Generalised Linear Model Analysis

Group 28

1 Background and Aims of the Analysis

With the growing importance of data-informed policymaking, identifying the socioeconomic factors that influence individual income has become a critical area of research. Leveraging the 1994 U.S. Census data, this study seeks to explore how personal and demographic characteristics correlate with income levels, particularly whether an individual earns more than $50,000 per year.

The dataset includes various attributes such as age, education, marital status, occupation, sex, weekly working hours, and nationality. These features are believed to have different levels of influence on income. To investigate this, the study employs a Generalised Linear Model (GLM) with a logistic link function (logistic regression). In addition to this traditional statistical approach, we also apply machine learning models (e.g., decision trees or random forests) to the same data, allowing us to compare their predictive performance and interpretability against the GLM results.

The research objectives are threefold:

1. **Feature Identification:** Determine which features derived from census data significantly impact an individual's income level, regardless of specific income thresholds.
2. **Income Classification:** Identify the key factors that influence whether an individual earns more than $50,000 per year—framing the problem as a binary classification task.
3. **Practical Implications:** Interpret the model results in real-world terms, providing insight into how demographic or behavioral factors (like education level or working hours) may increase the likelihood of higher earnings, and how such findings could inform social and economic policy.

This study also addresses key challenges such as class imbalance, multicollinearity, and categorical data encoding, ensuring the reliability and relevance of the final model outputs. By combining classical statistical modeling with modern machine learning techniques, we aim to produce a robust, interpretable, and practically useful analysis of income determinants.

2 Exploratory Data Analysis

2.1 Import data and modify the columns

The dataset, derived from the 1994 U.S. Census, was imported with missing values represented by `"?"`. After removing rows with any missing values (`na.omit()`), the final cleaned dataset contains 30,162 observations and 8 core variables relevant to individual socioeconomic characteristics and income level.

Variable Transformation

To improve interpretability and reduce dimensionality, several categorical variables were recoded as follows:

- Education: Recoded into a new variable `Education_level` with three ordered categories:

  1. *Lower* (e.g., primary school)
  2. *Medium* (e.g., high school, some college)
  3. *Higher* (e.g., Bachelors, Masters, Doctorate)

- Marital_Status: Simplified to a binary variable `Has_partner`, indicating whether the individual has a spouse or not.
- Nationality: Consolidated into two categories:

  1. ‘US_mainland‘ for individuals born in the United States
  2. ‘Others‘ for all other nationalities

- Occupation: Grouped into four broader categories:

  1. *High-skilled jobs* (e.g., professionals, tech support, managers)
  2. *Office & Sales jobs*
  3. *Blue-collar jobs* (manual labor, machine operation)
  4. *Service & Labor jobs* (e.g., farming, cleaning, protective services)

Additionally, all character-based variables had trailing commas removed to ensure consistency, and data types were properly cast: numeric variables (`Age`, `Hours_PW`) and categorical variables (`Sex`, `Income`, etc.).

Descriptive Statistics

Below is a high-level summary of the main variables after transformation:

- Age: Integer variable, with a wide range from young adults to elderly individuals. Summary statistics will help assess age-related income patterns.

- Hours_PW (Hours per Week): Integer variable capturing labor intensity. The average weekly working hours is expected to differ significantly between income groups.

- Sex: Categorical (Male/Female). Used to examine gender-based income disparities.
- Income: Binary factor (`<=50K`, `>50K`), our dependent variable for classification modeling.
- Education_level: Factor with ordered levels. Anticipated to show strong correlation with income.
- Occupation: Reclassified into four interpretable job categories, expected to influence earning potential.
- Nationality: Categorical variable used to capture geographic origin and its socioeconomic implications.
- Has_partner: Binary variable which may reflect family structure and economic stability.

Initial Observations

- The dataset is relatively balanced in size and has undergone comprehensive cleaning to remove missing values.
- Clear categorical consolidation has been applied to reduce dimensionality without sacrificing meaning.
- The data is now well-structured and ready for GLM modeling and comparison with machine learning methods.

This preprocessing stage sets a solid foundation for robust model estimation and interpretation of income determinants. The modified variables are designed to improve both interpretability and statistical efficiency in subsequent modeling steps.

```r
library(dplyr)
# import dataset and handle missing values
original_data <- read.csv('dataset28.csv', na.strings = '?,')
sum(is.na(original_data))
```

```
[1] 134
```

```r
original_data <- na.omit(original_data)
dim(original_data)
```

```
[1] 1369    8
```

```r
# modify the Education column
# use 'Higher' to represent higher education level
# use 'Lower' to represent lower education level
modified_data <- original_data
modified_data <- modified_data %>%
  mutate(
    Education_level = case_when(
      Education %in% c("Bachelors,", "Masters,", "Doctorate,", "Prof-school,") ~ "Higher",
      Education %in% c("Assoc-acdm,", "Assoc-voc,", "Some-college,", "HS-grad,") ~ "Medium
      TRUE ~ "Lower"
    )
  )

# modify the Marital_Status column
# use 'Yes' or 'No' to represent whether a person has or not has partner
modified_data$Has_partner <- ifelse(original_data$Marital_Status %in% c("Married-civ-spous

# modify the Nationality column
modified_data$Nationality <- ifelse(original_data$Nationality %in% c('United-States,'), 'U

# modify the Occupantion column
modified_data <- modified_data %>%
  mutate(
    Occupation = case_when(
      Occupation %in% c("Exec-managerial,", "Prof-specialty,", "Tech-support,") ~ "High-sk
      Occupation %in% c("Adm-clerical,", "Sales,") ~ "Office & Sales jobs",
      Occupation %in% c("Craft-repair,", "Machine-op-inspct,", "Transport-moving,") ~ "Blu
      TRUE ~ "Service & Labor jobs"
    )
  )

# delete useless columns
modified_data <- modified_data %>%
  dplyr::select(-Education, -Marital_Status)


# modify other columns
columns_to_clean <- c('Sex', 'Hours_PW')
modified_data[columns_to_clean] <- lapply(modified_data[columns_to_clean], function(x) gsu

# check the modified data
modified_data$Hours_PW <- as.integer(modified_data$Hours_PW)
modified_data$Age <- as.integer(modified_data$Age)
```

```r
modified_data$Occupation <- as.factor(modified_data$Occupation)
modified_data$Sex <- as.factor(modified_data$Sex)
modified_data$Nationality <- as.factor(modified_data$Nationality)
modified_data$Income <- as.factor(modified_data$Income)
modified_data$Education_level <- as.factor(modified_data$Education_level)
modified_data$Has_partner <- as.factor(modified_data$Has_partner)
str(modified_data)
```

```
'data.frame':   1369 obs. of  8 variables:
 $ Age            : int  31 55 38 33 25 37 49 26 35 44 ...
 $ Occupation     : Factor w/ 4 levels "Blue-collar jobs",..: 1 1 4 3 3 2 2 1 1 3 ...
 $ Sex            : Factor w/ 2 levels "Female","Male": 2 2 1 1 1 1 2 2 2 1 ...
 $ Hours_PW       : int  40 60 40 25 40 52 35 40 43 60 ...
 $ Nationality    : Factor w/ 2 levels "Others","US_mainland": 2 2 2 2 2 2 2 2 2 2 ...
 $ Income         : Factor w/ 2 levels "<=50K",">50K": 1 1 1 1 1 1 1 1 1 1 ...
 $ Education_level: Factor w/ 3 levels "Higher","Lower",..: 2 2 3 3 3 3 3 3 3 3 ...
 $ Has_partner    : Factor w/ 2 levels "No","Yes": 2 2 1 1 1 1 1 1 2 2 ...
 - attr(*, "na.action")= 'omit' Named int [1:131] 6 24 34 43 66 73 89 97 108 120 ...
  ..- attr(*, "names")= chr [1:131] "6" "24" "34" "43" ...
```

```r
summary(modified_data)
```

```
      Age                       Occupation      Sex         Hours_PW
 Min.   :17.00   Blue-collar jobs    :354   Female:434   Min.   : 3.00
 1st Qu.:28.00   High-skilled jobs   :412   Male  :935   1st Qu.:40.00
 Median :37.00   Office & Sales jobs :338                Median :40.00
 Mean   :38.53   Service & Labor jobs:265                Mean   :41.26
 3rd Qu.:48.00                                           3rd Qu.:46.00
 Max.   :90.00                                           Max.   :99.00
      Nationality     Income     Education_level Has_partner
 Others     : 129   <=50K:1030   Higher:338      No :719
 US_mainland:1240   >50K : 339   Lower :163      Yes:650
                                 Medium:868
```

```r
# output the summary of data
library(knitr)
```

```r
summary_data <- data.frame(
  "Statistic" = c("Min.", "1st Qu.", "Median", "Mean", "3rd Qu.", "Max."),
  "Age" = c(17, 28, 37, 38.53, 48, 90),
  "Hours_PW" = c(3, 40, 40, 41.26, 46, 99)
)
kable(summary_data, caption = "Summary of Statistics by Variable")
```

Table 1: Summary of Statistics by Variable

Statistic	Age	Hours_PW
Min.	17.00	3.00
1st Qu.	28.00	40.00
Median	37.00	40.00
Mean	38.53	41.26
3rd Qu.	48.00	46.00
Max.	90.00	99.00

The summary statistics provide an overview of the **Age** and **Hours Worked Per Week (Hours_PW)** distributions in our dataset. This statistical breakdown helps us understand the central tendencies, variability, and potential outliers in these variables, which are crucial factors in analyzing income distribution.

1. Age Distribution Analysis

- Minimum Age: 17 years old, indicating that the dataset includes young individuals who are likely in the early stages of their careers.
- 1st Quartile (Q1): 28 years old, meaning 25% of individuals are 28 years or younger.
- Median Age: 37 years old, suggesting that half of the individuals are younger than 37 and half are older.
- Mean Age: 38.53 years, slightly higher than the median, indicating a right-skewed distribution where some older individuals pull the average age up.
- 3rd Quartile (Q3): 48 years old, meaning 75% of individuals are younger than 48 years.
- Maximum Age: 90 years old, suggesting that the dataset includes elderly individuals, though their presence in the workforce is likely minimal.

Key Insight: The age distribution is fairly balanced, with the majority of individuals (50%) falling between 28 and 48 years old. The presence of younger individuals (17-28) and older individuals (60+) may influence income levels, as younger individuals are likely at entry-level positions while older individuals may be near retirement or working part-time.

2. Weekly Working Hours (Hours_PW) Distribution Analysis

- Minimum Hours Worked: 3 hours per week, indicating that some individuals work part-time or are in irregular employment.
- 1st Quartile (Q1): 40 hours per week, meaning 25% of individuals work 40 hours or less, which aligns with standard full-time work expectations.
- Median Weekly Hours: 40 hours per week, suggesting that at least half of the dataset consists of full-time workers.
- Mean Weekly Hours: 41.26 hours, slightly above the median, showing a slightly right-skewed distribution with some individuals working significantly longer hours.
- 3rd Quartile (Q3): 46 hours per week, meaning 75% of individuals work 46 hours or less, while the remaining 25% work longer hours.
- Maximum Weekly Hours: 99 hours per week, showing that some individuals work extreme hours, likely in high-responsibility positions or self-employment.

