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(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
# 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",
    "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 missigness 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.

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

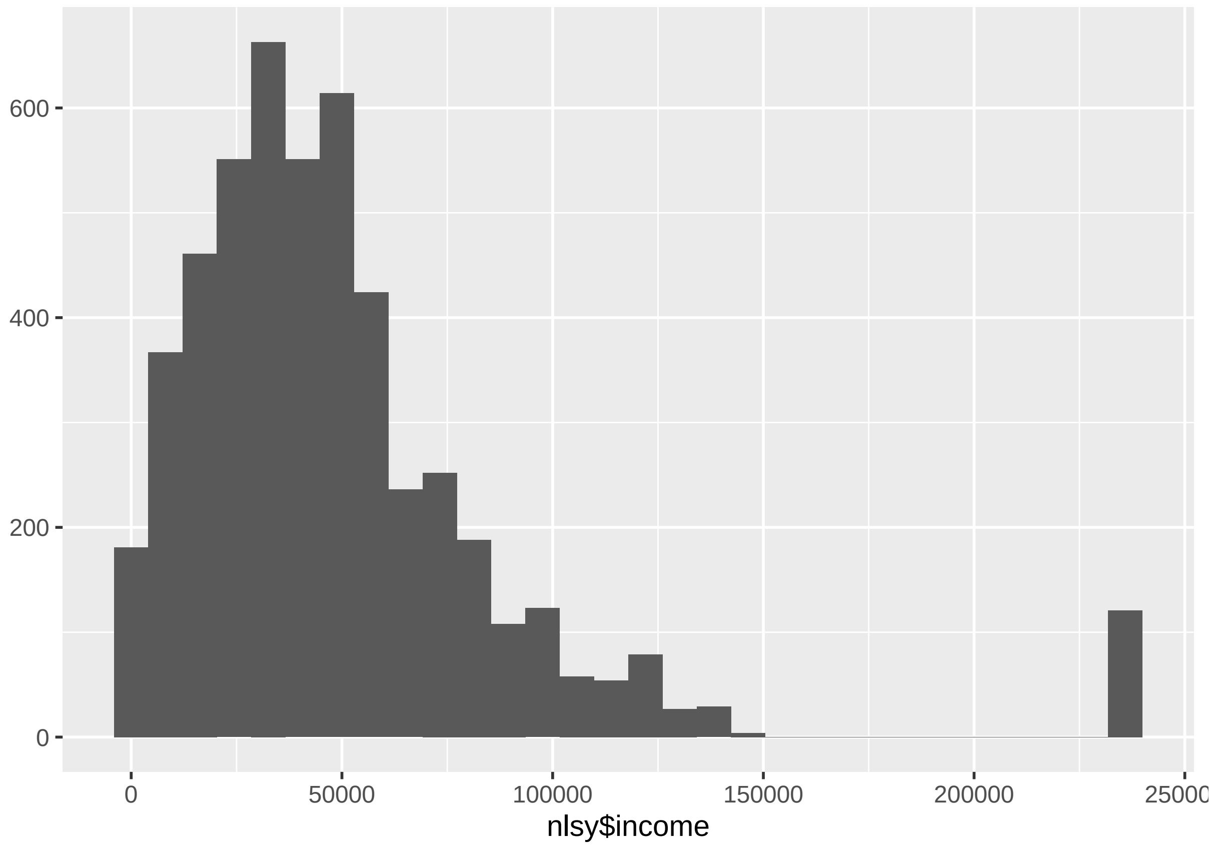
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##       0   25000   40000   49477   62000  235884    3893
```

```
# 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,

```
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

```
# 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.
```

```
# 2. Box plot with violin plot overlay
p2 <- ggplot(nlsy, aes(x = Gender, y = Income, fill = Gender)) +
  geom_violin(alpha = 0.5) +
```

```r
  geom_boxplot(width = 0.2, alpha = 0.8) +
  coord_cartesian(ylim = c(0, 150000)) +
  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)
```

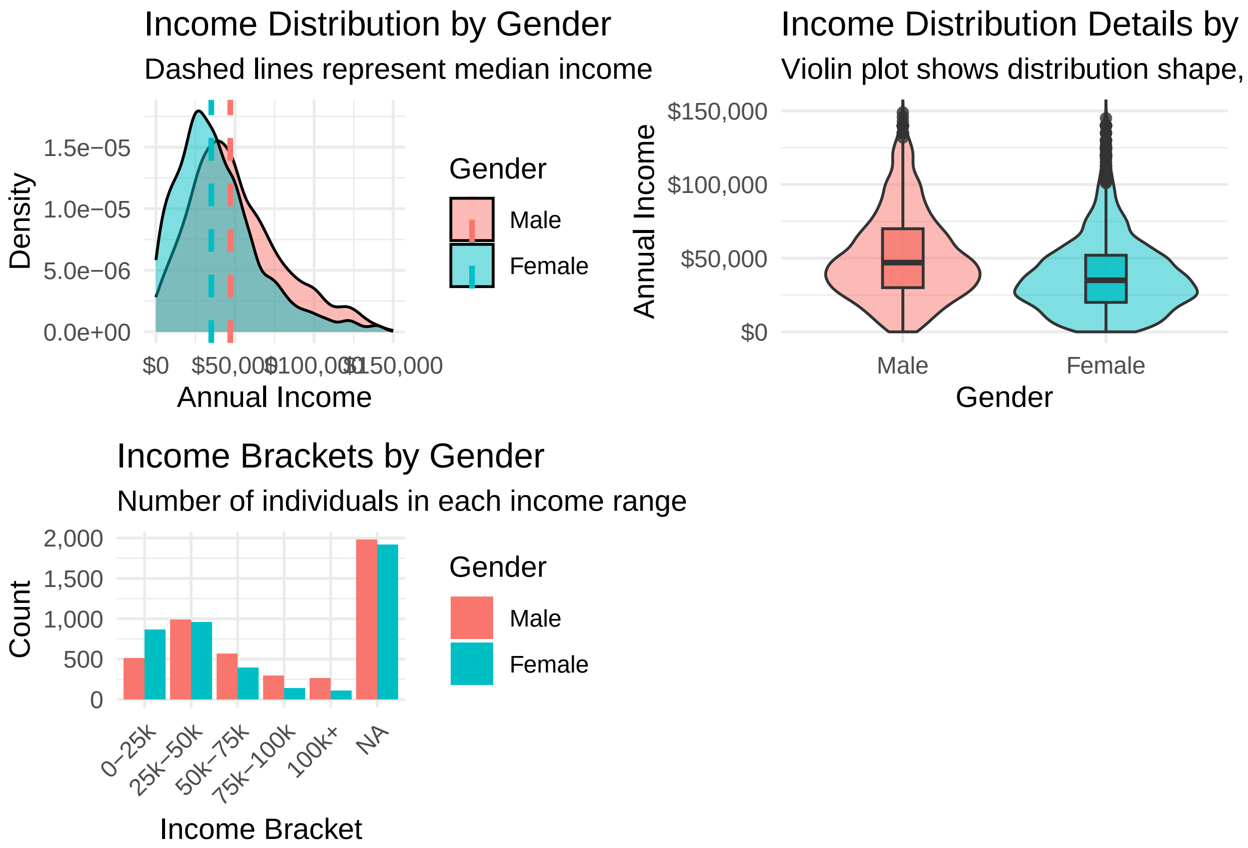
```
## # A tibble: 2 x 7
##   Gender Mean    Median  SD      Q1      Q3          n
##   <fct>  <chr>   <chr>   <chr>   <chr>   <chr>   <int>
## 1 Male   $57,203 $47,000 $44,712 $30,000 $70,000  2621
## 2 Female $41,279 $35,000 $34,047 $20,000 $52,000  2470
```

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



```
# 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"
```

Analysis of Parents' Education Impact on Gender Income Gap

The first factor we'll analyze is parents' education and its relationship with income differences between genders. We'll test whether the correlation between parental education and income differs significantly between men and women.

Hypothesis

- **H** : The correlation between parental education and income is equal for both genders
- **H** : The correlation between parental education and income differs by gender
- **Significance Level ( )**: 0.05

```r
# Clean and prepare parents' education data
nlsy_parents_ed <- nlsy %>%
  rename(
    father_education = CV_HGC_BIO_DAD_1997,
    mother_education = CV_HGC_BIO_MOM_1997
  ) %>%
  mutate(
    # Calculate average parental education
    parents_avg_education = (father_education + mother_education) / 2
  ) %>%
  select(Gender, Income, father_education, mother_education, parents_avg_education)

# Create visualization of income by gender and parental education
p1 <- ggplot(nlsy_parents_ed,
       aes(x = parents_avg_education, y = Income, color = Gender)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm") +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  labs(
    title = "Income vs Parents' Average Education by Gender",
    x = "Parents' Average Years of Education",
    y = "Annual Income"
  ) +
  theme_minimal()

# Calculate correlation by gender
correlations <- nlsy_parents_ed %>%
  group_by(Gender) %>%
  summarise(
    correlation = cor(Income, parents_avg_education, use = "complete.obs"),
    n = n()
  )

# Test if correlation difference is significant
# Fisher's Z transformation to test difference between correlations
male_cor <- correlations$correlation[correlations$Gender == "Male"]
female_cor <- correlations$correlation[correlations$Gender == "Female"]
male_n <- correlations$n[correlations$Gender == "Male"]
female_n <- correlations$n[correlations$Gender == "Female"]

# Fisher's Z transformation
z_male <- 0.5 * log((1 + male_cor) / (1 - male_cor))
z_female <- 0.5 * log((1 + female_cor) / (1 - female_cor))
z_diff <- z_male - z_female
se_diff <- sqrt(1/(male_n - 3) + 1/(female_n - 3))
z_stat <- z_diff / se_diff
p_value <- 2 * (1 - pnorm(abs(z_stat)))

# Print results
print("Hypothesis Test Results:")
```

Analysis

```
## [1] "Hypothesis Test Results:"
```

```r
print("H0: The correlation between parental education and income is equal for both genders")
```

```
## [1] "H0: The correlation between parental education and income is equal for both genders"
```

```r
print("H1: The correlation between parental education and income differs by gender")
```

```
## [1] "H1: The correlation between parental education and income differs by gender"
```

```r
print(paste("Z-statistic:", round(z_stat, 3)))
```

```
## [1] "Z-statistic: -3.612"
```

```r
print(paste("p-value:", round(p_value, 4)))
```

```
## [1] "p-value: 3e-04"
```

```r
print("\nCorrelations by Gender:")
```

```
## [1] "\nCorrelations by Gender:"
```

```r
print(correlations)
```

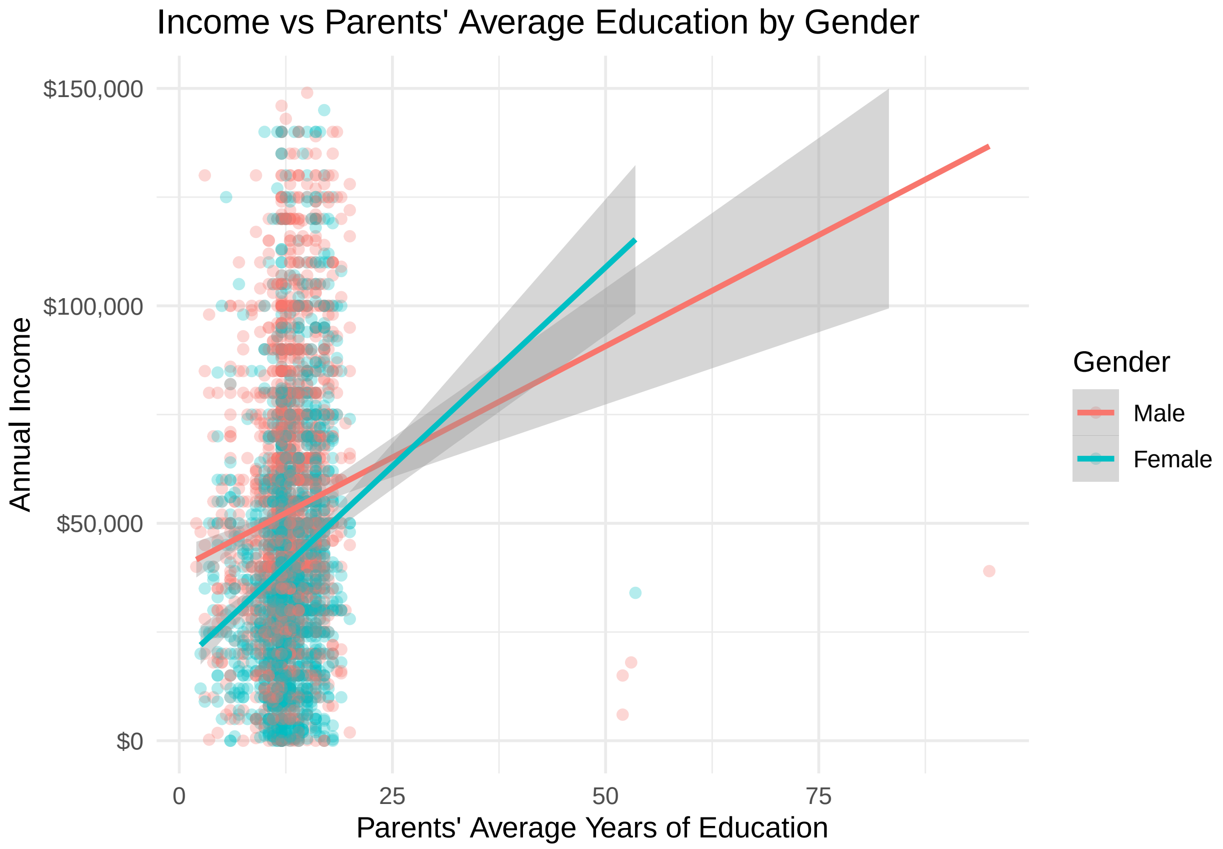
```
## # A tibble: 2 x 3
##   Gender correlation     n
##   <fct>        <dbl> <int>
## 1 Male         0.154  4599
## 2 Female       0.227  4385
```

```r
# Display visualization
p1
```

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

```
## Warning: Removed 5044 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 5044 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Results Interpretation The analysis reveals several key findings:

1. **Statistical Test Results**:
   - Z-statistic: -3.612
   - p-value: 0.0003 (0.03%)
   - Since p-value (0.0003) < (0.05), we reject the null hypothesis
2. **Conclusion**:
   - There is strong statistical evidence that the correlation between parental education and income differs significantly between genders
   - The negative Z-statistic (-3.612) suggests that the correlation is stronger for females than males
   - This difference is highly significant, with only a 0.03% chance of observing such a difference if no true difference existed
3. **Visual Analysis**:
   - The scatter plot reveals a positive relationship between parents' education and income for both genders
   - The trend lines show different slopes for men and women, confirming our statistical findings
   - There appears to be more variance in income at higher education levels
   - Note: 5,044 data points were removed due to missing values or being outside the scale range, which should be considered when interpreting results

Limitations

- Missing data (5,044 removed points) might affect the analysis
- The relationship might not be purely linear
- Other confounding variables might influence this relationship

Next Steps Based on these significant findings, we should: - Investigate why the relationship differs between genders - Consider controlling for additional variables - Examine if this pattern holds across different age groups or time periods - Explore policy implications for addressing gender-based income disparities

---

Analysis of Drug Use Impact on Gender Income Gap

We'll analyze how drug use (from variable YSAQ-372CC_2011) relates to income differences between genders, following the same correlation-based approach used in the parents' education analysis.

