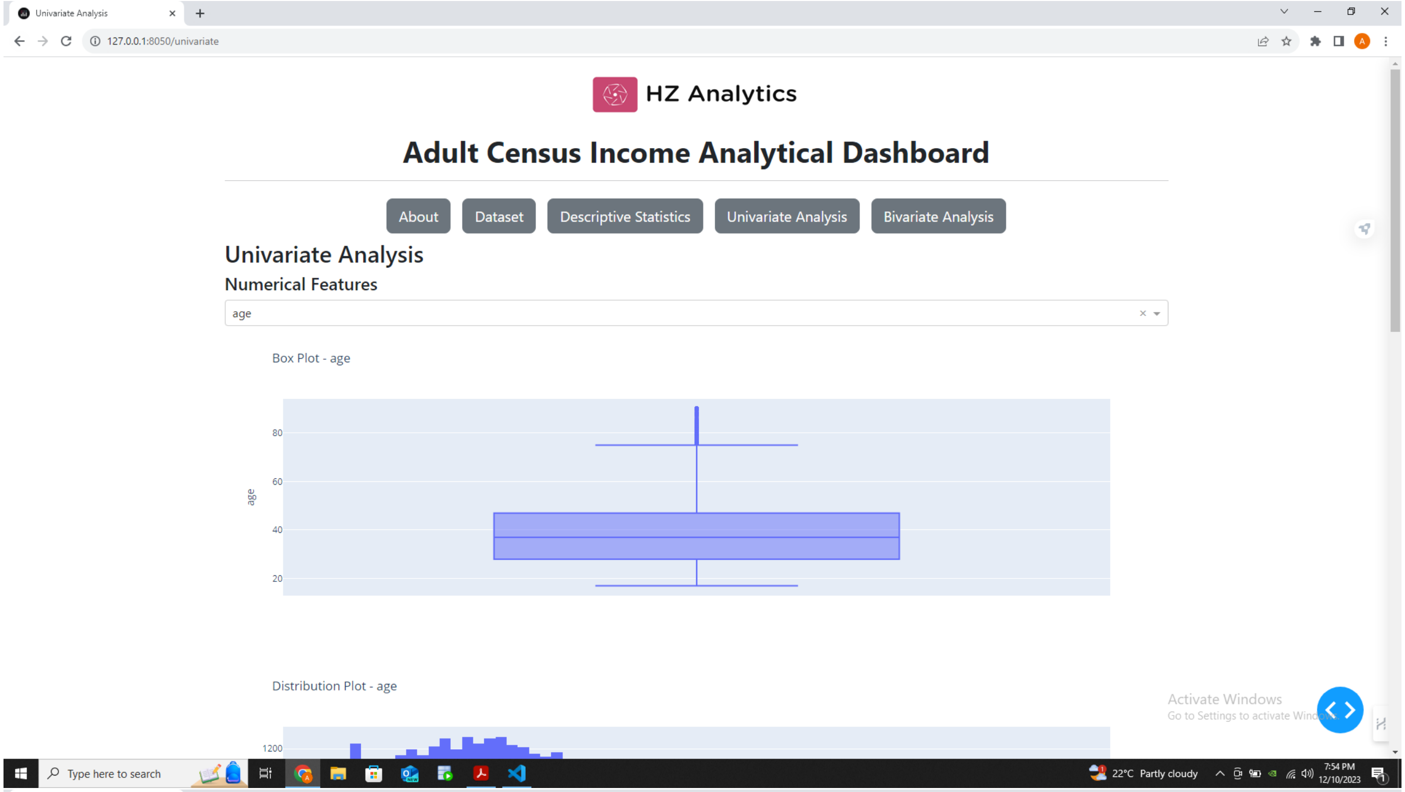
workclass	education	marital-status	occupation	relationship	race	gender	native-country
45175	45175	45175	45175	45175	45175	45175	45175
7	16	7	14	6	5	2	41
Private	HS-grad	Married-civ-spouse	Craft-repair	Husband	White	Male	United-States
33262	14770	21042	6010	18653	38859	30495	41256

