Exercises on Multiple Plots

STSM2634

2025-05-19

Q1. Use the iris dataset and plot histograms of Sepal length, Sepal width, Petal length, and Petal width in a single plot.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.2     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.1
## ── Conflicts ────────────────────────────────────────
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(Ecdat)
```

```
## Warning: package 'Ecdat' was built under R version 4.3.3
```

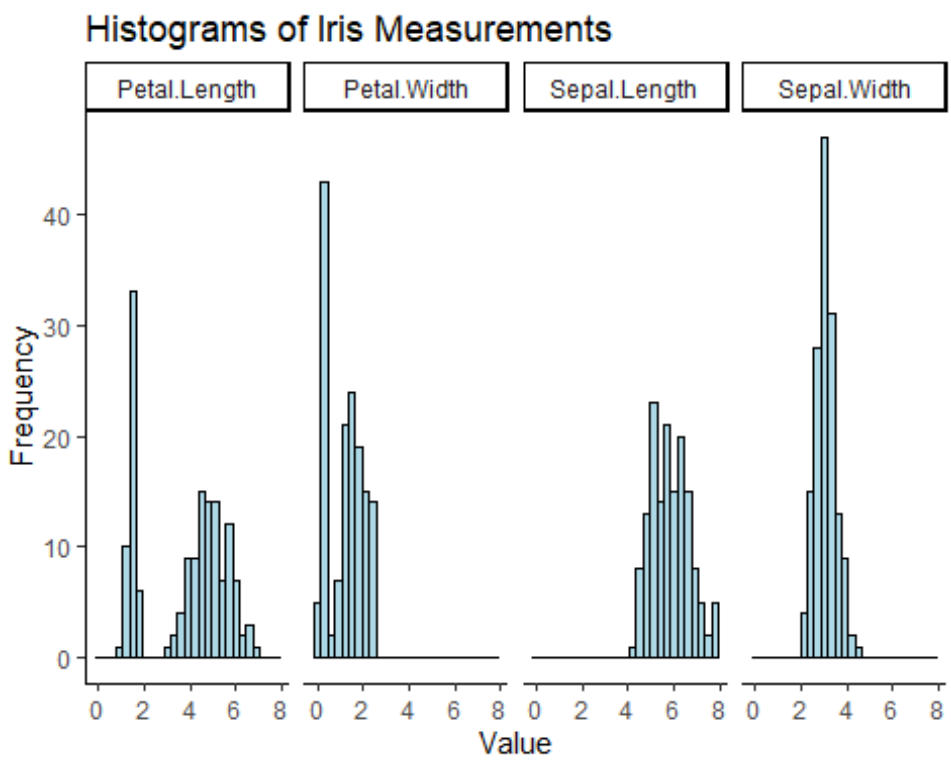
```
## Loading required package: Ecfun
```

```
## Warning: package 'Ecfun' was built under R version 4.3.3
```

```
##
## Attaching package: 'Ecfun'
##
## The following object is masked from 'package:base':
##
##     sign
##
```

```
##
## Attaching package: 'Ecdat'
##
## The following object is masked from 'package:MASS':
##
##     SP500
##
## The following object is masked from 'package:datasets':
##
##     Orange
```

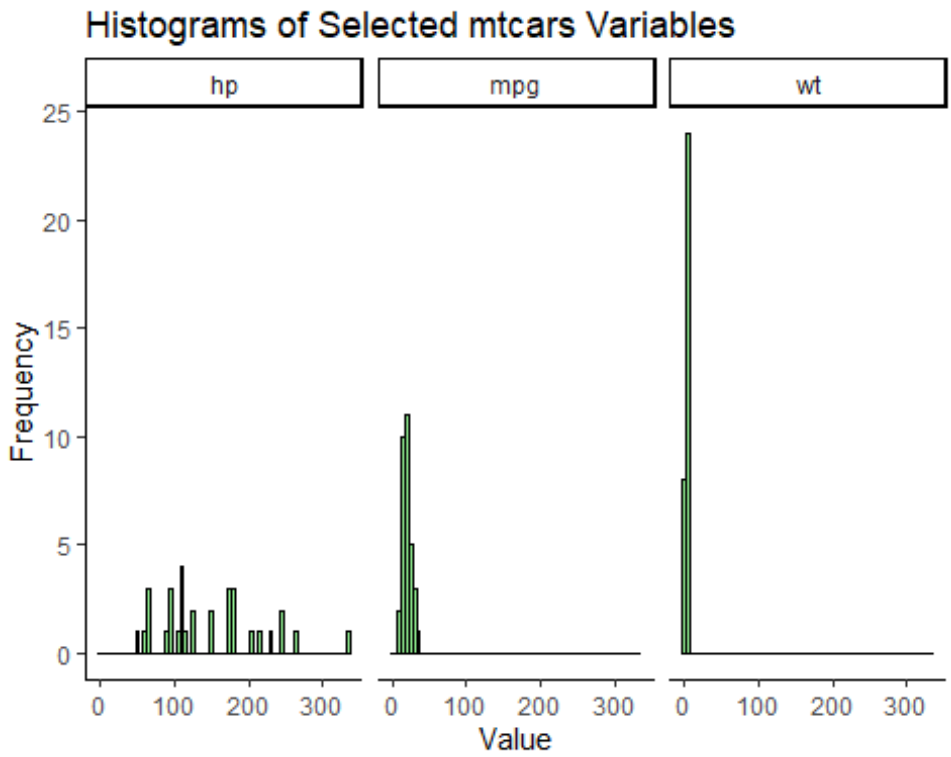
```
data = iris[, 1:4]  # exclude Species column

data_long = gather(data, key = "variable", value = "value")

