NLSY79 longitudinal project

AUTHOR
Cristian T

PUBLISHED
August 5, 2024

```
suppressMessages(library(tidyverse))
suppressMessages(library(tibble))
suppressMessages(library(dplyr))
suppressMessages(library(ggplot2))
suppressMessages(library(scales))
suppressMessages(library(RColorBrewer))

# Read the RData files into a data frame
load("physical_data_nlsy79.RData")
load("education_data_nlsy79.RData")
load("income_data_nlsy79.RData")
```

Join Data

Our analysis examines income, education, and physical characteristics in 2014, so we joined the relevant datasets and restricted the year.

```
data <- inner_join(income_data_nlsy79, education_data_nlsy79,
                   by=c("CASEID", "year"))
data <- inner_join(data, physical_data_nlsy79,
                   by=c("CASEID", "year")) %>%
  select(CASEID, year, income, education, hair, sex)
data <- data %>% filter(year==2014)
glimpse(data)
```

```
Rows: 12,686
Columns: 6
$ CASEID    <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1…
$ year      <int> 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, 2014, …
$ income    <int> NA, 21000, 40000, NA, NA, 112000, NA, 47000, 80000, NA, NA, …
$ education <int> NA, 12, 10, NA, NA, 16, NA, 14, 14, NA, NA, NA, NA, 18, 16, …
$ hair      <chr> NA, "light brown", "blond", "light brown", NA, "brown", "bro…
$ sex       <chr> "female", "female", "female", "female", "male", "male", "mal…
```

After joining and limiting the year to 2014, the filtered data has 12686 rows and six columns.

Exploration of the Income Variable

We will begin by running some basic summary statistics to provide us with a basic understanding of the distribution and characteristics of the income data in the dataset. We first run a check for missing values to assess data completeness. We further explore the measures of central tendency (mean, median), dispersion (standard deviation), and range (minimum and maximum values).

```
# Check for missing values in the income column
missing_incomes <- sum(is.na(data$income))
```

```
        # Mean
        mean_income <- mean(data$income, na.rm = TRUE)
        # Median
        median_income <- median(data$income, na.rm = TRUE)
        # Standard deviation
        sd_income <- sd(data$income, na.rm = TRUE)
        # Minimum
        min_income <-min(data$income, na.rm = TRUE)
        # Maximum
        max_income <- max_income <- max(data$income, na.rm = TRUE)
        # Printing the results
        cat("\nMissing Incomes:", missing_incomes, "\n")
```

```
Missing Incomes: 5903
```

```
        cat("\nMean Income:", mean_income, "\n")
```

```
Mean Income: 42403.01
```

```
        cat("Median Income:", median_income, "\n")
```

```
Median Income: 28000
```

```
        cat("Standard Deviation of Income:", sd_income, "\n")
```

```
Standard Deviation of Income: 60847.59
```

```
        cat("Minimum Income:", min_income, "\n")
```

```
Minimum Income: 0
```

```
        cat("Maximum Income:", max_income, "\n")
```

```
Maximum Income: 370314
```

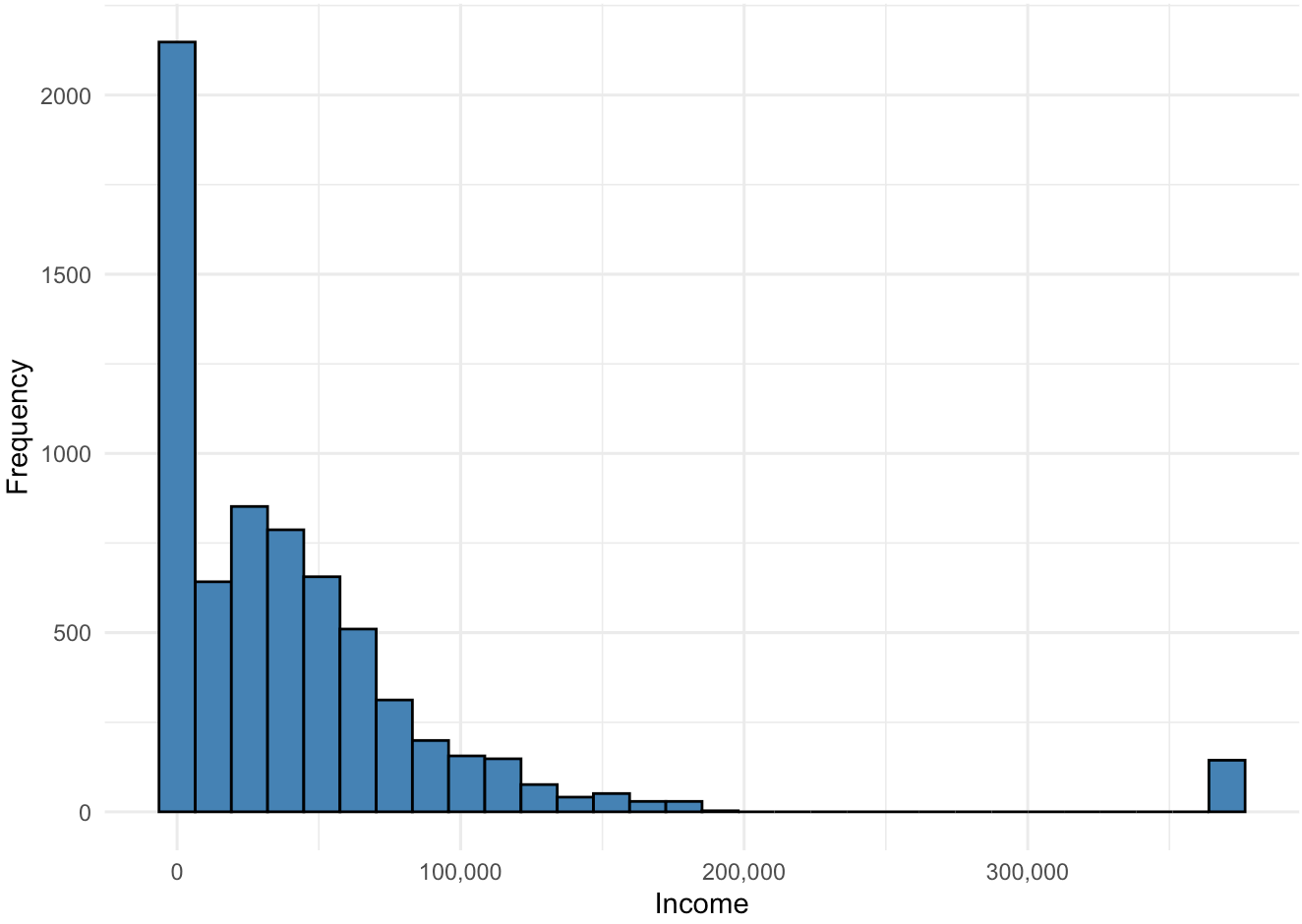
```
        #overview of Summary statistics
        summary(income_data_nlsy79$income)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
      0    1344   12000   19867   26000  370314   85628
```

Visually inspecting income we see the following plot.

```
        p <- ggplot(data, aes(x = income)) +
          geom_histogram(bins = 30, fill = "steelblue", color = "black") +  # Specify fill ar
          labs(x = "Income", y = "Frequency") +
          scale_x_continuous(labels = scales::comma) +
          theme_minimal()
```

```
ggsave("income.png", plot = p)
p
```



Based on the above graph, over 2000 responses show zero income in 2014, but we will keep our zero values. We want to explore the level of education and how it correlates with income.

Remove any rows with missing values in income.

```
data <- data %>%
  filter(!is.na(income))
```

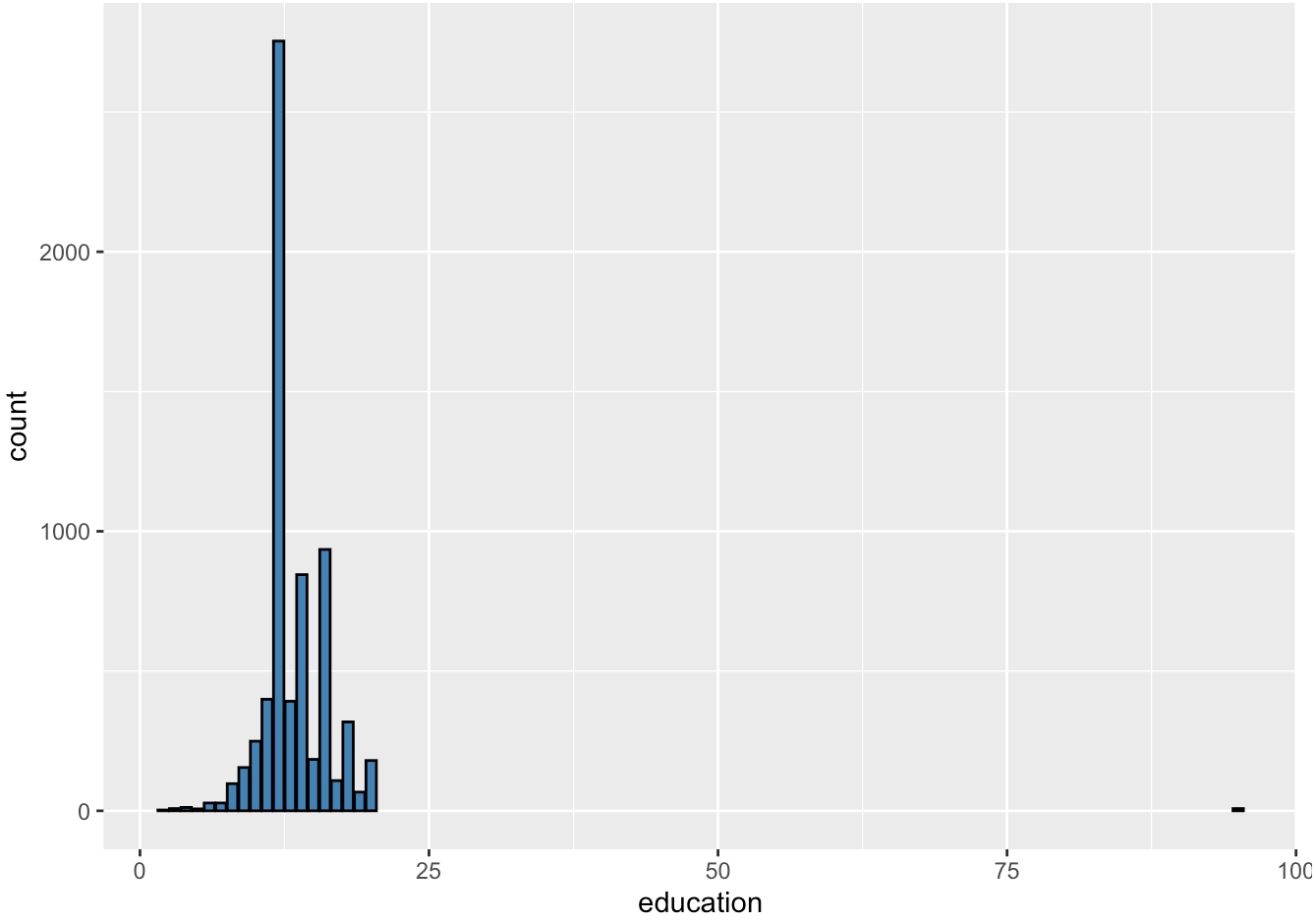
Study of Education

```
p <- ggplot(data = data, aes(x = education)) + geom_bar(fill = "steelblue", color = '
ggsave("education.png", plot = p)
```

```
Saving 7 x 5 in image

Warning: Removed 6 rows containing non-finite outside the scale range
(`stat_count()`).
```

```
p
```

```
Warning: Removed 6 rows containing non-finite outside the scale range
(`stat_count()`).
```