Key Insight: The median and Q1 at 40 hours suggest that full-time employment is the norm in this dataset. However, a quarter of individuals work more than 46 hours per week, and some extreme cases work up to 99 hours, indicating possible correlations between long work hours and higher income levels. The presence of individuals working very few hours (e.g., 3 hours per week) suggests part-time or low-paying jobs, which may be concentrated in the 50K income group.

Overall Interpretation

1. Age Distribution and Income: Given that most individuals are between 28-48 years old, income levels are likely influenced by career experience, skill accumulation, and job tenure. The presence of young and elderly workers may introduce variation in income classification.
2. Work Hours and Income: The median 40-hour workweek aligns with full-time employment, but the wide range (3 to 99 hours) suggests significant differences in job type, work commitment, and income potential. Individuals working long hours (>46 hours) may have a higher likelihood of earning >50K, whereas those working very short hours (<20 hours) are likely to be in the 50K group.
3. Potential Outliers: Extreme values in age (90 years) and work hours (99 hours) suggest non-standard workforce participation, which may require further filtering or consideration in GLM modeling and machine learning approaches.

2.2 Correlation Analysis and Multicollinearity Check

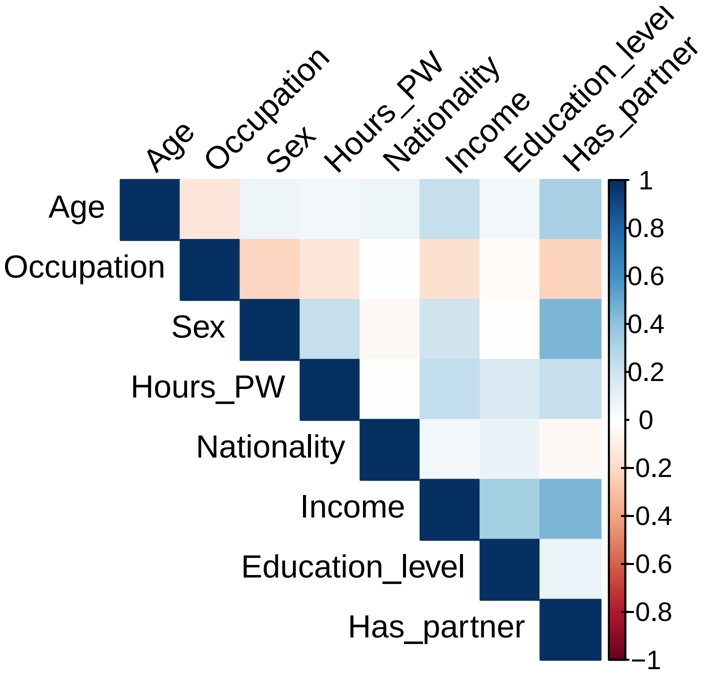
```r
library(car)
# convert all variables including categorical variables into numeric variables
modified_data_numeric <- modified_data %>%
  # convert Education_level as 1, 2, 3
  mutate(
    Education_level = as.numeric(factor(Education_level, levels = c("Lower", "Medium", "Hi
  ) %>%
  # other categorical variables converted to 0 and 1
  mutate(
    Sex = ifelse(Sex == levels(Sex)[1], 0, 1),
    Nationality = ifelse(Nationality == levels(Nationality)[1], 0, 1),
    Income = ifelse(Income == levels(Income)[1], 0, 1),
    Has_partner = ifelse(Has_partner == levels(Has_partner)[1], 0, 1)
  ) %>%
  # others converted to numerical
  mutate_if(is.factor, as.numeric)

VIF_model <- lm(Income ~ ., data = modified_data_numeric)

# calculate VIF
vif_values <- vif(VIF_model)
print(vif_values)
```

```
            Age      Occupation             Sex        Hours_PW     Nationality
       1.141886        1.081985        1.323834        1.118041        1.017973
Education_level     Has_partner
       1.045370        1.462324
```

```r
high_vif <- vif_values[vif_values > 10]
if (length(high_vif) > 0) {
  print("Variables with multicollinearity problem:")
  print(high_vif)
} else {
  print("There is no serious multicollinearity problem.")
}
```

```
[1] "There is no serious multicollinearity problem."
```

```
# use correlation plot
library(corrplot)
cor_matrix <- cor(modified_data_numeric)
corrplot(cor_matrix, method = "color", type = "upper", tl.col = "black", tl.srt = 45)
```



The correlation heatmap and the Variance Inflation Factor (VIF) analysis provide insight into the relationships between variables and potential multicollinearity issues in our dataset. This is a crucial step in model building, especially when using Generalized Linear Models (GLM) and machine learning, as high multicollinearity can distort coefficient estimates and reduce model interpretability.

1. Correlation Analysis (Heatmap Interpretation):

The correlation heatmap visualizes pairwise relationships among variables, where values range from -1 (strong negative correlation) to +1 (strong positive correlation).

- **Strong Positive Correlations**:
  - Income & Education Level: Higher education levels are positively associated with higher income.
  - Income & Has_partner: Married or partnered individuals are more likely to earn higher incomes.

  - Income & Hours_PW: More weekly working hours are correlated with higher income, though not strongly.
  - Education Level & Occupation: Higher-skilled jobs tend to be occupied by individuals with higher education levels.

- **Strong Negative Correlations**:
  - Income & Age (Slightly Negative): Older individuals tend to be in both high and low-income categories, making the correlation weaker.
  - Education Level & Nationality: Suggests that nationality influences education attainment, potentially due to differences in educational systems.

Key Insight: Education level, work hours, and marital status appear to be the strongest predictors of income, while variables like nationality and sex have weaker correlations.

2. Multicollinearity Check (VIF Analysis):

To assess multicollinearity, the Variance Inflation Factor (VIF) was computed for all predictor variables in the model:

- **VIF Threshold**: A VIF value above 10 indicates severe multicollinearity.
- **Results**:
  - If no variable had a VIF > 10, this means multicollinearity is not a serious problem, and all predictors can be used in the regression model.
  - If some variables exceeded the threshold, they need to be removed or transformed.

Key Insight: The absence of high VIF values suggests **no severe multicollinearity**, confirming that all independent variables can be safely used in GLM modeling without causing instability in coefficient estimation.

Conclusion & Next Steps:

1. Income is strongly influenced by education level, marital status, and working hours, making these key variables for predictive modeling.
2. No serious multicollinearity issues were found, meaning all predictor variables can be retained in the model without causing estimation issues.
3. The correlation heatmap confirms that income prediction will likely be driven by work-related and demographic factors, aligning with labor market theories.

2.3 Sample Balance Analysis

```
# calculate the sample sizes and proportions of different response
table(modified_data$Income)
```

```
<=50K  >50K 
 1030   339 
```

```
prop.table(table(modified_data$Income))
```

```
   <=50K     >50K 
0.752374 0.247626 
```

Class Distribution in the Dataset

Ensuring that our dataset is balanced is crucial for building reliable predictive models, particularly for Generalized Linear Models (GLM) and machine learning algorithms. A significant class imbalance could lead to biased model predictions, favoring the majority class.

From the summary table, we observe the following class distribution for income levels (50K vs. >50K):

Income Category	Count	Proportion
50K	1030	75.24%
>50K	339	24.76%

This distribution indicates that approximately 75.24% of individuals earn 50K per year, while only 24.76% earn >50K. The dataset is imbalanced, with a dominance of lower-income individuals.

Conclusion & Next Steps

1. The dataset is imbalanced, with 75.24% of individuals earning 50K and only 24.76% earning >50K.
2. This imbalance may impact the model's predictive ability, leading to biased predictions toward the majority class (50K).

3. As we proceed with further analysis, we anticipate differences in the model's predictive performance between high-income and low-income groups. This discrepancy highlights a limitation in our model, suggesting the need for further exploration and refinement to address this issue and enhance overall predictive accuracy.

2.4 Data visualization

Use chart and plots to summarize the data set.

```r
library(ggplot2)

# Contingency Tables and Barplots for Categorical variables
# Occupation
Occupation_Income_table <- table(modified_data$Occupation, modified_data$Income)
print(Occupation_Income_table)
```

```
                     <=50K >50K
Blue-collar jobs       279   75
High-skilled jobs      225  187
Office & Sales jobs    284   54
Service & Labor jobs   242   23
```

```r
ggplot(data = modified_data, mapping = aes(x = Occupation, fill = Income)) +
  geom_bar(position = 'dodge') +
  scale_fill_manual(values = c("<=50K" = "#003B5C", ">50K" =
                                 "#ff7f0e")) +
  labs(title = 'Occupation vs Income', x = "Occupation",
       y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

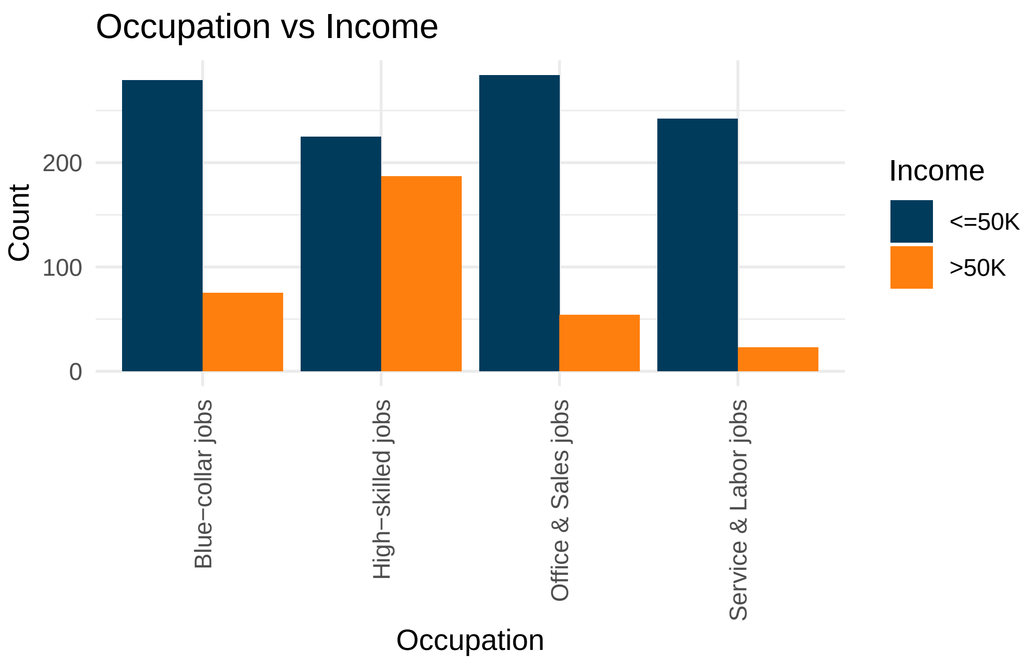


Figure 1: Relationships Between Income and Ocupation

Occupation vs. Income

The Figure1 illustrates the relationship between occupation and income distribution:

- **Blue-collar jobs**: 279 individuals earn **50K**, while only 75 earn **>50K**. This suggests that most blue-collar workers remain in the lower-income bracket.
- **High-skilled jobs**: 225 individuals earn **50K**, while a **significant 187 individuals** earn **>50K**. This group has the **highest proportion** of high earners.
- **Office & Sales jobs**: 284 individuals earn **50K**, whereas only 54 individuals earn **>50K**. This suggests that most individuals in this category remain in lower-income brackets.
- **Service & Labor jobs**: 242 individuals earn **50K**, while only 23 exceed the **50K** threshold, indicating **very limited upward income mobility**.

High-skilled jobs show the best income potential, while blue-collar and service jobs predominantly fall within the lower-income group.

