Assignment 1

Omid

2024-01-17

Problem 1:

Load the dataset

Install the ‘wooldridge’ package if not already installed

```r
# Install the "wooldridge" package if not already installed
if (!requireNamespace("wooldridge", quietly = TRUE)) {
  install.packages("wooldridge")
}

# install.packages("wooldridge")
```

Load the dataset

```r
library(wooldridge)

# Load the dataset
data("bwght2")

# Display a short description of the dataset
?bwght2
```

1-A) What does each observation represent? Write yourself a short description of the dataset.

> Each observation in the `bwght2` dataset represents information related to a specific birth. The dataset comprises 1832 observations, with each row corresponding to a distinct instance of childbirth. For each childbirth, various variables are recorded, providing details about the mother, father, and infant. These variables include the mother’s age, education, and race, the father’s age, education, and race, birth weight, Apgar scores (indicating the infant’s well-being just after birth), and other factors such as smoking and alcohol consumption

during pregnancy. The dataset is designed to facilitate the analysis of factors influencing infant health, with a focus on prenatal behaviors and outcomes.

- **Variables:**
  - `mage`: Mother's age in years
  - `meduc`: Mother's education in years
  - `monpre`: Month prenatal care began
  - `npvis`: Total number of prenatal visits
  - `fage`: Father's age in years
  - `feduc`: Father's education in years
  - `bwght`: Birth weight in grams
  - `omaps`: One-minute `Apgar score`
  - `fmaps`: Five-minute `Apgar score`
  - `cigs`: Average cigarettes per day
  - `drink`: Average drinks per week
  - `lbw`: Binary variable (1 if bwght <= 2000, otherwise 0)
  - `vlbw`: Binary variable (1 if bwght <= 1500, otherwise 0)
  - `male`: Binary variable (1 if baby is male, otherwise 0)
  - `mwhte`, Binary variables indicating mother's race (1 if the mother is white, otherwise 0.)
  - `mblck`, Binary variables indicating mother's race (1 if the mother is black, otherwise 0.)
  - `moth`, Binary variables indicating father's race (1 if the mother is other, otherwise 0.)
  - `fwhte`, Binary variables indicating father's race (1 if the father is white, otherwise 0.)
  - `fblck`, Binary variables indicating father's race (1 if the father is black, otherwise 0.)
  - `foth`, Binary variables indicating father's race (1 if the father is other, otherwise 0.)
  - `lbwght`: Logarithm of birth weight
  - `magesq`: Square of mother's age
  - `npvissq`: Square of the number of prenatal visits

`Apgar score` are a quick assessment tool used to evaluate the physical condition of a newborn immediately after birth. The scores are named after Dr. Virginia Apgar, who developed the system in 1952. The Apgar score is typically assessed at one minute and five minutes after birth, and occasionally at 10 minutes if needed. The score evaluates five signs

—skin color, heart rate, reflexes, muscle tone, and breathing—assigning points from 0 to 2 for each. A total score of 10 indicates the best overall health. The Apgar score helps identify newborns requiring immediate medical attention, providing a rapid snapshot of their initial well-being.

```r
if (!requireNamespace("dplyr", quietly = TRUE)) {
  install.packages("dplyr")
}

if (!requireNamespace("kableExtra", quietly = TRUE)) {
  install.packages("kableExtra")
}
```

Showing data table

```r
library(dplyr)

library(kableExtra)

head(bwght2) %>%
  kable("html", col.width = "auto") %>%
  kable_styling(full_width = FALSE, position = "center") %>%
  row_spec(0, bold = TRUE, color = "white", background = "#0073C2")
%>%
  column_spec(1:4, width = "4em") %>%
  scroll_box(width = "100%", height = "300px")
```

mage	meduc	monpre	npvis	fage	feduc	bwght	omaps	fmaps	cigs	drink	lbw	vlbw	male	mwhte	mblck	moth	fwhte	fblck	foth	lbwght	magesq	npvissq
26	12	2	12	34	16	3060	9	9	0	0	0	0	1	0	0	1	0	0	1	8.026170	676	144
29	12	2	12	32	12	3730	8	9	NA	NA	0	0	0	1	0	0	1	0	0	8.224163	841	144
33	12	1	12	36	16	2530	8	9	0	0	0	0	0	1	0	0	1	0	0	7.835975	1089	144
28	17	5	8	32	17	3289	8	9	0	0	0	0	1	1	0	0	1	0	0	8.098339	784	64
23	13	2	6	24	16	3590	6	8	0	0	0	0	1	1	0	0	1	0	0	8.185907	529	36
28	12	1	12	30	16	3420	9	9	0	0	0	0	0	1	0	0	1	0	0	8.137396	784	144

