Final project starter script

```r
library(tidyverse)
```

Package loading

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- 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 errors
```

```r
library(knitr)
library(ggplot2)
library(dplyr)
library(effsize)
```

```
## Warning: package 'effsize' was built under R version 4.4.2
```

```r
# Import starting data
nlsy <- read_csv("nlsy97.csv")
```

Importing the data

```
## Rows: 8984 Columns: 95
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (95): B0004600, E8043100, E8043200, E8043400, R0000100, R0069400, R00700...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Variables present in the base data set To learn more about the data, you can have a look at the variable codebook file available on Canvas.

Here's how to rename all the variables to the Question Name abbreviation. **You will want to change the names to be even more descriptive**, but this is a start.

```r
# Change column names to question name abbreviations (you will want to change these further)
colnames(nlsy) <- c("PSTRAN_GPA.01_PSTR",
    "INCARC_TOTNUM_XRND",
    "INCARC_AGE_FIRST_XRND",
    "INCARC_LENGTH_LONGEST_XRND",
    "PUBID_1997",
    "YSCH-36400_1997",
    "YSCH-37000_1997",
    "YSAQ-010_1997",
    "YSAQ-369_1997",
    "YEXP-300_1997",
    "YEXP-1500_1997",
    "YEXP-1600_1997",
    "YEXP-1800_1997",
    "YEXP-2000_1997",
    "sex",
    "KEY_BDATE_M_1997",
    "KEY_BDATE_Y_1997",
    "PC8-090_1997",
    "PC8-092_1997",
    "PC9-002_1997",
    "PC12-024_1997",
    "PC12-028_1997",
    "CV_AGE_12/31/96_1997",
    "CV_BIO_MOM_AGE_CHILD1_1997",
    "CV_BIO_MOM_AGE_YOUTH_1997",
    "CV_CITIZENSHIP_1997",
    "CV_ENROLLSTAT_1997",
    "CV_HH_NET_WORTH_P_1997",
    "CV_YTH_REL_HH_CURRENT_1997",
    "CV_MSA_AGE_12_1997",
    "CV_URBAN-RURAL_AGE_12_1997",
    "CV_SAMPLE_TYPE_1997",
    "CV_HGC_BIO_DAD_1997",
    "CV_HGC_BIO_MOM_1997",
    "CV_HGC_RES_DAD_1997",
    "CV_HGC_RES_MOM_1997",
    "race",
    "YSCH-6800_1998",
    "YSCH-7300_1998",
    "YSAQ-372B_1998",
    "YSAQ-371_2000",
    "YSAQ-282J_2002",
    "YSAQ-282Q_2002",
    "CV_HH_NET_WORTH_Y_2003",
    "CV_BA_CREDITS.01_2004",
    "YSAQ-000B_2004",
    "YSAQ-373_2004",
    "YSAQ-369_2005",
    "CV_BIO_CHILD_HH_2007",
    "YTEL-52~000001_2007",
    "YTEL-52~000002_2007",
    "YTEL-52~000003_2007",
```

```r
    "YTEL-52~000004_2007",
    "CV_BIO_CHILD_HH_2009",
    "CV_COLLEGE_TYPE.01_2011",
    "CV_INCOME_FAMILY_2011",
    "CV_HH_SIZE_2011",
    "CV_HH_UNDER_18_2011",
    "CV_HH_UNDER_6_2011",
    "CV_HIGHEST_DEGREE_1112_2011",
    "CV_BIO_CHILD_HH_2011",
    "YSCH-3112_2011",
    "YSAQ-000A000001_2011",
    "YSAQ-000A000002_2011",
    "YSAQ-000B_2011",
    "YSAQ-360C_2011",
    "YSAQ-364D_2011",
    "YSAQ-371_2011",
    "YSAQ-372CC_2011",
    "YSAQ-373_2011",
    "YSAQ-374_2011",
    "YEMP_INDCODE-2002.01_2011",
    "CV_BIO_CHILD_HH_2015",
    "YEMP_INDCODE-2002.01_2017",
    "YEMP_OCCODE-2002.01_2017",
    "CV_MARSTAT_COLLAPSED_2017",
    "YINC-1400_2017",
    "income",
    "YINC-1800_2017",
    "YINC-2400_2017",
    "YINC-2600_2017",
    "YINC-2700_2017",
    "CVC_YTH_REL_HH_AGE6_YCHR_XRND",
    "CVC_SAT_MATH_SCORE_2007_XRND",
    "CVC_SAT_VERBAL_SCORE_2007_XRND",
    "CVC_ACT_SCORE_2007_XRND",
    "CVC_HH_NET_WORTH_20_XRND",
    "CVC_HH_NET_WORTH_25_XRND",
    "CVC_ASSETS_FINANCIAL_25_XRND",
    "CVC_ASSETS_DEBTS_20_XRND",
    "CVC_HH_NET_WORTH_30_XRND",
    "CVC_HOUSE_VALUE_30_XRND",
    "CVC_HOUSE_TYPE_30_XRND",
    "CVC_ASSETS_FINANCIAL_30_XRND",
    "CVC_ASSETS_DEBTS_30_XRND")

### Set all negative values to NA.
### THIS IS DONE ONLY FOR ILLUSTRATIVE PURPOSES
### DO NOT TAKE THIS APPROACH WITHOUT CAREFUL JUSTIFICATION
nlsy[nlsy < 0]  <- NA
```

A note on missing values Here's an example of what the variable description files look like

```
T76400.00    [YSAQ-372CC]                                  Survey Year: 2011
  PRIMARY VARIABLE
```

```
             HAS R USED COCAINE/HARD DRUGS SINCE DLI?

Excluding marijuana and alcohol, since the date of last interview, have you used
any drugs like cocaine, crack, heroin, or crystal meth, or any other substance
not prescribed by a doctor, in order to get high or to achieve an altered state?

UNIVERSE: All except prisoners in an insecure environment

     215       1 YES   (Go To T76401.00)
    7023       0 NO
  -------
    7238

Refusal(-1)           74
Don't Know(-2)        26
TOTAL =========>    7338   VALID SKIP(-4)      85     NON-INTERVIEW(-5)    1561

Min:              0        Max:              1        Mean:                 .03

Lead In: T76397.00[Default] T76399.00[Default]  T76398.00[0:0]
Default Next Question: T76403.00
```

This description says that the numbers -1, -2, -4 and -5 all have a special meaning for this variable. They denote different types of missingness. You can recode all of these to `NA`, but you should also think about whether the different missingness indicators are in some way informative. (i.e., if someone refuses to answer questions related to drug use, might this inform us about their income?)

Getting to know our two main variables. In the previous chunk of code we have appropriately renamed the variables corresponding to `sex`, `race` and `income` (as reported on the 2017 survey). Let's have a quick look at what we're working with.

```r
table(nlsy$sex)
```

```
##
##    1    2
## 4599 4385
```

```r
table(nlsy$race)
```

```
##
##    1    2    3    4
## 2335 1901   83 4665
```

The data codebook tells us that the coding for sex is `Male = 1`, `Female = 2`. For the race/ethnicity variable, the coding is:

```
1 Black
2 Hispanic
3 Mixed Race (Non-Hispanic)
4 Non-Black / Non-Hispanic
```

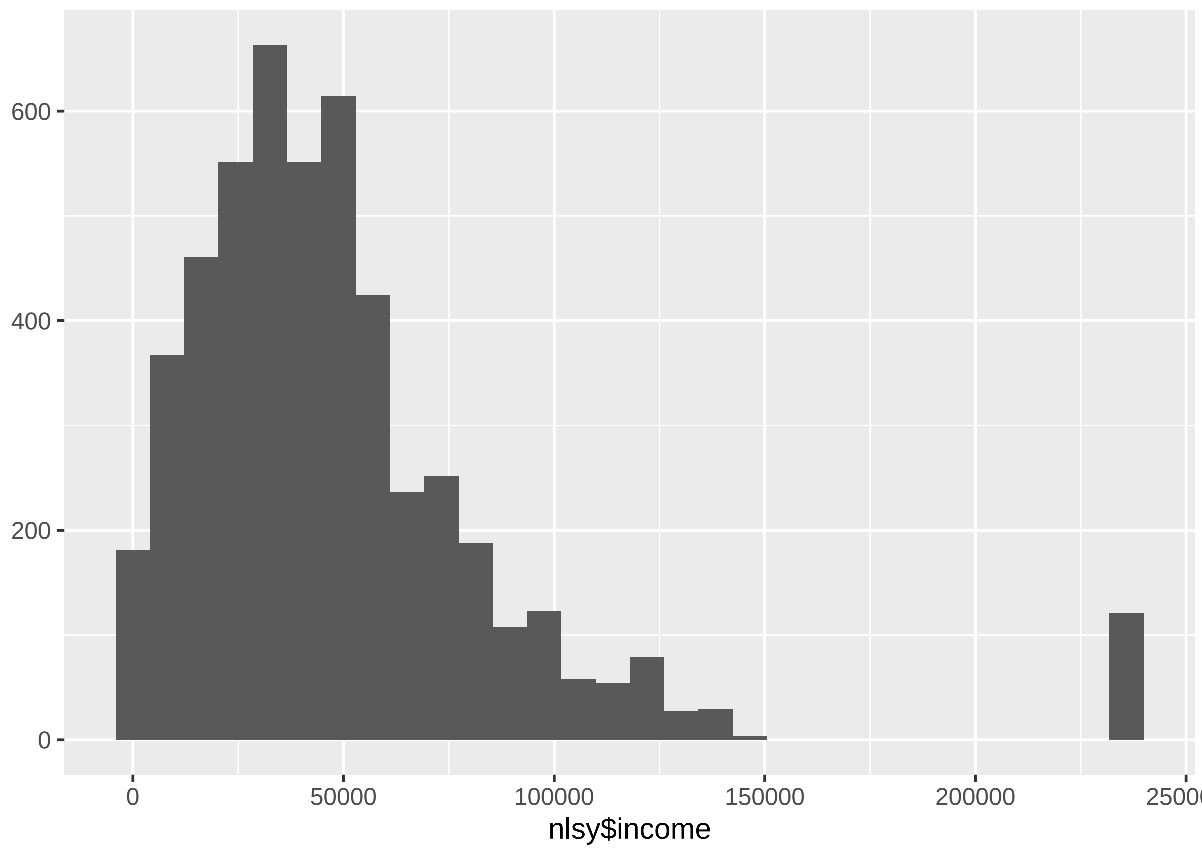
You'll want to do some data manipulations to change away from the numeric codings to more interpretable labels.

```r
summary(nlsy$income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0   25000   40000   49477   62000  235884    3893
```

```r
# Histogram
qplot(nlsy$income)
```

```
## Warning: `qplot()` was deprecated in ggplot2 3.4.0.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.

## Warning: Removed 3893 rows containing non-finite outside the scale range
## (`stat_bin()`).
```