Categorical Features Distribution

workclass × ▾

workclass	count
Private	33262
Self-emp-not-inc	3795
Local-gov	3100
State-gov	1946
Self-emp-inc	1645
Federal-gov	1406
Without-pay	21

Univariate Analysis



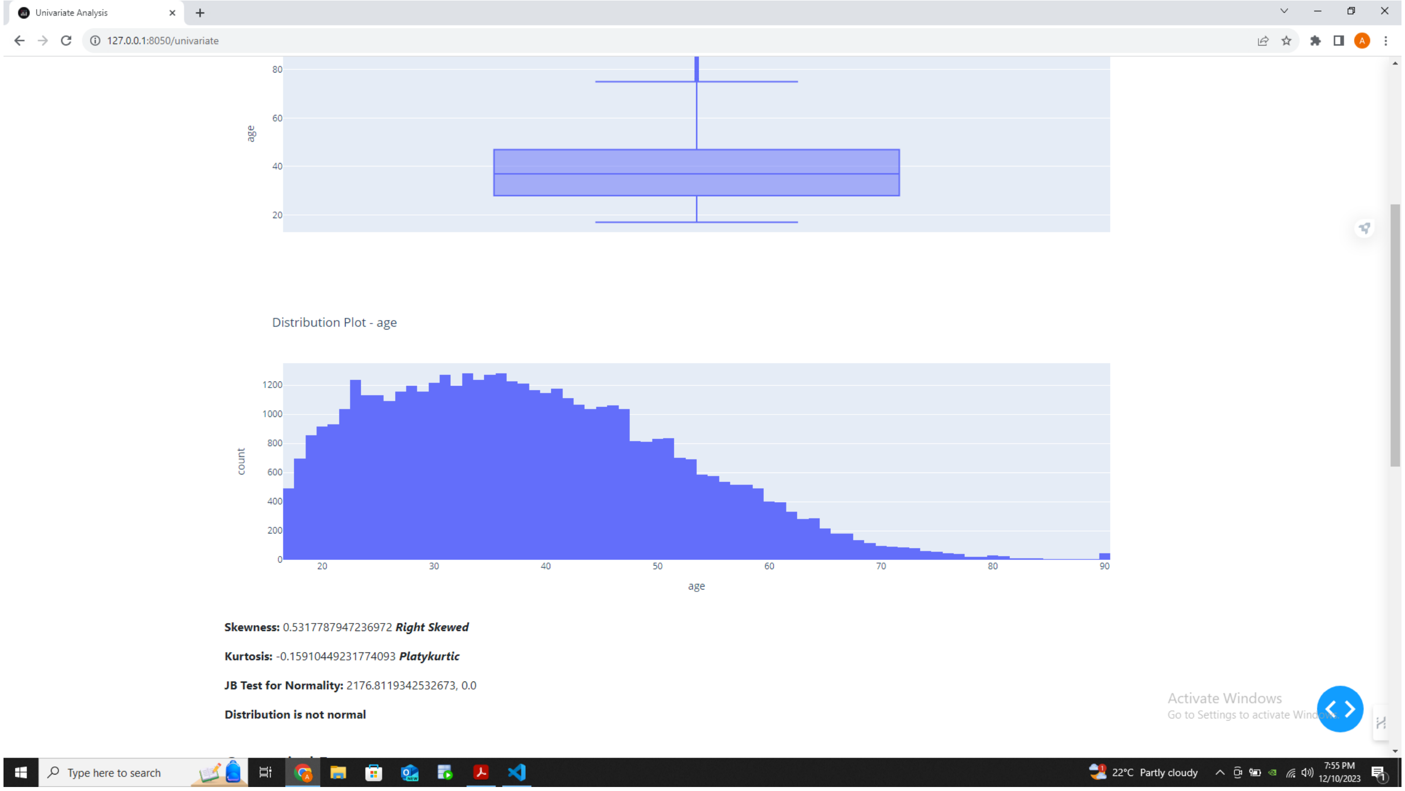
Adult Census Income Analytical Dashboard