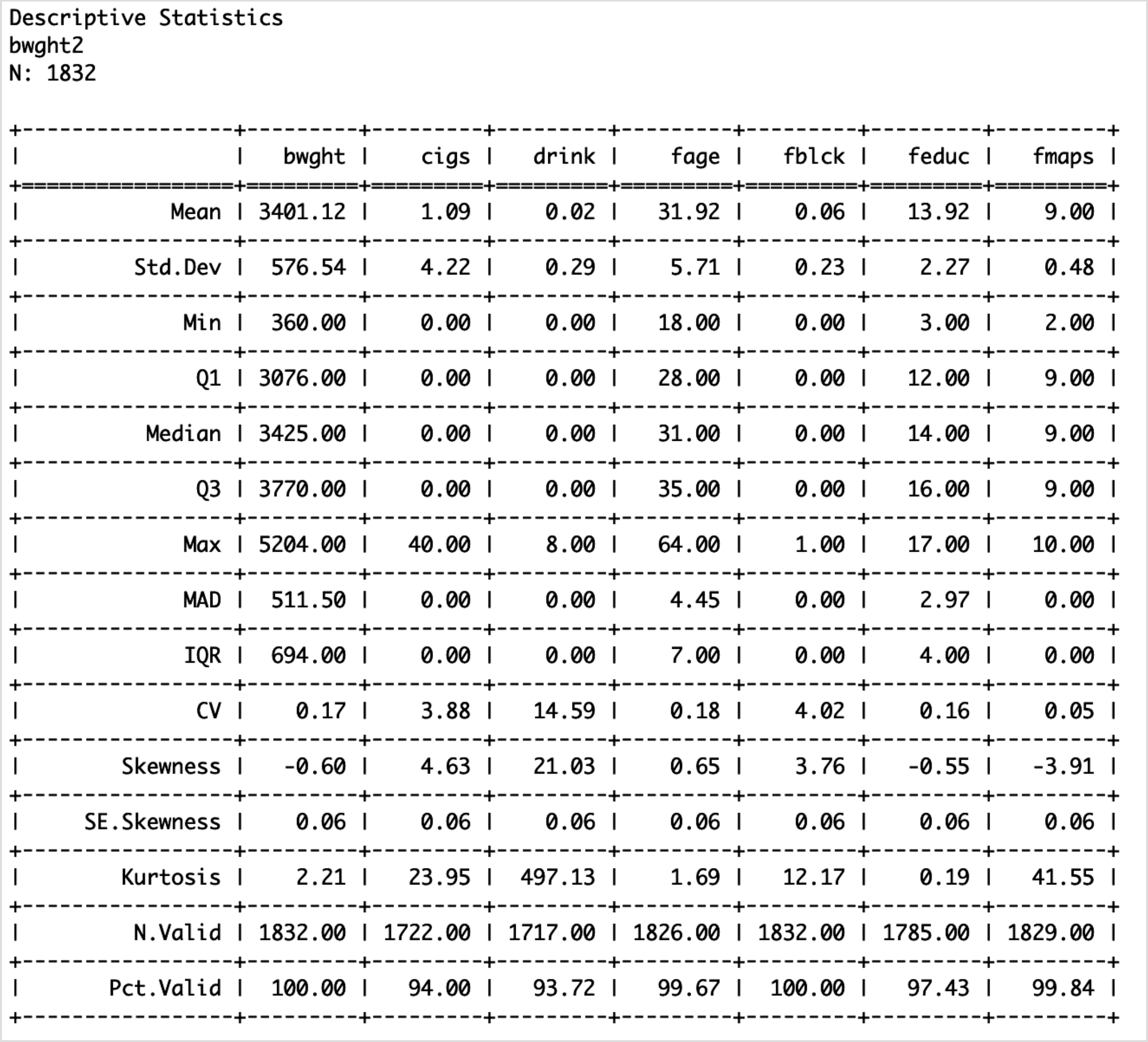
or

```r
View(bwght2)
```

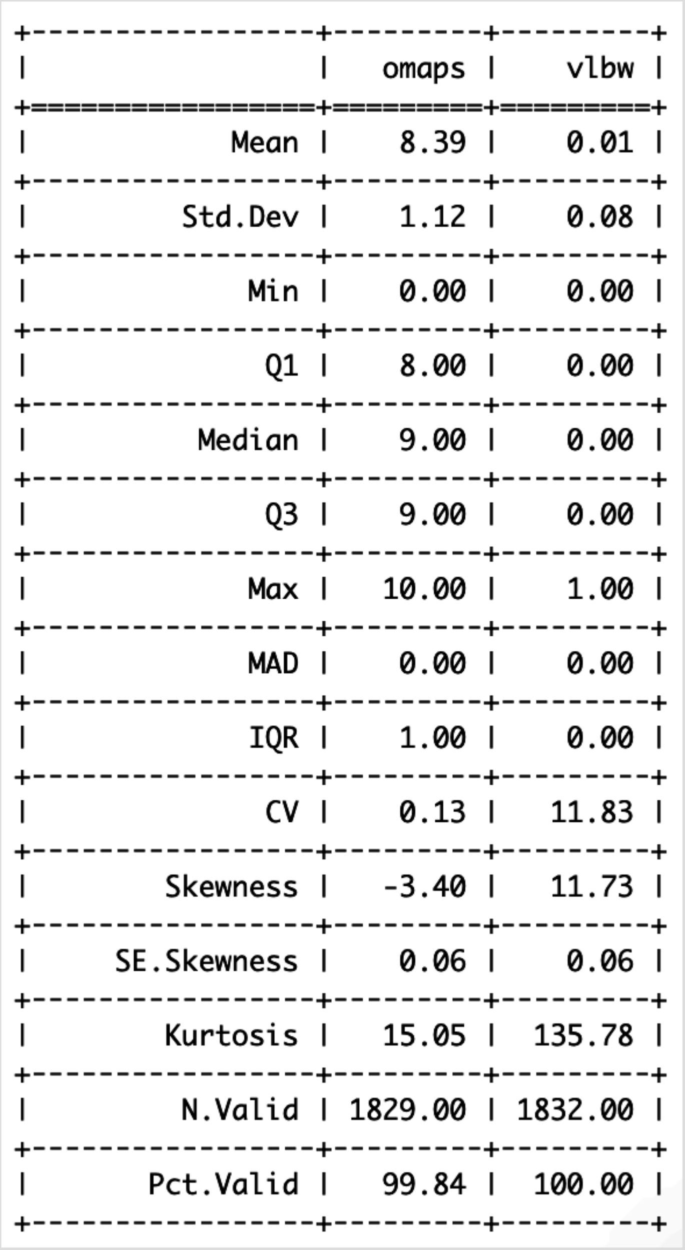
Summary of the dataset

```r
if (!requireNamespace("summarytools", quietly = TRUE)) {
  install.packages("summarytools")
}

library(summarytools)

# descr(bwght2)
descr(bwght2, style = “grid")
```

A portion of the output is:



```
Descriptive Statistics
bwght2
N: 1832

+-----------------+---------+---------+---------+---------+---------+---------+---------+
|                 |   bwght |    cigs |   drink |    fage |   fblck |   feduc |   fmaps |
+=================+=========+=========+=========+=========+=========+=========+=========+
|            Mean | 3401.12 |    1.09 |    0.02 |   31.92 |    0.06 |   13.92 |    9.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|         Std.Dev |  576.54 |    4.22 |    0.29 |    5.71 |    0.23 |    2.27 |    0.48 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|             Min |  360.00 |    0.00 |    0.00 |   18.00 |    0.00 |    3.00 |    2.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|              Q1 | 3076.00 |    0.00 |    0.00 |   28.00 |    0.00 |   12.00 |    9.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|          Median | 3425.00 |    0.00 |    0.00 |   31.00 |    0.00 |   14.00 |    9.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|              Q3 | 3770.00 |    0.00 |    0.00 |   35.00 |    0.00 |   16.00 |    9.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|             Max | 5204.00 |   40.00 |    8.00 |   64.00 |    1.00 |   17.00 |   10.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|             MAD |  511.50 |    0.00 |    0.00 |    4.45 |    0.00 |    2.97 |    0.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|             IQR |  694.00 |    0.00 |    0.00 |    7.00 |    0.00 |    4.00 |    0.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|              CV |    0.17 |    3.88 |   14.59 |    0.18 |    4.02 |    0.16 |    0.05 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|        Skewness |   -0.60 |    4.63 |   21.03 |    0.65 |    3.76 |   -0.55 |   -3.91 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|     SE.Skewness |    0.06 |    0.06 |    0.06 |    0.06 |    0.06 |    0.06 |    0.06 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|        Kurtosis |    2.21 |   23.95 |  497.13 |    1.69 |   12.17 |    0.19 |   41.55 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|         N.Valid | 1832.00 | 1722.00 | 1717.00 | 1826.00 | 1832.00 | 1785.00 | 1829.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|       Pct.Valid |  100.00 |   94.00 |   93.72 |   99.67 |  100.00 |   97.43 |   99.84 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
```



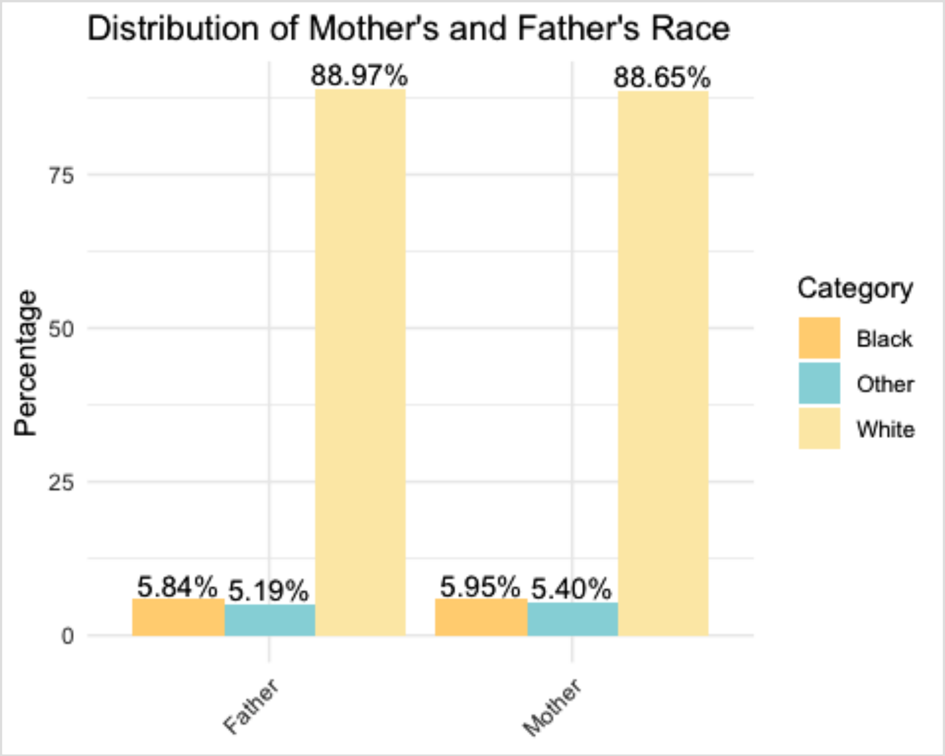
```
+-----------------+---------+---------+
|                 |   omaps |    vlbw |
+=================+=========+=========+
|            Mean |    8.39 |    0.01 |
+-----------------+---------+---------+
|         Std.Dev |    1.12 |    0.08 |
+-----------------+---------+---------+
|             Min |    0.00 |    0.00 |
+-----------------+---------+---------+
|              Q1 |    8.00 |    0.00 |
+-----------------+---------+---------+
|          Median |    9.00 |    0.00 |
+-----------------+---------+---------+
|              Q3 |    9.00 |    0.00 |
+-----------------+---------+---------+
|             Max |   10.00 |    1.00 |
+-----------------+---------+---------+
|             MAD |    0.00 |    0.00 |
+-----------------+---------+---------+
|             IQR |    1.00 |    0.00 |
+-----------------+---------+---------+
|              CV |    0.13 |   11.83 |
+-----------------+---------+---------+
|        Skewness |   -3.40 |   11.73 |
+-----------------+---------+---------+
|     SE.Skewness |    0.06 |    0.06 |
+-----------------+---------+---------+
|        Kurtosis |   15.05 |  135.78 |
+-----------------+---------+---------+
|         N.Valid | 1829.00 | 1832.00 |
+-----------------+---------+---------+
|       Pct.Valid |   99.84 |  100.00 |
+-----------------+---------+---------+
```

1-B) Choose 2-3 variables of interest here that might be related in some way. Write a paragraph or draw a diagram of your hypothesized relationship between the variables.