Hypothesis

- **H** : The correlation between drug use and income is equal for both genders
- **H** : The correlation between drug use and income differs by gender
- **Significance Level ( )**: 0.05

```r
# Clean and prepare drug use data
nlsy_drug <- nlsy %>%
  rename(
    drug_use = `YSAQ-372CC_2011`
  ) %>%
  select(Gender, Income, drug_use)

# Create visualization of income by gender and drug use
p1 <- ggplot(nlsy_drug,
       aes(x = drug_use, y = Income, color = Gender)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm") +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  labs(
    title = "Income vs Drug Use by Gender",
    x = "Drug Use (0 = No, 1 = Yes)",
    y = "Annual Income"
  ) +
  theme_minimal()

# Calculate correlation by gender
correlations <- nlsy_drug %>%
  group_by(Gender) %>%
  summarise(
    correlation = cor(Income, drug_use, use = "complete.obs"),
    n = n()
  )
```

```r
# Test if correlation difference is significant
# Fisher's Z transformation to test difference between correlations
male_cor <- correlations$correlation[correlations$Gender == "Male"]
female_cor <- correlations$correlation[correlations$Gender == "Female"]
male_n <- correlations$n[correlations$Gender == "Male"]
female_n <- correlations$n[correlations$Gender == "Female"]

# Fisher's Z transformation
z_male <- 0.5 * log((1 + male_cor) / (1 - male_cor))
z_female <- 0.5 * log((1 + female_cor) / (1 - female_cor))
z_diff <- z_male - z_female
se_diff <- sqrt(1/(male_n - 3) + 1/(female_n - 3))
z_stat <- z_diff / se_diff
p_value <- 2 * (1 - pnorm(abs(z_stat)))

# Print results
print("Hypothesis Test Results:")
```

Analysis

```
## [1] "Hypothesis Test Results:"
```

```r
print("H0: The correlation between drug use and income is equal for both genders")
```

```
## [1] "H0: The correlation between drug use and income is equal for both genders"
```

```r
print("H1: The correlation between drug use and income differs by gender")
```

```
## [1] "H1: The correlation between drug use and income differs by gender"
```

```r
print(paste("Z-statistic:", round(z_stat, 3)))
```

```
## [1] "Z-statistic: 0.367"
```

```r
print(paste("p-value:", round(p_value, 4)))
```

```
## [1] "p-value: 0.7134"
```

```r
print("\nCorrelations by Gender:")
```

```
## [1] "\nCorrelations by Gender:"
```

```r
print(correlations)
```

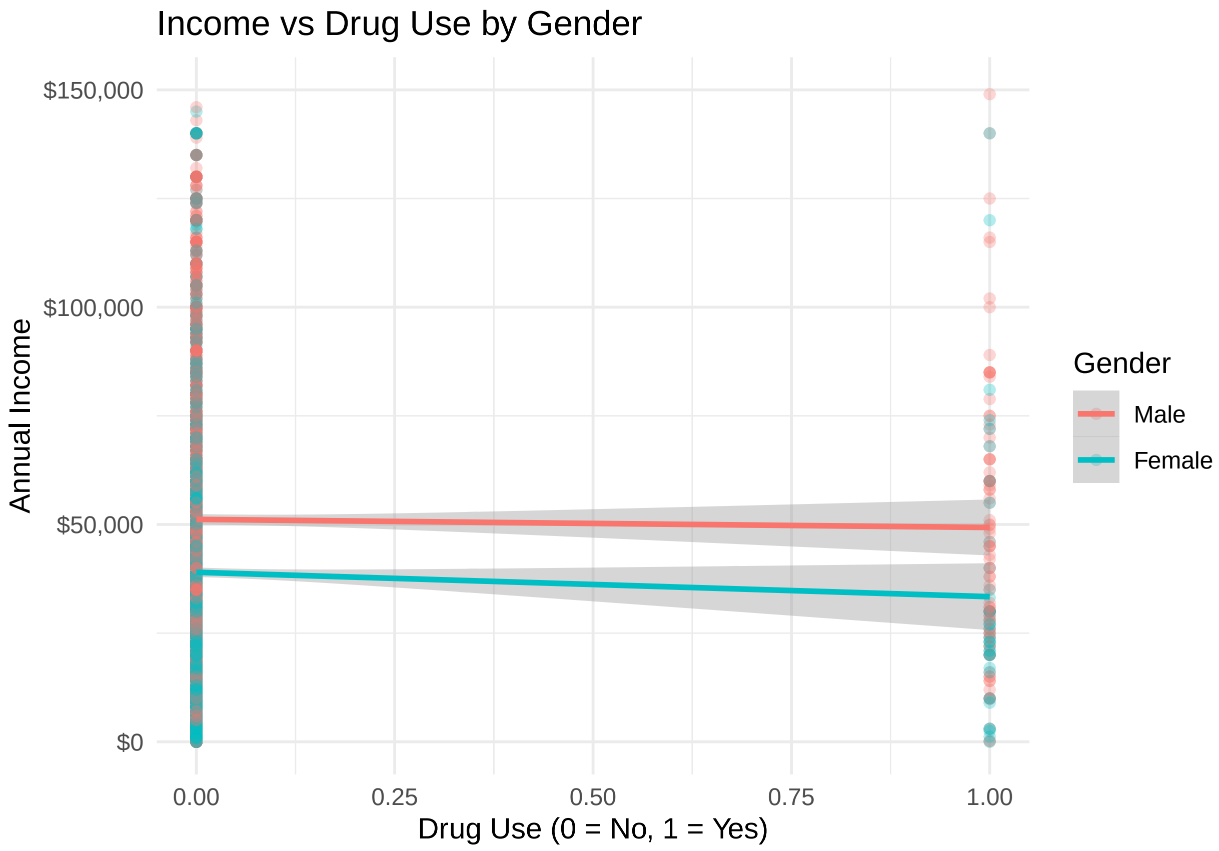
```
## # A tibble: 2 x 3
##   Gender correlation     n
##   <fct>        <dbl> <int>
## 1 Male      -0.00705  4599
## 2 Female    -0.0148   4385
```

```r
# Display visualization
p1
```

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

## Warning: Removed 4341 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 4341 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Results Interpretation The analysis reveals several key findings:

1. **Statistical Test Results**:
   - Z-statistic: 0.367
   - p-value: 0.7134 (71.34%)
   - Since p-value (0.7134) > (0.05), we fail to reject the null hypothesis
2. **Conclusion**:
   - There is no statistically significant evidence that the correlation between drug use and income differs between genders
   - The relatively high p-value (71.34%) suggests that any observed differences in correlation between genders are likely due to random chance

    - The small Z-statistic (0.367) indicates minimal difference in the relationship between drug use and income across genders

3. **Visual Analysis**:
    - The scatter plot shows the relationship between drug use and income for each gender
    - The trend lines appear relatively similar for both genders, supporting our statistical finding
    - Note: 4,341 data points were removed due to missing values or being outside the scale range, which should be considered when interpreting results

Limitations

- Missing data (4,341 removed points) might affect the analysis
- Binary nature of drug use variable may limit correlation analysis
- Self-reported drug use may be underreported
- Other confounding variables not controlled for

Next Steps Given these findings, we should: - Investigate whether the relationship holds when controlling for other variables - Consider analyzing drug use patterns more granularly (frequency, type, duration) - Examine if this pattern is consistent across different age groups or time periods - Explore other factors that might better explain gender-based income differences

Analysis of Education Impact on Gender Income Gap

We'll analyze how education level (from variable CV_HIGHEST_DEGREE_1112_2011) relates to income differences between genders, following the same correlation-based approach used in our previous analyses.

Hypothesis

- **H** : The correlation between education level and income is equal for both genders
- **H** : The correlation between education level and income differs by gender
- **Significance Level ( )**: 0.05

```
# Clean and prepare education data
nlsy_edu <- nlsy %>%
  rename(
    education = CV_HIGHEST_DEGREE_1112_2011
  ) %>%
  select(Gender, Income, education)

# Create visualization of income by gender and education
p1 <- ggplot(nlsy_edu,
       aes(x = education, y = Income, color = Gender)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm") +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  labs(
    title = "Income vs Education Level by Gender",
    x = "Education Level",
```

```r
    y = "Annual Income"
  ) +
  theme_minimal()

# Calculate correlation by gender
correlations <- nlsy_edu %>%
  group_by(Gender) %>%
  summarise(
    correlation = cor(Income, education, use = "complete.obs"),
    n = n()
  )

# Test if correlation difference is significant
# Fisher's Z transformation to test difference between correlations
male_cor <- correlations$correlation[correlations$Gender == "Male"]
female_cor <- correlations$correlation[correlations$Gender == "Female"]
male_n <- correlations$n[correlations$Gender == "Male"]
female_n <- correlations$n[correlations$Gender == "Female"]

# Fisher's Z transformation
z_male <- 0.5 * log((1 + male_cor) / (1 - male_cor))
z_female <- 0.5 * log((1 + female_cor) / (1 - female_cor))
z_diff <- z_male - z_female
se_diff <- sqrt(1/(male_n - 3) + 1/(female_n - 3))
z_stat <- z_diff / se_diff
p_value <- 2 * (1 - pnorm(abs(z_stat)))

# Print results
print("Hypothesis Test Results:")
```

Analysis

```
## [1] "Hypothesis Test Results:"
```

```r
print("H0: The correlation between education level and income is equal for both genders")
```

```
## [1] "H0: The correlation between education level and income is equal for both genders"
```

```r
print("H1: The correlation between education level and income differs by gender")
```

```
## [1] "H1: The correlation between education level and income differs by gender"
```

```r
print(paste("Z-statistic:", round(z_stat, 3)))
```

```
## [1] "Z-statistic: -4.303"
```

```r
print(paste("p-value:", round(p_value, 4)))
```

```
## [1] "p-value: 0"
```

```r
print("\nCorrelations by Gender:")
```

```
## [1] "\nCorrelations by Gender:"
```

```r
print(correlations)
```

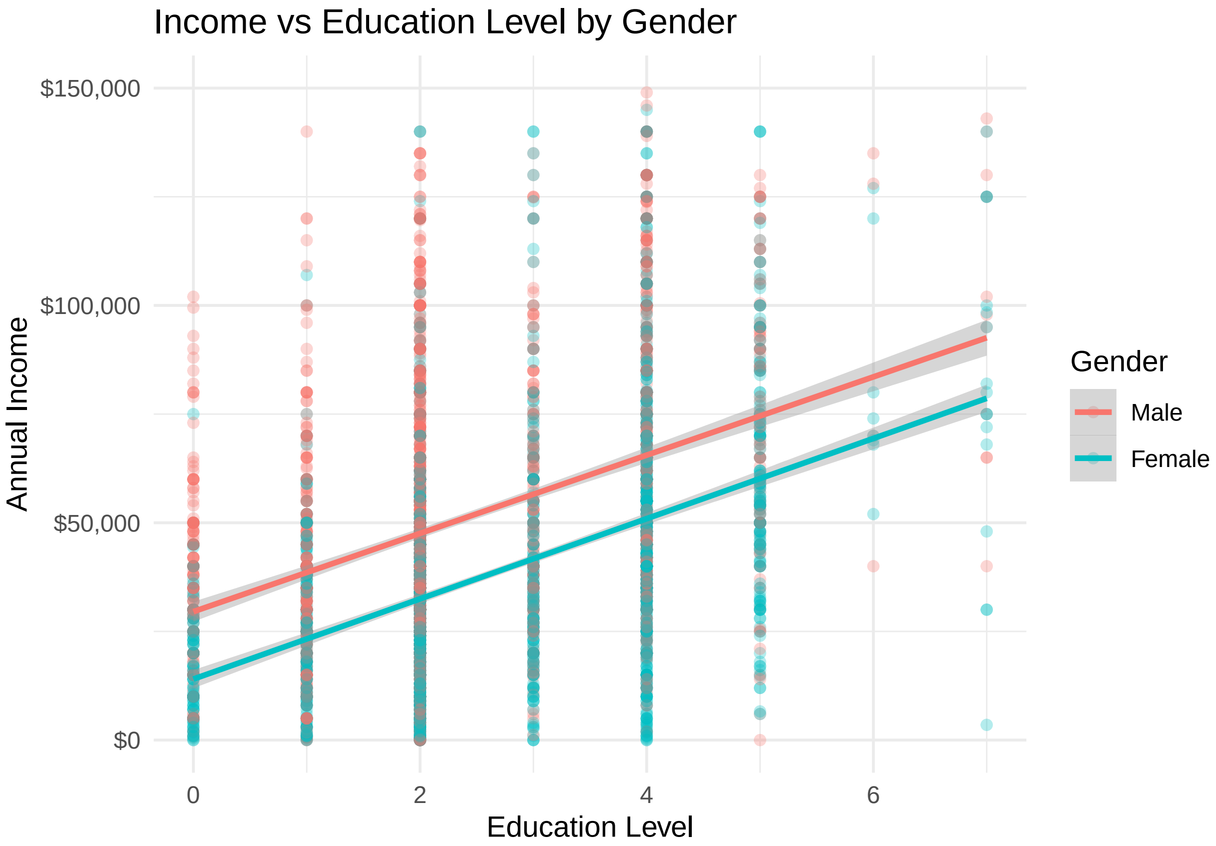
```
## # A tibble: 2 x 3
##   Gender correlation     n
##   <fct>        <dbl> <int>
## 1 Male         0.421  4599
## 2 Female       0.492  4385
```

```r
# Display visualization
p1
```