Univariate Analysis

Numerical Features

Box Plot - age

Distribution Plot - age



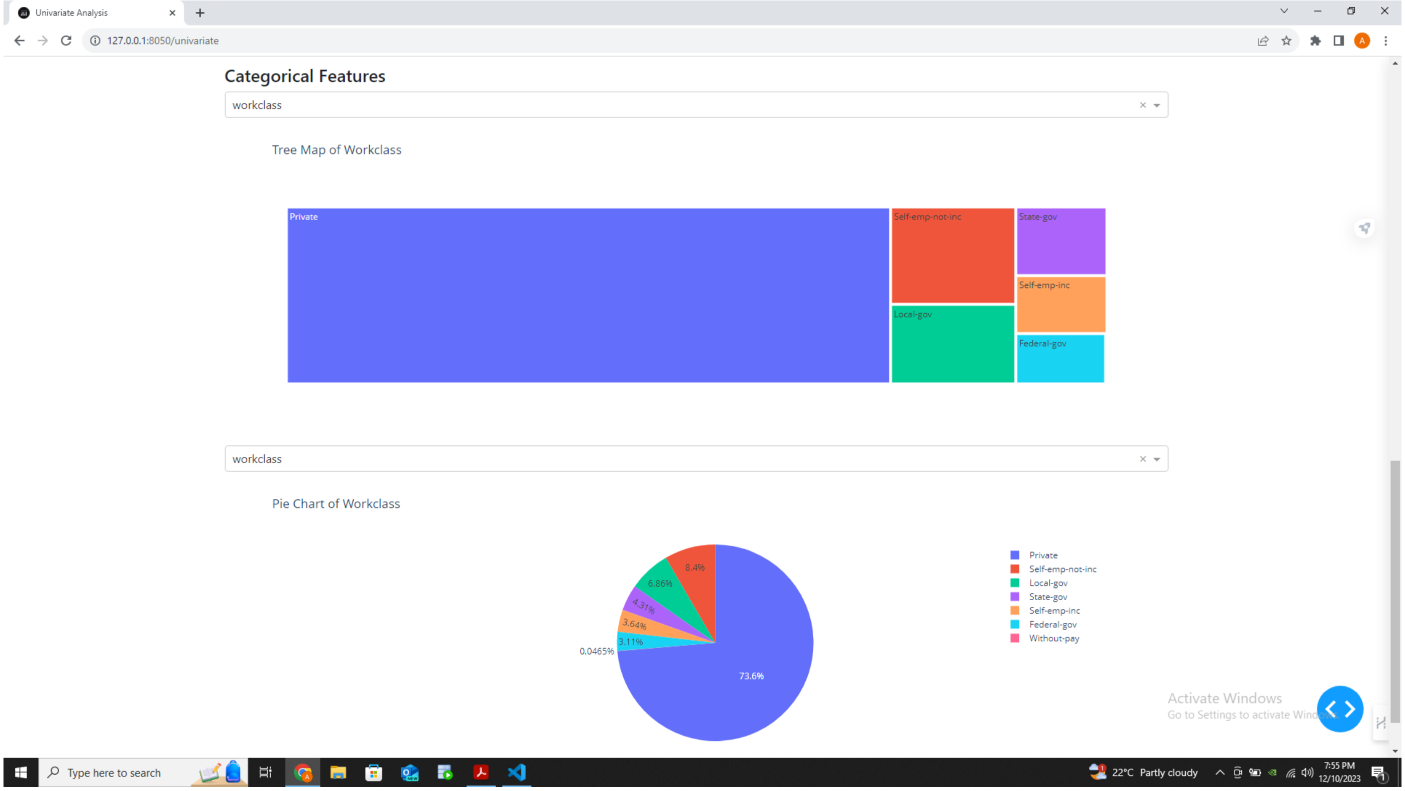
Distribution Plot - age

Skewness: 0.5317787947236972 ***Right Skewed***

Kurtosis: -0.15910449231774093 ***Platykurtic***

