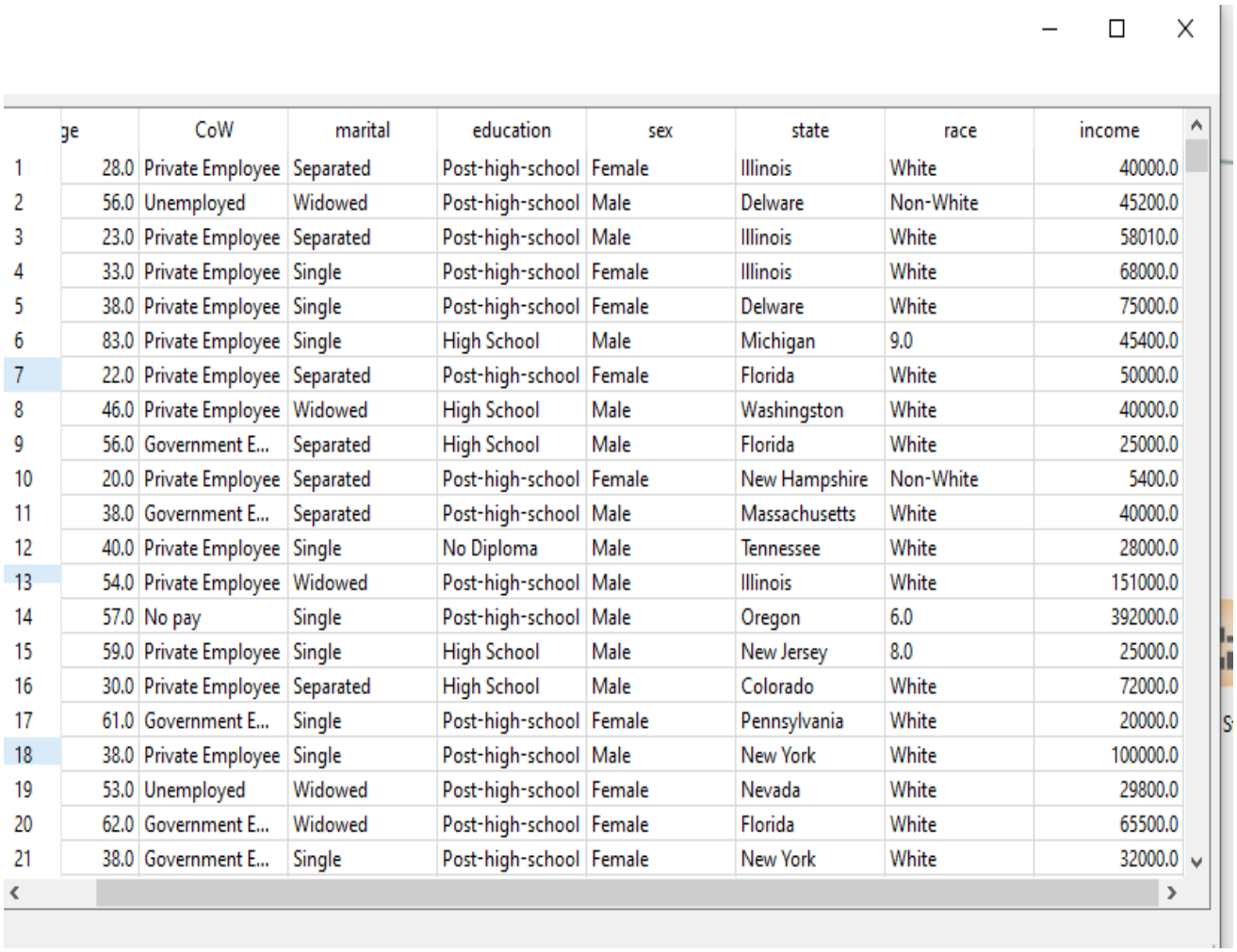
Report for Part 1: Pre-processing

1. Introduction

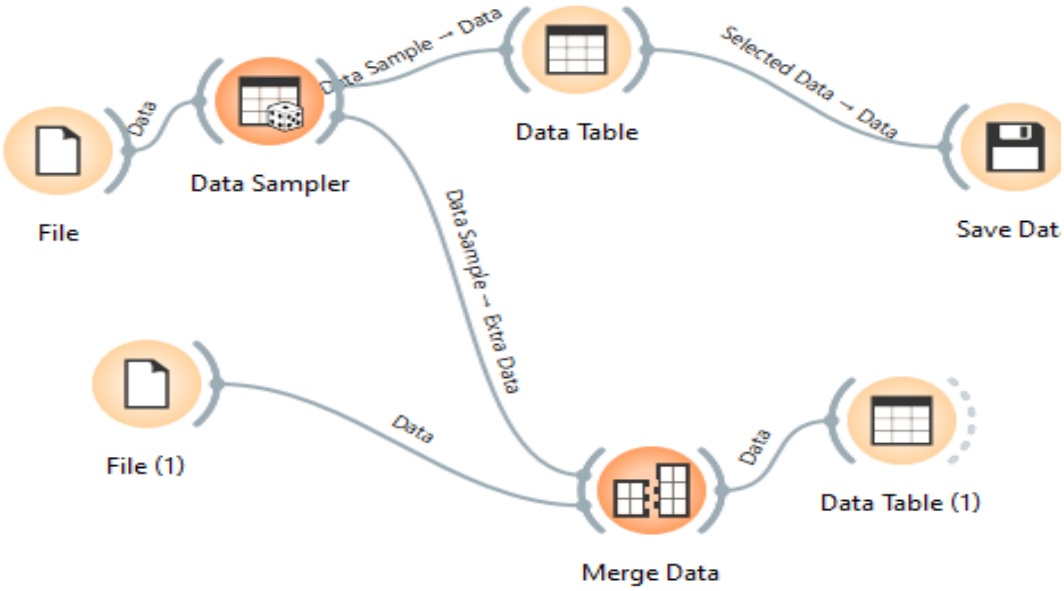
The dataset provided contains census data from the United States, including demographic, occupational, and income-related attributes. Before conducting further analysis, data pre-processing is necessary to improve readability and handle any inconsistencies. This section details the pre-processing steps taken, including data transformation, missing value handling, and exploratory analysis.