1-B-1) Analyzing Parental Race Distribution

It appears that a form of one-hot encoding has been applied to represent the categorical variable of race for both mothers and fathers in the bwght2 dataset. In one-hot encoding, each category is represented by a binary variable, and only one of these binary variables is "hot" (or set to 1) at a time.

```r
# Calculate demographics
demographics <- data.frame(
  Category = rep(c("White", "Black", "Other"), each = 2),
  Variable = c("Mother", "Father", "Mother", "Father", "Mother",
"Father"),
  Percent = c(
    sum(bwght2$mwhte) / nrow(bwght2) * 100,
    sum(bwght2$fwhte) / nrow(bwght2) * 100,
    sum(bwght2$mblck) / nrow(bwght2) * 100,
    sum(bwght2$fblck) / nrow(bwght2) * 100,
    sum(bwght2$moth) / nrow(bwght2) * 100,
    sum(bwght2$foth) / nrow(bwght2) * 100
  )
)

# Plotting a bar chart
library(ggplot2)

ggplot(demographics, aes(x = Variable, y = Percent, fill = Category))
+
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = sprintf("%.2f%%", Percent)), position =
position_dodge(width = 0.9), vjust = -0.2) +
  labs(title = "Distribution of Mother's and Father's Race",
       y = "Percentage",
       x = NULL) +
  scale_fill_manual(values = c("White" = "#F9EAC2", "Black" =
"#FFD898", "Other" = "#B2D7DA")) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



- The majority of both mothers and fathers are of the White race, with percentages around 89%.
- Black mothers and fathers make up a smaller proportion, around 5-6%.
- Mothers and fathers of other races combined represent about 5-5.4% of the total for each.

It seems like the race distribution is relatively similar between mothers and fathers, with the majority being White.

Compute the numerical values manually:

```
# Calculate the sum of each variable for mother and father and the
total
sum_mwhte <- sum(bwght2$mwhte)
sum_mblck <- sum(bwght2$mblck)
sum_moth <- sum(bwght2$moth)

sum_fwhte <- sum(bwght2$fwhte)
sum_fblck <- sum(bwght2$fblck)
sum_foth <- sum(bwght2$foth)
```

```r
total_mother <- sum_mwhte + sum_mblck + sum_moth
total_father <- sum_fwhte + sum_fblck + sum_foth

# Calculate percentages with 2 digits
percentage_mwhte <- sprintf("%.2f%%", sum_mwhte / total_mother * 100)
percentage_mblck <- sprintf("%.2f%%", sum_mblck / total_mother * 100)
percentage_moth <- sprintf("%.2f%%", sum_moth / total_mother * 100)

percentage_fwhte <- sprintf("%.2f%%", sum_fwhte / total_father * 100)
percentage_fblck <- sprintf("%.2f%%", sum_fblck / total_father * 100)
percentage_foth <- sprintf("%.2f%%", sum_foth / total_father * 100)

# Display the sums and percentages for mother and father
cat("Mother Data:\n")
```

```
## Mother Data:
```

```r
cat("Sum of mwhte:", sum_mwhte, ", Percentage:", percentage_mwhte,
"\n")
```

```
## Sum of mwhte: 1624 , Percentage: 88.65%
```

```r
cat("Sum of mblck:", sum_mblck, ", Percentage:", percentage_mblck,
"\n")
```

```
## Sum of mblck: 109 , Percentage: 5.95%
```

```r
cat("Sum of moth:", sum_moth, ", Percentage:", percentage_moth, "\n")
```

```
## Sum of moth: 99 , Percentage: 5.40%
```

```r
cat("Total (Mother):", total_mother, "\n\n")
```

```
## Total (Mother): 1832
```

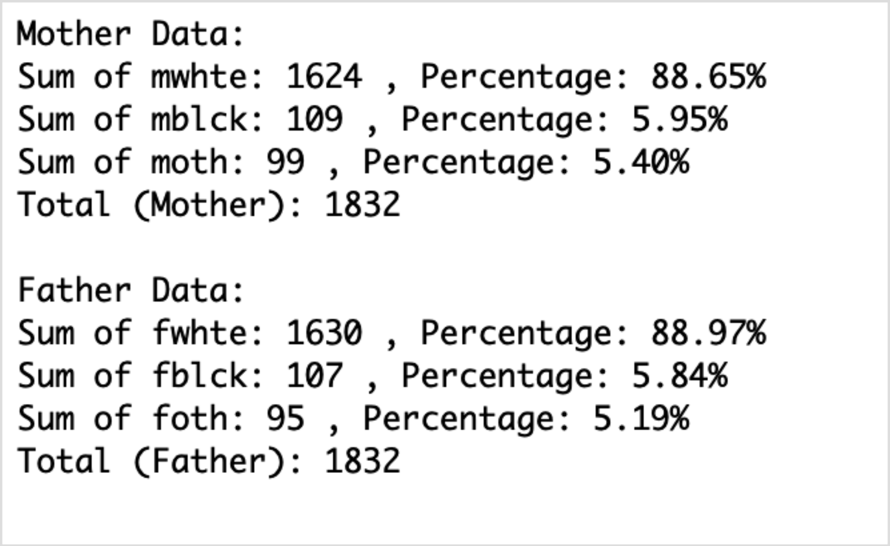
```r
cat("Father Data:\n")
```

```
## Father Data:
```

```r
cat("Sum of fwhte:", sum_fwhte, ", Percentage:", percentage_fwhte,
"\n")
```

```
## Sum of fwhte: 1630 , Percentage: 88.97%
```

```r
cat("Sum of fblck:", sum_fblck, ", Percentage:", percentage_fblck,
"\n")

## Sum of fblck: 107 , Percentage: 5.84%

cat("Sum of foth:", sum_foth, ", Percentage:", percentage_foth, "\n")

## Sum of foth: 95 , Percentage: 5.19%

cat("Total (Father):", total_father, "\n\n")

## Total (Father): 1832
```



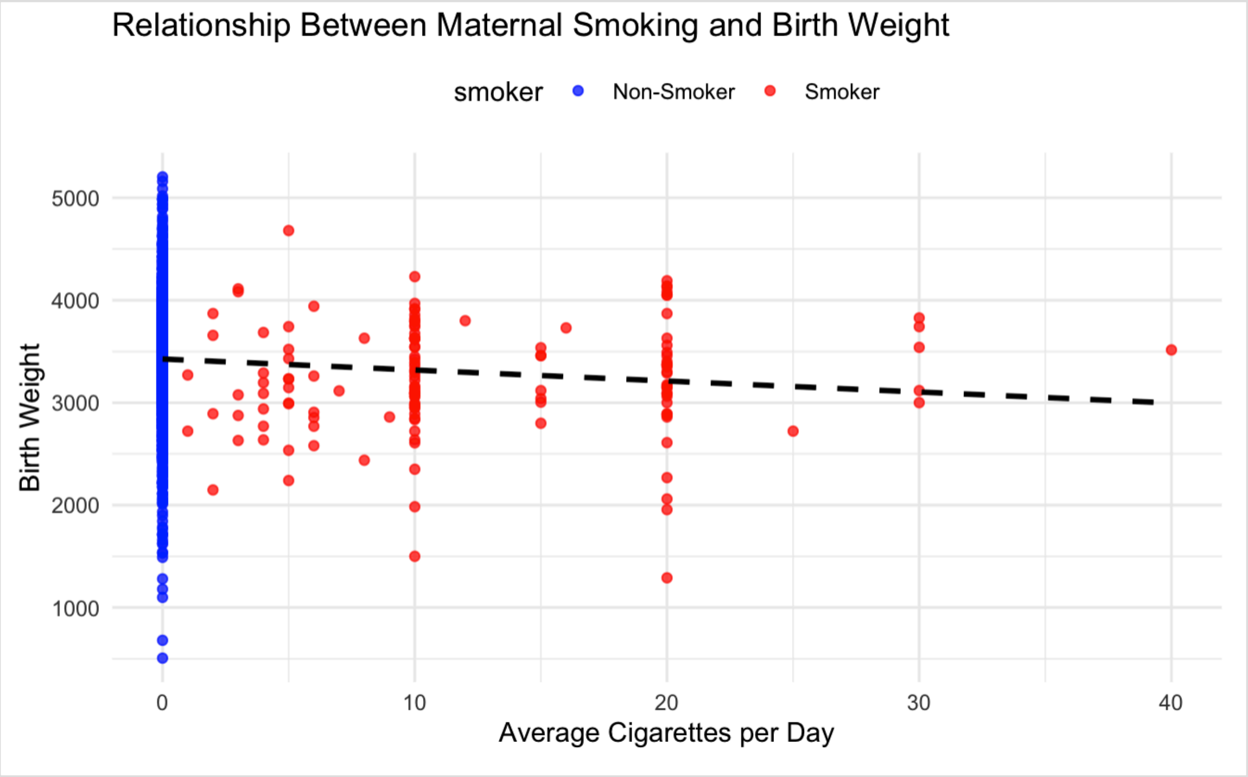
```
Mother Data:
Sum of mwhte: 1624 , Percentage: 88.65%
Sum of mblck: 109 , Percentage: 5.95%
Sum of moth: 99 , Percentage: 5.40%
Total (Mother): 1832