JB Test for Normality: 2176.8119342532673, 0.0

Distribution is not normal



Categorical Features

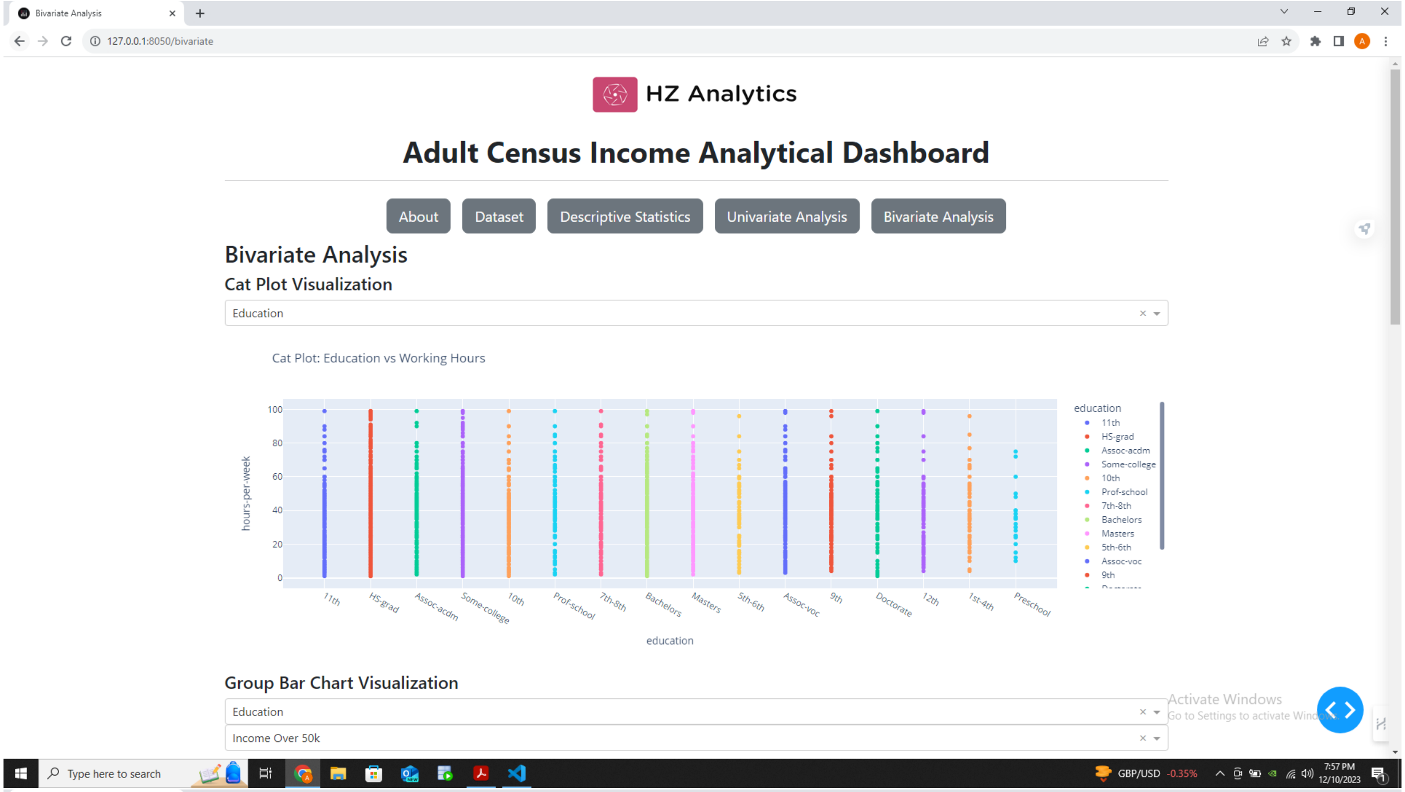
workclass

Tree Map of Workclass

workclass

Pie Chart of Workclass

Bivariate Analysis



HZ Analytics

Adult Census Income Analytical Dashboard