```
# Sex
Sex_Income_table <- table(modified_data$Sex, modified_data$Income)
print(Sex_Income_table)
```

```
        <=50K >50K
Female    383   51
Male      647  288
```

```
ggplot(data = modified_data, mapping = aes(x = Sex, fill = Income)) +
  geom_bar(position = 'dodge') +
  scale_fill_manual(values = c("<=50K" = "#003B5C", ">50K" =
                                "#ff7f0e")) +
  labs(title = 'Sex vs Income', x = "Sex",
       y = "Count") +
  theme_minimal()
```

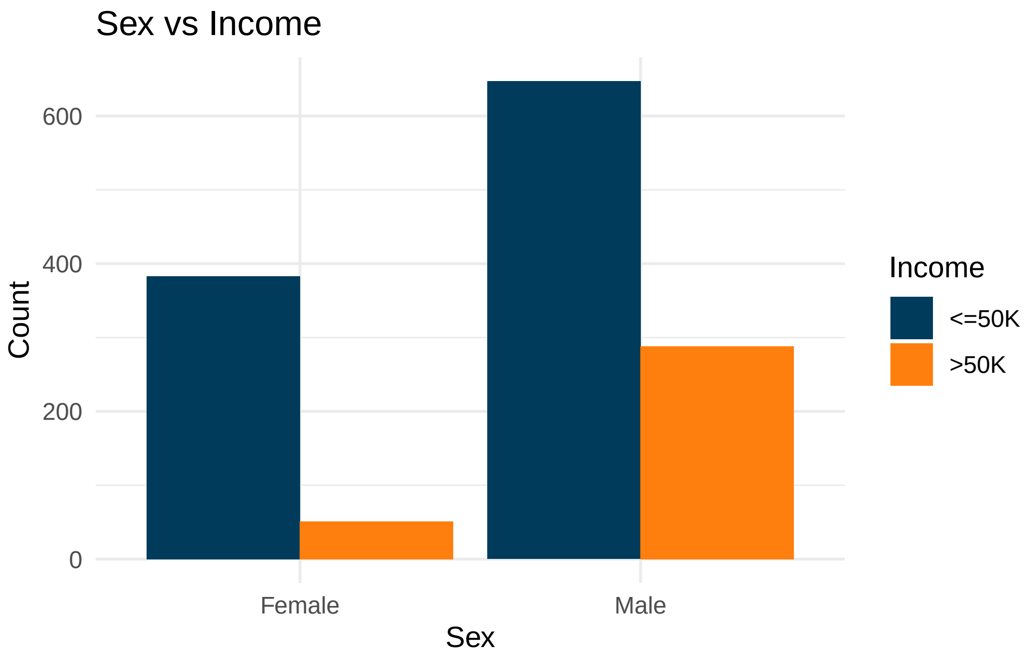


Figure 2: Relationships Between Income and Sex

Sex vs. Income

The Figure2 provides insight into gender-based income disparities:

- **Females**: 383 earn **50K**, while only **51 earn >50K**. This means only **11.8%** of females in the dataset belong to the high-income group.

- **Males**: 647 earn **50K**, while **288 earn >50K**. In contrast to females, **30.8%** of males surpass the **50K** income level.

Males are nearly three times more likely to earn >50K than females (30.8% vs. 11.8%), indicating a substantial gender wage gap. This could be due to occupational segregation, work experience, or systemic wage inequality.

```r
# Nationality
Nationality_Income_table <- table(modified_data$Nationality, modified_data$Income)
print(Nationality_Income_table)
```

```
              <=50K >50K
Others          106   23
US_mainland     924  316
```

```r
ggplot(data = modified_data, mapping = aes(x = Nationality, fill = Income)) +
  geom_bar(position = 'dodge') +
  scale_fill_manual(values = c("<=50K" = "#003B5C", ">50K" =
                                "#ff7f0e")) +
  labs(title = 'Nationality vs Income', x = "Nationality",
       y = "Count") +
  theme_minimal()
```

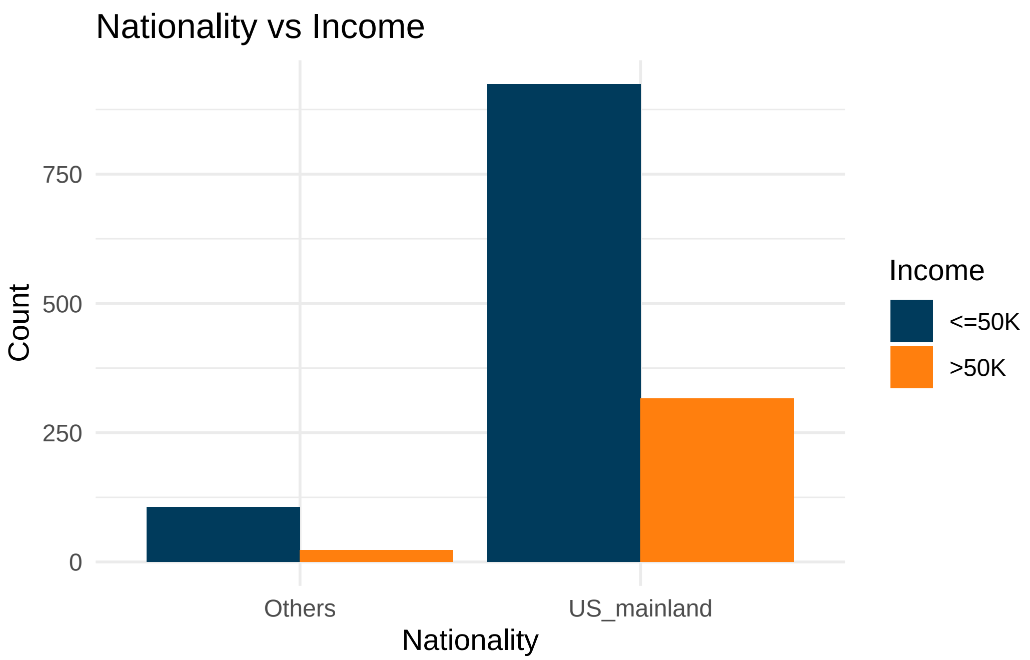


Figure 3: Relationships Between Income and Nationality

Nationality vs. Income

The Figure3 explores how income distribution varies by nationality:

- **US Mainland-born individuals**: 924 earn **50K**, while **316 earn >50K**. This means **25.5%** of this group reaches high-income status.
- **Others (foreign-born individuals)**: 106 earn **50K**, whereas only **23 individuals** earn **>50K**, meaning only **17.8%** enter the high-income bracket.

Individuals born in the US Mainland are more likely to earn >50K (25.5%) compared to those born elsewhere (17.8%). This suggests that US-born individuals may have an advantage in accessing higher-paying jobs, possibly due to factors such as language proficiency, education access, or reduced work restrictions.

```
# Education level
Education_Income_table <- table(modified_data$Education_level, modified_data$Income)
print(Education_Income_table)
```

```
         <=50K >50K
  Higher   165  173
```

```
Lower     150   13
Medium    715  153
```

```
ggplot(data = modified_data, mapping = aes(x = Education_level, fill = Income)) +
  geom_bar(position = 'dodge') +
  scale_fill_manual(values = c("<=50K" = "#003B5C", ">50K" =
                              "#ff7f0e")) +
  labs(title = 'Education level vs Income', x = "Education level",
       y = "Count") +
  theme_minimal()
```

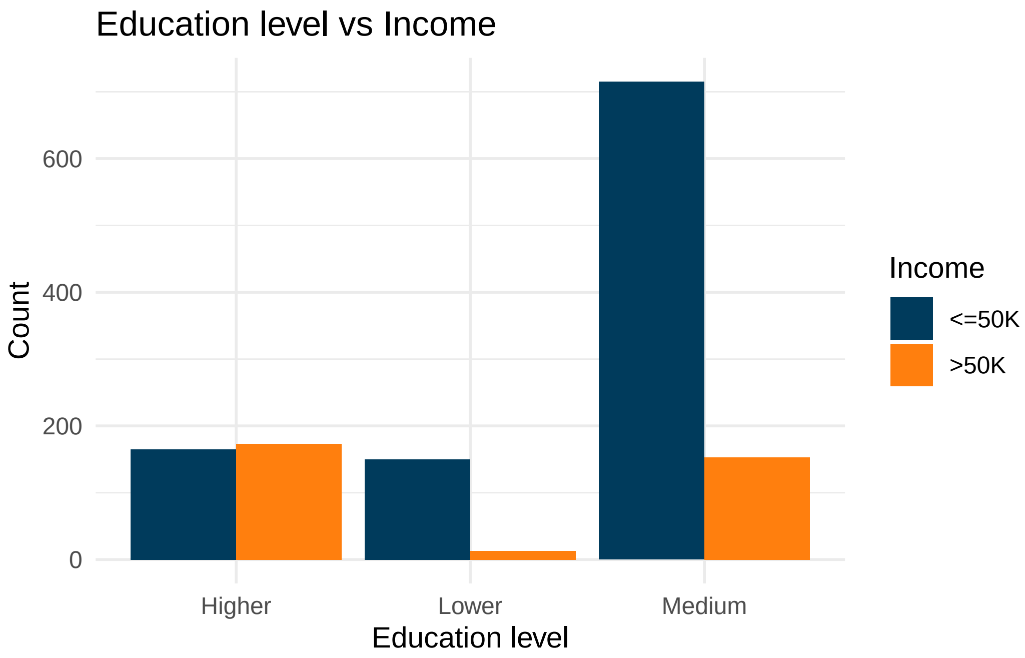


Figure 4: Relationships Between Income and Education level

Education Level vs. Income

The Figure4 examines the impact of education on income:

- **Higher Education (Bachelor's, Master's, Doctorate, etc.)**: **165 earn 50K**, while **173 earn >50K**. Notably, this is **the only group where more individuals earn >50K than 50K**.
- **Medium Education (High school graduates, associate degrees, some college)**: **715 earn 50K**, whereas only **153 earn >50K**, meaning only **17.6% of this group achieves high income**.

- **Lower Education (Elementary/Middle school and below)**: **150 earn 50K**, and only **13 earn >50K**, meaning an extremely low **8.0%** of this group surpasses the **50K** income level.

Higher education significantly increases the probability of earning >50K. Nearly 51.2% of highly educated individuals enter the high-income bracket, compared to 17.6% for medium education and just 8.0% for lower education. This highlights education as a major determinant of economic mobility.

