Fall 2016
MIS 413/572 - Introduction to Big Data Analytics

Exercise 1

Please create an R data frame from below vectors. Use any R packages/functions to answer the following data management questions. DO NOT use any R loop statements (e.g. *for* and *while*).

```
age = c(20,26,17,40,53,34,57,32,53,38,NA,65,39,27,19,63,69)
height = c(165,177,158,168,179,182,164,187,163,NA,172,173,168,153,169,175,189)
weight = c(53,57,48,67,75,NA,46,49,52,77,42,93,70,65,59,74,98)
```

1) Remove those observation(s) with any NA.

2) Write an R function that finds a mode (眾數) of a given vector.

3) Use your mode function and *Map*() / *apply*-family functions to find the mean and the mode of each column.

4) Create a new categorical variable *age_cat* by discretizing *age* with intervals “0-20” ,”21-40”, and “above 40”.

5) Add a new column *BMI*, where $BMI = \frac{weight(kg)}{height(meter)^2}$.

6) Create a new categorical variable *BMI_cat* by discretizing *BMI* with intervals on course slides *Unit 2*.

7) Create a crosstab of *age_cat* by *BMI_cat*.

8) Perform any simple statistical test of independence on the crosstab and see if there is any *connection* between the *age* and *BMI*. Did you find anything interesting?

9) Identify those people/observation(s) whose *age*, *height*, and *weight* are <u>all</u> in the 4th quartile.