Analysis of fruitfly data

This program reads data on fruit fly longevity. Find more information in the data dictionary.

This code was written by Steve Simon and Leroy Wheeler on 2024-11-07 and is placed in the public domain.

Load the tidyverse library

```
library(broom)
library(tidyverse)
```

Comments on the code

For most of your programs, you should load the tidyverse library. The broom library converts your output to a nicely arranged dataframe. The messages and warnings are suppressed.

List the variable names

```
vlist <- c(
  "id",
  "partners",
  "type",
  "longevity",
  "thorax",
  "sleep")
```

Comments on the code

When a dataset does not have variables on the first line, you need to specify them in the code.

Read the data and view a brief summary

```
fly <- read_fwf(
  "../data/fruitfly.txt",
  col_types="nnnnnn",
  fwf_widths(
    widths=c(2, 2, 2, 3, 5, 3),
    col_names=vlist))
glimpse(fly)
```

```
Rows: 125
Columns: 6
$ id        <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 1…
$ partners  <dbl> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, …
$ type      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …
$ longevity <dbl> 35, 37, 49, 46, 63, 39, 46, 56, 63, 65, 56, 65, 70, 63, 65, …
$ thorax    <dbl> 0.64, 0.68, 0.68, 0.72, 0.72, 0.76, 0.76, 0.76, 0.76, 0.76, …
$ sleep     <dbl> 22, 9, 49, 1, 23, 83, 23, 15, 9, 81, 12, 15, 37, 24, 26, 17,…
```

Comments on the code

The fruitfly dataset has a fixed width format (fwf). You need to specify the columns that each variable uses.

Notice that the two categorical variables, partners and type, are actually numbers rather than strings. To avoid having R treat these variables as if they were continuous, use the factor function in some of the code below.

Create a subset of fruitfly dataset

```
fly |>
  filter(type != 9) -> fly_subset
```

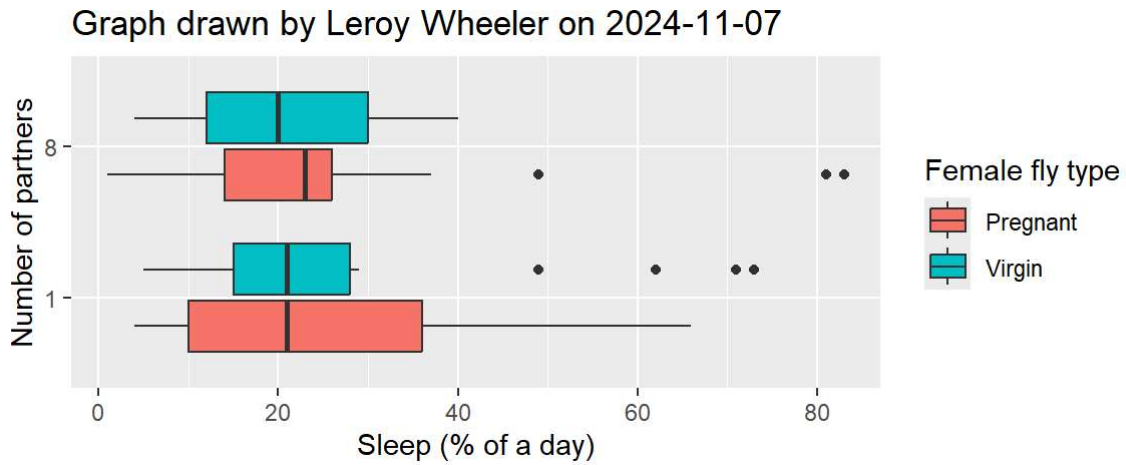
Comments on the code

If you exclude the pure control group (No females), you can analyze the two factors, partners and type individually. Partners has two category levels, 1 for when one female was included in the cage and 8 for when eight females were included in the cage. Type also has two category levels, 0 for pregnant female fly/flies and 1 for virgin fly/flies. A male

fly will not mate with a pregnant females, so you can think of this as a second level of controls. The two factors are crossed, meaning that every possible combination of partners and type has outcomes measured.

