Workshop 01 Overview

The provided code performs **data analysis, data cleaning, exploratory data analysis (EDA), statistical operations, and visualisation** on the **Adult Income Dataset (adult.csv)**.

Step-by-Step Breakdown

Importing Necessary Libraries

- **Pandas** is used for handling tabular data.
- **NumPy** helps with numerical computations.
- **Matplotlib** is used for creating visualizations.

Loading the Dataset

- Loads the adult.csv file into a Pandas DataFrame.

Adding Feature Names & Reloading the Data

- Assigns meaningful **column names** to the dataset.
- Reloads the dataset with the updated column names.

Displaying Data Samples

- Displays **sample rows** and the **shape** of the dataset.

Data Sampling (Creating a Subset)

- Selects **30,000 random records** from the dataset for analysis.
- Ensures reproducibility using random_state=236.

Statistical Summary

- Generates **summary statistics** for numerical columns.
- Counts occurrences of education-num and education.

Data Cleaning & Feature Selection

- Drops the **fnlwgt (final weight)** column, which is not helpful for analysis.

Exploratory Data Analysis (EDA)

Visualizing Age Distribution

- Creates a **boxplot** and **histogram** for age distribution.

Comparing Mean vs. Median Age

- Checks whether the **mean age is greater than the median**.

Counting Gender Distribution

- Counts the number of **male and female** individuals.

Workclass Distribution

- Counts the **occurrences of each workclass**.

Grouping Data and Aggregating Values

Average Age by Gender

- Computes the **mean age** for each gender.

Capital-Gain Analysis

- Computes **average capital-gain per occupation and gender**.

Filtering & Merging Data

- Separates the dataset into **male and female records**.
- Computes **total capital-gain per marital status**.
- Merges both dataframes for comparison.

Finding Maximum Age Across Races

- Finds the **oldest individual for each race**.

Visualising Capital-Gain & Education

Histogram & Boxplot for Capital-Gain

- Creates a **histogram** and **boxplot** for capital-gain.

Boxplot of Age by Education

- **Compares age distributions** across education levels.

Checking Missing Values

- Counts **missing values** for each column.

Applying Label Encoding to Categorical Data

- **Converts categorical values into numerical format** for machine learning.

Analysing Migration Patterns

- **Counts non-US migrants** and visualises them in a **bar chart**.

Identifying Male-Dominated Occupations

- Identifies **occupations with a higher percentage of males**.

List of Main Functions Used

Function	**Purpose**
pd.read_csv()	Reads the dataset into a Pandas DataFrame
data.head(n)	Displays the first n rows of the dataset
data.tail(n)	Displays the last n rows of the dataset
data.shape	Returns the dimensions of the dataset
data.describe().T	Generates summary statistics
data.drop()	Removes unnecessary columns

Function	Purpose
data.groupby()	Groups data by specified attributes
data.hist()	Creates histograms
data.boxplot()	Creates boxplots
data.value_counts()	Counts unique values in a column
plt.show()	Displays plots
LabelEncoder().fit_transform()	Converts categorical data into numerical form

---

Summary

This notebook performs **exploratory data analysis (EDA)** on the **Adult Income Dataset (adult.csv)** by:

✓ **Cleaning & transforming the data.**
✓ **Generating statistical summaries**
✓ **Visualizing age, income, and occupation distributions**
✓ **Identifying trends based on gender, education, and work class**
✓ **Applying encoding techniques for future machine learning models**