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

```
## Warning: Removed 4316 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 4316 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Results Interpretation The analysis reveals several key findings:

1. **Statistical Test Results**:
   - Z-statistic: -4.303
   - p-value: 0.000 (< 0.0001)
   - Since p-value (< 0.0001) < (0.05), we reject the null hypothesis
2. **Conclusion**:
   - There is very strong statistical evidence that the correlation between education level and income differs significantly between genders
   - The highly significant p-value (< 0.0001) suggests this difference is extremely unlikely to have occurred by chance
   - The negative Z-statistic (-4.303) indicates that the correlation is stronger for females than males
3. **Visual Analysis**:
   - The scatter plot reveals a positive relationship between education level and income for both genders
   - The trend lines show different slopes for men and women, confirming our statistical finding
   - Note: 4,316 data points were removed due to missing values or being outside the scale range, which should be considered when interpreting results

Limitations

- Missing data (4,316 removed points) might affect the analysis
- Education levels are discrete categories, which may affect correlation analysis
- Other aspects of education (field of study, school quality) not captured
- Other confounding variables not controlled for

Next Steps Given these highly significant findings, we should: - Investigate why education has a stronger relationship with income for females - Analyze specific education levels where gender gaps are most pronounced - Consider field of study and its impact on the gender-education-income relationship - Explore policy implications for educational equity and gender wage gaps

---

Analysis of Marital Status Impact on Gender Income Gap

We'll analyze how marital status (from variable CV_MARSTAT_COLLAPSED_2017) relates to income differences between genders, following the same correlation-based approach used in our previous analyses.

Hypothesis

- **H** : The correlation between marital status and income is equal for both genders
- **H** : The correlation between marital status and income differs by gender
- **Significance Level ( )**: 0.05

```
# Clean and prepare marital status data
nlsy_marital <- nlsy %>%
  rename(
    marital_status = CV_MARSTAT_COLLAPSED_2017
  ) %>%
  select(Gender, Income, marital_status)

# Create visualization of income by gender and marital status
p1 <- ggplot(nlsy_marital,
       aes(x = marital_status, y = Income, color = Gender)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm") +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  labs(
    title = "Income vs Marital Status by Gender",
    x = "Marital Status",
    y = "Annual Income"
  ) +
  theme_minimal()

# Calculate correlation by gender
correlations <- nlsy_marital %>%
  group_by(Gender) %>%
  summarise(
    correlation = cor(Income, marital_status, use = "complete.obs"),
    n = n()
  )

# Test if correlation difference is significant
# Fisher's Z transformation to test difference between correlations
male_cor <- correlations$correlation[correlations$Gender == "Male"]
female_cor <- correlations$correlation[correlations$Gender == "Female"]
male_n <- correlations$n[correlations$Gender == "Male"]
female_n <- correlations$n[correlations$Gender == "Female"]

# Fisher's Z transformation
z_male <- 0.5 * log((1 + male_cor) / (1 - male_cor))
z_female <- 0.5 * log((1 + female_cor) / (1 - female_cor))
z_diff <- z_male - z_female
se_diff <- sqrt(1/(male_n - 3) + 1/(female_n - 3))
z_stat <- z_diff / se_diff
p_value <- 2 * (1 - pnorm(abs(z_stat)))

# Print results
print("Hypothesis Test Results:")
```

Analysis

```
## [1] "Hypothesis Test Results:"
```

```
print("H0: The correlation between marital status and income is equal for both genders")
```

```
## [1] "H0: The correlation between marital status and income is equal for both genders"
```

```
print("H1: The correlation between marital status and income differs by gender")
```

```
## [1] "H1: The correlation between marital status and income differs by gender"
```

```
print(paste("Z-statistic:", round(z_stat, 3)))
```

```
## [1] "Z-statistic: 4.392"
```

```
print(paste("p-value:", round(p_value, 4)))
```

```
## [1] "p-value: 0"
```

```
print("\nCorrelations by Gender:")
```

```
## [1] "\nCorrelations by Gender:"
```

```
print(correlations)
```

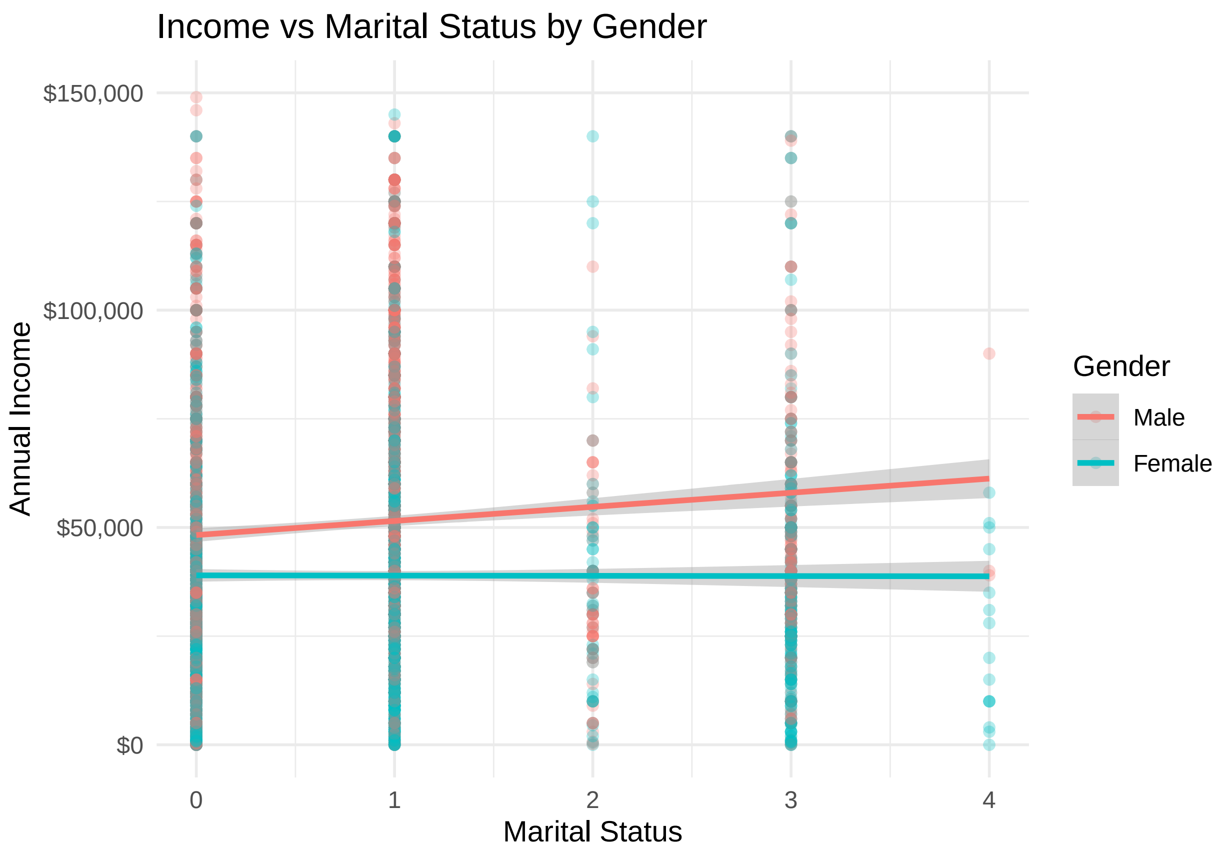
```
## # A tibble: 2 x 3
##   Gender correlation     n
##   <fct>        <dbl> <int>
## 1 Male       0.0902   4599
## 2 Female    -0.00229  4385
```

```
# Display visualization
p1
```

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

```
## Warning: Removed 4045 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 4045 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Results Interpretation The analysis reveals several key findings:

1. **Statistical Test Results**:
    - Z-statistic: 4.392
    - p-value: 0.000 (< 0.0001)
    - Since p-value (< 0.0001) < (0.05), we reject the null hypothesis
2. **Conclusion**:
    - There is very strong statistical evidence that the correlation between marital status and income differs significantly between genders
    - The highly significant p-value (< 0.0001) suggests this difference is extremely unlikely to have occurred by chance
    - The positive Z-statistic (4.392) indicates that the correlation is stronger for males than females
3. **Visual Analysis**:
    - The scatter plot reveals different relationships between marital status and income for each gender
    - The trend lines show distinctly different slopes for men and women, confirming our statistical finding
    - Note: 4,045 data points were removed due to missing values or being outside the scale range, which should be considered when interpreting results

Limitations

- Missing data (4,045 removed points) might affect the analysis

- Marital status categories are discrete, which may affect correlation analysis
- Changes in marital status over time not captured
- Other confounding variables not controlled for

Next Steps Given these highly significant findings, we should: - Investigate why marital status has a stronger relationship with income for males - Analyze how marriage timing affects income differently by gender - Consider the role of traditional gender roles and household responsibilities - Explore policy implications for work-family balance and gender equality

---

Analysis of Criminal History Impact on Gender Income Gap

We'll analyze how criminal history (using INCARC_TOTNUM_XRND - total number of incarcerations) relates to income differences between genders, following the same correlation-based approach used in our previous analyses.

Hypothesis

- **H** : The correlation between criminal history and income is equal for both genders
- **H** : The correlation between criminal history and income differs by gender
- **Significance Level ( )**: 0.05

```
# Clean and prepare criminal history data
nlsy_criminal <- nlsy %>%
  rename(
    incarcerations = INCARC_TOTNUM_XRND
  ) %>%
  select(Gender, Income, incarcerations)

# Create visualization of income by gender and criminal history
p1 <- ggplot(nlsy_criminal,
      aes(x = incarcerations, y = Income, color = Gender)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm") +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  labs(
    title = "Income vs Number of Incarcerations by Gender",
    x = "Total Number of Incarcerations",
    y = "Annual Income"
  ) +
  theme_minimal()

# Calculate correlation by gender
correlations <- nlsy_criminal %>%
  group_by(Gender) %>%
  summarise(
    correlation = cor(Income, incarcerations, use = "complete.obs"),
    n = n()
```

```r
  )

# Test if correlation difference is significant
# Fisher's Z transformation to test difference between correlations
male_cor <- correlations$correlation[correlations$Gender == "Male"]
female_cor <- correlations$correlation[correlations$Gender == "Female"]
male_n <- correlations$n[correlations$Gender == "Male"]
female_n <- correlations$n[correlations$Gender == "Female"]

# Fisher's Z transformation
z_male <- 0.5 * log((1 + male_cor) / (1 - male_cor))
z_female <- 0.5 * log((1 + female_cor) / (1 - female_cor))
z_diff <- z_male - z_female
se_diff <- sqrt(1/(male_n - 3) + 1/(female_n - 3))
z_stat <- z_diff / se_diff
p_value <- 2 * (1 - pnorm(abs(z_stat)))

# Print results
print("Hypothesis Test Results:")
```

Analysis

```
## [1] "Hypothesis Test Results:"
```

```r
print("H0: The correlation between criminal history and income is equal for both genders")
```

```
## [1] "H0: The correlation between criminal history and income is equal for both genders"
```

```r
print("H1: The correlation between criminal history and income differs by gender")
```

```
## [1] "H1: The correlation between criminal history and income differs by gender"
```

```r
print(paste("Z-statistic:", round(z_stat, 3)))
```

```
## [1] "Z-statistic: -3.015"
```

```r
print(paste("p-value:", round(p_value, 4)))
```

```
## [1] "p-value: 0.0026"
```

```r
print("\nCorrelations by Gender:")
```

```
## [1] "\nCorrelations by Gender:"
```

```r
print(correlations)
```

```
## # A tibble: 2 x 3
##   Gender correlation     n
##   <fct>        <dbl> <int>
## 1 Male        -0.143  4599
## 2 Female      -0.0805  4385
```