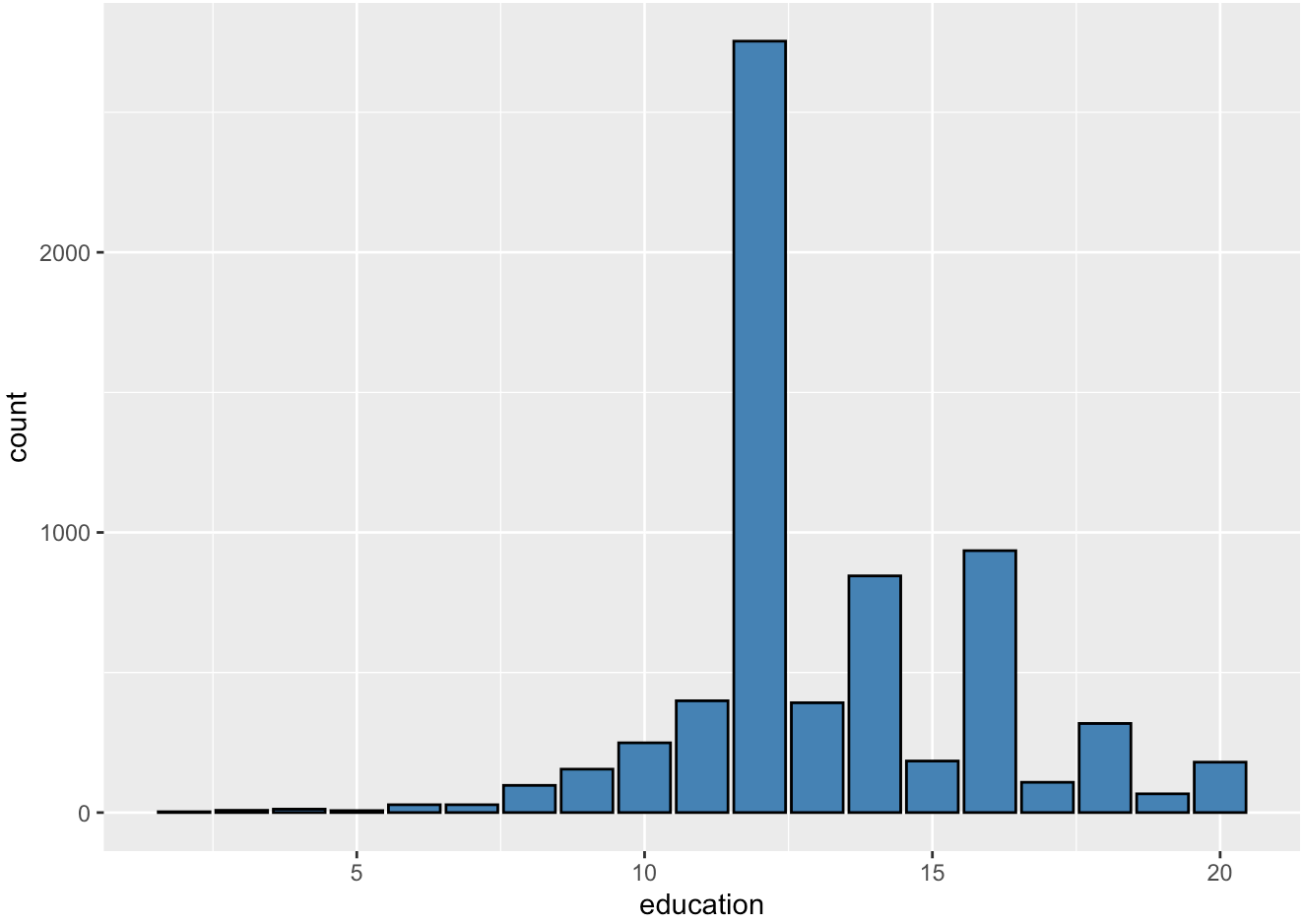


This plot reveals the spurious value 95, meaning some ungraded education. We'll exclude these from the analysis. Evidently, there are few of these values in the education data.

```
p <- ggplot(
  data = filter(data, education < 95),
  aes(x = education)
) + geom_bar(fill = "steelblue", color = "black")
ggsave("education.png", plot = p)
```

```
Saving 7 x 5 in image
```

```
p
```




```
        data <- data %>% filter(education<95)
```

```
        sum(is.na(data$education))
```

```
[1] 0
```

There is no missing value in education now.

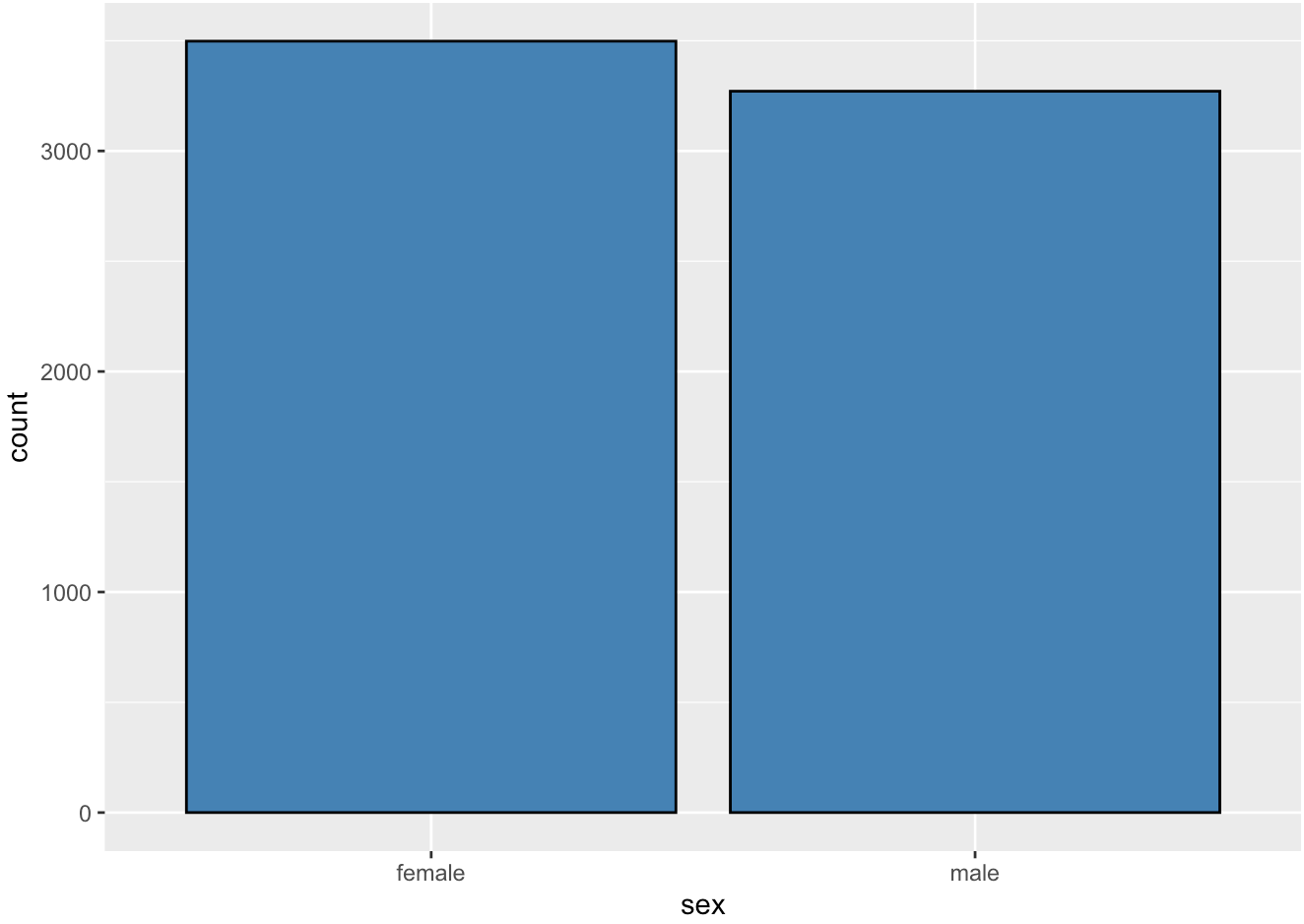
Study of Gender

Let's plot sex to examine the distribution.

```
        p <- ggplot(data = data, aes(x = sex)) + geom_bar(fill = "steelblue", color = "black'
        ggsave("sex.png", plot = p)
```

```
Saving 7 x 5 in image
```

```
        p
```




```
        sum(is.na(data$sex))
```

```
[1] 0
```

