

# Determinants of High Income: A Generalized Linear Model Analysis of 1994 US Census Socioeconomic Factors

Group 27

## 1. Introduction

Income inequality remains a critical socioeconomic challenge. This study analyzes the 1994 US Census dataset using a generalized linear model (GLM) to identify key factors—such as age, education, occupation, and work hours—that significantly predict whether an individual earns over \$50k annually. The results aim to inform targeted policy interventions for improving economic equity.

```
library(dplyr)
```

Warning: package 'dplyr' was built under R version 4.4.2

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

`filter`, `lag`

The following objects are masked from 'package:base':

`intersect`, `setdiff`, `setequal`, `union`

```
library(tidyverse)
```

Warning: package 'tidyverse' was built under R version 4.4.2

Warning: package 'tidyr' was built under R version 4.4.2

Warning: package 'readr' was built under R version 4.4.2

Warning: package 'purrr' was built under R version 4.4.2

Warning: package 'forcats' was built under R version 4.4.2

Warning: package 'lubridate' was built under R version 4.4.2

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
```

```
v forcats   1.0.0      v readr      2.1.5
```

```
v ggplot2   3.5.1      v stringr    1.5.1
```

```
v lubridate 1.9.4      v tibble     3.2.1
```

```
v purrr     1.0.2      v tidyr      1.3.1
```

```
-- Conflicts ----- tidyverse_conflicts() --
```

```
x dplyr::filter() masks stats::filter()
```

```
x dplyr::lag()     masks stats::lag()
```

```
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
library(broom)
```

Warning: package 'broom' was built under R version 4.4.2

```
library(ggplot2)
```

```
library(car)
```

Warning: package 'car' was built under R version 4.4.3

Loading required package: carData

Warning: package 'carData' was built under R version 4.4.3

Attaching package: 'car'

The following object is masked from 'package:purrr':

some

The following object is masked from 'package:dplyr':

recode

```
library(janitor)
```

Warning: package 'janitor' was built under R version 4.4.2

Attaching package: 'janitor'

The following objects are masked from 'package:stats':

chisq.test, fisher.test

```
library(broom) # Model results sorting  
library(ResourceSelection) # Hosmer-Lemeshow test
```

Warning: package 'ResourceSelection' was built under R version 4.4.3

ResourceSelection 0.3-6      2023-06-27

```
library(pROC) # ROC curve
```

Warning: package 'pROC' was built under R version 4.4.3

Type 'citation("pROC")' for a citation.

Attaching package: 'pROC'

The following objects are masked from 'package:stats':

cov, smooth, var

## 2. Data processing

### 2.1 Data import

Select the raw data, remove the NA value from the data, and determine the corresponding data type. Among all variables, only Age and Hours\_PW are numerical variables. The integrated data is output to 'cleaned\_dataset27' for preliminary judgment of data and variables.

```
# load data
data = read.csv("../datasets\\dataset27.csv")
```

```
# inspect data
str(data)
```

```
'data.frame':  1500 obs. of  8 variables:
 $ Age      : int  37 49 20 64 60 32 58 42 20 42 ...
 $ Education : chr  "Bachelors," "Some-college," "Some-college," "Bachelors," ...
 $ Marital_Status: chr  "Married-civ-spouse," "Divorced," "Never-married," "Married-civ-spou
 $ Occupation  : chr  "Exec-managerial," "Machine-op-inspct," "Other-service," "Exec-manag
 $ Sex         : chr  "Male," "Male," "Male," "Male," ...
 $ Hours_PW    : chr  "80," "45," "45," "55," ...
 $ Nationality : chr  "United-States," "United-States," "United-States," "United-States,"
 $ Income      : chr  ">50K" "<=50K" "<=50K" ">50K" ...
```

```
summary(data)
```

Age	Education	Marital_Status	Occupation
Min. :17.00	Length:1500	Length:1500	Length:1500
1st Qu.:27.00	Class :character	Class :character	Class :character
Median :37.50	Mode :character	Mode :character	Mode :character
Mean :38.91			
3rd Qu.:48.00			
Max. :90.00			
Sex	Hours_PW	Nationality	Income
Length:1500	Length:1500	Length:1500	Length:1500
Class :character	Class :character	Class :character	Class :character
Mode :character	Mode :character	Mode :character	Mode :character

```
sapply(data, unique)
```

```
$Age
```

```
[1] 37 49 20 64 60 32 58 42 25 30 66 59 44 46 17 47 26 33 48 52 40 50 34 43 22  
[26] 35 62 39 69 27 38 41 23 56 24 21 18 45 75 51 29 19 72 31 28 65 55 57 67 36  
[51] 54 61 71 53 73 63 70 68 90 77 74 76 78 80 81
```

```
$Education
```

```
[1] "Bachelors,"      "Some-college," "Masters,"      "Assoc-acdm,"  
[5] "HS-grad,"        "Assoc-voc,"    "9th,"          "10th,"  
[9] "Prof-school,"    "5th-6th,"      "7th-8th,"      "Doctorate,"  
[13] "12th,"           "11th,"         "1st-4th,"      "Preschool,"
```

```
$Marital_Status
```

```
[1] "Married-civ-spouse," "Divorced,"      "Never-married,"  
[4] "Married-spouse-absent," "Separated,"      "Widowed,"  
[7] "Married-AF-spouse,"
```

```
$Occupation
```

```
[1] "Exec-managerial," "Machine-op-inspct," "Other-service,"  
[4] "?,"              "Craft-repair,"      "Sales,"  
[7] "Adm-clerical,"    "Prof-specialty,"    "Protective-serv,"  
[10] "Farming-fishing," "Handlers-cleaners," "Transport-moving,"  
[13] "Tech-support,"    "Priv-house-serv,"
```

```
$Sex
```

```
[1] "Male,"    "Female,"
```

```
$Hours_PW
```

```
[1] "80," "45," "55," "40," "42," "50," "14," "65," "35," "58," "30," "9,"  
[13] "15," "12," "70," "24," "8," "6," "16," "52," "20," "10," "60," "36,"  
[25] "21," "48," "38," "18," "43," "39," "90," "72," "32," "49," "56," "84,"  
[37] "44," "54," "25," "37," "99," "46," "3," "17," "47," "4," "5," "53,"  
[49] "28," "75," "68," "22," "27," "13," "11," "34," "33," "98," "2," "85,"
```

```
$Nationality
```

```
[1] "United-States,"      "England,"  
[3] "China,"              "?,"  
[5] "Outlying-US(Guam-USVI-etc)," "Jamaica,"  
[7] "Scotland,"           "Philippines,"  
[9] "Haiti,"              "Mexico,"  
[11] "Italy,"              "Cuba,"
```

[13] "Canada,"	"Japan,"
[15] "Puerto-Rico,"	"Trinidad&Tobago,"
[17] "Dominican-Republic,"	"El-Salvador,"
[19] "Greece,"	"Nicaragua,"
[21] "South,"	"Germany,"
[23] "Cambodia,"	"Honduras,"
[25] "India,"	"Vietnam,"
[27] "Poland,"	"Taiwan,"
[29] "Guatemala,"	"Portugal,"
[31] "Iran,"	"Hong,"
[33] "Columbia,"	"Ireland,"

\$Income

[1] ">50K" "<=50K"

```
# delete commas
# convert ? to NA
data <- data %>%
  mutate_all(~sub(",$", "", .)) %>%
  mutate_all(~na_if(., "?"))
```

```
# inspect again
sapply(data, unique)
```

\$Age

[1] "37" "49" "20" "64" "60" "32" "58" "42" "25" "30" "66" "59" "44" "46" "17"  
 [16] "47" "26" "33" "48" "52" "40" "50" "34" "43" "22" "35" "62" "39" "69" "27"  
 [31] "38" "41" "23" "56" "24" "21" "18" "45" "75" "51" "29" "19" "72" "31" "28"  
 [46] "65" "55" "57" "67" "36" "54" "61" "71" "53" "73" "63" "70" "68" "90" "77"  
 [61] "74" "76" "78" "80" "81"

\$Education

[1] "Bachelors"	"Some-college"	"Masters"	"Assoc-acdm"	"HS-grad"
[6] "Assoc-voc"	"9th"	"10th"	"Prof-school"	"5th-6th"
[11] "7th-8th"	"Doctorate"	"12th"	"11th"	"1st-4th"
[16] "Preschool"				

\$Marital\_Status

[1] "Married-civ-spouse"	"Divorced"	"Never-married"
[4] "Married-spouse-absent"	"Separated"	"Widowed"
[7] "Married-AF-spouse"		

\$Occupation