Bivariate Analysis

Cat Plot Visualization

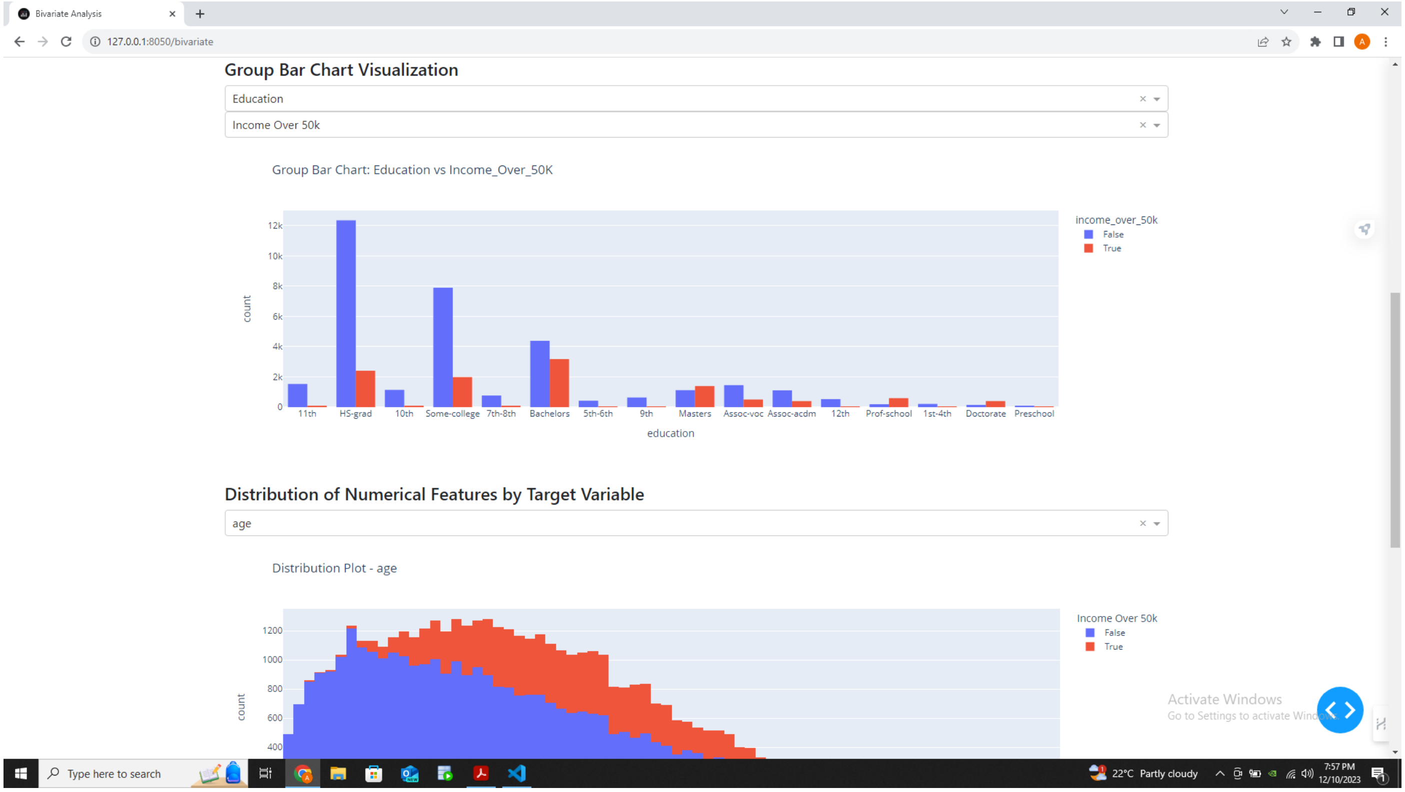
Education

Cat Plot: Education vs Working Hours

Group Bar Chart Visualization

Education

Income Over 50k



Group Bar Chart Visualization

Education