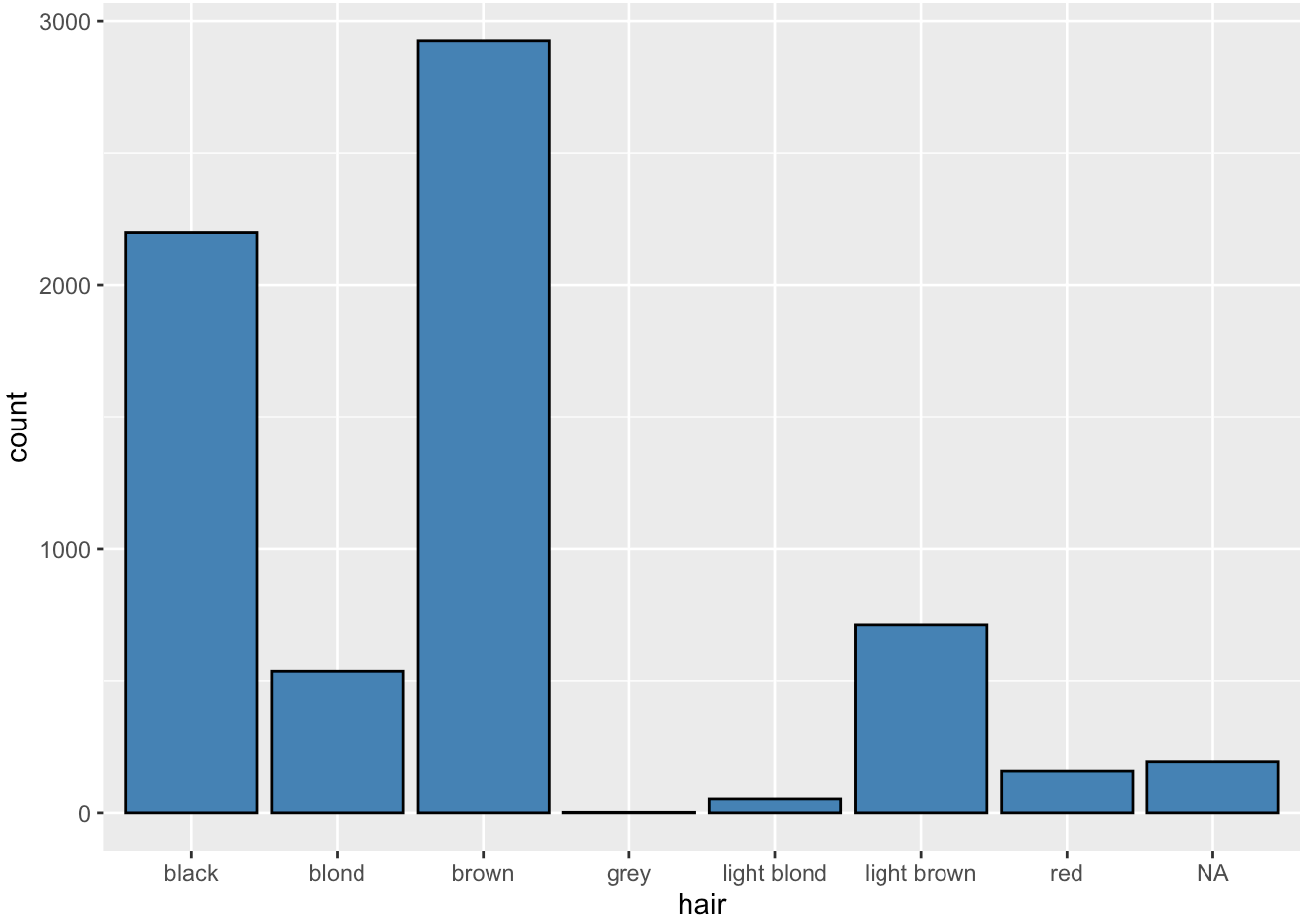
There is no missing value in sex now.

Study of Hair

```
        p <- ggplot(data = data,  aes(x = hair) ) + geom_bar(fill = "steelblue", color = "bla
        ggsave("hair.png", plot = p)
```

```
Saving 7 x 5 in image
```

```
        p
```



This graph shows that over 2,500 respondents have brown hair, and over 2,000 respondents have black hair.

```
sum(is.na(data$hair))
```

```
[1] 191
```

There are 191 missing values in hair.

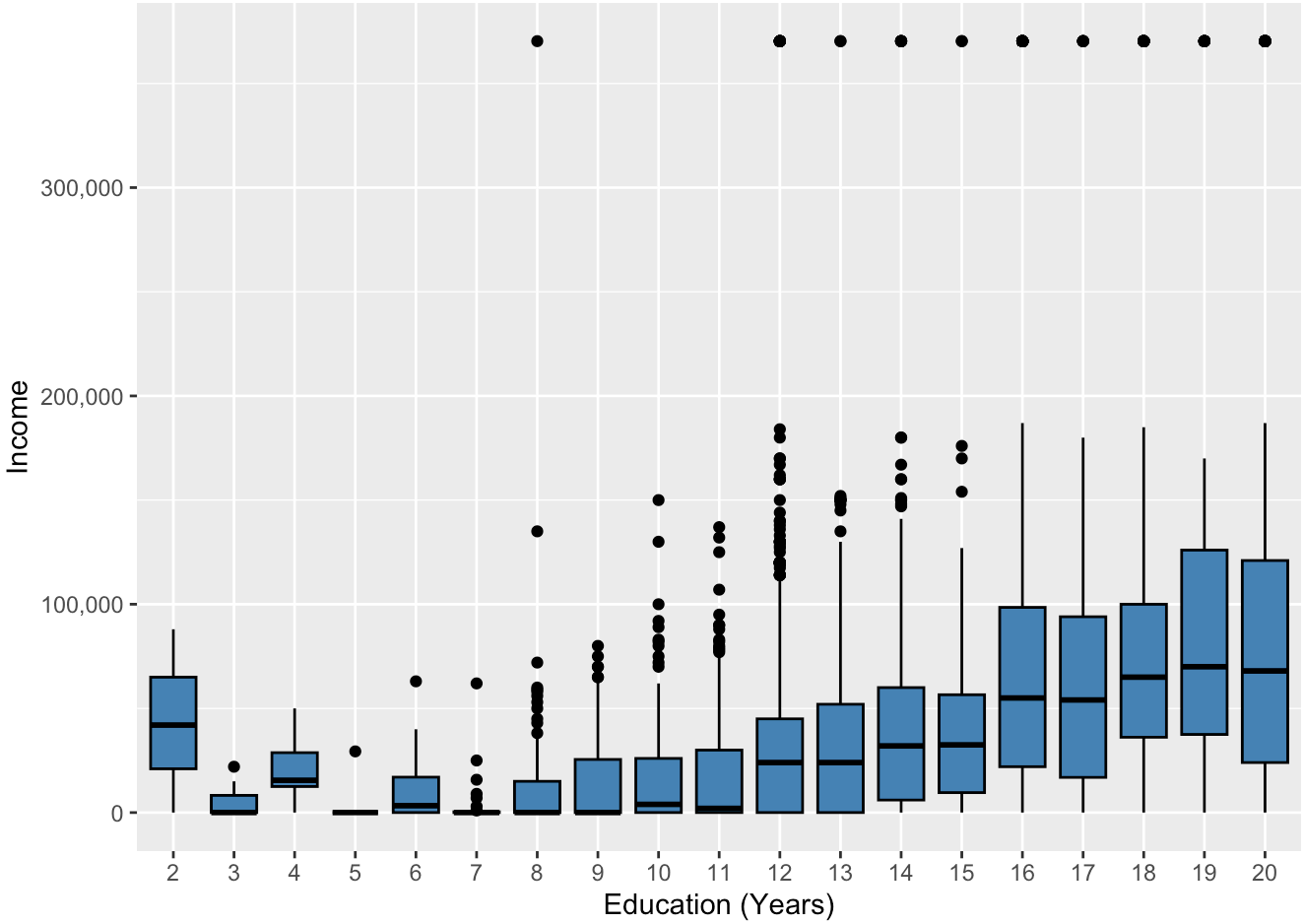
Explore the Effects of Education on Income

In this section, we are looking at the connection between education levels and income. We are trying to understand income distribution dynamics concerning different educational milestones and the individuals' mean income.

```
#plot showing distribution education and income
p <- ggplot(data, aes(x = as.factor(education), y = income)) +
  geom_boxplot(fill = "steelblue", color = "black") +
  labs(x = "Education (Years)", y = "Income") +
  scale_y_continuous(labels = scales::comma)
ggsave("income_education.png", plot = p)
```

```
Saving 7 x 5 in image
```

```
p
```



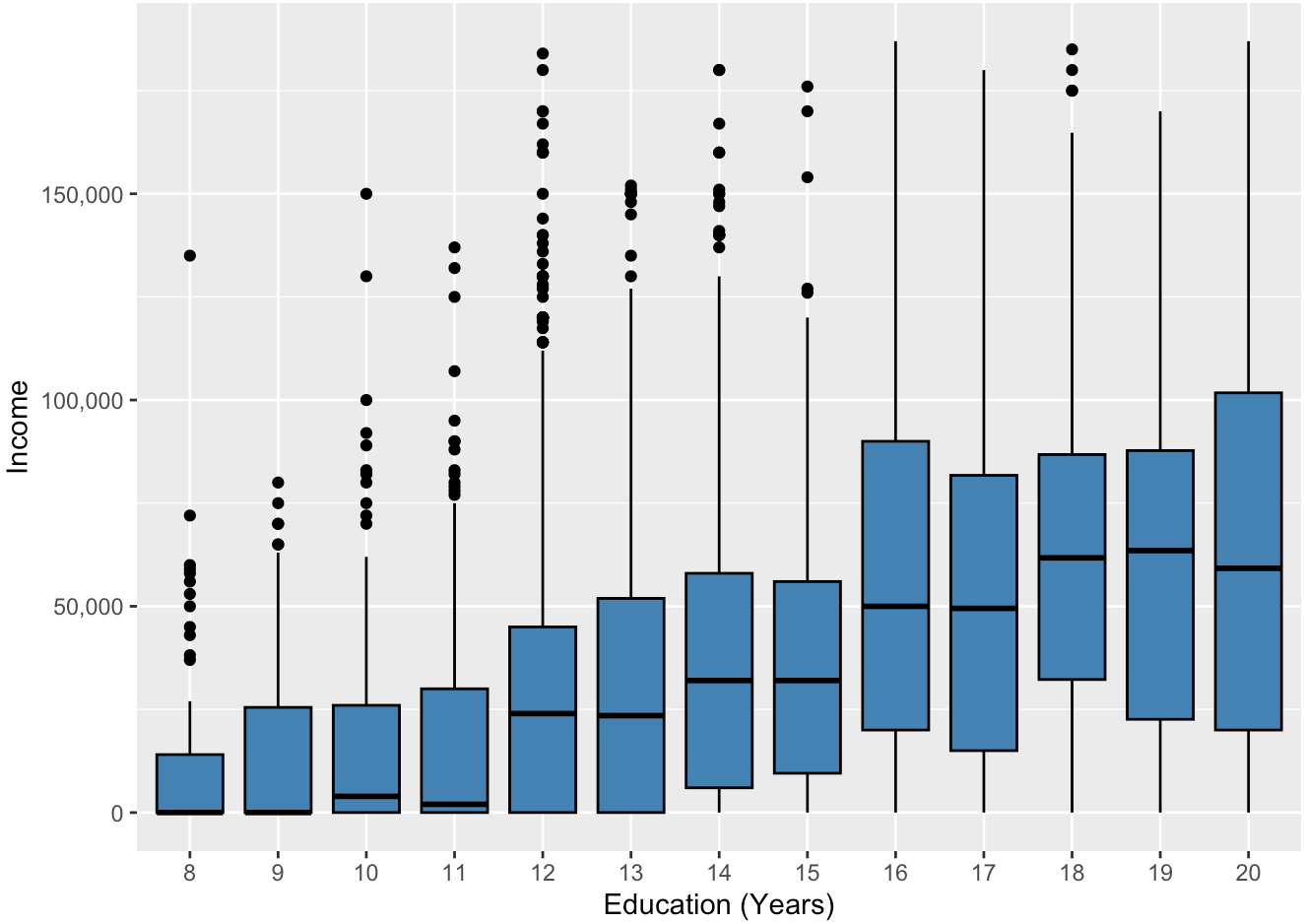
In the above graph, we are exploring the relationship between education levels and income. The chart presents a box plot showing income distribution across various education levels, using blue boxes to highlight the interquartile range of incomes within each group. The plot suggests that there is a positive association between income and years of education.