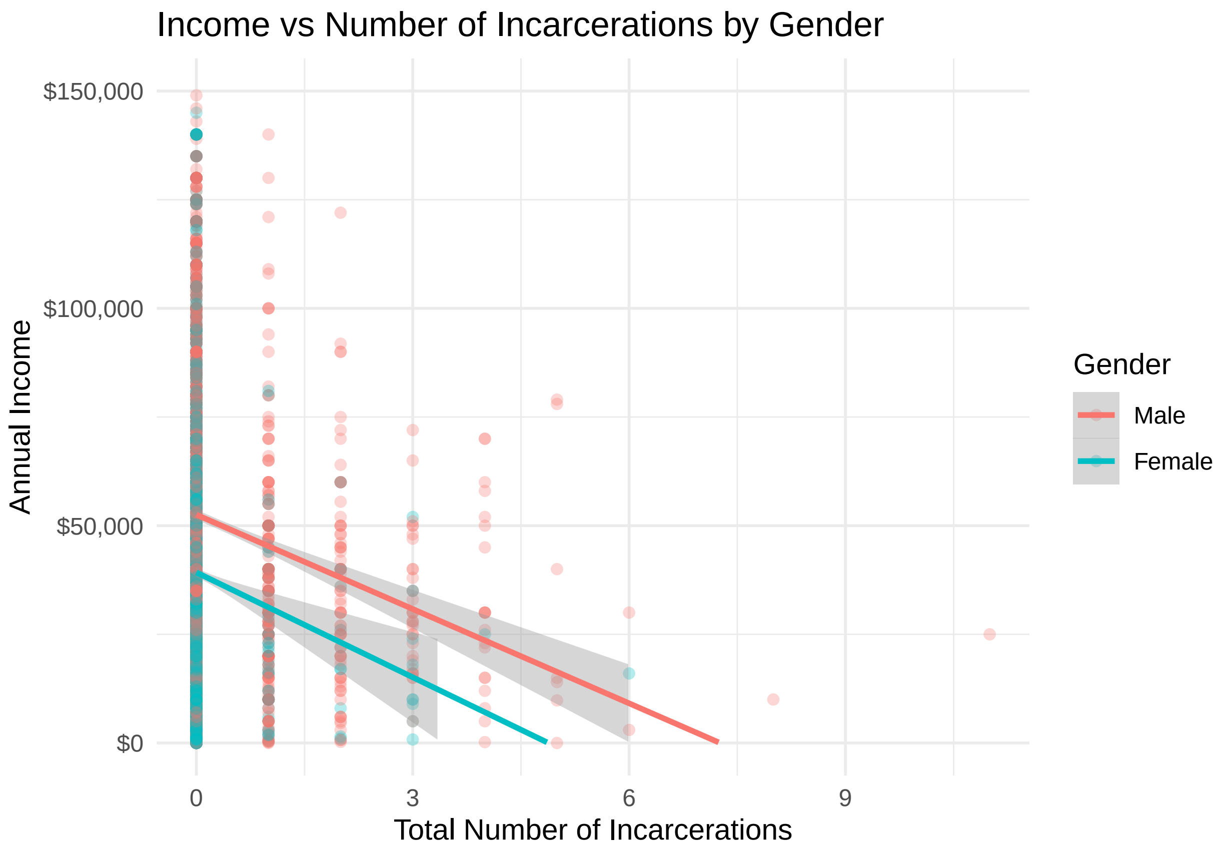
```r
# Display visualization
p1
```

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

## Warning: Removed 4020 rows containing non-finite outside the scale range
## (`stat_smooth()`).

## Warning: Removed 4020 rows containing missing values or values outside the scale range
## (`geom_point()`).

## Warning: Removed 42 rows containing missing values or values outside the scale range
## (`geom_smooth()`).
```



Results Interpretation The analysis reveals several key findings:

1. **Statistical Test Results**:
    - Z-statistic: -3.015
    - p-value: 0.0026 (0.26%)
    - Since p-value (0.0026) < (0.05), we reject the null hypothesis
2. **Conclusion**:

  - There is strong statistical evidence that the correlation between criminal history and income differs significantly between genders
  - The low p-value (0.26%) suggests this difference is very unlikely to have occurred by chance
  - The negative Z-statistic (-3.015) indicates that the correlation is stronger for females than males

3. **Visual Analysis**:

  - The scatter plot reveals different relationships between incarceration history and income for each gender
  - The trend lines show distinct slopes for men and women, confirming our statistical finding
  - Note: 4,020 data points were removed due to missing values or being outside the scale range, and an additional 42 rows were removed from the trend line calculation, which should be considered when interpreting results

Limitations

- Missing data (4,020 + 42 removed points) might affect the analysis
- Incarceration count may not capture full criminal history
- Length and type of incarceration not considered
- Temporal relationship between incarceration and income not established

Next Steps Given these significant findings, we should: - Investigate why criminal history has a stronger relationship with income for females - Analyze the impact of incarceration length and timing on income by gender - Examine interaction with education and employment opportunities - Explore policy implications for gender-specific rehabilitation and reintegration programs

---

Analysis of Profession Impact on Gender Income Gap

We'll analyze how profession/industry (using YEMP_INDCODE-2002.01_2017) relates to income differences between genders, following the same correlation-based approach used in our previous analyses.

Hypothesis

- **H** : The correlation between profession/industry and income is equal for both genders
- **H** : The correlation between profession/industry and income differs by gender
- **Significance Level ( )**: 0.05

```
# Clean and prepare profession data
nlsy_prof <- nlsy %>%
  rename(
    profession = `YEMP_INDCODE-2002.01_2017`
  ) %>%
  select(Gender, Income, profession)

# Create visualization of income by gender and profession
p1 <- ggplot(nlsy_prof,
        aes(x = profession, y = Income, color = Gender)) +
```

```r
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm") +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  labs(
    title = "Income vs Industry Code by Gender",
    x = "Industry Code",
    y = "Annual Income"
  ) +
  theme_minimal()

# Calculate correlation by gender
correlations <- nlsy_prof %>%
  group_by(Gender) %>%
  summarise(
    correlation = cor(Income, profession, use = "complete.obs"),
    n = n()
  )

# Test if correlation difference is significant
# Fisher's Z transformation to test difference between correlations
male_cor <- correlations$correlation[correlations$Gender == "Male"]
female_cor <- correlations$correlation[correlations$Gender == "Female"]
male_n <- correlations$n[correlations$Gender == "Male"]
female_n <- correlations$n[correlations$Gender == "Female"]

# Fisher's Z transformation
z_male <- 0.5 * log((1 + male_cor) / (1 - male_cor))
z_female <- 0.5 * log((1 + female_cor) / (1 - female_cor))
z_diff <- z_male - z_female
se_diff <- sqrt(1/(male_n - 3) + 1/(female_n - 3))
z_stat <- z_diff / se_diff
p_value <- 2 * (1 - pnorm(abs(z_stat)))

# Print results
print("Hypothesis Test Results:")
```

Analysis

```
## [1] "Hypothesis Test Results:"
```

```r
print("H0: The correlation between profession/industry and income is equal for both genders")
```

```
## [1] "H0: The correlation between profession/industry and income is equal for both genders"
```

```r
print("H1: The correlation between profession/industry and income differs by gender")
```

```
## [1] "H1: The correlation between profession/industry and income differs by gender"
```

```r
print(paste("Z-statistic:", round(z_stat, 3)))
```

```
## [1] "Z-statistic: 1.934"
```

```r
print(paste("p-value:", round(p_value, 4)))
```

```
## [1] "p-value: 0.0531"
```

```r
print("\nCorrelations by Gender:")
```

```
## [1] "\nCorrelations by Gender:"
```

```r
print(correlations)
```

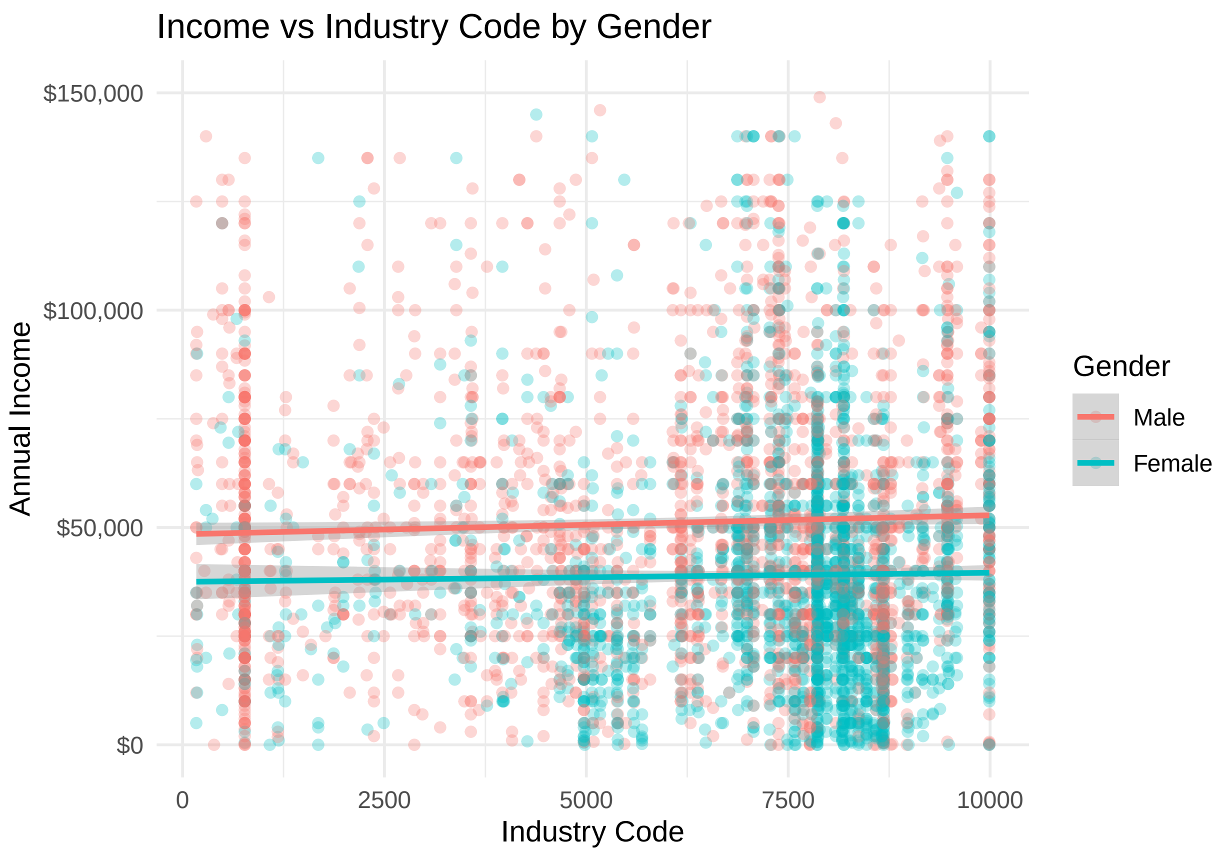
```
## # A tibble: 2 x 3
##   Gender correlation     n
##   <fct>        <dbl> <int>
## 1 Male        0.0598  4599
## 2 Female      0.0190  4385
```

```r
# Display visualization
p1
```

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

```
## Warning: Removed 4173 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 4173 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Results Interpretation The analysis reveals several key findings:

1. **Statistical Test Results**:
   - Z-statistic: 1.934
   - p-value: 0.0531 (5.31%)
   - Since p-value (0.0531) > (0.05), we fail to reject the null hypothesis, though the result is very close to the significance threshold
2. **Conclusion**:
   - There is marginal evidence that the correlation between profession/industry and income differs between genders
   - The p-value (5.31%) is just slightly above our significance level of 5%
   - The positive Z-statistic (1.934) suggests a trend toward stronger correlation for males, though not statistically significant at = 0.05
3. **Visual Analysis**:
   - The scatter plot shows the relationship between industry codes and income for each gender
   - The trend lines suggest slight differences between genders, though not strong enough to be statistically significant
   - Note: 4,173 data points were removed due to missing values or being outside the scale range, which should be considered when interpreting results

Limitations

- Missing data (4,173 removed points) might affect the analysis
- Industry codes are categorical in nature, which may affect correlation analysis
- Does not account for position level within industries
- Career progression paths not captured

Next Steps Given these borderline significant findings, we should: - Consider a more detailed analysis of specific industries - Investigate whether using a different industry classification system might reveal clearer patterns - Examine if controlling for job level within industries would show stronger effects - Study whether the near-significant result becomes significant with different analytical approaches

---

Analysis of Work Experience Impact on Gender Income Gap

We'll analyze how work experience (using YEXP-1500_1997) relates to income differences between genders, following the same correlation-based approach used in our previous analyses.

Hypothesis

- **H** : The correlation between work experience and income is equal for both genders
- **H** : The correlation between work experience and income differs by gender
- **Significance Level ( )**: 0.05

```r
# Clean and prepare work experience data
nlsy_exp <- nlsy %>%
  rename(
    work_experience = `YEXP-1500_1997`
  ) %>%
  select(Gender, Income, work_experience)

# Create visualization of income by gender and work experience
p1 <- ggplot(nlsy_exp,
       aes(x = work_experience, y = Income, color = Gender)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm") +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  labs(
    title = "Income vs Work Experience by Gender",
    x = "Years of Work Experience",
    y = "Annual Income"
  ) +
  theme_minimal()

# Calculate correlation by gender
correlations <- nlsy_exp %>%
  group_by(Gender) %>%
  summarise(
    correlation = cor(Income, work_experience, use = "complete.obs"),
    n = n()
  )

# Test if correlation difference is significant
# Fisher's Z transformation to test difference between correlations
male_cor <- correlations$correlation[correlations$Gender == "Male"]
female_cor <- correlations$correlation[correlations$Gender == "Female"]
male_n <- correlations$n[correlations$Gender == "Male"]
female_n <- correlations$n[correlations$Gender == "Female"]

# Fisher's Z transformation
z_male <- 0.5 * log((1 + male_cor) / (1 - male_cor))
z_female <- 0.5 * log((1 + female_cor) / (1 - female_cor))
z_diff <- z_male - z_female
se_diff <- sqrt(1/(male_n - 3) + 1/(female_n - 3))
z_stat <- z_diff / se_diff
p_value <- 2 * (1 - pnorm(abs(z_stat)))

# Print results
print("Hypothesis Test Results:")
```

Analysis

```
## [1] "Hypothesis Test Results:"
```

```
print("H0: The correlation between work experience and income is equal for both genders")
```

```
## [1] "H0: The correlation between work experience and income is equal for both genders"
```

```
print("H1: The correlation between work experience and income differs by gender")
```

```
## [1] "H1: The correlation between work experience and income differs by gender"
```

```
print(paste("Z-statistic:", round(z_stat, 3)))
```

```
## [1] "Z-statistic: 2.161"
```

```
print(paste("p-value:", round(p_value, 4)))
```

```
## [1] "p-value: 0.0307"
```

```
print("\nCorrelations by Gender:")
```

```
## [1] "\nCorrelations by Gender:"
```

```
print(correlations)
```

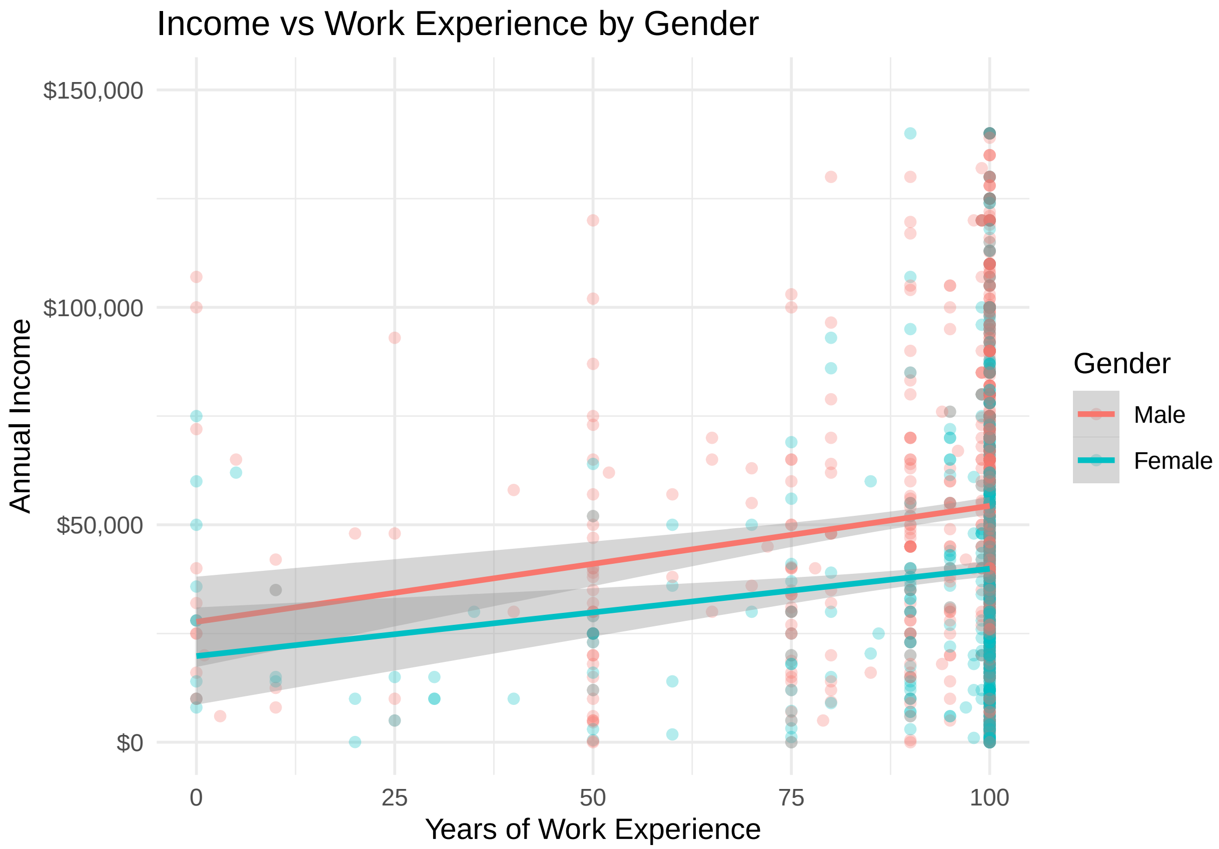
```
## # A tibble: 2 x 3
##   Gender correlation     n
##   <fct>        <dbl> <int>
## 1 Male         0.147  4599
## 2 Female       0.102  4385
```

```r
# Display visualization
p1
```