ggplot(data_long, aes(x = value)) +
  geom_histogram(binwidth = 0.3, color = "black", fill = "lightblue") +
  facet_wrap(~ variable, nrow = 1) +
  labs(x = "Value", y = "Frequency", title = "Histograms of Iris
Measurements")+
  theme_classic()
```



Q2. Plot histograms for numeric variables in ` mtcars' dataset.

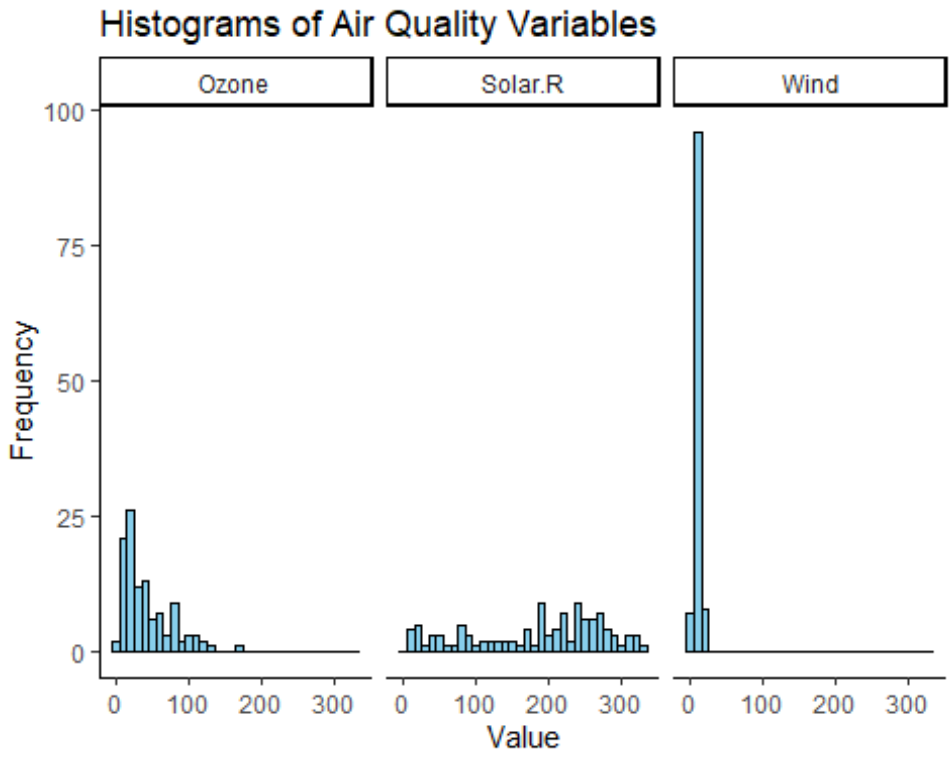
```
data = mtcars[, c("mpg", "hp", "wt")]

data_long = gather(data, key = "variable", value = "value")
```

```
ggplot(data_long, aes(x = value)) +
  geom_histogram(binwidth = 5, color = "black", fill = "lightgreen") +
  facet_wrap(~ variable, nrow = 1) +
  labs(x = "Value", y = "Frequency", title = "Histograms of Selected mtcars
Variables")+
    theme_classic()
```



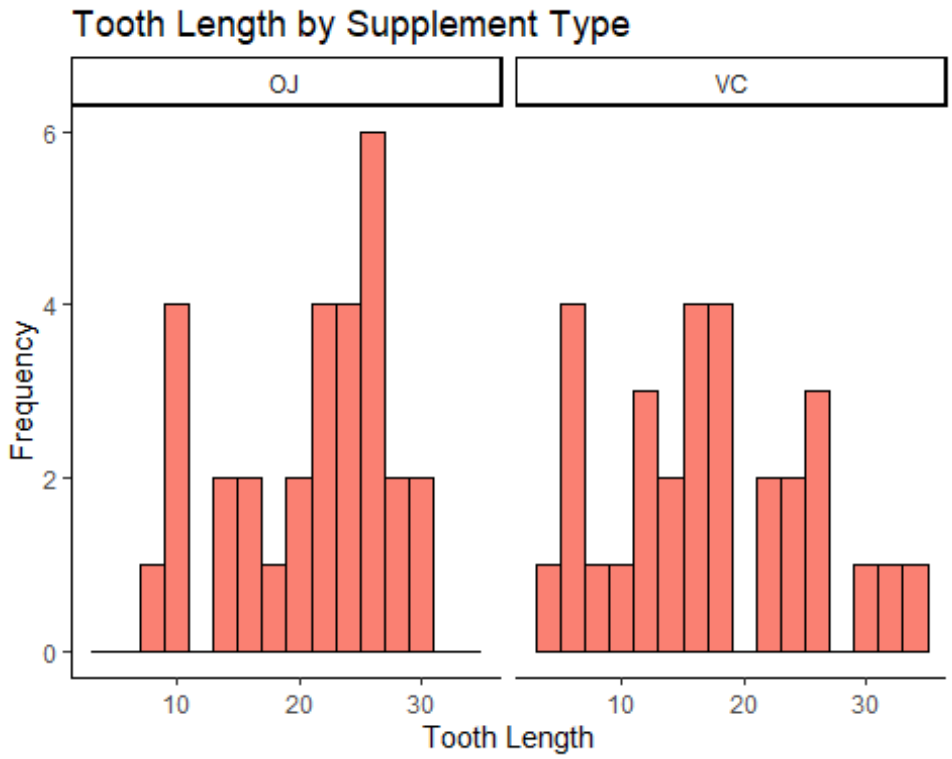
Q3. Use the airquality dataset to plot histograms of Ozone, Solar.R, and Wind (ignore NAs).

```
data = na.omit(airquality[, c("Ozone", "Solar.R", "Wind")])

data_long = gather(data, key = "variable", value = "value")

ggplot(data_long, aes(x = value)) +
  geom_histogram(binwidth = 10, color = "black", fill = "skyblue") +
  facet_wrap(~ variable, nrow = 1) +
  labs(x = "Value", y = "Frequency", title = "Histograms of Air Quality
Variables")+
    theme_classic()
```



Q4. Visualize histograms of len grouped by supp using facets.

