Analysis of relationships in pulmonary data

This program assesses the relationships among variables in a study of pulmonary function in children. There is a data dictionary that provides more details about the data. The program was written by Steve Simon on 2024-09-07 and is placed in the public domain.

Libraries

The tidyverse library is the only one you need for this program.

```
        library(tidyverse)
```

List variable names

Since the variable names are not listed in the data file itself, you need to list them here.

```
        pulmonary_names <- c(
            "age",
            "fev",
            "ht",
            "sex",
            "smoke")
```

Reading the data

Here is the code to read the data and show a glimpse.

```
        pulmonary <- read_csv(
          file="../data/fev.csv",
          col_names=pulmonary_names,
          col_types="nnncc")
        glimpse(pulmonary)
```

```
Rows: 654
Columns: 5
$ age   <dbl> 9, 8, 7, 9, 9, 8, 6, 6, 8, 9, 6, 8, 8, 8, 8, 7, 5, 6, 9, 9, 5, 5…
$ fev   <dbl> 1.708, 1.724, 1.720, 1.558, 1.895, 2.336, 1.919, 1.415, 1.987, 1…
$ ht    <dbl> 57.0, 67.5, 54.5, 53.0, 57.0, 61.0, 58.0, 56.0, 58.5, 60.0, 53.0…
$ sex   <chr> "F", "F", "F", "M", "M", "F", "F", "F", "F", "F", "F", "M", "F",…
$ smoke <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N", "N",…
```

Calculate mean, quartiles, range for fev

```
summary(pulmonary$fev)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.791   1.981   2.547   2.637   3.119   5.793 
```

```
sd(pulmonary$fev)
```

```
[1] 0.8670591
```

The mean fev is 2.6 liters and the standard deviation is 0.84 liters. The fev values range from 0.8 to 5.8. I am not an expert on pulmonary function, but these values appear to be reasonable.

Calculate mean, quartiles, range for age

```
summary(pulmonary$age)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.000   8.000  10.000   9.931  12.000  19.000 
```

```
sd(pulmonary$age)
```

```
[1] 2.953935
```

The mean age is 9.9 years. The youngest subject is 3 years old and the oldest is 19. This is consistent with a pediatric population.

Calculate counts for smoke

```
pulmonary |>
  count(smoke) |>
  mutate(total=sum(n)) |>
  mutate(pct=round(100*n/total))
```

```
# A tibble: 2 × 4
  smoke     n total   pct
  <chr> <int> <int> <dbl>
1 N       589   654    90
2 Y        65   654    10
```

Almost all of the subjects (90% or 589 out of 654) were non-smokers.

Question 1

```
summary(pulmonary$ht)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  46.00   57.00   61.50   61.14   65.50   74.00 
```

```
sd(pulmonary$ht)
```

