# REPORT

**1. Data Cleaning Insights**

- **Duplicates Removed:**
    - The dataset initially contained **24 duplicate rows**, which were dropped to ensure data integrity.
- **Missing Values Handled:**
    - **workclass**: **1836 missing values** were filled based on the most frequent category.
    - **occupation**: **1843 missing values** were filled using the mode (most common occupation in the respective workclass).
    - **native-country**: **583 missing values** were filled with the most frequent country in the dataset.
- **Outliers Identified & Treated:**
    - **Capital-Gain and Capital-Loss:**
        - The **capital-gain column had extreme values, with a maximum of 99,999**, which were verified as legitimate.
        - The **capital-loss column had a maximum of 4,356**, showing a similar pattern.
    - **Hours-Per-Week:**
        - The maximum work hours recorded were **99 hours per week**, which, while extreme, was retained as valid.

---

**2. Key Graph-Based Insights**

**Income Distribution**

- **75.92% of individuals** earned below the income threshold (<=50K), while **24.08% earned more than >50K**.
- The dataset is **highly imbalanced**, with a majority falling in the lower-income category.

**Age vs. Income**

- The **average age** of individuals in the dataset is **38.78 years**.
- Most high earners were aged **35 to 50 years**.
- Individuals under **25 years** and over **60 years** had significantly lower income proportions.

**Gender-Based Disparities**

- **Men make up 67.5% of high-income earners (>50K), while women account for only 32.5%)**.
- **Women are more concentrated in lower-paying occupations**, especially clerical and caregiving roles.

**Workclass & Income**

- **Private-sector employees** account for **73% of the dataset** and show the **widest income range**.
- **Government employees** (federal, state, and local) had **lower income variance** but stable earnings.
- **Self-employed individuals** showed **the highest variability**, with some earning significantly more, while others earned far less than salaried employees.

**Occupation-Based Income Trends**

- **Top earning professions:**
    - **Executive/Managerial**: **48.3% earn above >50K**.
    - **Professional Specialty**: **45.6% earn above >50K**.
    - **Tech-related fields** had a strong presence in the high-income category.
- **Lowest earning professions:**
    - **Clerical, Service, and Laborers**: Majority earn **<=50K**.
    - **Farming, Fishing, and Handlers** had the lowest proportion of high earners.

**Racial Disparities**

- **White individuals** make up **85.4%** of the dataset and have the highest proportion of high-income earners.
- **Black individuals** constitute **9.6%**, but only **11%** of them earn more than >50K.
- **Asian-Pacific Islanders and Native Americans** have **lower representation** but show **higher education levels**, leading to a slightly better income distribution.

**Native Country & Income Trends**

- **United States (91.4%)** dominates the dataset, with the highest number of high-income earners.
- Individuals from **India, Canada, and Germany** had relatively higher incomes compared to other non-U.S. countries.
- **Developing countries** (e.g., Mexico, Philippines, South America) had **a significantly lower percentage of high-income earners**.

**Work Hours & Income**

- **People working 40+ hours per week** had a significantly higher proportion of >50K earners.
- **Part-time workers (≤30 hours per week)** were largely in the <=50K category.
- **A notable anomaly:** Some individuals working **60+ hours per week** still earned <=50K, suggesting industry-based income limitations.

---

**Final Observations**

- **Education, workclass, and occupation were the strongest predictors of high income.**
- **Significant gender and racial disparities exist in income distribution.**
- **Working longer hours generally correlated with higher earnings, but some occupations still had limited upward mobility.**