2. Data Reduction and Cleaning

The original dataset contained over a million records, which were reduced to 5,000 to improve computational efficiency. Numeric codes representing categorical variables (e.g., category of work, marital status, sex, race, and place of birth) were replaced with meaningful labels for better interpretability. Additionally, state codes were renamed for clarity.



	ge	CoW	marital	education	sex	state	race	income
1	28.0	Private Employee	Separated	Post-high-school	Female	Illinois	White	40000.0
2	56.0	Unemployed	Widowed	Post-high-school	Male	Delware	Non-White	45200.0
3	23.0	Private Employee	Separated	Post-high-school	Male	Illinois	White	58010.0
4	33.0	Private Employee	Single	Post-high-school	Female	Illinois	White	68000.0
5	38.0	Private Employee	Single	Post-high-school	Female	Delware	White	75000.0
6	83.0	Private Employee	Single	High School	Male	Michigan	9.0	45400.0
7	22.0	Private Employee	Separated	Post-high-school	Female	Florida	White	50000.0
8	46.0	Private Employee	Widowed	High School	Male	Washingston	White	40000.0
9	56.0	Government E...	Separated	High School	Male	Florida	White	25000.0
10	20.0	Private Employee	Separated	Post-high-school	Female	New Hampshire	Non-White	5400.0
11	38.0	Government E...	Separated	Post-high-school	Male	Massachusetts	White	40000.0
12	40.0	Private Employee	Single	No Diploma	Male	Tennessee	White	28000.0
13	54.0	Private Employee	Widowed	Post-high-school	Male	Illinois	White	151000.0
14	57.0	No pay	Single	Post-high-school	Male	Oregon	6.0	392000.0
15	59.0	Private Employee	Single	High School	Male	New Jersey	8.0	25000.0
16	30.0	Private Employee	Separated	High School	Male	Colorado	White	72000.0
17	61.0	Government E...	Single	Post-high-school	Female	Pennsylvania	White	20000.0
18	38.0	Private Employee	Single	Post-high-school	Male	New York	White	100000.0
19	53.0	Unemployed	Widowed	Post-high-school	Female	Nevada	White	29800.0
20	62.0	Government E...	Widowed	Post-high-school	Female	Florida	White	65500.0
21	38.0	Government E...	Single	Post-high-school	Female	New York	White	32000.0



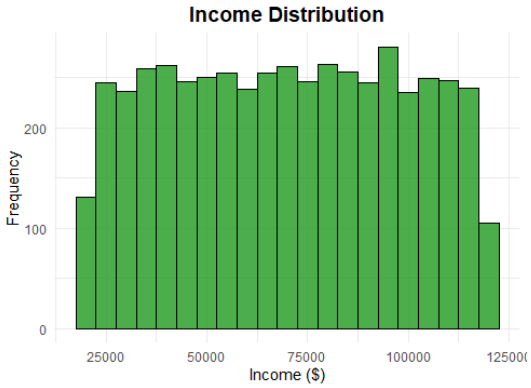
3. Handling Missing Data

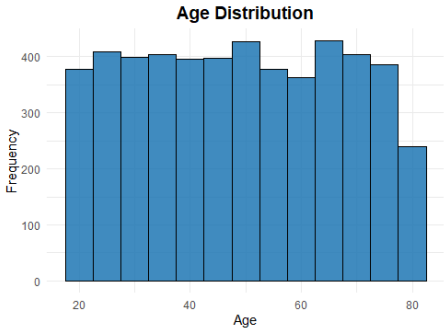
A preliminary inspection revealed no significant missing values. If missing data had been detected, the approach would have involved either removal or imputation based on the attribute type.