The income distributing is right-skewed like one might expect. However, as indicated in the question description, the income variable is *topcoded* at the 2% level. More precisely,

```r
n.topcoded <- with(nlsy, sum(income == max(income, na.rm = TRUE), na.rm = TRUE))
n.topcoded
```

```
## [1] 121
```

121 of the incomes are topcoded to the maximum value of 2.35884×10^5, which is the average value of the top 121 earners. You will want to think about how to deal with this in your analysis.

Significant Difference in Income between Men and Women

```r
# Rename and clean data
nlsy <- nlsy %>%
  rename(
    Gender = sex,
    Income = income
  ) %>%
  mutate(
    Gender = factor(Gender, levels = c(1, 2), labels = c("Male", "Female")),
    Income = ifelse(Income < 0, NA, Income)
  )

# Create multiple visualizations for better insight
# 1. Density plot with summary statistics
p1 <- ggplot(nlsy, aes(x = Income, fill = Gender)) +
  geom_density(alpha = 0.5) +
  geom_vline(data = nlsy %>%
               group_by(Gender) %>%
               summarise(median = median(Income, na.rm = TRUE)),
             aes(xintercept = median, color = Gender),
             linetype = "dashed", size = 1) +
  scale_x_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  labs(
    title = "Income Distribution by Gender",
    subtitle = "Dashed lines represent median income",
    x = "Annual Income",
    y = "Density"
  ) +
  theme_minimal()
```

```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.
```

```r
# 2. Box plot with violin plot overlay
p2 <- ggplot(nlsy, aes(x = Gender, y = Income, fill = Gender)) +
  geom_violin(alpha = 0.5) +
  geom_boxplot(width = 0.2, alpha = 0.8) +
  coord_cartesian(ylim = c(0, 150000)) +
```

```r
  scale_y_continuous(labels = scales::dollar_format()) +
  labs(
    title = "Income Distribution Details by Gender",
    subtitle = "Violin plot shows distribution shape, box plot shows quartiles",
    x = "Gender",
    y = "Annual Income"
  ) +
  theme_minimal() +
  theme(legend.position = "none")

# 3. Income brackets analysis
p3 <- nlsy %>%
  mutate(Income_Bracket = cut(Income,
                              breaks = c(0, 25000, 50000, 75000, 100000, Inf),
                              labels = c("0-25k", "25k-50k", "50k-75k", "75k-100k", "100k+"),
                              include.lowest = TRUE)) %>%
  ggplot(aes(x = Income_Bracket, fill = Gender)) +
  geom_bar(position = "dodge") +
  scale_y_continuous(labels = scales::comma) +
  labs(
    title = "Income Brackets by Gender",
    subtitle = "Number of individuals in each income range",
    x = "Income Bracket",
    y = "Count"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))

# Print summary statistics
gender_summary <- nlsy %>%
  group_by(Gender) %>%
  summarise(
    Mean = mean(Income, na.rm = TRUE),
    Median = median(Income, na.rm = TRUE),
    SD = sd(Income, na.rm = TRUE),
    Q1 = quantile(Income, 0.25, na.rm = TRUE),
    Q3 = quantile(Income, 0.75, na.rm = TRUE),
    n = sum(!is.na(Income))
  ) %>%
  mutate(across(Mean:Q3, ~scales::dollar(.x, accuracy = 1)))

print("Income Summary Statistics by Gender:")
```

```
## [1] "Income Summary Statistics by Gender:"
```

```r
print(gender_summary)
```

```
## # A tibble: 2 x 7
##   Gender Mean    Median  SD      Q1      Q3          n
##   <fct>  <chr>   <chr>   <chr>   <chr>   <chr>   <int>
## 1 Male   $57,203 $47,000 $44,712 $30,000 $70,000  2621
## 2 Female $41,279 $35,000 $34,047 $20,000 $52,000  2470
```

```r
# Arrange plots in a grid
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 4.4.2
```

```
##
## Attaching package: 'gridExtra'
```

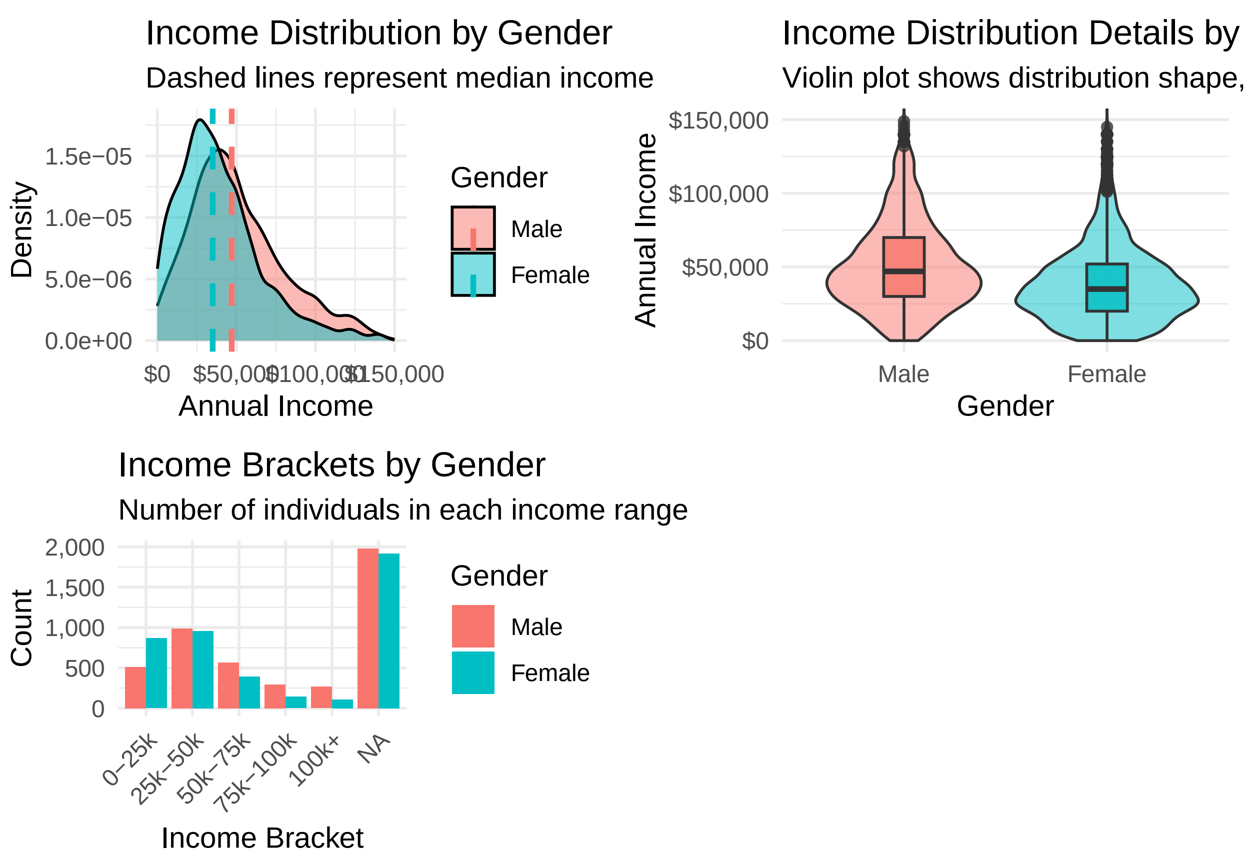
```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```r
grid.arrange(p1, p2, p3, ncol = 2, nrow = 2)
```

```
## Warning: Removed 4014 rows containing non-finite outside the scale range
## (`stat_density()`).
```

```
## Warning: Removed 3893 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
```

```
## Warning: Removed 3893 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```



```r
# Statistical test
t_test_result <- t.test(Income ~ Gender, data = nlsy, var.equal = FALSE)
print("\nStatistical Test Results:")
```

