Income Classification Using Logistic Regression: Data Preparation

Donnie Minnick

2025-06-25

# Objective

This project analyzes income prediction using logistic regression on the UCI Adult Income dataset. It explores the impact of class imbalance, feature selection, and decision threshold tuning to build interpretable and balanced classification models.

# Source Dataset

The Adult Income dataset originates from the 1994 U.S. Census database. It was created to predict whether an individual earns more than $50,000 per year based on demographic and employment-related attributes.

This is a binary classification dataset commonly used for logistic regression, decision trees, and fairness analysis.

It was created by Ronny Kohavi and Barry Becker, from the Census Bureau’s 1994 Current Population Survey (CPS).

It is hosted by the UCI Machine Learning Repository [here](https://archive.ics.uci.edu/ml/datasets/adult).

Load libraries.

library("dplyr")  
library("ggplot2")  
library("knitr")

# Load Raw Data

adult\_income\_clean <- readRDS("../data/raw/adult\_income\_raw.rds")  
  
str(adult\_income\_clean)

## 'data.frame': 32560 obs. of 15 variables:  
## $ age : int 50 38 53 28 37 49 52 31 42 37 ...  
## $ workclass : chr " Self-emp-not-inc" " Private" " Private" " Private" ...  
## $ fnlwgt : int 83311 215646 234721 338409 284582 160187 209642 45781 159449 280464 ...  
## $ education : chr " Bachelors" " HS-grad" " 11th" " Bachelors" ...  
## $ education\_num : int 13 9 7 13 14 5 9 14 13 10 ...  
## $ marital\_status: chr " Married-civ-spouse" " Divorced" " Married-civ-spouse" " Married-civ-spouse" ...  
## $ occupation : chr " Exec-managerial" " Handlers-cleaners" " Handlers-cleaners" " Prof-specialty" ...  
## $ relationship : chr " Husband" " Not-in-family" " Husband" " Wife" ...  
## $ race : chr " White" " White" " Black" " Black" ...  
## $ sex : chr " Male" " Male" " Male" " Female" ...  
## $ capital\_gain : int 0 0 0 0 0 0 0 14084 5178 0 ...  
## $ capital\_loss : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ hours\_per\_week: int 13 40 40 40 40 16 45 50 40 80 ...  
## $ native\_country: chr " United-States" " United-States" " United-States" " Cuba" ...  
## $ income : chr " <=50K" " <=50K" " <=50K" " <=50K" ...

Data frame contains 32,560 observations and 15 variables.

# Variable Definitions

| Variable | Type | Description |
| --- | --- | --- |
| age | Numeric | Age of the individual in years. |
| workclass | Categorical | Type of employer or employment sector (e.g., Private, Self-emp, Gov). |
| fnlwgt | Numeric | Final person weight: how many people the respondent represents in the survey. |
| education | Categorical | Highest level of education completed (e.g., Bachelors, HS-grad). |
| education\_num | Numeric | Numeric representation of education (ordinal scale from 1–16). |
| marital\_status | Categorical | Marital status (e.g., Married-civ-spouse, Divorced, Never-married). |
| occupation | Categorical | Kind of work performed by the individual (e.g., Exec-managerial, Sales). |
| relationship | Categorical | Person’s household role (e.g., Husband, Wife, Own-child). |
| race | Categorical | Race classification per Census (e.g., White, Black, Asian-Pac-Islander). |
| sex | Categorical | Biological sex of the respondent (Male, Female). |
| capital\_gain | Numeric | Profit from sale of capital assets (e.g., stocks, real estate). |
| capital\_loss | Numeric | Loss from sale of capital assets. |
| hours\_per\_week | Numeric | Number of hours the person works per week. |
| native\_country | Categorical | Country of birth (respondent is currently living in the U.S.). |
| income | Categorical | Target variable: whether income is <=50K or >50K per year. |

# Handle Missing Values

Trim extra white space from values and omit records that contain missing values.

adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(across(where(is.character), ~ stringr::str\_trim(.))) %>%  
 tidyr::drop\_na()

Record count reduced to 30,161 observations.

# Review and Encode Variables

## Age

Age of individual in years.

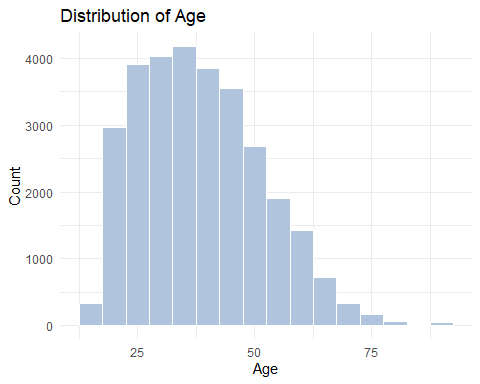
entries\_by\_age <- adult\_income\_clean %>%  
 group\_by(age) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_age,  
 col.names = c("Age", "Entries", "Percent"),  
 caption = "Entries By Age",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Age

| Age | Entries | Percent |
| --- | --- | --- |
| 17 | 328 | 0.01 |
| 18 | 447 | 0.01 |
| 19 | 594 | 0.02 |
| 20 | 629 | 0.02 |
| 21 | 621 | 0.02 |
| 22 | 674 | 0.02 |
| 23 | 824 | 0.03 |
| 24 | 752 | 0.02 |
| 25 | 799 | 0.03 |
| 26 | 745 | 0.02 |
| 27 | 789 | 0.03 |
| 28 | 808 | 0.03 |
| 29 | 774 | 0.03 |
| 30 | 813 | 0.03 |
| 31 | 851 | 0.03 |
| 32 | 789 | 0.03 |
| 33 | 837 | 0.03 |
| 34 | 836 | 0.03 |
| 35 | 828 | 0.03 |
| 36 | 852 | 0.03 |
| 37 | 828 | 0.03 |
| 38 | 791 | 0.03 |
| 39 | 785 | 0.03 |
| 40 | 765 | 0.03 |
| 41 | 769 | 0.03 |
| 42 | 741 | 0.02 |
| 43 | 743 | 0.02 |
| 44 | 704 | 0.02 |
| 45 | 706 | 0.02 |
| 46 | 711 | 0.02 |
| 47 | 683 | 0.02 |
| 48 | 523 | 0.02 |
| 49 | 555 | 0.02 |
| 50 | 575 | 0.02 |
| 51 | 571 | 0.02 |
| 52 | 455 | 0.02 |
| 53 | 448 | 0.01 |
| 54 | 394 | 0.01 |
| 55 | 386 | 0.01 |
| 56 | 343 | 0.01 |
| 57 | 337 | 0.01 |
| 58 | 344 | 0.01 |
| 59 | 332 | 0.01 |
| 60 | 276 | 0.01 |
| 61 | 259 | 0.01 |
| 62 | 213 | 0.01 |
| 63 | 186 | 0.01 |
| 64 | 173 | 0.01 |
| 65 | 136 | 0.00 |
| 66 | 110 | 0.00 |
| 67 | 111 | 0.00 |
| 68 | 90 | 0.00 |
| 69 | 80 | 0.00 |
| 70 | 64 | 0.00 |
| 71 | 54 | 0.00 |
| 72 | 40 | 0.00 |
| 73 | 49 | 0.00 |
| 74 | 38 | 0.00 |
| 75 | 34 | 0.00 |
| 76 | 29 | 0.00 |
| 77 | 20 | 0.00 |
| 78 | 14 | 0.00 |
| 79 | 15 | 0.00 |
| 80 | 16 | 0.00 |
| 81 | 13 | 0.00 |
| 82 | 7 | 0.00 |
| 83 | 5 | 0.00 |
| 84 | 8 | 0.00 |
| 85 | 3 | 0.00 |
| 86 | 1 | 0.00 |
| 88 | 3 | 0.00 |
| 90 | 35 | 0.00 |