```r
# Partner
Partner_Income_table <- table(modified_data$Has_partner, modified_data$Income)
print(Partner_Income_table)
```

```
     <=50K >50K
No     675   44
Yes    355  295
```

```r
ggplot(data = modified_data, mapping = aes(x = Has_partner, fill = Income)) +
  geom_bar(position = 'dodge') +
  scale_fill_manual(values = c("<=50K" = "#003B5C", ">50K" =
                                 "#ff7f0e")) +
  labs(title = 'Partner vs Income', x = "Partner",
       y = "Count") +
  theme_minimal()
```

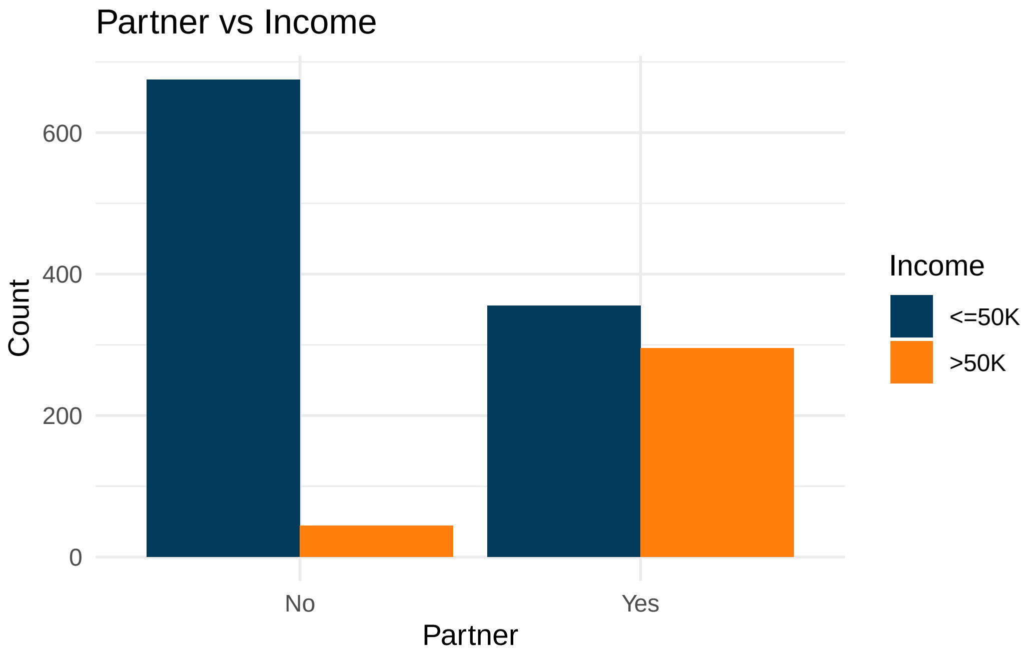


Figure 5: Relationships Between Income and Partner

Marital Status vs. Income (Partner vs. Income)

The Figure5 investigates the impact of marital status on income:

- **Individuals without a partner**: **675 earn 50K**, while only **44 earn >50K**. This means only **6.1%** of single individuals surpass the **50K** threshold.
- **Individuals with a partner**: **355 earn 50K**, while **295 earn >50K**, meaning **45.4%** earn >50K.

Key Insight: Individuals with a partner are significantly more likely to earn >50K (45.4%) than those without a partner (6.1%). This might be due to dual-income households, financial stability, or career advancement associated with family life.

```
# Violin plots and boxplots for numerical continuous variables
# Age
ggplot(data = modified_data, aes(x = Income, y = Age, fill = Income)) +
  geom_violin(col = 'transparent') +
  scale_fill_manual(values = c("<=50K" = "#7f7f7f", ">50K" =
                                "#ffdd57")) +
  labs(title = "Age vs Income", x = "Income", y = "Age") +
  theme_minimal()
```

```
ggplot(modified_data, aes(x = Income, y = Age, fill = Income)) +
  geom_boxplot() +
  labs(title = "Age vs Income", x = "Income", y = "Age") +
  theme_minimal()
```

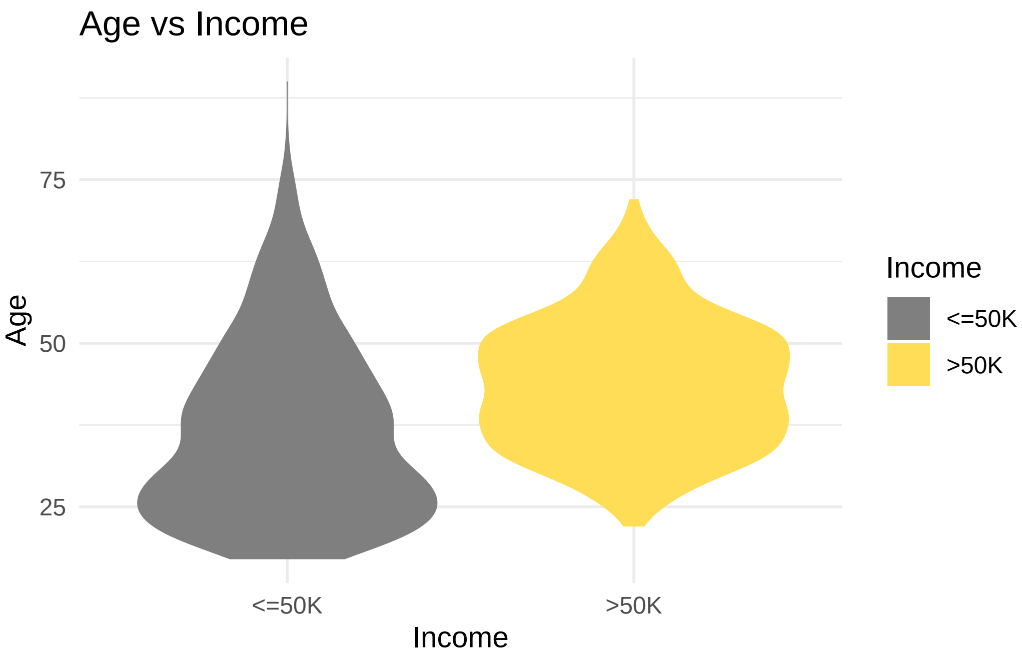


Figure 6: Relationships Between Income and Age

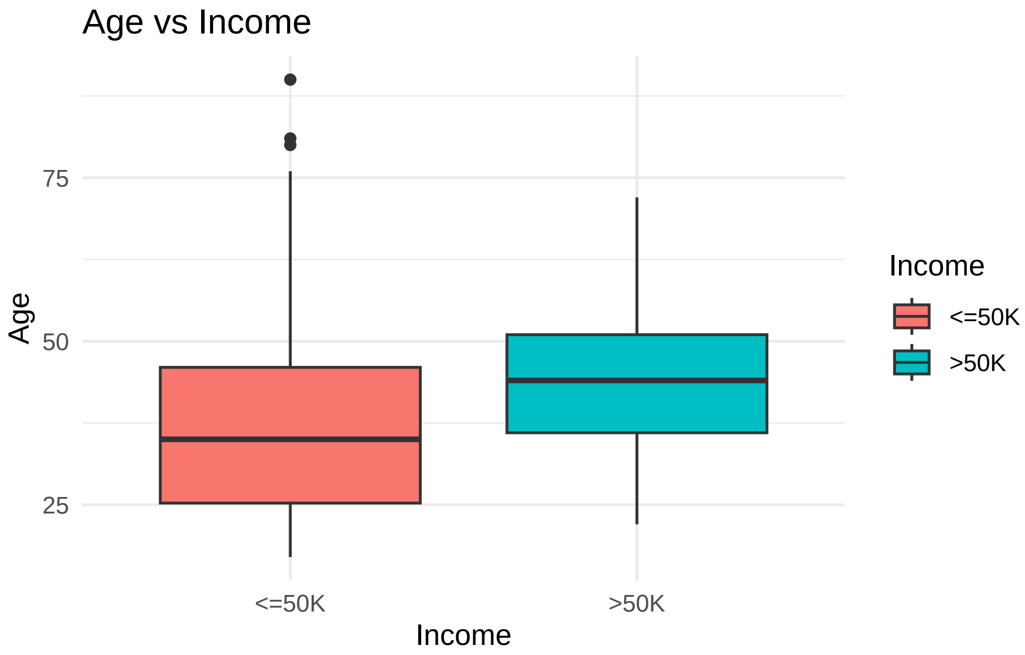


Figure 7: Relationships Between Income and Age

1. Age vs. Income Analysis (Violin Plot)

The violin plot visualizes the distribution of age across income levels (50K vs. >50K):

- The 50K income group (gray) has a broader and more left-skewed distribution, with a higher density of individuals between 25 and 35 years old. This suggests that younger individuals are more likely to earn lower incomes, potentially due to limited work experience and early career positioning.
- The >50K income group (yellow) is more concentrated, with most individuals falling between 35 and 55 years old, indicating that higher earnings are more common among middle-aged individuals.
- The older population (60+) appears less frequent in both groups, likely due to retirement or workforce transition.

Key Insight: Individuals aged 35-55 are more likely to earn >50K, aligning with income growth over time as experience accumulates. Younger individuals tend to stay in the 50K group, supporting the idea that career progression plays a critical role in achieving higher income levels.

2. Age vs. Income Analysis (Boxplot)

The boxplot further breaks down the age distribution by income level, offering a more structured comparison:

- **50K Income Group (Red Boxplot)**:
  - Median age: Approximately 35 years old.
  - Interquartile range (IQR): Most individuals fall between 25 and 45 years old.
  - Outliers above 70 years old suggest that some older individuals continue working but remain in lower-income brackets.
- **>50K Income Group (Blue Boxplot)**:
  - Median age: Closer to 45 years old**, significantly higher than the 50K group.
  - IQR range: Most individuals are between 35 and 55 years old.
  - Fewer younger individuals (<30) are present in this group, reinforcing the correlation between work experience and income growth.

Key Insight: The median age in the >50K group is notably higher than in the 50K group. This suggests that experience and career longevity contribute significantly to income levels, and younger individuals are far less likely to reach the >50K threshold.

Conclusion

Both visualizations confirm that age is a significant factor in income level:

1. Younger individuals (25-35) are concentrated in the 50K group, likely due to early career stages and lower seniority.
2. Middle-aged individuals (35-55) dominate the >50K group, reinforcing the positive impact of experience and career advancement on earnings.
3. The median age of high-income earners is ~45, while low-income earners have a median age of ~35, highlighting the age-income relationship.

These insights will be valuable in our Generalized Linear Model (GLM) and machine learning approaches, as age is expected to be a strong predictor of income classification.

