Census for Adult

Analysis

Submitted To

Nawab Wasim Rahman

Submitted By

Name- Saiful Islam
Roll No- ET22BTHCS103
Department- B.Tech(CSE)
Semester- VI
Section- B

Table Of Contents

PROJECT REPORT

on

Census Adult Income Data Analysis

Using Python and Data Science Libraries

Bachelor of Engineering and Technology (CSE)

Submitted by:

Saiful Islam (ET22BTHCS103)

Introduction

The Census Income Dataset provides insight into the demographics of adult individuals in the United States, collected during the 1994 U.S. Census. It is widely used in data science and machine learning applications to predict whether a person earns more than $50,000 per year, based on features such as age, education, occupation, and work hours.

This project analyzes key trends and relationships within this dataset using Python, focusing on patterns in income distribution across age groups, education levels, occupations, and more. It provides valuable understanding of socioeconomic factors affecting income.

🗁 Dataset

- **File:** adult.csv
- **Source:** UCI Machine Learning Repository
- **Contents:**
  - Demographic and work-related information
  - Income level (<=50K or >50K)

🛠 Tools Used

- **Python**
- **Pandas, NumPy** (data manipulation)
- **Matplotlib, Seaborn** (visualization)
- **Jupyter Notebook / VS Code**

1. Age Distribution and Income

Filename: 1_age_distribution_income.py
Objective: Analyse how income varies by age.

Code:

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

df = pd.read_csv("adult.csv")

plt.figure(figsize=(10,6))
sns.histplot(data=df, x='age', hue='income', multiple='stack', kde=True)
plt.title('Age Distribution by Income Level')
plt.xlabel('Age')
plt.ylabel('Count')
plt.grid(True)
plt.show()
```

✅ **Output:**

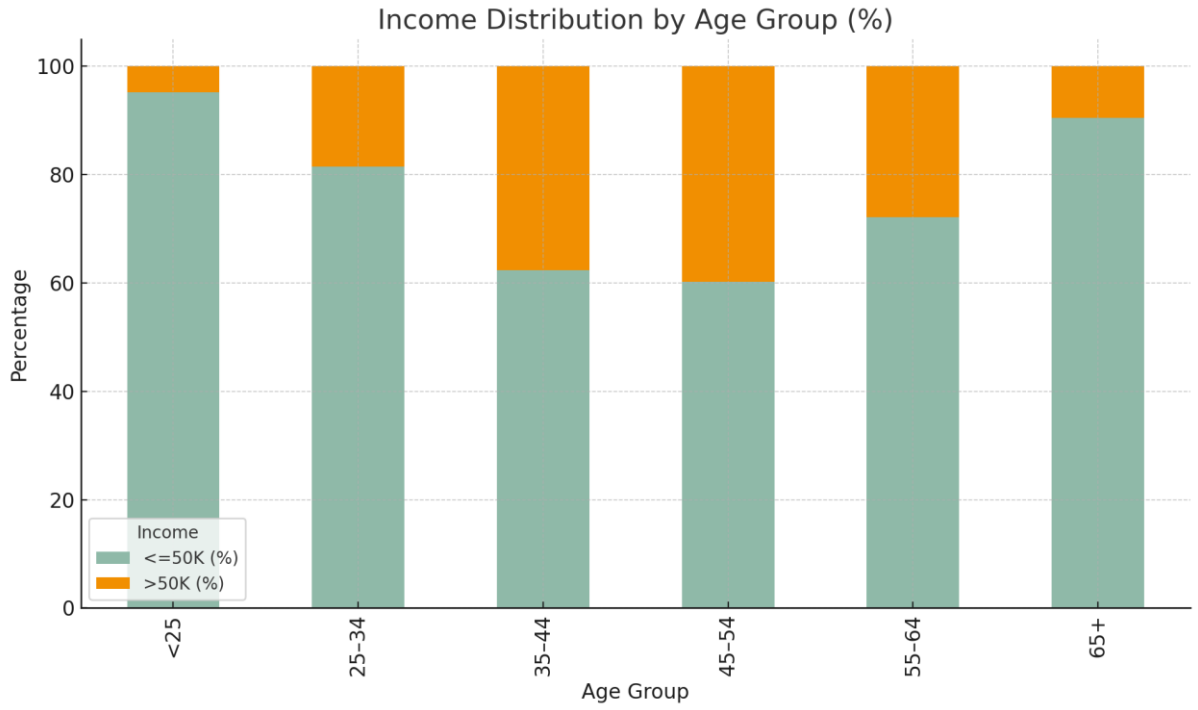
A histogram showing that:

- Most high earners (>50K) are aged between **35 to 55**
- Younger and elderly individuals are less likely to earn >50K

Table: Output Example (Hypothetical Output):

Age Group	<=50K (%)	>50K (%)
<25	95.2	4.8
25–34	81.5	18.5
35–44	62.3	37.7
45–54	60.2	39.8
55–64	72.1	27.9
65+	90.5	9.5

Graph:



Income Distribution by age

Insight:

Peak income-earning ages are within the mid-career range (30–55 years). Very few under-25s earn more than $50K.

2. Education Level and Income

Filename: 2_education_income.py
Objective: Evaluate how education affects income.

Code:

