XP Booster
April 27

For this challenge you will be working in GROUPS OF TWO and using the heart disease dataset (heart.csv). Be sure to clearly write all the code you use.

1. How many observations and variables are in the dataset?
2. What proportion of people have heart disease in the dataset?
3. Out of all males, what proportion have heart disease?
4. Out of all females, what proportion have heart disease?
5. If you randomly select a male and a female, which person has a higher probability of having heart disease?
6. Look at association between heart disease and gender using a relative boxplot.

Solution:

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

### Question 1: How many obaervations in the dataset, how many variables?
glimpse(heart)

### Question 2: What is the propoortion of people with heart disease
sum(heart$HeartDisease=="Yes")/length(heart$HeartDisease) # 0.08559546

### Question 3: Propoortion of males with heart disease
sum(heart$HeartDisease=="Yes" & heart$Gender=='Male')/length(heart$
    HeartDisease) # 0.05046671

### Question 4: Propoortion of females with heart disease
sum(heart$HeartDisease=="Yes" & heart$Gender=='Female')/length(heart$
    HeartDisease) # 0.03512875

### Question 5: The male proportion appears to be higher.

### Question 6: Heart disease Across Gender withrelative bar chart
heart %>% ggplot(aes(x=Gender, fill=HeartDisease)) +
  geom_bar(position='fill') +
  theme_classic()
```