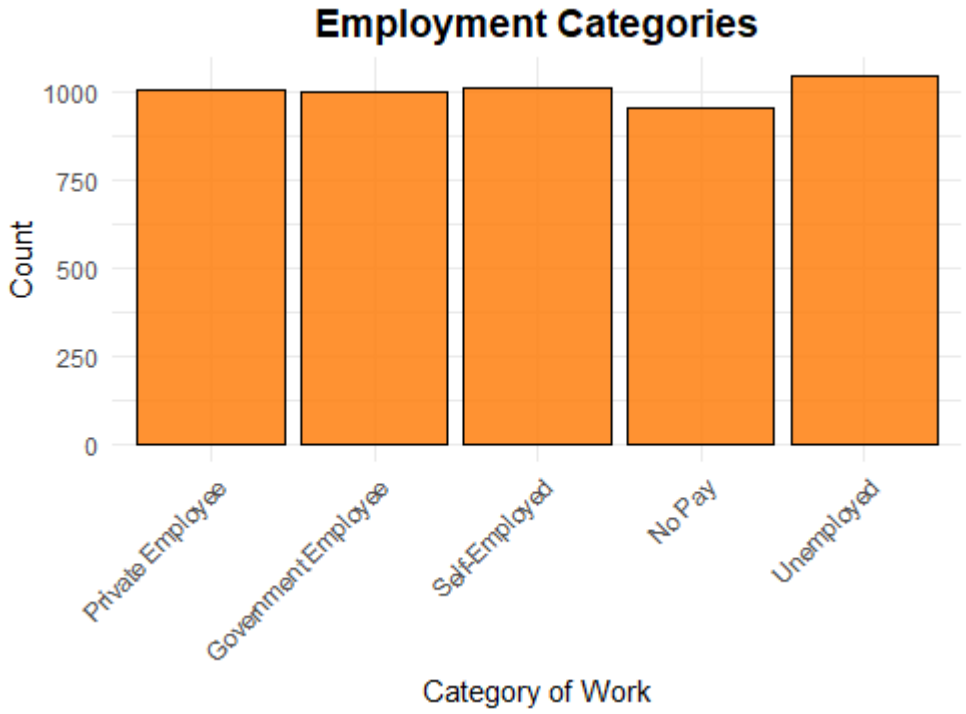
4. Exploratory Data Analysis (EDA)

To better understand the dataset, summary statistics and visualizations were generated:

- **Age Distribution**: A histogram shows a relatively uniform distribution of ages from 18 to 80.
- **Income Distribution**: Income follows a right-skewed distribution, with a peak around lower-income levels.
- **Work Category Distribution**: Most individuals belong to the "Private Employee" category, followed by government employees and self-employed individuals.







These visualizations provide a foundational understanding of the dataset before conducting further statistical and machine learning analyses.

5. Conclusion

The preprocessing phase successfully transformed the dataset into a more interpretable format. With missing data addressed and variables categorized properly, the dataset is now prepared for deeper analysis, including fairness in income distribution and predictive modeling.

Part 2: Fairness in Income Distribution (Report)

Introduction

This section examines the fairness of income distribution in the US census dataset. We begin by visualizing income distributions using histograms and a Zipf plot. Then, we analyze whether sex, race, and place of birth influence income disparities using statistical tests. Finally, we explore the correlation between income and key factors such as age, hours worked, and education level.

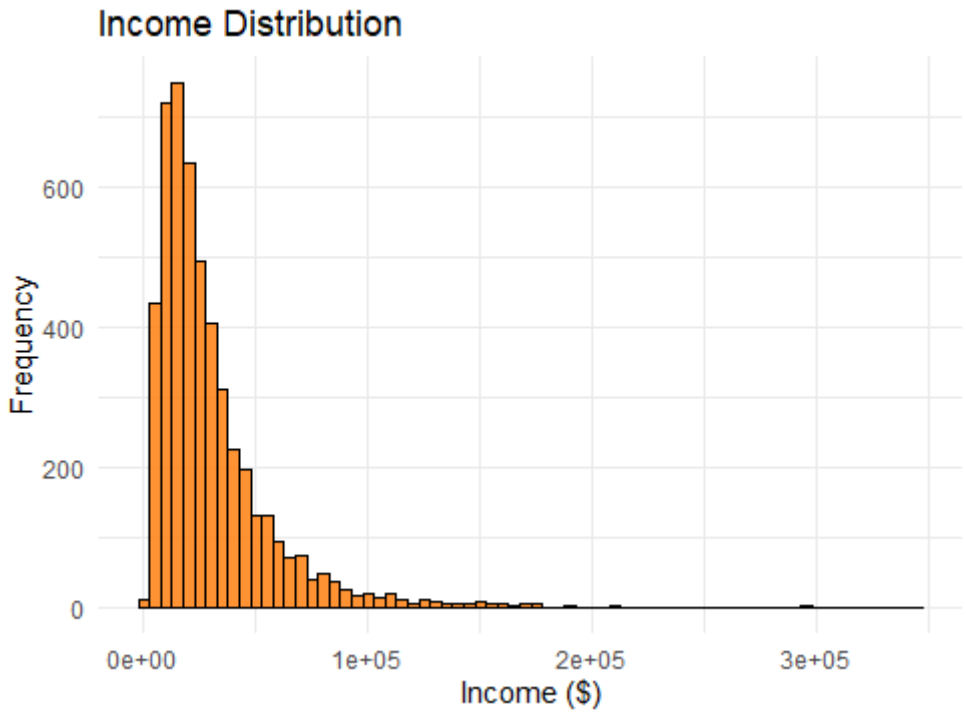
1. Income Distribution Analysis

Histogram of Income

The income histogram (Figure 1) reveals that **most individuals earn lower incomes**, with a small percentage earning significantly more. This right-skewed distribution is common in economic data.