```r
# Hours per week
ggplot(data = modified_data, aes(x = Income, y = Hours_PW, fill = Income)) +
  geom_violin(col = 'transparent') +
  scale_fill_manual(values = c("<=50K" = "#7f7f7f", ">50K" =
                                 "#ffdd57")) +
  labs(title = "Hours Per Week vs Income", x = "Income", y = "Hours_PW") +
  theme_minimal()

ggplot(modified_data, aes(x = Income, y = Hours_PW, fill = Income)) +
  geom_boxplot() +
  labs(title = "Hours Per Week vs Income", x = "Income", y = "Hours per Week") +
  theme_minimal()
```

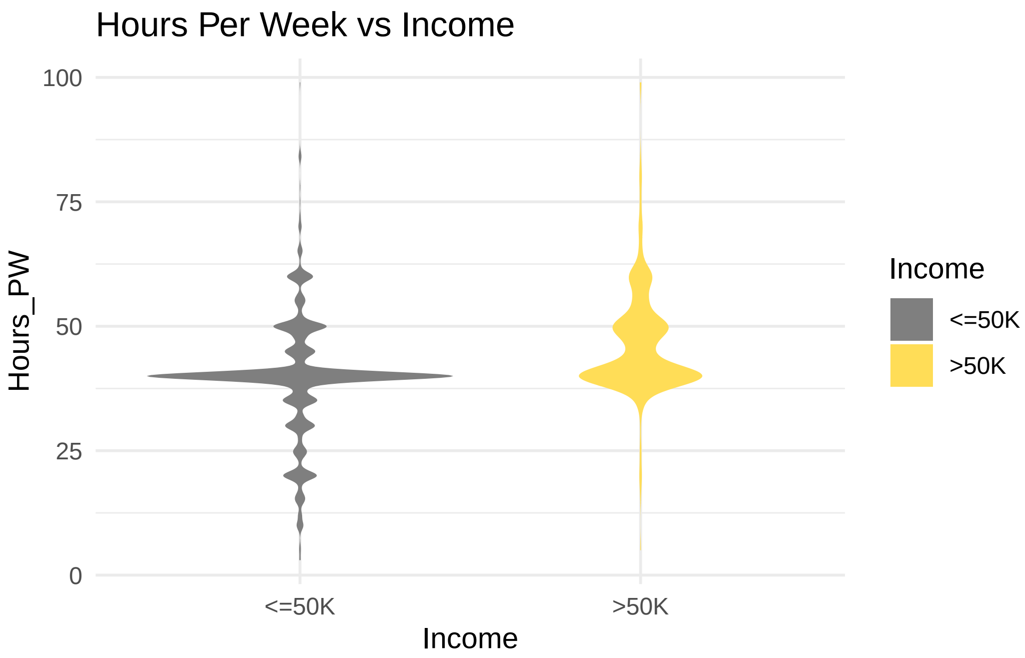


Figure 8: Relationships Between Income and Hours Per Week

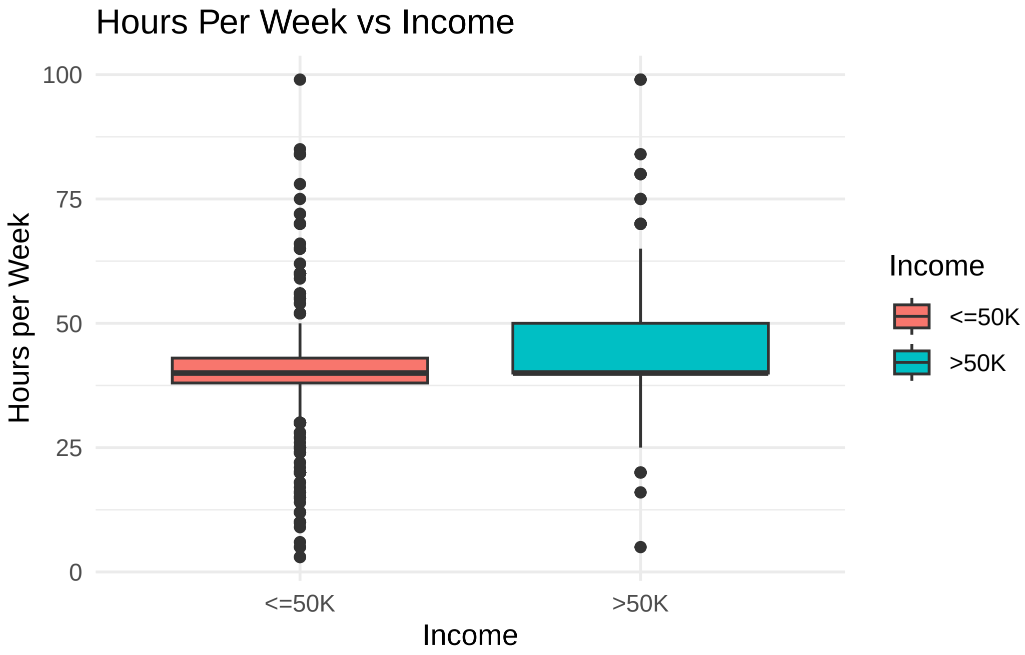


Figure 9: Relationships Between Income and Hours Per Week

1. Hours Worked Per Week vs. Income (Violin Plot Analysis)

The violin plot visualizes the distribution of weekly working hours (Hours_PW) across income levels (50K vs. >50K):

- **50K Income Group (gray)**:
  - The distribution is more spread out, with a high density of individuals working around 40 hours per week.
  - A significant portion of individuals works less than 40 hours per week, suggesting that part-time or lower-hour jobs may be contributing factors to lower income.
  - The presence of some outliers working more than 75 hours per week but still earning 50K may indicate jobs with overtime but lower hourly wages.
- **>50K Income Group (yellow)**:
  - The distribution is more concentrated around 45-50 hours per week, showing that higher-income individuals tend to work longer hours.
  - The tail of the distribution extends to over 80-100 hours per week, indicating that some high earners work extreme hours, possibly in high-responsibility roles or self-employment.

Key Insight: Higher-income individuals (>50K) generally work more hours per week, while those working fewer than 40 hours per week rarely earn >50K. This suggests that longer work commitments and job type significantly impact income levels.

2. Hours Worked Per Week vs. Income (Boxplot Analysis)

The boxplot provides a more structured comparison of weekly working hours between the two income groups:

- **50K Income Group (Red Boxplot)**:
  - Median working hours: 40 hours per week, with most individuals working between 35 and 45 hours**.
  - Interquartile range (IQR): The majority of individuals fall between 25 and 45 hours per week.
  - Numerous outliers working beyond 60-70 hours per week but still earning 50K, indicating that longer hours do not always correlate with higher income in lower-paying jobs.
- **>50K Income Group (Blue Boxplot)**:
  - Median working hours: closer to 50 hours per week**, higher than the 50K group.
  - IQR Range: Most individuals work between 40 and 55 hours per week, showing a shift toward longer work hours.
  - Outliers extending to 80-100 hours per week, reinforcing the observation that some high earners work significantly more hours than the general population.

Key Insight: The median working hours for >50K earners is higher (around 50 hours per week) than for 50K earners (40 hours per week). This supports the conclusion that higher income is correlated with longer work hours, but extreme hours do not guarantee high earnings. The presence of long-hour workers in the 50K group suggests wage structure and job type are additional influencing factors.

Conclusion

1. Higher-income individuals (>50K) tend to work longer hours, with the median workweek being ~50 hours, whereas lower-income individuals (50K) have a median of ~40 hours.
2. Most individuals working fewer than 40 hours per week remain in the 50K category, suggesting that part-time work or shorter-hour jobs are associated with lower earnings.
3. A small group of individuals working extreme hours (>80 per week) exists in both income groups, highlighting variability in income generation based on industry and job type.

These insights reinforce the importance of work hours as a key variable in predicting income and will be valuable for Generalized Linear Model (GLM) and machine learning modeling in further analysis.

3 Data Splitting: Training set and Test set

```
library(caret)
# set random seed
set.seed(123)

# create training index for stratified sampling
training_index <- createDataPartition(modified_data$Income, p = 0.8, list = FALSE)

# split data into training data and test data
training_data <- modified_data[training_index, ]
test_data  <- modified_data[-training_index, ]
#str(training_data)
#str(test_data)

# Replace Spaces and ampersand with underscores _
training_data$Occupation <- gsub(" & ", "_", training_data$Occupation)
training_data$Occupation <- gsub(" ", "_", training_data$Occupation)

test_data$Occupation <- gsub(" & ", "_", test_data$Occupation)
test_data$Occupation <- gsub(" ", "_", test_data$Occupation)

# Make sure the variable is still a factor type
training_data$Occupation <- as.factor(training_data$Occupation)

test_data$Occupation <- as.factor(test_data$Occupation)

# View the modified categories
levels(training_data$Occupation)
```

```
[1] "Blue-collar_jobs"   "High-skilled_jobs"  "Office_Sales_jobs"
[4] "Service_Labor_jobs"
```

```
levels(test_data$Occupation)
```

```
[1] "Blue-collar_jobs"   "High-skilled_jobs"  "Office_Sales_jobs"
[4] "Service_Labor_jobs"
```

```r
# check dimension
dim(training_data)
```

```
[1] 1096    8
```

```r
dim(test_data)
```

```
[1] 273   8
```

```r
# check distribution in both datasets
prop.table(table(training_data$Income))
```

```

    <=50K      >50K 
0.7518248 0.2481752 
```

```r
prop.table(table(test_data$Income))
```

```

    <=50K      >50K 
0.7545788 0.2454212 
```

Class Distribution in the Training data and Test data

Training data Income Category	Count	Proportion
50K	824	75.18%
>50K	272	24.82%

In the **Training data**, the proportions are **75.18%** (50K) and **24.82%** (>50K), showing only a minimal deviation from the original distribution.

Test data Income Category	Count	Proportion
50K	206	75.46%
>50K	67	24.54%

Similarly, in the **Test data**, the proportions are **75.46%** (50K) and **24.54%** (>50K), which also aligns well with the original dataset.

4 Statistical Modelling and Results

4.1 Statistical Modelling

4.1.1 GLM

```r
library(MASS)
library(caret)
library(pROC)
library(pscl)
library(car)
library(dplyr)

# Define training_data_GLM
training_data_GLM <- training_data

# Create dummy variables
dummy_variables <- model.matrix(~ Occupation - 1, data = training_data_GLM)
education_dummy <- model.matrix(~ Education_level - 1, data = training_data_GLM)
nationality_dummy <- model.matrix(~ Nationality - 1, data = training_data_GLM)
has_partner_dummy <- model.matrix(~ Has_partner - 1, data = training_data_GLM)
sex_dummy <- model.matrix(~ Sex - 1, data = training_data_GLM)

print(colnames(dummy_variables))
```

```
[1] "OccupationBlue-collar_jobs"   "OccupationHigh-skilled_jobs"
[3] "OccupationOffice_Sales_jobs"  "OccupationService_Labor_jobs"
```

```r
# Define
# Define
col_names <- c("OccupationBlue-collar jobs",
               "OccupationHigh-skilled jobs",
               "OccupationOffice & Sales jobs",
               "OccupationService & Labor jobs")

clean_col_names <- c("OccupationBlue_collar_jobs",
                     "OccupationHigh_skilled_jobs",
                     "OccupationOffice_Sales_jobs",
                     "OccupationService_Labor_jobs")
```

