Homework 2

MTH 3270 Data Science

Due Mon., Wed. Feb. 12

Read These Chapters of the Book	**Then Do These Exercises**
Appendix B 2	Problems 1-7 (below) 2.2*, 2.5**, 2.6***

* For Problem 2.2, the url is:
`http://www.nytimes.com/2012/04/15/sunday-review/coming-soon-taxmageddon.html`

The two graphs are accessed via links at the bottom of the page.

** For Problem 2.5, the url is:
`http://mdsr-book.github.io/exercises.html#exercise_25`

*** For Problem 2.6, the url is:
`http://tinyurl.com/nytimes-unplanned`

1

a) Write R commands that create three vectors:

- `TrueAndMissing` containing values `TRUE` and `NA` (at least one of each in any order).
- `FalseAndMissing` containing values `FALSE` and `NA`.
- `Mixed` containing values `TRUE`, `FALSE` and `NA`.

Report your R commands.

b) Apply the functions `any( )` and `all()` to each of the vectors of part *a* and report the results.

2 Consider the following data on populations, illiteracy rates, and murder rates for 10 states of the United States: population in thousands, percent illiteracy, and murders per 100,000 population.

State	Population	Illiteracy	Murder
Alabama	3615	2.1	15.1
Alaska	365	1.5	11.3
Arizona	2212	1.8	7.8
Arkansas	2110	1.9	10.1
California	21198	1.1	10.3
Colorado	2541	0.7	6.8
Connecticut	3100	1.1	3.1
Delaware	579	0.9	6.2
Florida	8277	1.3	10.7
Georgia	4931	2.0	13.9

First create the following vectors:

```
illit <- c(2.1, 1.5, 1.8, 1.9, 1.1, 0.7, 1.1, 0.9, 1.3, 2.0)
murder <- c(15.1, 11.3, 7.8, 10.1, 10.3, 6.8, 3.1, 6.2, 10.7, 13.9)
pop <- c(3615, 365, 2212, 2110, 21198, 2541, 3100, 579, 8277, 4931)
state <- c("Alabama", "Alaska", "Arizona", "Arkansas", "California",
             "Colorado", "Connecticut", "Delaware", "Florida", "Georgia")
```

Now write commands involving the comparison operators (`>`, `<`, `==`, etc.) and either square brackets `[ ]` or the `subset()` function that do the following:

a) Return the names of the states whose populations are less than 2,500 (thousand).

b) Return illiteracy rates that are greater than the median illiteracy rate. Use `median()` to get the median illiteracy rate.

c) Return the murder rates for states whose illiteracy rate is greater than the median illiteracy rate.

Report your R commands.

3 Consider again the data on murder rates from the previous exercise.

a) Write a command involving `which()` that determines the indices of the murder rates that are greater than 12.

b) The following command does the same thing as `which(murder > 12)`:

```
(1:10)[murder > 12]
```

Why do you think the parentheses are included in the expression? Experiment a little.

4 For each of the following `"logical"` expressions, guess whether it will evaluate to `TRUE` or `FALSE`, then check your answer:

a) `(4 > 5 & 8 < 9) | 2 > 3`

b) `4 > 5 & (8 < 9 | 2 > 3)`

c) `(4 > 5 & 2 > 3) | 8 < 9`

d) `(4 > 5 & (2 > 3 | 8 < 9)) | 7 < 6`

5 An exception to the rule that `NA`s in operators necessarily produce `NA`s is given by the `&` and `|` operators.

a) What are the values of:

```
NA | TRUE
FALSE & NA
```

Why?

b) What are the values of:

```
NA | FALSE
TRUE & NA
```

Why?

6 Consider the vector:

```
x <- c(12, 4, 8, NA, 9, NA, 7, 12, 13, NA, 10)
```

Recall that `is.na(x)` indicates whether or not each element of `x` is a missing value (`NA`). Use `is.na()`, square brackets `[ ]`, and `!` (R's version of "not") to write one or more commands that extract all the *non-missing* values (non-`NA`'s) from `x`. For this problem, do not assume that you know where in `x` the `NA`s are.

7 The ***geometric mean*** can be defined as the nth root of the product of n positive numbers $x_1, x_2, \ldots, x_n$, i.e.

$$\text{Geometric mean} \;=\; (x_1 \times x_2 \times \cdots \times x_n)^{\frac{1}{n}}$$

(i.e. `prod(x)^(1/length(x))`, where `x` is a vector of positive values.)

Write a function `gm()` that takes a vector argument `x` containing positive numbers and returns their geometric mean, but gives appropriate feedback when the input is not numeric or contains non-positive values (use `if(!is.numeric(x) | any(x <= 0))` followed by `stop()`).

Report the R code for your `gm()` function.

Test your `gm()` function by passing it *a*) a numeric vector of positive values, *b*) a vector containing one or more non-positive values, and *c*) a character vector.