Ages range from 17 to 90. Plot age.

ggplot(adult\_income\_clean, aes(x = age)) +  
 geom\_histogram(binwidth = 5, fill = "lightsteelblue", color = "white") +  
 labs(title = "Distribution of Age", x = "Age", y = "Count") +  
 theme\_minimal()



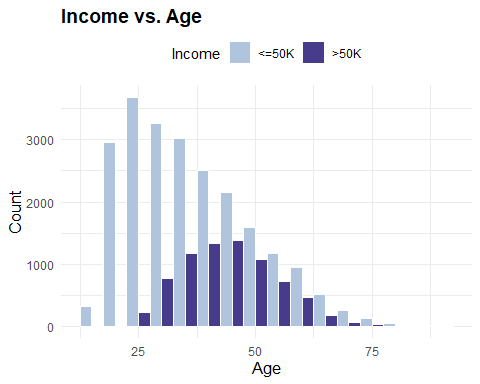
Most observations fall in the 25-50 age range, and there is a long tail extending into older ages, especially past age 60, but these are rare. There a very few individuals over age 70, and the count drops steeply after age 55.

The peak aligns with prime working years. The sparse representation for very young and very old individuals is likely due to school enrollment and retirement.

What is the trend between age and income?

Plot age versus income.

ggplot(adult\_income\_clean, aes(x = age, fill = income)) +  
 geom\_histogram(binwidth = 5, position = "dodge", color = "white") +  
 scale\_fill\_manual(values = c("<=50K" = "lightsteelblue", ">50K" = "slateblue4")) +  
 labs(title = "Income vs. Age", x = "Age", y = "Count", fill = "Income") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(size = 14, face = "bold"),  
 axis.title = element\_text(size = 12),  
 legend.title = element\_text(size = 11),  
 legend.position = "top")



There are diverging age-income distributions here.

The <=50k group is dominant in younger ages (20-40) and steadily declines with age. The reflects early-career positions, part-time workers and students, and perhaps structural wage inequality by sex, race, or education.

The >50k group is barely present until around ages 28-30, grows and peaks between ages 35-50, then declines. This reflects mid-career professionals, and the effects of education, experience and seniority. The income plateau for this group is near peak career age.

Both groups decline in a manner that mirrors general workforce retirement patterns.

Age does not appear to be linearly related to income. There’s a threshold effect, i.e. higher income doesn’t gradually rise from age 18 onward. It ramps sharply post-age 30.

And income inequality is dynamic by age, i.e. the >50k group becomes more dominant only after career experience accumulates.

## Workclass

Represents the type of employer or employment sector.

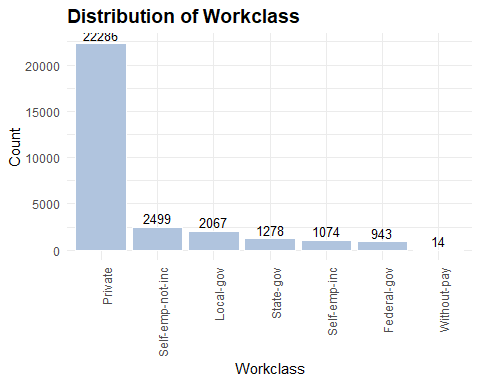
entries\_by\_workclass <- adult\_income\_clean %>%  
 group\_by(workclass) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_workclass,  
 col.names = c("Workclass", "Entries", "Percent"),  
 caption = "Entries By Workclass",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Workclass

| Workclass | Entries | Percent |
| --- | --- | --- |
| Federal-gov | 943 | 0.03 |
| Local-gov | 2,067 | 0.07 |
| Private | 22,286 | 0.74 |
| Self-emp-inc | 1,074 | 0.04 |
| Self-emp-not-inc | 2,499 | 0.08 |
| State-gov | 1,278 | 0.04 |
| Without-pay | 14 | 0.00 |

Plot workclass.

ggplot(entries\_by\_workclass, aes(x = forcats::fct\_reorder(workclass, entries, .desc = TRUE), y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Workclass",  
 x = "Workclass",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



Consolidate from 7 groups to 4 to reduce model complexity and align with more meaningful labor categories.

* **Public**: Federal, state or local goverment,
* **Private**: Most traditional employers,
* **Self-Employed**: Entrepreneurs and contractors, and
* **Not Employed**: No income.

