ANALYZING SWISS DATA SET

About Dataset

The "swiss" dataset is a dataset containing socio-economic data from various regions of Switzerland in the late 19th century. This dataset was compiled by Ernst Engel in 1888.

The variables in the "swiss" dataset and their descriptions are as follows:

- Fertility: Fertility rate of women aged 35-79 in 1888 (number of children per birth).
- Agriculture: Percentage of total land area devoted to agricultural products.
- Examination: Results of military tests in 1888 for men of military age (%).
- Education: Literacy rate of men aged 20-24 in 1888 (%).
- Catholic: Percentage of Catholic population (%).
- Infant.Mortality: Infant mortality rate (number of infants who died per 1000 births).

This dataset can be used to examine the effects of socio-economic factors on fertility, agriculture, education, and other variables. For example, it can be used to investigate the relationship between fertility rate and agriculture or the impact of education level on the number of children per birth.

Packages used in this study:

```
install.packages("tidyverse")
```

```
Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
(as 'lib' is unspecified)
```

```
install.packages("ggplot2")
```

```
Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.3'
(as 'lib' is unspecified)
```

```
library(ggplot2)
library(tidyverse)
```

```
-- 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 lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.1

-- 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 beco
```

Loading dataset:

```
data(swiss)
```

Examining the structure of the dataset:

```
str(swiss)
```

```
'data.frame':   47 obs. of  6 variables:
 $ Fertility       : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...
 $ Agriculture     : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...
 $ Examination     : int  15 6 5 12 17 9 16 14 12 16 ...
 $ Education       : int  12 9 5 7 15 7 7 8 7 13 ...
 $ Catholic        : num  9.96 84.84 93.4 33.77 5.16 ...
 $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...
```

View the first 6 observations:

```
head(swiss)
```

```
             Fertility Agriculture Examination Education Catholic
Courtelary        80.2        17.0          15        12     9.96
Delemont          83.1        45.1           6         9    84.84
Franches-Mnt      92.5        39.7           5         5    93.40
Moutier           85.8        36.5          12         7    33.77
Neuveville        76.9        43.5          17        15     5.16
Porrentruy        76.1        35.3           9         7    90.57
             Infant.Mortality
Courtelary               22.2
```

```
Delemont                22.2
Franches-Mnt            20.2
Moutier                 20.3
Neuveville              20.6
Porrentruy              26.6
```

Get summary statistics:

```
summary(swiss)
```

```
  Fertility      Agriculture     Examination      Education
 Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00
 1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00
 Median :70.40   Median :54.10   Median :16.00   Median : 8.00
 Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98
 3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00
 Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00
    Catholic       Infant.Mortality
 Min.   :  2.150   Min.   :10.80
 1st Qu.:  5.195   1st Qu.:18.15
 Median : 15.140   Median :20.00
 Mean   : 41.144   Mean   :19.94
 3rd Qu.: 93.125   3rd Qu.:21.70
 Max.   :100.000   Max.   :26.60
```

Check for missing observations:

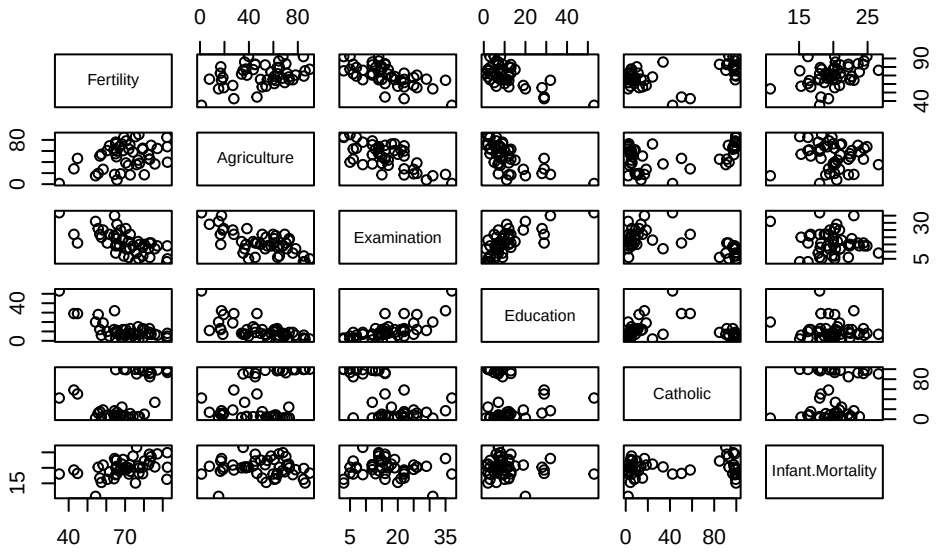
```
any(is.na(swiss))
```

```
[1] FALSE
```