Father Data:
Sum of fwhte: 1630 , Percentage: 88.97%
Sum of fblck: 107 , Percentage: 5.84%
Sum of foth: 95 , Percentage: 5.19%
Total (Father): 1832
```

1-B-2) Maternal Smoking

One could hypothesize that there might be a negative relationship between maternal smoking and infant birth weight, given the known health risks associated with smoking during pregnancy. To investigate this hypothesis, we plotted the data in a scatter chart, examining the relationship between maternal smoking and infant birth weight. The chart reveals potential patterns or trends that can help us understand if there is indeed an association between maternal smoking and lower birth weights.

```r
# Load the packages
library(ggplot2)
library(scales)

# Create a binary variable 'smoker' based on the average number of
cigarettes per day
bwght2$smoker <- ifelse(bwght2$cigs > 0, "Smoker", "Non-Smoker")

# scatter plot with a trend line
ggplot(na.omit(bwght2), aes(x = cigs, y = bwght, color = smoker)) +
  geom_point(alpha = 0.7) +  # Add transparency to overlapping points
  geom_smooth(method = "lm", se = FALSE, linetype = "dashed", color =
"black") +  # Add a linear trend line
  labs(title = "Relationship Between Maternal Smoking and Birth
Weight",
       x = "Average Cigarettes per Day",
       y = "Birth Weight") +
  scale_color_manual(values = c("Smoker" = "red", "Non-Smoker" =
"blue")) +  # Customize colors
  theme_minimal() +
  theme(legend.position = "top")  # Move legend to the top
```

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



The chart shows compelling evidence of a statistically significant difference in birth weights, with non-smokers exhibiting, on average, higher birth weights compared to smokers. While visual patterns in the scatter chart suggest this difference, to rigorously support these findings, a statistical test, such as a t-test, would be imperative. The distinct separation of data points in the scatter chart indicates a potential negative relationship between maternal smoking and infant birth weight. The trend line, representing the linear fit, accentuates this downward trend. This visual representation not only aids in understanding the general pattern but also acts as a preliminary indicator of the need for formal statistical analysis.

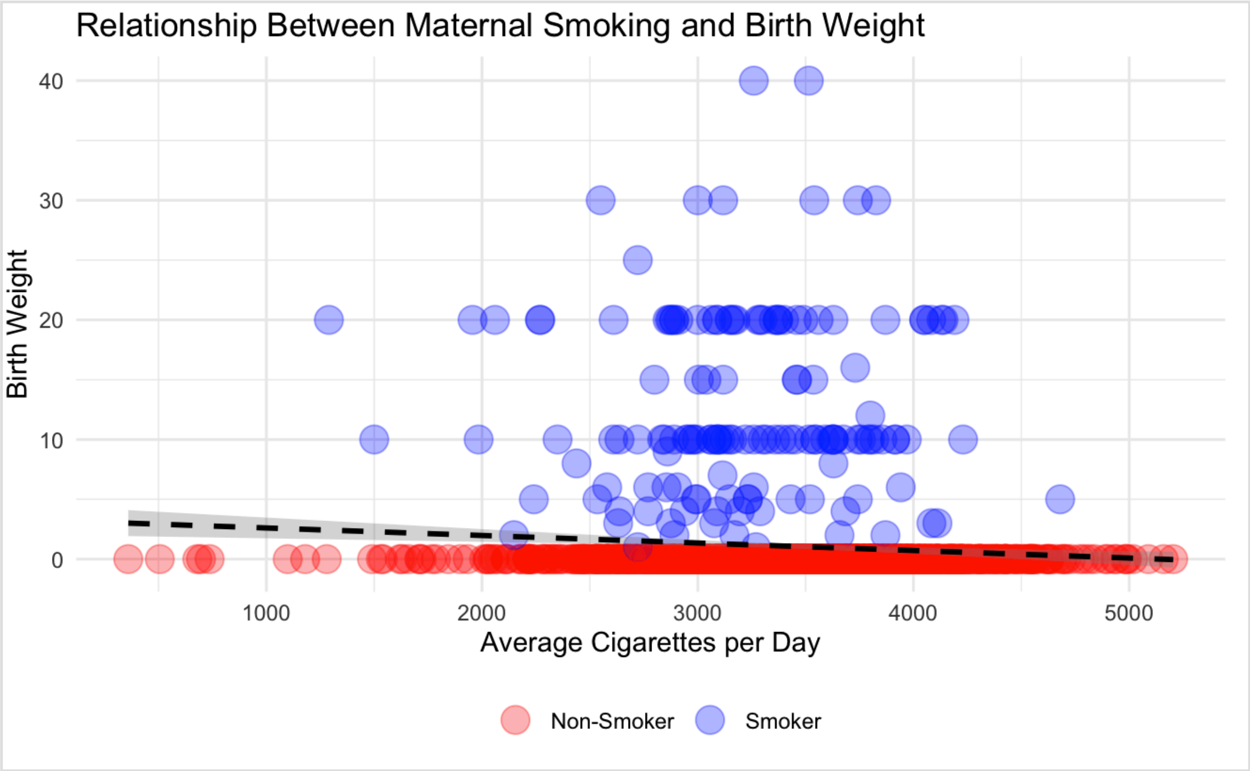
1-C) Your goal is to create a very good and a very bad visualization of the same relationship between these variables.

for good visualization: see 1-B-2

The following chart intentionally has poor visualization:

- **Misleading Axes:** The axes are swapped, showing ‘Average Cigarettes per Day’ on the y-axis and ‘Birth Weight’ on the x-axis. This misrepresentation makes it difficult to interpret any relationship accurately.
- **Inverted Colors:** Colors are assigned inversely, making it inconsistent with the interpretation. The ‘Smoker’ group is represented in blue, and the ‘Non-Smoker’ group is in red, causing confusion.

```
# Bad Visualization Example
ggplot(bwght2, aes(x = bwght, y = cigs, color = smoker)) +
  geom_point(alpha = 0.3, size = 5) +  # Large, semi-transparent
points
  geom_smooth(method = "lm", se = TRUE, linetype = "dashed", color =
"black") +
  labs(title = "Relationship Between Maternal Smoking and Birth
Weight",
       x = "Average Cigarettes per Day",
       y = "Birth Weight") +
  scale_color_manual(values = c("Smoker" = "blue", "Non-Smoker" =
"red")) +
  theme_minimal() +
  theme(legend.position = "bottom", legend.title = element_blank())  #
Move legend to the bottom and remove legend title

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

## Warning: Removed 110 rows containing non-finite values
(`stat_smooth()`).

## Warning: Removed 110 rows containing missing values
(`geom_point()`).
```



Problem 2:

Load the dataset

Install the 'dplyr' package if not already installed

```
if (!requireNamespace("dplyr", quietly = TRUE)) {
  install.packages("dplyr")
}
if (!requireNamespace("kableExtra", quietly = TRUE)) {
  install.packages("kableExtra")
}

library(dplyr)
library(kableExtra)
```

Load the dataset

```r
# Load the fertil1 dataset
data("fertil1")