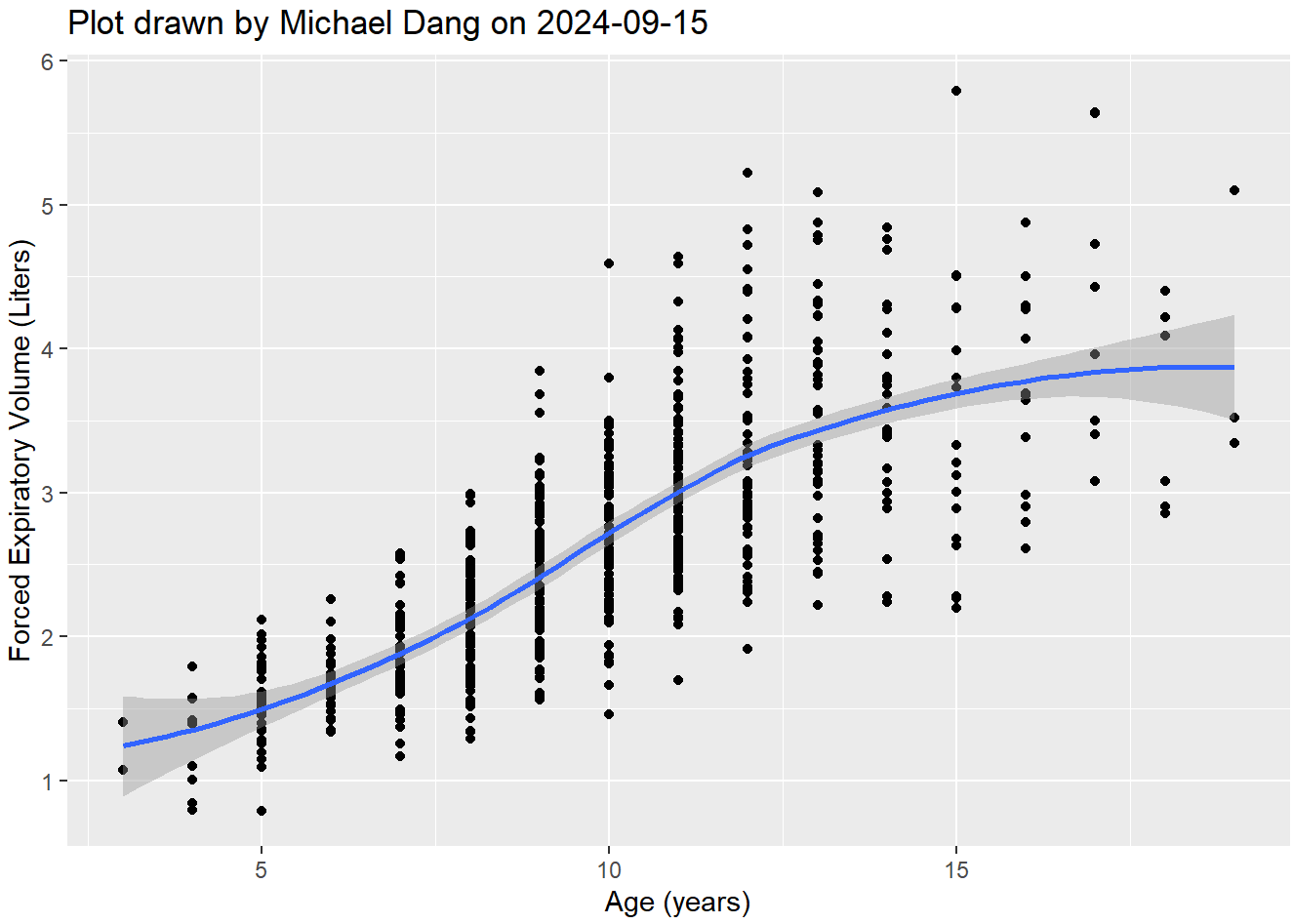
```
[1] 5.703513
```

The mean of the height is 61.15 inches and the standard deviation is 0.84 inches. The height values range from 47 to 74.

Plot age versus fev

```
pulmonary |>
  ggplot(aes(age, fev)) +
    geom_point() +
    geom_smooth() +
    xlab("Age (years)") +
    ylab("Forced Expiratory Volume (Liters)") +
    ggtitle("Plot drawn by Michael Dang on 2024-09-15")
```