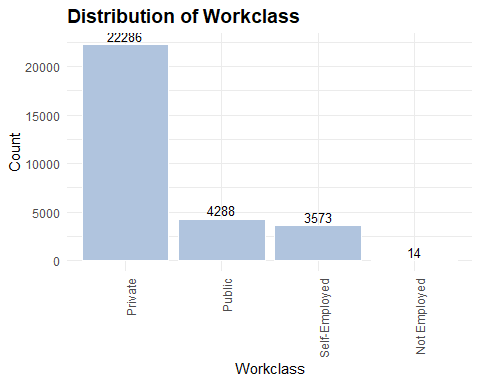
adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(workclass\_group = case\_when(  
 workclass %in% c("Federal-gov", "Local-gov", "State-gov") ~ "Public",  
 workclass == "Private" ~ "Private",  
 workclass %in% c("Self-emp-inc", "Self-emp-not-inc") ~ "Self-Employed",  
 workclass == "Without-pay" ~ "Not Employed",  
 TRUE ~ NA\_character\_))  
  
entries\_by\_workclass\_group <- adult\_income\_clean %>%  
 group\_by(workclass\_group) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_workclass\_group,  
 col.names = c("Workclass Group", "Entries", "Percent"),  
 caption = "Entries By Workclass Group",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Workclass Group

| Workclass Group | Entries | Percent |
| --- | --- | --- |
| Not Employed | 14 | 0.00 |
| Private | 22,286 | 0.74 |
| Public | 4,288 | 0.14 |
| Self-Employed | 3,573 | 0.12 |

Plot workclass groups.

ggplot(entries\_by\_workclass\_group, aes(x = forcats::fct\_reorder(workclass\_group, entries, .desc = TRUE), y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Workclass",  
 x = "Workclass",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



## Final Weight

Final weight is a person-level sampling weight provided by the US Census Bureau to indicate how many people in the overall population the sampled individual represents.

It exists because the Census Bureau uses stratified sampling and not every individual had an equal chance of being selected. So each respondent was assigned a weight to adjust for over- or under-sampling in demographic groups and populatiuon representation by region, age, race, gender, etc.

If someone has a fnlwgt of 148780, it means that person represents 148,780 people in the full population.

I don’t need to adjust for survey design, so I will exclude the variable.

adult\_income\_clean <- adult\_income\_clean %>%  
 select(-fnlwgt)

## Education

Highest level of education completed (e.g., Bachelors, HS-grad).

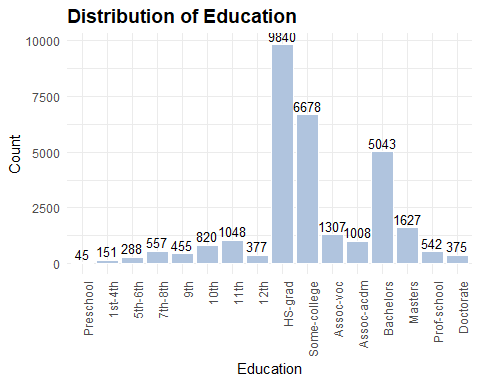
entries\_by\_education <- adult\_income\_clean %>%  
 group\_by(education) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2)) %>%  
 mutate(education = factor(education,  
 levels = c("Preschool",  
 "1st-4th",   
 "5th-6th",   
 "7th-8th",   
 "9th",   
 "10th",   
 "11th",  
 "12th",  
 "HS-grad",  
 "Some-college",  
 "Assoc-voc",  
 "Assoc-acdm",  
 "Bachelors",  
 "Masters",  
 "Prof-school",  
 "Doctorate")))  
  
kable(entries\_by\_education,  
 col.names = c("Education", "Entries", "Percent"),  
 caption = "Entries By Education",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Education

| Education | Entries | Percent |
| --- | --- | --- |
| 10th | 820 | 0.03 |
| 11th | 1,048 | 0.03 |
| 12th | 377 | 0.01 |
| 1st-4th | 151 | 0.01 |
| 5th-6th | 288 | 0.01 |
| 7th-8th | 557 | 0.02 |
| 9th | 455 | 0.02 |
| Assoc-acdm | 1,008 | 0.03 |
| Assoc-voc | 1,307 | 0.04 |
| Bachelors | 5,043 | 0.17 |
| Doctorate | 375 | 0.01 |
| HS-grad | 9,840 | 0.33 |
| Masters | 1,627 | 0.05 |
| Preschool | 45 | 0.00 |
| Prof-school | 542 | 0.02 |
| Some-college | 6,678 | 0.22 |

Plot education.

ggplot(entries\_by\_education, aes(x = education, y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Education",  
 x = "Education",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



Consolidate the 16 groups into a smaller set of 7 more meaningful categories.

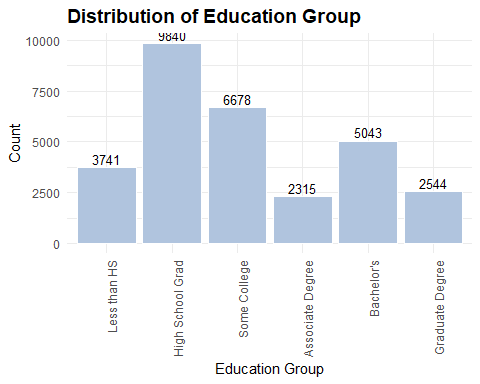
adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(education\_group = case\_when(  
 education %in% c("Preschool", "1st-4th", "5th-6th", "7th-8th",  
 "9th", "10th", "11th", "12th") ~ "Less than HS",  
 education == "HS-grad" ~ "High School Grad",  
 education == "Some-college" ~ "Some College",  
 education %in% c("Assoc-acdm", "Assoc-voc") ~ "Associate Degree",  
 education == "Bachelors" ~ "Bachelor's",  
 education %in% c("Masters", "Doctorate", "Prof-school") ~ "Graduate Degree",  
 TRUE ~ NA\_character\_))  
  
entries\_by\_education\_group <- adult\_income\_clean %>%  
 group\_by(education\_group) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2)) %>%  
 mutate(education\_group = factor(education\_group,  
 levels = c("Less than HS",   
 "High School Grad",   
 "Some College",   
 "Associate Degree",   
 "Bachelor's",   
 "Graduate Degree")))  
  
kable(entries\_by\_education\_group,  
 col.names = c("Education Group", "Entries", "Percent"),  
 caption = "Entries By Education Group",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Education Group

| Education Group | Entries | Percent |
| --- | --- | --- |
| Associate Degree | 2,315 | 0.08 |
| Bachelor’s | 5,043 | 0.17 |
| Graduate Degree | 2,544 | 0.08 |
| High School Grad | 9,840 | 0.33 |
| Less than HS | 3,741 | 0.12 |
| Some College | 6,678 | 0.22 |

Plot education group.

ggplot(entries\_by\_education\_group, aes(x = education\_group, y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Education Group",  
 x = "Education Group",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



Create a binary variable that indicates whether an individual has a college degree or not.

adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(has\_college\_degree = ifelse(education\_group %in% c("Bachelor's", "Graduate Degree"), 1, 0))

## Education Number

Numeric representation of education (ordinal scale from 1–16).