```
data = ToothGrowth

# You only need to use len and supp; no need to reshape
ggplot(data, aes(x = len)) +
  geom_histogram(binwidth = 2, color = "black", fill = "salmon") +
  facet_wrap(~ supp) +
  labs(x = "Tooth Length", y = "Frequency", title = "Tooth Length by
Supplement Type")+
    theme_classic()
```



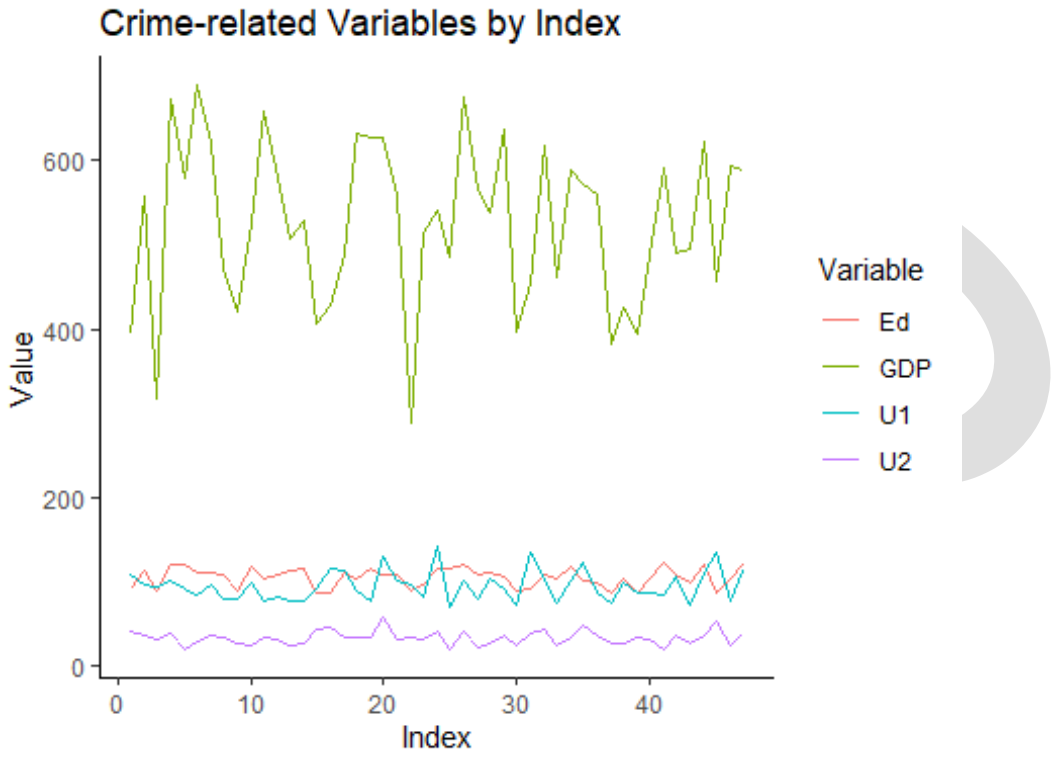
Q5. For the UScrime data from Package MASS, plot crime rates over income and education.

```
data(UScrime)
str(UScrime)
```

```
## 'data.frame':    47 obs. of  16 variables:
##  $ M   : int  151 143 142 136 141 121 127 131 157 140 ...
##  $ So  : int  1 0 1 0 0 0 1 1 1 0 ...
##  $ Ed  : int  91 113 89 121 121 110 111 109 90 118 ...
##  $ Po1 : int  58 103 45 149 109 118 82 115 65 71 ...
##  $ Po2 : int  56 95 44 141 101 115 79 109 62 68 ...
##  $ LF  : int  510 583 533 577 591 547 519 542 553 632 ...
##  $ M.F : int  950 1012 969 994 985 964 982 969 955 1029 ...
##  $ Pop : int  33 13 18 157 18 25 4 50 39 7 ...
##  $ NW  : int  301 102 219 80 30 44 139 179 286 15 ...
##  $ U1  : int  108 96 94 102 91 84 97 79 81 100 ...
##  $ U2  : int  41 36 33 39 20 29 38 35 28 24 ...
##  $ GDP : int  394 557 318 673 578 689 620 472 421 526 ...
##  $ Ineq: int  261 194 250 167 174 126 168 206 239 174 ...
##  $ Prob: num  0.0846 0.0296 0.0834 0.0158 0.0414 ...
##  $ Time: num  26.2 25.3 24.3 29.9 21.3 ...
##  $ y   : int  791 1635 578 1969 1234 682 963 1555 856 705 ...
```

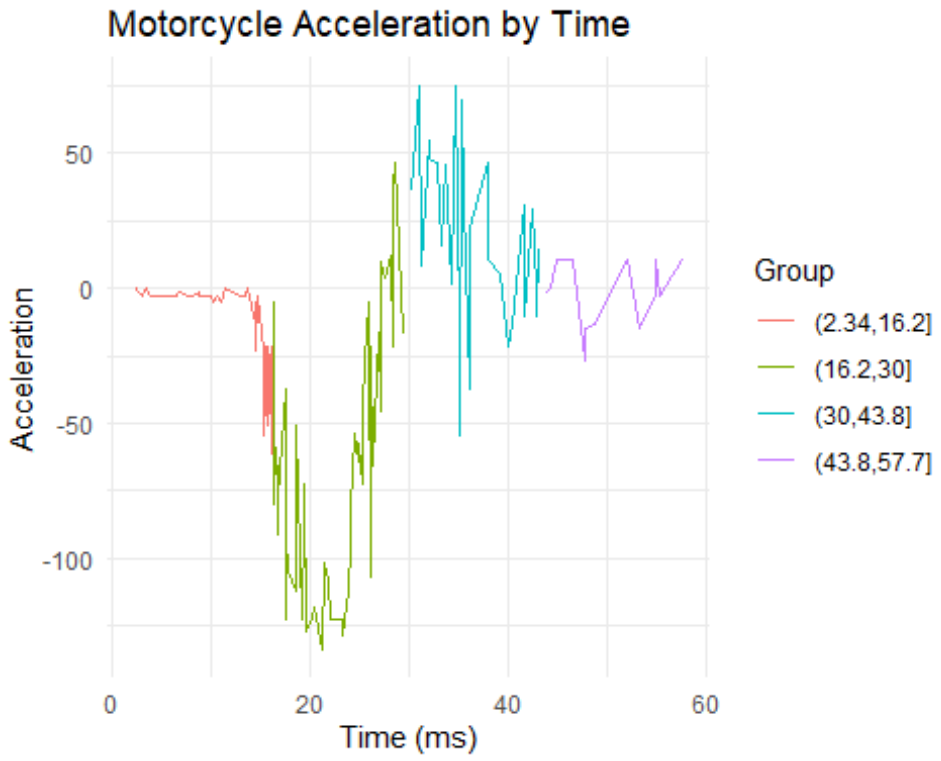
```
# Select continuous variables to compare
df = UScrime[, c("Ed", "U1", "U2", "GDP")]
df$id = 1:nrow(df)
```

```r
# Reshape
df_long = pivot_longer(df, cols = -id, names_to = "Variable", values_to =
"Value")

# Plot
ggplot(df_long, aes(x = id, y = Value, color = Variable)) +
  geom_line() +
  labs(title = "Crime-related Variables by Index", x = "Index", y = "Value")+
  theme_classic()
```



Q6. Use the ` mcycle' dataset and plot head acceleration curves.