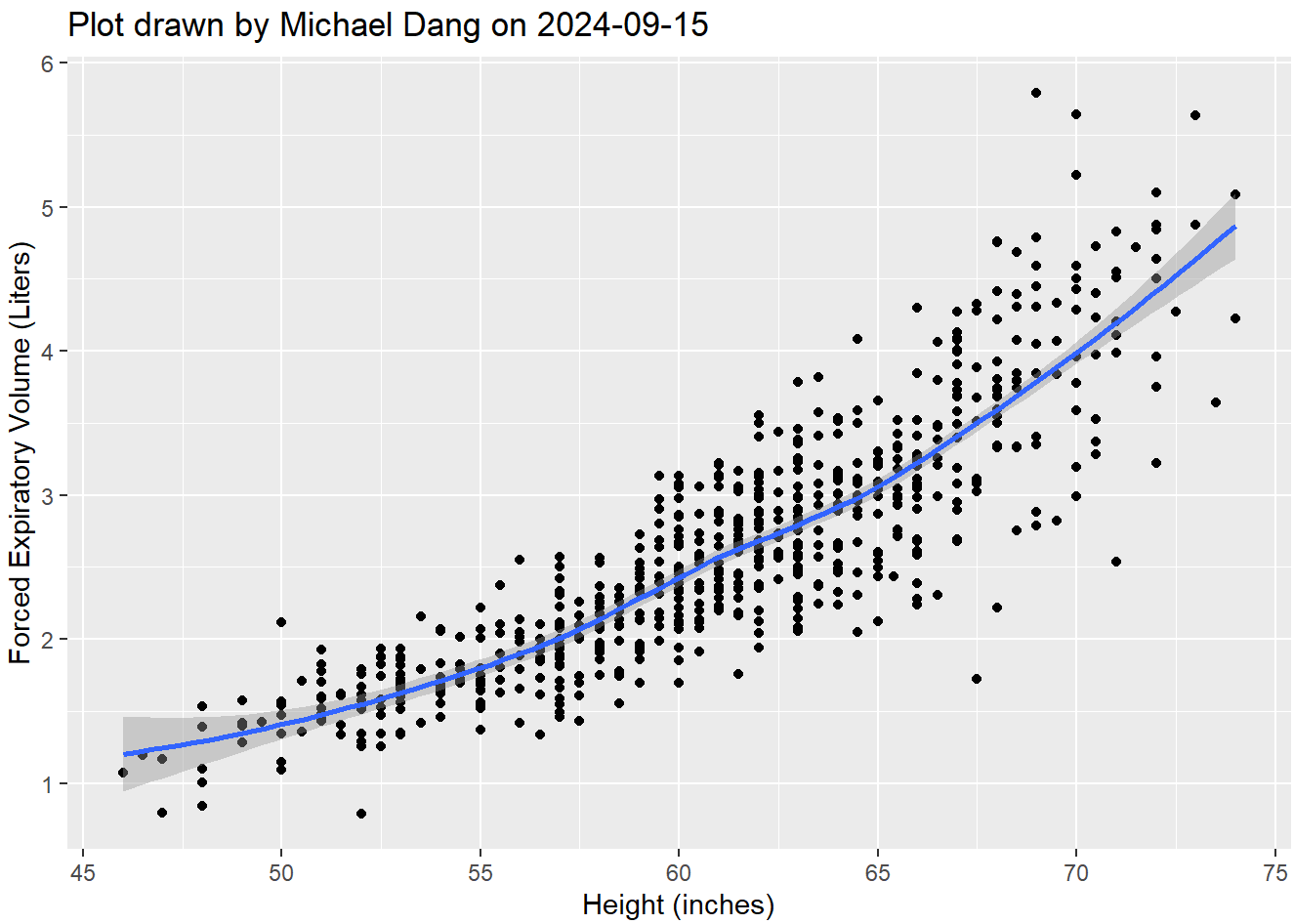
```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



Question 2

```
pulmonary |>
  ggplot(aes(ht, fev)) +
    geom_point() +
    geom_smooth() +
    xlab("Height (inches)") +
    ylab("Forced Expiratory Volume (Liters)") +
    ggtitle("Plot drawn by Michael Dang on 2024-09-15")
```

```
`geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



The plot demonstrates a clear positive trend, suggesting that taller individuals generally have greater Forced Expiratory Volume (fev), a measure commonly used in pulmonary function tests to assess lung health.

Correlation between age and fev

```
cor(pulmonary$age, pulmonary$fev)
```

```
[1] 0.756459
```

The correlation, 0.75, and the plot both show a strong positive association between age and fev.

Question 3

```
cor(pulmonary$ht, pulmonary$fev)
```

```
[1] 0.868135
```

The correlation is 0.87, and the plot both show a strong positive association between height and fev.

