R workshop session 1 exercise

Exercise 1: Road crossing example

Let's assume you have 5% chance of a fatal accident when crossing a road. What are your chances of surviving ten such crossings?

SOLUTION:

```
Let p be the probability of death crossing a road.
So the probability of surviving a road is 1 - p.

Prob. of surviving 10 road crossing
 = Prob. of surviving 1st road x
    Prob. of surviving 2nd road x
     ... x
      Prob. of surviving 10th road
 = (1-p)^10
```

Thus the probability of surviving all ten crossings is:

```r
p <- 0.05
(1-p)^10
```

[1] 0.5987369

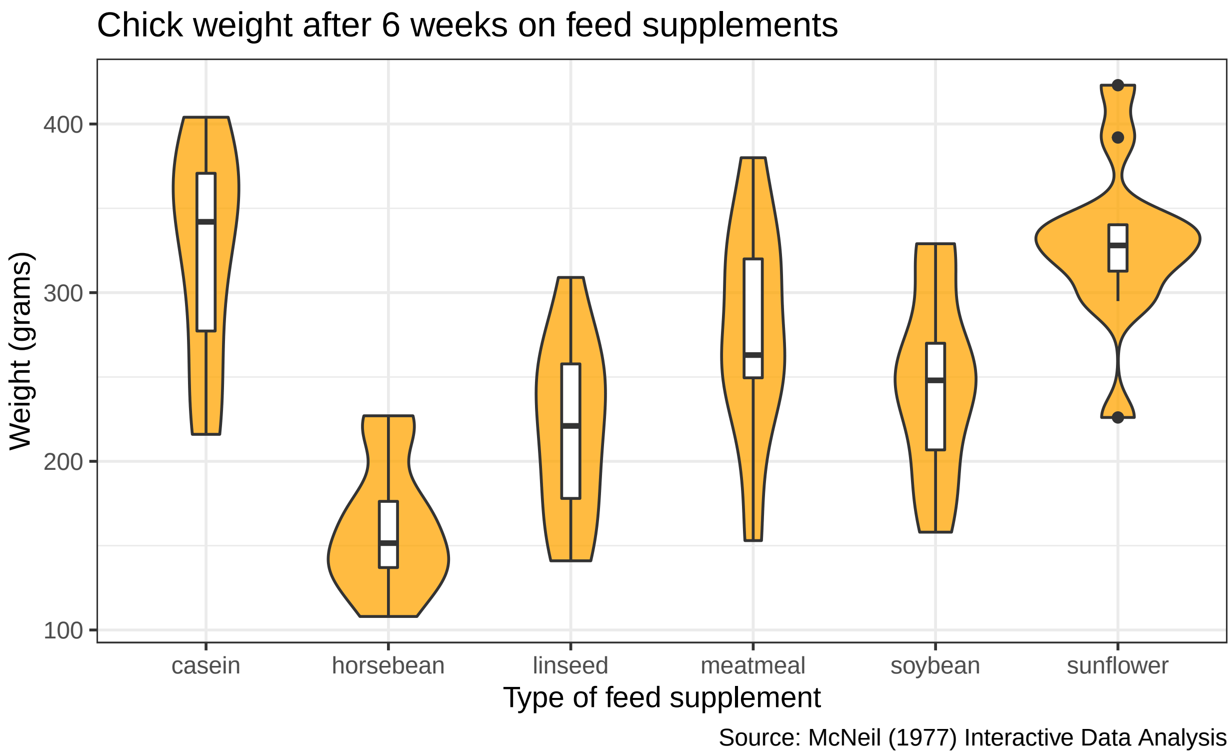
Exercise 2: Chick weights

Background: Newly hatched chicks were randomly allocated into six groups, and each group was given a different feed supplement. Their weights in grams after six weeks are given along with feed types. *Source:* McNeil (1977) Interactive Data Analysis

Exercise 2: Read in the data from the chick_weight sheet in the session1_data.xlsx file. Reproduce the following graph as closely as possible. Name your script chick_weights.R