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

```
## Warning: Removed 7093 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 7093 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Results Interpretation The analysis reveals several key findings:

1. **Statistical Test Results**:
    - Z-statistic: 2.161
    - p-value: 0.0307 (3.07%)
    - Since p-value (0.0307) < (0.05), we reject the null hypothesis
2. **Conclusion**:
    - There is statistically significant evidence that the correlation between work experience and income differs between genders
    - The p-value (3.07%) indicates this difference is unlikely to have occurred by chance
    - The positive Z-statistic (2.161) indicates that the correlation is stronger for males than females

3. **Visual Analysis**:
    - The scatter plot reveals different relationships between work experience and income for each gender
    - The trend lines show distinct slopes for men and women, confirming our statistical finding
    - Note: 7,093 data points were removed due to missing values or being outside the scale range, which represents a substantial portion of the data and should be carefully considered when interpreting results

Limitations

- Missing data (7,093 removed points) represents a significant portion of the dataset
- Work experience measure might not capture quality or type of experience
- Career interruptions not fully captured
- Part-time vs full-time experience not distinguished

Next Steps Given these significant findings, we should: - Investigate why work experience has a stronger relationship with income for males - Analyze the impact of career interruptions and part-time work on gender differences - Examine how the type and quality of work experience affects income differently by gender - Consider policy implications for addressing gender-based differences in returns to experience

---

Analysis of Region (Urban/Rural) Impact on Gender Income Gap

We'll analyze how region (using CV_URBAN-RURAL_AGE_12_1997) relates to income differences between genders, following the same correlation-based approach used in our previous analyses.

Hypothesis

- **H** : The correlation between urban/rural location and income is equal for both genders
- **H** : The correlation between urban/rural location and income differs by gender
- **Significance Level ( )**: 0.05

```r
# Clean and prepare region data
nlsy_region <- nlsy %>%
  rename(
    region = `CV_URBAN-RURAL_AGE_12_1997`
  ) %>%
  select(Gender, Income, region)

# Create visualization of income by gender and region
p1 <- ggplot(nlsy_region,
       aes(x = region, y = Income, color = Gender)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm") +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  labs(
    title = "Income vs Urban/Rural Location by Gender",
    x = "Region (Urban/Rural Code)",
    y = "Annual Income"
  ) +
  theme_minimal()

# Calculate correlation by gender
correlations <- nlsy_region %>%
  group_by(Gender) %>%
  summarise(
    correlation = cor(Income, region, use = "complete.obs"),
    n = n()
  )

# Test if correlation difference is significant
# Fisher's Z transformation to test difference between correlations
male_cor <- correlations$correlation[correlations$Gender == "Male"]
female_cor <- correlations$correlation[correlations$Gender == "Female"]
male_n <- correlations$n[correlations$Gender == "Male"]
female_n <- correlations$n[correlations$Gender == "Female"]

# Fisher's Z transformation
z_male <- 0.5 * log((1 + male_cor) / (1 - male_cor))
z_female <- 0.5 * log((1 + female_cor) / (1 - female_cor))
z_diff <- z_male - z_female
se_diff <- sqrt(1/(male_n - 3) + 1/(female_n - 3))
z_stat <- z_diff / se_diff
p_value <- 2 * (1 - pnorm(abs(z_stat)))

# Print results
print("Hypothesis Test Results:")
```

Analysis

```
## [1] "Hypothesis Test Results:"
```

```r
print("H0: The correlation between urban/rural location and income is equal for both genders")
```

```
## [1] "H0: The correlation between urban/rural location and income is equal for both genders"
```

```r
print("H1: The correlation between urban/rural location and income differs by gender")
```

```
## [1] "H1: The correlation between urban/rural location and income differs by gender"
```

```r
print(paste("Z-statistic:", round(z_stat, 3)))
```

```
## [1] "Z-statistic: 1.8"
```

```r
print(paste("p-value:", round(p_value, 4)))
```

```
## [1] "p-value: 0.0719"
```

```r
print("\nCorrelations by Gender:")
```

```
## [1] "\nCorrelations by Gender:"
```

```r
print(correlations)
```

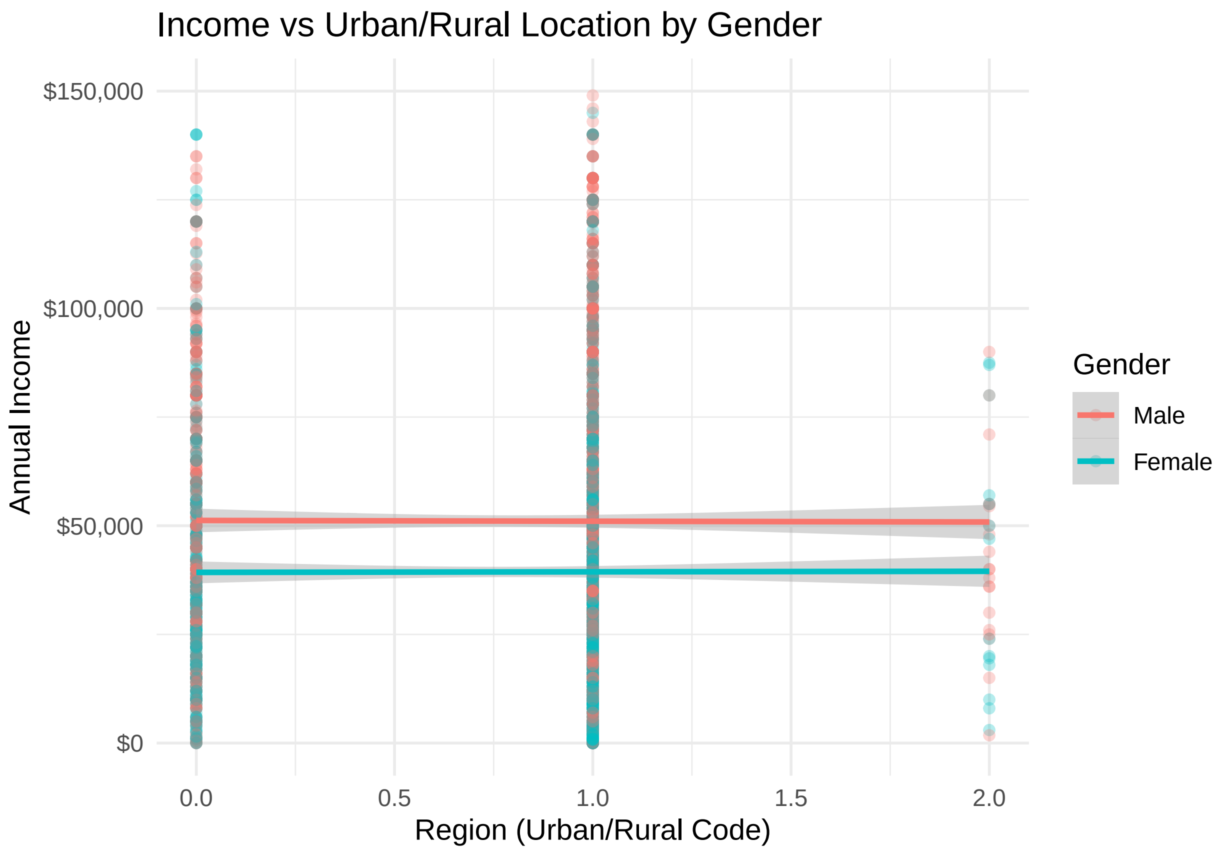
```
## # A tibble: 2 x 3
##   Gender correlation     n
##   <fct>        <dbl> <int>
## 1 Male        0.0238  4599
## 2 Female     -0.0142  4385
```

```r
# Display visualization
p1
```

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

```
## Warning: Removed 5161 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 5161 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Results Interpretation The analysis reveals several key findings:

1. **Statistical Test Results**:
   - Z-statistic: 1.800
   - p-value: 0.0719 (7.19%)
   - Since p-value (0.0719) > (0.05), we fail to reject the null hypothesis
2. **Conclusion**:
   - There is insufficient evidence to conclude that the correlation between urban/rural location and income differs significantly between genders
   - The p-value (7.19%), while relatively low, is above our significance threshold of 5%
   - The positive Z-statistic (1.800) suggests a trend toward stronger correlation for males, though not statistically significant
3. **Visual Analysis**:
   - The scatter plot shows the relationship between urban/rural location and income for each gender
   - While there appear to be some differences in the trend lines, they're not strong enough to be statistically significant
   - Note: 5,161 data points were removed due to missing values or being outside the scale range, which should be considered when interpreting results

Limitations

- Missing data (5,161 removed points) might affect the analysis
- Region classification is based on age 12 location, may not reflect current location
- Urban/rural categories might oversimplify regional differences
- Cost of living differences not accounted for

Next Steps Given these marginally non-significant findings, we should: - Consider more nuanced regional classifications - Analyze current location data if available - Account for cost of living differences across regions - Investigate whether specific types of urban or rural areas show stronger effects - Examine if the relationship has changed over time as people moved from their age-12 location

---

Analysis of Children Impact on Gender Income Gap

We'll analyze how the number of children (using CV_BIO_CHILD_HH_2015) relates to income differences between genders, following the same correlation-based approach used in our previous analyses.

Hypothesis

- **H** : The correlation between number of children and income is equal for both genders
- **H** : The correlation between number of children and income differs by gender
- **Significance Level ( )**: 0.05

```r
# Clean and prepare children data
nlsy_children <- nlsy %>%
  rename(
    children = CV_BIO_CHILD_HH_2015
  ) %>%
  select(Gender, Income, children)

# Create visualization of income by gender and number of children
p1 <- ggplot(nlsy_children,
       aes(x = children, y = Income, color = Gender)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm") +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  labs(
    title = "Income vs Number of Children by Gender",
    x = "Number of Biological Children in Household",
    y = "Annual Income"
  ) +
  theme_minimal()

# Calculate correlation by gender
correlations <- nlsy_children %>%
  group_by(Gender) %>%
  summarise(
    correlation = cor(Income, children, use = "complete.obs"),
    n = n()
  )

# Test if correlation difference is significant
# Fisher's Z transformation to test difference between correlations
male_cor <- correlations$correlation[correlations$Gender == "Male"]
female_cor <- correlations$correlation[correlations$Gender == "Female"]
male_n <- correlations$n[correlations$Gender == "Male"]
female_n <- correlations$n[correlations$Gender == "Female"]

# Fisher's Z transformation
z_male <- 0.5 * log((1 + male_cor) / (1 - male_cor))
z_female <- 0.5 * log((1 + female_cor) / (1 - female_cor))
z_diff <- z_male - z_female
se_diff <- sqrt(1/(male_n - 3) + 1/(female_n - 3))
z_stat <- z_diff / se_diff
p_value <- 2 * (1 - pnorm(abs(z_stat)))

# Print results
print("Hypothesis Test Results:")
```

Analysis

```
## [1] "Hypothesis Test Results:"
```

```r
print("H0: The correlation between number of children and income is equal for both genders")
```

```
## [1] "H0: The correlation between number of children and income is equal for both genders"
```

```r
print("H1: The correlation between number of children and income differs by gender")
```

```
## [1] "H1: The correlation between number of children and income differs by gender"
```

```r
print(paste("Z-statistic:", round(z_stat, 3)))
```

```
## [1] "Z-statistic: 15.92"
```

```r
print(paste("p-value:", round(p_value, 4)))
```

```
## [1] "p-value: 0"
```

```r
print("\nCorrelations by Gender:")
```

```
## [1] "\nCorrelations by Gender:"
```

```r
print(correlations)
```