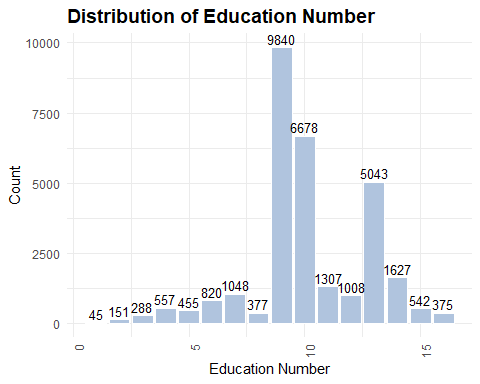
entries\_by\_education\_num <- adult\_income\_clean %>%  
 group\_by(education\_num) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_education\_num,  
 col.names = c("Education Number", "Entries", "Percent"),  
 caption = "Entries By Education Number",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Education Number

| Education Number | Entries | Percent |
| --- | --- | --- |
| 1 | 45 | 0.00 |
| 2 | 151 | 0.01 |
| 3 | 288 | 0.01 |
| 4 | 557 | 0.02 |
| 5 | 455 | 0.02 |
| 6 | 820 | 0.03 |
| 7 | 1,048 | 0.03 |
| 8 | 377 | 0.01 |
| 9 | 9,840 | 0.33 |
| 10 | 6,678 | 0.22 |
| 11 | 1,307 | 0.04 |
| 12 | 1,008 | 0.03 |
| 13 | 5,043 | 0.17 |
| 14 | 1,627 | 0.05 |
| 15 | 542 | 0.02 |
| 16 | 375 | 0.01 |

Plot education number.

ggplot(entries\_by\_education\_num, aes(x = education\_num, y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Education Number",  
 x = "Education Number",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



Since I’m using education group as a factor, and education number is simply numeric proxy for the same information, including both in the model will lead to multicollinearity or confounding.

Drop the education number column and use the education group column for clear odds comparison.

adult\_income\_clean <- adult\_income\_clean %>%  
 select(-education\_num)

## Marital Status

Indicates individual’s marital status.

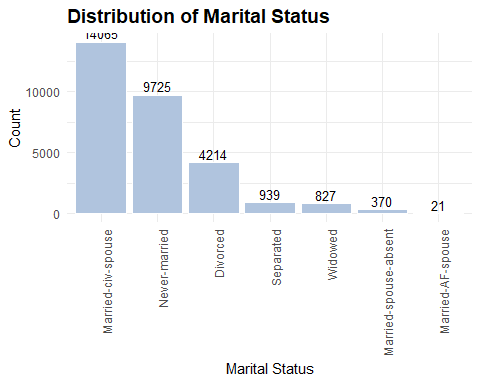
entries\_by\_marital\_status <- adult\_income\_clean %>%  
 group\_by(marital\_status) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_marital\_status,  
 col.names = c("Marital Status", "Entries", "Percent"),  
 caption = "Entries By Marital Status",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Marital Status

| Marital Status | Entries | Percent |
| --- | --- | --- |
| Divorced | 4,214 | 0.14 |
| Married-AF-spouse | 21 | 0.00 |
| Married-civ-spouse | 14,065 | 0.47 |
| Married-spouse-absent | 370 | 0.01 |
| Never-married | 9,725 | 0.32 |
| Separated | 939 | 0.03 |
| Widowed | 827 | 0.03 |

Plot marital status.

ggplot(entries\_by\_marital\_status, aes(x = forcats::fct\_reorder(marital\_status, entries, .desc = TRUE), y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Marital Status",  
 x = "Marital Status",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



Reduce 7 marital\_status levels into a few semantically and analytically meaningful categories.

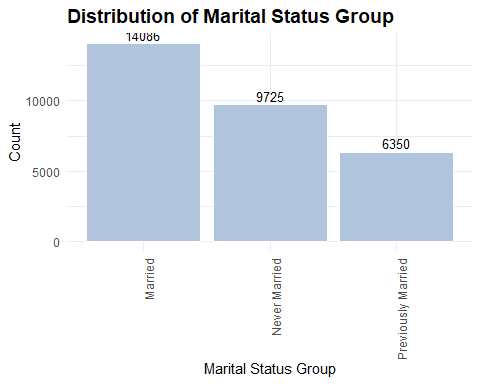
adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(marital\_status\_group = case\_when(  
 marital\_status %in% c("Married-civ-spouse", "Married-AF-spouse") ~ "Married",  
 marital\_status == "Never-married" ~ "Never Married",  
 marital\_status %in% c("Divorced", "Separated", "Widowed", "Married-spouse-absent") ~ "Previously Married",  
 TRUE ~ NA\_character\_))  
  
entries\_by\_marital\_status\_group <- adult\_income\_clean %>%  
 group\_by(marital\_status\_group) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_marital\_status\_group,  
 col.names = c("Marital Status Group", "Entries", "Percent"),  
 caption = "Entries By Marital Status Group",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Marital Status Group

| Marital Status Group | Entries | Percent |
| --- | --- | --- |
| Married | 14,086 | 0.47 |
| Never Married | 9,725 | 0.32 |
| Previously Married | 6,350 | 0.21 |

Plot marital status group.

ggplot(entries\_by\_marital\_status\_group, aes(x = forcats::fct\_reorder(marital\_status\_group, entries, .desc = TRUE), y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Marital Status Group",  
 x = "Marital Status Group",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



Create a binary variable that indicates whether an individual is married or not.

adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(is\_married = ifelse(marital\_status\_group == "Married", 1, 0))

## Occupation

Kind of work performed by the individual.

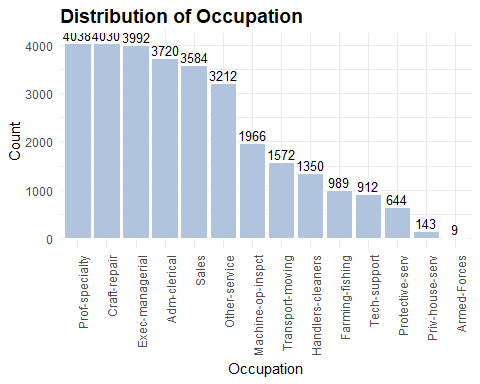
entries\_by\_occupation <- adult\_income\_clean %>%  
 group\_by(occupation) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_occupation,  
 col.names = c("Occupation", "Entries", "Percent"),  
 caption = "Entries By occupation",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By occupation

| Occupation | Entries | Percent |
| --- | --- | --- |
| Adm-clerical | 3,720 | 0.12 |
| Armed-Forces | 9 | 0.00 |
| Craft-repair | 4,030 | 0.13 |
| Exec-managerial | 3,992 | 0.13 |
| Farming-fishing | 989 | 0.03 |
| Handlers-cleaners | 1,350 | 0.04 |
| Machine-op-inspct | 1,966 | 0.07 |
| Other-service | 3,212 | 0.11 |
| Priv-house-serv | 143 | 0.00 |
| Prof-specialty | 4,038 | 0.13 |
| Protective-serv | 644 | 0.02 |
| Sales | 3,584 | 0.12 |
| Tech-support | 912 | 0.03 |
| Transport-moving | 1,572 | 0.05 |

Plot occupations.

ggplot(entries\_by\_occupation, aes(x = forcats::fct\_reorder(occupation, entries, .desc = TRUE), y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Occupation",  
 x = "Occupation",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



Group the detailed occupation values into broader, interpretable job sector categories to stabilize the model and improve interpretability.