```r
data(mcycle)

# Simulate splitting data by speed levels (quantiles)
mcycle$Group = cut(mcycle$times, breaks = 4)

ggplot(mcycle, aes(x = times, y = accel, color = Group)) +
  geom_line() +
  labs(title = "Motorcycle Acceleration by Time", x = "Time (ms)", y =
"Acceleration")+
  theme_minimal()
```



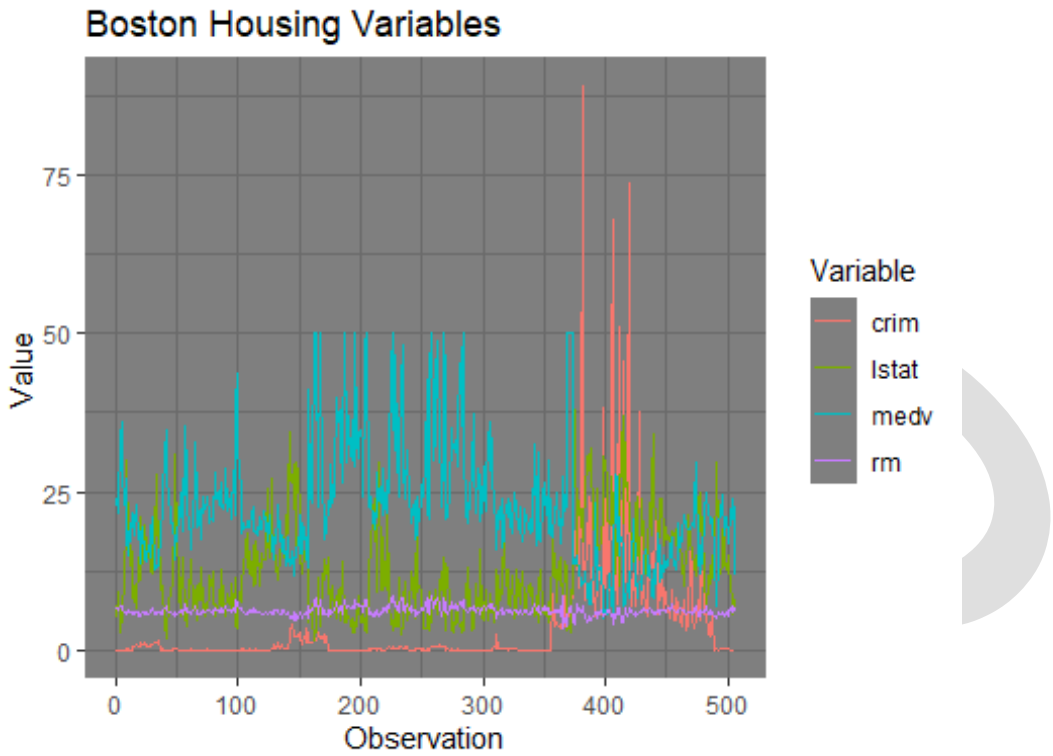
Q7. Use the `Boston' data and plot the median home value over other variables 'lstat', 'rm', and 'crim'.

```
data(Boston)
str(Boston)
```

```
## 'data.frame':    506 obs. of  14 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524
0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ black  : num  397 397 393 395 397 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

```
# Choose a few variables and add ID for plotting
df = Boston[, c("medv", "lstat", "rm", "crim")]
df$id = 1:nrow(df)
df_long = pivot_longer(df, cols = -id, names_to = "Variable", values_to =
```

```
"Value")

ggplot(df_long, aes(x = id, y = Value, color = Variable)) +
  geom_line() +
  labs(title = "Boston Housing Variables", x = "Observation", y = "Value")+
  theme_dark()
```



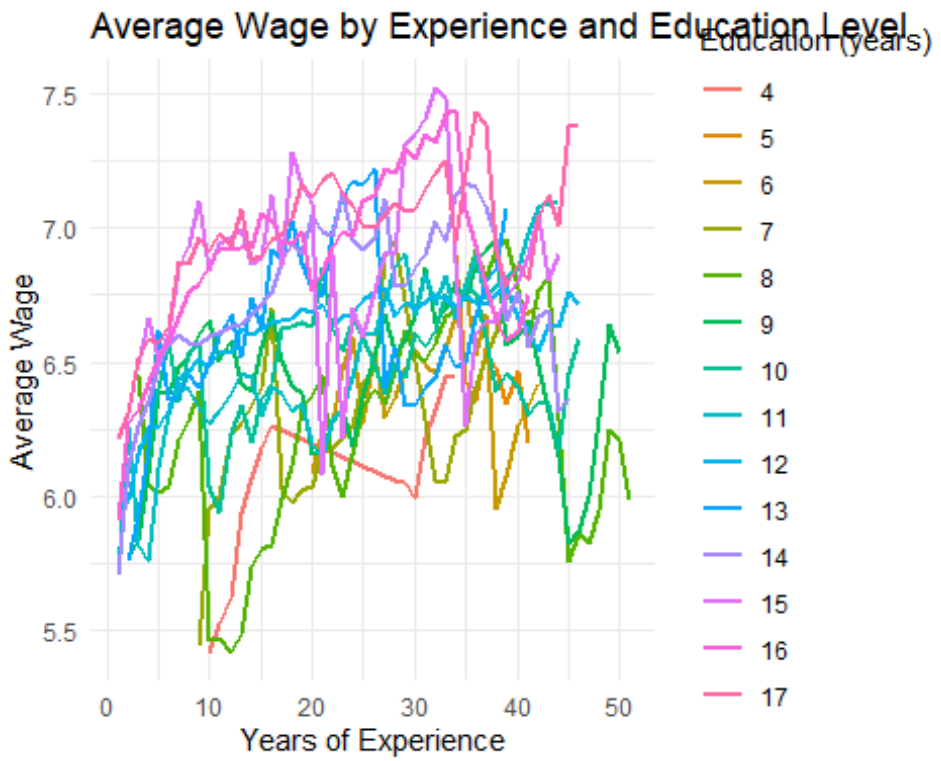
Q8. Use the Wages dataset and plot wages by experience and education.

```
# Load data
data(Wages)
str(Wages)
```

```
## 'data.frame':    4165 obs. of  12 variables:
##  $ exp    : int  3 4 5 6 7 8 9 30 31 32 ...
##  $ wks    : int  32 43 40 39 42 35 32 34 27 33 ...
##  $ bluecol: Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 2 2 ...
##  $ ind    : int  0 0 0 0 1 1 1 0 0 1 ...
##  $ south  : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 1 1 1 ...
##  $ smsa   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ married: Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ sex    : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ union  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 2 ...
##  $ ed     : int  9 9 9 9 9 9 9 11 11 11 ...
##  $ black  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ lwage  : num  5.56 5.72 6 6 6.06 ...
```

```r
# Group data by years of education and calculate average wage by experience
avg_wage = Wages %>%
  group_by(ed, exp) %>%
  summarise(mean_wage = mean(lwage, na.rm = TRUE), .groups = "drop")

# Plot multiple lines: one line per education level
ggplot(avg_wage, aes(x = exp, y = mean_wage, color = as.factor(ed))) +
  geom_line(size = 1) +
  labs(title = "Average Wage by Experience and Education Level",
       x = "Years of Experience", y = "Average Wage",
       color = "Education (years)") +
  theme_minimal()
```