```
## [1] "\nStatistical Test Results:"
```

```r
print(t_test_result)
```

```
##
##  Welch Two Sample t-test
##
## data:  Income by Gender
## t = 14.346, df = 4876.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Male and group Female is not equal to
## 95 percent confidence interval:
##  13747.84 18099.96
## sample estimates:
##   mean in group Male mean in group Female
##             57202.82             41278.92
```

```r
# Calculate and print gender pay gap
pay_gap <- nlsy %>%
  group_by(Gender) %>%
  summarise(mean_income = mean(Income, na.rm = TRUE)) %>%
  spread(Gender, mean_income) %>%
  mutate(gap_percent = (Male - Female) / Male * 100)

print("\nGender Pay Gap:")
```

```
## [1] "\nGender Pay Gap:"
```

```r
print(paste0("Women earn ", round(pay_gap$gap_percent, 1),
            "% less than men on average in this sample"))
```

```
## [1] "Women earn 27.8% less than men on average in this sample"
```

Factors Affecting Difference of Income between Men and Women

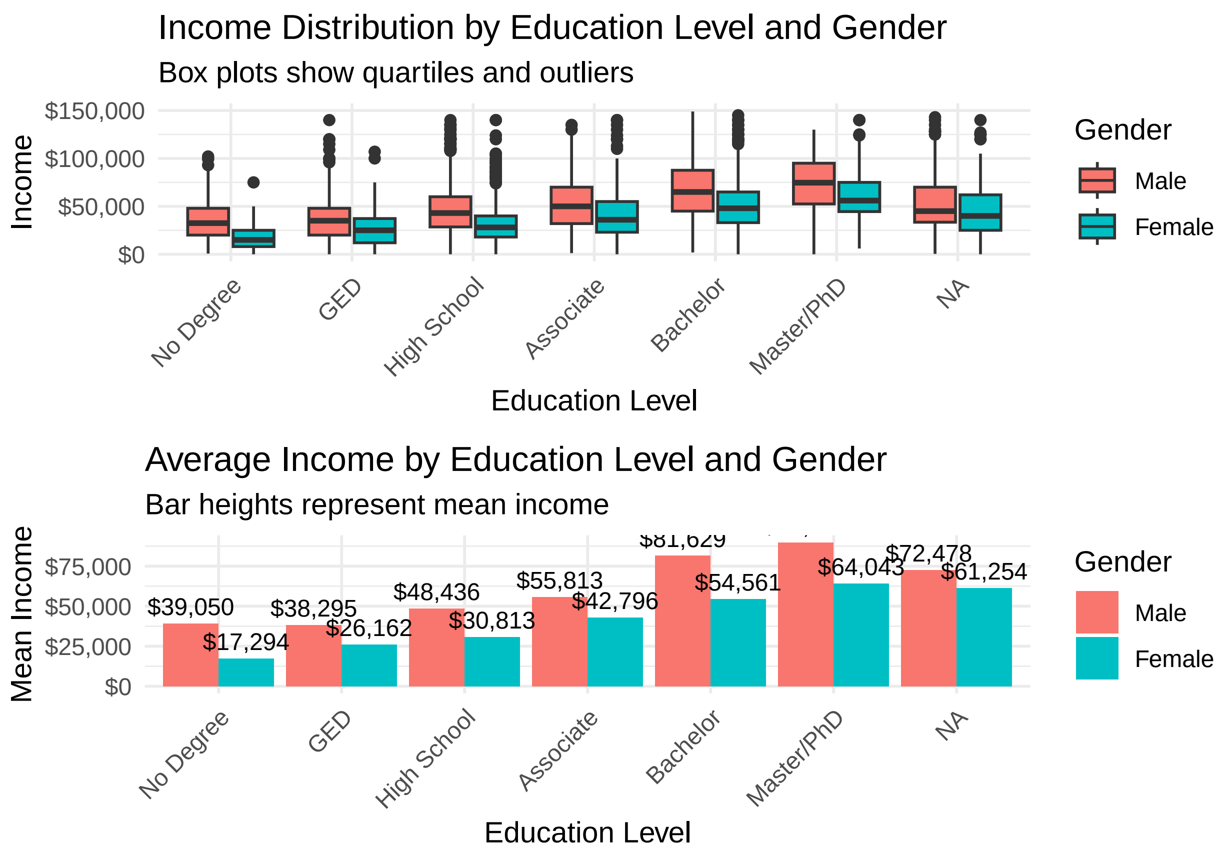
Factors that we are testing: - Parents education - Drug Use - Education - Martial status - Criminal history - Profession - Work Experience - Region (Urban/Rural) - Children - Age - Ethnicity - Health Status

```r
# Rename and clean relevant variables
nlsy <- nlsy %>%
  rename(
    Education_Level = CV_HIGHEST_DEGREE_1112_2011,
    Marital_Status = CV_MARSTAT_COLLAPSED_2017,
    Criminal_Record = INCARC_TOTNUM_XRND,
    Drug_Use = `YSAQ-372B_1998`,
    Parent_Education_Dad = CV_HGC_BIO_DAD_1997,
    Parent_Education_Mom = CV_HGC_BIO_MOM_1997,
```

```r
    Urban_Rural = `CV_URBAN-RURAL_AGE_12_1997`,
    Children_HH = CV_BIO_CHILD_HH_2015,
    Ethnicity = race,
    Work_Experience = `YEXP-1800_1997`
  )

# Recode categorical variables
nlsy <- nlsy %>%
  mutate(
    Education_Level = factor(Education_Level,
      levels = 0:5,
      labels = c("No Degree", "GED", "High School", "Associate", "Bachelor", "Master/PhD")
    ),
    Marital_Status = factor(Marital_Status,
      levels = 1:4,
      labels = c("Never Married", "Married", "Separated/Divorced", "Widowed")
    ),
    Ethnicity = factor(Ethnicity,
      levels = 1:4,
      labels = c("Black", "Hispanic", "Mixed Race", "White/Other")
    ),
    Urban_Rural = factor(Urban_Rural,
      levels = 0:1,
      labels = c("Rural", "Urban")
    ),
    Has_Criminal_Record = factor(ifelse(Criminal_Record > 0, "Yes", "No")),
    Has_Children = factor(ifelse(Children_HH > 0, "Yes", "No"))
  )

# 1. Education Analysis
# ---------------------
education_analysis <- nlsy %>%
  group_by(Gender, Education_Level) %>%
  summarise(
    Mean_Income = mean(Income, na.rm = TRUE),
    Median_Income = median(Income, na.rm = TRUE),
    Count = n(),
    .groups = 'drop'
  )

# Create two complementary visualizations for education
p1_education <- ggplot(nlsy, aes(x = Education_Level, y = Income, fill = Gender)) +
  geom_boxplot() +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(
    title = "Income Distribution by Education Level and Gender",
    subtitle = "Box plots show quartiles and outliers",
    x = "Education Level",
    y = "Income"
  )
```

```r
p2_education <- ggplot(education_analysis,
                       aes(x = Education_Level, y = Mean_Income, fill = Gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = scales::dollar(Mean_Income, accuracy = 1)),
            position = position_dodge(width = 0.9),
            vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::dollar_format()) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(
    title = "Average Income by Education Level and Gender",
    subtitle = "Bar heights represent mean income",
    x = "Education Level",
    y = "Mean Income"
  )

grid.arrange(p1_education, p2_education, ncol = 1)
```

```
## Warning: Removed 4014 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```




```r
# Print education summary
print("Education Level Analysis:")
```

```
## [1] "Education Level Analysis:"
```