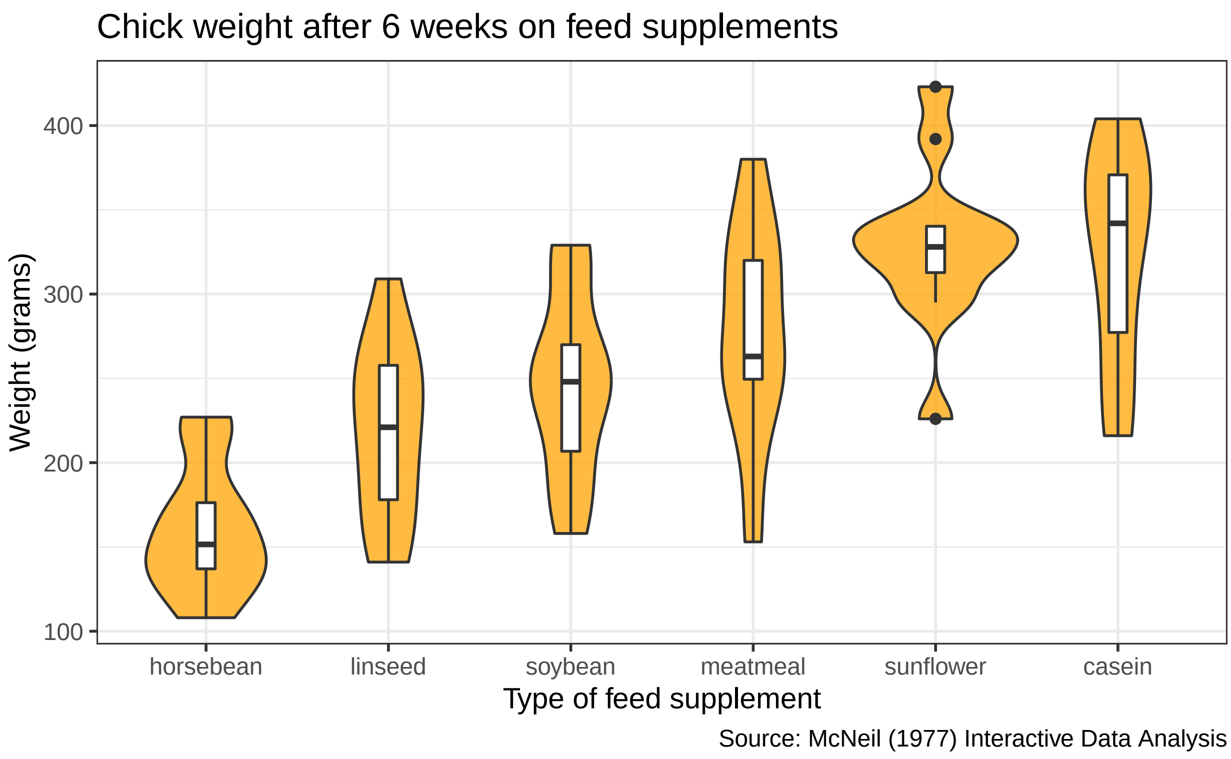
SOLUTION:

```r
setwd("C:/Users/aramasamy/Desktop/R_workshop")
pacman::p_load(tidyverse, readxl, gridExtra, janitor)

cw <- read_excel("data/session1_data.xlsx", sheet="chick_weight")

ggplot(cw, aes(x=feed, y=weight)) +
  geom_violin(fill="orange", alpha=0.75) +
  geom_boxplot(width=0.1) +
  labs(title="Chick weight after 6 weeks on feed supplements",
       caption="Source: McNeil (1977) Interactive Data Analysis",
       x="Type of feed supplement", y="Weight (grams)") +
  theme_bw()
```



Actually, we can generate a better plot by ordering the boxplots by median. We can do this by using reorder() on the aesthetics for x-axis as follows.

```
ggplot(cw, aes(x=reorder(feed, weight, FUN=median), y=weight)) +
  geom_violin(fill="orange", alpha=0.75) +
  geom_boxplot(width=0.1) +
  labs(title="Chick weight after 6 weeks on feed supplements",
       caption="Source: McNeil (1977) Interactive Data Analysis",
       x="Type of feed supplement", y="Weight (grams)") +
  theme_bw()
```



Exercise 3: Height vs weight by Gender (optional)

Background: 10000 measurements of height and weight for men and women. *Source:* Machine Learning for Hackers, Drew Conway & John Myles-While, O'Reilly Media.

