Exploring Cardio Fitness

March 2, 2022

The dataset used for this challenge is titled 'fitness.csv'.

What topic does the dataset cover?

The market research team at AdRight is assigned the task to identify the profile of the typical customer for each treadmill product offered by CardioGood Fitness. The market research team decides to investigate whether there are differences across the product lines with respect to customer characteristics. The team decides to collect data on individuals who purchased a treadmill at a CardioGoodFitness retail store during the prior three months.

Variables

- Product: Which treadmill is used.
- Age: How old person is.
- Gender: Gender of person.
- Education: Years of education
- MaritalStatus: Single/married
- Usage: Average number of times a person will use treadmill weekly (expected).
- Fitness: Physical fitness from 1-5 (self-rated, poor-excellent)
- Income: Income of individual.
- Miles: Average number of miles a person will run (expected).

Assignment

Note: Throughout this assignment the variable 'fitness' is sometimes treated as a numeric variable and sometimes treated as a categorical variable. You will need to convert the variable from numeric to categorical using the factor() function in several questions.

1. Visualize and describe the distribution for fitness.
2. Are the variables 'usage' and 'fitness' independent of each other? Use a relative barchart to support your answer.
3. Is the income level of people different across different levels of fitness? Use a pairwise boxplot to support your answer.
4. Is fitness level independent of martial status? Use a relative barchart to support your answer.
5. Do more fit people tend to run more? Use pairwise boxplots to support your answer.
6. Is the fitness of a person associated with the product they use? Use a relative barchart to support your answer.
7. Do people who use certain products tend to have higher incomes? Use pairwise boxplots to support your answer.

Solution:

```
  ### Load packages and dataset
library(ggplot2)
library(tidyverse)
cardio <- read.csv('/data/datasets/cardio.csv')

# Take a look at the data
glimpse(cardio)

### Question 1: Distribution for fitness
cardio %>% ggplot(aes(x=Fitness)) +
  geom_histogram() +
  theme_classic()

### Question 2: Usage relationship with fitness
cardio %>% ggplot(aes(x=factor(Usage), fill=factor(Fitness))) +
  geom_bar(position='fill') + theme_classic()

### Question 3: Income relationship with fitness
cardio %>% ggplot(aes(fill=factor(Fitness), y=Income))+
  geom_boxplot(alpha=0.5) +
  theme_classic()

### Question 4: Marital Status and Fitness relative barchart
cardio %>% ggplot(aes(x=Fitness, fill=MaritalStatus)) +
  geom_bar(position='fill') +
  theme_classic()

### Question 5: Do more fit people tend to run more?
cardio %>% ggplot(aes(fill=factor(Fitness), y=Miles))+
  geom_boxplot() + theme_classic()

### Question 6: Is the fitness of a person associated with product they use?
cardio %>% ggplot(aes(x=Fitness, fill=Product)) +
  geom_bar(position='fill') +
  theme_classic()

### Question: Do people who use certain products tend to have higher incomes?
cardio %>% ggplot(aes(y=Income, fill=Product)) +
  geom_boxplot()+
  theme_classic()
```