To further clean things up, we will exclude any education under the 8th grade. We are performing this cleanse to comply with inaccuracies and following the law that most states have some compulsory education laws and adjusting for potential age allowances that some states will allow one to drop out of school.

```
p <- ggplot(
  filter(data, education >= 8  & income <= 200000),
  aes(x = as.factor(education), y = income)) +
  geom_boxplot(fill = "steelblue", color = "black") +
  labs(x = "Education (Years)", y = "Income") +
  scale_y_continuous(labels = scales::comma)
ggsave("income_education_yrs.png", plot = p)
```

```
Saving 7 x 5 in image
```

```
p
```

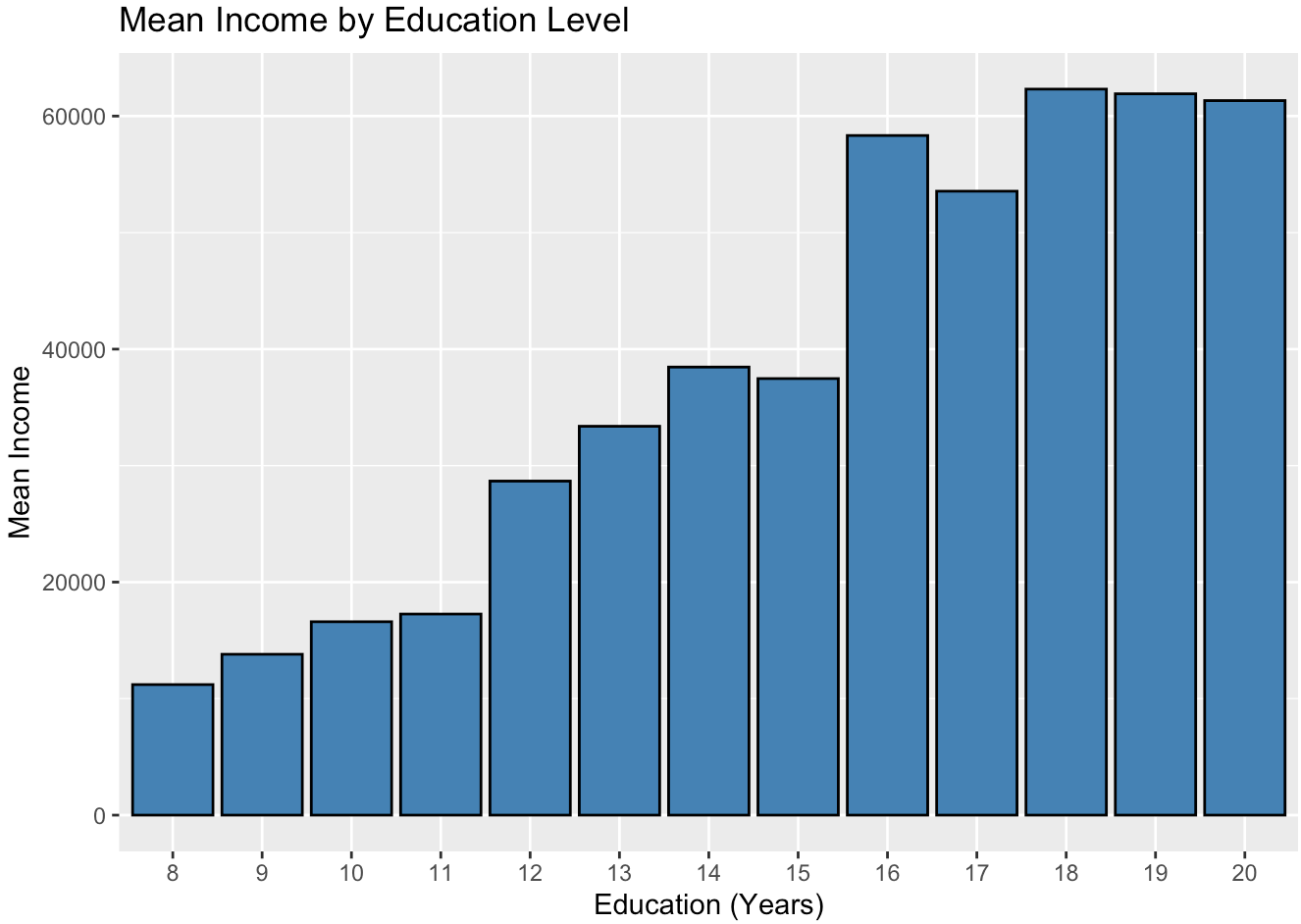


The positive linear relationship becomes more evident after removing the respondents with less than eight years of education, highlighting the influence of education level on income.

```
income_by_education_mean <- data %>%
  filter(income <= 200000, education >= 8) %>%
  group_by(education) %>%
  summarize(mean_income = mean(income, na.rm = TRUE))

p <- ggplot(income_by_education_mean,
       aes(x = as.factor(education), y = mean_income)) +
  geom_bar(stat = "identity", fill = "steelblue", color = "black") +
  labs(x = "Education (Years)", y = "Mean Income", title = "Mean Income by Education
ggsave("mean_income_by_years.png", plot = p)
```

```
Saving 7 x 5 in image
```

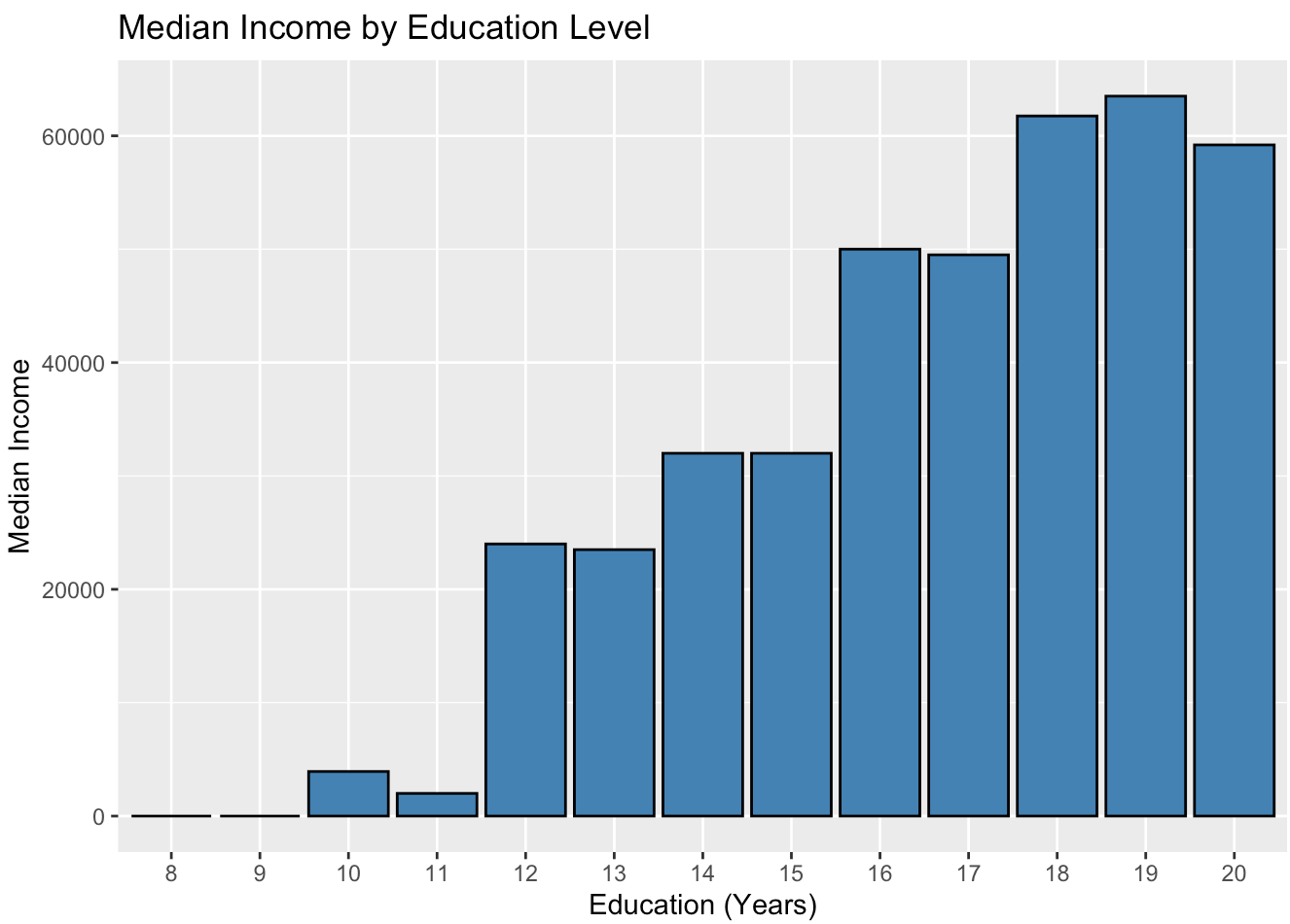
```
p
```




```
income_by_education_median <- data %>%
  filter(income <= 200000, education >= 8) %>%
  group_by(education) %>%
  summarize(median_income = median(income, na.rm = TRUE))

p <- ggplot(income_by_education_median, aes(x = as.factor(education), y = median_inco
  geom_bar(stat = "identity", fill = "steelblue", color = "black") +
  labs(x = "Education (Years)", y = "Median Income", title = "Median Income by Educat
ggsave("mean_income_by_level.png", plot = p)
```

```
Saving 7 x 5 in image
```

```
p
```

