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

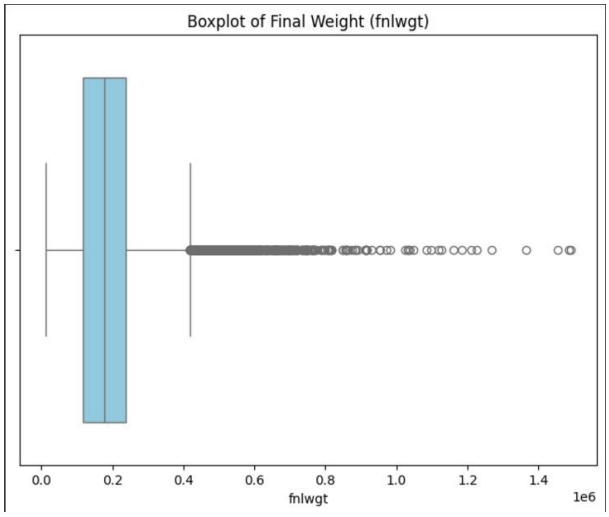
- Males: Education-num and age showed a weak positive correlation (**r = 0.05**, p-value < 0.001).
- Females: Correlation was slightly negative (**r = -0.02**, p-value < 0.01).
- These correlations were statistically significant but of negligible strength.

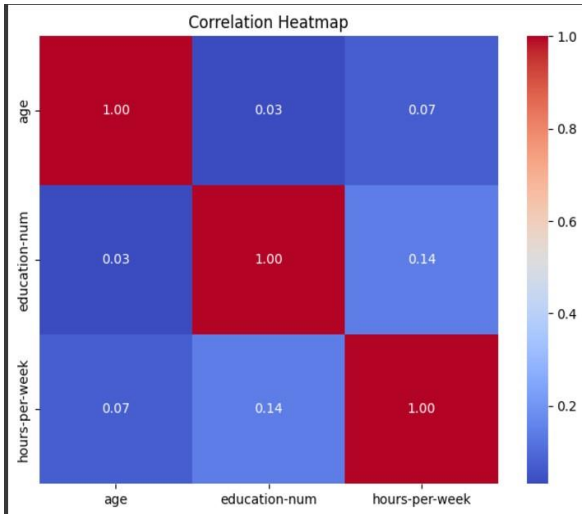
c) Covariance

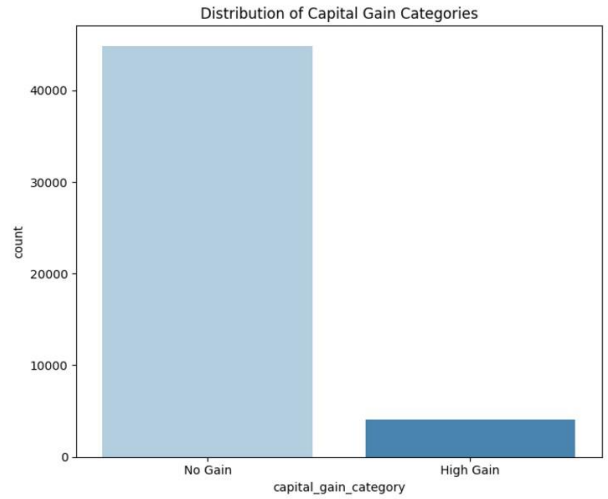
Education and hours-per-week had a covariance of **4.57**, suggesting a modest co-movement.