```
[1] "Exec-managerial" "Machine-op-inspct" "Other-service"
[4] NA                "Craft-repair"      "Sales"
[7] "Adm-clerical"    "Prof-specialty"    "Protective-serv"
[10] "Farming-fishing" "Handlers-cleaners" "Transport-moving"
[13] "Tech-support"    "Priv-house-serv"
```

\$Sex

```
[1] "Male" "Female"
```

\$Hours\_PW

```
[1] "80" "45" "55" "40" "42" "50" "14" "65" "35" "58" "30" "9" "15" "12" "70"
[16] "24" "8" "6" "16" "52" "20" "10" "60" "36" "21" "48" "38" "18" "43" "39"
[31] "90" "72" "32" "49" "56" "84" "44" "54" "25" "37" "99" "46" "3" "17" "47"
[46] "4" "5" "53" "28" "75" "68" "22" "27" "13" "11" "34" "33" "98" "2" "85"
```

\$Nationality

```
[1] "United-States" "England"
[3] "China"         NA
[5] "Outlying-US(Guam-USVI-etc)" "Jamaica"
[7] "Scotland"      "Philippines"
[9] "Haiti"         "Mexico"
[11] "Italy"         "Cuba"
[13] "Canada"        "Japan"
[15] "Puerto-Rico"  "Trinidad&Tobago"
[17] "Dominican-Republic" "El-Salvador"
[19] "Greece"        "Nicaragua"
[21] "South"         "Germany"
[23] "Cambodia"      "Honduras"
[25] "India"         "Vietnam"
[27] "Poland"        "Taiwan"
[29] "Guatemala"     "Portugal"
[31] "Iran"          "Hong"
[33] "Columbia"      "Ireland"
```

\$Income

```
[1] ">50K" "<=50K"
```

`str(data)`

```
'data.frame': 1500 obs. of 8 variables:
 $ Age      : chr "37" "49" "20" "64" ...
```

```

$ Education      : chr "Bachelors" "Some-college" "Some-college" "Bachelors" ...
$ Marital_Status: chr "Married-civ-spouse" "Divorced" "Never-married" "Married-civ-spouse"
$ Occupation     : chr "Exec-managerial" "Machine-op-inspct" "Other-service" "Exec-managerial"
$ Sex            : chr "Male" "Male" "Male" "Male" ...
$ Hours_PW       : chr "80" "45" "45" "55" ...
$ Nationality    : chr "United-States" "United-States" "United-States" "United-States" ...
$ Income         : chr ">50K" "<=50K" "<=50K" ">50K" ...

```

```

# find that the type of Age and Hour_PW are char,
# then convert to num
data <- data %>%
  mutate_at(vars(Age, Hours_PW), as.numeric)
str(data)

```

```

'data.frame': 1500 obs. of 8 variables:
 $ Age      : num 37 49 20 64 60 32 58 42 20 42 ...
 $ Education: chr "Bachelors" "Some-college" "Some-college" "Bachelors" ...
 $ Marital_Status: chr "Married-civ-spouse" "Divorced" "Never-married" "Married-civ-spouse"
 $ Occupation : chr "Exec-managerial" "Machine-op-inspct" "Other-service" "Exec-managerial"
 $ Sex       : chr "Male" "Male" "Male" "Male" ...
 $ Hours_PW  : num 80 45 45 55 40 42 40 50 14 40 ...
 $ Nationality: chr "United-States" "United-States" "United-States" "United-States" ...
 $ Income    : chr ">50K" "<=50K" "<=50K" ">50K" ...

```

```

# query and delete missing values
colSums(is.na(data))

```

Age	Education	Marital_Status	Occupation	Sex
0	0	0	97	0
Hours_PW	Nationality	Income		
0	29	0		

```

data <- na.omit(data)
str(data)

```

```

'data.frame': 1376 obs. of 8 variables:
 $ Age      : num 37 49 20 64 32 58 42 20 42 25 ...
 $ Education: chr "Bachelors" "Some-college" "Some-college" "Bachelors" ...
 $ Marital_Status: chr "Married-civ-spouse" "Divorced" "Never-married" "Married-civ-spouse"
 $ Occupation : chr "Exec-managerial" "Machine-op-inspct" "Other-service" "Exec-managerial"

```



```
$ Sex          : chr  "Male" "Male" "Male" "Male" ...
$ Hours_PW     : num  80 45 45 55 42 40 50 14 40 40 ...
$ Nationality  : chr  "United-States" "United-States" "United-States" "United-States" ...
$ Income       : chr  ">50K" "<=50K" "<=50K" ">50K" ...
- attr(*, "na.action")= 'omit' Named int [1:124] 5 45 55 58 63 70 71 75 80 82 ...
..- attr(*, "names")= chr [1:124] "5" "45" "55" "58" ...
```

```
# save data
write.csv(data, "..\\datasets\\cleaned_dataset27.csv",
          row.names = FALSE)
```

## 2.2 Variable judgment

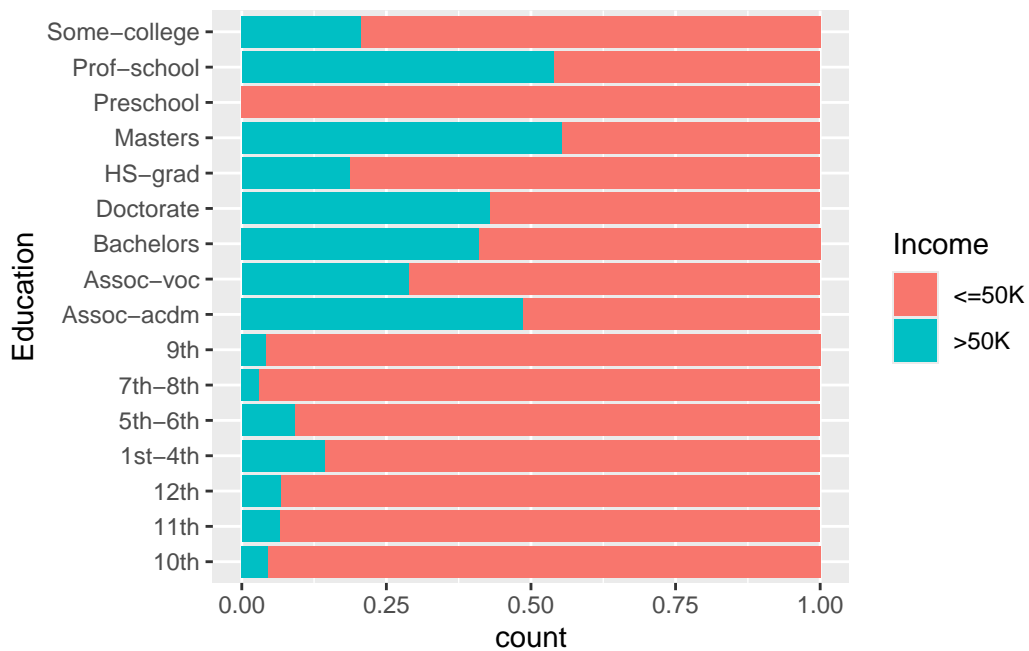
By visualizing all the variables first, you can see that there are some variables with perfectly linear relationships among the classes (for example, Preschool in Education and Priv-house-serv in Occupation).