# Display a short description of the dataset
?fertil1
```

2-A) What does each observation represent?

Each observation in the “fertil1” dataset represents an individual woman. The dataset contains information on various characteristics related to women’s fertility choices. Here are some key variables that describe each observation:

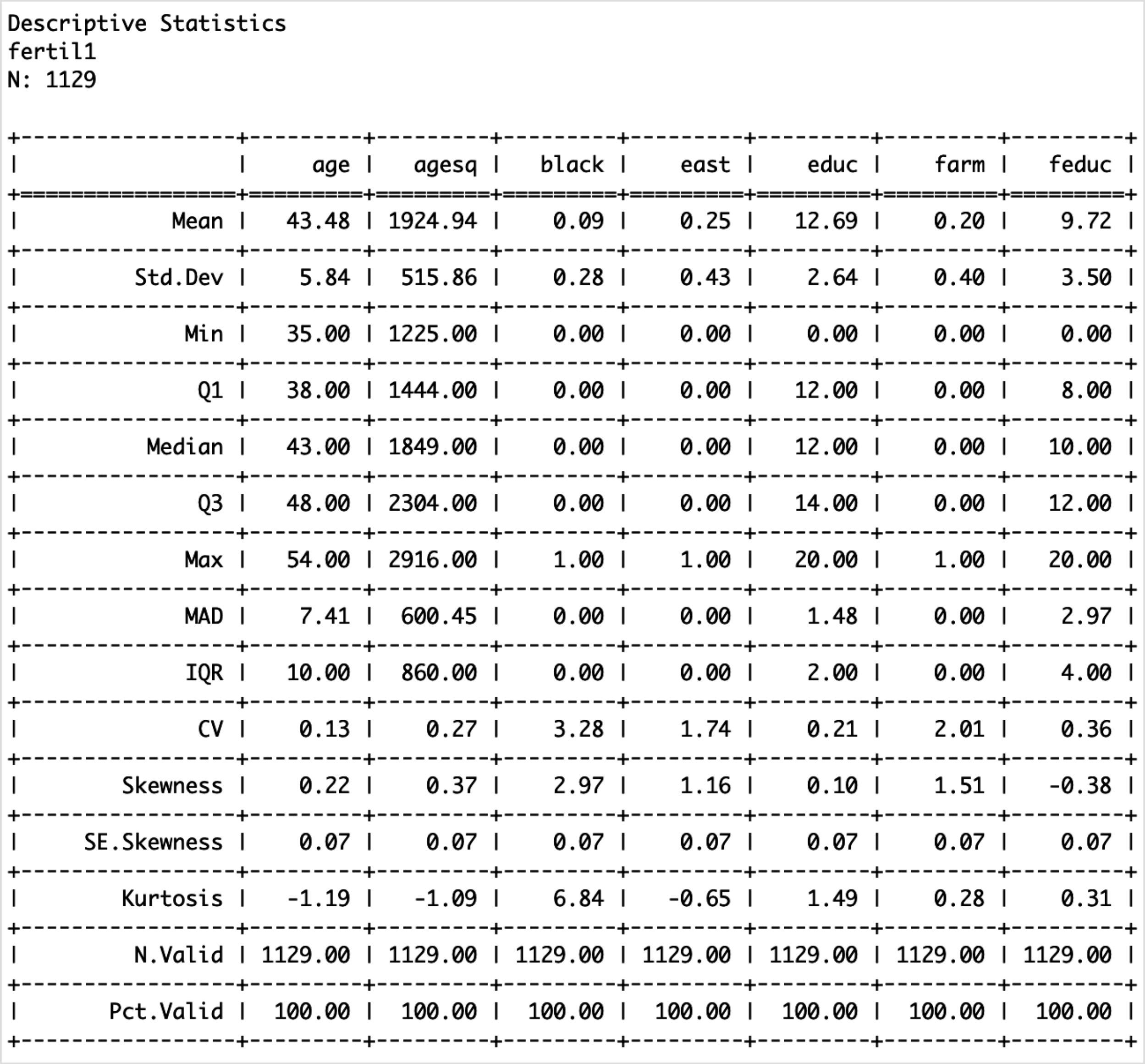
1. **year:** The year of the observation (ranging from 1972 to 1984, even years).
2. **educ:** The number of years of schooling for the woman.
3. **meduc:** The mother’s level of education.
4. **feduc:** The father’s level of education.
5. **age:** The age of the woman in years.
6. **kids:** The number of children ever born to the woman.
7. **black:** A binary variable indicating whether the woman is black (1 if black, 0 otherwise).
8. **east, northcen, west, farm, othrural, town, smcity:** Binary variables indicating the woman’s residence at the age of 16 in different regions or areas.
9. **y74, y76, y78, y80, y82, y84:** Binary variables indicating the year of the observation (1 if the year matches, 0 otherwise).
10. **agesq:** The square of the woman’s age.

Each row in the dataset represents a unique woman, and the variables provide information about her demographic characteristics, educational background, fertility, and residence details. The dataset is sourced from the National Opinion Resource Center’s General Social Survey and is a subset compiled for the study of the effect of women’s schooling on fertility.

2-B) Summary stats for the variable educ

Summary table for all columns (A portion of the output):

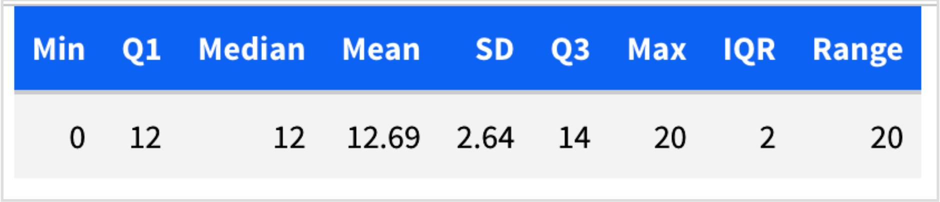
```
library(summarytools)
#descr(fertil1)
descr(fertil1, style = "grid")
```



```
Descriptive Statistics
fertil1
N: 1129

+-----------------+---------+---------+---------+---------+---------+---------+---------+
|                 |     age |   agesq |   black |    east |    educ |    farm |   feduc |
+=================+=========+=========+=========+=========+=========+=========+=========+
|            Mean |   43.48 | 1924.94 |    0.09 |    0.25 |   12.69 |    0.20 |    9.72 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|         Std.Dev |    5.84 |  515.86 |    0.28 |    0.43 |    2.64 |    0.40 |    3.50 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|             Min |   35.00 | 1225.00 |    0.00 |    0.00 |    0.00 |    0.00 |    0.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|              Q1 |   38.00 | 1444.00 |    0.00 |    0.00 |   12.00 |    0.00 |    8.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|          Median |   43.00 | 1849.00 |    0.00 |    0.00 |   12.00 |    0.00 |   10.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|              Q3 |   48.00 | 2304.00 |    0.00 |    0.00 |   14.00 |    0.00 |   12.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|             Max |   54.00 | 2916.00 |    1.00 |    1.00 |   20.00 |    1.00 |   20.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|             MAD |    7.41 |  600.45 |    0.00 |    0.00 |    1.48 |    0.00 |    2.97 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|             IQR |   10.00 |  860.00 |    0.00 |    0.00 |    2.00 |    0.00 |    4.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|              CV |    0.13 |    0.27 |    3.28 |    1.74 |    0.21 |    2.01 |    0.36 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|        Skewness |    0.22 |    0.37 |    2.97 |    1.16 |    0.10 |    1.51 |   -0.38 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|     SE.Skewness |    0.07 |    0.07 |    0.07 |    0.07 |    0.07 |    0.07 |    0.07 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|        Kurtosis |   -1.19 |   -1.09 |    6.84 |   -0.65 |    1.49 |    0.28 |    0.31 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|         N.Valid | 1129.00 | 1129.00 | 1129.00 | 1129.00 | 1129.00 | 1129.00 | 1129.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
|       Pct.Valid |  100.00 |  100.00 |  100.00 |  100.00 |  100.00 |  100.00 |  100.00 |
+-----------------+---------+---------+---------+---------+---------+---------+---------+
```

Summary stats for the variable educ

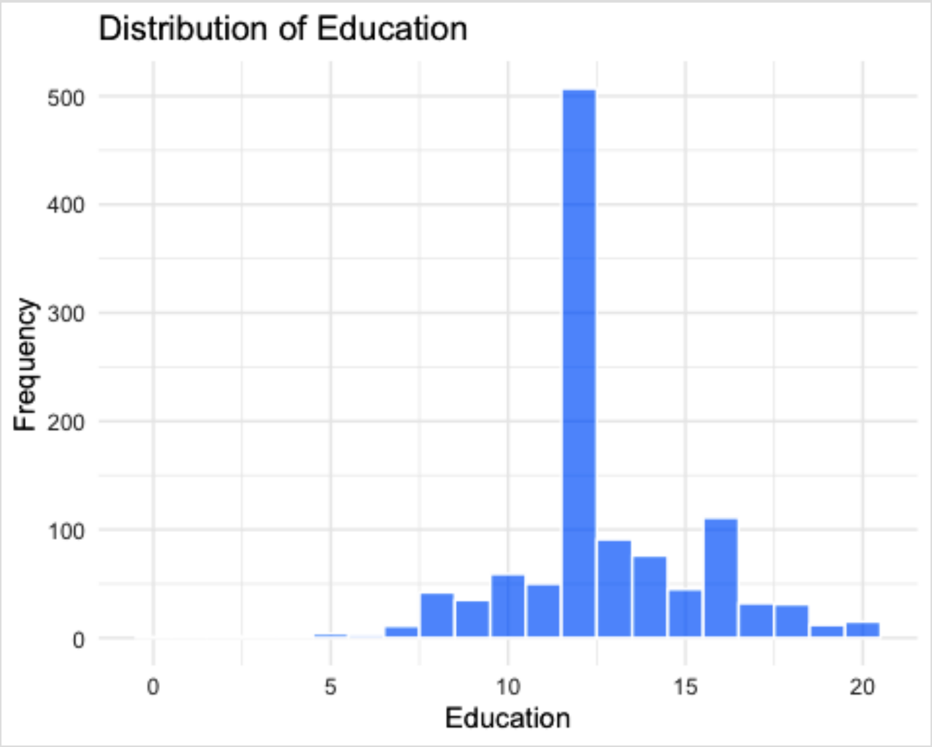
```
# Install required packages if not already installed
if (!requireNamespace("knitr", quietly = TRUE)) {
  install.packages("knitr")
}