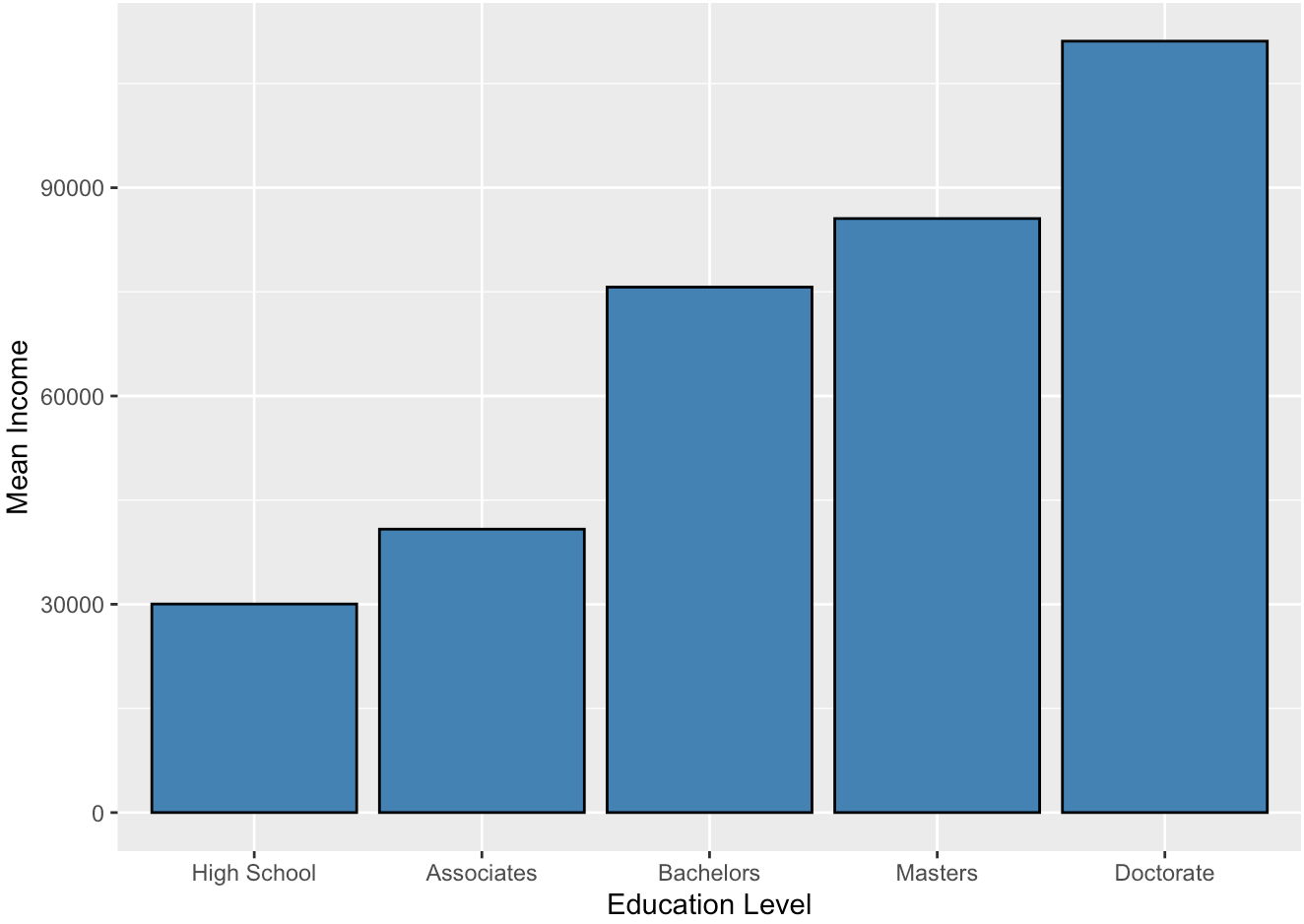


Continuing our examination, we look at the relationship between education levels and income by summarizing individuals' mean and median income better to understand the income distribution within each education group. The bar chart illustrates how mean and median income differs across these academic years. Looking at both of the graphs, we see how much those exceptionally high incomes have affected the overall mean income, while the median incomes appears to be reflective of the majority of individuals.

```
# Filter the data for the specific education levels
education_levels <- c(12, 14, 16, 18, 20)
income_by_education <- data %>%
  filter(education %in% education_levels)

# Create a bar plot for mean
p <- ggplot(income_by_education, aes(x = factor(education), y = income)) +
  geom_bar(stat = "summary", fun = "mean", fill = "steelblue", color = "black") +
  labs(x = "Education Level", y = "Mean Income") +
  scale_x_discrete(labels = c("12" = "High School", "14" = "Associates", "16" = "Bach
ggsave("mean_income_by_education.png", plot = p)
```

```
Saving 7 x 5 in image
```

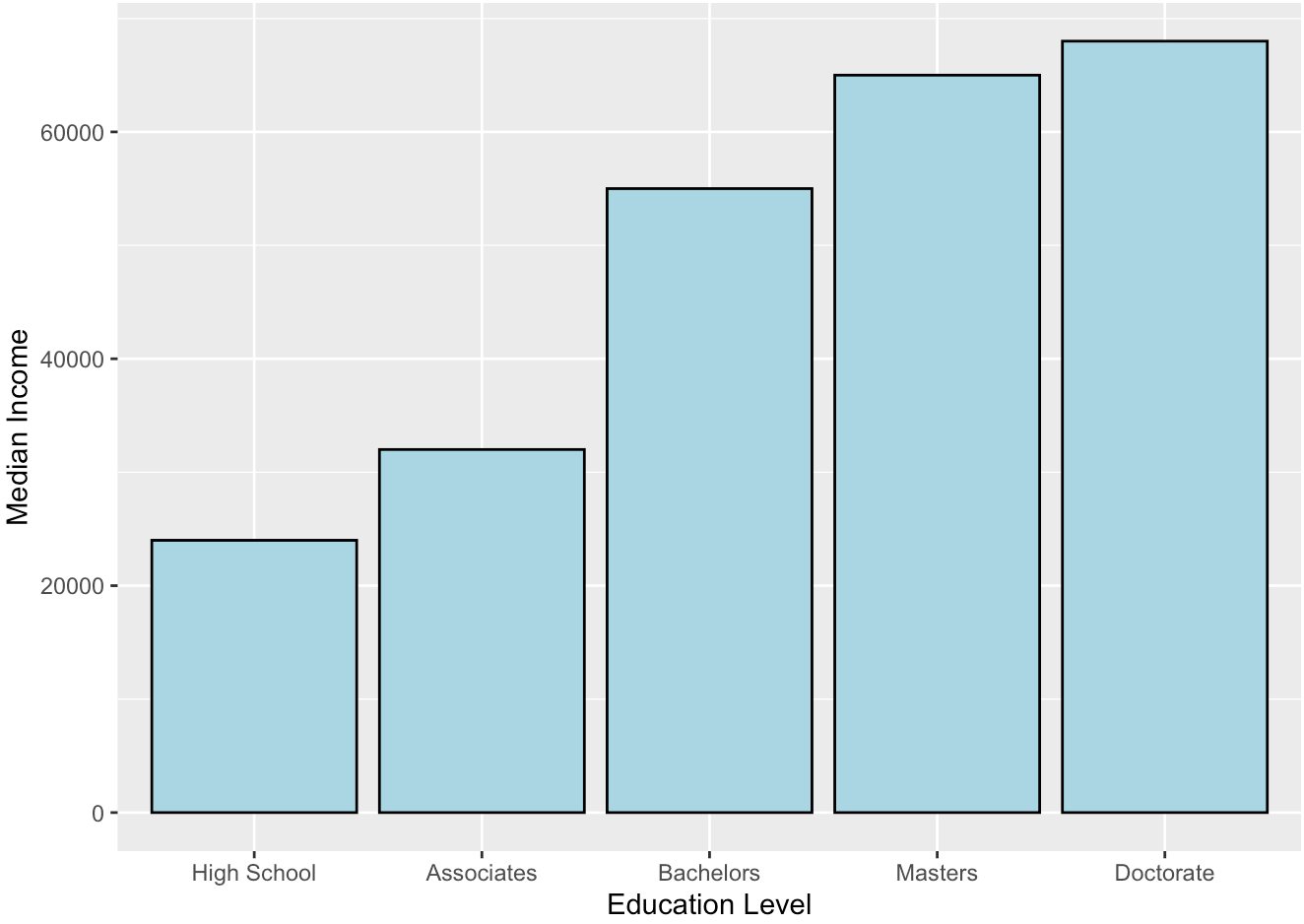
```
p
```




```
# Create a bar plot for median
q <- ggplot(income_by_education, aes(x = factor(education), y = income)) +
  geom_bar(stat = "summary", fun = "median", fill = "lightblue", color = "black") +
  labs(x = "Education Level", y = "Median Income") +
  scale_x_discrete(labels = c("12" = "High School", "14" = "Associates", "16" = "Bach
ggsave("mean_income_by_education.png", plot = q)
```

```
Saving 7 x 5 in image
```

```
q
```