EXPLORATORY DATA ANALYSIS:

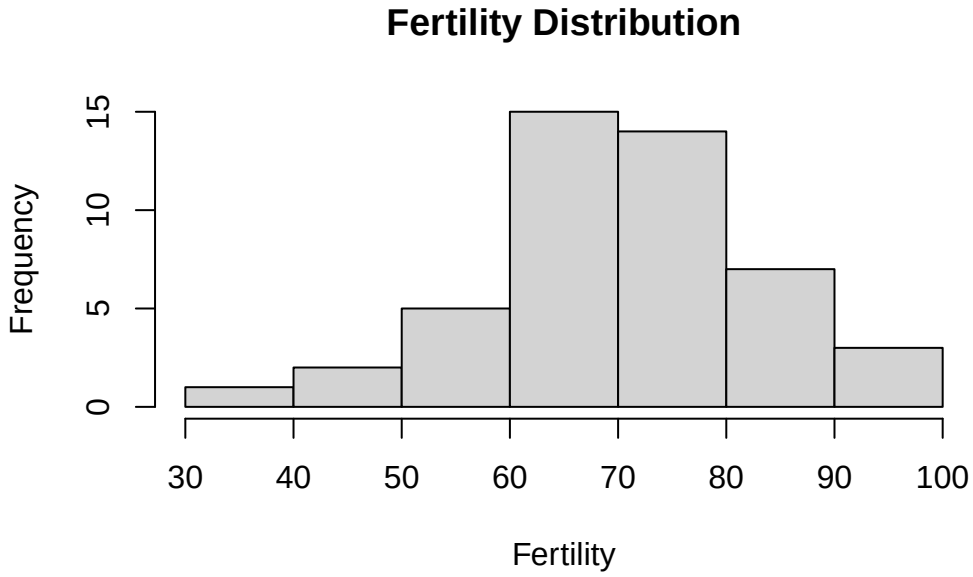
1. Scatter plots showing the relationship between variables:

```
pairs(swiss)
```