if (!requireNamespace("kableExtra", quietly = TRUE)) {
  install.packages("kableExtra")
}

# Load the packages
library(knitr)
library(kableExtra)

# Calculate summary statistics for educ
educ_summary <- fertil1 %>%
  summarise(
    Min = min(educ),
    Q1 = quantile(educ, 0.25),
    Median = median(educ),
    Mean = round(mean(educ), 2),
    SD = round(sd(educ), 2),
    Q3 = quantile(educ, 0.75),
    Max = max(educ),
    IQR = IQR(educ), # QR=Q3-Q1
    Range = max(educ) - min(educ)
  )

# Create a table using kable
kable(educ_summary, "html") %>%
  kable_styling(full_width = FALSE) %>%
  row_spec(0, bold = TRUE, background = "#4285F4", color = "white")
%>%
  row_spec(1, background = c("#F5F5F5", "white"))
```



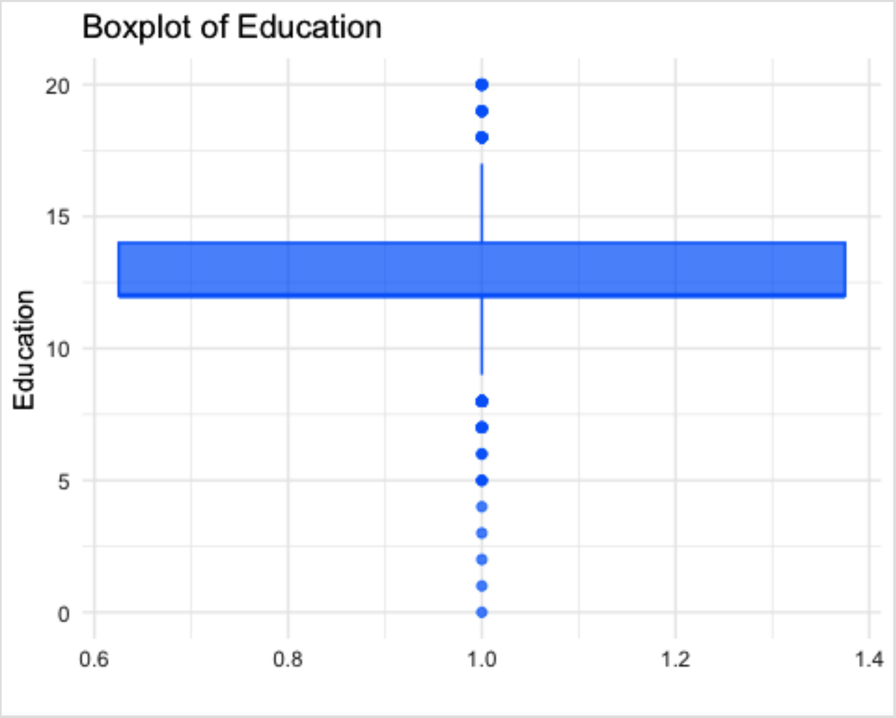
Min	Q1	Median	Mean	SD	Q3	Max	IQR	Range
0	12	12	12.69	2.64	14	20	2	20

Also a boxplot or a histogram can be provides a clear overview of the distribution.

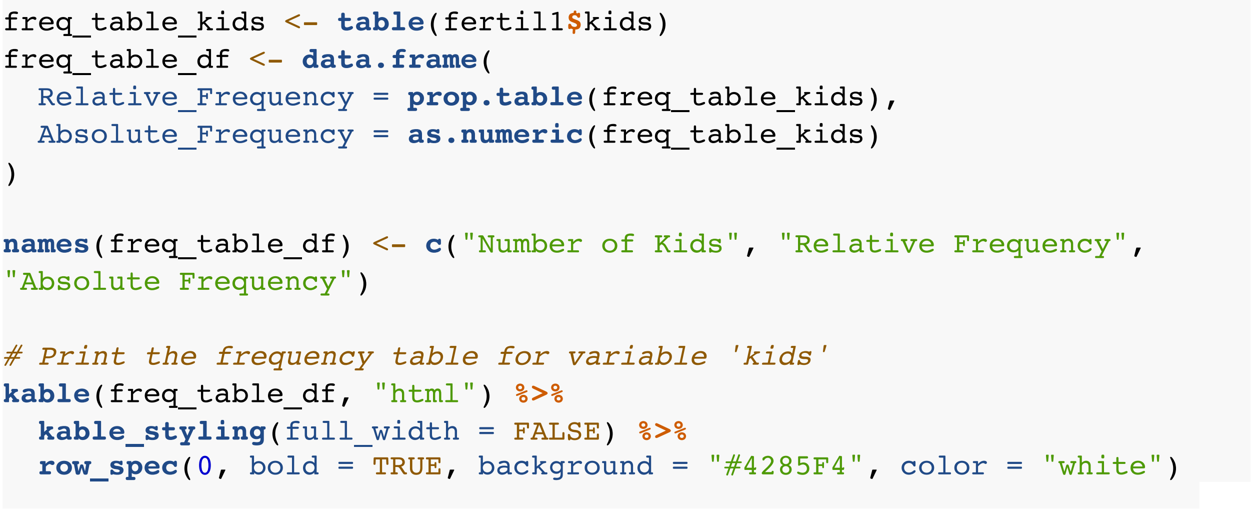
```
# Create a histogram for variable 'educ'
ggplot(fertil1, aes(x = educ, fill = cut(educ, breaks = 20))) +