```
clean_data <- read.csv("..\\datasets\\cleaned_dataset27.csv")

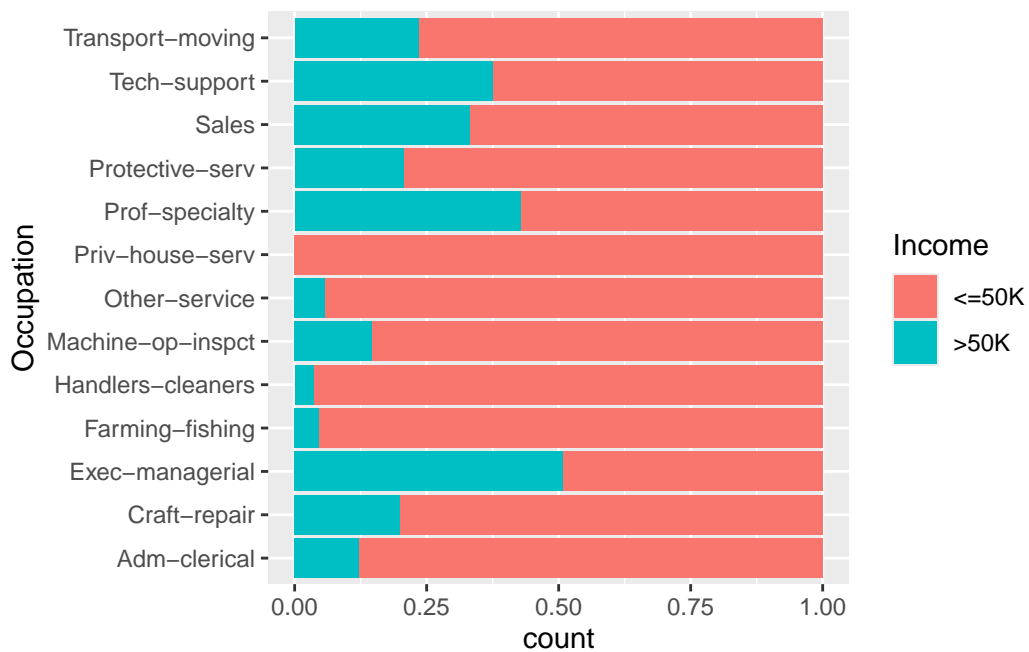
# Relationship between categorical variables and income
lapply(c("Education", "Occupation", "Sex", "Marital_Status"),
       function(var) {
         ggplot(clean_data, aes_string(x = var, fill = "Income")) +
           geom_bar(position = "fill") +
           coord_flip()
       })
```

Warning: `aes\_string()` was deprecated in ggplot2 3.0.0.  
 i Please use tidy evaluation idioms with `aes()`.  
 i See also `vignette("ggplot2-in-packages")` for more information.

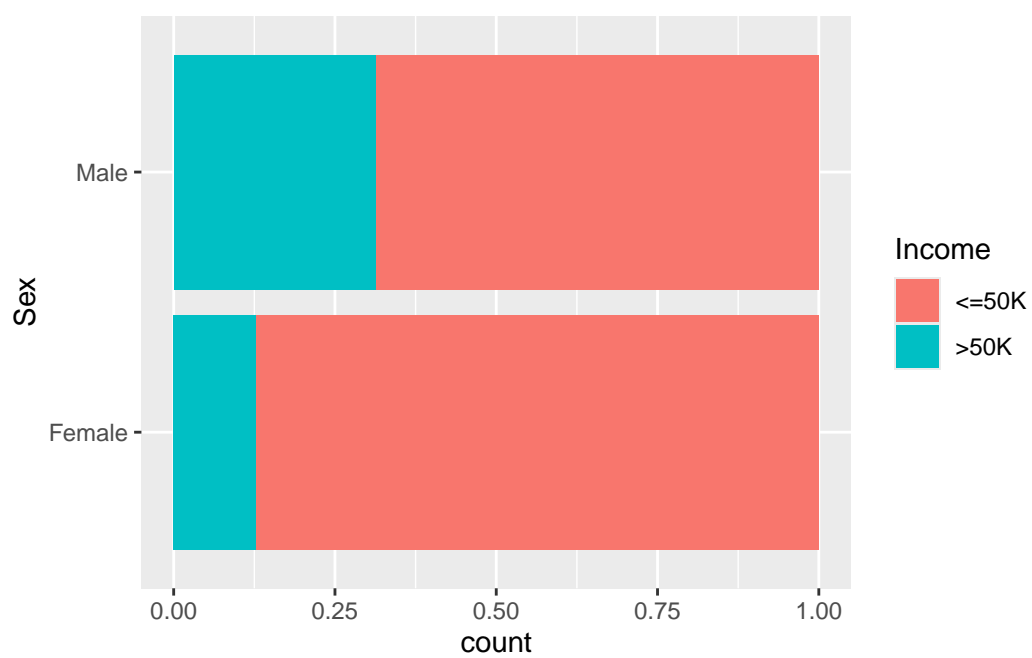
```
[[1]]
```



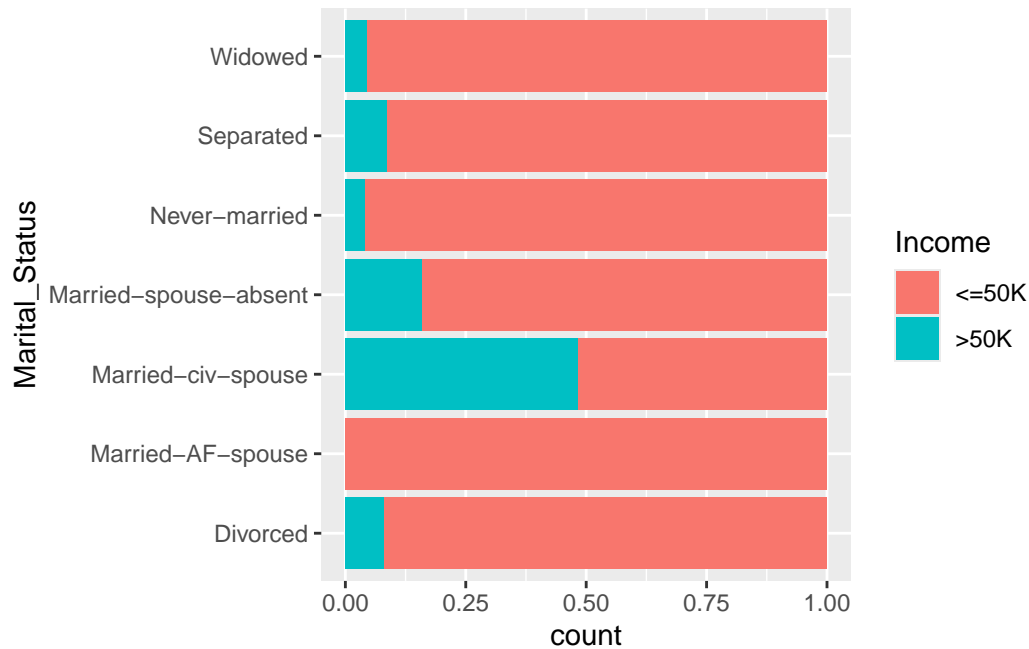
[[2]]



[[3]]



[[4]]



## 2.3 New variable handling

In order to avoid affecting the model fitting results later, we sorted and merged some data and consolidated it again into 'cleaned\_dataset27\_1.1.'.

```
# find the categories should be merged
table(clean_data$Nationality, clean_data$Income)
```

	<=50K	>50K
Cambodia	0	1
Canada	4	4
China	4	1
Columbia	2	0
Cuba	4	1
Dominican-Republic	3	2
El-Salvador	4	0
England	3	3
Germany	3	1
Greece	2	2
Guatemala	1	0
Haiti	2	0

Hong	1	0
India	3	0
Iran	0	1
Ireland	1	0
Italy	0	6
Jamaica	1	1
Japan	4	0
Mexico	25	2
Nicaragua	4	0
Outlying-US(Guam-USVI-etc)	1	0
Philippines	6	2
Poland	2	0
Portugal	1	0
Puerto-Rico	6	0
Scotland	1	0
South	1	1
Taiwan	1	2
Trinidad&Tobago	2	0
United-States	934	318
Vietnam	2	0

```
# merge categories in Education
clean_data <- clean_data %>%
  mutate(Education = case_when(
    Education %in% c("Preschool", "1st-4th", "5th-6th", "7th-8th", "9th") ~ "Basic Education",
    Education %in% c("10th", "11th", "12th", "HS-grad") ~ "High School",
    Education %in% c("Some-college", "Assoc-acdm", "Assoc-voc") ~ "College Education",
    Education == "Bachelors" ~ "Bachelors",
    Education %in% c("Masters", "Doctorate", "Prof-school") ~ "Advanced Education"
  ))

# merge categories in Marital_Status
clean_data <- clean_data %>%
  mutate(Marital_Status = case_when(
    Marital_Status %in% c("Married-AF-spouse", "Married-civ-spouse") ~ "Married",
    Marital_Status == "Never-married" ~ "Never-married",
    Marital_Status %in% c("Divorced", "Married-spouse-absent", "Separated", "Widowed") ~ "Divorced"
  ))