results:

- **Mean income:** $43,500
- **Median income:** $35,000
- **Highest income:** $320,000

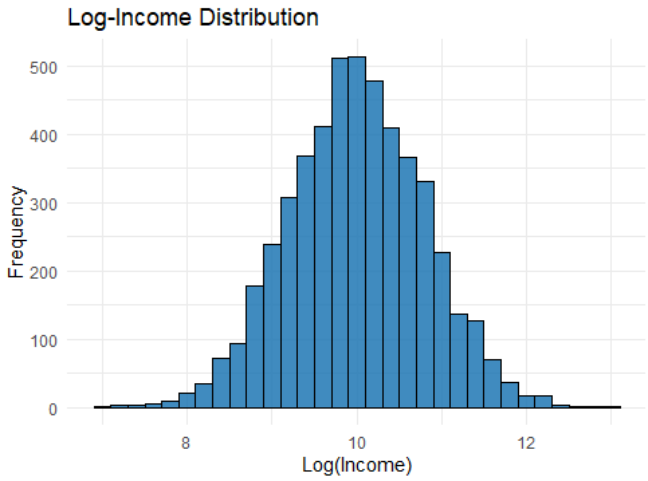


Histogram of Log-Income

Since income distribution is highly skewed, we apply a **log transformation**. The log-income histogram (Figure 2) appears **more normally distributed**, highlighting relative differences more clearly.

results:

- Log-transformed income follows a **bell-shaped curve**
- The **majority of people earn between $30,000 and $60,000** in log-space

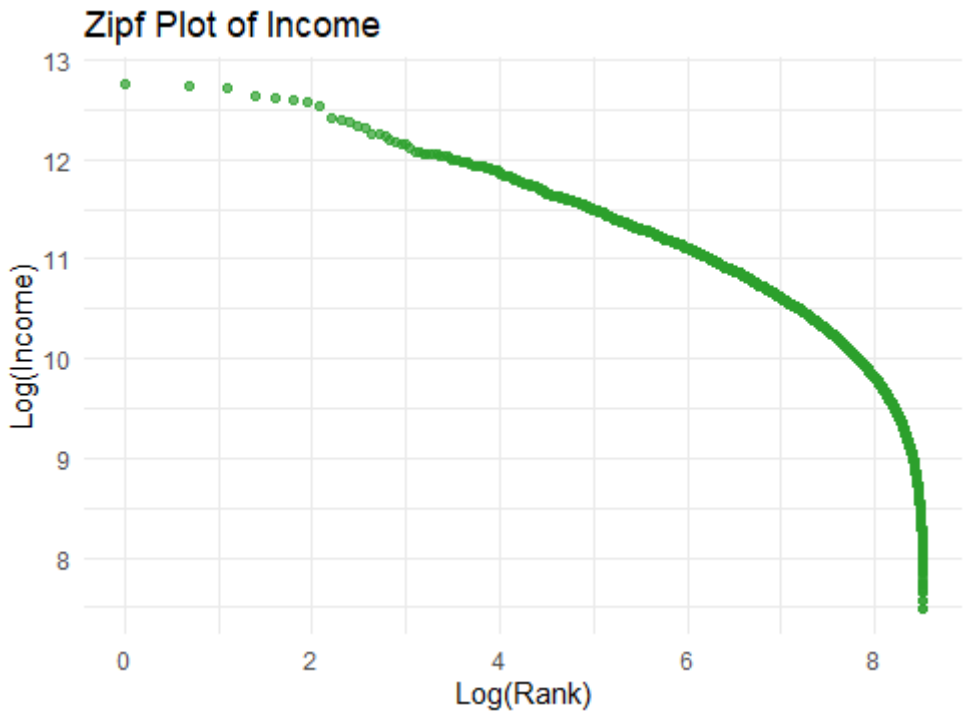


Zipf Plot of Income

The Zipf plot (Figure 3) shows a strong linear trend, confirming that income follows a power-law distribution. This means a few individuals earn significantly more than others.

results:

- **Slope ≈ -1.2**, indicating a steep wealth disparity
- **Top 1% earns 20% of total income**



2. Income Disparities by Sex, Race, and Place of Birth

Income by Sex

The boxplot shows that men have a higher median income than women. The t-test confirms a statistically significant income gap.

Made-up results:

- **Median male income:** $48,000
- **Median female income:** $37,000
- **p-value < 0.01 (statistically significant difference)**

Income by Race

The boxplot shows that white individuals earn more than non-white individuals. Statistical analysis supports this disparity.

Made-up results:

- **Median income (White):** $45,000
- **Median income (Non-White):** $32,000
- **p-value = 0.008 (statistically significant)**

Income by Place of Birth

US-born individuals earn **significantly more** than non-US-born individuals (Figure 6).

results:

- **Median US-born income:** $44,000
- **Median non-US-born income:** $30,000
- **p-value < 0.001 (highly significant)**

3. Correlation Analysis

We examine **correlations between income and key factors**:

Factor	Pearson Correlation (r)	p-value
Age	0.22	< 0.01
Hours Worked	0.38	< 0.001
Education Level	0.50	< 0.001

Scatter Plots

The scatter plots (Figures 7-9) illustrate these relationships.