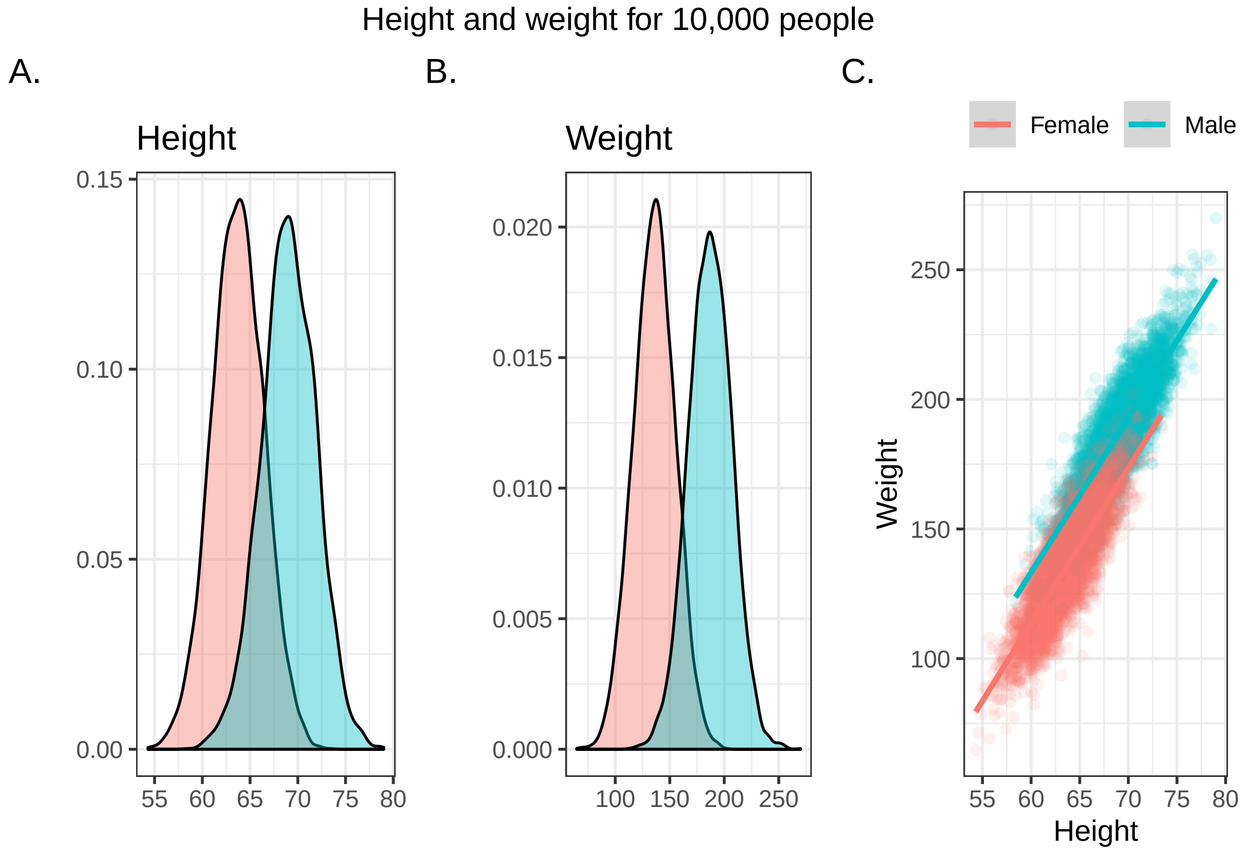
Exercise: Read in the data from the height_weight sheet in the session1_data.xlsx file. Name your script height_weights.R.

1. Reproduce the following graph as closely as possible. Hint: `grid.arrange()`
2. Does anything look strange?
3. How many men and women are there in this dataset?

SOLUTION:

```
hw <- read_excel("data/session1_data.xlsx", sheet="height_weight")

g.h <- ggplot(hw, aes(x=Height, fill=Gender)) +
  geom_density(alpha=0.4) +
  labs(x="", y="", title="\nHeight", tag="A.") +
  theme_bw() + theme(legend.position="none")

g.w <- ggplot(hw, aes(x=Weight, fill=Gender)) +
  geom_density(alpha=0.4) +
  labs(x="", y="", title="\nWeight", tag="B.") +
  theme_bw() + theme(legend.position="none")
```

```
g.hw <- ggplot(hw, aes(x=Height, y=Weight, col=Gender)) +
  geom_point(alpha=0.1) +
  geom_smooth(method="lm") +
  labs(x="Height", y="Weight", tag="C.", col=NULL) +
  theme_bw() + theme(legend.position="top")

grid.arrange(g.h, g.w, g.hw, nrow=1,
             top="Height and weight for 10,000 people")
```



The values given for height and weight look odd. Upon further check, you realize the height is recorded in inches and weight is recorded in pounds. You will learn how to convert it to metric units in session 2 and how to estimate the slopes and intercepts from a linear regression in session 3.

```
hw %>% tabyl(Gender)
##  Gender    n percent
##  Female 5000     0.5
##    Male 5000     0.5
rm(hw, g.h, g.w, g.hw)
```