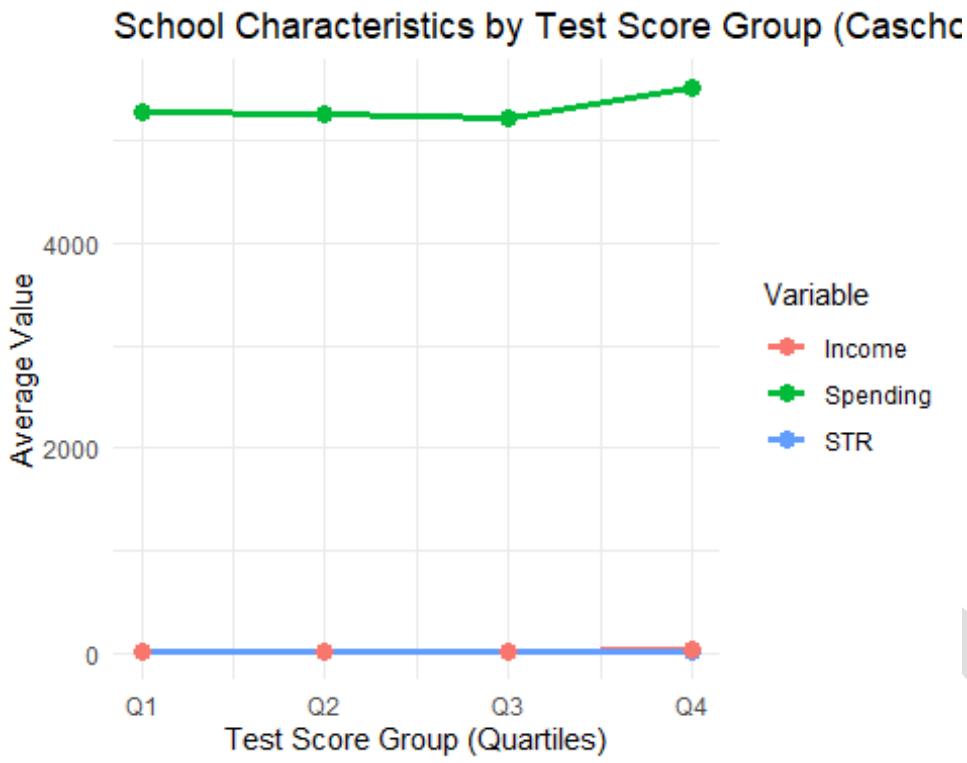
```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Q9. Use the BudgetFood dataset, and plot total food expenditures by income group.

```r
# Load dataset
data("Caschool")
str(Caschool)
```

```
## 'data.frame':    420 obs. of  17 variables:
##  $ distcod : int  75119 61499 61549 61457 61523 62042 68536 63834 62331
67306 ...
##  $ county  : Factor w/ 45 levels "Alameda","Butte",..: 1 2 2 2 2 6 29 11 6
25 ...
##  $ district: Factor w/ 409 levels "Ackerman Elementary",..: 362 214 367
132 270 53 152 383 263 94 ...
##  $ grspan  : Factor w/ 2 levels "KK-06","KK-08": 2 2 2 2 2 2 2 2 2 1 ...
##  $ enrltot : int  195 240 1550 243 1335 137 195 888 379 2247 ...
##  $ teachers: num  10.9 11.1 82.9 14 71.5 ...
##  $ calwpct : num  0.51 15.42 55.03 36.48 33.11 ...
##  $ mealpct : num  2.04 47.92 76.32 77.05 78.43 ...
##  $ computer: int  67 101 169 85 171 25 28 66 35 0 ...
##  $ testscr : num  691 661 644 648 641 ...
##  $ compstu : num  0.344 0.421 0.109 0.35 0.128 ...
##  $ expnstu : num  6385 5099 5502 7102 5236 ...
##  $ str     : num  17.9 21.5 18.7 17.4 18.7 ...
##  $ avginc  : num  22.69 9.82 8.98 8.98 9.08 ...
##  $ elpct   : num  0 4.58 30 0 13.86 ...
##  $ readscr : num  692 660 636 652 642 ...
##  $ mathscr : num  690 662 651 644 640 ...

# Create test score groups (quartiles)
Caschool$score_group = ntile(Caschool$testscr, 4)

# Group by score_group and compute averages
avg_stats = Caschool %>%
  group_by(score_group) %>%
  summarise(
    STR = mean(str, na.rm = TRUE),            # Student-Teacher Ratio
    Spending = mean(expnstu, na.rm = TRUE),  # Spending per student
    Income = mean(avginc, na.rm = TRUE),      # Average income
    .groups = "drop"
  )

# Reshape to long format for plotting
avg_stats_long = pivot_longer(avg_stats, cols = c(STR, Spending, Income),
                              names_to = "Variable", values_to =
"MeanValue")

# Plot
ggplot(avg_stats_long, aes(x = score_group, y = MeanValue, color = Variable))
+
  geom_line(linewidth = 1.2) +
  geom_point(size = 3) +
  scale_x_continuous(breaks = 1:4, labels = c("Q1", "Q2", "Q3", "Q4")) +
  labs(title = "School Characteristics by Test Score Group (Caschool)",
       x = "Test Score Group (Quartiles)",
       y = "Average Value") +
  theme_minimal()
```



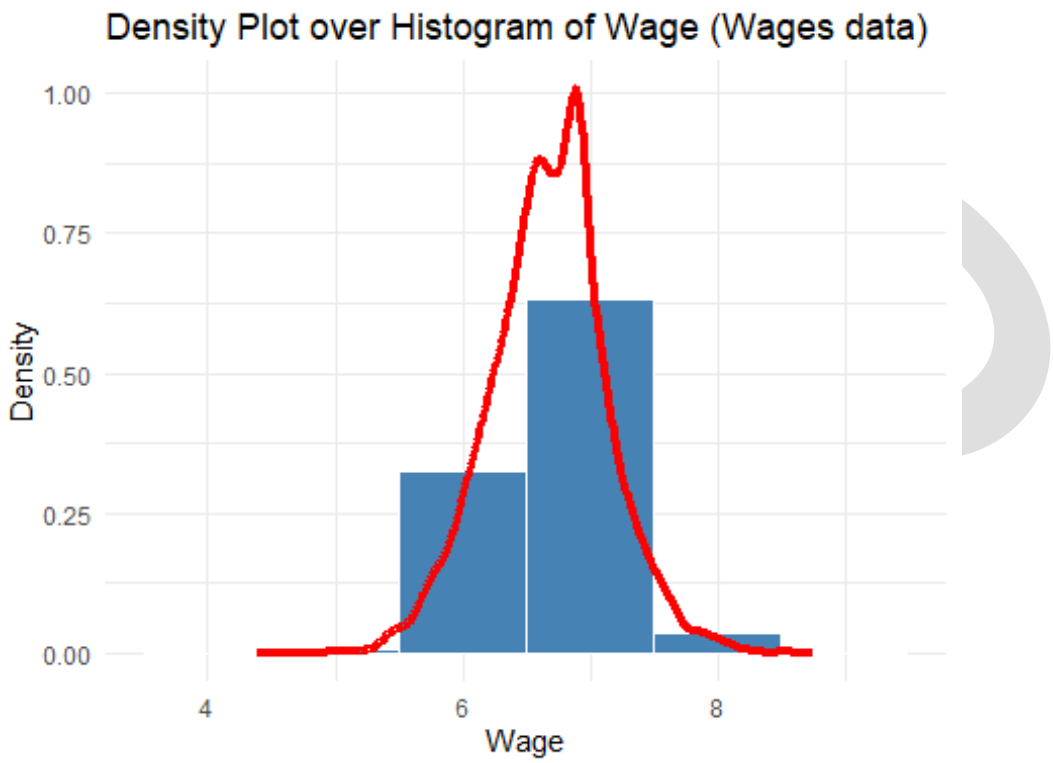
Q10. Plot the Wage Density from the Wages dataset.