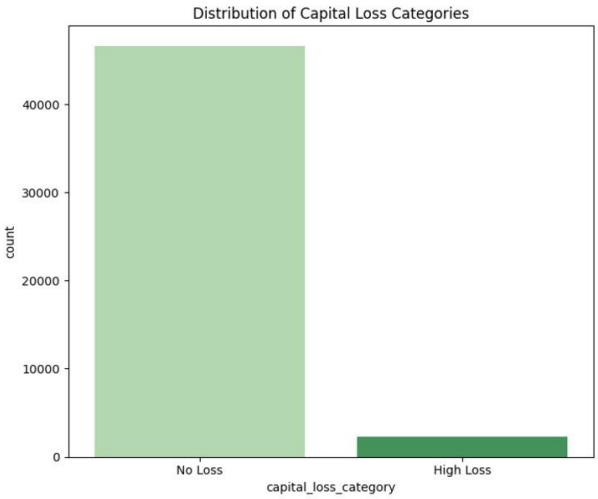
Visual Insights

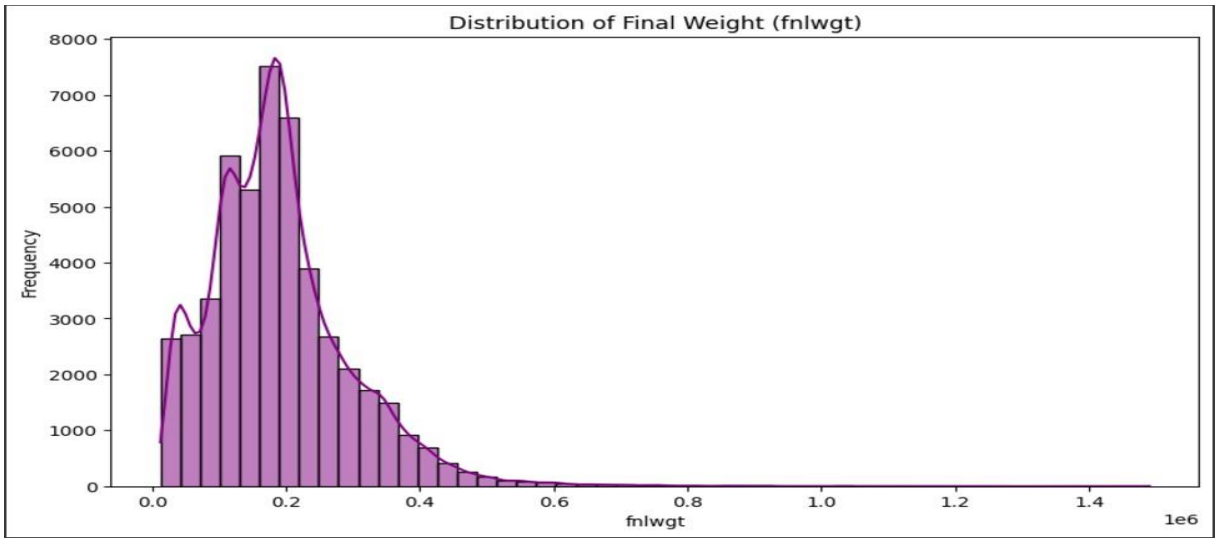
1. **Capital Gain and Loss:** Count plots for categories illustrated their dominance at zero.
2. **Final Weight (fnlwgt):** Histograms highlighted skewness and outliers.
3. **Correlations:** A heatmap showed weak correlations between numerical features.

