Homework 1
STAT 4410/8416 Section 001
Fall 2014
Due: September 12, 2014 by midnight

1. (a) What is data science?
   (b) Explain with an example what you mean by data product.
   (c) Carefully read the Cleveland's paper shown in lecture 2 and discuss what he suggested about the field of statistics and data science.
   (d) Explain in a short paragraph how data science is different from computer science.

2. In our **R** class we created the following function to get the square of a number. The function is written such that it gives us a text output `Big number` if the input is more than 100.

```
getSquare <- function(x){
  if(x>100)
    return("Big number") else
      return(x^2)
}
```

   We checked that the function is working as expected since we have

```
getSquare(5)

## [1] 25

getSquare(500)

## [1] "Big number"
```

   But the function does not work as expected for the following case. Instead of giving 'Big number' as an output it provides the actual square.

```
x <- c(25,200)
getSquare(x)

## [1]   625 40000
```

   Explain what is going wrong here. Also give a solution of this problem.

3. Write a program that will do the following. Include your codes and necessary outputs to demonstrate your work.
   (a) Generate 90000 random numbers from an exponential distribution with mean 30 and store these numbers in a vector called myVector. **Report** a histrogram of the numbers you just generated.
   (b) Convert myVector into a matrix of 900 columns and assign it to an object called myMatrix. **Report** the dimension of myMatrix.
   (c) Compute the column means of myMatrix. **Report** a histogram of those column means.
   (d) Explain why the two histograms you have created in questions (3a) and (3c) are different in shapes.

4. What are the very first few steps one should do once data is loaded onto **R**? Demonstrate that by loading tips data from http://www.ggobi.org/book/data/tips.csv