R with public health data

Abhishek S D

11/05/2020

Basic R

1. To check which directory you are working in:

```
getwd()
```

```
## [1] "E:/R for Data science/Files/Statatistics-and-Data-Analysis-in-public-health"
```

2. To import the data. You need to change the "file" location to where you've stored the data set

```
g <- read.csv(file = "D:/Courses/R/cancer data for MOOC 1.csv",header = TRUE, sep = ',')
```

3. To have a look at the first few rows of our data set:

```
head(g)
```

```
##   patient_id age gender      bmi smoking exercise fruit veg cancer
## 1          1  61      0 20.79721       2        0     1   2      0
## 2          2  68      1 27.30000       0        0     0   1      0
## 3          3  62      1 22.18310       0        0     1   3      0
## 4          4  61      1 35.26846       2        0     2   4      0
## 5          5  58      1 22.67334       1        0     3   1      0
## 6          6  46      1 27.69035       1        0     0   2      1
```

4. To inspect the `age` variable:

```
g$age
```

```
##  [1] 61 68 62 61 58 46 67 68 53 59 72 45 58 44 58 42 48 68 82 71 84 68 64 62 56
## [26] 72 72 54 57 62 64 72 41 89 39 73 54 73 63 84 61 44 51 56 74 59 75 40 54 40
## [51] 53 62 48 75 55 70 55 60 73 59 56 60 70 68 69 46
```

5. To display a summary of the genders of our patients:

```
table(g$gender)
```

```
##
##  0  1
## 33 33
```

6. To display a summary of the BMI of our patient:

```
summary(g$bmi)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.81   21.41   24.78   24.22   27.30   40.62
```

7. To display a summary of the smoking status of our patients:

```
table(g$smoking)
```

```
##
##  0  1  2
## 26 18 21
```

8. To display a summary of the exercise status of our patients:

```
table(g$exercise)
```

```
##
##  0  1  2
## 22 22 22
```

9. To create a new variable `fruitveg`, which sums the daily consumption of fruit and veg of each patient:

```
g$fruitveg <- g$fruit + g$veg
```

10. To display a histogram of the ages of our patients:

```
hist(g$age)
```

11. To create a new binary variable **five_a_day**, whether the patient eats at least 5 fruit or veg a day:

```
g$five_a_day <- ifelse(g$fruitveg >= 5, 1, 0)
```