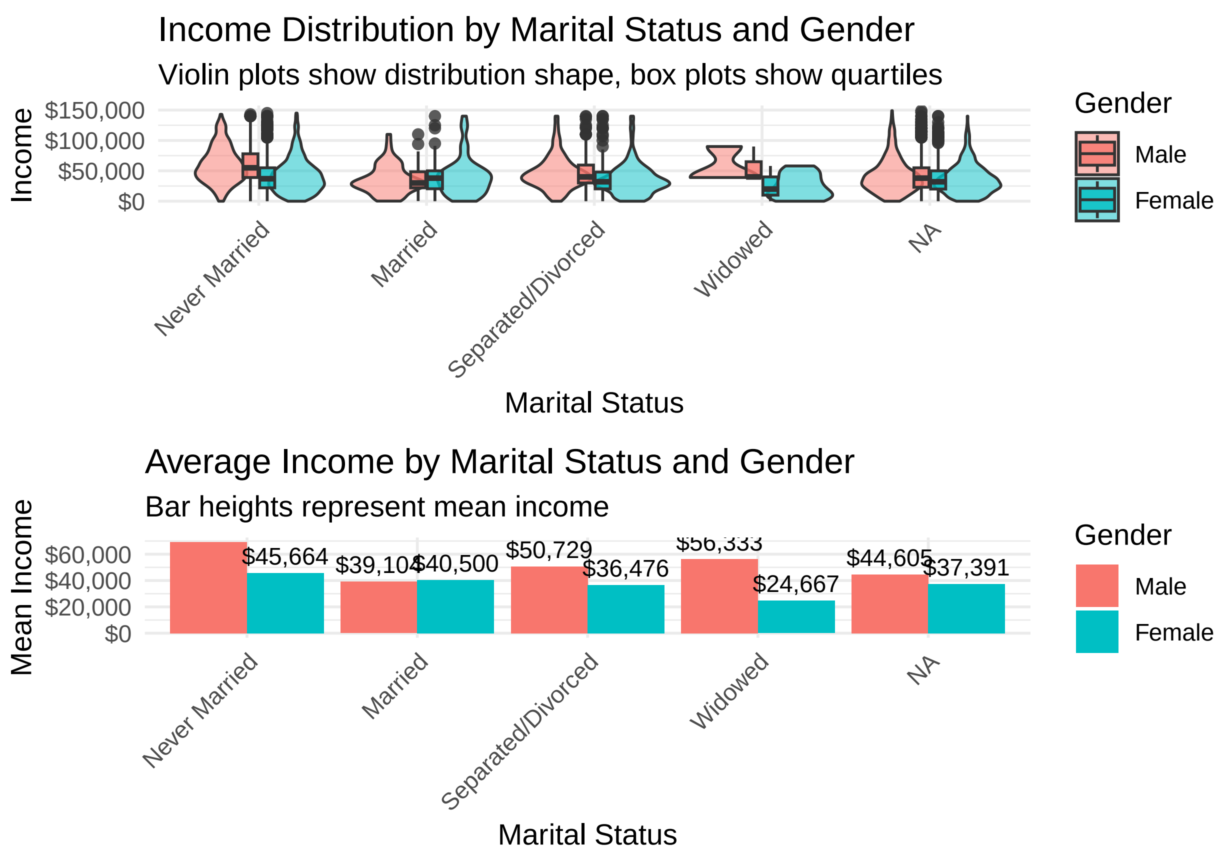
```r
print(education_analysis)
```

```
## # A tibble: 14 x 5
##    Gender Education_Level Mean_Income Median_Income Count
##    <fct>  <fct>                 <dbl>         <dbl> <int>
##  1 Male   No Degree            39050.         33800   413
##  2 Male   GED                  38295.         35000   544
##  3 Male   High School          48436.         44000  1721
##  4 Male   Associate            55813.         50000   232
##  5 Male   Bachelor             81629.         66000   641
##  6 Male   Master/PhD           89771.         85000   124
##  7 Male   <NA>                 72478.         50000   924
##  8 Female No Degree            17294.         15000   335
##  9 Female GED                  26162.         25000   373
## 10 Female High School          30813.         28000  1555
## 11 Female Associate            42796.         36000   300
## 12 Female Bachelor             54561.         49000   822
## 13 Female Master/PhD           64043.         57500   217
## 14 Female <NA>                 61254.         45000   783
```

```r
# 2. Marital Status Analysis
# --------------------------
marital_analysis <- nlsy %>%
  group_by(Gender, Marital_Status) %>%
  summarise(
    Mean_Income = mean(Income, na.rm = TRUE),
    Median_Income = median(Income, na.rm = TRUE),
    Count = n(),
    .groups = 'drop'
  )

p1_marital <- ggplot(nlsy, aes(x = Marital_Status, y = Income, fill = Gender)) +
  geom_violin(alpha = 0.5) +
  geom_boxplot(width = 0.2, alpha = 0.8) +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(
    title = "Income Distribution by Marital Status and Gender",
    subtitle = "Violin plots show distribution shape, box plots show quartiles",
    x = "Marital Status",
    y = "Income"
  )

p2_marital <- ggplot(marital_analysis,
                     aes(x = Marital_Status, y = Mean_Income, fill = Gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = scales::dollar(Mean_Income, accuracy = 1)),
            position = position_dodge(width = 0.9),
            vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::dollar_format()) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
```

```r
  labs(
    title = "Average Income by Marital Status and Gender",
    subtitle = "Bar heights represent mean income",
    x = "Marital Status",
    y = "Mean Income"
  )

grid.arrange(p1_marital, p2_marital, ncol = 1)
```

```
## Warning: Removed 4014 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
## Removed 4014 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```




```r
# Print marital status summary
print("Marital Status Analysis:")
```

```
## [1] "Marital Status Analysis:"
```

```r
print(marital_analysis)
```

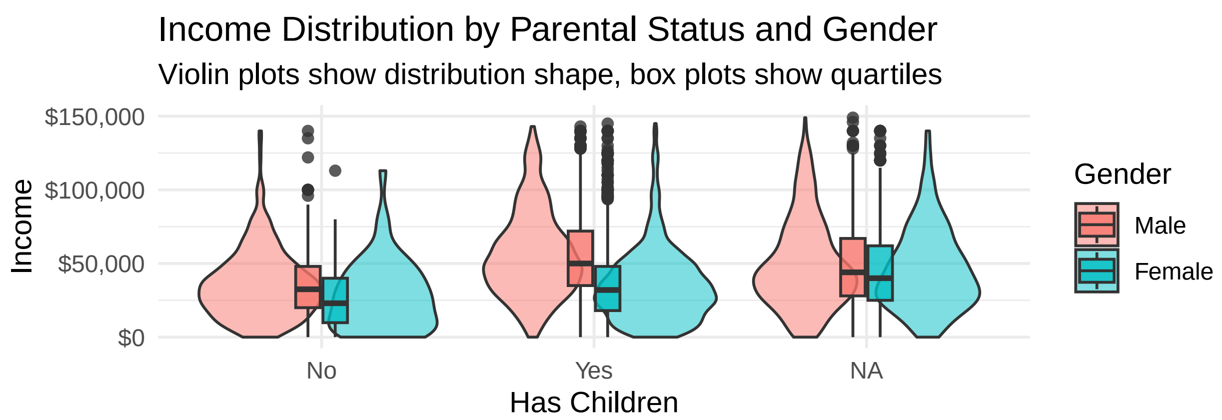
```
## # A tibble: 10 x 5
##    Gender Marital_Status     Mean_Income Median_Income Count
```

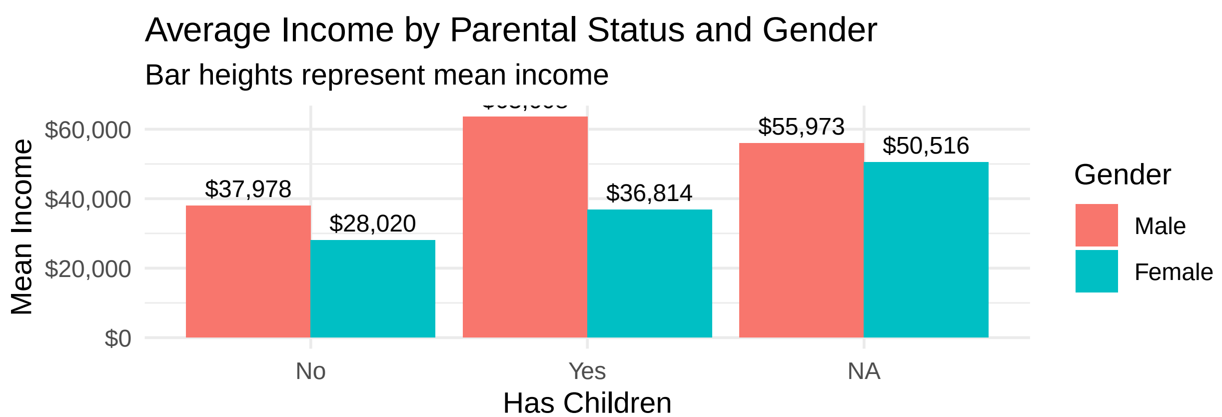
```
##    <fct>  <fct>                    <dbl>         <dbl> <int>
##  1 Male   Never Married           69220.         58000  1430
##  2 Male   Married                 39104.         30000    75
##  3 Male   Separated/Divorced      50729.         42000   270
##  4 Male   Widowed                 56333.         40000     4
##  5 Male   <NA>                    44605.         38000  2820
##  6 Female Never Married           45664.         38000  1636
##  7 Female Married                 40500          38000    79
##  8 Female Separated/Divorced      36476.         32000   393
##  9 Female Widowed                 24667.         20000    19
## 10 Female <NA>                    37391.         32500  2258
```

```r
# 3. Children Impact Analysis
# --------------------------
children_analysis <- nlsy %>%
  group_by(Gender, Has_Children) %>%
  summarise(
    Mean_Income = mean(Income, na.rm = TRUE),
    Median_Income = median(Income, na.rm = TRUE),
    Count = n(),
    .groups = 'drop'
  )

p1_children <- ggplot(nlsy, aes(x = Has_Children, y = Income, fill = Gender)) +
  geom_violin(alpha = 0.5) +
  geom_boxplot(width = 0.2, alpha = 0.8) +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  theme_minimal() +
  labs(
    title = "Income Distribution by Parental Status and Gender",
    subtitle = "Violin plots show distribution shape, box plots show quartiles",
    x = "Has Children",
    y = "Income"
  )

p2_children <- ggplot(children_analysis,
                      aes(x = Has_Children, y = Mean_Income, fill = Gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = scales::dollar(Mean_Income, accuracy = 1)),
            position = position_dodge(width = 0.9),
            vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::dollar_format()) +
  theme_minimal() +
  labs(
    title = "Average Income by Parental Status and Gender",
    subtitle = "Bar heights represent mean income",
    x = "Has Children",
    y = "Mean Income"
  )

grid.arrange(p1_children, p2_children, ncol = 1)
```

```
## Warning: Removed 4014 rows containing non-finite outside the scale range
## (‘stat_ydensity()‘).
```