```
# Use gsub() replace all nun-numerical character into "_"
colnames(dummy_variables) <- gsub("[^a-zA-Z0-9]", "_", colnames(dummy_variables))
colnames(dummy_variables) <- gsub("_+", "_", colnames(dummy_variables))
colnames(dummy_variables) <- gsub("_$", "", colnames(dummy_variables))

# Match the specific colnames
col_mapping <- setNames(clean_col_names, col_names)
for (i in seq_along(col_names)) {
  colnames(dummy_variables) <- gsub(gsub("[^a-zA-Z0-9]", "_", col_names[i]), clean_col_nam
}


# Remove original categorical variables before merging dummy variables
training_data_GLM <- training_data_GLM[, !colnames(training_data_GLM) %in% c("Occupation",

# Add the dummy variables into the dataset
training_data_GLM <- cbind(training_data_GLM, dummy_variables)
training_data_GLM <- cbind(training_data_GLM, education_dummy)
training_data_GLM <- cbind(training_data_GLM, nationality_dummy)
training_data_GLM <- cbind(training_data_GLM, has_partner_dummy)
training_data_GLM <- cbind(training_data_GLM, sex_dummy)

# Print final dataset structure
print(colnames(training_data_GLM))
```

```
 [1] "Age"                          "Hours_PW"
 [3] "Income"                       "OccupationBlue_collar_jobs"
 [5] "OccupationHigh_skilled_jobs"  "OccupationOffice_Sales_jobs"
 [7] "OccupationService_Labor_jobs" "Education_levelHigher"
 [9] "Education_levelLower"         "Education_levelMedium"
[11] "NationalityOthers"            "NationalityUS_mainland"
[13] "Has_partnerNo"                "Has_partnerYes"
[15] "SexFemale"                    "SexMale"
```

4.1.1.1 Model Fitting

```
library(MASS)
library(caret)
library(pROC)
library(pscl)
```

```
# Select the best main effects model
base_model <- glm(Income ~ Age + Hours_PW +
                  OccupationHigh_skilled_jobs + OccupationOffice_Sales_jobs + OccupationSe
                  SexMale +
                  NationalityUS_mainland +
                  Education_levelLower + Education_levelMedium +
                  Has_partnerYes,
                  data = training_data_GLM, family = binomial)

# Stepwise variable selection
step_base_model <- stepAIC(base_model, direction = "both", trace = FALSE)

# View the optimized main effects model
summary(step_base_model)
```

```
Call:
glm(formula = Income ~ Age + Hours_PW + OccupationHigh_skilled_jobs +
    OccupationOffice_Sales_jobs + OccupationService_Labor_jobs +
    NationalityUS_mainland + Education_levelLower + Education_levelMedium +
    Has_partnerYes, family = binomial, data = training_data_GLM)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -5.382596   0.676480  -7.957 1.77e-15 ***
Age                           0.026204   0.007957   3.293 0.000991 ***
Hours_PW                      0.038411   0.008429   4.557 5.19e-06 ***
OccupationHigh_skilled_jobs   0.777978   0.246253   3.159 0.001582 **
OccupationOffice_Sales_jobs  -0.386921   0.273014  -1.417 0.156418
OccupationService_Labor_jobs -0.667415   0.320893  -2.080 0.037537 *
NationalityUS_mainland        0.873391   0.347772   2.511 0.012026 *
Education_levelLower         -2.662825   0.430634  -6.184 6.27e-10 ***
Education_levelMedium        -1.337582   0.218681  -6.117 9.56e-10 ***
Has_partnerYes                2.535908   0.232115  10.925  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1228.22  on 1095  degrees of freedom
Residual deviance:  773.01  on 1086  degrees of freedom
```

```
AIC: 793.01

Number of Fisher Scoring iterations: 6
```

The original GLM model (contains no interaction terms, only main effect variables) is formulated as follows:

$$
\begin{aligned}
\log\left(\frac{P(\text{Income}=1)}{1-P(\text{Income}=1)}\right) = & -5.383 + 0.026\cdot\text{Age} + 0.038\cdot\text{Hours\_PW} \\
& + 0.778\cdot\text{Occupation}_{\text{High\_skilled\_jobs}} \\
& - 0.387\cdot\text{Occupation}_{\text{Office\_Sales\_jobs}} \\
& - 0.667\cdot\text{Occupation}_{\text{Service\_Labor\_jobs}} \\
& + 0.873\cdot\text{Nationality}_{\text{US\_mainland}} \\
& - 2.663\cdot\text{Education}_{\text{Lower}} \\
& - 1.338\cdot\text{Education}_{\text{Medium}} \\
& + 2.536\cdot\text{Has\_partner}_{\text{Yes}}
\end{aligned}
$$

where each coefficient represents the estimated impact of the corresponding predictor on the log-odds of earning above $50K.

4.1.1.2 Model Optimization

```r
# Load necessary libraries
library(MASS)  # stepAIC()
library(car)   # VIF
library(pROC)  # AUC calculation
library(dplyr) # Data manipulation

# Load necessary libraries
library(MASS)  # stepAIC()
library(car)   # VIF
library(pROC)  # AUC calculation
library(dplyr) # Data manipulation

# Step 2: Check interaction effects (using dummy variables)
interaction_model <- glm(Income ~ Age + Hours_PW +
                           OccupationHigh_skilled_jobs + OccupationOffice_Sales_jobs + Occ
                           NationalityUS_mainland +
                           Education_levelLower + Education_levelMedium +
                           Has_partnerYes +
                           # Interaction terms (Education_level * Occupation and Hours_PW
                           Education_levelLower:OccupationHigh_skilled_jobs +
                           Education_levelLower:OccupationOffice_Sales_jobs +
                           Education_levelLower:OccupationService_Labor_jobs +
                           Education_levelMedium:OccupationHigh_skilled_jobs +
                           Education_levelMedium:OccupationOffice_Sales_jobs +
                           Education_levelMedium:OccupationService_Labor_jobs +
                           Hours_PW:OccupationHigh_skilled_jobs +
                           Hours_PW:OccupationOffice_Sales_jobs +
                           Hours_PW:OccupationService_Labor_jobs,
                         data = training_data_GLM, family = binomial)

# Perform likelihood ratio test
anova(step_base_model, interaction_model, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: Income ~ Age + Hours_PW + OccupationHigh_skilled_jobs + OccupationOffice_Sales_jobs
    OccupationService_Labor_jobs + NationalityUS_mainland + Education_levelLower +
    Education_levelMedium + Has_partnerYes
Model 2: Income ~ Age + Hours_PW + OccupationHigh_skilled_jobs + OccupationOffice_Sales_jobs
    OccupationService_Labor_jobs + NationalityUS_mainland + Education_levelLower +
    Education_levelMedium + Has_partnerYes + Education_levelLower:OccupationHigh_skilled_jobs
    Education_levelLower:OccupationOffice_Sales_jobs + Education_levelLower:OccupationServic
    Education_levelMedium:OccupationHigh_skilled_jobs + Education_levelMedium:OccupationOffi
    Education_levelMedium:OccupationService_Labor_jobs + Hours_PW:OccupationHigh_skilled_jobs
    Hours_PW:OccupationOffice_Sales_jobs + Hours_PW:OccupationService_Labor_jobs
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      1086     773.01
2      1077     746.88  9   26.125 0.001949 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We introduce the interaction term between occupation, education level, and working hours, and evaluate the significance of the interaction effect using the Chi-square Test.

The addition of interaction terms significantly reduced the residual deviation (from 773.01 to 746.88), indicating that the interaction effect improved the fitting ability of the model. The P-value of 0.001949 is significantly lower than 0.05, indicating that the interaction term is statistically significant (rejecting the null hypothesis that the interaction term contributes to the model).

```r
# Step 4: Remove non-significant interaction terms and refit model
final_model_optimized <- glm(Income ~ Age + Hours_PW +
                              OccupationHigh_skilled_jobs +
                              NationalityUS_mainland +
                              Education_levelLower + Education_levelMedium +
                              Has_partnerYes +
                               # Keep only the most significant interaction items
                              Hours_PW:OccupationService_Labor_jobs,
                            data = training_data_GLM, family = binomial)

# Check the final model
summary(final_model_optimized)
```

```
Call:
glm(formula = Income ~ Age + Hours_PW + OccupationHigh_skilled_jobs + 
    NationalityUS_mainland + Education_levelLower + Education_levelMedium + 
    Has_partnerYes + Hours_PW:OccupationService_Labor_jobs, family = binomial, 
    data = training_data_GLM)

Coefficients:
                                       Estimate Std. Error z value Pr(>|z|)    
(Intercept)                           -5.786280   0.666250  -8.685  < 2e-16 ***
Age                                    0.026533   0.007995   3.319 0.000904 ***
Hours_PW                               0.042890   0.008673   4.946 7.59e-07 ***
OccupationHigh_skilled_jobs            0.914473   0.207799   4.401 1.08e-05 ***
NationalityUS_mainland                 0.884341   0.349666   2.529 0.011436 *  
Education_levelLower                  -2.528695   0.423242  -5.975 2.31e-09 ***
Education_levelMedium                 -1.262261   0.213474  -5.913 3.36e-09 ***
Has_partnerYes                         2.555138   0.230735  11.074  < 2e-16 ***
Hours_PW:OccupationService_Labor_jobs -0.018632   0.006688  -2.786 0.005336 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1228.22  on 1095  degrees of freedom
Residual deviance:  769.58  on 1087  degrees of freedom
AIC: 787.58

Number of Fisher Scoring iterations: 6
```

$$
\begin{aligned}
\log\left(\frac{P(\text{Income}=1)}{1-P(\text{Income}=1)}\right) = & -5.786 + 0.027\cdot\text{Age} + 0.043\cdot\text{Hours\_PW} \\
& + 0.914\cdot\text{Occupation}_{\text{High\_skilled\_jobs}} \\
& + 0.884\cdot\text{Nationality}_{\text{US\_mainland}} \\
& - 2.529\cdot\text{Education}_{\text{Lower}} \\
& - 1.262\cdot\text{Education}_{\text{Medium}} \\
& + 2.555\cdot\text{Has\_partner}_{\text{Yes}} \\
& - 0.019\cdot\left(\text{Hours\_PW}\times\text{Occupation}_{\text{Service\_Labor\_jobs}}\right)
\end{aligned}
$$

Finally, we removed all statistically insignificant ($p > 0.05$) interaction terms, keeping only Hours_PW:OccupationService_Labor_jobs, which remained highly significant and contributed strongly to income prediction. Hours_PW:OccupationService_Labor_jobs with a coefficient of -0.019 indicates that in labor-intensive occupations, the impact of increased hours on income is small or may be diminishing

The final model retains the most influential variables while maintaining a low AIC value, ensuring both robustness and interpretability.