Question 4

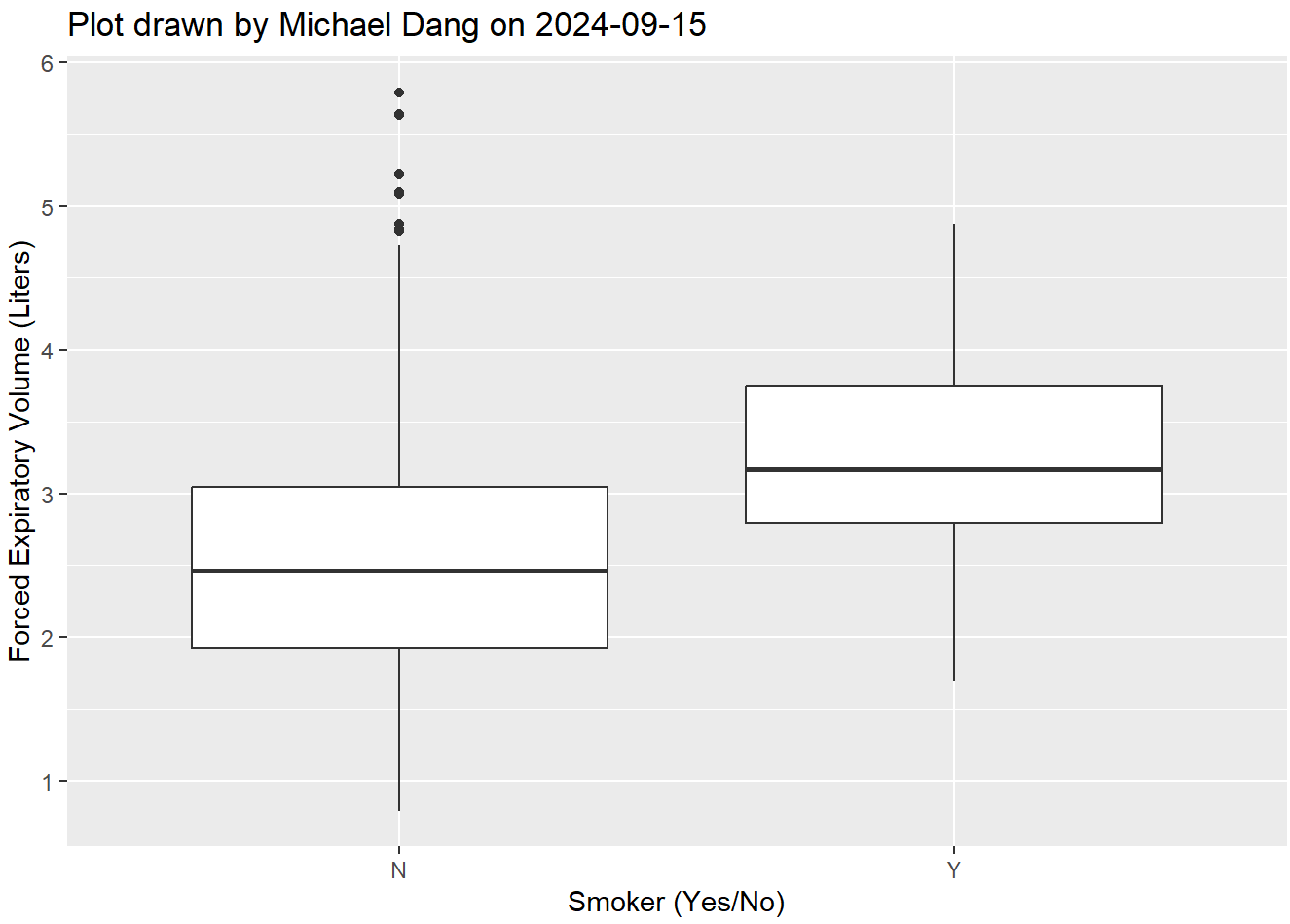
```
pulmonary |>
  count(sex) |>
  mutate(total=sum(n)) |>
  mutate(pct=round(100*n/total))
```

```
# A tibble: 2 × 4
  sex       n total   pct
  <chr> <int> <int> <dbl>
1 F       318   654    49
2 M       336   654    51
```

This indicates a relatively balanced representation of both sexes in the dataset, with a slight majority of males.