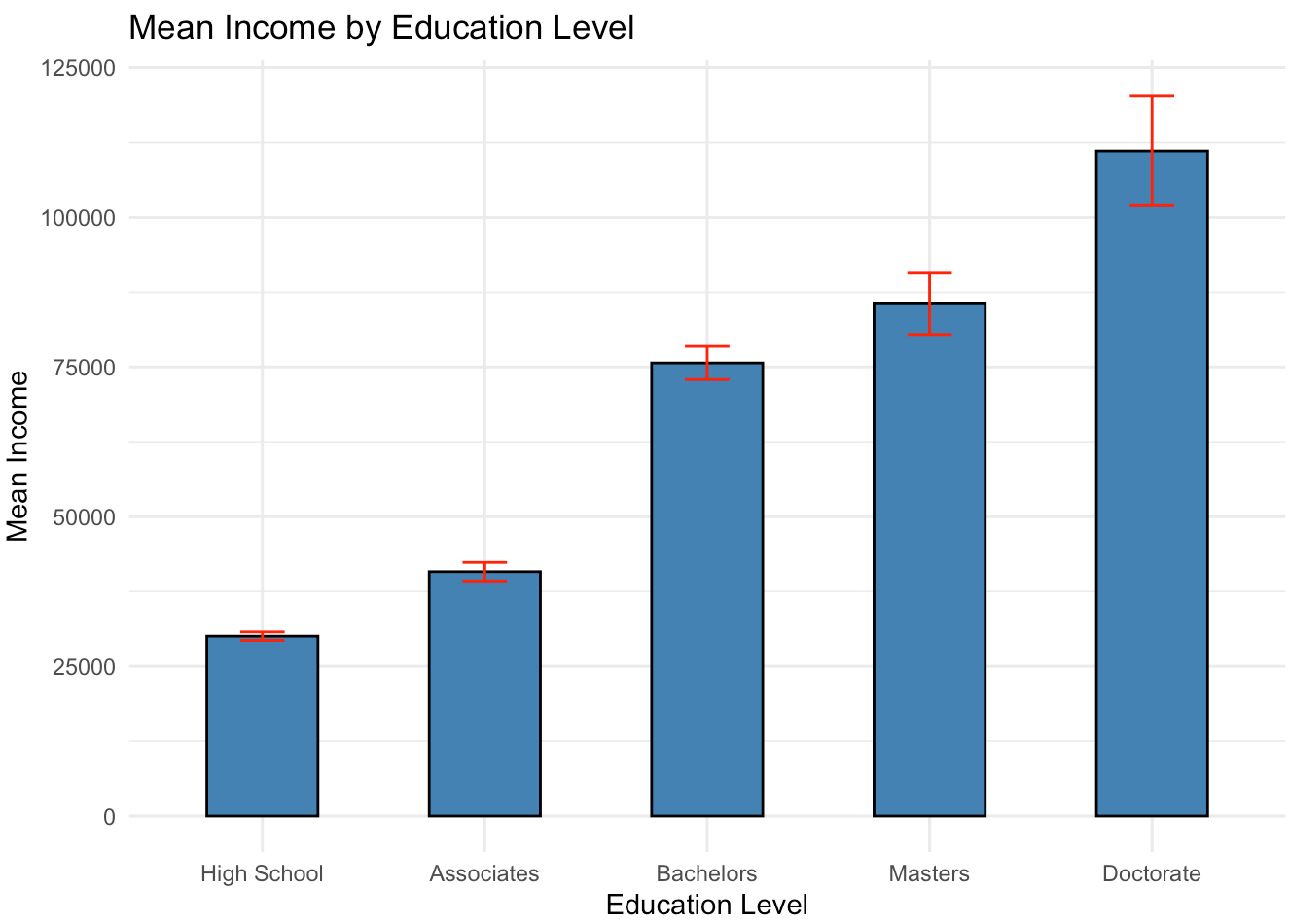


This second series of graphs focuses on specific education levels (12, 14, 16, 18, 20) corresponding to milestones such as high school, associate's degrees, bachelor's degrees, MBAs, and doctorates. We calculate and visualize the mean and median income for individuals with these distinct education levels.

```
# Calculate summary statistics for income by education level
summary_stats <- income_by_education %>%
  group_by(education) %>%
  summarise(mean_income = mean(income, na.rm = TRUE),
            sd_income = sd(income, na.rm = TRUE),
            n = n())

# Create a bar plot with error bars
p <- ggplot(summary_stats, aes(x = factor(education), y = mean_income)) +
  geom_bar(stat = "identity", fill = "steelblue", color = "black", width = 0.5) +
  geom_errorbar(aes(ymin = mean_income - sd_income / sqrt(n), ymax = mean_income + sd
                width = 0.2, color = "red") +
  labs(x = "Education Level", y = "Mean Income", title = "Mean Income by Education Le
  scale_x_discrete(labels = c("12" = "High School", "14" = "Associates", "16" = "Bach
  theme_minimal()
ggsave("mean_income_by_education_level.png", plot = p)
```

```
Saving 7 x 5 in image
```

```
p
```



We are looking at the mean income for each education level achieved in 2014. This graph shows the average income for individuals with each corresponding education level. We added the red error bars to provide insights into the reliability of these income estimates, showing us the variability in income within each education level and representing the uncertainty or variability in the average income estimates for each education level. The height of these bars reflects the degree of variability within each education level's income, with taller bars signifying greater variability and shorter bars indicating less uncertainty in the income estimate

Comparing Income to Physical Attributes

We will begin exploring the relationship between income and physical attributes, focusing on hair color.

```
# Group hair colors into broader categories
data <- data %>%
    filter(!is.na(hair), !is.na(income)) %>%
  mutate(broad_hair_color = case_when(
    hair %in% c("blond", "light blond") ~ "blond",
    hair %in% c("brown", "light brown") ~ "brown",
    hair %in% c("grey", "other_low_count_category1", "other_low_count_category2") ~ '
    TRUE ~ as.character(hair)  # Keep other categories as is
  ))

# we want to see a new count of our distribution of hair color, count the number of c
hair_color_counts_broad <- data %>%
  group_by(broad_hair_color) %>%
  summarize(observations = n())
```

```
#display counts
hair_color_counts_broad
```

broad_hair_color	observations
<chr>	<int>
black	2196
blond	588
brown	3636
other	2
red	156

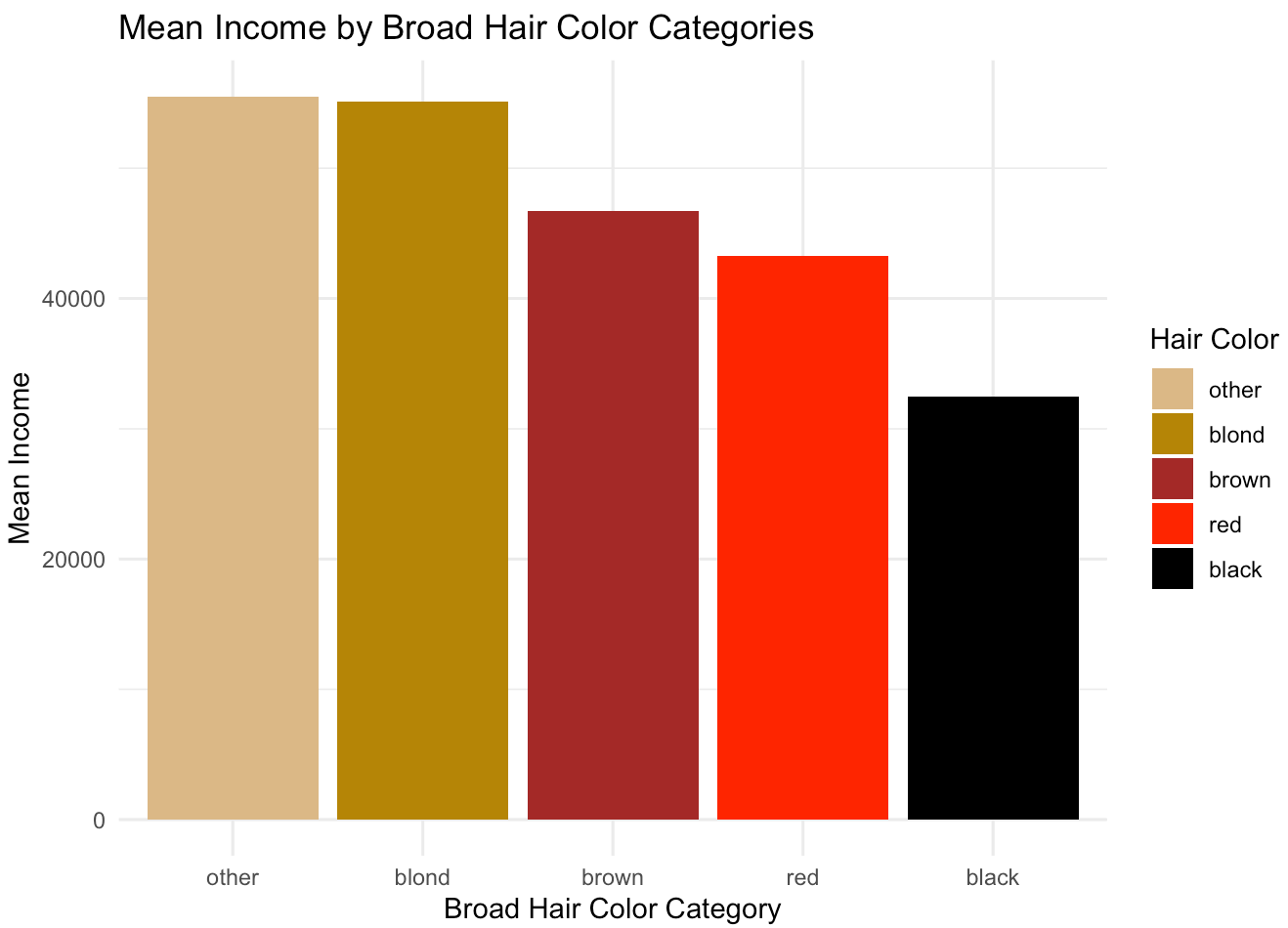
5 rows

```
#our observations are now updated but now we see black and brown consisting of the ma
```

```
# Calculate the mean income for each broad hair color category
mean_income_by_color <- data %>%
  group_by(broad_hair_color) %>%
  summarize(mean_income = mean(income, na.rm = TRUE))  %>%
  mutate(broad_hair_color = reorder(broad_hair_color, -mean_income))

# Define custom colors for each hair color category
hair_color_colors <- c(
  "other" = "burlywood",
  "blond" = "darkgoldenrod",
  "brown" = "brown",
  "black" = "black",
  "red" = "red"
)

# Create a bar plot for mean income by broad hair color
p <- ggplot(mean_income_by_color,
       aes(x = broad_hair_color, y = mean_income, fill = broad_hair_color)) +
  geom_bar(stat = "identity") +
  labs(title = "Mean Income by Broad Hair Color Categories", x = "Broad Hair Color Ca
  scale_fill_manual(values = hair_color_colors) +  # Use custom colors
  theme_minimal() + labs(fill = "Hair Color")  # Rename the legend

ggsave("mean_income_by_color.png", plot = p)
```

```
Saving 7 x 5 in image
```

```
p
```