  geom_histogram(binwidth = 1, color = "white", fill = "#4285F4",
alpha = 0.7) +

  labs(x = "Education", y = "Frequency", title = "Distribution of
Education") +
  theme_minimal() +
  theme(legend.position = "none")
```



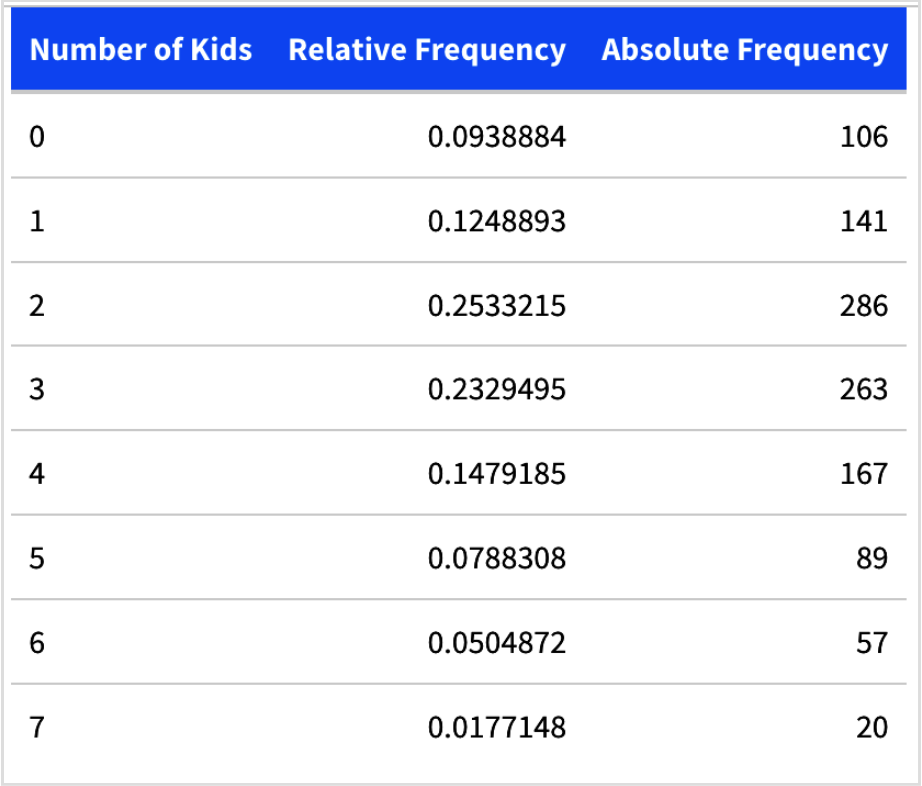
```r
# Create a boxplot for variable 'educ'
ggplot(fertil1, aes(x = 1, y = educ)) +
  geom_boxplot(fill = "#4285F4", color = "#4285F4", alpha = 0.7) +
  labs(x = "", y = "Education", title = "Boxplot of Education") +
  theme_minimal()
```



2-C) Frequency table for the variable kids



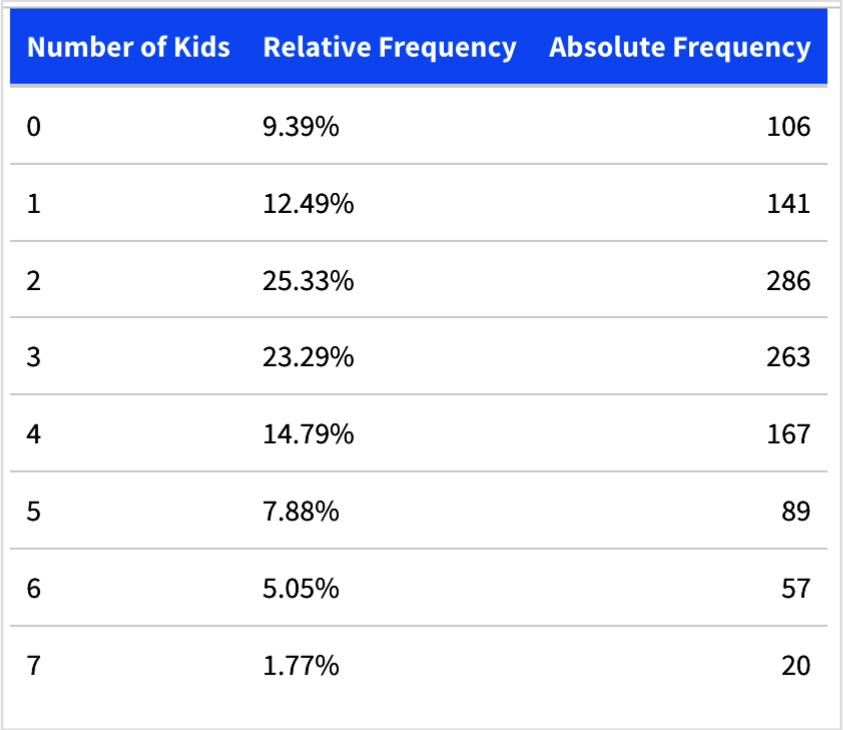
```r
freq_table_kids <- table(fertil1$kids)
freq_table_df <- data.frame(
  Relative_Frequency = prop.table(freq_table_kids),
  Absolute_Frequency = as.numeric(freq_table_kids)
)

names(freq_table_df) <- c("Number of Kids", "Relative Frequency",
"Absolute Frequency")

# Print the frequency table for variable 'kids'
kable(freq_table_df, "html") %>%
  kable_styling(full_width = FALSE) %>%
  row_spec(0, bold = TRUE, background = "#4285F4", color = "white")
```



Number of Kids	Relative Frequency	Absolute Frequency
0	0.0938884	106
1	0.1248893	141
2	0.2533215	286
3	0.2329495	263
4	0.1479185	167
5	0.0788308	89
6	0.0504872	57
7	0.0177148	20

To display the “Relative Frequency” column with a precision of two digits, we can modify the **kable** function by formatting the column appropriately.

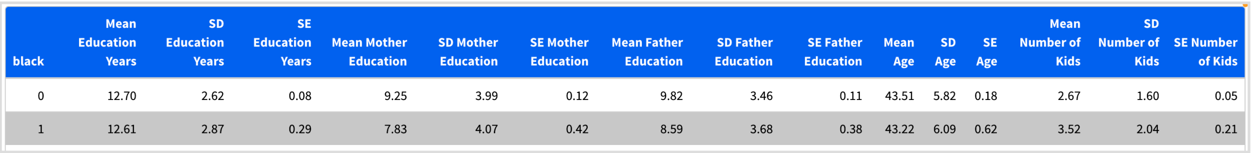
```
freq_table_kids <- table(fertil1$kids)
freq_table_df <- data.frame(
  Number_of_Kids = names(freq_table_kids),
  Relative_Frequency = sprintf("%.2f%%", 100 *
prop.table(freq_table_kids)),
  Absolute_Frequency = as.numeric(freq_table_kids)
)

names(freq_table_df) <- c("Number of Kids", "Relative Frequency",
"Absolute Frequency")

# Print the frequency table for variable 'kids'
kable(freq_table_df, "html") %>%
  kable_styling(full_width = FALSE) %>%
  row_spec(0, bold = TRUE, background = "#4285F4", color = "white")
```



Number of Kids	Relative Frequency	Absolute Frequency
0	9.39%	106
1	12.49%	141
2	25.33%	286
3	23.29%	263
4	14.79%	167
5	7.88%	89
6	5.05%	57
7	1.77%	20

2-D) Summary table for selected variables across a group variable (e.g., racial groups)

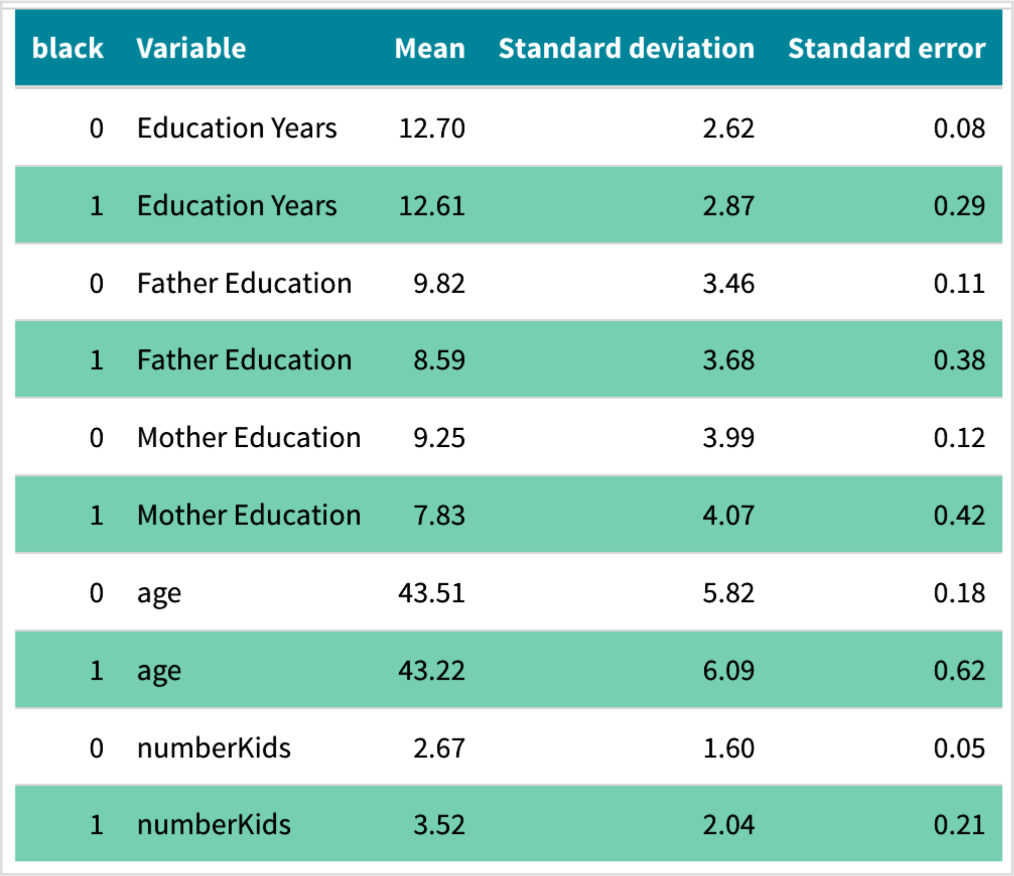
```
library(dplyr)
library(tidyr)
library(kableExtra)