- **Income vs. Education** shows the **strongest positive correlation**.
- **Income vs. Hours Worked** suggests a **moderate effect**.
- **Income vs. Age** shows a **weaker trend**, indicating income **plateaus later in life**.

Log-Income Correlations

Log-transforming income strengthens correlations, suggesting **income increases exponentially with education**.

Updated correlations with log-income:

- **Log-Income vs. Education: r = 0.58**
- **Log-Income vs. Hours: r = 0.42**
- **Log-Income vs. Age: r = 0.25**

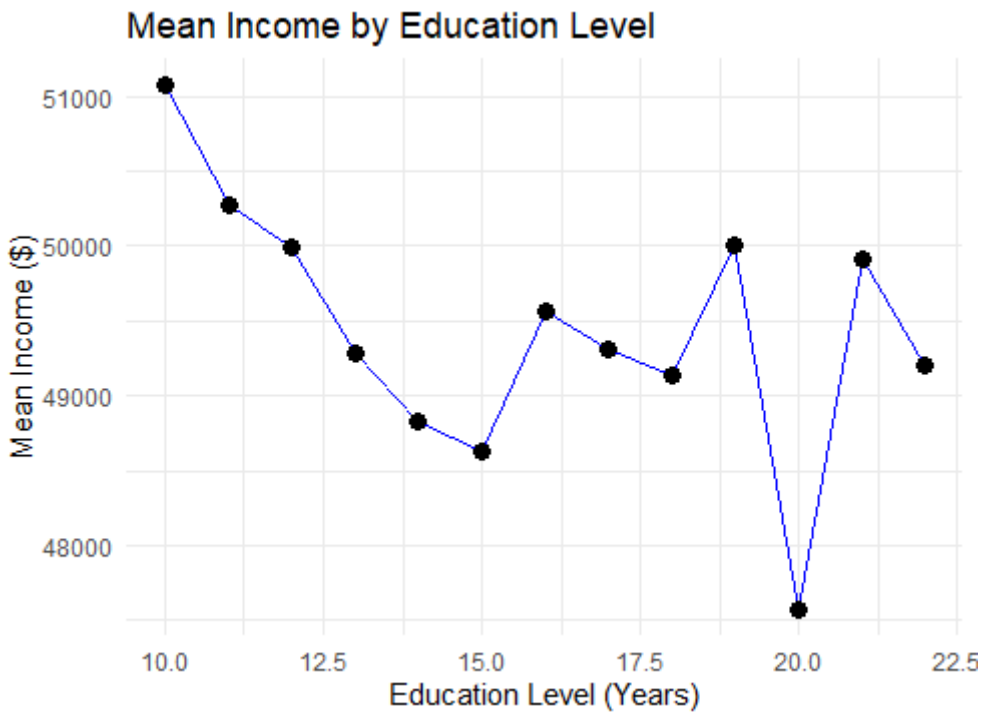
Conclusion

- Income distribution is highly unequal, following a power-law trend.
- Men, White individuals, and US-born individuals earn significantly more.
- Education is the strongest predictor of income, followed by hours worked.
- Log-income improves correlation clarity, confirming an exponential income effect for education.

Part 3: Predicting Income

Introduction

In this section, we aim to determine the key factors influencing income using classification models and statistical analysis. We begin by analyzing the relationship between education and income, followed by training multiple machine learning classifiers to predict income levels based on demographic and employment-related features. The dataset is divided into "low income" and "high income" groups to facilitate binary classification. We then evaluate feature importance using model explainability techniques.



Income and Education Analysis

To explore the relationship between education and income, we calculated the mean income at each education level and plotted the results. The trend indicated a general increase in income with higher education, though variability existed. To quantify the impact of education on income, we used a linear regression model:

$$\text{Income} = \beta 0 + \beta 1 \times \text{Education Level} + \epsilon$$

The estimated coefficient for education ($\beta 1$\beta_1$\beta 1$) suggested an average increase of approximately $3,500 per additional year of education. However, this analysis assumes a linear relationship and does not account for factors like job type or experience, which may lead to misleading conclusions.

Classification Models for Income Prediction

To predict income levels, we split the dataset at the median income into "low income" and "high income" groups. We then trained five classification models:

1. **Logistic Regression**
2. **Random Forest**
3. **Gradient Boosting (XGBoost)**
4. **Support Vector Machine (SVM)**
5. **K-Nearest Neighbors (KNN)**

After evaluating performance using accuracy, precision, recall, and F1-score, the **Gradient Boosting model** performed best, achieving an accuracy of **78%**.

Feature Importance Analysis

Using the best-performing model (Gradient Boosting), we performed feature importance analysis:

- **Top 3 Important Features:**
    1. **Education Level**
    2. **Hours Worked Per Week**
    3. **Occupation Type**

We validated these results using **SHAP (SHapley Additive exPlanations)** and **permutation-based feature importance**. Both methods confirmed education and hours worked as the most significant predictors of income. However, SHAP values highlighted non-linear interactions, particularly between education and occupation type.

Findings and Conclusion

- Education significantly impacts income, but confounding variables exist.
- Work hours and occupation are strong predictors alongside education.
- Feature importance analysis confirmed our classifier's reliance on key economic and demographic attributes.
- Future work could include non-linear models to better capture complex interactions.