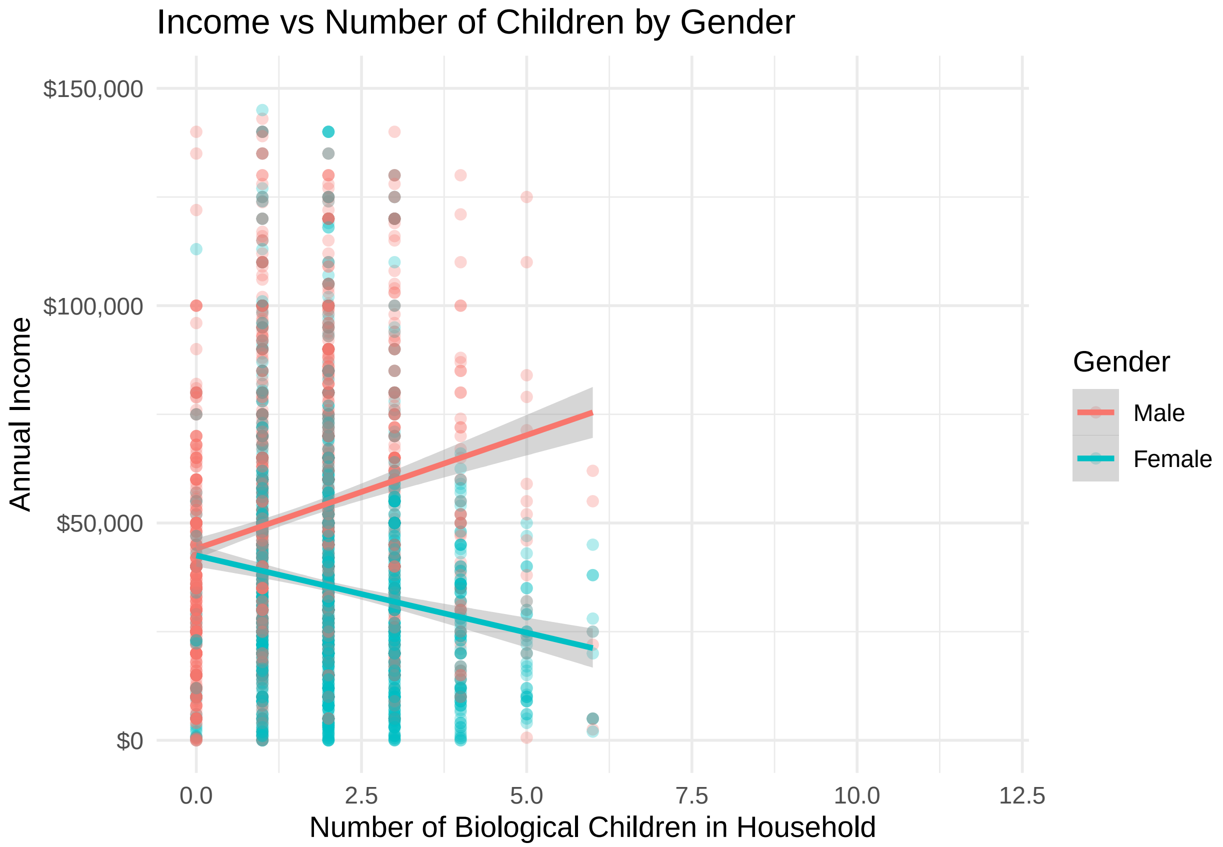
```
## # A tibble: 2 x 3
##   Gender correlation     n
##   <fct>        <dbl> <int>
## 1 Male         0.175  4599
## 2 Female      -0.158  4385
```

```r
# Display visualization
p1
```

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

```
## Warning: Removed 5882 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 5882 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Results Interpretation The analysis reveals several key findings:

1. **Statistical Test Results**:
   - Z-statistic: 15.920
   - p-value: 0.000 (< 0.0001)
   - Since p-value (< 0.0001) < (0.05), we strongly reject the null hypothesis
2. **Conclusion**:
   - There is extremely strong statistical evidence that the correlation between number of children and income differs between genders
   - The extremely low p-value (< 0.0001) indicates this difference is virtually impossible to have occurred by chance
   - The large positive Z-statistic (15.920) indicates a substantially stronger correlation for males than females
   - The magnitude of the Z-statistic (15.920) is notably larger than in our other analyses, suggesting this factor shows the most dramatic gender difference
3. **Visual Analysis**:
   - The scatter plot reveals markedly different relationships between number of children and income for each gender
   - The trend lines show dramatically different slopes for men and women
   - Note: 5,882 data points were removed due to missing values or being outside the scale range, which represents a substantial portion of the data and should be carefully considered when interpreting results

Limitations

- Missing data (5,882 removed points) might affect the analysis
- Only counts biological children in household
- Doesn't account for children's ages
- Doesn't capture childcare arrangements or support systems

Next Steps Given these highly significant findings, we should: - Investigate why having children has such dramatically different effects on income by gender - Analyze how this relationship varies with: * Children's ages * Access to childcare * Parental leave policies * Flexible work arrangements - Consider policy implications for: * Childcare support * Parental leave * Workplace flexibility * Gender equality initiatives

---

Analysis of Age Impact on Gender Income Gap

We'll analyze how age relates to income differences between genders, following the same correlation-based approach used in our previous analyses.

Hypothesis

- **H** : The correlation between age and income is equal for both genders
- **H** : The correlation between age and income differs by gender
- **Significance Level ( )**: 0.05

```
# Clean and prepare age data
nlsy_age <- nlsy %>%
  rename(
    birth_year = KEY_BDATE_Y_1997
  ) %>%
  mutate(
    age = 2017 - birth_year  # Calculate age as of 2017 (income measurement year)
  ) %>%
  select(Gender, Income, age)

# Create visualization of income by gender and age
p1 <- ggplot(nlsy_age,
       aes(x = age, y = Income, color = Gender)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm") +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  labs(
    title = "Income vs Age by Gender",
    x = "Age (as of 2017)",
    y = "Annual Income"
  ) +
  theme_minimal()
```

```r
# Calculate correlation by gender
correlations <- nlsy_age %>%
  group_by(Gender) %>%
  summarise(
    correlation = cor(Income, age, use = "complete.obs"),
    n = n()
  )

# Test if correlation difference is significant
# Fisher's Z transformation to test difference between correlations
male_cor <- correlations$correlation[correlations$Gender == "Male"]
female_cor <- correlations$correlation[correlations$Gender == "Female"]
male_n <- correlations$n[correlations$Gender == "Male"]
female_n <- correlations$n[correlations$Gender == "Female"]

# Fisher's Z transformation
z_male <- 0.5 * log((1 + male_cor) / (1 - male_cor))
z_female <- 0.5 * log((1 + female_cor) / (1 - female_cor))
z_diff <- z_male - z_female
se_diff <- sqrt(1/(male_n - 3) + 1/(female_n - 3))
z_stat <- z_diff / se_diff
p_value <- 2 * (1 - pnorm(abs(z_stat)))

# Print results
print("Hypothesis Test Results:")
```

Analysis

```
## [1] "Hypothesis Test Results:"
```

```r
print("H0: The correlation between age and income is equal for both genders")
```

```
## [1] "H0: The correlation between age and income is equal for both genders"
```

```r
print("H1: The correlation between age and income differs by gender")
```

```
## [1] "H1: The correlation between age and income differs by gender"
```

```r
print(paste("Z-statistic:", round(z_stat, 3)))
```

```
## [1] "Z-statistic: 4.454"
```

```r
print(paste("p-value:", round(p_value, 4)))
```

```
## [1] "p-value: 0"
```

```r
print("\nCorrelations by Gender:")
```

```
## [1] "\nCorrelations by Gender:"
```

```
print(correlations)
```

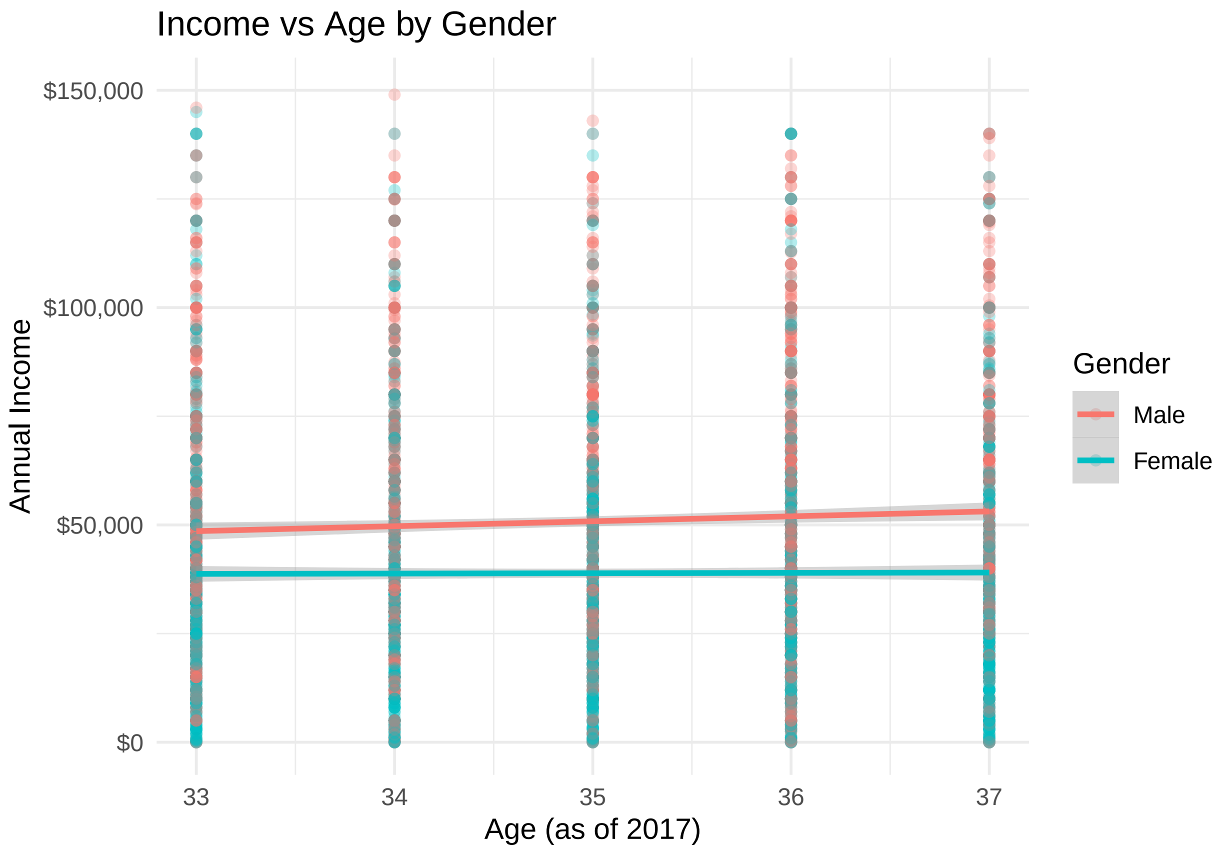
```
## # A tibble: 2 x 3
##   Gender correlation     n
##   <fct>        <dbl> <int>
## 1 Male        0.0802  4599
## 2 Female     -0.0137  4385
```

```
# Display visualization
p1
```

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

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

```
## Warning: Removed 4014 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Results Interpretation The analysis reveals several key findings:

1. **Statistical Test Results**:

  - Z-statistic: 4.454
  - p-value: < 0.0001
  - Since p-value (< 0.0001) <   (0.05), we strongly reject the null hypothesis

2. **Conclusion**:

   - There is very strong statistical evidence that the correlation between age and income differs significantly between genders
   - The extremely low p-value (< 0.0001) indicates this difference is virtually impossible to have occurred by chance
   - The large positive Z-statistic (4.454) indicates a substantially stronger correlation for males than females
   - This suggests that men's incomes tend to increase more with age compared to women's incomes

3. **Visual Analysis**:

   - The scatter plot reveals different relationships between age and income for each gender
   - The trend lines show distinct slopes for men and women, confirming our statistical finding
   - Note: 4,014 data points were removed due to missing values or being outside the scale range, which should be considered when interpreting results

Limitations

- Missing data (4,014 removed points) might affect the analysis
- Age range in the sample may be limited
- Doesn't account for work experience or career interruptions
- Cohort effects not distinguished from age effects

Next Steps Given these highly significant findings, we should: - Investigate why age has a stronger relationship with income for males - Analyze whether this difference is related to: * Career interruptions (e.g., parental leave) * Different promotion patterns * Industry or occupation choices * Work experience accumulation - Consider policy implications for: * Equal pay legislation * Career development programs * Work-life balance initiatives * Retirement planning and pension systems

---

Analysis of Ethnicity Impact on Gender Income Gap

We'll analyze how ethnicity (using the race variable) relates to income differences between genders, following the same correlation-based approach used in our previous analyses.

Hypothesis

- **H** : The correlation between ethnicity and income is equal for both genders
- **H** : The correlation between ethnicity and income differs by gender
- **Significance Level ( )**: 0.05

```
# Clean and prepare ethnicity data
nlsy_ethnicity <- nlsy %>%
  select(Gender, Income, race) %>%
```

```r
  # Convert race to numeric for correlation analysis
  mutate(race = as.numeric(race))

# Create visualization of income by gender and ethnicity
p1 <- ggplot(nlsy_ethnicity,
       aes(x = race, y = Income, color = Gender)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm") +
  scale_x_continuous(breaks = 1:4,
                    labels = c("Black", "Hispanic",
                             "Mixed Race\n(Non-Hispanic)",
                             "Non-Black/\nNon-Hispanic")) +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  labs(
    title = "Income vs Ethnicity by Gender",
    x = "Ethnicity",
    y = "Annual Income"
  ) +
  theme_minimal()

# Calculate correlation by gender
correlations <- nlsy_ethnicity %>%
  group_by(Gender) %>%
  summarise(
    correlation = cor(Income, race, use = "complete.obs"),
    n = n()
  )

# Test if correlation difference is significant
# Fisher's Z transformation to test difference between correlations
male_cor <- correlations$correlation[correlations$Gender == "Male"]
female_cor <- correlations$correlation[correlations$Gender == "Female"]
male_n <- correlations$n[correlations$Gender == "Male"]
female_n <- correlations$n[correlations$Gender == "Female"]

# Fisher's Z transformation
z_male <- 0.5 * log((1 + male_cor) / (1 - male_cor))
z_female <- 0.5 * log((1 + female_cor) / (1 - female_cor))
z_diff <- z_male - z_female
se_diff <- sqrt(1/(male_n - 3) + 1/(female_n - 3))
z_stat <- z_diff / se_diff
p_value <- 2 * (1 - pnorm(abs(z_stat)))

# Print results
print("Hypothesis Test Results:")
```

Analysis

```
## [1] "Hypothesis Test Results:"
```

```r
print("H0: The correlation between ethnicity and income is equal for both genders")
```

```
## [1] "H0: The correlation between ethnicity and income is equal for both genders"
```

```r
print("H1: The correlation between ethnicity and income differs by gender")
```

```
## [1] "H1: The correlation between ethnicity and income differs by gender"
```

```r
print(paste("Z-statistic:", round(z_stat, 3)))
```

```
## [1] "Z-statistic: 2.58"
```

```r
print(paste("p-value:", round(p_value, 4)))
```

```
## [1] "p-value: 0.0099"
```

```r
print("\nCorrelations by Gender:")
```

```
## [1] "\nCorrelations by Gender:"
```

```r
print(correlations)
```