Question 1: Create a subset of the fruitfly data by removing the age where type equals 9. Draw a clustered boxplot with sleep as the outcome and partners and type as the categorical predictors. Interpret this graph. Is there evidence of non-normality?

```
fly_subset |>
  ggplot(aes(factor(partners), sleep, fill=factor(type))) +
    geom_boxplot() +
    xlab("Number of partners") +
    ylab("Sleep (% of a day)") +
    ggtitle("Graph drawn by Leroy Wheeler on 2024-11-07") +
    labs(fill="Female fly type") +
    scale_fill_discrete(labels=c("Pregnant", "Virgin")) +
    coord_flip()
```



Interpretation of the output for question 1.

Sleep time appear similar in all groups and it also appears that assumptions of homogeneity and normality hold up ok.

Question 2: Calculate descriptive statistics for sleep (mean, standard deviation, and sample size) by the combination of the two categorical predictors, partners and type. Is there evidence of heterogeneity?

```
fly_subset |>
  group_by(type, partners) |>
  summarize(
    sleep_mn=mean(sleep),
    sleep_sd=sd(sleep),
    n=n()) -> fly_means
fly_means
```

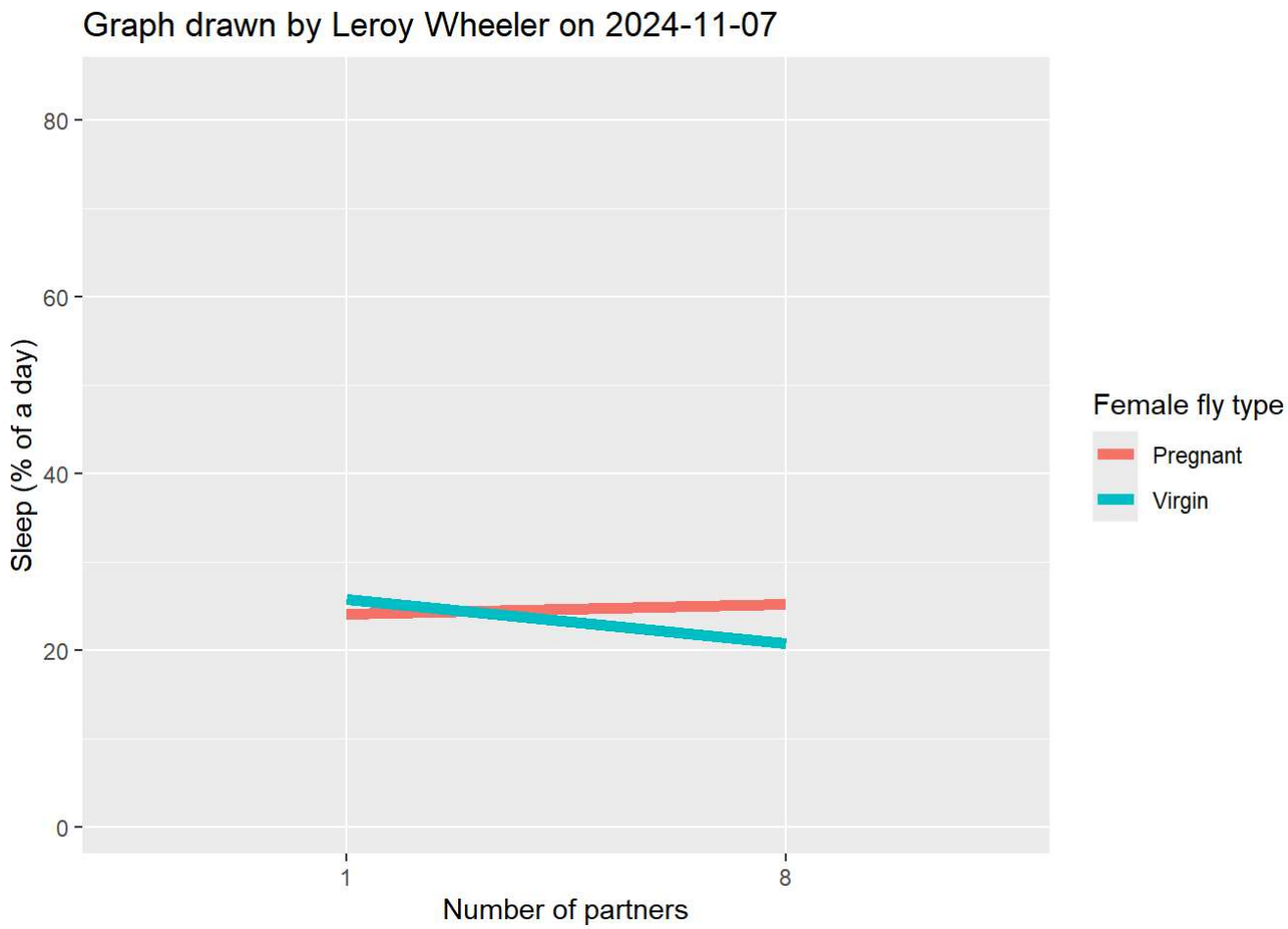
```
# A tibble: 4 × 5
# Groups:   type [2]
   type partners sleep_mn sleep_sd     n
  <dbl>    <dbl>    <dbl>    <dbl> <int>
1     0        1     24.1     16.7    25
2     0        8     25.2     19.8    25
3     1        1     25.8     18.4    25
4     1        8     20.8     10.7    25
```

