Introduction to Educational Statistics

Assignment #4 - Computer Lab

Exercise 1

- How many cases are there in this dataset? 20000
- How many variables? 9
- For each variable, identify its data type
    - Genhlth: ordinal
    - Exerany: nomial
    - Hlthplan: nominal
    - Smoke100: nominal
    - Height: continuos
    - Weight: continuous
    - Wtdesire: discrete variable
    - Age: discrete variable
    - Gender: nomial

Exercise 2

Create a numerical summary for height and age, and compute the interquartile range for each. Compute the relative frequency distribution for gender and exerany. How many males are in the sample? What proportion of the sample reports being in excellent health?

Var	Min.	1st Qua.	Median	Mean	3rd Qua.	Max.	**Inter Qua.**
Height	48	64	67	67.18	70	93	**6**
Age	18	31	43	45	57	99	**26**

- Compute the interquartile range for each:
    - Height: 6

    - Age: 26

- Compute the relative frequency distribution for gender and exerany:

    - table(cdc$gender)/20000: **m(0.47845) - f (0.52155)**

    - table(cdc$exerany)/20000: **0(0.2543) - 1 (7457)**

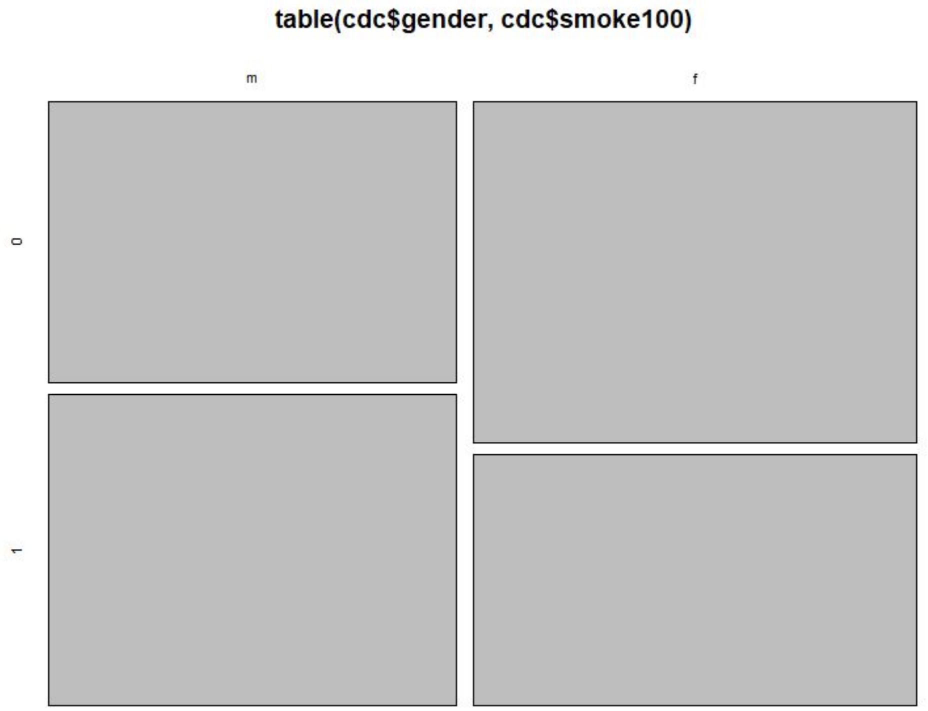
- How many males are in the sample? table(cdc$gender) => **m = 9569**

- What proportion of the sample reports being in excellent health?

    - table(cdc$genhlth)/20000 => **excellent = 0.23285**

Exercise 3

What does the mosaic plot reveal about smoking habits and gender? Males tend to smoke more than females