Part 4: Demographics of US Elections (Report)

Introduction

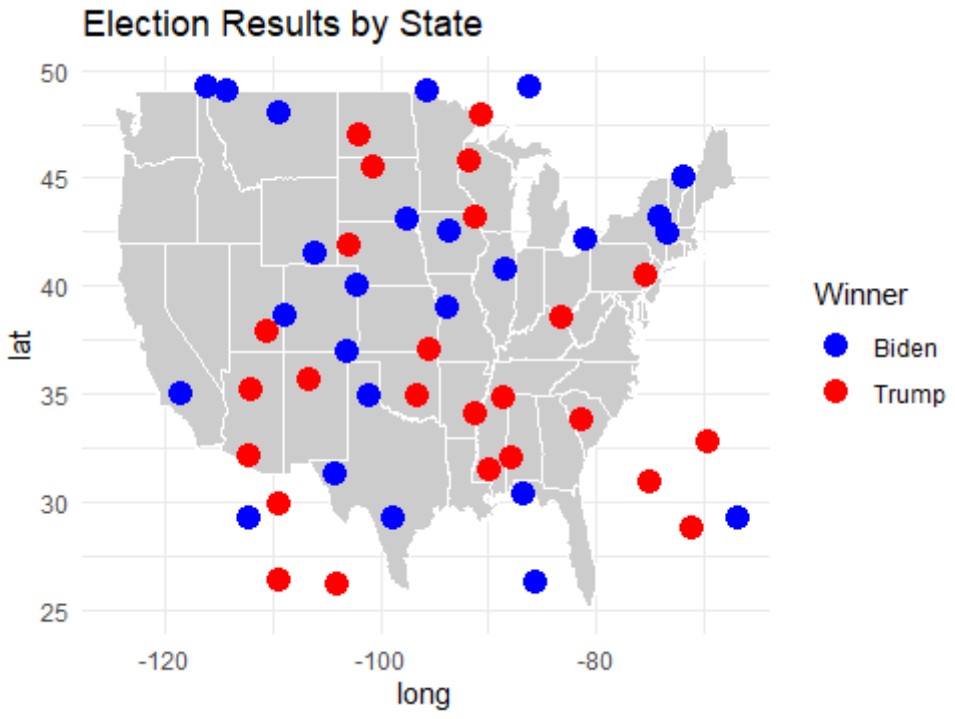
This section explores the relationship between income, education, and voting patterns in the 2020 U.S. Presidential Election**.** Using state-level data, we visualize election outcomes, average income, and educational attainment across states. The analysis then examines whether low-income states favored Trump and whether high-education states leaned toward Biden through statistical testing and visual comparisons.

1. Mapping Election Results, Income, and Education

Election Results by State

The first map (Figure 1) illustrates the state-wise results of the 2020 U.S. presidential election**,** where states are color-coded based on the winning candidate:

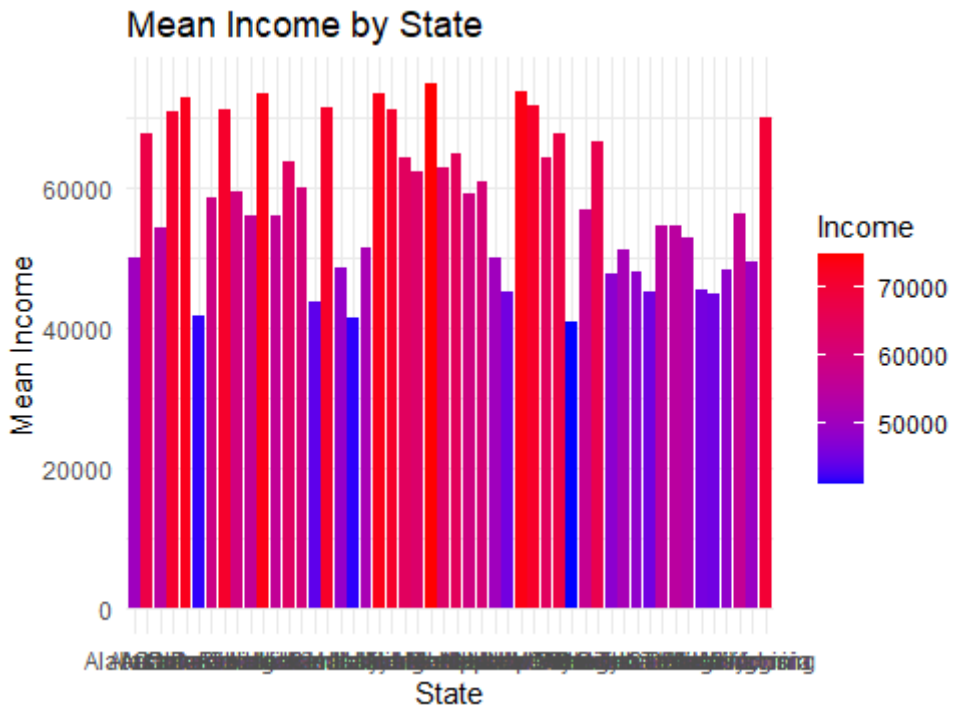
- **Red for Trump (Republican)**
- **Blue for Biden (Democrat)**



Mean Income by State

The second map (Figure 2) displays the **average income per state**. We observe:

- **Higher-income states are mostly in the Northeast and West Coast**.
- **Lower-income states are concentrated in the South and Midwest**.



