Class04

Fan Wu(PID:A15127541)

Section 1: BRFSS as an example data.frame

> The Behavioral Risk Factor Surveillance System (BRFSS) is an annual telephone survey of 350,000 people in the United States. The survey is designed to identify risk factors in the adult population and report emerging health trends.

This dataset is a sample of 20,000 people from the survey done in 2000

```
source("http://thegrantlab.org/misc/cdc.R")
```

To view the dataset itself, either by Environment panel, or from the command below:

```
View(cdc)
```

> Key-point: The $ operator in R is used to access variables (i.e. columns) within a data.frame; for example, cdc$height tells R to look in the cdc data.frame for the height variable:

```
head(cdc$height)
```

```
[1] 70 64 60 66 61 64
```

> Q1. How would you “argue” with the tail() function to print out the last 20 weight values? Provide your code below:

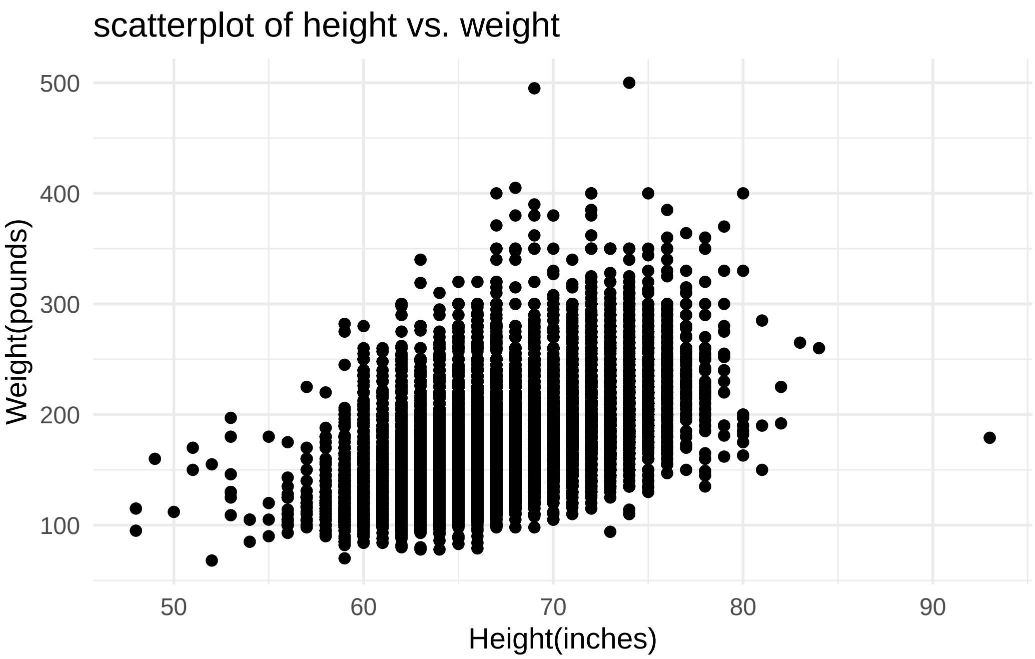
```
tail(cdc$height, 20)
```

```
 [1] 75 74 63 71 69 64 63 72 74 73 71 72 71 63 69 66 73 65 67 69
```

> Q2. Make a scatterplot of height vs weight using the plot() function. Add the code you used to generate this plot here:

```
library(ggplot2)

ggplot(cdc) +
  aes(x = height,
      y = weight) +
  geom_point() +
  labs(title = "scatterplot of height vs. weight",
       x = "Height(inches)",
       y = "Weight(pounds)") +
  theme_minimal()
```



Q3.Do height and weight appear to be associated? If so are they positively associated or negatively associated?

Yes, positively associated in trend. Higher height, higher weight in trend.