```
# Load data
data("Wages")
str(Wages)
```

```
## 'data.frame':    4165 obs. of  12 variables:
##  $ exp    : int  3 4 5 6 7 8 9 30 31 32 ...
##  $ wks    : int  32 43 40 39 42 35 32 34 27 33 ...
##  $ bluecol: Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 2 2 ...
##  $ ind    : int  0 0 0 0 1 1 1 0 0 1 ...
##  $ south  : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 1 1 1 ...
##  $ smsa   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ married: Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ sex    : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ union  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 2 ...
##  $ ed     : int  9 9 9 9 9 9 9 11 11 11 ...
##  $ black  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ lwage  : num  5.56 5.72 6 6 6.06 ...
```

```
# Use the 'wage' variable
data1 = data.frame(x = Wages$lwage)

# Estimate density
density_data = density(data1$x, na.rm = TRUE)
data2 = data.frame(ex = density_data$x, prob1 = density_data$y)
```

```r
# Combined histogram + density line plot
combined_plot = ggplot() +
  geom_histogram(data = data1, aes(x = x, y = after_stat(density)), binwidth
= 1, fill = "steelblue", color = "white") +
  geom_line(data = data2, aes(x = ex, y = prob1), color = "red", linewidth =
1.5) +
  labs(x = "Wage", y = "Density", title = "Density Plot over Histogram of
Wage (Wages data)") +
  theme_minimal()

combined_plot
```

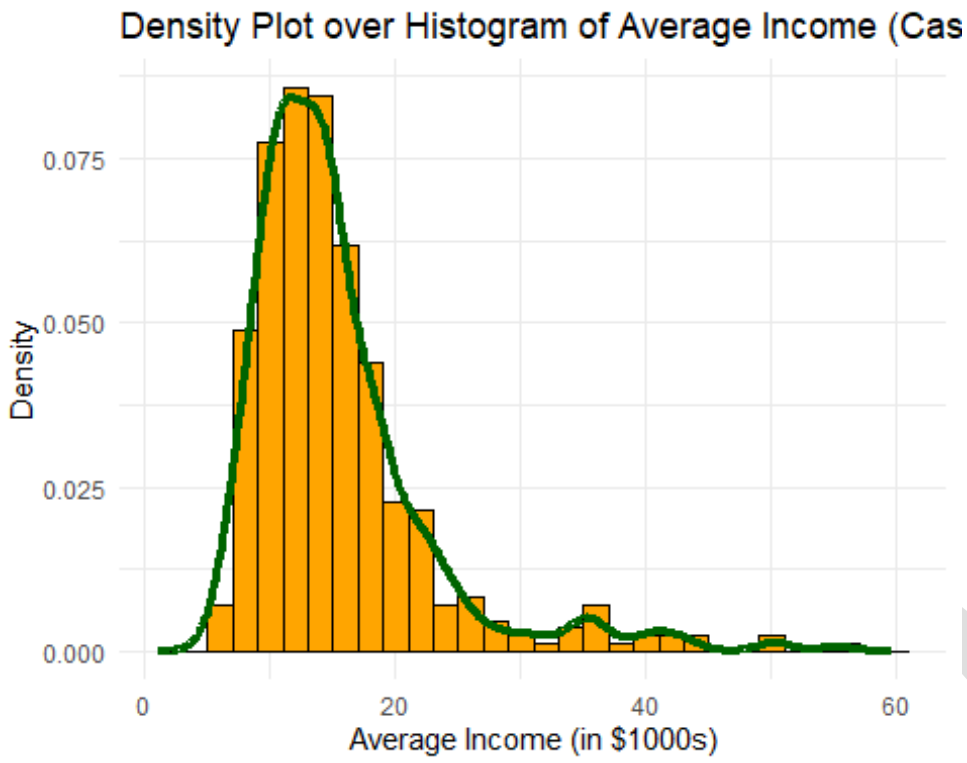


Q11. Plot the density of the Average Income from Caschool dataset.

```r
# Load data
data("Caschool")
str(Caschool)
```

```
## 'data.frame':    420 obs. of  17 variables:
##  $ distcod : int  75119 61499 61549 61457 61523 62042 68536 63834 62331
67306 ...
##  $ county  : Factor w/ 45 levels "Alameda","Butte",..: 1 2 2 2 2 6 29 11 6
25 ...
##  $ district: Factor w/ 409 levels "Ackerman Elementary",..: 362 214 367
132 270 53 152 383 263 94 ...
##  $ grspan  : Factor w/ 2 levels "KK-06","KK-08": 2 2 2 2 2 2 2 2 2 1 ...
##  $ enrltot : int  195 240 1550 243 1335 137 195 888 379 2247 ...
```

```
##  $ teachers: num  10.9 11.1 82.9 14 71.5 ...
##  $ calwpct : num  0.51 15.42 55.03 36.48 33.11 ...
##  $ mealpct : num  2.04 47.92 76.32 77.05 78.43 ...
##  $ computer: int  67 101 169 85 171 25 28 66 35 0 ...
##  $ testscr : num  691 661 644 648 641 ...
##  $ compstu : num  0.344 0.421 0.109 0.35 0.128 ...
##  $ expnstu : num  6385 5099 5502 7102 5236 ...
##  $ str     : num  17.9 21.5 18.7 17.4 18.7 ...
##  $ avginc  : num  22.69 9.82 8.98 8.98 9.08 ...
##  $ elpct   : num  0 4.58 30 0 13.86 ...
##  $ readscr : num  692 660 636 652 642 ...
##  $ mathscr : num  690 662 651 644 640 ...