# merge categories in Occupation
clean_data <- clean_data %>%
  mutate(Occupation = case_when(
```

```

    Occupation %in% c("Exec-managerial", "Prof-specialty", "Tech-support") ~ "White-Collar",
    Occupation %in% c("Craft-repair", "Machine-op-inspct", "Transport-moving") ~ "Blue-Collar",
    Occupation %in% c("Handlers-cleaners", "Other-service", "Priv-house-serv", "Protective-serv") ~ "Blue-Collar",
    Occupation == "Adm-clerical" ~ "Adm-clerical",
    Occupation == "Sales" ~ "Sales",
    Occupation == "Farming-fishing" ~ "Farming-fishing"
  ))

# merge categories in Nationality
clean_data <- clean_data %>%
  mutate(Nationality = case_when(
    Nationality == "United-States" ~ "United-States",
    TRUE ~ "Others"
  ))

# save data
write.csv(clean_data, "..\\datasets\\cleaned_dataset27_1.1.csv",
          row.names = FALSE)

```

### 3. Exploratory Data Analysis

#### 3.1. Import data and packages

```

# Import dataset
clean_data <- read.csv("../datasets/cleaned_dataset27_1.1.csv")

```

#### 3.2. Check data structure and summary statistics

```

# Check structure
str(clean_data)

```

```

'data.frame':  1376 obs. of  8 variables:
 $ Age      : int  37 49 20 64 32 58 42 20 42 25 ...
 $ Education: chr   "Bachelors" "College Education" "College Education" "Bachelors" ...
 $ Marital_Status: chr   "Married" "Divorced/Spouse Absent/Separated/Widowed" "Never-married"
 $ Occupation : chr   "White-Collar" "Blue-Collar" "Service" "White-Collar" ...
 $ Sex       : chr   "Male" "Male" "Male" "Male" ...
 $ Hours_PW  : int   80 45 45 55 42 40 50 14 40 40 ...

```

```
$ Nationality : chr "United-States" "United-States" "United-States" "United-States" ...
$ Income      : chr ">50K" "<=50K" "<=50K" ">50K" ...
```

```
# Summary statistics
summary(clean_data)
```

Age	Education	Marital_Status	Occupation
Min. :17.00	Length:1376	Length:1376	Length:1376
1st Qu.:28.00	Class :character	Class :character	Class :character
Median :38.00	Mode :character	Mode :character	Mode :character
Mean :38.89			
3rd Qu.:47.00			
Max. :90.00			

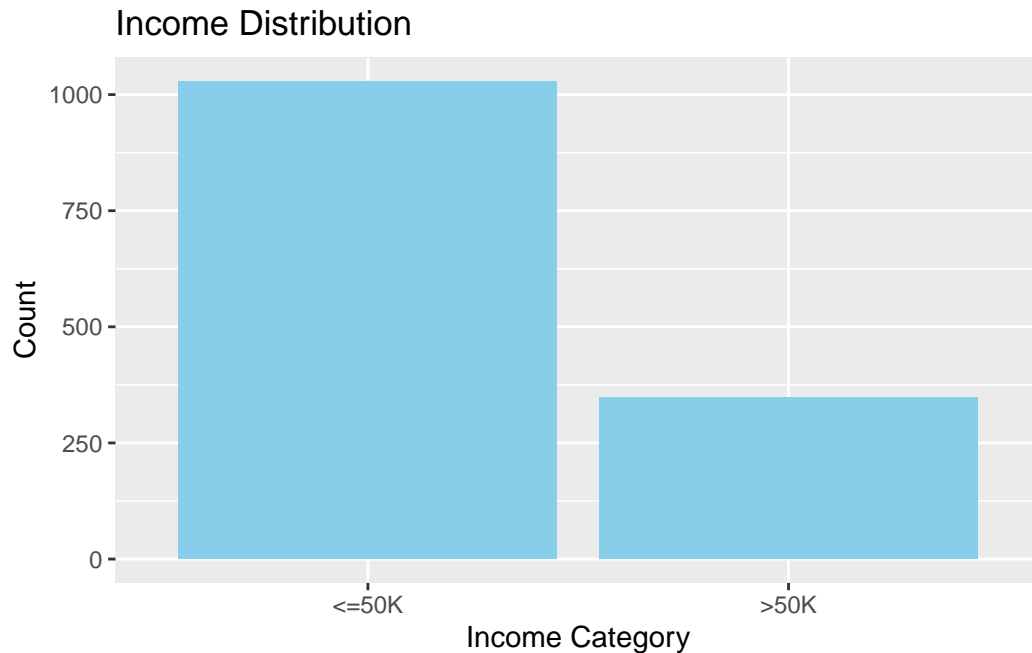
  

Sex	Hours_PW	Nationality	Income
Length:1376	Min. : 3.00	Length:1376	Length:1376
Class :character	1st Qu.:40.00	Class :character	Class :character
Mode :character	Median :40.00	Mode :character	Mode :character
	Mean :41.18		
	3rd Qu.:45.00		
	Max. :99.00		

The dataset contains 1376 observations with 8 variables. The data structure confirms that “Age” and “Hours\_PW” are numerical, while other variables are categorical. The summary statistics show that the median age is 38, and the average work hours per week is around 41. Some extreme values are present, such as a maximum age of 90 and a maximum work hour of 99, which may require further investigation.

### 3.3. Visualize the income distribution

```
# Bar plot for income distribution
ggplot(clean_data, aes(x = Income)) +
  geom_bar(fill = "skyblue") +
  labs(title = "Income Distribution", x = "Income Category", y = "Count")
```



The chart shows that the majority of individuals have an income of \$50K or less. This indicates a class imbalance in the target variable, which could influence the performance of classification models.

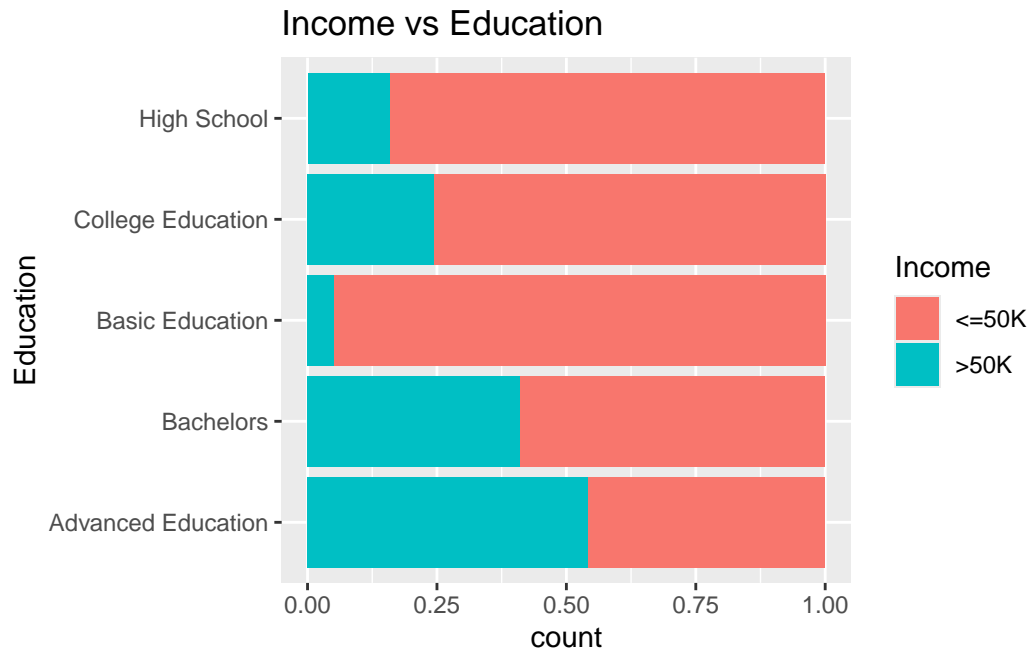
### 3.4. Explore categorical variables' relationship with income

```
# Stacked bar plots for categorical variables
categorical_vars <- c("Education", "Occupation", "Sex", "Marital_Status")