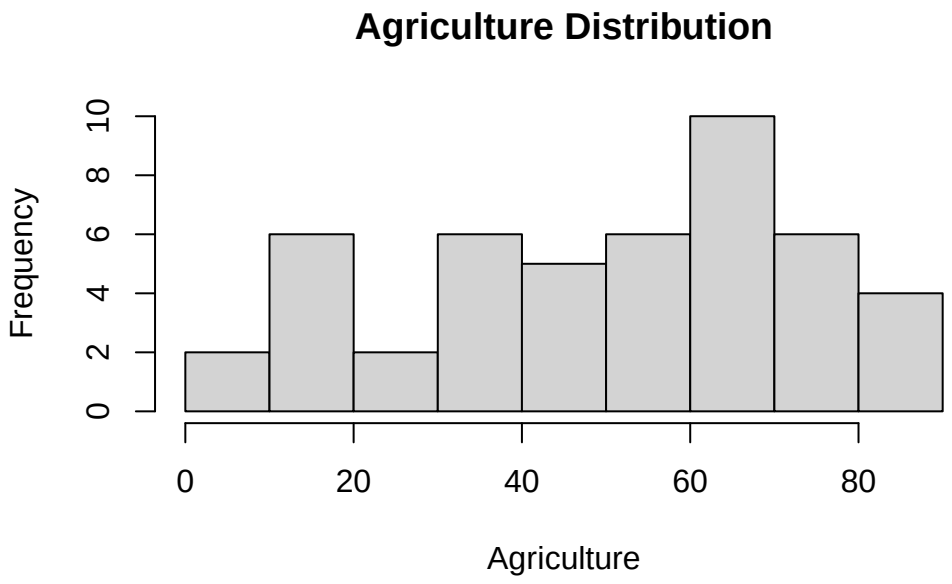


2. Histograms showing the distribution of variables:

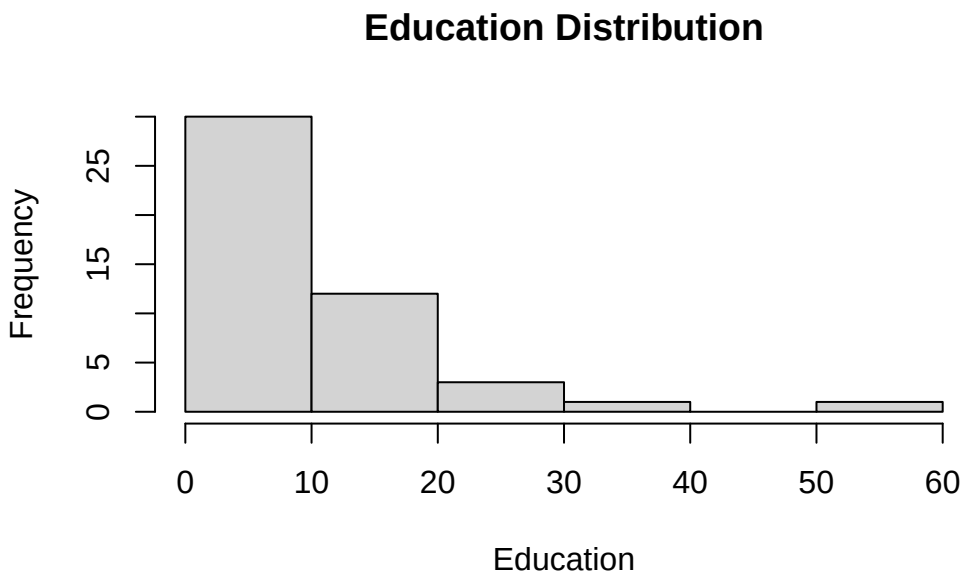
```
hist(swiss$Fertility, main = "Fertility Distribution", xlab = "Fertility")
```



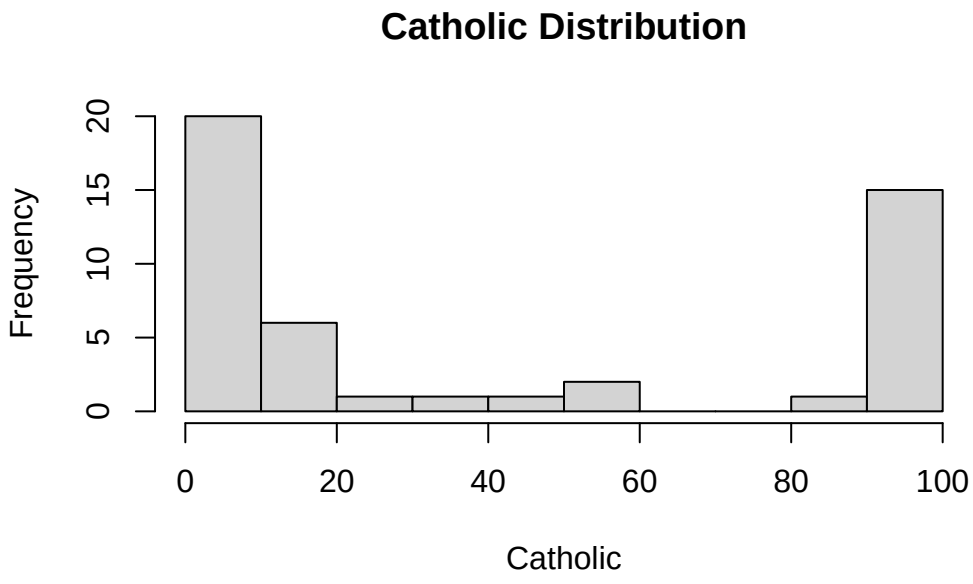

```
hist(swiss$Agriculture, main = "Agriculture Distribution", xlab = "Agriculture")
```