# Use 'avginc' (average district income)
data1 = data.frame(x = Caschool$avginc)

# Density estimate
density_data = density(data1$x, na.rm = TRUE)
data2 = data.frame(ex = density_data$x, prob1 = density_data$y)

# Plot histogram + density line
combined_plot2 = ggplot() +
  geom_histogram(data = data1, aes(x = x, y = after_stat(density)), binwidth
= 2,
                 fill = "orange", color = "black") +
  geom_line(data = data2, aes(x = ex, y = prob1), color = "darkgreen",
linewidth = 1.5) +
  labs(x = "Average Income (in $1000s)", y = "Density",
       title = "Density Plot over Histogram of Average Income (Caschool
data)") +
  theme_minimal()

combined_plot2
```



Q12. Plot the density of the nonlabor income from the SwissLabor dataset (use AER package).

```
library(AER)
```

```
## Warning: package 'AER' was built under R version 4.3.3
```

```
## Loading required package: car
```

```
## Warning: package 'car' was built under R version 4.3.1
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'carData'
```

```
## The following object is masked from 'package:Ecdat':
##
##     Mroz
```

```
##
## Attaching package: 'car'
```

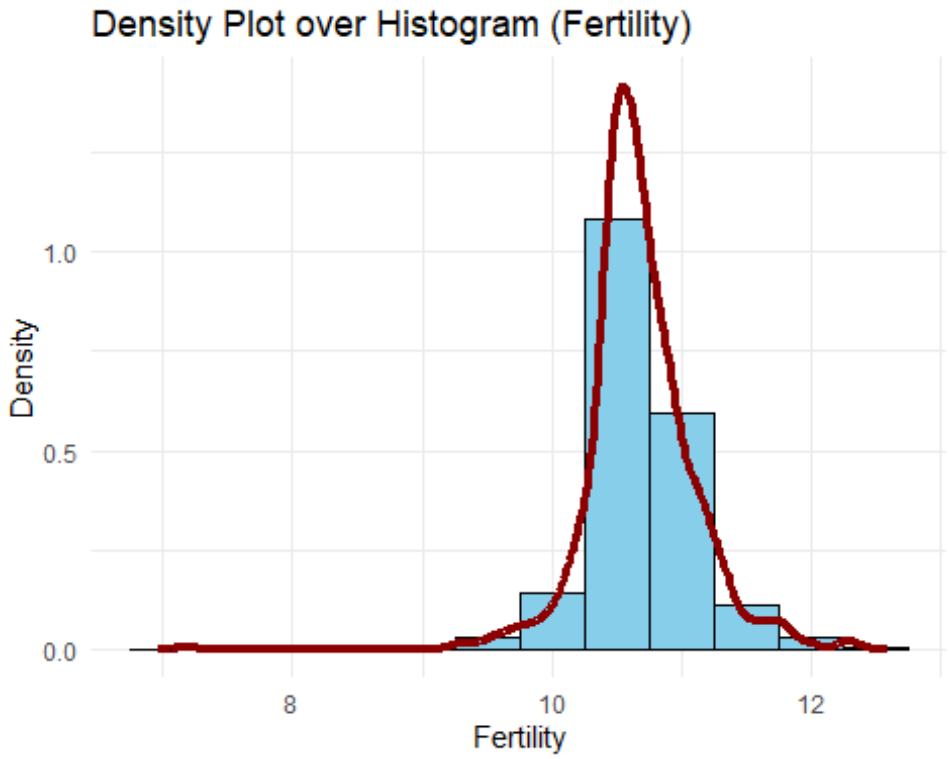
```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
## The following object is masked from 'package:purrr':
##
##     some

## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 4.3.1

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival
```

```
# Load data
data("SwissLabor")
str(data)
```

```
## 'data.frame':    60 obs. of  3 variables:
##  $ len : num  4.2 11.5 7.3 5.8 6.4 10 11.2 11.2 5.2 7 ...
##  $ supp: Factor w/ 2 levels "OJ","VC": 2 2 2 2 2 2 2 2 2 2 ...
##  $ dose: num  0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 0.5 ...
```

```
# Use 'fertility' variable
data1 = data.frame(x = SwissLabor$income)

# Compute density
density_data = density(data1$x, na.rm = TRUE)
data2 = data.frame(ex = density_data$x, prob1 = density_data$y)

# Plot
combined_plot3 = ggplot() +
  geom_histogram(data = data1, aes(x = x, y = after_stat(density)), binwidth
= 0.5,
                 fill = "skyblue", color = "black") +
  geom_line(data = data2, aes(x = ex, y = prob1), color = "darkred",
linewidth = 1.5) +
  labs(x = "Fertility", y = "Density", title = "Density Plot over Histogram
(Fertility)") +
  theme_minimal()

combined_plot3
```



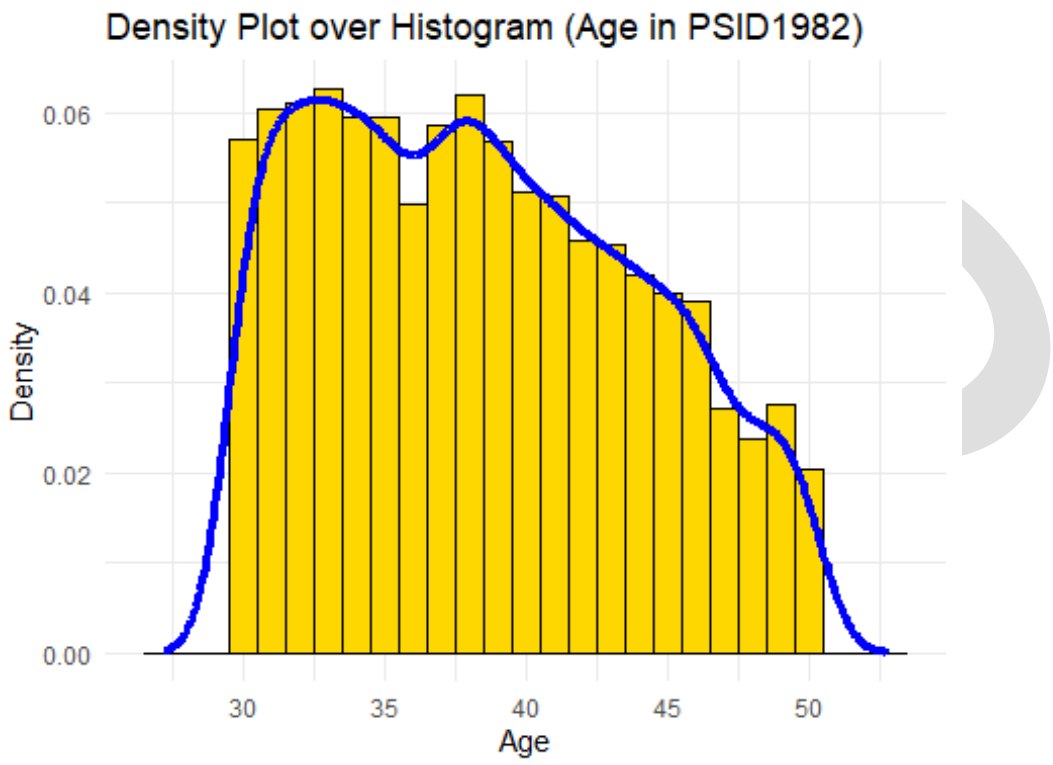
Q13. Create a density plot for the Age of Women from the PSID1982 dataset.