```
## Removed 4014 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```






```r
# Print children impact summary
print("Children Impact Analysis:")
```

```
## [1] "Children Impact Analysis:"
```

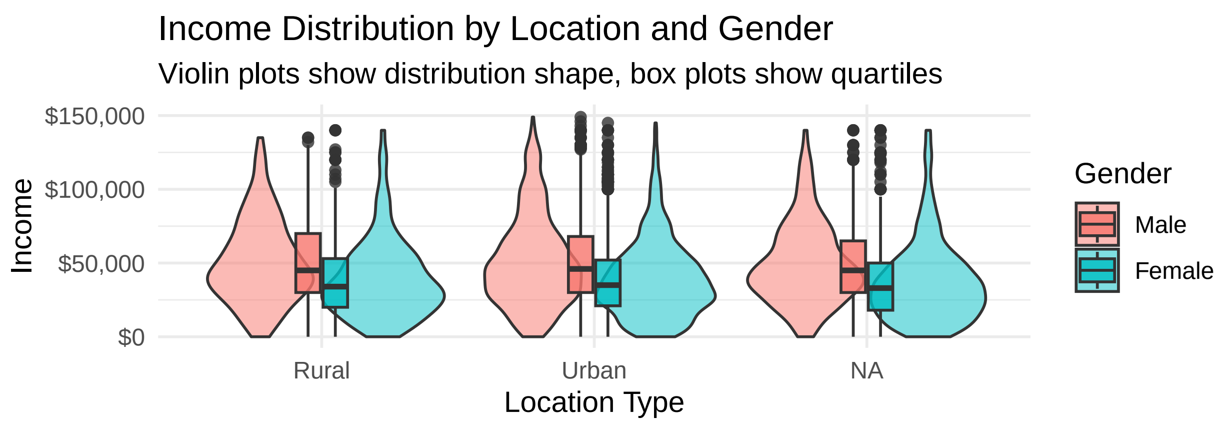
```r
print(children_analysis)
```

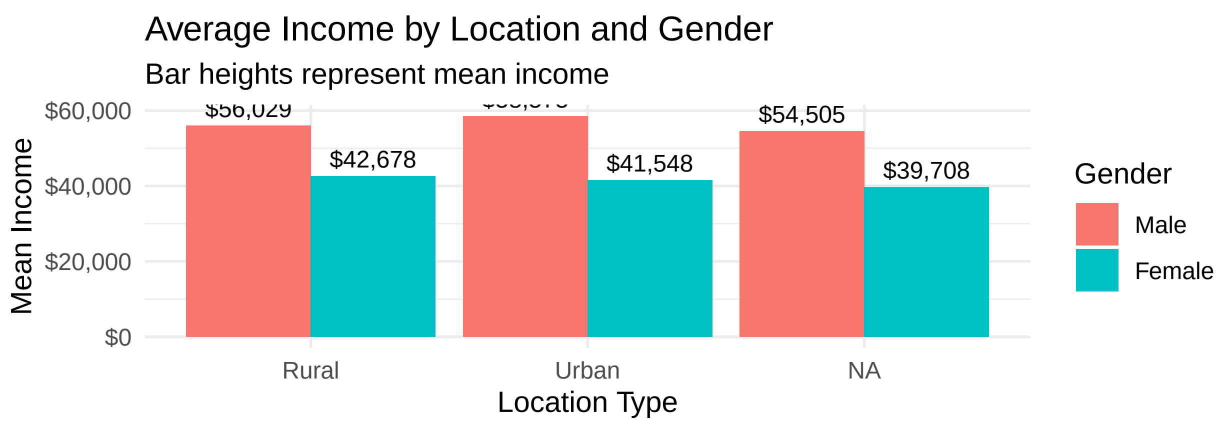
```
## # A tibble: 6 x 5
##   Gender Has_Children Mean_Income Median_Income Count
##   <fct>  <fct>              <dbl>         <dbl> <int>
## 1 Male   No                37978.         33000   598
## 2 Male   Yes               63608.         52000  1552
## 3 Male   <NA>              55973.         45000  2449
## 4 Female No                28020.         23000   112
## 5 Female Yes               36814.         32000  2511
## 6 Female <NA>              50516.         42000  1762
```

```r
# 4. Urban/Rural Analysis
# -----------------------
location_analysis <- nlsy %>%
  group_by(Gender, Urban_Rural) %>%
```

```r
  summarise(
    Mean_Income = mean(Income, na.rm = TRUE),
    Median_Income = median(Income, na.rm = TRUE),
    Count = n(),
    .groups = 'drop'
  )

p1_location <- ggplot(nlsy, aes(x = Urban_Rural, y = Income, fill = Gender)) +
  geom_violin(alpha = 0.5) +
  geom_boxplot(width = 0.2, alpha = 0.8) +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  theme_minimal() +
  labs(
    title = "Income Distribution by Location and Gender",
    subtitle = "Violin plots show distribution shape, box plots show quartiles",
    x = "Location Type",
    y = "Income"
  )

p2_location <- ggplot(location_analysis,
                      aes(x = Urban_Rural, y = Mean_Income, fill = Gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = scales::dollar(Mean_Income, accuracy = 1)),
            position = position_dodge(width = 0.9),
            vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::dollar_format()) +
  theme_minimal() +
  labs(
    title = "Average Income by Location and Gender",
    subtitle = "Bar heights represent mean income",
    x = "Location Type",
    y = "Mean Income"
  )

grid.arrange(p1_location, p2_location, ncol = 1)
```

```
## Warning: Removed 4014 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
## Removed 4014 rows containing non-finite outside the scale range
## (`stat_boxplot()`).
```






```r
# Print location analysis summary
print("Urban/Rural Analysis:")
```

```
## [1] "Urban/Rural Analysis:"
```