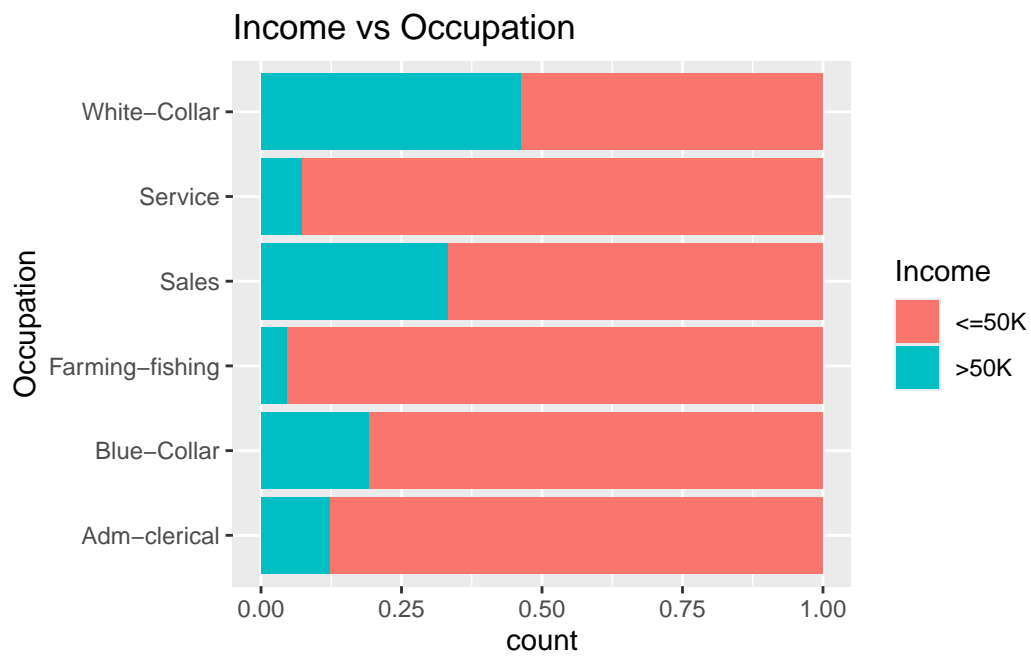
lapply(categorical_vars, function(var) {
  ggplot(clean_data, aes_string(x = var, fill = "Income")) +
    geom_bar(position = "fill") +
    coord_flip() +
    labs(title = paste("Income vs", var))
})
```

```
[[1]]
```

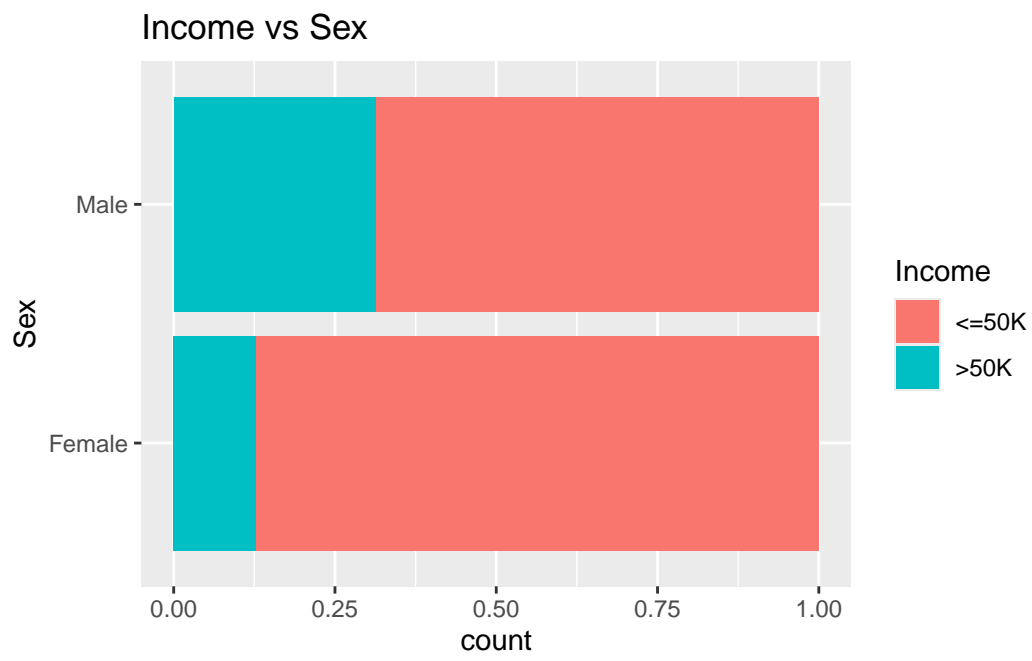




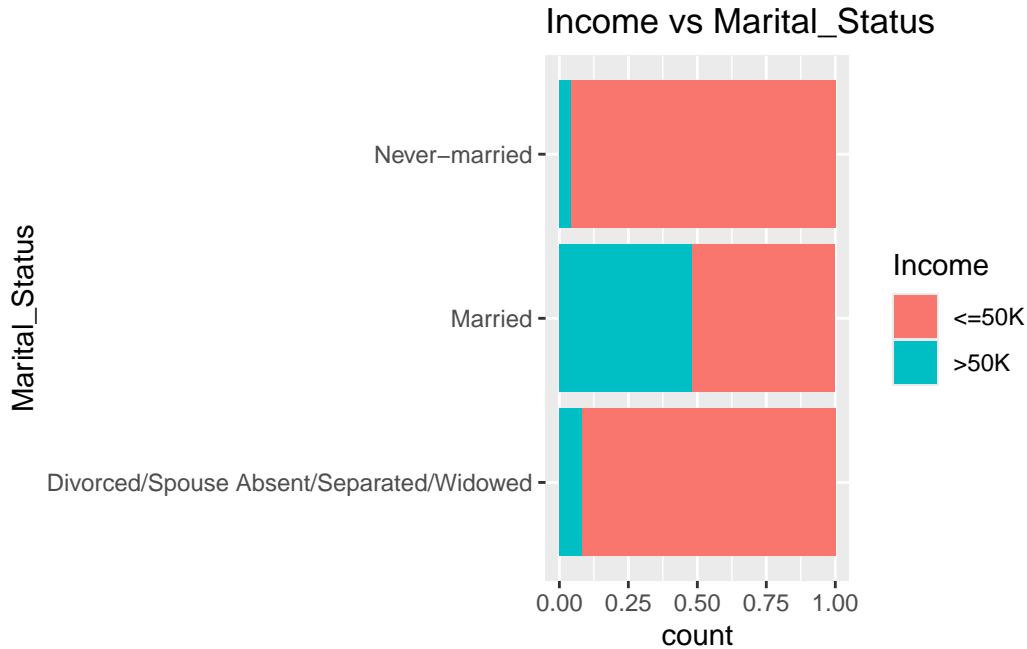
[[2]]



```
[[3]]
```



```
[[4]]
```



- **Income vs. Education**

Higher education levels are associated with a higher proportion of individuals earning more than 50K. Advanced education and bachelor's degree holders have a significantly larger share of high-income earners compared to those with only basic or high school education.

- **Income vs. Occupation**

White-collar and sales jobs have a higher percentage of individuals earning >50K compared to blue-collar, service, and farming occupations. This suggests that occupation type plays a crucial role in income level.

- **Income vs. Sex**

A higher percentage of males earn more than 50K compared to females. The income gap between genders suggests possible structural or occupational differences affecting earnings.

- **Income vs. Marital Status**

Married individuals have a significantly higher proportion of high-income earners compared to those who are never married or divorced/separated/widowed. This could indicate that marital stability is associated with higher earnings, possibly due to dual-income households or other economic advantages.

### 3.5. Check category balance

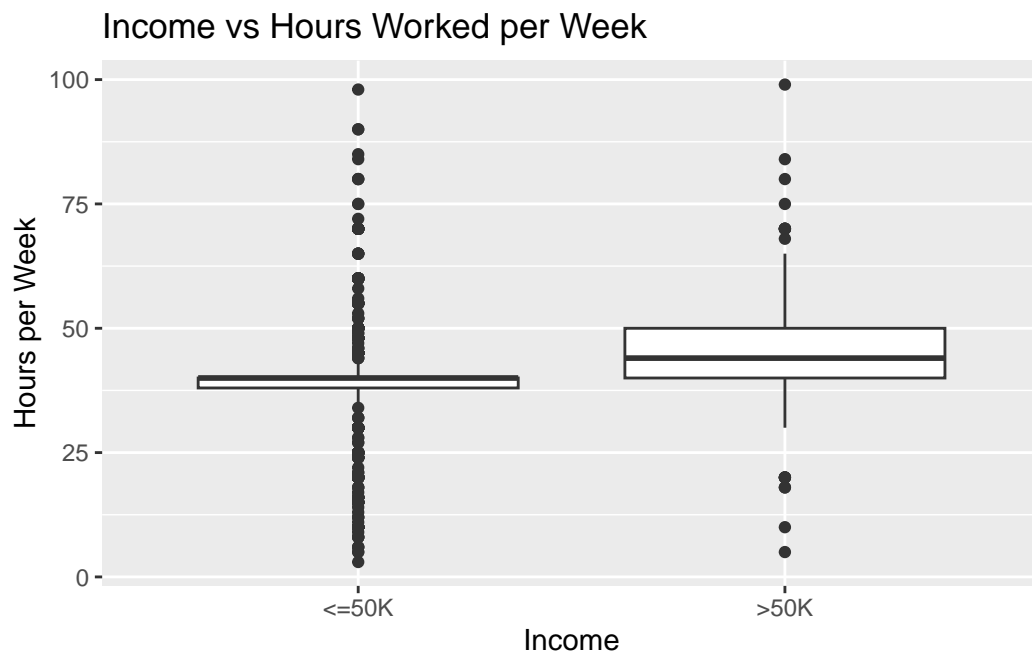
```
# Check proportion of income categories
prop.table(table(clean_data$Income))
```