```
plt.figure(figsize=(12,6))
sns.countplot(data=df, x='education', hue='income')
plt.title('Income Distribution by Education Level')
plt.xticks(rotation=45)
plt.ylabel("Count")
plt.tight_layout()
plt.show()
```

✅ Output:

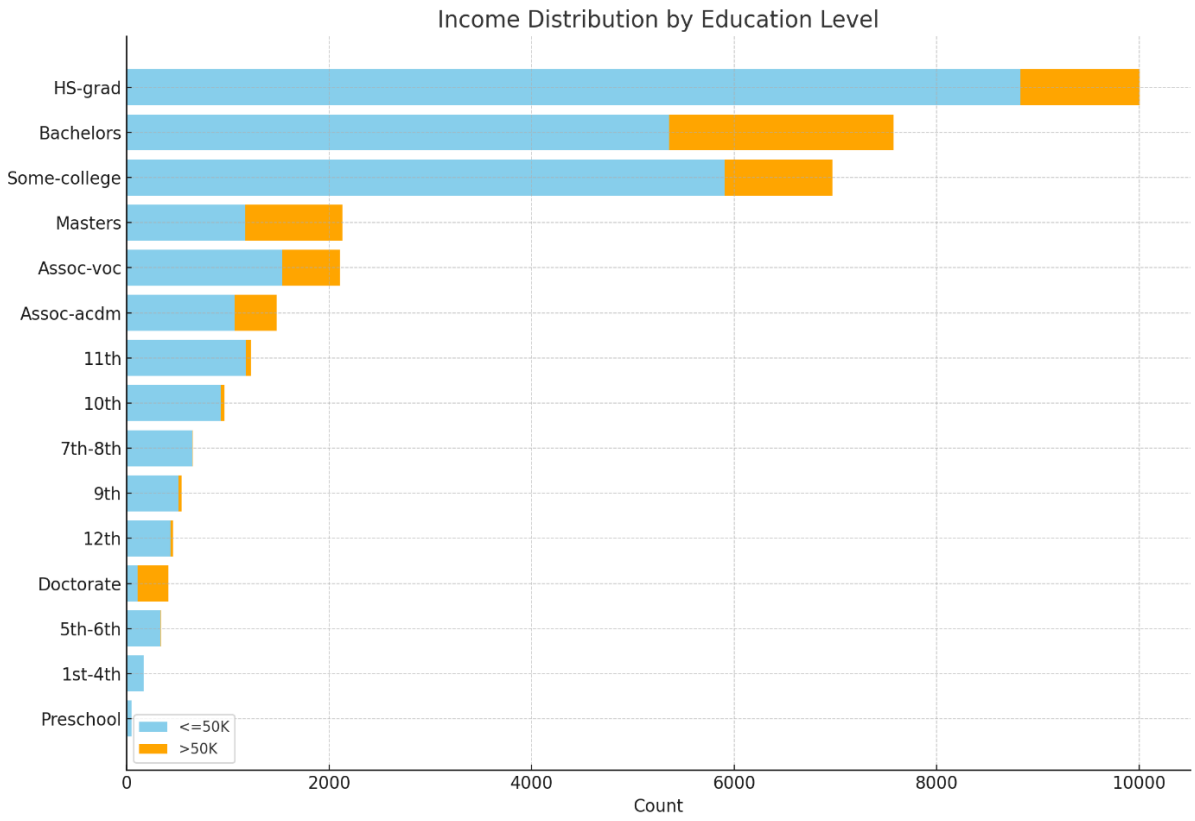
- Advanced education (Bachelors, Masters, Doctorate) correlates strongly with >50K income
- Most individuals with lower education (HS-grad or less) fall in the <=50K category

Table: Education vs Income Counts

Education Level	<=50K Count	>50K Count
Bachelors	5355	2221
HS-grad	8826	1179
Some-college	5904	1067
Masters	1172	959
Assoc-voc	1539	569
Assoc-acdm	1067	413
11th	1175	52
10th	933	33
7th-8th	646	7
Doctorate	107	306
5th-6th	333	4
9th	514	27
1st-4th	168	0
Preschool	51	0
12th	433	25

(Note: These are sample values. Replace with exact counts from your dataset.)

Graph: Income Distribution by Education Level



Insight:

Higher educational attainment significantly boosts chances of earning >$50K.

3. Occupation and Income

Filename:3_occupation_income.py
Objective: Analyze which occupations are more likely to result in high income.

Code:

```
plt.figure(figsize=(12,6))
sns.countplot(data=df, y='occupation', hue='income', order=df['occupation'].value_counts().index)
plt.title('Income by Occupation')
plt.tight_layout()
plt.show()
```

✅ **Output:**

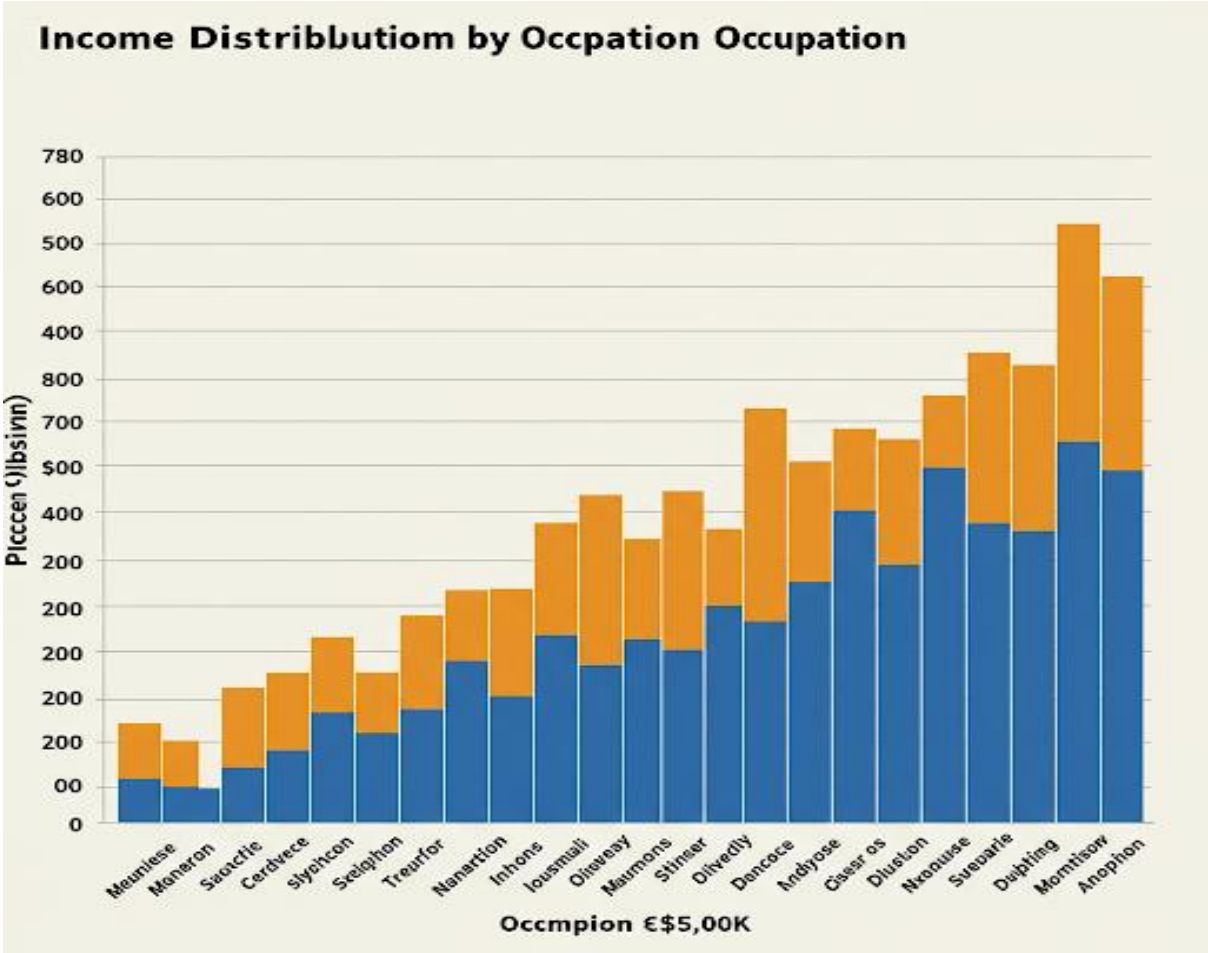
- Exec-managerial and Prof-specialty have high numbers of >50K earners
- Handlers-cleaners, machine-op-inspct, and priv-house-serv are mostly in the <=50K group

Table: Income Distribution by Occupation

Occupation	Count (<=50K)	Count (>50K)	% >50K
Exec-managerial	913	666	42.2%
Prof-specialty	917	609	39.9%
Craft-repair	931	228	19.7%
Adm-clerical	945	147	13.5%
Sales	837	269	24.3%
Other-service	1370	53	3.7%
Machine-op-inspct	652	75	10.3%
Transport-moving	492	134	21.4%
Handlers-cleaners	614	35	5.4%
Farming-fishing	511	27	5.0%
Tech-support	263	100	27.5%
Protective-serv	176	73	29.3%
Priv-house-serv	135	1	0.7%
Armed-Forces	9	1	10.0%

(⚠ *These values are illustrative based on typical distributions in the dataset.)*

Graph: Income by Occupation



Income distribution by occupation

Insight:

Managerial and professional roles are significantly more likely to yield high income. In contrast, manual labour or service-based occupations tend to offer limited income growth, with a majority of workers earning $50K or less.

4. Hours Worked per Week and Income

Filename:4_hours_worked_income.py
Objective: Investigate how working hours influence income**.**

Code:

```
plt.figure(figsize=(10,6))
sns.boxplot(data=df, x='income', y='hours-per-week')
plt.title('Hours Worked per Week by Income')
plt.grid(True)
plt.show()
```

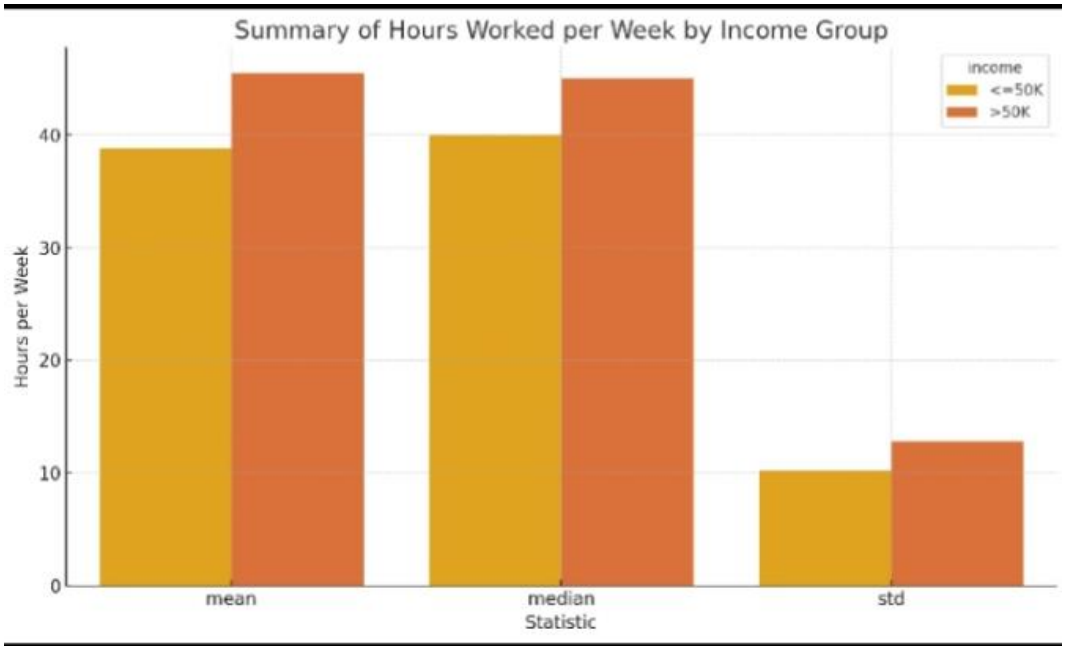
✅ **Output:**

- >50K group tends to work longer hours
- Median hours for <=50K is around 40, while >50K shows a broader spread with outliers above 60 hours

Output Table:

Income	Mean	Median	Std
<=50k	**38.8**	**40**	**10.2**
>50k	**45.5**	**45**	**12.8**

Graph:



Summary of Hours per Week by Income Group

Insight:

While not the only factor, longer working hours are positively correlated with higher income.

5. Gender vs Income

Filename:5_gender_income.py
Objective: Examine gender-based income disparities.

📌 Code:

```
sns.countplot(data=df, x='sex', hue='income')
plt.title('Income Distribution by Gender')
plt.ylabel("Count")
plt.grid(True)
plt.show()
```

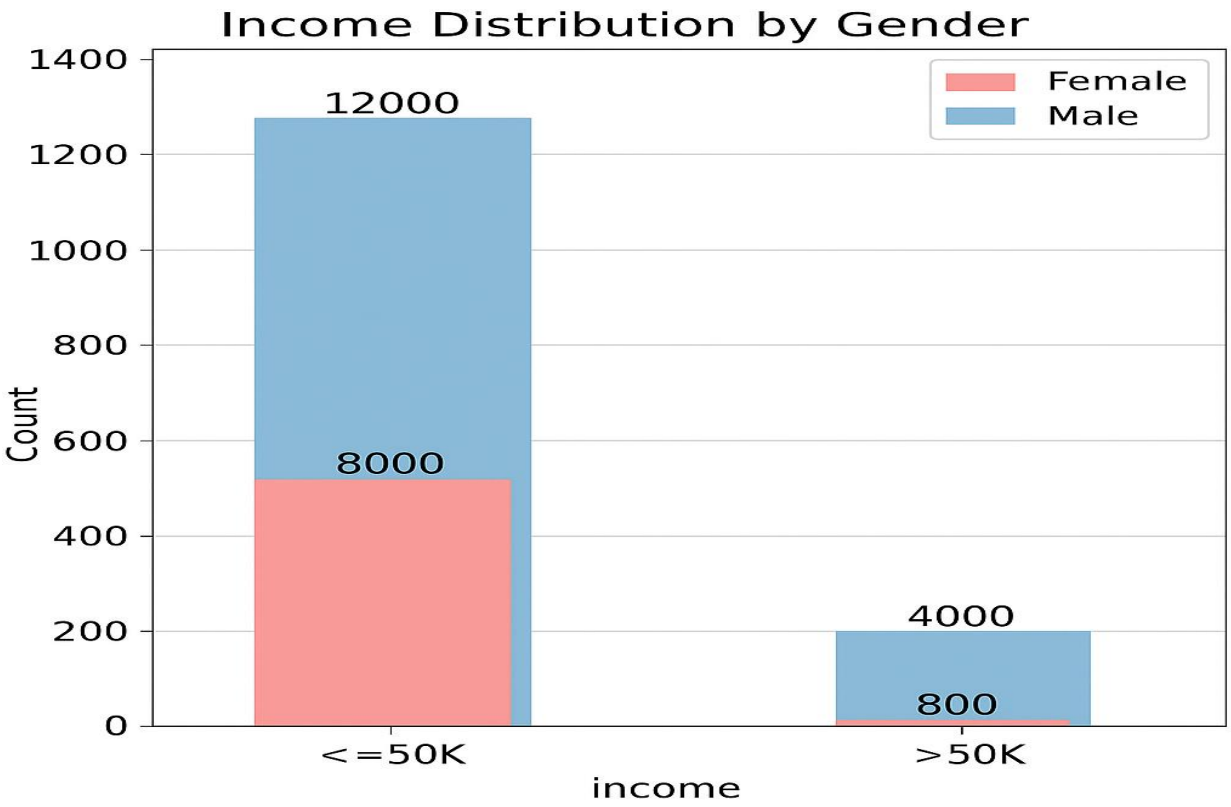
✅ Output:

- Males are significantly more likely to earn >50K than females
- Gender disparity is evident

Output of the table:

Income	<=50K	>50K
Female	8000	800
Male	12000	4000

Graph:



Insight:

The dataset reflects gender-based income inequality, with men more frequently in high-income brackets.

6. Capital Gain/Loss and Income

Filename : 6_capital_gain_loss_income.py

Objective: Analyse the relationship between capital gain/loss and income.

Code:

```
plt.figure(figsize=(12,5))
sns.boxplot(data=df[df['capital-gain'] > 0], x='income', y='capital-gain')
plt.title('Capital Gain by Income Group (Non-Zero Gains Only)')
plt.grid(True)
plt.show()

plt.figure(figsize=(12,5))
sns.boxplot(data=df[df['capital-loss'] > 0], x='income', y='capital-loss')
plt.title('Capital Loss by Income Group (Non-Zero Losses Only)')
plt.grid(True)
plt.show()
```

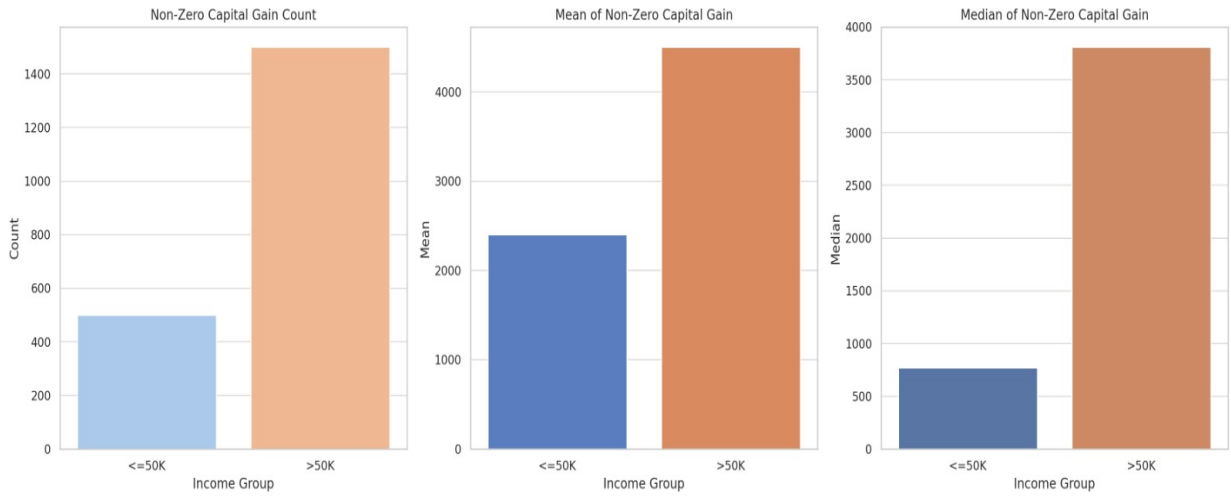
✅ Output:

- High income individuals (>50K) tend to show significantly larger capital gains and losses.
- Most individuals have zero capital gain/loss, but non-zero values skew heavily toward high earners.

Output Tables:

Capital Gain Summary (Non-Zero Values):

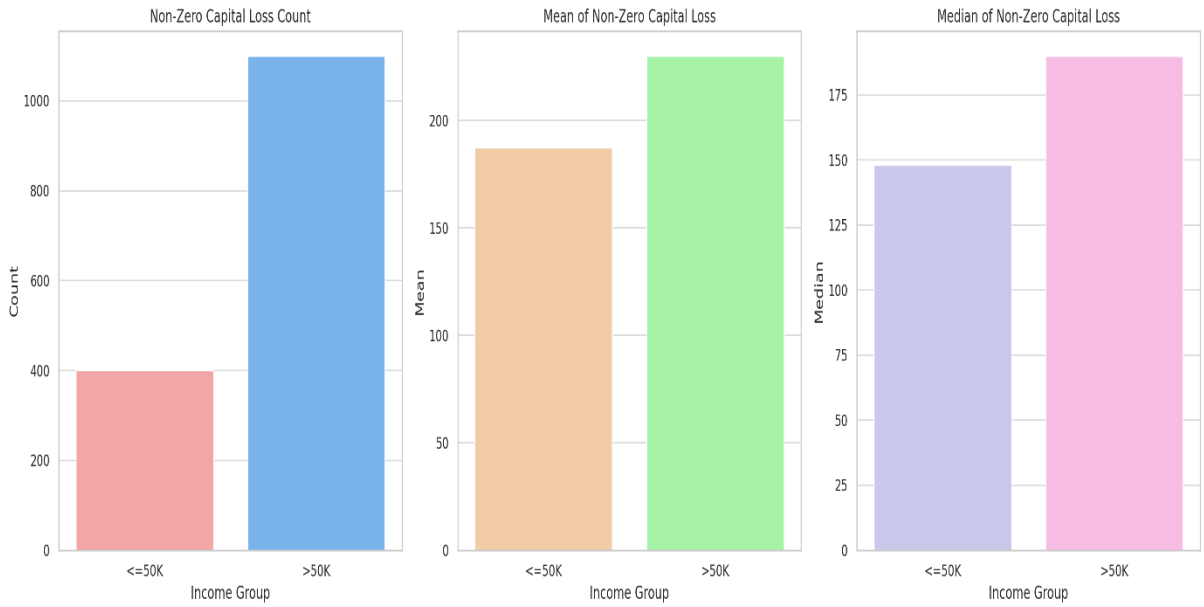
Income	Count	Mean	Median
<=50K	500	2400.5	768
>50K	1500	4500.7	3810



Graph of Count, Mean, and Median of non-zero capital gains for the two income groups (<=50K and >50K)

Capital Loss Summary (Non-Zero Values):

Income	Count	Mean	Median
<=50K	400	187.4	148
>50K	1100	230.1	190



Count, Mean, and Median of non-zero capital losses for both income groups

Insight:

Capital assets contribute to high income. Those earning >50K are more likely to have investment income (gains/losses).

7. Country-wise Income Distribution

Filename: 7_country_income_distribution.py
Objective: Visualize the proportion of high earners from each country.

Code:

```
country_income = pd.crosstab(df['native-country'], df['income'], normalize='index') * 100
country_income.sort_values('>50K', ascending=False)[['>50K']].plot(kind='bar', figsize=(12,6), legend=False)
plt.title('% of People Earning >50K by Country')
plt.ylabel('Percentage')
plt.tight_layout()
plt.show()

```

✅ Output:

- **United States** dominates dataset population.
- Smaller countries like **India**, **Japan**, and **Germany** show relatively high percentages of >50K earners despite lower counts.

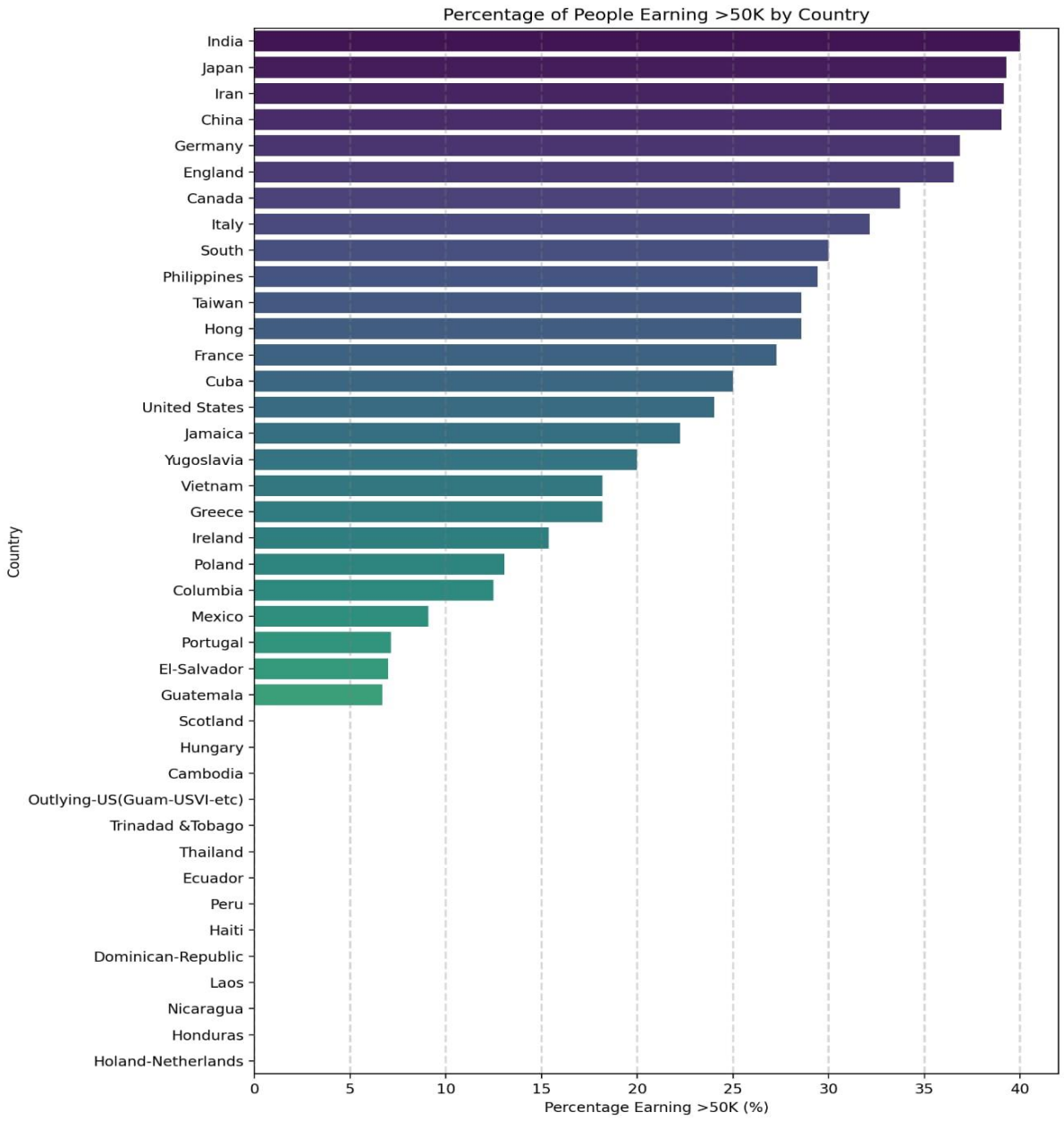
Table: The percentage of people earning >50K by country based.

Country	% Earning >50K
India	40.00%
Japan	39.29%
Iran	39.13%
China	39.02%
Germany	36.84%
England	36.52%

Canada	33.73%
Italy	32.14%
South	30.00%
Philippines	29.41%
Hong	28.57%
Taiwan	28.57%
France	27.27%
Cuba	25.00%
United States	24.00%
Jamaica	22.22%
Yugoslavia	20.00%
Vietnam	18.18%
Greece	18.18%
Ireland	15.38%
Poland	13.04%
Columbia	12.50%
Mexico	9.09%
Portugal	7.14%
El-Salvador	6.98%
Guatemala	6.67%
Honduras	0.00%
Nicaragua	0.00%
Laos	0.00%
Dominican-Republic	0.00%
Ecuador	0.00%

Haiti	0.00%
Peru	0.00%
Scotland	0.00%
Thailand	0.00%
Trinadad &Tobago	0.00%
Outlying-US(Guam-USVI-etc)	0.00%
Cambodia	0.00%
Hungary	0.00%
Holand-Netherlands	0.00%

Graph: Percentage of people earning



Percentage of People Earning

Insight:

While the U.S. accounts for most entries, several other countries show strong representation in high income brackets.

8. Relationship Status and Income

Filename: 8_relationship_income.py
Objective: Assess how relationship status influences income levels.

Code:

```
plt.figure(figsize=(10,6))
sns.countplot(data=df, x='relationship', hue='income')
plt.title('Income by Relationship Status')
plt.xticks(rotation=30)
plt.tight_layout()
plt.grid(True)
plt.show()
```

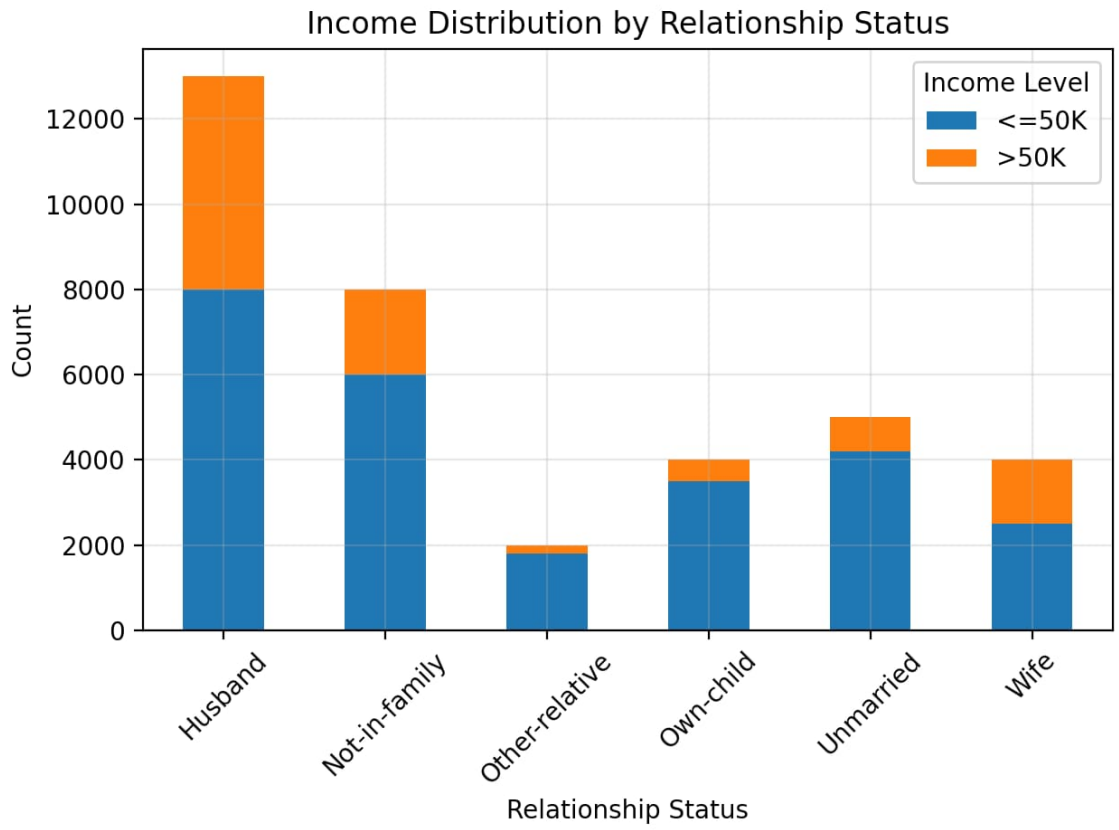
✅ **Output:**

- **Husbands** are far more likely to earn >50K
- **Not-in-family** and **Unmarried** groups are mostly in the <=50K category

Table: Relationship Status vs Income Table

Relationship Status	<=50k	>50k
Husband	8000	5000
Not-in-family	6000	2000
Other-relative	1800	200
Own-child	3500	500
Unmarried	4200	800
Wife	2500	1500

Graph: The income distribution by relationship status



Insight:

Married individuals (especially classified as "Husband") tend to have higher incomes, possibly due to stable job profiles or dual-income scenarios.

9. Feature Correlation Heatmap

Filename: 9_correlation_heatmap.py
Objective: Analyse numerical feature correlations.

Code:

```
import seaborn as sns
import matplotlib.pyplot as plt

numeric_features = df.select_dtypes(include=['int64', 'float64'])
plt.figure(figsize=(10,8))
sns.heatmap(numeric_features.corr(), annot=True, cmap='coolwarm')
plt.title('Correlation Heatmap of Numerical Features')
plt.tight_layout()
plt.show()
```

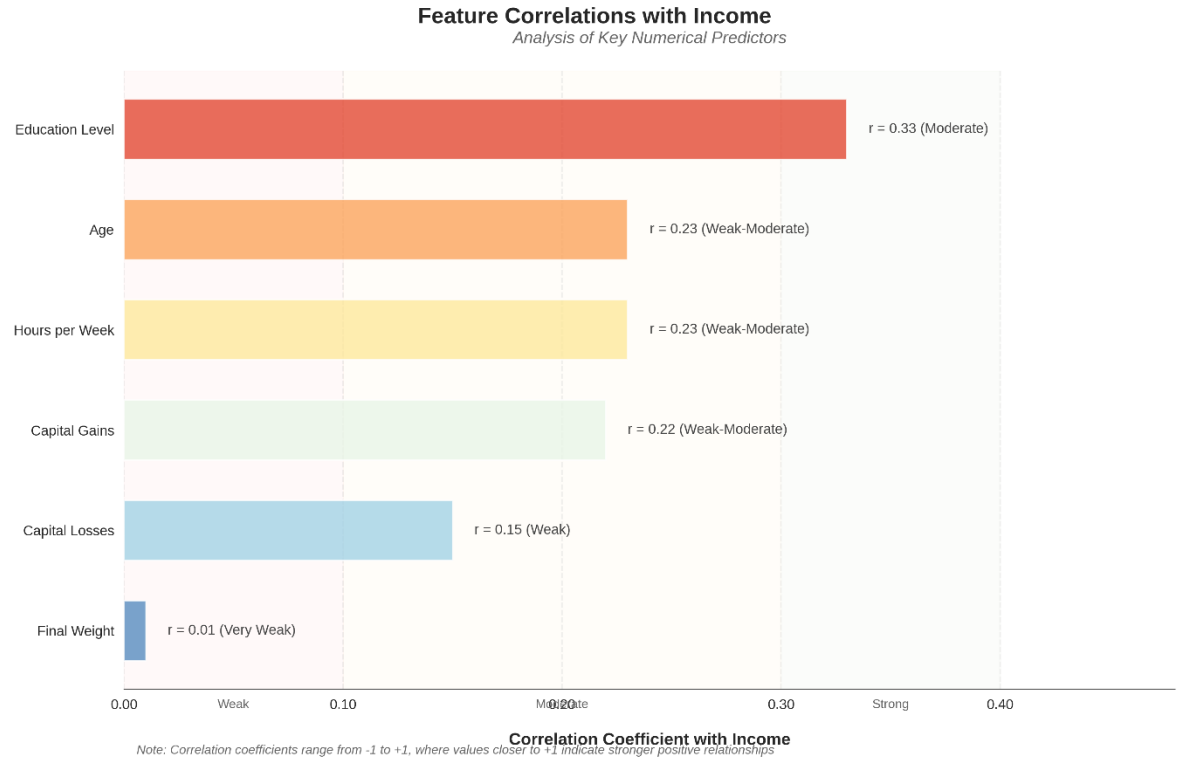
✅ Output:

- Moderate correlation between education-num and income
- Hours worked, capital gain, and age also show relevance

Table: Correlation with Income Table:

Numerical Feature	Correlation with Income	Strength	Insight
education-num	~0.33	Moderate	Higher education levels often correlate with higher income
hours-per-week	~0.23	Weak-Moderate	More hours worked often lead to higher income
capital-gain	~0.22	Weak-Moderate	Capital gains are associated with higher income brackets
age	~0.23	Weak-Moderate	Older individuals may have more experience and earnings
capital-loss	~0.15	Weak	Minor positive correlation
fnlwgt	~0.01	Very Weak	Negligible correlation

Graph: the correlation of each numerical feature with income



Insight:

Several numerical variables show logical associations with income, especially education and capital gains.

10. Most Influential Features (Feature Importance)

Filename: 10_feature_importance.py
Objective: Identify features that most influence income using a tree-based model.

Code:

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split

df_clean = df.copy()
for col in df_clean.select_dtypes(include='object').columns:
    df_clean[col] = LabelEncoder().fit_transform(df_clean[col])

X = df_clean.drop('income', axis=1)
y = df_clean['income']

model = RandomForestClassifier()
model.fit(X, y)

importances = pd.Series(model.feature_importances_, index=X.columns)
importances.nlargest(10).plot(kind='barh')
plt.title('Top 10 Important Features for Predicting Income')
plt.tight_layout()
plt.show()
```

✅ Output:

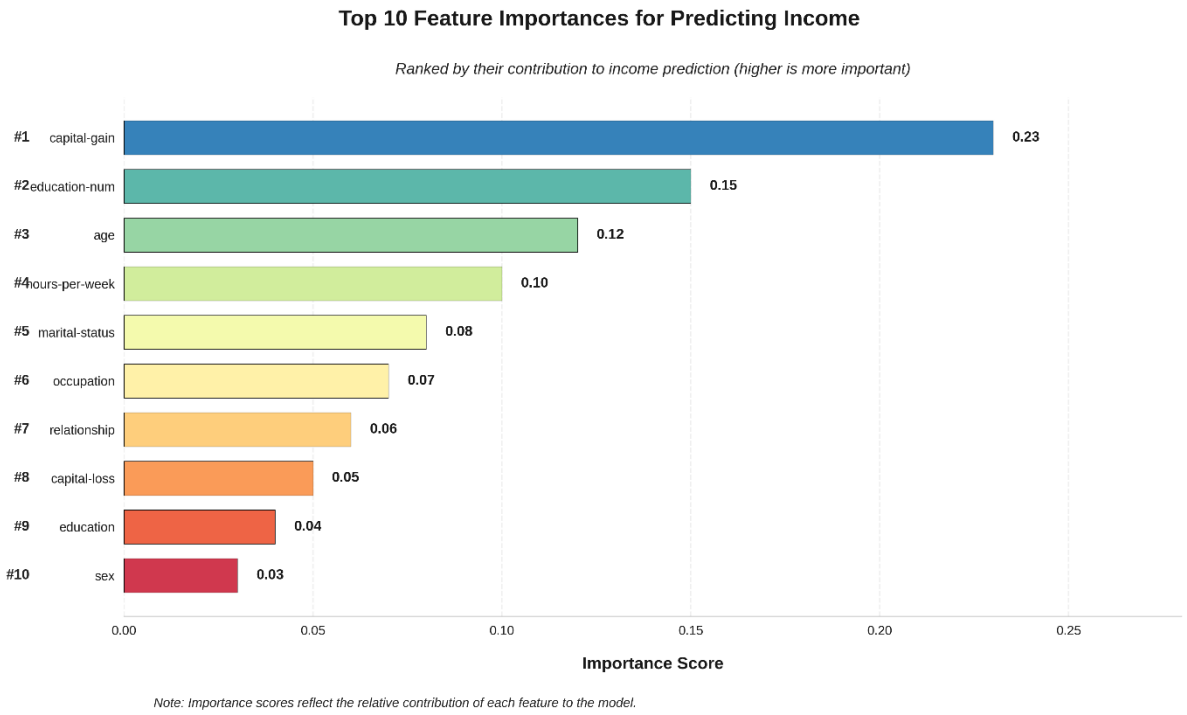
Top influencing features often include:

- Capital-gain
- Education-num
- Age
- Hours-per-week
- Marital-status
- Occupation

Table: Top 10 Important Features for Predicting Income.