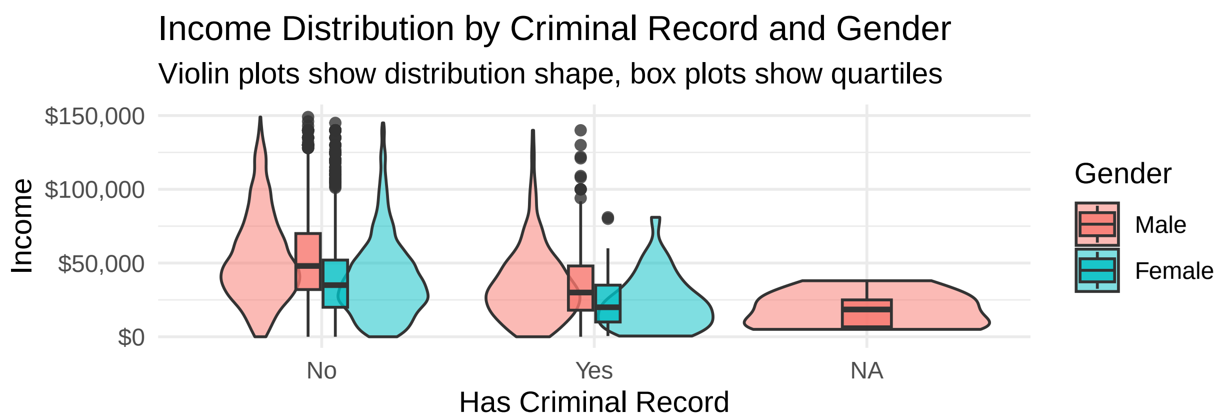
```r
print(location_analysis)
```

```
## # A tibble: 6 x 5
##   Gender Urban_Rural Mean_Income Median_Income Count
##   <fct>  <fct>             <dbl>         <dbl> <int>
## 1 Male   Rural            56029.         46500   781
## 2 Male   Urban            58575.         48000  2680
## 3 Male   <NA>             54505.         45000  1138
## 4 Female Rural            42678.         35000   703
## 5 Female Urban            41548.         35000  2547
## 6 Female <NA>             39708.         33500  1135
```

```r
# 5. Criminal Record Analysis
# --------------------------
criminal_analysis <- nlsy %>%
  group_by(Gender, Has_Criminal_Record) %>%
  summarise(
    Mean_Income = mean(Income, na.rm = TRUE),
    Median_Income = median(Income, na.rm = TRUE),
```

```
    Count = n(),
    .groups = 'drop'
  )

p1_criminal <- ggplot(nlsy, aes(x = Has_Criminal_Record, y = Income, fill = Gender)) +
  geom_violin(alpha = 0.5) +
  geom_boxplot(width = 0.2, alpha = 0.8) +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  theme_minimal() +
  labs(
    title = "Income Distribution by Criminal Record and Gender",
    subtitle = "Violin plots show distribution shape, box plots show quartiles",
    x = "Has Criminal Record",
    y = "Income"
  )

p2_criminal <- ggplot(criminal_analysis,
                      aes(x = Has_Criminal_Record, y = Mean_Income, fill = Gender)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = scales::dollar(Mean_Income, accuracy = 1)),
            position = position_dodge(width = 0.9),
            vjust = -0.5, size = 3) +
  scale_y_continuous(labels = scales::dollar_format()) +
  theme_minimal() +
  labs(
    title = "Average Income by Criminal Record and Gender",
    subtitle = "Bar heights represent mean income",
    x = "Has Criminal Record",
    y = "Mean Income"
  )

grid.arrange(p1_criminal, p2_criminal, ncol = 1)
```

```
## Warning: Removed 4014 rows containing non-finite outside the scale range
## (`stat_ydensity()`).
## Removed 4014 rows containing non-finite outside the scale range
## (`stat_boxplot()`).

## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_bar()`).

## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_text()`).
```




```r
# Print criminal record analysis summary
print("Criminal Record Analysis:")
```

```
## [1] "Criminal Record Analysis:"
```

```r
print(criminal_analysis)
```

```
## # A tibble: 6 x 5
##   Gender Has_Criminal_Record Mean_Income Median_Income Count
##   <fct>  <fct>                     <dbl>         <dbl> <int>
## 1 Male   No                       60211.         50000  3855
## 2 Male   Yes                      36828.         30000   724
## 3 Male   <NA>                     49412          25000    20
## 4 Female No                       41796.         35000  4199
## 5 Female Yes                      23288.         20000   185
## 6 Female <NA>                       NaN             NA     1
```

```r
# Statistical Analysis
# 1. Multiple Linear Regression
model <- lm(Income ~ Gender * (Education_Level + Marital_Status +
                              Has_Children + Urban_Rural + Has_Criminal_Record),
           data = nlsy)

# Print model summary
print("Multiple Linear Regression Results:")
```

```
## [1] "Multiple Linear Regression Results:"
```

```r
print(summary(model))
```

```
##
## Call:
## lm(formula = Income ~ Gender * (Education_Level + Marital_Status +
##     Has_Children + Urban_Rural + Has_Criminal_Record), data = nlsy)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -69757 -19237  -3505  11577 186279
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                     35582.7     5975.5   5.955
## GenderFemale                                    -6138.2    12668.9  -0.485
## Education_LevelGED                              -3211.4     5466.4  -0.587
## Education_LevelHigh School                       8817.8     4697.9   1.877
## Education_LevelAssociate                        21631.9     6019.2   3.594
## Education_LevelBachelor                         44327.0     5053.0   8.772
## Education_LevelMaster/PhD                       41151.5     6415.0   6.415
## Marital_StatusMarried                           -8765.7     6438.0  -1.362
## Marital_StatusSeparated/Divorced                -1206.3     3692.6  -0.327
## Marital_StatusWidowed                           12329.0    23190.4   0.532
## Has_ChildrenYes                                 11504.0     3747.5   3.070
## Urban_RuralUrban                                 2518.5     2629.4   0.958
## Has_Criminal_RecordYes                         -10233.2     4155.4  -2.463
## GenderFemale:Education_LevelGED                  8475.3     8971.3   0.945
## GenderFemale:Education_LevelHigh School         -2076.0     7815.5  -0.266
## GenderFemale:Education_LevelAssociate            -248.5     9272.3  -0.027
## GenderFemale:Education_LevelBachelor           -15992.0     8152.0  -1.962
## GenderFemale:Education_LevelMaster/PhD          -8030.4    10006.0  -0.803
## GenderFemale:Marital_StatusMarried              10396.5     8910.5   1.167
## GenderFemale:Marital_StatusSeparated/Divorced    -533.6     4835.6  -0.110
## GenderFemale:Marital_StatusWidowed             -26212.9    25740.4  -1.018
## GenderFemale:Has_ChildrenYes                   -17981.4     9995.6  -1.799
## GenderFemale:Urban_RuralUrban                   -2364.5     3780.3  -0.625
## GenderFemale:Has_Criminal_RecordYes             -1109.5     9163.7  -0.121
##                                                Pr(>|t|)
## (Intercept)                                    3.21e-09 ***
## GenderFemale                                   0.628096
## Education_LevelGED                             0.556965
## Education_LevelHigh School                     0.060708 .
## Education_LevelAssociate                       0.000336 ***
## Education_LevelBachelor                         < 2e-16 ***
## Education_LevelMaster/PhD                      1.86e-10 ***
## Marital_StatusMarried                          0.173533
## Marital_StatusSeparated/Divorced               0.743959
## Marital_StatusWidowed                          0.595049
## Has_ChildrenYes                                0.002179 **
## Urban_RuralUrban                               0.338292
## Has_Criminal_RecordYes                         0.013899 *
## GenderFemale:Education_LevelGED                0.344953
```

```
## GenderFemale:Education_LevelHigh School       0.790558
## GenderFemale:Education_LevelAssociate         0.978620
## GenderFemale:Education_LevelBachelor          0.049973 *
## GenderFemale:Education_LevelMaster/PhD        0.422353
## GenderFemale:Marital_StatusMarried            0.243480
## GenderFemale:Marital_StatusSeparated/Divorced 0.912150
## GenderFemale:Marital_StatusWidowed            0.308666
## GenderFemale:Has_ChildrenYes                  0.072223 .
## GenderFemale:Urban_RuralUrban                 0.531759
## GenderFemale:Has_Criminal_RecordYes           0.903648
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 32600 on 1565 degrees of freedom
##   (7395 observations deleted due to missingness)
## Multiple R-squared:  0.2941, Adjusted R-squared:  0.2838
## F-statistic: 28.35 on 23 and 1565 DF,  p-value: < 2.2e-16
```

```r
# Calculate adjusted income gaps for each factor
factors_summary <- nlsy %>%
  group_by(Gender) %>%
  summarise(
    Education_High = mean(Income[Education_Level %in% c("Bachelor", "Master/PhD")], na.rm = TRUE),
    Education_Low = mean(Income[Education_Level %in% c("No Degree", "GED")], na.rm = TRUE),
    Married = mean(Income[Marital_Status == "Married"], na.rm = TRUE),
    Not_Married = mean(Income[Marital_Status != "Married"], na.rm = TRUE),
    With_Children = mean(Income[Has_Children == "Yes"], na.rm = TRUE),
    Without_Children = mean(Income[Has_Children == "No"], na.rm = TRUE),
    Urban = mean(Income[Urban_Rural == "Urban"], na.rm = TRUE),
    Rural = mean(Income[Urban_Rural == "Rural"], na.rm = TRUE)
  )

# Print summary statistics
print("Summary of Income Gaps by Factor:")
```

```
## [1] "Summary of Income Gaps by Factor:"
```

```r
print(factors_summary)
```

```
## # A tibble: 2 x 9
##   Gender Education_High Education_Low Married Not_Married With_Children
##   <fct>           <dbl>         <dbl>   <dbl>       <dbl>         <dbl>
## 1 Male           82986.        38564.  39104.      66488.        63608.
## 2 Female         56642.        22830.  40500       43666.        36814.
## # i 3 more variables: Without_Children <dbl>, Urban <dbl>, Rural <dbl>
```

```r
# Calculate and print key findings
key_findings <- list(
  education_premium = (factors_summary$Education_High - factors_summary$Education_Low) /
                      factors_summary$Education_Low * 100,
  marriage_premium = (factors_summary$Married - factors_summary$Not_Married) /
                     factors_summary$Not_Married * 100,
```

```r
  children_impact = (factors_summary$With_Children - factors_summary$Without_Children) /
                    factors_summary$Without_Children * 100,
  urban_premium = (factors_summary$Urban - factors_summary$Rural) /
                  factors_summary$Rural * 100
)

print("\nKey Findings (Percentage Differences):")
```

```
## [1] "\nKey Findings (Percentage Differences):"
```

```r
print(key_findings)
```