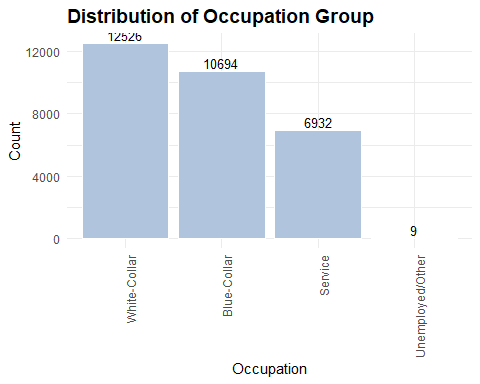
adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(occupation\_group = case\_when(  
 occupation %in% c("Exec-managerial", "Prof-specialty", "Tech-support", "Sales") ~ "White-Collar",  
 occupation %in% c("Craft-repair", "Machine-op-inspct", "Transport-moving", "Handlers-cleaners",  
 "Farming-fishing", "Priv-house-serv", "Protective-serv") ~ "Blue-Collar",  
 occupation %in% c("Adm-clerical", "Other-service") ~ "Service",  
 occupation %in% c("Armed-Forces", "?") ~ "Unemployed/Other",  
 TRUE ~ NA\_character\_))  
  
entries\_by\_occupation\_group <- adult\_income\_clean %>%  
 group\_by(occupation\_group) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_occupation\_group,  
 col.names = c("Occupation Group", "Entries", "Percent"),  
 caption = "Entries By Occupation Group",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Occupation Group

| Occupation Group | Entries | Percent |
| --- | --- | --- |
| Blue-Collar | 10,694 | 0.35 |
| Service | 6,932 | 0.23 |
| Unemployed/Other | 9 | 0.00 |
| White-Collar | 12,526 | 0.42 |

Plot occupation group.

ggplot(entries\_by\_occupation\_group, aes(x = forcats::fct\_reorder(occupation\_group, entries, .desc = TRUE), y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Occupation Group",  
 x = "Occupation",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



Create a binary variable indicating whether an individual has a white-collar occupation or not.

adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(is\_white\_collar = ifelse(occupation\_group == "White-Collar", 1, 0))

## Relationship

The relationship column in the Adult Income dataset represents the individual’s role within their household, based on how they relate to the person who filled out the Census form. It describes the respondent’s family role or household position. It’s not about romantic relationships per se, but rather household structure.

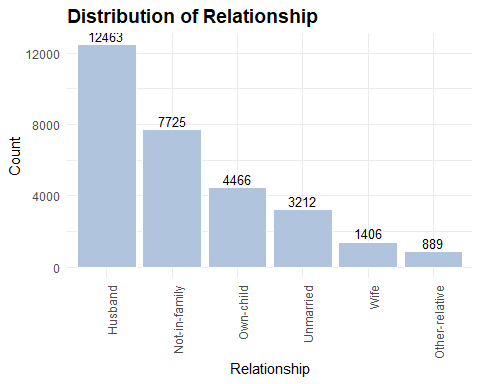
entries\_by\_relationship <- adult\_income\_clean %>%  
 group\_by(relationship) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_relationship,  
 col.names = c("Relationship", "Entries", "Percent"),  
 caption = "Entries By Relationship",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Relationship

| Relationship | Entries | Percent |
| --- | --- | --- |
| Husband | 12,463 | 0.41 |
| Not-in-family | 7,725 | 0.26 |
| Other-relative | 889 | 0.03 |
| Own-child | 4,466 | 0.15 |
| Unmarried | 3,212 | 0.11 |
| Wife | 1,406 | 0.05 |

Plot relationship.

ggplot(entries\_by\_relationship, aes(x = forcats::fct\_reorder(relationship, entries, .desc = TRUE), y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Relationship",  
 x = "Relationship",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



Here’s a common, modeling-friendly grouping.

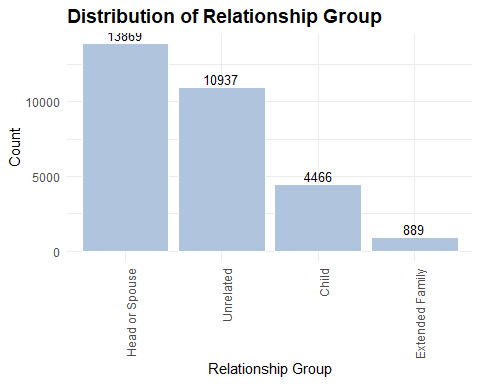
adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(relationship\_group = case\_when(  
 relationship %in% c("Husband", "Wife") ~ "Head or Spouse",  
 relationship == "Own-child" ~ "Child",  
 relationship == "Other-relative" ~ "Extended Family",  
 relationship %in% c("Not-in-family", "Unmarried") ~ "Unrelated",  
 TRUE ~ NA\_character\_))  
  
entries\_by\_relationship\_group <- adult\_income\_clean %>%  
 group\_by(relationship\_group) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_relationship\_group,  
 col.names = c("Relationship Group", "Entries", "Percent"),  
 caption = "Entries By Relationship Group",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Relationship Group

| Relationship Group | Entries | Percent |
| --- | --- | --- |
| Child | 4,466 | 0.15 |
| Extended Family | 889 | 0.03 |
| Head or Spouse | 13,869 | 0.46 |
| Unrelated | 10,937 | 0.36 |

Plot relationship group.

ggplot(entries\_by\_relationship\_group, aes(x = forcats::fct\_reorder(relationship\_group, entries, .desc = TRUE), y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Relationship Group",  
 x = "Relationship Group",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



## Race

Race classification per census data.

