**Assignment 2**
**Submission deadline: Friday Sep 23, 2016 by 8 PM**
**Submission format: upload document in Canvas**

1. **(20 points)** Consider the Iris data set from assignment 1 problem 3. Construct the following plots in R.

(a) Plot of Petal Length (x-axis) vs Petal Width (y-axis). Briefly describe the relation between petal length and petal width as you observe from the plot.
(b) Plot of Petal Length (x-axis) vs Petal Width (y-axis), with different colors for the different classes of plants.
(c) Plot of Sepal Length (x-axis) vs Sepal Width (y-axis), with different colors for the different classes of plants.
(d) Observing the plots in (b) and (c), if you had to distinguish between classes by using either petal dimensions or sepal dimensions, which one would you choose --- petals or sepals, and why?

2. **(20 points)** Consider the babynames data from assignment 1 problems 4,5.

(a) Create a subset of the data with female babies named “Mary” from 1880-2014.
(b) Create a subset of the data with female babies named “Sophia” from 1880-2014.
(c) Construct a plot of the proportion of female babies named “Mary” from 1880-2014. On the same plot, add/overlay a plot of the proportion of female babies named “Sophia” from 1880-2014.
(d) Briefly describe your interpretation of the plot.

**Hint:** You can use points() or lines() to overlay two or more plots in R, i.e., two or more y-series with the same x-series. Run the following code for illustration.

```
plot(x=1:100,y=sin(1:100/5),col="red",ylab="",xlab="")
points(x=1:100,y=cos(1:100/5),col="blue")
lines(x=1:100,y=cos(1:100/5)^2,col="green")
```

3. Run the following code segments in R, describe what is wrong and how to correct it. Report only the R code with correction, do not report output.
(a) **(10 points)**

```
a = rnorm(100)
b = sqrt(a)
print(b)
```

(b) **(10 points)**

```
a = c("1","2","3","4","5")
b = mean(a)
print(b)
```

4. **(20 points)** Load the anscombe dataset in R. (Hint: data.anscombe = anscombe)

(a) Fit linear regression of (i) y1 on x1 (ii) y2 on x2 (iii) y3 on x3 and (iv) y4 on x4. Write down the four fitted regression lines.
(b) Construct the following plots (i) y1 vs x1 (ii) y2 vs x2 (iii) y3 vs x3 and (iv) y4 vs x4.
(c) Use your judgement and describe the discrepancy between the plots and the regression lines.

5. **(20 points)** For each of the four cases in the anscombe dataset, explain whether a linear regression model is appropriate.
   **Hint:** Look at regression diagnostics like plot of residuals against x, plot of leverage (or influence), etc.

**Assignment instructions:**

1. **Honor code:** The Virginia Tech honor pledge for assignments is as follows:
   **"I have neither given nor received unauthorized assistance on this assignment."**

   The pledge is to be written out on all graded assignments at the university and signed by the student. Type up your name to sign.

2. Submit your assignment as a document (word, pdf or similar) to Canvas, clearly marked with student's name and assignment number, eg. Sengupta_Srijan_HW2.pdf. Your submission should include R code and answers to problems.

3. Late assignments will not be accepted. Check Canvas regularly for assignments and submission dates.

4. You are free to discuss assignment problems with your classmates, but submitted work (answers and codes) **must** be your own work. Students are not allowed to copy computer codes or answers from each other, and must write their own codes and answers.