Data Exploration and Insights Report

Overview

This report explores the UCI Adult dataset, focusing on descriptive statistics, variable distributions, correlations, and insights into categorical and continuous features. The dataset's goal is to predict income levels based on demographic and employment-related attributes.

Key Findings

1. Dataset Metadata and Variables

The dataset includes:

- **Features (X):** 14 columns, including demographics (age, race, sex), education, occupation, and financial attributes.
- **Target (y):** Binary indicator of income ("<=50K" or ">50K").
- Missing values were present in "workclass," "occupation," and "native-country."

2. Data Types and Missing Values

All columns matched expected data types. Missing values accounted for:

- **workclass:** 2,799 missing entries (5.9%)
- **occupation:** 2,809 missing entries (5.9%)
- **native-country:** 857 missing entries (1.8%) These missing values were replaced with NaN for analysis.

3. Continuous Variables Analysis

a) Capital Gain and Capital Loss

- Highly skewed distributions with most values at zero.
- Median values: **0** for both variables.
- Categories created:
  - **Capital Gain:** "No Gain," "High Gain"
  - **Capital Loss:** "No Loss," "High Loss"
- Insights:
  - 91.7% had no capital gain, and 95.3% had no capital loss.

b) Final Weight (fnlwgt)

- Distribution showed significant skewness (0.64).
- Outliers: 1,453 entries (2.9%).
- Comparison by gender revealed slightly higher mean fnlwgt for males.

4. Categorical Variables Analysis

Key attributes like "workclass," "occupation," and "native-country" showed rich categorical diversity:

- Example: "workclass" included "Private," "Self-emp-not-inc," and "Unknown" (NaN).

5. Correlations

a) Overall Correlations

- Education level and weekly working hours showed a weak positive correlation (**r = 0.14**, p-value < 0.001).
- Age correlated weakly with hours-per-week (**r = 0.07**) and education-num (**r = 0.03**).

b) Gender-Specific Correlations

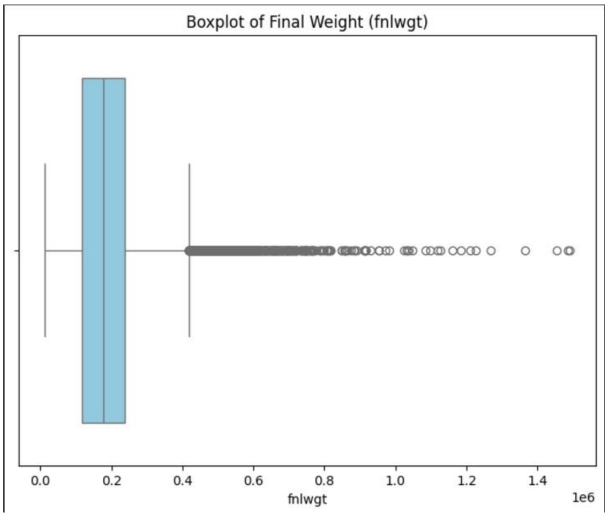
- Males: Education-num and age showed a weak positive correlation (**r = 0.05**, p-value < 0.001).
- Females: Correlation was slightly negative (**r = -0.02**, p-value < 0.01).
- These correlations were statistically significant but of negligible strength.

c) Covariance

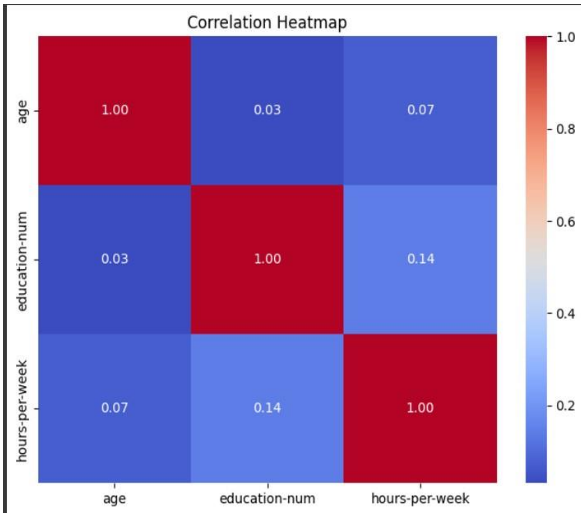
Education and hours-per-week had a covariance of **4.57**, suggesting a modest co-movement.

Visual Insights

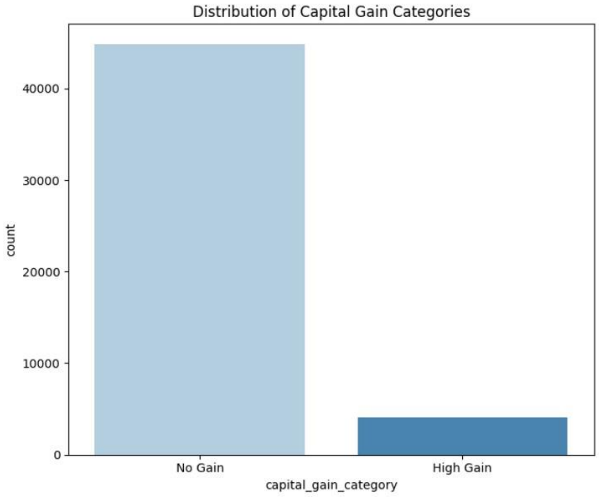
1. **Capital Gain and Loss:** Count plots for categories illustrated their dominance at zero.
2. **Final Weight (fnlwgt):** Histograms highlighted skewness and outliers.
3. **Correlations:** A heatmap showed weak correlations between numerical features.



