Lab 4 : Normal probability

1. Use the CDC dataset you explored in the previous lab.
2. Create a subset dataset with only 3 variables: weight, height and gender.
3. Create two additional datasets fdims and hdims: one with only men and another with only women where weight and height in the created datasets are in Kg and Cm respectively.
4. Use the created datasets to make a histogram of men's heights and a histogram of women's heights. How would you compare the various aspects of the two distributions?
5. Compute the mean and standard deviation of female heights:

   ```
   fhgtmean <- mean(fdims$hgt)
   fhgtsd   <- sd(fdims$hgt)
   ```

6. Plot a normal distribution curve on top of the histograms to see how closely the data follow a normal distribution: Make a density histogram to use as the backdrop and use the lines function to overlay a normal probability curve. We use dnorm to calculate the density of each of x-values in a distribution that is normal with mean fhgtmean and standard deviation fhgtsd. (To adjust the y-axis you can add a third argument to the histogram function: ylim = c(0, 0.06)):

   ```
   hist(fdims$hgt, probability = TRUE)
   x <- 140:190
   y <- dnorm(x = x, mean = fhgtmean, sd = fhgtsd)
   lines(x = x, y = y, col = "blue")
   ```

7. Based on the plot, does it appear that the data follow a nearly normal distribution? To verify this, construct a normal probability plot, also called a normal Q-Q plot (for "quantile-quantile"). A data set that is nearly normal will result in a probability plot where the points closely follow the line. Any deviations from normality leads to deviations of these points from the line:

   ```
   qqnorm(fdims$hgt)
   qqline(fdims$hgt)
   ```

8. What do probability plots look like for data that I *know* came from a normal distribution? Simulate data from a normal distribution using rnorm.

   ```
   sim_norm <- rnorm(n = length(fdims$hgt), mean = fhgtmean, sd = fhgtsd)
   ```

9. Compare the shapes of the simulated data set, sim_norm, **as well as its normal probability plot** with those obtained earlier.
10. Assuming female heights follow a normal distribution, what is the "theoretical" probability that a randomly chosen young adult female is taller than 182 cm (use pnorm)?
11. Now calculate the probability empirically, by determining how many observations fall above 182 and then dividing this number by the total sample size:
    sum(fdims$hgt > 182) / **length**(fdims$hgt)

    Compare this result with that in 10.