12. To summarise the **five_a_day** variable:

```
table(g$five_a_day)
```

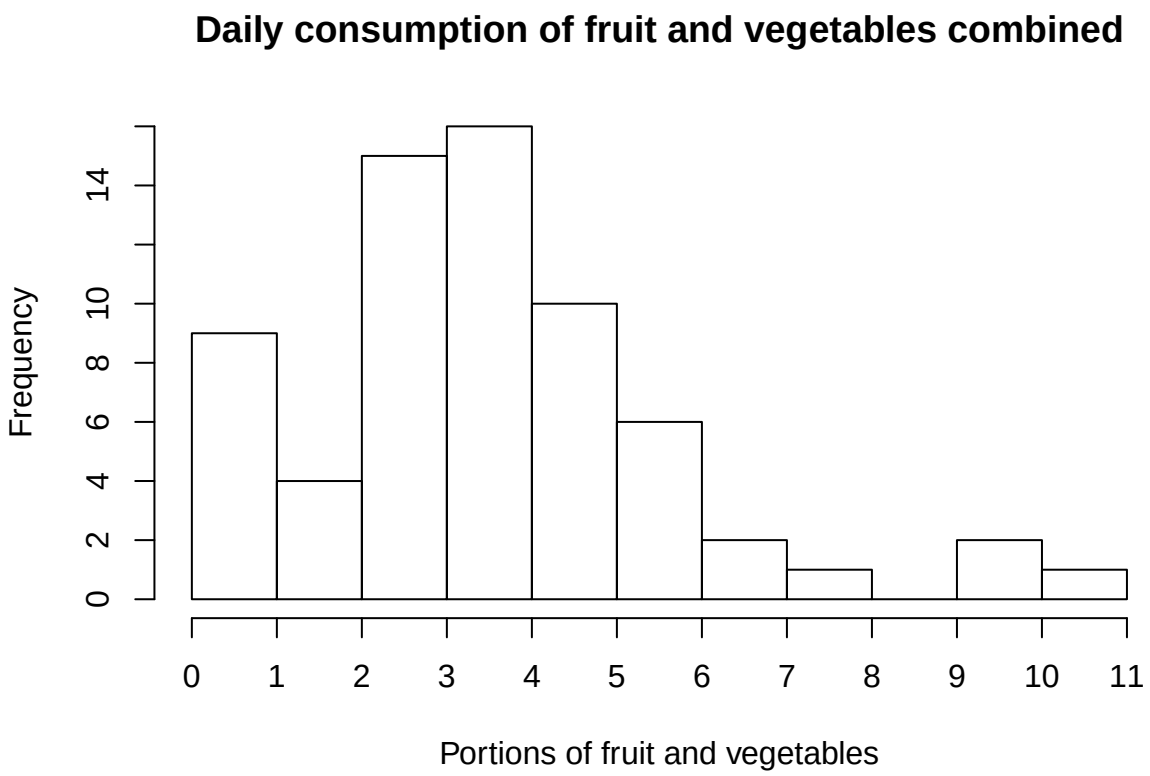
```
##
##  0  1
## 44 22
```

13. To display a histogram of the daily fruit and veg consumption of our patients, including a title and proper axes:

```
hist(g$fruitveg, xlab = "Portions of fruit and vegetables", main = "Daily consumption of fruit and veget

axis(side = 1, at = seq(0, 11, 1))

axis(side = 2, at = seq(0, 16, 2))
```



14. To create a new binary variable **healthy_BMI**, whether the patient has a healthy BMI or not:

```
g$healthy_BMI <- ifelse(g$bmi > 18.5 & g$bmi < 25, 1, 0)
```

15. To summarise **healthy_BMI**:

```
table(g$healthy_BMI)
```

```
##
##  0  1
## 40 26
```

16. To run a chi-squared test to look for an association between eating five or more fruit and veg a day and cancer:

```
chisq.test(x = g$five_a_day, y = g$cancer)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  g$five_a_day and g$cancer
## X-squared = 2.4265, df = 1, p-value = 0.1193
```

16. To run a (two-tailed) t-test to see whether the mean BMI of those with cancer is different from the mean BMI of those without cancer:

```
t.test(g$bmi ~ g$cancer)
```

```
##
##  Welch Two Sample t-test
##
## data:  g$bmi by g$cancer
## t = 0.90034, df = 21.878, p-value = 0.3777
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.735200  4.396204
## sample estimates:
## mean in group 0 mean in group 1
##         24.5198         23.1893
```

17. To run a (two-tailed) t-test to see whether the mean BMI of those with cancer is different from the mean BMI of those without cancer, where the variances are equal:

```
t.test(g$bmi ~ g$cancer, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  g$bmi by g$cancer
## t = 0.92959, df = 64, p-value = 0.3561
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.528819  4.189823
## sample estimates:
## mean in group 0 mean in group 1
##         24.5198         23.1893
```

18. To run a t-test to see whether the mean BMI of all patients is different from 25:

```
t.test(g$bmi, mu = 25)
```

```
##
##  One Sample t-test
##
## data:  g$bmi
## t = -1.3061, df = 65, p-value = 0.1961
## alternative hypothesis: true mean is not equal to 25
## 95 percent confidence interval:
##  23.02077 25.41406
## sample estimates:
## mean of x
##  24.21742
```

19. To run a chi-squared test to see whether there is an association between eating five or more fruit a day and having cancer:

```r
chisq.test(x = g$five_a_day, y = g$cancer)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  g$five_a_day and g$cancer
## X-squared = 2.4265, df = 1, p-value = 0.1193
```

20. To create a new binary variable, whether overweight or not according to their BMI:

```r
g$overweight <- ifelse(g$bmi >= 25, 1, 0)
```

21. To summarise the `overweight` variable:

```r
table(g$overweight)
```

```
##
##  0  1
## 34 32
```

22. To run a chi-squared test to see whether there is an association between being overweight and cancer:

```r
chisq.test(x = g$overweight, y = g$cancer)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  g$overweight and g$cancer
## X-squared = 0.20625, df = 1, p-value = 0.6497
```

That's the end of basics in R.