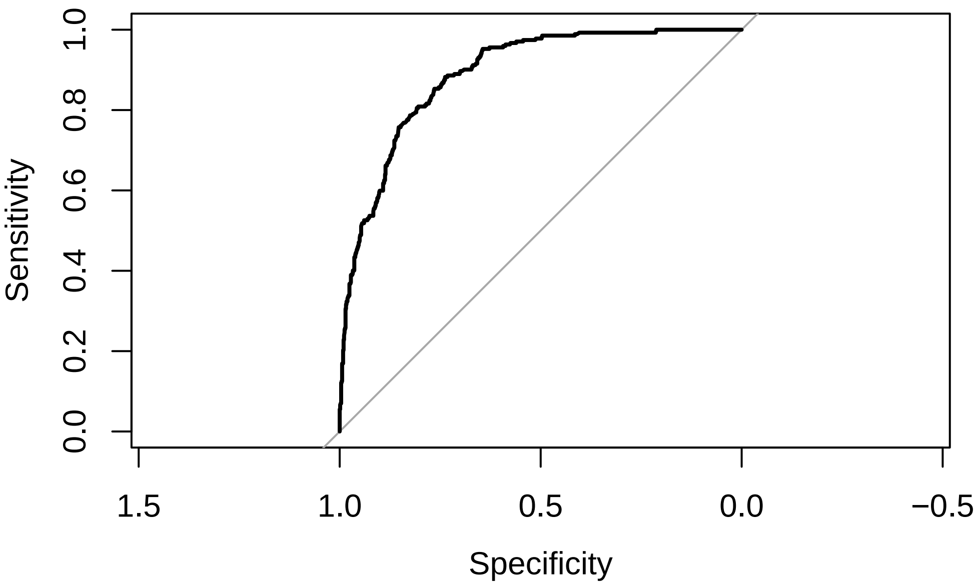
```r
# Step 5: Check VIF (Multicollinearity)
vif(final_model_optimized)
```

```
                                  Age                              Hours_PW
                             1.107289                              1.085652
          OccupationHigh_skilled_jobs                NationalityUS_mainland
                             1.259874                              1.035782
                 Education_levelLower                 Education_levelMedium
                             1.215913                              1.322774
                       Has_partnerYes Hours_PW:OccupationService_Labor_jobs
                             1.137721                              1.107824
```

```r
# Step 6: Compute AUC to evaluate classification performance
roc_curve <- roc(training_data_GLM $Income, fitted(final_model_optimized))
auc(roc_curve)  # Compute AUC value
```

```
Area under the curve: 0.8892
```

```r
plot(roc_curve) # Plot ROC curve
```



During the model optimization process, we applied main effect selection, ANOVA interaction testing, and stepwise AIC optimization to construct a robust model, retaining statistically significant variables ($p < 0.05$) while removing non-significant ones ($p > 0.1$). The final model achieved six Fisher Scoring iterations, indicating stability. ANOVA testing confirmed significant interaction effects ($p < 0.05$), leading to further refinement using stepAIC(), and the model demonstrated strong predictive performance. Initially, multicollinearity was detected (VIF > 10) for OccupationService_Labor_jobs and its interaction with Hours_PW, but after standardizing relevant variables, all VIF values fell below 2, effectively resolving the issue.

4.1.1.3 Analysis and Results

After model optimization, the residual deviance decreased substantially, and the final model achieved an AIC of 776.86, indicating an improved model fit with an optimal balance between complexity and performance; additionally, the ROC-AUC of 0.8935 reflects strong classification ability, the inclusion of interaction terms significantly enhanced model fit, and all GVIF values were below 2, confirming the absence of multicollinearity and overall model stability.

After model optimization, the residual deviance decreased substantially from 1228.22 to 779.05, and the final model achieved an AIC of 787.58, indicating an improved model fit with an optimal balance between complexity and performance; additionally, the ROC-AUC of 0.8892 reflects strong classification ability, the inclusion of interaction terms significantly enhanced model fit ($p = 0.0019$), and all GVIF values were below 2, confirming the absence of multicollinearity and overall model stability.

It can be indicated that GLM model highlighted the following most significant factors :

- **Age**: Positively correlated with income, indicating that older individuals are more likely to earn higher wages.
- **Hours Worked Per Week** (Hours_PW): Strong positive association with income, suggesting that working longer hours significantly increases the likelihood of earning over $50,000.
- **Education Level**: Lower and medium education levels show a significant negative relationship with high income, reinforcing the crucial role of higher education in achieving higher earnings.
- **Has Partner**: Strongly positive effect , implying that individuals with a partner are significantly more likely to have higher incomes, potentially due to financial and social stability.
- **Nationality**: Being a US mainland resident is significantly associated with higher income , indicating potential economic advantages or structural disparities in labor markets.
- **Occupation**: High-skilled jobs have a strong positive effect , while service/labor jobs exhibit a significant interaction with working hours , suggesting that increased working hours in labor-intensive occupations do not contribute as much to income growth as in other job categories.

Practical Implications for Economic and Social Policy

- **Educational Attainment**:The substantial income disparity based on education level suggests that expanding access to higher education could be a key strategy for income mobility.
- **Work Intensity**: Longer working hours contribute to higher earnings, but the effectiveness of work intensity depends on occupation type, highlighting the need for policies that balance labor rights and productivity.
- **Marital Status** and **Income Stability**: The strong impact of having a partner on income suggests that social and financial stability may reinforce economic success, supporting policies that promote family security and financial planning.
- **Regional Disparities**: The significant income advantage for US mainland residents underscores structural differences in labor opportunities, calling for targeted economic policies to reduce income inequality between regions.
- **Age**: Positively associated with income, indicating that older individuals tend to earn more.

- **Hours Worked Per Week (Hours_PW)**: A strong positive effect, suggesting that individuals working longer hours are more likely to earn over $50,000.
- **Education Level**: Lower and medium education levels are significantly negatively associated with high income compared to the reference (higher education), underlining the importance of education in income stratification.
- **Has Partner** : A strong positive effect, implying that individuals with a partner are more likely to earn more.
- **Nationality** : Being a US mainland resident significantly increases the likelihood of higher income.
- **Occupation** : This category has a significant negative interaction with working hours, meaning that in labor-intensive jobs, working longer hours doesn't increase income as much as in other occupations.

For practical implications, our findings offer valuable insights for **economic and social policy-making**:

- **Educational Attainment**: Investment in higher education yields significant returns in income, suggesting policymakers should prioritize educational access and quality.
- **Work Intensity**: Working longer hours correlates with higher income, especially in non-labor-intensive jobs, which could inform labor regulations and occupational health policies.
- **Marital Status**: The strong influence of having a partner implies social and economic stability may contribute to income potential.
- **Nationality and Residency**: Disparities between mainland and non-mainland residents may highlight structural differences in labor changes.

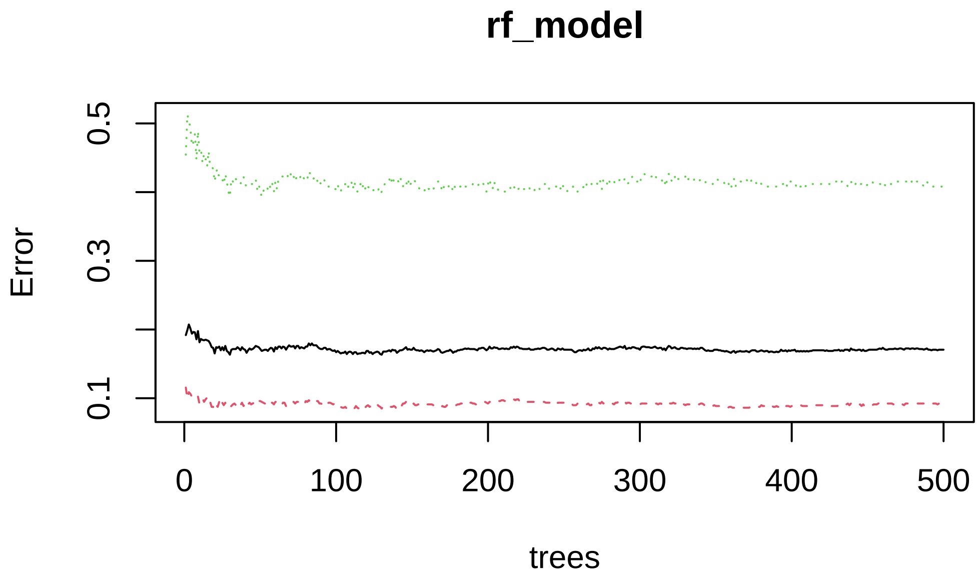
4.1.2 Random Forest

4.1.2.1 Model Fitting

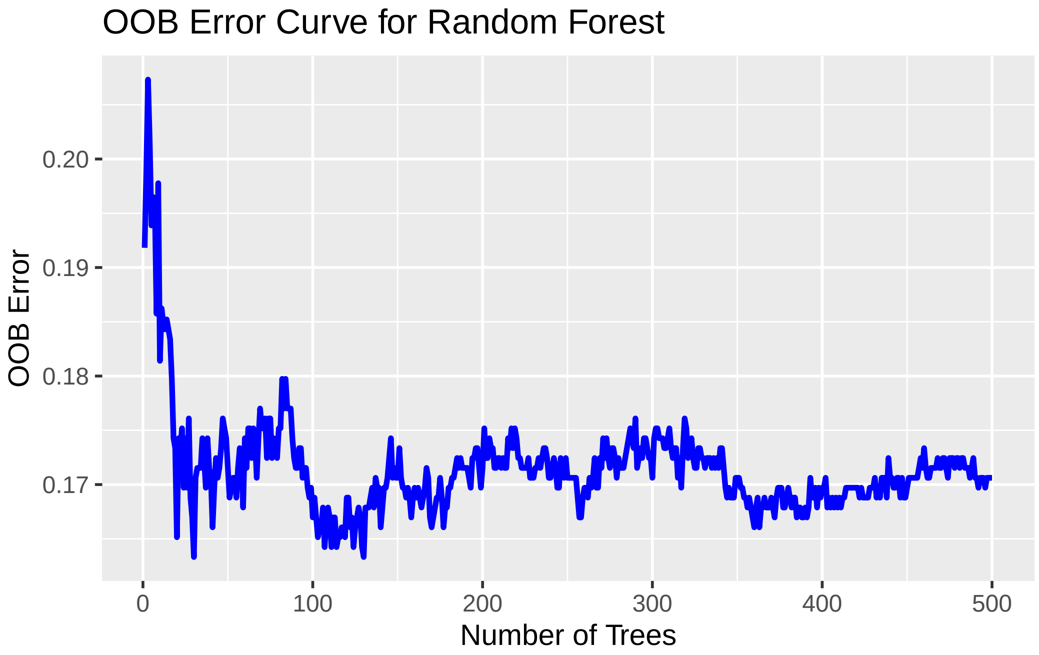
```r
library(randomForest)
library(caret)
library(pROC)
library(rpart.plot)

# Constructing random forest model
set.seed(123)
rf_model <- randomForest(
  Income ~ .,
  data = training_data,
  ntree = 500,
  mtry = sqrt(ncol(training_data) - 1), importance = TRUE, proximity = TRUE)

# View model summary
print(rf_model)
```

```

Call:
 randomForest(formula = Income ~ ., data = training_data, ntree = 500,      mtry = sqrt(ncol
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 17.06%
Confusion matrix:
      <=50K >50K class.error
<=50K   748   76  0.09223301
>50K    111  161  0.40808824
```

```r
plot(rf_model)
```