Interpretation of the output for question 2

The mean values of sleep time are all similar and the standard deviations are also similar so no evidence of heterogeneity is obvious.

Question 3: Draw a line graph for the mean sleep levels compared by type and partners. Is there evidence of an interaction?

```r
fly_means |>
  ggplot(aes(
    factor(partners),
    sleep_mn,
    group=factor(type),
    color=factor(type))) +
    geom_line(linewidth=2) +
    expand_limits(y=range(fly_subset$sleep)) +
  xlab("Number of partners") +
  ylab("Sleep (% of a day)") +
  ggtitle("Graph drawn by Leroy Wheeler on 2024-11-07") +
labs(color="Female fly type") +
  scale_color_discrete(labels=c("Pregnant", "Virgin"))
```



Interpretation of the output for question 3

Even though the two lines cross, we see from the Anova table from question 4 that there is no significant interaction between the number of partners and the partner type.

Question 4: Analyze the sleep variable using a two factor analysis of variance with an interaction. Present and interpret the analysis of

variance table.

```
m1 <- aov(sleep ~ factor(partners)*factor(type), data=fly_subset)
anova(m1)
```

```
Analysis of Variance Table

Response: sleep
                            Df  Sum Sq Mean Sq F value Pr(>F)
factor(partners)             1    96.0   96.04  0.3408 0.5607
factor(type)                 1    46.2   46.24  0.1641 0.6863
factor(partners):factor(type)  1   231.0  231.04  0.8198 0.3675
Residuals                   96 27054.3  281.82
```

Interpretation of the output for question 4

There is a no statistically significant interaction between the number of female partners and the type of partners (pregnant or virgin). Because there is no interaction, we can also interpret the individual categorical predictors. We see that each categorical shows no evidence of significant influence on sleep time.

Question 5: What factors might make you consider using a log transformation for the sleep variable? Do not run such an analysis but tell us whether you think the data would warrant such a transformation?

For the virgin female category there are some outliers to the right in the one partner cage so I might want to run a log transformation on that category.

For the pregnant female category, there are also some outliers in the eight partner cage, however, there might be some left skewing of that subcategory so I would have to run the log transformation and see if it improves our outcomes.