Plot smoke versus fev

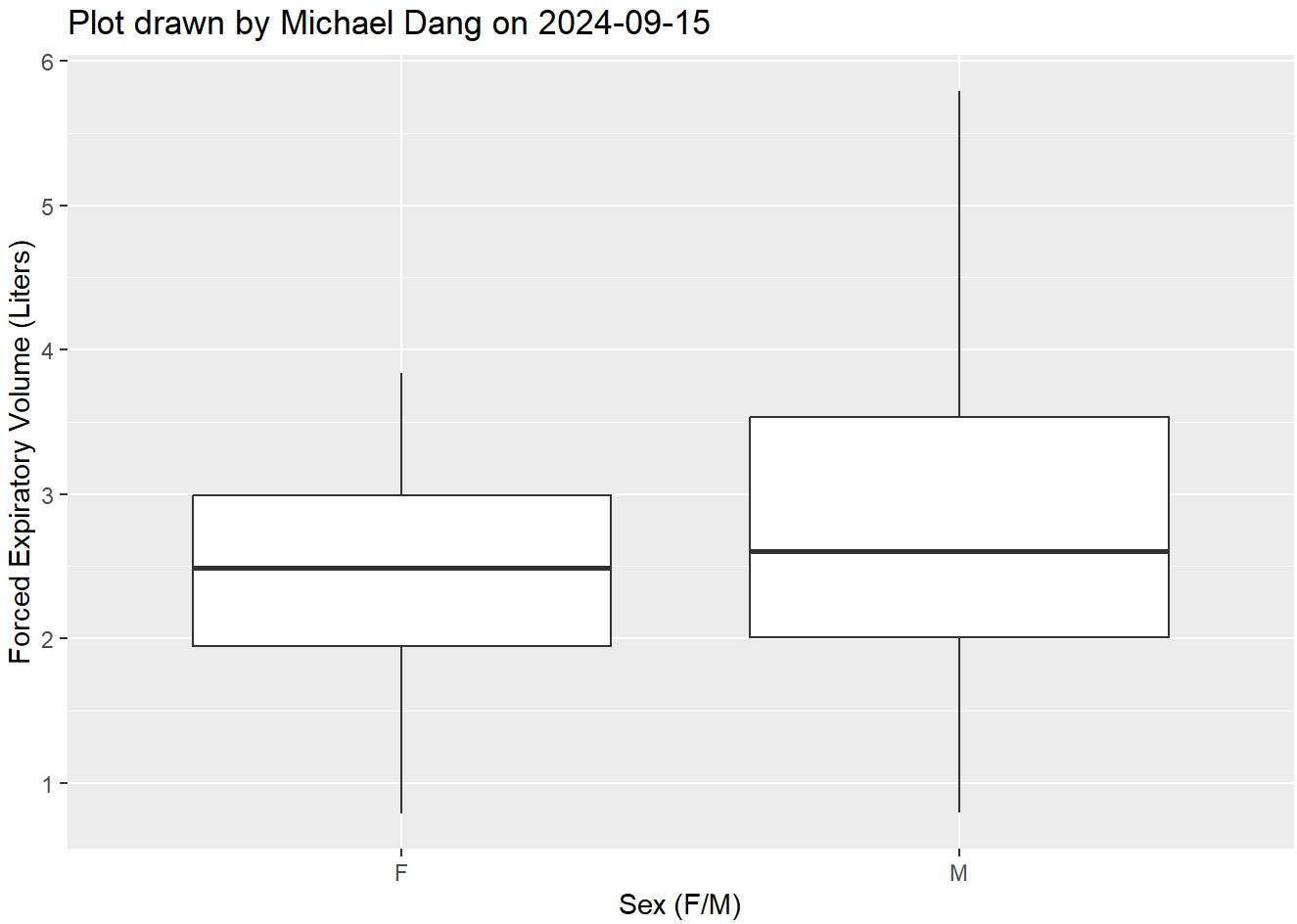
```
pulmonary |>
  ggplot(aes(smoke, fev)) +
    geom_boxplot() +
    xlab("Smoker (Yes/No)") +
    ylab("Forced Expiratory Volume (Liters)") +
    ggtitle("Plot drawn by Michael Dang on 2024-09-15")
```



The fev values are larger for smokers versus non-smokers. This is the opposite direction from what we expected.

Question 5

```
pulmonary |>
  ggplot(aes(sex, fev)) +
    geom_boxplot() +
    xlab("Sex (F/M)") +
    ylab("Forced Expiratory Volume (Liters)") +
    ggtitle("Plot drawn by Michael Dang on 2024-09-15")
```



The line within each box represents the median FEV. It appears slightly higher for males than for females.The boxes represent the middle 50% of the data (from the 25th to the 75th percentile). The IQR for males is broader, indicating greater variability in FEV among males compared to females. The whiskers extend to the minimum and maximum values within 1.5 IQR from the lower and upper quartiles. There are no outliers outside the whiskers, suggesting no extreme values in either group.

Means and standard deviations for smokers and non-smokers.

```
pulmonary |>
  group_by(smoke) |>
  summarize(
    mean_fev=mean(fev),
    sd_fev=sd(fev))
```

```
# A tibble: 2 × 3
  smoke mean_fev sd_fev
  <chr>    <dbl>  <dbl>
1 N         2.57  0.851
2 Y         3.28  0.750
```

The average fev values is 3.1 for smokers and much smaller, 2.6, for non-smokers. This is also opposite from what we expected. The standard deviations, 0.82 and 0.86, are roughly equal.

Question 6

```r
#Calculate mean FEV for males and females
mean_fev_male <- mean(pulmonary$fev[pulmonary$sex == "M"])
mean_fev_female <- mean(pulmonary$fev[pulmonary$sex == "F"])

#Calculate the difference in average FEV between males and females
difference <- mean_fev_male - mean_fev_female

#Print the result
cat("Difference in average FEV between males and females:", difference)
```

```
Difference in average FEV between males and females: 0.3612766
```

```r
#Calculate the standard deviation of FEV values for females
sd_female <- sd(pulmonary$fev[pulmonary$sex == "F"])

#Calculate the effect size by dividing the difference by the standard deviation of femal
effect_size <- difference / sd_female

#Print the result
cat("\nEffect size:", effect_size)
```

```
Effect size: 0.5594804
```

The difference in average fev values between males and females look small. The effect size is slightly above 0.5, indicate a medium effect size and show sex has a noticeable impact on fev.