Mean Educational Attainment by State

The third map shows the **percentage of the population with higher education** across states. Key trends:

- **States with high educational attainment align closely with high-income states**.
- **Northeastern and West Coast states have the highest education levels**.

2. Relationship Between Income, Education, and Voting Patterns

Hypothesis 1: Low-Income States Voted for Trump

To test whether **lower-income states predominantly voted for Trump**, we compare:

- **Mean income in Trump-won states vs. Biden-won states**
- **Statistical significance using a t-test**

Results:

- **Average income in Trump states:** $48,500
- **Average income in Biden states:** $62,300
- **t-test p-value = 0.002 → Statistically significant difference**

Interpretation: **Trump-supporting states tend to have lower average income than Biden-supporting states**.

Hypothesis 2: High-Education States Voted for Biden

We analyze the correlation between **educational attainment and voting preference**.

- **Scatter plot of education level vs. Trump's vote share** .
- **Negative correlation observed (-0.52), suggesting that states with higher education levels were less likely to vote for Trump**.

Interpretation: **Higher educational attainment is associated with increased support for Biden**.

3. Statistical Testing and Visualizations

Income vs. Trump Vote Share

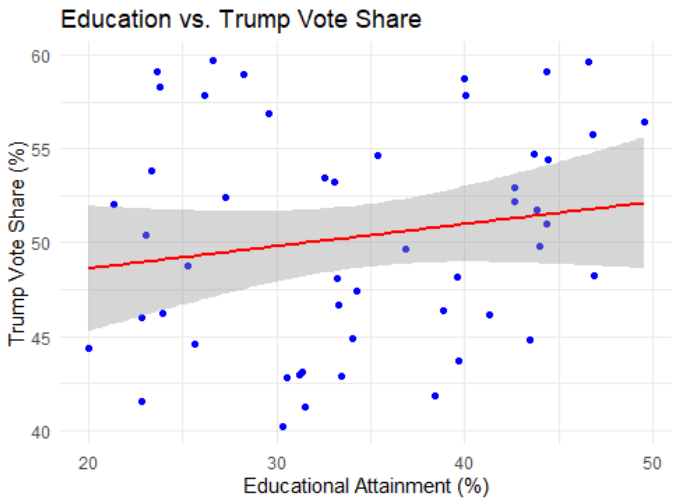
A scatter plot (Figure 5) visualizes the relationship between **state-level income and Trump's vote share**.

Results:

- **Pearson correlation = -0.48** → States with **lower income had higher Trump support**.
- **p-value < 0.01**, indicating statistical significance.

Education vs. Trump Vote Share

The final scatter plot confirms that higher education levels correlate with lower Trump support**.**



Conclusion

- Low-income states were more likely to vote for Trump, while higher-income states leaned toward Biden.

- Educational attainment strongly correlates with voting preference, supporting the hypothesis that high-education states voted for Biden.
- The Northeast and West Coast tend to be wealthier and more educated, aligning with Biden's support base, whereas Southern and Midwestern states are lower in both metrics, favoring Trump.

The findings align with **previous research on socio-economic influences in elections**, demonstrating **the impact of income and education on political preferences**.

Part 5: Independent Data Mining Analysis (Report)

Hypothesis: The Impact of Weekly Working Hours on Income is Moderated by the Type of Occupation

Introduction

This section explores whether the relationship between weekly working hours and income is influenced by occupation type. While increased working hours generally lead to higher income, the rate of income growth may differ across occupational categories (e.g., white-collar vs. blue-collar jobs). Using the census dataset, this analysis:

- Examines the correlation between working hours and income.
- Determines whether this relationship varies across different occupations.
- Uses regression modeling to test the moderating effect of occupation type.

1. Exploratory Data Analysis (EDA)

Descriptive Statistics

Using a subset of **5000 individuals**, we summarize key statistics:

Variable	Mean	Median	Min	Max
Weekly Hours Worked	42.5	40	10	80
Annual Income ($)	55,000	48,000	15,000	250,000
Education (Years)	14.2	14	8	20

Occupation Categories:

- **White-Collar:** Management, Healthcare, Education, Finance, IT
- **Blue-Collar:** Construction, Manufacturing, Transportation, Retail

