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",…
```

Question 1: Update the program to calculate descriptive statistics (mean, standard deviation, minimum, and maximum) for ht. Interpret these statistics.

```
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)
```

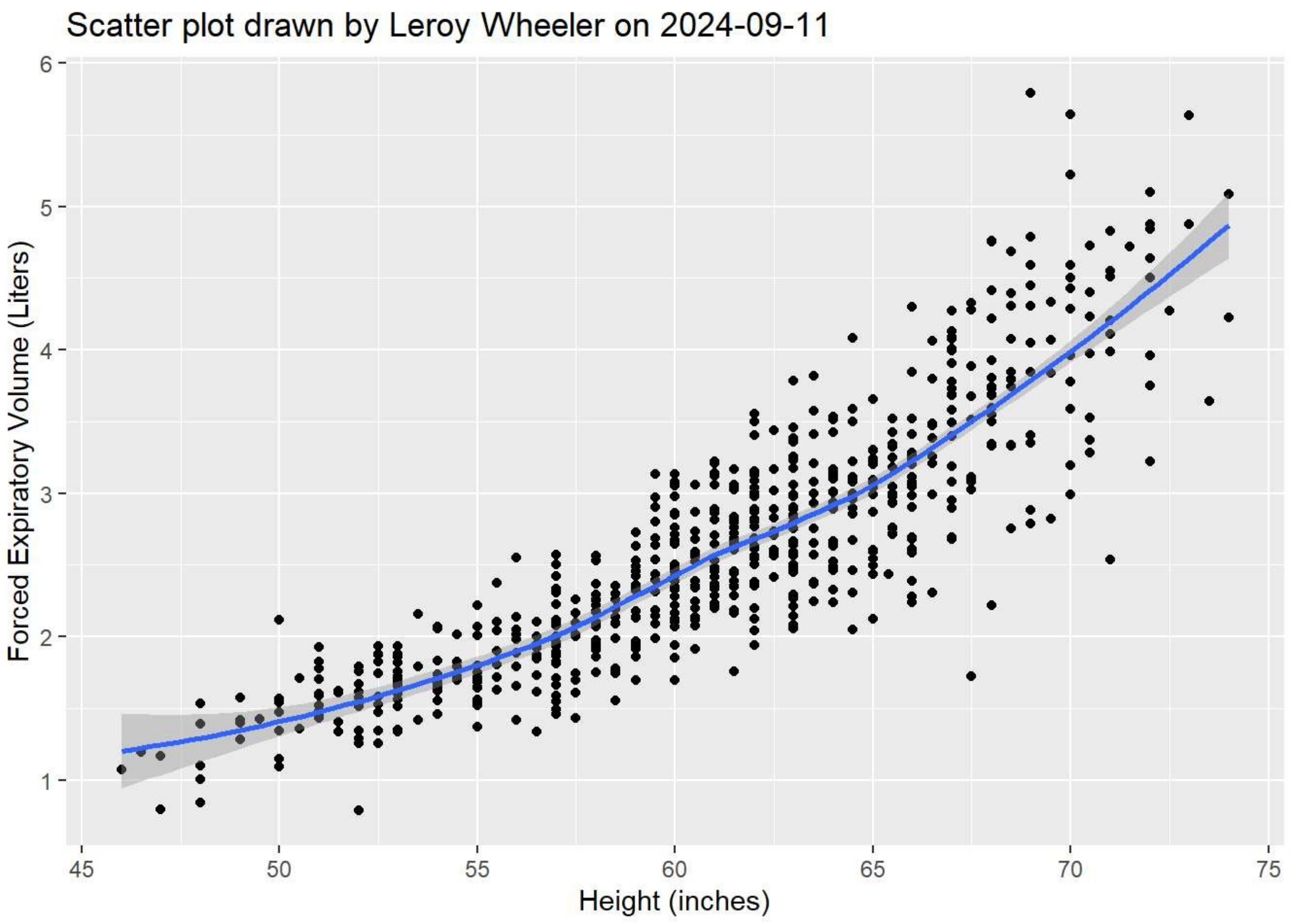
```
[1] 5.703513
```

The mean height is about 61 inches with a standard deviation of almost 6 inches. Height ranges from 46 to 74 inches, which is consistent with a pediatric population.

Question 2: Draw a scatterplot of ht versus fev. Place ht on the x-axis and fev on the y-axis. Interpret this plot.