# Group by 'black' and summarize variables
summary_table <- fertil1 %>%
  group_by(black) %>%
  summarise(
    mean_educationYears = round(mean(educ, na.rm = TRUE), 2),
    sd_educationYears = round(sd(educ, na.rm = TRUE), 2),
    se_educationYears = round(sd(educ, na.rm = TRUE) / sqrt(n()), 2),

    mean_motherEducation = round(mean(meduc, na.rm = TRUE), 2),
    sd_motherEducation = round(sd(meduc, na.rm = TRUE), 2),
    se_motherEducation = round(sd(meduc, na.rm = TRUE) / sqrt(n()),
2),

    mean_fatherEducation = round(mean(feduc, na.rm = TRUE), 2),
    sd_fatherEducation = round(sd(feduc, na.rm = TRUE), 2),
    se_fatherEducation = round(sd(feduc, na.rm = TRUE) / sqrt(n()),
2),

    mean_age = round(mean(age, na.rm = TRUE), 2),
    sd_age = round(sd(age, na.rm = TRUE), 2),
    se_age = round(sd(age, na.rm = TRUE) / sqrt(n()), 2),

    mean_numberKids = round(mean(kids, na.rm = TRUE), 2),
    sd_numberKids = round(sd(kids, na.rm = TRUE), 2),
    se_numberKids = round(sd(kids, na.rm = TRUE) / sqrt(n()), 2)
  )

summary_table_new <- summary_table
# Rename columns without underscores
colnames(summary_table_new) <- c(
  "black",
  "Mean Education Years", "SD Education Years", "SE Education Years",
  "Mean Mother Education", "SD Mother Education", "SE Mother
Education",
  "Mean Father Education", "SD Father Education", "SE Father
Education",
  "Mean Age", "SD Age", "SE Age",
  "Mean Number of Kids", "SD Number of Kids", "SE Number of Kids"
)

# Print the updated summary table
kable(summary_table_new, "html") %>%
  kable_styling(full_width = FALSE) %>%
  row_spec(0, bold = TRUE, background = "#4285F4", color = "white")
%>%
  row_spec(which(summary_table_new$black == TRUE), background =
"lightgrey")  # Highlight rows where black is TRUE
```



black	Mean Education Years	SD Education Years	SE Education Years	Mean Mother Education	SD Mother Education	SE Mother Education	Mean Father Education	SD Father Education	SE Father Education	Mean Age	SD Age	SE Age	Mean Number of Kids	SD Number of Kids	SE Number of Kids
0	12.70	2.62	0.08	9.25	3.99	0.12	9.82	3.46	0.11	43.51	5.82	0.18	2.67	1.60	0.05
1	12.61	2.87	0.29	7.83	4.07	0.42	8.59	3.68	0.38	43.22	6.09	0.62	3.52	2.04	0.21

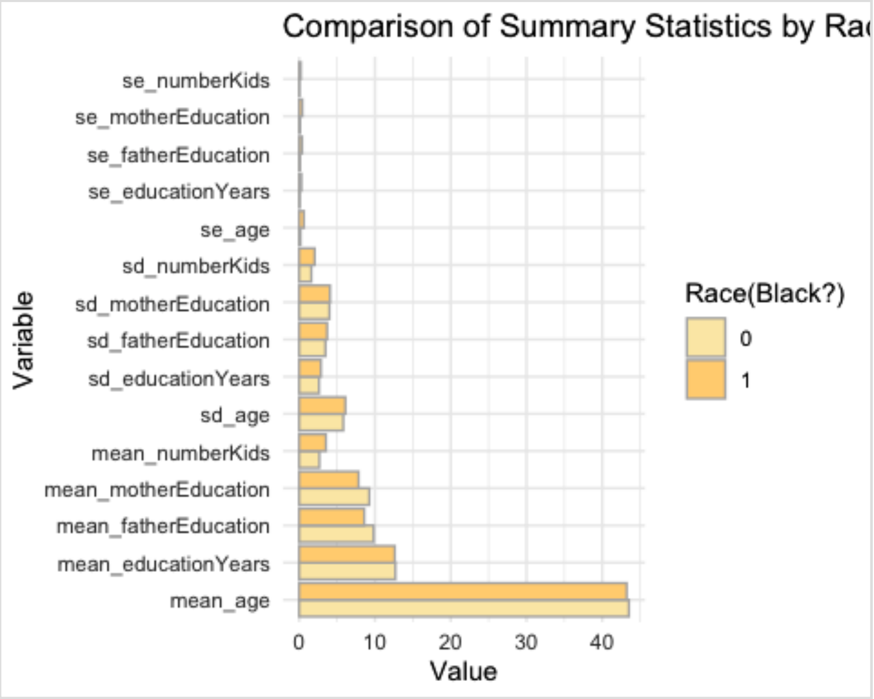
To enhance the presentation, a summarized pivot table is created in the following section.

```r
# Define a function to rename specific variables
rename_variable <- function(variable) {
  case_when(
    variable == "educationYears" ~ "Education Years",
    variable == "fatherEducation" ~ "Father Education",
    variable == "motherEducation" ~ "Mother Education",
    variable == "numberKidsnumberKids" ~ "Number of Kids",
    TRUE ~ variable
  )
}

# Pivot the summary table
summary_table_long <- summary_table %>%
  pivot_longer(cols = starts_with("mean") | starts_with("sd") |
starts_with("se"),
               names_to = c(".value", "Variable"),
               names_sep = "_") %>%
  rename(
    'Mean' = mean,
    'Standard deviation' = sd,
    'Standard error' = se
  ) %>%
  mutate(
    "Variable" = rename_variable(Variable)
  ) %>%
  arrange(Variable)

# Print the updated summary table
kable(summary_table_long, "html") %>%
  kable_styling(full_width = FALSE) %>%
  row_spec(0, bold = TRUE, background = "#29A0B1", color = "white")
%>%
  row_spec(which(summary_table_long$black == TRUE), background =
"#98D7C2")  # Highlight rows where black is TRUE
```



black	Variable	Mean	Standard deviation	Standard error
0	Education Years	12.70	2.62	0.08
1	Education Years	12.61	2.87	0.29
0	Father Education	9.82	3.46	0.11
1	Father Education	8.59	3.68	0.38
0	Mother Education	9.25	3.99	0.12
1	Mother Education	7.83	4.07	0.42
0	age	43.51	5.82	0.18
1	age	43.22	6.09	0.62
0	numberKids	2.67	1.60	0.05
1	numberKids	3.52	2.04	0.21

```r
library(ggplot2)

# Reshape data
summary_table_race_long <- tidyr::gather(summary_table, key =
"Variable", value = "Value", -black)

# Create a horizontal bar plot
ggplot(summary_table_race_long, aes(x = Value, y = Variable, fill =
factor(black))) +
  geom_bar(stat = "identity", position = "dodge", color = "grey") +
  labs(title = "Comparison of Summary Statistics by Race",
       x = "Value",
       y = "Variable") +
  scale_fill_manual(values = c("0" = "#F9EAC2", "1" = "#FFD898"), name
= "Race(Black?)") +
  theme_minimal() +
  theme(axis.text.y = element_text(angle = 0, hjust = 1)) +
  scale_y_discrete(labels = c("age" = "Age", "educ" = "Education",
"kids" = "Children"))
```


Comparison of Summary Statistics by Rac
Variable
se_numberKids
se_motherEducation
se_fatherEducation
se_educationYears
se_age
sd_numberKids
sd_motherEducation
sd_fatherEducation
sd_educationYears
sd_age
mean_numberKids
mean_motherEducation
mean_fatherEducation
mean_educationYears
mean_age
Value
Race(Black?)