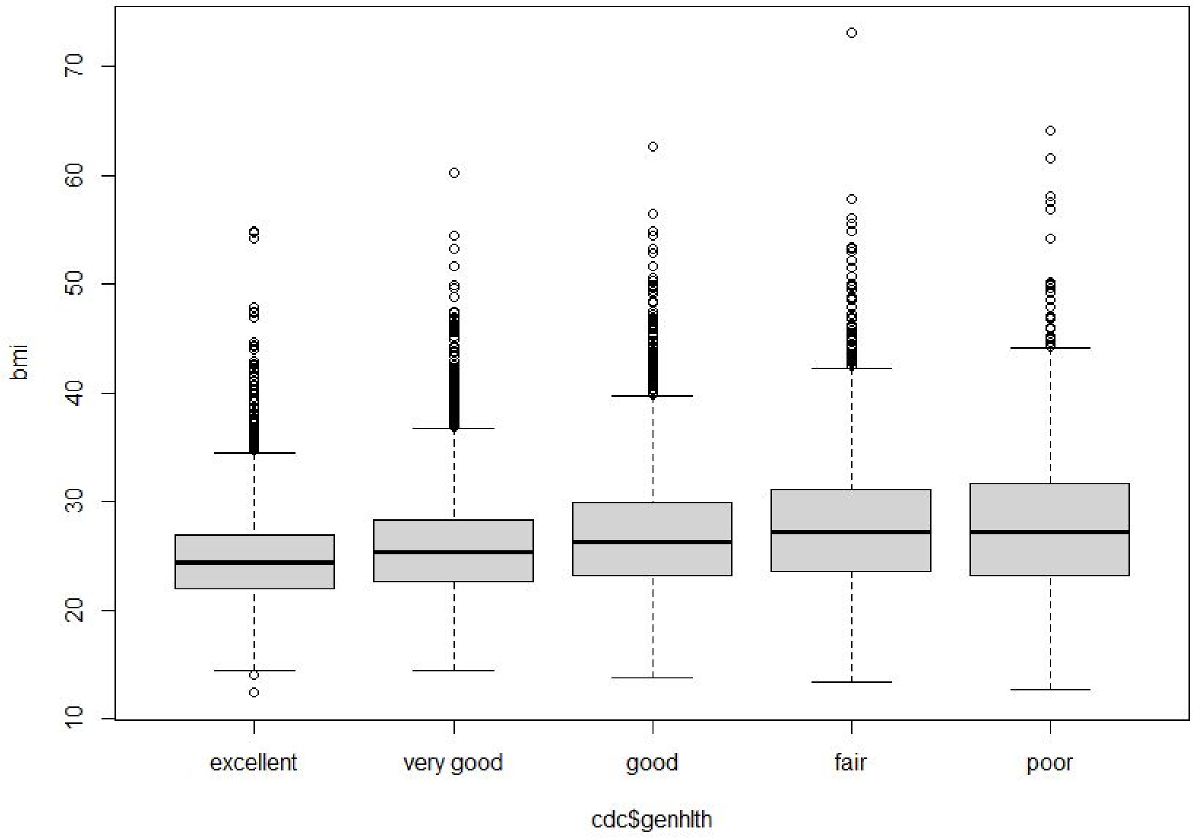
Exercise 4

Create a new object called under23 and smoke that contains all observations of respondents under the age of 23 that have smoked 100 cigarettes in their lifetime

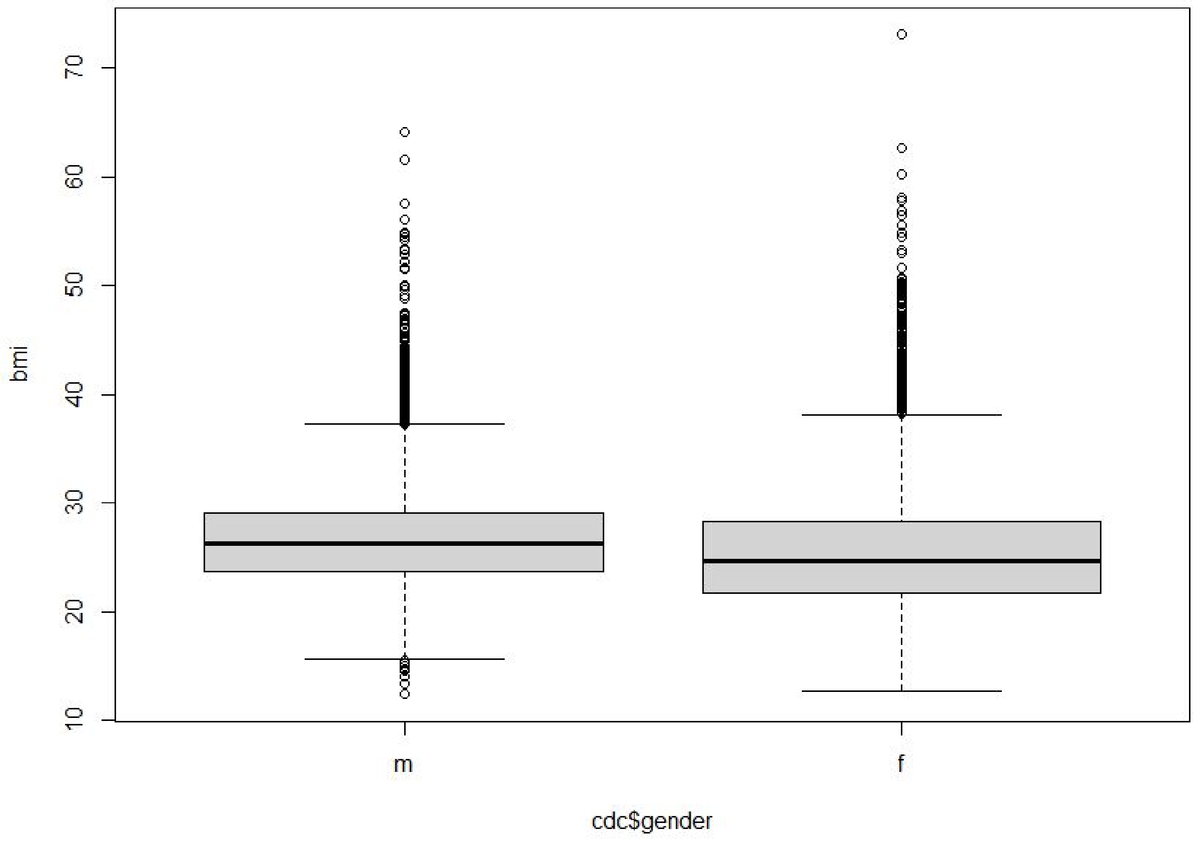
Answer: > under23_and_smoke <-subset(cdc,cdc$smoke100==1 & cdc$age <23)

Exercise 5

What does this box plot show? It compares the distribution of numerical values across categories.



Pick another categorical variable from the dataset and see how it relates to the BMI: boxplot(bmi ~ cdc$gender). Similar to previous plot, there are a lot of extreme values in the second plot


bmi
m
f
cdc$gender