```
pulmonary |>
  ggplot(aes(ht, fev)) +
              geom_point() +
              geom_smooth() +
              xlab("Height (inches)") +
              ylab("Forced Expiratory Volume (Liters)") +
              ggtitle("Scatter plot drawn by Leroy Wheeler on 2024-09-11")
```

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



There is a positive linear association between height and fev. Calculation of r will likely confirm this observation.

Question 3: Calculate the correlation between ht and fev. Interpret this correlation.

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

```
[1] 0.868135
```

A correlation value of r=0.87 confirms the strong positive relationship between height and fev in this data set.

Question 4: Calculate counts and percentages for sex. Please be sure to convert sex from the numeric codes into a factor. Interpret these statistics.

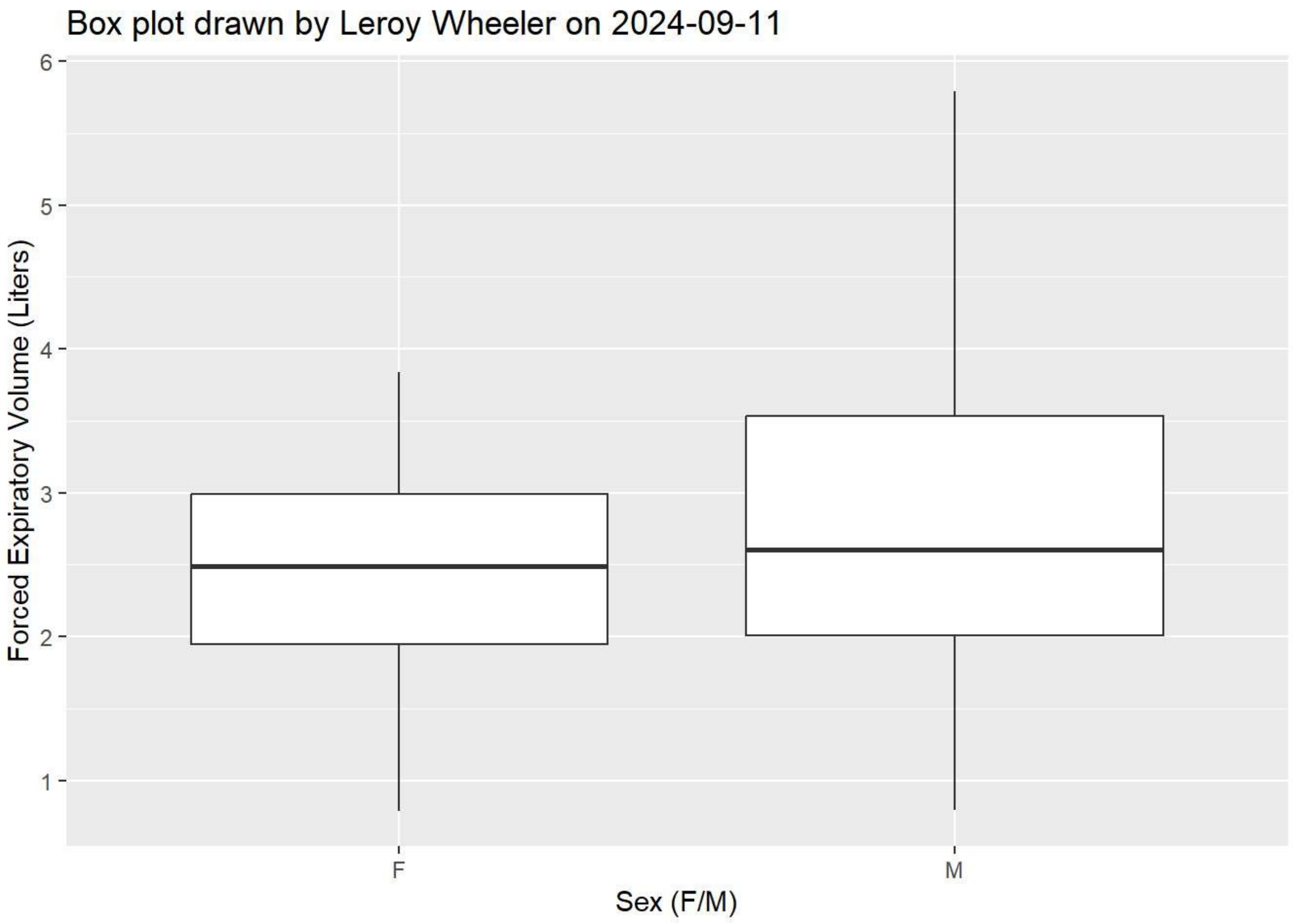
```
pulmonary |>
  count(sex) |>
  mutate(total=sum(n)) |>
  mutate(pct=round(100*n/total))
```

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

The data set was roughly split in half according to sex with 51% males and 49% females.

Question 5: Draw a boxplot for sex and fev. Interpret this boxplot