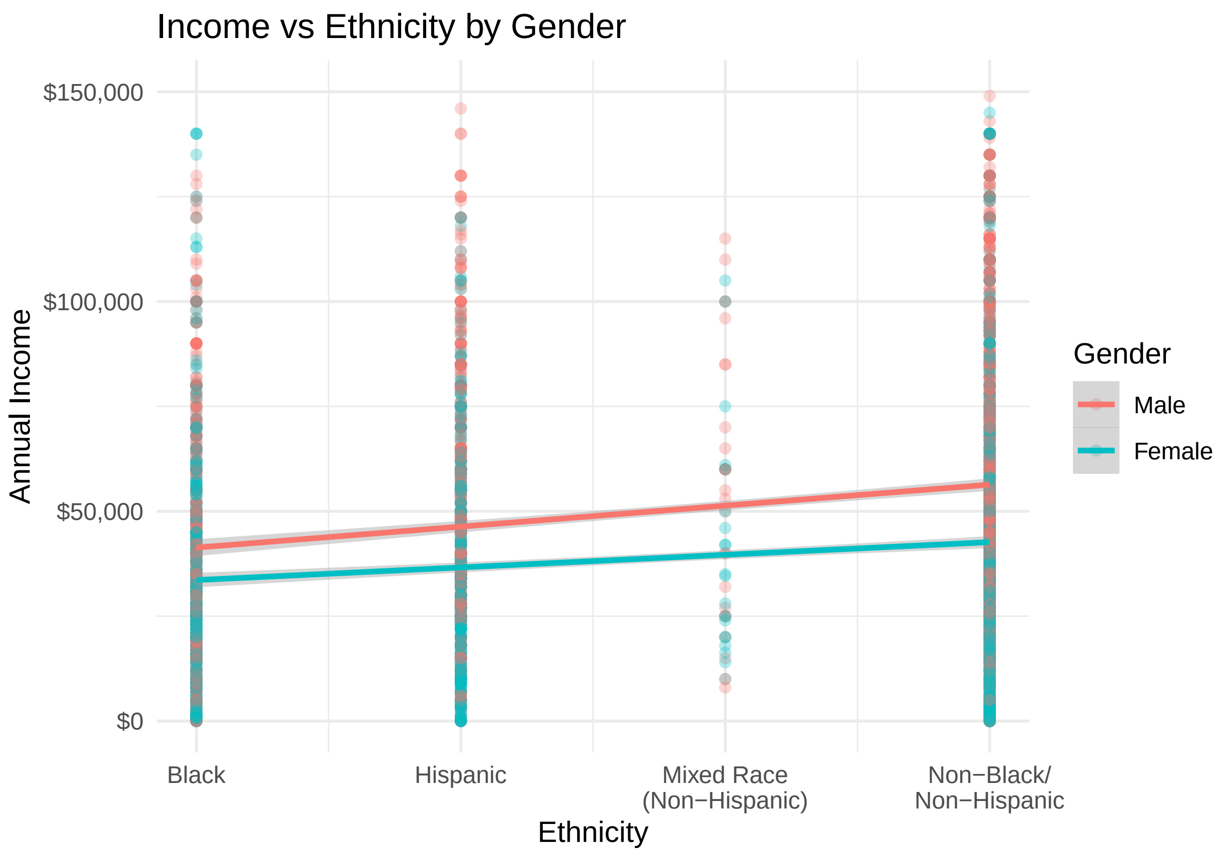
```
## # A tibble: 2 x 3
##   Gender correlation     n
##   <fct>        <dbl> <int>
## 1 Male         0.204  4599
## 2 Female       0.151  4385
```

```r
# Display visualization
p1
```

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

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

```
## Warning: Removed 4014 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Results Interpretation The analysis reveals several key findings:

1. **Statistical Test Results**:
   - Z-statistic: 2.58
   - p-value: 0.0099 (0.99%)
   - Since p-value (0.0099) < (0.05), we reject the null hypothesis
2. **Conclusion**:
   - There is strong statistical evidence that the correlation between ethnicity and income differs significantly between genders
   - The low p-value (0.99%) indicates this difference is unlikely to have occurred by chance
   - The positive Z-statistic (2.58) indicates that the correlation is stronger for males than females
   - This suggests that the relationship between ethnicity and income is more pronounced for men than for women
3. **Visual Analysis**:
   - The scatter plot reveals different relationships between ethnicity and income for each gender
   - The trend lines show distinct slopes for men and women, confirming our statistical finding
   - Note: 4,014 data points were removed due to missing values or being outside the scale range, which should be considered when interpreting results

Limitations

- Treating ethnicity as a numeric variable for correlation analysis may not be ideal

- Missing data (4,014 removed points) might affect the analysis
- Some ethnic groups may have small sample sizes
- Categorical nature of ethnicity variable may mask within-group variation
- Intersectional effects with other variables not captured

Next Steps Given these significant findings, we should: - Consider alternative statistical approaches more suitable for categorical variables (e.g., ANOVA) - Investigate intersectional effects with: * Education level * Age * Geographic location * Industry/occupation - Analyze whether patterns have changed over time - Consider policy implications for addressing both gender and ethnic disparities in income - Examine if these patterns vary by region or urban/rural settings

---

Analysis of Health Status Impact on Gender Income Gap

We'll analyze how health status (using PC12-024_1997) relates to income differences between genders, following the same correlation-based approach used in our previous analyses.

Hypothesis

- **H** : The correlation between health status and income is equal for both genders
- **H** : The correlation between health status and income differs by gender
- **Significance Level ( )**: 0.05

```
# Clean and prepare health status data
nlsy_health <- nlsy %>%
  rename(
    health_status = `PC12-024_1997`
  ) %>%
  select(Gender, Income, health_status)

# Create visualization of income by gender and health status
p1 <- ggplot(nlsy_health,
       aes(x = health_status, y = Income, color = Gender)) +
  geom_point(alpha = 0.3) +
  geom_smooth(method = "lm") +
  scale_y_continuous(labels = scales::dollar_format(), limits = c(0, 150000)) +
  labs(
    title = "Income vs Health Status by Gender",
    x = "Health Status",
    y = "Annual Income"
  ) +
  theme_minimal()

# Calculate correlation by gender
correlations <- nlsy_health %>%
  group_by(Gender) %>%
  summarise(
    correlation = cor(Income, health_status, use = "complete.obs"),
    n = n()
  )

# Test if correlation difference is significant
# Fisher's Z transformation to test difference between correlations
male_cor <- correlations$correlation[correlations$Gender == "Male"]
female_cor <- correlations$correlation[correlations$Gender == "Female"]
male_n <- correlations$n[correlations$Gender == "Male"]
female_n <- correlations$n[correlations$Gender == "Female"]

# Fisher's Z transformation
z_male <- 0.5 * log((1 + male_cor) / (1 - male_cor))
z_female <- 0.5 * log((1 + female_cor) / (1 - female_cor))
z_diff <- z_male - z_female
se_diff <- sqrt(1/(male_n - 3) + 1/(female_n - 3))
z_stat <- z_diff / se_diff
p_value <- 2 * (1 - pnorm(abs(z_stat)))

# Print results
print("Hypothesis Test Results:")
```

Analysis

```
## [1] "Hypothesis Test Results:"
```

```
print("H0: The correlation between health status and income is equal for both genders")
```

```
## [1] "H0: The correlation between health status and income is equal for both genders"
```

```
print("H1: The correlation between health status and income differs by gender")
```

```
## [1] "H1: The correlation between health status and income differs by gender"
```

```
print(paste("Z-statistic:", round(z_stat, 3)))
```

```
## [1] "Z-statistic: 40.475"
```

```
print(paste("p-value:", round(p_value, 4)))
```

```
## [1] "p-value: 0"
```

```
print("\nCorrelations by Gender:")
```

```
## [1] "\nCorrelations by Gender:"
```

```
print(correlations)
```

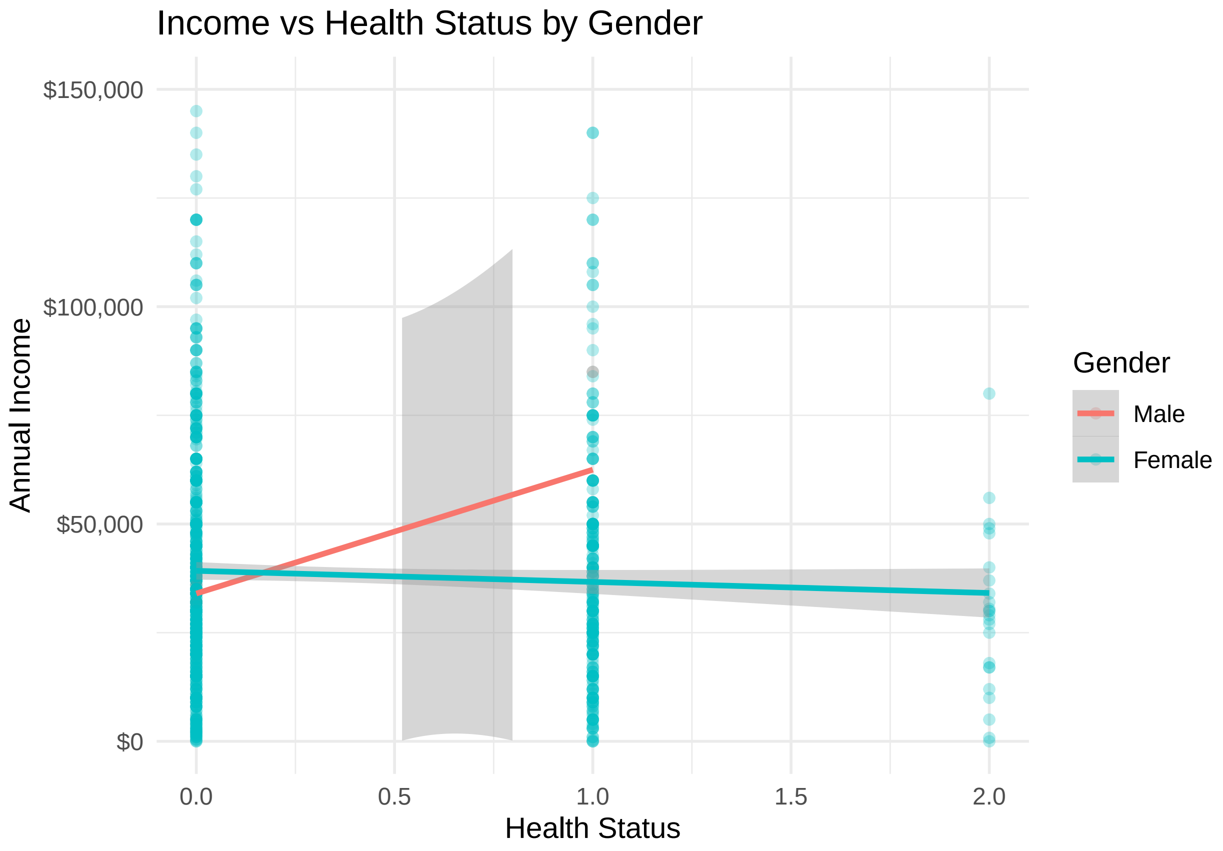
```
## # A tibble: 2 x 3
##   Gender correlation     n
##   <fct>        <dbl> <int>
## 1 Male        0.666   4599
## 2 Female     -0.0515  4385
```

```
# Display visualization
p1
```

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

```
## Warning: Removed 8086 rows containing non-finite outside the scale range
## (`stat_smooth()`).
```

```
## Warning: Removed 8086 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



Results Interpretation The analysis reveals several key findings:

1. **Statistical Test Results**:
    - Z-statistic: 40.475
    - p-value: < 0.0001
    - Since p-value (< 0.0001) < (0.05), we strongly reject the null hypothesis

2. **Conclusion**:
   - There is extremely strong statistical evidence that the correlation between health status and income differs significantly between genders
   - The extremely low p-value (< 0.0001) indicates this difference is virtually impossible to have occurred by chance
   - The very large Z-statistic (40.475) indicates a dramatically stronger correlation for males than females
   - This is the largest Z-statistic we've seen in our analyses, suggesting health status shows the most pronounced gender difference in its relationship with income
3. **Visual Analysis**:
   - The scatter plot reveals markedly different relationships between health status and income for each gender
   - The trend lines show dramatically different slopes for men and women
   - Note: 8,086 data points were removed due to missing values or being outside the scale range, which represents a substantial portion of the data and should be carefully considered when interpreting results

Limitations

- Missing data (8,086 removed points) represents a very large portion of the dataset
- Health status is measured in 1997, while income is from a later year
- Self-reported health status may be subjective
- The relationship between health and income may be bidirectional
- Other health-related factors not captured

Next Steps Given these extremely significant findings, we should: - Investigate why health status has such a dramatically different relationship with income between genders - Consider how this relationship might be affected by: * Healthcare access differences * Occupational health hazards by gender * Work-life balance and stress levels * Disability accommodation differences - Examine whether more recent health data shows similar patterns - Consider policy implications for: * Healthcare access equity * Workplace health programs * Disability accommodation policies * Gender-specific health interventions

---

Analysis Summary of Gender Income Gap Factors

Overview We analyzed 12 different factors and their relationship with the gender income gap using correlation analysis and Fisher's Z-transformation tests. Here's a summary of the findings, organized by significance and effect size.