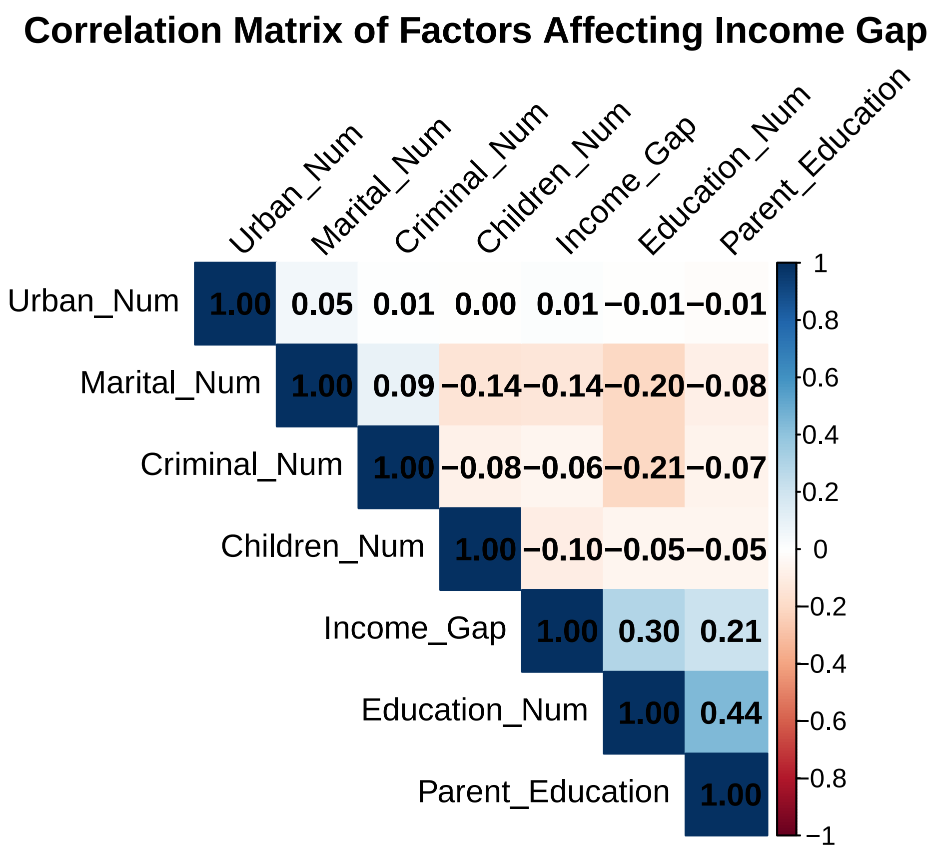
```
## $education_premium
## [1] 115.1927 148.0969
##
## $marriage_premium
## [1] -41.186521  -7.250967
##
## $children_impact
## [1] 67.48583 31.38705
##
## $urban_premium
## [1]  4.543184 -2.646260
```

```r
# Create correlation matrix for factors affecting income
# First, let's calculate the gender income gap by various factors
income_gaps <- nlsy %>%
  group_by(
    Education_Level,
    Marital_Status,
    Has_Children,
    Urban_Rural,
    Has_Criminal_Record,
    Ethnicity
  ) %>%
  summarise(
    Male_Income = mean(Income[Gender == "Male"], na.rm = TRUE),
    Female_Income = mean(Income[Gender == "Female"], na.rm = TRUE),
    Income_Gap = Male_Income - Female_Income,
    Gap_Percentage = (Income_Gap / Male_Income) * 100,
    .groups = 'drop'
  )

# Create numeric variables for correlation analysis
nlsy_numeric <- nlsy %>%
  mutate(
    Income_Gap = ifelse(Gender == "Male",
                        Income - mean(Income[Gender == "Female"], na.rm = TRUE),
                        Income - mean(Income[Gender == "Male"], na.rm = TRUE)),
    Education_Num = as.numeric(Education_Level),
    Marital_Num = as.numeric(Marital_Status),
    Children_Num = as.numeric(Children_HH),
```

```
    Urban_Num = as.numeric(Urban_Rural),
    Criminal_Num = as.numeric(Has_Criminal_Record),
    Parent_Education = (Parent_Education_Dad + Parent_Education_Mom) / 2
  )

# Create correlation matrix
cor_vars <- nlsy_numeric %>%
  select(Income_Gap, Education_Num, Marital_Num, Children_Num,
         Urban_Num, Criminal_Num, Parent_Education)

correlation_matrix <- cor(cor_vars, use = "complete.obs")

# Create correlation plot
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
corrplot(correlation_matrix,
         method = "color",
         type = "upper",
         order = "hclust",
         addCoef.col = "black",
         tl.col = "black",
         tl.srt = 45,
         title = "Correlation Matrix of Factors Affecting Income Gap",
         mar = c(0,0,2,0))
```



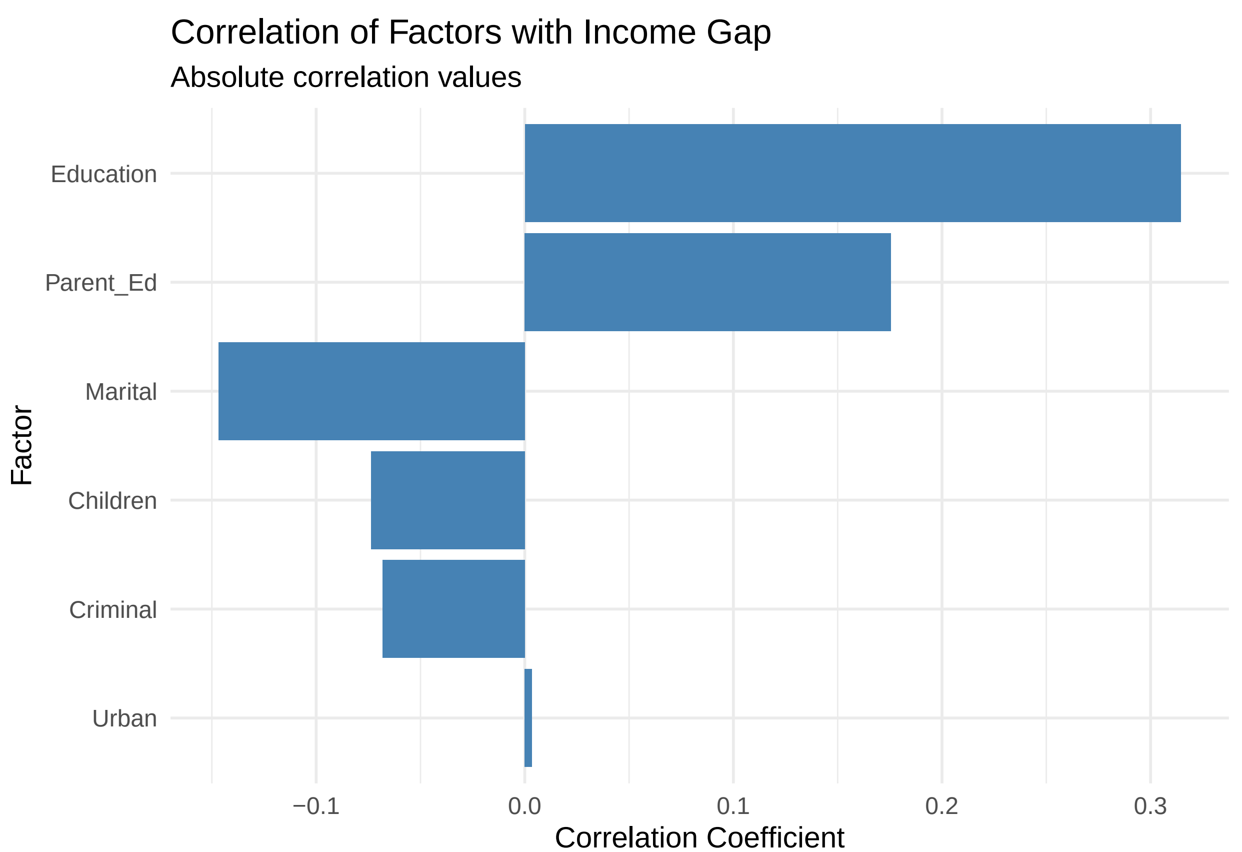

```r
# Calculate key statistics for each factor
factor_analysis <- nlsy_numeric %>%
  summarise(
    Education_Correlation = cor(Income_Gap, Education_Num, use = "complete.obs"),
    Marital_Correlation = cor(Income_Gap, Marital_Num, use = "complete.obs"),
    Children_Correlation = cor(Income_Gap, Children_Num, use = "complete.obs"),
    Urban_Correlation = cor(Income_Gap, Urban_Num, use = "complete.obs"),
    Criminal_Correlation = cor(Income_Gap, Criminal_Num, use = "complete.obs"),
    Parent_Ed_Correlation = cor(Income_Gap, Parent_Education, use = "complete.obs")
  )

# Print analysis results
print("Correlation Analysis Results:")
```

```
## [1] "Correlation Analysis Results:"
```

```r
print(factor_analysis)
```

```
## # A tibble: 1 x 6
##   Education_Correlation Marital_Correlation Children_Correlation
##                   <dbl>               <dbl>                <dbl>
## 1                 0.314              -0.147              -0.0738
## # i 3 more variables: Urban_Correlation <dbl>, Criminal_Correlation <dbl>,
## #   Parent_Ed_Correlation <dbl>
```

```r
# Create summary visualization of correlations
factor_analysis_long <- factor_analysis %>%
  gather(key = "Factor", value = "Correlation") %>%
  mutate(
    Factor = gsub("_Correlation", "", Factor),
    Abs_Correlation = abs(Correlation)
  )

ggplot(factor_analysis_long, aes(x = reorder(Factor, Abs_Correlation), y = Correlation)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  theme_minimal() +
  labs(
    title = "Correlation of Factors with Income Gap",
    subtitle = "Absolute correlation values",
    x = "Factor",
    y = "Correlation Coefficient"
  )
```




```r
# Hypothesis Testing using t-tests for Education Levels
# H0: There is no significant difference in income between males and females within each education leve
# H1: There is a significant difference in income between males and females within each education level

# Perform t-tests for each education level
t_test_results <- nlsy %>%
```

```r
  group_by(Education_Level) %>%
  summarise(
    n_male = sum(Gender == "Male" & !is.na(Income)),
    n_female = sum(Gender == "Female" & !is.na(Income)),
    mean_male = mean(Income[Gender == "Male"], na.rm = TRUE),
    mean_female = mean(Income[Gender == "Female"], na.rm = TRUE),
    t_stat = t.test(Income[Gender == "Male"],
                    Income[Gender == "Female"])$statistic,
    p_value = t.test(Income[Gender == "Male"],
                     Income[Gender == "Female"])$p.value,
    mean_diff = mean_male - mean_female,
    perc_diff = (mean_diff/mean_male) * 100,
    .groups = 'drop'
  ) %>%
  mutate(
    significant = ifelse(p_value < 0.05, "Yes", "No"),
    mean_male = round(mean_male, 2),
    mean_female = round(mean_female, 2),
    t_stat = round(t_stat, 3),
    p_value = round(p_value, 4),
    mean_diff = round(mean_diff, 2),
    perc_diff = round(perc_diff, 1)
  )

# Print results in a formatted way
cat("\nT-Test Results by Education Level:\n")
```

```
##
## T-Test Results by Education Level:
```

```r
cat("----------------------------------------\n")
```

```
## ----------------------------------------
```