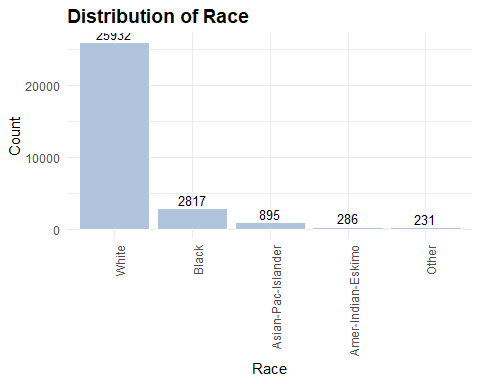
entries\_by\_race <- adult\_income\_clean %>%  
 group\_by(race) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_race,  
 col.names = c("Race", "Entries", "Percent"),  
 caption = "Entries By Race",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Race

| Race | Entries | Percent |
| --- | --- | --- |
| Amer-Indian-Eskimo | 286 | 0.01 |
| Asian-Pac-Islander | 895 | 0.03 |
| Black | 2,817 | 0.09 |
| Other | 231 | 0.01 |
| White | 25,932 | 0.86 |

Plot race.

ggplot(entries\_by\_race, aes(x = forcats::fct\_reorder(race, entries, .desc = TRUE), y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Race",  
 x = "Race",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



## Sex

Biological sex of the individual.

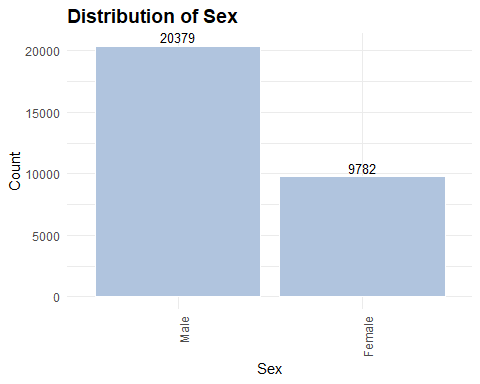
entries\_by\_sex <- adult\_income\_clean %>%  
 group\_by(sex) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_sex,  
 col.names = c("Sex", "Entries", "Percent"),  
 caption = "Entries By Sex",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Sex

| Sex | Entries | Percent |
| --- | --- | --- |
| Female | 9,782 | 0.32 |
| Male | 20,379 | 0.68 |

Plot sex.

ggplot(entries\_by\_sex, aes(x = forcats::fct\_reorder(sex, entries, .desc = TRUE), y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Sex",  
 x = "Sex",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



Are females sufficiently represented here? How does this compare to the US population as a whole?

According to U.S. Census Bureau estimates for the 1994–1995 period (which aligns with the dataset’s timeframe), the actual U.S. adult population was approximately 50–51% female and 49–50% male. So yes — women are underrepresented in this dataset by about 18–19 percentage points.

Why would this be the case?

Employment-focused sampling: The dataset is built from the Census Bureau’s Current Population Survey (CPS) — specifically tailored for labor market studies. Women were less likely to be in full-time or high-wage employment in the 1990s, which might have affected inclusion.

Income filter bias: If the dataset emphasizes workers with reported income, many part-time or unpaid caregivers (disproportionately women) might be excluded.

Historical gender roles: In 1994–1995, the labor force participation rate for women was about 58%, compared to 75%+ for men — this drives skewed representation in income-based surveys.

Impact of this?

If the model learns patterns that favor male-majority patterns, it may underfit or misclassify women or reinforce bias if deployed in real decision-making.

What do I do about this?

Check for model performance disparities and/or consider weighting or rebalancing to match the real-world gender population.

Create binary variable indicating whether an individual is a male or not.

adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(is\_male = ifelse(sex == "Male", 1, 0))

## Capital Gain and Capital Loss

These fields represent the profit or loss from the same of capital assets in the past year. They are reported in US dollars and apply to as very small subset of the population.

entries\_by\_capital\_assets <- adult\_income\_clean %>%  
 mutate(has\_capital\_gain = if\_else(capital\_gain > 0, 1, 0),  
 has\_capital\_loss = if\_else(capital\_loss > 0, 1, 0)) %>%  
 group\_by(has\_capital\_gain, has\_capital\_loss) %>%  
 summarise(entries = n())

## `summarise()` has grouped output by 'has\_capital\_gain'. You can override using  
## the `.groups` argument.

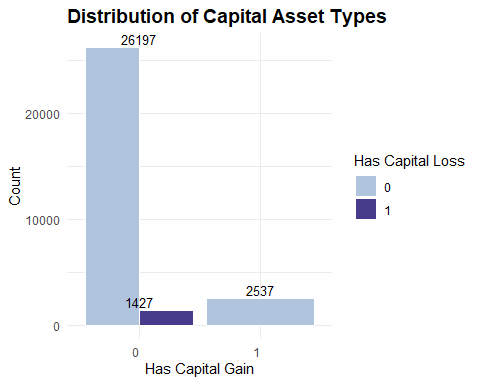
kable(entries\_by\_capital\_assets,  
 col.names = c("Has Capital Gain", "Has Capital Loss", "Entries"),  
 caption = "Entries By Capital Assets",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Capital Assets

| Has Capital Gain | Has Capital Loss | Entries |
| --- | --- | --- |
| 0 | 0 | 26,197 |
| 0 | 1 | 1,427 |
| 1 | 0 | 2,537 |

Plot capital asset occurrences.

ggplot(entries\_by\_capital\_assets, aes(x = factor(has\_capital\_gain),   
 y = entries,   
 fill = factor(has\_capital\_loss))) +  
 geom\_col(position = "dodge", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, hjust = "center", size = 3.5) +  
 labs(title = "Distribution of Capital Asset Types",  
 x = "Has Capital Gain",  
 y = "Count",  
 fill = "Has Capital Loss") +  
 scale\_fill\_manual(values = c("0" = "lightsteelblue", "1" = "slateblue4")) +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 0, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



26,197 people have no gain or loss reported, and 3,964 reported either a gain or a loss.

The best approach here is to convert the columns to binary flags, which will simplify the model to asking, “Did this person have a gain or loss?” This also avoids the effects of extreme values.

Add the binary columns.

adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(has\_capital\_gain = if\_else(capital\_gain > 0, 1, 0),  
 has\_capital\_loss = if\_else(capital\_loss > 0, 1, 0))

## Hours Per Week

Number of hours the person works per week.

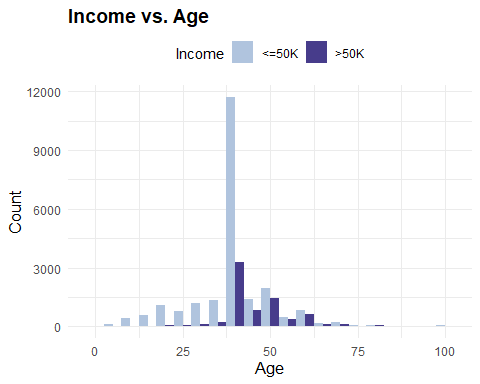
More hours equals higher likelihood of full-time employment, and thereby higher income. May be especially meaningful near thresholds like 35–40 hours (full-time vs. part-time).