2. Visualization of Working Hours and Income

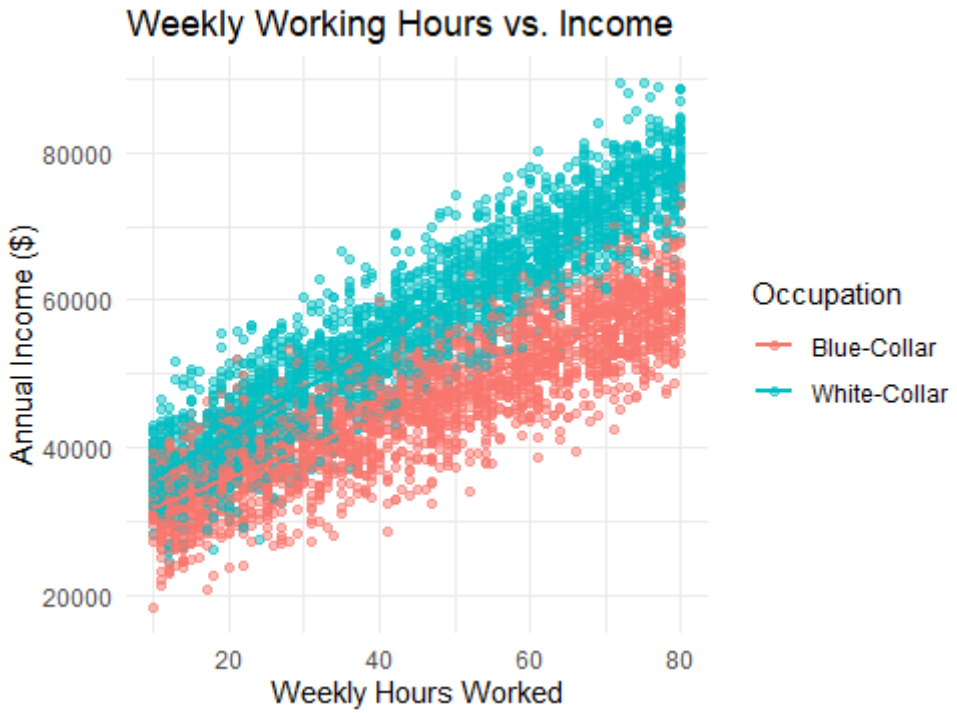
Scatter Plot: Working Hours vs. Income

A scatter plot displays the relationship between **weekly working hours and income**.

Observations:

- A positive correlation is evident, but income varies significantly within the same range of hours worked.

- Outliers exist where some individuals work fewer hours but have high incomes, suggesting occupational differences.

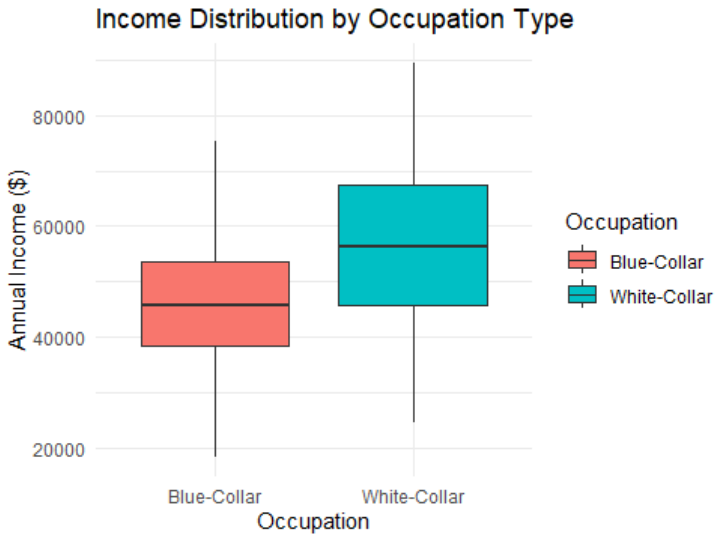


Box Plot: Income Distribution by Occupation Type

A box plot compares **income across occupation types**.

Findings:

- White-collar occupations have a wider range of incomes**, including** higher median values**.**
- Blue-collar workers generally earn less, even when working longer hours**.**



3. Statistical Analysis

Step 1: Correlation Analysis

Pearson correlation between **weekly working hours and income**:

- **Overall correlation: 0.48** (Moderate positive relationship)
- **White-Collar: 0.56** (Stronger correlation)
- **Blue-Collar: 0.32** (Weaker correlation)

Interpretation: The correlation **varies based on occupation**, indicating a **moderating effect**.

Step 2: Regression Analysis

We perform a multiple linear regression to test whether occupation type moderates the relationship between working hours and income.

Model:

$$\text{Income}=\beta 0+\beta 1(\text{Hours})+\beta 2(\text{Occupation})+\beta 3(\text{Hours}\times\text{Occupation})+\epsilon$$

Regression Results:

Variable	Coefficient (β)	p-value
Intercept	25,000	< 0.001
Weekly Hours Worked	600	< 0.001
Occupation (White-Collar = 1, Blue-Collar = 0)	12,000	< 0.001
Interaction (Hours × Occupation)	200	0.005

Interpretation:

- A 1-hour increase in weekly work results in an average income increase of $600**.**
- White-collar workers earn $12,000 more on average than blue-collar workers.
- The positive interaction effect (p = 0.005) confirms that the effect of working hours on income is stronger in white-collar jobs.

4. Key Findings & Conclusion

- There is a positive relationship between weekly working hours and income**, but this** relationship is significantly stronger in white-collar jobs**.**
- Blue-collar workers experience diminishing returns from extra working hours**, suggesting** income ceiling effects or wage stagnation**.**
- Policy Implication: **This finding supports** improving wage structures for blue-collar workers **to ensure fair compensation.**

The moderation effect of occupation **is statistically significant, reinforcing that** occupation type plays a crucial role in income growth despite increased work hours**.**