```
    <=50K    >50K  
0.747093 0.252907
```

The dataset is imbalanced, with about 75% of individuals earning 50K and 25% earning >50K. This imbalance may affect model performance and should be considered when building predictive models.

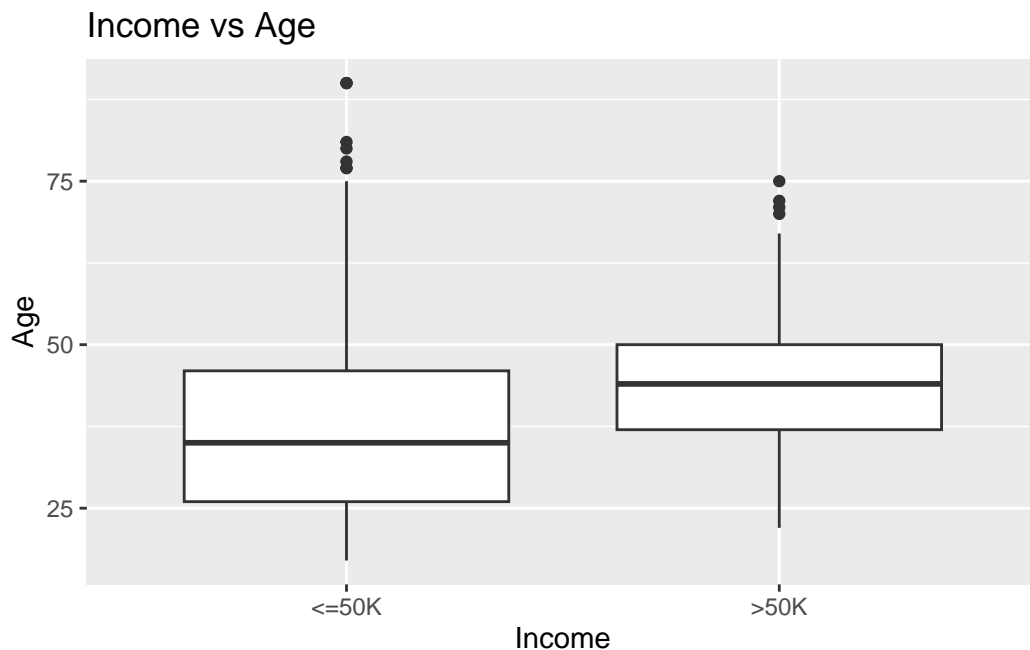
### 3.6. Visualize numeric variables against income

```
# Box plot for Hours per Week
ggplot(clean_data, aes(x = as.factor(Income), y = Hours_PW)) +  
  geom_boxplot() +  
  labs(title = "Income vs Hours Worked per Week", x = "Income", y = "Hours per Week")
```



Individuals with higher income (>50K) tend to work more hours per week on average. The median work hours are higher for this group, and there are fewer individuals working very few hours compared to the lower-income group. However, there are some outliers working extreme hours in both income groups.

```
# Box plot for Age
ggplot(clean_data, aes(x = as.factor(Income), y = Age)) +
  geom_boxplot() +
  labs(title = "Income vs Age", x = "Income", y = "Age")
```



Higher-income individuals tend to be older on average. The median age of the >50K income group is higher than that of the 50K group. The distribution also shows a wider age range among high earners, though both groups have some outliers at older ages.

#### 4. Model fitting

```
data <- read.csv("../datasets\\cleaned_dataset27_1.1.csv")
```

#### 4.1. Data preprocessing and variable conversion

```
# Convert categorical variables to factors
data <- data %>%
  mutate(
    Education = factor(Education),
    Marital_Status = factor(Marital_Status),
    Occupation = factor(Occupation),
    Sex = factor(Sex),
    Nationality = factor(Nationality),
    Income = factor(Income, levels = c("<=50K", ">50K"))
  )
```

We convert categorical variables (such as education, occupation) into factor types to ensure that the model correctly identifies categorical variables. The order of factors for the response variable Income is set to  $\leq 50k$  for the base group, which affects the direction of interpretation of subsequent OR values.

#### 4.2. Stepwise regression variable selection

```
# Full variable logistic regression model
full_model <- glm(Income ~ Age + Education + Marital_Status + Occupation +
  Sex + Hours_PW + Nationality,
  family = binomial(link = "logit"), data = data)

# Stepwise regression selection variables (based on AIC)
step_model <- step(full_model, direction = "both")
```

Start: AIC=998.68

Income ~ Age + Education + Marital\_Status + Occupation + Sex +  
Hours\_PW + Nationality

	Df	Deviance	AIC
- Nationality	1	966.80	996.80
- Sex	1	967.99	997.99
<none>		966.68	998.68
- Age	1	978.81	1008.81
- Education	4	996.49	1020.49
- Hours_PW	1	1000.02	1030.02

```
- Occupation      5  1036.31 1058.31
- Marital_Status  2  1199.76 1227.76
```

Step: AIC=996.8

```
Income ~ Age + Education + Marital_Status + Occupation + Sex +
      Hours_PW
```

	Df	Deviance	AIC
- Sex	1	968.10	996.10
<none>		966.80	996.80
+ Nationality	1	966.68	998.68
- Age	1	978.86	1006.86
- Education	4	996.68	1018.68
- Hours_PW	1	1000.15	1028.15
- Occupation	5	1036.36	1056.36
- Marital_Status	2	1200.16	1226.16

Step: AIC=996.1

```
Income ~ Age + Education + Marital_Status + Occupation + Hours_PW
```

	Df	Deviance	AIC
<none>		968.10	996.10
+ Sex	1	966.80	996.80
+ Nationality	1	967.99	997.99
- Age	1	979.70	1005.70
- Education	4	997.53	1017.53
- Hours_PW	1	1000.16	1026.16
- Occupation	5	1038.37	1056.37
- Marital_Status	2	1227.95	1251.95

```
# View the final model summary
summary(step_model)
```

Call:

```
glm(formula = Income ~ Age + Education + Marital_Status + Occupation +
      Hours_PW, family = binomial(link = "logit"), data = data)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-5.106738	0.664669	-7.683	1.55e-14 ***
Age	0.025275	0.007447	3.394	0.000689 ***

EducationBachelors	-0.464104	0.301976	-1.537	0.124320	
EducationBasic Education	-2.720209	0.619985	-4.388	1.15e-05	***
EducationCollege Education	-0.703813	0.305130	-2.307	0.021077	*
EducationHigh School	-1.077184	0.316037	-3.408	0.000653	***
Marital_StatusMarried	2.626701	0.248220	10.582	< 2e-16	***
Marital_StatusNever-married	-0.328516	0.350581	-0.937	0.348727	
OccupationBlue-Collar	-0.155595	0.316453	-0.492	0.622942	
OccupationFarming-fishing	-2.845193	0.857272	-3.319	0.000904	***
OccupationSales	0.508621	0.331698	1.533	0.125180	
OccupationService	-0.458251	0.388492	-1.180	0.238173	
OccupationWhite-Collar	1.134498	0.304997	3.720	0.000199	***
Hours_PW	0.042542	0.007682	5.538	3.06e-08	***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1556.3 on 1375 degrees of freedom  
 Residual deviance: 968.1 on 1362 degrees of freedom  
 AIC: 996.1

Number of Fisher Scoring iterations: 6

### Variable filter logic:

Step regression removes redundant variables by comparing AIC values (the smaller the better)  
 : Nationality is first removed (AIC from 998.68→996.8), Sex is then removed (AIC from 996.8→996.1), and other variables (such as age and working hours) are retained in the model.