Also helps differentiate between low-wage full-time and high-wage part-time work.

Its impact can varies by sex, occupation, and education. For example, a woman working 60 hours per week in tech may earn more than a man working 30 hours per week in clerical.

Plot hours per week.

ggplot(adult\_income\_clean, aes(x = hours\_per\_week, fill = income)) +  
 geom\_histogram(binwidth = 5, position = "dodge") +  
 labs(title = "Income vs. Hours Worked per Week") +  
 scale\_fill\_manual(values = c("<=50K" = "lightsteelblue", ">50K" = "slateblue4")) +  
 labs(title = "Income vs. Age", x = "Age", y = "Count", fill = "Income") +  
 theme\_minimal() +  
 theme(plot.title = element\_text(size = 14, face = "bold"),  
 axis.title = element\_text(size = 12),  
 legend.title = element\_text(size = 11),  
 legend.position = "top")



There is a strong peak at 40 hours per week; this is expected since 40 hours is the standard for full-time work in the US. It dominates both groups but especially the <=50k group.

Beyond 40 hours per week, the proportion of higher earners increases, suggesting that working longer hours is correlated with higher income. Lower hours lean toward <=50k and those working under 35 hours per week are much more likely to be in this group. This could reflect part-time gigs or under-employment situations.

Will keep the variable; it’s obviously informative.

Will create a categorical version by binning the values then check the proportions by group.

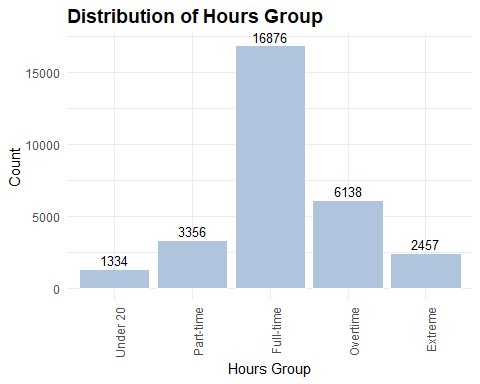
adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(hours\_group = case\_when(hours\_per\_week < 20 ~ "Under 20",  
 hours\_per\_week < 35 ~ "Part-time",  
 hours\_per\_week < 45 ~ "Full-time",  
 hours\_per\_week < 60 ~ "Overtime",  
 TRUE ~ "Extreme"))  
  
entries\_by\_hours\_group <- adult\_income\_clean %>%  
 group\_by(hours\_group) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2)) %>%  
 mutate(hours\_group = factor(hours\_group,  
 levels = c("Under 20",   
 "Part-time",   
 "Full-time",   
 "Overtime",   
 "Extreme")))  
  
kable(entries\_by\_hours\_group,  
 col.names = c("Hours Group", "Entries", "Percent"),  
 caption = "Entries By Hours Group",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Hours Group

| Hours Group | Entries | Percent |
| --- | --- | --- |
| Extreme | 2,457 | 0.08 |
| Full-time | 16,876 | 0.56 |
| Overtime | 6,138 | 0.20 |
| Part-time | 3,356 | 0.11 |
| Under 20 | 1,334 | 0.04 |

Plot hours groups.

ggplot(entries\_by\_hours\_group, aes(x = hours\_group, y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Hours Group",  
 x = "Hours Group",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))

 Create binary variable to indicate whether a person works full-time or not.

adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(works\_full\_time = ifelse(hours\_group %in% c("Full-time", "Overtime"), 1, 0))

## Native Country

The country where the respondent was born (not necessarily their citizenship). All individuals in the dataset are living in the U.S. at the time of the survey.

Some are likely naturalized U.S. citizens, others are permanent residents, temporary workers, or possibly undocumented (though CPS doesn’t explicitly categorize that).

This variable often proxies for immigrant status, which correlates with income, education, language fluency, and occupational access.

It’s especially relevant when modeling income, because language and cultural barriers may impact earning potential, some immigrant groups may face systemic labor market disadvantages or advantages. However, it does not distinguish between naturalized vs. non-citizens.

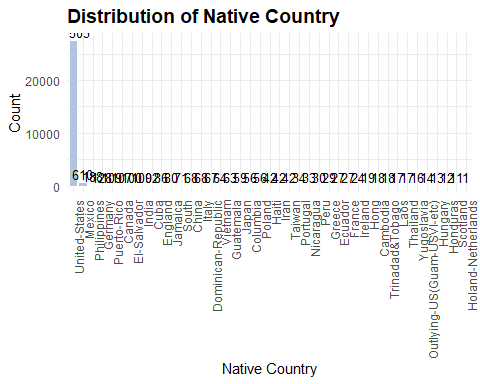
entries\_by\_native\_country <- adult\_income\_clean %>%  
 group\_by(native\_country) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_native\_country,  
 col.names = c("Native Country", "Entries", "Percent"),  
 caption = "Entries By Native Country",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Native Country

| Native Country | Entries | Percent |
| --- | --- | --- |
| Cambodia | 18 | 0.00 |
| Canada | 107 | 0.00 |
| China | 68 | 0.00 |
| Columbia | 56 | 0.00 |
| Cuba | 92 | 0.00 |
| Dominican-Republic | 67 | 0.00 |
| Ecuador | 27 | 0.00 |
| El-Salvador | 100 | 0.00 |
| England | 86 | 0.00 |
| France | 27 | 0.00 |
| Germany | 128 | 0.00 |
| Greece | 29 | 0.00 |
| Guatemala | 63 | 0.00 |
| Haiti | 42 | 0.00 |
| Holand-Netherlands | 1 | 0.00 |
| Honduras | 12 | 0.00 |
| Hong | 19 | 0.00 |
| Hungary | 13 | 0.00 |
| India | 100 | 0.00 |
| Iran | 42 | 0.00 |
| Ireland | 24 | 0.00 |
| Italy | 68 | 0.00 |
| Jamaica | 80 | 0.00 |
| Japan | 59 | 0.00 |
| Laos | 17 | 0.00 |
| Mexico | 610 | 0.02 |
| Nicaragua | 33 | 0.00 |
| Outlying-US(Guam-USVI-etc) | 14 | 0.00 |
| Peru | 30 | 0.00 |
| Philippines | 188 | 0.01 |
| Poland | 56 | 0.00 |
| Portugal | 34 | 0.00 |
| Puerto-Rico | 109 | 0.00 |
| Scotland | 11 | 0.00 |
| South | 71 | 0.00 |
| Taiwan | 42 | 0.00 |
| Thailand | 17 | 0.00 |
| Trinadad&Tobago | 18 | 0.00 |
| United-States | 27,503 | 0.91 |
| Vietnam | 64 | 0.00 |
| Yugoslavia | 16 | 0.00 |

Plot native countries.

ggplot(entries\_by\_native\_country, aes(x = forcats::fct\_reorder(native\_country, entries, .desc = TRUE), y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Native Country",  
 x = "Native Country",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



Group countries into regional economic categories.