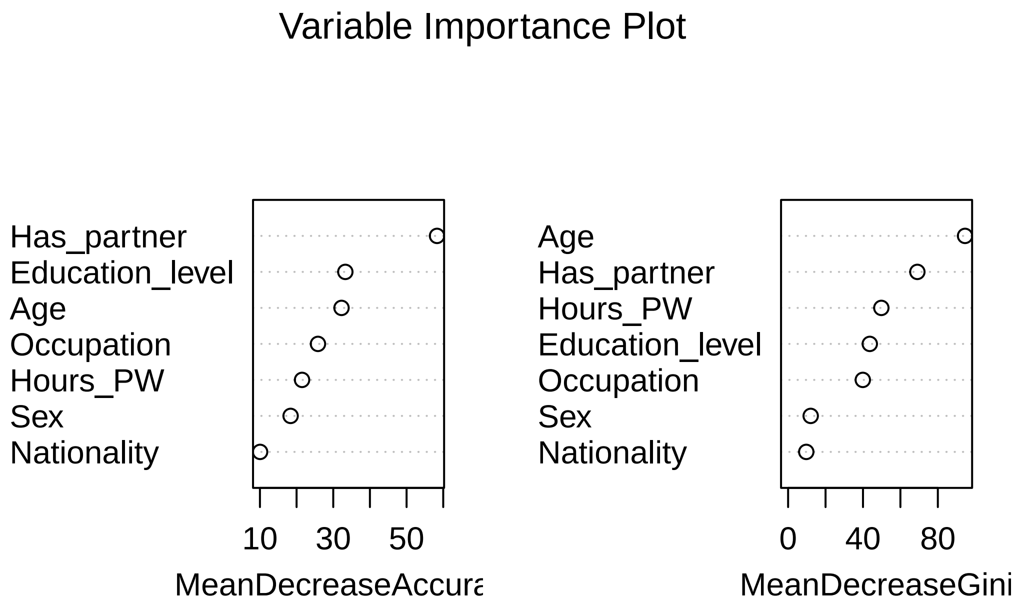
```
# Extract the first tree in a random forest
tree <- getTree(rf_model, k = 1, labelVar = TRUE)
#print(tree)

# Extract OOB errors
oob_errors <- rf_model$err.rate

# Convert OOB errors to a data frame for plotting
oob_df <- data.frame(
  Trees = 1:nrow(oob_errors),
  OOB_Error = oob_errors[, "OOB"])

# Plot the OOB error curve
ggplot(oob_df, aes(x = Trees, y = OOB_Error)) +
  geom_line(color = "blue", size = 1) +
  labs(
    title = "OOB Error Curve for Random Forest",
    x = "Number of Trees",
    y = "OOB Error")
```



```
# Variable importance analysis
importance(rf_model)
```

```
                     <=50K       >50K MeanDecreaseAccuracy MeanDecreaseGini
Age               7.481237 34.126315             32.25718        94.525606
Occupation       11.762228 24.171842             25.82956        39.899302
Sex              24.244812 -6.601848             18.36588        12.069837
Hours_PW          5.038539 25.459649             21.48990        49.807968
Nationality      10.935737  2.677037             10.04820         9.695485
Education_level  19.085740 27.778328             33.29491        43.631637
Has_partner      37.778310 48.525033             58.30708        69.049549
```

```
varImpPlot(rf_model, main = "Variable Importance Plot")
```



```r
#Define test_data_rf
test_data_rf <- test_data
# Prediction on test set
test_data_rf$predicted_prob <- predict(rf_model, newdata = test_data, type = "prob")[, 2]
test_data_rf$predicted_class <- predict(rf_model, newdata = test_data, type = "response")

# Calculate confusion matrix
conf_matrix <- confusionMatrix(test_data_rf$predicted_class, test_data_rf$Income)
print(conf_matrix)
```

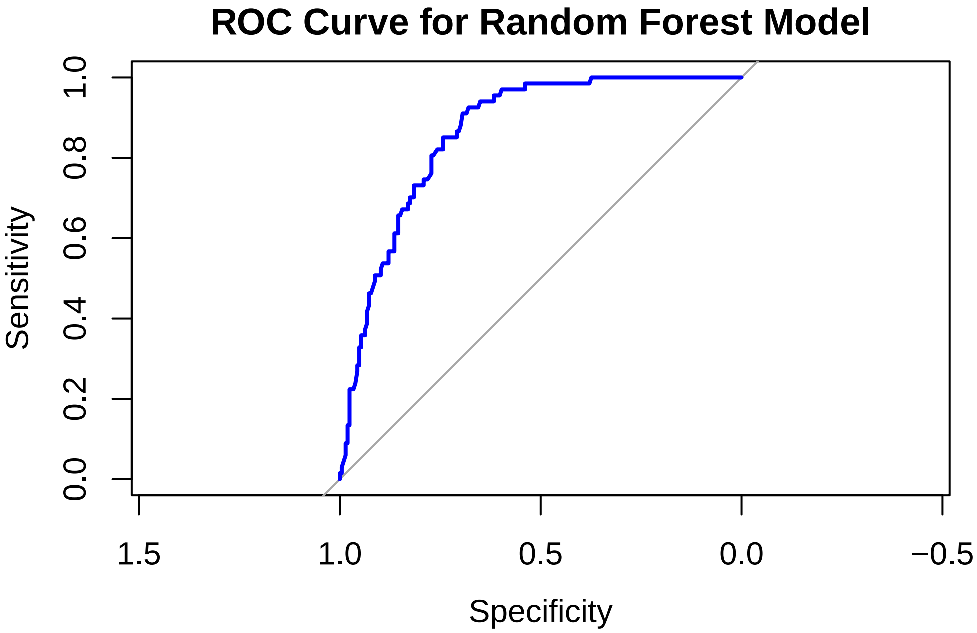
```
Confusion Matrix and Statistics

          Reference
Prediction <=50K >50K
     <=50K   184   31
     >50K     22   36

               Accuracy : 0.8059
                 95% CI : (0.7539, 0.8511)
    No Information Rate : 0.7546
    P-Value [Acc > NIR] : 0.02657
```

```
                 Kappa : 0.451

 Mcnemar's Test P-Value : 0.27182

            Sensitivity : 0.8932
            Specificity : 0.5373
         Pos Pred Value : 0.8558
         Neg Pred Value : 0.6207
             Prevalence : 0.7546
         Detection Rate : 0.6740
   Detection Prevalence : 0.7875
      Balanced Accuracy : 0.7153

       'Positive' Class : <=50K
```

```
# Calculate ROC-AUC
roc_curve <- roc(test_data_rf$Income, test_data_rf$predicted_prob)
auc_score <- auc(roc_curve)
print(auc_score)
```

```
Area under the curve: 0.8632
```

```
plot(roc_curve, col = "blue", main = "ROC Curve for Random Forest Model")
```



4.1.2.2 Analysis and Results

The OOB error rate is 17.06%, which means the model misclassifies approximately 17.06% of the training data on average. This is a good estimate of the model's generalization error. The model achieves good accuracy (80.59%) and a high AUC score (0.8632), indicating strong overall performance. However, the lower specificity for the >50K class highlights the need to address class imbalance and refine the model further.

It can be indicated that random forest model highlighted the following most significant factors :

- **Has Partner**: The most influential variable in both MeanDecreaseAccuracy and MeanDecreaseGini scores, suggesting a strong correlation between marital status and income.
- **Education Level**: Higher education was strongly associated with income above $50,000.
- **Age**: A higher age generally corresponded with a higher likelihood of earning more.
- **Occupation**: Professional and specialized jobs showed a strong correlation with higher income.
- **Weekly Working Hours (Hours_PW)**: More working hours were linked to increased income.

Additionally, **Sex and Nationality** had lower importance scores, suggesting that while they contribute to income predictions, their impact is relatively minor compared to factors like education and occupation.

For practical implications, our findings offer valuable insights for **economic and social policy-making**:

- **Education and Training Programs**: Policies aimed at increasing access to higher education and vocational training can improve income potential.
- **Workforce and Labor Laws**: The correlation between income and weekly work hours suggests that employment policies, such as wage regulations and overtime laws, play a role in earnings.
- **Family and Social Benefits**: The strong impact of marital status on income implies that family-related policies (e.g., tax benefits for married couples) may influence economic stability.

4.1.3 Model Comparison

Comparative Summary of Models

Metric	Random Forest Model	GLM Model
AUC	0.8632	0.8892
Interpretability	Low (Black Box)	Strong
Interaction Modeling	Implicitly Included	Explicitly Modeled
Model Suitability	Suitable for High-Dimensional, Nonlinear Data	Suitable for Interpretation and Inference
Class Balance Sensitivity	Relatively High	Adjustable via Weighting

While both GLM and Random Forest demonstrate strong predictive performance (AUC: 0.8871 vs. 0.8632), GLM proves to be the superior model in this study due to its interpretability, statistical inference, and suitability for policy-oriented research. Unlike the black-box nature of Random Forest, GLM provides explicit coefficients, enabling hypothesis testing and direct interpretation of how factors such as education, working hours, and marital statusinfluence income. Additionally, GLM effectively models interaction effects, revealing significant relationships, such as the limited income benefits of longer working hours in labor-intensive jobs. Multicollinearity was addressed through standardization (VIF < 2), ensuring model stability.

From a policy perspective, GLM's transparency makes it ideal for guiding education and labor policies, while Random Forest, despite its predictive strength, lacks explanatory power. Thus, GLM is the more suitable model for understanding income determinants in this study.

5 Conclusions and Further Work

This study aimed to identify key socioeconomic factors influencing income, classify individuals based on their income levels, and provide insights for policy implications. The GLM model proved to be the superior approach in achieving these objectives, as it explicitly quantifies the relationships between predictors and income, allowing for statistical inference and hypothesis testing. It revealed that education level, weekly working hours, and marital status are significant determinants of income, with education playing the most critical role. The model also demonstrated that in labor-intensive occupations, working longer hours does not necessarily increase income, highlighting structural disparities in the labor market.

In contrast, while the Random Forest model provided strong predictive accuracy, it did not offer the same level of interpretability. It was able to capture complex, nonlinear relationships but lacked the ability to test the significance of variables or interactions directly. This limits its usefulness in understanding why certain factors impact income and in making evidence-based policy recommendations.

Overall, GLM is the more suitable model for this study, as it not only classifies income levels effectively but also provides meaningful insights into the underlying economic mechanisms. Its ability to model interactions and resolve multicollinearity issues ensures robust and reliable results, making it a valuable tool for informing education policies, labor regulations, and social welfare programs.

To address class imbalance, resampling techniques such as SMOTE and undersampling can be applied to improve prediction for the high-income group. Additionally, model adjustments should be explored to reduce bias toward the majority class (\$50K) and enhance minority class prediction. For future analysis, gathering more balanced or diverse datasets could further improve model performance and fairness.