```
# Load data
data("PSID")
str(PSID)
```

```
## 'data.frame':    4856 obs. of  8 variables:
##  $ intnum  : int  4 4 4 4 5 6 6 7 7 7 ...
##  $ persnum : int  4 6 7 173 2 4 172 4 170 171 ...
##  $ age     : int  39 35 33 39 47 44 38 38 39 37 ...
##  $ educatn : int  12 12 12 10 9 12 16 9 12 11 ...
##  $ earnings: int  77250 12000 8000 15000 6500 6500 7000 5000 21000 0 ...
##  $ hours   : int  2940 2040 693 1904 1683 2024 1144 2080 2575 0 ...
##  $ kids    : int  2 2 1 2 5 2 3 4 3 5 ...
##  $ married : Factor w/ 7 levels "married","never married",..: 1 4 1 1 1 1
1 4 1 1 ...
```

```
# Use 'age' variable
data1 = data.frame(x = PSID$age)

# Density estimate
density_data = density(data1$x, na.rm = TRUE)
data2 = data.frame(ex = density_data$x, prob1 = density_data$y)

# Plot
```

```
combined_plot4 = ggplot() +
  geom_histogram(data = data1, aes(x = x, y = after_stat(density)), binwidth
= 1,
                 fill = "gold", color = "black") +
  geom_line(data = data2, aes(x = ex, y = prob1), color = "blue", linewidth =
1.5) +
  labs(x = "Age", y = "Density", title = "Density Plot over Histogram (Age in
PSID1982)") +
  theme_minimal()

combined_plot4
```



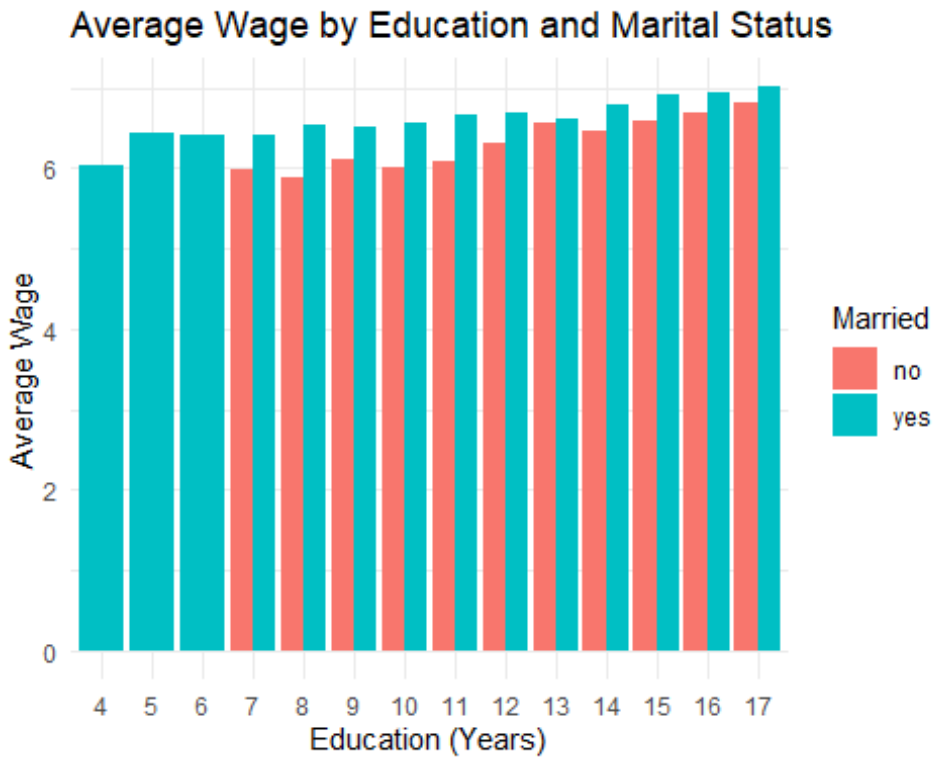
Q14. Show the average wage by education level and marital status from the Wages data.

```
data("Wages")
str(Wages)
```

```
## 'data.frame':    4165 obs. of  12 variables:
##  $ exp    : int  3 4 5 6 7 8 9 30 31 32 ...
##  $ wks    : int  32 43 40 39 42 35 32 34 27 33 ...
##  $ bluecol: Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 2 2 2 ...
##  $ ind    : int  0 0 0 0 1 1 1 0 0 1 ...
##  $ south  : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 1 1 1 ...
##  $ smsa   : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ married: Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
```

```
##  $ sex    : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ union  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 2 ...
##  $ ed     : int  9 9 9 9 9 9 9 11 11 11 ...
##  $ black  : Factor w/ 2 levels "no","yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ lwage  : num  5.56 5.72 6 6 6.06 ...
```

```r
# Group and summarize
df1 <- Wages %>%
  group_by(ed, married) %>%
  summarise(mean_wage = mean(lwage, na.rm = TRUE), .groups = "drop")

# Plot grouped bar chart
ggplot(df1, aes(x = as.factor(ed), y = mean_wage, fill = as.factor(married)))
+
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Education (Years)", y = "Average Wage", fill = "Married",
       title = "Average Wage by Education and Marital Status") +
  theme_minimal()
```



Q15. Compare number of cars by number of cylinders and gear types from the mtcars dataset.

```r
df3 <- mtcars %>%
  count(cyl, gear)

# Plot
```

```r
ggplot(df3, aes(x = as.factor(cyl), y = n, fill = as.factor(gear))) +
  geom_bar(stat = "identity", position = "dodge", color = "black") +
  scale_fill_brewer(palette = "Set2", name = "Gears") +  # Improved color
palette
  labs(x = "Number of Cylinders", y = "Number of Cars",
       title = "Distribution of Cars by Cylinders and Gears") +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    legend.position = "top"
  )
```