Q4. What is the Pearson correlation value for height and weight?

```
cor(cdc$height,cdc$weight)
```

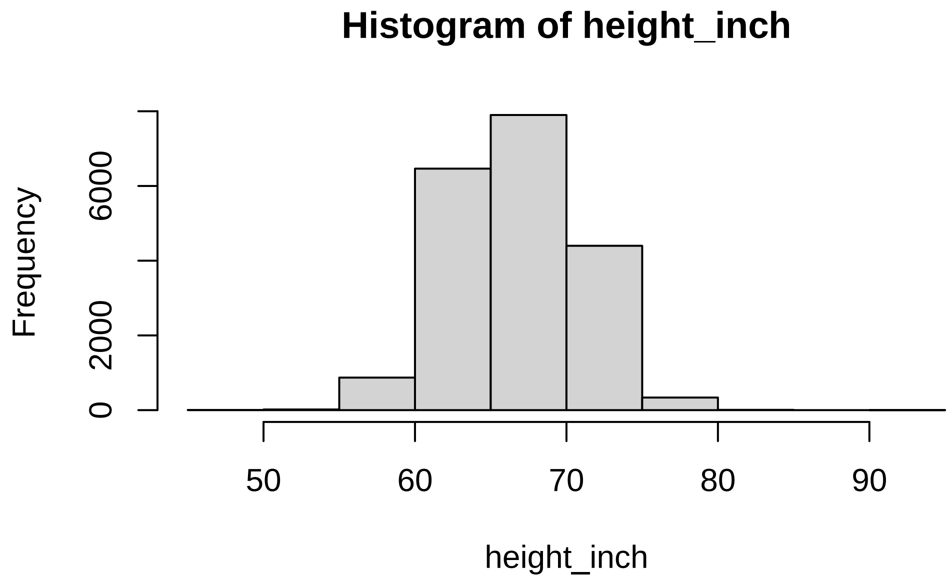
```
[1] 0.5553222
```

Side-Note: We can use the `cor()` function to calculate the Pearson correlation of height and weight. A correlation coefficient of **0.1** is thought to represent a weak or small association; a correlation coefficient of **0.3** is considered a moderate correlation; and a correlation coefficient of **0.5 or larger** is thought to represent as strong or large correlation.

Key-Point: When we use the $ notation we extract a vector from the data.frame

Many "base R" graphics functions work with vectors as input just like the plot() function.

```
height_inch <- cdc$height
hist(height_inch)
```



Creating new vector >Q5 Create a new object weight_kg that records weight in kilograms

1 in = .0254 m; 1 lb = .454 kg.

```
height_m <- cdc$height * 0.0254

weight_kg <- cdc$weight *0.454
```

Section 2: BMI

BMI is calculated as weight in kilograms divided by height in meters squared

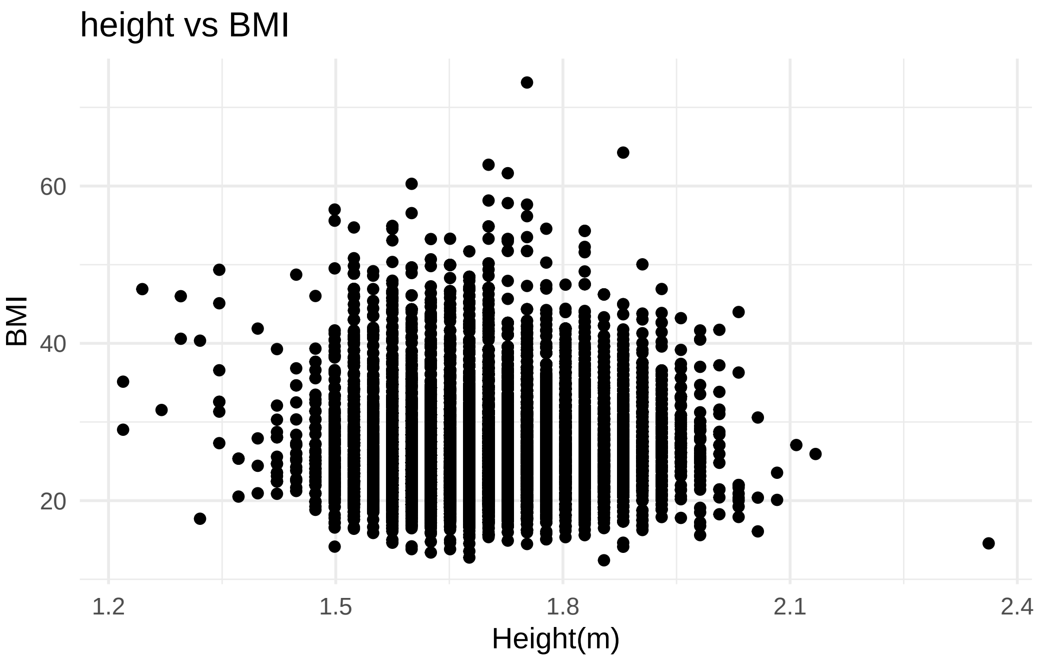
> Q6. Create a new object bmi and make a plot of height vs BMI. Provide your code below and comment on whether height and BMI seem to be associated?