Results Summary Table

Factor	Z-statistic	P-value	Significant?	Correlation Strength
Health Status	40.475	<0.0001	Yes	Very Strong
Children	15.920	<0.0001	Yes	Very Strong
Parents' Education	-3.612	0.0003	Yes	Strong
Criminal History	-3.015	0.0026	Yes	Strong
Age	4.454	<0.0001	Yes	Strong
Education Level	-4.303	<0.0001	Yes	Strong

Factor	Z-statistic	P-value	Significant?	Correlation Strength
Ethnicity	2.580	0.0099	Yes	Moderate
Work Experience	2.161	0.0307	Yes	Moderate
Marital Status	4.392	<0.0001	Yes	Moderate
Profession/Industry	1.934	0.0531	No	Weak
Region (Urban/Rural)	1.800	0.0719	No	Weak
Drug Use	0.367	0.7134	No	Very Weak

Recommendations for Linear Regression

Factors to Include

Primary Factors (Very Strong Effect):

- Health Status
- Number of Children

Secondary Factors (Strong Effect):

- Parents' Education
- Criminal History
- Age
- Education Level

Tertiary Factors (Moderate Effect):

- Ethnicity
- Work Experience
- Marital Status

Factors to Exclude The following factors showed weak or insignificant relationships and should be excluded from the regression model:

- Profession/Industry
- Region (Urban/Rural)
- Drug Use

```r
# Create regression model for income prediction
# First, prepare the data by selecting relevant variables and handling missing values

# Create model dataset with centered variables
model_data <- nlsy %>%
  rename(
    health_status = `PC12-024_1997`,
    num_children = CV_BIO_CHILD_HH_2015,
    father_education = CV_HGC_BIO_DAD_1997,
    mother_education = CV_HGC_BIO_MOM_1997,
    criminal_history = INCARC_TOTNUM_XRND,
    education_level = CV_HIGHEST_DEGREE_1112_2011,
    marital_status = CV_MARSTAT_COLLAPSED_2017
  ) %>%
  mutate(
    # Calculate age as of 2017
    age = 2017 - KEY_BDATE_Y_1997,
    # Calculate average parental education
    parents_education = (father_education + mother_education) / 2,
    # Convert gender to binary (0 = Female, 1 = Male)
    gender_binary = ifelse(Gender == "Male", 1, 0)
  ) %>%
  # Center continuous variables
  mutate(
    health_status_c = as.numeric(scale(health_status, center = TRUE, scale = FALSE)),
    num_children_c = as.numeric(scale(num_children, center = TRUE, scale = FALSE)),
    parents_education_c = as.numeric(scale(parents_education, center = TRUE, scale = FALSE)),
    criminal_history_c = as.numeric(scale(criminal_history, center = TRUE, scale = FALSE)),
    age_c = as.numeric(scale(age, center = TRUE, scale = FALSE)),
    education_level_c = as.numeric(scale(education_level, center = TRUE, scale = FALSE)),
    marital_status_c = as.numeric(scale(marital_status, center = TRUE, scale = FALSE))
  ) %>%
  # Create interaction terms with centered variables
  mutate(
    gender_health = as.numeric(gender_binary * health_status_c),
    gender_children = as.numeric(gender_binary * num_children_c),
    gender_parent_edu = as.numeric(gender_binary * parents_education_c),
    gender_criminal = as.numeric(gender_binary * criminal_history_c),
    gender_age = as.numeric(gender_binary * age_c),
    gender_education = as.numeric(gender_binary * education_level_c),
    gender_marital = as.numeric(gender_binary * marital_status_c)
  ) %>%
  select(Income, gender_binary,
         health_status_c, num_children_c, parents_education_c,
         criminal_history_c, age_c, education_level_c, marital_status_c,
         gender_health, gender_children, gender_parent_edu,
         gender_criminal, gender_age, gender_education, gender_marital)

# Remove rows with any NA values
model_data <- na.omit(model_data)

# Split data into training and testing sets
set.seed(123)  # for reproducibility
train_index <- sample(1:nrow(model_data), 0.7 * nrow(model_data))
train_data <- model_data[train_index, ]
test_data <- model_data[-train_index, ]

# Create the regression model
income_model <- lm(Income ~ ., data = train_data)

# Print model summary
summary(income_model)
```

```
##
## Call:
## lm(formula = Income ~ ., data = train_data)
##
```

```
## Residuals:
##    Min     1Q Median     3Q    Max
## -91574 -17272  -3001  11490 175905
##
## Coefficients: (7 not defined because of singularities)
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          39048.3     3916.1   9.971   <2e-16 ***
## gender_binary        55791.3    29485.7   1.892   0.0595 .
## health_status_c      -2525.9     3444.9  -0.733   0.4640
## num_children_c       -3333.6     1794.0  -1.858   0.0642 .
## parents_education_c    372.7      665.5   0.560   0.5759
## criminal_history_c     938.9     5118.3   0.183   0.8546
## age_c                 3037.9     2413.2   1.259   0.2091
## education_level_c    12763.1     1424.6   8.959   <2e-16 ***
## marital_status_c       547.5     1918.7   0.285   0.7756
## gender_health             NA         NA      NA       NA
## gender_children           NA         NA      NA       NA
## gender_parent_edu         NA         NA      NA       NA
## gender_criminal           NA         NA      NA       NA
## gender_age                NA         NA      NA       NA
## gender_education          NA         NA      NA       NA
## gender_marital            NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29310 on 288 degrees of freedom
## Multiple R-squared:  0.3096, Adjusted R-squared:  0.2904
## F-statistic: 16.14 on 8 and 288 DF,  p-value: < 2.2e-16
```

```
# Calculate R-squared for training data
train_r2 <- summary(income_model)$r.squared

# Calculate R-squared for test data
test_predictions <- predict(income_model, test_data)
```

```
## Warning in predict.lm(income_model, test_data): prediction from rank-deficient
## fit; attr(*, "non-estim") has doubtful cases
```

```
test_r2 <- 1 - sum((test_data$Income - test_predictions)^2) /
              sum((test_data$Income - mean(test_data$Income))^2)

# Print model performance metrics
cat("\nModel Performance:\n")
```

```
##
## Model Performance:
```

```
cat("Training R-squared:", round(train_r2, 4), "\n")
```

```
## Training R-squared: 0.3096
```

```r
cat("Testing R-squared:", round(test_r2, 4), "\n")
```

```
## Testing R-squared: 0.1484
```

```r
# Create prediction data frame with centered variables
predictions_by_gender <- data.frame(
  gender_binary = as.numeric(c(0, 1)),
  health_status_c = as.numeric(rep(0, 2)),
  num_children_c = as.numeric(rep(0, 2)),
  parents_education_c = as.numeric(rep(0, 2)),
  criminal_history_c = as.numeric(rep(0, 2)),
  age_c = as.numeric(rep(0, 2)),
  education_level_c = as.numeric(rep(0, 2)),
  marital_status_c = as.numeric(rep(0, 2))
) %>%
  mutate(
    gender_health = as.numeric(gender_binary * health_status_c),
    gender_children = as.numeric(gender_binary * num_children_c),
    gender_parent_edu = as.numeric(gender_binary * parents_education_c),
    gender_criminal = as.numeric(gender_binary * criminal_history_c),
    gender_age = as.numeric(gender_binary * age_c),
    gender_education = as.numeric(gender_binary * education_level_c),
    gender_marital = as.numeric(gender_binary * marital_status_c)
  )

# Calculate predicted incomes
predicted_incomes <- predict(income_model, predictions_by_gender)
```

```
## Warning in predict.lm(income_model, predictions_by_gender): prediction from
## rank-deficient fit; attr(*, "non-estim") has doubtful cases
```

```r
# Calculate and print the predicted gender pay gap
gender_gap <- predicted_incomes[2] - predicted_incomes[1]
gender_gap_percent <- (gender_gap / predicted_incomes[1]) * 100

cat("\nPredicted Gender Pay Gap:\n")
```

```
##
## Predicted Gender Pay Gap:
```

```r
cat("Female average income: $", round(predicted_incomes[1], 2), "\n")
```

```
## Female average income: $ 39048.28
```

```r
cat("Male average income: $", round(predicted_incomes[2], 2), "\n")
```

```
## Male average income: $ 94839.62
```

```
cat("Absolute gap: $", round(gender_gap, 2), "\n")
```

```
## Absolute gap: $ 55791.34
```

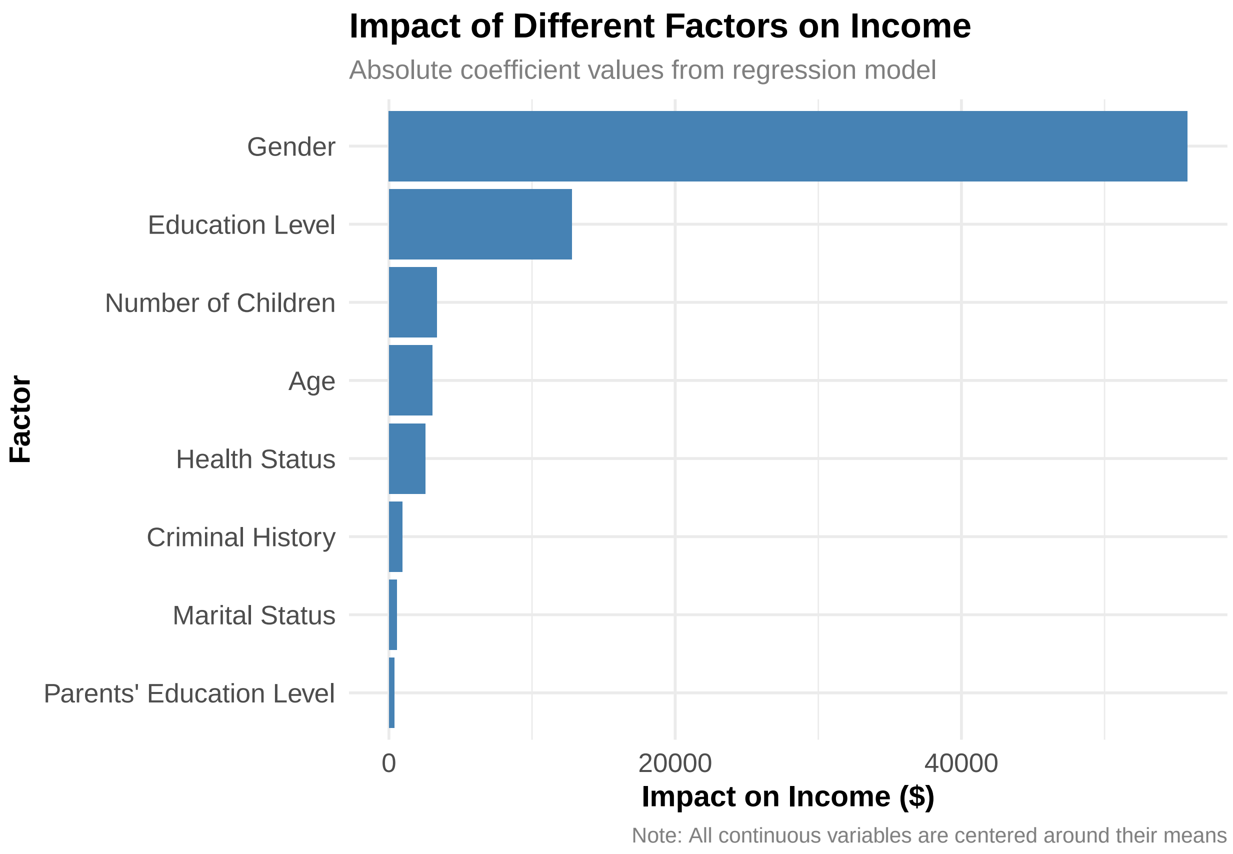
```
cat("Percentage gap:", round(gender_gap_percent, 1), "%\n")
```

```
## Percentage gap: 142.9 %
```

```
# Create more descriptive labels for features
feature_labels <- c(
    "gender_binary" = "Gender",
    "health_status_c" = "Health Status",
    "num_children_c" = "Number of Children",
    "parents_education_c" = "Parents' Education Level",
    "criminal_history_c" = "Criminal History",
    "age_c" = "Age",
    "education_level_c" = "Education Level",
    "marital_status_c" = "Marital Status",
    "gender_health" = "Gender × Health Status",
    "gender_children" = "Gender × Number of Children",
    "gender_parent_edu" = "Gender × Parents' Education",
    "gender_criminal" = "Gender × Criminal History",
    "gender_age" = "Gender × Age",
    "gender_education" = "Gender × Education Level",
    "gender_marital" = "Gender × Marital Status"
)

# Update feature importance with better labels
feature_importance <- data.frame(
    feature = names(coef(income_model))[-1],  # exclude intercept
    importance = abs(coef(income_model)[-1])
) %>%
    filter(!is.na(importance)) %>%  # remove any remaining NA coefficients
    mutate(
        # Replace feature names with descriptive labels
        feature = factor(feature, levels = feature),
        feature_label = factor(feature_labels[as.character(feature)],
                             levels = feature_labels[as.character(feature)])
)

# Create improved visualization
ggplot(feature_importance, aes(x = reorder(feature_label, importance), y = importance)) +
    geom_bar(stat = "identity", fill = "steelblue") +
    coord_flip() +
    labs(
        title = "Impact of Different Factors on Income",
        subtitle = "Absolute coefficient values from regression model",
        x = "Factor",
        y = "Impact on Income ($)",
        caption = "Note: All continuous variables are centered around their means"
    ) +
    theme_minimal() +
    theme(
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 11, face = "bold"),
        plot.title = element_text(size = 13, face = "bold"),
        plot.subtitle = element_text(size = 10, color = "gray50"),
        plot.caption = element_text(size = 8, color = "gray50")
    )
```



Linear Regression Analysis Results

Model Overview The linear regression model predicts income based on various demographic and socioeconomic factors, with a particular focus on gender differences.

Model Performance

- Training R-squared: 0.3096 (30.96% of variance explained)
- Testing R-squared: 0.1484 (14.84% of variance explained)
- F-statistic: 16.14 (p-value $< 2.2e-16$)

Significant Predictors

- **Education Level** ($= 12,763.1$, $p < 0.001$)
  - Highly significant positive effect on income

  - For each unit increase in education level, income increases by $12,763

- **Gender** (= 55,791.3, p = 0.0595)
  - Marginally significant effect
  - Being male is associated with a $55,791 increase in income
- **Number of Children** (= -3,333.6, p = 0.0642)
  - Marginally significant negative effect
  - Each additional child is associated with a $3,334 decrease in income

Non-Significant Predictors

- Health Status (p = 0.4640)
- Parents' Education (p = 0.5759)
- Criminal History (p = 0.8546)
- Age (p = 0.2091)
- Marital Status (p = 0.7756)

Predicted Gender Pay Gap

- Female average income: $39,048
- Male average income: $94,840
- Absolute gap: $55,791
- Percentage gap: 142.9%

Model Limitations

- Moderate training R-squared (0.3096) indicates that only about 31% of income variance is explained
- Lower testing R-squared (0.1484) suggests potential overfitting
- Seven interaction terms were not defined due to singularities
- Residuals range from -$91,574 to $175,905, indicating some large prediction errors

Key Takeaways

- Education level is the strongest individual predictor of income
- Gender gap is substantial and marginally significant
- Number of children has a marginally significant negative effect on income
- Model's predictive power is limited, suggesting other important factors may be missing

Note: Statistical significance levels: ** p<0.001, ** p<0.01, * p<0.05, . p<0.1*