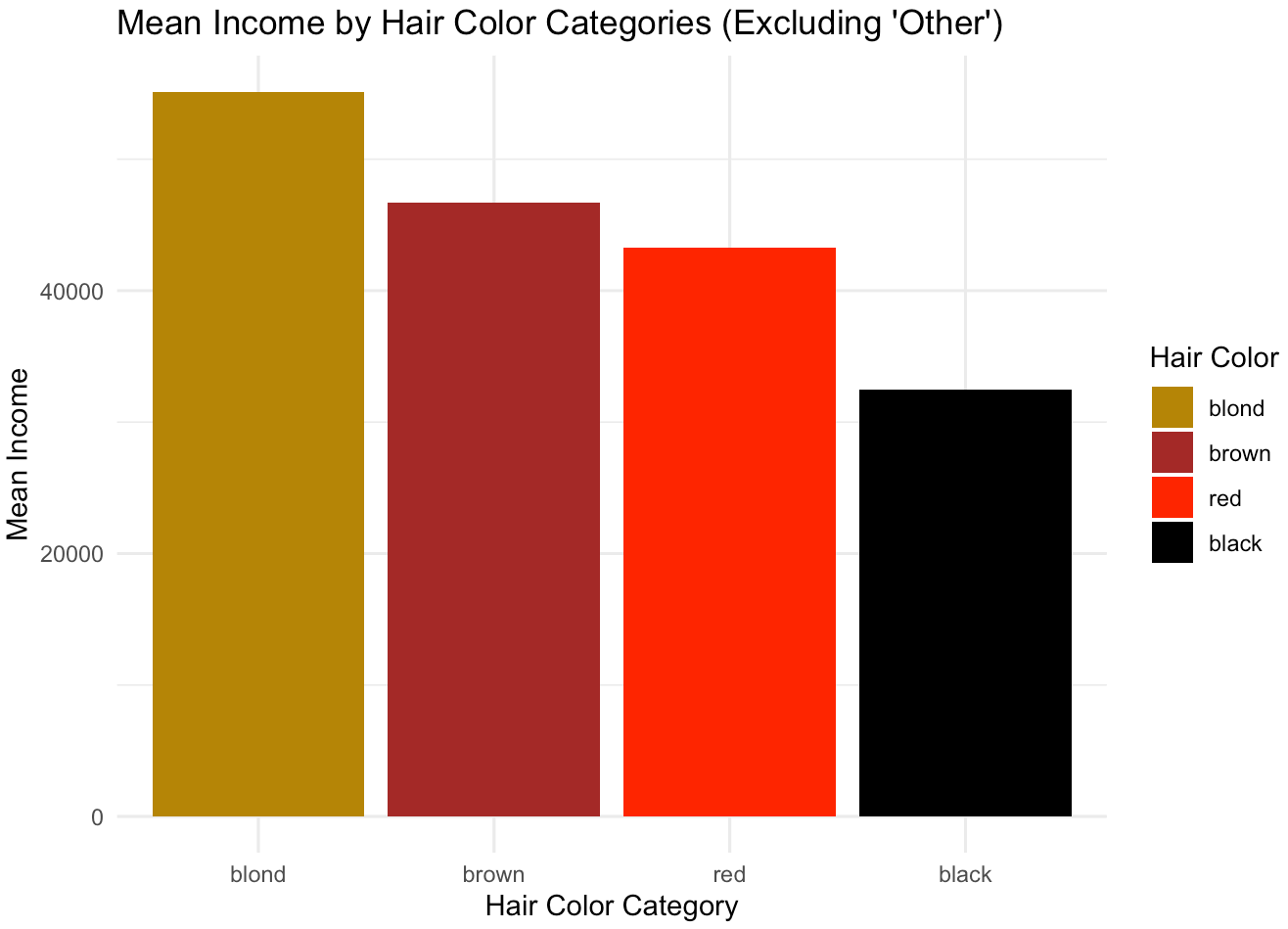
We examine the distribution of hair colors in the dataset and find that the hair colors appear imbalanced. Black and brown have the most observations, but we are limited in variety for the colors red, grey, blond, and light blond. These results will ultimately create an imbalance and affect the statistical significance and reliability when analyzing the relationship between hair color and income.

Mean Income by Broad Hair Color Categories

```
#the other category looks misleading so we should filter this out
# Calculate the mean income for each broad hair color category, excluding "other"
mean_income_by_color_exclude_other <- data %>%
  filter(broad_hair_color != "other") %>%
  group_by(broad_hair_color) %>%
  summarize(mean_income = mean(income, na.rm = TRUE))

# Reorder the broad hair color categories by mean income in descending order
mean_income_by_color_exclude_other <- mean_income_by_color_exclude_other %>%
  mutate(broad_hair_color = reorder(broad_hair_color, -mean_income))

# Create a bar plot for mean income by broad hair color with custom colors
p <- ggplot(mean_income_by_color_exclude_other, aes(x = broad_hair_color, y = mean_ir
  geom_bar(stat = "identity") +
  labs(title = "Mean Income by Hair Color Categories (Excluding 'Other')", x = "Hair
  scale_fill_manual(values = hair_color_colors) +  # Use custom colors
 theme_minimal() + labs(fill = "Hair Color")  # Rename the legend

ggsave("mean_income_by_color_exclude_other.png", plot = p)
```

```
Saving 7 x 5 in image
```

```
p
```




```
# Find the median income for the black hair color
reference_median <- mean_income_by_color_exclude_other %>%
  filter(broad_hair_color == "black") %>%
  pull(mean_income)

# Calculate the percentage difference compared to black category
mean_income_by_color_exclude_other <- mean_income_by_color_exclude_other %>%
  mutate(percentage_difference = ((mean_income - reference_median) / reference_median

# Sort the data in ascending order of percentage difference
  arrange(percentage_difference)

mean_income_by_color_exclude_other
```

broad_hair_color	mean_income	percentage_difference
<fct>	<dbl>	<dbl>
black	32471.66	0.00000
red	43301.60	33.35195
brown	46683.38	43.76650
blond	55116.43	69.73700

4 rows

To compensate for this, we grouped our hair colors into broader categories, such as combining different shades of blond and brown while preserving other types as they were. We acknowledge the small count in the “other” category. We will make further adjustments, setting the stage for further analysis of income across these newly defined hair color groups.

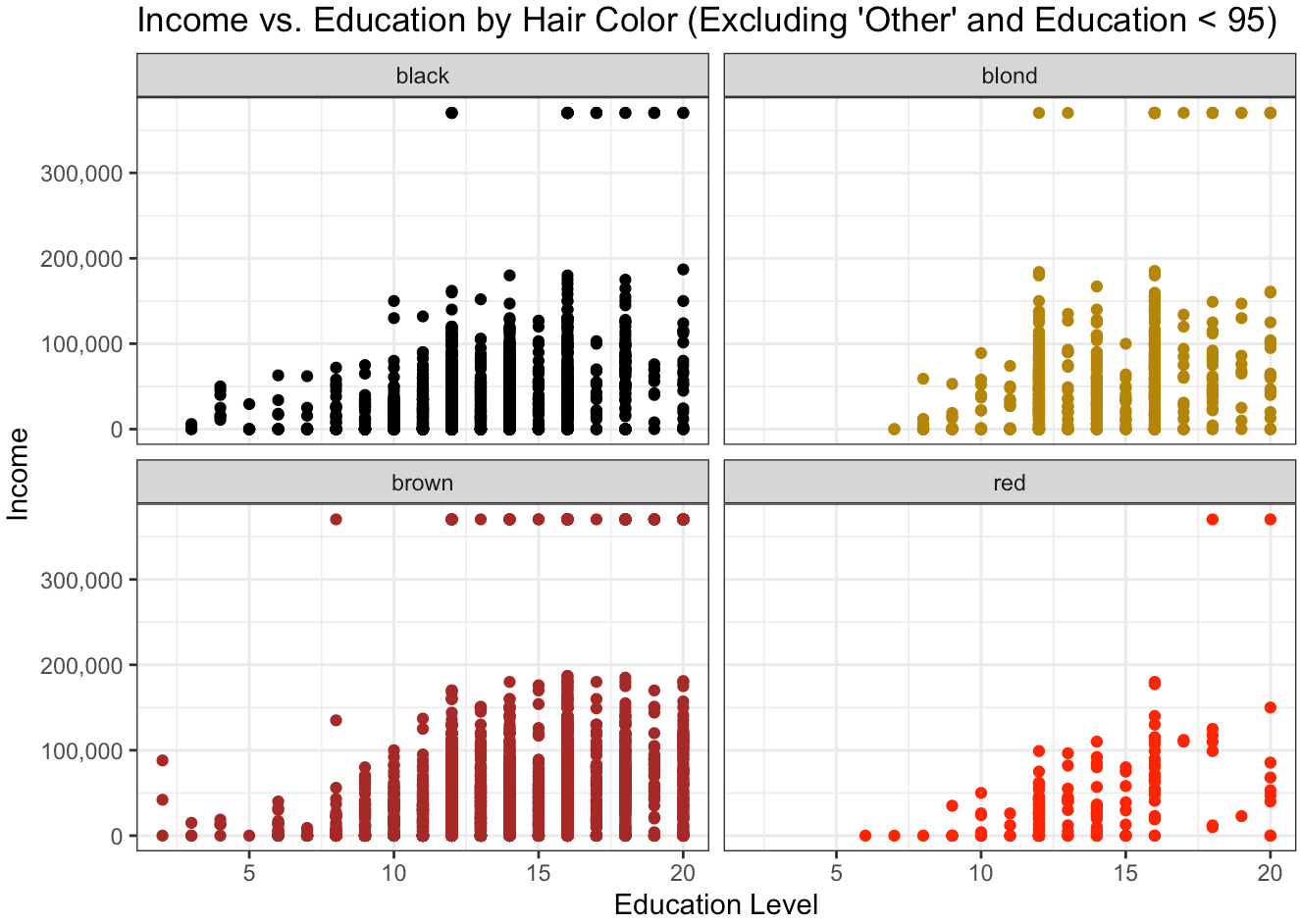
To begin analyzing this information, we calculate the average income for the different hair color categories we are looking at. We create a bar plot to visualize the mean income. As we mentioned earlier, the “other” category was included in our dataset, and when it is graphed, it appears misleading. We created a new plot and excluded the “other” classification. We recalculated the mean income for the remaining colors. Based on the information in our dataset, we saw that people with blond hair, on average, have a higher mean income than those with black, brown, or red hair.

Combined Effects of Education and Hair on Income

```
# Create a faceted scatterplot of income vs. education
p <- ggplot(data %>% filter(broad_hair_color != "other",  !is.na(income)),
       aes(x = education, y = income)) +
  geom_point(aes(color = broad_hair_color)) +
  facet_wrap(~broad_hair_color) +
  labs(x = "Education Level", y = "Income", title = "Income vs. Education by Hair Col
  scale_color_manual(values = hair_color_colors) +
  theme_bw() +
  guides(color = "none") +  # Remove the color legend
  scale_y_continuous(labels = scales::comma)

ggsave("Income_vs_Education_by_HairColor.png", plot = p)
```

```
Saving 7 x 5 in image
```

```
p
```



In this analysis, we created a faceted scatterplot to explore the relationship between income and education. We see how income and education levels vary across different hair color groups. We exclude the “other” hair color category to enhance clarity. While the scatter plot reveals that individuals with blond hair, on average, tend to have lower education completion than those with black or brown hair, they consistently have a higher mean income. Additionally, it’s worth noting that higher education levels generally correlate with higher incomes across all hair color categories. Further exploration of this data could unveil more intriguing patterns and insights.