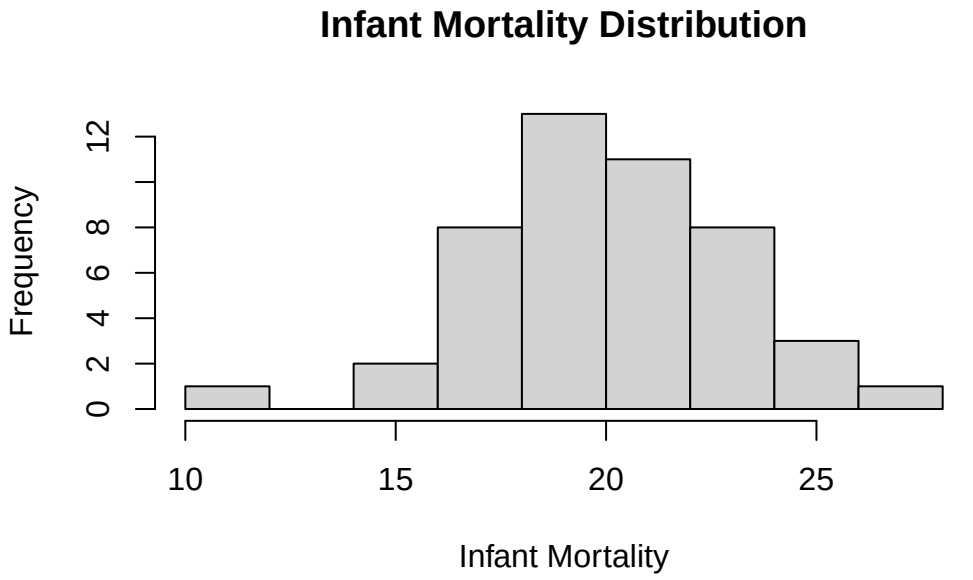

```
hist(swiss$Education, main = "Education Distribution", xlab = "Education")
```




```
hist(swiss$Catholic, main = "Catholic Distribution", xlab = "Catholic")
```




```
hist(swiss$Infant.Mortality, main = "Infant Mortality Distribution", xlab = "Infant Mort
```