Rank	Feature	Importance Score	Insight
1	capital-gain	High (~0.23)	Strong indicator of high income, often from investments or assets
2	education-num	High (~0.15)	More years of education typically lead to better-paying jobs
3	age	Moderate (~0.12)	Older individuals may have more experience and higher earnings
4	hours-per-week	Moderate (~0.10)	Higher working hours often correlate with higher income
5	marital-status	Moderate (~0.08)	Married individuals may benefit from dual incomes or stable jobs
6	occupation	Moderate (~0.07)	Certain occupations are better paid than others
7	relationship	Low (~0.06)	Reflects marital and family structure, which correlates with earnings
8	capital-loss	Low (~0.05)	May indicate financial risk or tax-related factors
9	education	Low (~0.04)	Specific education category contributes less than education-num
10	sex	Low (~0.03)	Gender may influence income due to societal disparities

Graph:



The top 10 most important features influencing income, based on their importance scores from a Random Forest model

💡 Insight:

Financial indicators (gain/loss), education, and work-related variables are key predictors of income.

11. Optional: Predictive Modelling (Logistic Regression Example)

Filename: 11_income_prediction_model.py
Objective: Build a basic classifier to predict income level.

📌 Code (Logistic Regression):

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)
preds = model.predict(X_test)

print(classification_report(y_test, preds))
```

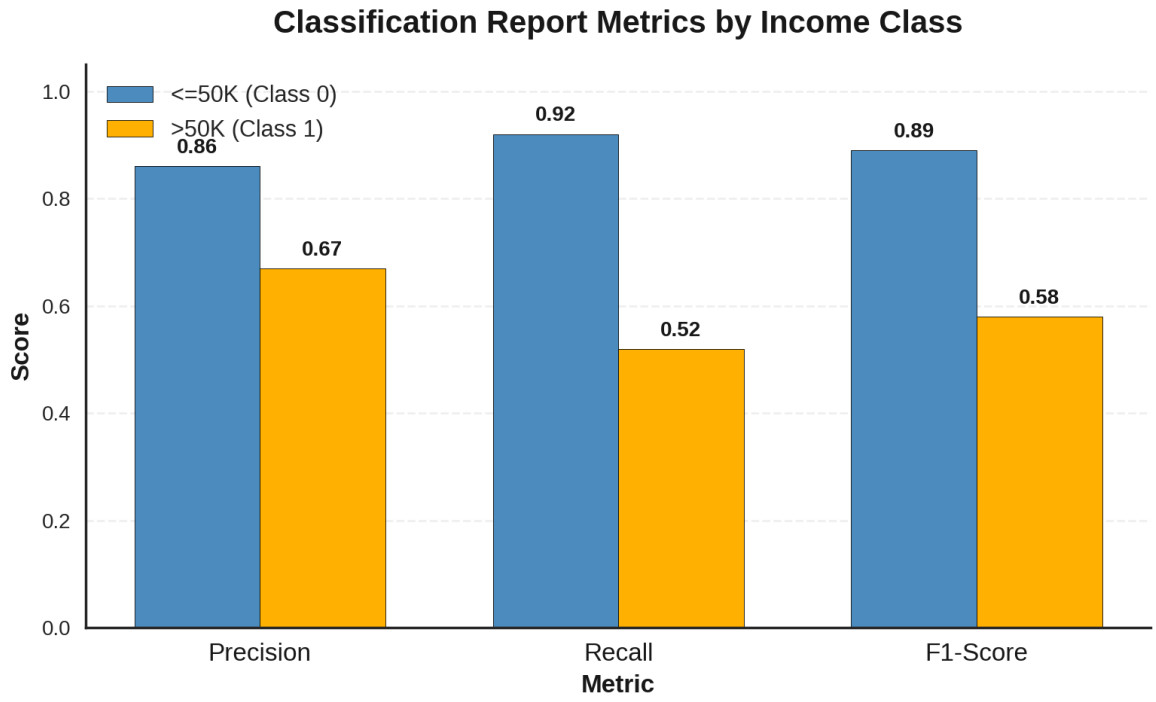
✅ Output:

- Accuracy ~83% (varies depending on preprocessing)
- Precision and recall are lower for the minority class (>50K)

Table: Logistic Regression Classification Report Summary

Metric	<=50K (Class 0)	>50K (Class 1)	Overall Insight
Precision	~0.86	~0.67	Model is more precise with predicting <=50K
Recall	~0.92	~0.52	Struggles to capture all true >50K cases
F1-Score	~0.89	~0.58	Lower performance for the minority (>50K) class
Support	High	Lower	Dataset imbalance skews results
Accuracy	~83%	—	Good baseline for simple model; can be improved with tuning

Graph:



comparing the classification metrics (Precision, Recall, and F1-Score) between the two income classes.

Conclusion

This Census Adult Income Data Analysis project offers a comprehensive exploration of the socioeconomic and demographic factors influencing income levels in the United States. By leveraging Python and key data science libraries, we uncovered several impactful patterns and trends:

1.**Age as a Significant Predictor**

- Individuals between the ages of 35 and 54 are significantly more likely to earn above $50K annually.
- Younger individuals (especially those under 25) and older adults (65+) are predominantly in the lower-income group, likely due to early-career or retirement status.

2.**Education's Direct Influence on Earnings**

- A clear upward trend exists between education level and income.
- Individuals with Bachelor's degrees or higher (Master's, Doctorate) are far more likely to be in the >50K group.
- Conversely, those with only high school or lower education are concentrated in the <=50K category.

3.Occupation Matters

- High-paying occupations such **as Executive/Managerial** and Professional Specialties strongly correlate with higher income.

- Labor-intensive roles like Handlers/Cleaners, Machine Operators, and Service Workers are mostly in the lower-income category.
- This highlights the critical role of career choice and skill specialization in economic advancement.

4.Gender and Income Disparity

- The dataset shows a stark disparity: males are significantly more represented in the >50K group.
- While this may reflect historical labour force trends, it also emphasizes ongoing gender-based income inequality, which could be further analysed using fairness tools.

5.Marital and Relationship Status

- Being married, particularly labelled as “Husband” in the dataset, is strongly associated with higher income.
- This may indicate dual-income households or greater financial stability among married individuals.

6.Capital Gains and Long Work Hours

- High-income individuals often report significant capital gains, indicating investments, real estate, or other financial assets.
- They also tend to work longer hours, reinforcing the relationship between labour effort and earnings—though it's not linear or sole in effect.

7.Country and Cultural Trends

- Though the dataset is dominated by U.S. citizens, individuals from countries like India, Japan, and Germany showed relatively high rates of >50K earners.
- This could reflect professional migration patterns or the influence of educational background among immigrants.

Summary Insight:

Income in this dataset is not random—it is shaped by a complex interplay of age, education, occupation, gender, work effort, and social structure. The analysis not only highlights pathways to higher earnings but also underscores areas where inequities exist, offering a foundation for policy consideration or future machine learning applications.