```
pulmonary |>
  ggplot(aes(sex, fev)) +
    geom_boxplot() +
    xlab("Sex (F/M)") +
        ylab("Forced Expiratory Volume (Liters)") +
        ggtitle("Box plot drawn by Leroy Wheeler on 2024-09-11")
```



The fev values are a little larger for males when compared to females. The variability for the male data is also slightly higher as well. These results are not surprising.

Question 6: Calculate the difference in average fev values between males and females. Is this a large or a small difference? Calculate the

effect size by dividing by the standard deviation of the females. Is this a small, medium, or large effect size?

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

```
# A tibble: 2 × 3
  sex   mean_fev sd_fev
  <chr>    <dbl>  <dbl>
1 F         2.45  0.646
2 M         2.81  1.00 
```

The average fev values for males is 2.8 which is larger than that observed in females, which is 2.5. Males also have a standard deviation of 1, which is also higher than the standard deviation of 0.6 seen in females.

The effect size between males and females is approximately 0.6 standard deviations.

Analysis of gardasil shots by demographic factors

This program reads data on Gardasil vaccinations in young women. Find more information in the data dictionary.

The program was written by Steve Simon on 2024-09-07 and is placed in the public domain.

Load the tidyverse library

For most of your programs, you should load the tidyverse library. The messages and warnings are suppressed.

```r
library(tidyverse)
```

Read the data and view a brief summary

Use the read_csv function to read the data. The glimpse function will produce a brief summary. Use tolower to convert uppercase to lowercase.

```r
gard <- read_csv(
  file="../data/gardasil.csv",
  col_names=TRUE,
  col_types="nnnnnnnnnn")
names(gard) <- tolower(names(gard))
glimpse(gard)
```

```
Rows: 1,413
Columns: 10
$ age           <dbl> 21, 21, 20, 14, 17, 11, 17, 15, 13, 18, 17, 22, 16, 13, …
$ agegroup      <dbl> 1, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0,…
$ race          <dbl> 0, 0, 0, 0, 3, 1, 0, 3, 3, 0, 1, 0, 3, 1, 1, 0, 1, 1, 0,…
$ shots         <dbl> 3, 3, 1, 3, 2, 1, 1, 3, 3, 3, 2, 2, 1, 2, 1, 1, 1, 3, 3,…
$ completed     <dbl> 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1,…
$ insurancetype <dbl> 3, 3, 1, 3, 3, 0, 3, 1, 1, 2, 1, 3, 1, 3, 0, 1, 1, 1, 1,…
$ medassist     <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,…
$ location      <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,…
$ locationtype  <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
$ practicetype  <dbl> 1, 1, 1, 0, 1, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 1, 1, 1, 1,…
```

Question 7: First create factors for medassist

The factor function identifies a variable as categorical and assigns labels to number codes. You don't necessarily need to use factor if the data you read in is character strings, as R automatically treats those variable as categorical.

```
gard$medassist <- factor(
  gard$medassist,
  levels=0:1,
  labels=c(
    "No medical assistance",
    "Received medical assistance"))
```

Question 7: Summarize and interpret the percentage of patients receiving medical assistance. Be sure to convert the number codes for this variable into labels using the factor function

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

```
# A tibble: 2 × 4
  medassist                       n total   pct
  <fct>                       <int> <int> <dbl>
1 No medical assistance        1138  1413    81
2 Received medical assistance   275  1413    19
```

Eighty one percent of patients received at least some medical assistance while the remaining 19% did not.

Create factors for shots

It is a bit silly to replace 1, 2, 3 with One, Two, Three. The main reason is to clearly identify shots as categorical rather than continuous.

```
gard$shots <- factor(
  gard$shots,
  levels=1:3,
  labels=c(
    "One",
    "Two",
    "Three"))
```

Counts and percentages for shots

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