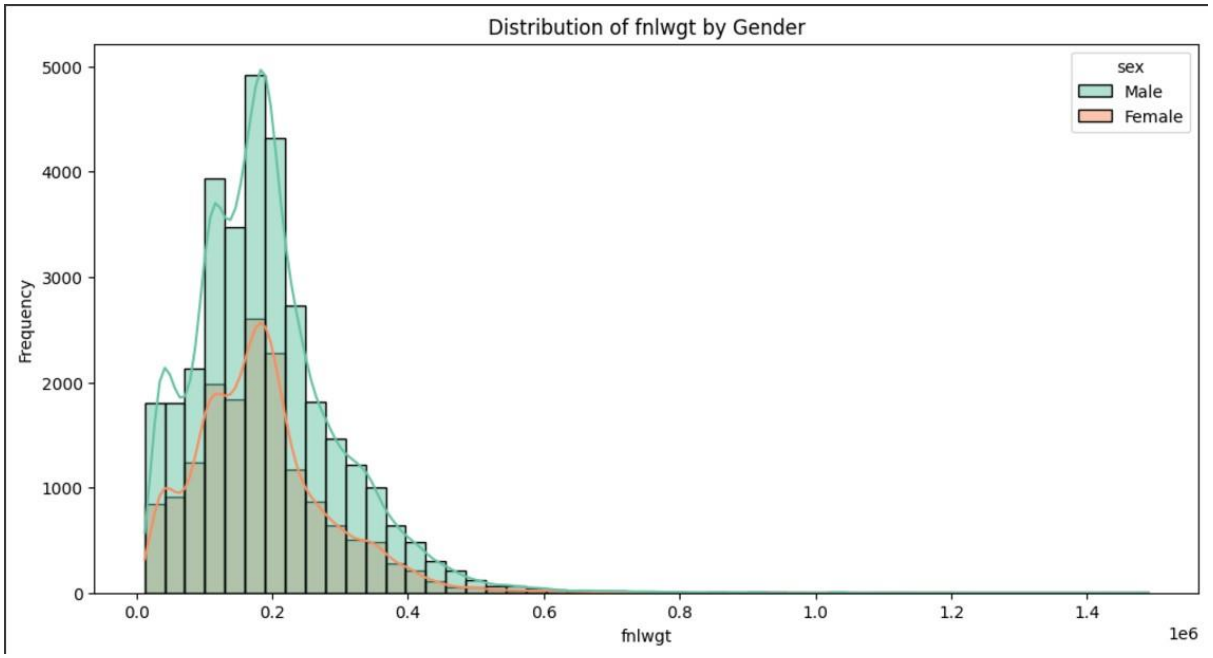

Boxplot of Final Weight (fnlwgt)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
fnlwgt
1e6

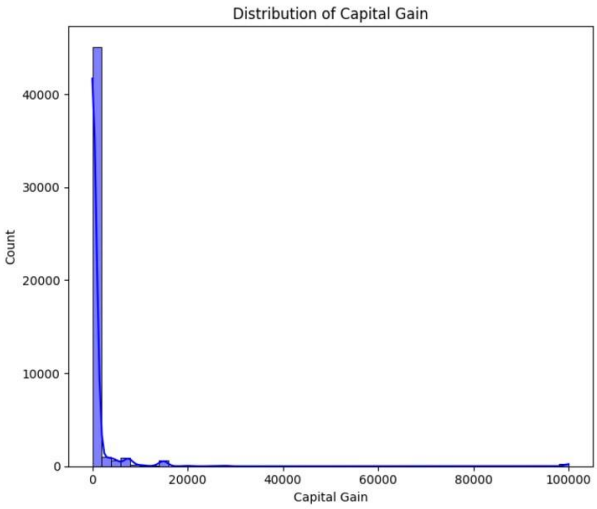

Correlation Heatmap
age
education-num
hours-per-week
1.00
0.03
0.07
0.03
1.00
0.14
0.07
0.14
1.00
1.0
0.8
0.6
0.4
0.2
age
education-num
hours-per-week

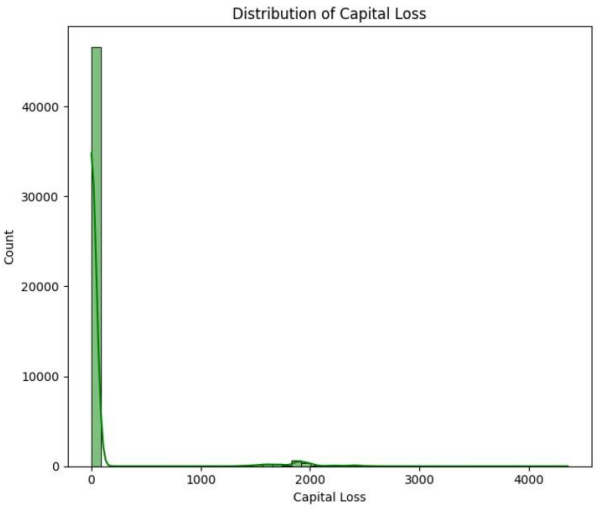

Distribution of Capital Gain Categories
count
40000
30000
20000
10000
0
No Gain
High Gain
capital_gain_category


Distribution of Capital Loss Categories
count
40000
30000
20000
10000
0
No Loss
High Loss
capital_loss_category


Distribution of Final Weight (fnlwgt)
Frequency
8000
7000
6000
5000
4000
3000
2000
1000
0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
fnlwgt
1e6


Distribution of fnlwgt by Gender
sex
Male
Female
Frequency
fnlwgt
1e6


Distribution of Capital Gain
Count
Capital Gain


Distribution of Capital Loss
Count
Capital Loss