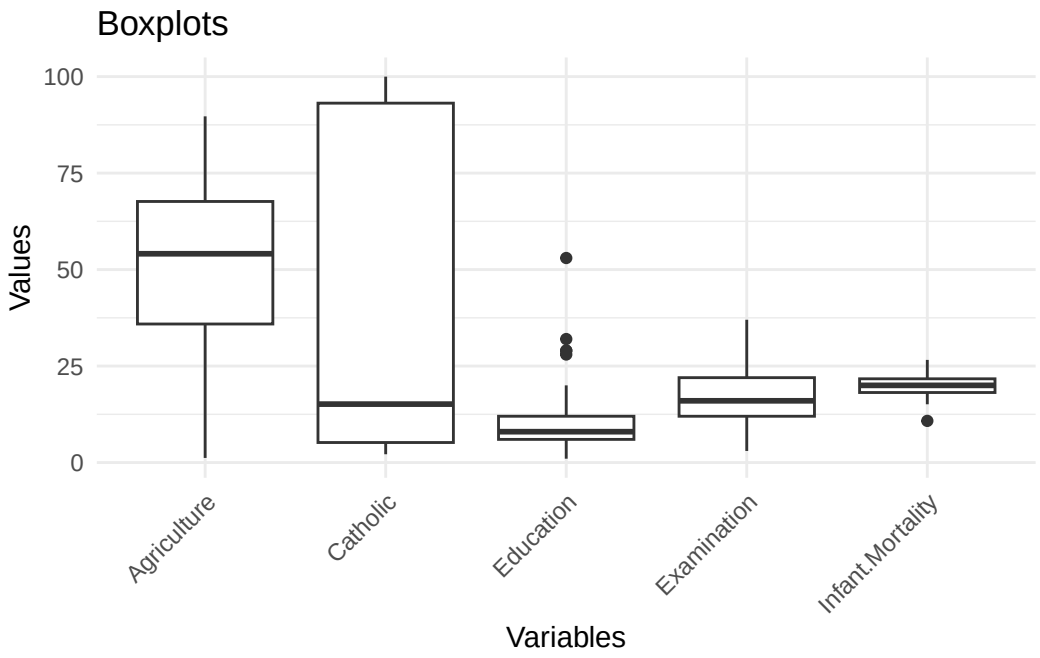


3. Boxplots:

```
library(tidyverse)
# Convert the dataset to appropriate format
swiss_df <- swiss %>%
  gather(key = "variable", value = "value", -1)

ggplot(swiss_df, aes(x = variable, y = value)) +
```

```
  geom_boxplot() +
  labs(x = "Variables", y = "Values", title = "Boxplots") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Interpretation of the Results:
According to the above graphs, we can obtain the following results:

- As the fertility rate decreased, the area allocated to agriculture increased.
- There is a negative relationship between education level and fertility, that is, as the education level increases, the fertility rate decreases.
- As the level of education increased, the area allocated to agriculture decreased.
- As the area allocated to agriculture increases, the fertility rate decreases.
- No obvious relationship seems to be observed between the examination and other variables.
- No clear trend is evident between examination and other variables.
- There is no clear trend between the variables and the proportion of the Catholic population.
- The infant mortality rate seems to decrease as education level increases.

REGRESSION ANALYSIS:

Examining the relationship between fertility and education:

```
reg_model <- lm(Fertility ~ Education, data = swiss)
summary(reg_model)
```

```
Call:
lm(formula = Fertility ~ Education, data = swiss)

Residuals:
    Min      1Q  Median      3Q     Max
-17.036  -6.711  -1.011   9.526  19.689

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  79.6101     2.1041  37.836  < 2e-16 ***
Education    -0.8624     0.1448  -5.954 3.66e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.446 on 45 degrees of freedom
Multiple R-squared:  0.4406,    Adjusted R-squared:  0.4282
F-statistic: 35.45 on 1 and 45 DF,  p-value: 3.659e-07
```

ANOVA:

Evaluating the effect of the "Catholic" variable on "Fertility":

```
anova <- aov(Fertility ~ as.factor(Catholic), data = swiss)

summary(anova)
```

```
                   Df Sum Sq Mean Sq F value Pr(>F)
as.factor(Catholic) 45   7140  158.67   4.193  0.372
Residuals            1     38   37.84
```

Because the p value equals to 0.372 we can say that this model is not significant and it is obvious that there is not any effect of the "Catholic" variable on "Fertility".

ANCOVA:

Examining the religious affiliation (Catholic) of the cantons by checking the relationship between fertility and education:

```r
model_ancova <- lm(Fertility ~ Education + Catholic, data = swiss)

summary(model_ancova)
```

```
Call:
lm(formula = Fertility ~ Education + Catholic, data = swiss)

Residuals:
    Min      1Q  Median      3Q     Max
-15.042  -6.578  -1.431   6.122  14.322

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 74.23369    2.35197  31.562  < 2e-16 ***
Education   -0.78833    0.12929  -6.097 2.43e-07 ***
Catholic     0.11092    0.02981   3.721  0.00056 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.331 on 44 degrees of freedom
Multiple R-squared:  0.5745,    Adjusted R-squared:  0.5552
F-statistic:  29.7 on 2 and 44 DF,  p-value: 6.849e-09
```

- The estimated coefficient for education level (-0.78833) is negative, indicating that the fertility rate decreases as education level increases. The estimated coefficient for the Catholic population proportion (0.11092) is positive, indicating that the fertility rate increases as the Catholic population proportion increases.
- Coefficients marked with three asterisks (***) are statistically significant (p-value $<$ 0.001). This shows that the effect of education level and Catholic population proportion on fertility is statistically significant.
- The F-statistic is 29.7 and the p-value is very small ($p < 0.001$), indicating that the model is generally significant.
- R-squared indicates that the independent variables explain approximately 57.45% of the variance in the fertility rate. Adjusted R-squared shows the adjusted R-squared for each added independent variable in the model.