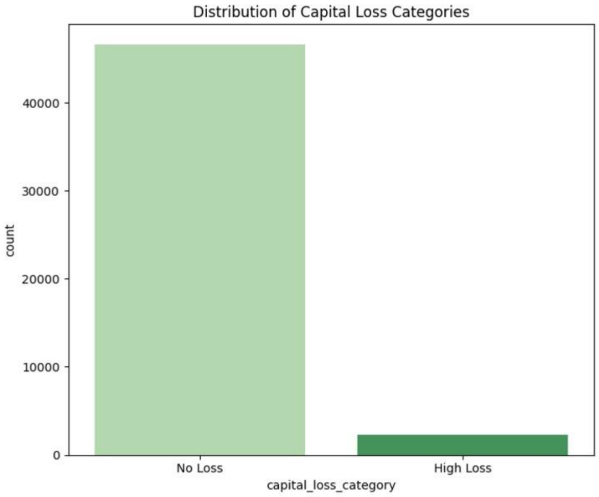
Boxplot of Final Weight (fnlwgt)



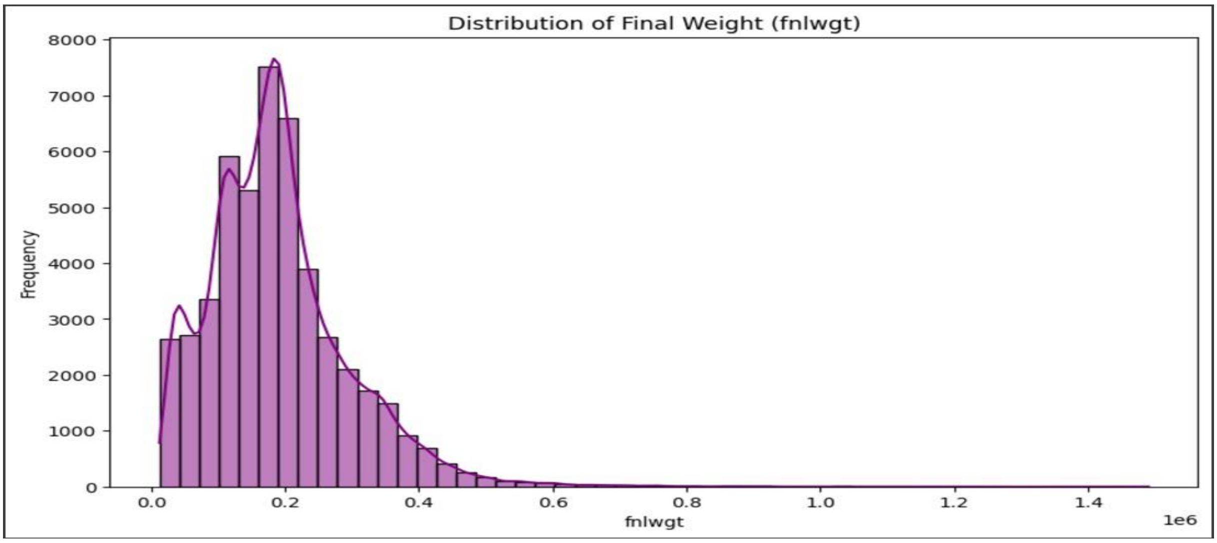
Correlation Heatmap



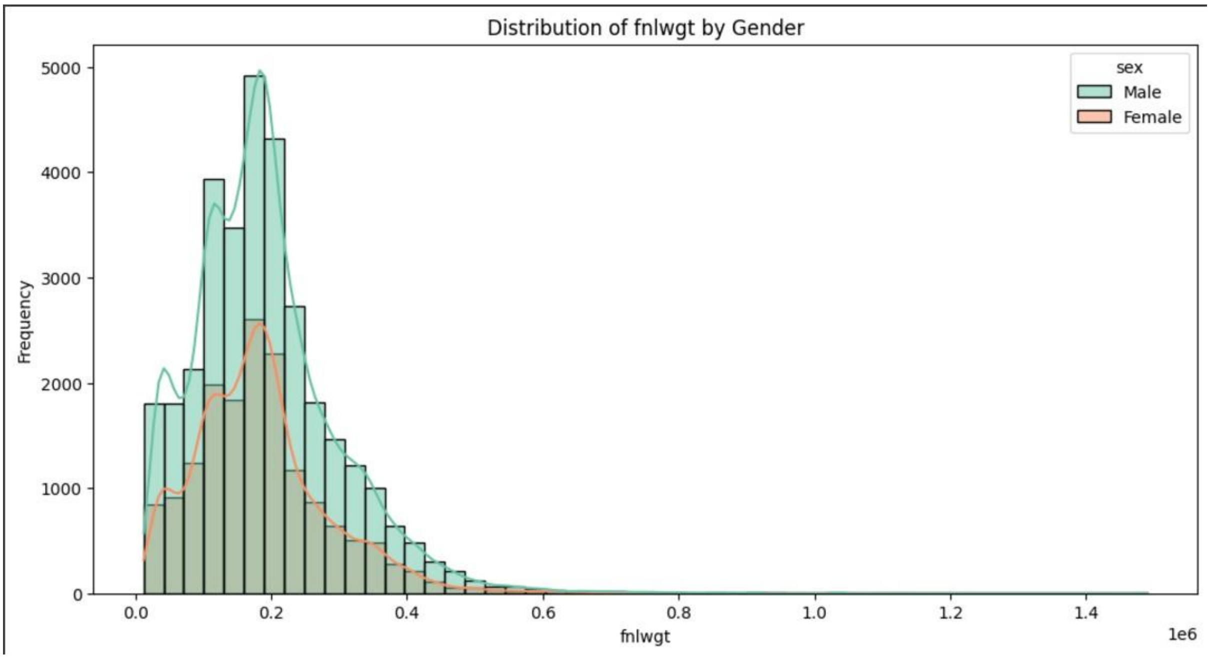
Distribution of Capital Gain Categories

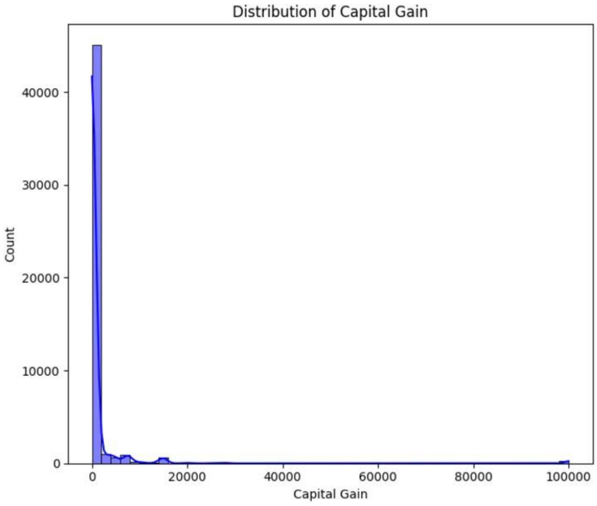


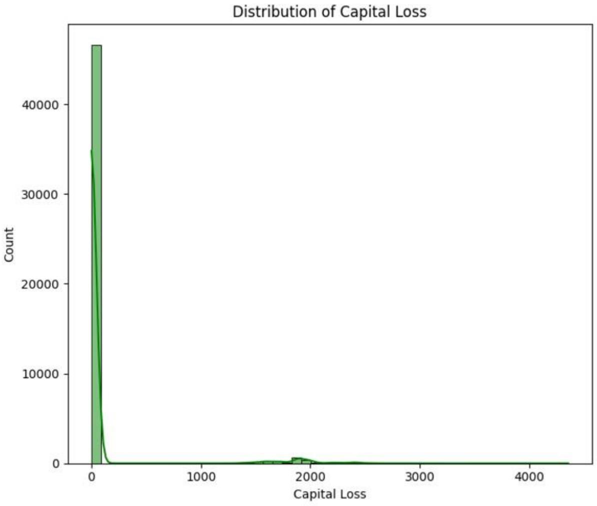
Distribution of Capital Loss Categories



Distribution of Final Weight (fnlwgt)







Issues for Discussion in Analysis

a) Who Works More Hours, Men or Women?

An analysis of hours worked per week by gender reveals:

- **Men:** Median hours per week = **40**, slightly higher than women.
- **Women:** Median hours per week = **37.5**, with greater variability in working hours compared to men.
- Distribution plots show that men are overrepresented in higher work-hour categories.

b) Trend in Hours Worked: Controlling for Education Level

When using **education-num** (years of education) as a control variable:

- The trend of men working more hours than women persist.
- However, the difference narrows slightly when education level is accounted for. This suggests that **education-num** moderates, but does not eliminate, the gender difference in work hours.
- Testing the significance of education-num:
    - A regression analysis shows **education-num** to be statistically significant (**p-value < 0.05**) in predicting hours worked.
    - The positive coefficient indicates that higher education is associated with more hours worked.

Comparing Models:

Model 1 (gender alone) will have the lowest explanatory power (Adjusted R^2) and higher AIC/BIC, as it doesn't account for important variables like education or income.

Model 2 (gender + education-num) will show an improvement in fit (higher Adjusted R^2, lower AIC/BIC), as **education-num** is a significant predictor of hours worked.

Model 3 (gender + education-num + gross_income_group) should show the best performance, with the highest Adjusted R^2 and the lowest AIC/BIC, as **gross_income_group** adds significant explanatory power regarding work h