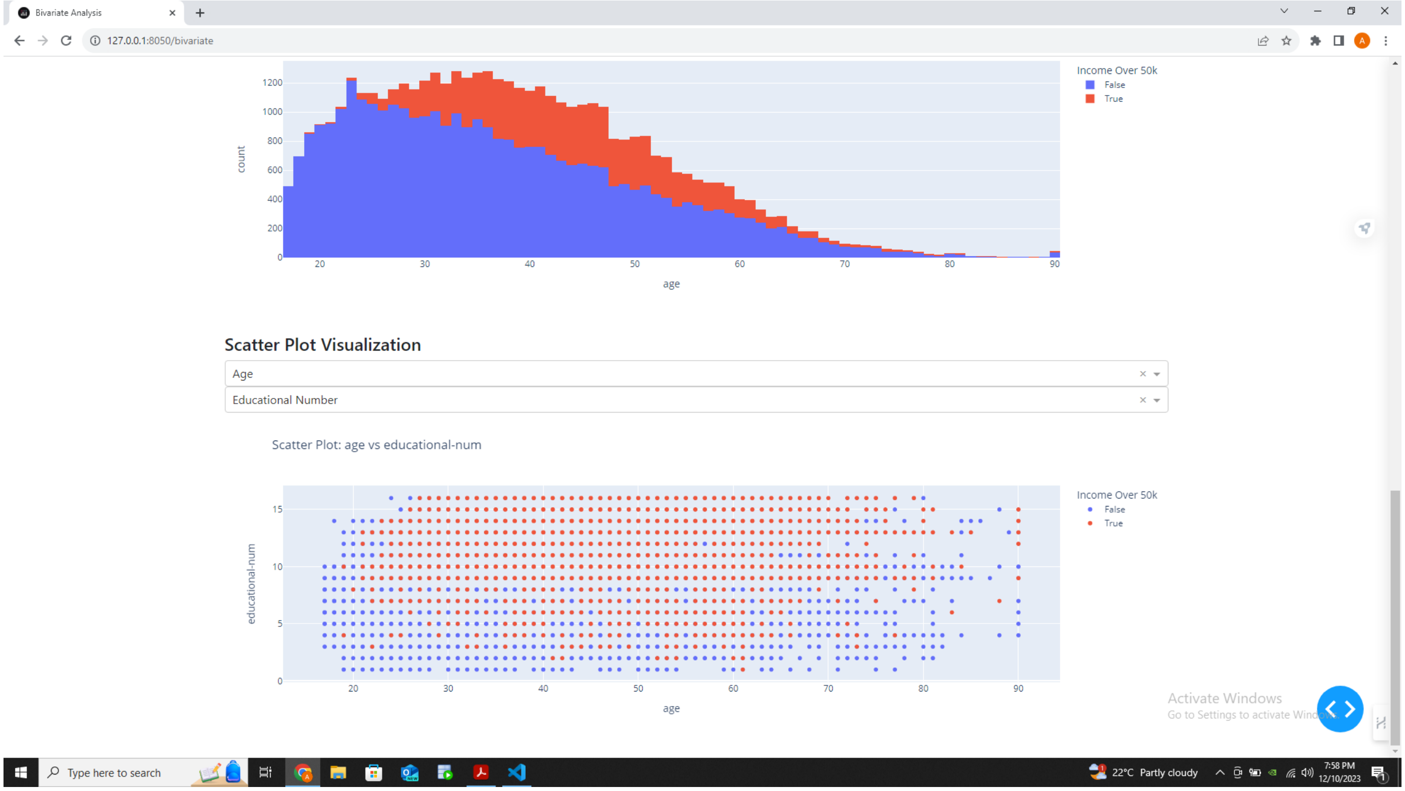
Income Over 50k

Group Bar Chart: Education vs Income_Over_50K

Distribution of Numerical Features by Target Variable

age

Distribution Plot - age



Scatter Plot Visualization

Age

Educational Number

Scatter Plot: age vs educational-num