create a new object "BMI"

```r
BMI <- (weight_kg)/(height_m^2)
```

make a plot of height vs BMI

```r
ggplot(cdc) +
  aes(x = height_m,
      y = BMI) +
  geom_point() +
  labs(title = "height vs BMI",
       x = "Height(m)",
       y = "BMI") +
  theme_minimal()
```



> Q7.Are height and BMI strongly associated? What are their correlation value?

```
cor(x = height_m, y = BMI)
```

```
[1] 0.03251694
```

Using logical vectors to count and subset

```
head(BMI >= 30, 100)
```

```
  [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE
 [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
 [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
 [61] FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [73] FALSE  TRUE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
 [85] FALSE FALSE FALSE FALSE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
 [97]  TRUE FALSE FALSE FALSE
```

A very useful trick that we will turn to over and over again is to treat logical vectors as equivalent to zero and one values. For example:

```
obese <- c(BMI >=30)
sum(obese)
```

```
[1] 3897
```

> Q8. Can you use this summing of a logical vector approach to find out how many obese individuals there are in the dataset?

```
sum(BMI>=30)
```

```
[1] 3897
```

To find the proportion of obese individuals we can use the following code: `sum()` here finds the number of obese people `length()`here finds the total number of participants

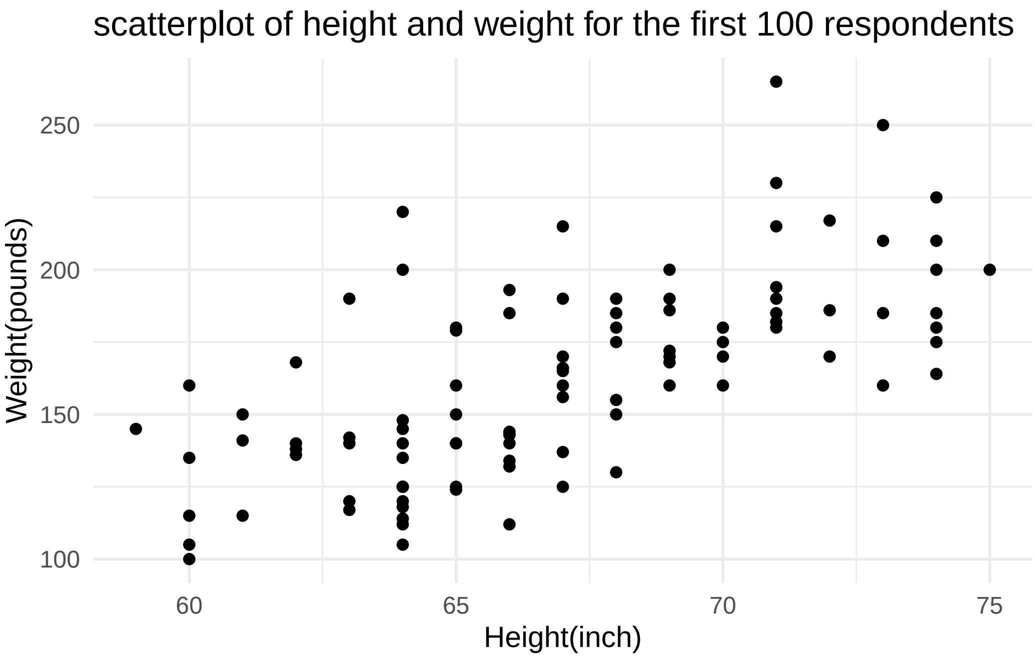
```
sum(BMI >= 30)/length(BMI)
```

```
[1] 0.19485
```

Section 3: Accessing subsets of data using row and column indices

> Q9. Use bracket notation to make a scatterplot of height and weight for the first 100 respondents.

```
cdc_first100 <- head(cdc, 100)
ggplot(cdc_first100)+
  aes(x = height,
      y = weight)+
  geom_point() +
  labs(title = "scatterplot of height and weight for the first 100 respondents",
       x = "Height(inch)",
       y = "Weight(pounds)")+
  theme_minimal()
```



Section 4 Using `table()` function

Q10. How many obese individuals are male in the full dataset?

- How do I first get at the data I need (i.e. the gender column)?
- Then how do I subset this vector to have only bmi >= 30 folks?
- Then finally, how do I pass this to the table() function to count up the m and f values.