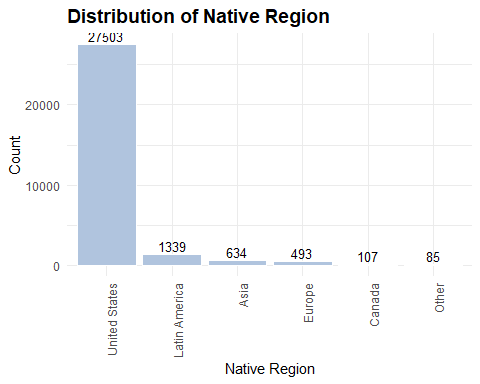
adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(native\_region = case\_when(  
 native\_country == "United-States" ~ "United States",  
 native\_country %in% c("Mexico", "El-Salvador", "Guatemala", "Honduras", "Nicaragua", "Columbia",   
 "Ecuador", "Peru", "Cuba", "Puerto-Rico", "Dominican-Republic",   
 "Trinadad&Tobago", "Jamaica", "Haiti") ~ "Latin America",  
 native\_country %in% c("China", "India", "Japan", "Vietnam", "Philippines", "Iran",   
 "Taiwan", "Thailand", "Cambodia", "Hong", "Laos") ~ "Asia",  
 native\_country %in% c("England", "France", "Germany", "Greece", "Hungary", "Ireland",   
 "Italy", "Poland", "Portugal", "Scotland", "Yugoslavia", "Holand-Netherlands") ~ "Europe",  
 native\_country == "Canada" ~ "Canada",  
 TRUE ~ "Other"))  
  
entries\_by\_native\_region <- adult\_income\_clean %>%  
 group\_by(native\_region) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_native\_region,  
 col.names = c("Native Region", "Entries", "Percent"),  
 caption = "Entries By Native Region",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Native Region

| Native Region | Entries | Percent |
| --- | --- | --- |
| Asia | 634 | 0.02 |
| Canada | 107 | 0.00 |
| Europe | 493 | 0.02 |
| Latin America | 1,339 | 0.04 |
| Other | 85 | 0.00 |
| United States | 27,503 | 0.91 |

Plot native regions.

ggplot(entries\_by\_native\_region, aes(x = forcats::fct\_reorder(native\_region, entries, .desc = TRUE), y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Native Region",  
 x = "Native Region",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 90, hjust = 1),  
 plot.title = element\_text(size = 14, face = "bold"))



## Income

The target variable, whether income is <=50K or >50K per year.

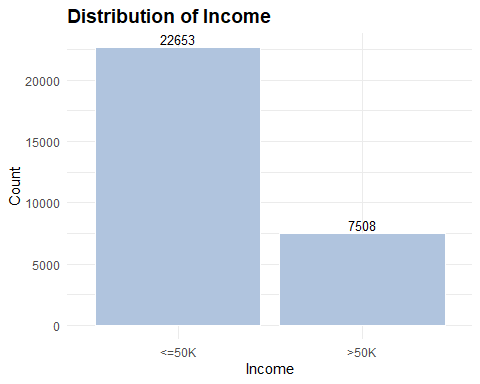
entries\_by\_income <- adult\_income\_clean %>%  
 group\_by(income) %>%  
 summarise(entries = n()) %>%  
 mutate(percent = round(entries / sum(entries), 2))  
  
kable(entries\_by\_income,  
 col.names = c("Income", "Entries", "Percent"),  
 caption = "Entries By Income",  
 format.args = list(big.mark = ","),  
 align = c("l", "r", "r"))

Entries By Income

| Income | Entries | Percent |
| --- | --- | --- |
| <=50K | 22,653 | 0.75 |
| >50K | 7,508 | 0.25 |

Plot income.

ggplot(entries\_by\_income, aes(x = income, y = entries)) +  
 geom\_col(fill = "lightsteelblue", color = "white") +  
 geom\_text(aes(label = entries), vjust = -0.3, size = 3.5) +  
 labs(title = "Distribution of Income",  
 x = "Income",  
 y = "Count") +  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 0),  
 plot.title = element\_text(size = 14, face = "bold"))



There’s a class imbalance here, i.e. a three-to-one imbalance between the two groups, which is perhaps expected. But what does this mean for this model?

The distribution makes sense. The data reflects 1994-1995 US income and, adjusted for inflation, $50k then is approximately $100k today. It is also important to note that CPS oversamples lower-income households for poverty and economic policy research.

A naive model could predict that everyone earns less than $50k and still be 75% accurate; we don’t learn anything useful from this. High accuracy doesn’t mean high model quality when the classes are imbalanced.

Logistic regression estimates the probability of the positive class >50k and with fewer examples of high earners, the model may under-predict the probability of <50k because of the bias toward the majority class.

If we look at just overall accuracy, we’ll miss how the model performs for each class. We need to use class-sensitive metrics to confirm how good the model is at identifying high earners as well as majority class members.

Create a binary version of income.

adult\_income\_clean <- adult\_income\_clean %>%  
 mutate(income\_binary = ifelse(income == ">50K", 1, 0))

# Save Datasets

The final dataset includes 28 variables, which includes grouped and binary variables I engineered for purposes of model testing and comparison.

saveRDS(adult\_income\_clean, "../data/processed/adult\_income\_clean.rds")

Create training and test sets with a 70/30 random split.

set.seed(91210)  
  
trainIndex <- caret::createDataPartition(adult\_income\_clean$income\_binary, p = 0.7, list = FALSE)  
  
adult\_income\_train <- adult\_income\_clean[trainIndex, ]  
  
adult\_income\_test <- adult\_income\_clean[-trainIndex, ]  
  
saveRDS(adult\_income\_train, "../data/processed/adult\_income\_train.rds")  
  
saveRDS(adult\_income\_test, "../data/processed/adult\_income\_test.rds")