Combined Effects of Gender and Hair on Income

```
# Filter the data to exclude "other"
filtered_data <- data %>%
  filter(broad_hair_color != "other")

# Group the filtered data by broad_hair_color and gender, and calculate mean income
income_by_color_gender <- filtered_data %>%
  group_by(broad_hair_color, sex) %>%
  summarize(mean_income = mean(income, na.rm = TRUE))
```

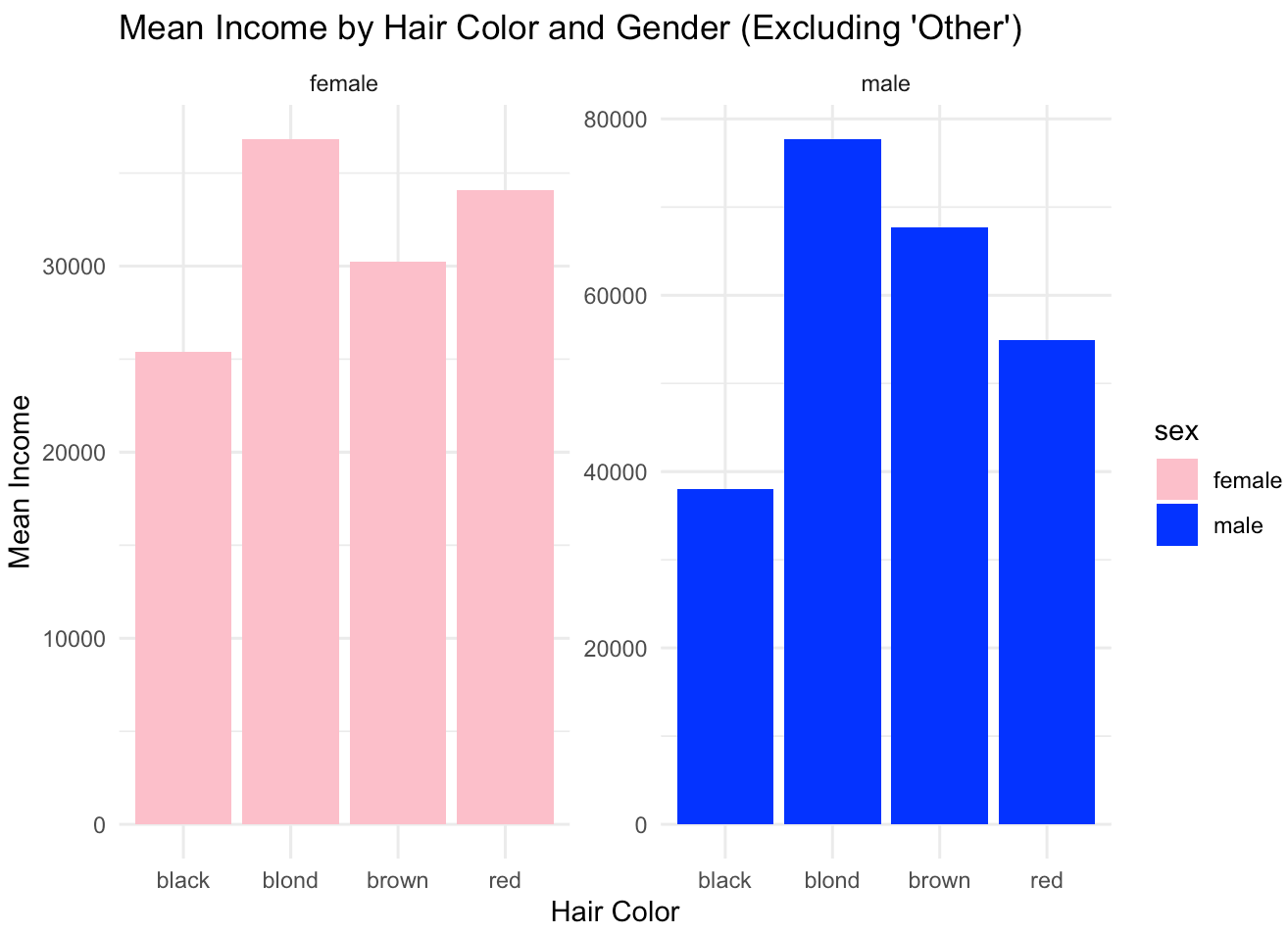
```
`summarise()` has grouped output by 'broad_hair_color'. You can override using
the `.groups` argument.
```

```
# Define custom colors for gender
sex_color <- c(
  "male" = "blue",
```

```
  "female" = "pink"
)

# Create a faceted bar plot for mean income by hair color and gender
p <- ggplot(income_by_color_gender, aes(x = broad_hair_color, y = mean_income, fill =
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Hair Color", y = "Mean Income", title = "Mean Income by Hair Color and Ge
  scale_fill_manual(values = sex_color) +  # Custom colors
  theme_minimal() +
  facet_wrap(~sex, scales = "free_y")  # Facet by gender with independent y-axes
ggsave("income_by_color_gender.png", plot = p)
```

```
Saving 7 x 5 in image
```

```
p
```




```
# Group the data by broad_hair_color, sex, and calculate mean income
mean_income_by_color_gender <- data %>%
  group_by(broad_hair_color, sex) %>%
  summarize(mean_income = mean(income, na.rm = TRUE))
```

```
`summarise()` has grouped output by 'broad_hair_color'. You can override using
the `.groups` argument.
```

```
# Check the summary
```

```
mean_income_by_color_gender
```

broad_hair_color	sex	mean_income
<chr>	<chr>	<dbl>
black	female	25402.55
black	male	38064.69
blond	female	36829.21
blond	male	77714.70
brown	female	30226.17
brown	male	67718.90
other	male	55500.00
red	female	34074.34
red	male	54935.96

9 rows

```
# Calculate the mean income for each combination of broad_hair_color and sex
mean_income_by_color_gender <- data %>%
  group_by(broad_hair_color, sex) %>%
  summarize(mean_income = mean(income, na.rm = TRUE))
```

```
`summarise()` has grouped output by 'broad_hair_color'. You can override using
the `.groups` argument.
```

```
# Calculate the percentage difference
percentage_difference <- mean_income_by_color_gender %>%
  pivot_wider(names_from = sex, values_from = mean_income) %>%
  mutate(percentage_difference = ((male - female) / female) * 100)

# View the resulting data frame
percentage_difference
```

broad_hair_color	female	male	percentage_difference
<chr>	<dbl>	<dbl>	<dbl>
black	25402.55	38064.69	49.84595
blond	36829.21	77714.70	111.01377
brown	30226.17	67718.90	124.04063
other	*NA*	55500.00	*NA*
red	34074.34	54935.96	61.22381

5 rows

We created a faceted plot by filtering the hair color data to exclude the “other” category in the dataset. We then grouped the variables “broad_hair_color” and “sex” to calculate the mean income for each resulting group. Upon reviewing the graph, it is evident that there are gender-based income disparities. In most hair color categories, males tend to have higher mean incomes than females. To provide a more detailed understanding, we present the calculations in terms of percentage variations in the subsequent tables, illustrating these differences.

Grouped Analysis of Summary Statistics

Mean and Median Income by Gender

```
data %>%
  group_by(sex) %>% summarise(Mean=mean(income, na.rm=T),
                              Median=median(income, na.rm=T))
```

sex	Mean	Median
<chr>	<dbl>	<dbl>
female	29583.82	21310
male	56745.00	40000

2 rows

Males seem to have higher average and median income compared with females.

Mean and Median Income by Hair

```
data %>%
  group_by(hair) %>% summarise(Mean=mean(income, na.rm=T),
                              Median=median(income, na.rm=T))
```

hair	Mean	Median
<chr>	<dbl>	<dbl>
black	32471.66	21000
blond	54444.28	36000
brown	46643.14	31000
grey	55500.00	55500
light blond	62044.69	34500
light brown	46848.35	33000
red	43301.60	28000

7 rows

Respondents with light blond color seem to have the highest average and median income. Respondents with black color seem to have the lowest average and median income.

Mean and Median Income by Hair and Gender

```
data %>%
  group_by(sex, hair) %>% summarise(Mean=mean(income, na.rm=T),
                              Median=median(income, na.rm=T))
```

```
`summarise()` has grouped output by 'sex'. You can override using the `.groups`
argument.
```

sex	hair	Mean	Median
<chr>	<chr>	<dbl>	<dbl>
female	black	25402.55	18000
female	blond	37290.65	23000
female	brown	29726.75	22000
female	light blond	32291.67	13000
female	light brown	32042.25	23194
female	red	34074.34	22800
male	black	38064.69	25000
male	blond	75441.46	52000
male	brown	67101.35	47000
male	grey	55500.00	55500

1-10 of 13 rows Previous **1** 2 Next

Proportion With Income More Than 200000 by Race

```
data %>%
  group_by(sex) %>% summarise(Proportion=mean(income>200000, na.rm=T))
```

sex	Proportion
<chr>	<dbl>
female	0.002630041
male	0.041825095

2 rows

The proportion of income over $200,000 is higher for males than females.

Proportion With Income More Than 200000 by Hair Color

```
data %>%
  group_by(hair) %>% summarise(Proportion=mean(income>200000, na.rm=T))
```

hair	Proportion
<chr>	<dbl>
black	0.01001821
blond	0.03544776
brown	0.02702703
grey	0.00000000
light blond	0.05769231
light brown	0.02244039
red	0.01282051

7 rows

The respondent with light blond color has the highest proportion of income, more than $200,000.

Hypotheses for Further Analysis

Hypothesis 1: Education and Income

We hypothesize that there is a positive relationship between education level and income. In our analysis, we saw that individuals with higher levels of education tended to have higher incomes. We expect that as years of education increase, so does the average income.

Hypothesis 2: Hair Color and Income

Based on our analysis of hair color and income, we hypothesize that certain hair colors may be associated with higher incomes when compared to others. Specifically, our data suggests that individuals with blond hair tend to have a higher mean and median income than those with black, brown, or red hair. This hypothesis may require further investigation to test its statistical significance.

Hypothesis 3: Gender-Based Income

We propose that gender-based income disparities exist in our dataset. Specifically, we hypothesize that males, on average, have higher incomes than females. Our hypothesis is drawn from our exploratory data analysis, which reveals that the mean and median income for males is consistently higher than that of females across various education levels.