Assignment Report: Exploratory Data Analysis (EDA) on Adult Income Dataset

# 1. Introduction

The objective of this assignment is to perform **Exploratory Data Analysis (EDA)** on the Adult Income dataset. The dataset contains demographic information such as age, education, occupation, race, sex, workclass, etc., along with a target variable indicating whether an individual earns **>50K** or **≤50K** annually.

The aim is to understand the **distribution of attributes, detect patterns, relationships, and trends**, and prepare the dataset for further **predictive modeling**.

# 2. Dataset Overview

* **Dataset Size:** ~32,561 rows × 15 columns
* **Type of Data:** Mixed (categorical & numerical)
* **Key Columns:**
  + **age (numeric):** Age of individuals
  + **workclass (categorical):** Employment type (Private, State-gov, Self-employed, etc.)
  + **education / education\_num (categorical & numeric):** Education level and its numeric equivalent
  + **marital\_status (categorical):** Marital status of individuals
  + **occupation (categorical):** Job type
  + **relationship (categorical):** Relation within the family
  + **race (categorical):** Race of individuals
  + **sex (categorical):** Gender
  + **hours\_per\_week (numeric):** Working hours per week
  + **income (target, categorical):** ≤50K or >50K annual income.

# 3. Data Cleaning & Preprocessing

* **Missing Values:** Some categorical variables like *workclass* and *occupation* had "?" as missing markers, which were handled.
* **Duplicates:** Dataset was checked for duplicates.
* **Encoding:** For machine learning, categorical features would require encoding (e.g., One-Hot Encoding, Label Encoding).
* **Outliers:** Age and working hours contained outliers (e.g., age > 90, hours\_per\_week = 99), which were investigated.

# 4. Exploratory Data Analysis (EDA)

## 4.1 Univariate Analysis

* **Age:** Most individuals fall between **20–50 years**, with a peak around 30–40.
* **Education:** The dataset is dominated by *HS-grad* and *Bachelors*.
* **Workclass:** Majority are from the *Private* sector, followed by *Self-emp-not-inc*.
* **Occupation:** The most frequent occupations include *Prof-specialty*, *Craft-repair*, and *Exec-managerial*.
* **Income:** Around **75% of individuals earn ≤50K**, while only 25% earn >50K.

## 4.2 Bivariate Analysis

* **Age vs Income:** Older individuals (>35) tend to earn more than younger individuals.
* **Education vs Income:** People with higher education (*Bachelors, Masters, Doctorate*) are more likely to earn >50K.
* **Workclass vs Income:** Government and self-employed individuals show higher chances of >50K compared to private employees.
* **Gender vs Income:** Males are more likely to earn >50K than females (gender bias observed).
* **Hours-per-week vs Income:** Individuals working **>40 hours/week** are more likely to fall into the >50K category.

## 4.3 Categorical Insights

* **Race vs Income:** The dataset is predominantly White, and income disparity is observed across racial groups.
* **Marital Status vs Income:** Married individuals, especially *Married-civ-spouse*, are more likely to earn >50K.

# 5. Key Insights

1. Higher **education** strongly correlates with higher income.
2. **Work experience (age + hours worked)** positively influences income.
3. **Gender disparity** exists: males have a higher proportion of >50K income.
4. **Private sector employees** dominate the dataset but are less likely to earn >50K compared to other sectors.
5. **Marital status** is a strong predictor of income levels.

# 6. Conclusion

The EDA provided meaningful insights into the Adult Income dataset:

* Education and occupation are crucial predictors of income.
* Demographic factors such as gender, marital status, and race influence income levels.
* A significant **class imbalance** exists in the target variable (more ≤50K than >50K), which should be addressed in predictive modeling using techniques like **SMOTE or class weighting**.

This analysis establishes a strong foundation for building **classification models** (e.g., Logistic Regression, Decision Trees, Random Forests) to predict income levels.