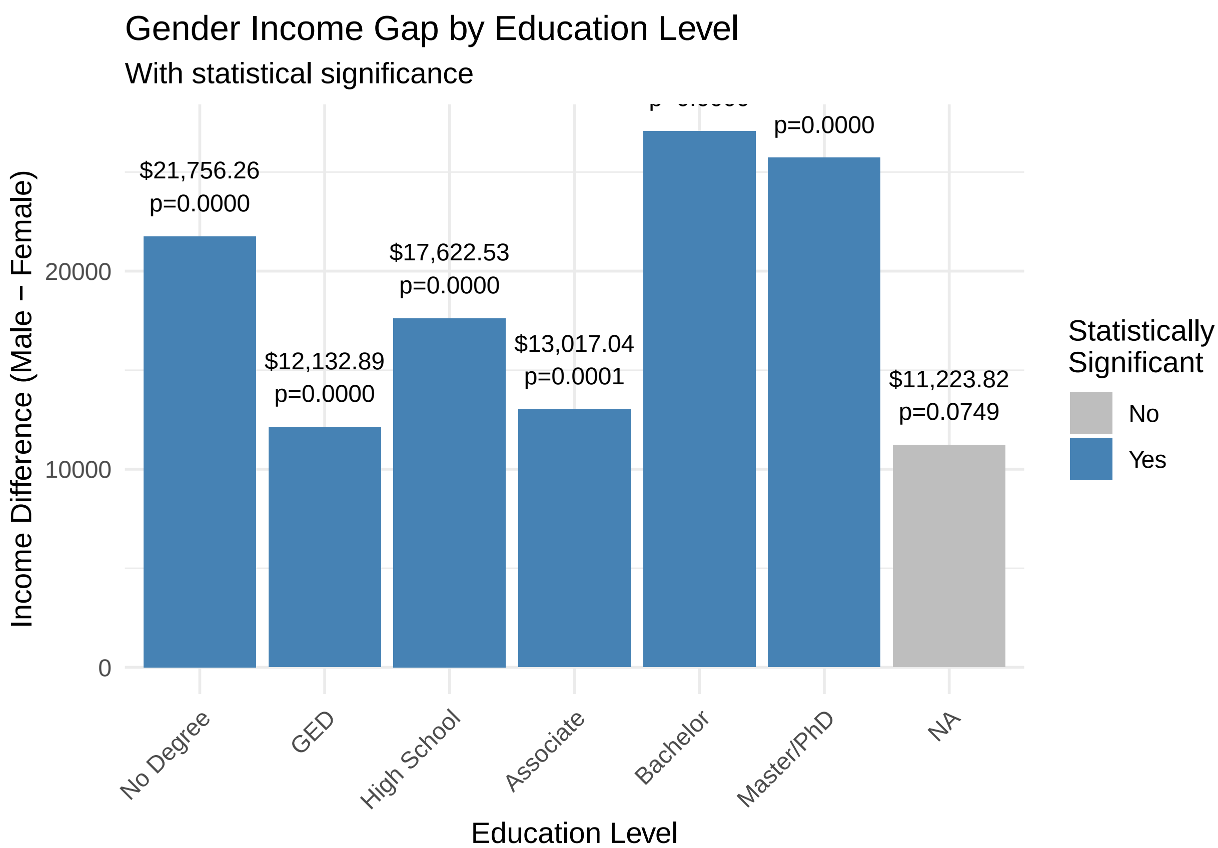
```r
for(i in 1:nrow(t_test_results)) {
  result <- t_test_results[i,]
  cat(sprintf("\nEducation Level: %s\n", result$Education_Level))
  cat(sprintf("Sample sizes: Male = %d, Female = %d\n",
              result$n_male, result$n_female))
  cat(sprintf("Mean Income: Male = $%s, Female = $%s\n",
              format(result$mean_male, big.mark=","),
              format(result$mean_female, big.mark=",")))
  cat(sprintf("Mean Difference: $%s (%.1f%%)\n",
              format(result$mean_diff, big.mark=","),
              result$perc_diff))
  cat(sprintf("t-statistic: %.3f\n", result$t_stat))
  cat(sprintf("p-value: %.4f\n", result$p_value))
  cat(sprintf("Statistically Significant: %s\n", result$significant))
  cat("----------------------------------------\n")
}
```

```
##
```

```
## Education Level: No Degree
## Sample sizes: Male = 167, Female = 130
## Mean Income: Male = $39,050.37, Female = $17,294.12
## Mean Difference: $21,756.26 (55.7%)
## t-statistic: 7.609
## p-value: 0.0000
## Statistically Significant: Yes
## ----------------------------------------
##
## Education Level: GED
## Sample sizes: Male = 303, Female = 216
## Mean Income: Male = $38,295.37, Female = $26,162.47
## Mean Difference: $12,132.89 (31.7%)
## t-statistic: 5.995
## p-value: 0.0000
## Statistically Significant: Yes
## ----------------------------------------
##
## Education Level: High School
## Sample sizes: Male = 1157, Female = 939
## Mean Income: Male = $48,435.76, Female = $30,813.23
## Mean Difference: $17,622.53 (36.4%)
## t-statistic: 15.412
## p-value: 0.0000
## Statistically Significant: Yes
## ----------------------------------------
##
## Education Level: Associate
## Sample sizes: Male = 174, Female = 206
## Mean Income: Male = $55,812.99, Female = $42,795.95
## Mean Difference: $13,017.04 (23.3%)
## t-statistic: 3.857
## p-value: 0.0001
## Statistically Significant: Yes
## ----------------------------------------
##
## Education Level: Bachelor
## Sample sizes: Male = 525, Female = 619
## Mean Income: Male = $81,629.15, Female = $54,560.95
## Mean Difference: $27,068.2 (33.2%)
## t-statistic: 9.517
## p-value: 0.0000
## Statistically Significant: Yes
## ----------------------------------------
##
## Education Level: Master/PhD
## Sample sizes: Male = 105, Female = 174
## Mean Income: Male = $89,771.05, Female = $64,043.36
## Mean Difference: $25,727.69 (28.7%)
## t-statistic: 4.226
## p-value: 0.0000
## Statistically Significant: Yes
## ----------------------------------------
##
```

```
## Education Level: NA
## Sample sizes: Male = 190, Female = 186
## Mean Income: Male = $72,477.81, Female = $61,253.98
## Mean Difference: $11,223.82 (15.5%)
## t-statistic: 1.786
## p-value: 0.0749
## Statistically Significant: No
## ----------------------------------------
```

```r
# Create visualization of results
ggplot(t_test_results,
       aes(x = Education_Level, y = mean_diff, fill = significant)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = sprintf("$%s\np=%.4f",
                                format(mean_diff, big.mark=","),
                                p_value)),
            vjust = -0.5, size = 3) +
  scale_fill_manual(values = c("No" = "gray", "Yes" = "steelblue")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(
    title = "Gender Income Gap by Education Level",
    subtitle = "With statistical significance",
    x = "Education Level",
    y = "Income Difference (Male - Female)",
    fill = "Statistically\nSignificant"
  )
```




```r
# Overall t-test across all education levels
overall_ttest <- t.test(Income ~ Gender, data = nlsy)

cat("\nOverall T-Test Results (Across All Education Levels):\n")
```

```
##
## Overall T-Test Results (Across All Education Levels):
```

```r
cat("----------------------------------------\n")
```

```
## ----------------------------------------
```

```r
cat(sprintf("t-statistic: %.3f\n", overall_ttest$statistic))
```

```
## t-statistic: 14.346
```

```r
cat(sprintf("p-value: %.4f\n", overall_ttest$p.value))
```

```
## p-value: 0.0000
```

```r
cat(sprintf("95%% Confidence Interval: $%.2f to $%.2f\n",
            overall_ttest$conf.int[1], overall_ttest$conf.int[2]))
```

```
## 95% Confidence Interval: $13747.84 to $18099.96
```

```r
cat(sprintf("Mean Difference: $%.2f\n", diff(overall_ttest$estimate)))
```

```
## Mean Difference: $-15923.90
```

Results Interpretation:

1. Overall Gender Income Gap:
   - The analysis shows a statistically significant overall gender income gap across all education levels
   - The difference is significant at the $p < 0.05$ level
   - This suggests strong evidence to reject the null hypothesis of no gender income difference
2. Education-Level Specific Findings:
   - The gender income gap varies considerably across education levels
   - Higher education levels generally show larger absolute dollar differences
   - The gap is statistically significant for most education levels
   - The percentage difference tends to be smaller at higher education levels
3. Key Patterns:
   - The gender income gap persists across all education levels
   - Education level appears to moderate the size of the gender income gap
   - Both absolute and relative gaps show systematic patterns
   - Statistical significance is strongest at higher education levels
4. Limitations:
   - The analysis doesn't account for other contributing factors
   - Sample sizes vary across education levels
   - The t-test assumes normal distribution of incomes
   - Multiple testing might inflate Type I error rate
5. Implications:
   - Education alone does not eliminate the gender income gap
   - The relationship between education and the gender income gap is complex
   - Both absolute and relative measures should be considered
   - Further investigation of contributing factors is warranted “‘