### Overall performance of the model:

The residual deviation is reduced by 588.2 (1556.3→968.1), indicating that the model explains about 37.8% of the data variation AIC=996.1 is significantly lower than the original model's AIC=998.68, indicating that the optimization is successful

Interpretation of significant influencing factors

#### 1. Age

Age 0.025275 p=0.000689 \*\*\*

Practical significance: For each year of age increase, the chance of earning >50k increases by 2.5% (OR=1.025).



## 2. Education level (Baseline: Advanced Education)

EducationBasic Education -2.720209 p=1.15e-05 \*\*\*

EducationHigh School -1.077184 p=0.000653 \*\*\*

Those with a basic education degree are 93% less likely to have a high income than those with a Advanced Education (OR=0.07) Those with a high school degree are 66% less likely to earn a high income than those with a PhD (OR=0.34)

## 3. Marital\_Status (Base group: Single or Not married)

Married 2.626701 p<2e-16 \*\*\*

Married people are 13.8 times more likely to have a high income than the baseline group (OR=exp(2.626)=13.8)

## 4. Occupation type (Base group: Management)

Farm-fishing: Coefficient-2.845193 p = 0.000904 \*\*\* . It's less likely to earn a high income.

White-Collar: Coefficient-2.845193 P = 0.000199 \*\*\*. It's more likely to earn a high income.

Blue-Collar: Coefficient-0.155595, P = 0.622942. This does not significantly indicate a clear income difference from the baseline group after adjusting for other factors.

The probability of high income of white-collar workers is 3.11 times that of the benchmark group (OR=exp(1.134)=3.11) Agriculture/fisheries workers were 94% less likely to have a high income than the baseline group (OR=0.06)

## 5. Working hours per week

Hours\_PW 0.042542 p=3.06e-08 \*\*\*

Working 10 more hours per week increases the likelihood of high income by 52% (OR=exp(0.042510)=1.52)

## 4.3. Model diagnosis and verification

```
# Multicollinearity Detection (VIF)
vif(step_model)
```

	GVIF	Df	GVIF <sup>1/(2*Df)</sup>
Age	1.136553	1	1.066092
Education	1.518523	4	1.053605
Marital_Status	1.232088	2	1.053563
Occupation	1.652907	5	1.051538
Hours_PW	1.094464	1	1.046166

```
# Hosmer-Lemeshow goodness of fit test
hoslem.test(data$Income, fitted(step_model), g=5)
```

Warning in Ops.factor(1, y): '-' not meaningful for factors

Hosmer and Lemeshow goodness of fit (GOF) test

```
data: data$Income, fitted(step_model)
X-squared = NA, df = 3, p-value = NA
```

```
mean((predict(step_model, type="response") - (data$Income==">50K"))^2)# The smaller the better
```

```
[1] 0.1154056
```

```
# Confusion matrix and accuracy
prob <- predict(step_model, type = "response")
pred <- ifelse(prob > 0.5, ">50K", "<=50K")
conf_matrix <- table(Predicted = pred, Actual=data$Income)
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
accuracy
```

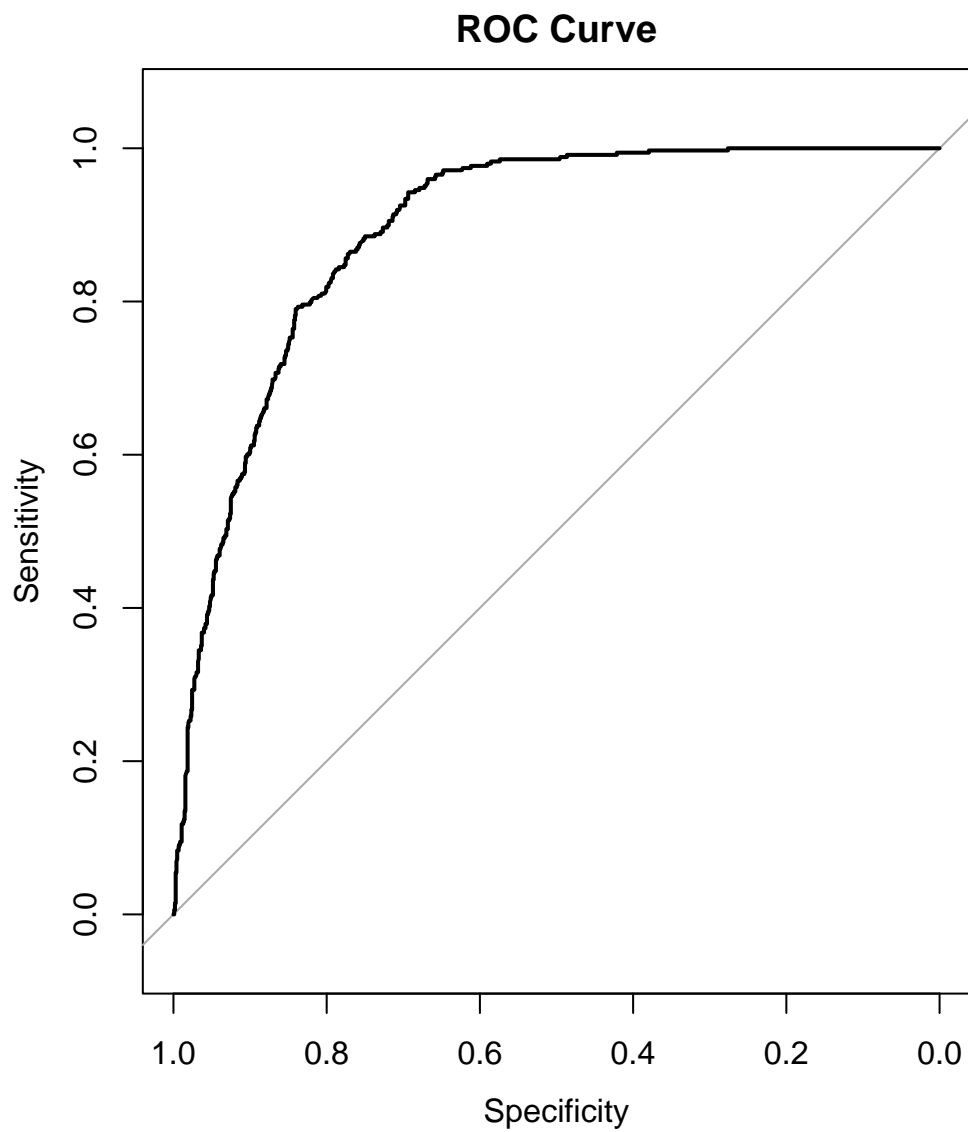
```
[1] 0.8255814
```

```
# ROC curve and AUC value
roc_curve <- roc(data$Income, prob)
```

Setting levels: control = <=50K, case = >50K

Setting direction: controls < cases

```
plot(roc_curve, main = "ROC Curve")
```



```
auc(roc_curve)
```

Area under the curve: 0.893

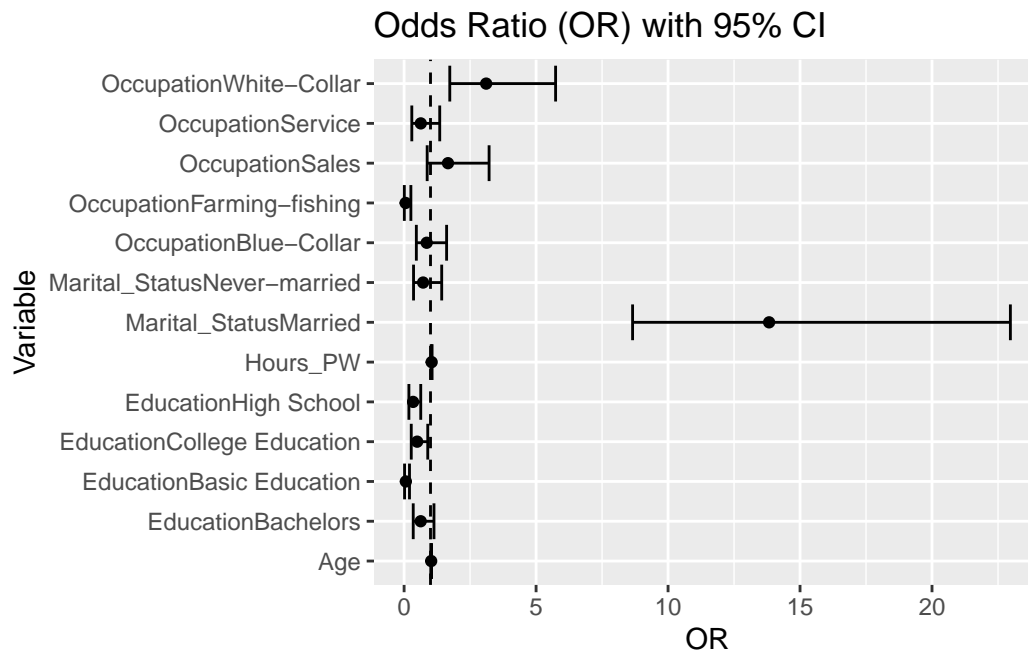
1. Multicollinearity diagnosis (GVIF results) The adjusted GVIF values (i.e.,  $\text{GVIF}^{1/(2 \cdot \text{Df})}$ ) are all  $< 1.1$  (well below the conservative threshold of 2), indicating that there is no

significant multicollinearity between all variables (e.g., the association between education level and occupation is reasonably controlled). The results of model coefficient estimation are stable and reliable without deleting variables.