```
# A tibble: 3 × 4
  shots     n total   pct
  <fct> <int> <int> <dbl>
1 One     440  1413    31
2 Two     436  1413    31
3 Three   537  1413    38
```

Slightly more patients got three shots than one or two shots, but this is still less than half of the patients overall.

Question 8: First calculate the percentages for number of shots received by whether the patient received medical assistance. Interpret this chart.

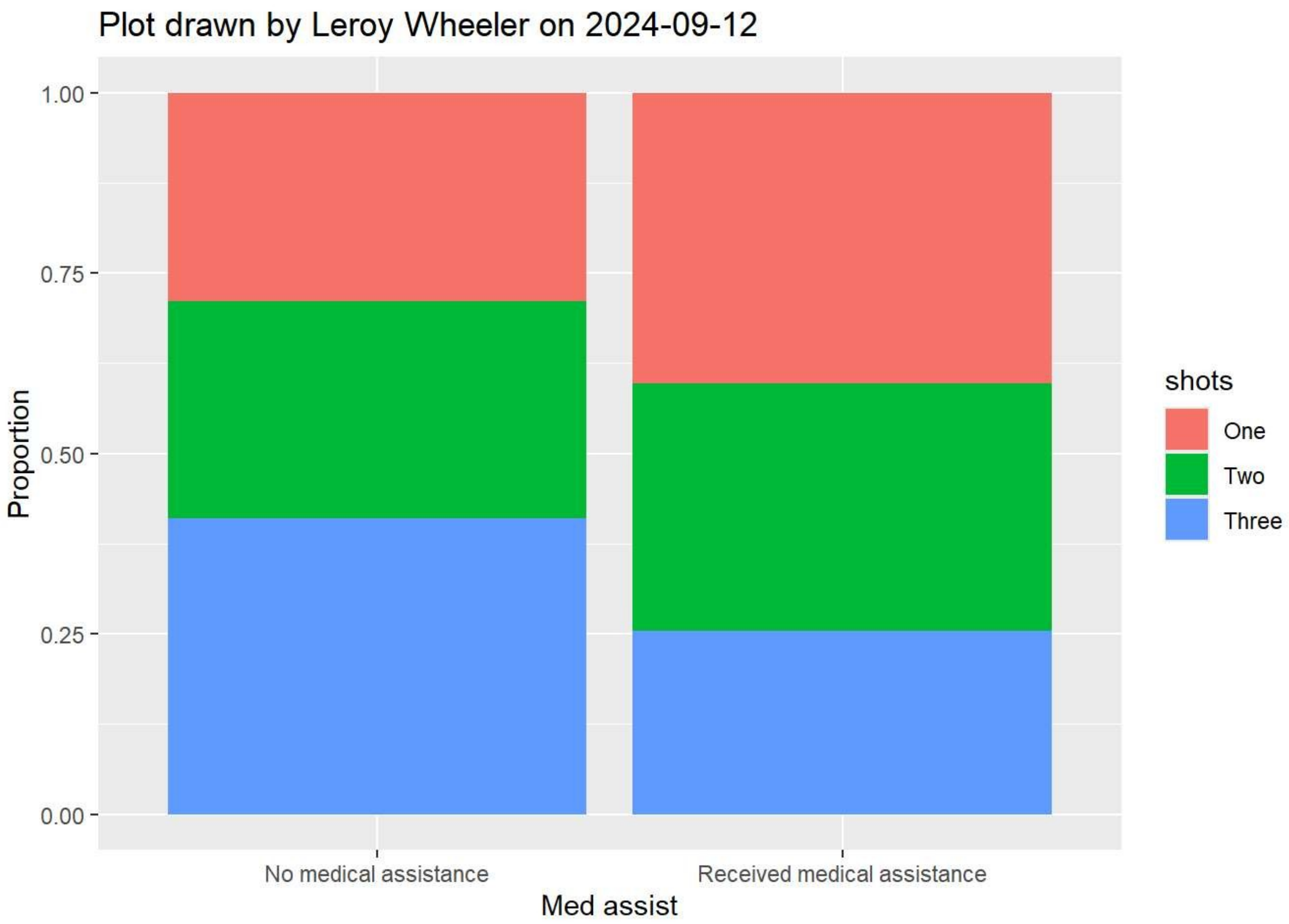
```
gard |>
  count(medassist, shots) |>
  group_by(medassist) |>
  mutate(row_total=sum(n)) |>
  mutate(pct=round(100*n/row_total))
```

```
# A tibble: 6 × 5
# Groups:   medassist [2]
  medassist                    shots     n row_total   pct
  <fct>                        <fct> <int>     <int> <dbl>
1 No medical assistance        One     329      1138    29
2 No medical assistance        Two     342      1138    30
3 No medical assistance        Three   467      1138    41
4 Received medical assistance One     111       275    40
5 Received medical assistance Two      94       275    34
6 Received medical assistance Three    70       275    25
```

Surprisingly 41% of patients who did not receive medical assistance received all three shots when compared to the 25% of patients who received medical assistance.

Question 8: Draw a bar chart showing the percentages for number of shots received by whether the patient received medical assistance. Interpret this chart.

```
gard |>
  ggplot(aes(x=medassist, fill=shots)) +
    geom_bar(position="fill") +
    xlab("Med assist") +
    ylab("Proportion") +
    ggtitle("Plot drawn by Leroy Wheeler on 2024-09-12")
```



Patients who did not receive medical assistance were more likely to complete the full round of three Gardisil shots compared to patients who received some medical assistance.