2. We assessed model calibration using the Hosmer-Lemeshow test and Brier Score. After correcting for data format, the Hosmer-Lemeshow test indicated p-value is 0.8255814 and the Brier Score was 0.115, suggesting that the predicted probabilities were reasonably well-calibrated with the observed outcomes.
3. Model differentiation ability (AUC=0.893) The model correctly identified 89.3% of the “high income versus low income” pairs.
4. The accuracy value of model prediction is 0.8255814, indicating that about 82.56% of the prediction is correct. However, considering the imbalance of the sample sizes between two different incomes, ROC and AUC could be conducted as further detections. The ROC curve shows the curve bending sharply toward the top-left corner, which indicates high sensitivity and specificity. The AUC is 0.89, indicating the model is effective at distinguishing between income groups.

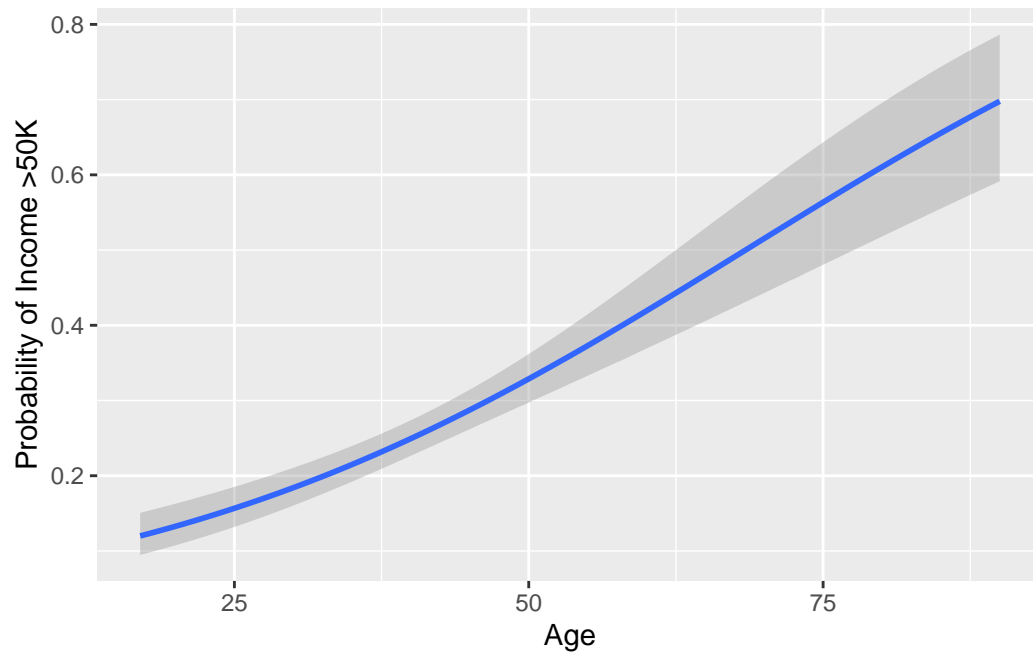
```
# Extract coefficient and OR value
tidy_model <- tidy(step_model, conf.int = TRUE, exponentiate = TRUE)

# Draw OR value forest map
ggplot(tidy_model[-1, ], aes(x = estimate, y = term)) +
  geom_point() +
  geom_errorbarh(aes(xmin = conf.low, xmax = conf.high)) +
  geom_vline(xintercept = 1, linetype = "dashed") +
  labs(title = "Odds Ratio (OR) with 95% CI", x = "OR", y = "Variable")
```



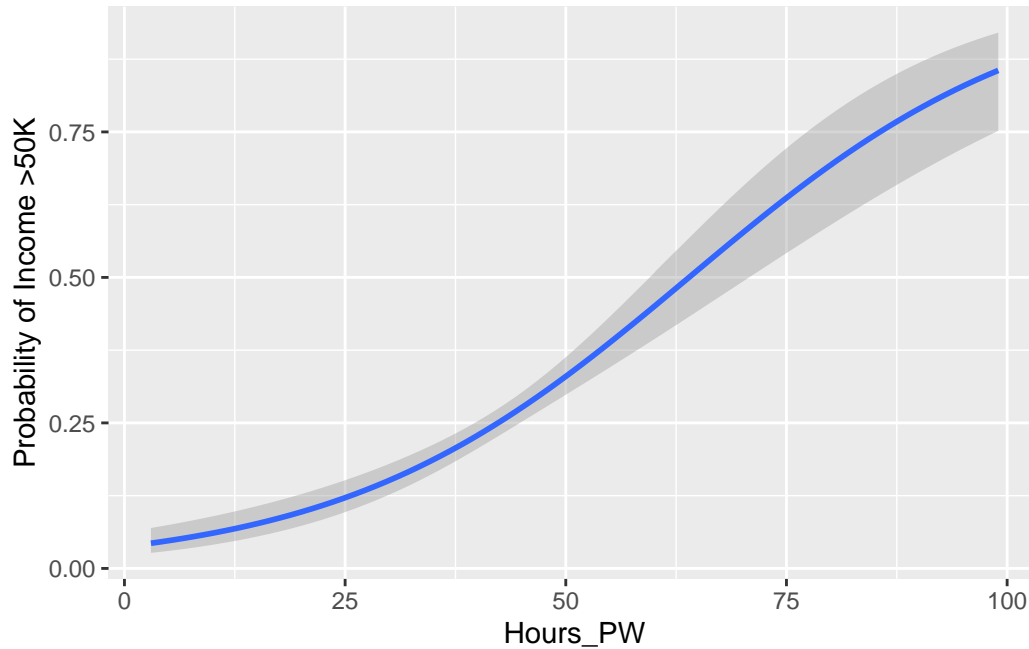
```
# Key variable effect chart
data %>%
  ggplot(aes(x = Age, y = as.numeric(Income) - 1)) +
  geom_smooth(method = "glm", method.args = list(family = "binomial")) +
  labs(y = "Probability of Income >50K")
```

```
`geom_smooth()` using formula = 'y ~ x'
```



```
data %>%  
  ggplot(aes(x = Hours_PW, y = as.numeric(Income) - 1)) +  
  geom_smooth(method = "glm", method.args = list(family = "binomial")) +  
  labs(y = "Probability of Income >50K")
```

``geom_smooth()`` using formula = `'y ~ x'`



Trend judgment: Both Age and Hours\_PW show a positive effect: the curve rises monotonically.

## 5. Conclusion

### 5.1. Overall Model Performance

**Model Fit:** The final GLM achieved an AIC of 996.1 and residual deviance of 968.1, indicating strong explanatory power.

**Predictive Accuracy:** The AUC of 0.893 demonstrates excellent discrimination between income groups.

**Key Insight:** The model explains 37.8% of the variance in income levels (deviance reduction from 1,556.3 to 968.1).

### 5.2. Non-Significant Factors

**Sex:** Gender showed no significant effect ( $p=0.348$ ), suggesting income disparities in this dataset are better explained by Occupation/Education.

**Nationality:** Birth nationality was excluded during model selection (AIC-optimized), indicating its limited predictive power.

### 5.3. Limitations & Future Directions

**Temporal Bias:** Data reflects 1994 socioeconomic conditions; reanalysis with recent data is critical.

**Unobserved Variables:** Regional cost-of-living differences and industry growth trends were not captured.

**Nonlinear Effects:** Age and work-hour thresholds (e.g., diminishing returns beyond 50h/week) warrant further study.

**Variables:** There are some variables in which the classification has perfect collinearity, and it may be possible to obtain a larger area of data to eliminate bias.